Red Wine quality

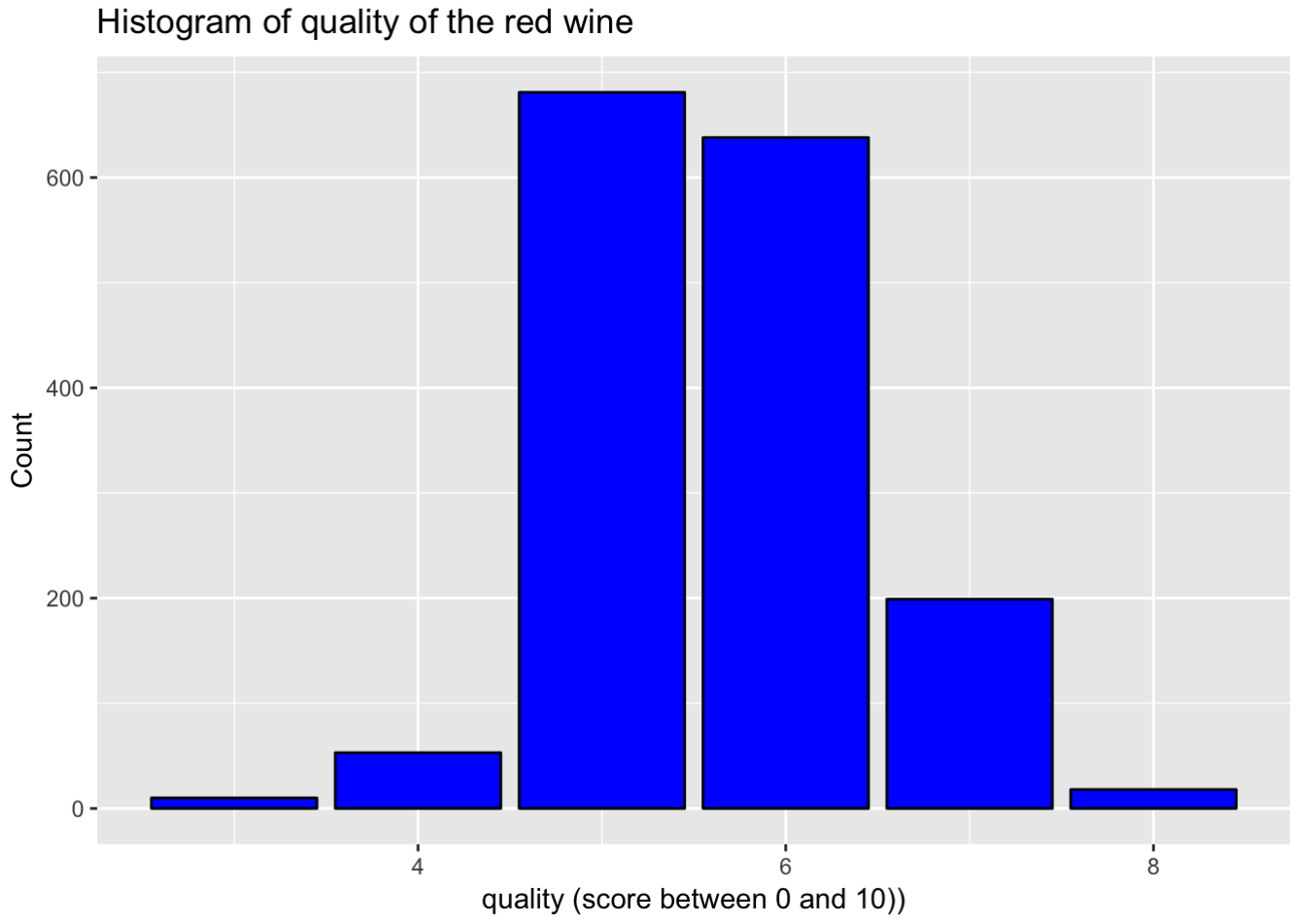
by

Shruti Tiwari

==========

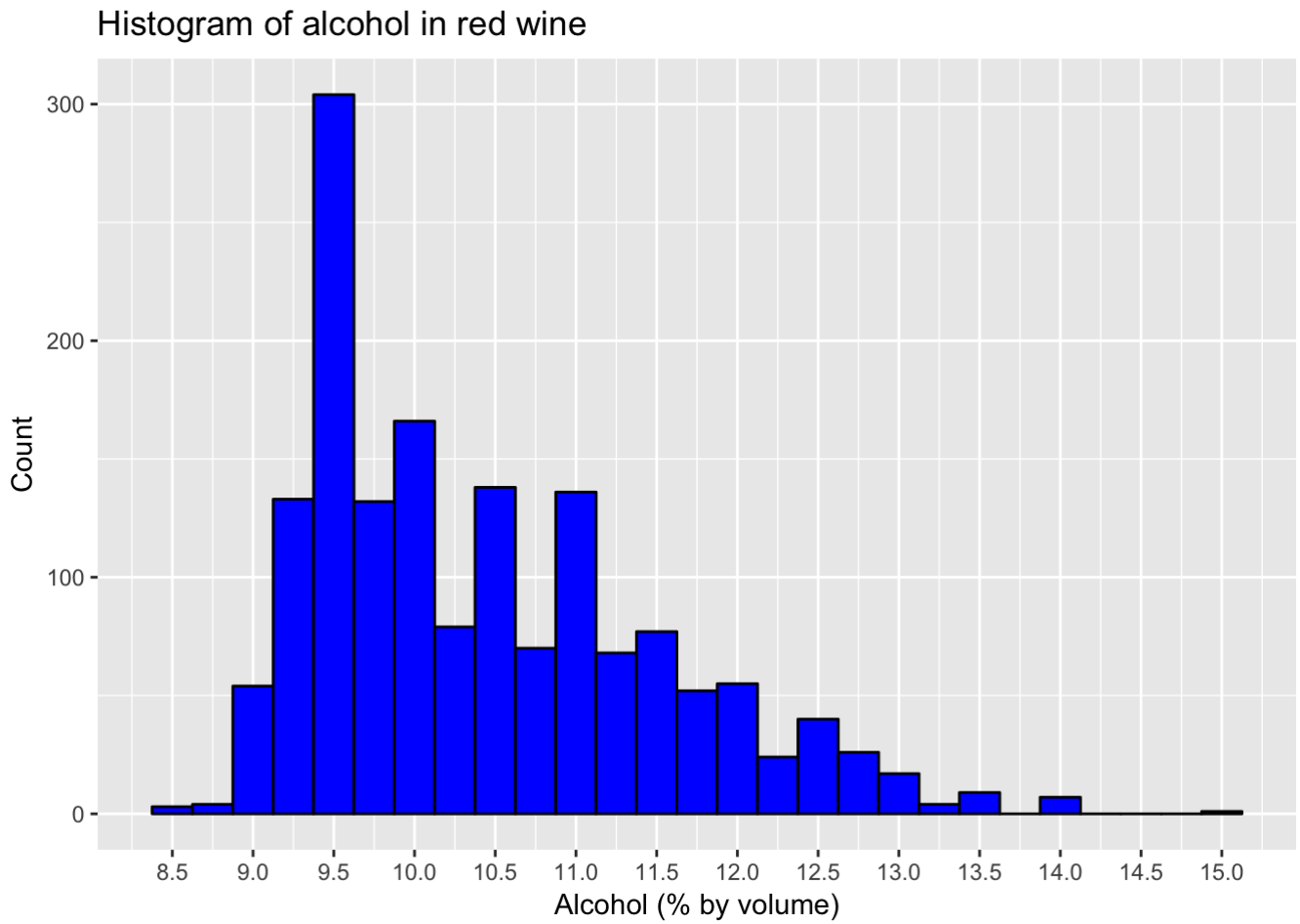
This dataset of quality of the red wine consists of 13 variables and 1599 observations[1]. The first variable X is the unique identity number which I have removed in this study. Quality is the main output feature and other 11 variables are fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, and alcohol. I have added another variable other.acid by subtracting citric.acid from fixed.acidity to study the effect of acids other than citric.acid and acetic acid. Please refer to the text file[2] for the details of these variables.

Univariate Plots Section




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

In this dataset the minimum quality is 3, maximum is at 8, median is 6 and mean is 5.64. The histogram of quality looks like normal distribution which is expected.

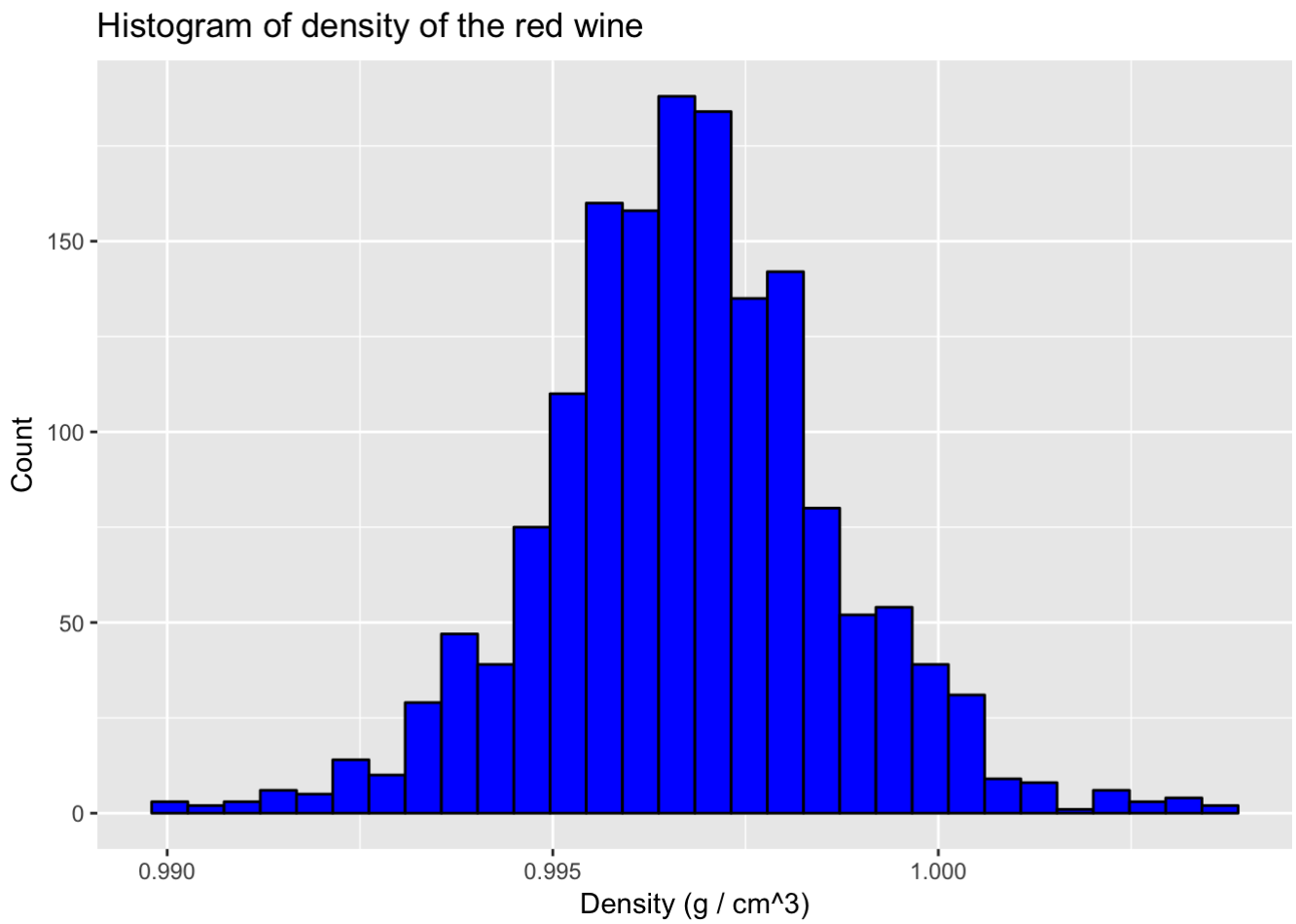



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

The histogram of alcohol shows a normal distribution with a few outliers. The summary of alcohol shows that minimum value of alcohol is 8.40 and maximum value is 14.90 The mean and median are 10.20, 10.42. The very small difference in these values supports the normal distribution.

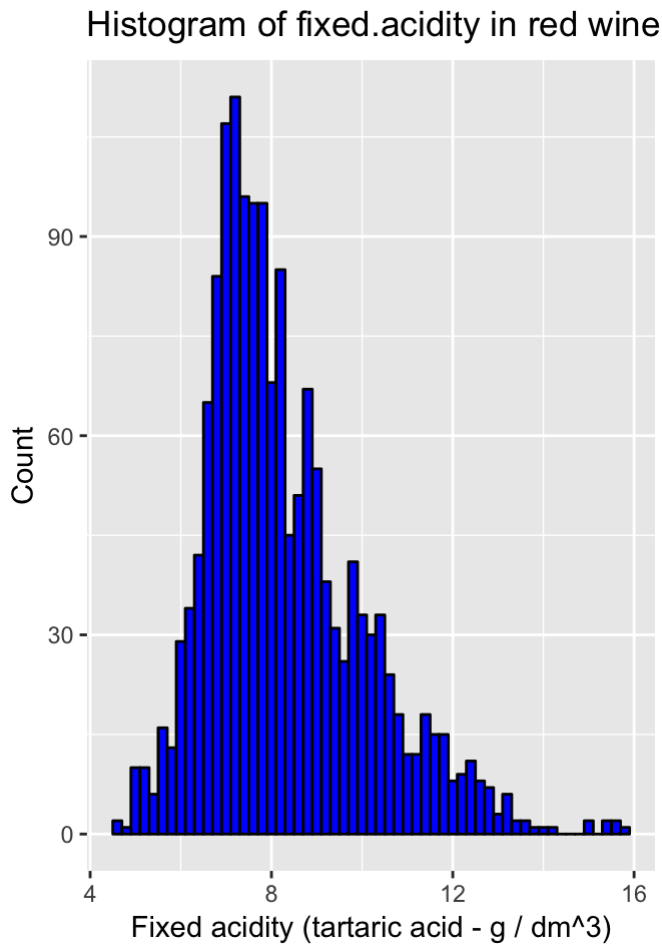

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

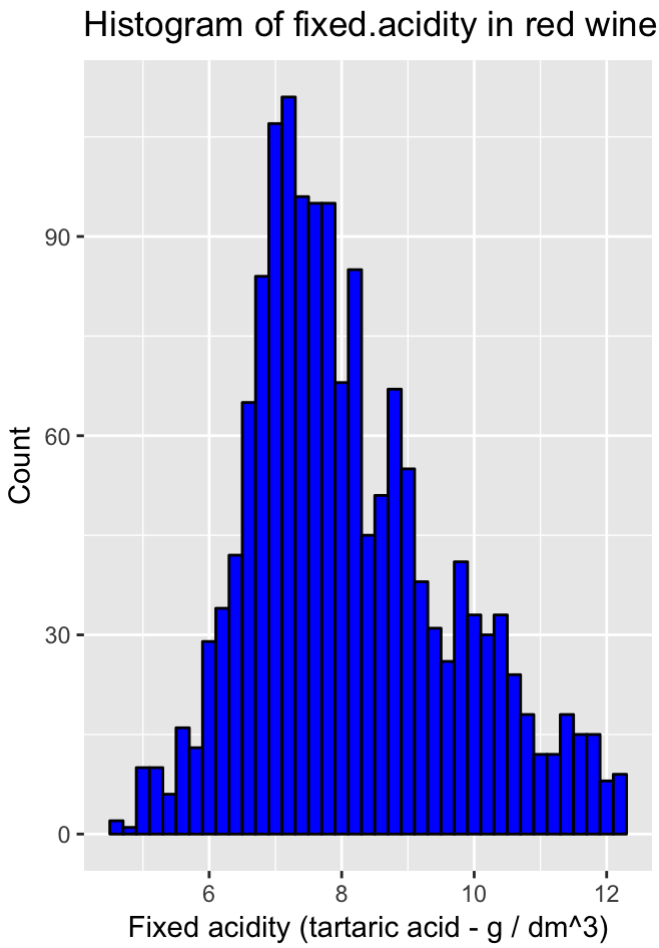
The histogram of residual sugar shows a positive skewed distribution with many far lying outliers. To detect these outliers a box plot has been shown which shows that data points above 3 g/dm^3 and below 1.6 are outliers. We see a peak around 2 and most of the data lies within the range of 1.6 and 3 g/dm^3. A red wine with sugar content of 3~4 g/dm^3 is considered medium sweet and that of above 5 g/dm^3 is sweet. Hence we can safely say that most of the red wine in this data set is below medium sweet or better named as off dry. This sugar level indicates that most of the sugar of grapes has been converted to alcohol.




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

The histogram of the density of red wine shows a perfect normal distribution with it's mean and median being approximately equal around .997 g/cm^2. In this data set the density ranges from 0.99 g/cm^2 to 1g/cm^2.

