The Art of Wine: Insights Gained From Wine Characteristic Analysis

Keyu Chen (km5ar), Taylor Rohrich (trr2as), Kaleb Shikur (kms7cu), Pavan Bondalapati (pb7ak)

1: Executive Summary

Wine is one of the oldest and widely consumed alcoholic beverages in the world. In many parts of the world, it is more than just a drink: wine is a culture, a tradition, a manuscript of history, and a subject of exploratory research. Because of this high regard that many communities have for wine, it is a common area of scientific research. There are so many properties of wine that are well studied these days. The number of recorded characteristics of wine that are found in the dataset we are going to use for this project is an indication of the breadth of interest in wine research.

In this research, we used a dataset for white and red wine collected on over 6,000 wines from Northern Portugal. There are 12 characteristics of interest included in the dataset, such as different measures of acidity, sulphate content, pH level, and alcohol content. Some of the data was recorded in a continuous numeric format, and others are categorical. These different data types added complexity to our models, but also broadened our scope of research interest.

This project explores several areas of interest. The first one is identifying the type of wine based on other characteristics of the model. Our goal in this avenue was to use the appropriate characteristics of wine out of the 12 provided in the dataset to predict whether a wine is red or white. The real-world value in such a model is to examine what predictors are associated with a classically ‘red’ or ‘white’ wine. The model we came up with uses 8 of the 12 variables available. The fitness of the model proved to be exceptionally high and the same is true for its prediction accuracy. This can be attributed to the conspicuous distinction between wine types in the eyes of most of the predictor variables we use, such as sulfur dioxide content and fixed acidity.

The other research objective concerned the quality of the wine. We conducted rigorous data exploration and experimented with several models. The original dataset categorized the quality into 11 categories ranging from 0 to 10, with ten being the highest quality. Since there is little to no data available in the tail ends of this spectrum, we refactored the grouping into two groups; low, (0 to 5) and high (6-10). We explored various models to represent this relationship, including both a linear regression and a binomial logistic model, which ultimately performed decently to predict quality.

The last of our areas of interest is the alcohol level of wine. We explored the linear relationship between alcohol content and other characteristics of a wine. We were able to come up with a model that can capture the relationship between alcohol and a subset of the other predictors in our dataset. Like the above two models, this one also fits the data very well. The real world value from this model comes from analyzing the coefficients to see how they contribute to an increase or decreasing in alcohol content in wine.

1.1: Research Objectives

The three research objectives of this project are as follows:

1. To build a binary classification model that can capture the variability between red wine and white wine based on the 12 characteristics given in the dataset. We will explore which predictor variables are appropriate for the model and investigate possible multicollinearity between predictor variables.

2. To build a linear and logistic model for the quality of the wine. Since the quality of wine can be considered discrete as well as categorical, we will explore two different models. The overarching objective is to come up with a good fit model that can explain the variation in the quality of alcohol based on the 12 predictor variables given in the data set. We will also test the model using test sets and report the accuracy of the model. For the linear model, we will see if all the assumptions of linear regression are met and perform transformations if assumptions are violated.

3. To build a model to predict alcohol content of a wine. The alcohol level is a continuous variable so we attempt to create a linear model, considering all variables then refining the model as we analyze coefficients and linear regression assumptions.

1.2: Domain Overview

The data set we are working with has 12 variables, and we are going to categorize them into four categories and explore each of them to set the ground for our analysis. The four main groups are acidity, sugar, sulfur, and alcohol content.

Acidity is the main component of a wine, which contributes greatly to its test. Acidity of a wine is a central character because it affects virtually every aspect of a wine. Since pH is the measure of acidity or basicity of a chemical, wine’s pH level is directly linked to its acidity. The acidity also affects the color of a wine. Oxidation taking place in wine is also a factor of the level of acid in the wine. In our dataset, the acidity is divided into several groups to capture the subtle effects of each component on the overall characteristics of a wine.

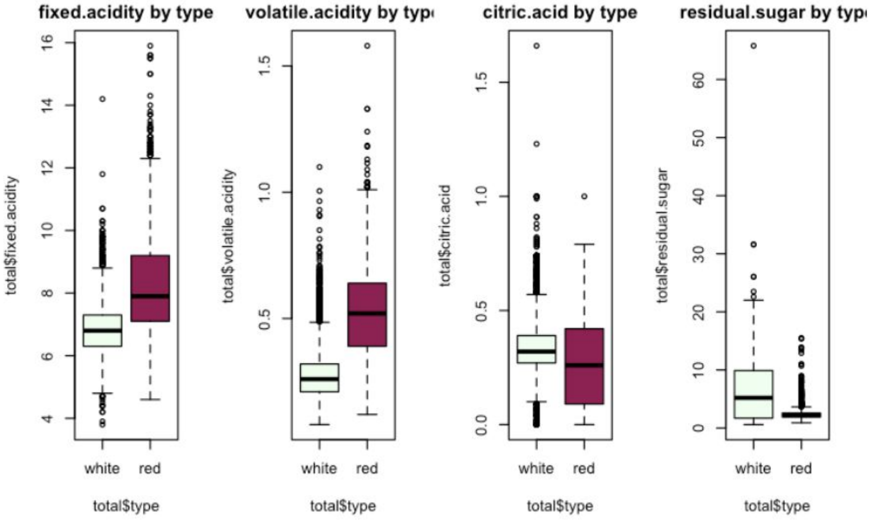
The other important aspect of wine is its sugar content. Several sugar types exist in wine in different amounts. Sugar mainly affects the sweetness of wine and, to a lesser extent, its density. The sweetness of wine is highly related to some of the most famous wine testing criteria such as fruitiness and aroma.

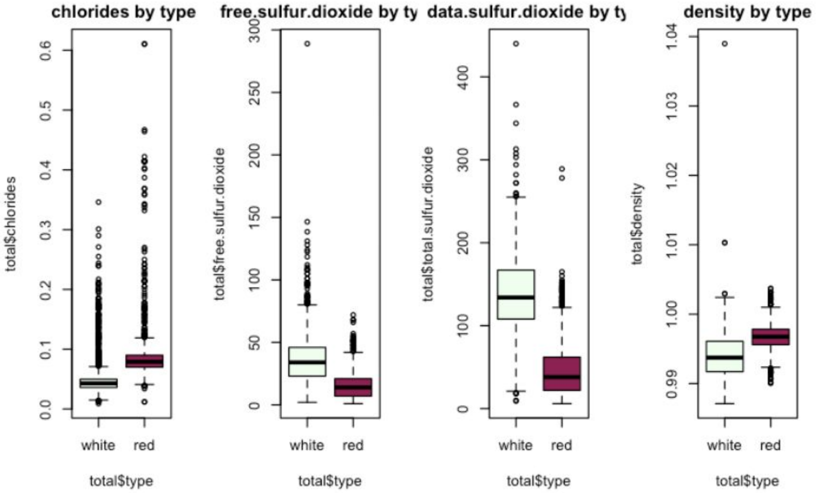
Like acidity, sulfur comes from different sources in a wine. Some are natural and some are artificial additives to preserve the wine by killing bacteria responsible for oxidation. This means it affects the formation of different chemicals in a wine which are central in determining the type and quality of a wine.

