

Exploratory Data Analysis on Red Wine Quality

Getting to Know the Database

Viewing the first 6 rows in the dataset to see how the data looks like.

```
##   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
```

There are 11 chemicals in the table that can influence the quality of red wine. For instance acidity level, sugar, PH level, alcohol, etc. The last column in the table specifies the quality (score between 0 to 10) of the wine based on all the chemicals.

Now let's get some info on different types of data in the dataset. According to the table below, **row number** and **quality** are of type integer, while the rest of the variables are numeric.

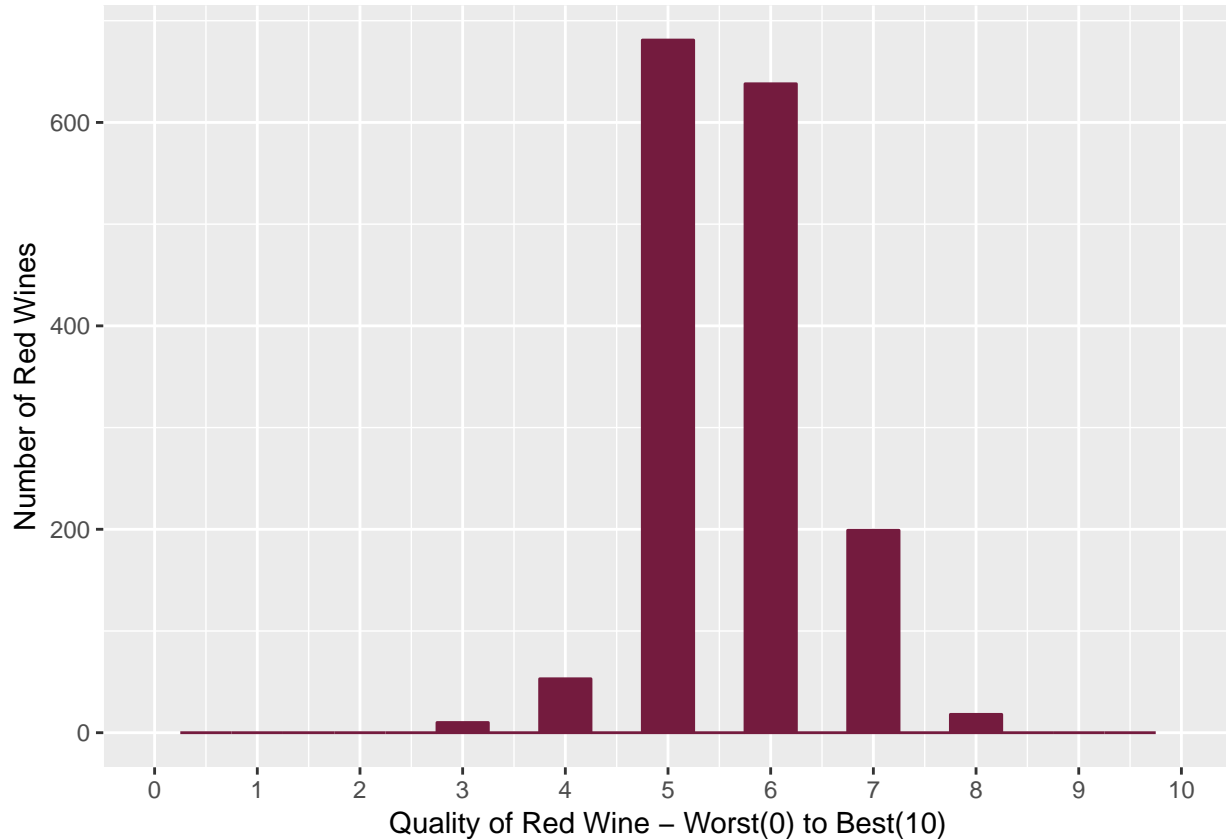
```
## 'data.frame':   1599 obs. of  13 variables:
## $ X              : int  1 2 3 4 5 6 7 8 9 10 ...
## $ fixed.acidity   : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
## $ volatile.acidity : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
## $ citric.acid     : num  0 0 0.04 0.56 0 0 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 ...
```

I would like to remove the **X** vector from the dataset as it only holds row numbers in the table which is no use for me during my investigation.

Univariate Plots Section

I create some basic histograms for each chemical property in the table.

First I like to see how the **quality** property is distributed in the table. Quality of red wine is one of the main point of interest for me in this dataset, as I'd like to see what properties in red wine can influence the quality.



Based on the plot the majority of the red wines -about more than 1200- have either 5 or 6 ranking for their quality. About 200 red wines have quality ranking as 7, but the number is not as significant as 5 and 6. The rest are either 3,4 or 8.

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

The average quality of red wines in the database is 5.63, with the maximum ranking set to 8 and the minimum set to 3. Therefor, there is no red wine in the database with a quality of 10 or 0.

