**STAT 350 – Group 16**
**Professor: Derek Bingham**
**Report by:**
**McKeowen Watts**
**Eui Jeong (Stephanie) Chung**
**Surbhi Negi**
**Tedmond Christo**

# Exploratory Data and Regression Analysis

## *Wine Quality data set*

# Contents

# Abstract

This exploratory data analysis will be examining data pertaining to the red and white varieties of the Portuguese “Vinho Verde” wine. The data provided presents eleven unique physicochemical characteristics of the wines. Using these independent variables and armed with the knowledge we gained in STAT 350, we will attempt to determine the relationship (if any) that these variables have with the dependent variable quality. Our analysis will include graphical representation of the data, a linear regression model to determine the significance of each variable, and residual analysis to understand the intricacies of the data. Once an initial model is fit, we performed variable selection to determine the most optimal model, and cross validation to test its efficacy.

# Introduction

Wine is an alcoholic beverage made from grape varieties. It is consumed by many people all around the world, and praised for its versatility. Additionally, there are countless different varieties, meaning there is a unique type for each unique person!

At the physicochemical level, wines don’t seem to be all that different. For all unique different varieties, a similar chemical formula is followed. The natural sugars present in grapes are combined with yeast, and the fermentation process begins. The amount of sugar available for the fermentation process is a result of the age and species of grapes. As the grape ages and ripens, the natural sugar levels rise. As the yeast consumes sugar, carbon dioxide and ethanol are released. This ethanol gives the wine its alcoholic content. From there, the wine is stored in barrels and aged. Depending on the temperature of the wine during this fermentation process, the length of the fermentation varies from one to two weeks. During this time, the natural acids present in the wine are mellow, and the flavour begins to really develop.

The physicochemical variables present in wine are all a product of this fermentation process. Depending on the type of grape, the length of fermentation, and many other variables, the tangible values for the chemical variables change.

But what truly defines the quality of wine? ***Can the quality of a wine be determined just from its physicochemical make up? Or do we need more information to be able to determine a good wine from a bad wine?*** In this project, we will conduct a multiple linear regression on wine quality datasets to get an answer to this question.

# Data Description

We obtained the wine quality dataset from the UCI Machine Learning Repository.

Before initiating our data analysis, we should take a more careful look at the datasets. Red wine data contains 12 variables and 1599 observations, while white wine data contains 12 variables and 4898 observations. There are no missing values for both datasets.

| | fixed.acidity <dbl> | volatile.acidity <dbl> | citric.acid <dbl> | residual.sugar <dbl> | chlorides <dbl> | free.sulfur.dioxide <dbl> | total.sulfur.dioxide <dbl> | density <dbl> | pH <dbl> | sulphates <dbl> | alcohol <dbl> | quality <int> |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11 | 34 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |
| 2 | 7.8 | 0.88 | 0.00 | 2.6 | 0.098 | 25 | 67 | 0.9968 | 3.20 | 0.68 | 9.8 | 5 |
| 3 | 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15 | 54 | 0.9970 | 3.26 | 0.65 | 9.8 | 5 |
| 4 | 11.2 | 0.28 | 0.56 | 1.9 | 0.075 | 17 | 60 | 0.9980 | 3.16 | 0.58 | 9.8 | 6 |
| 5 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11 | 34 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |
| 6 | 7.4 | 0.66 | 0.00 | 1.8 | 0.075 | 13 | 40 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |

**Table 1.** First 6 observations of red wine data

| | fixed.acidity <dbl> | volatile.acidity <dbl> | citric.acid <dbl> | residual.sugar <dbl> | chlorides <dbl> | free.sulfur.dioxide <dbl> | total.sulfur.dioxide <dbl> | density <dbl> | pH <dbl> | sulphates <dbl> | alcohol <dbl> | quality <int> |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.0 | 0.27 | 0.36 | 20.7 | 0.045 | 45 | 170 | 1.0010 | 3.00 | 0.45 | 8.8 | 6 |
| 2 | 6.3 | 0.30 | 0.34 | 1.6 | 0.049 | 14 | 132 | 0.9940 | 3.30 | 0.49 | 9.5 | 6 |
| 3 | 8.1 | 0.28 | 0.40 | 6.9 | 0.050 | 30 | 97 | 0.9951 | 3.26 | 0.44 | 10.1 | 6 |
| 4 | 7.2 | 0.23 | 0.32 | 8.5 | 0.058 | 47 | 186 | 0.9956 | 3.19 | 0.40 | 9.9 | 6 |
| 5 | 7.2 | 0.23 | 0.32 | 8.5 | 0.058 | 47 | 186 | 0.9956 | 3.19 | 0.40 | 9.9 | 6 |
| 6 | 8.1 | 0.28 | 0.40 | 6.9 | 0.050 | 30 | 97 | 0.9951 | 3.26 | 0.44 | 10.1 | 6 |

**Table 2**. First 6 observations of white wine data

## Variable Description

| Variable | Description | Units |
|---|---|---|
| **Fixed acidity** | Fixed or nonvolatile acids involved with wine | tartaric acid – $g/dm^3$ |
| **Volatile acidity** | Amount of acetic acid in wine – large amounts may give wine a vinegary taste | acetic acid – $g/dm^3$ |
| **Citric acid** | Acid that adds freshness and flavor to wines, normally found in small quantities | $g/dm^3$ |
| **Residual sugar** | The amount of sugar remaining after fermentation stops | $g/dm^3$ |
| **Chlorides** | Amount of salt in the wine | sodium chloride - $g/dm^3$ |
| **Free sulfur dioxide** | Free form of sulfur dioxide that prevents microbial growth and oxidation of wine | $mg/dm^3$ |
| **Total sulfur dioxide** | Amount of free and bound forms of sulfur dioxide | $mg/dm^3$ |
| **Density** | Density of water depending on the percent alcohol and sugar content | $g/cm^3$ |
| **pH** | Acidity of wine on a scale from 0 to 14 – most wines are between 3 and 4 | pH |
| **Sulphates** | Wine additive that acts as an antimicrobial and antioxidant | potassium sulphate - $g/dm^3$ |
| **Alcohol** | Percent alcohol content of wine | % by volume |
| **Quality** | Quality of the wine, based on the chemical properties listed above | Integer value from 0-10 |

The first 11 variables are all measurable chemical amounts present in each wine. The final variable quality is our dependent variable, and is represented by an integer value.

# Data Visualization

Now that we know what our variables are, let us further examine and visualize the relation between the chemical properties (independent variables) and quality (dependent variable).

## *Histograms*

First, we will visualize the distribution of each variable in the red wine and white wine quality dataset.

### *Red wine data*

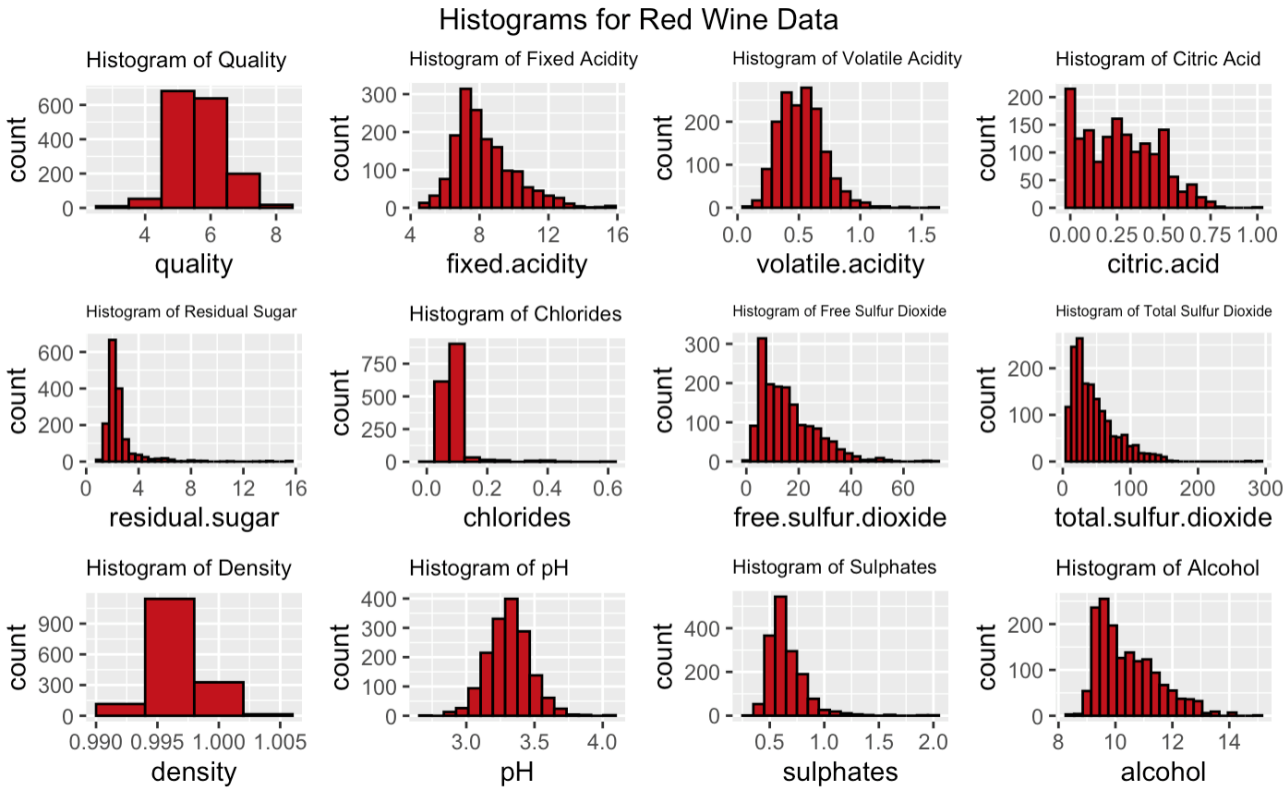



**Figure 1.** Histograms of variables from red wine data

From Figure 1, we observe that the quality variable is normally distributed with most of the observations lying on quality scales 5 and 6. The trend of the quality variable indicates that there are much more normal wines than excellent or poor ones.

