INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

INSA Toulouse

Big Data Report : Wine Review analysis What are the factors impacting the rate of a wine? Are they other factors to consider when buying wine?

Auteurs:
Aurélien FLOUTARD
Lucile FONTAINE

Professor:
Gilles TREDAN
Roberto PASQUA

January 27, 2020

Contents

1 Introduction

As wine enthusiastic people, we have chosen to study a dataset that gathered reviews and various information about wines from 42 different countries. We found the data on the online platform Kaggle. The original dataset is composed of 150,930 observations scrape from WineEnthusiast catalog in November 2017. This analysis highlights the factors that influenced the wine rating such as price or origin. We also tried to identify keys and patterns of the wine consumption that can help us to choose the best wine.

2 Is there a correlation between the price and the rating?

2.1 The rating distribution

First of all, we realized that most of the time, when we bought a bottle, we referred to the price to assess its quality. It is why we tried to find out if the wine rating is correlated to its price. One thing you may know is that the ratings in the dataset are going from 80 to 100. In the figure below, you may find the distribution of ratings.

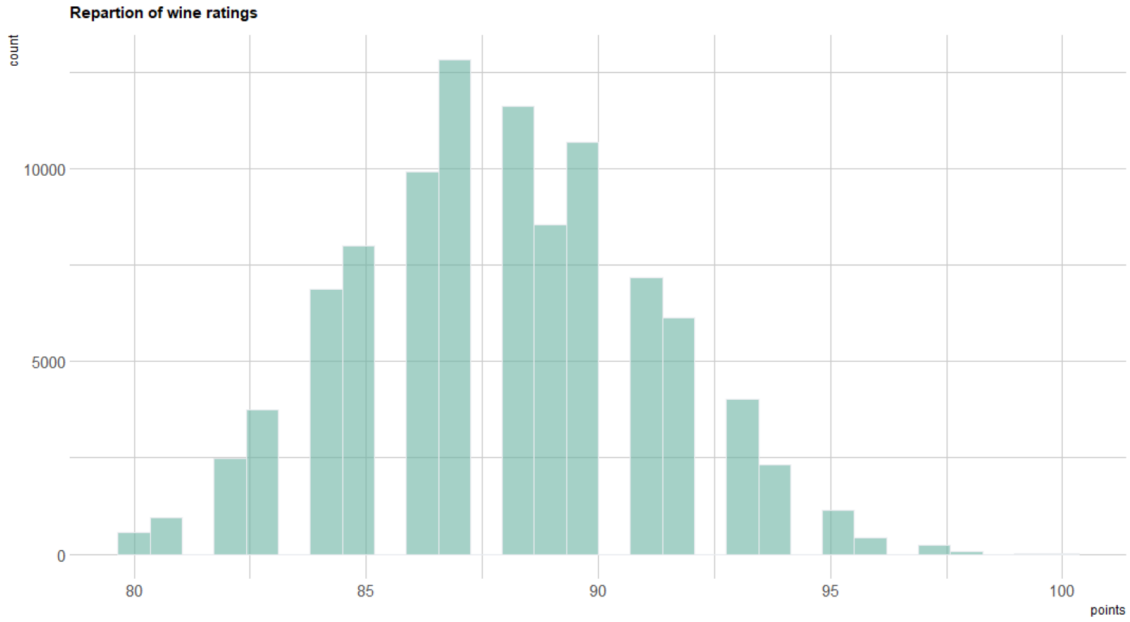

Figure 1: Distribution of rating between 80 and 100

The Figure 1 reveals that very good wines, it means more than 98, are extremely rare. In fact, only 103 out of 97 821 wines get a rating higher than 98. Most of the wines get a rating between 85 and 90.

2.2 Focus on the US, France and Italy

To have a realistic idea of the correlation between the price and the rating of wines, we decided to focus only on the three countries most represented in the dataset. As a matter of fact, you can see in the Figure 2 below that US (40531 reviews), France (14847 reviews) and Italy (14452 reviews) are preponderant in the dataset.The data has been extracted from a american magazine so it is not so surprising to see that The United States represents 42% of all reviews, France is 17% and Italy is 15%.

