

# Exploring Red Wines by OZGUN BALABAN

## Overview

In this project a sample set of red wines are analyzed for 12 different variables. I selected red wines as a study since I am curious in learning what makes of a good wine. We will first start with summarizing the data and identifying every parameter. In a data study it is important to make a background study to understand the phenomena.

## Summary of the data

```
summary(red)
```

```
##           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
```

Analyzing the data, we have 1599 observations with 13 variables. These are;

**X** -> unique id

**fixed.acidity** -> acid produced in the body from sources other than CO2

**volatile.acidity** -> acids produced by yeast and bacteria

**citric.acid** -> rarely found in wine, can be used to boost wine acidity

**residual.sugar** -> determines sweetnees of the wine

**chlorides** -> saltiness of the wine (not a good thing)

**free.sulfur.dioxide** -> sulfur oxide prevents bacteria growth

**total.sulfur.dioxide** -> total amount

**density** -> density of the wine - 1 is same as water

**pH** -> acidity level (it is alkaline if over 7 but all the values in set under 7)

**sulphates** -> an additive to prevent wine fault

**alcohol** -> alcohol level of the wine

**quality** -> scores by sommeliers higher is better

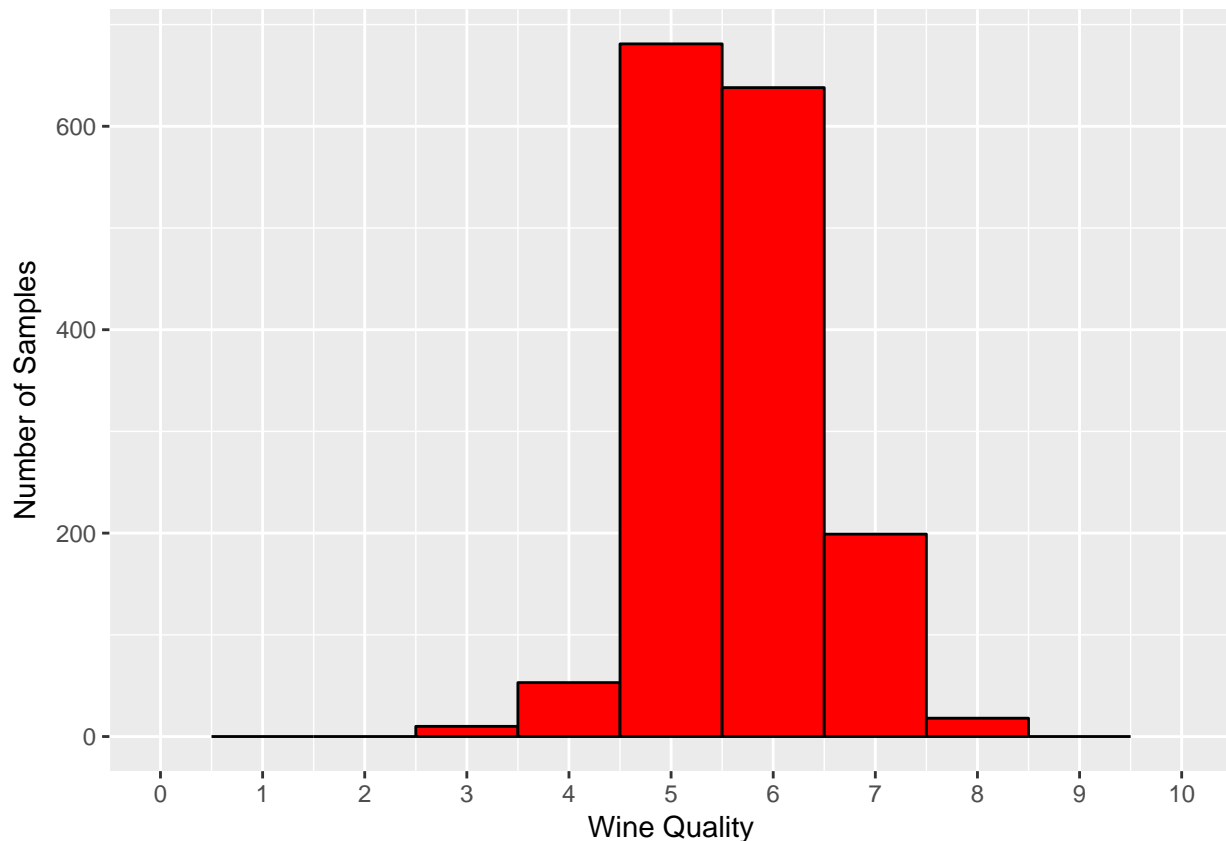
This listing of the parameters reveals some interesting relations. Acidity is one of the important things in wine. There are different opinions as to what level of volatile acidity is appropriate for higher quality wine. Although too high a concentration is sure to leave an undesirable, 'vinegar' tasting wine, some wine's acetic acid levels are developed to create a more 'complex', desirable taste. The renowned 1947 Cheval Blanc is widely recognized to contain high levels of volatile acidity.

Sugar level is also important in the 'taste' of the wine. Too much sugar might hide the complexity of the flavors. So looking on these parameters it is important to find out how these affect the overall taste. Some interesting questions that can be asked;

## Univariate Plots Section

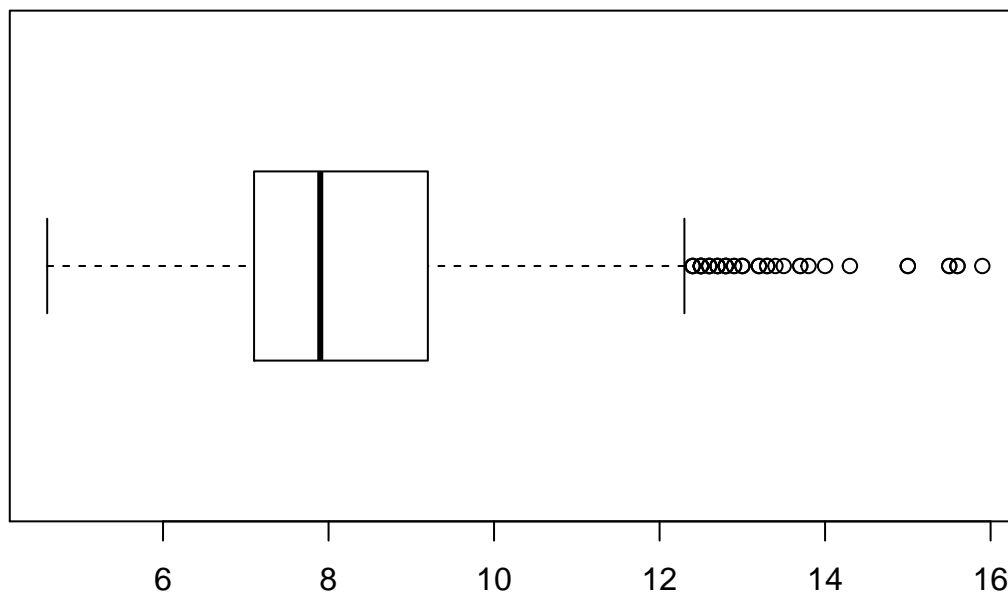
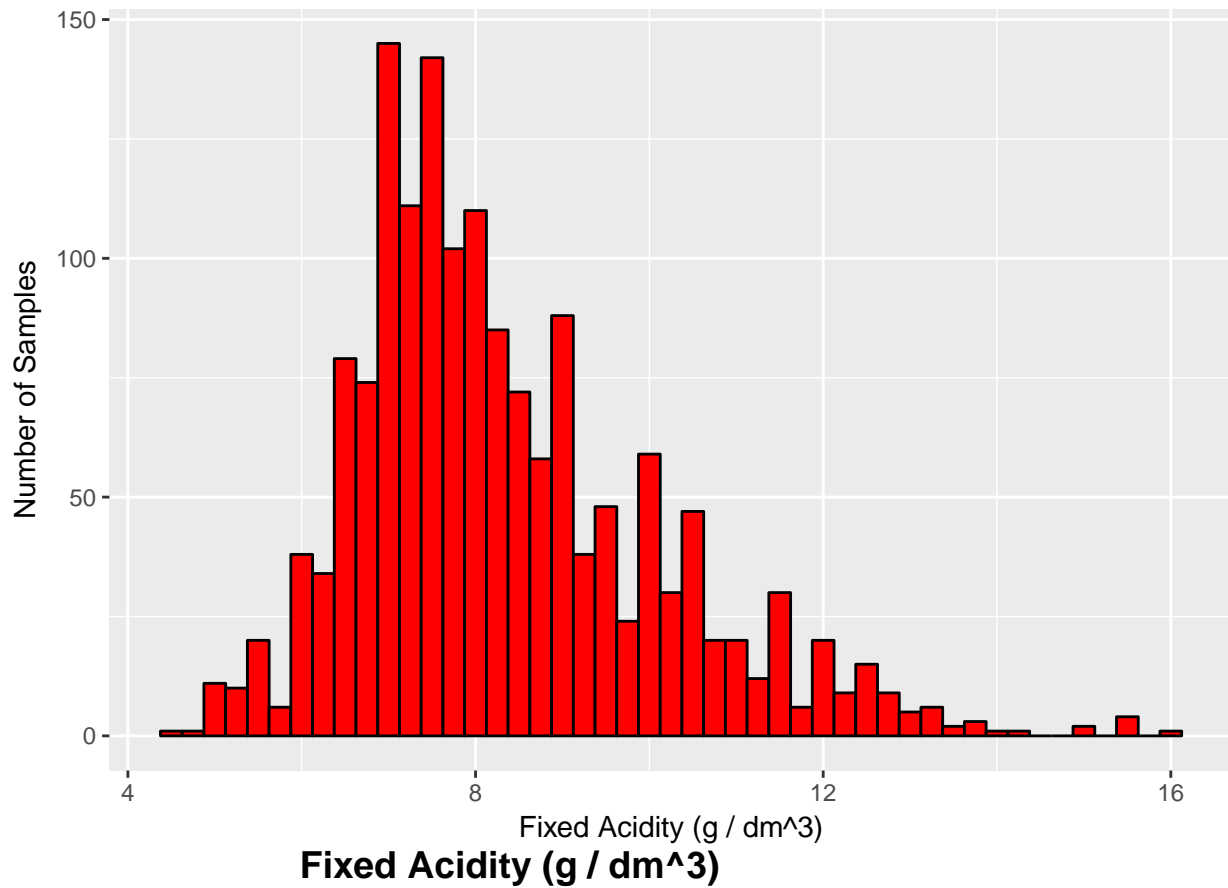
### Wine Quality

Wine quality is supplied by tasters. It can be in the range from 0 to 10. Plotting the wine quality;