Variables including residual sugar, chlorides, free sulfur dioxide, and total sulfur dioxide have a left-skewed distribution. However, it does seem plausible to obtain a normal distribution for most of the variables if we successfully detect and fix the outlier observations.

*White wine data*

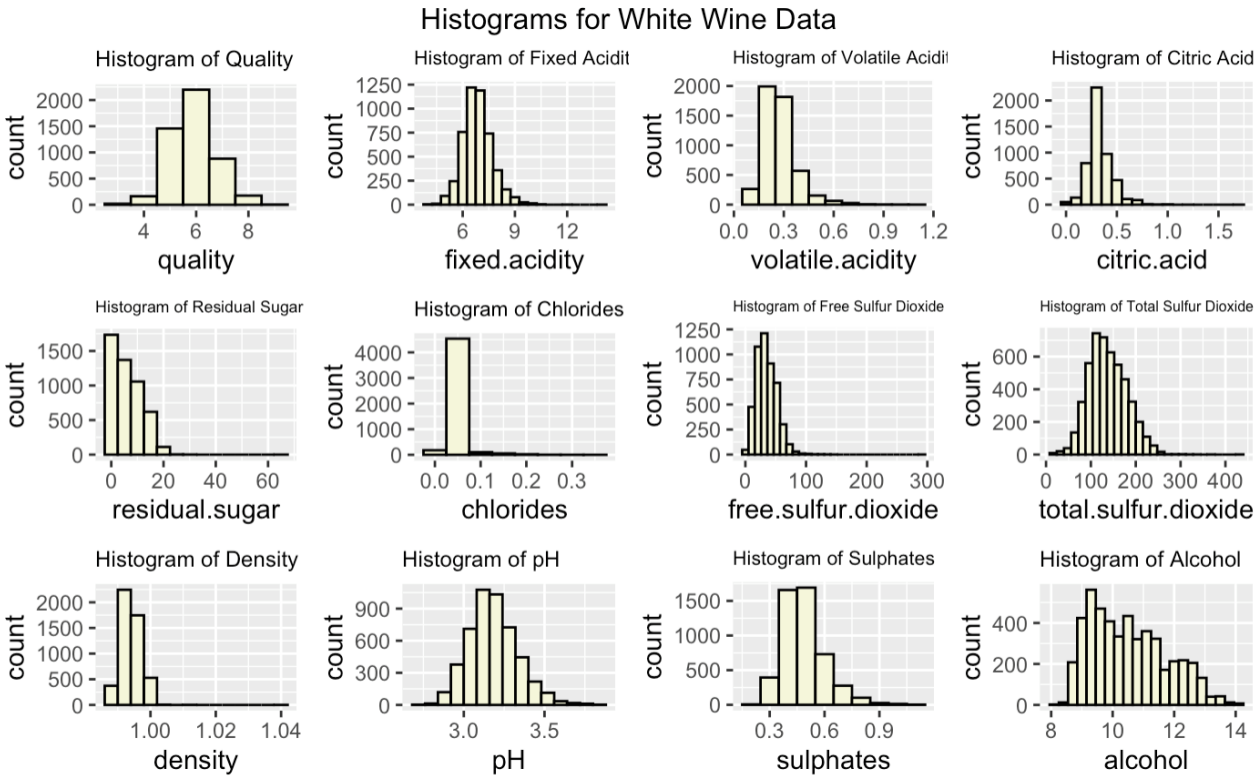


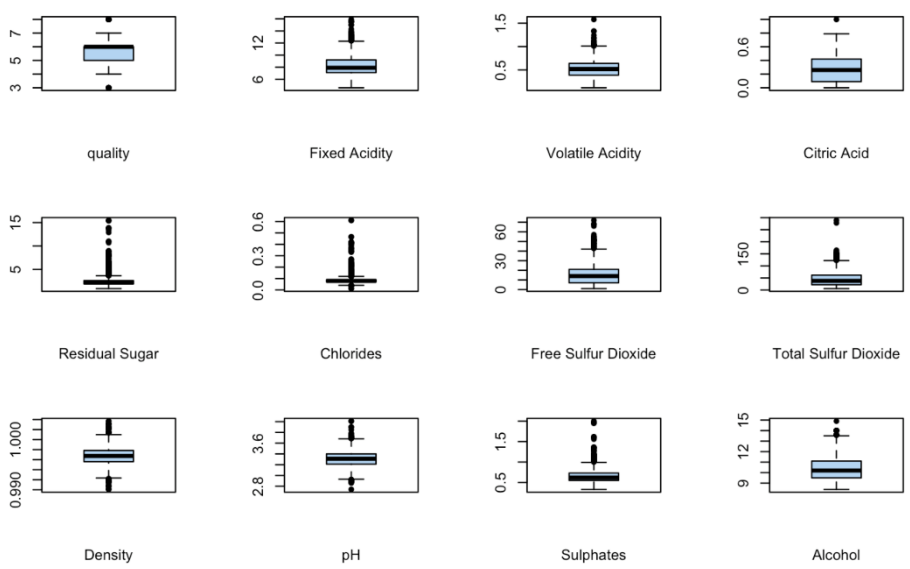

**Figure 2.** Histograms of variables from white wine data

Figure 2 indicates that quality variable has an approximate normal distribution with most values concentrated around the values of 5, 6, and 7.

Additionally, most of the variables have a very heavy right-tail, except pH and alcohol. However, some of the skewed data such as fixed acidity, volatile acidity, or citric acid could be normally distributed if the outliers were detected and fixed.

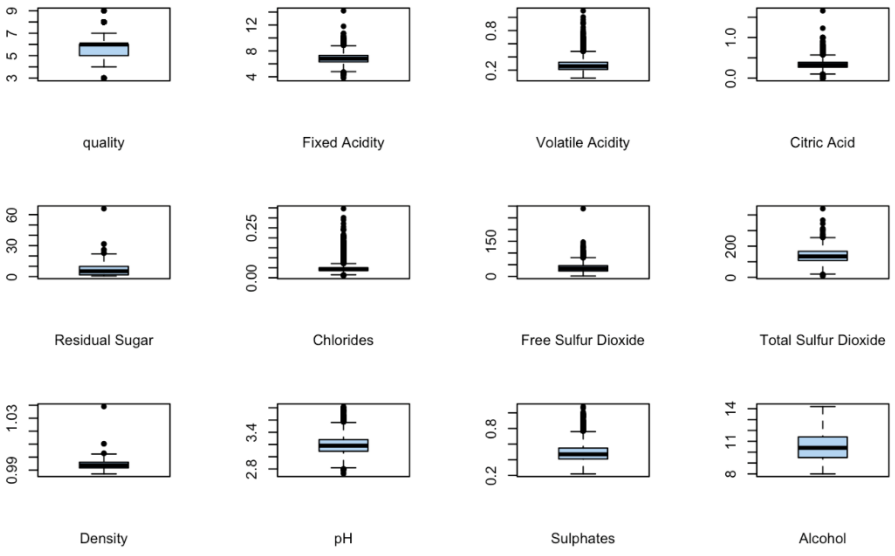
## *Box Plots*

Boxplots give us a visual idea of whether a point is within the interquartile range.




In the red wine dataset, we observe that citric acid, free sulfur dioxide, total sulfur dioxide, and alcohol are right-skewed. We can visually confirm from Figure 3 that most of the outliers have a greater value than the third quartile. In fact, we can clearly notice the additional small outlier that we added.

**Figure 3.** Boxplots of variables from red wine data




In Figure 4, we observe that residual sugar, density, and alcohol are right-skewed. Also, for most of the variables, detecting the outliers and accounting for them will make the distribution more symmetric. Just as before, most of the outliers are greater than the interquartile range.

