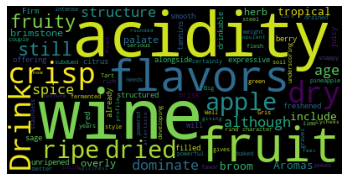
Rohini Shrivastava

HW 2: Semi-Structured Data

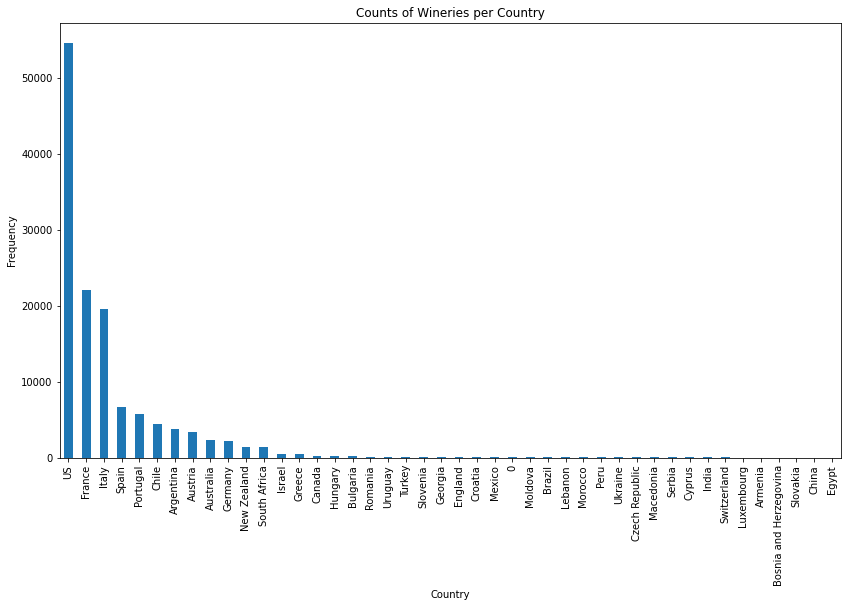
The *Wine Review*s data set was found on Kaggle. This data set contains a two csv and one json file. The data set looks at the price of the wine, location of the winery, overall customer satisfaction taken from various wineries. The data was originally scraped from WineMag.com. Overall, the goal of this is to see if there is a correlation between satisfaction and price. In Google Colab, the json file was loaded in and all N/A values were replaced with 0.

Before looking into the overall question, a few basic analyses were done. First a word cloud was generated using the description of the wine to understand what the most common themes were *(Figure 1).* Looking at this figure, it is possible to tell that most of the wines were acidic and tasted of fruit. There were various flavors, with apple being one of the fruity ones.

**

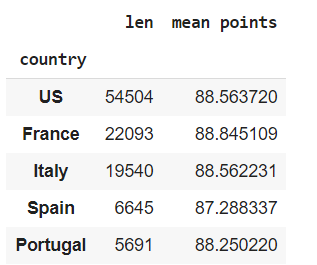
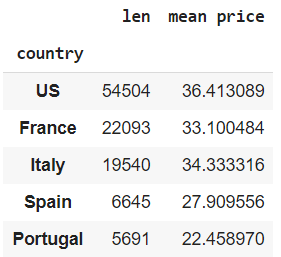
*Figure 1: Word Cloud of Wine Description*

Overall, the average price of wine was $32.91. and the average satisfaction was 88.45. According to Figure 2, the most wineries visited were in the US at a value of over 55,000. The second most was France which was closer to 24,000 and then in third place was Italy at 20,000.



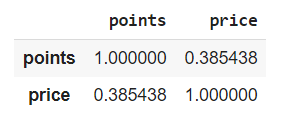
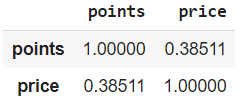
*Figure 2: Counts of Wineries per Country*

Looking at the average reviews per country, the top five countries with the most wine tasting had around the same satisfaction rate. Spain was the highest at 87.29, while the rest of the four countries were at 88. (*Figure 3).* The average price of wine was also looked at for the same five countries. The most expensive was the US at $36.41 and the lease was Portugal at $22.45.

*Figure 3: Country vs Average Review and Price*

Finally, a correlation was run between points (satisfaction) and price of the wine. Two methods were used: Pearson (*Figure 4)* and Kendall (*Figure 5)*. Overall, there was no correlation between satisfaction and price. Both correlation methods had a correlation of 0.385, which shows there is not much of a positive correlation.

*Figure 4: Pearson Method for correlation Figure 5: Kendall Method for correlation*

While there was no correlation between price and points, in the future, it would be interesting to create a regression model to see if country and variety of the wine can help in predicting the satisfaction consumers have. There may be some inherent bias in the models dependent on location however, as most of the reviews are collected from the United States. It would be best to have about the same number of reviews per country when looking into the regression modeling.