What Determines Expert Appreciation of Wine

Robert Ernst: 443776

February 10, 2021

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master Thesis: Data Science and Marketing Analytics
Supervisor: A. Alfons
Assessor: N.M. Almeida Camacho

Abstract

Who does not like a good glass of wine? However, how many people actually know which wines they will like beforehand? As most people do not know the qualities of all different wines on offer, they go to look for expert advice. One of these places to find such advice is the Wine Enthusiast Magazine, a magazine that posts over 25,000 wine reviews each year. Every wine-maker would like their wines to be appreciated and must therefore be interested in what these experts have to say. Doing well on such a platform is something that all wine-makers would like. This paper will therefore perform an in-depth analysis of these wine reviews and give advice to the wine-maker to ensure that their wines are well-received. [1]

[1]The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

Approximately 8,000 years ago, in what is now known as Georgia, people started fermenting grapes in order to produce wine (McGovern et al., 2017). Only later, in ancient Greek and Rome, drinking wine became part of life. Along with the rise of the Roman Empire and the spread of Christianity, drinking wine became a general practice all over Europe. The practice was supported by the Roman Catholic Church, which promoted the use of wine in one of their main sacraments, the Eucharist. A culture of wine arose in Europe, which was slowly exported to the rest of the world when the Europeans set sail in the 15^{th} century. Slowly but surely, people started to produce wine in these countries as well. Wine found its way in world culture.

Initially, wine-makers learned from their peers and predecessors how wine was made. The key process being the fermentation of grapes, which turns sweet non-alcoholic grape juice into the delicious alcoholic beverage, was not well understood. The real base of scientific knowledge behind the microbiological mechanics of the wine production only started to grow after Louis Pasteur's research on yeast and fermentation. Since then, countless numbers of papers have been written on all the different microbiological aspects of the wine-making process.

As a result of this vast base of scientific knowledge on how to produce wine, there exists an extremely wide range of wine aromas and experiences. For example, when ordering a beer in a restaurant or bar, one will know what it will taste like and how much it will approximately cost. For wine, this generally is not so much the case. Not only are there great differences between wines produced by different wine-makers, there are also great differences between wines produced in different years by a single wine-maker. In order to get the wine that one desires, people often ask the waiter or the sommelier for consultation. These people will most likely explain how these different wines will taste and through this explanation one can choose which experience they are most looking for.

However, when going to the supermarket or ordering wine online, one can not always ask someone for advice. What one can do, is search for experiences that other people have of certain wines. One place that many people go to for information is the Wine Enthusiast Magazine, which is the world's largest magazine of its kind and is based on the wine ratings system of Robert Parker. In this magazine, a group of experts blindly review over 25,000 wines per year and give them a rating between 80-100 (*Wine Enthusiast Magazine*, 2020).

Research already showed that there is a positive relationship between willingness to pay and perceived quality of a wine (Landon & Smith, 1997). What this might quite reasonably imply, is that most people would also want to pay more for a better rated wine. Combining this with the immense competition in the wine-making field, producers are required to better understand what makes a great quality wine (Swiegers & Pretorius, 2005). Using the knowledge of what people appreciate in a wine with the mechanics of making it, seems to be the only way for a wine-maker to survive in the 21^{st} century. Given the ever increasing body of scientific literature on wine production, it is the wine-makers duty to apply this knowledge on his production process (Kosseva, Joshi, & Panesar, 2016). How to do this, is greatly reliant on the following question:

What do wine experts value in wine?

A better understanding of the answer to this question will give wine-makers the edge over their competition. In order to answer this question, I will first go over the literature on the sensory and non-sensory characteristics of wine and hypothesize how these characteristics influence the rating of a wine bottle. Subsequently, I will examine the data and discuss the transformations that have to be done to test the formulated hypotheses. In the methodology, I will explain how several methods will be applied to the data so that the hypotheses can be tested. In the results section I will show the results of applying these methods to the obtained data and come to conclusions about the hypotheses.

2 Theoretical Framework

Many papers have discussed wine in a variety of ways. Studies have looked into factors determining prices, for instance between wines in a specific province (Cardebat & Figuet, 2004). Some have dived into the distinction between the New World and the Old World of wine (Thorpe, 2009). Others looked for a hedonistic explanation of wine prices (Benfratello, Piacenza, & Sacchetto, 2009).

In this section, I will split the literature in two parts. Firstly, I will consider the literature on the non-sensory characteristics of wine. For example, the influence of the year or place in which it was produced on its reception. Is a wine produced in Italy generally perceived better than a wine produced in France? This first section will try to establish a baseline as to global differences in wine reception, differences that a single wine-maker can not really influence. Secondly, I will go over the literature on the influence of sensory characteristics in wine tasting. For example, by looking at the process in which wine is tasted and by looking at the different types of aromas that can be present in wine. Getting a clear overview of how wine is tasted and what can be tasted will give direction to the possibilities that wine-makers have in shaping their wines.

2.1 Non-Sensory Characteristics

In this section I will discuss three non-sensory characteristics of wine. Firstly, I will discuss the price of a wine bottle, which can be several dollars in the supermarket or even several hundred thousand dollars at auction. Secondly, I will discuss the theory on the location in which a bottle was produced, by looking at the country in which it was produced. Thirdly, I will look into the literature on the influence of the year of production on the quality of a wine. Why do people speak of a good year or a bad year? In order to examine these non-sensory characteristics, an OLS regression and a random forest will be used.

2.1.1 Price

Prices of a wine bottle can vary greatly between different bottles. Most bottles that one can purchase at the local supermarket cost less than ten dollars. Whereas when going to a wine retailer, wine prices can reach several hundred dollars or even several thousand dollars. At auctions, the prices of a wine bottle can easily exceed this. At an auction in 2018, a bottle from 1945 was sold for the staggering amount of $558,000 (Frank, 2018). So, what drives this great difference in the price of a bottle of wine?

The basic explanation from economics would start to look from a perspective of supply and demand. When looking at this from the demand side, people are most likely willing to pay more for a bottle that they perceive of as excellent than that of which they perceived as ordinary. When looking at this from supply side, things get slightly more complicated. On the one hand, the quantity in which a certain bottle is produced varies greatly between wine-makers. Some wineries have significantly more land to grow the grapes than others, making some wines rarer than others. On the other hand, production processes differ between wine-makers. In the Champagne region in France, it is for instance required by law that the grapes are handpicked (Comité-Champagne, 2020). Whereas wine-makers in many other regions are allowed to use machines for grape picking, which implies a lower cost of production. Furthermore, some wine-makers choose to make their wines organically, meaning to not spray the grapes with chemicals. Which will result in having to throw away many of the grapes as a results of nature messing with the grapes.

Therefore, the price appears to capture the underlying quality as perceived by both the consumer and the wine-makers. From which I arrive at the following hypothesis:

H_1: Expensive wines are perceived better than cheap wines.

2.1.2 Geography

A wine is produced at a winery, but in a more general sense a wine is produced in a certain region, province or country. For example, a bottle of the Italian wine-maker Ornellaia comes from the region called Bolgheri in the province of Tuscany. One could easily argue that wines even from the same region can taste completely different. So, to what extent does the geographical origin of a wine bottle determine its perception?

The broadest division of the geographical location of wine is the distinction of New World and Old World (Robinson & Harding, 2015). The New World consists of countries such as Australia, Chile and the USA, whereas the Old World consists of countries such as France, Italy and Spain. The greatest difference between these two worlds is the climate. Where Old World wines are mostly produced in cooler climates and New World wines in warmer, which can greatly determine the way in which a wine will taste (Puckette, 2020a).

Moreover, these differences in climate also persist between countries. These climates can be further categorized into three categories: Mediterranean, Continental and Maritime (Swan, 2019). The Mediterranean climate is a climate in which there are hot and dry summers and cool and wet winters, this is the climate in for instance the Italian province of Tuscany (Augustyn, 2019). The Continental climate, which can be found in for instance the Italian Piedmont, is characterized by lower temperatures and more rainfall than in the Mediterranean climate (Joshua, 2020). The Maritime climate can be found in regions close to sea, such as the Bordeaux in France, this climate is in terms of temperature and precipitation somewhere in between the other two climates (Jones, 2015).

Given the great impact that the geographic location seems to have on the way in which the wine is produced, I come to the following hypothesis:

H_2: The geographical location of production has an impact on the rating of a wine.

2.1.3 Year

Most people are aware of the fact that there are 'good years' and 'bad years' in wine (Tuttle, 2011). A good year implies that the conditions for grape growth were very good in that year and that the wine is usually very tasty and expensive. Whereas a bad year, does not necessarily mean that a wine does not taste good, but the conditions for growth were sub-optimal in that year. The main reason for a year to be considered a 'good year', was that the climate was favourable for grape growth in that year. A study done in the north and central parts of Italy already showed that a quality improvement could partially be linked to year-on-year increase in temperature and decrease in rainfall (Dalu et al., 2013). In this section I will further entertain the possibility of an influence of a wine's year of production on consumer perception.

Given the strong influence of climate on determining whether a year is good or bad, one needs to realize that climates are different and change differently in different regions. Therefore, it would be very hard to argue that one year would be good or bad for all different wines. For example, when looking at the past 30 years of wine from the Bordeaux region, the years 2009, 2016 and 2018 are considered to be good years (Leve, 2019). Whereas, for example, the best years for wines in the Coonawarra region in Australia were 2001, 2005 and 2012 (Edison, 2018). Whereas the best years for wine from Sonoma region in California are 2007, 2009 and 2010.

Given these individual examples of good wine years, one might expect that experts systematically appreciate wines from one country differently over different years. This thought gives to the following:

H_3: The local existence of a good year causes a higher appreciation for a wine of that year.

2.2 Sensory Characteristics

In economics, goods and services are classified into three categories: search goods, experience goods and credence goods (Ford, Smith, & Swasy, 1988). Many economists classify wine as an experience good, which is a good of which the real value can only be known after consumption (Ashton, 2014). After tasting the wine, not only does one think about what they exactly taste, but even more importantly whether they like it. In this section I will discuss the ways in which sensory characteristics could influence the wine experience.

This section will be divided into the wine characteristics and the wine aromas. I will first go over the wine-tasting process and its constituent elements. The influence of elements such as color, taste and smell on the the experience of wine-tasting will be thoroughly investigated. Furthermore, the wine aromas will be divided into three different categories, which will be further analysed to get to specific answers of what a wine should taste like according to experts. Finally, hypotheses of different levels, between and within these fundamental aspects will be formulated.

2.2.1 Wine Characteristics

In order to consistently review wine, one needs a sequence in which to carefully do so. Most of these sequences consist of three types of evaluation, namely through sight, smell and taste (Cozzolino et al., 2008). In the Wine Enthusiast Magazine a detailed process containing these three types of evaluation is given (Gregutt, 2020).

For evaluation through sight, the magazine suggests a process consisting of four steps. Firstly, one looks into the glass from the top to get a sense of the depth of the color. Secondly, one views the wine from the side in order observe the wine's clarity. Thirdly, the observant tilts the glass slightly in order to get an idea about the age of the wine and the level of alcohol. Finally, the wine should be swirled in the glass and one should look for tears after the wine has re-stabilized on the side of the glass, which can give an indication of both the level of sugar and alcohol in the wine.

For evaluation of wine by smell, there are also several steps to be done to better understand the properties of the wine. In the first place, one takes a few quick sniffs in order to possibly find some basic flaws in the wine. For example, sometimes a wine will smell like "a musty old attic" which could mean that is corked (Gregutt, 2020). After checking whether the wine is off, one can take a deeper dive into the smell and look for fruity or floral aromas. Besides these aromas that mostly have to do with the grape itself, the observant can pick up aromas that could give a hint on the type of barrel that was used. For example, whether it was contained in an old or a new barrel.

Lastly, one gets to evaluate the wine by tasting it. One tastes the wine by taking a sip of wine and letting it circulate through your mouth and taking small gasps of air. From this sip, the observant can judge on five different aspects. First, whether the wine is balanced, which means whether there is a balance of the basic tastes like sweet and sour. Second, whether the wine is harmonious, which means checking if all flavours are presented in a proportionate way. Third, one can judge the complexity of the wine, whether a wine consists of many different and unique flavours. Fourth, especially in these complex wines, they can evolve in the glass over time as they are in contact with the air, possibly yielding new flavours. Finally, one can judge whether a wine is complete, which basically is a measure of the combination of the other four measurements.

As a result of this process one can come to a great variety of conclusions. These conclusions mostly contain information on five different and distinct wine characteristics (Puckette, 2020a). Firstly, sweetness, wines can either be sweet or dry, or somewhere in between. An example of a very sweet wine is for instance a Sauternes, which is usually paired to a sweet dessert (Brewczynski, 2018). Whereas wines made from the Sauvignon Blanc grape, are often perceived as rather dry (Robbilard, 2020). Secondly, acidity, the level of which determines the way the wine feels in the mouth. Acidity is usually also

prescribed to the Sauvignon Blanc grape. Thirdly, wines have a level of tannin, which are residues left from the wine production, the degree of which determines the level of bitterness in the wines. Expensive wines often have greater levels of tannin and the tannin in older wines is usually a lot smoother (Puckette, 2017). Fourthly, the level of alcohol present in the wine, which is usually somewhere between the 11 and 13 percent. Finally, there is a characteristic named body, which captures a great variety of factors, which allows one to say whether a wine is full-bodied or light bodied. Wines with higher levels of alcohol are often also full-bodied wines, these wines tend to combine well with fatty foods.

When all these five wine characteristics are at the right level, one can speak of a harmonious wine. In a 2009 study, Benfratello et al. investigated the impact of sensory characteristics on the prices of Italian wines. The authors found harmony to be the single most important and statistically significant. This paper will take a closer look at the components that together make harmony and see in what ways harmony can impact the rating of wines in general.

