Exploratory Data Analysis - Red Wine Dataset by Joseph Cayetano

Introduction: The Red Wine Dataset consists of 1599 red wines with 12 variables associated to them. I will be taking a look at the univariate, bivariate, and multivariate relationships between the 12 variables in the Red Wine Dataset using Exploratory Data Analysis (EDA). The dataset has already been cleaned and prepared for data analysis.

Univariate Plots Section

Variables of wine dataset

```
##  [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"
```

Structure of wine dataset

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

Summary of dataset

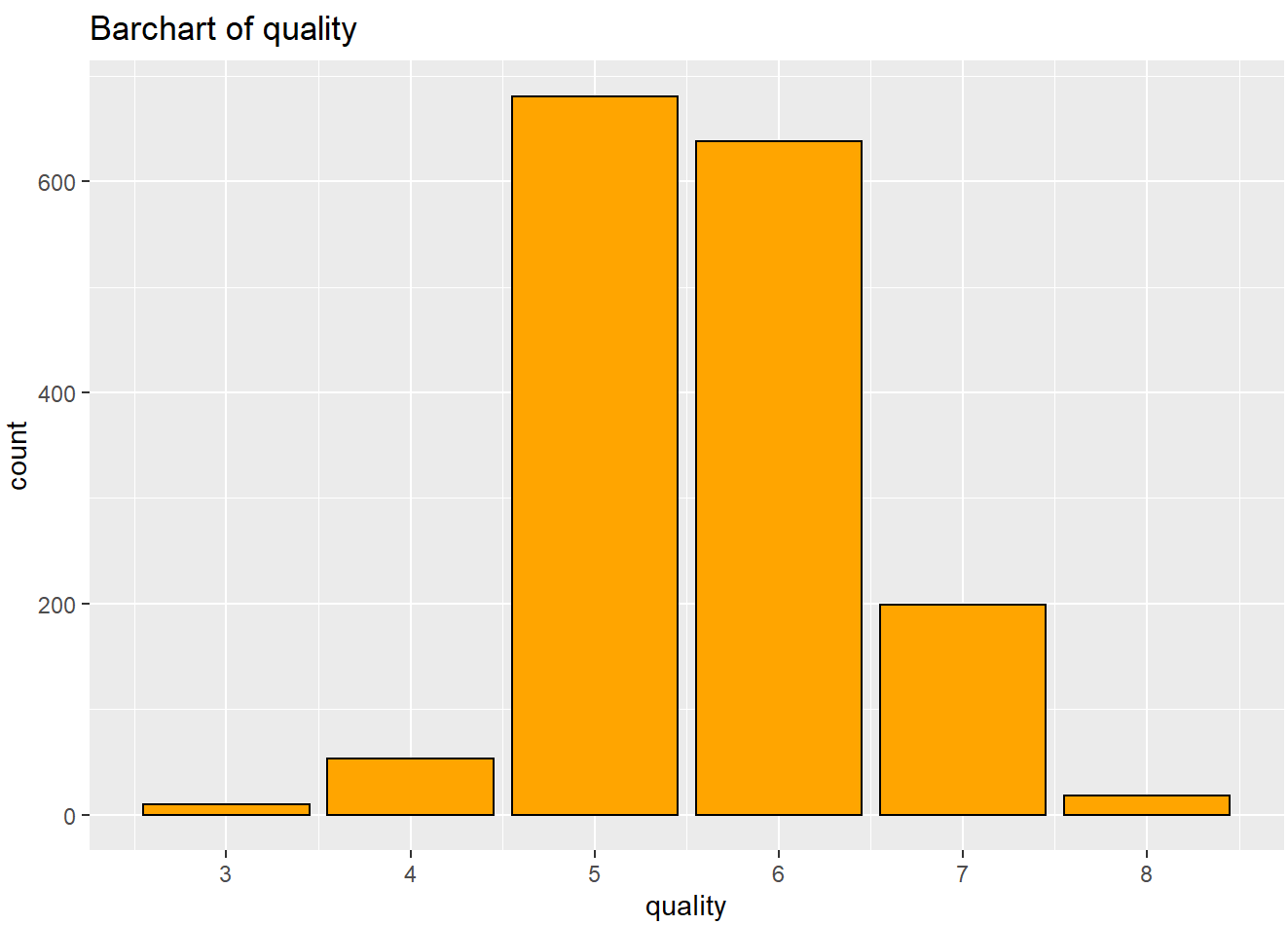
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

The maximum number given to quality is 8, while the minimum is 3. Basically, 8 is great, while 3 is bad.

