Red Wine Uncorked: How Chemistry Shapes Quality

Cynthia Hampton

2025-05-12

# Introduction

As representatives of Winebow, a U.S.-based importer and distributor with a deep-rooted passion for fine wine and spirits, we are committed to delivering an exceptional experience to our customers. Our mission extends beyond distribution—we aim to curate a thoughtful selection that highlights the craftsmanship of our producers while enhancing their global exposure and reach. In the world of wine, quality is a concept shaped not only by tradition and terroir but also by science. A wine’s chemical composition plays a crucial role in defining its aroma, flavor, texture, and overall appeal. Understanding these chemical underpinnings allows producers, distributors, and connoisseurs alike to better evaluate and appreciate wine at every stage—from vineyard to glass. This analysis explores how the chemical composition of red wine influences its overall quality.

# Data Description

The data used in this study come from red wine samples of the Portuguese “Vinho Verde” variety. These samples include physicochemical inputs, which measure the physical and chemical properties that influence a wine’s quality, composition, and behavior during production, aging, and consumption. Additionally, a sensory variable is included to assess the wine’s quality on a scale from 0 to 10. Multiple tests were conducted to explore the relationships between these physicochemical properties and the sensory quality score. For this project, we randomly selected several key properties to examine their influence and potential for optimization. The selected variables includes: Fixed Acidity refers to the wine’s organic acids that do not easily evaporate and contribute to the overall taste profile. Residual Sugar is the amount of sugar remaining after fermentation, which affects the sweetness and balance of the wine. Total Sulfur Dioxide is the combined amount of free and bound sulfur dioxide in the wine. Free sulfur dioxide (SO₂) is an additive not chemically bound to other compounds and is used for its antioxidant and preservative properties. Bound sulfur dioxide is chemically combined with other elements such as sugar or acetaldehyde. Citric Acid is a minor organic acid that influences the wine’s taste. In small doses, it adds freshness and a fruity aroma, while higher concentrations can result in a tart or sour taste. Alcohol Content indicates the amount of alcohol produced during fermentation when yeast converts sugars. Generally, lower alcohol content (below 10%) results in a sweeter wine, while higher alcohol content (above 10%) tends to produce a stronger, more pungent taste. This investigation aims to understand how the composition of these variables impacts the overall quality of red wine.

## Setting up R file

## R Markdown

Association: Between Two Quantitative Variables Y: Quality X: Wine Composition metrics Fixed Acidity Residual Sugar Total Sulfur Dioxide Citric Acid Alcohol

## Loading file and checking dataset

#Loading file  
getwd()

## [1] "/Users/cynbandz"

list.files()

## [1] "01 MatPlotLib simple plot.ipynb"   
## [2] "ch09a Financial Crisis.ipynb"   
## [3] "Chat 9 Plotting.ipynb"   
## [4] "Desktop"   
## [5] "Documents"   
## [6] "Downloads"   
## [7] "Hampton\_Assignment1102.ipynb"   
## [8] "Hampton\_Crime\_and\_Incarceration(DataClean).ipynb"   
## [9] "Hampton\_Red Wine Quality.Rmd"   
## [10] "Hampton\_Red-Wine-Quality.docx"   
## [11] "Hampton\_Red-Wine-Quality.Rmd"   
## [12] "Hampton\_STA9750 Assignment1.pdf"   
## [13] "Library"   
## [14] "Median Asking Price CH.csv"   
## [15] "Median Asking Rent.csv"   
## [16] "Median Rent Panel Dashboard.ipynb"   
## [17] "Movies"   
## [18] "MTA Delay.ipynb"   
## [19] "mta\_data.csv"   
## [20] "MTA\_Subway\_Train\_Delays.csv"   
## [21] "MTA\_Subway\_Trains\_Delayed\_\_Beginning\_2020\_20250226.csv"   
## [22] "Music"   
## [23] "NYPD\_Arrest\_Data\_\_Year\_to\_Date\_ - NYPD\_Arrest\_Data\_\_Year\_to\_Date\_.csv"  
## [24] "Pictures"   
## [25] "Project Vis.ipynb"   
## [26] "Public"   
## [27] "Red Wine Quality .csv"   
## [28] "spx(1).csv"

wine <-read.csv('Red Wine Quality .csv')  
  
#Checking dataset  
head(wine)

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

names(wine)

## [1] "fixed.acidity" "volatile.acidity" "citric.acid"   
## [4] "residual.sugar" "chlorides" "free.sulfur.dioxide"   
## [7] "total.sulfur.dioxide" "density" "pH"   
## [10] "sulphates" "alcohol" "quality"

str(wine)

## 'data.frame': 1599 obs. of 12 variables:  
## $ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...  
## $ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...  
## $ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...  
## $ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...  
## $ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...  
## $ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 17 ...  
## $ total.sulfur.dioxide: num 34 67 54 60 34 40 59 21 18 102 ...  
## $ density : num 0.998 0.997 0.997 0.998 0.998 ...  
## $ pH : num 3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...  
## $ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...  
## $ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...  
## $ quality : int 5 5 5 6 5 5 5 7 7 5 ...

