Wine Data Analysis

Paul-Jason Mello

12/6/2021

Introduction

I have selected this data set because I am interested in learning about which properties have a significant impact on the quality of a wine. I have begun researching the world of wines independently and this variety specifically comes from the north of Portugal. As a descendant of two Portuguese immigrants, who just began entering the world of wines, I thought it would be a great data set to work with. I hope that once my analysis is finished I will have deciphered what makes a good quality wine. Additionally, I highly recommend those who are interested to read the paper developed by Paulo Cortez http://www3.dsi.uminho.pt/pcortez/wine5.pdf He details out many interesting methods in developing a solid algorithm for wine classification. It should be noted that despite the intricate methods presented, his best accuracy for the classification of wine qualities was 89% when using an SVM (support-vector machine) to classify red wines specifically.

Data

Github Repository Link for this Project: https://github.com/pauljmello/Wine-Data-Analysis

Data source: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Data collection: This data set is a collection of wine data sampled from the northern end of Portugal by Paulo Cortez in 2009. Specifically, it comes from a variety of grapes called "Vinho Verde". Cortez collected these samples as part of an observational study "to model wine quality based on physicochemical tests."

Units of observations:

Each row represents a purchasable wine. The wine name, price, and other identifying aspects have been removed for privacy reasons.

The column names are denoted below with a brief summary describing the unit of observation.

1. fixed acidity - Ratio (double), 4 Types of standard acids found in wine.
2. volatile acidity - Ratio (double), Measure of a wines gaseous acid content.
3. citric acid - Ratio (double), Measure of wines citric acid content.
4. residual sugar - Ratio (double), Measure of natural grape sugar leftovers after fermentation.
5. chlorides - Interval (double), Measures the saltiness of a wine. Consistent with the salt in water used during the wine making process.
6. free sulfur dioxide - Ratio (integer), Sulfur which is added to the wine to regulate bacterial stability.
7. total sulfur dioxide - Ratio (integer), The total amount of sulfur that is found in the wine.
8. density - Ratio (double), A measure of how dense the wine is.

9. pH - Interval (double), A measure of how acidic or basic the wine is. ph < 7 = acidic, ph > 7 = basic, ph $= 7$ = neutral.
10. sulfates - Ratio (double), A salt that forms when citric acid reacts with other chemicals.
11. alcohol - Ratio (double), The alcohol content of the wine.
12. quality - Ordinal (integer), The quality of the wine as determined and evaluated by 3 wine experts.

Additional derived columns will be added to each data set consisting of the following descriptions:

13. total.free - Ratio (double), total sulfur dioxide - free sulfur dioxide. This is in essence sulfur which is considered "dead" because it is no longer fighting bacteria growth.

XX. color - Nominal (string), The color of the observed wine. *Only found in the aggregate wine data set.

XXI. ratings - Binary (integer), Wines which were rated 7-10 will be considered good (1), while wines from 0-6 will be considered bad (0).

Variables: I plan to study the wine data set by using all of the variables given except for density. I believe that the density of wine is so similar it will have no effect on any aspect or analysis of the data.

Important Notes

Wines are made in a process composed of many parts over potentially very long periods of time. It is a complex task that may not be easily modeled by physiochemical traits as there are many external variables which contribute to the quality of a wine. Wine attributes have diminishing returns in each column. For example drinking pure sulfates would be toxic, not tasty. Additionally, quality is a subjective measure, even when rated by experts.

Questions

1. Is there a strong correlation in determining the quality of a wine when looking at the ratios between total sulfur dioxide and free sulfur dioxide?
2. Can we develop a regression model to accurately predict the ratings of a wine?
3. Which aspects of acidity play an important role when determining the variation in the data?
4. Can we model wine color by its physiochemical properties?

Data Preprocessing

Loading and formatting data.

```
redWine <- read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequal
whiteWine <- read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/wineq
```

The data set claims to have no missing observations So we will check to ensure all observations are present.

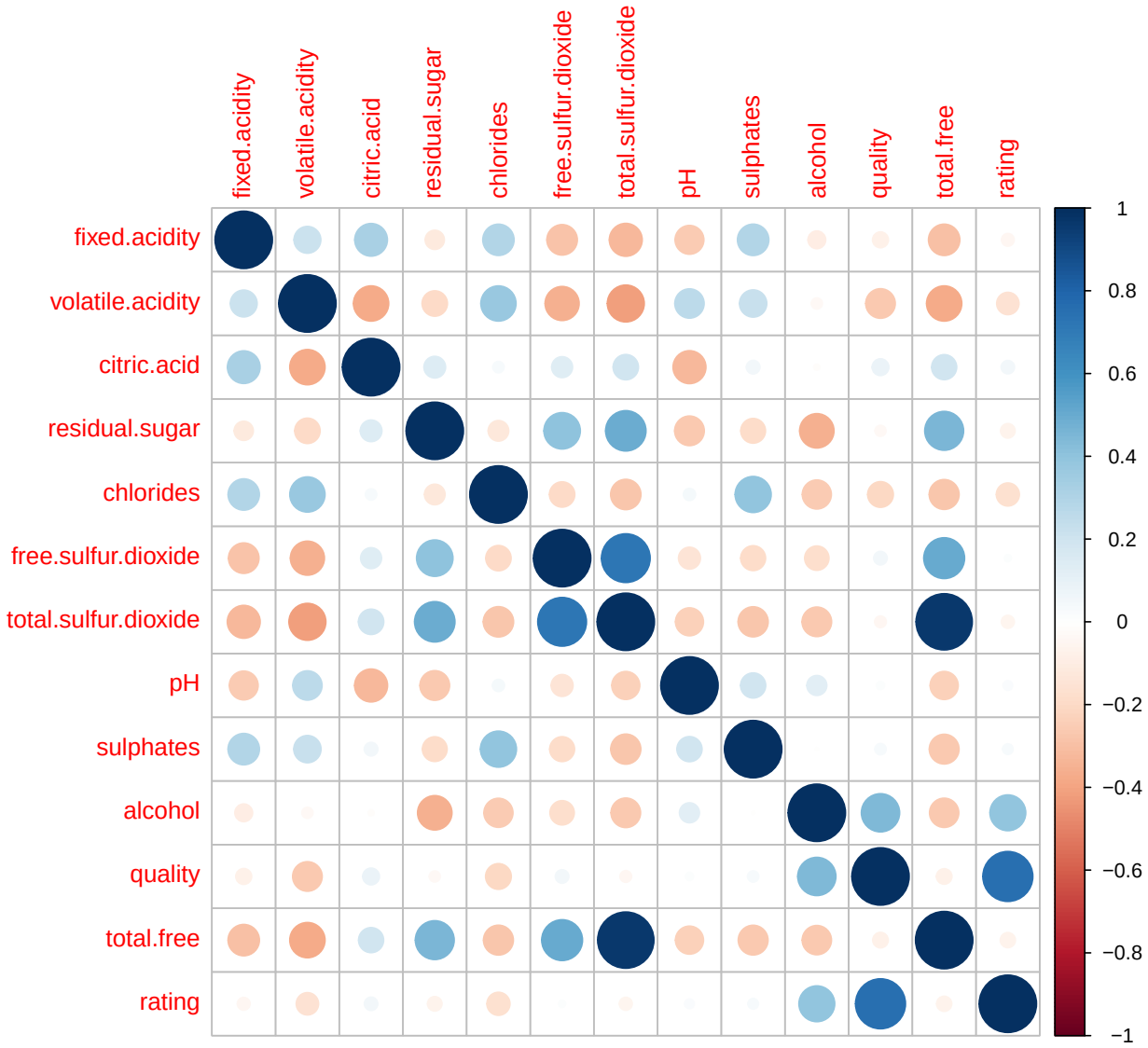
```
## [1] "Missing Values of Red Wine: 0"
```

```
## [1] "Missing Values of White Wine: 0"
```

Next we want to derive the columns and create the aggregated data set we need. One of these columns being created represents the quantity of the current "dead" sulfates that are no longer actively fighting bacterial growth, but remain lingering in the background. We will also create another column to describe the color of the wine in the aggregated data set.

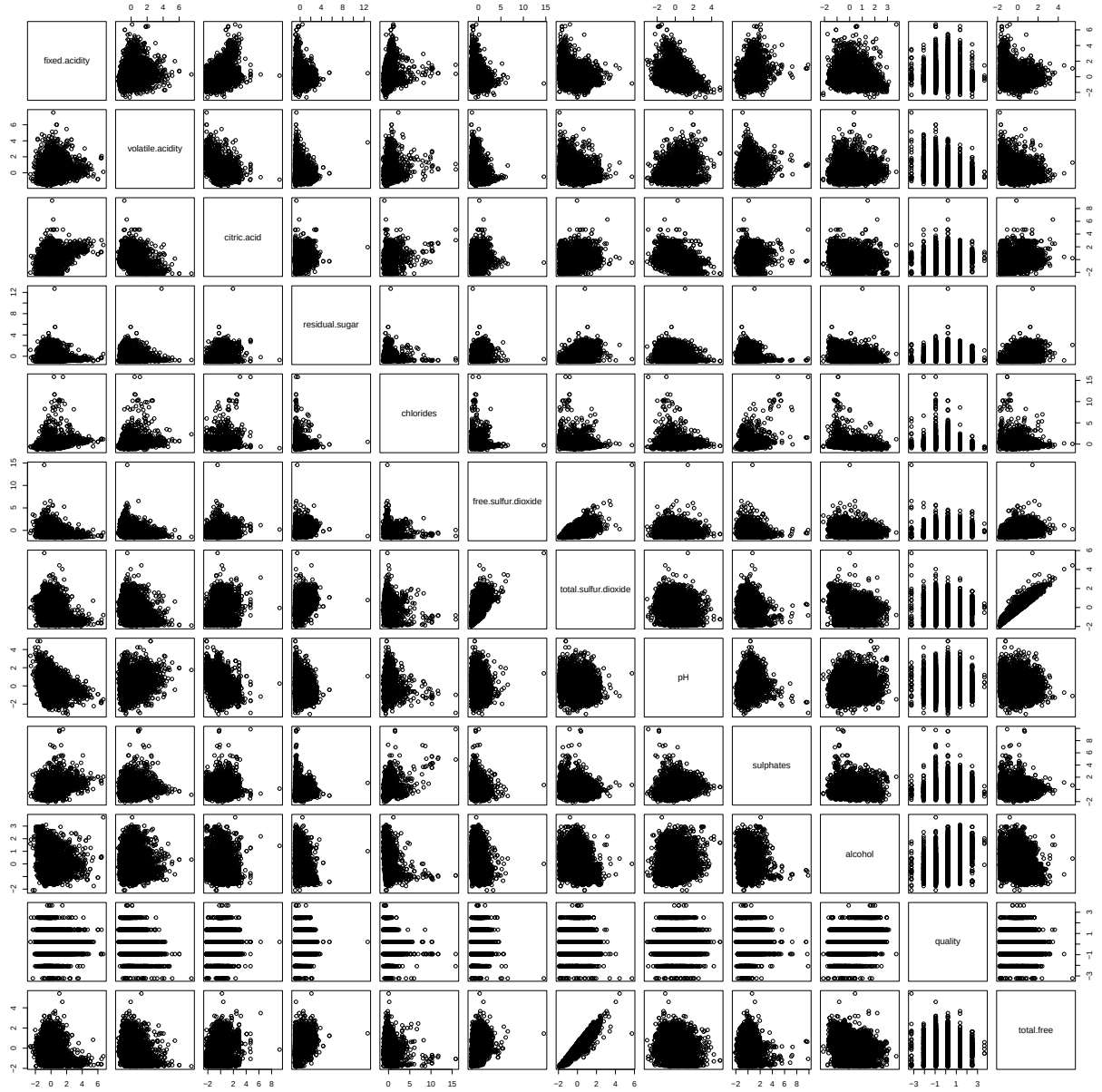
Visualizations of Data

Below we can create some initial visualizations. This includes a correlation matrix and a general plot of variable. The correlation will become useful in our analysis and especially when we get to regression later on.



