Project Details

This project deals with the EDA of the Red Wine data set. I will use my skills gathered here in the course to come up with some data analysis that would help us find some pattern or just give us more information about the dataset. Let us first start by loading the dataset in R.

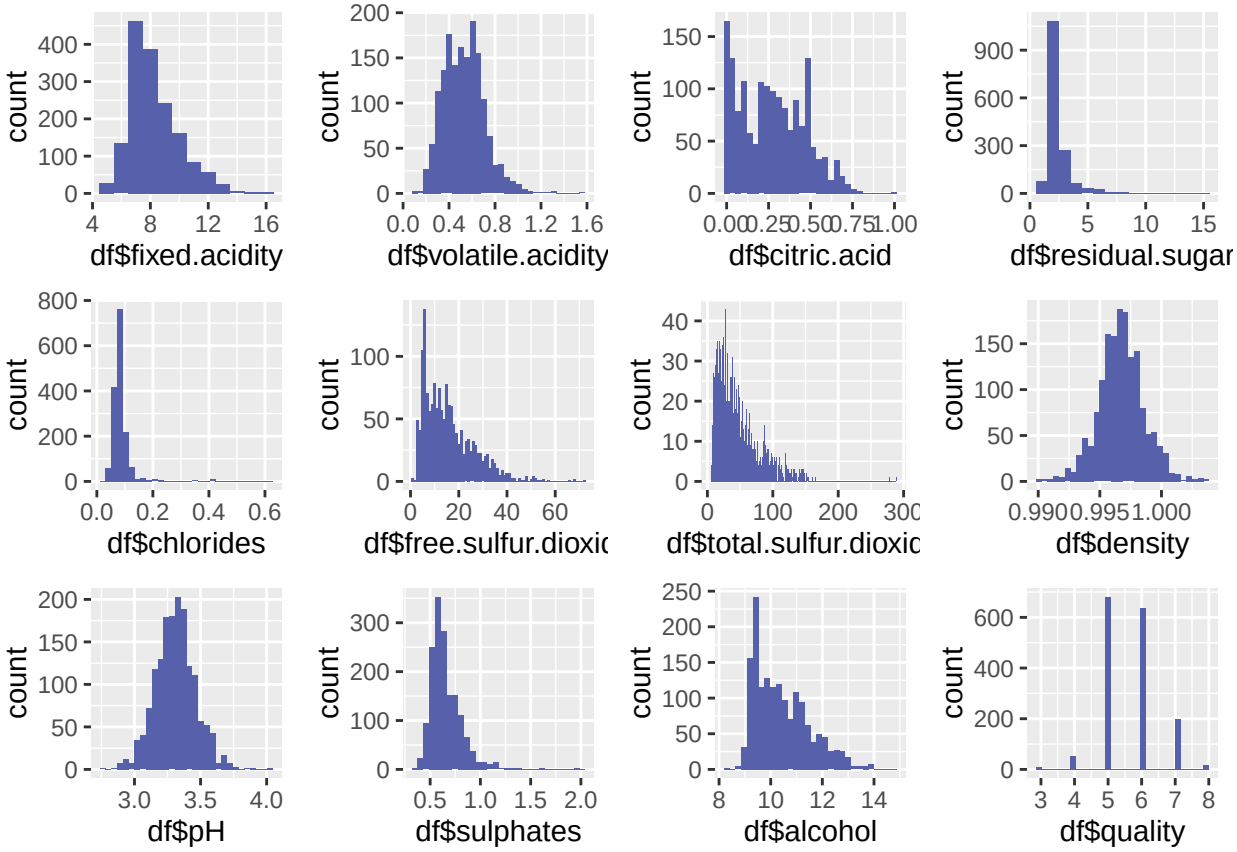
Our data was loaded successfully. Now let us see what type of data we have in the dataset and try runnning the basic statistics on it.

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 ...
```

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

The dataset has 1599 observations and 13 variables. Our dependent varibale is the Quality variable that describes the quality of a wine. This is a ranked system that has been rated by at least 3 judges and according

to the documentation, it scales from 0-10. If we look at the summary of the quality variable, we see that the minimum score is 3, the maximum is 8 and the mean is 5.636. Another variable that is intersting is the alcohol variable, with the mean alcohol content being 10.42, which generally us quite strong! The variable is X is the serial number or a column that contains the row number.



Plotting the basic histograms, we can see that I have not added binwidths to a couple of variables, the primary reason for that is we have distorted/data that does not make sense if we add binwidths to these plots. We see that pH follows a very nice normal distribution curve, so does density. Quality gives us a somewhat normally distributed graph too.

Univariate Variable Analysis:

Quality:

The quality varibale can be used to create a new varibale. We can discretize the quality variable as follows:

1. Wine below rating 4 should be labelled as 'Poor'
2. Wine between 4-7 should be labelled as 'Ideal'
3. 7 and above should be labelled as 'Good'

Alcohol:

The alcohol content in a wine can also be discretized and added as a new variable. Following are the rules:

1. Below 9.50, 'Mild'

2. Between 9.50-11.10, 'Classic'
3. Above 11.10, 'Strong'

```
##  Poor Ideal  Good
##    63  1319   217
```

```
##    Mild Classic  Strong
##     297     895     407
```

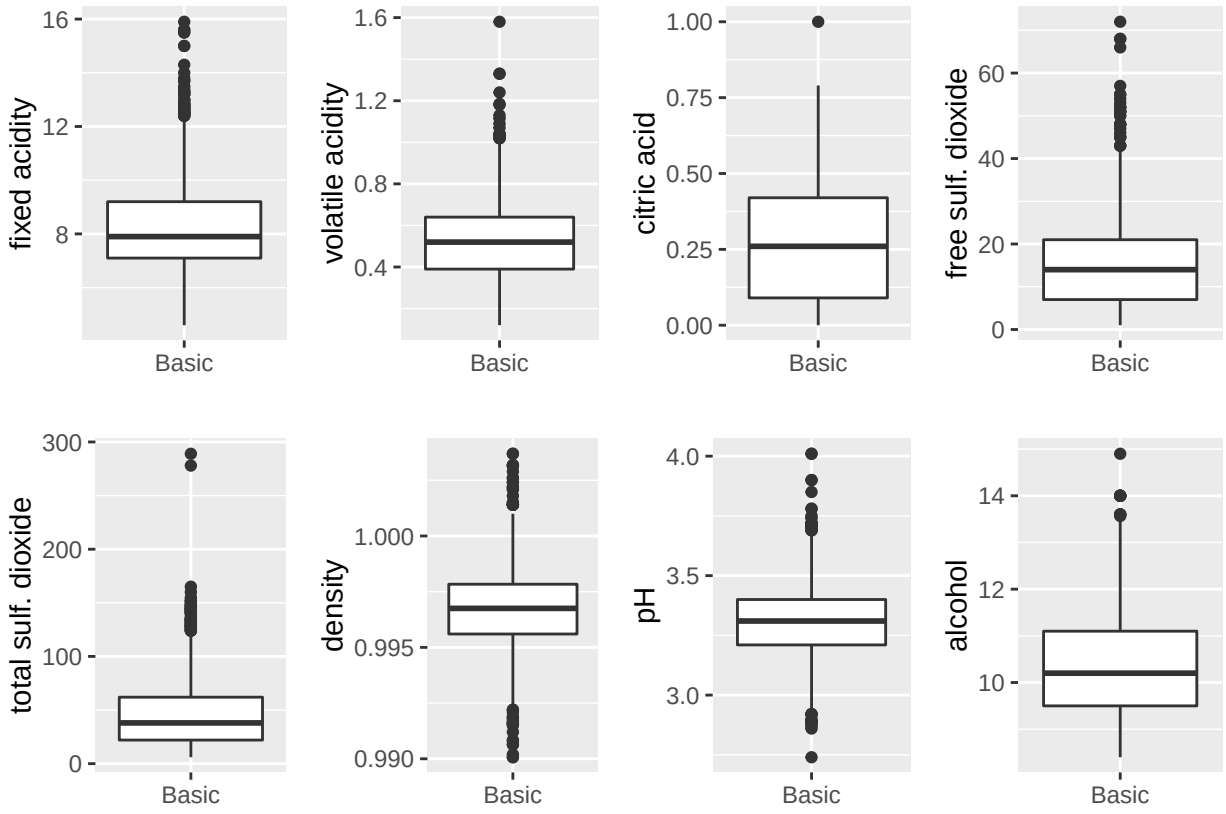
Rating of the wine can be broken down as:

1. Poor: 63
2. Ideal: 1319
3. Good: 217

The wines can be divided on the basis of their alcohol rating as:

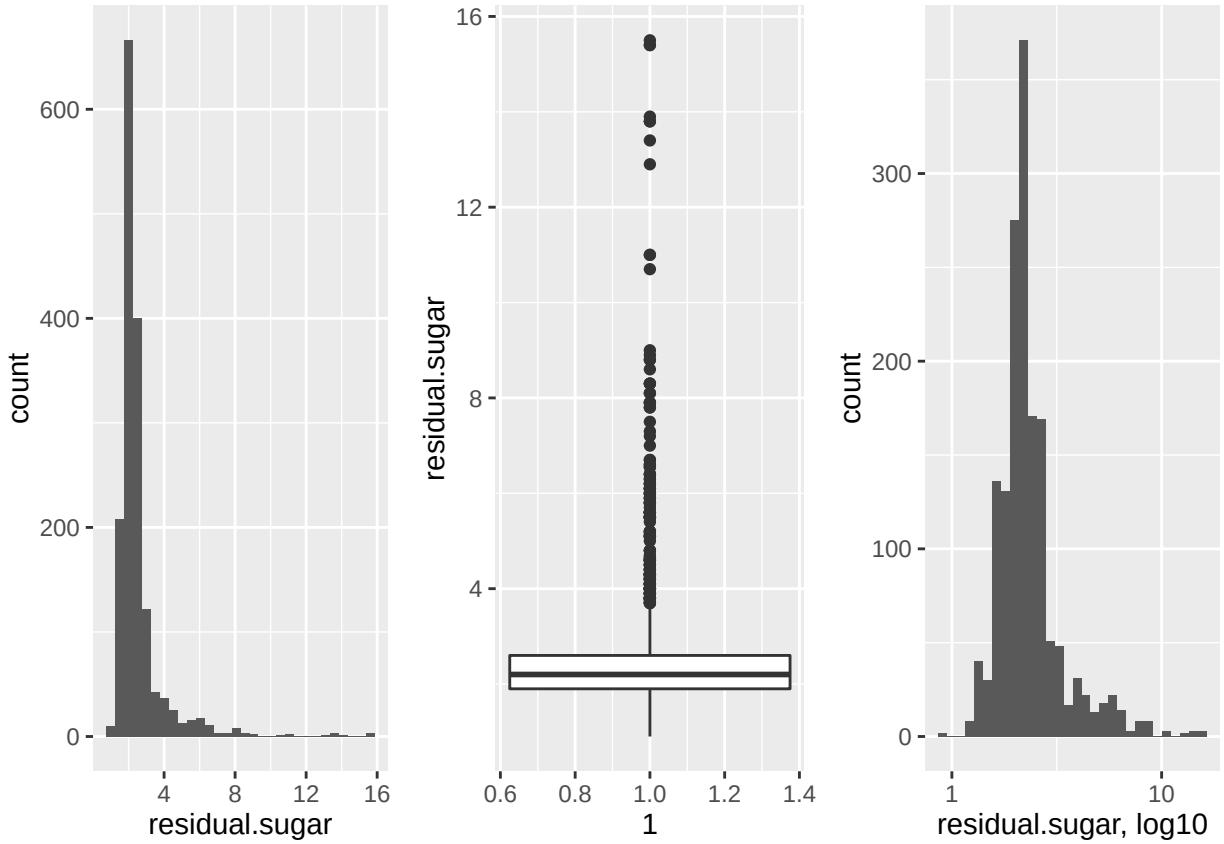
1. Mild: 297
2. Classic: 895
3. Strong: 407

Checking outliers:



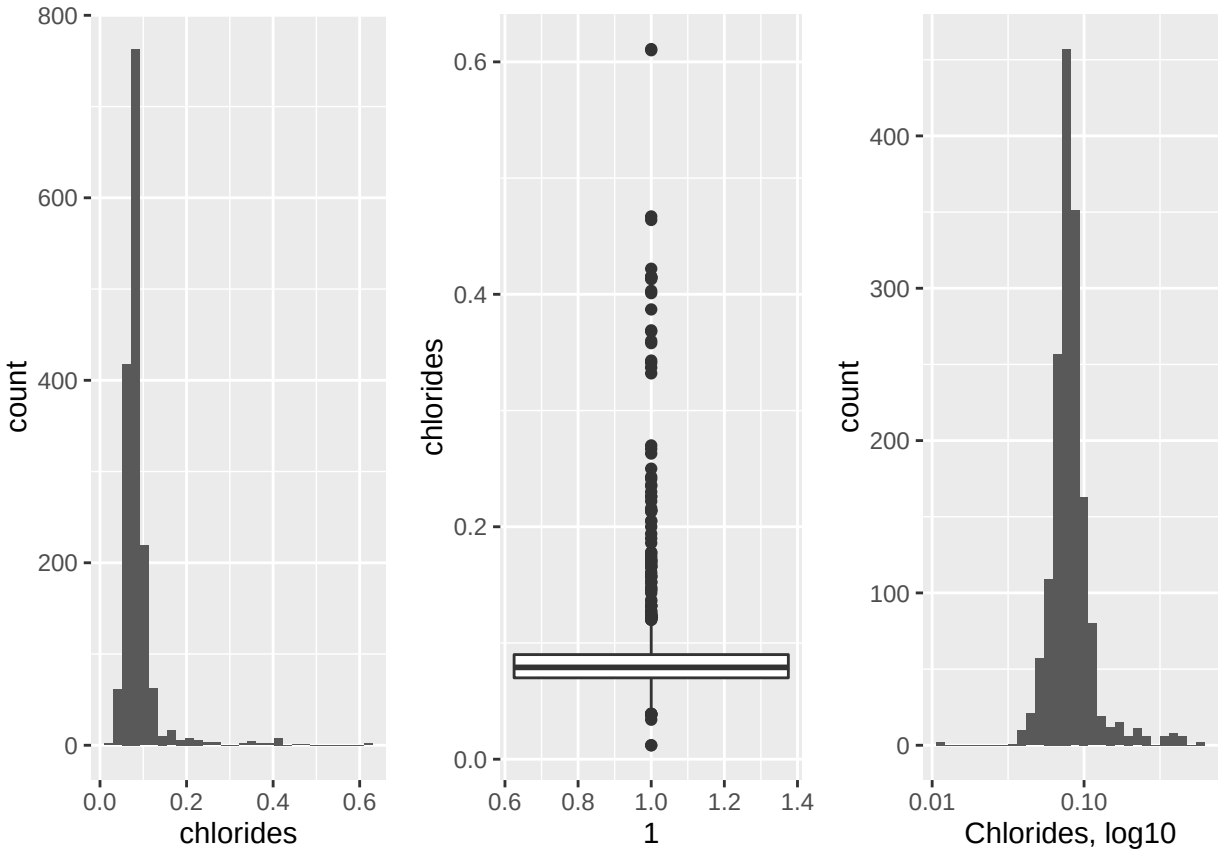
From the boxplots, it is clear that chlorides, residual sugar and sulphates have a lot of outliers. This can also be made out from the histograms that they have long tail ends, specifying that they contain a lot of

outliers. We can closely examine each of the variable by plotting the 95th percentile and log scale graph for the following variables.



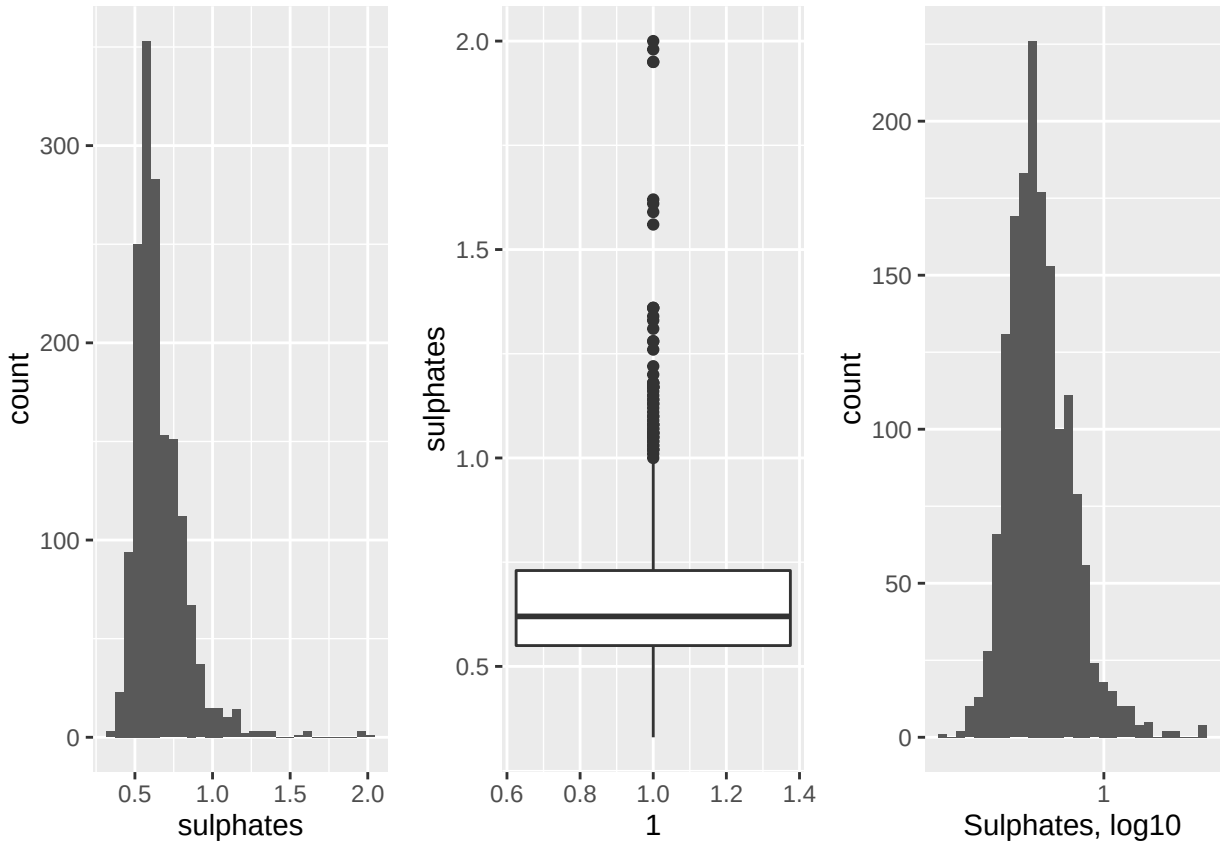

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log10(df$residual.sugar)
## W = 0.85507, p-value < 2.2e-16
```




