Abstract

Wine is a complex and revered alcoholic beverage enjoyed by millions around the world. Wine quality has been an elusive concept to me ever since I began exploring different kinds of wine; what exactly makes a wine “good”. In this study I plan to examine the different contributing factors to wine production and how it in turns the quality of the wine. The main purpose of this experiment is to predict wine quality based on physicochemical data. The dataset I will be analyzing contains 4898 instances of white wine, with 11 different categories of physicochemical data which are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol.

Introduction

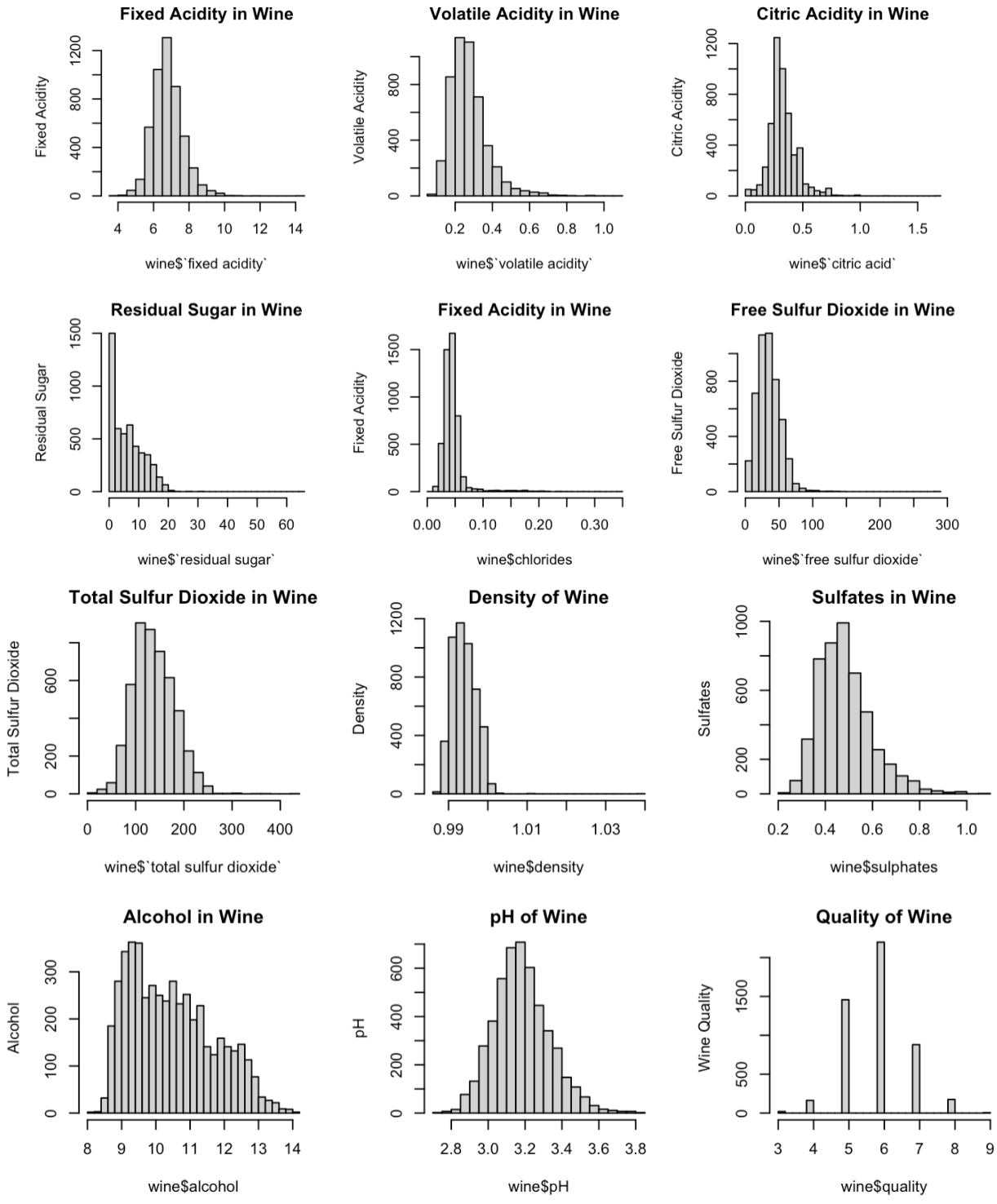
The dataset is about the physicochemical variables that affect the quality of the Portuguese “Vinho Verde” White wine, more description about the dataset can be found below. In this experiment we will be using R to estimate the quality of white wine by different variables based on their chemical composition. Both linear and non-linear methods of prediction were used in this paper. Two different final models were achieved with different levels of accuracy.