After developing our correlation graph above we should now create a normalized version of the data to

demonstrate any patterns that may be relevant. I have removed all binary variables for easier viewing.



After looking at the raw data and its derived attributes we should now look to see what quantity of data is considered to be an outlier. From here we can decide weather or not the outlier data is still relevant. Notably we removed ratings and color from determining outliers as these should not relevant to outlier analysis.

```
## [1] "Number of Outliers In Red Wine Data: 465"
```

```
## [1] "Number of Outliers In White Wine Data: 1068"
```

```
## [1] "Number of Outliers In Wine Data: 1667"
```

As it turns out, there are a significant amount of outliers throughout every data set. I have decided to remove only a handful of outliers because this significance is an indication that wine comes in many different

varieties due to its process. This will likely affect the PCA analysis later. Despite keeping nearly every data point I have removed the most egregious outliers because they were many times outside an acceptable range for their respective data sets. Interestingly, the outliers which were removed tended to be lower quality wines. I believe this is a very strong indication that the principal of having too much of something present ruins the experience in something as complex as wine. The only attribute which did not follow this trend was citric acid, which had an acceptable level of quality.

Exploratory Data Analysis

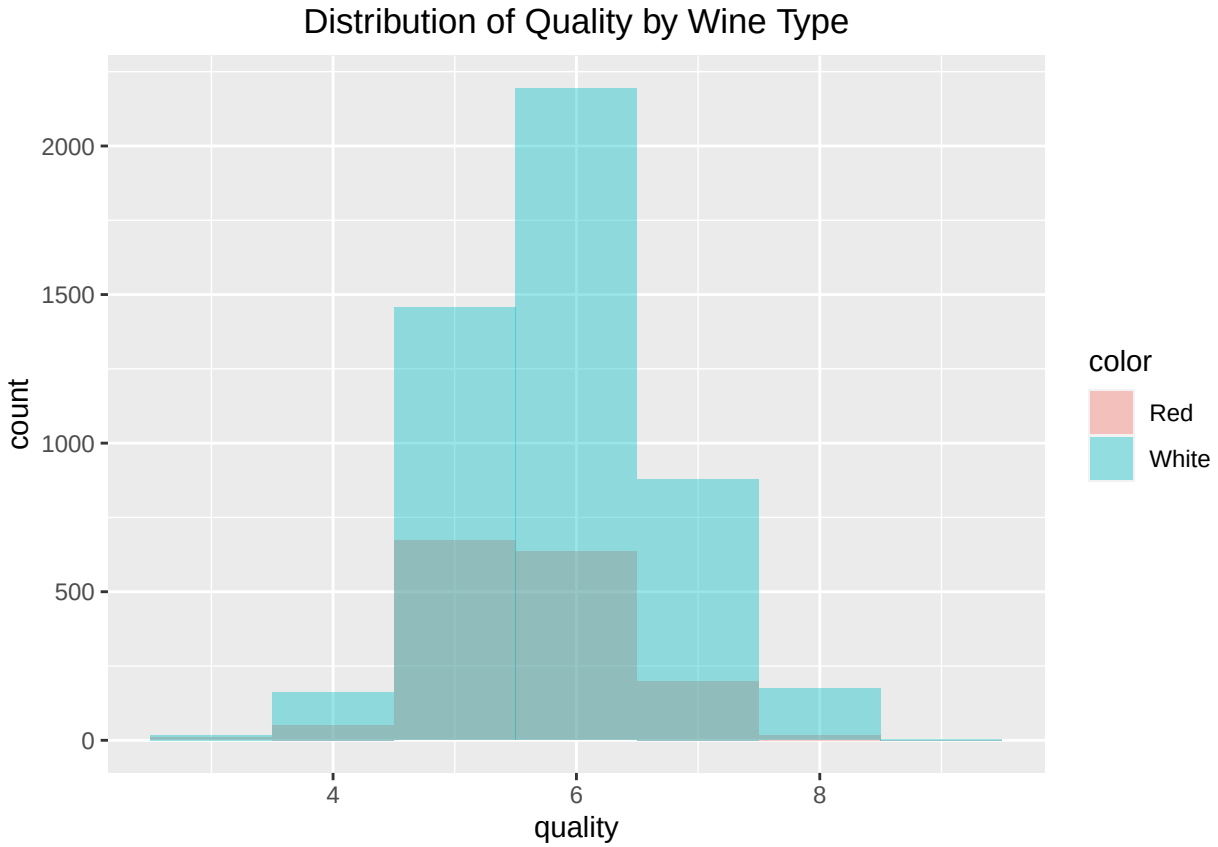
Here we can see an aggregated summary of the wine data. This gives us some meaningful insights regarding the statistics of the data. We have removed color and rating because their statistical properties are irrelevant.

```
summary(Wine %>% select(-color, -rating)) #Aggregated Wine Data Summary
```

```
##  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   : 7.213   Mean   :0.3387   Mean   :0.3182   Mean   : 5.439
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.1850   Max.   :1.0000   Max.   :31.600
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide       pH
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :2.720
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.5        1st Qu.:3.110
##  Median :0.04700   Median : 29.00      Median :118.0        Median :3.210
##  Mean   :0.05573   Mean   : 30.47      Mean   :115.6        Mean   :3.219
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:3.320
##  Max.   :0.46700   Max.   :131.00      Max.   :313.0        Max.   :4.010
##    sulphates         alcohol         quality        total.free
##  Min.   :0.2200   Min.   : 8.00   Min.   :3.000   Min.   :  3.00
##  1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   1st Qu.: 55.00
##  Median :0.5100   Median :10.30   Median :6.000   Median : 86.00
##  Mean   :0.5296   Mean   :10.49   Mean   :5.821   Mean   : 85.13
##  3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000   3rd Qu.:116.00
##  Max.   :1.3600   Max.   :14.90   Max.   :9.000   Max.   :251.50
```

As we can see there still appears to be some significant outliers present in the summary statistics. However, these are considerably closer when compared to the outliers which were removed.

Now that we have cleaned and summarized our data its time to compare and contrast the two data sets by visualizing the inherent differences. Note that the white wine data set contains approximately 3x as many variables as red wine does.

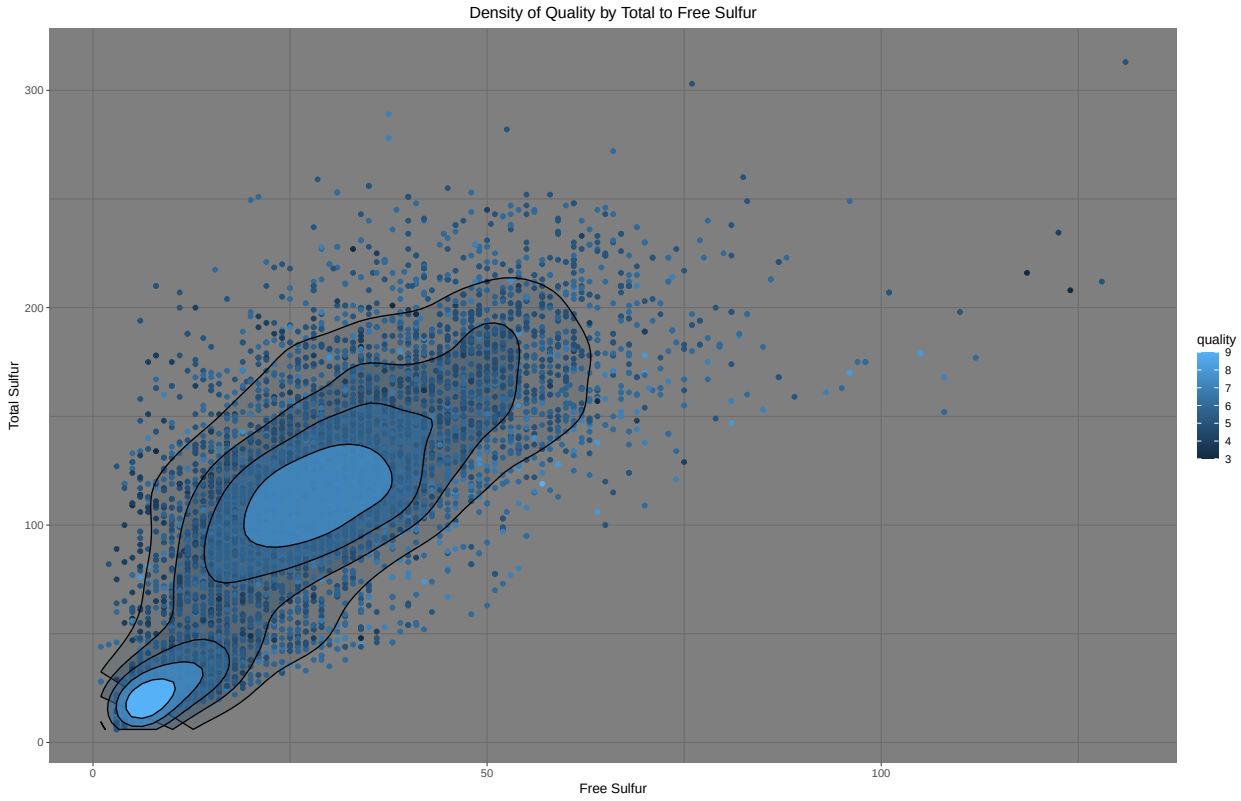


As we can see there are clear differences in the quality counts, but more importantly, we can begin to see how the quality of white wines tends to be greater than red wines. Further investigation of the red wine distribution demonstrates there are more red wines rated to be quality five than six. In addition the red wine subset is skewed to the right, while white wine appears to be normally distributed. This maybe caused by the grape variety being better suited for white wines than red wines. However, because quality is a measure of sensory data we can not be certain of this conclusion.

Following the initial visualizations we can begin to assess each question in depth to form a better understanding of the wine data set.

Question 1:

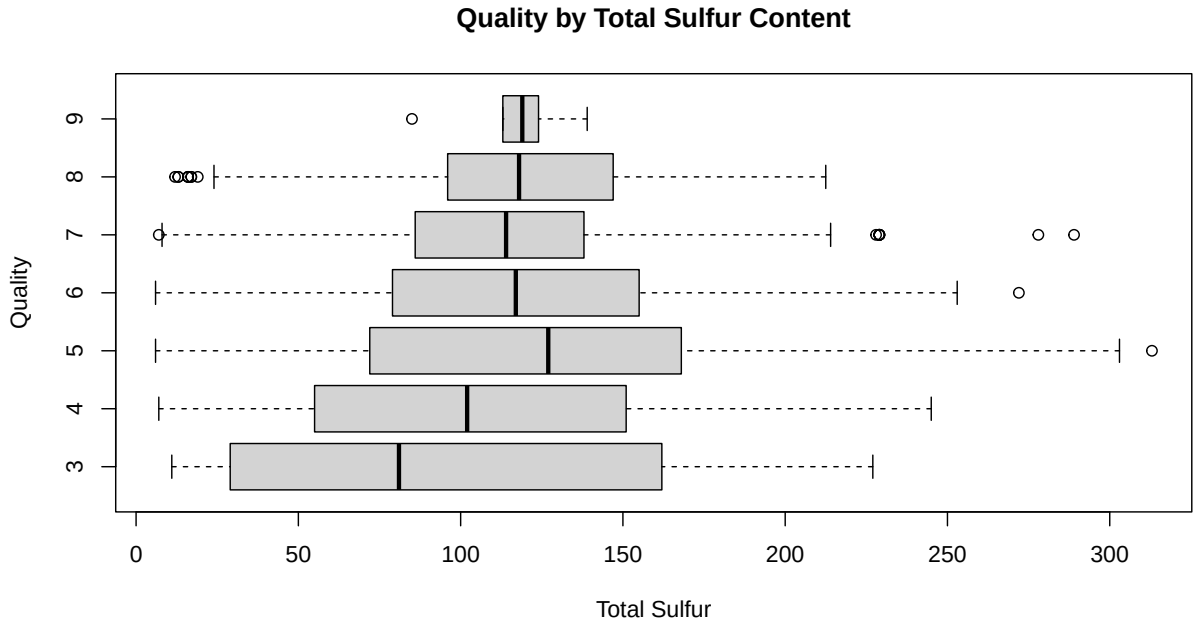
"Is there a strong correlation in determining the quality of a wine when looking at the ratios between total sulfur dioxide and free sulfur dioxide?" Given our initial correlation plot we could see that the highest correlations, outside of our derived attributes, appeared to be between total sulfur dioxide and free sulfur dioxide. These two metrics are inherently similar as they describe the sulfur dioxide content; However, there is an indication that these attributes may be in part responsible for the overall rated quality of a wine.