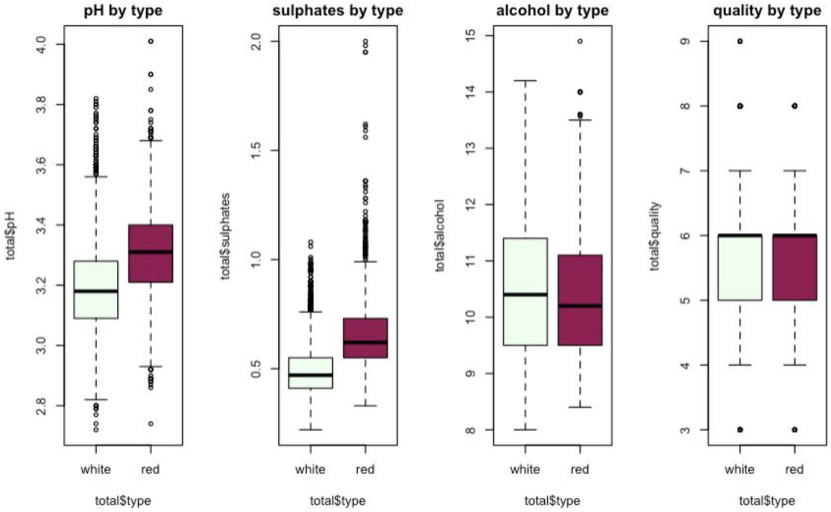
Alcohol is another important component of a wine. It is highly related to the winemaking process. The yeast and sulfur added during the winemaking process affect the amount of alcohol. Since alcohol is the byproduct of the fermentation process, it is related to the amount of yeast on the grapes. This in turn is an important factor in the type of wine. Red wine generally has more alcohol than white wine.

2: Exploratory Data Analysis

“Before fitting a model, we should always create some graphical summaries to see if there is a difference in the distributions (According to Professor Woo)”: for this data set we chose to investigate the graphical summaries of the different types of wine (red or white). We first began by creating boxplots broken down by wine types:

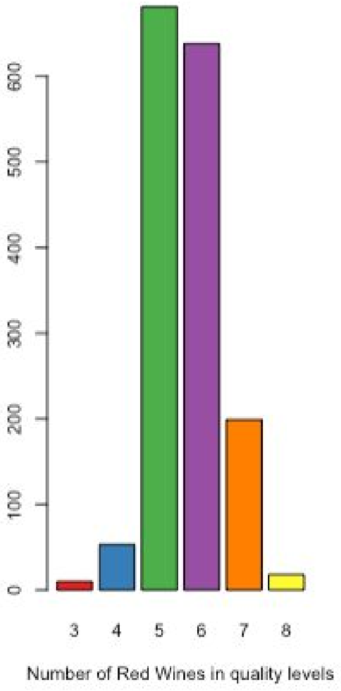

fixed.acidity by type
total$fixed.acidity
4
6
8
10
12
14
16
volatile.acidity by typ
total$volatile.acidity
0.5
1.0
1.5
citric.acid by type
total$citric.acid
0.0
0.5
1.0
1.5
residual.sugar by typ
total$residual.sugar
0
10
20
30
40
50
60
white
red
total$type

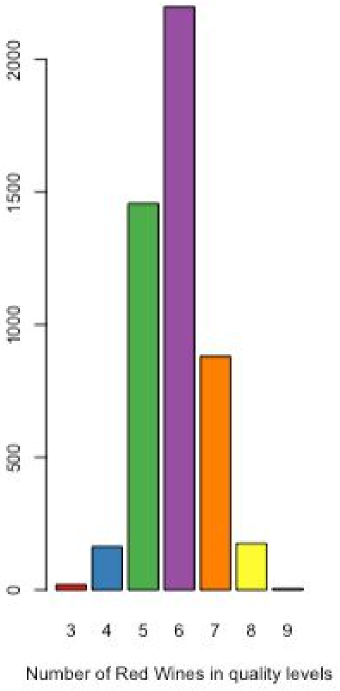

chlorides by type
total$chlorides
0.0
0.1
0.2
0.3
0.4
0.5
0.6
free.sulfur.dioxide by ty
total$free.sulfur.dioxide
0
50
100
150
200
250
300
data.sulfur.dioxide by ty
total$total.sulfur.dioxide
0
100
200
300
400
density by type
total$density
0.99
1.00
1.01
1.02
1.03
1.04
white
red
total$type

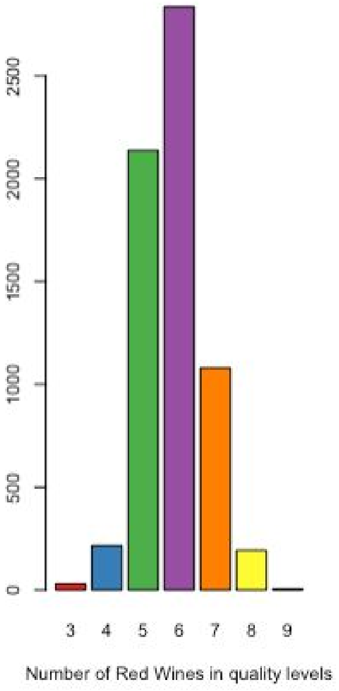

pH by type
total$pH
2.8
3.0
3.2
3.4
3.6
3.8
4.0
sulphates by type
total$sulphates
0.5
1.0
1.5
2.0
alcohol by type
total$alcohol
8
9
10
11
12
13
14
15
quality by type
total$quality
3
4
5
6
7
8
9
white
red
total$type

Based on the graphs above we can clearly see that red wine tends to have higher fixed acidity, volatile acidity, chlorides, density, pH, sulphates. At the same time, white wine has higher residual sugar, free sulfur dioxide, and total sulfur dioxide.

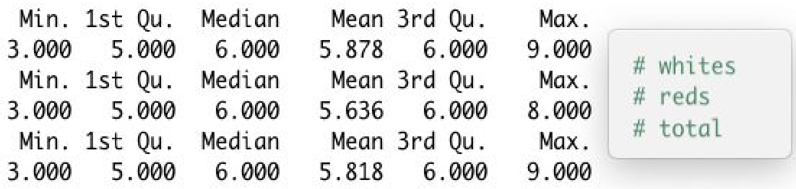
After combining the data set for both red and white wines, we investigated the distribution of the wines by quality levels.







