White Wine Quality Analysis

##### This report explores a dataset containing the 12 attributes for approximately 4,898 white wine samples. 11 variables describe the chemical properties of the wine. 1 variable is an output variable based on the sensory perception of the wines.

##### Below is the description for each of the 12 attributes. This description was directly obtained from the dataset provided by Udacity.

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3 - citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines

4 - residual sugar: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5 - chlorides: the amount of salt in the wine

6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7 - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content

9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant

11 - alcohol: the percent alcohol content of the wine

Output variable (based on sensory data): 12 - quality (score between 0 and 10)

### The data was obtained from the following links:

<https://classroom.udacity.com>

<https://docs.google.com/document/d/1qEcwltBMlRYZT-l699-71TzInWfk4W9q5rTCSvDVMpc/pub?embedded=true>

# Univariate Plots Section

wines <- read.csv ("wineQualityWhites.csv", stringsAsFactors = FALSE)  
  
summary(wines)

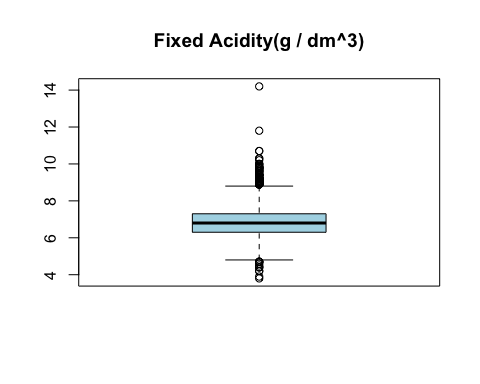
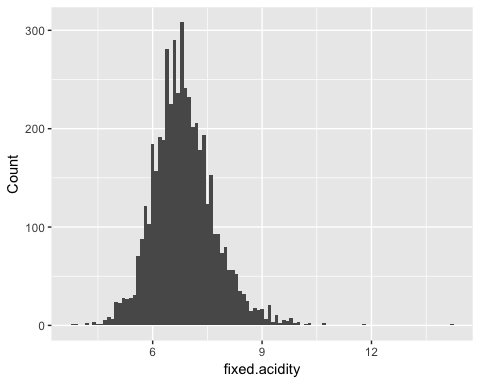
## X fixed.acidity volatile.acidity citric.acid   
## Min. : 1 Min. : 3.800 Min. :0.0800 Min. :0.0000   
## 1st Qu.:1225 1st Qu.: 6.300 1st Qu.:0.2100 1st Qu.:0.2700   
## Median :2450 Median : 6.800 Median :0.2600 Median :0.3200   
## Mean :2450 Mean : 6.855 Mean :0.2782 Mean :0.3342   
## 3rd Qu.:3674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

str(wines)

## 'data.frame': 4898 obs. of 13 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ fixed.acidity : num 7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...  
## $ volatile.acidity : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...  
## $ citric.acid : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...  
## $ residual.sugar : num 20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...  
## $ chlorides : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...  
## $ free.sulfur.dioxide : num 45 14 30 47 47 30 30 45 14 28 ...  
## $ total.sulfur.dioxide: num 170 132 97 186 186 97 136 170 132 129 ...  
## $ density : num 1.001 0.994 0.995 0.996 0.996 ...  
## $ pH : num 3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...  
## $ sulphates : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...  
## $ alcohol : num 8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...  
## $ quality : int 6 6 6 6 6 6 6 6 6 6 ...

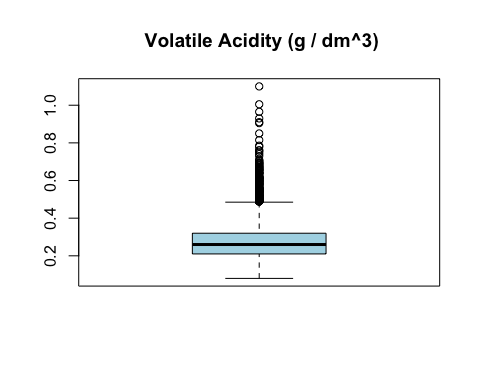
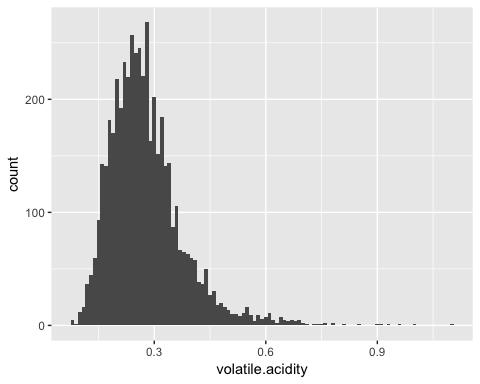
For the following univariate plots, histograms and box plots are used to analyze the 12 variables and 4898 observations.

