Wine Data Analysis

Weronika Tomanek, Mateusz Geca, Jedrzej Sarna
AGH WMS

November 2022

1 Introduction

Wine is one of the top most popular beverages in the world. With the big variety of the product everyone makes a decision about choosing a certain one based on some features that are important for them. In our project we focus on analysing these features, checking the connections between them and their influence on the final rating. In the end we will try to find the best candidates for different tastes. The results of our work (including an app) can be found on GitHub under this link.

2 Preparing the dataset

For our analysis we chose the Vivino.com data from year 2020 available on the website Kaggle.com. As the world's largest online wine marketplace and most downloaded wine app, the Vivino community is made up of millions of wine drinkers from around the world. Each user can rate the wine according to their taste and find on the website all the details about each wine.

The dataset we have chosen contains variables such as name of the wine, origin country, origin region or province, origin winery, average rating, number of ratings, price in euro and year of production. We downloaded the csv files for every style of the wine and loaded them into the python environment creating data frames using pandas package. At the modeling step, we decided to work with data as a single data frame, so we joined our data creating a feature for wine styles. Wines that have been made after 2018 year were labeled as "Non vintage" without a precise year, so we decided to simply substitute for it the integer 2022. Another significant change was transformating the price of the wine to the difference between the price and minimum price in the data set.

For the most clear results we decided to use python environment, R and Statistica for different parts of our analysis.

3 Basic data analysis

At this point we decided to get a better understanding of our data and find interesting trends we wanted to explore further. We created our first plots and ran some basic tests. Moreover, we created several different heat maps representing the spread of wines over the world based on the region of origin and specific conditons (available on GitHub).

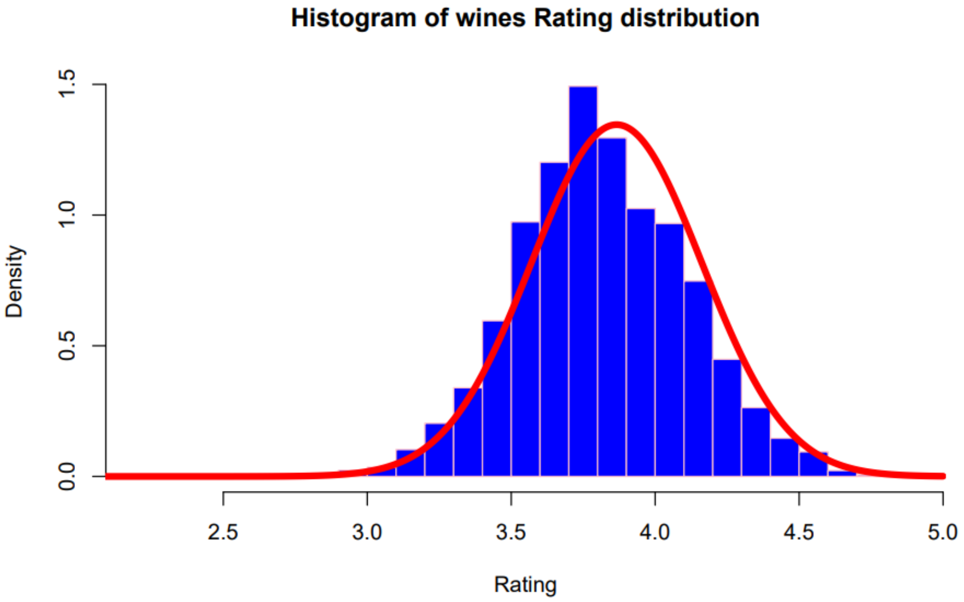


Figure 1: Histogram of wines rating distribution

From the histogram we see that the rating is likely to have a normal distribution. After running a shapiro test on the data the outcome seems to be rather surprising for us. The p-value is almost equal to zero. Although the bell-shaped curve is indicating the gaussian distribution, the essential conditions are not satisfied.

Let us extend our analysis of rating by including other variables. To get more primary information about our data and we created the following box plots.

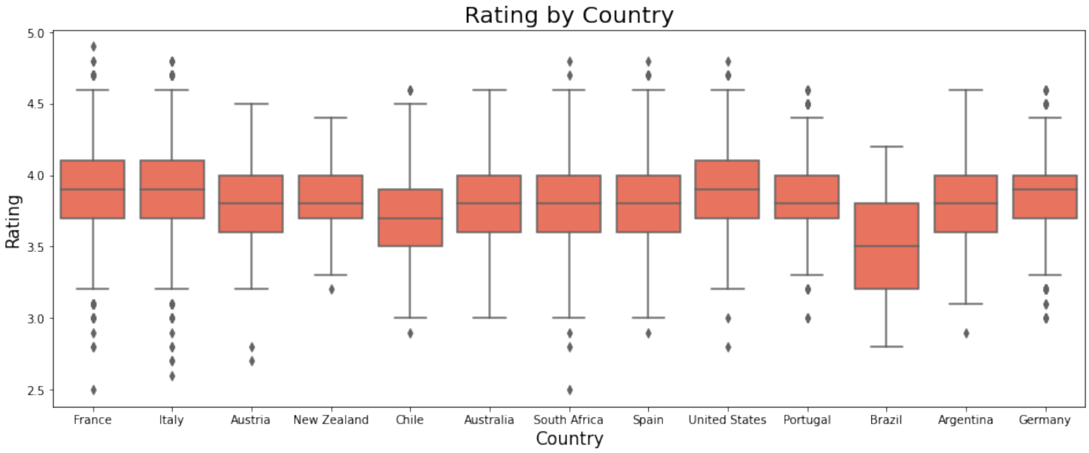


Figure 2: Rating by Country

We decided to consider only the countries that produced 50 wines or more to make the plot interpretable. The rating of wines from France and Italy fluctuates the most, probably as the majority of wines in our dataset comes from these countries. Brazilian wines do not have a good reputation on average and New Zealand does not have wines rated poorly.

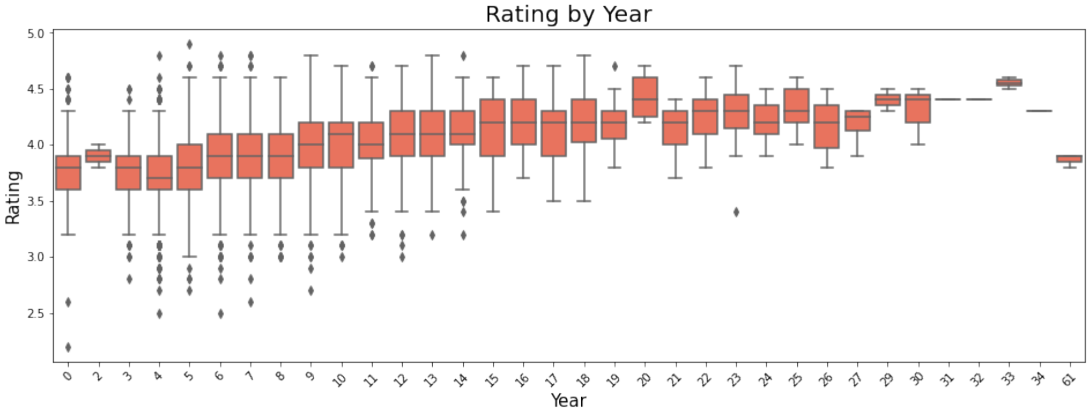


Figure 3: Rating by Year

The trend is quite clear. To some point rating of the wine increases with age, but with the oldest wines the average rating drastically drops. We face outcome like that due to the small quantity of wines in this age group. Important observation is that among younger wines the spread of rating is big and it becomes smaller with increasing age.

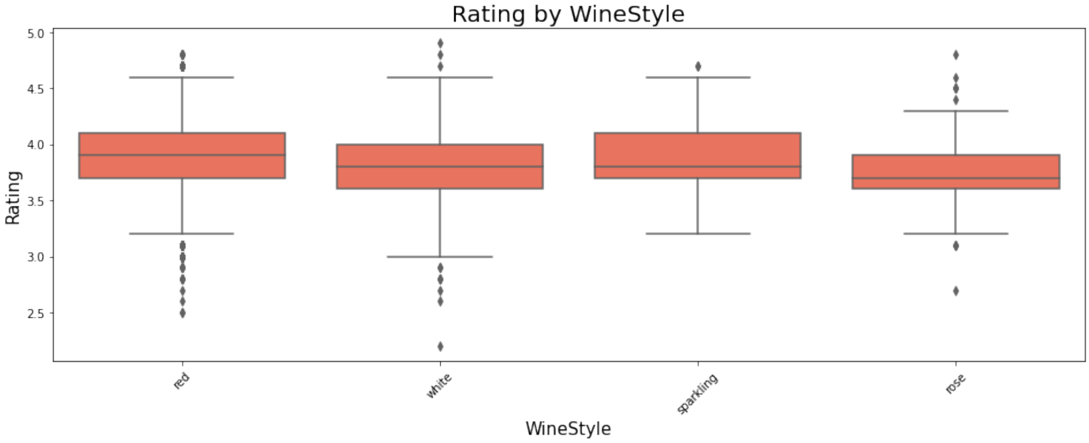


Figure 4: Rating by WineStyle

The average rating for each of the styles is rather similar. Sparkling wines do not have neither good nor bad outlying observations. With red and white wines it can be said that there is a bit more uncertainty of the product quality as the rating reaches extreme values.
The next test we decided to run was checking the correlation between variables.

From the plot we can see that there exists a significant correlation between the rating and the price. We decided to explore the relationship between these variables as one of the first problems of our project.

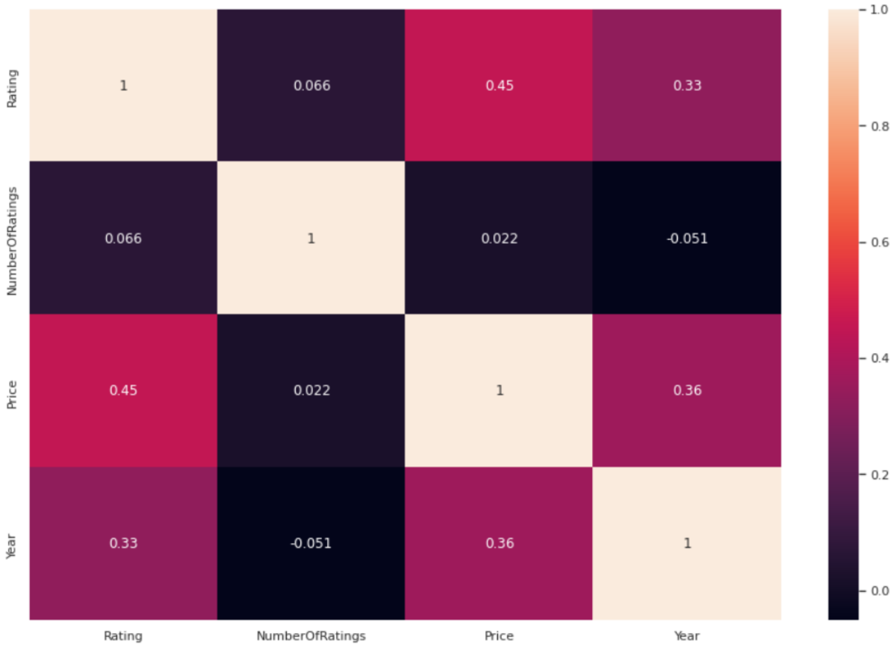


Figure 5: Heatmap for numerical columns of the data set

3.1 Price analysis

First, let us check what the price distribution looks like. The following plots do not include prices shifted by the minimum to represent the true conditions.

