Vinho Verde Portugal Wine-Predicting Quality

Ben Wilson

8/12/2021

Executive Summary

Vinho Verde comes from a small province in Northern Portugal known for its white, red, and rosé wines. These wines are known for their zapping acidity, carbonation, and lower alcohol content. This project provides an analysis and evaluation to explore how quality is related to the physicochemical and sensory variables and how such variables relate to one another. Our overall goal for this project is to use corresponding visualizations and regression to answer three questions about the wine focused on addressing the impact of attributes on the type of wine (white versus red), the quality of wine (scores 1-10) and how the wine ranks in quality score (high quality versus mid to low quality). The datasets used describe more than 6,000 red and white Vinho Verde wines with roughly 75% equating to white wines and 25% equating to red wines. Our dataset provided was clean and complete, so outside of consolidating the wines into a single file we proceeded to perform exploratory data analysis (EDA) through visualizations of the various relationships between the 12 attributes. From our EDA, the key learnings which influenced our modelling process included determining how several of the predictors related to one another on a surface level as well as identifying stark differences in the attributes representing wine quality for red and white. These learnings led our team to modelling red and white wine separately and examining each dataset for interactions between variables.

After performing a series of analyses on the data in order to answer our three questions, the following results were found:

- The color of wine can be predicted with 99% accuracy when using volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, alcohol, and quality rating as predictors. As citric acid, residual sugar, and total sulfur dioxide increase in amount, the log odds of a wine being white increase. As volatile acidity, chlorides, density, alcohol content, and quality increase in amount, the log odds of a wine being white decrease.
- The quality of wine on a 1-10 scale can be predicted using a combination of the original 12 predictors, with a different approach for red wines than white wines
    - Volatile acidity, chlorides, and total sulfur dioxide negatively affect the quality of both white and red wines.
    - Free sulfur dioxide, sulphates, and alcohol content positively affect the quality of both red and white wines.
    - White wine quality is also positively affected by its residual sugar and higher pH values, and it is negatively affected by fixed acidity.
    - An increase in pH of a red wine negatively affects its quality.
- Quality scores, high or mid-to-low, should also be predicted differently for red and white wines
    - Red wine scores can be predicted with 85.98% accuracy when using volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol content as predictors. Increases in volatile acidity, chlorides, and total sulfur dioxide increase the log odds of a wine being white. Increases in free sulfur dioxide, sulphates, and alcohol content decrease these odds.
    - White wine scores can be predicted with 77.09% accuracy when using fixed acidity, volatile acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol content as predictors. An increase in volatile acidity, chlorides and total sulfur dioxide increase log odds of a wine being white. Fixed acidity, residual sugar, free sulfur dioxide, pH, sulphates, and alcohol content decrease those odds when increased.

In conclusion, we can state with a high degree of certainty, all else held equal, that we can predict the color and quality (in terms of high quality vs. low-to-mid quality) of Vinho Verde wine using the 12 physicochemical and sensory attributes provided.

Research Questions

The research questions we explored for the purposes of our study of Vinho Verda wine were divided into three parts given the nature of the data:

- Can we predict the color of wine (white versus red) based on the physicochemical and sensory attributes?
- How do the attributes of the wine impact the level of quality as a ranking?
- Which attributes will lead to high quality wine (score 7-10) versus low-to-mid quality wine (1-6)?

The attributes referenced include the physicochemical and sensory attributes as well as the numerical wine ranking.

Data Description

The wine quality dataset was sourced from the UCI Machine Learning Repository. Two datasets were included for the population: red and white Vinho Verde wine samples from the north of Portugal. Due to privacy and logistical circumstances, the datasets do not include information on the type of grapes nor the brand of wine. Attributes were subset into input variables based on physiochemical responses and output variables based on sensory data. The input variables consisted of the following attributes:

- Fixed acidity - Acids involved with wine which are fixed or unable to evaporate easily
- Volatile acidity - Amount of acetic acid within the wine (high levels may lead to vinegar taste)
- Citric acidity - Adds 'freshness' and flavor to wines (generally found in small quantities)
- Residual sugar - Amount of sugar remaining after the fermentation process completes
- Chlorides - Amount of salt within the wine
- Free sulfur dioxide - Free form of SO2 that exists in equilibrium between molecular SO2 and biosulfite ion (prevents microbial growth and oxidation of wine)
- Total sulfur dioxide - Amount of free and bound forms of SO2
- Density - Observes density depending on the percent alcohol and sugar content
- PH - Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)

The output variables consisted of the following attributes:

- Quality- based on sensory data determines the quality of wine determines by a scale of (0-10)

Exploratory Data Analysis (EDA)

In order to assist our team in the investigation of the relationship between quality ranking and the associated physicochemical variables, we initially conducted a series of exploratory data analysis (EDA) steps to expand our understanding of the relationships and trends that exist. Additionally, we set out to understand what similarities high quality wine has with the given attributes versus low-to-mid quality wine.

Given that the red and white Portuguese Vinho Verde wine was provided in two separate data frames, our team initially consolidated each into a single frame for analysis throughout the project, creating a new 'Color' attribute in turn which identified the 'red' and 'white' color. Prior to developing a series of informative visualizations, our team deemed it necessary to provide categorical labels to the wine quality rankings which are determined by Semantic Scholar when using quality rankings 1-10. The categorical rankings include:

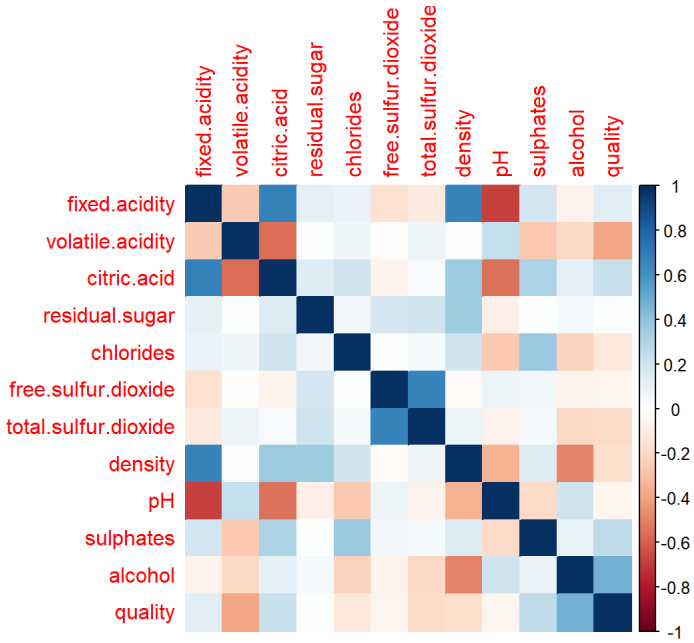
- Quality = 1-4: Undrinkable
- Quality = 5: Pretty Bad
- Quality = 6: Fair
- Quality = 7: Quaffable
- Quality = 8: Very Good
- Quality = 9-10: Excellent

For our third question, we deemed those wines which ranked 7-10 as our high-quality wines whereas all other wines of a lower rank were considered low-to-mid quality wines.