Besides these overall impressions of wine, captured by the wine characteristics and the level of harmony, one can also taste specific flavours in wine. In the next section, these flavours will be subdivided into three different types of aromas of three different origins.

2.2.2 Wine Aromas

The perception of wine quality is greatly influenced by wine aromas. These aromas are highly complex and are formed by what is known as volatile compounds, such as alcohols, acids and esters (Villamor & Ross, 2013). These compounds can have an effect on wine aromas at even relatively small dosage. In wine, the effect of the presence of these compounds on the aroma is threefold. Their presence has an effect on what is known as the primary, secondary and tertiary aromas (Mercer, 2020). In this section I will go over these three different types of aromas and why they matter.

The primary aromas mainly arise from the grapes themselves, the terroir and partly how these grapes are processed (Rapp & Mandery, 1986). The majority of aroma derived from the grape is located in the skin of the grape and different types of aromas reside in different varieties of grapes (Villamor & Ross, 2013). The aroma of the Chardonnay grape for instance, can often be associated with apples and lemons. While most of the wines are made from a single grape, for example 100% Chardonnay, whereas other wines can be made from multiple types of grapes, which is called a blended wine. Through blending these different types of grapes in certain proportions, wine-makers can to some extent control the aroma of the wine.

Besides the choice of grapes, the nature of the terroir is of great significance to the taste of wine (Puckette, 2020b). Terroir is an old French term used to describe the environment in which the grapes are grown (Puckette, 2020b). Primary points of concern when talking about terroir are the climate of the region, the composition of the soil and several other aspects of the terrain of the vineyard. A warmer climate will cause the grapes to have a higher sugary content which will in turn ensure that the wine is higher on alcohol. Similarly, a cooler climate will cause lower levels of sugar and higher levels of acidity.

These primary aromas can mainly be classified in three sub-categories: fruity, herbal and floral (McKirdy, 2019). With fruity flavours such as blackberry, cherry and citrus. Herbal flavours such as thyme, bay leaves and parsley. Among the floral flavours are flavours such as elder-flower, jasmine and lavender. My expectation is that all of these flavours matter, but that the fruity flavours supersede given that grapes are a fruit in themselves.

The secondary aromas of wine arise when the fermentation process is initiated and later when wine is aged in a barrel (Tilden, 2020). The primary aromas that these grapes contain in themselves are further transformed when the wine-maker adds yeast to the grapes. Adding yeast to the grapes starts the fermentation process and creates many flavour compounds, which extrapolate to these secondary aromas.

In the fermentation process, yeast converts glucose and fructose present in the grapes

to both ethanol and CO_2 (Swiegers & Pretorius, 2005). What remains consists mostly of water and alcohol. Just 1% of the wines usually consists of flavour compounds produced by the fermentation process. Swiegers and Pretorius (2005) name the 6 most important flavour compounds produced by this process: esters; higher alcohols; carbonyl compounds; volatile acids; volatile phenols and sulfure compounds. Where these esters are responsible for most of the fruity flavours created in this process. The carbonyl compounds are mostly responsible for creating apple and citrus-like flavours on the one hand, and buttery aromas in higher concentrations and nutty aromas in lower concentrations on the other hand (Swiegers & Pretorius, 2005). These volatile phenols should be present in low concentration, as in high concentrations it will cause the wine to smell like a stable. Like the phenols, the sulfur compounds can create off-flavours. In the case of sulfur, the wine could smell of rotten eggs and garlic. For a wine-maker it is important to know about these compounds and how they should be altered in order to get to the desirable end-state

The secondary aromas that are generated in this process can be subdivided into three categories: yeast, brettanomyces and malolactic fermentation (WineScribble.com, 2020a). In the first category, yeast, there are aromas to be found such as bread, cheese and beer. In the second category, brettanomyces, which are yeast that are not added on purpose, flavour as cinnamon, black pepper and bacon can arise. In the final category, malolactic fermentation, which is caused by some bacteria, aromas such as hazelnut, cream and chocolate arise (WineScribble.com, 2020a).

The tertiary aromas, also known as the bouquet, are the result of aging the wine. This aging occurs primarily in oak barrels, but continues inside the bottle. The result of this aging-process in the barrel is mainly dependent upon the wine's exposure to oxygen, the provenance of the oak barrel and whether the barrel is new or re-used. The effects of oxygen are mostly the production of certain nutty aromas such as that of walnuts and hazelnuts. The maturation process in the oak barrels gives of woody aromas to the wine (WineScribble.com, 2020b). Aromas such as vanilla, cinnamon and cedar can be brought forward. Inside the bottle, some fruity flavours might further develop, into flavours such as that of dried fruits, smoke and tobacco (McKirdy, 2019). The wine in a sense gains in complexity over time.

2.2.3 Hypotheses of Sensory Characteristics

The hypotheses for the sensory characteristics will be constructed in several levels. These levels will be ordered from global differences to aroma-level differences. The first expectation is that wines that are graded far apart from each other exhibit vastly different characteristics and aromas. The hypothesis is stated as follows:

H_4: Experts talk differently about wine characteristics and aromas in different point-categories.

The point-categories are the sum of all wines that lie within a certain range of points. For example, wines that score between 80 and 86 points, will be classified as a good wine. Whereas wines with over 90 points will be classified as an excellent wine.

In order to make any statements on this hypothesis, a Latent Dirichlet Allocation will be executed. The LDA will show the latent topics that are present in these different point-categories (as suggested by the Wine Enthusiast Magazine). A visual representation of these different latent topics will guide the conclusions about this hypothesis.

As suggested by Benfratello et al. (Benfratello et al., 2009), harmony is a driving factor for the very expensive wines. One would expect that for the highest classes of wine, people would experience a symphony of structure and complexity in their mouth. Whereas for the mediocre levels of wine several more basic specific flavours would be expected, such as that of a particular fruit. This gives rise to the following hypothesis:

H_5: As wine quality increases, the discussion of wine characteristics will dominate those of specific aromas.

If this hypothesis were to be true, the implications could be observed in several ways. A close look into the different latent topics could show a different relative presence for these two aspects in the different point-categories. Furthermore, a combination of the results of the OLS and the random forest could indicate the relative appreciation and importance of indicators of these two aspects.

Since the mean and median score (88/100) are closer to being a acceptable/good wine than to an excellent/superb wine, it is perhaps more interesting to look at the drivers of quality in the lower half of the point-scheme. As earlier hypothesised, the wine aromas are expected to dominate this part of the distribution. From the discussion of the wine aromas, the variation of the use of grapes appears to be the most fundamental. Followed by the degree of which the fermentation process can be altered and to some the degree the way the wine is stored. From which the following hypothesis follows:

H_6: The level of importance of the wine aromas are in the same order as they are numbered.

This would imply that the primary aromas are the most important and the tertiary aromas are the least important. There are several ways to check this hypothesis. They all start with recognising the underlying words used to described these aromas. For example, the word "apple" would denote a primary aroma, whereas the word "oak" would denote a tertiary aroma. The presence of these words can be checked in the results of the LDA. These words are present in the components of the PCA. Both the random forest and the OLS can show the impact of unigrams or bigrams representing these words and through components that cover multiple signifiers of a particular aroma.

To go down one more level, a closer inspection of the primary aromas will performed, as they are deemed most important. As wines are made of grapes, which are fruits. One would expect that identified aromas would be closest to other fruits. Which is why the following hypothesis is stated.

H_7: Fruity aromas are the most important aromas of the primary aromas.

In order to test this hypothesis, the unigrams, bigrams and principal components will be carefully examined. If the hypothesis were to be true, one would expect a relative over-representation of fruity flavours in the results of the different models. A reading of the coefficients and levels of importance of these different fruity flavours might yield specific advice to the wine-maker as to what specific aromas are appreciated and how these aromas can be achieved.

3 Data

The data for this thesis was obtained from Kaggle: `https://www.kaggle.com/zynicide/wine-reviews?select=winemag-data-130k-v2.csv`. Kaggle is an online community owned by Google and offers data sets to data scientists for non-commercial purposes. The data set originally consisted of 129,970 rows and 14 columns, where every row represents a review of a particular wine featured in the Wine Enthusiast Magazine. After removing the duplicates, the data consists of 119,988 unique wine reviews and after removing the variables containing the name of the taster and his Twitter name were removed only 12 columns remained.

The Wine Enthusiast Magazine is an American wine magazine that publishes thousands of wine reviews a year. People can send their wines to the company's headquarters, where a wine-tasting expert will review the wine and give it a score. This score is captured by

the *points* variable, which is the dependent variable in this paper. The score ranges from 80 to a 100 and can be further divided into several classes (Winemag.com, 2020):

Points	80-82	83-86	87-89	90-93	94-97	98-100
Descriptions	Acceptable	Good	Very Good	Excellent	Superb	Classic

Table 1: This table shows the scoring system that is used by the Wine Enthusiast Magazine

Both the mean and median number of points scored is approximately equal to 88 and moreover, the acceptable and classic class are sparsely populated, with respectively 2900 and 129 reviews. In order to get meaningful categories of approximately equal size, the data was split into three categories, as can be seen in the following table:

Points	80-86	87-89	90-100
Descriptions	Good	Very Good	Excellent
Number of Reviews	32681	41717	45590

Table 2: This table shows the point-categories used to test the fourth hypothesis

Using the *sample.split* function from the *caTools* package in R, the full data set is equally split over the dependent variable on a train and test set to be used for the OLS regression. Using this same function, the data was split on a 1:3 ratio (train/test) for the random forest as to decrease computational time. In the following sub-sections the other relevant variables will be discussed.

3.1 Data Transformations for Non-Sensory Characteristics

In this section will be shown what data transformations and reductions have to be performed in order to test the first four hypotheses.

3.1.1 Price

An important variable in the data set is the *price* variable, which resembles the purchase price in dollars for a particular bottle of wine. The prices in this dataset range from $ 4 to $ 3300, with a mean price of $ 35 and a median price of $25, which suggest that the price distribution is skewed to the right. This suspicion is proven in the following density plot:

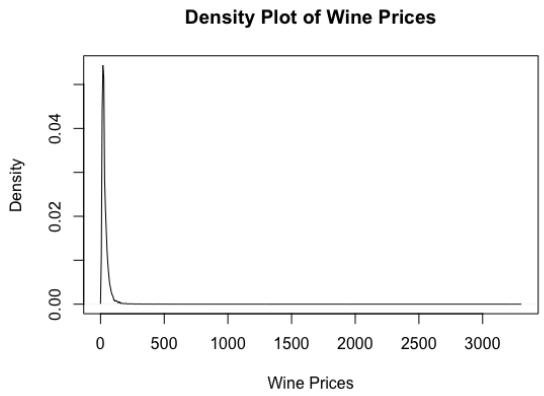

Figure 1: This is a density plot of the wine prices in the data set

The distribution of the price variable is skewed to the right, as some wines are excessively more expensive than the typical wines. For the random forest this will not be an issue, as it will find a threshold of price to split the data over. For the OLS regression, this skewness

will cause the coefficient estimate to be biased. Therefore, it will also be wise to take the logarithm of the price variable to counter this bias. Besides, several 0 values were present for the price variable, these were imputed with the median level of $ 25.

3.1.2 Location

The dataset contained multiple variables indicating the location at which a wine was produced: *country*, *province*, $region_1$, $region_2$ and *winery*. However, in order to test the effect that the location has on the rating of wine, one has to construct meaningful levels of location. It is of course interesting to know that wines from one particular region score very well, but for statistical testing and the bigger picture, having thousands of regions is not so beneficial. Therefore, only the differences between countries will be considered.

In order to see if there are differences between countries, I decided to only consider the countries that had many wines tested by the magazine. Countries like China, Egypt and Slovakia have only produced one bottle that is reviewed by the *Wine Enthusiast Magazine*. Therefore, I wanted to reduce the number of countries to those that have a contribution of more than 1% wines, which reduces the number of unique countries from 43 to 12. All wines from a country with less than 1% presence were bundled in one separate category called 'Other'. These countries are coded as separate dummy variables for the random forest.

These 12 unique countries can be split up into countries from the Old World and countries from the New World. The New World wines mostly come from these countries:

Country	USA	Argentina	Chile	Australia	New Zealand	South Africa
#	49585	3502	4159	2130	1270	1286

Table 3: New World Wines

What comes to the eye is that American wines constitute nearly half of the wines present in the sample. This most likely has to do with the fact that it is an American magazine and people mostly send their wines to the magazine in order to be reviewed. Whereas all other New World countries have rather similar numbers of wines reviewed. This distribution is rather different for the wines from the Old World countries:

Country	France	Italy	Spain	Portugal	Austria	Germany
#	18909	17022	5656	4960	2996	1978

Table 4: Old World Wines

France and Italy have rather similar numbers of wines reviewed, with quite a gap between themselves and the rest of the Old World countries. A total of 2266 wines are included in another category resembling all the countries with relatively low levels of presence.

