Red Wine Dataset Evaluation

Table of Contents

Brief Overview

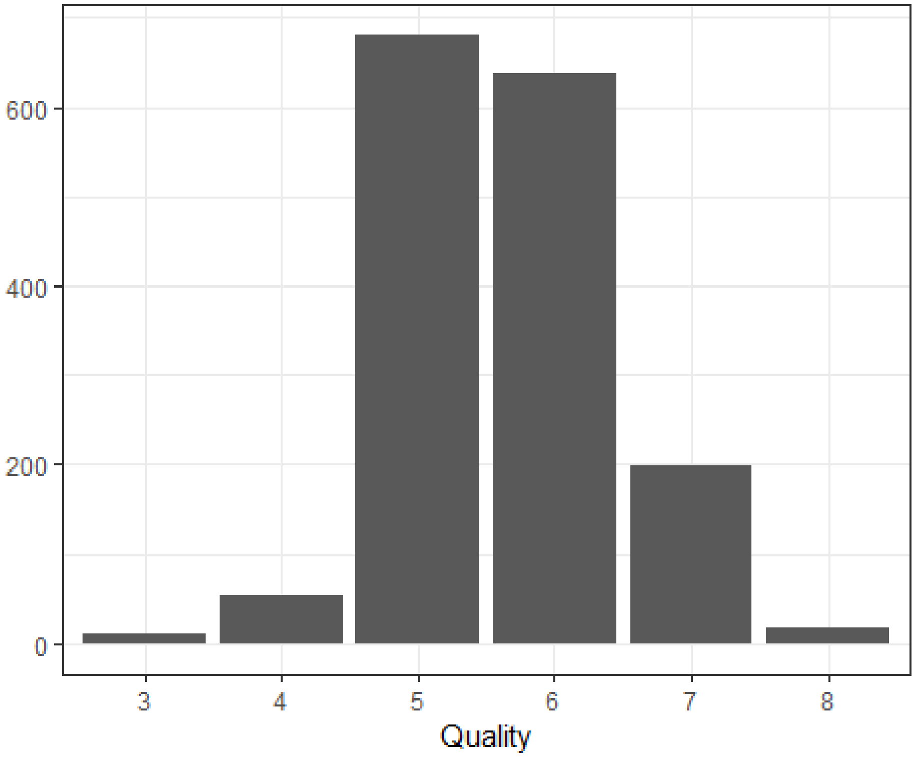
The red wine data set has -

-12 variables

-1599 observations

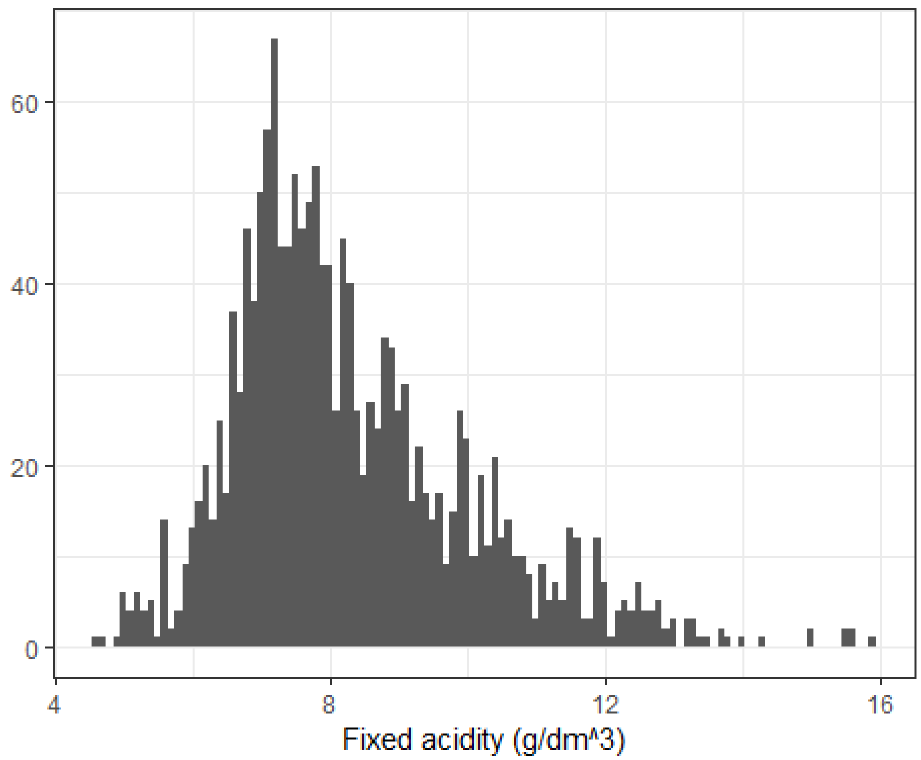
Note - The quality variable is the only discrete variable while all the others are continuous.

Univariate Plots Section

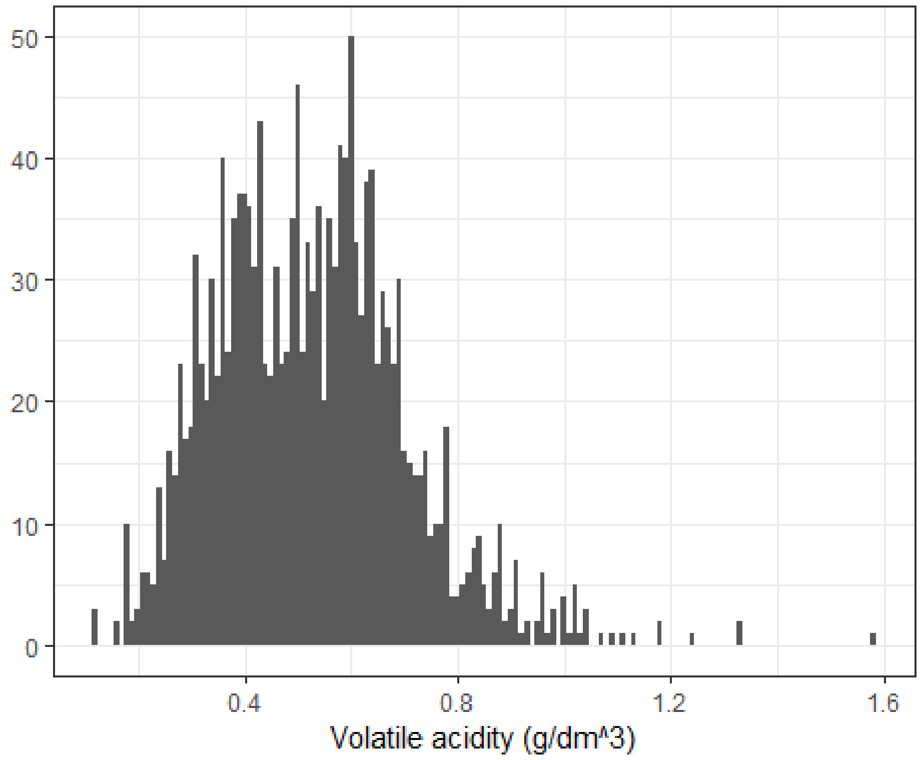


```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.000   5.000   6.000   5.636   6.000   8.000
```

The quality of wines are somewhat normally distributed, and are very concentrated around the values of 5 and 6.

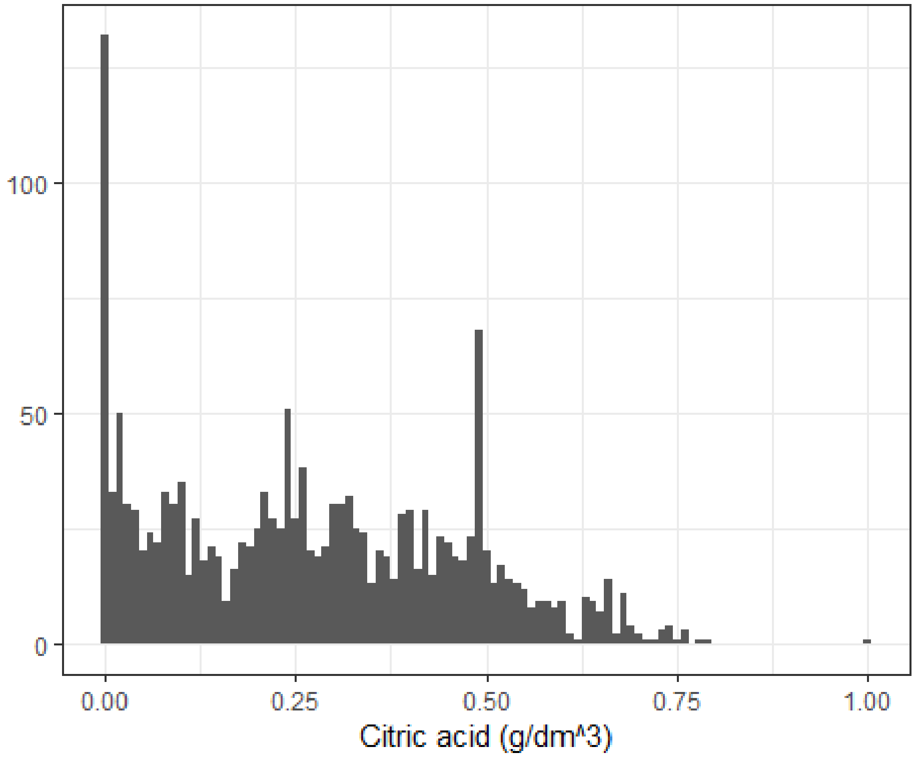


This graph shows that fixed acidity is right skewed. The graph congregates around 7.7.



