RED WINE EXPLORATION by PIERRE CONREAUX

Introduction

Our goal is to get a better understanding of which chemical properties influence ***the quality of red wines.***

We will use a public dataset available which come from: *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.*

For more information visit the following link: https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt

First, we load ***the necessary packages***:

Then, we ***load the data:***

To confirm the loading, we ***read the head*** of the data:

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

Univariate Plots Section

Before plotting, we need to understand our data sets and its variables.

After a quick description and a summary, we will plot each variables.

Data set

```
## [1] 1599   12
```

This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine, and 1 variable on the quality of the wine.

Regarding the quality variable, at least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

Noticed that our data is already tidy, so we do not have to clean it.

Variables

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

Description of variables:

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3 - citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines

4 - residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5 - chlorides: the amount of salt in the wine

6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7 - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content

9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

11 - alcohol: the percent alcohol content of the wine

Output variable (based on sensory data): 12 - quality (score between 0 and 10)

data summary

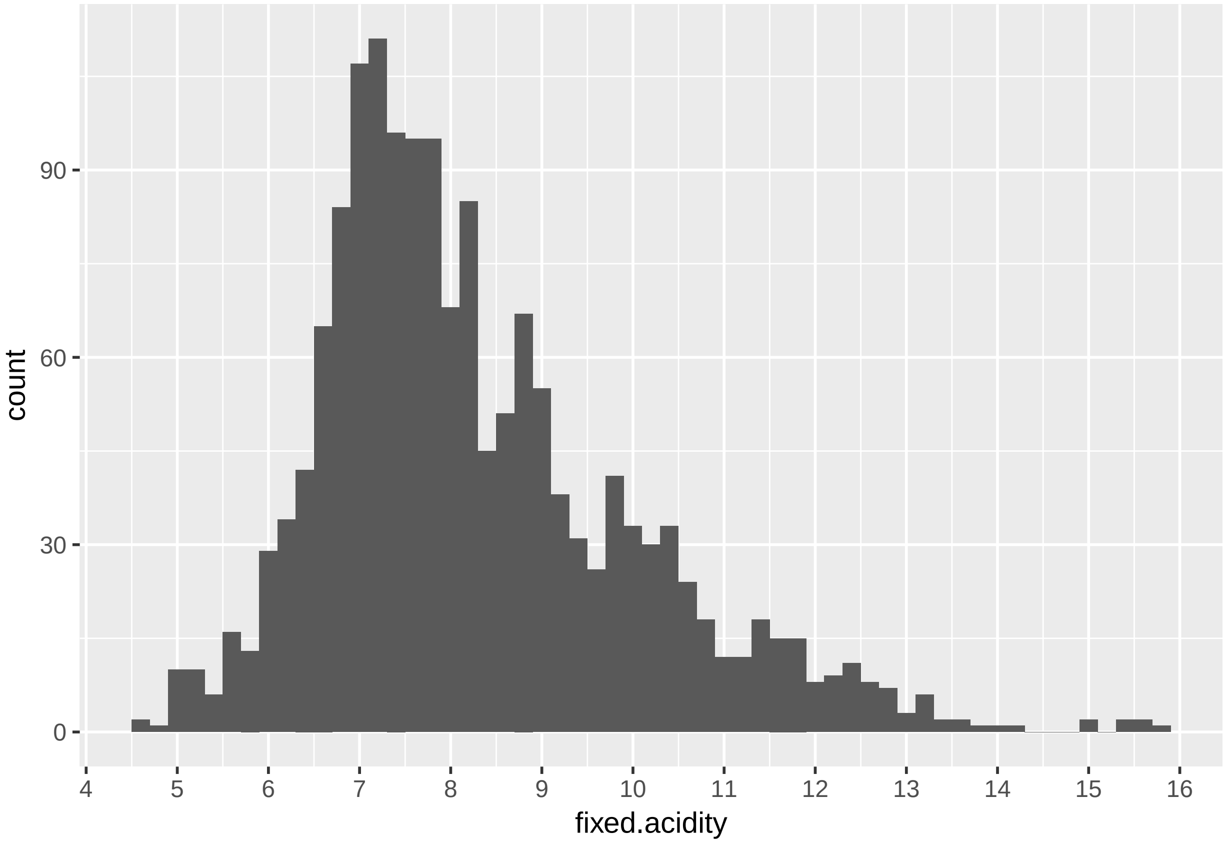
```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : 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 ...
```

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

Plots

We start by create a ***function to create histogram***:

Fixed Acidity

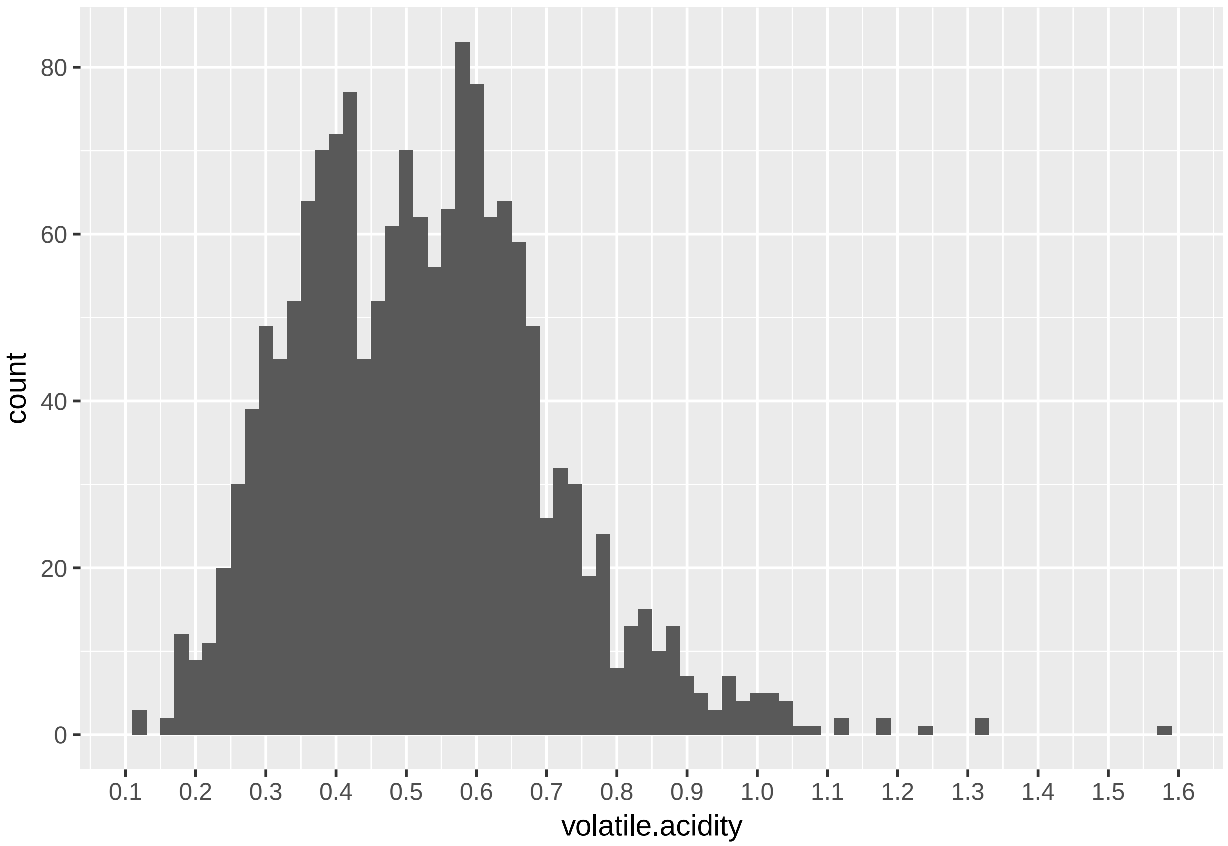


The graph depicts ***the distribution of fixed acidity.***

We can see that the distribution is ***right skewed.***

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

Volatile Acidity

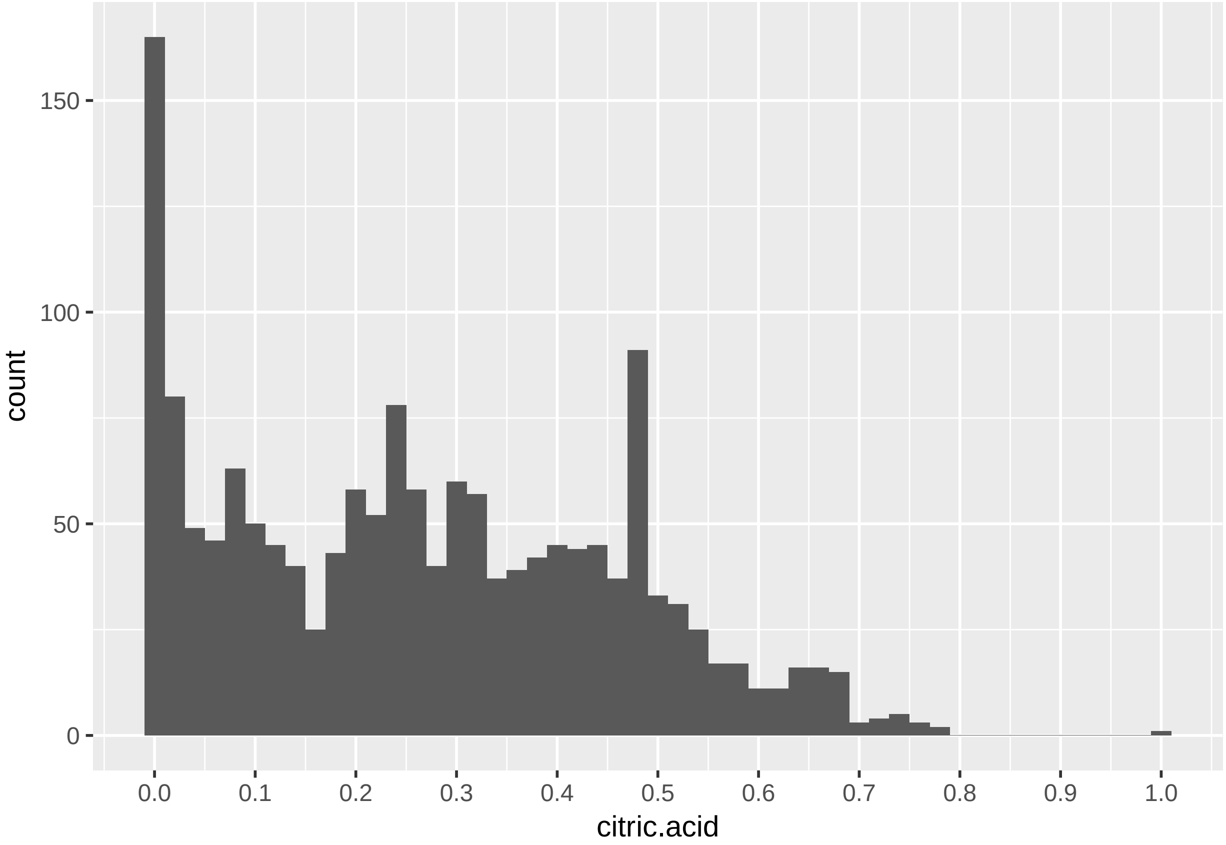


The graph depicts ***the distribution of volatile acidity.***

We can see that the distribution is approximately ***normally skewed.***

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

Citric acid

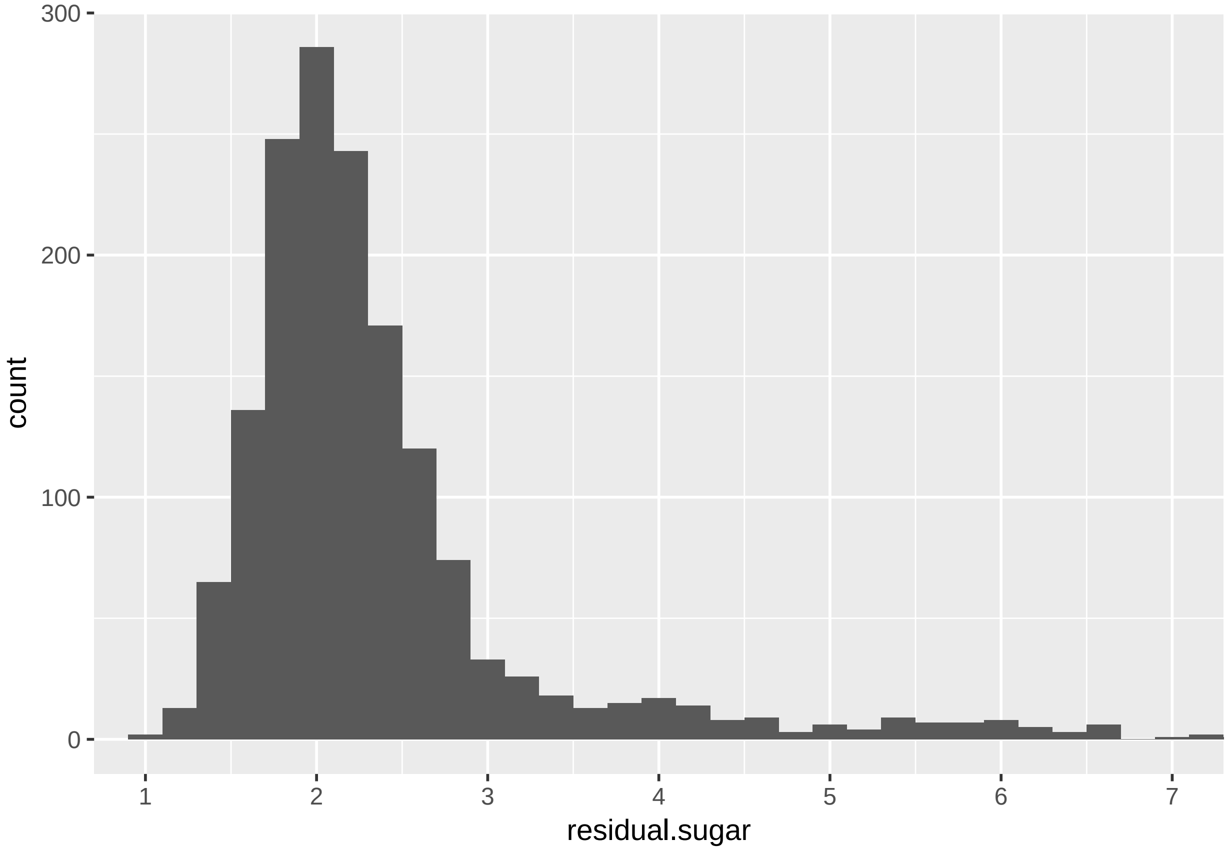


The graph depicts ***the distribution of citric acid.***

We can see that the distribution is ***multi-modal.***

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

Residual Sugar

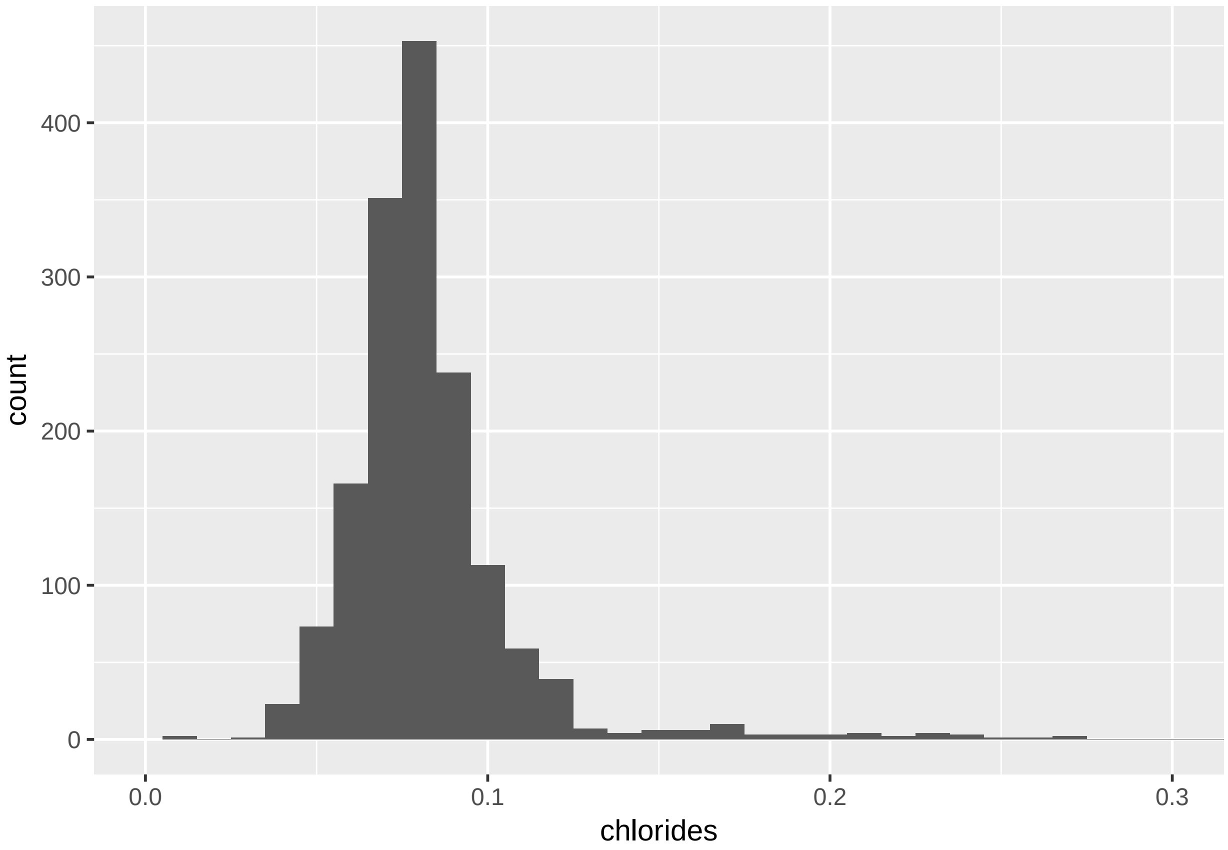


The graph depicts ***the distribution of residual sugar.***

We can see that the distribution is approximately ***right skewed.***

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

Chlorides

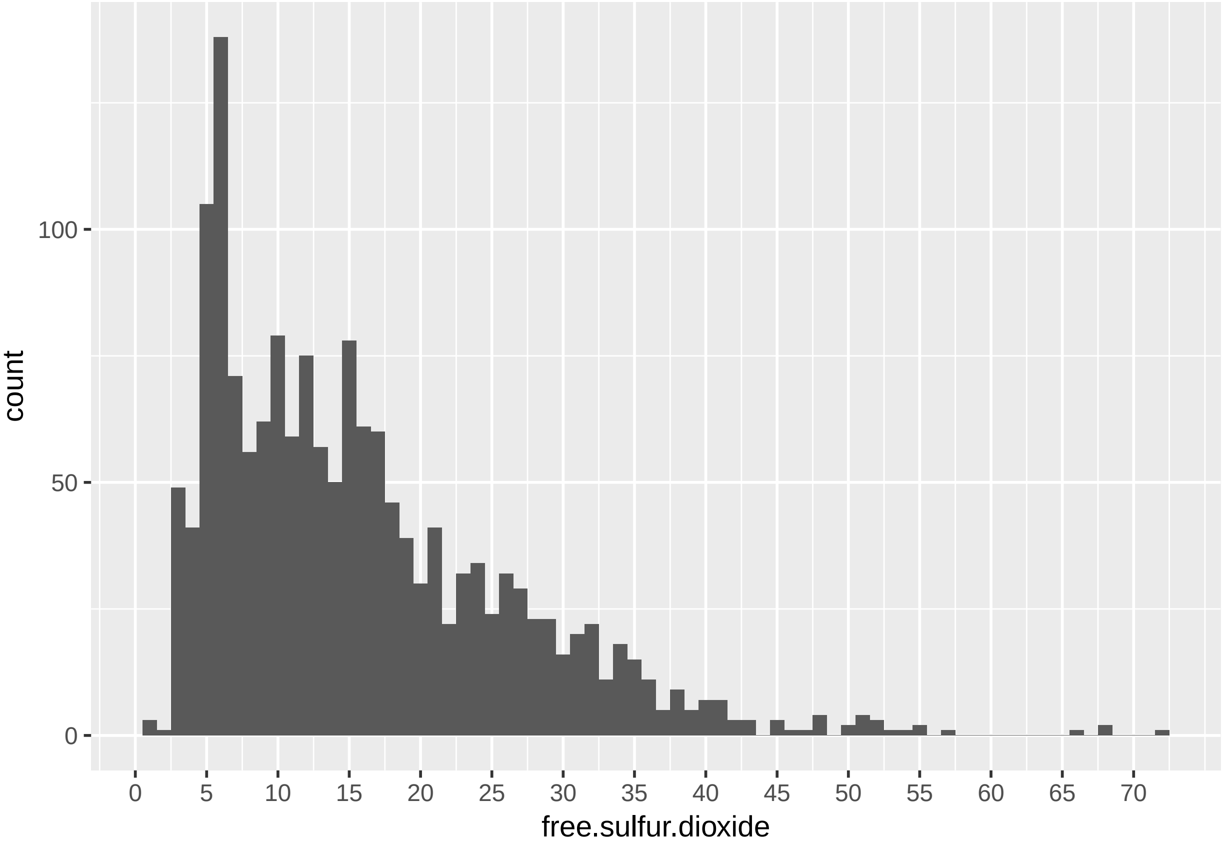


The graph depicts ***the distribution of chlorides.***

We can see that the distribution is ***normal.***

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

Free sulfur dioxide

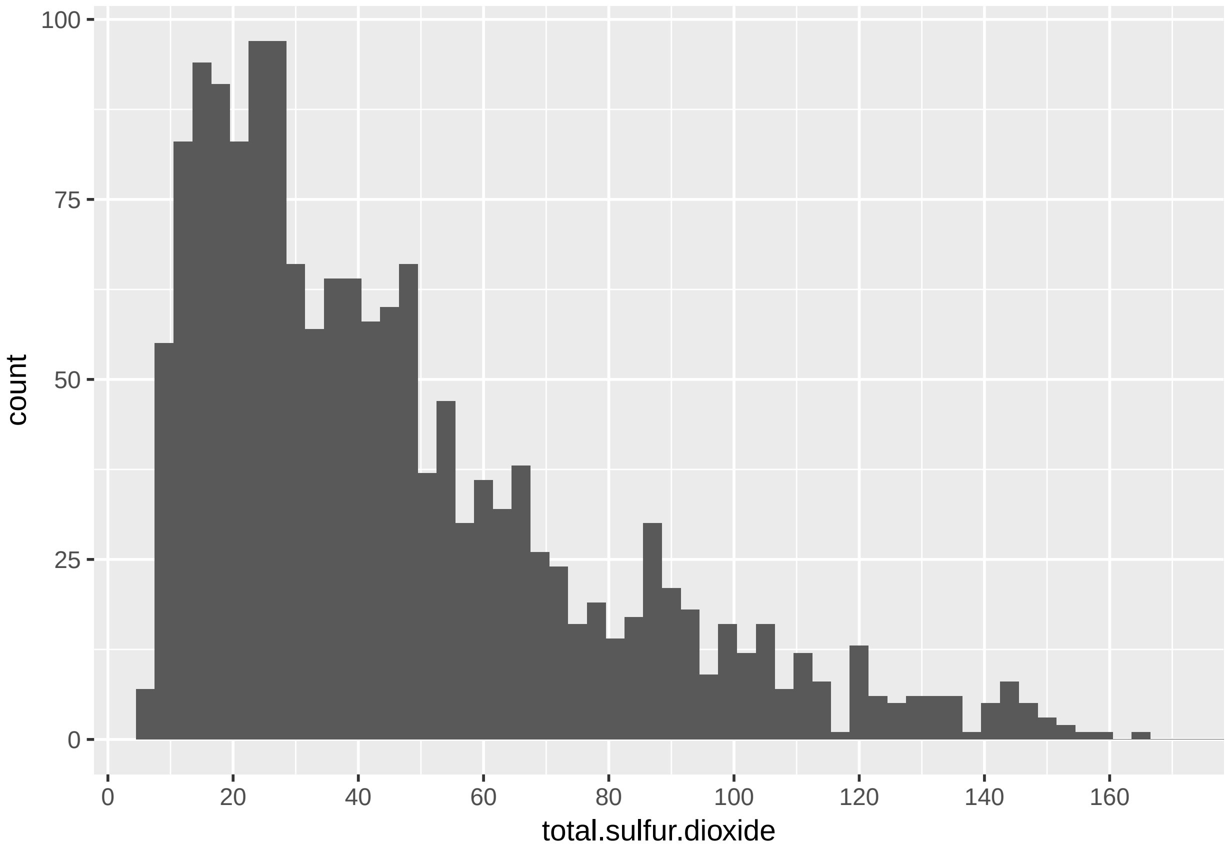


The graph depicts ***the distribution of free sulfur dioxide.***

We can see that the distribution is ***right skewed.***

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

Total sulfur dioxide

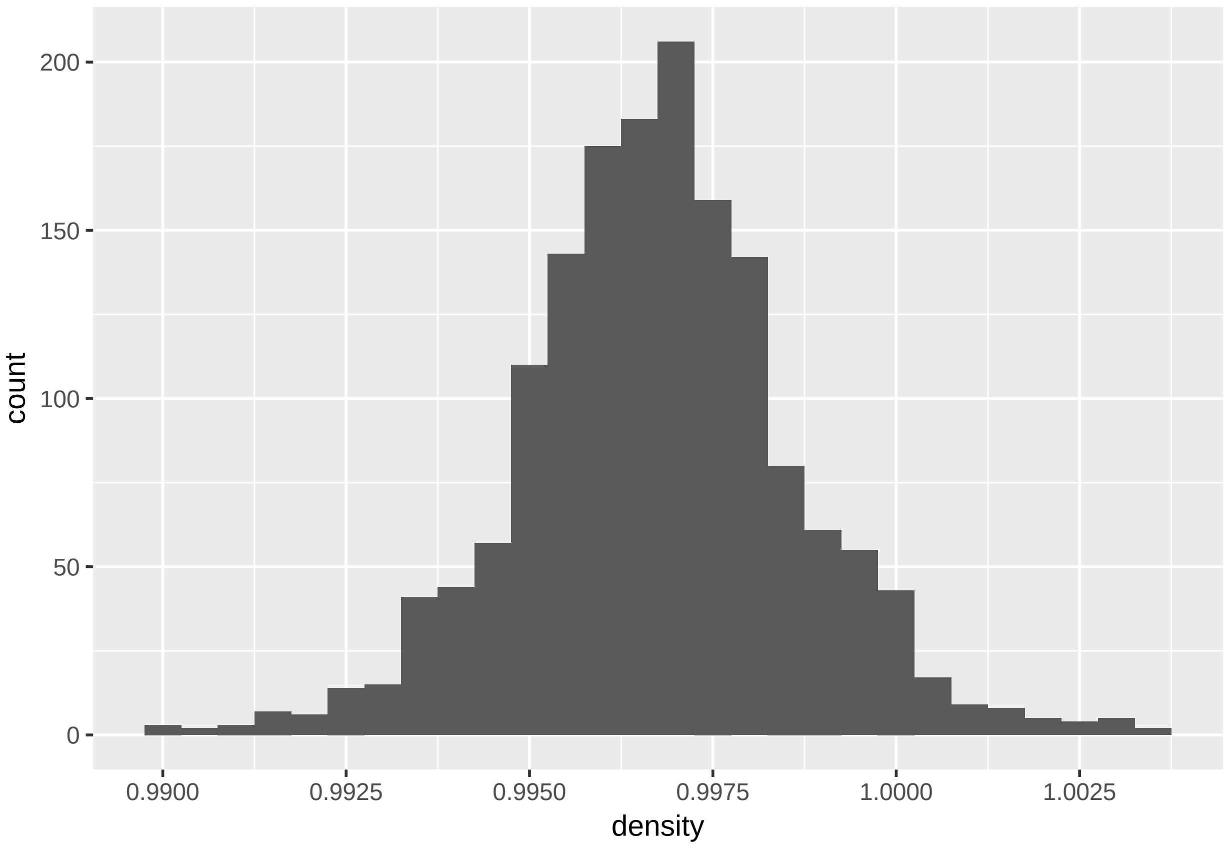


The graph depicts ***the distribution of total sulfur dioxide.***

We can see that the distribution is ***right skewed.***

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

Density

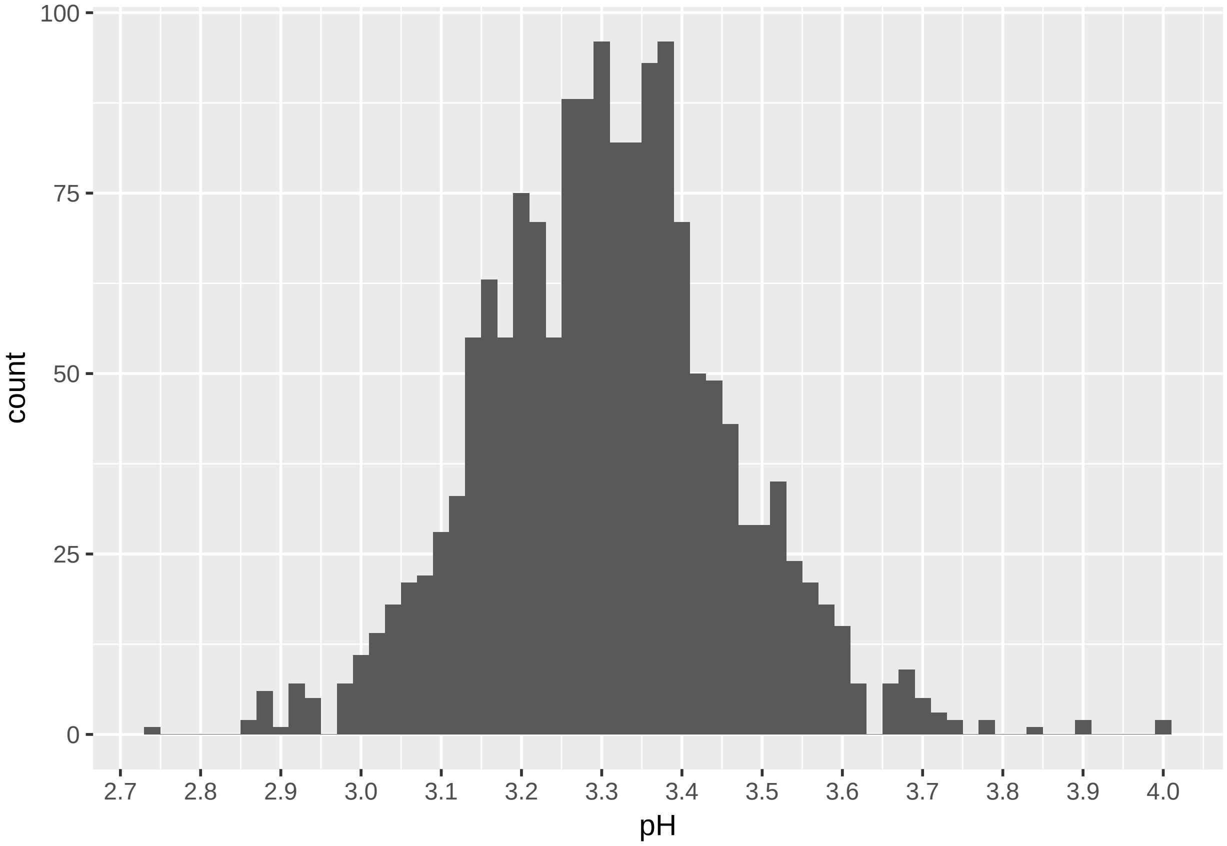


The graph depicts ***the distribution of density.***

We can see that the distribution is ***normal.***

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

pH

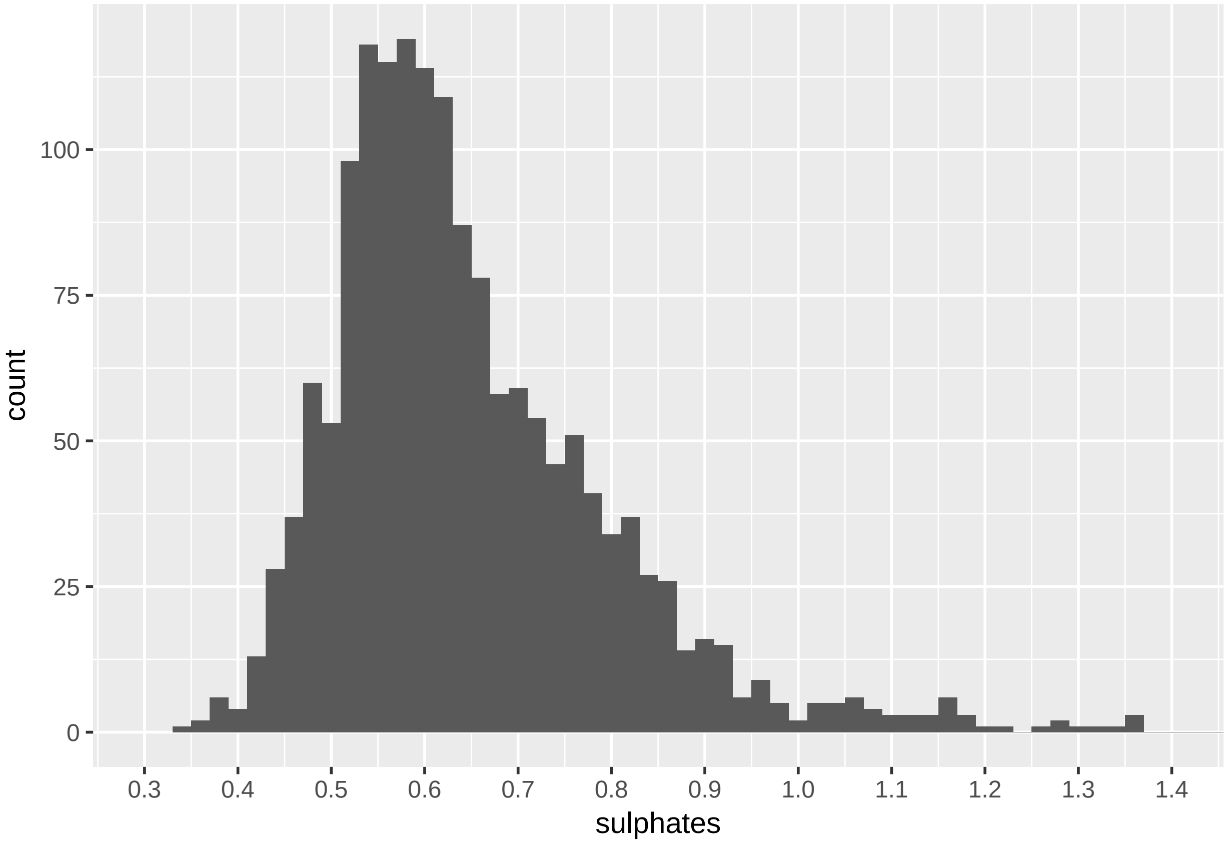


The graph depicts ***the distribution of pH.***

We can see that the distribution is ***normal.***

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

Sulphates

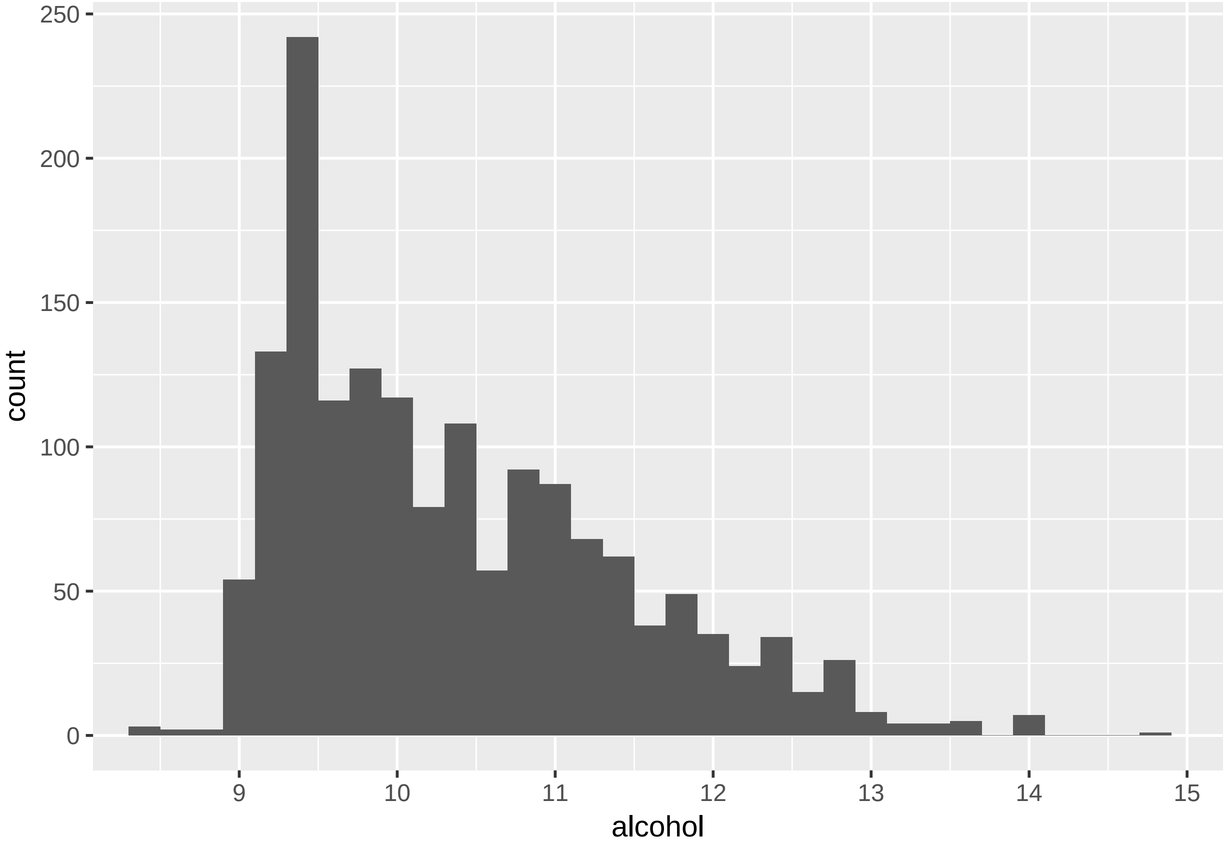


The graph depicts ***the distribution of sulphates.***

We can see that the distribution is ***right skewed.***

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

Alcohol

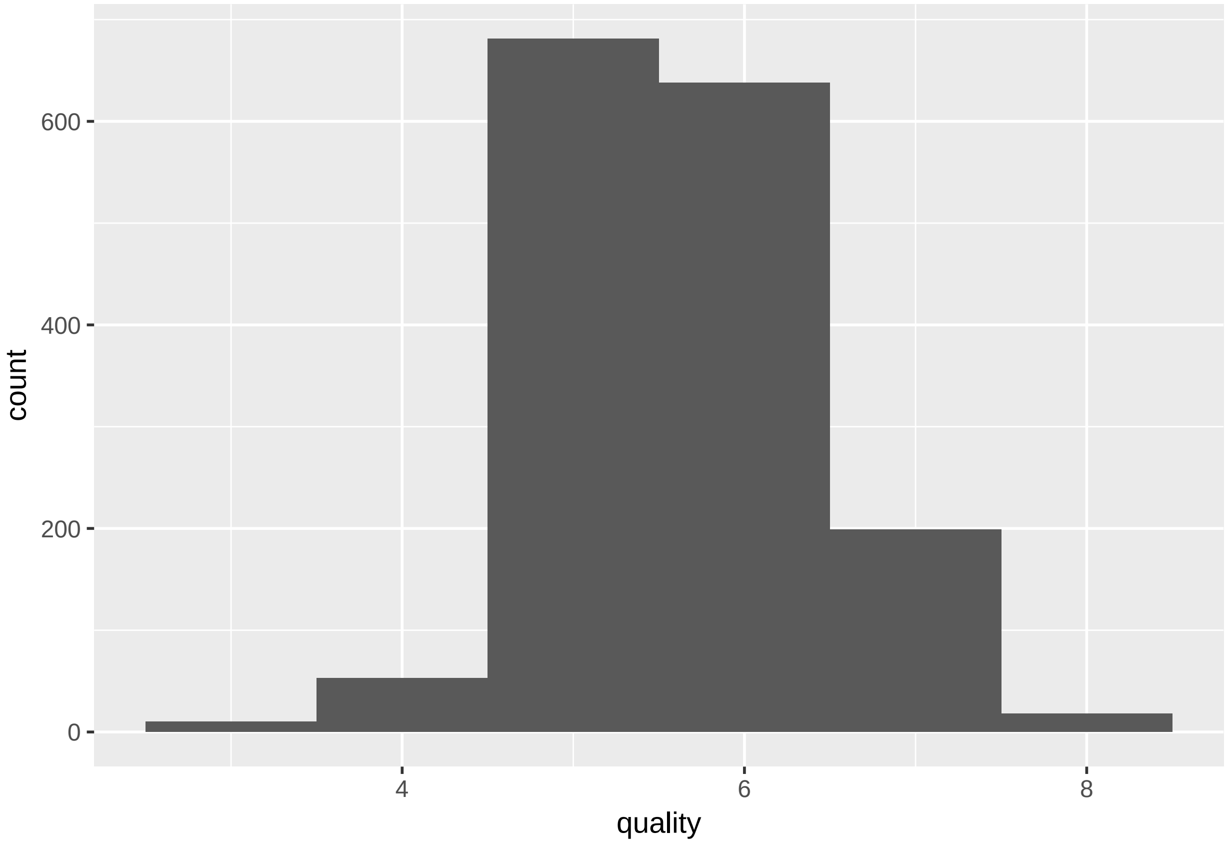