**Figure 4.** Boxplots of variables from white wine data

## *Pairs Plots*

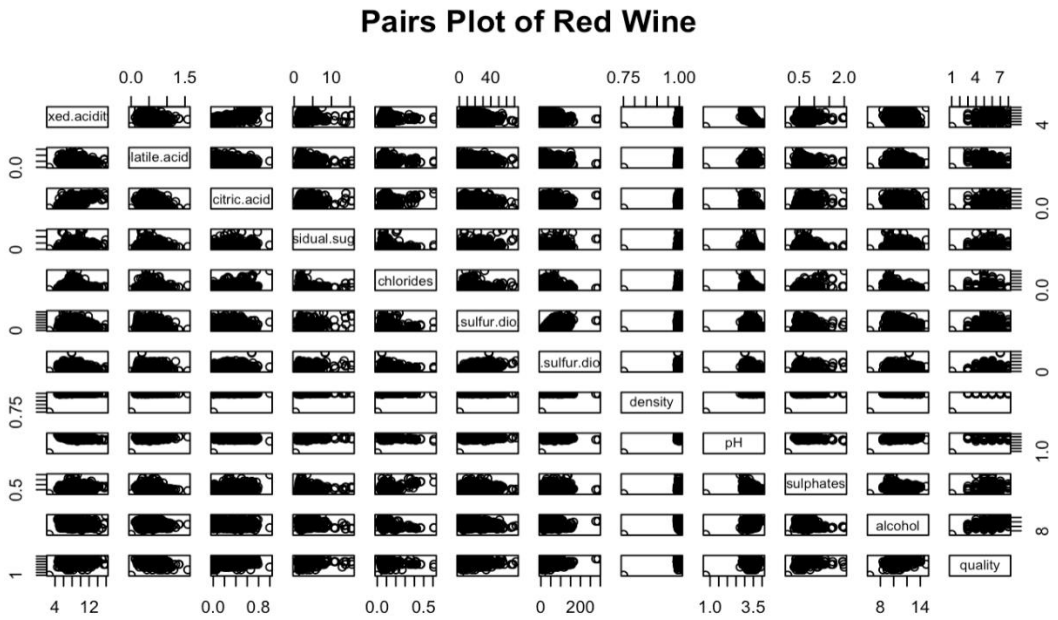



**Figure 5.** Pairs plot of red wine data

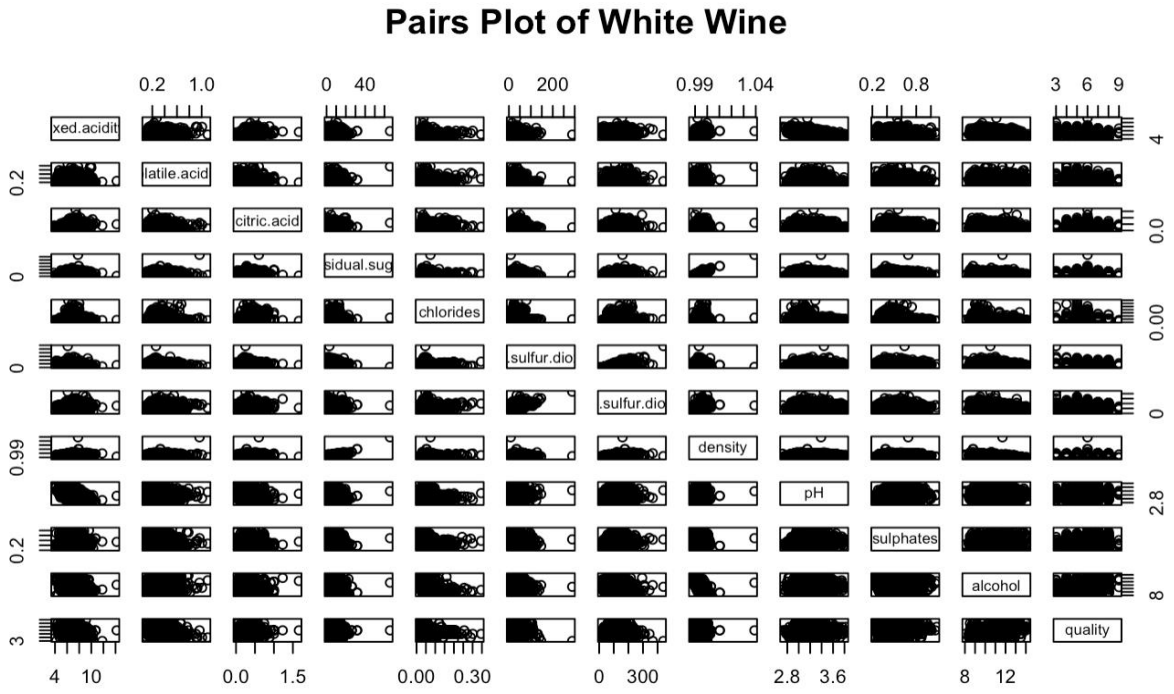



**Figure 6.** Pairs plot of white wine data

Figure 5 and Figure 6 suggest multicollinearity being present among independent variables. Fitting a linear model does not seem fully appropriate. However, we will try to fit a model and conduct hypothesis tests to confirm the adequacy of our model.

### *Additional data point*

Before we step into our main analysis, we will add an extra data point to each dataset.

For the **red wine** data, as we saw from the boxplots, most of the outliers are greater than the third quartile of the data. In other words, there are fewer outliers that are smaller than the minimum of each variable. In order to make the outliers of the dataset more diverse, we added a data point with very small values into the red wine data.

For the **white wine** data, we focused on the pairs plot and the presence of multicollinearity between different variables. Since the presence of multicollinearity among dependent variables may dramatically impact the usefulness of our model, we decided to add a point that will break multicollinearity. We applied the concept of centering the variables, which is one of the ways to fix multicollinearity. To construct our new data point, we took the mean and the third quartile of each variable. The mean was then subtracted from the third quartile for each of our 11 independent variables, and the resulting values were taken as our added data point.

Now that we have a good understanding of both datasets, we will conduct a multiple linear regression against the quality of the red and white wine.

# Methods

The analysis was done on red wine quality data and white wine quality data. After getting an overview of the data through visual representation, a linear regression model was fit to both datasets. Next, using the *summary()* command, we tested the significance of the overall model and individual regressors.

In the next part of the analysis, a diagnosis for model assumption and model adequacy was done using residual plots and Cook's distance. Transformations including log transformation and square root transformation were done to improve the low adjusted R-squared value on the revised model. Subset regression, forward selection, backward elimination, and stepwise regression variable selection tactics were used to find the best regression model for both datasets. Overall, each model that was constructed was compared based on their adjusted R-squared values, and the final models were obtained. Finally, residual analysis was done on the final models to verify that the models are adequate.

Finally, both datasets were separated into a train dataset and a test dataset to conduct cross validation. The results of the prediction were plotted and it was concluded that the prediction power of the models was mostly weak. Libraries and additional documentation of the analysis can be found in the appendix.

# Results

## Fitting the model

To answer the question of what makes wine good or bad, we first have to find out how each physicochemical level affects the quality of wines. We will begin our analysis by fitting the two datasets into a linear model.

```
Call:
lm(formula = quality ~ ., data = red.wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.65794 -0.36368 -0.04403  0.45568  2.05281

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.624e+01  3.049e+00  -5.328 1.14e-07 ***
fixed.acidity        -1.001e-02  1.746e-02  -0.573  0.56658
volatile.acidity     -1.111e+00  1.203e-01  -9.239  < 2e-16 ***
citric.acid          -1.845e-01  1.473e-01  -1.253  0.21053
residual.sugar        2.469e-04  1.214e-02   0.020  0.98377
chlorides            -1.928e+00  4.186e-01  -4.606 4.44e-06 ***
free.sulfur.dioxide   4.672e-03  2.166e-03   2.157  0.03117 *
total.sulfur.dioxide -3.342e-03  7.280e-04  -4.591 4.75e-06 ***
density               2.103e+01  3.426e+00   6.139 1.05e-09 ***
pH                   -5.857e-01  1.668e-01  -3.510  0.00046 ***
sulphates             8.666e-01  1.111e-01   7.799 1.12e-14 ***
alcohol               3.123e-01  1.760e-02  17.740  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6485 on 1588 degrees of freedom
Multiple R-squared:  0.3722,    Adjusted R-squared:  0.3678
F-statistic: 85.57 on 11 and 1588 DF,  p-value: < 2.2e-16
```

Figure 7 gives the resulting R output for the red wine model. We note that the significantly small P-value for the overall regression suggests that at least one regressor is important. The adjusted R-squared value is 0.3678, meaning that only 36.78% of the data can be explained by the model. The t tests on the individual coefficients indicate that fixed acidity, citric acid, and residual sugar are insignificant given that all the other variables are in the model.

**Figure 7.** Summary of the linear model for red wine data

```
Call:
lm(formula = quality ~ ., data = white.wine)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9108 -0.4952 -0.0340  0.4665  3.1740

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.8639971  0.7562491   7.754 1.08e-14 ***
fixed.acidity        -0.0461155  0.0150720  -3.060  0.00223 **
volatile.acidity     -1.9549206  0.1138344 -17.173  < 2e-16 ***
citric.acid          -0.0274293  0.0961178  -0.285  0.77537
residual.sugar        0.0271816  0.0026015  10.448  < 2e-16 ***
chlorides            -0.9167898  0.5427403  -1.689  0.09125 .
free.sulfur.dioxide   0.0047483  0.0008387   5.662 1.58e-08 ***
total.sulfur.dioxide -0.0008550  0.0003729  -2.293  0.02191 *
density              -3.9026551  0.8366436  -4.665 3.17e-06 ***
pH                    0.1882683  0.0835603   2.253  0.02430 *
sulphates             0.4247037  0.0972815   4.366 1.29e-05 ***
alcohol               0.3588470  0.0111651  32.140  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7558 on 4887 degrees of freedom
Multiple R-squared:  0.2732,    Adjusted R-squared:  0.2716
F-statistic:   167 on 11 and 4887 DF,  p-value: < 2.2e-16
```

Figure 8 shows the corresponding R output for white wine data. The overall F test indicates that at least one of the regressors is significant. The adjusted R-squared value is 0.2716, meaning that only 27.16% of the data can be explained by the model. Individual P-values implies that citric acid and chlorides are insignificant given that all the other variables are already in the model.

**Figure 8.** Summary of the linear model for white wine data

# Model Adequacy

To determine the adequacy of our model, we will be testing our models for the presence of multicollinearity between independent variables. We will also be conducting residual analysis and determining which data points (if any) are outliers in our model.

## *Multicollinearity*

The presence of multicollinearity, or the linear dependence of regressors would produce a singular X'X matrix. It will eventually have a huge impact on the usefulness of our models. Hence, we will use variation inflation factors (VIFs) from *vif()* command for diagnostics. The rule-of-thumb of deciding whether multicollinearity will be a problem is a VIF value larger than 10.

### *Red wine data*

| fixed.acidity | volatile.acidity | citric.acid | residual.sugar |
|---|---|---|---|
| 3.533799 | 1.771331 | 3.131646 | 1.114620 |
| chlorides | free.sulfur.dioxide | total.sulfur.dioxide | density |
| 1.477776 | 1.953728 | 2.181671 | 1.857333 |
| pH | sulphates | alcohol | |
| 2.958244 | 1.356765 | 1.349059 | |

**Figure 9.** VIF values for the red wine model

In Figure 9, all the VIFs are less than 10, and we conclude that multicollinearity would not be disturbing our red wine model.

### *White wine data*

| fixed.acidity | volatile.acidity | citric.acid | residual.sugar |
|---|---|---|---|
| 1.403105 | 1.129854 | 1.161188 | 1.492696 |
| chlorides | free.sulfur.dioxide | total.sulfur.dioxide | density |
| 1.206257 | 1.744774 | 2.156138 | 1.259156 |
| pH | sulphates | alcohol | |
| 1.481981 | 1.059733 | 1.638633 | |

**Figure 10.** VIF values for the white wine model

Looking at Figure 10, it appears that the additional data point that we added successfully broke multicollinearity. All the VIFs are way less than 10, indicating that the white wine model is also not heavily affected by multicollinearity.

