Red Wine Dataset

by Adarsh Pawar

All of the packages that I end up using in my analysis in this code chunk.

Loading the Wine data in the R and review of the head of the data set.

```
## [1] "/Users/adarshpawar/Desktop/FP2"

##   X fixed.acidity volatile.acidity citric.acid residual.sugar
chlorides
## 1 1           7.4             0.70        0.00            1.9
0.076
## 2 2           7.8             0.88        0.00            2.6
0.098
## 3 3           7.8             0.76        0.04            2.3
0.092
## 4 4          11.2             0.28        0.56            1.9
0.075
## 5 5           7.4             0.70        0.00            1.9
0.076
## 6 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
```

Looking to the data structuer and data type.

```
## '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 ...
```

Getting the information about the NA rows in the dataset.

```
## [1] 0
```

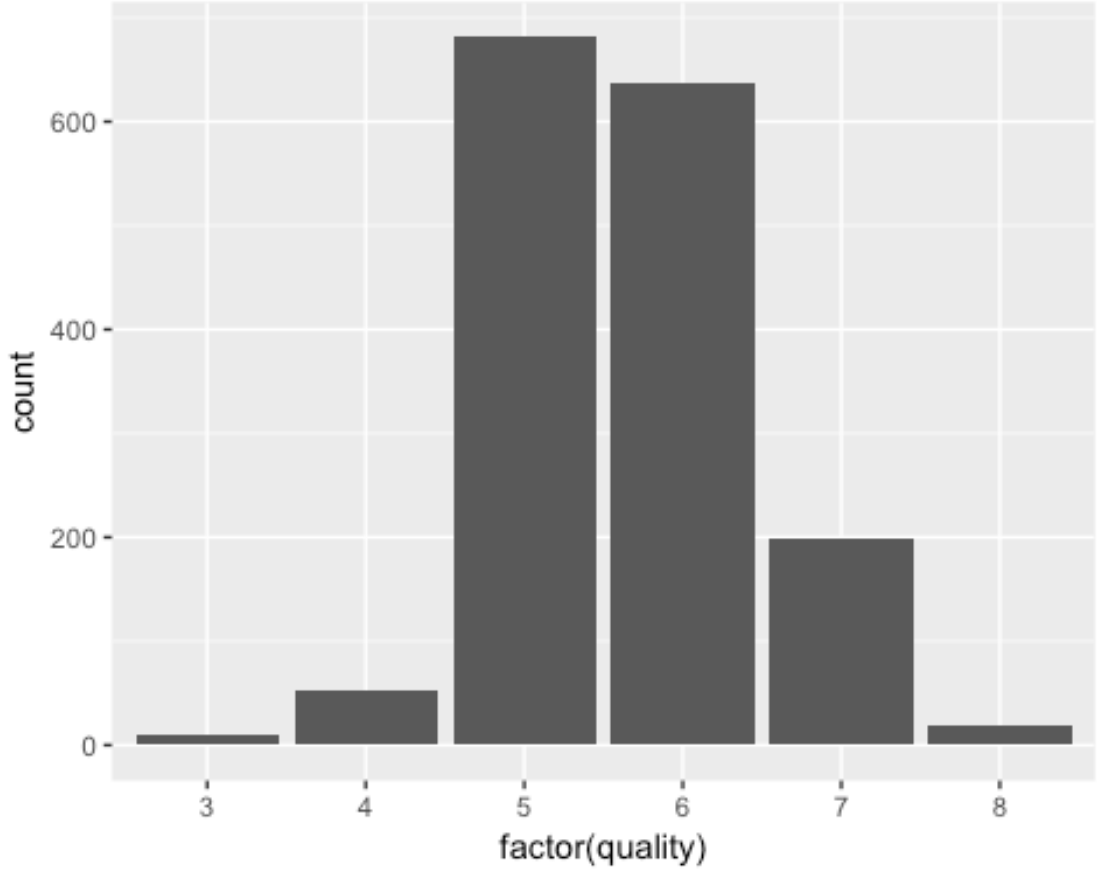
Introduction:

Among all dataset that Udacity provides, I found the Wine Data Set compatible and interesting for me to explore. There is one more reason I have to choose that dataset because I like to know more about wine just like I did with the Diamonds in the previous lessons with Udacity. It was amazing to experience that type of exploring data though, I learned lots of things. The same I

want to do in this Wine Dataset. The dataset consists of 13 variables and 1599 observations integer and numeric data types containing data about different features of red wines for example acidity, sugar, and alcohol content and quality. I'm going to analyze the relationship between all the variables and see what type of effect they make in wine quality.

Univariate Plots Section

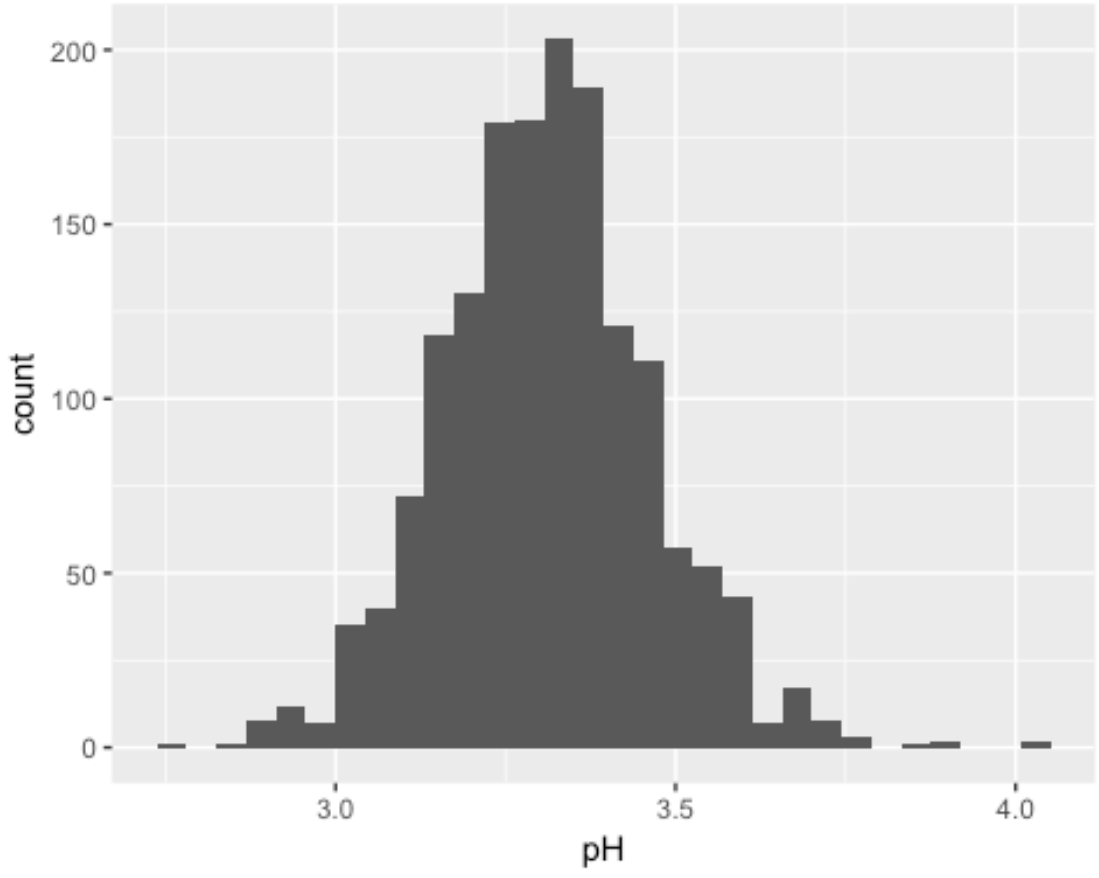
I am going to start my analyze by creating the plot to get the information about the most important variable which is the quality rate in the data.



Getting the Summary of the quality variable in the dataset.

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

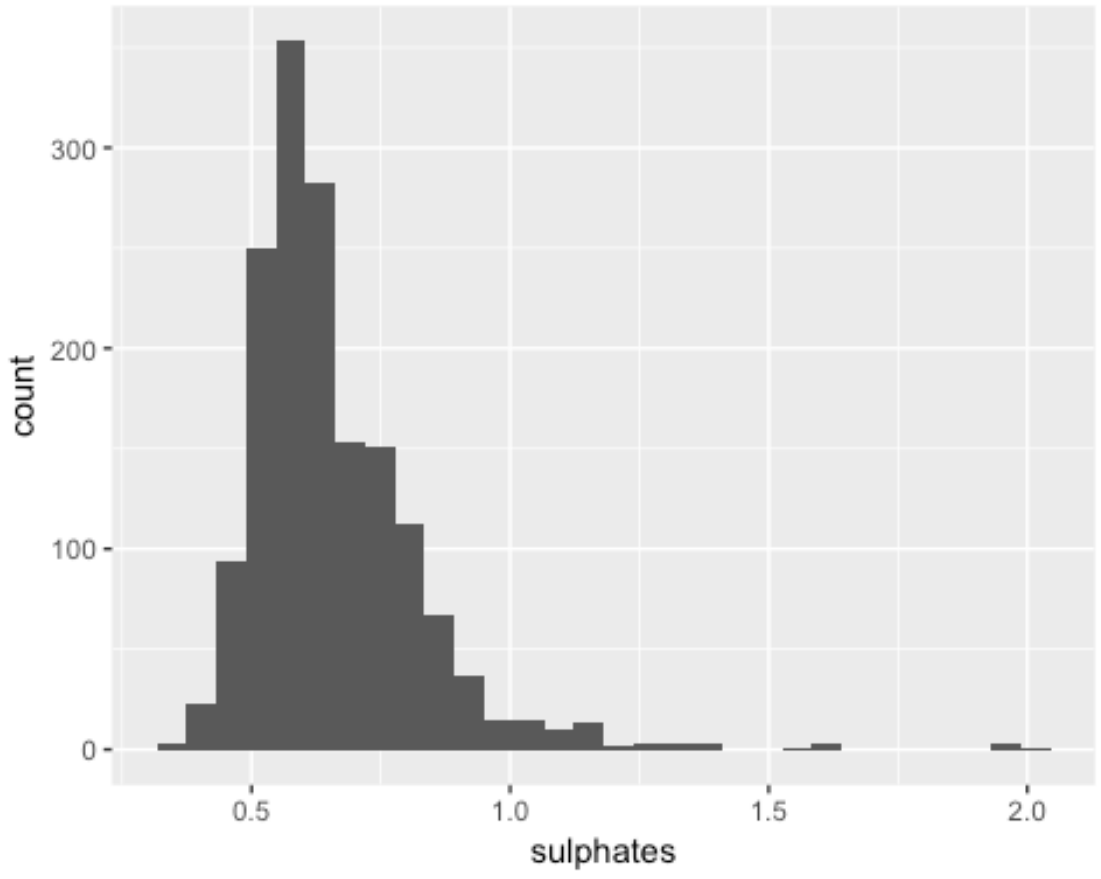
I got a normally distributed histogram, most wine have a quality between 5 and 7, the tallest clusters of bars is 6, representing the most common quality. The mean value is 5.636, the min quality is 3 and the max quality is 8.



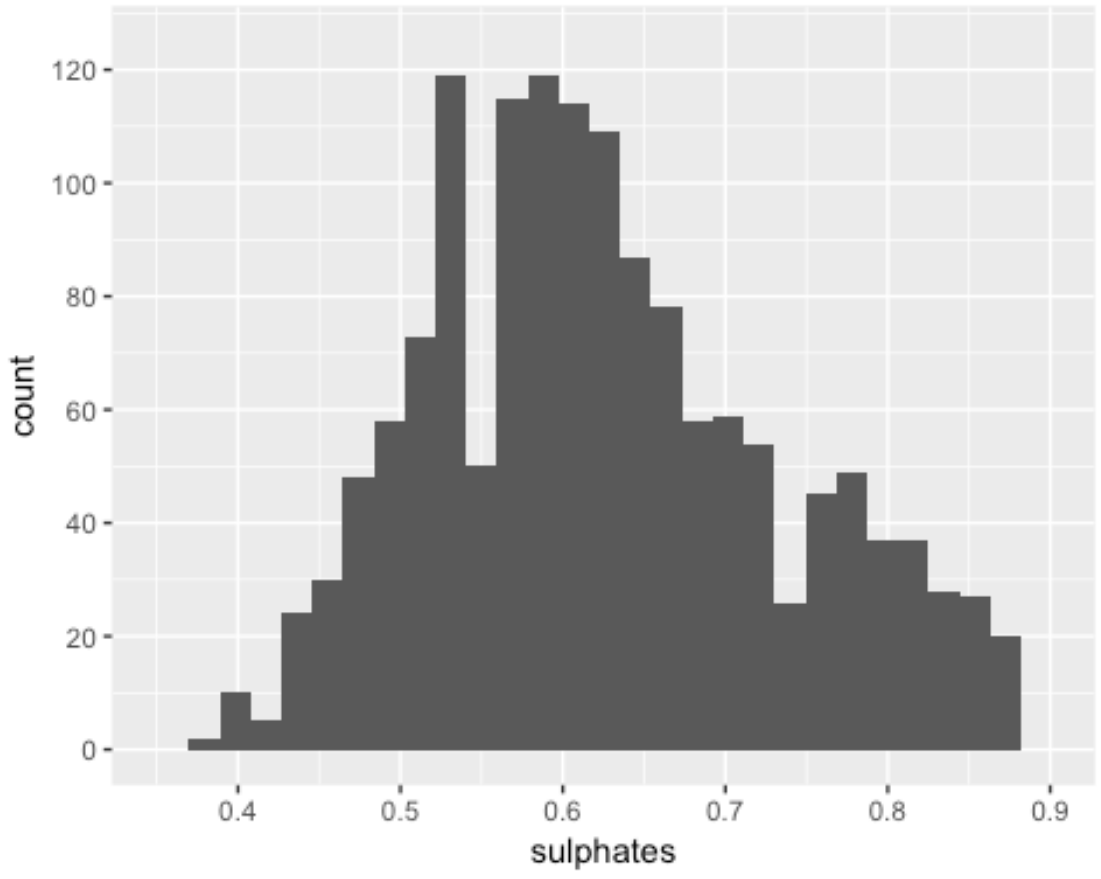
Summary of the pH plot.

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

The histgoram of pH is also normally distibuted and concentrated around 3.27. The min and max values are 2.74 and 4.01, the median is 3.310.



