

Department of Information Science and Technology

Back to the past to charter the vinyl electronic market

Sara Carolina Augusto Lousão

Dissertation presented in partial fulfillment of the requirements for the Degree of
Master in Management of Information Systems

Supervisor:

Doctor Pedro Ramos, Associate Professor, ISCTE-IUL

Co-supervisor:

Doctor Sérgio Moro, Assistant Professor, ISCTE-IUL

September 2019

Acknowledgements

The opportunity to do this kind of research work and truly learn how to apply data mining to real world data is something I have been wanting to accomplish for a long time. Being able to write this thesis and to go through the entire process was a demanding yet rewarding experience. Therefore, I am deeply grateful to all the people who supported me throughout one of the most important phases in my life.

First of all, I would like to express my gratitude to both my supervisors, Professor Pedro Ramos and Professor Sérgio Moro for their guidance and their constant availability and prompt feedback, which were crucial to accomplishing this study's goals.

I would like to thank my parents and my sister for their unconditional support, as well as the opportunities they provided me, while always wanting the best for me and encouraging me to do better.

I am extremely grateful to Henrique, for his patience in stressful times and for sharing his ideas and inputs. Without him, I would not have been able to achieve this.

Resumo

Nas últimas décadas, o ritmo alucinante da evolução tecnológica e a massificação do uso de dispositivos digitais avançados, forçou múltiplas empresas e mercados inteiros a escolher entre reinventarem a sua estratégia ou perecerem perante a possibilidade de se tornarem obsoletos. Um ótimo exemplo de uma destas indústrias é o ramo do Entretenimento, onde setores inteiros foram substituídos por plataformas de conteúdo digitais.

Esta tese foca-se, no formato de média provavelmente mais icónico de todos os tempos, o vinil. Sendo um dos primeiros formatos para reprodução de áudio, foi criado por volta de 1920 e dominou o mercado até aos anos 80, altura em que a invenção do *Compact Disc* substituiu finalmente o vinil. Esta mudança foi, principalmente devida ao baixo custo de produção do CD, bem como, um formato mais portátil, facilitando assim a distribuição. Contudo, mais recentemente, o vinil recuperou alguma da sua popularidade, em grande parte por causa de colecionadores dedicados que criaram uma comunidade de entusiastas que mantiveram o formato vivo.

Com este estudo pretende-se perceber, quais dos fatores envolvidos na compra e venda, são os mais importantes para o valor de venda de um vinil. De modo a conseguir avaliar esta hipótese, foram criados quatro conjuntos de dados para representar registos antigos e recentes de dois géneros musicais diferentes, *Rock* e *Jazz*. Os dados de base foram obtidos através de *webscraping* do mercado *online* no site *Discogs* e no *ranking Hot 100* da *Billboard*. Durante todo o presente estudo, a metodologia de trabalho escolhida foi a CRISP-DM e os programas de *software* usados para a análise dos dados foram o *SAS Enterprise Guide* e o *SAS Enterprise Miner*. Esta abordagem revelou que, tanto a presença de um artista nos *charts*, como a editora correspondente pertencer a uma das “*Big Three*”, não garante que o preço dos discos seja muito elevado. Os resultados mostraram também que as variáveis que medem a popularidade tornam-se mais relevantes na “era” em que o género do disco é mais popular e que as grandes editoras têm vindo a perder quota de mercado para um número maior de editoras independentes.

Palavras-Chave: Vinil, Preço, Ciência dos Dados.

Abstract

For the past decades, the astounding rhythm of technological evolution and the massification of advanced digital devices, have forced several companies and entire market sectors to choose between reinventing themselves or perishing into obsolescence. A great example is the Entertainment industry, where entire sectors were replaced by digital content platforms.

This thesis focuses on, perhaps the most iconic media format of all time, the vinyl record. Vinyl was one of the first formats for audio reproduction, created around 1920, managing to become increasingly more popular until the 80's, when the invention of the Compact Disc finally replaced the vinyl. This was mainly due to the lower costs of production of the CD, as well as requiring less space and maintenance, becoming easier to distribute. Interestingly, vinyl made a small comeback, mainly due to avid collectors who unknowingly created a community that kept the format alive.

The goal of this study is to understand which factors, involved in the buying and selling of vinyl, influenced its price, with the initial hypothesis considering record labels and popular rankings to be some of the most contributing variables. To be able to evaluate this, four datasets were created in an endeavor to represent recent and past records of two different genres, Rock and Jazz, by extracting data from Discogs' marketplace and Billboard's Hot 100 chart. For this research, the chosen work methodology was CRISP-DM and the software programs for data analysis were SAS Enterprise Guide and SAS Enterprise Miner. Such approach allowed unveiling that an artist's presence in the charts and their labels belonging to one of the 'Big three', do not always dictate their records at highest prices. The results also showed that features which measure popularity become more relevant in the 'era' where the record's genre is more popular and that big record labels have been losing market share to an increasing number of independent labels.

Keywords: Vinyl record, Price, Data Science.

Index

ACKNOWLEDGEMENTS	I
RESUMO	II
ABSTRACT	III
INDEX	IV
LIST OF TABLES	V
LIST OF FIGURES	VI
LIST OF ACRONYMS AND ABBREVIATIONS	VII
CHAPTER 1 – INTRODUCTION	1
1.1. TOPIC AND RESEARCH PROBLEM	1
1.2. TOPIC MOTIVATION	1
1.3. HIPOTHESYS AND RESEARCH GOALS	2
1.4. METHODOLOGY	2
1.5. STRUCTURE AND ORGANIZATION OF THE DISSERTATION	2
CHAPTER 2 – LITERATURE REVIEW	3
2.1. VINYL ORIGIN AND REEMERGENCE	3
2.2. MARKET CHARACTERIZATION	4
2.3. THE IMPORTANCE OF RECORD LABELS IN THE 1960S – 1980S	5
2.4. CHARTS	6
2.5. ELECTRONIC MARKETS	7
2.6. DATA ANALYSIS OF E-MARKETS	2
2.7. RELATED LITERATURE	5
CHAPTER 3 – MATERIALS AND METHODS	7
3.1. METHODOLOGY	7
3.2. WEB SCRAPING	9
3.3. DATA TRANSFORMATIONS	14
3.3.1. <i>Discogs Transformations</i>	14
3.3.2. <i>Billboard Transformations</i>	16
3.4. DATA MODELING	17
3.4.1. <i>Partitioning</i>	18
3.4.2. <i>Models</i>	18
3.4.3. <i>Variable Selection</i>	19
3.4.4. <i>Model Comparison</i>	20
CHAPTER 4 – RESULTS AND DISCUSSION	21
4.1. DATA ANALYSIS RESULTS	21
4.2. DATA MODELING RESULTS	27
4.3. DISCUSSION	29
CHAPTER 5 – CONCLUSIONS AND RECOMMENDATIONS	31
5.1. CONCLUSIONS	31
5.2. RESEARCH LIMITATIONS	33
5.3. FUTURE RESEARCH	34
REFERENCES	35
APPENDIX	40

List of Tables

Table 1 - Methods and Algorithms in Data Mining. Adapted: Camilo & Silva (2009) ...	4
Table 2 - Data Mining Algorithm Applications	4
Table 3 - Contributions of related literature.	6
Table 4 - List of features.....	13
Table 5 - Conversions of media and sleeve condition.	15
Table 6 - Comparison of artist and label for Rock Past.....	22
Table 7 - Comparison of artist and label for Rock Present.	23
Table 8 - Comparison of artist and label for Jazz Past.	24
Table 9 - Comparison of artist and label for Jazz Present.	25
Table 10 - Labels ownership and dataset distribution examples.	26

List of Figures

Figure 1 - U.S. Music Revenues by Format, Adjusted for Inflation (U.S. Sales Database, 2017).....	3
Figure 2 - U.S. Music Revenues by Format (U.S. Sales Database, 2017).	4
Figure 3 - Worldwide vinyl sales from 1997 to 2013 (Richter F., 2014).	3
Figure 4 - U.S. Music Revenues by Format (U.S. Sales Database, 2017).	4
Figure 5 - U.S. Music Revenues by Format, Adjusted for Inflation (U.S. Sales Database, 2017).....	4
Figure 6 - Phases of the CRISP-DM (Wirth & Hipp, 2000).	7
Figure 7 – Location for the vinyl marketplace features extracted from Discogs.	11
Figure 8 - Location for the vinyl marketplace features in release page from Discogs...	11
Figure 9 - Location for the chart raking features from Billboard.	12
Figure 10 - Example diagram of the tasks applied.	18
Figure 11 - Variable importance.....	28

List of Acronyms and Abbreviations

API - Application Programming Interface

ASCII - American Standard Code for Information Interchange

BMG – Bertelsmann Music Group

CD – Compact Disc

CRISP-DM – Cross-Industry Standard Process for Data Mining

DJ – Disc Jockey

EMI – Electric and Musical Industries Ltd

Hi-Fi – High Fidelity

HTML - Hypertext Markup Language

HTTP - Hypertext Transfer Protocol

IBM – International Business Machines Corporation

LP – Long Playing

SAS – Statistical Analysis System

SEMMA – Sample Explore Modify Model and Assess

UK – United Kingdom

URL - Uniform Resource Locator

USA – United States of America

XML - Extensible Markup Language

XPath - XML Path Language

Chapter 1 – Introduction

1.1. Topic and Research Problem

The thesis will focus on studying the vinyl market, mainly covering the theme of price formation on past and current popular albums based on data scraped from available sources such as web platforms and charts. Consequently, the influence of an existent record label monopoly is taken into consideration to verify if its loss of power has been reflected in records sales and rankings.

The vinyl format origin dates back to the beginning of the XX century, being the most commonly used format until the 1980's where it was replaced by the Compact Disc. During its following years, the vinyl was kept alive by nostalgics and fans that carried on the legacy of collecting and trading. In the past ten years, the format's popularity has been rising, driven in part by the emergence of online communities and platforms that ease the process of searching and trading records.

It is this recent rise in popularity that requires a closer look to understand the evolution of the prices in the vinyl market since its peak, where only three record labels had more than 50% of the profits of total sales (Day, 2011), to the present day, where the shift in music delivery methods has changed deeply with online platforms and electronic commerce which begin to undermine the monopoly preestablished by the "Big Three".

1.2. Topic Motivation

The choice of this topic was mainly motivated by an interest in data mining and music. Ever since music availability has been democratized and songs have become a daily part of people's quotidian, there has been a huge market surrounding this form of entertainment. With the need to appease the audience's demands, it was fundamental to achieve production and distribution methods that could reach millions of listeners.

This was the problem solved by the big record labels that offered the artists the possibility to focus only on their music, delegating the "technical concerns" for a part of the profits, thus creating a dependency between artist and label that has lasted to the present day, allowing bigger players to charge higher fees as their services had a greater ease in reaching larger audiences.

Nowadays, the vinyl has lost the better part of its popularity and the need for production and mass distribution is mitigated by the creation of online streaming music services. As such, the analysis of the vinyl market until the current day allows a special perspective on the first reproducible music format and chart its course through data analysis software.

1.3. Hypothesis and Research Goals

With the recent reemergence of vinyl, along with the advent of the Internet, much is changing in this old market. A vinyl record can be bought for anything between five and tens of thousands of dollars and, where in the old days the good deals were found in garage sales and old stores, nowadays looking for or purchasing a particular record is just a few clicks away.

To analyze and better understand the differences between past and present, it is necessary to step back and take a look at the bigger picture. This thesis proposal is to research how the price evolution of vinyl records were affected by, first and foremost, the popularity achieved during their peak periods, registered in the charts of the corresponding time and, secondly, the influence of their record labels.

As such, the main goals of the following dissertation are to provide a deeper knowledge into what shapes price in the vinyl electronic commerce while taking into consideration sales and popularity, as well as verifying if the past influence of a powerful monopoly of three record labels has started to lose its strength, through the application of data mining to data gathered from available sources.

1.4. Methodology

The goals proposed above will require data from record sales and their corresponding artist, genre, label, etc. which is freely available online. The first step will be to gather all the possibly relevant data through web crawlers and save it in a database in a standardized model for the different sources.

After harnessing the raw information and applying some small transformations to fit it all in the single database model, this dataset will be processed by Data Mining

algorithms to extract the underlying knowledge on price formation. These results will then be evaluated to determine their accuracy and, therefore, their validity.

1.5. Structure and Organization of the Dissertation

In Chapter 1, a comprehensive introduction of the approached topics as well as the methodologies that will be used and the research goals is provided, enabling a quick overview on the main focus of the document and its constituent parts. Chapter 2 will, subsequently, contextualize the topic by reviewing the existent literature, starting from the origin of the vinyl record to its recent reemergence, while also covering more technical aspects of the topic as electronic markets and their data analysis. Chapter 3 addresses the materials and methods used, describing in detail the decisions made in the process of gathering and mining data. In Chapter 4, the results obtained will be evaluated and their meaning discussed to consider different perspectives on the possible conclusions. Finally, Chapter 5 will be formed by the conclusions extracted from the study while addressing research limitations and unforeseen scenarios, closing with possible next steps for future related work.

Chapter 2 – Literature Review

2.1. Vinyl origin and reemergence

Vinyl records were originally created in the early 20th century and dominated the music market up until mid 1980's (Figure 1), even after the appearance of the portable Compact Cassettes in the 1960's. As the dominating music format, with an approximate average of 2 billion dollars revenue in the U.S. and a market share of approximately 70% between 1973 and 1983 (U.S. Sales Database, 2017), vinyl brought to life the albums of artists such as The Beatles, Pink Floyd, The Rolling Stones, Miles Davis and John Coltrane, among many iconic others. While the invention of the cassette may have started vinyl's demise due to its portability (Plasketes, 1992), it was not until the 1990's with the invention and rise in popularity of Compact Discs (CD's), that vinyl's market dominance came to an end, going from the most popular of music formats to being purchased and used mostly by nostalgics, collectors and Disc Jockeys (DJ's) as will be reviewed in this chapter.

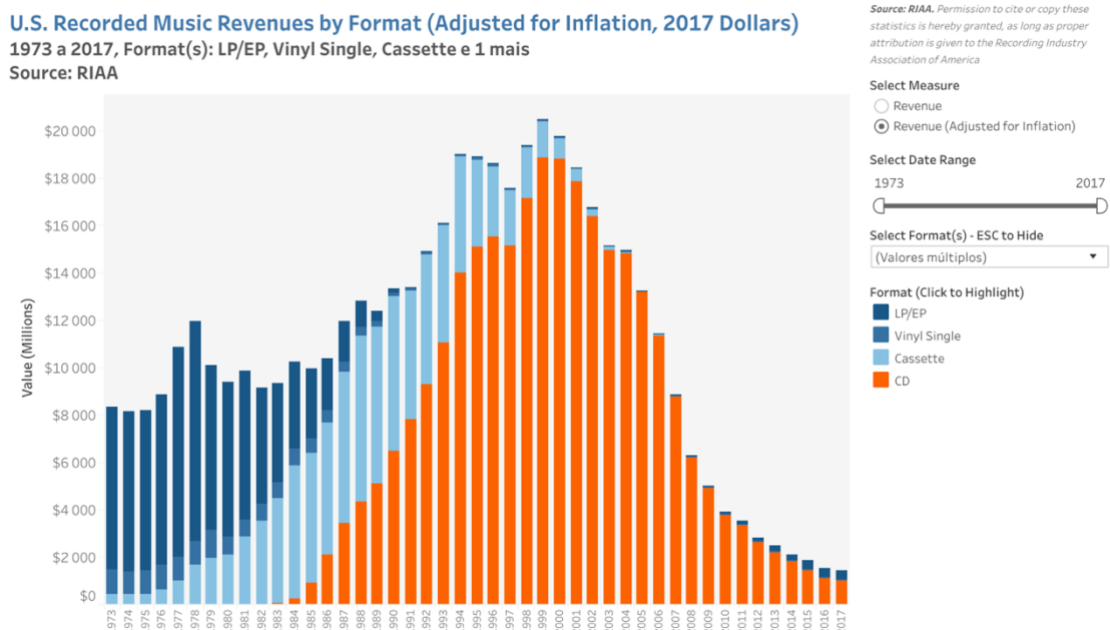


Figure 1 - U.S. Music Revenues by Format, Adjusted for Inflation (U.S. Sales Database, 2017).

It was mainly through the persistence of record stores and the involvement of the surviving community in events like Record Store Day, an annual event created in the USA, that vinyl records reached, in 2008, its highest sales number since 1991 and reemerged as one of the preferred music formats (Figure 2). This went against the

expectations for an era where MP3 and online streaming platforms, like YouTube, already existed. Nonetheless, the sudden boom in vinyl sales could be explained by their intrinsic high fidelity (hi-fi) sound, its physicality, tactile and aesthetic appeal when compared to digital audio files and, last but not least, their vintage feel in a market that was and still is partially driven by old-fashion consumerism (Sarpong, D., Dong, S., & Appiah, G., 2016).

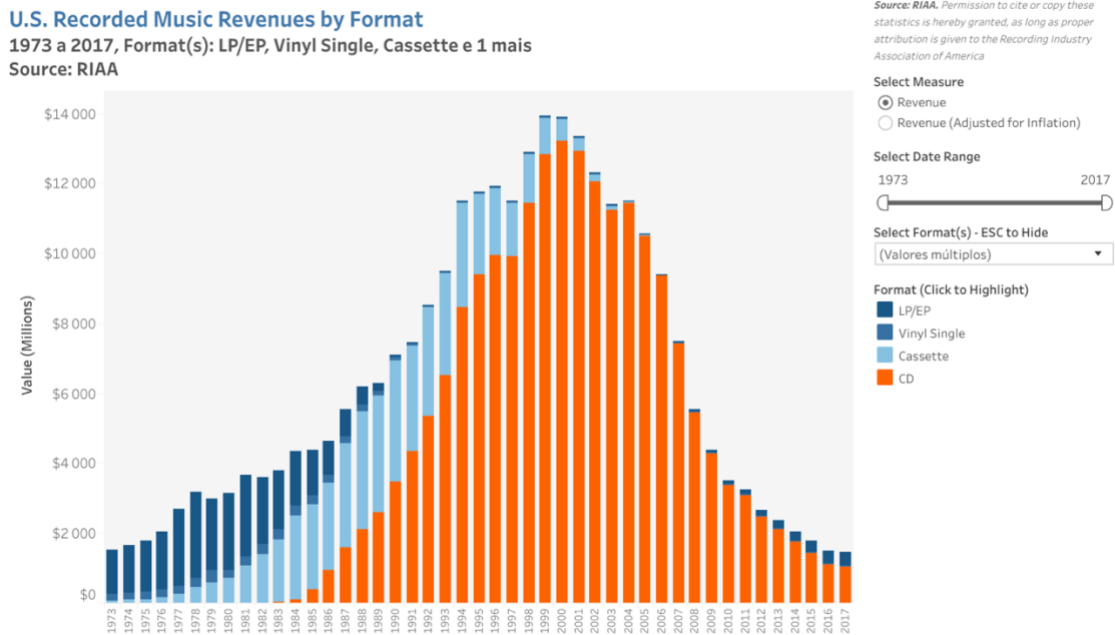


Figure 2 - U.S. Music Revenues by Format (U.S. Sales Database, 2017).

2.2. Market characterization

According to Sarpong *et al.* (2016), to better understand the vinyl market and its consumers, it is possible to separate and characterize its audience into diverse groups, namely, the baby boomers (1946 - 1964) who grew up during vinyl's peak in popularity and chose to keep collecting records such as nostalgic collectors, label bosses, record dealers and DJ's, and the generation Y (1981 - 1996) who grew up in a digital era surrounded by CD's and digital audio files but still preferred vinyl as their predilect music format contributing largely to vinyl sales. Some examples of these generation Y consumers are the young enthusiasts, the new buyers, the romantic musicians and "sighing skeptics". It is through these diverse groups of enthusiasts that the vinyl subculture finds its representation and, only through their passion, can a format that has seen little to no technological innovation, survive and grow in popularity.