Dataset

The data set I will be using is publicly available for research purposes from https://archive.ics.uci.edu/ml/datasets/Wine+Quality, which studies the Portuguese “Vinho Verde” wine. The dataset explores 4898 different variants of white wine, and the quality score attached to each one was assessed by wine experts also known as sommeliers. The quality rating is based on a sensory test carried out by these sommeliers. They rated each wine from 0 - very bad to 10 – very excellent. The features included in the dataset are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol. These will be our predictor variables and quality will be our dependent variable. With this data I aim to predict the rating that an expert will give to a wine sample, using a range of physicochemical properties, such as density and alcohol percentage. Description of each variable are as follows:

Fixed Acidity: Amount of Tartaric Acid in wine, measured in g/dm3
Volatile Acidity: Amount of Acetic Acid in wine, measured in g/dm3
Citric Acid: Amount of citric acid in wine in g/dm3. Contributes to crispness of wine.
Residual Sugar: amount of sugar left in wine after fermentation. Measured in in g/dm3
Chlorides: amount of Sodium Chloride (salt) in wine. Measured in g/dm3
Free Sulfur Dioxide: Amount of SO2 in free form. Measured in mg/dm3
Total Sulfur Dioxide: Total Amount of SO2. Too much SO2 can lead to a pungent smell. SO2 acts as an antioxidant and antimicrobial agent.
Density: Density of Wine in g/dm3
pH: pH of Wine on a scale of 0-14. 0 means highly Acidic, while 14 means highly basic.
Sulfates: Amount of Potassium Sulfate in wine, measured in g/dm3.Contributes to the formation of SO2.
Alcohol: alcohol content in wine (in terms of % volume)
Quality: Wine Quality graded on a scale of 1 - 10 (Higher is better)

To further examine this data we construct univariate plots to get an idea of the distribution of our dataset. I decided to use histograms for each variable to examine its distribution.



Something to consider about this dataset is that the dependent variable we are predicting is based on human input. Wine sommeliers ranked the wines based on their individual taste. That being said I expect sugar, alcohol and density to be significant factors for our analysis. In the dataset we see that residual sugar had