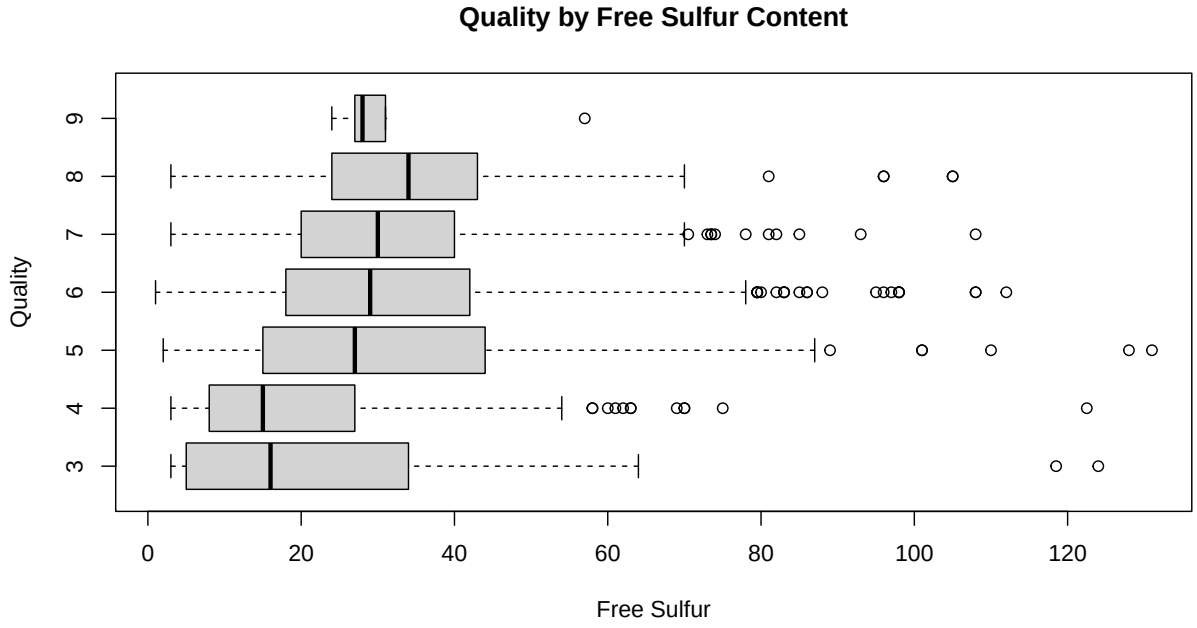
Through the graph above we can see how quality is related to the correlation between free sulfur and total sulfur. We can think of this as a topological graph where the data points of total and free sulfur are elevated based on the quality of the wine. Interestingly, we can see some relationships emerge where it is evident that higher quality wines tend to be centralized in two specific clusters. Particularly, they appear when sulfur contents are approximately (10, 20) and (30, 110). This demonstrates that there is a necessary balance between the sulfur types in a wine that helps develop a wine's complex flavors through bacterial growth.

This alone is not indicative of a relationship, but it helps demonstrate the emerging relation between a wines sensory quality and its physiochemical properties.

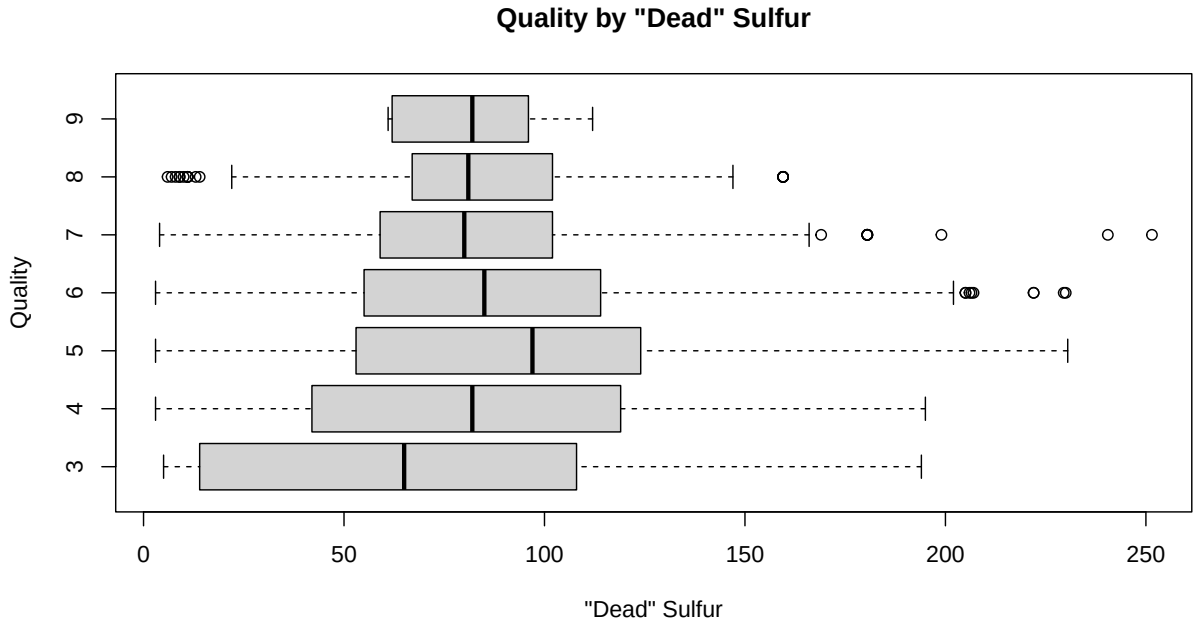
As a result we can expand into the next section and cover the summary statistics of each sulfur based attribute as labelled by its quality. This may give us key insight into the underlying patterns present in each wine quality tier.



We can see in this first box plot that there is a sweet spot for total sulfur when making a high quality wine. Lower quality wines tend to have far greater variation and far different means. We can also see that as wine quality increases the mean of each quality level begins to "tighten" itself around the higher quality wine's mean.



Meanwhile, in this box plot the highest quality wines are incredibly concise in respect to the amount of free sulfur. This is important because free sulfur is specific to sulfur which is added to the wine making process and not already naturally occurring. This implies that there is an ideal amount of sulfur which should be added into a wine to improve the overall quality.



Finally, as we can see in this box plot, the highest quality wines tend to balance sulfur far better by having tighter margins. Interestingly, we can see that this chart models our total sulfer box plot in roughly the same way. Specifically we can see the means start low, overextend and then return to a good balance as the quality increases.

We can conclude that there is some significance in sulfur which plays an important role in the quality of a wine. Moreover, we can expand this by attempting to fit a linear regression model to our data set in an attempt to see which variables are statistically significant in determining the rating of a wine.

Question 2:

Can we fit an accurate linear regression model to the rating of a wine given its physiochemical properties? It is important to keep in mind that the rating attribute we will be attempting to predict is entirely dependent on quality. Wine qualities from 0-6 are considered bad while wines from 7-10 will be considered good. We will condense our wine data into all numeric by one-hot encoding any attributes which are not already numeric.

```
##
## Call:
## lm(formula = rating ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     sulphates + alcohol + color, data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93244 -0.22755 -0.09230  0.05287  1.09159
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.3578806  0.0868633 -15.632  < 2e-16 ***
## fixed.acidity         0.0100603  0.0055125   1.825 0.068070 .
## volatile.acidity     -0.2996945  0.0477049  -6.282 3.65e-10 ***
```

```
## citric.acid          -0.0089997  0.0469346  -0.192 0.847946
## residual.sugar        0.0065581  0.0013911   4.714 2.50e-06 ***
## chlorides            -0.3343943  0.1982835  -1.686 0.091779 .
## free.sulfur.dioxide   0.0016968  0.0004549   3.730 0.000194 ***
## total.sulfur.dioxide -0.0005501  0.0001855  -2.966 0.003032 **
## sulphates             0.2349607  0.0446428   5.263 1.48e-07 ***
## alcohol               0.1363343  0.0054185  25.161  < 2e-16 ***
## color                 0.0351195  0.0263497   1.333 0.182657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3582 on 4520 degrees of freedom
## Multiple R-squared:  0.1904, Adjusted R-squared:  0.1886
## F-statistic: 106.3 on 10 and 4520 DF,  p-value: < 2.2e-16
```

From this initial model we can see which attributes are statistically significant in determining the rating of our wine. Note that both sulfur attributes are playing a statistically significant role in determining the rating of the wine. From here we can extract each significant attribute which affects the predicted rating of a wine and attempt to increase our models R-Squared accuracy.

```
##
## Call:
## lm(formula = rating ~ (volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     total.sulfur.dioxide + sulphates + alcohol), data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88949 -0.22503 -0.09228  0.05180  1.08944
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.2963454  0.0673855 -19.238  < 2e-16 ***
## volatile.acidity     -0.3461335  0.0369305  -9.373  < 2e-16 ***
## residual.sugar        0.0069907  0.0013778   5.074 4.06e-07 ***
## free.sulfur.dioxide   0.0015179  0.0004473   3.393 0.000696 ***
## total.sulfur.dioxide -0.0004225  0.0001512  -2.794 0.005233 **
## sulphates             0.2051703  0.0391551   5.240 1.68e-07 ***
## alcohol               0.1397499  0.0049272  28.363  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3584 on 4524 degrees of freedom
## Multiple R-squared:  0.189,  Adjusted R-squared:  0.1879
## F-statistic: 175.7 on 6 and 4524 DF,  p-value: < 2.2e-16
```

After extraction we have actually increased our F-statistic by double, gained no difference in our P-value, and lost a negligible bit of our R-Squared. However, before continuing, it would be interesting to see a forward and backwards step-wise selection to compare our models.

```
##          (Intercept)              alcohol     volatile.acidity
##        -1.2069354327         0.1314721363        -0.3186459647
##            sulphates       residual.sugar  free.sulfur.dioxide
##         0.2573664644         0.0060704475         0.0017633527
```

```
## total.sulfur.dioxide            chlorides
##        -0.0005394486        -0.5064975858
```

As we can see from developing our forward step feature selection it has simply included all significant variables. When conducting mine I took into account the R-Squared loss and for a model this large I would consider a few of these variables entirely unnecessary to save space and time complexity. However, this does clearly label the important factors associated with modelling the rating of alcohol and sulphates.

Next we will check a backwards step-wise selection and see if it offers us any difference.

