# *Wine Ratings – Which Wine is the Best?*

Andrea Fox

Summer 2021

<https://github.com/anfox86/Wine-Ratings-Which-Wine-is-the-Best>

**Introduction**

For this project I wanted to explore the world of wine and wine ratings. I have never been much of a wine person, more a beer gal, so I wanted to learn more about wine since all my friends are self-proclaimed winos. The biggest advice I have always been given is that since I like sweet things to go with a sweet wine, but I was wondering if ratings were something I should look at more when determining choice. According to an article I read, “wine ratings actually drive price, sales, and reputation.” (Is Robert Parker high? Comparing wine raters, 2017) So, I decided to look at wine ratings by different wine types, region, winery, and country to see which one is the best, and if it is possible, I plan to buy it to try.

**Business Problem/Hypothesis**

My approach for this topic is like my last one, where I am the client who is trying to find the best rated wine. I went through Kaggle and found a few different datasets that might work, but ultimately chose the dataset Wine Rating & Price (Budnyak, 2020). This dataset consisted of 5 csv files that contained the following: varieties, red, white, sparkling, and rose. Varieties only had one column of different names, so I opted not to use that one and focus on the other four.

Some of my initial questions are exploratory data analysis driven such as does the vintage year influence the rating? I also wanted to see if there was any kind of relationship between wine and the price. I also wanted to see if country or region had any affect on ratings, since many areas are known exclusively for wine. Then I want to dive into some machine learning methods and look to see if I can predict the rating of a wine based on location, price, or rating. I plan to do something like project one and score each method to see which would be the best for this project then dive into the machine learning. My last topic that I want to research may be stretching my capabilities some, but I want to know if there are types of wines from a specific region that generally score well but have an off year. My thought process here is Napa Valley in California is known for being wine country, but maybe they have a bad year due to wildfires that consistently happen in California and surrounding areas. For the last question I’ll need to find the wines that rate the best then see if there are any patterns in the area bottled. If there are any that score well normally but have an off year or two then I could look up the region to see if anything happened that might cause an issue.

**Clean-up**

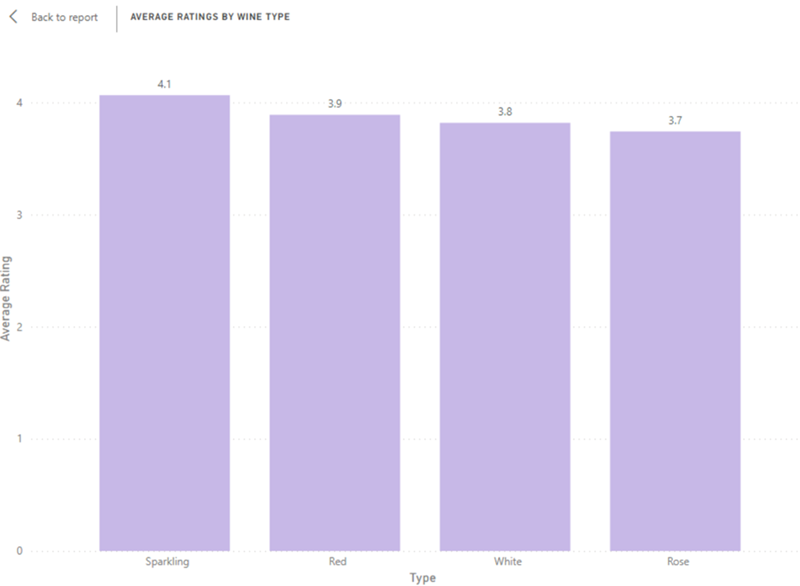
The first thing I did was go to Kaggle and download the four different csv files I intended to use. I then read all four into my Jupyter notebook. I started with Rose since it looked to take up the smallest amount of storage and worked my way to red wines. For each file I printed them out to look at variables and get a general idea of what the data in each looked like. All four of them had the same 8 columns named the same, so I did not need to remove any columns since they were all relevant to my analysis. I then looked at all four dataframes to see if any had null values. This part blew my mind because I had no null values in the dataframes, so I did not have to do any clean-up there. I then looked at each dataframe by describe so I could see count, mean, standard deviation, minimum, and maximum counts. This was nice to see face up because I could see the mean for both rating and price by each wine type.