3.1.3 Year

When looking at the years in which the wines were produced, one finds 91 different years of productions. When reducing this number of years to the years in which over 1% of the bottles were produced, the number of unique years drops to just 13. Now the data set only consists of wines made between 2004 and 2016 and a category which contains the wines outside this domain, which are some wines from 2017 and all of the years before 2004. The wines are distributed over these years as follows:

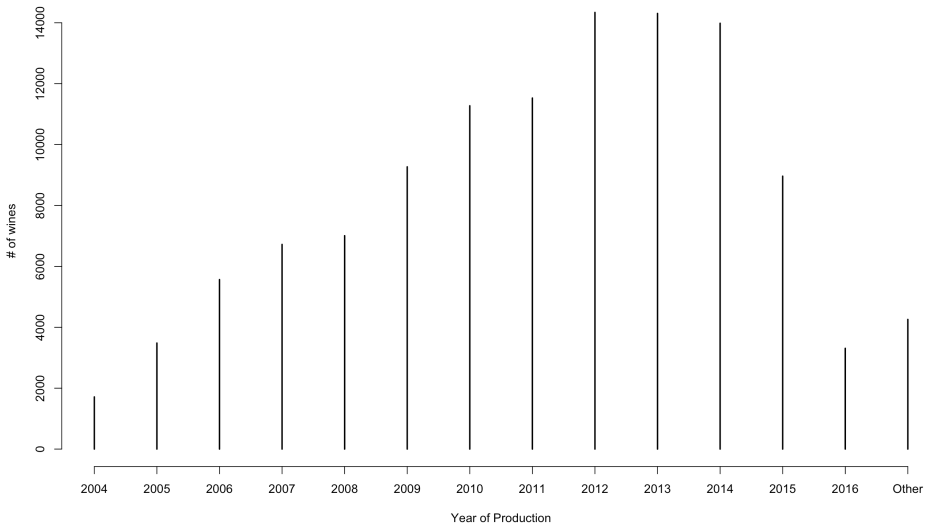

Figure 2: This figure shows the number of wines produced between 2004 and 2016

This figure shows that the most common years for a wine to be produced in are 2012-2014 and that the number of wines selected from a specific year rise quite steadily from 2004-2012. In order to test the hypothesis related to the year of production, the interaction effects are taken between the country variables and the dummy variables for each separate year.

3.2 Data Transformations for Sensory Characteristics

In this section I will explain how the data is prepared to investigate the sensory characteristics. Two models that are used specifically for this purpose are LDA and PCA, the exact functioning of which will be further discussed in the methodology section. These models both require a document-term-matrix as inputs, what this is and how it is obtained is explained in the next section. Lastly, n-grams will be used to test the impact of specific aromas or characteristics, what they are and how they are obtained will be explained afterwards.

3.2.1 Document-term-matrix

The *description* variable in the dataset contains all descriptions of all the reviewed wines. In order to perform the LDA and the PCA, the descriptions that are given have to be pre-processed. All punctuation marks and numbers are to be removed and all words have to be stemmed. To be stemmed meaning that all words are reduced to their base form. For example, 'fruity', 'fruits' and 'fruit' would all be stemmed to 'fruit'. After having done this a document-term matrix is created. A document-term-matrix is a matrix where the rows represent all the unique reviews and the columns represent all stemmed terms present in all of the descriptions in the dataset. In R, this is primarily achieved by applying the *tidytext* and the *tm* package using the *dtm* function. A DTM would look as follows:

DTM	$term_1$	$term_2$	$term_3$
doc_1	2	0	1
doc_2	0	1	1
doc_3	0	0	1

Table 5: Example of a Document-Term-Matrix

Where the values in this matrix resemble the number of times a specific term is used in a

specific document. In this example, the first term appears two times in the first review and the second term appears once in the second review.

For the LDA, the first step is to find the optimal number of topics, k and the optimal level of α. In order to find these optimal levels, the datasets for the three categories are split into a training and validation set, in a 1:1 ratio using the *sample* function.

3.2.2 Unigrams and Bigrams

In order to test for the impact of specific aromas or characteristics on the rating of a wine, unigrams and bigrams are created. These are examples of n-grams, where the unigram represents one word and a bigram a combination of two words appearing next to each other. These n-grams are added as variables to the data set if they surpass a certain threshold, in this case if it is mentioned over a 1,000 times. How this works is, for example, an expert speaks of 'white peach' in a review, than the bigram 'white peach' will take the value 1.

These n-grams are created by using the *unnest_tokens* function from the *tidytext* package in R.

4 Methodology

In order to test my hypotheses, I have to apply certain methods and models to my data. In order to reduce the dimensionality of the vast number of words, both a Principal Component Analysis and a Latent Dirichlet Allocation will be performed. The factors created by the PCA will be used as an input for both the OLS regression and the random forest. Whereas the latent topics generated by the LDA will be used to visualize the underlying structure of the reviews.

In order to test the effect that certain features may have on the rating of wine both an Ordinary Least Squares regression and a Random Forest will be executed. In this section I will go over the functioning of these four models.

4.1 Principal Component Analysis

In this section I will go over the Principal Component Analysis. Firstly, I will explain the steps that are taken to create these principal components. Secondly, I will go over the nature of these principal components and how they ought to be interpreted.

4.1.1 Calculating the Principal Components

One way to analyse the expert reviews, is to consider the words that they use in their reviews. However, when considering over 100,000 reviews, carefully reading all the reviews is not really an option. One way of going about this, is to find out which words tend to appear together in a review, a method capable of doing so is PCA. The end goal of PCA is to create principal components that explain most of the variations of the use of the words in the body of reviews. The starting point of PCA is a document-term-matrix, which is a matrix containing all the documents, which are the reviews, and terms, which are all the stemmed words in the reviews. If a term appears once in a document, the value of the corresponding cell in the matrix is 1. An example of such a document-term-matrix is the following:

DTM	$\mathbf{term_1}$	$\mathbf{term_2}$	$\mathbf{term_3}$
$\mathbf{doc_1}$	2	3	1
$\mathbf{doc_2}$	4	4	3
$\mathbf{doc_3}$	6	5	5

Table 6: An example of a document-term-matrix

What this matrix shows is that, for instance, the first term appears twice in the first document and the second term appears 4 times in the second document. In order to get from this document-term-matrix to the principal components, several steps need to be performed (Dubey, 2018).

When trying to find the underlying patterns in the use of words, one is essentially interested in deviations from the mean. For this, the mean for all dimensions in the document-term matrix need to be obtained, which are represented by $\bar{\mathbf{A}} = \begin{bmatrix} 4 & 4 & 3 \end{bmatrix}$. This would imply that the first term on average, appears four times in the three documents.

In order to find the deviations from the mean use of a particular term, a covariance matrix needs to be computed. What a covariance matrix shows is the direction of the linear relationships between the different terms (Saha, 2018). Where a positive covariance implies that terms are more likely to appear together in a document and negative covariance implies that terms are less likely to appear together in a review.

In order to exactly calculate the covariance between the different terms, the following formula can be used:

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=n}^{n} (\mathbf{x}_i - \bar{x})(\mathbf{y}_i - \bar{y}) \tag{1}$$

Where $\mathbf{x}$ and $\mathbf{y}$ denote the vectors of values for specific terms, with their respective n, number of elements x_i and y_i. The $\bar{x}$ and $\bar{y}$, which are the sample means, are subtracted respectively and these differences are multiplied and summed to get the covariance between the terms. Computing this for all terms yields a covariance matrix, which for the above shown example would be:

$$\begin{array}{c} \text{terms} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \mathbf{4} & 2 & 4 \\ 2 & \mathbf{1} & 2 \\ 4 & 2 & \mathbf{4} \end{bmatrix} \end{array}$$

Where the three values in the diagonal represent the variance of the three terms. From which can be seen that the first and third term vary the greatest. Moreover, the first and third term tend to vary with the same sign, indicating that for this example, if term 1 was used more often, than term 3 was also used more often.

From this covariance matrix, the eigenvectors and eigenvalues, which lie at the heart of PCA, can be computed (Raschka, 2015). Where the eigenvector represents the principal components, and of which the eigenvalues determine the magnitude. A 2-dimensional representation of an eigenvector is the following:

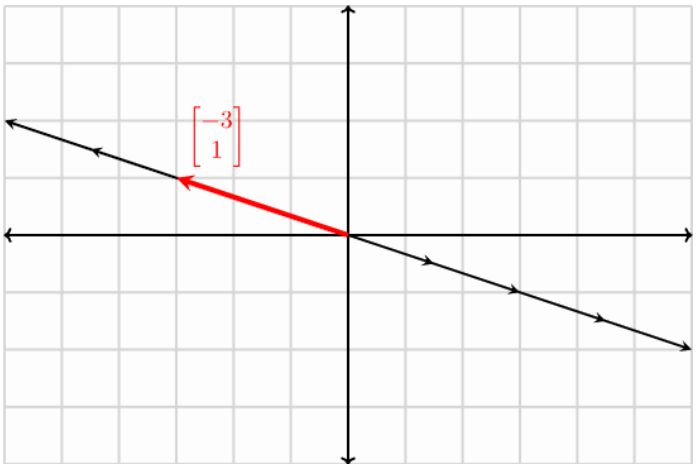

Figure 3: Example of a vector plotted on the XY plane showing it's span (Adewumi, 2019)

The eigenvector in this case represents a line through the origin and the coordinate (-3,1). The height of the eigenvalue consequently determines to what extent this line is stretched.

The properties of these eigenvalues, λ, and the eigenvector, $\mathbf{v}$, need to satisfy the following equation $\mathbf{Av} = \lambda\mathbf{v}$, where $\mathbf{A}$ represents the covariance matrix.

In order to solve this equation, both sides are multiplied with the identity matrix, $\mathbf{I}$, combined and equated to zero, which yields: $\mathbf{A} - \lambda\mathbf{Iv} = \mathbf{0}$. To find the value for λ, the following must be solved: $\mathbf{det}(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$.

After finding all different levels of λ, the corresponding eigenvectors can be found by calculating $\mathbf{Av} = \lambda\mathbf{v}$ for all λ. The eigenvectors are then to be sorted by their eigenvalues. So, that the first principal component is the eigenvector with the highest eigenvalue and the second principal component is the eigenvector with the second highest eigenvalue. These two eigenvectors show the directions in which the data has the highest levels variance. Graphically, the eigenvector with the second highest eigenvalue lies perpendicular to the eigenvector with the highest eigenvalue:

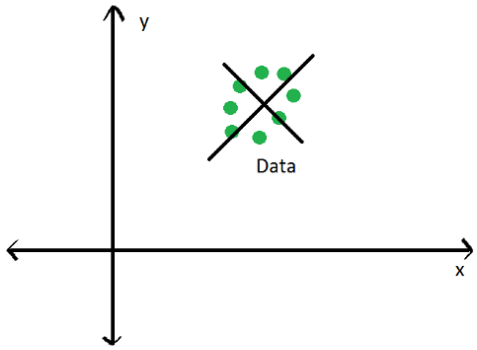

Figure 4: Illustration of the subspace among which the samples are spread out in relation to the principal components (Adewumi, 2019)

In order to more easily interpret the reviews with respect to these principal components, the data can be multiplied with the eigenvectors. This can be done in R with the *varimax* function from the *stats* package. As a results the correlation of the reviews with these eigenvectors is pushed furthest towards a correlation of 1 or 0. A loading close to 1 would indicate that a review is positively correlated to a component. Whereas a loading of 0 would indicate that there is very little correlation between a component and a review.

After rotation, the importance of these words can be sorted by their factor loadings. The words having the highest loadings, can be seen as the factor labels. Using these factor labels, the factors can be characterized as pertaining to specific aspects of the wine-tasting procedure.

In order to find the optimal number of factors to used, the proportion of variance explained is plotted over the number of principal components. When the proportion of variance explained does not decrease significantly by adding an extra principal component, this can be viewed as a suitable cut-off point. This can be visualized using a scree plot, where the cut-off point is known as the elbow (Mangale, 2020).

The optimal number of factors are used as input variables for both the regression and the random forest. If one or more of these factors turn out to be of importance, their specific meaning can be read off from the factor labels. If a factor would have a positive and significant regression coefficient, it could be argued that the words that load high on these factors might have a positive impact on the rating of a wine.

In the next section I will go over the other dimensionality reducing technique, the Latent Dirichlet Allocation.

4.2 Latent Dirichlet Allocation

In order to find out what the experts are talking about in different classes of wine, a LDA can be performed. What LDA does is finding latent topics in a group of texts. In order to perform the LDA, this group of texts is transformed into a document-term-matrix. Within this document-term-matrix, the assumption is that there are latent distribution of documents over topics and topics over terms. A LDA checks for words that are often mentioned together in a document. These groups of words that are often mentioned in the same documents represent the latent topics. First, the functioning of the model will be explained. Second, the way of choosing the optimal model is discussed.

4.2.1 The Model

The model tries to calculate the probability that a particular document is generated. This model can be presented mathematically like this:

$$P(\underbrace{\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}}_{\text{Variables}}; \overbrace{\boldsymbol{\alpha}, \boldsymbol{\beta}}^{\text{HP}}) = \underbrace{\prod_{j=1}^{M} P(\theta_j; \alpha)}_{\text{part 1}} \underbrace{\prod_{i=1}^{K} P(\varphi_i; \beta)}_{\text{part 2}} \underbrace{\prod_{t=1}^{N} \overbrace{P(Z_{j,t}|\theta_j)}^{\text{a}} \overbrace{P(W_{j,t}|\varphi_{Z_{j,t}})}^{\text{b}}}_{\text{part 3}} \tag{2}$$

The model determines the probability that a particular document is returned based on several distributions. The multinomial distributions $\boldsymbol{W}$ and $\boldsymbol{Z}$, which follow from the Dirichlet distributions $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$.

The model can be fine-tuned by setting different values for α and β, which are known as the hyperparameters. Where α in this case refers to the concentration of the Dirichlet distribution, θ. Where β represents the topic/word density, where a low level implies that k number topics, consists of a small number of words, and a high level implies the opposite.

Part 1 of the equation displays the Dirichlet probability distribution of the documents over the topics. It generates a distribution of probabilities of a document belonging to a specific topic. For example, document 1 is 70% topic 1 and 30% topic 2. When aggregating all these Dirichlet distributions, a multinomial distribution $Z_{j,t}$ of all the topics can be generated (part 3a).

Part 2 of the equation displays the Dirichlet probability distribution of the topics over the terms. It generates a distribution of probability of a topic belonging to a specific term. For example, topic 1 is 60% word 1 and 40% word 2. Using this Dirichlet probability distribution and the multinomial distribution $Z_{j,t}$, the multinomial distribution of the word assignment $W_{j,t}$ can be generated (part 3b).