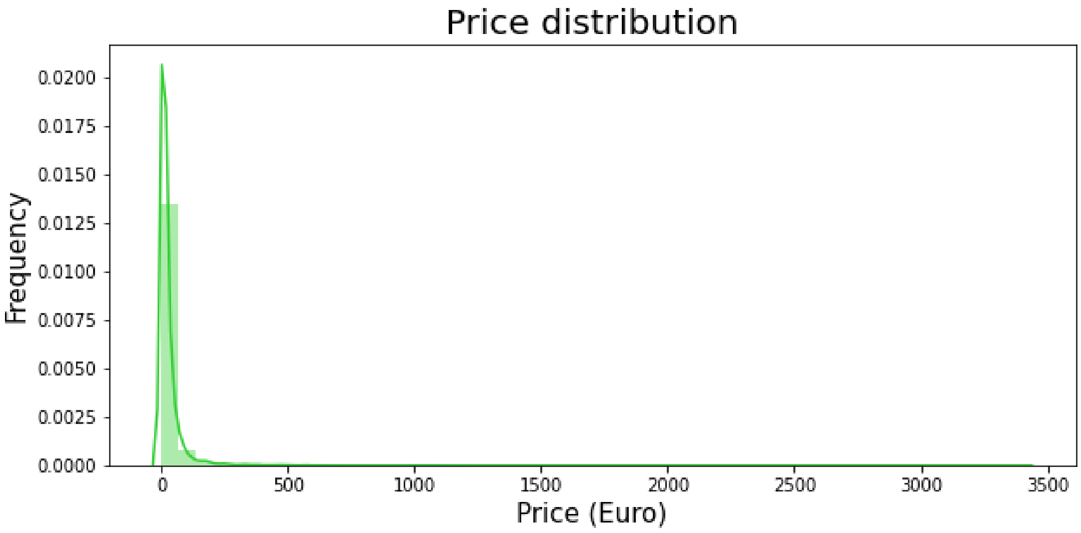


Figure 6: Price Distribution

We can not tell much from this chart due to the very high price of one of the wines (outlier). The wine named *'Pomerol 2012'* costs 3410.79 Euros, almost 2000 Euros more than the second most expensive wine in the dataset. To make the plot more readable and make it look more similar to the normal distribution, we use the natural log transformation.

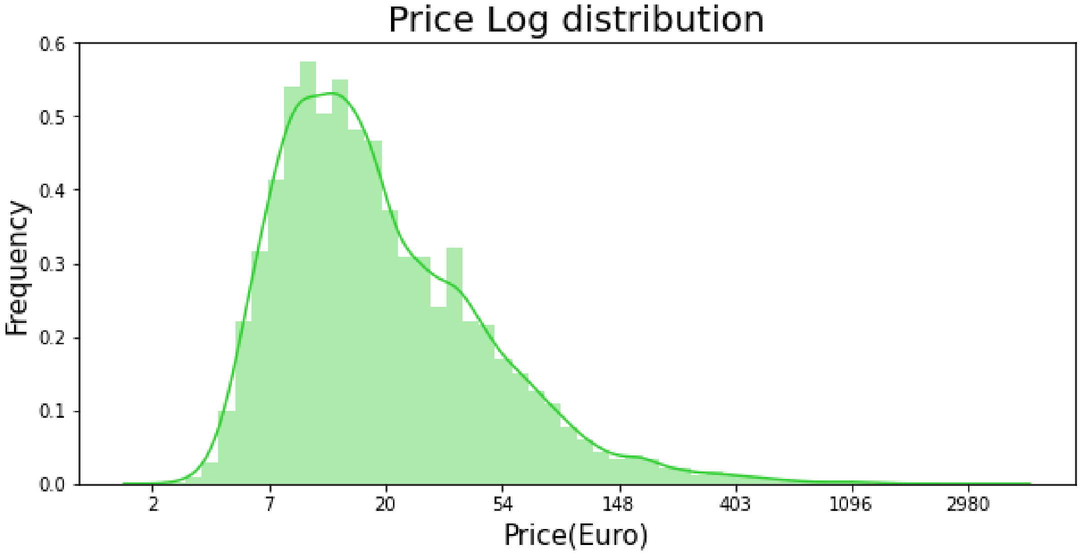


Figure 7: Price Log Distribution

Now we are able to draw our first conclusions. We can see that our dataset is dominated by cheaper wines. This observation will be important in our further considerations. Let us move on with our analysis by exploring how the price behaves after taking into account other variables.

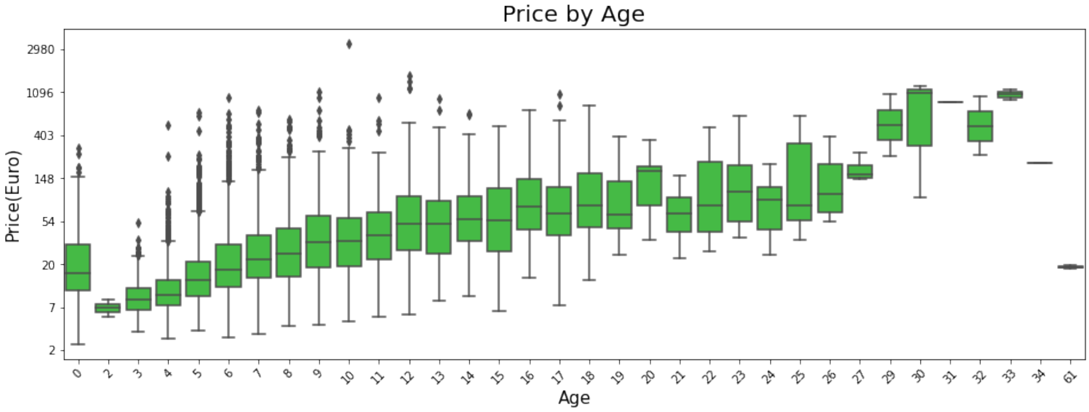


Figure 8: Price by Age

Here we can clearly see a dependence. Confirming our beliefs, graph shows that the older the wine the more expensive it tends to be. However, there are some wines that deviate from the existing trend. We will focus on these outliers later in the project.

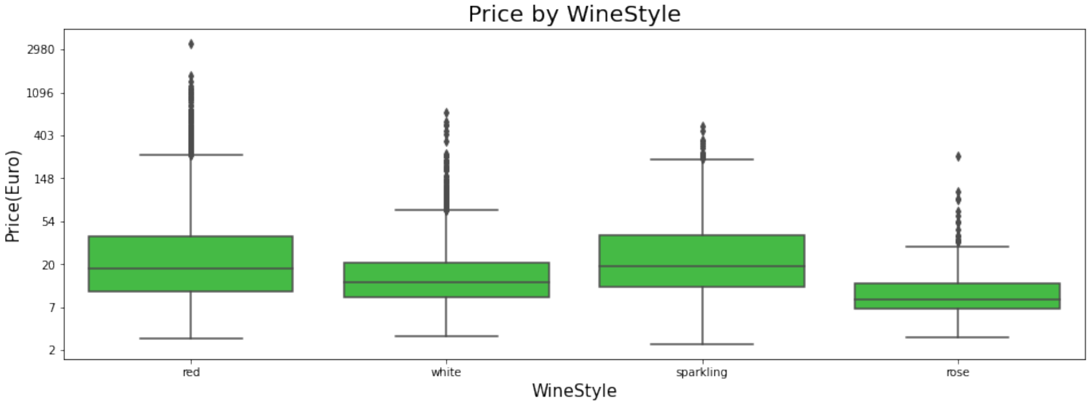


Figure 9: Price by WineStyle

From this histogram we can also draw some important conclusions. In general rose wines are the cheapest, the average price for white wine is relatively low and red and sparkling wines seem to have similar averages. It is worth noticing that the most expensive wines are red. In fact red wines have the most flavor elements and people are willing to pay a lot for a very good red wine.

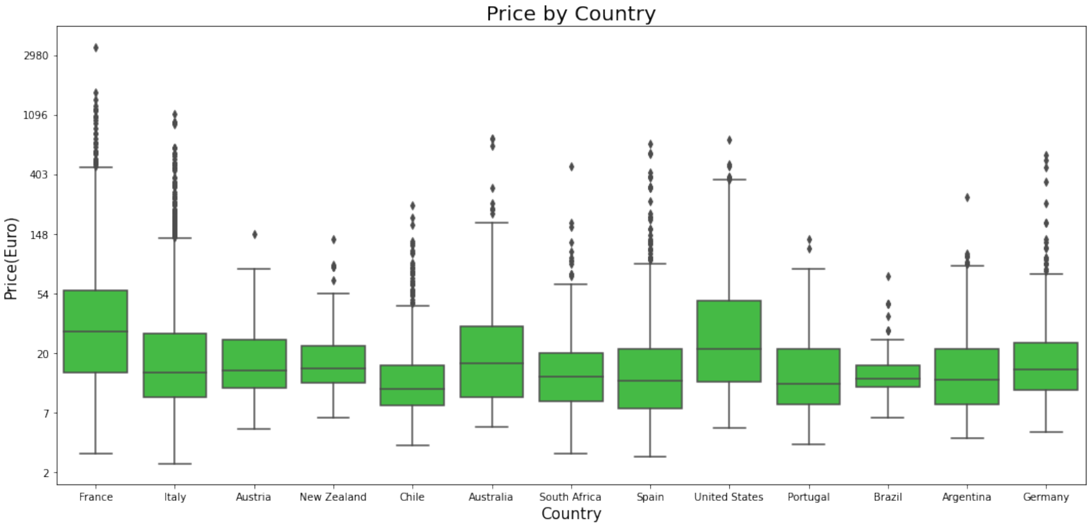


Figure 10: Price by Country

For this plot once again we decided to take into consideration only countries with a relatively large amount of exported wines. We can see that France and the United States have the most expensive wines on average when Brazil has the cheapest ones. In addition, the most expensive wines in our entire dataset are from France.

Now, after more detailed analysis of the Price variable, we can proceed to explore the relationship between Price and Rating. First, let us apply a simple linear regression model.

X-Price
Y-Rating

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, ..., n$$

This is the form of our model. There are n equations that can be stacked together and written in matrix notation as

$$Y = X\beta + \epsilon,$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

For us $n = 13834$ which is the number of different wines in the dataset. Now we compute the vector of estimators for our data using the formula introduced during the class.