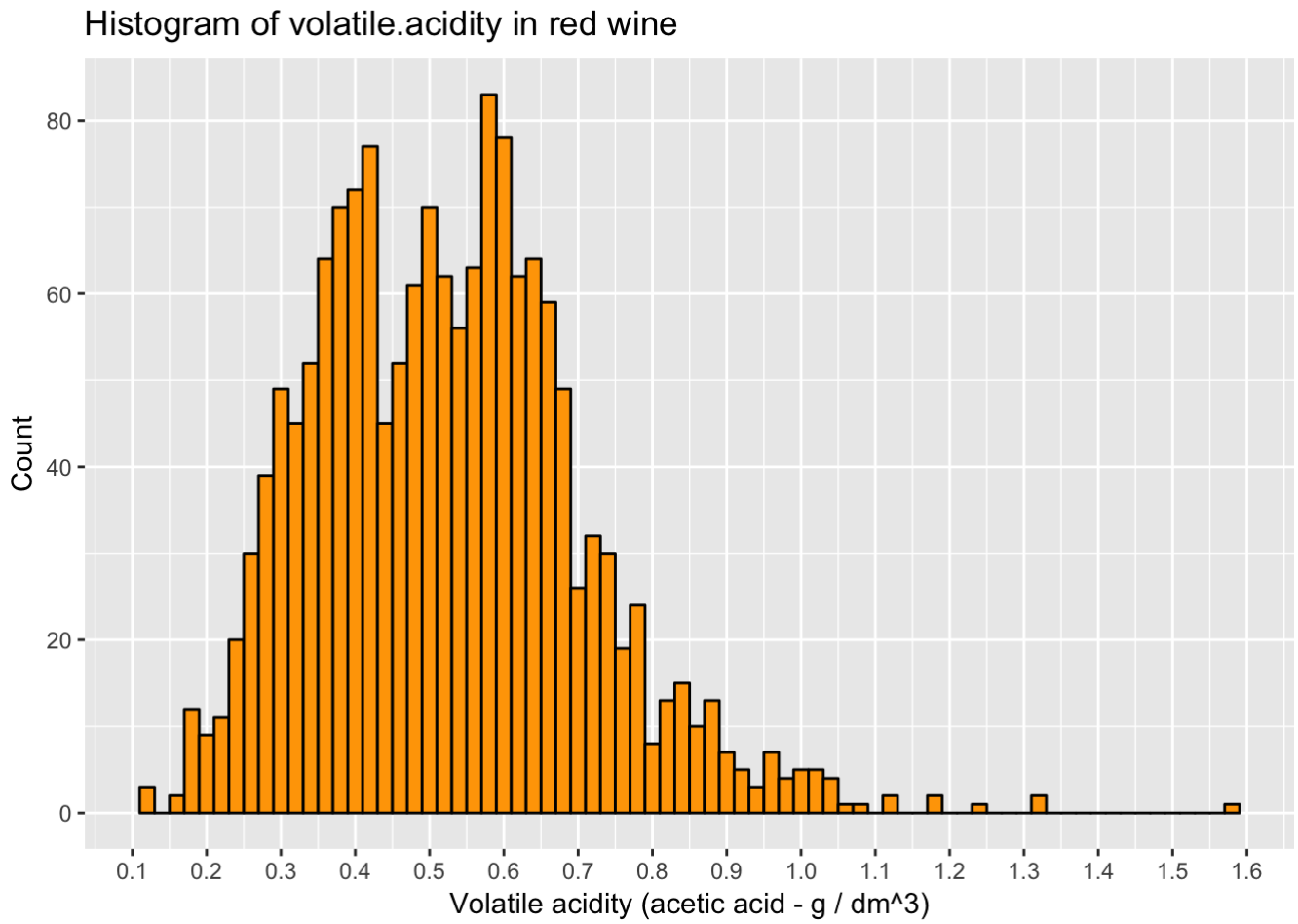





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

```
## [1] 12.4
```

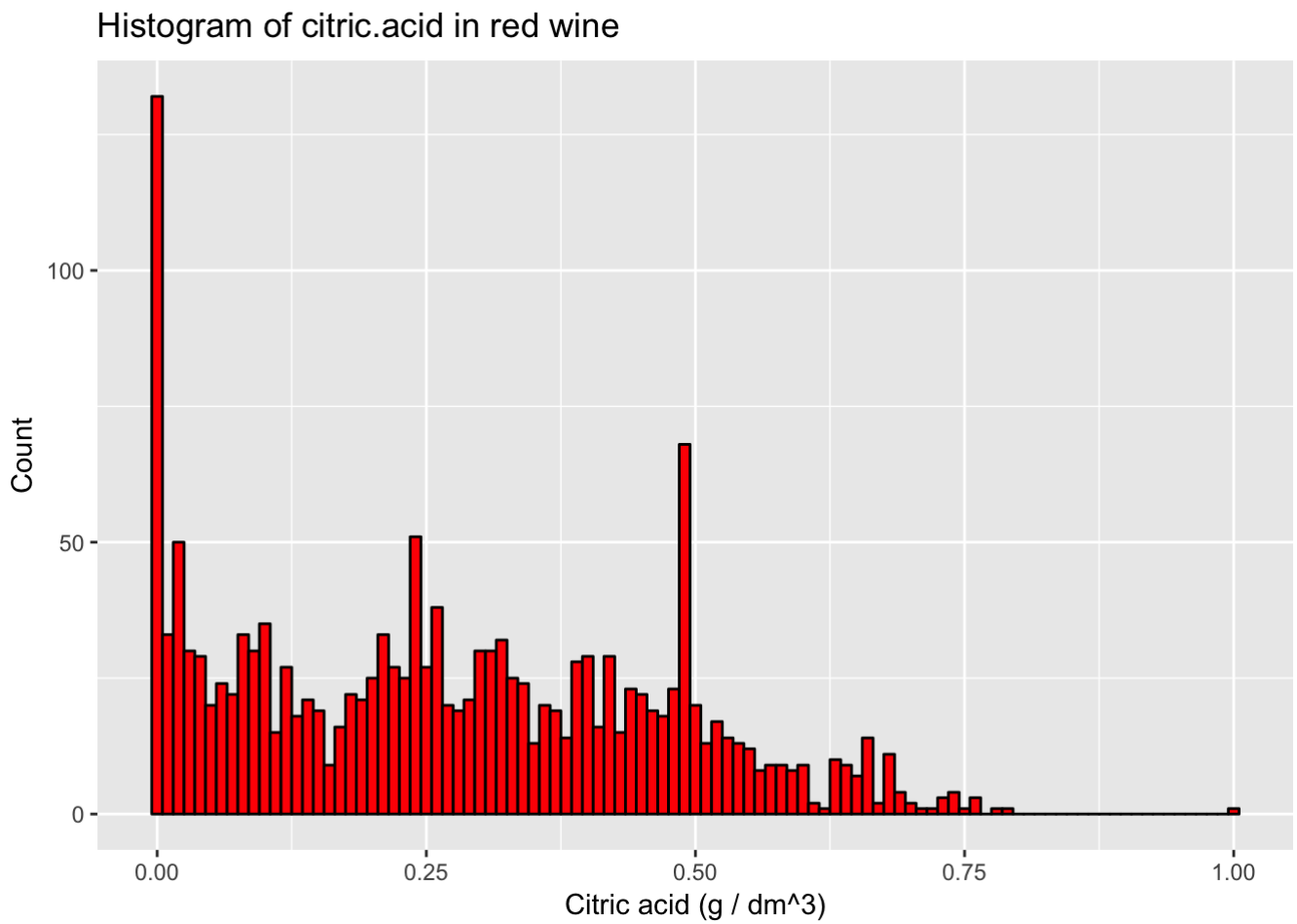
The fixed acidity histogram shows a long tailed data which looks like is coming because of outliers. On outlier detection we find that datas above 12.4 g/dm^3 are actually outliers. After removing the outliers the distribution looks more normal.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

```
## [1] 1.02
```

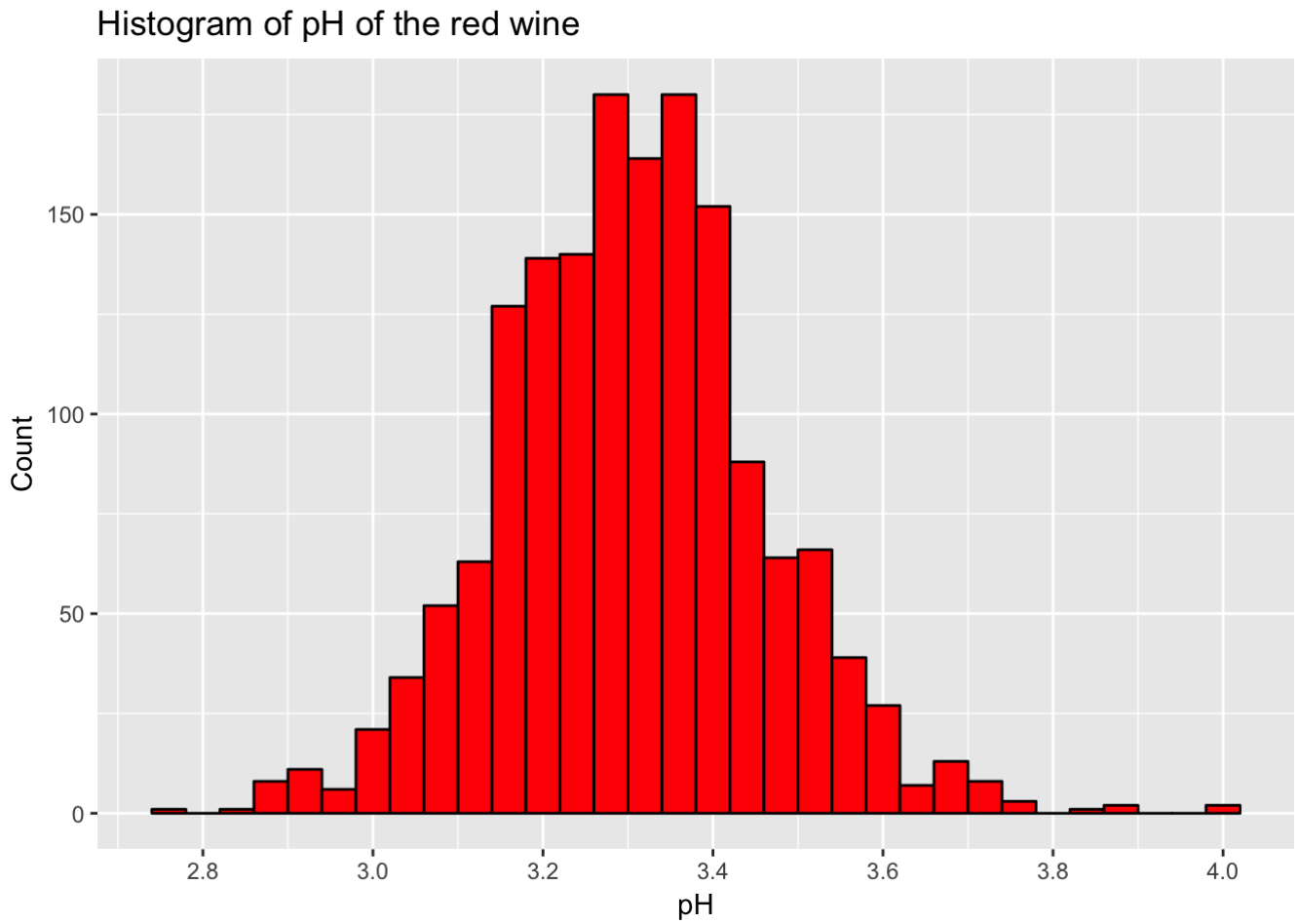
The histogram of volatile acidity shows a normal distribution with a few far lying outliers. Data points after 1.2 g/dm^3 are outliers.




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

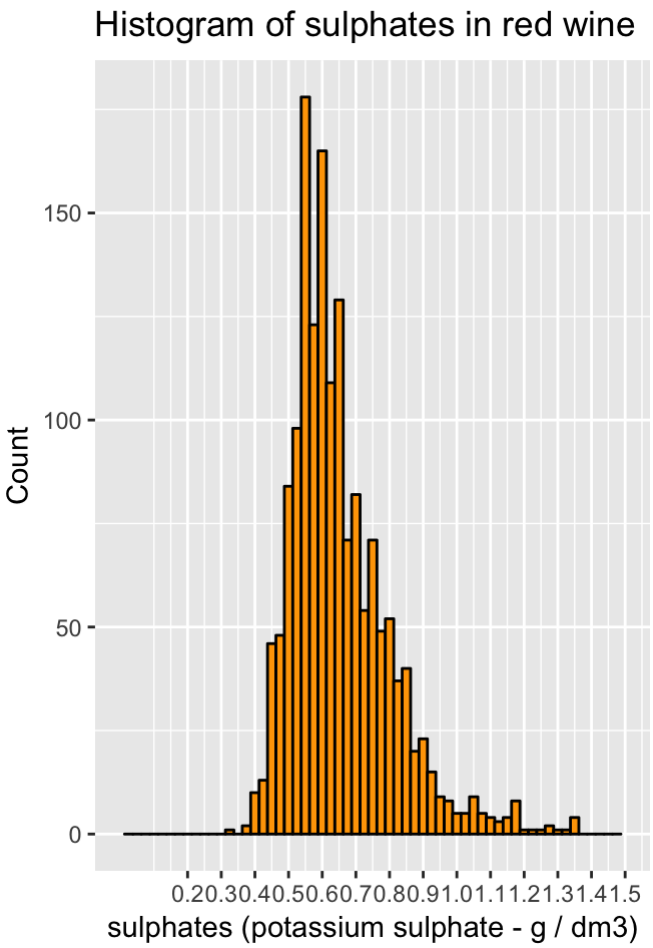
```
## [1] 1
```

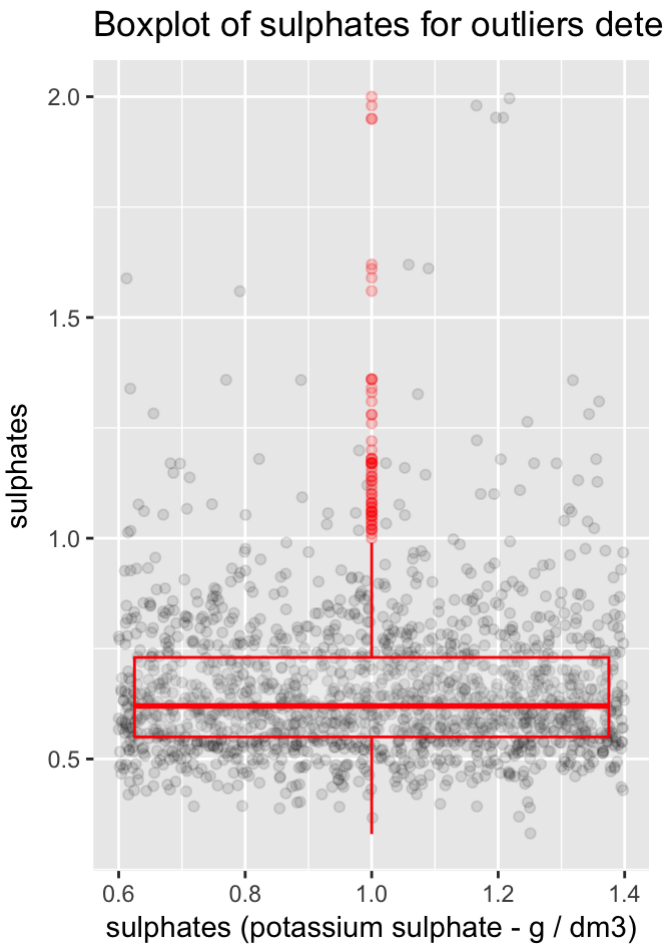
The histogram of citric acid is quite interesting which seems like almost uniform distribution except three modes at 0.0, 0.25 and 0.5 g/dm^3. It might be because of rounding off at data entry. There is just one far lying outlier at 1 g/dm^3 which also supports the rounding of the number at the time of entry. The mean is at 0.271 and median is at 0.26 and mode is at 0.




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

Acidity of a wine is one of the important feature of wine. The histogram of pH value of red wine shows good normal distribution with few far lying outliers. The mean and median is at 3.31 showing most of the wines in this data sets have average acidity.

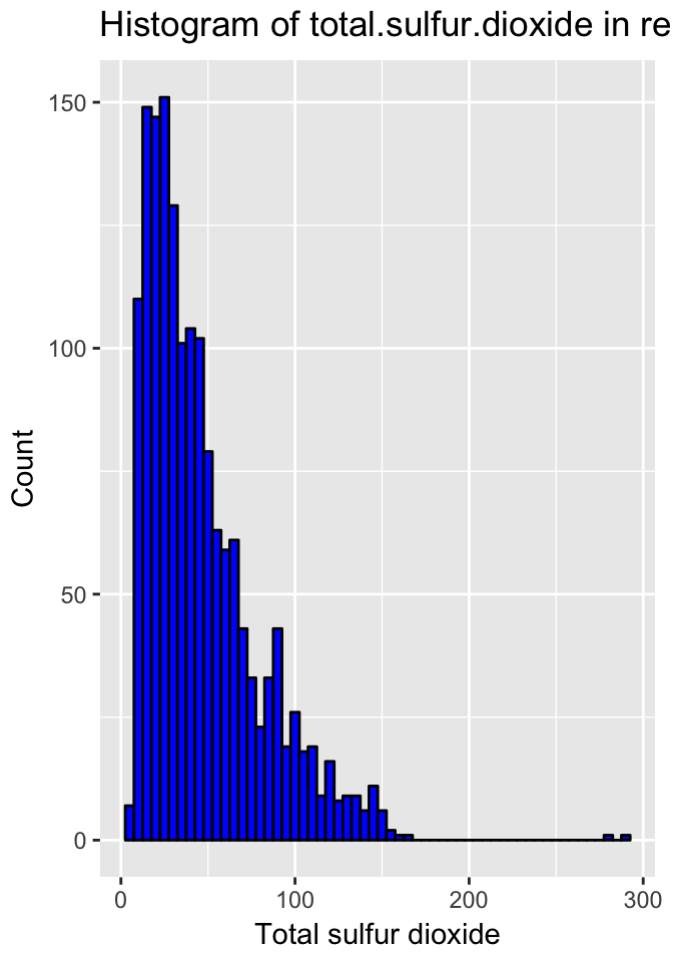


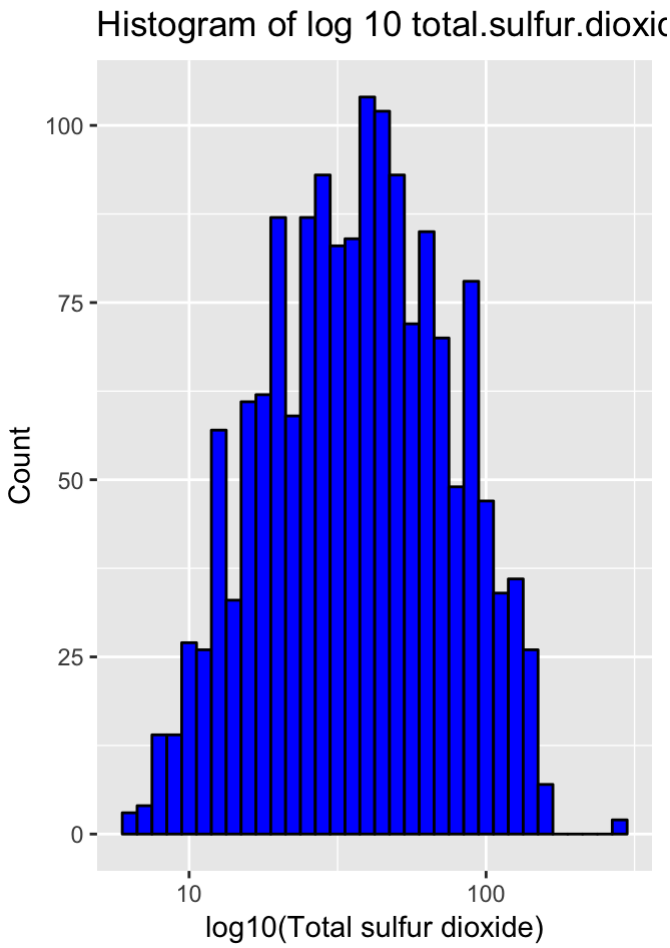



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

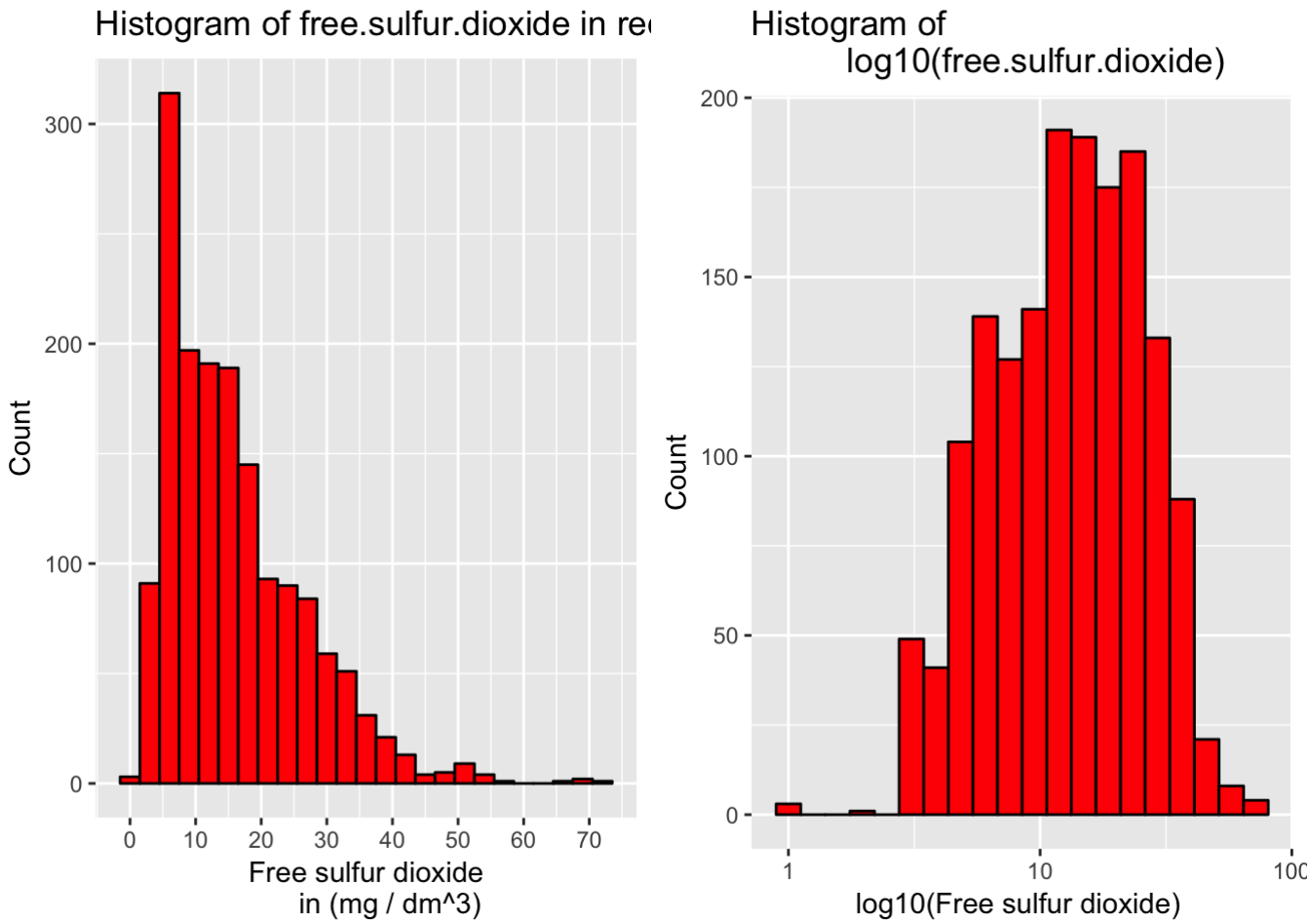
```
## [1] 1
```

The histogram of sulphates are nearly normal distribution with some outliers. Idid boxplot outliers detection and found that data points for value of sulphates below 0.51 and above 0.75 are outliers. We will use this information for outliers removal in bivariate and multivariate analysis.