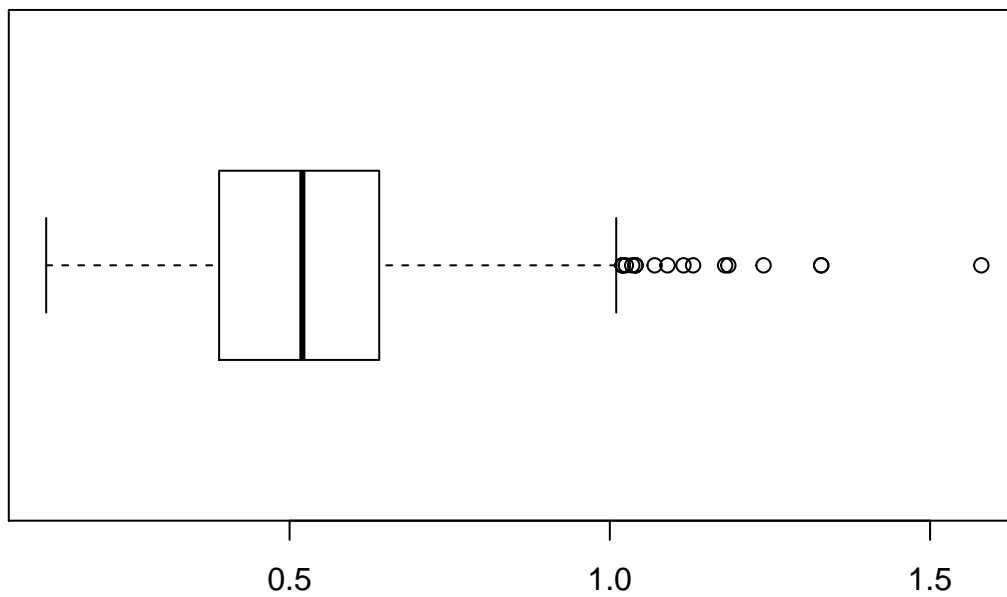
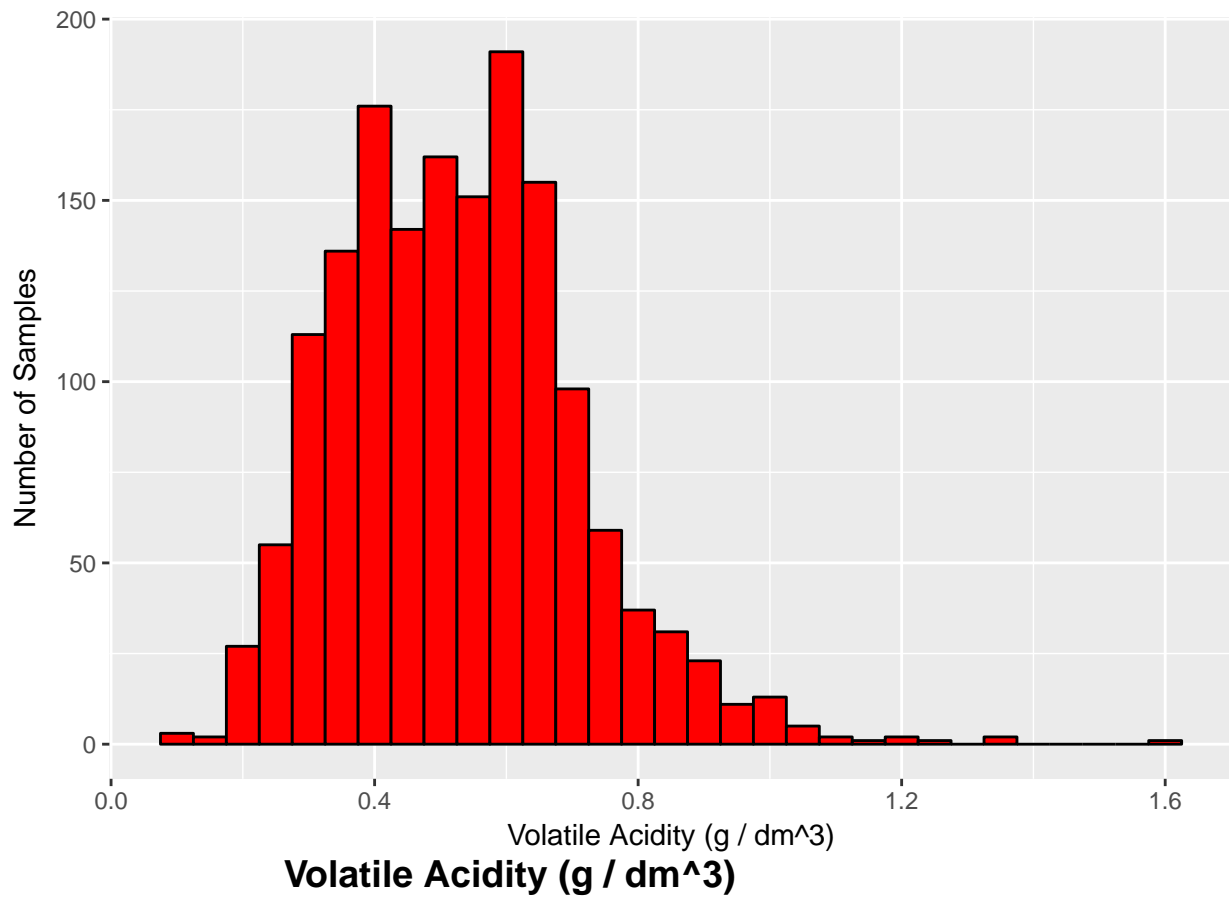
There is no outliers here. Minimum value is 3, maximum value is 8. We can see most of the wines are stacked around 5-6 and a little bit of 7.

## Fixed Acidity



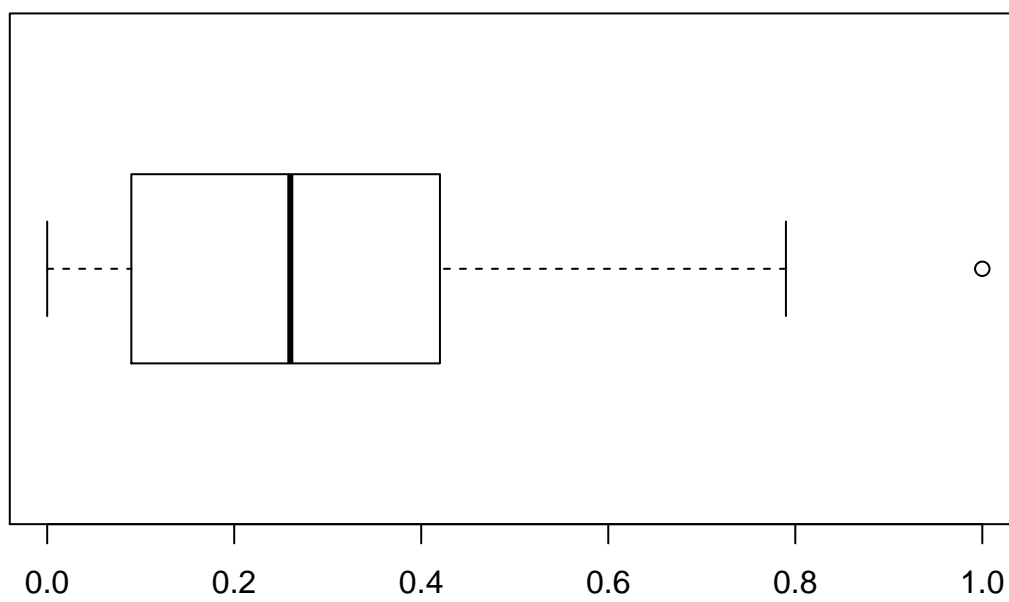
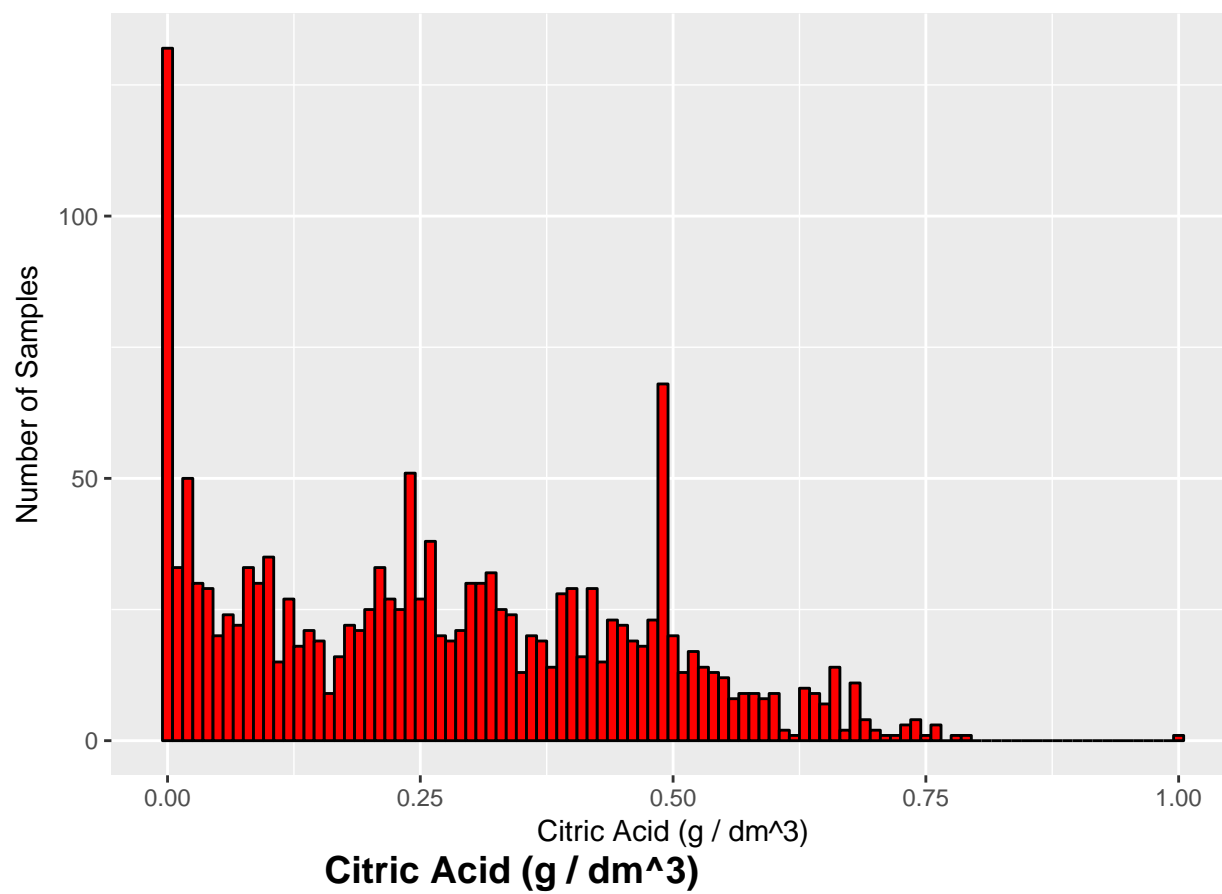
Here we see the mean around 8. There are few outliers in the right side of the data. From this graph we can conclude wines are generally more acidic than basic.

## Volatile Acidity

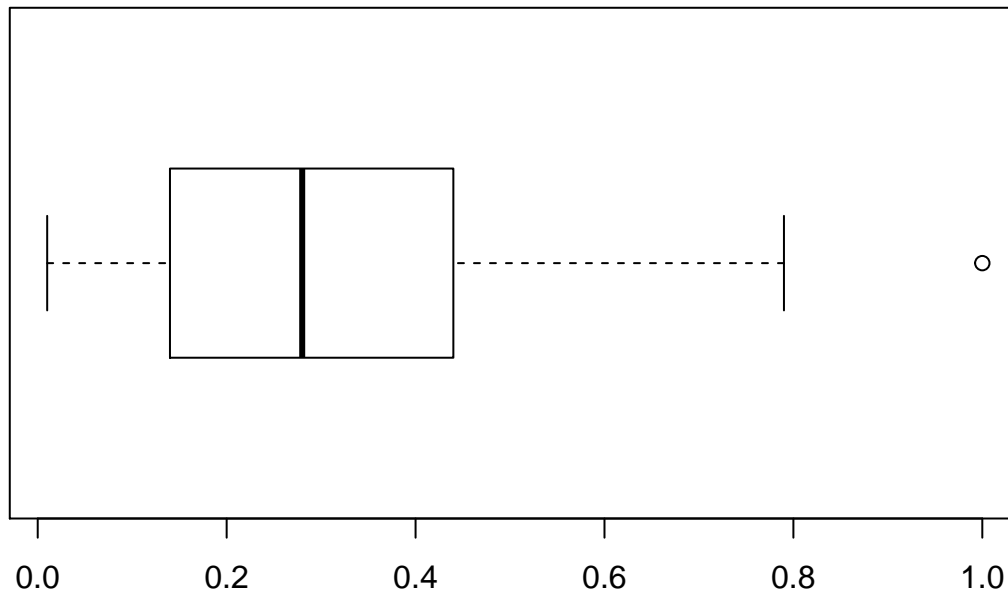


The mean here is around 0.5. There are some outliers again in the right side.

## Citric Acid



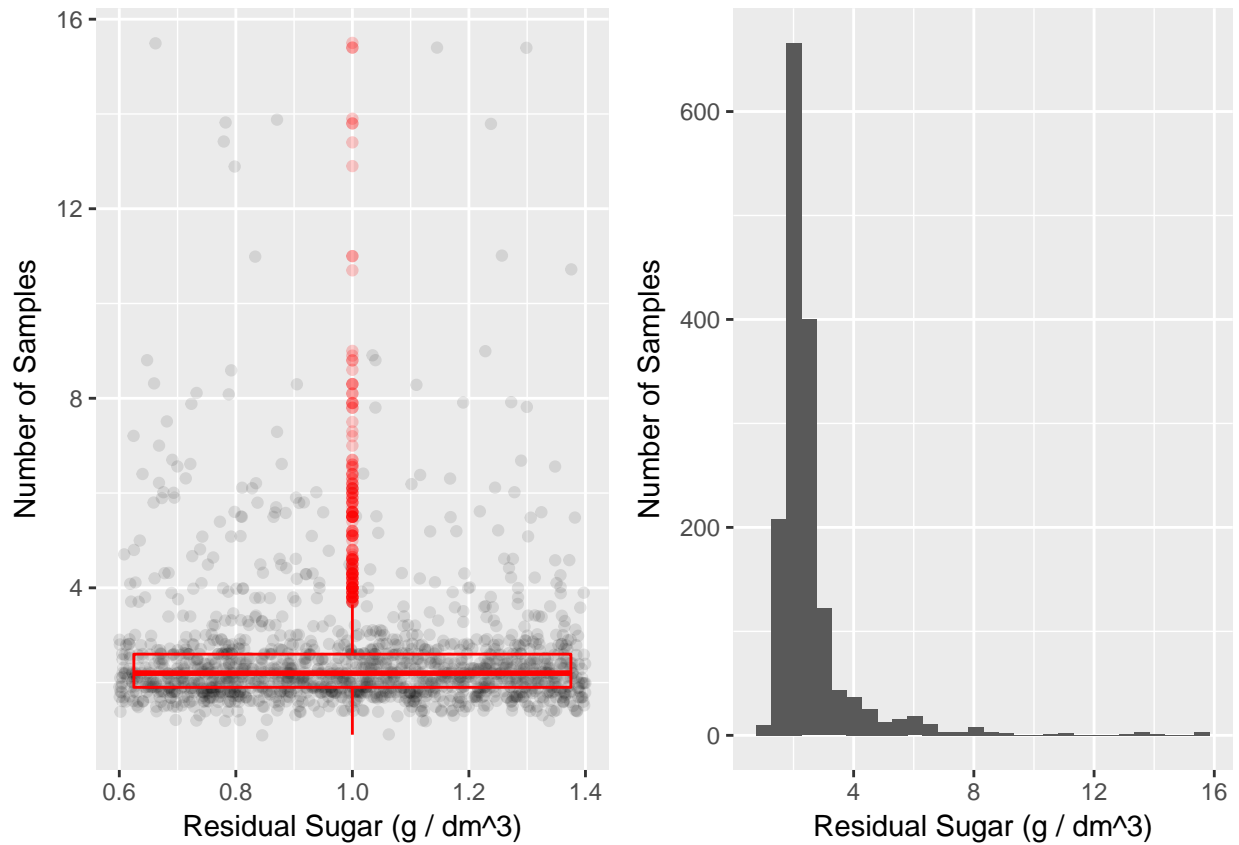
## [1] 132



From the previous observation we know that citric acid is rarely used to adjust acidity of wines. We have 132 wines with no citric acid. Searching on internet reveals that in EU using citric acid is banned. So it would be nice if we had the data of the origin of the wines used in this study.