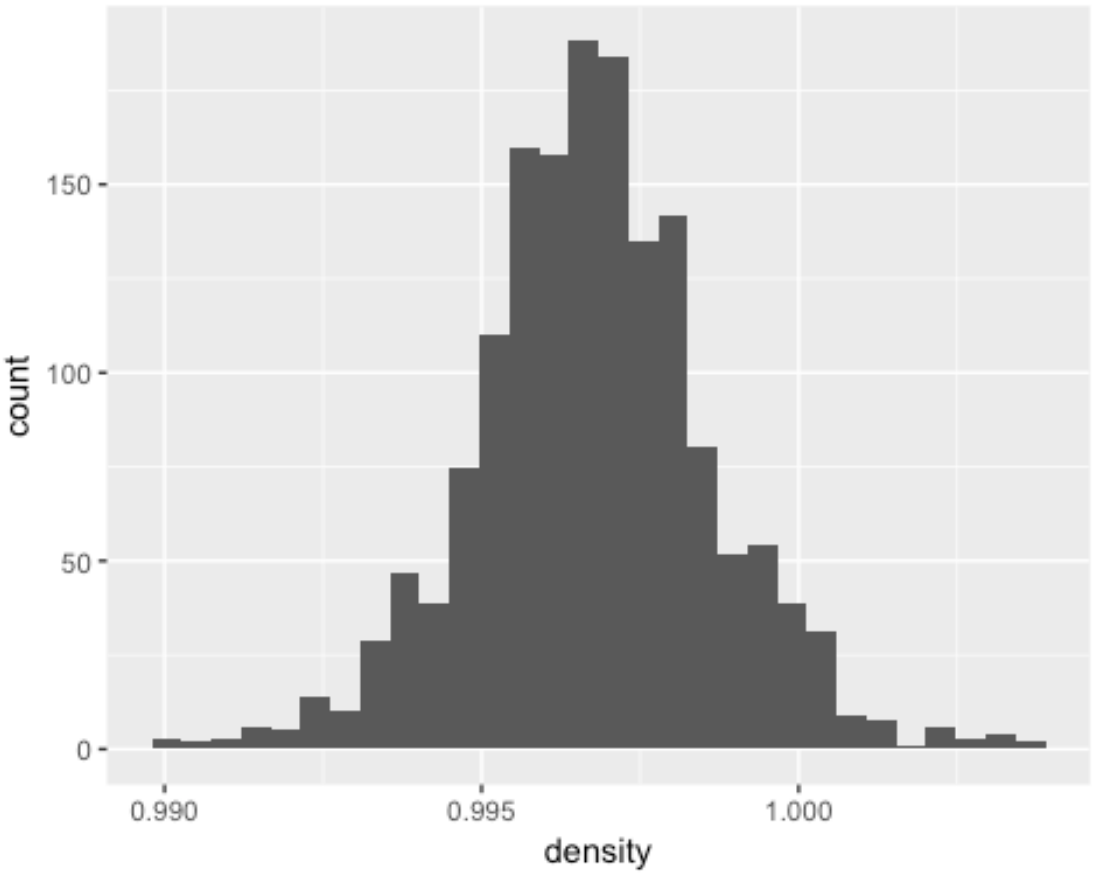
A sulphates plot with the ignoring values after 0.9 by setting the x-axis limit in the plot.

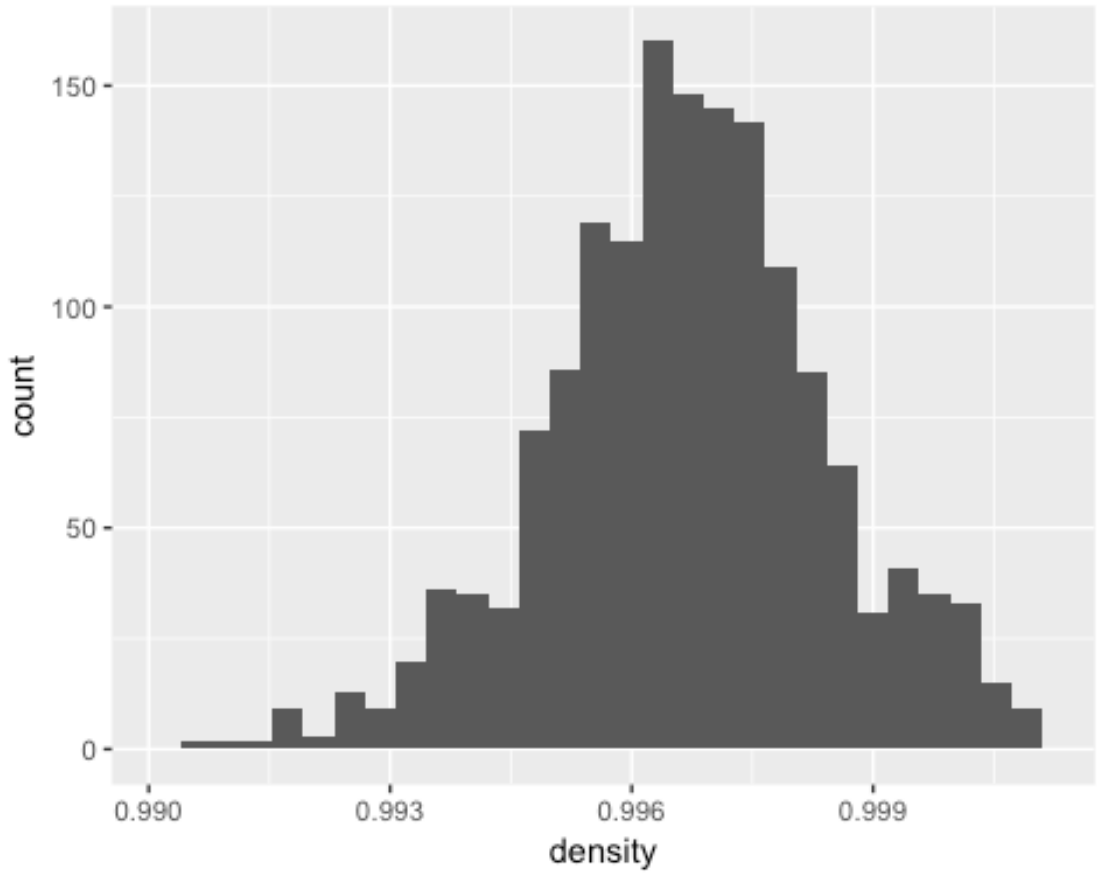


Summary of the sulphates plot.

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

Thw histogram of sulphates contains the most ofttен occuring values are between 0.4 and 0.9, the peak is about 120. For imporove readability and remove outstanding values I used ‘xlim’ function.

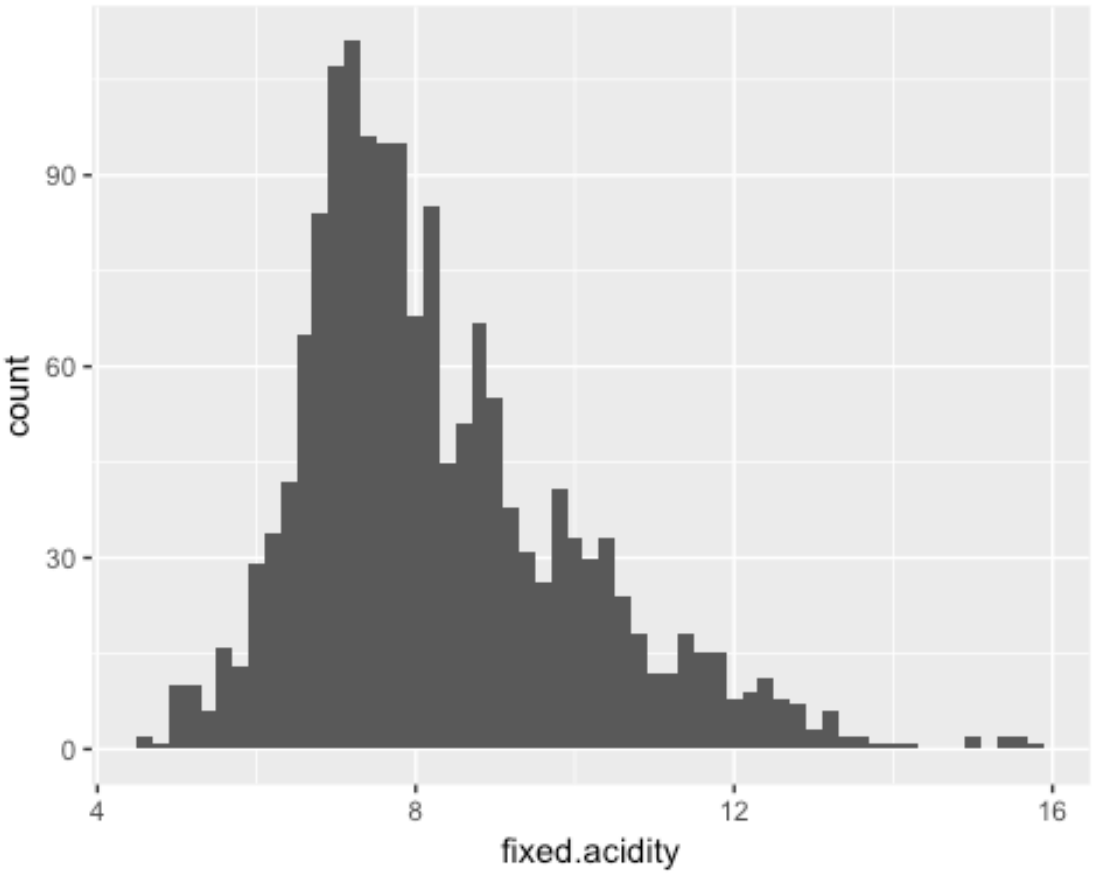

150
100
50
0
count
0.990
0.995
1.000
density

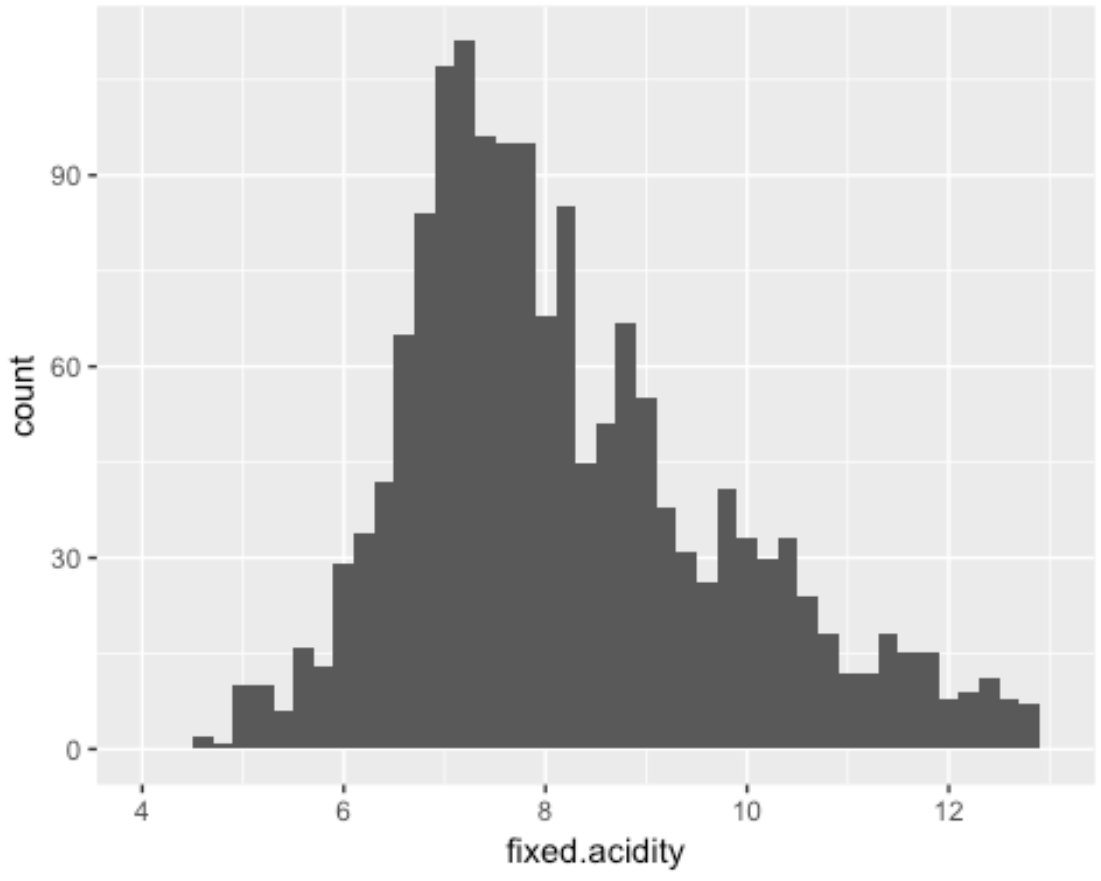


Summary of the density plot.

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

The first histogram of the density variable is normal distrybuted. The differance between the min value and max value is too low. In sencond plot I used the xlim fuction to remove the outstanding values to make the plot more readable.

