# STA363 Final Project Report

*Matt Sherrick*

*12/10/2019*

**Abstract**

In this report, I will be looking at a dataset that includes several physicochemical properties within red wine. Our goal is to determine which of these factors makes a wine considered "good." We set a threshold on the response variable, quality, to answer this by saying anything with a quality of 6 or over is deemed "good." In order to determine what properties of wine were most closely related to good wine quality, we use a combination of the best subset selection and random forest methods. Using visualization to explore our data, we determined that some of the more important predictors were alcohol percentage, volatile acidity, sulphates, and citric acid. We then fit both linear and logistic regression models using best subset selection to choose a model with the best goodness-of-fit and prediction ability. This model included the influential physicochemical indexes that we determined would be important for prediction. We then create trees by fitting a random forest model to determine what predictors are important in each tree. Our results indicated that the random forest model produced the lower test MSE and thus was the better model for prediction. This model also supported our earlier findings that the three most important physicochemical properties that positively contribute to good wine quality are alcohol content, volatile acidity, and sulphates.

# Contents

# Section 1 : Motivation and Background

## 1.1 : Motivation of choosing this dataset

For my report, I chose to analyze the dataset titled Red Wine Quality acquired from kaggle.com [2]. It includes data about the numerous aspects within wine that affects its quality. I had the motivation to use this dataset after a thought I had this past Thanksgiving. My dad was a big wine collector while he was still alive. My mom has kept some of my dad's nicest wines stored in our house, and she brought out a couple of these bottles for our Thanksgiving dinner. I always wondered what makes a specific bottle of wine considered "good." I felt that working with this dataset would help me find the answer. From working through this dataset, I believe lot of insight can be found through data analysis that can be important to any wine connoisseur out there.

## 1.2 : Analyzing the peer-review journal article

There has been several statistical contributions to the topic of wine quality prediction. As a part of this section, I will be reviewing the article "Quality Evaluation Based on Multivariate Statistical Methods" written by Shen Yin, Xiangping Zhu, and Hamid Reza Karimi. Their work involves working on a dataset of wine quality and physicochemical indexes of grapes and creating quality prediction models using the multivariate statistical methods ordinary least squares regression (OLSR), principal component regression (PCR), partial least squares regression (PLSR), and modified partial least squares regression (MPLSR). The goal of their analysis is to help make predictions on wine quality and make better decisions towards "grape selection, wine classification, and target marketing" before the wine making process [4].

The article begins with an introduction about wine quality and previous methods that have been used in this area of study, such as principal component analysis. Yin goes on to talk about some of the important factors based on these previous studies. Most importantly, the grape is the "most basic and important factor" [4] in this process; its physicochemical indexes have a significant relationship to the quality of the wine, and the sugar has an impact on the alcohol fermentation. It is was also found in one study that the variety of grapes have a significant impact on wine quality [4]. One of the most prevalent issues that arise when studying wine quality is multicollinearity. The statistical methods used in this journal are aimed to counter this problem. The next section discusses the four methods used and the mathematics behind them all, as well as the purpose of using each one. Following this methodology is a section on the winemaking process that includes a six-step task: grape harvesting, making the must, fermenting the wine, bottling, aging, then pressing and setting [4]. This section is purely to inform the reader of the important things to realize during all these steps.

We now shift to the results and discussion section, which is split into three subsections: processing the data, model comparison, and choosing the grape physicochemical indexes. The data itself includes the wine quality and fifty grape physicochemical indexes. Samples were taken from ten human experts to evaluate the wine quality based on four categories: "presentation (15 points), fragrance (30 points), mouth feel (44 points), and overall feeling (11 points)" [4]. The four major points are summed up and then divide by ten to represent each human expert in order to get the wine quality. Then, the samples are divided up where 27 samples represent the calibration set and 10 samples make up the verification set. The calibration set is used to fit the quality prediction models while the verification set is used to assess the predictive ability [4]. For the methods themselves, there were a set of assumptions made to ensure that any potential differences in wine quality are due to the differences in grapes. First, it is assumed that identical vinification processes were followed, meaning that the same materials were added during the winemaking process for all the wines in the data. Second, the winemakers all had similar vinification processes. Lastly, the wine quality was assessed by the human experts right after the vinification.

When it comes to utilizing each method, Yin begins by deeming OLSR unable to fit a good model due to multicollinearity and the fact that the number of samples is less than the number of predictor variables [4]. He proceeds with the other methods, PCR, PLSR, and MPLSR, in order to build the models for the report. To compare the models, Yin uses two measures. The first one is the root mean squared error of calibration, or RMSEC, which measures the fitting ability using the calibration samples; the second measure is the root

mean squared error of prediction, or RMSEP, which determines the model's predictive ability using the verification samples [4]. We want these two values to be small. To use PCR, the cumulative percent variance (CPR) is set to 95% to obtain sixteen principal components. Using cross-validation, four latent variables were chosen for PLSR. Since MPLSR uses about all the information from the data, it does not require a selection of principal components or latent variables. Based on figures and measures provided in the report, it is found that the PCR method has the weakest fitting ability with a RMSEC of 4.53. The best method among the three is the MPLSR, which has a RMSEC of 0 meaning the fitted wine qualities equaled the actual qualities. This is probably because of the fact that MPLSR avoids the obstacle of picking principal components and latent variables. Yin then turns to the RMSEP to compare the predictive ability between the three created models. Similar results are found when computing this measure; the PCR had the highest value and the MPLSR had the smallest. This led to the conclusion that the model created used the MPLSR method has the best fitting ability and prediction ability.

The final subsection of the results section deals with choosing among the fifty grape physicochemical indexes that effectively predict wine quality. A measure called the contribution ratio (CR) is computed using the MPLSR model to collect values that indicates a physicochemical index's contribution to quality. A negative value indicates a "negative contribution to the wine quality" [4]. The table of results show the most influential factors of wine; alanine, total sugar, and soluble solids have a negative contribution to quality while protein and dry matter content have a positive contribution to quality. In order to cut out the indexes that have little to no influence, a CR threshold is placed, following which produces a new set with 37 physicochemical indexes. From the results in this set, similar indexes from the first model are present with the addition of new factors that positively contribute to wine quality: fructose, pH, and titratable acid. Dry matter content went from having high influence on quality to having very little impact in the new model as well. There was no difference in RMSEC values between the two models and a very small addition of 0.27 to the RMSEP value from the old model to the new model.

In the concluding section, it is determined that because of the strong RMSEC and RMSEP values, the model created using the MPLSR statistical method yields to the best fitting ability and prediction ability of all the methods considered. We can see that the grape physicochemical indexes of protein, fructose, pH, and titratable acid positively contribute to good wine quality, while total sugar and soluble solids negatively contribute to the wine quality.

## Section 2 : Cleaning

Our steps of cleaning our dataset involve checking to see if there is any missing data, followed by looking for variables that cannot be used. We can first check the summary of the dataset. We begin with 1,599 rows and 12 columns, with each variable being numeric. It is important to note that the wine quality variable ranges from integer values of 3 to 8, while the rest of the variables are represented as decimals. As for the cleaning itself, the dataset is very clean as it is. There were no missing values. Looking at the variables individually, we tend to see that some of variables have a higher range than others. For instance, total sulfur dioxide has a minimum of 6 and a maximum of 289. We can check for outliers in the dataset, however I believe that it may not be necessary. This is because we can make the assumption that these recorded physicochemical indexes are correct measurements. In addition, there were no missing values, which gives me further confidence that no data points that could be considered outliers were recorded due to experimental error.

