POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Dominik Filipiak

Exploring features of paintings for modelling preferences of buyers on the Polish art market

Badanie cech obrazów na potrzeby modelowania preferencji nabywców na polskim rynku sztuki

Doctoral dissertation

PhD Supervisor: Prof. Dr. hab. Witold Abramowicz

Auxiliary PhD Supervisor: Dr. hab. Agata Filipowska, prof. UEP

Date of submission:

Supervisor's signature

Institute of Informatics and Quantitative Economics, Department of Information Systems

Poznań 2022

Rodzicom

Acknowledgments

First of all, I would like to thank prof. Witold Abramowicz for guiding me as a supervisor during all my years spent in Department of Information Systems. I would also like to thank my secondary supervisor, dr hab. Agata Filipowska for encouragement, motivation, and numerous late evening conversations, during which many of presented ideas were conceived.

I would like to express my gratitude to my fiancée Ada, since this dissertation would not be possible to finish without her unconditional support and encouragement during the writing. Finally, I would like to thank my parents, Zbigniew and Elżbieta, to whom this dissertation is dedicated.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

cGAN Conditional Generative Adversarial Network

CIE Commission Internationale de l'Éclairage

CLD Colour Layout Descriptor

CMY Cyan, Magenta, Yellow

CMYK Cyan, Magenta, Yellow, Black

CQD Colour Quantisation Descriptor

CSD Colour Structure Descriptor

CV Computer Vision

DCD Dominant Colour Descriptor

DCT Discrete Cosine Transform

GAN Generative Adversarial Network

GoF/GoP Group of Frames/Group of Pictures

HMMD Hue-Max-Min-Diff

HR Hedonic regression

HSI Hue, Saturation, Intensity

HSL Hue, Saturation, Lightness

HSV Hue, Saturation, Value

MMCQ Modified Median Cut Quantisation

MPEG Moving Picture Experts Group

MSE Mean Squared Error

OLS Ordinary Least Squares

PCA Principal Component Analysis

PSNR Peak Signal-to-Noise Ratio

RGB Red, Green, Blue

RMSE Root Mean Squared Error

RSR Repeated-sale regression

SCD Scalable Colour Descriptor

SOM Self-Organising Map

sRGB Standard Red, Green, Blue

Chapter 1

Introduction

Plattner (1998) has written about *a market where producers don't make work primarily for sale, where buyers often have no idea of the value of what they buy, and where middlemen routinely claim reimbursement for sales of things they've never seen to buyers they've never dealt with.* Being described by these peculiar features and paradoxes, the art market managed to attract the attention of scholars in the domain of quantitative economics decades ago (Frey & Pommerehne, 1989). Since the fall of the Polish People's Republic in 1989, the Polish art market is growing fast – Artnet put Desa Unicum among the top 25 European auction houses (considering their total sales value of paintings and sculptures) in 2021 at the 8th position[1]. There is no single definition of art. It can be of an arbitrary form, with examples ranging from drawings and sculptures to conceptual and performance arts. This study is devoted to the particular form of visual art: paintings. This dissertation tries to investigate preferences of buyers and price determinants for paintings in the Polish Art market – using traditional quantitative methods, as well as more modern ensemble machine learning techniques. Special attention is given to colour-related features, as we hypothesise that the used colours can make a difference in the hammer price.

1.1 Motivation and Contribution

Leaving its *invaluable* features aside, art is an important class of alternative investments. Whether as planned portfolio diversification or from genuine interest, fine art attracts buyers for decades.

[1] `https://www.bankier.pl/wiadomosc/Desa-wsrod-najwiekszych-domow-aukcyjnych-w-Europie-8279938.html` [accessed on March, the 26th, 2022]

This places art among stocks, bonds or gold and makes it a particularly interesting asset class during uncertain times – worldwide and in Poland. The Polish art market is relatively young. It can be characterised by its huge potential for fast growth, which is demonstrated by the recent report figures (artinfo.pl, 2019). The same report states that 302 auctions were held in 2018, reaching 252,000,000 PLN turnover (+17.7% annual increase compared to 2017 with 284 auctions). Compared to 2001 with 47 auctions, an increase can be clearly observed. The number of auction participants in Poland can be attributed to the fact that the market attracts professional investors, as well as casual middle-class buyers.

Since investors allocate their capital to expensive paintings and subsequent price records are covered by the press, a set of natural questions arises – what factors drive hammer prices? What is the rate of return? Is it better to invest in art, or maybe would it be safer to buy stocks or bonds? Naturally, these questions make the art market a subject of interest for economists. In terms of *Journal of Economic Literature* (abbreviated as JEL) classification, quantitative art market studies fall into at least two categories. Cultural economics is a branch of economics devoted to studying general works of art and their influence on the economy. Its subclass, *economics of the arts and literature* is recognised with its own code (JEL Z11). Since this dissertation is concerned with price determinants of paintings, it falls into this category. Since art is often compared to the other forms of investments, quantitative art market research is usually heavily concerned with price indices. This falls to the subcategory of econometric and statistical methods: *Index Numbers and Aggregation* (JEL C43) – however, the indices are treated rather instrumentally in this work. This work also contributes to minimising the phenomena of information asymmetry. Entering art markets require a very specific domain knowledge, which might hamper potential buyers – the Polish one is no different. Identifying price determinants would certainly facilitate making decisions and understanding the market in general.

This work contributes to the existing body of knowledge in several ways. In general, a handful of papers are devoted to the quantitative study of the Polish art market. Several papers analysing different approaches to hedonic regression using the data from 2007-2010 have been published (Kompa & Witkowska, 2013; Witkowska, 2014; Lucińska, 2013, 2015). Witkowska and Lucińska (2015) also examined a sample with an extended period (2007-2013) in terms of different types of art market indices. More recently, Białowąs, Potocki, and Rogozińska (2018), as well as Szyszka and Białowąs (2019) conducted research on the largest dataset, which consist of observations from 1989-2012. However, the collected dataset allows only for conducting repeated-sales regression,

which does not directly analyse price determinants. All these datasets are not available publicly – the data can be compiled manually from auction house pages and catalogues. In this work, we prepared two new datasets - the Top 10 Painters dataset, which considers the most popular painters in the considered years, and the Young Art dataset. As the name suggests, the latter consists of paintings of young artists, which is an important segment of the Polish art market. Even though lots at these auctions start at relatively low prices, the growing importance of such works is manifested by the increasing number of auctions and their turnover (artinfo.pl, 2019).

To this date, a small number of publications concerned with colour-related features for quantitative analysis of painting prices has been published (Stepanova, 2015; Pownall & Graddy, 2016; Charlin & Cifuentes, 2018). However, to the best of our knowledge, this work is the first one that evaluates different algorithms used for colour quantisation in paintings to minimise the resulting error and maximise colour diversity. This is also the first work to investigate the importance of colour-related features of art using explainable artificial intelligence methods. Finally, this work employs a state-of-the-art machine learning algorithm for the art market analysis – XGBoost. While normally used for predictive analysis tasks, here it is paired with explainable artificial intelligence methods, such as PDP-plots or SHAP.

1.2 Problem, Thesis, and Research Questions

With the aforementioned motivation from Section 1.1, the **goal** of this dissertation is to explore features of paintings in order to model buyers' preferences on the Polish art market In this dissertation, this is understood as the description of the sold lots themselves (since they were chosen by their buyers), as well as exploration of the range of factors that influence the hammer price. Some scholars already tried to understand the important qualities behind paintings prices in the Polish art market using quantitative methods (Kompa & Witkowska, 2013; Lucińska, 2015; Witkowska & Lucińska, 2015). However, the topic of the influence of colour-related features on the Polish art market remains unexplored, to the best of our knowledge.

Therefore, the main **research problem** is: *Not knowing which features (in particular colour-related visual features) of paintings are important for modelling preferences of buyers on the Polish art market.* As mentioned earlier, some scholars have already proven the significance of particular features of paintings. However, the problem of colours was not yet tackled in the context of the Polish Art Market.

The main research thesis (in which we focus on the novelty of this research) is formulated as follows:

Thesis 1 (T1): *An application of colour quantisation with Algorithm 18 in order to extract features of paintings increases the explained variance of models representing buyer's preferences and price determinants on the Polish art market.*

Colour quantisation is understood here as a process of reducing the number of colours by unifying similar ones in order to facilitate the economic analysis of paintings (see Chapter 3). This leads to a problem of comparing and selecting the right quantisation method for this task, which is later applied to paintings analysis using Algorithm 18 (also called Artefact 1 in a Design research sense – see Section 1.3). The thesis is evaluated using the (adjusted) coefficient of determination in linear models and validated with XGBoost models with variable importance assessment techniques. The impact of overall colourfulness is also tested.

In order to solve the research problem, the following *research questions* needs to be posed:

Q1: Which methods can be used to assess the importance of paintings' features for the hammer price?

Q2: How to extract colour-related information from paintings?

Q3: What is the best colour quantisation algorithm for paintings?

Q4: Which features of paintings are important for buyers on the Polish art market?

To answer these questions and prove the thesis, we distinguish the following *research objectives*:

O1: Prepare datasets allowing conducting the experiment.

O2: Develop a method for extracting colour-related features from paintings (Artefact 1).

O3: Evaluate the method for extracting colour-related features on Polish art market data.

O4: Discuss which features of paintings are important for buyer's preferences on the Polish art market.

O5: Discuss price determinants for paintings on the Polish art market.

Table 1.1 maps chapters of this dissertation to the corresponding research questions and objectives. More details on the structure itself are provided in Section 1.4.

1.3 Methodology

Since the presented thesis is a juxtaposition of quantitative economics and data science, this dissertation benefits from various research frameworks. The relation to the economics was already

Table 1.1: Chapters of this dissertation with corresponding research questions and objectives.

	Research questions	Research objectives
Chapter 1	–	–
Chapter 2	Q1	–
Chapter 3	Q2	–
Chapter 4	Q1	–
Chapter 5	Q3	O1, O2
Chapter 6, Section 6.1	Q4	O1, O4
Chapter 6, Section 6.2	Q4	O3, O5
Chapter 7	–	–

Source: own study.

explained in the motivation (Section 1.1) – we explore buyer's preferences and price determinants for paintings sold in Polish auction houses, which is an important market for alternative investments. Finally, data science is understood here as an interdisciplinary field for extracting knowledge from data – some of its concepts (such as tree ensemble models, explanatory AI methods, or CRISP-DM process) are used to explain buyers preferences and price determinants on the Polish art market for paintings, whereas various feature extraction methods (such as colour quantisation) are used to quantitatively describe paintings. The rationale behind this work is to shed light on buyers preferences and price determinants of paintings sold in Polish auction houses. In general, this might put this research in the behavioural science context, as an explanation of human behaviour. While this is partially achieved by using already known methods (linear models, tree ensembles), the process of colour-oriented feature engineering is relatively uncommon in the literature (a few similar cases are mentioned in Section 3.4). Since we create artefacts in a form of a method (e.g. Algorithm 18) and models (see Section 6.2), this positions the research as design science. Such an intersection of economics and data science with a developed methods makes *design science for information systems* (Hevner et al., 2004) a good match in terms of research methodology.

As explained by Hevner et al. (2004), Design science (also stylised as Design-science) research guidelines consists of 7 steps. The first one is *design as an artefact*. An artefact is understood as a solution to an unsolved problem or an application of existing knowledge in an innovative way – not necessarily as an instantiation, but also as conceptual work. In this research, we deploy several artefacts. For instance, the algorithm for calculating the share of representative colours for paintings (Algorithm 18) is one of them. We also constructed several models for explaining the behaviour of the Polish market for paintings (in Section 6.2).

The second guideline is *problem relevance*. The artefacts are developed to resolve important scientific and business problems. The importance of our research goal and objectives has already been justified in Section 1.1, with a detailed theoretical foundations shown in chapters 2 3, and 4.

The third guideline (*design evaluation*) is focused on testing the utility, quality and efficacy of designed artefacts. Depending on a specific artefact, this is done by using several metrics. Mean squared error and colour diversity measures were used for evaluating a particular colour quantisation function. For the constructed models, mean squared error is also used, as well as R^2 and a couple more metrics. In terms of the notions used by Hevner et al. (2004), such a way of evaluation falls to the category of *controlled experiments*.

The fourth guideline considers *research contributions*. These have been clearly stated in Section 1.1. Hevner et al. understand research contributions as either the design artefact, foundations, or methodologies. Algorithm 18 matches the latter definition, as it combines advanced colour quantisation methods in order to explore buyers' preferences on the art market. Foundations include constructs, models and methods. The last category considers methodologies. In this dissertation only known methodologies are used and there are no contributions to this category.

The fifth guideline is *research rigour*. During our effort to search non-trivial but usable truths, we maintained scientific formalism whenever possible. At the same time, we try to enforce logical simplicity and epistemic soundness. The rigour was ensured by using design science principles with additionally employing CRISP-DM and research literature strategies, which are explained later in this section.

The sixth guideline is *design as a search process*. It refers to the iterative nature of the artefact development. It is reflected in our research – for example, in Section 5.1, in which we search for the most suitable colour quantisation algorithm. Similarly, we used hyperparameter tuning in order to build models in Section 6.2.

Finally, *communication of research* forms the last design science guideline and states that the results of the research process should not only be addressed to the scientific community, but also to technology-oriented and managerial audiences. It is manifested in the dissertation in chapter summaries and particularly in Chapter 7, which concludes the dissertation. All of these realisations of design science guidelines can be found in Table 1.2. The overall design science for information systems research framework is depicted in Figure 1.1 – here, the business needs can be understood as shaping the buyer's preferences.

Table 1.2: Design science research guidelines proposed by Hevner et al. (2004) followed by their realisation in the dissertation.

	Guideline	Realisation
1	Design as an artefact	Artefact 1 (Algorithm 18), models from Section 6.2
2	Problem relevance	Described in Section 1.1 and explored in chapters 2 3, and 4
3	Design evaluation	Controlled experiments in Chapter 5 and 6
4	Research contributions	Enlisted in Section 1.1 and Chapter 7
5	Research rigour	Methodologies nad Guidelines explained in Section 1.3
6	Design as a search process	Manifested in Section 5.1 and 6.2
7	Communication of research	Summary and conclusions in Chapter 7

Source: own study.

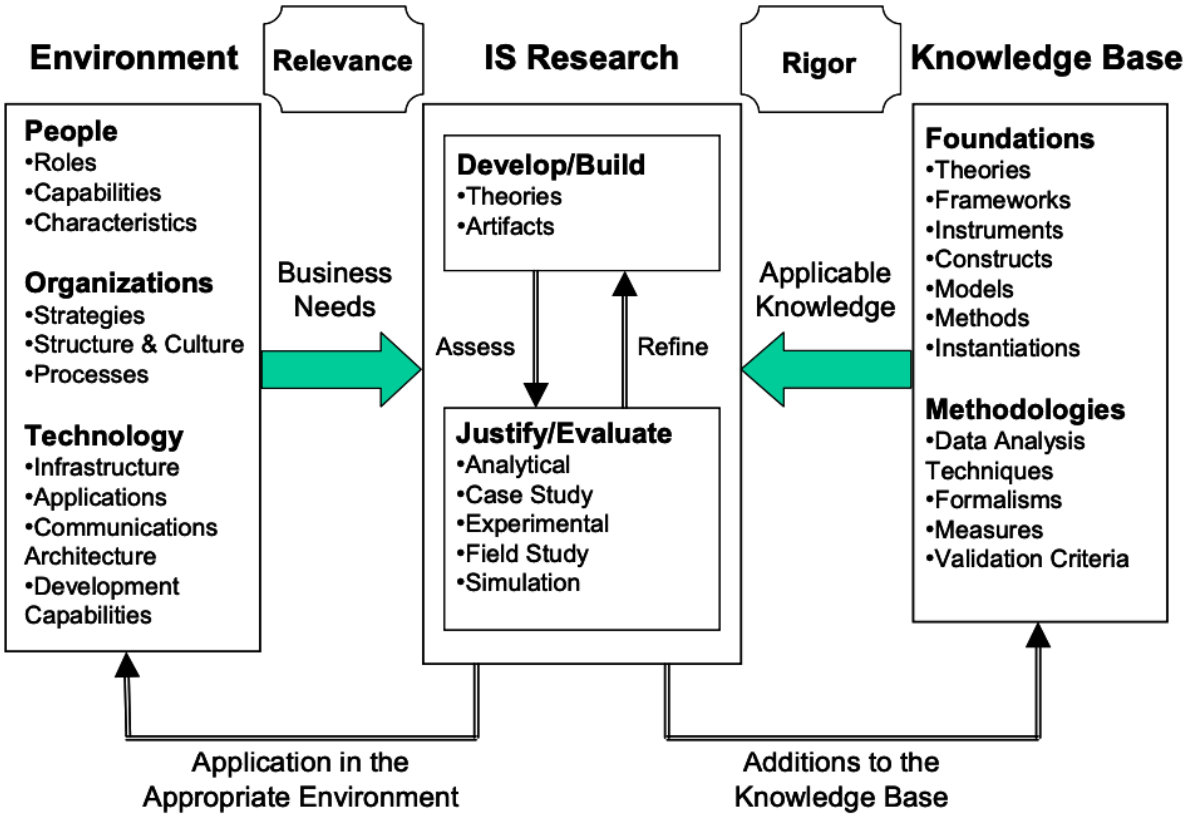


Figure 1.1: Design science for information systems research framework.
Source: Hevner et al. (2004)

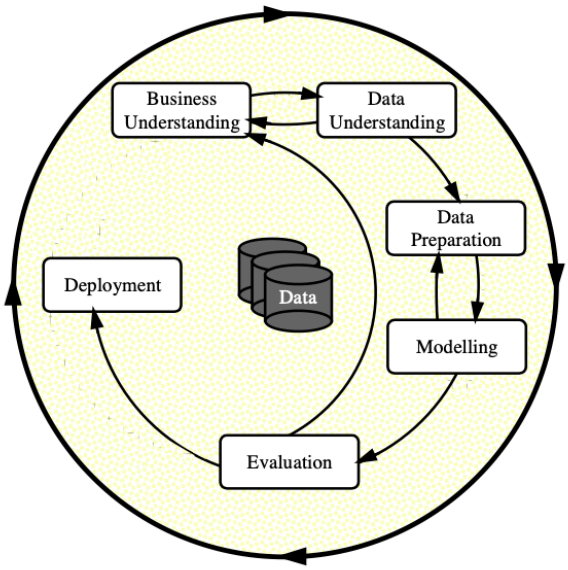


Figure 1.2: CRISP-DM process diagram.
Source: Levy and Ellis (2006)

Two additional sources of research guidelines are followed – *systems approach to conduct an effective literature review* and CRISP-DM. For conducting the literature review, we follow the principles outlined by Levy and Ellis (2006). Techniques such as keyword-, forward-, and backward search have been used throughout the body of knowledge available in the Internet databases, such as Google Scholar, SpringerLink, Elsevier Science Direct, and ACM digital library. The relevant articles were later analysed and synthesised, which resulted in chapters 2, 3, and 4. Introduced by Wirth and Hipp (2000), CRISP-DM (an abbreviation from CRoss Industry Standard Process for Data Mining) is a popular life cycle model for data mining activities. It is depicted in Figure 1.2. The important parts of the research process presented in this dissertation can be framed in this methodology. Originally, CRISM-DM consists of six phases. The *business understanding* phase focuses on the problem definition (e.g. Section 1.2), whereas *data understanding* provides first insights from the collected data (partially Section 6.1). *Data preparation* is concerned with constructing the final dataset (Section 5.2). *Modelling* and *Evaluation* are covered in Section 6.2, in which the experiments are performed and described. The final phase (*deployment*) is however beyond the scope of this work, as it is not needed in research (contrary to production, for which CRISP-DM was primarily made).

1.4 Structure of Dissertation

The dissertation is organised into 7 different chapters (including this one) and the appendix. Each of the chapters is concluded by a short summary (excluding this and the final ones). This chapter is devoted to the introduction of the research problem and the motivation beyond. It outlines the goals, provides the motivation and methodology for the research, and explains the structure of the dissertation. Regarding the rest, chapters 2, 3, and 4 survey the existing knowledge base in different domains, whereas chapters 5 and 6 provide new experiments, analyses, and findings. Chapter 7 summarises the dissertation.

Chapter 2 provides an important contextual information about quantitative methods used for art market research. Section 2.1 explores general mechanisms and structure of the art market – the global, as well as the Polish one. In Section 2.2, the traditional means employed for quantitative art market analysis are described. The analysis of the colour-related features needs a theoretical introduction, which is provided in Chapter 3. Section 3.1 offers a comprehensive overview of the notion of colour and its representation on digital devices, such as RGB and CIE $L^*a^*b^*$. Section 3.2 introduces the notions of features (in a image processing/computer vision sense) and descriptors. Section 3.3 provides a survey of popular colour quantisation and palette design algorithms. Finally, Section 3.4 surveys quantitative art market research focused on colour-related studies, as this dissertation aims at similar research goals. Chapter 4 is a comprehensive overview of modern tree-based machine learning techniques supplied with explainable artificial intelligence methods. Section 4.1 presents classic decision trees, ensemble methods and the XGBoost algorithm. Section 4.2 provides an overview of selected concepts related to explainable artificial intelligence, such as permutation importance, partial-dependence profiles nad SHAP values.

The remaining chapters offer new findings and form the main contribution of this dissertation. Before performing quantitative analyses of the Polish art market, Chapter 5 focuses on colour-related feature engineering methods. In Section 5.1, an evaluation of colour quantisation algorithms is presented. The algorithms are evaluated in terms of their mean squared error and resulting colour diversity. Section 5.2 presents two engineered features, which will be further used in the data analysis – colour share with Chang's k-means quantisation and colourfulness. Chapter 6 details the findings of the data analysis. Section 6.1 presents the statistics about the two analysed datasets (*Young Art* and *Top 10 Painters*). Section 6.2 presents the search for

important features of the paintings, which are understood as the key factors behind buyers preferences and hammer prices. Finally, Chapter 7 provides a summary, along with short conclusions and a future work outline. The dissertation is concluded by the appendix, in which additional tables and figures are placed.

Chapter 2

Quantitative Research on Fine Art Auctions

As Anselm Kiefer once said: *Art is difficult. It's not entertainment. There are only a few people who can say something about art – it's very restricted. When I see a new artist I give myself a lot of time to reflect and decide whether it's art or not. Buying art is not understanding art.* Yet there is a lot of people interested in buying art. Whether they are professional investors or casual art admirers, new participants seem to be attracted every year. This chapter is devoted to the quantitative perspective on buying and selling art – with a special focus on art auctions. It is concerned with answering the research question Q1 (*Which methods can be used to assess the importance of paintings' features for the hammer price?*). Section 2.1 presents the global art market, as well as its general mechanism and concepts fundamental for art trading. The Polish art auctions are discussed in Section 2.1.2, in which the major auction houses and local characteristics are presented. While experts' appraisal is the most popular and trusted way to put a price tag on a painting, a number of scholars have intensively explored the topic of price determinants for fine art. Traditionally, the art market was studied using regression methods used in econometrics. More recently, some researchers turned to algorithms known from data science in a broader sense, though regression analysis is still the most popular tool. The existing literature on this topic is presented in Section 2.2, which is directly related to the research question Q1. Finally, a short summary concludes this chapter.

2.1 Art Market

Britannica defines the art market as a physical or figurative venue, in which art is traded. Although this short definition covers the subject matter, it hides the complexity of this market. In the eyes of financial experts, several qualities make art unique. For instance, Zboroń (2018) argues that when we discuss the art market, we should consider *homo aestheticus* instead of *homo oeconomicus*, as the law of diminishing marginal utility seems to not hold for the art market. Żaglewska (2016) claims that the art market is tightly coupled with the cultural capital of buyers rather than supply and demand. When it comes to bubbles, art as an investment seems to be not that much different from other markets. Works of Damien Hirst are perhaps the most known and quite recent example of such a rapid increase in price due to speculative transactions[1]. The artist known for e.g. his stuffed animals in formaldehyde is widely considered as a marketing mastermind, which certainly had an impact on high prices. Eventually, the attraction of speculative investments caused the prices to plummet. All these peculiarities make the art market an interesting subject for economic research. This section explains the most basic mechanisms in the global market. Special attention is given to auctions, as they are the main subject of this dissertation.

2.1.1 Art Market in General

Art investment is a part of a broader category of alternative investments, which is often chosen to diversify an investment portfolio. Borowski (2013, 2015) argues that the market for alternative investments – so for the art as well – can be characterised by a number of qualities. First, there is the problem of *illiquidity* – even taking into account the advent of internet auctions, a good deal of sales still takes place during dedicated auctions or private gallery viewings, which is a serious obstacle for quick liquidation. Some say that this is the 3D market, which, in this context, is an abbreviation from death, debts, and divorce – three main factors of selling collectables at auctions. Other features of the alternative investment market are *problems with appraisal* (since there's no way to know the *real* market price of art) and *long investment horizon* (for example, this equals 10 years on average before selling for some art markets). This makes alternative investments an *irregular source of income*. Other features mentioned by Borowski are *requirement for domain knowledge* and *information asymmetry* (which partially results from

[1] `https://www.bloomberg.com/news/articles/2012-11-21/for-collectors-with-hirst-comes-pain` [accessed on August, the 2nd, 2020]

the former). An interesting quality of this market is the *occurrence of non-rational investors* – some participants are collectors, which are more interested in aesthetic qualities rather than high return. Alternative investments are *prone to trends*, such as the Impressionist bubble in the 1980s or the aforementioned Damien Hirst case. Finally, *problems with storage* implicate additional costs stemming from conservation, protection, and insurance.

Auction houses, art galleries, and individual art sales are traditionally the central places for the art market. The last two are sometimes referred to as private sales. Additionally, art is traded at special annual or biannual art fairs, which usually also get a lot of media coverage – such as the famous Art Basel fair. Nowadays, the growing importance of art fairs and especially online art trading can be observed. Some of the digital auctions are attributed to the legacy auction houses, however some entities work primarily or even exclusively online as well. Sometimes, a distinction between the primary and secondary market is introduced – the first one is concerned with works being sold for the first time, whereas the latter is for works being traded before at least once. In general, the works on the secondary market can be characterised with a lower risk, however they are less innovative than those on the primary one at the time of sale (Zorloni, 2005). Participants in the art market are often classified into sellers, buyers, and middlemen. Middlemen are (usually) individual or institutional art dealers. Sometimes, sellers on the primary market are the artist themselves, though many of them prefer to be represented by a gallery.

Art auctions are a significant part of the whole art market – TEFAF (2014) estimates they account for 47% of sales. Numerous scholars in their quantitative analyses focus on auctions, as they are the only source of data accessible in the Internet. On the contrary, private sales are virtually impossible to track in a scientifically rigorous way. Therefore, this work focuses only on auction houses. An art auction is held by an *auctioneer*. During an auction, a number of *lots* are offered. Following Sotheby's art market glossary[2], a lot is a single object (sometimes a group of objects), which is presented for sale at auction as a single unit. An auctioneer communicates what is the asking price for a given lot and waits for the offers. Then, a potential buyer (*bidder*) can signalise that they can pay the price – this is called a *bid*. Then, the auctioneer communicates the price needed to top the current one and waits for the next bid. The process is continued up to the point at which there are no new bids. After the bidding, a lot might be sold or *bought in* – the second term means that there were no bidders interested in obtaining it or it failed to meet the auction guarantees (i.e. the minimum price set up by the seller).

[2] `https://www.sothebys.com/en/glossary` [accessed on August, the 2nd, 2020]

One may think that the last bid is the final price to be paid for a given lot, but this matter is a bit more complicated. Discussing art appraisal and its determinants requires defining a number of different price types functioning in art markets. Everything starts with the *asking price* – this is the level from which the bidding starts. Sometimes, *estimations* are provided. These two figures indicate the expected price range for a given lot and usually are determined by experts with extensive art market knowledge. Although these numbers are only anticipation, sometimes they act as a strong predictor of the final price range (Habalová, 2018). The asking price does not have to be the lowest one at which a given lot can be sold – this is determined by the *reserve price* or *guarantee*. If a lot fails to reach this level during the bidding, the transaction is not guaranteed. This price is known to the seller and the auction house. The highest (winning) bid is called the *hammer price*. This price, however, does not contain additional fees and charges (*buyer's premium*) set by a given auction house. The price which includes that is called *premium price*, and it varies among auction houses, though it is often the hammer price increased by approximately 20%. Sometimes, other charges may apply, such as local taxes or *droit de suite*, which is a fee for the author of a given lot. Sales which takes place in the evening are considered to be more prestigious. In many art markets, December is traditionally the month of the highest turnover and the number of sold lots.

Even with an endless variety and non-homogenous nature of art, lots sold at auctions can still be characterised by a number of fixed features, rather applicable regardless of their form – starting from their author, which is usually the most important price determinant (Kräussl & Elsland, 2008; Borowski, 2015). Perhaps one of the most frequent parts of the description is medium and technique. Traditionally, *paintings* denote an artwork made on canvas, contrary to *works on paper*. The latter is an umbrella term for graphics, drawings, or sketches. Graphics can be further subdivided into other categories (such as etchings, lithographies, woodcuts...). In describing art, a technique commonly references a group of used materials (*oil on canvas* or *acrylic on paper*, for example). Other important forms of art sold at auctions are, for example, photography and sculptures. Sometimes, handicrafts (such as plates or cutlery) and furniture can be encountered as well. Cars and watches are the subjects of dedicated, special auctions. One of the most intrinsic features of a lot is its size. The larger artworks are sold for higher prices, which is supported by numerous econometric models (see the next sections). Lots can also be characterised by the presence of a signature (or lack of thereof) – sometimes with a date. Another important feature – provenance – denotes a history of ownership and location of a given

artwork. If a lot has been displayed at a prestigious institution, it is expected to impact its price positively. Provenance is an important feature for the transparency of the market. In practice, however, it is often difficult to collect information about previous owners (Gramlich, 2017).

In terms of the investment language, a painting, due to its uniqueness, can generally be considered as a heterogeneous good (Borowski, 2015). Things are a bit more complicated for works on paper – it might vary and actually needs introducing some additional definitions in terms of the technique of their production, which we will borrow from glossaries of printmaking terms[3]. Drawings and sketches on paper, for instance, are unique. *Prints* – defined as images impressed (usually on paper) using matrices using a process which can be repeated – are not. In the process of printmaking, preparing a *matrix* is an important step. It can be defined as a base (for example – wooden or metal) from which prints are made. Essentially, matrices are the objects crafted by artists. A single print is called an *impression*. A set of all impressions made without any change to the matrix is called a *state*. All impressions published at the same time form an *edition*. First impressions are traditionally perceived as better ones and therefore they are more expensive, though in practice their quality might vary. In general though, works on paper are much cheaper compared to paintings due to their general non-uniqueness and using machines in the production process. It is worth mentioning a *monotype*, which is a print with only one issue.

Usually, lots are grouped into specific art auction types. *Old Masters* usually denotes European artists, which works dates back from the Renaissance up to the 18th century. *Modern* art auctions deals with lots from the 18th century up to the first half of the 20th century. Then, *contemporary* art auctions date back to the second half of the 20th century – sometimes, *post-war* periods are distinguished. A special category of auctions entails *emerging art* – it considers the work of the youngest artist, which usually has just entered the market and their art are sold at affordable prices. Sometimes specific artistic periods and movements receive their devoted auctions (such as Impressionism). All these definitions are rather flexible, especially in terms of the considered dates – they may vary on different national markets.

In terms of investment returns, it is debatable whether the art market is a good choice. Contrary to eye-catching auction records at Christie's or Sotheby's, some scholars argue that art historically yields relatively low returns (Baumol, 1986; Frey & Pommerehne, 1989). More recent

[3] `https://zam.umaine.edu/wp-content/uploads/sites/96/2013/09/Printmaking-Glossary.pdf` [accessed on August, the 2nd, 2020], `http://www.philaprintshop.com/diction.html` [accessed on August, the 2nd, 2020]

research shows that art yield returns similar to the ones from S&P 500. The report delivered by Deloitte (2019) provides insightful information on the global art market. For example, artnet's (see the next paragraph) top 100 artists produced an 8% compound annual growth rate in 2000-2018, whereas it was only 3% for S&P500. This group of artists, however, captures only the best performing part of the market. There is also an effect of a 10-years holding period – works with a long holding period tends to be sold at higher prices. Another report, prepared by Art Basel and UBS (McAndrew, 2020) provides less optimistic figures – the global art sales in 2019 was down 5% (year-on-year) to $64.1 billion (of which $24.2 billion considers sales at auction), which is the same level as in 2017. The US, China, and the UK are the largest art market in terms of sales – they share 84% of the market combined altogether. In 2019, post-war and contemporary art accounted for 53% of the value on the market.

Auction catalogues are the traditional data source of asking prices (sometimes accompanied by estimates). With the advent of the Internet, these catalogues are often available online on the web pages of auction houses. Post-auction hammer prices are often published as well. However, for art market professionals, it might be hard to follow countless art trading institutions. Therefore, companies aggregating data from multiple auction houses emerged. Perhaps the most known examples are ArtPrice[4] and artnet[5]. They provide an access to auction databases, which facilitates the research on appraisal and provenance. While the aforementioned pages consider the global art market, local markets have often dedicated services, such as artinfo[6] for the Polish one. With the constantly growing amount of information provided by historic auction sales, one may experience *analysis paralysis*.

Though there is much effort to make the art market transparent, it is not free from various ethical issues. For instance, Sotheby's and Christie's, two widely recognised auction houses, were involved in a price-fixing scandal[7]. It happened more than 15 years ago, but bidders seem to always have legitimate trust issues with auction houses. The first thing is appraisal – how can one trust evaluation prices since they are set up by people working for organisations which would like to maximise their profits? The next thing considers so-called *chandelier bidding*. In this situation, during the bidding, the auctioneer shouts the price while pointing at a non-existing bidder (to the ceiling, or more generally – to a chandelier), just to raise prices and encourage

[4] `http://www.artprice.com/` [accessed on August, the 2nd, 2020]
[5] `http://www.artnet.com` [accessed on August, the 2nd, 2020]
[6] `https://artinfo.pl` [accessed on August, the 2nd, 2020]
[7] `https://www.theguardian.com/uk/2002/oct/31/arts.artsnews` [accessed on March, the 5th, 2020]

actual bidders to make snap decisions. The next unethical technique is *collusion*. Some bidders gathered around a given seller form a cliqué, in order to bid against low prices. In the worst-case scenario, in which this lot is bought by a member of that cliqué, the auctioneer loses some money for the auction house fee. However, such an auction result might make prices of a given artist rise, which is intended by owners of their lots. While such an outcome is a win-win situation for owners, it's highly unethical and might cause market bubbles. There is a number of other aspects to consider, such as possibly shady provenance (such as stolen), or chances of buying a fake piece of art.

Finally, art as an investment might raise some interesting sociological questions. Entering the art market requires three different forms of capital: financial, cultural, and social. Without financial capital, investments are not possible for obvious reasons (though emerging art auctions aim at lowering this barrier). From the market point of view, cultural capital plays an important role in decision making. In fact, this is a form of *information asymmetry* preventing potential buyers from entering the market. David, Oosterlinck, and Szafarz (2013) argue that the art market is inherently inefficient due to the information asymmetry – namely, the lack of information about unsold artworks. The last one (social capital) is actually not a requirement per se, but it is tightly coupled with cultural capital. Participation in cultural events is attributed to people with higher social status (Domański, 2015). Sometimes, they even know artists themselves. All this build up their cultural capital and facilitates entering the market. This relationship was explored thoroughly for the Polish art market by Żaglewska (2016). All in all, monetising *invaluable* features or relation between artistic value and market appraisal may be questioned. Zboroń (2018) tries to grapple with these issues, although they still seem to remain open for debate. This, however, does not change the fact that art is being traded – for speculative and investment purposes as well – and this makes it an interesting subject for financial and economic research.

2.1.2 Art Auctions in Poland

Whereas the recent figures do not look great for the global market, the Polish part seems to be on a continuous and stable growth trajectory. In pre-1989 Poland, there was only one institution having a monopoly for trading art – DESA, which is an abbreviation for *Dzieła Sztuki i Antyki* (pol. artworks and antiques), was founded in 1950. Prices of art were relatively low, mainly due to two reasons (Białowąs et al., 2018). The first one was generally the low purchasing power of

Table 2.1: Share and turnover on the Polish art market in 2018, 2019 and 2020.

auction house	turnover [mln. PLN]			market share [%]		
	2018	2019	2020	2018	2019	2020
Desa Unicum	115.5	152.0	196.3	45.8	51.5	51.6
Polswiss Art	53.6	54.6	69.7	21.3	18.5	18.3
Agra-Art	18.8	24.0	35.6	7.5	8.1	9.3
Sopocki Dom Aukcyjny	22.5	21.5	24.7	8.9	7.3	6.5
Rempex	6.9	8.4	7.7	2.7	2.8	2.0
Libra	8.1	7.6	5.0	3.2	2.6	1.3
Polski Dom Aukcyjny	3.3	4.6	5.7	1.3	1.5	1.5
Art in House	2.2	3.1	5.8	0.9	1.1	1.5
Face to Face Art	1.3	3.0	2.4	0.5	1.0	0.6
Desa Sp. z o.o.	1.8	2.3	2.5	0.7	0.8	0.7

Source: artinfo.pl (2019),
`https://artinfo.pl/artinformacje/artinfo-oglasza-wyniki-rynku-sztuki-za-2019-rok`
[accessed on March, the 5th, 2020], and
`https://artinfo.pl/artinformacje/artinfo-pl-oglasza-wyniki-rynku-sztuki-za-2020-rok`
[accessed on March, the 26th, 2022]

the Polish society. The second one considers the facilitation of buying art for cultural heritage institutions. The symbolic end of the state-controlled economics in Poland in 1989 makes the market more than 30 years old now. As of 2018, there are 25 regular auction houses in Poland. Traditionally, the most known five are Desa Unicum (not to be confused with Cracow-based Desa Sp. z o.o. or the aforementioned DESA), Polswiss Art, Sopocki Dom Aukcyjny, Agra-Art, and Rempex. All of them conduct auctions since the late 80s or 90s. Libra, which entered the market in 2017, seems to be the only auction house that can compete with the aforementioned legacy ones. Following the report provided by artinfo.pl (2019), the financial results confirm the growing tendency for the Polish art market. The specific figures for 2018–2020 are presented in Table 2.1. The unquestionable leadership of Desa Unicum is manifested by its market share. It has even been ranked by artnet as the 11th European auction house by total sales value of paintings and sculptures in 2018. In 2019, it claimed up to the 9th position in the rank, which was later advanced to the 8th in 2020 and 2021[8].

The number of auctions held in Poland every year has a growing trend as well. In 2018[9] there were almost six times more auctions than in 2000 (Figure 2.1). Following artinfo.pl (2019),

[8] `https://www.bankier.pl/wiadomosc/Desa-wsrod-najwiekszych-domow-aukcyjnych-w-Europie-8279938.html` [accessed on March, the 26th, 2022]

[9] The data analysed in later chapters range to 2018. While this dissertation was submitted in 2022 and there are some newer figures on the market (which we occasionally report), we mainly focus on the snapshot of the Polish art market up to 2018 due to this reason.

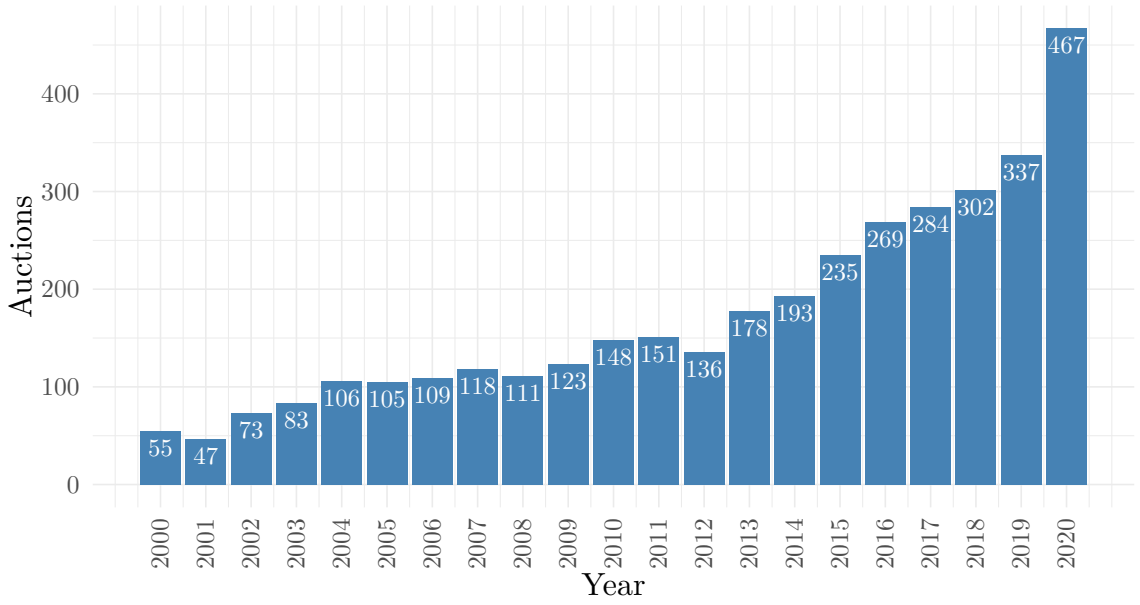


Figure 2.1: A number of fine art auctions in Poland between 2000 and 2020.
Sources: artinfo.pl (2019, 2020, 2021)

in 2018 there were 122 traditional and modern art auctions (turnover: 207.7m PLN), 92 young and contemporary art auctions (9.1m PLN), 6 sculpture auctions (12.2m PLN), and 6 auctions devoted to photography (1.5m PLN). Most of them were held in Warsaw. Similarly to other art markets, most of the sales takes place in December. There is an increasing demand for modern art, which is the most important segment of the Polish art market (artinfo.pl, 2019). Hammer prices for paintings from artists such as Wojciech Fangor, Ryszard Winiarski, or Henryk Stażewski can go well above 1 million PLN. As of 2022[10], the most expensive painting sold at a Polish auction house was a portrait of a lady by Peter Paul Rubens, sold at Desa Unicum for 12,000,000 PLN (hammer price excluding buyer's premium) in 2022. The most expensive Polish painting sold at a Polish auction house was "Dwie mężatki" by Andrzej Wróblewski. The painting was sold for 11,200,000 PLN (hammer price excluding buyer's premium) on contemporary art auction. The highest price of a Polish artist is attributed to Magdalena Abakanowicz – her set of sculptures "Bambini" was sold in 2021 for 11,300,000 PLN (excluding premium) at Polswiss Art[11].

Similarly to the global one, the Polish art market isn't free from various ethical issues. Some

[10] `https://www.bankier.pl/wiadomosc/Nowy-rekord-na-polskim-rynku-sztuki-Chodzi-o-obraz-Rubensa-8300710.html` [accessed on March, the 26th, 2022]

[11] `https://rynekisztuka.pl/2021/12/08/rekord-aukcyjny-bambini-magdalena-abakanowicz/` [accessed on March, the 26th, 2022]

sellers are reporting the evidence of allegedly sold artworks, which are returned to the owners[12]. Such practice is conducted to create the illusion of high demand and increase future prices. Bidders at young art auctions in Poland have to face the problem of plagiarism, such as in Figure 2.2. Another example considers the famous case of Art-B, a company operating in the 1990s. It was best known for its misuse of banking system inefficiencies, which allowed its owners to make fortune. Art-B also traded blue-chip artworks – it turned out that it was just a tool for money laundering (Borowski, 2013).

George Dikhamindjia, *untitled*, 2017 Karolina Fox, *Stand by me*, 2018

Figure 2.2: An example of plagiarism (on the right) at the Polish Young Art auction. Source: `https://www.facebook.com/bekazmlodejsztuki/` [accessed on October, the 17th, 2019]

However, the most notorious recent case is the story of Abbey House. Technically speaking, it wasn't an auction house – a private gallery would describe it better, though they referred to themselves as *artbanking*. Abbey House gained attention after selling a painting of an unknown young artist for a price, which reached the levels of Polish blue-chip artists and was considered way too high (Figure 2.3). As it turned out, their business model was constructed around contracts forcing artists to deliver a number of artworks monthly, which were later sold at artificially bloated prices in order to create an intended bubble. Such an approach could not last long and the business finally collapsed. While the case of Abbey House is widely recognised, there's also a less known follow-up. After the aforementioned scandal, Abbey House Group S.A. bought ARTNews, a widely known art journal[13]. The former name disappeared, leaving little trust in the transparency of the art market. With the rise of online art auctions, it is expected

[12] `https://pieniadze.rp.pl/inwestycje/sztuka/22102-rynek-sztuki-w-polsce-2019-2` [accessed on August, the 2nd, 2020]

[13] `https://dziennikpolski24.pl/wielka-klapa-na-polskim-rynku-sztuki/ar/c3-3545117` [accessed on March, the 5th, 2020]

that the markets will become more transparent[14].

Figure 2.3: Agata Kleczkowska, untitled, 2010, oil and acrylic on canvas – sold for 160,000 PLN in 2010 at Abbey House.
Source: `http://rynekisztuka.pl` [accessed on October, the 17th, 2019]

2.1.3 Buyers' preferences and price determinants

Buyers' preferences can be described as a set of desirable features of a given good. These features can determine a willingness to purchase it and might affect its price. In a market of heterogeneous goods determining these values might not be straightforward. Paintings fall into this category, as they are incomparable in a direct way. However, one can explore the motivation behind buying paintings and provide descriptive statistics about the market. Price determinants might be used to quantify the impact of a given feature.

One way to try to determine buyers' preferences is by exploring the motivation behind buying art. Artsy surveyed online art collectors for core buying motivations[15]. According to this report, core online motivation driving buyers are (in this order) buying art as a home decoration (important for 71%), as an inspiration (67%), to build a collection, to provide support for artists, for a gift, to donate it to other institutions, to resell it to other clients. What's more interesting, only 17% of surveyed people said that the content of the work or the artist background does not matter when it comes to buying decisions. The report also reveals slight gender differences. More women are willing to buy art as an inspiration or support known artists, whereas men are more interested in treating it as an investment or building a collection. A more recent report from ArtBasel and UBS (McAndrew, 2020) surveyed high-net-worth (HNW) buyers worldwide for motivations behind buying art. These buyers mentioned:

[14] `https://news.artnet.com/market/price-transparency-art-market-1915145` [accessed on March, the 26th, 2022]
[15] `https://www.artsy.net/article/artsy-editorial-drives-art-buyers` [accessed on August, the 2nd, 2020]

- aesthetics/devorative considerations (important for 95% surveyed),
- passion/expression of personality (93%),
- support of artists and culture (92%),
- social-contacts and friendship (87%),
- family traditions and heritage (87%),
- social reasons (86%),
- portfolio diversification (85%),
- expected return (85%),
- status/cultural credibility (84%),
- hedging against inflation (81%).

While decorative aspects seem necessary to HNW buyers, one can notice the importance of financial and social aspects compared to online buyers. The latter group is in line with findings provided by Żaglewska (2016). These motivations give some insight into preferences but do not fully reveal them.

A more straightforward approach to assess buyers' preferences would consider descriptive statistics considering sales data in auction houses, such as the frequency of buying a given artist and its price. The turnover and transaction share of different types of art for the Polish art market is presented in Table 2.2. Compared to the global market, the classification and nomenclature of art auctions in Poland are slightly different. Old Masters (including XIXth century and modernism) auctions often offer works dated to World War II. More recent lots made before the 1990s are considered contemporary, with special categories such as op-art and conceptual art. In Poland, emerging artists can be divided into young art and new post-1989 generation. The latter considers young but established artists who have already gained some recognition, contrary to young art auctions (pol. aukcje młodej sztuki), which is dedicated to debutants. Young art (also called emerging art) accounts for an essential part of the Polish art market and is even a subject of dedicated studies (Wójtowicz, 2019). Desa Unicum held the first such auction in 2008. For many collectors, the formula of selling works of young and unknown artists with a low initial price (500 PLN) facilitated entering the market. The report provided by artinfo.pl (2019) sheds some light on the figures describing the market for young art auctions in Poland in 2018. The total turnover at young art auctions was equal to 9,100,100 PLN, which accounted for 42.4% of lots sold at art auctions. The mean hammer price for a single lot on such auctions was 1685 PLN, whereas the record for the highest one belongs to Daniel Pawłowski (48,000 PLN). For

Table 2.2: Turnover and transaction share by art type in Poland in 2018.

	% of turnover	% of transactions
paintings	83.8	63.5
sculptures	7.5	3.7
works on paper	3.2	15.1
handicrafts	2.9	11.4
photography	1.1	3.2
others	1.6	3.1

Source: artinfo.pl (2019)

24% of lots at young art auctions, initial prices turned out to be their hammer prices. Desa Unicum and Art in House both have around one-quarter of this market's total turnover. More descriptive statistics about paintings on the Polish art market can be found in Section 6.1 – both top painters (in terms of the number of sold works) and young art are described there.

Determining which features can describe a painting and which values of these features are popular among buyers is the first step of assessing buyers' preferences. The second one considers treating them as *price determinants* – along with quantifying their influence on the hammer price. Borowski (2013) enlists features that influence hammer prices on the art market:

- author's recognition,
- prices of other works of this author,
- meaning of a given work in author's portfolio,
- artists popularity on global markets,
- authenticity of a given work,
- quality of a given work,
- presence in art catalogues,
- general condition of a given work,
- creation whereabouts,
- current trends,
- number of exhibitions which presented a given work,
- names of publishers, which published reproductions of a given work,
- belonging to a famous collection.

In terms of quantitative analysis, many of these features are hard to obtain. For instance, all information related to provenance is often a subject of discretion. Other ones, however, are often provided in auction catalogues – such as size. Kräussl and Elsland (2008) compared determinants

of paintings hammer prices used in numerous quantitative art market research studies. They are presented in Table 2.3, where "+" means the positive influence on the price, "-" stands for a negative one, and finally "+/-" represents both positive and negative influence (depending on the categorical variable value). However, the table does not explain the level of attribution to the price. These determinants have been obtained using quantitative art market analysis, to which the next section is dedicated. One can also observe that some features positively influence the hammer price in some datasets, whereas others negatively affect others (such as surface). Therefore, one needs to be careful when generalising these insights, as they are tightly coupled with the analysed datasets. Section 6.2 describes price determinants for the aforementioned Polish datasets of the top 10 painters and young art. The methods for quantifying the influence on the price are described in the following subsection.

2.2 Quantitative Art Market Analysis

Since art can be perceived as an investable asset class, a number of scholars carried out studies on measuring its performance and comparing it to other forms of investments. Drawing from traditional price-tracking methods, many art market indices have been created to serve this purpose. Triplett (2004) argues that conventional price index methodologies should not be used with heterogeneous goods, such as art (in general, excluding prints etc.). There is no agreement among scholars on how these indices for the art market should be built, though a number of approaches exist. Popular ways for constructing such indices and determining price determinants in a quantitative manner are examined in this subsection. The initial findings contained here has been published in a separate paper (Filipiak & Filipowska, 2016).

Following Ginsburgh, Mei, and Moses (2006), art market indices have four main purposes: outlining general market trends, measuring them, examining the influence of external impacts, and appraisal. The first one helps in measuring market returns in a similar way to Dow Jones Industrial Average and facilitates comparisons with other financial instruments. An index is also a tool to examine the market's volatility and its correlation with another form of investments, which are crucial for risk diversification. They can also be used for measuring the influence of external factors, such as inflation. Finally, they can be employed for the appraisal of artworks. Ginsburgh et al. (2006) also written about desirable features of art market indices. Following them, a good art market index must rely on publicly available price databases. It has to address

Table 2.3: Price determinants in different art market studies.

Variable	Chanel, Gérard-Varet, and Ginsburgh (1992)	Czujack (1997)	Renneboog and Van Houtte (2002)	Hodgson and Vorkink (2004)	Biey and Zanola (2005)	Worthington and Higgs (2006)	Kräussl and Schellart (2007)
Year of sale	+/-		+/-	+	+/-	+	+/-
Month						+/-	
School							
Width	+		+	+			-
Height	+		-	-			+
Width^2	-						
Height^2	+						
Surface	-	+	+	-	-	+	-
Surface^2		+				+	
Technique		+/-		+	+/-	+/-	+
Support		+/-		+			+
Place of sale		+/-	+				+
Auction house	+/-	+/-	+/-	+	+/-	+	+
Painter	+/-			+/-		+/-	
Signed?		-	+		-		-
Painter alive?						-	+
Painter age						+	
Painter age^2						+	
Painter age^3						+	
Painter age^4						+	
Art current			+/-				
Average price							+
Publication		+					
Exhibitions		+/-					
Periods		+					
Provenance		-					
Estimate							+
Period	1855-1970	1963-1994	1970-1997	1968-2001	1988-1995	1973-2003	1986-2006
Sample size	1,972	921	10,598	12,821	1,665	30,227	1,688
Artists	46	1	71	152	1	50	23

Source: Kräussl and Elsland (2008)

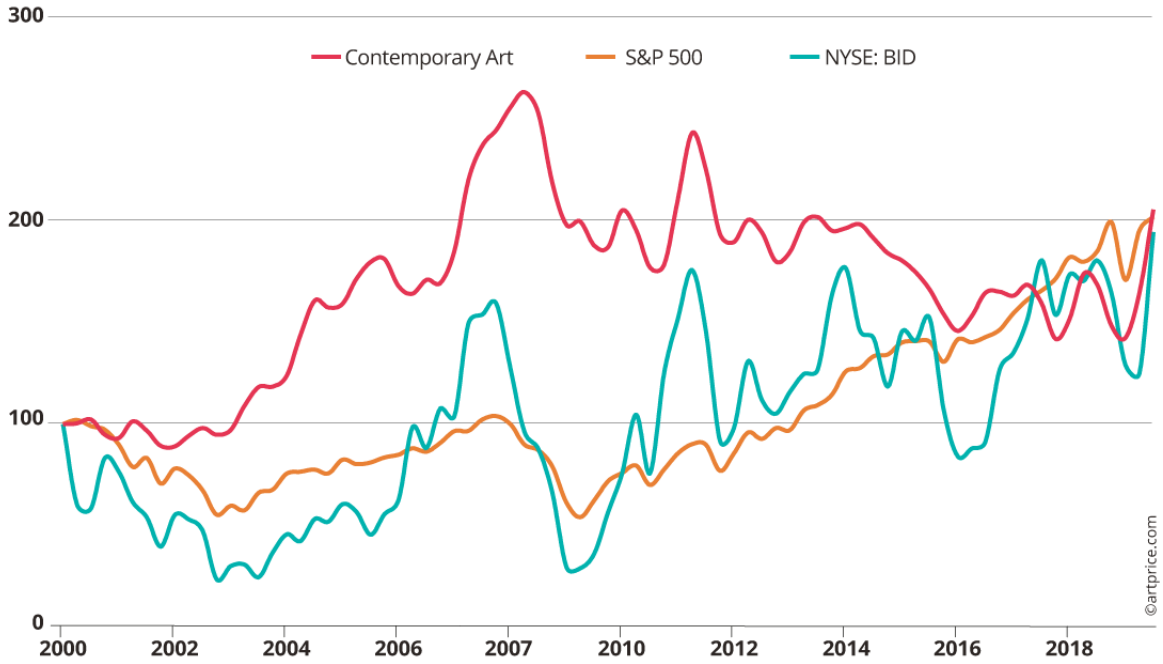


Figure 2.4: Artprice's Contemporary Art Price Index vs financial markets.
Source: `https://www.artprice.com` [accessed on March, the 14th, 2020]

the issue of heterogeneity and avoid selection bias at the same time. It should also take into account different types of artworks. Finally, updates of the index should be provided in an equal time interval. Figure 2.4 presents the contemporary art index values compared to S&P 500 and Sotheby's at New York Stock's Exchange (hence the NYSE:BID symbol), whereas Figure 2.5 depicts the artnet index for the top 10 and 100 artists, compared to the gold fixing price and S&P 500 (both considers the global art market).

However, art market indices are criticised for a number of issues. For instance, they suffer from selection bias. It is not only about the selection of artists which would be taken into account – even though usually only popular ones are considered in a given index. The main problem is the thing that indices are built on the best performing part of the market since only sold lots are considered (i.e. bought-ins are not taken into the account). Another problem stems from the fact that some part of the market is unobservable. While auction houses often provide auction results, transaction prices in numerous private galleries are not publicly available. Following the last available TEFAF report (Pownall, 2017), the ratio of sales in private galleries to auction houses varies in different regions. Nevertheless, an important part of the examined market is non-observable.