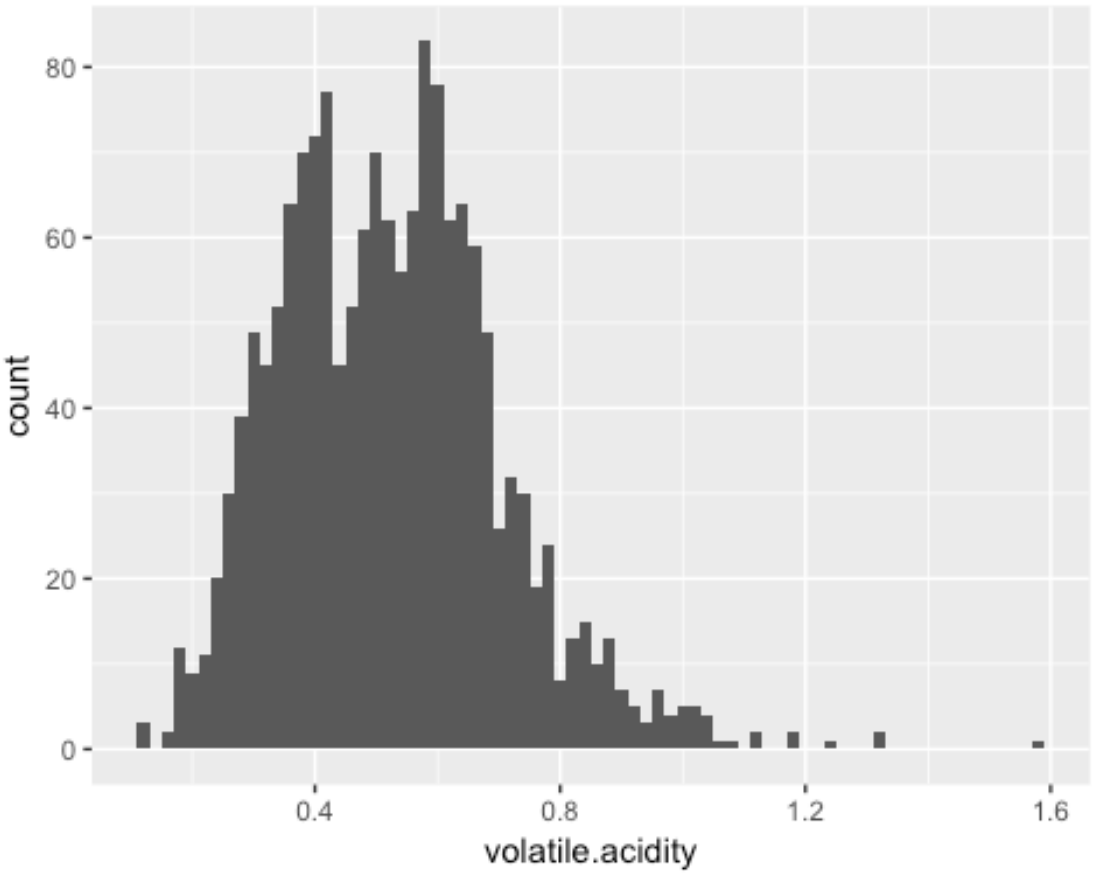

90
60
30
0
count
4
8
12
16
fixed.acidity

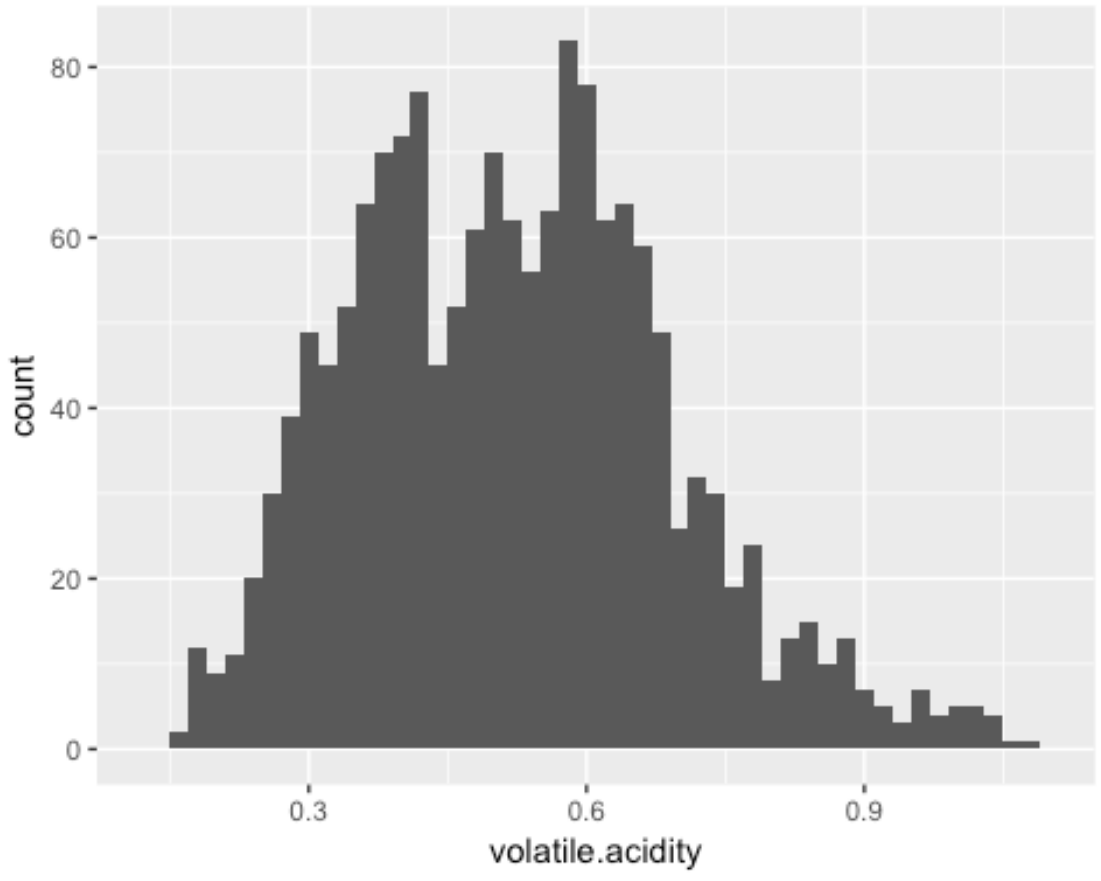


Summary of the fixed acidity.

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

The ditribution of fixed acidity has its peak around 7 and skeewed to the right. In the second plot the outliner is removes using xlim function. Fixed acidity variable have 4.60 min valuve and 15.90 max value for new plot the max value is 13.

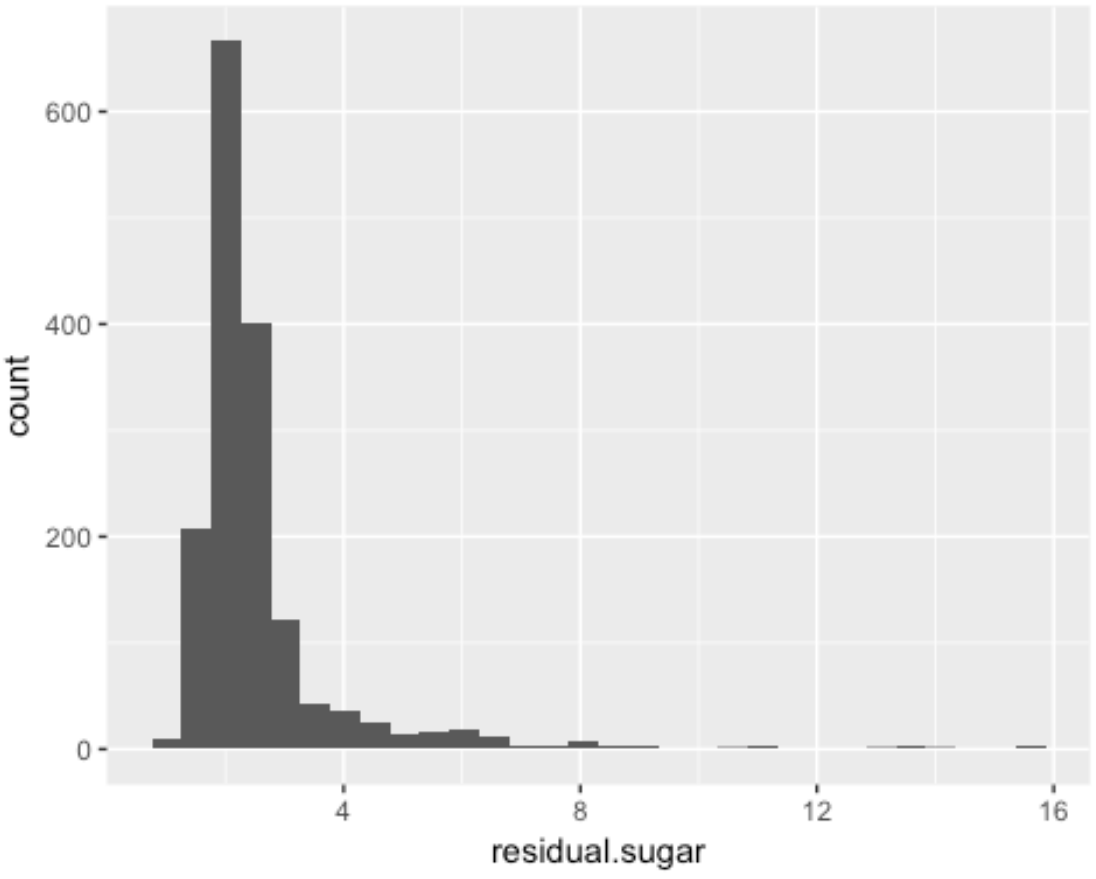

80
60
40
20
0
count
0.4
0.8
1.2
1.6
volatile.acidity

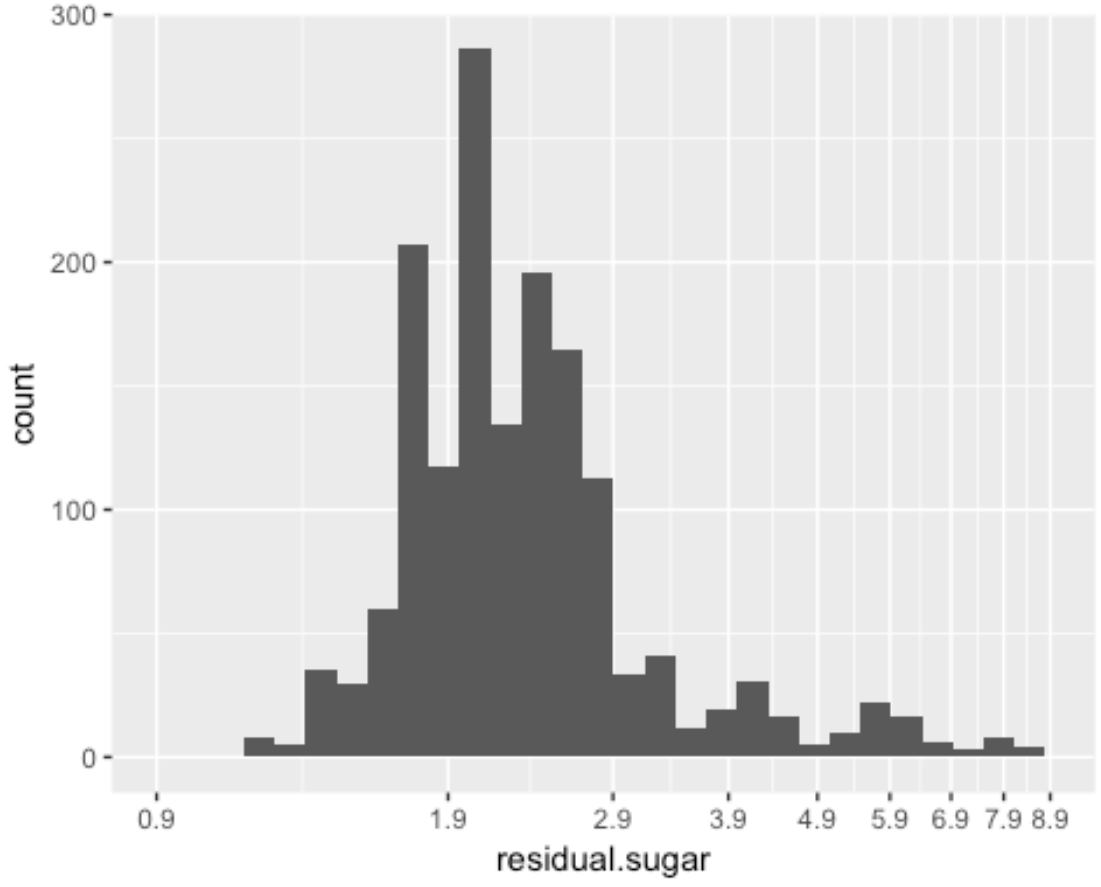


Summary of the volatile acidity.

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

The histogram of volatile acidity is concentrated around 0.6 and skewed to the right. The most values are lacated between 0.3 and 0.8. I transformed the second plot with xlim to limit the outliers.

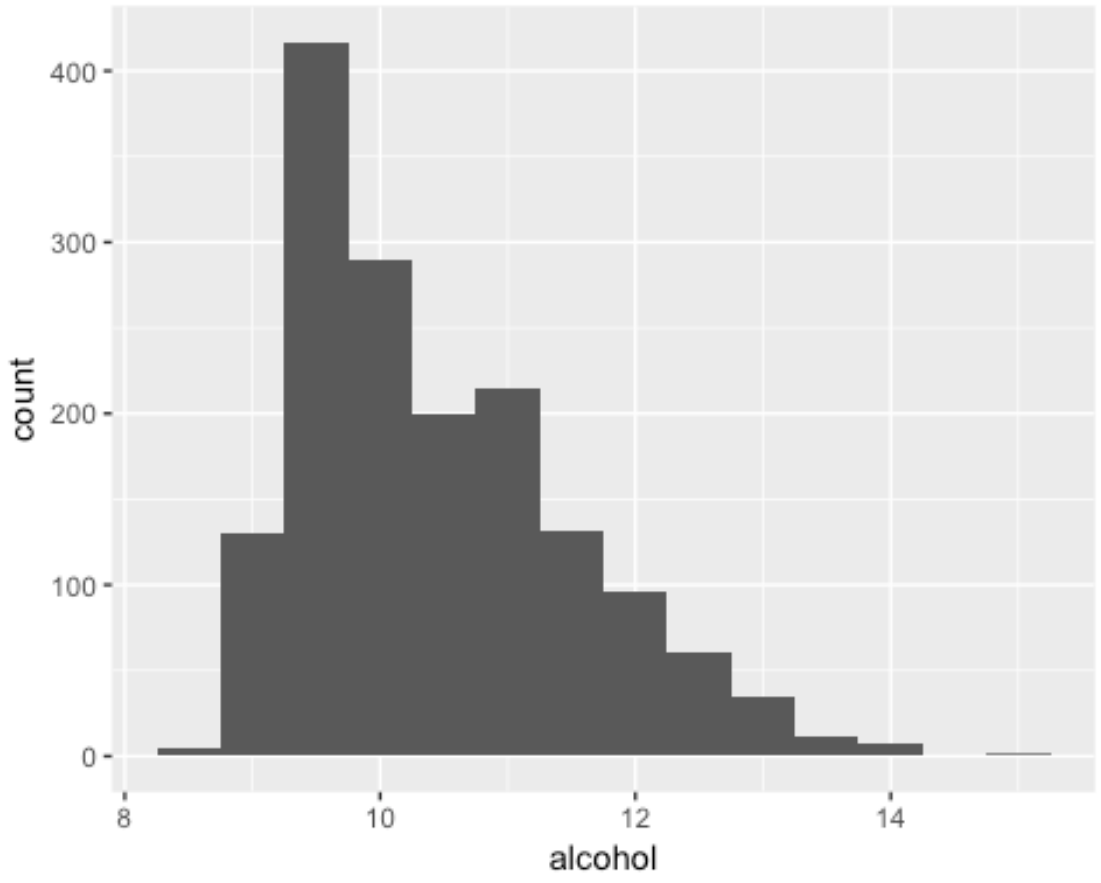

600
400
200
0
count
4
8
12
16
residual.sugar



Summary of the residual sugar.