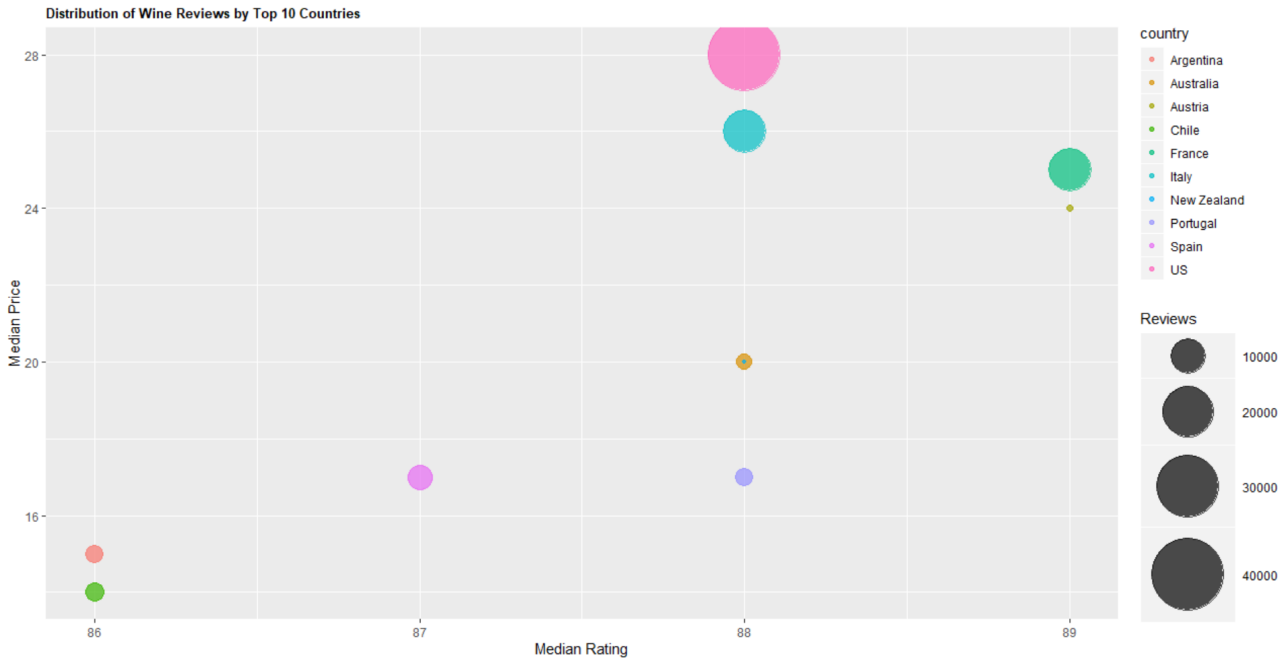

Figure 2: Mapping of the top 10 countries

We have chosen to do the mapping using the median values because it is more significant than using the average values. For instance, the average price of French wine is \$45 while the median price is only 25. A minority of very expansive wine is influencing the average price whereas half of wines is less than \$25. This reasoning is also applicable for American and Italian wines.

2.3 The correlation between price and ratings

The plot of the relationship between prices and ratings is the following :

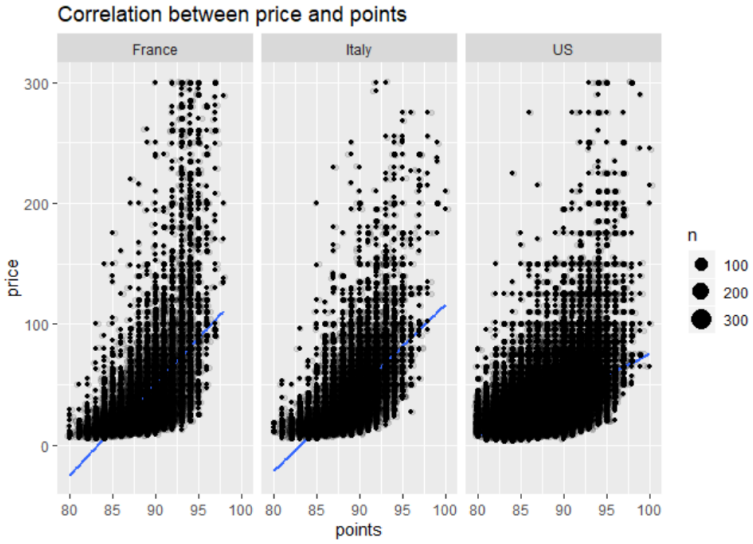

Figure 3: Correlation between price and rating

Unsurprisingly, in these three countries, higher-priced wines tend to have higher ratings. Nonetheless, we remark also that wines that cost between 10\$ and 100\$ can have an excellent grade above 95.