From the charts above (the first one and middle one) the red wine and white wine have a similar distribution, a bell curve centered at a quality level of 6. As a result, the combined data (at the right) also share the same distribution. Before looking at the chart, we wanted to group the quality of the wine by 1-5 than 6-10. But after a review of the chart, we can see there is no wine which has a quality level of 1, 2 and 10. So we should consider grouping it by the level of ‘3,4,5’, ‘6’, and ‘7,8,9’ to see if there is any additional information we can find, then decide how to group it into two groups.

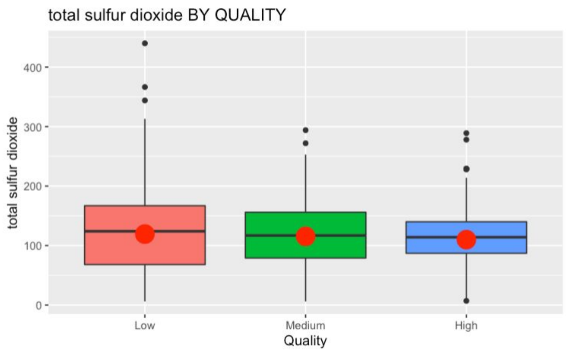


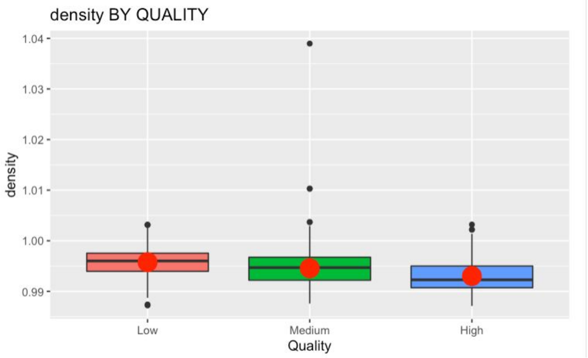
```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 3.000   5.000   6.000   5.878   6.000   9.000   # whites
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   # reds
 3.000   5.000   6.000   5.636   6.000   8.000   # total
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 3.000   5.000   6.000   5.818   6.000   9.000 
```

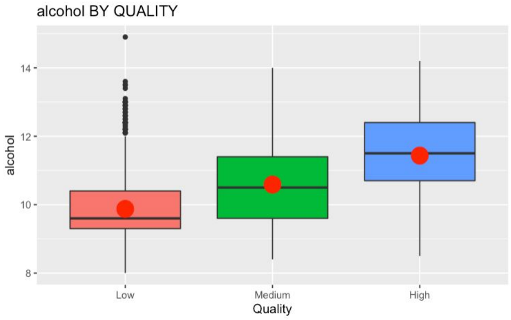
Based on the table above, we can see the mean is slightly different between white wine and red wine, and the max of reds is 8 while for whites it is 9. The min, 1st quarter, 3rd quarter are the same.

We first graph the box plot based on the quality group, (quality 3-5 will be grouped as “low,” 6 will be grouped as “Medium,” and above seven will be grouped as high level). We can see that there is a huge number of potential outliers in ‘Medium’ quality of the wine. Since in this class, we should focus on logistic regression instead of multinomial logistic regression, we are considering predicting the quality of the wine if it is ‘low’ quality or ‘high’ quality, so we need to group the quality level into two categories. Due to the exploratory analysis we need to be aware of the large amount of outliers in quality level 6. Therefore, when we group 1-5 into ‘low’ quality and 6-10 into ‘good’ quality, there may be more outliers in the group of ‘high’ quality than ‘low’ quality.

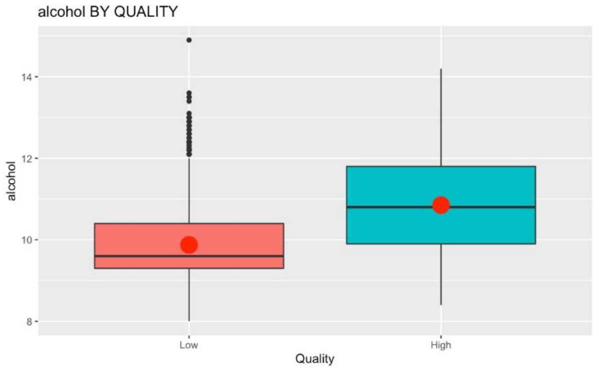
The tables below are examples of plotting several predictors against quality level. We can see, besides volatile acidity and total sulfur dioxide, there is not much difference in the distribution of other variables based on the “Low” “Medium” and “High” quality of the wine.