Another very important part of analyzing the vinyl subculture and subjacent market is understanding which genres are the most listened to. This knowledge allows a better focus on the fractions of the marketplace which will yield the most relevant data. Among the many music genres existent nowadays, the highest registered vinyl record sales during 2018 in the USA (Watson, 2019) were, from first to last, Rock (41.7%), Pop (25.6%), R&B (7.9%), Rap/Hip-Hop (6.6%), Stage & Screen (5.7%) and Jazz (4.3%), with the remaining 9 other genres having a 8.3% share of the market.

After understanding the genre distribution, another relevant factor is sales volume and how it has evolved throughout the years. Since 2008, when vinyl reemerged as a relevant music format in the industry, record sales have seen dramatic increases each year. According to the British Phonographic Industry (an entity representing the UK's recorded music industry), in 2011, sales rose by 43.7%, with what was called a "modest resurgence" (Bartmanski & Woodward, 2015) and, in 2013, the UK watched vinyl sales soar by 101%, corresponding to over 780,000 records sold, peaking for the first time in 15 years (Sarpong *et al.*, 2016).

In the USA, during 2010, the popular pop and rock magazine Rolling Stone ran with the headline that read "Vinyl sales increase despite industry slump". This article used Nielsen's SoundScan charts to report that "though overall album sales dropped by 13% in 2010, sales of vinyl increased by 14% over the previous year, with around 2.8 million units sold" (Bartmanski & Woodward, 2015). This ominous turn in album sales versus vinyl sales culminated in 2014, when the USA witnessed a staggering rise of 52% of records sold in comparison to the previous year (Negus, 2015), with the numbers from Nielsen's SoundScan mid-year report showing that, in the first semester alone, 4 million units had already been sold (Sarpong *et al.*, 2016). Global sales accompanied this trend, reaching an 18-year high with labels, record stores and vinyl only clubs sustaining themselves by attracting a small but devoted niche of music fans as shown in Figure 1-5 (Negus, 2015).

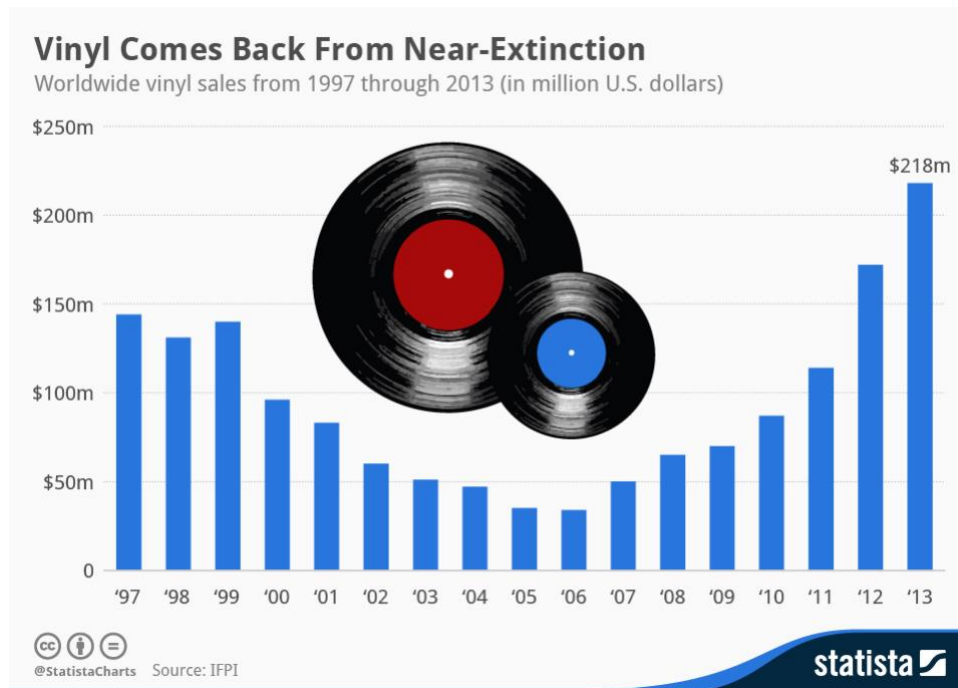


Figure 3 - Worldwide vinyl sales from 1997 to 2013 (Richter F., 2014).

Looking past country and worldwide sales figures, perhaps one of the greatest indicators of how people perceive vinyl as a “popular, ‘authentic’ way to experience the pop and rock canon” are the numbers of records sold for the top selling LP’s. For example, the record *Abbey Road* by The Beatles was the recordist in LP sales from 2009 to 2011, selling over 100,000 units all over the world, clearly showing how LP's have reemerged as a popular music format in the past decade (Bartmanski & Woodward, 2015).

As mentioned above, another great factor for the allure observed in this marketplace is the history behind vinyl and, with its authenticity comes the perspective of seeing these records as an investment in a collectible (Bartmanski & Woodward, 2015). An example of this can be found in an issue of the *New Musical Express* music magazine where it listed the top 20 vinyl releases which contained an original pressing of The Quarrymen (group that featured members of The Beatles, prior to the creation of the band) with an estimated worth of £100,000 as well as The Sex Pistols’ “God Save The Queen” single in an estimation of £8,000 (Bartmanski & Woodward, 2015).

U.S. Recorded Music Revenues by Format

1973 a 2017, Format(s): LP/EP, Vinyl Single, Download Album e 3 mais
 Source: RIAA

Source: RIAA. Permission to cite or copy these statistics is hereby granted, as long as proper attribution is given to the Recording Industry Association of America

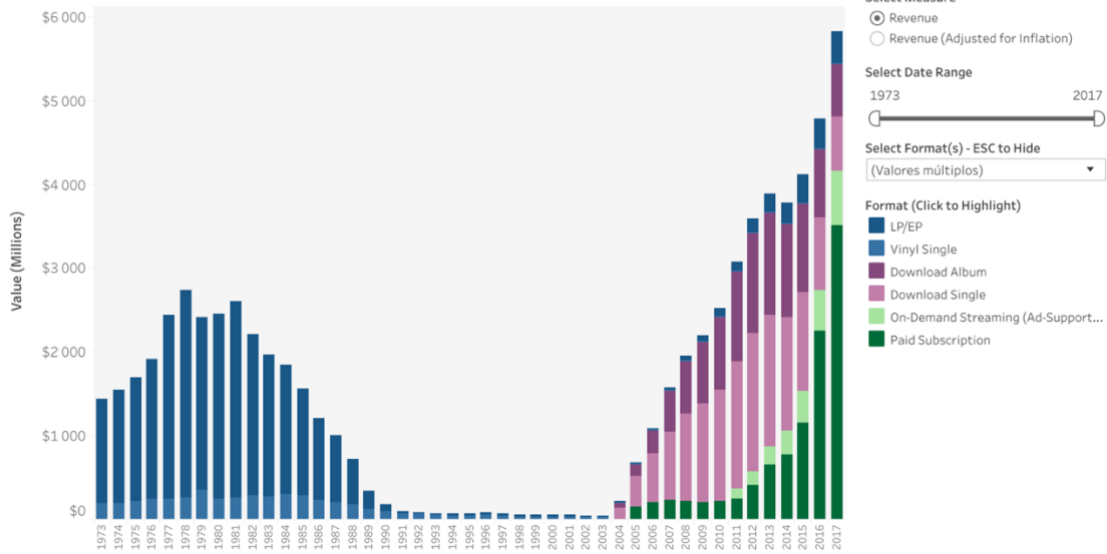


Figure 4 - U.S. Music Revenues by Format (U.S. Sales Database, 2017).

U.S. Recorded Music Revenues by Format (Adjusted for Inflation, 2017 Dollars)

1973 a 2017, Format(s): LP/EP, Vinyl Single, Download Album e 3 mais
 Source: RIAA

Source: RIAA. Permission to cite or copy these statistics is hereby granted, as long as proper attribution is given to the Recording Industry Association of America

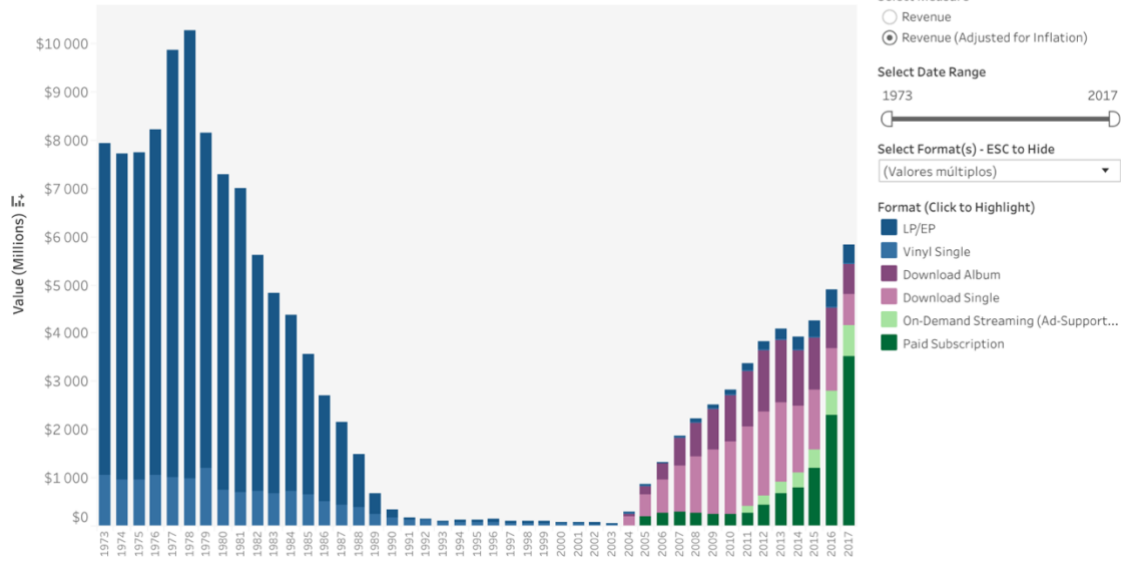


Figure 5 - U.S. Music Revenues by Format, Adjusted for Inflation (U.S. Sales Database, 2017).

2.3. The importance of record labels in the 1960s – 1980s

Although, for most of the population, listening to music may fall under the category of entertainment, for the music industry and all its players, representing a total of 38 billion dollars in annual global sales, it is very much a serious business. Despite the fact that producing music as a saleable product is a somewhat recent phenomenon, the creation of a music market was initiated by the sale of music sheets in the nineteenth century, only really taking off a century later when the demand for recorded music formats started to emerge, being supplied vinyl records, cassettes, CD's, among others (Gillett, 2011) transforming the music industry in a global business.

Through all the changes that the physical product for the delivery of music has undertaken, the industries division of labour and hierarchies have remained relatively stable. Artists are the ones that create music whereas record labels help them produce, promote and distribute the final product, which will finally be consumed by the audience. In the USA, the music market in 2007, was mainly dominated by only five major record labels, so-called “big five”, which were, namely, EMI, Sony, Universal-Vivendi, Time Warner and Bertelsmann BMG, controlling both production and distribution of music records. In October 2008, the music market share was split mainly by three of these 5, with Universal Music Group leading the pack, owning 35.12% of the market, followed by Sony Music Entertainment with 22.79% and Warner Music Group with 21.12%, leaving EMI Group with 8.35% and the remaining players with a mere 12.61% share (Day, 2011).

This market share division became even more favorable for the soon to be “big three”, when, in 2008, Sony bought the remaining 50% stake held by BMG, thus acquiring them (Kreps, 2008). In the following years, EMI also met its demise, ultimately selling most of their branches to Time Warner, Universal (Perpetua, 2011) and Sony (Wang, 2018). Some believe that it is due to the existence of a monopoly capable of controlling the entire supply chain, that artists have been prevented of independently producing and distributing their own material. This explains the fact that record labels are the ones who get to keep around 85 to 90% of the profit generated from music sales, using their influence to generate changes in technology to enhance their dominant position.

Luckily for the consumers, with the creation of the internet and its dissemination paving the way for online file sharing and the subsequent explosion in popularity for illegal music downloads, the traditional supply chain model that heavily relied on

physical distribution is starting to face some serious challenges. In 2004, the global piracy industry was estimated to be worth around 4.8 billion dollars with reported numbers of 4.5 million counterfeit CD's sold each year in the U.K. alone. Another challenge created by the internet was the creation of online services, such as Amazon.com, that specialize in worldwide product distribution, mitigating the need to pay for distribution services provided by record labels (Graham *et al.*, 2004).

2.4. Charts

According to Hakanen (1998), representing information in a chart can have multiple functions, from being applied as a marketing tool in the music business to radio stations using them to create and evolve playlists that could increase their appeal to their audience or retailers relying on them to order stock and organise their stores, as well as recommending customer purchases. Generically speaking, charts serve their audiences as an information tool. As said in Attali (1985) "Charts give value, channel, and select things that would otherwise have none, that would float undifferentiated".

In the music industry, beyond these marketing uses, a popular music chart conveys the complexity of relationships among several factors such as business, musicians, music and consumer. It directly shows the state of the music business even to the most alienated of consumers. Music charts not only define what is popular but, more importantly, help shape the definition of popularity itself. On the downside, they leniently try to organize art into categories, contributing to the loss of their unique and unquantifiable properties. As in any type of list, charts mitigate complexity with 'graphical simplicity, aggregating different forms of relationships between "pairs" into all-embracing reality' (Hakanen, 1998). Being able to have an album or a single reach one of the top charts is a very important goal in an artist's career since it brings important economic implications and influences the public's awareness, perception and, most importantly, the profits yielded by the artists' work (Bhattacharjee *et al.*, 2007).

Nowadays, popular music charts work as a form of templating for the ranking system, being one of the oldest, as well as one of the most referenced and recognized systems in popular culture. Consumer Reports take advantage of rankings reminiscent of Billboard's charts whereas local newspapers feature movie rankings with the most recent releases and, even in everyday discourse, we define someone's success as "rising to the

top of the charts". All these references reinforce the idea of popular music charts as a strong ideal type of information representation.

The most pertinent chart, that is constantly mentioned across the available literature, is the particular case of the Billboard magazine's charts which currently stands as a model of a universal ranking system. In these charts, it is possible to observe numbered slots forming a list of artists and their music, publisher and recording company, accompanied by its history in the form of numbers describing their previously occupied positions (one week ago, one month ago, etc.). Billboard magazine has harnessed such a reputation by providing chart information based on sales music recordings from as early as 1913 to the present date, assembling their weekly Top 100 albums from retail-store sales reports collected, compiled and provided by Nielsen SoundScan (Bhattacharjee *et al.*, 2007).

The use of this ranking systems is globally widespread, ranging from local newspapers, television and radio top-hits to country-wide charts aggregated by sales, genre, or even lack of popularity, depending on what the audience wants to know. It is because of the absence of one global authority or reference that each entity may choose to create their own chart to convey what themselves believe to be important. This variety of sources and diversity of evaluating parameters, creates a need to focus on the entity that seems to be the most representative of the industry. On that note, this thesis will mainly take into consideration the Billboard magazine's charts, when analysing trends and tendencies, due to its immense popularity and abundance of information (Hakanen, 1998).

2.5. Electronic markets

With the evolution and development of the Internet, the vinyl marketplace, like many other markets that started taking some form online, was suddenly able to reach larger audiences and create spaces where followers or particular niches, brands or products can cluster and create connections, enabling them to share experiences, tastes and contributing to the evolution and preservation of vinyl records. Consumers of any kind were finally able to enjoy music of better quality and authenticity, since LP's became available online. This also made it possible for people to only purchase records they were genuinely interested in, because they could listen to entire albums digitally before committing to buying the vinyl version (Sarpong *et al.*, 2016).

Simultaneous to the emergence of this e-market, contemporary independent labels started releasing store-only series of their vinyl, motivating the more driven enthusiasts to visit local stores to expand their collections. It is also the case that local shops sell unique vinyl at lower prices which can also happen at antiques and vintage markets, something that is much harder to come by at websites such as Discogs (Bartmanski & Woodward, 2015).

Regarding electronic markets, there are three main players in the business of reselling vinyl records. Amazon with 900,000 entries registered in their U.S.A. website, eBay with 2.3 million items and Discogs leading with 5.7 million (Rosenblatt, 2018). Discogs is a great example of one of these electronic markets, since it is one of the most important web resources for anyone who wants to identify, locate, sell or buy any physically recorded media, due to its extensive database coverage of vinyl records and almost 140,000 contributors in 2016 (Diggin' Into Discogs Data, 2016). In 2017, this online platform alone, reported almost 8 million in vinyl sales, an 18,81% increase in comparison to the previous year (State of Discogs 2017, 2018).

2.6. Data analysis of e-markets

One of the most valuable assets of any business is the information collected about how their costumers interact with one's products or services. Among the large amounts of data that result from these interactions, resides powerful information that can help shape a business or organization in their eternal quest for competitive advantages over the market. To uncover this hidden knowledge and draw the big picture of whatever insights a business needs to succeed, it is necessary to continuously collect, store, process and analyze the vast datasets that result from daily operation.

Looking at the numbers on the amount of data produced each year, it is possible to observe a tremendous growth, implying an increasing need of unearthing the latent potential of an organization's data. For example, according to Ahmed (2004), in 1989, the estimated number of databases in the United States was around 5 million whereas, 15 years later, that number has doubled. However, the interesting phenomenon presents itself in the past couple of years, where the exponential explosion in data generation accounts for 90% of all existing data (Marr, 2018), implying the existence of a huge variety of datasets that go from a few megabytes to several exabytes, which require the application of increasingly complex statistical tools combined with machine learning algorithms that

enable the classification, discovery of patterns and trends or prediction of possible outcomes (Silva, A. T., Moro, S., Rita, P., & Cortez, P., 2018).

The usage of these automated processes has earned the designation of Data Mining and its results can be potentially applied to many different problems such as decision support, prediction, forecasting and estimation, providing crucial aid in important business decisions (Ahmed, 2004). The roots of Data Mining lie on statistics and data analysis which have been greatly improved through applying machine learning methods and techniques. This concept, despite its prior existence as an evolving process, was only coined with its current name in the 1990's and has been gaining on popularity since then. However, it is only in the past few years that it has seen its major boom, mainly due to the increasingly large amounts of data, known as Big Data, produced every second from a very diverse array of sources that ranges from sensors and devices to social media interactions and applications (Canito *et al.*, 2018).

Nowadays, mining data is mostly used for improving customer acquisition and retention, reducing fraud, identifying inefficiencies in internal processes to help reformulate operations and, finally, mapping the unexplored terrain of the Internet. Discovering patterns of customer purchases, patterns of fraudulent credit card usage or identifying loyal customers are just some examples of these uses. The primary types of tools usually revolve around machine learning techniques such as artificial neural networks, genetic algorithms, decision trees, rule induction and nearest neighbor method together with data visualization. With these tools, depending on the target problem, Data Mining mainly produces six types of information: classification, regression, association, summarization, segmentation and visualization (Table 1 and 2).

Based on the previous examples, it is fair to say that these processes have a huge variety of applications in a multitude of sectors such as consumer product sales, finance, manufacturing, health, banking and insurance, among many others. This list keeps growing with the cost of data acquisition, storage and processing getting cheaper by the year, democratizing the technology for the masses, making it one of the most useful tools for the business community in the next century (Ahmed, 2004).