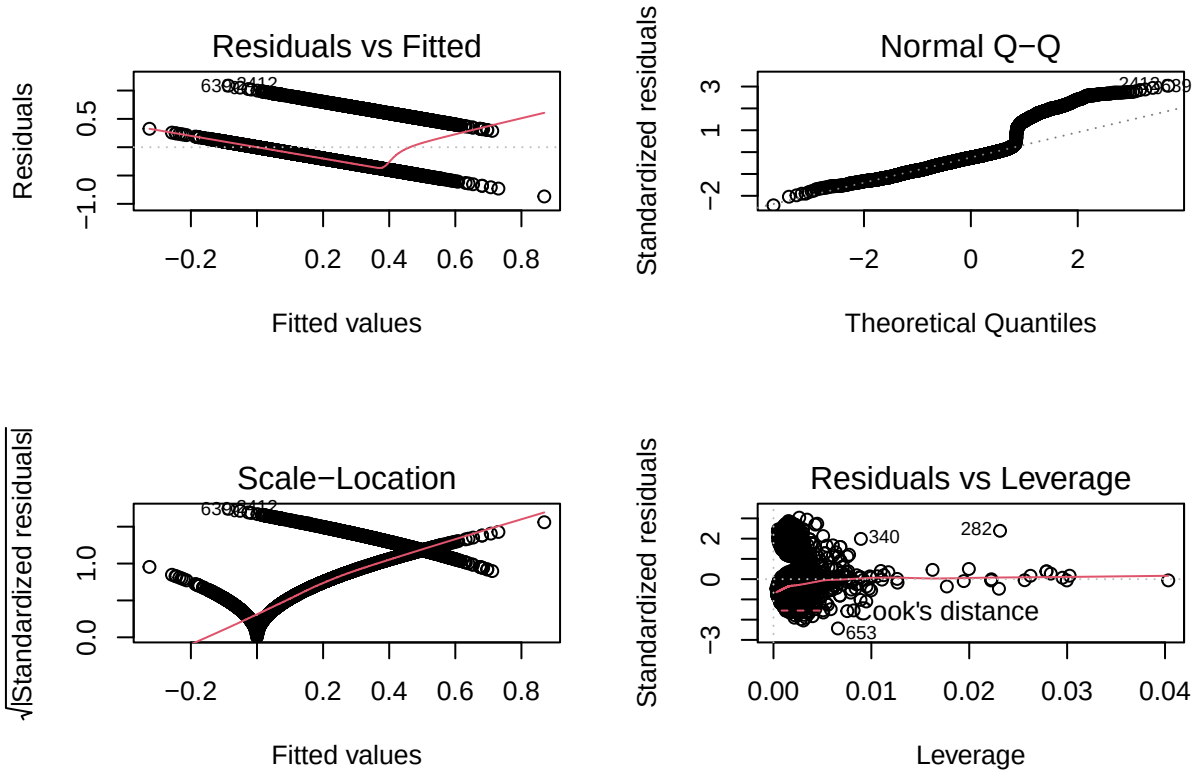
```
intercept <- lm(rating ~ 1, data = Wine)
backward <- step(intercept, direction='backward', scope=formula(WineModel0), trace=0)
backward$coefficients
```

```
## (Intercept)
##   0.1970961
```

Our backwards selection has nothing outside of the intercept. Thus moving forward we will continue with the our forward step-wise selection model.

```
WineModel <- lm(rating ~ volatile.acidity + residual.sugar + chlorides + free.sulfur.dioxide + total.su
summary(WineModel)
```

```
##
## Call:
## lm(formula = rating ~ volatile.acidity + residual.sugar + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + sulphates +
##     alcohol, data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86913 -0.22736 -0.09226  0.05102  1.08607
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.2548678  0.0711021 -17.649  < 2e-16 ***
## volatile.acidity     -0.3287392  0.0381325  -8.621  < 2e-16 ***
## residual.sugar        0.0067779  0.0013824   4.903 9.77e-07 ***
## chlorides            -0.3494119  0.1915496  -1.824 0.068198 .
## free.sulfur.dioxide   0.0015369  0.0004473   3.436 0.000596 ***
## total.sulfur.dioxide -0.0004590  0.0001525  -3.009 0.002633 **
## sulphates             0.2271559  0.0409585   5.546 3.09e-08 ***
## alcohol               0.1364439  0.0052488  25.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3583 on 4523 degrees of freedom
## Multiple R-squared:  0.1896, Adjusted R-squared:  0.1884
## F-statistic: 151.2 on 7 and 4523 DF,  p-value: < 2.2e-16
```



There appears to be some interesting splits in the "Residuals vs Fitted" and "Scale-Location" graphs. More importantly, the data appears to be linear in both of these cases which leads me to the conclusion that when creating this model there was something I am not aware of that is greatly damaging the modelling. This may be the result of using sensory data and the many complexities of the wine making process. One important note is that it appears our data tends to be normally distributed, but it isn't as a result of the data after our first standard deviation.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1525  329
##          1   34   55
##
##                Accuracy : 0.8132
##                  95% CI : (0.7951, 0.8303)
##     No Information Rate : 0.8024
##     P-Value [Acc > NIR] : 0.121
##
##                   Kappa : 0.1709
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9782
##             Specificity : 0.1432
##          Pos Pred Value : 0.8225
```

```
##          Neg Pred Value : 0.6180
##              Prevalence : 0.8024
##          Detection Rate : 0.7849
##    Detection Prevalence : 0.9542
##       Balanced Accuracy : 0.5607
##
##        'Positive' Class : 0
##
```

After testing our modelling accuracy we can see that we are approximately 80% accurate when determining whether a wine is rated good or bad. Although okay, it is not strong enough to be considered accurate or precise.

As a result I would like to try this model again, but this time I will remove the outliers from the data set and see if they are having a large affect on the modelling accuracy.

```
##
## Call:
## lm(formula = rating ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     sulphates + alcohol + color, data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93079 -0.22713 -0.09228  0.05355  1.09155
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.3554281  0.0857350 -15.810  < 2e-16 ***
## fixed.acidity         0.0095769  0.0048878   1.959 0.050136 .
## volatile.acidity     -0.2967260  0.0451120  -6.578 5.33e-11 ***
## residual.sugar        0.0065412  0.0013876   4.714 2.50e-06 ***
## chlorides            -0.3405010  0.1955289  -1.741 0.081675 .
## free.sulfur.dioxide   0.0016977  0.0004549   3.732 0.000192 ***
## total.sulfur.dioxide -0.0005536  0.0001846  -2.999 0.002722 **
## sulphates             0.2342525  0.0444908   5.265 1.47e-07 ***
## alcohol               0.1362134  0.0053724  25.355  < 2e-16 ***
## color                 0.0345898  0.0262168   1.319 0.187110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3582 on 4520 degrees of freedom
## Multiple R-squared:  0.1904, Adjusted R-squared:  0.1887
## F-statistic: 118.1 on 9 and 4520 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = rating ~ (volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     total.sulfur.dioxide + sulphates + alcohol), data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88956 -0.22506 -0.09235  0.05190  1.08949
##
```

```
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.2965617  0.0674150 -19.233  < 2e-16 ***
## volatile.acidity     -0.3461035  0.0369352  -9.371  < 2e-16 ***
## residual.sugar        0.0069936  0.0013782   5.075 4.04e-07 ***
## free.sulfur.dioxide   0.0015184  0.0004474   3.394 0.000695 ***
## total.sulfur.dioxide -0.0004226  0.0001513  -2.794 0.005227 **
## sulphates             0.2052143  0.0391609   5.240 1.68e-07 ***
## alcohol               0.1397646  0.0049292  28.355  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3584 on 4523 degrees of freedom
## Multiple R-squared:  0.189,  Adjusted R-squared:  0.1879
## F-statistic: 175.7 on 6 and 4523 DF,  p-value: < 2.2e-16
```

Once again we will display the step-wise selections below to assess how we determined our models final selection.

```
forward <- step(intercept, direction='forward', scope=formula(NewWineModelTest), trace=0)
forward$coefficients
```

```
##          (Intercept)              alcohol     volatile.acidity
##        -1.2069354327         0.1314721363        -0.3186459647
##            sulphates       residual.sugar  free.sulfur.dioxide
##         0.2573664644         0.0060704475         0.0017633527
## total.sulfur.dioxide            chlorides
##        -0.0005394486        -0.5064975858
```

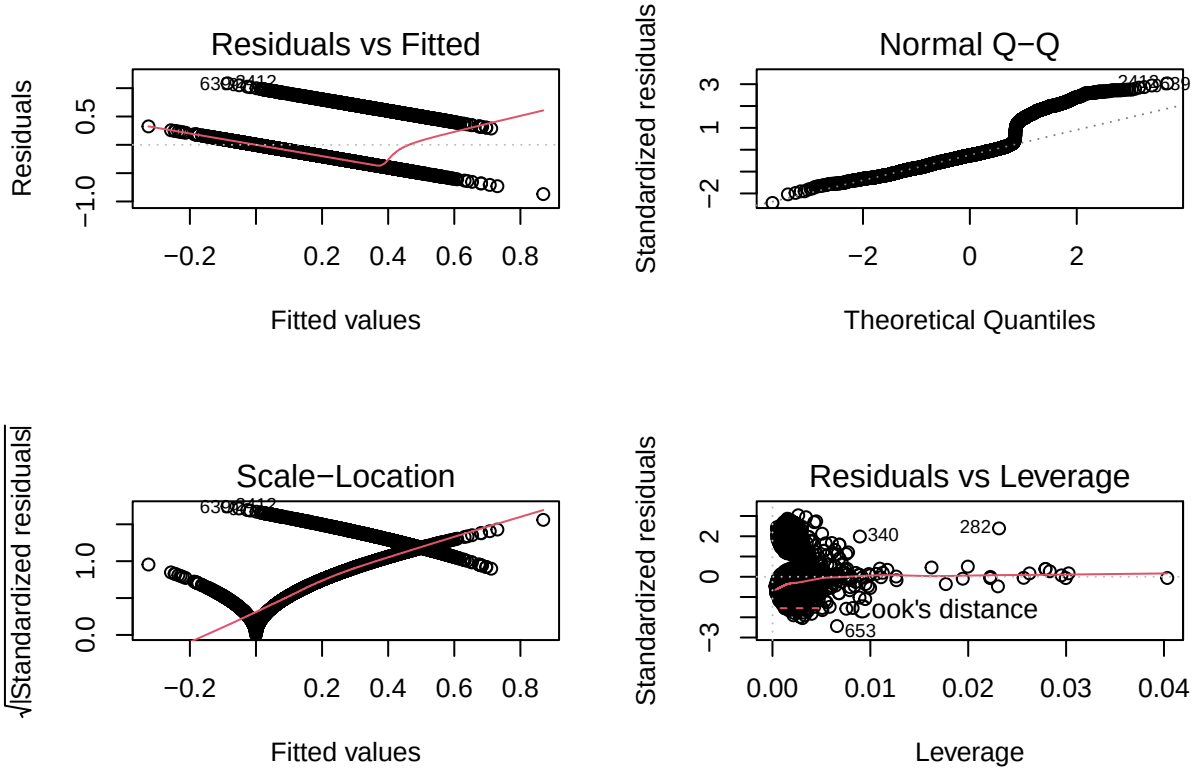
```
backward <- step(intercept, direction='backward', scope=formula(NewWineModelTest), trace=0)
backward$coefficients
```

```
## (Intercept)
##   0.1970961
```

One again, we can see that after conducting our step-wise selections on the new data set without outliers we observe no change across any of the coefficients. This means that there is little to no change in the model with or without outliers. In learning of this I am now far more confident with my decision to keep the outliers.