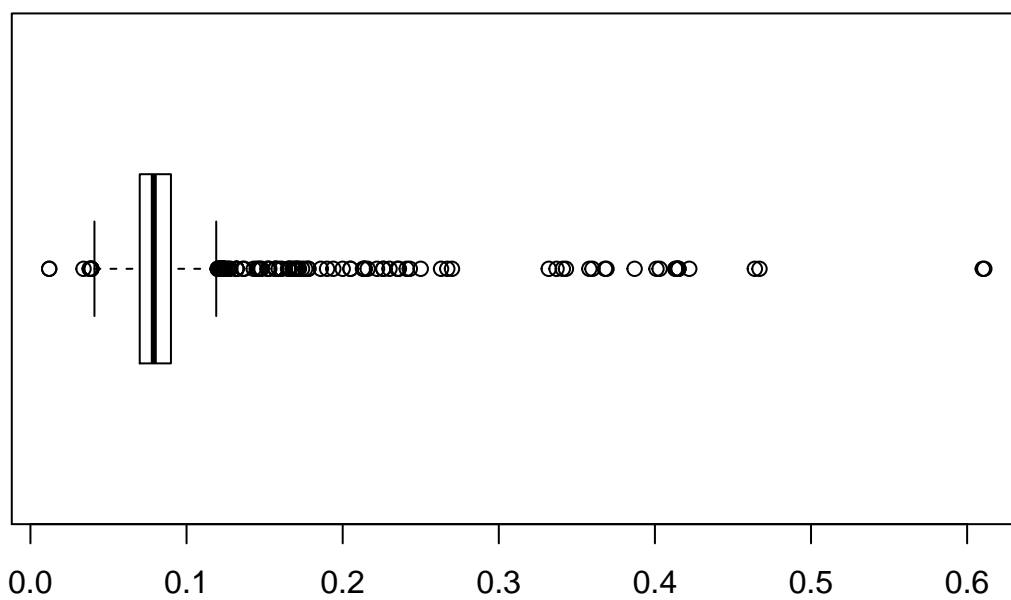
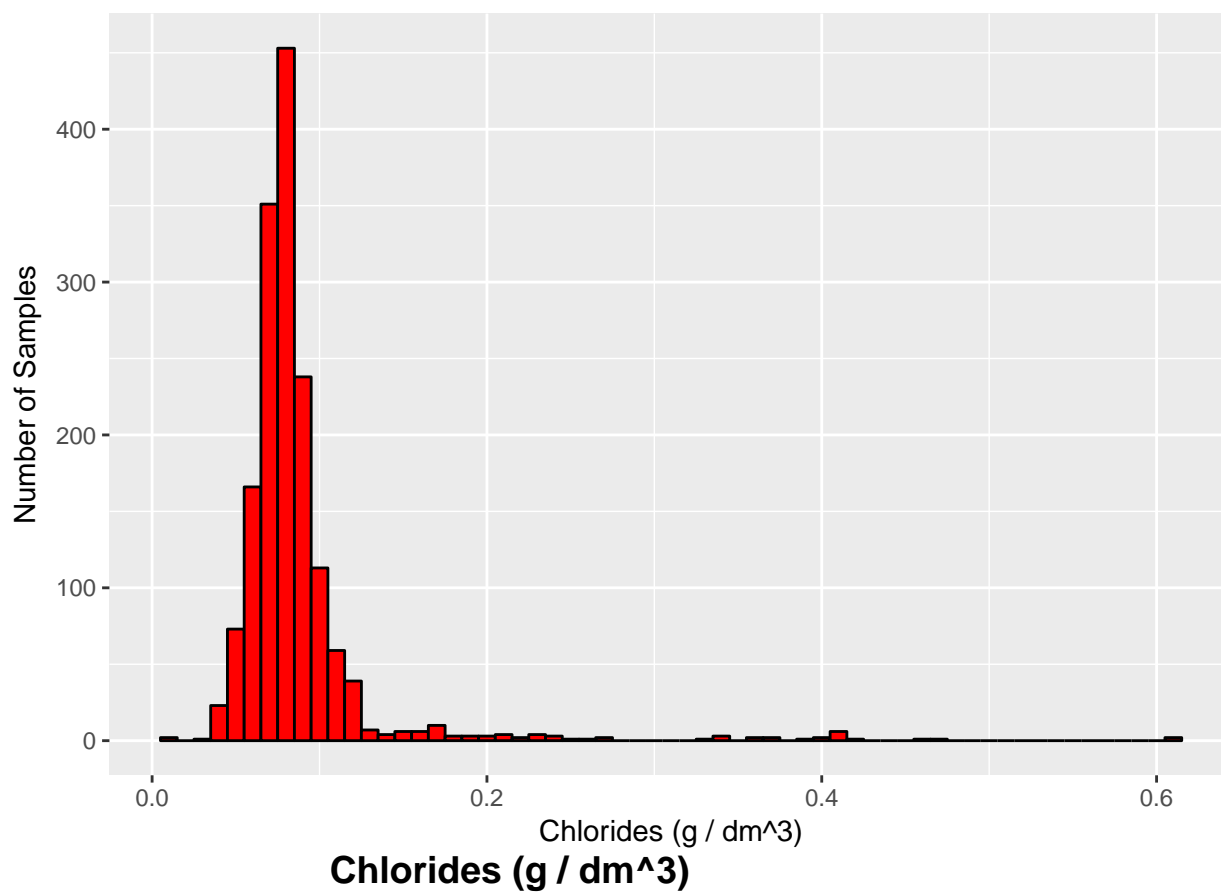
## Residual Sugar

```
grid.arrange(ggplot(red, aes( x = 1, y = residual.sugar ) ) +
              geom_jitter(alpha = 0.1 ) + xlab('Residual Sugar (g / dm^3)')+
              ylab("Number of Samples") +
              geom_boxplot(alpha = 0.2, color = 'red' ) ,
              ggplot(red, aes( x = residual.sugar ) ) +
              xlab('Residual Sugar (g / dm^3)')+
              ylab("Number of Samples") +
              geom_histogram(bins=30 ),ncol=2)
```



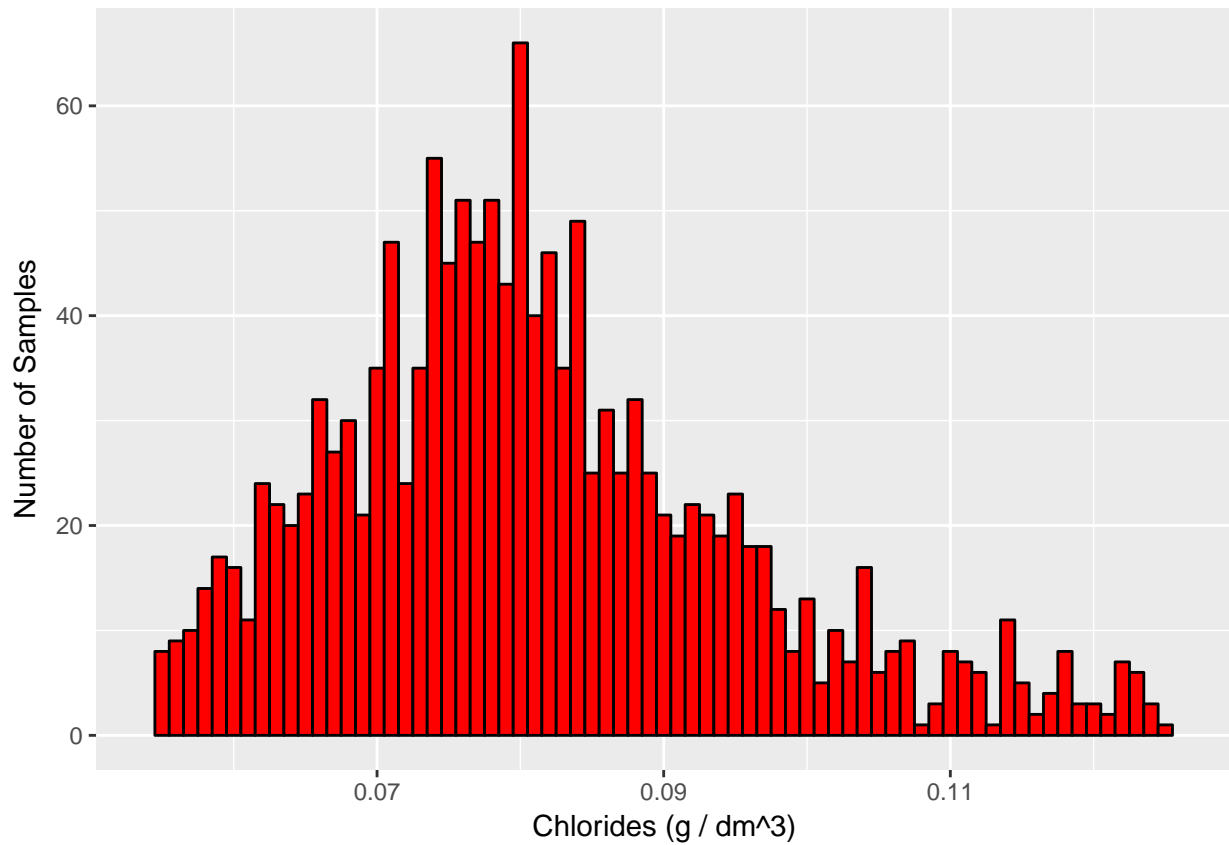
Residual sugar has many outliers. Removing the outliers I plotted again. Sugar level will probably influence the taste of wine. So many outliers also mean that sugar level is open to varied taste and experimentation. People generally have varied opinions about sweet wines.

