CUNY Baruch College

Term Project: Red Wine Quality

Group Submission

Yula Ko

Farzana Manjra

Natalia Gomez

Victoria Wayda

JiaRui (Jesse) Shao

STA 3920 NFA

Yuanfeng Cai

11 December 2018

1. Introduction

The wine in this dataset was evaluated and their quality determined by wine professionals. We plan to analyze the data and evaluate which chemical properties classify a wine as 'good’ quality by utilizing various methods. We believe the target audience that may find this useful could be anyone who is looking to determine the quality of wine based on these variables. This dataset consists of 12 variables and 1,599 observations. Quality is our response variable and we plan to determine which of the other variables have a greater effect in determining wine quality.

2. Exploratory Data Analysis & Data Visualization

In order to give some meaning to the attributes, we have provided a brief description of each one. This aids our audience in understanding how these variables could have an effect on the quality. Fixed acidity refers to the acids that are present and do not evaporate easily. Volatile acidity could indicate spoilage or mistakes in processing, such as using damaged grapes. This causes acetic acid bacteria to grow and make the wine taste unpleasant. Citric acid can add a fresh taste to the wine and it’s usually added as a preservative. Residual sugar is sugar still present after fermentation. Chlorides refer to the amount of salt and free sulfur dioxide prevents oxidation. The density of the wine is dependent on the sugar content and the pH varies depending on the acidity. Tart red wine generally has a lower pH and softer tasting red wine has a higher pH. Sulphates contribute to the SO_2 levels and alcohol is the percentage of alcoholic content. With some newfound knowledge on wine we’re going to start exploring our data.

Despite being a numeric value, the quality variable, or y-response, is actually not a quantitative variable but rather a qualitative, categorical variable. We created a bar chart to visualize how the wine in this dataset falls into one of the following six quality categories: 3, 4, 5, 6, 7, and 8. With 855 good quality wines and 744 bad quality wines, we saw a normal distribution in our bar chart, meaning the data is equally represented (Exhibit 1). Instead of having six responses for the quality variable we converted it to a binary response falling into the two descriptive categories, good quality and bad quality. If the quality is greater than six the wine is categorized as good quality. If the quality is less than six the wine is considered bad quality. We also created a summary table with the following statistics for each variable:

minimum, maximum, mean, standard deviation, and median. This revealed some interesting observations. For example, we saw outliers in free sulfur dioxide and total sulfur dioxide. We also saw that pH levels varied, indicating there are tart and sweet wines present (Exhibit 2).

Exploratory data analysis was conducted by creating scatter plots and boxplots to study the data. Since all of the variables were continuous except for the quality variable, we made a few pairwise scatter plots to explore our data and see if there was a possibility of any correlation between the continuous variables. In Exhibit 3, Graph A you will see a scatter plot of some of the continuous variables, not all, because it would be easier to see with few variables per pairwise scatter plot so we made more than one. If you look at Graph A you will see a pattern in the density and citric acid scatter plot and since it's gradually increasing, it's a sign of a possible positive correlation. You can also see a possible positive correlation between variables density and fixed acidity, and citric acid and fixed acidity. However, there are more scatter plots where you can see a pattern like the pH and citric acid plot, the pH and fixed acidity plot and even the density and pH plot. Here the pattern gradually decreases which means a possible negative correlation. The other plots you don't see any patterns so we can't say much about them except that the plots look quite random indicating a possibility of no correlation. If you look at Graph B, another pairwise scatter plot, which has more scatter plots of the other variables we can see some patterns in some of the plots as well. You can see a possible negative correlation between variables alcohol and density since the pattern is decreasing and also a possible positive correlation between free sulfur dioxide and total sulfur dioxide because you can see it increasing. The rest of the scatter plots don't show any pattern so there's a possibility of no correlation. The heatmap(Graph D) shows correlation but in numerical form. It basically reads the pairwise scatter plots but just numerically. The numbers closer to 1 or -1 mean high correlation and closer to 0 means no correlation. The graph you see has red hue and blue hue, red means positive correlation and blue mean negative correlation. Pairwise scatterplots and heat maps are very useful to explore data and see patterns.