One of the many areas of application that has greatly benefited from the insights that Data Mining processes can yield is electronic commerce or e-commerce, which basically translates to commercial trades made online through websites or mobile applications. The knowledge found in mining transactional and clickstream data contributes to improving many aspects of these businesses, including site design and

experience, strategies for personalized product recommendation, customer loyalty and overall profitability, as well as detecting and preventing fraudulent behaviors. For example, good user experiences, either physically or virtually, largely enhance the average purchase size, the number of customers that return to the shop and, ultimately, the value of the brand, whereas bad experiences can hurt a brands' name and image much more than the immediate revenue loss.

Besides supporting online transactions, an e-commerce website can fulfill many of the necessary roles for such businesses. One of these roles is creating a place for customers to get information about products and services. As an example, in the year 2000, IBM estimated savings of 2 billion dollars in costs just by offering online customer support and information. Another role is the detection of emerging buying patterns such as highly searched products that yield no results, suggesting that the product should be made available by the supplier, or products that are often bought together can be more easily suggested. Furthermore, due to the non-physical nature of these stores, the possibility of testing new products or ads or even presenting the appropriate messages within the right timings, becomes much easier and, more importantly, much faster to reverse should some experiment not yield the expected results (Kohavi, 2001). The best proof of this are today's large online e-commerce websites such as eBay, Amazon and Alibaba that have already surpassed even the biggest retail stores, not only in customers or sales, but also in technological innovation (Silva *et al.*, 2018).

Table 1 - Methods and Algorithms in Data Mining. Adapted: Camilo & Silva (2009)

Class	Method	Goal	Algorithms
Prediction	Classification	Supervised learning approach where the goal is to attribute one of several discrete classification classes to each of the given instances of data.	Decision Trees, Rule Induction, Neural Networks, Genetic Algorithms
	Regression	Supervised learning algorithms used to predict values using dependencies and relationships between numeric and non-categorical variables.	Decision Trees, Rule Induction, Neural Networks, Genetic Algorithms
Description	Association	Unsupervised learning technique designed to uncover rules to describe relations between sets of data.	Rule Induction, Neural Networks
	Summarization	Unsupervised learning method used to robustly describe the main tendencies of subsets of data.	Decision Trees, Genetic Algorithms, Neighborhood Roughs
	Segmentation	Another unsupervised learning class of algorithms that focus on grouping the most similar instances of data, to represent their underlying relation.	Decision Trees, Rule Induction, Neural Networks, Genetic Algorithms, Neighborhood Roughs
	Visualization	Visual component of the process, enabling scientists to portray the results through graphs, diagrams and charts.	Decision Trees

Table 2 - Data Mining Algorithm Applications

Algorithms	Example applications
Decision Trees	Automatic generation of decision trees considering possible relevant variables to characterize vinyl price drivers.
Rule Induction	Induction of rules based on vinyl sales data to detect patterns and relations between market characteristics and price.
Neural Networks	Creation of neural networks to model and predict fluctuation in the price of vinyl.
Genetic Algorithms	Application of genetic algorithms to create a predictive model for vinyl price changes.
Neighborhood Roughs	Characterization of the market by calculating clusters of albums, artists, genres or price classes.

2.7. Related literature

Regarding the objective of this thesis, which is to analyze vinyl record price formation, using web scraping and market research to gather data which will then be processed by data mining techniques, the related work research was mainly focused on data mining analysis of price on related subjects.

On this topic, several articles were found that applied data and text mining to diverse areas, like stock or electricity prices focusing on different goals and using a plethora of analysis techniques (Table 3). Nonetheless, the scarcity of results specifically applied to vinyl or music markets indicates that a new, more technological approach of analysis may lead to interesting conclusions.

Table 3 - Contributions of related literature.

References	Business	Data	Contributions
Goolsbee & Chevalier (2002)	Online sales	Sales ranks of approximately 20,000 books from Amazon and Barnes and Noble.	- Price sensitivity of the studied merchants proved to be significant in both but demand at Barnes and Noble had greater elasticity of price.
Etzioni <i>et al.</i> (2003)	Air travel	12000 airline fare observations from six different companies.	- Showed significant simulated savings resulting from the application of data mining to price data.
Kaur <i>et al.</i> (2011)	Online auction	Complete bidding records from 149, 7-day auctions that were transacted on eBay between March and June 2003.	- Improvements in auction's end price prediction for clusters who supported the clustering-based model proposed by the bid selector.
Lu <i>et al.</i> (2005)	Electricity market	9 months of data gathered from the Queensland electricity market.	- Proposed a forecast model able to predict forecasted price spike, level of spike and associated forecast confidence level.
Patel <i>et al.</i> (2015)	Stock market	10 years of historical data from 2003 to 2012 from two stocks and two stock price indices.	- Trend deterministic data preparation as a way of improving prediction accuracy of stock prices.
Moro <i>et al.</i> (2018)	Hotel booking	3137 booking simulations (Booking.com) complemented with data from other online sources (Google, Tripadvisor, Facebook).	- All online sources play a significant role in prices (social media has the most relevance).
Huang <i>et al.</i> (2012)	Electricity market	Year long observations from markets of Ontario, Alberta and New York.	- From the three feature selection methods and the four classifiers tested, correlation-based feature selection scored highest and decision trees scored lowest in 24-hour ahead electricity price prediction, respectively.
Kilian & Vega (2011)	Energy market	Daily data on crude oil and gasoline prices combined with component news of U.S. macroeconomics (1983-2008).	- Found no compelling evidence of a strong relationship between energy prices and macroeconomic news.

Chapter 3 – Materials and Methods

3.1. Methodology

There are some methodologies that have been developed in an attempt to standardize data mining processes. The two most popular are CRISP-DM and SEMMA. SEMMA was developed by the SAS Institute, initially as a suggested method for working with SAS Enterprise Miner and it consists in the following five stages: sample, explore, modify, model and assess (Azevedo & Santos, 2008). This methodology will not be described in detail here since the present research will follow the Cross-Industry Standard Process for Data Mining, most commonly known as, CRISP-DM, due to its huge number of use cases and its better fit for an academic research that requires some business understanding before beginning the data sampling. The chosen standard is a comprehensive data mining methodology and process model that provides a blueprint for anyone to conduct a project in the area. This process model organizes the necessary tasks to properly handle data into six main phases (Figure 6), being those the following: business understanding, data understanding, data preparation, modeling, evaluation and deployment (Shearer, 2000).

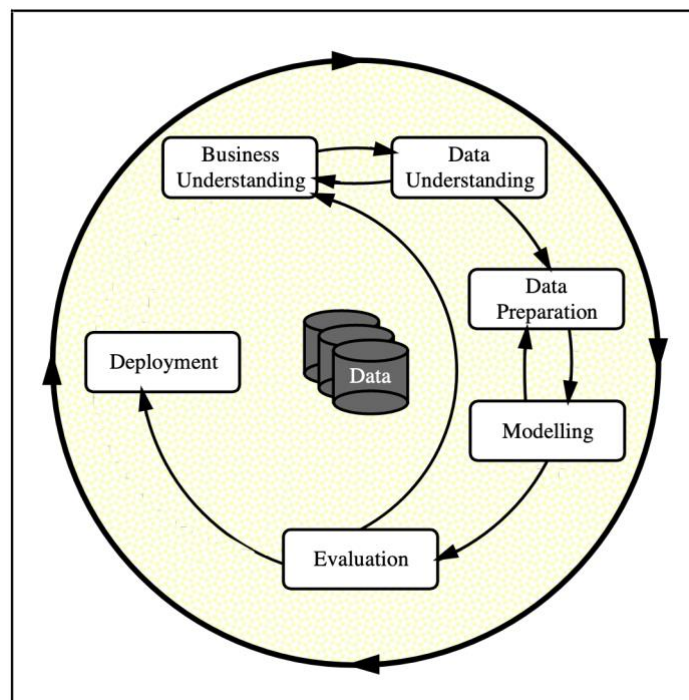


Figure 6 - Phases of the CRISP-DM (Wirth & Hipp, 2000).

The first phase of this methodology is perhaps the most important in the entire process since it consists in understanding the context of the problem for which a solution is trying to be found. In order to understand which data should be gathered or used, every data practitioner needs to first comprehend how the different variables might be related. To achieve this, there are some key steps that are usually followed such as determining business objectives, assessing the situation, determining the data mining goals and, finally, producing the project's plan.

Once there is enough information to build a plan, phase two is initiated. In understanding the data, the first step is to collect it, resorting to polls, web scraping, public datasets or api's, etc. These initial samples will be grossly evaluated by the data practitioner in order to assess if the data complies with the minimum requirements for the experiment. For example, if height is an important parameter and the initial dataset does not provide relevant information on height fluctuation values, then it is important to search on other sources. Finally, the data can be explored through queries, visualization and reporting and its quality assessed for invalid values or relevant incongruencies.

The next phase consists in preparing for the following modeling phase. It involves selecting the most representative features of the data sample before cleaning it using such techniques as attempting to estimate missing values or removing features altogether, among others. As all the relevant features are ready to use, among the various gathered sources, the obvious step is the integration of data into a single final dataset that will fuel the construction of a data model.

After the preparation and analysis, the data is ready to be modeled. By selecting which modeling techniques will be used, extra transformations might be necessary to comply with some mandatory format of the chosen technique. Next, the dataset is divided into training and test sets with the purpose of training the model on one of the sets and testing it on a completely different set, also assessing how unbiased the produced model can be. This is an iterative and repetitive process that requires testing several different models and tuning different parameters to achieve the most fitting solution to the problem at hands.

As mentioned above, before this model can be used in the real world, it needs to be carefully evaluated to check if it fulfills all the initial business requirements. In this phase, everything is rechecked starting on the results, and retracing the steps from the

choice of model and model parameters all the way through the transformations applied, looking for something that might have been overlooked.

At last, if the level of confidence on the produced model is high enough, it means it is now ready to be deployed into the business which can be done by planning the deployment of the model and its subsequent monitoring and maintenance, as well as producing reviews and reports that prove the accuracy and reliability of the produced work.

These six phases are the keystone of CRISP-DM and have been applied across industries and companies for more than twenty years and it is its simplicity, comprehensiveness and overflow of documentation and reviews that make it a suitable choice for the research process of this thesis.

3.2. Web Scraping

The procedure hereby designated as web scraping, is described by the automatic crawling of a website with the intention of collecting data, either by using scripts written for a specific task or by resorting to tools developed to extract information from the web. In comparison to a more traditional approach to data gathering such as surveys, the amount of information available is much larger and the time it takes to retrieve it can be reduced dramatically, depending only on the available computational resources. These great advantages also generate a problem of an excess of unstructured data that needs to be processed in order to become useful (Moro *et al.*, 2019).

There are three main steps in the process of scraping websites for data. The first step is to make a request to a website's available endpoint through the HTTP protocol (a stateless text-based Internet protocol that usually coordinates transactions between a Web browser or some other client, and a Web server). These requests usually need to simulate a "User-Agent" coherent with what would be an expected user request, in order to retrieve the desired response, since some websites try to differentiate between "bots" and users to prevent automatic retrieval abuse. Another way of preventing this kind of abuse, relies on the responsibility of the creators of these scrapers to make requests parsimoniously and only to endpoints indicated in the "robots.txt" file, minimizing the possibility of affecting the server.

Once the response is retrieved from the request, it needs to be parsed for its content to be extracted. Automated data extraction is only possible due to mark-up languages,

such as HyperText Markup Language (HTML) or eXtensible Markup Language (XML), that provide a common underlying nested structure and also due to the work done on creating standards for the interpretation of web-based content that help in bringing coherence and consistency to user experience among Websites and Web browsers. Despite the existence of these formats and the tools that facilitate its parsing like regular expression matching or HTML/XML parsing libraries such as XPath, this process is still highly dependent on the organization chosen by the website's creators, forcing the developer to thoroughly investigate the code organization of the desired target (Landers *et al.*, 2016).

Finally, it is necessary to organize all the gathered information into a standardized format, so it can be properly stored and further analysed in the future. Even though this process takes place on the previously scraped data, it can be considered an essential part of the scraping procedure since, without post-processing the information into a unifying format, the results would be much less usable, therefore rendering the process useless (Glez-Peña *et al.*, 2013).

For this research's web scraping process, the chosen data sources were two music genres from the vinyl marketplace Discogs (Figure 7 and 8) and the Hot 100 chart from the ranking website Billboard (Figure 9). For Discogs, the goal was to gather relevant data about the marketplace for two different genres in order to establish a comparison between two different audiences. The chosen genres were Rock and Jazz since they had the largest amount of available data. With the Billboard website, the main objective was to assess which artists and genres were the most popular in the vinyl golden era and since its reemergence in 2008. This data source provides insights on artist popularity which can be a determining factor on vinyl price. Since only the Hot 100 chart had relevant information on the vinyl golden era (starting from 1958 to the 1980's) the choice was simple.

The available release year and label for a specific release might not correspond to either the release year of the original record or its original label (or labels in the case of old international releases), since there are records that have been re-edited and remastered multiple times. Given this incongruence, an additional Discogs webpage was consulted, where an ascending sort by year could be applied, in order to retrieve the lowest release year registered and its corresponding label for the gathered releases. This page was obtained through the release page (Figure 8) and can be found in Appendix 1.

Discogs Search artists, albums and more... Explore Marketplace Community

Marketplace All Items Items I Want Purchases Cart Buyer Settings Search Marketplace

You Selected: **Shop Jazz Vinyl Records**

Format: Vinyl Genre: Jazz

401 - 425 of 2,673,399 < Prev Next >

Sort By: Listed, Condition, Artist, Title, Label Seller Price

Frank Sinatra - Some Nice Things I've Missed (LP, Album)
 Label: Reprise Records
 Cat#: F 2195
 Media Condition: Very Good Plus (VG+)
 Sleeve Condition: Very Good Plus (VG+)
 Disc is close to NM, Still in Shrink!
 View Release Page

CambridgeMusic
 98.1% 55 ratings
 Ships From: United States
 \$3.90
 about €3.51 + shipping

Jeff Beck - There & Back (LP, Album, San)
 Label: Epic
 Cat#: FE 35684
 Media Condition: Near Mint (NM or M-)
 Sleeve Condition: Very Good Plus (VG+)
 Great album! Like new!
 View Release Page

CambridgeMusic
 98.1% 55 ratings
 Ships From: United States
 \$7.50
 about €6.75 + shipping

Currency
 EUR (€) 1,217,075
 USD (\$) 994,030
 GBP (£) 371,079
 CAD (CA\$) 25,211
 AUD (A\$) 20,682

Genre
 Pop 472,598
 Funk / Soul 398,360
 Rock 286,085
 Stage & Screen 181,780

Figure 7 – Location for the vinyl marketplace features extracted from Discogs.

Discogs Search artists, albums and more... Explore Marketplace Community Log In Register

Frank Sinatra - Some Nice Things I've Missed
 Label: Reprise Records - F 2195
 Format: Vinyl, LP, Album
 Country: US
 Released: 1974
 Genre: Jazz, Pop
 Style: Easy Listening, Vocal

Tracklist

ID	Track Name	Duration
A1	You Turned My World Around Written-By - Bert Kaempfert, Dave Ellingson, Herbert Rehbein, Kim Carnes	2:45
A2	Sweet Caroline Written-By - Neil Diamond	2:42
A3	The Summer Knows Music By - Michel Legrand Words By - Alan & Marilyn Bergman	2:41
A4	I'm Gonna Make It All The Way Written-By - Floyd Huddleston	2:50
A5	Tie A Yellow Ribbon Round The Ole Oak Tree Written-By - Irwin Levine, Larry Brown	3:05

Release (2568871)
 Edit Release
 All Versions of this Release
 Review Changes

Marketplace 47 For Sale from €0.88
 Buy Vinyl Sell Vinyl

Statistics
 Have: 475
 Want: 44
 Avg Rating: 4.15 / 5
 Ratings: 26
 Last Sold: 23 Jun 19
 Lowest: €1.80
 Median: €4.40
 Highest: €18.00

Figure 8 - Location for the vinyl marketplace features in release page from Discogs.

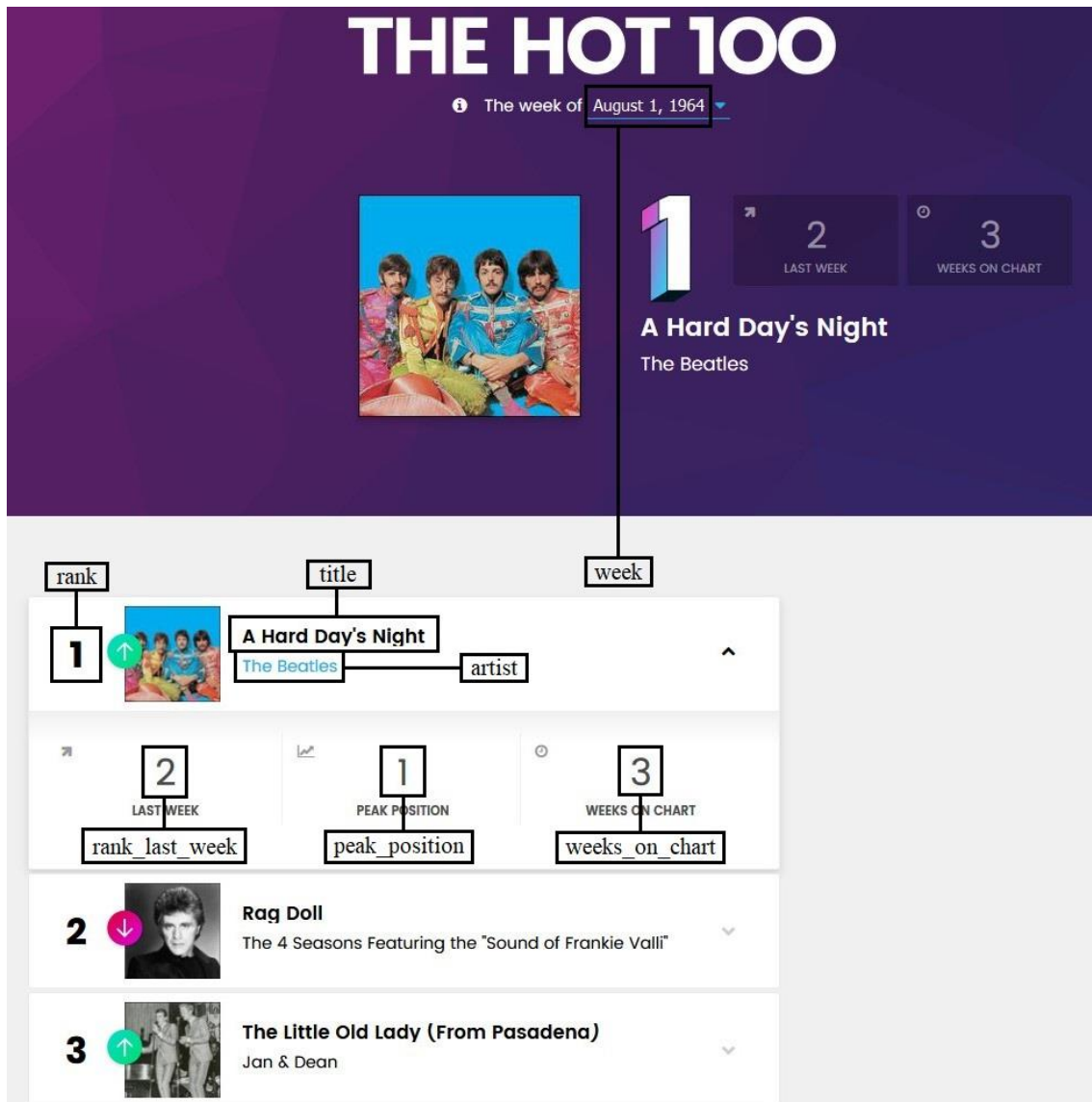


Figure 9 - Location for the chart raking features from Billboard.