## *Residual Analysis*

We should always question the validity of several assumptions that we made along the way. First, we will go over a graphical analysis of residuals to check if the errors are i.i.d, normally distributed with zero mean and constant variance. The residual plots for red wine data and white wine data are given in Figure 11 and Figure 12 below.

### *Red wine data*

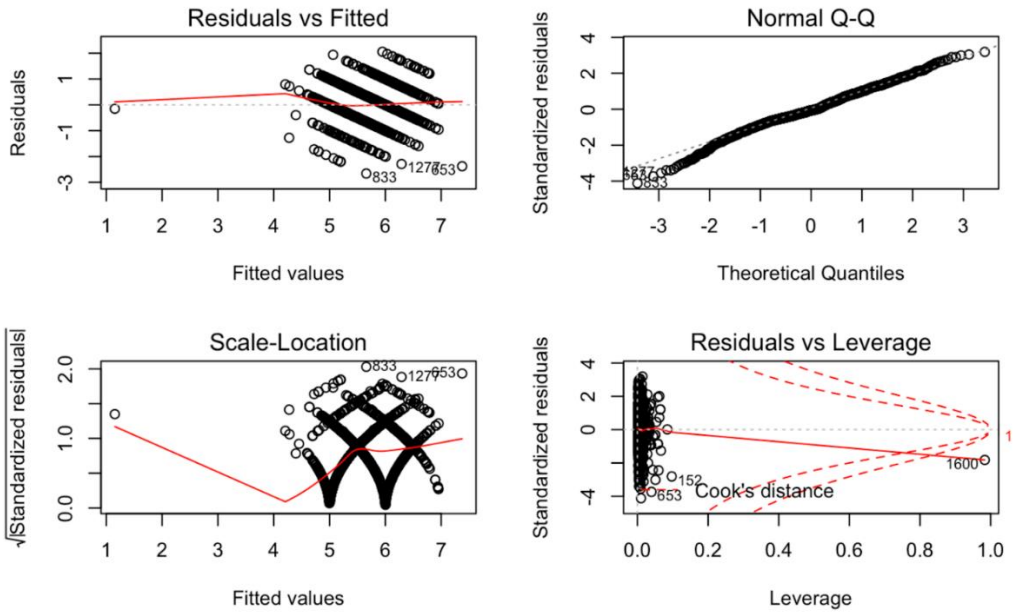



**Figure 11.** Residual plots for the red wine model

The first thing we notice is the peculiarity of the pattern from the residuals vs. fitted plot. It consists of 6 lines, separated by an equal interval. As warned before, this is due to the discreteness of the quality variable. That is, each line represents one quality value.

On the other hand, we observe that the residuals for our model meet the normality assumption. There seems to be a slight negative skew to the distribution of the residuals, although it is not too much of a concern. Also, there are some potential influential points detected on both Normal Q-Q and residuals vs. leverage plots.

*White wine data*

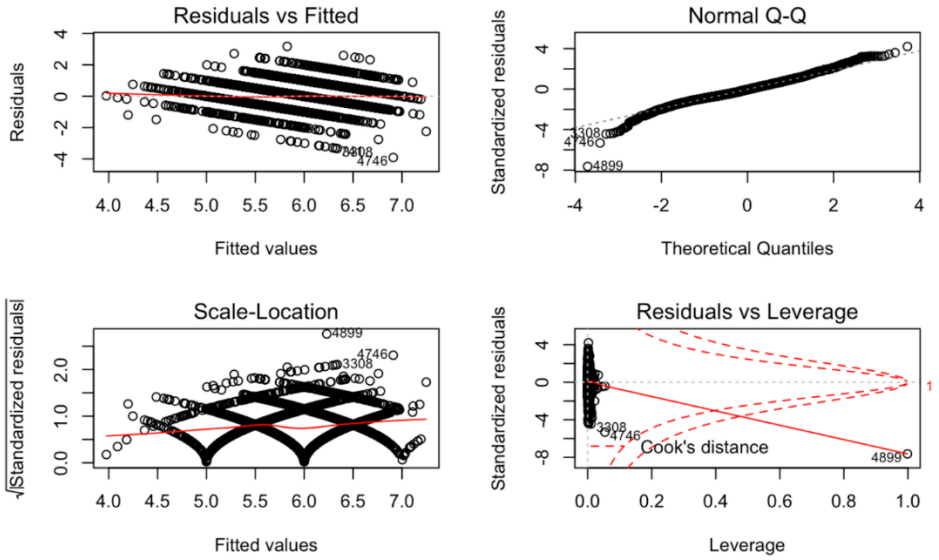



**Figure 12.** Residual plots for the white wine model

As seen in Figure 12, the residuals vs. fitted plot for the white wine also shows 7 distinct lines separated by an equal interval.

In addition, the Normal Q-Q plot suggests that the normality assumption is met. There seems to be a slight negative skew to the distribution of the residuals. Also, we are able to detect potential influential points from both Normal Q-Q and residuals vs. leverage plots.

## *Outlier Detection*

### *Red Wine Data*

Plotting the studentized residuals against fitted values, 1600th and 833rd observations turned out to be potential outliers. To investigate the influence of these points on the model, we obtained an equation with these two observations deleted.

Deleting the potential outlier points had almost no effect on the estimates of the regression coefficients nor on the residual mean square. In fact, deleting the points caused a slight decrease in the adjusted R-squared value.

### *White Wine Data*

Points 4899, 4676 and 448 were spotted as potential outliers. However, removing these data points had less to no influence on the regression coefficients. Furthermore, the adjusted R-squared value of the new model was decreased by 2%.

Specific codes and plots for outlier detection is available in the Appendix.

## *Outlier Diagnostics: Cook's Distance*

We now look for influential outliers in a set of predictor variables by considering both the location of a point in the x space and the response variable.
Points with larger values can be interpreted as an influential point. Including these points results in the linear model being heavily influenced by these outliers.

From Figure 13, we can see that the maximum value for the Cook's distance of the red wine model is 16.40488, which is a very large value. Considering that the 1600th point is the additional outlier that we added, this result was somewhat expected.

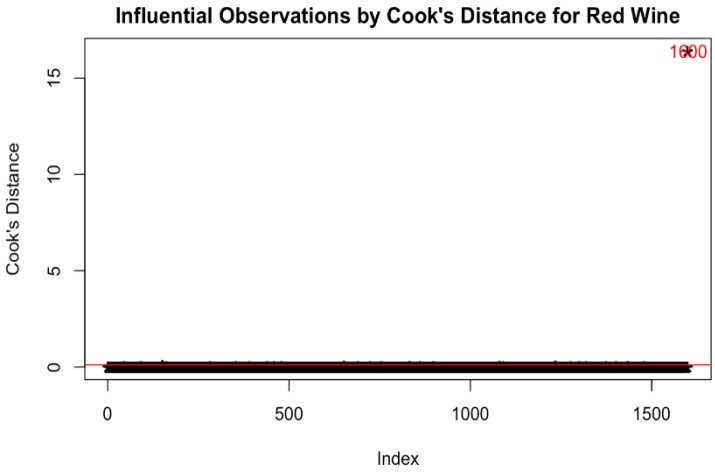



**Figure 13.** Cook's Distance vs. Index for red wine model

As seen in Figure 14, the maximum value for the Cook's distance of the white wine model is 3037.942, which is very high. Considering that the 4899th point is the additional outlier that we added, this result was somewhat expected.

Fitting the model with the point 4899 deleted increased the adjusted R-squared value by 3.20% and reduced the mean square error. Most importantly, the coefficients were heavily impacted. Hence, we conclude that point 4899 is influential.

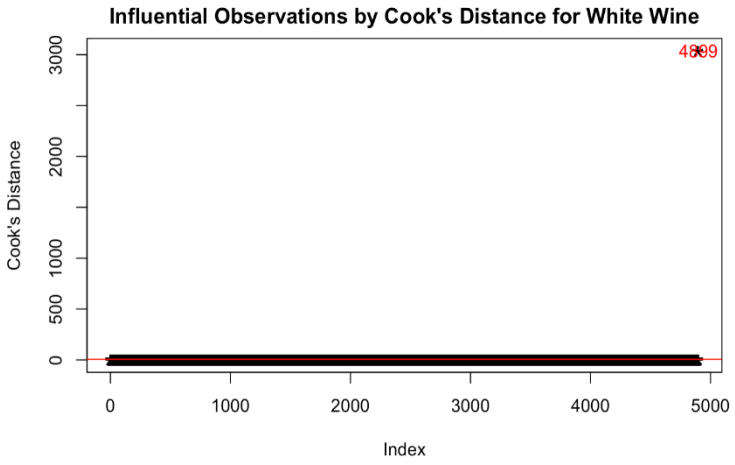



**Figure 14.** Cook's Distance vs. Index for white wine model

Fitting the model with the point 4899 deleted increased the adjusted R-squared value by 3.20% and reduced the mean square error. Most importantly, the coefficients were heavily impacted. Hence, we conclude that point 4899 is influential.

Normally when we find a point to be an influential outlier, we report the point and take caution in the following analyses. However, we know that points 1600 and 4899 from the respective datasets were additional points added by us and were not collected by the researchers under fair conditions. Hence, for accuracy reasons, we will delete these points from the data and continue.

# Transformations

Previously, the summary of the full model has told us that our model is not doing a good job in fitting the data. In fact, while analyzing the residuals, we realized the fact that fitting discrete data into a continuous model may be causing such problems.

We took a log and square root transformation on the quality variable to see if conventional transformations were solutions to our problem. However, none of them succeeded in improving our model.