Adding a New Column 'Rating' to My DataFrame

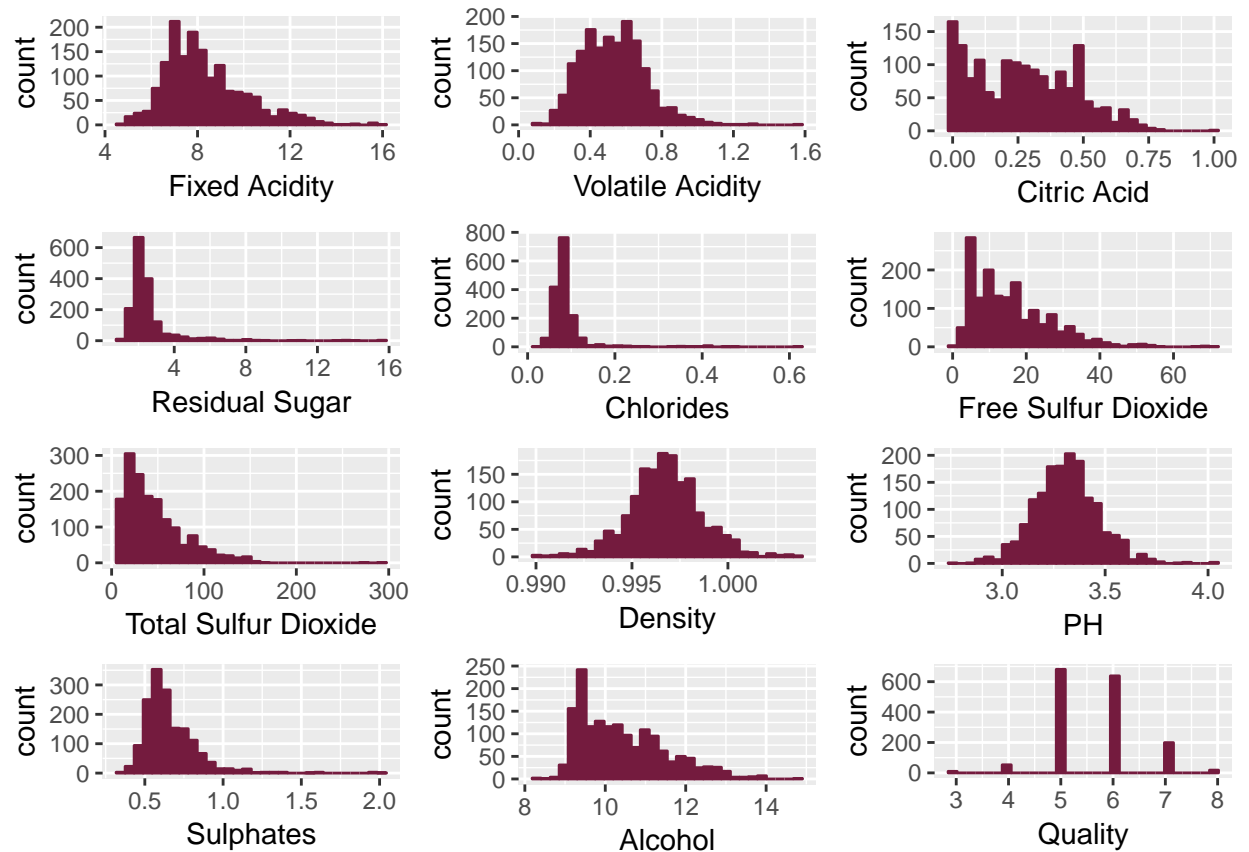
Since the quality rating is from 0-10, I like to create a rating system, from 1-10 (as there are no wines with rating zero in the data), and categories these numbers into ratings like this: 1-2 : Poor 3-4 : Below Average 5-6 : Average 7-8 : Above Average 9-10 : Excellent

The first 20 rows of my dataset for the **rating** column now looks like this:

```
## [1] "Average"      "Average"      "Average"      "Average"
## [5] "Average"      "Average"      "Average"      "Above Average"
## [9] "Above Average" "Average"      "Average"      "Average"
## [13] "Average"      "Average"      "Average"      "Average"
```

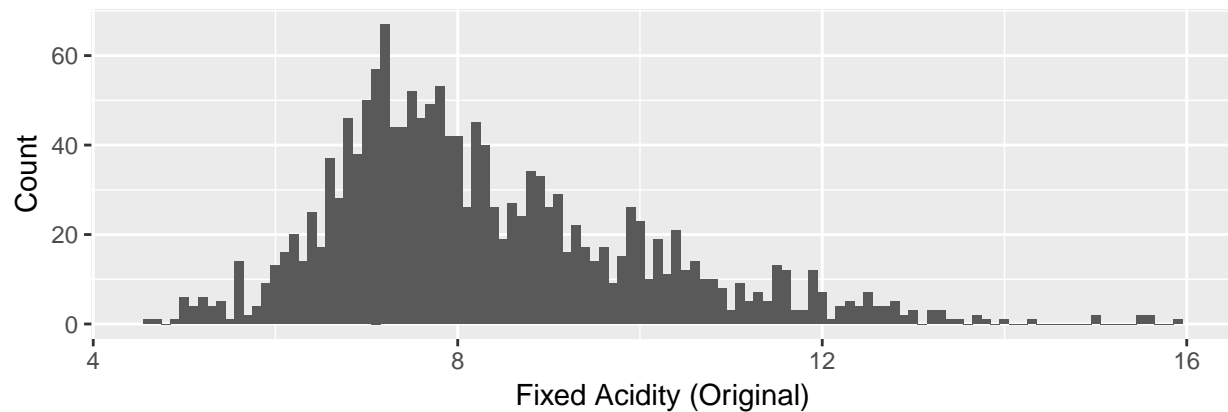
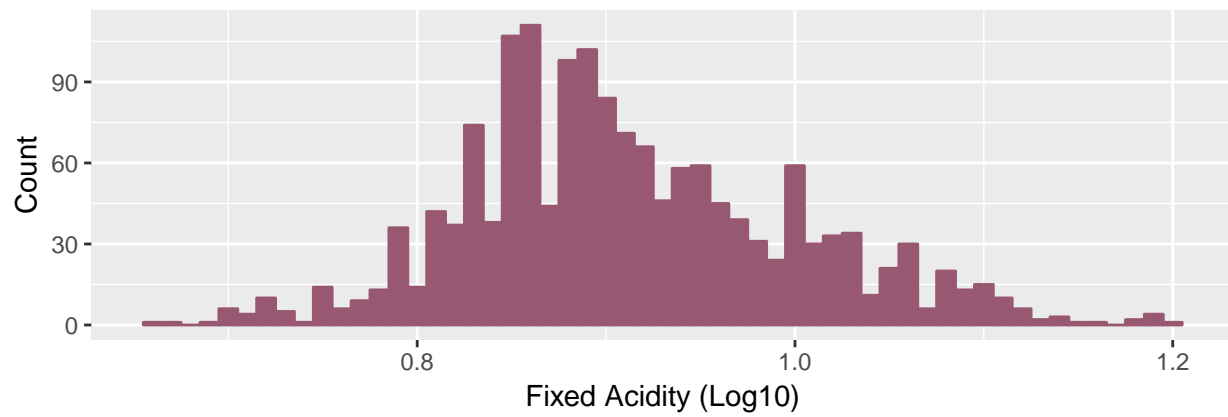
```
## [17] "Above Average" "Average"          "Below Average" "Average"
```

After looking at the quality distribution, I'd like to get a histogram of all the chemical properties - as well as quality - in the table. These histograms can also help me in finding out if any of values in the plots need to be changed to e.g. log10 for a smoother distribution.