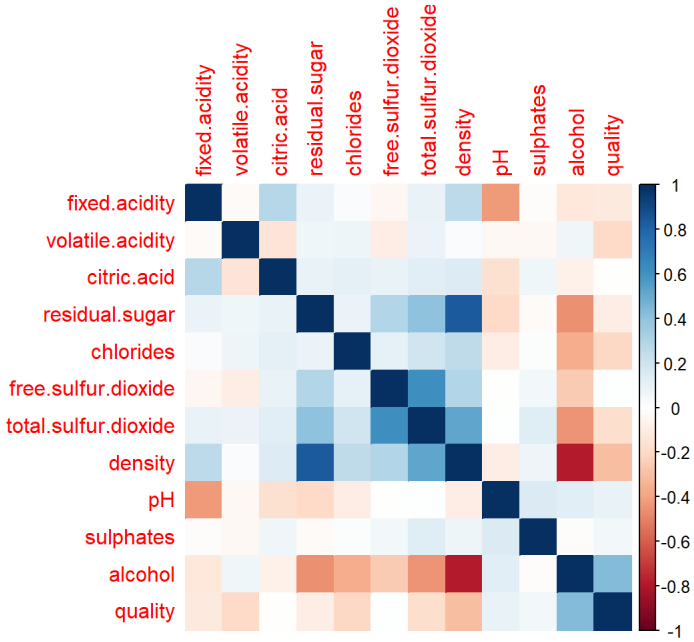
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923
##  Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970
##  Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.210   Median :0.5100   Median :10.30   Median :6.000
##  Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
##  3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
##     color
##  Length:6497
##  Class :character
##  Mode  :character
##
##
##
```

In reviewing the summarized consolidated data, initial observations determined that there were 1599 red wine scores versus 4898 white wine scores. Given the differences in population size as well as the differences we found in the physicochemical inputs between white and red wines, we noted the higher white wine observation count as an important point for further exploration during EDA.

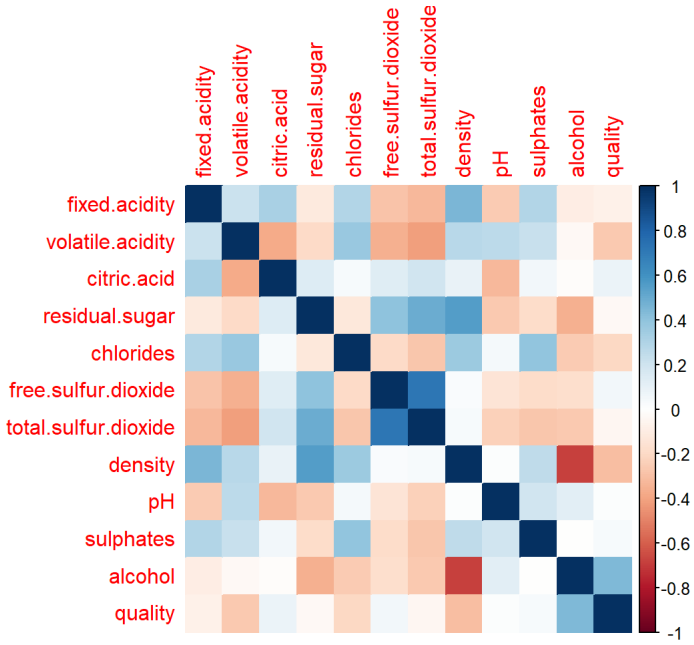
Correlation of red wine attributes:



Correlation of white wine attributes:



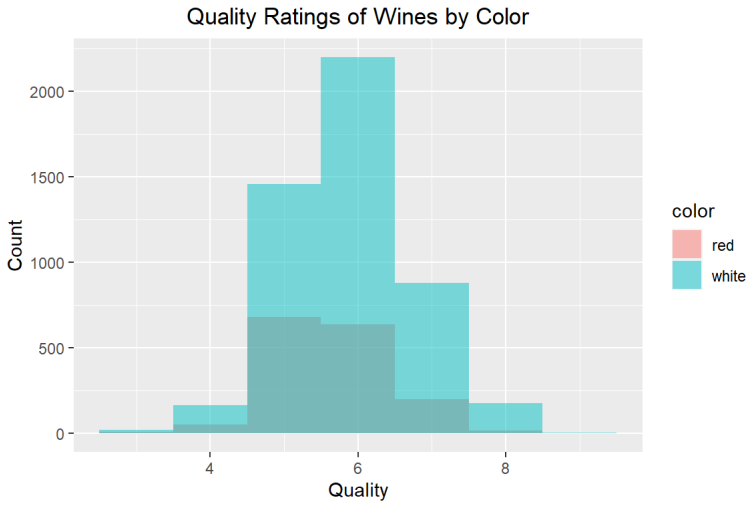
Correlation of all wine attributes (red and white):



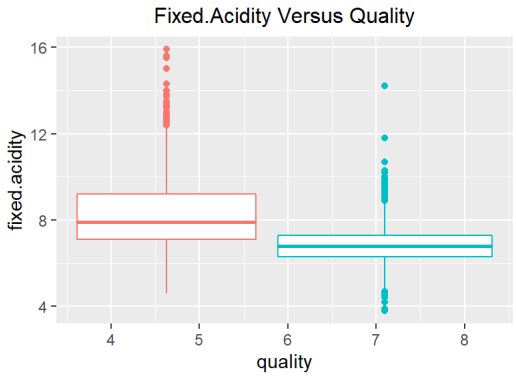
When examining red and white wine feature correlations, we saw a strong negative correlation between alcohol and density as well as a strong positive correlation between alcohol and quality. The second of the two surprised our team as we did not expect the level of alcohol to have an impact on the quality rating. A second strong positive correlation that we did happen to expect existed between free sulfur dioxide and total sulfur dioxide, given that as free sulfur dioxide increases, total sulfur dioxide increases proportionally.

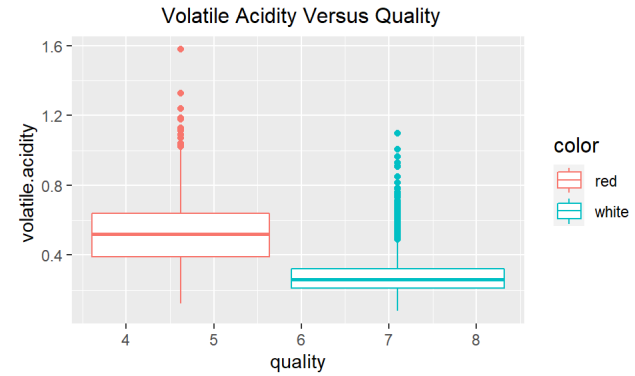
In reviewing feature correlation for red wines, we saw a strong negative correlation between pH levels and fixed acidity and a strong positive correlation between citric acid and fixed acidity (as one acidity level increases so to does another) as well as density and fixed acidity. Saying as one type of acidity increases so does another, we saw to be incorrect with red wines though given a strong negative relationship between citric acidity and volatile acidity. The takeaway was that acidity will be a crucial factor in examining red wine color.

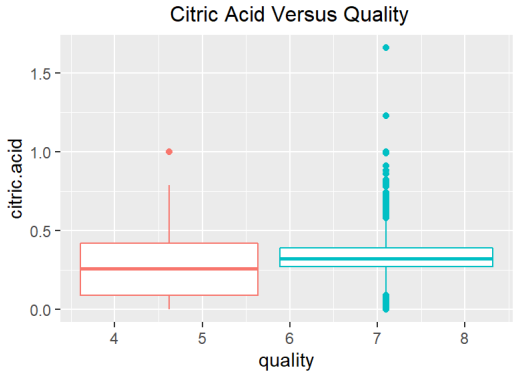
With white wines, unsurprisingly we saw a high correlation between residual sugar and density (white wine known for being higher in sugar) as well as a strong negative correlation between alcohol and density (as white wine is known less for its 'body' compared to red wine). The takeaway was that density will be a crucial factor in differentiating between white and red wine colors. The outcome from the correlation plots for our team was an understanding that multicollinearity may exist between pairs of attributes as well as that we may need to have multiple equations when predicting quality of wine given that high quality reds have a different balance of attributes compared to high quality whites.

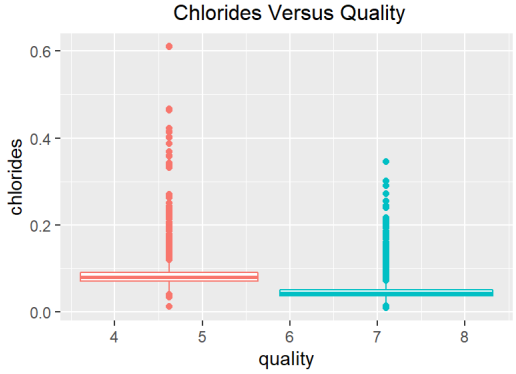


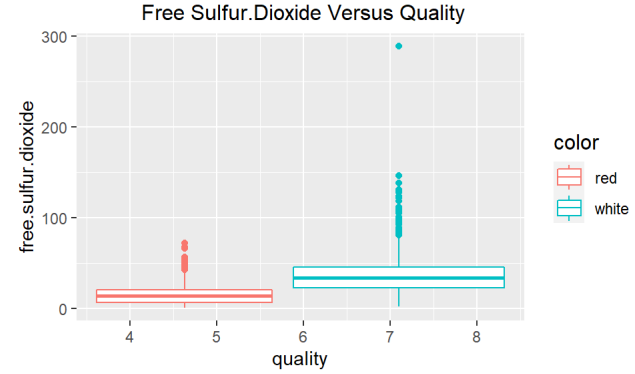
From viewing the quality distributions by color of wine within a histogram, we saw a similar distribution of quality rankings per wine color. Although those attributes for white wine will have a greater weight given that white wine observations encompass 75% of total observations, we were reassured to discover that the quality rankings between colors remained evenly distributed. Additionally, the data appears to be fairly normally distributed with no clear skew impacting the quality rankings.

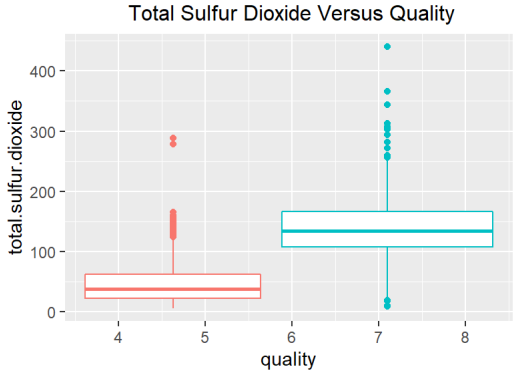


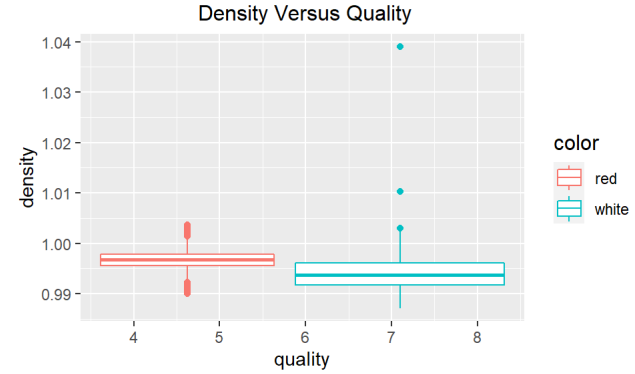


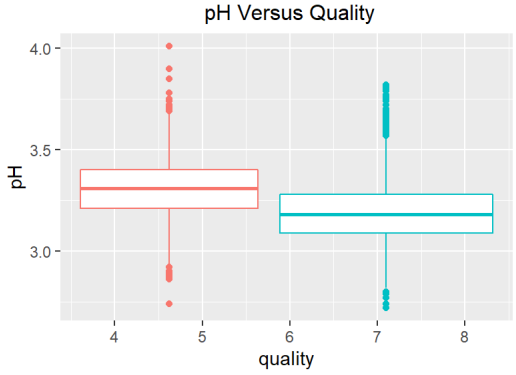


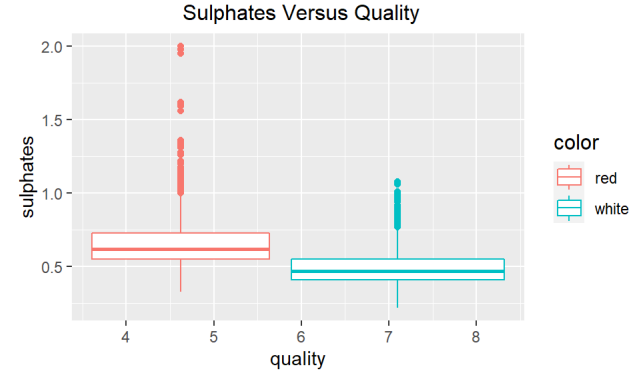




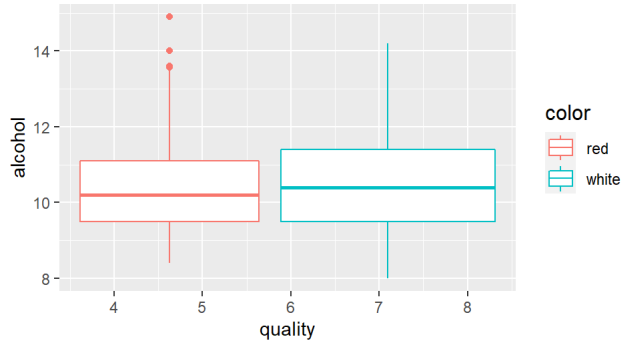








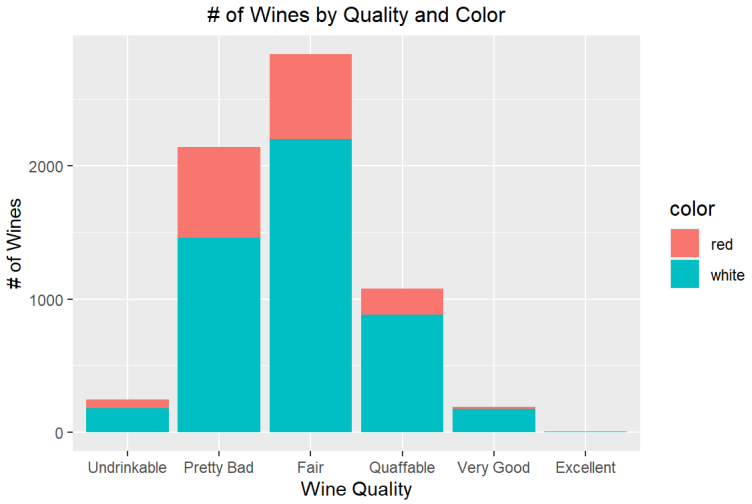
Alcohol Versus Quality