The only real work I ended up doing during my clean-up was to add an additional column to each dataframe called ‘Type’. For each dataframe I added the additional column and filled the values with the corresponding type. For example, I added the column Type into the dataframe for red wines and then repeated the values in Type as red. This way when I combined the four dataframes into one I still knew which wines matched up to the name, country, region, ratings, and price. After adding the column into each dataframe I then created the variable wine and used pd.concat to stack each dataframe on top of each other. This was not hard to do since the headers were all the same and the columns all fell into the same order. I found when creating visuals that I had missed one step in clean up, so I came back to replace the values N.V. in the year with 0. This is because N.V. is not an integer and was creating some issues when reading into PowerBi. Once this was complete, I saved my new dataframe as a csv to use for visuals and processing.

**Visual Analysis**

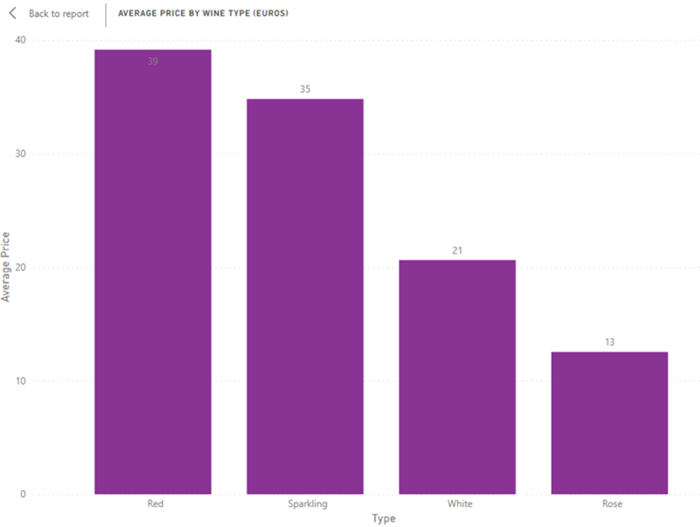
For my visualizations I used PowerBi as I find it the most intuitive tool to use when it comes to creating charts. I created three dashboards one named Overview, the second one named Location, and the third named Country Specifics. On the Overview dashboard I really spent time looking at rating by different areas as well as price and number of ratings. On the first visual it was a simple look at Average Ratings by Wine Type which showed Sparkling wines had the highest average rating, while Rose wines had the lowest average ratings. I was surprised that by average ratings, Sparkling wine did so well as usually I have either red or white recommended to me. I also wonder here if because there are less names in sparkling or number of reviews whether this had any affect on the average.

*Figure 1 – Average Ratings by Wine Type*



I then looked at the average number of ratings by wine type and saw that sparkling had a much higher number than any other wine type. Sparkling sat at 1506.08 while the next closest wine type only had 415.29 wine ratings. This probably did skew the data on my first chart some, but there really is no good way to lower that without affecting the validity of the data. I also wanted to look at prices of wine because I was curious if wine ratings and price had any relationship. Here Red wines had a higher price at 39 euros, while the lowest price wine type is Rose at 13 euros.