We can observe a basic model with all the predictors to check the significance of each variable and observe whether or not there exists multicollinearity. First off, lets look at the summary of a standard linear model of our data.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 21.9652 | 21.1946 | 1.04 | 0.3002 |
| fixed.acidity | 0.0250 | 0.0259 | 0.96 | 0.3357 |
| volatile.acidity | -1.0836 | 0.1211 | -8.95 | 0.0000 |
| citric.acid | -0.1826 | 0.1472 | -1.24 | 0.2150 |
| residual.sugar | 0.0163 | 0.0150 | 1.09 | 0.2765 |
| chlorides | -1.8742 | 0.4193 | -4.47 | 0.0000 |
| free.sulfur.dioxide | 0.0044 | 0.0022 | 2.01 | 0.0447 |
| total.sulfur.dioxide | -0.0033 | 0.0007 | -4.48 | 0.0000 |
| density | -17.8812 | 21.6331 | -0.83 | 0.4086 |
| pH | -0.4137 | 0.1916 | -2.16 | 0.0310 |
| sulphates | 0.9163 | 0.1143 | 8.01 | 0.0000 |
| alcohol | 0.2762 | 0.0265 | 10.43 | 0.0000 |

We begin with a $R^2_{adj}$ value of 0.3561 and a F-statistic of 81.35 obtained from the summary of the model. In the table of the model, it is interesting to observe that some of the predictors have a p-value of over 0.05, meaning they are not statistically significant. One value we can use to detect multicollinearity is the VIF.

| | VIF |
|---|---|
| fixed.acidity | 7.77 |
| volatile.acidity | 1.79 |
| citric.acid | 3.13 |
| residual.sugar | 1.70 |
| chlorides | 1.48 |
| free.sulfur.dioxide | 1.96 |
| total.sulfur.dioxide | 2.19 |
| density | 6.34 |
| pH | 3.33 |
| sulphates | 1.43 |
| alcohol | 3.03 |

From this table, we see that fixed acidity and density have relatively high VIF values of 7.77 and 6.34 respectively. These variables also had the highest p-values in the model summary. In general, with VIF values this high, we tend to avoid including these variables in our model. We can check to see if we obtain a stronger $R^2_{adj}$ value when excluding these two predictors and observe any differences in the VIF values for the variables.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.6583 | 0.4611 | 10.10 | 0.0000 |
| volatile.acidity | -1.0815 | 0.1164 | -9.29 | 0.0000 |
| citric.acid | -0.1426 | 0.1229 | -1.16 | 0.2462 |
| residual.sugar | 0.0094 | 0.0120 | 0.78 | 0.4343 |
| chlorides | -1.9616 | 0.4030 | -4.87 | 0.0000 |
| free.sulfur.dioxide | 0.0046 | 0.0022 | 2.13 | 0.0335 |
| total.sulfur.dioxide | -0.0034 | 0.0007 | -4.89 | 0.0000 |
| pH | -0.5465 | 0.1332 | -4.10 | 0.0000 |
| sulphates | 0.8969 | 0.1104 | 8.12 | 0.0000 |
| alcohol | 0.2917 | 0.0172 | 16.95 | 0.0000 |

| | VIF |
|---|---|
| volatile.acidity | 1.65 |
| citric.acid | 2.18 |
| residual.sugar | 1.09 |
| chlorides | 1.37 |
| free.sulfur.dioxide | 1.94 |
| total.sulfur.dioxide | 2.01 |
| pH | 1.61 |
| sulphates | 1.33 |
| alcohol | 1.28 |

As a result, there was not significant change to the $R^2_{adj}$ value, as it went from 0.3561 to 0.3565. The VIF values for all the variables in the new model are low like we want them to be. However, since the removal of fixed acidity and density made little change to our model's goodness-of-fit, we will keep these variables in the dataset moving forward. It is certainly important to keep these findings in mind when using statistical methods later in this report, as it will help us understand why these predictors may or may not be included.

We need some way to decide on whether or not a wine is good. We can achieve this by assigning some threshold to the quality variable. For the purpose of this report, I will say that any data point with a wine quality of 6 or over can be considered a "good" wine. To do this, I add a new wine quality indicator variable to the dataset, with 1 representing that a wine is good (quality greater than or equal to 6) and 0 indicating otherwise. Thus, we move forward with our dataset consisting of 1,599 rows still and now 13 columns.

## Section 3 : Exploratory Data Analysis

In this section, we will be diving into the data itself and looking for relationships between variables and other information that will help us in our analysis of good wine quality. As we saw in the previous section, density and fixed acidity gave us something to think about. We ended up keeping these variables in our dataset, but now we can analyze any significant relationships between all of our predictors and quality. We can begin by first looking at the correlations between the variables.

The correlation plot displayed in Figure 1 shows us the correlation between all the variables in the dataset. We specifically want to look at the variables associated with quality in order to observe what factors are important in good wine. As we can see, quality seems to have a slightly strong correlation with volatile acidity and alcohol and some correlation with citric acid and sulphates. For the other variables, we can see a relatively strong correlation between fixed acidity and citric acid, fixed acidity and density, and fixed acidity and pH. Additionally, alcohol and density are negatively correlated, as is citric acid with volatile acidity and pH. We can create a table to identify the specific correlation coefficients for the variables we are interested in.

| | fixed.acidity | volatile.acidity | citric.acid | residual.sugar | chlorides | free.sulfur.dioxide | total.sulfur.dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fixed.acidity | 1.00 | -0.26 | 0.67 | 0.11 | 0.09 | -0.15 | -0.11 | 0.67 | -0.68 | 0.18 | -0.06 | 0.12 |
| volatile.acidity | -0.26 | 1.00 | -0.55 | 0.00 | 0.06 | -0.01 | 0.08 | 0.02 | 0.23 | -0.26 | -0.20 | -0.39 |
| citric.acid | 0.67 | -0.55 | 1.00 | 0.14 | 0.20 | -0.06 | 0.04 | 0.36 | -0.54 | 0.31 | 0.11 | 0.23 |
| residual.sugar | 0.11 | 0.00 | 0.14 | 1.00 | 0.06 | 0.19 | 0.20 | 0.36 | -0.09 | 0.01 | 0.04 | 0.01 |
| chlorides | 0.09 | 0.06 | 0.20 | 0.06 | 1.00 | 0.01 | 0.05 | 0.20 | -0.27 | 0.37 | -0.22 | -0.13 |
| free.sulfur.dioxide | -0.15 | -0.01 | -0.06 | 0.19 | 0.01 | 1.00 | 0.67 | -0.02 | 0.07 | 0.05 | -0.07 | -0.05 |
| total.sulfur.dioxide | -0.11 | 0.08 | 0.04 | 0.20 | 0.05 | 0.67 | 1.00 | 0.07 | -0.07 | 0.04 | -0.21 | -0.19 |
| density | 0.67 | 0.02 | 0.36 | 0.36 | 0.20 | -0.02 | 0.07 | 1.00 | -0.34 | 0.15 | -0.50 | -0.17 |
| pH | -0.68 | 0.23 | -0.54 | -0.09 | -0.27 | 0.07 | -0.07 | -0.34 | 1.00 | -0.20 | 0.21 | -0.06 |
| sulphates | 0.18 | -0.26 | 0.31 | 0.01 | 0.37 | 0.05 | 0.04 | 0.15 | -0.20 | 1.00 | 0.09 | 0.25 |
| alcohol | -0.06 | -0.20 | 0.11 | 0.04 | -0.22 | -0.07 | -0.21 | -0.50 | 0.21 | 0.09 | 1.00 | 0.48 |
| quality | 0.12 | -0.39 | 0.23 | 0.01 | -0.13 | -0.05 | -0.19 | -0.17 | -0.06 | 0.25 | 0.48 | 1.00 |

Supporting our findings from the correlation plot, we get a coefficient of 0.48 between quality and alcohol and a value of -0.39 between quality and volatile acidity. When looking at fixed acidity and density, we see a small, positive correlation between quality and fixed acidity and a small, negative correlation between quality and density. This supports our insights from the last section where we determined that these two predictors may not be of big significance for our model. From these observations, we can say that higher