We next further evaluated the distribution of our dataset based on a five number summary (minimum, Q1, median, Q3, and maximum) within boxplots for various attributes, differentiating for wine color. After running boxplots with the respective variables, we noted similar interactions between red and white wines for chlorides, alcohol, and sulphates. The interquartile range for alcohol, sulphates, and chlorides is approximately the same for both red and white wines. However, the median for red wines is slightly higher than that for white wines with sulphates and chlorides. We saw a similar distribution for the median of the alcohol versus quality boxplot as well with sulphates and chlorides presenting outliers in both red and white wines that caused variability in the distribution.

Although we note similarities between some attributes, the rest of the box plot distributions indicate significant variability in the distributions of red versus white wines. Attributes like sulfur dioxide, citric acid, residual sugar, and fixed acidity show the medians vary greatly in red and white wines and outliers are present in both. The interquartile range for sulfur dioxide, citric acid, residual sugar, and fixed acidity is also drastically different for both red and white wines.

Since the similar interactions could be due to a result of multicollinearity or variable interaction effects, we identified the need to test for such an issue within the feature selection process. Additionally, we chose to proceed with evaluating the red and white wine models separately when assessing quality as the attributes for each wine type appear materially different between each color.



Similar to our distribution chart previously, our team noticed a fairly even distribution of red and white wines by wine quality. Additionally, when reviewing those wines that will appear in our high quality category (scores 7-10) versus low-to-mid quality category (1-6), the distribution evens out as well. Such even distributions provided us confidence in having sufficient sample sizes to perform our upcoming categorical regression without the need to oversample as can often be the case for low sample observations of one category. When reviewing the data by quality categories, we saw as well that it kept a fairly normal shape, with a slight bias toward more lower quality wines although not enough to be concerned with for the purposes of our modelling.

Model Development

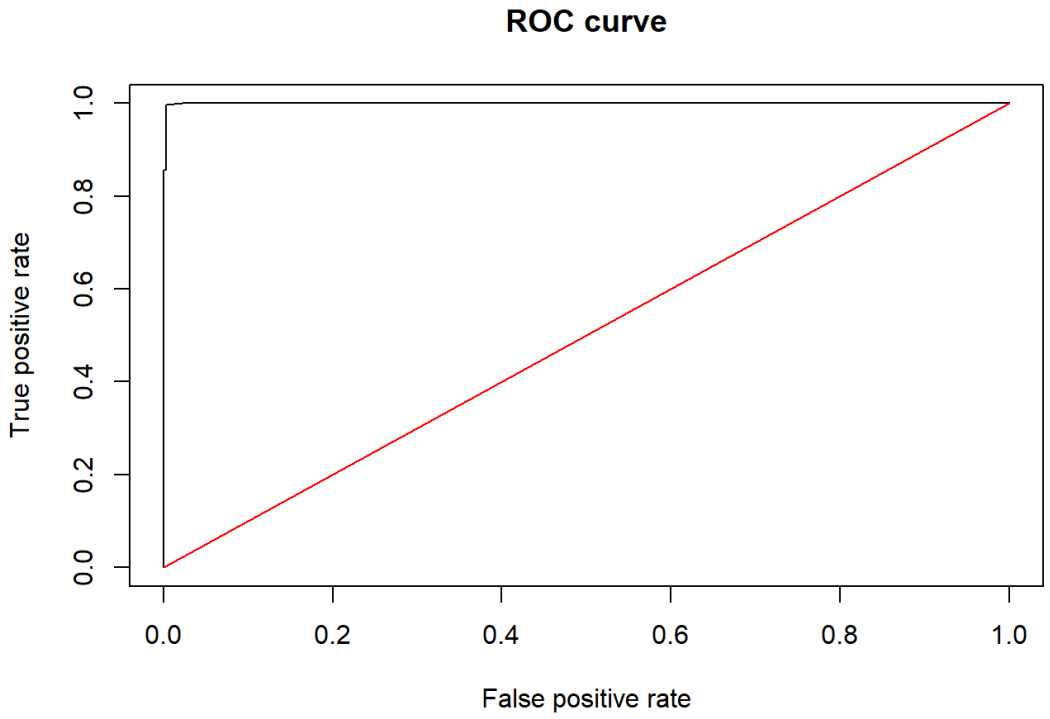
Predicting Wine Color

Following our initial EDA to enhance our understanding of the data attributes relationships, we focused on solving our first question 'Can we predict the color of wine (white versus red) based on its physiochemical and sensory attributes?' Although we included a 'Color' attribute upon binding of datasets, we had to factor this column for the purposes of R reading it a binary attribute (0 or 1) for our logistic regression. Logistic regression was chosen for this analysis due to the problem being categorical in nature (predicting red or white).

The full wine population of red and white wine data was randomly split into train and test portions, 80% of observations splitting into train whereas 20% of observations splitting into test. This was done so that the model can train on the same attributes we would receive for future data while minimizing the effects of data discrepancies in order to understand the patterns that exist for determining whether the wine is of a red or white color. In performing the logistic regression, we conducted Wald tests on individual predictors. The predictors fixed.acidity and pH were removed due to low significance in producing an enhanced modelling output, resulting in the following variable equation after nine modelling iterations:

$$color = 1701 - 6.445volatile.acidity + 3.373citric.acid + 0.9005residual.sugar - 25.82chlorides - 0.04934free.sulfur.dioxide + 0.04662total.sulfu$$

To review our model performance, our team used the Receiver Operating Characteristic (ROC), which tested the classification at all classification thresholds by plotting the True Positive against the False Positive rate, as well as the confusion matrix, visualizing how often our model confuses the two classes.




