Myranda Swartzwelter
DSC-530 Final Project

Wine Quality Factors

As a wine lover, I can tell you about differences in grape varieties, New World vs Old World wines, and what a tannin is, but I do not have much experience with the science behind what makes good wine. For centuries, farmers have been carefully cultivating their grapes to produce the best quality wine. I was able to find a data set that has 11 different qualities of wine - from fixed and volatile acidity to alcohol content, and the scaled quality of the wine. As I did more digging into this subject, I found a paired datasets of the Portuguese Vinho Verde wine - one for the red variety of this wine and one for the white variety. I combined these datasets to answer the questions: Which of the variables are best correlated with wine quality? Does this differ between red and white wines?

The outcome of my EDA was a multivariate regression analysis which found that the top 3 correlated variables for wine quality differed slightly between red and white wines. The top 3 variables correlated with red wine quality were alcohol content, volatile acidity, and sulphates. The top 3 variables correlated with white wine quality were alcohol content, volatile acidity, and chlorides. These 3 variables for the respective wine types did a better job in a regression analysis than any single variable, or when combining the red and white wine datasets together.

I think in this analysis, one thing that was missing was a variety of the qualities of the wine. Most of the wines in the dataset were in the quality range 4-7, and the extremely low quality and extremely high-quality wines were missing from the dataset. I think having a wider range of quality would have led to stronger conclusions about the variables that impact wine quality.

Many of the variables in the dataset were related, so even though at first, I thought the different variables would be full of insight, I had to exclude some of the 11 because of multicollinearity. However, I think adding the grape varietal to the dataset could have helped with the analysis, as I believe that certain varietals are associated with higher-quality wine. I would be curious to see if that is true.

At first, I thought I could assume that red wine and white wine quality would be due to the same factors, so I thought I could combine the dataset together and only run the analyses on one dataset. As I went on, I discovered that there were different correlations between red and white wines, so I ended up re-separating the datasets.

I faced the challenge of multicollinearity, which I did not fully understand what to do with. I had to circle back to the analysis several times to account for the relationships between my dependent variables. For the next analysis, I would recommend further delving into the multicollinearity present in this dataset to resolve it and make better models for the prediction of wine quality.