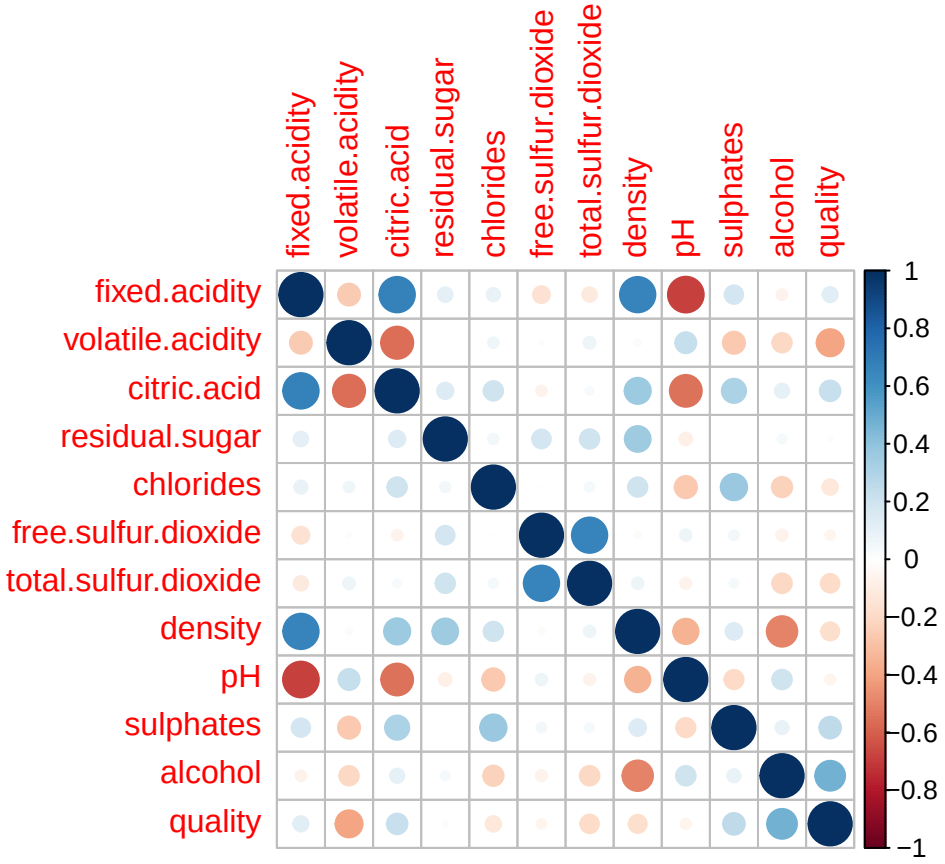


Figure 1: Correlation Plot

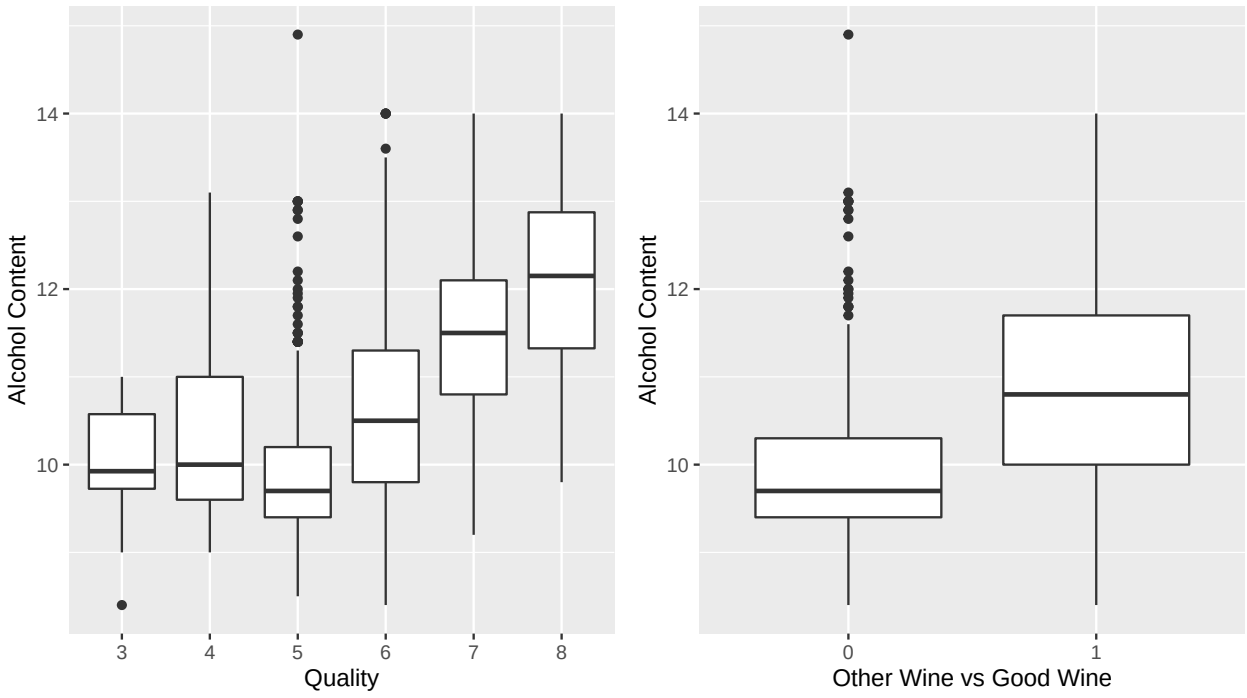


Figure 2: Boxplot of Alcohol Content and Quality

alcohol percentages lead to a higher quality of wine, and higher volatile acidity results in poorer wine quality. We also see that citric acid and sulphates have correlation coefficients of 0.23 and 0.25 respectively with quality. With these variables in mind, we can visually analyze any trends found between these variables and the quality of wine using our newly added quality indicator variable. We can also observe the rest of the variables to see if we find other patterns not discovered through correlation.

In Figure 2, we see a side-by-side boxplot of the alcohol content for all the different measures of wine quality, as well as the boxplots for the wines that are considered "good" based on the indicator. In the left boxplot, our claim is supported that lower alcohol leads to lower quality measures. From the right boxplot, we can conclude that there is a general trend in better quality with higher alcohol content.

Figure 3 shows us another pattern between volatile acidity and wine quality as supported by our findings from Figure 1. Looking at all the individual recorded qualities, as well as the wines that were marked "good," both boxplots display the similar trend that lower volatile acidity indicates better wine quality. This makes sense as in general, high volatile acid levels could cause a wine to go bad. It leads to a sharp, nail polish smell, and also causes the wine to have a vinegar-like taste.

In Figure 4, we see the boxplots of sulphates and quality. Sulphate represents the sulfur dioxide levels ($SO_2$) in the wine. It is an additive used in the winemaking process that acts as an antioxidant and antibacterial. As supported by the positive correlation coefficient, we see a general upward trend in quality as sulphate increases. We do tend to see a lot of outliers, but within the first and third quartiles of the data, there is not much of a difference in values between the good wine and the other data points.

The last of our variables of interest is citric acid. We can observe the boxplots of citric acid and wine quality in Figure 5. There is similar variability throughout each of the boxplots for each quality score, but we do see a general upward trend in the median of the citric acidity levels as quality increases. We also see that in general, good wines have higher levels of citric acid. This physicochemical index adds freshness and a nice taste to wines, which makes sense that is positively correlated with good wine.

From these plots, we can conclude that good wine quality is associated with high alcohol content, high sulphate levels, high citric acidity, and low volatile acidity. Looking back at the table of correlation coefficients,