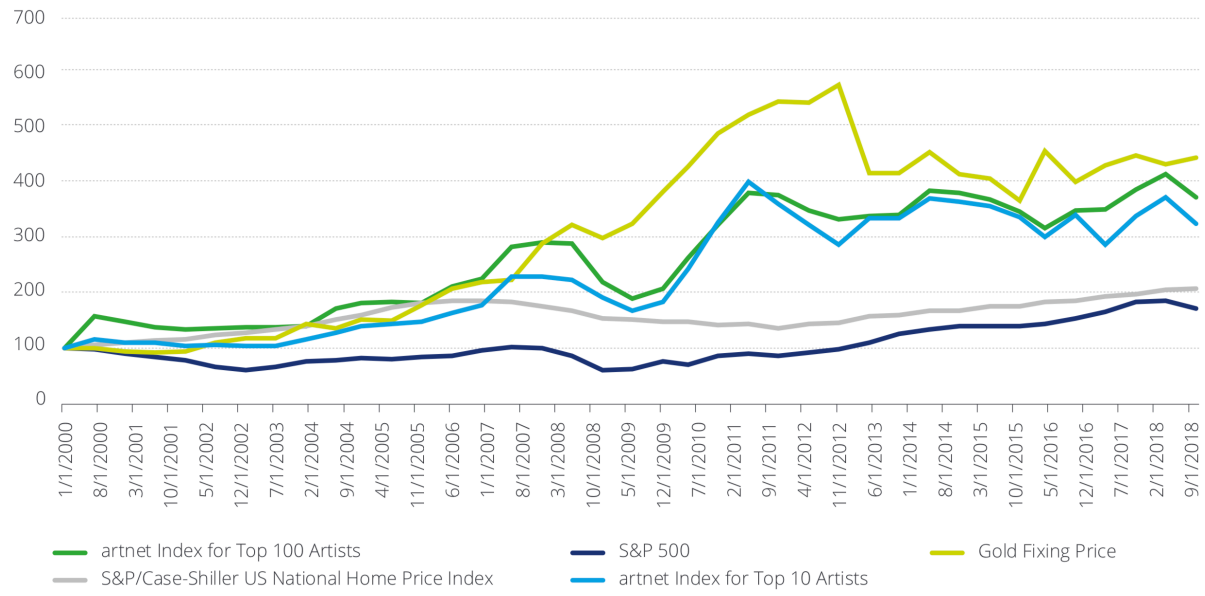


Figure 2.5: artnet price index for top 100 artists.
Source: Deloitte (2019)

There is a number of approaches for art market index construction. The most basic one is the so-called *naïve index*, which is calculated basing on mean/median prices in consecutive years. For instance, an index for some period $t+1$ would be calculated as follows:

$$\text{Index}_{t+1} = \frac{\prod_{i=1}^{n} (p_{i,t+1})^{1/n}}{\prod_{i=1}^{m} (p_{i,t})^{1/m}}. \tag{2.1}$$

Essentially, it is the quotient of two geometric means of artwork prices p_i for two separate periods $t+1$ and t. Since it is unlikely to have the same number of observations for these two periods, two variables m and n are introduced to represent the number of observations in each one. The naiveness of this index type stems from the assumption of equal distribution of artworks' features, which is a condition practically impossible to satisfy in real-world transactions data. Therefore, it is rather not used in practice. Usually, art market indices base on one of these two approaches: *hedonic regression* (hereafter HR) and *repeated-sales regression* (abbreviated as RSR). The former approach is probably the most known from house prices indices. Hedonic regression indices rely on quantifying artworks' features and examining their influence on price, which also enables tracking particular price determinants. Repeated-sales regression indices examine price differences for lots sold at least twice, which can be more accurate. Their drawback stems from the fact that this approach significantly reduces the dataset to analyse – for example,

artnet Analytics (2014) states that lots sold at least two times constitute only 10% of their dataset at that time. Interestingly, Ginsburgh et al. (2006) argue that HR and RSR are positively correlated given a long enough period with a sufficient number of sold artworks. Other index types also exist (such as e.g. *composite indices*), though they seem to be rarely used and therefore they are not extensively discussed here.

The remainder of this section is structured as follows. Section 2.2.1 provides a concise overview of perhaps the most popular technique for building art market indices and examining the impact of particular features – hedonic regression. Another important technique, repeated-sales regression is discussed in Section 2.2.2. Section 2.2.3 is focused on the process of index construction. Finally, Section 2.2.4 offers a brief review of other quantitative techniques used in art market research.

2.2.1 Hedonic Regression

Hedonikos is a Greek word used for pleasure. Following Chau and Chin (2003), in this context pleasure can be viewed as a utility (or satisfaction) of the consumption of goods and services. Hedonic regression assumes that the price of goods can be expressed as a linear combination of their attributes and qualities. Such models are especially popular for building house model prices. This market can be characterised by durability, heterogeneity, and spatial fixity of examined goods (Chau & Chin, 2003). Traditionally, hedonic regression is a popular way to build an index for the art market as well, since it shares the first two qualities with the property market. The purpose of hedonic regression is twofold. With the usage of time dummies, the aforementioned art market index can be created. On the other hand, models generated in such a way enable us to evaluate the influence and statistical significance of particular qualities of artworks, which enables us to underline their price determinants.

Before analysing different approaches to art market research, it will be noteworthy to say that hedonic regression is just a form of plain linear regression. Following Dougherty (2011), a linear regression tries to capture a relationship between dependent (explained) variable y_i and a linear combination of n independent (explanatory) variables $x_{i,t}$ (each with weight β_t), which is assumed to have the following formula:

$$y_i = \beta_0 + \sum_{t=1}^{n} x_{i,t}\beta_t + \varepsilon_i, \tag{2.2}$$

where β_0 is the intercept. The random component ε_i for the observation i is called the disturbance term. The true values for β_t are usually not known and they are a subject of estimation. These values are then used by the following equation:

$$\hat{y}_i = b_0 + \sum_{t=1}^{n} x_{i,t} b_t, \tag{2.3}$$

where $\hat{y}$ is the prediction of dependent (explained) variable for an observation i, β_0 is the intercept, $x_{i,t}$ is the t-th independent (explanatory) variable for the i-th observation multiplied by its estimated weight β_t. This is often written in a shorter, vectorised form (with $\mathbf{X}_{0,t} = 1$):

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}. \tag{2.4}$$

The parameters in linear regression can be estimated using the ordinary least squares method (abbreviated as OLS). This technique relies on minimising the sum of squared residuals (abbreviated as RSS). A residual e_i for i-th observation is the difference between the real and fitted value, i.e. $y_i - \hat{y}_i$:

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{t=1}^{n} x_{i,t} b_t \right)^2. \tag{2.5}$$

Therefore, the OLS method minimises the following:

$$\begin{aligned} \arg\min_{\mathbf{b}} \mathrm{RSS} &= \arg\min_{\mathbf{b}} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{t=1}^{n} x_{i,t} b_t \right)^2 \\ &= \arg\min_{\mathbf{b}} \left\| \mathbf{y} - \mathbf{X}\mathbf{b} \right\|^2 \end{aligned} \tag{2.6}$$

The solution for this minimisation problem is given by the following formula in a matrix-vector notation:

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{2.7}$$

Leaving the assumption of linear dependence aside, there's still a number of conditions that have to be fulfilled in order to use OLS. For instance, the disturbance term has to fulfil the Gauss-Markov conditions (Dougherty, 2011). The first condition considers its expected value for every observation:

$$\forall_i \ \mathbb{E}(\varepsilon_i) = 0. \tag{2.8}$$

The second condition assumes constant population variance σ_ε^2 for all observations:

$$\forall_i \ \sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2. \tag{2.9}$$

This implies homoscedasticity. The third condition regards independence of disturbance terms:

$$\forall_{i,j,i\neq j} \ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0. \tag{2.10}$$

The fourth condition is about independent distribution of explanatory variables, i.e. the population covariance between the explanatory variable and the error term is equal to zero:

$$\forall_i \ \mathrm{Cov}(x_i, \varepsilon_i) = 0. \tag{2.11}$$

Aside from Gauss-Markov conditions, there exist a number of other requirements. For instance, it is assumed that the distribution of the disturbance term is normal. Another condition assumes a lack of multicollinearity, which is a near-linear relation between (at least) two dependent variables. In practice, it is hard to meet all of these criteria and some of these conditions are violated, which negatively impacts the estimates to some extent.

The coefficient of determination, perhaps better known as R^2, is a single number with a maximum value of 1 used to evaluate linear regression. Following Dougherty (2011), it measures *goodness of fit*. It can be also interpreted as the percent of variance explained by the model. Coefficient of determination is calculated as follows:

$$R^2 = \frac{\mathrm{Var}(\hat{\mathbf{y}})}{\mathrm{Var}(\mathbf{y})}. \tag{2.12}$$

To examine the significance of particular regression coefficients b_i, p-values at the significance level (usually, α=0.05 or 0.01) are used in order to test the null hypothesis ($H_0\colon b_i = 0$ versus $H_1\colon b_i \neq 0$). However, there is an ongoing debate[16] about the usefulness of p-values, as there are known cases in which they are misused. Despite being even banned in some journals, p-values still seem to be widely used in the scientific community.

One of the earliest works devoted to the econometrics for the art market was presented by Schneider and Pommerehne (1983), in which they developed a simple model for modern art. They

[16] `https://www.the-scientist.com/news-opinion/drop-statistical-significance--scientists-say-65635` [accessed on August, the 2nd, 2020]

started with assuming that the art market is competitive to some extent (i.e. supply and demand model the prices) and gallery owners try to maximise their profits. The authors of this paper tried to identify supply and demand factors influencing the price of a representative artwork, which was assumed to be non-random. They used a two-step procedure to model artwork sales in the 1970s using two relationships. Aesthetics evaluation was performed estimated using OLS in the first model:

$$\mathrm{AES}_i = \sum_{j=1}^{n} \alpha_j x_{i,j} + \varepsilon_{\mathrm{AES},i}, \tag{2.13}$$

where AES_i is the aesthetics score of an artwork i, whereas $x_{i,j}$ represents mostly the dummy variables for different artistic styles (such as *Pop Art*, *Optical Art*, *New Realism*, or *Conceptual Art*) and other variables (awarded art prizes, number of one-man exhibitions, moving average of past prices etc.) with a weight α_j and $\varepsilon_{\mathrm{AES},i}$ is the disturbance term. The fitted model reached $R^2 = 0.55$. The second equation models the relation between prices and numerous characteristics and it is estimated using GLS (general least squares):

$$\ln p_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{i,j} + \varepsilon_i. \tag{2.14}$$

In the second equation (estimated using GLS), p_i is the price of an artwork i, β_0 denotes the intercept, $x_{i,j}$ is the j-th characteristics of the painting i (e.g. estimated AES_i from Equation (2.13) or rates of other sources of investment) with a weight β_j, and ε_i is the disturbance term. The second equation has been run twice. In the first run, p_i has been based on experts' appraisal, whereas in the second one real auction house prices were used – these models achieved $R^2 = 0.6$ and $R^2 = 0.66$ respectively.

In hedonic models, independent variables are traditionally referred to as hedonic variables. Kräussl and Elsland (2008) examines popular hedonic variables used in the art market research literature. They enlist year of sale, month of the year, school, width, height, surface, technique, support, place of sale, auction house, painter, presence of a signature, the fact that the artist is dead, painter age, year, art current, reputation (average price), publication, number of times exhibited, working periods, provenance, and prior price estimates. Some of these variables are transformed using squares or taking logarithms. Perhaps the most common form of a hedonic regression equation (sometimes called time-dummy form) in art market research literature is

given as follows:

$$\ln P_{i,t} = \alpha + \sum_{j=1}^{z} \beta_j X_{ij} + \sum_{t=0}^{\tau} \gamma_t D_{it} + \varepsilon_{it} \tag{2.15}$$

in which $\ln P_{i,t}$ represents the natural logarithm of a price of a given painting $i \in \{1, 2, ..., N\}$ at time $t \in \{1, 2, ..., \tau\}$, and regression coefficients are given by α (intercept), β (for hedonic variables) and γ (for time dummies), X_{ij} stands for hedonic variables in the model. Time dummies variables are represented by D_{it} – it is equal to one only if a given paining i was sold in a period t (otherwise it is equal to zero). The first considered period is treated as a base one and usually it is excluded from the OLS estimation. After performing the OLS, the time dummy coefficients γ_t can be used to calculate the index for a period t (here with the base set at 100):

$$\text{Index}_t = 100e^{\gamma_t}. \tag{2.16}$$

More complicated ways to construct an index are presented in Section 2.2.3. Other hedonic coefficients captured in β_j can be used to determine the effect and statistical significance of a given quality in terms of the pricing.

Hedonic regression has been used in countless global art market studies. For instance, the extensive research carried out by Renneboog and Spaenjers (2013) provided an insight to the five art markets (France, Italy, the Netherlands, the UK, and the US) from 1970 to 2008. The index built in this paper is presented in Figure 2.6. Hedonic methods are also used to inspect different parts of the art market. For example, Czujack (1997) used HR to explore price determinants for Picasso's paintings, Habalová (2018) devoted her study to fine art photography, whereas the market for sculptures was investigated by Locatelli-Biey and Zanola (2002). Some scholars try to explain particular phenomena using HR – for instance, Etro and Stepanova (2015) analysed historical auction records traded in Paris and proved the existence of the "death effect" – a spike of the prices shortly after the death of a given artist.

Regarding the Polish art market, there is a handful of scholars who used HR in their research. For example, Lucińska (2012, 2013) used hedonic regression to examine the relationship between hammer prices and artist's age. The considered sample consisted of popular Polish artists. For Leon Wyczókowski, Wojciech Kossak and Julian Fałat there is a visible negative correlation between their age and hammer prices. No clear pattern was found in terms of the artist age of the most expensive work sold. In another paper, Lucińska (2015) used hedonic regression to build the Polish art index for 2007-2010. Kompa and Witkowska (2013) used the 2-step hedonic

Figure 2.6: Art index for France, Italy, the Netherlands, the UK, and the US.
Source: Renneboog and Spaenjers (2013)

approach (see Section 2.2.3) presented by Kräussl and Elsland (2008). In their research, 10,400 artworks of 2,938 painters sold in 2007-2010 have been used. They constructed 25 different hedonic models in total, some of which had relatively high $R^2 > 0.95$ The most important hedonic variables were *artist alive*, *size* and *price class*. In another paper, Witkowska (2014) analysed naive, hedonic and average hedonic indices on the same dataset.

Chau and Chin (2003) argues that there are some drawbacks of hedonic regression, starting from the assumption of a linear relationship between dependent variable and features. Another problem stems from the choice of basic functional forms, which can be linear, reciprocal, semi-log and log-log. Misspecification of included qualities can lower the quality of the model as well. Perhaps the most important drawback of hedonic regression is selection bias since it captures only the best part of the market – the lots which have been actually sold. In order to mitigate this problem, Collins, Scorcu, and Zanola (2009) proposed using the chained Fisher price index along with the Heckman 2-step estimation procedure. Despite these simplifications and drawbacks, hedonic regression is still a popular tool for scholars and financial experts for examining markets. It is especially useful for goods that are traded infrequently in long intervals – such as in the art market.

2.2.2 Repeated-Sales Regression

A method considering the prices of the same object sold at least twice seems very natural for the price index calculation. Following Hill (2011), repeated-sales regression (hereafter RSR) in its most basic form presented by Bailey, Muth, and Nourse (1963) is estimated using OLS and it looks as follows:

$$\ln p_{t,h} - \ln p_{s,h} = \sum_{\tau=0}^{T} \beta_\tau D_{\tau,h} + \varepsilon_h, \tag{2.17}$$

where $p_{t,h}$ and $p_{s,h}$ are the consecutive prices of the same object with an index h for periods t and s, The time dummies $D_{\tau,h}$ are equal to -1 if $\tau = t$, 1 if $\tau = s$, and 0 in other cases. The index can be calculated similarly as in HR methods:

$$\text{Index}_t = e^{\hat{\beta}_t}. \tag{2.18}$$

Repeated-sales regression does not need data about additional features of considered items, which is an advantage and disadvantage at the same time – it facilitates the analysis, but the impact of particular price determinants cannot be determined. Compared to HR, the selection bias

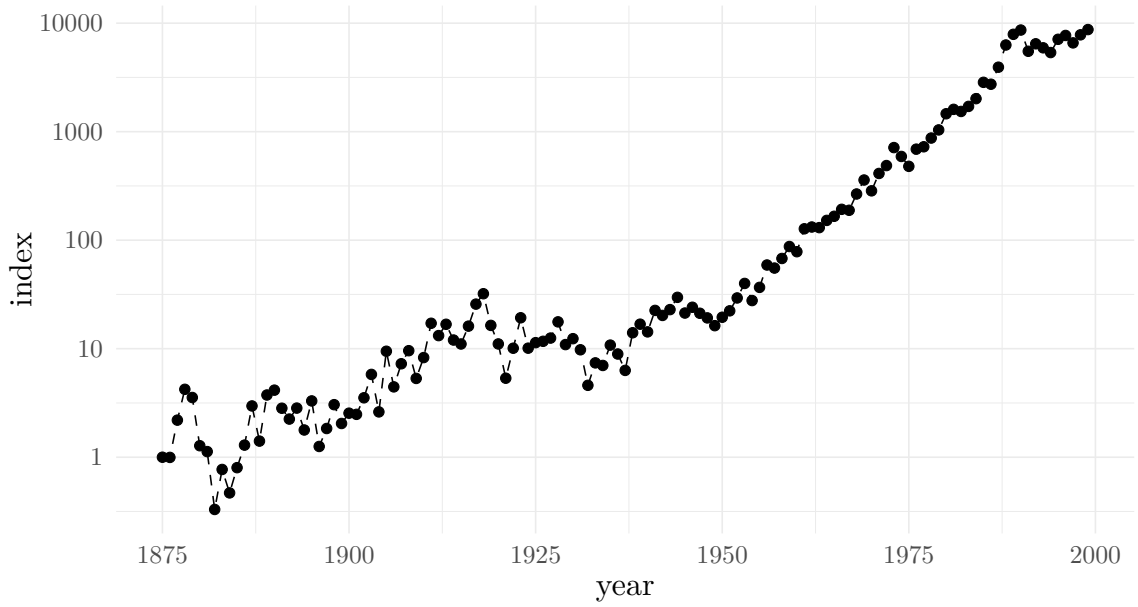


Figure 2.7: Mei & Moses All Art index values between 1875 and 1999 (log scale).
Source: Mei and Moses (2002)

in RSR is manifested even more since the considered goods are assumed to be traded relatively frequently (at least twice), which certainly makes them not a random sample (Graddy, Hamilton, & Pownall, 2012). Another criticism stems from the fact that two goods will not necessarily stay the same – for instance, a flat can be renovated (Hill, 2011). More sophisticated techniques for RSR do exist – for instance, Case and Shiller (1987) suggested a three-stage approach, which uses generalised least squares (GLS).

Mei Moses Fine Art Index is perhaps the most famous art market index. It is named after their creators – Mei and Moses (2002). The index was built in 2001 and is an example of the usage of repeated-sales regression. In the beginning, the index was built from the data collected from Christie's and Sotheby's for lots sold from 1950 to 2000. For some lots, provenance information had been given. This enabled to track previous sales even to 1875. For each database entry (i.e. sold lot), two prices are associated – *purchase price* and *sale price*. The whole database contained 4,896 pairs of prices (899 for American art, 1,709 for Impressionist, and 2,288 for Old Masters). The index is presented in Figure 2.7. The dataset was further developed (Mei & Moses, 2005; Ginsburgh et al., 2006) – in 2016, it entailed 45,000 lots sold at least twice. It was acquired by Sotheby's[17] in the same year.

[17] `https://www.sothebys.com/en/articles/sothebys-acquires-the-mei-moses-art-indices` [accessed on March, the 5th, 2020]

Following Mei and Moses (2002), it is assumed that the index for some interval $t = 1 \dots T$ can be calculated from the following equation:

$$r_{i,t} = \mu_t + \eta_{i,t} \tag{2.19}$$

$r_{i,t}$ denotes a return for an art asset $i = 1 \dots N$ is time t, μ_t is described as the average return in time t for paintings, and $\eta_{i,t}$ is the error term for the asset i is time t. The indices are stored in $\boldsymbol{\mu}$, which is a T-dimensional vector. A single observation i can be characterised by its purchase price $P_{i,b}$ (on date b_i) and sale price $P_{i,s}$ (on date s_i). Return for a single item is measured as follows:

$$r_i = \ln\left(\frac{P_{i,s}}{P_{i,b}}\right). \tag{2.20}$$

Using Equation (2.19) and (2.20) one would obtain the following:

$$r_i = \ln\left(\frac{P_{i,s}}{P_{i,b}}\right) = \sum_{t=b_i+1}^{s_i} r_{i,t} = \sum_{t=b_i+1}^{s_i} \mu_t + \sum_{t=b_i+1}^{s_i} \eta_{i,t}. \tag{2.21}$$

Using $\mathbf{r}$ as a N-dimensional vector of all observations, a maximum-likelihood estimate of $\boldsymbol{\mu}$ is given by the following equation:

$$\hat{\boldsymbol{\mu}} = (\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{r}, \tag{2.22}$$

where $\mathbf{X}$ denotes a $N \times T$ matrix which elements represent time for each considered lot i, and $\boldsymbol{\Omega}$ stands for a weighting matrix based on times between sales (Goetzmann, 1993). The index is given by the following equation:

$$\text{Index}_t = \exp(\mu_t + \sigma^2/2) - 1, \tag{2.23}$$

where σ^2 is estimated using the procedure provided by Case and Shiller (1987).

Regarding the Polish art market, repeated-sales regression methods are not popular among researchers due to the relatively young age of the post-1989 auction houses, which results in scarcity of available data samples. Białowąs et al. (2018) performed perhaps the largest study devoted to the Polish art market using RSR. They used Art&Business magazine data paired with Repeated-Sales Regression to analyse Polish auctions – the research included 28,951 lots sold from 1989 to 2012, of which 1142 (around 4%) was sold at least twice in this period. The

examined dataset does not include the data from auction houses that discontinued operations in this period. The resulting index is visualised in Figure 2.8, showing relatively stable and long-term growth in this period. Particularly, the periods of economic growth in Poland (1995-2000 and 2003-2008) are directly reflected in the index values. Białowąs et al. (2018) went further and compared art to other sources of investment. Polish art yielded similar returns to treasury bonds – this resemblance was captured by scholars for other markets as well (Renneboog & Spaenjers, 2013; Pesando & Shum, 2008). However, stock and gold outperformed art in this period. The results of this comparison are presented in Figure 2.9. The same dataset of 1142 artworks sold at least twice was later investigated by Szyszka and Białowąs (2019) in terms of the existence of the death effect on the Polish art market (which, as it turned out, affects the price).

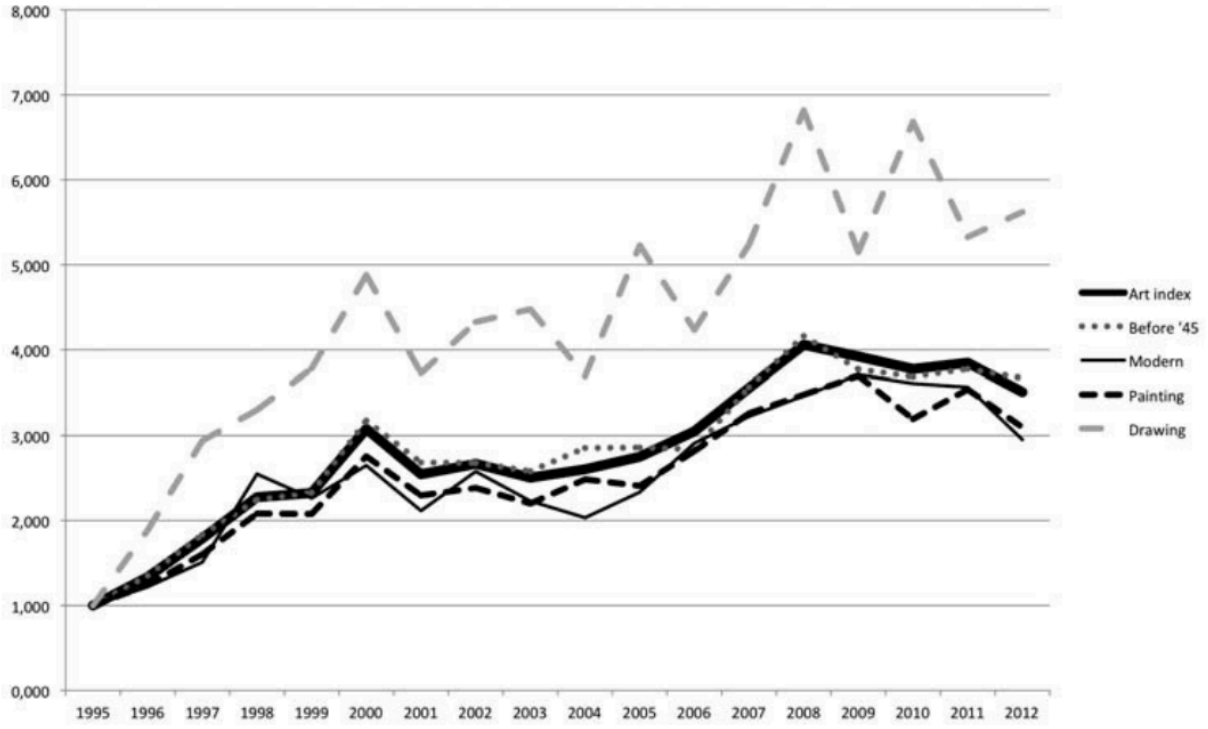


Figure 2.8: The Polish art index between 1995 and 2012 with subcategories.
Source: Białowąs et al. (2018)

2.2.3 Art Market Indices

Times-Sotheby's index was perhaps the first dedicated art market index. It was published in The Times between 1967 and 1971, certainly helping to perceive art as an asset class[18]. This section

[18] `https://hyperallergic.com/476003/your-money-is-safe-in-art-how-the-times-sotheby-index-transformed-the-art-market/` [accessed on August, the 2nd, 2020]

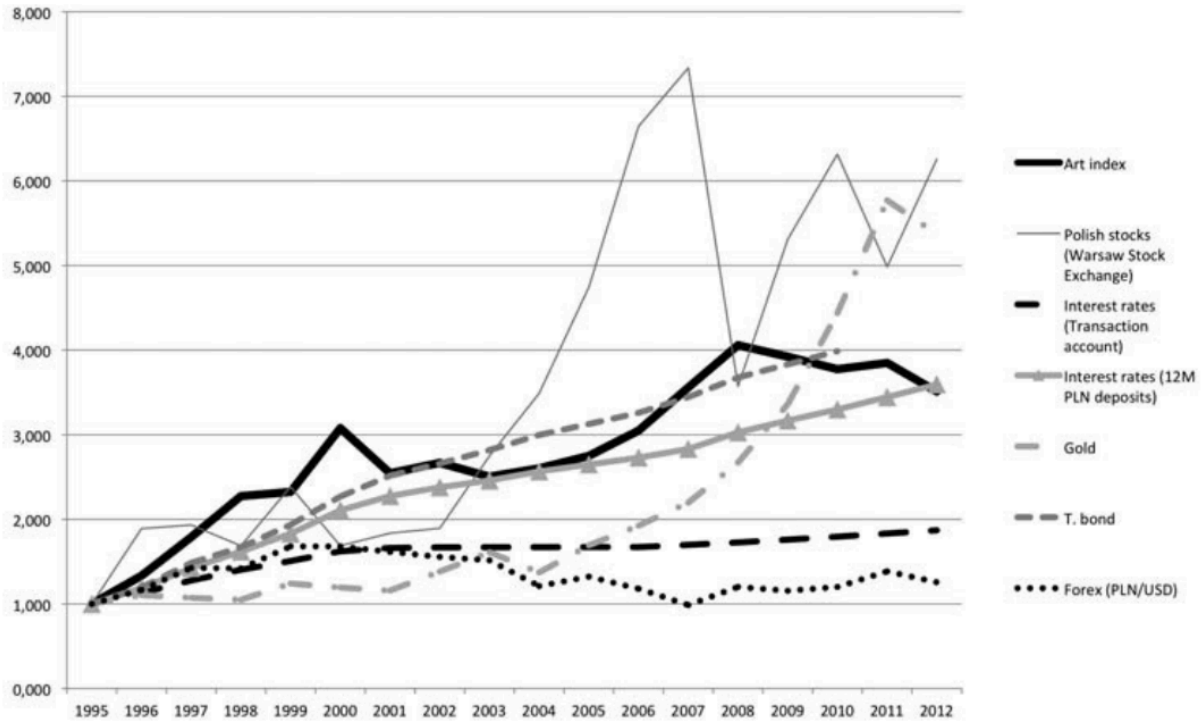


Figure 2.9: The Polish art index between 1995 and 2012 compared to the other investments.
Source: Białowąs et al. (2018)

is dedicated to the variety of indices themselves. In the introduction to this section, we have already presented the naïve index in Equation (2.1), which can provide quite noisy estimates due to assumptions, which are hard to meet in practice. The most popular form of an index was also mentioned several times. Using year dummy coefficients – for example γ_t from Equation (2.15) – the direct index with a base 100 is given by this formula:

$$\text{Index}_t = 100e^{\gamma_t}. \tag{2.24}$$

The index in this form is called direct, since it is calculated straight from the OLS coefficients, with no further formulas. The index base at 100 can be an arbitrary number, though it's a popular choice for this value. For the base period ($t = 1$), the coefficient is equal to 100 since γ_1 is not estimated and is treated as 0 – therefore $\text{Index}_t = 100e^0 = 100$.

Sometimes, however, the index is given in a different form. For creating his Middle Eastern & Northern African art market index, Kräussl (2015) used the following equation:

$$\text{Index}_t = 100\frac{e^{\gamma_{t+1}}}{e^{\gamma_t}}. \tag{2.25}$$

Locatelli-Biey and Zanola (2002), as well as Białowąs et al. (2018) built their indices with the following recursive formula:

$$\text{Index}_{t+1} = \begin{cases} \text{Index}_t \frac{e^{\gamma_{t+1}}}{e^{\gamma_t}} & \text{if} \quad t > 1, \\ 100 & \text{if} \quad t = 1. \end{cases} \tag{2.26}$$

This can be simplified as follows:

$$\text{Index}_t = 100 \frac{e^{\gamma_t}}{e^{\gamma_1}} = 100 e^{\gamma_t - \gamma_1}. \tag{2.27}$$

Usually, the majority of explained variance in hedonic models comes from the artist variable. It often significantly reduces the analysed number of records, as researchers have to choose a small number of artist dummies in order to provide meaningful models. This enforces selection bias. Using an example of the German art market, Kräussl and Elsland (2008) provided a 2-step hedonic approach, which was aimed at removing the aforementioned obstacle. It uses quality adjustment (Triplett, 2004). The authors analysed 61,135 German auction records from the period between 1985 and 2007 to create the German art index, which was built using the following equation:

$$\text{Index}_{t+1} = \frac{\prod_{i=1}^{n} (P_{i,t+1})^{1/n} / \prod_{i=1}^{m} (P_{i,t})^{1/m}}{\text{HQA}} \tag{2.28}$$

It resembles naïve index from Equation (2.1), but there is one additional element. HQA is an abbreviation for hedonic quality adjustment and represents the mean change of paintings characteristics' influence on a price. It which is given by the following equation:

$$\text{HQA} = \exp \left[\sum_{j=1}^{z} \beta_j \left(\sum_{i=1}^{n} \frac{\mathbf{X}_{ij,t+1}}{n} - \sum_{i=1}^{m} \frac{\mathbf{X}_{ij,t}}{m} \right) \right]. \tag{2.29}$$

Now, by combining equations (2.28) and (2.29) one would obtain

$$\text{Index}_{t+1} = \frac{\prod_{i=1}^{n} (P_{i,t+1})^{1/n} / \prod_{i=1}^{m} (P_{i,t})^{1/m}}{\exp \left[\sum_{j=1}^{z} \beta_j \left(\sum_{i=1}^{n} \frac{\mathbf{X}_{ij,t+1}}{n} \right) \right]}, \tag{2.30}$$

which is quality-corrected hedonic index for the period $t + 1$. The 2-step hedonic approach presented by Kräussl and Elsland (2008) considers estimation of (2.15) on a sub-sample of artists and using the obtained β_j coefficients in Equation (2.30). After some changes, a similar method

can be used to create an artists-specific index. One has to replace prices per period t by prices per artist y and remove the artists dummy from $\mathbf{X}$. The final index for the artist y has the following form:

$$\text{Index}_y = \frac{\prod_{i=1}^{n}(P_{i,y})^{1/n} / \prod_{i=1}^{m}(P_{i,y-1})^{1/m}}{\exp\left[\sum_{j=1}^{z}\beta_j\left(\sum_{i=1}^{n}\frac{\mathbf{X}_{ij,t+1}}{n}\right)\right]}. \tag{2.31}$$

This equation, however, needs to be calculated manually instead of using OLS. Kräussl and Elsland (2008) argue that this approach diminishes the impact of selection bias.

2.2.4 Other techniques

While HR and RSR are the most popular tools for quantitative research of art markets, several other methods facilitating this task also exist. Bocart and Hafner (2012) proposed a heteroskedastic hedonic regression model, which captures time-varying volatility. Instead of time dummies, they used a local likelihood estimator. Using a right-tailed unit root test with forward recursive regressions, Kräussl, Lehnert, and Martelin (2016) argues that there is a speculative bubble in the main art markets since the 2010s. In other research, due to limitations of hedonic regression in modelling some phenomena, Førsund and Zanola (2006) used Data Envelopment Framework (often abbreviated as DEA). This benchmarking technique is known from operations research and was used to evaluate the performance of auction houses selling Picasso's works. In another example, Charlin and Cifuentes (2014) introduced Artistic Power Value (abbreviated as APV), which is dubbed by its authors as an investor-oriented metric – it bases on a price per area unit. Using a data sample of 716 Picasso paintings, Scorcu and Zanola (2011) examined quantile regression. Contrary to OLS procedure which minimises the RSS as in Equation (2.5), this method employs the minimisation of a weighted sum of the absolute values of residuals:

$$\min_{\{\beta_j\}_{j=0}^{k}} \sum_i \left| y_i - \sum_{j=0}^{k}\beta_j x_{j,i} \right| h_i, \tag{2.32}$$

where $y_i - \sum_{j=0}^{k}$ denotes the residual and h_i is the weight of observation i, drawn from the following equation:

$$h_i = \begin{cases} 2q & \text{if} \quad y_i - \sum_{j=0}^{k}\beta_j x_{j,i} > 0, \\ 2-2q & \text{if} \quad y_i - \sum_{j=0}^{k}\beta_j x_{j,i} \leq 0. \end{cases} \tag{2.33}$$

The quantile $0 < q < 1$ is a subject of estimation.

Artnet, which is famous for its vast database of sold artworks, estimates that only 10% can be used in RSR (artnet Analytics, 2014), which is due to lack of provenance information and long time between consecutive sales. Nevertheless, they presented their own index as a blend of RSR and HR. They also demonstrate that the former can be viewed as a nested case of the latter – i.e. RSR can be derived from HR. Following artnet Analytics (2014), to prove that one can start with a hedonic regression for an item i with h characteristics at time t at price p_{i,t_1}, which can be formulated as follows:

$$\ln(p_{i,t_1}) = \sum_{k=1}^{K} \alpha_k h_{i,k} + \sum_{\tau=0}^{T-1} \beta_\tau d_{i,\tau} + \eta_{i,t_1}, \tag{2.34}$$

where $h_{i,k}$ is the k-th characteristics of an item i, and $d_{i,\tau}$ is the indicator variable ($d_{i,0} = 1$ for every item i). Now, a similar equation can be formulated for another item j sold at time t_2 ($t_1 < t_2$) at price p_{j,t_2}:

$$\ln(p_{j,t_2}) = \sum_{k=1}^{K} \alpha_k h_{j,k} + \sum_{\tau=0}^{T-1} \beta_\tau d_{j,\tau} + \eta_{j,t_2}. \tag{2.35}$$

If one would assume that i and j represent the same item and their price difference is the explained variable, by subtracting (2.35) from (2.34) one will obtain the following equation:

$$\ln(p_{i,t_1}) - \ln(p_{j,t_2}) = \sum_{k=1}^{K} \alpha_k h_{i,k} + \sum_{\tau=0}^{T-1} \beta_\tau d_{i,\tau} + \eta_{i,t_1} - \left(\sum_{k=1}^{K} \alpha_k h_{j,k} + \sum_{\tau=0}^{T-1} \beta_\tau d_{j,\tau} + \eta_{j,t_2} \right), \tag{2.36}$$

which – since $\sum_{k=1}^{K} \alpha_k h_{i,k}$ is the same in both equations – after simplification can be written as follows:

$$\ln\left(\frac{p_{i,t_1}}{p_{j,t_2}}\right) = \sum_{\tau=0}^{T-1} \beta_\tau d_{i,\tau} - \sum_{\tau=0}^{T-1} \beta_\tau d_{j,\tau} + \zeta, \tag{2.37}$$

where $\zeta = \eta_{i,t_1} - \eta_{j,t_2}$, i.e. it represents the error term for both items. This equation is a standard RSR form, therefore it shows that HR can be derived from RSR, i.e. RSR is a special case of HR.

Artnet indices operate on *comparable sets*, which can be perceived as sales data for a single artist grouped in an internal review process. These sets are to gather similar lots (both from statistician's and art historian's point of view) and their purpose is to extend the dataset in RSR-based analysis (artnet Analytics, 2014). Figure 2.10 presents a sample comparable set, in which one can observe a group of similar artworks – both in terms of their aesthetics and price.

Figure 2.10: Sample comparable set
Source: artnet Analytics (2014)

Artnet produced two two types of indices: equal-weighted and cap-weighted index.

Following artnet Analytics (2014), theirs equal-weighted index resembles the S&P Equal-Weighted Index, since all entities included are equally important – contrary to the cap-weighted index. It is is constructed assuming a linear relationship in the following equation:

$$\ln(p_{i,s,t}) = \alpha + \frac{1}{N_S}\sum_{t=1}^{T}\sum_{j=1}^{N_{s,t}}\ln(p_{j,s,t}) + \sum_{t=1}^{T}\gamma_t D_{i,t} + \sum_{s=1}^{S}\delta_s C_{i,s} + \varepsilon_{i,s,t}, \tag{2.38}$$

where $\ln(p_{i,s,t})$ denotes the price for a lot $i \in \{1 \dots N\}$ in time $t \in \{1 \dots T\}$ in a comparable set $s \in \{1 \dots S\}$, α is the intercept, N_s is the number of lots in the comparable set s, $N_{s,t}$ similarly marks the number of lots in the comparable set s but limited to year t, $D_{i,t}$ and $C_{i,s}$ are time and comparable set related dummies respectively, γ_t and δ_s are coefficients, which denotes a marginal impact of, consecutively, time t on the logged price and comparable set s, and $\varepsilon_{i,s,t}$ is the corresponding error term. Moving the average of logged prices for time t in a comparable set s to the left one would obtain:

$$\ln(p_{i,s,t}) - \frac{1}{N_S}\sum_{t=1}^{T}\sum_{j=1}^{N_{s,t}}\ln(p_{j,s,t}) = \alpha + \sum_{t=1}^{T}\gamma_t D_{i,t} + \sum_{s=1}^{S}\delta_s C_{i,s} + \varepsilon_{i,s,t}. \tag{2.39}$$

The left-hand side now forms a price corrected by the mean log price in the considered comparable set, which is defined as $Y_{i,s,t}$ by the following equation:

$$Y_{i,s,t} = \ln(p_{i,s,t}) - \frac{1}{N_S}\sum_{t=1}^{T}\sum_{j=1}^{N_{s,t}}\ln(p_{j,s,t}). \tag{2.40}$$

Expected values for equations (2.39) and (2.40) altogether will look as follows:

$$\mathbb{E}(Y_{i,s,t}) = \mathbb{E}\left(\alpha + \sum_{t=1}^{T} \gamma_t D_{i,t} + \sum_{s=1}^{S} \delta_s C_{i,s},\right), \tag{2.41}$$

which can be written in the matrix form as follows:

$$\mathbb{E}(\mathbf{Y}) = \mathbb{E}(\mathbf{\Delta\Gamma}). \tag{2.42}$$

Provided that Gauss-Markov theorem assumptions are satisfied, marginal influences of time dummies can be obtained using the OLS estimator:

$$\hat{\mathbf{\Gamma}} = (\mathbf{\Delta}^\top \mathbf{\Delta})^{-1}(\mathbf{\Delta}^\top \mathbf{Y}). \tag{2.43}$$

Finally, the index (based at 100) for time t can be obtained with the following formula:

$$\text{Index}_t = 100e^{\hat{\gamma}_t}. \tag{2.44}$$

The second index type created by artnet is a cap-weighted index. It resembles S&P Market Cap-Weighted Index, in which more valuable entities are more important for the index, i.e. they have higher weights. This index is described by artnet Analytics (2014) as similar to the equally-weighted one, except for one important difference – the existence of a diagonal weighting matrix $\mathbf{\Omega}$. This matrix is calculated as follows:

$$\Omega_{i,s,t} = \frac{\frac{1}{N_{s,t}} \sum_{j=1}^{N_{s,t}} p_{j,s,t}}{\sum_{s=1}^{S_t} \sum_{j=1}^{N_{s,t}} p_{j,s,t}} \tag{2.45a}$$

$$\mathbf{\Omega} = \text{diag}(\Omega_{1,1,1}, ..., \Omega_{N,S,T}) \tag{2.45b}$$

The rest of calculations is the same, except the estimation of time dummies in (2.43), which are obtained in a slightly different way using $\mathbf{\Omega}$:

$$\hat{\mathbf{\Gamma}} = (\mathbf{\Delta}^\top \mathbf{\Omega\Delta})^{-1}(\mathbf{\Delta}^\top \mathbf{\Omega Y}). \tag{2.46}$$

The usage of weighting matrix $\mathbf{\Omega}$ is similar to Equation (2.22), which was used in Mei Moses Fine Art Index.

The aforementioned artnet indices operate on a single artist. The more holistic market view can be obtained basing on *Artnet C50 Index* (top 50 contemporary artists), *Impressionists Art Index*, and *Modern Art Index*. All of them are called *composite* indices, as they evaluate the performance of multiple artists in a given market sector. For each of these sectors, the list of included artists is obtained using a ranking. Following artnet Analytics (2014), at first, all artists are grouped by their style, date birth & death, and art movement. Then, one has to calculate number of sold lots $N_{a,t}$ for an artist a in year t and their yearly median price (without prints) in that year $\widetilde{M}_{a,t}$. These two numbers are then multiplied:

$$m_{a,t} = \widetilde{M}_{a,t} N_{a,t}. \tag{2.47}$$

Now, one can calculate five years exponential decay based rank for artist a in year t:

$$\text{rank}_{a,t} = m_{a,t-1}e^{0} + m_{a,t-2}e^{-1} + m_{a,t-3}e^{-2} + m_{a,t-4}e^{-3} + m_{a,t-5}e^{-4} + m_{a,t-6}e^{-5}. \tag{2.48}$$

Artists with the highest rank are selected for calculating the index.

Now, one can generate the composite index basing on Equation (2.34). To do so, an equal-weight matrix has to be prepared:

$$\xi_{a,t} = \frac{N_t}{n_{a,t}A_t}, \tag{2.49a}$$

$$\boldsymbol{\Xi} = \text{diag}(\xi_{a,t})_{N\times N}, \tag{2.49b}$$

where N_t is the number of all artworks sold in time t, is the number of all artworks sold $N_{a,t}$ for an artist a in time t, and A_t is the number of artists with at least one transaction in time t. Just as in Equation (2.43), one can use OLS to estimate parameters for equal-weighted composite index:

$$\hat{\boldsymbol{\Gamma}} = (\boldsymbol{\Delta}^\top \boldsymbol{\Xi} \boldsymbol{\Delta})^{-1}(\boldsymbol{\Delta}^\top \boldsymbol{\Xi} \mathbf{Y}). \tag{2.50}$$

Similarly, the procedure for cap-weighted composite index resembles Equation (2.46):

$$\hat{\boldsymbol{\Gamma}} = (\boldsymbol{\Delta}^\top \boldsymbol{\Xi} \boldsymbol{\Omega} \boldsymbol{\Delta})^{-1}(\boldsymbol{\Delta}^\top \boldsymbol{\Xi} \boldsymbol{\Omega} \mathbf{Y}). \tag{2.51}$$

While the whole procedure for composite market generation is interesting, there are some controversies regarding this methodology – namely in the case of C50, the index for the top 50

contemporary artists. Salmon (2012) criticises C50 for its changing group of included artists – it appears to drop out some names twice time faster than S&P 500. This, however, is not the biggest problem – as Salmon reports, the artists are added to the index *after* receiving a wide recognition and achieving high prices. This constitutes selection and survivorship bias since C50 captures only the best performing part of the market. Salmon also notes that S&P 500 would outperform C50 if one would take dividends into account. He also argues that the changing nature of the contemporary art market does not make it homogeneous or investable and C50 can't be treated as a true benchmark. As of 2020, it is hard to find information on C50, which suggests that artnet might have stopped using this index. The remaining two (Impressionists Art Index and Modern Art Index) seem to be discontinued as well, albeit a similar age-based index division is used to this date (for example, in the recent Citibank report[19]).

ArtRank[20] (previously known as SellYouLater) introduced flipping at scale to the art market (Velthuis, 2014). In general, flipping is a speculative practise of buying goods just to sell them at a higher price, preferably in a short time. This practice is well known in the housing market. While speculative actions are nothing new for the art market, treating particular artists in such a way is something new. The site was created in 2014 and the authors claim that it generated a 4200% return in a 16-month period. ArtRank provides rankings on a quarterly basis, in which particular artists are labelled as "sell now", "buy now", or "liquidate". This is also a demonstration of advanced data science and machine learning techniques for the art market, though the exact mechanism of ranking preparation is not known. A brief description on ArtRank's page states: *"The algorithm is comprised of six exogenous components: Presence, Auction results, market Saturation, market Support, Representation and Social mapping (PASSRS). Each component is qualitatively weighted in service of defining a vector or 'artist trajectory'. We compare past trajectories to help forecast early emerging artists' future value. (...) Our purpose-coded machine-learning algorithm extracts relevant explanatory metrics from over three million historic data points including auction results, representation, collectors, and museums. These weighted qualitative metrics work in conjunction with our classification algorithm to identify prime artist prospects based on known trajectory profiles"*. Due to treating artists as a commodity, the site gained a lot of negative press[21] and was criticised in the community. It does seem inactive since

[19] `https://www.privatebank.citibank.com/newcpb-media/media/documents/insights/Citi-GPS-Art-report-Dec2020.pdf` [accessed on March, the 26th, 2022]

[20] `http://artrank.com` [accessed on March, the 14th, 2020]

[21] `http://www.theguardian.com/artanddesign/2014/jun/23/artrank-buy-sell-liquidate-art-market-website-artists-commodities` [accessed on March, the 5th, 2020], `http://www.nytimes.com/2015/02/08/`

2017, though.

2.3 Summary

In this chapter, we sought to understand the phenomenon of the art market analysis and answered the research question Q1 (*Which methods can be used to assess the importance of paintings' features for the hammer price?*). Section 2.1 provided a concise overview of the art market. Section 2.2 presented quantitative methods used to investigate it. Finally, Section 3.4 was devoted to the colour-related quantitative art market analysis.

Section 2.1 provided important contextual information about the art market. The structure of the global art market, as well as some basic auction mechanisms have been introduced in Section 2.1.1. Section 2.1.2 sheds light on the Polish auction houses. The data clearly shows that the Polish auction market is on the rising trend, successfully competing with the other European markets, which makes it an interesting subject of research.

Section 2.2 contains a theoretical introduction to the quantitative analysis of the art market. The main types of used methods has been discussed:

- hedonic regression models (Section 2.2.1),
- repeated-sales regression models (Section 2.2.2).

Different methods for index construction have been briefly discussed in Section 2.2.3. A selection of other techniques used in quantitative analysis of art market data were documented in Section 2.2.4. A good deal of researchers seems to prefer to use hedonic regression in its simplest form. It allows to include a larger dataset in the analysis compared to repeated-sales regression and enables to determine qualities influencing the price. However, both HR and RSR suffer from selection bias.

`magazine/art-for-moneys-sake.html?_r=0` [accessed on March, the 5th, 2020]

Chapter 3

Image Processing Techniques for Colour Analysis

As Wittgenstein (1977) said, "*Our ordinary language has no means for describing a particular shade of colour. Thus it is incapable of producing a picture of this colour*". Fortunately, a number of methods and techniques provide the right *language* to examine the influence of the visual appearance of a painting in terms of colours and answer the research question Q2 (*How to extract colour-related information from paintings?*). However, a broad range of subjects needs to be introduced first. This chapter contains a literature survey on image processing and computer vision methods related to colour analysis, and it comprises three parts. Section 3.1 provides a reasonably comprehensive overview of colour models and spaces, which will be used throughout the rest of this dissertation. Perhaps the two most popular families (CIE and RGB colour spaces) are described. Section 3.2 provides some basic contextual information regarding image processing and computer vision. The notions of features and descriptors are described since some of them constituted a strong encouragement for doing this study. Since the quantisation-related features were particularly inspiring, Section 3.3 is devoted to this phenomenon. The most popular quantisation algorithms are discussed, such as k-means clustering, median cut, and octrees. A subset of quantitative art market analysis papers considering features related to colour is particularly important for the subject of this dissertation. These works are discussed in Section 3.4, which also partially answers the research question Q2.

3.1 Colour Representation

Before discussing any image processing or computer vision techniques for colour analysis, it would be worthy of defining what colour actually is and how it is represented in a digital world. There exist a number of different methods for achieving this, of which RGB is perhaps the most known one. Colour can be defined as a visible feature of light, which can be characterised by its wavelength. A human eye is equipped with three groups of cone cells, which are responsible for *seeing* different wavelengths (Choudhury, 2014). However, for the purpose of this dissertation, we are more concerned with digital means of representing colours. Nevertheless, it would be impossible to describe colour spaces without relating to plain physics. A *colour space* is a way of expressing a given colour by its numerical representation. Usually, it is represented as a three-dimensional vector since a human eye has three aforementioned distinct colour receptors and therefore receives three different stimuli every time it sees something. The notion of a tristimulus colour system refers to this quality.

A *colour model* is an abstract mathematical model used to precisely describe colours, such as RGB or CMYK – using three-dimensional and four-dimensional vectors, accordingly. Colour models can be *additive* (such as RGB) or *subtractive* (such as CMYK). Additivity refers to the following model quality: the higher all the values are, the closer to the white will it be. Subtractive colour models work the other way around – the higher the values are, the more dark the colour is (Szeliski, 2011).

A colour space is a more specific term – it can be perceived as an implementation of a colour model, which maps abstract vectors from the model to real colours. An example of a colour space is sRGB, built around the RGB colour model. In practice, these two terms are used interchangeably, and there is a fine line between them. In the literature, colour spaces usually are divided to two groups: *device-dependent* and *device-independent*. A device-dependent colour space is a colour space in which the perception of a given colour depends on the used display device, such as CMYK- or RGB-related spaces. They are discussed in Section 3.1.2. On the contrary, an object which colour is represented by a device-independent colour space will always look the same, regardless of the used device – CIE colour spaces can be a good example for this category (see Section 3.1.1). Choudhury (2014) enumerates the third category – *internal colour spaces*, which lies somewhere between the aforementioned group. Internal colour spaces standardise a device-independent colour space. He enlists sRGB and Adobe RGB as examples

(we describe them in Section 3.1.2). Another way to divide colour spaces is by relating to how they look to the human perception: they can be perceived as *linear* or *non-linear*. In this case, linearity refers to the difference between two colours in a given space. Some of the CIE colour spaces aim to have this feature.

3.1.1 CIE Colour Spaces

Commission Internationale d'Éclairage (eng. International Commission on Illumination, abbreviated as CIE) is the Vienna-based international organisation concerned with the standards related to light and lighting, such as vision, photometry, or colourimetry. CIE was started in 1913 as a continuation of Commission Internationale de Photométrie. The organisation is widely known for its contribution to colour science – particularly for the CIE XYZ and CIE L*a*b* colour spaces, as well as for the numerous ΔE functions for measuring the colour difference (Schanda, 2007).

CIE XYZ

CIE XYZ is a device-independent colour space, which was presented in 1931 by CIE (Smith & Guild, 1931). However, it is non-linear to human perception (Burger & Burge, 2009a). CIE XYZ is sometimes referenced as the most fundamental and low-level colour space since it is based on how a human eye *sees* and any colour can be expressed by its absolute values, i.e. without a reference to other colours (Choudhury, 2014). It is based on three imaginary colours X, Y and Z, where Y is meant to represent luminance. Following Fairchild (2005), these values are calculated from the following set of equations:

$$X = k \int_{\lambda} \overline{x}(\lambda)\Phi(\lambda)\, d\lambda, \tag{3.1}$$

$$Y = k \int_{\lambda} \overline{y}(\lambda)\Phi(\lambda)\, d\lambda, \tag{3.2}$$

$$Z = k \int_{\lambda} \overline{z}(\lambda)\Phi(\lambda)\, d\lambda, \tag{3.3}$$

where λ is the wavelength in nanometres, $\overline{x}$, $\overline{y}$, $\overline{z}$ are colour matching functions related to the standard observer, $\Phi(\lambda)$ denotes the spectral power distribution of the stimulus, and k is the normalising factor (in absolute colorimetry $k = 683$ lumen/W). The purpose of the CIE 1931 standard observer is to mimic the behaviour of an average human eye by the associated matching

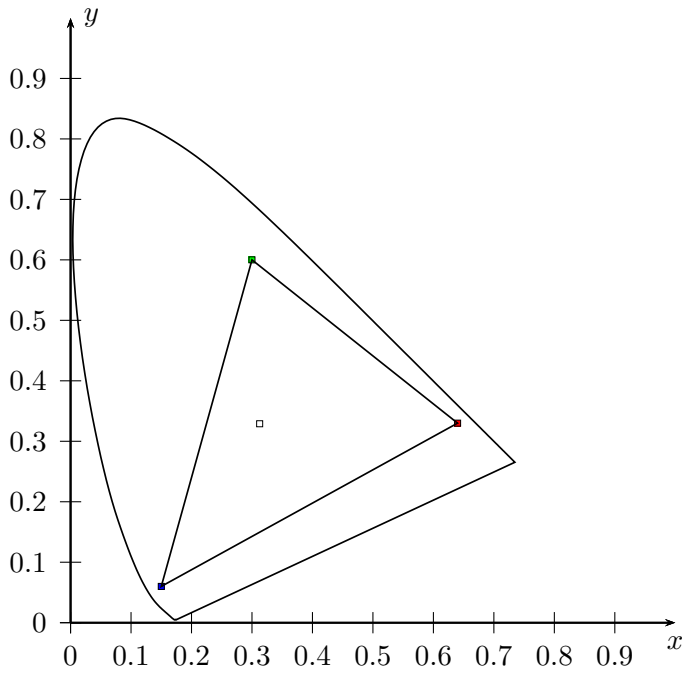


Figure 3.1: CIE chromaticity diagram with the sRGB gamut.

functions:

$$\overline{x}(\lambda)\colon\ \mathbb{R} \mapsto \mathbb{R}_+, \qquad \overline{y}(\lambda)\colon\ \mathbb{R} \mapsto \mathbb{R}_+, \qquad \overline{z}(\lambda)\colon\ \mathbb{R} \mapsto \mathbb{R}_+, \tag{3.4}$$

each with the λ range between 360 and 830 nm. Values of these functions are set by CIE as a standard. Following Fairchild (2005), $\Phi(\lambda)$ depends on the context – it can be defined as a spectral radiance or a relative spectral power distribution for self-luminous stimuli.

Now, three chromaticity values x, y, and z can be derived from these values by using the following projection:

$$x = \frac{X}{X+Y+Z}, \quad y = \frac{Y}{X+Y+Z}, \quad z = \frac{Z}{X+Y+Z} = 1 - x - y. \tag{3.5}$$

Since $x + y + z = 1$, the last variable can be skipped. Drawing the first two values on a two-dimensional plane results in the famous horseshoe-shaped chromaticity diagram. These diagrams are often accompanied by a triangle-shaped *gamut*, which represents the extent of an available range for a given colour space. Figure 3.1 depicts an example of a chromaticity diagram paired with the sRGB gamut.

The neutral point ($x = y = \frac{1}{3}$) is also marked in this figure – it is one of the *standard illuminants*, which are artificial constructs made to mimic light sources in common calculations

Table 3.1: Standard illuminants in CIE XYZ colour space compared.

	Temp.	X	Y	Z	x	y	z
D50	5000° K	0.964296	1.000000	0.825105	0.3457	0.3585	0.2958
D65	6500° K	0.950456	1.000000	1.088754	0.3127	0.3290	0.3583
E	5400° K	1.000000	1.000000	1.000000	0.3333	0.3333	0.3333

Source: Burger and Burge (2009a)

(Fairchild, 2005). Each standard illuminant is described by spectral power distribution and the correlated colour temperature. This one is called E, an absolute neutral point in CIE XYZ. Burger and Burge (2009a) have drawn an interesting conclusion from this diagram – complementary colours lie on the straight lines running through E. There are also other families standard illuminants. Perhaps the most popular one is the D family, with its two representatives – D50 and D65 (Judd et al., 1964). They are meant to represent natural daylight. Hence they are prefixed by the capital D. Technically, D50 and D65 represent daylight with the correlated colour temperature equal to 5000K and 6500K accordingly (Fairchild, 2005). Table 3.1 compares these standard illuminants in terms of the CIE XYZ colour space.

It is worth mentioning that there is also a colour space that directly measures the values received by three stimuli in a human eye. It is called the LMS colour space, from the light, medium, and long spectral sensitivity to the wavelengths, each for the separate stimulus. Because of this quality, this colour space is particularly interesting for studying colour blindness. A D65-normalised mapping between LMS and CIE XYZ can be done with the following linear transformation (Fairchild, 2005):

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.400 & 0.708 & -0.081 \\ -0.226 & 1.165 & 0.046 \\ 0.000 & 0.000 & 0.918 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}. \tag{3.6}$$

CIE L*a*b*

CIE L*a*b* (pronounced sea-lab, sometimes called CIE LAB, often referred simply as Lab) is a device-independent colour space specified by CIE in 1976. Its name refers to the three dimensions: L* denotes the lightness, a* is the red-green component, and b* is the yellow-blue component (depicted in Figure 3.2). The range for L^* is $[0, 100]$, whereas for a^* and b^* it is $[-127, 127]$. In mathematical terms, CIE L*a*b* is a transformation of CIE XYZ to a metric space (Choudhury,

2014). In colour science, it is called uniform colour space, due to the fact that the Euclidean distance ΔE between two colours *should* be the same as the perceived colour difference for a human eye[1]. This concept is called *perceptual uniformity*. Following Burger and Burge (2009a) and Ganczarski (2004), the transformation from CIE XYZ and CIE L*a*b* is standardised by the ISO norm 13655. It is done in the following way:

$$L^* = 116 \cdot Y \ - 16, \qquad a^* = 500 \cdot (X' - Y'), \qquad b^* = 200 \cdot (Y' - Z'), \tag{3.7}$$

where

$$X' = f_1\left(\frac{X}{X_{ref}}\right), \qquad Y' = f_1\left(\frac{Y}{Y_{ref}}\right), \qquad Z' = f_1\left(\frac{Z}{Z_{ref}}\right), \tag{3.8}$$

$$f_1(c) = \begin{cases} c^{1/3} & \text{for } c > 0.008856, \\ 7.787c + \frac{16}{116} & \text{for } c \leq 0.008856. \end{cases} \tag{3.9}$$

Similarly, the CIE L*a*b* to CIE XYZ transformation can be performed using these equations:

$$X = X_{ref} \cdot f_2\left(\frac{a^*}{500} + Y'\right), \qquad Y = Y_{ref} \cdot f_2\left(Y'\right), \qquad Z = Z_{ref} \cdot f_2\left(Y' - \frac{b^*}{200}\right), \tag{3.10}$$

where

$$Y' = \frac{L^* + 16}{116}, \qquad f_2(c) = \begin{cases} c^3 & \text{for } c^3 > 0.008856, \\ \frac{c - 16/116}{7.787} & \text{for } c^3 \leq 0.008856. \end{cases} \tag{3.11}$$

In the both sets of equations, X_{ref}, Y_{ref}, and Z_{ref} denote the reference white point (typically D65).