3 Varietal rating and prices by country

In this part, we focused mainly on the variety of the wine. We took into account the ratings according to the variety, the prices but also the origin country of the wine.

3.1 What are the top varieties on the dataset?

The graph below is presenting the mean ratings and prices of the varieties most represented in the dataset. These varieties are all ranked near 90. Nevertheless, there is a lot of disparities in terms of price that varied from 16$ for the Portugese Red to 56$ for the Nebbiolo.

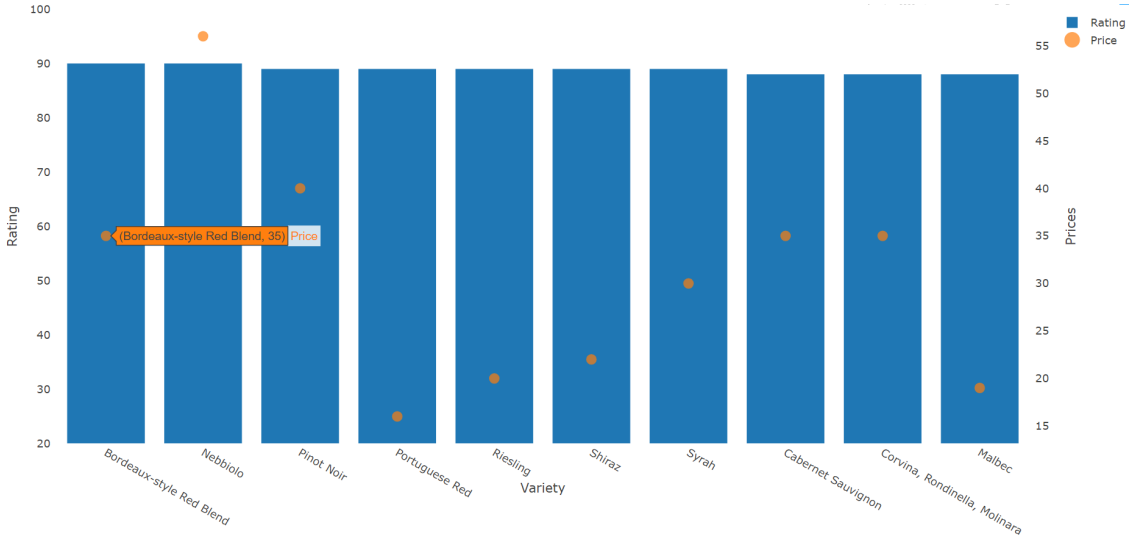

Figure 4: Rating and prices of top varieties

3.2 The origin of wine varieties and the visible changes depending of countries

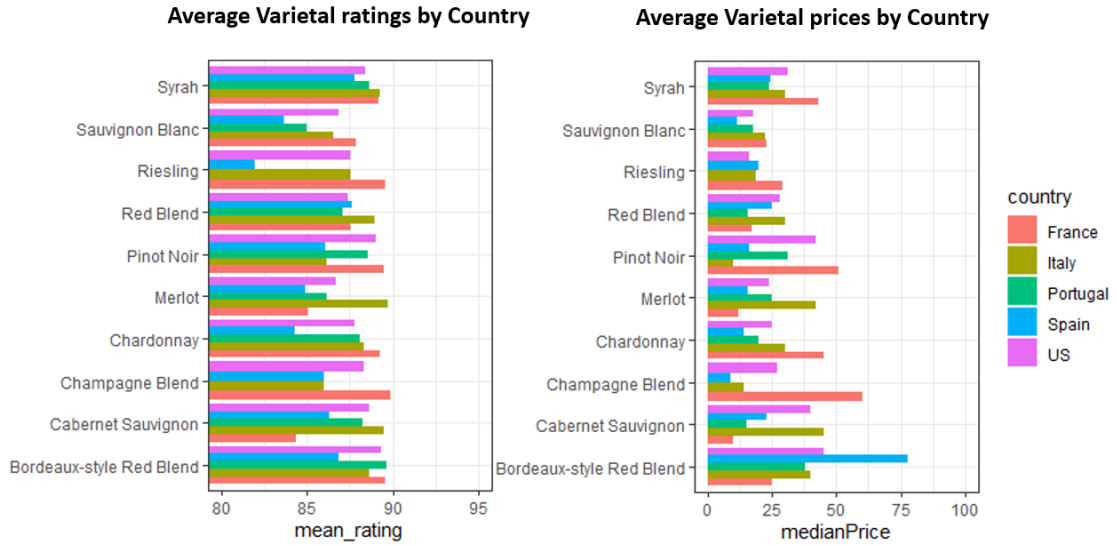

Figure 5: Average rating and prices of top variety in the Top 5 countries

The Figure 6 is giving us some interesting information. For instance, The "Bordeaux-style Red Blend" is rated between 87 and 90 in nearly all the countries selected. Nonetheless, looking at the prices, we note that the "Bordeaux-style Red Blend" in Spain is three times more expansive in Spain than in France. According to our dataset, if you want the best value for money when buying your bottle of Bordeaux, you should choose the french one.