FIXED ACIDITY VS COUNT



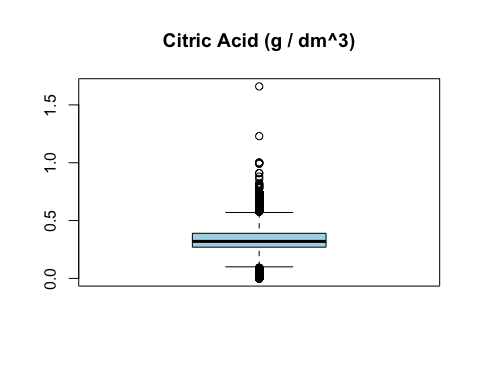
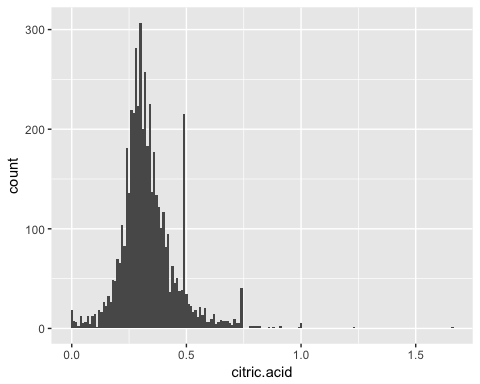
The mean is 6.855 and the median is 6.8 for the fixed acidity variable. These two plots above show a fairly normal distribution because the median and the mean are very close together.

ACIDITY(VOLATILITY) VS COUNT



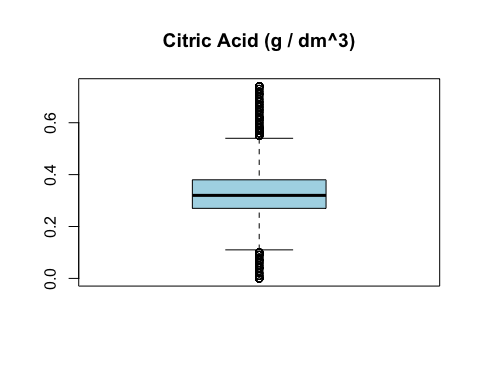
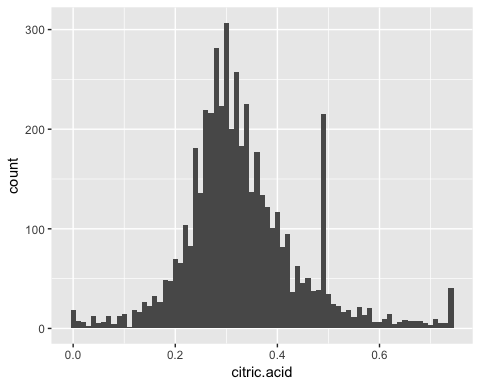
Acidity (Volatility) has a fairly normal distribution with a mean of 0.27 and a median of 0.26.

CITRIC ACID VS COUNT



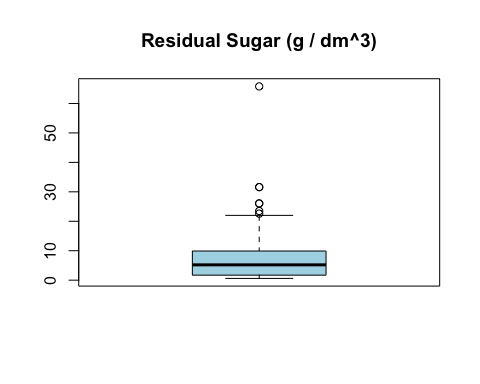
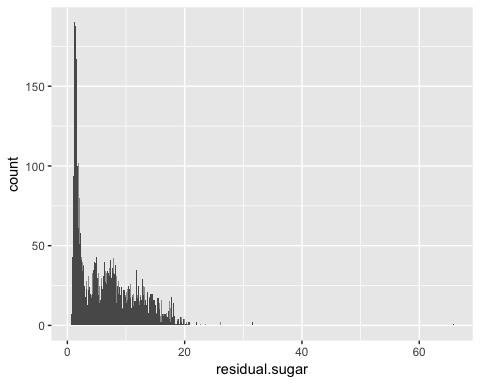
Because citric acid contained outlier values, graphs depicting the data with citric acid level strictly less than 0.75 were created below.