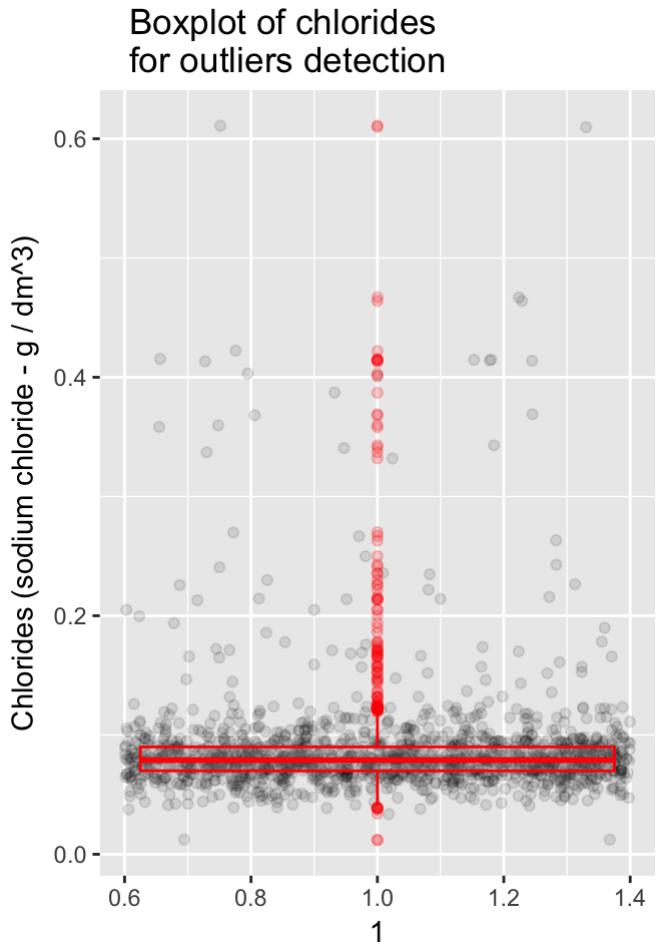


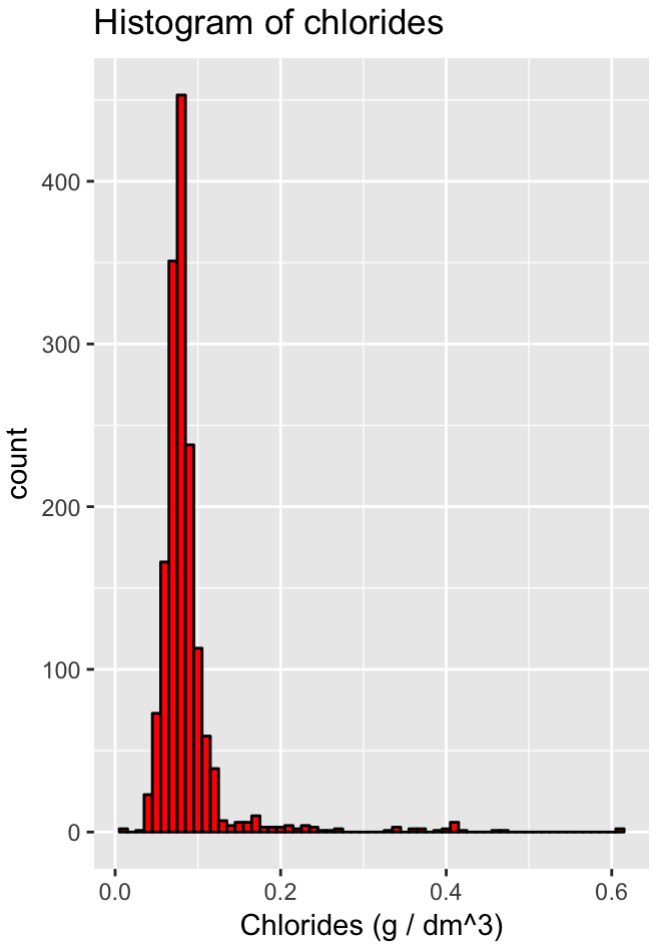
The histogram of total sulfur oxide is strongly positive skewed with a few far lying outliers. To get a more normal distribution we transform x axis at log10.




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

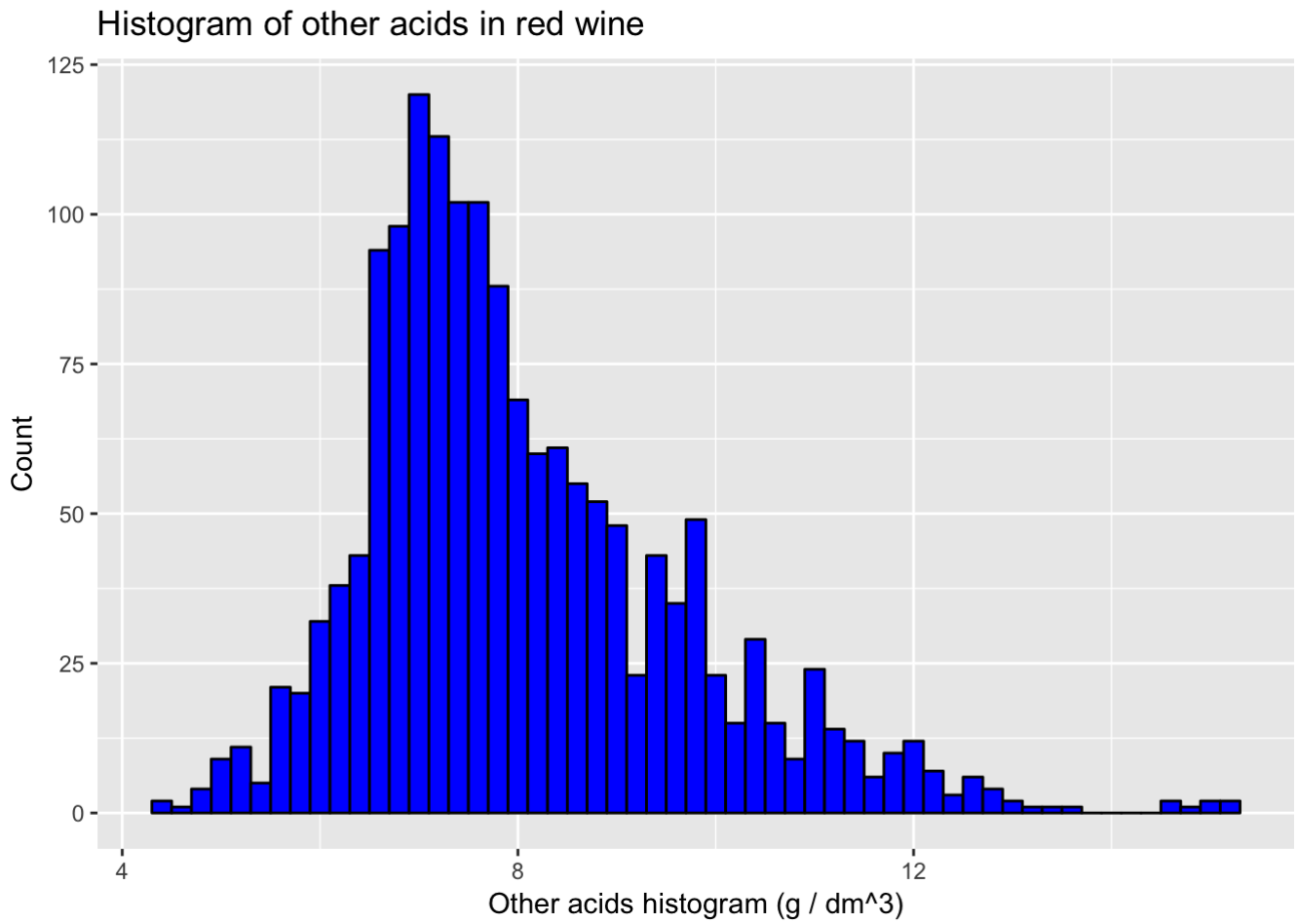
Similar to total sulfur oxide, the histogram of free sulfur oxide is also positively skewed with few far lying outliers. That is why plotting the histogram at log10 x axis gives more normal distribution.






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

If we ignore the many far lying outliers the histogram of chlorides is a peaked normal distribution. With this kind of distribution, we may not find any reliable correlation of chlorides with other features of red wine data




```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 --    1
```

In order to study the effect of tartaric acid on red wine, I made a new variable called other acid which is a difference of fixed acidity and citric acid. The histogram shows a normal distribution with few far lying outliers.

Univariate Analysis

The first feature of interest in univariate analysis is Quality of the red wine which is measured from 1 to 10. In this dataset the minimum quality is 3, maximum is at 8, median is 6 and mean is 5.64. The histogram of quality looks like normal distribution which is expected. But the data structure here seems to be monopolizedon average quality (5 and 6) red wine. 82.49% of the data lies in quality 5 and 6. This is definitely going to affect the outcomes of this analysis as there are very few data for higher quality and lower quality wine. The histogram of alcohol shows a normal distribution with a few outliers. The summary of alcohol shows that minimum value of alcohol is 8.40 and maximum value is 14.90 The mean and median are 10.20, 10.42. The very small difference in these values supports the normal distribution. density, sulphates, volatile acidity, fixed acidity and oher.acid have normal distributions. Chlorides and residiual.sugar have positive skewed nature. I suspect it is because of the presence of far lying outliers. To confirm that I plot a boxplot which shows that in case of sugar the data points above 4 g / dm^3 are outliers if we remove them we find a normal distribution of residual sugar. Similarly, the boxplot of chlorides show data points after 0.1 g/dm^3 are outliers. When we remove these outliers we find a normal distribution of chlorides. The histogram of total.sulfur.oxide and free.sulfur.oxide on linear scale are positively skewed structure so I plot them on log10 axis to get normalize distribution.

What is the structure of your dataset?

There are 13 variables and 1599 observations.

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

Quality as an output feature is the most important feature of this dataset.

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

Alcohol, sulphates, citric.acid and volatile acidity are the main features to understand the catagorization of the red wine.

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

I created other.acid feature by subtracting citric.acid from fixed.acidity.

Of the features you investigated, were there any unusual distributions?

The histogram of free.sulfur.oxide and total.sulfur.oxide have positive skewed distributions. I transformed the x axis to log10 to achieve a more normalized distribution.

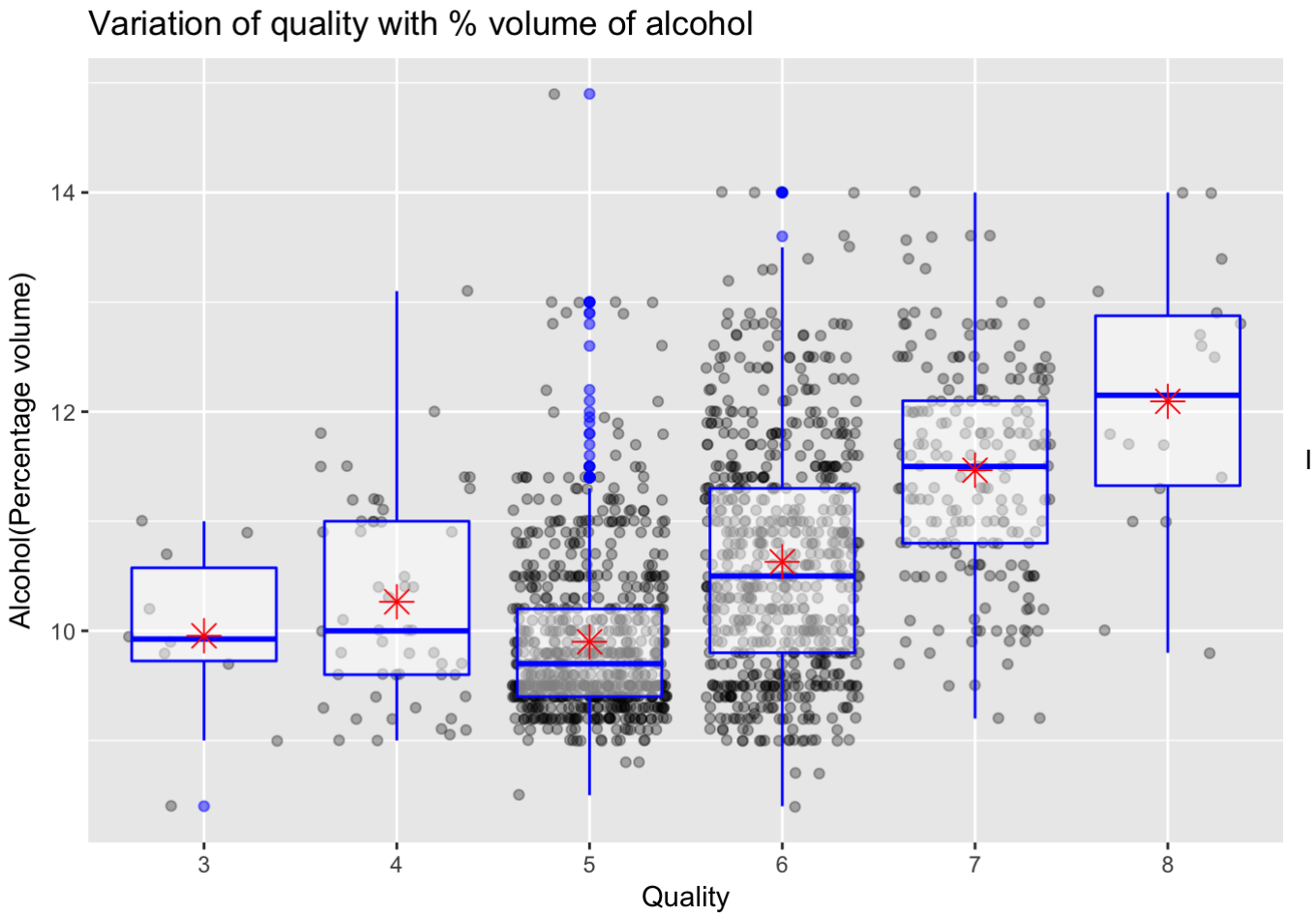
Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I did not have to change any data.

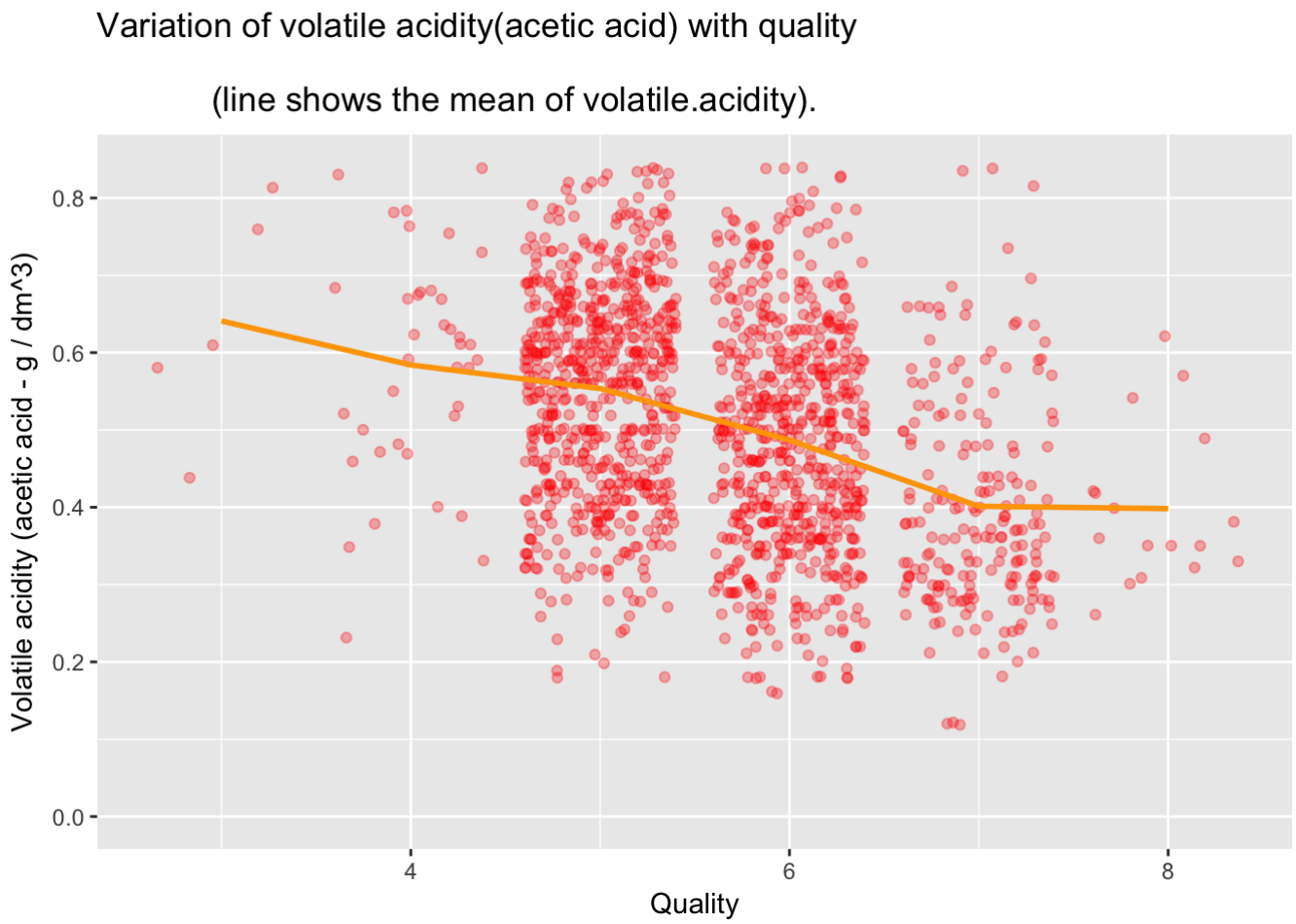
Bivariate Plots Section

Before starting the bivariate analysis, it would be helpful to get the idea of correlations between the variables. I plot scatterplot matrix and list the variables with significant correlations (cor. coeff. >0.2) fixed.acidity is highly correlated with density, citric acid and pH.

```
## [1] 0.4761663
```

