EXECUTIVE SUMMARY OF WINE QUALITY DATA ANALYSIS

Brief information on the dataset

The dataset provides information on some of the chemical properties of wine produced in Portuguese which include wine quality rating and wine color. The dataset has a total of 6497 records. The variables that were identified in the dataset includes: fixed-acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality and color. These variables were measured to see if they could help predict the quality of the wine. The analysis will be more descriptive and might not be strong enough to conclude if these variables are good predictors or identify the best predictor. However, the result might be a good starting point to show the factors that could probably affect or predict wine quality.

Questions

The following questions were answered to provide insight as the quality of the wine.

1. Describe the distribution of the wine quality based on the ratings (quality data).
2. Does the wine quality differs by the five main factors: color, alcohol content, pH value, density and residual sugar.
3. It there any correlation between wine quality and all the chemical properties?

Results

1. The wine quality data is normally distributed with a mean value of 5.82 while the wine quality rating with the highest frequency is 6 (See Figure 1). The standard deviation (0.873) also indicates that the wine quality rating is distributed around the mean value. Analysis also showed that the maximum rating is 9, indicating that none of the wine received the highest possible rating (10). Furthermore, it was revealed that majority (63.3%) of the wine received a rating that is higher than the mean value which is suggests that the wines tend to receive a fair quality rating.

2. Considering the result of the analysis in categorizing the factors (color, alcohol content, pH value, density and residual sugar). The factors that record a noticeable difference in wine quality based on their categories were wine alcohol content and density (See Figure 2a). Wines that have high alcoholic content tend to receive better quality rating compared to those with low alcohol content. For the density, wine with low density had better quality rating compared to those with high density. This suggests that wine that are light and that has high alcoholic content tend to be preferred. No striking difference in wine quality rating was observed considering other factors (color, residual sugar, and pH value).

3. The correlation result revealed: fixed acid(-0.076), volatile acid (-0.265), citric acid (0.085), residual sugar (-0.036), chlorides (-0.201), free sulfur (0.055), total sulfur (0.041), density(-0.305), pH(0.019), sulphates (0.038), alcohol(0.444). The result showed that alcohol has the highest correlation followed by density, volatile acid and chloride, in that order. The correlation between wine quality and other variables is negligible. Furthermore, it was observed that the correlation between alcohol content and wine quality is positive (the higher the alcohol content, the higher the wine quality rating), while the correlation between (density, volatile acid, chloride) and wine quality is negative (the quality of wine tends to increase as the variables decreases).

Conclusion

Generally speaking the wines quality received a fair rating with an average rating of 5.8, and 6 being the most frequent rating. The wine chemical properties that are most likely good predictors of the wine quality are mainly its alcoholic content and the wine density: the customers tend to have a good taste for light alcoholic drinks. Other chemical properties to also watch out for are the volatile acid (smell and taste of vinegar) and chloride (saltiness). However, ascertaining the predictor(s) and level of determination is beyond scope of this analysis.

VISUALIZATIONS

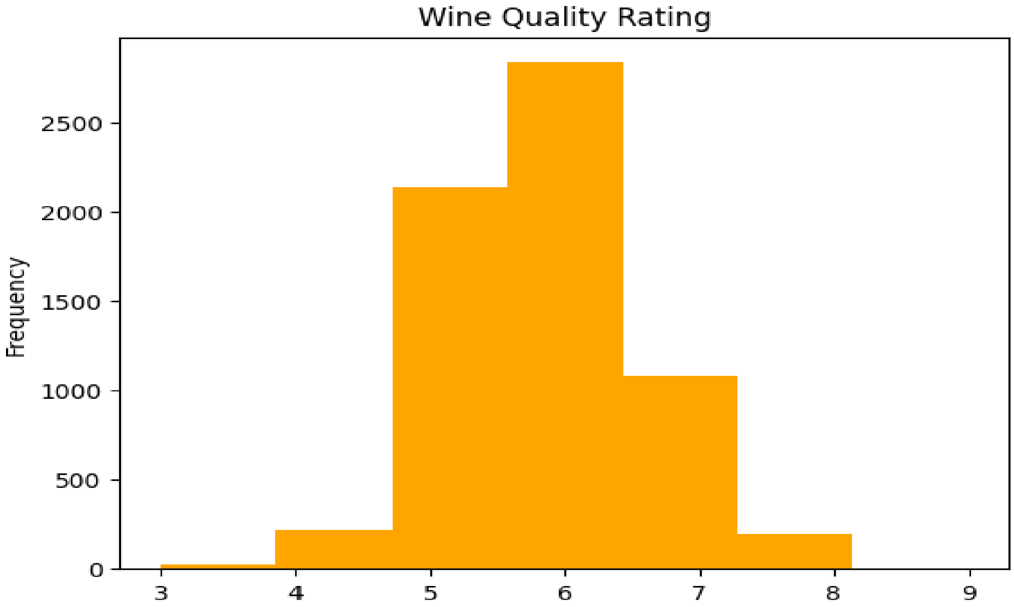

Figure 1: Wine Quality Rating

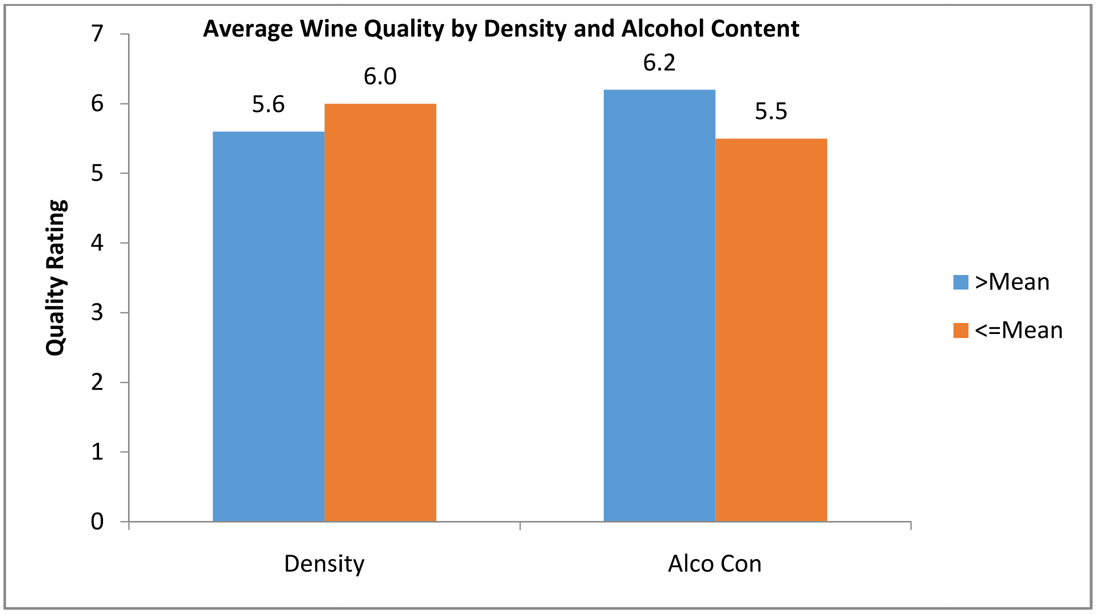

Figure 2: Wine Quality Rating by Density and Alcohol Content