4.2.2 Perplexity

The goal of the model is to ascribe a high probability to a correct sentence and a low probability to a false sentence (Campagnola, 2020). In order to achieve this, the k number of topics and the level of α must be optimized. To optimize these these two inputs, the *perplexity* function from the *topicmodels* package in R is used. Perplexity is a method of evaluating language models, such as the LDA. The function operates using the following formula:

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}}$$

Where the perplexity (PP) of a full sentence (W) is the multiplicative inverse of the probability that the language model assigns to that specific sentence, which is then normalized by the N number of words in the validation set (Gandhi, 2018). Where $w_1 w_2 ... w_N$ are all words present in the validation set, the higher the probability of predicting a sentence from the validation set, the lower the level of perplexity. Therefore, the levels of k and α need to found for all three categories, such that the perplexity is minimized.

In the next section both the functioning and the assumptions of the Ordinary Least Squares will be examined.

4.3 Ordinary Least Squares

OLS is a statistical method that can be used to estimate a linear relationship between one dependent variable and one or more independent variables. In this section, I will explain how this model works, what assumptions must be satisfied and how to perform variable selection.

4.3.1 The model

The basic equation to be estimated with OLS with multiple variables is:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i \tag{3}$$

Where $i = 1, ..., n$, with n number of observations. y_i is the predicted value for a particular observation of the dependent variable, *points*. β_0 represents the intercept of the regression which is also known as α. Where β_{i1} is the estimated coefficient for the ith observations first variable, x_1. Furthermore, the β for all k number of variables is estimated at all n number of observations. The error term at each observation is represented by ϵ. The goal of the OLS regression is to minimize the sum of squared residuals, which is to minimize the following (Bremer, 2012):

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}) \tag{4}$$

Through the *lm* function in R, the levels of β that minimize this equation can be found.

4.3.2 Feature Selection

In the full linear regression, a total of 375 independent variables were selected. These variables, except for *logprice*, can be classified among the following categories:

Category	Unigrams	Bigrams	Countries	Years	Countries*Years	Factors
#	50	95	13	14	182	20

Table 7: Number of eligible variables per category

This is a great number of variables and it would seem to be highly unlikely that all of them are of significant importance in determining the dependent variable. In a paper on high-dimensional variable selection (Wasserman & Roeder, 2009), a multi-stage method is introduced. In this case, one part of the dataset is used to select the variables of greatest importance using LASSO regression. LASSO stands for Least Absolute Shrinkage and Selection Operator and uses a penalized regression to find the level of β that minimizes the following (Tibshirani, 2011):

$$\sum_{i=1}^{n} (y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5}$$

Where the left hand side is like the OLS regression earlier and the right hand side represents the penalty term. By adding the penalty term of a certain level of , which dictates the level of penalisation, the β for the least contributing variables go to zero. Through cross-validation, which is a way of re-sampling a specific data set into k number of cross-validated samples, two levels of λ are obtained. The level of λ that minimizes the root mean squared error

and the level of λ that is one standard deviation away from this point. One of these two levels can be used to select the optimal number of variables.

These selected variables are then used in a separate regression on the other part of the data. The results of this feature selection will be presented in the results section.

4.3.3 Assumptions

To correctly interpret the model and the limitations that is has, it is key to understand the model's underlying assumptions and to what extent they are satisfied (Tranmer & Elliot, 2008). Some basic assumptions about the model are quite easy to check. Is the dependent variable continuous? Are the explanatory variables either continuous or binary? The more complicated assumptions are those on the model errors.

The first of these assumptions of the model errors is : **Zero Conditional Mean of Errors**. If this assumption is satisfied, the mean of the errors at any value of the predictor variable is zero (Williams, Grajales, & Kurkiewicz, 2013). If this assumption is violated, it may be the case that the coefficients are biased. This violation could result from underlying non-linearity or measurement errors.

The second assumption of the model errors is: **Independence of Errors**. If this assumption is satisfied, the errors are uncorrelated with each other. If this assumption is violated, the errors are correlated with each other. One example of this is autocorrelation, in which case observations in a time series are correlated with lagged values (Williams et al., 2013).

The third assumption of the model errors is: **Homoscedacity of Errors**. If this assumption is satisfied, the errors have a constant and finite variance across all levels of the dependent variable. If the variance of the errors is not constant, the regression coefficients will be consistent and unbiased, but not efficient.

The fourth and final assumption of the model errors is: **Normal Distribution of Errors**. This assumption is particularly of importance in smaller samples and is required for the trustworthiness of the significance tests (Williams et al., 2013). If this assumption is not satisfied, the coefficients can still be efficient, unbiased and consistent.

The validity of these assumptions will be further discussed in the results section.

4.4 Random Forest

In the regression analysis as used in this thesis, linearity is a big assumption. It is quite reasonably possible that this assumption might not always hold in the analysis of wine reviews. Perhaps some types of flavour are valued in one point category, but not so much valued in another point-category. To work around this assumption and further supplement the analysis, a random forest can be used. In this section, the functioning of the random forest will be explained.

Firstly, an explanation of how a decision tree works and why it is flawed. Secondly, how a collection of decision trees can make a random forest and how this can overcome the limitations of a decision tree. Thirdly, the tuning of the hyperparameters needed to calibrate the random forest is discussed. Finally, a discussion on the interpretation of the results of the variable importance.

4.4.1 Decision Tree

Decision trees have been around for a long time, but it really gained traction in 1986, when J.R. Quinlan published *Induction of Decision Trees* (Quinlan, 1986). He invented the ID3 algorithm that allowed a computer to perform an induction task. The induction task, was trying to decide what class a particular object belonged to, on the basis of the object's attributes. The collection of objects is divided into a training set and a test set. The decision tree is grown on the training set, where objects are split over certain attributes on

specific nodes, until all objects end up in a leaf node. This can be graphically represented like the following:

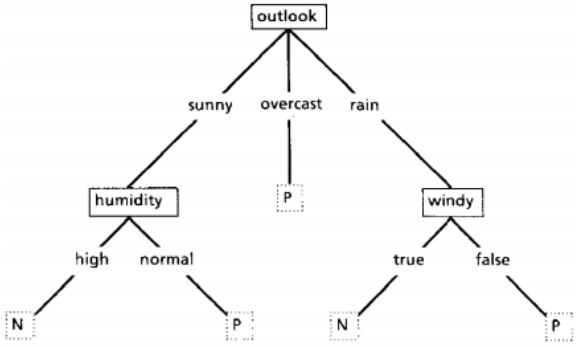

Figure 5: An example of a simple decision tree (Quinlan, 1986)

In this weather-based recommendation, at the root of the tree, we divide the objects based on their attribute values for the weather outlooks. The objects can then be further divided over their values for the other attributes until all objects are in leaf nodes (Quinlan, 1986). A tree can be constructed such that all objects are correctly classified. However, a highly complicated tree is quite unlikely to also classify all objects in a test set correctly. The essence of the task at hand is to induce what underlying principals are fundamental in the classification of an object into a class (Quinlan, 1986). The algorithm should therefore be written such that the most parsimonious decision tree is chosen.

Besides classification, decision trees can also be used in a regression setting, which is the setting of this research. The leaf nodes in this case, do not represent classes, but numerical values. An example of a regression tree is the following:

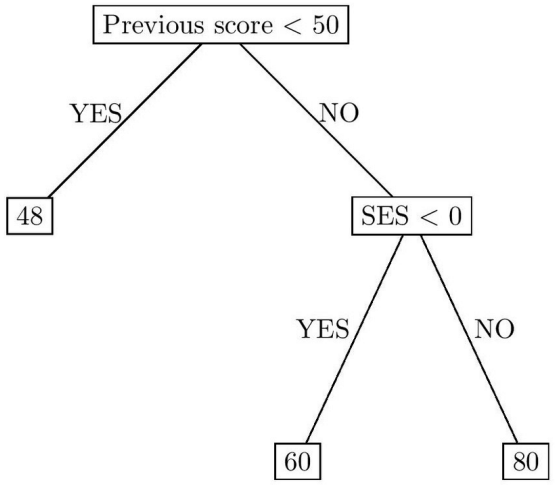

Figure 6: An example of a regression tree (Schiltz, Masci, Agasisti, & Horn, 2018)

In this example, we are interested in a particular score and this score is determined over several attributed. The way of evaluating this type of decision tree is through RMSE (Root Mean Squared Error). Where the test set values are compared to the test set predicted values. The RMSE formula is: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$

Where the actual value y_i, is subtracted from the predicted value $\hat{y}_i$ to form the error. This error is squared and all n errors are then summed and divided over the number of objects in order to get the mean squared error. The square root of this number is then taken to get a general sense of how far the predictions are off.

In practice, the RMSE of a regression tree is rather high, which is a result of overfitting.

Overfitting in this case means that the tree mostly resembles the structure present in the training set, but not so much the underlying structure of the fundamental nature of these objects. In order to get rid off overfitting, a random forest can be used.

4.4.2 Random Forest

After the introduction of the decision tree, many more accurate methods of induction were formulated, one of which was the random forest as proposed by Leo Breiman in 2001 (Breiman, 2001). Breiman defined a random forest as a collection of decision trees with a random selection of features at each node (Breiman, 2001).

These n number of decision trees are run on samples generated through bagging. Bagging is short for bootstrap aggregating and means generating bootstrap replicates from the training set (Breiman, 1996). What this entails is generating new training set by sampling from the original training set with replacement. So that it is possible that some of the objects in the training sample are not present in the bootstrap sample, or occur more than once in the bootstrap sample. On all the nodes on the n number of decision trees, features are randomly chosen to split the data on (Breiman, 2001).

After having grown all individual trees, a majority vote among these trees decides the class an individual object is in. Or in case of the regression trees, the average value predicted. As a result of both bagging, random feature selection and the Strong Law of Large Numbers, the predictions converge, thereby minimizing the generalization error (Breiman, 2001). As these predictions converge, they get rid of the overfitting problem that a single decision tree might have and get closer to the fundamental way in which an object is accurately classified.

In order to get the most accurate predictions, the random forest can be tuned, which will be explained in the following section.

4.4.3 Tuning the Hyperparameters

In order to improve the predictive power of the random forest, several hyper-parameters can be tuned. Unlike model parameters, which are obtained by the model's algorithm, for instance the slope and intercept of OLS, hyperparameters have to be set in advance. One way of doing this is through a grid search. In a grid search different levels of a hyper-parameters are plugged into the random forest, after which the level with the lowest RMSE can be picked. For instance, the number of trees to be used in the forest or the maximum number of variables to be considered when splitting on a node (Koehrsen, 2018).

4.4.4 Variable Importance

Besides the level of predictive performance of the random forest, it is quite interesting to know which variables are of greatest importance in driving the outcomes (Hoare, 2019). This importance can be obtained either through checking how much the predicitive accuracy decreases when a variable is left out. In this paper the *varImp* function of the *caret* package is used. This function accumulates the sum of the decrease in predictive power across all trees in the forest whenever a specific variable is used to split. From this function a list is generated with the relative importance of all variables in determining the prediction. The 30 most important variables will be plot, from which their relative importance can be deduced.

5 Results

In this section the results to the hypotheses will be discussed in the order of appearance in the theoretical framework. To make statements about these hypotheses, the results from OLS, random forest, PCA and LDA will be used. The results from the OLS are attached in

the appendix, from which significance and coefficients of variables will be obtained for the relevant hypotheses. The results from the LDA will give some insight in the latent topics present in the different categories of wine. The major result from the random forest is the variable importance. By considering the relative importance of the variables the findings from both the OLS and LDA can be further specified. Most of these variables are self-evident, but the factors not so much, whereas they do have a great variable importance in the random forest. In the first part of the results section, the specifications of the different models will be discussed. In the second part answers to the different hypotheses will be formulated.

5.1 Models

In this section, several characteristics and model descriptions will be given. Firstly, some comments about the factor labels of the PCA will be formulated. Secondly, the optimal models for the LDA will be discussed. Thirdly, the assumptions for the OLS regression will be checked. Finally, the predictive performance and the important variables of the random forest will be highlighted.

5.1.1 PCA

The first part to find out, was the optimal number of factors to be used. When plotting the proportion of variance explained over the number of principal components, the following was obtained:

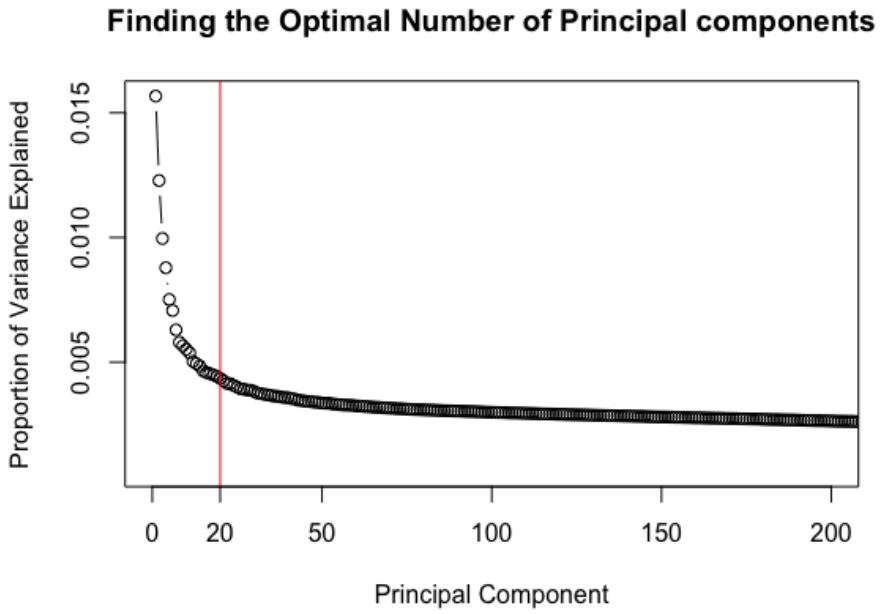

Figure 7: Finding the optimal number of factors to used in the PCA

After around 20 factors the proportion of variance explained by an extra factor does not drop a lot further. This point is known as the elbow point and is generally the best place to set the number of factors. The total proportion of variance explained is rather low for these first 20 principal component, namely 13.2%. Which might suggest that their predictive power is rather low.