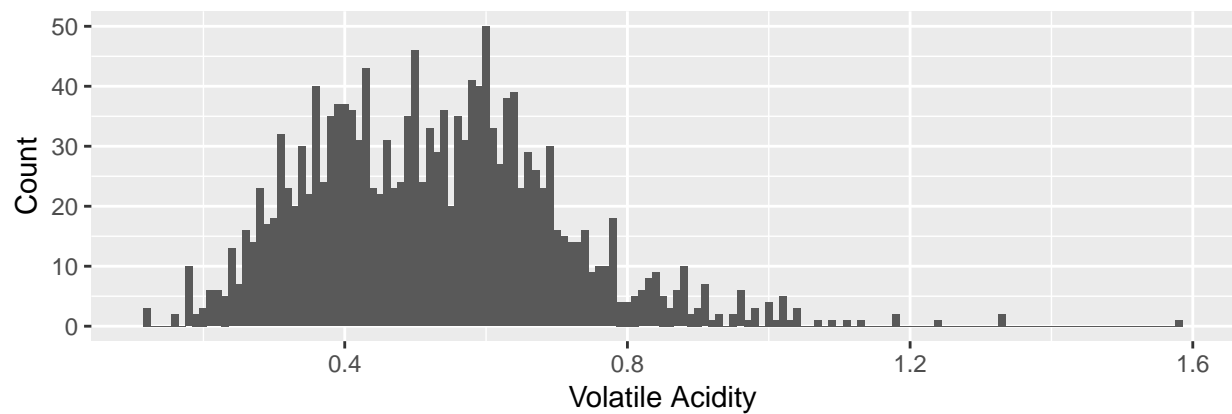
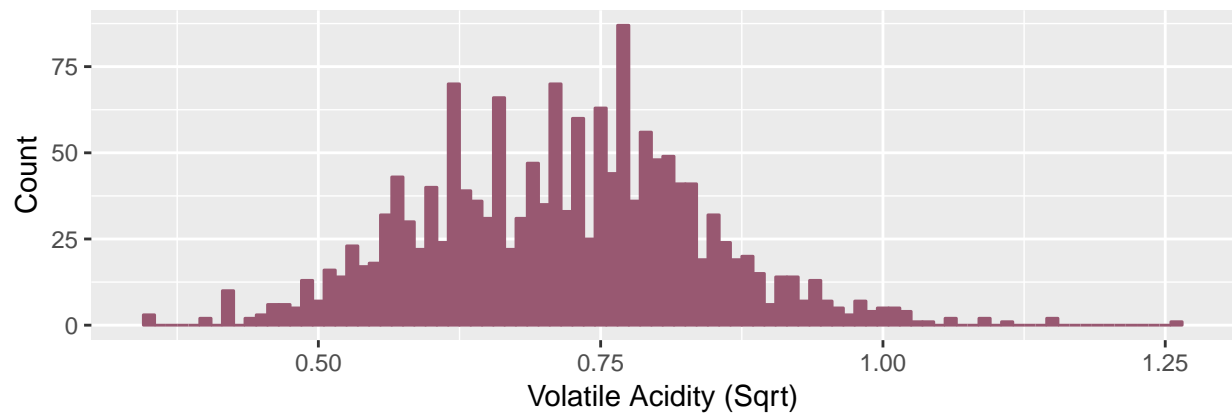
Fixed Acidity, Volatile Acidity and Citric Acid

Fixed Acidity distribution has most of its values from 5-12, with a few more than 12, and some outliers around 16. I apply the log10 function to its x-axis to make the distribution look normal.



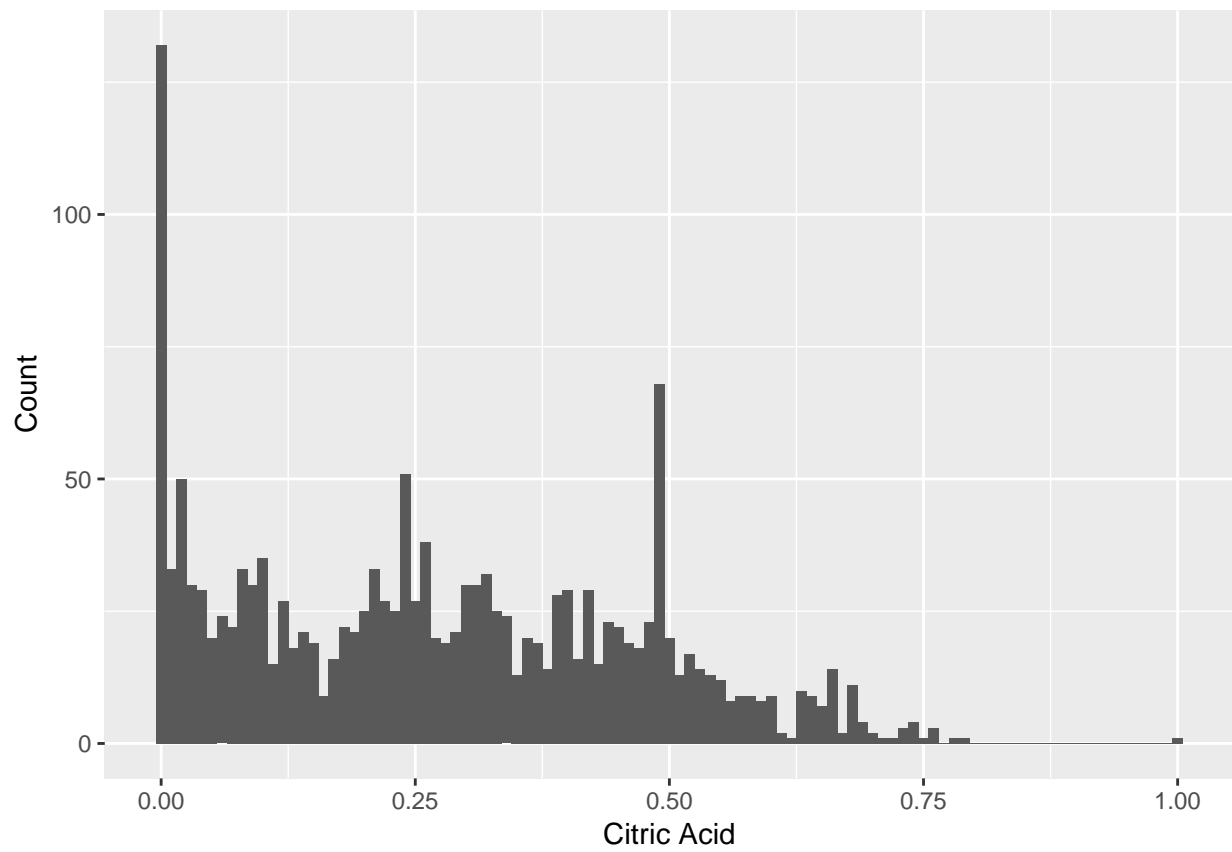
Volatile Acidity distribution has most of its values between 0 and 0.8, with a few going to 1, and then outlier after 1 to 1.6.