#Checking for missing values  
sum(is.na(wine))

## [1] 0

colSums(is.na(wine))

## fixed.acidity volatile.acidity citric.acid   
## 0 0 0   
## residual.sugar chlorides free.sulfur.dioxide   
## 0 0 0   
## total.sulfur.dioxide density pH   
## 0 0 0   
## sulphates alcohol quality   
## 0 0 0

#Checking for duplicates  
sum(duplicated(wine))

## [1] 240

#Checking Correlations with quality  
all\_correlations(wine, interest = "quality", sorted = "magnitude")

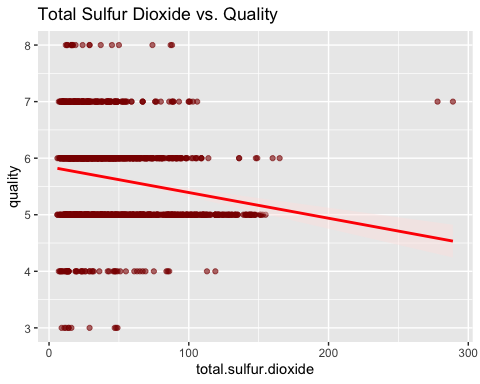
## var1 var2 correlation pval  
## 1 alcohol quality 0.47616632 2.831477e-91  
## 2 volatile.acidity quality -0.39055778 2.051715e-59  
## 3 sulphates quality 0.25139708 1.802088e-24  
## 4 citric.acid quality 0.22637251 4.991295e-20  
## 5 total.sulfur.dioxide quality -0.18510029 8.621703e-14  
## 6 density quality -0.17491923 1.874957e-12  
## 7 chlorides quality -0.12890656 2.313383e-07  
## 8 fixed.acidity quality 0.12405165 6.495635e-07  
## 9 pH quality -0.05773139 2.096278e-02  
## 10 free.sulfur.dioxide quality -0.05065606 4.283398e-02  
## 11 residual.sugar quality 0.01373164 5.832180e-01

# Variables for Association Analysis

###Total Sulfur Dioxide and Quality

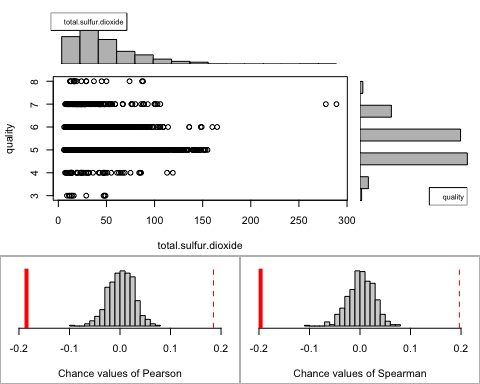
#Plot Graph   
ggplot(wine, aes(x = total.sulfur.dioxide, y = quality)) +   
 geom\_point(color = "darkred", alpha = 0.6) +   
 geom\_smooth(method = "lm", color = "red", fill = "mistyrose", se = TRUE) +   
 ggtitle("Total Sulfur Dioxide vs. Quality")

## `geom\_smooth()` using formula = 'y ~ x'



#Checking for Association   
associate(quality~total.sulfur.dioxide, data= wine)

## Association between total.sulfur.dioxide (numerical) and quality (numerical)  
## using 1599 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r -0.1851003 0  
## Spearman's rank correlation -0.1967351 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

#Checking for Correlation   
all\_correlations(wine, interest = "total.sulfur.dioxide", sorted= "magnitude")

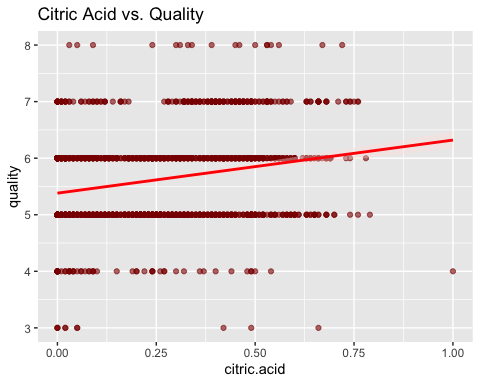
## var1 var2 correlation pval  
## 1 free.sulfur.dioxide total.sulfur.dioxide 0.66766645 6.404723e-207  
## 2 total.sulfur.dioxide alcohol -0.20565394 9.890520e-17  
## 3 residual.sugar total.sulfur.dioxide 0.20302788 2.449285e-16  
## 4 total.sulfur.dioxide quality -0.18510029 8.621703e-14  
## 5 fixed.acidity total.sulfur.dioxide -0.11318144 5.709033e-06  
## 6 volatile.acidity total.sulfur.dioxide 0.07647000 2.213857e-03  
## 7 total.sulfur.dioxide density 0.07126948 4.354284e-03  
## 8 total.sulfur.dioxide pH -0.06649456 7.818341e-03  
## 9 chlorides total.sulfur.dioxide 0.04740047 5.809120e-02  
## 10 total.sulfur.dioxide sulphates 0.04294684 8.601835e-02  
## 11 citric.acid total.sulfur.dioxide 0.03553302 1.555454e-01

**Analysis:**According to the Pearson correlation (r= -0.185) total sulfur dioxide and wine quality have a weak negative linear relationship. This suggests that while there is a slight negative correlation between the two variables, the relationship is weak. P- values are less than 0.05, so the correlation is statistically significant. Total sulfur dioxide serves as a preservative to prevent oxidation, but its influence on wine quality appears to be minimal compared to other variables.