The tools used in this process were the R programming language and an XPath module that enabled the extraction of data from the HTML nodes, obtained in each request to the chosen website. Despite the difficulties in being able to run long data extraction processes, due to network instability and server request throttling for too many requests, the scripts managed a decent amount of records for each dataset. For the Discogs extractions regarding Rock and Jazz, the number of observations gathered was 6,200 and 7,900 respectively, whereas, for the Hot 100 chart, gathered records for the time interval between 1958 and 1980 were 116,785 while, for the 2008 to 2019 interval, were 58,200. The extracted features from each source can be analyzed in detail in Table 4.

Table 4 - List of features.

Feature name	Source	Data type	Description	Status
title	Discogs	Character	Title and artist for the record.	Approved
label	Discogs	Character	Label that published the release.	Approved
release_page	Discogs	Character	Link to vinyl release page.	Removed
avg_vinyl_rating	Discogs	Numeric	Average rate of the release.	Approved
seller_rating	Discogs	Numeric	Rating of the seller.	Removed
nr_users_have	Discogs	Integer	Number of users that have the release.	Approved
nr_users_want	Discogs	Integer	Number of users that want the release.	Approved
n_seller_ratings	Discogs	Integer	Number of ratings given to the seller.	Removed
price	Discogs	Numeric	Price of the record.	Removed
media_condition	Discogs	Character	Condition of the vinyl record.	Approved/Converted
sleeve_condition	Discogs	Character	Condition of the vinyl's sleeve.	Approved/Converted
lowest_price	Discogs	Numeric	Lowest price for the release.	Removed
median_price	Discogs	Numeric	Median price for the release.	Approved
highest_price	Discogs	Numeric	Highest price for the release.	Removed
release_year	Discogs	Integer	Year of the release.	Approved
title	Billboard	Character	Title of the song in the chart.	Removed
artist	Billboard	Character	Name of the artist in the chart.	Approved
rank	Billboard	Integer	Rank of the song in the chart.	Removed
rank_last_week	Billboard	Integer	Last week's rank for the song in the chart.	Removed
peak_position	Billboard	Integer	Peak position for the song in the chart.	Removed/Converted
weeks_on_chart	Billboard	Integer	Number of weeks for the song in the chart.	Removed/Converted
week	Billboard	Date	Date (DD-MM-YYYY) the song was in the chart.	Removed/Converted
artist	Computed	Character	Extraction of artist from Discogs title feature.	Approved
min_peak_position	Computed	Integer	Best position in rank achieved by an artist with a song.	Approved
max_weeks_on_chart	Computed	Integer	Maximum number of weeks an artist stayed in rank with a song.	Approved
year_on_chart	Computed	Integer	Year extracted from week feature.	Approved

3.3. Data Transformations

As previously mentioned, in order to efficiently analyse and model data, it is of the utmost necessity to process the data until it is ready for the intended analysis. As such, all the gathered datasets went through a transformation process that will be documented in this section.

Before these transformations, some features were filtered out as seen in Table 4's status column. This filter was applied post data gathering and removed the following features: 'release_page', 'seller_rating', 'n_seller_ratings', Billboard's 'title', 'rank_last_week', 'rank', 'peak_position', 'weeks_on_chart' and 'week'. The variable 'release_page' was used to extract 'highest_price', 'median_price', 'lowest_price' and 'release_year' but had no relevance to the price. 'Seller_rating' along with 'n_seller_ratings' showed a low variation of rating among the sellers in Discogs. Billboard's 'title' was impossible to match with Discog's data since it referred to a specific song and not a record title, while 'rank_last_week' and 'rank' were less relevant since 'peak_position' was the most relevant to extract the best position achieved by an artist in the Hot 100 chart, thus establishing a one to one correlation by 'artist' with Discogs. Finally, 'peak_position', 'weeks_on_chart' and 'week' yielded 'min_peak_position', 'max_weeks_chart' and 'year_on_chart' respectively.

3.3.1. Discogs Transformations

In the datasets obtained from Discogs, the first transformation applied consisted in correcting the encodings of the extracted currency symbols. Because this platform is an international marketplace, several records are uploaded in different currencies, represented by special characters out of the traditional ASCII scope, making it necessary, for readability purposes, to convert these special characters to the same encoding as the rest of the data. These different currencies were then converted to euro to facilitate the comparison.

The next transformation aimed at removing rows where missing values existed in important columns. This is an important step to guarantee that the data is solid and coherent and that any model trained on this dataset will not be misled by invalid values. Since most of the missing values were under the sleeve condition column, once they were eliminated, it was possible to convert the nominal scale used to represent media and sleeve

conditions into an ordinal scale from 0 to 9, where 0 and 1 stand for the values “No Cover” and “Generic” respectively (and are only applicable to sleeve condition) and the remaining values from 2 to 9 range from “Poor (P)” to “Mint (M)” condition (Table 5).

Table 5 - Conversions of media and sleeve condition.

Nominal	Ordinal
Mint (M)	9
Near Mint (NM or M-)	8
Very Good Plus (VG+)	7
Very Good (VG)	6
Good Plus (G+)	5
Good (G)	4
Fair (F)	3
Poor (P)	2
Generic	1
No Cover	0

The final one before cleaning the data from irrelevant and low variance features and removing duplicate entries consisted in extracting the artists name from the record title. This was only feasible because Discogs uses a standardized format for record names where the title and artist are separated by a hyphen. As such, by scraping only the record name, it was possible to extract both the record title and artist name. Finally, the removed features were number of seller ratings, seller rating and release page. The number of ratings of a seller was considered irrelevant as well as the release page url because they obviously have no direct impact on price. This was not the case with seller rating, which could have an impact on price. However, due to the low variance of the gathered values (probably caused by an elevated repetition of the same seller in the data), it also had to be excluded.

As mentioned in the web scraping section, due to the existing mismatches between the extracted data and the original release years and labels, the data for the master release page was extracted, providing more accurate values for the incongruent features. This extraction resulted in a set with the release page used for correspondence with the previously obtained data and the newly retrieved original release year and label. This data was integrated in the final dataset by performing a left join on release page followed by a replacement of the existing release years and labels by the correct ones wherever was

possible. For the cases where the new release year or label held no values, the previous ones were kept.

After these transformations, the Rock and Jazz datasets extracted from Discogs had 2,589 and 2,685 records respectively.

3.3.2. Billboard Transformations

Regarding the datasets extracted from Billboard's Hot 100, the applied transformations were a bit more complex despite the simpler dataset when compared to the one obtained from Discogs. The initial transformation was applied to deal with missing values. Unlike the method applied in the previous datasets, that consisted in removing rows with missing values. In this case, the features containing the missing values were, in fact, very relevant for the problem at hand. For that reason, the missing values were replaced by dummy ones. For example, in the case of the peak position, since a lower number represents a better position, the missing values for this column were given the value of 9999 to represent an invalid instance. The opposite was done with number of weeks on chart where a larger number means more time on charts, which was the desirable criteria. As such, the missing values were replaced by 0.

Secondly, new variables for year and week number were extracted from the chart date. This provided a good temporal notion without feeding date formats to the model. Consequently, the column 'week' was disregarded.

Once the two datasets extracted were ready, they were merged into a single dataset, not only to ease the posterior integration with the Discogs datasets for each gathered genre, but also because both datasets resulted from the same chart, having the same context. This step resulted in a single dataset with 174,985 records.

It was only in this moment that it was possible to extract the desired information from these entries. The goal was to know which was the peak position each artist had been able to achieve in this rank and, at the same time, check the maximum number of weeks they had been in the rank. These seemed to be the most relevant variables in this dataset that could be able to influence price in some way. It is important to point out that both peak position and number of weeks on chart are connected to a specific song by an artist and the calculated result was applied to the artist alone. This means that a given artist might have reached its peak position with one song while beating the record of weeks on chart with another.

Last but not least, in order to be able to remove the repeated artist occurrences, while keeping the best position in chart for an artist and the longest time that artist remained in the chart with a specific song, the resulting dataset from Billboard was filtered by selecting only the rows where the number of weeks on chart matched the maximum number of weeks on chart, resulting in a reduction in the number of records to 6,471. This finally enabled the merge between these results and each of the generated datasets for Discogs, attributing minimum peak position, maximum weeks on chart and the year on chart to the artists in Discogs Rock and Jazz data that also appeared in the resulting Billboard's Hot 100 records. For the remaining artists, that were not among the Billboard's Hot 100 artists, these three new columns acquired dummy values of 9999 (for year on chart and minimum peak position) and 0 (for maximum weeks on chart).

In the end, both Rock and Jazz datasets were divided by 'release_year' to consider the golden era of vinyl between 1958 and 1980 and its reemergence post 2008. This resulted in two datasets for each genre with 2,642 entries for Jazz's past records and 100 for more recent releases while the datasets for Rock's past and present records comprised 1,832 and 831 entries respectively. As mentioned earlier, the goal of this research is to understand what has conditioned record prices in the past and which factors have been relevant in the last decade. To achieve this, a comparison was done between the two aforementioned eras and two main musical genres, Rock and Jazz. The golden era of vinyl for Rock was compared with its corresponding reemergence and the same was done with Jazz's time intervals. After this, the data for vinyl's prime time was compared between Rock and Jazz and a similar comparison was done between these two genres for the recent growth in popularity.

3.4. Data Modeling

The process of data modeling usually consists in assembling a pipeline-like structure of tasks that select or transform data to be used as input to several algorithms that will try to model it and create a predictive or descriptive model. For this research, a simple set of tasks to partition the data and select variables was applied before generating the data models. These techniques were applied to each of the considered genres and "eras" in order to obtain a better knowledge of, not only what conditions price, but also what changes between old releases and new releases for these two music genres. The computation work described in this section was performed with SAS Enterprise Miner.

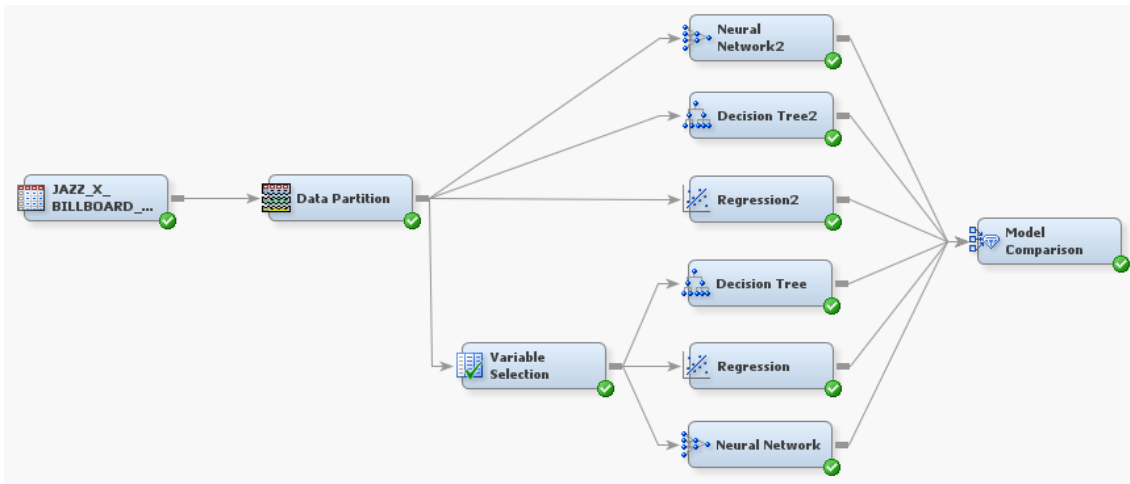


Figure 10 - Example diagram of the tasks applied.

3.4.1. Partitioning

When modeling data, usually a three-step process occurs where the data is divided in sets that will be given in a specific order to the algorithm analyzing it. These steps are commonly nominated as train, validation and test. The train subset is the first part of the data to be processed by the algorithm and is, as the name suggests, where the model is trained to best describe or predict the target. The second subset is the validation one. This is where the algorithm validates the adequacy of the generated model and also the phase for fine tuning of the generated models. Finally, the test step is applied to obtain an unbiased estimate of the model's generalization error, showing how the model could behave in a real scenario. In this thesis, the technique for data partitioning was a percentage split where the train set consisted in 70% of the data and the remainder 30% where divided 15-15 for validation and test sets.

3.4.2. Models

For this problem, the three different algorithms selected were Neural Network, Decision Tree and Linear Regression. Each of these algorithms can, as mentioned above, be used in many different contexts with many different purposes since they are built to use relationships between the input features to create a model that describes or predicts the chosen target.

Neural Network is a type of algorithm, primarily inspired by biological neural networks, where individual nodes (called artificial neurons) can receive, store, process

and emit a signal to other nodes. This node organization allows the network to learn from received input, with little guidance from the researcher, as the information is locally stored inside the nodes and the weights of the connections among them are readjusted to propagate new information (Hecht-Nielsen, 1992). In SAS Enterprise Miner, the default Neural Network node's architecture is the multilayer perceptron which consists of a network that can contain any number of inputs, any number of units inside its hidden layer (node layer, or layers, between input and output layers) and any number of outputs. This architecture uses a normal distribution error function (also denominated as least-squares or mean-squared-error) as default for interval type targets, which matches the data's target type (Neural Network Node: Reference, 2017).

Decision Tree consists in a tree-like organization of nodes that can represent either tests or attributes (Quinlan, 2014). A test node computes a value based on the attribute values of an instance usually resulting in two or more sub-trees that can contain more test nodes or leaves which, if encountered, halts the execution, yielding the predicted value. SAS implementation of this algorithm uses input type to find multi-way splits for nominal, ordinal and interval target inputs. For this research, the evaluation of a splitting rule was based on a statistical significance test called F-test and the default value of six was used to limit the maximum depth of the generated tree (Decision Tree Node, 2017).

Finally, Linear Regression, aims at representing a target variable by some linear combination of its associated features. When only one feature is present to explain the target, the method is called simple linear regression whereas, if multiple features are available (such is the case of this research), it is denominated as multiple linear regression. As for the Enterprise Miner details, the Regression node allows for target and input variables of the binary, nominal, ordinal or interval types. In this thesis, a linear regression was used, which implicates the use of an identity link function and a normal distribution error function. Since no selection method was applied, all the available features were used in this model (Regression Node, 2017).

3.4.3. Variable Selection

In this step, the goal was to discover the importance of the variables in relation to the target. The selected variables were then provided as input to the algorithms as to see if it was possible to generate better models than the ones generated from the entire set of features. There are several methods that can be applied when selecting the best inputs,

such as R-Square or Chi-Square selection which use different processes for assessing variable importance. The available method for interval-type targets is R-Square which consists in forward stepwise least squares regression that maximizes the model R-square value (Variable Selection Node, 2017).

3.4.4. Model Comparison

The final step is comparing the models generated to assess their performance using benchmarking techniques. The Model Comparison node of SAS (Model Comparison Node, 2017) implements many of these processes, making it the suitable node to use next. Due to the nature of the target, the chosen metric was Average Squared Error, in order to provide a good comparison between the models. This implementation takes the previously designated partitions for test and validation, and assesses how the different generated models perform by choosing the lowest average squared error, while also testing each model's generalization capacity.

Chapter 4 – Results and Discussion

4.1. Data Analysis Results

To better understand the datasets generated and, more specifically, what really influences the price of a record, it was necessary to first choose which feature better represented the price of a record (target variable) and then try to discover which of the remaining features had the most direct relation with the target. In order to accomplish this, the variable chosen for analysis was ‘median_price’, which results from a calculation made by Discogs for the average price of a release.

The next step was the comparison between this ‘median_price’ and the remaining features gathered from Discogs and Billboard’s Hot 100 chart. To achieve this, SAS Enterprise Guide was used to generate histograms and filter data, in order to help visualize which features are best related to price for each genre and each considered time interval. The plots contributed to understanding what some of the important factors for price on the Discogs marketplace were, such as ‘avg_vinyl_rating’ and ‘nr_users_want’, as can be expected from any market, since the quality and popularity of the product are always the main contributors for price.

However, the goal of this research is more focused in trying to discover if the popularity of the artist and the influence of its record label contributed to a higher pricing of their records. As such, each of the four datasets was filtered to create a smaller sample where only the entries that presented a ‘median_price’ above 30 were considered. This value was chosen as the lower bound for the “expensive” records filter, since according to Palm (2019), the average price of a record is around 25 dollars, and the goal of this step was to single out the records priced well above average. To enable a more detailed analysis into the most valuable records in the available data sample, Table 6-9 were created. This showed that, although there were a few extremely popular artists, where some even registered a ‘min_peak_position’ for the time intervals considered, these did not “dominate” the samples, as there are also less popular ones in these “above 30” listings.

Regarding the ‘label’ diversity in the tables below, such different labels would easily lead to the assumption that “the big three” lobby was not obvious in the gathered data. Nevertheless, by researching these recording companies and looking at their histories, their huge market share becomes very clear, as a lot of these businesses (Table

10) were eventually purchased by either one of the three. Although these major record labels are also represented for “present” records, it is a lot more common in the “past” datasets, since the music industry has changed drastically throughout the years and, many of the first labels did not grow big or fast enough to resist being acquired by the few ones that did. In addition, it is also important to report that the decrease in the presence of “big” recording companies is also clearly contrasted by a much larger number of independent labels, self-released records and a select few that achieved such success, managing to create their own companies and retrieve all the rights to their records as is the case for Metallica, Radiohead or Jack White, for example.