By getting the square root of the values, I can change the distribution towards a normal one. However, the outliers still exist after 1.00



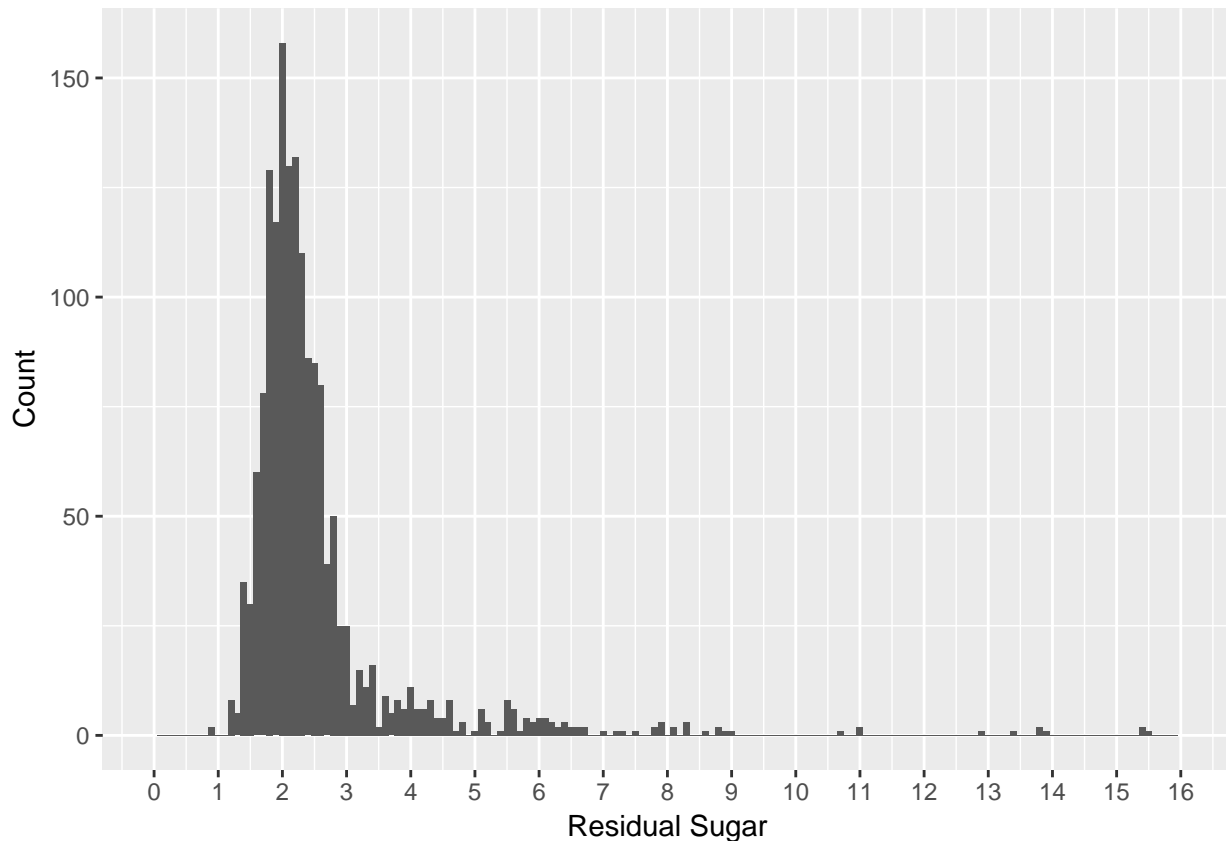
Citric Acid values are distributed along the plot on its x-axis, starting from zero (with its most count) to 0.75. Seems there are outliers for citric acid level 1.00

The plot has a its rise and fall throughout the x-axis. The most count contains citric acid level of zero. There are also two more visible peaks in count around 0.25 and 0.5 values.



Residual Sugar

Residual Sugar distribution has the majority of the count around 1.5-2.5. There are many outliers around 7 and onwards.



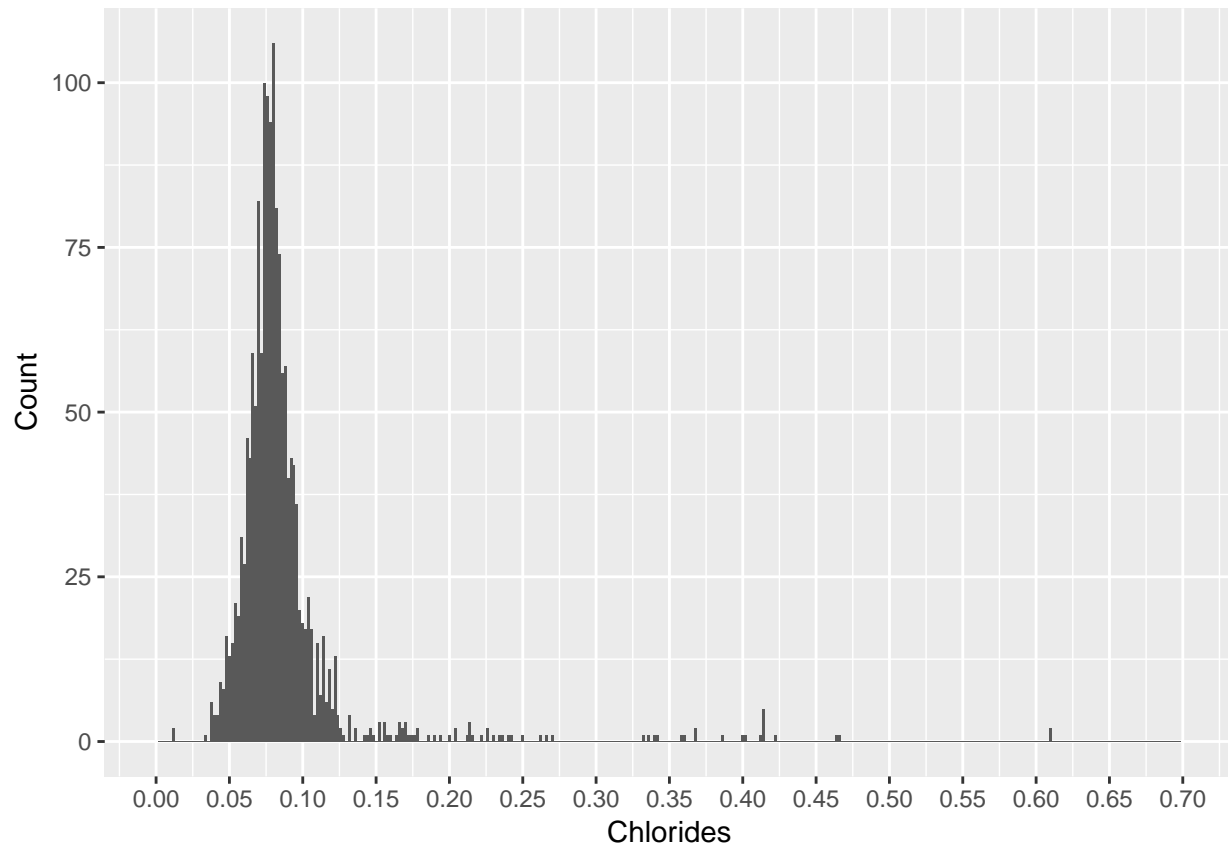
The outliers can have a direct influence on change the average value for residual sugar. Let's take a look at the summary for residual sugar to find out how the average is influenced.