The graph depicts ***the distribution of alcohol.***

We can see that the distribution is ***right skewed.***

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

Quality



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

The graph depicts ***the distribution of quality.***

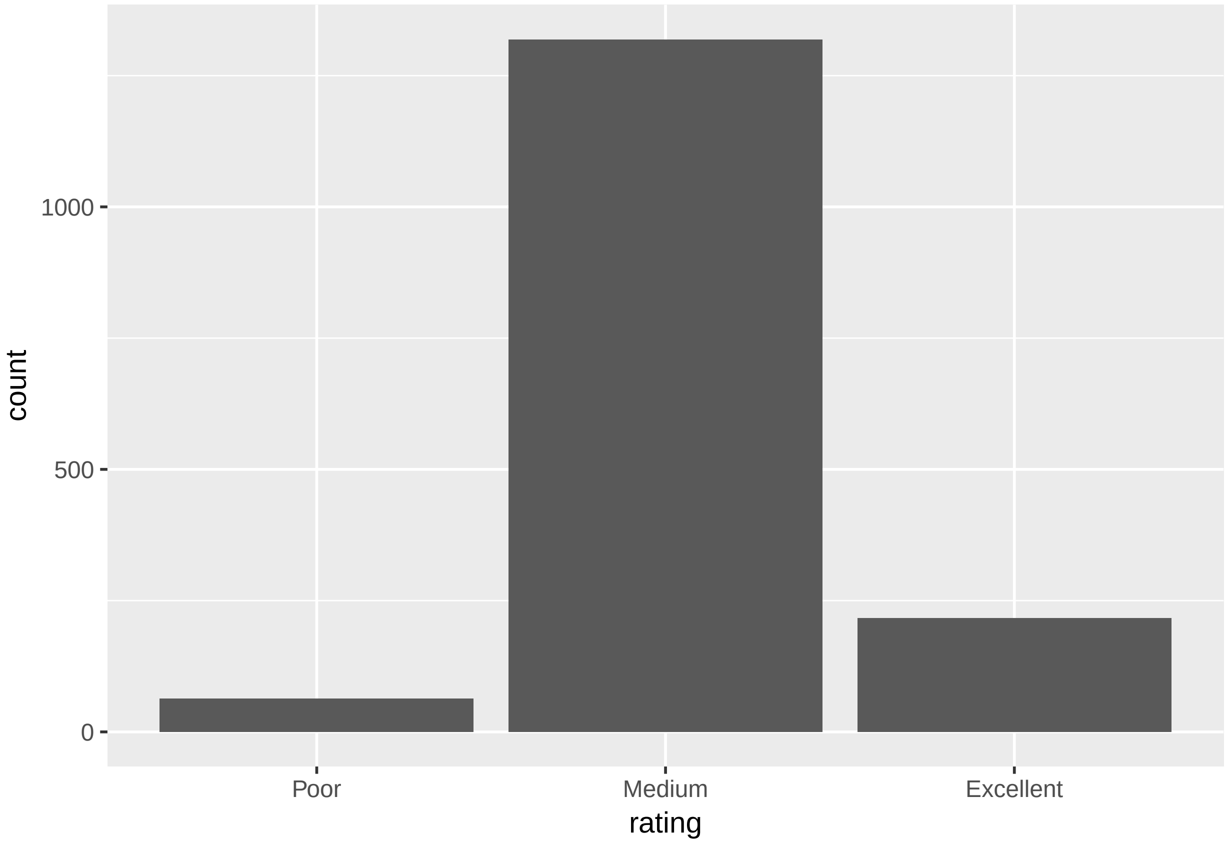
We can see that the distribution is ***normal.***

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

Rating

To ensure a better visibility in multivariate section, we create ***rating, a discrete variable***, to gather quality by levels.

To plot a discrete variable, as rating, we need to modify the histogram code.



```
##      Poor    Medium Excellent
##        63      1319       217
```

Univariate Analysis

What is the structure of your dataset?

This data set contains ***1,599 red wines with 13 columns.***

11 numeric variables on the chemical properties of the wine and ***2 quality variable.***

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

The main feature is the ***quality*** of the wine.

What other features in the dataset do you think will help support your

investigation into your feature(s) of interest?

Each numeric variables represent a charateristic of the wine, so ***each variables have to be investigating***.

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

We created a rating variable that we will use in multivariate plots section.

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?

The data set is already tidy, so we not have to perform any operations to tidy it.

Bivariate Plots Section

In this section, we will start by analysis the relation between quality and the other variables.

After, we will study interaction between features.

quality

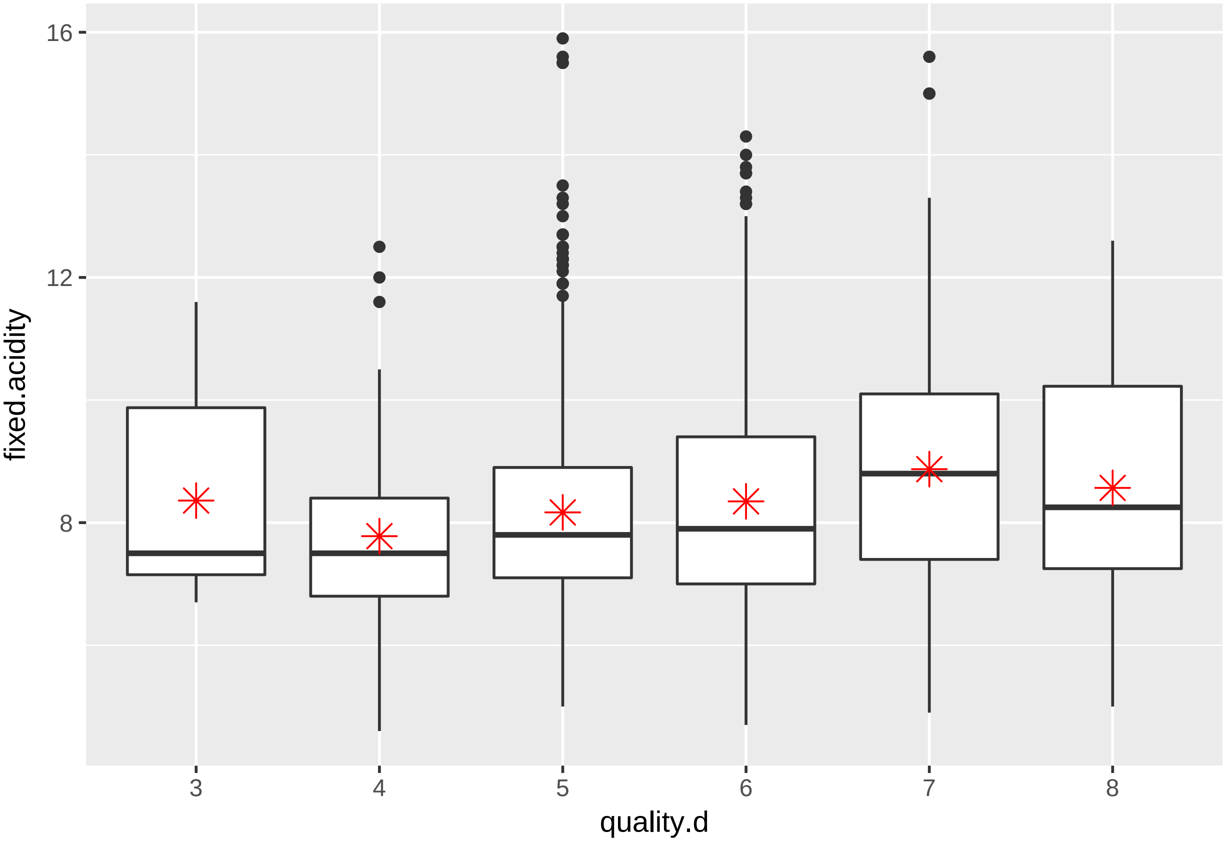
In order to be able to create boxplots we need to make ***a discrete variable for quality.***

We will do this in using ***cut:***

Now we can create a ***function to make boxplots.***

Finally, we can ***combine our functions.***

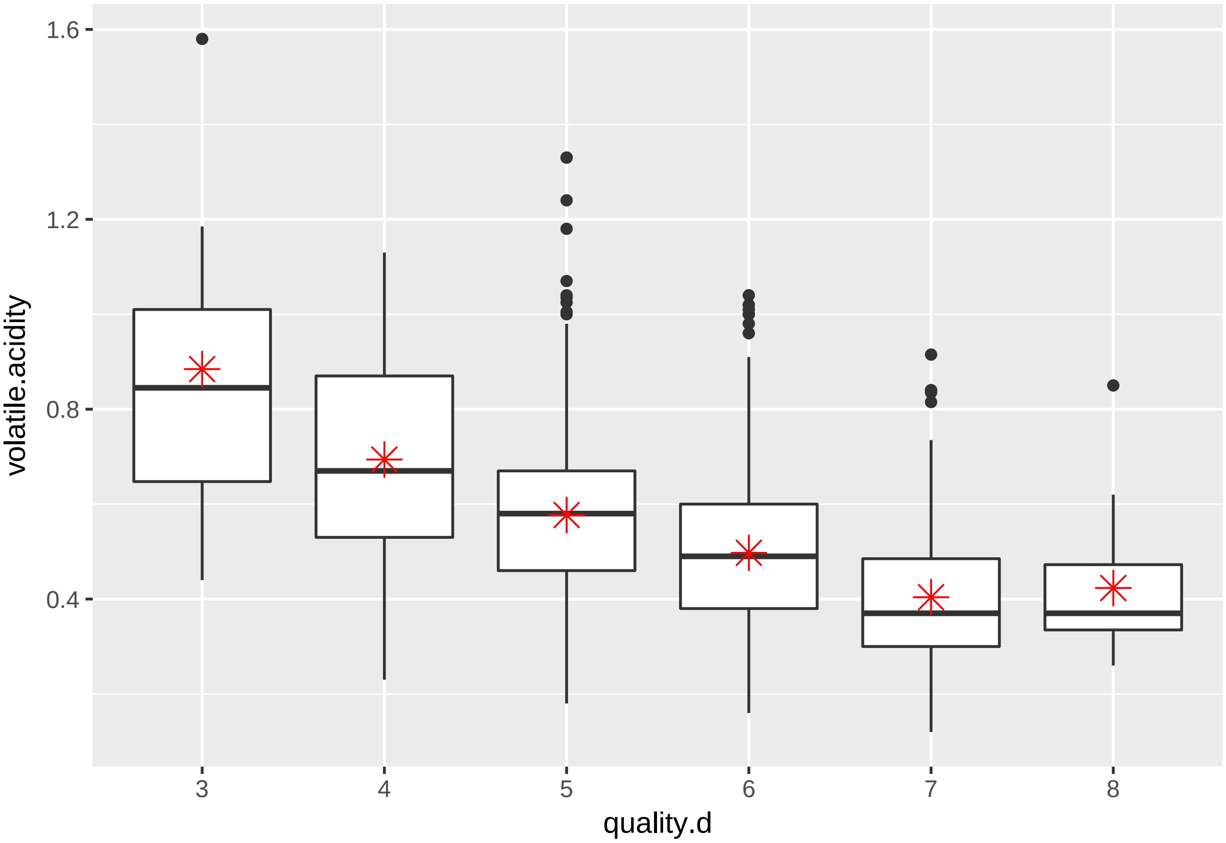
Quality vs fixed.acidity



The median level of fixed acidity seems close between the quality rating. We can notice a sligh rise between 6 and 7.

```
## [1] 0.1240516
```

Quality vs volatile.acidity

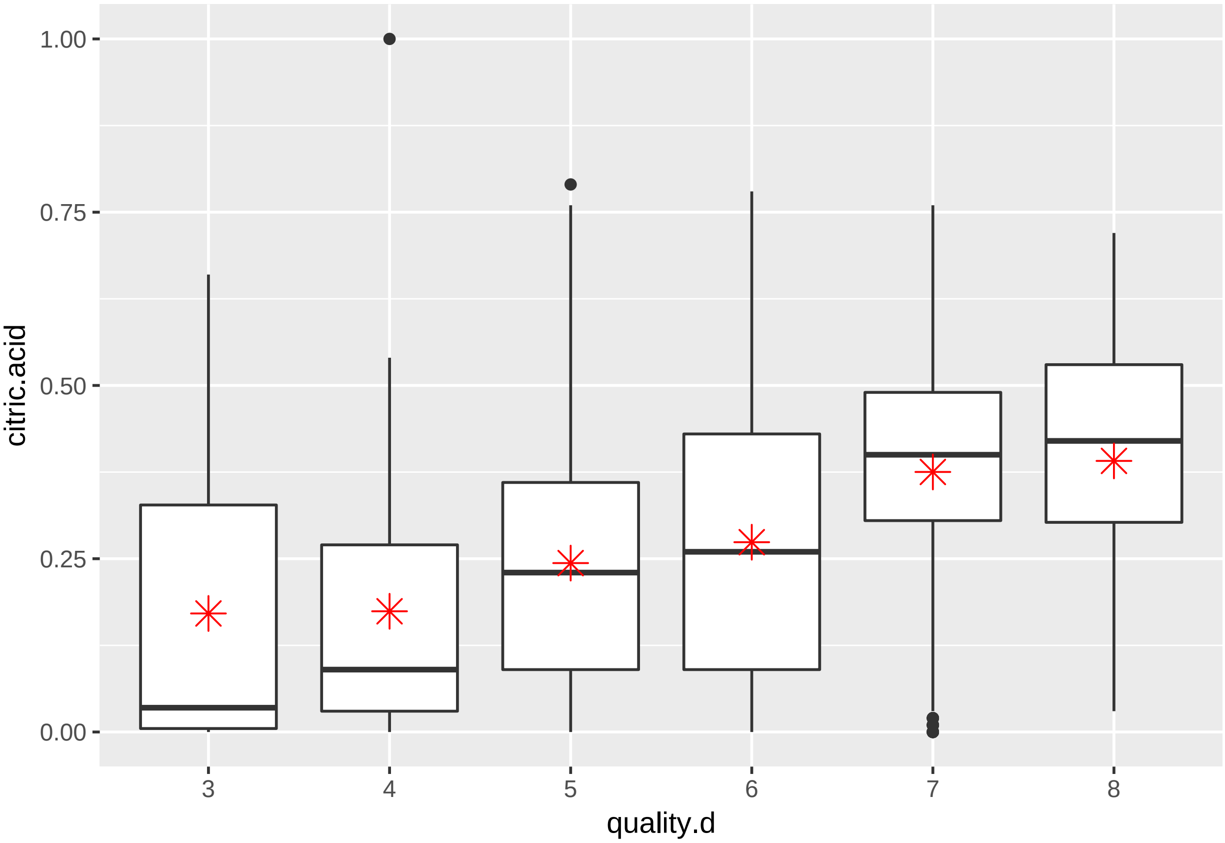


The volatile acidity is negatively correlated to the quality.