Table 6 - Comparison of artist and label for Rock Past.

artist	label	min_peak_position	median_price
Led Zeppelin	Atlantic	4	80
Steely Dan	ABC Records	4	44
Deep Purple	Parlophone	4	31,79
Jefferson Airplane	RCA Victor, RCA Victor	5	30,48
Jackson Browne	Asylum Records	8	96,1
The Who	Track Record, Polydor	9	168,08
Jethro Tull	Chrysalis, Island Records	11	58,6
Small Faces	Decca	16	48,71
Cymande	Janus Records	48	34,73
Janet Jones	Midas Recordings	9999	202,63
The Lou Reichner Band	E.S.R. Records	9999	98,42
Ugly Custard	Pussy	9999	70
Floating Bridge	Vault, Vault	9999	57,64
Jimmy Carter and Dallas County Green	BOC	9999	57,2
The Old Man & The Sea	Sonet	9999	42,5
Den Za Den	RTV Ljubljana	9999	40
City Preachers	Decca	9999	40
Gordon Lowe Featuring Laurel Ward	Yorkville	9999	35,79
Pascal Comelade	Parasite	9999	35,1
Scorpions	RCA Victor, RCA Victor	9999	33,55
The Pentangle*	Transatlantic Records	9999	32,39
The Slits	Island Records	9999	30,63

Table 7 - Comparison of artist and label for Rock Present.

artist	label	min_peak_position	median_price
Metallica	Blackened Recordings	31	30,8
Radiohead	Ticker Tape Ltd.	37	61,6
Sleaford Mods	Salon Alter Hammer, Anker (2)	9999	192,63
Witch (3)	Now-Again Records	9999	140,8
The Boxer Rebellion	Absentee Recordings	9999	139,9
Hot Mulligan	Save Your Generation Records	9999	115,79
The Flaming Lips	Warner Bros. Records	9999	86,84
The Temper Trap	Infectious Records	9999	80,99
RSI-MSK	Drumetrics	9999	80,13
HIM (2)	Sire (2), Reprise Records (2)	9999	61,6
Daughters	Ipecac Recordings	9999	52,79
Jack White And The Bricks	Third Man Records	9999	52,79
The Legendary Pink Dots	Beta-lactam Ring Records	9999	50,6
The Mystery Lights	Wick Records	9999	44
Pallbearer	Profound Lore Records, Profound Lore Records	9999	43,34
Monster Magnet	Spinning Goblin Productions	9999	42
Sniffing Glue	Plastic Bomb Records	9999	40
Handful Of Snowdrops	Domestica	9999	39,9
The Schoettes	Not On Label	9999	39
Chicano Batman	Drop Shadow Records	9999	38,73
Brother O'Brother	Self Released (Brother O'Brother)	9999	37,5
Miles Kane	Virgin EMI Records	9999	37,05
King Gizzard And The Lizard Wizard	Not On Label (King Gizzard And The Lizard Wizard Self-released)	9999	37
Sun City Girls	Abduction	9999	36,13
Boss Hog	Bronze Rat Records	9999	35,2
Snail Mail (2)	Matador	9999	33,75
Fucked Up	Tankcrimes	9999	33,15
Kishi Bashi	Joyful Noise Recordings, Joyful Noise Recordings	9999	33
Braveyoung	The End Records	9999	33
Opeth	Nuclear Blast Entertainment, Moderbolaget Records	9999	33
Sroeng Santi / Narong Rurachbuadang	Paradise Bangkok	9999	30,68
Warpaint	Rough Trade	9999	30,05

Table 8 - Comparison of artist and label for Jazz Past.

artist	label	min_peak_position	median_price
William DeVaughn	Roxbury Records	4	40
George Benson	Paul Winley Records	4	30
Joe Henderson	Blue Note	8	43,99
The Rainbow-Orchestra	Colorit	9999	182,5
Sonny Rollins	Blue Note	9999	172,75
Libra (6)	Cinevox	9999	156,2
Hideo Shiraki Quintet + 3 Koto Girls	SABA	9999	130,97
Dewan Motihar Trio, Irene Schweizer Trio*, Manfred Schoof, Barney Wilen	SABA	9999	101,5
Alessandro Alessandroni	SR Records (6)	9999	95,04
Horace Tapscott With The Pan-Afrikan Peoples Arkestra	Nimbus West Records, Nimbus West Records	9999	88,5
The Dave Pike Set	MPS Records, MPS Records	9999	74,5
Flim & The BB's	Sound 80	9999	70,4
Billy May	20th Century Fox Records	9999	66
John Tchicai And Cadentia Nova Danica	MPS Records	9999	62,92
Miles Davis + 19	Columbia	9999	52,8
Manfredo Fest Trio	Fermata	9999	52
The Soft Machine*	Probe, Probe	9999	50
Wes, Buddy & Monk Montgomery* Featuring Harold Land & Freddie Hubbard	Pacific Jazz	9999	49,33
Rundfunk-Tanzorchester Berlin	AMIGA	9999	48,5
George Braith	Blue Note	9999	48,4
Art Ensemble Of Chicago (AACM)*	Nessa Records	9999	45,03
Alfred Hitchcock	Imperial	9999	44,5
Harvey Mason	Arista, Arista	9999	44
Miles Davis	CBS	9999	43,96
Albert Ayler Trio	ESP Disk	9999	43,71
Sideline (3)	JA Records (2)	9999	40
Gary Bartz NTU Troop	Milestone (4)	9999	39,6
Grant Green	Blue Note	9999	39,06
Ernest Ranglin	MPS Records, BASF	9999	39
Billie Holiday	Storyville (3)	9999	37,1

Sonny Rollins	RCA Victor, RCA Victor	9999	35,2
Lonnie Liston Smith & The Cosmic Echoes*	Flying Dutchman	9999	35
Art Farmer	King Records, CTI Records	9999	34,95
The Diddys Featuring Paige Douglas	Bam-Buu Records, Bam-Buu Records	9999	34,95
The L.A. Jazz Ensemble	PBR International	9999	34,73
Thad Jones	Blue Note	9999	34
Pharoah Sanders	Impulse!	9999	32
Miles Davis	Columbia	9999	31,68
Howard Wales & Jerry Garcia	Douglas, Douglas	9999	31,58
Don Cherry	BYG Records	9999	30,8
Billy Harper	Denon	9999	30,4

Table 9 - Comparison of artist and label for Jazz Present.

artist	label	min_peak_position	median_price
Kamasi Washington	Young Turks	9999	38,72
Kamasi Washington	Brainfeeder	9999	34,88

Table 10 - Labels ownership and dataset distribution examples.

Owner	Label	Rock past	Jazz past	Rock present	Jazz present
Sony Music Entertainment	RCA Victor	X			
	Roxbury Records		X		
	Columbia		X		
	AMIGA		X		
	Arista		X		
	Flying Dutchman		X		
Warner Music Group	Atlantic	X			
	Asylum Records	X			
	Parlophone	X			
	Chrysalis	X			
	Sire			X	
	Reprise Records			X	
Universal Music Group	Polydor	X			
	Island Records	X			
	Decca	X			
	Blue Note		X		
	Imperial		X		
	Impulse!		X		
	Virgin EMI Records			X	
Independent	Track Record	X			
	Transatlantic Records	X			
	Blackened Recordings			X	
	Ticker Tape Ltd.			X	
	Ipecac Recordings			X	
	Third Man Records			X	
	Beta-lactam Ring Records			X	
	Profound Lore Records			X	
	Bronze Rat Records			X	
	Tankcrimes			X	
	Joyful Noise Recordings			X	
	Nuclear Blast Entertainment			X	
	Young Turks				X
	Brainfeeder				X

4.2. Data Modeling Results

One of the main disadvantages of using black box models (computer generated mathematical models that are very hard to interpret), is the inherent difficulty in understanding what or why the model was generated in a certain way and how the target's relationship with its characterizing features has shaped that same model. The necessity to obtain this information in order to better understand and evaluate these results, created a new research branch dedicated to mitigating this problem. Of this effort, techniques such as sensitivity analysis or rule induction from networks, were developed, enabling a better importance assessment of the input variables (Silva *et al.*, 2018).

In accordance with the data presented in Figure 11, obtained from the variable selection node, it is possible to observe some patterns in what variables play the largest and the smallest role in the evaluated datasets. It is important to mention that the variables 'nr_users_want' and 'avg_vinyl_rating' are not present in this plot since they were overshadowing the remaining contributing factors with much higher values and occupying the place of the two most important variables, though changing in order, for almost all datasets.

For the case of past Rock records, the variables 'sleeve_condition' and 'media_condition' present an interesting difference among the remaining features. In contrast, for its present counterpart, these sleeve and media conditions were the least significant, with 'min_peak_position', 'max_weeks_on_chart' and 'year_on_chart' appearing as more relevant, which shows some interesting differences between "eras". As for the Jazz datasets, for the one representing past releases, the most interesting pattern is in how similarly to the present Rock records, 'min_peak_position', 'year_on_chart' and 'max_weeks_on_chart' also represent important factors. Oppositely, in the set of data representing Jazz present, it is possible to, just like in Rock past, see 'media_condition' and 'sleeve_condition' maintaining some relevance.

These similarities across genres and eras motivate a comparison between the Rock and Jazz datasets. As mentioned, it is possible to observe in Figure 11, by focusing on Jazz past and present or Jazz and Rock past, how the variables 'min_peak_position', 'year_on_chart' and 'max_weeks_on_chart' tend to switch places in importance with 'sleeve_condition' and 'media_condition' across genres and eras. This interesting pattern will be discussed in further detail in the next section.

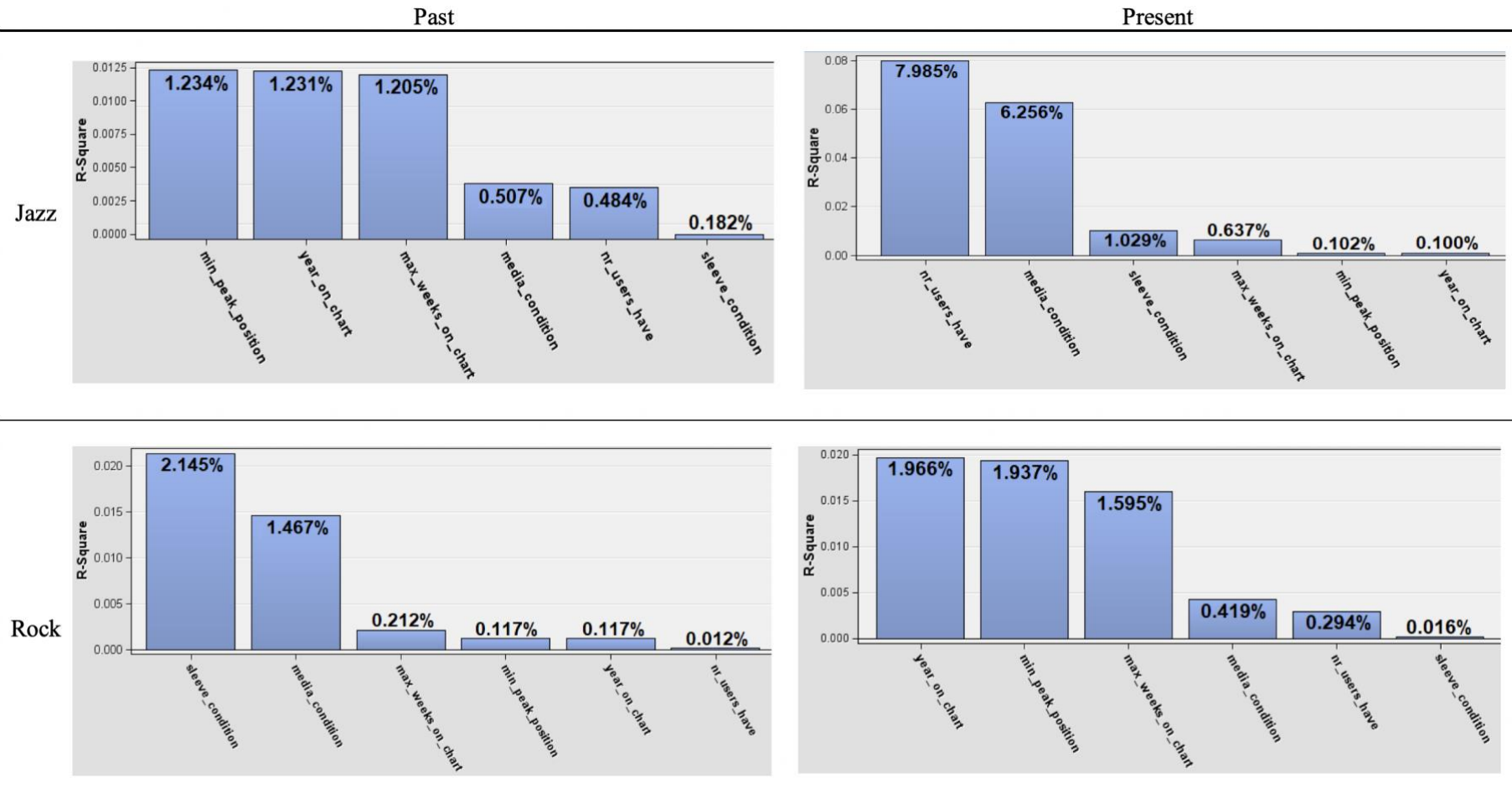


Figure 11 - Variable importance.

4.3. Discussion

Given the achieved results presented above, it is possible to observe some patterns and theorize on which factors may or may not be plausible explanations for what was reported while providing useful insights in characterizing what influences a vinyl record market value. It is important to mention that, since the data was gathered from Discogs and Billboard, any conclusions withdrawn from the data will be primarily applicable in the context provided by these platforms and, secondarily, to the overall music market.

With these limitations in mind, the most interesting results are the ones that are deeply connected to the goal of this study, which focuses, primarily, on an artist's popularity and his presence in the charts, as a possibly important price driver and, secondly, in the record label behind the success. Regarding these hypotheses, it is possible to conclude that both these factors can definitely be relevant to price, in some cases (as can be seen in Table 6) with the examples of Deep Purple or Led Zeppelin, which show a good 'min_peak_position' in Billboard's Hot 100 as well as labels that eventually became part of one of the "big three". Nevertheless, they do not always dictate which artists will have their records at highest prices. This can be explained by many possible scenarios, one of which could be a record release with a reduced number of produced records that becomes a collectible, making it a lot more valuable without necessarily having an immensely famous artist or a very powerful label behind it.

Focusing more deeply on the results presented on the record labels, it is interesting to observe the ongoing change that becomes obvious once one realizes which labels belong to which other labels. This change shows how the big record labels have been losing market share to an increasing number of independent labels, implying a paradigm shift on the industry going from a more label-centered to a more artist-centered market, where new alternatives to music production and distribution are made available every day and where more and more musicians become successful enough to build their own recording companies, giving artists in general more rights in the managing part business.

Besides these more goal-oriented results, the data mining process also uncovered some patterns that are worth discussing. The first noteworthy detail in the generated results is that, for each dataset, the models that achieved the best performance were relatively different, hinting on the subtle differences in the market for each genre and each era. In the variable selection process, it is also possible to see these differences (as shown in Figure 11), pointing to the changes that have been occurring in the music industry and,

consequently, at the generalized change in perception, from the artists to the audience, regarding the vinyl records. These variations, also show another interesting pattern.

As reported in the end of section 4.2 and shown in Figure 11, the variables ‘min_peak_position’, ‘year_on_chart’ and ‘max_weeks_on_chart’ appear to be more relevant in Jazz past and Rock present. Then, in their respective counterparts, Jazz present and Rock past, where ‘media_condition’ and ‘sleeve_condition’ tend to take the place of the first three. Such is worth considering, as the most obvious thing in common between Jazz past and Rock present is that, in the past, Jazz was one of the most popular genres, just like Rock is nowadays, which suggests that the features extracted from Billboard to measure popularity become more relevant in the “eras” where the record’s genre is more popular. It is also interesting to see what happens when the genres are not as appealing to the masses, with ‘sleeve_condition’ and ‘media_condition’ becoming more relevant, indicating that the records might become more perceived as collectibles.

Focusing on the discrepancies among the selections of variables and models for each dataset, the data for recent Jazz records has shown the most different results in several aspects. For example, the selected model for this dataset was the only selected model generated from the selected variables. This is interesting because, overall, the models that performed better were the ones that processed the complete set of input variables, suggesting that, despite the low R-Square values achieved by some of the features, they were still useful for the model to describe the target. Another example of these discrepancies can be seen in the R-Square values for the Jazz present dataset, where ‘nr_users_have’ has the second-best score. In comparison, this feature was consistently in the least relevant variables for all the other datasets.

Chapter 5 – Conclusions and Recommendations

5.1. Conclusions

This research, like many others, results from an endeavor to better understand some part of the present reality, gathering information about the subject and trying to extract knowledge from it. The initial objective of this thesis started as an effort to determine what factors influenced the price of vinyl records. The initial hypotheses consisted on two strong candidates as big contributors, the first being their popularity and the second being the label that originally released the record. For this, four datasets were built by scraping data from Discogs, extracting a sample of old and new records for both Rock and Jazz genres, and from Billboard's Hot 100 chart, where the ranking positions extracted were matched by artist with the entries from Discogs. The variables obtained from Billboard were used to assess the likely effect of popularity while the labels were extracted from Discogs along with several other features, including the target 'median_price'.

With these datasets, it was possible to generate plots to visualize the distribution of the data and existing visible correlations, showing that the record labels and the ranking positions, although present among the most expensive records (Table 6-9), did not seem to be the majority of the cases, indicating that many other factors might also influence the price of a record. Neural Network, Decision Tree and Linear Regression algorithms were also applied to the data, in order to generate mathematical models able to describe price in function of the other available inputs, a process that pointed out some potentially interesting changes in how the records seem to be perceived by the buyers.

Finally, the variable selection process, used to filter the inputs for three of the models generated for each dataset, suggests that variables like 'avg_vinyl_rating', the ranking on Discogs for the record's release, and 'nr_users_want', the number of users wanting the same release, seem to be the more relevant factors influencing price. Besides those obvious features, there is also another observable pattern regarding 'max_weeks_on_chart' and 'min_peak_position'. These variables seem to be more relevant in the periods of more popularity for Jazz, with the past "era", and Rock, with the "present" era. This means that when Jazz was one of the most popular genres, the artist's popularity would have a larger contribution to the price of his records. The same can be said about Rock and the relevance of its more recent popularity to price. Despite

the clear difference in relevance among the considered inputs, the models that best performed were still the ones that used an unfiltered set features, meaning that, regardless of which factors influence price, the remaining ones were still useful in finding a better solution.

In summary, this thesis suggests a possible set of variables that can be relevant to a vinyl record's price and a data mining approach to analyzing the datasets, creating an overview of what the main factors to be considered are and showing how the genre's popularity impacts the importance of an artist's popularity. For example, if nowadays an artist becomes popular within the Jazz genre, which has a relatively smaller fan base when compared to Rock, then the influence of his fame on the price of his records will actually be less than the influence of the record's condition. However, for more popular genres, with larger communities of followers and wider media coverage, the opposite becomes true. As such, the work presented in this research along with the knowledge yielded from the accomplished results, opens up several paths for further investigations as can be read in the last section of this chapter, as well as providing companies and interested people some tools to better understand how to evaluate and attribute a price to a record they might want to sell.

5.2. Research Limitations

In the present day, the availability of publicly accessible data on the Internet, as well as the tools provided by technology, enable faster information extracting and processing while also helping to understand the underlying knowledge. However, the over abundance of information also means the presence of many irrelevant sources creating the need to filter these sources and choose only the relevant ones becomes more evident. For this thesis, Discogs marketplace and Billboard Hot 100 chart were chosen as reliable data sources, building a dataset representing a sample of the existing online vinyl market.

The available data and the way it is structured online represents the first limitation to many studies resorting to web scraping since unstructured and missing data can be frequently found. Among the gathered variables, the ‘release_year’ and ‘label’ were particularly hard to reliably determine since there were many records listed on Discogs that did not refer to the original release, listing different release years and labels. To solve this problem, an attempt to increase the accuracy of these variables was done by finding the web address corresponding to the original record webpage in Discogs and retrieving the values listed there.

The second relevant limitation for this research is linked to the nature of the tools chosen to model and analyze the datasets. Since machine learning algorithms are essentially mathematical models that rely on the given inputs to be able to compute a possible solution, there are usually specific rules as to the way these inputs can be processed. In this case, all the non-discrete text inputs were disregarded when generating the models, automatically excluding artists and labels, conditioning the first look at the data and forcing a deeper research on the recording companies’ histories.

Lastly, it is worth mentioning that, in every data science study, the more reliable and more representative the data samples are, the better the models generated are able to perform, meaning that, despite the fact that some very interesting conclusions regarding the online vinyl market can be inferred from the datasets assembled, these same conclusions need further investigation and future work, in order to really be able to generalize to a larger amount of markets, genres or even formats.

5.3. Future Research

As introduced in the end of the previous section, the question that fueled this dissertation is a wide and complex question, opening many possibilities for future work. There are several fairly obvious next steps. The first would be to expand the number of datasets considered, by including more genres, the number of entries in each dataset, extracting more entries from Discogs, and adding new relevant variables, possibly performing text mining on Discogs' user comments, thus enabling a more complete overview on this online vinyl market.

Another interesting path would be to explore other formats from tape to online streaming services. As mentioned previously, online music streaming services represent nowadays the largest share of the music industry, eliminating the necessity of mass producing physical records and distributing them across the world and profoundly changing the process of finding and listening to music. This alone should be a good enough reason to try to better understand the changes the music business, and entertainment in general, are undertaking.

Thirdly, the way these profound changes affect the recording companies, how the “big three” monopoly was built, and the solutions developed to begin undermining it, also seem to be a very enticing line of reasoning to explore. The way physical records and digital music coexist in the present day, make room for countless possibilities of a future, where labels, much as music streaming services will have to adapt if they intend to survive, as the offer on the market increases and the costs for production and distribution decrease.

Last but not least, a lot more can be done regarding the models generated and any future ones. In this study, the parameters chosen for each algorithm were picked for being the most commonly used in the field to provide a solid base for future comparison. For that reason and the fact that sometimes, with careful tuning of some parameters by professionals that deeply understand these markets and these algorithms, it is possible to achieve models that produce incredibly useful results, this might also be an interesting research path to follow.

