Analysis of Vinho Verde Red and White Wines

Nikki Aaron, Colleen Callahan, Michael Pajewski, Pantea Ferdosian

Executive Summary

Paulo Cortez of the University of Minho, Guimarães, Portugal released a dataset of approximately 6500 white and red variants of Portuguese “vinho verde” wine.[1] This slightly fizzy, high acidity fruit flavored wine is most often of the white variety, with reds sometimes available in limited supply.[2] Consequently, the data set contains a much larger proportion of white wines. The data contains 11 chemical characteristics of each wine, then lists a quality rating from wine tasting experts. Our analysis of this data set focused on determining the effect of each characteristic on wine quality and wine type (red or white).

Exploration of the data concluded that some of the included chemical characteristics were correlated to one another, investigation was conducted to determine if these characteristics could be disregarded in our final models. These included fixed acidity, total acidity, volatile acidity, and pH; density, residual sugars, and alcohol content; and total and free sulfur dioxide.

When examining the wine quality class (scores 1-5 being classified as ‘bad’ wine and 6-10 being classified as ‘good’ wine) of the red and white wine data sets individually, it is clear that different predictors show an effect on each data set differently. Thus, we concluded that two separate models would produce a more accurate prediction of wine quality. Alcohol content is the one predictor that appeared positively correlated with quality. Redundant measures of acidity were removed and volatile acidity was kept as it was most useful in estimating quality. In both red and white wines, volatile acidity trends with bad quality, likely due to the unpleasant flavor and aroma that come from evaporating acids. Predictors that were not useful in predicting red wine quality, but were useful for white wine quality were pH and residual sugar. Levels of sulfur dioxide were not significant and were all removed from the white wine quality analysis.

To determine wine type, quality was the least useful predictor and was dropped along with some of the aforementioned redundant characteristics. Prediction for wine type based on density, residual sugars, volatile acidity, and other remaining chemical factors was found to be 99% accurate for wines like those in this data set.

Vinho Verde Wines

The datasets analysed in this report contain white and red variants of Portuguese “vinho verde” wine. The Vinho Verde region was the first local to export Portuguese wines to European markets. Today, Vinho Verde is one of the most known and largest wine regions in the world.[3] Vinho Verde is a region located in the hills of northern Portugal. They also mention in their website that the origin behind their name is the idea that it is harvested early and should be consumed “young”.[4] An interesting fact to note is that Vinho Verde is known for its high acidity.

Exploratory Data Analysis

Our team was provided two data sets of wine one containing 4898 white wines and one 1599 red wines. Each data set contains 11 variables of the chemical properties of the different wines, and each wine is rated between 0 (lowest quality) and 10 (best quality). Our two goals are (1) to find which objective variables that can define a pattern that predicts the quality of a given wine; and (2) see if we can predict if a wine is a red or white wine based on the objective variables. The table below shows the 11 variables included in both red and wine datasets:

Variables	**Description**
Fixed Acidity	Most acids contained in wine are fixed or nonvolatile
Volatile acidity	Acids that readily evaporate, high levels lead to an unpleasant vinegar taste
Citric acid	Citric acid can add ‘freshness’ and flavor to wines
Residual sugar	Sugar remaining after fermentation between 1 gram/liter and 45 grams/liter
Chlorides	Amount of salt in the wine
Free sulfur dioxide	Prevents microbial growth and the oxidation of wine
Total sulfur dioxide	In low concentrations is undetectable, but over 50 ppm becomes evident in the nose and taste of the wine
Density	Depends on the percent content of water, alcohol, and sugar
PH	ph is on a scale from 0 (acidic) to 14 (basic) most wines are between 3-4 on the pH scale. Related to acidity, but does not vary nearly as much.
Sulphates	antimicrobial and antioxidant
Alcohol (% volume)	Percent alcohol content of the wine
Quality	Qualitative score between 0-10

Summary of Red Wine Data:

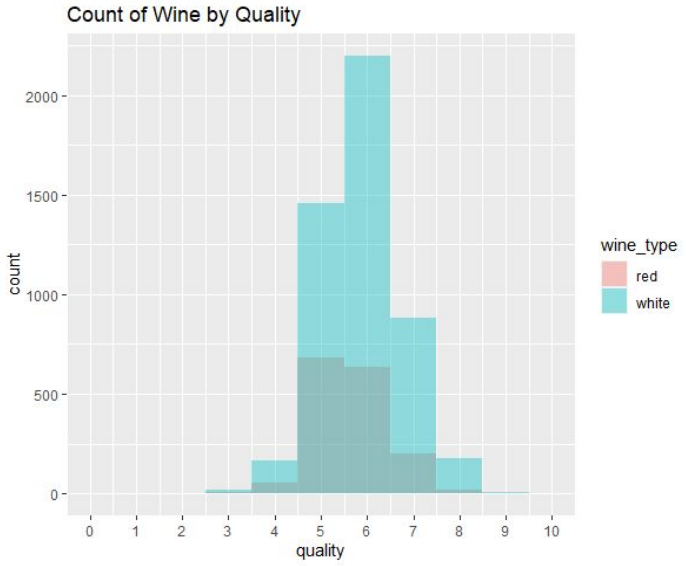
```
       X         fixed.acidity   volatile.acidity  citric.acid    residual.sugar
Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density             pH
Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901   Min.   :2.740
1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210
Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968   Median :3.310
Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967   Mean   :3.311
3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400
Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037   Max.   :4.010
   sulphates         alcohol         quality
Min.   :0.3300   Min.   : 8.40   Min.   :3.000
1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
Median :0.6200   Median :10.20   Median :6.000
Mean   :0.6581   Mean   :10.42   Mean   :5.636
3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Summary of White Wine Data:

```
       X        fixed.acidity    volatile.acidity  citric.acid     residual.sugar
Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
Median :2450   Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density             pH
Min.   :0.00900   Min.   :  2.00       Min.   :  9.0        Min.   :0.9871   Min.   :2.720
1st Qu.:0.03600   1st Qu.: 23.00       1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090
Median :0.04300   Median : 34.00       Median :134.0        Median :0.9937   Median :3.180
Mean   :0.04577   Mean   : 35.31       Mean   :138.4        Mean   :0.9940   Mean   :3.188
3rd Qu.:0.05000   3rd Qu.: 46.00       3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280
Max.   :0.34600   Max.   :289.00       Max.   :440.0        Max.   :1.0390   Max.   :3.820
   sulphates         alcohol         quality
Min.   :0.2200   Min.   : 8.00   Min.   :3.000
1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
Median :0.4700   Median :10.40   Median :6.000
Mean   :0.4898   Mean   :10.51   Mean   :5.878
3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

As mentioned above in the tables, we see that the count of observations for white wine far exceeds the number of observations for red wine. The data set includes 1599 observations for red wine and but only 4898 observations for white wine.