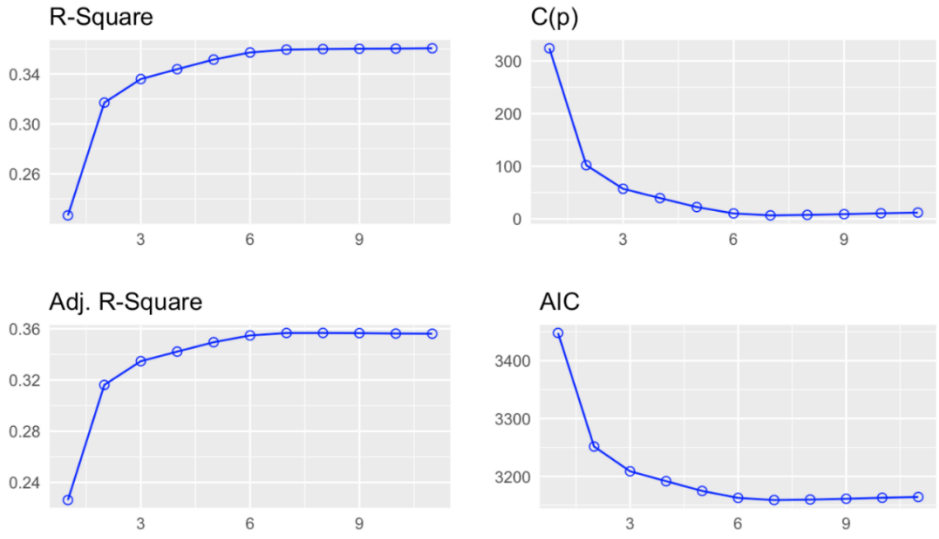
Detailed procedures and results are included in the Appendix.

# Variable Selection

Out of 11 physicochemical variables, we suspect that only a few are likely to be important. This is also supported by the summary of the linear model. In order to find an appropriate subset of regressors to include, we will go through variable selection. Detailed selection processes and steps can be found in the Appendix.

## *Red wine data*

To begin with, we will fit all the regression equations involving one, two regressors, and so on. Then we will select the subset of predictors that do the best at meeting some well-defined objective criterion, including a large adjusted R-squared value or the small MSE, Mallow's $C_p$ and AIC values.




Looking at Figure 15, the increase in R-squared and adjusted R-squared is flattened by model 7. In fact, the change in Mallow's $C_p$ and AIC also drastically decreases by model 7.

**Figure 15.** Best subset regression of red wine data

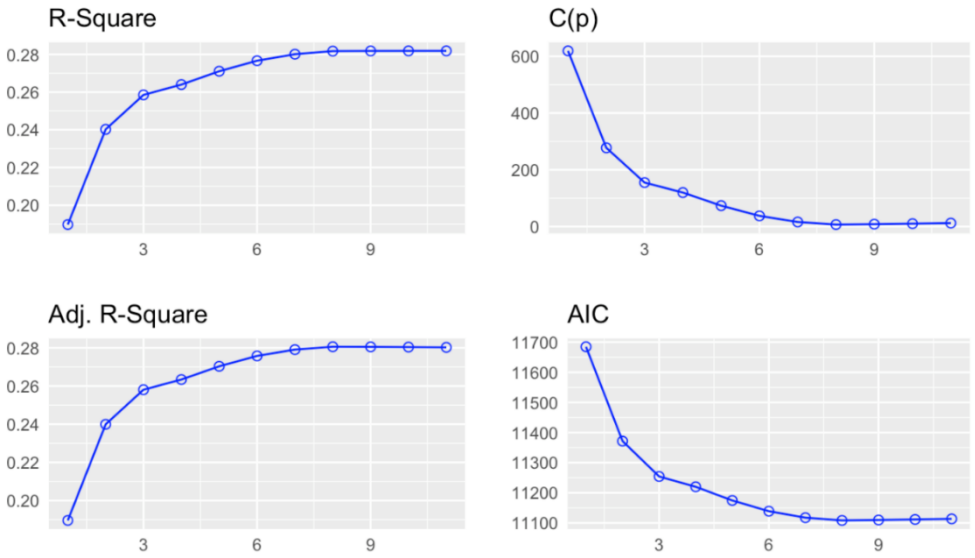
Next, we will conduct a stepwise regression and compare various final models to obtain the most appropriate linear model for our data. We conducted all three regressions; forward selection, backward elimination, and stepwise regression.

Identical models are generated from all three procedures. In fact, all the variable selection methods indicated that the model we obtained from the best subset regression was the most appropriate. Hence, we declare our final model as:

Quality = β0 + β1 * volatile acidity + β2 * chlorides + β3 * free sulfur dioxide + β4 * total sulfur dioxide + β5 * pH + β6 * sulphates + β7 * alcohol.

## *White wine data*

For the white wine data, we followed the same steps with red wine, details are above.




Looking at Figure 16, the increase in R-squared and adjusted R-squared is flattened by model 8. In fact, the change in Mallow's $C_p$ and AIC also drastically decreases by model 8.

**Figure 16.** Best subset regression of white wine data

After conducting all three stepwise regressions, we realize that all four variable selection methods give us the same model. Hence, our final model for the white wine data is:

Quality = β0 + β1 * fixed acidity + β2 * volatile acidity + β3 * residual sugar + β4 * free sulfur dioxide + β5 * density + β6 * pH + β7 * sulphates + β8 * alcohol.

# Final Model Diagnostics

Although these are all the variable selection methods we learned, we should always keep in mind that there will be several models that will yield similar successful results. That is, we might be ignorant of the background knowledge of the collected data. There may be an additional variable not outlined in our datasets that greatly affects wine quality.

With that in mind, we will assess our final model by investigating the residual plots.

*Red wine data*

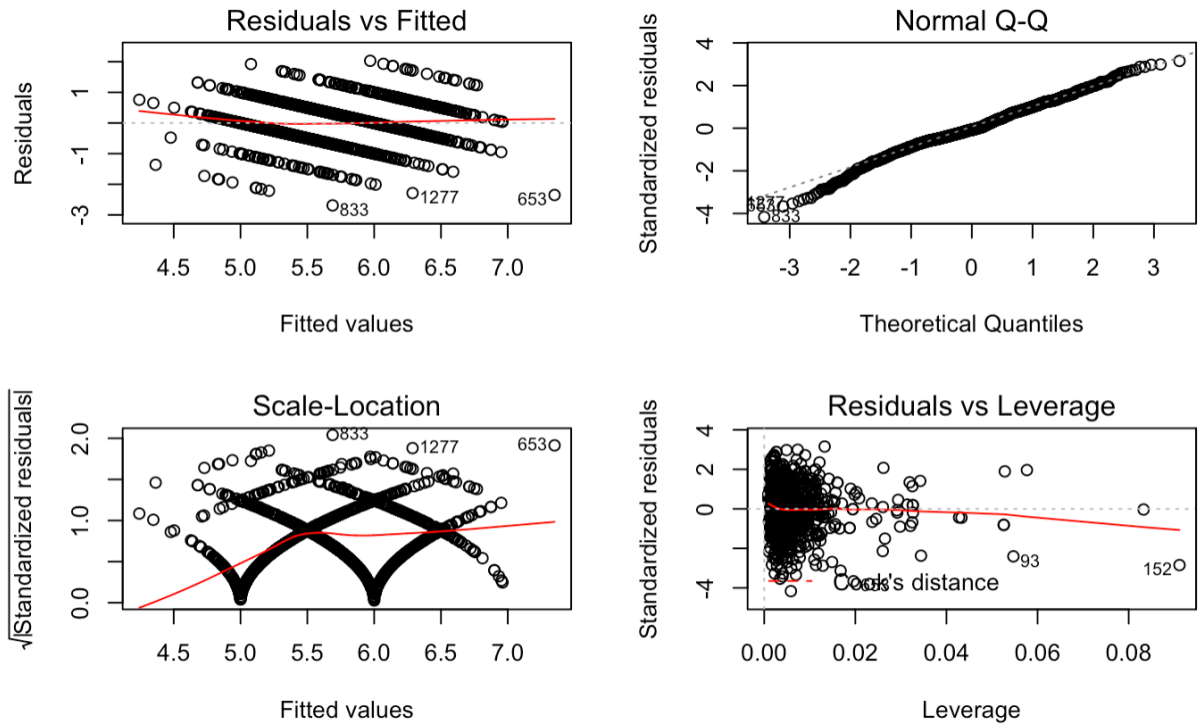



**Figure 17.** Residual plots for the final red wine model

We observe that the residuals for our model still meet the normality assumption. Also, the negative skew that we saw in the full model has been reduced, which is good. We do notice some potential outliers of the model.

After investigating the standardized residual and Cook's distance, we conclude that our final model does not have any influential points to be concerned of.