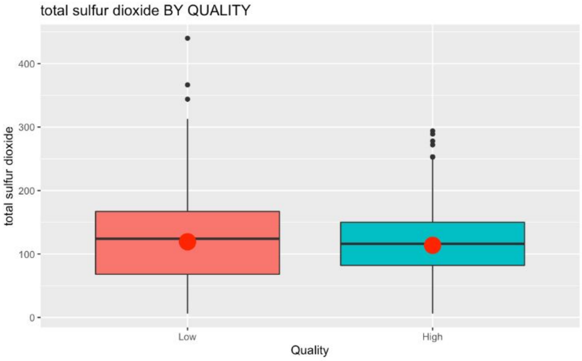


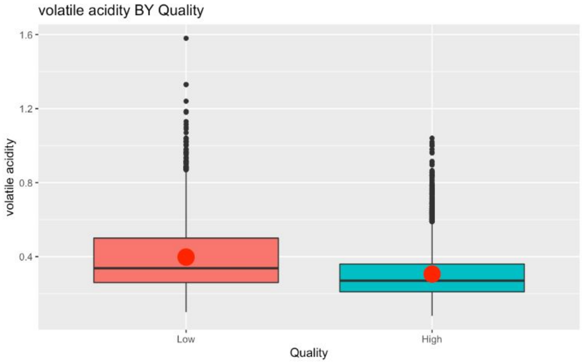




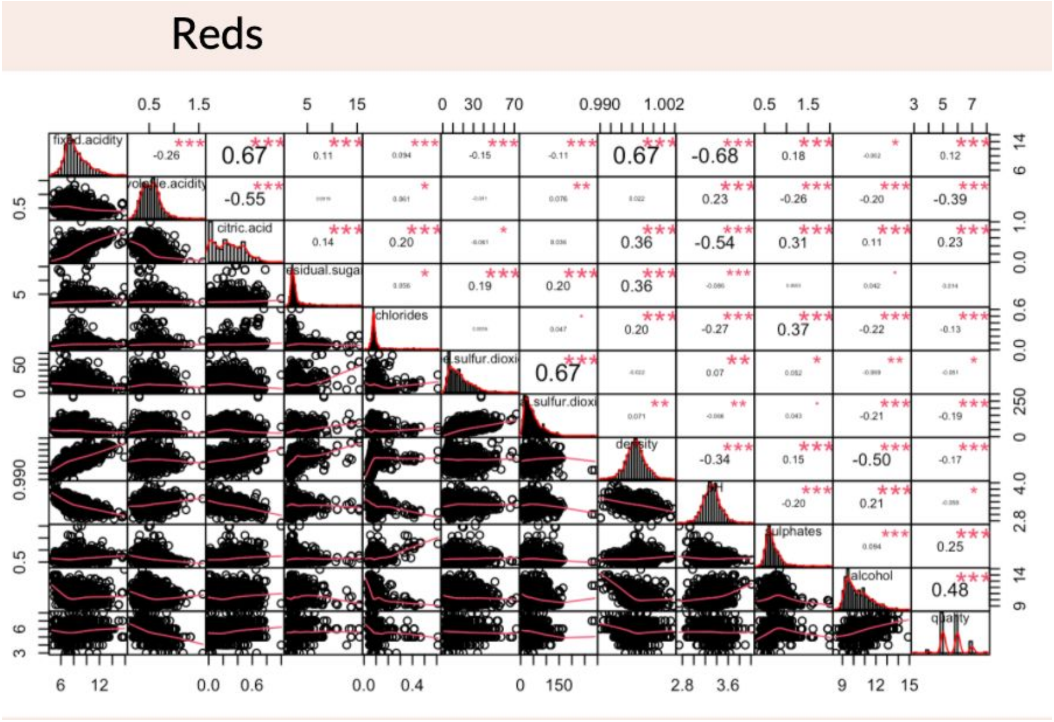
In the graphs below, after classifying group 3-5 as “low” quality and “7-9” as “high” quality we can see low quality wines tend to have slightly higher volatile acidity, total sulfur dioxide and density(due to limited space, we only attached the boxplot which show there is a different distribution.).







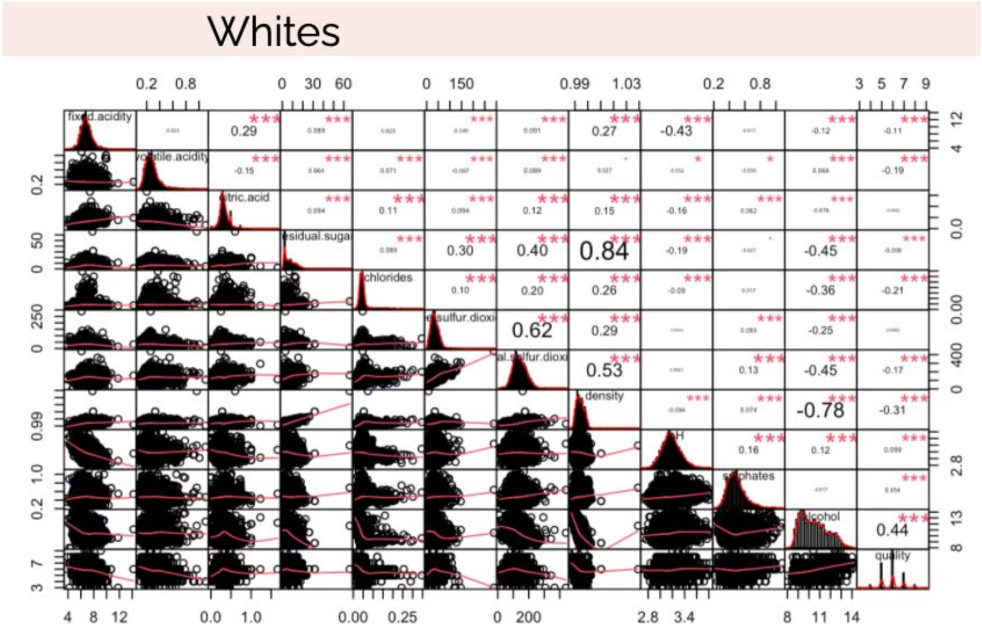
We also are interested in examining the correlation between all variables, therefore we can create a correlation heat map, in this case, since we can look at the correlation between different variables.


Reds

Based on the correlation plot of red wine, we can see the highest correlations with the following:

1. Quality with alcohol and volatile acidity.
2. Fixed acidity with citric acid, density and pH.
3. Volatile acidity with citric acid.
4. Citric acid with density and sulphates.

There may be multicollinearity between those variables.


Whites

Based on the correlation plot of white wine, we can see the highest correlations with the following:

1. Quality with alcohol and density
2. Fixed acidity with citric acid, density and pH
3. Residual sugar with free sulfur dioxide, total sulfur dioxide, density and alcohol
4. Chlorides with density and alcohol
5. Free Sulfur dioxide with total sulfur dioxide, density and alcohol
6. Total sulfur dioxide with density

There may be effects of multicollinearity between those variables, which must be carefully considered during our experimental analysis. It seems like both white wine and red wine have slightly different predictors that are correlated with each other. We will look at those variables more carefully in our analysis phase to decide what variables we should keep and what should be dropped.

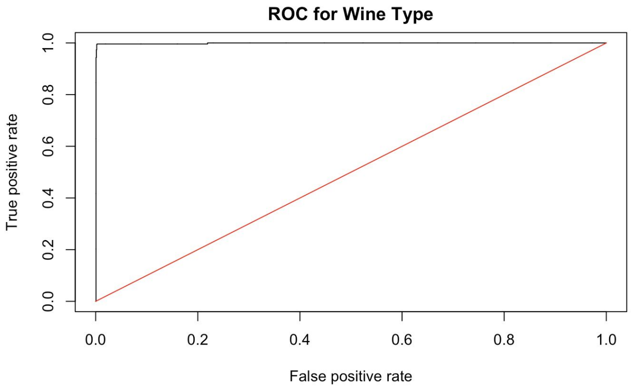
3: Detailed Analysis

3.1: Predicting Red vs. White Wine

From a real-world perspective, a winemaker would be interested in finding what variables contribute to making a wine classically "red" or "white". To explore this relationship, we will create a binomial logistic regression model to determine if we can accurately predict whether a wine is red or white based solely on its characteristics. Afterward, we will be able to investigate what characteristics contribute to these results. To begin, we merge together rows of the red and white datasets, append a column that describes if a wine is 'red' or 'white', and then randomly split the data into a 70, 30 train-test split, respectively. For this scenario, we also merged the categories of 'quality' into two levels: 'low' and 'high' corresponding to wine quality ratings of 1-5, 6-10, respectively.