The difference between two colours $C_1 = \begin{bmatrix} L_1^* & a_1^* & b_1^* \end{bmatrix}^\intercal$ and $C_2 = \begin{bmatrix} L_2^* & a_2^* & b_2^* \end{bmatrix}^\intercal$ is usually called ΔE. In the most basic form, it can be calculated using the standard Euclidean distance and it is called ΔE_{ab}^*:

$$\Delta E_{ab}^* = \|C_1 - C_2\|_2 = \sqrt{(L_1^* - L_2^*)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2} \tag{3.12}$$

However, CIE L*a*b* is not *perfectly* perceptually uniform and better formulas for ΔE do exist,

[1] It turned out to be not so uniform, though – later in this section we present measures better for this purpose than plain Euclidean distance.

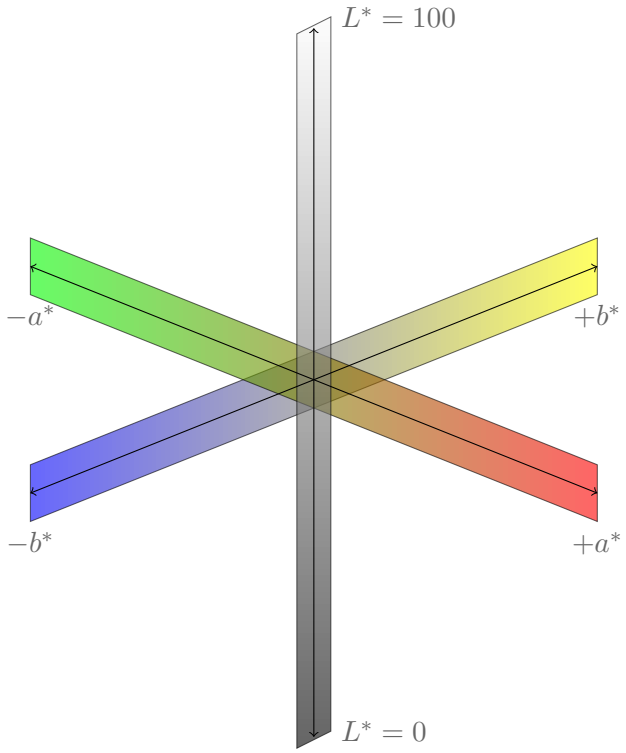


Figure 3.2: CIE L*a*b* colour space – a conceptual perspective.
Based on: Vilson Vieira's code

such as CIE94 and CIEDE2000 (Sharma & Bala, 2002; Habekost, 2013). The subject of colour difference ΔE was studied extensively by Mokrzycki and Tatol (2011). In 1994, CIE tried to tackle the problem of non-uniformity by introducing the new, much more complicated colour difference measure – this time called ΔE^*_{94} (or CIE94):

$$\Delta E^*_{94} = \sqrt{\left(\frac{\Delta L^*}{k_L S_L}\right)^2 + \left(\frac{\Delta C^*_{ab}}{k_C S_C}\right)^2 + \left(\frac{\Delta H^*_{ab}}{k_H S_H}\right)^2}, \tag{3.13}$$

where particular elements are calculated as follows:

$$\Delta L^* = L^*_1 - L^*_2, \qquad C^*_1 = \sqrt{a^{*2}_1 + b^{*2}_1}, \qquad C^*_2 = \sqrt{a^{*2}_2 + b^{*2}_2}, \tag{3.14}$$

$$\Delta C^*_{ab} = C^*_1 - C^*_2, \qquad \Delta a^* = a^*_1 - a^*_2, \qquad \Delta b^* = b^*_1 - b^*_2, \tag{3.15}$$

$$\Delta H^*_{ab} = \sqrt{\Delta E^{*\,2}_{ab} - \Delta L^{*2} - \Delta C^{*\,2}_{ab}} = \sqrt{\Delta a^{*2} + \Delta b^{*2} - \Delta C^{*\,2}_{ab}}, \tag{3.16}$$

$$S_L = 1, \qquad S_C = 1 + K_1 C^*_1, \qquad S_H = 1 + K_2 C^*_1. \tag{3.17}$$

The formula has the form of modified Euclidean distance – there are two noticeable differ-

ences. The first one is the LCH notation (instead of L*a*b*). The second one lies in the introduced two families of scaling coefficients. Following Habekost (2013), the value of k-coefficients $(k_C, k_H, k_L, K_1, K_2)$ depends on the analysed medium, whereas S-coefficients (S_L, S_C, S_K) try to minimise the effect of non-uniformity. The first two of k-coefficients (k_C, k_H) are often set to 1. Therefore, they can be omitted in calculations. The value of the remaining three needs to be known before performing the calculations.

While better than the original ΔE^*_{ab}, the formula for calculating ΔE^*_{94} is not perfect as well – for example, it can perceive the difference between black and green as the same as white and green[2]. In 2000, CIE described another way to measure colour difference – ΔE^*_{00} (also known as CIEDE2000). Following Sharma, Wu, and Dalal (2005) and Z. Schluessler[3], it can be calculated as follows:

$$\Delta E^*_{00} = \sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + R_T \frac{\Delta C'}{k_C S_C}\frac{\Delta H'}{k_H S_H}}. \tag{3.18}$$

This set of meticulous formulas can be viewed as an expansion to ΔE^*_{94} – especially considering the fact that it again uses the LCH notation. The first three terms under the square root look the same as well. There are differences, though – for example, in the calculation of scaling factors, which are calculated as follows:

$$\Delta L' = L^*_2 - L^*_1, \qquad \bar{L} = \frac{L^*_1 + L^*_2}{2}, \qquad \bar{C} = \frac{C^*_1 + C^*_2}{2}, \tag{3.19}$$

$$a'_1 = a^*_1 + \frac{a^*_1}{2}\left(1 - \sqrt{\frac{\bar{C}^7}{\bar{C}^7 + 25^7}}\right), \qquad a'_2 = a^*_2 + \frac{a^*_2}{2}\left(1 - \sqrt{\frac{\bar{C}^7}{\bar{C}^7 + 25^7}}\right), \tag{3.20}$$

$$\bar{C}' = \frac{C'_1 + C'_2}{2}, \qquad \Delta C' = C'_2 - C'_1, \qquad C'_1 = \sqrt{a_1'^2 + b_1^{*2}}, \qquad C'_2 = \sqrt{a_2'^2 + b_2^{*2}}, \tag{3.21}$$

$$h'_1 = \operatorname{atan2}(b^*_1, a'_1) \bmod 360°, \qquad h'_2 = \operatorname{atan2}(b^*_2, a'_2) \bmod 360°, \tag{3.22}$$

$$\Delta h' = \begin{cases} h'_2 - h'_1 & |h'_1 - h'_2| \leq 180° \\ h'_2 - h'_1 + 360° & |h'_1 - h'_2| > 180°, h'_2 \leq h'_1 \\ h'_2 - h'_1 - 360° & |h'_1 - h'_2| > 180°, h'_2 > h'_1 \end{cases}, \tag{3.23}$$

[2] `http://zschuessler.github.io/DeltaE/learn` [accessed on October, the 17th, 2019]
[3] Ibid.

$$\Delta H' = 2\sqrt{C'_1 C'_2}\sin(\Delta h'/2), \quad \bar{H}' = \begin{cases} (h'_1 + h'_2)/2 & |h'_1 - h'_2| \leq 180^\circ \\ (h'_1 + h'_2 + 360^\circ)/2 & |h'_1 - h'_2| > 180^\circ, h'_1 + h'_2 < 360^\circ \\ (h'_1 + h'_2 - 360^\circ)/2 & |h'_1 - h'_2| > 180^\circ, h'_1 + h'_2 \geq 360^\circ \end{cases} \tag{3.24}$$

Similarly to ΔE^*_{94}, k-coefficients (k_L, k_C, k_H) are often set to 1. The main difference between ΔE^*_{94} and CIEDE2000 lies in the last term of these equations, which acts as an extra factor for ensuring perceptual uniformity. S-coefficents stands for compensations for lightness (S_L), chroma (S_C), and hue (S_H) respectively. They are calculated as follows:

$$T = 1 - 0.17\cos(\bar{H}' - 30^\circ) + 0.24\cos(2\bar{H}') + 0.32\cos(3\bar{H}' + 6^\circ) - 0.20\cos(4\bar{H}' - 63^\circ), \tag{3.25}$$

$$S_L = 1 + \frac{0.015\left(\bar{L} - 50\right)^2}{\sqrt{20 + \left(\bar{L} - 50\right)^2}}, \qquad S_C = 1 + 0.045\bar{C}', \qquad S_H = 1 + 0.015\bar{C}'T, \tag{3.26}$$

The hue rotation term R_T is concentrated on ensuring correct calculations for the blue region:

$$R_T = -2\sqrt{\frac{\bar{C}'^7}{\bar{C}'^7 + 25^7}}\sin\left[60^\circ \cdot \exp\left(-\left[\frac{\bar{H}' - 275^\circ}{25^\circ}\right]^2\right)\right]. \tag{3.27}$$

Due to the fact of perceptual uniformity, CIE L*a*b* is applied in Chang's k-means, one of the colour quantisation algorithms considered in this dissertation (see Section 3.3.3). CIEDE2000 equation provides a reliable method of measuring colour distance, and it is applied later in this dissertation in the assessment of colour quantisation algorithms (see Chapter 5).

CIE L*u*v*

Not only CIE L*a*b* was released by CIE in 1976. CIE L*u*v* (often dubbed as CIE LUV, pronounced as sea-love) is another device-independent colour space (L* represents the lightness, u* is for the red-green axis and v* spans from yellow to blue). While it tries to be perceptually uniform as well, CIE L*u*v* differs from the CIE L*a*b* in the whitepoint adaptation – the former uses Judd-style adaptation, whereas the latter uses von Kries transform (Choudhury,

2014). The conversion from CIE XYZ to CIE L*u*v* can be done as follows (Poynton, 2012):

$$L^* = \begin{cases} 903.3\frac{Y}{Y_{ref}} & \frac{Y}{Y_{ref}} \leq 0.008856 \\ 116\left(\frac{Y}{Y_{ref}}\right)^{\frac{1}{3}} - 16 & \frac{Y}{Y_{ref}} > 0.008856 \end{cases}, \qquad u^* = 13L(u' - u'_{ref}), \qquad v^* = 13L(v' - v'_{ref}), \tag{3.28}$$

where the intermediate values u' and v' are calculated as follows:

$$u' = \frac{4X}{X + 15Y + 3Z}, \qquad v' = \frac{9Y}{X + 15Y + 3Z}. \tag{3.29}$$

Similarly to the CIE L*a*b* transformation, X_{ref}, Y_{ref}, and Z_{ref} denotes CIE XYZ coordinates of the reference white point (such as D65). Following Poynton (2012), going from CIE L*u*v* to CIE XYZ requires obtaining intermediate values first:

$$u' = \frac{u^*}{13L^*} + u'_{ref}, \qquad v' = \frac{v^*}{13L^*} + v'_{ref}, \tag{3.30}$$

where u'_{ref} and v'_{ref} refer to the reference white point. Now the x, y coordinates can be calculated:

$$x = \frac{9u'}{6u' - 16v' + 12}, \qquad y = \frac{4v'}{6u' - 16v' + 12}. \tag{3.31}$$

Alternatively, a conversion to CIE L*u*v* values can be done with the following equations:

$$X = Y\frac{9u'}{4v'}, \qquad Y = \begin{cases} Y_{ref}L^*\left(\frac{3}{29}\right)^3, & L^* \leq 0.008856 \cdot 903.3 \\ Y_{ref}\left(\frac{L^*+16}{116}\right)^3, & L^* > 0.008856 \cdot 903.3 \end{cases}, \qquad Z = Y\frac{12 - 3u' - 20v'}{4v'}. \tag{3.32}$$

3.1.2 RGB Colour Spaces

The RGB colour model denotes perhaps one of the most popular and widely known ways of describing colours in a digital world. The name is an abbreviation from the red, green, and blue tristimulus components. RGB is an additive colour model in the sense that each component represents a different light spectrum but added together they make a final colour (see Figure 3.3). Colour in this model is represented by a three-dimensional vector, in which each component has an integer value between 0 and 255. This gives $256^3 = 16,777,216$ possible colours in total. As

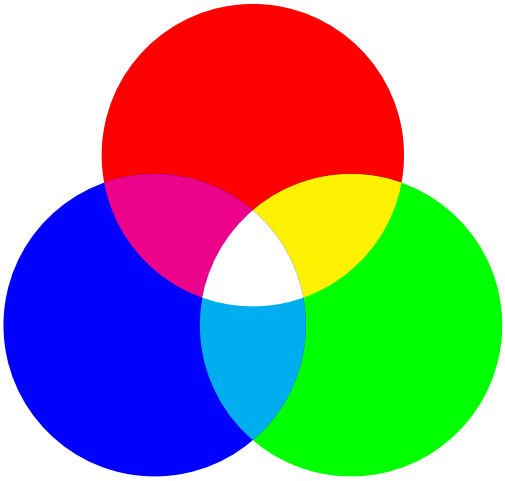


Figure 3.3: Visualisation of the additive nature of RGB.
Source: `http://www.texample.net/tikz/examples/rgb-color-mixing/` [accessed on October, the 17th, 2019]

a closely related colour model, RGBA comes with an additional alpha parameter, which denotes opacity. RGB is related to an infinite number (Choudhury, 2014) of colour spaces – all of which are device-dependent. Bruce Lindbloom[4] enlists a number of popular RGB colour spaces, such as Adobe RGB, Apple RGB, Best RGB, Beta RGB, CIE RGB, ColorMatch TGB, Don RGB 4, ECI RGB v2, Ekta Space PS5, NTSC RGB, PAL/SECAM RGB, ProPhoto RGB, SMPTE-C RGB, sRGB, and Wide Gamut RGB. However, two particular spaces seem to be especially popular and widely used – sRGB and Adobe RGB. They are described in the subsections below.

sRGB

Created jointly by Hewlett-Packard and Microsoft in 1996, sRGB (standard RGB) is often considered as a *default* colour space, due to its ubiquitousness in the Internet and a wide range of applications (e.g. printers, monitors, or digital cameras). It is an international standard[5]. Confusingly, the term RGB is often used as a synonym for sRGB. Following Burger and Burge (2009a), sRGB is a nonlinear colour space with regard to CIE XYZ. These nonlinear components are often denoted as R', G', B'. To transform CIE XYZ values to sRGB, one needs to obtain the linear values first (here denoted as R, G, B). The transformation from CIE XYZ to (linear) sRGB can be done with a linear transformation (the reference white point is D65), using the

[4] `http://brucelindbloom.com/index.html?WorkingSpaceInfo.html` [accessed on October, the 17th, 2019]
[5] `https://webstore.iec.ch/publication/6169` [accessed on October, the 17th, 2019]

transformation matrix $\mathbf{M}_{\mathrm{XYZ}\to\mathrm{RGB}}$:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \mathbf{M}_{\mathrm{XYZ}\to\mathrm{RGB}} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{3.33}$$

where

$$\mathbf{M}_{\mathrm{XYZ}\to\mathrm{RGB}} = \begin{bmatrix} 3.240479 & -1.537150 & -0.498535 \\ -0.969256 & 1.875992 & 0.041556 \\ 0.055648 & -0.204043 & 1.057311 \end{bmatrix}, \tag{3.34}$$

To obtain non-linear components R', G', and B', the gamma correction has to be performed. The transformation from CIE XYZ to sRGB uses the modified gamma correction f_γ ($\gamma = 2.4$, effectively $\gamma \approx 2.2$):

$$R' = f_\gamma(R), \qquad G' = f_\gamma(G), \qquad B' = f_\gamma(B), \tag{3.35}$$

$$f_\gamma(c) = \begin{cases} 12.92c & c \leq 0.0031308 \\ 1.055c^{1/\gamma} - 0.055 & c > 0.0031308 \end{cases} \tag{3.36}$$

The final results need to be scaled from $[0, 1]$ to the $[0, 255]$ range. To convert sRGB to CIE XYZ, the process needs to be reversed. Following Burger and Burge (2009a), one has to start with the inverted gamma correction f_γ^{-1} of non-linear R', G', B' values (each in the range of $[0, 1]$):

$$R = f_\gamma^{-1}(R'), \qquad G = f_\gamma^{-1}(G'), \qquad B = f_\gamma^{-1}(B'), \tag{3.37}$$

$$f_\gamma^{-1}(c') = \begin{cases} 12.92c'^{-1} & c' \leq 0.0031308 \\ \frac{c'+0.055}{1.055}^{\gamma} & c' > 0.0031308 \end{cases}. \tag{3.38}$$

After obtaining the linear values, a linear transformation with inverted $\mathbf{M}_{\mathrm{XYZ}\to\mathrm{RGB}}$ is applied (again, using D65):

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{M}^{-1}_{\mathrm{XYZ}\to\mathrm{RGB}} \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \tag{3.39}$$

where

$$\mathbf{M}_{\mathrm{RGB}\to\mathrm{XYZ}} = \mathbf{M}^{-1}_{\mathrm{XYZ}\to\mathrm{RGB}} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix}. \tag{3.40}$$

Adobe RGB

Despite its omnipresence, sRGB has one significant drawback – a relatively small gamut (see Figure 3.1 in Section 3.1.1). In 1998, Adobe tried to tackle this problem by creating its own colour space – Adobe RGB[6]. It has a significantly larger gamut, which makes it popular in e.g. printing industry, or among professionals photographers. Coordinates in Adobe RGB can be obtained from CIE XYZ values using Equation 3.33, but with different transformation matrices $\mathbf{M}$. Another difference is the gamma correction (now with $\gamma = 2.199$). The transformation matrices $\mathbf{M}$ (again, the reference white is D65) are given below:

$$\mathbf{M}_{\mathrm{XYZ}\to\mathrm{RGB}} = \begin{bmatrix} 2.0413690 & -0.5649464 & -0.3446944 \\ -0.9692660 & 1.8760108 & 0.0415560 \\ 0.0134474 & -0.1183897 & 1.0154096 \end{bmatrix}, \tag{3.41}$$

$$\mathbf{M}_{\mathrm{RGB}\to\mathrm{XYZ}} = \mathbf{M}^{-1}_{\mathrm{XYZ}\to\mathrm{RGB}} = \begin{bmatrix} 0.5767309 & 0.1855540 & 0.1881852 \\ 0.2973769 & 0.6273491 & 0.0752741 \\ 0.0270343 & 0.0706872 & 0.9911085 \end{bmatrix}. \tag{3.42}$$

RGB models are ubiquitous – also in cultural economics. Few researchers even measure RGB values directly in colour-related quantitative art market research (see Chapter 2). Numerous colour quantisation algorithms also use them (see Section 3.3). In this dissertation, the RGB colour model is chosen to describe representative colours (see Chapter 5 and 6).

3.1.3 Other Colour Models

Arguably, RGB and CIE models are the most widely used ones. However, both of them are not universal. Their definition is usually considered as far from intuitive for human viewers. To overcome this approach, a family of different models has been invented – HSV, HSL, and HSI. On contrary, specific industries prefer colour models different from the ones already presented. To name a few, CMYK is very popular in e.g. home printers, whereas YUV and YCbCr is standard

[6] `https://www.adobe.com/digitalimag/pdfs/AdobeRGB1998.pdf` [accessed on October, the 17th, 2019]

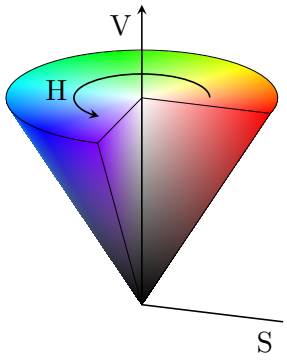


Figure 3.4: The HSV cone – a conceptual representation.
Source: `https://tex.stackexchange.com/a/330274` [accessed on October, the 17th, 2019]

for colour television. This subsection briefly introduces the aforementioned colour models.

HSV, HSL, and HSI

Instead of RGB, the same gamut can be obtained using more intuitive colour models: HSV (hue, saturation, value), HSL (hue, saturation, lightness), or HSI (hue, saturation, intensity). All of them can be obtained using specific transformations on the RGB model. These colour spaces are sometimes called cylindrical since their graphic representation can be fitted into this geometric figure. These colour spaces are not standardised and perceived as non-linear, but they are relatively easy to understand (Choudhury, 2014). HSV (sometimes called HSB, where B is for brightness) and HSL (sometimes referred to as HLS) were developed in the 1970s. In both models, the hue attribute refers to the same value and is given in degrees, as it represents the angle around the vertical axis (see Figure 3.4, which depicts the cone-shaped representation of HSV).

The transformation from RGB to HSV can be defined as follows[7]. At first, one need to normalise RGB values to $[0, 1]$:

$$R' = \frac{R}{255}, \qquad G' = \frac{G}{255}, \qquad B' = \frac{B}{255}. \tag{3.43}$$

After that, the difference Δ between the maximum and minimum colour can be obtained:

$$C_{max} = \max(\{R', G', B'\}), \qquad C_{min} = \min(\{R', G', B'\}), \qquad \Delta = C_{max} - C_{min}. \tag{3.44}$$

[7]Based on `https://www.rapidtables.com/convert/color/rgb-to-hsv.html` [accessed on October, the 17th, 2019]

Finally, the HSV values can be determined by the following set of equations:

$$H = \begin{cases} 0^\circ, & \Delta = 0 \\ 60^\circ \cdot \left(\frac{G'-B'}{\Delta} \bmod 6 \right), & C_{max} = R' \\ 60^\circ \cdot \left(\frac{B'-R'}{\Delta} + 2 \right), & C_{max} = G' \\ 60^\circ \cdot \left(\frac{R'-G'}{\Delta} + 4 \right), & C_{max} = B' \end{cases}, \quad S = \begin{cases} 0, & C_{max} = 0 \\ \frac{\Delta}{C_{max}}, & C_{max} \neq 0 \end{cases}, \quad V = C_{max}. \tag{3.45}$$

The conversion from HSV to RGB is given by the following set of equations[8]. One has to start with the following calculations:

$$C = V \cdot S, \qquad X = C \left(1 - \left| \frac{H}{60^\circ} \bmod 2 - 1 \right| \right), \qquad m = V - C. \tag{3.46}$$

After that, intermediate values R', G', B' can be calculated depending on the hue H – final values are obtained by adding m and scaling the result to the proper RGB range:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{cases} \begin{bmatrix} C & X & 0 \end{bmatrix}^\intercal, & 0^\circ \leq H < 60^\circ \\ \begin{bmatrix} X & C & 0 \end{bmatrix}^\intercal, & 60^\circ \leq H < 120^\circ \\ \begin{bmatrix} 0 & C & X \end{bmatrix}^\intercal, & 120^\circ \leq H < 180^\circ \\ \begin{bmatrix} 0 & X & C \end{bmatrix}^\intercal, & 180^\circ \leq H < 240^\circ \\ \begin{bmatrix} X & 0 & C \end{bmatrix}^\intercal, & 240^\circ \leq H < 300^\circ \\ \begin{bmatrix} C & 0 & X \end{bmatrix}^\intercal, & 300^\circ \leq H < 360^\circ \end{cases}, \qquad \begin{bmatrix} R \\ G \\ B \end{bmatrix} = 255 \begin{bmatrix} R' + m \\ G' + m \\ B' + m \end{bmatrix}. \tag{3.47}$$

Regarding HSL, values in this colour model can be obtained in a partially similar way[9]. The RGB values have to be normalised in the same fashion as in Equation 3.43. Then, the difference between extrema is calculated exactly as in Equation 3.44. Since H is the very same as in HSV, this value is obtained just as in Equation 3.45. The difference lies in the calculation of the remaining components. While S stands for saturation once again, its value is different from the S in HSV. The last component L represents the lightness. These two can be calculated in the

[8]Based on `https://www.rapidtables.com/convert/color/hsv-to-rgb.html` [accessed on October, the 17th, 2019]

[9]Based on `https://www.rapidtables.com/convert/color/rgb-to-hsl.html` [accessed on October, the 17th, 2019]

following way:

$$S = \begin{cases} 0, & \Delta = 0 \\ \frac{\Delta}{1-|2L-1|}, & \Delta \neq 0 \end{cases}, \qquad L = \frac{C_{max} - C_{min}}{2}. \tag{3.48}$$

The backward conversion – HSL to RGB – once again resembles the process of going from HSV to RGB[10]. The supporting parameter X is the same as in 3.46, whereas C and m are calculated differently:

$$C = (1 - |2L - 1|) \cdot S, \qquad m = L - \frac{C}{2}. \tag{3.49}$$

The actual RGB values are now calculated exactly as in Equation 3.47.

HSI (an abbreviation from hue, saturation, intensity) is a model, which is worth mentioning due to decoupling colour intensity from the colour itself. This feature makes this model particularly interesting for various computer vision purposes (Choudhury, 2014). Following Luo, Lin, Yu, and Chen (2013), there exist a number of ways to calculate HSI values. Nevertheless, hue and saturation are calculated differently compared to HSV and HSL colour models. For example, Hoy (1997) proposed RGB to HSI transformation, which is given be the following set of equations:

$$H = \arctan\left(\frac{\sqrt{3}(G-B)}{(R-G)+(R-B)}\right), \qquad S = 1 - \left(\frac{\min(\{R,G,B\})}{I}\right), \qquad I = \frac{R+G+B}{3}. \tag{3.50}$$

CMY and CMYK

A number of other colour models exist. For example, in the printing industry, one can encounter CMY (abbreviation from Cyan, Magenta, Yellow) colour model. Contrary to RGB, it is a subtractive colour model, which means that the primary colours are subtracted from black. CMY colour space is rather uncommon – since it can't produce a true black (Choudhury, 2014), much more popular CMYK (Cyan, Magenta, Yellow, Black) was introduced. To obtain CMYK values from RGB, these values need to be normalised to the range $[0, 1]$. After that, the actual

[10]Based on `https://www.rapidtables.com/convert/color/hsl-to-rgb.html` [accessed on October, the 17th, 2019]

conversion[11] can take place:

$$R' = \frac{R}{255}, \qquad G' = \frac{G}{255}, \qquad B' = \frac{B}{255}, \tag{3.51}$$

$$C = \frac{1 - R' - K}{1 - K}, \qquad M = \frac{1 - G' - K}{1 - K}, \qquad Y = \frac{1 - B' - K}{1 - K}, \qquad K = 1 - \max\left(\left\{R', G', B'\right\}\right). \tag{3.52}$$

Going from CMYK to RGB[12] is even simpler:

$$R = 255(1 - C)(1 - K), \qquad G = 255(1 - M)(1 - K), \qquad B = 255(1 - Y)(1 - K). \tag{3.53}$$

YUV and YCbCr

Another two colour spaces defined on the basis of RGB are YUV and YCbCr. The first one – YUV – is popular in analogue colour TV broadcasting (Choudhury, 2014). Y represents the luminance, whereas U and V are the colour components. It is derived from RGB using the following set of equations (Choudhury, 2014):

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \mathbf{M}_{\text{RGB}\to\text{YUV}} \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \qquad \mathbf{M}_{\text{RGB}\to\text{YUV}} = \begin{bmatrix} 0.2989 & 0.5866 & 0.1145 \\ -0.147 & 0.289 & 0.436 \\ 0.615 & -0.515 & -0.100 \end{bmatrix}. \tag{3.54}$$

While YUV is used for analogue purposes, YCbCr is more popular in the digital world. Y represents luminance, Cr denotes the red-to-green component, and Cb is blue-to-yellow. The transformation from RGB can be done in the following way (Choudhury, 2014):

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \mathbf{M}_{\text{RGB}\to\text{YCbCr}} \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \qquad \mathbf{M}_{\text{RGB}\to\text{YCbCr}} = \begin{bmatrix} 0.2989 & 0.5866 & 0.1145 \\ -0.1688 & -0.3312 & 0.5000 \\ 0.5000 & -0.4184 & -0.0816 \end{bmatrix}. \tag{3.55}$$

Few quantitative art market researchers have noticed and used a natural interpretation of some RGB-based colour models attributes, such as hue, saturation or lightness. They measured these values on their datasets and incorporated them into their research (see Section 3.4).

[11] Based on `https://www.rapidtables.com/convert/color/rgb-to-cmyk.html` [accessed on October, the 17th, 2019]

[12] Based on `https://www.rapidtables.com/convert/color/cmyk-to-rgb.html` [accessed on October, the 17th, 2019]

3.2 Colour-Related Features and Descriptors in Computer Vision and Image Processing

Computer vision – often abbreviated as CV – is a multidisciplinary field concerned with automated *understanding* pictures and videos by computers. It can be viewed as a subfield of artificial intelligence, with a number of applications such as object recognition, image segmentation, or face detection. Computer vision often uses various image processing methods, which – in general – take an image as an input and produce a new image as an output (image segmentation is a notable example). Computer vision and image processing are frequently used interchangeably in colloquial contexts – they often use the very same methods and the only difference lies in their goals (*understanding* versus *obtaining* a new image). This section is concerned with colour-related *features* and *descriptors*, which lays the foundation in numerous CV tasks.

3.2.1 Features and Descriptors

The notion of a *feature* (often accompanied by detection and extraction) plays a vital role in computer vision and image processing (Szeliski, 2011). It seems that there is no agreement among researchers regarding the exact feature definition. Essentially, it means roughly the same thing as in other machine learning applications – though in CV there exists an informal list of pre-defined features. There is no standard group of features for every CV task, and it is rather an evolving set. Some of them, however, are well-defined (Chowdhury, Verma, Tom, & Zhang, 2015). The classical CV pipeline often includes feature detection (extraction) and feature description. Following Szeliski (2011), the former stage addresses the problem of searching for features within the image, whereas the latter converts it to a *descriptor*. Fisher et al. (2013) explain the difference between features and descriptors in the following way. An image feature is defined as a general term describing some interesting image structure, such as particular points, curves, edges, surfaces, etc. An image descriptor[13] is a set of short vectors, which should be compact and invariant to popular transformations (for example, invariant to affine transformations) and can be used for a comparison with other images. In this context, a feature is a broader term than a descriptor. Descriptors (or features as well) can be later used for e.g. image-based queries (Royo, 2010). An example of a popular descriptor is Scale Invariant Feature Transform, widely known as SIFT (Lowe, 2004).

[13]Related terms, such as shape or gist image descriptors are beyond the scope of this work – in this dissertation, by *descriptor* we mean an *image descriptor*.

Salahat and Qasaimeh (2017) describe the desired characteristics of a hypothetical ideal feature: *distinctiveness*, *locality*, *quantity*, *accuracy*, *efficiency*, *repeatability*, *invariance*, and *robustness*. *Distinctiveness* ensures that there is a decent number of variations of the resulting feature. *Locality* means that it should be concerned only with a small fragment of a given image. *Quantity* of detected features is sufficient if it is large enough (but not too large) to capture the image characteristics. *Accuracy* relates to the ability for finding features under different conditions. *Efficiency* is connected to the order of time complexity of used algorithms. *Repeatability* ensures reproducible features. *Invariance* minimises the effect of large deformation, whereas *robustness* deals with the smaller ones.

While there is no standard set of features, a number of scholars gathered the most popular ones according to the purpose. They can be divided into local and global ones – the former describes low-level (or pixel-level) features, whereas the latter relates to the whole image (Wei, Phung, & Bouzerdoum, 2016). For instance, Deselaers, Keysers, and Ney (2008) examined features in the problem of content-based image retrieval. They enlisted appearance-based image features, colour histograms, Tamura features, global texture descriptor, Gabor features, invariant feature histograms, local image descriptors, and MPEG-7 features. They can refer to different image qualities – colour, texture, shape, or local features. From this list, the colour-related ones are appearance-based image features, colour histograms, invariant feature histograms, and some of the local- and MPEG-7-related features. While the overall list is certainly non-exhaustive, this dissertation is significantly concerned with the notion of colours. Therefore, the remainder of this section is devoted to this particular group of features.

Following Deselaers et al. (2008), *appearance-based image features* can be treated as baseline for some purposes (such as medical radiographs). Two images are scaled to the same size (such as 32×32 pixels). The similarity between images is calculated using the Euclidean distance. More sophisticated comparison measures exist as well, such as the image distortion model (Keysers, Deselaers, Gollan, & Ney, 2007). Since pixels directly correspond to band values in a given colour space, this type of feature is clearly colour-related.

Colour histograms constitute a popular approach in numerous computer vision tasks (Reinhard & Pouli, 2011; Munisami, Ramsurn, Kishnah, & Pudaruth, 2015; Wong et al., 2016). They are often used as a baseline for more sophisticated methods as well. A colour histogram is a distribution of some colour bands within the image (such as separate RGB channels). It is a statistic concerned solely with the colour occurrence – shape and texture qualities are omitted in this

feature. Jensen-Shannon divergence (abbreviated as JSD) may be used to evaluate the similarity of two histograms (Puzicha, Buhmann, Rubner, & Tomasi, 1999). It is calculated as follows:

$$d_{\mathrm{JSD}}(H, H') = \sum_{m=1}^{M} H_m \log \frac{2H_m}{H_m + H'_m} + \log H'_m \frac{2H'_m}{H_m + H'_m}, \tag{3.56}$$

where H and H' are the examined histograms and m represents a single bin (out of M bins) in a given histogram.

A feature is called *invariant* if popular transformations (such as rotation, translation, or scaling) do not affect it. In the domain of computer vision, this property is very desirable since it substantially facilitates many tasks (e.g. two photographs of the same building from different angles can be matched in a comparison). As a special case of histograms, Deselaers et al. (2008) mention *invariant feature histograms*. They explore pixel intensities using monomial and relational functions.

Deselaers et al. (2008) also describe *local image descriptors*, which are concerned with *image patches*. A patch is just a small part of a given image. There are three methods described related to these descriptors. In the first one, patches at salient points are extracted (Deselaers, Keysers, & Ney, 2005) and clustered. After that, PCA is applied to reduce the dimensionality and create a 2048-dimensional vector. The second method (Mikolajczyk et al., 2005) also makes the feature clusters but calculates their statistics, such as mean and variance. The last method (Paredes, Pérez, Juan, & Vidal, 2001) places all features of all considered images in a KD-tree for efficient neighbour search.

Naturally, there are many other colour-related features described in the literature. For instance, Shahbahrami, Borodin, and Juurlink (2008) describe colour moments and colour coherence vectors. In another example, Wei et al. (2016) carefully examines other state-of-the-art visual descriptor types. The MPEG-7 defines even more of them – since it is a well-known standard, the next subsection is devoted to its set of features.

3.2.2 MPEG-7 Colour Descriptors

Starting in 2002, Moving Picture Experts Group (abbreviated as MPEG) published a set of ISO/IEC standards for describing multimedia content, MPEG-7, which was meant to provide interoperability between audio-visual systems (Shih-Fu Chang, Sikora, & Purl, 2001). The standard was later amended end extended multiple times, though the work on the first documents

dates back to 1999. It was meant to be a continuation of MPEG-1, MPEG-2, and MPEG-4. MPEG-7 consists of e.g. a set of *description schemes* and *descriptors*. This time, however, a descriptor means something slightly broader than in the previous definition. Albeit it is used as a kind of synonym for a feature, it can be used to describe much more general characteristics (such as colour space – see below).

The MPEG-7 standard enlists four groups of descriptors: colour descriptors, texture descriptors, shape descriptors, and motion descriptors. Since this dissertation is concerned mostly with colours, we describe only the colour-related ones. Following Ohm et al. (2002), a number of colour-related descriptors are defined in the MPEG-7 standard: *Colour Space Descriptor* (CSD) with *Colour Quantisation Descriptor* (CQD), *Dominant Colour Descriptor* (DCD), *Scalable Colour Descriptor* (SCD), *Group of Frames/Group of Pictures Descriptors* (GoF/GoP), *Colour Structure Descriptor* (CSD), and *Colour Layout Descriptor* (CLD). While the standard defines these descriptors, it does not exactly specify how particular values are calculated.

Colour Space Descriptor (CSD) is not a feature in a strict computer vision sense – it specifies which colour space is used. This information is important for the other descriptors. According to the MPEG-7, the available colour spaces are RGB, YCbCr, HSV, HMMD, monochrome, and any linear transformation of the RGB colour model in the form of 3×3 matrices (Martínez, 2003). The first three spaces are described in Section 3.1. The monochrome colour space is just the Y component from the YCbCr colour space. Manjunath, Salembier, and Sikora (2002) define the HMMD (an abbreviation for Hue-Max-Min-Diff) colour space, which bases on the transformation of the RGB colour space and consists of four components: hue H, the colours with the highest and lowest values (Max and Min respectively), and the difference between them ($Diff$). The hue component is obtained exactly as in HSV. The rest is calculated using the following set of equations:

$$Max = \max\{R, G, B\}, \qquad Min = \min\{R, G, B\}, \qquad Diff = Max - Min. \tag{3.57}$$

To discretise continuous colour values, *Colour Quantisation Descriptor* (CQD) specifies the number of bins for uniform quantisation for a chosen CSD (see the next subsection for an extensive literature review on the topic of colour quantisation). This applies to all colour spaces, except for HMMD, in which the standard defines four non-uniform methods of quantisation (Ohm et al., 2002). CQD is connected to *Dominant Colour Descriptor* (DCD), which gives an overview

of the representative colours of a given image. Following Ohm et al. (2002), it can be defined as follows:

$$F = \{\{\mathbf{c}_i, p_i, v_i\}, s\}, \tag{3.58}$$

where, for N dominant colours, $\mathbf{c}_i$ is a colour vector denoting the dominant colour in a colour space given by CSD, whereas p_i represents the share (in percents, normalised) of this colour such that $0 \leq p_i \leq 1, \sum_i p = 1$. The colour variance v_i is an optional parameter, which should describe the variations of the original colours (i.e. before the process of quantisation). The last parameter s represents the overall spatial coherency. Ohm et al. (2002) suggest the usage of the Generalised Lloyd's Algorithm for the process of DCD extraction (see Section 3.3).

Scalable Colour Descriptor (SCD) is just a colour histogram made using the HSV colour space with applied Haar wavelet transform. The histogram values are normalised to 4-bit integers. Such a transformation uses the Haar matrix, which is a $2^n \times 2^n$ square matrix, where 2^n ($n \in \mathbb{N}$) is the number of histogram bins (Porwik & Lisowska, 2004). Since the MPEG-7 standard bounds the number of bins to the power of 2, the Haar transform in the form $\mathbf{h}_{\text{SCD}} = \mathbf{H}_{2^n \times 2^n} \mathbf{h}_{\text{HSV}}$ can be used ($\mathbf{H}_{2^n \times 2^n}$ is the Haar matrix of the appropriate size). For example, for a 4-bin histogram $\mathbf{h}_{\text{HSV}}$, the transformation is conducted as follows:

$$\mathbf{h}_{\text{SCD}} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ \sqrt{2} & -\sqrt{2} & 0 & 0 \\ 0 & 0 & \sqrt{2} & -\sqrt{2} \end{bmatrix} \mathbf{h}_{\text{HSV}}. \tag{3.59}$$

Group of Frames/Group of Pictures (GoF/GoP) is another histogram-based descriptor dedicated for videos (GoF) and a group of pictures (GoP). Essentially, GoF/GoP is an aggregation of multiple SCD-s for separate frames/pictures (Manjunath et al., 2002). Three types of aggregations are available: average, median, and intersection. The first two works using classical statistics, mean and median, for a given set of histograms. The last one, the intersection aggregation, uses the minimum value in a group of histograms for particular bins.

Colour Layout Descriptor (CLD) is designed to capture the spatial distribution of colours. Following Manjunath et al. (2002) and Rayar (2017), it uses the YCbCr colour space and consists of four stages. At first, the image is divided into 64 blocks. In the next step, a representative colour is select for each block (an average is recommended). These colours are later represented

as $8 \times 8 \times 3$ tensor and transformed using the discrete cosine transform (DCT), each band as a separate 8×8 matrix. After that, they are zig-zag[14] scanned and quantised using the AC and DC coefficients from DCT. The resulting compact feature can be later used for e.g. fast image retrieval purposes.

Colour Structure Descriptor (CSD) tries to capture colour distribution and its spatial structure (Buturovic, 2005). Following Manjunath et al. (2002), at first a histogram $\mathbf{h}_{\text{CSD}}$ (in HMMD colour space, quantised) is created. It consists of 2^n elements ($n \in \{5, 6, 7, 8\}$) This descriptor uses a structuring element, which resembles the behaviour of a filter in convolutional neural networks. A moving square of 8×8 size is sliding through the whole image with a unit stride (there's no padding). For non-standard size images, subsampling is used. The subsampling factor $K = 2^p$ is calculated using the following equation:

$$p = \max\left\{0, \left\lfloor \log_2 \sqrt{wh} - 7.5 \right\rfloor\right\}, \tag{3.60}$$

where w and h represent picture width and height. The second quantisation divides the HMMD colour spaces into five subspaces using the pre-defined intervals. Each subspace is later uniformly quantised.

The last-mentioned colour descriptor of the MPEG-7 standard becomes an inspiration for describing an image in terms of its colours, which is one of the central ideas behind this dissertation. There are many more ways to choose colours representative of a given picture histogram-based methods, to which Section 3.3 is dedicated.

3.3 Colour Quantisation and Palette Design

Thanks to the colour vision ability, a human eye is capable of distinguishing approximately 200 intensity levels of red, green, and blue, which results in roughly 10 million colours (Gervautz & Purgathofer, 1988). RGB colour space can represent even more. To carry a full RGB colours palette, 24 bits are needed, each 8 for every channel (often referred as *true colour*). As we presented in Section 3.1, this will result in $256^3 = 16,777,216$ colours in total, which exceeds the human perception by a huge margin. However, for decades, both numbers were too large for computer display and memory capabilities. There was a need for a method, which will reduce

[14] `http://www.cmlab.csie.ntu.edu.tw/cml/dsp/training/coding/jpeg/jpeg/encoder.htm` [accessed on October, the 17th, 2019]

the number of colours. For example, at some point video memory allowed to use 16 colours (later this was increased to 256). Colour quantisation methods aim to tackle these problems. Computer graphics hardware is far more advanced nowadays, though the problem of reducing the number of colours is still valid. For instance, GIF, a popular image format, operates on a 256-bit palette[15]. It is also an important part of the JPEG 2000 standard (Marcellin et al., 2002).

Colour quantisation can be defined as a process of reducing the number of colours in digital images with minimal visual distortion, which can also be formulated as lossy image compression (Brun & Trémeau, 2003). In the more general context, the notion of *quantisation* means the digitalisation of a continuous signal in the domain of signal processing (Xiang, 2007). Formally the problem of colour quantisation can be formulated as follows. Consider an array of n vectors $\mathbf{I}_C = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \dots & \mathbf{p}_n \end{bmatrix}$ in some colour space (for example $\forall_i \mathbf{p}_i \in [0, 255]^3$ for RGB) representing pixel colours in the original image $\mathbf{I}_C$ (flattened to 1D for simplicity here). There are d unique vectors in $\mathbf{I}_C$ (i.e. $\mathbf{I}_C$ has d distinct colours). One wants to obtain a new image $\mathbf{I}_Q$, which has the same number of pixels n, but with only k distinct colours (typically $k \ll d$). The goal of colour quantisation is finding a mapping from $\mathbf{I}_C$ to $\mathbf{I}$, which minimises the distortion between them (sometimes called quantisation error). The distortion quantifies the difference between the new and original image. Roughly speaking, the colour quantisation process consists of two phases: palette design (sometimes referred as search for representatives) and pixel mapping (Ozturk, Hancer, & Karaboga, 2014).

There are two topics closely related to the problem of colour quantisation: colour segmentation and palette design. The problem of segmentation considers partitioning images in a way such that resulting disjoint and homogenous regions represent different depicted objects (Lucchese & Mitra, 2001). Following Lucchese and Mitra (2001), methods for colour image segmentation can be feature-space based (clustering, histogram thresholding), image-domain based (split-and-merge, region growing, edge-based, and neural network techniques), or physics-based. Some algorithms in these methods are used in colour quantisation (for example, k-means in clustering or octree in split-and-merge methods). Due to the different problem formulations, the majority of the methods require different approaches (such as edge detection, which can be ignored in colour quantisation). The second topic, palette generation, is rather loosely formulated and considers

[15]Though GIF uses a reduced 256-bit colour palette, obtaining true colour is possible by using multiple image blocks, each with a different palette.

creating a set of colours, which is visually appealing, well-harmonised and can be reused for design purposes (Obrador, 2006; Morse, Thornton, Xia, & Uibel, 2007). Colour harmonisation is also a subject of separate research (Cohen-Or, Sorkine, Gal, Leyvand, & Xu, 2006). Palette can be generated from existing images. In fact, this is the first part of colour quantisation, which is the main part of the resemblance of these two domains. There is no need to minimise quantisation error, however, so results can be quite different. Some scholars generate their palettes and use them to transfer colours to other images (Lin & Hanrahan, 2013; Chang, Fried, Liu, DiVerdi, & Finkelstein, 2015).

Although the topic has been explored by scientists at least since the 1970s, there exist only a few extensive literature reviews devoted to the topic of colour quantisation. The one provided by Brun and Trémeau (2003) covers the topic in detail. Similarly, the review written by Xiang (2007) explored the problem of approaches colour quantisation by treating it as an approximation problem. Both surveys discuss a wide range of algorithms. Burger and Burge (2009b) discuss only the most popular methods, but they provide extensive code listings with corresponding figures, which facilitates the implementation. Scheunders (1997) meticulously analyses clustering algorithms applied to the problem of colour quantisation. While not being strict literature surveys, some publications also provide a comprehensive review of the means of colour quantisation. These algorithms are often compared in terms of their quantisation error. For example, Celebi (2011) investigated possible ways of improvement of k-means. A number of variations of this approach were compared with 11 other colour quantisation algorithms. In another example, a more recent survey was conducted by Ozturk et al. (2014), in which he shortly reviews the available algorithms before comparing theirs with them. Other scholars investigate the usage of colour quantisation in a particular context, such as in the JPEG 2000 standard (Marcellin et al., 2002). It is also worth mentioning *The Graphics Gems Series*, in which numerous colour quantisation algorithms (such as octrees or uniform methods) have been described (Glassner, 1990; Arvo, 1991; Kirk, 1992; P. S. Heckbert, 1994; Paeth, 1995).

There are several ways to classify colour quantisation algorithms. Brun and Trémeau (2003) presented a very nuanced classification. They have written about *pre clustering* and *post clustering* methods, in which the difference lies in the number of times the representative colours are calculated – once in pre clustering methods, while the latter type enforces recalculations and iterative improvements. Xiang (2007) enlists two main categories of colour quantisation: *Image-independent* and *image-dependent*. The latter category can be divided to *context-free* and

context-sensitive methods. Another distinction focuses on the quality of being colour-space neutral (or not). Following Ozturk et al. (2014), there exist a number of algorithms that can be applied to the colour quantisation problems, but most of them can be classified into two kinds of methods: *splitting* and *clustering-based*.

Among image-independent colour quantisation methods, Xiang (2007) enlists uniform quantisation, trellis-coded quantisation, and sampling by Fibonacci lattice. As image-dependent methods (context-free), they enumerate popularity method, detecting peaks in histograms, Peano scan, median cut, center cut, octree quantisation, agglomerative clustering, k-means, minimising total quantisation error, and minimising maximum intercluster distance. Finally, for the context-dependant, they mention dithered image-dependent methods and feedback-based quantisation.

The problem of colour quantisation focuses on minimising the quantisation error. Quantisation algorithms can also be compared using their time and space complexity. Sometimes, execution time is also reported – especially in the past, where computational power was rather modest and this was a critical factor. Usually, the evaluation of quantisation algorithms uses the standard mean squared error formula (which is to be minimised):

$$\mathrm{MSE}(\mathbf{I}_C, \mathbf{I}_Q) = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{p}_i - \hat{\mathbf{p}}_i||_2^2, \tag{3.61}$$

where $\mathbf{I}_C$ and $\mathbf{I}_Q$ denote the original and quantised images, consisting of n pixels each ($\mathbf{p}_i \in \mathbf{I}_C$ and $\hat{\mathbf{p}}_i \in \mathbf{I}_Q$). Sometimes researchers also use peak signal-to-noise ratio (abbreviated as PSNR), which is calculated directly from MSE:

$$\mathrm{PSNR}(\mathbf{I}_C, \mathbf{I}_Q) = 20 \log_{10} \left(\frac{255}{\sqrt{\mathrm{MSE}(\mathbf{I}_C, \mathbf{I}_Q)}} \right). \tag{3.62}$$

Contrary to the MSE, the higher PSNR, the better. While the problem of quantisation is formulated to minimise MSE value, an intentional noise is often introduced. This process is called dithering and is used to reduce banding, which may seem unnatural for a human eye. As for the test data, there is a number of image processing and computer vision datasets available, depending on a specific task. Most of the scholars seems to use standard examples displaying a wide range of possible image features, such as *Lena*, *Peppers*, or *Pool* – for example these images were used by Celebi (2011) and Schaefer (2014). The first one – a picture of Lena Söderberg – is perhaps the most popular image in the image processing community. However, its usage is

discouraged nowadays due to the history and context of the original image ("On alternatives to Lenna", 2017). It is even banned in Nature Nanotechnology[16].

Three quantisation algorithms seem to be especially popular – k-means, median cut, and octrees. They are described in detail in the following subsections, along with other techniques (not strictly limited to the colour quantisation *per se* – some relevant palette design techniques are described as well). Some of them are not used nowadays, but they are worth mentioning for having a broader context. The other ones use state-of-the-art techniques, such as Generative Adversarial Networks.

3.3.1 Uniform Quantisation

Presumably the oldest technique in this section, uniform quantisation is a method (or family of methods) in which the goal is to obtain a universal colour palette, which is independent of the original distribution of colours. Usually, this term means truncating the least significant bits for each colour component (Xiang, 2007). However, following Xiang (2007), other scholars suggest using different forms of uniform colour quantisation. Trellis-coded quantisation, a method stemming from the domain of telecommunication, was successfully applied to this task (Ungerboeck, 1982; Marcellin & Fischer, 1990; An & Cai, 2008). As an example of a completely different approach, Mojsilovic and Soljanin (2001) presented a method of sampling by the Fibonacci lattice, which is inspired by the golden ratio. In this section, we focus on the simplest form, which relies on the aforementioned bit truncation.

Algorithm 1 Uniform quantisation, 3-3-2 variant for RGB.

1: **function** UniformQuantisation($\mathbf{p}$)
2: $R \leftarrow \mathbf{p}.R$ & `0x00FF0000` $\gg 16$
3: $G \leftarrow \mathbf{p}.G$ & `0x0000FF00` $\gg 8$
4: $B \leftarrow \mathbf{p}.B$ & `0x000000FF`
5: **return** $(R$ & `0xE0`$) \mid (G$ & `0xE0`$) \gg 3 \mid (B$ & `0xC0`$) \gg 6$

Based on: Burger and Burge (2009b)

Uniform quantisation does not necessarily mean the same number of bits cut off from each colour component. For example, if one would like to carry colour information on 8 bits instead of 24 needed for true colour RGB, it is obvious that colour components can't be of the same length. For example, we may encode the original colour information with 3 bits for the red component, 3 for green, and 2 for blue. Consider a following example with the RGB colour pixel

[16] `https://www.nature.com/articles/s41565-018-0337-2` [accessed on October, the 17th, 2019]

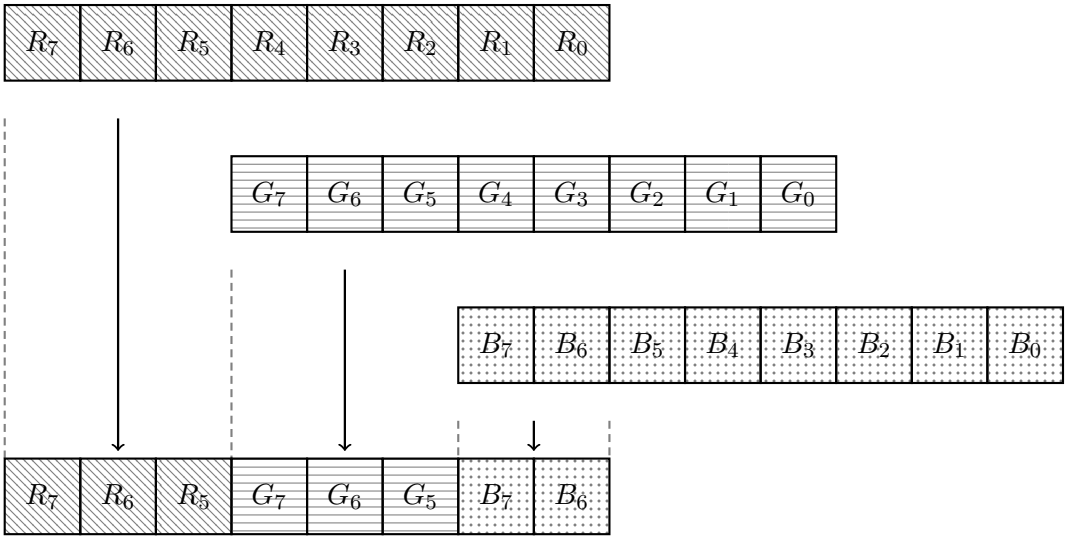


Figure 3.5: Uniform quantisation, 3-3-2 variant for RGB.
Based on: Burger and Burge (2009b)

$\mathbf{p} = [255, 127, 32]$. Treating the components of $\mathbf{p}$ as binary numbers, we will get 11111111_2, 01111111_2, and 00100000_2 accordingly. After the truncation of the least significant bits, we will obtain 111_2, 011_2, and 00_2. This can be encoded as 11101100_2 (which is 236 as a decimal number). The quantised value of $\mathbf{p}$ will then be $[11100000_2, 01100000_2, 00000000_2]$ ($[224, 96, 0]$ in decimal). Such information can be stored only in 8 bits instead of 24, which fulfils the quantisation goal. The procedure is shown in Algorithm 1 and Figure 3.5. Similarly, Xiang (2007) provides an interesting example for 15 bits. Instead of removing the three least significant bits from each colour component, he argues that the standard formula for calculating luminance[17] might constitute an inspiration for the division. Xiang argues that taking 5 bits for the red component, 6 for the blue, and 4 for the red one would capture the colour differences in a better way, since RGB is not a perceptually uniform colour space.

Uniform quantisation assumes a uniform distribution of colours in the considered image, which is rather rarely encountered in practice. This can result in a large quantisation error and (subjectively) ugly outcomes – especially with a large number of bits truncated. However, the simplicity of uniform quantisation brings something else to the table. Since uniform quantisation is image-independent, there are no memory requirements. Hence, its space complexity is of $\mathcal{O}(0)$ and makes it a good choice for memory-less devices. The algorithm is linear in k (the number of colours we want to obtain) in the search of representatives phase. Similarly, it is linear in the

[17] The luminance is derived as follows: $Y = 0.299R + 0.587G + 0.114B$.

number of pixels n in the mapping phase. The time complexity is of $\mathcal{O}(k)$ and $\mathcal{O}(n)$ respectively.

3.3.2 Popularity Method

The popularity (sometimes called populosity) method conceptually is a simple algorithm, in which we draw the k most popular colours from a given image colour histogram. In a way, it extends uniform quantisation. Following P. Heckbert (1982), it was simultaneously invented in two different institutions in 1978 by Tom Boyle & Andy Lippman from MIT's Architecture Machine Group and Ephraim Cohen from the New York Institute of Technology. The efforts of both groups of scholars do not appear to be published in any academic sources, though the procedure has been described by P. Heckbert (1980).

The core algorithm is simple and can be summarised as generating a histogram and picking k most popular colours. However, Xiang (2007) suggests that this approach does not perform well on true colour images, as well as in the case of relatively small k. A simple pre-processing trick can alleviate this problem – one need to generalise image pixels. Since a true colour image comes with 256^3 different values in the most extreme case (assuming RGB), exploring such a huge number of buckets to create a histogram is not really convenient. The number of colours can be drastically reduced using uniform quantisation – for example, the 3-3-2 variant from Algorithm 1 may be used. In such a reduced set of colours, the difference between representatives should be larger. The algorithm still omits small but distinctive regions, which can carry important details.

Algorithm 2 employs these modifications. The algorithm is fast – its overall time complexity is dominated by the sorting operation. P. Heckbert (1982) suggests the selection sort algorithm (of $\mathcal{O}(dk)$), though any sorting algorithm will do the trick. In practice, these numbers are not that big for modern computers. The space complexity is (of $\mathcal{O}(r)$), where r denotes the size of the list of reduced colours. The mapping phase is pretty straightforward – one needs to replace colours in the original image with the nearest one using the Euclidean distance.

Algorithm 2 Popularity method – palette generation.

```
1: function PopularityMethod(I_C = {p_1, p_2, ..., p_n}, k)
2:     reducedColours ← List()
3:     for each p in I_C do
4:         reducedColours.Append(UniformQuantisation(p_i))
5:     reducedColours ← reducedColours.Sort()
6:     return reducedColours.Take(k)
```

Based on: P. Heckbert (1980) and Xiang (2007)

3.3.3 *k*-means Clustering

Denoting perhaps one of the most recognised unsupervised machine learning algorithms, the term k-means dates back to 1967, when it has been coined up by James MacQueen (1967). Tracing its roots, however, is a bit more complicated since MacQueen referenced a slightly different algorithm from what everyone understands under that term. Following Bock (2007), the history of k-means considers several approaches to different problems and can be sketched as follows. The very first explicit formulation of k-means clustering problem in its continuous form is attributed to Steinhaus (1956). However, it was Lloyd (1982) who proposed a standard algorithm for the continuous version of k-means (often called Lloyd's algorithm). The algorithm was ready in 1957, though it was not published outside Bell Labs until 1982 (Bock, 2007). The earliest discrete version of k-means is attributed to Forgy (1965), though the first publication belongs to Jancey (1966).