```
##
## Call:
## lm(formula = rating ~ volatile.acidity + residual.sugar + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + sulphates +
##     alcohol, data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86920 -0.22751 -0.09232  0.05106  1.08612
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          -1.2550734  0.0711385 -17.643  < 2e-16 ***
## volatile.acidity     -0.3287273  0.0381368  -8.620  < 2e-16 ***
## residual.sugar        0.0067804  0.0013828   4.903 9.75e-07 ***
## chlorides            -0.3491613  0.1915863  -1.822 0.068449 .
## free.sulfur.dioxide   0.0015372  0.0004474   3.436 0.000595 ***
## total.sulfur.dioxide -0.0004590  0.0001525  -3.009 0.002633 **
## sulphates             0.2271759  0.0409634   5.546 3.09e-08 ***
## alcohol               0.1364583  0.0052512  25.986  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3583 on 4522 degrees of freedom
## Multiple R-squared:  0.1896, Adjusted R-squared:  0.1883
## F-statistic: 151.1 on 7 and 4522 DF,  p-value: < 2.2e-16
```



After running the new model without outliers we have seen no change in any aspect of the modelling. While my hopes of developing a linear regression model for adequately predicting the rating of a wine has been dashed, I am now confident that our outliers were not affecting the model or the data at large. I have reinforced my emerging understanding that wine qualities are significantly more complex than the sum of their physiochemical properties. Ones that the residuals would likely need in order to account for the immense variability which is not described inherently in our data set. Despite this I still want to check to see how accurate this classifier is.

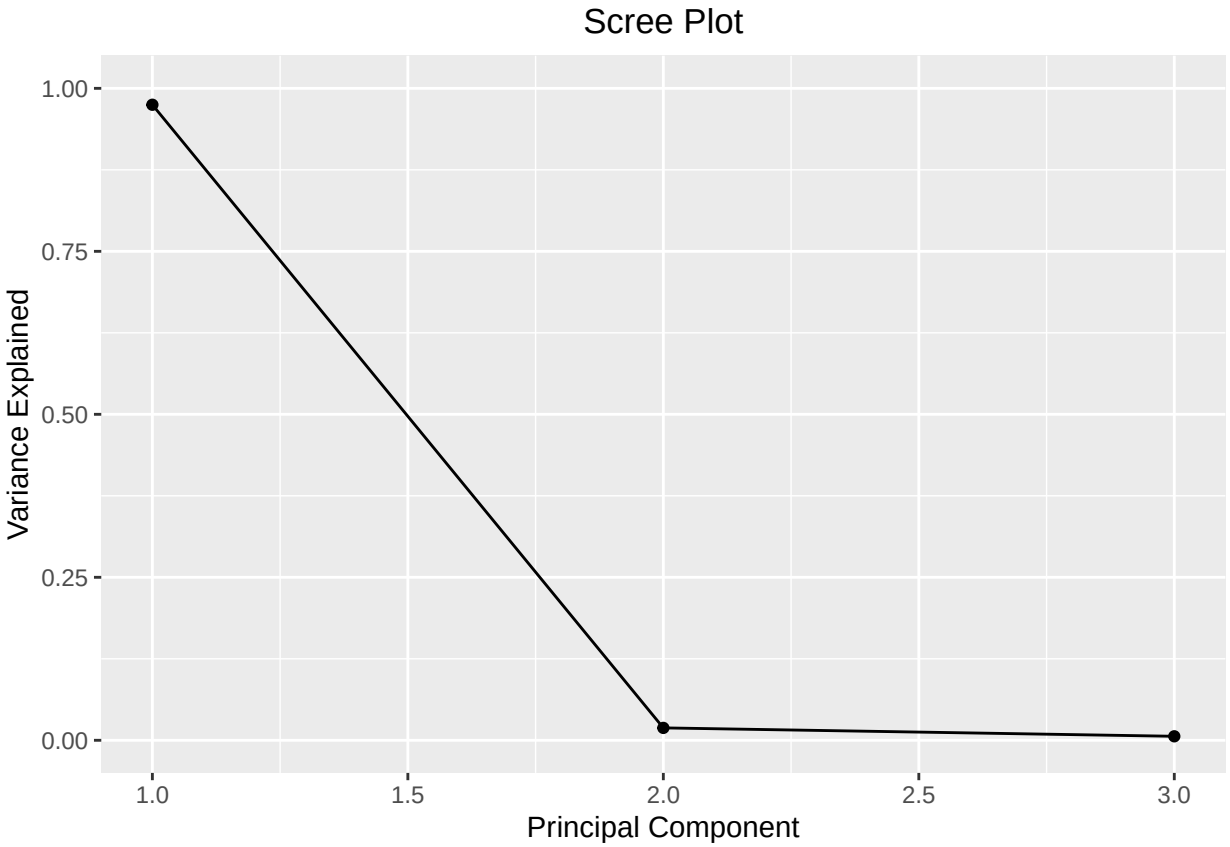
```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction    0    1
##          0 1524  329
##          1   34   55
##
##                Accuracy : 0.8131
##                  95% CI : (0.795, 0.8302)
##     No Information Rate : 0.8023
##     P-Value [Acc > NIR] : 0.121
##
##                   Kappa : 0.1709
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9782
##             Specificity : 0.1432
##          Pos Pred Value : 0.8225
##          Neg Pred Value : 0.6180
##              Prevalence : 0.8023
##          Detection Rate : 0.7848
##    Detection Prevalence : 0.9542
##       Balanced Accuracy : 0.5607
##
##        'Positive' Class : 0
##
```

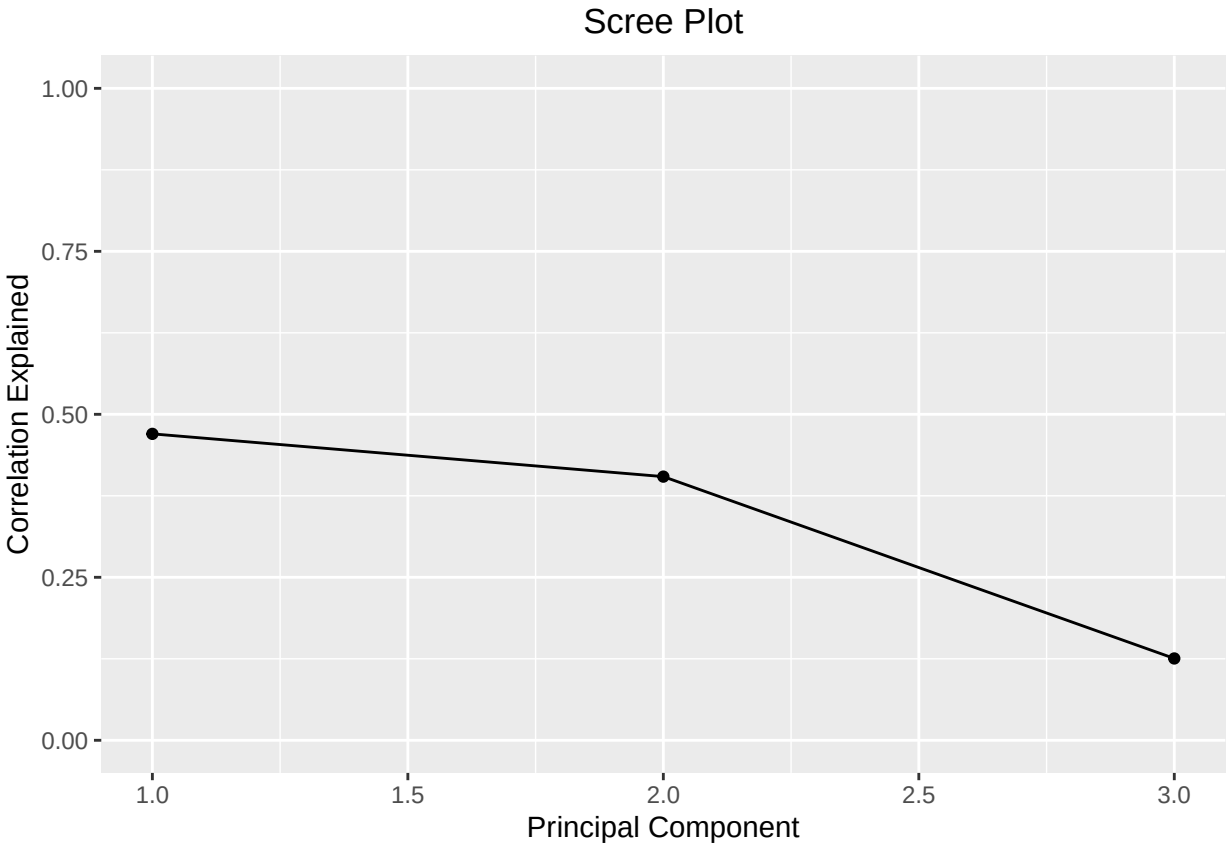
After modeling rating for the wine data set, we were only able to correctly identify approximately 80% of wine ratings. This however is within the confidence interval of our previous model's accuracy as well. As a result we must conclude that there is statistically no change in predicting accuracy when removing outliers. Despite this we can still say that our model is better than simply guessing, but not good enough for any real world applications.

Question 3:

Which acidity attribute plays an important role? After preparing each respective data set by scaling the data, and collecting their eigenvalues, it is now time to determine the percentages of our checks. Here we will be able to see which aspect of acidity is playing a crucial role.



It is clear that the first column (fixed acidity) is playing the only significant role when it comes to covariance as the other columns are almost 0 respectively. As a result we can state that fixed acidity can explain a total of 97% of the total variance we find in this analysis.



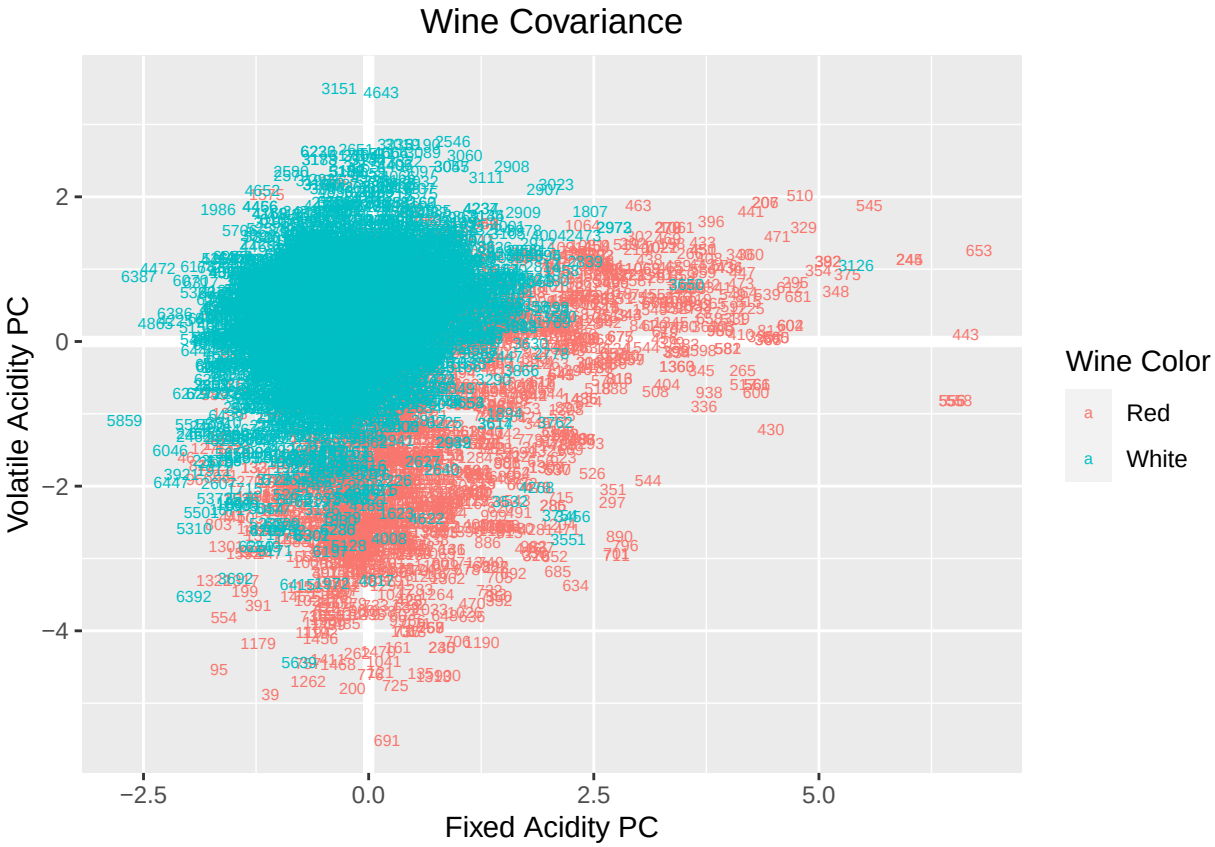
Unlike the covariance matrix the correlation matrix presents us with an interesting scree plot which demonstrates that the first two principal components are playing a very similar role in determining the correlation.

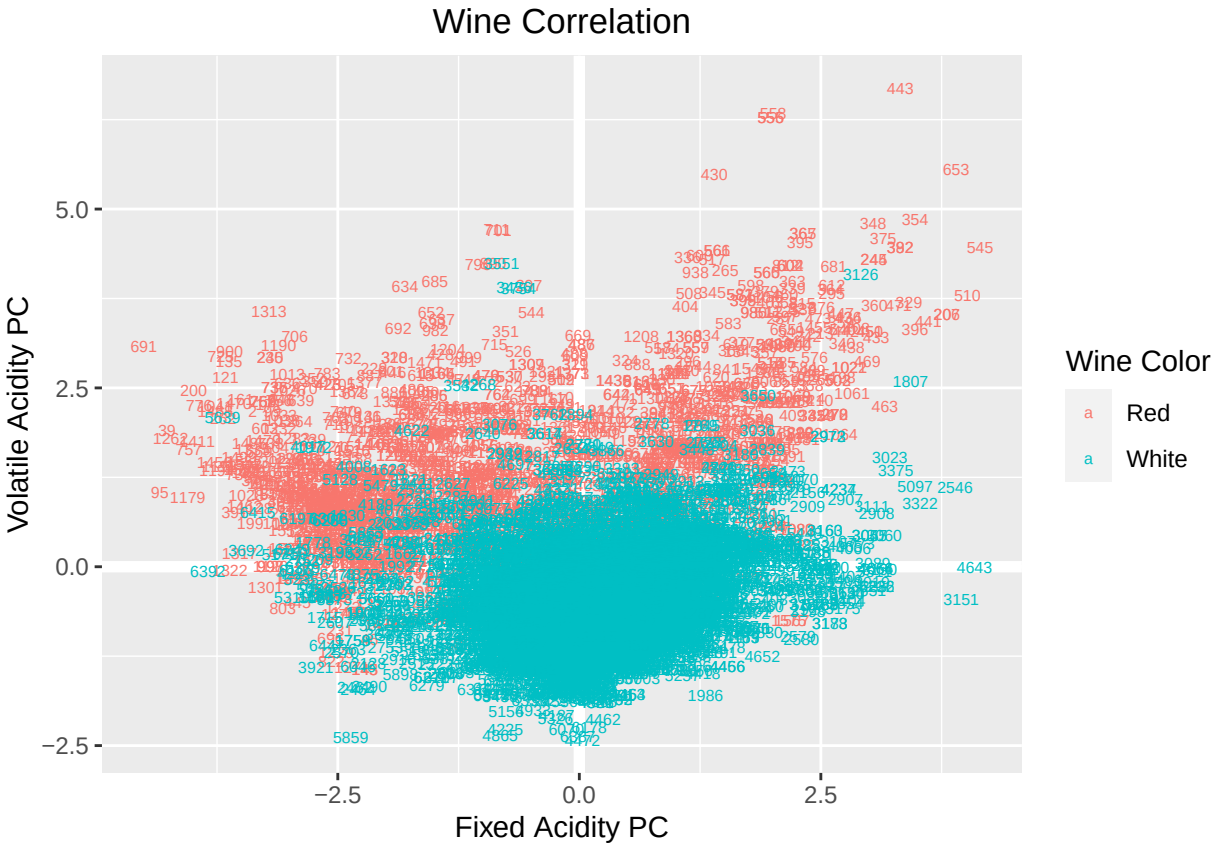
Next we want to perform some checks to ensure the sums of our values still total to one and are thus valid.