## Chlorides



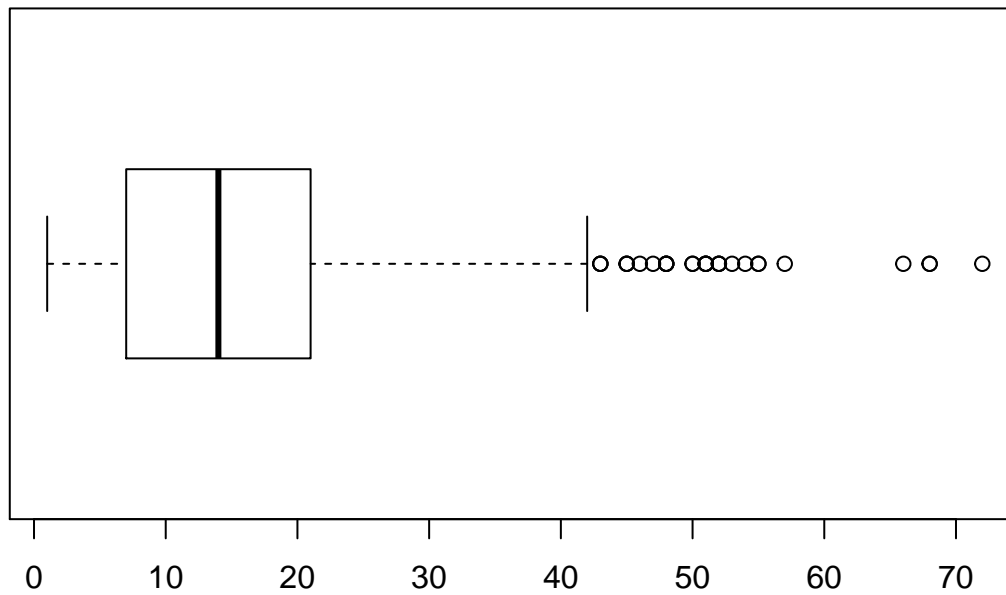
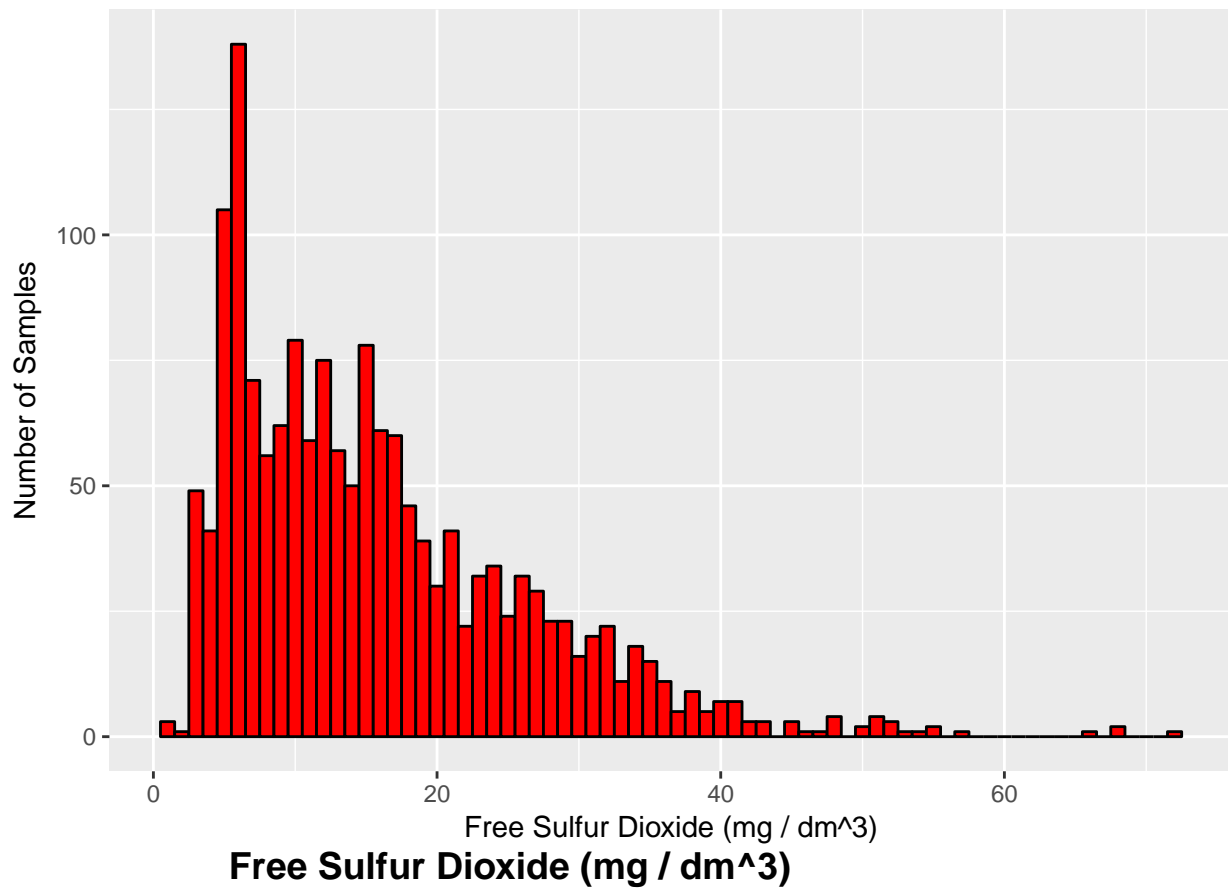
## Warning: Removed 158 rows containing non-finite values (stat\_bin).





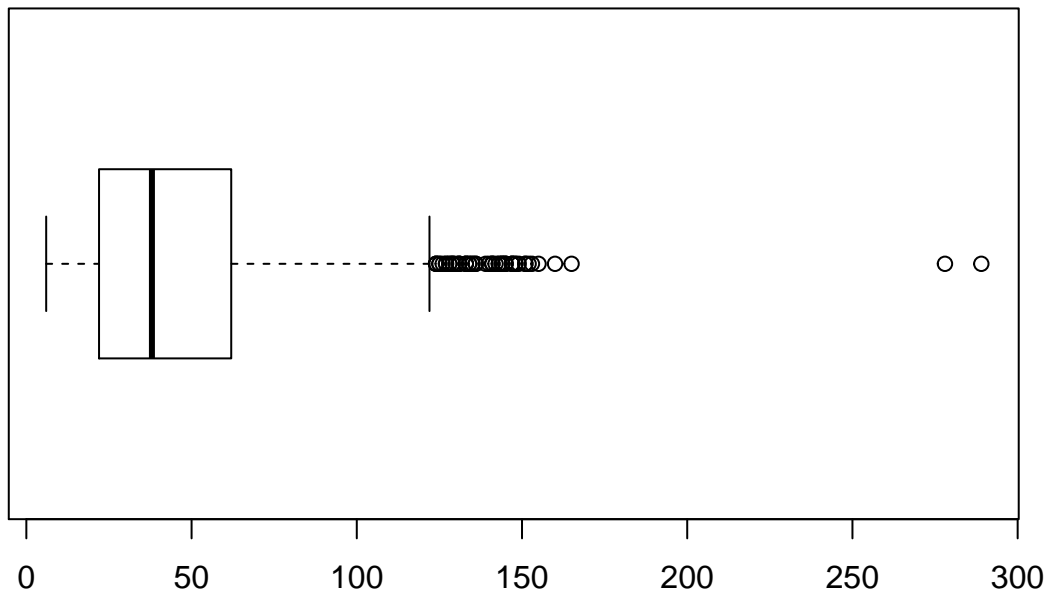
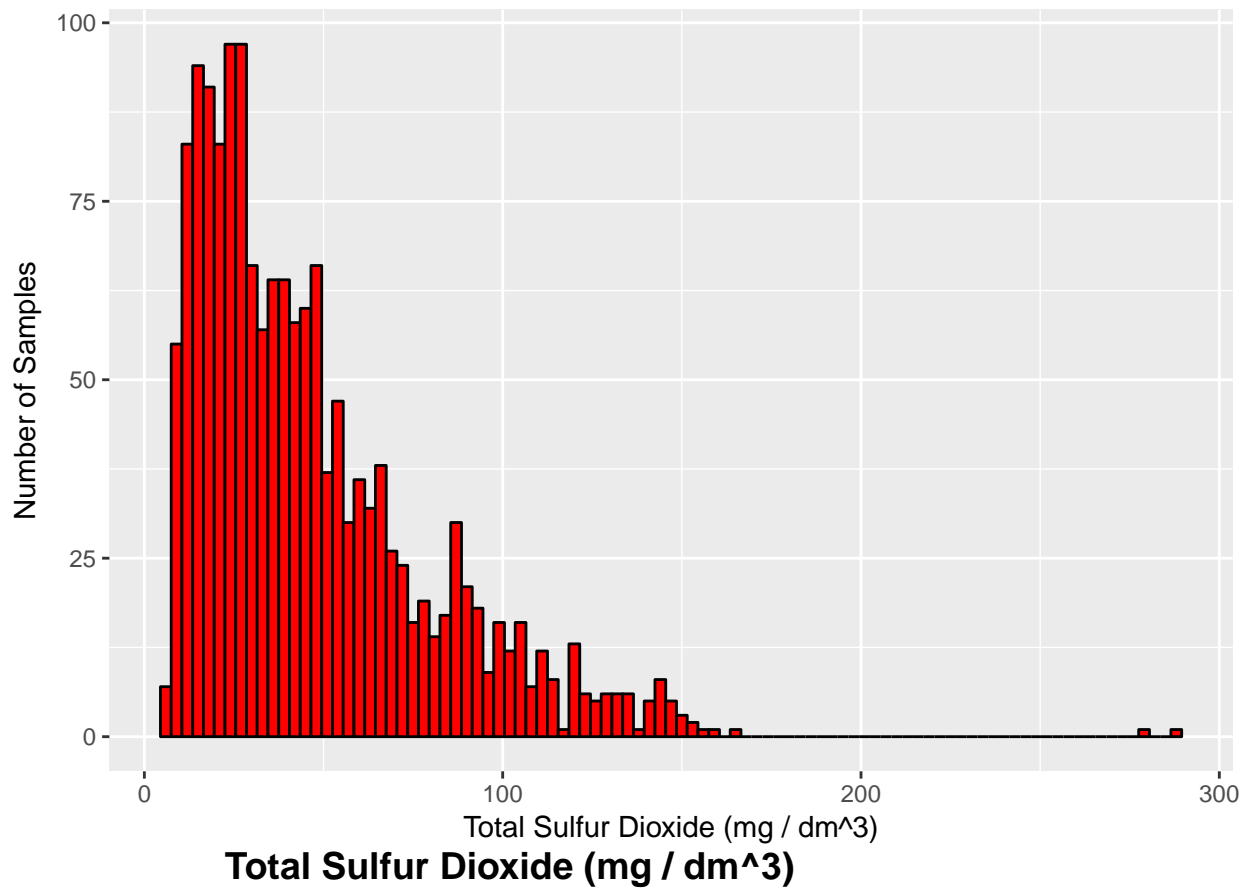
Chlorides have many outliers as well. Removing the extreme values, I plot it again. Chlorides give unwanted salty smell but in turn it protects the wine, so too much Chloride might make the wine taste bad but it might stop spoilage.

## Free Sulfur Dioxide



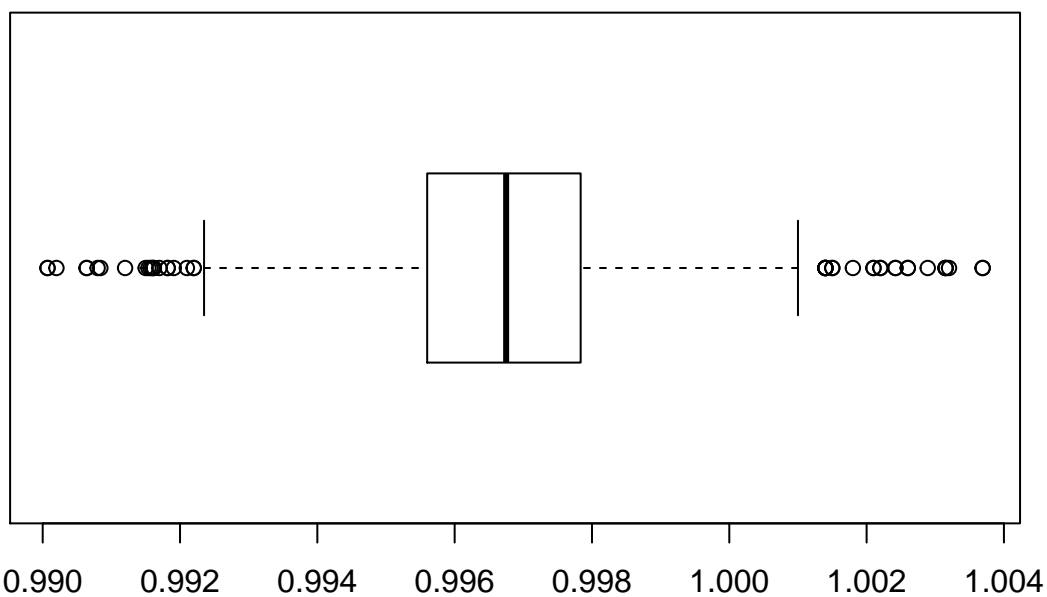
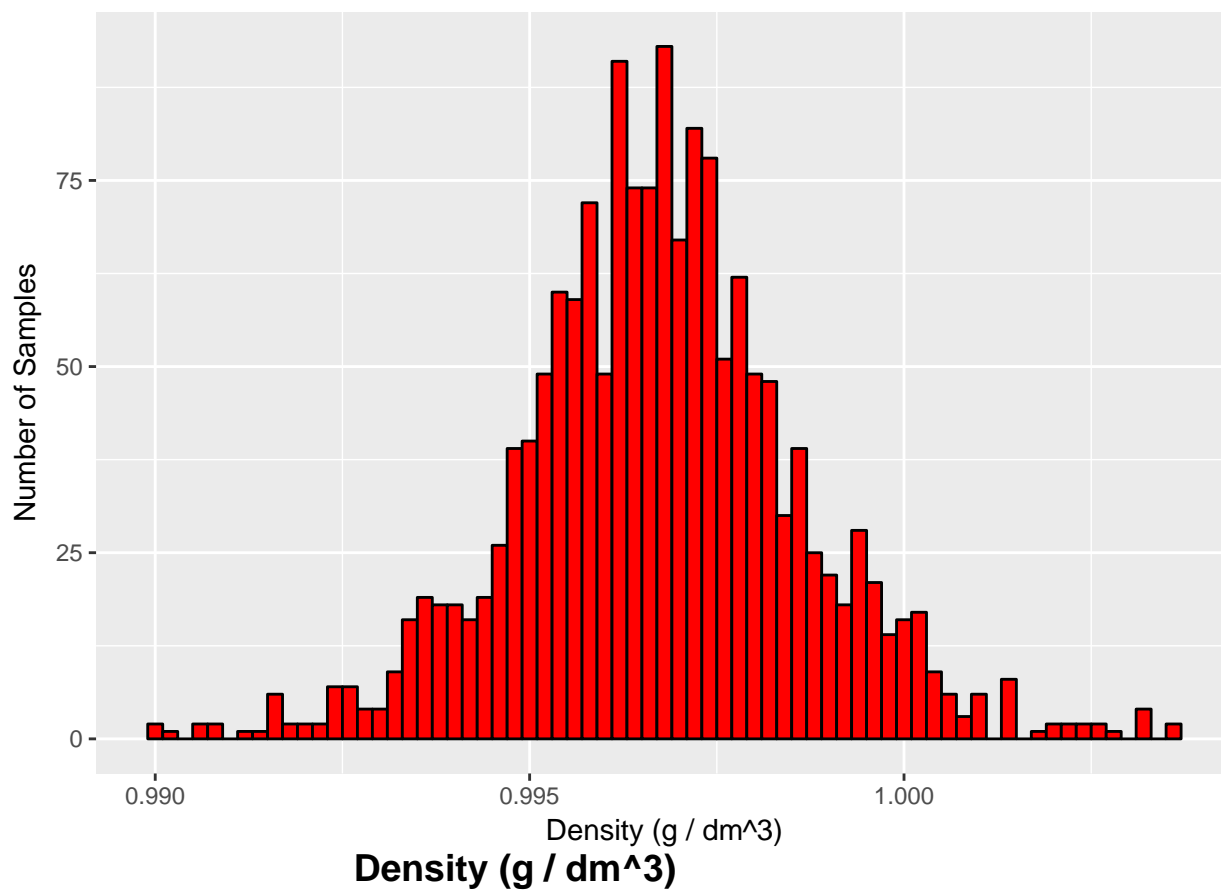
Average Free Sulfur Dioxide is around 15 with few outliers. Though outliers can be quite away from the mean value.