CITRIC ACID VS COUNT (Trimmed Version)



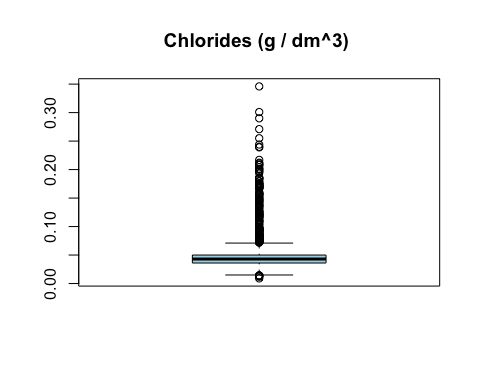
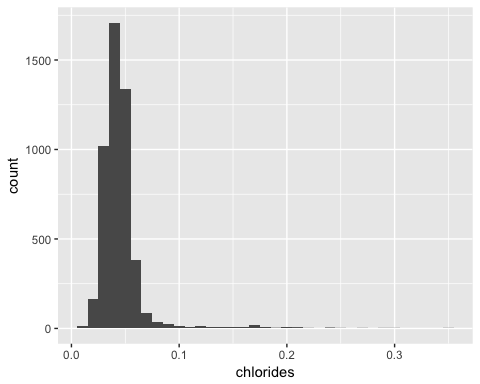
Before trimming the data for citric acid, the mean was 0.3342 and the median was 0.32. After removing the outliers, the mean resulted as 0.3315 and the median remained the same as 0.32. The distance between the mean and the median slightly decreased after the trimming.

RESIDUAL SUGAR VS COUNT



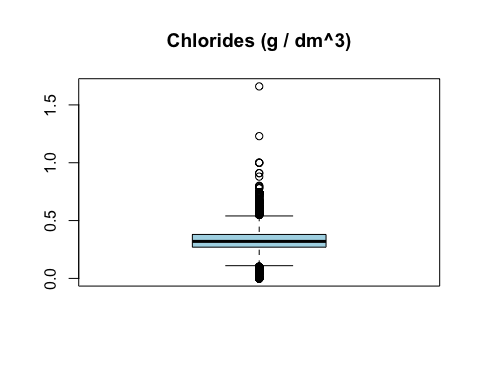
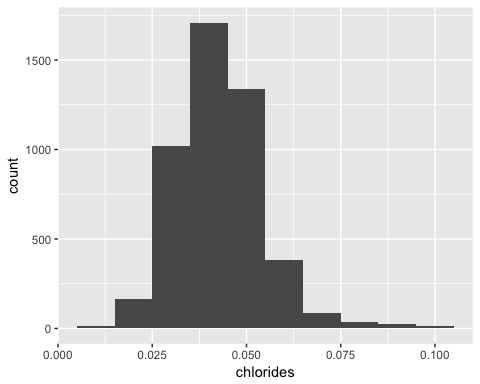
Residual sugar has a mean of 6.391 and a median of 5.200. The mean value lies to the right of the median value, indicating a skew to the right.

CHLORIDES VS COUNT



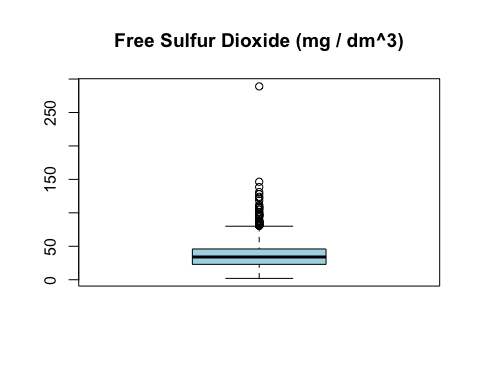
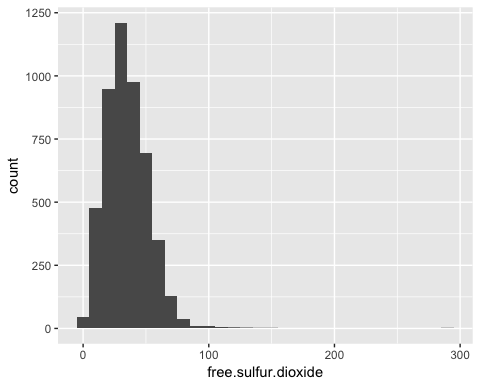
Since we observed that the plot for chlorides vs count is skewed, we want to filter out the outliers as shown below.

CHLORIDES VS COUNT (Trimmed Version)



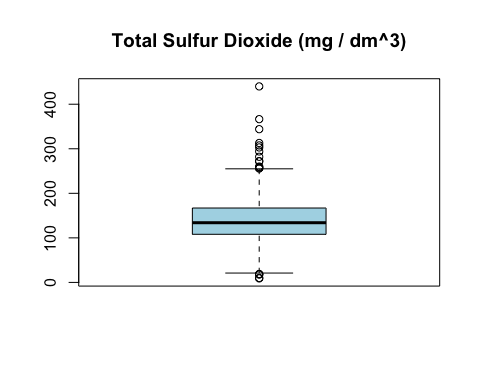
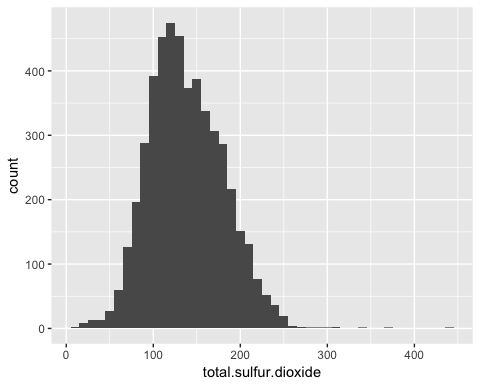
The graphs above display the data with chlorides level less than 0.1. The median is now 0.04250 and the mean is 0.04312, slightly decreasing the skew compared to the plots showing the unfiltered data.