We can see a stability of the level of volatile acidity between 7 and 8, nevertheless this is not significant, because a small number of wines have a rating quality of 8.

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

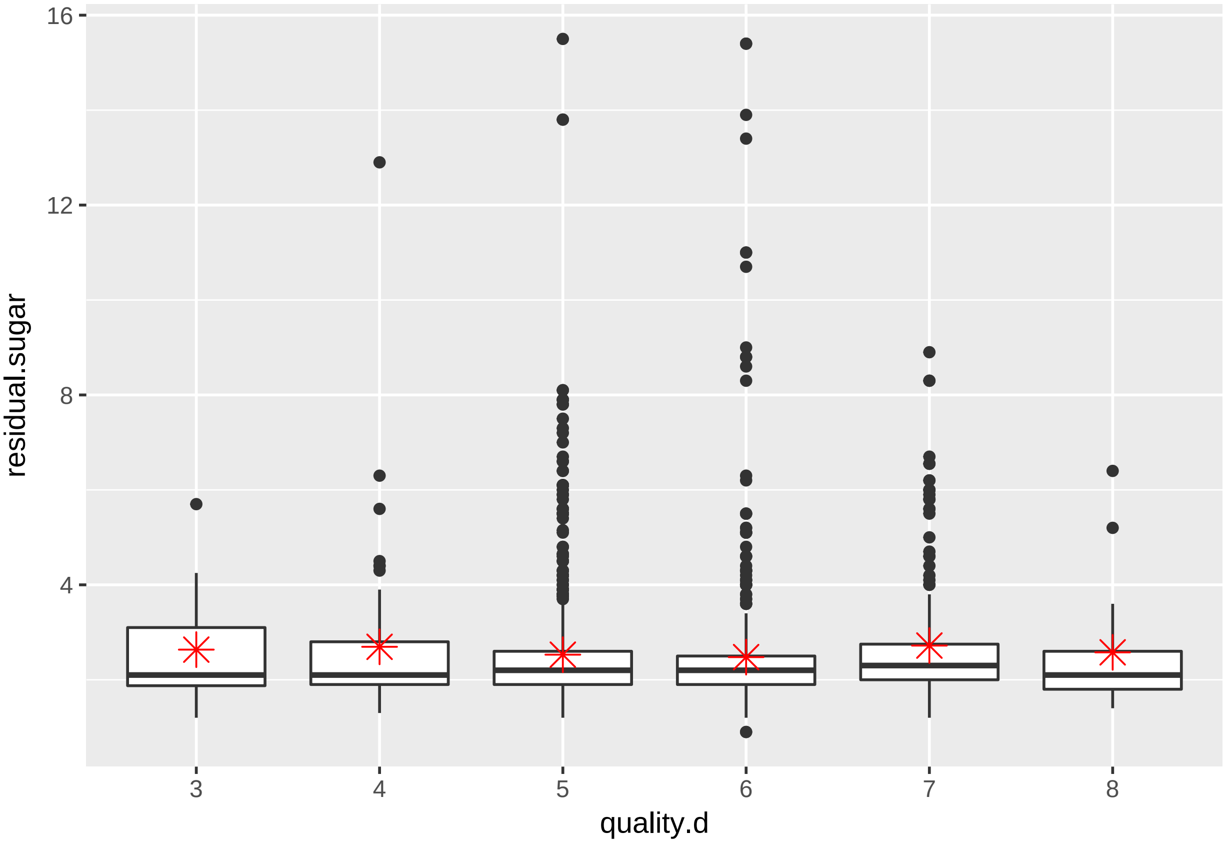
Quality vs citric.acid



The citric acid is positively correlated with quality.

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

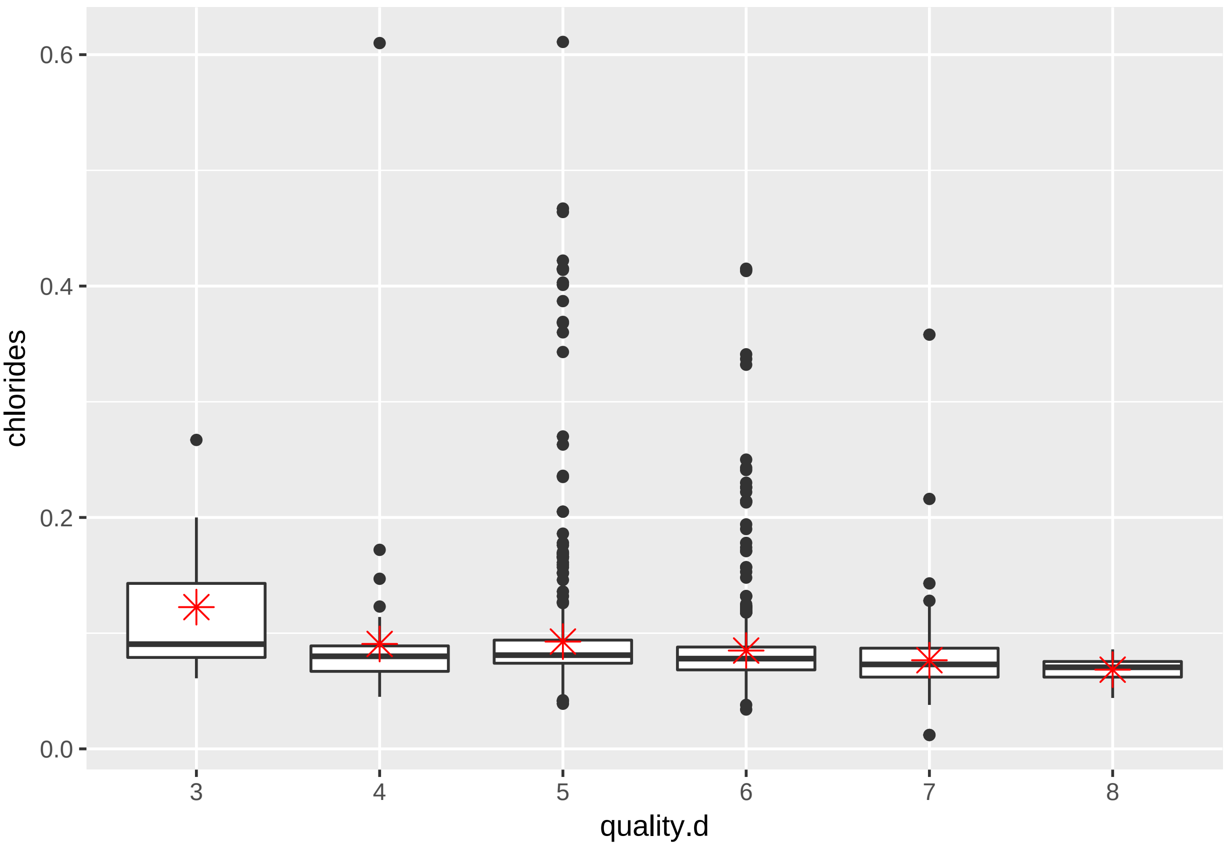
Quality vs residual sugar



The residul sugar seems weakly correalated to quality.

```
## [1] 0.01373164
```

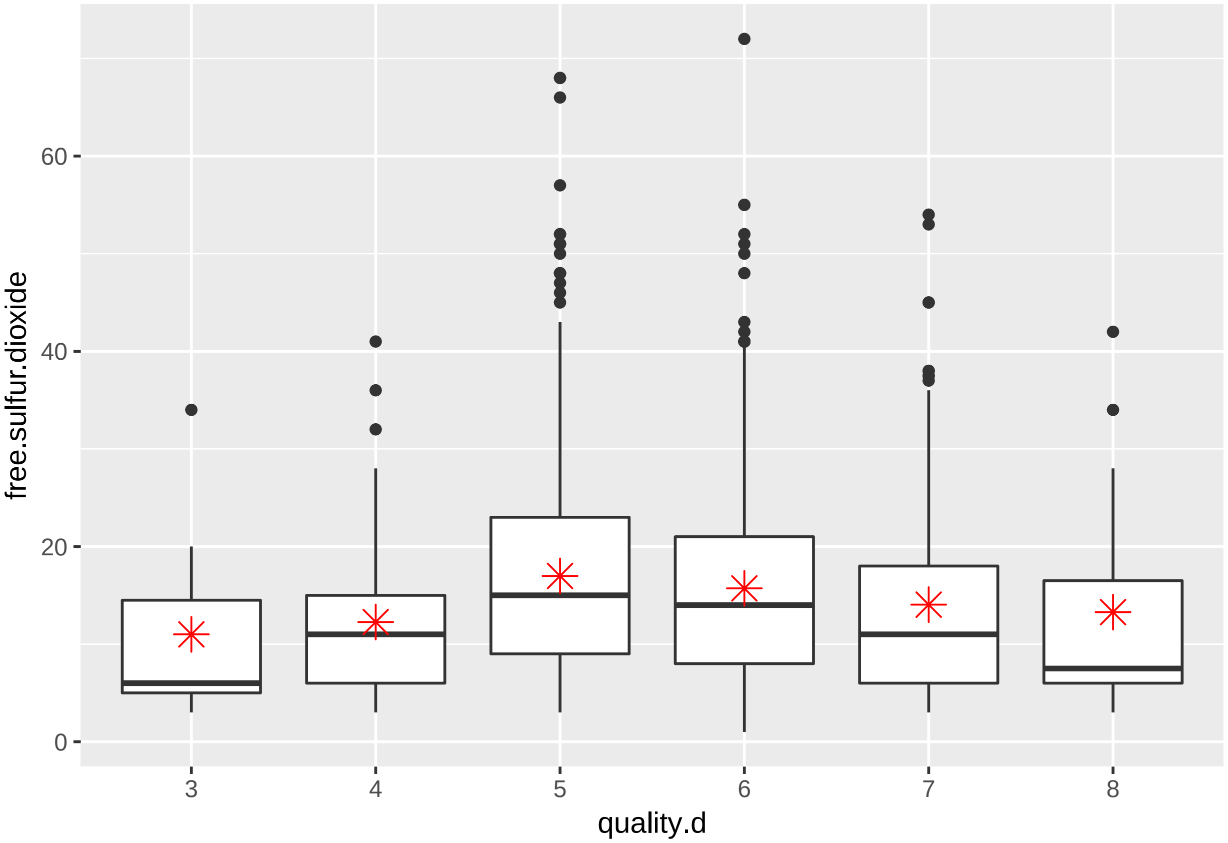
Quality vs chlorides



The chlorides seem weakly correalated to quality.

```
## [1] -0.1289066
```

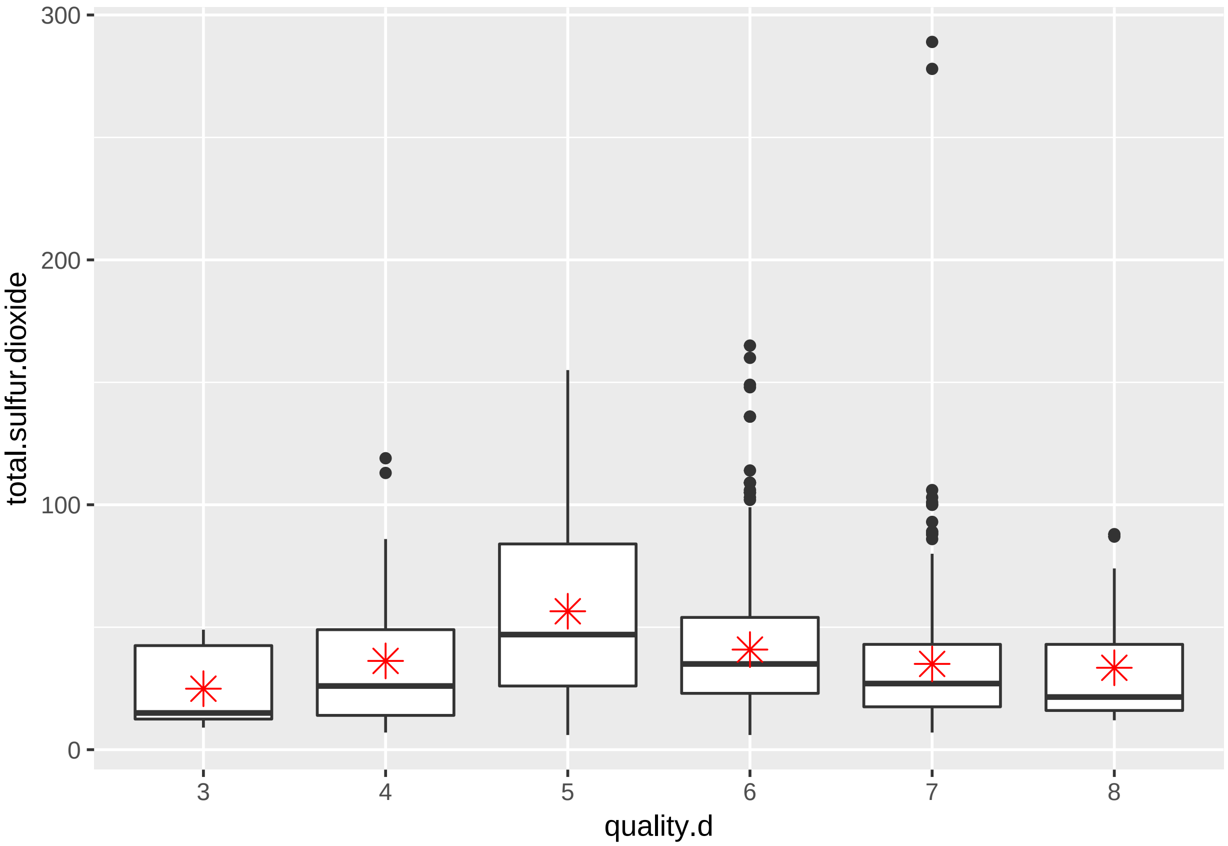
Quality vs free.sulfur.dioxide



Free sulfur dioxide and quality seem positively correlated between 3 and 5, then negatively correlated.

```
## [1] -0.05065606
```