## Total Sulfur Dioxide



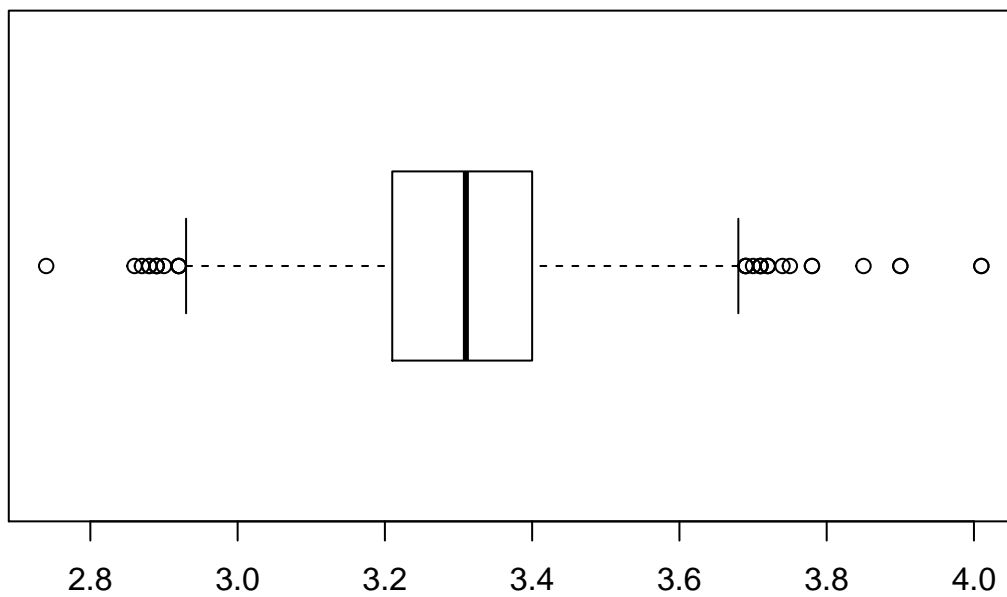
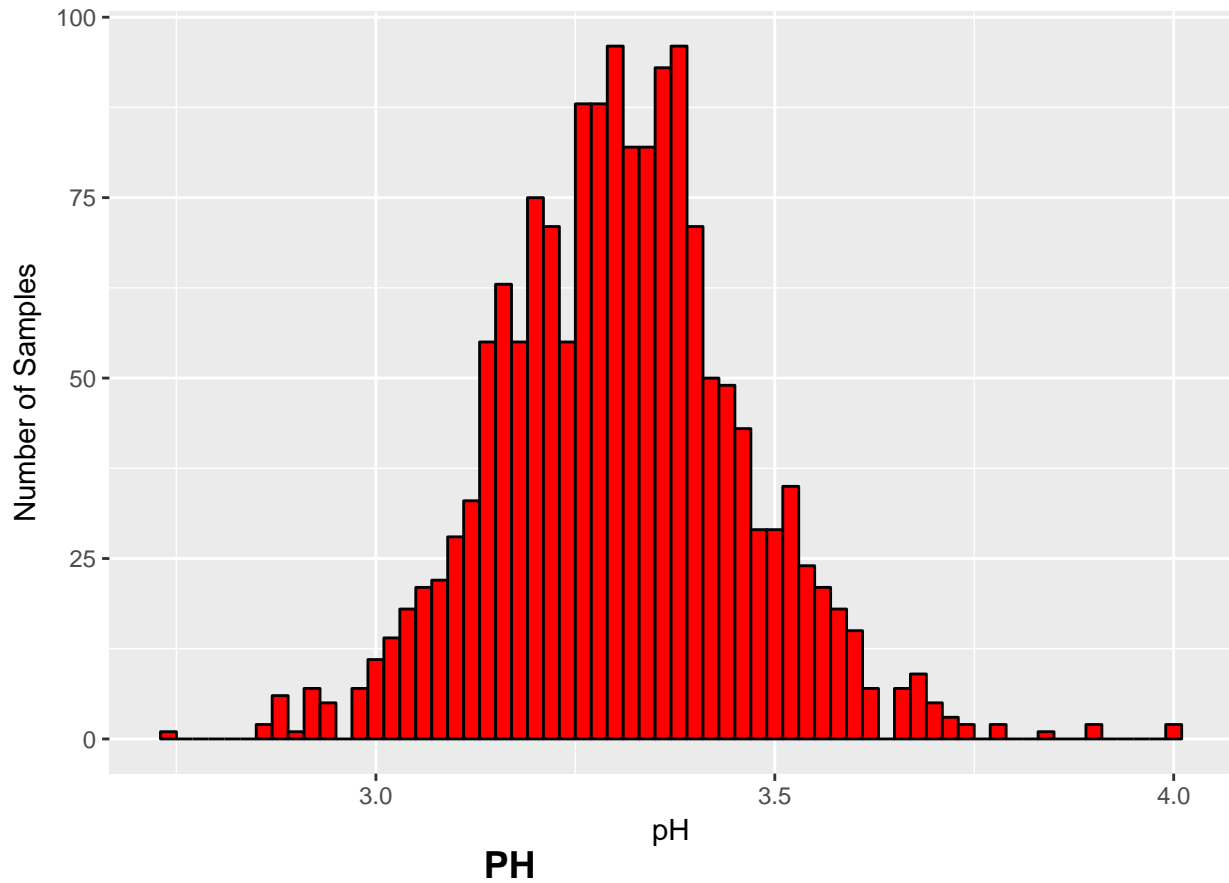
Total Sulfur Dioxide has many outliers on the right hand side of the data. And there are two points that are extremely away from the mean value.

## Density



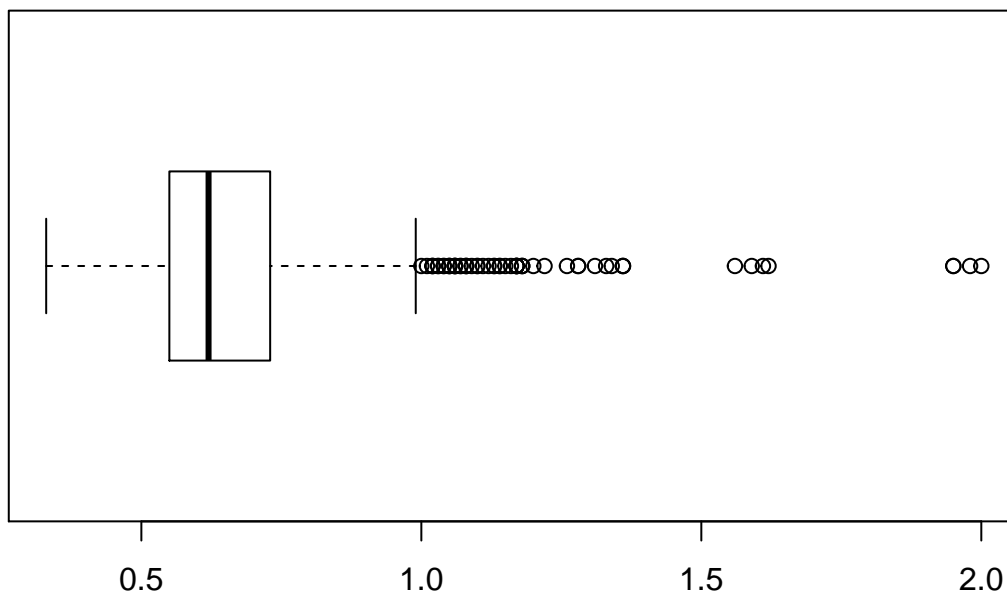
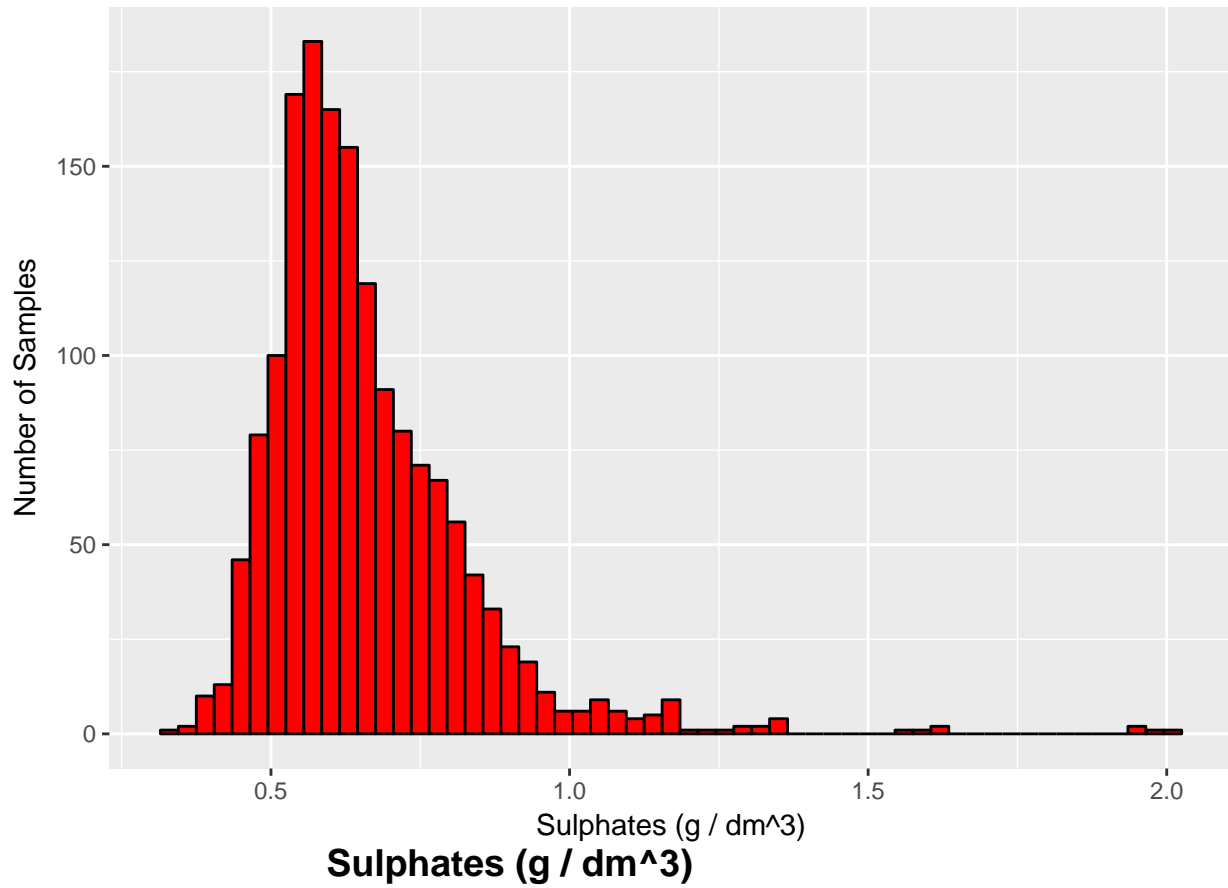
Density values look very uniformly distributed, even the distances around the points are close, so I don't think it has a big effect on the taste.