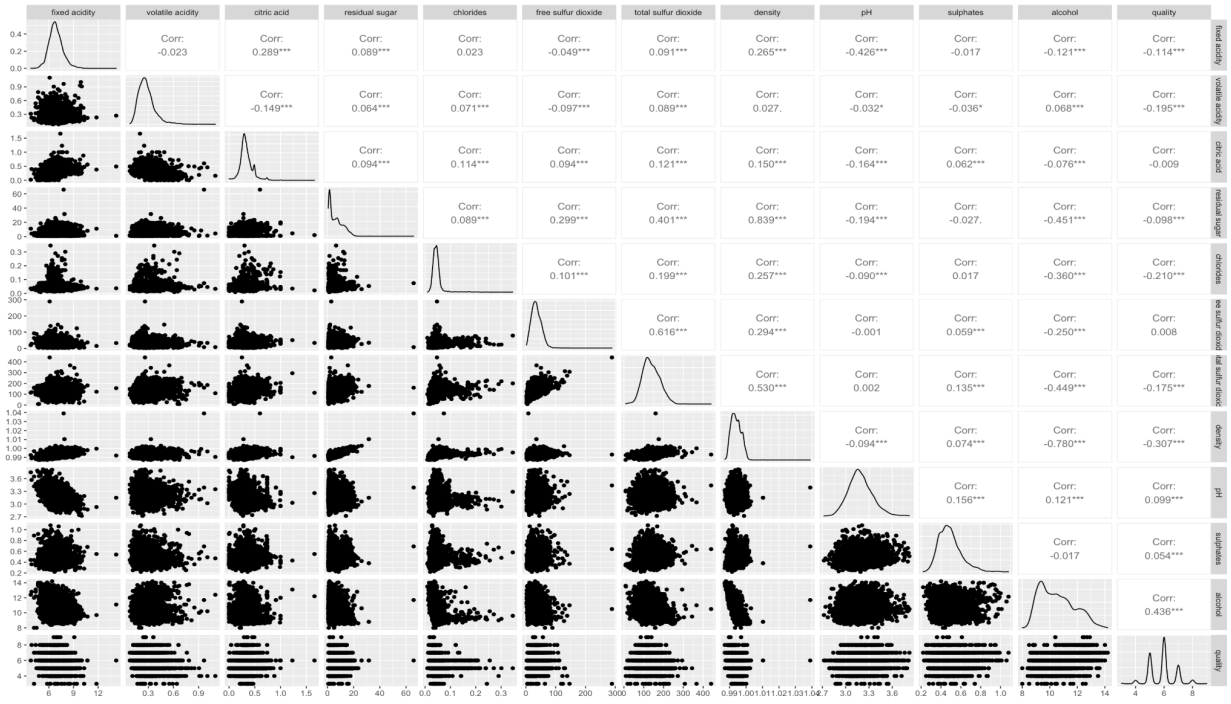
two peaks at 2 and 10, so we can say that most sommeliers either prefer really sweet or really dry wine.

Data Mining Methodology

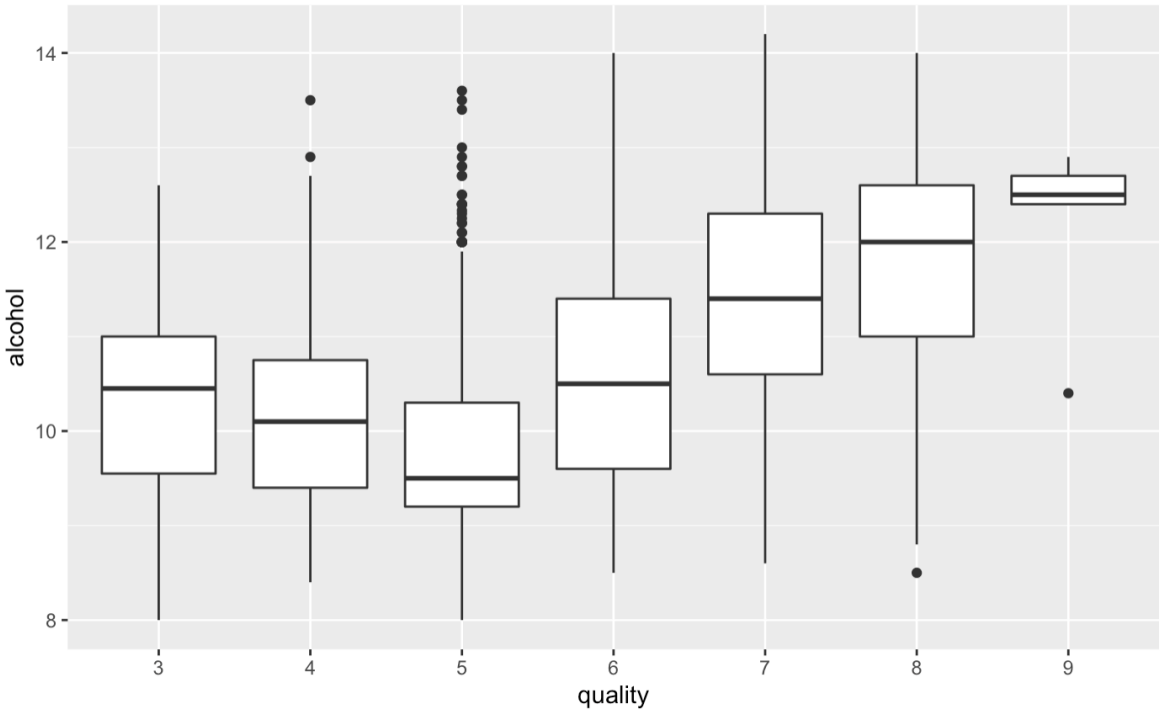
In this experiment I will be using data mining techniques to determine the dependency of wine quality on other variables and then use this knowledge to create a model to predict white wine quality. First, we must ensure our variables are normally distributed and do not suffer from a multicollinearity problem. Then I will assess which variables are statistically significant regarding quality prediction. I will also use Vector Inflation Factor analysis to quantify multicollinearity within the model. I will be using R language for the analysis, along with the following libraries: readxl, ggplot2, car, carDATA, rpart, randomForest, rpart.plot, caret, and psych.