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

According to the plot for residual sugar, with most of the values concentrated around 1.5-2.5, I was expecting the mean to fall around this range as well. However, the existence of the outliers was an interesting factor for me to check to see how much they can influence the mean. According to the summary, the mean still falls around 2.5 inspite of the outliers. This can make sense as the density of the plot is more significant in 1.5-2.5 range, than around the outliers.

Chlorides

Chlorides distribution has most of the values concentrated around 0.03-0.15. There are a few outliers especially after 0.35. I would like to check the summary to see how the mean value is influenced by the outliers.



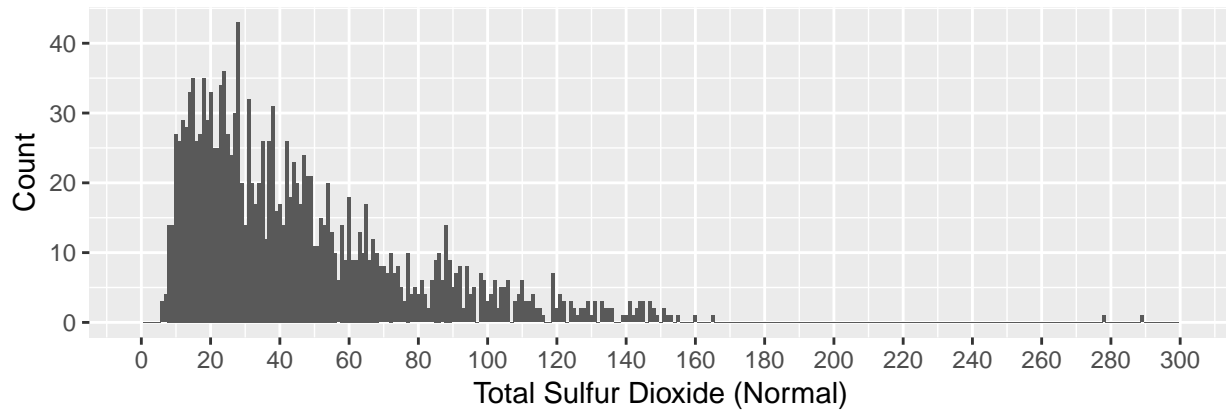
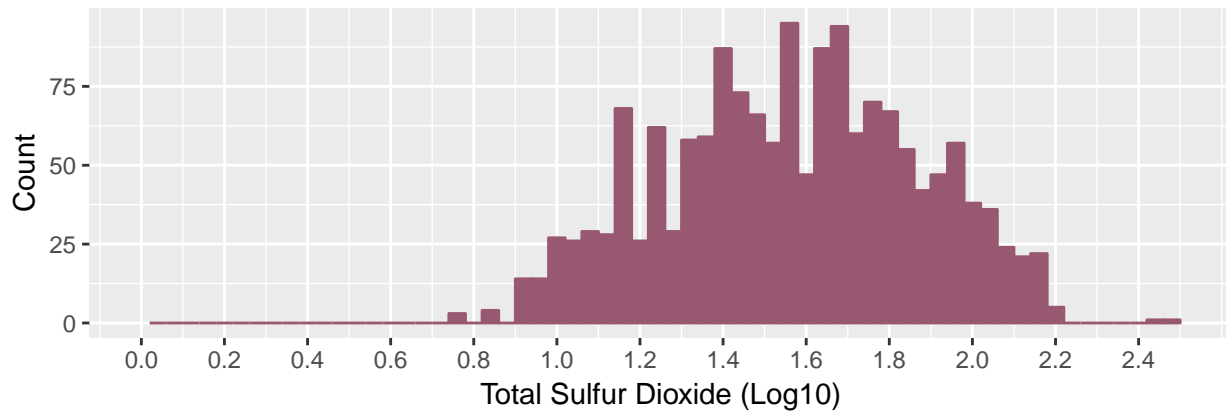
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

I expected that the mean value would fall somewhere between 0.05-0.10, and according to the summary, it's correct. The mean value of 0.08 is in my expected range.

Free Sulfur Dioxide and Total Sulfur Dioxide

Total Sulfur Dioxide distribution is skewed to left. The rise and fall of the distribution is not that much; however, there are a few noticeable peaks in count around 30-40, 60-80, and 80-90.