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Given all the data the results are as follows:

$$\hat{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 3.80937565 \\ 0.00188415 \end{pmatrix}$$

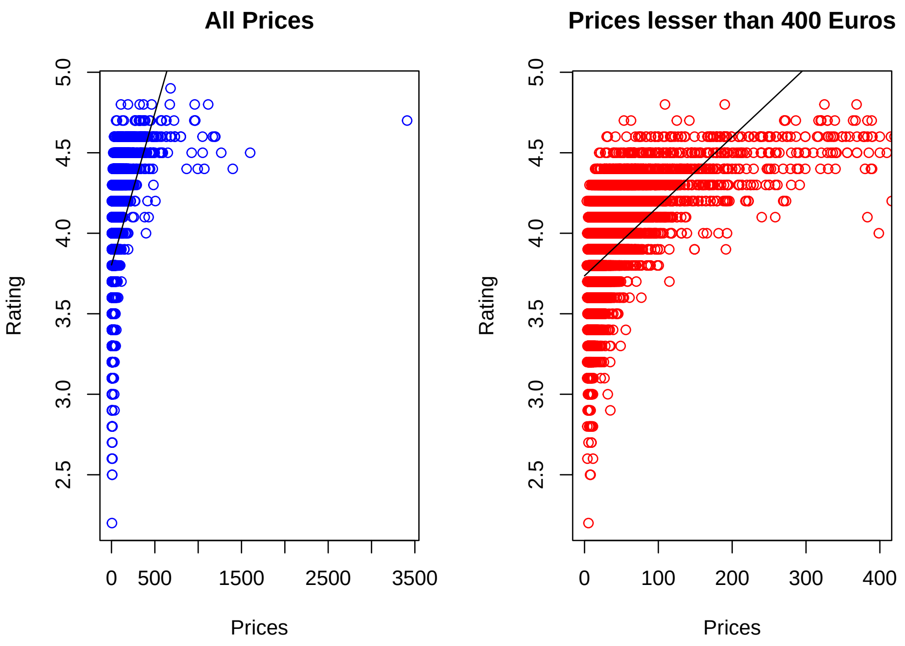


Figure 11: Simple Linear Regression

After plotting our linear regression function we can already see some outliers in the first graph. Since it makes it difficult to determine whether the line fits the data well, we cap the price to 400 Euros in second graph to see the pattern better.

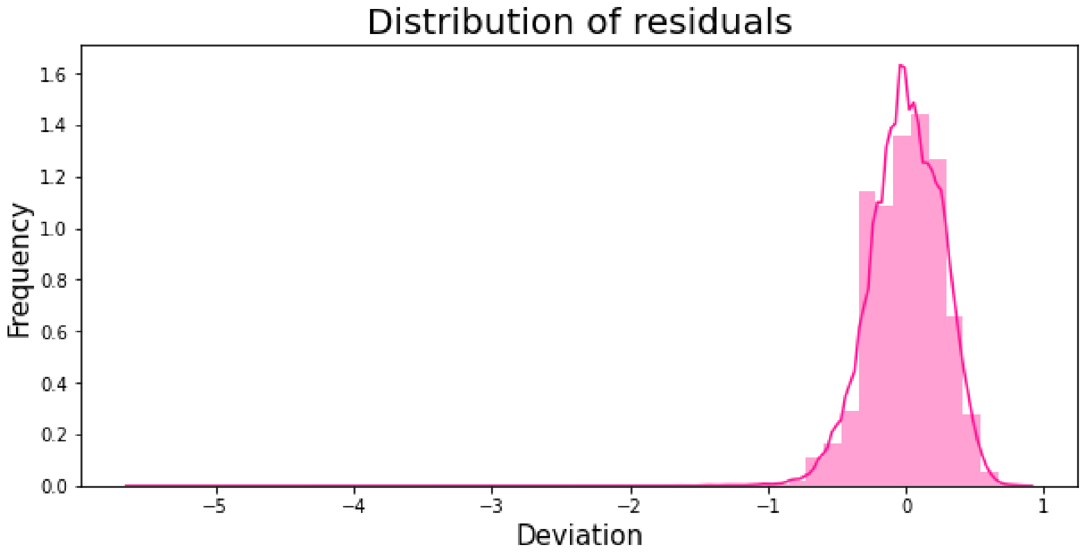


Figure 12: Distribution of residuals

We plot the distribution of residuals to see how good our estimation is. We can see that the error is quite low and it fluctuates around 0 which means that the estimators are well fitted. However, from the previous plot we saw that the points on our 'Price x Rating' plane line up in a characteristic way and it seems to be a good idea to use a model with a linear spline. Now our goal is to find the knot which will help us to fit the data properly.
A linear spline with a knot at τ_0 is:

$$\psi_{\tau_0}(x) = (x - \tau_0)_+ = \begin{cases} 0, & x < \tau_0 \\ (x - \tau_0), & x \geq \tau_0 \end{cases}$$

We want our function to be described as:

$$Y_i = \beta_0 + \beta_1 X_i + \alpha \psi_{\tau_0}(X_i) = \begin{cases} \beta_0 + \beta_1 X_i, & X_i < \tau_0 \\ \beta_0 + \beta_1 X_i + \alpha(X_i - \tau_0), & X_i \geq \tau_0 \end{cases}$$

To find the knot τ_0 we select a set of candidate estimates of τ_0 and calculate the mean squared error for each candidate. To be more precise, for each candidate for the knot we calculated their corresponding candidate estimates of β_0, β_1, α and found a set of values $\hat{\beta}_0, \hat{\beta}_1, \hat{\alpha}$ that minimizes the mean squared error. After these calculations we are able to select the proper estimator $\tau_0 = 37$. In summary, our function has the form:

$$Y_i = \beta_0 + \beta_1 X_i + \alpha \psi_{\tau_0}(X_i) = \begin{cases} 3.5878 + 0.0153 X_i, & X_i < 37 \\ 3.5878 + 0.0153 X_i + (-0.0147)(X_i - 37), & X_i \geq 37 \end{cases}$$

Now we can plot our function to see whether it fits the data better.

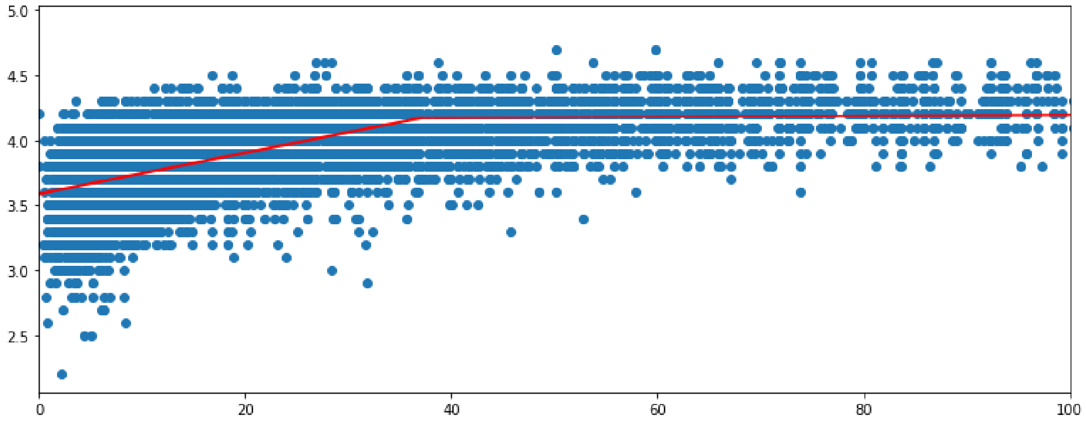


Figure 13: Linear Spline

We can see that the knot significantly improved the fit of our line, especially for the wines under 100 Euros. We chose that interval because it includes a large part of our data and most likely is a reasonable price range for a typical consumer.

3.2 Number Of Ratings Analysis

Let us explore the NumberOfRating variable more precisely. The best way to start is to see its distribution.

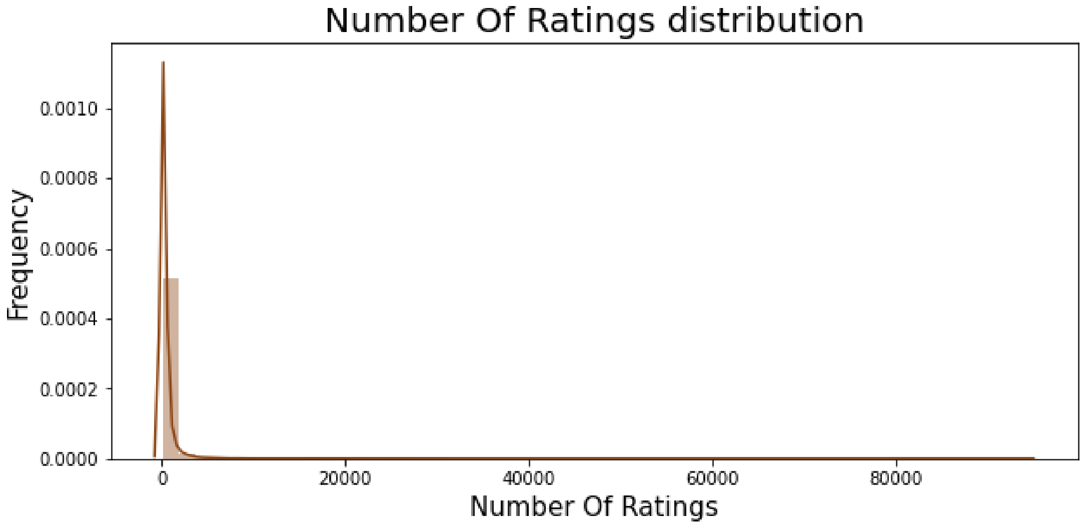


Figure 14: Number of Ratings distribution

Important fact is that our dataset only has wines with a number of ratings greater than or equal to 25. The plot may not be very detailed but it shows us that here we also have some outliers, some single very popular wines, but overall the average number of ratings is low. We use the natural log transformation once again to make the chart more readable. We present plots for all wines and the ones with number of ratings less than 1000.