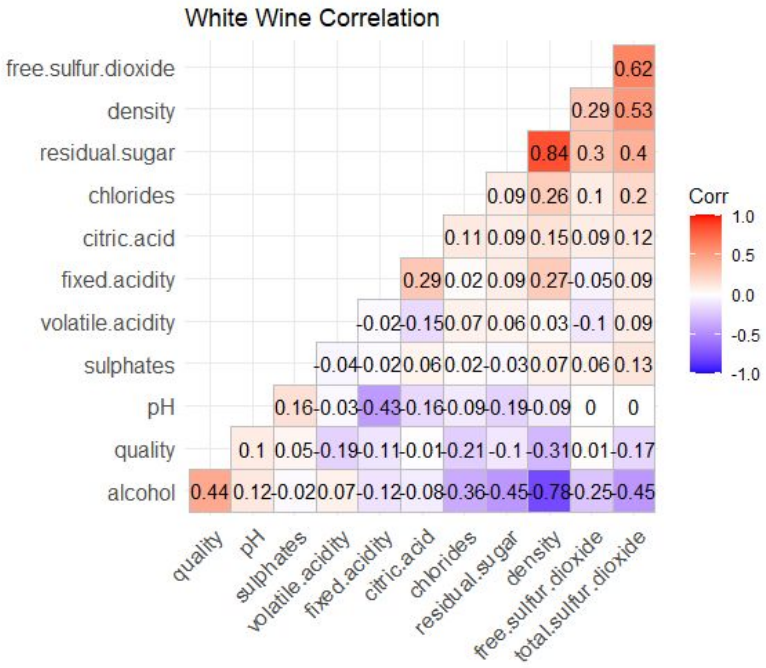
The table also includes distribution of the properties of red and white wines. There are a few extreme values in the dataset. The variables residual.sugar, total.sulfur.dioxide, sulphates, and chlorides have maximum values which are significantly higher than the mean and median values of those predictors. We may consider removing these outliers in the next steps for a more adequate model.



The Count of Wine by Quality chart shows the quality of red and white wine is approximately normally distributed. In the data set wines are graded with a quality between 0 being very bad and 10 being excellent quality. The majority of the both red and white wines fall between a quality rating of 5 and 7. There are no wines in the data set below a quality of 3 and or above a quality of 9. The lack of wines lower than a quality of 3 and above the quality of 8 may make it difficult or unreliable for our model to predict very low or very high quality wines.

Our group tackled the issue of correlation of the continuous variables in the data sets for white and red wines separately. Our team used correlation plots to identify highly correlated variables to consider for removal from the future analysis. The correlation plot for white wine shows the following variables are highly correlated.

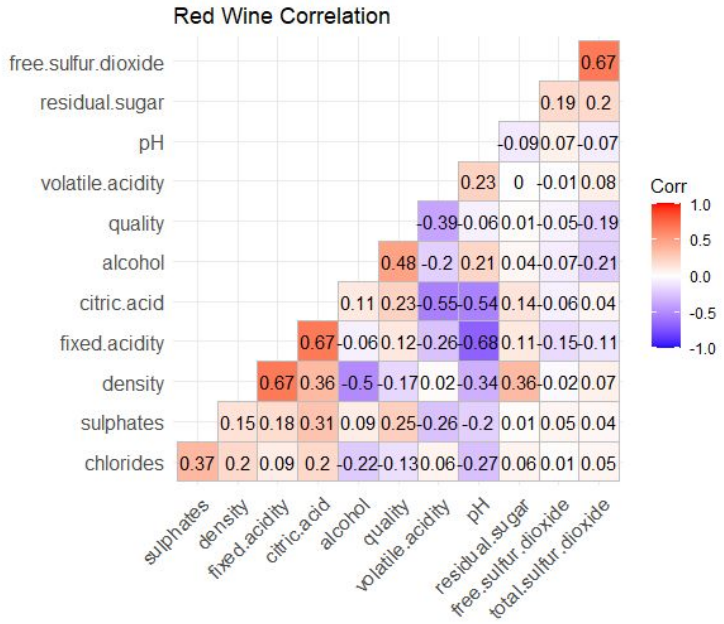
1. Quality and alcohol
2. Residual sugar and density
3. Alcohol and density



4. Free sulfur dioxide and total sulfur dioxide

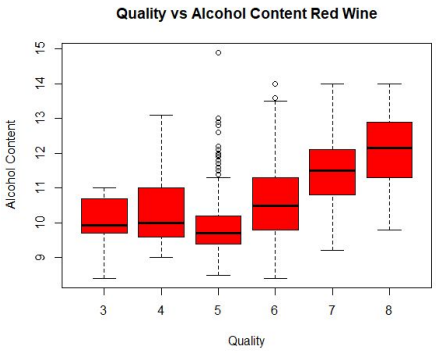
The correlation plot for red shows following variables are highly correlated.

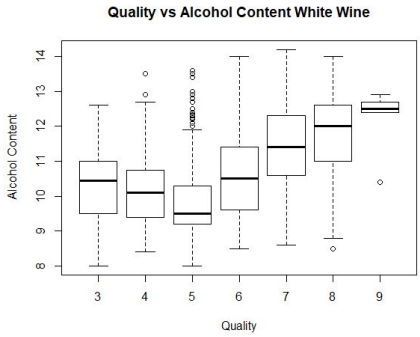
1. Quality and alcohol
2. Density and fixed acidity
3. Alcohol and density
4. Fixed acidity and citric acid
5. Volatile acidity and citric acid
6. Fixed acidity and ph
7. Free sulfur dioxide and total sulfur dioxide.



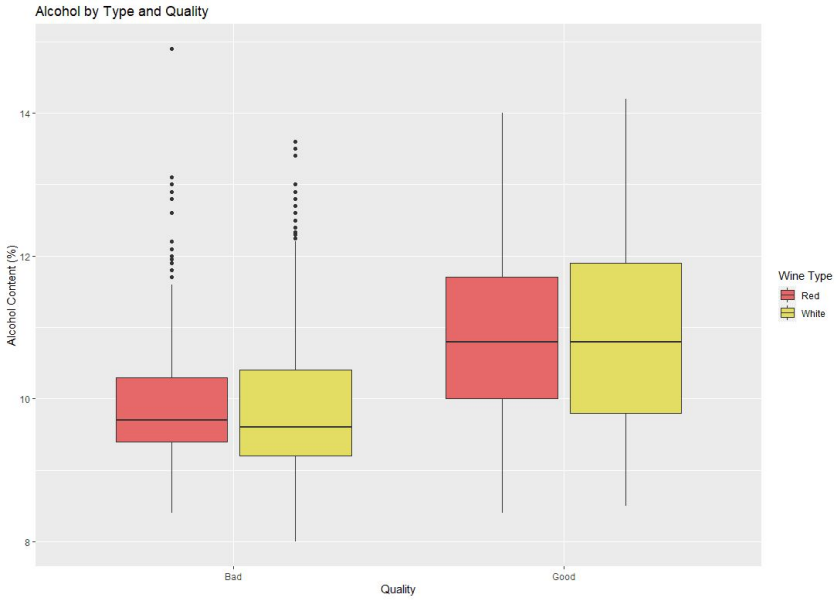
For both red and white wine it is understandable that there is correlation between many of the variables. For example, red and white wine both show negative correlation for alcohol and density. This is not surprising because wine is approximately 85% water and up to 14% alcohol making up 99% of the volume of wine. [5] Alcohol is less dense than water so it is understandable that the more alcohol the lower the density of the wine. For Red, fixed acidity with citric acid, fixed acidity with ph, and free sulfur dioxide with total sulfur dioxide are all correlated. Each of these pairs of variables are related by definition to each other so it is not surprising there is a correlation between each of them.

In both the red and white datasets alcohol content is somewhat positively correlated with quality and all of the variables are mostly uncorrelated with quality. It can be observed in the quality vs alcohol content histograms for both Red and White wines there is a somewhat positive trend between the median alcohol content for each quality of wine. There is a dip in alcohol content for both red and white wines with a quality of 5 affects the strength of the positive relationship. When looking at alcohol content's relationship to quality we can not assume there is a direct correlation between quality and alcohol content we need to consider the limits of the data set and the influence of the other factors on quality.