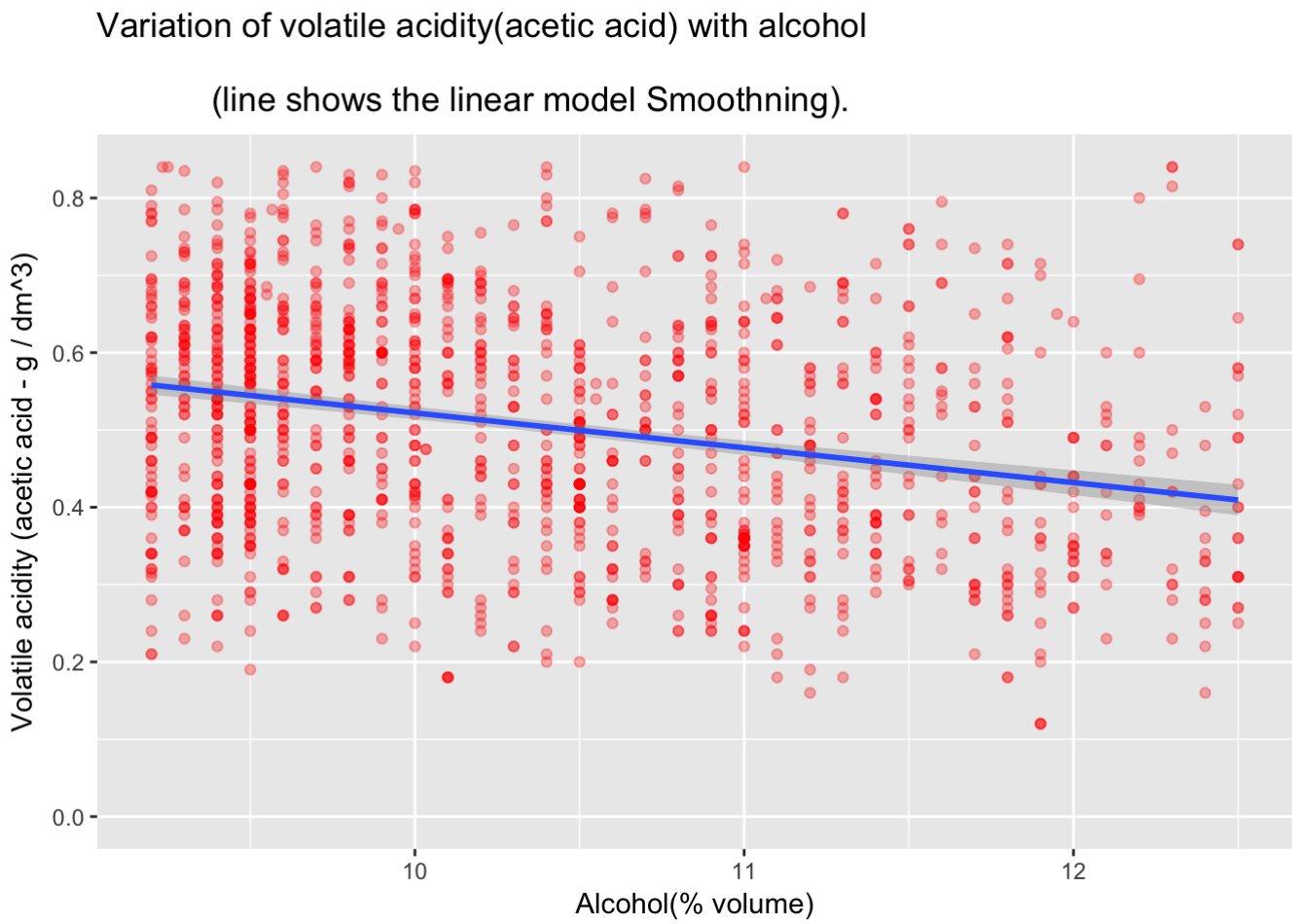


observed some outliers in the alcohol histogram which I removed for bivariate analysis. Looking at the mean of alcohol curve we observe the quality is increasing linearly with alcohol for the range between 5 to 7 while for the range 3-4 and 7-8 the mean of alcohol is constant and for the range 4-5 there is actually a decrease in the mean of alcohol.



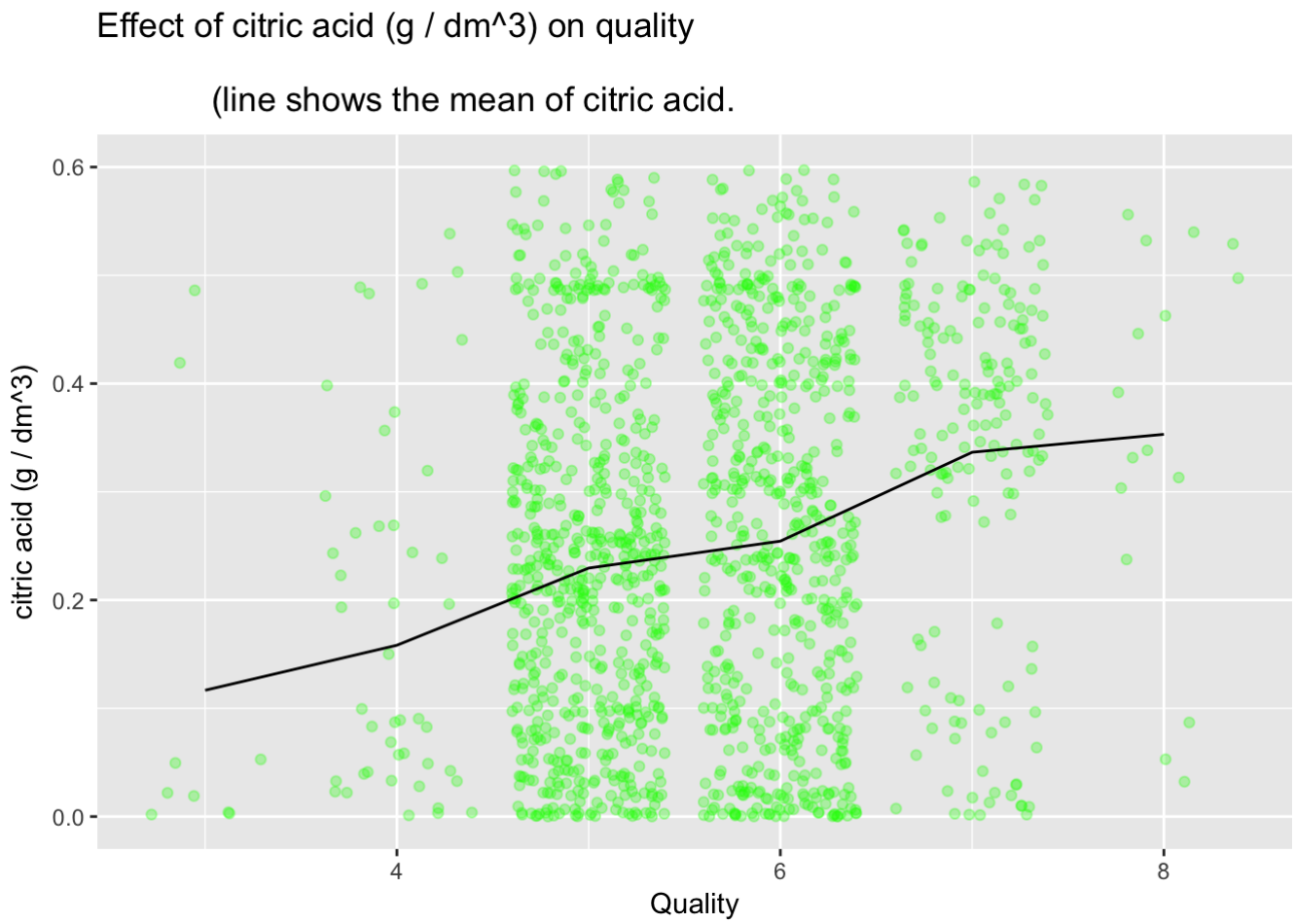

```
## [1] -0.3905578
```

The plot shows that above a threshold value higher volatile acidity decreases the quality of the wine. This relationship is linear in nature.



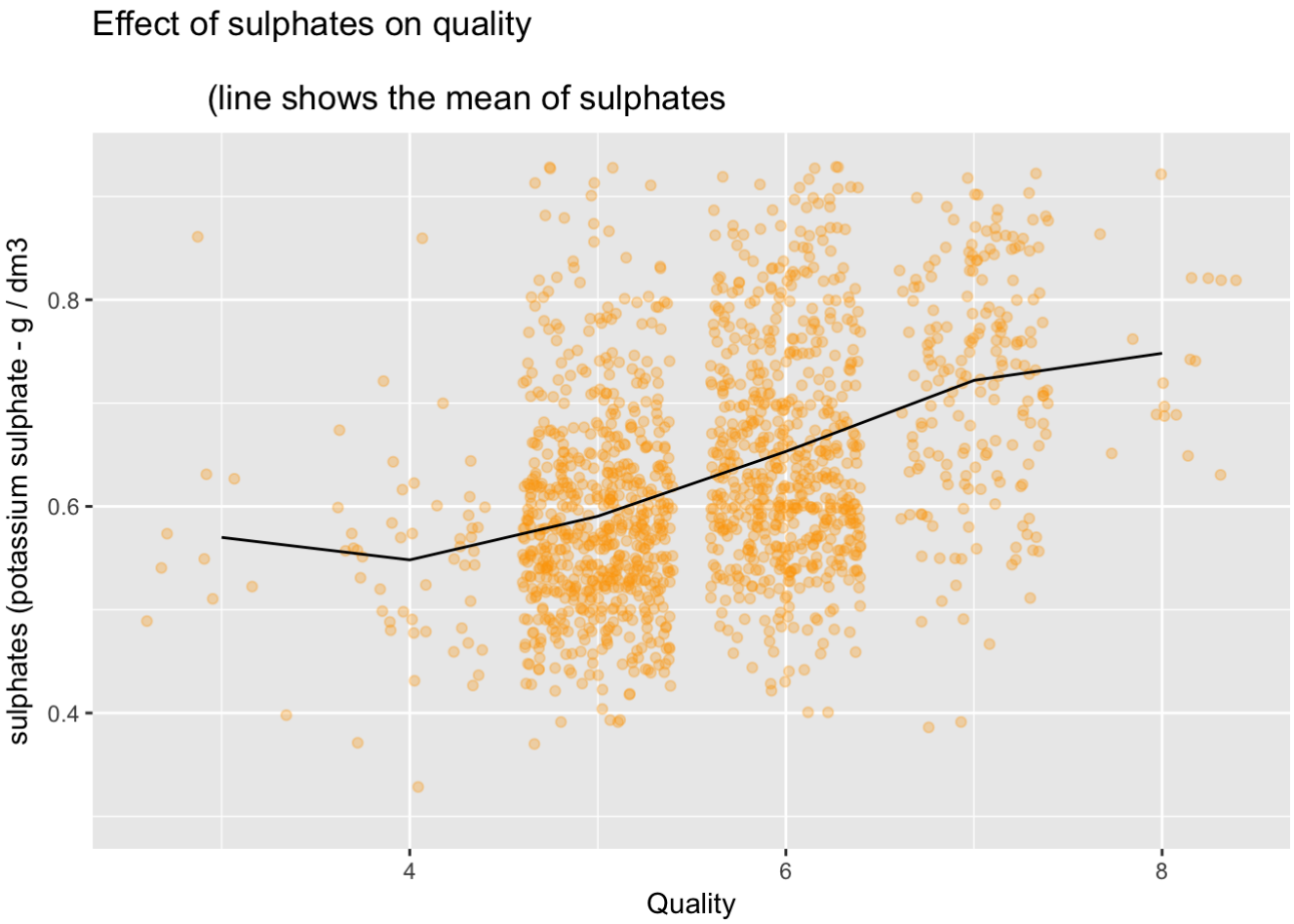

```
## [1] -0.202288
```

The smoothning method has been set at linear model('lm'). The plot shown negative linear correlation above a threshold acetic acid value of ~ o.4 g.dm^3.

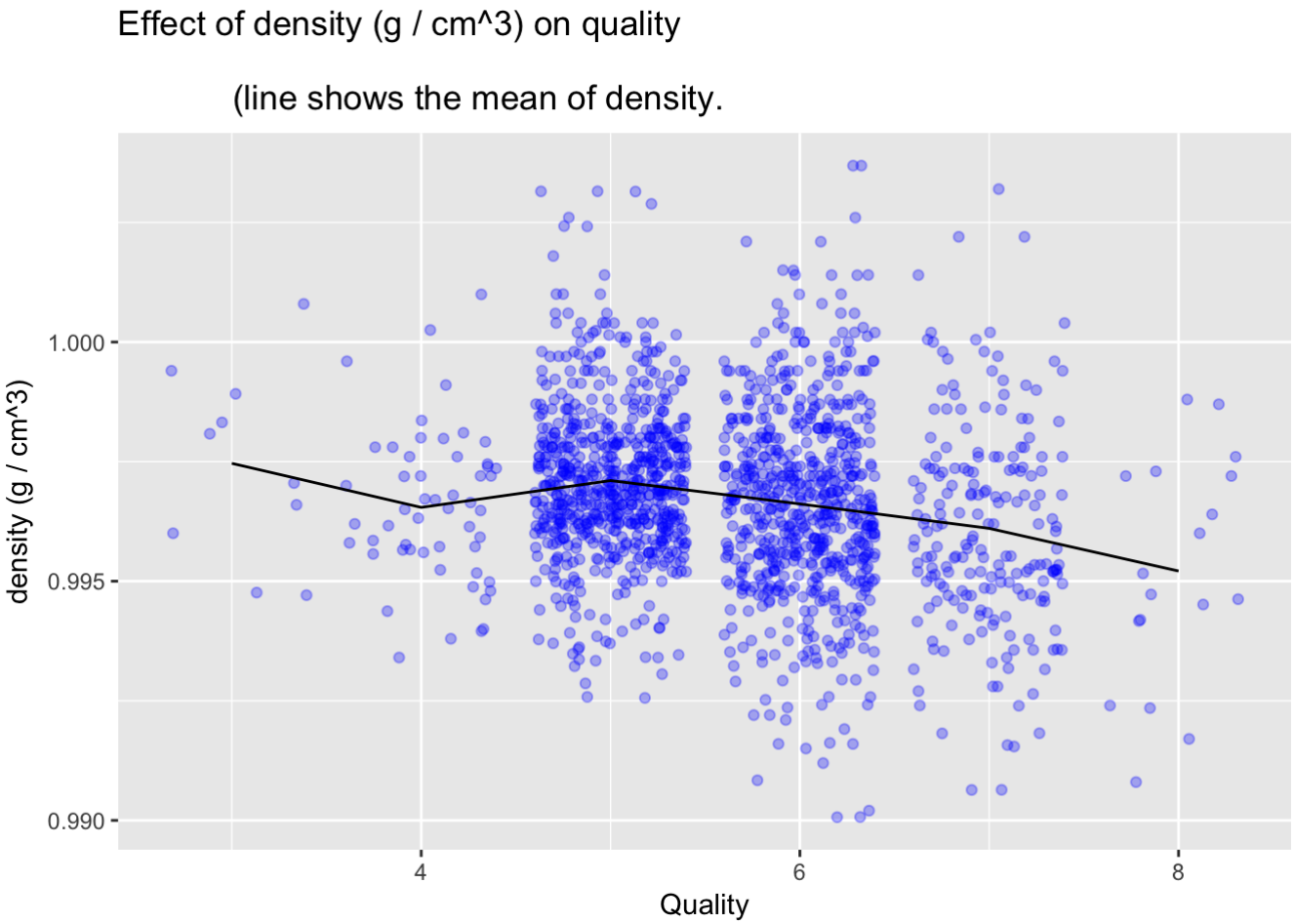



```
## [1] 0.2263725
```

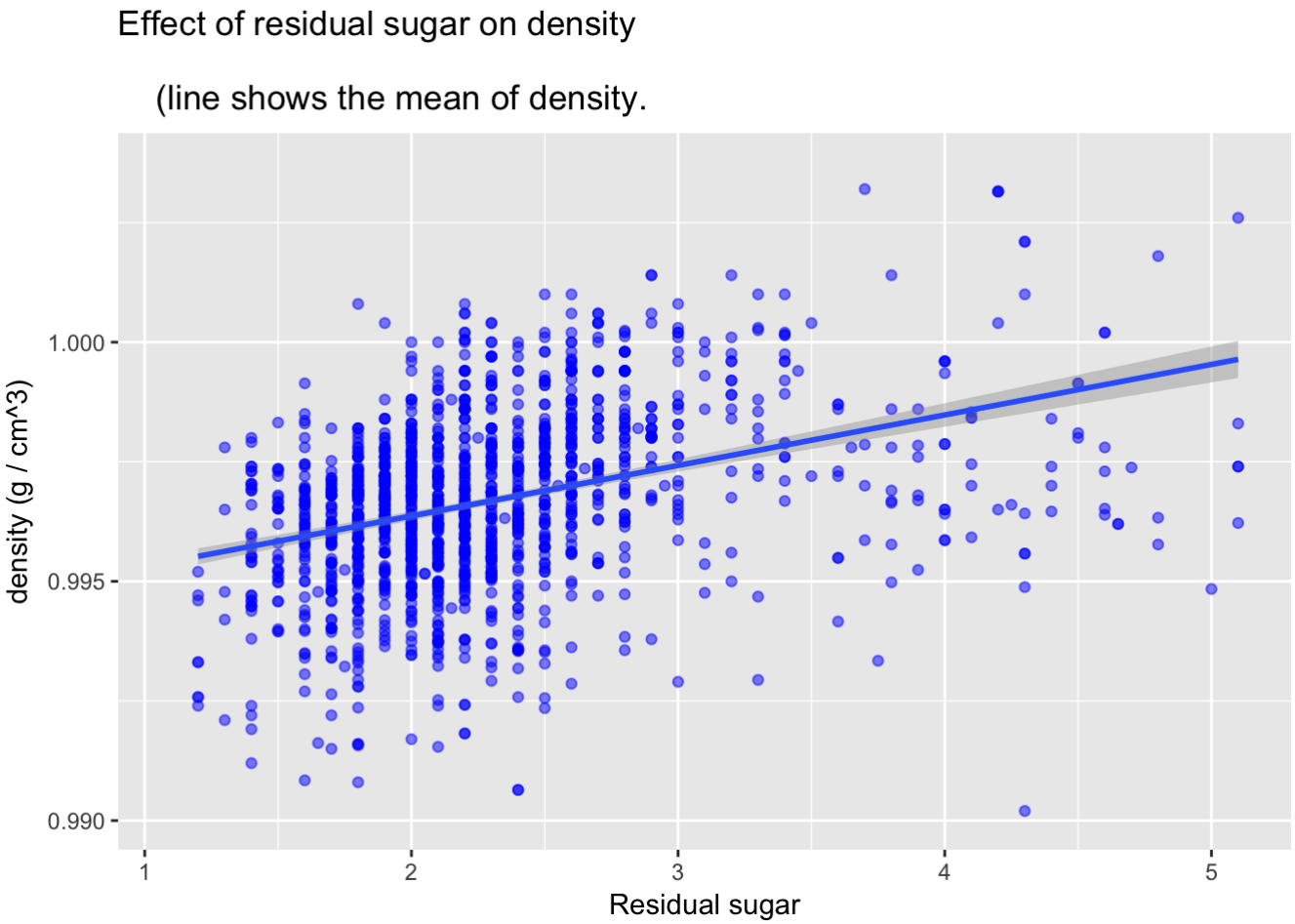
The citric acid vs quality indicates that the quality of the red wine increases with increase in citric acid.