FREE SULFER DIOXIDE VS COUNT



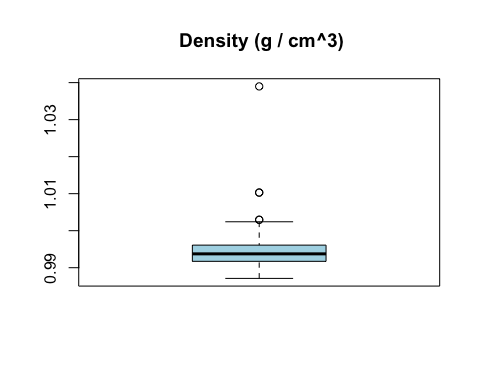
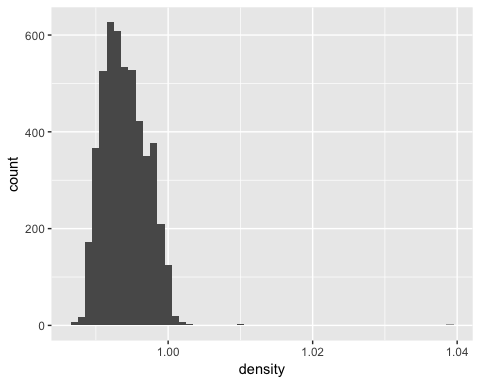
Free sulfur dioxide appears to have a slight skew to the right in its distribution.

TOTAL SULFER DIOXIDE VS COUNT



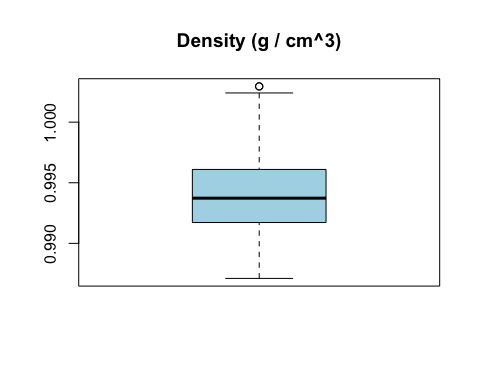
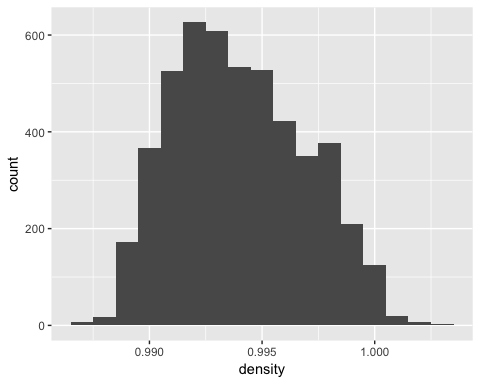
Total sulfur dioxide versus count displayed a fairly normal distribution.

DENSITY VS COUNT



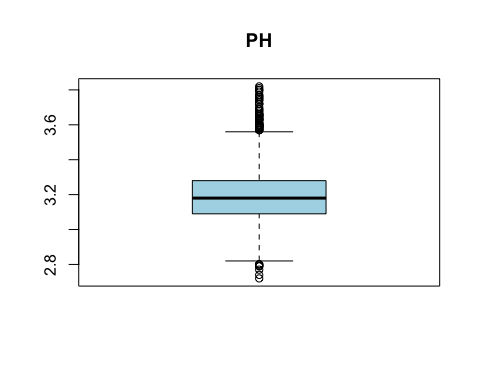
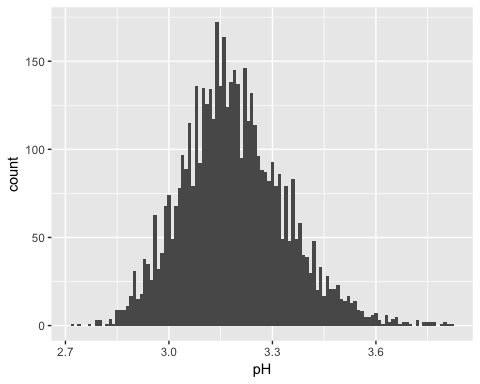
The above plots display the unfiltered data. We’re now going to remove the outliers for the density variable.

