Capstone Project 1 Milestone Report:

Which physicochemical properties make an excellent quality wine?

There are over 10,000 species of wine grapes in the world. This large variety of options could be overwhelming for a small start up winery. Sourcing grape species that produce quality wines could be the difference between success and failure. For this reason, my theoretical client is a vineyard who wishes to identify a species of grapes that has the physicochemical properties that result in a quality wine. My analysis will show them the properties that are prevalent in a quality wine and will allow them to analyse which species exhibits these qualities in their own wines. Additionally, it will give them insight by classifying which physicochemical properties make a quality wine.

In my analysis, I will use data that comes from the machine learning repository wine quality dataset which has 1599 red wines and 4898 white wines. In the data set both the red and the white wines are from the grape species vitis vinifera and are from the Vihno Verde area of Portugal. This area has over 30 species of vitis vinifera which is the grape used in all fine wines according to their tourism website. The data set has no missing values and 13 attributes. 12 of these are physicochemical properties such as alcohol, density, pH, etc. The last is the output classification variable ‘quality’ which is a discrete value between 1 and 10.

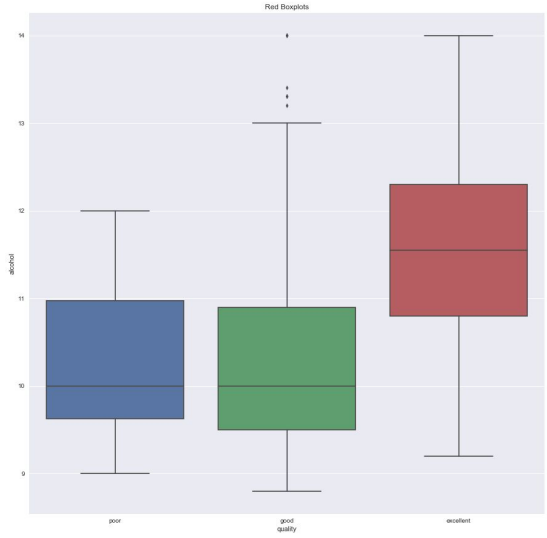
Solving this problem will require Logistic Regression or Classification to connect which physicochemical properties are associated with quality wine in both the red and white varieties. This classification will identify which features impact wine quality by classifying qualities within the discrete output observations.

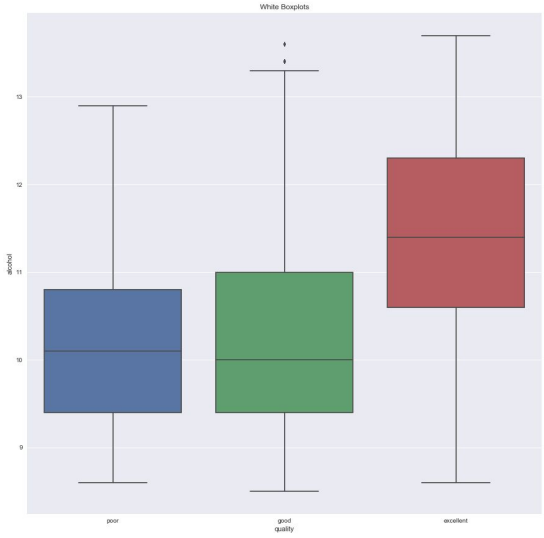
Data Wrangling:

Both the red and white data frames contained no missing values. The 11 observed physicochemical properties are represented by floats and the discrete categorical variable quality uses a 1-10 scale. In the data wrangling process, I first created a train test split. I then grouped the discrete categorical variable quality into three values: poor (values 1-4), good (values 5,6), and excellent (values 7-10). Last, I removed the outliers that were outside two standard deviations from the mean for each of the 11 observed properties. After removing these outliers the white train dataframe had 3675 observation and the red train dataframe had 1,199 observations.