For the 20 factors obtained, the terms that are most greatly associated with them are represented in the following table:

factor	$label_1$	$label_2$	$label_3$	$label_4$	$label_5$
1	palat	tannin	cherri	black	aroma
2	fruit	ag	tannin	sweet	rich
3	finish	aroma	berri	flavor	plum
4	fruit	plum	berri	aroma	sweet
5	acid	balanc	rich	palat	sweet
6	cherri	raspberri	appl	aroma	red
7	palat	nose	tannin	aroma	acid
8	finish	flavor	dry	berri	aroma
9	flavor	tannin	drink	dry	appl
10	drink	flavor	readi	fruiti	tannin
11	black	pepper	currant	cherri	tannin
12	ripe	dry	tannin	textur	plum
13	cabernet	sauvignon	blend	merlot	franc
14	spice	tannin	bake	flavor	ripe
15	oak	vanilla	toast	appl	tannin
16	note	aroma	tannin	rich	finish
17	fresh	cherri	palat	oak	light
18	rich	aroma	tannin	textur	acid
19	red	berri	fruit	currant	aroma
20	appl	acid	tannin	sweet	blackberri

Table 8: Factor Labels

There are many words that are greatly associated with multiple factors, *tannin* for instance, is named in 12 of the factors. The most unique of these 20 factors is factor 13, which appears to cover the types of grapes used in a wine. In order to let the PCA guide in deciding on the hypotheses, it is useful to classify these factors as belonging to certain aspects of the wine.

Where some factors, such as 2,5,12 and 16 predominantly cover wine characteristics. For example: sweet, balance, rich, ripe and finish. Whereas other factors are about specific aromas. Factors such as 6 and 19 cover the primary aromas, for example: cherry, raspberry, fruit and currant. Only factor 11 appears to concern the secondary aromas, as the two most associated words are black and pepper. The tertiary aromas are captured mostly by factors 15 and 17, who speak of oak and vanilla.

Knowing the difference between these groups of factors will help guiding the interpretation of the results of OLS and the Random Forest. In the next section, the model specifications of the LDA will be discussed.

5.1.2 LDA

After having created the document-term matrices for the good, very good and excellent categories, a model needs to be generated to find the latent topics in the reviews. In order to find the best model, the model will have to be estimated for different levels of k and different levels of α. These value that minimize the level of perplexity of the model on the validation set, is presented in the following table:

k	45	50	55	45	50	55	45	50	55
Category	Good			Very Good			Excellent		
$\alpha = 0.01$	458.4	454.6	457.4	456.9	456.3	456.3	542.5	548.0	542.5
$\alpha = 0.05$	433.9	**421.3**	421.9	391.4	**391.0**	391.4	519.0	**515.5**	518.8
$\alpha = 0.1$	438.7	437.9	441.3	444.2	440.4	442.4	522.1	521.5	523.0

Table 9: This table contains the level of perplexity of the validation sets using different numbers of α and k for the different categories

These perplexity levels show that for all three categories, the optimal level of α is 0.05, whereas the k optimal number of latent topics is equal to 50. The perplexity level of the middle category is the lowest, which might have to do with the fact that this category only contains wines with a score between 87 and 89. Quite to the contrary, the category with the highest rated wines also has the highest level of perplexity, but also the greatest point-range, namely 90-100.

5.1.3 OLS

In order to reduce the number of variables in the regression, feature selection is performed through LASSO regression. The results are presented in the following graph:

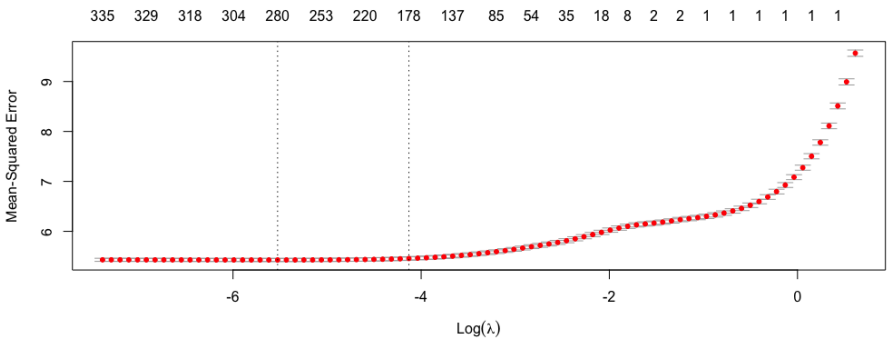

Figure 8: 10-Fold Cross-Validated Variable Selection using LASSO regression

The minimum level of Mean-Squared Error results into 281 non-zero coefficients. The point that is one standard error away from this point is at 179 variables. One can select either of these amounts, in this case the smallest number variables was chosen. When aggregating these variables over the earlier described categories, the following is obtained:

Category	Unigrams	Bigrams	Countries	Years	Countries*Years	Factors
Full	50	95	13	14	182	20
Restricted	3	68	6	11	94	3

Table 10: Number of eligible variables per category

In comparison to the full model with 50 unigrams present, only 18 unigrams make it to the final model. Of the bigrams a far greater number remain, 86 out of 95. For the countries and years, most of these remain significant enough to be included in the full model and 94 out of the original 128 interactions remain. An interesting observation is the fact that only a total of 3 out of 20 of the factors are selected for the final model. This observation was to be expected when considering the low level of variance explained by the selected factors.

In order to determine how valid the results from the final model are, the validity of the four assumptions on the model errors must be checked. One starting point is plotting the residuals over the predicted values:

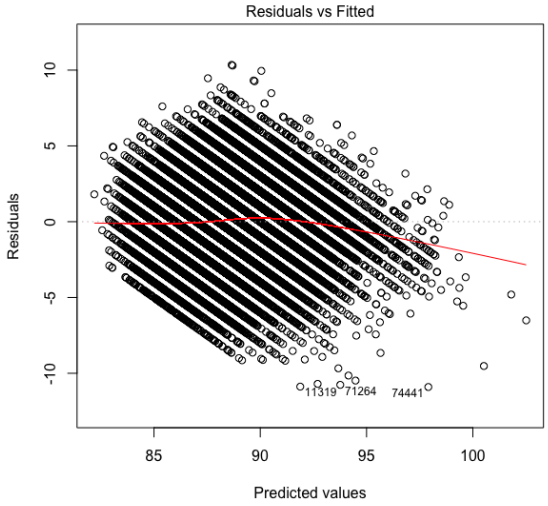

Figure 9: Plotting the residuals over the fitted values

The first assumption, **Zero Conditional Mean of Errors**, seems to hold until somewhere between 90 and 95 points. After that point, the residuals are more negative on average, meaning that higher values are predicted than the actual values. This means that there might be some upward bias in this higher point range. The second assumption, **Independence of Errors**, is rather likely to hold, some words might appear more together than others, however given the vast size of the words base it is rather likely to hold. The third assumption, **Homoscedacity of Errors**, does not seem to fully hold. This might have to do with the fact that the points awarded are not normally distributed over the interval from 80 to a 100. The fourth assumption, **Normal Distribution of Errors**, can be viewed with respect to the following plot:

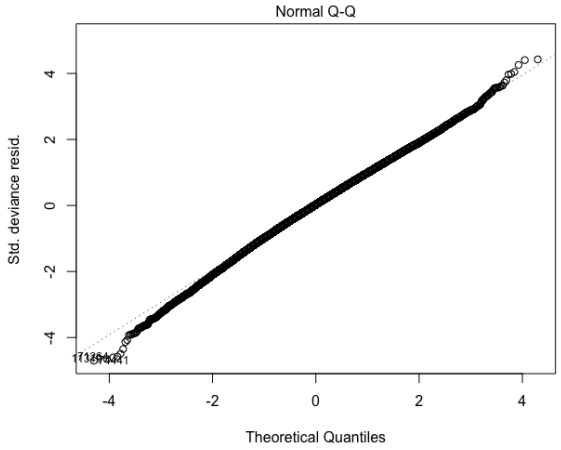

Figure 10: Visualising the distribution of the errors

The errors appear to be rather normally distributed, but at both edges of the outward quantiles they deviate from the diagonal. This implies that the predictions at the outer bounds must be analysed with great scrutiny. This is most likely caused by the fact that the points rated by the experts range from 80-100, whereas OLS is not limited to this range.

In the next section, the specifications of the random forest are discussed.

5.1.4 Random Forest

The goal of the random forest is to be able to predict what types of wine will score well. In order to optimize this prediction, several hyperparameters were tuned. For this research,

the *mtry* and *ntree* hyperparameters were tuned. The *mtry* hyperparameter determines the number of variables to be chosen from at each decision node. Whereas the *ntree* hyperparameter determines the number of trees to be used in the random forest. As the random forest used for this research is a collection of regression trees, the metric to evaluate the different models is the RMSE (Root Mean Squared Error). A grid search was performed over both hyperparameters in order to minimize the out of sample RMSE. A grid search comes to down to comparing the RMSE for different configurations of *ntree* and *mtry*. The following was obtained:

mtry	1	2	5	10	25	50
ntree = 2000	3.097	3.120	3.221	3.340	3.488	3.581
ntree = 2500	**3.096**	3.118	3.221	3.341	3.489	3.583
ntree = 3000	3.097	3.119	3.219	3.338	3.489	3.581

Table 11: Out of sample RMSE of different of *ntree* and *mtry*

This table shows that the optimal value for *mtry* is 1, since this minimizes the RMSE. The RMSE for the three levels of *ntree* is practically the same, with the model with an *ntree* of 2500 minimising it at 3.096. This implies that for every new review plugged into the model, the model on average is 3 points of. Considering that the point range is 80-100, this is quite a considerable error. Even so, this model might be rather helpful in placing new instances into the larger point-categories.

After having selected the model that minimizes the RMSE, it is interesting to know which variables are of greatest importance in the prediction. The importance of the 30 most important variables is shown in the following plot:

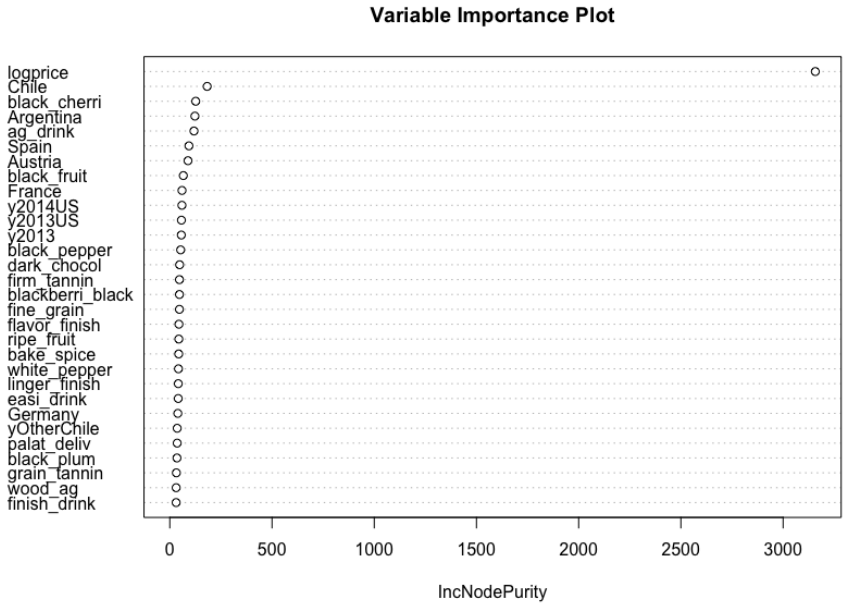

Figure 11: Variable Importance Plot

The first thing that comes to the eye is that the *logprice* variable is the most important, by far. The other 29 variables appear to agree with the variable selection as can be seen in table 10. 19 of the 30 most important variables are bigrams, their importance can be important in determining what types of characteristics and aromas are key in predicting wine quality. Moreover, several countries and years of wine that is produced in a particular country appear to be of importance. Similar to the variable selection of the LASSO regression, unigrams and factors appear to contribute relatively little to the overall prediction. The

importance of the 30 most important variables will be further investigated in the coming sections.

5.2 H_1

The first non-sensory characteristic of wine considered in the theoretical framework was the price. The hypothesis was stated:

H_1: Expensive wines are perceived better than cheap wines.

When plugging the *logprice* variable into the restricted model, the following was found:

Coefficient	**Estimate**	**Std. Error**	**Significance**
logprice	2.796	0.017	***

Table 12: Regression outcome for *logprice* (see Appendix for all results)
Note: *p<0.1; **p<0.05; ***p<0.01

This means that the null-hypothesis of the price of no effect can be rejected at a 0.01 significance level , meaning that the data favors a relationship between the price and the rating of the wine. Furthermore, the *logprice* variable is by far the most important variable in the random forest (see figure 11). Therefore H_1 appears to be very likely and price appears to be a good proxy of the underlying quality of a wine. What this result exactly means for a wine-maker is quite ambiguous. People might perceive a more expensive wine as being of greater quality, whereas quality is not changed. Furthermore, investing more in the production of a bottle can increase the wine's quality, but this does not have to be the case.

In the next section a closer look will be taken at the importance of the country of production on the appreciation of wine.

5.3 H_2

The second non-sensory characteristic of wine considered in the theoretical framework was the location. The hypothesis was stated:

H_2: The geographical location of production has an impact on the rating of a wine.