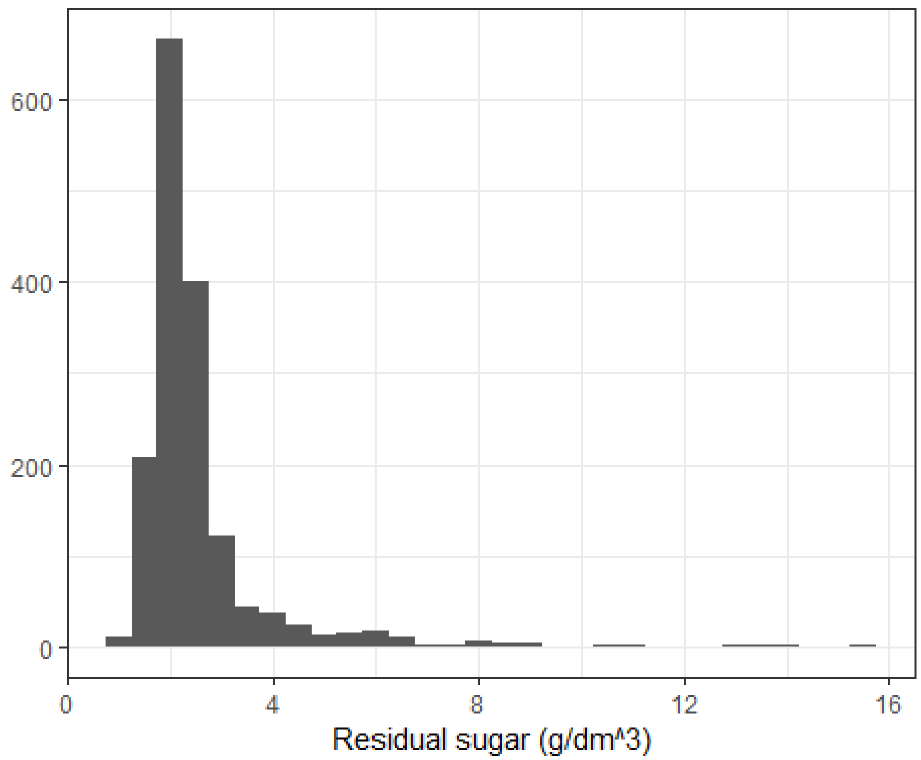
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

Volatile acid seems to be right skewed, and congregates around .5.



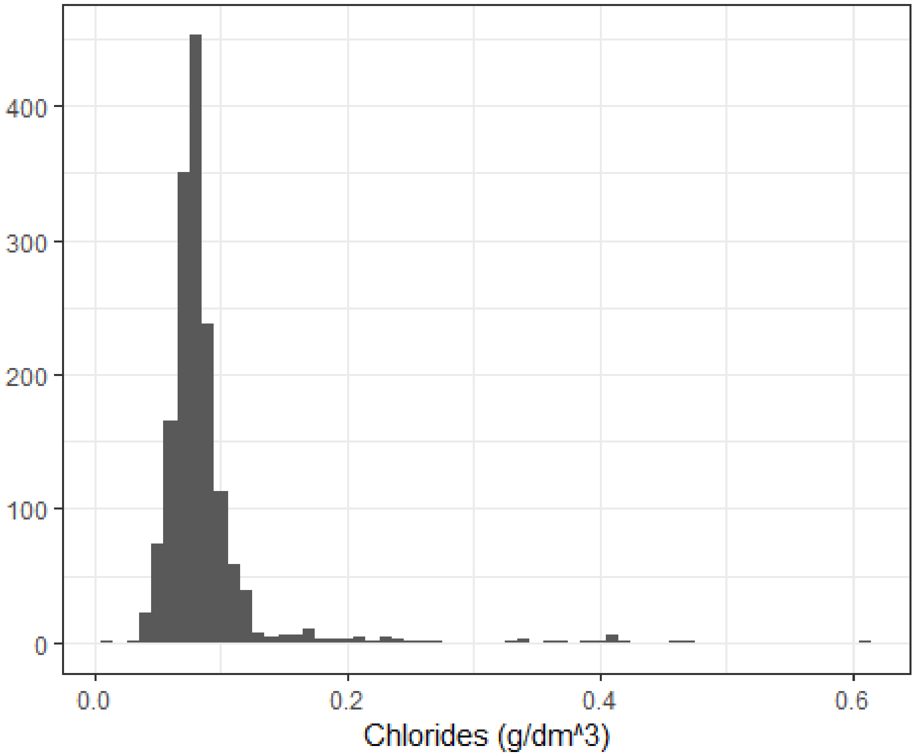
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

Citric acid is extremely anormal and cannot really be categorized.



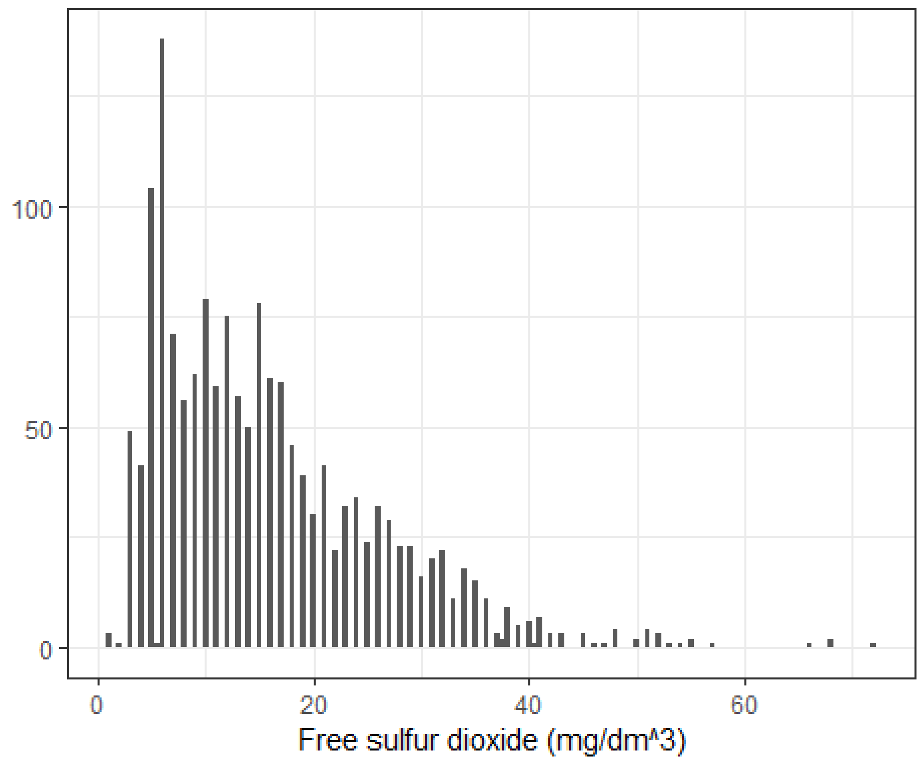
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

This graph is predominantly right skewed with outliers.



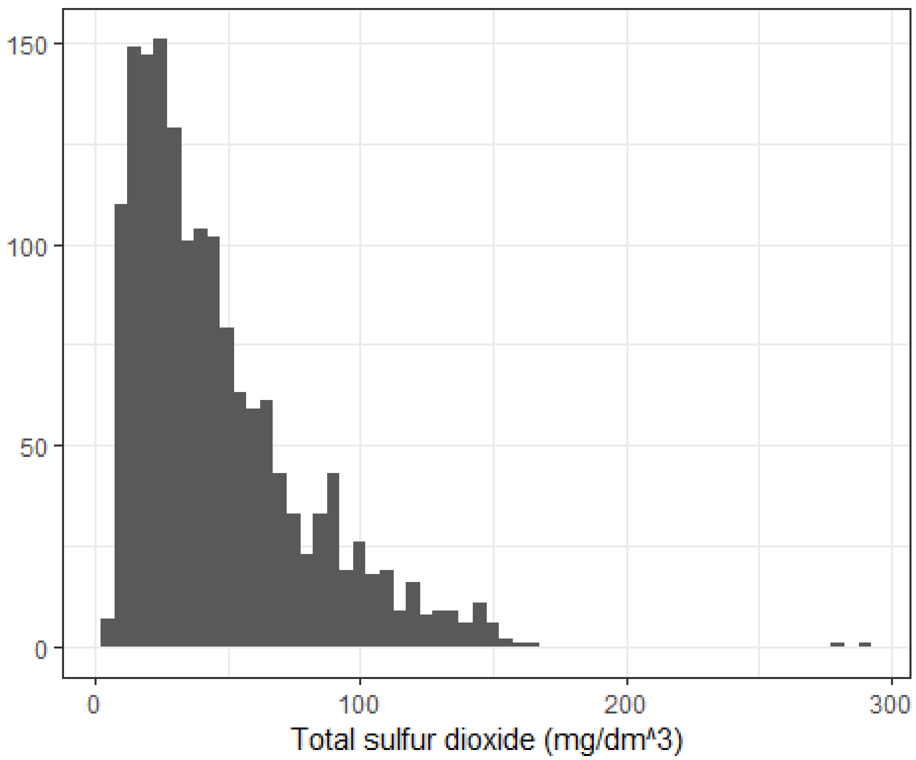
```
##     Min.  1st Qu.  Median     Mean  3rd Qu.     Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

Chlorides seem to be normal, however the outliers have a tendency to be extreme.



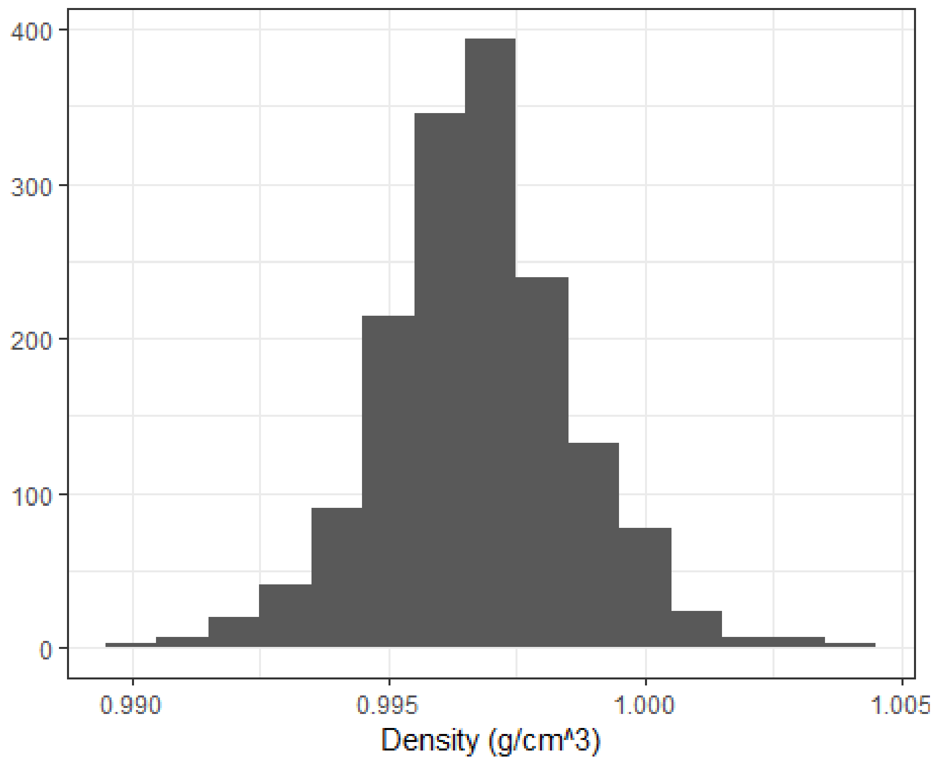
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    7.00   14.00   15.87   21.00   72.00
```

The distribution of free sulfur dioxide is right skewed.