###Citric Acid Quality

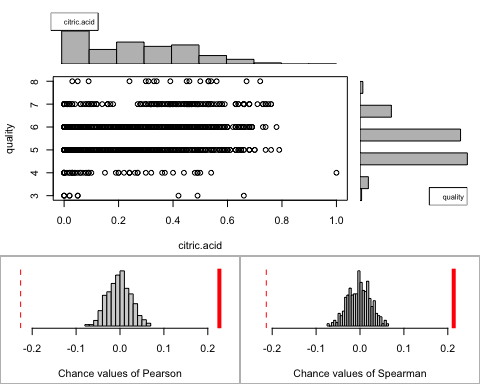
#Plot Graph   
ggplot(wine, aes(x = citric.acid, y = quality)) +   
 geom\_point(color = "darkred", alpha = 0.6) +   
 geom\_smooth(method = "lm", color = "red", fill = "mistyrose", se = TRUE) +   
 ggtitle("Citric Acid vs. Quality")

## `geom\_smooth()` using formula = 'y ~ x'



#Association  
associate(quality~citric.acid, data= wine)

## Association between citric.acid (numerical) and quality (numerical)  
## using 1599 complete cases



## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.2263725 0  
## Spearman's rank correlation 0.2134809 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

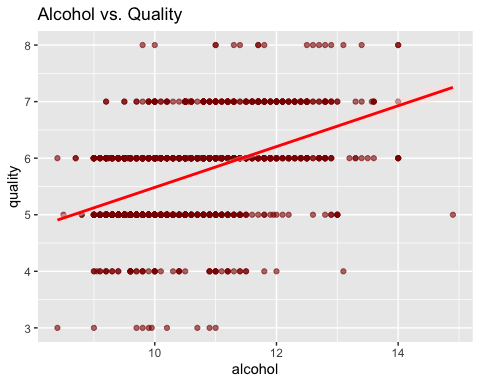
#Correlation  
all\_correlations(wine, interest = "citric.acid", sorted= "magnitude")

## var1 var2 correlation pval  
## 1 fixed.acidity citric.acid 0.67170343 2.535321e-210  
## 2 volatile.acidity citric.acid -0.55249568 1.805663e-128  
## 3 citric.acid pH -0.54190414 1.007201e-122  
## 4 citric.acid density 0.36494718 1.478795e-51  
## 5 citric.acid sulphates 0.31277004 1.265262e-37  
## 6 citric.acid quality 0.22637251 4.991295e-20  
## 7 citric.acid chlorides 0.20382291 1.863705e-16  
## 8 citric.acid residual.sugar 0.14357716 8.083723e-09  
## 9 citric.acid alcohol 0.10990325 1.059462e-05  
## 10 citric.acid free.sulfur.dioxide -0.06097813 1.473916e-02  
## 11 citric.acid total.sulfur.dioxide 0.03553302 1.555454e-01

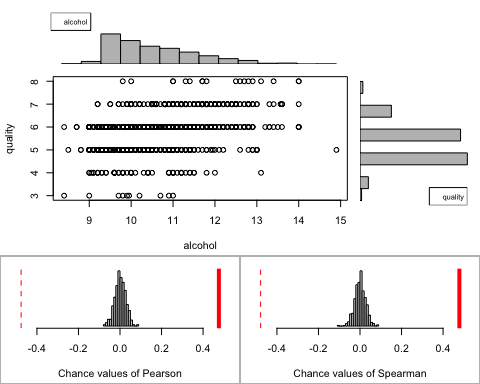
**Analysis:** Pearson’s correlation was (r= 0.2263), indicating there is a weak positive linear relationship between citric acid and quality. P-values are less than 0.05, indicating the correlations are statistically significant. However, citric acid and wine quality linear relationship isn’t meaningful. The correlation between the two variables is weak, indicating that citric acid is not a reliable predictor of wine quality. While citric acid may play a role in the taste or acidity balance of wine, its impact on overall quality appears to be minimal based on this analysis.

### Alcohol and Quality

## `geom\_smooth()` using formula = 'y ~ x'



## Association between alcohol (numerical) and quality (numerical)  
## using 1599 complete cases



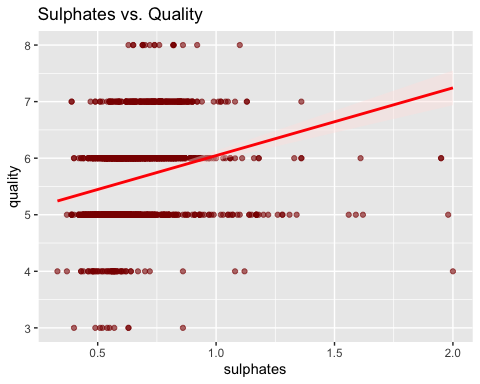
## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.4761663 0  
## Spearman's rank correlation 0.4785317 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

## var1 var2 correlation pval  
## 1 density alcohol -0.49617977 3.938835e-100  
## 2 alcohol quality 0.47616632 2.831477e-91  
## 3 chlorides alcohol -0.22114054 3.654950e-19  
## 4 total.sulfur.dioxide alcohol -0.20565394 9.890520e-17  
## 5 pH alcohol 0.20563251 9.964498e-17  
## 6 volatile.acidity alcohol -0.20228803 3.155190e-16  
## 7 citric.acid alcohol 0.10990325 1.059462e-05  
## 8 sulphates alcohol 0.09359475 1.783053e-04  
## 9 free.sulfur.dioxide alcohol -0.06940835 5.492314e-03  
## 10 fixed.acidity alcohol -0.06166827 1.364868e-02  
## 11 residual.sugar alcohol 0.04207544 9.258425e-02