Like we mentioned before we did not only have all continuous variables, we also had a qualitative variable also known as a categorical variable which was our quality variable also our y-response. Here we used box plots to explore our data since we had a continuous and a categorical variable. If you look at Exhibit 3 Graph C, you will see three boxplots, here we graphed quality with alcohol, citric acid and volatile acidity. These three graphs if you look separately there averages between good and bad quality are quite different and vary. This is a possible indication of correlation, but we cannot confirm yet. If you look at Graph D you will see that these box plots with three other variables show that the averages are the same for good and bad quality wine which possibly means that these variables don't have an impact on quality of wine.

3. Logistic Regression

We are interested in predicting what variables make wine quality is good. We are going to apply logistic regression to estimate the probability of good wine and probability of bad wine. First we fit the first model which includes all 11 independent variables. We used alpha of 0.05 as our cut off for significance. Based on the results we can see that not every independent variable is statistically significant. These variables have high p-values: p-value for fixed acidity is 0.17, for residual sugar is 0.3, for density is 0.53 and for ph is 0.6 (Exhibit 4).Then we fit the second model which includes 7 independent variables such as alcohol, volatile acidity, chlorides, citric acid, free sulfur dioxide, total sulfur dioxide and sulphates, but based on Exhibit 4 we can see that p-value for citric acid is 0.45 which is much higher than alpha of 0.05 and it is not statistically significant. Next we fit the third model which includes 6 variables such as alcohol, volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide and sulphates. Based on the results we see that all variables are statistically significant. Their associated p-values are very low (Exhibit 4). We also checked for multicollinearity in each model by computing variance inflation factor. The VIF score for the predictor fixed acidity is very high (VIF=7.97) and VIF score for the predictor density is also high (VIF=5,77) in the first model. This might be problematic. VIF scores for predictors in the second model and the third model are below 5 which indicates we don’t have multicollinearity.

Then we estimated these models by computing their AIC and BIC. The Exhibit 5 shows that the first model gets 1679.625 on AIC, 1744.151 on BIC, the second model gets 1680.155 on AIC, 1723.172 on BIC, and the third gets 1678.716 on AIC and 1716.346 on BIC. The third model has smaller AIC and BIC. This is an evidence that the third model is a better fit to the data than the first model and the second model.

Next we evaluated the prediction accuracy of these models. First we used holdout method. We randomly selected 80% of the observations for training and the remaining in the test set. We evaluated the performance of the models using test set and we produced a confusion matrix to determine how many observations were correctly classified. The accuracy for the first model and the second model is 77.50% and the accuracy for the third model is 77.81% (Exhibit 6). The accuracy for the third model is slightly higher compare to the other models. Then we evaluated the prediction accuracy of these models by performing 10-fold cross-validation. The accuracy for the third model is 74.63% which is higher than for the first model (74.29%) and the second model (74.61%) (Exhibit 7). These results indicate that the third model is better than the first model and the second model.

4. K-NN

We could also perform K-NN on this dataset using all the 11 variables as predictors. First, we used 5-fold cross-validation to select the best K. In this case, the best k is 1 and the corresponding overall accuracy of the model is 75.91%.

5. Classification

Next, we are going to implement decision tree method. We performed 10-fold cross-validation to find the best subtree. The Exhibit 8 tells us in what cases wine quality will be good, in what cases wine quality will be bad. It indicates alcohol more than 10.525 always leads to good wine, and alcohol less than 10.525 follows various paths. This plot shows us that the level of alcohol in wine is what matters for making good wine. The overall accuracy of the model is 70.94% and recall rate is 63.83% (Exhibit 9). Single decision tree can tend to over-fit because of their high variance. Random forest reduces this variance by averaging many trees. So next, we implemented random forest method which combines a lot of decision trees. The accuracy of the model is 84.06%. We successfully improved the model from the one in decision tree. In the plot of the feature importance of the variable (Exhibit 9), we see that alcohol is the most important variable affecting the quality, and following sulphates with the highest mean decrease in Accuracy and Gini.

6. Conclusion

We used Logistic Regression, K-NN, and Classification exploring this dataset to determine the factors that are important for wine quality. According to the results we got, we can see that the random forest gave us the highest accuracy that is 84.06%, followed by the 77.81% rate by logistic regression and then followed by 75.91% accuracy rate by K-NN and finally decision tree with 70.94% of accuracy. Thus we learned that alcohol and sulphates are the two most essential factors when it comes to wine quality.

