## Introduction

The following analysis is a prediction of wine quality based on 11 physiochemical test measures and grape type. The data used includes 6497 wines which are all variants of the Portuguese "Vinho Verde". Each wine has been given a quality rating on a scale from one to ten (worst to best); this variable will be changed to a binary response of poor (1-5) or fine (6-10) for the purpose of this project. All physiochemical measures are represented as continuous variables and are as follows: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. There is an additional categorical variable for grape type where 0 represents red grapes and 1 represents white grapes. The objective of this analysis is to examine the relationships between the predictor variables and wine quality in the hope to identify important measures that can predict a wine’s quality so that the subjective assessment of wine quality and pricing can become more controlled. The ultimate goal is to develop an objective model using some or all of these variables to predict wine quality.

## Exploratory Analysis

This analysis begins by first looking at histograms (Appendix A) of the variables to assess for any deviations in normality or potential outliers in the data. Of utmost concern is the response variable as it a discrete variable based on an ordinal ranking. Looking at this plot, it can be seen that the full range is [3, 9] where the majority of wine rankings fall between the 5 to 7 range and that the data do not appear to deviate far from a normal distribution. Add in chart for binary response and description.

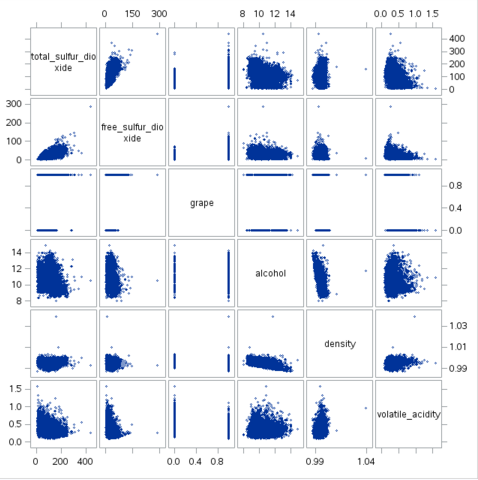
Many of the explanatory variables appear to have large outliers which means transformations might be needed with model fitting. Fixed acidity, volatile acidity, and citric acidity appear to have a relatively normal and symmetric distribution but have large outliers; chlorides, free sulfur dioxide, and total sulfur dioxide have similar characteristics but are not as symmetric.

Residual sugar is skewed right which could indicate the need for a logarithmic transformation. Density and alcohol have irregular shapes but do not appear to have outliers. pH and sulfates show signs of symmetry but both have a right skew; neither appear to have outliers.

Now that the individual variables have been examined, the relationships between the predictor variables themselves and response variable need to be examined. This can be done by examining scatterplots and correlations in order to determine data patterns and assure there is not multicollinearity within the model.

In Appendix B, Pearson correlations between all variables are shown. Any correlation larger than 0.4 has been highlighted as this is indicative of a high correlation; as it can be seen there are many correlations which are higher than this threshold. This needs to be considered when the model is built as any underlying linear patterns between the predictor variable will cause redundancy within the model and potentially inflate model performance.

The largest of these correlations are between total sulfur dioxide and free sulfur dioxide at 0.72093, total sulfur dioxide and grape at 0.70036, alcohol and density at -0.68675, and grape and volatile acidity at -0.65304 (this likely okay to include since one of the variables is categorical and a linear relationship is very unlikely); in fact the all response variable combinations with high correlation include at least one of the previously mentioned variables. It is likely than any combination of these variables within a model will cause fit issues including inflation.

Additionally, the scatterplot between these variables show a linear trend between total sulfur dioxide and free sulfur dioxide and positive fanning patterns between both total sulfur dioxide and volatile acidity and free sulfur dioxide and volatile acidity. All of this evidence shows that including both total sulfur dioxide and free sulfur dioxide in the model will cause inflation. Of note, a scatter plot of all variables was not included due to SAS limitations.

When looking for any high correlation between the non-transformed response variable and a predictor variable the only high correlation seen is between alcohol and quality at 0.44432. There are also a couple moderate correlations between density and quality at -0.30586 and volatile acidity and quality at -0.26570.