DENSITY VS COUNT (Trimmed Version)



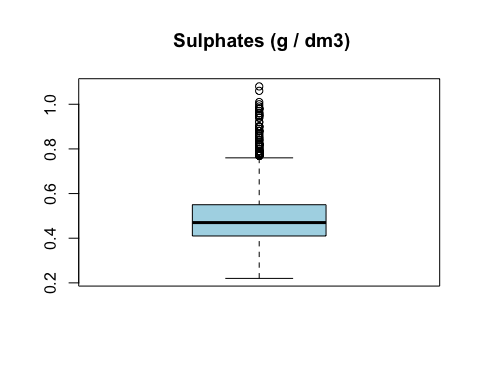
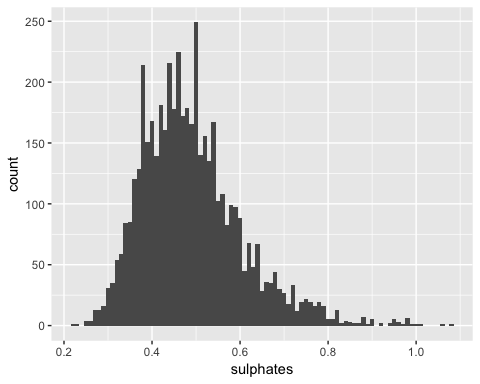
In the plots above, the data was trimmed to display density to be less than 1.01, filtering out the outliers.

PH LEVEL VS COUNT



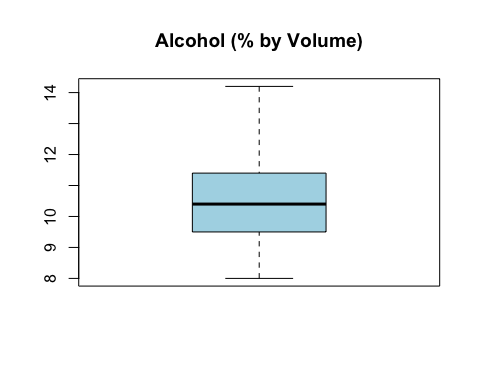
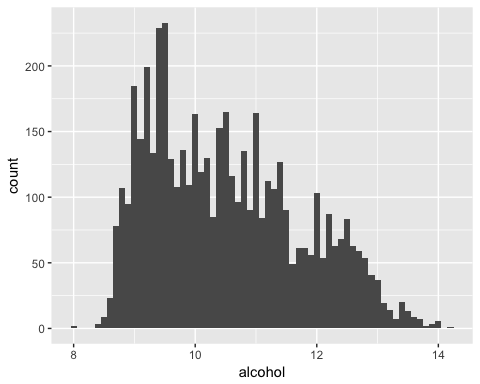
In the plot displaying the Ph Level vs Count, the mean is 3.188 and the median is 3.180. The mean and the median are extremely close, indicating a normal distribution.

SULPHATES VS COUNT



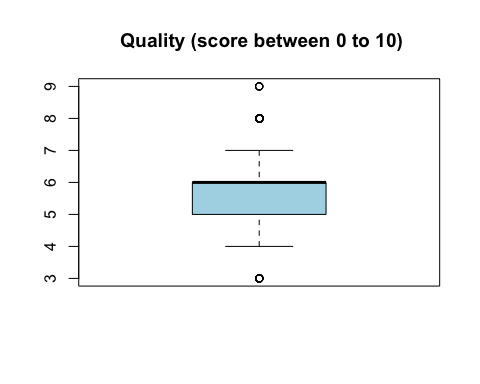
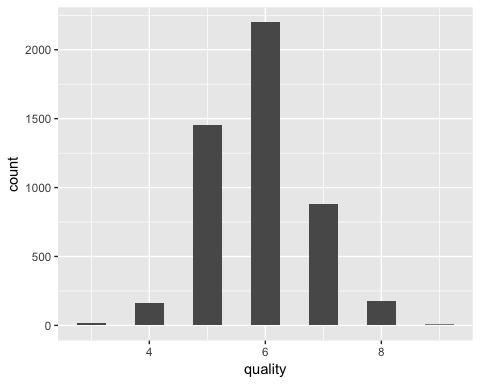
Sulphates vs Count data displays a distribution which appears to have a slight skew to the right.

ALCOHOL VS COUNT



In the plots above, different levels of alcohol are well distributed across all samples.

QUALITY LEVEL VS COUNT



Quality versus Count displays a normal distribution. Most observations are concentrated near the quality level of 6.

## Univariate Analysis

Our dataset consisted of 4898 samples of white wines with 12 characteristics: fixed.acidity, volatile, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, and quality.