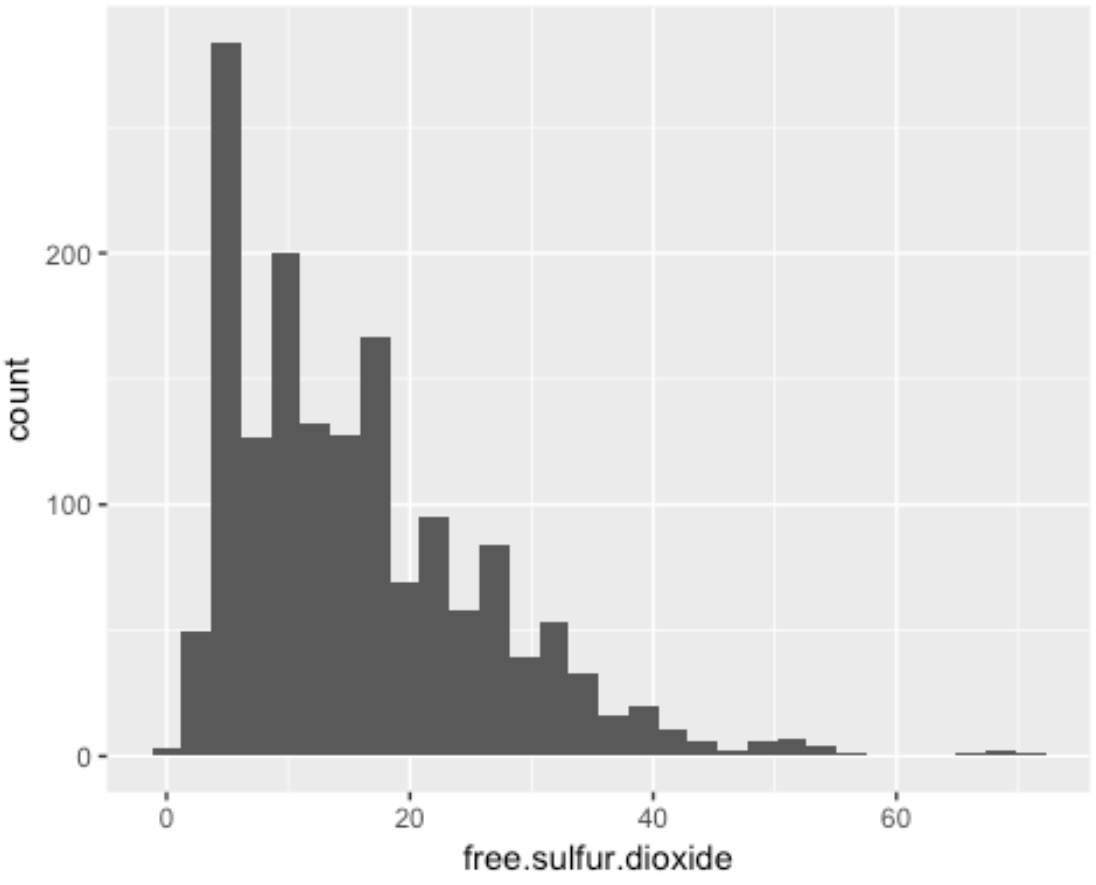
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

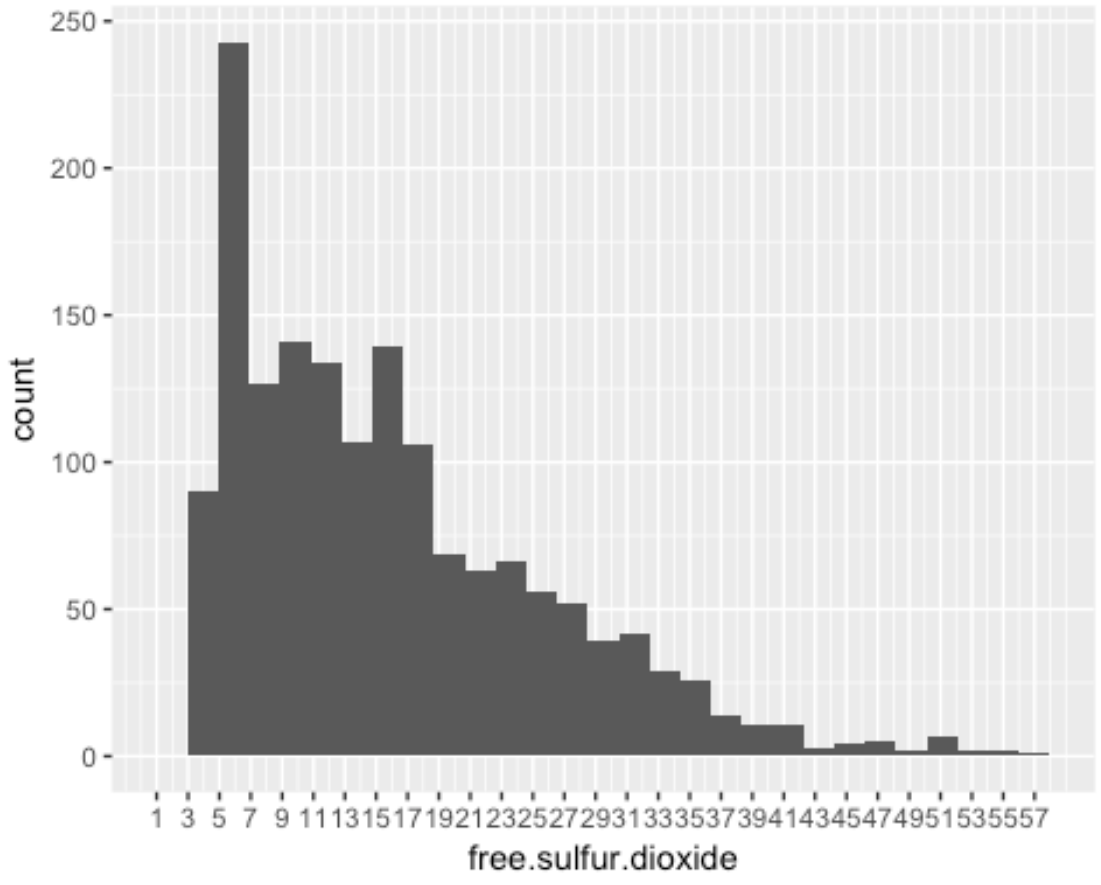
The distribution of residual sugar is skewed to the right. Based on the summary and the histogram, the distribution of the sugar content is relatively divided the min value is 0.9 and the max value is 15.5. There are only few values after the 8.9. So i used the log10 by using the scale_x_log10. There are wins which contains less sugar.




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

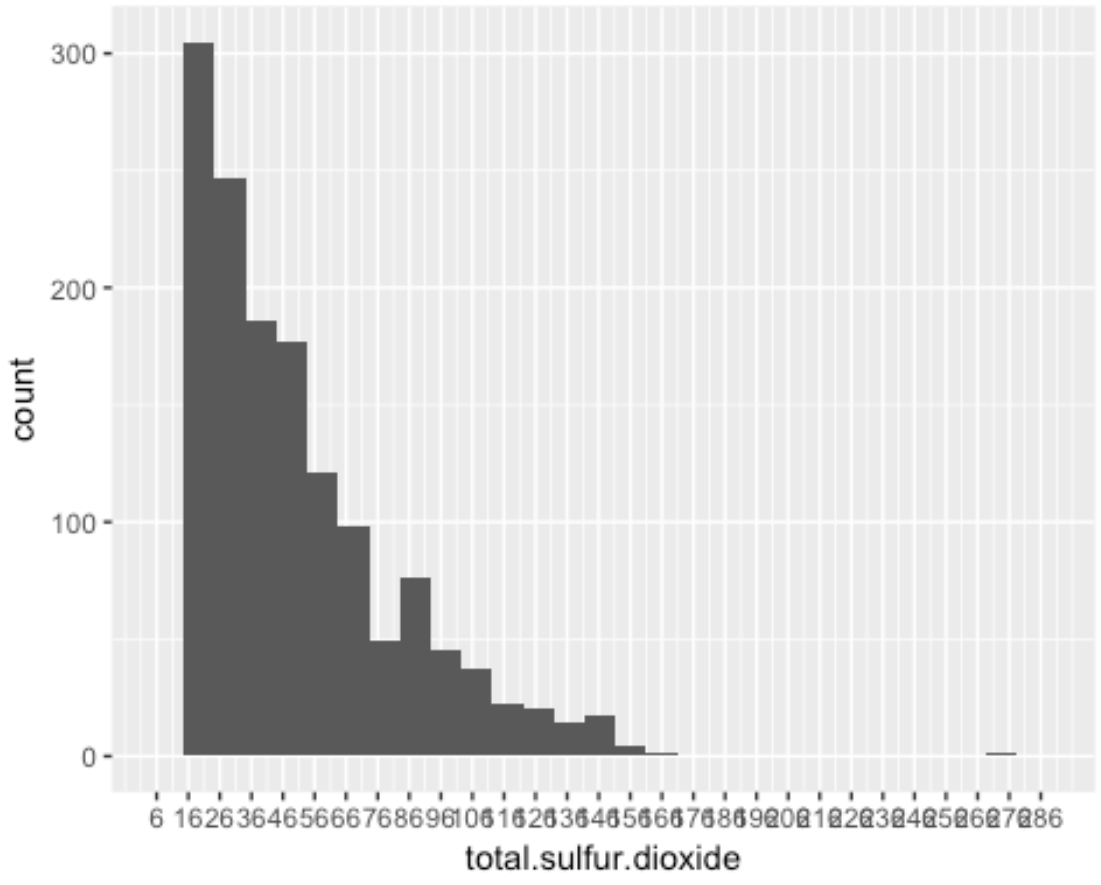
There is a right skewed distribution of alcohol plot, all most all the wine have their alcohol content between 8.5 and 12. Obviously there is no wine without alcohol. The max value is 14.90 and min value is 8.40. I set the binwidth to the 0.5 so that we have a better look.

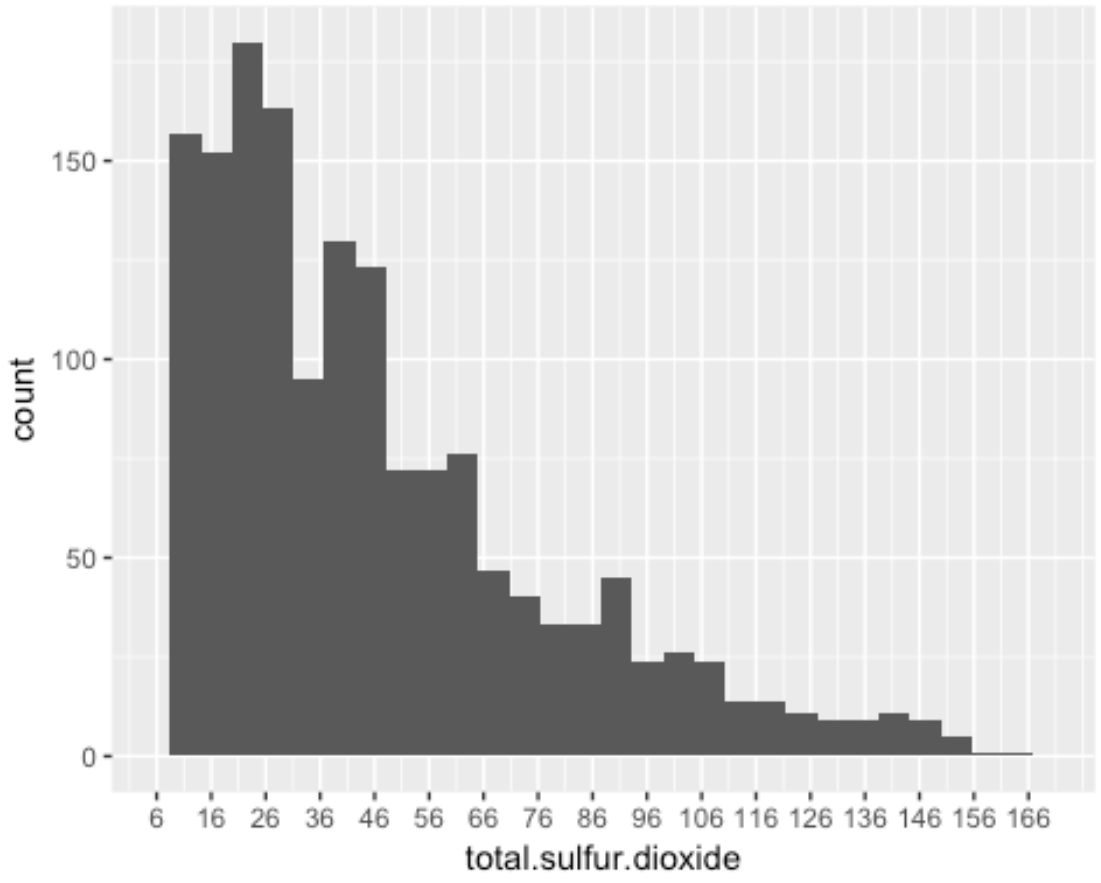

count
200
100
0
0
20
40
60
free.sulfur.dioxide




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

According to the plot we have the highest count of 3 to 19 free.sulfur.dioxide. There is a right-skewed distribution of a free.sulfur.dioxide plot. After trimming the x-axis 1 to 58 in the second plot to have a better look on data.


300
200
100
0
count
6 16
total.sulfur.dioxide




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

After trimming the plot 6 to 170 with help of the first plot. As I notice that in this plot min value is 6 and max value is 289 but the mean value is 46.47 which is very low and median is 38. That's why I decided to trim this plot.

Univariate Analysis

What is the structure of your dataset?

After completing my univariate exploration on the Wine dataset:- The dataset contains 13 variable and 1599 observations. This dataset is based on all the ingredients and their quantities which also effect on the quality of the wine. It data about different features of red wines, be made

up of pH value, rates of acidity, sugar content and quality rates. All the variables are a quantitative value which shows quantity or quality of the product. All the variables are represented by an integer and numeric data type. I check for any missing or 'NA' values in the observation but there not any.