We initially fit a logistic model with all 12 variables predicting type ('red', 'white'), After doing so, we were able to see that the Wald Test for several predictors were insignificant: sulphates, pH, citric acid, and the quality level. We performed a delta ΔG^2 test to see if we could drop all of these predictors together from the model. The null hypothesis was that $\beta_{\text{sulphates}}=\beta_{\text{pH}}=\beta_{\text{citric acid}}=\beta_{\text{quality level}}= 0$: these predictors can be dropped altogether from our model. The alternative hypothesis is that at least one β_i above is not equal to 0: we cannot drop these predictors all together from the model. After comparing the residual deviance of the reduced model with the residual deviance of the full model, the ΔG^2 test statistic was 8.86, which corresponds to a p-value of .0647. Because this p-value is greater than an appropriate alpha level of .05, we fail to reject the null hypothesis: we can drop sulphates, pH, citric acid, and quality level all together from the model. The chosen, reduced model is described below, where π is the probability of a wine being ‘red’:

$$
\begin{aligned}
log(\frac{\pi}{1-\pi}) = {} & -2149.61 + 2.36 * alcohol + 2142.71 * density + \\
& (-0.05) * total.\,sulfur.\,dioxide + 0.06 * free.\,sulfur.\,dioxide + \\
& 20.45 * chlorides + (-0.87) * residual.\,sugar + 6.66 * volatile.\,acidity + (-0.68) * fixed.\,acidity
\end{aligned}
$$



To test the validity of the model, we briefly created a ROC curve and AUC: The resulting ROC plot gives us confidence that the model is performing well. The accompanying AUC is .99. However, we must hesitate to automatically assume this means the model works well because the sample sizes of the two classes (red, white) are very uneven. As a more appropriate test of this model, we will create a confusion matrix based on testing the model on the 30% test split of the data. To create the confusion matrix, we set the threshold to 0.5, because the significance of a false positive vs. a false negative in this scenario is neutral. The resulting confusion matrix is:

	FALSE	TRUE
White	1453	3
Red	3	491

With a false positive rate of .002 and a false negative rate of .006, and overall accuracy of .997 on new data, our model has proven to be robust and has predictive ability on new data. Thus, we can now comment on the real-world applicability of this model by looking at its coefficients to determine some of the distinctions between red and white wines by their characteristics. Because the coefficients of the predictors alcohol, density, free sulfur dioxide, chlorides, and volatile acidity are positive, and π is the probability of a wine being red, higher levels of these predictors (increasing one at a time in the presence of the others) are associated with red wines. On the contrary, because the coefficients of total sulfur dioxide, residual sugar, and fixed acidity are negative, higher levels of these predictors (increasing one at a time in the presence of the others) are associated with classically white wines. For a novice winemaker, these insights can help with the production of new wines, depending on if they are trying to make a red or white wine. Below we briefly show the results for 10-fold cross-validation of the full vs. reduced model:

Average Accuracy per Logistic Model (Wine Type) for 10-Fold Cross-Validation	
Complete Logistic Regression Model	0.99432
Reduced Logistic Regression Model	0.99432

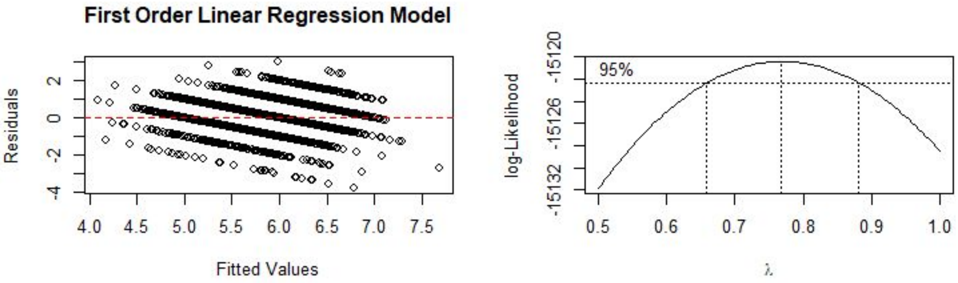
Average AUC per Logistic Model (Wine Type) for 10-Fold Cross-Validation	
Complete Logistic Regression Model	0.99578
Reduced Logistic Regression Model	0.99582

3.2 : Predicting Quality of Wine: Performance of a Linear vs. Logistic Model

Of great importance in the world of wine is crafting a wine that is considered of the highest quality. Therefore, it is essential to understand what characteristics of the wine are generally associated with better quality wines and what mistakes to avoid to prevent a ‘bad batch’. To predict wine quality, we started with creating a linear regression model to predict quality, viewing it as a quantitative variable. Once this approach proved to not have high accuracy, we then created a binomial logistic regression model to predict a ‘high’ or ‘low’ quality wine, with better performance.

3.2.1: Performance of Wine using a Linear Regression Model

In this section, we utilize a linear regression model in order to predict the quality of wine. From the wine data, the response variable is a discrete numerical variable that ranges from 3 to 9. First, we fitted a first order multiple linear regression model with the 12 predictors (no interaction) from the data set. This model has a residual standard error of 0.7331 on 6484 degrees of freedom, a multiple R-squared value of 0.2965, and a p-value less than 2.2e-16. From this p-value, we can conclude that this first order model linear regression is useful in predicting quality; however, only 29.65% of the variation in quality scores is explained by our model.

