TITLE: Investigation of the chemical composition of red wine.

RED WINE EXPLORATION

Armindo Cuamba

==

Red wine quality

The purpose of this work is to use the data analysis technique to investigate the quality of the manufactured products. As an example, I will explore the data set of the 'red wine' to identify the parameters that determine the wine quality. The data used in this works is composed of a different dataset of the chemical properties that make up the wine.The technique used in this work can be used to investigate the quality of any type of product.

Univariate Plots Section

In order to get the statistics of each factor in the data set I will get the summary

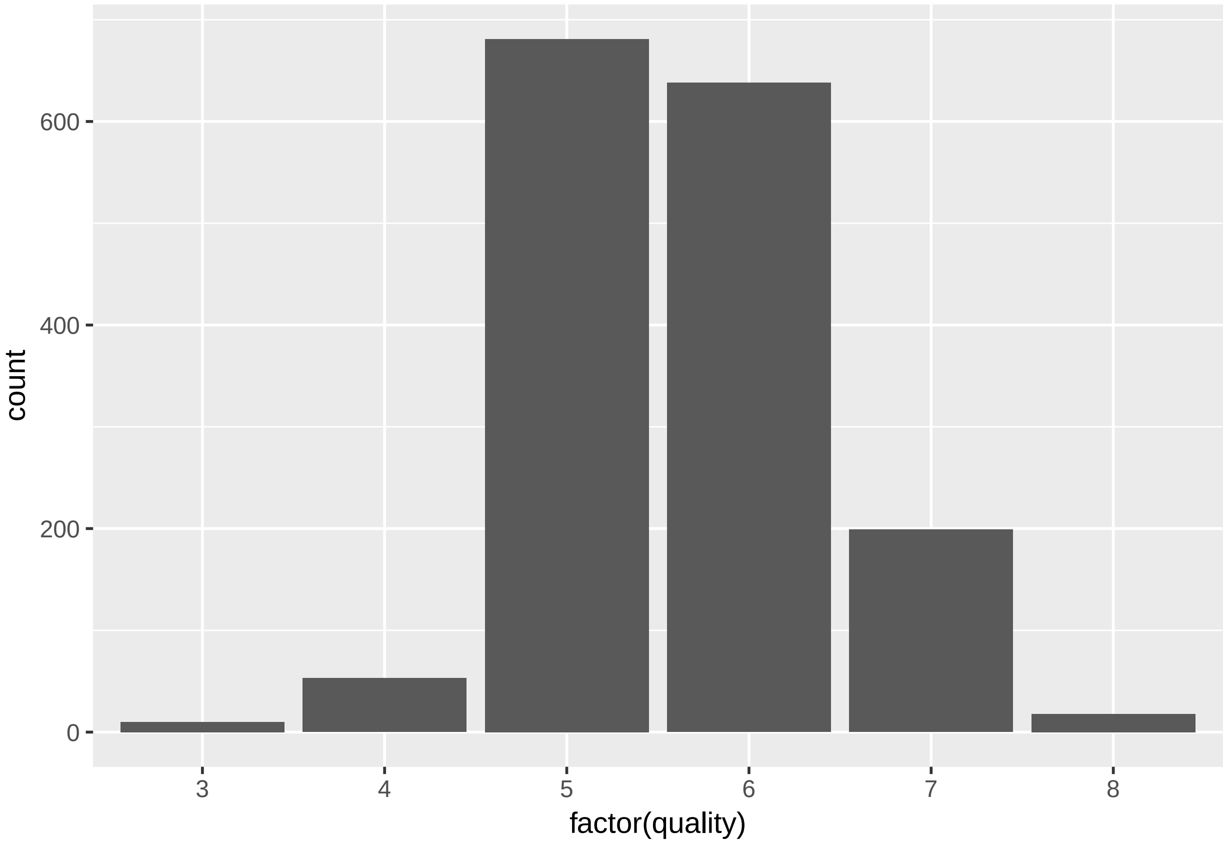
```
##        X          fixed.acidity   volatile.acidity  citric.acid   
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000  
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090  
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260  
##  Mean   : 800.0   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
```

