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. 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

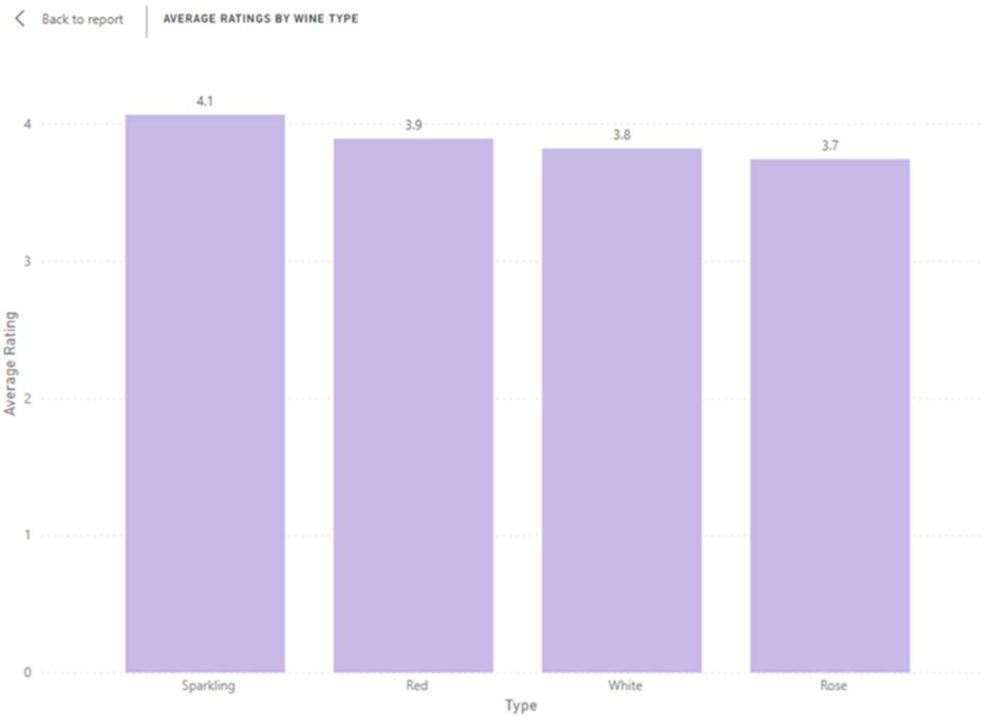
The first thing I did was go to Kaggle and download the four different csv files I intended to use. I then read all four into my Jupyter notebook. I started with Rose since it looked to take up the smallest amount of storage and worked my way to red wines. For each file I printed them out to look at 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

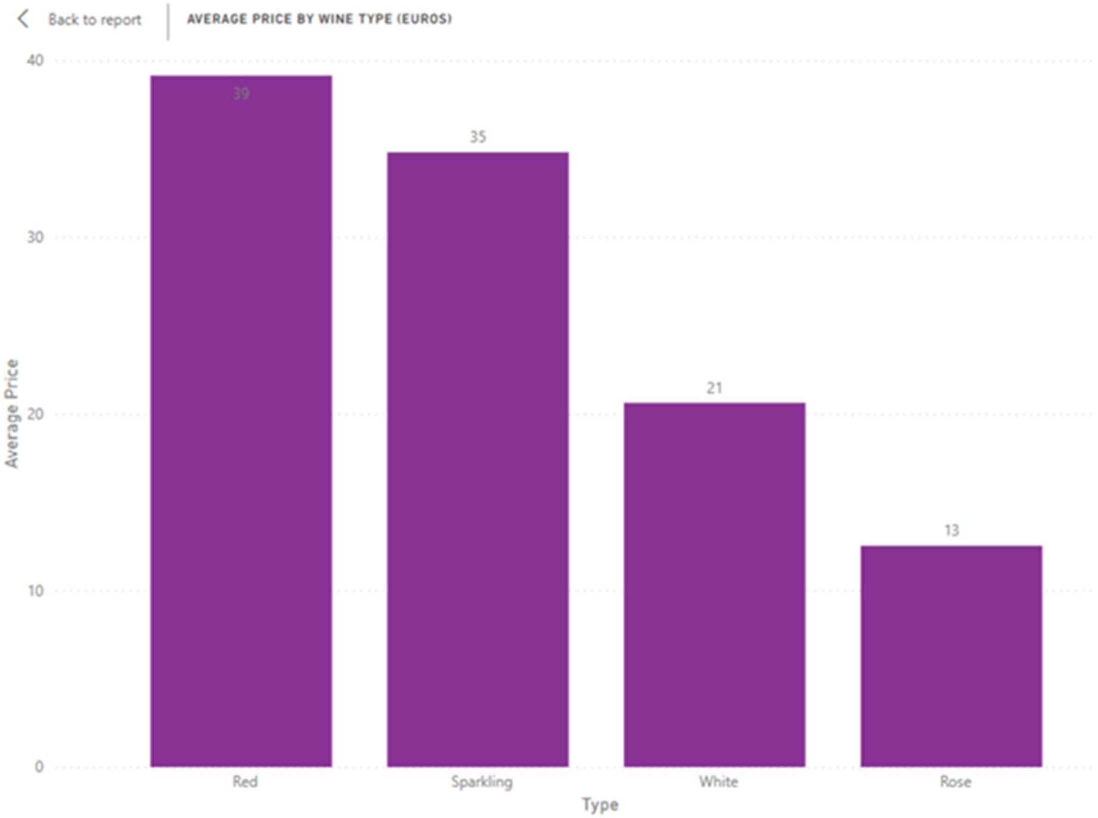
For my visualizations I used PowerBi as I find it the most intuitive tool to use when it comes to creating charts. I created two dashboards one named Overview and the second one named Location. On the Overview dashboard I really spent time looking at rating by different areas as well as price and number 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)