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

```
##
##  Shapiro-Wilk normality test
##
## data:  log10(df$chlorides)
## W = 0.82836, p-value < 2.2e-16
```

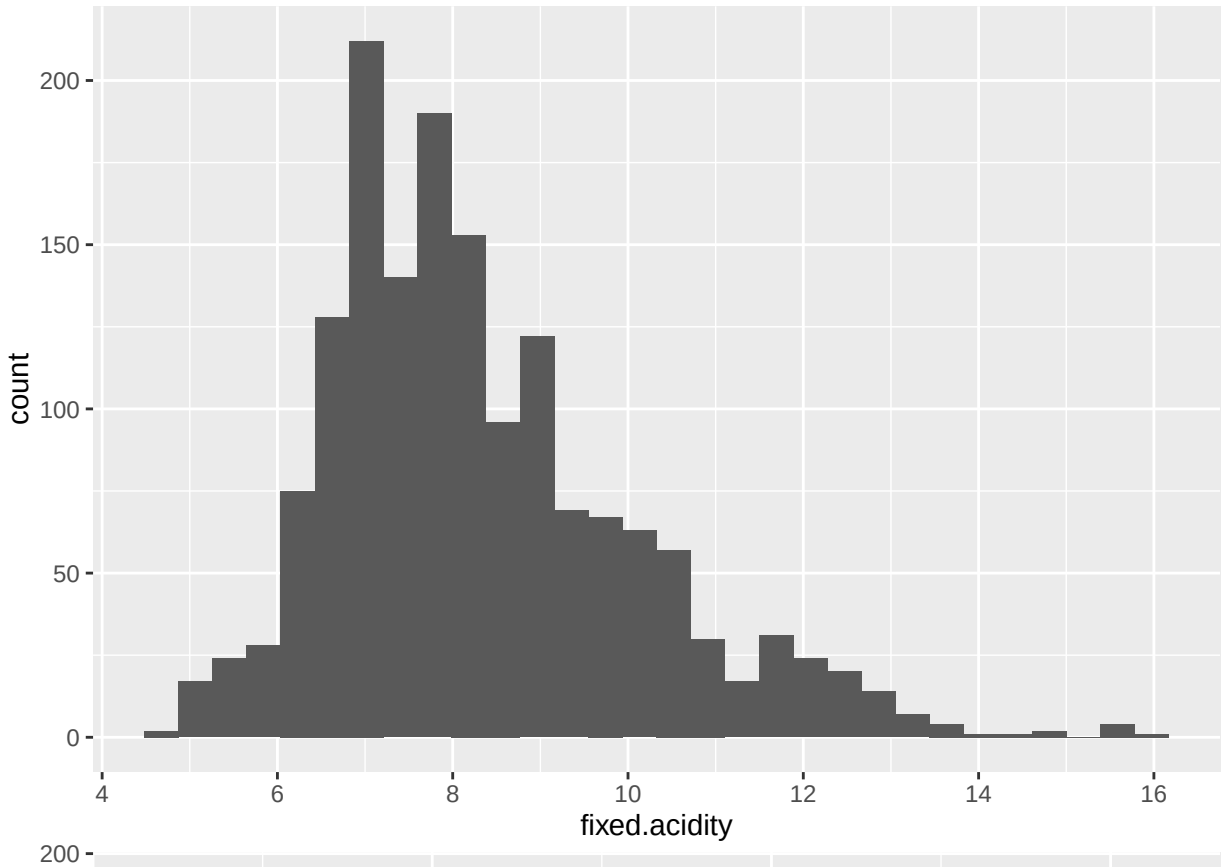


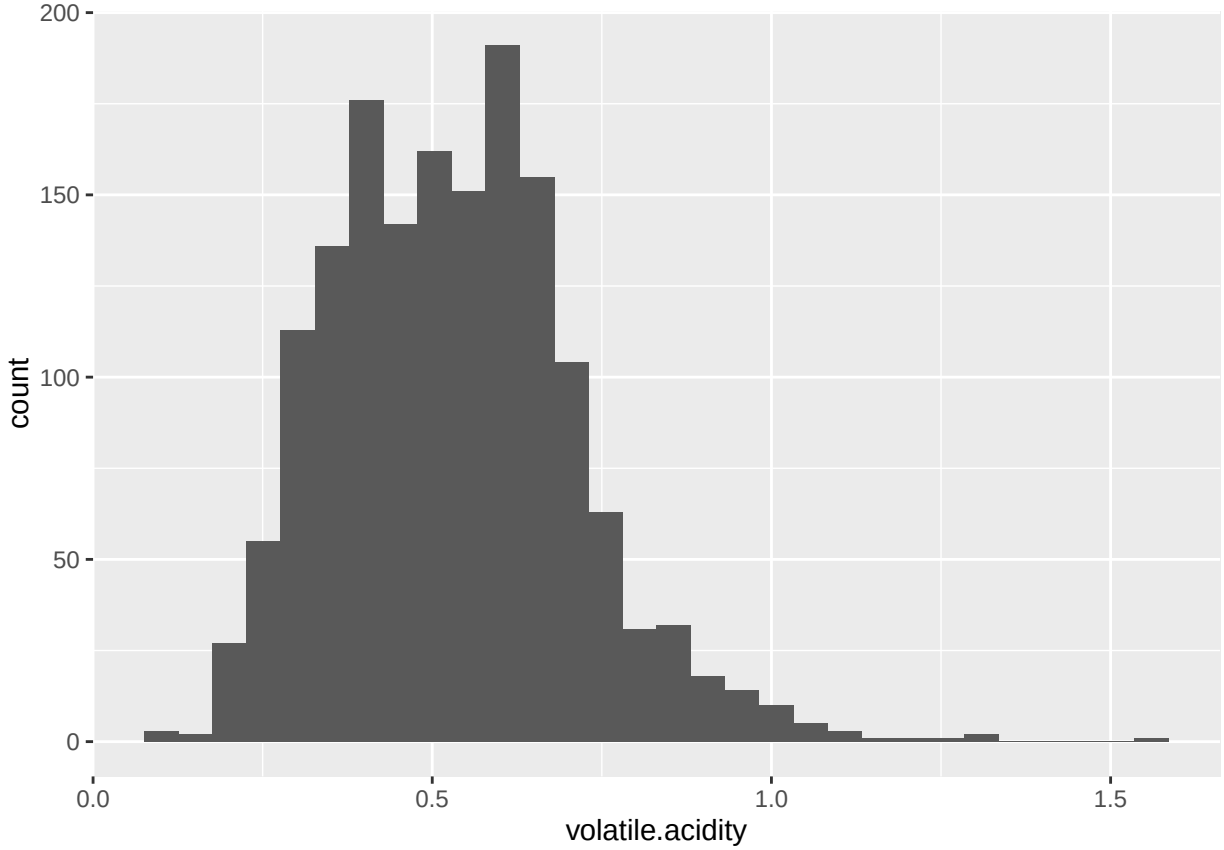

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

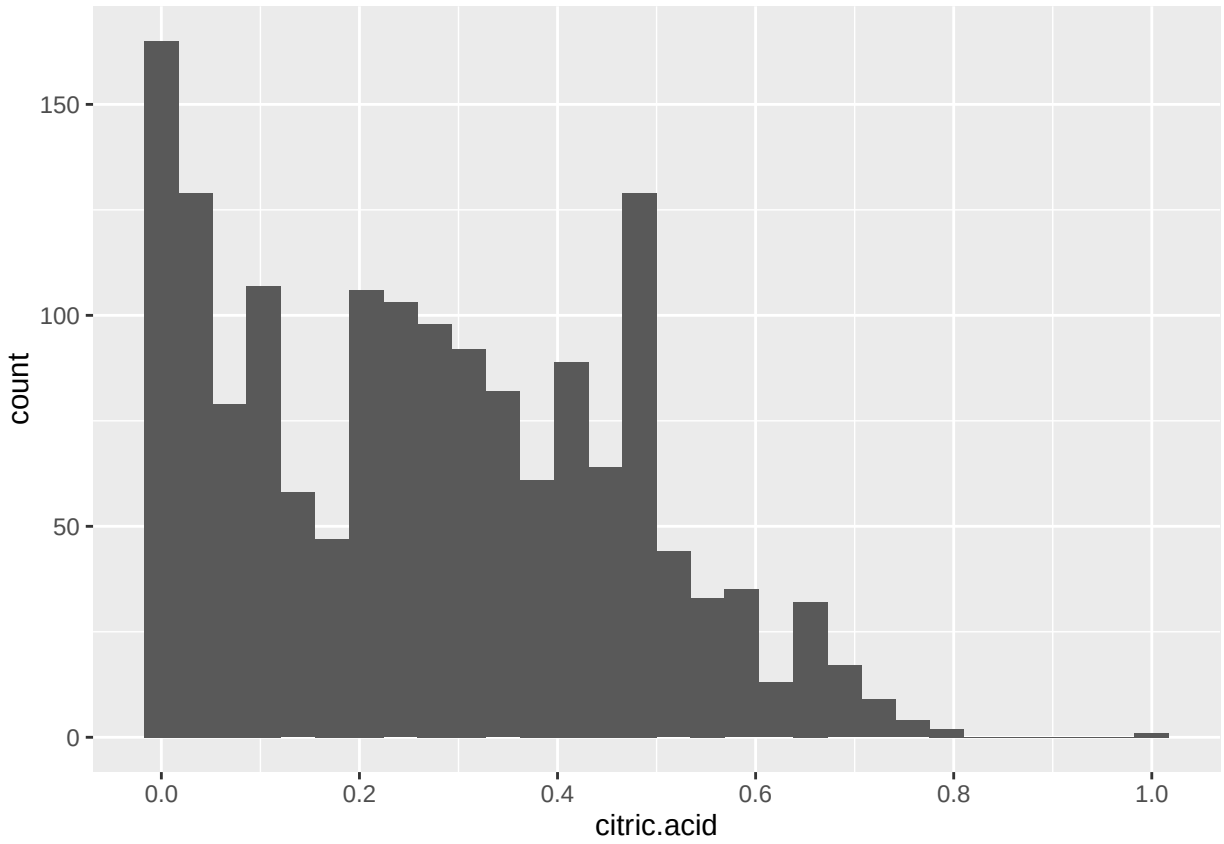
```
##
##  Shapiro-Wilk normality test
##
## data:  log10(df$sulphates)
## W = 0.95889, p-value < 2.2e-16
```

None of them appear to have a normally distributed curves as the p value is less than 0.05 (i.e. 95% confidence interval).

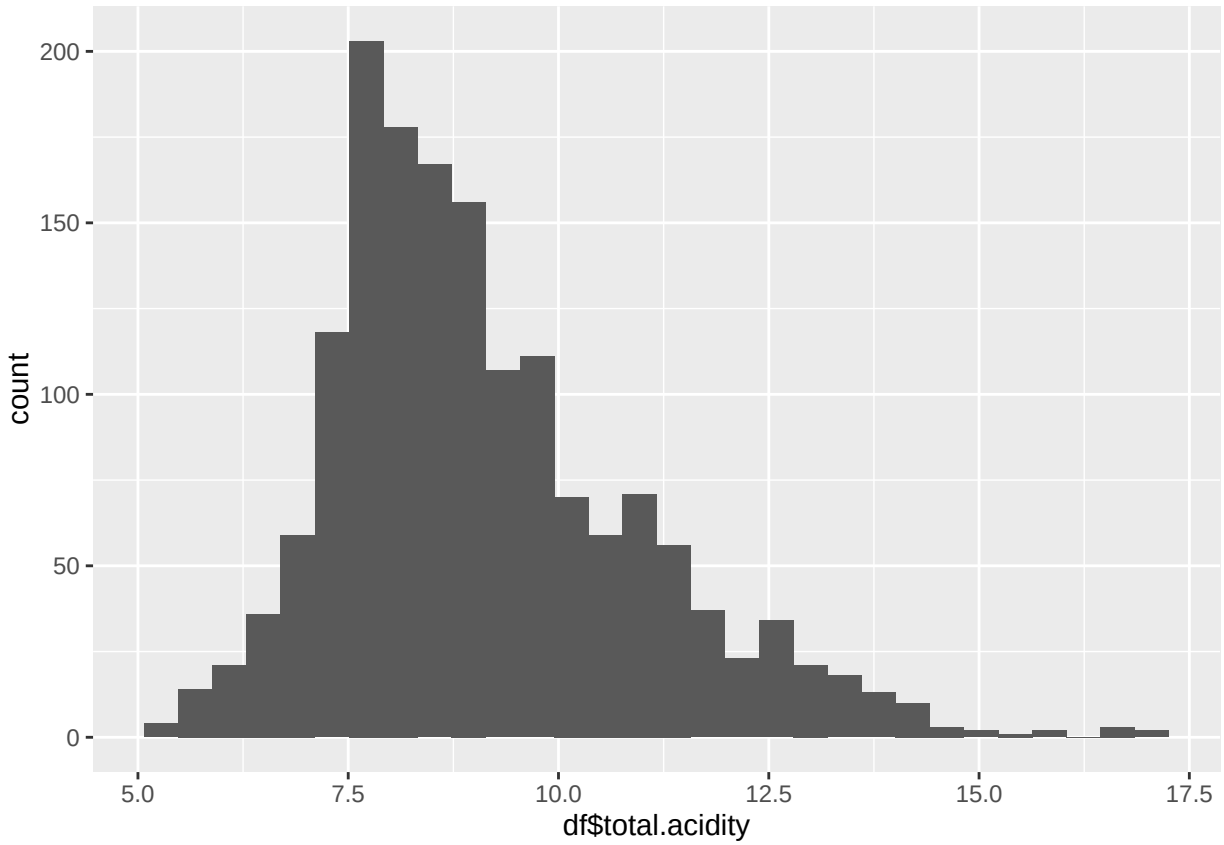