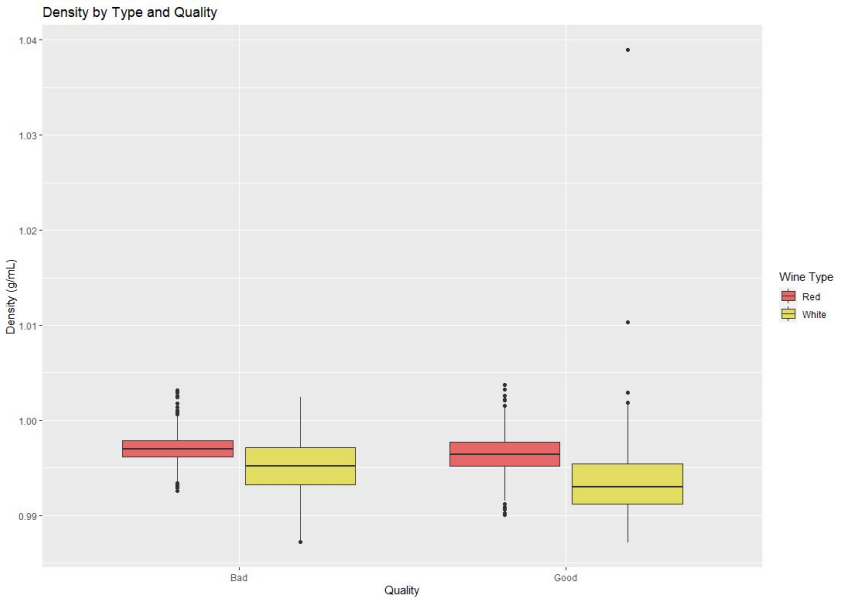




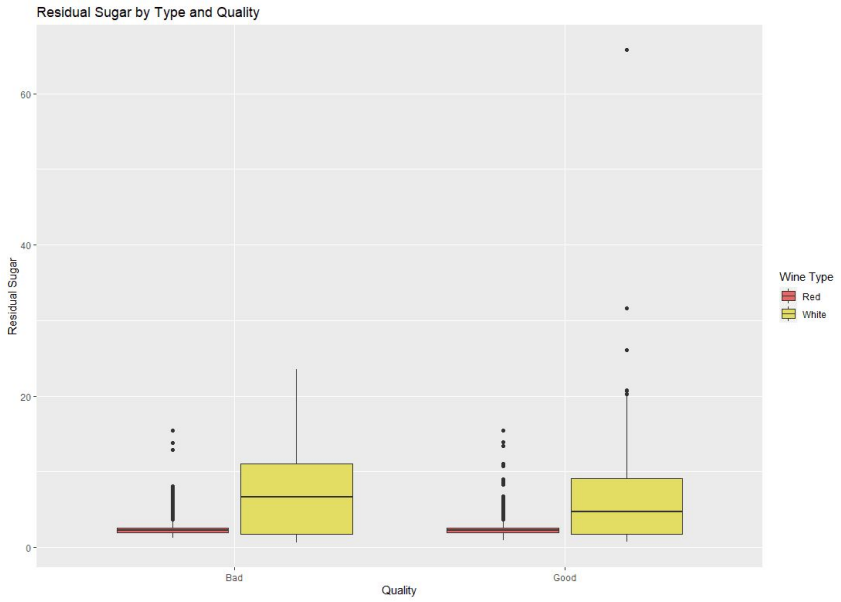
We can see in the “Quality by Type” boxplot that the mean and variance of quality ratings are virtually the same between wine types. It appears that quality and wine type will not be useful for predicting one another. All box-plots to follow will show predictors grouped by quality and colored by wine type.



Density is determined by the water (density = 0.863 g/mL) , alcohol (density = 0.789 g/mL), and sugar (density = 1.587 g/mL) content of a wine. Sugar is the densest of these, but also the smallest component of wine. Consequently, density has a strong inverse correlation with alcohol content (r = -0.67), and is significantly higher in red wines with lower mean alcohol content.

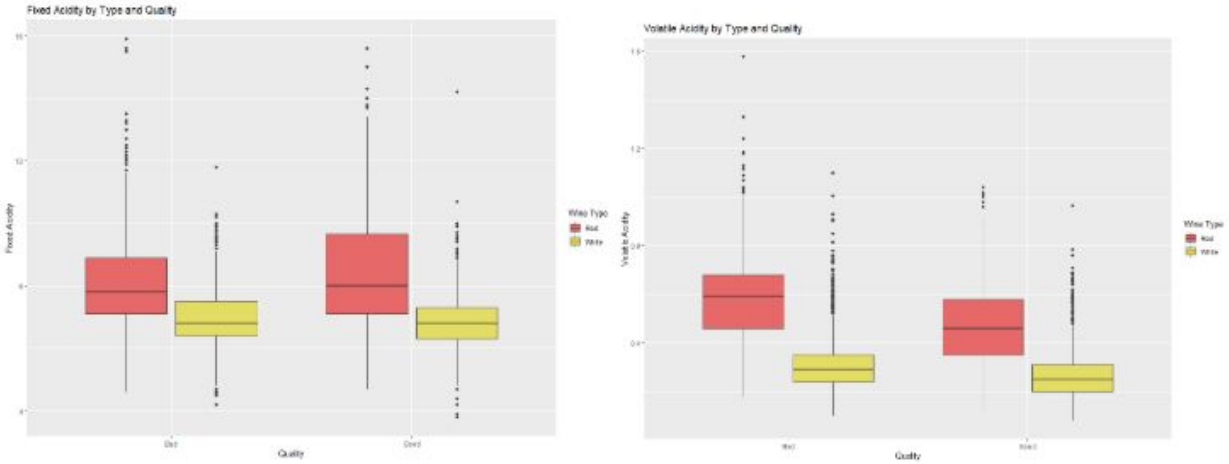


Though its percent content in wines is relatively low, residual sugar does appear to be an important factor in predicting wine type. Assuming that residual sugar is the main factor in wine sweetness, red wines tend to have a low sweetness level with very little variance, whereas white wines vary widely from the level of reds up to very sweet flavors. It appears that “bad” white wines tend to be too sweet.

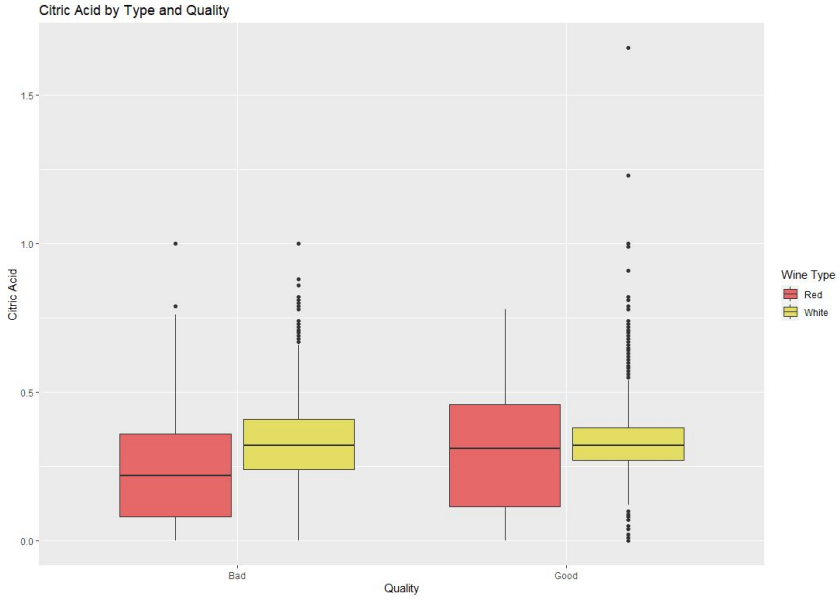


Wine acidity is another important factor in determining wine type and flavor quality. Red wines tend to contain much higher fixed and volatile acidity than whites. Wines in their respective flavor categories with relatively high volatile acidity and low fixed acidity are rated as “bad” much