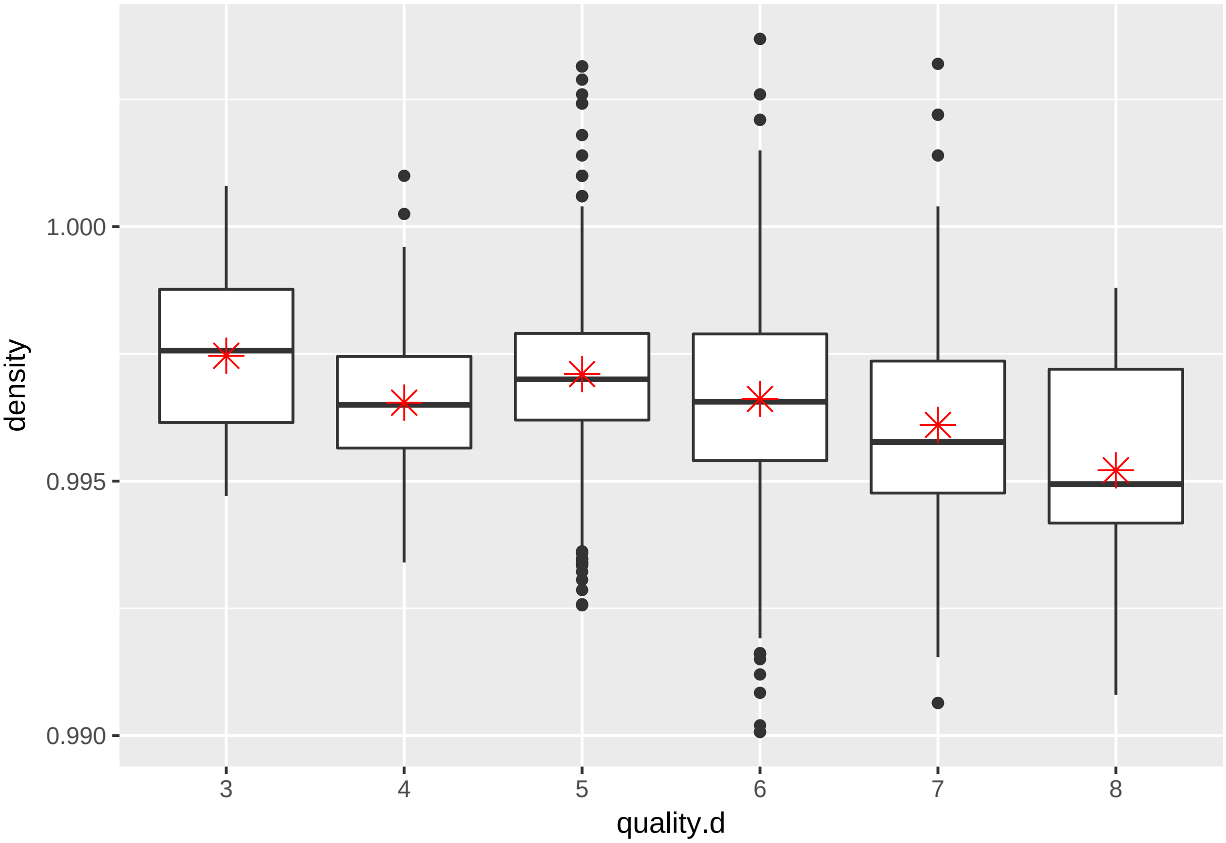
Quality vs total.sulfur.dioxide



Total sulfur dioxide and quality seem positively correlated between 3 and 5, then negatively correlated.

```
## [1] -0.1851003
```

Quality vs density

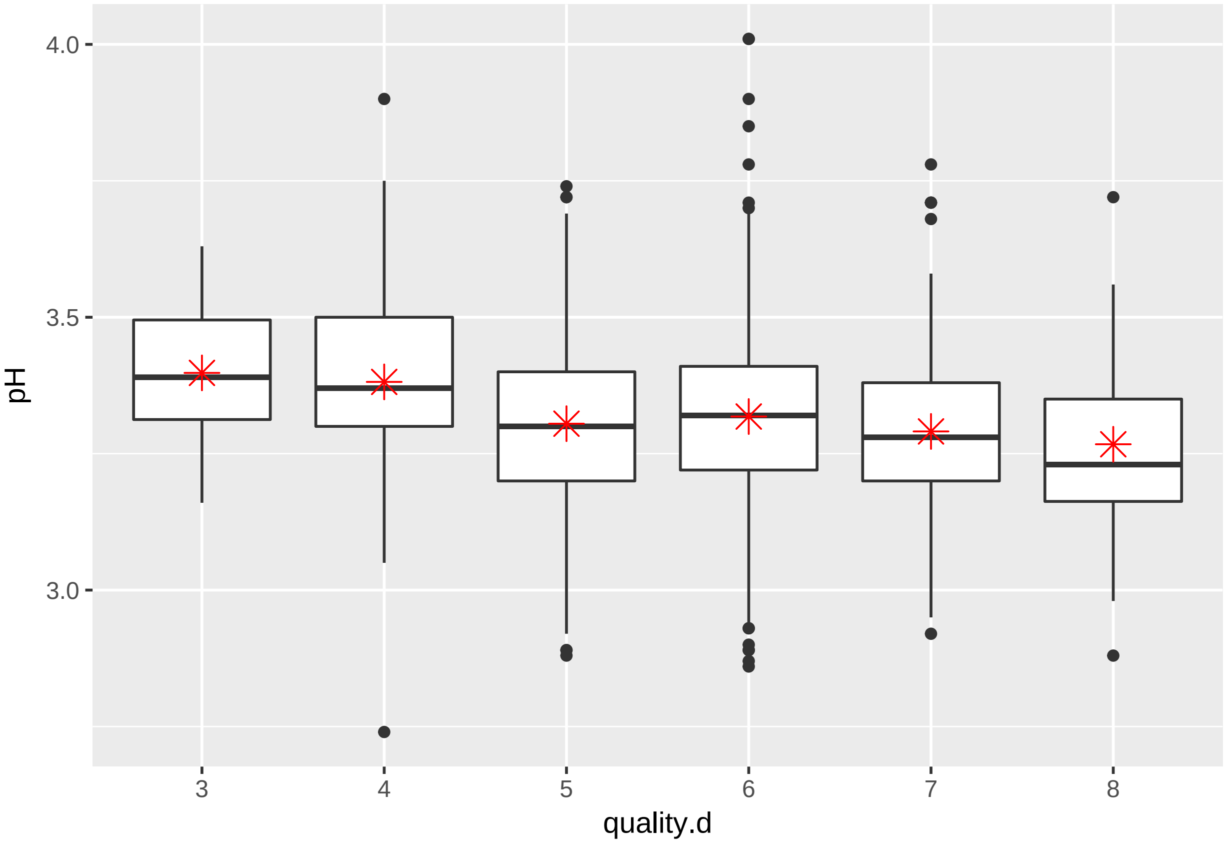


Density seems to be negatively correlated with quality.

Nevertheless, we can notice a slight increase between 4 and 5.

```
## [1] -0.1749192
```

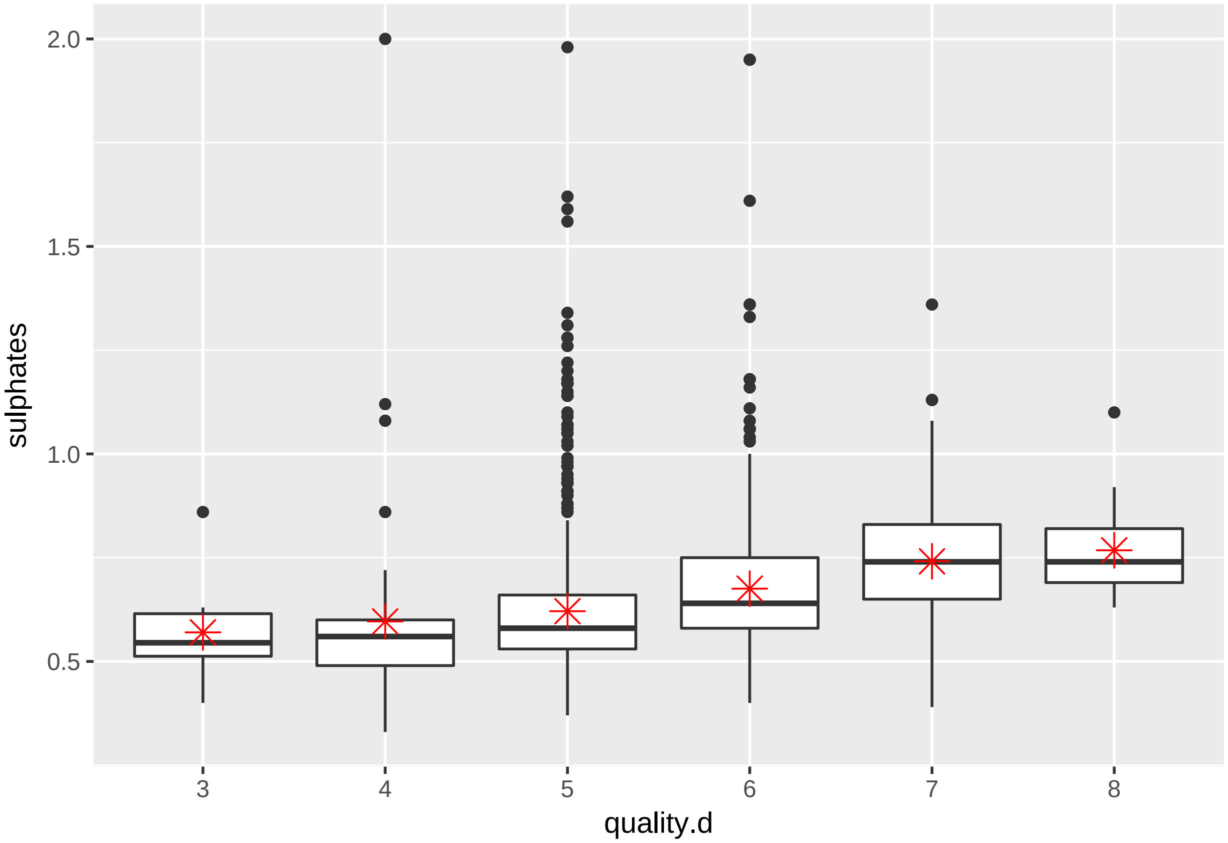
Quality vs pH



PH seems to be negatively correlated with quality.

```
## [1] -0.05773139
```

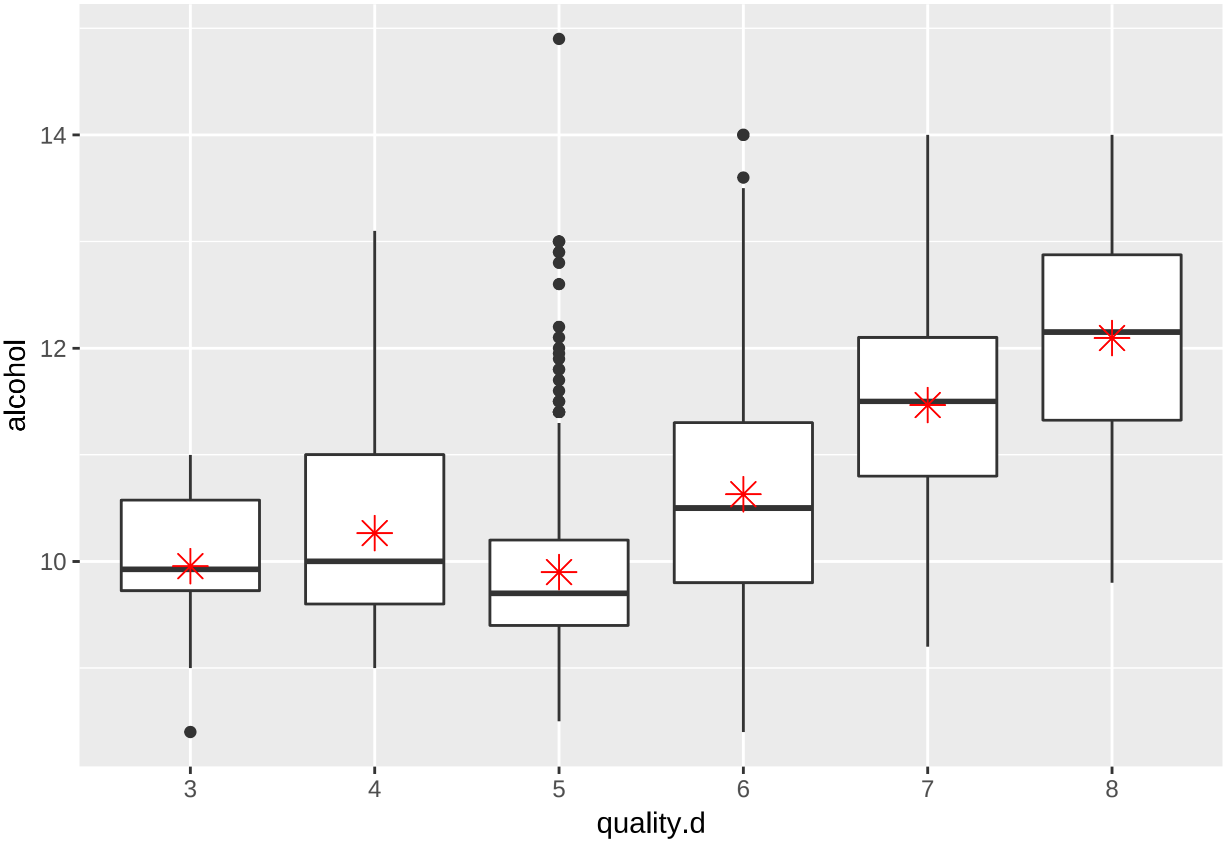
Quality vs sulphates



Sulphates and quality seems to be positevely correlated.

```
## [1] 0.2513971
```

Quality vs alcohol

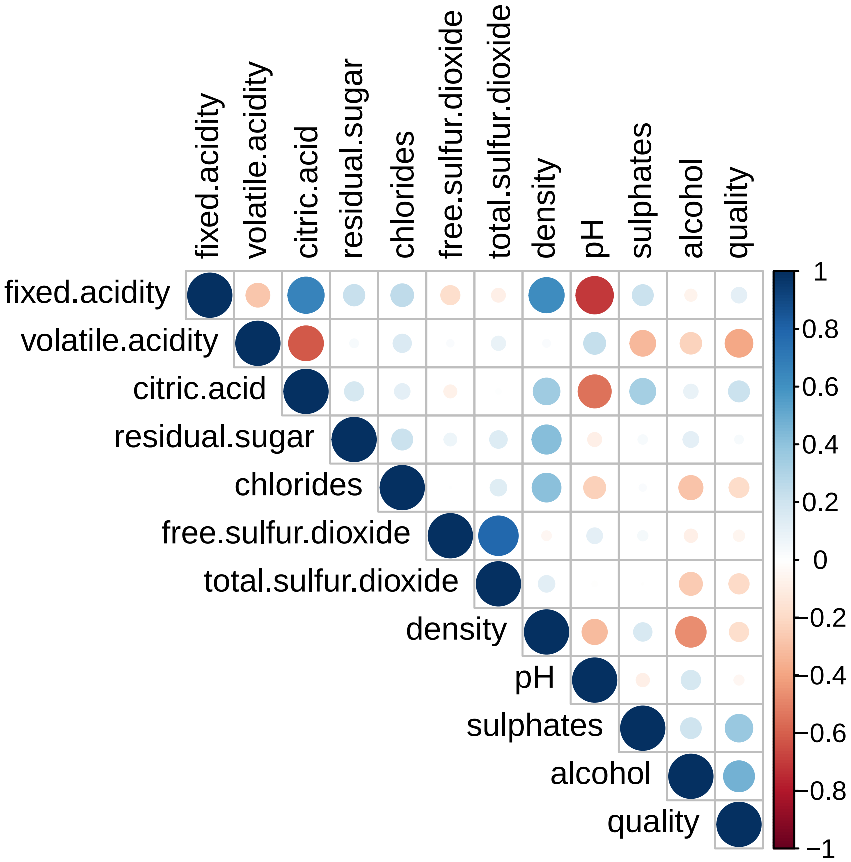


Alcohool and quality seems to be positevely correlated, except for the rate 5 which have a lot of ouliers.

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

Complementary bivariate plots

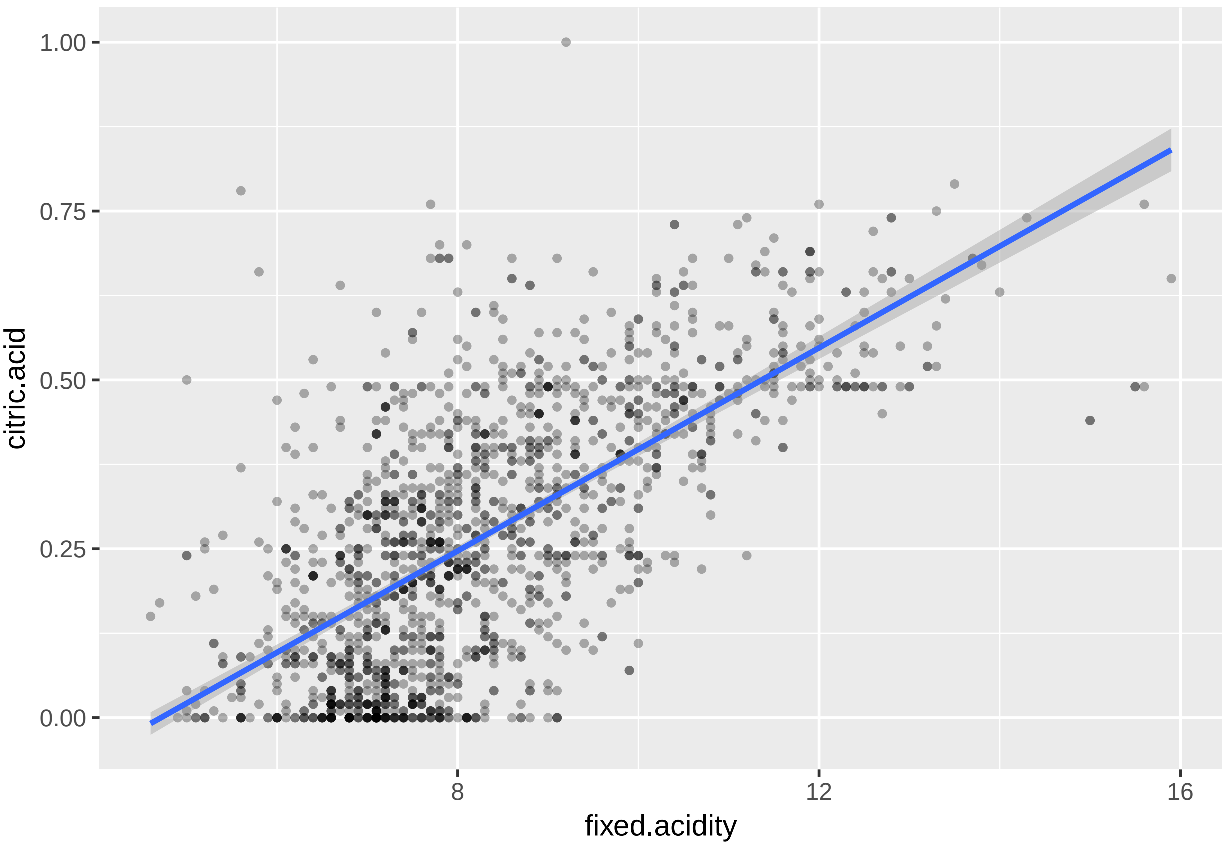
In order to choose some bivariate plots, we will start to create a correlation matrix, then visualize it.