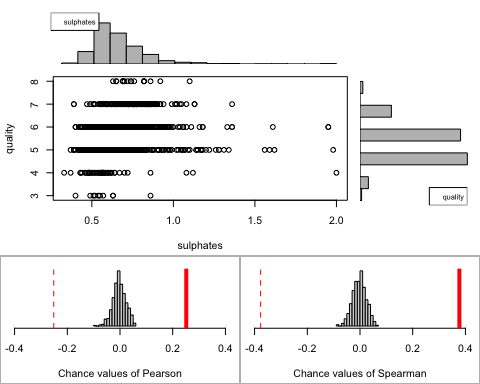
**Analysis:** The Pearson correlation (r= 0.4761), there is a moderate linear relationship between the quality of wine and alcohol. P-values are less than 0.05, so the correlations are statistically significant between the two variables.During the fermentation period, producers may allow longer fermentation for higher-quality wines or certain styles, such as bold reds, which naturally result in higher alcohol content.

### Sulphates and Quality

## `geom\_smooth()` using formula = 'y ~ x'



## Association between sulphates (numerical) and quality (numerical)  
## using 1599 complete cases



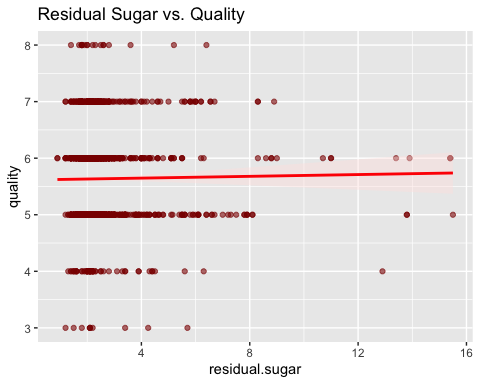
## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.2513971 0  
## Spearman's rank correlation 0.3770602 0  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0 and 0.007   
## the p-value of Spearman's rank correlation is between 0 and 0.007   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

## var1 var2 correlation pval  
## 1 chlorides sulphates 0.371260481 1.986310e-53  
## 2 citric.acid sulphates 0.312770044 1.265262e-37  
## 3 volatile.acidity sulphates -0.260986685 2.606926e-26  
## 4 sulphates quality 0.251397079 1.802088e-24  
## 5 pH sulphates -0.196647602 2.106734e-15  
## 6 fixed.acidity sulphates 0.183005664 1.648652e-13  
## 7 density sulphates 0.148506412 2.418474e-09  
## 8 sulphates alcohol 0.093594750 1.783053e-04  
## 9 free.sulfur.dioxide sulphates 0.051657572 3.888321e-02  
## 10 total.sulfur.dioxide sulphates 0.042946836 8.601835e-02  
## 11 residual.sugar sulphates 0.005527121 8.252134e-01

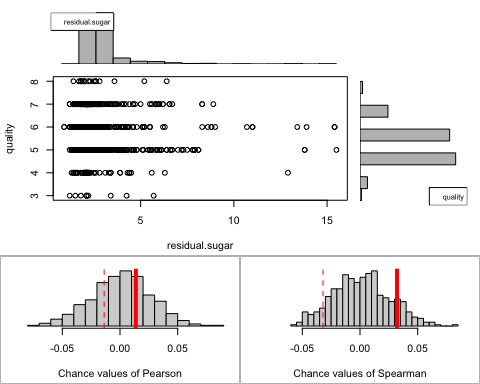
**Analysis:**According to the Pearson correlation (r = 0.25), there is a moderate positive linear relationship between the quality of wine and sulphates. As sulphates increase, the quality of wine tends to increase slightly, but it is not a very strong correlation. P-values are lower than 0.007 for both quality and sulphates, so the correlations are significant between the two and sulphates are meaningfully associated with the quality of wine, even if the linear relationship is weaker. Sulphates appear to correlate strongly with other variables such as citric acid, suggesting that sulphates is more present in the chemical composition of the wine rather than the quality.

### Residual Sugar and Quality

## `geom\_smooth()` using formula = 'y ~ x'



## Association between residual.sugar (numerical) and quality (numerical)  
## using 1599 complete cases

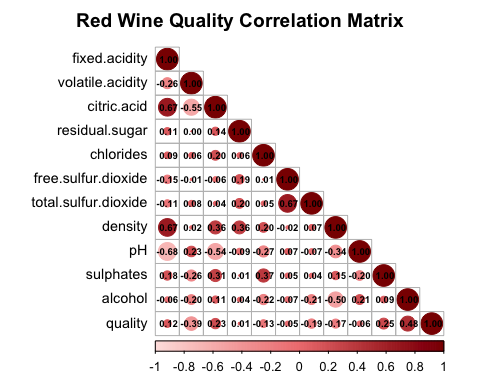


## Permutation procedure:  
## Value Estimated p-value  
## Pearson's r 0.01373164 0.576  
## Spearman's rank correlation 0.03204817 0.216  
## With 500 permutations, we are 95% confident that:  
## the p-value of Pearson's correlation (r) is between 0.531 and 0.62   
## the p-value of Spearman's rank correlation is between 0.181 and 0.255   
## Note: If 0.05 is in this range, increase the permutations= argument.  
##   
##   
##   
## Advice: If stream of points is well described by an ellipse, use Pearson's r.  
## Otherwise, as long as stream is monotonic, use Spearman's rank correlation  
## or try logs, e.g. associate( log10(y)~log10(x) )