pH



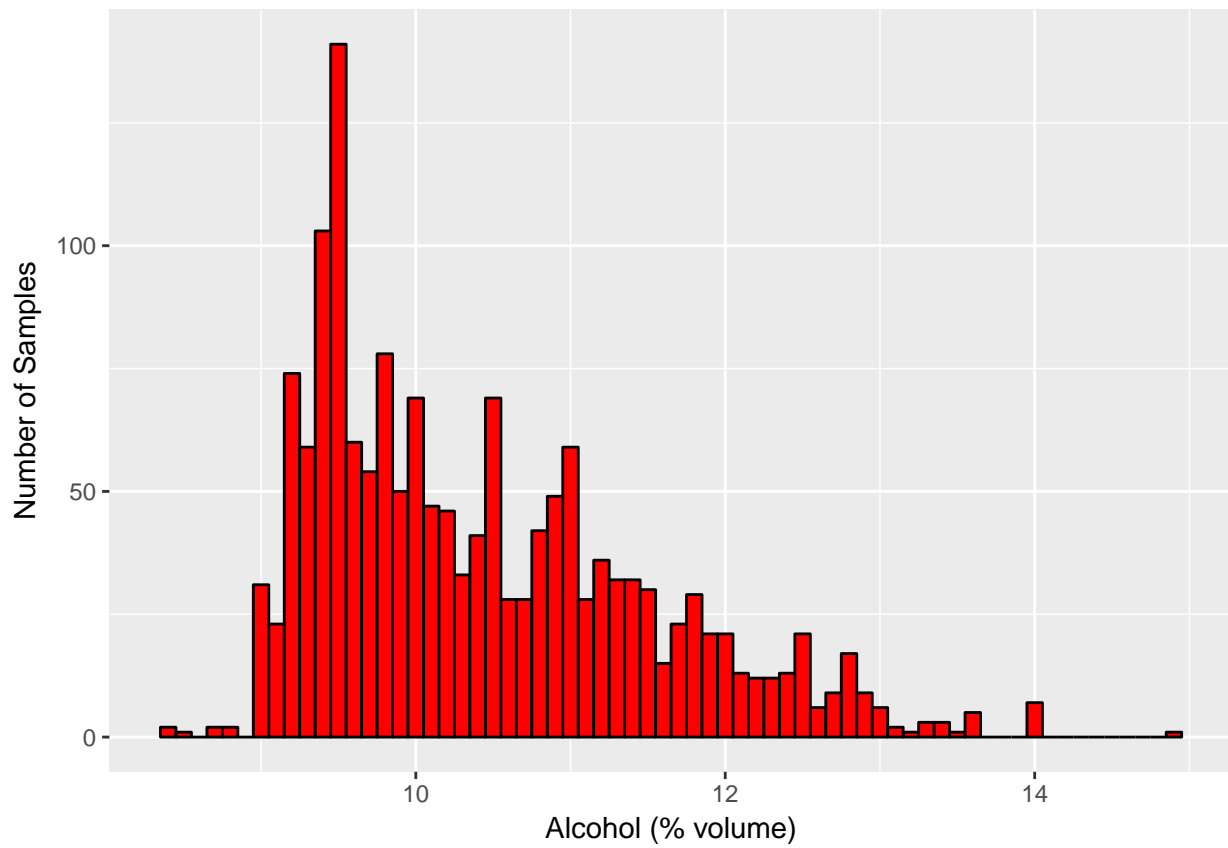
Observing Ph values we can see its mean is around 3.3 and it has few outliers. We can also observe since any pH lower than 7 is acidic, all the wines in this study is acidic. Since it is well spread towards both right and left, I wonder if this effects the taste.

## Sulphates



Many outliers on the right side, Sulphates are naturally occurring molecules, because of the fermentation. Too much sulphates might give a foul taste as well.

## Alcohol

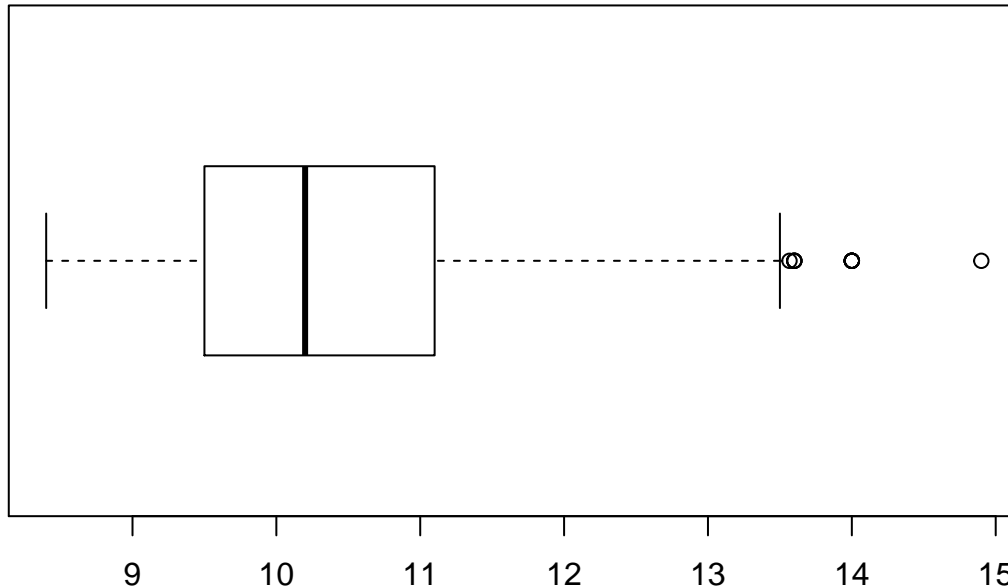


## [1] 8.4

## [1] 14.9

## [1] 10.42298

## Alcohol (% volume)



Observing alcohol level shows minimum alcohol level is 8.4 and maximum level is 14.9 with mean around 10.4.

## Univariate Analysis

For a successful data analysis, univariate analysis of different parameters is crucial. In this stage we can already notice there is a lot variance in certain variables whereas some are same for most of the wines such as densities. Having a univariate analysis at the beginning of a study, makes it easier to understand the context.

What is the structure of your dataset?

```
str(red)

## '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 ...
```



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

In this study I am more interested in how different factors effect taste of wine.

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

Many features of the data will be related. Analizing the features with respect to quality we find this relationship;

```
##                                [,1]
## fixed.acidity                 0.12405165
## volatile.acidity             -0.39055778
## citric.acid                   0.22637251
## residual.sugar                0.01373164
## chlorides                    -0.12890656
## free.sulfur.dioxide          -0.05065606
## total.sulfur.dioxide         -0.18510029
## density                      -0.17491923
## pH                           -0.05773139
## sulphates                     0.25139708
## alcohol                       0.47616632
```

So Fixed Acidity, Citric Acid, Residual Sugar, Sulphates and Alcohol is positively corelated. Alcohol has a strong corelation. Whereas Volatile Acidity, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density and pH is negatively corelated. Volatile Acidity has strong negative corelation.

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

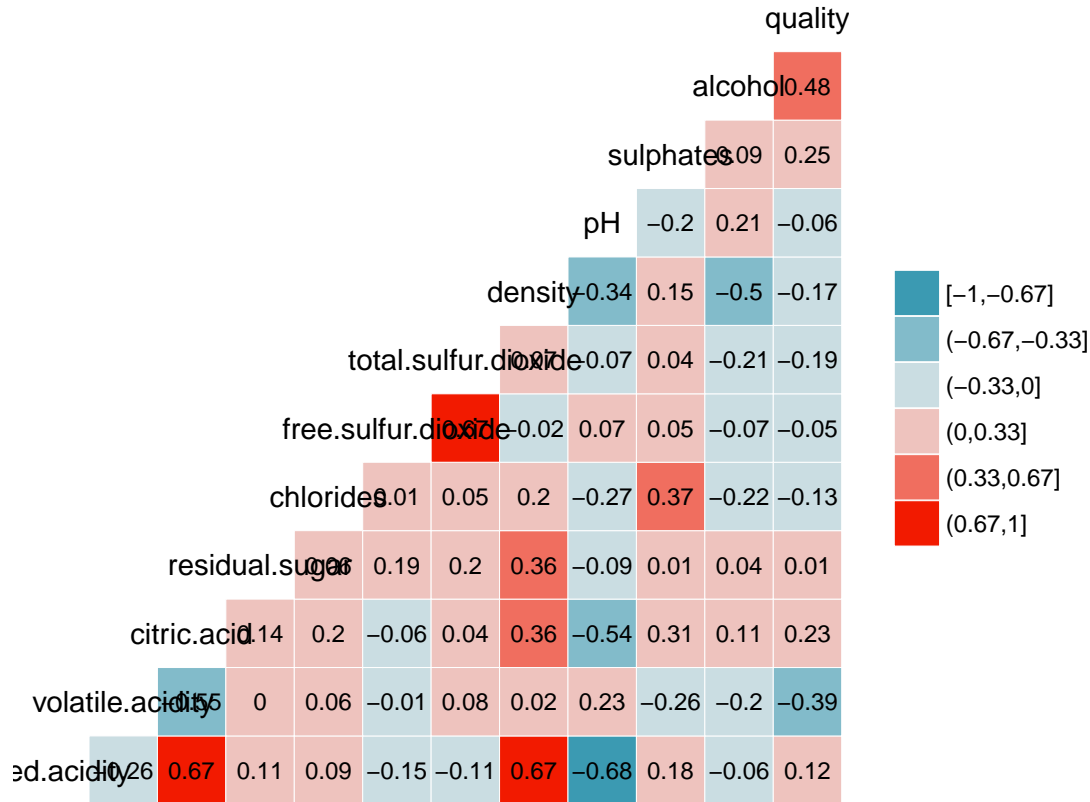
No, no new variable is created.

**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?**

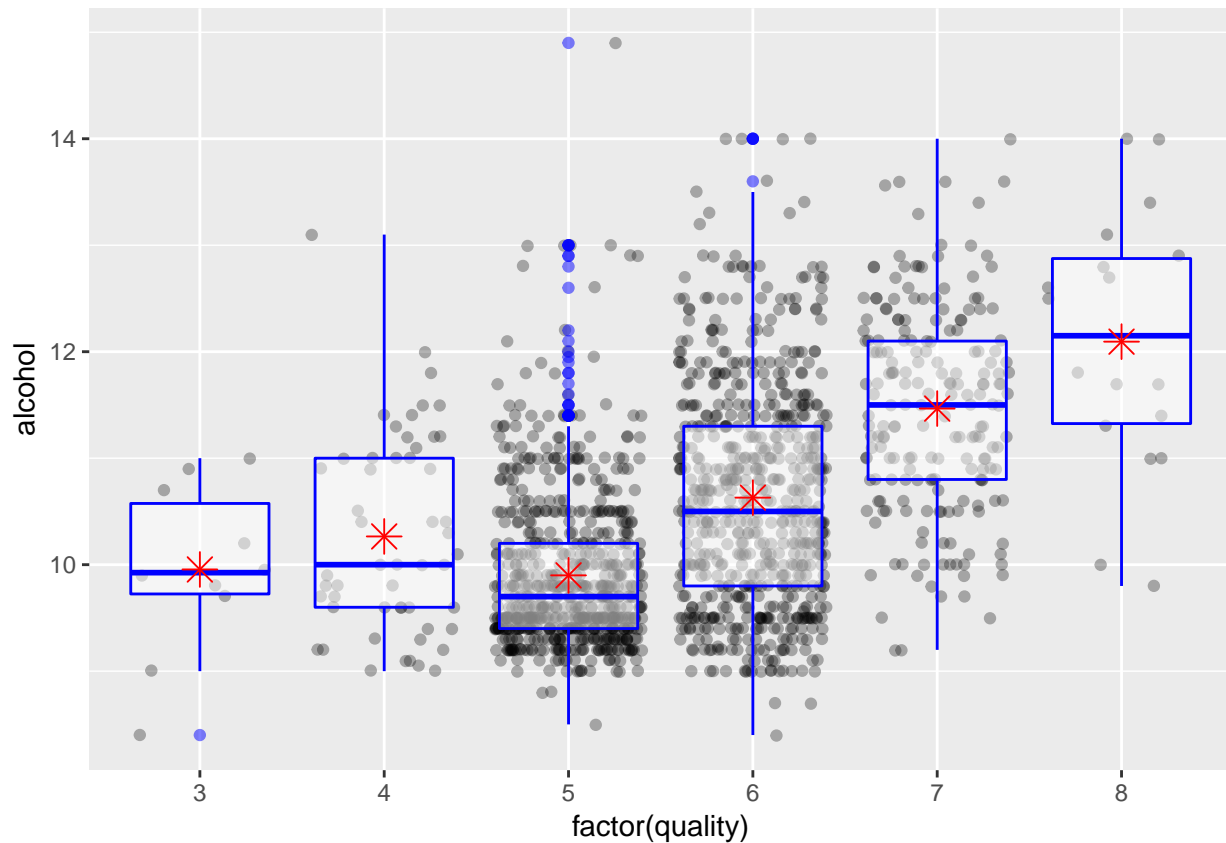
As explained in previous section, Residual sugar and chlorides have outliers, those are plotted again after removing the top and bottom extremes. Citric acid has many zero values.