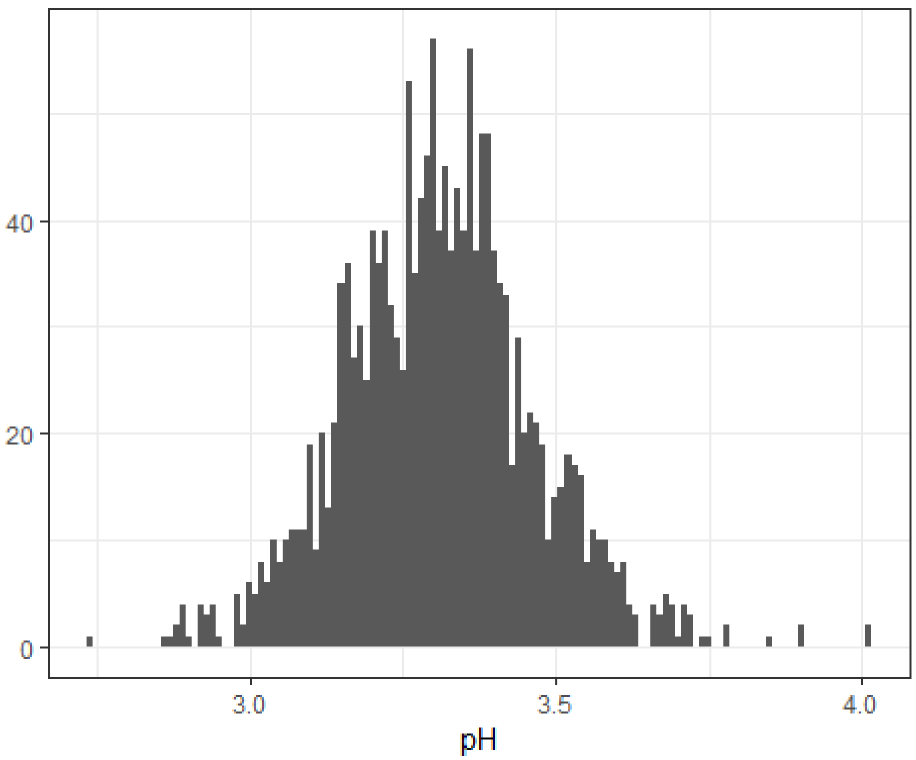
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

The distribution of total sulfur dioxide is right skewed with a few outliers in the plot.



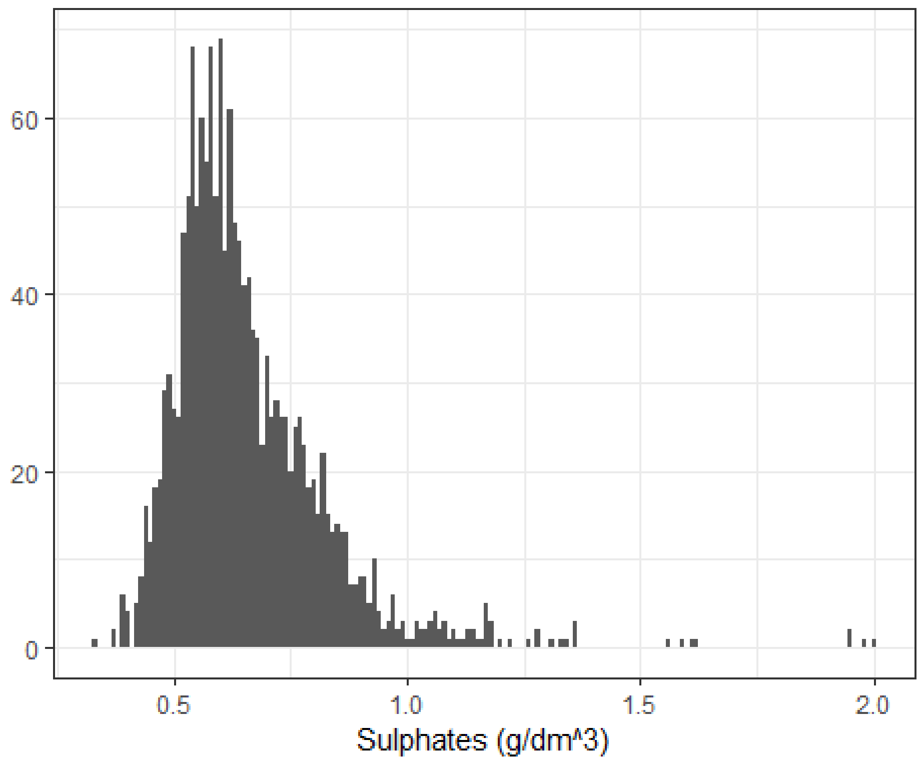
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

The distribution of density is normal.



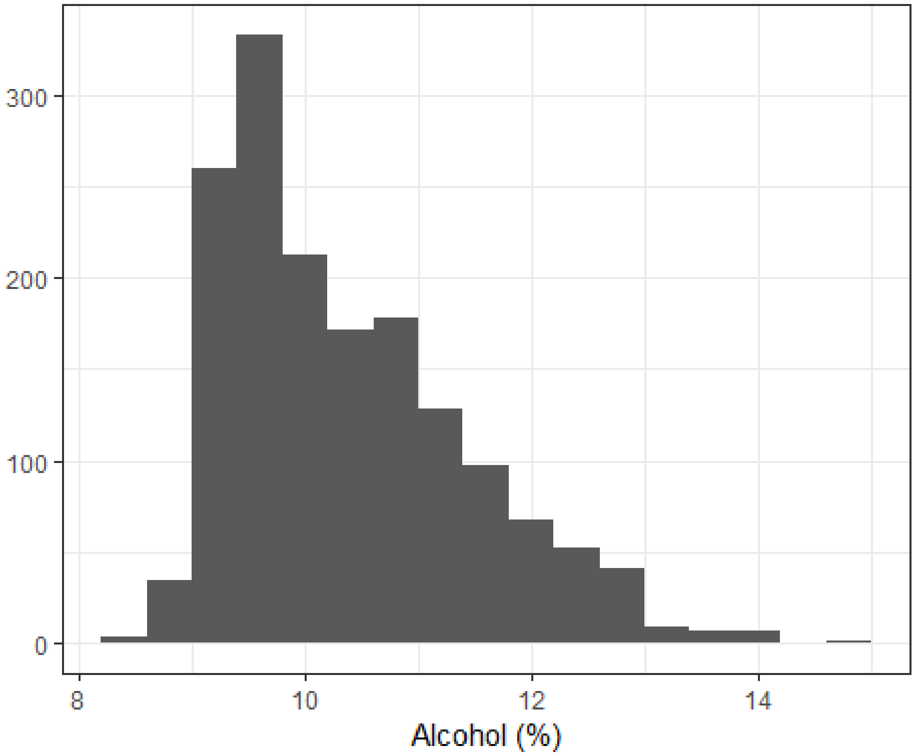
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

The distribution of pH is also normal.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

The distribution of sulphates is right skewed and the plot has some outliers.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

The distribution of alcohol is right skewed.

To examine our data in a better light we will be splitting our quality data into three categories. When we examine the quality graph we see that the data primarily spans from the value of 3 to the value of 8. So we will group values together as:

- 3 & 4 correlate to Low Quality
- 5 & 6 correlate to Moderate Quality
- 7 & 8 correlate to High Quality

Univariate Analysis

What is the structure of your dataset?

In total there are 1,599 different red wines that have been logged into this data base and they are all scaled on 12 different attributes.

- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulfur Dioxide
- Total Sulfur Dioxide
- Density
- pH
- Sulphates
- Alcohol
- Quality

The mean, median, and mode of each of these attributes has been included in the table above.

What is/are the main feature(s) of interest in your dataset?

With the dataset we have been provided, it seems most interesting the examine the correlation quality would have with other attributes. Things such pH (Is it more basic or acidic?), or fixed acidity would likely play a large part in determining exactly how the wines have been connected to certain values of quality.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

Attributes such as the types of volatilities and density would be itneresting to explore further.

Did you create any new variables from existing variables in the dataset?