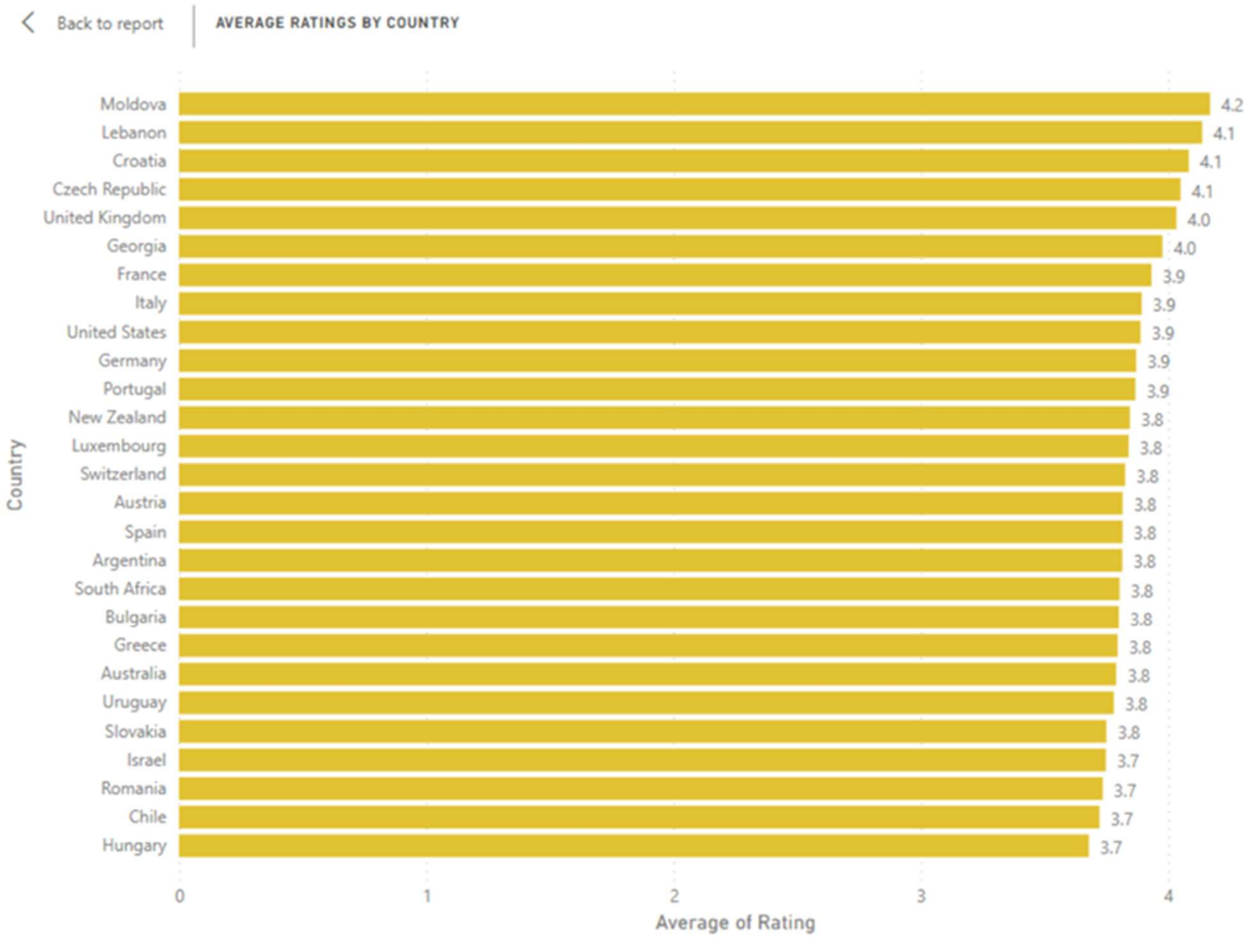


One thing I did not consider here is what is the difference between Sparkling wine and Champagne. I know in my local grocery store champagne is stored under the sparkling wine sign, so I had to go look this up since I know champagne can be somewhat pricey. I found an article that explained that champagne 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 really 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. The winery that had the highest average rating is Masseto which is located in Italy.