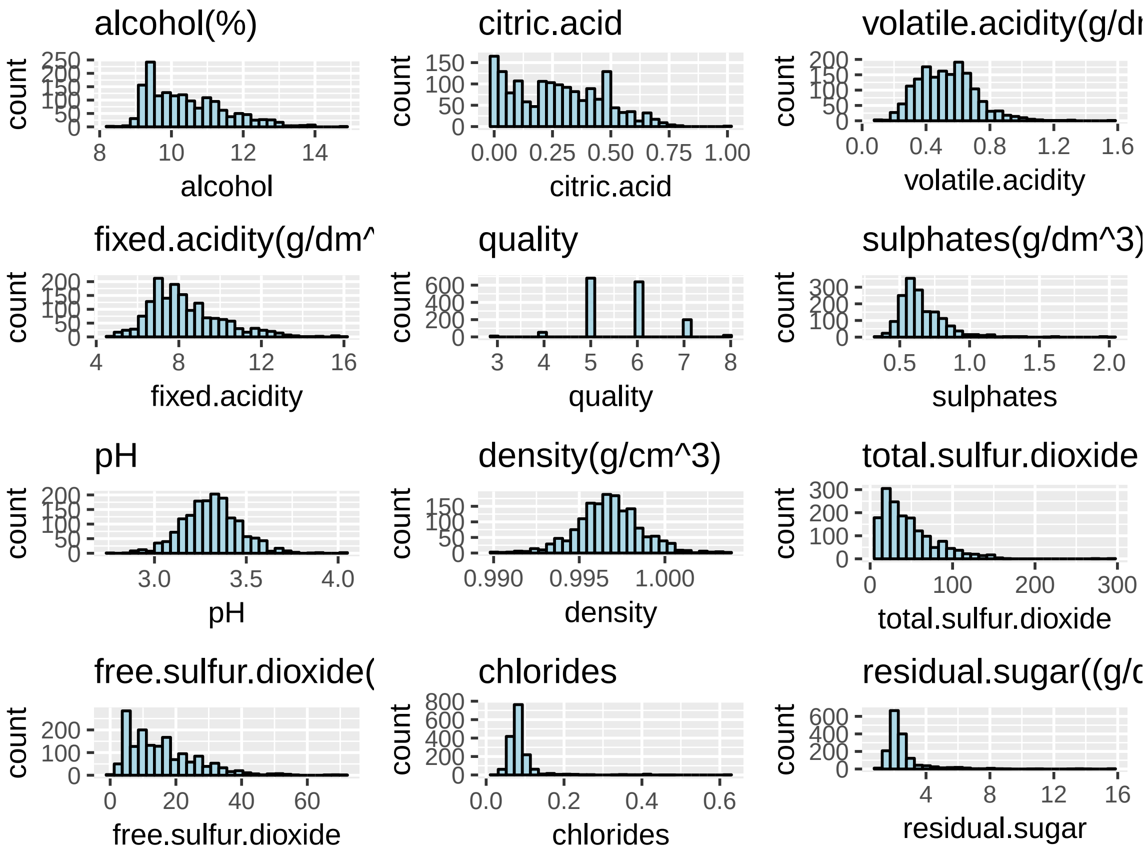
The red-wine data is composed on 1599 observables which and divided by 13 numeric observables.

```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"    
##  [4] "citric.acid"          "residual.sugar"       "chlorides"           
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"             
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

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

The quality is the elements that I am investigating and its summary is given by the table above. It is a discrete quantity ranging from 3 to 8. The data set of the red-wine has the median 6 and the mean is 5.636. The histogram below shows the distribution of the quality, where quality=5 is the highest peak.





Normal distribution is defined as the is a symmetric continuous distribution

Skewness distribution is asymmetric distribution where one tail dominates the distribution. It is a positive skew when the right tail is long and is called negative skew when the left tail is long.

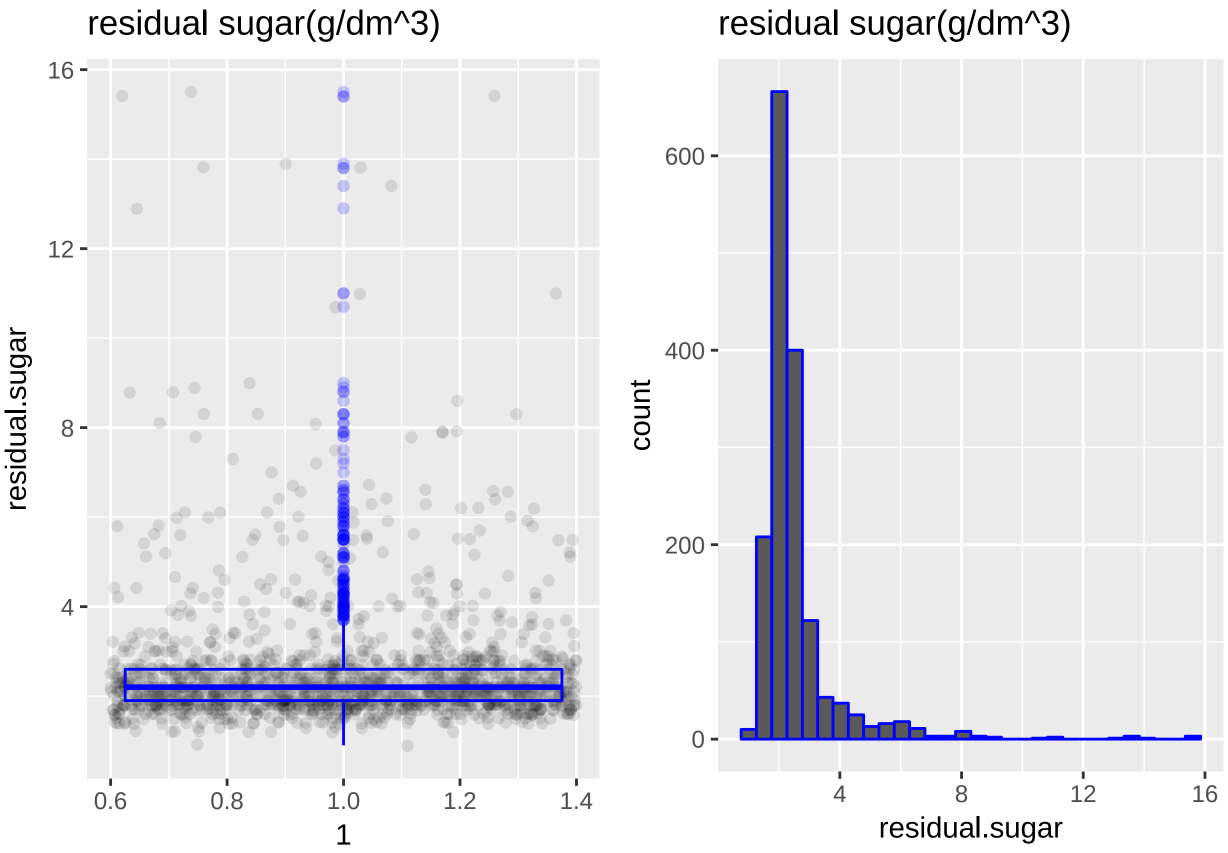
Long tail is the distribution with a mixture of high frequency at the origin and low frequency at the infinity.

The graphic above shows a lot of interesting features. All the histogram indicates that the chemical parameters are continuous except the quality which is a discrete parameter.

The histogram of the density, and pH show a normal distribution,

Residual sugar

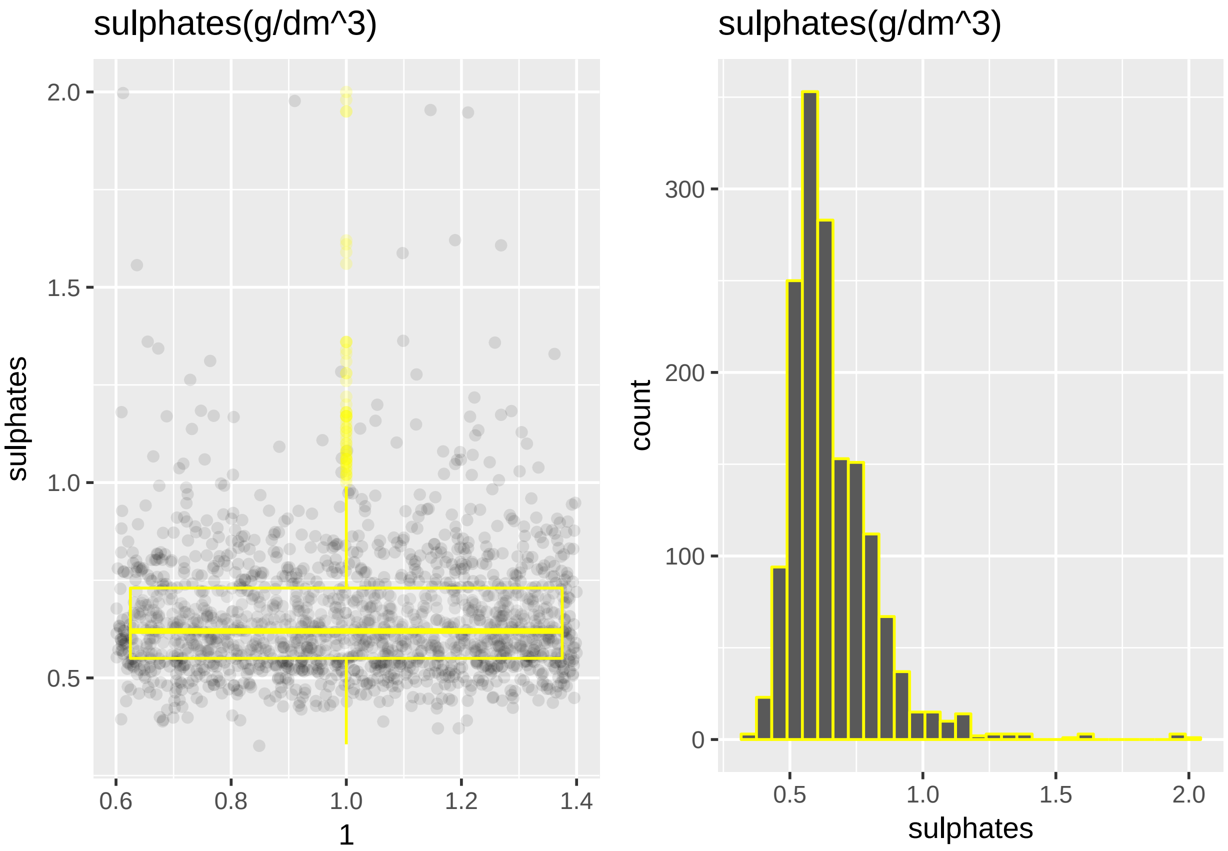
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```



The residual sugar has the long tail distribution. The mean value of the residual sugar is 2.539 with 2.2 median. This tells us that the majority of wines does not have a huge quantity of residual sugar. Which implies that the residual sugar may not determine the quality of the wine.

Sulphates

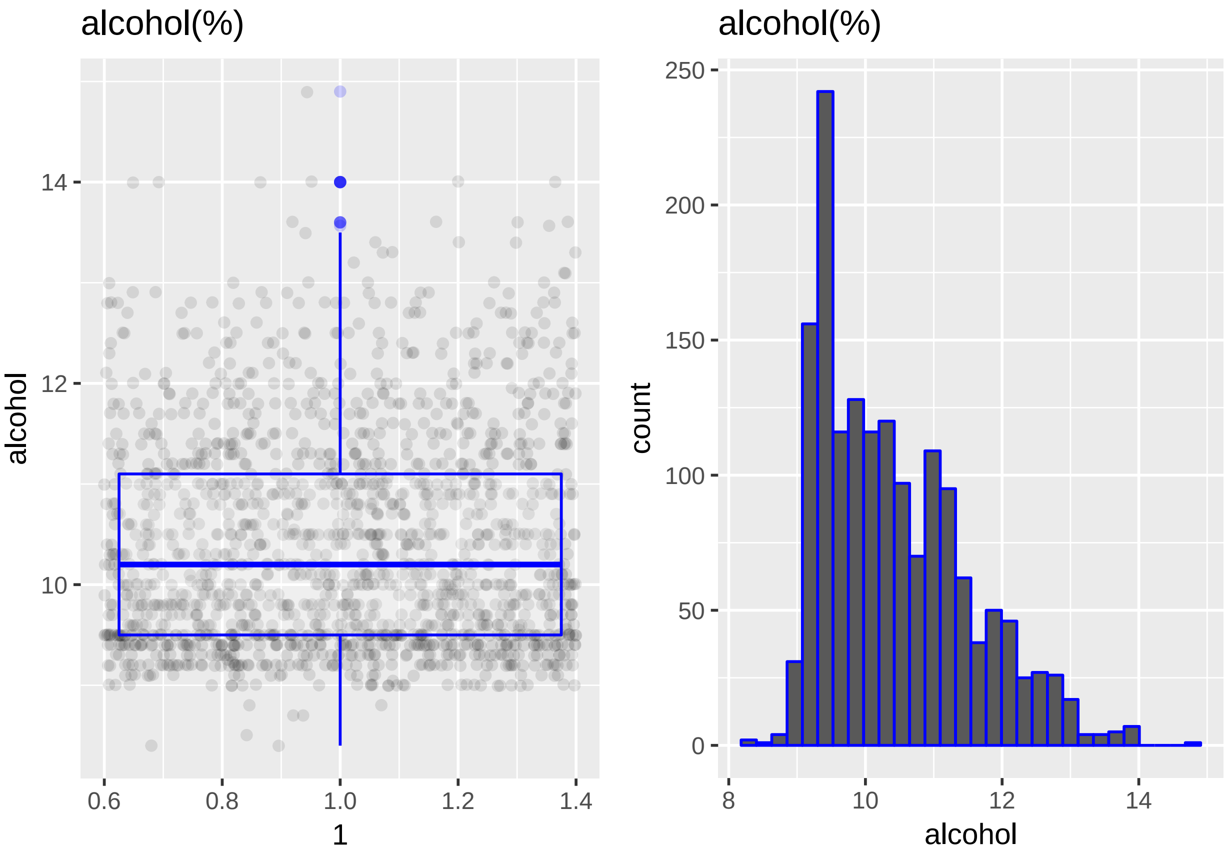
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```



The sulphates distribution is asymmetric. The maximum of the distribution is not at the center. The mean value of the sulphates is around 0.6581 and the median is 0.6200. From this graphics is difficult to determine the contribution of the sulphates on the quality of the wine.

Alcohol

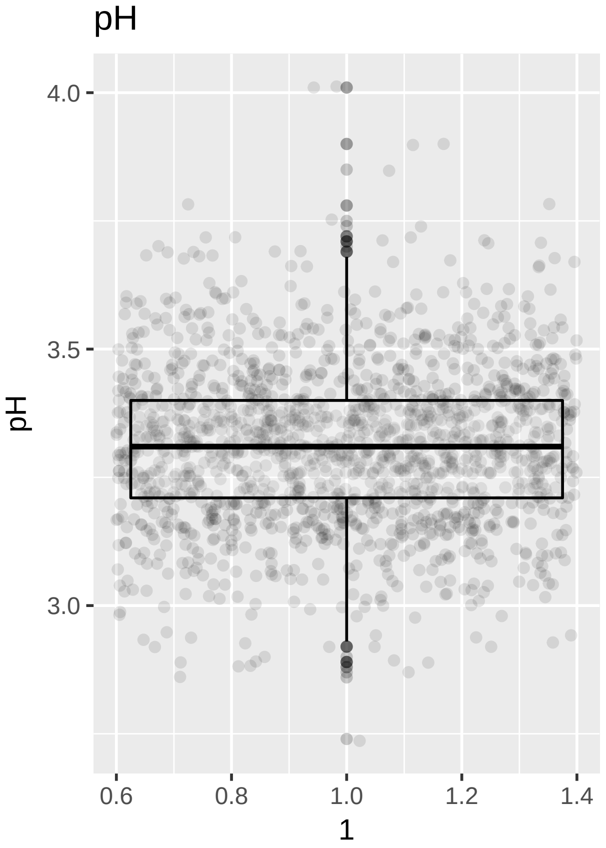
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

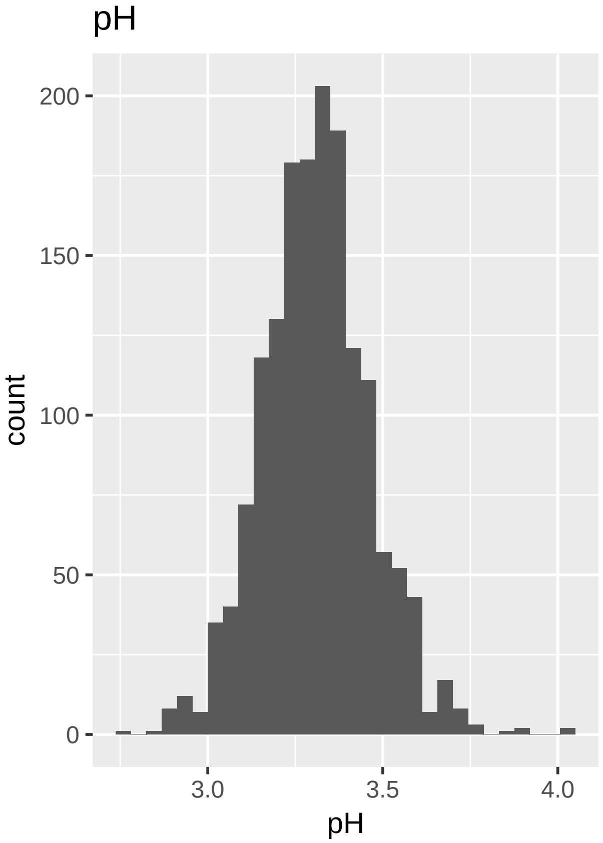


The distribution of the alcohol content is a long tail. The alcohol concentration has the highest peak around 9.5 %. #### The alcohol distribution show some interesting features. The percentage of the alcohol decreases with the increases of the count of the varies types of wines. This implies that the majority of wines count have less alcohol concentration, and few wine counts have more alcohol. We could suspect that these few wine count with more alcohol conent have high quality.

pH

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

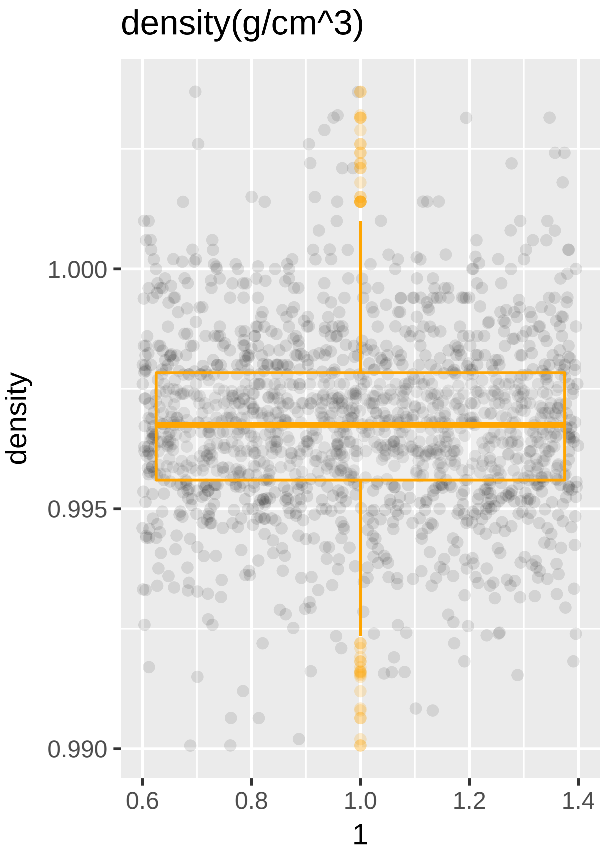


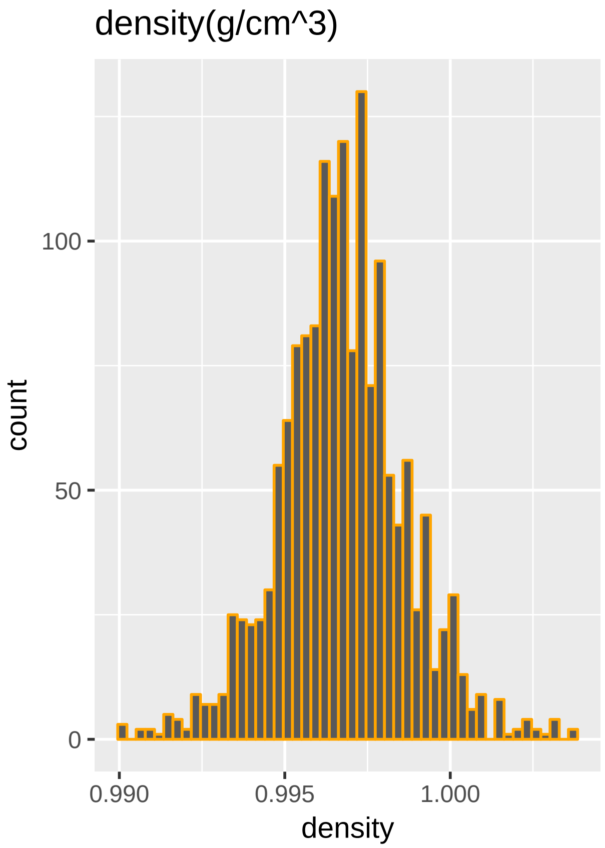


The normal distribution of pH makes the analysis more difficult to tell how it will affect the quality of the wine. It is centered at the median. The median value is 3.310 and the mean is 3.311 respectively

Density

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

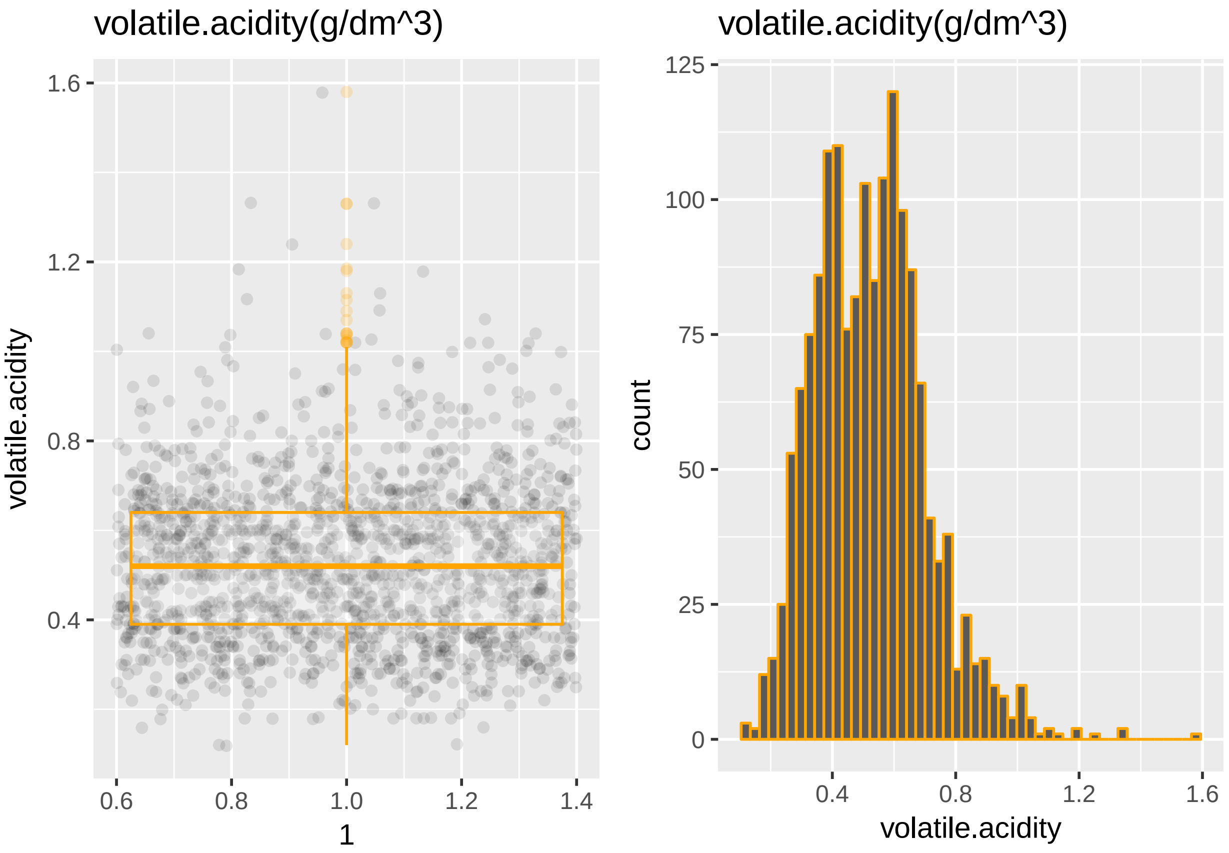




Density has the normal distribution centered at the median. The value of the mean and median and almost equal. The median is 0.9968 and the mean is 0.9967. Is not easier to tell how the density can affect the quality.

Volatile acidity

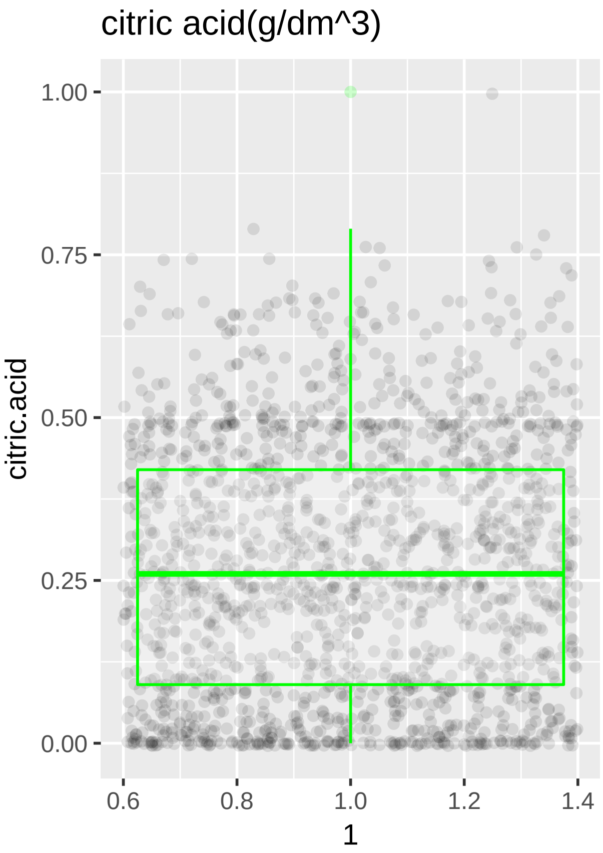
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

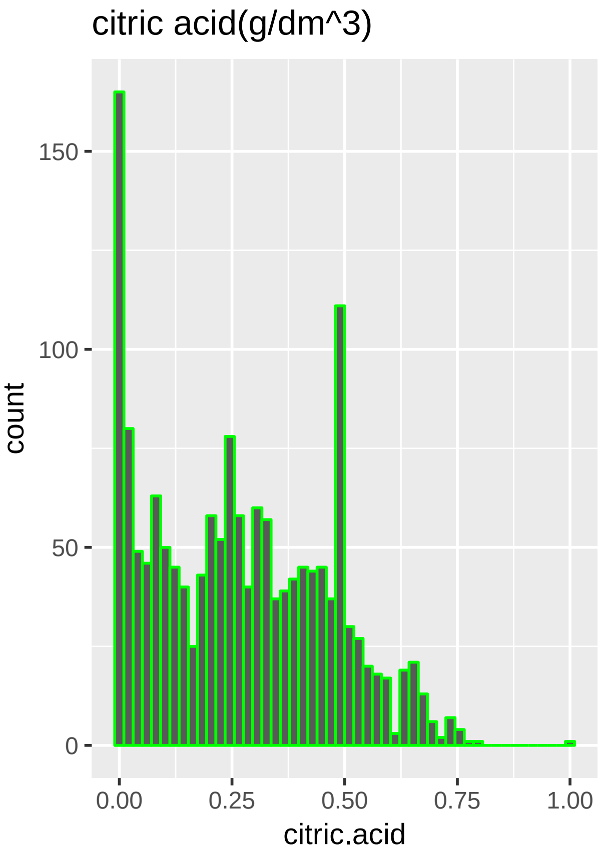


There are few counts with more volatile acidity. The distribution has two maximums.

Citric acid

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

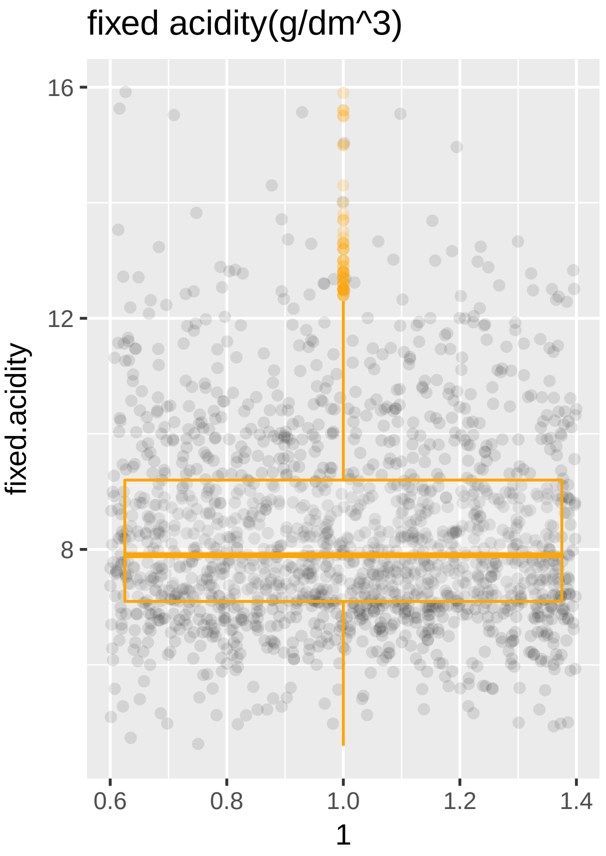


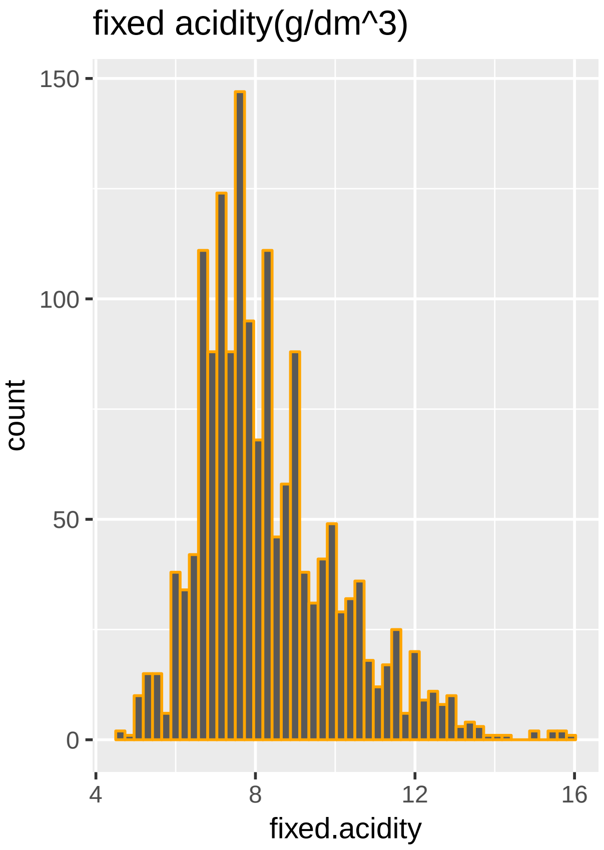


The citric acid distribution indicates that the citric acid is important for wine production. The concentration is well distributed in the majority of the wines. In addition, the corresponding median is 0.260 and which is less than the mean value of 0.271. This may indicate that the citric acid has major contribution on the wine quality.

Fixed acidity

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

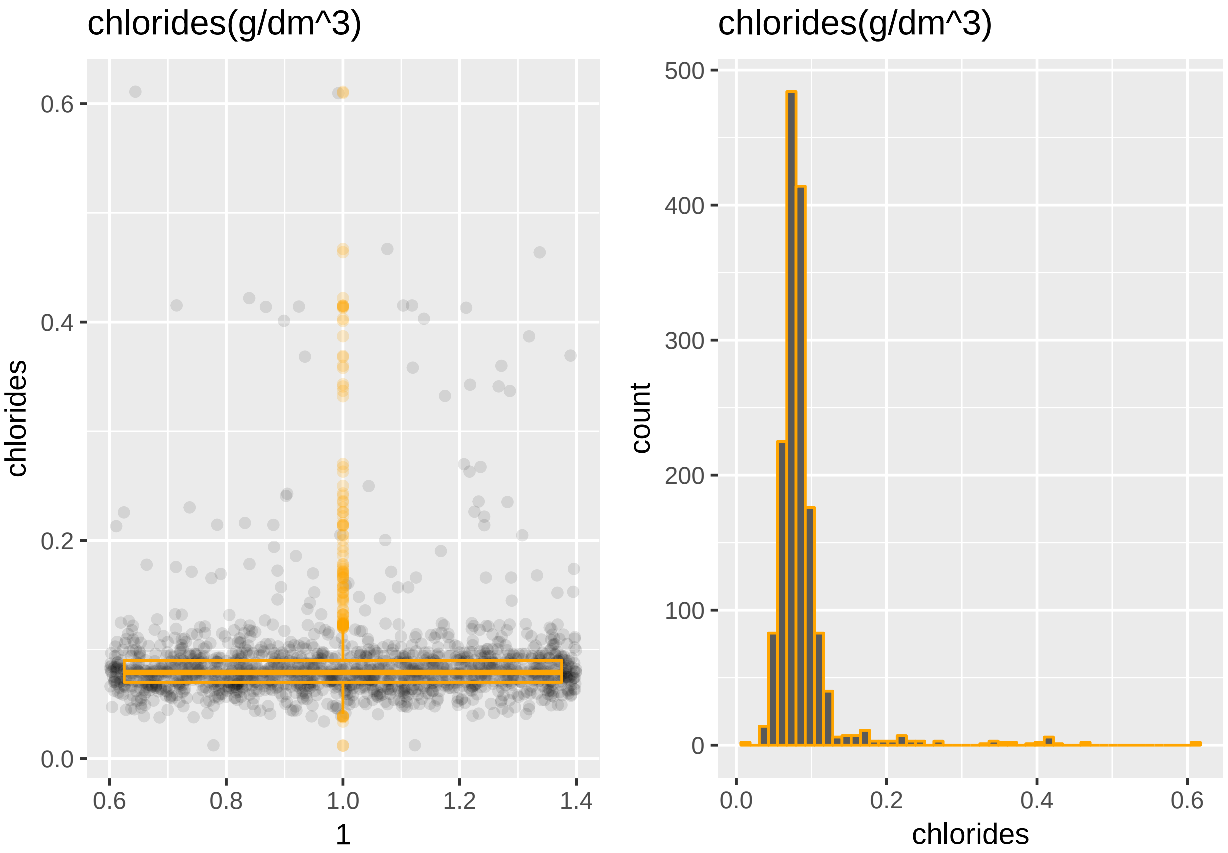




The median value of fixed acidity is 7.90 and the corresponding mean is 8.32. These values a little close so more investigation is requered to determine the contribution of the fixed acidity to the quality of the wine.

Chlorides

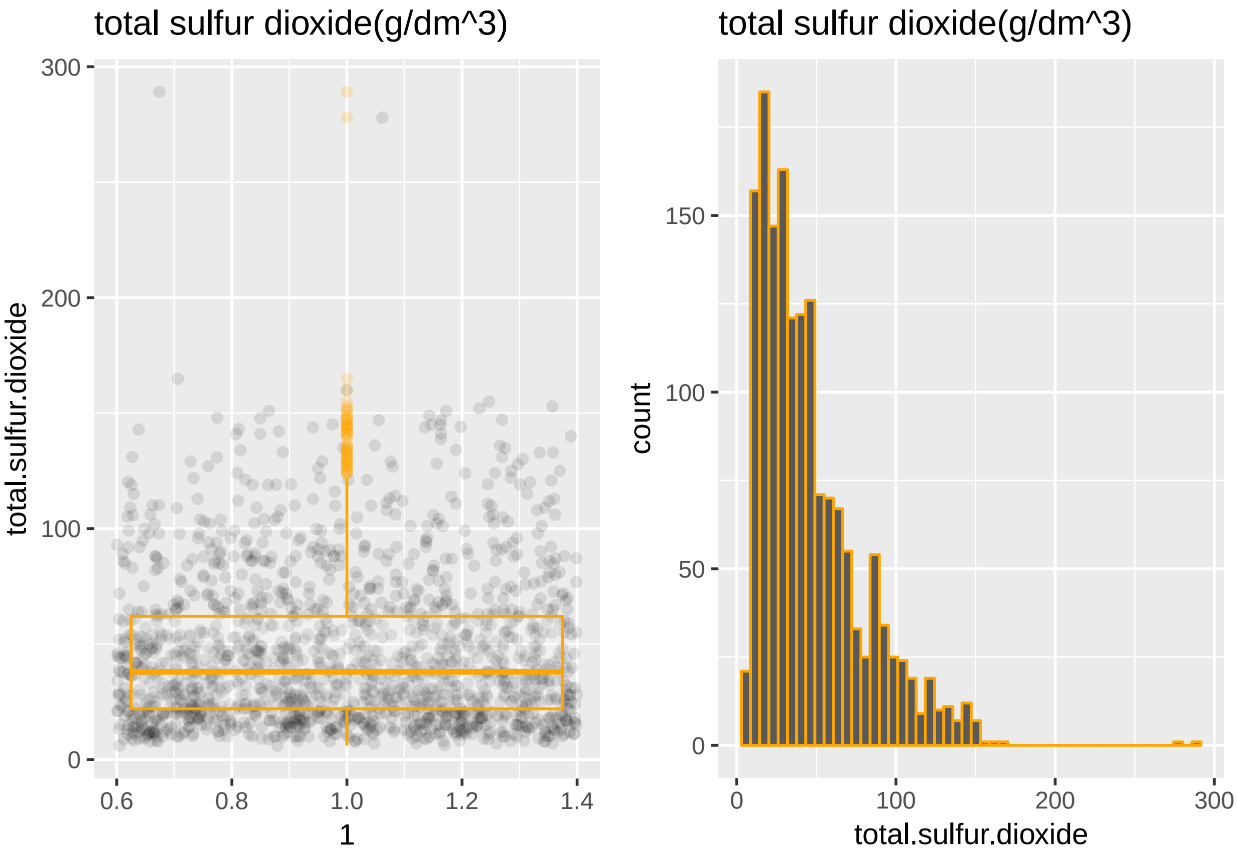
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```



This distribution shows that the chlorides do not have a considerable contribution to the wine quality. The concentration used is very few. The mean value is around 0.07900 and the median is 0.08747.

Total sulfur dioxide

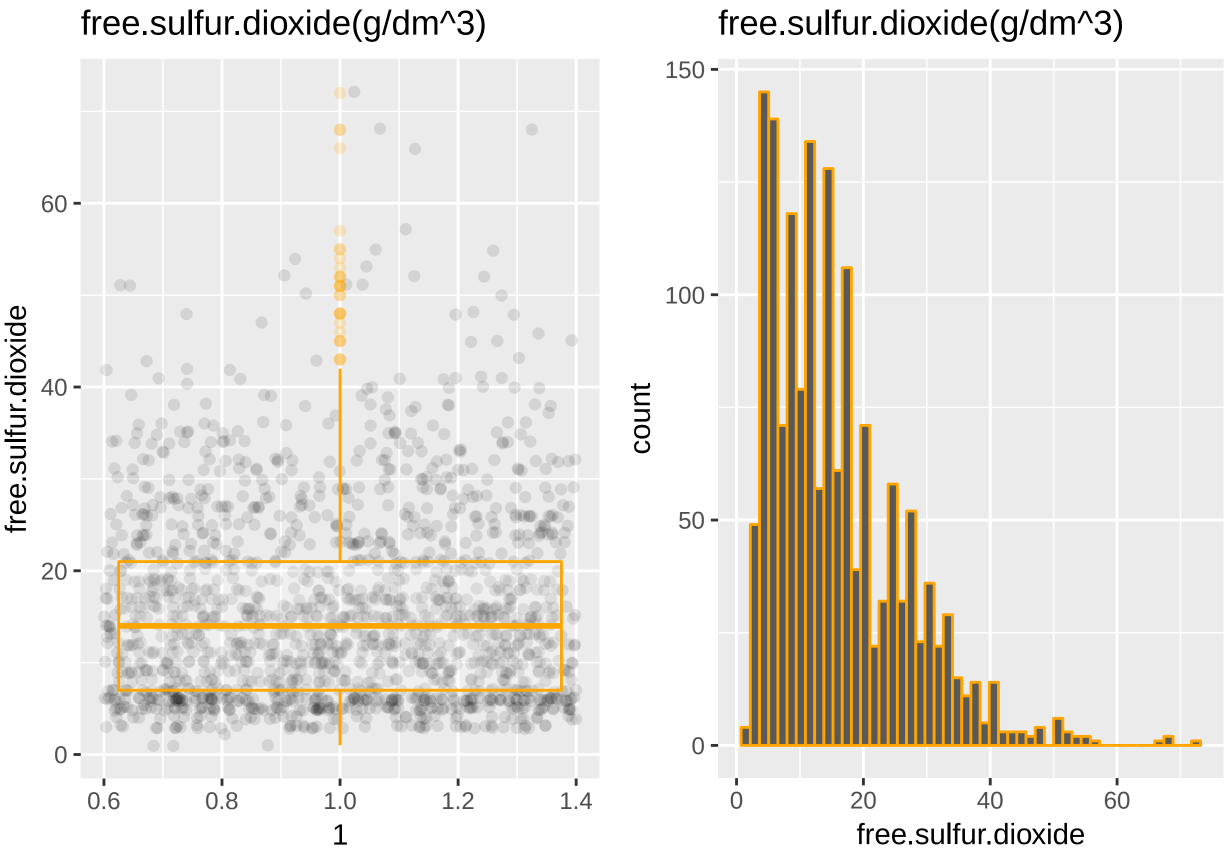
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.00   22.00   38.00   46.47   62.00  289.00 
```



Asymmetry distribution of the total sulfur dioxide. The maximum of the distribution is located at the low concentration of the total sufur dioxide.
The median is 38.00 and the mean value is 46.47.