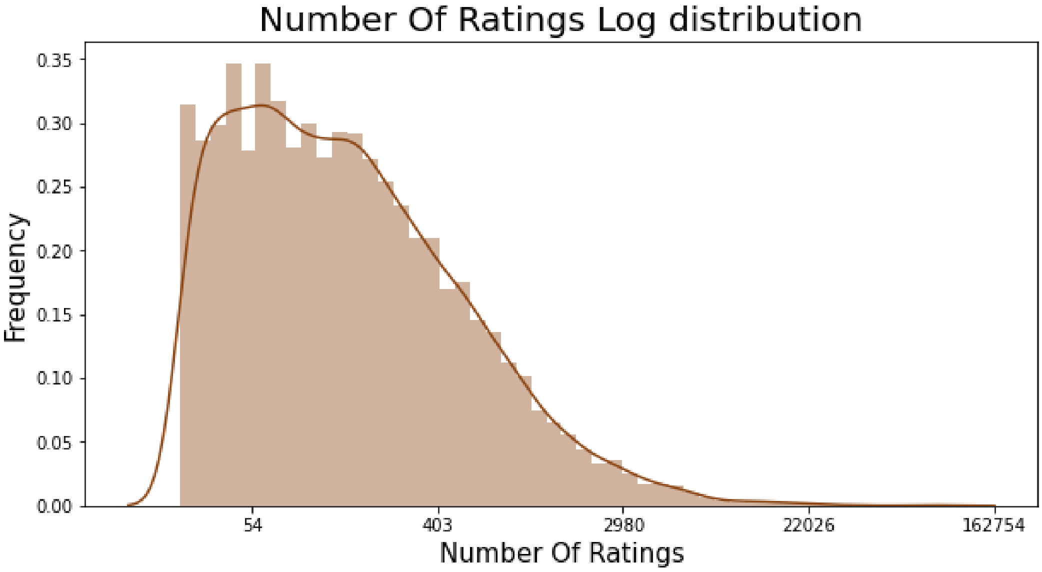


Figure 15: Number of Ratings Log distribution

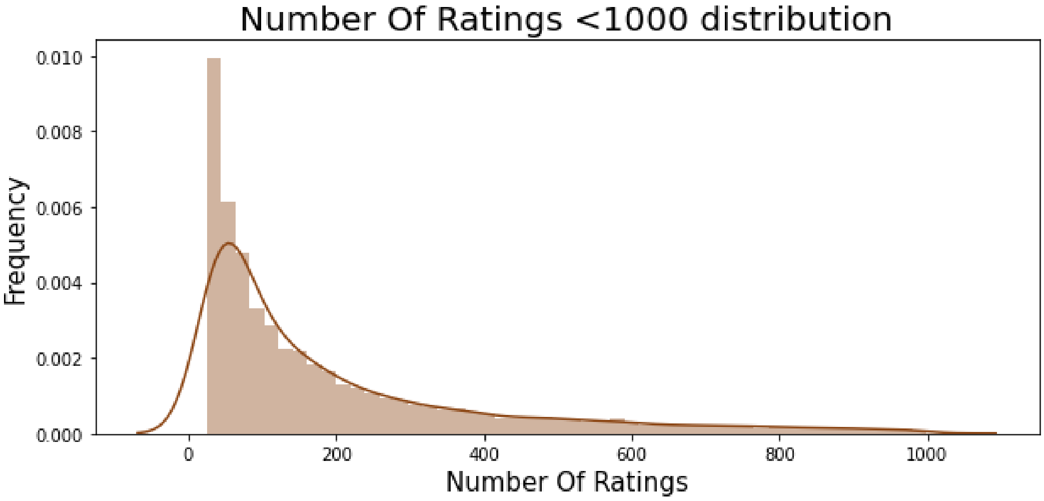
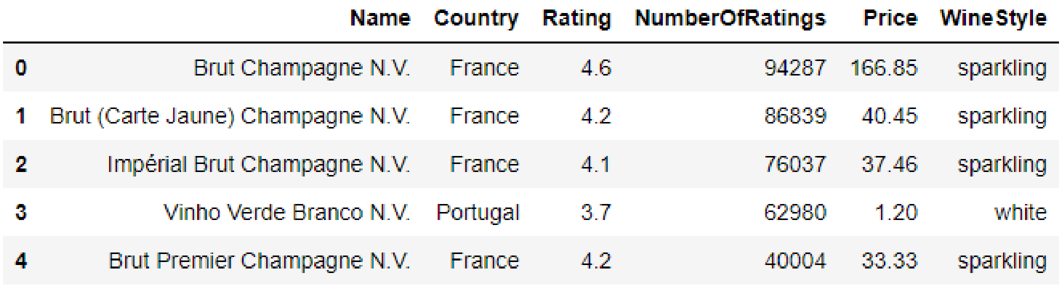


Figure 16: Number of Ratings less than 1000 distribution

To satisfy the reader's curiosity, we also present top 5 most popular wines in the Vivino app that cause the big spread for number of ratings.



	Name	Country	Rating	NumberOfRatings	Price	WineStyle
0	Brut Champagne N.V.	France	4.6	94287	166.85	sparkling
1	Brut (Carte Jaune) Champagne N.V.	France	4.2	86839	40.45	sparkling
2	Impérial Brut Champagne N.V.	France	4.1	76037	37.46	sparkling
3	Vinho Verde Branco N.V.	Portugal	3.7	62980	1.20	white
4	Brut Premier Champagne N.V.	France	4.2	40004	33.33	sparkling

Figure 17: Top 5 most popular wines

Now we are certain that a big part of our wines is close to the minimum number of ratings required to be stated in the app, which is 25 ratings. The conclusion is that this app may not be very popular or most of the wines in our dataset are rather unique.

3.3 Age Analysis

Let us check how the age of our wines is distributed.

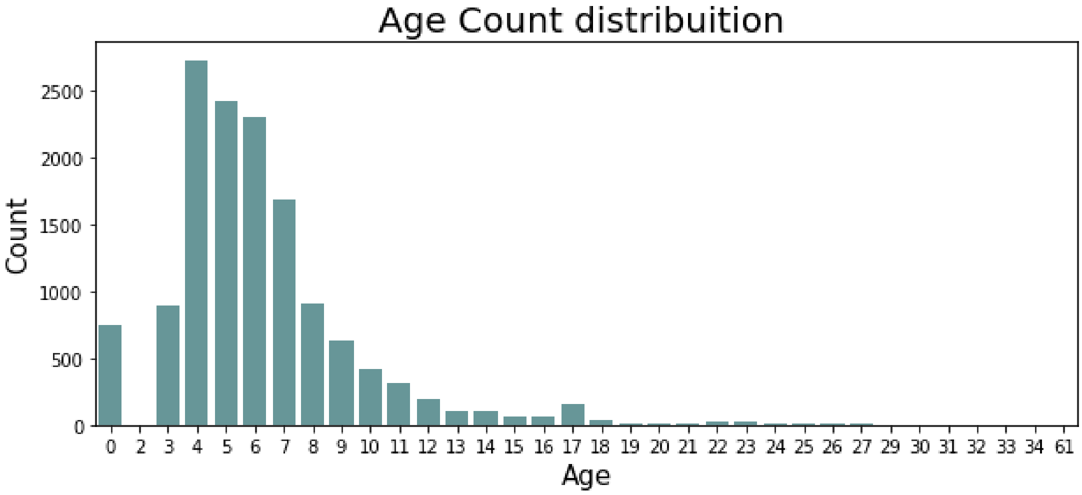


Figure 18: Age distribution

Most of the wines are quite young, almost 78% was produced after year 2014. The youngest vintage wines in our dataset were produced in 2019. To calculate the age we subtracted the year of manufacture from the current year 2022. Moreover, by age equal to 0 we denoted the Non-Vintage wines, which are usually a blend from the produce of two or more years. In general, it can be seen that older wines are more and more unique.

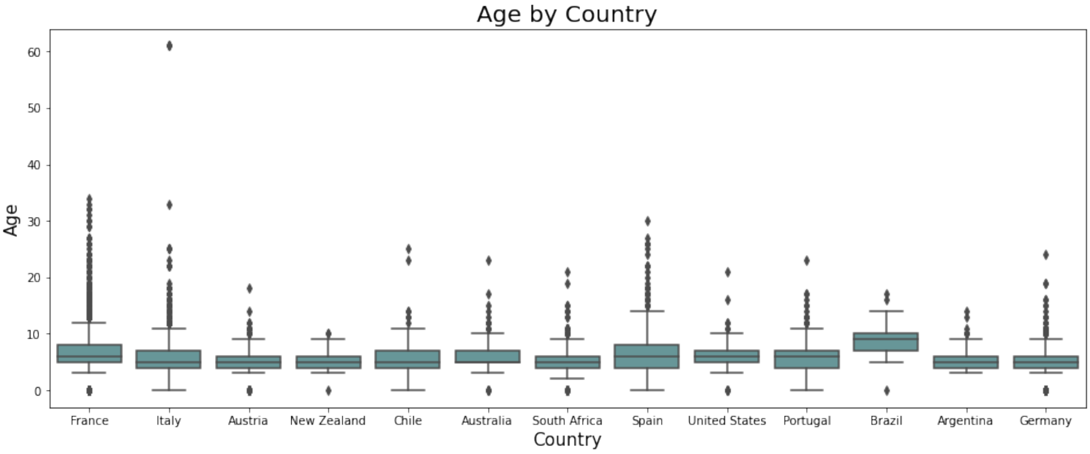


Figure 19: Age by Country

This graph does not tell us much, as the averages for individual countries are similar. Brazil stands out slightly, where on average the wine is older. One more interesting conclusion from this chart is that the oldest wine comes from Italy.

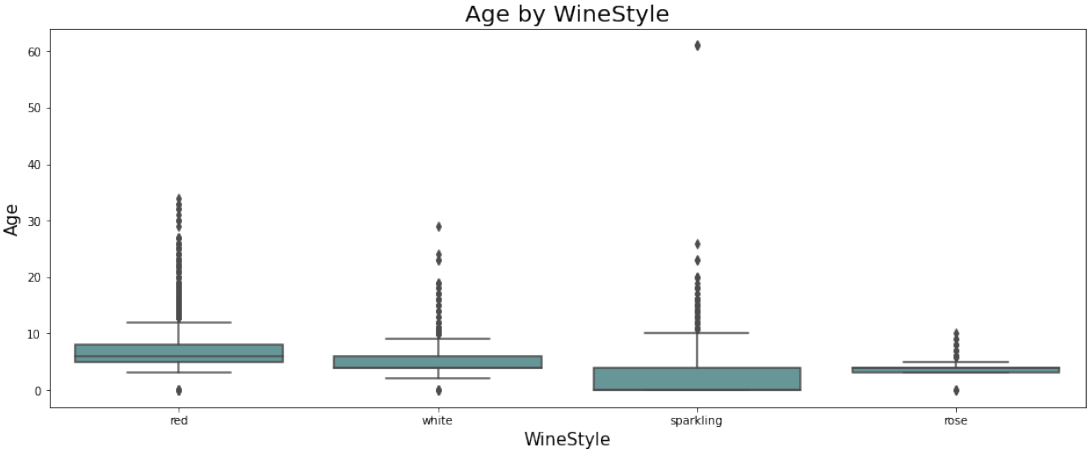


Figure 20: Age by WineStyle

It can be seen that the distribution of age for red and white wines is similar. On average, sparkling wines are the youngest, but among this type there is also the oldest wine, which is an interesting observation. The age of rose wines does not fluctuate much, most likely due to the small amount of wines of this type in our dataset.

4 ANOVA

For our ANOVA analysis we decided to split our dataset into groups based on their price, age, quantity of ratings, country of origin and wine style. We divided each of the variables into the following groups:

Price (Euros)			
very cheap	cheap	medium	expensive
<10	[10,20)	[20,100)	>100

Age (Years)			
very young	young	middle aged	old
<4	[4,6)	[6,17)	>17

Number Of Ratings		
small	medium	big
<90	[90,900)	>900

Wine Style			
red	white	sparkling	rose

In terms of countries, we only considered countries that export more or exactly 50 different wines. Applying the above divisions led us to the following interesting results.

4.1 One -way ANOVA for the price

Among the very cheap wines one country scored an outstanding result – in this price category New Zealand has the wines with the highest rating. The country that performed the worst in this analysis is Brazil, which apparently does not have good and very cheap wines. We observe really good results for the wines from Portugal, Germany, Italy and Austria.