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. For the rest of this dashboard, I 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.

Technical Analysis

Not much to mention here as I have had to rethink my technical analysis strategy. I need to go through my references and see how it was done on similar datasets to see if it can be done here or if I will have to follow a similar method to my last project.

Conclusion

I still have work to do to complete this project. So far though I have had a few surprises in ratings by location as well as by price and number of ratings. With my visual analysis I have not yet clearly identified what I think will be the best wine for me to try, but hopefully with some of the technical analysis I can come to a more solid answer.

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

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

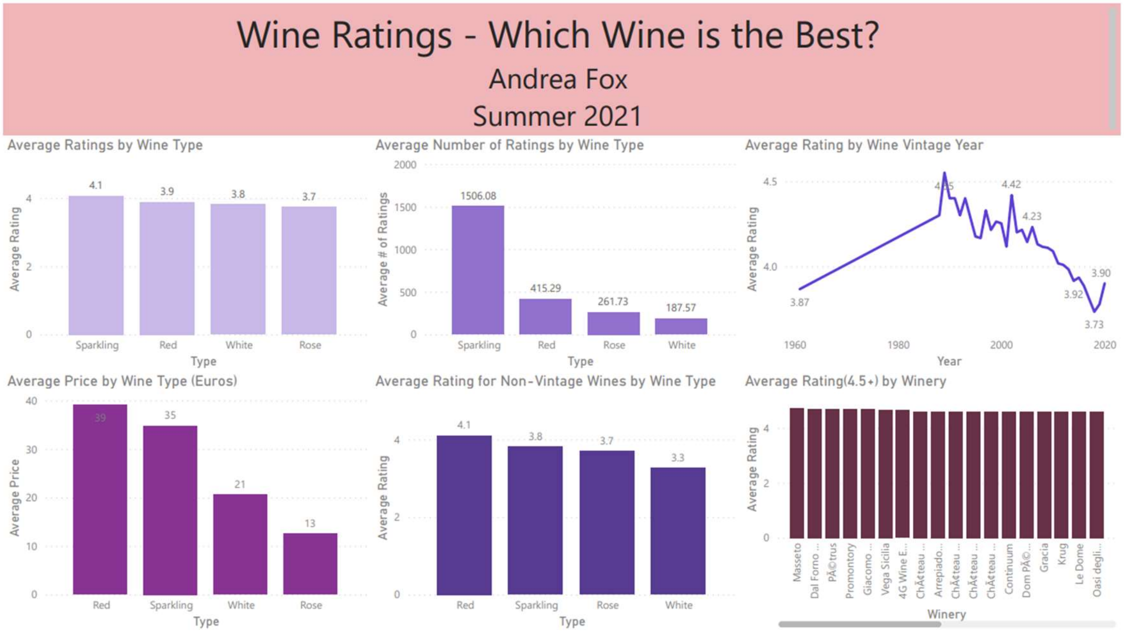
Heid, M. (2020, December 10). *The Difference Between Champagne and Sparkling Wine*. Retrieved from Food & Wine: https://www.foodandwine.com/wine/champagne-sparkling-wine/difference-between-champagne-and-sparkling-wine

Appendix

Table 1 – Variables from Dataset (Budnyak, 2020)