The sum of squares criterion is the central idea behind k-means. It can be formulated continuously and discretely (Bock, 2007). The former version is given by the optimisation problem of exhaustive and mutually exclusive partitioning X to k clusters:

$$\min_{C} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C} ||\mathbf{x}_i - \mathbf{c}_j||_2^2, \tag{3.63}$$

where k denotes a number of desired clusters, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^d$ represents the input data to be partitioned within clusters, $c_k \in C$ is the centre of cluster S_k, and $||\cdot||_2$ stands for the Euclidean norm. Following Bock (2007), the problem can be formulated in continuous space:

$$\min_{\mathcal{B}} \sum_{i=1}^{k} \int_{B_i} ||x - \mathbb{E}[X|X \in B_i]||_2^2 dP(x), \tag{3.64}$$

where $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^d$, in which $\mathbf{x}_i$ stands for a realisation of a random vector with distribution P, and $\mathcal{B} = (B_1, B_2, \ldots, B_k) \in \mathbb{R}^d$. Actually, the discrete version from Equation 3.63 can be derived from the continuous version.

A typical k-means procedure is presented in Algorithm 3. The algorithm in its basic form takes two input variables: k, which denotes the desired number of output clusters (or colours in this context), and $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^d$, a set of n d-valued vectors $\mathbf{x}_i$. The first step of k-means considers the preparation of the initial set of cluster centres (INITIALISE-CLUSTERS(X)). As this step is much more important than it seems, we discuss it later in this section. After the initial assignment, the main loop runs until satisfying the termination criteria (TERMINATE?), with two loops in every iteration. The first one goes n times and assigns every data sample $\mathbf{x}_i$ to the nearest cluster – $m[i]$ stores the index of that cluster. The second one is run k times to recalculate cluster centres. S_j is a set of points $\mathbf{x}_i$ with the smallest distance to the centre $\mathbf{c}_j$. The new cluster centres $\mathbf{c}_j$ are given by the mean of the points in S_j. The termination criteria (tested by TERMINATE? in Algorithm 3) can be chosen among, for example, no progress in convergence, lack of variance improvement (considering a specific threshold), or simply by reaching the maximum number of iterations. After satisfying the termination criterion, the algorithm yields $C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\} \in \mathbb{R}^d$, which is a set of k d-dimensional cluster centres $\mathbf{c}_j$.

While the exact solution to the k-means problem is proven to be NP-hard starting from k=2 or d=2 (Aloise, Deshpande, Hansen, & Popat, 2009), the heuristic provided by Lloyd (1982) can converge quickly to a local optimum. With a fixed d, the algorithm is linear in n and k per iteration, i.e. its time complexity equals to $\mathcal{O}(nkt)$, where t is the number of iterations (Celebi, 2011). Sometimes, it is referred to as $\mathcal{O}(nktd)$, to include the impact of the number of dimensions d. When it comes to the space complexity of k-means, it is equal to $\mathcal{O}(n(d+k))$. While these values are considered reasonable for standard clustering problems, applying Lloyd's algorithm as a colour quantisation method is considered rather slow and space inefficient – especially compared to fast algorithms such as median cut or octrees. However, k-means can provide high-quality results, which makes this algorithm still a popular choice for less time-sensitive applications.

Algorithm 3 Lloyd's algorithm for k-means clustering.

1: **function** K-MEANS(k, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^d, \mathbf{C}$)
2: **if** $\neg\mathbf{C}$ **then** $\mathbf{C} \leftarrow$ INITIALISE-CLUSTERS(X)
3: **while** $\neg$ TERMINATE? **do**
4: **for** $(i \leftarrow 1; i \leq n; i \leftarrow i + 1)$ **do**
5: $m[i] = \underset{j \in \{1,2,\ldots,k\}}{\operatorname{argmin}} ||\mathbf{x}_i - \mathbf{c}_j||_2^2$
6: **for** $(j \leftarrow 1; j \leq k; j \leftarrow j + 1)$ **do**
7: $\mathbf{S}_j \leftarrow \{\mathbf{x}_i | m[i] = j\}$
8: $\mathbf{c}_j \leftarrow \frac{1}{|S_j|} \sum_{\mathbf{x}_i \in S_j} \mathbf{x}_i$
9: **return** $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\} \in \mathbb{R}^d$

Based on: Celebi (2011)

The algorithm is proven to be very sensitive to the chosen cluster initialisation method (INITILISE-CLUSTERS(X) in Algorithm 3) (Selim & Ismail, 1984). Celebi (2011) states that wrongly initialised clusters may result in empty clusters, slow convergence, and generally getting stuck in local minima. There exist a decent number of initialisation methods for k-means to choose from (Table 3.2). The simplest approach is Forgy's method, which assigns random centres to the clusters with $\mathcal{O}(k)$ complexity (Forgy, 1965). A similar approach is proposed by (MacQueen, 1967) – centres are also assigned randomly, though this time they are selected solely from the input data points. Perhaps the most popular approach so far is k-means++. While obtaining the initial clusters takes extra time of complexity of $\mathcal{O}(nkd)$, the k-means algorithm is expected to find a solution in $\Theta(\log k)$ using this initialisation method. To achieve this, k-means++ uses probability distributions in the calculations – the authors call this step D^2 *weighting*. Many k-means implementations use this initialisation method as a default one.

So far, we have discussed a standard approach to Lloyd's algorithm and k-means problem. Puzicha, Held, Ketterer, Buhmann, and Fellner (1998, 2000) proposed a different approach to the problem of colour quantisation. Contrary to the majority of methods, in which dithering methods often follow the process of quantisation, they combined them into a single algorithm. Their approach, called *spatial colour quantisation*, bases on a cost function, which performs quantisation and digital halftoning at the same time. It relies on a generalised k-means criterion. Two optimisation methods are presented: iterative conditional mode (which is similar to k-means) and digital annealing (which resembles simulated annealing). The authors report that the latter seems to avoid the problem of getting stuck on poor local minima. They also applied multiscale optimisation to reduce the search space. The algorithm was later implemented and

Table 3.2: Initialisation methods for k-means with their complexities.

Method	Time complexity	Reference
Random (Forgy's)	$\mathcal{O}(k)$	Forgy (1965)
Random (MacQueen's)	$\mathcal{O}(k)$	MacQueen (1967)
Bradley and Fayyad's	$\mathcal{O}(d \cdot \textsc{Iter}(d))$	Bradley and Fayyad (1998)
Splitting	$\mathcal{O}(nk)$	Linde, Buzo, and Gray (1980)
Minmax	$\mathcal{O}(nk)$	Hochbaum and Shmoys (1985), Gonzalez (1985), Katsavounidis, Kuo, and Zhang (1994)
Density-based	$\mathcal{O}(n)$	Al-Daoud and Roberts (1996)
Maximum variance	$\mathcal{O}(n \log n)$	Al-Daoud (2005)
Subset-farthest first	$\mathcal{O}(k^2 \ln k)$	Turnbull and Elkan (2005)
k-means++	$\mathcal{O}(nkd)$	Arthur and Vassilvitskii (2007)
Var-Part	$\mathcal{O}(nkd)$	Su and Dy (2007)
PCA-Part	$\mathcal{O}(nkd^2)$	Su and Dy (2007)

Abbreviations: k – number of clusters, n – number of pixels, d – number of dimensions (e.g. 3 for RGB), $\textsc{Iter}(d)$ – the numbers of iterations required by algorithm for clustering. Sources: Celebi (2011); Celebi, Kingravi, and Vela (2013)

published by D. Coetzee and called *scolorq*[18]. He argues that the algorithm works very well on a small number of representatives (between 4 and 8) but does not handle large numbers (such as 256) and continuous-tone images very well.

Bezdek (1981) proposed fuzzy c-means, a generalisation of k-means, in which a data point can belong to more than one cluster. The algorithm has also been adapted to the problem of colour quantisation, for example by Schaefer and Zhou (2009). However, Wen and Celebi (2011) argue that fuzzy c-means, while being much slower in terms of execution time, is not superior to k-means for the colour quantisation applications. The k-means algorithm can also be reformulated to include weights $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$. Required modifications are presented in Algorithm 4 – notice the changes in lines 5 and 8.

[18] `https://people.eecs.berkeley.edu/~dcoetzee/downloads/scolorq/` [accessed on October, the 17th, 2019]

Algorithm 4 Weighted k-means algorithm.

1: **function** WEIGHTEDK-MEANS(k, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^d, \mathbf{w} = \{w_1, w_2, \ldots, w_n\}, \mathbf{C}$)
2: **if** $\neg\mathbf{C}$ **then** $\mathbf{C} \leftarrow$ INITIALISE-CLUSTERS(X)
3: **while** $\neg$ TERMINATE? **do**
4: **for** $(i \leftarrow 1; i \leq n; i \leftarrow i + 1)$ **do**
5: $m[i] = \underset{j \in \{1,2,\ldots,k\}}{\operatorname{argmin}} \mathbf{w}_i ||\mathbf{x}_i - \mathbf{c}_j||_2^2$
6: **for** $(j \leftarrow 1; j \leq k; j \leftarrow j + 1)$ **do**
7: $\mathbf{S}_j \leftarrow \{\mathbf{x}_i | m[i] = j\}$
8: $\mathbf{c}_j \leftarrow \frac{1}{\sum_{\mathbf{x}_i \in S_j} \mathbf{w}_i} \sum_{\mathbf{x}_i \in S_j} \mathbf{w}_i \mathbf{x}_i$
9: **return** $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\} \in \mathbb{R}^d$

Based on: Celebi (2011) and `http://www.cs.fsu.edu/~ackerman/CIS5930/notes/Weighted%20clustering.pdf` [accessed on October, the 17th, 2019]

Celebi (2011) proposed the weighted sort-means algorithm, which is a faster version of k-means tailored to the problem of colour quantisation. The time complexity of this algorithm is of $\mathcal{O}(k^2 + k^2 \log k + d\gamma)$, where k is the desired number of representatives, d stands for the number of distinct colours in the original image, and γ will be explained in a moment. While quadratic formulas may seem to make it slower compared to the original k-means, the complexity is dominated by the last parameter $d\gamma$, since $k \ll d$. The speed-up was possible to achieve by the usage subsampling, sample weighting, and the sort-means (Phillips, 2002) algorithm. The variable γ comes from sort-means and, following Phillips (2002), represents *the average of all points* $\mathbf{p}$ *of the number of means that are no more than twice as far as* $\mathbf{p}$ *is from the mean* $\mathbf{p}$ *was assigned to in the previous iteration.* Celebi claims that the weighted sort-means method gives the same results as the original k-means.

Chang et al. (2015) tackles the problem of photo recolouring, in which the first step considers palette extraction using the weighted variation of k-means. While the problem of colour transfer is beyond the scope of this work, they approached the extraction in a way that not only yields a small quantisation error but also will diversify colours in the resulting palette, which makes it particularly interesting for the objectives of this dissertation. Their approach was later used in Google Art Palette[19], which allows to upload an image, extract its palette and find artworks with a similar palette. The source code is available on GitHub[20]. This approach bases on a modified version of weighted k-means (Algorithm 5). The first difference lies in the pre-processing step – instead of the original image, its histogram $\mathbf{L}$ of size s^3 constitutes the input for weighted

[19] `https://experiments.withgoogle.com/art-palette` [accessed on October, the 17th, 2019]
[20] `https://github.com/googleartsculture/art-palette` [accessed on October, the 17th, 2019]

k-means. For each bin in $\mathbf{L}$, its mean colour is computed in the CIE L*a*b* space. In the original paper $s = 16$ is used, hence the histogram $\mathbf{L}$ contains 4096 bins. Weights $\mathbf{w}$ corresponds to the number of pixels associated with each bin. The exact procedure for obtaining $\mathbf{L}$ and $\mathbf{w}$ is presented in Algorithm 6.

The authors of this method have been aware of the sensitivity of k-means on the cluster initialisation method and proposed their own one, which is presented in Algorithm 7. The method was formulated with two goals in mind – to eliminate randomness and increase the difference between the initial cluster centres (i.e. make the representative colours more different from each other at the beginning). The first cluster centre represents the bin with the highest weight. After that assignment, the rest of weights is enhanced by a factor of $1 - \exp\left(||\mathbf{c}_s, \mathbf{c}_t||_2^2/\sigma^2\right)$ in order to penalise close bins. The parameter σ represents a falloff – in the original paper, $\sigma = 80$, whereas in the Google Arts implementation, it is close to 60. The process is repeated until k centres are found. The authors of this method report a significant speed-up over k-means, as well as brighter colours in the resulting palette.

Algorithm 5 Chang's et al. palette extraction algorithm.

function ChangsPaletteExtraction(s, k, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^d$)
    $\mathbf{L}, \mathbf{w} \leftarrow$ CreateHistogramAndWeights(s, k, $\mathbf{X}$)
    $\mathbf{C} \leftarrow$ WeigthedK-Means(k, $\mathbf{L}$, $\mathbf{w}$) ▷ with Algorithm 7 for Initialise-Clusters
    return $\mathbf{C}$

Based on: Chang et al. (2015) and `https://github.com/googleartsculture/art-palette` [accessed on October, the 17th, 2019]

Algorithm 6 Chang's et al. palette extraction algorithm – histogram creation.

function CreateHistogramAndWeights(s, k, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^d$)
    $\mathbf{w} \leftarrow \underbrace{[0, 0, \ldots, 0]}_{s^3}$
    $\mathbf{L} \leftarrow [\,]$
    for $\mathbf{x}$ in $\mathbf{X}$ **do**
        $r \leftarrow \mathbf{x}$.R, $g \leftarrow \mathbf{x}$.G, $b \leftarrow \mathbf{x}$.B
        $\mathbf{x}_{\text{Lab}} \leftarrow$ RgbToLab($\mathbf{x}$)
        $i \leftarrow s\,(\lfloor r/s \rfloor + \lfloor g/s \rfloor) + \lfloor b/s \rfloor$
        if $\neg(i$ in $\mathbf{L})$ **then**
            $\mathbf{L}[i] \leftarrow \mathbf{x}_{\text{Lab}}$
        else
            $\mathbf{L}[i] \leftarrow \mathbf{L}[i] + \mathbf{x}_{\text{Lab}}$
        $\mathbf{w}[i] + +$
    return $\mathbf{L}, \mathbf{w}$

Based on: Chang et al. (2015) and `https://github.com/googleartsculture/art-palette`

Algorithm 7 Chang's et al. palette extraction algorithm – cluster centres initialisation.

1: **function** InitialiseClusters(s, σ, $\mathbf{L} = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n\} \in \mathbb{R}^d$)
2: $\quad\mathbf{w}_{\text{mut}} = \text{Clone}(\mathbf{w})$
3: $\quad\mathbf{S} = [\,]$
4: $\quad i_{\max} \leftarrow 0$
5: $\quad$**for** $(i \leftarrow 1; i \leq k; i \leftarrow i + 1)$ **do**
6: $\quad\quad i_{\max} \leftarrow \text{argmax}_i\, \mathbf{w}_{\text{mut}}[i]$
7: $\quad\quad$**if** $\mathbf{w}_{\text{mut}}[i] = 0$ **then break**
8: $\quad\quad \mathbf{c}_s \leftarrow \frac{1}{\mathbf{w}[i_{\max}]}\mathbf{L}[i_{\max}]$
9: $\quad\quad \mathbf{w}_{\text{mut}}[i_{\max}] \leftarrow 0$
10: $\quad\quad$**for** $(j \leftarrow 1; j \leq s^3; j \leftarrow j + 1)$ **do**
11: $\quad\quad\quad$**if** $\mathbf{w}_j > 0$ **then**
12: $\quad\quad\quad\quad \mathbf{c}_t \leftarrow \frac{1}{\mathbf{w}[j_{\max}]}\mathbf{L}[j_{\max}]$
13: $\quad\quad\quad\quad \mathbf{w}[j] \leftarrow \mathbf{w}[j]\left(1 - \exp\left(||\mathbf{c}_s, \mathbf{c}_t||_2^2/\sigma^2\right)\right)$
$\quad\quad\quad \mathbf{S}.\text{Append}(\mathbf{w}[i_{\max}])$
14: $\quad$**return S**

Based on: Chang et al. (2015) and `https://github.com/googleartsculture/art-palette` [accessed on October, the 17th, 2019]

3.3.4 Median Cut

The median cut algorithm had its first appearance in print in the early 1980s (P. Heckbert, 1980, 1982), though it was invented by the author of these publications in 1979. The algorithm showed a significant improvement compared to uniform quantisation and the popularity method, which were widely used back then. The original papers, however, only briefly describe the ideas in the algorithm and lack strict formulation. Since the exact steps are unclear, there is a number of variations of the median cut algorithm (Kruger, 1994; Bloomberg, 2008). Thomas G. Lane in his implementation in JPEG library[21] argues that various median cut quantisers vary in a method for choosing the largest box and its median point. Bloomberg (2008) also points out dissimilarities among different implementations, namely the methods for queueing boxes for the division, deciding the axis to split on, and deciding in which box a median pixel belongs. Eventually, as an answer to these issues, he proposed his Modified Median Cut Algorithm (abbreviated as MMCQ), which is a part of his image processing and analysis library called Leptonica[22].

To demonstrate median cut, we rely on a straightforward implementation[23], which is presented in Algorithm 8 in Python-like pseudocode. Suppose that our picture we want to quantise an image containing n pixels. These pixels can be represented by their colours in a three-

[21] `http://libjpeg.sourceforge.net` [accessed on October, the 17th, 2019]
[22] `http://www.leptonica.org` [accessed on October, the 17th, 2019]
[23] `https://github.com/mvanveen/mcut` [accessed on October, the 17th, 2019]

dimensional RGB space. As for the result of the quantisation, we want to obtain k colours. At first, one needs to search for extrema for all possible RGB values. After that, a rectangular box, which edges intersect these points, is created. The longest edge indicates the dimension to split on. Conducting such action (see algorithm 9) results in two boxes, preferably with the same number of points. To calculate the exact split point, all the values of the given dimension need to be sorted. The median constitutes a split point through which the dividing plane is conducted (it is orthogonal to the remaining axes) – hence the algorithm name. The boxes are then recursively divided until the desired number of boxes k is reached. After that, for each box in the resulting set, its RGB values are averaged in order to obtain representatives for the reduced palette.

Algorithm 8 Median cut algorithm.

```
1:  function MedianCut(pixels, k)
2:      colours ← GetColours(pixels)
3:      boxes ← [Box(colours)]
4:      while boxes.size < k do
5:          globalMaxSize ← 0
6:          for index, box in enumerate(boxes) do
7:              size ← box.size
8:              maxSize ← max(size)
9:              maxDim ← size.index(max_size)
10:             if maxSize > globalMaxSize then
11:                 globalMaxSize ← maxSize
12:                 maxBox ← index
13:         splitBox ← boxes[maxBox]
14:         box_a, box_b ← splitBox.split(maxDim)
15:         boxes ← boxes[: maxBox] + [box_a, box_b] + boxes[maxBox + 1 :]
16:     return [c.average() for c in boxes]
```

Based on: `https://github.com/mvanveen/mcut` [accessed on October, the 17th, 2019]

Algorithm 9 Median cut algorithm – box's split method.

```
1:  function Split(axis)
2:      self.shrink()
3:      self.colors = sortedByAxis(self.colors, axis)
4:      medianIdx = int((self.colors).size/2)
5:      return Box(self.colors[: medianIdx])), Box(self.colors[medianIdx :])
```

Based on: `https://github.com/mvanveen/mcut` [accessed on October, the 17th, 2019]

One can say that the median cut algorithm shows some resemblance to the k-d tree algorithm (Bentley, 1975). Figure 3.6 illustrates this similarity by presenting the whole process using a toy

example. Suppose we want to quantise a 3×4 image P to 4 colours with the following set of pixels represented by its RGB values:

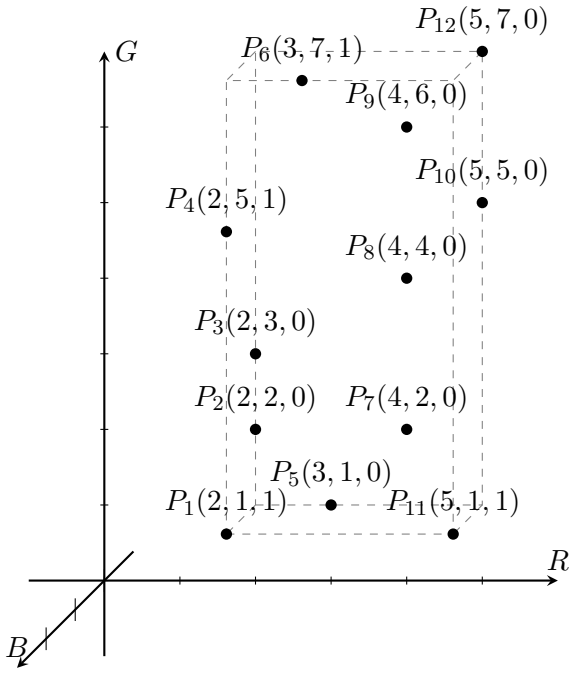
$$\begin{aligned} P = \{&(2,1,1),(2,2,0),(2,3,0),\\ &(2,5,1),(3,1,0),(3,7,1),\\ &(4,2,0),(4,4,0),(4,6,0),\\ &(5,5,0),(5,1,1),(5,7,0)\} \end{aligned}$$

At first, we set a rectangular box, which tightly encapsulates all the points in RGB space (Figure 3.6a). Since the green forms the longest axis, we cut along this one. The goal is to have the same or, if not possible, almost the same number of points in two new sets, so we form a dividing plane between P_3 and P_8. Notice that the two resulting boxes (Figure 3.6b) shrink to tightly encapsulate their areas, just like the first one. In the next step, we split the lower box along the red axis, which is the longest one in this case. The median point is between $P2$ and $P5$. This results in the two new boxes (Figure 3.6c). In this case, the leftmost one is flat, since only two dimensions are needed to entangle its points. For the fourth step, the upper box can be split along the red or green axis, since both of them are the longest ones. Figure 3.6d shows the result after using the red axis. At this point, we have a required number of boxes, which represent clusters of colours. For each of these clusters, we average its RGB values to obtain representatives for the quantised palette (R_1-R_4 in Figure 3.7).

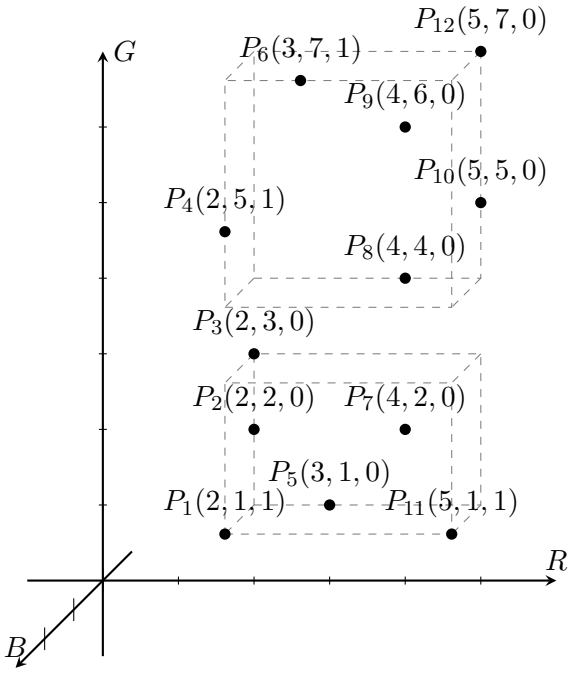
Gervautz and Purgathofer (1988) report that the search for representatives in this algorithm is quasilinear ($\mathcal{O}(n \log_2 k)$), where n is the number of pixels and k denotes the number of representatives. The time complexity of the mapping step is identical. The space complexity of median cut is linear in the number of unique colours d – $\mathcal{O}(d)$.

3.3.5 Octrees

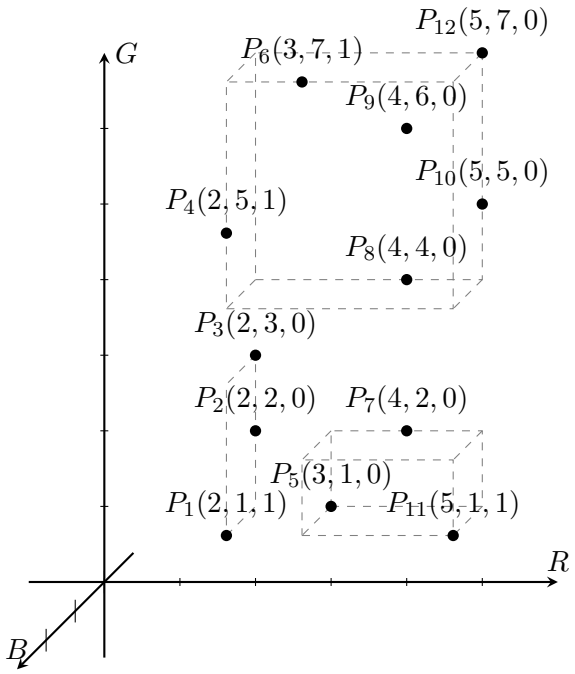
In the domain of spatial partitioning, a quadtree is a popular data structure, which provides a means for recursively dividing a two-dimensional space into four parts called quadrants (Finkel & Bentley, 1974). Meagher (1982) extended this approach to the third dimension by inventing octrees – as the name suggests, space is divided into the 8 equals parts (this time called octants). Formally, an octree is a tree, in which non-leaf nodes have exactly eight children (Figure 3.8). It took several years to transfer this idea to the problem of colour quantization, which is attributed



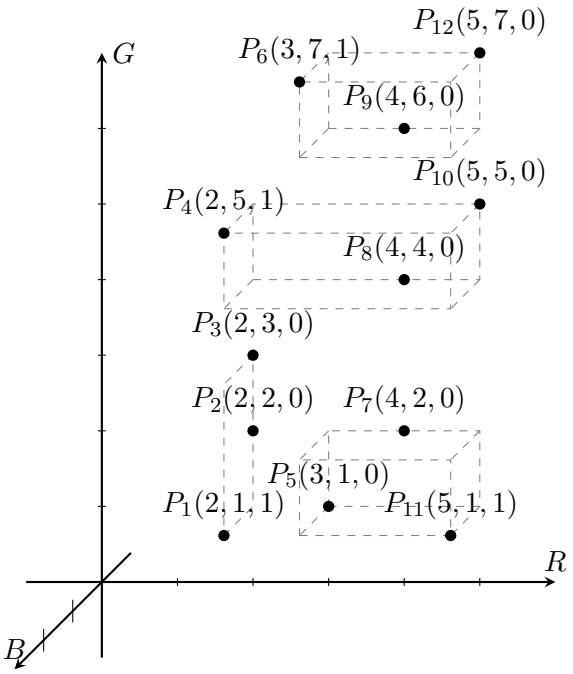
(a) First iteration.



(b) Second iteration.



(c) Third iteration.



(d) Fourth iteration.

Figure 3.6: Median cut quantisation with $k = 4$.

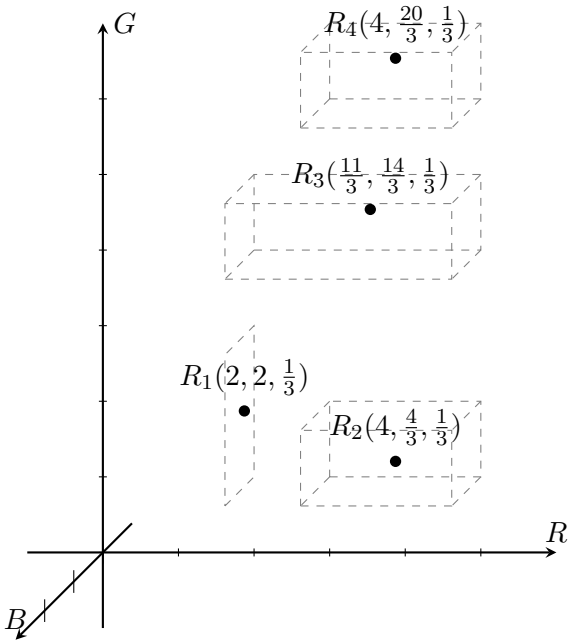


Figure 3.7: Median cut quantisation with $k = 4$ (Figure 3.6 continued) – choosing representatives.

to Gervautz and Purgathofer (1988).

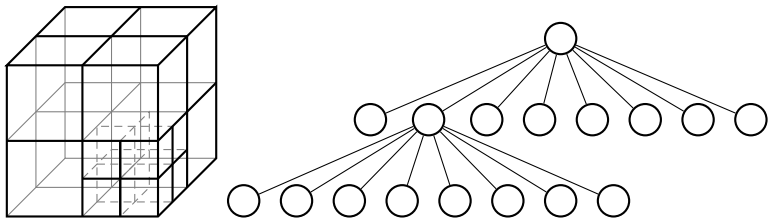


Figure 3.8: Illustration of the division scheme and the corresponding octree. The three-dimensional cube on the left is recursively divided into 8 equal parts, which constitutes the graph representation on the right-hand side.

Algorithm 10 Octree quantisation – the main loop.

```
1: function BuildOctree(pixels, k)
2:     octree ← BuildEmptyOctree()
3:     for p in pixels do
4:         octree.Insert(p)
5:     while octree.size > k do octree.Reduce( )
6:     return octree
```

Based on: Gervautz and Purgathofer (1988) and `https://github.com/delimitry/octree_color_quantizer` [accessed on October, the 17th, 2019]

The algorithm relies on building the memory-efficient representation of colours in a given image and consists of three phases: an octree construction, its reduction, and palette building. Consider an image with n pixels with d distinct colours, which are to be reduced to k colours in the quantisation process ($n \leq d \leq k$). At first, an empty tree is present. Pixels are processed one by one and nodes representing their colour are added to the tree. A tree node contains the colour occurrence counter, so if a given colour representation already exists, its value is incremented. An RGB pixel will be represented by a leaf node at level 8. The higher-level branches of the octree represent clusters of colours, which actually averages these colours. The node placement within the octree is calculated using the function GetColourIndex() in Algorithm 11. The parameter p represents a single pixel with its corresponding RGB values, whereas *level* denotes the considered level of the octree (which is 8 at most). Notice that the algorithm uses bitwise operators for so-called bit masking.

Algorithm 11 Octree quantisation – colour indexing.

1: **function** GetColourIndex($p \in [0, 256]^3$, *level*)
2: $i \leftarrow 0$
3: $m \leftarrow 2^7 \gg level$ ▷ mask depends on *level*
4: **if** $p.R \,\&\, m$ **then** $i \leftarrow i \mid 2^2$
5: **if** $p.G \,\&\, m$ **then** $i \leftarrow i \mid 2^1$
6: **if** $p.B \,\&\, m$ **then** $i \leftarrow i \mid 2^0$
7: **return** i

Based on: Gervautz and Purgathofer (1988) and `https://github.com/delimitry/octree_color_quantizer` [accessed on October, the 17th, 2019]

The intuition behind the GetColourIndex() method can be explained as follows. Suppose we have a single pixel p with its RGB values 121, 112, and 131. In a vector notation, that would be written as:

$$\begin{bmatrix} 121 & 112 & 131 \end{bmatrix}^\intercal .$$

Now if we rewrite each number as its binary representation and treat every digit as a separate column, we will get the following matrix:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} .$$

The resulting matrix contains the indices in its column. For example, the last column $\begin{bmatrix}1 & 0 & 1\end{bmatrix}^\intercal$ after transposition and treating columns as a single number will yield 5. Applying this method for the whole matrix, we will get:

$$\begin{bmatrix}1 & 6 & 6 & 6 & 4 & 0 & 1 & 5\end{bmatrix},$$

which is the colour index. The leftmost value (1) denotes the number of the first branch to expand (from the root), then there is the 6th branch on the second level etc. On the last level, the leaf nodes have counters, initially set to 0. The value of the leaf corresponding to the generated index is incremented. Figure 3.9 visualises this process. It is repeated for encoding all the pixels.

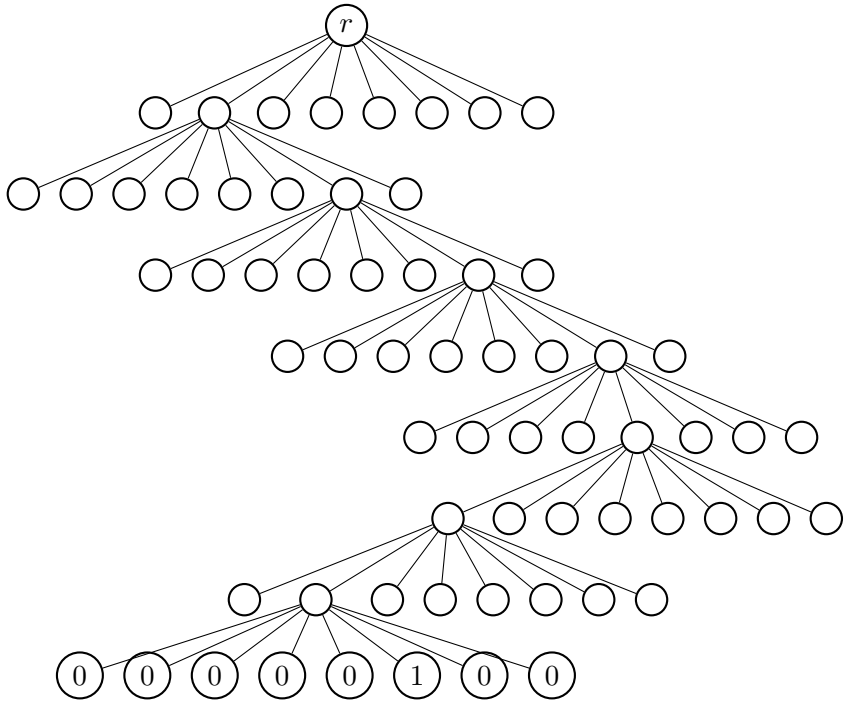


Figure 3.9: Appending $p = [121, 112, 131]$ to an empty octree. Nodes $[1, 6, 6, 6, 4, 0, 1, 5]$ are expanded, and the value at leaf is incremented.

Pixels p are successively inserted (INSERT(p) in Algorithm 12) into the octree – the algorithm requires a full single pass over the whole image. If the number of distinct colours d (which actually is the number of leaf nodes) exceeds the number of desired colours k, a reduction needs to be performed (REDUCTION in Algorithm 12). To proceed with such an operation, a set of reducible nodes needs to be selected. Gervautz and Purgathofer (1988) suggest the following method: if there's only one node at the deepest level, this node is going to be reduced. Otherwise, there's a need to choose among several candidates at the deepest level. The authors of the algorithm

proposed two methods for achieving that. The first one looks up for the node with the smallest number of pixels, which is an intuitive way of discarding the least popular colours. This method should result in a small sum of errors. The second method assumes removal of the most popular nodes, which will result in a slightly different palette and larger error, but it will save shadings. Other researchers proposed different methods. For example, there's a popular approach[24] in which a node at the nest to the last level is chosen, and all its children are removed at once. Regardless of the method, the colours and pixel counts from the removed nodes are added to the parent node and then averaged. The process is repeated until the threshold of k colours is reached.

As for the time and space complexity of the octree algorithm, Gervautz and Purgathofer (1988) report the following data. The search for representatives, as well as the mapping step, is linear in the number of pixels n – $\mathcal{O}(n)$. Octrees are very efficient in terms of memory as well, since their space complexity is linear in the number of pixels k – $\mathcal{O}(k)$.

Algorithm 12 Octree quantisation – insertion and reduction.

```
1:  function INSERT(p)
2:      index ←GETCOLOURINDEX(p, 7)
3:      if octree.HASNODE(index) then
4:          octree.APPEND(index)
5:          octree.GETNODE(index).counter ← 1
6:      else
7:          octree.GETNODE(index).counter ← octree.GETNODE(index).counter + 1
8:      octree.size ← octree.size + 1
9:
10: function REDUCE
11:     deepestLevel ← octree.GETDEEPESTLEVEL
12:     nodes ← octree.GETREDUCIBLENODESONLEVEL(deepestLevel)
13:     if nodes.SIZE == 1 then
14:         octree.REMOVENODE(nodes[0])
15:     else
16:         for node in nodes do
17:             node.GETPARENT().ADDCOLOURS(node.colours)
18:             octree.REMOVENODE(node)
19:         octree.size ← octree.size − 1
```

Based on: Gervautz and Purgathofer (1988) and `https://github.com/delimitry/octree_color_quantizer` [accessed on October, the 17th, 2019]

[24] See `https://observablehq.com/@tmcw/octree-color-quantization` [accessed on October, the 17th, 2019]or `https://github.com/delimitry/octree_color_quantizer` [accessed on October, the 17th, 2019].

3.3.6 Neural Networks

Neural Networks, together with the famous backpropagation algorithm, constitute a well-known approach to solving numerous problems in applied artificial intelligence by simulating the way a human brain works (Russell & Norvig, 2016). In contrast to standard neural networks, self-organising maps (abbreviated as SOM, sometimes called Kohonen's networks) employ unsupervised competitive learning instead of error correction (such as a backpropagation algorithm paired with gradient descent). The idea of self-organising maps was created by Kohonen (1990). Formally, SOM defines a mapping $\mathbb{R}^n \mapsto \mathbb{R}^m$, where $n \leq m$ and the mapping is continuous on almost all of its domain (Dekker, 1994). Contrary to classical neural networks, self-organising maps preserve the spatial structure of a given input – in a traditional network it is flattened. Another difference lies in the learning paradigm. Instead of backpropagation, Kohonen's networks use competitive learning (namely, the winner-takes-all strategy). They rely on multidimensional scaling (abbreviated as MDS) for making a fixed-size representation of an unknown dimensionality of an input vector.

Kohonen's SOM has been successfully applied to the problem of colour quantisation by Dekker (1994) in his *NeuQuant* method. The sampling factor is a key parameter since it controls the speed and quality of the results. However, high-quality results demand more time. The algorithm is reported to perform well for a large number of representatives ($n = 64$ and more), though its results for small palettes (e.g. $n = 8$) are modest. In the original paper (Dekker, 1994), the network consists of a 1-D array of 256 neurones, each containing a weight vector $[R_i, G_i, B_i]$. The network is a mapping $[0, 255]^3 \mapsto [0, 255]$, where the left-hand side represents an index of a given representative.

Deep Learning (LeCun, Bengio, & Hinton, 2015) and Convolutional Neural Networks (Krizhevsky, Sutskever, & Hinton, 2012) are perhaps one of the most important techniques in recent developments of artificial intelligence, computer science, or maybe even in the whole STEM in the 2010s. The progress in domains such as autonomous vehicles, voice recognition, or medical image processing gained momentum. While the concept of the neural network had been known for decades, the progress was possible because of the aforementioned techniques and the more powerful GPU cards, which provide a constantly increasing number of FLOPS available for AI researchers. Initially, Deep Learning became popular thanks to its application to the *discriminative* models, which handle tasks such as image classification. This approach is contrary to the *generative* group, which is – as the name suggests – all about generating new content. There are

a plethora of different approaches to neural networks, but one of them seems to be particularly clever and interesting.

Dubbed by Facebook's chief AI scientist Yann LeCun as *the coolest idea in machine learning in the last twenty years*, Generative Adversarial Networks – or simply GANs – are widely known for generating state-of-the-art results in numerous domains (Goodfellow et al., 2014; Goodfellow, Bengio, & Courville, 2016). For example, the recent StyleGAN (Karras, Laine, & Aila, 2019), which inferred ideas from style transfer networks (Huang & Belongie, 2017), provide impressive results on the task of mixing facial features from two different images. This results in a new image of a non-existing person, inheriting facial features from both of the input pictures. This architecture was later used in `thispersondoesnotexist.com`[25], a web service for generating pictures of people who do not exist in reality, which later became an Internet viral in 2019.

Following Goodfellow et al. (2016), the main idea of GANs relies on the scenario in which two different neural networks compete with each other. In the context of game theory, this can be perceived as a zero-sum game and resembles the famous minimax algorithm. These two aforementioned networks are named *generator* $\mathcal{G}$ and *discriminator* $\mathcal{D}$ – the generator prepares fake data, whereas discriminator's job is to tell whether a received sample is real ($\mathbf{x}_r$) or not ($\mathbf{x}_f$). In a computer vision setting, the generator $\mathcal{G}\colon Z \rightarrow X$ network creates fake images $\mathbf{x}_f \in X$ from some random noise vector $\mathbf{z} \in Z$ – its objective is to make them as *realistic* as possible. The discriminator receives an image $\mathbf{x}$ coming either from the real or fake distribution and tries to get better at guessing whether it's a real one over time (in other words, it is a simple binary classification task). However, GANs are notoriously hard to train due their lack of convergence (Goodfellow et al., 2014) – minimising loss for $\mathcal{G}$ and $\mathcal{D}$ does not guarantee to reach an equilibrium. To tackle this problem, a concept which is not a zero-sum game and uses logarithmic probability was formalised by Goodfellow et al. (2014) in the following equation (see also Figure 3.10):

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{\mathbf{x}} \left[\log(\mathcal{D}(\mathbf{x}))\right] + \mathbb{E}_{\mathbf{z}} \left[\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z})))\right] \tag{3.65}$$

Isola, Zhu, Zhou, and Efros (2017) described the concept of Conditional Adversarial Networks (abbreviated as cGANs). The main difference between GAN and cGAN is the generator $\mathcal{G}\colon \{Z, X\} \rightarrow X$, which now is aware of the current real sample $\mathbf{x}_r \in X$. We can now reformulate

[25] `https://thispersondoesnotexist.com`

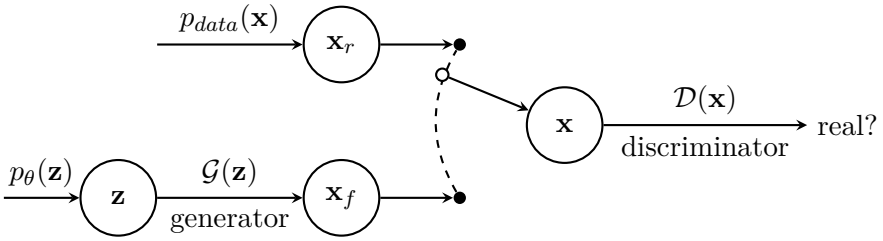


Figure 3.10: Generative Adversarial Network – the general architecture.
Based on: `https://github.com/PetarV-/TikZ/tree/master/Generative%20adversarial%20network` [accessed on October, the 17th, 2019]

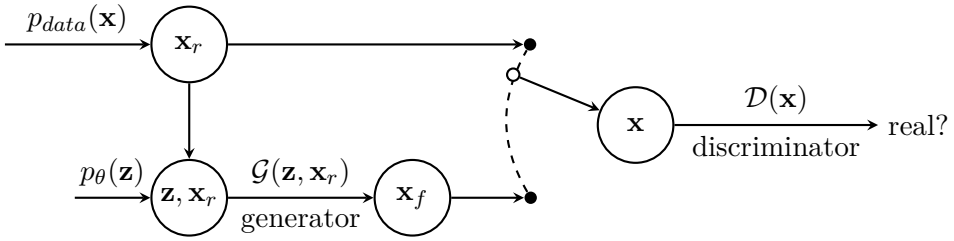


Figure 3.11: Conditional Generative Adversarial Network – the general architecture
Based on: `https://github.com/PetarV-/TikZ/tree/master/Generative%20adversarial%20network` [accessed on October, the 17th, 2019]

Equation 3.65 to the following one:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{\mathbf{x}_r, \mathbf{x}_f} \left[\log(\mathcal{D}(\mathbf{x}_r, \mathbf{x}_f)) \right] + \mathbb{E}_{\mathbf{x}_r, \mathbf{z}} \left[\log(1 - \mathcal{D}(\mathbf{x}_r, \mathcal{G}(\mathbf{x}_r, \mathbf{z}))) \right] \tag{3.66}$$

The whole process is depicted in Figure 3.11. A generative model needs to be used for $\mathcal{G}$ – for example, Encoder-Decoder (Cho et al., 2014), used in e.g. the domain of machine translation (as an example of a sequence-to-sequence task). This cGAN performs image-to-image tasks, such as turning grayscale images into colourful ones or generating photorealistic images from user-provided sketches. Its $\mathcal{G}$ is built using the U-Net architecture, which extends the idea in Encoder-Decoder by using so-called skip connections (Ronneberger, Fischer, & Brox, 2015).

How is this related to the problem of colour quantisation, though? The authors of the aforementioned cGAN paper published their source code, which Jack Qiao later adapted to create Colormind[26], an online application that can extract colour palettes from photos. Whereas there is no scientific publication as an outcome of this project, a part of the resulting source code is available on GitHub[27]. The approach used in this project, dubbed by the author as a Generative-

[26] `http://colormind.io`
[27] `https://github.com/Jack000/pix2pix`

MMCQ, provides a means for palette generation. As the author focuses on palette generation instead of colour quantisation, Generative-MMCQ aims at generating a visually attractive and contrasting set of colours, which together can be re-used for various design-related purposes.

The implementation details are only partially described by the author[28], but in general, this solution combines, as the name suggests, the features of MMCQ and cGAN. At first, the cGAN (with minor changes) presented by Isola et al. (2017) is trained using the data from Adobe Color, as well as with a selection of palettes from Dribble. The main palette generation algorithm firstly uses MMCQ results as ground truth data ($\mathbf{x}_r$). The generator $\mathcal{G}$ receives this image as well and generates a number of palettes using the random value $\mathbf{z}$. The final classifier evaluates the resulting palettes and selects the best one (the final result is pseudorandom due to $\mathbf{z}$). It is worth noticing that this classifier is not the discriminator $\mathcal{D}$, which turned out to be too good in selecting MMCQ palettes, as the author reports. It was trained by a hand-picked selection of images and colour palettes, which were chosen by the author.

While not really connected to the problem of colour quantisation, due to the topic of this dissertation, there was one more GAN application which perhaps is worth mentioning. Pierre Fautrel, Hugo Caselles-Dupré, and Gauthier Vernier formed a Paris-based artistic collective *Obvious*. In 2018, *their* work *Edmond de Belamy, from La Famille de Belamy* (Figure 3.12) was sold at Christie's for $432,500 (premium price), which exceeded the initial estimates ($7,000-$10,000) by an order of magnitude[29]. Instead of a typical signature, the artwork contains GAN's objective function (Equation 3.65). They were not the first artists to use GANs (Cohn, 2018), but they received extensive media coverage due to this specific auction. François Chollet, the creator of a popular neural network library called Keras, even suggested a term *GANism* name this artistic movement.

3.3.7 Other Algorithms and Relevant Issues

This section covered the most popular quantisation algorithms. There exist a number of other interesting approaches, which were not discussed in this work. For example, Ozturk et al. (2014) use the artificial bee colony algorithm. Schaefer, Agarwal, and Celebi (2018) presented an algorithm for colour reduction which is based on the grey wolf optimisation. Other scholars also tackled somewhat related problems. For example, Glasbey, van der Heijden, Toh, and Gray

[28] `http://colormind.io/blog/extracting-colors-from-photos-and-video/`

[29] `https://www.christies.com/lotfinder/Lot/edmond-de-belamy-from-la-famille-6166184-details.aspx` [accessed on October, the 17th, 2019]

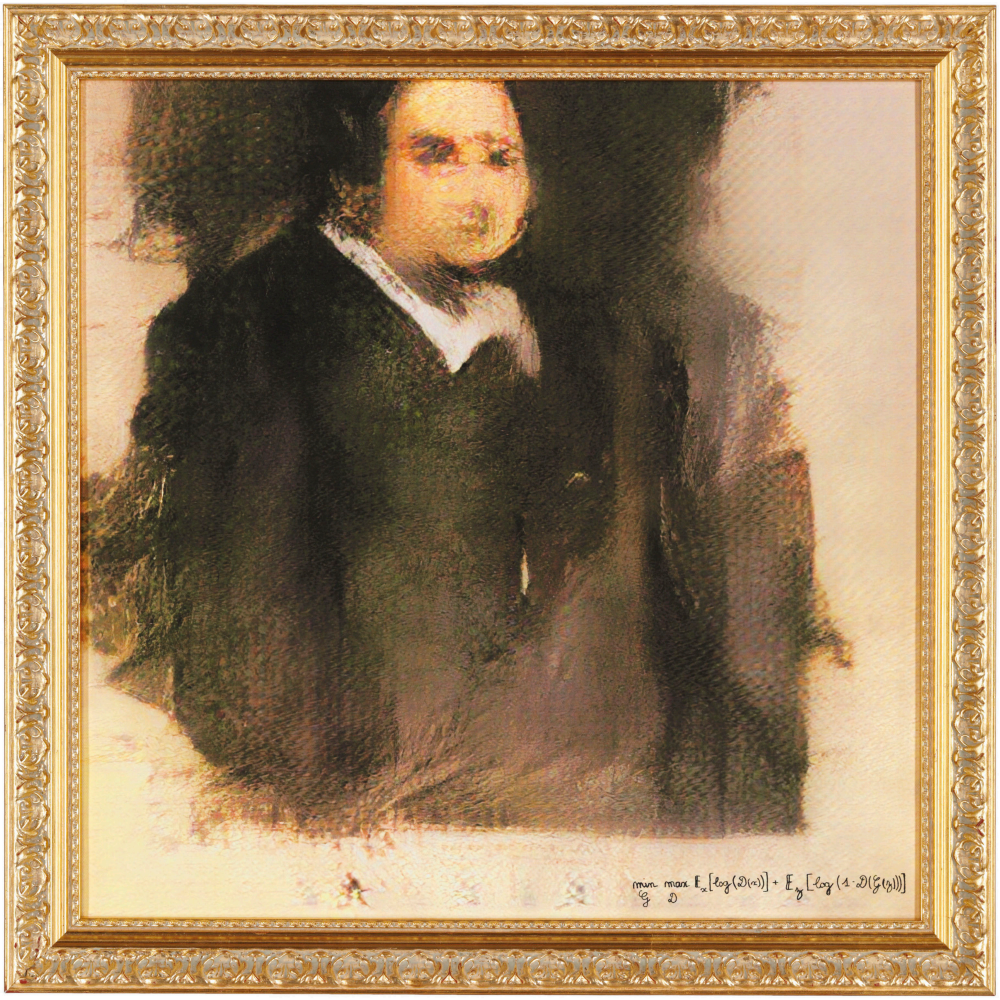


Figure 3.12: Obvious, *Edmond de Belamy, from La Famille de Belamy*, Generative Adversarial Network print on canvas, 2018.

Source: Christie's

(2007) investigated the topic of finding the maximally distinct set of colours. Formally, the problem is to iteratively find a set of colours c_i such that:

$$c_{n+1} = \arg\max_{c} \left(\min_{i=1,\ldots,n} D\left(c_i, c\right) \right), \tag{3.67}$$

where D is the distance function and c denotes any colour. They presented two approaches – one which is sequential, and the second, more sophisticated, based on simulated annealing.

It is worth mentioning that two important topics related to quantisation were not discussed in this section. While we covered the topic of a search of representatives, dithering and the mapping phase were omitted. Dithering is a process of intentional increasing the quantisation error to minimise the grain effect on quantised images and make them look more *smooth*, which is more visually appealing for a human eye. Floyd-Steinberg dithering (Floyd & Steinberg, 1975) is perhaps one of the most widely known examples of such algorithms. They are omitted in this section since this work is concerned with an accurate generalisation of colour palette, which makes such methods irrelevant.

On the other hand, we did not discuss all the possible algorithms for the mapping phase, assuming using the Euclidean distance where it was possible – the potential speedup in this phase was not the goal of this work. The mapping phase lies in a function that takes a single point and returns its representative from the set of all representatives. The most straightforward method is just finding the closest colour using the Euclidean distance and build an inverse colour map. The time complexity of such solution is of $\mathcal{O}(kn)$. There exist a number of faster and more sophisticated methods to achieve that, such as k-d trees, locally sorted search, and Voronoï diagrams. This topic was examined by Brun and Trémeau (2003).

3.4 Colours in Quantitative Art Market Research

This section explores research related to the topic of colours and the price of artworks. A few scholars sought to understand the phenomena of the impact of colour-related features on the price of artworks – their effort is described here. Some notions relating to econometrics are used in this section (such as hedonic models). They are described in detail in Chapter 2 – especially in Section 2.2).

Perhaps Stepanova (2015) delivered one of the earliest works considering the impact of colours on price. She analysed two datasets – one consisting only of Picasso's paintings, whereas the

other contained colour field abstract expressionists works (Mark Rothko and Hans Hoffman, both analysed separately). She used 259 paintings made by Picasso and sold between 1998 to 2014, 128 made by Hans Hoffman sold between 2001 and 2015, and 98 of Mark Rothko (1998-2015). All the data comes from Christie's and Sotheby's auction houses (both NY). The standard log-linear time dummy hedonic price model was used. Hedonic variables were divided into sale and intrinsic characteristics. The former group consists of *year of sale*, *auction house*, *evening auction* (dummy). The latter group features *canvas/wooden support* (dummy), *size*, *signature*, and *working period* categorised based on the work of Czujack (1997). The rest of the intrinsic characteristics contains colour-related variables.

Having used the RGB colour model, Stepanova pointed out that each image in the dataset consists of 2000 different colours on average, which cannot be analysed using the hedonic model. Therefore, she proposed to use colour quantisation (unfortunately, the selected algorithm is not specified in the paper). For each image, the 10 most dominant colours were extracted. Each pixel was then assigned to one of them. The quantised colour analysis of the available Picasso artworks in the RGB colour space revealed two high-concentration areas: orange and blue-teal clusters. Stepanova also tackles the notion of the impact of *colour diversity* in her research. Here, diversity is defined as the average Euclidean distance between colours in the RGB colour space. The examples of high- and low-diversity paintings of Picasso are presented in Figure 3.13.

For the Picasso dataset, three models (all data, the *blue and red* period excluded, and artist age instead of his artistic period) were prepared – all can be characterised with an R^2 around 70%. In these models, the *surface* was the statistically significant variable, which is in line with similar research previously mentioned in this work. Other significant variables were the *evening sale* and *more than 2 book mention* dummies. From the perspective of this work, the most interesting variables are the colour-related ones. The logarithm of the *diversity of colours*, as well as the share of the aforementioned clusters showed strong and statistically significant positive correlations with the price. Following Stepanova (2015), a 1000 cm^2 increase of the area occupied by the blue-teal colours will result in a 21% increase in the price. As the author suggests, this might stem from the fact that Picasso's *blue period* resulted in his most expensive works. Regarding diversity, a 1% increase of this value results in a price increase of 0.58%. This variable was also statistically significant for both Rothko and Hoffman – an increase of 1% in the diversity results in a 0.34% price increase. However, the analysis of colour histograms has not revealed any clusters correlated with prices.

(a) Pablo Picasso, *Buste de femme assise sur une chaise*, 1939. Source: `http://christies.com` [accessed on August, the 2nd, 2020]

(b) Pablo Picasso, *Bouteille, vere et pipe*, 1914. Source: `http://mutualart.com` [accessed on August, the 2nd, 2020]

Figure 3.13: High (left) and low (right) colour diversity artworks, as presented by Stepanova (2015).

In another work, Pownall and Graddy (2016) analysed prints made by Andy Warhol in order to examine the influence of colour intensity and lightness on price. Using the RGB colour model and CIE L*a*b* colour space, they argue that intense colours have a positive impact on the final price, and the darkness is more valued than lightness. In numerous hedonic models, the artist dummy is often one of the most important hedonic variables. To reduce this effect and focus on colour-related attributes, Pownall & Graddy investigated the works of a single artist. Andy Warhol was chosen, since he produced numerous colourful pop-art works in the 1960s, on which he depicted popular celebrities (Marilyn Monroe for instance). Since these works were printed, they had been issued many times in limited collections. Moreover, in Warhols' works the same image was often used several times, but with different colours. These features make his *catalogue raisonné* a perfect dataset for controlling for colours. In addition to that, the dataset was narrowed down to works of similar size, genre and auction location. The used dataset consists of 178 observations sold in 2012 in two major auction houses (Christie's and Sotheby's), with an average of $40,000 ($218,500 was the highest price, whereas $2,250 – the lowest). The size of each image was reduced to 200×200, resulting in 4k pixels in a single one.

The experiment carried out by Pownall & Graddy was twofold. In the first run, the RGB colour model was used, whereas the second was performed with CIE L*a*b*. Both were used for

measuring the intensity of colours and lightness. For the RGB part, the intensity was measured simply by the mean $R/G/B$ value for an image (the lower the RGB, the darker the image). Concerning lightness, Matlab's `rgb2gray` function was used to convert RGB to greyscale. It removes hue and saturation and retains the luminance using the following function:

$$\texttt{rgb2gray}(R, G, B) = 0.2989 \cdot R + 0.5870 \cdot G + 0.1140 \cdot B. \tag{3.68}$$

The lightness is calculated as an average value of the aforementioned formula. Regarding CIE L*a*b*, the intensity was measured as the average values of a* and b* components. As for the lightness, there were no further calculations needed, as the L* component stands for this attribute (yet again the mean value was taken).

There were three sets of 12 models in total, each of which was a standard hedonic price model in a semi-logarithmic form. The models were controlled for the following independent variables: a number of editions, size, age, place of sale (dummy: London or New York), auction house (dummy: Christie's or Sotheby's). For editions, size and age a natural logarithm was taken from the original values. Since all observations came from the same artist and were sold in the same year, there was no need to use *artist* and *year* dummy variables. The first set of 4 models controlled the average of R, G, B, and all of them. All separate values turned out to be statistically significant ($\alpha = 0.05$) with a negative coefficient, which means that the price increases by 0.3% per one unit intensity. However, using all of them resulted in multicollinearity and none was statistically significant. The lightness was tested by grey mean (statistically significant negative impact at $\alpha = 0.01$) and standard deviation (no statistical significance). Regarding the four CIE L*a*b* models (all components were used alone, except the final model, in which all was used), only the L* component was significant ($\alpha = 0.01$) – similarly to RGB, the coefficient was negative.

The values of the rest of the estimated coefficients were similar among the models. In general, the greater the total number of editions, the lesser the price is, as it follows the intuition in which uniqueness has its price. This was confirmed in the experiment, as in the majority of models the coefficient for editions has a negative value and was statistically significant at the 0.1 level. Age and size turned out to positively impact price as well ($\alpha = 0.01$). All the models had R^2 within the range of 0.20-0.23. The fact that these values are relatively low stems from a simple truth. As the authors of this paper suggest, most of the explained variance in similar models comes

from the *artist* dummy. Since only Pollock is considered and hence this dummy is missing in the model, the coefficient of determination is relatively low. Interestingly, the authors mentioned k-means clustering in the paper, but it was used only for showing the R/G/B colour clusters in a single image.