## Appendix A (Histograms)





## Appendix B (Correlation Coefficients)

| **Pearson Correlation Coefficients, N = 6497 Prob > |r| under H0: Rho=0** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **fixed\_acidity** | **volatile\_acidity** | **citric\_acid** | **residual\_sugar** | **chlorides** | **free\_sulfur\_dioxide** |
| **fixed\_acidity** | 1.00000 | 0.21901 <.0001 | 0.32444 <.0001 | -0.11198 <.0001 | 0.29819 <.0001 | -0.28274 <.0001 |
| **volatile\_acidity** | 0.21901 <.0001 | 1.00000 | -0.37798 <.0001 | -0.19601 <.0001 | 0.37712 <.0001 | -0.35256 <.0001 |
| **citric\_acid** | 0.32444 <.0001 | -0.37798 <.0001 | 1.00000 | 0.14245 <.0001 | 0.03900 0.0017 | 0.13313 <.0001 |
| **residual\_sugar** | -0.11198 <.0001 | -0.19601 <.0001 | 0.14245 <.0001 | 1.00000 | -0.12894 <.0001 | 0.40287 <.0001 |
| **chlorides** | 0.29819 <.0001 | 0.37712 <.0001 | 0.03900 0.0017 | -0.12894 <.0001 | 1.00000 | -0.19504 <.0001 |
| **free\_sulfur\_dioxide** | -0.28274 <.0001 | -0.35256 <.0001 | 0.13313 <.0001 | 0.40287 <.0001 | -0.19504 <.0001 | 1.00000 |
| **total\_sulfur\_dioxide** | -0.32905 <.0001 | -0.41448 <.0001 | 0.19524 <.0001 | 0.49548 <.0001 | -0.27963 <.0001 | 0.72093 <.0001 |
| **density** | 0.45891 <.0001 | 0.27130 <.0001 | 0.09615 <.0001 | 0.55252 <.0001 | 0.36261 <.0001 | 0.02572 0.0382 |
| **pH** | -0.25270 <.0001 | 0.26145 <.0001 | -0.32981 <.0001 | -0.26732 <.0001 | 0.04471 0.0003 | -0.14585 <.0001 |
| **sulphates** | 0.29957 <.0001 | 0.22598 <.0001 | 0.05620 <.0001 | -0.18593 <.0001 | 0.39559 <.0001 | -0.18846 <.0001 |
| **alcohol** | -0.09545 <.0001 | -0.03764 0.0024 | -0.01049 0.3977 | -0.35941 <.0001 | -0.25692 <.0001 | -0.17984 <.0001 |
| **quality** | -0.07674 <.0001 | -0.26570 <.0001 | 0.08553 <.0001 | -0.03698 0.0029 | -0.20067 <.0001 | 0.05546 <.0001 |
| **grape** | -0.48674 <.0001 | -0.65304 <.0001 | 0.18740 <.0001 | 0.34882 <.0001 | -0.51268 <.0001 | 0.47164 <.0001 |

| **Pearson Correlation Coefficients, N = 6497 Prob > |r| under H0: Rho=0** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **total\_sulfur\_dioxide** | **density** | **pH** | **sulphates** | **alcohol** | **quality** | **grape** |
| **fixed\_acidity** | -0.32905 <.0001 | 0.45891 <.0001 | -0.25270 <.0001 | 0.29957 <.0001 | -0.09545 <.0001 | -0.07674 <.0001 | -0.48674 <.0001 |
| **volatile\_acidity** | -0.41448 <.0001 | 0.27130 <.0001 | 0.26145 <.00010.47164 | 0.22598 <.0001 | -0.03764 0.0024 | -0.26570 <.0001 | -0.65304 <.0001 |
| **citric\_acid** | 0.19524 <.0001 | 0.09615 <.0001 | -0.32981 <.0001 | 0.05620 <.0001 | -0.01049 0.3977 | 0.08553 <.0001 | 0.18740 <.0001 |
| **residual\_sugar** | 0.49548 <.0001 | 0.55252 <.0001 | -0.26732 <.0001 | -0.18593 <.0001 | -0.35941 <.0001 | -0.03698 0.0029 | 0.34882 <.0001 |
| **chlorides** | -0.27963 <.0001 | 0.36261 <.0001 | 0.04471 0.0003 | 0.39559 <.0001 | -0.25692 <.0001 | -0.20067 <.0001 | -0.51268 <.0001 |
| **free\_sulfur\_dioxide** | 0.72093 <.0001 | 0.02572 0.0382 | -0.14585 <.0001 | -0.18846 <.0001 | -0.17984 <.0001 | 0.05546 <.0001 | <.0001 |
| **total\_sulfur\_dioxide** | 1.00000 | 0.03239 0.0090 | -0.23841 <.0001 | -0.27573 <.0001 | -0.26574 <.0001 | -0.04139 0.0008 | 0.70036 <.0001 |
| **density** | 0.03239 0.0090 | 1.00000 | 0.01169 0.3463 | 0.25948 <.0001 | -0.68675 <.0001 | -0.30586 <.0001 | -0.39065 <.0001 |
| **pH** | -0.23841 <.0001 | 0.01169 0.3463 | 1.00000 | 0.19212 <.0001 | 0.12125 <.0001 | 0.01951 0.1159 | -0.32913 <.0001 |
| **sulphates** | -0.27573 <.0001 | 0.25948 <.0001 | 0.19212 <.0001 | 1.00000 | -0.00303 0.8071 | 0.03849 0.0019 | -0.48722 <.0001 |
| **alcohol** | -0.26574 <.0001 | -0.68675 <.0001 | 0.12125 <.0001 | -0.00303 0.8071 | 1.00000 | 0.44432 <.0001 | 0.03297 0.0079 |
| **quality** | -0.04139 0.0008 | -0.30586 <.0001 | 0.01951 0.1159 | 0.03849 0.0019 | 0.44432 <.0001 | 1.00000 | 0.11932 <.0001 |
| **grape** | 0.70036 <.0001 | -0.39065 <.0001 | -0.32913 <.0001 | -0.48722 <.0001 | 0.03297 0.0079 | 0.11932 <.0001 | 1.00000 |

## Appendix (SAS Code)

data wine;

infile '/folders/myfolders/Experimental Statistics I/winequality.csv' firstobs = 2 delimiter = ';';

input fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides free\_sulfur\_dioxide total\_sulfur\_dioxide density pH sulphates alcohol quality grape;

proc univariate data = wine;

histogram;

run;

/\*close to normal distribution, slight right skew. Enough levels to consider like a cont var./\*

/\*Included in write up\*/

proc sgscatter data = wine;

matrix total\_sulfur\_dioxide free\_sulfur\_dioxide grape alcohol density volatile\_acidity;

run;

proc corr data = wine plots(maxpoints=none)=scatter;

var fixed\_acidity volatile\_acidity citric\_acid residual\_sugar chlorides free\_sulfur\_dioxide total\_sulfur\_dioxide density pH sulphates alcohol quality grape;

run;