APPENDIX

EXHIBIT 1: Bar Chart

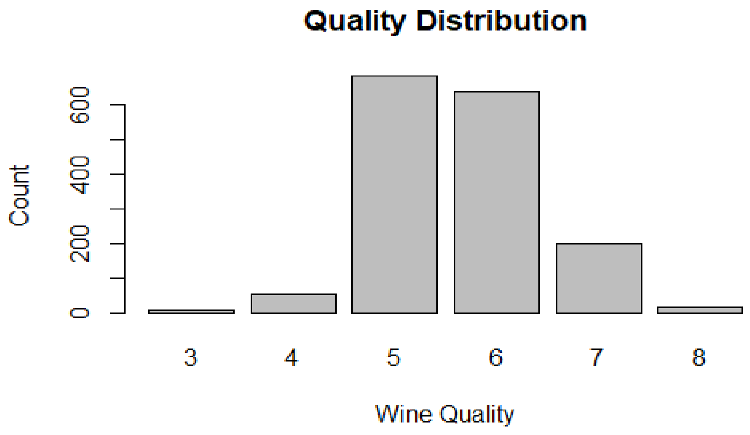
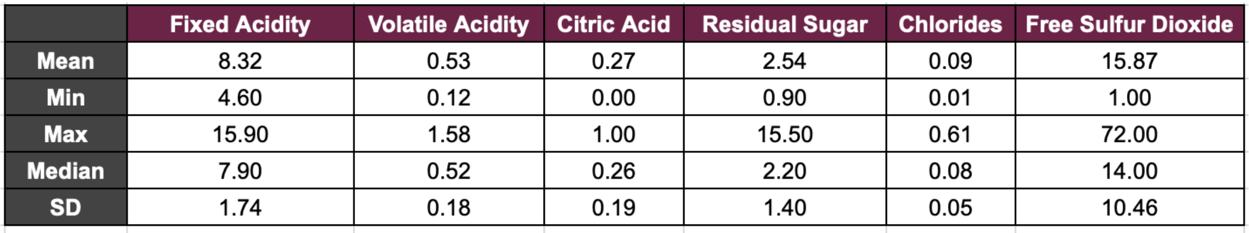
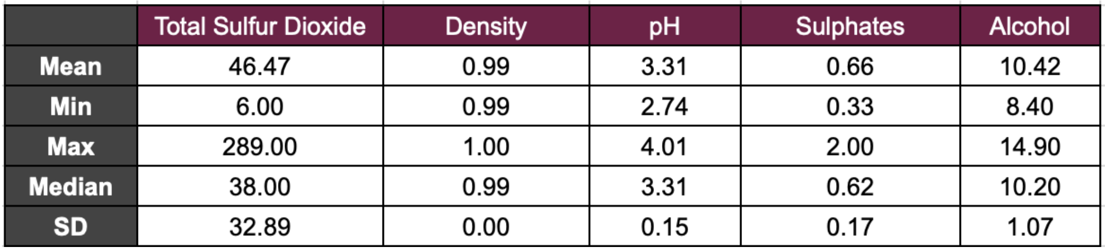