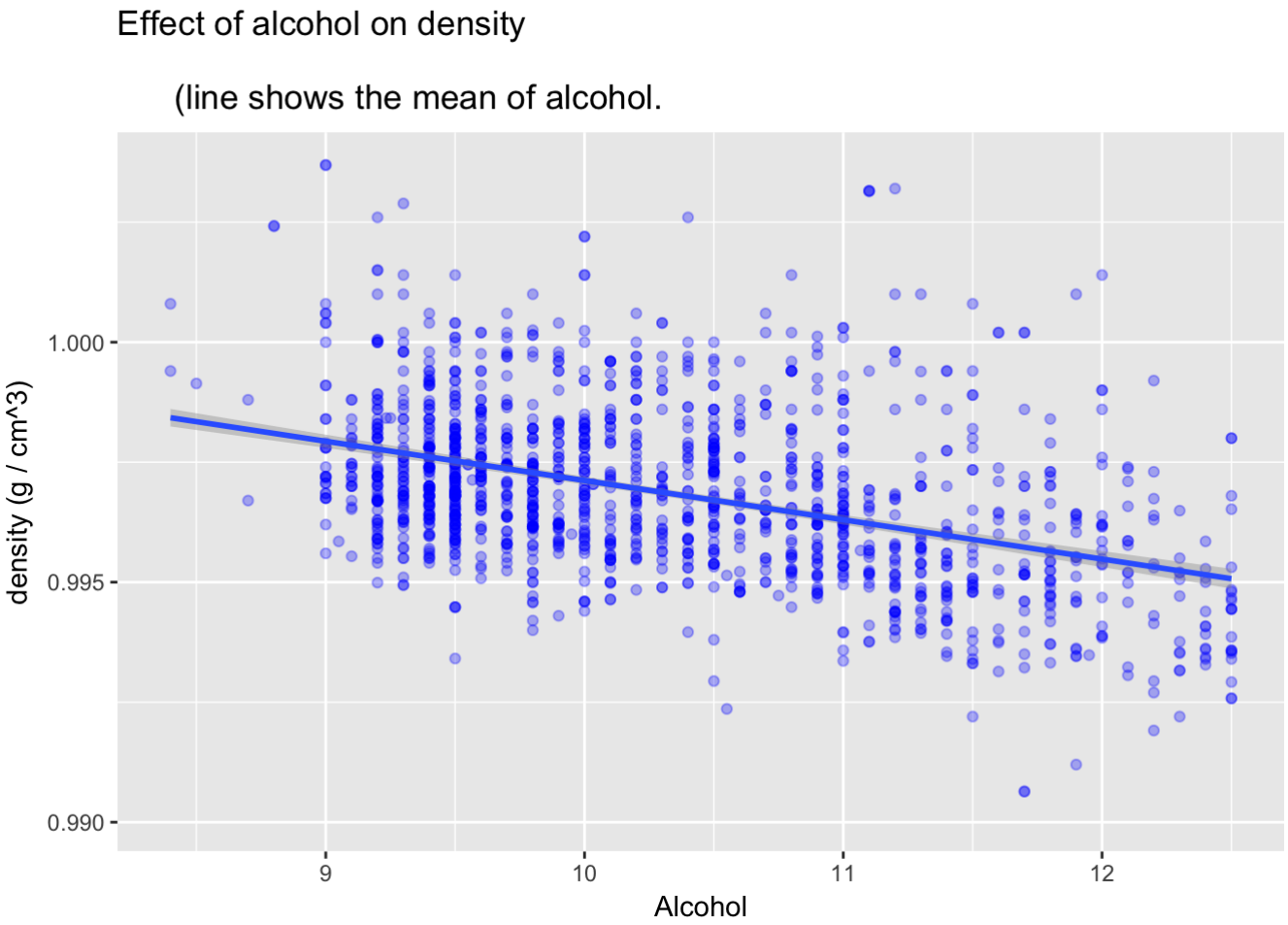
The plot of sulphates vs quality suggests that higher content of sulphates is favourable for quality of the wine though in low quality wine the change in sulphates amount is not that evident.



As the density decreases the quality gets better it might be because of increase in alcohol will reduce the density of wine.

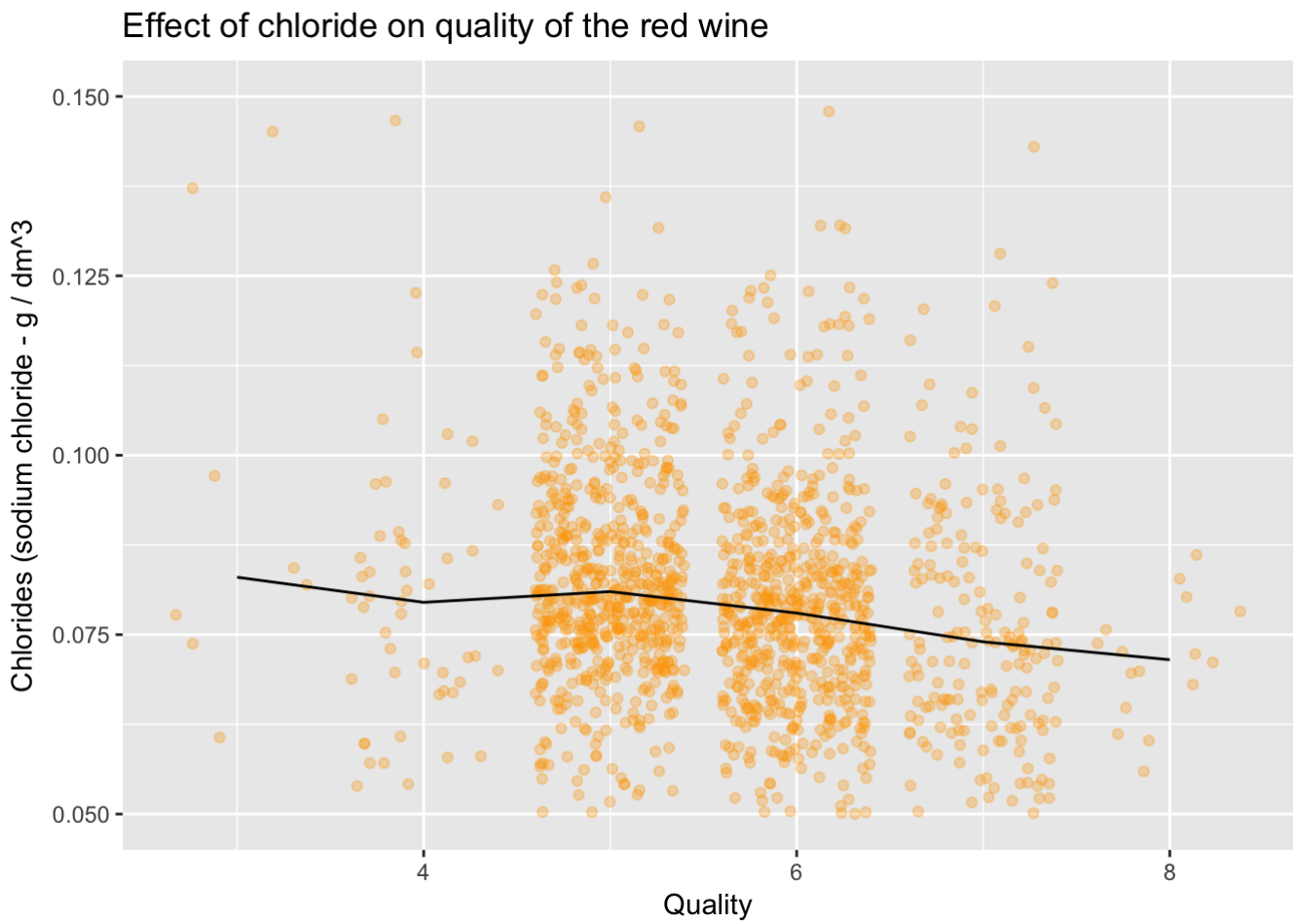


This bivariate plot shows that density of wine is linearly proportional to content of residual sugar in red wine.

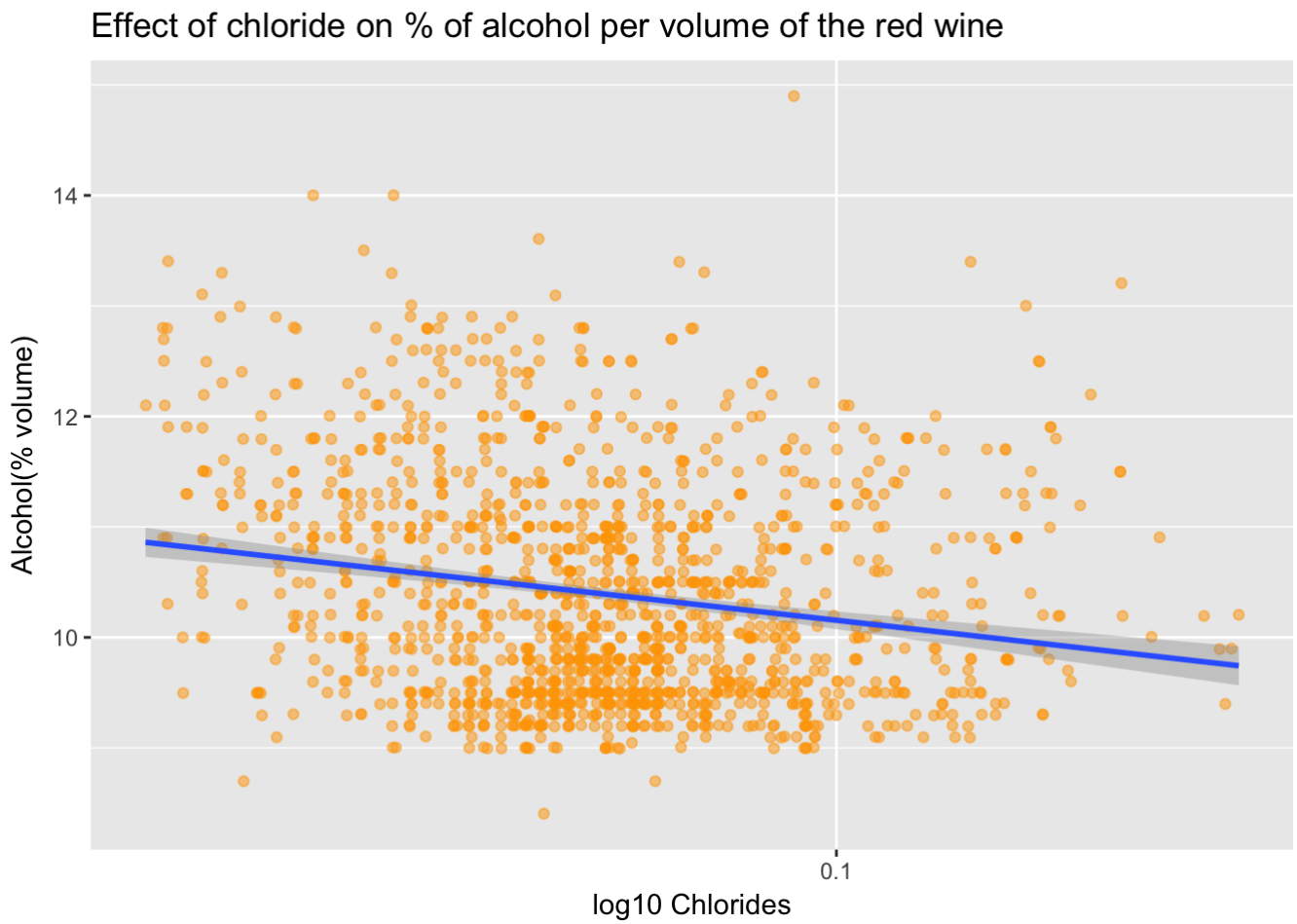


The density of red wine decrease with increase in alcohol content. This is expected as we know the density of alcohol is less than water.

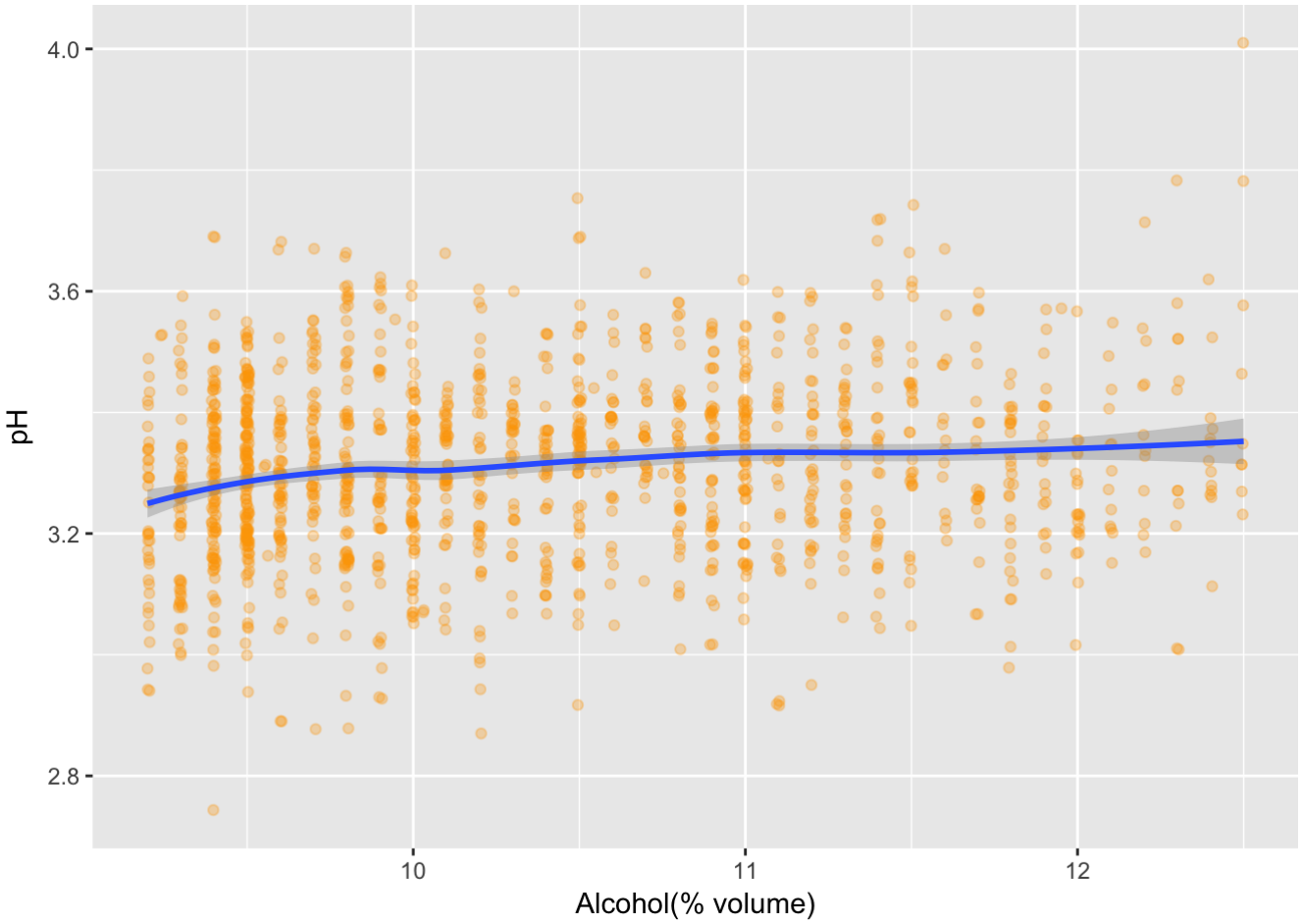
The quality of wine is slighly decreasing with pH. THis can be better understood with the bivariate analysis of acidities which contribute to pH of red wine.



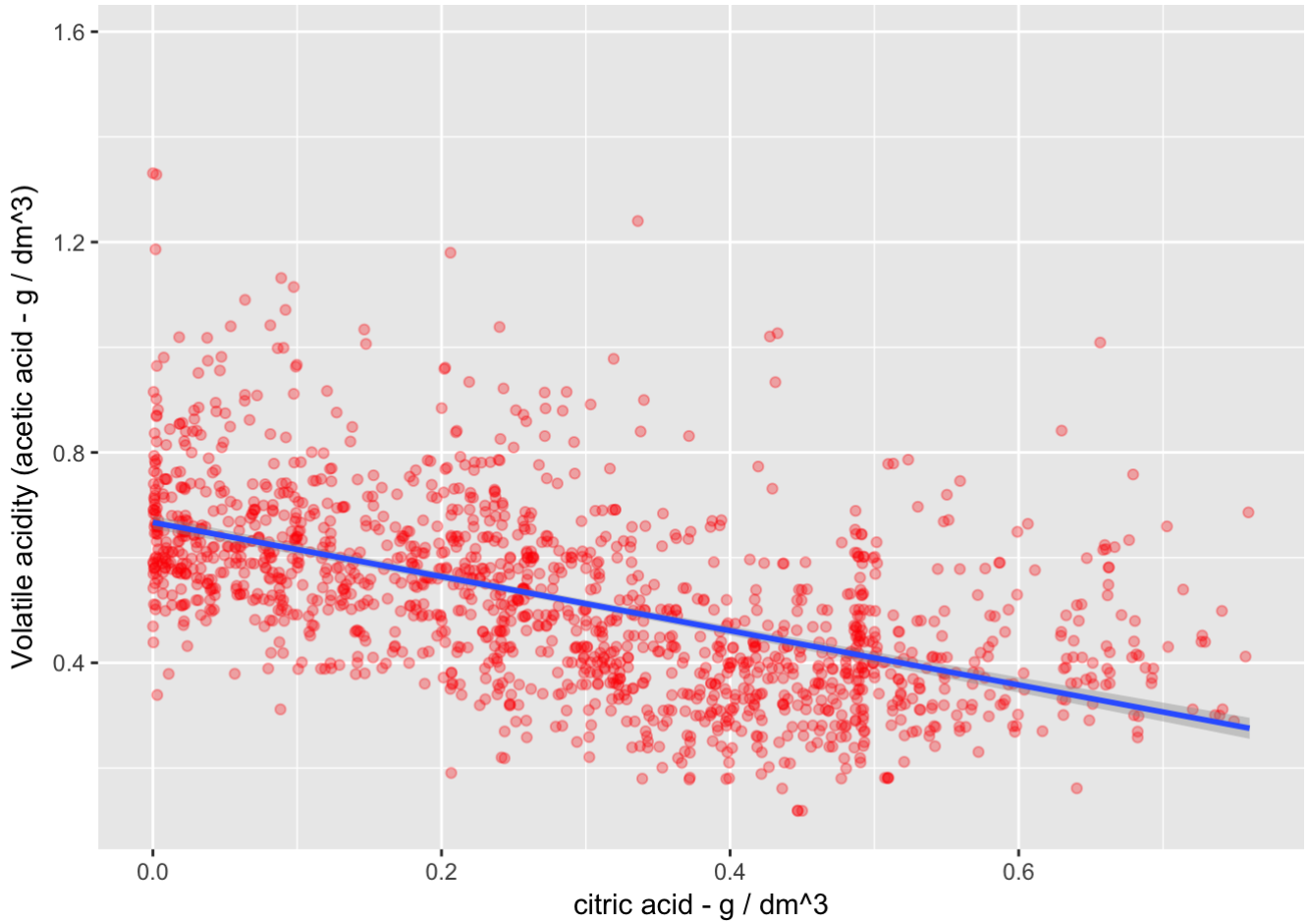
The quality of wine is slightly decreasing by chlorides in red wine. We have removed the outliers in this analysis.