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

While working with this wine dataset I am interested in seeing that which variable or ingredient is affecting more on wine quality. In order to find out, I go through the all the variables and try to find out the possible correlation between wine quality and other components, like residual, alcohol content, a rate of acidity and density of the wine.

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

According to my research, I found that a good wine contains the ideal amount of sugar which do affect the quality of the wine if it has lower or higher sugar levels. The good quality wines have lower sugar content and the lower quality wine have higher sugar contents. I will also look through the other variables also and analyze the correlation between alcohol content, density, and quality. https://www.youtube.com/watch?v=4UJmB3EqhU0

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

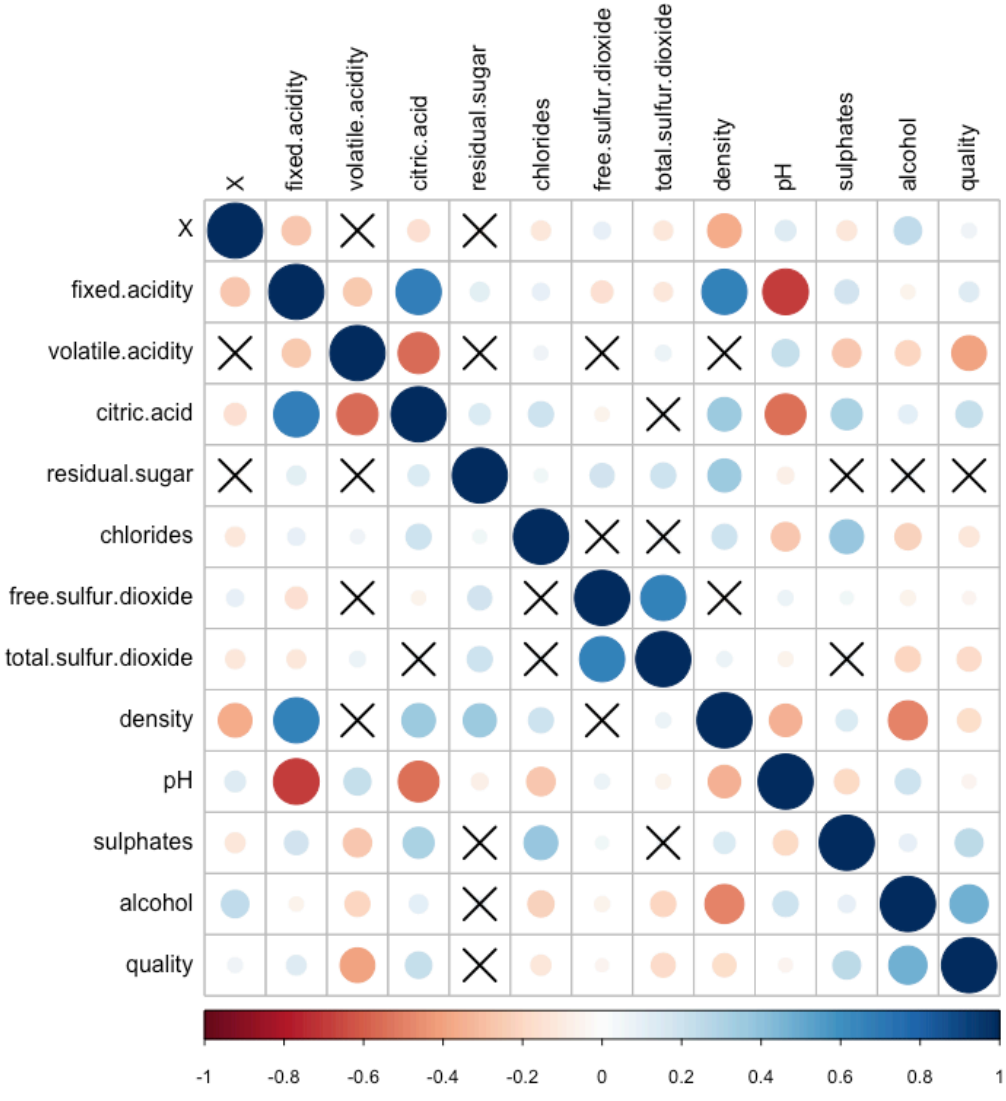
I did not create any new variables in the database. I think the quality variable is enough for doing an investigation. It would be great if I have a price variable so, it may help me to make analyze and predictions.

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?

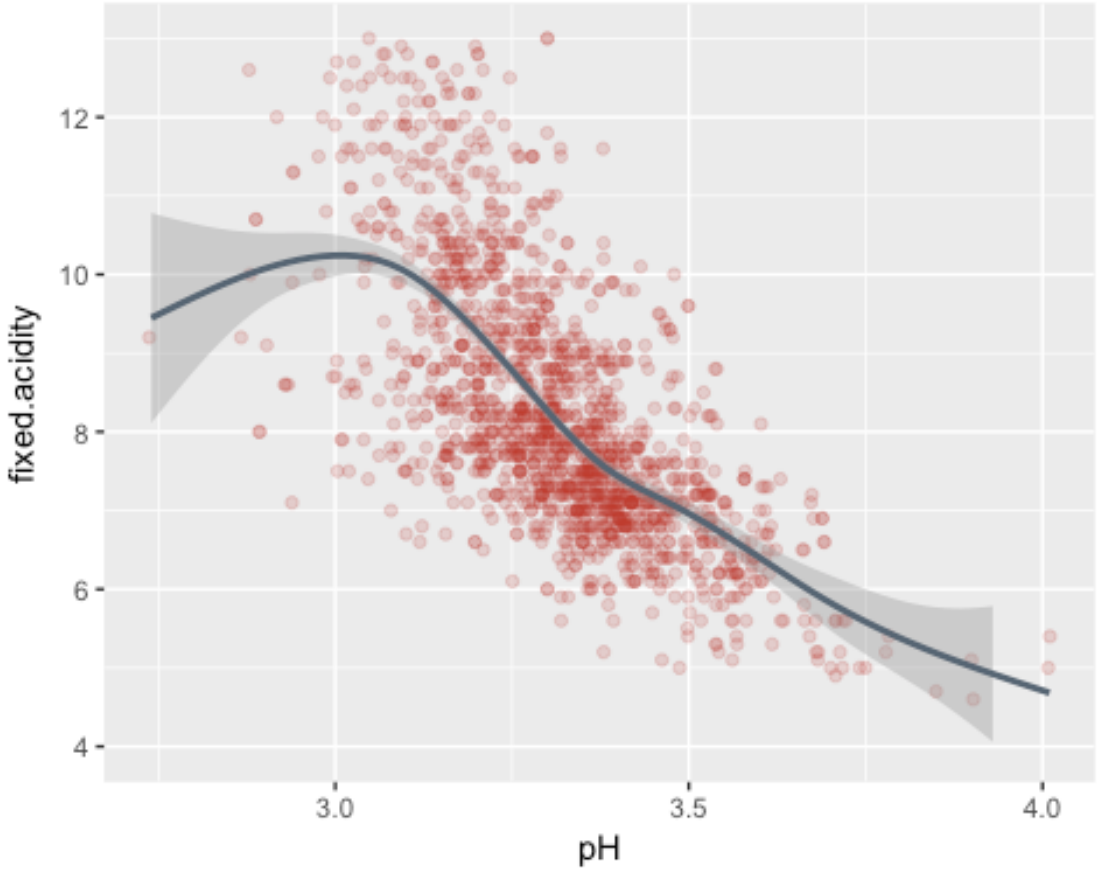
The Red Wine Dataset is provided by Udacity and it's already a tidy dataset so, I didn't need to make any type of adjustment or perform any cleaning operation.

Bivariate Plots Section

By plotting all the variable of the dataset in one place so that I can get a quick visualization about which value is more relevant in predicting the wine quality. According to this matrix, presume alcohol content to correlate with quality. There are more connections between the variables. I set 0.95 confidence level and a 0.05 significance and marked the corresponding variables with a black X.



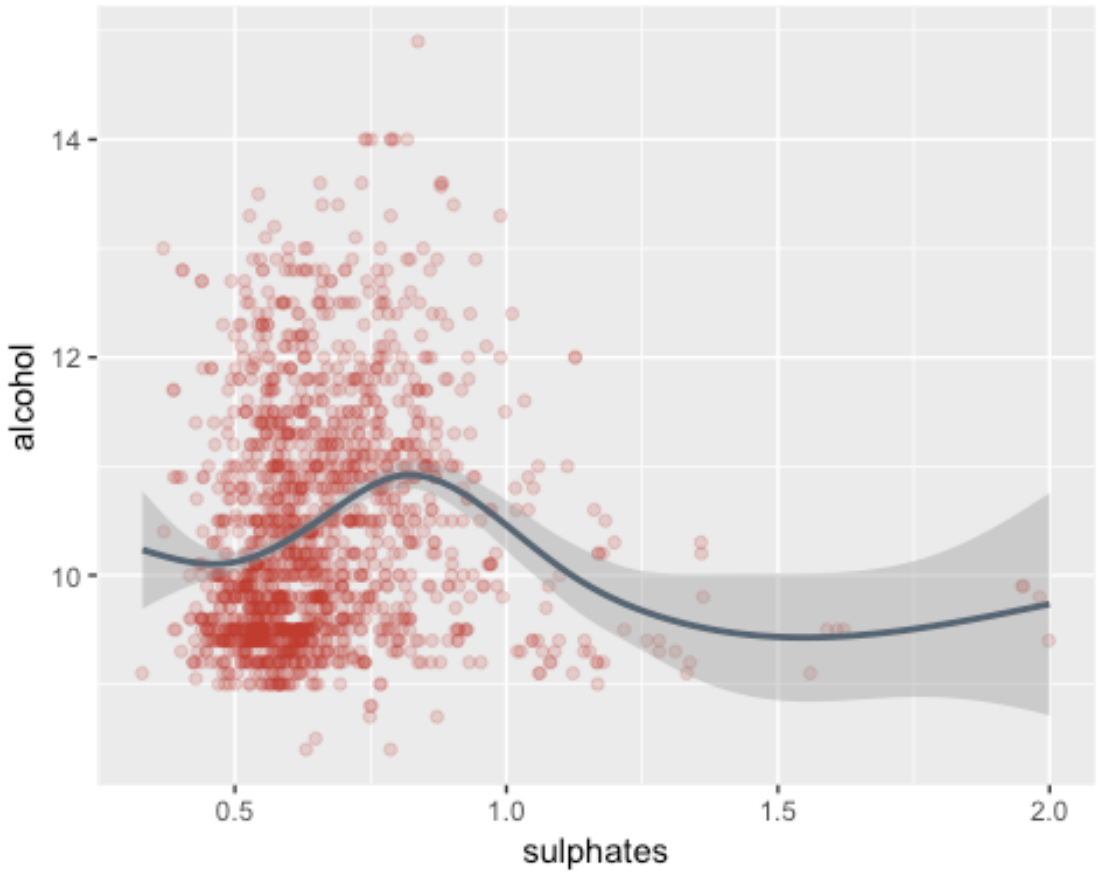
This scatters plot shows the relationship between the fixed acidity and pH. According to the plot visualization, pH has a relationship between the fixed acidity because when pH increases there is also increase in fixed acidity and when fixed decrease there is also decrease in pH. I used the stat_smooth() function to represent the correlation. I found this article very useful for my exploring about the fixed acidity and pH variables in the wine dataset.:- http://winefolly.com/review/understanding-acidity-in-wine/



The correlation between the two variables and found a moderate negative relationship (-0.6829782).

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

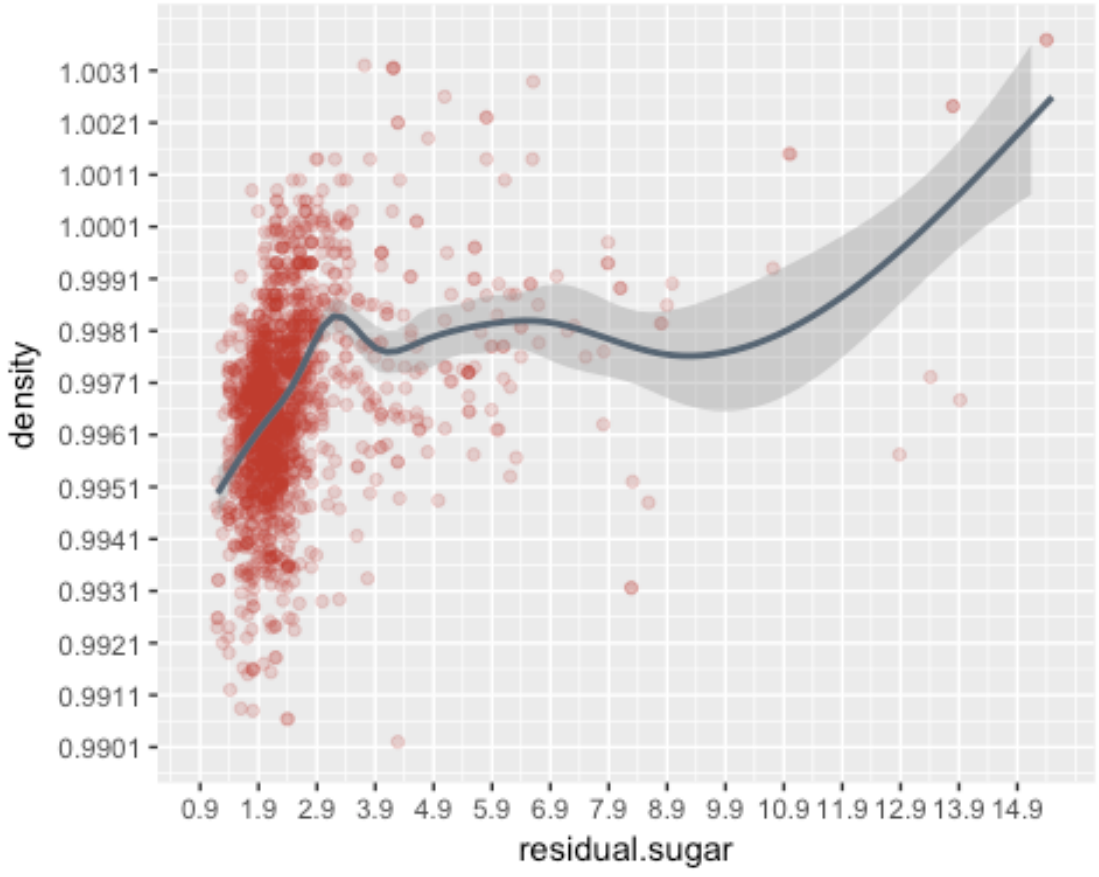
In this scatter plot we have sulphates and alcohol variables. As we can see there is no relationship between them. Correlation coefficient number, their correlation is just 0.09359475.