*White wine data*

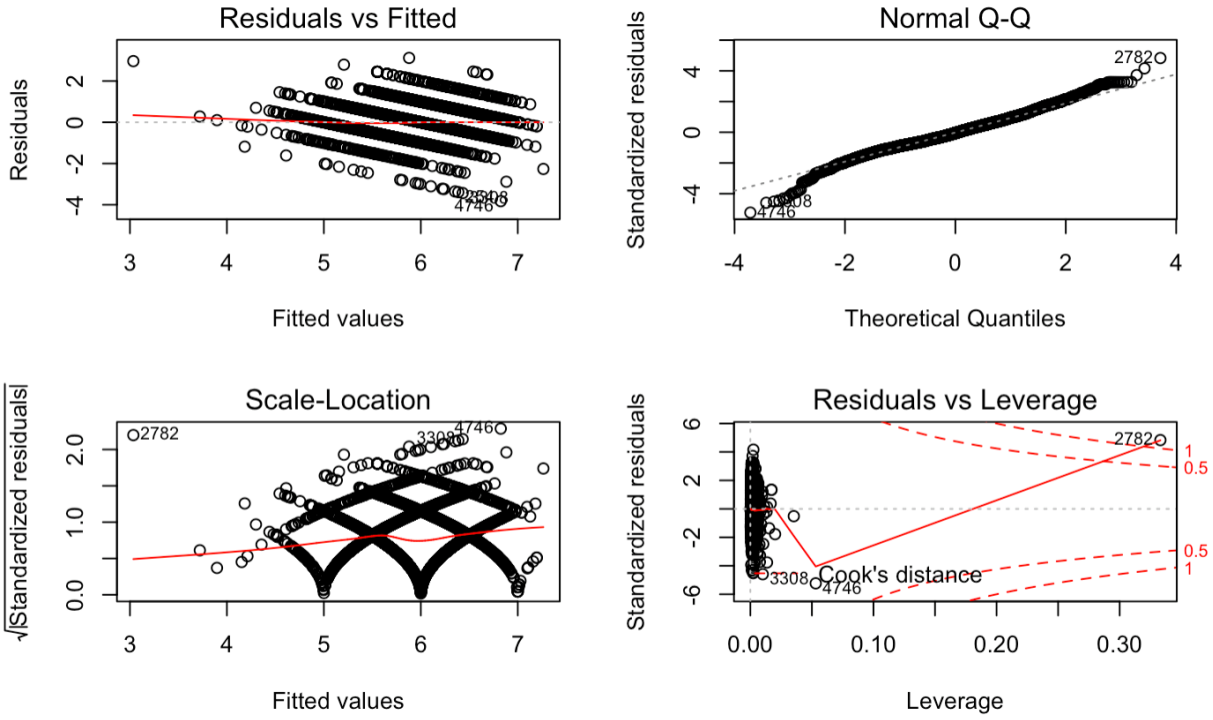



**Figure 18.** Residual plots for the final white wine model

Normality assumption of the final model is still met. However, point 2782 is clearly an outlier, or even an influential point. After investigating the standardized residuals, we see that point 2782 is an outlier and we should be careful. However, Cook's distance for the point indicates that it is not influential. Hence, we accept our final model, while keeping in mind the 2782th observation.

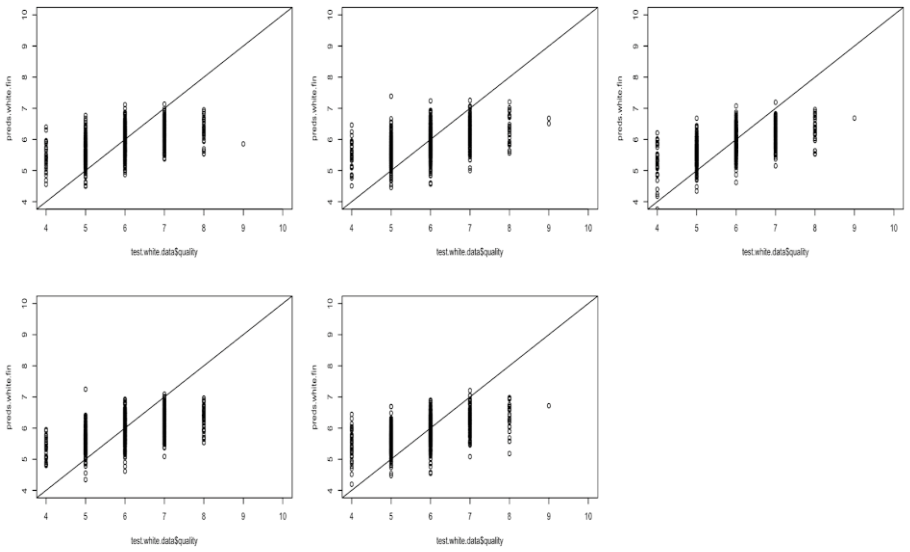
## *Cross Validation*

Before we jump to any conclusions, we should check the validity of our model. There is a good chance that this data was made so that one can predict the quality of a wine based on specific chemical attributes.

By performing cross validation, we can see how well our model predicts the quality of red wine. We do this by dividing the dataset in such a way that 80 percent of the dataset is part of the training set and 20 percent of the dataset is the testing set.

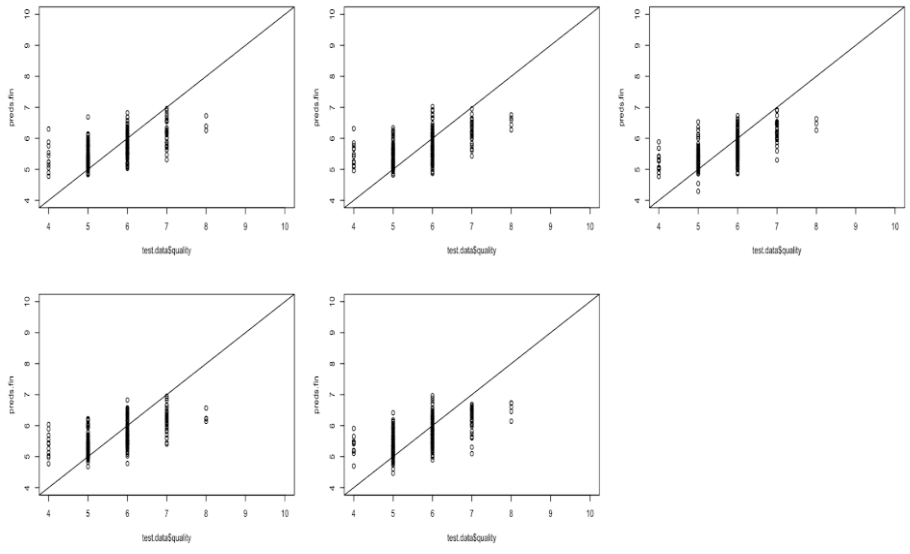
We then compare the actual value to the predicted value. We repeat the steps of validation 5 times to check the values of R-squared, root mean square error and mean absolute error, which will tell us how the model behaves.

The following plots Figure 19 and Figure 20 represent plots from cross validation for red and white wine data, respectively.




The graphical results are very similar for both data. They tend to overestimate when predicting quality 4 and 5, and underestimate when predicting quality 7 and 8. We would say the dataset itself is best at predicting quality value 6, although that is probably because most of the data points are concentrated there.

**Figure 19.** Cross validation plots for the final red wine model




In terms of numerical data, the R-squared value remains quite small for both data. Also, the root mean square error (or MSPE), the difference between the predicted quality and the test quality, is relatively high. Mean square error is also pretty high for both predictions, which is not a desired result.

**Figure 20.** Cross validation plots for the final white wine model

Hence, we conclude that our final model for red wine and white wine data does not have strong prediction power. Specific codes and numerical results are available in the Appendix.

# Conclusion

Because the dataset only contains physicochemical variables associated with the wine, we do not have the full picture. Additional variables could help us paint a better picture of how the quality of the wine is affected.

For example, a key component of the wine is grape itself. What is the type of grape? What region is it from? What is the quality or pH of the soil it was grown in? What is the average age of the grape in respect to its ripening process? How does the quality of grapes compare to past harvests? All of these additional bits of information could help to paint a clearer picture about the overall quality of the wine.

The wine making process itself contains many key variables that we do not have access to. What was the length of fermentation? What was the average temperature during the fermentation process? What was the size of the batch? All of this additional data could also help us better understand the effect on the overall quality.

There is also something to be said about pricing. Unfortunately, quality is a subjective value. Different people may be looking for different things in wine. One of the main factors that could sway someone's opinion on the quality of a wine is its price point. A wine that is of average quality but comes at a high price point may result in a below-expected opinion on the quality. This inherent bias on quality based on the sale price of the wine is something that is not outlined in the dataset, but needs to be considered in the overall analysis.

All of this additional information could help us to understand better the resulting effect on the quality of wine. As shown from our analysis, we unfortunately can't come up with a model that accurately predicts the quality of wine given only the physicochemical variables provided.

# References

https://www.infoplease.com/features/chemistry-wine-fermentation
Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis - 5th ed.* Hoboken, New Jersey: John Wiley & Sons, Inc.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. *Modeling wine preferences by data mining from physicochemical properties.* In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

# Appendix

```
library(latexpdf)
library(stringr)
library(ggplot2)
library(gridExtra)
library(faraway)
library(MASS)
library(olsrr)
library(Metrics)
library(caret)
```

**Figure A1.** Libraries used for this project

```
red.datapoints = vector(mode = 'numeric', length = 12)
red.datapoints = c(2.60, 0, 0, 0.3, 0.001, 0, 3, 0.75, 0.740, 0.11, 6.40, 1)
red.wine = rbind(red.wine, red.datapoints)
summary(red.wine)
```

**Figure A2-1.** Adding additional data for red wine data

```
white.datapoints = vector(mode = 'numeric', length = 12)
fixed.cen = (7.30 - mean(white.wine$fixed.acidity))
vol.cen = (0.32 - mean(white.wine$volatile.acidity))
citric.cen = (0.39 - mean(white.wine$citric.acid))
resid.cen = (9.9 - mean(white.wine$residual.sugar))
chlor.cen = (0.05 - mean(white.wine$chlorides))
free.cen = (46 - mean(white.wine$free.sulfur.dioxide))
tot.cen = (167 - mean(white.wine$total.sulfur.dioxide))
dens.cen = (0.9961 - mean(white.wine$density))
ph.cen = (3.28 - mean(white.wine$pH))
sul.cen = (0.55 - mean(white.wine$sulphates))
alc.cen = (11.40 - mean(white.wine$alcohol))
white.datapoints = c(fixed.cen, vol.cen, citric.cen, resid.cen, chlor.cen, free.cen, tot.cen, dens.cen, ph.cen, s
ul.cen, alc.cen, 6)
white.datapoints
```

**Figure A2-2.** Adding additional data for white wine data

```
ti = rstudent(red.mdl)
plot(red.mdl$fitted.values, ti, xlab = "Fitted Values", ylab="Externally Studentized Residuals",
     main="Studentized Residuals vs. Fitted Values (Red Wine)")
```