In another publication, Charlin and Cifuentes (2018) argue that colours play a vital role in art – they mention Anish Kapoor is known for her special kind of black, or Yves Klein, which has its own type of blue. However, this relation is relatively unexplored, as the authors claim. They prepared a dataset consisting of 169 works of Mark Rothko produced between 1950 and 1970, sold mostly at Christie's and Sotheby's. Premium prices were adjusted to 2018 USD. A standard log-linear HR model was used, using the following explanatory variables: *area*, log(*area*), *work on paper*, *evening sale*, *year of sale*. The yielded adjusted R^2 was equal to 0.86. Average R/G/B and H/S/V coordinates added to the equation turned out to be not statistically significant, though adding only saturation or value individually from the latter colour model turned out to be significant. When added to the model, average L*/a*/b* values resulted in statistically significant dependent variables, though.

The authors of this paper focused on two colour-related features: contrast and diversity. For determining the contrast of a given image, the k-means algorithm was used to extract at most 10 dominant colours from paintings. Then, for each painting, the three most dominant colours (considering pixel share) represented by their centroids c_i are chosen. The contrast measure D is calculated as a sum of distances of each cluster centres, given with the following formula:

$$\mathrm{D} = \mathrm{d}(c_1, c_2) + \mathrm{d}(c_2, c_3) + \mathrm{d}(c_1, c_3). \tag{3.69}$$

The greater the distance between centroids, the higher the contrast measure is. Colour diversity was measured using Herfindahl Index (abbreviated as HI), which is known for its usage in portfolio diversification scenarios as a concentration measure (Woerheide & Persson, 1992). For the L* component was partitioned to $[0, 20]$, $[21, 40]$, $[41, 60]$, $[61, 80]$, $[81, 100]$, whereas for a* and b* the intervals were set as follows: $[-128, -78]$, $[-77, -27]$, $[-26, 25]$, $[26, 76]$, $[77, 127]$. Thus, the space was partitioned to $5^3 = 125$ cells – λ_j denotes the percentage of pixels in a j-th colour cell. Herfindahl Index is then given by the following equation:

$$\mathrm{HI} = \frac{1 - \sum_{j=1}^{125} \lambda_j^2}{1 - \frac{1}{125}}. \tag{3.70}$$

The index value ranges from 0 (all pixels in a single cell) to 1 (perfect diversification). In conclusion, Charlin and Cifuentes (2018) claim that contrast and diversity – defined as above – have a positive impact on price, at least in terms of the aforementioned dataset with Rothko's paintings. Surprisingly, the colour itself (here as a hue) does not have a statistically significant impact on price.

Habalová (2018) has carried out an extensive study on the price determinants for art photography, in which colour-related features were considered as well. The used dataset was collected from Sotheby's and Philips auction houses, containing 368 artworks from 147 different authors and sold between 2016 and 2017. For the data imputation process, mean values were used. Habalová tried to explain price determinants using OLS, model averaging, regression trees, and random forests. To compare these, the mean average precision estimate (abbreviated as MAPE) was used.

The first set of OLS models series has been built around expert price estimates. Sole estimates in Model 0 turned out to be quite a good predictor ($R^2 = 0.81$). Then, Habalová tried adding mean R/G/B values (Model 1, $R^2 = 0.79$), dominant R/G/B values (Model 2, $R^2 = 0.81$), dominant H/S/V values (Model 3, $R^2 = 0.81$), and the colour diversity (Model 4, $R^2 = 0.81$) – as in Stepanova (2015). All the models can be characterised with relatively high R^2 coming mostly from the *mean estimate* variable. Therefore, this variable was excluded in the next set of models in favour of a set of the following hedonic variables: *year of birth*, *nationality*, *sex*, *formal education in art* (dummy), *number of words in their biography at Wikipedia*, theirs *Artfacts rank*, *number of searches on Artforum*, and *age*. The R^2 of these five models (without estimates) was consecutively 0.21, 0.20, 0.21, 0.21, and 0.21. This shows that most experts' estimates are a powerful predictor, whereas the evidence of the impact of colour is mixed in this dataset.

In order to improve MAPE, Habalová explored three machine learning techniques: model averaging, regression trees and random forests. Akaike Information Criterion (usually abbreviated as AIC) was used for the model averaging procedure, in which the importance of the colour-related variables turned out to be smaller than in a traditional OLS model building. The CART algorithm (Breiman, Friedman, Stone, & Olshen, 1984) was used for building regression trees. They turned out to have greater predictive performance, especially when *estimates* were omitted in the model. In the case of estimate-based models, the algorithm discarded other variables. For the random forest algorithm, 500 trees were used. It had the highest MAPE – around 15% – when all the variables were incorporated. All in all, the obtained results suggest the effect of

colours seemed to be marginal in this research, whereas the most of explanatory power came from experts' estimates.

Other research related to colour clustering in the art also exist – for instance, Moosburger (2017) explored colour quantisation for paintings, whereas Kim, Son, and Jeong (2014) used low-level techniques for putting art history terms (such as *sfumato*) in image processing context. These works, however, do not explore the relations of these features to the price. This section presented research relevant to the subject of this dissertation. Although a small number of researchers tried to grapple with the issue of colour-related price determinants for artworks, it seems that this is still a relatively unexplored topic. The presented papers often incorporated rather small samples and some of them yielded contrasting results, which prevented drawing more general conclusions. The usage of relatively simple methods for colour quantisation also leaves room for potential improvement.

3.5 Summary

In this chapter, two intertwined topics were the subject of our analysis. The first section presented a broad range of popular colour models and spaces since they were needed to form a basis for a more advanced discussion on colour-related topics. These two enabled us to answer the research question Q2 (*How to extract colour-related information from paintings?*). In the first section, we presented a broad range of popular colour models and spaces, since they were needed to form a basis for a more advanced discussion on colour-related topics. Two leading families of colour models are RGB- and CIE-related. The first one is prevalent and can be very descriptive (e.g. the saturation attribute in HSV), despite relatively small gamuts in some cases. The main drawback of these colour spaces is the fact that they are not perceptually uniform. CIE-related colour spaces, such as CIE $L^*a^*b^*$ colour space addresses this problem. Later in this dissertation, we extensively use sRGB and CIE $L^*a^*b^*$ colour spaces. The first will be used to describe colours (for example, in Chapter 6), whereas the second one is used within one of the considered quantisation algorithms (Chang's k-means). CIE $L^*a^*b^*$ is also a part of the CIEDE2000 formula from Equation (3.18), which will be used in Section 5.1 in Equation (5.1) in order to compare diversity of colours during colour quantisation algorithm comparison.

Section 3.2 introduced the notions of features and descriptors in the domain of computer vision and image processing. Since colour-related attributes have special attention in this dis-

Table 3.3: Comparison of selected colour quantisation methods (with complexity notation borrowed from Section 3.3).

Method	Author	Complexity		Key features
		Time	Space	
Uniform	–	$\mathcal{O}(k)$	$\mathcal{O}(0)$	very fast, uniformly selects representatives, large error
Popularity	Boyle & Lippman, Cohen	$\mathcal{O}(dk)$	$\mathcal{O}(r)$	histogram-like representative selections, large error
k-means	Lloyd (1982)	$\mathcal{O}(nkt)$	$\mathcal{O}(n(d+k))$	minimises sum of squares
Chang's k-means	Chang et al. (2015)	$\mathcal{O}(nkt)$	$\mathcal{O}(n(d+k))$	k-means with increased colour diversity
Median Cut	P. Heckbert (1980)	$\mathcal{O}(n \log_2 k)$	$\mathcal{O}(d)$	resembles k-d tree
Octree	Gervautz and Purgathofer (1988)	$\mathcal{O}(n)$	$\mathcal{O}(k)$	tree-based

Source: own study.

sertation, this was reflected in this section – it provided broad guidance on popular features connected to colours. We also described some of the MPEG-7 descriptors since one of its groups are devoted to colours. A number of these features and descriptors used colour quantisation. This served as an inspiration to explore this topic in greater detail in the next section, in which we discussed the most popular colour quantisation algorithms. A selection of them is briefly summarised in Table 3.3. Three of them seem to be especially popular due to their ability to generate decent results – median-cut, octrees, and k-means. Variations of weighted k-means with enhanced colour distinctiveness turned out to be a good fit for the purpose of colour-related painting features. These algorithms are evaluated in Chapter 5 in terms of generating representatives for paintings.

Finally, in Section 3.4 other colour-related research was analysed. Experiments carried out by Stepanova (2015), Pownall and Graddy (2016), Charlin and Cifuentes (2018) and Habalová (2018) have been presented and discussed. While these works seem to pioneer the topic of the impact of colours on the price, they were conducted on rather small data samples. There is also room for improvement in terms of the used quantisation techniques and usage of their results.

Chapter 4

Modern Explainable Artificial Intelligence Methods with Decision Trees

Andrew Ng once famously said that *"Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform in the next several years"*. The last decade in artificial intelligence was dominated by machine learning (often abbreviated as ML), which resulted in numerous methods and algorithms. Paired with explainable artificial intelligence techniques, machine learning broadens the range of answers to the research question Q1 (*Which methods can be used to assess the importance of paintings' features for the hammer price?*) presented in Chapter 2. One can use them to generate models that are more complex than traditional quantitative art market analysis tools and compare their outcomes. While deep neural networks are prevalent for solving tasks related to unstructured data (such as computer vision and natural language processing), decision trees models are often used for tabular data due to their effectiveness. Since the latter category is suitable for art market data, tree-based methods are described in Section 4.1. This work, however, is concerned with the explanatory analysis rather than sole prediction. Therefore, selected explainable AI techniques are described in Section 4.2. They are later used to analyse traditional and especially tree-based models, as the former does not have a straightforward interpretation. Finally, this chapter is concluded by a summary in Section 4.3.

4.1 Solving Machine Learning Tasks with Decision Trees

Machine learning is a field of study concerned with algorithms and models for detection and extrapolation of patterns (Russell & Norvig, 2016) in order to solve some *task* based on training data, without being explicitly programmed to do so (Samuel, 1959). Machine learning can be divided and classified in many ways. One of the basic approaches considers three categories, depending on the task they are trying to solve:

- *supervised learning* – the training dataset contains observations and the target variable (labels or numerical values) and it is used to predict new targets on unseen data,
- *unsupervised learning* – the training dataset contains observations without the target variable and it is used to generalise (often using patterns or clusters) new targets on unseen data,
- *reinforcement learning* – some *agent* is trained to perform *actions* that maximise its reward function in a given *environment*.

One can enlist more specialised tasks (such as semi-supervised, continual, or few-shot learning), though these three are typically perceived as the most fundamental way to ML approaches. As the target variable is available in our case (the hammer price of a sold lot – see Chapter 6), we are concerned with supervised learning.

There are numerous approaches to this task and enlisting them all is beyond the scope of this work. To name a few, one can use k-nearest neighbour (Fix & Hodges, 1989), support vector machines (Boser, Guyon, & Vapnik, 1992), neural networks (including deep learning approaches and convolutional neural networks) (Goodfellow et al., 2016), or decision tree based approaches. Recently, machine learning competitions have been dominated by deep neural networks and tree-based approaches. Deep neural networks offer state-of-the-art results on unstructured data (such as images, video, natural language, or speech). Tree-based models (with a notable example of the XGBoost algorithm) in general outperform deep neural networks in structured, tabular data (Shwartz-Ziv & Armon, 2021). They also offer easier hyper-parameter tuning and are relatively easier to interpret. On the other side, some deep learning tricks, such as convolutions, might be ineffective in the tabular world.

Auction catalogues and results can be presented as structured, tabular data. This property makes them a perfect subject for tree-based methods. Decision trees are tree-like flowcharts, which try to model some phenomena and support decision making. In the context of supervised

machine learning, a decision tree is a structure used for solving the prediction tasks. The structure of a tree is the subject of the learning process. This section describes different approaches to building tree-based models – starting from classic decision trees (Section 4.1.1), we then describe ensemble methods and state-of-the-art techniques, such as XGBoost (Section 4.1.2).

4.1.1 Classic Decision Trees

As the name suggests, a decision tree can be represented as a classical data structure – a tree. Following Cormen, Leiserson, Rivest, and Stein (2009), a tree is an abstract non-linear data structure, which can be perceived as a specific acyclic graph. The "first" node (i.e. a node without an ancestor) is called the root, whereas nodes without children are called leaf nodes – each node can have an arbitrary number of its children. A decision tree is a representation of a function of a vector of features, which returns a single value for the target variable – discrete for classification and continuous for regression problems (Russell & Norvig, 2016). In this context, all nodes except the leaves represent intermediate decisions – starting from the root node. For categorical variables, the number of children (i.e. possible choices) depends on the number of categories. If a given variable is numerical, a set of intervals are chosen first, based on hyper-parameters called *splitters*. The one a given value belongs denotes the category. Leaf nodes represent final decisions (specific numbers or categories).

While decision trees structure appears clear and interpretable in general, their construction is not trivial. The problem of decision tree learning falls into the category of supervised machine learning tasks, which means that it is incrementally built from labelled data samples. A generic heuristics for learning decision trees is presented in Algorithm 13, which is a greedy divide-and-conquer strategy – it searches for the most important attribute to split on using the Importance method. The tree is built starting from the root. The child nodes are added subsequently with recursive calls of the algorithm. It can be done as long as there is only one category left in a given branch, though the height of the tree is limited in practice. The values for leaf nodes are given by their Plurality-Value. In classification problems, they contain the mode of the target variable with the qualities of all ancestor nodes. For regression problems, the mean is used. However, this is just the generic approach to the problem – numerous variations present different modifications of this method.

Algorithm 13 Decision tree learning algorithm.

function DECISION-TREE-LEARNING(*examples, attributes, parentExamples*)
  if *examples* is empty **then return** PLURALITY-VALUE(parentExamples)
  else if all *examples* have the same classification **then return** the classification
  else if *attributes* is empty **then** PLURALITY-VALUE(*examples*)
  else
    $A \leftarrow \underset{a \in attributes}{\arg\max}\ \text{IMPORTANCE}(a, examples)$
    $tree \leftarrow$ a new decision tree with root test A
    for each value v_k of A **do**
      $exs \leftarrow \{e | e \in examples$ **and** $e.\text{A} = v_k\}$
      $subtree \leftarrow$ DECISION-TREE-LEARNING$(exs, attributes - A, examples)$
      add a branch to *tree* with label $A = v_k$ and subtree *subtree*
  return *tree*

Source: Russell and Norvig (2016)

Considering n binary attributes, there are 2^{2^n} possible decision trees which can be built (Russell & Norvig, 2016). Moreover, finding an optimal binary decision tree is proven to be NP-complete (Laurent & Rivest, 1976). Therefore, choosing the right heuristics for building a tree is a crucial step. At the heart of the presented algorithm lies the definition of the IMPORTANCE method, which is often called *information gain*. It is based on the concept of *impurity*, which can be perceived as a measure of the homogeneity of a given split. There exist several methods to calculate this value. The most popular methods are information gain using entropy, the Gini impurity, and variance reduction. The first two are dedicated to classification tasks, whereas the latter method is for regression.

The first method for classification uses the notion of entropy. The idea was introduced by Shannon (1948) in his famous work. Entropy H can be perceived as a measure of uncertainty. It can be defined as follows:

$$H(T) = -\sum_{i=1}^{n} p_i \log_2 p_i, \tag{4.1}$$

where p_i denotes the is the percentage of data samples belonging to each class on a given split for the i-th variable out of n in the set of labelled learning examples $T = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ and $\sum_{i=1}^{n} p_i = 1$. The logarithm base is connected to units resulting from this equation – for instance, base 2 gives bits. The impurity I_E is simply defined as its entropy:

$$I_E(p_i, \ldots, p_n) = H(T). \tag{4.2}$$

The information gain IG (which is in fact KullbackLeibler divergence) of T for the attribute

$1 \leq a \leq m$ (split candidate) is then calculated as follows:

$$\mathrm{IG}(T, a) = \mathrm{H}(T) - \mathrm{H}(T|a). \tag{4.3}$$

This equation uses also the notion of the conditional entropy, which is defined as follows:

$$\mathrm{H}(T|a) \ = -\sum_{v \in V} \frac{|\{\mathbf{x} \in T | x_a = v\}|}{|T|} \cdot \mathrm{H}\left(\{\mathbf{x} \in T | x_a = v\}\right), \tag{4.4}$$

which operates on the same notation as the aforementioned equations (V denotes the set all the possible values v for the attribute a). Such an approach is sometimes called simply information gain (without mentioning entropy) and it is used in popular decision tree learning algorithm, such as C4.5 (Quinlan, 2014).

The Gini impurity I_G uses simpler equation, defined as follows:

$$I_G(T) = 1 - \sum_{k=1}^{m} p_k^2, \tag{4.5}$$

where p_i is the share of items belonging to a given class. It is used for instance in the CART algorithm (an abbreviation from classification and regression trees), which was presented by Breiman et al. (1984). In classification problems, the Gini impurity appears to be a slightly more popular approach – perhaps thanks to the fact that it is more computationally feasible, as a simple sum of squares is calculated faster than more expensive operations involving logarithms in the entropy equation. This does not necessarily mean that the Gini impurity yields better results than entropy.

For regression problems, there are different criteria for impurity. For instance, one can use variance reduction techniques I_V. There are several approaches to achieving this. For example, variance reduction based on mean square error is defined as follows:

$$I_V(T) = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \mu\right)^2, \tag{4.6}$$

where y_i is the value of a given instance from the training set, m denotes the number of instances, and the mean μ is given by $\frac{1}{m} \sum_{i=1}^{m} y_i$. Variance reduction is used in the aforementioned CART algorithm for solving regression problems.

Decision trees often provide decent results for many classification and regression tasks. The

used data do not need any normalisation nor any general assumptions about them, which makes them immune to problems such as co-linearity. The ease of use and their interpretability makes decision trees a popular choice for numerous problems. However, they are often criticised for being prone to over-fitting. This is a classic supervised machine learning problem, in which a given model can be characterised with a relatively good training error, but poor performance with previously unseen data (test error). This problem can be mitigated by controlling the hyper-parameters (e.g. the maximum height of the tree or the impurity criterion), or using methods such as k-fold cross-validation.

4.1.2 Ensemble Methods for Decision Tree Learning

In machine learning, ensemble methods constitute an umbrella term for combining predictions from multiple models to improve results. More formally, these methods can be perceived as meta-algorithms that select an ensemble of hypotheses from the hypothesis space instead of yielding only one (Russell & Norvig, 2016). In the context of ensemble methods, a single machine learning model is often called a *weak learner*. On the other hand, ensemble models are dubbed as *strong learners*. In general, these techniques are focused on finding a better bias-variance trade-off compared to standard techniques. They are especially popular paired with decision trees, which are notorious for their high variance.

Bagging (also known as *bootstrap aggregation*) is an example of ensemble machine learning technique. The method was introduced by Breiman (1996). It can be used for regression and classification problems as well. This technique consists of two stages – bootstrapping and aggregation. The first stage of bagging – bootstrapping – is just a sampling method (with replacement) for the original dataset. The sample size should be large enough to provide a good representation of the observed data distribution. Conducting sampling with replacement is assumed to make each sample independent of the other, which is generally the desired property in statistical-related problems. Then, weak learners are trained – this process can be conducted in parallel. In the aggregation stage, some averaging is applied to the obtained models. For example, it can be the mean output for the regression problem or the most popular class for the classification.

Breiman (2001) introduced *random forests*, which is a special kind of decision tree ensembles. They extend the idea of bagging by using only a subset of available features for every sample. The subset of considered attributes is chosen at random. This approach turned out to be very effective and produce better strong learners. Random forests can be applied to classification and

regression problems as well. The algorithm became very popular in the academic and business community as well, being used to this day.

Another approach to ensemble learning is *boosting* – contrary to bagging, this process can be perceived as sequential. In general, single weak learners are trained iteratively and misclassified data samples are given higher priority. There exist several approaches to conducting this process and merging intermediate hypotheses. Two notable examples are adaptive and gradient boosting. The first one changes the weight of particular samples from the training set to promote difficult examples in the forthcoming iterations. The latter method changes the actual values of the observation to achieve the same goal.

AdaBoost (an abbreviation from adaptive boosting) was introduced by Freund and Schapire (1997). Following Russell and Norvig (2016), it can conceptually be seen as a sequential learning process, which uses the concept of a weighted training set – each training example $(\mathbf{x}_i, y_i)$ has now the corresponding element – w_i, which denotes the sample weight. At first, the training set is treated equally. After learning the first model, the weights are adjusted to increase the importance of wrongly predicted data. The process of learning and adjusting is repeated a given number of times to obtain a strong learner. The procedure is presented in Algorithm 14, where $\mathbf{T} = \{(\mathbf{x}_i, y_i)\}$ is the training set consisting of N examples, L denotes the learning algorithm, and M is the number of hypotheses in the ensemble. Within the algorithm, $\mathbf{w}$ stands for the sample weights (initially, $\forall_i\ \mathbf{w}[i] = \frac{1}{N}$) and $\mathbf{h}$ represents the vector of hypotheses with their corresponding weights stored in $\mathbf{z}$.

Algorithm 14 AdaBoost.

function ADABOOST($\mathbf{T}, L, M$)
  for $m \leftarrow 1$ **to** M **do**
    $\mathbf{h}[m] \leftarrow L(\mathbf{T}, \mathbf{w})$
    $\epsilon \leftarrow 0$
    for $j \leftarrow 1$ **to** N **do**
      if $\mathbf{h}[m](x_j) \neq y_j$ **then** $\epsilon \leftarrow \epsilon + \mathbf{w}[j]$
    for $j \leftarrow 1$ **to** N **do**
      if $\mathbf{h}[m](x_j) = y_j$ **then** $\mathbf{w}[j] \leftarrow \mathbf{w}[j]\frac{\epsilon}{1-\epsilon}$
    $\mathbf{w} \leftarrow$ NORMALISE($\mathbf{w}$)
    $\mathbf{z}[m] \leftarrow \ln \frac{1-\epsilon}{\epsilon}$
  return WEIGHTED-MAJORITY($\mathbf{h}, \mathbf{z}$)

Source: Russell and Norvig (2016)

Another approach to this type of ensemble method is called gradient boosting (often abbrevi-

ated as GBM, where M stands for machine), which can be perceived as a more general approach since it allows different loss functions. The first reported works on this topic were conducted independently by Friedman (2001) and Mason, Baxter, Bartlett, and Frean (2000). Gradient boosting is a greedy strategy, which updates the model in an additive manner. Following Parr and Howard[1], the prediction $\hat{y}$ based on an input feature vector $\mathbf{x} \in \mathbf{X}$ is assumed to be a sum of M sub-functions $f_m(\mathbf{x})$:

$$\hat{y} = \sum_{m=1}^{M} f_m(\mathbf{x}) = F_M(\mathbf{x}). \tag{4.7}$$

In this approach, each f_m is considered as a weak model, added stage-wise. The number of stages M is a hyper-parameter. The aforementioned relation can be presented in a recursive form:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + f_m(\mathbf{x}), \tag{4.8}$$

At first, the model F_0 is initialised with some value. The structure of the weak learner consist of two actual elements:

$$F_m(\mathbf{X}) = F_{m-1}(\mathbf{X}) + \eta\Delta_m(\mathbf{X}), \tag{4.9}$$

where $\Delta_m(\mathbf{X})$ is the actual weak model. The size of a change is controlled by the η hyper-parameter, which is called the learning rate. At the heart of gradient boosting lies the arbitrary loss function, which has to be optimised. For instance, the popular L_2 loss function (MSE) is given by the following definition:

$$L(\mathbf{y}, F_M(\mathbf{X})) = \sum_{i=1}^{N} \left(y_i - F_M(\mathbf{x}_i)\right)^2. \tag{4.10}$$

Although the original MSE equation should also be multiplied by $\frac{1}{N}$ to be an actual mean, this fraction is treated here as a constant and dropped in further calculations, since it does not affect

[1] `https://explained.ai/gradient-boosting/descent.html` [accessed on August, the 2nd, 2020]

them. To minimise the loss, its gradient has to be calculated:

$$\begin{aligned}\frac{\partial}{\partial \hat{y}_j} L(\mathbf{y}, \hat{\mathbf{y}}) &= \frac{\partial}{\partial \hat{y}_j} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\ &= \frac{\partial}{\partial \hat{y}_j} (y_j - \hat{y}_j)^2 \\ &= 2 (y_j - \hat{y}_i) \frac{\partial}{\partial \hat{y}_j} (y_j - \hat{y}_i) \\ &= -2 (y_j - \hat{y}_i) .\end{aligned} \tag{4.11}$$

The gradient $\nabla_{\hat{\mathbf{y}}} L(\mathbf{y}, \hat{\mathbf{y}})$ is therefore equal to $-2 (\mathbf{y} - \hat{\mathbf{y}})$. The scalar can be left aside, which reduces the gradient to a simple residual vector $\mathbf{y} - \hat{\mathbf{y}}$. It is called pseudo-residual in the original paper and is used to train the next weak learner. Algorithm 15 sums up the procedure and presents the boosting algorithm for regression trees using L_2 loss function. The algorithm can work with other loss functions, such as L_1 loss, also known as the mean absolute error (MAE). The general, more complicated approach involving the explicit gradient of the arbitrary loss function is presented in Algorithm 16.

Algorithm 15 Gradient boosting for regression trees with L_2 loss.

1: **function** GBM-L2($\mathbf{X}, \mathbf{y}, M, \eta$)
2: $F_0 (\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} y_i$
3: **for** $m \leftarrow 1$ **to** M **do**
4: $\mathbf{r}_{m-1} \leftarrow \mathbf{y}_{m-1} - F_{m-1}(\mathbf{X})$
5: Train regression tree Δ_m on $\mathbf{r}_{m-1}$, minimise squared error
6: $F_m(\mathbf{X}) \leftarrow F_{m-1}(\mathbf{X}) + \eta \Delta_m(\mathbf{X})$
7: **return** F_m

Source: Friedman (2001) and `https://explained.ai/gradient-boosting/descent.html` [accessed on August, the 2nd, 2020]

Algorithm 16 Gradient boosting for regression trees with arbitrary loss.

function GBM($\mathbf{X}, \mathbf{y}, M, \eta$)
  $F_0(\mathbf{x}) = \underset{v}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, v)$
  for $m \leftarrow 1$ **to** M **do**
    $\hat{\mathbf{y}}_{m-1} = F_{m-1}(\mathbf{X})$
    $\mathbf{r}_{m-1} \leftarrow \nabla_{\hat{\mathbf{y}}_{m-1}} L(\mathbf{y}, \hat{\mathbf{y}}_{m-1})$
    Train regression tree Δ_m on $\mathbf{r}_{m-1}$, minimise squared error
    for each l **in** Δ_m **do**
      $w \leftarrow \sum_{i \in l} L(y_i, F_{m-1}(\mathbf{x}_i + w))$
      Alter l to predict w
    $F_m(\mathbf{X}) \leftarrow F_{m-1}(\mathbf{X}) + \eta \Delta_m(\mathbf{X})$
  return F_m

Source: Friedman (2001) and `https://explained.ai/gradient-boosting/descent.html` [accessed on August, the 2nd, 2020]

Chen and Guestrin (2016) presented XGBoost (an abbreviation from Extreme Gradient Boosting), which is a scalable machine learning system for tree boosting, implemented in popular programming languages. XGBoost can be seen as an enriched gradient boosting framework, adding some important practical and algorithmic features to the original meta-algorithm. For instance, it employs regularisation, which is a technique for penalising complex models and avoiding over-fitting. It also has advanced algorithms for split finding for tree building. On the technical side, XGBoost can be characterised by its scalability, parallelism, or distributed computing. The algorithm is also ready to work with real-world sparse data. The authors provided the out-of-core computation mechanism for large datasets, which doesn't fit into memory at once.

Following Chen and Guestrin (2016), in XGBoost the loss function at the step t is now given by the following equation:

$$\mathcal{L}^{(t)} = \sum_i L\left(y_i, \hat{y}_i\right) + \sum_k \Omega(f_k), \tag{4.12}$$

where L is a differentiable convex loss function and $\Omega(f_t)$ is the regularisation part of the equation and is defined as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2. \tag{4.13}$$

Here T is a number of the leaves in the tree, w is the weight of a score in the corresponding leaf, and γ and λ are hyper-parameters that control how much the model is penalised for its complexity. The learning parameter η (not shown in the equation directly) is dubbed as shrinkage by the authors. The rest of the notation is the same as in the gradient boosting example. XGBoost also uses column sub-sampling, which is a mechanism known from random forests. During deci-

sion tree learning, much time is devoted to finding good split candidates for numerical features. The authors of XGBoost extend the popular exact greedy and approximate algorithms for split finding, enhanced with mechanisms for finding weighted quantiles and sparsity-awareness. As of 2020, XGBoost and deep neural networks are often employed in numerous machine learning competitions to provide state-of-the-art results in various tasks, such as classification or regression, regardless of the considered domain. While the latter is popular for computer vision and natural language processing tasks, XGBoost often outperforms other algorithms in numerous supervised machine learning tasks on structured, tabular data.

4.2 Selected Explainable Artificial Intelligence Frameworks

Linear regression offers p-values and easy to interpret coefficient estimates, making the models transparent to examination. While simple tree-based models can also be easily investigated and visualised, the ones created with XGBoost are too complex to understand. With the advent of artificial intelligence and the presence of increasingly complex models such as ensemble methods or deep neural networks, there is an increasing need for methods explaining these models. *Explainable Artificial Intelligence* (abbreviated as XAI) is the discipline concerned with techniques used for making black-box models understandable by humans and improving the general interpretability of models. Biecek and Burzykowski (2020) classify these methods into two categories: instance and dataset-level explanations. As the names suggest, the former focuses on explaining the model output for a single observation, whereas the latter examines the model behaviour in the presence of all considered observations. This subsection explains a selection of useful XAI methods briefly, as they are used[2] in later sections of this work.

For some machine learning approaches, there exist a number of model-specific tools for assessing the importance of the used variables. For instance, p-values are commonly used with linear regression. The more general concept of feature importance was introduced by Breiman (2001) along with his random forest algorithm. As explained in the previous sections, this algorithm needs some function to measure the decrease of impurity. Values obtained in the training process are later reused for calculating variable importance as a mean decrease of the impurity among all the trees. As a consequence, all the functions considered as impurity measures can be used for calculating variable importance, including the functions presented in the previous

[2]Biecek (2018) also provided a convenient R/python package `DALEX`, which is used for XAI tasks in this work.

subsection – such as Gini impurity or variance reduction. However, the main drawback of pure feature importance is its bias towards the features, which have a high cardinality. Therefore, there was a need to come up with a more robust method.

Permutation importance was introduced by Altmann, Toloşi, Sander, and Lengauer (2010). Following Biecek and Burzykowski (2020), it is a model-agnostic approach used for model exploration. The method for obtaining permutation importance values is presented in Algorithm 17. It operates on arbitrary model f with its $\mathcal{L}$ loss function (such as MSE), whereas $\mathbf{X}$ and $\mathbf{y}$ denote the feature matrix with corresponding real values and $\hat{\mathbf{y}}$ are the result of $f(\mathbf{X})$. The idea of the algorithm is simple – it permutes the values of each column in the feature matrix and then checks the influence of that permutation using the loss function. In the last step, the loss values can also be divided by themselves instead using subtraction, i.e. L_{*j}/L_0 can be used. Naturally, the higher vip_j is, the more important is the j-th feature. The main drawback of this method is its randomness stemming from the $\textsc{Permute}(\mathbf{X}_j)$. To minimise its impact on the calculation, the algorithm should be run multiple times – the results should be then averaged.

Algorithm 17 Permutation importance algorithm.

function PermutationImportance($f, \mathcal{L}, \hat{\mathbf{y}}, \mathbf{X}, \mathbf{y}$)
    $L_0 \leftarrow \mathcal{L}(\hat{\mathbf{y}}, \mathbf{X}, \mathbf{y})$
    for $\mathbf{X}_j$ **in** $\mathbf{X}$ **do**
        $\mathbf{X}_{*j} \leftarrow \textsc{Permute}(\mathbf{X}_j)$
        $\hat{\mathbf{y}}_{*j} \leftarrow f(\mathbf{X}_{*j})$
        $L^{*j} = \mathcal{L}(\hat{\mathbf{y}}_{*j}, \mathbf{X}_{*j}, \mathbf{y})$.
        $\text{vip}_j = L_{*j} - L_0$
    return vip

Source: Biecek and Burzykowski (2020)

Introduced along with gradient boosting machines by Friedman (2001), *partial-dependence profiles* (abbreviated as PD profiles) is a tool for dataset-level explanatory model analysis. Following Biecek and Burzykowski (2020), PD profiles intuitively depict the expected value of the target variable as a function of a given explanatory variable. They can be used for comparing models and investigate similarities and discrepancies between them. Technically, they are averaged *ceteris paribus profiles*, which makes them share the same disadvantages – they are not suited for correlated variables, for instance. Given a $n \times m$ feature matrix and a model f, a PD

plot for feature j is given by the following function:

$$\hat{g}_j(z) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_{k,1}, \ldots, \mathbf{X}_{k,j-1}, z, \mathbf{X}_{k,j+1}, \ldots, \mathbf{X}_{k,m}), \tag{4.14}$$

where the range of z should reflect the range of j-th feature in $\mathbf{X}$. For linear regression models, $\hat{g}_j$ forms a straight line.

Shapley additive explanations (abbreviated as SHAP) were introduced by Lundberg and Lee (2017) and further investigated by Lundberg et al. (2020). Using the hierarchy of Biecek and Burzykowski (2020), SHAP falls to a category of instance-level explanations. The idea is inspired by the concept from game theory – Shapley values (Shapley, 1953). Notice that Shapley values are not the same as SHAP values. Following Lundberg and Lee (2017), an explanation model for complex models (such as XGBoost) should be a simplified version of the original model. They propose an additive feature attribution model. Provided that f is the original model and $f(\mathbf{x})$ is its prediction, g is the explanation model operating on a simplified input $\mathbf{x}'$, which is given by $\mathbf{x} = h_{\mathbf{x}}(\mathbf{x}')$, where $h_{\mathbf{x}}$ is the mapping function. Linear function g of $\mathbf{z}' \in \{0,1\}^M$ represents the simplified model, whereas $\phi_i \in \mathbb{R}$ and M is the dimensionality of simplified input:

$$g(\mathbf{z}') = \phi_0 + \sum_{i=1}^{M} \phi_i \mathbf{z}'_i. \tag{4.15}$$

For $\mathbf{z}' \approx \mathbf{x}'$, local models should ensure that $g(\mathbf{z}') \approx f(h_{\mathbf{x}}(\mathbf{z}'))$. In their work, Lundberg and Lee (2017) compare different approaches for building additive explanations models (such as LIME or classical Shapley values), which try to set the values for ϕ_i They argue that there is exactly one model g which satisfies the properties of *local accuracy*, *missingness* and *consistency* and its coefficients are given by the following formula:

$$\phi_i\left(f, \mathbf{x}\right) = \sum_{\mathbf{z}' \subseteq \mathbf{x}'} \frac{|\mathbf{z}'|!(M - |\mathbf{z}'| - 1)!}{M!} \left(f_{\mathbf{x}}(\mathbf{z}') - f_{\mathbf{x}}(\mathbf{z}' \setminus i)\right), \tag{4.16}$$

where $|\mathbf{z}'|$ is the cardinality of $\mathbf{z}'$ excluding zero entries, $\mathbf{z}' \subseteq \mathbf{x}'$ stands for all $\mathbf{z}'$ in which non-zero values are a subset of non-zero values of $\mathbf{x}'$, $\mathbf{z}' \setminus i$ denotes $\mathbf{z}'$ with setting $\mathbf{z}'_i = 0$, and $f_{\mathbf{x}}(\mathbf{z}') = f(h_{\mathbf{x}}(\mathbf{z}'))$. Biecek and Burzykowski (2020) notice that SHAP values are based on additive contributions of the explained features, which might result in biased results for non-additive models. Another drawback is the time of calculating the values in model-agnostic

applications, though it can be done efficiently for tree-based models. Nevertheless, SHAP is considered a state-of-the-art observation-level explanation method.

4.3 Summary

This chapter described modern tree-based approaches for machine learning, along with selected explainable artificial intelligence. They were described in order to extend research methods to answer the research question Q1 (*Which methods can be used to assess the importance of paintings' features for the hammer price?*) better. Traditional decision trees paired with ensemble methods result in a powerful machine learning framework, which tops many data science competitions – XGBoost. We also presented a selection of explainable artificial intelligence frameworks, such as permutation-based importance and SHAP values. Tree ensembles, as well as XAI methods are used in the later chapters of this work. The XGBoost algorithm is employed to prepare a model for painting prices and compared to the traditional linear regression-based approach. Using these tools alone, the comparison would be possible in terms of traditional evaluation metrics (such as RMSE). However, it cannot provide much insight into price determinants, as boosted trees are too complex to interpret. Therefore, the presented explainable AI methods are used to shed light on the models and understand factors driving prices for the Polish art market.

Chapter 5

Extracting Colour-Related Features from Paintings

The goal of this dissertation is to examine buyers' preferences and price determinants for paintings on the Polish art market – not only using the features available in auction catalogues, but also the engineered ones. While the detailed statistics about the used dataset are provided in the next chapter, this one is dedicated to the description of the methods used for feature engineering. This technique is used to extend the number of variables available for the analysis. While standard features of paintings (such as size, author, etc.) are used in the analysis, we try to leverage the visual information stemming from the painting itself – the focus is on colours. Section 5.1 contains a large scale comparison of selected algorithms used for colour quantisation of paintings. The dataset used in this test fulfils the research objective O1 (*Prepare datasets allowing conducting the experiment*), though is discussed in detail in Section 6.1. These algorithms have been examined in terms of their mean squared error and colour diversity. This directly answers the research question Q3 (*What is the best colour quantisation algorithm for paintings?*). Section 5.2 describes two methods used for feature engineering – the selected colour quantisation algorithm for calculating colour share (Artefact 1/Algorithm 18) and the method for obtaining the colourfulness of a painting. The realisation of Artefact 1 fulfils the research objective O2 (*Develop a method for extracting colour-related features from paintings (Artefact 1)*). Like other chapters, this one is also concluded by a short summary in Section 5.3.

5.1 Comparing Colour Quantisation Algorithms for Feature Extraction from Paintings

The goal of this section is to examine which colour quantisation algorithms can be used for determining how colours influence hammer prices of artworks or buyer's decisions on the art market in general. Popular colour quantisation algorithms are compared in terms of quantisation error and colour diversity. For measuring quantisation error, a standard mean squared error formula is used. Recall Equation (3.61):

$$\mathrm{MSE}(\mathbf{I}_C, \mathbf{I}_Q) = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{p}_i - \hat{\mathbf{p}}_i||_2^2,$$

where $\mathbf{I}_C$ and $\mathbf{I}_Q$ denote the original and quantised images, consisting of n pixels each ($\mathbf{p}_i \in \mathbf{I}_C$ and $\hat{\mathbf{p}}_i \in \mathbf{I}_Q$). The lesser the MSE, the more similar the quantised image is (on average).

However, some artworks might present important details on relatively small regions using a distinct colour. As a consequence, these pixels will be outnumbered by larger areas and will not have their representative. A small colourful outbreak in one of his artworks can make it very distinct. This is especially true for artists, who use a relatively small colour palette. To illustrate this case, consider the following example of a portrait made by Stanisław Ignacy Witkiewicz, widely known as Witkacy (Figure 5.1). Some colour quantisation algorithms might have a tendency to omit the purple tie area, as it is not relatively large compared to the rest of this painting. If there's a small number of representatives to find, that tie might be omitted since it won't contribute much to MSE. For a human observer, however, that tie definitely stands out compared to the rest of the painting area.

Therefore, the algorithms are tested for *diversity* of generated representatives. Intuitively, it represents the average colour difference between all possible pairs of unique representatives. The higher the value, the more contrasting colours are in the image. Diversity is measured as follows:

$$\mathrm{div}(\mathbf{I}_Q) = \frac{1}{n}\sum_{(\mathbf{r}_1,\mathbf{r}_2)\in P}^{n} \Delta E^*_{00}\left(\mathbf{r}_1, \mathbf{r}_2\right), \tag{5.1}$$

where ΔE^*_{00} is the CIEDE2000 formula defined in Equation (3.18) for measuring colour difference, r_1 and r_2 are representatives in CIE L*a*b*, and P is a set containing $n = \binom{n_r}{2}$ combinations of n_r representatives in the quantised image $\mathbf{I}_Q$.

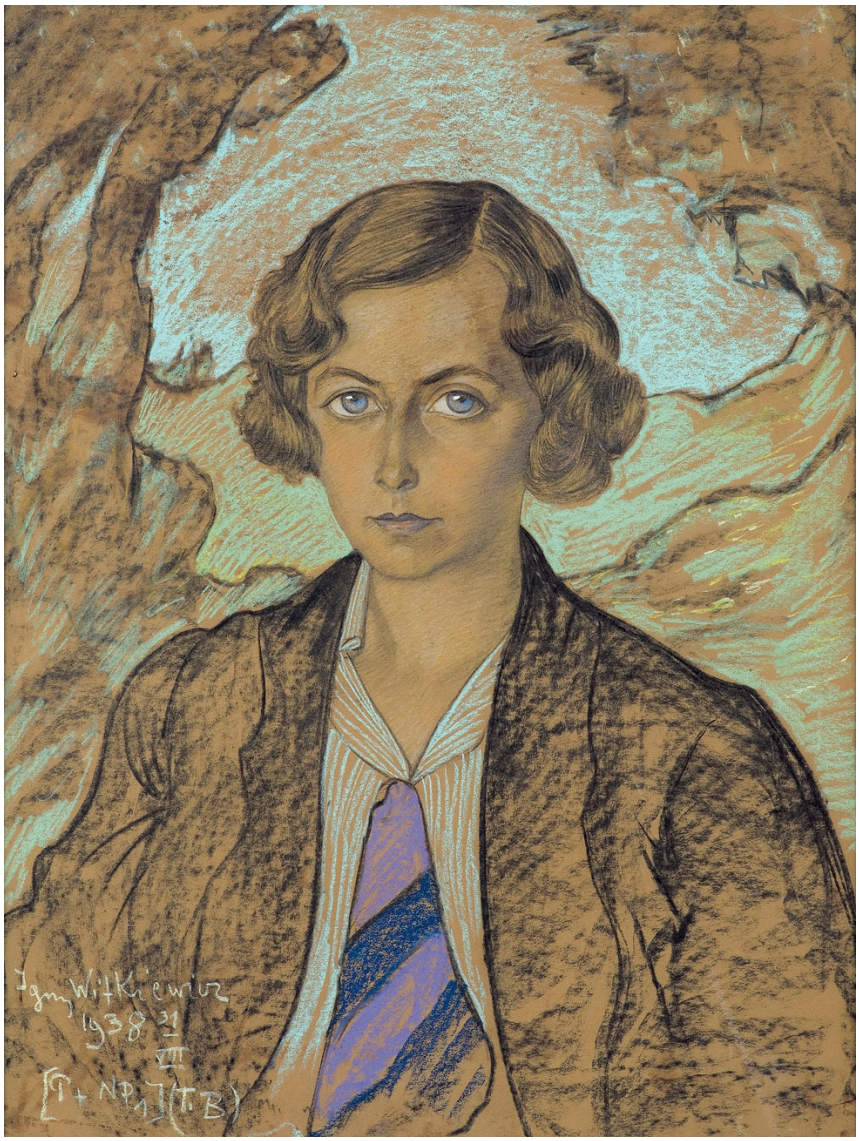


Figure 5.1: Stanisław Ignacy Witkiewicz, *Portret Ireny Kanafoskiej-Dembickiej*, 1938. Source: Agra-Art, `https://sztuka.agraart.pl/licytacja/357/23787` [accessed on October, the 17th, 2019]

To illustrate the different behaviour of colour quantisation algorithms, their quantisation errors and diversities were measured on the aforementioned Witkacy's portrait (Figure 5.1). Four algorithms were tested – median cut (see Section 3.3.4), octrees (Section 3.3.5), k-means, and Chang's k-means (both described in Section 3.3.3). The first three are popular quantisation algorithms, which are still used in many cases. The last one, introduced by Chang et al. (2015), is a modified k-means. Even though it was made for a different purpose (colour transfer), this algorithm is included in the test since it was designed to increase the diversity of the resulting colour palette. For the median cut and octree algorithms implementations Pillow[1], a popular fork of PIL (an abbreviation from Python Image Library), was used. Another Python library, Scikit-learn[2] provided a k-means implementation. For Chang's k-means, a JavaScript based implementation from Google was employed (available on GitHub[3]). All of the algorithms were set to discover 5 representative colours.

Table 5.1 presents the results of this comparison in terms of quantisation error and colour diversity. Figure 5.2 illustrates the results associated with a resulting palette of representatives, which are later presented in RGB colour space in Figure 5.3. In terms of quantisation error, there is no significant difference between the worst (Chang's k-means) and the best (median cut) algorithm – all of them provided similar quality. With regard to diversity, however, the best algorithm (Chang's k-means) scored two times better results than the last one (octree). In fact, Chang's version is better than the second-best algorithm (k-means) by 35%. This example provides preliminary evidence that Chang's k-means might strike a good balance between low quantisation error and high colour diversity.

Table 5.1: A comparison of the performance of quantisation algorithms for Figure 5.1.

algorithm	MSE	div
Median cut	68.24	45.40
Octree	70.60	35.47
k-means	71.74	52.85
Chang's k-means	75.86	72.10

A closer look at Figure 5.2 reveals that all algorithms except Chang's yielded similar palettes. The quality of the visual appearance of a quantised image differs, though. For the median cut, the resulting picture has good contours resembling the original shape, though the important purple

[1] `https://pillow.readthedocs.io` [accessed on October, the 17th, 2019]
[2] `https://scikit-learn.org` [accessed on October, the 17th, 2019]
[3] `https://github.com/googleartsculture/art-palette` [accessed on October, the 17th, 2019]

details are lost. As for the weakest algorithm in this comparison (octree), there is some noise introduced, perhaps due to the tree-based replacement phase and a low number of representatives. Supposedly, if the number of chosen representatives is higher, the algorithm would perform better compared to each other. Both k-means and Chang's k-means yielded sharp and visually pleasing results, though only the latter kept the purple colour instead of another shade of brown, which seems to be a small price to pay. A three-dimensional visual comparison of representatives in Figure 5.3 reveals that their position in RGB colour space differs significantly from the rest, which shows a similar pattern.

To check whether it is possible to generalise these findings for different paintings, a quantitative analysis was performed – a large-scale test on a dataset of 750 paintings from Polish auction houses. Similarly to the analysis in Figure 5.1, algorithms performance has been tested on MSE and diversity (both averaged this time). The set of colour quantisation algorithms is the same: median cut, octree, k-means (with k-means++ initialisation), and Chang's k-means. The test is repeated for different number of colours (6, 8, 10) and different size of input images (128×128, 256×256, 512×512). The images were rescaled to a given size (instead of, for instance, cropping).

The test dataset is a subset of the initial dataset of 88,435 artworks from Polish auction houses (see Chapter 6). The original dataset contains various types of lots sold at auction houses, such as paintings, graphics, or sculptures. Since this work focuses on paintings, the initial dataset was filtered by the used medium and technique. Considering only such pairs of these attributes with at least 50 observations narrowed the dataset down to 30,995 matching rows. There is no single standard definition of painting. It can be distinguished in terms of used medium and technique, however. The following subset is choosen: *oil on canvas*, *oil on cardboard*, *acrylic on canvas*, *oil on hardboard*, *oil on plywood*, *oil on fibreboard*, *oil on double canvases*, *mixed technique on canvas*, *oil and acrylic on canvas*, *oil on canvas on cardboard*, and *acrylic on fibreboard*. Now, the dataset consists of only paintings. In the next step, lots were grouped by year of the auction – the ones with less than 1000 observations were ruled out (leaving data from the period of 2008-2018). Then, the top 15 artists in terms of lots count were chosen (Edward Dwurnik, Eugeniusz Eibisch, Jacek Malczewski, Jan Szancenbach, Jerzy Kossak, Jerzy Nowosielski, Stanisław Kamocki, Tadeusz Dominik, Wiktor Korecki, Wlastimil Hofman, Wojciech Kossak, Wojciech Weiss, Wojciech Ćwiertniewicz, Włodzimierz Terlikowski). For each of them, 50 lots were drawn – resulting in 750 lots total.

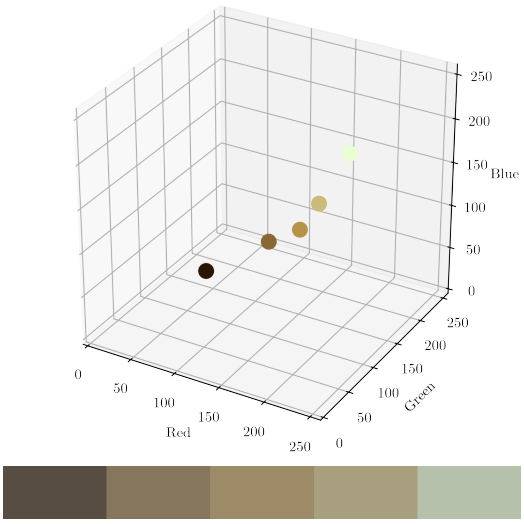
(a) Median cut

(b) Octree

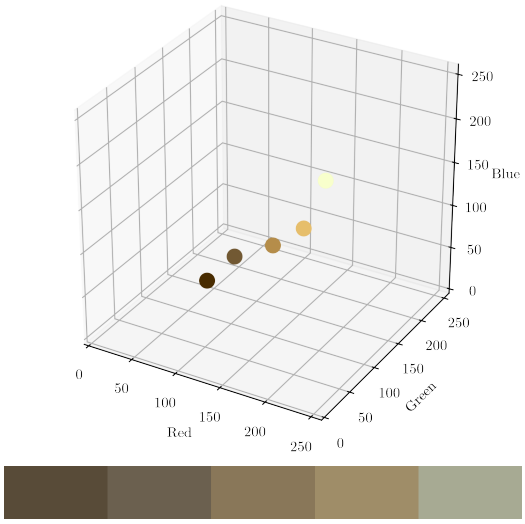
(c) k-means

(d) Chang's k-means

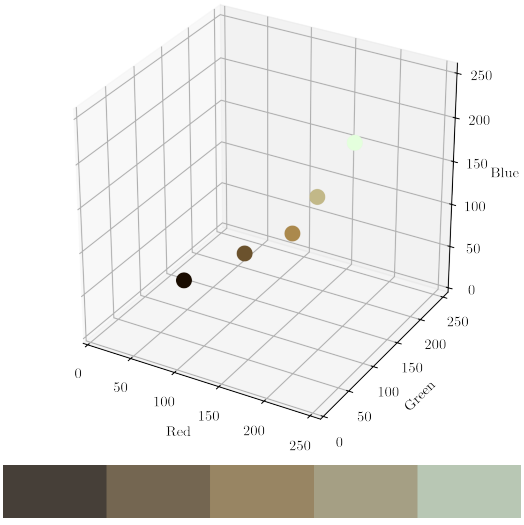
Figure 5.2: Witkacy's portrait from Figure 5.1 after colour quantisation along with the generated palettes.



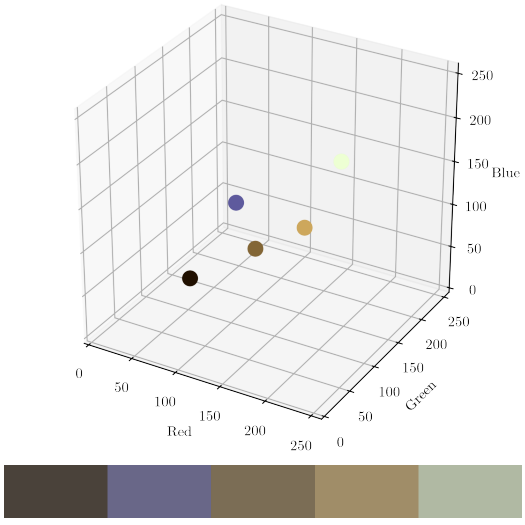
(a) Median cut



(b) Octree



(c) k-means



(d) Chang's k-means

Figure 5.3: Colour palettes presented in RGB colour space for Witkacy's portrait from Figure 5.1 after colour quantisation.

Figure 5.4 provides a selection of paintings and their corresponding results for quantisation algorithms and parameters. In many cases, median cut returned satisfying results which resemble the original paintings, though the returned colours may seem quite bland, especially compared with both variants of k-means. Results yielded by octrees were somehow noisy and grainy. Sometimes important parts of paintings were lost in the process of colour quantisation. This effect is particularly visible in Figure 5.4c – the train almost disappeared. Both k-means and Chang's k-means provided good results, the perceived difference often lies in small details – like the framed painting visible in Figure 5.4d.

Table 5.2 shows means and standard deviations for MSE and diversity calculated for the tested algorithms. In terms of average MSE, k-means won most of the time, followed closely by Median Cut – the difference was usually very small, no larger than 2 in general. Chang's k-means was third, whereas octree's performance was the worst. A possible explanation for the low score of octrees might be the specific way of choosing representatives, for which such a small number of colours might not be suitable. As intuitively expected, all the algorithms performed better with an increasing number of colours, but their MSE scores were slightly worse with increasing size (as there were more possibilities to make errors). Standard deviations were the lowest for octrees (followed by Chang's k-means).

Regarding average colour diversity, Chang's k-means yielded the highest results regardless of size or number of colours, which means that this algorithm yields the most diverse palette of representatives. At the same time, median cut provided the worst results in this category for most of the time, followed closely by octrees. The traditional k-means algorithm was somewhere in the middle. Interestingly, standard deviations were usually the lowest for octrees once again, although in 4 cases, Chang's k-means had the lowest values in this category. The highest standard deviations were yielded by the median cut algorithm.

The data provides convincing evidence in favour of the most of conclusions derived from the analysis of the behaviour of the algorithms for Figure 5.1. Namely, there's no single algorithm that tops these two metrics – all have their own trade-offs. In terms of MSE, k-means can provide the best results. While this algorithm performed the best in this context, median cut is often used in numerous practical applications due to its lower computational complexity, which makes it easy to apply on a low-end device and still provide fast results. That was especially the case in the past, where scarce resources were available, but there was a high demand for displaying reduced bitrate colours due to technological limitations back then. As for the colour

diversity, Chang's variant performs best since it was specifically designed to provide a wider palette of colours. Interestingly, it's not the worst one regarding MSE, which may be surprising since choosing *wrong* colours (in terms of MSE) will negatively impact this factor. Therefore, this algorithm might be considered as a balanced choice in terms of these two metrics, though the original k-means would be a good selection as well.

5.2 Feature Engineering

Feature engineering is understood here as a process of preparing the variables for analysis. This process is often crucial for successful data analysis. Regardless of the used algorithms and techniques, the explanatory power of models is limited by the used set of variables. This section sums up the process of shaping the available art market data for the experiments. For analysing price determinants, the target variable is the `price_final`, which is a numeric (float) value in the PLN currency representing the hammer price. With the same reasoning as with the hedonic regression, one can assume that a lot can be characterised by its qualities. Categorical variables, treated as one-hot encoded dummies are `auction_date_year`, `auction_house`, and `technique`. `Size` (in metres) is a numeric variable, which was derived as a geometric mean from the measurements provided by the auction house. The non-trivially derived features are 16 `Rx_Gx_Bx` variables (see Section 5.2.1) representing the share of each representative colour on a given lot, and its overall `colourfulness` which is discussed in Section 5.2.2.

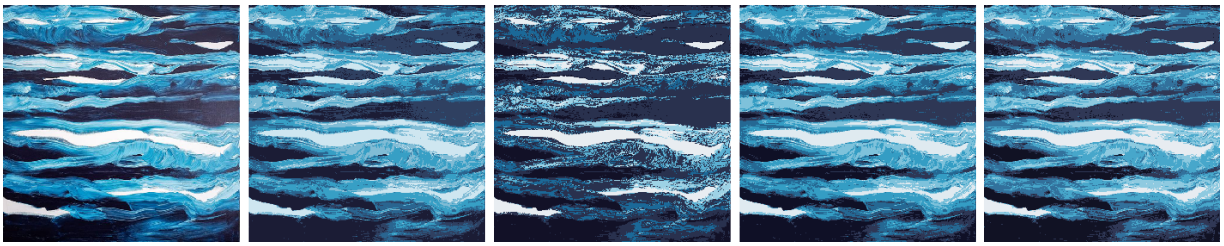
5.2.1 Representative Colours Share

The original set of variables available for every painting is extended by k new ones – each represents the share of the presence of particular representative colours. These colours have to be obtained first in the process of colour quantisation performed on the whole dataset (actually, only the search of representatives is performed – the actual quantisation takes place later). Due to the results obtained in the previous section, Chang's k-means (see 5 and 6 in Section 3.3.3) is the algorithm of choice, as it maintains the best colour diversity score among the tested algorithms with relatively low quantisation error.

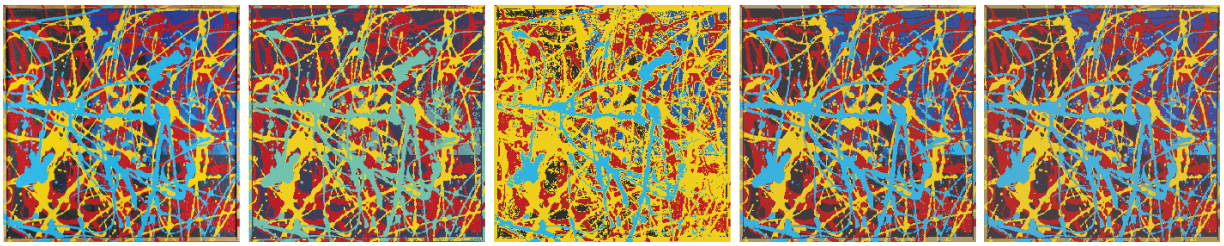
The process of obtaining the share of particular representative colours is presented in Algorithm 18. The set of all pictures is denoted as $\mathcal{P}$. The images of paintings are firstly merged into one large image $\mathbf{X}$ containing all of them. Since paintings differ in size, each image was

Table 5.2: A comparison of performance (mean MSE and mean diversity) of quantisation algorithms for the most popular 15 painters (750 paintings in total).