*Figure 2 – Average Price by Wine Type (Euros)*

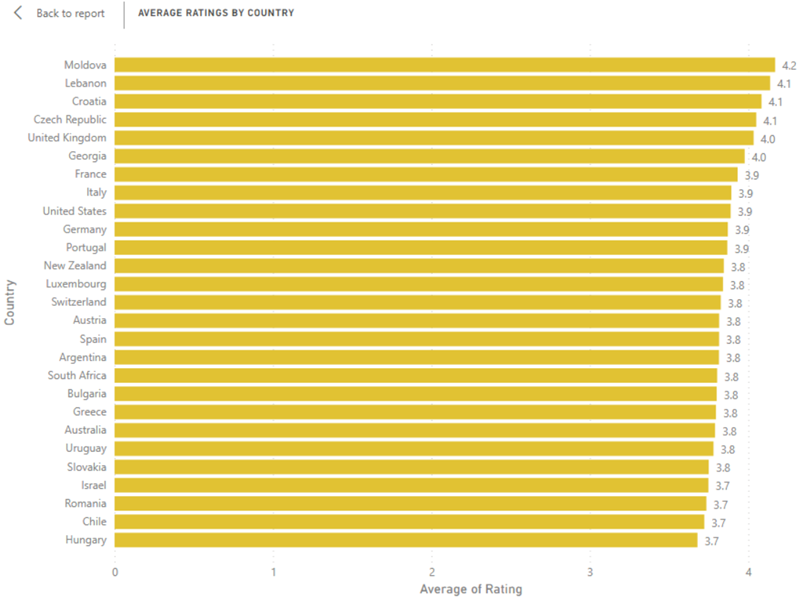


One thing I did not consider here is what is the difference between Sparkling wine and Champagne. I know in my local grocery store champagne is stored under the sparkling wine sign, so I had to go look this up since I know champagne can be somewhat pricey. I found an article that explained that champagne is sparkling wine, but earned the name based on the region in France called Champagne. It is a mistake on the US and other places to call sparkling wine champagne because true champagne comes from Champagne France, and they must use specific grapes and processes to make it. (Heid, 2020) I guess this means when I make mimosas using California “champagne” I am only using sparkling wine.

The next thing to see here was average ratings by vintage year and non-vintage year. Non-vintage means it is a combination of different wine vintages so that could easily span more than one year. Here I found Red wines rated higher followed by Sparkling, Rose, and White. For vintage year, I looked at the average rating by vintage year instead of specific types. The year that had the highest rating with a 4.55 was 1989. I have a feeling this is a limitation on my dataset and most likely year would play a much larger role in rating if I had any wines that went farther back than 1960. Lastly, on the Overview tab I looked at average ratings (4.5 and higher) by Wineries. I ended up replacing the last visual on the Overview during my technical analysis. This is because I wanted to look at country by average rating and wine type. I was kind of surprised that there is a equal mixture of wine types by country. There were only a few that had either one or two wine types. The countries with only one wine type are Czech Republic - White, United Kingdom - Sparkling, and Mexico – Red. For the Czech Republic and the United Kingdom, the average rating for their wine types was typically higher than those same wine types associated with countries known for wine. However, Mexico did not have an average rating higher for other countries with Red wines.

On the second dashboard I looked at wine ratings by various locations. I first looked at the average ratings by county.

*Figure 3 – Average Ratings by Country*



I was surprised by this one because I fully expected either France or Italy to have the highest ratings by country since both are well known for wine. Another surprise is that the United States, France, and Italy all have the same average rating at 3.9. I then looked at average ratings by region for the different continents. I chose doing it by continent because otherwise I had an overwhelming number of lines in my charts that would just look messy. However, I did combine Asia, Africa, and Oceania because those regions did not have a ton of ratings. When broken down into continents looking at the highest rating is in Western Europe specifically in France sitting at 4.9. The highest rating in each is North America the US at 4.6, South America Argentina at 4.5, Eastern Europe Moldova at 4.2, Western Europe France 4.9, and Asia/Africa/Oceania Lebanon at 4.4. I enjoyed looking at the wine ratings by countries because some countries were listed that I would never have associated with winemaking like Georgia or Israel.

The last dashboard I have was added later because I wanted to see totals for each country by type, and then take look closer at the top 4 countries (by total). France, Spain, and Italy had a higher total count for Red wines, while Germany had a higher total count for White wines. Spain and Italy had a higher average rating for their red wines, while France has a higher average rating for Sparkling wines. Germany had the highest average rating under white wines. It looks like in most cases if a country has a higher count of a specific type of wine it will most likely have a higher average rating for that wine type.

**Technical Analysis**

My technical analysis ended up being like my last project where I’m really looking at the numbers. I read in my new csv and removed the extra column that it always seems to add called Unnamed: 0. I started the actual analysis looking at the correlation between year and rating. There was a weak, positive result so this did not give me much. I decided to start looking at numbers by looking at unique values for Ratings. I then looked at the rows for the highest, second highest, lowest, and second lowest ratings. I did not look like year really played any role looking at the specific rows, but for the most part the higher rated wines also had higher prices. This led me to looking at wine.describe() so I could look at the rows with the highest and lowest prices. For the highest price at 3.410.79 euros, was a red wine that was rated 4.7. The lowest price at 3.15 euros had two wines at 4.2 and 3.8 ratings. I also saw with the lowest price that both wines came from the same country, region, and winery. Last thing I did was look at the wineries with lowest and highest prices to see if there were any other wines produced by those wineries. There were no other wines for those wineries in the dataset other than the ones I found for lowest and highest prices.