Among the cheap wines no country outstands the others. The differences are noticeable but not significant. Again the worst performance was by Brazilian wines. The interesting fact is that even though there are not many Moldavian wines, they are undeniably the best in this price category. Moreover, the very cheap New Zealand wines from the previous price group perform worse only than Moldavian wines and better than every other country in the category of cheap wines.

In the category of medium price we do not have many observations for the Moldavian wines, despite that according to the ANOVA they are the best wines for that price. Except for this country, another good choice would be a wine from Argentina, Chile, Portugal or Spain. The cheap Moldavian wines still score the same as the medium priced wines from the listed countries. Very cheap wines from New Zealand do not perform much worse than the middle priced wines.

Expensive wines are undeniably better than cheaper wines, regardless of the country of origin. Only middle priced Moldavian wines can be compared with expensive wines.

The best average ratio of the wine rating to the wine price is for the Mexican and Bulgarian wines.

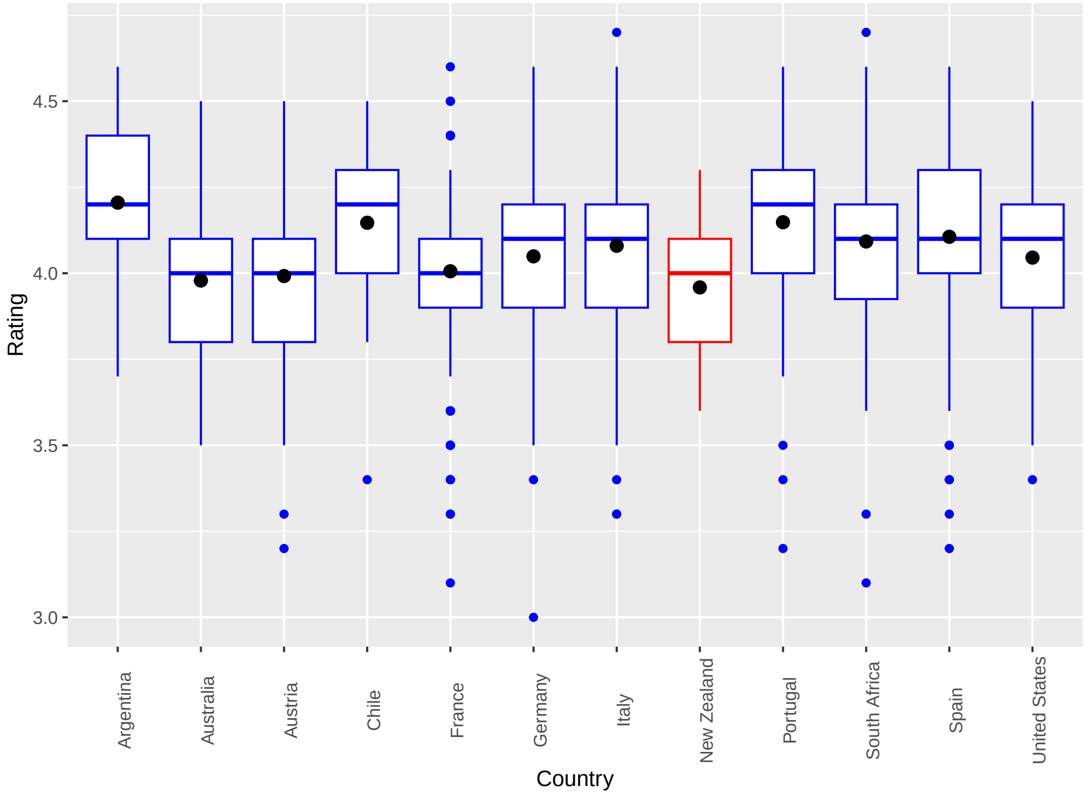


Figure 21: Comparison of very cheap New Zealand wines (in red) and rest of wines (in blue) in medium price category

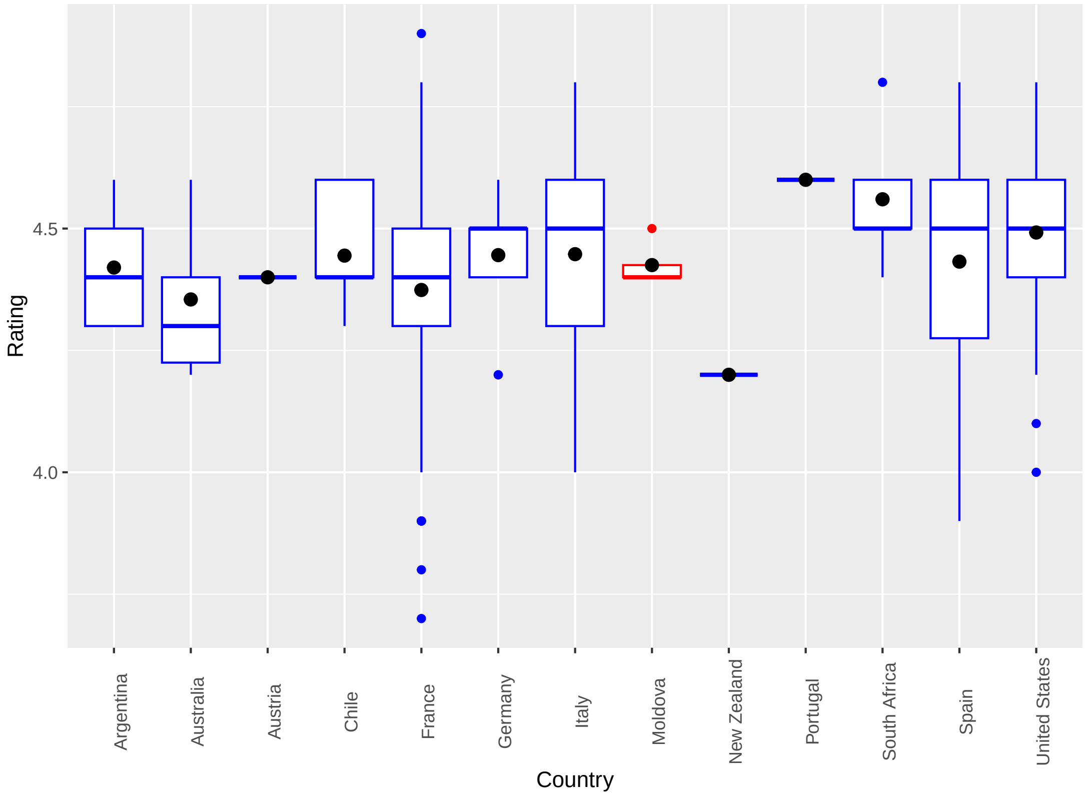


Figure 22: Comparison of medium price Moldovian wines (in red) and rest of wines (in blue) in expensive category

4.2 Two-way ANOVA for age and country

When it comes to the age of the wine it is clear that the older the wine the better. The only groups that performed similarly were young and very young wines.

4.3 Two-way ANOVA for number of ratings and country

The quantity of ratings also has an impact on the final rating. The more ratings the wine has the higher the final score is.
The countries with the biggest variety of wine are Portugal and Spain.

4.4 Three-way ANOVA for price, age and country

If we combine the analysis of the impact of the country of origin, price and age of the wine on the rating we see, that the New Zealands wines are outstanding even more not only in the category of very cheap but also very young wines. Another good choice are very young wines from Germany. It is worth noticing that very young and old wines perform better than young and middle aged wines.

Among cheap wines again the outstanding ones are young wines from New Zealand. They are comparable only with very cheap wines from the same country. In this category the worst performance is by Brazilian middle aged wines – they are overall one of the worst wines without looking at the price. The worst wines again are from the young and middle aged wines categories. Among middle aged wines there are no big differences between groups. Regardless of the country of origin and the age of the wine they are clearly better than cheap wines and worse than the expensive ones (except for the cheaper wines from New Zealand).

Expensive wines are comparable to each other, but significanly better than the rest are young and old Italian wines and middle aged American wines. The worst in this price category are the young wines from France and old wines from Spain.

4.5 Three-way ANOVA for price, quantity of ratings and country

If we base our groups on the country of origin, price and quantity of ratings we see that again the best performing group are very cheap wines from New Zealand, especially the ones with medium and big amount of ratings. In this price category they are comparable only with each other. Even if we add expensive wines to the comparison only Australian wines with the big amount of ratings can be compared to them. We also see again that there are no good wines in Brazil – no matter which variables we take into consideration, it performs equally bad. The top wines with medium price are wines from South America and South Africa with medium or big amount of ratings. This group can be compared with expensive wines. For the highest price the best ratings were scored by Italian and Spanish wine. Important observation is that for most of the countries better ratings are scored by wines with the bigger amount of ratings.

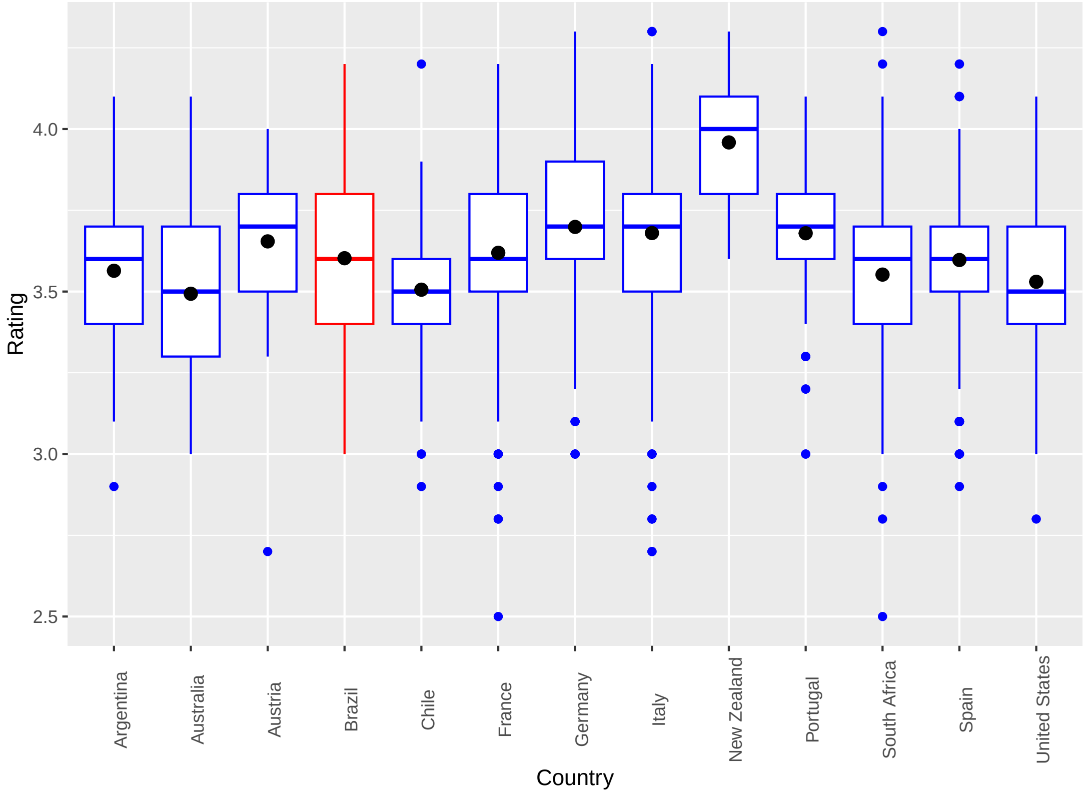


Figure 23: Comparison of medium price wines from Brazil (in red) and rest of wines (in blue) in very cheap category

4.6 Three-way ANOVA for price, age and number of ratings

If we combine the analysis of price, age and quantity of ratings at the first sight we see that for every price range except for the cheapest the dominating wines are the ones with more ratings, regardless of the year of production. Among the cheapest wines we see that the dominating group are the very young wines, regardless of the quantity of ratings. It is worth noticing that the difference between the groups for every price is very small. It is a rather interesting result, considering that when analysing connections only between price and year and only between price and quantity of ratings we had a result that year should be more significant variable than quantity of wines. For this grouping we do not observe any cheaper wines comparable with the expensive ones.