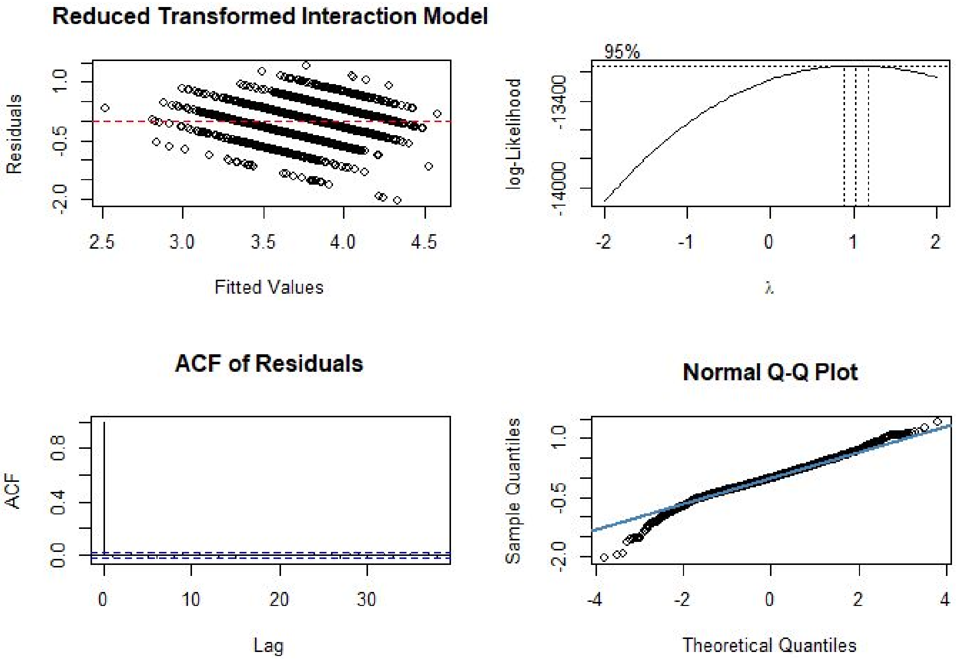


A forward stepwise regression was performed on the first-order model. The stepwise model recommended dropping “citric acid” in the prediction of quality scores. From the residual plot of the reduced first-order model, their correlation between the fitted values and the residuals is 9.42e-17. Our residuals are normally distributed and appear homoscedastic. Next, we checked for any multicollinearity between the predictors in our reduced model. The correlation matrix for all 12 predictors shows that free sulfur dioxide and total sulfur dioxide share a correlation of 0.7209; moreover, density and alcohol share a correlation of -0.6867. From our first order model, we dropped “total sulfur dioxide” and “density” since they had a lower t-value as compared to their correlated counterparts (i.e. free sulfur dioxide and alcohol, respectively). A forward stepwise regression was performed again; this time, “pH” was insignificant in improving our reduced model. A box cox normality plot was utilized to determine if our response variable could use a transformation. The 95% confidence interval for the lambda estimate of the log-likelihood was from 0.65 to 0.85, with the maximum at 0.75. Given this lambda estimate, a transformed first-order model fitted the eight predictors from the reduced model with the response variable raised to the 3/4th power (i.e. $y' = y^{3/4}$). The final, transformed first-order model is as follows:

$$\begin{aligned} y' = 2.0632 - 0.00786 fixed.acidity - 0.7861 volatile.acidity + 0.0099 residual.sugar - \\ 0.5479 chlorides + 0.0015 free.sulfur.dioxide + 0.2599 sulphates + 0.1672 alcohol \\ + 0.1534 typered \end{aligned}$$

Compared to the full first-order model, the transformed first-order model has a substantially lower residual standard error of 0.3568 on 6488 degrees of freedom. The multiple R-squared value decreased to 0.2861, and the model’s overall p-value remains less than 2.2e-16. This transformed first-order model is still significant in predicting quality scores for wines. Moreover, this model meets the requirements for ordinary least squares regression.

The transformed first-order model has denoted eight important predictors for the prediction of wine quality. Next, the full interaction model incorporates the reduced first-order model as well as all 28 interaction terms between the predictors. A backwards stepwise model was run to determine the significant variables from a total of 36 predictors. After assessing the box cox normality plot, the response variable was likewise raised to the 3/4th power. Our final, transformed interaction model incorporates 25 predictors and has a residual standard error of 0.3498 on 6471 degrees of freedom, multiple R-squared values of 0.3155, and a p-value of less than 2.2e-16.



Below we briefly show the results of 10-fold cross-validation for the models examined above:

Average RMSE per Linear Model (Quality) for 10-Fold Cross-Validation	
First Order Linear Regression Model	0.73309
Reduced Linear Regression Model	0.73727
Reduced Transformed Linear Regression Model	0.73733
Complete Interaction Linear Regression Model	0.72641
Transformed Interaction Linear Regression Model	0.7265
Reduced Transformed Interaction Linear Regression Model	0.72515

Due to the low R-squared value above, we decided that linear regression may not be the ideal way to model quality. We then proceed to create a binomial logistic model to see if this model better captures quality.

3.2.2: Performance of Wine using a Binomial Logistic Model

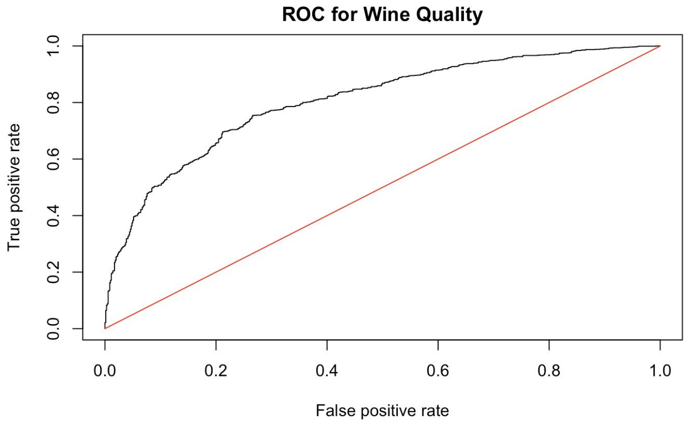
Similar to section 2.1 above, to convert the category of quality into a binomial problem, we releveled the categories such that qualities of 0 through 5 are considered ‘low’ quality and qualities 6 through 10 are considered ‘high’ quality. Based on the histograms of quality discussed in exploratory data analysis, we believe this to be a fair split of the data into approximately equal classes. We will train our model on the 70% train split of the data so that we can validate our model with the test data.

To begin, we fit a naive model with all 12 predictors predicting wine quality. Looking at a summary of the model, we were able to use a Wald Test to drop chlorides from the model: chlorides had a Z value of -1.781, which corresponded to a p-value of .075. Because this p-value is greater than an appropriate alpha of .05, we fail to reject the null hypothesis that $\beta_{chlorides}$ = equal to 0: chlorides is not useful in this model in predicting the quality of wine in the presence of the other predictors; therefore it can be dropped.

The final, reduced model is listed below, where π is the probability of a wine being ‘high’ quality as defined above:

$$\begin{aligned} log(\frac{\pi}{1-\pi}) = {} & 148.12 + 0.8 \ast alcohol + 1.97 \ast sulphates + 1.02 \ast pH + (-160.81) \ast density + \\ & (-0.01) \ast total.\,sulfur.\,dioxide + 0.01 \ast free.\,sulfur.\,dioxide + 0.14 \ast residual.\,sugar + (-0.76) \ast citric.\,acid + \\ & (-4.86) \ast volatile.\,acidity + 0.15 \ast fixed.\,acidity + 0.66 \ast typered \end{aligned}$$

To validate the model, we initially created an ROC plot:



However, similar to analysis above, although the ROC plot appears to demonstrate that our model performs decently better than random, with an accompanying AUC of .806, the reality is that the two classes (‘low’,’high’) for wine quality do not have equal sample sizes, so the results from this plot can be misleading. To have a more accurate measure of validation, we will create a confusion matrix, predicting the quality of wines in the test data split:

	FALSE	TRUE
Low	396	303
High	203	1048

The false positive rate of our model is .43 and the false negative rate is .16, and the total accuracy is .74. Indeed, this model is far from perfect, and has particularly high false positive rates, but the overall accuracy suggests that this is a somewhat decent model: indeed an AUC of .806 is much higher than guessing at random, which is an AUC of 0.5. Although not ideal, this model does appear to capture the distinction between higher and lower quality wines better than the linear model discussed in 3.2.1: we sacrifice some granularity (only being able to predict ‘high’ or ‘low’ quality instead of a number) for better accuracy. Below are the results of 10-fold cross-validation of the full vs. reduced model discussed above.