Acidity plays a vital role in determining if a wine is highy rated or not. In order to add more features to the existing dataset, we can add the acidity of all the acidic factors in the data set namely fixed.acidity, volatile.acidity and citric.acid. But first, let us examine the distribution scale of the acids.


count
200
150
100
50
0
4
6
8
10
12
14
16
fixed.acidity


count
200
150
100
50
0
0.0
0.5
1.0
1.5
volatile.acidity

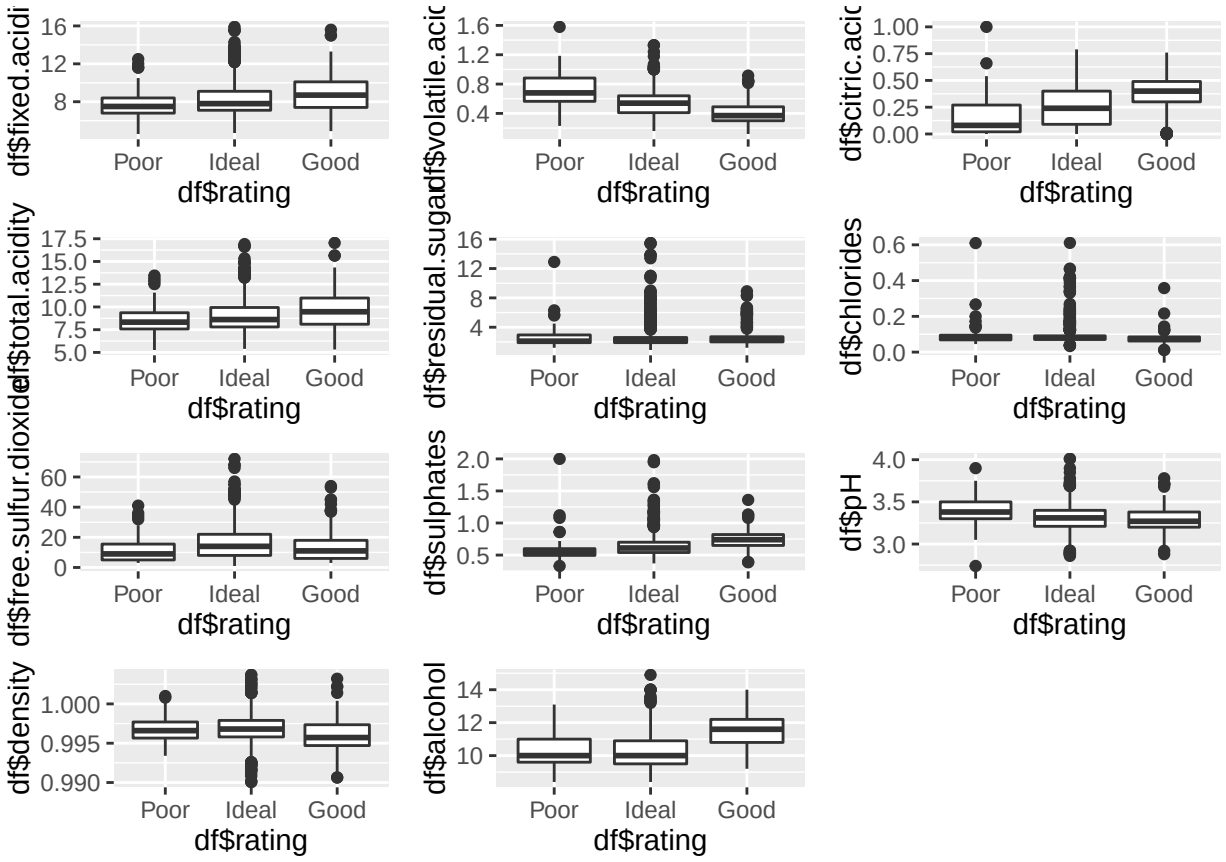


We notice that a lot of observation for the citric acid contains 0 zero count. Now defining the new total variable:

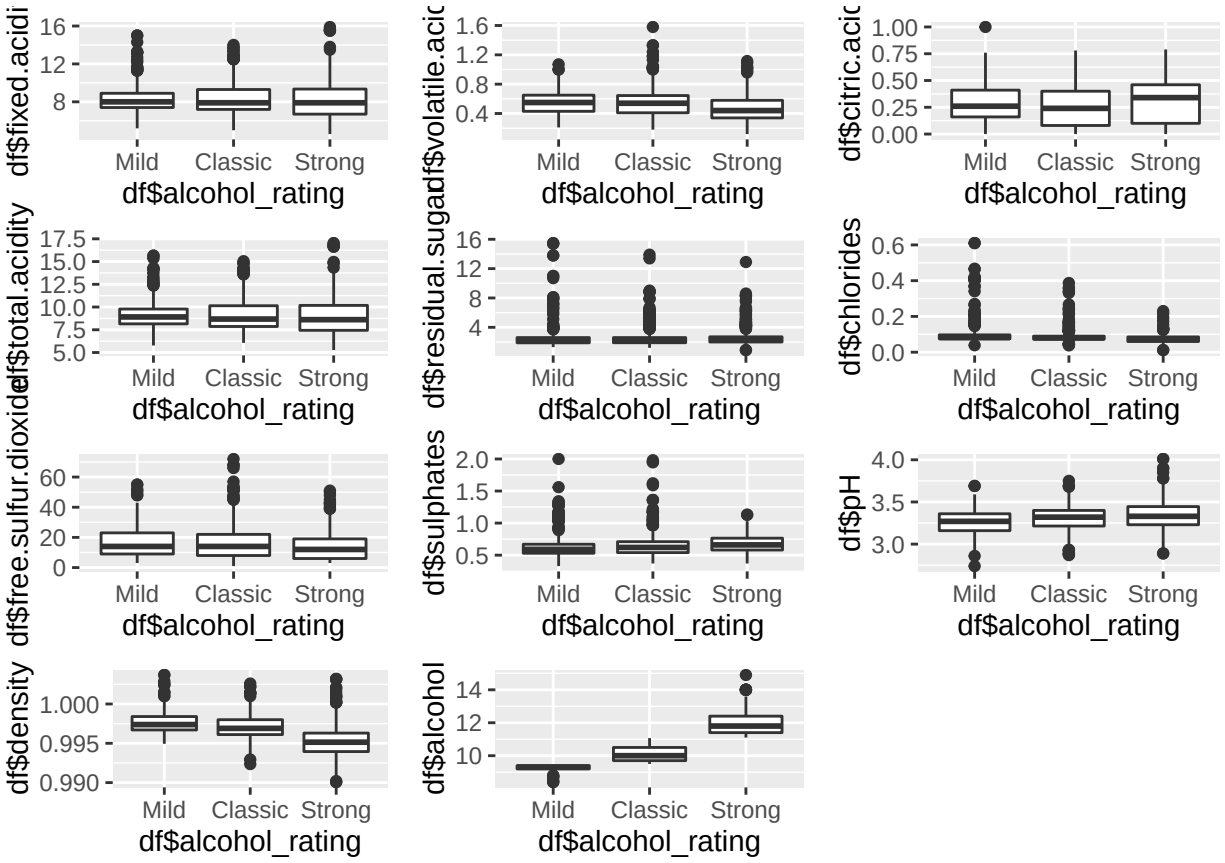


This plot turns out to be a very well formed normally distributed curve.