References

Ahmed, S. R. (2004, April). Applications of data mining in retail business. In *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on* (Vol. 2, pp. 455-459). IEEE.

Attali, J. (1985). *Noise: The political economy of music* (Vol. 16). Manchester University Press.

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Bartmanski, D., & Woodward, I. (2015). The vinyl: The analogue medium in the age of digital reproduction. *Journal of consumer culture*, 15(1), 3-27.

Bhattacharjee, S., Gopal, R. D., Lertwachara, K., Marsden, J. R., & Telang, R. (2007). The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. *Management Science*, 53(9), 1359-1374.

Camilo, C. O., & da Silva, J. C. (2009). Data mining: Concepts, tasks, methods and tools. *Institute of Computer Science Federal University of Goiás, Technical Report*, 4.

Canito, J., Ramos, P., Moro, S., & Rita, P. (2018). Unfolding the relations between companies and technologies under the Big Data umbrella. *Computers in Industry*, 99, 1-8.

Day, B. (2011). In defense of copyright: Record labels, creativity, and the future of music. *Seton Hall J. Sports & Ent. L.*, 21, 61.

Decision Tree Node. (2017, August 30). Retrieved from <https://go.documentation.sas.com/?docsetId=emref&docsetTarget=n0cx4ud03paymdn1kargegadueml.htm&docsetVersion=14.3&locale=en#p1hdxmdrosd4izn1m36fnn5l7sdi>

Diggin' Into Discogs Data. (2016). Retrieved from <https://blog.discogs.com/en/diggin-into-discogs-album-releases/>

Dixon, T., & Marston, A. (2002). UK retail real estate and the effects of online shopping. *Journal of Urban Technology*, 9(3), 19-47.

Etzioni, O., Tuchinda, R., Knoblock, C. A., & Yates, A. (2003, August). To buy or not to buy: mining airfare data to minimize ticket purchase price. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 119-128). ACM.

Gillett, C. (2011). *The sound of the city: The rise of rock & roll*. Souvenir Press.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. *Briefings in bioinformatics*, 15(5), 788-797.

Goolsbee, A., & Chevalier, J. (2002). *Measuring prices and price competition online: Amazon and Barnes and Noble* (No. w9085). National Bureau of Economic Research.

Graham, G., Burnes, B., Lewis, G. J., & Langer, J. (2004). The transformation of the music industry supply chain: A major label perspective. *International Journal of Operations & Production Management*, 24(11), 1087-1103.

Hakanen, E. A. (1998). Counting down to number one: The evolution of the meaning of popular music charts. *Popular Music*, 17(1), 95-111.

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65-93). Academic Press.

Huang, D., Zareipour, H., Rosehart, W. D., & Amjady, N. (2012). Data mining for electricity price classification and the application to demand-side management. *IEEE Transactions on Smart Grid*, 3(2), 808-817.

Kannan, K. S., Sekar, P. S., Sathik, M. M., & Arumugam, P. (2010, March). Financial stock market forecast using data mining techniques. In *Proceedings of the International Multiconference of Engineers and computer scientists* (Vol. 1, p. 4).

Kaur, P., Goyal, M., & Lu, J. (2011, April). Data mining driven agents for predicting online auction's end price. In *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on* (pp. 141-147). IEEE.

Kreps, D. (2008, October 2). Sony buys out Bertelsmann, Ending Sony BMG. Retrieved from <https://www.rollingstone.com/music/music-news/sony-buys-out-bertelsmann-ending-sony-bmg-101255/>

Kilian, L., & Vega, C. (2011). Do energy prices respond to US macroeconomic news? A test of the hypothesis of predetermined energy prices. *Review of Economics and Statistics*, 93(2), 660-671.

Kohavi, R. (2001, August). Mining e-commerce data: the good, the bad, and the ugly. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 8-13). ACM.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological methods*, 21(4), 475.

Lu, X., Dong, Z. Y., & Li, X. (2005). Electricity market price spike forecast with data mining techniques. *Electric power systems research*, 73(1), 19-29.

Marr, B. (2018, May 21). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. Retrieved from <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#1d54150460ba>

Model Comparison Node. (2017, August 30). Retrieved from <https://documentation.sas.com/?docsetId=emref&docsetTarget=p01jgc9rmzsg37n1lfncp67t0unm.htm&docsetVersion=14.3&locale=en#n0yplgk5vt8tu9n1scvy34w5n7tx>

Moro, S., Ramos, P., Esmerado, J., & Jalali, S. M. J. (2019). Can we trace back hotel online reviews' characteristics using gamification features?. *International Journal of Information Management*, 44, 88-95.

Moro, S., Rita, P., & Oliveira, C. (2018). Factors influencing hotels' online prices. *Journal of Hospitality Marketing & Management*, 27(4), 443-464.

Negus, K. (2015). Digital divisions and the changing cultures of the music industries (or, the ironies of the artefact and invisibility). *Journal of Business Anthropology*, 4(1), 151-157.

Neural Network Node: Reference. (2017, August 30). Retrieved from <https://documentation.sas.com/?docsetId=emref&docsetTarget=p0zbgj1tu3h1uhn1x6regixbdg7v.htm&docsetVersion=14.3&locale=en#p02hypwi3y88p1n17izqizaynoto>

Palm, M. (2019). Vinyl Records after the Internet. *The Dialectic of Digital Culture*, 149.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1), 259-268.

Perpetua, M. (2011, November 11). Universal Music Group Purchases EMI Music. Retrieved from <https://www.rollingstone.com/music/music-news/universal-music-group-purchases-emi-music-233091/>

Plasketes, G. (1992). Romancing the Record: The Vinyl De-Evolution and Subcultural Evolution. *The Journal of Popular Culture*, 26(1), 109-122.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Regression Node. (2017, August 30). Retrieved from <https://documentation.sas.com/?docsetId=emref&docsetTarget=n1jqzz8cssr9m2n1ktx2iyv87q56.htm&docsetVersion=14.3&locale=en#p1ro3qcktfvnyin130qbq984o4k9>

Richter F. (2014, November 19). Vinyl Comes Back From Near-Extinction. Retrieved from <https://www.statista.com/chart/2967/worldwide-vinyl-sales/>

Rosenblatt, B. (2018, September 18). Vinyl Is Bigger Than We Thought. Much Bigger. Retrieved from <https://www.forbes.com/sites/billrosenblatt/2018/09/18/vinyl-is-bigger-than-we-thought-much-bigger/#59e38e371c9c>

Sarpong, D., Dong, S., & Appiah, G. (2016). 'Vinyl never say die': The re-incarnation, adoption and diffusion of retro-technologies. *Technological Forecasting and Social Change*, 103, 109-118.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4), 13-22.

Silva, A. T., Moro, S., Rita, P., & Cortez, P. (2018). Unveiling the features of successful eBay smartphone sellers. *Journal of Retailing and Consumer Services*, 43, 311-324.

State of Discogs 2017. (2018). Retrieved from <https://blog.discogs.com/en/state-of-discogs-2017/>

Subramani, M., & Walden, E. (2001). The impact of e-commerce announcements on the market value of firms. *Information Systems Research*, 12(2), 135-154.

U.S. Sales Database. (2017). Retrieved from <https://www.riaa.com/u-s-sales-database/>

Variable Selection Node. (2017, August 30). Retrieved from <https://documentation.sas.com/?docsetId=emref&docsetTarget=n1m7rvh6yyb3mmn0zavezsher4ml.htm&docsetVersion=14.3&locale=en>

Wang, A. (2018, May 22). Sony Doubles Its Song Catalog – and Its Control of the Music Industry. Retrieved from <https://www.rollingstone.com/music/music-news/sony-doubles-its-song-catalog-and-its-control-of-the-music-industry-629865/>

Watson A. (2019, January 10). Share of vinyl album sales in the United States in 2018, by genre. Retrieved from <https://www.statista.com/statistics/694926/vinyl-album-sales-genre/>

Wirth, R., & Hipp, J. (2000, April). 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). Citeseer.

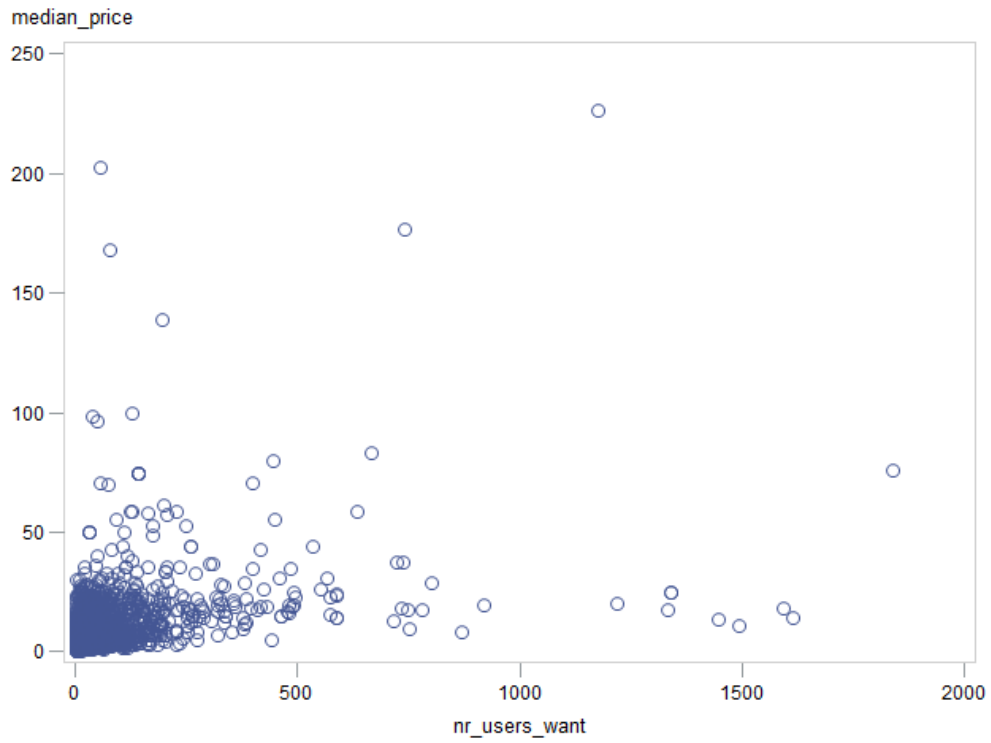
Appendix

Appendix 1 – Variables retrieved from Discogs Master page.

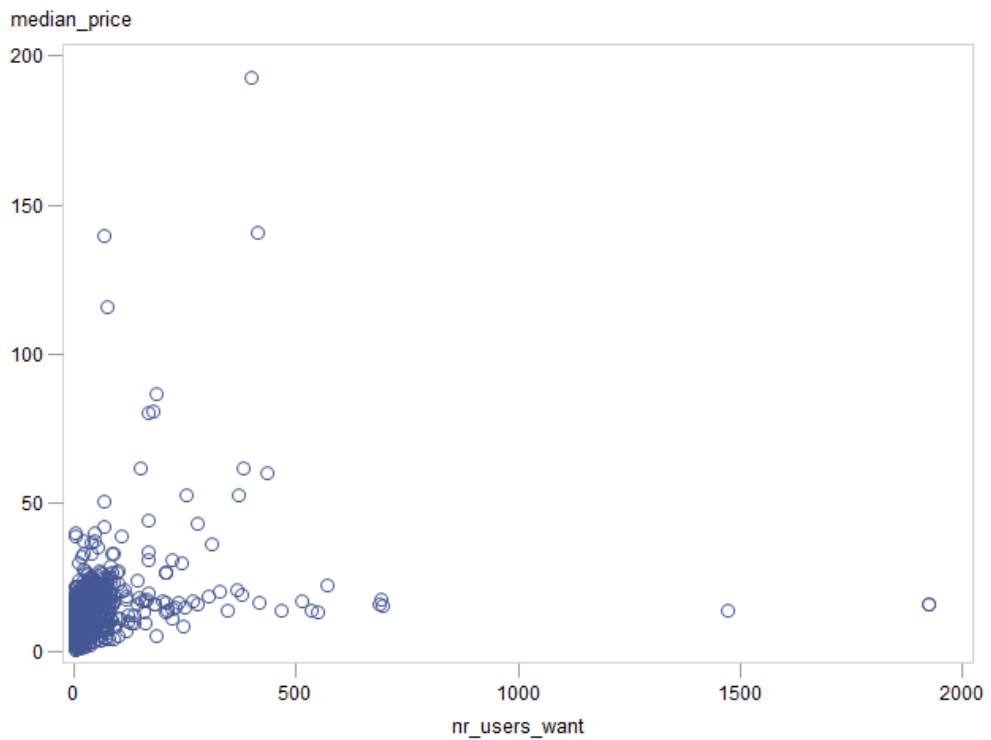
The screenshot shows the Discogs website interface. At the top, the browser address bar displays the URL: `https://www.discogs.com/master/211169?sort=year&sort_order=asc`. The Discogs logo and navigation menu are visible. The main content area features a profile picture of Frank Sinatra and the album title: **Frank Sinatra – Some Nice Things I've Missed**. Below the title, the genre is listed as **Jazz, Pop**, the style as **Easy Listening, Vocal**, and the year as **1974**. A tracklist follows, listing ten songs with their durations. Below the tracklist is a section titled **Versions (38)** with a search box labeled "Find Your Version". The table below lists four versions of the album, with annotations for `new_label` and `new_release_year`.

Title (Format)	Label	Cat#	Country	Year ^
Some Nice Things I've Missed (LP, Album)	Reprise Records	F 2195	US	1974
Some Nice Things I've Missed (LP, Album)	Reprise Records	F-2195	El Salvador	1974
Some Nice Things I've Missed (LP, Album, Quad)	Reprise Records, Reprise Records	S4-2195, F4-2195	US	1974
Some Nice Things I've Missed (LP, Album)	Reprise Records	54.020	France	1974

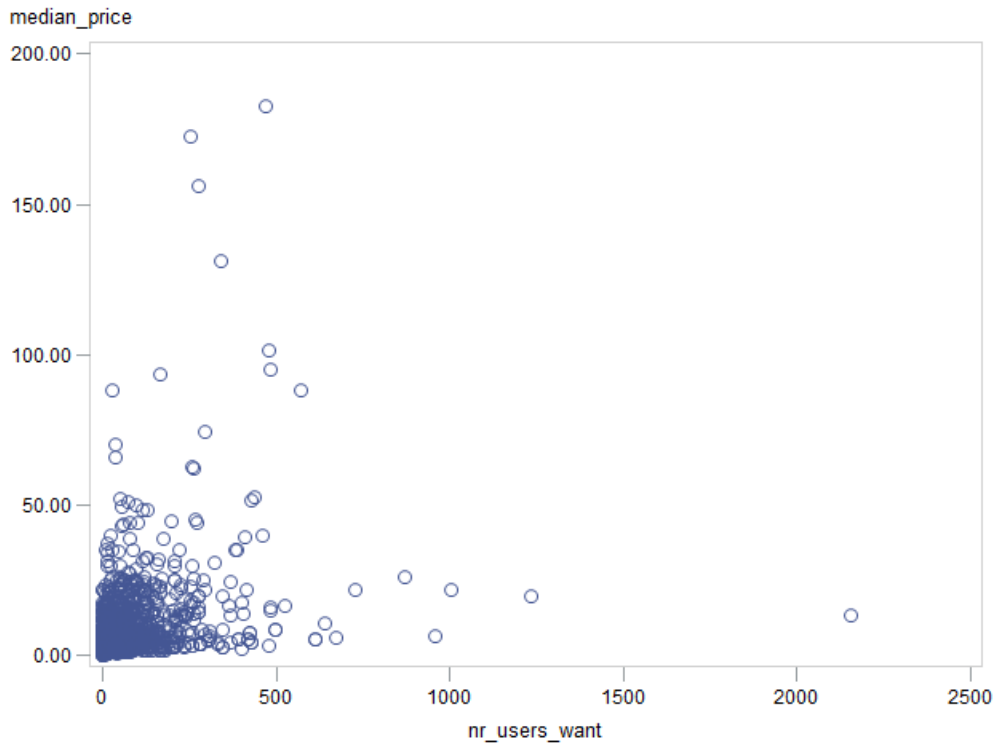
Appendix 2 – Scatter plot comparing 'nr_users_want' with 'median_price' for Rock past data.



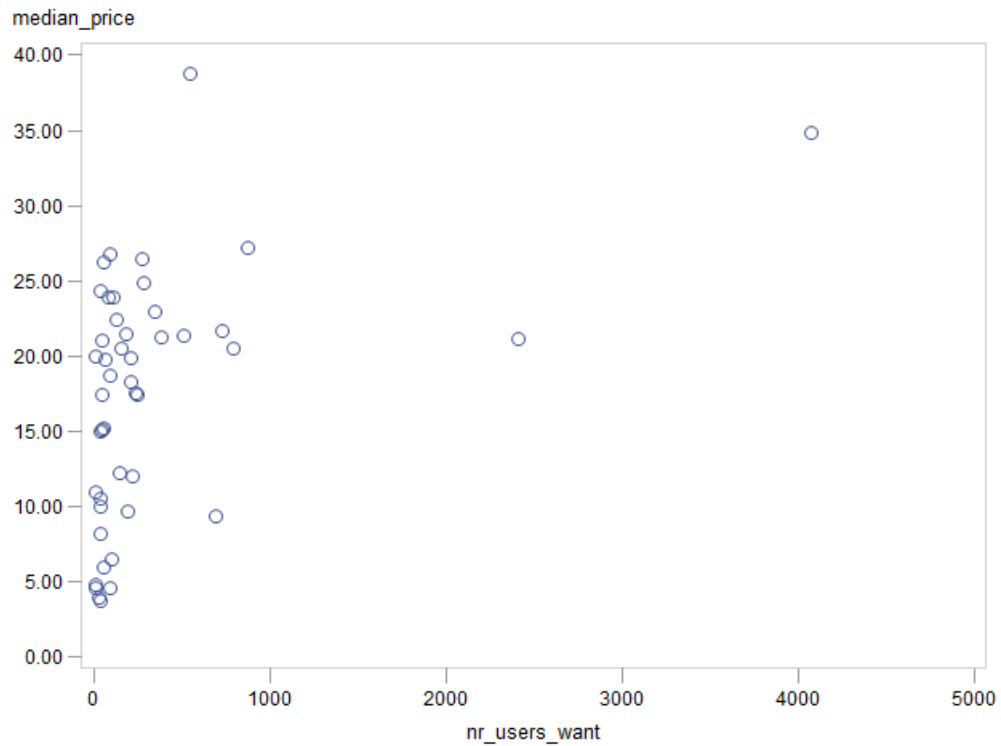
Appendix 3 – Scatter plot comparing 'nr_users_want' with 'median_price' for Rock present data.