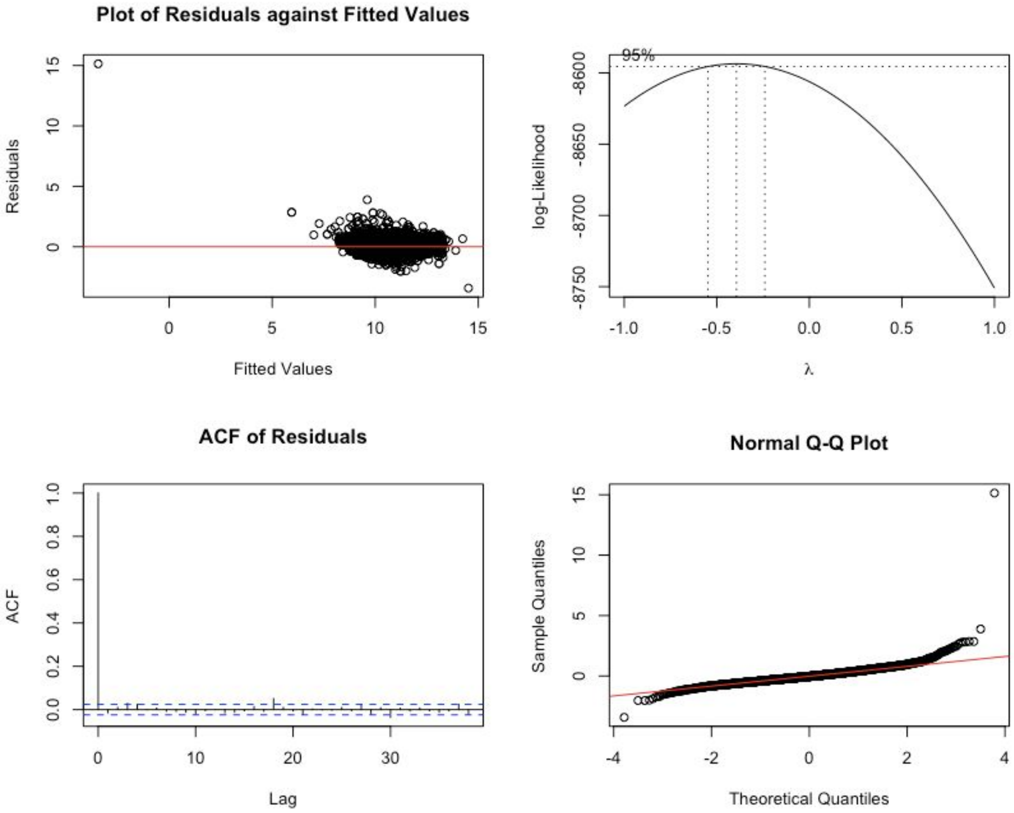
Average Accuracy per Logistic Model (Quality) for 10-Fold Cross-Validation	
Complete Logistic Regression Model	0.74099
Reduced Logistic Regression Model	0.7413

Average AUC per Logistic Model (Quality) for 10-Fold Cross-Validation	
Complete Logistic Regression Model	0.80287
Reduced Logistic Regression Model	0.80302

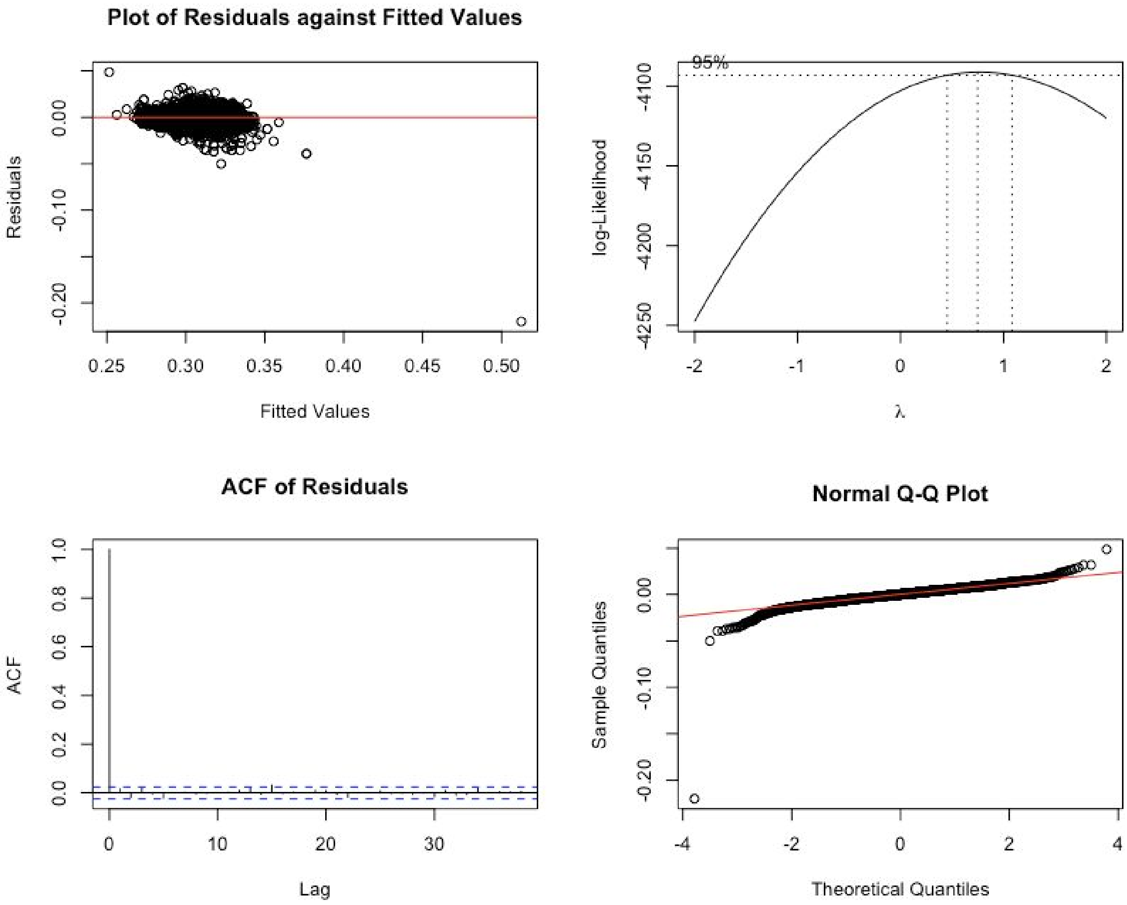
3.3: Prediction of Alcohol Content Using a Linear Regression Model

Another key real-world measure of wine is the alcohol content. The average wine contains 11.6 percent alcohol, and depending on the particular type of wine a winemaker is creating they would want to be able to methodically control if a wine has more or less alcohol than this average.