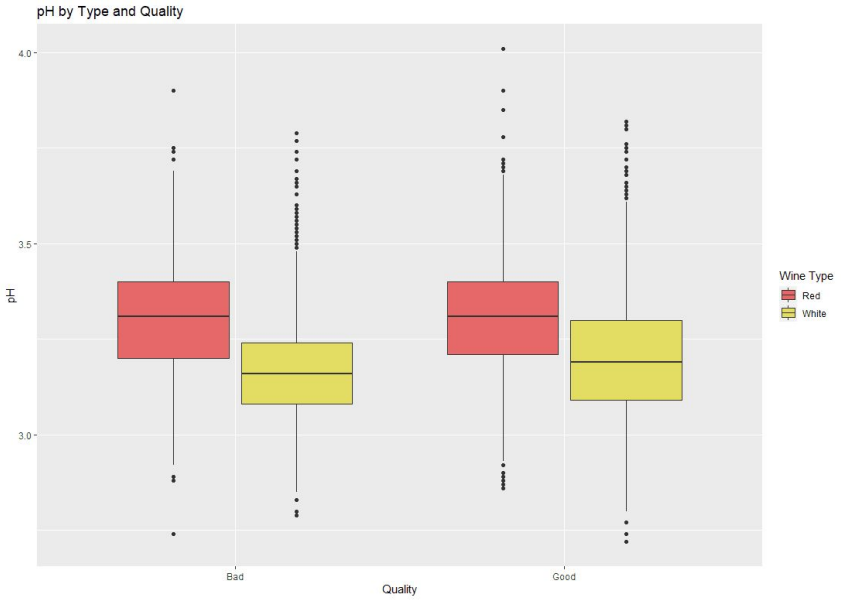
more often. Acids evaporate from them much more readily and it leaves an unpleasant sensation in the nose, stronger odor, and unbalanced flavor.



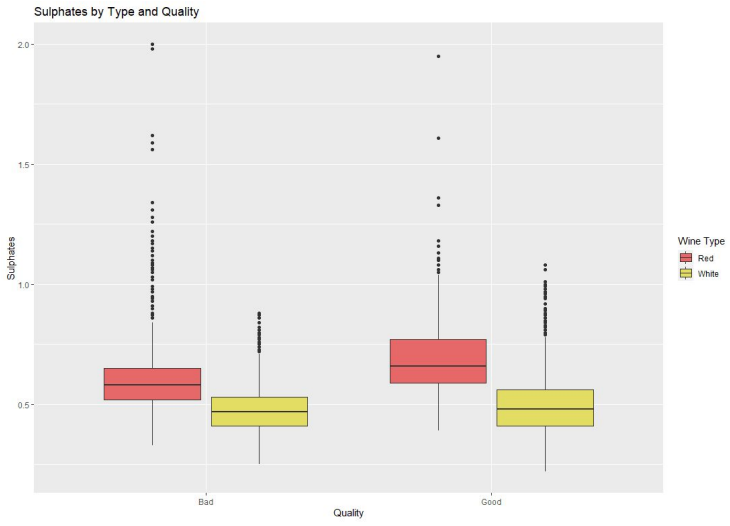
Citric acid is one of the acids contained in wines and lending to its flavor. The level of this citrus fruit acid is higher on average for White wines both “good” and “bad”. Red wines have a much larger variation in citric acid content, with lower average levels found in “bad” red wines. I think these reds would be described as “dull” rather than “bright”.



The pH variable measures acidity and is moderately correlated with citric acid (r = -0.33), volatile acidity (r = 0.26), and fixed acidity (r = -0.25). The pH of wines within each type does not vary too widely, and we see little difference between the quality levels.

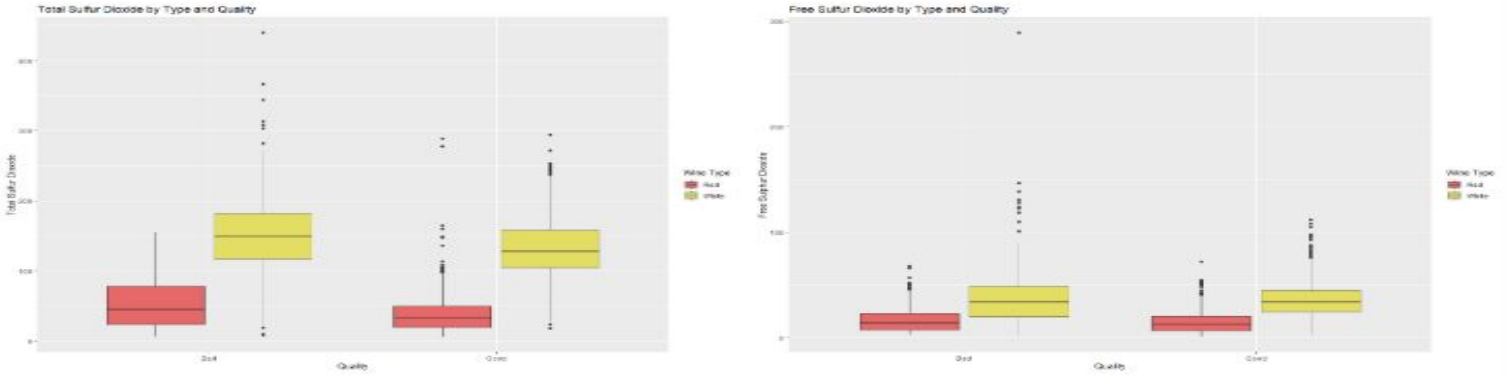


Sulphates are molecules containing sulfur -- mainly sulfite ions and the aforementioned sulfur dioxide in wines. They have antioxidant properties and also affect wine flavor. It is generally considered that high sulfates make a “dull” wine, though this is not reflected in our data set. We do see that sulphates are correlated with chloride (r = 0.39) and density (r = 0.26).

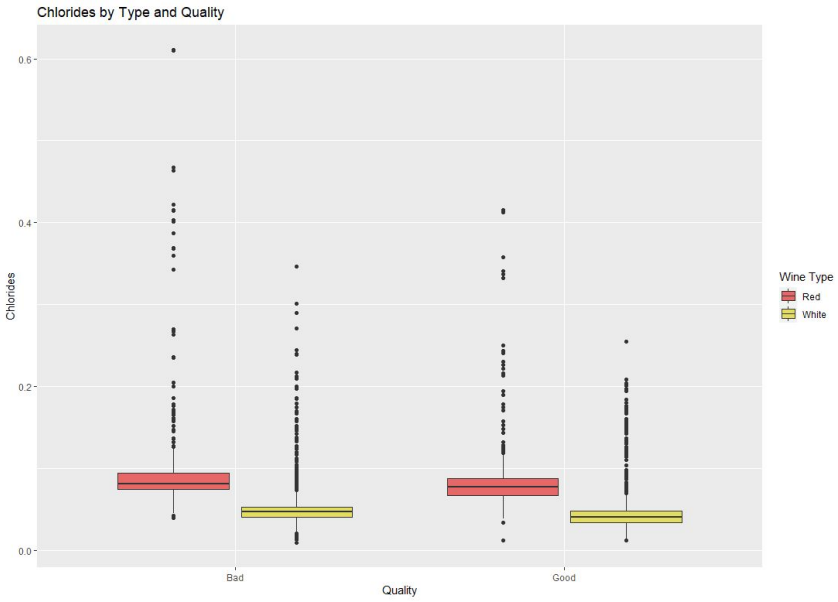


Sulfur dioxide is added during the winemaking process to prevent oxidation. Free and total sulfur dioxide levels are highly correlated (r = 0.72), and likely exhibit multicollinearity. Some quick research reveals that tannins in red wines act as a natural preservative, so much less

sulfur dioxide needs to be added to them as compared with white wines, which are more likely to oxidize. The substance is flavorless in concentrations less than 50 ppm, and undesirable in larger amounts. These facts explain well why the highest levels of both free and total sulphur dioxide are found in our “bad” white wine category.