EXHIBIT 2: Summary Table

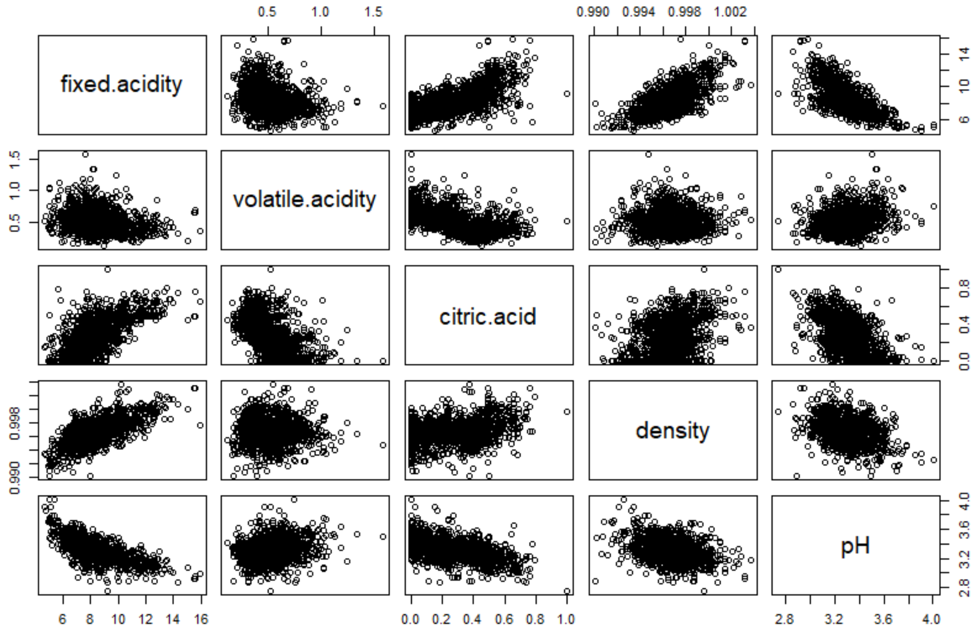


	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulfur Dioxide
Mean	8.32	0.53	0.27	2.54	0.09	15.87
Min	4.60	0.12	0.00	0.90	0.01	1.00
Max	15.90	1.58	1.00	15.50	0.61	72.00
Median	7.90	0.52	0.26	2.20	0.08	14.00
SD	1.74	0.18	0.19	1.40	0.05	10.46

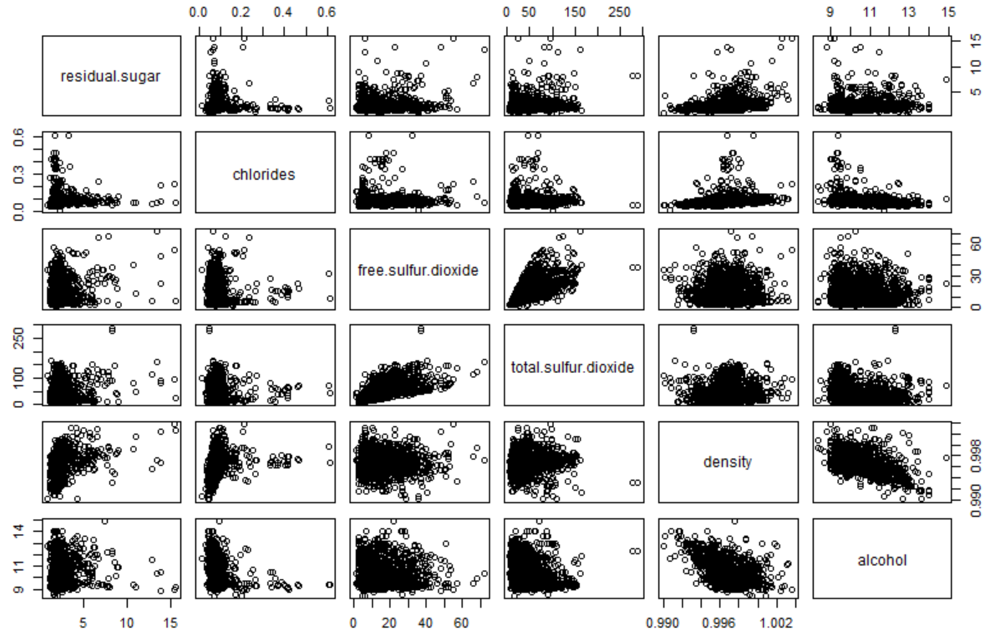


	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol
Mean	46.47	0.99	3.31	0.66	10.42
Min	6.00	0.99	2.74	0.33	8.40
Max	289.00	1.00	4.01	2.00	14.90
Median	38.00	0.99	3.31	0.62	10.20
SD	32.89	0.00	0.15	0.17	1.07

