Hadiza Mamman

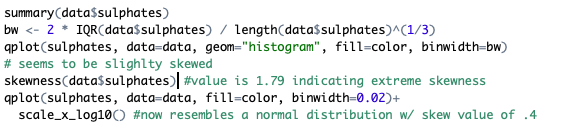
Wine Quality Score – Predictive Modeling

**Abstract**

Red and white wine from the Vinho Verde region of Portugal were given overall assessment scores, averaged from professional wine tasters, based on several chemical tests. These chemical tests included the content of fixed acidity, volatile acidity, citric acid, residual sugar, chloride, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol in each wine.

**Pre-processing**

Once the dataset was loaded into the R script, I created a histogram for each variable, in order to identify different attributes, to better understand the data. Most of the variables resembled a normal distribution, with the exception of alcohol, free sulfur dioxide, sulphate, chloride, residual sugar, and fixed acidity. Subsequently, I performed a log transformation to combat skewness.

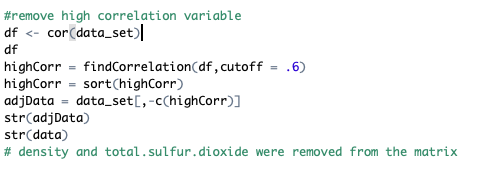
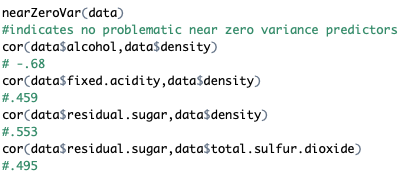




Another noticeable finding was the significant difference in levels of fixed acidity, and total sulfur dioxide between red and white wine.

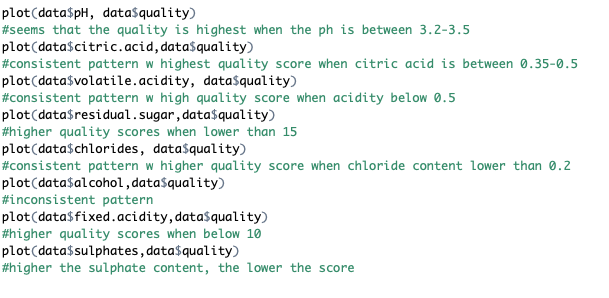


After gathering conclusions on the preliminary data, I created a correlation matrix to see which variables exuded collinearity, possibly indicating redundant information. I also used two different correlation functions to double check collinearity between variables.



The findCorrelation function identified density and total sulfur dioxide as the variables with the most correlation. Thus, the two variables were ruled out for the model and were removed from the dataset.

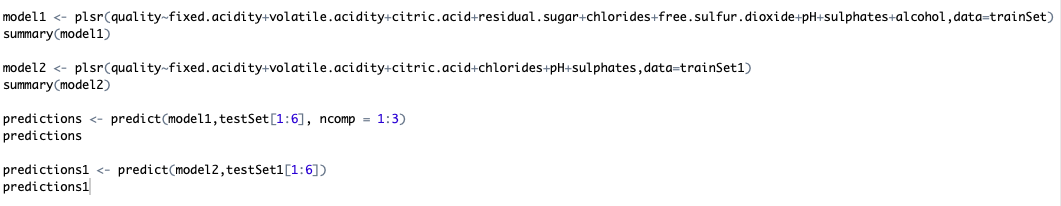
I performed diagnostic tests and created more histograms with the remaining variables to analyze more patterns and attributes in the modified dataset.



These diagnostic tests offered insight on possible variables that can be included in a model. Citric acid, volatile acidity, residual sugar, chloride, sulphate, and fixed acidity appeared to affect the quality scores the most.

**Data splitting & model building**

In order to make the partial least squares model more stable, I center and scaled the modified dataset before splitting it into the train and test sets. Once the splits were made, I created two models based on my conclusions. The first model included all of the variables that did not display collinearity. The second model included only the variables that I believed affected the quality score the most, based on the diagnostic tests.



**Conclusion**

After comparing both models and their predictions on the test set, I believed the first model was significantly more accurate than the second model. Subsequently, it did not support my hypotheses on certain variables substantially affecting the quality score. To confirm my model, I checked the residuals, and the RMSE of the training and test sets.