```
## [1] 0.9748274 0.9938477 1.0000000
```

```
## [1] 0.4700191 0.8744647 1.0000000
```

After the initial development it is clear that columns 1 and 2 have the highest impact. However, column 1's covariance is significantly higher than the other two, while correlation is spread out relatively evenly between the first two columns. These columns are consistent with the attributes of fixed.acidity and volatile.acidity.


Wine Covariance
Volatile Acidity PC
Fixed Acidity PC
2
0
−2
−4
−2.5
0.0
2.5
5.0
Wine Color
a
Red
a
White



After representing all the information in the covariance and correlation matrices, we conducted eigen decomposition on the resulting matrices, and have reached our conclusion. We can see that in both cases our data is clustered around the mean of (0, 0). This PCA analysis has helped us see how the two types of acidity appear to have some influence. Particularly fixed acidity has a greater impact on the correlation side while volatile acidity impacts the covariance more. Intuitively this makes sense due to the individual meanings of fixed and volatile. As a reminder, fixed acidity is representative of 4 types of acids and volatile represents the gaseous components of these acids. More graphs related to each wine color demonstrating tight clustering can be found in the appendix section. In those cases we can see more evidently how red wine differs from white wine.

Question 4:

Can we model wine color by its physiochemical properties?

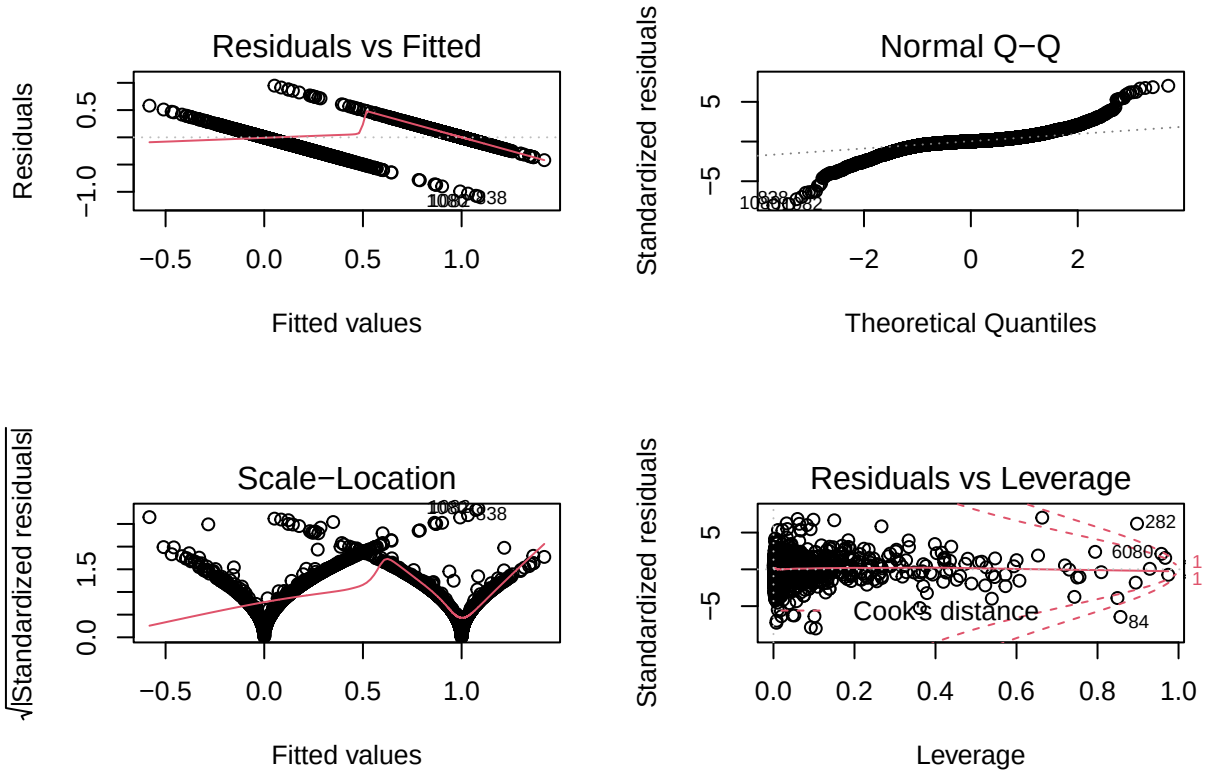
```
##
## Call:
## lm(formula = color ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol + quality, data = trainingData,
##     family = binomial)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54520 -0.10383  0.00821  0.11362  0.67276
##
## Coefficients:
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.911e+00  8.644e-02  33.672  < 2e-16 ***
## fixed.acidity        -8.892e-02  2.964e-03 -30.001  < 2e-16 ***
## volatile.acidity     -7.720e-01  2.309e-02 -33.435  < 2e-16 ***
## citric.acid           1.177e-01  2.465e-02   4.777 1.84e-06 ***
## residual.sugar       -3.174e-04  7.512e-04  -0.422   0.6727
## chlorides            -1.731e+00  1.070e-01 -16.178  < 2e-16 ***
## free.sulfur.dioxide  -2.661e-03  2.378e-04 -11.191  < 2e-16 ***
## total.sulfur.dioxide  3.557e-03  8.526e-05  41.724  < 2e-16 ***
## pH                   -5.028e-01  2.128e-02 -23.634  < 2e-16 ***
## sulphates            -4.580e-01  2.336e-02 -19.602  < 2e-16 ***
## alcohol               3.640e-02  3.105e-03  11.724  < 2e-16 ***
## quality              -7.310e-03  3.850e-03  -1.899   0.0576 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.189 on 4519 degrees of freedom
## Multiple R-squared:  0.8053, Adjusted R-squared:  0.8048
## F-statistic:  1699 on 11 and 4519 DF,  p-value: < 2.2e-16
```

```
decantedModel <- lm(color ~ (free.sulfur.dioxide * total.sulfur.dioxide) + (sulphates * chlorides * alc
# summary(decantedModel)
```

For the sake of this paper's length I will not include a summary of the model, but the residual error is .1462 and the R-Squared is .888. For those interested I encourage you to run the code for yourself and un-comment the summary code.

After some further tuning I was able to find a model which was lower than the original base model in regards to its residual error. Although this model still displays a high residual deviance, it is lower than our starting model. Additionally, we were able to see a marginal increase in our R-Squared.



Just as we saw earlier in question 2 our residuals are plotted in more or less the same way. The difference this time is that our models fit the data very well, almost scarily well, given the interesting shapes we see. We can see that our Q-Q plot gives us an understanding that the model is not perfectly normally distributed, but each of the tail ends help to balance the model out.

Continuing forward we will determine if we can make accurate predictions regarding the color of the wine given our previous findings.

```
##      reference
## data    0    1
##   -2    1    0
##   -1    2    0
##   0   472   10
##   1    18 1440
```

Interestingly, there appears to be some data that ends up being classified as a negative number. I am not sure what to make of this, but I've incorporated them into the False Positive count below.