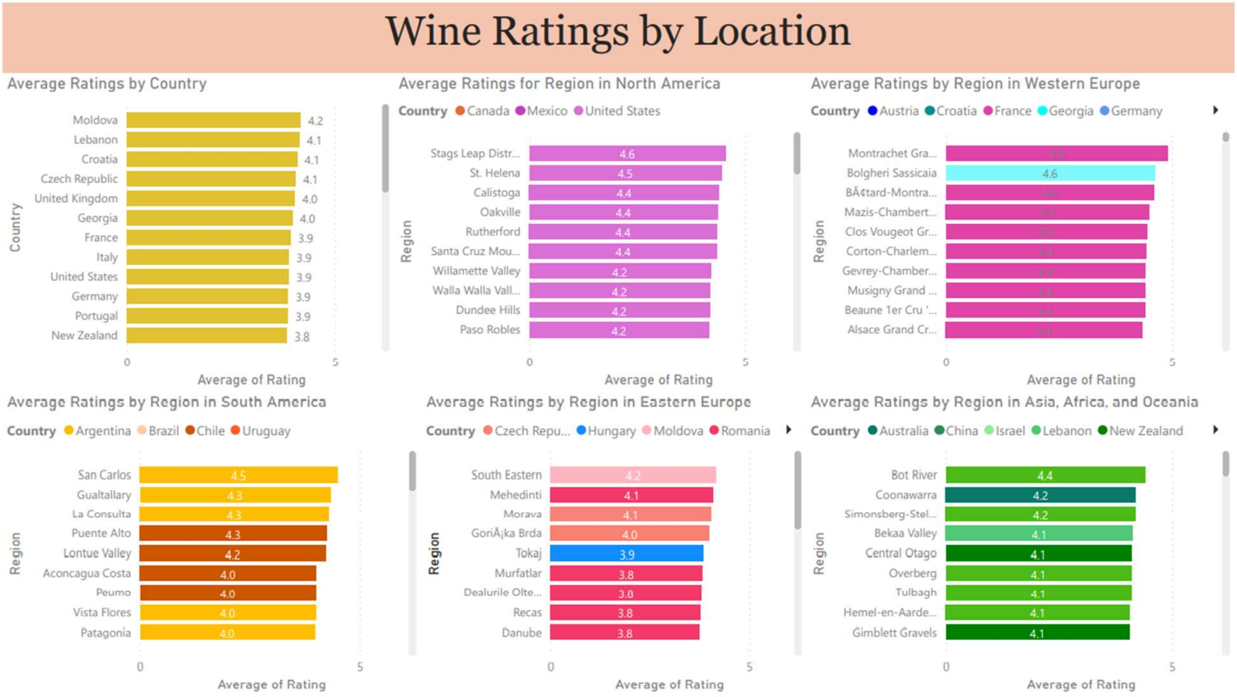
Variable	**Description**
Country	Country of orgin
Region	Origin region or province
Winery	Origin winery
Rating	Average rating of wine
NumberOfRatings	Number of people who rated the wine
Price	Price of wine in Euros
Year	Vintage year

Figure 1 – Overview Dashboard



Wine Ratings - Which Wine is the Best?
Andrea Fox
Summer 2021

Figure 2 – Wine Ratings by Location dashboard


Wine Ratings by Location
Average Ratings by Country
Country
Moldova 4.2
Lebanon 4.1
Croatia 4.1
Czech Republic 4.1
United Kingdom 4.0
Georgia 4.0
France 3.9
Italy 3.9
United States 3.9
Germany 3.9
Portugal 3.9
New Zealand 3.8
Average of Rating
Average Ratings for Region in North America
Country ● Canada ● Mexico ● United States
Region
Stags Leap Distr... 4.6
St. Helena 4.5
Calistoga 4.4
Oakville 4.4
Rutherford 4.4
Santa Cruz Mou... 4.4
Willamette Valley 4.2
Walla Walla Vall... 4.2
Dundee Hills 4.2
Paso Robles 4.2
Average of Rating
Average Ratings by Region in Western Europe
Country ● Austria ● Croatia ● France ● Georgia ● Germany
Region
Montrachet Gra... 4.9
Bolgheri Sassicaia 4.6
BĂ¢tard-Montra... 4.6
Mazis-Chambert... 4.5
Clos Vougeot Gr... 4.5
Corton-Charlem... 4.4
Gevrey-Chamber... 4.4
Musigny Grand ... 4.4
Beaune 1er Cru '... 4.4
Alsace Grand Cr... 4.4
Average of Rating
Average Ratings by Region in South America
Country ● Argentina ● Brazil ● Chile ● Uruguay
Region
San Carlos 4.5
Gualtallary 4.3
La Consulta 4.3
Puente Alto 4.3
Lontue Valley 4.2
Aconcagua Costa 4.0
Peumo 4.0
Vista Flores 4.0
Patagonia 4.0
Average of Rating
Average Ratings by Region in Eastern Europe
Country ● Czech Repu... ● Hungary ● Moldova ● Romania
Region
South Eastern 4.2
Mehedinti 4.1
Morava 4.1
GoriÅ¡ka Brda 4.0
Tokaj 3.9
Murfatlar 3.8
Dealurile Olte... 3.8
Recas 3.8
Danube 3.8
Average of Rating
Average Ratings by Region in Asia, Africa, and Oceania
Country ● Australia ● China ● Israel ● Lebanon ● New Zealand
Region
Bot River 4.4
Coonawarra 4.2
Simonsberg-Stel... 4.2
Bekaa Valley 4.1
Central Otago 4.1
Overberg 4.1
Tulbagh 4.1
Hemel-en-Aarde... 4.1
Gimblett Gravels 4.1
Average of Rating