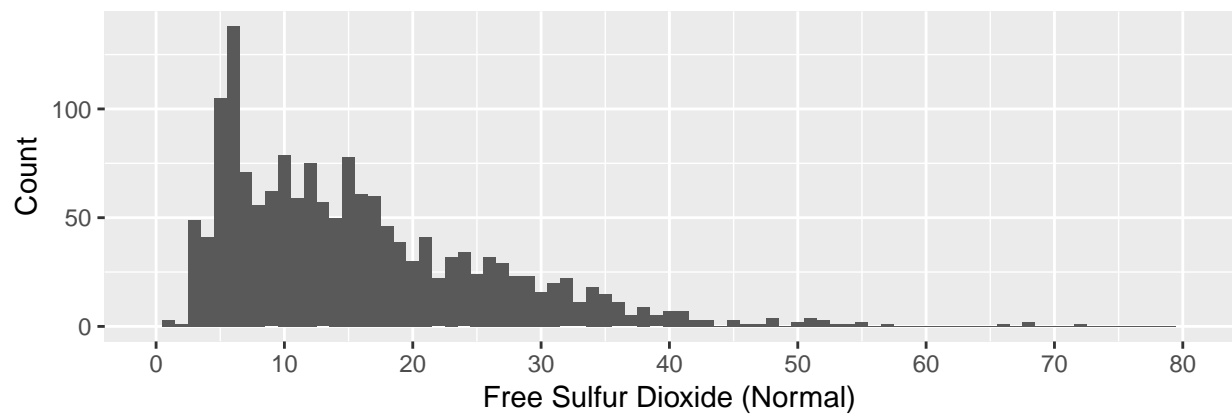
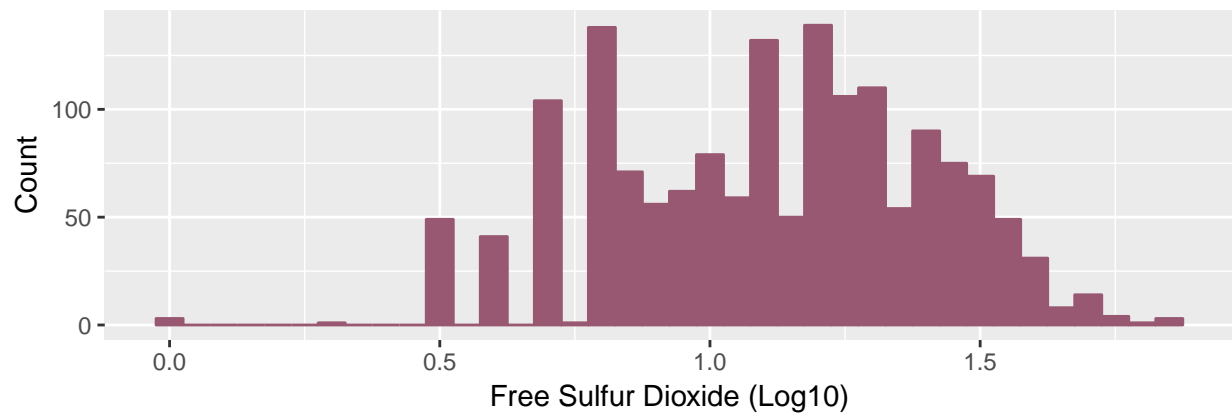
I can use \log_{10} on the x-axis to change the distribution to normal. Let's see how this will look like.



Based on the plots for **Total Sulfur Dioxide**, using `log10` on the x-axis helped the distribution to look normal-ish. There are still a few peaks in counts, but the overall shape of the distribution looks normal.

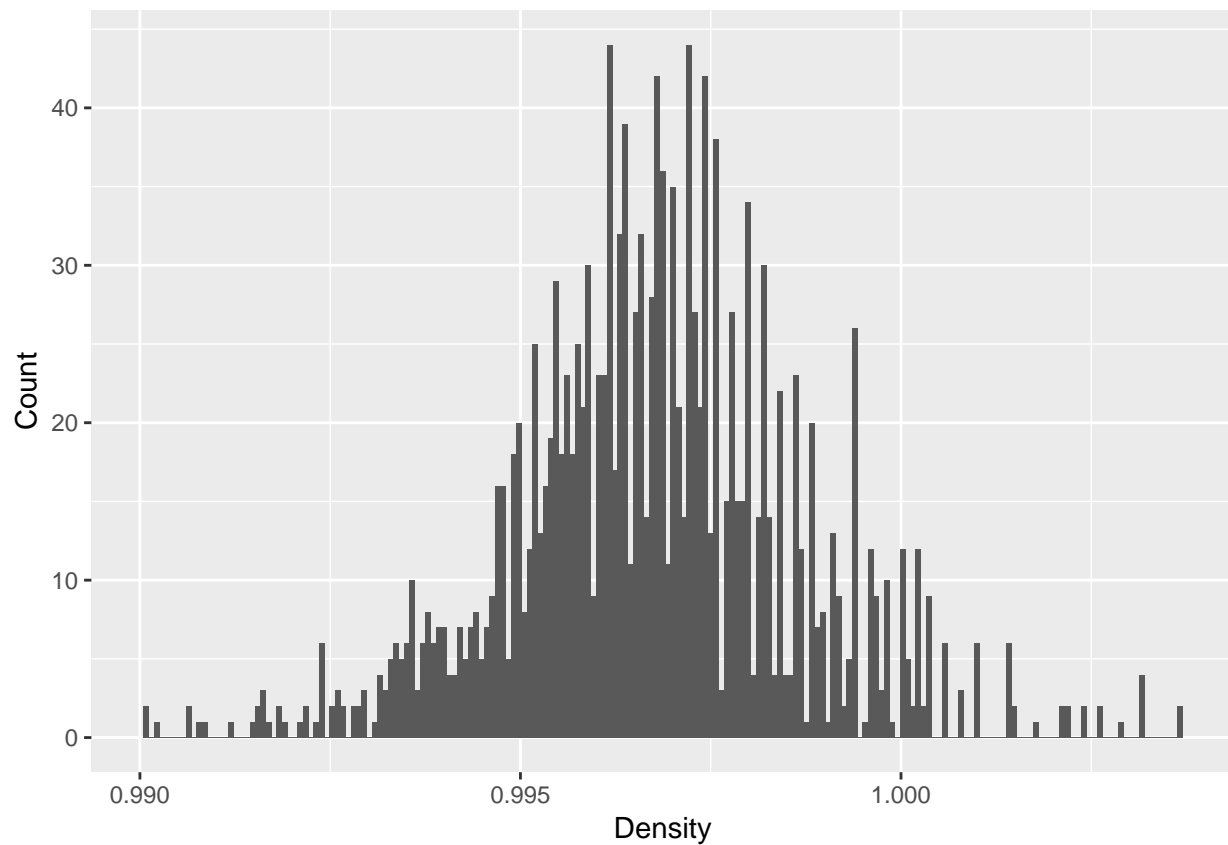
Free Sulfur Dioxide distribution is also skewed to the left, with most of the values concentrated around 5-15.