While some attributes may quantify the chemical properties of each wine, attributes such as citric acid, residual sugar, alcohol, volatile acidity, and quality may have greater impact on the overall taste.

citric acid - freshness and flavor

residual sugar - sweetness level

alcohol - alcohol level

volatile acidity - pleasant taste

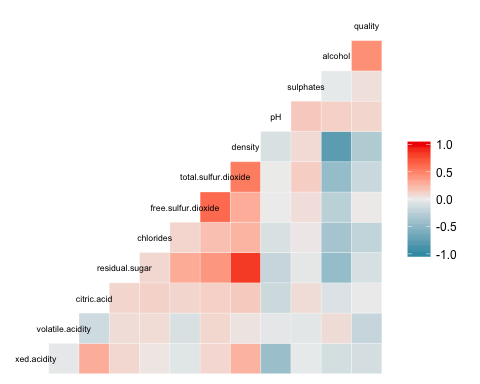
quality - overall quality level of the wine

Citric acid, residual sugar, alcohol, and volatile acidity levels are determined by physicochemical tests, and can also affect taste. It is unclear whether the quality level is determined by physicochemical tests or is classified based on perceived sensory experiences.

When analyzing the univariate plots, outliers were trimmed to obtain better visualizations.

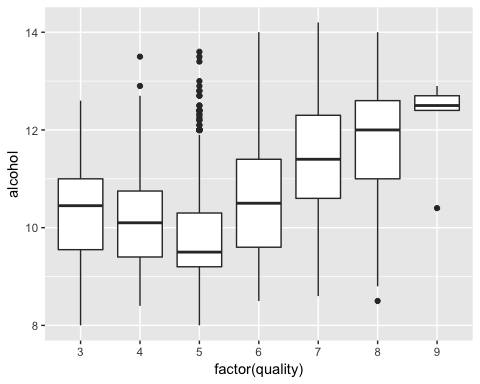
# Bivariate Plots Section

Correlation Plot



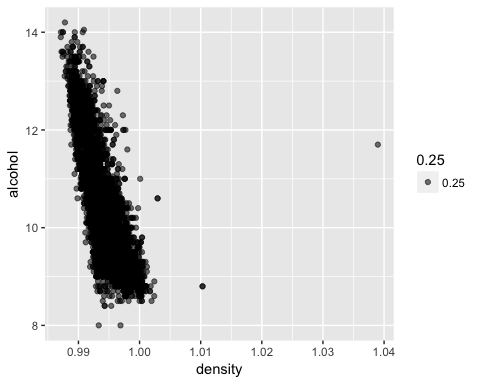
The correlation plot above helps us to see the different levels of correlation between the variables through the range of colors. Darker colors indicate stronger correlations between the variables.

Quality Versus Alcohol Level in %



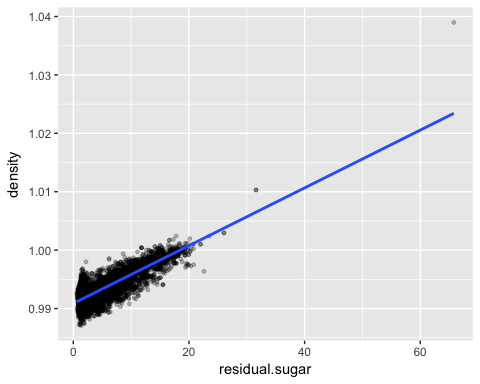
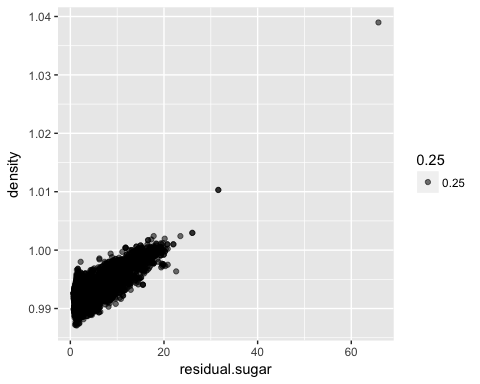
Quality and alcohol levels are important variables to consider in this wine analysis. Correlation between quality and alcohol level is 0.43557. It is showing a moderately positive correlation between these two variables. The box plots displaying the quality levels of 6 and 7 show normal distributions.

Density in g/cm^3 Versus Alcohol in %



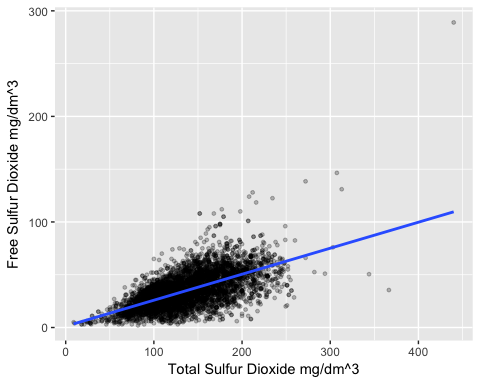
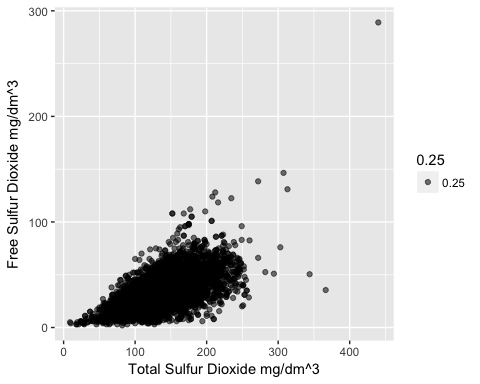
The correlation between density and alcohol is -0.78013762. As seen in the plot above, we can observe a strong negative correlation which is close to -1, indicating an inverse relationship between the density and alcohol.