Total count of each quality

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

Quality Barchart



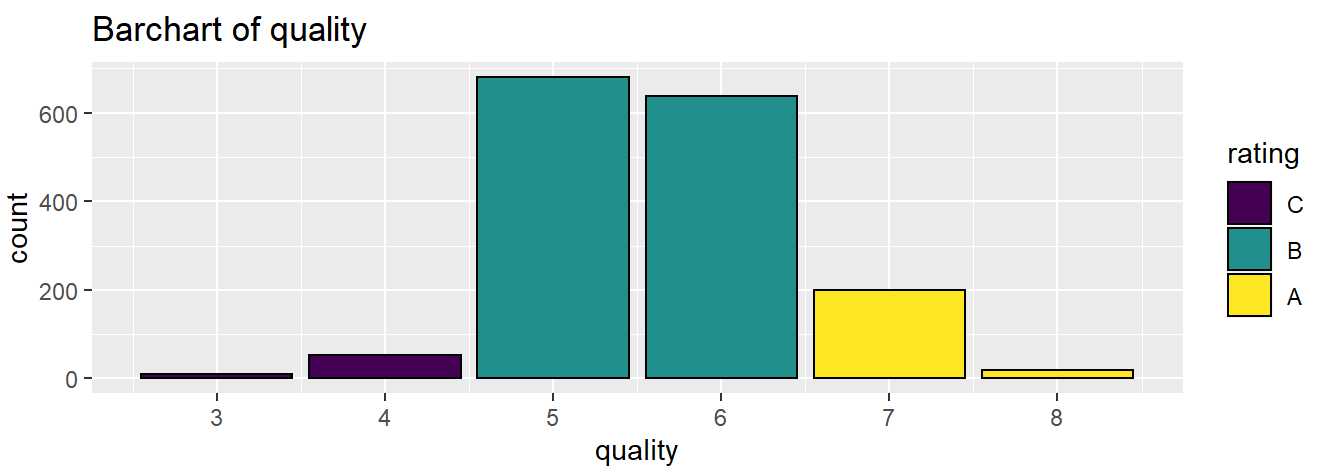
Looking at the bar chart, we can see that most wines in the dataset have a quality of 5 or 6. Also, wines with a quality of 3 or 8 rarely occur in the dataset.

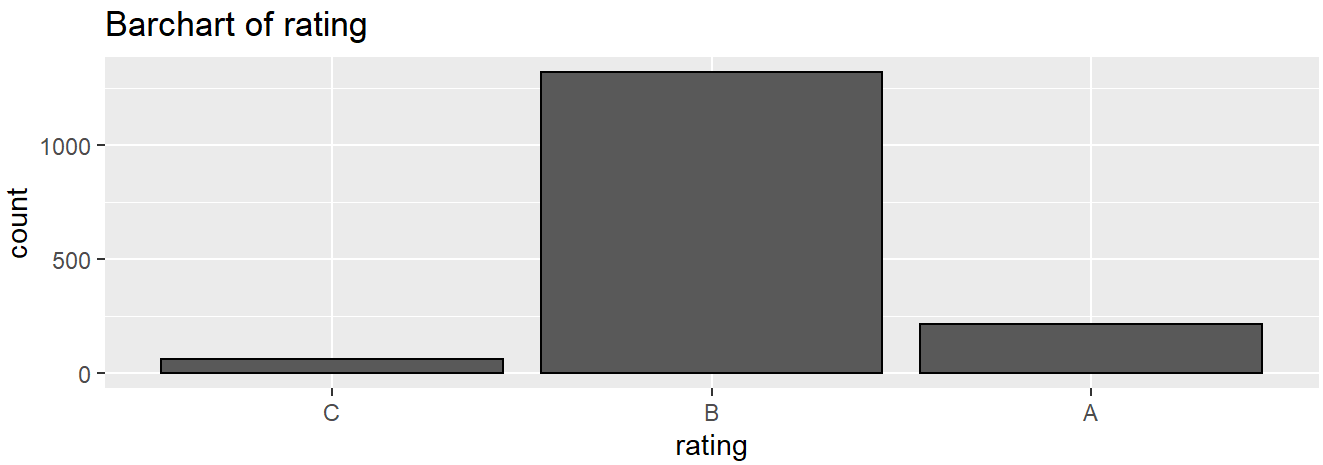
Creating ‘rating’ variable.

```
##    C    B    A
##   63 1319  217
```

I have made a ‘rating’ variable based on the variable ‘quality.’ ‘rating’ divided the ‘quality’ into 3 rating levels: A, B or C. Wines that have a quality of 3 or 4 are rated C. Wines that have a quality of 5 or 6 are rated B. Wines that have a quality of 7 or 8 are rated A.