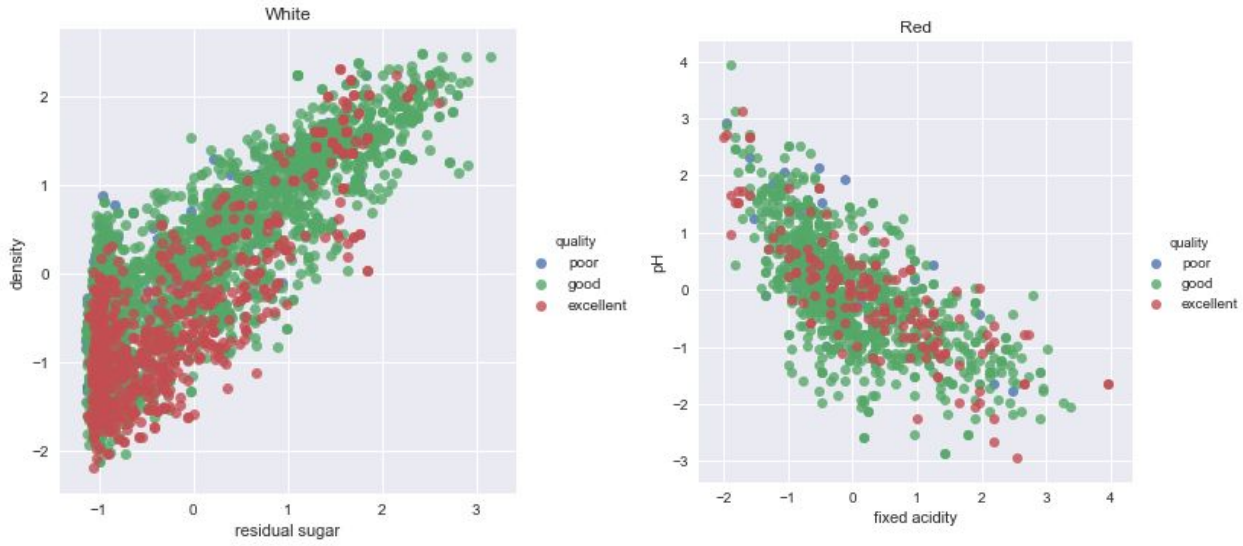
Data Story:

Higher alcohol seems to be a prevalent feature of excellent wines in both red and white varieties. Additionally, there appears to be differences in means and spread between good and excellent quality wines in many other physicochemical properties as well. To know if this difference in means in significant null hypothesis testing will be needed. The below box plot illustrates the the difference in shape and spread between good and excellent wines in their mean alcohol content. Again this evidence supports the idea that alcohol content has an influence on a quality wine.





In addition to these differences in means, several variables appear to be correlated. In both red an white wines, alcohol and density appear to inversely correlated. Residual sugar and density also seem to be correlated in white wines. The below scatter plots show the normalized correlation between alcohol and density in white wines and fixed acidity and pH in red wines. Again, inferential statistics will be needed to make definitive answers regarding the influence these factors have on the output variable quality.



Inferential Statistics:

Strong correlations exist in white wines between density and alcohol with an R-Squared value of 0.65 and density and residual sugar with an R-squared value of 0.7. No strong correlations can be found in the physicochemical properties of red although ph and fixed acidity have a R-Squared of 0.49, as well as citric acid and fixed acidity have an R-Squared value of 0.47.

After grouping the wines into good and excellent quality groups, I then performed null hypothesis z-tests between the groups on the physicochemical properties with an alpha of 3/22000. This small alpha was a reaction to the number of input variables in my dataframes and I believe it helped reveal interesting insights into the data. First, the only three properties whose null hypothesis was rejected in common for both red and white wines were Density, Total Sulphur Dioxide, and Alcohol. This may suggest there are few commonalities between quality in Red and White wines besides these three input variables. Additionally, the null hypothesis test showed that there are physicochemical properties that do not appear to have an affect on the output variable quality. These properties may be excess information that we can remove to simplify the model when using Logistic Regression.