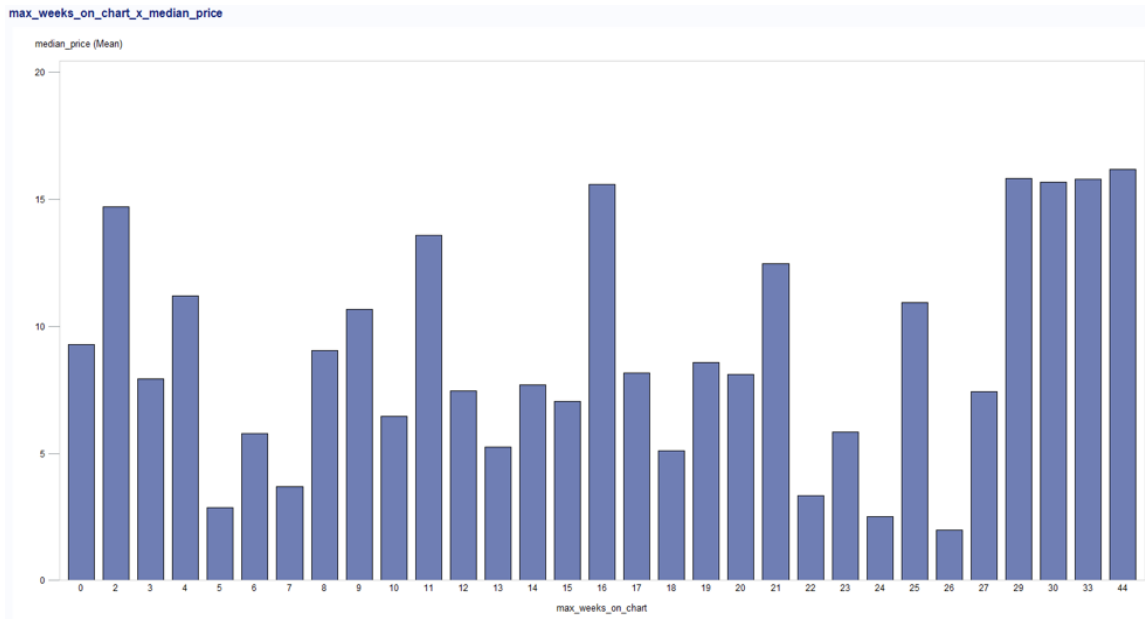
Appendix 4 – Scatter plot comparing ‘nr_users_want’ with ‘median_price’ for Jazz past data.



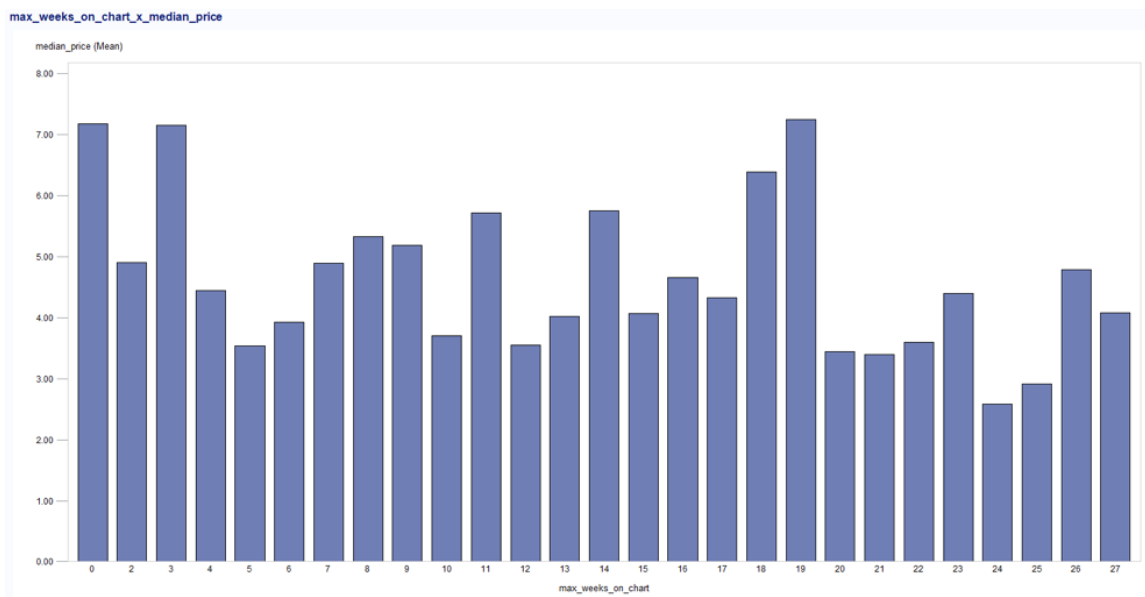
Appendix 5 – Scatter plot comparing ‘nr_users_want’ with ‘median_price’ for Jazz present data.



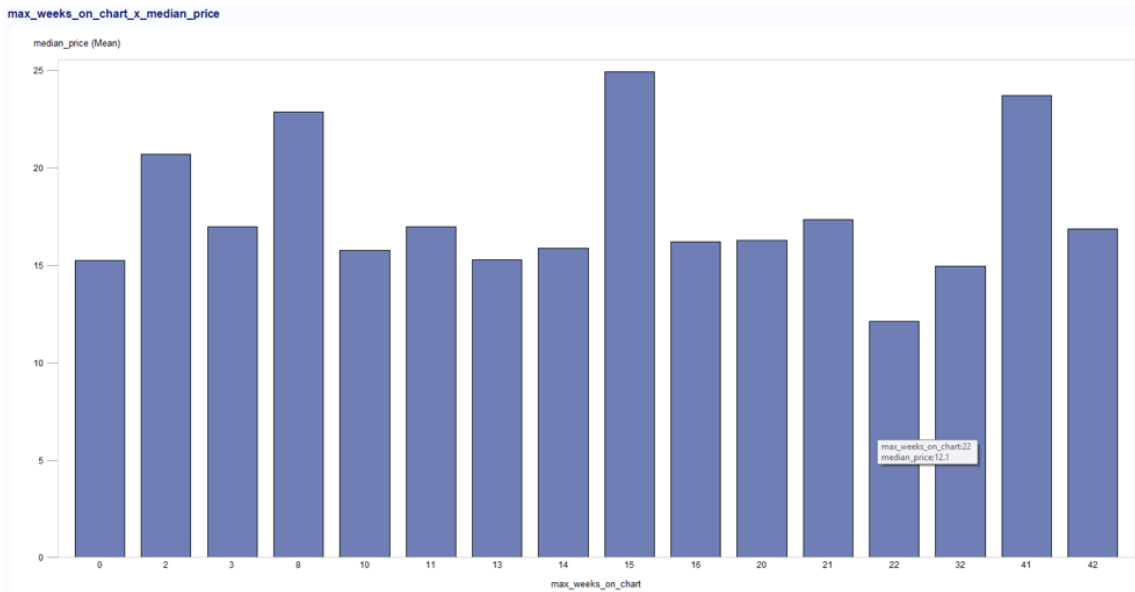
Appendix 6 – Comparison of 'max_weeks_on_chart' with 'median_price' for Rock past.



Appendix 7 – Comparison of 'max_weeks_on_chart' with 'median_price' for Jazz past.



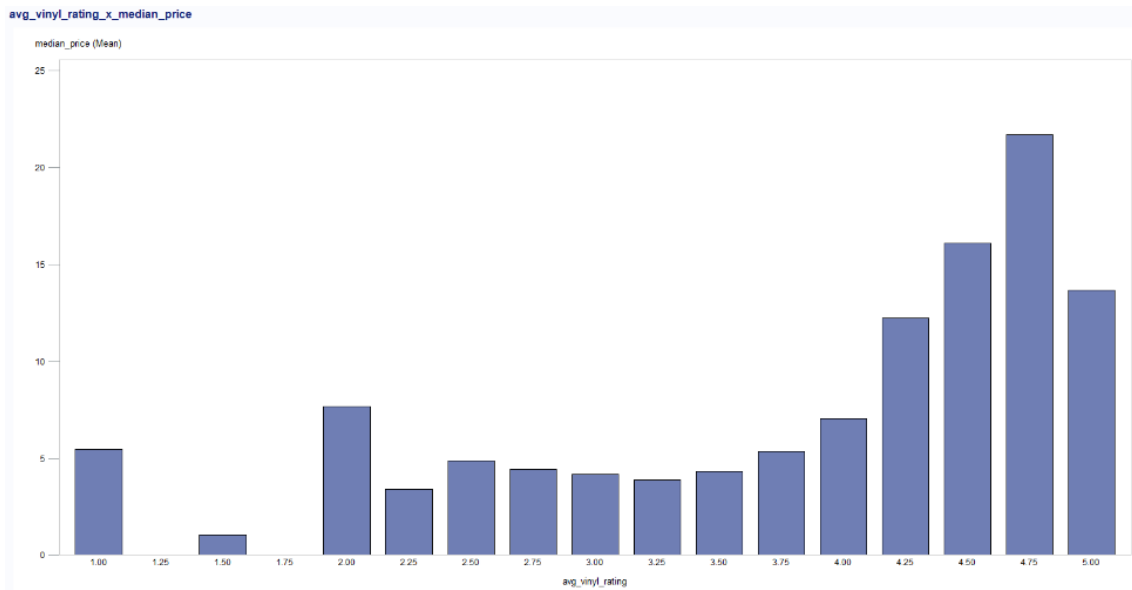
Appendix 8 – Comparison of ‘max_weeks_on_chart’ with ‘median_price’ for Rock present.



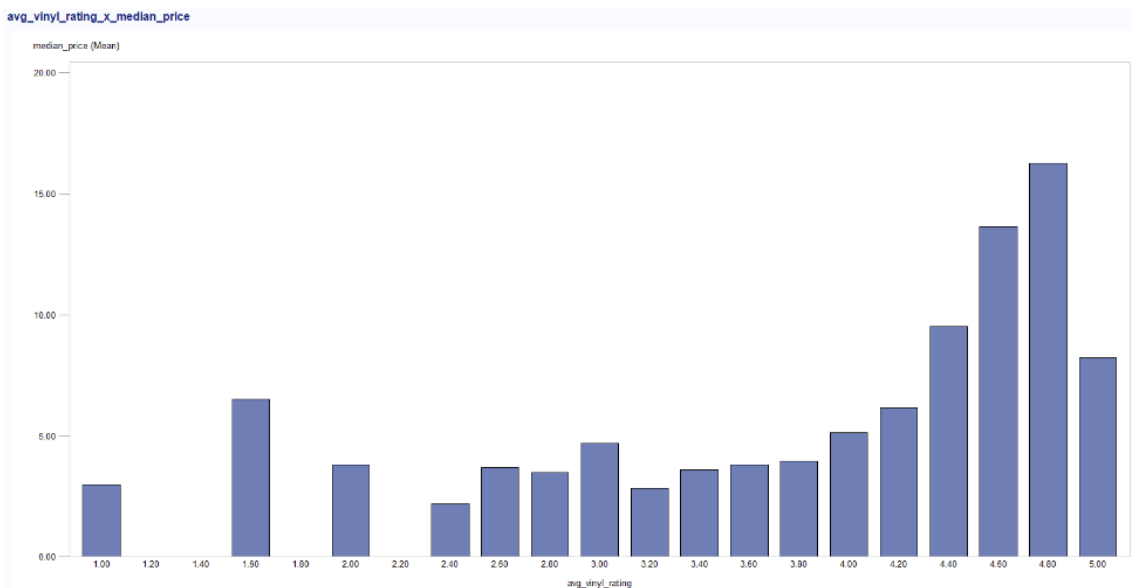
Appendix 9 – Comparison of ‘max_weeks_on_chart’ with ‘median_price’ for Jazz present.



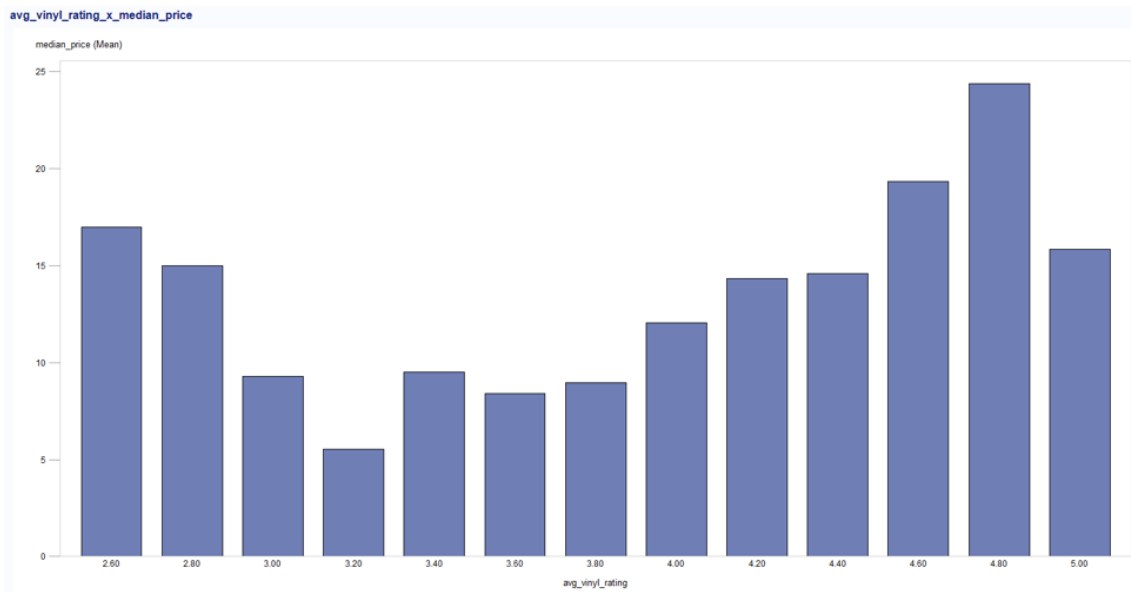
Appendix 10 – Comparison of 'avg_vinyl_rating' with 'median_price' for Rock past.



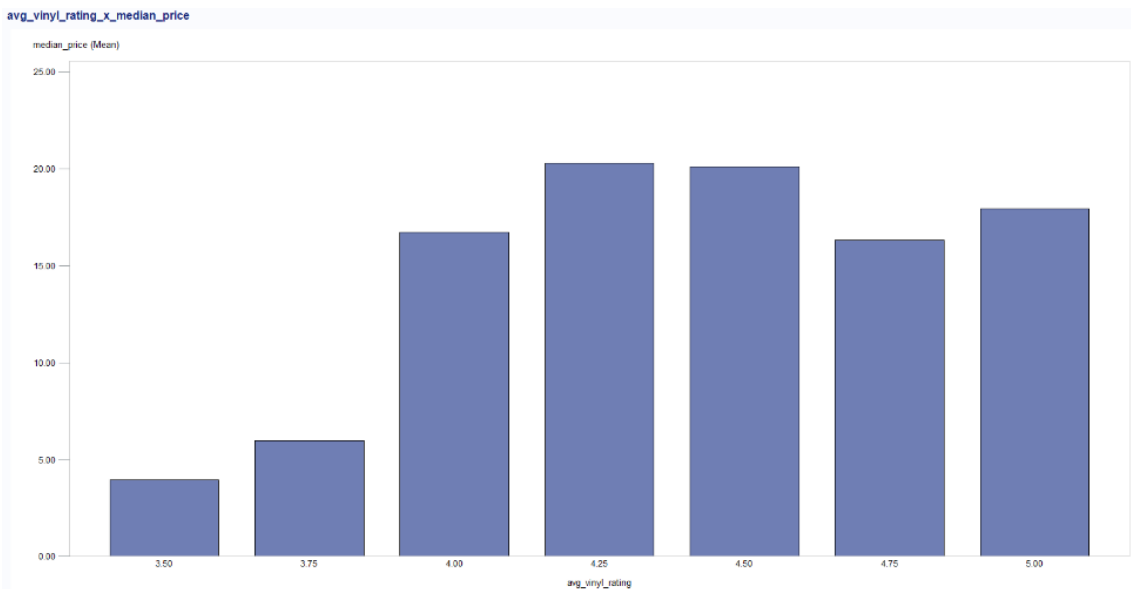
Appendix 11 – Comparison of 'avg_vinyl_rating' with 'median_price' for Jazz past.



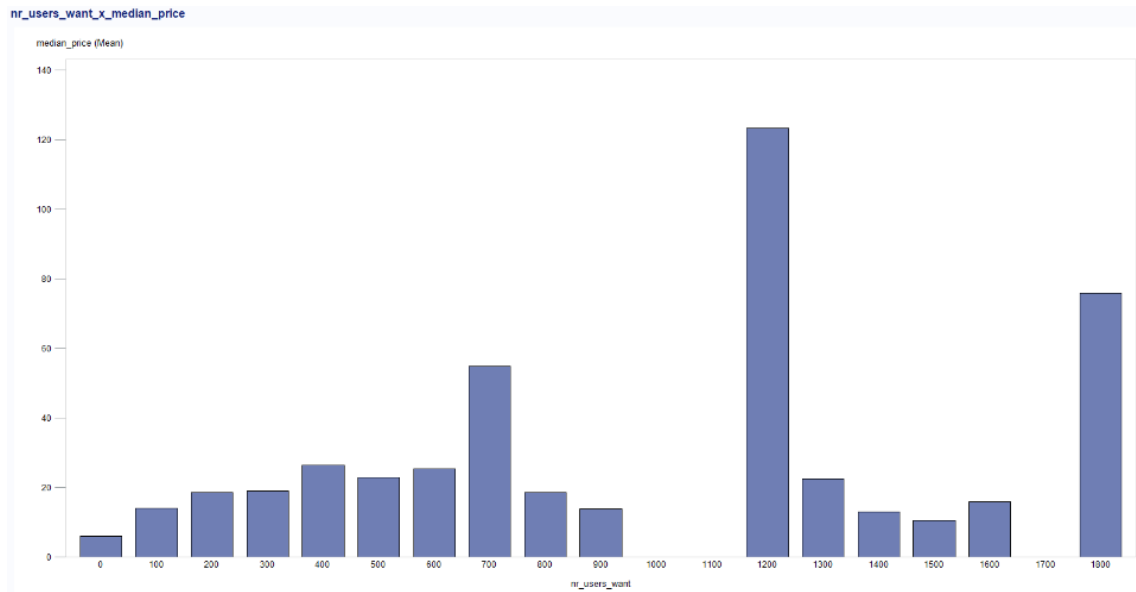
Appendix 12 – Comparison of ‘avg_vinyl_rating’ with ‘median_price’ for Rock present.



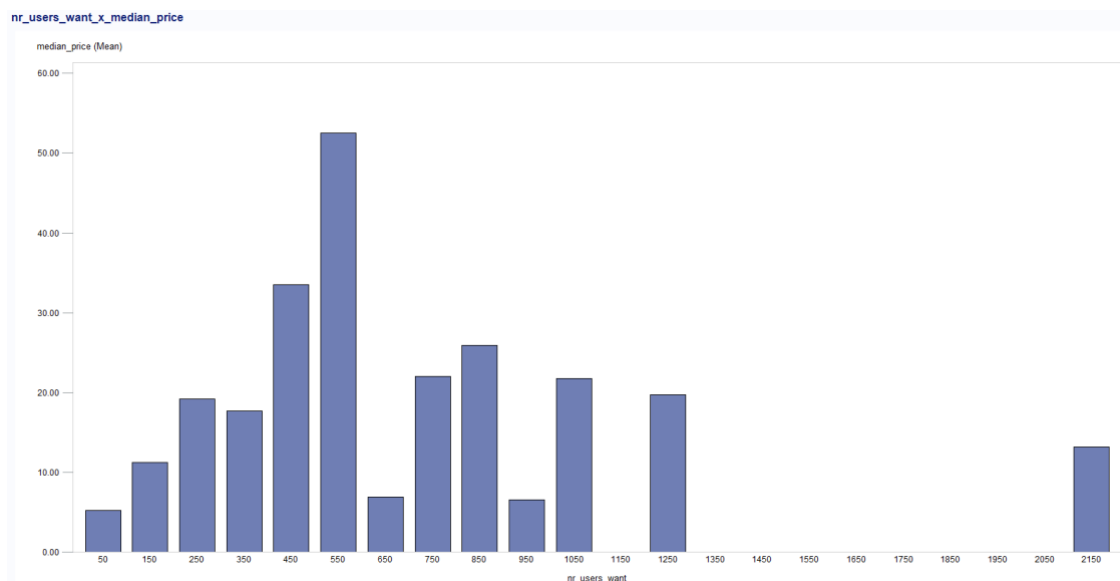
Appendix 13 – Comparison of ‘avg_vinyl_rating’ with ‘median_price’ for Jazz present.