Free sulfur dioxide

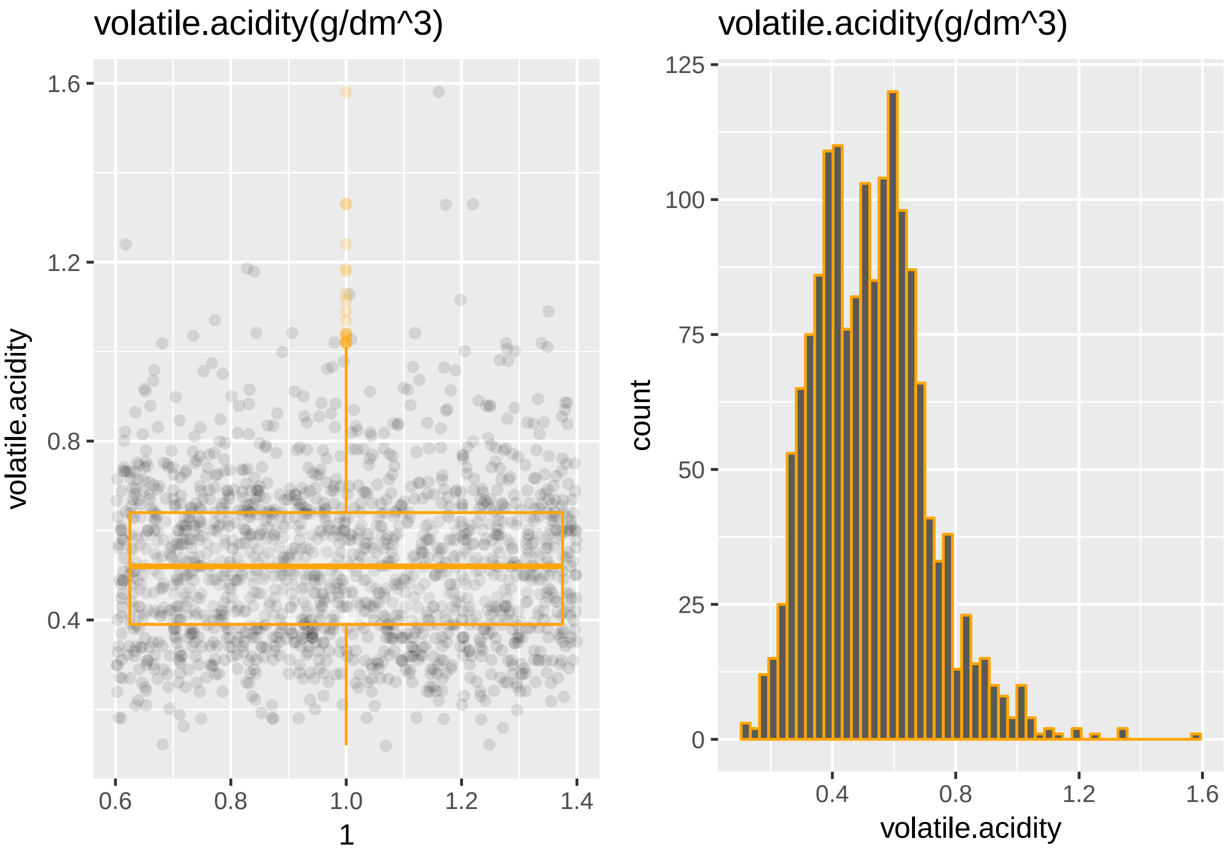
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```



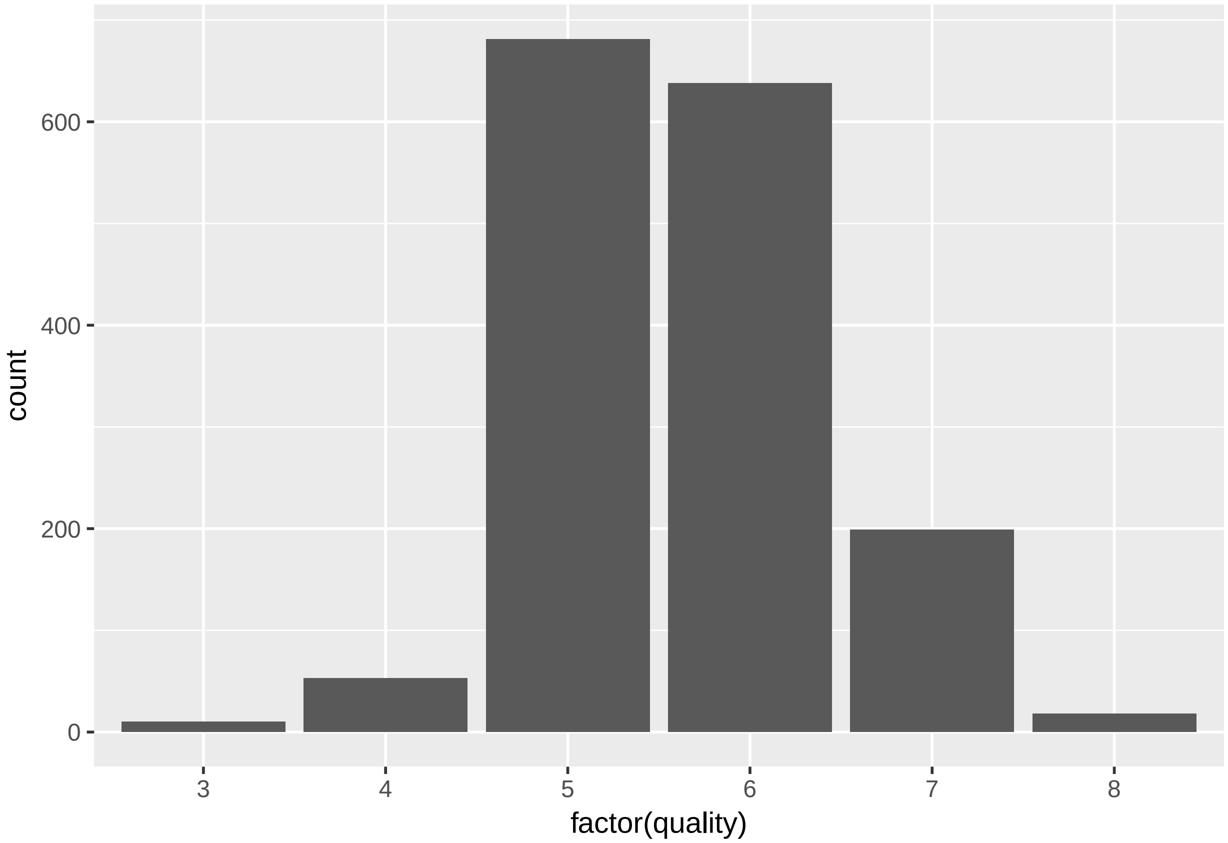
The distribution is asymmetric and the majority of wine counts have an average value of free sulfur dioxide(f-s-d). The median is 14.00, and the mean is 15.87. The mean is above the median which indicates that majority of the wines have considerable amount of f-s-d.

Volatile acidity

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



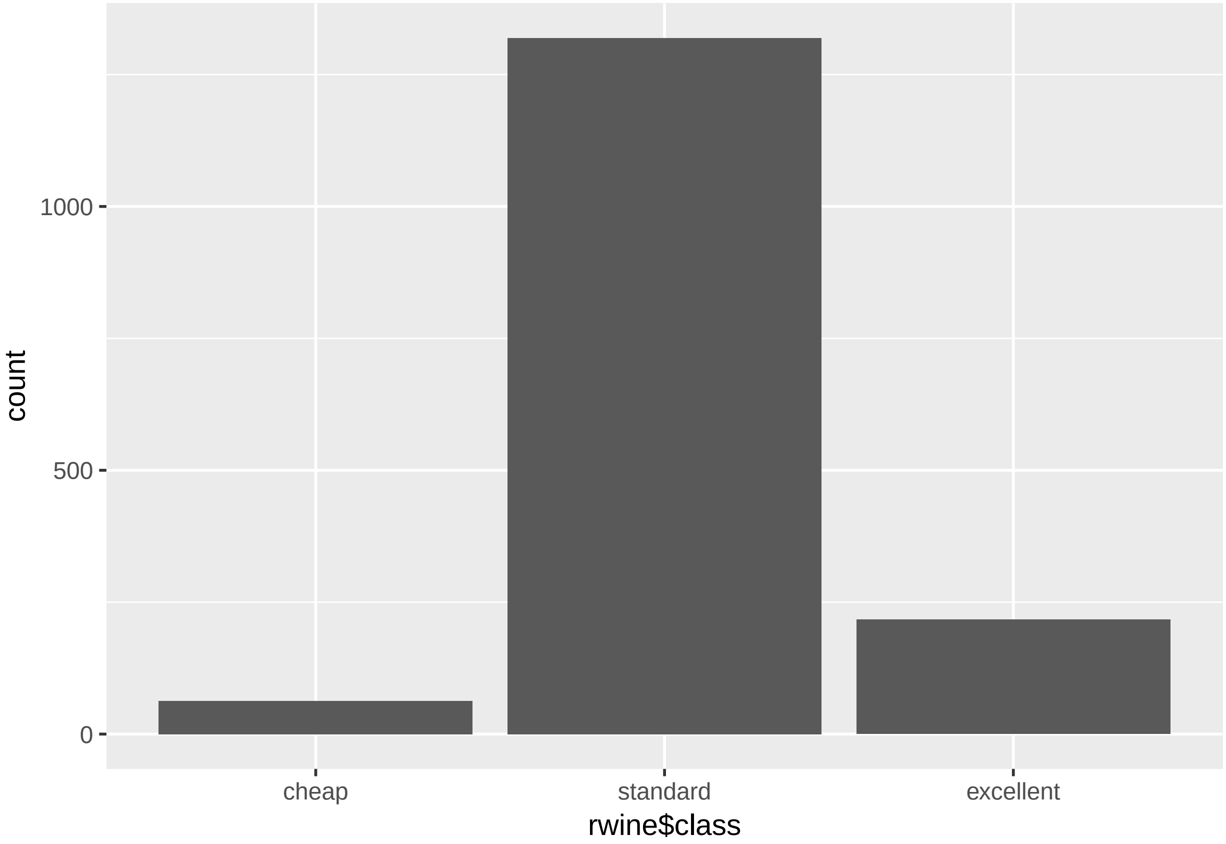
The distribution of volatile acidity has two peaks (bimodal distribution).
The median value is 0.5200 and the mean value is 0.5278. So far, this is giving a neutral contribution for the determination of the quality of the wine, because the difference between the mean and median is very small. The big peak is located around 0.6.