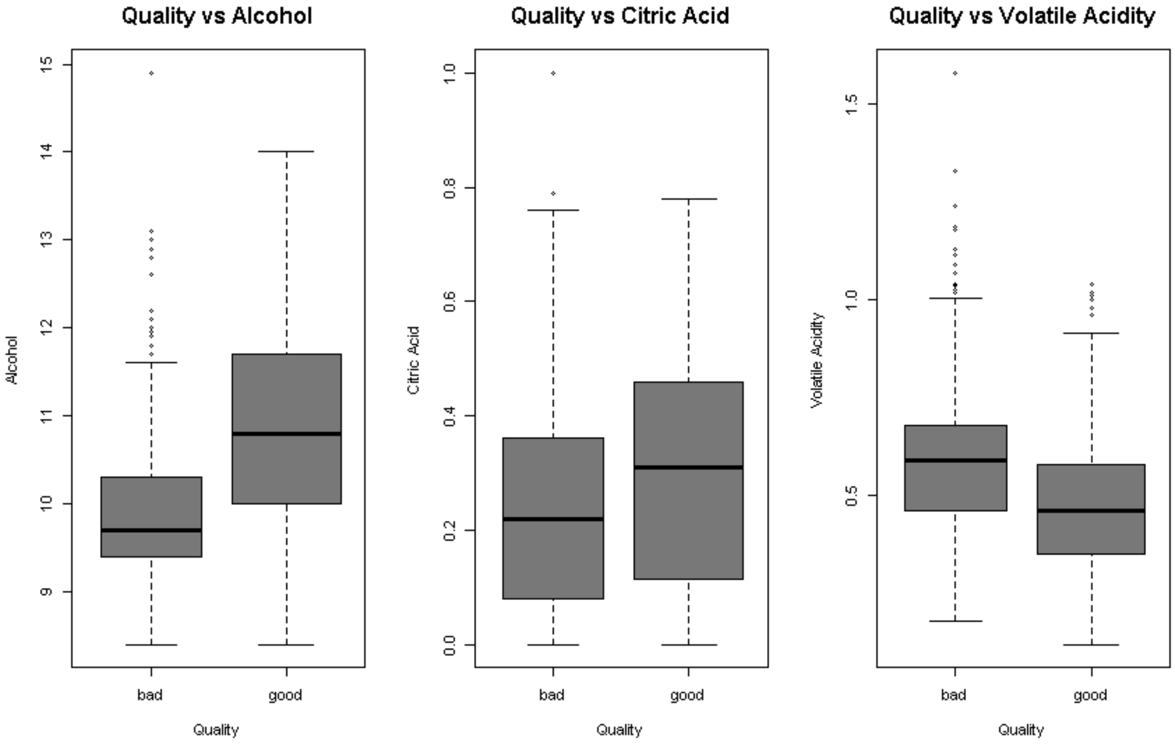
EXHIBIT 3: Exploratory Analysis Visualizations