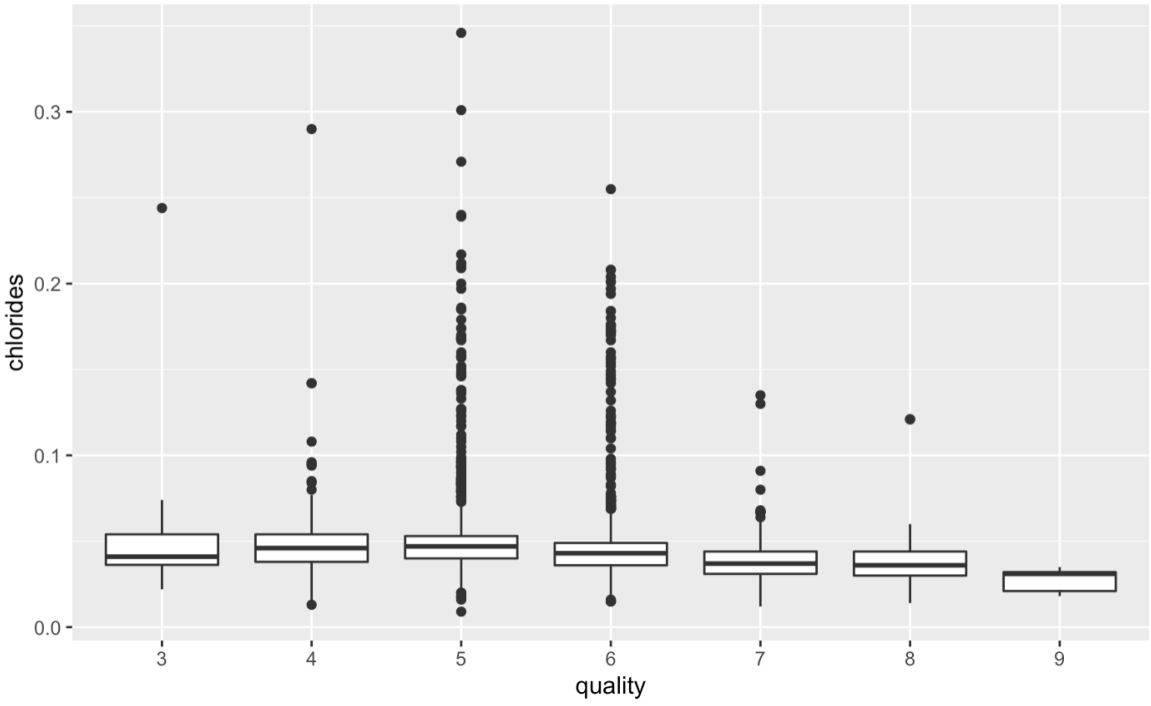
Correlation Analysis

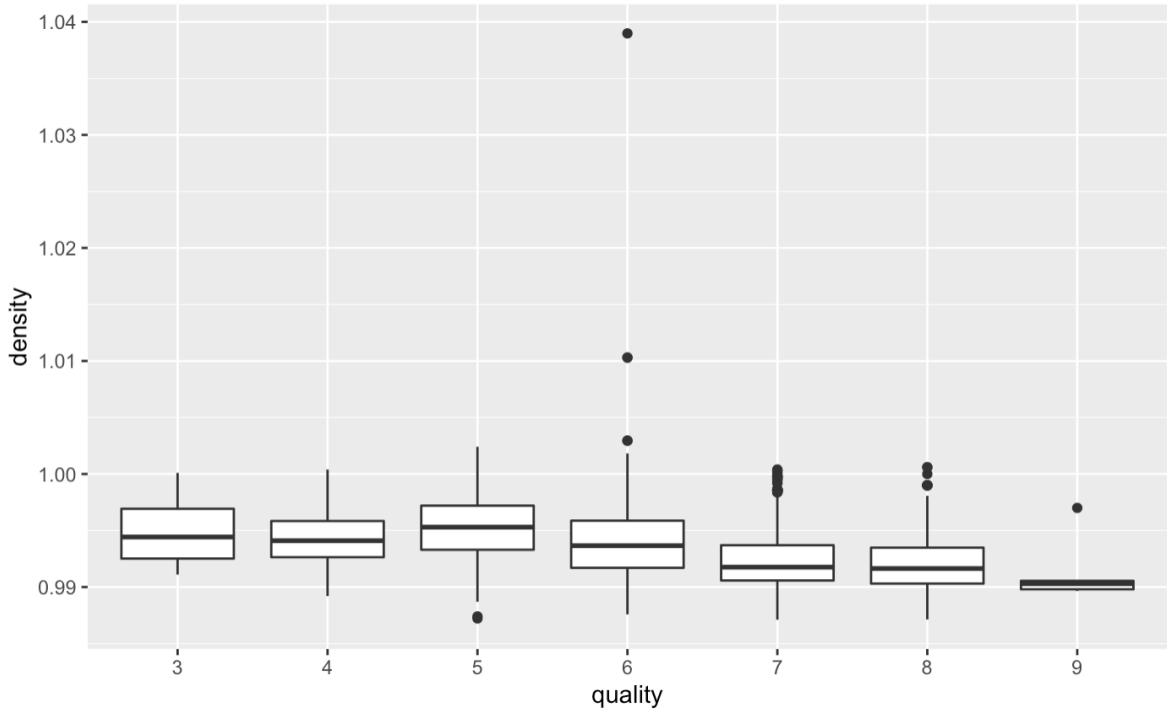
To quantify the affects the predictor variables have on wine quality. I computed the pairwise Pearson's correlation coefficients between each physicochemical attribute against wine quality. The summary of this analysis was then put in a correlation matrix.



From here we see that quality is mostly correlated with alcohol, density, and chlorides. Aslo alcohol is highly correlated with chlorides and sugar. To explore this further we construct grouped boxplots to analyze this phenomenon.


alcohol
14
12
10
8
3
4
5
6
7
8
9
quality


chlorides
0.3
0.2
0.1
0.0
3
4
5
6
7
8
9
quality



From the boxplots we can clearly see that higher quality wines tend to have high alcohol levels, are low density and have fewer chlorides. We can draw the conclusion that these variables will be statistically significant in our upcoming analysis.