```
## integer(0)
```

Area Under the Curve (AUC) Output

```
## [[1]]
## [1] 0.9995305
```

Confusion Matrix Output

```
##
##         FALSE TRUE
##   red     327    2
##   white     3  968
```

With our ROC curve reflecting a 99% performance when viewing the Area Under the Curve (AUC) metric (more on this metric in question 3), our model is almost perfectly predicting red versus white wine types. From the confusion matrix (using a 0.5 threshold) that it has a False Positive rate of 1% and a True Positive rate approaching 99%. The threshold of 0.5 was chosen as it is the default in industry to start our analysis and we did not deem a higher threshold necessary given the level of our risk for an incorrect prediction being quite low (wine type versus a high-risk prediction example of say cancer). Once again, from prior EDA review, we expected a high success rate in predicting the wine color given the trending differences in attributes for each wine color. Understanding this clearly from our model prediction though, we deemed it necessary to perform feature testing and selection next prior to predicting quality of red and white wines independently.

Feature Testing

Following our EDA, we discovered hints of multicollinearity while predicting wine color based on the physicochemical and sensory attributes, creating a need to look for interaction effects. An important note to address is that two modelling processes were developed to predict wine quality in our research question 2 and research question 3, one model for red wine and one model for white wine. The reason for two models was a result of the EDA performed where there was a clear distinction between the physicochemical and sensory attributes for each color wine. Additionally, from a contextual perspective, the attributes that tend to make a high-quality red wine (full body aka density, low sugar content, bold flavor) are different from those that make a high-quality white wine (balanced sugar content, subtle flavor, crisp florescent taste). This decision to balance two models through feature selection then became the basis for our third question in predicting high-quality wine. The reason for testing whether one or more predictor variables in our multiple regression model are linearly predicted from other variables with a high degree of accuracy is to simplify our model with fewer attributes and enhance our remaining variable coefficients once the effect is removed.

White Wines

To test for multicollinearity, we identified the variance inflation factors (VIFs) for each predictor, which calculate the severity of multicollinearity through an estimation of the response variables against each other parameter. Our team used a threshold of 5 to determine if the predictor should be removed. For white wine, a high VIF of 28.23 was produced for density as well as a high VIF of residual sugar of 12.6. Density was dropped and residual sugar was kept to view the results upon being rerun which led to a desired result less than 5 (VIF of 2.2), confirming the role of density in multicollinearity. Additionally, no leveraged observations were returned from the analysis of variance (ANOVA) test performed, testing the impact of each variable by comparing the means of different samples, which would cause influential differences to our model for white wine.

VIFs for All Predictors of White Wine

```
##        fixed.acidity     volatile.acidity          citric.acid
##             2.691435             1.141156             1.165215
##       residual.sugar            chlorides  free.sulfur.dioxide
##            12.644064             1.236822             1.787880
## total.sulfur.dioxide              density                   pH
##             2.239233            28.232546             2.196362
##            sulphates              alcohol
##             1.138540             7.706957
```

Red Wines

In testing for multicollinearity within red wine, our VIF test produced a similar high VIF for fixed acidity (score of 7.767) which was removed, leading to no further observations producing a similar high VIF score exceeding the threshold of 5. Additionally, no leveraged observations were returned from the ANOVA test performed which would cause influential differences to our model for red wine.

VIFs for All Predictors of Red Wine