**Conclusion**

Overall, I was not able to find the right wine for myself because of one big missing piece from this dataset. Since there is not a column that talks about flavor for each wine, it would be hard to pick one that might entice me. I found this article that mentioned when looking for a wine you must first ask yourself “How do you like your tea?” (Augur, 2016) For me, I know I typically like sweet and fruity things which looking up the wine that had the highest rating it is more floral with some fruit flavor. Without flavor I could easily pick a wine that would not appeal to me. I was also not able to do as much technical analysis since other examples I had were different datasets that could easily be used in different methods. Lastly, it stands that the higher the wine rating, the higher the wine price. Even though the wine that rated the highest was not the most expensive wine it was still 681.37 euros. With all of this in mind I think I’ll stick with the sparkling wine I get from the grocery store to make mimosas.

**10 Questions**

1. Do you think the data will be skewed since some countries and regions had a small number of ratings versus others?
2. What criteria are you using to consider the best? Is it rating? Price? Combination?
3. Is it possible that events during the wine vintage year would influence the overall rating?
4. Do ratings by wine type differ by the region? For example, if Napa Valley is known for producing more Red wines would those ratings be higher?
5. Does a higher price equal a better wine?
6. Would it make more sense to find the best wine by type instead of overall?
7. Do you think flavor (sweet, dry) will influence wine rating by type?
8. Price is done in Euros, is there a way to find out average price by US dollar?
9. What are the odds that you end up trying the wine that is overall best wine or best by types?
10. How has this project changed your perception of wine?

# References

Augur, H. (2016, March 14). *The Perfect Pairing: Machine Learning and Wine*. Retrieved from Dataconomy: https://dataconomy.com/2016/03/the-perfect-pairing-machine-learning-and-wine/

Budnyak, A. (2020, August 27). *Wine Rating & Price*. Retrieved from Kaggle: https://www.kaggle.com/budnyak/wine-rating-and-price

Heid, M. (2020, December 10). *The Difference Between Champagne and Sparkling Wine*. Retrieved from Food & Wine: https://www.foodandwine.com/wine/champagne-sparkling-wine/difference-between-champagne-and-sparkling-wine

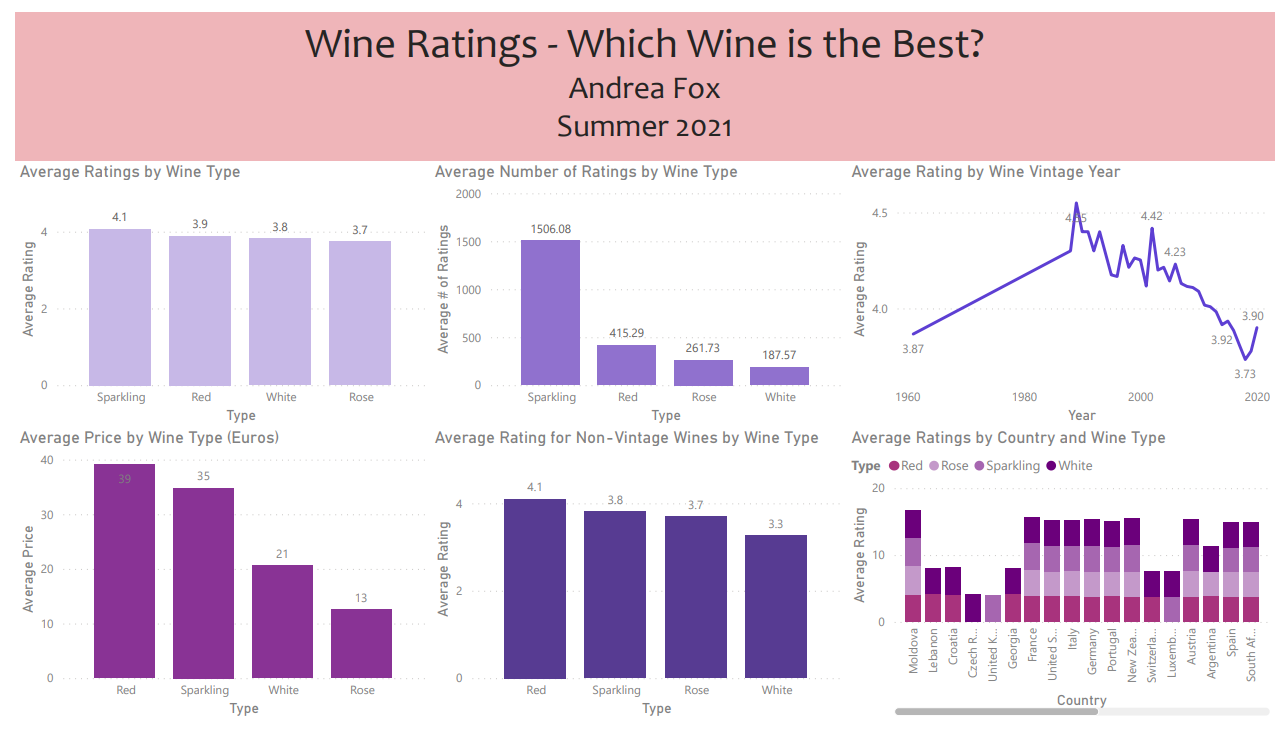
*Is Robert Parker high? Comparing wine raters*. (2017, July 19). Retrieved from Data Science of Wine: https://datascience.wine/2017/07/19/is-robert-parker-high-comparing-wine-raters/