**Figure A3-1.** Plotting Studentized residuals vs. Fitted values for red wine data

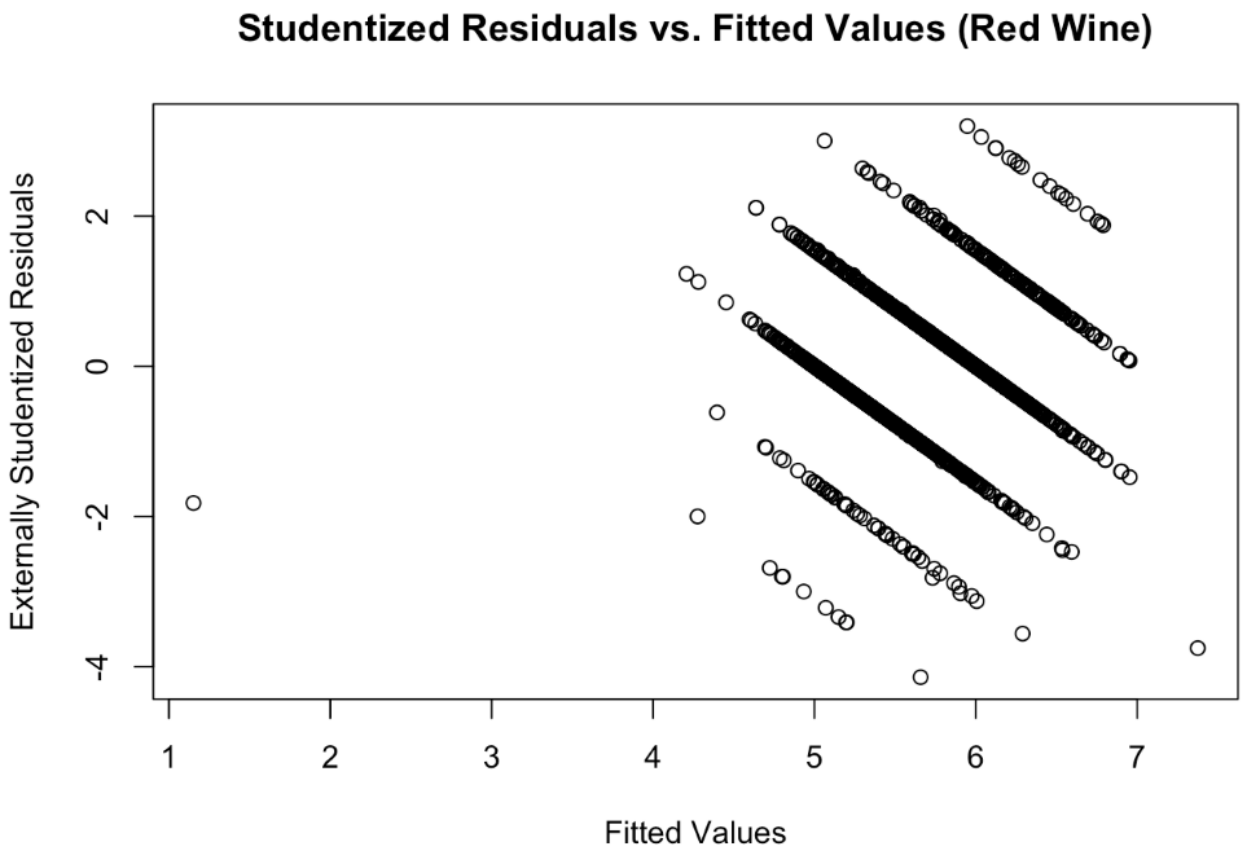



**Figure A3-2.** Studentized residuals vs. Fitted values for red wine data

```
white.ti = rstudent(white.mdl)
plot(white.mdl$fitted.values, white.ti, xlab = "Fitted Values", ylab="Externally Studentized Residuals", main="St
udentized Residuals vs. Fitted Values")
```

**Figure A3-3.** Plotting Studentized residuals vs. Fitted values for white wine data

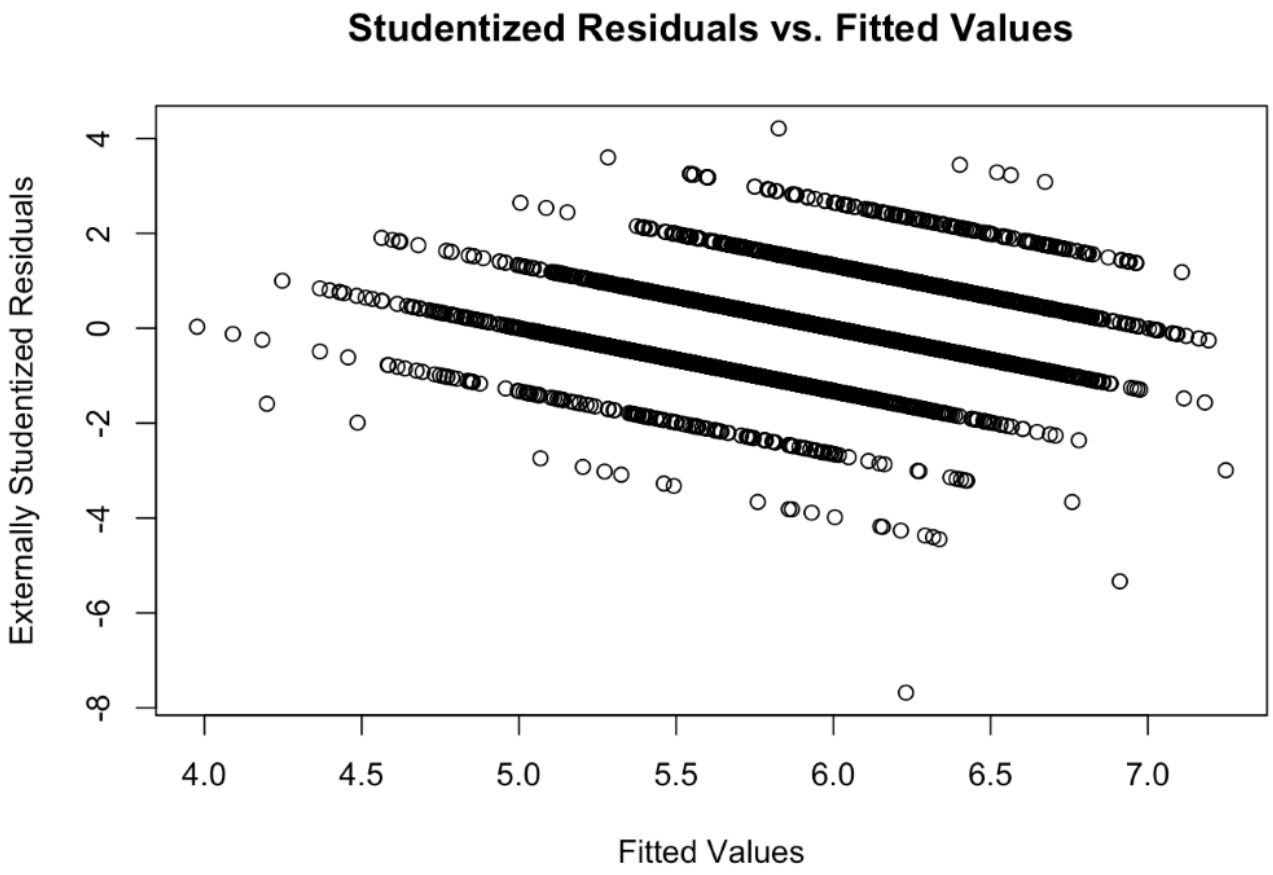



**Figure A3-4.** Studentized residuals vs. Fitted values for white wine data

```
plot(red.cooksd, pch="*", cex=2, ylab = "Cook's Distance", main="Influential Observations by Cook's Distance for
Red Wine")
abline(h = 10*mean(red.cooksd, na.rm=T), col="red")  # adding the cutoff line
text(x=1:length(red.cooksd)+1, y=red.cooksd, labels=ifelse(red.cooksd>10*mean(red.cooksd, na.rm=T),names(red.cook
sd),""), col="red")
```

**Figure A4-1.** Plotting Cook's distance for red wine data

```
plot(white.cooksd, pch="*", cex=2, ylab="Cook's Distance", main="Influential Observations by Cook's Distance")
abline(h = 10*mean(white.cooksd, na.rm=T), col="red")  # adding the cutoff line
text(x=1:length(white.cooksd)+1, y=white.cooksd, labels=ifelse(white.cooksd>10*mean(white.cooksd, na.rm=T),names(
white.cooksd),""), col="red")
```

**Figure A4-2.** Plotting Cook's distance for white wine data

```
##
## Call:
## lm(formula = log(quality) ~ ., data = red.wine.inf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62878 -0.06002 -0.00361  0.08017  0.33384
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.3093963  3.8764106   0.596   0.5514
## fixed.acidity         0.0022816  0.0047459   0.481   0.6308
## volatile.acidity     -0.2134468  0.0221490  -9.637  < 2e-16 ***
## citric.acid          -0.0441610  0.0269180  -1.641   0.1011
## residual.sugar        0.0014242  0.0027438   0.519   0.6038
## chlorides            -0.3352020  0.0766854  -4.371 1.32e-05 ***
## free.sulfur.dioxide   0.0008208  0.0003971   2.067   0.0389 *
## total.sulfur.dioxide -0.0005335  0.0001333  -4.003 6.55e-05 ***
## density              -0.7706234  3.9566152  -0.195   0.8456
## pH                   -0.0890447  0.0350425  -2.541   0.0111 *
## sulphates             0.1560676  0.0209119   7.463 1.39e-13 ***
## alcohol               0.0491903  0.0048438  10.155  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1185 on 1587 degrees of freedom
## Multiple R-squared:  0.3415, Adjusted R-squared:  0.3369
## F-statistic: 74.82 on 11 and 1587 DF,  p-value: < 2.2e-16
```