```
##        fixed.acidity     volatile.acidity          citric.acid
##             7.767512             1.789390             3.128022
##       residual.sugar            chlorides  free.sulfur.dioxide
##             1.702588             1.481932             1.963019
## total.sulfur.dioxide              density                   pH
##             2.186813             6.343760             3.329732
##            sulphates              alcohol
##             1.429434             3.031160
```

Predicting Wine Quality

After concluding our initial question, we examined our second question 'How do the attributes of the wine impact the level of quality as a ranking?' Given that quality is a numerical ranking, our team leveraged linear regression as a starting point to fit a model against the predictors of quality. In order to do so, we employed automated search procedures for feature selection using three directions (backward elimination, forward selection, and stepwise regression) to reduce the number of potential models considered. This meant declaring an intercept-only model as well as a full model with all predictors (except those removed for red or white during the feature selection process previously). For forward selection, model attributes were repeatedly added to the intercept only model and any prior predictors until the Akaike information criterion (AIC), an estimator of prediction error and measure of statistical model quality, could not be reduced further whereas the backward elimination begins with the full model and repeatedly removes predictors until our AIC is minimized and step-wise combines each type to arrive at a minimal AIC.

White Wine

For white wines, forward, backward, and stepwise selection all produced the same models. This model included the following nine predictors: fixed.acidity ,volatile.acidity, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates, and alcohol.

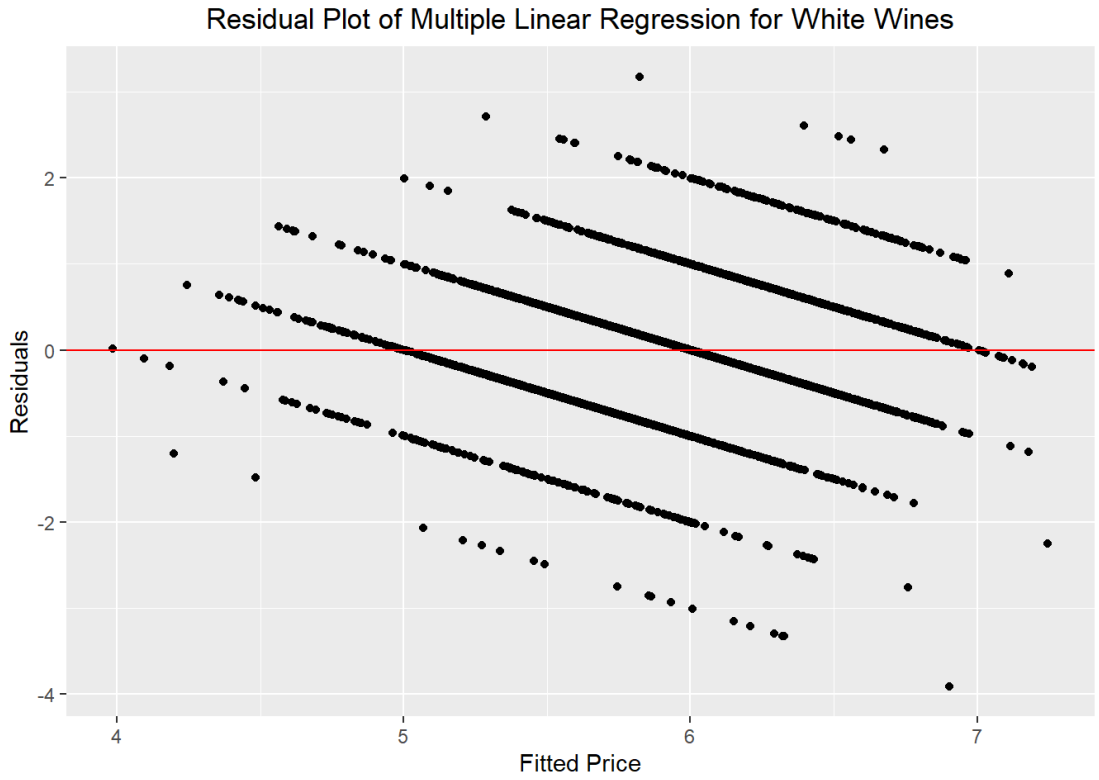
We proceeded to fit a multiple linear regression manually to see if this method would produce a model different from the one generated through the previous selection processes. When determining whether individual predictors had a significant effect on quality, we set a threshold for alpha of 0.05. From a multiple linear regression including all predictors, citric acid, chlorides, and total.sulfur.dioxide were removed as they were deemed to be insignificant. After dropping the insignificant predictors, the model was refitted. pH was found to be insignificant as well and was dropped from the model. The model was refitted once more, and the remaining predictors were all significant.

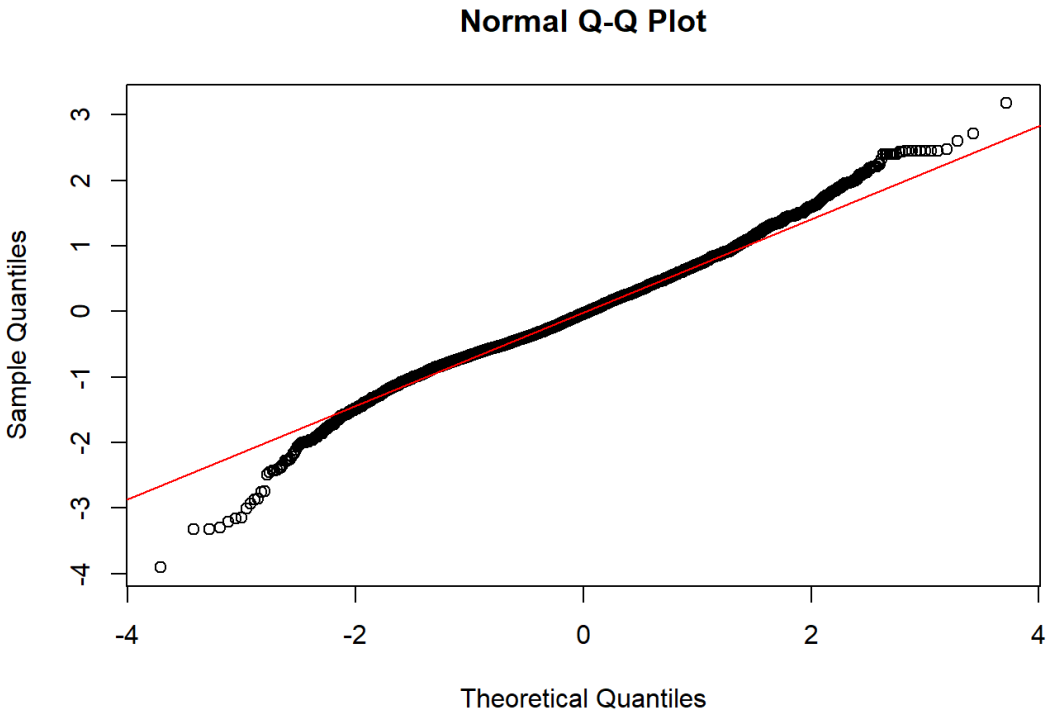
The remaining predictors included fixed.acidity, volatile.acidity, residual.sugar, free.sulfur.dioxide, sulphates, and alcohol with an adjusted R^2 value of 0.2699. The R^2 value for the step function method as a comparison was 0.2714. It is important to note that the only predictors that vary from the t-test and step function approach were chlorides, total.sulfur.dioxide, and pH. Given that our goal is to optimize the model for high performance, we proceeded with the higher R^2 value from the step function process for research question 3. If the model were to be scaled and used for production purposes, we would have next performed a cost-benefit analysis as to the impact of running a model with a few additional variables that enhances the model by ~1% to determine whether the benefit of additional performance power were worth the cost of running the model.

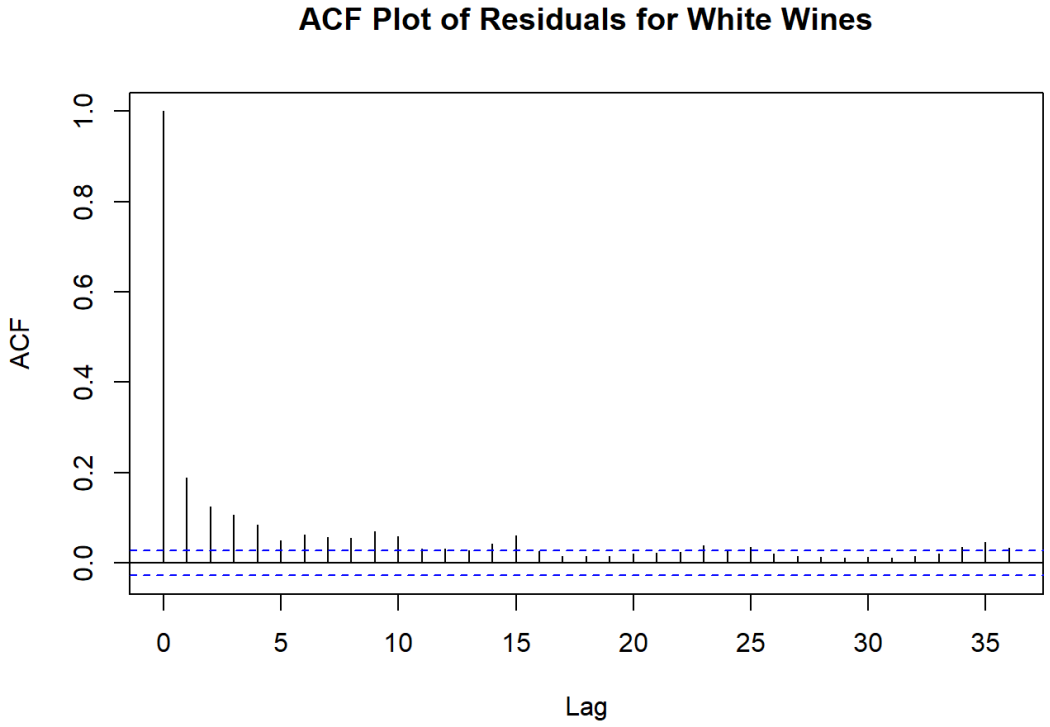
Therefore, our final model chosen had the following predictors: fixed.acidity ,volatile.acidity, residual.sugar , chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates, and alcohol. Given that output, our final linear regression equation was:

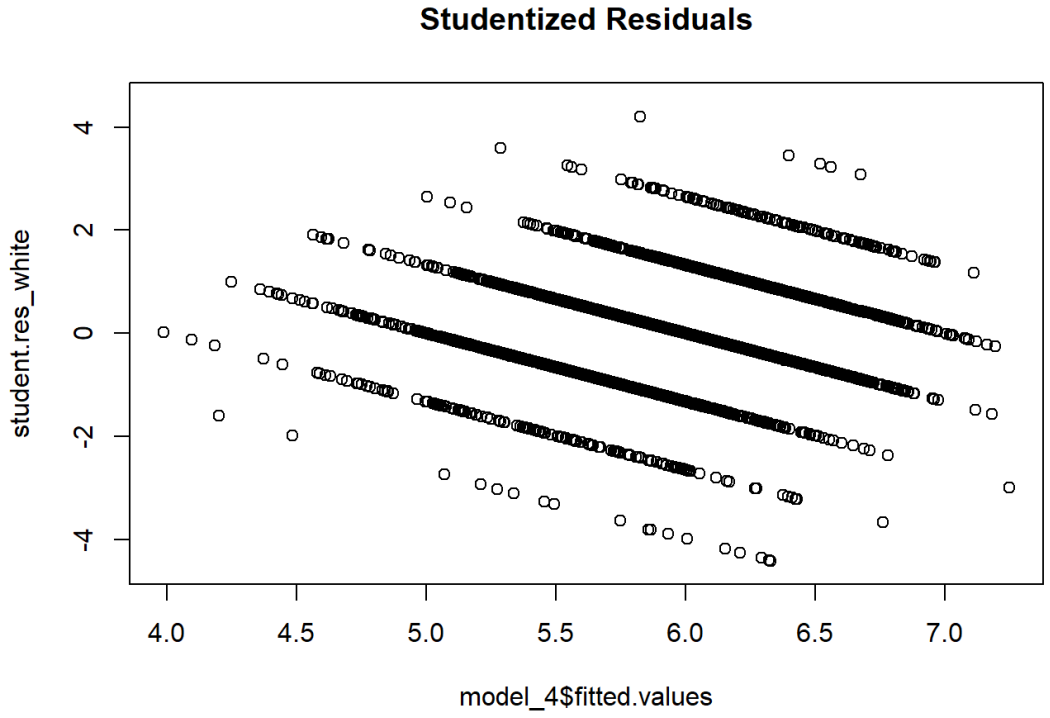
$$quality = 2.0615 - 0.05140 fixed.acidity - 1.9526 volatile.acidity + 0.02560 residual.sugar - 0.9721 chlorides + 0.004761 free.sulfur.dioxide - 0.00$$

we can interpret that these attributes are significant in impacting the level of quality in white wines. Generally, we see that fixed acidity, volatile acidity, chlorides, and total sulfur dioxide all have a negative effect on quality rating, while residual sugar, free sulfur dioxide, pH, sulphates, and alcohol have a positive effect on quality rating.


Residual Plot of Multiple Linear Regression for White Wines
Residuals
2
0
-2
-4
4
5
6
7
Fitted Price


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





As a check of our model prior to proceeding, our team produced a Residual Plot for the multiple linear regression, which represents the differences between observed and predicted values (residuals) of our observation and highlights any points not evenly distributed around the horizontal axis. The pattern of the chart depicted no issues with residuals (evenly distributed residual observations), nor did the Normal Q-Q plot, depicting any variance of observations from a theoretically perfect linear relationship, or the Studentized Residuals Plot, which examines the residuals without a constant variance. Our ACF plot shows that the observations are not as independent as we would like them to be, which may be a result of sampling from the same location. However, there is not enough dependence to believe that has a major impact on the model.