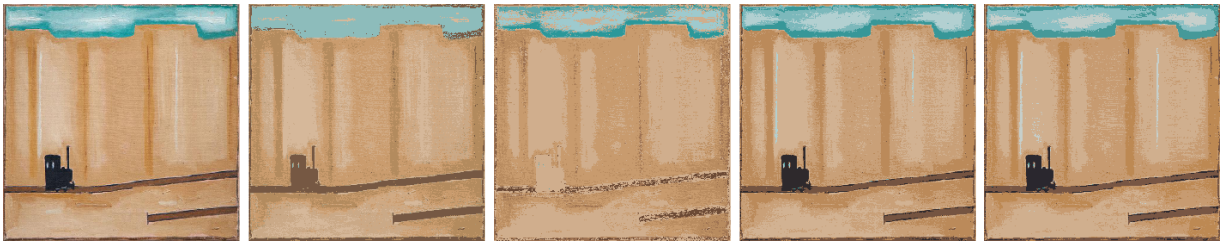
			MSE		diversity	
size	colours	algorithm	mean	σ	mean	σ
128	6	k-means	69.50	14.59	62.02	19.05
		Chang's k-means	75.45	13.26	66.25	18.90
		Median cut	69.59	15.32	55.17	21.44
		Octree	75.61	11.95	56.61	18.60
	8	k-means	63.02	15.43	62.10	17.94
		Chang's k-means	71.52	13.38	66.66	17.46
		Median cut	63.41	15.95	57.73	20.35
		Octree	72.33	11.67	59.07	16.66
	10	k-means	57.97	15.91	62.14	17.31
		Chang's k-means	68.71	13.44	67.36	16.26
		Median cut	59.40	15.70	56.20	19.46
		Octree	70.08	11.35	61.31	15.81
256	6	k-means	70.36	14.50	62.40	18.74
		Chang's k-means	75.83	13.33	67.57	17.86
		Median cut	70.44	15.20	55.97	21.51
		Octree	76.09	11.96	56.63	18.35
	8	k-means	64.06	15.32	62.69	17.85
		Chang's k-means	71.71	13.47	67.92	16.50
		Median cut	64.38	15.89	58.49	20.37
		Octree	72.89	11.76	59.17	16.45
	10	k-means	59.14	15.83	62.74	17.10
		Chang's k-means	68.44	13.72	68.32	15.58
		Median cut	60.49	15.72	57.02	19.50
		Octree	70.69	11.43	61.36	15.64
512	6	k-means	71.48	14.23	62.69	18.75
		Chang's k-means	76.64	13.01	68.32	17.37
		Median cut	71.30	15.09	56.46	21.69
		Octree	77.04	11.86	57.11	18.19
	8	k-means	65.31	15.17	62.84	17.71
		Chang's k-means	72.36	13.43	68.48	16.24
		Median cut	65.51	15.82	59.35	20.54
		Octree	73.90	11.66	59.29	16.41
	10	k-means	60.48	15.77	62.92	17.06
		Chang's k-means	69.23	13.40	69.08	15.25
		Median cut	61.63	15.63	57.69	19.65
		Octree	71.79	11.33	61.27	15.45



(a) Edward Dwurnik, *Ocean*, 2016 (512 pixels, 6 colours). Source: Sopocki Dom Aukcyjny, `http://www.sda.pl/ocean-2016,18760,en.html` [accessed on October, the 17th, 2019]



(b) Edward Dwurnik, *Obraz nr 229*, 2004 (512 pixels, 6 colours). Source: Rempex, `https://rempex.com.pl/wydarzenia/89-59-aukcja-sztuki-wspolczesnej/przedmioty/10052-obraz-nr-229-2004` [accessed on October, the 17th, 2019]



(c) Jerzy Nowosielski, *Pociąg*, 1972 (512 pixels, 8 colours). Source: Sopocki Dom Aukcyjny, `http://sda.pl/jerzy-nowosielski-pociag-1972,14108,pl.html` [accessed on October, the 17th, 2019]

(d) Jerzy Nowosielski, *Akt we wnętrzu*, 1965 (512 pixels, 8 colours). Source: Desa Unicum, `https://desa.pl/pl/wyniki/aukcja-sztuki-wspolczesnej-ip87/akt-we-wnetrzu-okolo-1965-r-7exz/` [accessed on October, the 17th, 2019]

(e) Wojciech Kossak, *Kirasjer 23 pułku na wedecie*, 1903 (512 pixels, 10 colours). Source: Rempex, `https://esensja.pl/varia/galeria/tekst.html?id=14964` [accessed on October, the 17th, 2019]

Figure 5.4: Colour quantisation algorithms comparison. Each row from the left: original image, median cut, octree, k-means, Chang's k-means.

scaled to the same size. This provides a fair influence on each image in the dataset. The size of shrunk images is controlled by n – each image is downsized to $n \times n$ pixels (in this work $n = 256$). Note that images containing framed paintings were manually altered – frames has been removed, as their presence would bias the results towards some specific colours. The palette of k representatives $\mathbf{C}$ is then obtained using Chang'sPaletteExtraction() – in this work, we use $s = 16$, as in the original paper for the quantisation algorithm. After that, each of the original but resized images $\mathbf{x}$ is quantised to $\mathbf{q}$ and the percentage share of particular colours $\mathbf{c}$ in $\mathbf{q}$ is calculated. The algorithm yields the share of each representative colour for every painting in the dataset in the form of a matrix $\mathbf{R}$, which has the size of $k \times \#\mathcal{P}$.

Algorithm 18 Calculation of the share of representative colours for paintings.

```
1: function QuantisePaintings(P, n, k, s)
2:     X ← ResizeAndMerge(P, n)
3:     C ← Chang'sPaletteExtraction(s, k, X)          ▷ see Algorithm 5
4:     for x in X do
5:         q ← Quantise(x, C)
6:         for c in C do
7:             R_{x,c} ← CountPixelsOfColour(q, c)/n
8:     return R
```

Figure 5.5 illustrates this process. A set of 16 paintings of different sizes were "squeezed" into one image in a way that each original painting occupies the same amount of space (Figure 5.5a). The process of search for 16 representatives with Chang's k-means algorithm resulted in a palette summarising the used colours (Figure 5.5b). Then, each of the original images is quantised using that palette (Figure 5.5c) and the share of each colour is calculated. The share of every colour is represented as a `Rx_Gx_Bx` variable, which denotes the colour coordinates in the RGB colour space.

5.2.2 Measuring Colourfulness

Colourfulness – a measure that tries to capture how colourful a given entity is – is a subjective concept. Notice that this notion is different from colour diversity, as a colour-diverse image might not be perceived as colourful. Hasler and Suesstrunk (2003) examined a number of image-related features in order to present a method for measuring the colourfulness of an image for natural scenes. They surveyed 20 participants in order to rate 84 images in terms of their colourfulness and provided three different metrics in order to estimate the colourfulness. The first two ($\hat{M}^{(1)}$

(a) Resized paintings combined into a single image in order to perform colour quantisation. Sources: Desa Unicum, Sopocki Dom Aukcyjny, Polswiss Art

(b) Representatives palette consisting of 16 colours

(c) Quantised images, using which one can calculate the share of each colour occupying a given image

Figure 5.5: Steps for calculating the share of representative colours.

and $\hat{M}^{(2)}$) are linear combinations of image features:

$$\begin{aligned} \hat{M}^{(1)} &= \sigma_{ab} + 0.37\mu_{ab}, \\ \hat{M}^{(2)} &= \sigma_{ab} + 0.94\mu_{C}, \end{aligned} \tag{5.2}$$

where σ_a and σ_b are the standard deviations of the a* and b* axes, which are used to calculate $\sigma_{ab} = \sqrt{\sigma_a^2 + \sigma_b^2}$. The centre of gravity in space to the neutral axis is denoted as μ_{ab}, whereas μ_C is the mean chroma value.

Hasler and Suesstrunk (2003) proposed the third metric, which is computationally fast, being correlated with around 95% of the conducted survey. It relies on features obtained form the sRGB colour space:

$$\begin{aligned} \mathbf{D}_{rg} &= |\mathbf{R} - \mathbf{G}|, \\ \mathbf{D}_{yb} &= \left| \frac{1}{2} (\mathbf{R} + \mathbf{G}) - \mathbf{B} \right|, \end{aligned} \tag{5.3}$$

where $\mathbf{R}$, $\mathbf{G}$, and $\mathbf{B}$ are the $m \times n$ matrices for consecutive RGB values of a given image. Then, for both resulting matrices, theirs means (μ_{rg} and μ_{yb}) and standard deviations (σ_{rg} and σ_{yb})

Table 5.3: Colourfulness attributes and corresponding metric (M) values.

Attribute	$\hat{M}^{(1)}$	$\hat{M}^{(2)}$	$\hat{M}^{(3)}$
not colourful	0	0	0
slightly colourful	6	8	15
moderately colourful	13	18	33
averagely colourful	19	25	45
quite colourful	24	32	59
highly colourful	32	43	82
extremely colourful	42	54	109

Source: Hasler and Suesstrunk (2003)

are used for the following calculations:

$$\begin{aligned} \sigma_{rgyb} &= \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}, \\ \mu_{rgyb} &= \sqrt{\mu_{rg}^2 + \mu_{yb}^2}. \end{aligned} \tag{5.4}$$

Finally, the colourfulness $\hat{M}^{(3)}$ is obtained from this equation:

$$\hat{M}^{(3)} = \sigma_{rgyb} + 0.3\mu_{rgyb}. \tag{5.5}$$

The corresponding values for all three metrics are presented in Table 5.3. The example can be found in Figure 5.6. A grayscale-like painting of Arkadiusz Mężyński (Figure 5.6a) is not colourful at all ($\hat{M}^{(3)} = 0$). On the contrary, Figure 5.6b presents a highly colourful painting by Kamil Jakóbczak ($\hat{M}^{(3)} = 162.841$). These values follow intuition. This particular formula ($\hat{M}^{(3)}$) is used to obtain new feature about a given painting in the dataset and is referred as `colourfulness`. Note that $\hat{M}^{(3)}$ is calculated using the original image of a painting – not the quantised one.

5.3 Summary

In this chapter, the following research objectives and questions have been a subject of analysis:

- the research question Q3 (*What is the best colour quantisation algorithm for paintings?*) through selecting the best colour quantisation algorithm for calculating colour share in paintings,
- the research objective O1 (*Prepare datasets allowing conducting the experiment*) through

(a) Arkadiusz Mężyński, *Złota*, oil on canvas, 2016 ($\hat{M}^{(3)} = 0$). Source: `sda.pl`

(b) Kamil Jakóbczak, *Akt*, acrylic on canvas, 2015 ($\hat{M}^{(3)} = 162.841$). Source: `desa.pl`

Figure 5.6: Example of low and high colourfulness in Polish emerging art.

preparing the art market dataset for answering Q3,

- the research objective O2 (*Develop a method for extracting colour-related features from paintings (Artefact 1)*) through delivering Algorithm 18 (Artefact 1) using two colour-related features.

Two used methods for feature engineering for paintings have been presented. The first one was focused on obtaining the representative colours for a painting dataset and calculating the share of every such colour. The second one tries to measure the concept of colourfulness. Both methods are used in the next chapter for extending the set of variables suitable for analysing price determinants for Polish paintings. To calculate the colour share, we chosen Chang's k-means algorithm, as it achieved a good balance between MSE and colour diversity (Section 5.1). It is not ideal in terms of the error, but it provides the best palette in terms of its diversification. Then, we presented the overall algorithm to leverage that palette and calculate the shares. For the colourfulness measure, the $\hat{M}^{(3)}$ formula provided by Hasler and Suesstrunk (2003) has been chosen, as its fast, reliable, and easy to interpret.

Chapter 6

Buyers' Preferences & Price Determinants for Polish Paintings

This chapter uses various quantitative methods and machine learning techniques in order to solve the research problem for this dissertation, which is determining the features which have an impact on artworks price with providing measures of that impact. Solving this problem is equivalent to answering the research question Q4 (*Which features of paintings are important for buyers on the Polish art market?*). The analysis is conducted using two datasets, which were collected in order to satisfy the research objective O1 (*Prepare datasets allowing conducting the experiment*). The first one consists of young art lots. The second one gathers artworks from the 10 most popular painters available in auction catalogues. Examining these datasets enables to fulfil the research objective O4 (*Discuss which features of paintings are important for buyer's preferences on the Polish art market*). In addition to checking the influence of features' characteristics that are typical for art market research, some colour-related features (created as described in Section 5) have been analysed. The idea is to check whether some specific colours might be a price-driving factor – just like Picasso and his famous Blue Period (Stepanova, 2015). To carry out the experiments, we use standard hedonic regression and the XGBoost algorithm, both paired with Artefact 1 (Algorithm 18). The resulting models are later analysed in terms of their performance and output, which fulfils research objectives O3 (*Evaluate the method for extracting colour-related features on Polish art market data*) and O5 (*Discuss price determinants for paintings on the Polish art market*) respectively.

6.1 Datasets Description and Preferences of Buyers

The market for paintings can be stratified in several ways, which was also mentioned in Section 2.1.2. For instance, the report provided by artinfo.pl (2019) uses three types: old art (pre-1945), contemporary art (post-1945), and young art (usually contemporary art made by often young debutants). For quantitative analysis, knowing who created a given painting (the `author` feature) is often crucial (see Chapter 2). However, in the case of young art, this information has low value, as such creators are unknown to buyers, and this information has little or no impact on the price. At the same time, using this information requires either a categorical variable with many categories or substantially narrowing down the set of analysed artists based on some criterion. The first option makes the analysis harder due to the lack of generalisation (especially with the linear models), whereas the criterion for the second option is unclear. In such cases, omitting the information about the creator provides an intuitive distinction from pre- and post-1945 art made by well-established names. One could further divide the latter. However, only four artists had more than 100 sold lots in such datasets. Splitting it further would result in underrepresented artists in the post-1945 dataset. Therefore, these two categories together has been treated as one. Future studies might consider splitting them, provided that more observations would be available. To sum up, two primary datasets are examined in this study – the *Young Art* dataset and the *Top 10 Painters* dataset.

Before describing these datasets in detail, we present their common characteristics. Both have been made from data publicly available at auction houses websites in the middle of 2018. Note that the observations from the last year are therefore incomplete. They were collected from the five main auction houses and resulted in 88,435 observations, which consisted of sold lots and bought-ins. The oldest observations date back to 2008. Since this study focuses on paintings, the dataset was narrowed down to 30,955 lots. By paintings, we mean lots made with acrylic, oil, or mixed paints a material such as a canvas, board, hardwood, paperboard, or plywood.

The hammer price is represented by `price_final` feature. It will be used as the explained/-target variable in the following sections. Nominal price values are in Polish złotys (PLN). Since the lots were sold in different years, these prices are subject to inflation/deflation. Therefore, they were adjusted to 2018 prices with the `priceR` package in R using World Bank inflation data. In quantitative art market research, a common transformation considers the natural logarithm

of the original price. Therefore, `ln_price_final` was used:

$$\texttt{ln\_price\_final} = \ln(\texttt{price\_final}).$$

Usually, auction houses provide the artwork dimensions in the $x \times y$ format. Occasionally, a lot depth was also provided – this value was ignored, as only the *flat* paintings are considered in this study. In this research, the geometric mean of the first two values is considered:

$$\texttt{size} = \sqrt{\texttt{size\_x} \cdot \texttt{size\_y}}.$$

The size of the sold lot is presented as `size` in cm. The place of sale is represented by the `auction_house` categorical variable. It includes the five biggest auction houses in Poland in the considered years: Desa Unicum, Rempex, Agra Art, Polswiss Art, and Sopocki Dom Aukcyjny. Similarly, `auction_date_year` is a categorical variable representing the year of sale. The next attribute is `technique`, which describes the combination of used technique and medium for a given lot. The common paints include acrylic, oil, and combinations of them. Regarding the medium, the canvas is a popular choice for paintings. In order to exclude underrepresented combinations of used materials and medium, we set a threshold of at least 50 observations per technique in the further analysis.

Excluding the transformations to the size and hammer price-related attributes, the aforementioned variables were explicitly obtained from the auction catalogues. The rest consist of features derived from painting visual appearance in the process of feature engineering described in Section 5.2. These variables considers a set of attributes named `Rx_Gy_Bz`, where `x`, `y`, and `z` denotes the coordinates in the RGB colour space (Algorithm 18). The last engineered variable is `colourfulness`, which is calculated using the $\hat{M}^{(3)}$ formula provided by Hasler and Suesstrunk (2003) and described in Section 5.2.2, Equation (5.5). Feature extraction for the colour-related variables was conducted on sold/not sold datasets. The further analyses include only sold lots, which were substantially smaller. Techniques with less than 50 observations were removed from the datasets, as well as lots with missing data.

6.1.1 Young Art

The first dataset considers paintings from young art auctions. It was obtained by searching for "młoda" (Polish for young) in the auction title (including inflexion) in the dataset of 30,955

Figure 6.1: Palette of 16 representatives generated for the young art dataset.

lots mentioned in the previous section. This resulted in 5,292 available observations, of which 4,657 (86%) are sold lots. The latter subset with the only sold lots will be later referred to as the Young Art dataset. In art market research, often the most of the explained variance comes from the author-related variables. Therefore, this was an incentive to prepare the dataset, which consists of paintings made by painters that are rather not known to the general audience and therefore diminish the effect of the painter name (to check whether other features are sufficiently explaining the variance). Some of the presented statistics mention `price_final` variable, but in the experiments it was omitted in favour of `ln_price_final`. Figure A1 in Appendix shows the difference in the distribution of price before and after taking the logarithm.

The colour palette for all paintings was generated using Chang's k-means algorithm, where $k = 16$. Each picture was scaled to 256 $\times$ 256 pixel – such value was chosen, since some of the pictures were not larger than this. The resulting palette is presented in Figure 6.1. The palette provides some vivid colours, such as ■ `R191_G108_B127` – they are not present with in the Top 10 Painters dataset (see the next subsection). Low prices are the key components of the young art auctions – they often start at 500 PLN. Three auction houses offered young art at auction in this dataset: Desa Unicum (since 2011), Sopocki Dom Aukcyjny (since 2014), and Polswiss (only in 2014) in the dataset. Indeed, the distribution of sold lots shows that 2014 was the year in which the trend of selling more young art at auctions has started. All the works were prepared on canvas regarding the technique, and they are almost evenly split between acrylic and oil paints. The colourfulness of the Young Art dataset can be characterised by right-skewed single-peaked distribution, averaging at moderately colourful in the original scale. Tables 6.1, 6.2, 6.3, and 6.4 present the detailed descriptive statistics about the Young Art dataset. No significant (i.e. higher than 0.8 or lower than -0.8) correlations between the variables have been found. The correlation matrix is presented in Figure A3 in the appendix. The distributions of years, size, and colourfulness are presented in Figure A2 in the appendix.

Table 6.1: Summary statistics for numerical variables in the Young Art dataset.

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
price_final	1,949.651	1,837.127	500.000	814.500	2,477.700	21,679
ln_price_final	7.274	0.742	6.215	6.703	7.815	9.984
size	88.447	24.252	15.000	72.100	101.600	200.000
colourfulness	49.755	29.402	0.000	26.576	68.415	170.624
R30_G27_B27	0.116	0.157	0.000	0.008	0.159	0.963
R52_G58_B132	0.029	0.068	0.000	0.00001	0.026	0.882
R56_G62_B70	0.101	0.115	0.000	0.020	0.144	0.882
R72_G120_B168	0.048	0.091	0.000	0.00001	0.055	0.797
R96_G148_B85	0.013	0.048	0.000	0.000	0.005	0.914
R105_G84_B58	0.049	0.070	0.000	0.005	0.064	0.771
R107_G114_B116	0.086	0.109	0.000	0.015	0.115	0.992
R122_G46_B38	0.027	0.058	0.000	0.00005	0.029	0.819
R122_G179_B202	0.051	0.107	0.000	0.00002	0.050	0.959
R168_G168_B167	0.109	0.137	0.000	0.017	0.147	0.998
R191_G108_B127	0.023	0.044	0.000	0.001	0.027	0.718
R192_G129_B57	0.034	0.066	0.000	0.0001	0.037	0.831
R200_G176_B137	0.056	0.089	0.000	0.002	0.070	0.790
R208_G57_B36	0.025	0.072	0.000	0.000	0.015	0.951
R214_G192_B55	0.022	0.069	0.000	0.000	0.010	0.955
R214_G216_B213	0.210	0.237	0.000	0.024	0.328	0.999

Table 6.2: Price statistics for auction houses in the Young Art dataset.

auction_house	N	Mean	St..Dev.	Min	Max
Desa Unicum	3,391	2,179.537	1,942.883	500.000	19,614.770
Polswiss	231	1,203.634	812.130	511.666	8,186.659
Sopocki Dom Aukcyjny	1,035	1,362.970	1,425.758	500.000	21,679.490

Table 6.3: Price statistics for years in the Young Art dataset.

auction_date_year	N	Mean	St..Dev.	Min	Max
2011	173	1,583.381	1,237.643	535.428	10,708.560
2012	198	1,822.887	1,266.574	517.020	6,204.240
2013	194	1,972.558	1,442.334	511.942	9,214.949
2014	521	1,734.755	1,722.464	511.666	17,396.650
2015	784	2,149.514	2,314.814	516.178	21,679.490
2016	1,200	1,948.752	1,844.850	519.633	15,588.980
2017	1,112	2,008.059	1,815.916	509.065	14,253.810
2018*	475	1,897.895	1,566.742	500.000	12,000.000

Table 6.4: Price statistics for techniques in the Young Art dataset.

technique	N	Mean	St..Dev.	Min	Max
acrylic, canvas	2,059	1,980.464	1,750.058	500	13,420.640
oil, acrylic, canvas	190	1,521.958	1,291.515	500	11,256.660
oil, canvas	2,048	1,888.408	1,939.224	500	21,679.490
mixed, canvas	360	2,347.546	1,900.410	500	12,471.180

6.1.2 Top 10 Painters

The second dataset entails the top 10 painters, in a sense of the highest number of observations in the original dataset of 30,995 paintings mentioned at the beginning of this section. Excluding the representative colours obtained in the process of quantisation, the dataset has almost the same variables set as the young art dataset. There is one additional column – `author`, which represents the categorical variable indicating the creator of the given painting. These painters are: Alfons Karpiski (184 paintings, 80 sold), Edward Dwurnik (179/121), Jacek Malczewski (209/102), Jerzy Kossak (317/194), Jerzy Nowosielski (192/145), Wiktor Korecki (180/97), Wlastimil Hofman (325/153), Wodzimierz Terlikowski (160/77), Wojciech Kossak (226/117), and Wojciech Weiss (242/79). Most of these artists lived between 19 and 20th centuries, except Jerzy Nowosielski (1923-2011) and Edward Dwurnik (1943-2018). The dataset contains 2,214 samples in total, of which 1,056 are sold (with no missing data). Similarly to the young art dataset, it is called the Top 10 Painters dataset in the rest of the dissertation. The descriptive statistics about the Top 10 Painters dataset are presented in tables 6.5, 6.6, 6.7, 6.8, and 6.9.

The prices in the Top 10 Painters dataset are much higher since the sold lots are made by established and well-known artists – the mean is almost 54,000 PLN, whereas the highest price in the dataset is 1,703,388 PLN. Similarly to the Young Art dataset, the original `price_final` (adjusted to inflation in 2018) is present in some statistics, though in the later experiments only `ln_price_final` is used. The distribution of the first is skewed right with a long tail. After taking the natural logarithm, the distribution is almost single-modal and close to symmetric, with a mild right tail. Both distributions are presented in Figures A4a and A4b respectively. The analysed lots were sold between 2008 and 2018. The number of sold lots started to grow rapidly in 2014. As in the Young Art dataset, for 2018 the data consist of only some of the lots sold in the first half of the year. Given that numerous auction houses have their most popular auctions in December, this might bias the results for this year. The distributions for the years,

size, and colourfulness are presented in Figure A5. The distributions for price per author are depicted in Figure A6 in the appendix.

The average size (as a single dimension) of artwork in the dataset is 62.505 cm, whereas the mean colourfulness equals 39.136. This is considerably lower than in the Young Art dataset, which is also reflected in the generated set of representatives. As with the Young Art dataset, the colour palette was generated with Chang's k-means algorithm with $k = 16$ and pictures scaled to 16×16. A palette using $k = 8$ has been generated as well, but this number turns out to be too small to cover some primary colours (such as green or yellow). Therefore, $k = 16$ was chosen. Both palettes are presented in Figure 6.2. The colours seem darker, compared to the much more vivid palette obtained in the Young Art dataset (Figure 6.1) – for instance, there is no representation for the pink colour (R191_G108_B127). The distributions for colourfulness per author are presented in Figure A7 in the appendix. For most of the painters, the distributions are normal-like and symmetric averaging at moderate colourfulness, though Edward Dwurnik stands out from the crowd since his works display a whole range of different levels of colourfulness. Among all the variables, no significant correlations were found. The correlation matrix for the Top 10 Painters dataset is presented in Figure A8 in the appendix.

(a) 8 representatives

(b) 16 representatives

Figure 6.2: Palettes of 8 and 16 representatives generated for the Top 10 Painters painters dataset.

6.2 Price Determinants

This section shows the usage of linear regression and XGBoost paired with the presented datasets and explainable artificial methods in order to find the price determinants for the Polish art market. As explained in Section 2, the majority of art market research uses hedonic and repeated-sales regression. Since the available datasets span for a limited number of years and the focus is on finding the price determinants, the OLS-based hedonic regression was chosen to be the

Table 6.5: Summary statistics for numerical variables in the Top 10 Painters dataset.

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
price_final	53,967.440	111,861.100	1,228	9,873.0	47,888.8	1,703,388
ln_price_final	10.070	1.185	7.113	9.198	10.777	14.348
size	62.505	30.012	12.649	41.735	77.356	258.023
colourfulness	39.136	17.844	4.051	27.302	47.172	130.132
R40_G36_B35	0.087	0.110	0.000	0.016	0.111	0.902
R49_G75_B127	0.020	0.061	0.000	0.000	0.008	0.587
R71_G67_B57	0.109	0.095	0.000	0.038	0.153	0.722
R86_G102_B106	0.048	0.056	0.000	0.010	0.066	0.603
R91_G142_B181	0.027	0.058	0.000	0.000	0.023	0.500
R101_G125_B81	0.030	0.047	0.000	0.002	0.042	0.530
R111_G72_B50	0.078	0.085	0.000	0.018	0.110	0.693
R127_G115_B101	0.106	0.087	0.000	0.039	0.151	0.556
R151_G175_B185	0.054	0.083	0.000	0.001	0.077	0.597
R171_G115_B69	0.049	0.068	0.000	0.008	0.062	0.789
R173_G170_B158	0.113	0.107	0.000	0.034	0.163	0.653
R181_G161_B92	0.028	0.042	0.000	0.003	0.036	0.478
R185_G63_B49	0.023	0.084	0.000	0.000	0.007	0.972
R185_G156_B123	0.083	0.079	0.000	0.025	0.119	0.727
R202_G182_B59	0.008	0.031	0.000	0.000	0.002	0.382
R212_G211_B203	0.138	0.162	0.000	0.017	0.197	0.942

Table 6.6: Price statistics for auction houses in the Top 10 Painters dataset.

auction_house	N	Mean	St..Dev.	Min	Max
AgraArt	241	46,367.730	89,834.690	1,227.879	779,448.900
Desa Unicum	360	63,582.500	93,995.140	1,247.118	733,053.300
Polswiss	82	153,196.700	282,053.600	4,884.547	1,703,388.000
Rempex	274	28,328.910	40,538.510	1,766.751	322,172.200
Sopocki Dom Aukcyjny	99	26,273.190	30,181.750	2,200.000	157,810.100

Table 6.7: Price statistics for years in the Top 10 Painters dataset.

auction_date_year	N	Mean	St..Dev.	Min	Max
2008	63	93,713.870	127,724.900	5,348.340	618,030.500
2009	52	58,039.160	97,999.240	4,122.218	492,376.000
2010	67	31,928.180	44,036.370	1,227.879	301,388.400
2011	67	29,475.700	31,766.480	3,212.567	171,336.900
2012	65	40,061.890	57,739.650	4,136.160	258,510.000
2013	77	42,940.600	60,965.780	3,481.203	368,598.000
2014	95	52,295.520	114,640.900	2,967.664	1,002,866.000
2015	135	75,899.620	171,587.400	2,787.363	1,703,388.000
2016	203	51,363.250	88,739.510	1,247.118	779,448.900
2017	177	53,345.960	120,096.200	2,443.511	1,119,942.000
2018	55	53,810.910	163,181.500	2,200.000	1,200,000.000

Table 6.8: Price statistics for techniques in the Top 10 Painters dataset.

technique	N	Mean	St..Dev.	Min	Max
oil, canvas	576	65,001.130	116,017.800	1,662.824	1,703,388.000
oil, board	52	40,331.540	70,157.660	1,227.879	332,564.900
oil, plywood	60	29,266.960	71,857.370	4,283.422	492,376.000
oil, paperboard	368	42,651.380	113,674.500	1,247.118	1,119,942.000

Table 6.9: Price statistics for author in the Top 10 Painters dataset.

author	N	Mean	St..Dev.	Min	Max
Alfons Karpiski	77	22,826.900	17,358.390	7,238.280	98,233.790
Edward Dwurnik	87	21,749.240	20,560.620	1,662.824	104,965.800
Jacek Malczewski	91	223,057.600	291,229.300	5,152.773	1,703,388.000
Jerzy Kossak	187	16,321.740	13,929.260	3,767.079	77,944.890
Jerzy Nowosielski	131	131,716.100	84,725.100	5,247.193	495,531.100
Wiktor Korecki	95	6,600.749	3,843.173	1,227.879	22,581.880
Wlastimil Hofman	140	26,493.030	23,051.250	2,967.664	144,765.600
Wodzimierz Terlikowski	74	22,572.150	22,278.500	4,072.518	134,206.300
Wojciech Kossak	100	40,030.570	34,201.620	5,090.648	171,336.900
Wojciech Weiss	74	36,824.920	27,539.990	7,126.907	148,659.300

baseline method in this study. However, HR models do not necessarily correctly capture the relationship between the explained and explanatory variables, which in general might not be linear. Therefore, the XGBoost algorithm (introduced in Section 4.1) is used, which is proven to be the state-of-the-art method for numerous machine learning tasks.

Albeit the main usage of XGBoost is the prediction in classification and regression tasks, it can be paired with the explainable artificial intelligence methods. It is also expected to generalise the relationship between the target variable and available features better since this algorithm can capture non-linear relationships, contrary to hedonic regression. Nevertheless, these two can be compared using the methods from explainable artificial intelligence. Following Biecek and Burzykowski (2020), this enables us to examine feature importance, their influence on the final prediction, or general performance of the models. This also enables to validate the agreement between the models, which can be reassuring about the way they capture important qualities of considered artworks.

Both models are evaluated against several metrics – mean squared error, root mean square error, coefficient of determination, and mean average deviation. Mean squared error (MSE) is calculated as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $y_i - \hat{y}_i$ is the difference between the true and predicted value for i-th out of n observation. Root mean square error is obtained simply by taking the square root in order to use the same units as the target variable:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}.$$

The coefficient of determination was already introduced in Section 2.2:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\mathrm{Var}(\hat{\mathbf{y}})}{\mathrm{Var}(\mathbf{y})},$$

where $\bar{y}$ is the mean observation value. Additionally, the adjusted R^2 is calculated for linear models in order to diminish the effect of increasing R^2 when adding new independent variables. These new variables are not necessarily useful, and this is now reflected by the metrics:

$$\text{Adjusted } \mathrm{R}^2 = 1 - \left(\frac{(1-\mathrm{R}^2)(n-1)}{n-k-1}\right),$$

where n is the number of observations and k is the number of independent variables. The last

used metrics is mean absolute deviation (MAD), calculated as follows:

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - \bar{y}| \,.$$

Since the main goal of this work focuses on explanatory analysis instead of prediction, the usual train/test split was not used. Instead, all the available data samples were used for training and the metrics. While this is not the case for linear regression, tree-based models are notorious for their tendency to overfitting. While the XGBoost algorithm provides some internal mechanism to combat this problem, the k-cross-validation procedure was used to ensure that the model provides a good degree of generalisation. It is a resampling process in which the dataset is partitioned to k even subsets. Then, the k different folds are created – in each, one of the subsets is the validation set, whereas the rest is used for the training. We set $k = 10$, which is a standard value used in numerous research. Each fold was used then in the hyperparameter tuning process, in which every possible set of 432 combinations have been tested, which paired with 10-fold CV results in 4320 models to be tested. The combination of hyper-parameters which has the lowest MSE for every fold on average, is later chosen as the final one and used to train the model on a full dataset. The set of values for hyper-parameters used is presented in the following list: `eta` $= \{0.01, 0.05, 0.05, 0.1\}$, `max_depth` $= \{4, 6, 8\}$, `gamma` $= \{0.2, 0, 0.1\}$, `colsample_bytree` $= \{1, 0.5\}$, `min_child_weight` $= \{1, 2\}$, and `subsample` $= \{0.7, 1\}$. Such a number of possible hyper-parameter combinations was enough to find a good fit for both datasets and noticeably improve MSE compared to the linear models, which is presented in the following subsections.

6.2.1 Young Art Dataset

Two linear models were created using the hedonic regression, with parameters estimated by OLS. The first one, which uses the set of variables obtained directly from the auction catalogue, and the second one, which also uses the colour-related variables obtained in the feature engineering process described in Section 5.2 – `colourfulness` and 16 variables representing the share of colours. To avoid singularities and multicollinearity, the following dummy variables were removed from regression analysis: `auction_date_year 2011`, `auction_house Desa Unicum`, `technique acrylic canvas`, and `R214_G216_B213`. Table 6.10 compares both of the linear models. The model without colour-related variables reached $R^2 = 0.137$ and adjusted $R^2 = 0.135$. After adding colourfulness and the share of representatives, the second linear model got $R^2 = 0.150$

and adjusted $R^2 = 0.144$. Unless stated otherwise, the second model (the one with colour features) is referred to simply as linear model in this section. Adding the colour-related variables enabled to explain additional 1% of the variance. However, all these values still suggest a poor fit – the linear relation fails to capture the relation between the price and considered variables. Regression diagnostics on residual vs fitted plot showed that residuals have a linear-like pattern (Figure A9). While rigorous Shapiro-Wilk test ($W = 0.98625$, $p < 2.2e-16$) and Kolmogorov-Smirnov test ($D = 0.046865$, $p = 2.611e-09$) both rejected their null hypotheses (that is, the normality of model residuals) at $\alpha = 0.05$, the visual inspection of normal quantile-quantile plot (Figure A10) reveals a light tail and was consistent with approximate normality in terms of residuals. A Scale-location plot has been used to assess homoscedasticity (Figure A11). While the spread of residuals increased slightly after 2000 PLN, homoscedasticity assumptions were not severely violated. Generalised variance inflation factors (abbreviated as GVIF) (Fox & Monette, 1992) showed no severe multicollinearity. That is, for every inspected dependant variable, $\text{GVIF}^{1/(2\text{DF})} < \sqrt{5}$, as suggested Buteikis (2018).

The most important variable in terms of the statistical significance was `size` (average hammer price increase by $100 \cdot (e^{0.007} - 1) = 0.7\%$ per every unit), which does not come as a surprise, as this is in line with other results in quantitative art market research. Both included auction houses (Polswiss and SDA) turned out to be statistically significant ($p < 0.01$) with a negative impact on the final price, which means that works of Young Art sold at Desa Unicum are associated with higher prices. Technique-related dummies were statistically insignificant, except the oil on canvas, which surprisingly had a negative impact on the price. Surprisingly, the hammer price diminishes with the colourfulness, which was not significant in a statistical sense. Regarding the colour variables, the ones which were statistically significant are ■ `R30_G27_B27` ($e^{-0.228}-1 = -0.2\%$ price drop per 1 percentage point increase of this colour on average according to the second model), ■ `R56_G62_B70` (-0.25% per 1pp increase), ■ `R96_G148_B85` (-0.483% per 1pp increase), ■ `R191_G108_B127` (-0.4% per 1pp increase), ■ `R192_G129_B57` ($+0.54\%$ per 1pp increase).

Three of the year-related explanatory variables turned out to be statistically significant (at most at 0.1). Figure 6.3 presents the index with a base equal to 1 build from values in Table 6.11. The index was built from the standard formula:

$$\text{Index} = e^{\gamma t}, \tag{6.1}$$

Table 6.10: Regression results for the Young Art dataset.

	Dependent variable: ln(price_final)	
	Model 1	Model 2
auction_date_year 2012	0.038 (0.072)	0.026 (0.072)
auction_date_year 2013	0.104 (0.072)	0.096 (0.072)
auction_date_year 2014	0.156** (0.067)	0.148** (0.067)
auction_date_year 2015	0.242*** (0.059)	0.233*** (0.059)
auction_date_year 2016	0.111* (0.057)	0.097* (0.057)
auction_date_year 2017	0.098* (0.057)	0.081 (0.058)
auction_date_year 2018	0.100 (0.062)	0.088 (0.062)
auction_house Polswiss	−0.441*** (0.061)	−0.417*** (0.061)
auction_house Sopocki Dom Aukcyjny	−0.441*** (0.026)	−0.440*** (0.026)
technique oil, acrylic, canvas	−0.064 (0.053)	−0.071 (0.053)
technique oil, canvas	−0.046** (0.022)	−0.043** (0.022)
technique mixed, canvas	0.112*** (0.040)	0.075* (0.040)
size	0.007*** (0.000)	0.007*** (0.000)
colourfulness		−0.001 (0.001)
R30_G27_B27		−0.228*** (0.077)
R52_G58_B132		−0.162 (0.174)
R56_G62_B70		−0.299*** (0.102)
R72_G120_B168		−0.084 (0.147)
R96_G148_B85		−0.659*** (0.220)
R105_G84_B58		0.151 (0.165)
R107_G114_B116		0.094 (0.106)
R122_G46_B38		0.166 (0.198)
R122_G179_B202		−0.073 (0.111)
R168_G168_B167		0.013 (0.092)
R191_G108_B127		−0.512** (0.249)
R192_G129_B57		0.432** (0.183)
R200_G176_B137		−0.199 (0.130)
R208_G57_B36		0.042 (0.197)
R214_G192_B55		−0.277 (0.191)
Constant	6.652*** (0.063)	6.780*** (0.072)
Observations	4,657	4,657
R^2	0.137	0.150
Adjusted R^2	0.135	0.144
Residual Std. Error	0.690 (df = 4643)	0.687 (df = 4627)
F Statistic	56.863*** (df = 13; 4643)	28.048*** (df = 29; 4627)
Note:		*p<0.1; **p<0.05; ***p<0.01

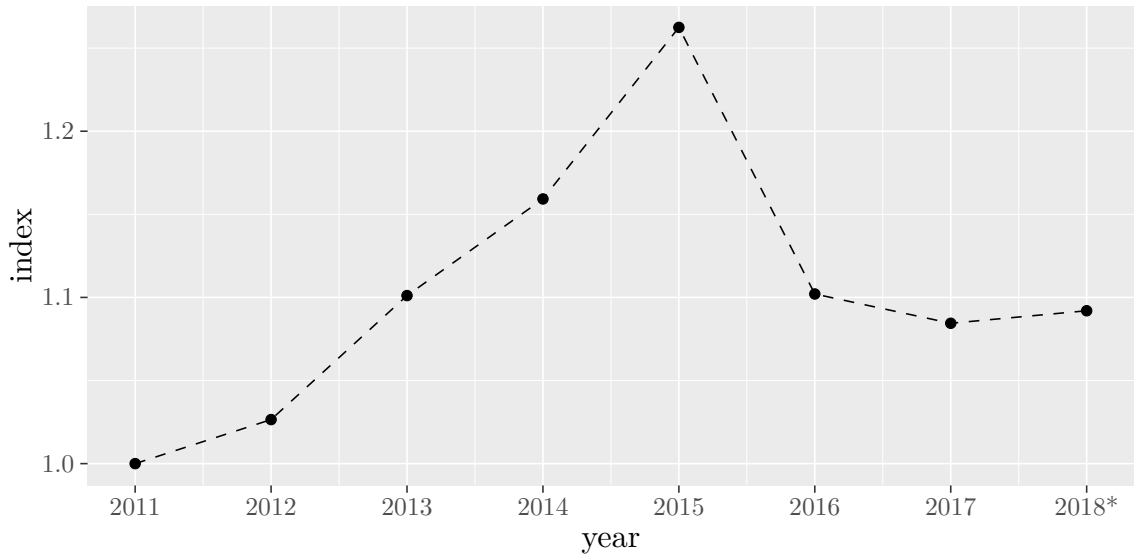


Figure 6.3: Art market index for the Young Art dataset.

where γ_t is the HR Model 2 coefficient value obtained for each year and $\gamma_{2011} = 0$. From 2011, the index for young artists was on the rise up to 2015. In 2016, a correction had occurred, which brings it down to levels from 2013. Up to 2018, The index remained on a relatively stable trajectory since this date.

However, the linear model with such low R^2 and high residual standard error cannot be considered as reliable explanatory tool. Therefore, an XGBoost model has been prepared. The best set of parameters obtained with 10-fold cross validation in terms of RMSE was `nrounds` = 1000, `max_depth` = 8, `eta` = 0.01, `gamma` = 0, `colsample_bytree` = 0.5, `min_child_weight` = 1, `subsample` = 1. Trained on the same data as the linear model, the XGBoost model nearly doubled its performance in terms of RMSE. Detailed model performance for linear regression and XGBoost is presented in Table 6.11. Figure 6.4 depicts boxplots of absolute values of residuals in the datasets. The red dot stands for the root mean square of residuals, which is considerably lower for the linear regression. The XGBoost residual distribution is normal-like, with a much narrower inter-quartile range than the linear one.

Table 6.11: Model performance comparison for the Young Art dataset.

Model	MSE	RMSE	R^2	MAD
Linear regression	0.469	0.685	0.150	0.508
XGBoost	0.106	0.325	0.808	0.209

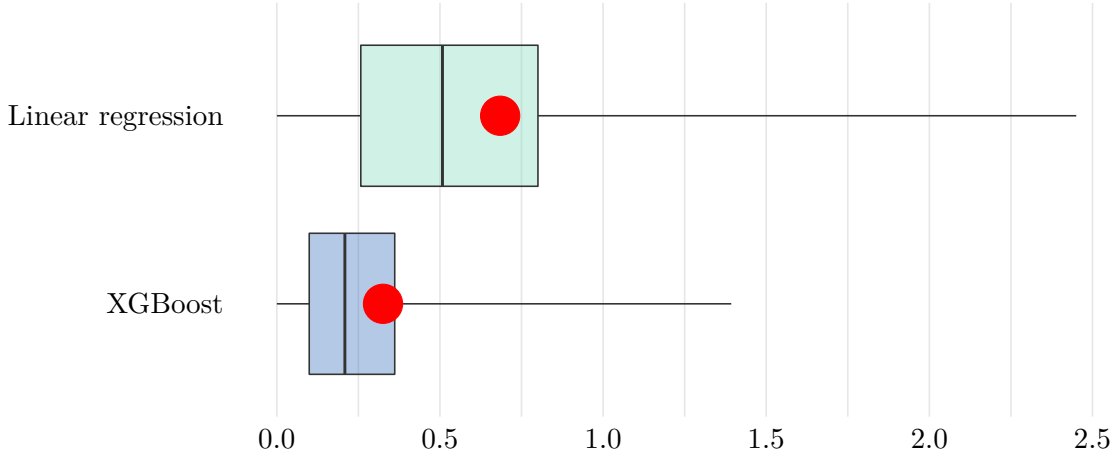


Figure 6.4: Boxplots of absolute values of residuals in the Young Art models.

Figure 6.5 presents feature importances comparison for both of the models after 50 permutations. For the linear model, dropping `auction_house`, `size`, and `auction_date_year` resulted in the highest RMSE increase. The rest of the variables seem to be of marginal importance. This was not the case for the XGBoost model. The aforementioned variables – along with `colourfulness` – are connected to the highest RMSE increase. Regarding the colour variables, all the statistically significant ones are of the highest importance in the XGBoost model (except one colour): R192_G129_B57, `R30_G27_B27`, `R168_G168_B167`, R56_G62_B70, `R96_G148_B85`, and `R191_G108_B127`. However, all the variables seem to be important for the XGBoost model, as even without the least important one, the change in the RMSE is still visible. The exact values for mean RMSE dropout losses are given in Table A.1 in Appendix.

To investigate it further, we analyse partial dependence plots for selected variables. In both models, `size` can be considered as the most important variable. However, the XGBoost model captured its relation to the hammer price as linear-like only for a certain interval (ranging roughly between 60 and 150). In a broader context, this relation seems to have more of a sigmoidal shape. The linear model suggested a slight decrease in the hammer price with an increase in colourfulness. This is somehow confirmed by the XGBoost, where the least colourful lots carry the highest value – although for the very work with the highest colourfulness, a price increase can be observed. Figure 6.6 shows both of these relations. Figure 6.7 presents partial-dependence profiles for colour representatives. Numerous discrepancies between linear and XGBoost models

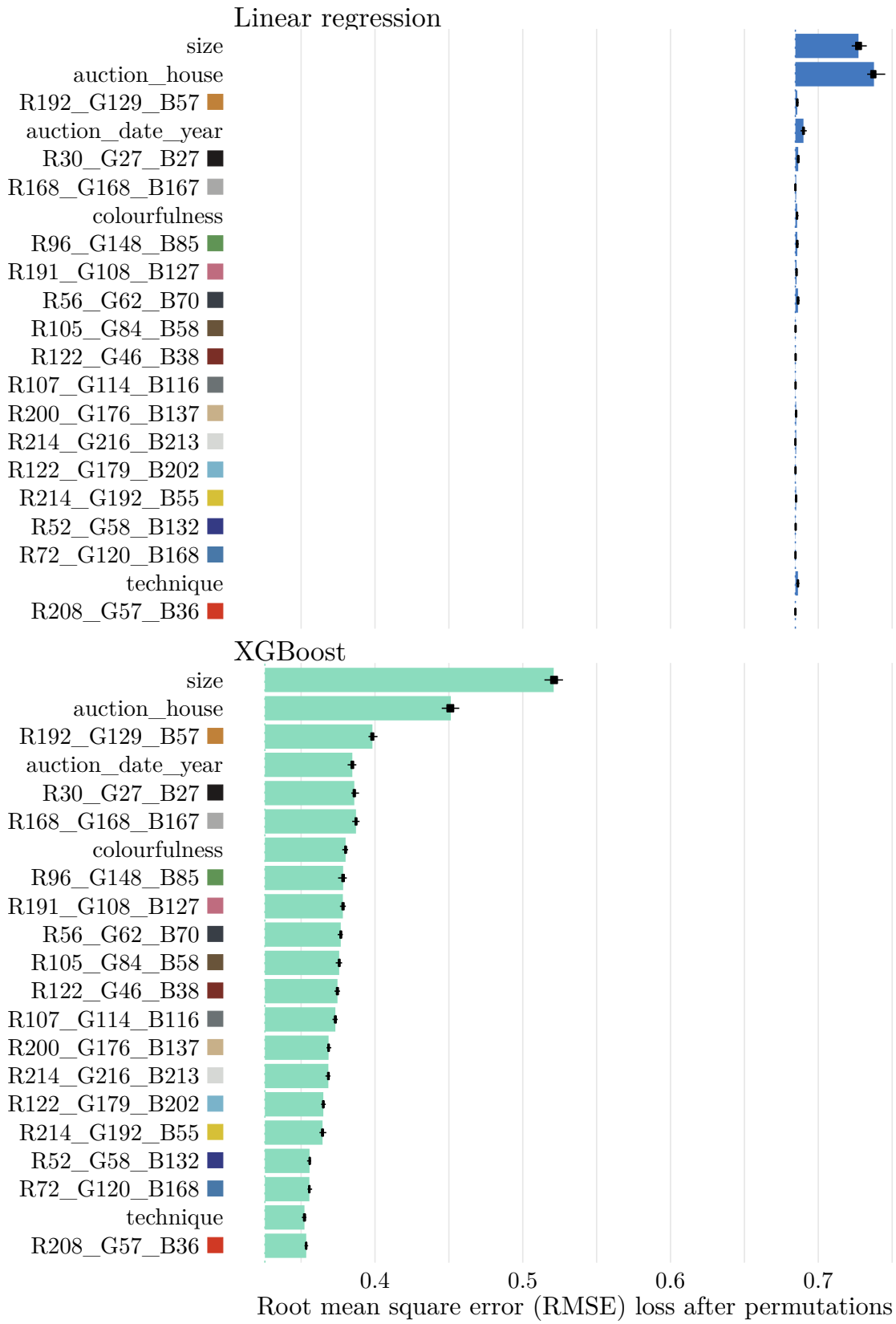


Figure 6.5: Feature importance in the Young Art models after 50 permutations.

Figure 6.6: Partial-dependence profiles for size and colourfulness in the Young Art models.

can be observed. It seems that there are no colours that drastically change the hammer price on the young art markets, although there are several relationships, which are worth describing. For instance, for the green (R96_G148_B85) the works which have up to several percent, there were priced slightly above the average, which forms an L-shaped relationship. Similarly, a small (up to 10%) share of R191_G108_B127 seems to be valued by art buyers. For R192_G129_B57, the relationship is linear and decreasing up to roughly 50% share, after which it goes up for the highest values. Somewhat surprisingly, the share of red R208_G57_B36 does not seem to make much difference for buyers.

6.2.2 Top 10 Painters Dataset

Similarly to the Young Art dataset, two linear models were created using OLS – the first with *standard* variables and the second with additional colour-related features, as described in Section 5.2. Due to potential multicollinearity, these variables were not a subject of the regression analysis: `author Alfons Karpiński`, `auction_date_year 2008`, `auction_house Agra Art`, `technique oil canvas`, and R212_G211_B203. This time, a much higher coefficient of determination values has been obtained. The main reason for such a change might be attributed to the inclusion of the `author` variable, from which usually the most of explained variability comes – along with `size`. The first model got R^2=0.784 (adjusted R^2=0.778), whereas the second one got R^2=0.791 (adjusted R^2=0.782). This time adding the colour-related variables can be accounted for an

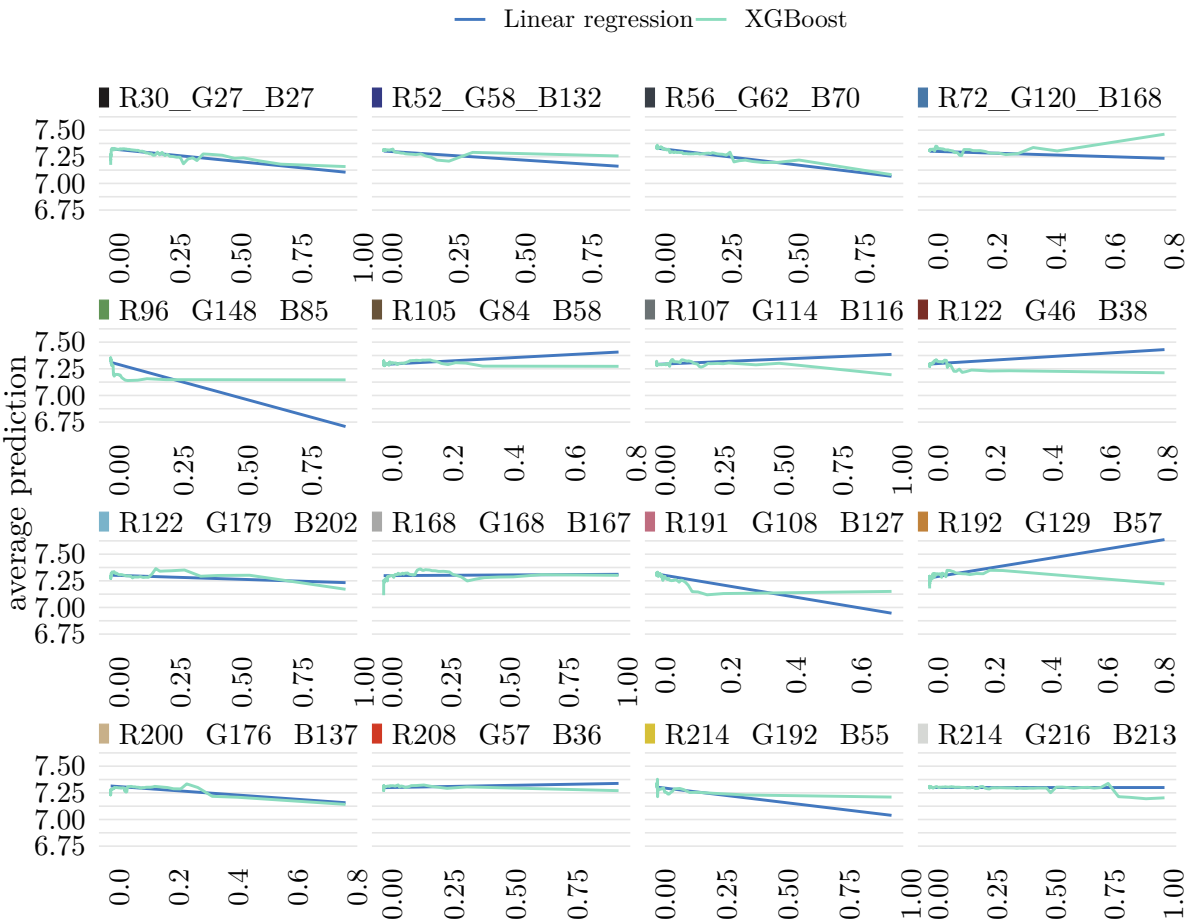


Figure 6.7: Partial-dependence profiles for colours in the Young Art models.

additional 0.5% of variance explained. The coefficient of determination values might suggest a good fit and be considered good enough in quantitative art market research, though it is worth noticing that in terms of RMSE there is still room for improvements. Similarly to the Young Art linear model, regression diagnostics have been performed. Residual vs fitted plot displayed a linear pattern in residuals (Figure A12). Again Shapiro-Wilk test ($W = 0.97424$, $p < 9.424\text{e}{-13}$) and Kolmogorov-Smirnov test ($D = 0.053754$, $p = 0.004474$) both rejected their null hypotheses at $\alpha = 0.05$, but the visual inspection of normal quantile-quantile plot suggested an approximate normality in terms of residuals with a heavy tail (Figure A13). Homoscedasticity was assessed with a scale-location plot (Figure A14), revealing an acceptable slight increase in fitted values. In terms of multicollinearity, no dependant variable violated the rule of thumb ($\text{GVIF}^{1/(2\text{DF})} < \sqrt{5}$), although colourfulness was close with $\text{GVIF}^{1/(2\text{DF})} \approx 2.208262$.

Regression results are compared in Table 6.12, where Model 1 and Model 2 are the models without and with the colour variables respectively. As in the previous subsection, we now refer to the second one as the linear model. The intercept equal to 9.126 can be interpreted as setting the hammer price of an *standard* lot in the dataset to $e^{9.126} = 9,191.18$ PLN. Regarding the `size` variable, an increase of 1cm results in $100 \cdot (e^{0.019} - 1) \approx 2\%$ price gain. Almost all artists turned out to be statistically significant. The same can be said about Polswiss (positive impact on price) and Rempex (negative impact) auction houses. In terms of `technique`, only the oil on board was the significant variable, which accounts for $100 \cdot (e^{-0.235} - 1) \approx -21\%$ price drop on average. This time the colourfulness is account for a positive impact on the log of the hammer price, averaging at $100 * (e^{0.005} - 1) \approx 0.5\%$ price change with every unit added. The only statistically significant colour at 1% level is ■ `R101_G125_B81` which had a negative impact on price – each additional percent of green in the lot results in $1 * (e^{-1.172} - 1) \approx -0.69\%$ price drop on average.

Nearly all variables related to the auction year turned to be statistically significant at a 5% level. Figure 6.8 presents the index for the Top 10 Painters dataset. Just as in the Young Art dataset, the standard formula based at 1 was used:

$$\text{Index} = e^{\gamma_t}, \tag{6.2}$$

where γ_t stands for the coefficient values obtained from HR Model 2 and $\gamma_{2008} = 0$. The sharp drop at the beginning of the index might be attributed to the global financial crisis in 2007, though

Table 6.12: Regression results for the Top 10 Painters dataset.