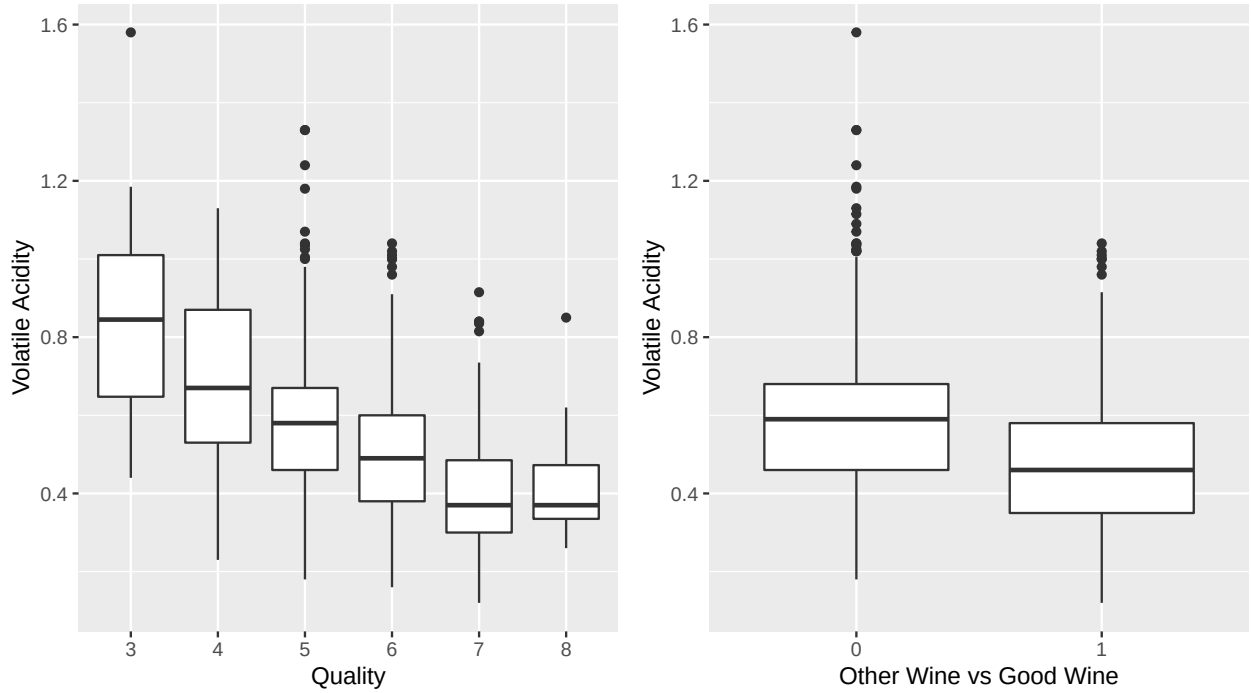


Figure 3: Boxplot of Volatile Acidity and Quality

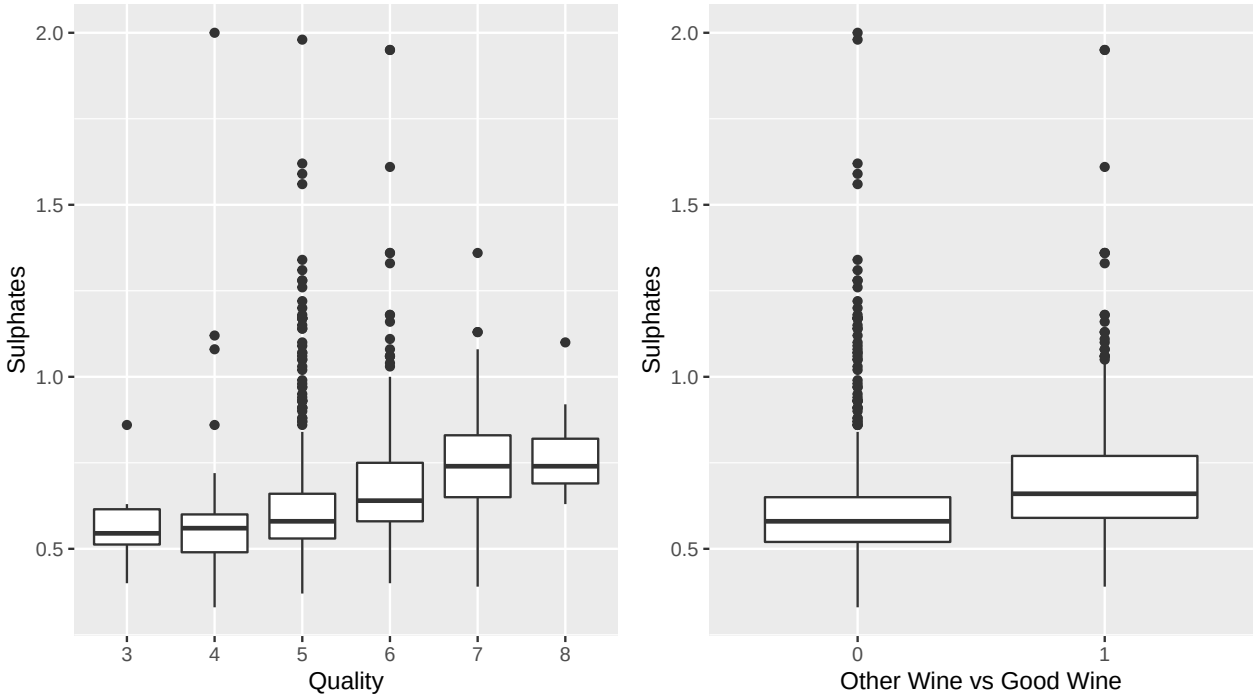


Figure 4: Boxplot of Sulphates and Quality

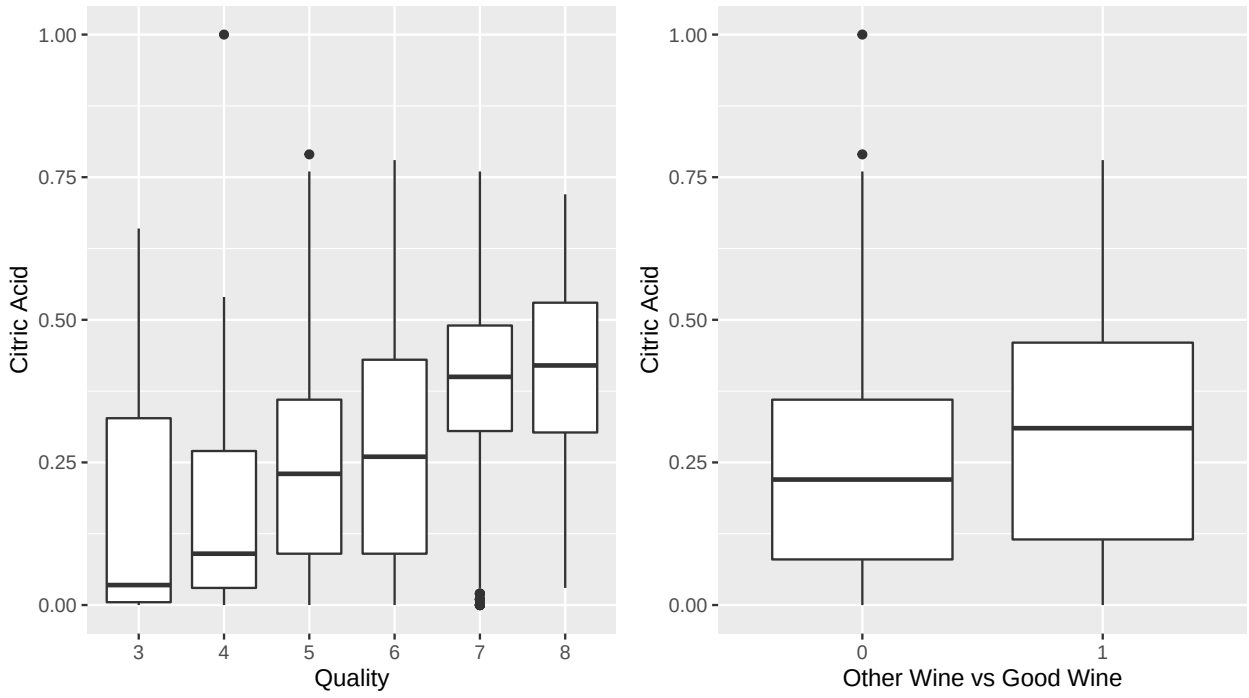


Figure 5: Boxplot of Citric Acid and Quality

variables such as residual sugar, free sulfur dioxide, and pH had very little to no relationship to quality. We can now check some other variables that had small correlations with wine to see if there are any other conclusions we can draw about wine quality. For example, we can look back at fixed acidity and density to either confirm or deny our former claims.

We can see the relationship between quality and the rest of our predictors in Figure 6. We see the relationship of these variables with the good wines versus other wines, followed by the boxplots of each degree of wine quality. First looking at fixed acidity and density, there is some variation when observing the boxplots with quality on the x-axis, but overall, there is not clear pattern. As we suspected earlier, these two variables don't have a significant impact on the quality of the wine. We can all see this in the boxplots using the wine quality indicator. The same can be said about the relationship between quality and total sulfur dioxide and chlorides.

Based on our findings through exploratory data analysis, the correlation plot and the boxplots created give us a good indication of which physicochemical indexes have a lot of influence on wine quality. Among the most important factors are alcohol content, volatile acidity, citric acid, and sulphates. We saw that the rest of the predictors either had little to no correlation, or we observed in their respective boxplots that there was not much difference throughout each wine quality rating.

## Section 4 : Method 1 - Best Subset Selection

The first statistical method I will be using to help determine what physicochemical indexes positively contribute to the quality of wine is **best subset selection**, or **BSS**. This procedure involves fitting several models trying different combinations of all the predictor variables. The goal of BSS is to fit the best model with a combination of predictors that yields the best statistical measure. With this method, we can look at the $C_p$ and $R^2_{adj}$ values. Mallows $C_p$ measures the fit of regression, meaning a good value represents strong coefficient estimation and predictive ability. In general, we want this value to be small for a model. The $R^2_{adj}$ indicates how well the data fits a line. We use this technique because it will choose the best combination of predictors which will then allow us to make accurate prediction with a well-fit model.