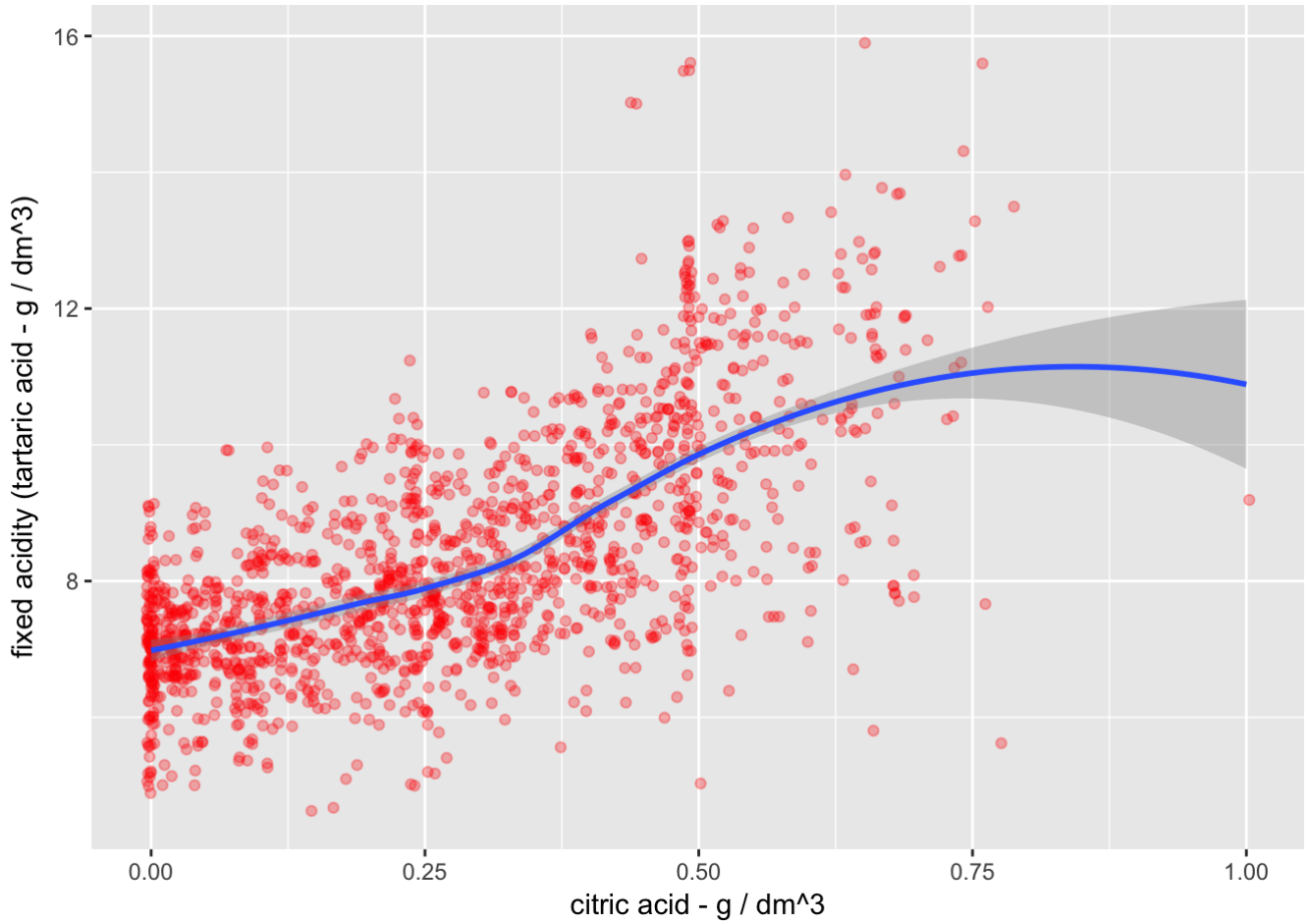
In this bivariate plot we see a linear decrease in alcohol with log10 of chlorides which means exponential decrease of alcohol with chlorides.



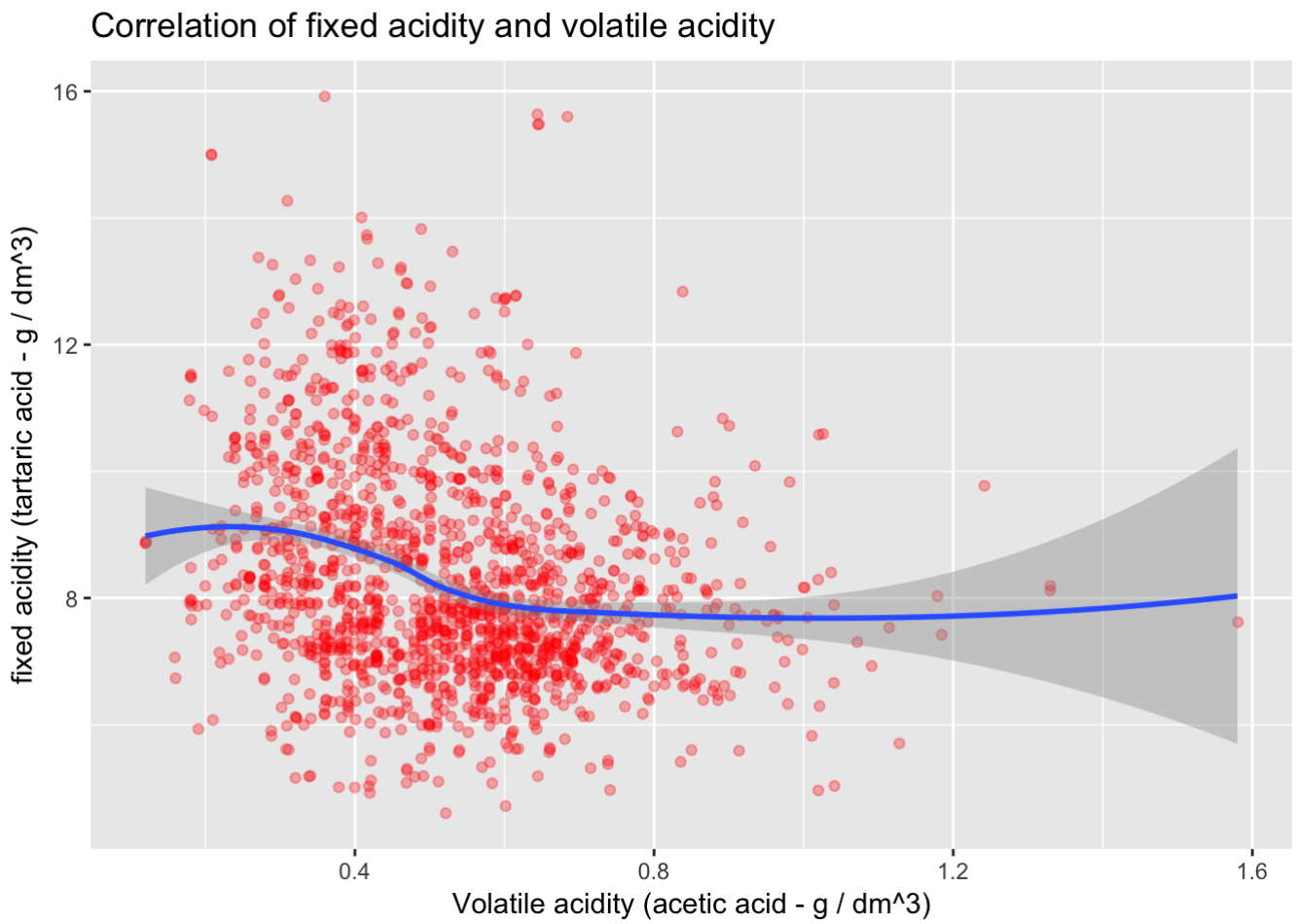
This plot shows no significant change in the pH value of wine with increase in alcohol percentage.



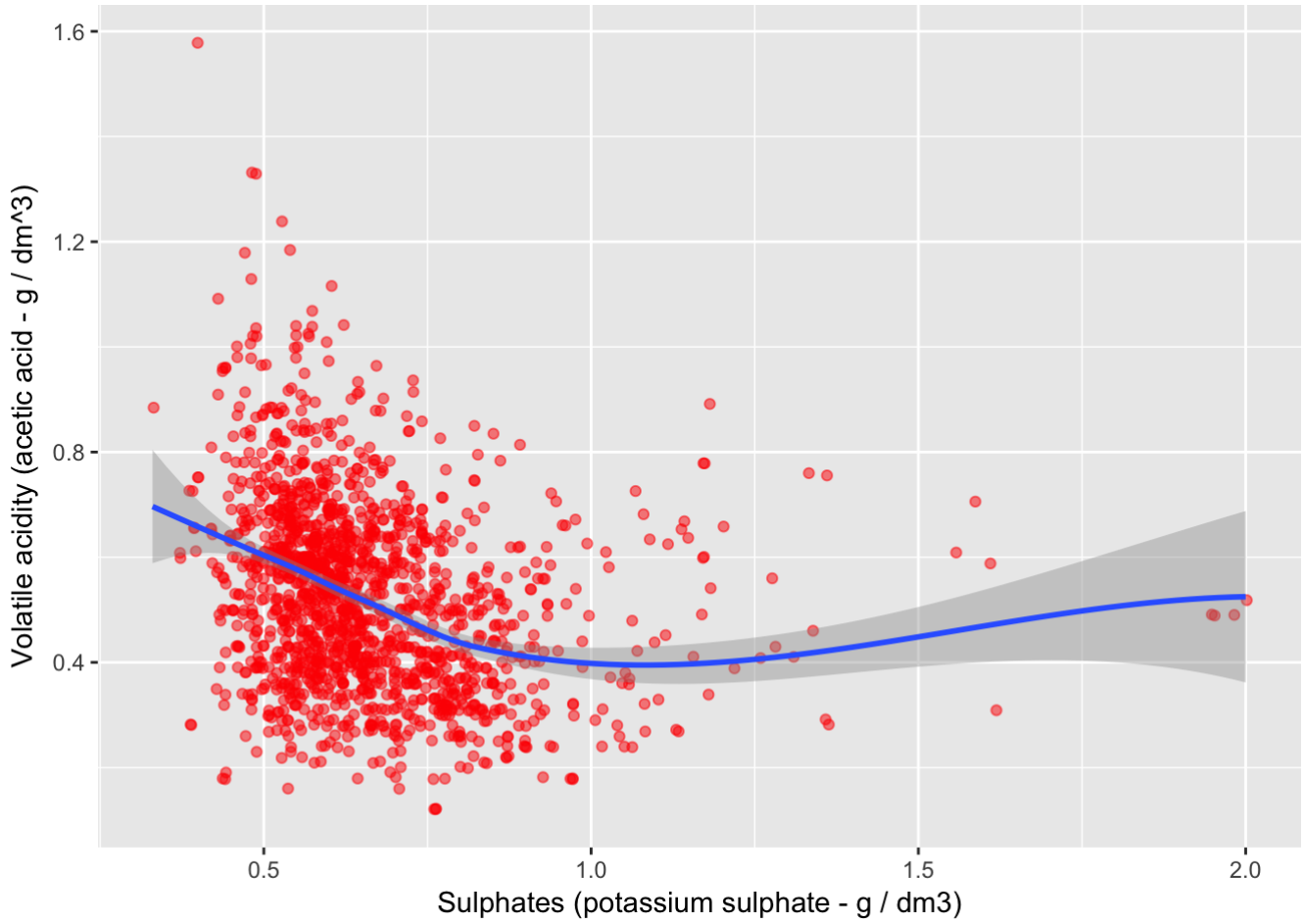
The volatile acidity decreases linearly with citric acid. We will analyze this relationship in bivariate analysis section



The fixed acidity increases linearly with citric acid.

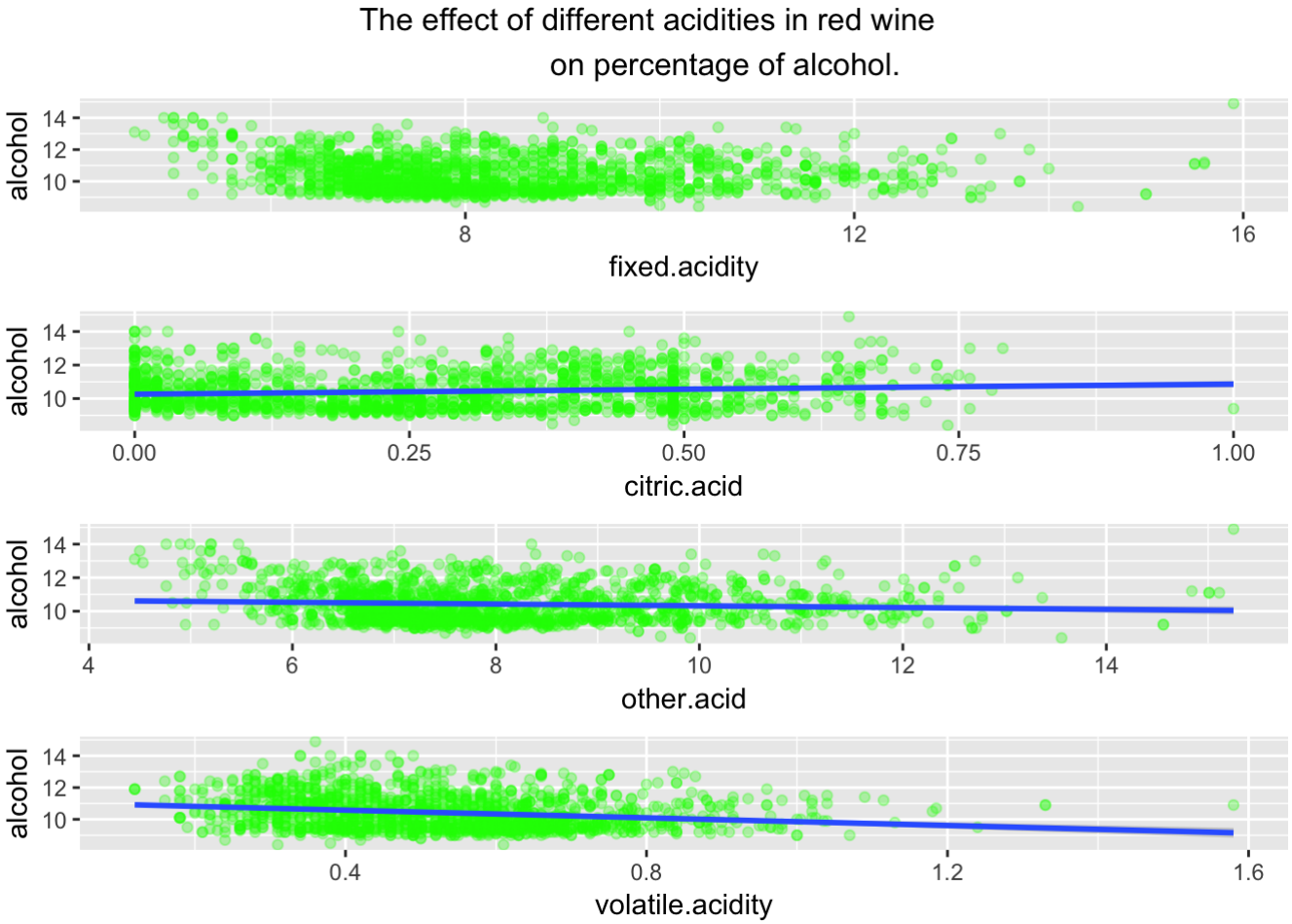

Correlation of fixed acidity and volatile acidity
fixed acidity (tartaric acid - g / dm^3)
16
12
8
0.4
0.8
1.2
1.6
Volatile acidity (acetic acid - g / dm^3)

Simialr to volatile acidity vs citric acid plot fixed acidity also show slight decrease in volatile acidity.



This bivariate analysis of volatile acidity to sulphates shows negative correlation. This observation meets with our expectation as the Sulphates are added in red wine to contain the production of volatile acidity.

```
## geom_smooth: na.rm = FALSE
## stat_smooth: na.rm = FALSE, method = lm, formula = y ~ x, se = TRUE
## position_identity
```



Here we plot effect of various acidity on alcohol percentage. Citric acid does not have any effect on alcohol while increase in tartaric acid in wine corresponds with very little decrease in the alcohol percentage. The volatile acidity has a significantly negative impact on the production of alcohol in red wine.

Bivariate Analysis

The first and foremost plot of interest is quality vs alcohol, as the correlation between these two features is highest. This is an interesting graph as unlike the expectations the alcohol does not have constant increase over the quality of the red wine. The outcome of the analysis of this plot is that there are other features playing significant role in the quality of the red wine. For the next plot, we study next significantly correlated relationship i.e. quality vs volatile acidity which is the amount of acetic acid in wine. The plot shows that above a threshold value higher volatile acidity decreases the quality of the wine. This relationship is linear in nature. Interestingly alcohol is also negatively correlated with volatile acidity. Now the question one may ask is how much correlation of alcohol with acetic acid constitutes the correlation of quality vs acetic acid. we will address this question in multivariate analysis section. For now our next object of interest is alcohol vs volatile.acidity plot. The smoothning method has been set at linear model('lm'). The plot shown negative linear correlation above a threshold acetic acid value of ~ o.4 g.dm^3. The next features correlated with quality are sulphates and citric acid.

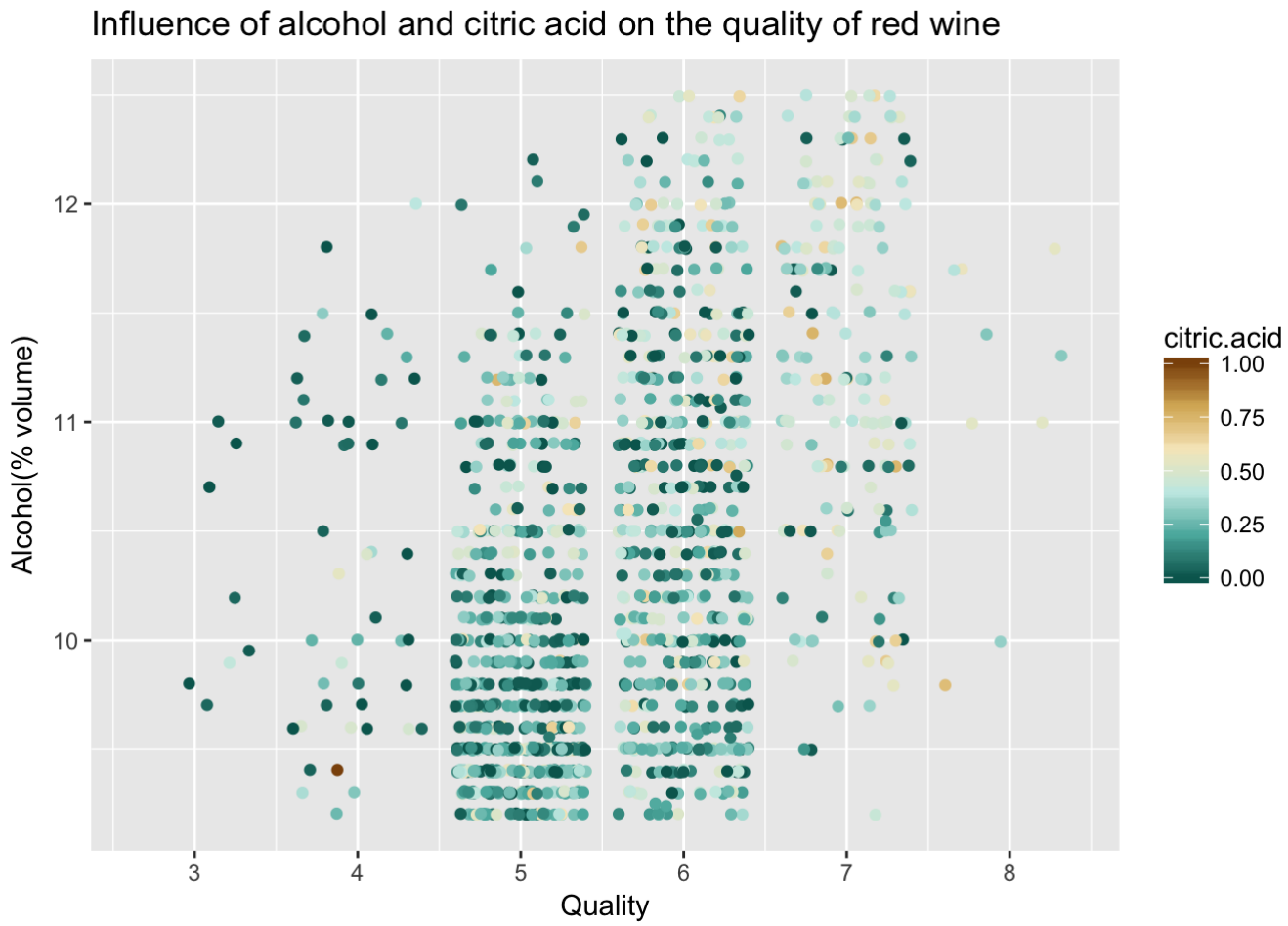
The plot of sulphates vs quality suggests that higher content of sulphates is favourable for quality of the wine though in low quality wine the change in sulphates amount is not that evident. Similarly, the citric acid vs quality indicates that the quality of the red wine increases with increase in citric acid. This observation is expected as citric acid gives the freshness to red wine. The density of the wine is strongly correlated with alcohol and residual sugar. So, next we analyze alcohol vs density, residual sugar vs density and quality vs density. I observed an exponential decrease in alcohol with chloride so I transformed the x axis on log10 and that gives a linear decreasing relationship.