## var1 var2 correlation pval  
## 1 residual.sugar density 0.355283371 9.013042e-49  
## 2 residual.sugar total.sulfur.dioxide 0.203027882 2.449285e-16  
## 3 residual.sugar free.sulfur.dioxide 0.187048995 4.684735e-14  
## 4 citric.acid residual.sugar 0.143577162 8.083723e-09  
## 5 fixed.acidity residual.sugar 0.114776724 4.199465e-06  
## 6 residual.sugar pH -0.085652422 6.065915e-04  
## 7 residual.sugar chlorides 0.055609535 2.617079e-02  
## 8 residual.sugar alcohol 0.042075437 9.258425e-02  
## 9 residual.sugar quality 0.013731637 5.832180e-01  
## 10 residual.sugar sulphates 0.005527121 8.252134e-01  
## 11 volatile.acidity residual.sugar 0.001917882 9.389168e-01

**Analysis:**According to the Pearson correlation (r = 0.0137), there is nearly no linear relationship between the quality of wine and residual sugar. P-values are higher than 0.05, so the correlations are not statistically significant. We fail to reject the null hypothesis, meaning residual sugar is not associated with the quality of wine. In the graph, there is no visible pattern and the data points are concentrated at low sugar levels, with no clear slope. Residual sugar has some correlation with density, as sugar adds weight to the wine. Overall, there is no meaningful correlation between the quality of wine and residual sugar.

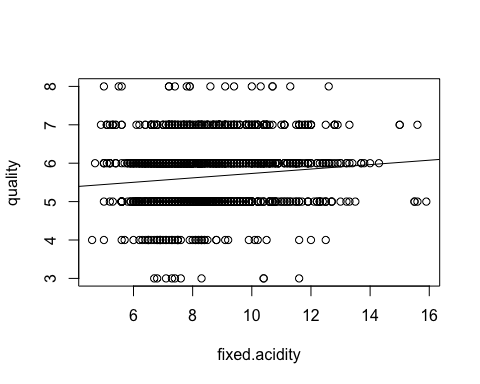
### Correlations Visualization

 **Analysis:** This correlation matrix creates a visualization of the linear relationships between the quality of wine and other variables in the chemical properties. The darker red shows a stronger positive correlation, and the lighter colors show weaker correlations. Negative correlations are shown in lighter colors. The bigger circles show stronger correlation, whether it is positive or negative, and smaller circles show weak or no correlation. Alcohol appears to have the strongest positive correlations with quality at 0.48, meanwhile volatile acidity shows a weak negative correlation with quality at -0.26. Residual sugar and chlorides show little or no correlation with quality. Citric acid and fixed acidity are strongly correlated with each other, and weakly correlated with pH. Residual sugar, chlorides, and sulphates have either little or no direct relationship with quality of wine.

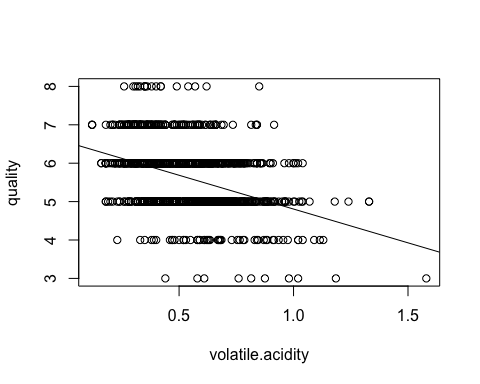
# Multi Linear Regression

##   
## The downloaded binary packages are in  
## /var/folders/lp/k23vsrvn1qvfj3\_d27962w8r0000gn/T//RtmpDMOvKs/downloaded\_packages

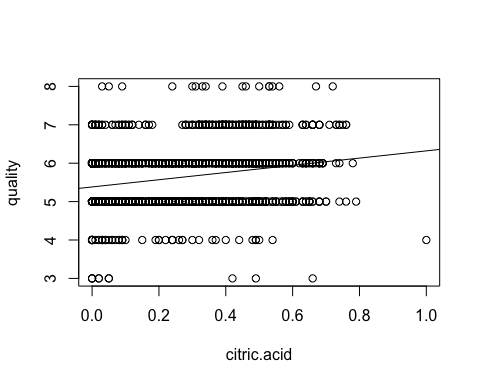
## [1] "fixed.acidity"