```
## Analysis of Variance Table
##
## Model 1: quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
## Model 2: quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4888 2793.4
## 2   4887 2793.4  1  0.051817 0.0907 0.7634
```

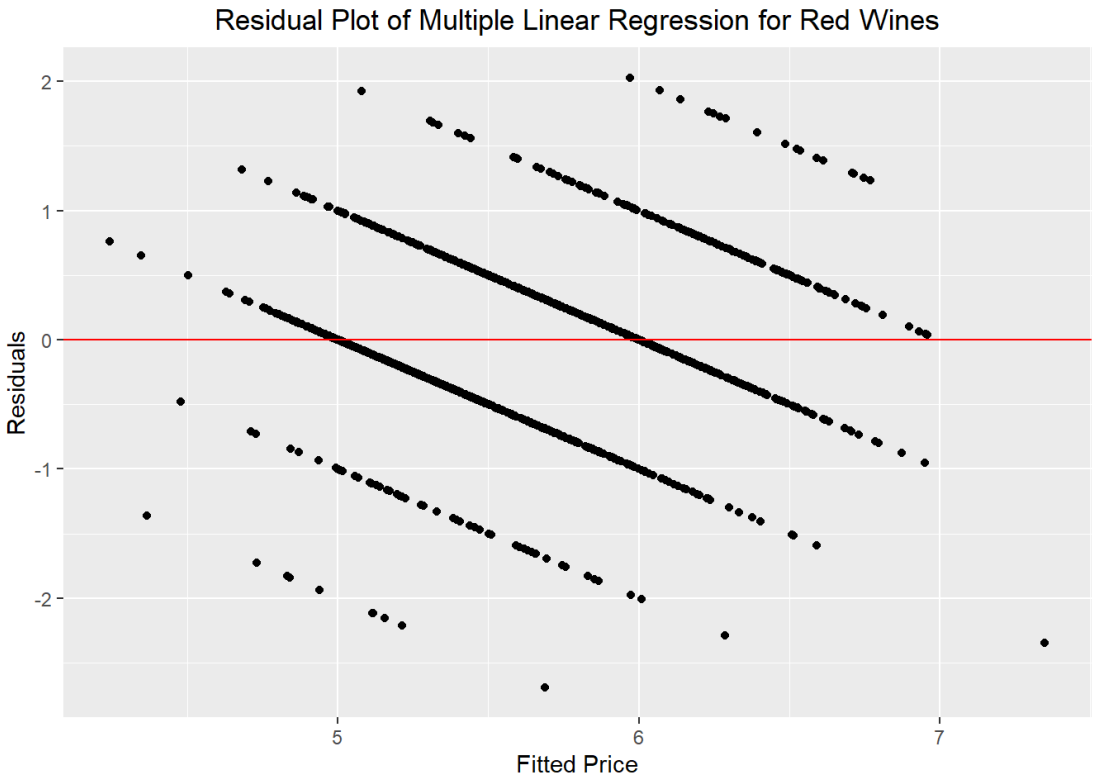
Upon confirmation of the ANOVA results identifying no other issues, as a final test to confirm observations are not impacted materially by leverages and influential points, our team used Levene's test (testing whether samples are of equal variance), Cook's distance (measuring the effect of deleting a given variable), DFFITs (measuring how much the fitted value of each observation changes as it is removed from the model), and DFBETAs (which measures how much the estimated coefficient changes when each observation is removed from the regression). Each test produced no observations of note to remove. A partial f-test was conducted to confirm that our model with a reduced number of predictors is useful compared to a full model, including all features (excluding those removed due to multicollinearity). Due to the high p-value associated with this test (0.7634), we can reject the null hypothesis and confirm that the reduced model should be used.

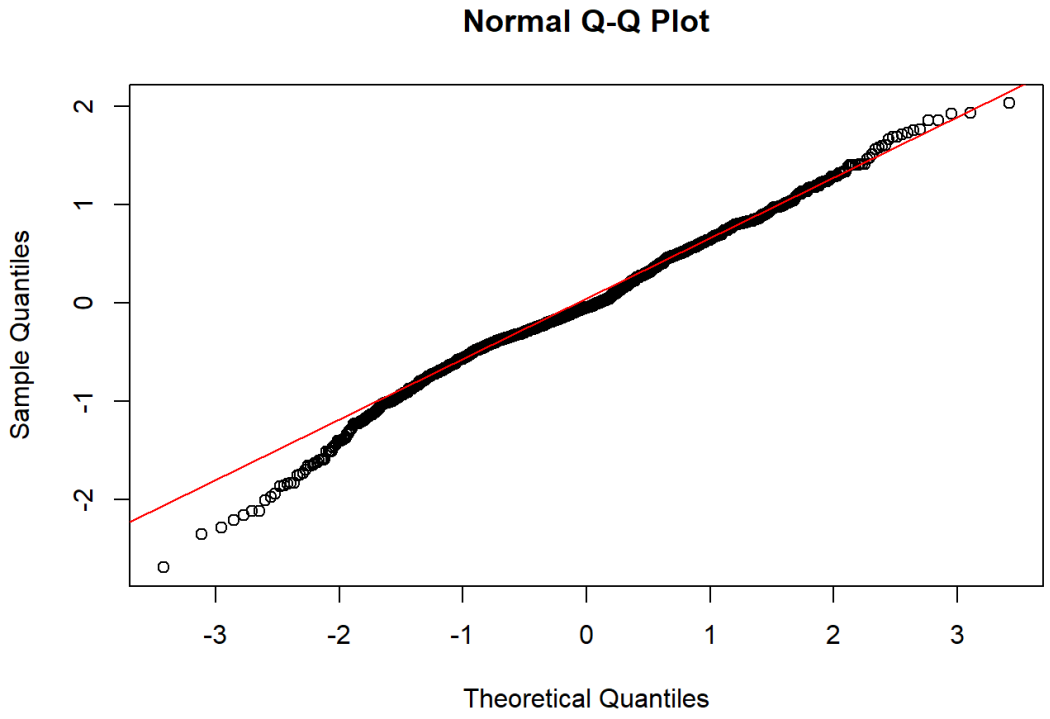
Red Wine

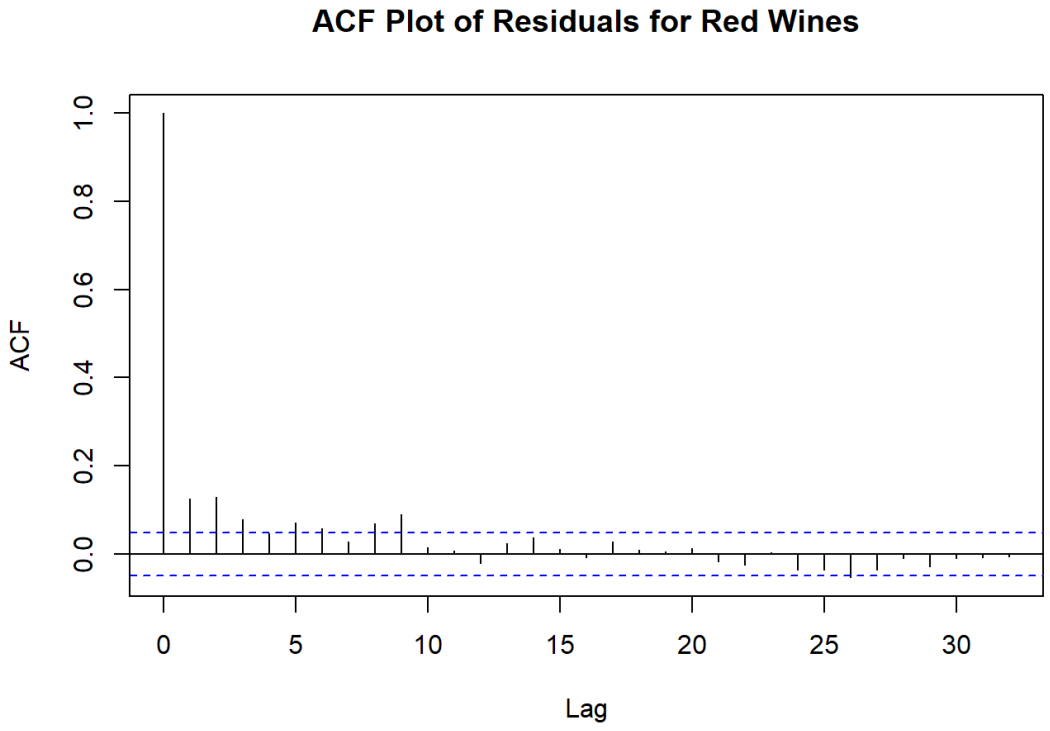
We began our red wine modelling process by creating an intercept only model, and a regression which includes all possible predictors other than fixed acidity (which was removed during feature selection). We then initiated the step function automated search procedures referenced previously. Each direction results in the same seven predictors of volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol as optimal for predicting quality of red wine. Next, we created a model using t-tests and dropping predictors with p-values greater than 0.05. Selecting predictors this way results in optimal predictors of volatile acidity, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol. Note that the only difference between the t-test approach and the step function is the chlorides predictor. However, we see that the step function approach results in an R^2 of 0.3567 while the t-test approach has an R^2 of 0.3466. We are optimizing for higher performance, so we select the model from using the step function. Our main predictors for quality of red wines are therefore: volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol.

$$quality = 4.4301 - 1.0128volatile.acidity - 2.0178chlorides + 0.005077free.sulfur.dioxide - 0.003482total.sulfur.dioxide - 0.4827pH + 0.8827su$$

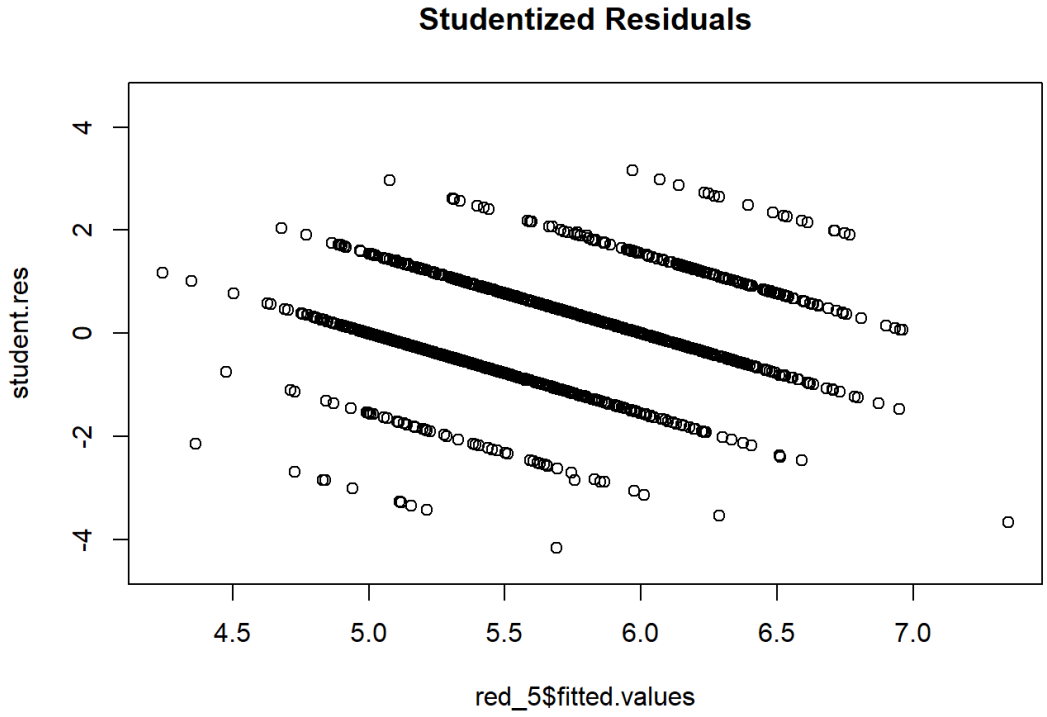
In general, we see that volatile acidity, chlorides, total sulfur dioxide, and pH have a negative effect on quality while free sulfur dioxide, sulphates, and alcohol have a positive effect on quality.







Once more our team produced the Residual Plot, Normal Q-Q Plot, and Studentized Residuals Plot without highlighting residuals of issue. Similar to white wines, our ACF plot shows that the observations are not as independent as we would like them to be. However, there is not enough dependence to believe that has a major impact on the model. Additionally, Levene's test, Cook's distance, DFFIT, and DFBETA tests were performed again with no observations highlighted for inquiry or removal. Since multicollinearity was reduced to a non-material level (given that it can never be entirely removed from a model) for red or white wine as confirmed by the VIF results and review of prior correlation plots during EDA, our team proceeded with the model development.