Applying `log10` on the x-axis does not work for **Free Sulfur Dioxide**, though. It changes the distribution, but seems the distribution goes more to the right, rather than getting normally distributed.



Density

The **Density** distribution looks like a normal distribution.



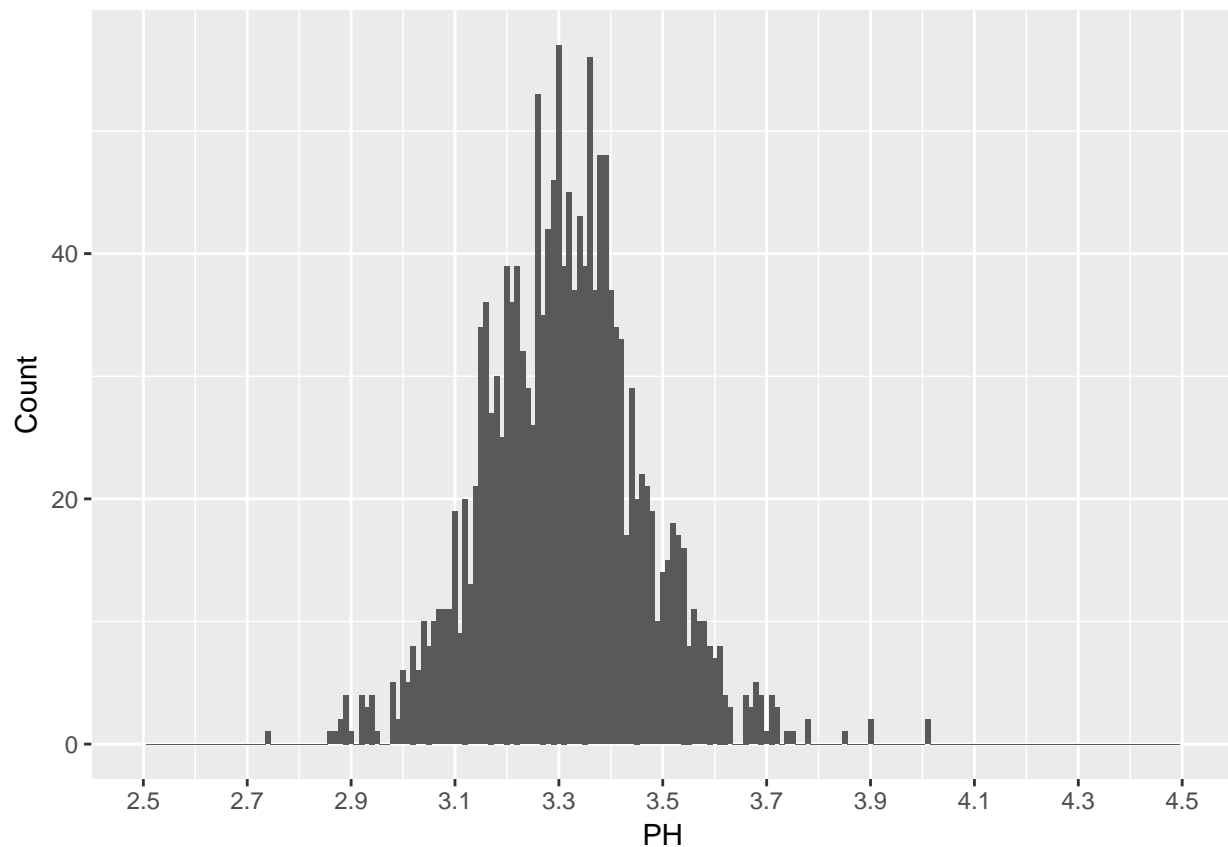
Getting the summary of the distribution, I have:

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

The mean of the distribution is 0.9967 which looks like it's almost in the middle of the distribution. The median is different than mean by 0.0011.

PH

The PH distribution looks normal, with a few outliers.

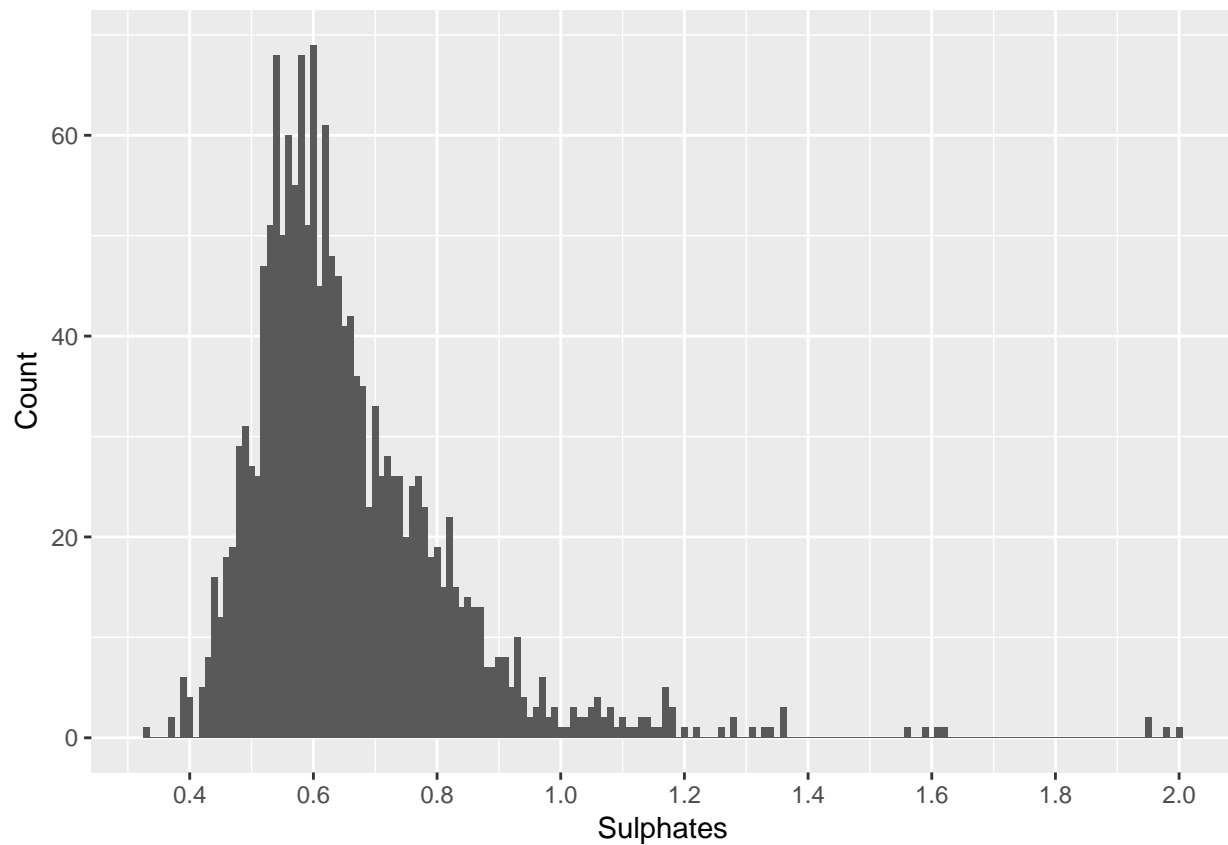


```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      2.740   3.210   3.310   3.311   3.400   4.010
```