## Bivariate Plots Section



First we check all the variables for correlation. We see which variables are strongly correlated with each other. This function is found from this website <https://briatte.github.io/ggcorr/>

```
ggplot(aes(factor(quality),
            alcohol),
      data = red) +
  geom_jitter( alpha = .3) +
  geom_boxplot( alpha = .5,color = 'blue')+
  stat_summary(fun.y = "mean",
              geom = "point",
              color = "red",
              shape = 8,
              size = 4)
```

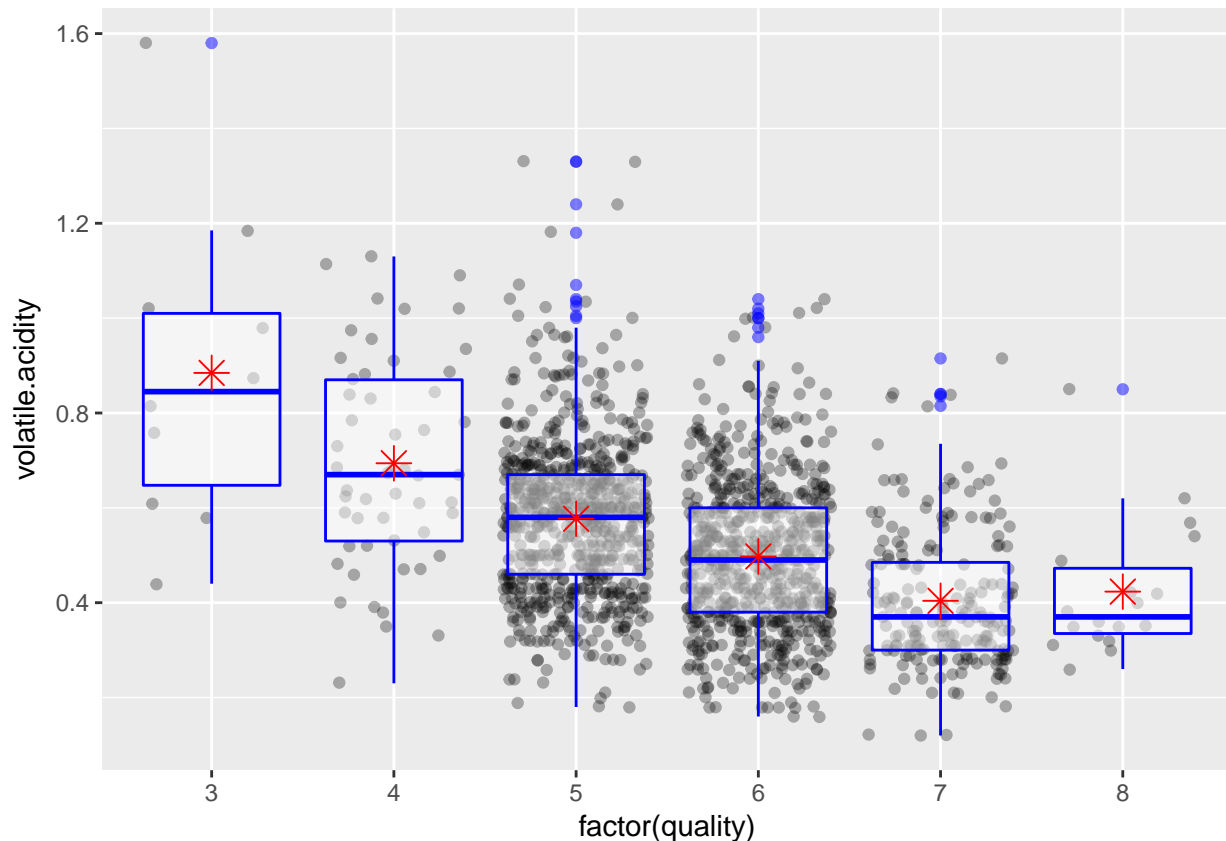


I want to first investigate relation of alcohol content and quality of the wine. I suspect there should be strong correlation.

Checking this graph we see good wines (quality>5) have more alcohol content. Whereas wines that are not so good (quality<6) have varying alcohol content.

We can notice positive correlation between alcohol content and quality.

```
ggplot(aes(factor(quality),
             volatile.acidity),
       data = red) +
  geom_jitter( alpha = .3) +
  geom_boxplot( alpha = .5,color = 'blue')+
  stat_summary(fun.y = "mean",
               geom = "point",
               color = "red",
               shape = 8,
               size = 4)
```



This shows the relationship between volatile acidity and quality. More volatile acidity causes the wine to taste like vinegar. We can see an opposite correlation between volatile acidity and quality.

## 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?**

As we have discussed earlier there is a strong correlation between perceived quality of the wine and alcohol content and volatile acidity. Unexpected previously there is no correlation between the sugar content and free sulfur level and quality of the wine, I was expecting a relationship between these values.

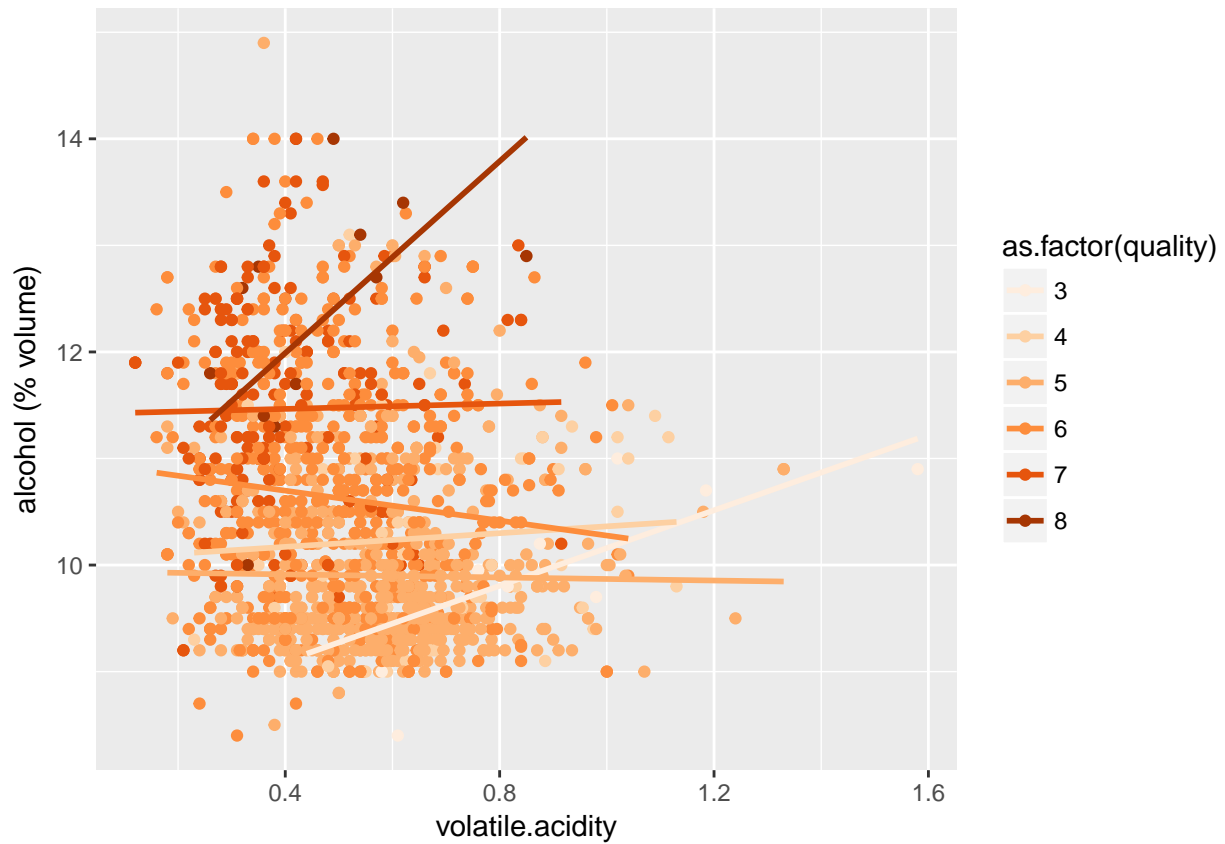
**Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?**

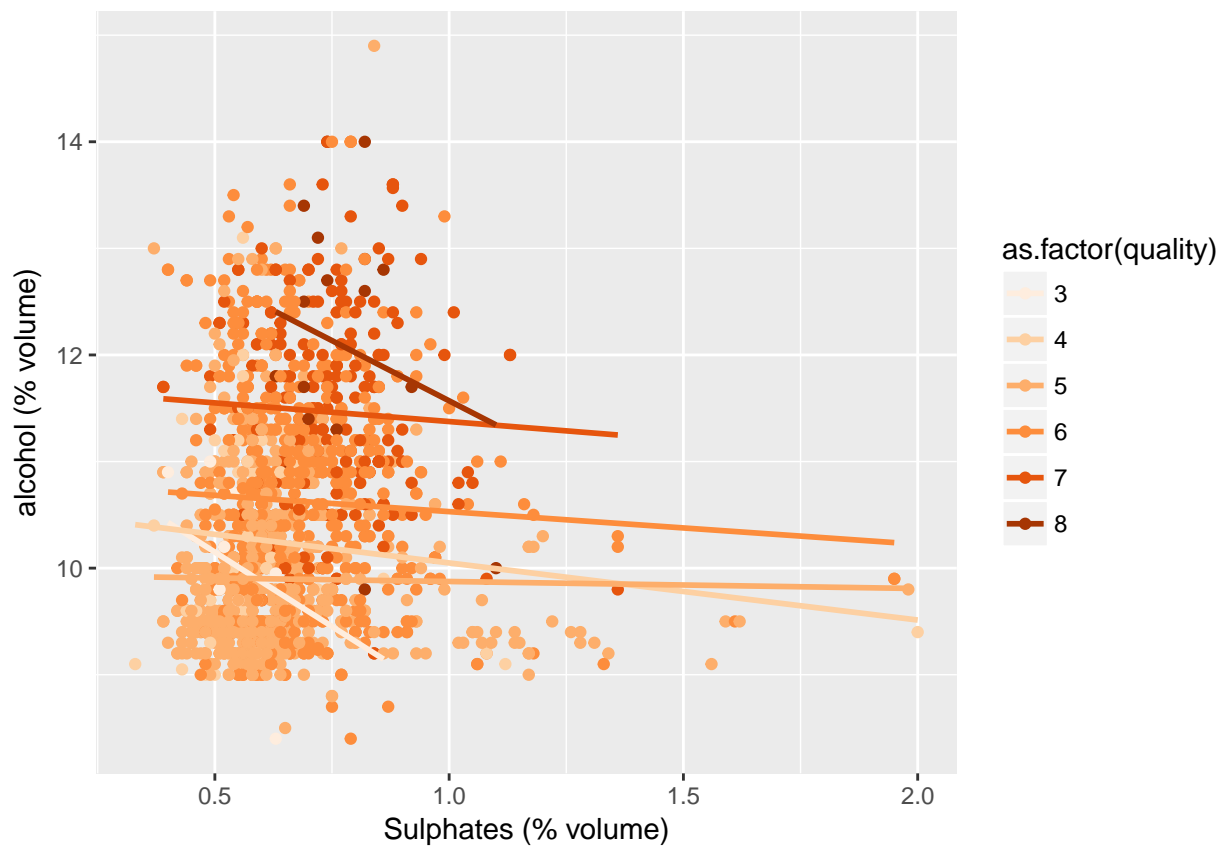
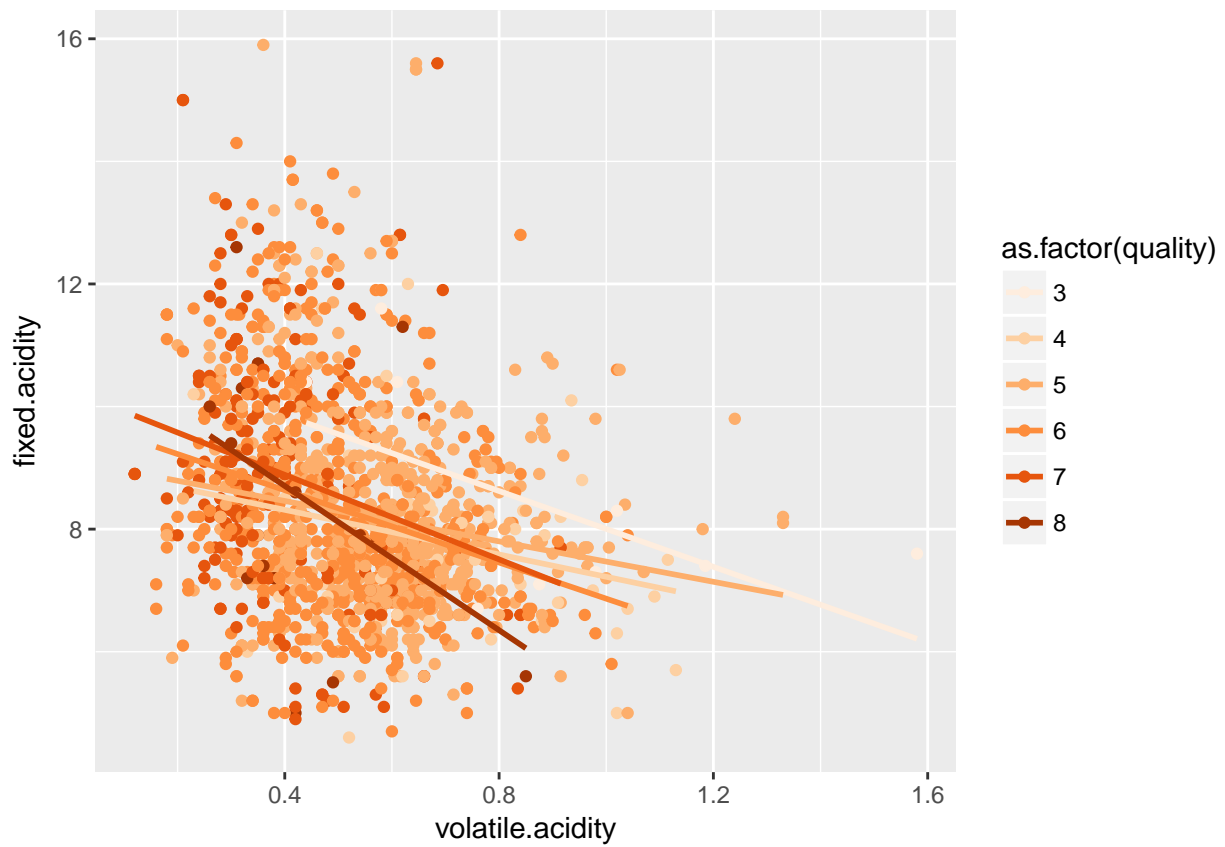
Here I noted a relationship between volatile acidity, fixed acidity and citric acid level. Citric acid level is positively correlated with fixed acidity and negatively correlated with volatile acidity. So my observation is that wine producers add citric acid to modify their acidity and lowering volatile acidity. But some wine makers cannot do that (as discussed I found out that EU producers cannot use citric acid, it would be good to check this.) Plotting some graphs shows interesting results for wine that doesnot use any citric acid, those are slightly worse in ratings as they contain more volatile acids.