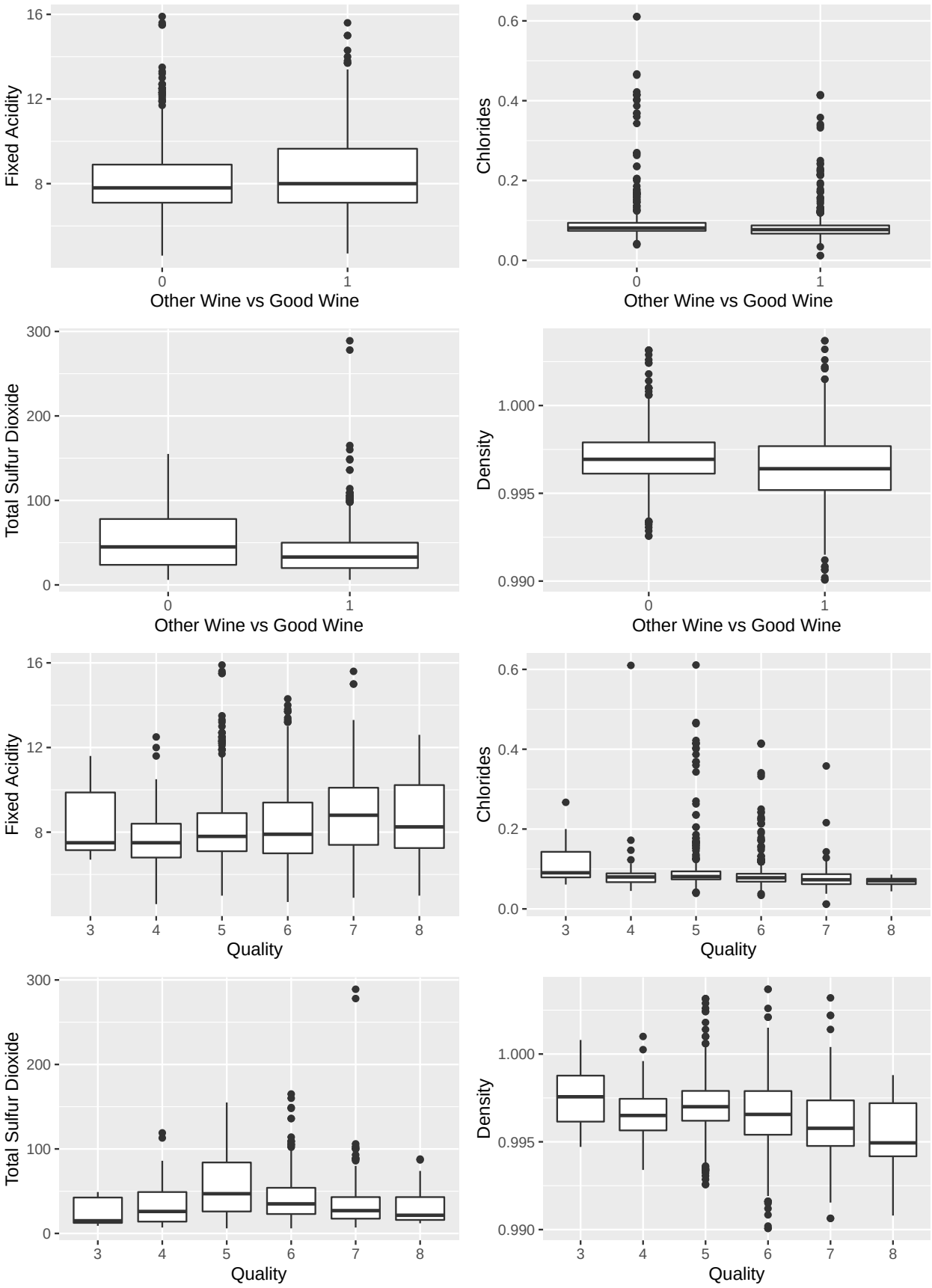


Figure 6: Boxplots of Quality with Fixed Acidity, Chlorides, Total Sulfur Dioxide, and Density

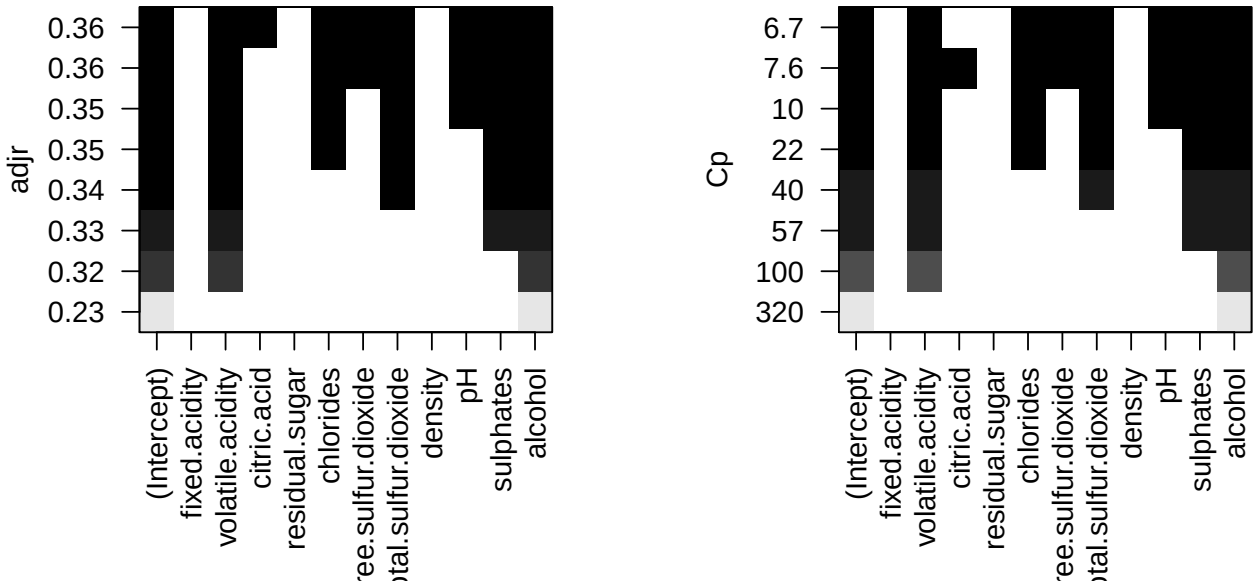


Figure 7: Best Model based on Adjusted $R^2$ and $C_p$

From a mathematical perspective, there are several steps associated with running best subset selection. The first thing to note is that the number of models that we will fit is related to the number of predictors. It can be represented as $2^p$ models. The procedure itself begins with setting the null model with 0 predictors to $M_0$. Then, we fit models using $\binom{p}{n}$ predictors with $n = 1...p$ total predictors. For instance, $M_1$ is represented with all the models choosing $\binom{p}{1}$ predictors, $M_2$ is the model choosing $\binom{p}{2}$ predictors, and so on. When forming these models, we want to choose the best model based on some statistical criteria. The most common measures to find and compare with other models are the smallest residual sum of squares or highest $R^2_{adj}$ value. We repeat this process until we find the model that fits $\binom{p}{n}$ predictors and yields the best statistical metric of interest [3].

In order to be certain on which factors are related to a higher wine quality, we can run the BSS command using both linear and logistic regression. This will involve using the continuous quality variable for the linear model and the quality indicator variable for the logistic model. While we did see from exploratory data analysis what variables most closely relate to good wine quality, we will use all the predictors when running the BSS command.

We begin by running the BSS command using quality as the response variable. For this report, the command will gather all the predictors to fit the model to produce the best $C_p$ and $R^2_{adj}$ value. We can analyze a plot that shows us the resulting models that meet this criteria.