The quality of the wine can be used to create a new variable called “class” to classify the wine as “cheap”=3-4, “standard”=5-6 and “excellent”=7-8. So I will create a new table to add to the dataframe “rwine”

```
##     cheap  standard excellent
##        63      1319       217
```

```
##        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            class
##  Min.   : 8.40   Min.   :3.000   cheap    :  63
##  1st Qu.: 9.50   1st Qu.:5.000   standard :1319
##  Median :10.20   Median :6.000   excellent: 217
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```



Univariate Analysis

What is the structure of your dataset?

The dataset is composed of a total of 1599 wines and consist of
12 parameters. These parameters are used for the investigation of the quality of the wine. One interesting feature of the dataset is that the scaling of each parameter used is different. This lead to a difference in the distribution of the corresponding histograms.

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

The purpose of this investigation is to used the 12 parameters to determine the quality as the mean feature. The quality is a discrete
parameter while the other are continuous. Each chemical element has contribution on the determination of the quality. There some histogram which show normal distribution and it is not easier to decide wether the corresponding parameter have dominant contribution on the determination of the wine quality.

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

The long tail and skewed distribution may not have a dominant contribution of the the quality of the wine. In fact, this distribution has a peak on one side simplify the investigation of their contribution on the wine quality,

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

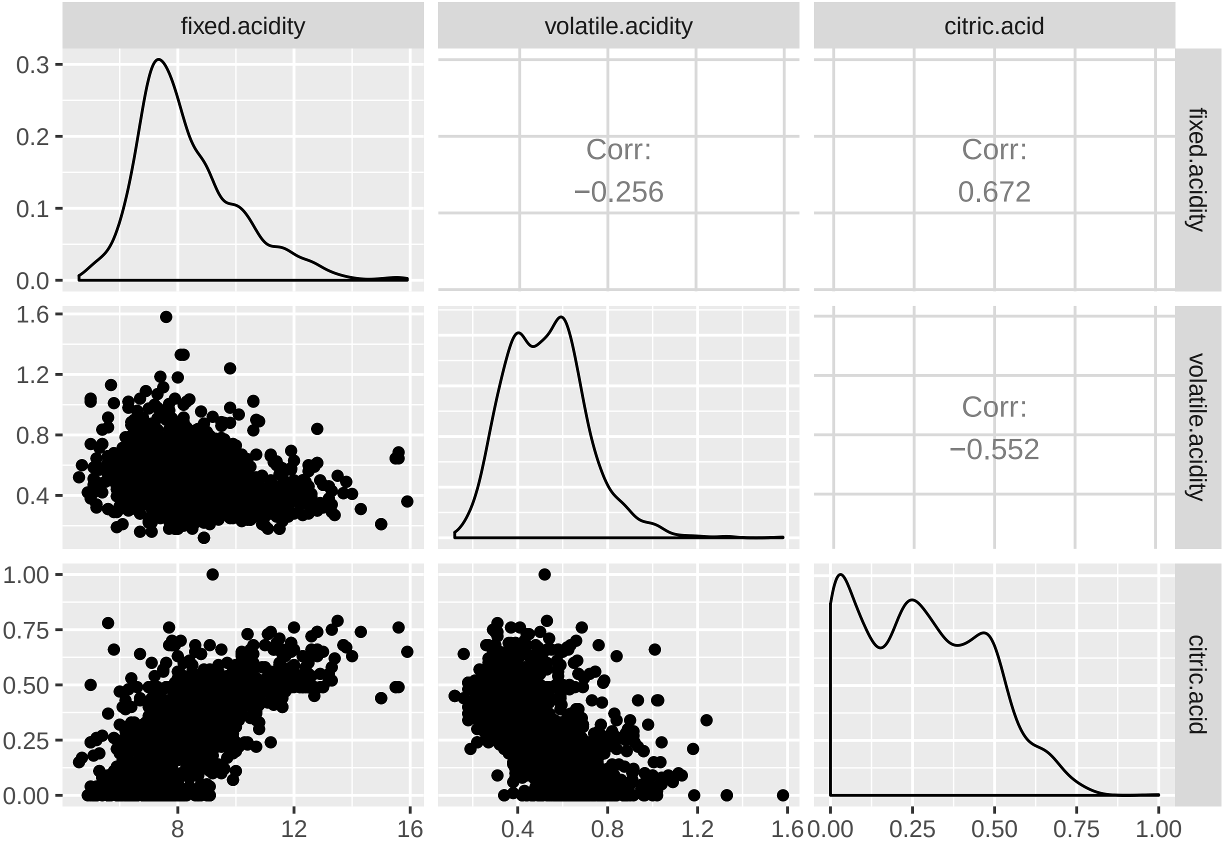
The mean variable of study is the quality. In order to determine the wine with the highest quality we scale the quality by class corresponding to cheap=3-4, standard=5-6 and excellent=7-8

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?

he histogram of sulphates was transformed into a log10 in order to get the distribution used for the analysis. It is not simple to use one plot and determine the parameter that effects the quality of the wines. This issue may be solved by inclusing two or more plots and determines its corresponding correlation. In the next section, We will do the analysis of two plot and evaluate the corresponding correlation function.

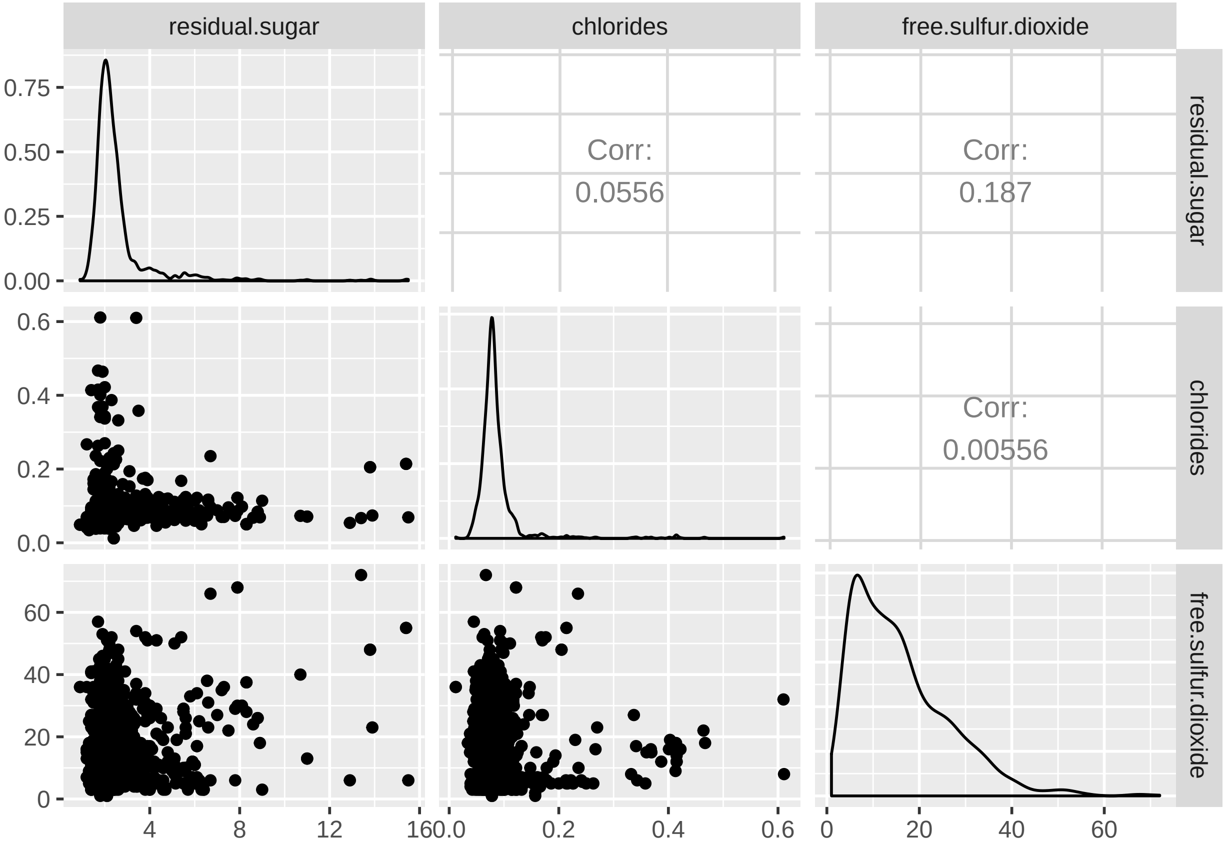
Bivariate Plots Section

I will calculated the of fixed acidety, volatile acidety and citric acid.



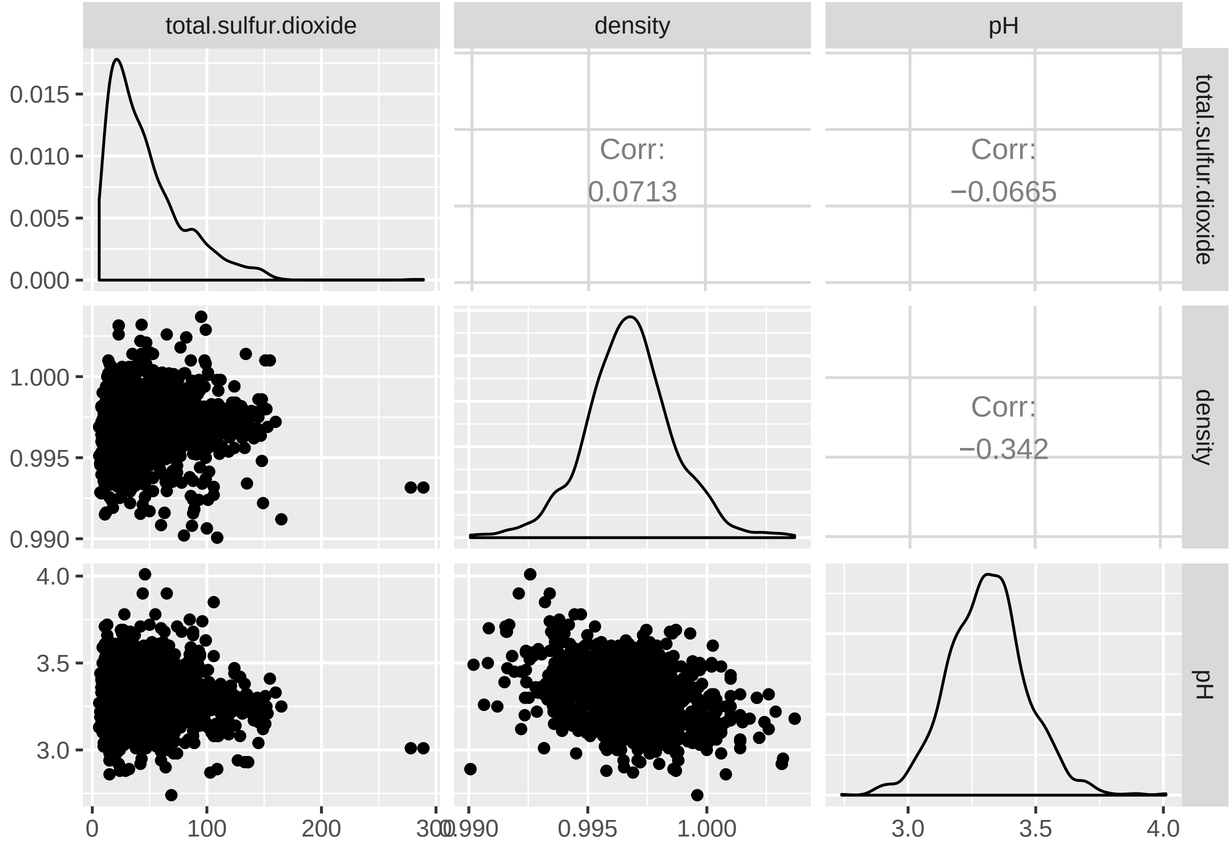
The volatile acidity and the fixed acidity have opposite effect on the quality of the wine because are not correlated. #### Similar effect happens between the citric.acid and volatile acidity, the corresponding correlation is negative.This implies that the citric acid and fixed acidity are correlated. This feature can ne seen from the table above, their correlation value positive.

Corelation of residual sugar, chlorides and free sulfur dioxide

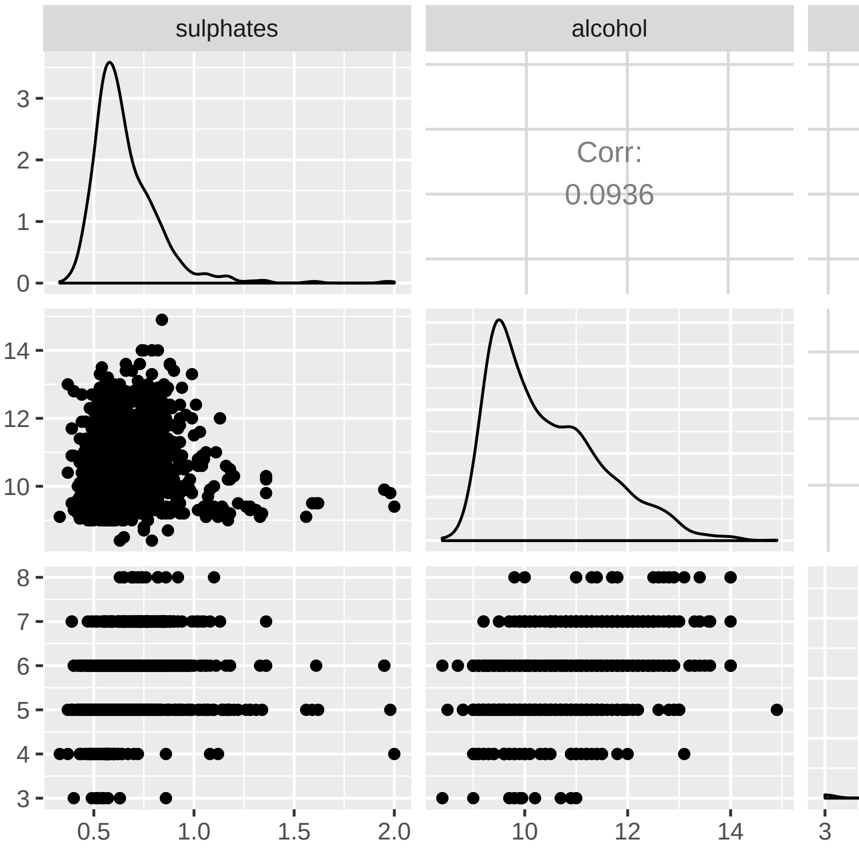


One intersting feature, is that all the quantities in the table anove are correlated(residual sugar, chlorides and free sulfur dioxide). This indicates that they have similar effect on the determination of the quality of the wine.

Corelation amongs tota sulfur dioxide density and pH

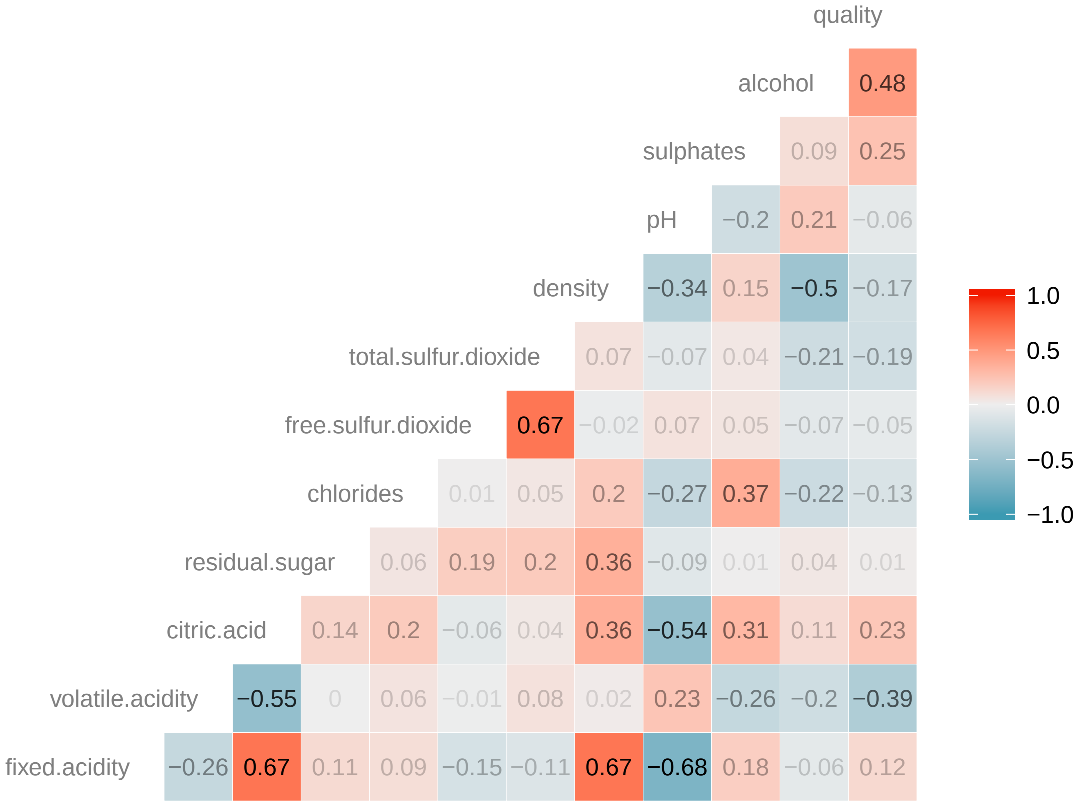


The pH is not correlated to both density and total sulfur dioxide. As a result, the latter and the former are strongly correlated.
This can also be seen from their position correlation value.



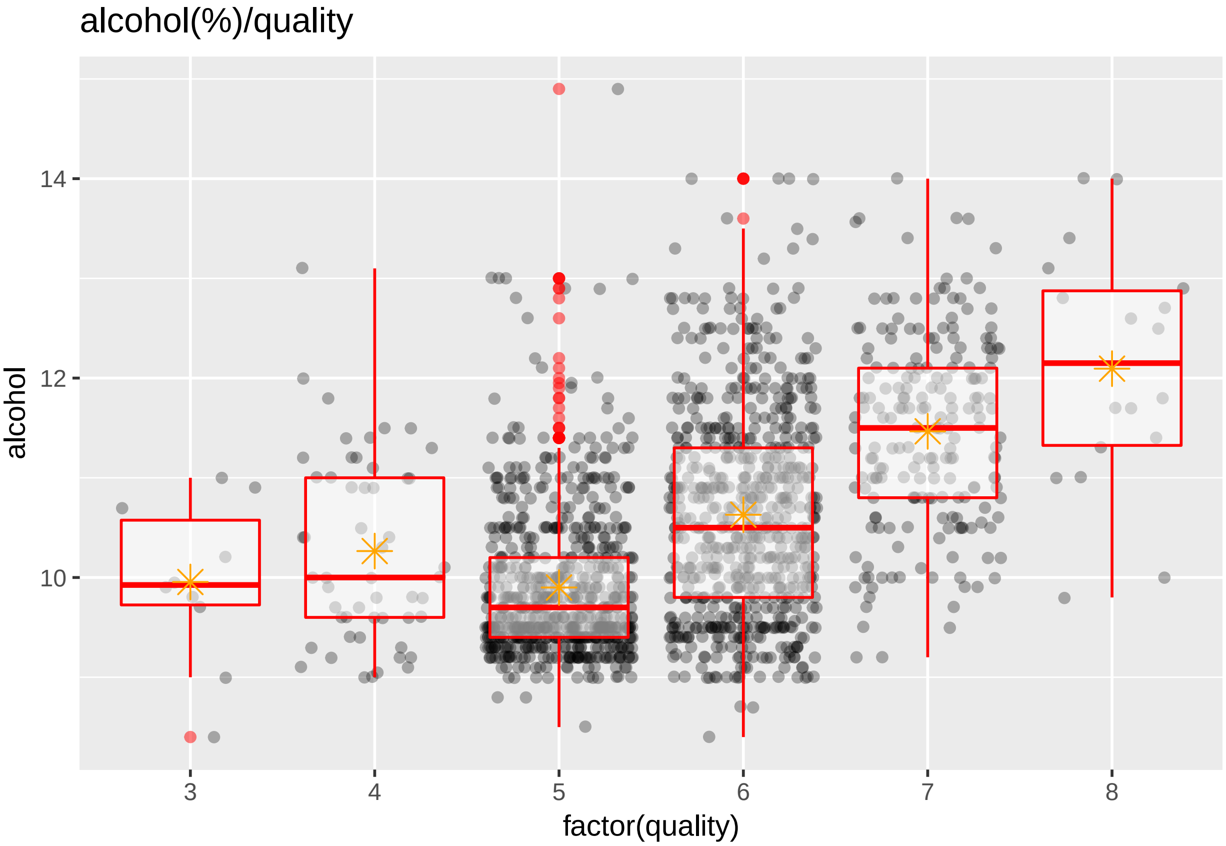
####Correlation of sulphates, alcohol and quality

The quality is strongly correlated to both alcohol and sulphates.

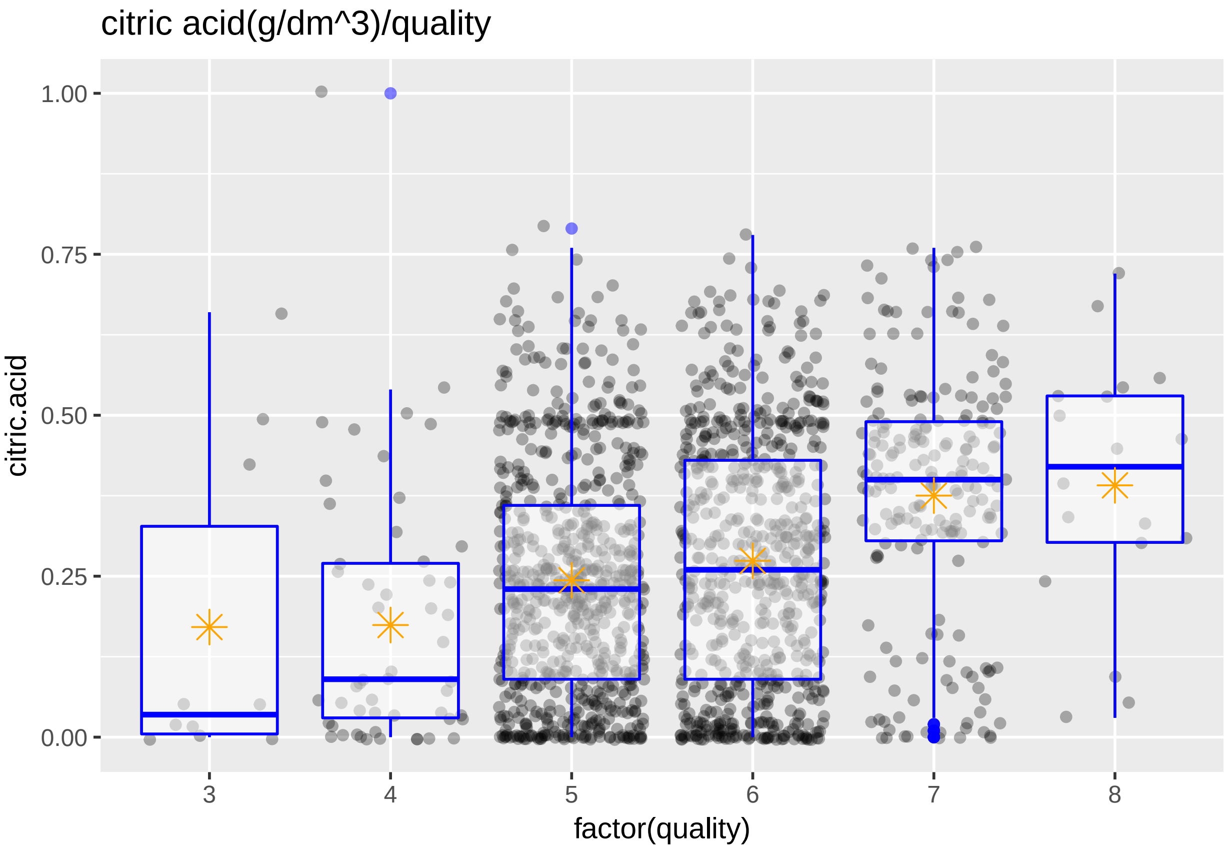


From the table above of the correlation with different parameters, we see that the quality is strongly correlated with: alcohol, sulphates, citric.acid.

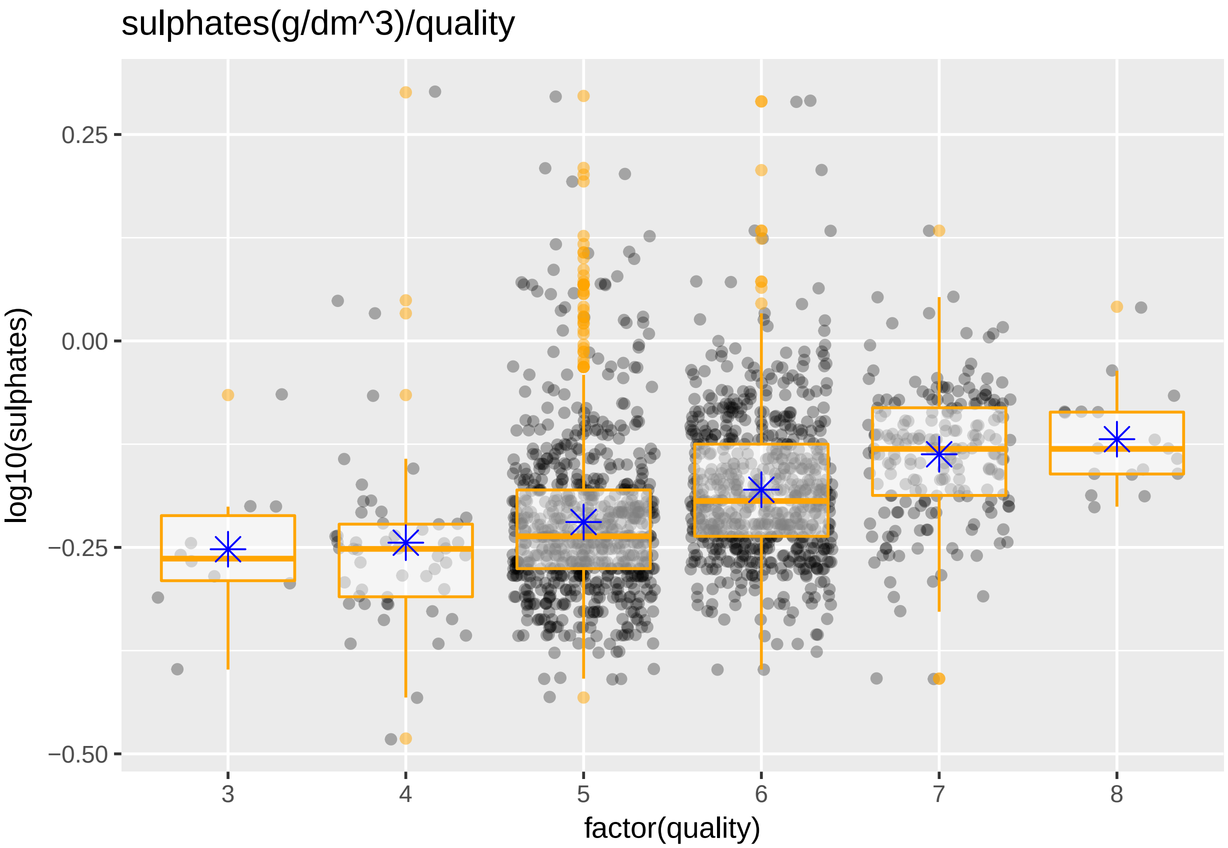
So far we can speculate that this two quantities may determine the quality of the wine. In order to confirm this hypothesis, we will plot differents graphics of the quality as a function of the 12 parameters above.



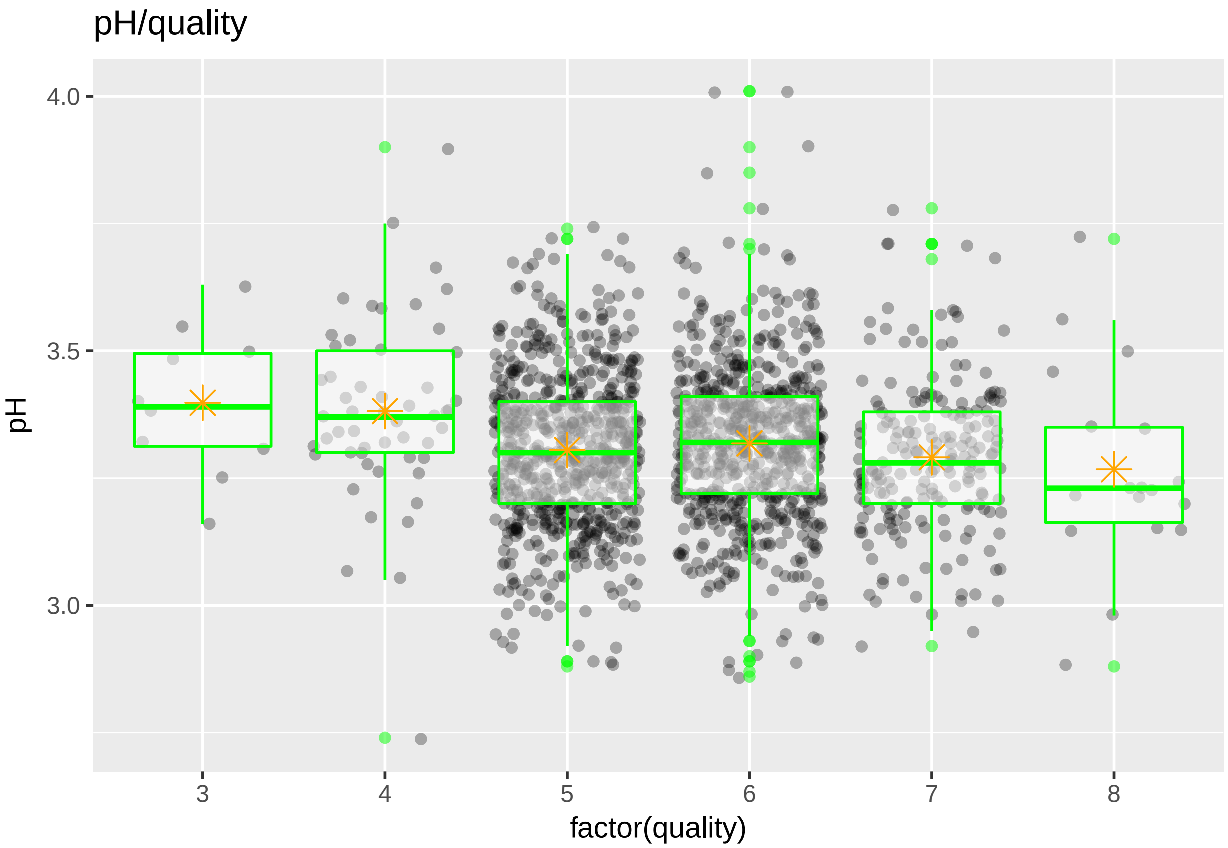
This graphics clarify what we observed earlier about the feature of the histogram of the alcohol. This result clearly illustrates the effect of the alcohol for the determination the quality of the wine. #### The high quality wines has more alcohol than all other wines.



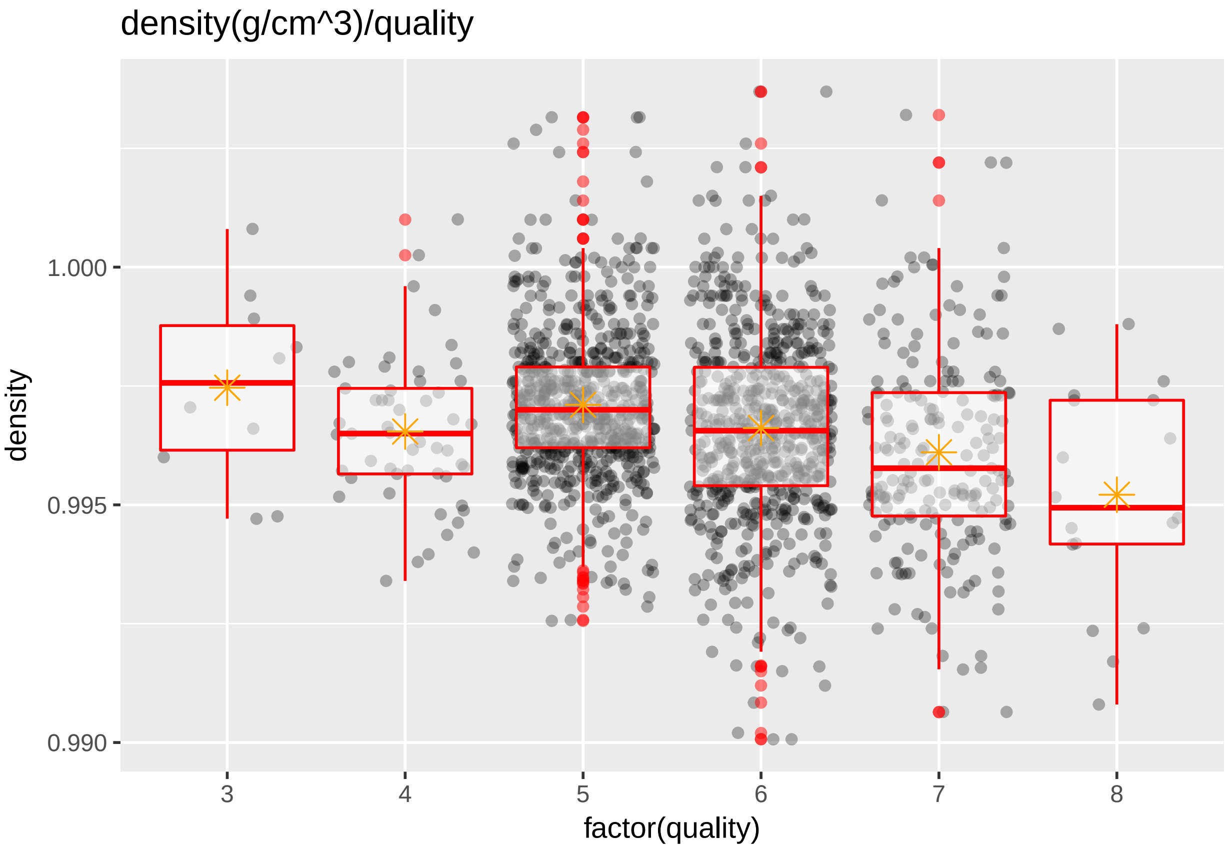
The citric acid also has a positive impact on the determination of the wine quality.



The sulphates increases the quality of the wine. I would be interesting to compare the sulphates wit the alcohol, citric acid to see which parameter have dominant contribution on the determination of the wine quality.



The pH decreases the quality of the wines. Lower wine quality have high pH levels.



The density has a negative impact on the quality of the wine. The quality of the wine increases with the decreases of pH concentration.

The average of the alcohol content in all wines is between 10 to 11 %. The citric.acid, sulphates also have an impact on the determination of the quality of the alcohol, and their corresponding graphics are plotted below.

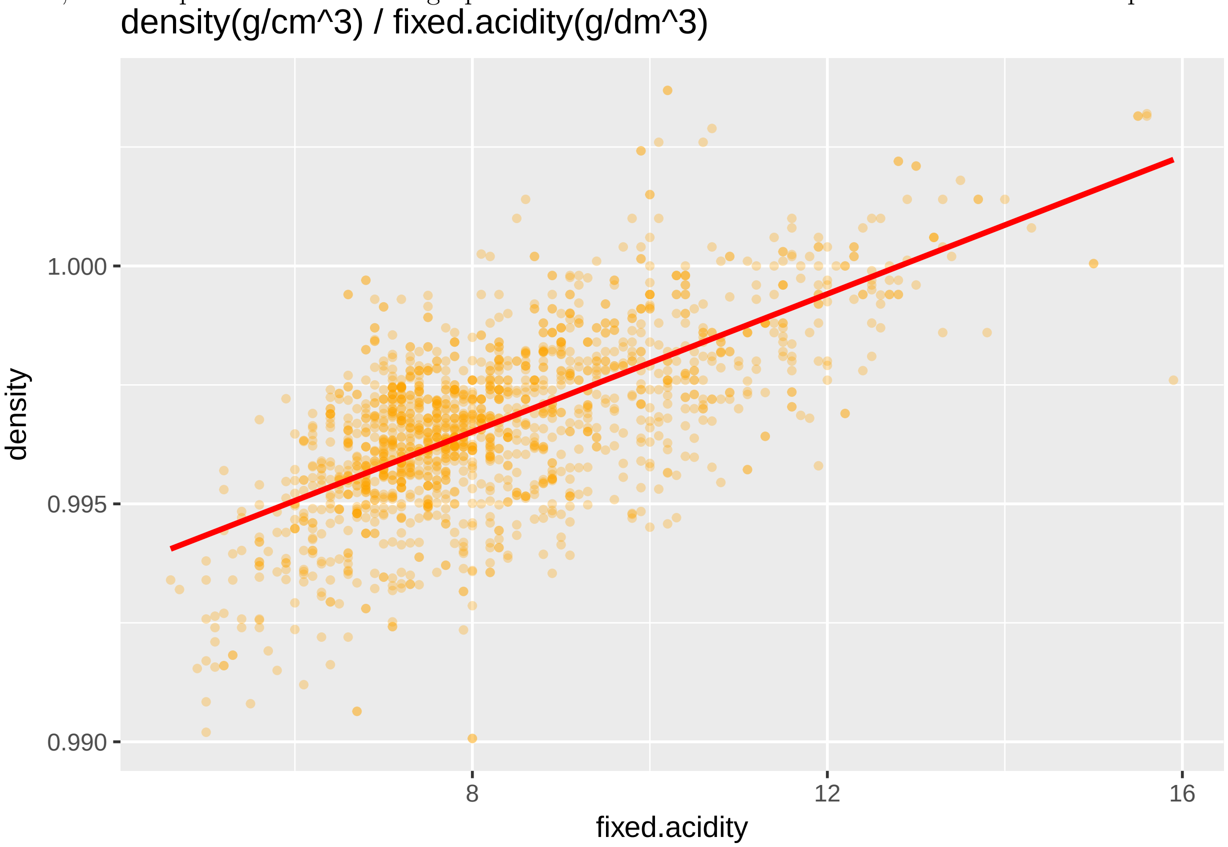
The box_plot gives the quantities which high affect the quality of the wines. In this plots is evident that the higher quantity of the alcohol, sulphates and the citric.acid a give higher quality of the alcohol. In order to check this result, I consulted the price of the wine in my local shop.I realized that the wine with alcohol percent below 10 was cheaper than the wine with high alcohol content. This conclusion is also valid for the determination of the quality of the beer. The graphic below illustrates the parameters that most influence on the wine quality. Another parameter increases the wine quality is the lower pH, residual sugar, and density respectively.