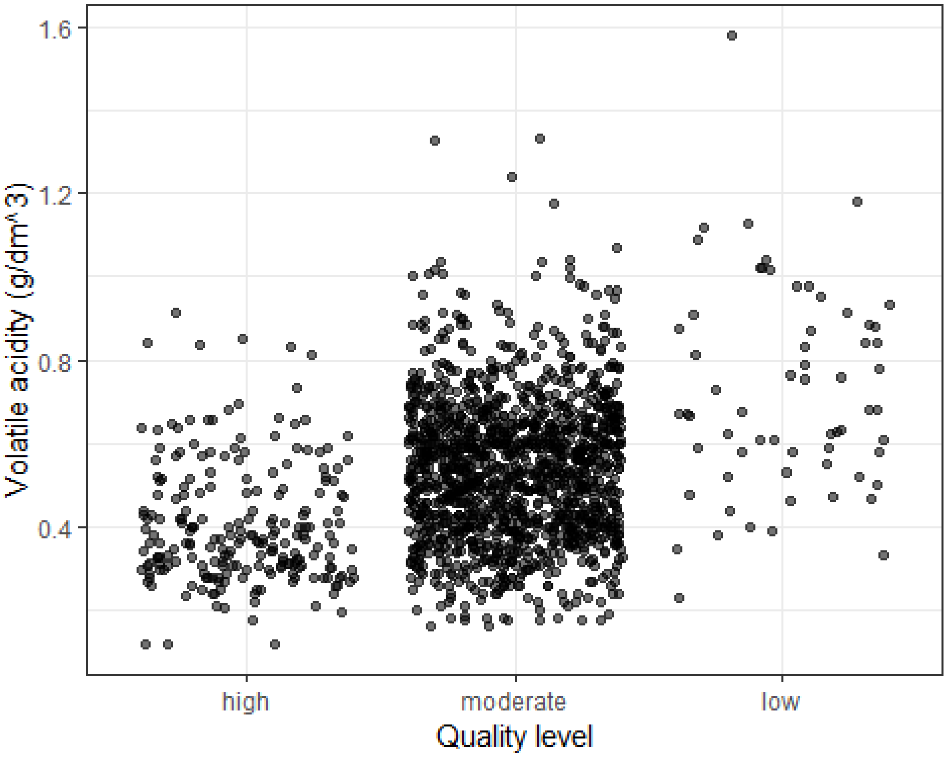
Earlier we seperated the quality attribute into three levels -> low, moderate, high. This will be used later as a variable.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

Some of the graphs (citric acid) had a tendency to form odd distributions and will be investigated.

There was no need to clean this data as it was provided tidied. Thereby the data was simply downloaded and then loaded into R.

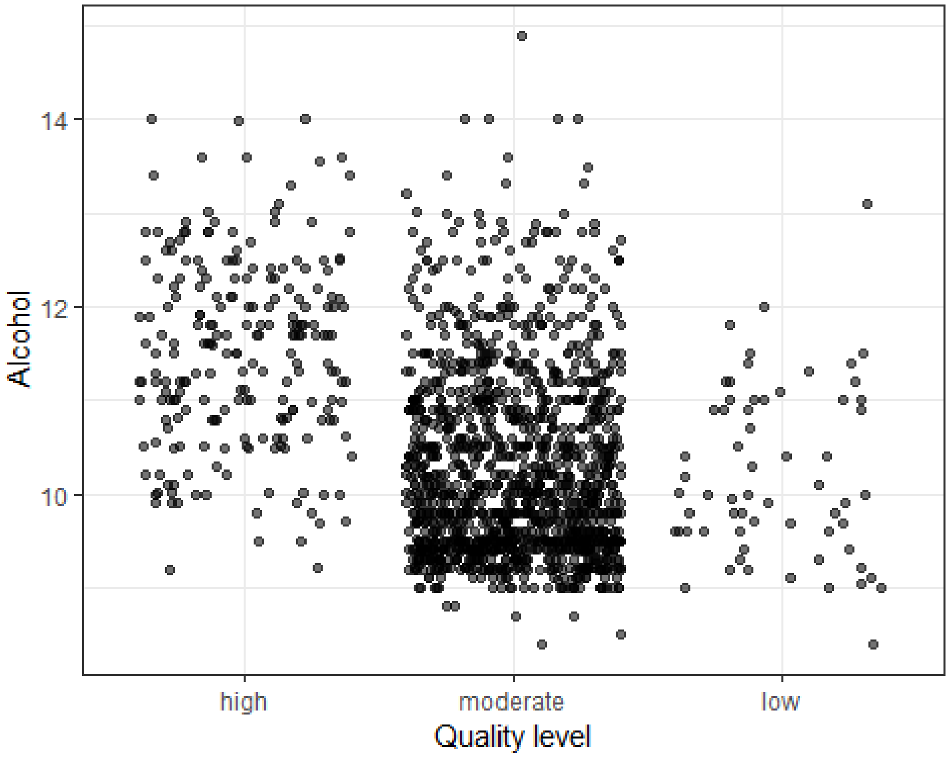
Bivariate Plots Section



```
##
## 	Pearson's product-moment correlation
```

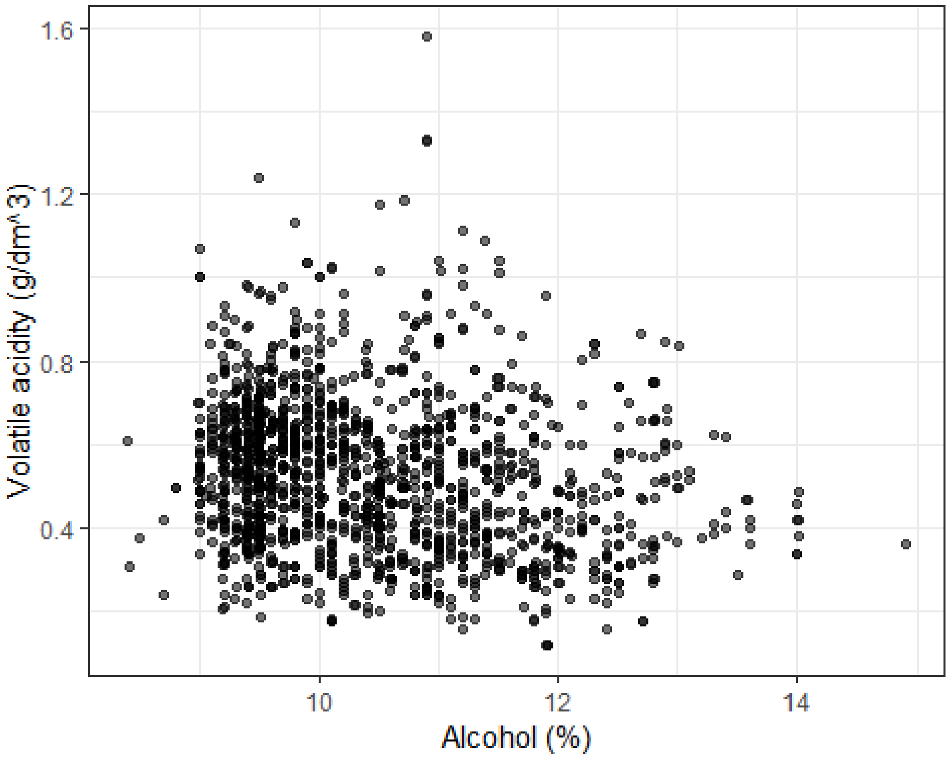
```
##
## data:  wine$quality and wine$volatile.acidity
## t = -16.954, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4313210 -0.3482032
## sample estimates:
##        cor
## -0.3905578
```

This graph quite clearly points to a strong negative correlation between quality and volatile acidity. It is important to note that the graph has somehow had its x axis flipped so the values go from high to low instead of low to high. This correlation coefficent comes out to -0.39. If we were to examine this trend we can see that this is an expected outcome. Volatile acidity refers to the “steam distillable acids present in wine”. These acids extend to but are not limited to lactic, formic, butyric, and propionic acids. These are all also prevalent in acid which makes sense considering that such a taste is not desired in wine.



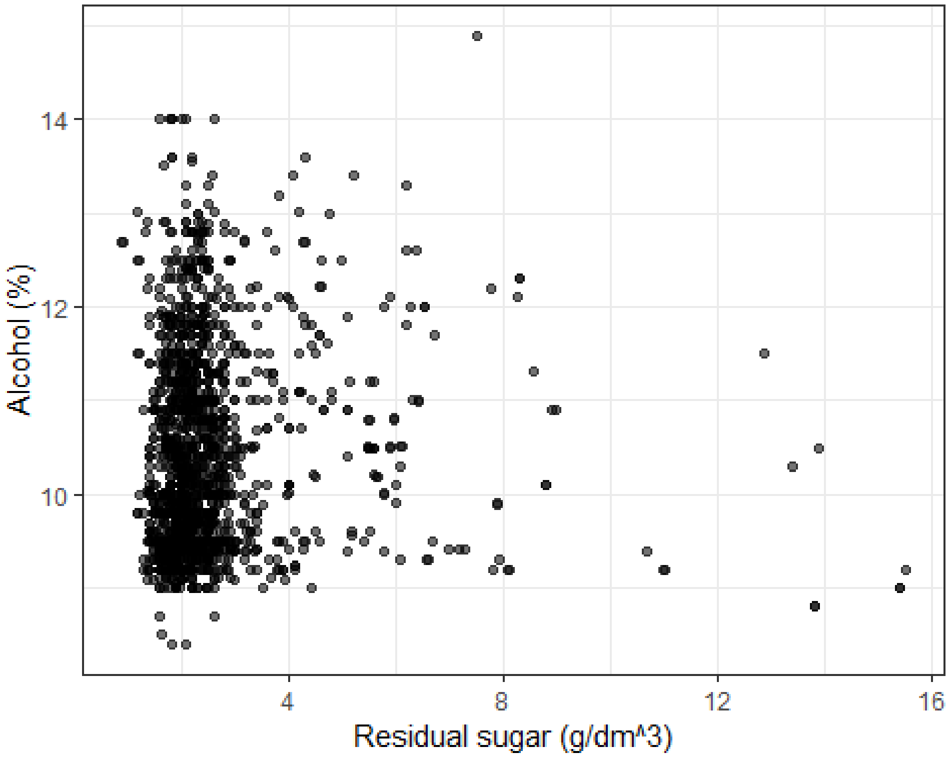
```
##
##  Pearson's product-moment correlation
##
## data:  wine$quality and wine$alcohol
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

There is a positive correlation between alcohol levels and the deemed quality of wine. The correlation coefficient is 0.476. Average quality and low quality wines have their percent alcohol contents concentrated around 10 whereas high quality wines have their percent alcohol contents concentrated around 12. This is likely due to the fact that a higher alcohol content gives wine a more desired effect.