4.7 Three-way ANOVA for age, price and wine style

If we look at the groups created by age, price and style of the wine we see that among the very cheap wines there is one outstanding category – young sparkling wines. They are similar to the best of the cheap wines and comparable with the worst of the medium priced wines of every style. For the category of medium priced wines the outstanding ones are very young rose and red wines. The latter group draws our attention due to the great number of elements/wines in this category. Unfortunately they are still not even comparable to the worst wines from the group of expensive wines. For the highest price the best ratings are for sparkling wines at every age. Apart from these results, the very good performing group are young pink wines, unfortunately we do not have enough wines in this category to take it into consideration even if we take every possible price.

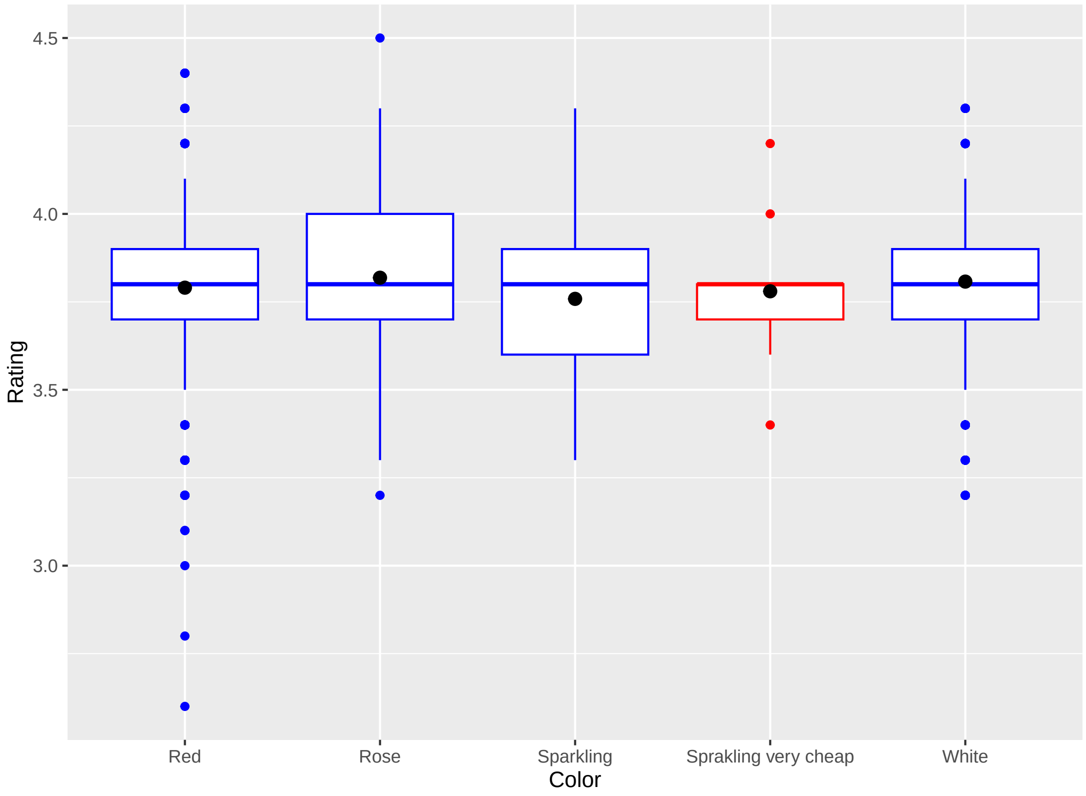


Figure 24: Comparison of very cheap sparkling wines (in red) and cheap wines (in blue)

4.8 Three-way ANOVA for wine style, age, price country

If we consider wine style, age, price and country of origin we see that among the cheapest wines again the outstanding ones are wines from New Zealand. Among very cheap and cheap wines the best performing categories were very young white wines from New Zealand and Spain. The interesting result of the analysis is the group of the worst cheap wines – it also includes some New Zealands and Australian wines, mostly red. Important lesson from this ANOVA is that we cannot base our choice only on country of origin but also on the style of the wine. For the medium priced wines except for the wines occurring in the previous analysis we can also observe a good performance of very young rose wines and medium aged sparkling wines from France. It is the first time French wines are dominating the category of medium priced wines. Among the expensive wines the best ones are old red wines from Italy and sparkling wines of every age from France. The worst wines are from France, they are comparable with wines from lower price category.

4.9 Four-way ANOVA for price, style, age, number of ratings

If we take into consideration price, age, number of rating and wine style we see that for the very cheap and cheap wines very good performance is by white, very young wines with many ratings. They are comparable with wines from medium price category and with none wines from the cheap category except for wines within the same group. For the wines in medium price category the dominating ones are medium aged sparkling wines with many ratings. Worth mentioning is the fact, that the worst performance in this category was also by sparkling wines but with small number of ratings. It is an important lesson that we should usually choose the wines with many ratings, even if the rest of the parameters is the same. For the expensive wines again the outstanding group are sparkling wines with many ratings.

4.10 ANOVA for all parameters – price, age, style, number of ratings, country of origin

It is really difficult to mention all the groups that are performing better than the others. For the cheapest wines except for the wines from New Zealand the good wines are also the very young wines of different styles with big or medium number of ratings. For cheap wines additional information to what we already know is the good performance of Spanish red and white wines with multiple ratings. For the medium priced wines we observe a surprisingly good performance of young red wines from Argentina with medium amount of ratings. They are significantly better than any other group in this price category and comparable with expensive wines. For the highest price category we observe the domination of sparkling wines and few groups of red wines – popular, middle aged wines from Italy, Spain and less popular wines from USA. We see that we should avoid expensive red wines from France since we can easily find equally good or better substitutes for smaller price.

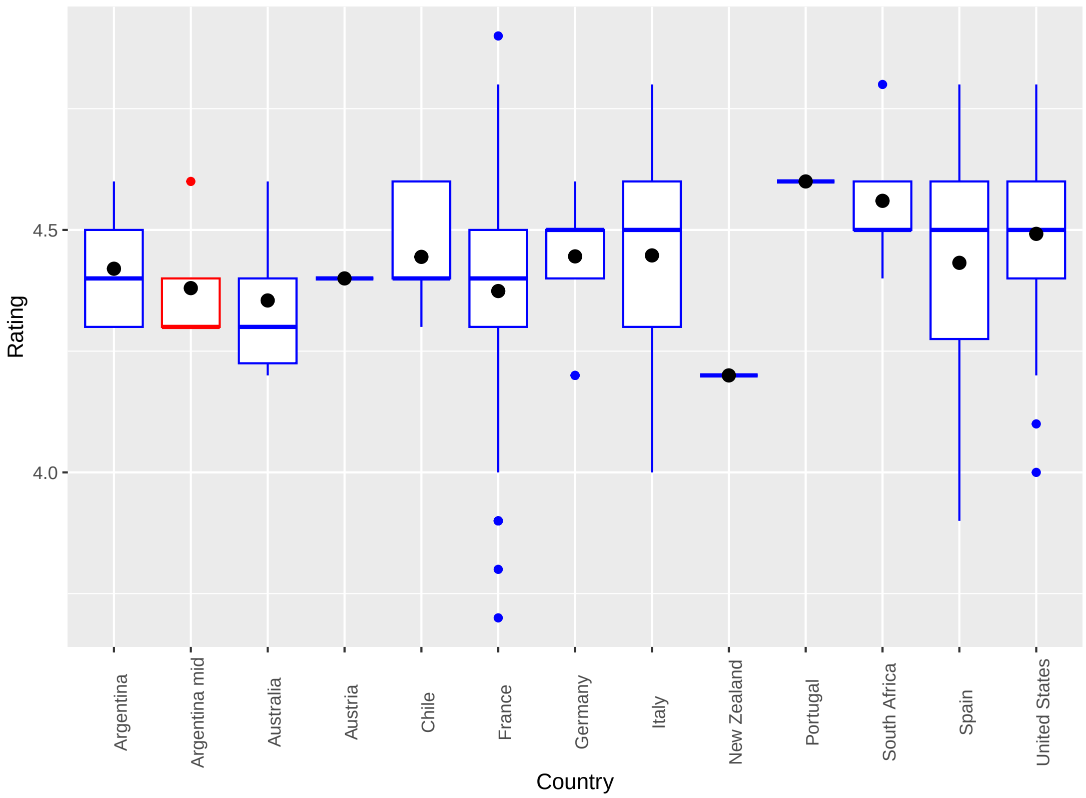


Figure 25: Comparison of medium price, medium amount of ratings, young red wines from Argentina (in red) and rest of wines (in blue) in expensive category

5 General Model

After our advanced analysis, now we are able to construct our general model. The model will take into account all of our analyzed variables, so our first task is to divide the dataset into groups with relatively similar number of observations. This setup will certainly improve the fit of our model to the data.

5.1 Preparing the groups

The subsets created from the variables in our dataset are not always equinumerous. We had to make our own division and here we show the groups we have chosen.

Red-Fr-Y-U	689	Red-It-O-P	527	Red-RoW-O-U	181	Oth-Fr-Y-P	359	Oth-Eur-Y-P	307
Red-Fr-Y-P	263	Red-Eur-Y-U	640	Red-RoW-O-P	98	Oth-Fr-O-U	154	Oth-Eur-O-U	214
Red-Fr-O-U	661	Red-Eur-Y-P	372	Red-Ame-Y-U	229	Oth-Fr-O-P	73	Oth-Row-Y-U	361
Red-Fr-O-P	643	Red-Eur-O-U	567	Red-Ame-Y-P	394	Oth-It-Y-U	793	Oth-Row-Y-P	143
Red-It-Y-U	829	Red-Eur-O-P	345	Red-Ame-O-U	125	Oth-It-Y-P	366	Oth-Ame-Y-U	161
Red-It-Y-P	577	Red-RoW-Y-U	334	Red-Ame-O-P	245	Oth-It-O	110	Oth-Ame-Y-P	136
Red-It-O-U	717	Red-RoW-Y-P	230	Oth-Fr-Y-U	594	Oth-Eur-Y-U	1291	Oth-O	75

At first glance, this table may seem intimidating, so we hasten to explain each symbol. First, we group by type of wine, dividing into red wines '-Red-' and the rest (white, rose, sparkling) denoted by '-Oth-'. We divide by nationality, where:

- '-Fr-' is France,
- '-It-' is Italy,
- '-Eur-' is Europe without France and Italy,
- '-Ame-' is North America and South America,
- '-RoW-' is the rest of world containing Africa, Asia and Australia.