```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"              "class"
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165           0.47616632           0.25139708
##  free.sulfur.dioxide total.sulfur.dioxide              density
##          -0.05773139          -0.17491923          -0.18510029
##                   pH              alcohol                 <NA>
##          -0.05065606          -0.12890656           0.01373164
##                 <NA>                 <NA>
##           0.22637251          -0.39055778
```

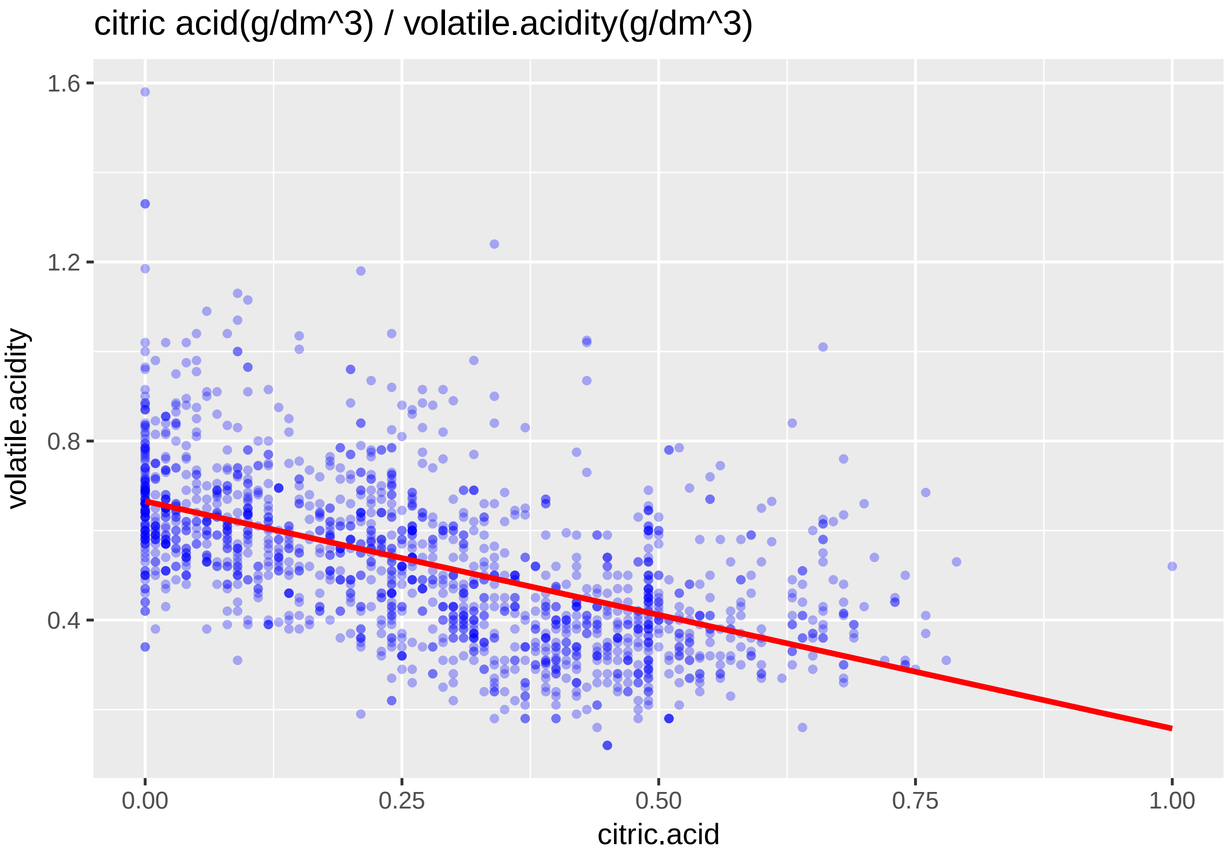
This result gives the information about the quantities which have high correlation are the following

Next, I will plot the scattered graphic to visualize the correlation between different parameter.



The density and the fixed acidity have positive correlation.

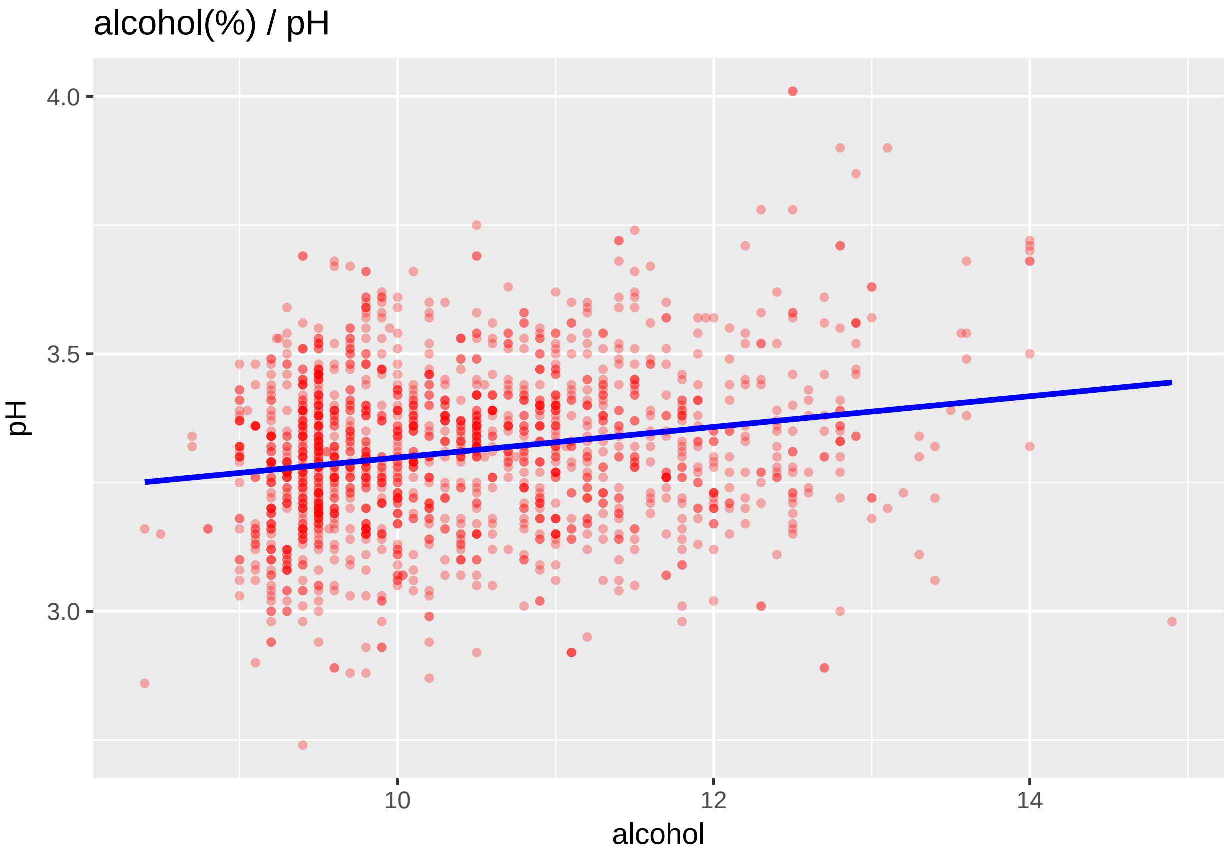
```
## [1] 0.6680473
```



The volatile acidity and citric acid have opposite effects, and negative correlation.

```
## [1] -0.5524957
```

The citric acid increases the wine quality, while the volatile acidety decreases the quality of the wine.

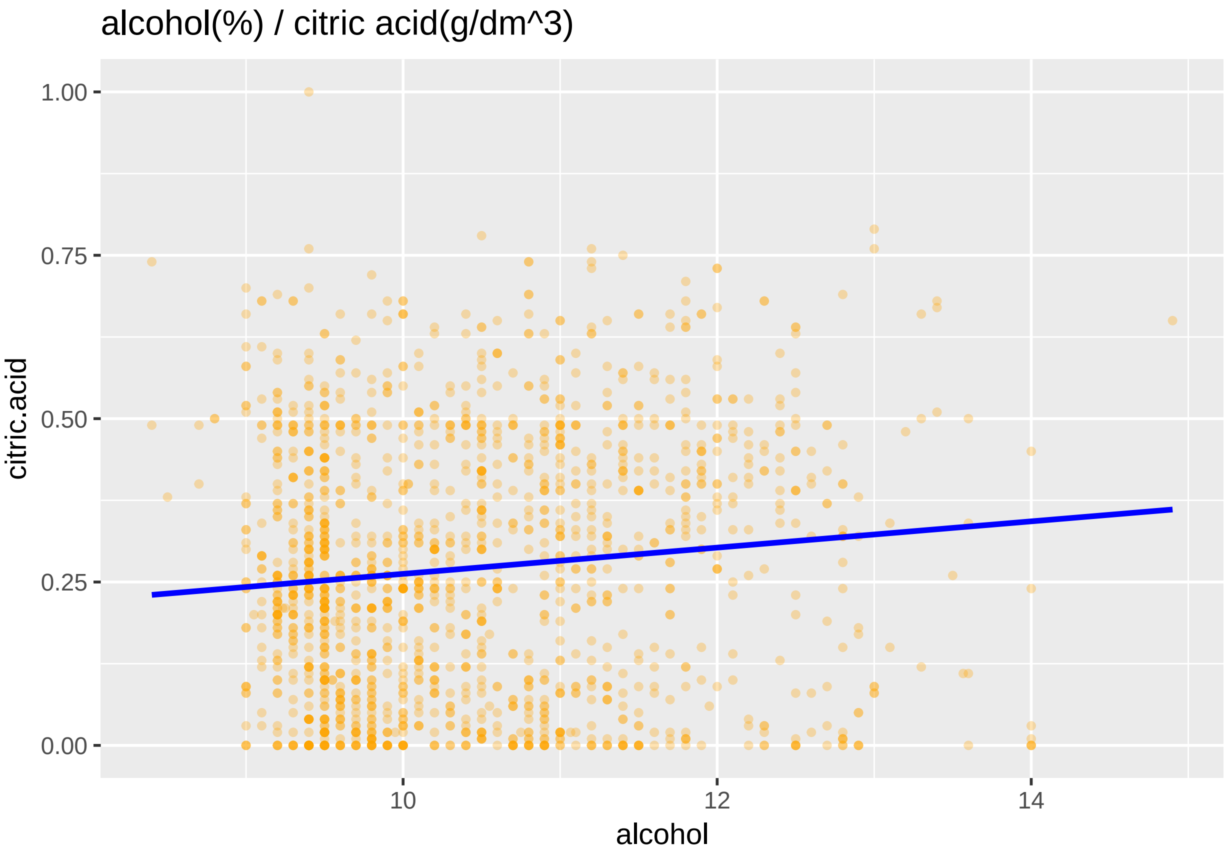


The graphic of the pH and the alcohol have positive inclinataion which

indicates the existence of a posetive correlation.

```
## [1] 0.2056325
```

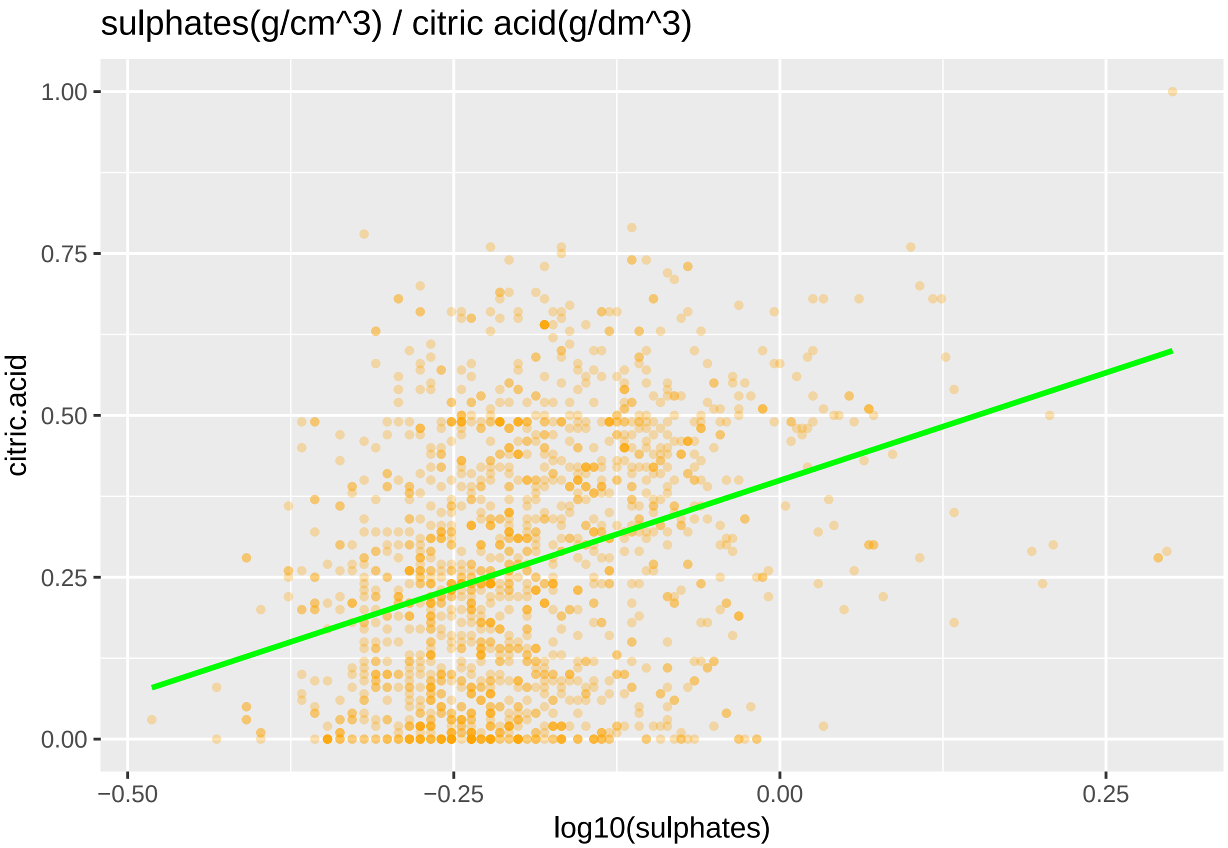
This result suggests that theses two parameter have positive effect on the quality of the wine. The multiplot graphycs, may conform or reject this idea.



The citric acid and the alcohol content both have positive effect on the quality.