```
##
##  Pearson's product-moment correlation
##
## data:  wine$alcohol and wine$volatile.acidity
## t = -8.2546, df = 1597, p-value = 3.155e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2488416 -0.1548020
## sample estimates:
##       cor
## -0.202288
```

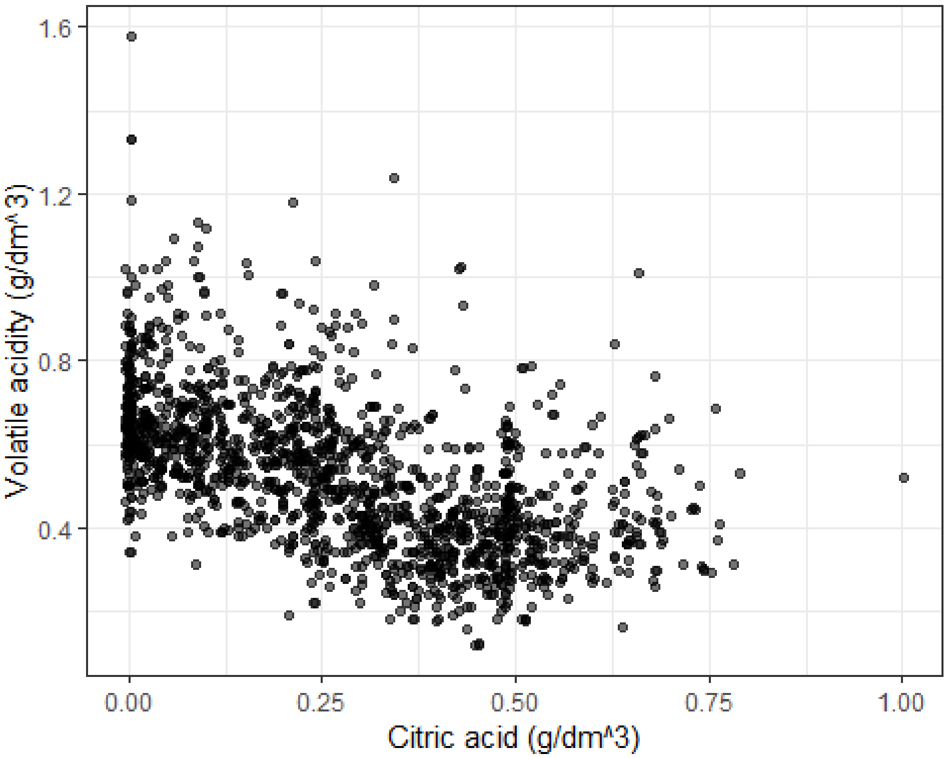
A weak negative correlation of -0.2 exists between percent alcohol content and volatile acidity. So we can’t really draw the conclusion that a buildup of acid is due to higher alcohol content.



```
##
##  Pearson's product-moment correlation
##
```

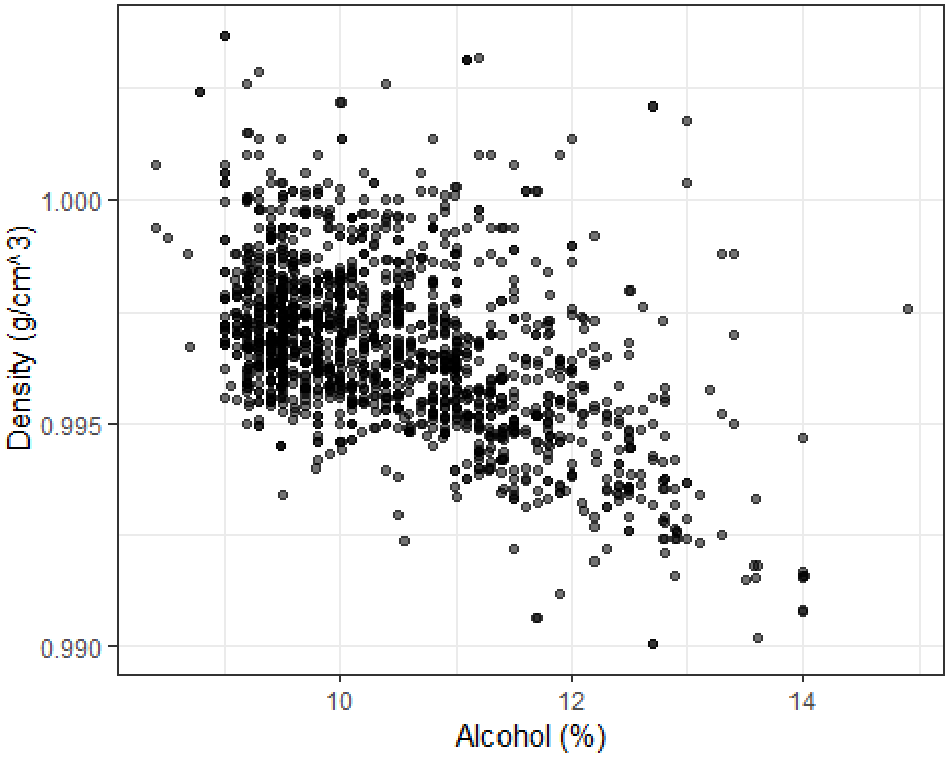
```
## data:  wine$alcohol and wine$residual.sugar
## t = 1.6829, df = 1597, p-value = 0.09258
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006960058  0.090909069
## sample estimates:
##        cor
## 0.04207544
```

In the case of alcohol vs residual sugar, the correlation coefficent is a paltry .04. This is understandable as steps can be taken to regulate sugar and alcohol levels to a winemaker’s tastes.



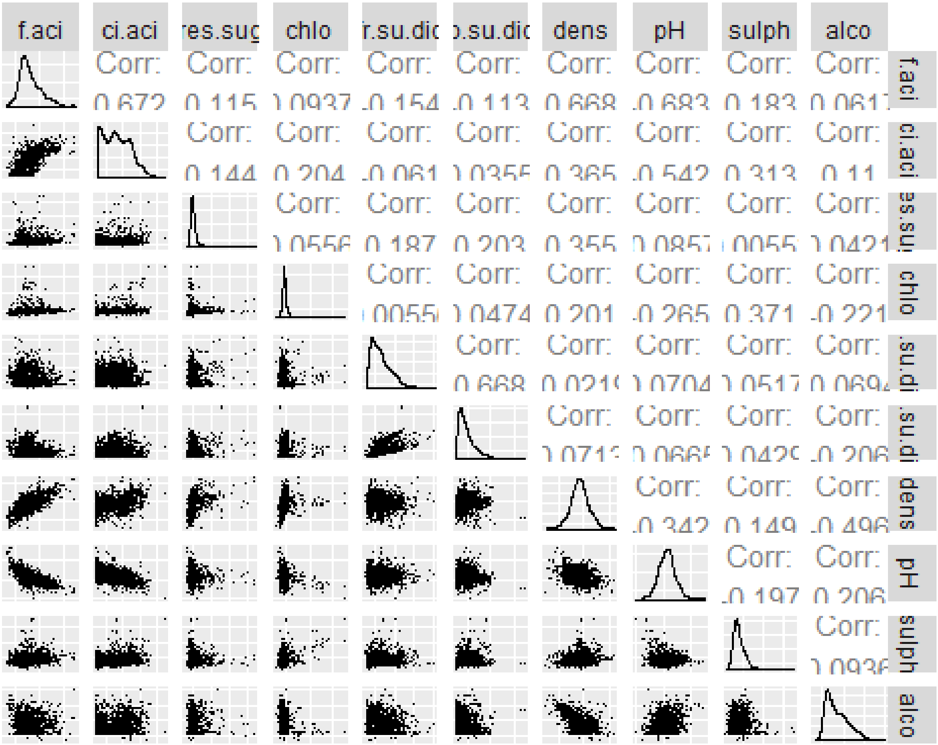
```
##
##  Pearson's product-moment correlation
##
## data:  wine$citric.acid and wine$volatile.acidity
## t = -26.489, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5856550 -0.5174902
## sample estimates:
##        cor
## -0.5524957
```

There is a negative correlation of -.55 between citric acid and volatile acidity.