The correlation between the sulphates and alcohol variables and found a moderate relationship.

```
## [1] 0.09359475
```

In this plot both the variables is an important ingredient for a wine. If we use a little bit of our chemistry and plot we getting here we know the fact that more sugar means the more density. Here we have low sugar which means in most of the win we have very low sugar levels which means we may get a good quality of wine at a low sugar level. And also I got a strong positive correlation between these two variables which is ‘0.3552834’.

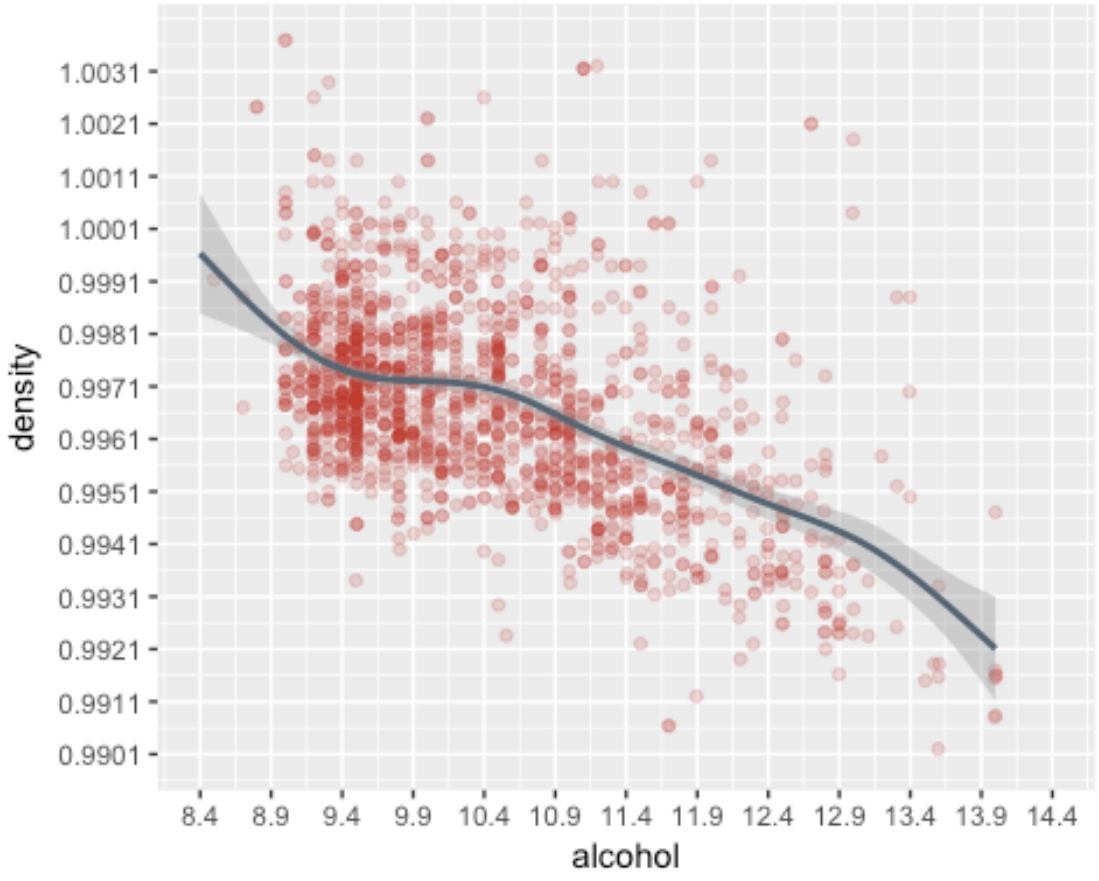



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

The correlation between the density and residualsugar variables and found a moderate relationship.

```
## [1] 0.3552834
```

As we know there would be no wine without the alcohol. Now the question is how much alcohol effect to the density of the wine. As we can see in the plot the density decreases with increase in the amount of alcohol in wine. Now we have to find which of the relationship is affecting more on the quality of the wine. I found the strong negative correlation coefficient between density and alcohol variables which is '-0.4961798'.

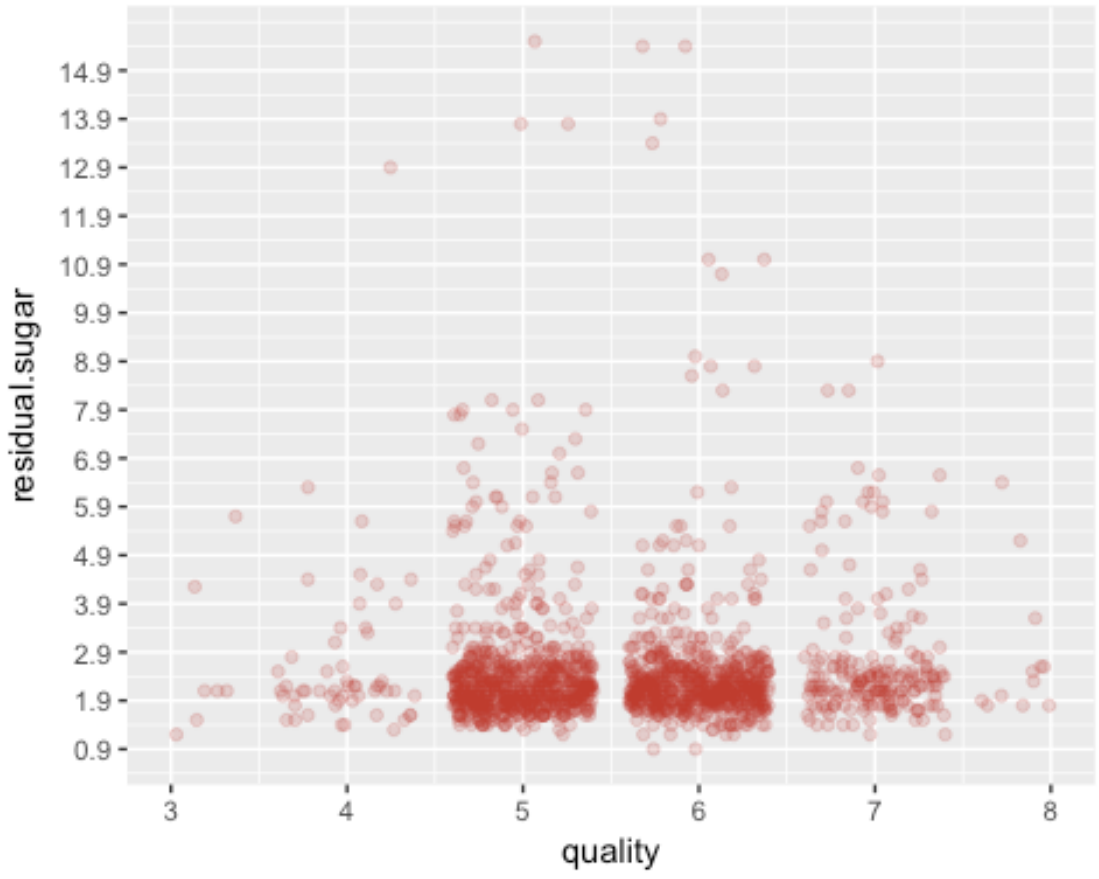



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

The correlation between the density and alcohol variables and found a moderate relationship.

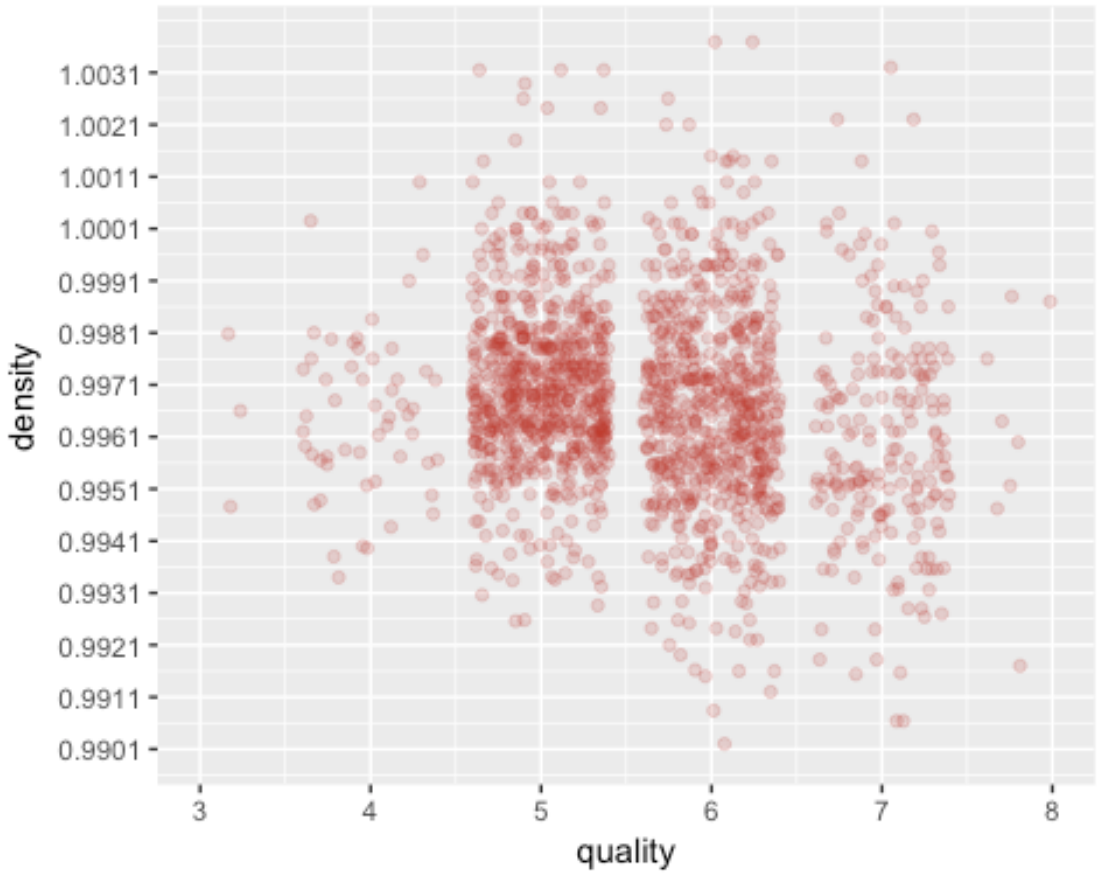
```
## [1] -0.4961798
```

I thought that sugar is a more important factor in the quality of the wine but it turns out to be wrong! According to the plot, the sugar I didn't find any relationship between them and also I got a very low correlation coefficient which is only '0.01373164'. This is not I was expecting so I have to try quality relationships with any other variables or ingredients.




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

In this plot of between wine quality and density variables, it turns out to be a negative relationship. The correlation coefficient between the two variables is ‘-0.1749192’. The best quality wine must have the lowest mean density.



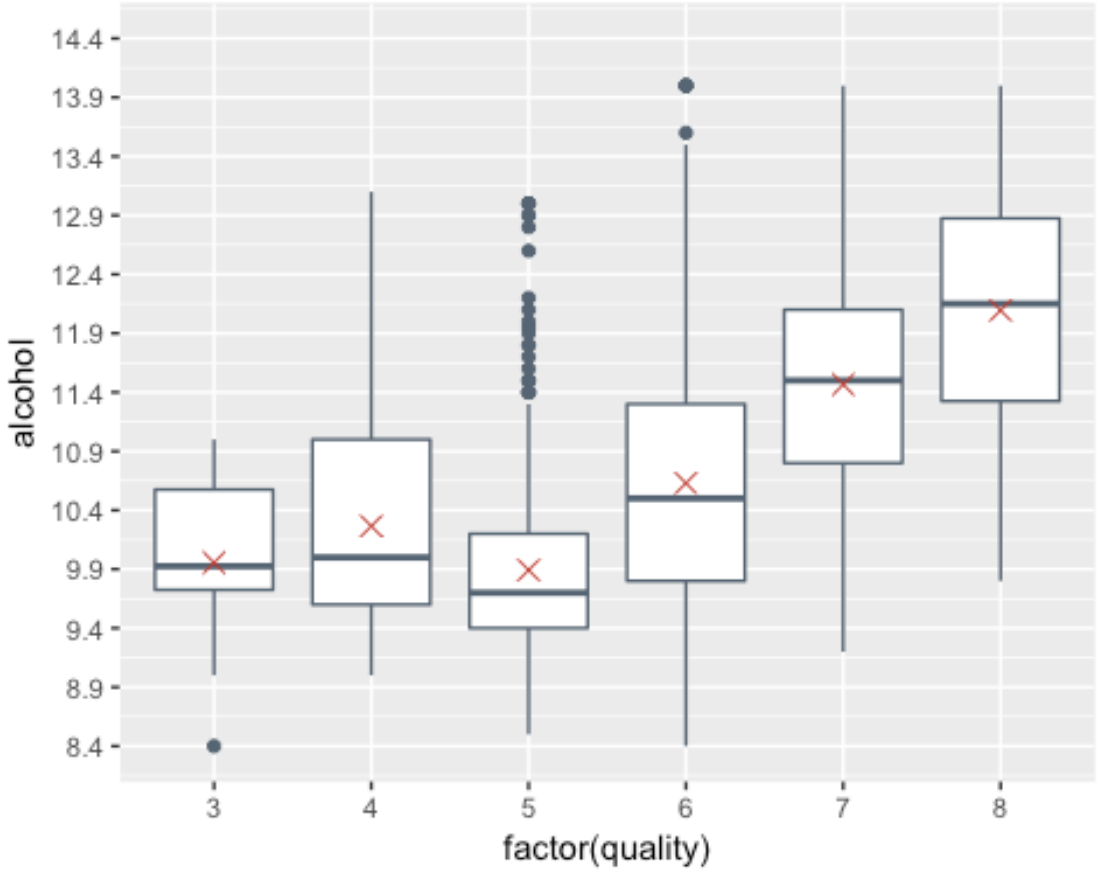
The correlation between the density and quality variables and found a moderate relationship.