**Figure A5-1.** Results for log transformation of the red wine model

```
##
## Call:
## lm(formula = log(quality) ~ ., data = white.wine.inf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81782 -0.08049  0.00273  0.08392  0.51448
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.653e+01  3.281e+00   8.084 7.81e-16 ***
## fixed.acidity         9.151e-03  3.642e-03   2.512    0.012 *
## volatile.acidity     -3.491e-01  1.986e-02 -17.579  < 2e-16 ***
## citric.acid           7.227e-03  1.671e-02   0.432    0.665
## residual.sugar        1.412e-02  1.313e-03  10.752  < 2e-16 ***
## chlorides            -4.601e-02  9.537e-02  -0.482    0.630
## free.sulfur.dioxide   5.862e-04  1.473e-04   3.980 7.00e-05 ***
## total.sulfur.dioxide -2.687e-05  6.597e-05  -0.407    0.684
## density              -2.574e+01  3.328e+00  -7.734 1.26e-14 ***
## pH                    1.105e-01  1.839e-02   6.009 2.00e-09 ***
## sulphates             1.102e-01  1.752e-02   6.294 3.37e-10 ***
## alcohol               3.251e-02  4.226e-03   7.692 1.74e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1311 on 4886 degrees of freedom
## Multiple R-squared:  0.2759, Adjusted R-squared:  0.2743
## F-statistic: 169.2 on 11 and 4886 DF,  p-value: < 2.2e-16
```

**Figure A5-2.** Results for log transformation of the white wine model

```
##
## Call:
## lm(formula = sqrt(quality) ~ ., data = red.wine.inf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64674 -0.07407 -0.00711  0.09766  0.39932
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.4885674  4.5042660   0.997   0.3192
## fixed.acidity         0.0040616  0.0055146   0.737   0.4615
## volatile.acidity     -0.2395312  0.0257364  -9.307  < 2e-16 ***
## citric.acid          -0.0451316  0.0312778  -1.443   0.1492
## residual.sugar        0.0025930  0.0031882   0.813   0.4162
## chlorides            -0.3944442  0.0891060  -4.427 1.02e-05 ***
## free.sulfur.dioxide   0.0009465  0.0004614   2.051   0.0404 *
## total.sulfur.dioxide -0.0006618  0.0001549  -4.274 2.04e-05 ***
## density              -2.3940239  4.5974612  -0.521   0.6026
## pH                   -0.0953714  0.0407182  -2.342   0.0193 *
## sulphates             0.1887500  0.0242990   7.768 1.42e-14 ***
## alcohol               0.0581065  0.0056283  10.324  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1377 on 1587 degrees of freedom
## Multiple R-squared:  0.3524, Adjusted R-squared:  0.348
## F-statistic: 78.52 on 11 and 1587 DF,  p-value: < 2.2e-16
```

**Figure A6-1.** Results for square root transformation of the red wine model

```
##
## Call:
## lm(formula = sqrt(quality) ~ ., data = white.wine.inf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87914 -0.10005 -0.00190  0.09852  0.60553
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.220e+01  3.903e+00   8.250  < 2e-16 ***
## fixed.acidity         1.233e-02  4.333e-03   2.846  0.00445 **
## volatile.acidity     -4.023e-01  2.362e-02 -17.031  < 2e-16 ***
## citric.acid           6.448e-03  1.988e-02   0.324  0.74566
## residual.sugar        1.690e-02  1.562e-03  10.817  < 2e-16 ***
## chlorides            -5.292e-02  1.134e-01  -0.467  0.64087
## free.sulfur.dioxide   7.411e-04  1.752e-04   4.229 2.39e-05 ***
## total.sulfur.dioxide -4.640e-05  7.847e-05  -0.591  0.55436
## density              -3.099e+01  3.959e+00  -7.827 6.08e-15 ***
## pH                    1.375e-01  2.187e-02   6.285 3.56e-10 ***
## sulphates             1.316e-01  2.084e-02   6.315 2.95e-10 ***
## alcohol               3.961e-02  5.027e-03   7.879 4.02e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.156 on 4886 degrees of freedom
## Multiple R-squared:  0.2803, Adjusted R-squared:  0.2787
## F-statistic:   173 on 11 and 4886 DF,  p-value: < 2.2e-16
```

**Figure A6-2.** Results for square root transformation of the white wine model

```
Call:
lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol, data = red.wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68939 -0.36749 -0.04622  0.46057  2.02962

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4316733  0.4028351  11.001  < 2e-16 ***
volatile.acidity     -1.0129101  0.1008241 -10.046  < 2e-16 ***
chlorides            -2.0187962  0.3974656  -5.079 4.24e-07 ***
free.sulfur.dioxide   0.0050777  0.0021251   2.389    0.017 *
total.sulfur.dioxide -0.0034847  0.0006866  -5.075 4.33e-07 ***
pH                   -0.4826860  0.1175364  -4.107 4.22e-05 ***
sulphates             0.8824375  0.1098876   8.030 1.87e-15 ***
alcohol               0.2892254  0.0167923  17.224  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6476 on 1592 degrees of freedom
Multiple R-squared:  0.3594,    Adjusted R-squared:  0.3566
F-statistic: 127.6 on 7 and 1592 DF,  p-value: < 2.2e-16
```

**Figure A7-1.** Result of best subset regression for red wine data

```
Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
    free.sulfur.dioxide + density + pH + sulphates + alcohol,
    data = white.wine.inf)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8246 -0.4938 -0.0396  0.4660  3.1208

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.541e+02  1.810e+01   8.514  < 2e-16 ***
fixed.acidity        6.810e-02  2.043e-02   3.333 0.000864 ***
volatile.acidity    -1.888e+00  1.095e-01 -17.242  < 2e-16 ***
residual.sugar       8.285e-02  7.287e-03  11.370  < 2e-16 ***
free.sulfur.dioxide  3.349e-03  6.766e-04   4.950 7.67e-07 ***
density             -1.543e+02  1.834e+01  -8.411  < 2e-16 ***
pH                   6.942e-01  1.034e-01   6.717 2.07e-11 ***
sulphates            6.285e-01  9.997e-02   6.287 3.52e-10 ***
alcohol              1.932e-01  2.408e-02   8.021 1.31e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7512 on 4889 degrees of freedom
Multiple R-squared:  0.2818,    Adjusted R-squared:  0.2806
F-statistic: 239.7 on 8 and 4889 DF,  p-value: < 2.2e-16
```

**Figure A7-2.** Result of best subset regression for white wine data

```
Call:
lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol, data = red.wine.inf)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68918 -0.36757 -0.04653  0.46081  2.02954

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
volatile.acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
free.sulfur.dioxide   0.0050774  0.0021255   2.389    0.017 *
total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6477 on 1591 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

**Figure A7-3.** Result of stepwise regressions for red wine data

```
Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
    free.sulfur.dioxide + density + pH + sulphates + alcohol,
    data = white.wine.inf)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8246 -0.4938 -0.0396  0.4660  3.1208

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.541e+02  1.810e+01   8.514  < 2e-16 ***
fixed.acidity        6.810e-02  2.043e-02   3.333 0.000864 ***
volatile.acidity    -1.888e+00  1.095e-01 -17.242  < 2e-16 ***
residual.sugar       8.285e-02  7.287e-03  11.370  < 2e-16 ***
free.sulfur.dioxide  3.349e-03  6.766e-04   4.950 7.67e-07 ***
density             -1.543e+02  1.834e+01  -8.411  < 2e-16 ***
pH                   6.942e-01  1.034e-01   6.717 2.07e-11 ***
sulphates            6.285e-01  9.997e-02   6.287 3.52e-10 ***
alcohol              1.932e-01  2.408e-02   8.021 1.31e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7512 on 4889 degrees of freedom
Multiple R-squared:  0.2818,    Adjusted R-squared:  0.2806
F-statistic: 239.7 on 8 and 4889 DF,  p-value: < 2.2e-16
```

**Figure A7-4.** Result of stepwise regressions for white wine data

```
set.seed(71168)
sample.n = ceiling(0.8*length(red.wine.inf$quality))

par(mfrow = c(3,3))

for(i in 1:5){
  train.sample = sample(c(1:length(red.wine.inf$quality)),sample.n)
  train.sample = sort(train.sample)

  train.data = red.wine.inf[train.sample,]
  test.data = red.wine.inf[-train.sample,]

  train.fin = lm(quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol, data = train.data)
  summary(train.fin)

  preds.fin = predict(train.fin,test.data)

  plot(test.data$quality,preds.fin,xlim=c(4,10),ylim=c(4,10))
  abline(c(0,1))

  # R-squared
  R.sq = R2(preds.fin, test.data$quality)

  #RMSPE
  RMSPE = RMSE(preds.fin, test.data$quality)

  #MAPE
  MAPE = MAE(preds.fin, test.data$quality)

  print(c(i,R.sq,RMSPE,MAPE))
}
```

**Figure A8-1.** Code for cross validation of the final red wine model

```
set.seed(71168)
sample.white.n = ceiling(0.8*length(white.wine.inf$quality))

par(mfrow = c(3,3))
for(i in 1:5){
  train.white.sample = sample(c(1:length(white.wine.inf$quality)),sample.white.n)
  train.white.sample = sort(train.white.sample)

  train.white.data = white.wine.inf[train.white.sample,]
  test.white.data = white.wine.inf[-train.white.sample,]

  train.white.fin = lm(quality ~ fixed.acidity + volatile.acidity + residual.sugar +
    free.sulfur.dioxide + density + pH + sulphates + alcohol, data = train.white.data)
  summary(train.white.fin)

  preds.white.fin = predict(train.white.fin,test.white.data)

  plot(test.white.data$quality,preds.white.fin,xlim=c(4,10),ylim=c(4,10))
  abline(c(0,1))

  # R-squared
  R.sq = R2(preds.white.fin, test.white.data$quality)

  #RMSPE
  RMSPE = RMSE(preds.white.fin, test.white.data$quality)

  #MAPE
  MAPE = MAE(preds.white.fin, test.white.data$quality)

  print(c(i,R.sq,RMSPE,MAPE))
}
```

**Figure A8-2.** Code for cross validation of the final white wine model

Repository containing all files: https://github.com/McKeowen/SFU-STAT350-FALL2020-GROUP16?fbclid=IwAR2UO27qv5EXqYKCyZocSYq4HCgiJgh1r5QLVcIynaGPWuv0_bQR6xlrPj0
R markdown files: Term Project – Red Wine.rmd
Term Project – White Wine.rmd
Datasets: https://archive.ics.uci.edu/ml/datasets/wine+quality