	Dependent variable: ln(price_final)	
	Model 1	Model 2
author Edward Dwurnik	−1.169*** (0.098)	−1.087*** (0.115)
author Jacek Malczewski	1.594*** (0.089)	1.617*** (0.091)
author Jerzy Kossak	−0.361*** (0.078)	−0.316*** (0.085)
author Jerzy Nowosielski	1.333*** (0.091)	1.330*** (0.097)
author Wiktor Korecki	−1.275*** (0.091)	−1.230*** (0.095)
author Wlastimil Hofman	0.127 (0.083)	0.137 (0.087)
author Wodzimierz Terlikowski	−0.346*** (0.100)	−0.336*** (0.106)
author Wojciech Kossak	0.309*** (0.087)	0.362*** (0.090)
author Wojciech Weiss	0.386*** (0.097)	0.403*** (0.100)
auction_date_year 2009	−0.155 (0.106)	−0.141 (0.105)
auction_date_year 2010	−0.327*** (0.100)	−0.319*** (0.100)
auction_date_year 2011	−0.361*** (0.099)	−0.355*** (0.099)
auction_date_year 2012	−0.408*** (0.100)	−0.387*** (0.100)
auction_date_year 2013	−0.353*** (0.096)	−0.344*** (0.096)
auction_date_year 2014	−0.435*** (0.093)	−0.403*** (0.094)
auction_date_year 2015	−0.299*** (0.087)	−0.287*** (0.088)
auction_date_year 2016	−0.293*** (0.083)	−0.259*** (0.085)
auction_date_year 2017	−0.377*** (0.086)	−0.356*** (0.086)
auction_date_year 2018	−0.450*** (0.105)	−0.440*** (0.105)
auction_house Desa Unicum	−0.024 (0.050)	−0.041 (0.051)
auction_house Polswiss	0.466*** (0.077)	0.422*** (0.080)
auction_house Rempex	−0.236*** (0.052)	−0.247*** (0.054)
auction_house Sopocki Dom Aukcyjny	0.118 (0.073)	0.080 (0.076)
technique oil, board	−0.233*** (0.088)	−0.235*** (0.088)
technique oil, plywood	0.121 (0.086)	0.105 (0.086)
technique oil, paperboard	−0.055 (0.051)	−0.055 (0.051)
size	0.019*** (0.001)	0.019*** (0.001)
colourfulness		0.005** (0.002)
R40_G36_B35		0.308 (0.212)
R49_G75_B127		−0.395 (0.437)
R71_G67_B57		−0.204 (0.243)
R86_G102_B106		−0.242 (0.383)
R91_G142_B181		−0.830* (0.483)
R101_G125_B81		−1.172*** (0.421)
R111_G72_B50		−0.243 (0.259)
R127_G115_B101		0.390 (0.255)
R151_G175_B185		−0.219 (0.278)
R171_G115_B69		−0.543 (0.331)
R173_G170_B158		−0.072 (0.231)
R181_G161_B92		−0.878* (0.517)
R185_G63_B49		−0.191 (0.310)
R185_G156_B123		−0.134 (0.286)
R202_G182_B59		−1.300* (0.692)
Constant	9.176*** (0.120)	9.126*** (0.183)
Observations	1,056	1,056
R^2	0.784	0.791
Adjusted R^2	0.778	0.782
Residual Std. Error	0.558 (df = 1028)	0.554 (df = 1012)
F Statistic	138.169*** (df = 27; 1028)	88.962*** (df = 43; 1012)
Note:		*p<0.1; **p<0.05; ***p<0.01

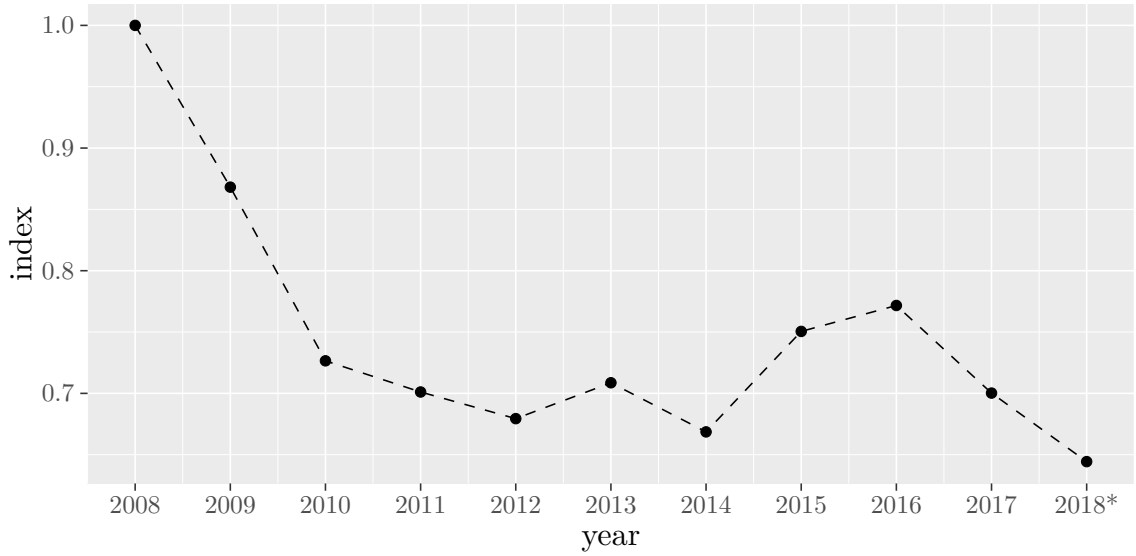


Figure 6.8: Art market index for the Top 10 Painters dataset.

Poland was rather not affected by it – at least compared to other European countries. However, this might be the case for the entities operating on the global market (including domestic and foreign capital), which invest in Polish art. The index stabilises itself in 2010 and oscillates around that value, not returning to the initial number. Does this mean that the Polish art market is going down? Not necessarily, if we look at how the index was constructed. Recently, more modern- and contemporary-oriented auctions have been held. The growing interest in these works is confirmed in the record-breaking prices of artists such as Roman Opałka and Wojciech Fangor. The majority of the artists included in the index were living between the 19th and 20th century, so the poor performance of the index might be attributed to the decreasing interest in such lots. This also sets an interesting direction for future work, as it might be expected to see more modern and contemporary art at the dedicated auctions.

While this time the linear model performed considerably better, the results can still be improved with the XGBoost model. As with the previous example, a 10-fold cross validation in terms of RMSE was run, resulting in the following set of the best parameters: `nrounds` = 1000, `max_depth` = 4, `eta` = 0.01, `gamma` = 0.2, `colsample_bytree` = 1, `min_child_weight` = 1, `subsample` = 0.7. Similarly to the Young Art dataset, the XGBoost doubled its performance in terms of RMSE. The detailed metrics for a comparison of both models can be found in Table 6.13. Figure 6.9 depicts the box plots of residuals for the models, with a much narrower interquartile distance in favour of XGBoost again (the red dot denotes root mean square of residuals).

Table 6.13: Model performance comparison for the Top 10 Painters dataset.

Model	MSE	RMSE	R^2	MAD
Linear regression	0.294	0.542	0.791	0.299
XGBoost	0.083	0.288	0.941	0.186

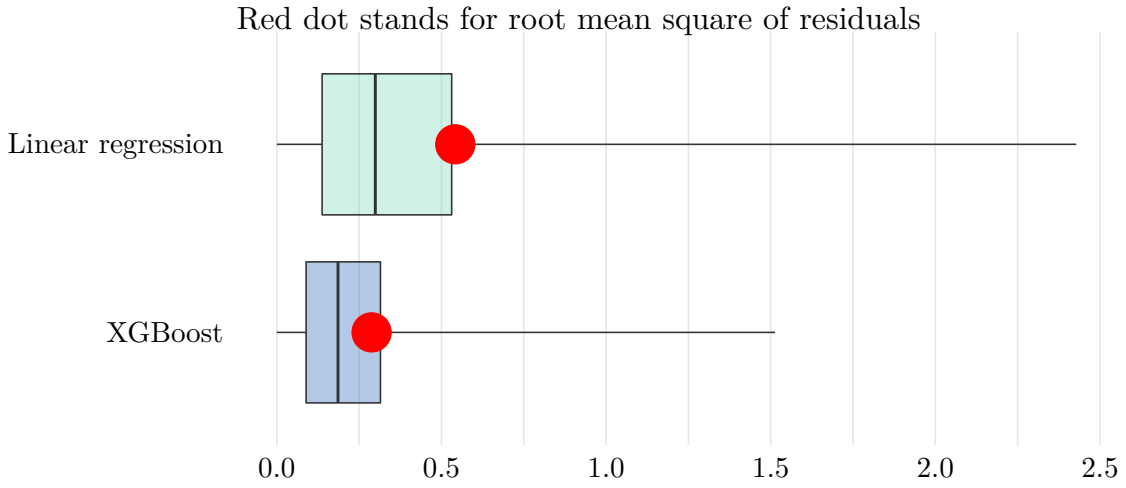


Figure 6.9: Boxplots of residuals in Top 10 Painters models.

Feature importance for both of the models are presented in Figure 6.10. For the linear model, leaving out `author` and `size` resulted in significant performance drops in terms of RMSE. A small, but noticeably difference can be observed with removing `auction_house`, `auction_date_year`, or `colourfulness`. The rest of the variables does not show significant importance. Among colours, ■ R101_G125_B81 has the highest importance, though it is relatively unnoticeable. Comparing all of this to the XGBoost model, some slight discrepancies might be observed. While `author` and `size` turned out to play the key role as well, `technique` was the third most important variable. However, the influence of `colourfulness` was small. Once again, ■ R101_G125_B81 was the most important colour, although this time it was followed closely by several other ones. The exact values for mean RMSE dropout losses are given in Table A.2 in Appendix.

Just as with the Young Art dataset, partial-dependence profiles have been prepared in order to investigate model outputs. Such profiles for `size` and `colourfulness` are depicted in Figure 6.11. Up to some point, the XGBoost model displayed the linear relationship in terms of `size` and the log of the hammer price. However, after reaching 100, the price becomes indifferent with additional centimetres. Naturally, this was beyond the capabilities of the linear model. In terms of `colourfulness`, the linear model showed a slight price increase with the increase of this measure. On contrary, the XGBoost model remains rather indifferent for this value.

Finally, Figure 6.12 depicts partial-dependence profiles for the colour share in the Top 10 Painters dataset. Once again, several discrepancies between models can be observed. In general, the XGBoost model has a tendency to diminish most of the colour influence, since for the majority of the chart the average predictions oscillate around the mean in most of the range – like in the case of ■ `R202_G182_B59`, which was practically ignored by the XGBoost model, but it is associated with a strong price decline in the linear model. However, a notable exception does exist – ■ `R101_G125_B81`, which has the highest importance for both models. In the linear model, the increase of the share of this colour is associated with the highest drop in the amount of log price. The XGBoost confirms it up to a certain point – the strong downward slope is present up to roughly 20% of this colour, whereas the further increase does not change the price much (in fact, there's even a small positive effect). It is worth noticing that buyers seem to display a similar negative attitude towards ■ `R96_G148_B85` in the Young Art dataset. A somewhat similar pattern with log-like curves for ■ `R40_G36_B35`, ■ `R49_G75_B127`, and ■ `R212_G211_B203` can be observed, which may be associated a strong effect with the the initial value range. The effects diminish with higher values.

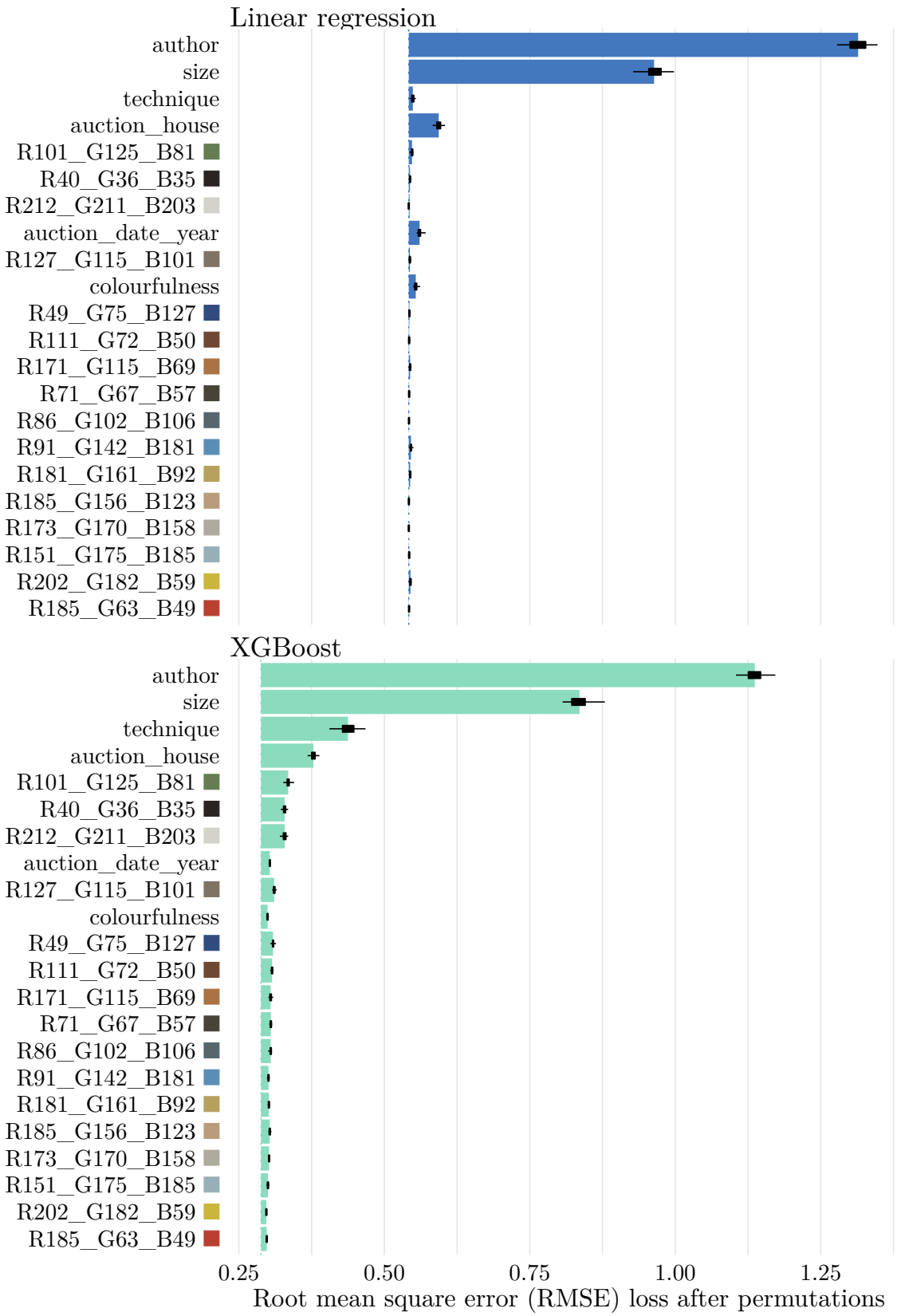


Figure 6.10: Feature importance in Top 10 Painters models after 50 permutations.

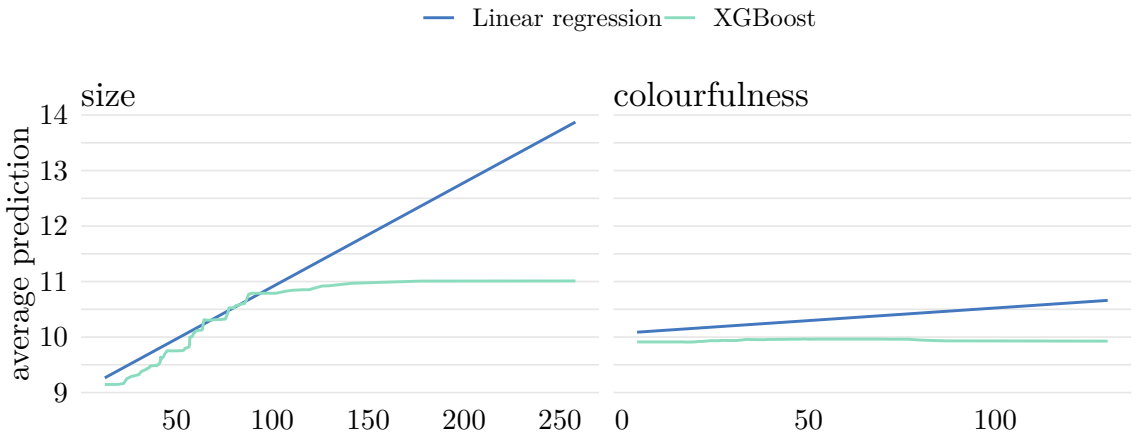


Figure 6.11: Partial-dependence profiles for size and colourfulness in Top 10 Painters models.

To further examine the effect of ■ R101_G125_B81 compared to other factors, SHAP values for the XGboost model were employed to examine the "greenest" (the share of this colour around 53%) painting in the dataset (Figure 6.14), made by Alfred Karpiński and sold for 22,000 PLN in Rempex. Since the ordering of the examined features using the SHAP method does matter while obtaining contributions (Biecek & Burzykowski, 2020), an average of 25 random orderings has been used (hence the boxplots in the figure). The results presented in Figure 6.13 suggest that this is indeed the most significant colour in terms of SHAP (−0.085), though its contribution to the expected value does not stand out compared to e.g. technique (−0.084) or auction house (−0.108) – not to mention small size (−0.621). ■ `R101_G125_B81` had almost always the negative contribution for the final price. The more detailed values (also for variables not included in the chart) are provided in Table A.3 in Appendix.

6.3 Discussion

The results of all experiments yielded some interesting findings, which are worth summing up. We collected and analysed two datasets – the Young Art dataset and the Top 10 Painters dataset. Both were analysed in terms of two factors – buyer's preferences and price determinants. The data on the sold paintings on the Young Art auctions collected from 2011 to (roughly) the half of 2018. In terms of buyer's preferences, the analysis showed a fast increase in the number of sold paintings in the young art sector, which was particularly visible in 2016. Acrylic and oil on

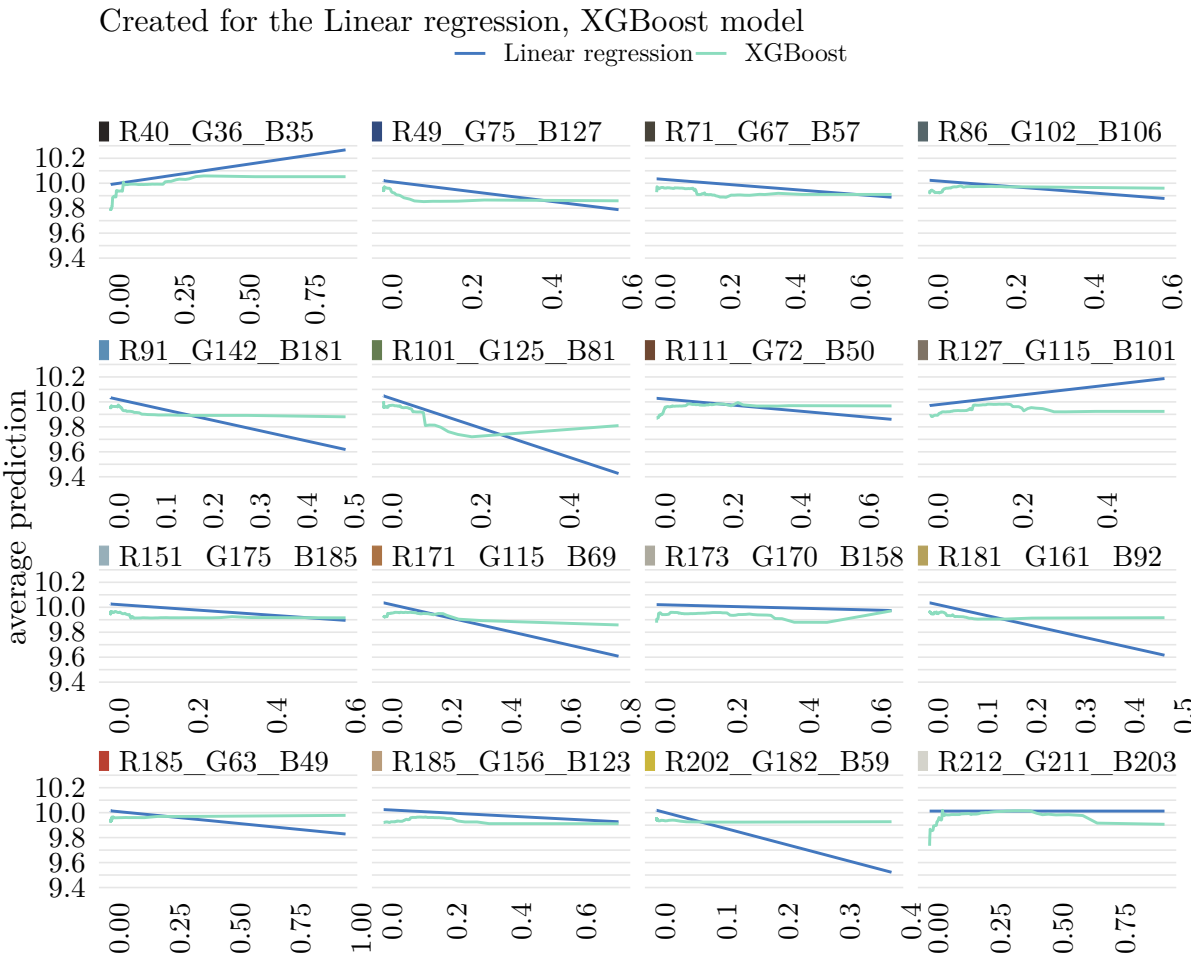


Figure 6.12: Partial-dependence profiles for colours in Top 10 Painters models.

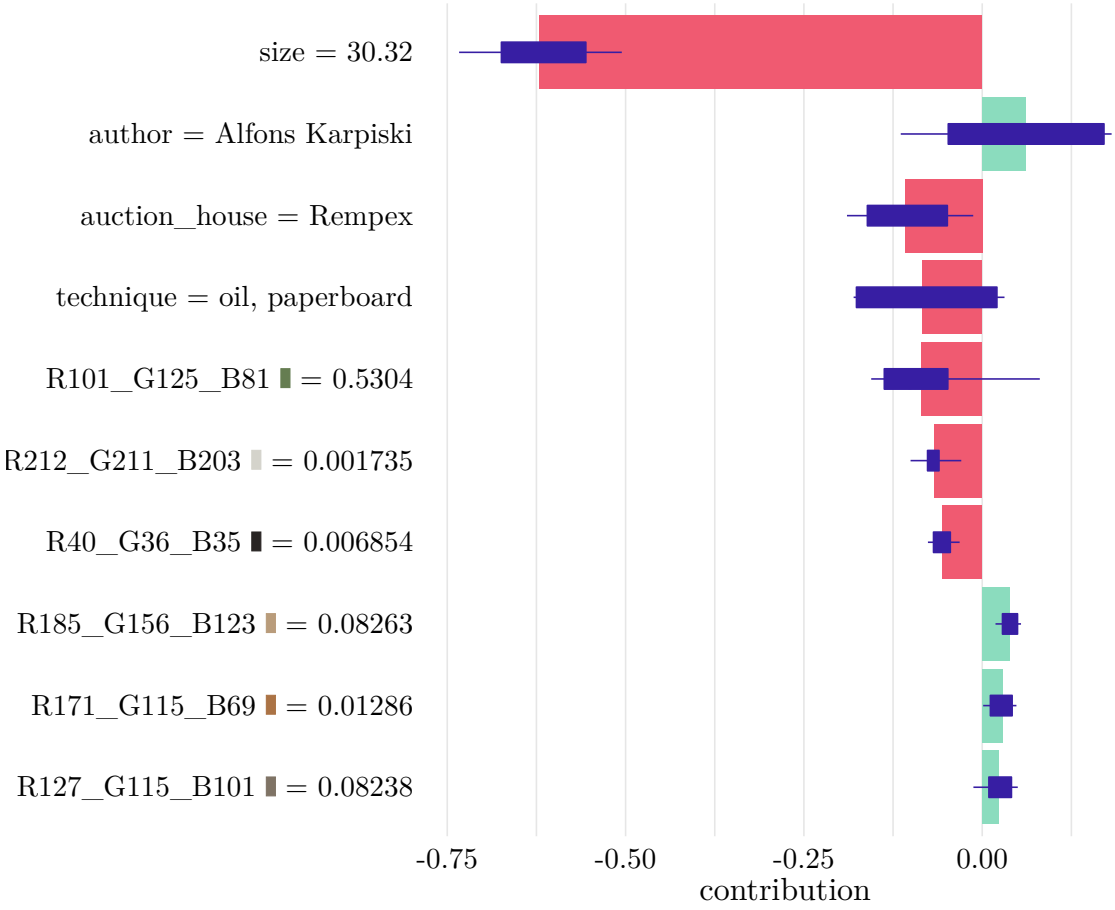


Figure 6.13: Average SHAP values attributions on 25 random orderings for Karpiński's *Wiejska droga w Bronowicach*.

Figure 6.14: Alfred Karpiński, *Wiejska droga w Bronowicach*, 1903.
Source: `rempex.com.pl`

canvas were almost equally popular techniques among the buyers of young art. Most of the young art has been sold in Desa Unicum and Polswiss Art auction houses. The average hammer price was 1949.65 PLN (adjusted to 2018 inflation). The average size of the painting turned to be 88 $\times$ 88 cm^2, whereas the colourfulness averaged at 50. The data for the Top 10 Painters painters regards sold lots in years 2008-2018 (again, up to roughly the first half of this year). These top 10 painters (in terms of the number of sold works) turned out to be A. Karpiński, E. Dwurnik, J. Malczewski, J. Kossak, J. Nowosielski, W. Korecki, W. Hofman, W. Terlikowski, W. Kossak, and W. Weiss. This list solely says a lot about the buyers' preferences in these years – the majority of the painters in the dataset lived between 19 and 20th centuries. This, however, seemed to change in recent years, as the younger generation of artists made some eye-catching price records at the Polish auction houses (Wojciech Fangor, Tamara Łępicka, Małgorzata Abakanowicz...). Naturally, the hammer price within these well-established artists was much higher, with the mean at ≈54,000 PLN (adjusted to 2018 inflation). The average size of a sold painting was this time slightly smaller: $\approx 62.5 \times 62.5$ cm^2. The same applies to the colourfulness of sold paintings (≈ 39 on average). Most of the works were sold in Desa Unicum, Rempex, and Agra Art. Interestingly, oil paints were the only ones used (mostly on canvas or paperboard). Oil paints are traditionally perceived as *the true and noble* ones (contrary to acrylic paints, for instance). However, their

omnipresence within the sold lots in the dataset might stem from the simple fact – they were invented around the 1940s and become more popular in the 1950s. In contrast, the majority of the painters in the dataset lived between 19 and 20th centuries.

The analysis of price determinants showed some interesting patterns. For each dataset, two kinds of models have been built: hedonic (which is in fact a linear regression appliance) and XGBoost model. Hedonic models are traditionally used in art market analysis due to their simplicity and ability to assess the impact of particular qualities of a painting. However, the relationship between the (logarithm of) hammer price and features of a given painting is not necessarily linear – not to mention other conditions which have to be met. Especially results from the Young Art linear model have to be treated with caution due to the relatively low value of explained variance. Therefore, we employed the XGBoost model in order to validate the results – features have been assessed using permutation importance. In terms of hedonic regression, we also reported the statistical significance of features. The coefficient values directly described the positive/negative impact of a given estimate. For XGBoost models, permutation importance and partial-dependence profiles were used to assess the influence of variables. In order to minimise the effect of overfitting, we used cross-validation during the training XGBoost models. For the Young Art dataset, size and auction house turned out to be the most important predictors of the price (in terms of permutation importance for both models). For the Top 10 Painters dataset, the name of the author and size was the strongest price determinants. These results are not surprising given the line of previous research.

We hypothesised that some colours might be preferred by buyers and this will be reflected in a positive influence on hammer prices. Similarly to other features, this was measured using statistical significance in linear models and feature importance in XGBoost models. Instead of finding the most valuable colour, we rather discovered the most "disliked" one – at least in the context of prices in the Young Art and Top 10 Painters datasets. Two slightly different shades of green appear to impact prices negatively: ■ R96_G148_B85 and ■ `R101_G125_B81` for the Young Art and the Top 10 Painters datasets. 150 paintings were having at least 10% of ■ `R96_G148_B85` in the young art dataset. On average, these paintings were cheaper by 432.27 PLN compared to all young art paintings. The geometric mean of the size of these paintings was smaller by 1.63 cm. However, they were more colourful (by 18.21). In terms of the Top 10 Painters dataset, 70 paintings were having at least 10% of ■ `R101_G125_B81` in the young art dataset. Interestingly, these paintings were more expensive (+6911.78 PLN on average) compared

to the mean price in the whole dataset. The geometric mean of their size was also smaller by 6.78 cm. Similarly to the young art, their "green" features slightly contributed to larger colourfulness (by 1.81). Since for this dataset the author distribution is available, it can be compared to the distribution of authors of having at least 10% of R101_G125_B81. Jacek Malczewski had 14.29% of paintings in the dataset with paintings with at least 10% green, compared to 8.62% in the whole dataset. For Wojciech Kossak, it was 17.14% versus 9.47%. A possible explanation of this might be the fact that W. Kossak is known from his landscapes, whereas the skin of people portrayed by Malczewski has a characteristic green tone. This hypothesis might be validated in future studies.

Our analysis also allowed us to create financial indices for the Polish art market. The index for the Young Art (Figure 6.3) dataset displays an upward trend up to a correction in 2016, in which its value returned to the levels from 2013 and remains stable up to 2018. A possible explanation for this correction might be connected to the Abbey House auction house scandal and the consequential loss of trust in young and emerging art from the side of investors. The index for the Top 10 Painters dataset (Figure 6.8) resulted in somewhat counter-intuitive figures, starting with a sharp fall in 2008, which may be attributed to the financial crisis. However, the market reports indicate that the Polish art market is on the rise in terms of numerous figures. A reason for this discrepancy might be the set of artists selected for the index construction. A closer look in the dataset suggests that mean prices for 2008 were significantly higher than for the other years. In 63 paintings sold in 2008, there are 10 paintings made by Jacek Malczewski ($\approx 15\%$) and 21 by Nowosielski ($\approx 33\%$). Given the fact that these two artists are much more expensive than the rest of the artists (in fact, it is an order of magnitude of the difference – see Table 6.7), the fact that these painters constituted almost a half of the available data might partially explain the highest index value in 2008. The reason for this overrepresentation is unknown – there's a possibility of a sell-out of some large collection, maybe in order to liquidate the paintings due to the financial crisis. However, a simple experiment for the Top 10 Painters index creation with excluded these two artists has shown that the shape of the index plot remains similar. How to interpret this figure then? Certainly, it does not mean that the Polish art market "is down" for almost ten years. This might suggest that the interest of buyers is shifting towards the more modern artist, such as Wojciech Fangor or Roman Opałka – and this is reflected with more and more auctions with op-art or non-figurative contemporary abstract art.

Up to a certain extent, our findings are in line with other results provided by other quanti-

tative researchers of the Polish art market. In their research, Białowąs et al. (2018) showed a peak in 2008 and a downward trajectory for 2008-2012 (see Figure 2.8). The results for the later years, however, are not available. Albeit a different methodology was used for the index creation, a somewhat similar pattern is visible. While Białowąs et al. (2018) used repeat-sales instead of hedonic regression, it might be assumed that painters from the Top 10 artist still constituted a significant share of sold lots. The same peak is also visible in other works. Witkowska and Lucińska (2015) analysed different indices for 2007-2013 using 22 painters – including all the ones from the Top 10 Painters dataset except W. Terlikowski. They created 5 models and tested 3 different types of indices. Interestingly, in all of the 15 pairs, the index value was the highest for 2008.

Most importantly, the usage of Artefact 1 (Algorithm 18) resulted in finding important colour-related features of paintings (such as the aforementioned share of green colour). Hedonic models without these features are characterised by smaller R^2 and adjusted R^2 (see figures in tables 6.10 and 6.12 for the Young Art and Top 10 Painters datasets respectively). This means the smaller share of explained variance for both of the analysed datasets without the usage of these colour-related variables. Additionally, the non-zero permutation importance values for colour variables in XGBoost models (tables A.1 and A.2) further indicates the importance of these features. Therefore, the validity of Thesis T1 (*An application of colour quantisation with Algorithm 18 in order to extract features of paintings increases the explained variance of models representing buyer's preferences and price determinants on the Polish art market.*) is supported.

6.4 Summary

In this chapter, we focused on describing buyers' preferences and price determinants for the Polish art market. The following research questions and objectives were a subject of this chapter:

- the research question Q4 (*Which features of paintings are important for buyers on the Polish art market?*) answered through describing buyers' preferences and price determinants,
- the research objective O1 (*Prepare datasets allowing conducting the experiment*) thanks two preparation of two datasets,
- the research objective O3 (*Evaluate the method for extracting colour-related features on Polish art market data*) due to empirically proven usefulness of Artefact 1,
- the research objective O4 (*Discuss which features of paintings are important for buyer's*

preferences on the Polish art market) obtained through descriptive statistics about the Polish art market,

- the research objective O5 (*Discuss price determinants for paintings on the Polish art market*) obtained through estimating price determinants for the Polish art market.

We constructed and presented two datasets, which are described in Section 6.1. Then, we conducted some experiments and provided a detailed analysis of built models in Section 6.2. Finally, Section 6.3 provides an analysis of these results in terms of preferences and price determinants. In Section 6.1.1, we introduced two datasets. The first one is the Young Art dataset, which gathers paintings from Polish young art auctions. The second one, named the Top 10 Painters dataset, was built using the most popular artists in the available data and presented in Section 6.1.2. Sections 6.2.1 and 6.2.2 describe our effort to construct classic hedonic regression models and confront them with high-performing XGBoost models for both datasets.

From an economics and financial perspective, some practical results have been delivered, which warrant further discussion. Hedonic regression models enabled to obtain art market indices. The market for young artists seem to be on the rise since 2011 (excluding the correction in years 2014-2015). The index for the Top 10 Painters dataset, however, displayed a downward trend. This might be attributed to the shifting preferences of Polish buyers – in the very recent auctions, modern artists gain more and more attention. In general, adding colour-related variables resulted in a small increase of explained variance for both linear models, which supports the validity of Thesis T1. As one might expect, most of the explained variance in the models came from the variable describing the artist. Since this was explicitly ignored in the Young Art dataset, the constructed model suffered from a poor fit. The size of artwork has high importance in both models. For the Top 10 Painters dataset, artists played a key role. However, the usage of explainable artificial intelligence methods for model comparison has revealed some discrepancies between models. Quite surprisingly, the models suggest that buyers tend to value less colourful artworks when buying young art. When it comes to the Top 10 Painters dataset, however, this feature seems to make no difference to the hammer price. The colour analysis does not show that there is a particular one guaranteeing high hammer prices. However, the analysis revealed which colour is certainly not attributed with auction records – as shown in both of the models, a large share of green is associated with price drops.

Chapter 7

Summary and Future Work

The final chapter of this dissertation can be conceptually divided into two parts. The first one is devoted to summarising our research in terms of the answered research questions, realised research objectives, and delivered research contributions. After that, we focus on possible directions for further work, as some follow-up questions remain to be answered.

This study used various quantitative techniques to elaborate on the stated research problem. This was realised by answering the following research questions:

- Q1 (*Which methods can be used to assess the importance of paintings' features for the hammer price?*),
- Q2 (*How to extract colour-related information from paintings?*),
- Q3 (*What is the best colour quantisation algorithm for paintings?*),
- Q4 (*Which features of paintings are important for buyers on the Polish art market?*).

The research question Q1 (*Which methods can be used to assess the importance of paintings' features for the hammer price?*) was answered by the literature review on quantitative art market analysis provided in Chapter 2. Chapter 4 described more recent approaches for explanatory data analysis. Traditional techniques (such as hedonic and repeated-sales regression) and more modern ones (various tree ensembles) have been summarised. The problem stated in the research question Q2 (*How to extract colour-related information from paintings?*) was dealt with in Chapter 3. While there are numerous algorithms suited for the problem of colour quantisation, the research question Q3 was concerned with choosing the right one for this research (*What is the best colour quantisation algorithm for paintings?*). Different algorithms have been tested in terms of their quantisation error and colour diversity in order the pick the most suitable one. For this research, Chang's k-means has been chosen, as it strikes a good balance between these two metrics. The

last research question Q4 (*Which features of paintings are important for buyers on the Polish art market?*) was answered in Chapter 6 – Section 6.1 provides summary statistics about sold paintings (i.e. buyer's preferences), whereas Section 6.2 tries to frame their particular features as price determinants. They are briefly summarised in the next paragraph.

Research objectives realised in this dissertation are:

- O1 (*Prepare datasets allowing conducting the experiment*),
- O2 (*Develop a method for extracting colour-related features from paintings (Artefact 1)*),
- O3 (*Evaluate the method for extracting colour-related features on Polish art market data*),
- O4 (*Discuss which features of paintings are important for buyer's preferences on the Polish art market*),
- O5 (*Discuss price determinants for paintings on the Polish art market*).

The research objective O1 (*Prepare datasets allowing conducting the experiment*) was realised by collecting an extensive dataset of paintings sold in the biggest Polish auction houses between 2008 and 2018. This dataset was narrowed down to two more fine-grained ones – the young art dataset and the top 10 painters dataset. The research objective O2 (*Develop a method for extracting colour-related features from paintings (Artefact 1)*) was realised by the development of Artefact 1 (Algorithm 18), which is an algorithm for the calculation of the share of representative colours for paintings. This algorithm uses Chang's k-means algorithm to get k representatives as the answer to the research question Q3. The research objective O3 (*Evaluate the method for extracting colour-related features on Polish art market data*) has been realised by the usage of statistical significance and practical importance (measured by XAI methods). With the usage of Artefact 1 (Algorithm 18), we found important colour-related features of paintings (such as the share of green colour). Linear models without these features were found to have e.g. smaller R^2 and adjusted R^2 (that is, the smaller share of explained variance). These conclusions support the validity of Thesis T1 (*An application of colour quantisation with Algorithm 18 in order to extract features of paintings increases the explained variance of models representing buyer's preferences and price determinants on the Polish art market.*) – as well as non-zero permutation importance values for colour variables in XGBoost models. The research objective O4 (*Discuss which features of paintings are important for buyer's preferences on the Polish art market*) has been fulfilled by collecting descriptive statistics of the paintings sold in the Polish auction houses. A *typical* painting belonging to the young art category has been sold with a hammer price of approximately 2,000 PLN at Desa Unicum or Sopocki Dom Aukcyjny. Acrylic on canvas or oil

on canvas paintings are the most common paintings in the young art auctions, sized 88 cm^2 each side. Jerzy Kossak, Wlastimil Hofman and Jerzy Nowosielski open the list of the top 10 Polish painters. Paintings belonging to this category were most often purchased at Desa Unicum for 54,000 PLN. A typical painting would be either made with oil on canvas or oil on paperboard, sized 60 cm^2 each side. These paintings also have a darker colour palette compared to the young art dataset. Finally, the research objective O5 (*Discuss price determinants for paintings on the Polish art market*) was reached by the usage of hedonic models with the datasets collected as the research objective O3 and the Artefact 1. For the young art dataset, the most of explained variance came from `auction_house` and `size` variables. In terms of colours, we noticed that green (■ R96_G148_B85) is associated with lower prices. Regarding the top 10 painters dataset, most of the explained variance came from the `author` and `size` variables, which is in line with other quantitative studies of art markets. Surprisingly, we also found that the share of green colour has a negative effect on hammer price – this time in a slightly different hue (■ R101_G125_B81).

During the writing of this dissertation, a number of research contributions have been made. The most important contributions are:

- Artefact 1, an algorithm for calculation of the share of representative colours for paintings,
- comparison of popular quantisation methods for extracting dominant colours in paintings,
- two datasets of paintings sold in the Polish auction houses,
- first hedonic analysis on the Polish art market in the period from 2007 to 2018,
- four models explaining painting's hammer price,
- first colour-related analysis in the Polish art market,
- first quantitative analysis of the influence of colours in the Polish art market,
- hedonic art market indices between 2007 and 2018.

Future studies may further our understanding of several art market phenomena in numerous ways. A certain limitation of this study is connected to the problem of selection bias since we only analysed sold works. For instance, one might concentrate on *sell determinants* (contrary to buyer's preferences and price determinants, as in this thesis). This might be achieved using binary classification (such as logit models) for sold and unsold lots. Such a method would enable to examine which features contribute to the likelihood of being a bought-in. Some researchers investigated it already, such as Seçkin and Atukeren (2012) or Bruno, Garcia-Appendini, and Nocera (2018). This, however, requires appropriate and complete data about bought-ins. At the moment of writing this thesis, some of the auction houses provide information about only

the sold lots in their results. Accessing external paid databases might help to fulfil this goal. A natural idea for the next research directions would be using more fine-grained datasets (in this work, we used two). This would enable the isolation of the effects of particular art periods and styles. The usual granulation provided by Polish auction houses (such as pre- and post-war art, op-art and conceptual art, or post-90s) can be used as an example. As a consequence, this would provide new indices.

Since we generated art market indices, it would be nice to compare them with other forms of investments, such as stock and bonds. This comparison would be especially interesting with an extension of this work up to 2020 – it is already known that the COVID-19 pandemia was an important source of shock to numerous markets. It might also be interesting to compare different index types, as Witkowska and Lucińska (2015) did. Since state-of-the-art tree ensemble algorithms have been used to generate the models, it might be tempting to check the predictive power. This has not been done in this thesis since we wanted to focus on explanatory power and compare the results with linear models, for which traditionally whole sample is used. However, there are no formal obstacles to perform a separate study on that in the future. Another future research direction might continue investigating the effect of colours. The share of surface occupied by colour from a quantised palette nor colourfulness does not exhaust the topic of colour-related features which can be extracted from paintings. More colour-related features (such as brightness or contrast) might also be investigated, as well general qualitative research explaining some colours in the context of art history research.

References

Al-Daoud, M. (2005). A new algorithm for cluster initialization. In *Wec'05: The second world enformatika conference.*

Al-Daoud, M., & Roberts, S. A. (1996). New methods for the initialisation of clusters. *Pattern Recognition Letters*, *17*(5), 451–455.

Aloise, D., Deshpande, A., Hansen, P., & Popat, P. (2009). Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, *75*(2), 245–248.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

An, J., & Cai, Z. (2008). Embedded trellis coded quantization for jpeg2000. *IEEE Transactions on Image Processing*, *17*(9), 1570–1573.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual acm-siam symposium on discrete algorithms* (pp. 1027–1035).

artinfo.pl. (2019). *Rynek sztuki w polsce. raport 2019.*

artinfo.pl. (2020). *Rynek sztuki w polsce. raport 2020.*

artinfo.pl. (2021). *Rynek sztuki w polsce. raport 2021.*

artnet Analytics. (2014). *artnet Indices White Paper* (Tech. Rep.).

Arvo, J. (Ed.). (1991). *Graphics Gems II.* Academic Press, Inc.

Bailey, M. J., Muth, R. F., & Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association*, *58*(304), 933–942.

Baumol, W. J. (1986). Unnatural value: or art investment as floating crap game. *The American Economic Review*, *76*(2), 10–14.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, *18*(9), 509–517.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms.* Norwell, MA,

USA: Kluwer Academic Publishers.

Białowąs, S., Potocki, T., & Rogozińska, A. (2018). Financial returns and cultural price determinants on the polish art market, 1991–2012. *Acta Oeconomica*, *68*(4), 591–615.

Biecek, P. (2018). Dalex: explainers for complex predictive models in r. *The Journal of Machine Learning Research*, *19*(1), 3245–3249.

Biecek, P., & Burzykowski, T. (2020). *Explanatory model analysis: Explore, explain and examine predictive models.*

Biey, M. L., & Zanola, R. (2005). The market for picasso prints: A hybrid model approach. *Journal of Cultural Economics*, *29*(2), 127–136.

Bloomberg, D. S. (2008). Color quantization using modified median cut. , 1–10.

Bocart, F. Y., & Hafner, C. M. (2012). Econometric analysis of volatile art markets. *Computational Statistics and Data Analysis*, *56*(11), 3091–3104. Retrieved from `http://dx.doi.org/10.1016/j.csda.2011.10.019` doi: 10.1016/j.csda.2011.10.019

Bock, H.-H. (2007). Clustering methods: a history of k-means algorithms. In *Selected contributions in data analysis and classification* (pp. 161–172). Springer.

Borowski, K. (2013). *Sztuka inwestowania w sztukę*. Difin SA.

Borowski, K. (2015). Rynek dzieł sztuki w latach 2012-2013 jako alternatywa innych form inwestowania. *Studia Ekonomiczne*, *239*, 7–24.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152).

Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for k-means clustering. In *Icml* (Vol. 98, pp. 91–99).

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brun, L., & Trémeau, A. (2003). Color quantization. In *Digital color imaging handbook* (pp. 589–638). CRC press.

Bruno, B., Garcia-Appendini, E., & Nocera, G. (2018). Experience and brokerage in asset markets: Evidence from art auctions. *Financial Management*, *47*(4), 833–864.

Burger, W., & Burge, M. J. (2009a). Colorimetric Color Spaces. In *Interactive image processing*

for machine vision (pp. 1–28). Retrieved from http://link.springer.com/10.1007/978-1-84800-195-4{_}6 doi: https://doi.org/10.1533/9780857099242.270

Burger, W., & Burge, M. J. (2009b). Color quantization. In *Principles of digital image processing: Core algorithms* (pp. 1–11). London: Springer London. Retrieved from https://doi.org/10.1007/978-1-84800-195-4_5 doi: 10.1007/978-1-84800-195-4_5

Buteikis, A. (2018). *Practical econometrics and data science*. Retrieved from http://http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/index.html

Buturovic, A. (2005). Mpeg 7 color structure descriptor for visual information retrieval project vizir 1.

Case, K. E., & Shiller, R. J. (1987). *Prices of single family homes since 1970: New indexes for four cities* (Tech. Rep.). National Bureau of Economic Research.

Celebi, M. E. (2011). Improving the performance of k-means for color quantization. *Image and Vision Computing*, *29*(4), 260–271.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, *40*(1), 200–210.

Chanel, O., Gérard-Varet, L.-A., & Ginsburgh, V. (1992). *The relevance of hedonic price indices: the case of paintings*. Groupe de Recherche en Economie Quantitative et Econométrie, UA CNRS.

Chang, H., Fried, O., Liu, Y., DiVerdi, S., & Finkelstein, A. (2015). Palette-based photo recoloring. *ACM Trans. Graph.*, *34*(4), 139–1.

Charlin, V., & Cifuentes, A. (2014). An investor-oriented metric for theart market. *The Journal of Alternative Investments*, *17*(1), 87–101.

Charlin, V., & Cifuentes, A. (2018). The Paintings of Mark Rothko: A Study of the Relationship Between Price and Color. *SSRN Electronic Journal*(October). Retrieved from https://www.ssrn.com/abstract=3262314 doi: 10.2139/ssrn.3262314

Chau, K. W., & Chin, T. (2003). A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Applications*, *27*(2), 145–165.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., &

Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Choudhury, A. K. R. (2014). *Using instruments to quantify colour*. doi: 10.1533/9780857099242.270

Chowdhury, S., Verma, B., Tom, M., & Zhang, M. (2015, jul). Pixel characteristics based feature extraction approach for roadside object detection. In *2015 international joint conference on neural networks (ijcnn)* (Vol. 2015-Septe, pp. 1–8). IEEE. Retrieved from http://ieeexplore.ieee.org/document/7280599/ doi: 10.1109/IJCNN.2015.7280599

Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., & Xu, Y.-Q. (2006). Color harmonization. In *Acm siggraph 2006 papers on - siggraph '06* (p. 624). New York, New York, USA: ACM Press. Retrieved from http://portal.acm.org/citation.cfm?doid=1179352.1141933 doi: 10.1145/1179352.1141933

Cohn, G. (2018). Ai art at christies sells for $432,500. *The New York Times*. Retrieved Aug. 26, 2019, from https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html

Collins, A., Scorcu, A., & Zanola, R. (2009). Reconsidering hedonic art price indexes. *Economics Letters*, *104*(2), 57–60.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms*. MIT press.

Czujack, C. (1997). Picasso paintings at auction, 1963–1994. *Journal of cultural economics*, *21*(3), 229–247.

David, G., Oosterlinck, K., & Szafarz, A. (2013). Art market inefficiency. *Economics Letters*, *121*(1), 23–25. Retrieved from http://dx.doi.org/10.1016/j.econlet.2013.06.033 doi: 10.1016/j.econlet.2013.06.033

Dekker, A. H. (1994, jan). Kohonen neural networks for optimal colour quantization. *Network: Computation in Neural Systems*, *5*(3), 351–367. Retrieved from https://www.tandfonline.com/doi/full/10.1088/0954-898X{_}5{_}3{_}003 doi: 10.1088/0954-898X_5_3_003

Deloitte. (2019). Art & Finance Report 2019.

Deselaers, T., Keysers, D., & Ney, H. (2005). Discriminative training for object recognition using image patches. In *2005 ieee computer society conference on computer vision and pattern recognition (cvpr'05)* (Vol. 2, pp. 157–162).

Deselaers, T., Keysers, D., & Ney, H. (2008). Features for image retrieval: an experimental comparison. *Information retrieval*, *11*(2), 77–107.

Domański, R. (2015). *Czy s w polsce klasy społeczne?* Krytyka Polityczna.

Dougherty, C. (2011). *Introduction to econometrics*. Oxford University Press.

Etro, F., & Stepanova, E. (2015). The market for paintings in p aris between r ococo and romanticism. *Kyklos*, *68*(1), 28–50.

Fairchild, M. D. (2005). *Color appearance models*. John Wiley & Sons.

Filipiak, D., & Filipowska, A. (2016). Towards data oriented analysis of the art market: survey and outlook. *e-Finanse*, *12*(1), 21–31.

Finkel, R. A., & Bentley, J. L. (1974). Quad trees a data structure for retrieval on composite keys. *Acta informatica*, *4*(1), 1–9.

Fisher, R. B., Breckon, T. P., Dawson-Howe, K., Fitzgibbon, A., Robertson, C., Trucco, E., & Williams, C. K. (2013). *Dictionary of computer vision and image processing*. John Wiley & Sons.

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, *57*(3), 238–247.

Floyd, R. W., & Steinberg, L. S. (1975). An adaptive algorithm for spatial gray scale..

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, *21*, 768–769.

Førsund, F. R., & Zanola, R. (2006). Dea meets picasso: The impact of auction houses on the hammer price. *Annals of Operations Research*, *145*(1), 149–165.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119–139.

Frey, B. S., & Pommerehne, W. W. (1989). *Muses and markets: Explorations in the economics of the arts*. Blackwell.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Ganczarski, J. (2004). *CIE LAB*. Retrieved from `http://lumen.iee.put.poznan.pl/do{_}pobrania/Wsp{_}chromat/CIE{_}Lab.pdf`

Gervautz, M., & Purgathofer, W. (1988). A simple method for color quantization: Octree quantization. In *New trends in computer graphics* (pp. 219–231). Springer.

Ginsburgh, V., Mei, J., & Moses, M. (2006). The computation of prices indices. *Handbook of the Economics of Art and Culture*, *1*, 947–979.

Glasbey, C., van der Heijden, G., Toh, V. F., & Gray, A. (2007). Colour displays for categorical images. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, *32*(4), 304–309.

Glassner, A. S. (Ed.). (1990). *Graphics Gems*. Academic Press, Inc. Retrieved from `https://linkinghub.elsevier.com/retrieve/pii/C20090213055`

Goetzmann, W. N. (1993). Accounting for taste: Art and the financial markets over three centuries. *The American Economic Review*, *83*(5), 1370–1376.

Gonzalez, T. F. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, *38*, 293–306.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Graddy, K., Hamilton, J., & Pownall, R. (2012). Repeat-sales indexes: Estimation without assuming that errors in asset returns are independently distributed. *Real Estate Economics*, *40*(1), 131–166.

Gramlich, J. (2017). Reflections on provenance research: Values–politics–art markets. *Journal for Art Market Studies*, *1*(2).

Habalová, V. (2018). Price determinants of art photography at auctions.

Habekost, M. (2013). Which color differencing equation should be used? *International Circular of Graphic Education and Research*(6).

Hasler, D., & Suesstrunk, S. E. (2003). Measuring colorfulness in natural images. In *Human vision and electronic imaging viii* (Vol. 5007, pp. 87–95).

Heckbert, P. (1980). *Color image quantization for frame buffer display* (Unpublished master's thesis). Massachusetts Institute of Technology.

Heckbert, P. (1982). *Color image quantization for frame buffer display* (Vol. 16) (No. 3). ACM.

Heckbert, P. S. (Ed.). (1994). *Graphics Gems IV*. Elsevier. Retrieved from `https://linkinghub.elsevier.com/retrieve/pii/C20130073604` doi: 10.1016/C2013-0-07360-4

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75–105.

Hill, R. (2011). Hedonic price indexes for housing.

Hochbaum, D. S., & Shmoys, D. B. (1985). A best possible heuristic for the k-center problem. *Mathematics of operations research*, *10*(2), 180–184.

Hodgson, D. J., & Vorkink, K. P. (2004). Asset pricing theory and the valuation of canadian paintings. *Canadian Journal of Economics/Revue canadienne d'économique*, *37*(3), 629–655.

Hoy, D. E. (1997). On the use of color imaging in experimental applications. *Experimental Techniques*, *21*(4), 17–19.

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the ieee international conference on computer vision* (pp. 1501–1510).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1125–1134).

Jancey, R. (1966). Multidimensional group analysis. *Australian Journal of Botany*, *14*(1), 127–130.

Judd, D. B., MacAdam, D. L., Wyszecki, G., Budde, H., Condit, H., Henderson, S., & Simonds, J. (1964). Spectral distribution of typical daylight as a function of correlated color temperature. *Josa*, *54*(8), 1031–1040.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4401–4410).

Katsavounidis, I., Kuo, C.-C. J., & Zhang, Z. (1994). A new initialization technique for generalized lloyd iteration. *IEEE Signal processing letters*, *1*(10), 144–146.

Keysers, D., Deselaers, T., Gollan, C., & Ney, H. (2007). Deformation models for image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(8), 1422–1435.

Kim, D., Son, S.-W., & Jeong, H. (2014). Large-scale quantitative analysis of painting arts. *Scientific reports*, *4*, 7370.

Kirk, D. (Ed.). (1992). *Graphics Gems III.* Academic Press, Inc.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. Retrieved from http://ieeexplore.ieee.org/document/58325/ doi: 10.1109/5.58325

Kompa, K., & Witkowska, D. (2013). Indeks rynku sztuki. *Badania pilotażowe dla wybranych malarzy polskich Art Price Index [Preliminary Investigation for Selected Polish Painters]. Zarządzanie i Finanse*, *11*(3), 33–50.

Kräussl, R. (2015). Art as an alternative asset class: Risk and return characteristics of the middle eastern and northern african art markets. *Cosmopolitan canvases: The globalization of markets for contemporary art*, 147–169.

Kräussl, R., & Elsland, N. v. (2008). *Constructing the true art market index: A novel 2-step hedonic approach and its application to the german art market* (Tech. Rep.). CFS working paper.

Kräussl, R., Lehnert, T., & Martelin, N. (2016). Is there a bubble in the art market? *Journal of Empirical Finance*, *35*, 99–109. Retrieved from http://dx.doi.org/10.1016/j.jempfin.2015.10.010 doi: 10.1016/j.jempfin.2015.10.010

Kräussl, R., & Schellart, E. (2007). Hedonic pricing of artworks: Evidence from german paintings. *Available at SSRN 968198*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kruger, A. (1994). Median-cut color quantization. *Dr Dobb's Journal-Software Tools for the Professional Programmer*, *19*(10), 46–55.

Laurent, H., & Rivest, R. L. (1976). Constructing optimal binary decision trees is np-complete. *Information processing letters*, *5*(1), 15–17.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science*, *9*.

Lin, S., & Hanrahan, P. (2013). Modeling how people extract color themes from images. In *Proceedings of the sigchi conference on human factors in computing systems - chi '13* (p. 3101). New York, New York, USA: ACM Press. Retrieved from http://dl.acm.org/citation.cfm?doid=2470654.2466424 doi: 10.1145/2470654.2466424

Linde, Y., Buzo, A., & Gray, R. (1980). An algorithm for vector quantizer design. *IEEE*

Transactions on communications, *28*(1), 84–95.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

Locatelli-Biey, M., & Zanola, R. (2002). The sculpture market: An adjacent year regression index. *Journal of Cultural Economics*, *26*(1), 65–78.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91–110.

Lucchese, L., & Mitra, S. K. (2001). Colour image segmentation: a state-of-the-art survey. *Proceedings-Indian National Science Academy Part A*, *67*(2), 207–222.

Lucińska, A. (2012). Rozwój artystyczny malarzy a ceny obrazów na polskim rynku sztuki. *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Finanse, Rynki Finansowe, Ubezpieczenia*(51), 715–725.

Lucińska, A. (2013). Wiek malarzy a ceny obrazów na rynku aukcyjnym w polsce. *Zarządzanie i Finanse*, *11*(3).

Lucińska, A. (2015). Zastosowanie narzędzi regresji hedonicznej do oceny poziomu stopy zwrotu i ryzyka inwestycji na rynku malarstwa polskiego. *Zarządzanie i Finanse*, *13*(4, cz. 2), 39–61.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, *2*(1), 2522–5839.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774).

Luo, H., Lin, D., Yu, C., & Chen, L. (2013). Application of different hsi color models to detect fire-damaged mortar. *International Journal of Transportation Science and Technology*, *2*(4), 303–316.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

Manjunath, B. S., Salembier, P., & Sikora, T. (2002). *Introduction to mpeg-7: multimedia content description interface*. John Wiley & Sons.

Marcellin, M. W., & Fischer, T. R. (1990). Trellis coded quantization of memoryless and gauss-markov sources. *IEEE transactions on communications*, *38*(1), 82–93.

Marcellin, M. W., Lepley, M. A., Bilgin, A., Flohr, T. J., Chinen, T. T., & Kasner, J. H. (2002). An overview of quantization in jpeg 2000. *Signal Processing: Image Communication*, *17*(1), 73–84.

Martínez, J. M. (2003). *Mpeg-7 overview* (Tech. Rep.). IEC JTC1/SC29/WG11.

Mason, L., Baxter, J., Bartlett, P. L., & Frean, M. R. (2000). Boosting algorithms as gradient descent. In *Advances in neural information processing systems* (pp. 512–518).

McAndrew, C. (2020). *The Art Market 2020. An Art Basel & UBS Report.*

Meagher, D. (1982). Geometric modeling using octree encoding. *Computer graphics and image processing*, *19*(2), 129–147.

Mei, J., & Moses, M. (2002). Art as an investment and the underperformance of masterpieces. *American Economic Review*, *92*(5), 1656–1668.

Mei, J., & Moses, M. (2005). Vested interest and biased price estimates: Evidence from an auction market. *The Journal of Finance*, *60*(5), 2409–2435.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., ... Van Gool, L. (2005). A comparison of affine region detectors. *International journal of computer vision*, *65*(1-2), 43–72.

Mojsilovic, A., & Soljanin, E. (2001). Color quantization and processing by fibonacci lattices. *IEEE transactions on image processing*, *10*(11), 1712–1725.

Mokrzycki, W. S., & Tatol, M. (2011). Color difference E - A survey. *Machine Graphics & Vision*, *20*(4), 383–411.

Moosburger, M. (2017). *Colour Labelling of Art Images Using Colour Palette Recognition* (Unpublished doctoral dissertation). Ludwig-Maximilians-Universität München.

Morse, B. S., Thornton, D., Xia, Q., & Uibel, J. (2007). Image-Based Color Schemes. In *2007 ieee international conference on image processing* (pp. III – 497–III – 500). IEEE. Retrieved from `http://ieeexplore.ieee.org/document/4379355/` doi: 10.1109/ICIP.2007.4379355

Munisami, T., Ramsurn, M., Kishnah, S., & Pudaruth, S. (2015). Plant leaf recognition using shape features and colour histogram with k-nearest neighbour classifiers. *Procedia Computer Science*, *58*, 740–747.