Bivariate Plots and Analysis:



The above boxplots help us understand what goes in the making of good wine. First observation we can make from the plots is that a 'good' rated wine has high fixed acidity. But we also observe that a lot of 'Ideal' wines also have high fixed acidity. We see that the number of outliers in the the 'Ideal' wines are very high in number, this could be explained by the fact that a lot of wine producers make inferences from the fact that 'Good' rated wines have higher fixed acidity. But, an interesting observation to note here is that the quantity of volatile acid decreases as the rating increases and this explains the above phenomenon of high fixed acid, but low rating. Another observation we can make here is that a 'Good' rated wine has higher alcohol content and lower pH value. Lower pH means that the wine is more acidic. A lot of good wines also have higher citric acid content, higher percentage of sulphates and lower density. Surprisingly, residual sugar did not have an impact on the quality of the wine because since I have brewed beer, sugar usually plays an important role in deciding the alcohol content of the beer. This was a learning curve! :)



The above boxplots were used to see if alcohol content was dependent on any other factor than the ones already discussed. Good wine generally has higher alcohol content and higher alcohol content usually follow the same pattern as good wines.

Correlation:

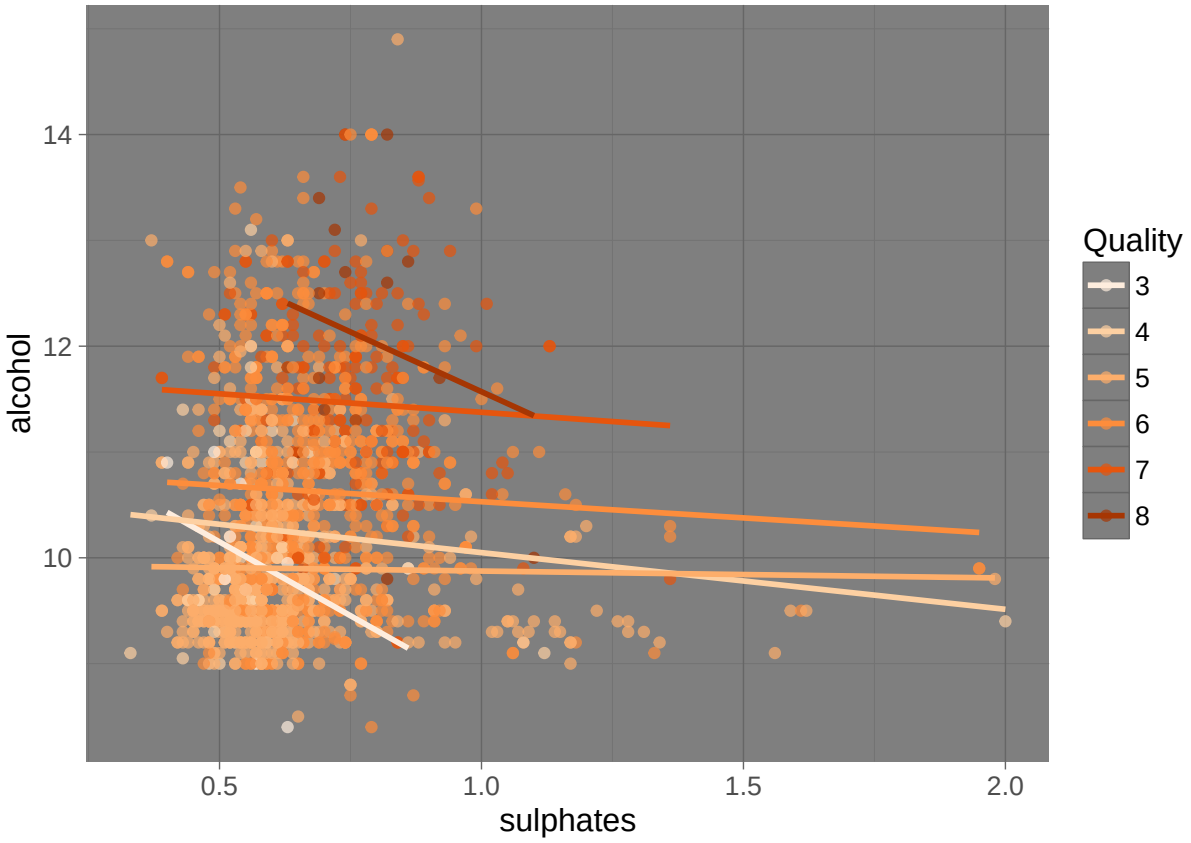
Below we calculate the correlation between quality and other variables. This would give us mathematical reasoning on what factors affect the quality of the wine.

```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165          -0.39055778           0.22637251
##        total.acidity log10.residual.sugar      log10.chlordies
##           0.10375373           0.02353331          -0.17613996
##  free.sulfur.dioxide total.sulfur.dioxide              density
##          -0.05065606          -0.18510029          -0.17491923
##                   pH      log10.sulphates              alcohol
##          -0.05773139           0.30864193           0.47616632
```

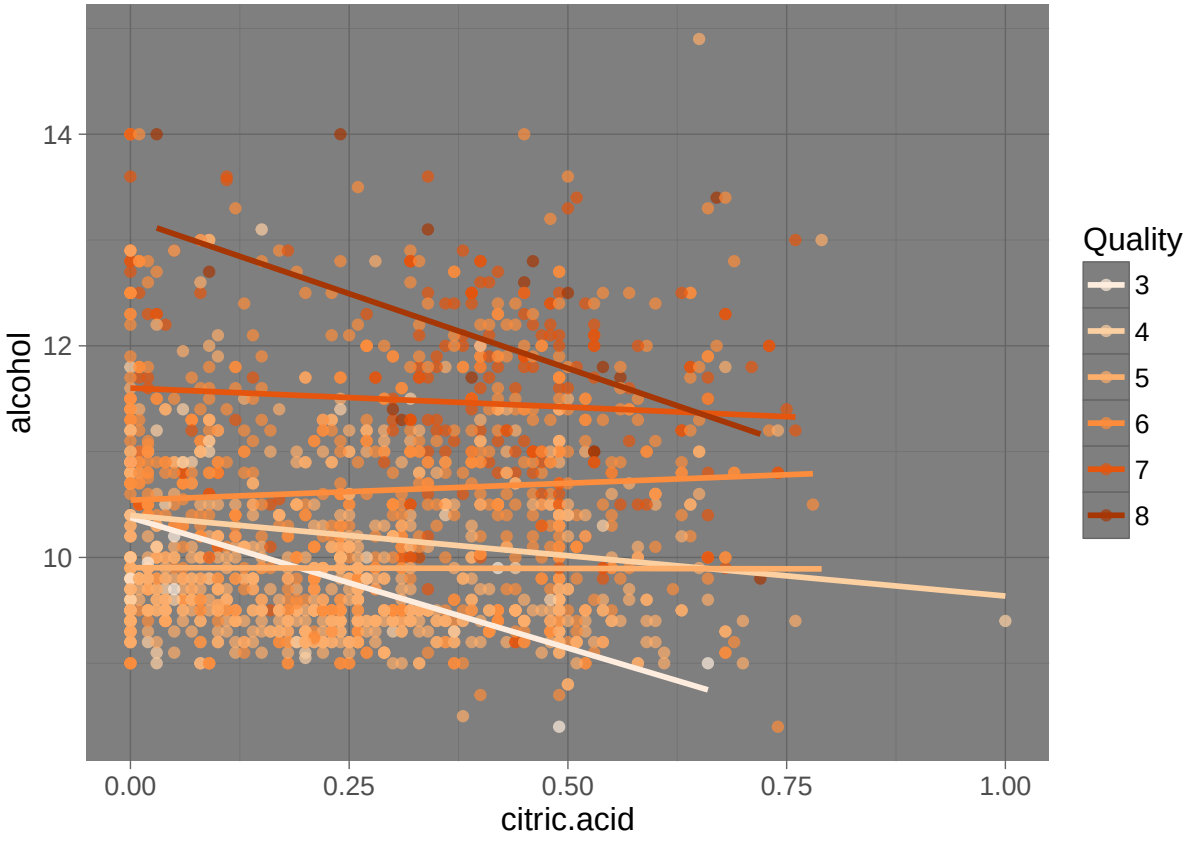
From the above experiment we can conclude that there are strong correlations between the quality of wine and:

1. Volatile Acid
2. Log10 Sulphates
3. Alocohol
4. Citric Acid

Let us try to examine if we can analyse from the below color brewed plots:

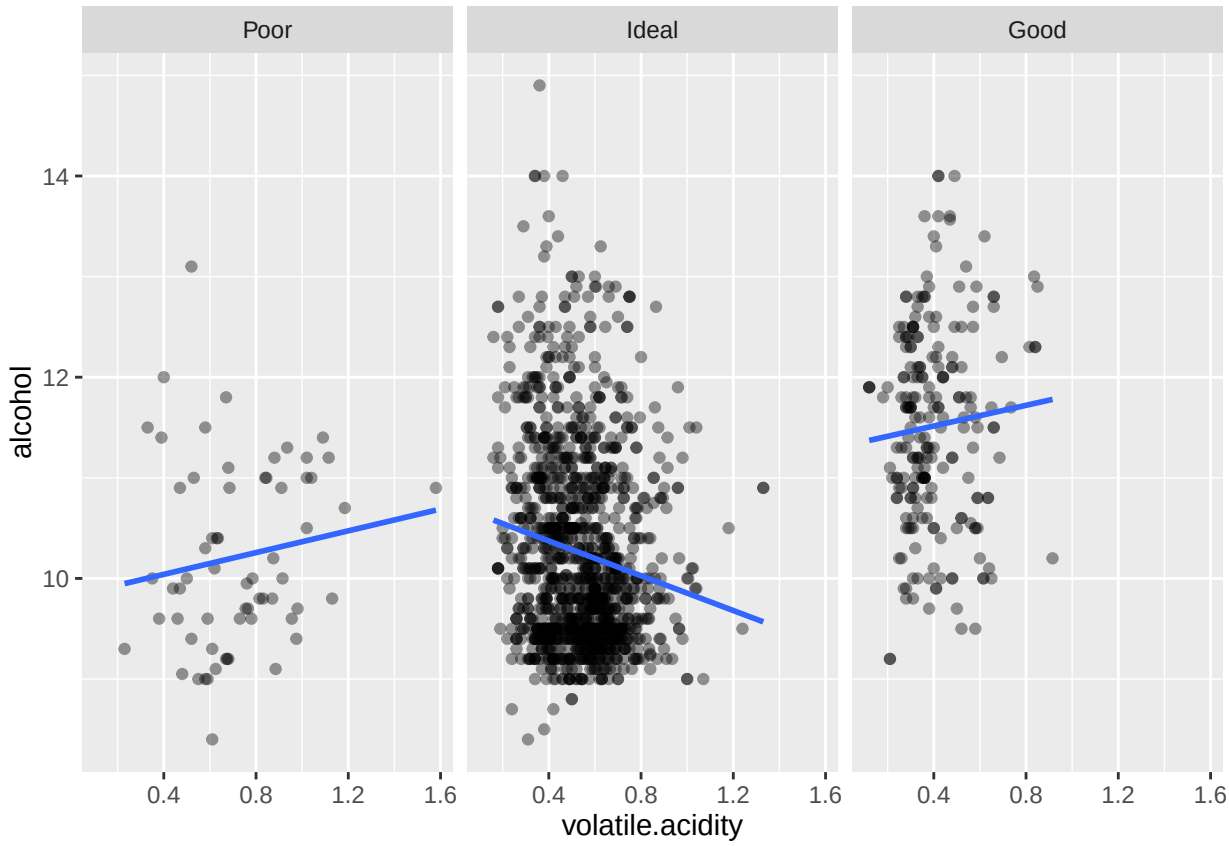


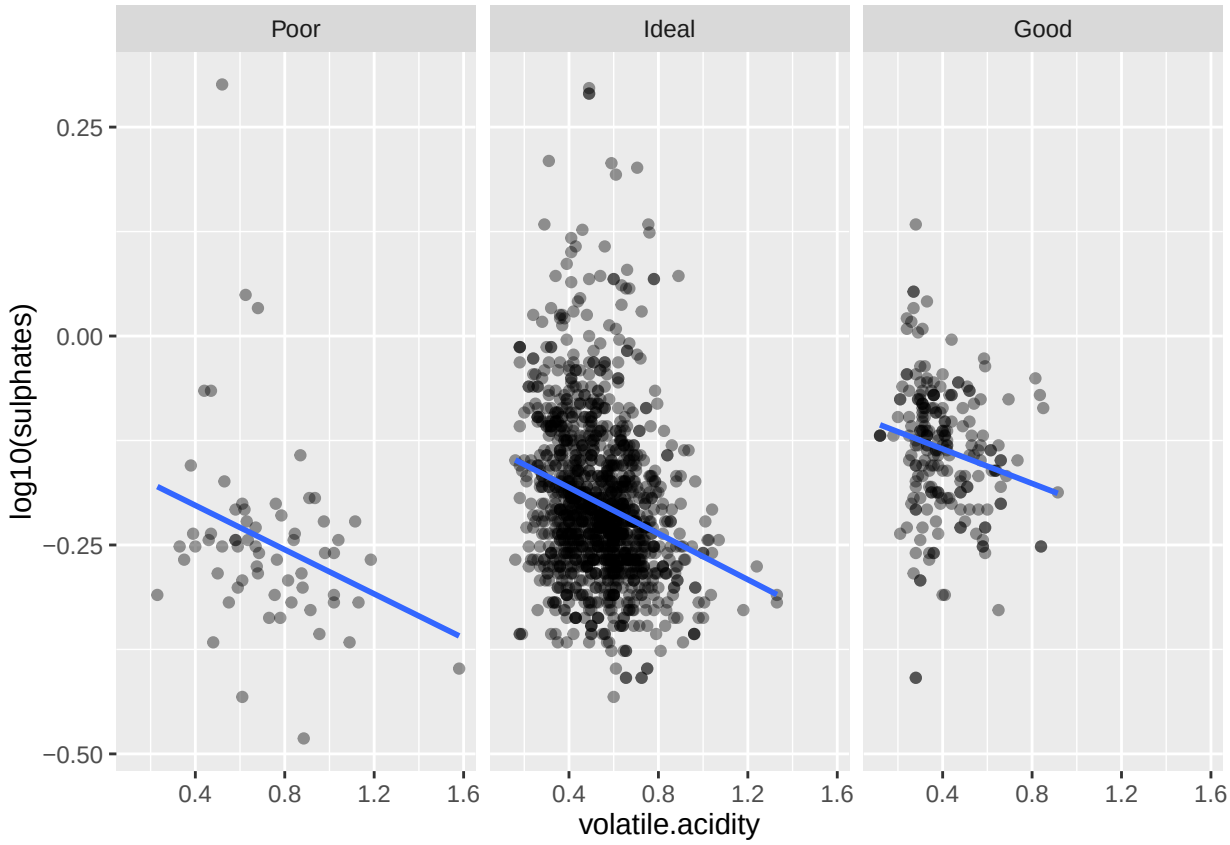

```
##
##  Pearson's product-moment correlation
##
## data:  df$sulphates and df$alcohol
## t = 3.7568, df = 1597, p-value = 0.0001783
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04477906 0.14196454
## sample estimates:
##       cor
## 0.09359475
```

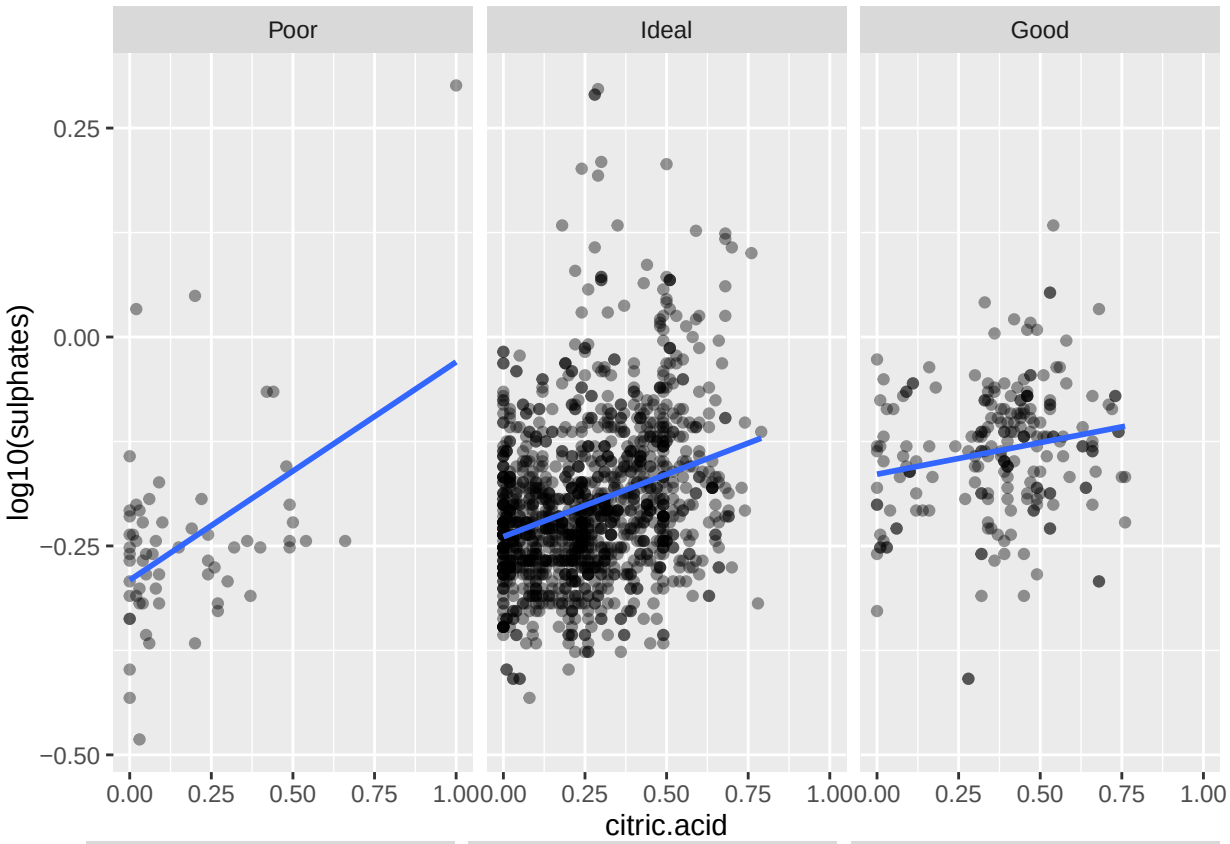


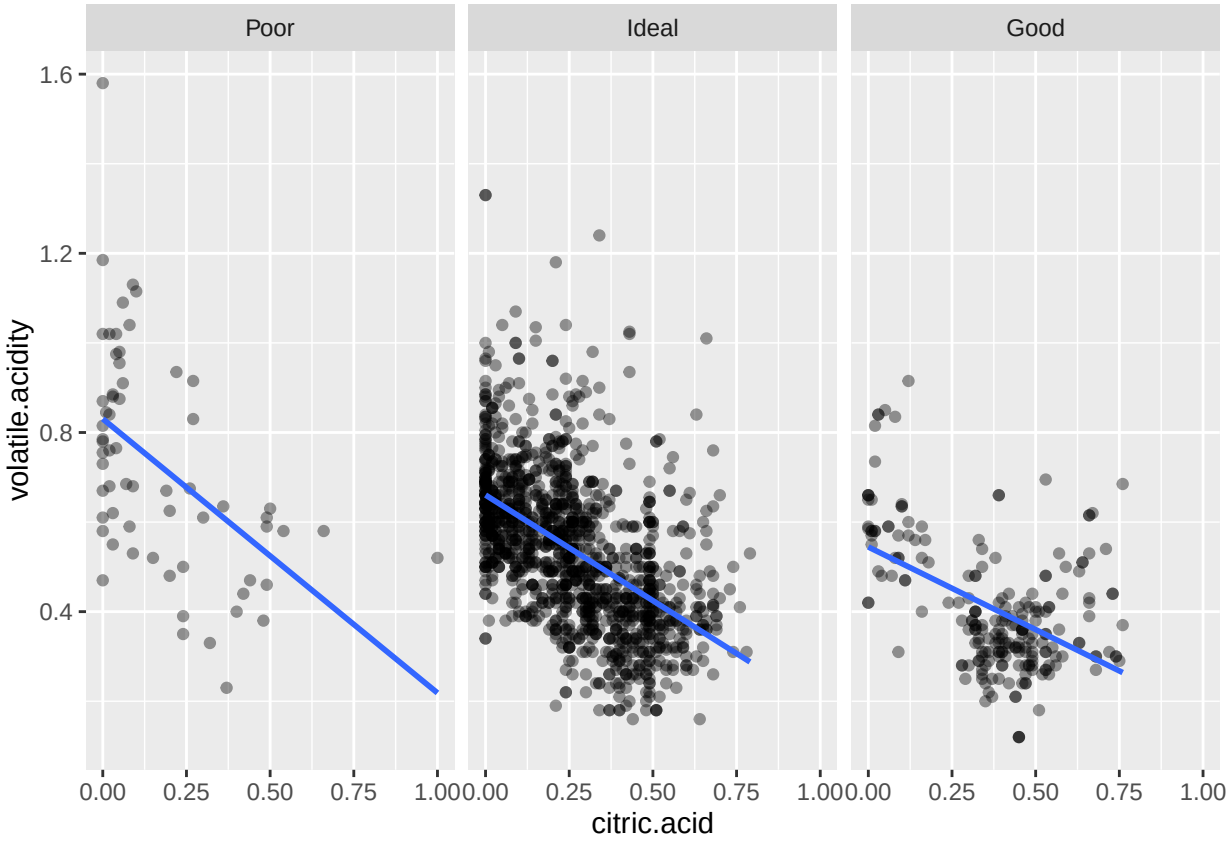

```
##
##  Pearson's product-moment correlation
##
## data:  df$citric.acid and df$alcohol
## t = 4.4188, df = 1597, p-value = 1.059e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.06121189 0.15807276
## sample estimates:
##       cor
## 0.1099032
```

The above plot is not self-explanatory but can sort of make a conclusion sayong that good quality wines have very low amounts of sulphates. The graph with citric acid is not very conclusive.The correlation helps us understand what goes in making the alcohol content higher in an alcohol. Since the correlation is very weak, we cannot make a call here. Now, I'll plot the graphs with facet as 'rating' and see what analysis we can make out of the dataset:


Poor
Ideal
Good
alcohol
14
12
10
0.4
0.8
1.2
1.6
volatile.acidity

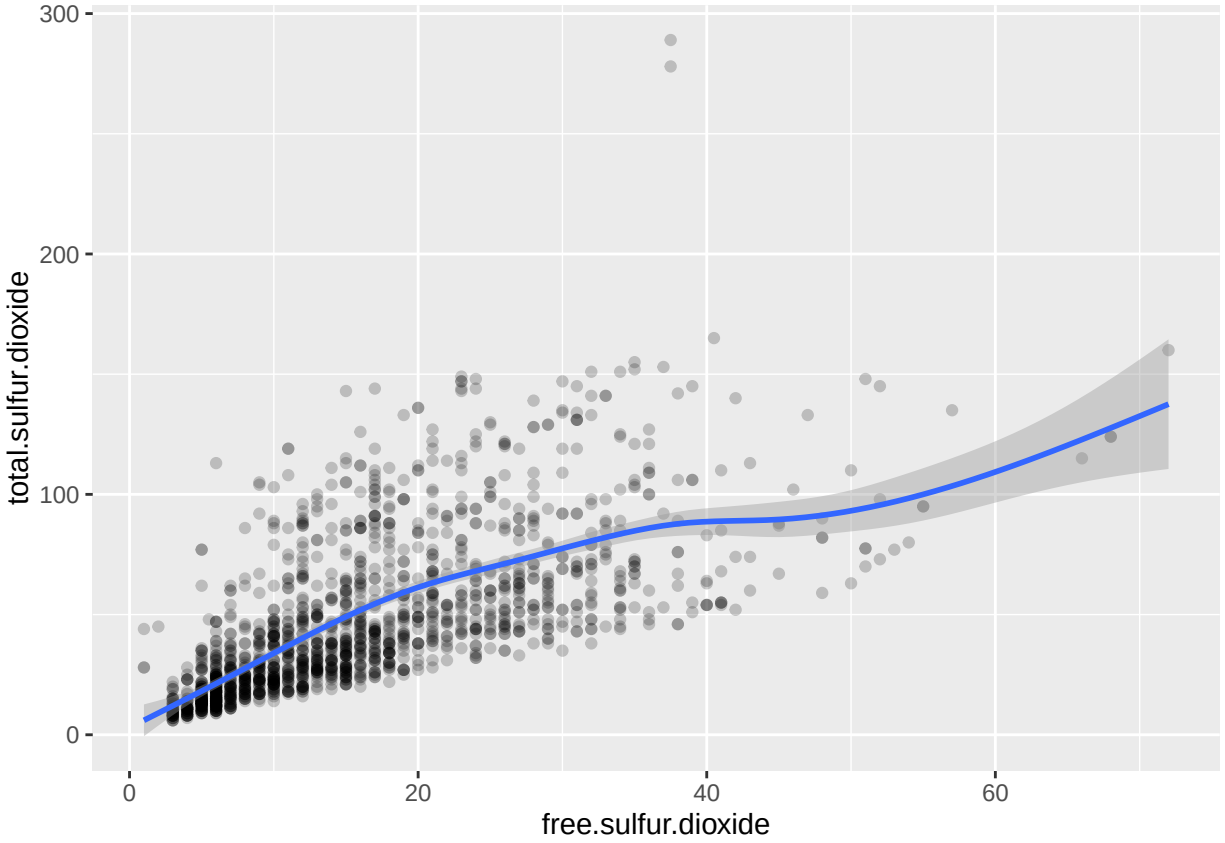

Poor
Ideal
Good
log10(sulphates)
0.25
0.00
−0.25
−0.50
0.4
0.8
1.2
1.6
volatile.acidity





Any sort of analysis is unclear from the above scatterplots. Only one thing that stands out is that volatile acid

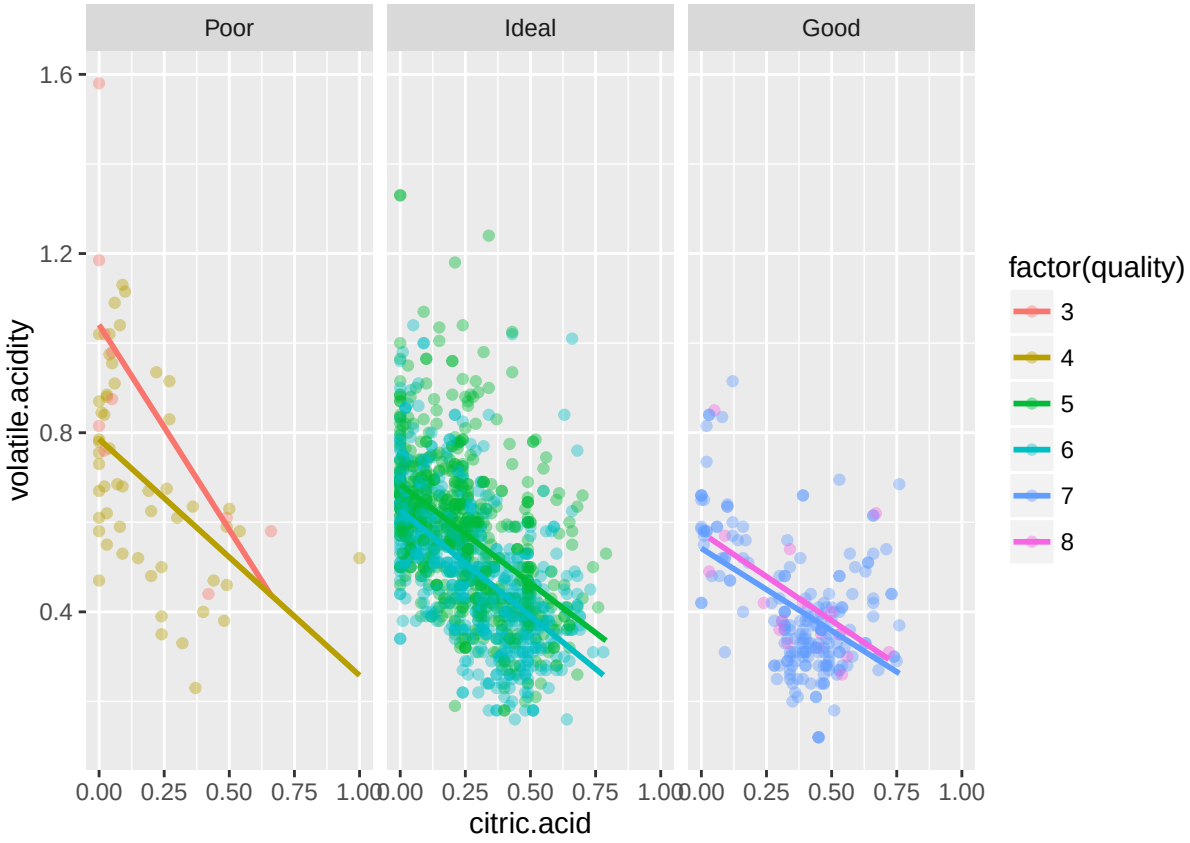
and citric are inversely proportional for a good rated wine. The regression line makes it easier to understand the plot and make our hypothesis. From the dataset, we can hypothesize that both the sulphur variables will be correlated to each other

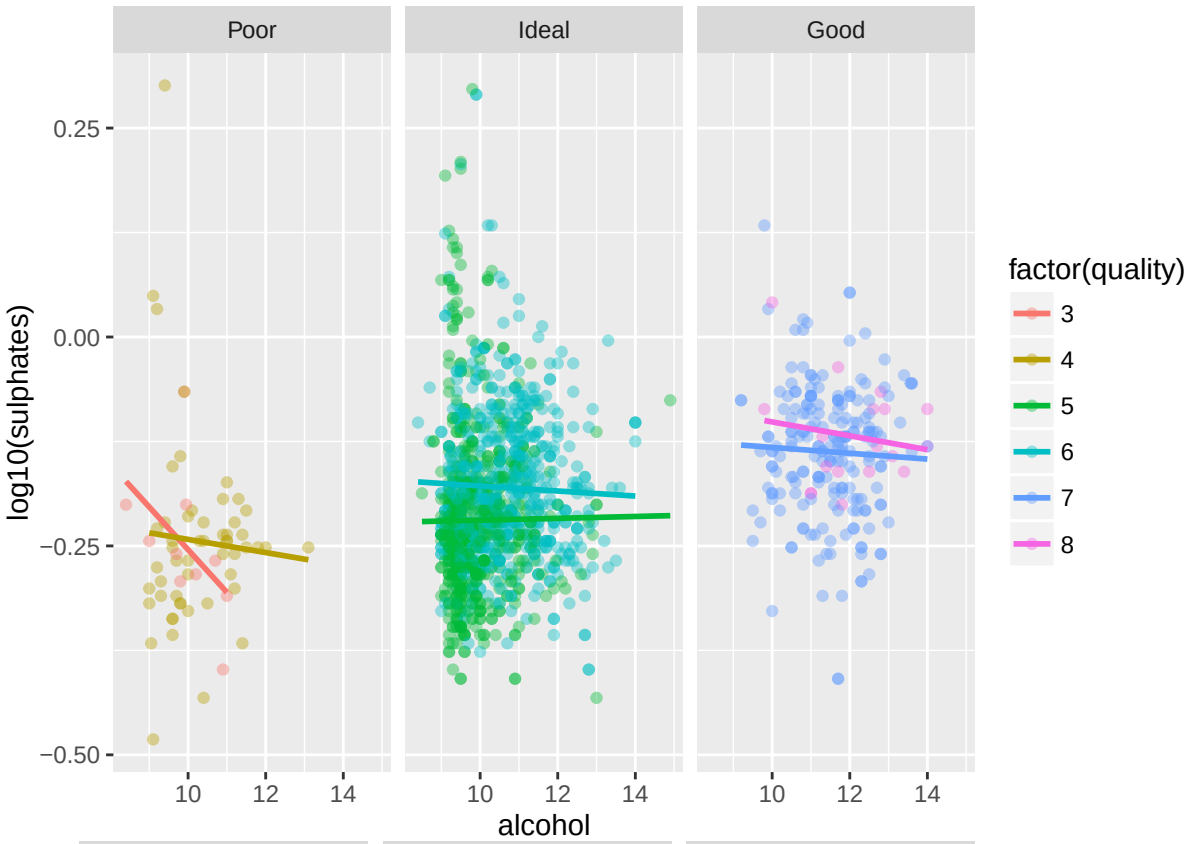


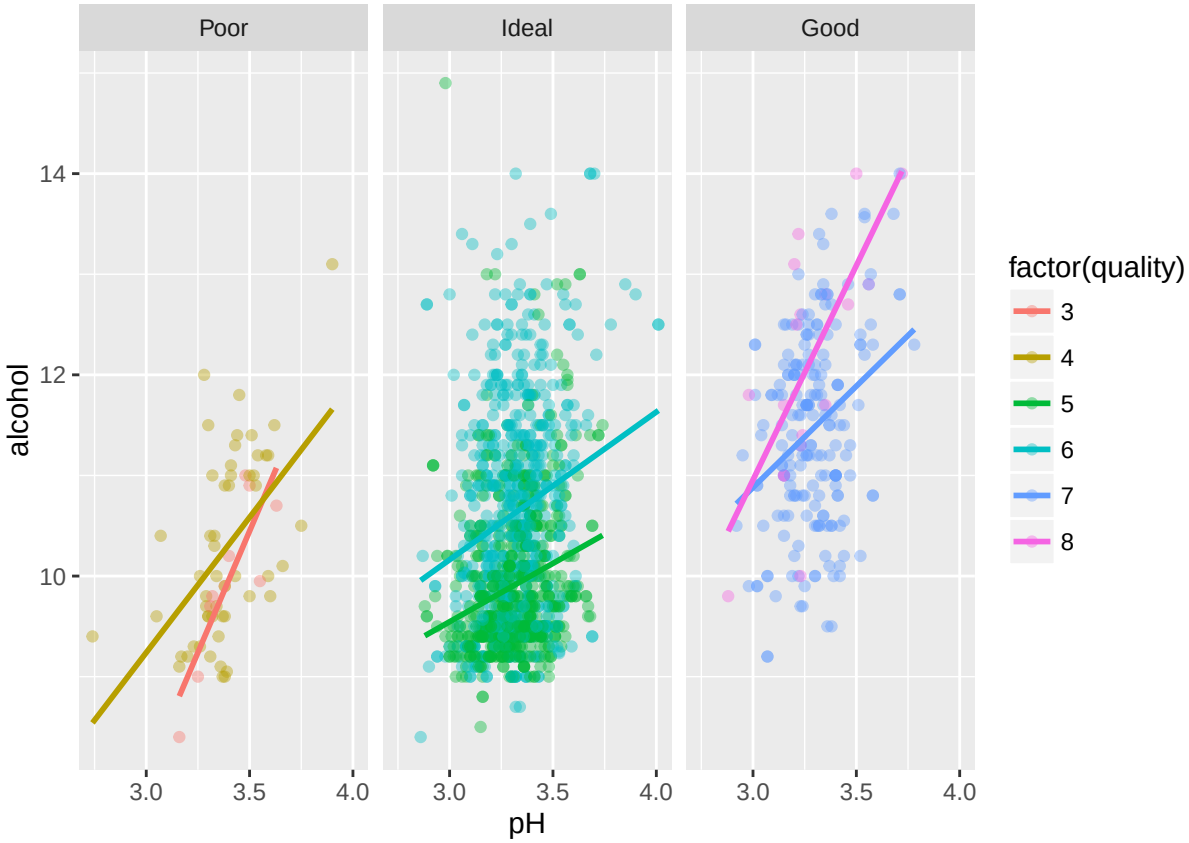
And from the plot, we get a very nice classifier that supports our hypothesis. We can also check for the correlation between the two variables:

```
##
##  Pearson's product-moment correlation
##
## data:  df$free.sulfur.dioxide and df$total.sulfur.dioxide
## t = 35.84, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6395786 0.6939740
## sample estimates:
##       cor
## 0.6676665
```

Multivariate Analysis:

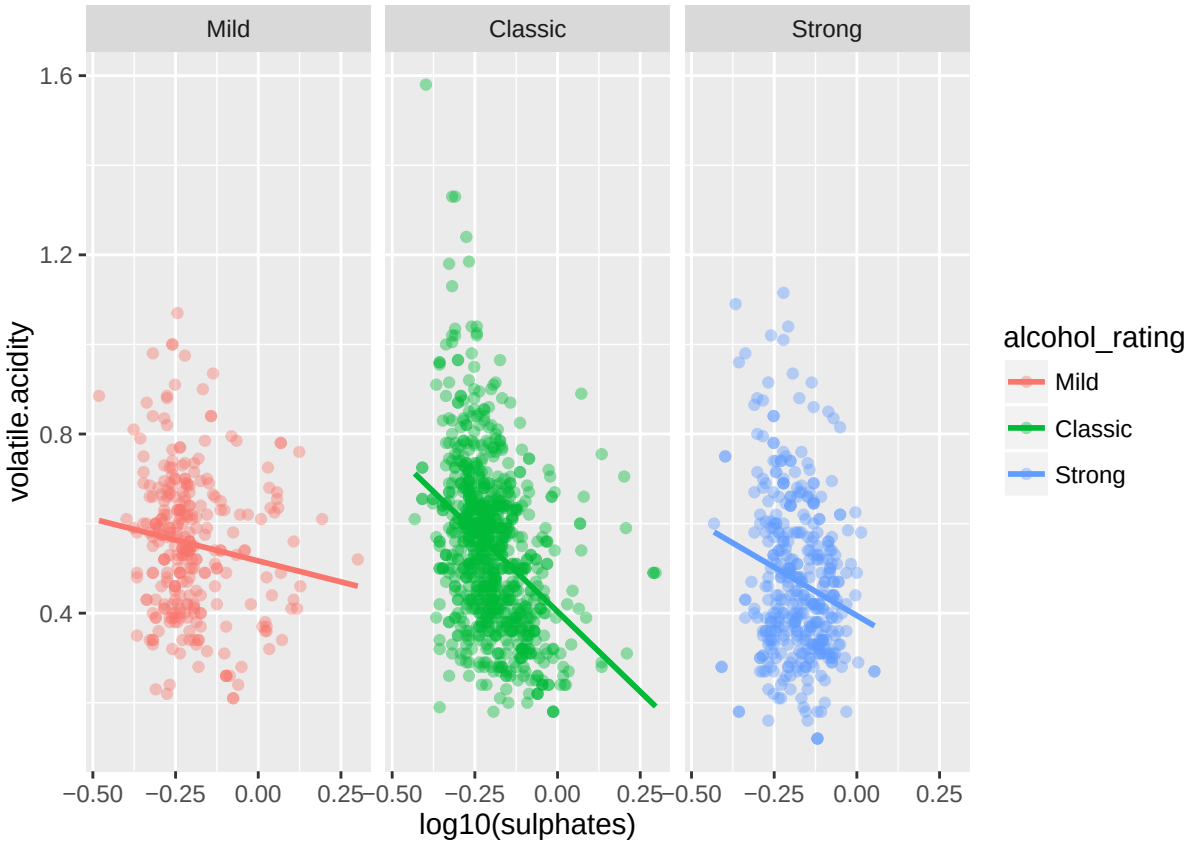


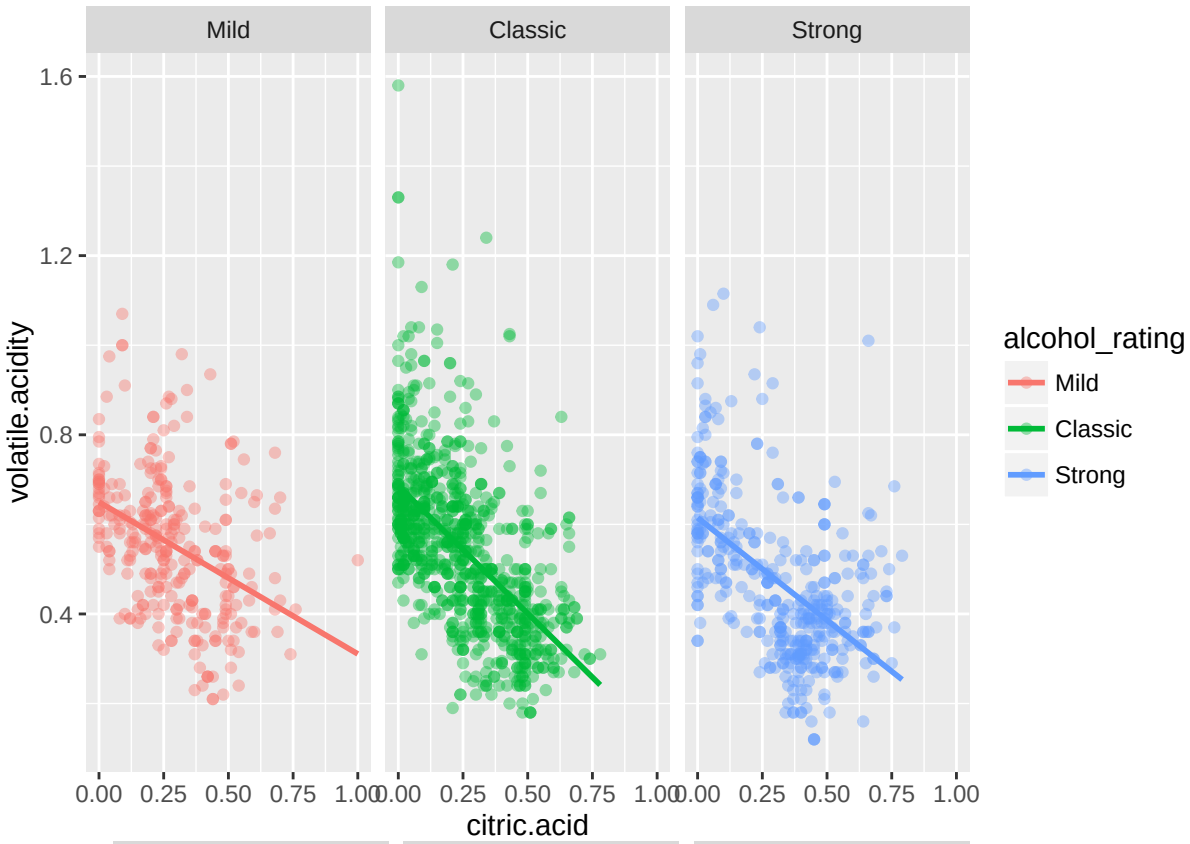


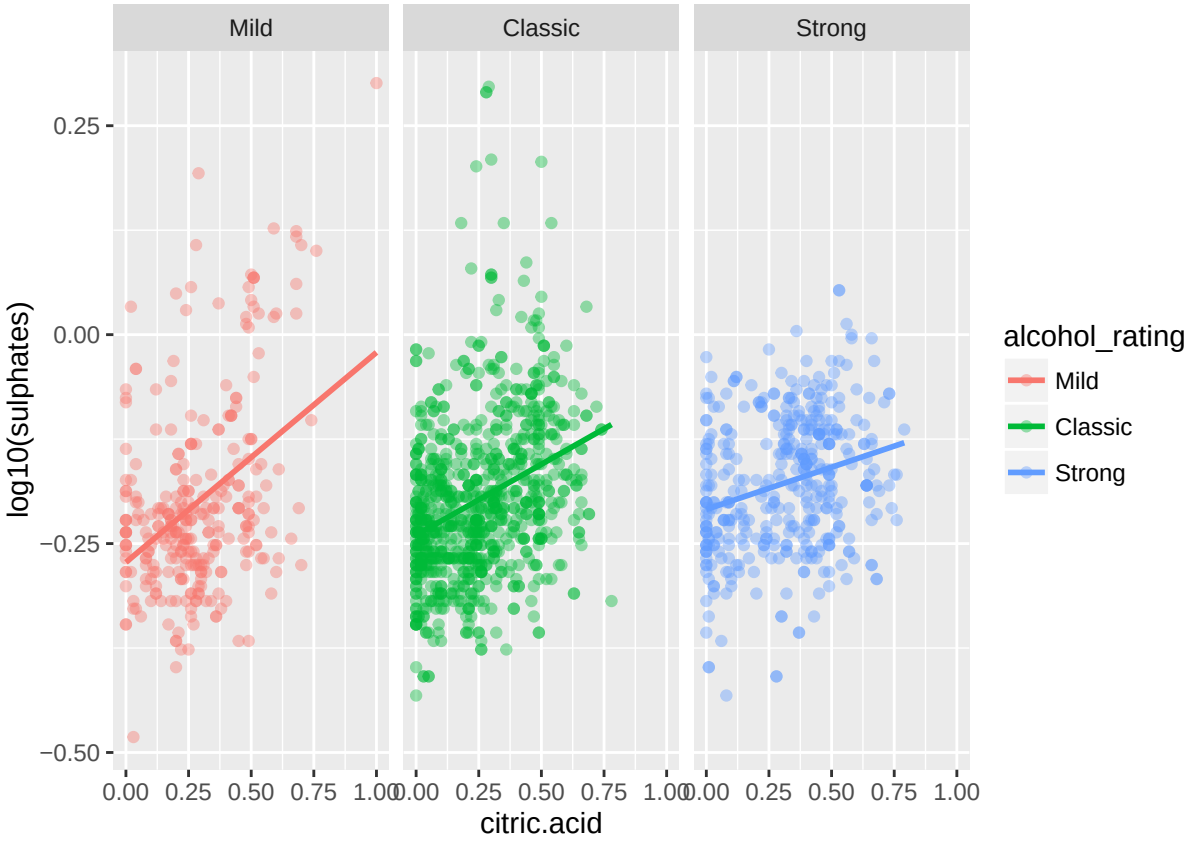


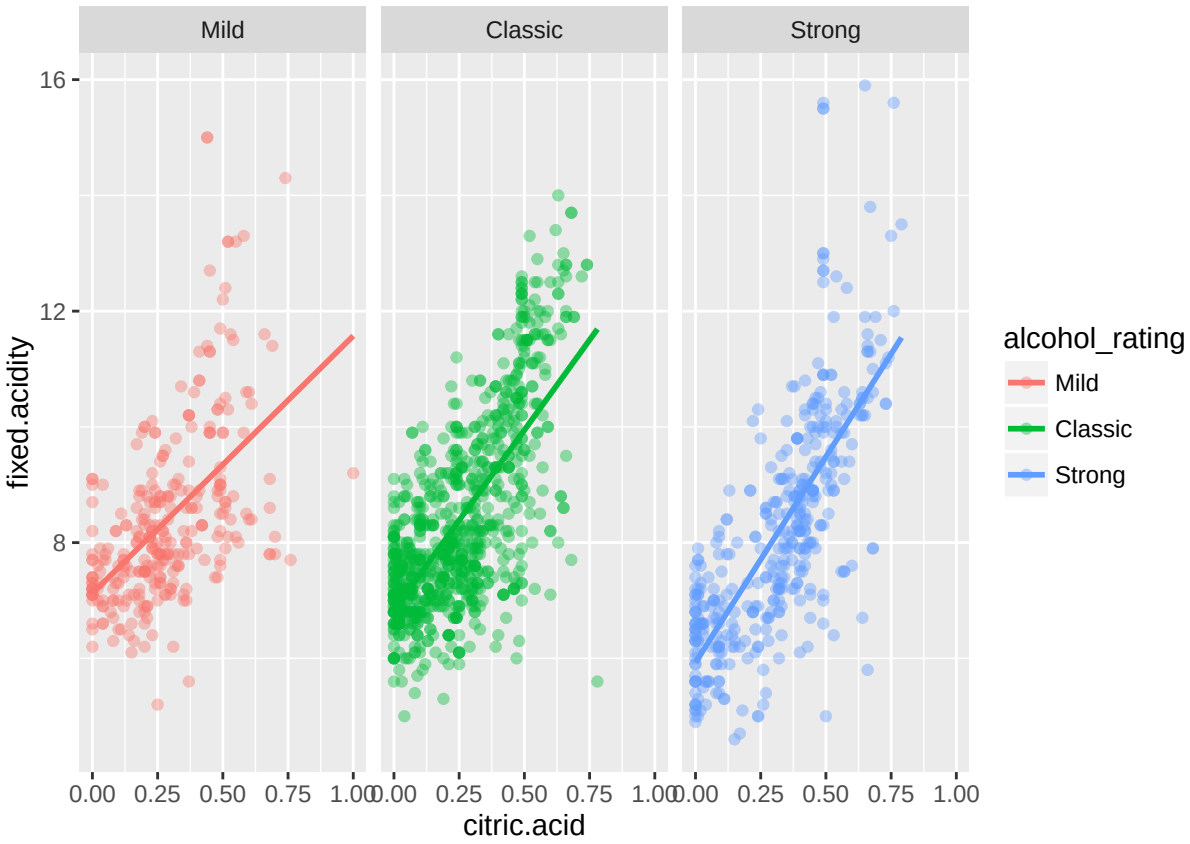
I have chosen variables based on their correlation score and produced the plots below. This clearly shows

that a good rated wine has high citric acid, low volatile acidity, higher sulphate and alcohol content. To see how acidity affects the wine's alcohol content, I plotted the below graphs:




Mild
Classic
Strong
volatile.acidity
1.6
1.2
0.8
0.4
0.00
0.25
0.50
0.75
1.00
citric.acid
alcohol_rating
Mild
Classic
Strong


Mild
Classic
Strong
log10(sulphates)
0.25
0.00
−0.25
−0.50
0.00
0.25
0.50
0.75
1.00
citric.acid
alcohol_rating
Mild
Classic
Strong

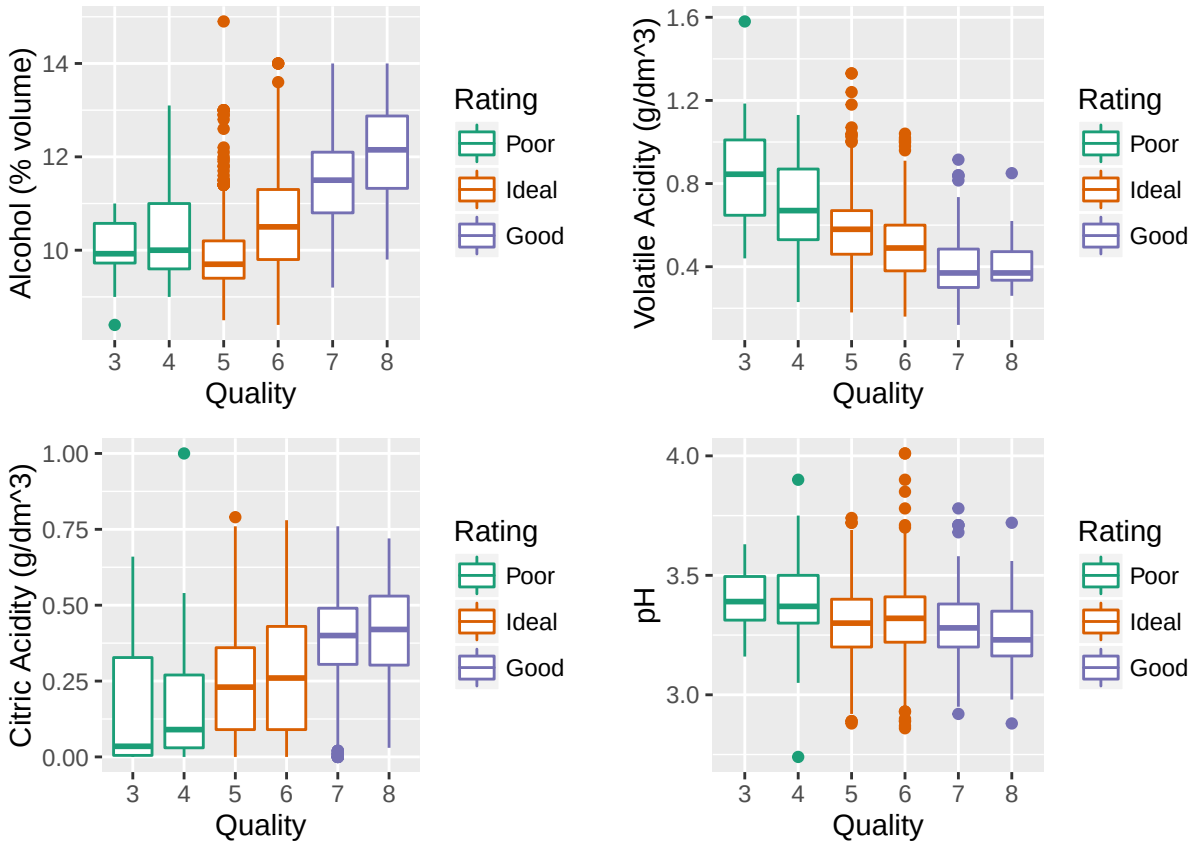


We get similar results like high citric acid, low volatile acidity all add to alcohol content.

Final Plots and Summary:

Plot 1:

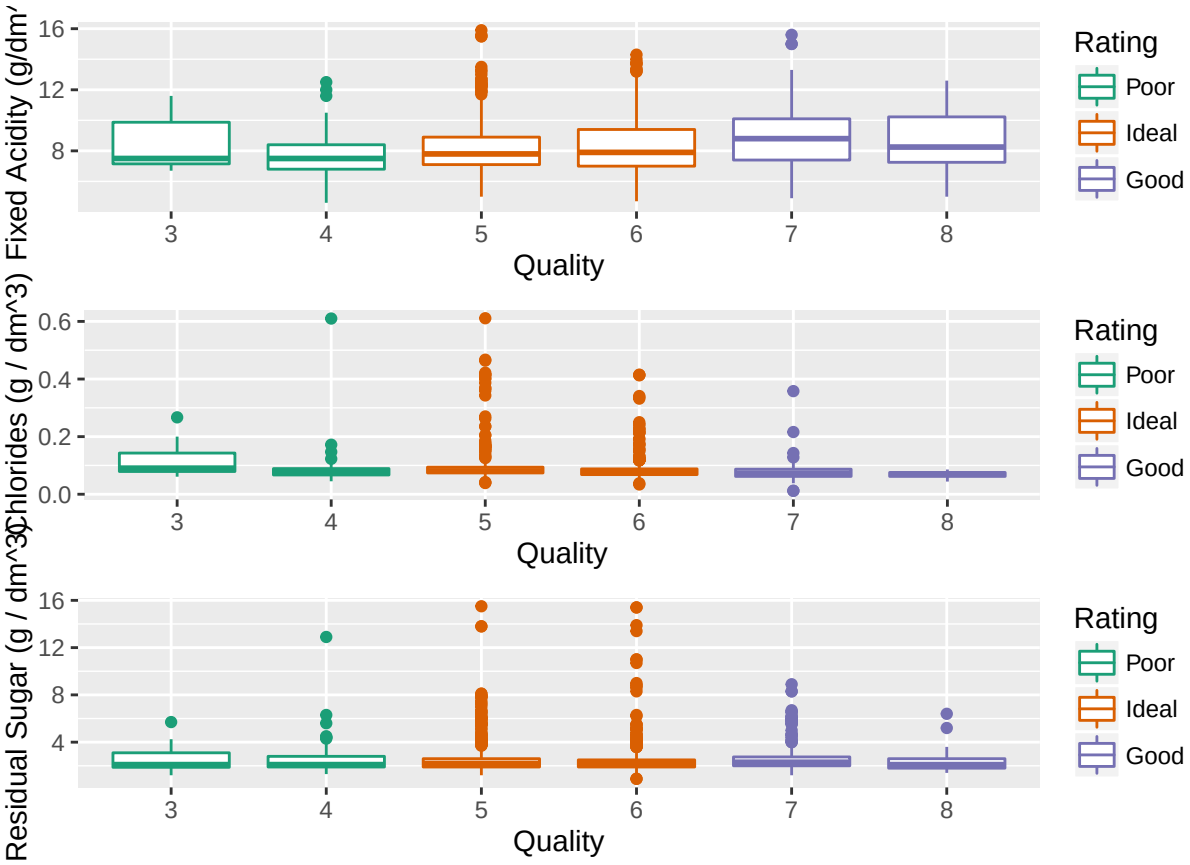
Factors Affecting Wine Quality:



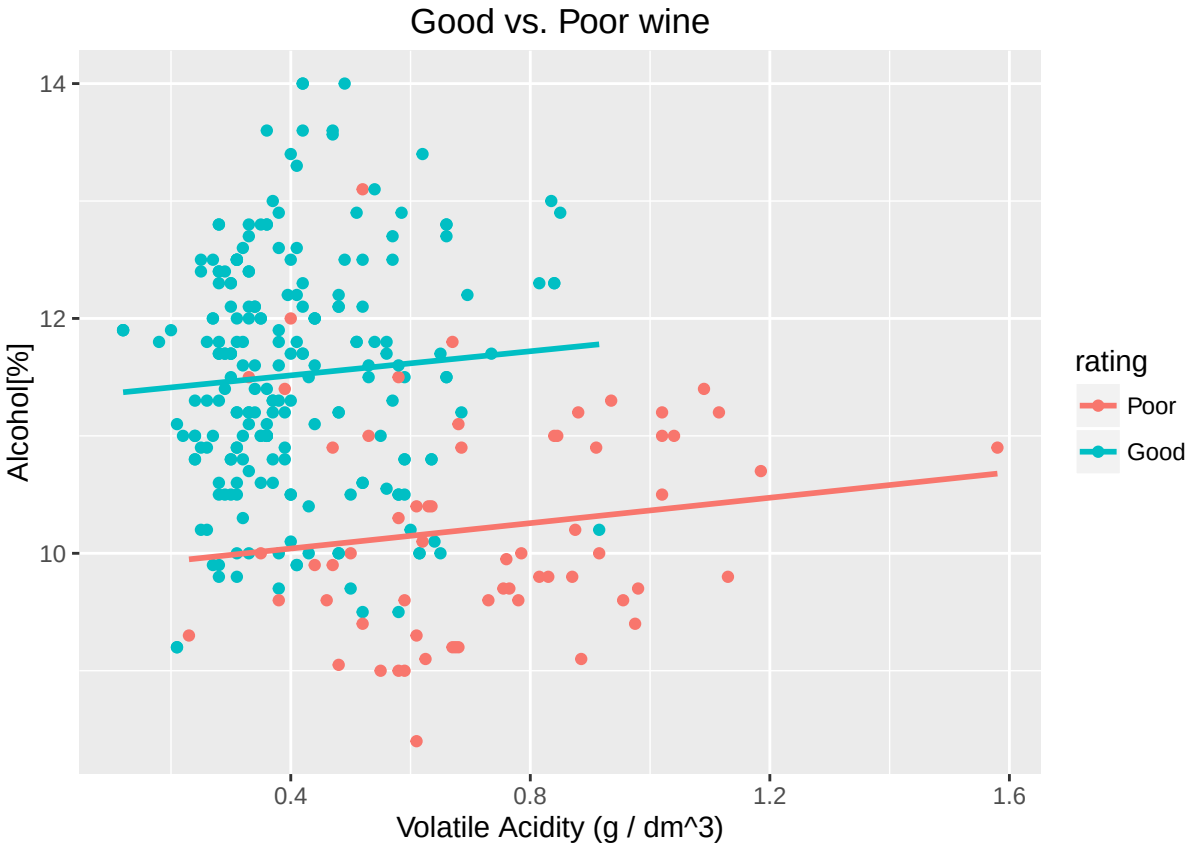
These plots were created to show the effect of alcohol content, acidity and pH on wine quality. Generally, higher acidity (or lower pH) is a characteristic of highly-rated wines. Citric acidity had a high correlation with wine quality, while fixed acid had a smaller impact. This plot helps us form a correlation that higher alcohol content usually has higher quality wine. Although the outliers in 'Ideal' section, seem to oppose this claim. This means that alcohol alone cannot justify the quality of the wine. Other boxplots demonstrate the effect of alcohol content on wine quality. Generally, higher alcohol content corresponded to higher quality wine, although the outliers and intervals negate this claim.This explains the facts that low volatile acidity, high citric acid and lower pH values all add to higher rating of a wine.

Plot 2:

Factors *not* affecting wine quality:



The above plots are designed to show what factors do not affect wine quality. Features like Fixed Acidity, Chlorides and Residual Sugar do not impact the quality of wine as the other factors. From the boxplots it is clear that these factors don't follow a pattern or have stronger correlation. Since we know what goes into making a good wine, it is also important to know what doesn't and that is why I have plotted the above boxplots. ##Plot 3: Distinguishing Good and Bad Wine:



The above plot is my final plot and it is a subset that does not contain 'Average Wine'. I'm trying to depict a picture here stating what distinguishes a good wine from a bad wine. From the graph it is quite evident that low volatile acidity and higher alcohol content produced better wines with a few outliers (exceptions).This shows us a visualy pleasing corelation that low volatile acid and higher alcohol content usually makes great wine which was also seen when we saw stronger correlation between quality-alcohol and quality-volatile acid. The plot above validates this claim.

Conclusion:

I chose this dataset because I tried brewing my own beer and that was one of the driving factors for choosing this dataset. I wanted to explore what affects the quality of wine, how acidity plays an important role and how wine brewing is an art. Although wine quality is conclusively a subjective factor and wine experts make most of the decision based on various factors, the correlations for the variables in the dataframe are within reasonable bounds. To further assist our plots, we could add inferential statistical tests that would measure our hypothesis of correlation insome way. In other words it would be interesting if we could come up with some statistical tests that can help us distinguish between 'Good' and 'Bad' quality wines. I had great fun learning and exploring this dataset. There were some stuggles along the way which included factors like me being new to R and trying to find a story with our data. This project helped me understand that we shouldn't be overwhelmed by the dataset and by starting small, we can definitely tell a story. For example, by just plotting histograms, I was able to get a great insight into the dataset. Happy Brewing!