Looking at the correlation coefficient we can deduce that the pH of the red wine does not affect the quality of the red wine. The positive correlation of pH with alcohol is studied in next plots along with variation of pH with different acidity in wine. There is a slight linear increase in pH with alcohol. As alcohol is neutral solution this observation only means that wine with higher alcohol have a little less acidity than the wine with lower percentage of alcohol.

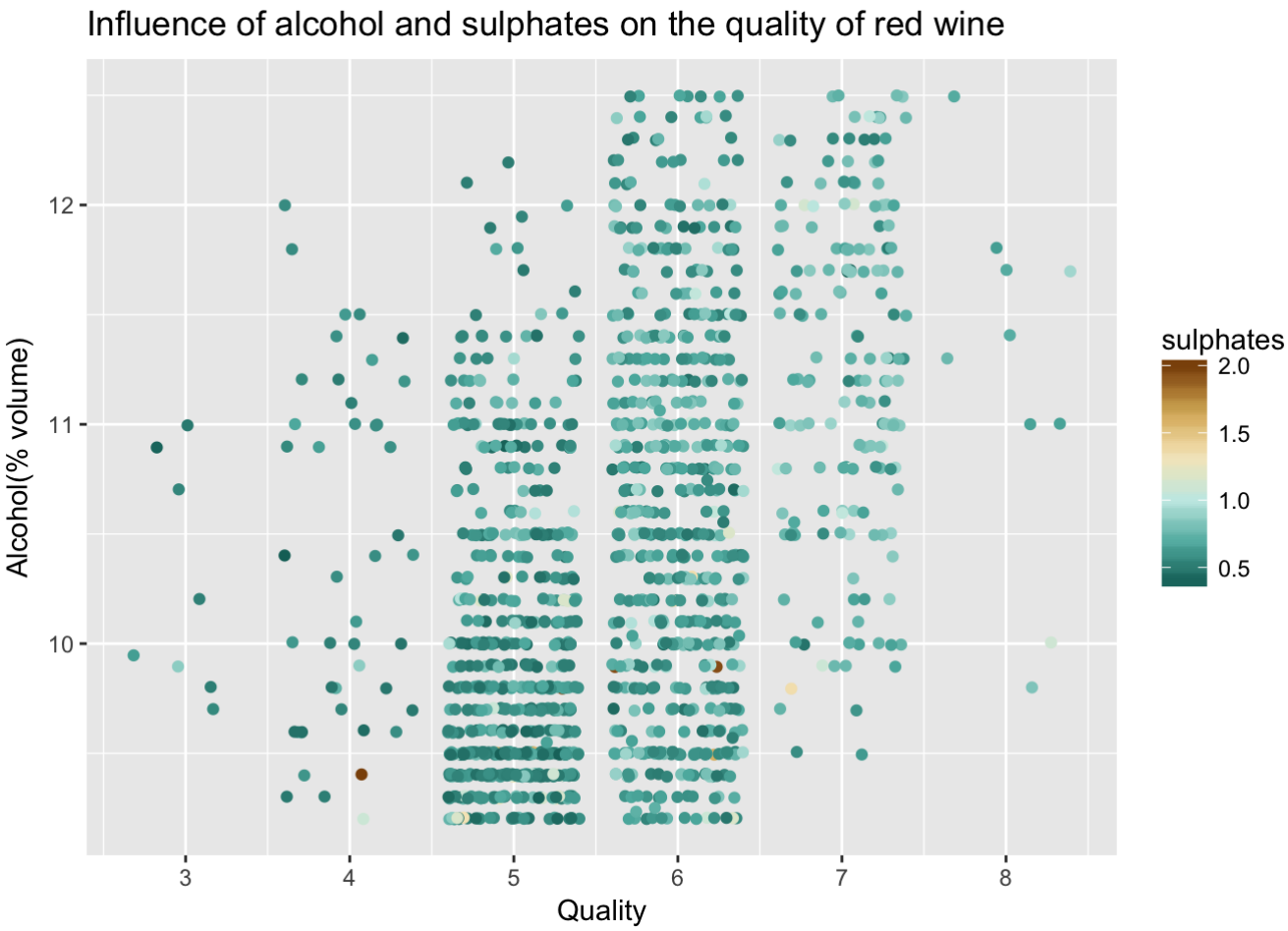
Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

In this data the main feature of the interest i.e. quality of the wine depends positively on the alcohol of the red wine.It also varies negatively with the amount of volatile.acidity or acetic.acid. The quality of the red wine improves with amount of sulphate as it is added to limit the production of acetic acid. Other than that the quality of the wine also depends on citric acid which adds freshness in the wine. ### Did you observe any interesting relationships between the other features
(not the main feature(s) of interest)? The most interesting feature in this data is positive correlation between volatile acidity and pH. As one might expect the negative correlation between the two features as pH decreases with increased acidity. My understanding is There is some lurking variable responsible for this observation that we have to figure out using multivariate analysis. Other than that alcohol depends negatively on volatile.acidity. This dependence can be explained as the presence of unwanted microbs that start decomposing alcohol in to acetic acid reducing the percentage of alcohol and increasing volatile acidity. ### What was the strongest relationship you found? The three strongest relationships are: citric.acid and acidity free.sulfur.oxide and total.sulfur.oxide quality and alcohol

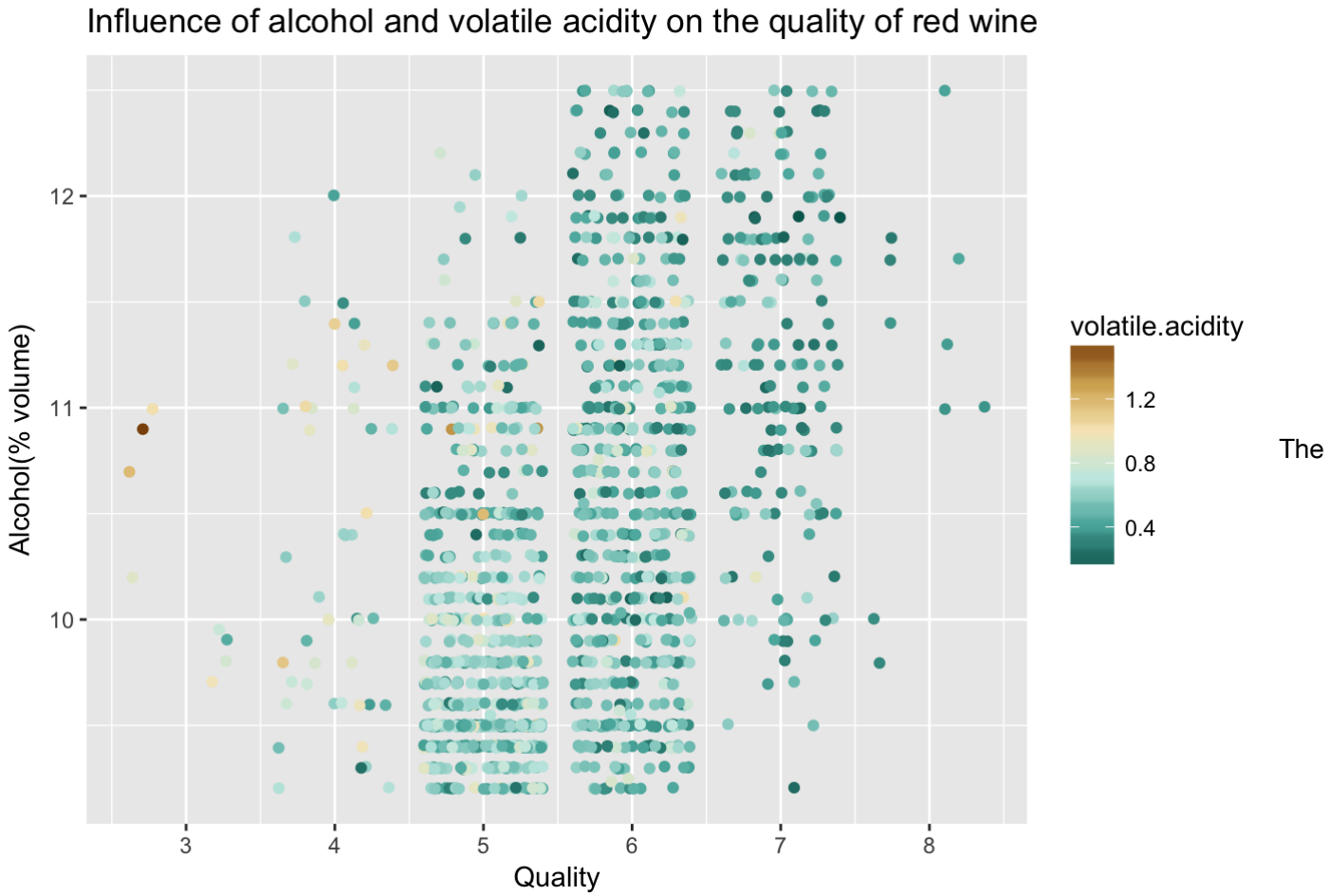
Multivariate Plots Section



In this analysis we observe in 7 and 8 quality wine some data points with low alcohol content have higher citric acid. This implies that high alcohol and/or high citric acid are features of higher quality of wine.

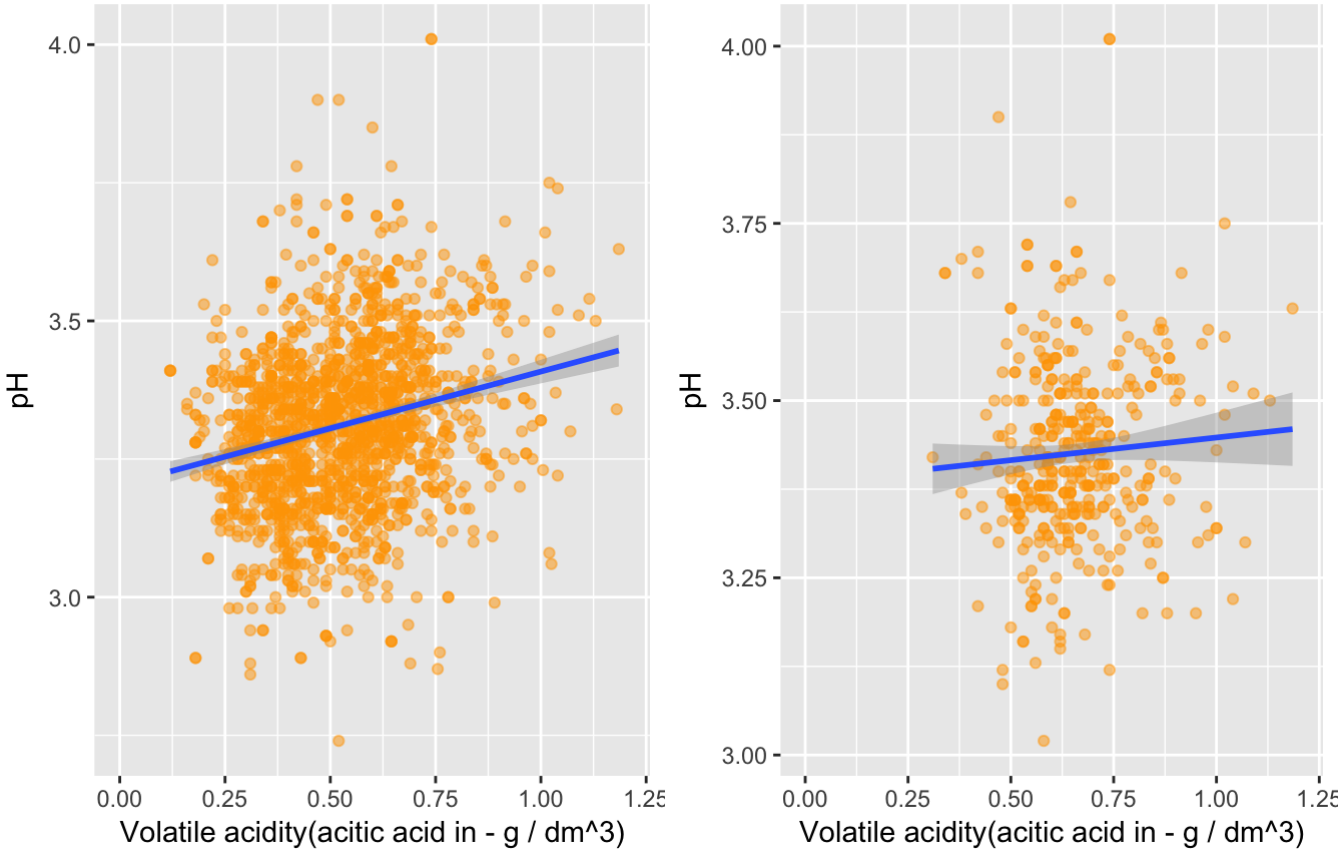

Influence of alcohol and sulphates on the quality of red wine
Alcohol(% volume)
12
11
10
3
4
5
6
7
8
Quality
sulphates
2.0
1.5
1.0
0.5

The multivariate plot of alcohol vs quality and sulphates inform that higher alcohol percentage usually go with higher sulphates and belong to higher quality of red wine.

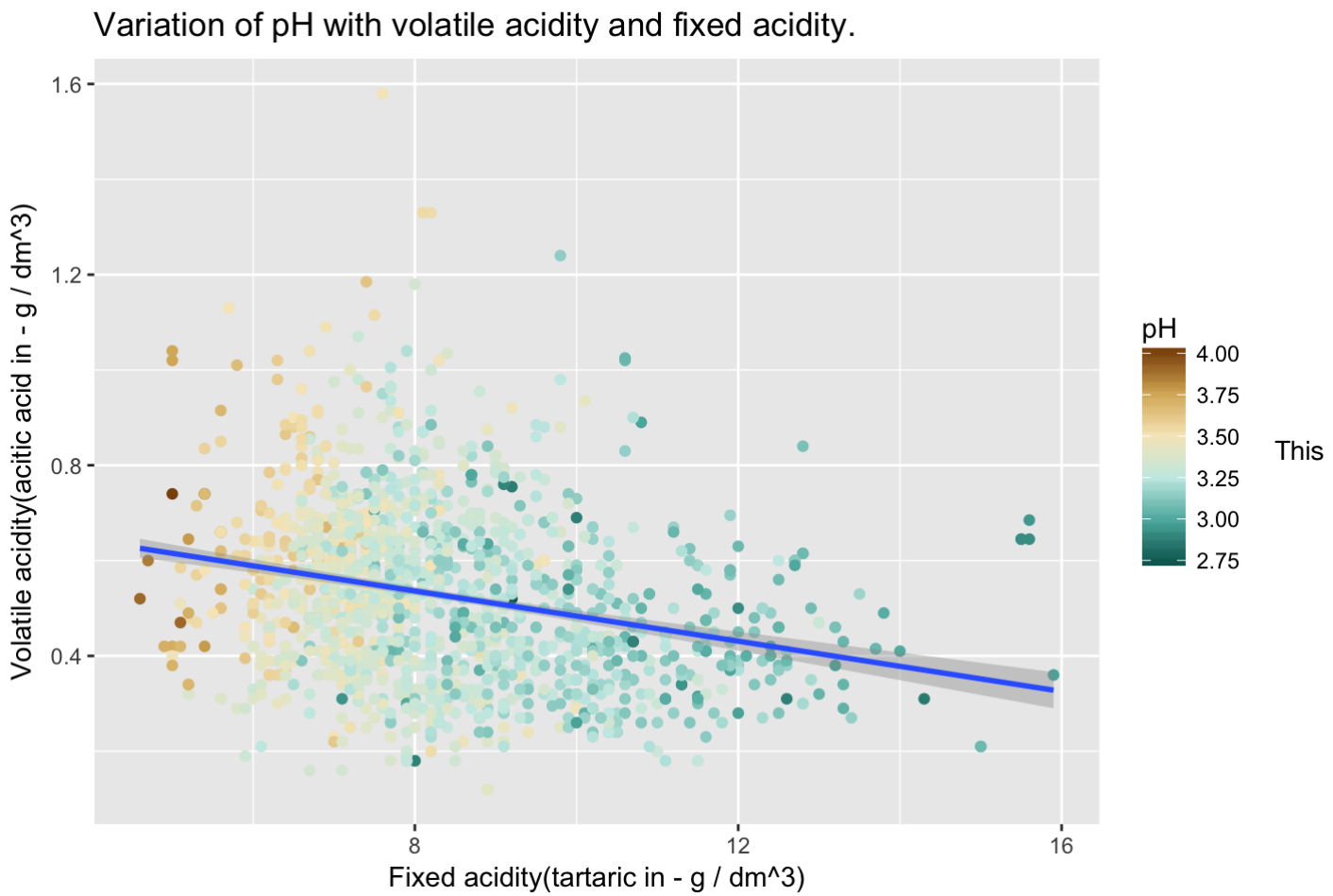


The effect of volatile acidity on alcohol and quality of red wine. We observe here that higher volatile acidity is found only in 5 and lower quality red wine. In this part we see that even with higher alcohol high content of volatile acidity reduces the quality of red wine.