Residual Sugar in (g/dm^3) Versus Density Level in (g/cm^3)



Residual sugar versus density has a correlation of 0.838966455 close to +1, –indicating a strong positive correlation between these two variables. Thus, as residual sugar level increases, density level also increases. In the plot above, outliers were not filtered. We can observe that the majority of the values fall below the residual sugar level of 30 and below the density level of 1.01.

Free Versus Total Sulfur Dioxide in (mg/dm^3)



## [1] "Below is the correlation coefficient."

## [1] 0.615501

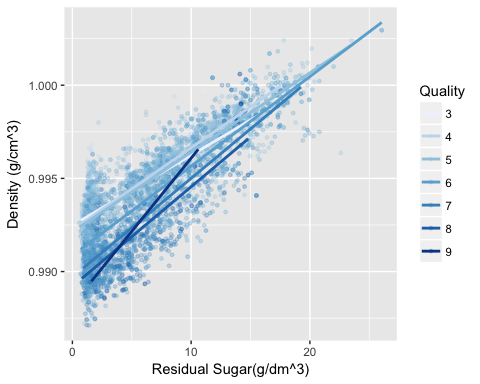
A scatter plot was first created when analyzing the relationship between the free versus the total sulfur dioxide levels. A fitted regression line was then added as shown in the second plot. The data points are moderately clustered around the regression line. The correlation coefficient is 0.615501 indicating a fairly strong positive correlation.

## Bivariate Analysis

According to the correlation plot, fairly strong correlations exist in the relationships between the following variables: quality versus alcohol, residual sugar versus density, density versus alcohol, and the free versus the total sulfur dioxide levels. A strong negative correlation was observed between the density and alcohol levels. The relationship between the residual sugar level versus density displayed a positive correlation. Positive correlations also existed between the quality versus alcohol levels, and between the free versus the total sulfur dioxide levels.

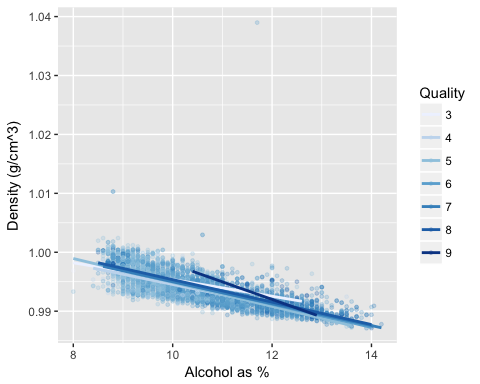
# Multivariate Plots Section

Residual Sugar(g/dm^3) vs Density(g/cm^3) and Quality



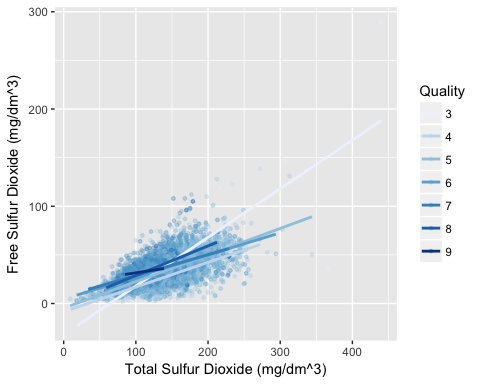
The data was trimmed removing the outliers to improve the visualization of trends. The plot above displays observations containing residual sugar level strictly less than 30. We see that as density increases, residual sugar also increases. However, quality is spread all throughout the different levels of density and residual sugar.

Alcohol as % vs Density (g/cm^3) vs Quality



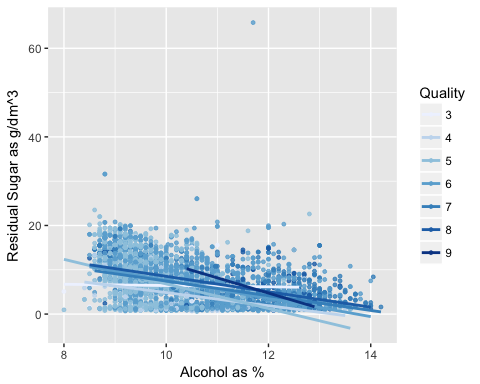
In the plot above, a negative correlation is observed between alcohol and density. As alcohol level increases, density decreases. Quality is spread all throughout the different levels of alcohol and density.