What was the strongest relationship you found?

Strongest relationships are the obvious ones such as pH and acid levels, total sulfur vs free sulfur etc. But there are also some strong relationship that are not that obvious such as alcohol level and volatile acidity levels and quality of wine.

## Multivariate Plots Section





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?**

These plots strengthened mainly the two topics I have been discussing, quality of wines and effect of alcohol and volatile acidity. And means of acidity (ie. volatile, fixed, citric) and quality. I conclude about the importance of the volatile acidity on the quality. And lastly I have plotted the sulphates and alcohol in a graph to see their effect on the quality.

**Were there any interesting or surprising interactions between features?**

The interesting conclusion from this study is that, wine is a very complex beverage and there is no easy way to describe what makes a good wine. Playing with the parameters we noted certain correlations but apart from alcohol level and volatile acidity, the affect of others are not so clear.

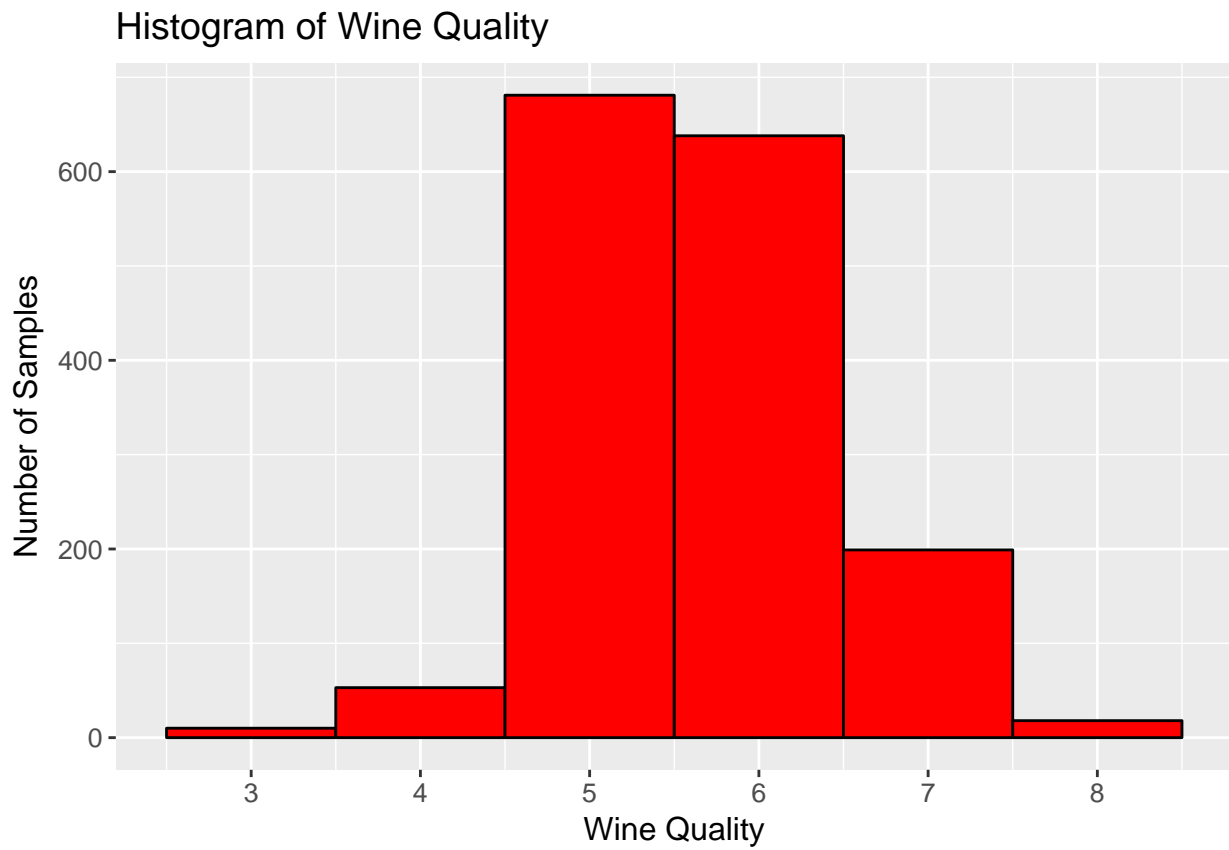
**OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.**

No I havent created a model.

---

## Final Plots and Summary

Plot One

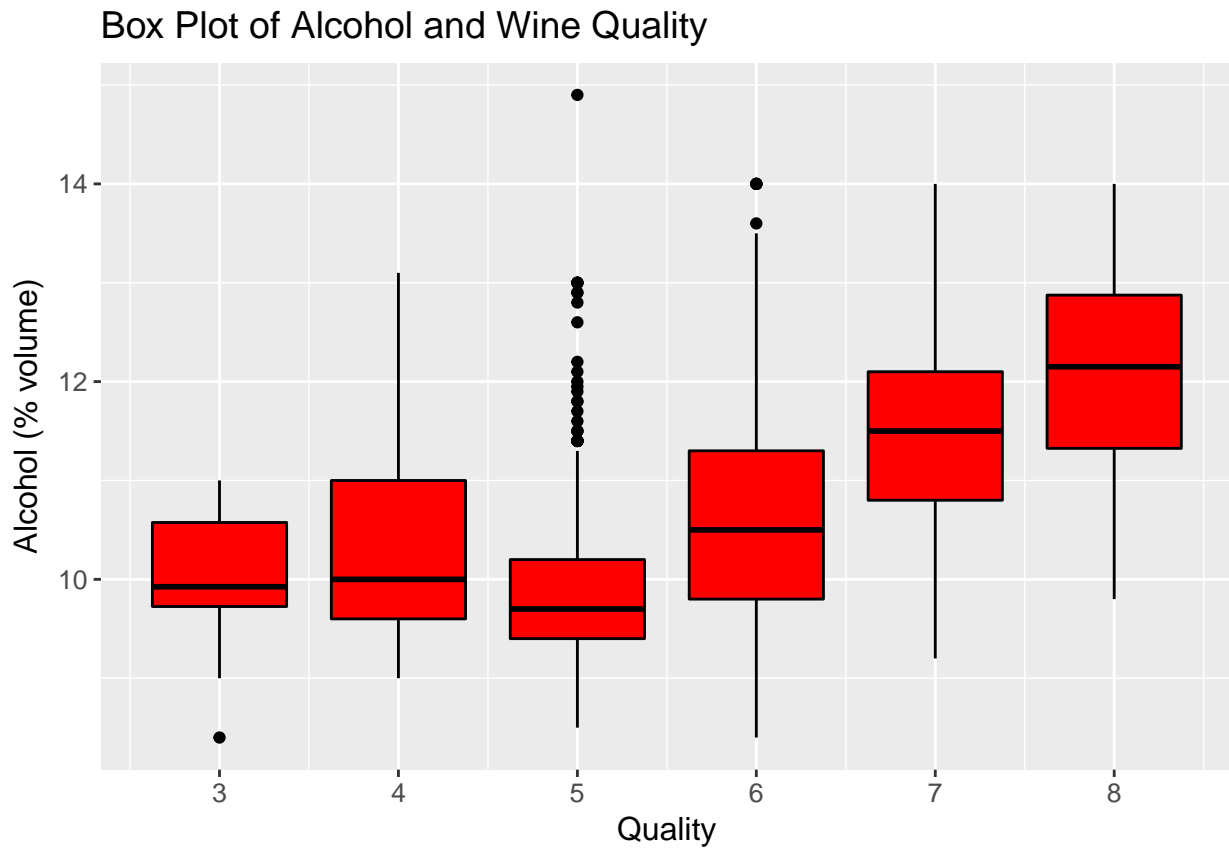


### Description One

This was the first plot I used in my study, as this is the main aspect that we are studying I put this in the concluding chapter. One can note from this graph majority of the wines are labeled as 5-6 rating whereas low quality or high quality wines are lesser in numbers.



Plot Two

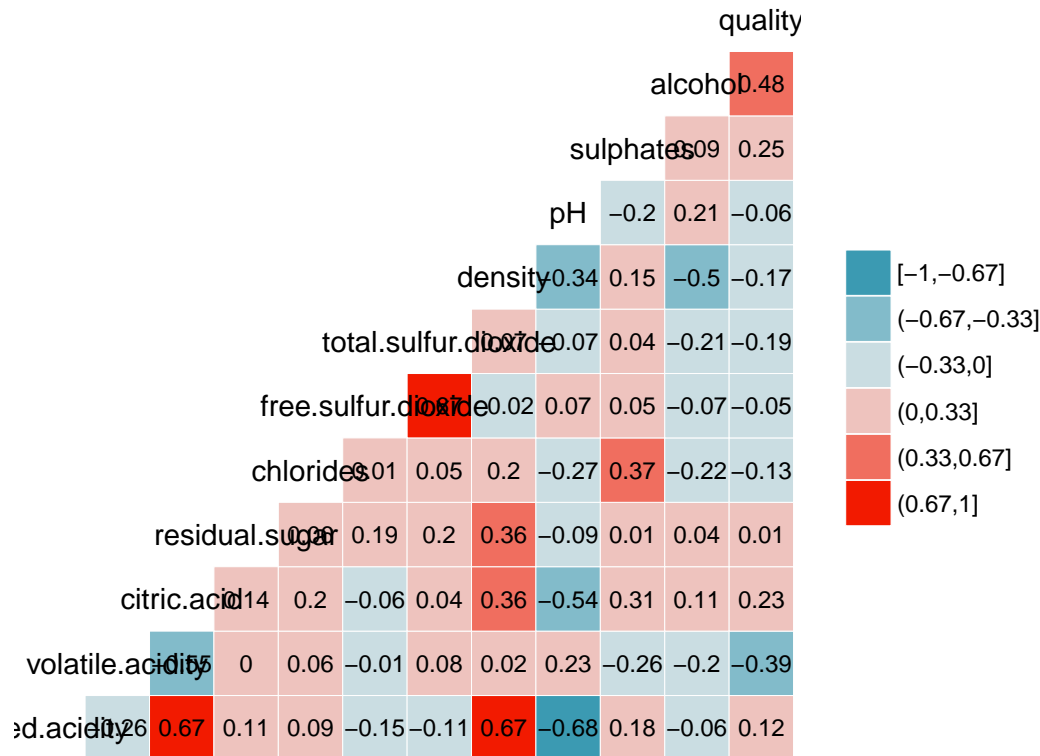


### Description Two

Alcohol is strongly correlated with wine quality especially on the better wines (Quality 7-8). This graph is showing this relationship. What I note is the interesting case of wines with rating 5. They are the ones with the lowest mean.

### Plot Three

#### Correlation of parameters



### Description Three

Here I put this plot as this summarises all the correlation of different parameters in one plot. There are some interesting things to note. At the start of the study I was certainly expecting some correlation between quality and residual sugar. But this plot and my previous examinations dont show any correlation. Also here it is easy to see the relationship between all the acidity parameters.

## Reflection

This was a very interesting study and it gave me a good chance to practice all the functions I learned from R environment. Also I had a chance to learn about red wines which will probably make me appreciate more of the taste of red wine next time. There were some obvious outcomes and not so obvious ones for me as well. The level of alcohol and quality of the wine was surprising for me. I was even expecting the opposite as I think the more alcohol in the wine it would me more “bitter”.

The relationship of volatile acidity and quality of wine is less suprising. As generally people dont like the vinegar taste of the volatile acids.

This study can be extended with the addition of location of the wine, year it has passed and it is fermentation process as the other parameters would probably be more useful. Such as less sulphates can make wines more susceptible to become faulty.

Some hardships I had during this study is about the legends and axis, I sometimes find it difficult to arrange the axis such that the information is readable. But in general I find this study, quite successful, as I was able to identify at least some parameters effecting wine taste, and maybe next time I am enjoying my wine I will appreciate more.

## **Resources**

[https://en.wikipedia.org/wiki/Acids\\_in\\_wine](https://en.wikipedia.org/wiki/Acids_in_wine)

[https://en.wikipedia.org/wiki/Wine\\_fault#Acetic\\_acid](https://en.wikipedia.org/wiki/Wine_fault#Acetic_acid)

<http://www.minitab.com/en-us/Published-Articles/Wine-Tasting-by-Numbers--Using-Binary-Logistic-Regression-to-Reveal-t>

<https://briatte.github.io/ggcorr/>