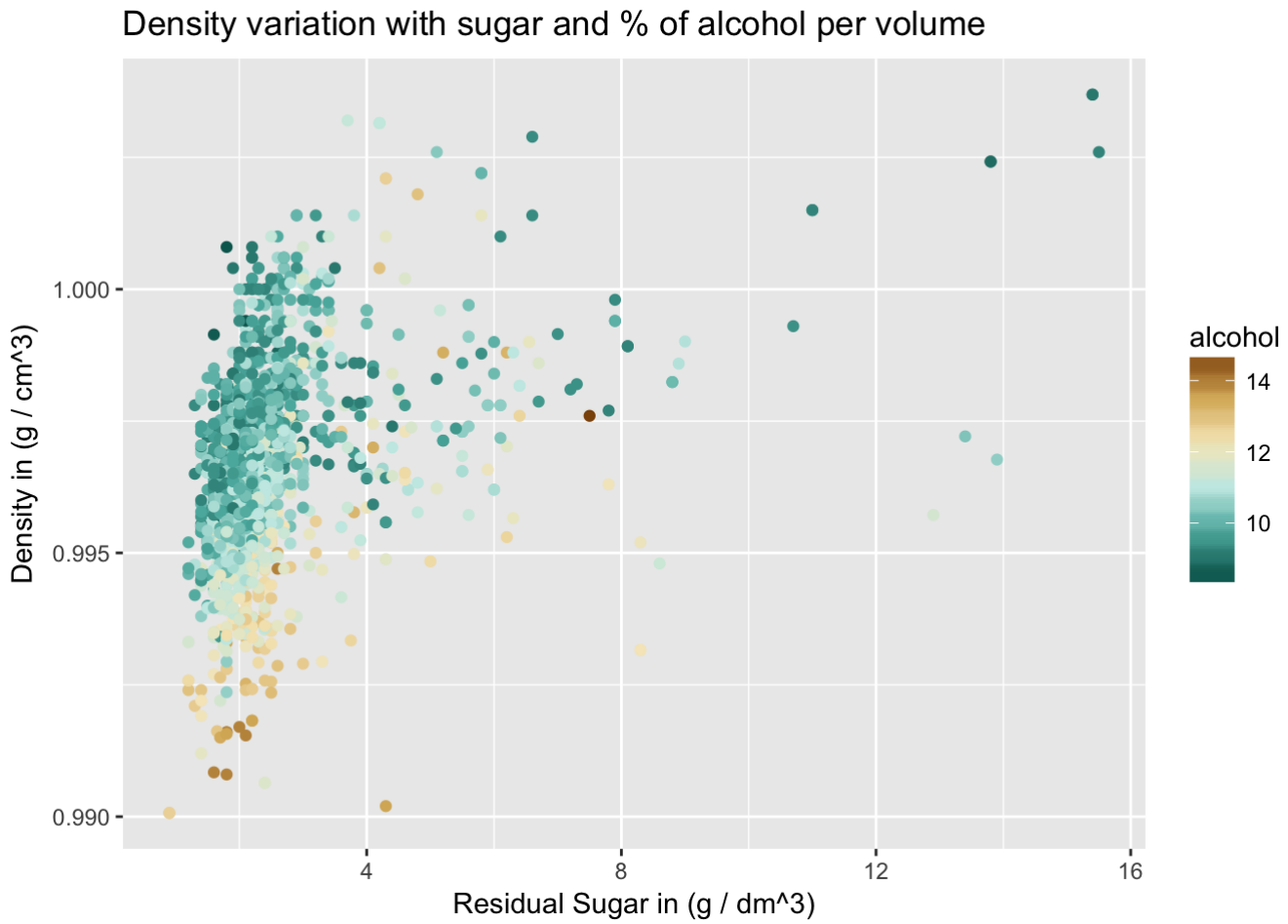
Change in correlation between pH and volatile acidity
in absence of citric acid.



The purpose of this plot is to find out if citric acid is the lurking variable resposible for positive correlation between volatile acidity and pH of red wine. The first plot shows the variation of pH with volatile acidity for all data points while the second plots has been plot for data points where value of citric acid is less than 0.3. We see the positive correlation is reduced in second graph.



plot helps us in analyzing the effect of fixed acidity and volatile acidity on pH. we see here the high value of pH belongs to the left side of the graph where fixed acidity is low. For a given value of fixed acidity we do not observe any increase in pH with increase in volatile acidity. Hence the positive correlation of volatile acidity and pH is observed because of their negative correlations with fixed acidity in red wine.



This plot shows how density of wine changes with residual sugar and percentage volume of alcohol.

Multivariate Analysis

In bivariate analysis of quality vs alcohol, we observed that in general quality is positively correlated with alcohol but there are still many data points with the high quality and low alcohol. To confirm this observation, we look at the summary of alcohol for quality= 3, 6 and 8 respectively.

```
{r echo=false, Multivariate_Plots}
with(subset(red, quality ==3), summary(alcohol))
with(subset(red, quality ==6), summary(alcohol))
with(subset(red, quality ==8), summary(alcohol))
```

Though, we see strong correlation between alcohol and quality of red wine, evidently, we can not predict the quality of the wine alone with the percentage of alcohol. Therefore, in order to find out the other features influencing the quality of the red wine we include the other three features affecting quality such as citric.acid, sulphates and volatile acidity. The alcohol vs quality plots with color as citric.acid shows that lower alcohol percentage in high quality red wine have high content of citric.acid. The multivariate plot of alcohol vs quality and sulphates inform that higher alcohol percentage usually go with higher sulphates and belong to higher quality of red wine. Still, these plots lack the consistency to state that the lack of alcohol in higher qualities is compensated by high sulphates and/or citric acid and/or low volatile acidity. Therefore in this multivariate analysis we replace sulphates by volatile.acidity as the primary goal to add sulphates in wine is to keep the volatile acidity minimal. Evidently, the plot of alcohol vs quality with volatile acidity consistely tells the negative correlation with both alohol and quality and clearly we can say that higher quality red wines have low content of volatile acidity and most of

the time higher volatile acidity results in to lower percentage of alcohol. Thus we conclude that lower volatile acidity is a reliable measurement of quality of the red wine. Still, if we observe all three plots we find many examples data points where similar proportion of alcohol/volatile.acidity/ citric acid lies in different quality, suggesting there are more features contributing to the quality of the red wine which has not been included in this dataset.

The change in density in red wine is an interplay between content of residual sugar and alcohol as can be seen in this multivariate analysis. In red wines with very low density, the content of alcohol is relatively high and that of residual sugar is low while for rather thick wines are relatively sweeter than the other wine with lower alcohol content. We gather from the data that most of the red wines (95%) have sugar content of less than 5.0 g/dm^3. Hence the change in density in red wine is caused more by the change in alcohol than change in sugar.

The most interesting analysis of this dataset is the positive correlation of volatile acidity with pH. As we observed in bivariate analysis, the pH of the red wine is increasing linearly with volatile.acidity. I suspect there is some lurking variable in play for this observation. my hypothesis is lower volatile.acidity is associated with presence of higher citric.acid or tartaric acid. This is possible if acid provides resistance to microbial infection. To study the role of citric acid in positive correlation of volatile acidity with pH. I compare the bivariate analysis of pH with volatile.acidity with and without citric.acid. We observe that for citric acid < 0.1 g.dm^3 the volatile acidity has almost no effect on pH. But we expect this relation to be negative that implies that tartaric acid is also a lurking variable in this scenario. There are no wines without tartaric acid so here we do multivariate analysis of volatile. acidity, fixed.acidity and pH. Here we observe that Higher pH corresponding to higher volatile acidity also corresponds to low fixed acidity and at fixed acidity greater than ~ 11 g/dm^3 there are no data points with pH higher than 3 and volatile acidity greater than 0.9.

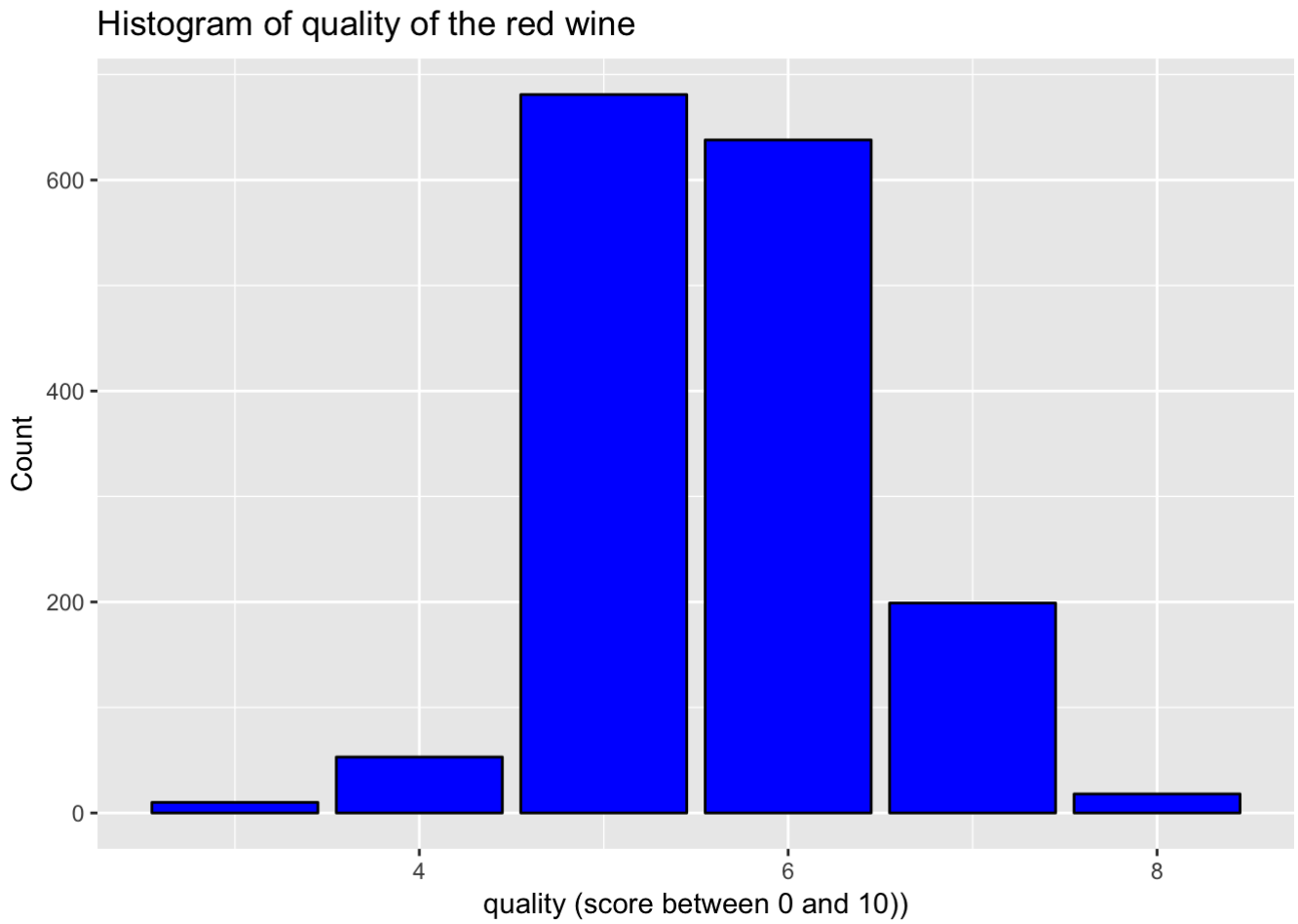
Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Were there any interesting or surprising interactions between features?

The interaction of volatile acidity with pH is interesting as it shows a positive correlation opposed to general expectations. The study reveals that citric acid and tartaric acid are working as lurking variables for this phenomena. It turned out to be a beautiful example of the Simpson's paradox.

Final Plots and Summary

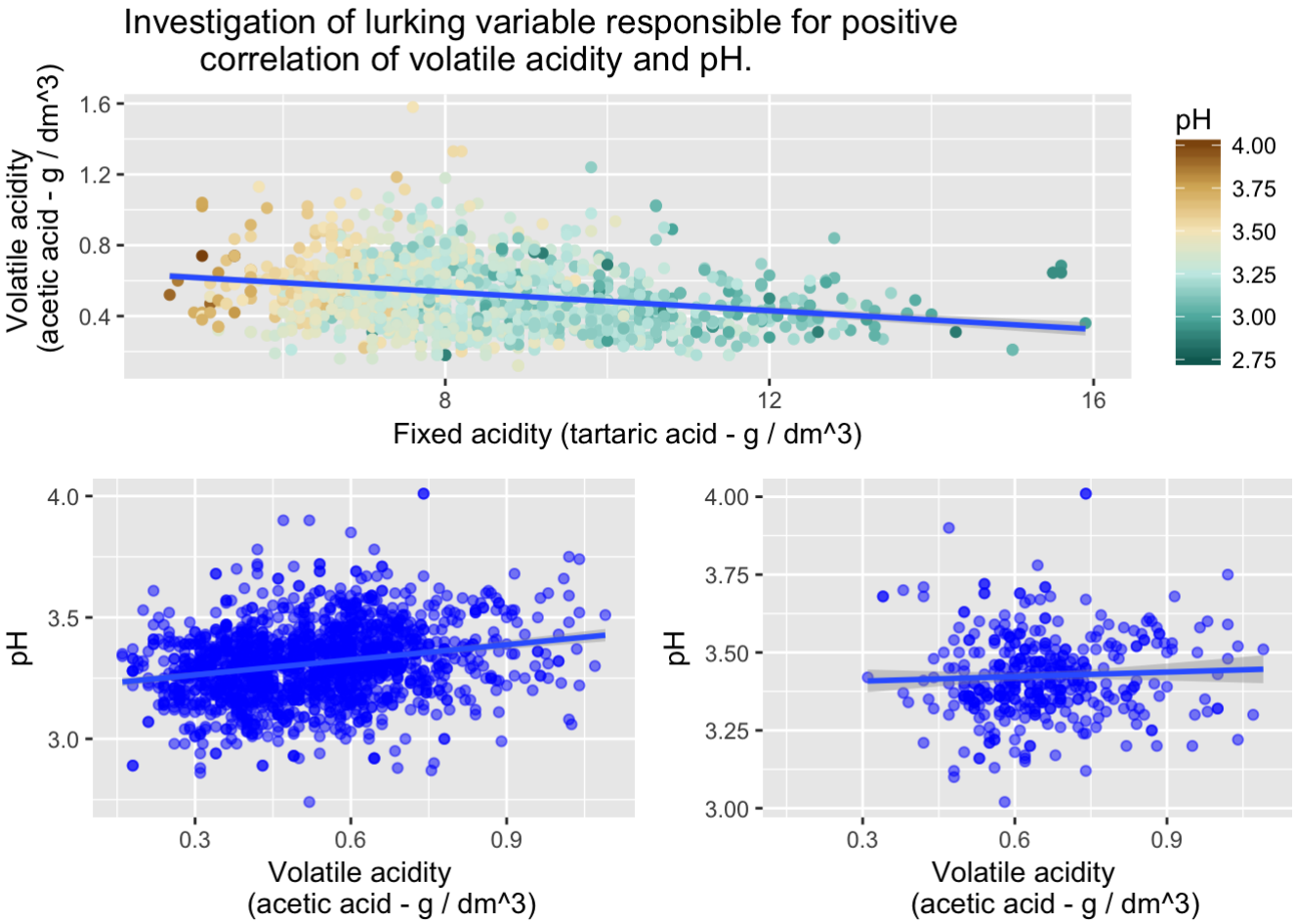
Plot One



Description One

I have added this plot to emphasize on the polarization of this sample. As we can see the counts of average quality is nearly 85% of all red wines. Hence any modeling for the quality of the red wine might not be true for high quality and lower quality red wines.

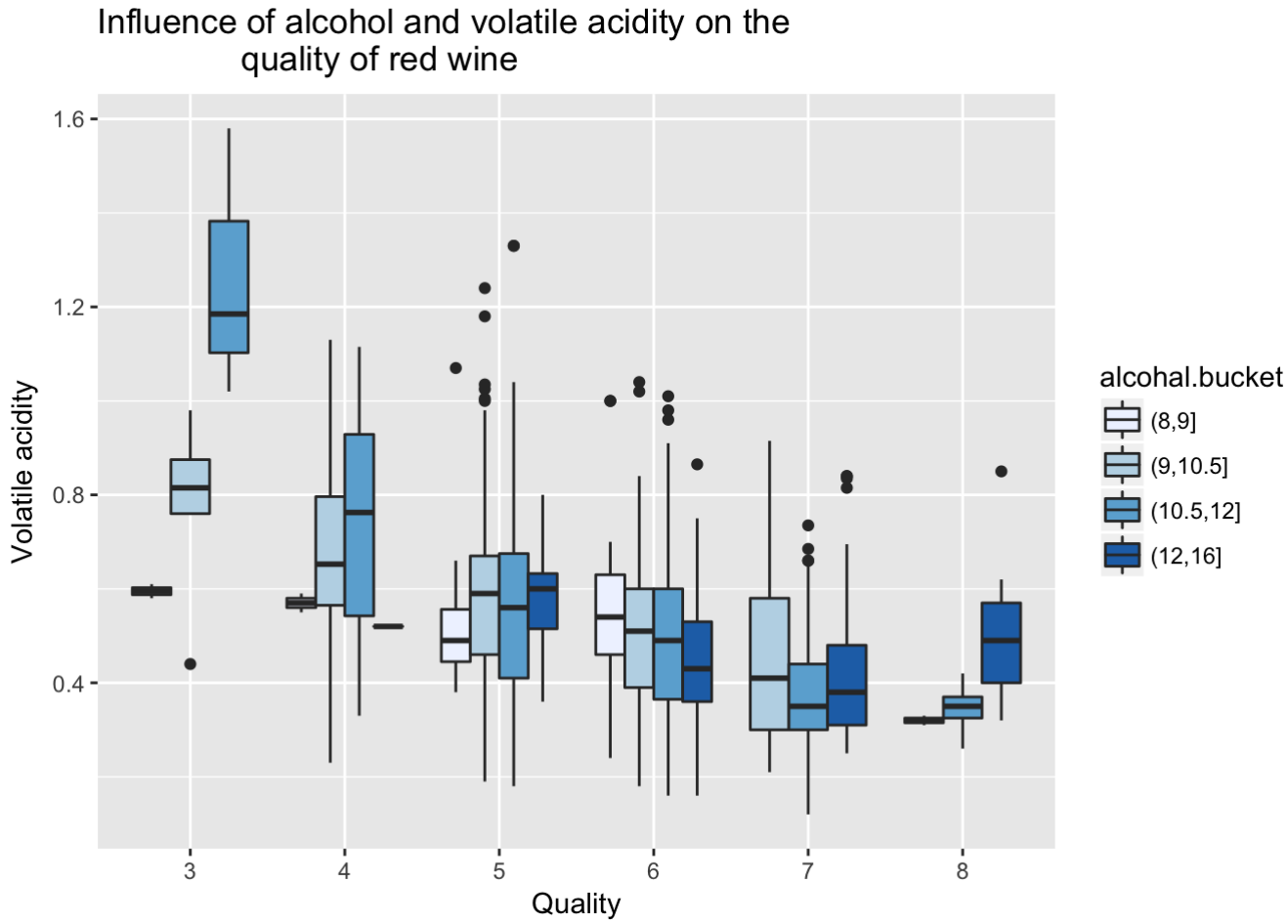
Plot Two



Description Two

with the helpof these plots I have tried to interpret the negative correlation of volatile acidity to pH and find the lurking variable. In the first graph we see that pH decreases linearly with the increase in fixed acidity. However, The variation of pH with volatile acidity is in fact not linear. Only below a threshold value of fixed acidity ~ 8 the higher value of volatile acidity correspond to higher pH. Since, for some reason, we have here more data points with high volatile acidity and low fixed acidity, we observe the negative correlation between volatile acidity and pH. To validate this point further, I plot pH vs volatile acidity and compare it with pH vs volatile acidity for ~ 0 content of citric acid (as citric acid is a part of the fixed acidity). In third plot we see the correlation of volatile acidity and pH is almost negligible. This shows that presence of citric acid is not only decrease the pH but also also production of acitic acid hence this effect appears in to positive correlation of volatile acidity to pH.

Plot Three



Description Three

This plot shows the two main variables alcohol and volatile acidity(acetic acid) contributing to the quality of the red wine. The plot supports that for a better quality of the red wine we need higher % of alcohol and lower volatile acidity. The fact that some data points, with nearly same proportion of alcohol, citric acid and acetic acid lie in different quality range suggests there may be more variables which constitute significantly in making a good red wine.

Reflection

I chose this data set of quality of red wine as this is a field, I am completely unaware of. By the time I finished it I can say I know more than many about what makes a red wine great. Just after analysing a few major features of this data sets I realized that this data set is missing the band width which is important to give a strong theory about the quality of wine. In this analysis, I found that higher quality red wine usually tends to have higher percentage of alcohol per volume but higher quality does not ensure higher percentage of alcohol. On the other hand, higher quality do ensure lower volatile acidity(acetic acid) in wine which is often attained by higher sulphates in red wine. The dependence of citric acid on quality of the red wine is also consistent. We can safely say that higher citric acid indicates better quality of the wine but not vice-versa.

In this dataset 4 variables seem to affect the quality of the wine but apparently there are more important variables are in play which had not been included in this dataset. The Age of the wine and tannine are couple of examples.

Reference

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

[2] https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt) ‘” ‘”