Free Sulfur Dioxide(mg/dm^3) vs Total Sulfur Dioxide(mg/dm^3) vs Quality



There exists a positive correlation between the free and the total sulfur dioxide levels. High quality wines are primarily observed within the region where total sulfur dioxide level is less than 200 mg/dm^3, and free sulfur dioxide is less than 100 mg/dm^3.

Residual Sugar (g/dm^3) vs Alcohol (%) vs Quality



In the plot above, a negative correlation exists between the sugar level and the alcohol % level. As sugar level increases, the alcohol % level decreases. However, good quality wines are not dependent upon these two variables.

## Multivariate Analysis

Including a third variable produces a more in depth observation of our analysis. Quality of wine does not depend upon residual sugar or density, because as we have already observed, good quality wines were spread all throughout the different levels of these two variables. However, there exists a positive correlation between the residual sugar and density.

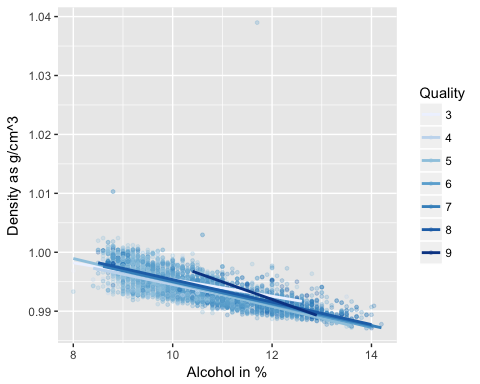
Quality is also not dependent upon alcohol levels; good quality wines were found in all levels of alcohol. However, a strong negative correlation was observed between alcohol and density.

In the third multivariate plot, high quality wines were not observed to be spread all throughout the different levels of free and sulfur dioxide. Instead, high quality wines were clustered in areas of total sulfur dioxide level less than 200 mg/dm^3, and free sulfur dioxide level less than 100 mg/dm^3.

In the fourth multivariate plot, a negative correlation is shown between the alcohol level and the residual sugar level; quality appears to be dependent upon these variables.

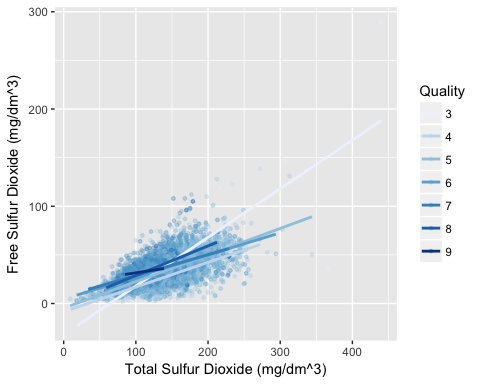
# Final Plots and Summary

Density in (g/cm^3) versus Alcohol in (%) vs Quality



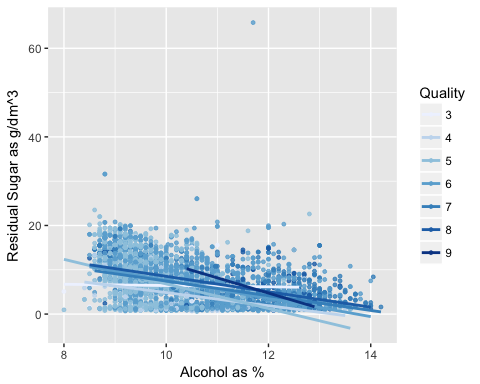
This plot was chosen to analyze how quality of wine is affected by the levels of alcohol and density. A negative correlation exists between the alcohol level and density; as alcohol level increases, density level decreases. However, we can still find wines of good quality as long as the density level remains below 0.9 g/cm^3. In general, the slope is steeper for wines of good quality.

Free Sulfur Dioxide(mg/dm^3) vs Total Sulfur Dioxide(mg/dm^3) vs Quality



The above plot was chosen because the independent and dependent variable (Total Sulfur Dioxide vs. Free Sulfur Dioxide) showed a moderately positive correlation of 0.615501. Wines of good quality in this case, did not appear to be evenly spread out. Good wines generally possessed total sulfur dioxide level less than 200 mg/dm^3, and free sulfur dioxide level less than 100 mg/dm^3.

Residual Sugar (g/dm^3) vs Alcohol (%) vs Quality



This plot shows a sequential color map to give the reader the ability to easily distinguish the differences between the varying quality levels. This plot is significant because it clearly displays the negative linear correlation between the sugar level and the alcohol % level. The quality is not dependent upon these variables as good quality wines can be found at all levels of sugar and alcohol % levels.

# Reflection

The objective of this study was to analyze the attributes that may affect the overall quality of white wines. The correlation plot was used to help make the final selections. One of the personal struggles I experienced while working on this project was getting familiarized with R. One limitation in performing this data analysis is that there is no information on the price of wines. If the price information became available in the future, it would be interesting to study how the price variable may affect the quality of the wines.