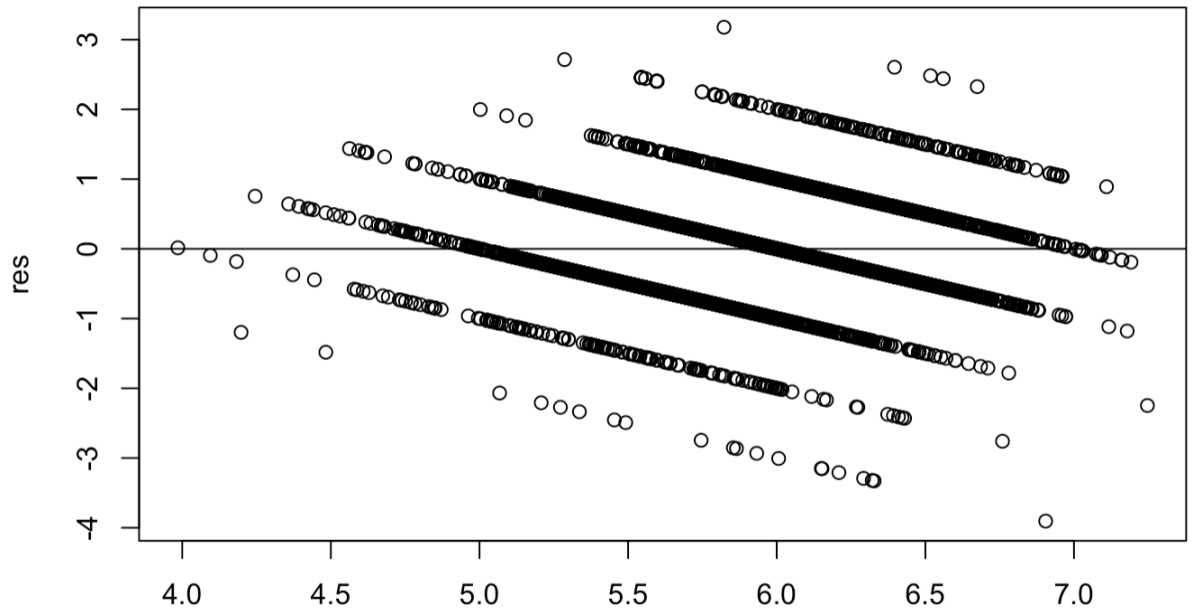
Linear Regression

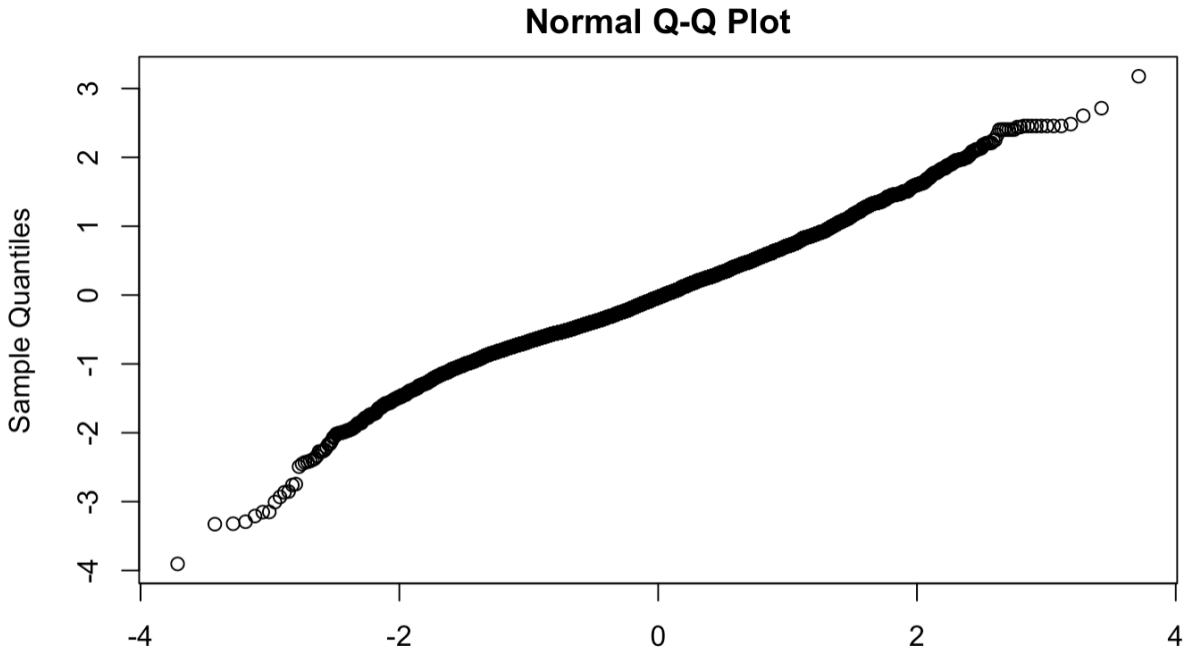
Initially I ran a regression model that included all 11 variables to get a preliminary understanding of the significance each predictor variable had on the dependent variable, “quality”. The adjusted R squared value of this model was 0.2803. Which means that the quality of wine is only ~28% reliant on the predictor variables. This value shows us that there are variables in this dataset that are statistically significant in predicting white wine quality. By examining the p-values of each variable we can select which variables predict wine quality. For this analysis I will be using a 95% confidence interval on all statistical testing. The variables Citric Acid, Chlorides, and Total Sulfur Dioxide have a p-value of over 0.05, which is outside of our significance threshold. After running the model with only 8 variables that were found to be statistically significant the Adjusted R squared is now 0.2806 which is only an increase of 0.003 and still not significant enough to accept.