The summary of the distribution shows the mean value falls at 3.31, very close to the median value.

Sulphates

The **Sulphates** distribution has most of its values concentrated around 0.5-0.7, and a few outliers around 1.6 and 2.0. I also check the summary of the table to see where the mean and median are placed.

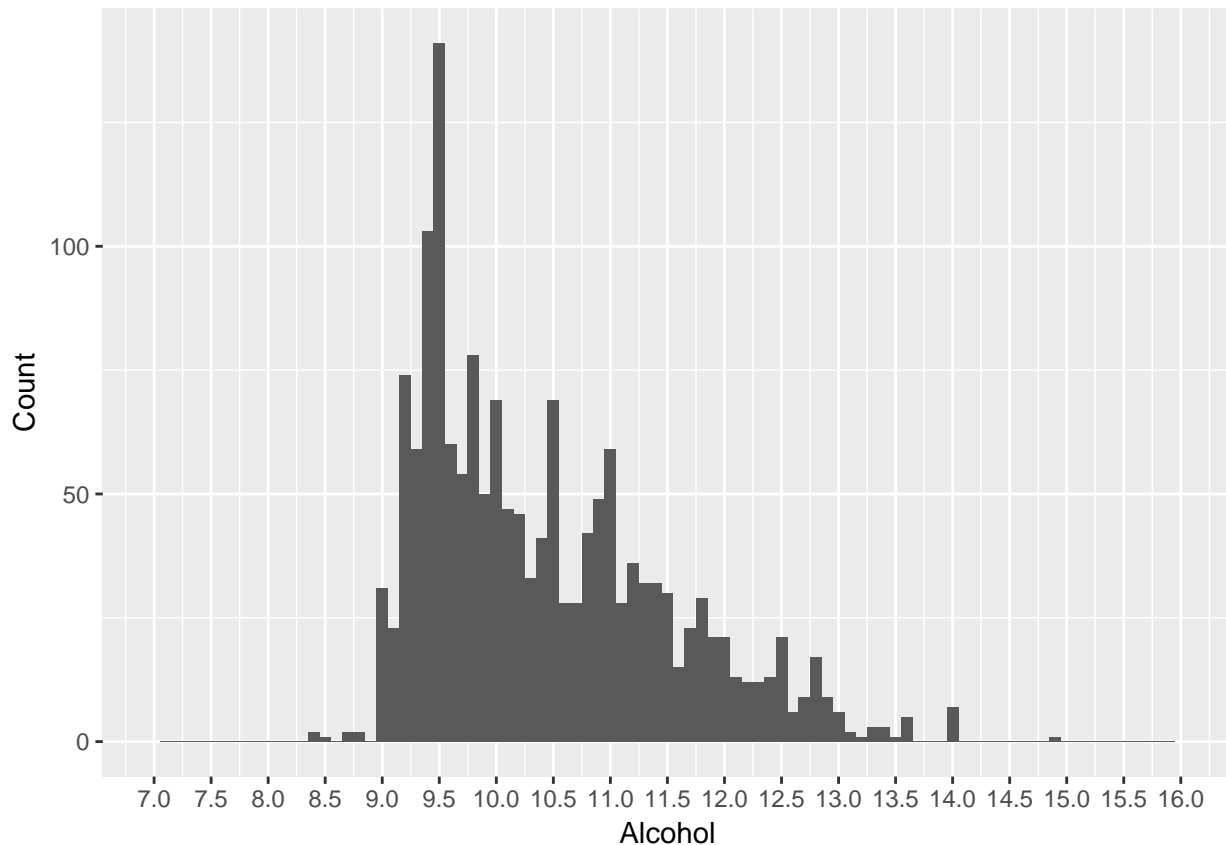


```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

The mean at 0.65 is within the range where most of the values are. The median at 0.62 looks also close to the mean.

Alcohol

The `Alcohol` distribution has its falls and rises, with its most values around 9.5% of alcohol.



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

The summary of the alcohol table shows the mean alcohol percentage is at 10.42, and the median is 10.20.

Univariate Analysis

Now, I'll answer a few questions regarding the database and my initial analysis.

What is the structure of the dataset?

The dataset has 1599 observations and 13 variables. Since I removed column `X`, and added column `rating`, the number of variables stay the same.

From the 13 variables, 11 of them are about chemical properties of red wines, 1 is about how the quality of red wine changes based on these chemical properties, and another one is where the red wines fall in the rating system based on their quality.

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

For me, the two important factors of interest are quality and rating. Further in the analysis I would like to see how the quality changes based on each chemical property.

Also, I would like to see if the alcohol level have any correlations with the quality of red wines.

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

For quality and rating, all the chemical properties can play an important role in supporting my investigation.

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

Yes, I created a new column `rating` which holds a rating system from `Poor` to `Excellent` based on the quality of red wines.

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?

1. I removed the `X` column since it was only holding the row numbers and was not playing any role in my investigation
2. I added the `rating` column to later categories the red wines in quality based on their ratings
3. Regarding the distributions, for Fixed Acidity, Volatile Acidity, Free Sulfur Dioxide and Total Sulfur Dioxide I applied the log-transformation on them to make the distribution more normal-ish.