```
## Analysis of Variance Table
##
## Model 1: quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
## Model 2: quality ~ volatile.acidity + citric.acid + residual.sugar + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
##     sulphates + alcohol
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1   1591 667.54
## 2   1588 666.80  3   0.73687 0.585 0.6249
```

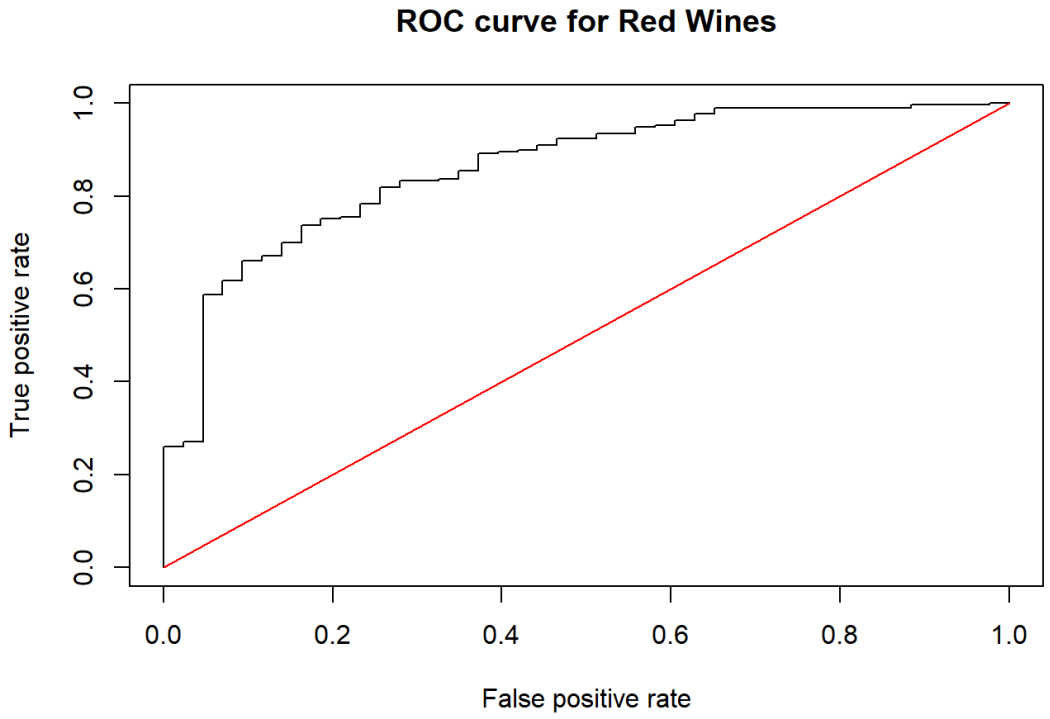
As was done for white wines, we conducted a partial f-test to confirm that our model with a reduced number of predictors is useful compared to a full model, including all features (excluding those removed due to multicollinearity). Due to the high p-value associated with this test (0.6249), we can reject the null hypothesis and confirm that the reduced model should be used.

Predicting High Quality

Once our team identified the feature selection required for examining the quality of wines as part of our second research question, we produced a logistic regression to answer our final research question 'Which attributes will lead to high quality wine (score 7-10) versus low-to-mid quality wine (1-6)?' In order to prepare the data for the regression, a binary attribute labelled 'qual_cat' was created and factored to categorize high quality wines (score 7-10) versus low to mid quality wines (1-6) as defined arbitrarily by our team.

Red Wine

To fit a logistic regression model to our red wine data, we started by once more splitting the data into training and testing sets (size of 80% and 20% respectively) as was done previously for predicting wine color although this time with the subset of red wine only observations. The model was fit using the same predictors identified as producing the optimal model for our linear model with research question 2 in order to predict wine quality, this time using our training set.




```
## integer(0)
```

Following the initial model being produced, we used the trained model to predict how well our model performs overall using split test data. The false positive rate of the model, where our model predicts that the test observation is a high-quality red wine when it is not (and vice-versa), against the true positive rate, where our model correctly predicts a high-quality red wine to be of high quality (and vice-versa), were plotted to form the ROC curve in order to assess model performance. Additionally, the ROC curve provided our team confirmation that the regression outperforms random guessing (depicted by the red linear line within the plot) for categorizing high-quality red wines by looking at the area under the ROC curve (AUC) aggregating to 0.8607 out of 1. The model is as follows:

$$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 5.1204 + 3.6526volatile.acidity + 6.9133chlorides - 0.0161free.sulfur.dioxide + 0.0151total.sulfur.dioxide + 2.2362pH - 3.1342sulph$$

Area Under the Curve (AUC) Output