Chlorides are mineral salts contained in wine, mostly sodium chloride, that can give the wine a salty flavor. Our data shows that salt level of wine is useful in predicting both wine type and quality. With higher salt levels in red wines and poor quality wines.



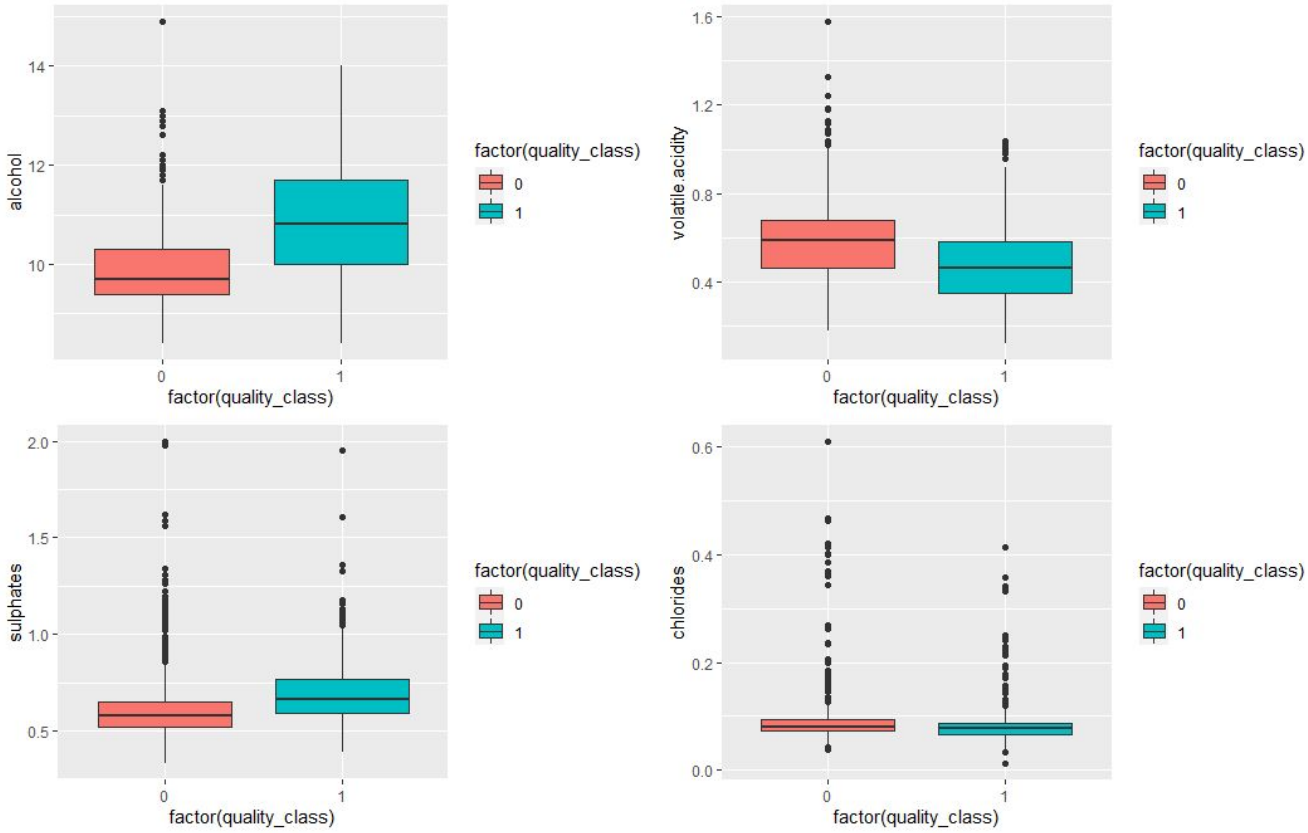
Detailed Analysis - Predicting Wine Quality

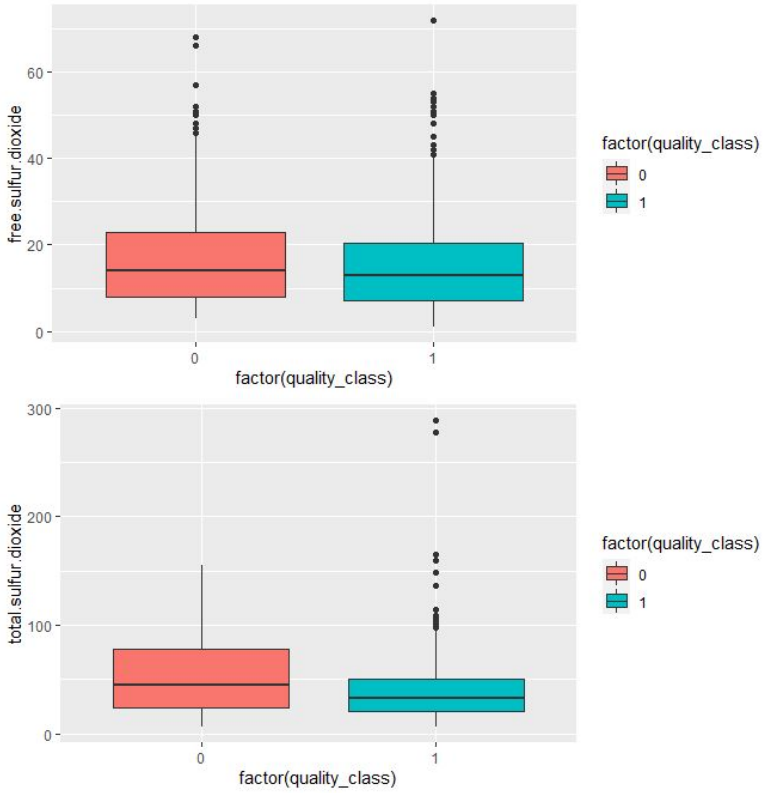
We performed two binary logistic regressions on the two wine datasets, one for white and one for red. We converted the *quality* of wine into a binomial categorical response variable, denoting 0 for 'Bad' wine (quality score less than 6) and 1 for 'Good' wine (quality score 6 or above. First, we worked with the red wine dataset to assess a binomial logistic regression.

In regressing quality class against all other predictor variables, the result told us that only seven predictors are significant in the presence of other predictors. The confusion matrix accuracy was only 77%. Our VIF plot shows two variables with possible collinearity.