We will plot the following bivariate correlations to get a better understanding:' - fixed acidity and citric acid - fixed acidity and density - fixed acidity and ph - free sulfur dioxide and total sulfur dioxide

Function to create scatterplot:

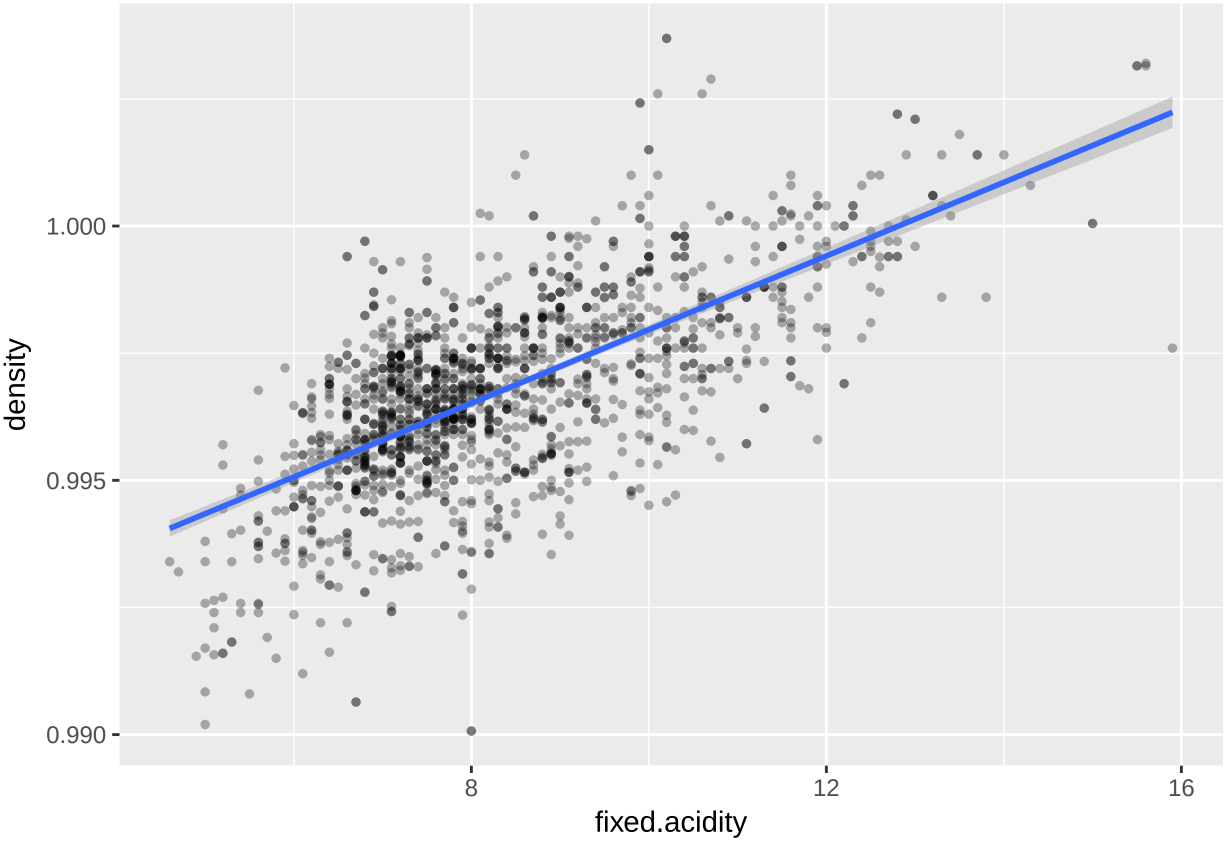
Fixed acidity vs citric acid



The scatterplot shows a linear positive moderate strong relation, nevertheless we can observe many outliers.

```
## [1] 0.6717034
```

Fixed acidity and density

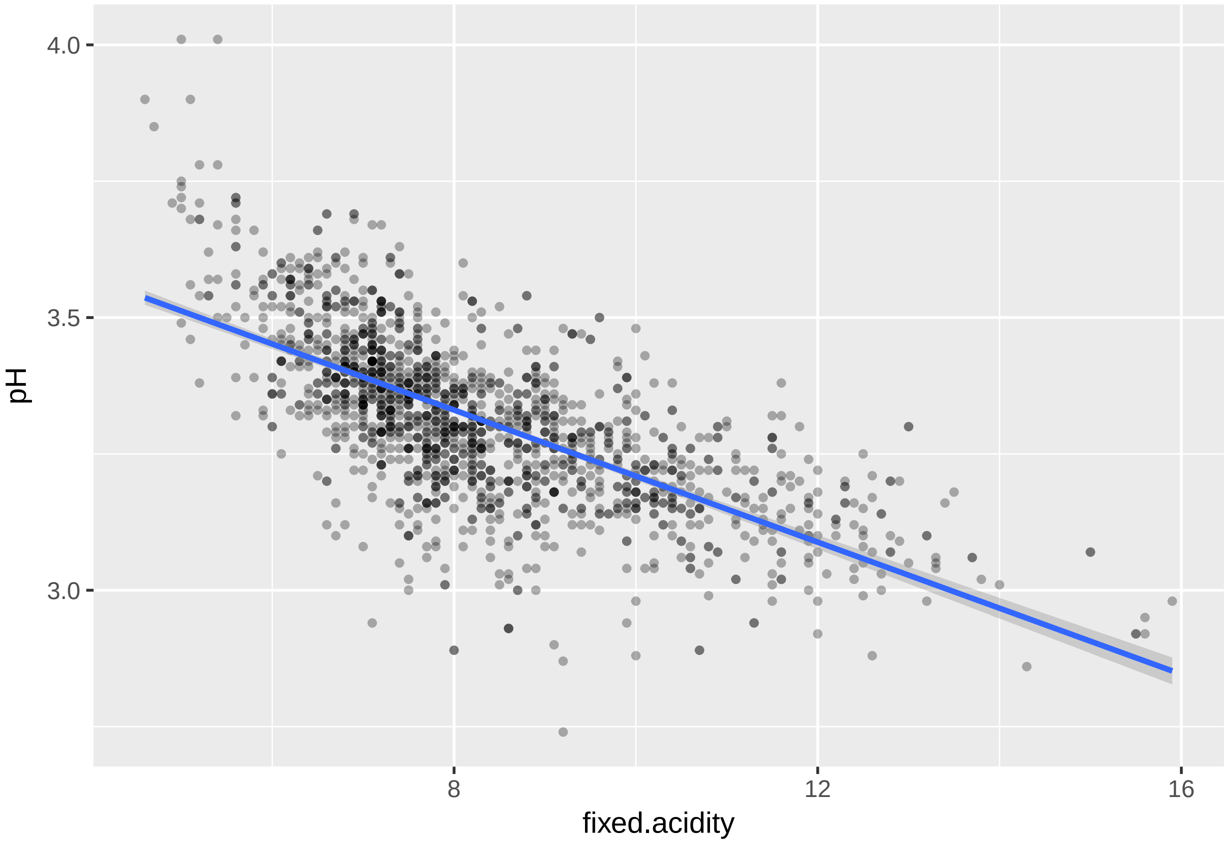


The scatterplot shows a linear positive strong relation.

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

Fixed acidity vs pH

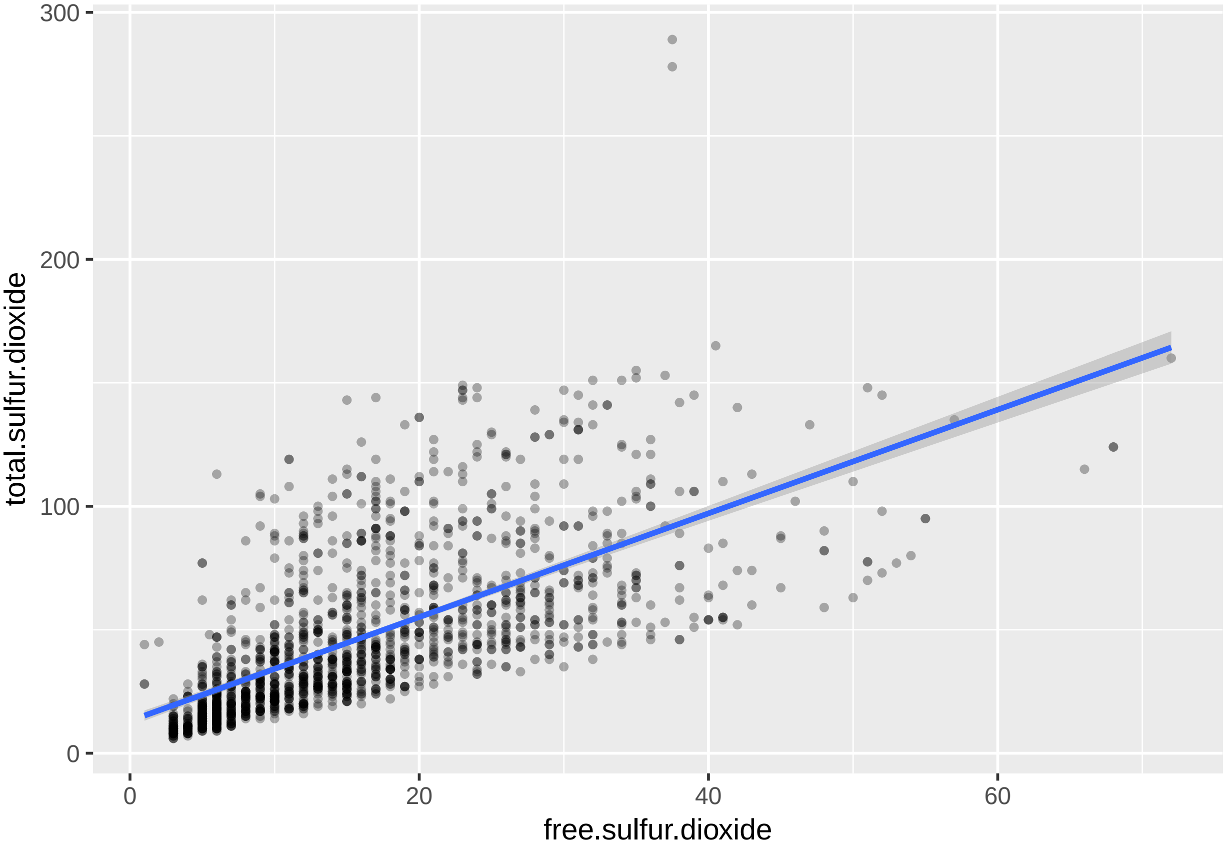
```
scatter_plot(x = "fixed.acidity",y="pH")
```



The scatterplot shows a linear negative strong relation.

```
## [1] -0.6829782
```

Free sulfur dioxide vs total sulfur dioxide

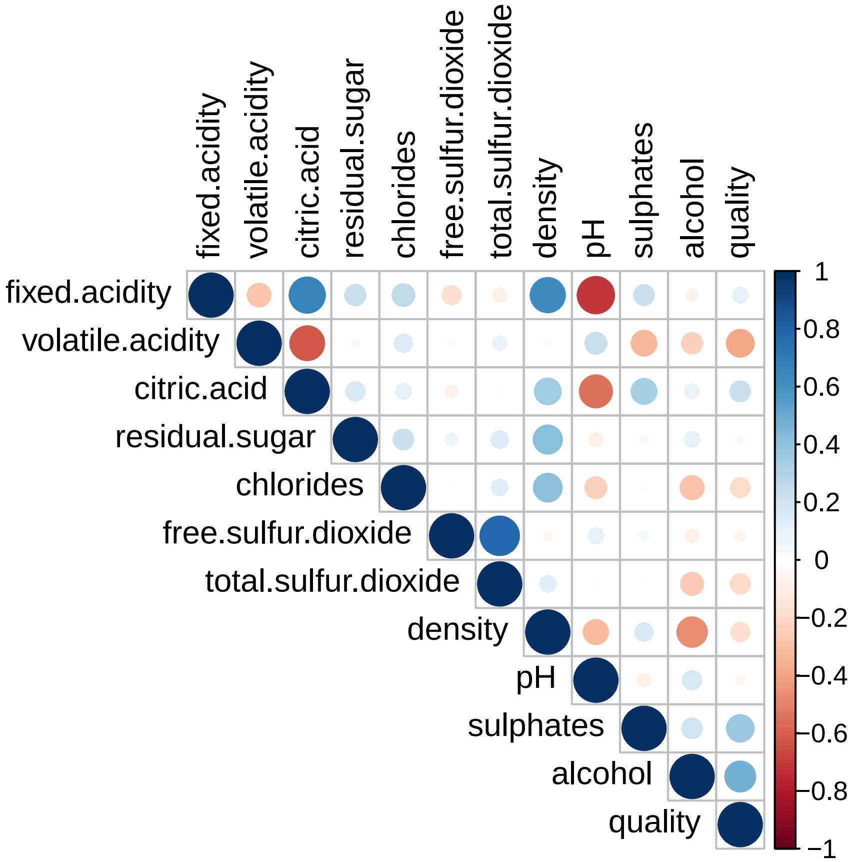


The scatterplot shows a linear positive strong relation.

```
## [1] 0.6676665
```

Bivariate Analysis

Correlation summary:



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?

Quality is positevely correlated with alcohol and sulpates and negatively with volatile acidity.

The other relations are not sufficiently correlated.

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)?

We observed the following relations: - a positive correlation between acid citric and fixed acidity. - a positve correlation between density and fixed acidity. - a negative correlation between pH and fixed acidity. - a stong positive relation between free sulfur dioxide and total sulfur dioxide.

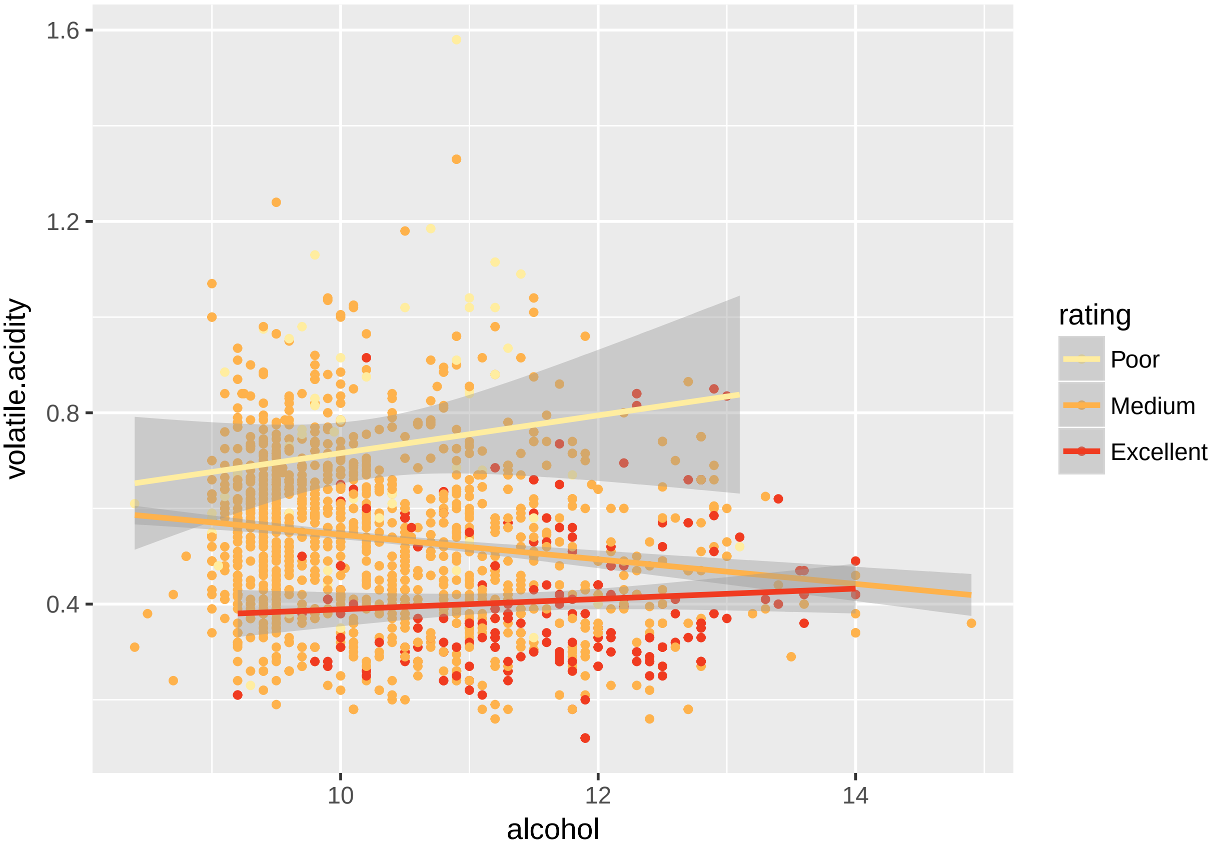
What was the strongest relationship you found?

The strongest relationship is between free sulfur dioxide and total sulfur dioxide, which is not really surprising, considering that total sulfur dioxide contains free sulfur dioxide.