We take into account the age of the wine, dividing it into young wines ('-Y-') if wine is not more than 6 years old and old wines ('-O-') when the age of wine is bigger than 6 years.

Finally, we consider the popularity of wines, dividing the groups into unpopular ('-U-') if the wine has less or exactly 200 ratings and popular ('-P-') if it has more. After defining each of the groups, we can proceed to implement the linear regression model for each individual group.

5.2 Applying linear regression

Let us write our equation in matrix form.

$$Y = X\beta + \epsilon,$$

For our particular model consisting of 35 groups, the matrix notation of the equation is as follows:

$$\left[\begin{array}{c} Y_1 \\ \vdots \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ \vdots \\ Y_{n_1+n_2} \\ \hline \vdots \\ \hline Y_{\sum_{i=1}^{34} n_i+1} \\ \vdots \\ Y_{\sum_{i=1}^{35} n_i} \end{array}\right] = \left[\begin{array}{cc|cc|c|cc} 1 & X_1 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & 0 & \vdots \\ 1 & X_{n_1} & 0 & 0 & \dots & 0 & 0 \\ \hline 0 & 0 & 1 & X_{n_1+1} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 1 & X_{n_1+n_2} & \dots & 0 & 0 \\ \hline \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ \hline 0 & 0 & 0 & 0 & \dots & 1 & X_{\sum_{i=1}^{34} n_i+1} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & X_{\sum_{i=1}^{35} n_i} \end{array}\right] \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{35} \end{bmatrix} + \varepsilon$$

We can use the formula that calculates the betas we need.

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

We received the betas and created graphs for each of the groups. To see how good the model is we plotted the distribution of resiudals. The sum of residuals was better than for the previous model. Therefore, it can be seen that the applied changes provided an improvement in the fit.

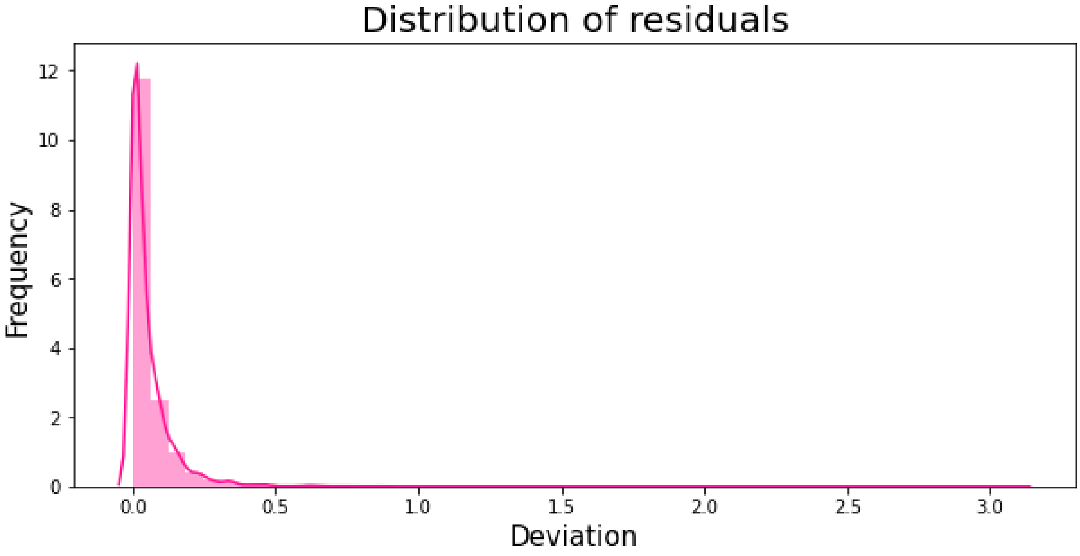


Figure 26: Residuals distribution

Again, we come to the conclusion that in each of the charts it is worth using the linear spline so we decide to use it.

5.3 Applying the linear spline to the model

Once again, we ended up searching for the formula:

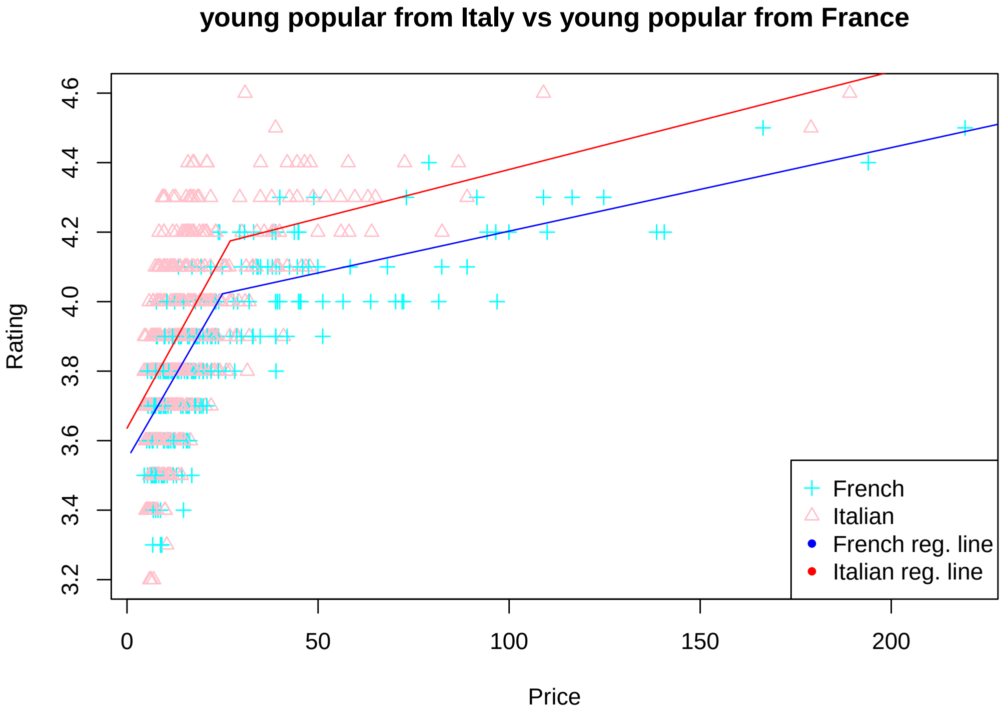
$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \alpha_j \psi_{\tau_{0j}}(X_{ij}) = \begin{cases} \beta_{0j} + \beta_{1j}X_{ij}, & X_{ij} < \tau_{0j} \\ \beta_{0j} + \beta_{1j}X_{ij} + \alpha_j(X_{ij} - \tau_{0j}), & X_{ij} \geq \tau_{0j} \end{cases}$$
$$\text{where} \quad j = 1, \ldots, 35, \quad i = 1, \ldots, n_j$$

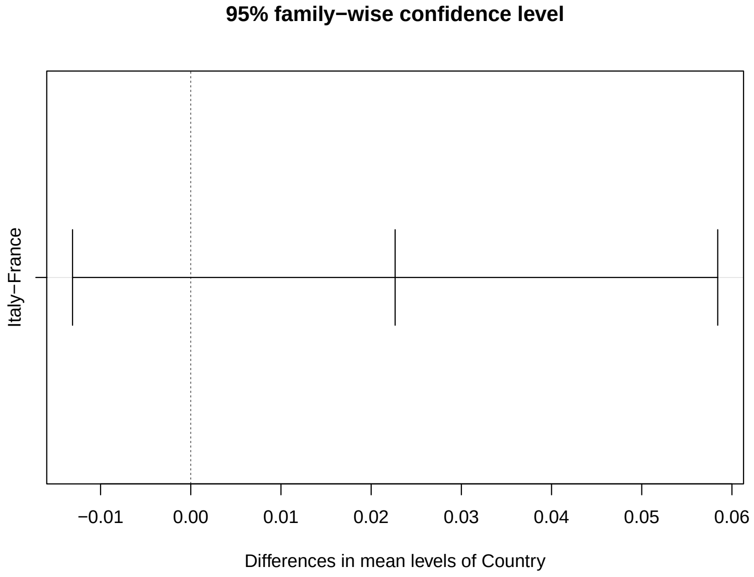
The key, however, is that this time this equation is searched for each of the groups. So we have a similar matrix equation but involving the knot.

$$\left[\begin{array}{c} Y_{1,1} \\ \vdots \\ Y_{n_1,1} \\ \hline Y_{1,2} \\ \vdots \\ Y_{n_2,2} \\ \hline \vdots \\ \hline Y_{1,35} \\ \vdots \\ Y_{n_{35},35} \end{array}\right] = \left[\begin{array}{cc|cc|c|cc} 1 & X_{1,1} & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & 0 & \vdots \\ 1 & X_{n_1,1} & 0 & 0 & \ldots & 0 & 0 \\ \hline 0 & 0 & 1 & X_{1,2} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 1 & X_{n_2,2} & \ldots & 0 & 0 \\ \hline \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ \hline 0 & 0 & 0 & 0 & \ldots & 1 & X_{1,35} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & X_{n_{35},35} \end{array}\right] \cdot \begin{bmatrix} \beta_{0,1} \\ \beta_{1,1} \\ \alpha_1 \\ \beta_{0,2} \\ \beta_{1,2} \\ \alpha_2 \\ \vdots \\ \beta_{0,35} \\ \beta_{1,35} \\ \alpha_{35} \end{bmatrix} + \varepsilon$$

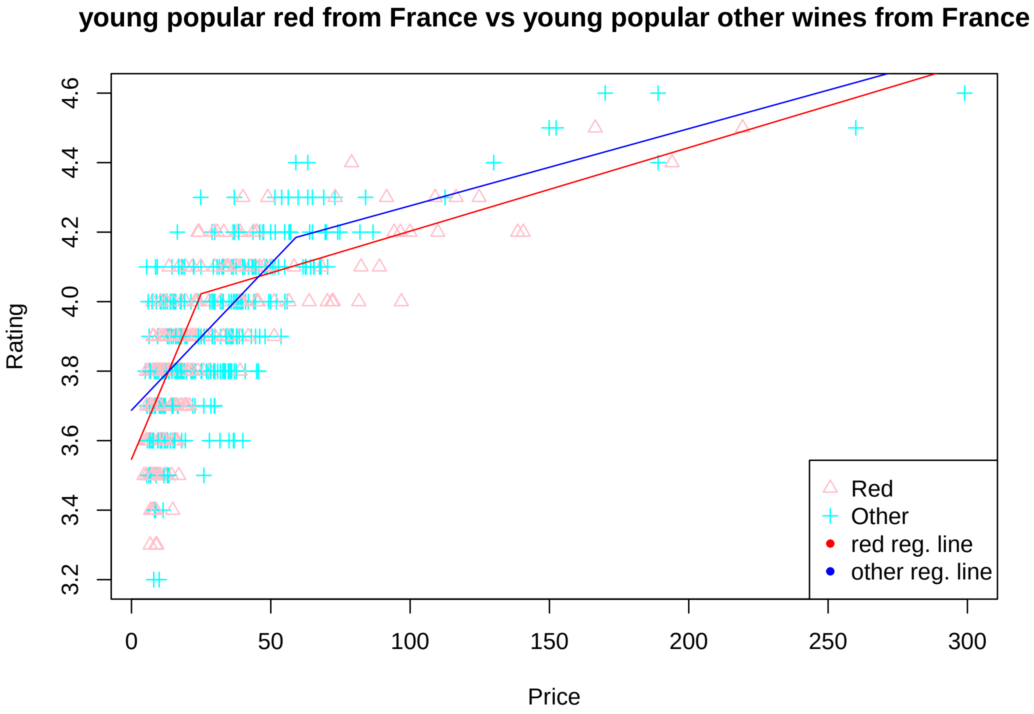
We managed to create a model for each of the group involving the linear spline. The sum of residuals was even smaller than for the previous one so the result was satisfactory. This version of the model is our general model.

