DS 219 Report

This analysis delves into two datasets, unraveling the component characteristics of both red and white variants of a Portuguese wine company, Vinho Verde wines. An array of key attributes were recorded, ranging from "fixed acidity" to the "alcohol" content; these datasets provide a lens through which we can unravel the distinct characteristics of these renowned wines.
The question at hand is which component affects the quality of wine the most. I hypothesize that alcohol concentration would affect the quality the most, as it is a substance that is very prominent on taste buds. The key methodology to this analysis is the comparison of r-squares. I will be utilizing Python with numpy and discuss the approach, key findings, and conclusions within this paper.
The initial approach was to pick out key features that I felt had the most influence. That was residual sugar, fixed acidity, and alcohol. I then created graphs, plotting the data with quality as my dependent variable. Utilizing seaborn to do a linear map plotting, for the red wines, though originally positive, as residual sugar increased past a certain point, it appeared that the quality went down. For the white wines, there was a linear decrease in quality as residual sugar went up. With the white wine, there were definitely indications of outliers that caused such a drastic change, but without those outliers, there wasn't enough variance in residual sugar to conclude direct association.
The next feature plotted was fixed acidity. For red wines, as fixed acidity went up, there seemed to be an increase in quality ratings, and for white wine, as fixed acidity went uo, there seemed to be a decrease at an exponential rate.
The last feature plotted from the original assumptions was alcohol. For both red and white wine, as alcohol went up, quality ratings went up. From plotting alone, alcohol and fixed acidity seemed to be the two features with the most effects from my hypothesis. To test that, I looked into the r-sqaures.
Using numpy, to discover r-square values, the results appeared for fixed acidity as red wine having an r-square of 15% and for white wine 13%. For alcohol in red wine, 23% and alcohol in white wine, 19%. Realizing there may have been a stronger feature that predicted quality, reanalysis of my data through summary statistics showed Sulphate to have the highest variance with a maximum of 2 and a minimum of .33 and an average of .65. The r-square value was 63%. This indicates that Sulphate alone may have the strongest correlation to alcohol quality. Plotting Sulphate, the spread was varying, with minimal outliers.
This led me to conclude that Sulphate concentration is positively associated with quality of alcohol. With research, Sulfur Dioxide (SO2), occurs naturally in the fermentation process to some extent, and is commonly added to wines as a preservative and antioxidant. This allows for winemakers to produce wine that can age gracefully without experiencing premature aging or spoilage.