```
sprintf("Accuracy Level: %f", (TP + TN) / (FP + FN + TP + TN))
```

```
## [1] "Accuracy Level: 0.984045"
```

```
sprintf("Inaccuracy Level: %f", (FP + FN)/(TP + TN))
```

```
## [1] "Inaccuracy Level: 0.016213"
```

It can correctly identify the color of the wine 98% of the time, which is far better than the human average of 51%. Often blind wine taste testing results in what is essentially a coin flip, but our model was able to distinguish between the two easily. We can see a strong indication that the color of the wine can be predicted accurately based on physiochemical properties. Additionally, and more importantly, given that this model has surpassed the 95% threshold we would desire for significance, we can state that their are important properties which can indicate what color of wine the data represents. This is intuitive as the the process for making white wine vs red wine is different. As a result the physical traits we would see in our data should also be affected and would inadvertently appear in the model as well.

Conclusion

After analyzing the data I feel comfortable in concluding that while these properties do play a role in the overall quality of a wine they are only a small part of a bigger picture. Additionally, it is important to consider these findings within the context of the data collection, notably that this data comes from a specific variety of grape. These considerations may not be applicable to the greater wine field as different grape types may yield significantly different properties than the Vinho Verde variety. Despite this, because Portugal is a top ten wine exporter, we can say that it is somewhat relevant in the greater analyzation of wine. We were able to see some distinct traits and trends in the data including our sulfur sweet spot. This small section indicates there is complexity in the system which is not captured by the data set, nor modeled in our attempts such as our PCA. This is only furthered by our inability to classify our rating regression assessments accurately. One proposed method by Paulo Cortez to tackle this shortcoming would be to build a linear regression model that gives a predicted wine quality a higher weight the closer it is to its rated quality. Future research may find success in attempting to increase the modelling dimensions to find clusters of closely related values of wine. Additionally, we may find that more extreme forms of modelling, such as deep learning, may be able to extract complexities in the physiochemical properties that we can not detect.

The approaches used also leave our analysis with a lot to be desired. Specifically when conducting our PCA it was notable that the quantity of outliers we had may have contributed to the poor PCA results. This may be because PCA has a hard time dealing with outliers. Additionally, the models we created and tested for our rating analysis did not produce desirable results. This maybe be because our approach to the problem is fundamentally flawed as quality is very subjective, even by professionals. As I have mentioned earlier, this may in turn be derived from the fact that our data set does not capture the many complexities of wine.

In conclusion, we were not able to develop a model which achieved any desirable accuracy for real world assessment of wine quality. Even if we successfully developed such a model with high accuracy and high precision we would have built it solely on the vinho verde grape variety which leads me to assume that we may not be able to use the model to predict quality in the greater wine field. This analysis has given me a lot of insight to the world of wines and its many complexities. If I have learned anything, it is that wine is incredibly subject and far more complex than the sum of its parts. Hopefully those who read this analysis have learned to appreciate wine the way I do now.

Appendix

Unused Methods:

```
summary(redWine)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 4.600   Min.   :0.1200   Min.   :0.0000   Min.   : 0.900
##  1st Qu.: 7.100   1st Qu.:0.3900   1st Qu.:0.0900   1st Qu.: 1.900
##  Median : 7.900   Median :0.5200   Median :0.2550   Median : 2.200
##  Mean   : 8.318   Mean   :0.5268   Mean   :0.2693   Mean   : 2.512
```

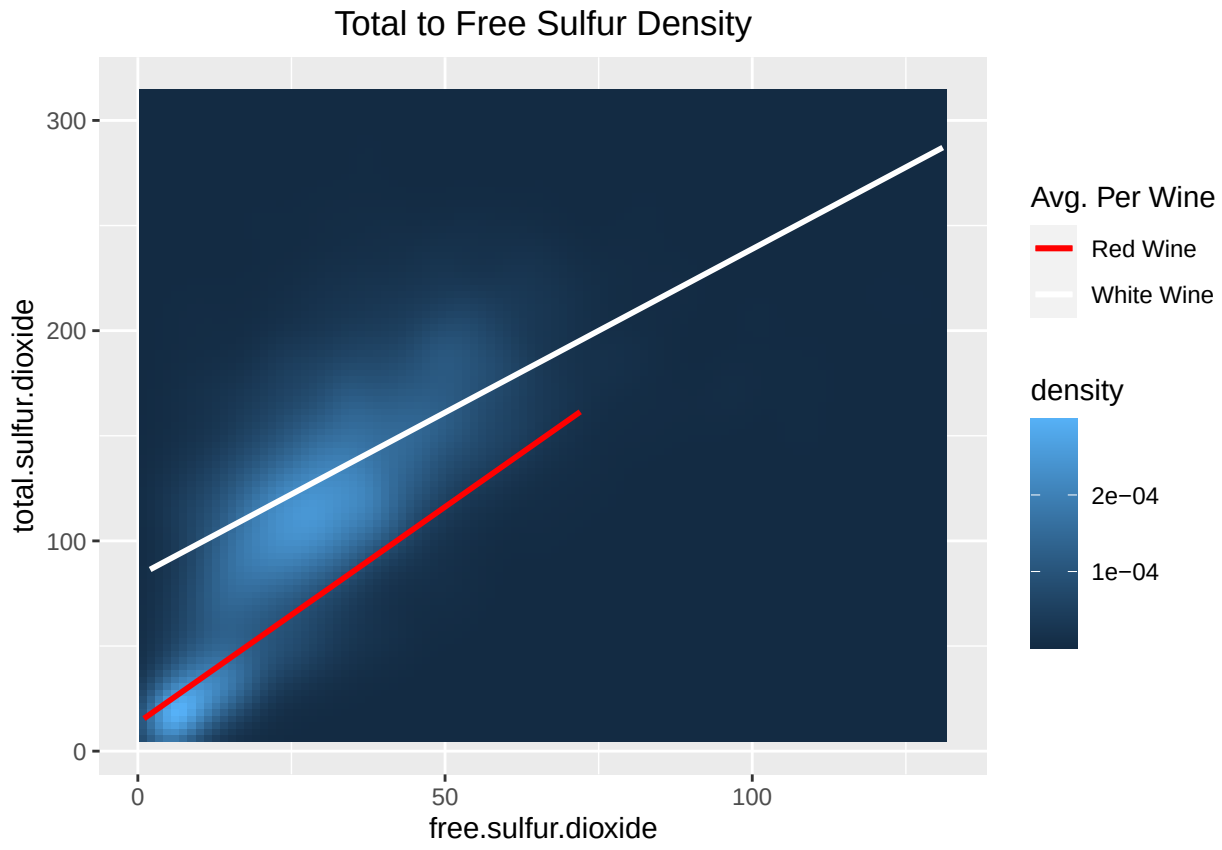
```
##  3rd Qu.: 9.200   3rd Qu.:0.6400   3rd Qu.:0.4200   3rd Qu.: 2.600
##  Max.   :15.900   Max.   :1.2400   Max.   :0.7900   Max.   :13.900
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide       pH
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :2.860
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:3.210
##  Median :0.07900   Median :13.50       Median : 38.00       Median :3.310
##  Mean   :0.08583   Mean   :15.79       Mean   : 45.91       Mean   :3.314
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 61.00       3rd Qu.:3.400
##  Max.   :0.42200   Max.   :72.00       Max.   :165.00       Max.   :4.010
##    sulphates         alcohol          quality       total.free
##  Min.   :0.3300   Min.   : 8.40   Min.   :3.00   Min.   :  3.00
##  1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.00   1st Qu.: 12.00
##  Median :0.6200   Median :10.20   Median :6.00   Median : 21.00
##  Mean   :0.6517   Mean   :10.43   Mean   :5.64   Mean   : 30.11
##  3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.00   3rd Qu.: 39.00
##  Max.   :1.3600   Max.   :14.90   Max.   :8.00   Max.   :128.00
##      rating
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1361
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