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide
7.896541	1.814588	3.302695	1.655045	1.376778	1.986458
total.sulfur.dioxide	density	pH	sulphates	alcohol	
2.268032	6.339640	3.237454	1.395900	2.985770	

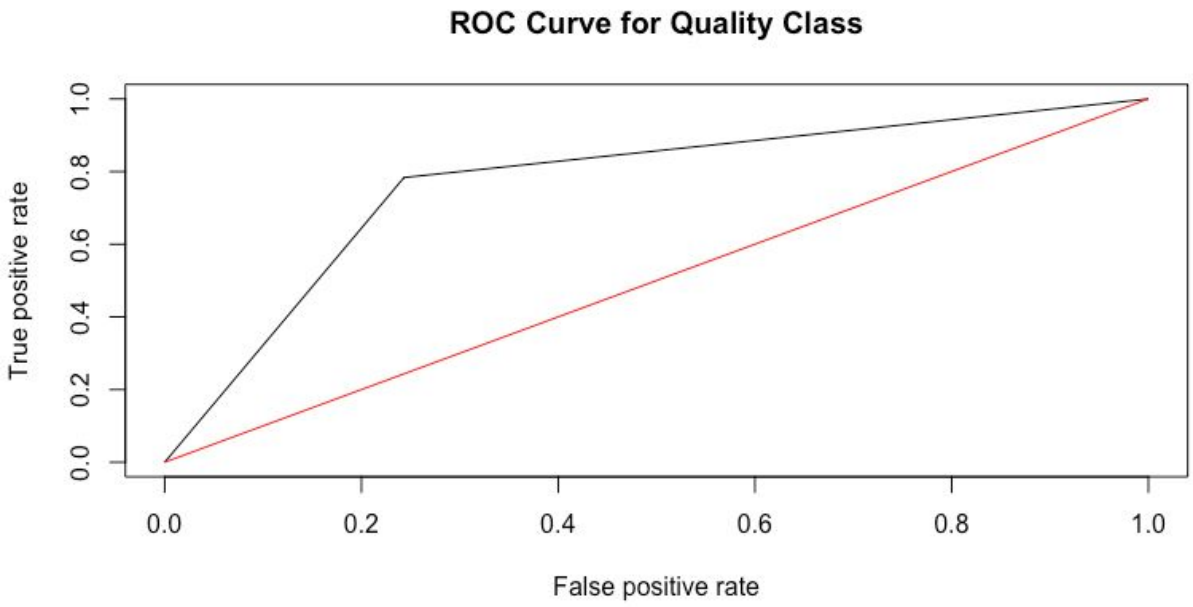
We produced 11 box plots, one for each predictor, to assess the relationship between quality class (good and bad wine) and the predictor.





We found that alcohol, free sulfur dioxide, sulphates, volatile acidity, total sulfur dioxide and chlorides showed significant differences between the two classes of quality for red wines. Then we fit a new model using only significant predictors. We evaluated the model using validation, and found that the confusion matrix shows 77% accuracy, very close to the full model. A partial G-squared test also confirms that the reduced model is acceptable.

We then produced a ROC curve for our model. The plot showed us that the model is better than random guessing at predicting quality class. The computed AUC value was 0.7701912 which confirms this conclusion.



The final model we selected for quality class of red wine is:

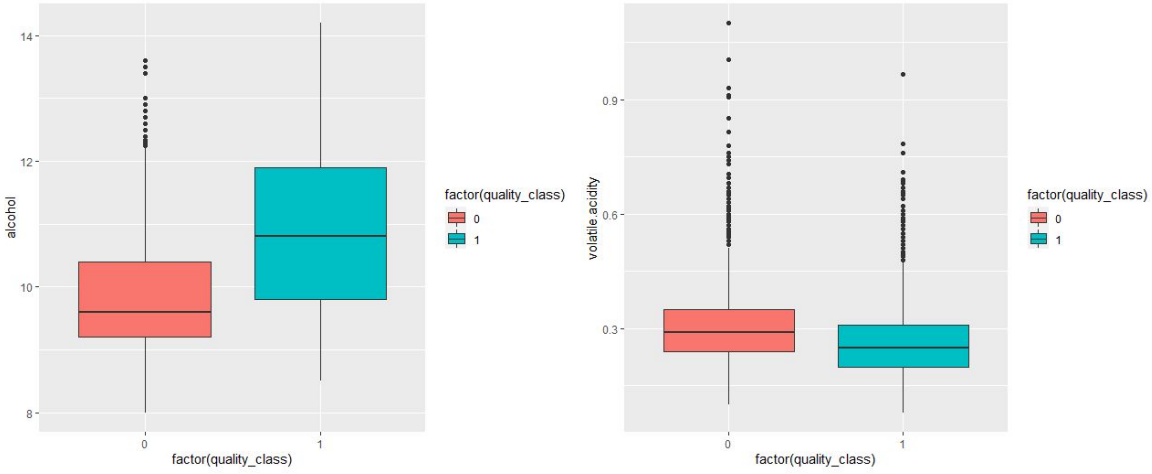
$$log(\pi/1-\pi) = -7.38 + 0.82 * alcohol - 3.13 * volatile\ acidity - 4.28 * chlorides$$
$$- 0.02 * total\ sulfur\ dioxide + 0.02 * free\ sulphur\ dioxide + 2.42 * sulphates$$

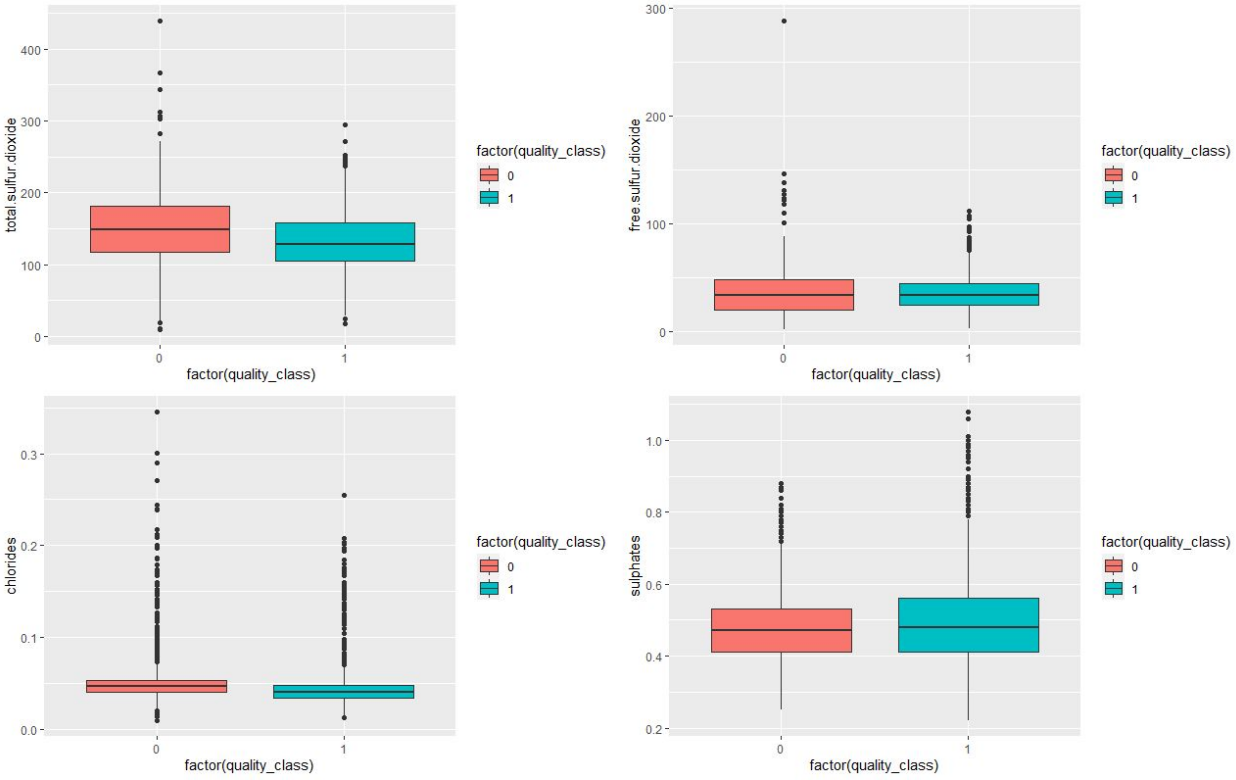
The summary for this model is as follows:

Accuracy: 77%	Precision: 79%	False Positive Rate: 30%
False Negative Rate: 28%	Recall: 78%	

```
Call:
glm(formula = quality_class ~ alcohol + volatile.acidity + chlorides +
    total.sulfur.dioxide + free.sulfur.dioxide + sulphates, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0259  -0.8914   0.3355   0.8597   2.2997

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -7.388636   0.955059  -7.736 1.02e-14 ***
alcohol               0.821395   0.083462   9.842  < 2e-16 ***
volatile.acidity     -3.129052   0.446953  -7.001 2.54e-12 ***
chlorides            -4.282749   1.669051  -2.566   0.0103 *
total.sulfur.dioxide -0.017315   0.003183  -5.440 5.31e-08 ***
free.sulfur.dioxide   0.020685   0.009630   2.148   0.0317 *
sulphates             2.416017   0.480138   5.032 4.86e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1543.9  on 1118  degrees of freedom
Residual deviance: 1185.5  on 1112  degrees of freedom
AIC: 1199.5

Number of Fisher Scoring iterations: 4
```

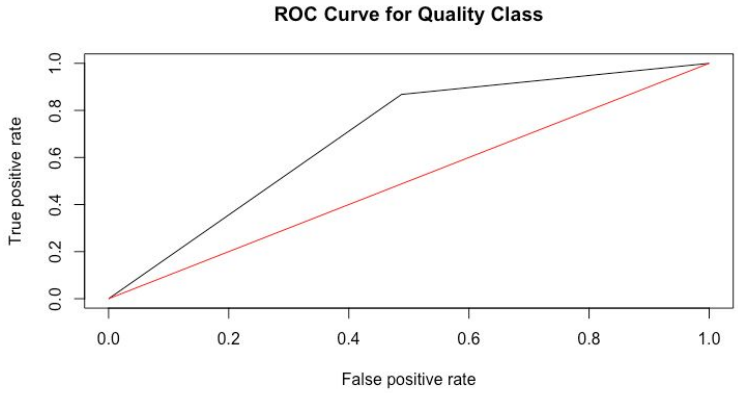
We repeated the quality class analysis for the white wine data set. The white wine analysis followed a similar pattern to the red wine analysis. When quality class was regressed on all other variables, the model found had 9 significant predictors and a confusion matrix accuracy of 75%. We reproduced the box plots for all the predictors by quality class and found that volatile acidity, residual sugar, free sulfur dioxide, density, sulphates, pH and alcohol showed significant differences between the two classes of quality.





We then refit a regression model using only significant predictors. The new confusion matrix accuracy is virtually the same as the full model, and a high partial delta G-squared test says we can keep the reduced model.

The ROC curve for this model shows that the model is better at predicting quality class than random guessing. The AUC for this model is 0.6900453 which confirms this, as well.



The final model we selected for quality class of white wine is:

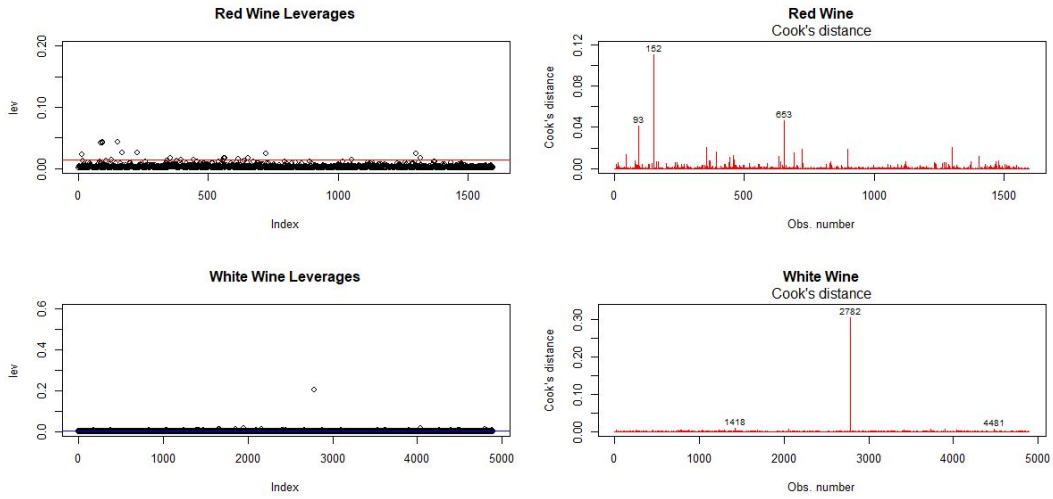
$$log(\pi/1-\pi) = 201.6 - 6.59 * volatile\ acidity + 0.14 * residual\ sugar + 0.02 * free\ sulphur\ dioxide - 213.6 * density + 0.97 * pH + 0.34 * sulphates + 0.80 * alcohol$$

The summary for this model is as follows:

Accuracy: 75%	Precision: 89%	False Positive Rate: 47%
False Negative Rate: 30%	Recall: 77%	

```
Call:
glm(formula = quality_class ~ volatile.acidity + residual.sugar +
    free.sulfur.dioxide + density + pH + sulphates + alcohol,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1210  -0.9005   0.4462   0.7995   2.7323

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        2.016e+02  5.334e+01   3.780 0.000157 ***
volatile.acidity  -6.590e+00  4.693e-01 -14.041  < 2e-16 ***
residual.sugar     1.406e-01  2.132e-02   6.595 4.27e-11 ***
free.sulfur.dioxide  1.032e-02  2.734e-03   3.773 0.000161 ***
density           -2.136e+02  5.341e+01  -3.999 6.36e-05 ***
pH                 9.671e-01  3.022e-01   3.200 0.001375 **
sulphates          1.337e+00  4.186e-01   3.194 0.001402 **
alcohol            8.016e-01  8.106e-02   9.888  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4374.0  on 3427  degrees of freedom
Residual deviance: 3451.9  on 3420  degrees of freedom
AIC: 3467.9

Number of Fisher Scoring iterations: 5
```

We use the Cook's distance to detect influential observations in the red and white data sets. Looking at the Cooks plot for both Red and White wines all the data points are well inside of the Cook's distance and do not exhibit high leverage so we do not need to investigate removing any data points from the data set.



Our conclusions based on the 75-77% accuracy for red wine and white wine quality estimation is that the models based on the predictors in this data set will be useful, but may not be relied on when extremely high accuracy is required.