As a result, we can see in Figure 7 that the two selected models that produce the best respective statistical measure are nearly identical. The model that has the highest $R^2_{adj}$ value includes the coefficients volatile acidity, citric acid, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol. The model with the lowest $C_p$ value has the same predictors except citric acid is removed. We can then look at the coefficients associated with these two models.

| | Model with best $R^2_{adj}$ |
|---|---|
| (Intercept) | 4.668 |
| volatile.acidity | -1.074 |
| citric.acid | -0.130 |
| chlorides | -1.950 |
| free.sulfur.dioxide | 0.005 |
| total.sulfur.dioxide | -0.003 |
| pH | -0.549 |
| sulphates | 0.891 |
| alcohol | 0.293 |

| | Model with best $C_p$ |
|---|---|
| (Intercept) | 4.430 |
| volatile.acidity | -1.012 |
| chlorides | -2.018 |
| free.sulfur.dioxide | 0.005 |
| total.sulfur.dioxide | -0.003 |
| pH | -0.483 |
| sulphates | 0.883 |
| alcohol | 0.289 |

At first glance, the coefficients are very close in value, even with one model having one more predictor than the other. It is important to note from Figure 7 that citric acid was added at a $C_p$ value of 7.6 but was then removed again when the value reached 6.7. A difference of 0.9 is not too significant, but another observation we can make is that a $R^2_{adj}$ value of 0.36 is obtained whether or not citric acid is added or not.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.6681 | 0.4608 | 10.13 | 0.0000 |
| volatile.acidity | -1.0736 | 0.1159 | -9.26 | 0.0000 |
| citric.acid | -0.1295 | 0.1218 | -1.06 | 0.2876 |
| chlorides | -1.9494 | 0.4027 | -4.84 | 0.0000 |
| free.sulfur.dioxide | 0.0048 | 0.0021 | 2.22 | 0.0267 |
| total.sulfur.dioxide | -0.0034 | 0.0007 | -4.84 | 0.0000 |
| pH | -0.5492 | 0.1331 | -4.12 | 0.0000 |
| sulphates | 0.8914 | 0.1102 | 8.09 | 0.0000 |
| alcohol | 0.2929 | 0.0171 | 17.10 | 0.0000 |

When observing the summary of the two fitted models, the $R^2_{adj}$ value was identical at 0.3567. Additionally, we see from the summary that citric acid has a p-value of 0.2876. This means that we would fail to reject a null hypothesis that the coefficient for citric acid is 0, as it is not statistically significant. With this finding and the support from our plots, we could say that leaving out the citric acid predictor would lead to the better model. One more way that will help us make a final decision is to find the test MSE of the models with and without citric acid. This would tell us whether or not the inclusion of this variable will result in stronger predictive ability.

When computing the MSE for the two models, we barely get a difference in value. The model that includes citric acid has a test MSE of 0.419536 while the model without the predictor has an MSE of 0.4195707. That is just a difference of 0.0000347, which I believe really won't make a significant change in the predictive ability. One last metric we can find in order to compare our model with the model created in the next section is the AIC, which represents for error when making any outside-of-sample predictions. In general, we want this value to be small.

As a result, we get an AIC of 3159.839 for the model with citric acid and 3158.977 for the model without. Again, we see a very small difference in value between the values, but the model that leaves citric acid out

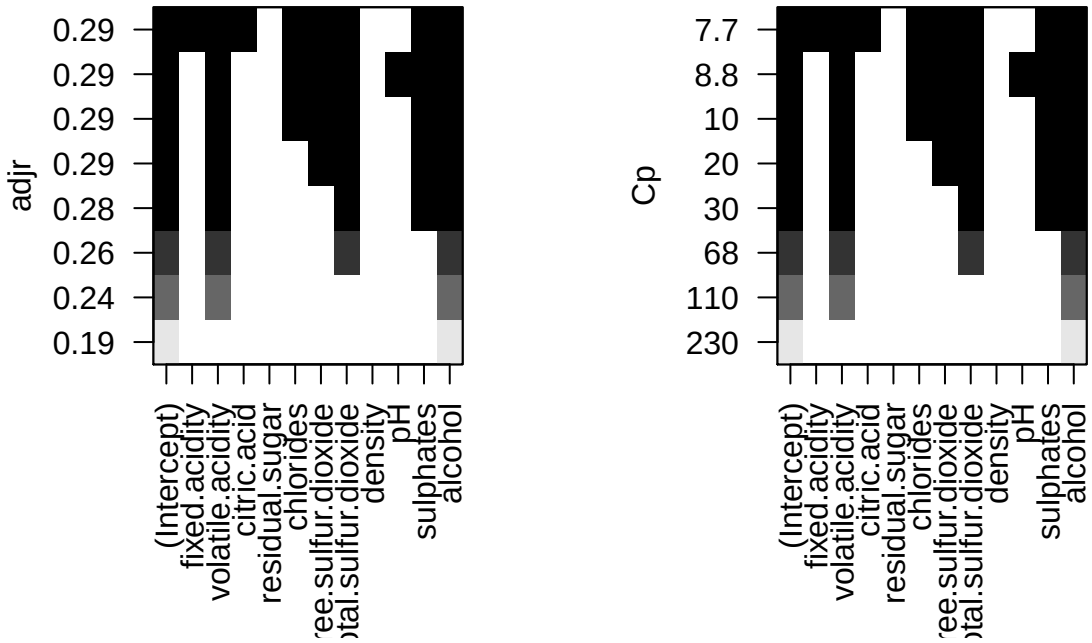


Figure 8: Best Model based on Adjusted $R^2$ and $C_p$

has a better value.

With all of our findings we can make a choice on our model. We saw that adding citric acid yields a lower MSE, but the difference between the models is almost not noticeable. In addition, the model without citric acid yields a lower AIC value. We also saw that the p-value of the citric acid coefficient was pretty high. Figure 7 showed us that there is no difference in the $R^2_{adj}$ value when citric acid is included or not, as well as the fact that excluding citric acid yields the lowest $C_p$. Thus, for our linear model using quality as the predictor, our final model is fit as follows:

$$\begin{aligned}\widehat{Quality} = 4.430 - 1.013\ Volatile\ Acidity - 2.018\ Chlorides + 0.0051\ Free\ Sulfur\ Dioxide \\ -0.0035\ Total\ Sulfur\ Dioxide - 0.483\ pH + 0.883\ Sulphates + 0.289\ Alcohol\end{aligned}$$

Now we can switch our focus to our other quality variable. With the intent of using logistic regression by setting the quality indicator as the response variable, we can run BSS again. To see if we get a different model in terms of the chosen predictors.

Based on Figure 8, we see that the same model has the highest $R^2_{adj}$ value and lowest $C_p$ value. We can already notice that these measures are weaker in this model than the model we choose for our linear model. Using logistic regression with quality indicator as the response variable gives us a model that includes the variables fixed acidity, volatile acidity, citric acid, chlorides, free sulfur dioxide, total sulfur dioxide, sulphates, and alcohol. We can fit this model and compare it to our linear model.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -9.2169 | 0.9500 | -9.70 | 0.0000 |
| fixed.acidity | 0.1273 | 0.0511 | 2.49 | 0.0127 |
| volatile.acidity | -3.3799 | 0.4780 | -7.07 | 0.0000 |
| citric.acid | -1.2604 | 0.5610 | -2.25 | 0.0247 |
| chlorides | -3.5291 | 1.5091 | -2.34 | 0.0194 |
| free.sulfur.dioxide | 0.0221 | 0.0082 | 2.70 | 0.0070 |
| total.sulfur.dioxide | -0.0156 | 0.0028 | -5.57 | 0.0000 |
| sulphates | 2.6863 | 0.4326 | 6.21 | 0.0000 |
| alcohol | 0.9054 | 0.0734 | 12.33 | 0.0000 |

As we can see, we get a fitted logistic regression model of:

$$\widehat{Quality\ Indicator} = -9.217 + 0.127\ Fixed\ Acidity - 3.380\ Volatile\ Acidity - 1.260\ Citric\ Acid - 3.529\ Chlorides + 0.022\ Free\ Sulfur\ Dioxide - 0.016\ Total\ Sulfur\ Dioxide + 2.686\ Sulphates + 0.905\ Alcohol$$

From the model summary, we get an AIC of 1675.8, which is significantly lower than the linear model. However, we did just see a lower $R^2_{adj}$ value and higher $C_p$ value from Figure 8. We should turn to the MSE in order to decide on what model is better.

Computing the MSE gives us a value of 48.92565. This is clearly a lot higher than our chosen linear model's low MSE of 0.41957. It is also important to notice that our logistic regression model includes some variables that, supported by my statistical measures and visualizations, were deemed not influential predictors of good wine quality. These included fixed acidity and citric acid. We can see that the negative citric acid coefficient in the model represents an inverse relationship to quality. This contradicts our findings from previous sections that stated that increasing citric acidity led to an increase in wine quality.

To summarize all of our results, I believe the final model I will choose is the following linear model:

$$\widehat{Quality} = 4.430 - 1.013\ Volatile\ Acidity - 2.018\ Chlorides + 0.0051\ Free\ Sulfur\ Dioxide - 0.0035\ Total\ Sulfur\ Dioxide - 0.483\ pH + 0.883\ Sulphates + 0.289\ Alcohol$$

I first came to this conclusion based on the resulting visualizations created from using the best subset selection technique. The combination of predictors in the model above were chosen using this method, as it was associated with the highest $R^2_{adj}$ value of 0.3567, lowest $C_p$ value of 6.7, and lowest AIC value of 3158.977. We observed that the model including citric acid produced a lower test MSE, but we determined that a difference of roughly 0.00003 between the two models was not enough evidence to choose that model over the model without citric acid. Compared to the logistic regression model, we saw that the linear model had better statistical metrics apart from AIC. The most significant difference between the linear and logistic model was the test MSE, with the linear model being much more superior in that regard. Thus, I can say that our chosen model would be the best goodness-of-fit and prediction ability.