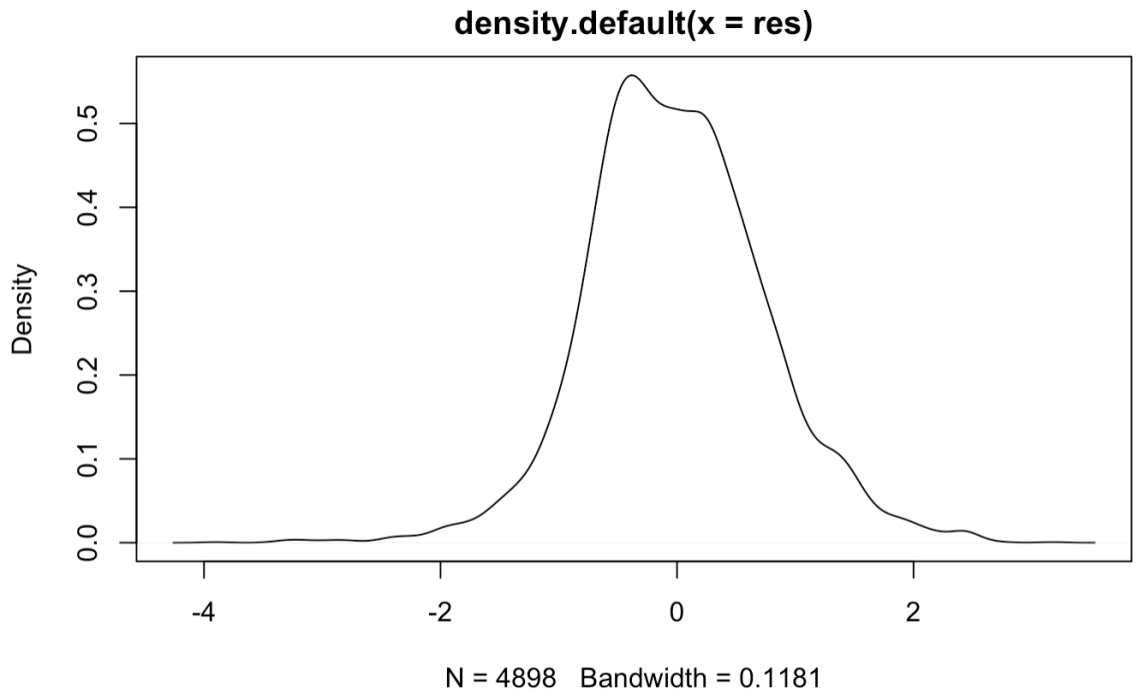
Vector Inflation Factor Analysis

The linear regression analysis helped in identifying which variables were statistically significant in the model. The next issue to address is multicollinearity, which is when two or more predictor variables might be correlated with each other. The next step in my experiment was to conduct a Variance Inflation Factory (VIF) analysis. This helps to further stabilize our model by quantifying how much the variance of its coefficient is inflated due to multicollinearity in the model. The smallest possible value of VIF is one which would indicate an absence of multicollinearity. A widely accepted rule of thumb within data science is that if a VIF value is greater than 5 it indicates a significant amount of collinearity, which would be detrimental to the model.

After all variables have been reduced to a VIF of under 5, we test for significance of each variable within the model by looking at the p values. Citric Acid was removed as it was not statistically significant. Unfortunately this only got our R-Squared value to 0.27. Here are the model validation plots.