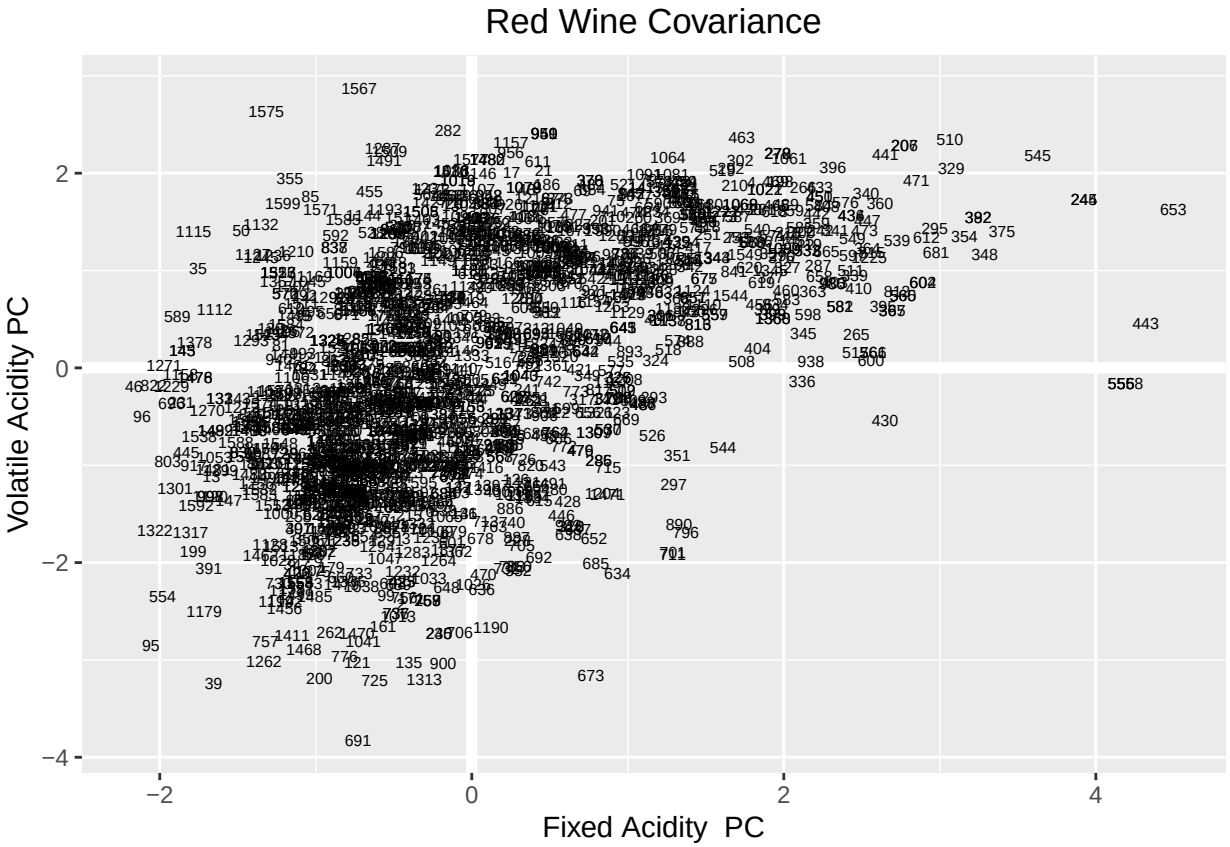
```
summary(whiteWine)
```

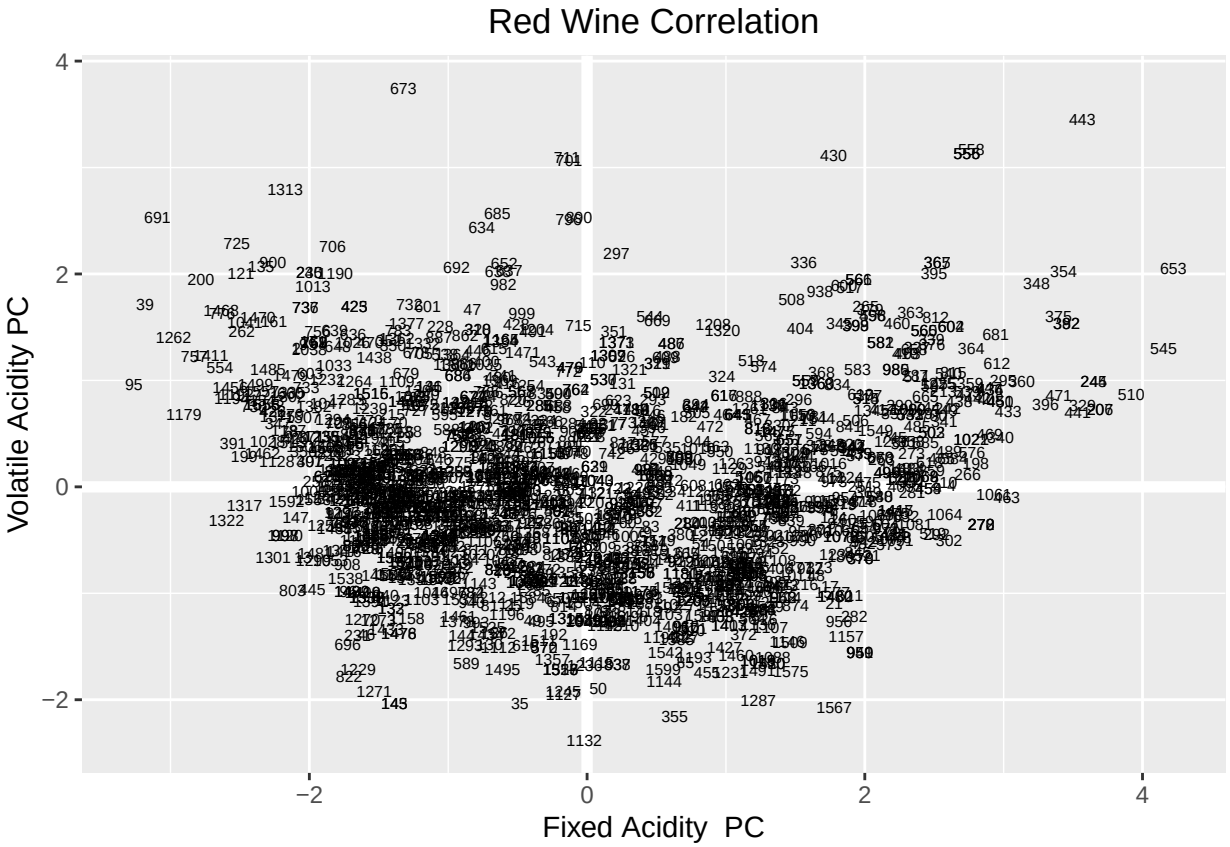
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.852   Mean   :0.2777   Mean   :0.3338   Mean   : 6.372
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :11.800   Max.   :1.0050   Max.   :1.0000   Max.   :26.000
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide       pH
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :2.720
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:3.090
##  Median :0.04300   Median : 34.00      Median :134.0        Median :3.180
##  Mean   :0.04566   Mean   : 35.21      Mean   :138.1        Mean   :3.188
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:3.280
##  Max.   :0.29000   Max.   :131.00      Max.   :313.0        Max.   :3.820
##    sulphates         alcohol          quality        total.free
##  Min.   :0.2200   Min.   :0.0800   Min.   :3.000   Min.   :  4.0
##  1st Qu.:0.4100   1st Qu.:0.2100   1st Qu.:5.000   1st Qu.: 78.0
##  Median :0.4700   Median :0.2600   Median :6.000   Median :100.0
##  Mean   :0.4897   Mean   :0.2777   Mean   :5.881   Mean   :102.9
##  3rd Qu.:0.5500   3rd Qu.:0.3200   3rd Qu.:6.000   3rd Qu.:125.0
##  Max.   :1.0800   Max.   :1.0000   Max.   :9.000   Max.   :230.0
##      rating
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.217
##  3rd Qu.:0.000
##  Max.   :1.000
```

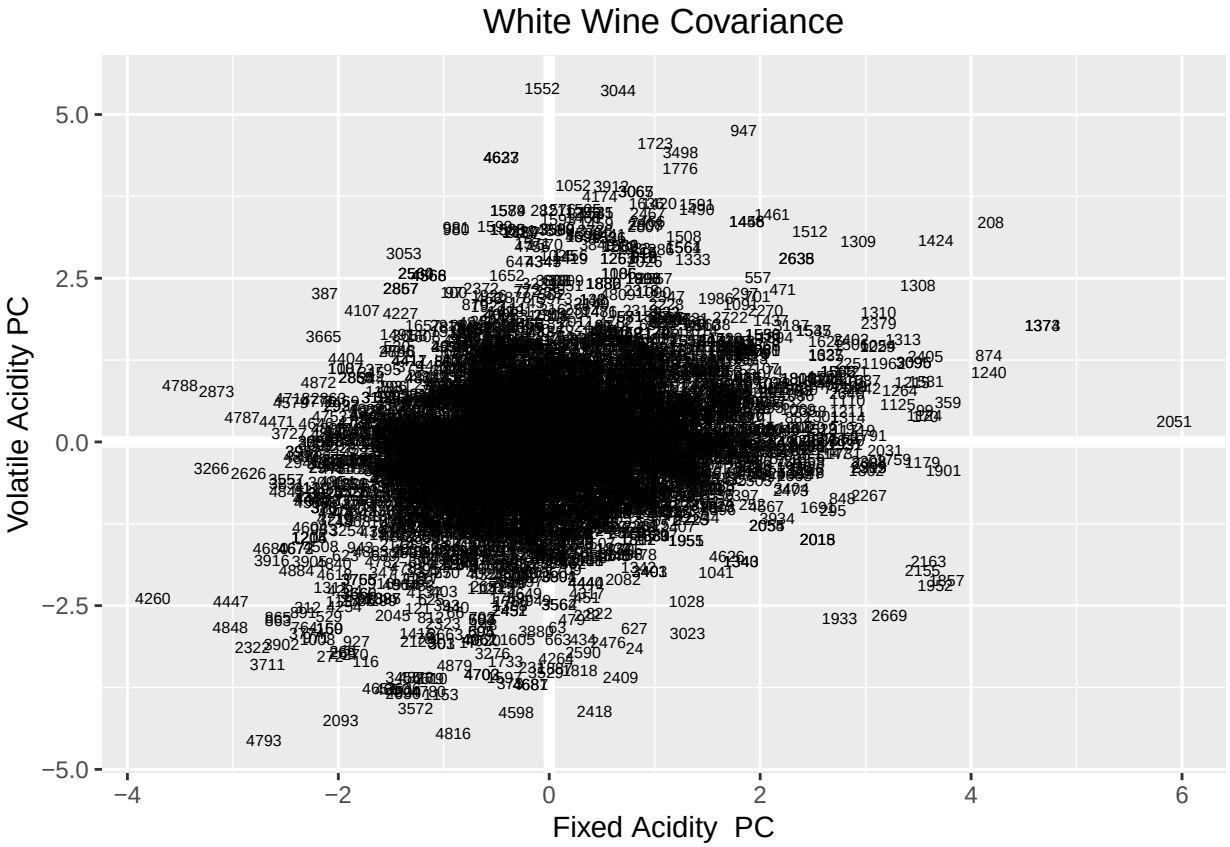
Unused Graphs:

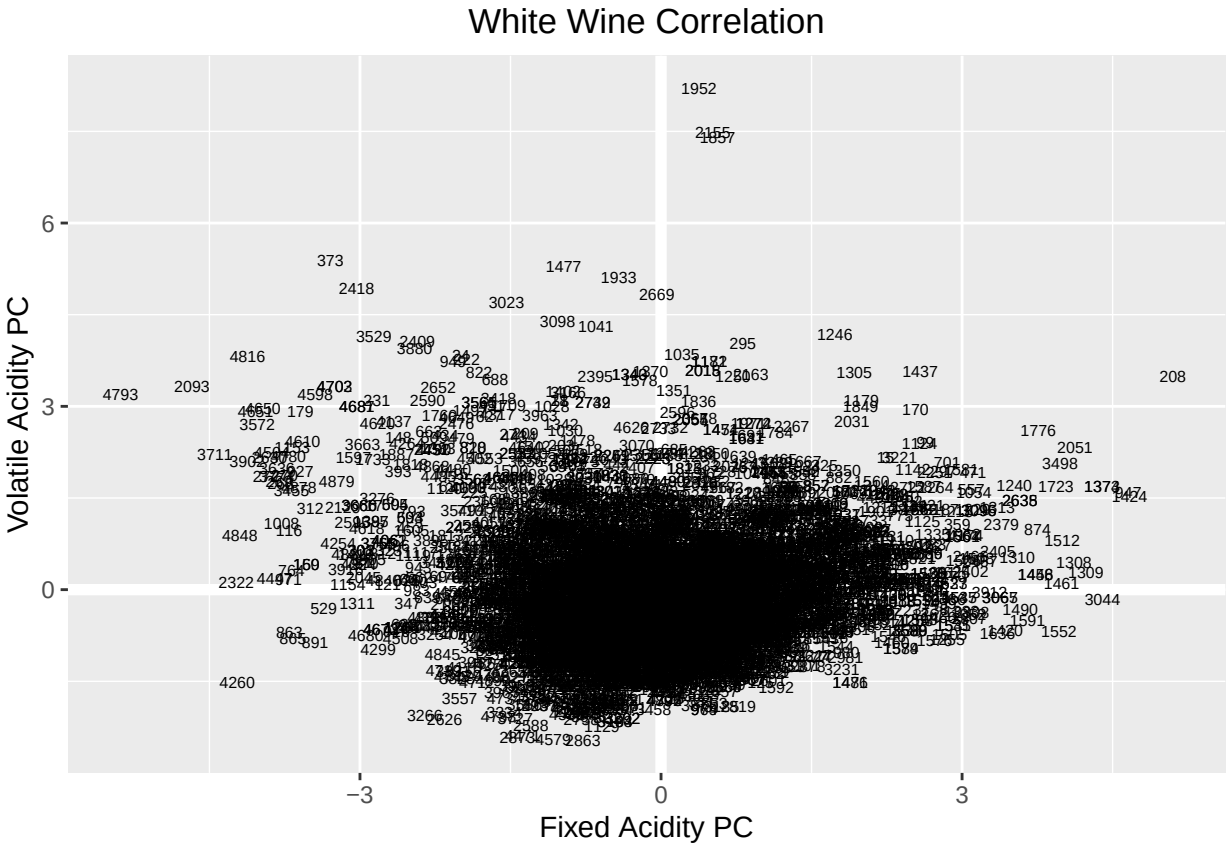
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




Red Wine Covariance
Volatile Acidity PC
Fixed Acidity PC
2
0
−2
−4
−2
0
2
4


Red Wine Correlation
Volatile Acidity PC
Fixed Acidity PC
4
2
0
−2
−2
0
2
4


White Wine Covariance
Volatile Acidity PC
Fixed Acidity PC
5.0
2.5
0.0
−2.5
−5.0
−4
−2
0
2
4
6



Unused Models:

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol, data = redWine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71561 -0.37065 -0.03892  0.43936  1.98403
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.8253558  0.6165090   7.827 9.13e-15 ***
## fixed.acidity        -0.0038516  0.0161395  -0.239   0.8114
## volatile.acidity     -1.0054606  0.1228569  -8.184 5.62e-16 ***
## citric.acid          -0.1706007  0.1473335  -1.158   0.2471
## residual.sugar        0.0010169  0.0131211   0.078   0.9382
## chlorides            -1.8713826  0.4655469  -4.020 6.10e-05 ***
## free.sulfur.dioxide   0.0046976  0.0021969   2.138   0.0327 *
## total.sulfur.dioxide -0.0036758  0.0007579  -4.850 1.36e-06 ***
## pH                   -0.6256087  0.1592985  -3.927 8.96e-05 ***
## sulphates             1.1921196  0.1256708   9.486  < 2e-16 ***
## alcohol               0.2848738  0.0174290  16.345  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6422 on 1569 degrees of freedom
## Multiple R-squared:  0.3681, Adjusted R-squared:  0.3641
## F-statistic: 91.42 on 10 and 1569 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol, data = whiteWine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4781 -0.6183 -0.0096  0.4848  3.2246
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.4372219  0.3550531  18.130  < 2e-16 ***
## fixed.acidity        -0.0802014  0.0165121  -4.857 1.23e-06 ***
## volatile.acidity      1.8181608 13.3623872   0.136   0.8918
## citric.acid           0.1853857  0.1074441   1.725   0.0845 .
## residual.sugar       -0.0055493  0.0027058  -2.051   0.0403 *
## chlorides            -7.3383460  0.5799321 -12.654  < 2e-16 ***
## free.sulfur.dioxide   0.0083597  0.0009457   8.840  < 2e-16 ***
## total.sulfur.dioxide -0.0043126  0.0003957 -10.899  < 2e-16 ***
## pH                    0.2247295  0.0909644   2.471   0.0135 *
## sulphates             0.4725223  0.1068933   4.421 1.01e-05 ***
## alcohol              -3.0608468 13.3566929  -0.229   0.8188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8288 on 4873 degrees of freedom
## Multiple R-squared:  0.1207, Adjusted R-squared:  0.1189
## F-statistic: 66.91 on 10 and 4873 DF,  p-value: < 2.2e-16
```

References

Data Set: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

Paper: https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377?via%3Dihub

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.