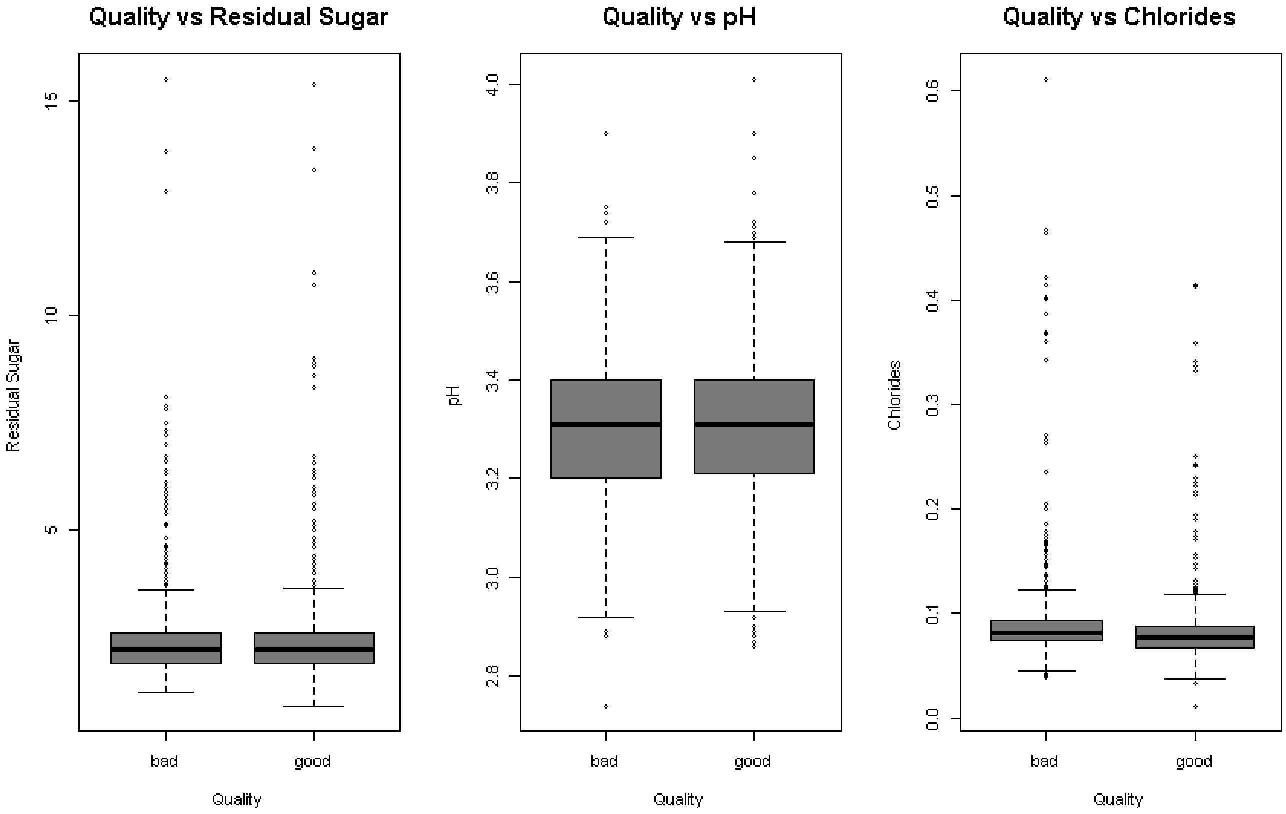
Graph A



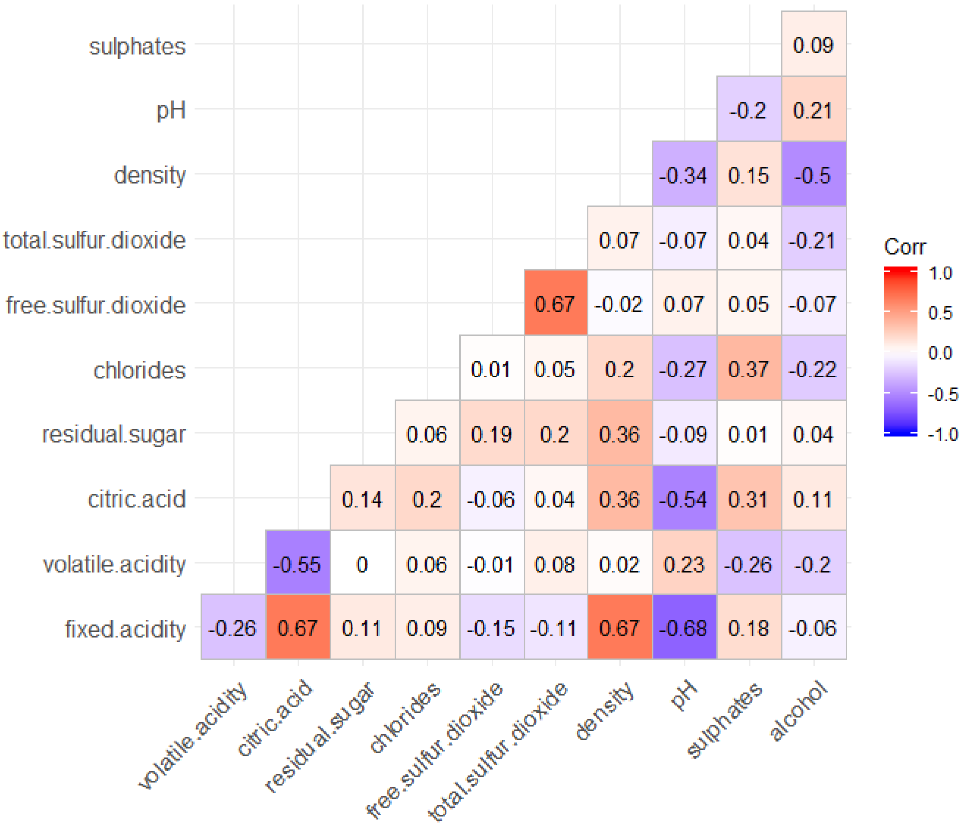
Graph B



Graph C

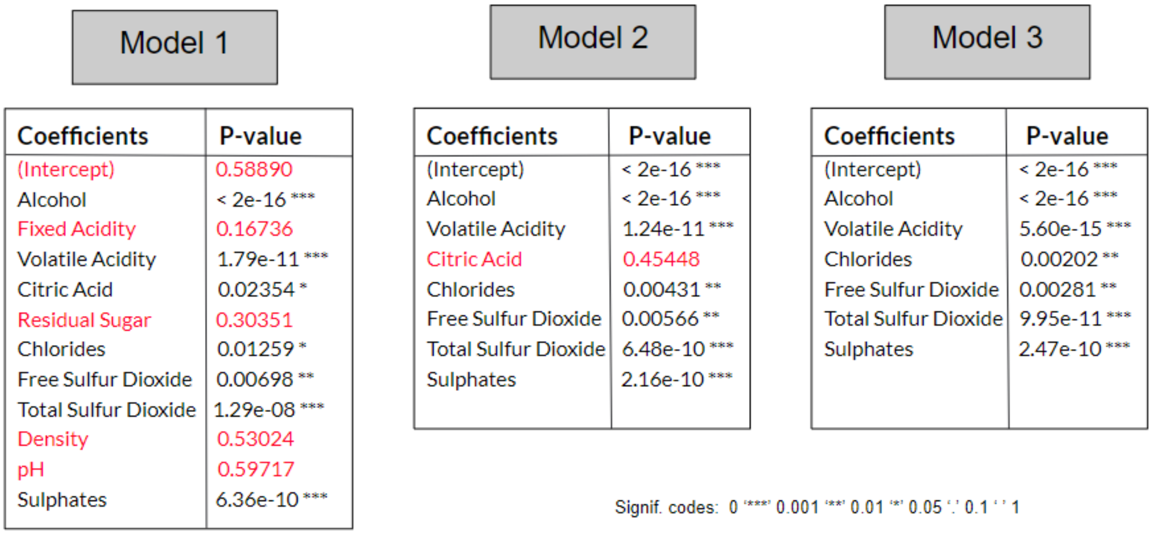

Quality vs Residual Sugar
Residual Sugar
15
10
5
bad
good
Quality
Quality vs pH
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
bad
good
Quality
Quality vs Chlorides
Chlorides
0.6
0.5
0.4
0.3
0.2
0.1
0.0
bad
good
Quality

Graph D



Graph E

EXHIBIT 4: Logistic Regression



Model 1

Coefficients	P-value
(Intercept)	0.58890
Alcohol	< 2e-16 ***
Fixed Acidity	0.16736
Volatile Acidity	1.79e-11 ***
Citric Acid	0.02354 *
Residual Sugar	0.30351
Chlorides	0.01259 *
Free Sulfur Dioxide	0.00698 **
Total Sulfur Dioxide	1.29e-08 ***
Density	0.53024
pH	0.59717
Sulphates	6.36e-10 ***

Model 2

Coefficients	P-value
(Intercept)	< 2e-16 ***
Alcohol	< 2e-16 ***
Volatile Acidity	1.24e-11 ***
Citric Acid	0.45448
Chlorides	0.00431 **
Free Sulfur Dioxide	0.00566 **
Total Sulfur Dioxide	6.48e-10 ***
Sulphates	2.16e-10 ***

Model 3

Coefficients	P-value
(Intercept)	< 2e-16 ***
Alcohol	< 2e-16 ***
Volatile Acidity	5.60e-15 ***
Chlorides	0.00202 **
Free Sulfur Dioxide	0.00281 **
Total Sulfur Dioxide	9.95e-11 ***
Sulphates	2.47e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

EXHIBIT 5: Model Estimation

	Model 1	Model 2	Model 3
AIC	1679.625	1680.155	1678.716
BIC	1744.151	1723.172	1716.356