```
## [1] 0.1099032
```

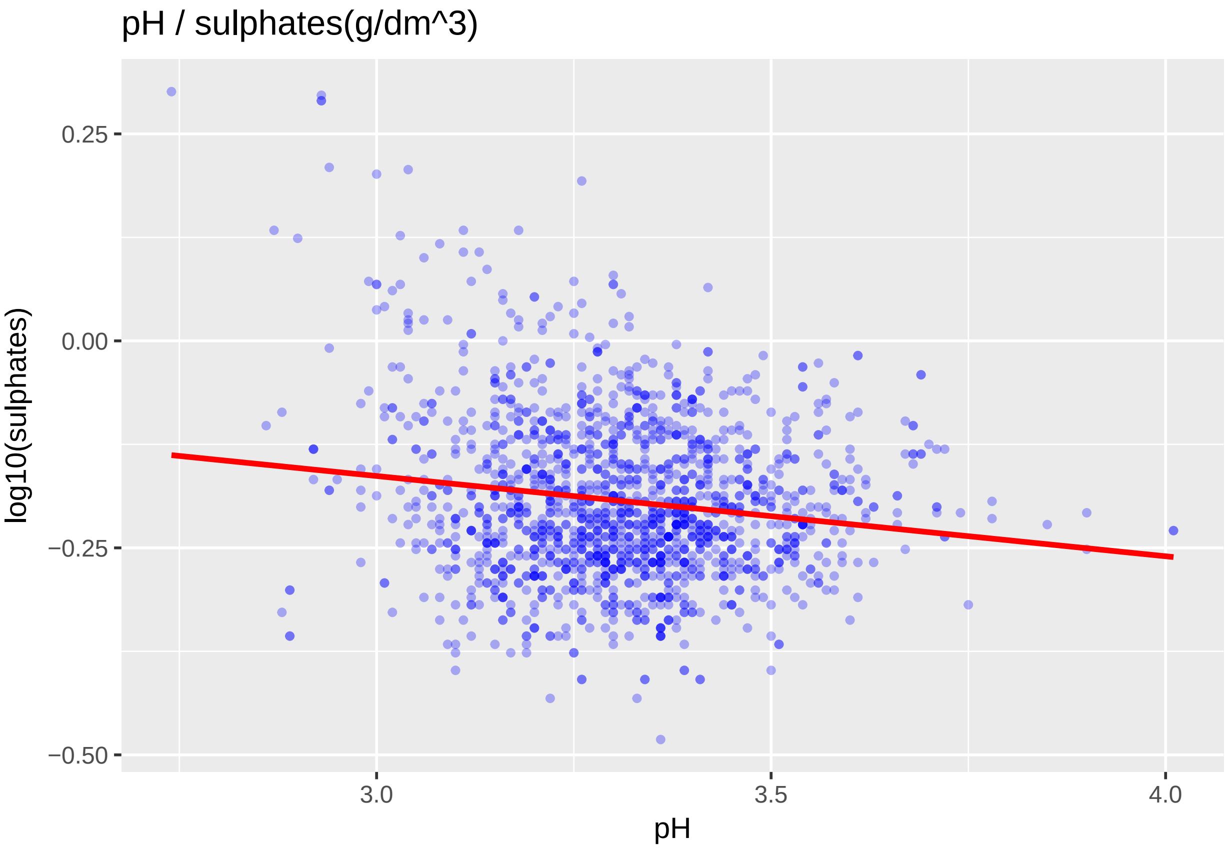
The slope of the graphic is positive which indicates a positive correlation.



The citric acid concentration increases linearly with the sulphates.

```
## [1] 0.31277
```

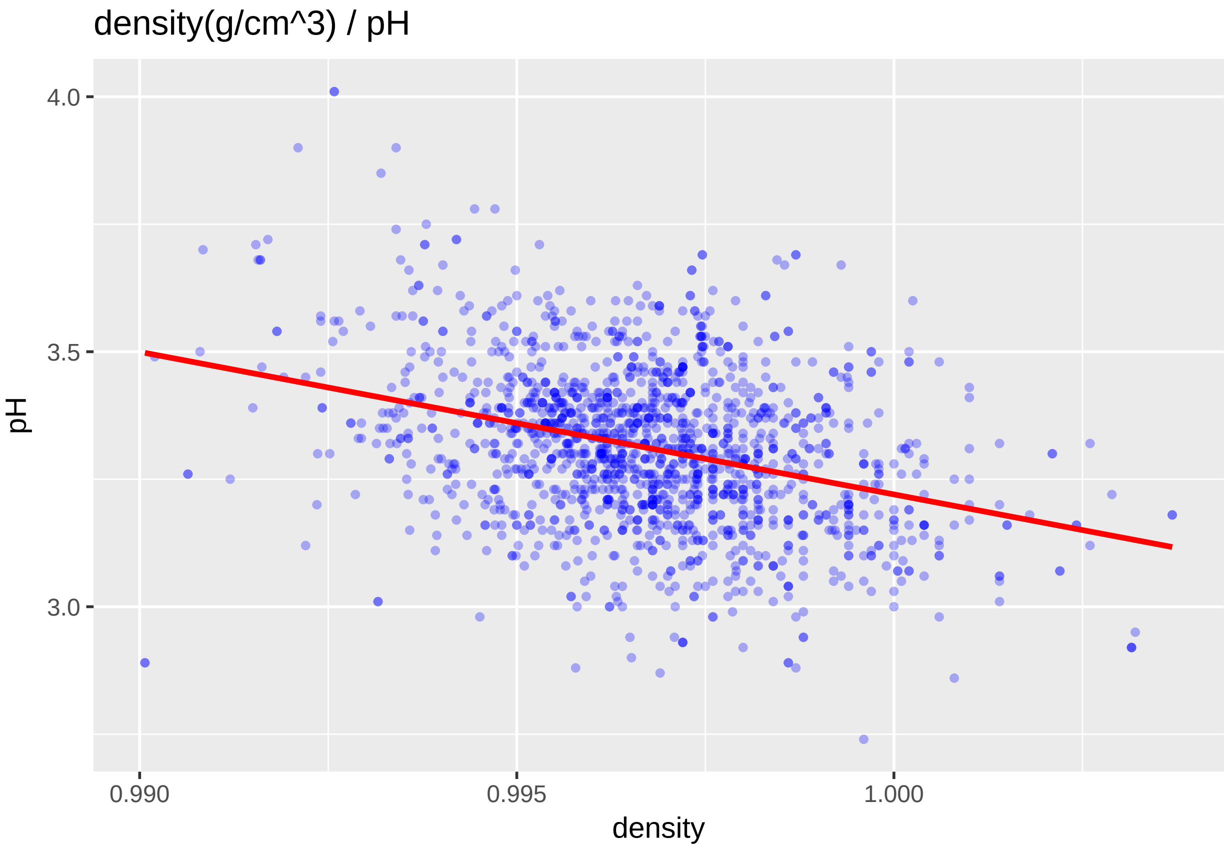
This indicates the existence of a Positive correlation,



The ph decreases the sulphates. However, the pH has a positive correlation with the alcohol

```
## [1] -0.1966476
```

This indicates the existence of a weak Negative correlation



The density decreases when pH increases. These quantities are expected to have opposites effect of the wine quality.

```
## [1] -0.3416993
```

This indicates the existence of moderate Negative correlation

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?

When the pH increases the and citric.acid decreases. The latter is more correlated with the alcohol content and also can be used to determine the quality of the wine.

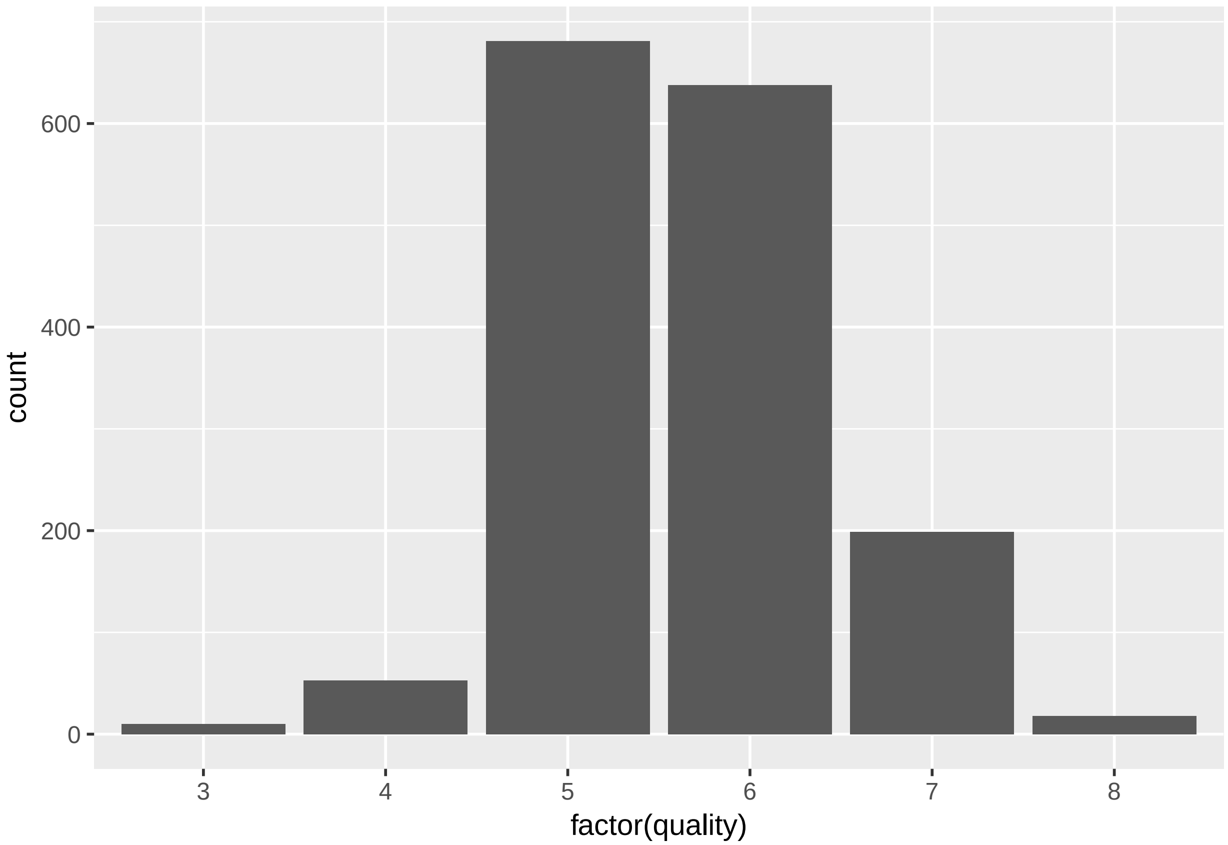
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

The relation between the alcohol and the log10(sulphates) is interesting correlation. The scatter point are concentrated at the average value of the two quantities which shows the strong correlation. The ph have positive correlation with the alcohol. However, it has negative correlation with both the sulphates and the citric acid. This feature will be discussed with details in the multivaries plots setion.

What was the strongest relationship you found?

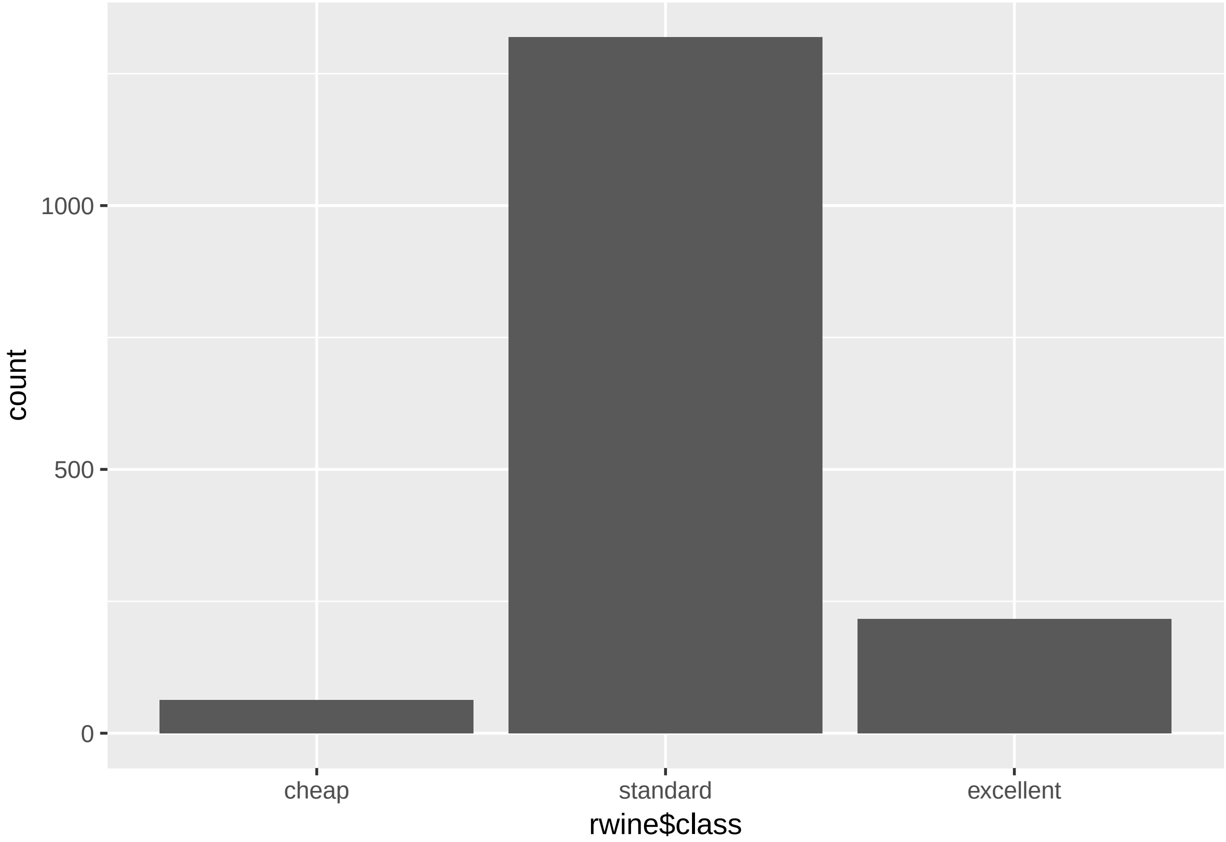
The alcohol and sulphates and the citric acid

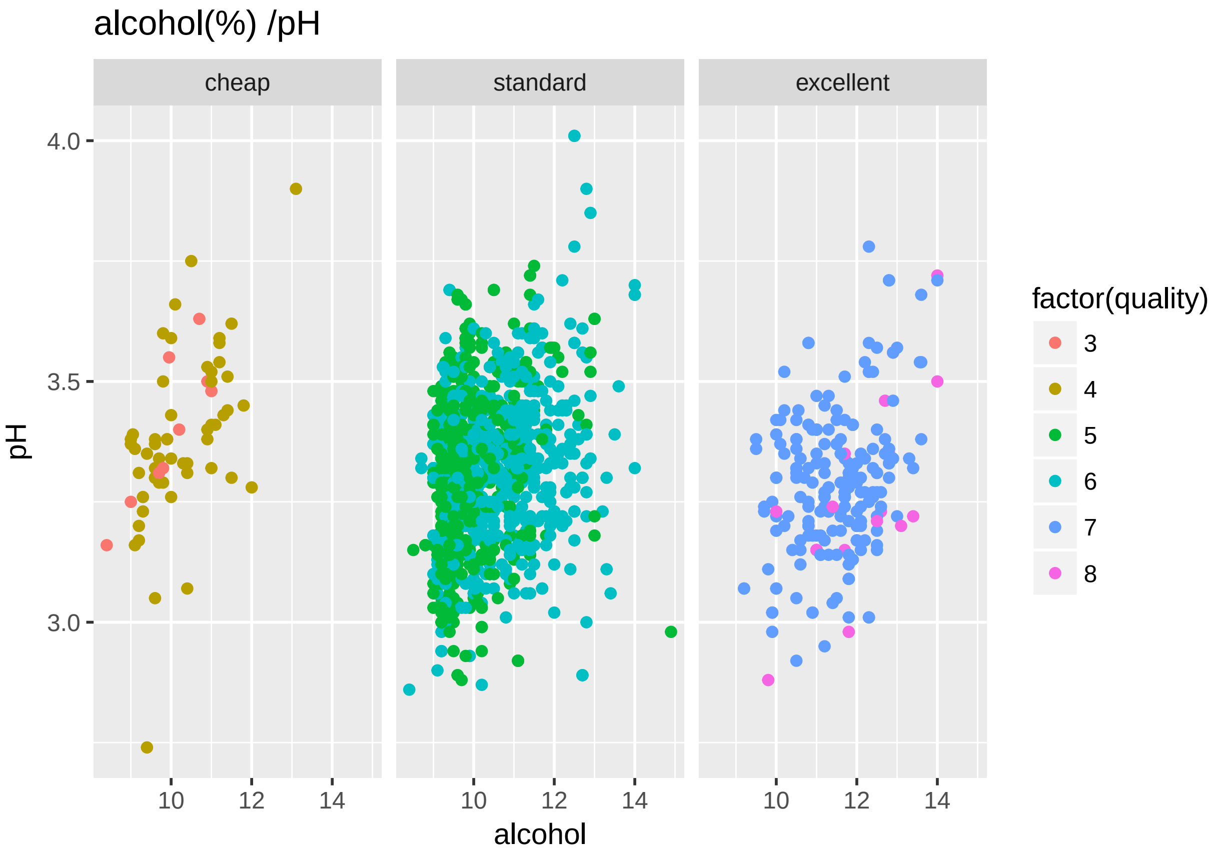
Multivariate Plots Section



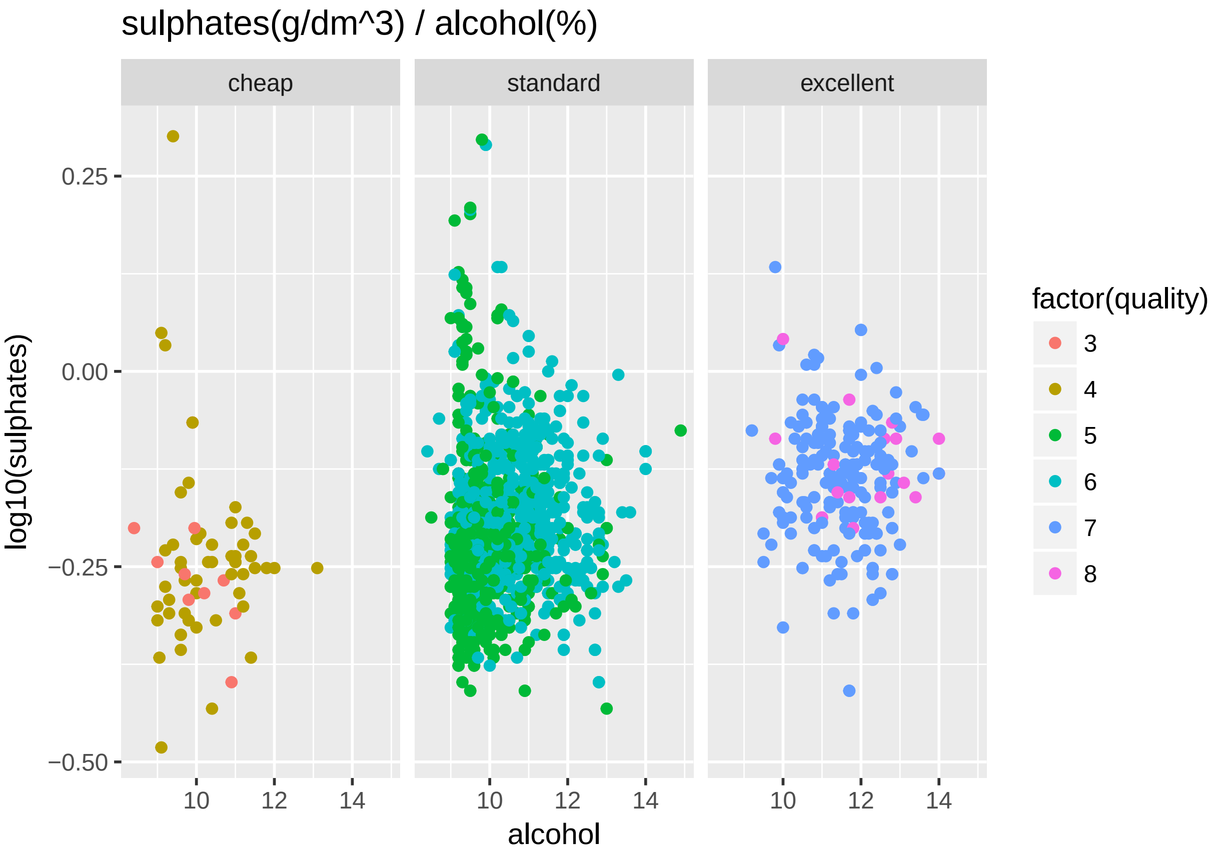
```
##     cheap  standard excellent
##        63      1319       217
```

```
##        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             class
##  Min.   : 8.40   Min.   :3.000   cheap    :  63
##  1st Qu.: 9.50   1st Qu.:5.000   standard :1319
##  Median :10.20   Median :6.000   excellent: 217
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