# Section 5 : Method 2 - Random Forest

The second method I will be using to help us decide which physicochemical indexes contribute to good wine quality is **random forest**. This is a common technique for fitting multiple tree-based models used for classification and regression. The advantage of creating multiple trees versus fitting just a single tree is that we can reduce overall variance and bias within each tree. The process for random forest is very similar to bagging, another multiple-tree creating method. The aspect that differentiates bagging and random forest is the selection of predictors. While bagging considers all the data's predictors when fitting models, random forest selects a random sample of predictors for each split. In practice, the most common way to determine our number of variables is out of all $p$ predictors, we select $m \approx \sqrt{p}$ predictors. This produces a benefit over bagging in that random forest decorrelates the trees. This helps reduce the possibility of having the response variable being highly correlated with one of our predictors.

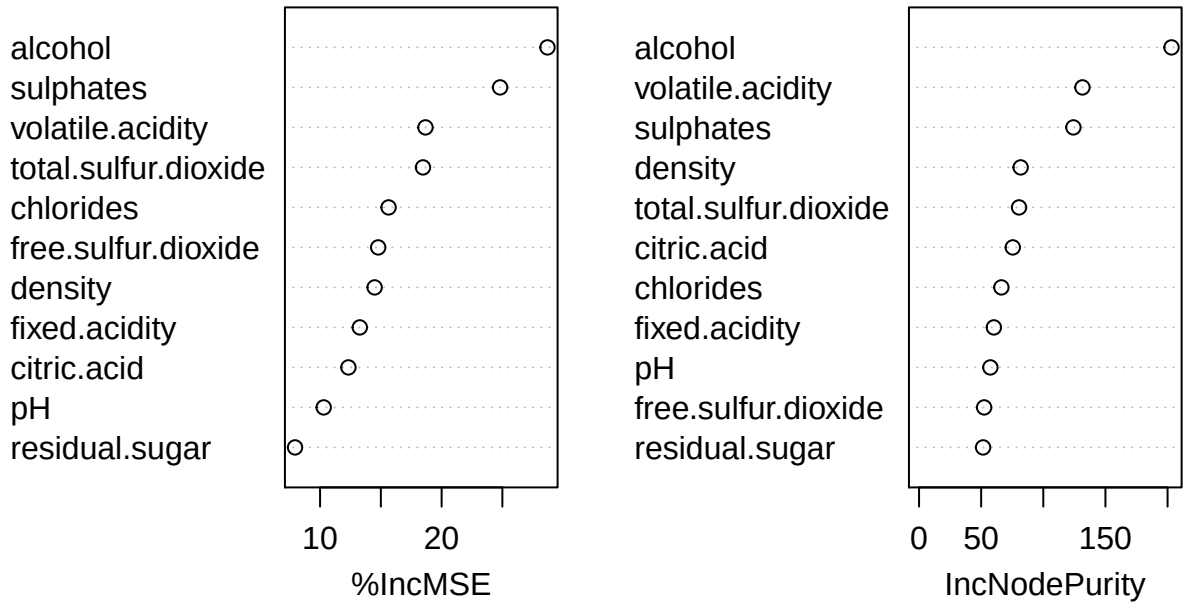


Figure 9: Plot of Variable Importance to Wine Quality

When fitting our model, I will be skipping over the quality indicator variable I created since I will be not creating classification trees. Instead, I will just be considering regression trees using quality as the response variable. The mathematics behind random forest follows similar steps as bagging. The first step is to select a number of $B$ bootstrap samples. We generally we want a large quantity of samples, so I will be choosing 100. Next, we randomly choose $\sqrt{p}$ predictors, choose the first variable to split at using recursive splitting, and then repeat the process with new samples of predictors until we reach the stopping point. With all the bootstrapped samples we acquire, we "train our method" [1], average the predcitions, and we get the following formula [1]:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

This represents the regression function that can predict outcomes based on our bootstrap samples.

Now we can fit our random forest model using quality as our response model. We fit 100 trees using $m = \sqrt{p}$ predictors. We can visualize our forest and make some observations about the predictors by looking at importance plots and partial plots. We'll begin by looking at the variable importance plot.

From Figure 9, we see two measures of importance: the mean decrease of prediction accuracy and the decrease in node impurity. This plot tells us that across all of created trees, we see that three of the most important variables that positively contribute to wine quality are alcohol, volatile acidity, and sulphates. This makes sense, as we have seen through data analysis and our best subset selection model that these characteristics of wine are among the most influential variables in wine quality. We can also confirm some of our other observations like how residual sugar and pH are not positive contributors to good wine quality. Of our variables of interest, we can also look at partial plots. These plots show the individual effect that a predictor has on the response variable and can really demonstrate the relationship of each variable as it relates to the

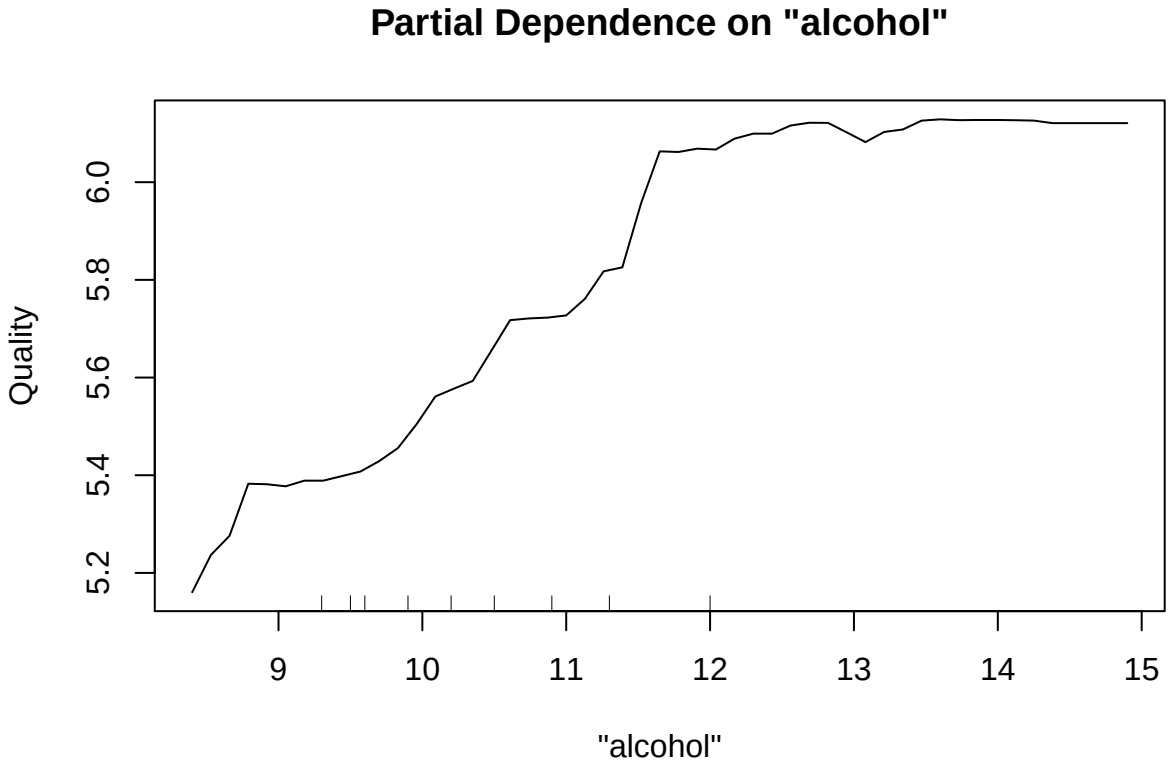


Figure 10: Partial Dependence of Alcohol

response variable.

The first variable we look at is alcohol, as we can reasonably conclude that is the most important variable in this dataset based on the importance plot. Figure 10 shows us the trend of how the steady increase in alcohol percentage improves the wine quality. We saw the same pattern in Figure 2. There are some spikes here and there, but ovreall, we see that higher alcohol content is a strong indicator of good wine. Next, we can look at volatile acidity.