Through the OLS variable selection, Italy, USA and Other were dropped. Meaning that they most likely produce your typical wine considering the number of points. The coefficient estimates can be viewed relative to the wines from these three nation categories. The results of the other nations are displayed in the following table:

Coefficient	**Estimate**	**Std. Error**	**Significance**
Argentina	−0.550	0.073	***
Australia	0.619	0.101	***
Austria	1.442	0.176	***
Chile	−0.568	0.107	***
France	0.655	0.050	***
Germany	0.997	0.111	***
New_Zealand	0.193	0.136	
Portugal	1.011	0.079	***
South_Africa	0.592	0.113	***
Spain	−0.192	0.079	**

Table 13: Coefficients for the selected countries (See Appendix for all coefficients)
Note: *p<0.1; **p<0.05; ***p<0.01

From this table, can be seen that except for New Zealand, all coefficient are statistically different from the three dropped nation categories. As of the three negative coefficient estimates, Argentina, Chile and Spain, wines from these countries are typically rated worse than those of the USA and Italy. All other countries seem to produce higher rated wines than those from the USA and Italy. Especially wines from Austria, Germany and Portugal are highly appreciated by the experts. Rather surprisingly, wines from France score just barely higher than those from the USA and Italy, whereas France is often viewed as the home of great wines. A great part of the reason why France, Italy and USA all have average ratings that are rather similar and of around zero, is the fact that they are by far the most tested wines, wines from these three countries comprise of around 75% of the wines sampled. So that the sample mean by construct is close to the mean of the wines produced from these three countries.

Some of these countries also appear to be important predictors in the random forest (see figure 11). Chile, Argentina, Spain, Austria, France and Germany are included in the 30 most important predictors. These results agree with the results of the OLS regression and point at H_2 being true. What this implies for the wine-makers is less clear cut. It seems likely that wines from some countries are perceived better than for others. However, perhaps in these countries the conditions for production are better or wine-makers simply have a longer tradition of wine-making. Moreover, it might be more expensive to produce wine in the countries that are perceived as superior. It is thinkable that the same piece of land might cost more in for instance Germany, than it would in Argentina.

In the next section, the variability of the perceived wine quality will be examined in a national context.

5.4 H_3

The third non-sensory characteristic of wine considered in the theoretical framework was the year of production. The hypothesis was stated:

H_3: The local existence of a good year causes a higher appreciation for a wine of that year.

As formulated in the theoretical framework, the existence of a good or bad year is most likely related to changes in the climate over the years. As climate is largely local, the wines are split over the New World category and the Old World category. The results for all the interaction effects can be found in the Appendix, the specific results for the New World wines are given by the following table:

Country	USA	Argentina	Chile	Australia	New Zealand	South Africa
Selected	12	4	4	4	6	3
Significant	12	1	2	3	3	1

Table 14: New World Wines years that are statistically significant at 5%

The first thing that comes to the eye is the big difference between the USA and the rest of the New World countries. Through the variable selection, the USA as a wine-making country was not selected as an important variable. As these wines from the USA comprise a large share of the total data set, it was not so surprising that on aggregate, the average American wine is rather similar to the sample mean. What is surprising is that 12 out of the 14 American wine years were passed through the variable selection and that all 12 of them were also statistically significant at 5%. The importance that a particular year can have on the perceived quality of American wines is further substantiated by the results from the random forest (see figure 11). These results point out that if an American wine was from 2013 or 2014, this might tell us a lot about the perceived quality.

For the other New World countries, it is a rather different story. On average only 4 years per country were selected, of which only 2 turned out to be significant at 5%. Hinting at

the notion that there might some difference in climate variability between the hemispheres. To further investigate this difference, the following results were obtained for the Old World countries:

Country	France	Italy	Spain	Portugal	Austria	Germany
Selected	6	8	8	6	11	5
Significant	5	7	5	3	6	3

Table 15: Old World Wine years that are statistically significant at 5%

Of these Old World countries, an average of 7 years were selected for each country, of which around 5 also turned out to be statistically significant at 5%. There appears to be a greater variability in the quality over different years in this part of the world than in the countries on the southern hemisphere. This difference is important to consider for wine-makers since it can have a great impact on their revenues. Wine-makers in the Old World should feel a greater need to insure against the risk of having a bad year that wine-makers in the New World.

In the next couple of sections, the impact of specific characteristics and aromas on the perceived quality of wine will be examined.

5.5 H_4 & H_5

The fourth and fifth hypotheses mostly concern the difference in broad differences between wines of greatly different perceived qualities. These broad differences in wine characteristics and wine aromas will be primarily investigated through the LDA. The fourth hypothesis was stated like this:

H_4: Experts talk differently about wine characteristics and aromas in different point-categories.

The expectation is that experts talk differently about these two aspects in general and about different objects within these aspects between the different point-categories. The fifth hypothesis in a sense builds on the fourth hypothesis:

H_5: As wine quality increases, the discussion of wine characteristics will dominate those of specific aromas.

The fourth hypothesis would predict that the different point categories handle different aspects of both the characteristics and aromas. Whereas the fifth hypothesis predicts that as wines receive greater rating, the importance of the aromas present in wine decreases and the specific characteristics are more thoroughly discussed. Where the aromas would be represented by specific flavours and the characteristics would be telling of something slightly more abstract, such as the body.

For each of the three point-categories, the six most most telling latent topics, as generated by the LDA, will be discussed. An interpretation of these latent topics, combined with some results from the OLS regression and the random forest, will help guide the formulation of a statement on these hypotheses.

5.5.1 Good Wines

The first category consists of the wines scoring between 80 and 86 points on the expert's evaluation. This category will be referred to as the good wines category for this analysis. When performing the LDA on the good wines, an optimal number of 50 latent topics were found through the LDA. The 6 most prominent topics are: 4, 22, 24, 36, 41 and 42. These topics can be described by the following figure:

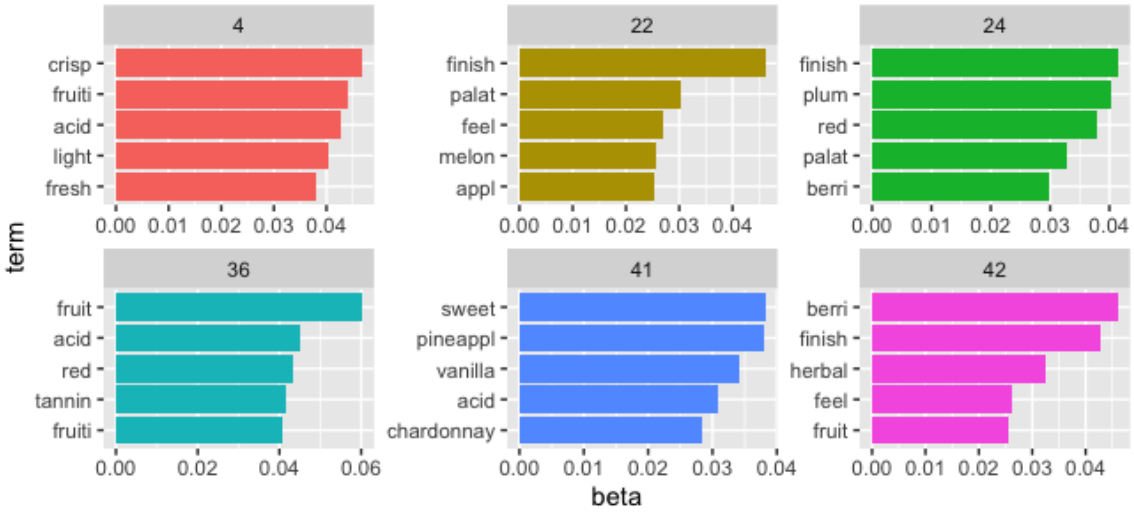

Figure 12: Top 6 Latent Topics for Good Wines

On the vertical axis the 5 most common words per topic are shown, so for example for topic 41: sweet, pineapple, vanilla, acid and chardonnay. Their relative importance within this topic is given by the beta parameter on the horizontal axis. By considering the six most frequent topics with their 5 most telling words, a general view on this first category can be formed.

Fruit appears to be a dominating theme in this first category, as every topic is in some way related to it either through the word fruit or the mentioning of a fruit. Secondly, acidity seems to play a role in the tasting of these wine, both through the word acid itself as through the word crisp which signifies a proper amount of acidity in the wine. Thirdly, the finish of the wine, sort of an overall impression after tasting the wine, seems to play a role in these good wines, as it is of great importance in topics 22, 24 and 42.

A total of 13 of 30 words clearly covers the wine aromas. Whereas 15 of 30 words clearly cover wine characteristics. In the next section, latent topics concerning wines with a slightly higher perceived quality will be examined.

5.5.2 Very Good Wines

The second category consists of only a relatively small range of points (87-89) as given by the experts. It is however a category in which many wines are situated, this category will be referred to as the very good wines. When performing the LDA on the very good wines category, an total of 50 latent topics are generated. The 6 most present topics are topics: 2, 8, 15, 20, 29 and 30. These topics can be described by the following figure:

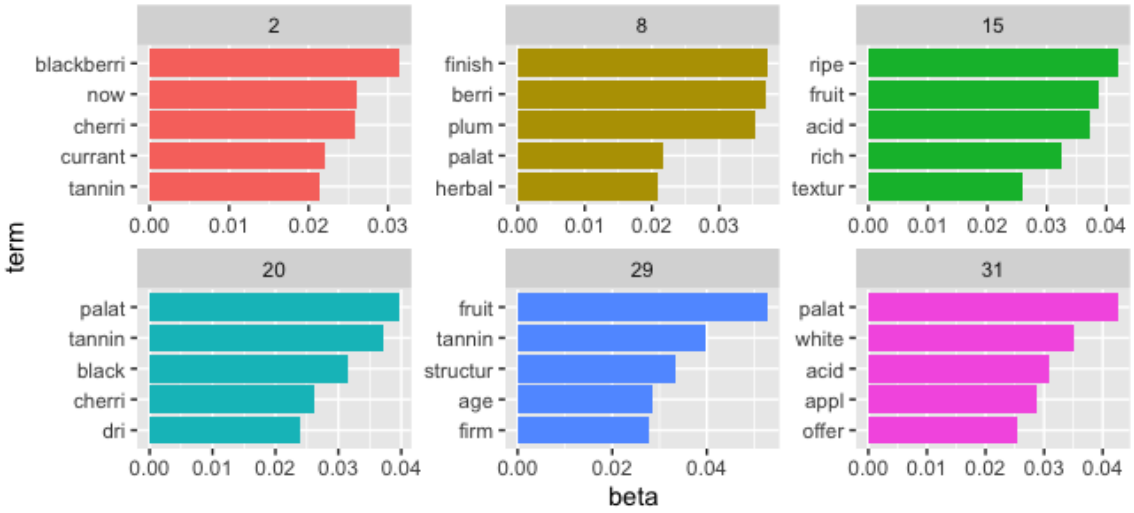

Figure 13: Top 6 Latent Topics for Very Good Wines

When comparing the second category with the first category, one notes that both fruity and acid aspects are less present in this category. Furthermore, there is greater mentioning of both the level of tannin in the wines and the wine palate. There appears to be a shift, with respect to category one, of specific flavour to overall sense of wine. This assumption is supported by the mentioning of ripe, rich, texture, dry, structure and firm. None of these 6 terms were mentioned in the first category, but are mentioned in the second category. This appears to be a sign of a shift in what experts value in different point categories. A shift from picking out specific, namely fruity flavours, to an experience in the mouth, a culmination of multiple sensations in the mouth.

A total of 12 out 30 words cover the wine aromas, whereas 15 words clearly cover wine characteristics. So a slight decrease of the number or words covering the wine aromas and a constant level of wine characteristics. In the next section, the latent topics of the wines that have the highest perceived quality will be examined.

5.5.3 Excellent Wines

The third and final category consists of the wines scoring 90 points and more, which are the excellent wines. When performing the LDA on the excellent wines, a total of 50 latent topics were generated. The 6 most frequent topics are: 1, 6, 9, 10, 20 and 27. These topics can be described by the following table:

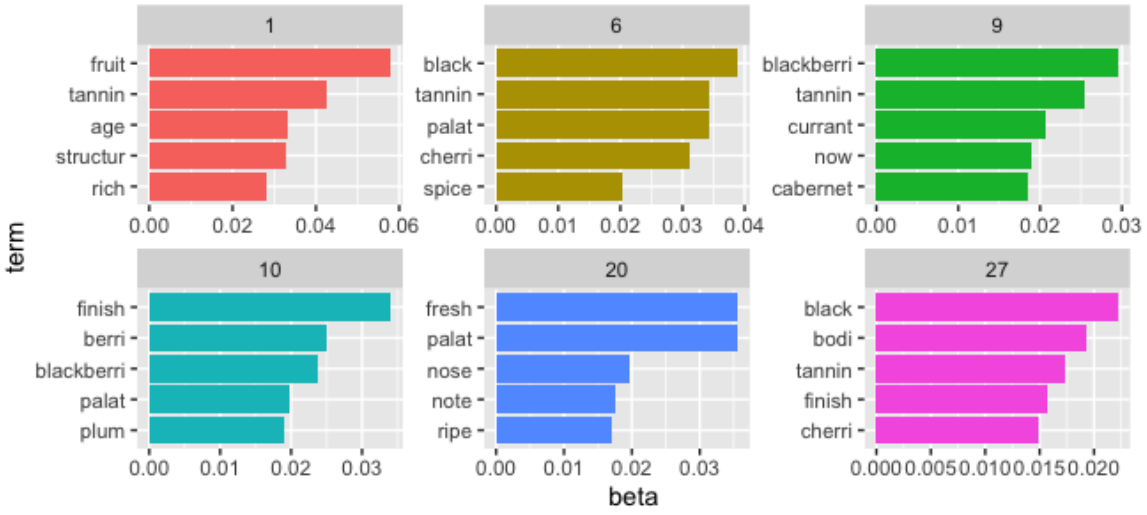

Figure 14: Top 6 Latent Topics for Excellent Wines

Like the second category, the level of tannin plays a big role in the excellent wines. The overall feel of the wine, as opposed to specific flavours, seems play even a bigger role in this class of excellent wines. This category discusses to a greater extent aspects such as ripening and aging of the wine, but also the structure and richness. In a sense this category is about completeness of the wine and to what extent different aromas come together to create a certain depth in the wine's flavour.