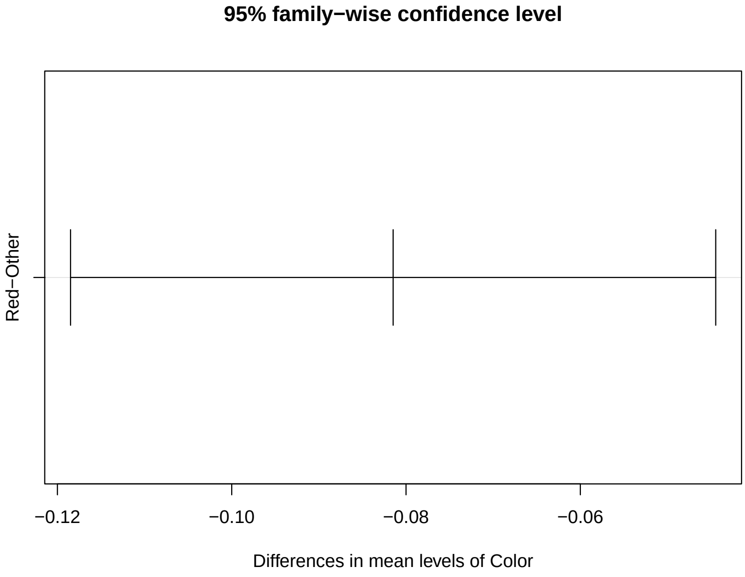
For visualization, we have a graph for each of the groups (thus 35), so showing all of them is pointless, because it is hard to draw the right conclusions. We selected certain groups for the analysis and compared them, placing them on one chart to see the differences clearly. In addition to each plot, we made confidence intervals to understand it better. Here are the more interesting comparisons.



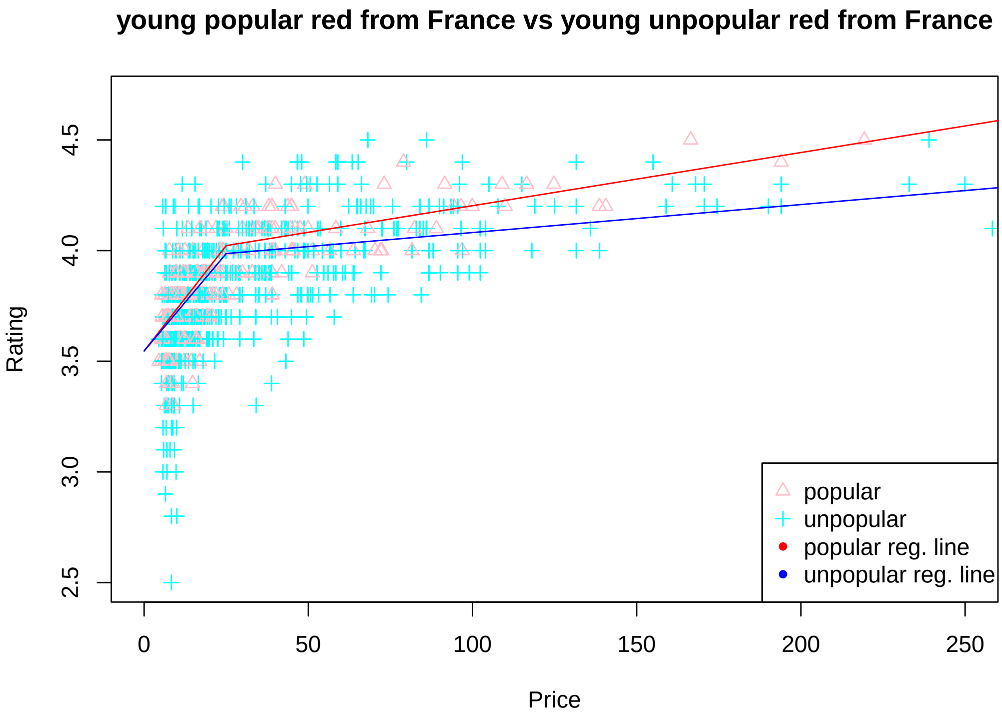


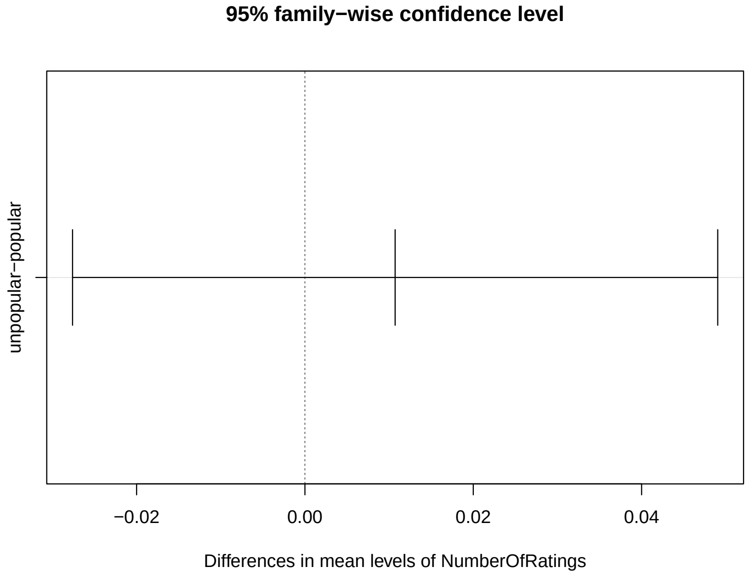
We see, that it in the category of young and popular wines it is wise to choose Italian wine over French wine, no matter the price. For a fixed price, almost all wines from France have multiple substitutes from Italy with better rating.



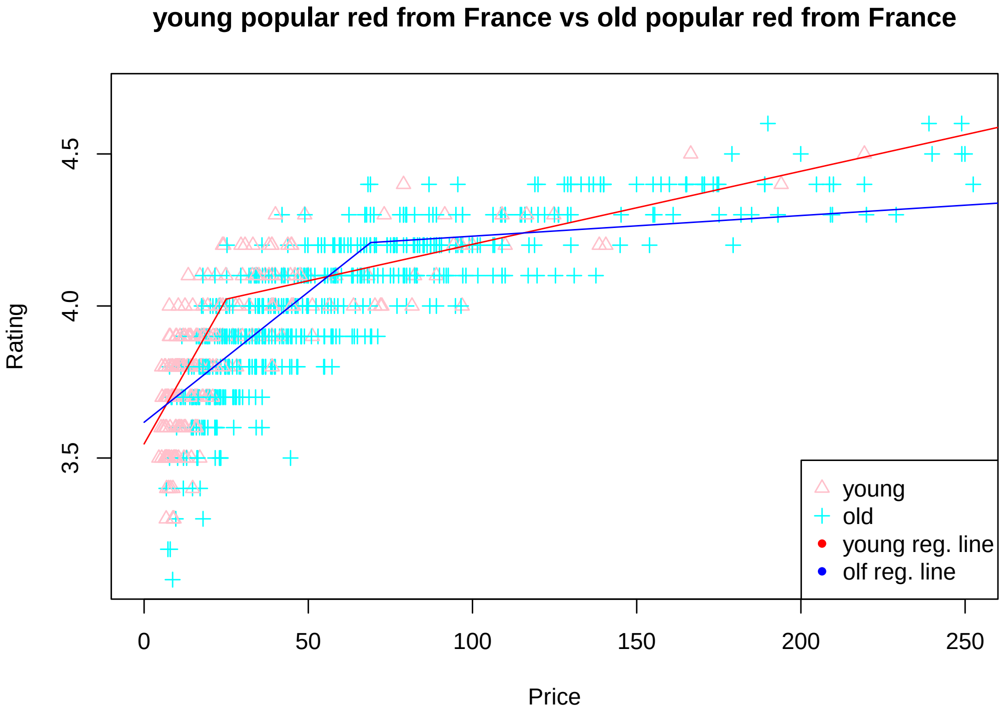


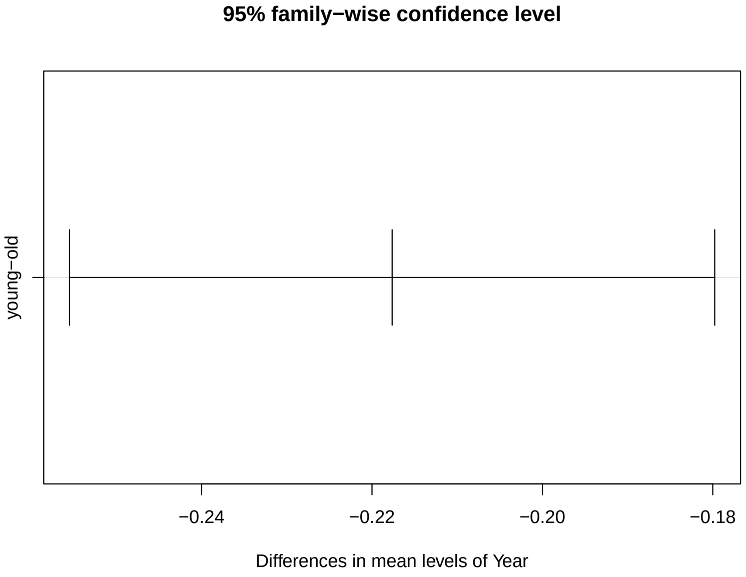
In a group of young, popular wines from France we see that if we look for a cheap wine we should choose a red one, since to the certain point rating of red wine increases faster than the other styles. For the higher price we see that white, rose and sparkling wines tend to be better than red.





For the young red wines from France if we decide on the more expensive wine, we should look for a popular one. Ratings for the wines drunk more often grow faster than for the unpopular wines.





Young popular red wines are usually better than the old ones. To avoid getting disappointed, no matter the price, between these categories we should choose the younger wine.