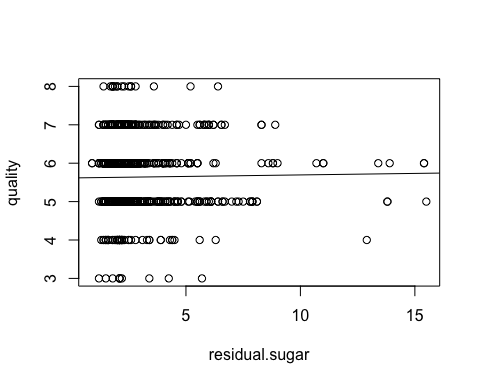
##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8248 -0.6061 0.1925 0.4341 2.5550   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.15732 0.09789 52.684 < 2e-16 \*\*\*  
## fixed.acidity 0.05754 0.01152 4.996 6.5e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8016 on 1597 degrees of freedom  
## Multiple R-squared: 0.01539, Adjusted R-squared: 0.01477   
## F-statistic: 24.96 on 1 and 1597 DF, p-value: 6.496e-07



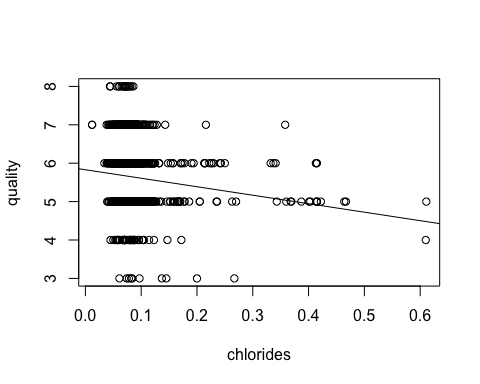
##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.79071 -0.54411 -0.00687 0.47350 2.93148   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.56575 0.05791 113.39 <2e-16 \*\*\*  
## volatile.acidity -1.76144 0.10389 -16.95 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7437 on 1597 degrees of freedom  
## Multiple R-squared: 0.1525, Adjusted R-squared: 0.152   
## F-statistic: 287.4 on 1 and 1597 DF, p-value: < 2.2e-16



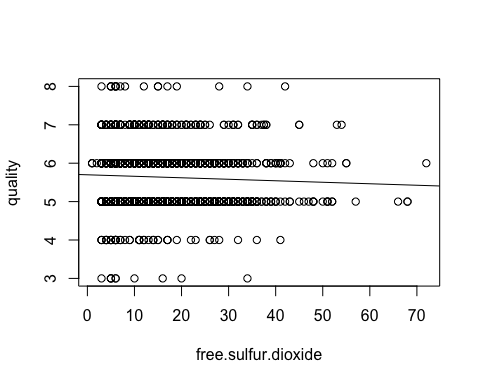
##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0011 -0.5976 0.1021 0.5057 2.5901   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.38172 0.03372 159.610 <2e-16 \*\*\*  
## citric.acid 0.93845 0.10104 9.288 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7869 on 1597 degrees of freedom  
## Multiple R-squared: 0.05124, Adjusted R-squared: 0.05065   
## F-statistic: 86.26 on 1 and 1597 DF, p-value: < 2.2e-16



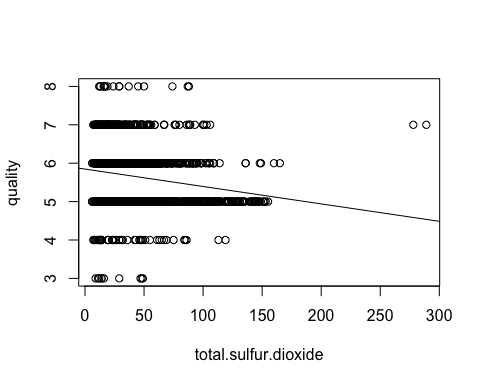
##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6609 -0.6334 0.3580 0.3690 2.3729   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.616055 0.041616 134.950 <2e-16 \*\*\*  
## residual.sugar 0.007865 0.014331 0.549 0.583   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8077 on 1597 degrees of freedom  
## Multiple R-squared: 0.0001886, Adjusted R-squared: -0.0004375   
## F-statistic: 0.3012 on 1 and 1597 DF, p-value: 0.5832



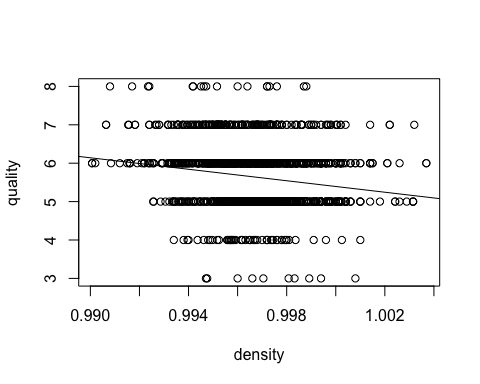
##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6946 -0.6503 0.3010 0.3607 2.3607   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.82948 0.04229 137.852 < 2e-16 \*\*\*  
## chlorides -2.21184 0.42578 -5.195 2.31e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8011 on 1597 degrees of freedom  
## Multiple R-squared: 0.01662, Adjusted R-squared: 0.016   
## F-statistic: 26.99 on 1 and 1597 DF, p-value: 2.313e-07