Figure 11 shows a negative relationship between volatile acidity and quality, which again confirms our findings from previous sections. Contrary to the plots we saw earlier, we get a better idea of the degree of volatile acidity that affects quality. We see that the quality flattens out when volatile acidity exceeds 1.0. This tells us that in general, we will see better wine quality when the volatile acidity is below 1. The last of the three most important variables we can look at is sulphates.

In Figure 12, we see the pattern of how the increase in sulphates in wine improves the quality. This shows us a pattern that were not expressed in other plots; we see that the quality shoots upward between sulphate levels of 0.5 and 1.0, after which it remains flat. So we can say when sulphates is less than 0.5, we will have a poor wine quality. When we have anything over 1, we can assume good quality.

Thus, from our random forest model, we get strong evidence that alcohol content, volatile acidity, and sulphates are among the most important variables. We made these exact same observations from our exploratory data analysis section. Additionally, we found that these variables were important enough to include in our model when using best subset selection. In order to determine how well our random forest model is at prediction we can look at the MSE value.

Looking at the the summary output of our random forest model, we get of 0.3209. We can see that in Figure 13, our MSE decreased as we created more trees and converges to about 0.32. Compared to the MSE of our model created through best subset selection, the random forest model's value is better.

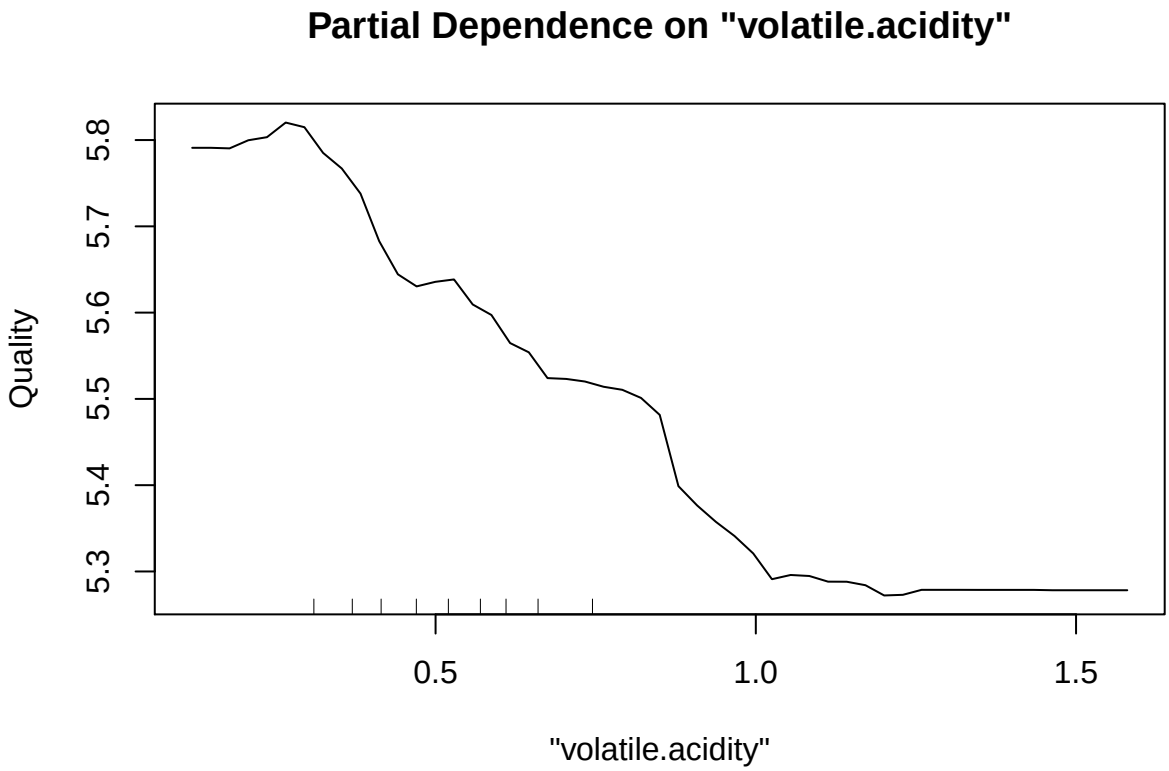


Figure 11: Partial Dependence of Volatile Acidity

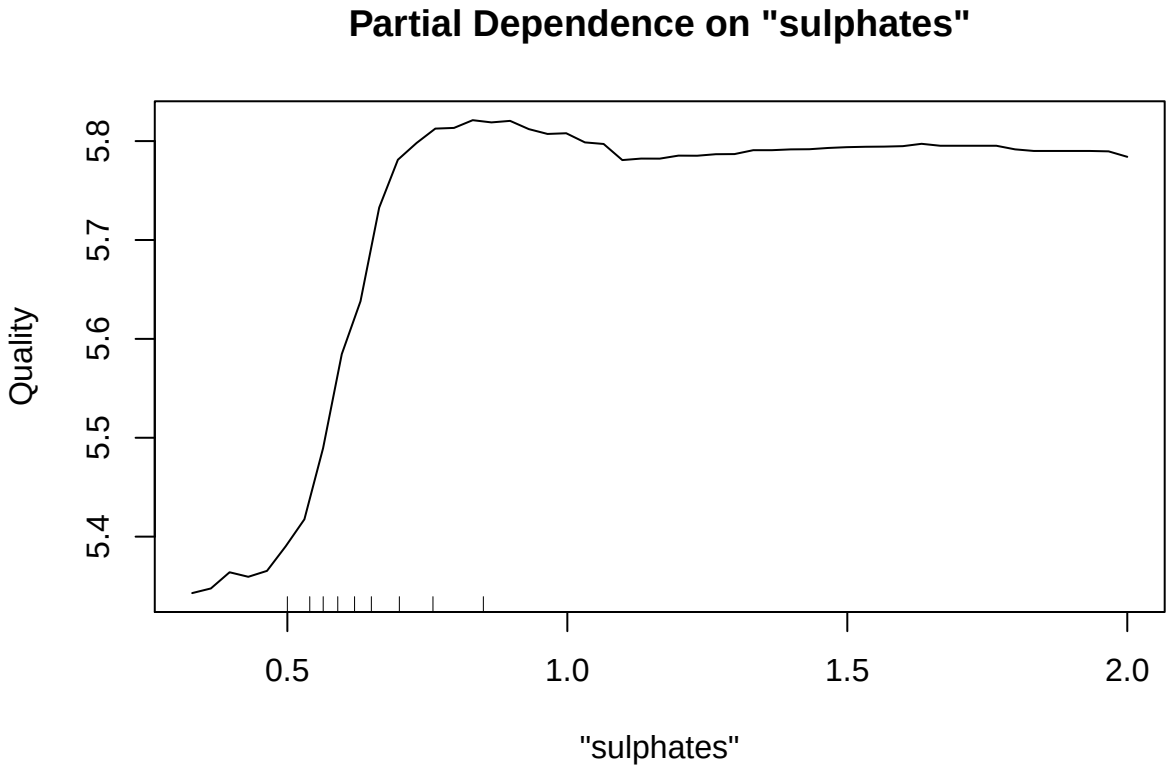


Figure 12: Partial Dependence of Sulphates


Figure 13: MSE as Number of Trees Increase

## Section 6 : Results

We considered predicting wine quality using two methods: best subset selection and random forest. Based on our findings from exploratory data analysis, we were able to determine that out of all the predictors, some of the most influential physicochemical indexes were alcohol, volatile acidity, sulphates, and citric acid. This helped us when working with our first method of best subset selection. We determined that citric acid was best left out of the final model, but we included the other three variables. Using $R^2_{adj}$ and $C_p$ values, we compared our model choices using both logistic and linear regression and ended up with a linear model with an MSE of 0.4196. We then moved on to our next method of random forest, where we fit a model and created several regression trees. Based on our visualizations, we gathered the same findings that alcohol, volatile acidity, and sulphates were the three most important variables in predicting good wine quality. When choosing between best subset selection and random forest, we looked to compare the test MSE. We found that we got a smaller MSE with our random forest. Thus, we can conclude that this is our best model in predicting wine quality.

$$MSE_{BSS} = 0.4196$$
$$MSE_{RF} = 0.3209$$

## Works Cited

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. (2017, November). Red Wine Quality. Version 2. Retrieved November 20, 2019 from https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009.

[3] Statistics - Best Subset Selection Regression [Gerardnico - The Data Blog]. https://gerardnico.com/data_mining/best_subset. Accessed 9 Dec. 2019.

[4] Yin, Shen, et al. "Quality Evaluation Based on Multivariate Statistical Methods." Mathematical Problems in Engineering, 2013, doi:10.1155/2013/639652.