Another interesting point visible on the graph is that Champagne is not only present in France but you can also find italian, american or spanish Champagne. The spanish and italian Champagne are much cheaper than the french one. But according to the dataset their quality (around 86) is lower than American and French Champagne. Nonetheless, it can be a good deal for those who don't want to spend to much money in a champagne bottle.

3.3 Data mining on Wine Reviews

This dataset is also adapted to do some text analysis. So we tried to extract information from wine descriptions. Before drawing any conclusion, it was necessary to prepare the text analysis by using the *tm package* in order to remove some words and characters useless for our analysis.

- Creation of a vector with the descriptions
- Load the descriptions data as a corpus
- Cleaning : remove stop words, white space, punctuation...
- Building of a "Term Document Matrix"

Thanks to this preparation, we were able to deduct that reviewers are describing often the flavors, the colour, the different notes of wines. The most popular words are the followings : "age", "aroma", "balance","palat"... We also find some words related to fruit : "blackberry", "fruit","cherries"...
Taking into account the great amount of text data on the dataset, we decided to only focus on portugese wine descriptions in our graphs. We made the same cleaning as described above and we focused on the top words :

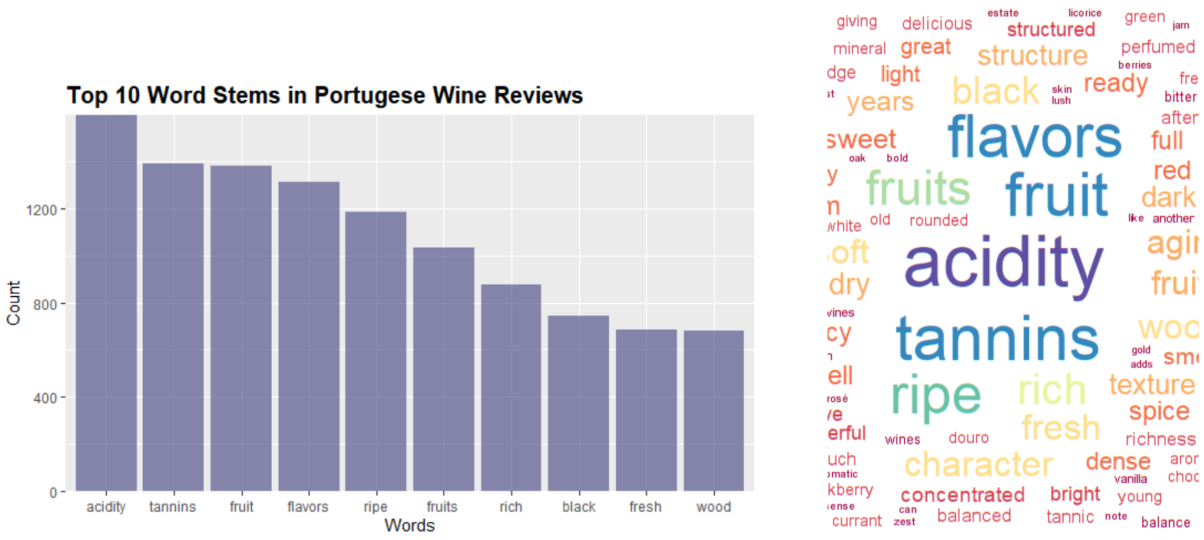

Figure 6: The Top 10 words and the most popular words in Portugese wines reviews

4 Conclusion

Thus, this analysis reveals that there is a strong correlation between prices and ratings of wines. Nonetheless, there are also some good wines that do not exceed $20. Considering this, we understand that the price is not the only factors to take into account when you buy a wine bottle. Other important aspects are the variety and the origin of wines. In fact, knowing more about the wine diversity can enable us to find the best value for money wine. Finally, this dataset enable us to try some basic data mining analysis. Looking at the top words present in the description reviews, it enable us to identify some patterns of consumption. We tried as mush as possible to vary the type of visualisation to extract the key information of the dataset.