# Appendix

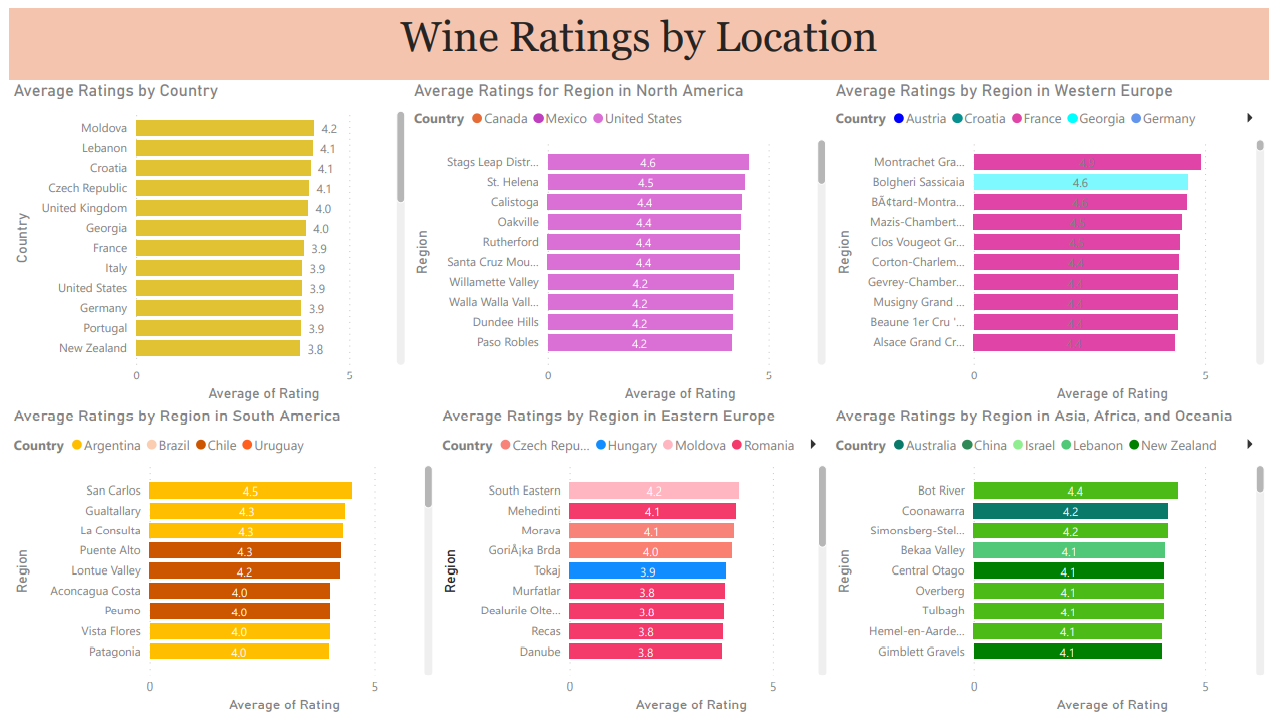
*Table 1 – Variables from Dataset* (Budnyak, 2020)

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Country | Country of origin |
| Region | Origin region or province |
| Winery | Origin winery |
| Rating | Average rating of wine |
| NumberOfRatings | Number of people who rated the wine |
| Price | Price of wine in Euros |
| Year | Vintage year |

*Figure 1 – Overview Dashboard*



*Figure 2 – Wine Ratings by Location dashboard*

**

*Figure 3 – Country Specifics*

**