A total of 11 out of 30 words cover the wine aromas. Whereas 16 out of 30 words clearly cover wine characteristics. When looking at the presence of aroma and characteristic specific words over the different categories, we obtain the following:

Category	Good	Very Good	Excellent
Aroma	13	12	11
Characteristics	15	15	16

Table 16: Two Main Sensory Aspects of Wine

One does observe a downward trend of the mentioning of wine aromas as the number of points a wine has been rewarded increases. This trend however is so small that one can not be sure that this constitutes anything fundamental. The results of the mentioning of the wine characteristics are rather ambiguous as well. The excellent category mentions the most of them, but that is only one more than in the other two categories. When taking a a closer look at the type of wine characteristics mentioned, differences start to get slightly larger. Where acidity was a relevant characteristic of the first category, it is no longer present in the excellent category. Similarly, characteristics having to do with body where not present in the first category, whereas they are largely present in the third category.

When look at the variable importance plot (see figure 11), 19 of the 30 most important variables turned out to be bigrams. Where some of these bigrams represent aromas, such as black fruit and dark chocolate. Whereas others refer to specific characteristics, such as lingering finish and firm tannin. These results are further supported by the variable selection done through the LASSO regression, where 68 out of the 95 bigrams were selected, the vast majority also being statistically significant at 5% (see Appendix).

In the next section, the relative importance of the three different types of aromas and their sub-categories will be examined.

5.6 H_6

In the theoretical framework, the three different types of aroma in wine were discussed. All three aromas can, in general, be attributed to specific parts of the wine-making process. Primary aromas arising from the type of grapes used. Secondary aroma result primarily from the fermentation process. Tertiary aromas result from aging wine in oak barrels and even in the bottle. The following was argued:

H_6: The level of importance of these categories is in the same order as they are numbered.

From the full OLS regression 3 out of 50 unigrams and 68 out of 95 bigrams were selected. Of these selected n-grams, a total of 62 bigrams turned out to be statistically significant at 5%, of which a total of 35 pertained to one of the three aroma categories. Whereas the primary aromas follow from the types of grapes used, for example fruity and herbal flavour. Secondary aromas follow from the fermentation process and include flavours such as bread and butter. Tertiary aromas are associated with the aging of the wine, such as vanilla and tobacco.

Of these 35 bigrams, a total of 32 pertained to primary aromas. Therefore the dominance of the primary aromas seems to be rather evident. When taking a closer look at the other three bigrams, they seem to best fit the tertiary aromas that results from the aging and oxidation processes:

wood_ag	*dark_chocol*	*oak_flavor*

Table 17: Tertiary aromas significant at 5% (See Appendix for all results)

Of these three flavours, the dark chocolate flavour also has a relatively high importance in the random forest's prediction (see figure 11). There appears to be no evidence that the secondary aromas are more important than tertiary. However, the evidence is too thin to state that tertiary aromas are more important than secondary aromas. When tasting, particular flavours of these two categories might come to mind and some of these particular flavours might be better appreciated then others. What is clear though, is that most of the aromas perceived by the experts, belong to the primary category.

In the next section, the constituents of this primary category will be further investigated.

5.7 H_7

In the previous section the evidence stacked up to provide plentiful evidence that the primary aromas are of greatest importance of the three. In order to provide more specific information to the wine-maker it is interesting to know which of these primary aromas are the most important. The following was hypothesised:

H_7: Fruity aromas are the most important aromas of the primary aromas.

The primary aromas can be split it up into fruity, floral and vegetable aromas. Of the 32 primary aromas that were significant at 5%, a total of 27 aromas can be considered to be fruity flavours:

red_fruit	*red_cherri*	*black_cherri*	*black_fruit*
dark_fruit	*berri_fruit*	*red_berri*	*orang_peel*
wild_berri	*white_peach*	*black_plum*	*lemon_lime*
black_currant	*green_appl*	*stone_fruit*	*red_currant*
blackberri_black	*cherri_plum*	*ripe_fruit*	*tropic_fruit*
yellow_fruit	*berri_aroma*	*cherri_raspberri*	*appl_pear*
cherri_flavor	*flavor_blackberri*	*blackberri_cherri*	

Table 18: Fruity flavours at 5% significance

21 of these aromas bear a positive coefficient between 0.147 and 0.844 (see Appendix). Meaning that most of the fruity aromas are appreciated, but that there effect is not very large. 6 of the 30 most important variables of the random forest were bigrams related to fruity aromas (see figure 11). The third most important variable was the *black_cherri* bi-gram, which also bears quite a hefty positive coefficient of 0.536 (see Appendix). So, one could turn to wine-makers and advise them on what types of grapes could be used, as to create desirable fruity aromas. However, as most grapes are highly sensitive to the particular climate they grow in, such a grape advice must always be made in conjunction with the climate of a particular area.

6 Conclusion

At the beginning of this thesis, I set out to find an answer to the question:

What do wine experts value in wine?

The answer will be structured along the lines of the seven hypotheses. The first one being:

H_1: Expensive wines are perceived better than cheap wines.

Even though the wine experts did not know the price of the bottle they tasted, the results show that there is a clear relationship between perceived wine quality and price. This relationship between quality and price most likely works both ways. On the one hand, people would most likely be willing to pay more for a better wine. On the other hand, wine-makers who invest a great deal in their wines, would feel inclined to ask an appropriate compensation for it. A correct understanding of this mechanism could help the wine-maker in maximizing his profits.

Whereas most wine-makers are location-bound, there are also many new entrepreneurs which enter the business each year. For instance, someone who decides to retire and spend part of his retirement savings on a winery. However, what is the best place to start a winery? In order to establish what place might be best suited, it would be interesting to know, if some locations are better than others. The following was hypothesised:

H_2: The geographical location of production has an impact on the rating of a wine.

When looking at the results for the different countries of production, there appear to be fundamental differences between wines produced in different countries. However, this difference can not easily be ascribed to one specific cause. Surely, climate plays a role, even so, countries with similar climates also show differences in quality between them. Old World producers seem to have a slight edge over the New World producers, but it might the case that production in the Old World is also more expensive. Besides, some countries have a greater history of producing wine and simply have more knowledge on the production process. A future study on wine-maker know-how between different countries might offer prove for this.

Another big concept in wine-making is the existence of good and bad wine years. Which, as discussed, mostly depends on the climate in a particular year and country. With an ever-increasing speed of global warming, climates could rapidly change in specific parts of the world. In order to find out the impact of these climate changes to come, it is interesting to know what impact a particularly good or bad climate has on the wine produced in a country in a specific year. The following was hypothesised:

H_3: The local existence of a good year causes a higher appreciation for a wine of that year.

There appear to be differences in the sensitivities between nations with regards to the year of production. Thew wines from the Old World countries appear to be more sensitive to

changes in the climate than those from the New World. From a business perspective, it might be beneficial to insure the each harvest against a bad climate, when knowing that production takes place in a volatile area.

The past three hypotheses dealt mainly with relatively unchangeable factors for most producers. For the wine-makers, it is most interesting to know what they can actually change in their production process. However, it is not the production process that is evaluated, but the wine characteristics and wine aromas. By defining the ways in which these characteristics and aromas influence wine perception, wine-makers can back-solve how to alter their production process. On the ground-level, the following was hypothesised:

H_4: Experts talk differently about wine characteristics and aromas in different point-categories

There are clearly different points of discussion in the different categories of wine appreciation. The good wines are mostly evaluated based on their fruity flavours and their levels of acidity and freshness. Whereas the importance of these fruity flavours diminishes in the excellent wines category and the levels of structure and richness are of greater importance. This is partly in accordance with the following hypothesis:

H_5: As wine quality increases, the discussion of wine characteristics will dominate those of specific aromas.

The importance of aromas appears to slightly decrease with the increase in wine quality. However, the importance of wine characteristics in general does not appear to change as wine quality improves. Rather, experts discuss different wine characteristics in wines of different qualities. Where good wines might be named light-bodied and easy to drink, the excellent wines rather are properly aged and have more complexity.

Whereas altering the production process in order to get balanced levels of acidity might be hard. Identifying the three stages of production and their adhering specific aromas is not. However, when one would like to focus on one part of production in particular, which stage is the most useful? The following was hypothesised:

H_6: The level of importance of these categories is in the same order as they are numbered.

The results clearly indicate that the primary aromas are the most important. Hence it makes the most sense for a wine-maker to really consider the kind of grapes to be used for their wines. The secondary aromas, quite surprisingly, do not appear to be of greater importance than the tertiary aromas. Even though the evidence is rather thin, the tertiary aromas, and the aging process for that matter, might even be of greater importance than the secondary aromas and thus the fermentation process.

However, it is the primary aromas and thus the choice of grapes that is deemed most important. Making the following hypothesis all the more relevant:

H_7: Fruity aromas are the most important aromas of the primary aromas.

All models point to this hypothesis being true. The detection of many unique fruity flavours lead to a slight increase in appreciation of a wine. Using different types of grapes and perhaps even different combinations of grapes might lead to achieving these particularly desirable flavours.

Obviously it takes great skill to produce an excellent wine and this is not something that can be achieved in just a couple of years. What is less hard to achieve, is to make a rather simple wine, with a couple of fruity aromas in it that most experts like. For a wine-maker it is important to know his own skill and judge what quality of wine is attainable. Once knowing what level of appreciation is the benchmark, the entire production process can be modified in such a way that the wine will bear just the right characteristics and aromas that best suit a wine in that category. Future research could further explore the exact chemicals present in different qualities of wine, to make a more precise guide to creating a wine of a certain quality.

References

Adewumi, J. (2019, March). Understanding the role of eigenvectors and eigenvalues in pca dimensionality reduction. Retrieved from https://medium.com/@dareyadewumi650/understanding-the-role-of-eigenvectors-and-eigenvalues-in-pca-dimensionality-reduction-10186dad0c5c

Ashton, R. H. (2014). Wine as an experience good: Price versus enjoyment in blind tastings of expensive and inexpensive wines. *Journal of Wine Economics*, *9*(2), 171–182. doi: 10.1017/jwe.2014.7

Augustyn, A. (2019, Aug). *Mediterranean climate.* Encyclopædia Britannica, inc. Retrieved from https://www.britannica.com/science/Mediterranean-climate

Benfratello, L., Piacenza, M., & Sacchetto, S. (2009). Taste or reputation: what drives market prices in the wine industry? estimation of a hedonic model for italian premium wines. *Applied Economics*, *41*(17), 2197–2209.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Bremer, M. (2012). Multiple linear regression. *Published in the Journal "Math 261A-Spring*.

Brewczynski, S. (2018, May). Sweet wine types. Retrieved from https://www.cellarswineclub.com/Sweet-Wine-Types.aspx

Campagnola, C. (2020, May). Perplexity in language models. Retrieved from https://towardsdatascience.com/perplexity-in-language-models-87a196019a94

Cardebat, J.-M., & Figuet, J.-M. (2004). What explains bordeaux wine prices? *Applied Economics Letters*, *11*(5), 293–296.

Comité-Champagne. (2020). Every harvest is unique. *Comité Champagne*. Retrieved from https://www.champagne.fr/en/from-vine-to-wine/vine-husbandry/grape-harvests

Cozzolino, D., Cowey, G., Lattey, K., Godden, P., Cynkar, W., Dambergs, R., ... Gishen, M. (2008). Relationship between wine scores and visible–near-infrared spectra of australian red wines. *Analytical and bioanalytical chemistry*, *391*(3), 975–981.

Dalu, J. D., Baldi, M., Dalla Marta, A., Orlandini, S., Maracchi, G., Dalu, G., ... Mancini, M. (2013). Mediterranean climate patterns and wine quality in north and central italy. *International journal of biometeorology*, *57*(5), 729–742.

Dubey, A. (2018, Dec). The mathematics behind principal component analysis. *Medium*. Retrieved from https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643

Edison, T. (2018, Oct). *What year wine is best? an overview of vintage years.* Retrieved from https://www.wineturtle.com/year-wine-best-vintage/

Ford, G. T., Smith, D. B., & Swasy, J. L. (1988). An empirical test of the search, experience and credence attributes framework. *ACR North American Advances*.

Frank, R. (2018, Oct). *Bottle of wine sells for a record $558,000.* CNBC. Retrieved from https://www.cnbc.com/2018/10/15/bottle-of-wine-sells-for-record-breaking-558000.html

Gandhi, M. (2018, November).

Gregutt, P. (2020, Apr). *How to taste wine - wine tasting tips from wine enthusiast magazine.* Retrieved from https://www.winemag.com/2015/08/25/how-to-taste-wine/

Hoare, J. (2019, Sep). How is variable importance calculated for a random forest? *Displayr*. Retrieved from https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/

Jones, G. (2015, August). Climate, grapes, and wine. Retrieved from https://www.guildsomm.com/public_content/features/articles/b/gregory_jones/posts/climate-grapes-and-wine

Joshua. (2020, May). Viticulture in piedmont: optimism and adaptation in the face of a changing climate. Retrieved from https://wordonthegrapevine.co.uk/piedmont-viticulture-climate-change/#:~:text=Although\%20sharing\%20similar\%20latitude\%20with,climate\%2C\%20and\%20significantly\%20lower\%20rainfall.&text=The\%20often\%20warm\%20summer\%20is,variation\%20throughout\%20the\%20growing\%20season.

Koehrsen, W. (2018, January). Hyperparameter tuning the random forest in python. Retrieved from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Kosseva, M. R., Joshi, V., & Panesar, P. S. (2016). *Science and technology of fruit wine production*. Academic Press.