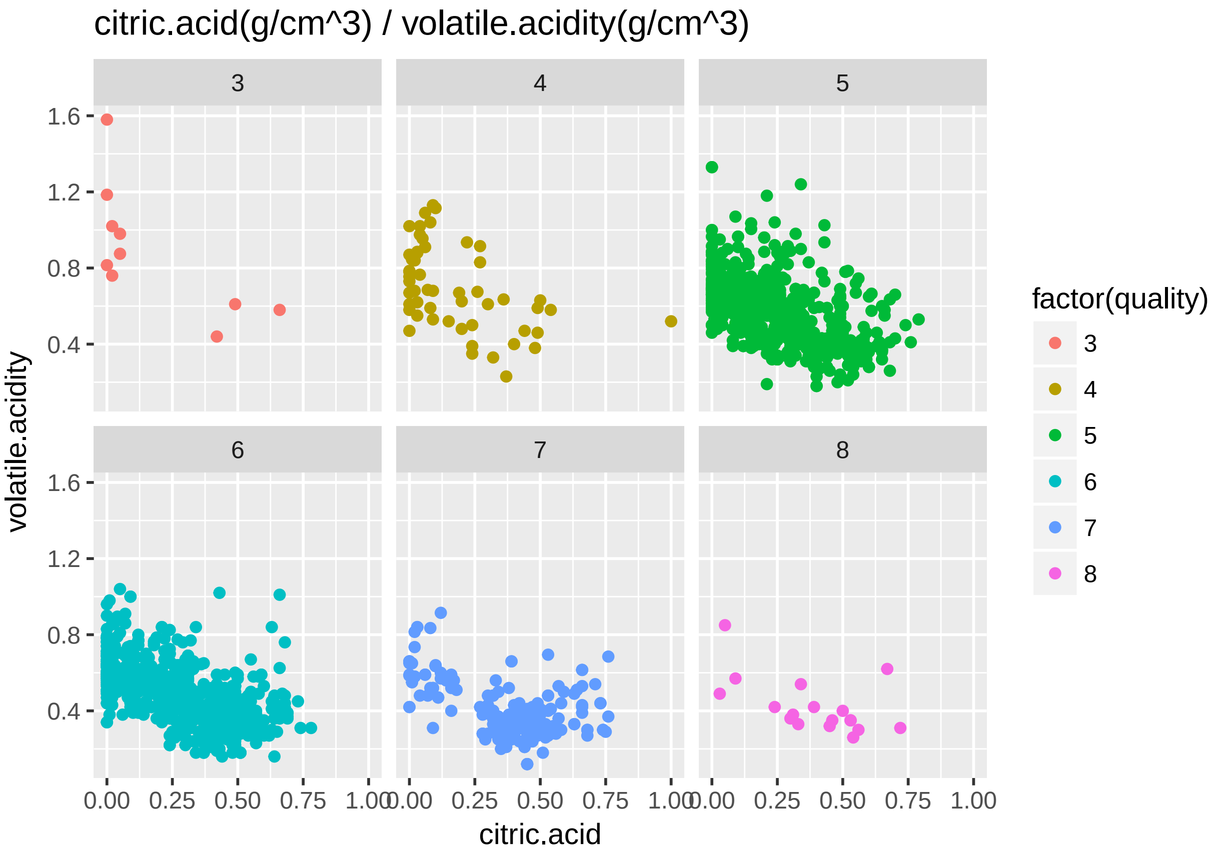




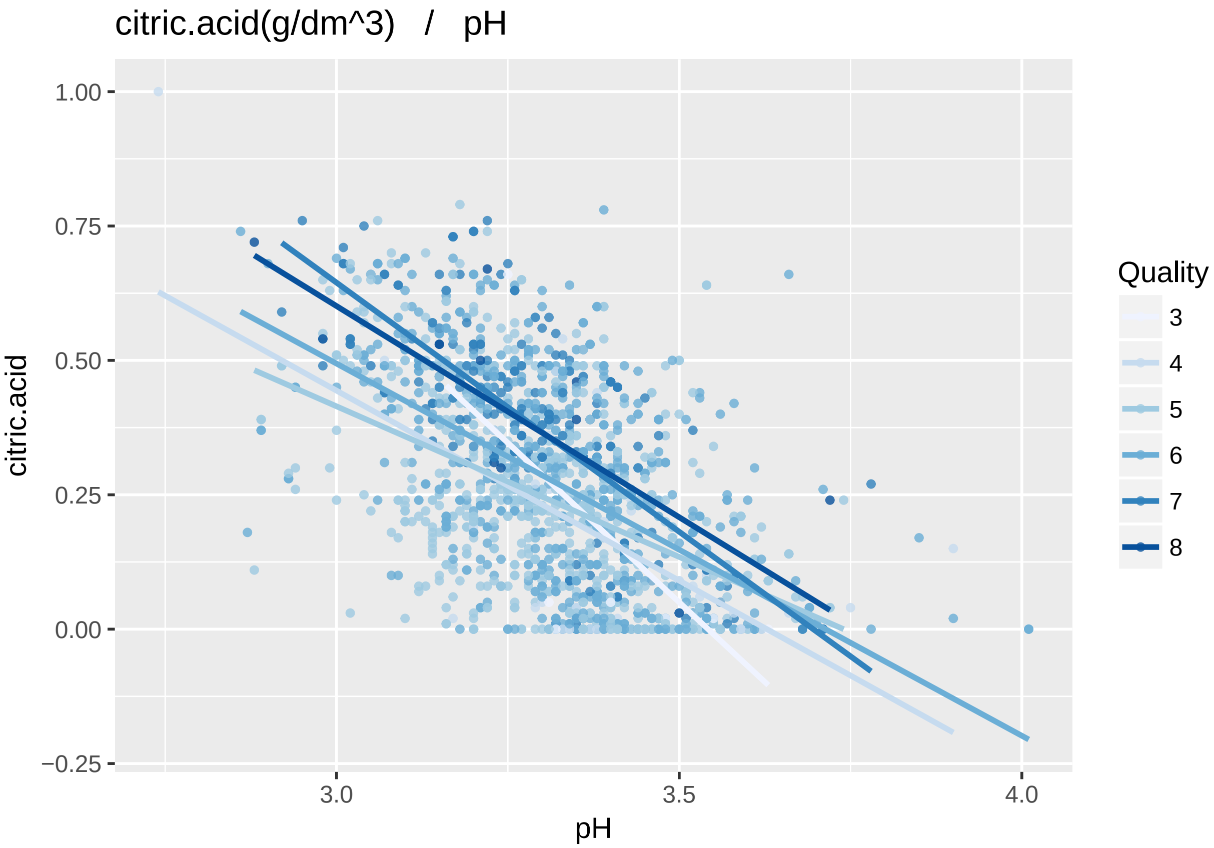
The concentration of pH is little in the high quality wine, while
the alcohol concetration tend to increase the quality wine. On the other hand the cheap wine have higher pH values and less acohol content. The standard wine has moderate values of both ph and alcohol. This is interesting because the pH has positive correlation with the alcohol, however it does not increases the quality of the wine.



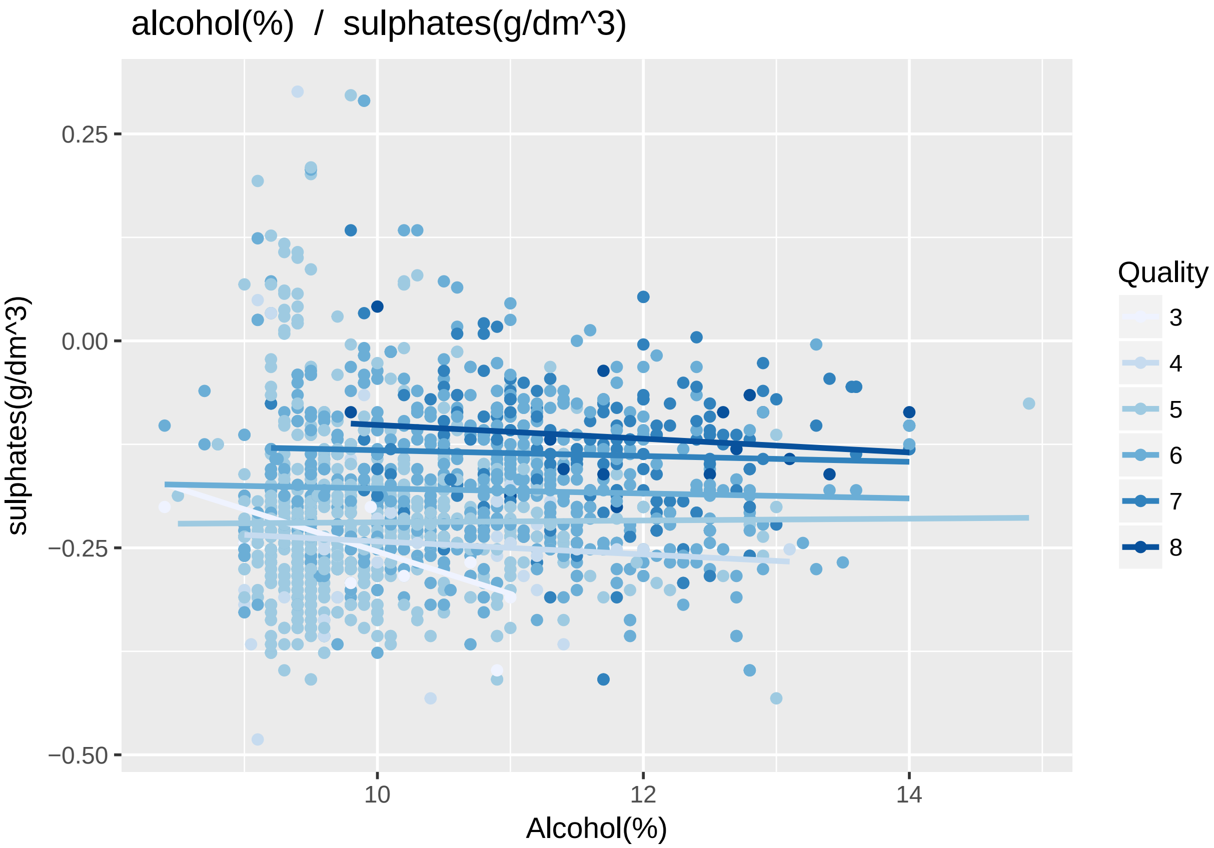
Here we clearly see the impact of the sulphates on the quality of the wine. The concentration of sulphates is lower in cheap wine and high in excellent wine. One interesting feature is that the standard wine have some points with indicates high sulpahte concentration with lower concentration of alcohol. This suggests that the alcohol concentration has more impact on the determination of the quality.



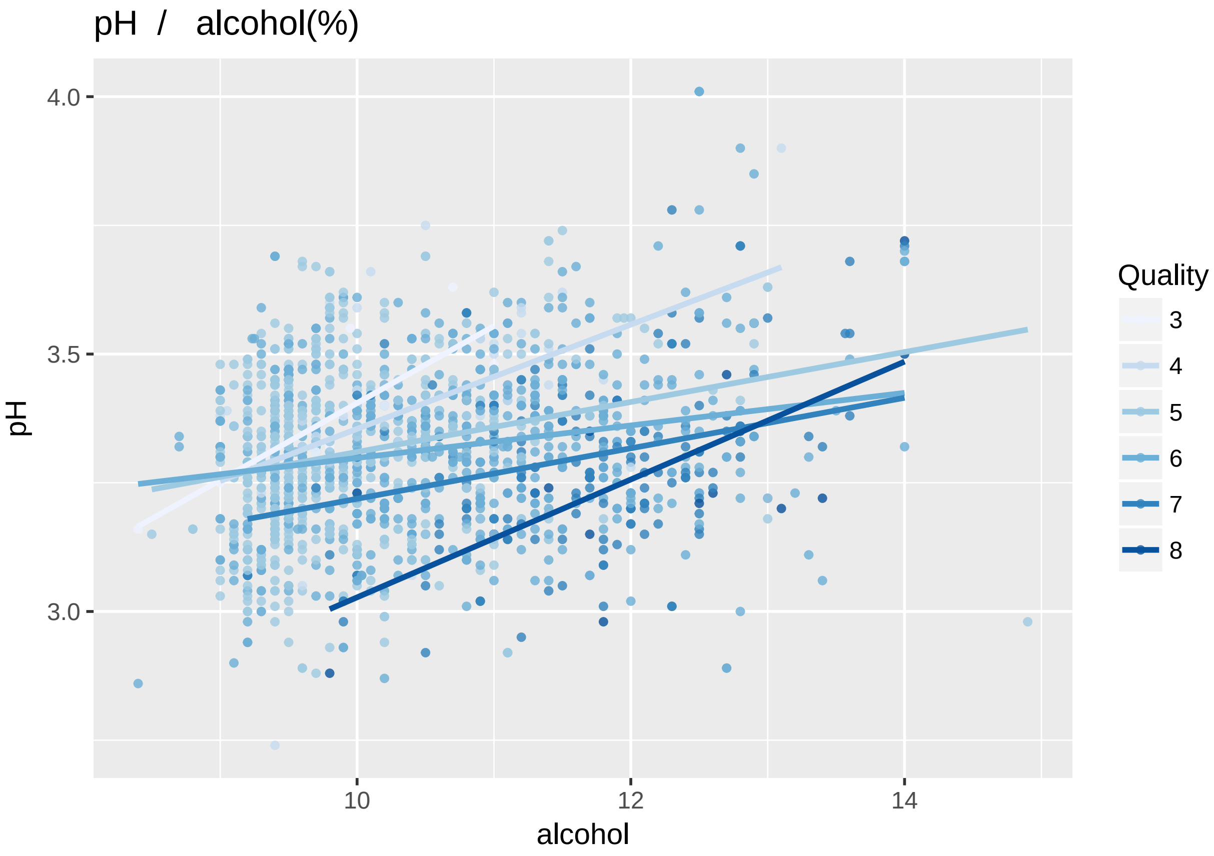
The increases of the volatile.acidity lower the quality of the wine. While the citric.acid has opposite impact. This can be seen from the cheaper wine whitch has more volatile acidity and lower citric acid.



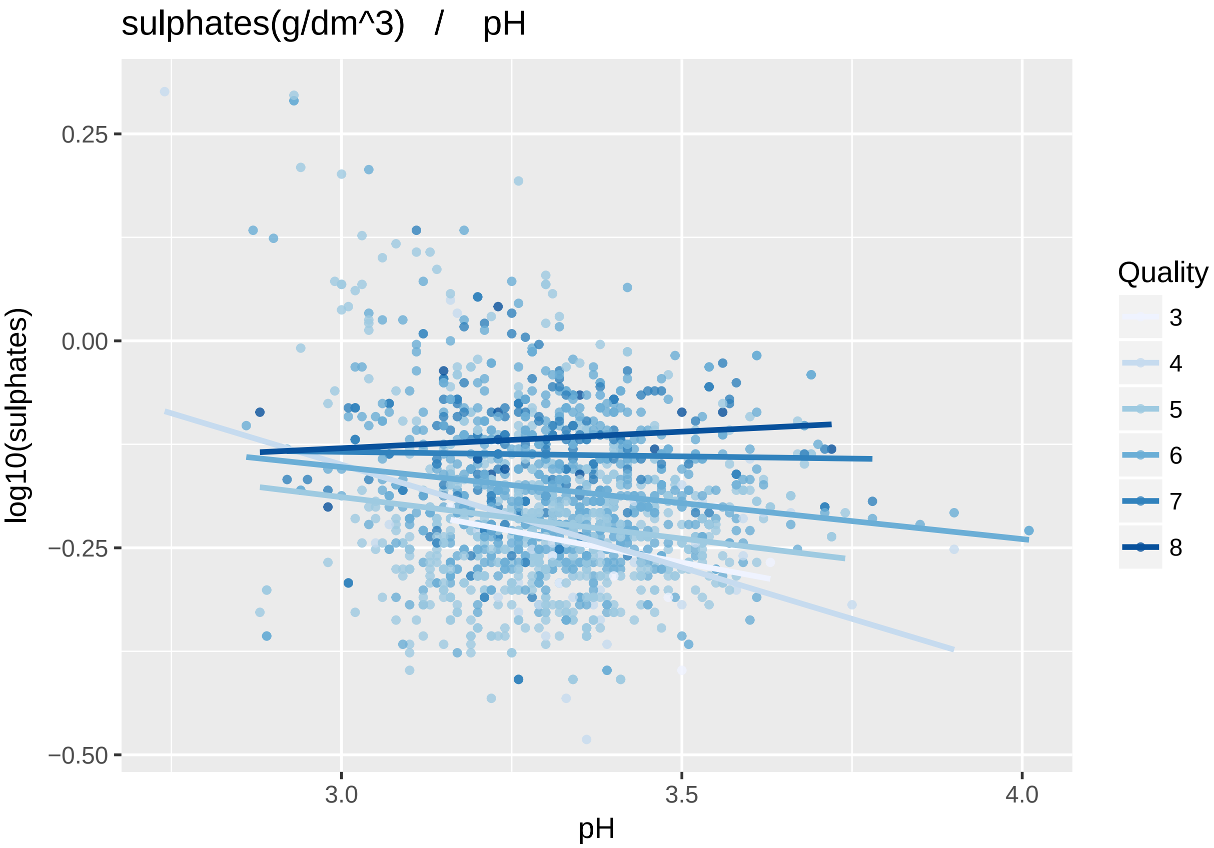
The pH decreases the quality, and the citric acid increases the quality.



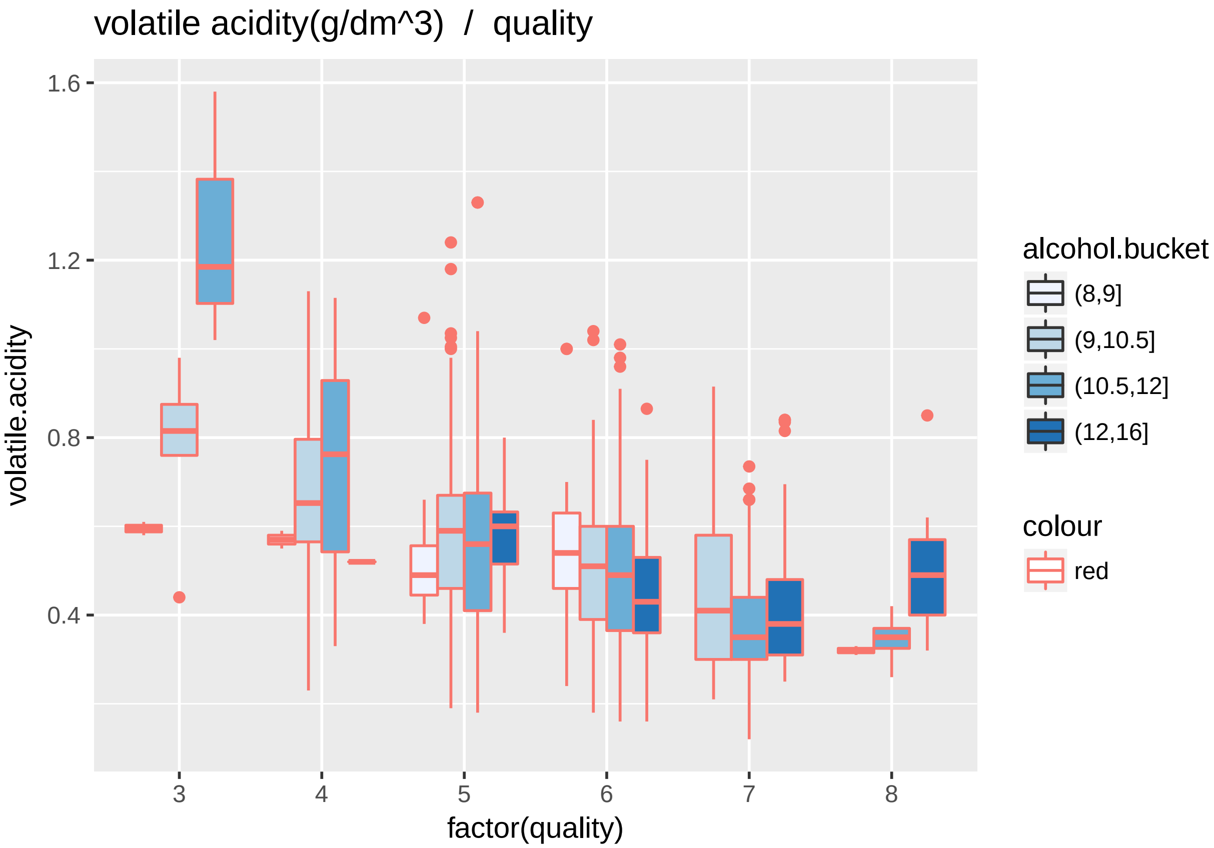
The alcohol has positive correlation with the sulphates. In addition these parameter have dominant contribution on the wine quality.



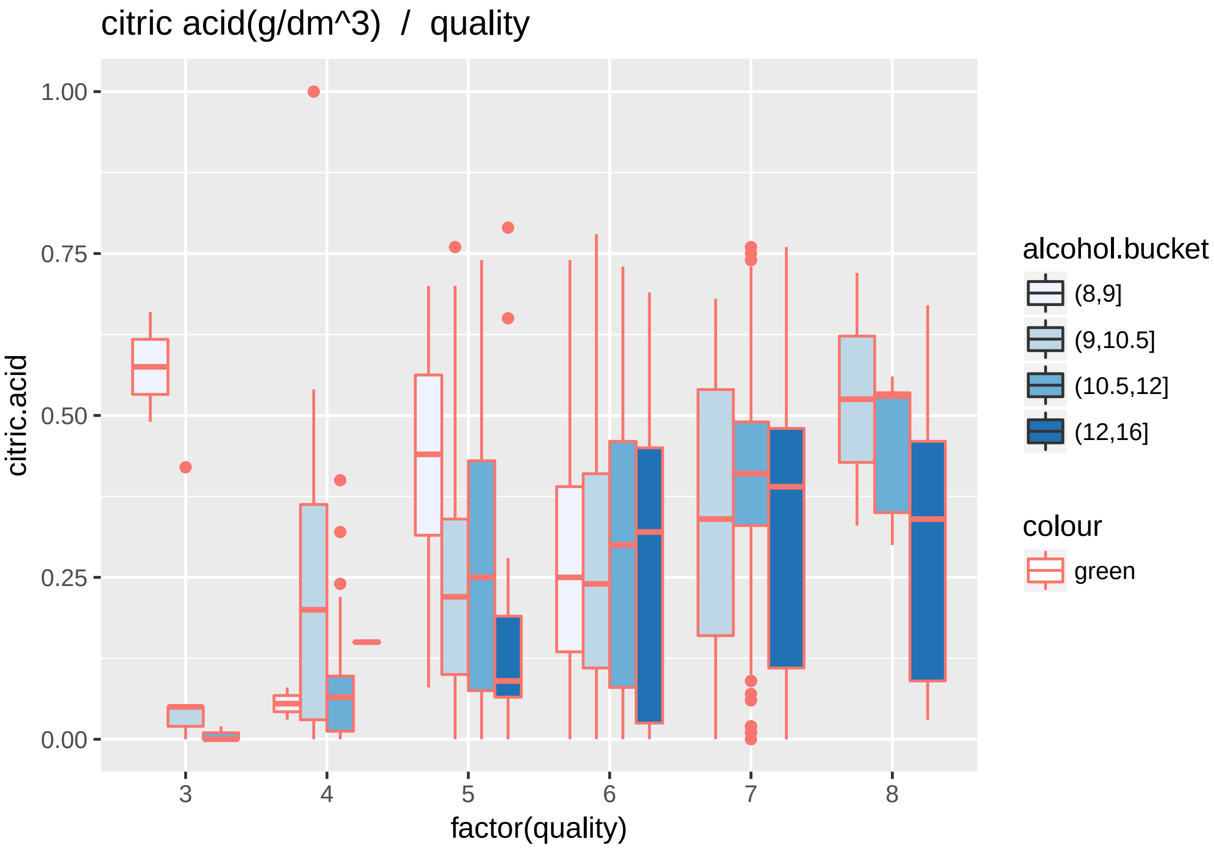
The alcohol has positive relationship with the pH. The high quality wine has lower pH.