To begin the analysis of predicting alcohol, we fit a basic multiple linear regression model with the 12 predictors: sulphates, pH, density, total.sulfur.dioxide, free.sulfur.dioxide, chlorides, residual.sugar, citric.acid,volatile.acidity, fixed.acidity, quality, and type (red or white). Before we can begin any analysis, we need to make sure that the linear regression assumptions are met for our model.



The plots above show the residual plot, Box Cox plot, ACF plot, and QQ plot of the untransformed equation. The constant error variance and error mean of 0 assumptions appear to be violated looking at the residual plot and Box Cox plot. There also appears to be correlation between the errors judging by the ACF plot. The only assumption that appears to be met is that the errors look approximately normal from the QQ plot. Because the lambda value in the Box Cox plot does not overlap with one (or zero) the first transformation applied should be y’ = $y^{-.5}$.

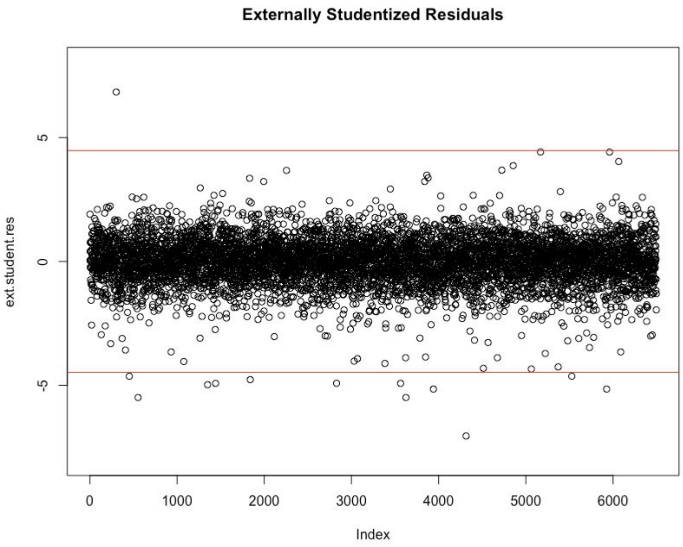


The plots above show the residual plot, Box Cox plot, ACF plot, and QQ plot of the transformed equation with $y' = y^{-.5}$. The residual plot shows that the error means of 0 assumption is approximately met (and thus linearity as well). The residual plot also demonstrates that the errors have an approximately constant variance. The normal QQ plot demonstrates that the errors are approximately normally distributed. The last assumption, uncorrelated errors, does not appear to be met judging by the ACF plot. I would argue that this has to do with the sampling method of the wine data. Indeed the wine data comes from red and white vinho verde wines from the north of Portugal -- this specificity of the way this data is collected suggests that the data is not from a random sample. Indeed, it is not surprising then that wines from a specific region of Portugal have correlated errors. Because this is the only assumption not met, we will stick with this current model due to the justification above. However, we must be cautious in the interpretation of results and predictions using this model.

Looking at a summary of this transformed model, it appears that not all of the coefficients are statistically significant. Indeed, we shall carry out a t test to determine if we can drop total sulfur dioxide from the equation. The null hypothesis is $\beta_{\text{total sulfur dioxide}}$= 0: total sulfur dioxide is not useful in predicting wine alcohol content given the other predictors. The alternative hypothesis is $\beta_{\text{total sulfur dioxide}}$!= 0: total sulfur dioxide is useful in predicting wine alcohol content given the other predictors. The resulting t test statistic is equal to 1.309, which corresponds to an p-value of .191. Because this p-value is significantly greater than a reasonable alpha of .5, we fail to reject the null hypothesis: total sulfur dioxide is not useful in predicting wine alcohol content given the other predictors. Therefore, we favor a reduced model not containing total sulfur dioxide. The adjusted R^2 is 0.8213, which suggests that a very high proportion of the variance in alcohol is explained by this model. The reduced multiple linear regression equation is:

$$\begin{aligned} y^{-0.5} = {} & -8.703 + (-0.01434) * sulphates + (-0.03823) * pH + 9.280 * density + \\ & 0.00005214 * free.sulfur.dioxide + 0.01525 * chlorides + -0.003146 * residual.sugar + \\ & -0.006261 * citric.acid + -0.00947 * volatile.acidity + -0.007612 * fixed.acidity + \\ & -0.001359 * quality + -0.01612 * typered \end{aligned}$$

Now that we have the final model, let us quickly consider potential outliers. The plot of the externally studentized residuals are below.



From the plot above, there do appear to be some potential high and low outliers. However, when running Cook’s Distance test, we only have one influential point. Given the size of the data set we will not remove this point but need to keep it in consideration when interpreting results.

In this scenario with the model above, we are less interested in predicting new values than analyzing the coefficients of the predictors to see how they contribute to the alcohol content of a wine. Because density, free sulfur dioxide, and chlorides have positive coefficients, increasing these characteristics (one at a time, holding the other predictors constant) in the wine-making process will increase y’ but **decrease** y (because $y' = y^{-.5}$): lowering alcohol content. If one is interested in making a more alcoholic wine, consider increasing (one at a time, holding the other predictors constant) sulphates, pH, residual sugar, citric acid, volatile acidity, and fixed acidity because these coefficients are negative: increasing these decrease the predicted value for the transformed model but **increase** y. Below, we briefly show the results of 10-fold cross-validation as we transformed and reduced the model.

Average RMSE per Linear Model (Alcohol) for 10-Fold Cross-Validation	
First Order Linear Regression Model	0.49526
Reduced Transformed Linear Regression Model	0.46363

4: Conclusion

In the report, we explored several questions that would provide real-world value to a winemaker interested in understanding the relationships between the numerous characteristics of the wine they are producing. We first explored a model predicting if a wine is ‘red’ or ‘white’, whose accuracy proved to be very high in predicting on new data. Such a model has real-world value because we examined the signs of the coefficients of the model to see how changing the predictors (one at a time) makes a wine more classically ‘red’ or ‘white’. We also explored what determines the alcohol content of a wine, and produced a model with a very high adjusted R^2. By analyzing the signs of the coefficients of the model we were able to understand how increasing or decreasing each predictor (holding the others constant) contributes to the overall alcohol content of the wine. Since this model had low levels for the VIF measure, we were confident that multicollinearity was not significantly present in our reduced model, so this interpretation of the coefficients was justifiable. Finally, we explored the ways to model quality: we began with a linear regression model which performed not ideally, then moved to a binomial logistic regression model that proved better, but not perfect, at distinguishing between ‘high’ and ‘low’ quality wines.