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

summaryBy function to get the mean, min and max values for all the quality and density variables.

```
##   quality density.mean density.min density.max
## 1       3    0.9974640     0.99471     1.00080
## 2       4    0.9965425     0.99340     1.00100
## 3       5    0.9971036     0.99256     1.00315
## 4       6    0.9966151     0.99007     1.00369
## 5       7    0.9961043     0.99064     1.00320
## 6       8    0.9952122     0.99080     0.99880
```

In this particular plot, I used boxplots to visualize the relationship between the alcohol and quality variable. By calculating the correlation coefficient we got a strong positive moderate correlation coefficient which is '0.4761663'. According to this plot, we can see the higher quality wines have a higher meadian alcohol content above 11 to 12. As low-quality wines have their median alcohol content around 10. By looking through the visualization, the worse wines are their max around 12 to 13 while the better around 14.



summaryBy function to get the mean, min and max values for all the quality and alcohol variables.

```
##   quality alcohol.mean alcohol.min alcohol.max
## 1       3     9.955000         8.4        11.0
## 2       4    10.265094         9.0        13.1
## 3       5     9.899706         8.5        14.9
## 4       6    10.629519         8.4        14.0
```

```
## 5       7    11.465913          9.2         14.0
## 6       8    12.094444          9.8         14.0
```

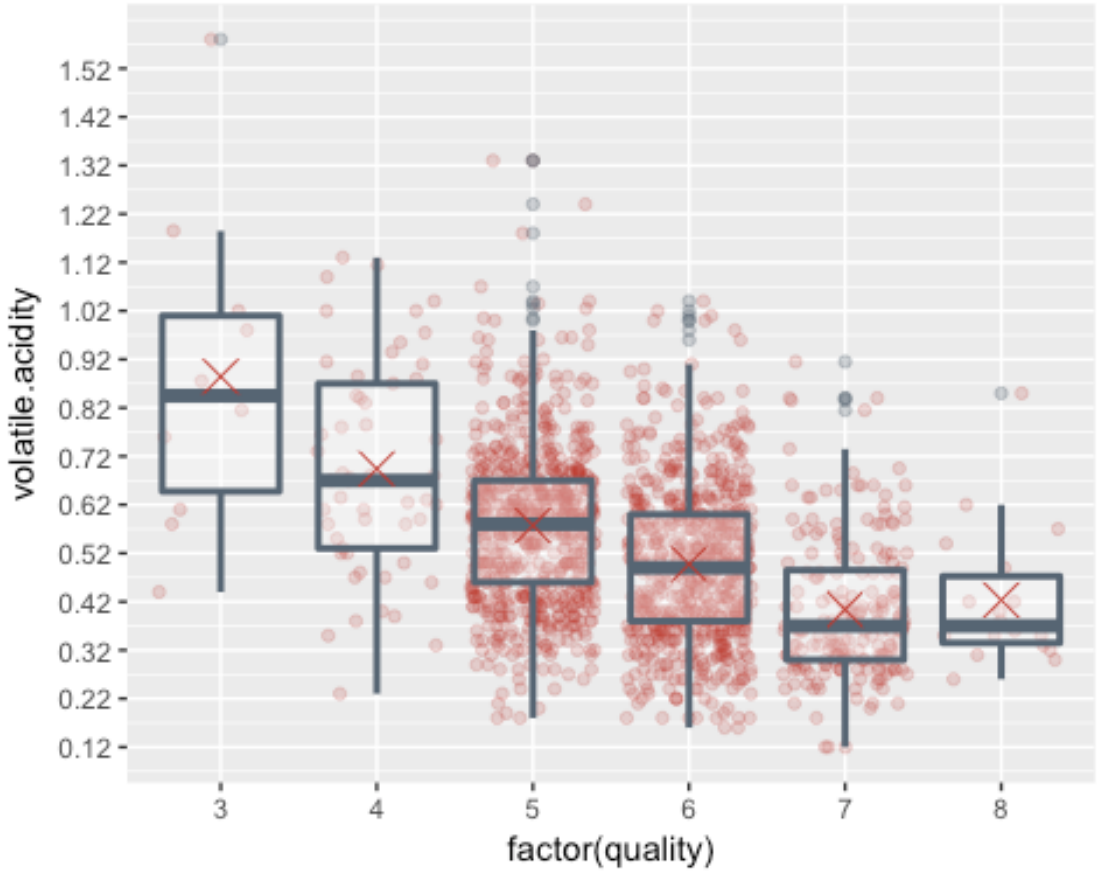
The correlation between the alcohol and quality variables and found a moderate relationship.

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

summaryBy function to get the mean values for all the quality and density variables.

```
##   quality alcohol.mean
## 1       3     9.955000
## 2       4    10.265094
## 3       5     9.899706
## 4       6    10.629519
## 5       7    11.465913
## 6       8    12.094444
```

For making this plot I used a jittering plot to visualize the distribution of the values contained by the volatile acidity and the quality of the win. As given in the database description that too high volatile acidity can lead to an unpleasant, vinegar test which affects the test of the wine quality. The correlation coefficient is not that much strong which supposed to be, it's in moderate negative value '-0.3905578'.
By looking in the plot we can see the low-quality wines have the high volatile acidity.



The correlation between the alcohol and quality variables and found a moderate relationship.

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

summaryBy function to get the mean, min and max values for all the quality and density variables.

```
##   quality volatile.acidity.mean volatile.acidity.min
volatile.acidity.max
## 1       3              0.8845000                  0.44
1.580
## 2       4              0.6939623                  0.23
1.130
## 3       5              0.5770411                  0.18
1.330
## 4       6              0.4974843                  0.16
```

```
1.040
## 5        7              0.4039196                    0.12
0.915
## 6        8              0.4233333                    0.26
0.850
```

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?

The most important variable I have an interest in was the quality, I did compare the quality variable with all the other variable and try to find the strong relationship which causes impact the quality of a wine. Throughout the results, I found that alcohol and density have a strong relationship with quality of a wine. - Residual sugar is supposed to have a relationship with quality as I thought but after plotting the graph, it seems that sugar having a very weak relationship with quality. The mean values are between 1 and 3. And the low level of sugar wine is having high quality. - Alcohol has a strong relationship with the quality variable. According to the plot the high quality of wine content the more amount of alcohol. The highest quality of wine content above 12 while the lowest quality of wine contents below 9.while - Density has the weak relationship with quality, it has the -0.1749192 correlation coefficient which is almost ignoring values. - Volatile acidity has the strongest relationship with quality because we got the strong correlation coefficient. The worst wines have a higher volatile acidity proportion.

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

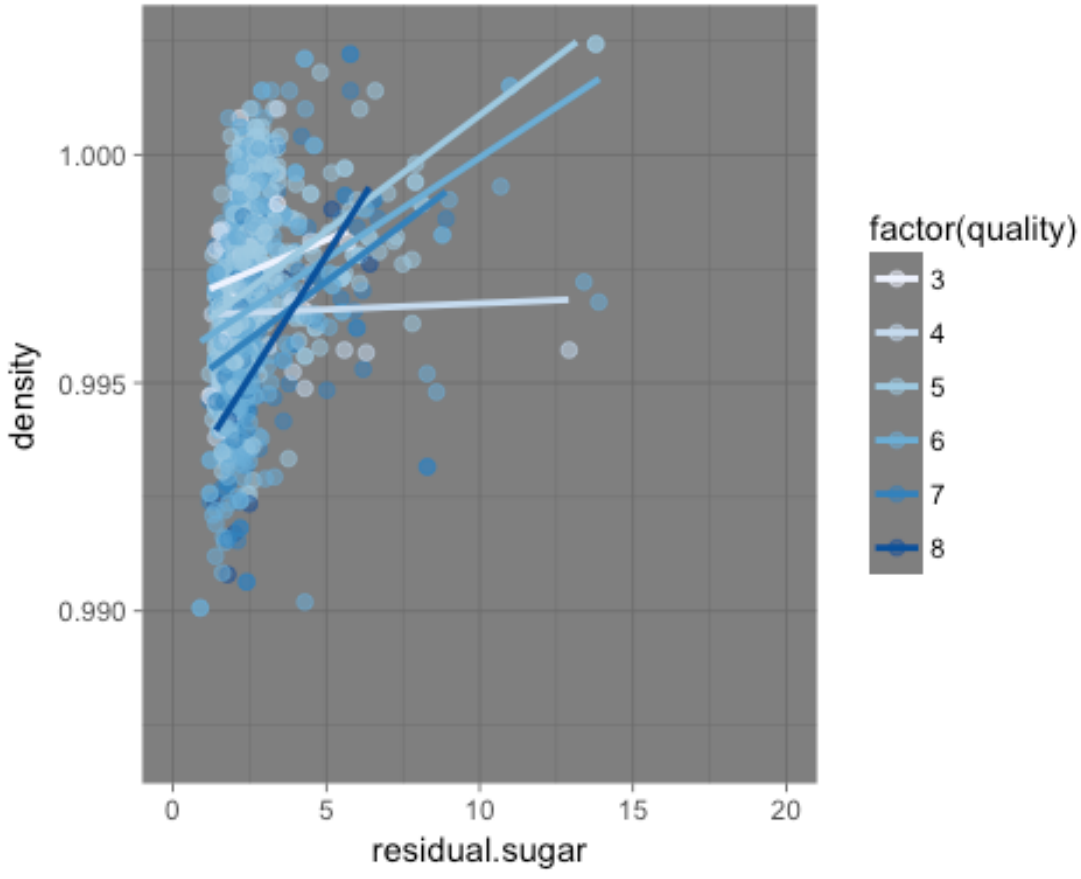
I observe a strong correlation between sugar and density, the more sugar in the wine have the more density and the less density wine have the less sugar. There is also the same relationship in the alcohol and density variables the more alcohol have less density and more density have less alcohol.

What was the strongest relationship you found?

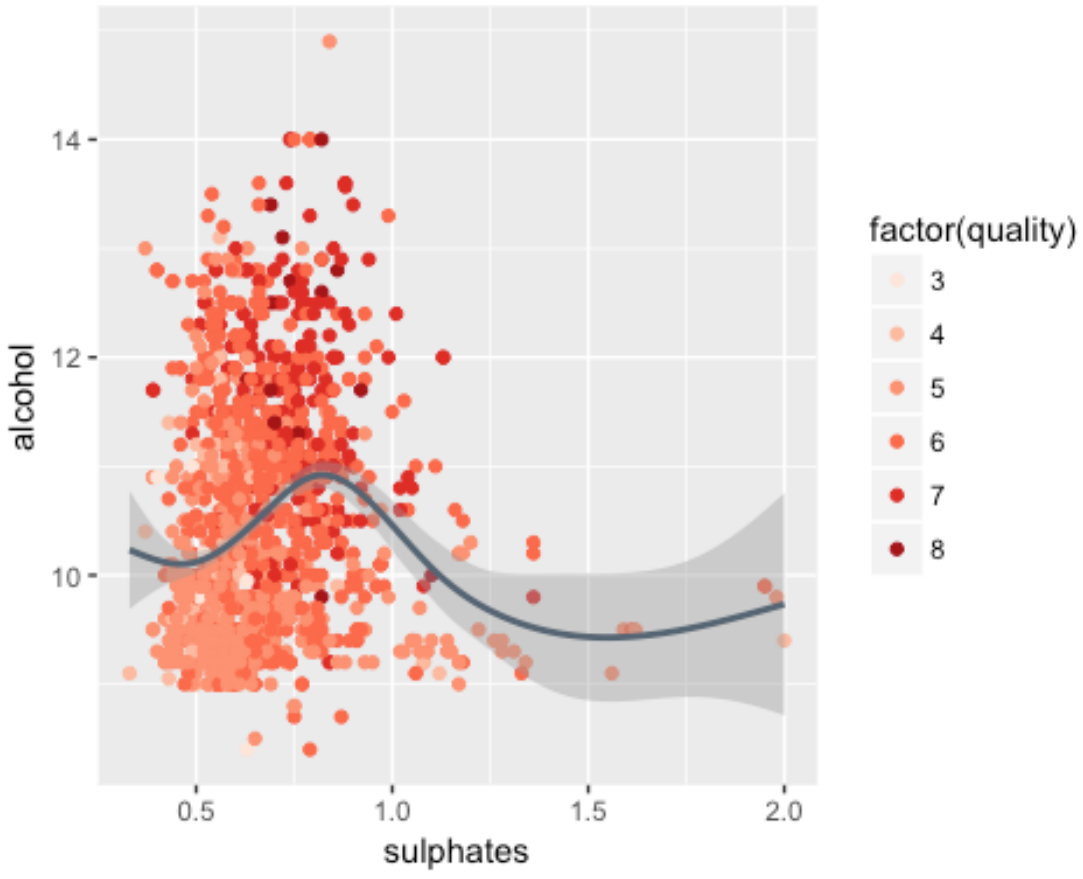
The strongest relationship I found was between alcohol content and quality which is 0.4761663

Multivariate Plots Section

In this plot, I created the scatter plot to have a better view between residual sugar, density. After grouped and colored the dots with the quality variables in order to show the correlation correlated scatter plot where dots, representing the better wines can be found on the bottom of the plot.

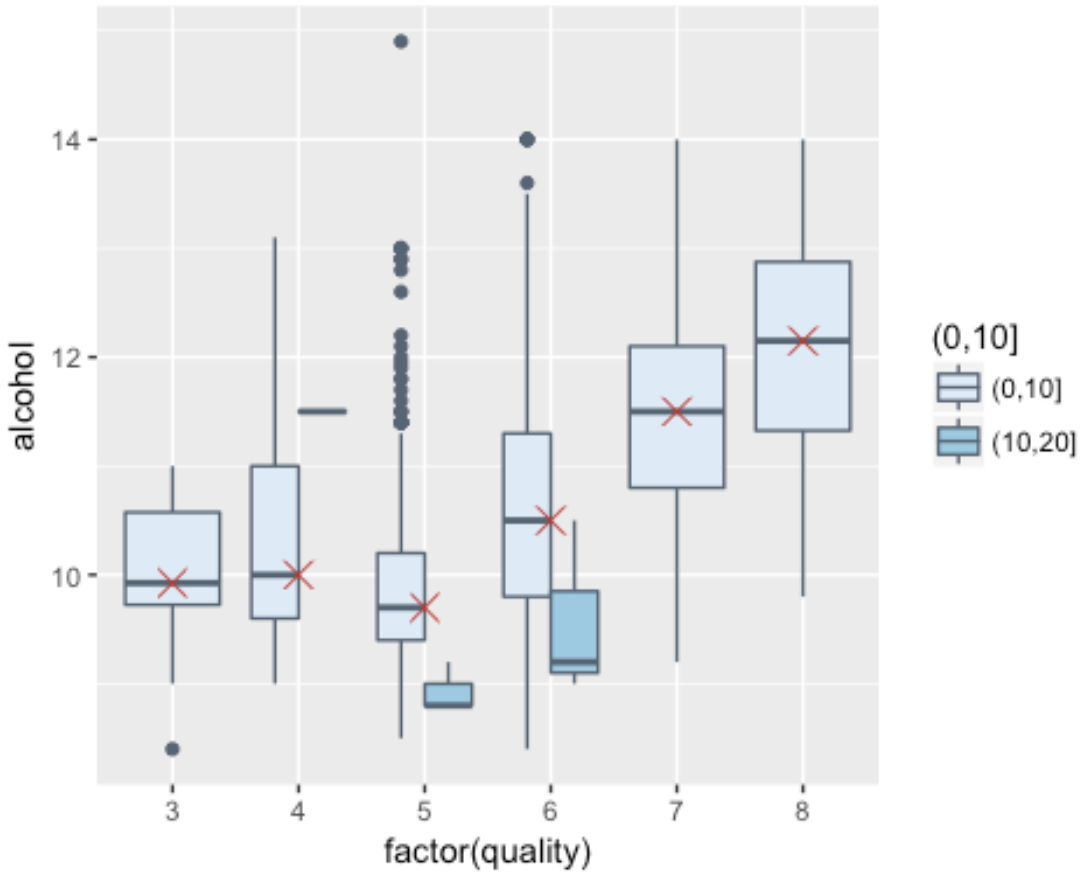


This plot is about the relationship between pH and sulphate, colored by a quality variable. To show the interesting trend for a nonlinear interpreted, I used stat_smooth() function in the plot. According to the color, the high quality of wine contents the high amount of alcohols but not the high sulphate, we can also notice the strong trend of distributionon of quality.