Multivariate Plots Section

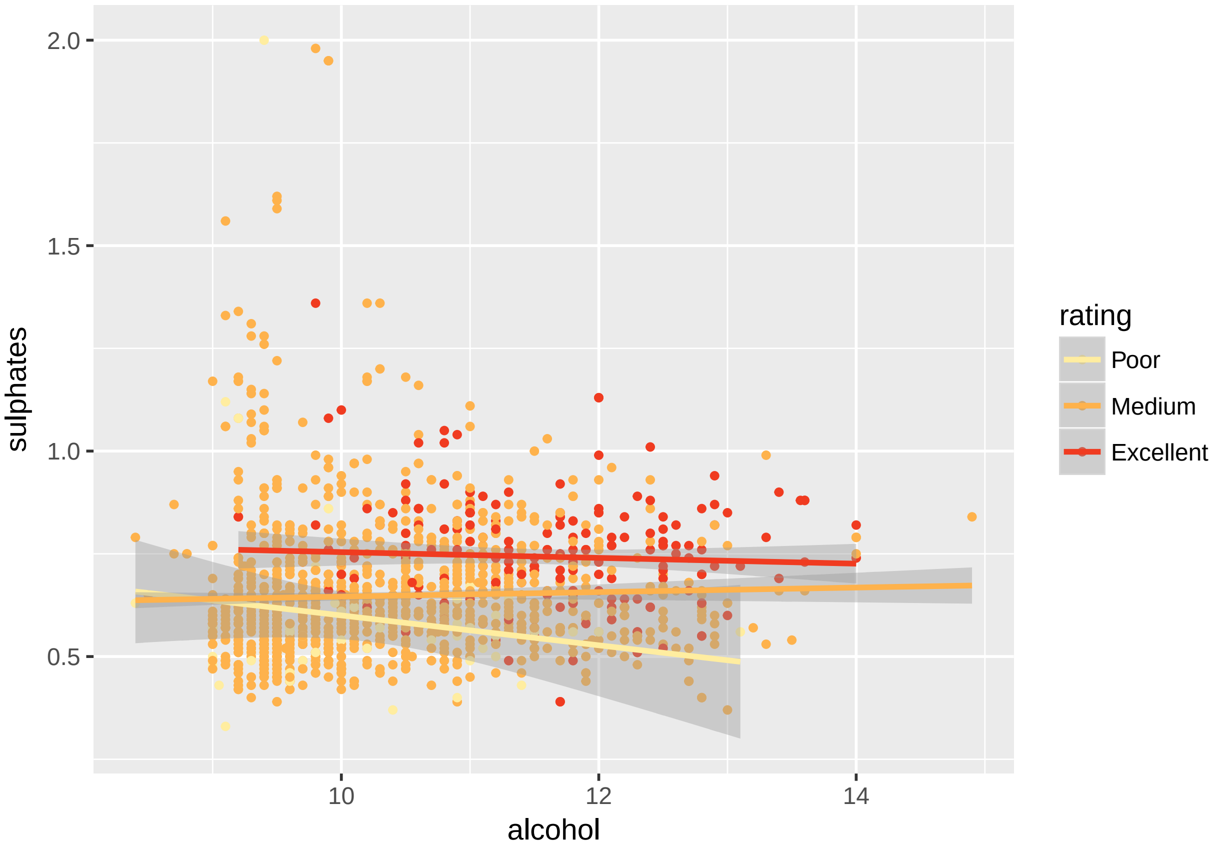
Funtion to create multivariate function:

Alcohol, volatile acidity and quality



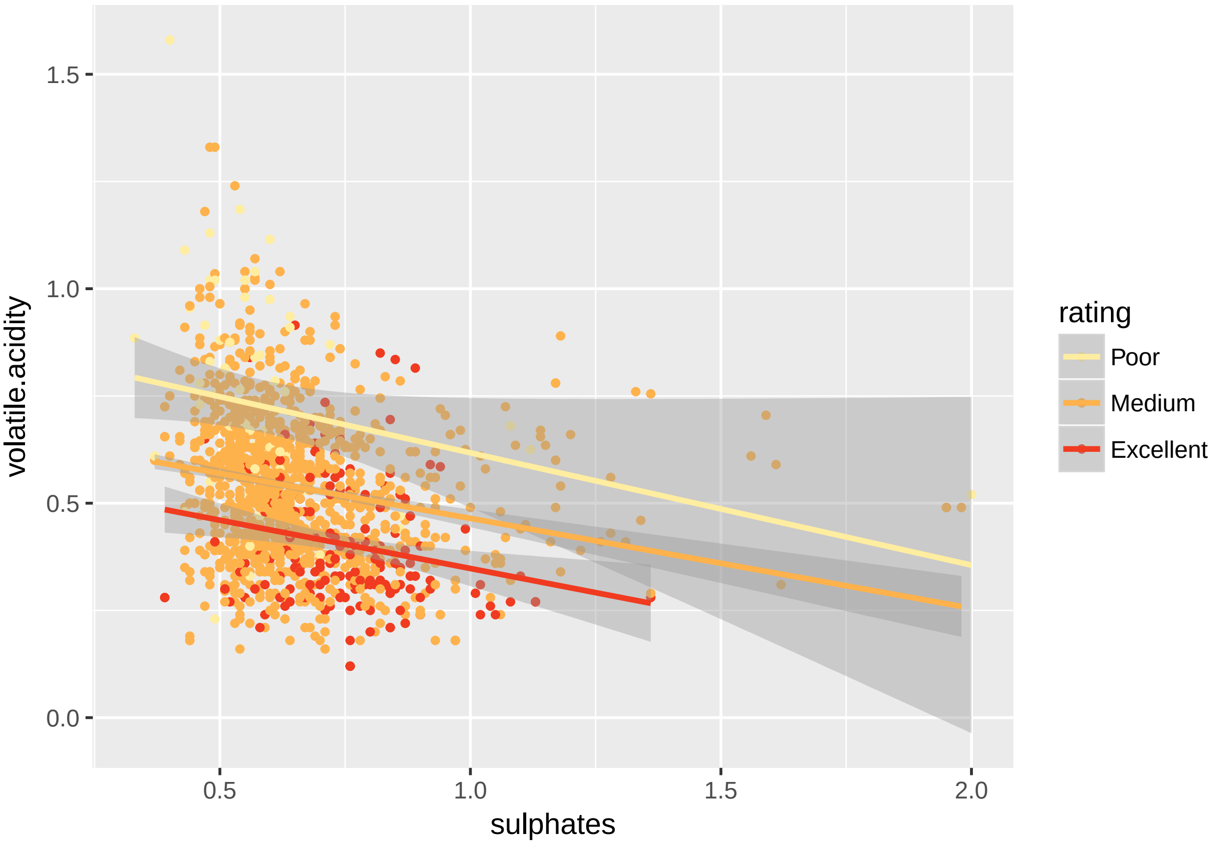
We can observe that excellent wines tend to have lower volatile acidity over the entire range of alcohol.

Alcohol, volatile acidity and quality



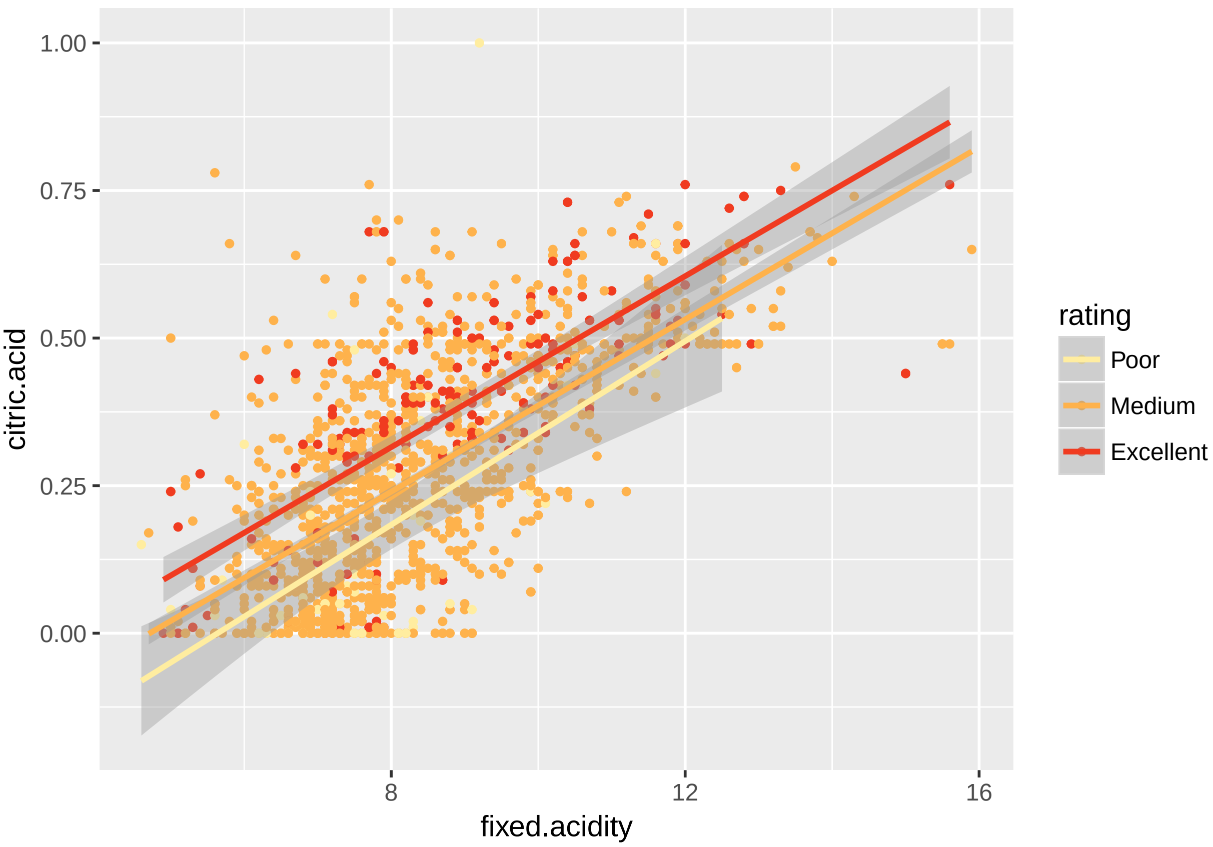
We can observe that excellent wines tend to have higher sulphates over the entire range of alcohol.

Sulphates, volatile acidiy, quality



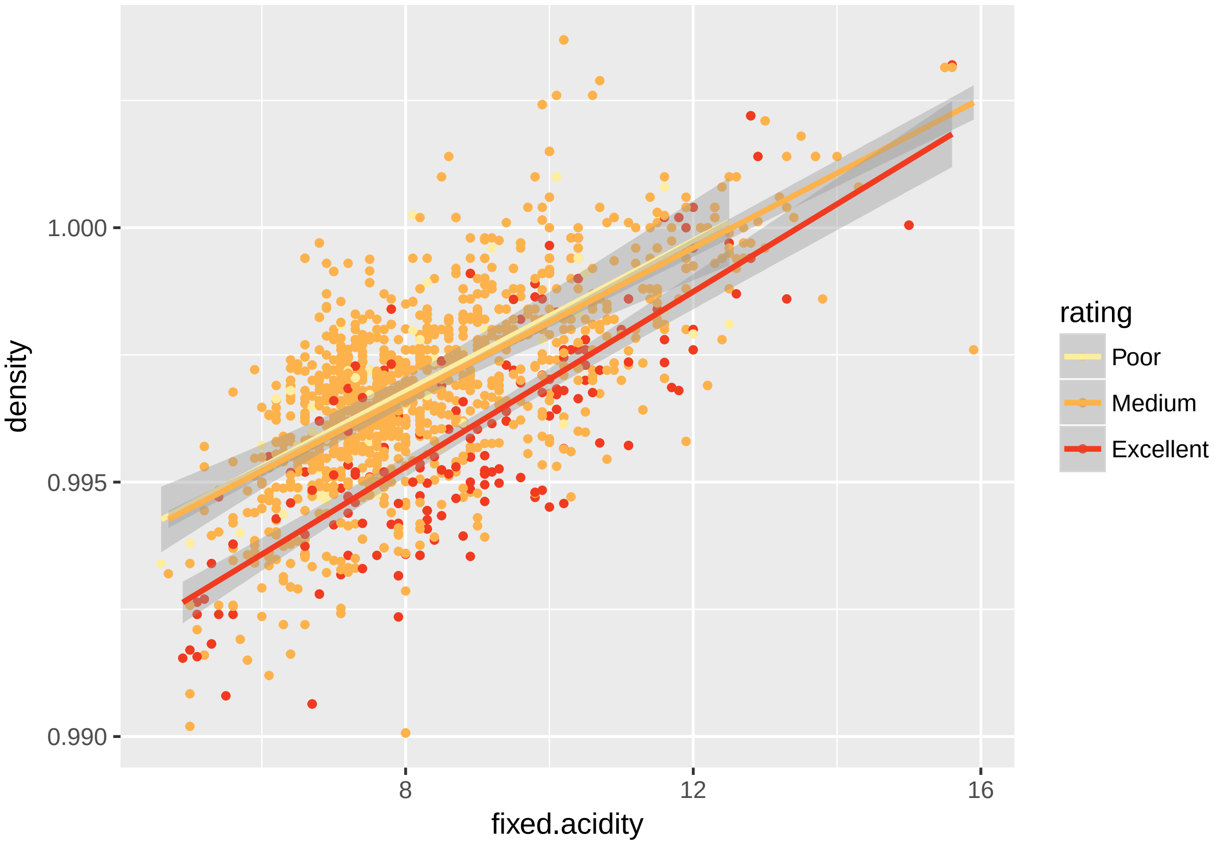
We can observe that excellent wines tend to have lower volatile acidity over the entire range of sulphates.

Fixed acidity, citric acid, quality



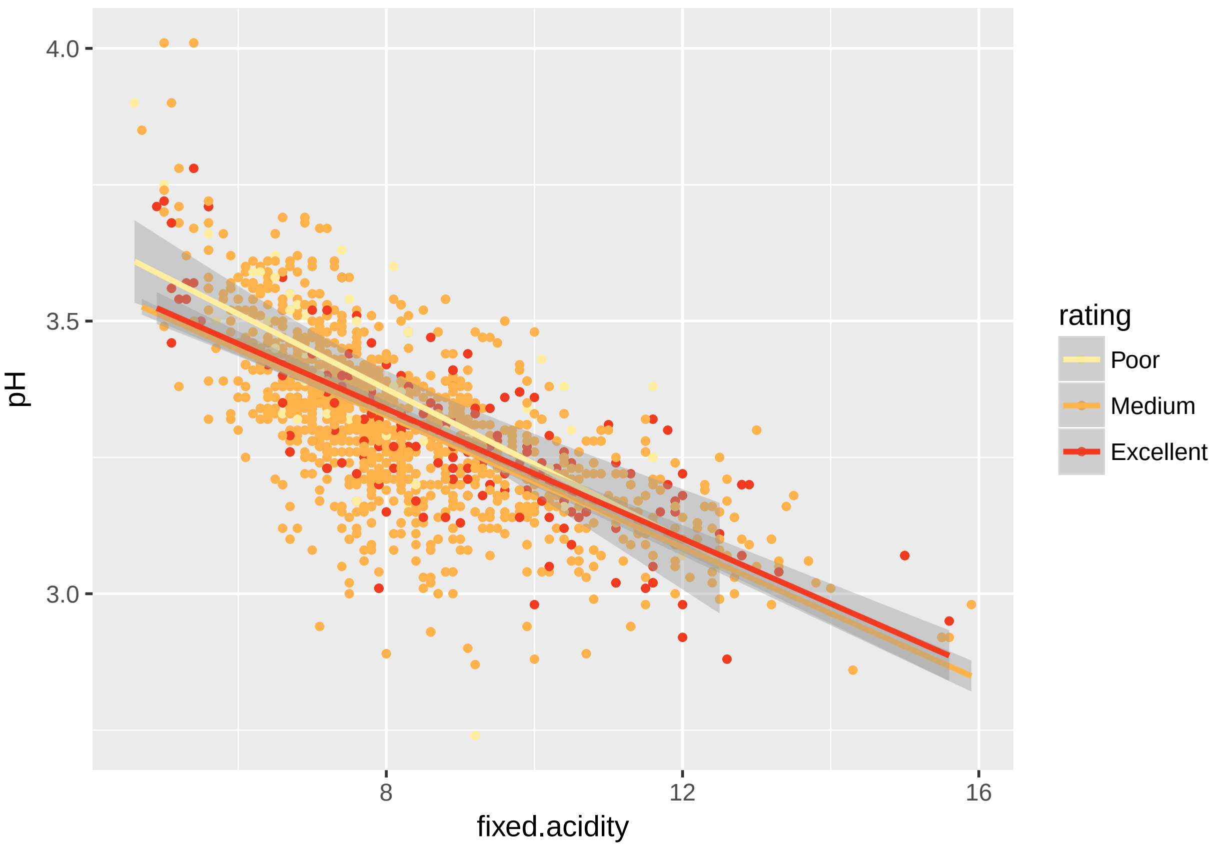
We can observe that excellent wines tend to have higher citic acid over the entire range of fixed acidity.

Fixed acidity, density, quality



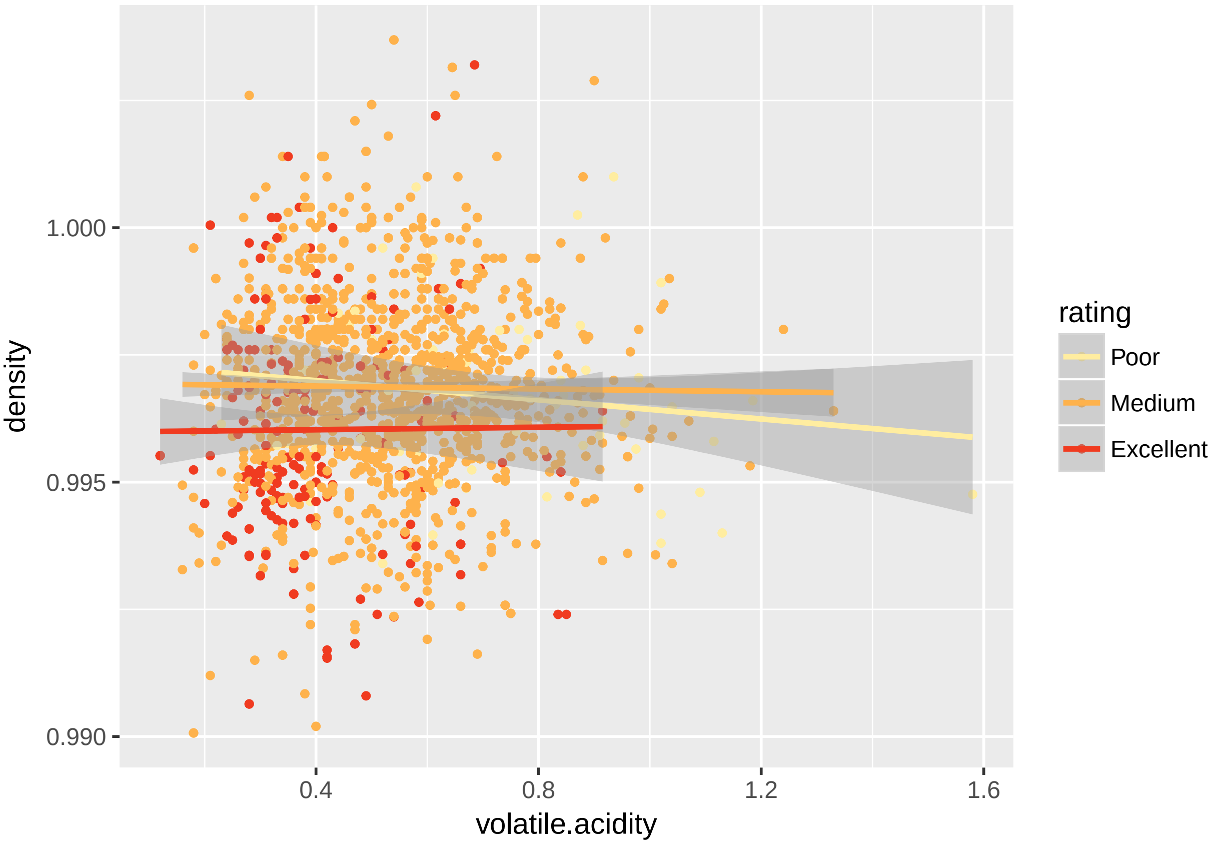
We can observe that excellent wines tend to have lower density over the entire range of fixed acidity.

