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.

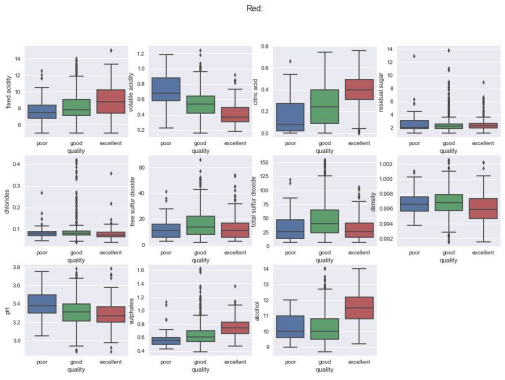
Solving this problem will require classification to connect which physicochemical properties are associated with quality wine in both the red and white varieties. This classification will identify which features impact wine quality by classifying qualities within the discrete output observations.

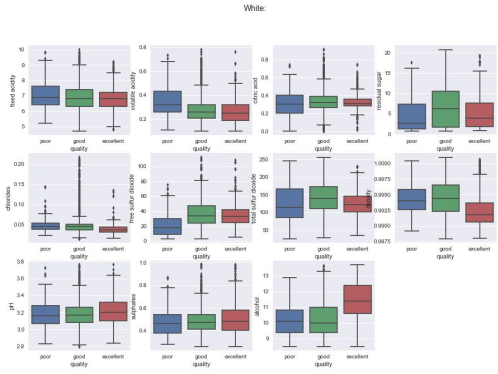
Data Wrangling:

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 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 dataframe had 2,935 observation and the red dataframe had 1,012 observations. I then grouped the discrete categorical variable quality into three values: poor (values 1-4), good (values 5,6), and excellent (values 7-10).

Data Story:

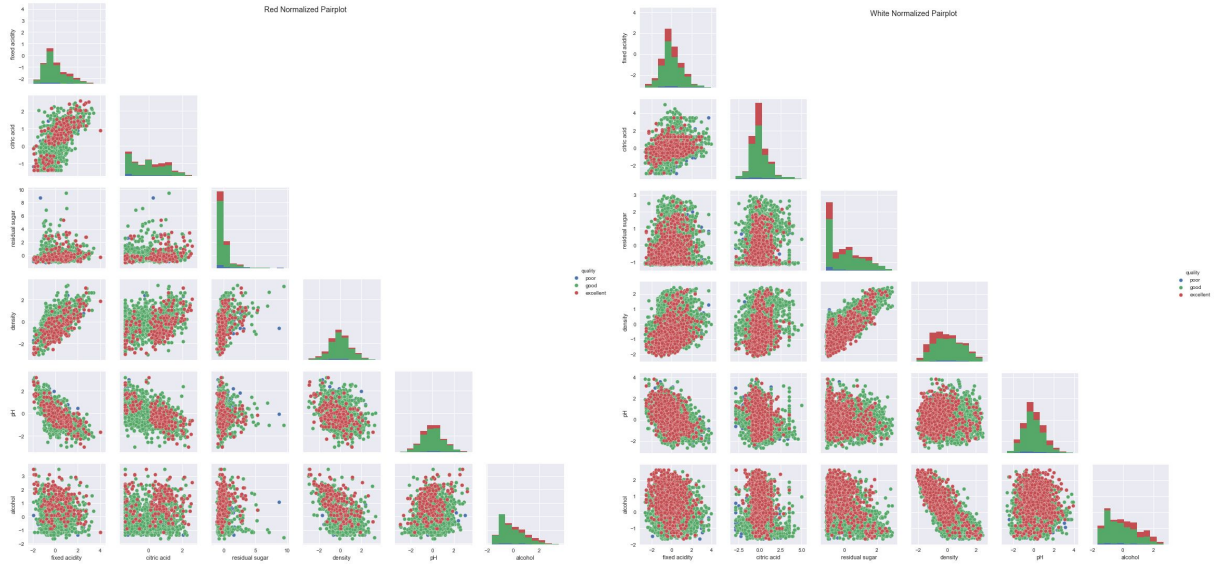
Higher alcohol seems to be a prevalent feature of excellent wines in both red and white varieties. The below boxplots show that there may be differences in mean and spread between good and excellent quality wines in many other physicochemical properties as well. To know if this difference in means in significant null hypothesis testing will be needed.





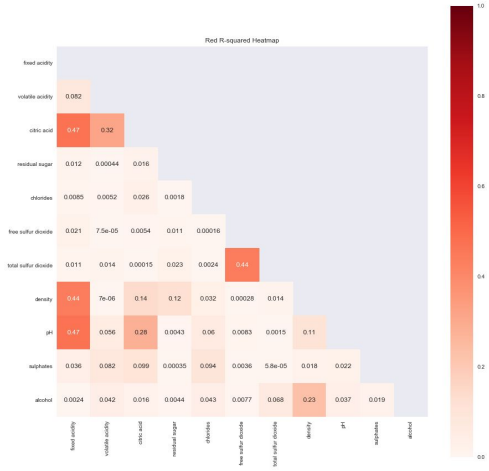
In addition to these differences in means, several variables appear to be correlated as can be seen in the pairplots below. In both red an white wines, alcohol

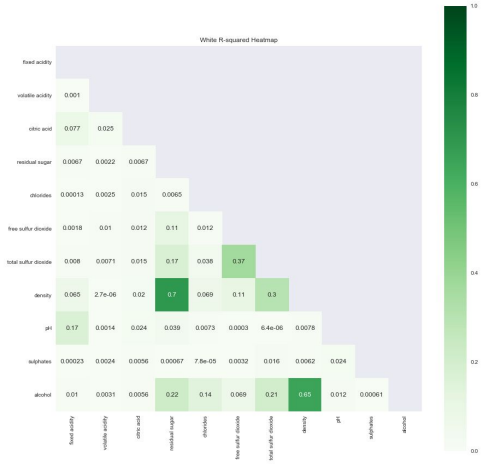
and density appear to inversely correlated. Residual sugar and density also seem to be correlated in white wines. Areas with concentraded red colored observations show that residual sugar may be an impactful property in red wines; conversely, citric acid may be an impactful property in white wine. Again, inferential statistics will be needed to make definitive answers regarding the influnce 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 .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 0.47, as well as citric acid and fixed acidity have an r dsquared value of 0.47. The rest of both heatmaps is relatively colorless showing that most physicochemical properties are independent of each other.





After grouping the wines into good and excellent quality groups and then performing null hypothesis tests between the groups on the physicochemical properties some interesting ideas arose. First, the only accepted null hypothesis in white wines is citric acid and in red wines the two accepted null hypothesis are residual sugar (just barely) and ph. Interestingly, citric acid in white wines had the same mean for both the excellent and good quality groups but fixed acidity rejected the null hypothesis and showed axcellent and quality wines. Citric acids are part of what makes up the the total acidity in wine so this may mean one of the other fixed acids, tartaric, malic, or succinic, has a greater influence on white wine quality than citric (http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity). Additionally, fixed acids and citric acid are not highly correlated in white wines so perhaps citric acids are even less significantly related to white wine quality. Each of the rejected null hypothesis test showed extremely significant p-values in both red and white wines and displays that there are differences in means of physicochemical properties between good and

excellent quality wines. Lastly, when looking at similarities between red and white wine properties, the shared rejected null hypothesis for both types are fixed acidity, volatile acidity, chlorides, free sulfer dioxide, total sulfer dioxide, density, sulphates, and alcohol. This means that any one of these properties may influence the quality of a wine in both red and white varieties.