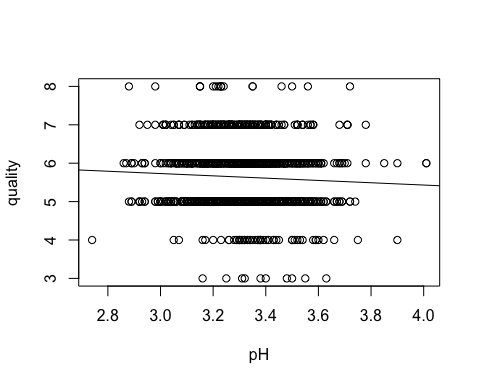
##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6864 -0.6394 0.3215 0.3762 2.4661   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.698107 0.036678 155.357 <2e-16 \*\*\*  
## free.sulfur.dioxide -0.003911 0.001929 -2.027 0.0428 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8068 on 1597 degrees of freedom  
## Multiple R-squared: 0.002566, Adjusted R-squared: 0.001941   
## F-statistic: 4.109 on 1 and 1597 DF, p-value: 0.04283



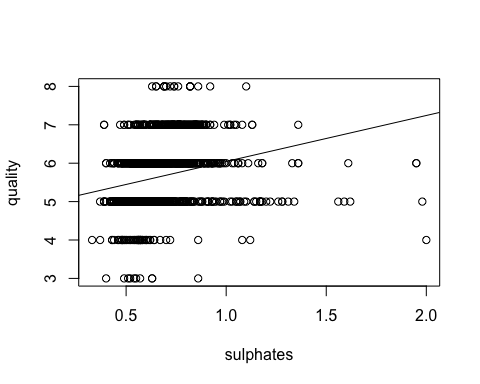
##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8063 -0.6336 0.2164 0.3800 2.5527   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.8471792 0.0343670 170.140 < 2e-16 \*\*\*  
## total.sulfur.dioxide -0.0045442 0.0006037 -7.527 8.62e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7939 on 1597 degrees of freedom  
## Multiple R-squared: 0.03426, Adjusted R-squared: 0.03366   
## F-statistic: 56.66 on 1 and 1597 DF, p-value: 8.622e-14



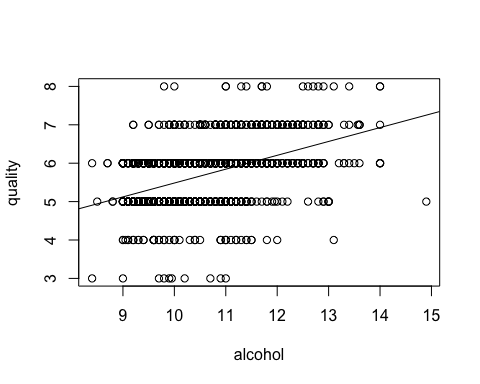
##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7885 -0.6216 0.1554 0.4271 2.5177   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 80.24 10.51 7.636 3.83e-14 \*\*\*  
## density -74.85 10.54 -7.100 1.87e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7954 on 1597 degrees of freedom  
## Multiple R-squared: 0.0306, Adjusted R-squared: 0.02999   
## F-statistic: 50.41 on 1 and 1597 DF, p-value: 1.875e-12



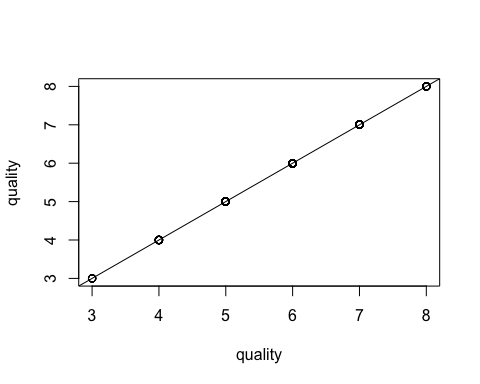
##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6817 -0.6394 0.3032 0.3878 2.4874   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.6359 0.4332 15.320 <2e-16 \*\*\*  
## pH -0.3020 0.1307 -2.311 0.021 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8065 on 1597 degrees of freedom  
## Multiple R-squared: 0.003333, Adjusted R-squared: 0.002709   
## F-statistic: 5.34 on 1 and 1597 DF, p-value: 0.02096



##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.2432 -0.5424 0.1102 0.4456 2.3977   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.84775 0.07842 61.82 <2e-16 \*\*\*  
## sulphates 1.19771 0.11539 10.38 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7819 on 1597 degrees of freedom  
## Multiple R-squared: 0.0632, Adjusted R-squared: 0.06261   
## F-statistic: 107.7 on 1 and 1597 DF, p-value: < 2.2e-16



##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8442 -0.4112 -0.1690 0.5166 2.5888   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.87497 0.17471 10.73 <2e-16 \*\*\*  
## alcohol 0.36084 0.01668 21.64 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7104 on 1597 degrees of freedom  
## Multiple R-squared: 0.2267, Adjusted R-squared: 0.2263   
## F-statistic: 468.3 on 1 and 1597 DF, p-value: < 2.2e-16

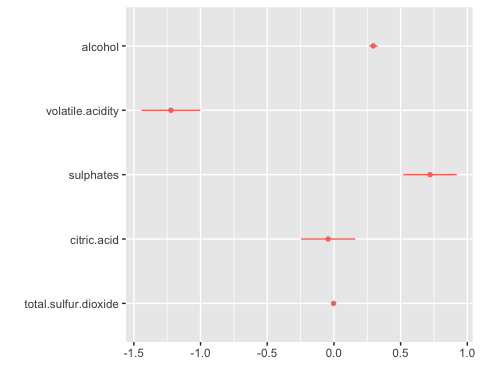


##   
## Call:  
## lm(formula = quality ~ value, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.044e-12 3.800e-16 3.800e-16 1.040e-15 2.657e-14   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.862e-13 4.613e-15 -4.037e+01 <2e-16 \*\*\*  
## quality 1.000e+00 8.102e-16 1.234e+15 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.615e-14 on 1597 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 1.524e+30 on 1 and 1597 DF, p-value: < 2.2e-16