```
##
##  Pearson's product-moment correlation
##
## data:  wine$alcohol and wine$density
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##        cor
## -0.4961798
```

Here the correlation coefficient is -0.5. The is due to the overall density of the wine getting diluted as more alcohol is added.



This graph shows different relationships and correlations that can be found in the data.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Volatile acidity vs. Quality is a relationship that was examined in this stage of the study. We noticed that as volatile acidity goes up (the wine becomes more acidic), quality goes down. With further examination into this phenomena we found that the acids responsible for volatile acidity are the same acids that give vinegar that iconic taste. In lieu of maintaining good flavor in wine it makes sense that this relationship exists.

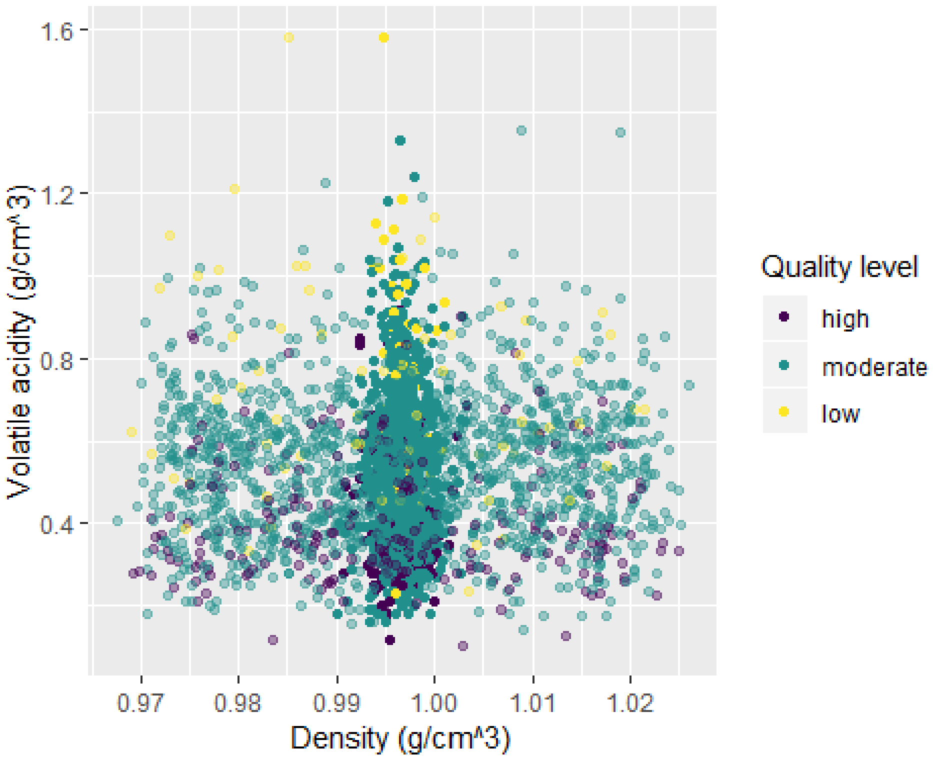
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

We noticed many relationships but the most interesting one I found was actually the lack of relationship between residual sugar and alcohol. Wine in its self is the fermentation of overripe fruit. The more fruit ripens that more alcohol can be derived from it, however at the same time more sugar and glucose is produced. In present times it seems entirely possible that such factors can be independently monitored, however during ancient times these two factors would have likely been correlated in some fashion.

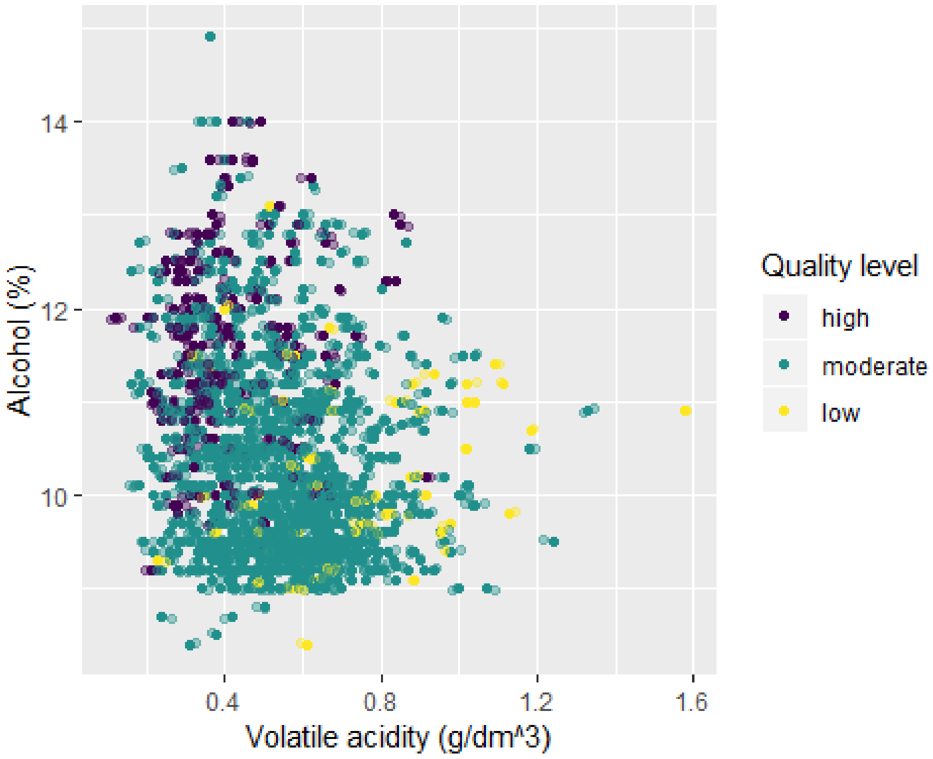
What was the strongest relationship you found?

Quality is strongly negatively correlated with volatile acidity.

Multivariate Plots Section

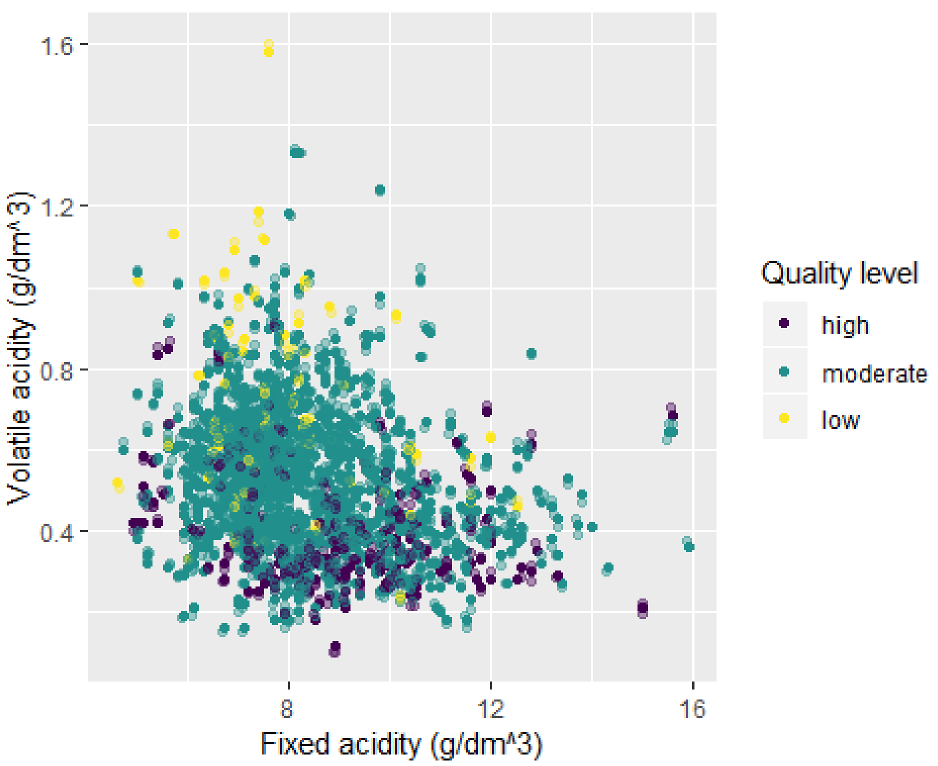


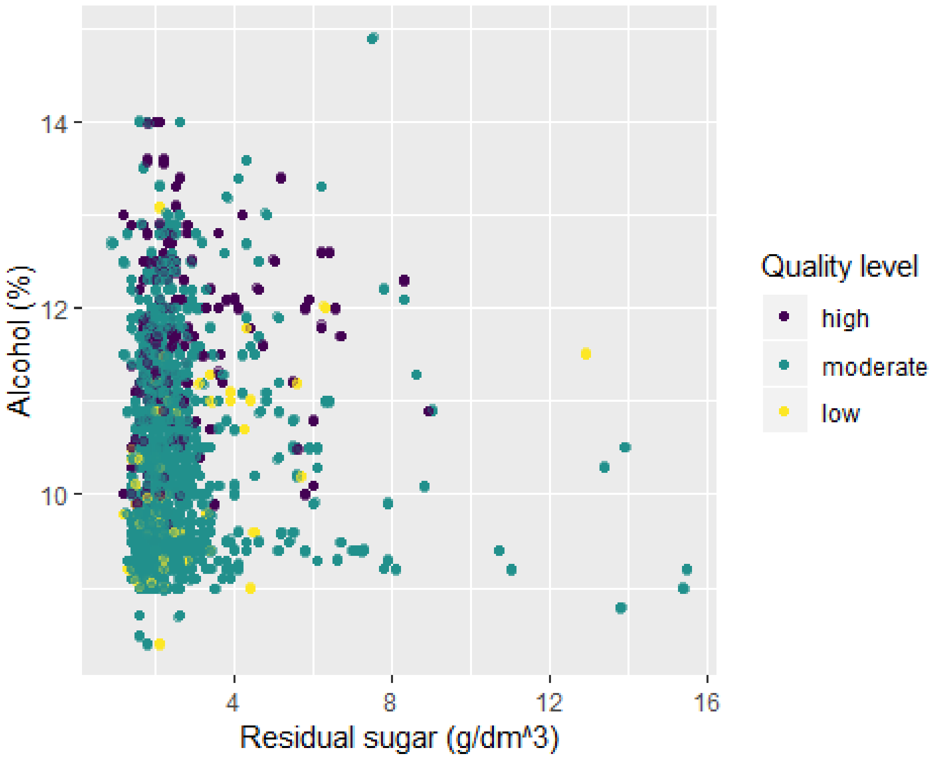
The densities of high quality wines are concentrated between 0.994 and 0.998, and the lower part of volatile acidity (y axis)

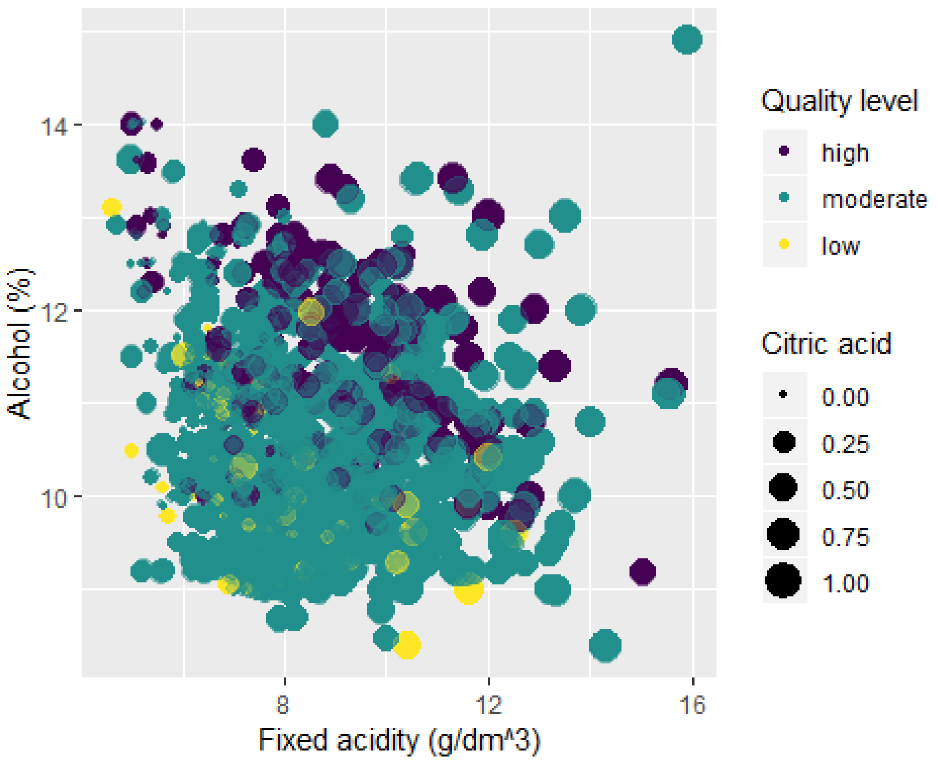


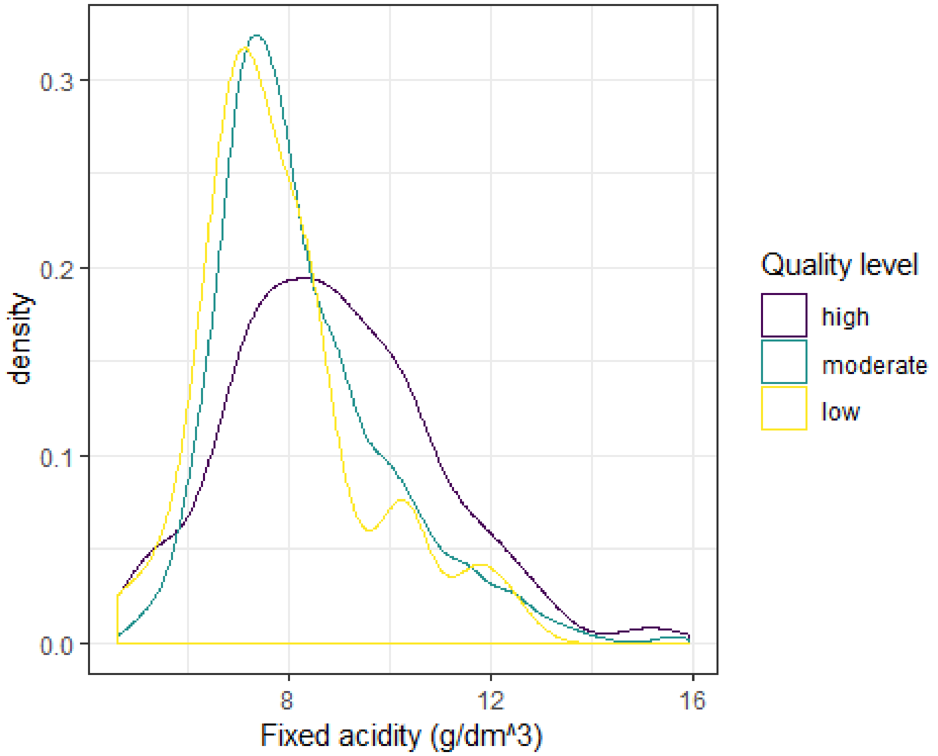
```
## [1] "Percent alcohol contents by quality level:"