```
## [[1]]
## [1] 0.860717
```

Confusion Matrix Output

```
##
##        FALSE TRUE
##   high     5   38
##   low      2  275
```

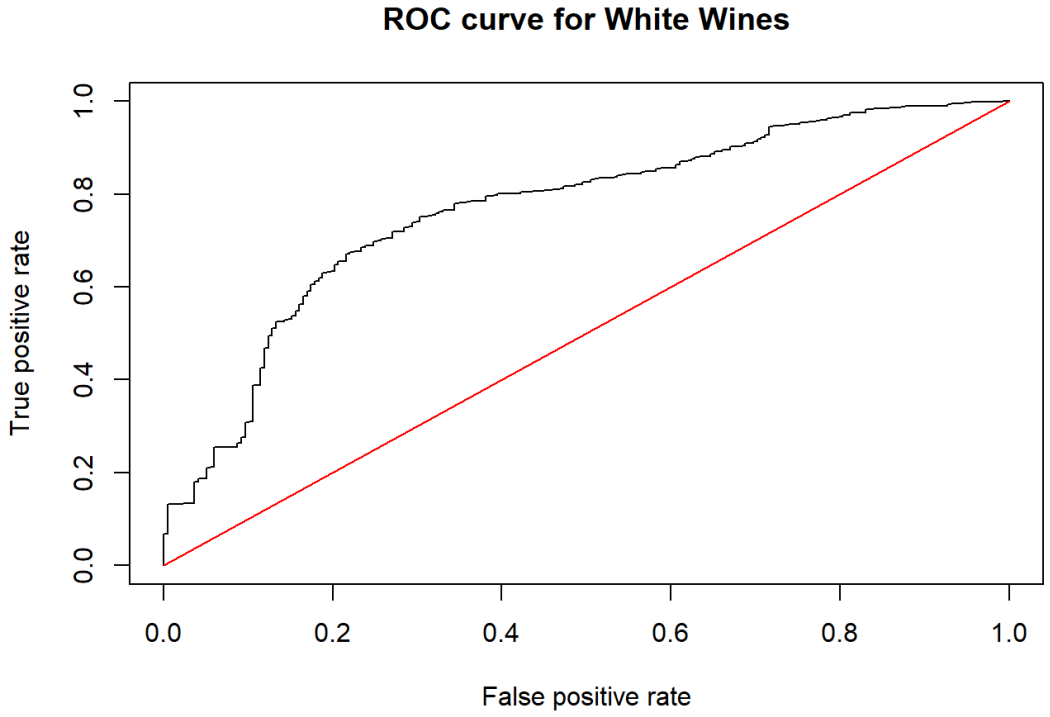
With a threshold of 0.3 defined by our team for guiding the confusion matrix in summarizing predicted results of our classification model, our confusion matrix resulted in an error rate of 0.1402, an accuracy rate of 0.8598, a false positive rate of 0.9167 and a false negative rate of 0.0037. It is worth noting that the confusion matrix results would differ given an adjusted threshold indicative of the level of risk in predicting observations incorrectly in favor of enhanced model performance. As the outputs from our model do not provide high-risk if incorrect (unlike for example incorrectly predicting cancer results), our team accepted more risk of possible incorrect predictions in favor of increased performance by using a lower threshold.

To further assess our model and determine whether all model predictors are representative of the red wine data population as produced by our data, our team ran a goodness of fit, chi-square test. Our null hypothesis was established that all the beta variables in our model were equal to zero whereas our alternative hypothesis was that at least one of these coefficients was not zero. By calculating a delta G squared test statistic and comparing it with a chi squared distribution with seven degrees of freedom, we found a test statistic of 319 with a p-value of 0. To summarize, given the p-value of 0 our team rejected the null hypothesis and concluded that our 7-predictor model should be chosen over the intercept-only model.

White Wine

Similar to the red wine modelling process, our team split the white wine data into train and test sets for performing the model fitting process. Based on the estimated linear regression model, the odds of a wine being of high-quality increased as volatile acidity, chlorides, total sulfur dioxide increase. Our final logistic regression equation for white wines is as follows:

$$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 13.8310 - 0.02815fixed.acidity + 3.5331volatile.acidity - 0.04687residual.sugar + 17.4986chlorides - 0.01344free.sulfur.dioxide + 0.$$




```
## integer(0)
```

Next, a ROC curve plot was produced to determine the performance of the model in classifying observations correctly within the test data split. The further the curve is from the diagonal line and the closer it is to the coordinate point (0,1) then the better the model is at classifying observations correctly. Since our model is indicative of being closer to the (0,1) coordinate, our model outperforms random chance.

Area Under the Curve (AUC) Output

```
## [[1]]
## [1] 0.7661634
```

Confusion Matrix Output

```
##
##        FALSE TRUE
##   high    14  204
##   low      4  758
```

Our white wine model AUC levels were computed to be 0.766, meaning that the regression continues to outperform random guessing. With a threshold of 0.03, our confusion matrix resulted in an error rate of 0.2291 with an accuracy rate of 0.7709. Our FPR rate was 0.9358 and our FNR rate was 0.0052.

For the final goodness of fit test, our null hypothesis was defined as all the beta variables are equal to zero whereas the alternative hypothesis was that at least one of the coefficients was not equal to zero. Upon calculation, the test statistic was $\Delta G^2 = 811.4406$ with a p-value equal to 0. In summary, we reject the null hypothesis as the data supports the claim that our model is useful, compared to the intercept only model.

Results & Conclusions

Model Interpretations & Context

Research Question 1

The selected model for research question 1 was developed using the following predictor variables and equation:

$color = 1701 - 6.445volatile.acidity + 3.373citric.acid + 0.9005residual.sugar - 25.82chlorides - 0.04934free.sulfur.dioxide + 0.04662total.sulf$

This reflected a ROC curve AUC performance of 99% and a confusion matrix FPR of 1% and TPR of 99%. Given the final logistic regression equation, we can interpret the model as, when predicting the color of wine holding all else constant, that there is a positive relationship between citric acid, residual sugar, and total sulfur dioxide whereas volatile acidity, chlorides, free sulfur dioxide, density, alcohol, and quality have a negative relationship with predicting the color of wine.

Research Question 2

For research question 2, two models were chosen to differentiate the prediction for white and red wine. The final model for red wine was chosen due to its superior adjusted R-squared. The model was developed using the following predictor variables and equation:

$quality = 4.43 - 1.0128(volatile.acidity) - 2.0178(chlorides) + 0.005(free.sulfur.dioxide) - .0035(total.sulfur.dioxide) - 0.4827(pH) + 0.8827(su$

Given the final linear regression equation, we can interpret that assuming all other variables remain constant, volatile acidity, chlorides and pH have a deleterious effect on the quality of red wine, while sulfur dioxide, sulphates and alcohol have a small positive effect.

The final model for white wine was developed using the following predictor variables and equation:

$quality = 2.06 - 0.051(fixed.acidity) - 1.9523(volatile.acidity) + 0.0256(residual.sugar) - 0.9721(chlorides) + 0.00476(free.sulfur.dioxide) - 0.($

Given the final linear regression equation, we can interpret that residual sugar, free sulfur dioxide, sulphates, and alcohol have a positive effect on the quality of white wine whereas fixed acidity, volatile acidity, chlorides, total sulfur dioxide, and pH have a negative impact on the quality of white wines.

Research Question 3

The final selected logistic regression models for research question 3 used to determine high quality wine for red and white wines can be seen as follows:

Red Wine

Our final red wine model allows us to conclude that all other variables being held constant, volatile acidity, chlorides, total sulfur dioxide and pH levels have a positive effect on determining quality of red wine, while free sulfur dioxide, sulphates and alcohol have a negative effect. Each coefficient can be interpreted in the following way, all else being equal, for a one unit increase in pH the odds of red wine being classified as high quality is multiplied by exp(2.2362) = 9.3577. Finally, the model has a final AUC performance of 0.861.

$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 5.1204 + 3.6526volatile.acidity + 6.9133chlorides - 0.0161free.sulfur.dioxide + 0.0151total.sulfur.dioxide + 2.2362pH - 3.1342sulph$

White Wine

Our final white wine model allows us to conclude that all other variables being held constant, volatile acidity, chlorides and total sulfur dioxide have a positive effect on determining quality of white wine, while fixed acidity, residual sugar, free sulfur dioxide, sulphates, and alcohol have a negative effect. Once again, coefficients can be interpreted in the following way, all else being equal, for a one unit increase in pH the odds of red wine being classified as high quality is multiplied by exp(-1.1718) = 0.3098. Finally, the model has a final AUC performance of 0.766.

$log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 13.8310 - 0.02815fixed.acidity + 3.5331volatile.acidity - 0.04687residual.sugar + 17.4986chlorides - 0.01344free.sulfur.dioxide + 0.$

Model Assumptions

Our models for the three research questions were developed using the following assumptions:

- Sample - The red and white sample data received from the UCI Machine Learning Repository in order to develop the given data used for the project was produced without error or bias.
- Independence - Observations within the sample dataset are independent of one another.
- Homoscedasticity - The variance of the residual is the same for any value of the predictive variable.
- Normality - Observations of the predictive and response variables are normally distributed.

Logistic regression

- Response variable is a Bernoulli random variable:
  - Our random variable was quality where $P(y_i = 1) = \pi_i$ and $P(y_i = 0) = 1 - \pi_i$

MLR and SLR Regression Assumptions

ϵ i.i.d $N(0, \sigma^2)$

- For each value of x, the errors have mean 0 for both red and white wines
- For each value of x, the errors have constant variance
- The observations are independent for both red and white wines.
- For each value of x, the errors follow a normal distribution for both red and white wines.

References

- Vinho Verde - `https://winefolly.com/deep-dive/vinho-verde-the-perfect-poolside-wine-from-portugal/#:~:text=Vinho%20Verde%20comes%20from%20a,a%20great%20ch`
- UCI Machine Learning Repository - `https://archive.ics.uci.edu/ml/datasets/wine+quality`)
- Semantic Scholar - `https://pdfs.semanticscholar.org/9ffc/11cd9bbe3f715e0f536fe3b789a60b112397.pdf`