Landon, S., & Smith, C. E. (1997). The use of quality and reputation indicators by consumers: the case of bordeaux wine. *Journal of Consumer Policy*, *20*(3), 289–323.

Leve, J. (2019, Apr). Bordeaux vintage chart 1959 to today, vintage rankings characteristics. *The Wine Cellar Insider*. Retrieved from https://www.thewinecellarinsider.com/wine-topics/bordeaux-wine-buying-guide-tasting-notes-ratings/bordeaux-wine-vintage-chart/

Mangale, S. (2020, August). Scree plot. Retrieved from https://medium.com/@sanchitamangale12/scree-plot-733ed72c8608

McGovern, P., Jalabadze, M., Batiuk, S., Callahan, M. P., Smith, K. E., Hall, G. R., ... others (2017). Early neolithic wine of georgia in the south caucasus. *Proceedings of the National Academy of Sciences*, *114*(48), E10309–E10318.

McKirdy, T. (2019, June). The differences between primary, secondary, and tertiary aromas, explained. Retrieved from https://vinepair.com/articles/wine-aromas-explained/#:~:text=Tertiary\%20aromas\%20and\%20flavors\%20arise,of\%20primary\%20aromas\%20and\%20flavors.&text=Bottle\%20age\%20also\%20develops\%20interesting\%20new\%20notes\%20in\%20wines.

Mercer, C. (2020, June). Primary vs tertiary wine aromas: what's the difference? Retrieved from https://www.decanter.com/learn/understanding-wine-aromas-329940/#:~:text=Primary\%20aromas\%2C\%20such\%20as\%20fruit,aromas\%20develop\%20as\%20wine\%20ages.

Puckette, M. (2017, July). Why you might prefer low tannin red wines. Retrieved from https://winefolly.com/tips/low-tannin-red-wines/#:~:text=Since\%20tannin\%20is\%20considered\%20a,less\%20bitter\%20in\%20older\%20wines.

Puckette, M. (2020a, May). *Real differences: New world vs old world wine.* Retrieved from https://winefolly.com/deep-dive/new-world-vs-old-world-wine/

Puckette, M. (2020b). Terroir definition for wine. Retrieved from https://winefolly.com/tips/terroir-definition-for-wine/#:~:text=sounds\%20like\%20\%E2\%80\%9Ctare\%20WAHr\%E2\%80\%9D,more\%20'terroir'\%20than\%20others.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81–106.

Rapp, A., & Mandery, H. (1986). Wine aroma. *Experientia*, *42*(8), 873–884.

Raschka, S. (2015, January). Principal component analysis. Retrieved from https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html#:~:text=The\%20eigenvectors\%20and\%20eigenvalues\%20of,the\%20eigenvalues\%20determine\%20their\%20magnitude.

Robbilard, H. (2020, September). Dry white wine (types, prices, best wines to buy in 2020). Retrieved from https://www.vinovest.co/blog/dry-white-wine

Robinson, J., & Harding, J. (2015). *The oxford companion to wine*. American Chemical Society.

Saha, S. (2018, October). Baffled by covariance and correlation??? get the math and the application in analytics for both the terms.. Retrieved from https://towardsdatascience.com/let-us-understand-the-correlation

-matrix-and-covariance-matrix-d42e6b643c22

Schiltz, F., Masci, C., Agasisti, T., & Horn, D. (2018, 06). Using regression tree ensembles to model interaction effects: A graphical approach. *Applied Economics*, *50*. doi: 10.1080/00036846.2018.1489520

Swan, F. (2019, May). Mediterranean climate — why it's great for winegrowing. Retrieved from http://www.fredswan.wine/2019/05/20/mediterranean-climate-winegrowing/

Swiegers, J., & Pretorius, I. (2005). Yeast and bacterial modulation of wine aroma and flavour. *Australian Journal of grape and wine research*, *11*(2), 139–173.

Thorpe, M. (2009). The globalisation of the wine industry: new world, old world and china. *China Agricultural Economic Review*.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282.

Tilden, M. (2020). How to understand the primary, secondary and tertiary aromas in wine. Retrieved from https://www.winemag.com/2020/07/27/primary-wine-aromas-guide/

Tranmer, M., & Elliot, M. (2008). Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, *5*, 30–35.

Tuttle, B. (2011, Nov). *Wines from 'bad years' are often great values.* Time. Retrieved from https://business.time.com/2011/11/04/wines-from-bad-years-are-often-great-values/

Villamor, R. R., & Ross, C. F. (2013). Wine matrix compounds affect perception of wine aromas. *Annual review of food science and technology*, *4*, 1–20.

Wasserman, L., & Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, *37*(5A), 2178.

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation*, *18*(1), 11.

Wine enthusiast magazine. (2020). Retrieved from https://www.winemag.com/about-us/

Winemag.com. (2020). Submit for rating. Retrieved from https://www.winemag.com/submit-for-rating/

WineScribble.com. (2020a). Secondary aromas in wine: An overview. Retrieved from https://winescribble.com/secondary-aromas-in-wine-overview/#:~:text=Secondary\%20aromas\%20in\%20wine\%20refer,are\%20further\%20developed\%20and\%20transformed.

WineScribble.com. (2020b). Tertiary aromas in wine: A quick overview. Retrieved from https://winescribble.com/tertiary-aromas-in-wine/#:~:text=Summary-,Tertiary\%20aromas\%20in\%20wine\%20refer\%20to\%20the\%20bouquet\%20of\%20smells,to\%20oxygen\%2C\%20oak\%20and\%20lees.

7 Appendix

Table 19: Regression Results from the Restricted Model

	Dependent variable:
	points
logprice	2.796*** (0.017)
y2012	0.222*** (0.050)
y2013	0.152** (0.068)
y2014	0.070 (0.055)
y2015	0.349*** (0.071)
y2016	0.065 (0.094)
yOther	0.016 (0.066)
Argentina	−0.550*** (0.073)
Australia	0.619*** (0.101)
Austria	1.442*** (0.176)
Chile	−0.386*** (0.075)
France	0.655*** (0.050)
Germany	0.997*** (0.111)
New_Zealand	0.193 (0.136)
Portugal	1.011*** (0.079)
South_Africa	0.295*** (0.108)
Spain	−0.192** (0.079)
y2005Argentina	0.512 (0.385)
y2008Argentina	−0.510** (0.219)
y2015Argentina	−0.200 (0.258)
yOtherArgentina	−0.777 (0.631)
y2012Australia	0.202 (0.326)
y2013Australia	0.856*** (0.262)
y2014Australia	0.639** (0.263)
yOtherAustralia	−0.952*** (0.237)
y2004Austria	−1.356* (0.798)
y2005Austria	−1.303* (0.671)
y2006Austria	−0.342 (0.351)
y2007Austria	−0.775** (0.304)
y2011Austria	−0.277 (0.252)
y2012Austria	0.325 (0.242)
y2013Austria	0.976*** (0.237)
y2014Austria	1.259*** (0.244)
y2015Austria	1.416*** (0.239)
y2016Austria	1.466*** (0.304)
yOtherAustria	−1.538** (0.612)
y2005Chile	1.035** (0.426)
y2010Chile	0.118 (0.164)
y2012Chile	0.016 (0.170)
y2015Chile	−0.545** (0.221)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 20: Regression Results from the Restricted Model

y2005France	0.142 (0.164)
y2006France	0.579*** (0.134)
y2008France	0.431*** (0.120)
y2009France	0.348*** (0.100)
y2010France	0.311*** (0.088)
y2013France	−0.298*** (0.105)
y2005Germany	−0.389 (0.562)
y2006Germany	−0.125 (0.329)
y2009Germany	−1.004*** (0.348)
y2014Germany	0.986*** (0.214)
y2015Germany	0.923*** (0.213)
y2004Italy	0.445*** (0.147)
y2006Italy	0.322*** (0.110)
y2007Italy	0.276*** (0.094)
y2009Italy	−0.560*** (0.098)
y2010Italy	−0.255*** (0.089)
y2011Italy	−0.276*** (0.086)
y2012Italy	−0.358*** (0.086)
y2016Italy	0.224 (0.226)
y2004New_Zealand	−2.127 (1.659)
y2006New_Zealand	−0.600 (0.539)
y2013New_Zealand	1.051*** (0.269)
y2014New_Zealand	0.654** (0.309)
y2015New_Zealand	0.394 (0.340)
y2016New_Zealand	1.072** (0.470)
y2004Other	−1.873*** (0.553)
y2005Other	−1.866*** (0.437)
y2006Other	−0.729** (0.341)
y2007Other	−1.668*** (0.296)
y2008Other	−0.567* (0.296)
y2009Other	−0.108 (0.233)
y2012Other	0.521*** (0.199)
y2013Other	0.729*** (0.209)
y2014Other	1.163*** (0.218)
y2015Other	0.836*** (0.256)
yOtherOther	−1.867*** (0.438)
y2007Portugal	0.717** (0.279)
y2008Portugal	0.005 (0.223)
y2011Portugal	0.300* (0.162)
y2012Portugal	−0.022 (0.157)
y2015Portugal	−0.457*** (0.164)
y2016Portugal	−0.473** (0.215)
y2014South_Africa	0.658* (0.378)
y2015South_Africa	0.427 (0.333)
y2016South_Africa	2.050*** (0.496)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 21: Regression Results from the Restricted Model

y2005Spain	0.212 (0.224)
y2007Spain	−0.085 (0.190)
y2008Spain	−0.541*** (0.181)
y2010Spain	0.269* (0.159)
y2011Spain	0.482*** (0.151)
y2013Spain	−0.450*** (0.170)
y2015Spain	−0.573*** (0.194)
yOtherSpain	−0.518** (0.225)
y2004US	−1.227*** (0.153)
y2005US	−1.027*** (0.095)
y2006US	−0.690*** (0.081)
y2007US	−0.535*** (0.077)
y2008US	−0.421*** (0.077)
y2009US	−0.288*** (0.069)
y2010US	−0.392*** (0.066)
y2011US	−0.312*** (0.068)
y2013US	0.439*** (0.077)
y2014US	0.784*** (0.070)
y2015US	0.913*** (0.093)
y2016US	1.116*** (0.144)
red_fruit	−0.216*** (0.052)
red_cherri	0.147** (0.066)
black_cherri	0.536*** (0.042)
pinot_noir	0.323*** (0.065)
black_fruit	0.671*** (0.048)
sauvignon_blanc	0.181** (0.081)
dark_fruit	0.488*** (0.084)
dry_tannin	−0.433*** (0.098)
berri_fruit	0.216*** (0.058)
berri_flavor	−0.086 (0.073)
red_berri	−0.273*** (0.063)
orang_peel	0.340*** (0.109)
white_pepper	0.975*** (0.072)
wild_berri	0.756*** (0.094)
wood_ag	0.369*** (0.083)
white_peach	0.844*** (0.089)
black_plum	0.482*** (0.073)
black_pepper	0.434*** (0.066)
lemon_lime	0.501*** (0.094)
black_currant	0.247*** (0.060)
green_appl	0.429*** (0.074)
cabernet_sauvignon	0.180*** (0.056)
stone_fruit	0.488*** (0.071)
aroma_flavor	−0.332*** (0.077)
red_currant	−0.249*** (0.079)
ripe_black	0.136 (0.091)
blackberri_black	0.641*** (0.103)
dark_chocol	0.807*** (0.083)
light_bodi	−0.603*** (0.102)
fruit_spice	0.615*** (0.104)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 22: Regression Results from the Restricted Model

easi_drink	-0.555^{***} (0.083)
cherri_plum	0.327^{***} (0.097)
ripe_fruit	0.518^{***} (0.070)
tropic_fruit	0.440^{***} (0.072)
yellow_fruit	0.471^{***} (0.106)
firm_tannin	0.529^{***} (0.068)
berri_aroma	0.406^{***} (0.092)
readi_drink	-0.532^{***} (0.059)
cherri_raspberri	0.321^{***} (0.085)
flavor_black	0.345^{***} (0.105)
tannin_acid	0.309^{***} (0.099)
palat_offer	-0.378^{***} (0.060)
cherri_blackberri	0.012 (0.101)
bake_spice	0.755^{***} (0.073)
appl_pear	0.249^{***} (0.089)
oak_flavor	-0.235^{**} (0.095)
cherri_flavor	-0.457^{***} (0.068)
medium_bodi	-0.332^{***} (0.058)
fruit_acid	0.063 (0.088)
finish_dry	0.579^{***} (0.105)
balanc_acid	0.637^{***} (0.095)
finish_drink	0.585^{***} (0.069)
aroma_lead	-0.196^{**} (0.099)
fresh_acid	0.514^{***} (0.077)
palat_deliv	0.322^{***} (0.082)
crisp_acid	0.316^{***} (0.073)
aroma_palat	-0.088 (0.095)
blend_cabernet	0.097 (0.103)
linger_finish	1.073^{***} (0.083)
ag_drink	0.970^{***} (0.077)
bright_acid	0.537^{***} (0.077)
flavor_blackberri	0.446^{***} (0.098)
tannin_drink	0.045 (0.088)
fine_grain	0.925^{***} (0.104)
blackberri_cherri	-0.227^{**} (0.095)
soft_tannin	-0.296^{***} (0.091)
lead_nose	-0.288^{***} (0.097)
white_flower	0.285^{***} (0.099)
acid_balanc	0.386^{***} (0.105)
note	-0.006 (0.023)
white	0.010 (0.030)
vanilla	-0.036 (0.034)
factor1	-0.013 (0.010)
factor5	-0.006 (0.010)
factor17	0.007 (0.010)
Constant	78.631^{***} (0.074)
Observations	57,858
R^2	0.431
Adjusted R^2	0.429
Residual Std. Error	2.338 (df = 57677)
F Statistic	242.256^{***} (df = 180; 57677)
Note:	*p<0.1; **p<0.05; ***p<0.01