## # A tibble: 3 x 3
##   quality.level  mean    sd
##   <ord>         <dbl> <dbl>
## 1 high           11.5 0.998
## 2 moderate       10.3 0.972
## 3 low            10.2 0.918

## [1] "Volatile acidities by quality level:"

## # A tibble: 3 x 3
##   quality.level  mean    sd
##   <ord>         <dbl> <dbl>
## 1 high          0.406 0.145
## 2 moderate      0.539 0.168
## 3 low           0.724 0.248
```

High quality feature seems to be associated with alcohol ranging from 11 to 13, volatile acidity from 0.2 to 0.5, and citric acid from 0.25 to 0.75

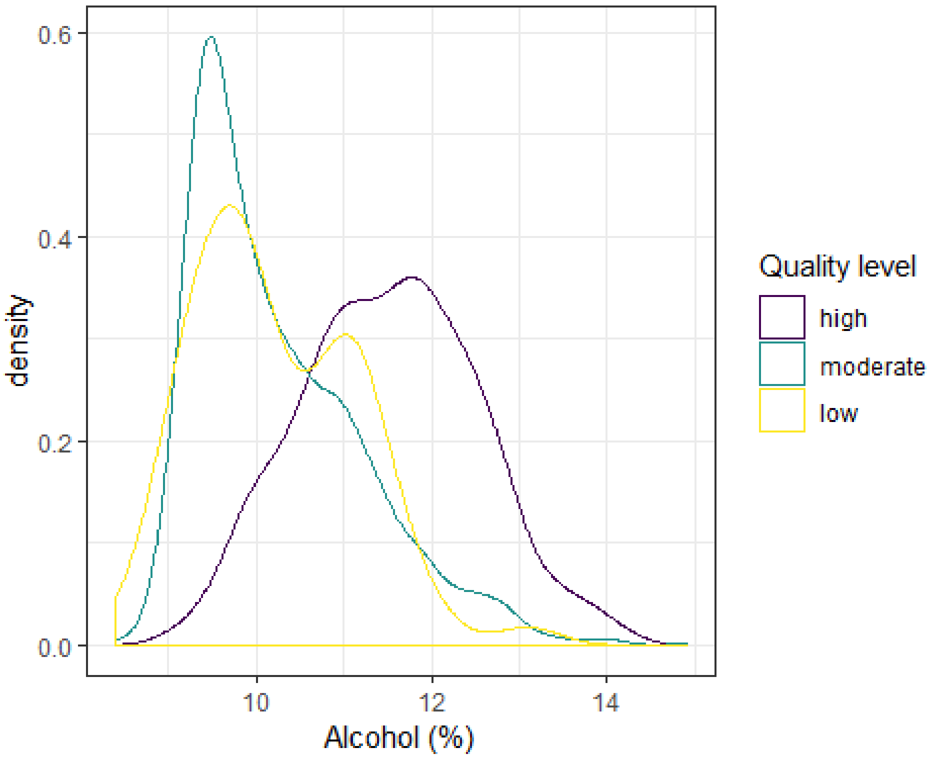

1.6
1.2
0.8
0.4
Volatile acidity (g/dm^3)
8
12
16
Fixed acidity (g/dm^3)
Quality level
high
moderate
low


14
12
10
Alcohol (%)
4
8
12
16
Residual sugar (g/dm^3)
Quality level
high
moderate
low


Alcohol (%)
14
12
10
8
12
16
Fixed acidity (g/dm^3)
Quality level
high
moderate
low
Citric acid
0.00
0.25
0.50
0.75
1.00


density
0.3
0.2
0.1
0.0
8
12
16
Fixed acidity (g/dm^3)
Quality level
high
moderate
low

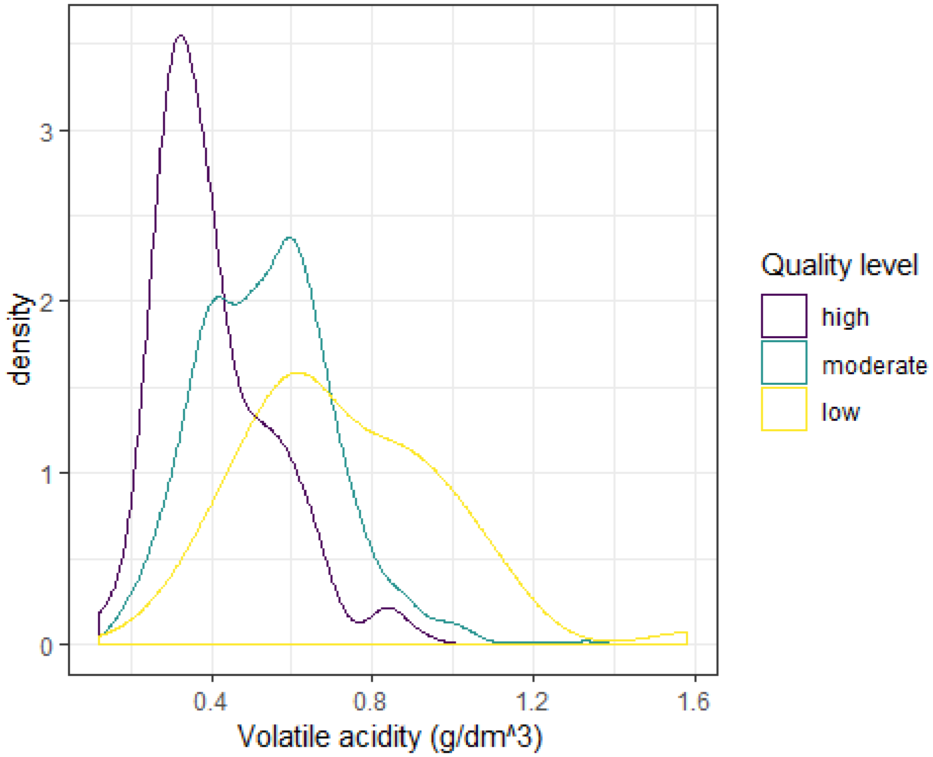
The distribution of low and average quality wines seem to be concentrated at fixed acidity values that are between 6 and 10. pH increases as fixed acidity decreases, and citric acid increases as fixed acidity increases.



```
##
## Call:
## lm(formula = quality ~ alcohol, 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
```

High quality wine density line is distinct from the others, and mostly distributed between 11 and 12.