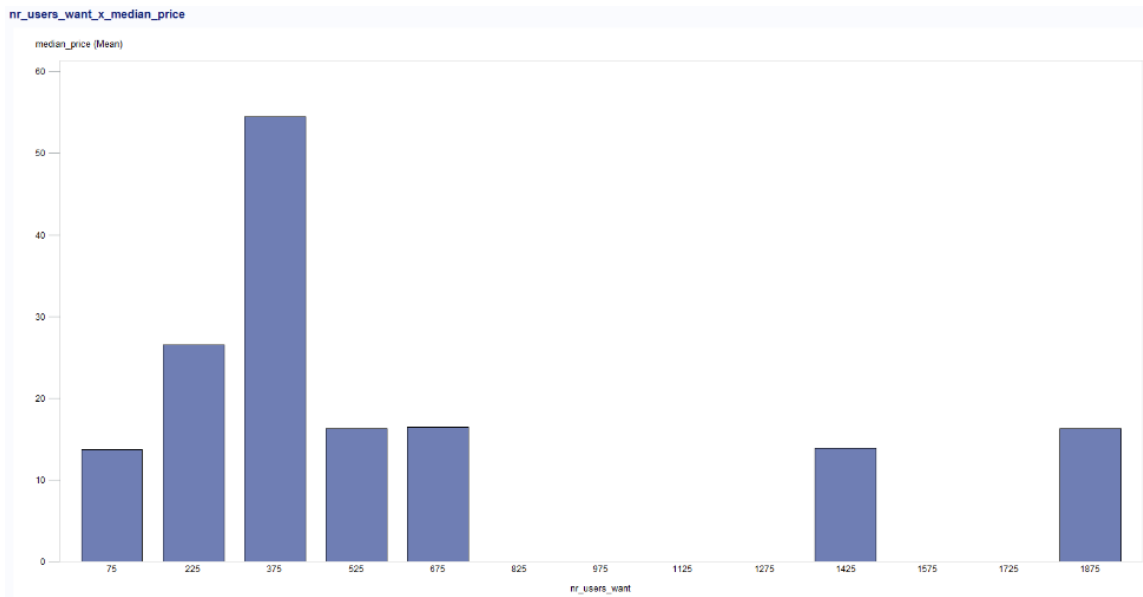
Appendix 14 – Comparison for 'nr_users_want' with 'median_price' for Rock past.



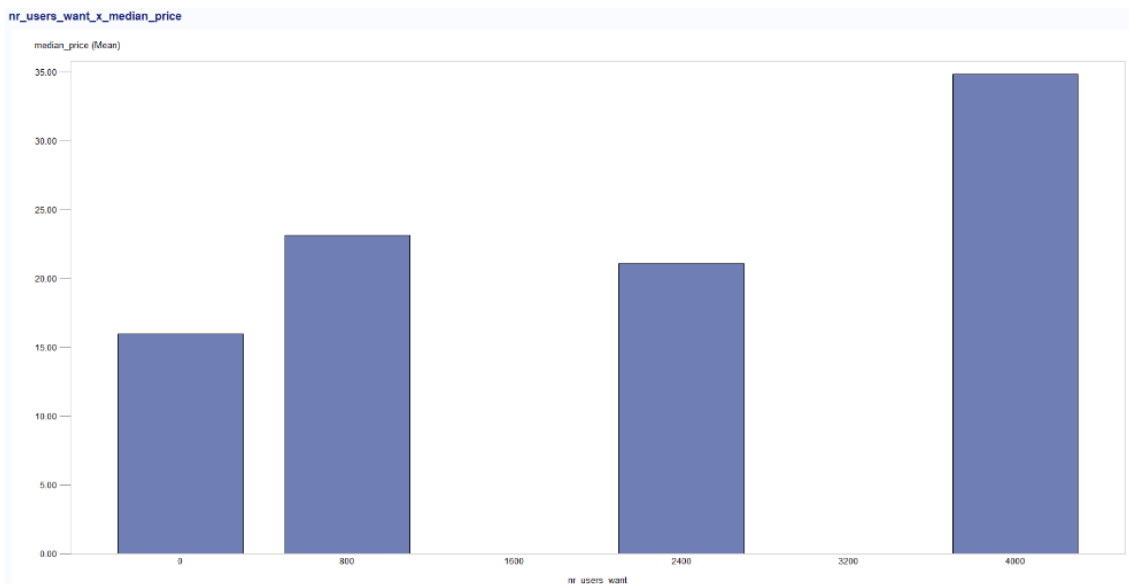
Appendix 15 – Comparison for 'nr_users_want' with 'median_price' for Jazz past.



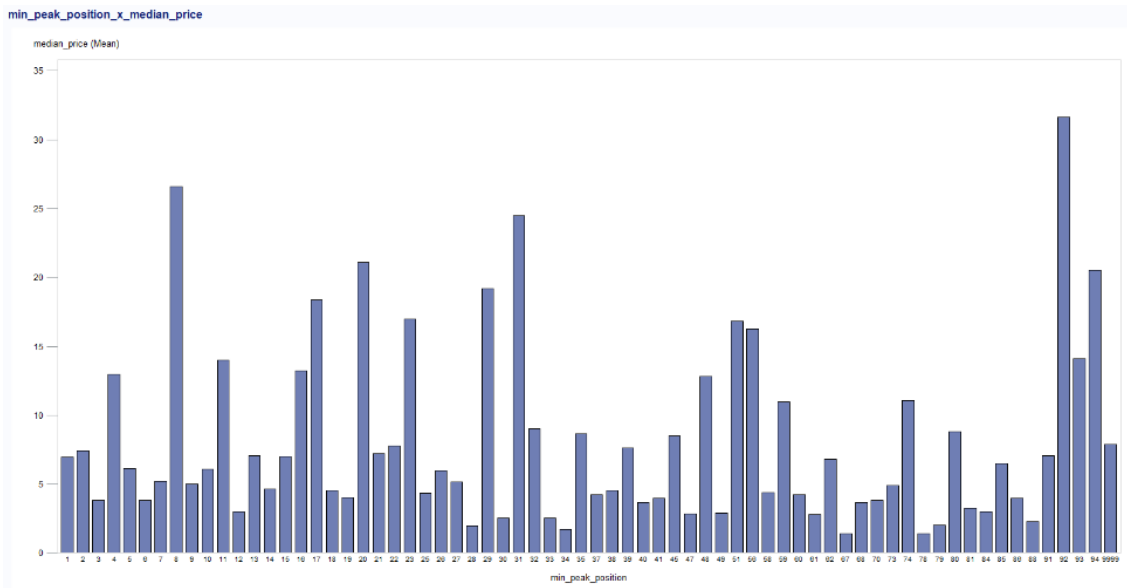
Appendix 16 – Comparison for 'nr_users_want' with 'median_price' for Rock present.



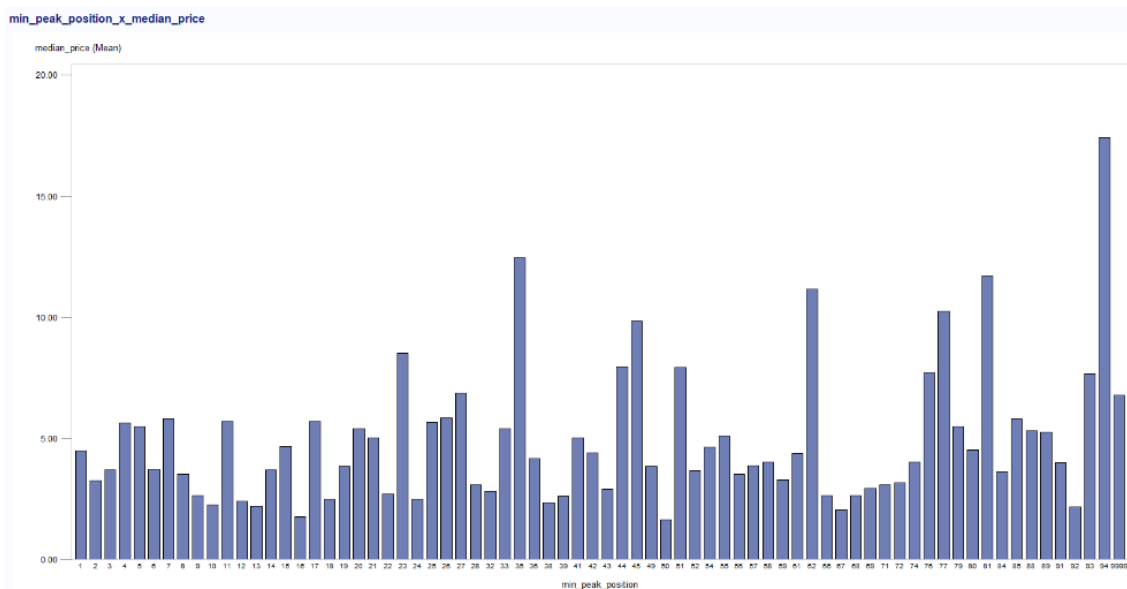
Appendix 17 – Comparison for 'nr_users_want' with 'median_price' for Jazz present.



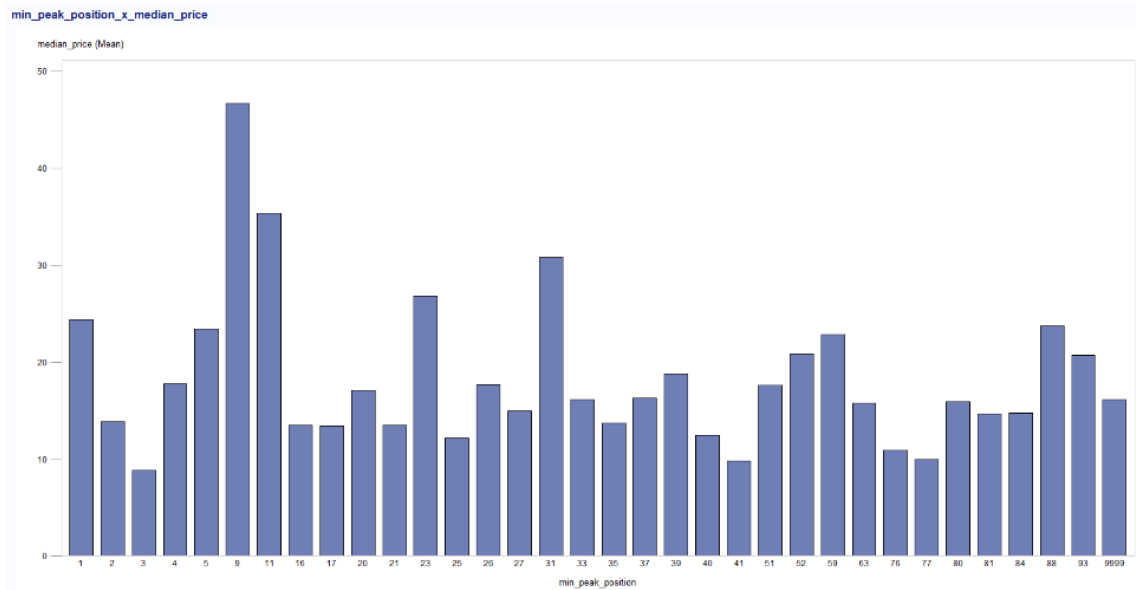
Appendix 18 – Comparison for ‘min_peak_position’ with ‘median_price’ for Rock past.



Appendix 19 – Comparison for ‘min_peak_position’ with ‘median_price’ for Jazz past.



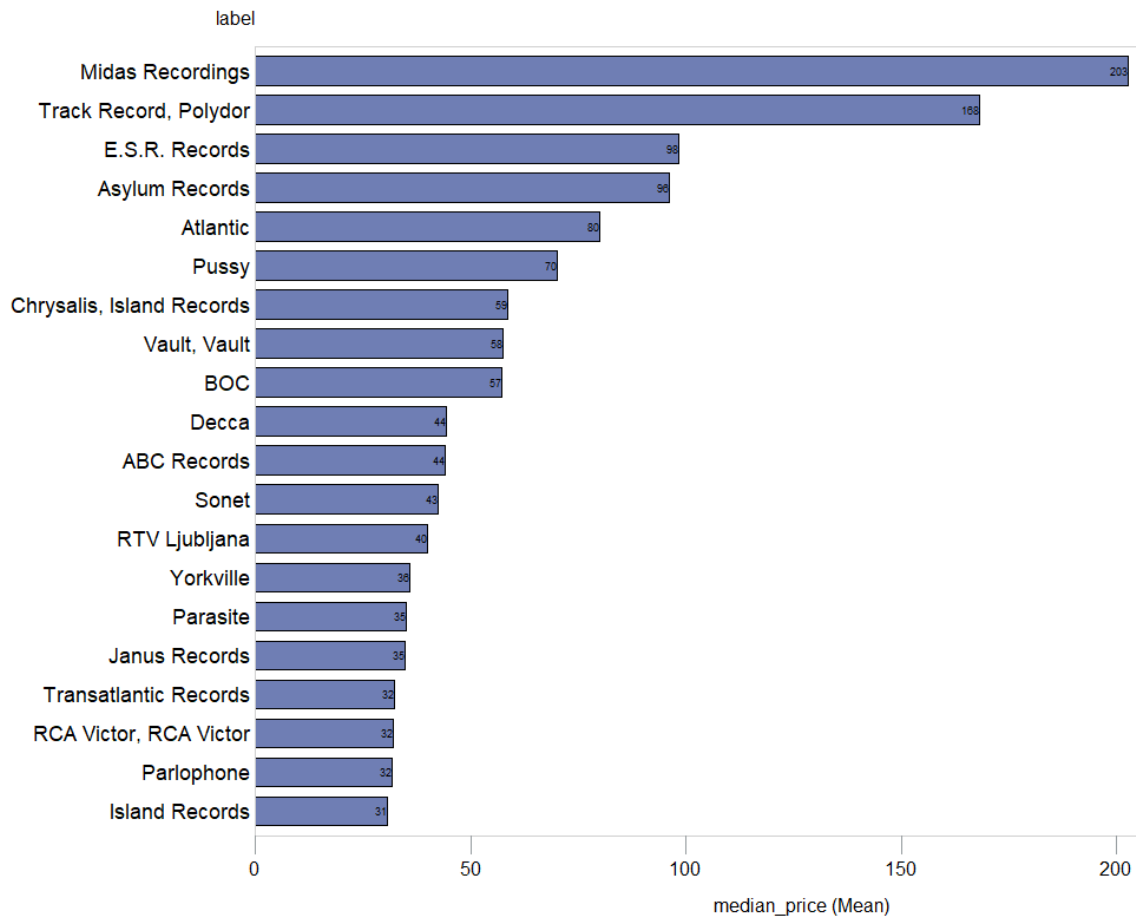
Appendix 20 – Comparison for ‘min_peak_position’ with ‘median_price’ for Rock present.



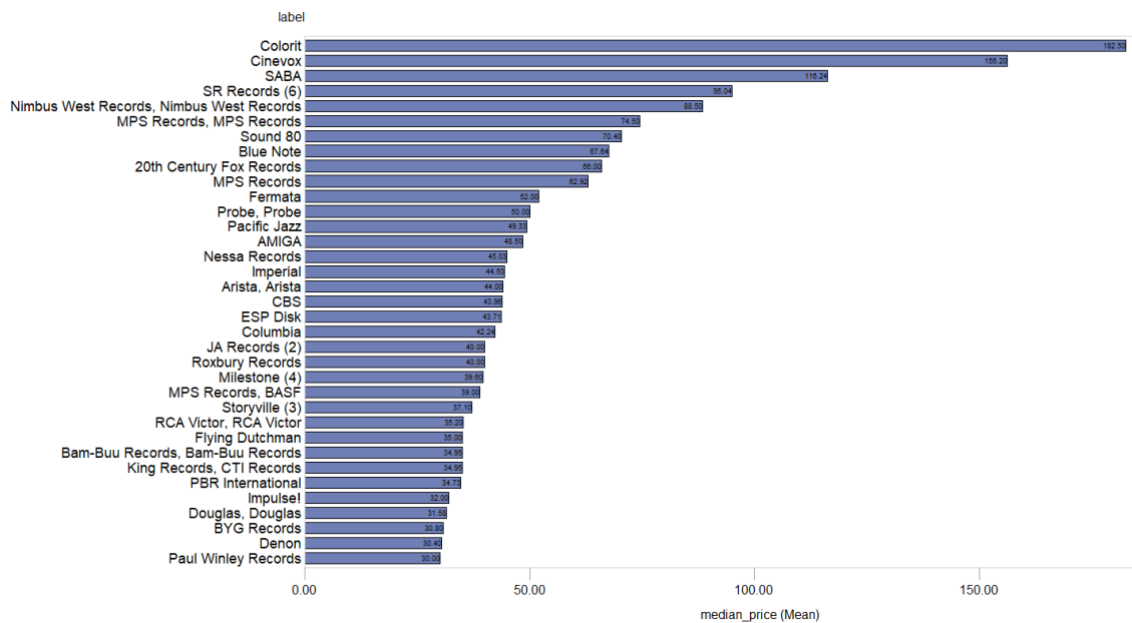
Appendix 21 – Comparison for ‘min_peak_position’ with ‘median_price’ for Jazz present.



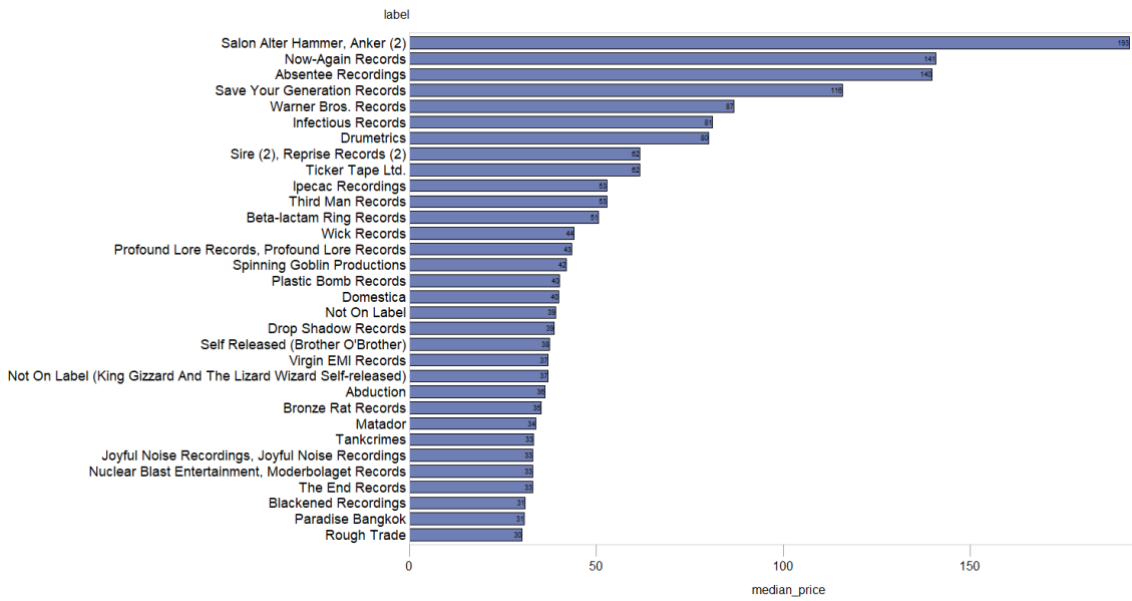
Appendix 22 – Comparison for ‘label’ with ‘median_price’ above 30 euros for Rock past.



Appendix 23 – Comparison for ‘label’ with ‘median_price’ above 30 euros for Jazz past.



Appendix 24 – Comparison for ‘label’ with ‘median_price’ above 30 euros for Rock present.



Appendix 25 – Comparison for ‘label’ with ‘median_price’ above 30 euros for Jazz present.