EXHIBIT 6: Holdout Method

Model 1

	True Bad	True Good
Predicted Bad	105	45
Predicted Good	27	143

Model 2

	True Bad	True Good
Predicted Bad	101	41
Predicted Good	31	147

Model 3

	True Bad	True Good
Predicted Bad	101	40
Predicted Good	31	148

EXHIBIT 7 : 10- Fold Cross-Validation

	Accuracy
Model 1	74.29
Model 2	74.61
Model 3	74.63

EXHIBIT 8: Decision Tree

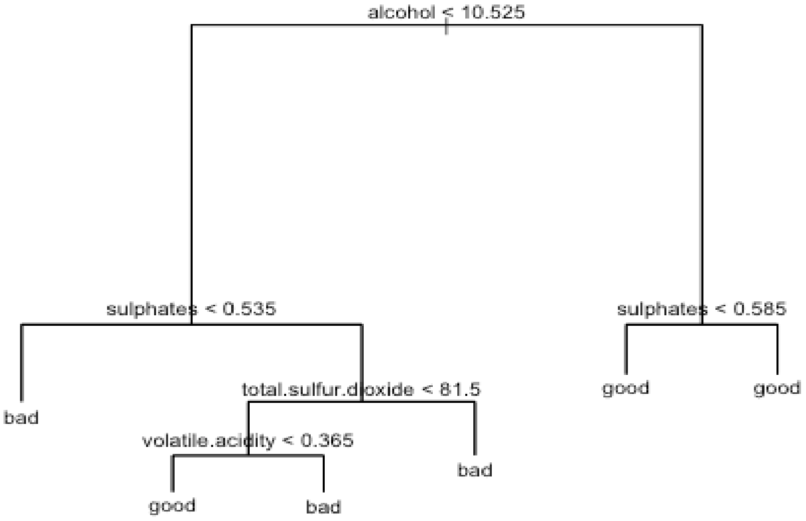
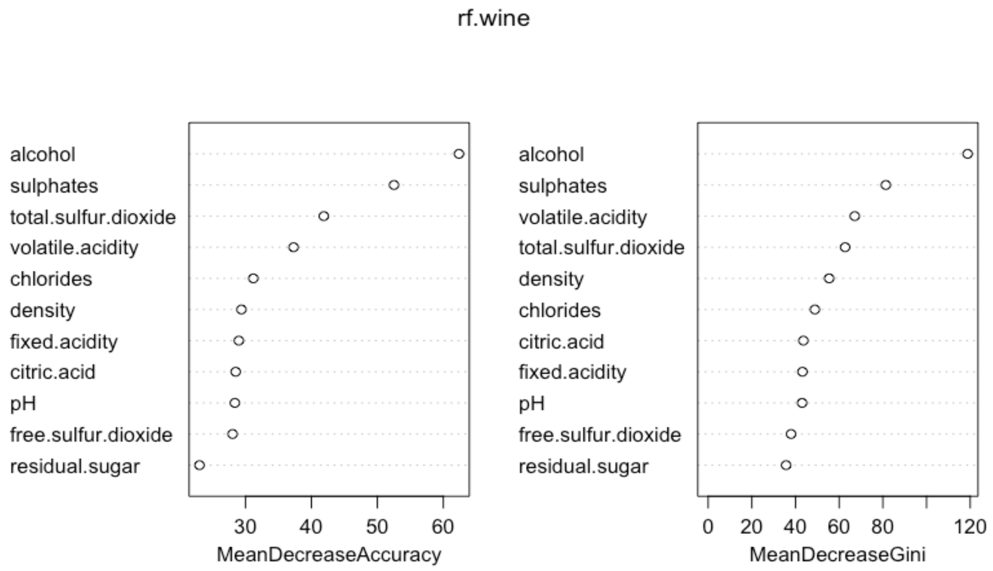


EXHIBIT 9: Confusion Matrix for Decision Tree

	True Bad	True Good
Predicted Bad	107	68
Predicted Good	25	120

EXHIBIT 9: Variance Importance


rf.wine
alcohol
sulphates
total.sulfur.dioxide
volatile.acidity
chlorides
density
fixed.acidity
citric.acid
pH
free.sulfur.dioxide
residual.sugar
30
40
50
60
MeanDecreaseAccuracy
alcohol
sulphates
volatile.acidity
total.sulfur.dioxide
density
chlorides
citric.acid
fixed.acidity
pH
free.sulfur.dioxide
residual.sugar
0
20
40
60
80
120
MeanDecreaseGini