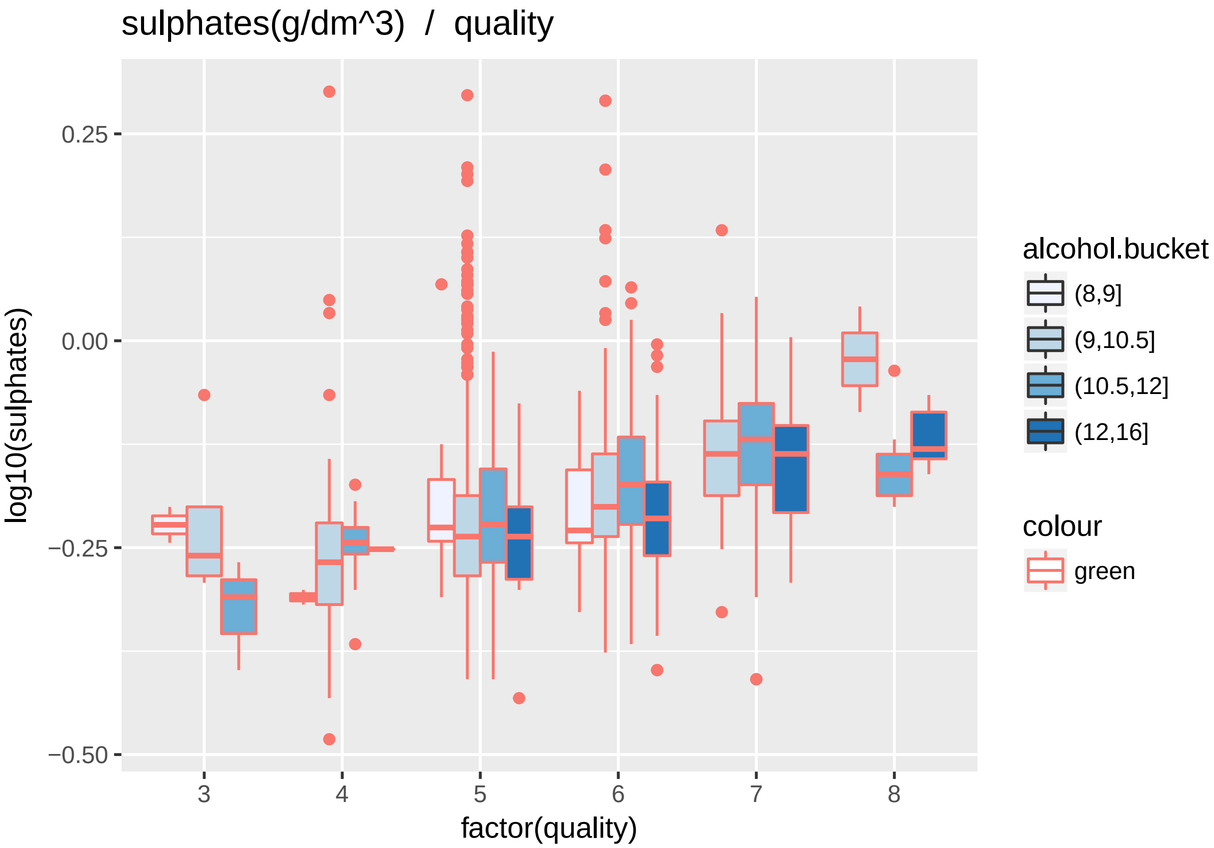
In the high quality wine both sulpahtes and the pH have positive relation. However, the pH have negative correlation for the other type of wines. This feature was observed earlier when we calculated the correlation value between the ph and the sulphates, which gave negative value.



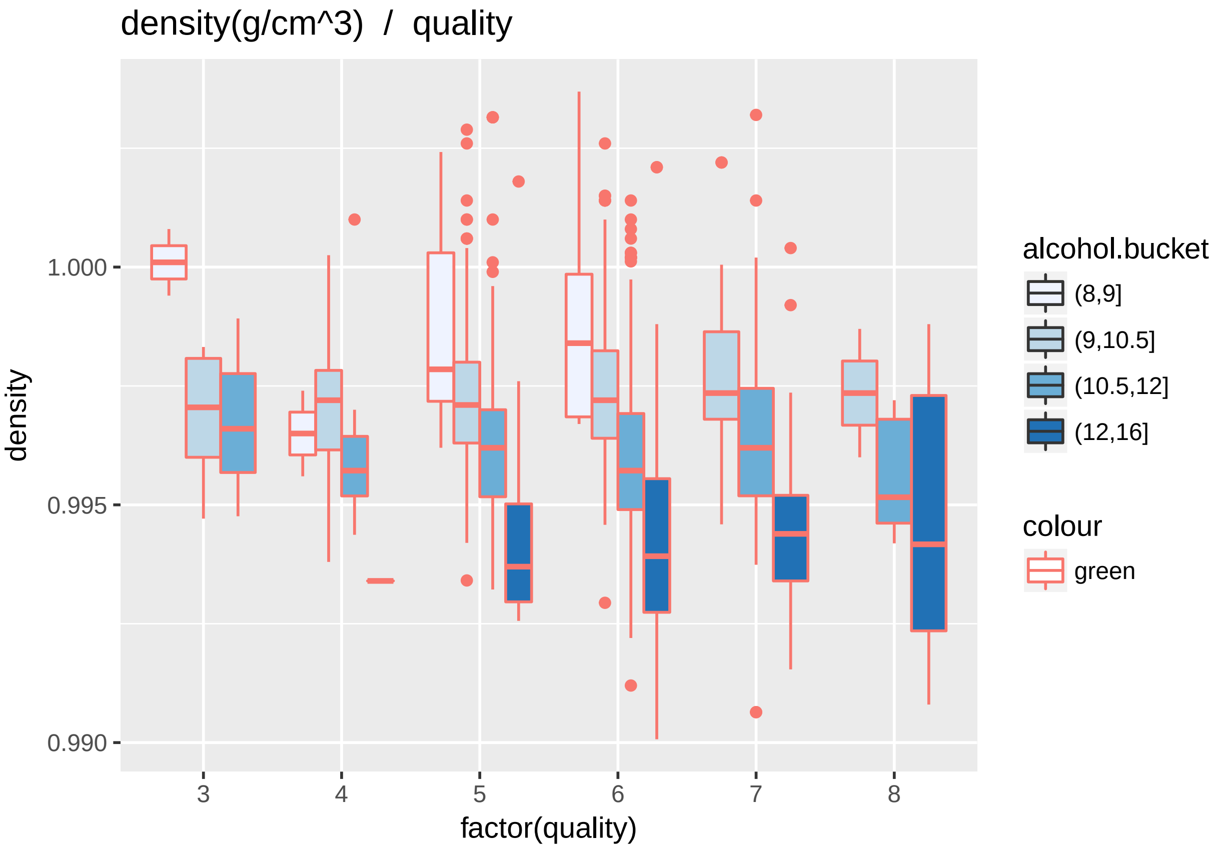
The volatile acidity decreases the quality of the wine. The wine with a high level of volatile acidity has less quality.



Both the citric acid have a positive effect on the wine quality.



Here we see the strong correlation between the alcohol content and the sulphates on the determination of the wine quality.



The high-quality wine have more alcohol and less density. So the latter decreases the wine quality.

```
##     cheap  standard excellent
##        63      1319       217
```

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

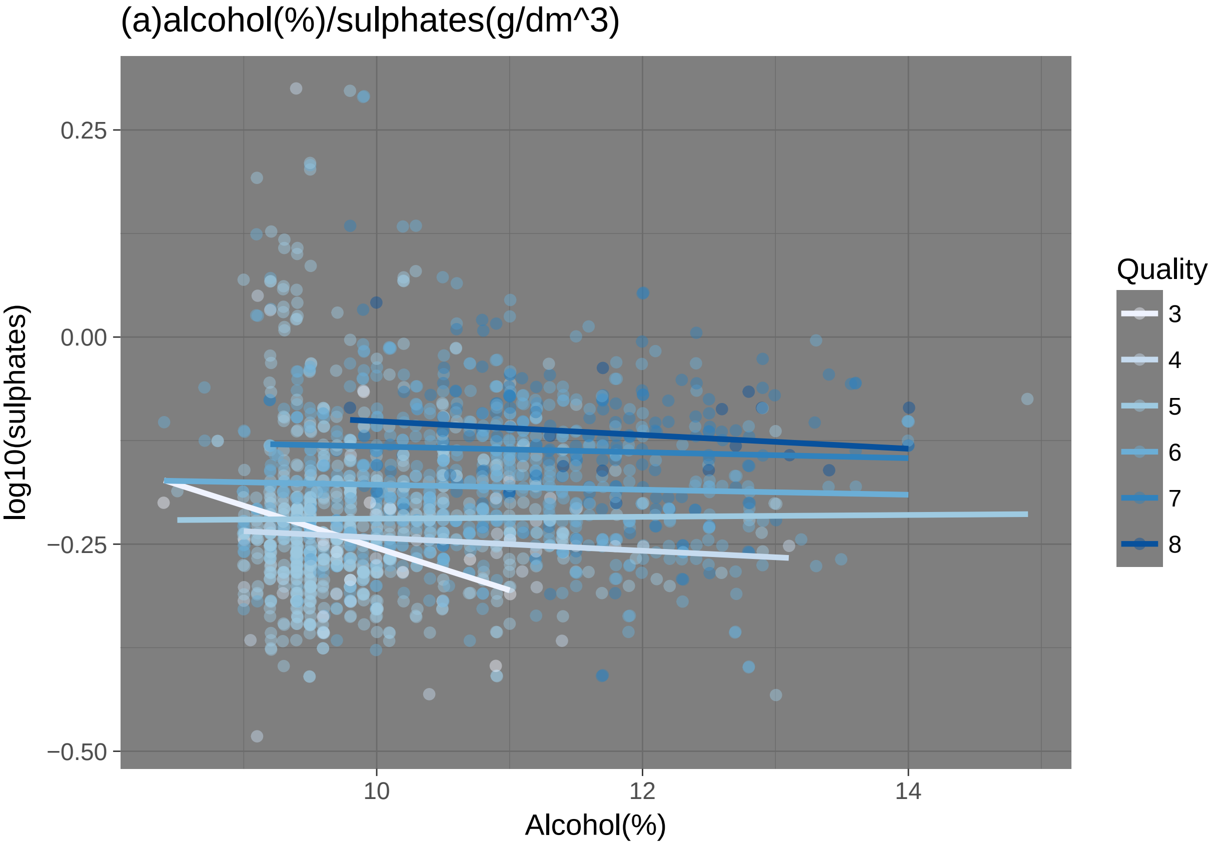
looking at your feature(s) of interest?

The multivariate plots simplify the investigation of the dataset. The sulphates tend to show a good positive effect on the determination of the quality of the wine. However, the alcohol contend dominated the contribution. The citric acidit also indicates a positive correlation with both the alcohol and the sulphates.

Were there any interesting or surprising interactions between features?

It was interesting to see that when pH decreases the value of the citric.acid and the quality of the wine. This feature was observed before but it becomes clear with many plots. Based on the chemistry, this is supposed to happen, and it was so interesting to see this phenomena and associate
it with the quality of the wine.

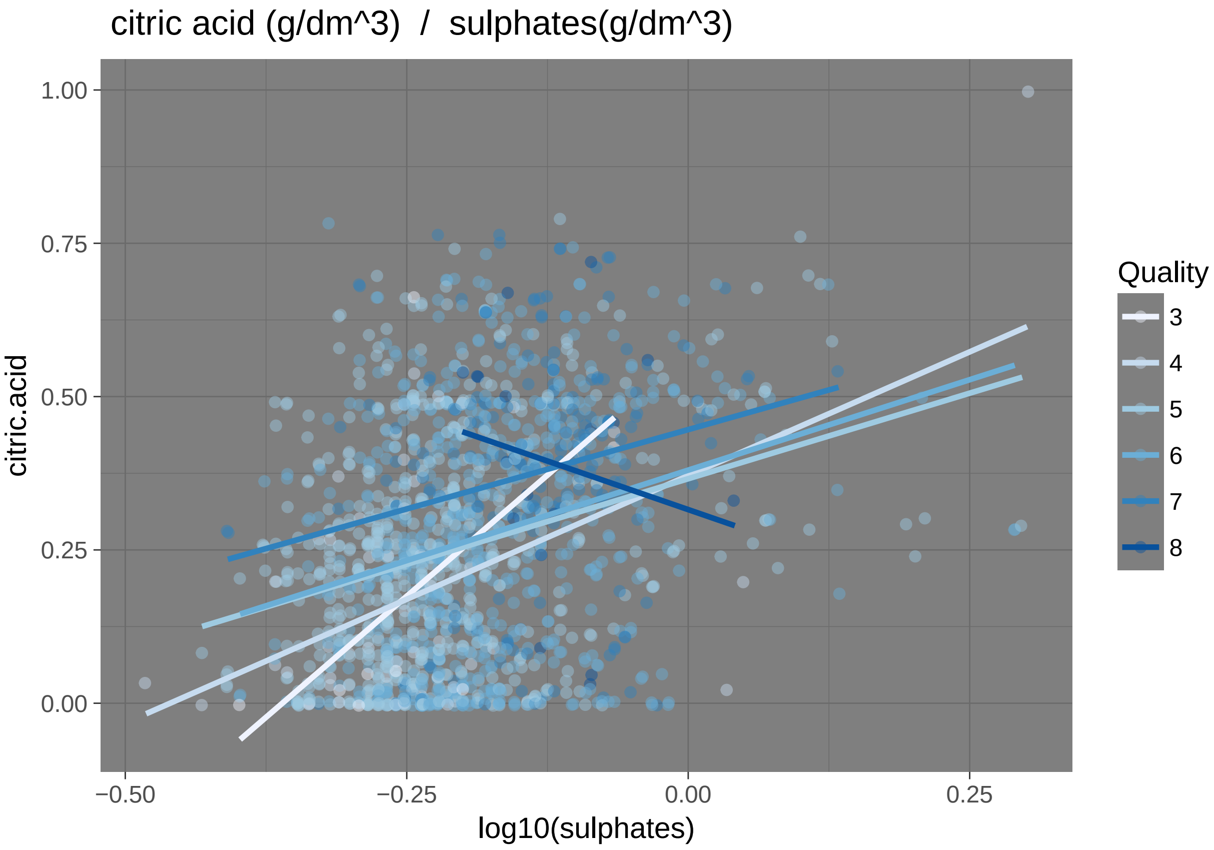
Final Plots and Summary



Description One

The graphic indicates that the quality increases with the increases of both alcohol and suphates concentration. In addition, the alcohol has more impact factor on the determination of the wine quality than the sulphates.

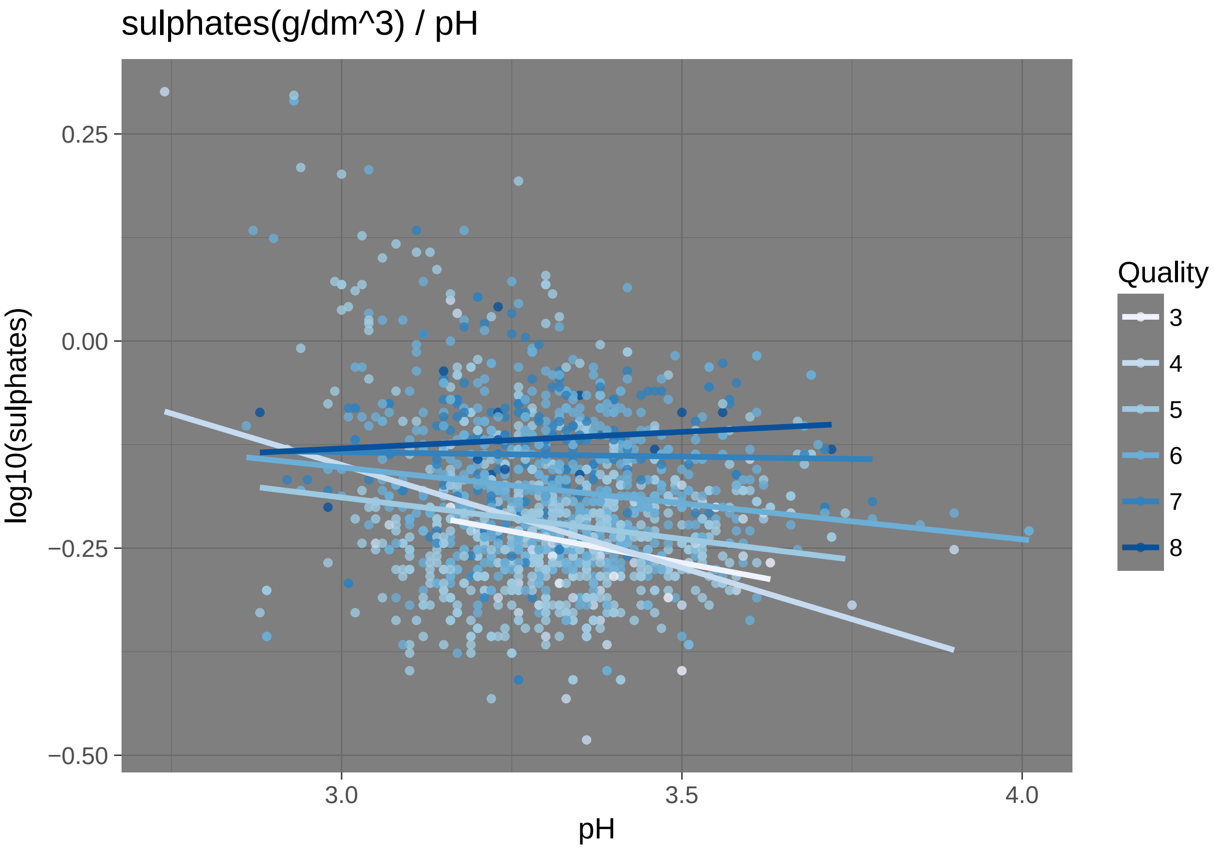
Plot Two



Description Two

Both the sulphates and the citric acid have apositive impact on the quality of the wine. In this plot we see that the sulphates contribute more than the citric acid on the quality of wine.

Plot Three



Description Three

The increase of the sulpahtes increases the quality of the wine. One interesting feature about this graphic is that the sulphates and
the pH have a negative correlation. But for the high quality wine we see that both sulphates and the pH have a graphic with a positve slop. So the pH has negative effect on the wine with lower wine quality.

Reflection

The purpose of this investigation is to study the properties which influence the quality of the wine. This investigation is based on the data set of 1599 observation and 11 chemical properties, used to determine the quality of the wine. The chemical material added to make wine have different composition and the quantity is also different. This makes the investigation
more interesting because we are comparing different quantities to come with one that has more impact on the wine quality. Our detailed analysis gives the following results.

(a) The quality is a discrete variable and is determined by each chemical product used for wine production. The pH and residual sugar, density do not have a dominant contribution on the determination on the quality of the wine.

(b) The alcohol content has a considerable contribution on the determination of the quality of the wine. It has a strong correlation with the sulphates and citric acid which also increases the quality. The less quality wine has little concentration of both sulphate and alcohol content. I visited a shop near my home and checked the price of the wine. The more wine expensive wine had high alcohol %, whicc is around 11% of the alcohol. I also concluded that beer has an average 6% of the alcohol while most the wine are around 11% af the alcohol. This statement supports that more alcohol concentration will give high quality of the wine.

(c) Not all acids have good correlation with the quality of the wine. For example, the citric.acid had a good impact on the determination of the quality of the wine. On the other hand, when the concentration of the
volatile acid decreases the quality of the wine decreases.

It was interesting to see that the pH have a negative correlation of the suphates. However, for the high quality wine both pH and the sulphates had a positive correlation. The alcohol on the hand, have a positive correlation with the pH. While, the citric acid have negative correlation with pH.

The density, volatile acidity have a nagaive effect on the determination of the quality of the wine. The mean value of the residual sugar is 2.539 with 2.2 median. This indicates that majority of the wines do not have high quantity of the residual sugar.

From this investigation, we have concluded that the alcohol, sulphates and the citric acid have dominant impact on determination of the quality of the wine. However, we do not know how much this parameters determine the price of the alcohol compared to other.

One parameter that would be incorporated in the study is the price of each substance used in the production of the wine. this would give a better understanding of the investigation of the alcohol quality and the price of the final product. If we have a certain quantity that is inserted with the corresponding price in the dataset, it would be interesting to see how much this chemical product effect both the quality and the price.

References

https://www.rstudio.com