Fixed acidity, pH, quality



We do not see a clear pattern between the three variables.

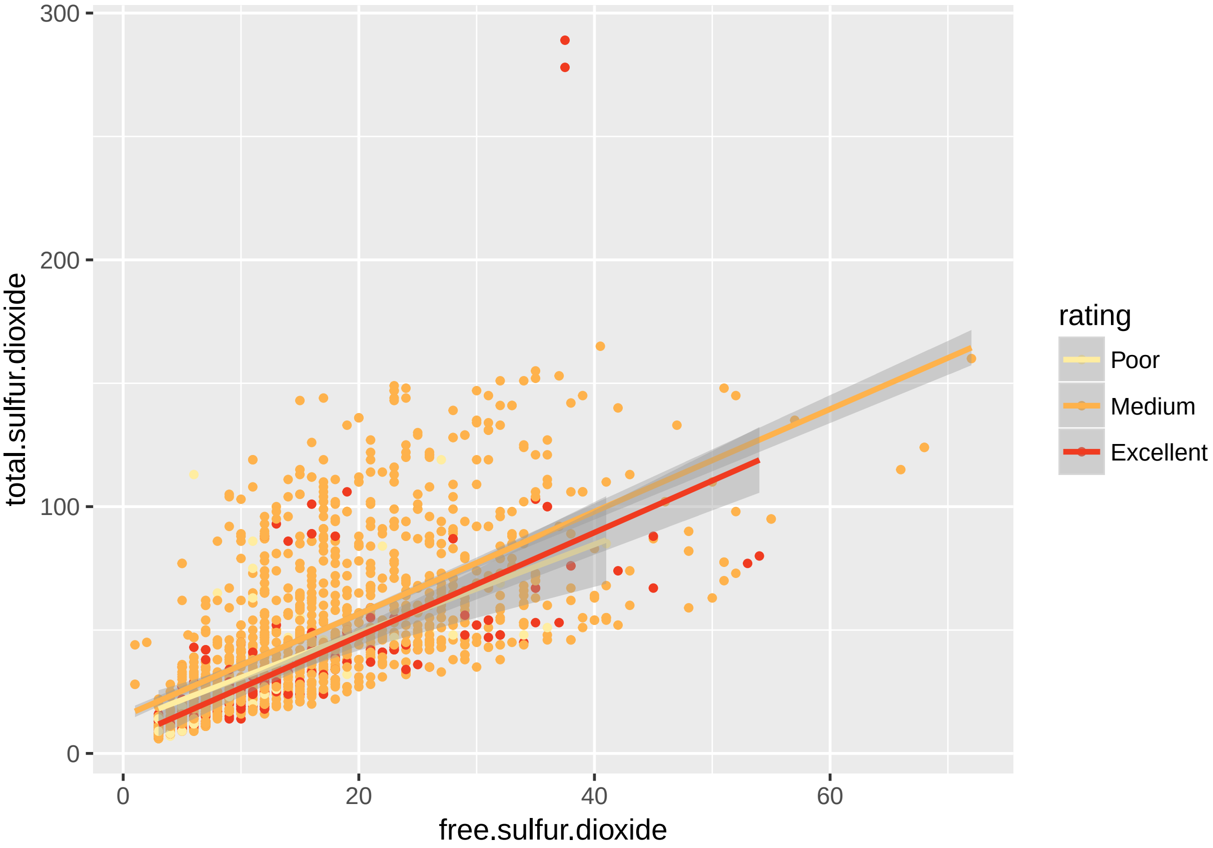
Volatile acidity, density and quality



We can observe that excellent wines tend to have lower density over the entire range of volatile acidity.

We can also see that excellent wines are represented on a lower range of volatile acidity than medium and poor.

Free sulfur dioxide, total sulfur dioxide, quality



We do not see a clear pattern between the three variables.

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?

We principally explored the relationships which have the highest correlation with quality.

We observed that higher alcohol and lower volatile acidity means generally quality wines, that excellent wines tend to have higher citic acid over the entire range of fixed acidity.

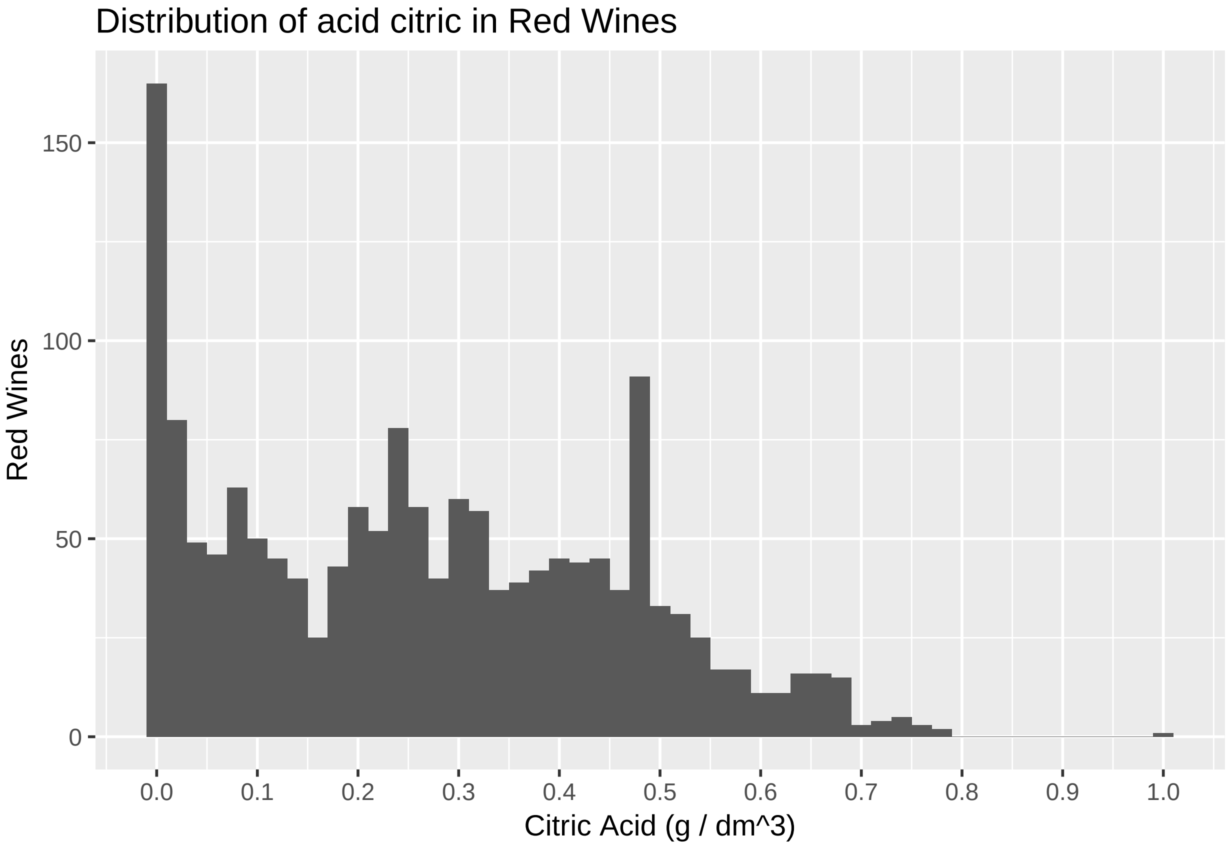
We also observed that excellent red wines tend to have higher citic acid over the entire range of fixed acidity, and lower density over the entire range of volatile acidity.

Were there any interesting or surprising interactions between features?

Low density and low volatile acidity contain the better red wines.

Final Plots and Summary

Plot One



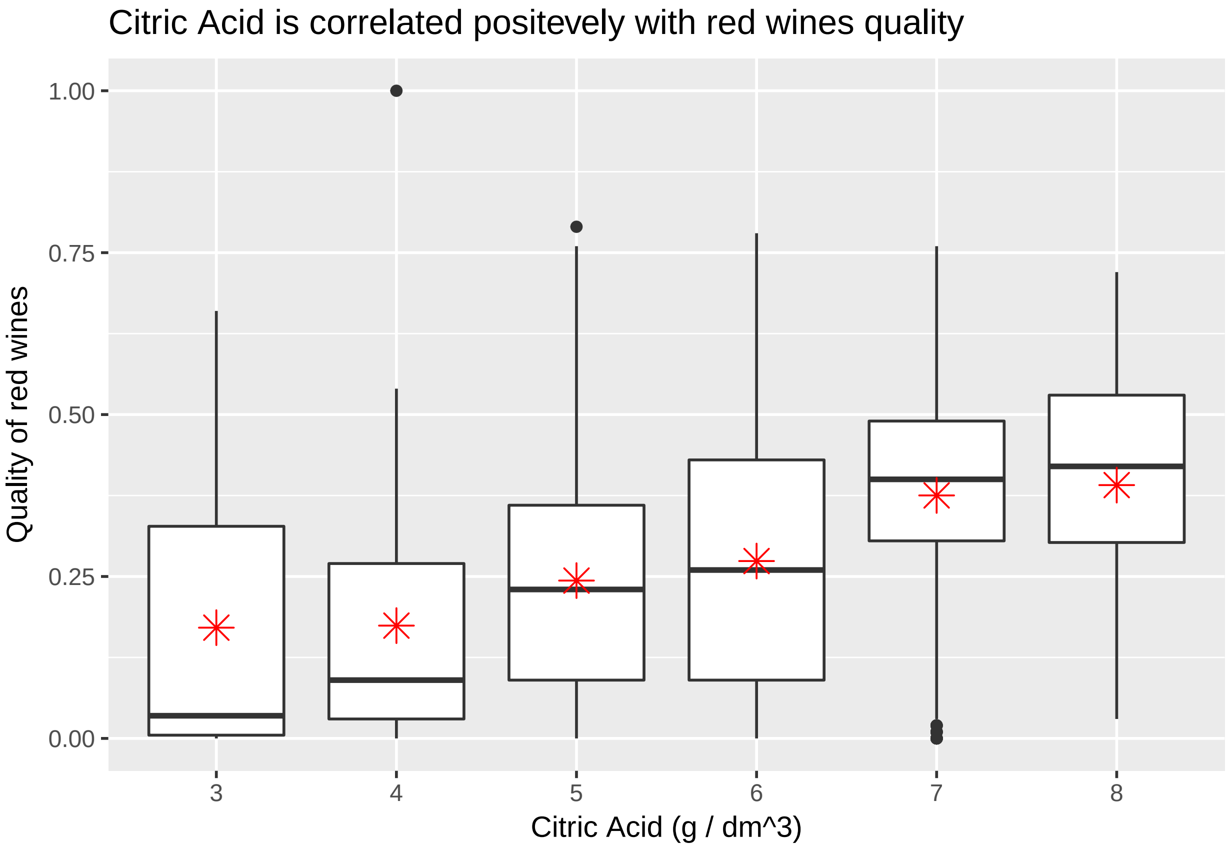
Description One

The graph depicts the distribution of acid citric in Red Wines.

This distribution wakes up our interest, because is the only muli-modal distribution among the variables.

Thus, it will be interesting to explore is relation with the quality variable.

Plot Two



Description Two

The box plot describes the relation between citric acid and quality of red wines.

Citric acid and qualiry of quaity of red wines have a postive correlation.

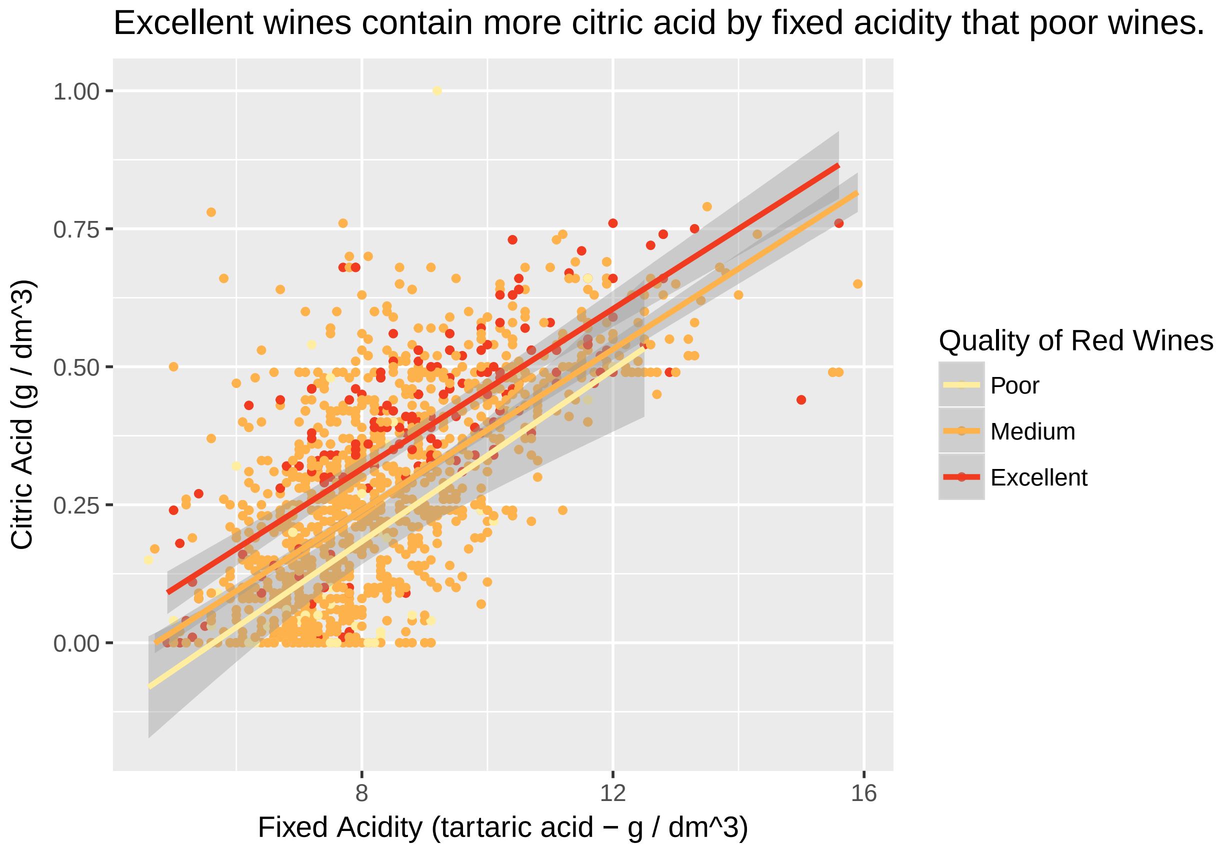
We can calcul the correlation to deepen our understanding of the two variables:

```
## [1] 0.2134809
```

The correlation is positive but not really strong.

But it can be interesting to see the relation between citric acid and other variables.

Plot Three



Description Three

The multivariate graph represents the relation between citric acid, fixed acidity and quality.

We can observe that for each level of fixed acidity, excellent wines contain more citric acid than the two other categories (medium and poor).

Excellent wines conatin more citric acid by fixed acidity that poor wines.

Reflection

Our project was to explore the variables composing red wines, and particullary their impacts on the quality of red wine.

In a first time, we plotted each distributions variables. The vast majority of variable were right skewed, but we found out a mutivariate disbrution, citric acid, which attrated our curiosity. We also found out a limitation at our analysis: the distribution of the wines quality is largely composed by two rating: 5 and 6

In a second time, in order to not miss something, we plotted each variables with the quality variable. We discover the following cues: - a positive correlation with acid citric, alcohol and sulphates. - a negative correaltion with volatile acidity.

In parallel we observed the following relations: - a positive correlation between acid citric and fixed acidity. - a positve correlation between density and fixed acidity. - a negative correlation between pH and fixed acidity. - a stong positive relation between free sulfur dioxide and total sulfur dioxide.

At this step it was difficult to analyze the factor to a good wines, because the correlation between the variables and the quality was not strong enough.

In order to get a better understanding, we chosen a number of multivariate to plot. The most notables were: - Alcohol, volatile acidity and quality - Alcohol, volatile acidity and quality - Sulphates, volatile acidiy, quality - Fixed acidity, density, quality - Volatile acidity, density and quality

And of course Fixed acidity, citric acid and quality which ended our reflection on citric acid and quality.

We observed that for each level of fixed acidity, excellent wines containned more citric acid than the two other categories of red wines (medium and poor).

To improve our analysis we can collect more data, especially wines with a rating different of 5 and 6.

Ressources

Info on the data set: https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt Plot: https://docs.google.com/document/d/1-f3wM3mJSkoWxDmPjsyRnWvNgM57YUPloucOIl07l4c/pub Histogram: http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization#basic-histogram-plots Correlation: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-form Multivariate analysis: http://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/src/multivariateanalysis.html