Detailed Analysis - Predicting Wine Type

To find out which predictors are useful for determining wine type, we created a binary logistic regression model using a 70%-30% train-test split of the data. A full model was fitted, but we could see from the p-values for each predictor that a few may be able to be dropped from the model for insignificance. We also noticed that a VIF plot showed evidence of multicollinearity between a few predictors, especially density (see output below).

```
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
            4.942690            2.025418            1.599768            7.567620            1.611537
  free.sulfur.dioxide total.sulfur.dioxide             density                  pH           sulphates
            2.187147            3.018782           16.149845            2.528557            1.556976
             alcohol             quality
            4.912481            1.403671
```

We could see from box-plots that quality likely had no predictive value, so it was dropped first. Next, since fixed, total, and volatile acidity are related, we tried dropping fixed acidity. Then through trial and error, we found that citric acid and pH could also be dropped. After each round of reduction, we used calculation of delta G-squared to get the p-value from a likelihood ratio test to verify that the dropped predictors were insignificant. High p-values each time indicated that we could not reject the null hypothesis, and that the reduced model was acceptable. Delta

G-squared for the reduced model gave a p-value of 0, indicating that the model is useful. A summary of the reduced model is shown below.

```
> anova(reduced, result, test="LRT")
Analysis of Deviance Table

Model 1: wine_type ~ density + residual.sugar + total.sulfur.dioxide +
    volatile.acidity + chlorides + sulphates + alcohol + free.sulfur.dioxide
Model 2: wine_type ~ density + residual.sugar + total.sulfur.dioxide +
    volatile.acidity + pH + chlorides + sulphates + alcohol +
    free.sulfur.dioxide + citric.acid + fixed.acidity + quality
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4538     332.57
2      4534     324.87  4   7.7021   0.1031
```

```
Call:
glm(formula = wine_type ~ density + residual.sugar + total.sulfur.dioxide +
    volatile.acidity + chlorides + sulphates + alcohol + free.sulfur.dioxide,
    family = binomial, data = data.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2863  -0.0580  -0.0159  -0.0006   5.6931

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.456e+03  1.351e+02 -10.779  < 2e-16 ***
density               1.447e+03  1.342e+02  10.786  < 2e-16 ***
residual.sugar       -8.933e-01  1.024e-01  -8.727  < 2e-16 ***
total.sulfur.dioxide -5.666e-02  5.721e-03  -9.904  < 2e-16 ***
volatile.acidity      7.457e+00  1.002e+00   7.443 9.87e-14 ***
chlorides             2.202e+01  4.115e+00   5.350 8.79e-08 ***
sulphates             3.546e+00  1.287e+00   2.755  0.00586 **
alcohol               1.488e+00  2.535e-01   5.869 4.37e-09 ***
free.sulfur.dioxide   7.097e-02  1.462e-02   4.855 1.20e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5074.50  on 4546  degrees of freedom
Residual deviance:  332.57  on 4538  degrees of freedom
AIC: 350.57

Number of Fisher Scoring iterations: 9
```

This reduction greatly improved our problematic VIF values by removing collinear predictors.

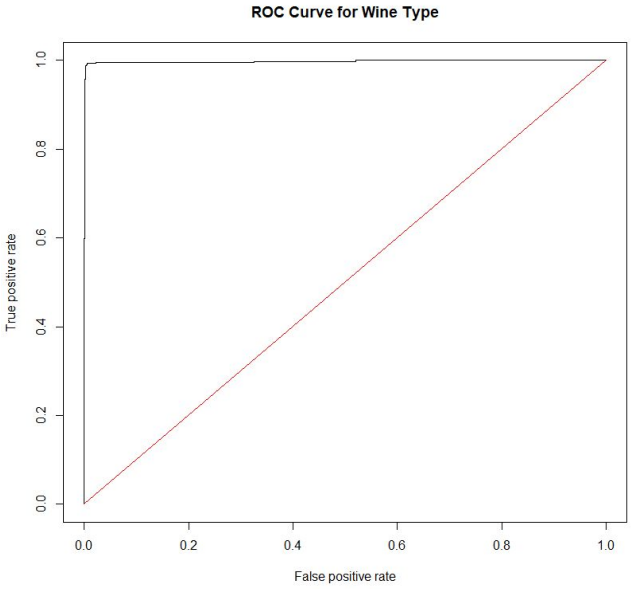
```
> vif(data.train[,c(2,4:8,10:11)])
    volatile.acidity       residual.sugar            chlorides  free.sulfur.dioxide
            1.528192             3.202998             1.526281             2.151217
total.sulfur.dioxide              density            sulphates              alcohol
            2.879756             5.108210             1.483397             2.645104
>
```

Using the resulting logistic model of the following equation,

$$log(\pi/1-\pi) = -1456 + 1447 * density - 0.89 * residual\ sugar - 0.06 * total\ sulfur\ dioxide$$

$$+\ 7.46 * volatile\ acidity\ 22.02 *\ chlorides\ +\ 3.55 * sulphates + 1.49 * alcohol + 0.071 * free\ sulfur\ dioxide$$

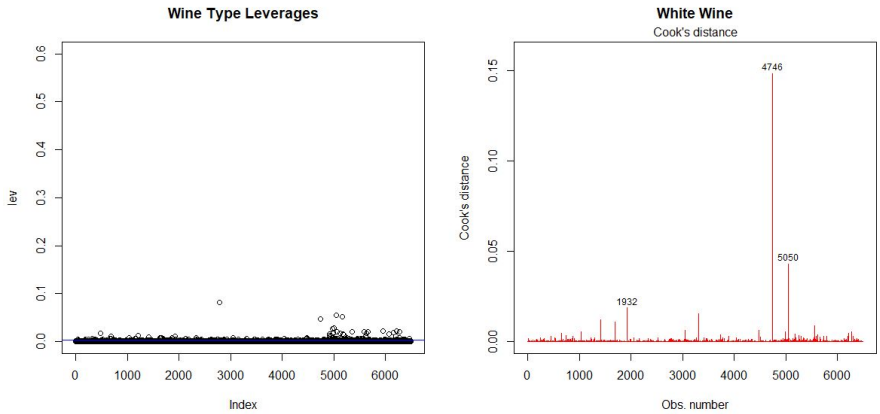
we could compute the accuracy of our model with the remaining 30% test data. The results show an almost perfect ROC curve and a confusion with only 1-2 more false positives and false negatives than was obtained using the full model. The model has great performance with an accuracy and precision of 99%, and very low false positive and negative rates.




```
> table(data.test$wine_type, preds>0.5)

     FALSE TRUE
  0  1466    4
  1     6  474
>
> table(data.test$wine_type, preds>0.7)

     FALSE TRUE
  0  1466    4
  1    15  465
>
```

Accuracy: 99%	Precision: 99%	False Positive Rate: 0.3%
False Negative Rate: 3.2%	Recall: 99%	

We again use the Cook's distance to detect influential observations in the model. For the wine type prediction model all the data points are well inside of the Cook's distance and do not exhibit high leverage so we do not need to investigate removing any data points from the data set.



Using our models, we created the following table showing the G-squared value and probability prediction for each model for two randomly selected data points.

	Red Wine Quality	**White Wine Quality**	**Wine Type**
Delta G-squared p-value	0	0	0
Sample 1 Prediction (Actual: Good Quality Red)	95% Probability Good	--	99.6% Probability Red
Sample 2 Prediction (Actual: Good Quality White)	--	95% Probability Good	1.5% Probability Red

All three models are useful for prediction according to the 0 p-values computed from delta G-squared. We can see from two sample points that odds prediction is correctly identifying actual values with high probabilities. Wine type has a bit better performance than predictors for wine quality.

References

[Dataset] 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.

[1] (2020 August 01). *Wine Quality Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[2] (2020 August 01). *101 Guide To Vinho Verde Wine from Portugal*. Wine Folly. https://winefolly.com/deep-dive/vinho-verde-the-perfect-poolside-wine-from-portugal/

[3] VINHO VERDE. *About Vinho Verde*. Retrieved from Vinho Verde: https://www.vinhoverde.pt/en/about-vinho-verde

[4] VINHO VERDE. *History of Vinho Verde*. Retrieved from Vinho Verde: https://www.vinhoverde.pt/en/history-of-vinho-verde

[5] Vivino. *Wine Chemistry 101: What is Wine Made of?* https://www.vivino.com/wine-news/wine-chemistry-101-what-is-wine-made-of