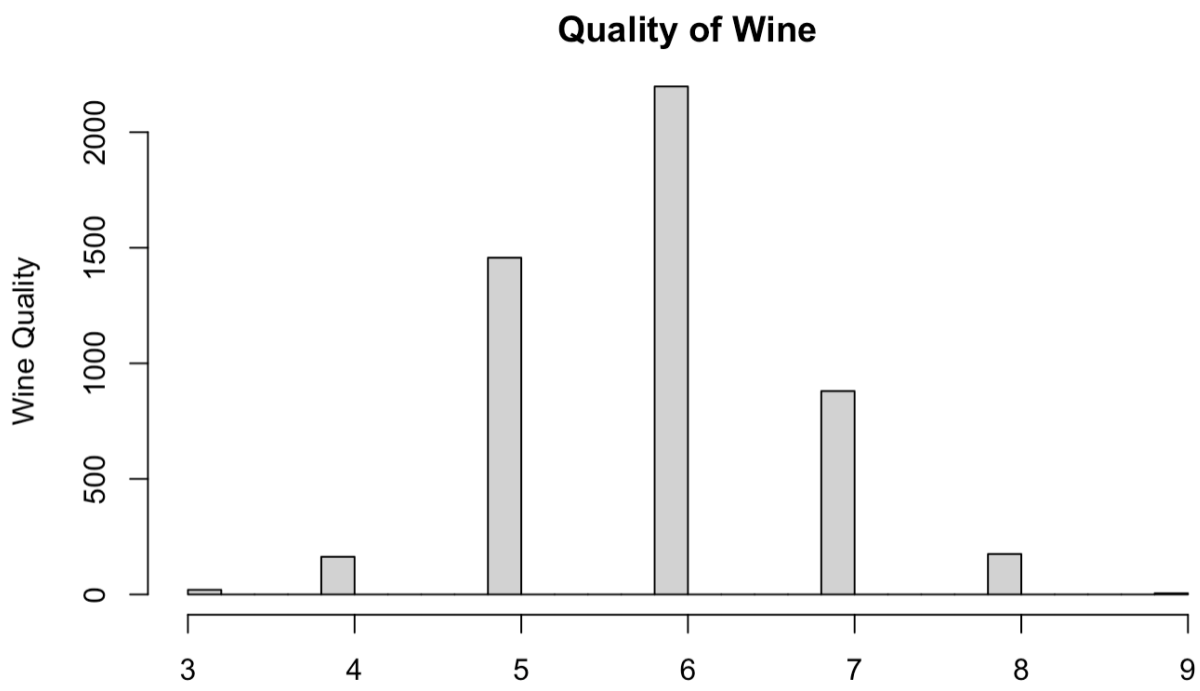

res
3
2
1
0
-1
-2
-3
-4
4.0
4.5
5.0
5.5
6.0
6.5
7.0


Normal Q-Q Plot
Sample Quantiles
3
2
1
0
-1
-2
-3
-4
-4
-2
0
2
4



Random Forest Algorithm

After using two linear methods, I decided to try a non-linear approach to construct a prediction model. For this approach I went with a random forest algorithm to analyze the link between the predictor factors in this dataset and the response variable, since the relationship between them may be more complex than I initially thought. After looking at the distribution of data points within the quality variable.



We can see here, there are a lot more wines with a quality of 6 as compared to the others quality levels. So we can draw the conclusion that there are a lot more normal wines than excellent or poor ones. For the purpose of this experiment, I will classify the wines into high, low, and medium based on their quality. Where if a wine is greater than six it is marked as 'high quality', equal to six is "medium quality" and anything lower than six is "low quality". After creating and testing our model we achieved 69.6% accuracy. This is very promising and nearly 3x more accurate than linear methods of prediction.

Conclusion

This dataset has almost 5000 wine data points, classified into 12 variables, and were graded by 3 different expert wine sommeliers. We found that the quality of the wine is positively correlated with alcohol, free.sulfur.dioxide, sulfate and pH and it is negatively correlated with the other physicochemical input variables. We explored the different variables and analyzed the relationship between the different chemical properties. To build our prediction model we first tried a linear regression approach which yielded a poor result of 27% accuracy. Then we used the Random

Forest algorithm to predict which quality bucket wines would fall in. This approach yielded a ~70% accurate model.