## [1] 0.01362

## [1] 0.99007

##   
## Call:  
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates +   
## citric.acid + total.sulfur.dioxide, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.72463 -0.38380 -0.06689 0.44606 2.14550   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.8431068 0.2050732 13.864 < 2e-16 \*\*\*  
## alcohol 0.2953419 0.0160375 18.416 < 2e-16 \*\*\*  
## volatile.acidity -1.2223102 0.1124774 -10.867 < 2e-16 \*\*\*  
## sulphates 0.7207881 0.1027039 7.018 3.32e-12 \*\*\*  
## citric.acid -0.0427246 0.1035810 -0.412 0.68   
## total.sulfur.dioxide -0.0022182 0.0005126 -4.327 1.60e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6552 on 1593 degrees of freedom  
## Multiple R-squared: 0.3439, Adjusted R-squared: 0.3418   
## F-statistic: 167 on 5 and 1593 DF, p-value: < 2.2e-16

 **Analysis:**Of the coefficients all are statistically significant except for citric acid using a standard 95% confidence interval (p <.05). From the significant coefficients two of them are positive and two are negative with alcohol level and sulfates both providing a positive impact though the model shows that sulphates have a higher impact on the wines quality. From this wine makers could conclude that focusing on a wine that emphasizes sulphates over one that does not would result in a better end product. On the other hand, volatile acidity has a very negative impact on wine quality. The presence of excess citric acid causes an undesirable taste in the wine and because this coefficient is high it shows that it has a large impact on how wine is made and should be considered for wine makers Total sulfur dioxide value is low indicating that even though it is statistically significant the impact that it has is negligible. This shows that sulfur dioxide may be unnoticeable to people who are drinking the wine and is not something that may need to be considered for wine makers. The R squared value for this model indicates that it explains an adequate amount of the variance but there may be other factors that contribute to a wines quality outside of what is in the dataset currently such as vineyard or vintage.

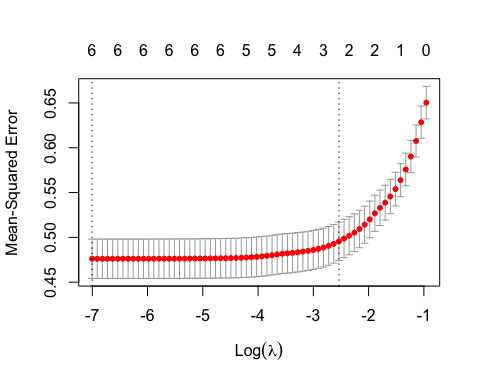
# Lasso Regression

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 4.1-8



## Best lambda: 0.0009089502

## 7 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) 1.965992330  
## fixed.acidity 0.017095992  
## residual.sugar -0.010543468  
## total.sulfur.dioxide -0.004043125  
## free.sulfur.dioxide 0.008307316  
## citric.acid 0.696161310  
## alcohol 0.328305234

**Analysis:** In order to have a better understanding of the factors influencing red wine quality, we utilized a Lasso Regression model. Using the Lasso Regression was helpful in predicting the outcomes and performing variable selection, notably by shrinking the less significant coefficients closer to zero. By using cross-validation to determine the lambda value, we found that it was approximately 0.0009. The mean squared error was minimized at this value, which shows that there was a food balance with the model accuracy. The results of this regression show that all the selected variables contributed to the prediction, and none of the variables were eliminated. Citric acid had the strongest positive influence on wine quality, suggesting that wines with higher amounts of citric acid have higher ratings. Alcohol also has a strong positive relationship with quality, which supports the perception that higher levels of alcohol enhance the quality and flavor of wine. Residual sugar and total sulfur dioxide had negative coefficients which implies that increasing these variables could have a negative effect on the quality of wine. Fixed acidity and free sulfur dioxide showed small positive effects, implying there is a small but slightly significant impact on quality. Overall, the Lasso Regression reiterated the discoveries from the other tests we conducted in regards to the impact of the chosen variables in comparison to wine quality.

# Conclusion

In consideration with all of our conducted tests, we conclude that alcohol, sulphates, and citric acid have the strongest positive correlations with quality amongst all the variables, suggesting that having higher levels of these variables can have a positive effect on wine. These variables had significant associations and coefficients as shown in the correlation matrix and the multi linear regression and lasso regression. On the other hand, residual sugar and total sulfur dioxide had a negative association with quality, suggesting that having higher levels of these variables can have a negative effect on wine. Other variables such as fixed acidity, free sulfur dioxide, and volatile acidity had more weak or inconsistent relationships, showing that their effect on wine quality is not as significant. As representatives of Winebow, we can use these insights to confirm that wine quality is not only determined by tradition and terroir, but also by the chemical makeup. Wine producers can enhance their production process to bring out the most desirable properties of wine and distributors can better position wine products to consumers. Ultimately, this analysis creates a connection between wine craftsmanship and science, supporting our mission to elevate wine experiences through data-driven selection.