In this plot, first of all, I made the bucket of residual sugar to have a better look at the plot with quality and alcohol. Then I designed a boxplot to reflect the relationship between quality, alcohol and residual sugar. These are the important variables for my visualization. With the help of this, we can differentiate between alcohol content by quality. The better the wine the higher the median alcohol content is for each level of quality (the medians are marked by a red cross). After going through the plot according to me the high quality of wine having the more amount of alcohol and the less amount of sugar.

```
##    (0,10]   (10,20]   (20,40] (40,65.8]
##      1588        11          0         0
```



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?

Density, alcohol and residual sugar these are the in the 1st plot established each other as when I added color to explore some of the data from it. The same is applicable when I plotted the last plot with alcohol, residual sugar to quality.

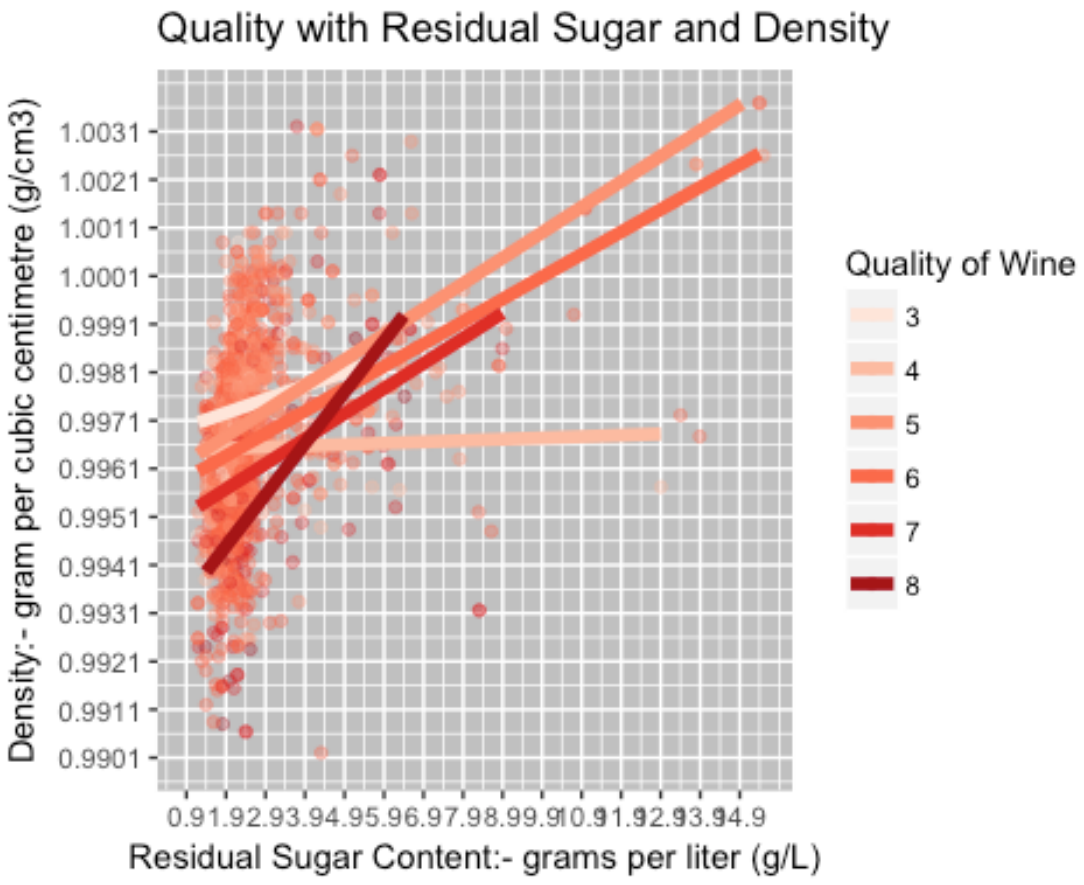
Were there any interesting or surprising interactions between features?

There are some interesting and interactions between features like alcohol and sulphate. I tried to find out about the relationship between this two variable but I didn't find anything interesting, I will continue my research on it. The alcohol and residual sugar having a very interesting relationship. While exploring the wine data I get to know about these things make so much difference for a wine quality. To be honest, until now I thought that the quality of wine only depends upon how much old they are.

Final Plots and Summary

Plot One

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 3 rows containing missing values (geom_smooth).
```

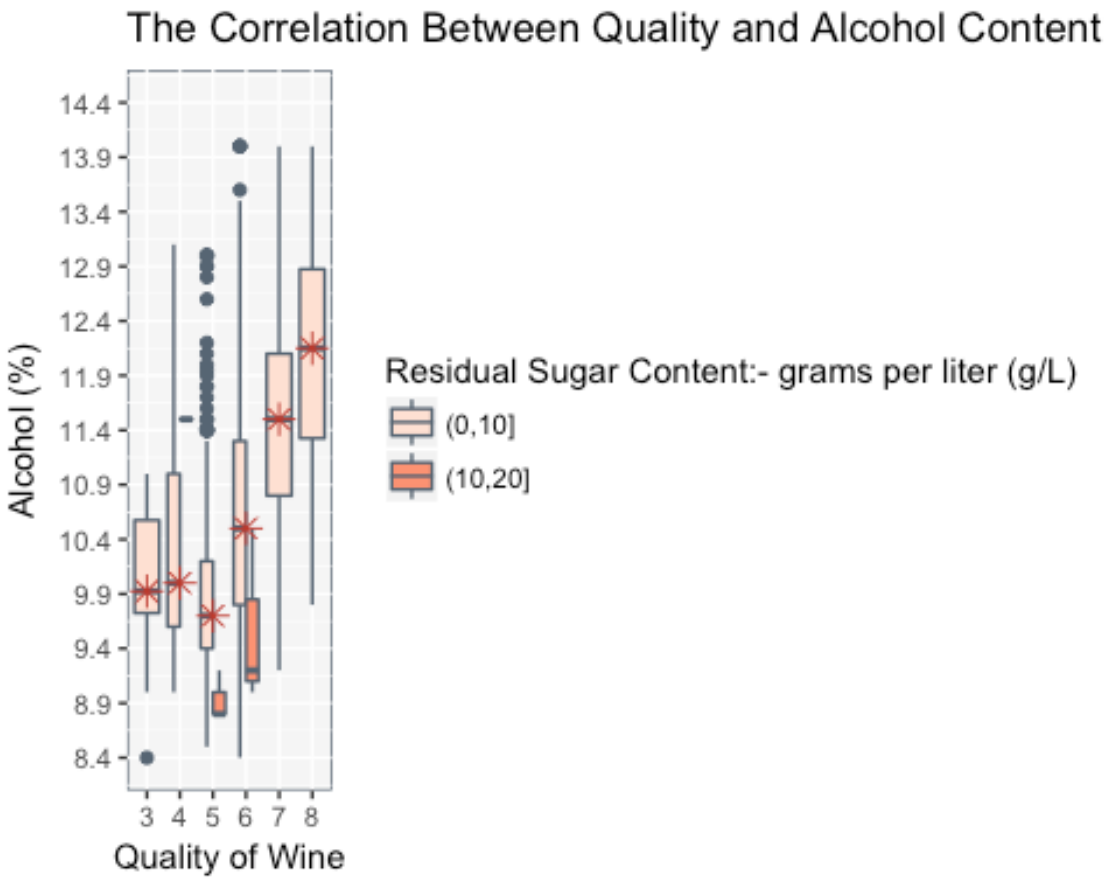


Description One

It shows that the relationship between residual sugar and density are strongly correlated. The quality is marked as a color also manage to correlate with density. The lower the density is, the higher the quality

Plot Two

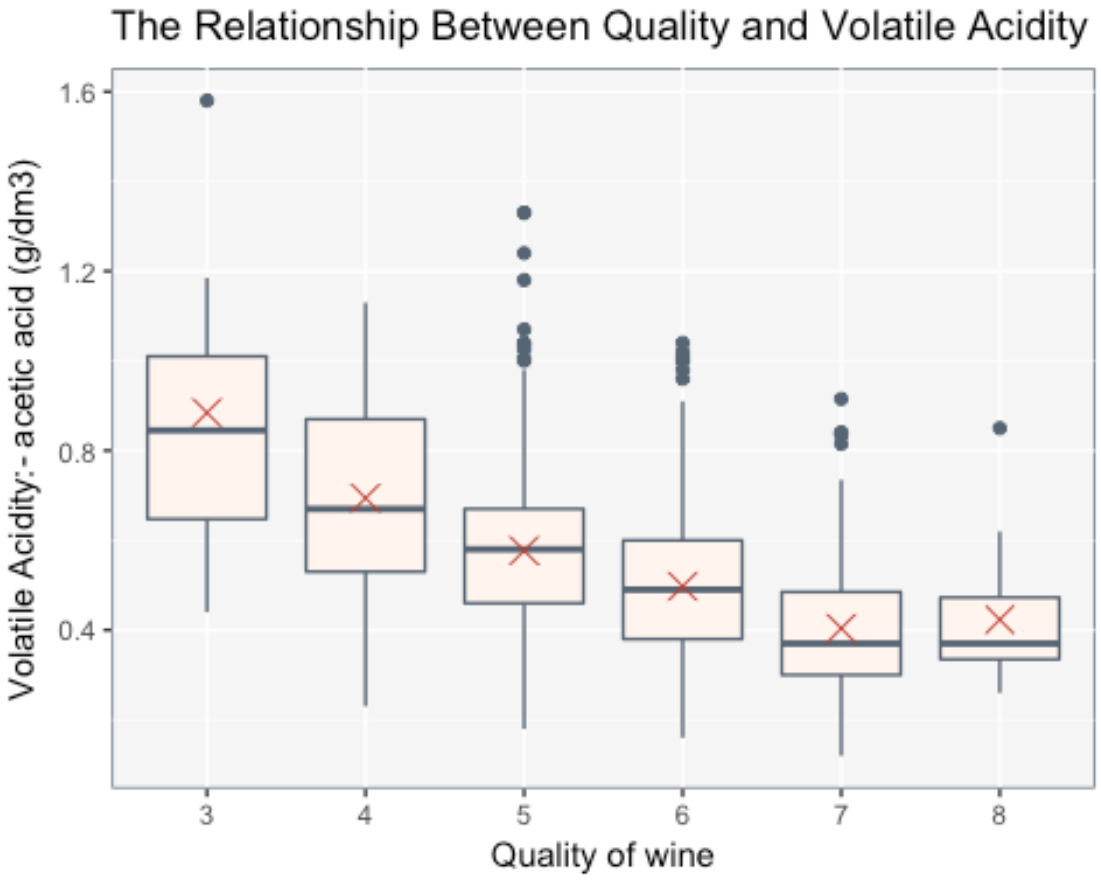
```
## Warning: Removed 1 rows containing non-finite values
(stat_boxplot).

## Warning: Removed 1 rows containing non-finite values
(stat_summary).
```



Description Two

In this plot, I used previously bucket value and the box to represent the very important factor which affects the wine quality which is alcohol. With the help of this graph, it is cleared that the good quality wine has a high level of alcohol and lower level of residual sugar. In the beginning, we found that sugar is not directly affecting the quality of the wine but sugar and alcohol and alcohol have strong correlation and alcohol indicates wine quality.

Plot Three



Description Three

In this plot, we can see that wine with higher volatile acidity having a low quality of wine and the low volatile acidity having the high-quality wine. When wine having high volatile acidity it may change the test of the wine and make it worst. As we can see the worst wines (quality 3 and 4) have higher volatile acidity while the better ones (quality 5 and 6) have lower values. This inclination does not correspond to the best wines (quality 7 and 8).

Reflection

From the very beginning, I wanted to work on the quality variable and try to make sense with other variables. I found some variables which tend to have a strong relationship with quality of a wine, like alcohol and density. They having a very strong relationship with quality of a wine, and there are some more variables which affect the quality of wine indirectly example residual sugar - when its high, density will be also high and high density makes the low quality of wines. Likewise, there are few more variables are there which takes an effect on the quality of wine indirectly. To get more information, I made some histograms for the separate variables. After getting the general idea about all the variables. In bivariate analysis, I tried to get a better knowledge of the factors which affect the quality. By taking every variable one by one and add some functions for exploring the relationship of the variables for one another and to know how they influence with wine quality. While doing that I show very weak correlation between sugar and quality which I didn't expected. But not directly but also indirectly sugar affect the quality of the wine. At the ending of my visualization, I highlighted some of my interesting plots and concepts to reflect the main components that influence the quality of a wine. It was my very first project in R language. Throughout the project, I was very excited and always doing research on my dataset that's the only reason I visualize it in a proper way. I think there are still so many things to explore like sulphates and alcohol etc. A good thing that I did not need to perform any cleaning operation or another kind of adjusting, it saves lots of time. R is a tricky language to pick up, but once I understood the patterns and syntax, I enjoyed the level of control it gave me over the visualizations. I found really challenging and exciting to work on this dataset.