Obrador, P. (2006). Automatic color scheme picker for document templates based on image analysis and dual problem. *Digital Publishing*, *6076*, 607609. doi: 10.1117/12.647075

Ohm, J.-R., Cieplinski, L., Kim, H. J., Krishnamachari, S., Manjunath, B. S., Messing, D. S.,

& Yamada, A. (2002). The MPEG-7 color descriptors. In *Introduction to mpeg-7.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.148.7760{&}rep=rep1{&}type=pdf

On alternatives to lenna. (2017). *Journal of Modern Optics*, *64*(12), 1119-1120. Retrieved from https://doi.org/10.1080/09500340.2016.1270881 doi: 10.1080/09500340.2016.1270881

Ozturk, C., Hancer, E., & Karaboga, D. (2014). Color image quantization: a short review and an application with artificial bee colony algorithm. *Informatica*, *25*(3), 485–503.

Paeth, A. W. (Ed.). (1995). *Graphics Gems V.*

Paredes, R., Pérez, J., Juan, A., & Vidal, E. (2001). Local representations and a direct voting scheme for face recognition. In *In workshop on pattern recognition in information systems.*

Pesando, J. E., & Shum, P. M. (2008). The auction market for modern prints: Confirmations, contradictions, and new puzzles. *Economic Inquiry*, *46*(2), 149–159.

Phillips, S. J. (2002). Acceleration of k-means and related clustering algorithms. In *Workshop on algorithm engineering and experimentation* (pp. 166–177).

Plattner, S. (1998). A most ingenious paradox: The market for contemporary fine art. *American anthropologist*, *100*(2), 482–493.

Porwik, P., & Lisowska, A. (2004). The haar-wavelet transform in digital image processing: its status and achievements. *Machine graphics and vision*, *13*(1/2), 79–98.

Pownall, R. A. (2017). *TEFAF Art Market Report 2017* (Tech. Rep.).

Pownall, R. A., & Graddy, K. (2016). Pricing color intensity and lightness in contemporary art auctions. *Research in Economics*, *70*(3), 412–420.

Poynton, C. (2012). *Digital video and hd: Algorithms and interfaces.* Elsevier.

Puzicha, J., Buhmann, J. M., Rubner, Y., & Tomasi, C. (1999). Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of the seventh ieee international conference on computer vision* (Vol. 2, pp. 1165–1172).

Puzicha, J., Held, M., Ketterer, J., Buhmann, J., & Fellner, D. (2000, apr). On spatial quantization of color images. *IEEE Transactions on Image Processing*, *9*(4), 666–682. Retrieved from http://ieeexplore.ieee.org/document/841942/ doi: 10.1109/83.841942

Puzicha, J., Held, M., Ketterer, J., Buhmann, J. M., & Fellner, D. (1998). *On Spatial Quantization of Color Images* (Tech. Rep.). Bonn: Rheinische Friedrich-Wilhelms-Universität.

Quinlan, J. (2014). *C4. 5: programs for machine learning.* Elsevier.

Rayar, F. (2017). ImageNet MPEG-7 Visual Descriptors - Technical Report. , 21–23. Retrieved from http://arxiv.org/abs/1702.00187

Reinhard, E., & Pouli, T. (2011). Colour spaces for colour transfer. In *International workshop on computational color imaging* (pp. 1–15).

Renneboog, L., & Spaenjers, C. (2013). Buying beauty: On prices and returns in the art market. *Management Science*, *59*(1), 36–53.

Renneboog, L., & Van Houtte, T. (2002). The monetary appreciation of paintings: From realism to magritte. *Cambridge Journal of Economics*, *26*(3), 331–358.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Royo, C. V. (2010). *Image-Based Query by Example Using MPEG-7 Visual Descriptors* (Doctoral dissertation). Retrieved from http://hdl.handle.net/2099.1/9453

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Salahat, E., & Qasaimeh, M. (2017, mar). Recent advances in features extraction and description algorithms: A comprehensive survey. In *2017 ieee international conference on industrial technology (icit)* (pp. 1059–1063). IEEE. Retrieved from http://ieeexplore.ieee.org/document/7915508/ doi: 10.1109/ICIT.2017.7915508

Salmon, F. (2012). *Artnet's silly indices.* https://www.reuters.com/article/idUS204436343920120524. ([accessed on March, the 14th, 2020])

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210–229.

Schaefer, G. (2014, dec). Soft computing-based colour quantisation. *EURASIP Journal on Image and Video Processing*, *2014*(1), 8. Retrieved from https://jivp-eurasipjournals.springeropen.com/articles/10.1186/1687-5281-2014-8 doi: 10.1186/1687-5281-2014-8

Schaefer, G., Agarwal, P., & Celebi, M. E. (2018). Effective colour reduction using grey wolf optimisation. In J. M. R. Tavares & R. Natal Jorge (Eds.), *Vipimage 2017* (pp. 170–178). Cham: Springer International Publishing.

Schaefer, G., & Zhou, H. (2009). Fuzzy clustering for colour reduction in images. *Telecommunication Systems*, *40*(1-2), 17.

Schanda, J. (2007). *Colorimetry: understanding the cie system*. John Wiley & Sons.

Scheunders, P. (1997). A comparison of clustering algorithms applied to color image quantization. *Pattern Recognition Letters*, *18*(11-13), 1379–1384.

Schneider, F., & Pommerehne, W. W. (1983). Analyzing the market of works of contemporary fine arts: An exploratory study. *Journal of Cultural Economics*, 41–67.

Scorcu, A. E., & Zanola, R. (2011). The Right Price for Art Collectibles: A Quantile Hedonic Regression Investigation of Picasso Paintings . *The Journal of Alternative Investments*(January 2010), 110819050219003. doi: 10.3905/jai.2011.2011.1.012

Seçkin, A., & Atukeren, E. (2012). A heckit model of sales dynamics in turkish art auctions: 2005-2008. *Review of Middle East Economics and Finance*, *7*(3), 1–32.

Selim, S. Z., & Ismail, M. A. (1984). K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*(1), 81–87.

Shahbahrami, A., Borodin, D., & Juurlink, B. (2008). Comparison between color and texture features for image retrieval. In *Proc. 19th annual workshop on circuits, systems and signal processing (prorisc 2008), veldhoven, the netherlands.*

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317.

Sharma, G., & Bala, R. (2002). *Digital color imaging handbook*. CRC press.

Sharma, G., Wu, W., & Dalal, E. N. (2005, feb). The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, *30*(1), 21–30. Retrieved from `http://doi.wiley.com/10.1002/col.20070` doi: 10.1002/col.20070

Shih-Fu Chang, Sikora, T., & Purl, A. (2001, jun). Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *11*(6), 688–695. Retrieved from `http://ieeexplore.ieee.org/document/927421/` doi: 10.1109/76.927421

Shwartz-Ziv, R., & Armon, A. (2021). Tabular data: Deep learning is not all you need. *arXiv preprint arXiv:2106.03253*.

Smith, T., & Guild, J. (1931). The cie colorimetric standards and their use. *Transactions of the optical society*, *33*(3), 73.

Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bulletin of the Polish Academy of Sciences*, *4*(3), 801–804.

Stepanova, E. (2015). The impact of color palettes on the prices of paintings. *Empirical Economics*, 1–19.

Su, T., & Dy, J. G. (2007). In search of deterministic methods for initializing k-means and gaussian mixture clustering. *Intelligent Data Analysis*, *11*(4), 319–338.

Szeliski, R. (2011). *Computer Vision: Algorithms and Applications* (Vol. 73). London: Springer London. Retrieved from `http://link.springer.com/10.1007/978-1-84882-935-0` doi: 10.1007/978-1-84882-935-0

Szyszka, A., & Białowąs, S. (2019). Prices of works of art by living and deceased artists auctioned in poland from 1989 to 2012. *Economics and Business Review*, *5*(4), 112–127.

TEFAF. (2014). *TEFAF Art Market Report 2014*.

Triplett, J. (2004). Handbook on hedonic indexes and quality adjustments in price indexes.

Turnbull, D., & Elkan, C. (2005). Fast recognition of musical genres using rbf networks. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4), 580–584.

Ungerboeck, G. (1982). Channel coding with multilevel/phase signals. *IEEE transactions on Information Theory*, *28*(1), 55–67.

Velthuis, O. (2014). ARTRANK AND THE FLIPPERS: APOCALYPSE NOW? *Texte zur Kunst*(96).

Wei, X., Phung, S. L., & Bouzerdoum, A. (2016). Visual descriptors for scene categorization: experimental evaluation. *Artificial Intelligence Review*, *45*(3), 333–368. doi: 10.1007/s10462-015-9448-4

Wen, Q., & Celebi, M. E. (2011). Hard versus fuzzy c-means clustering for color quantization. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 118.

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29–39).

Witkowska, D. (2014). An application of hedonic regression to evaluate prices of polish paintings. *International Advances in Economic Research*, *20*(3), 281–293.

Witkowska, D., & Lucińska, A. (2015). Hedoniczny indeks cen obrazów sprzedanych na polskim rynku aukcyjnym w latach 2007–2013. *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Finanse. Rynki finansowe. Ubezpieczenia (75 Rynek kapitałowy: skuteczne inwestowanie)*,

515–527.

Wittgenstein, L. (1977). *Remarks on colour*. Oxford: Blackwell.

Woerheide, W., & Persson, D. (1992). An index of portfolio diversification. *Financial services review*, *2*(2), 73–85.

Wójtowicz, A. (2019). *Aukcje młodej sztuki w świetle teorii Pierre'a Bourdieu* (Unpublished doctoral dissertation). Akademia Sztuk Pięknych w Warszawie.

Wong, C. Y., Jiang, G., Rahman, M. A., Liu, S., Lin, S. C.-F., Kwok, N., ... Wu, T. (2016). Histogram equalization and optimal profile compression based approach for colour image enhancement. *Journal of Visual Communication and Image Representation*, *38*, 802–813.

Worthington, A. C., & Higgs, H. (2006). A note on financial risk, return and asset pricing in australian modern and contemporary art. *Journal of Cultural Economics*, *30*(1), 73–84.

Xiang, Z. (2007). Color quantization. In T. F. Gonzalez (Ed.), *Handbook of approximation algorithms and metaheuristics.* Chapman and Hall/CRC.

Żaglewska, D. (2016). Wpływ kapitału kulturowego na polski rynek antykwaryczny.

Zboroń, H. (2018). Sztuka a ekonomia. o ekonomicznych wyborach artystów. *Studia Ekonomiczne*, *371*, 137–149.

Zorloni, A. (2005). Structure of the Contemporary Art Market and the Profile of Italian Artists. *International Journal of Arts Management*, *8*(1), 61–71.

Appendix A1

Supplementary Tables and Figures

Table A.1: Feature importances for the Young Art dataset as mean dropout RMSE for models after 50 permutations.

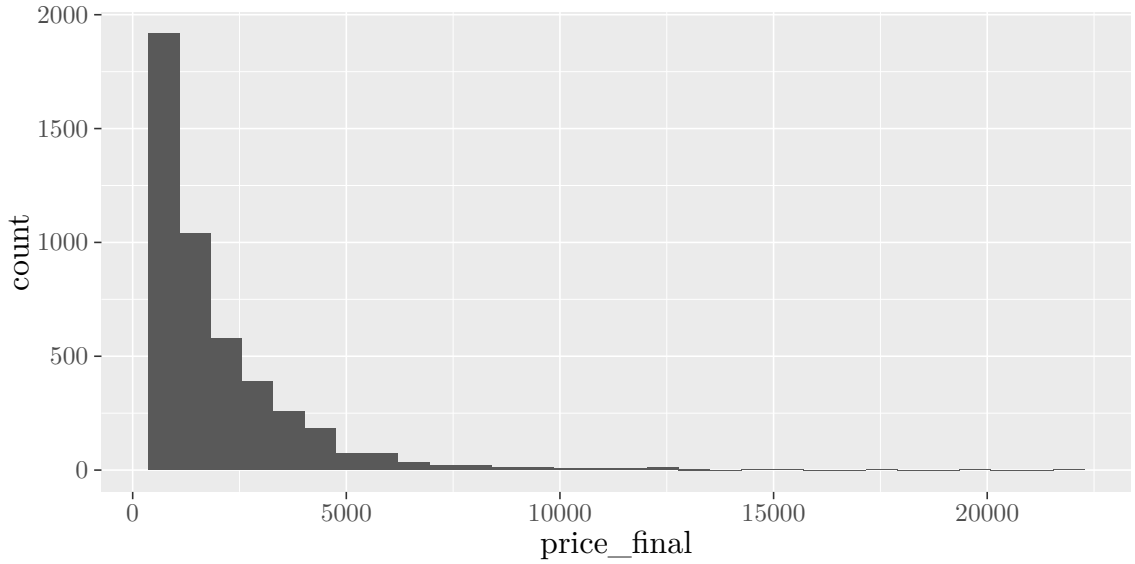
variable	Linear Model	XGBoost Model
full model	0.6845071	0.3254843
R214_G216_B213	0.6845071	0.3681401
R168_G168_B167	0.6845129	0.3870852
R208_G57_B36	0.6845241	0.3534719
R72_G120_B168	0.6845901	0.3555530
R122_G179_B202	0.6845994	0.3650241
R105_G84_B58	0.6846432	0.3757973
R122_G46_B38	0.6846603	0.3745327
R107_G114_B116	0.6846851	0.3728666
R52_G58_B132	0.6847072	0.3557020
R200_G176_B137	0.6849897	0.3685750
R214_G192_B55	0.6851225	0.3643169
R191_G108_B127	0.6853089	0.3781799
R192_G129_B57	0.6856487	0.3983196
colourfulness	0.6857518	0.3800922
R96_G148_B85	0.6858868	0.3783399
R56_G62_B70	0.6862403	0.3766801
technique	0.6863414	0.3521599
R30_G27_B27	0.6865752	0.3858511
auction_date_year	0.6898069	0.3847360
size	0.7271620	0.5205881
auction_house	0.7375620	0.4507589

Table A.2: Feature importances for the Top 10 Painters dataset as mean dropout RMSE for models after 50 permutations.

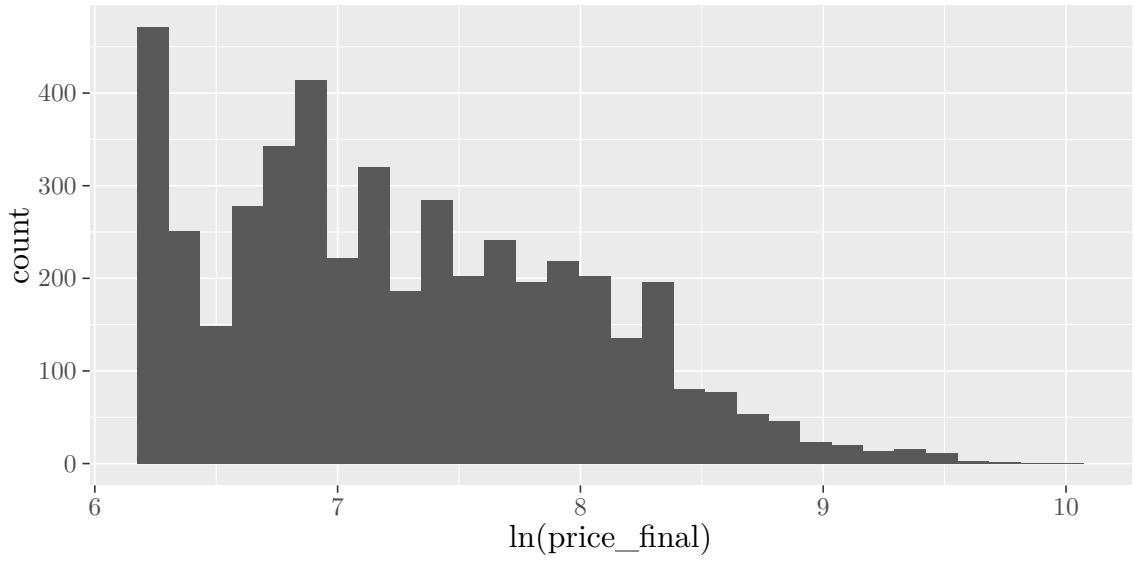
variable	Linear Model	XGBoost Model
full model	0.5419354	0.2878372
R212_G211_B203	0.5419354	0.3290035
R173_G170_B158	0.5420194	0.3014022
R185_G156_B123	0.5420426	0.3027465
R86_G102_B106	0.5422989	0.3047685
R185_G63_B49	0.5423681	0.2978281
R151_G175_B185	0.5424377	0.3001250
R71_G67_B57	0.5425672	0.3049566
R111_G72_B50	0.5427234	0.3068266
R49_G75_B127	0.5430543	0.3086191
R40_G36_B35	0.5439558	0.3289583
R127_G115_B101	0.5439652	0.3103065
R181_G161_B92	0.5443219	0.3015055
R171_G115_B69	0.5446291	0.3047417
R202_G182_B59	0.5448412	0.2969699
R91_G142_B181	0.5459894	0.3009139
R101_G125_B81	0.5472462	0.3349658
technique	0.5491550	0.4382097
colourfulness	0.5535299	0.2991040
auction_date_year	0.5608957	0.3030017
auction_house	0.5938087	0.3768391
size	0.9668608	0.8360037
author	1.3201908	1.1385939

Table A.3: SHAP attributions on 25 random orderings for Karpiński's *Wiejska droga w Bronowicach*.

feature	min	q1	median	mean	q3	max
auction_date_year = 2015	0.00923	0.01495	0.01644	0.016484	0.01968	0.02199
auction_house = Rempex	-0.18960	-0.16104	-0.11693	-0.108378	-0.04887	-0.01261
author = Alfons Karpiski	-0.11423	-0.04762	0.10729	0.061344	0.17073	0.18142
colourfulness = 34.97	0.00136	0.00932	0.01800	0.015690	0.02117	0.02581
R101_G125_B81 = 0.5304	-0.15564	-0.13729	-0.10149	-0.085022	-0.04819	0.08073
R111_G72_B50 = 0.03156	0.01561	0.01919	0.02065	0.022283	0.02532	0.03187
R127_G115_B101 = 0.08238	-0.01235	0.00934	0.02356	0.023077	0.04100	0.04982
R151_G175_B185 = 4.744e-05	0.00116	0.01402	0.01819	0.018174	0.02305	0.03043
R171_G115_B69 = 0.01286	0.00132	0.01143	0.03485	0.029067	0.04197	0.04794
R173_G170_B158 = 0.1148	0.00309	0.00649	0.00876	0.008676	0.01116	0.01361
R181_G161_B92 = 0.04557	-0.01556	-0.00779	0.00217	0.006117	0.02122	0.02923
R185_G156_B123 = 0.08263	0.01862	0.02838	0.04035	0.038336	0.04953	0.05441
R185_G63_B49 = 3.713e-05	-0.00255	0.00762	0.01197	0.014552	0.02404	0.02995
R202_G182_B59 = 2.682e-05	0.00231	0.00590	0.00708	0.007001	0.00834	0.01158
R212_G211_B203 = 0.001735	-0.10048	-0.07668	-0.06951	-0.067230	-0.06055	-0.02934
R40_G36_B35 = 0.006854	-0.07611	-0.06831	-0.05597	-0.055903	-0.04469	-0.03166
R49_G75_B127 = 0	-0.02003	-0.01296	0.00404	0.000373	0.01096	0.02257
R71_G67_B57 = 0.08545	0.01034	0.01395	0.01625	0.016651	0.01853	0.02587
R86_G102_B106 = 0.005586	-0.01031	-0.00409	0.00176	0.002231	0.00720	0.01750
R91_G142_B181 = 3.713e-05	-0.00208	0.00532	0.00907	0.008687	0.01318	0.01641
size = 30.32	-0.73342	-0.67413	-0.61062	-0.620856	-0.55570	-0.50533
technique = oil, paperboard	-0.18027	-0.17633	-0.14590	-0.083786	0.02063	0.03122

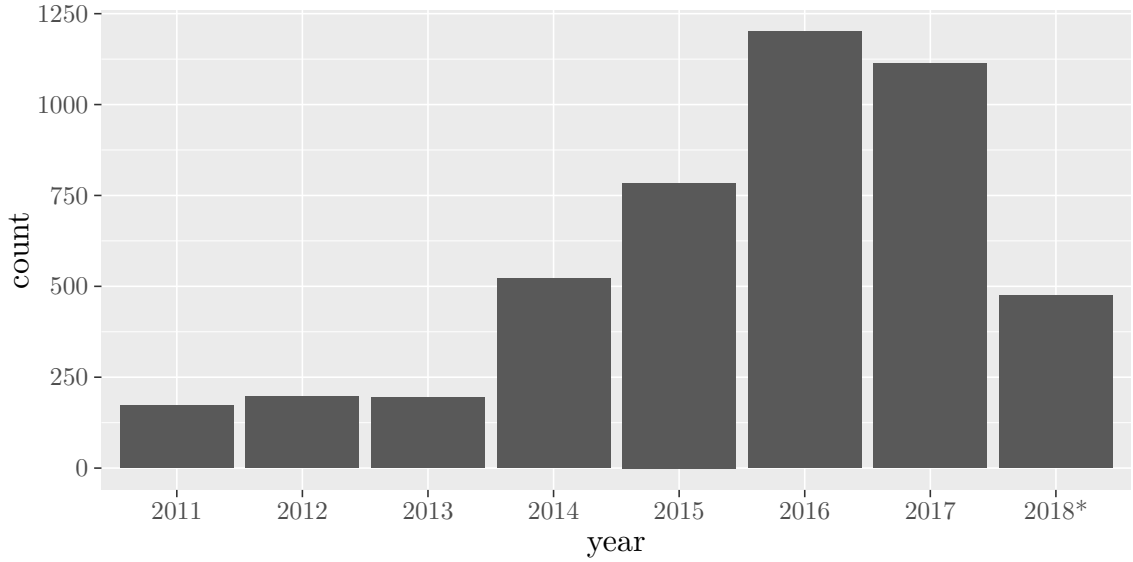


(a) Histogram of prices for the Young Art dataset

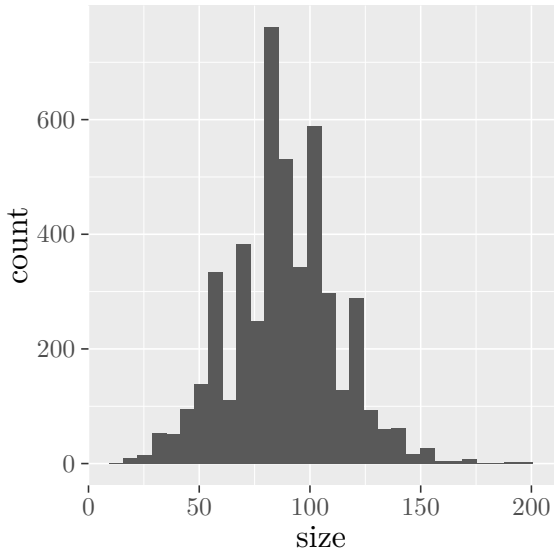


(b) Histogram of the natural logarithm of prices for the Young Art dataset

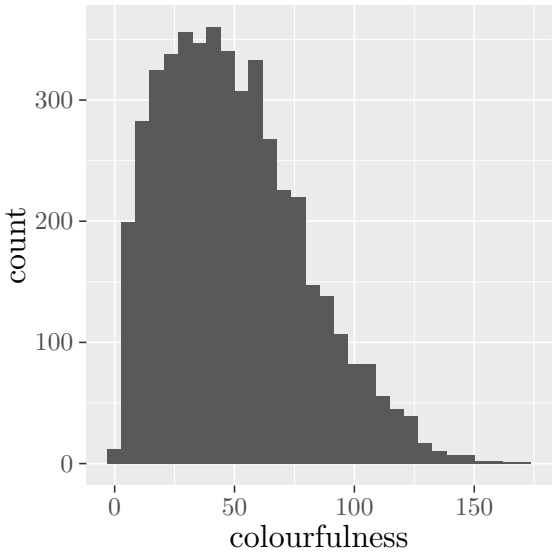
Figure A1: A comparison of price distribution histograms before and after taking natural logarithm in the Young Art Dataset.



(a) Year



(b) Size



(c) Colourfulness

Figure A2: Year, size and colourfulness histograms in the Young Art Dataset.

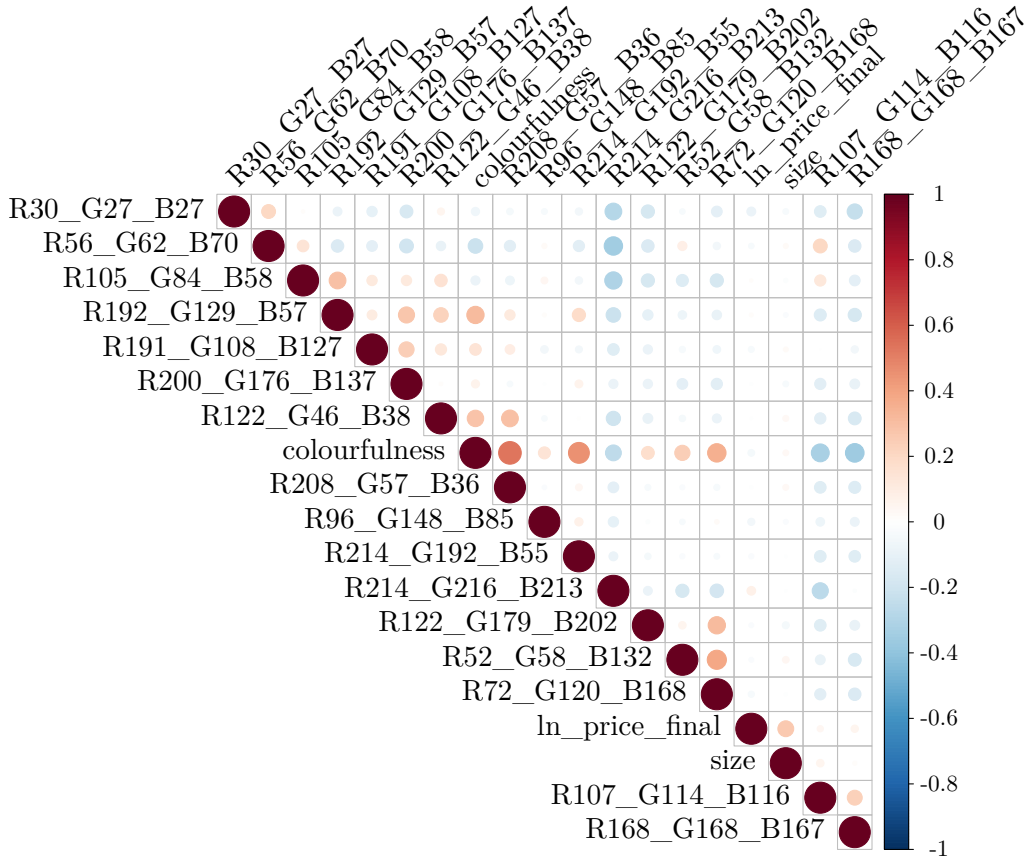
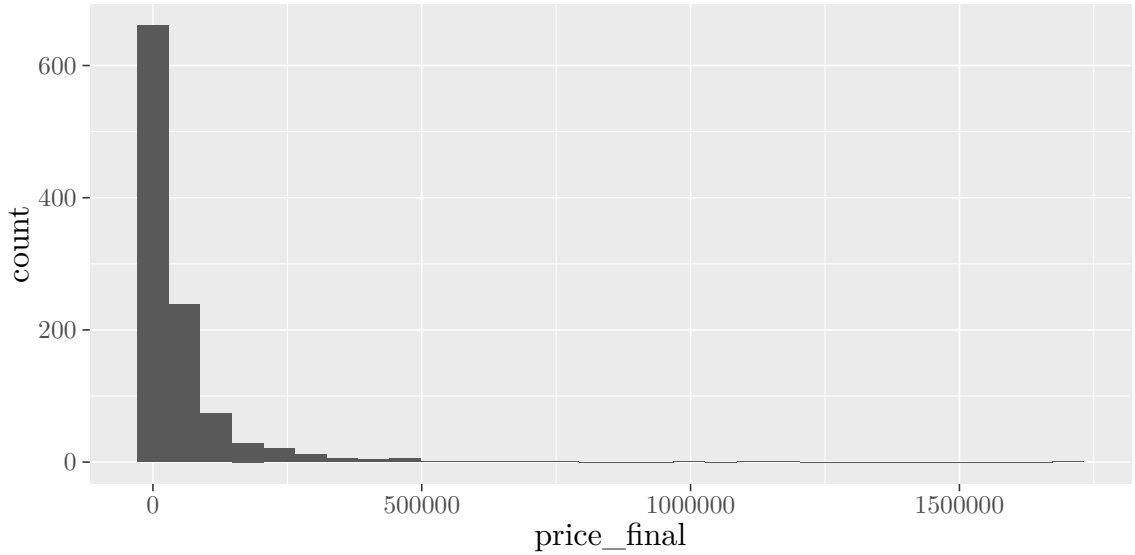
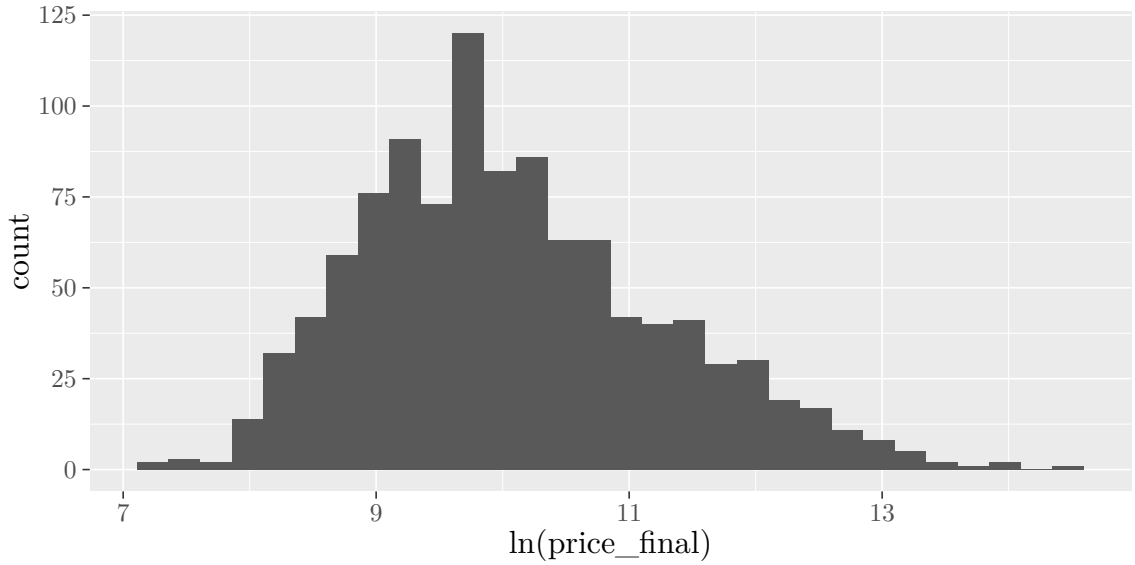


Figure A3: Correlation matrix for the Young Art dataset.

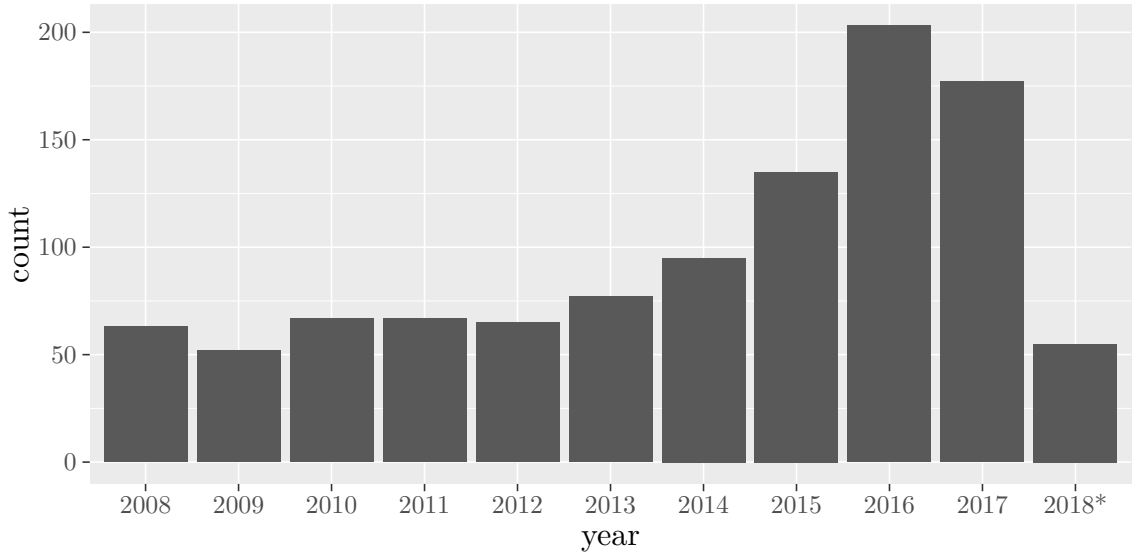


(a) Histogram of prices for the Top 10 Painters dataset

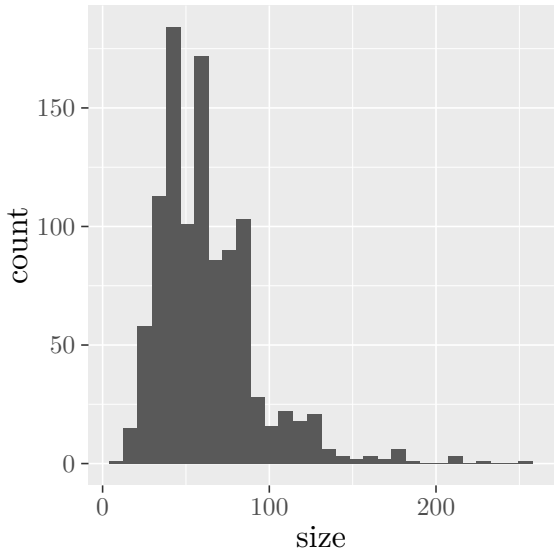


(b) Histogram of the natural logarithm of prices for the Top 10 Painters dataset

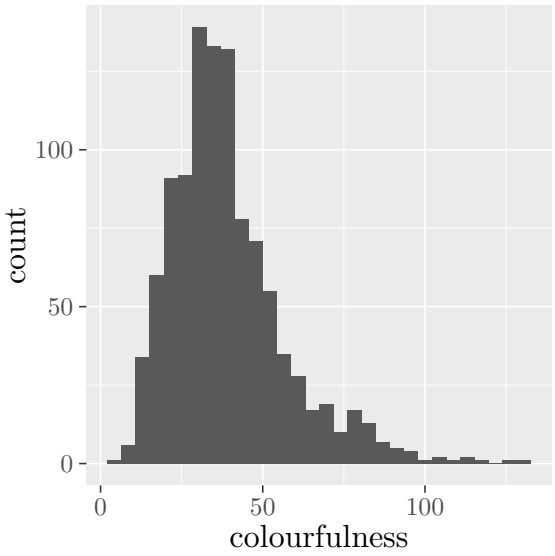
Figure A4: A comparison of price distribution histograms before and after taking natural logarithm in the Top 10 Painters Dataset.



(a) Year



(b) Size



(c) Colourfulness

Figure A5: Year, size and colourfulness histograms in the Top 10 Painters Dataset.

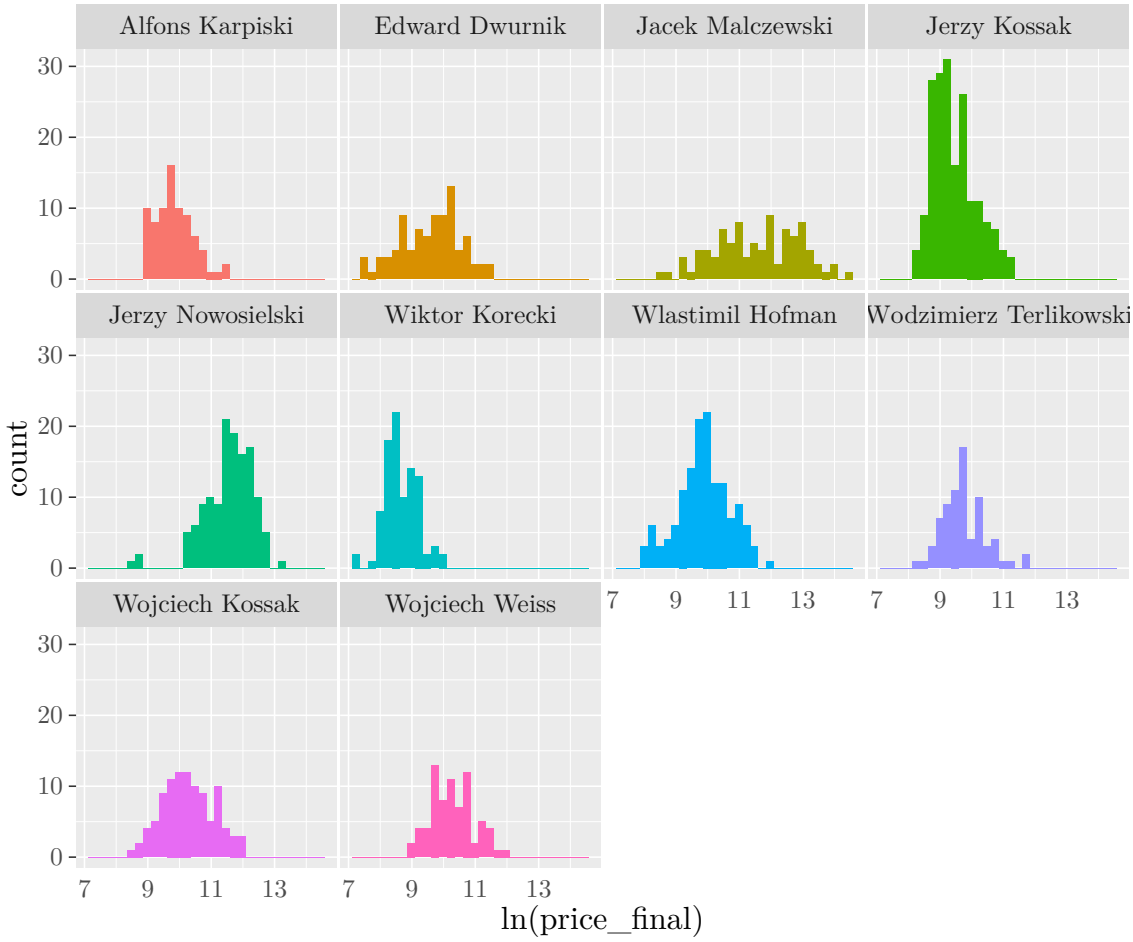


Figure A6: Price per author in the Top 10 Painters dataset.

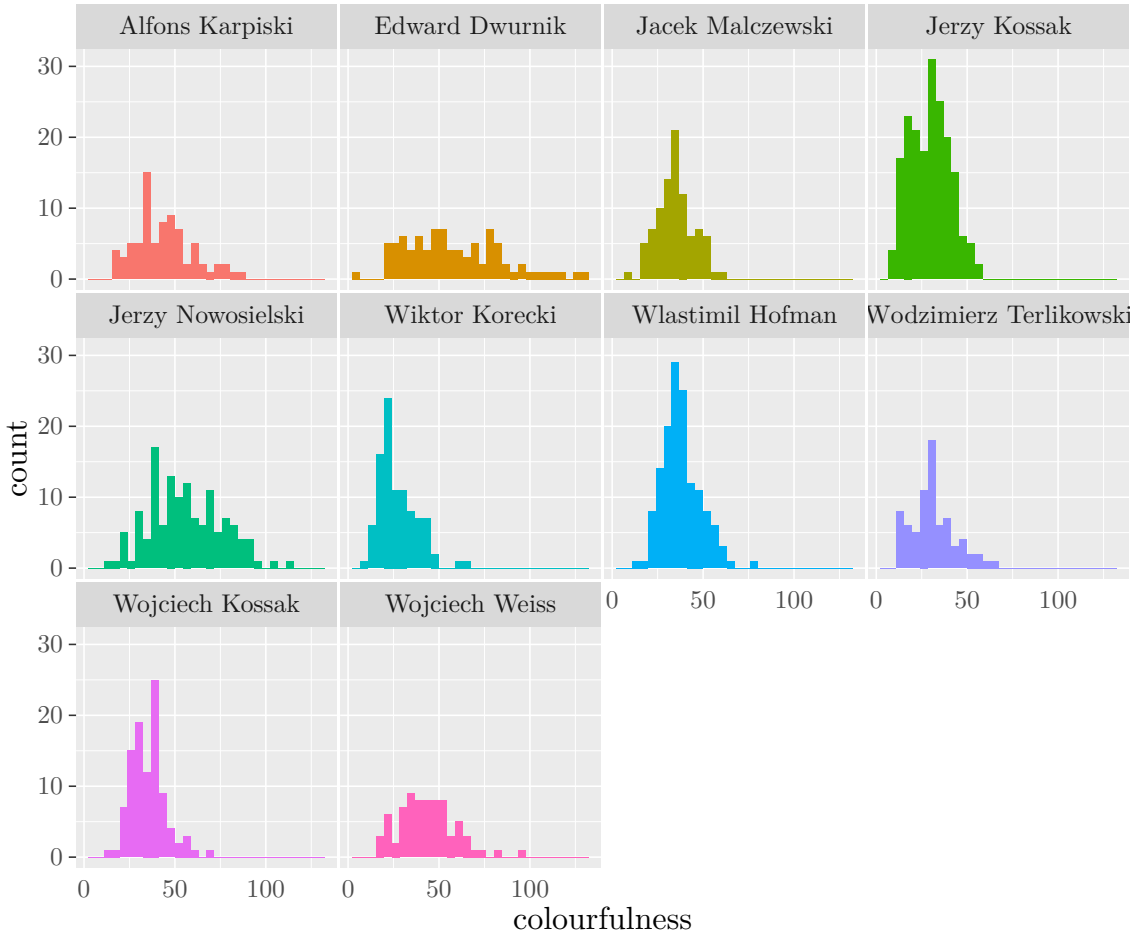


Figure A7: Colourfulness per author in the Top 10 Painters dataset.

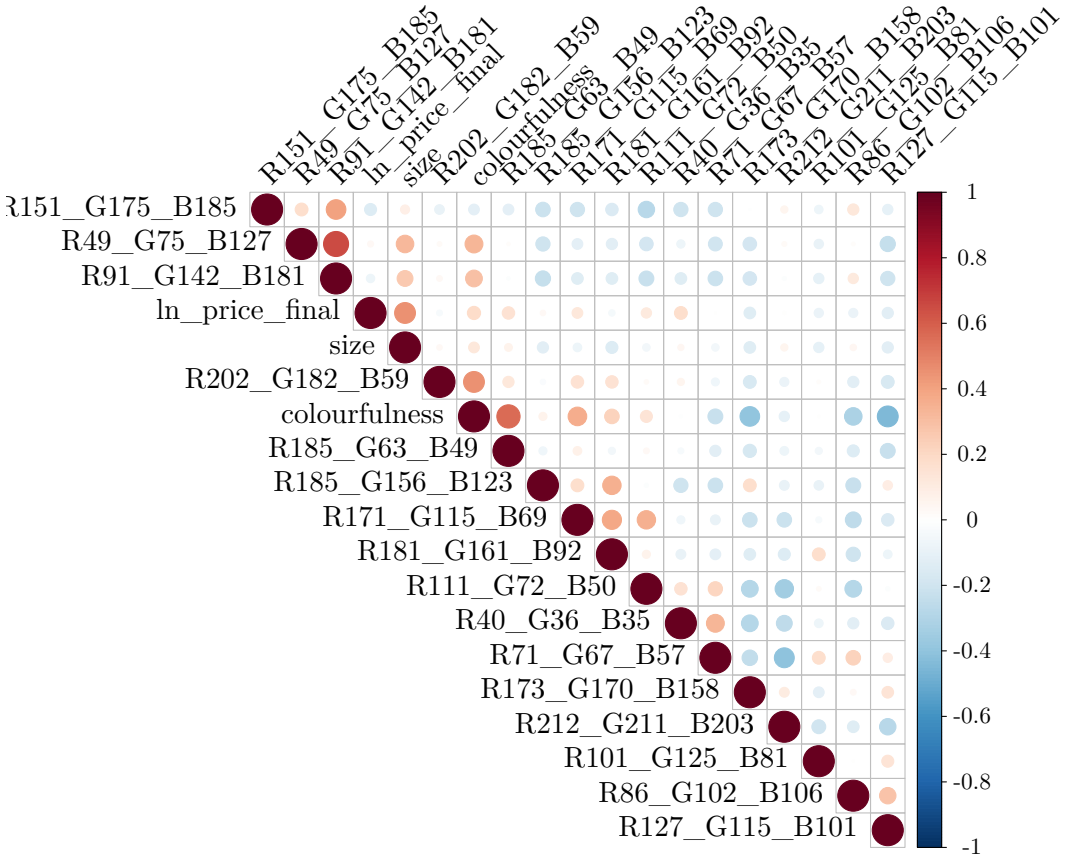


Figure A8: Correlation matrix for the Top 10 Painters dataset.

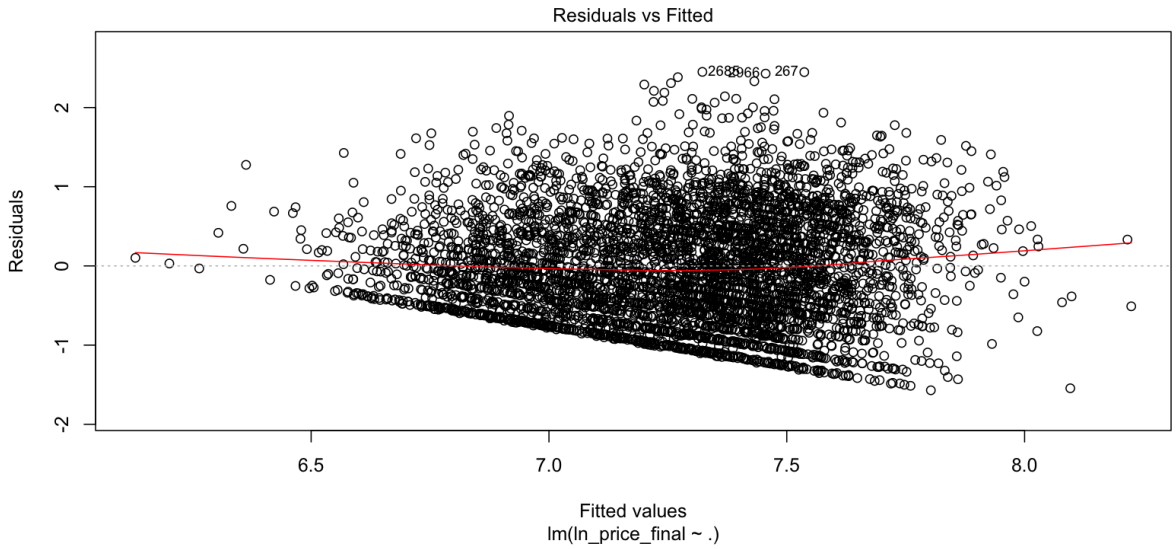


Figure A9: Regression diagnostics – residuals vs fitted plot for the linear model for the Young Art dataset.

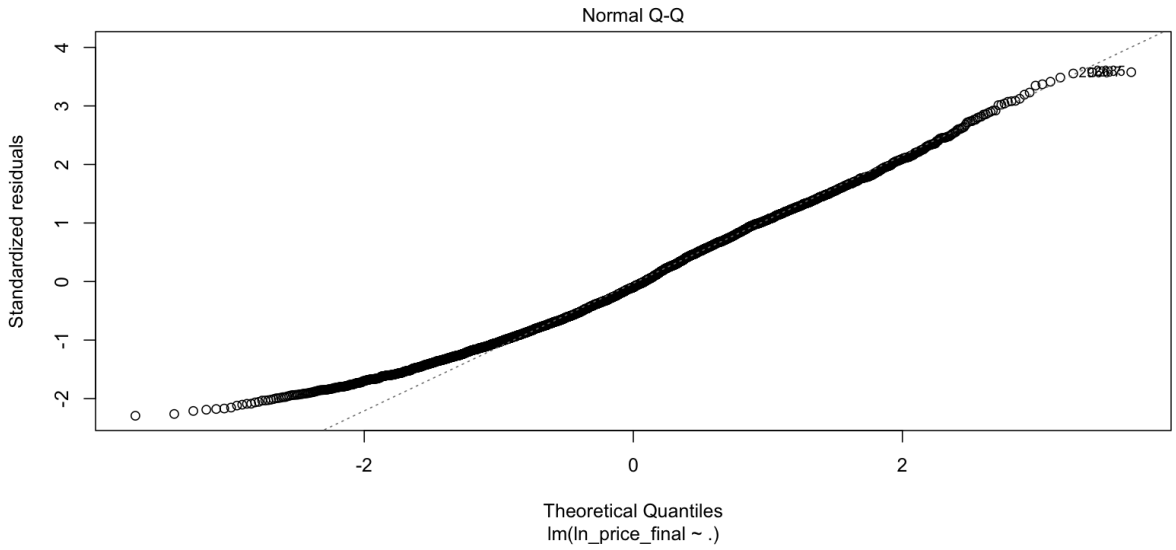


Figure A10: Regression diagnostics – normal quantile-quantile plot for the linear model for the Young Art dataset.

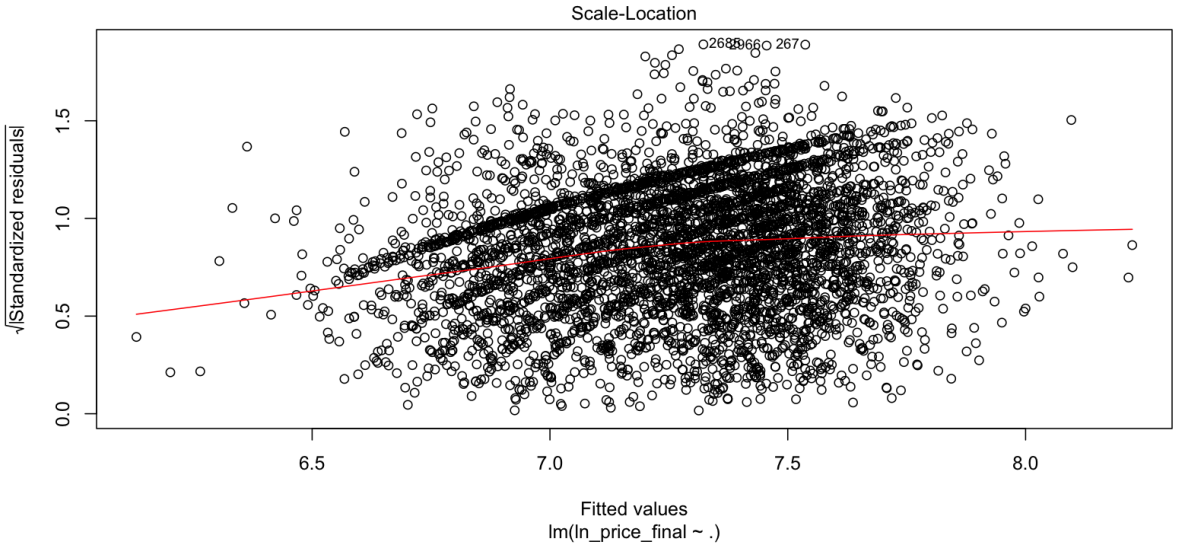


Figure A11: Regression diagnostics – homoscedasticity-assessing scale-location plot for the linear model for the Young Art dataset.

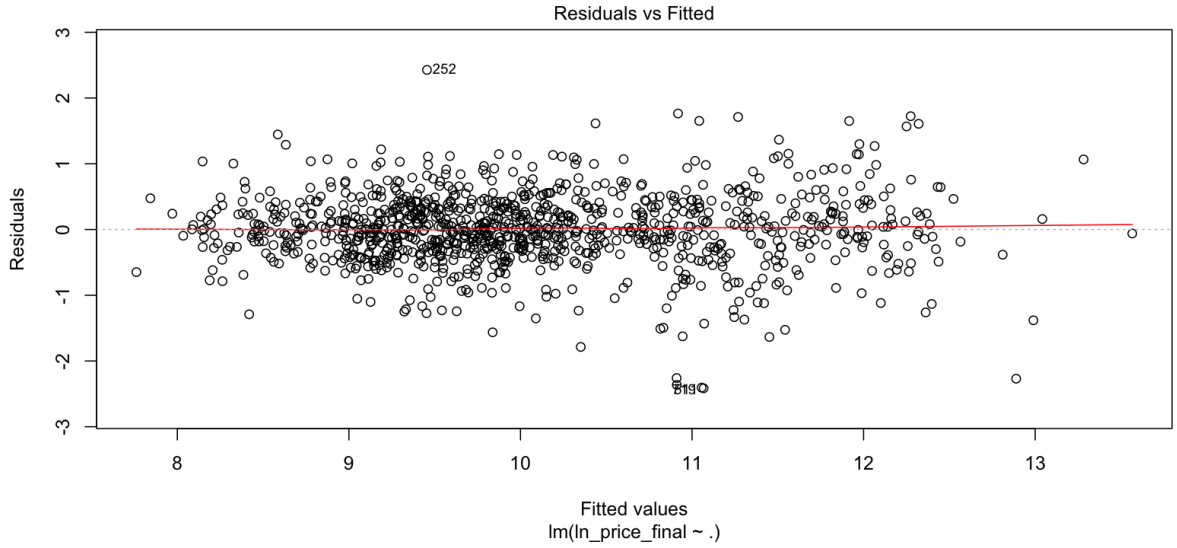


Figure A12: Regression diagnostics – residuals vs fitted plot for the linear model for the Top 10 Painters dataset.

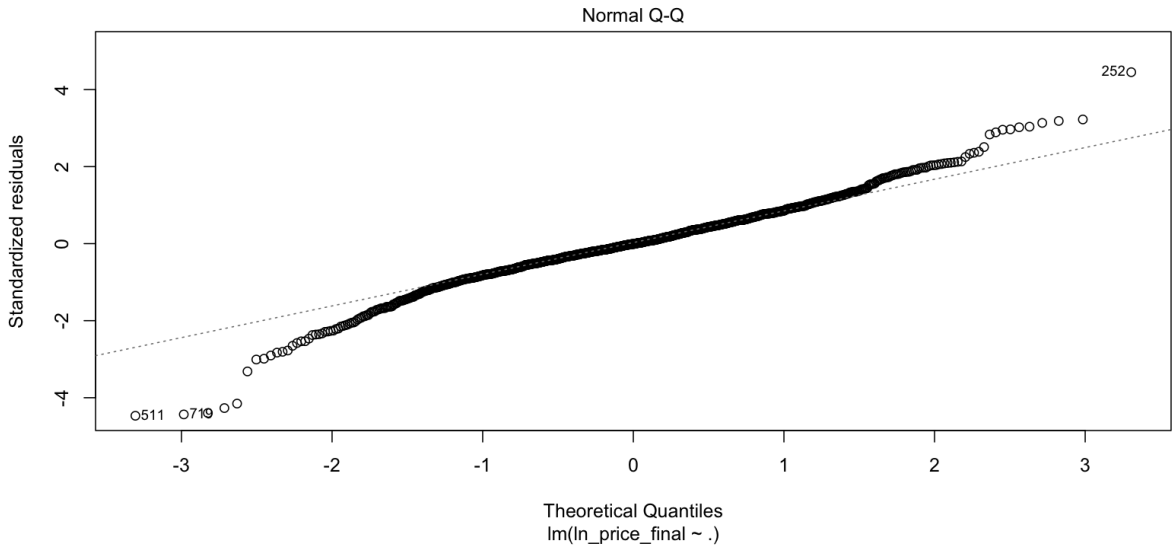


Figure A13: Regression diagnostics – normal quantile-quantile plot for the linear model for the Top 10 Painters dataset.

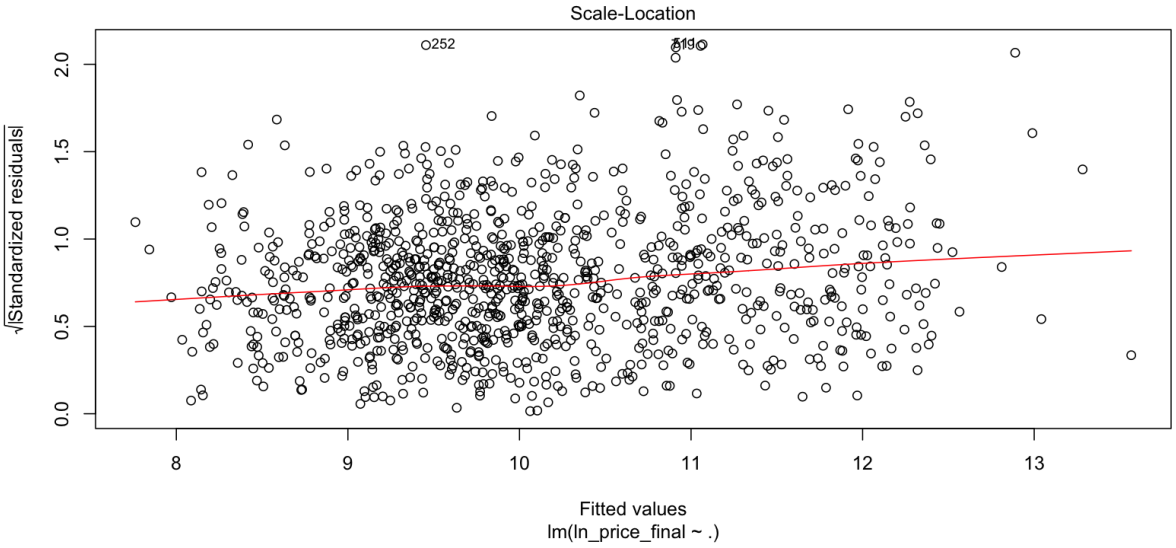


Figure A14: Regression diagnostics – homoscedasticity-assessing scale-location plot for the linear model for the Top 10 Painters dataset.