```
##
## Call:
## lm(formula = quality ~ volatile.acidity, 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
```

In this chart we see that as volatile acidity increases quality decreases.

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + alcohol, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59342 -0.40416 -0.07426  0.46539  2.25809
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.09547    0.18450   16.78   <2e-16 ***
## volatile.acidity -1.38364    0.09527  -14.52   <2e-16 ***
## alcohol           0.31381    0.01601   19.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6678 on 1596 degrees of freedom
## Multiple R-squared:  0.317,  Adjusted R-squared:  0.3161
## F-statistic: 370.4 on 2 and 1596 DF,  p-value: < 2.2e-16
```

When we add alcohol to the linear model we can see that the Rsquared value has been doubled.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

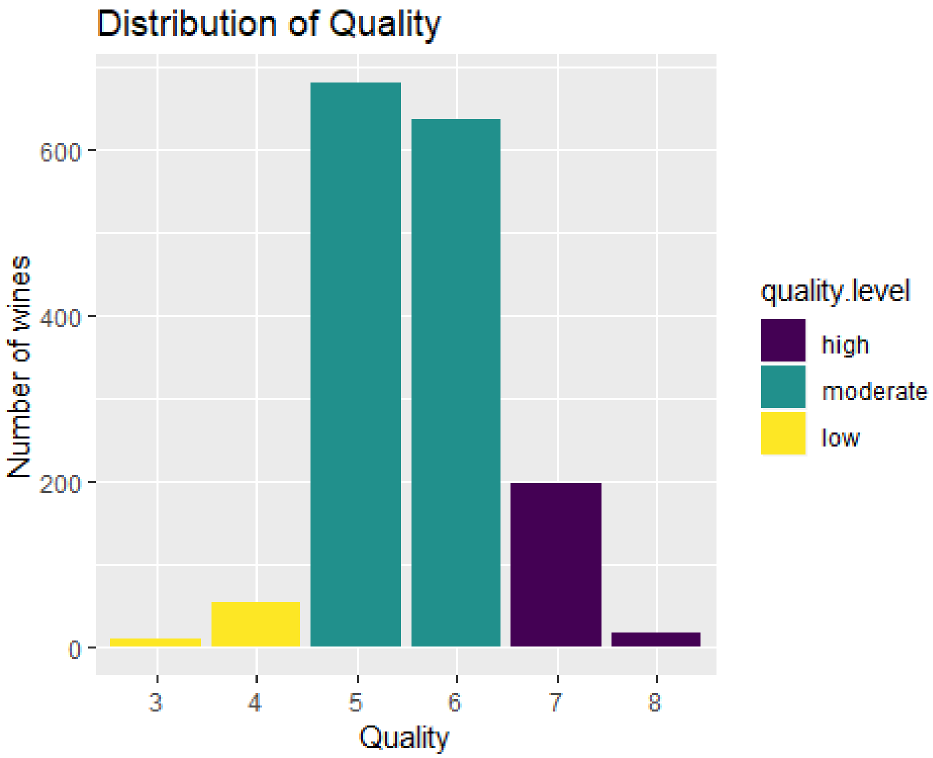
Quality doesn't really appear to have a significant relationship to alcohol, despite the correlation coefficient earlier showing a slight relationship, but the p-values show a level of importance. Examining the graph we can see that high quality wine has a tendency to have a higher density to volatile acidity ratio.

Were there any interesting or surprising interactions between features?

Quite surprisingly certain aspects had little to no effect on the quality of wine such as residual sugar. One would expect that the sweetness or sugary aspect of wine would affect it one way or another.

Final Plots and Summary

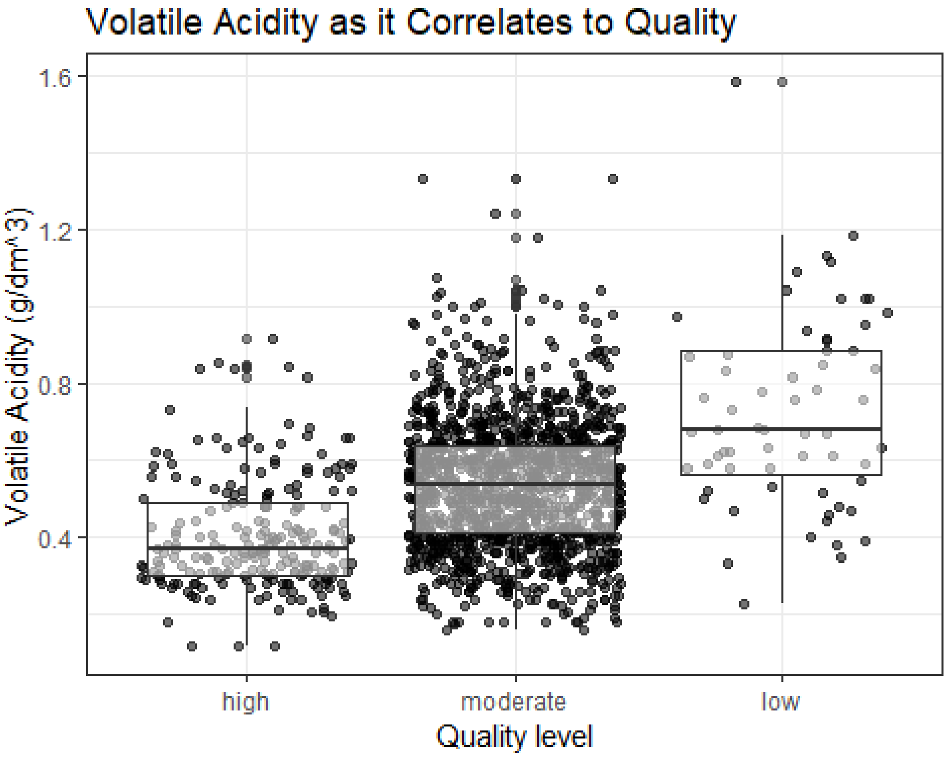
Plot One



Description One

I chose this plot because it provides a simple overview of the dataset from one of our more important attributes - Quality. Here we can see that wine quality is scaled on a normal distribution scale and the values extend from 3 to 8. I would presume that wine ratings actually encompass a scale of 1 - 10, however such values are not prevalent in the current dataset.

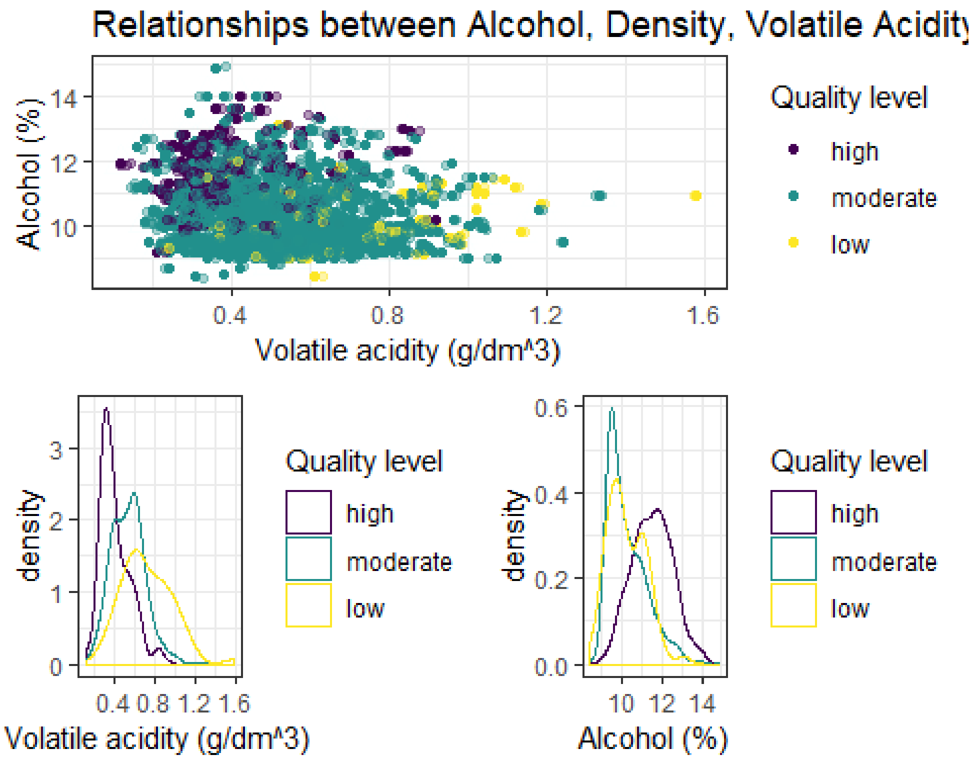
Plot Two



Description Two

One of my favorite things to research in this study was the correlation between volatile acidity and red wine quality. I didn’t know that certain enzymes such as lactic acid were prevalent in wines, and such acids were keynotes in vinegars. In this graph we can see that there is a strong negative correlation between volatile acidity and red wine quality, once again we can attribute this to the unwanted flavor such vinegar components would give wine.

Plot Three



Description Three

Now we can add more variables to examine in the case we had with volatile acidity and quality. Here we added alcohol and decided to examine what combination of volatile acidity and alcohol contribute to better wines. The graphs show that a low volatile acidity to a moderate - high alcohol content lead to better quality wines.

Reflection

This wine data set contained 1,599 different wines that were graded on 12 different attributes. After using three different types of analysis (univariate, bivariate, and multivariate) we can conclude certain relationships between these attributes.

For example we found that volatile acidity negativly affects the quality of red wines, whereas a slighly higher alcohol content positively affects it. This was certainly interesting as I am personally someone who has little to no knowledge of wine attributes and qualities.

As far as some of my hypotheses, I had earlier predicted (not recorded) that alcohol content would have no effect on quality, and that volatile acidity would have a positive effect. Going through the report now I can say that both of these hypotheses can be deemed false. In the future I would love to expand upon my research and really dig into what makes a good wine, some predictions I have are as follow:

-Fixed acidity would have a positive relationship

-Ph would have an odd relationship (Would it be something of a parabola-shaped relation ship as neither a predominantly basic, nor a predominantly acidic compound would have good place in wine?)

-Total Sulfur Dioxide would have a negative correlation as sulfur is an extremely pungent element and would give wine a strong flavor that is likely not desired.

To improve our dataset we could include wines that have some of the more extreme quality values, 1-2 and 9-10. This would give us a better insight into what makes an extremely poor or excellent wine. We could then use those attribute values and compare them to the values in more moderate values.

Some limitations this dataset does have include that the wines are perhaps not indicative of a large enough sample size, and havign more opinions may remove any trace bias.

Things I struggled with included finding the right relationships to graph, and graphing them in an appropriate manner. At times certain variables were more tricky to play around with than others.