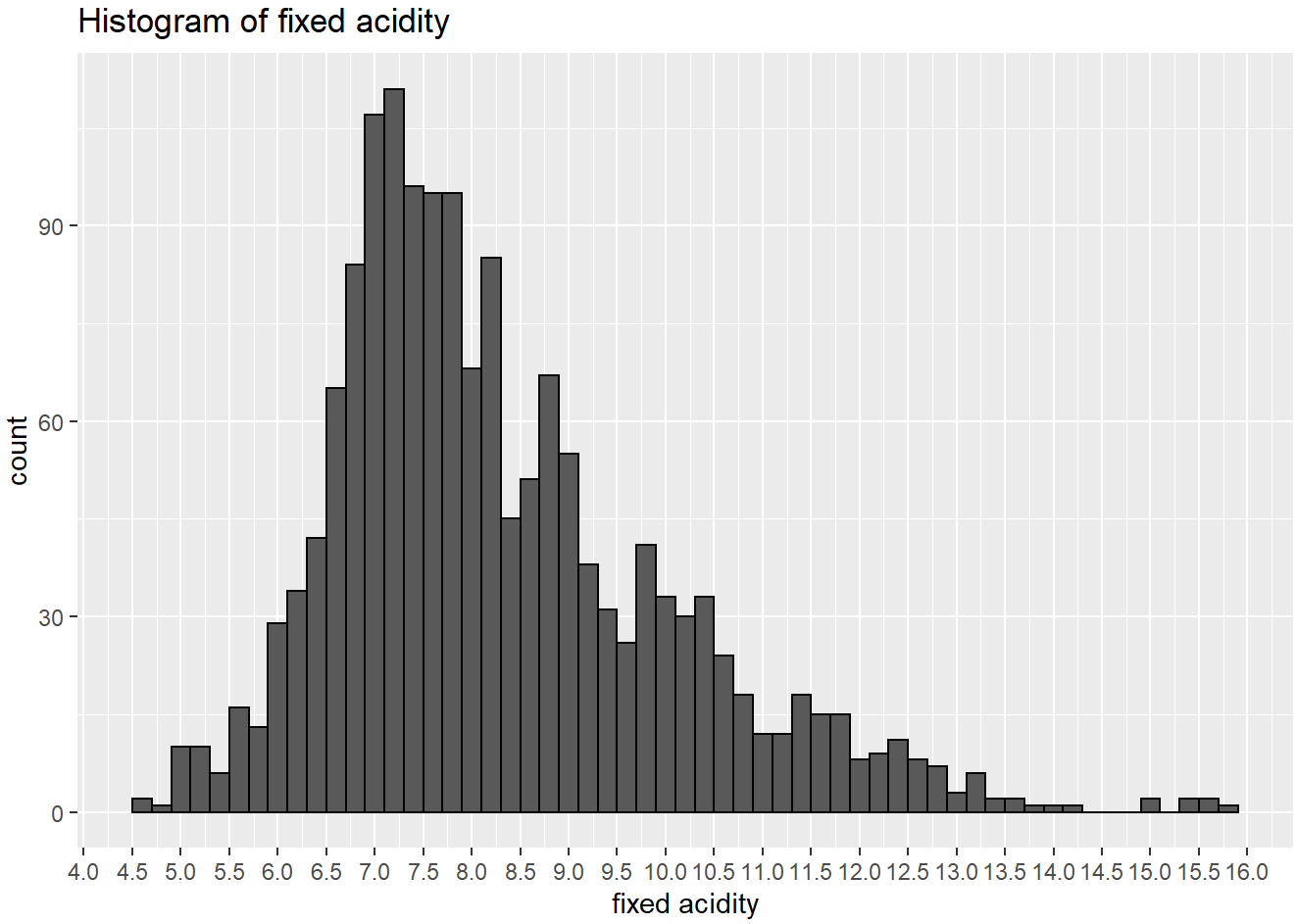
Quality and rating barchart comparison





As we can see, we now have a 'rating' bar chart that is equivalent to the 'quality' bar chart. Most wines have a B rating, which tells that their quality is either 5 or 6.

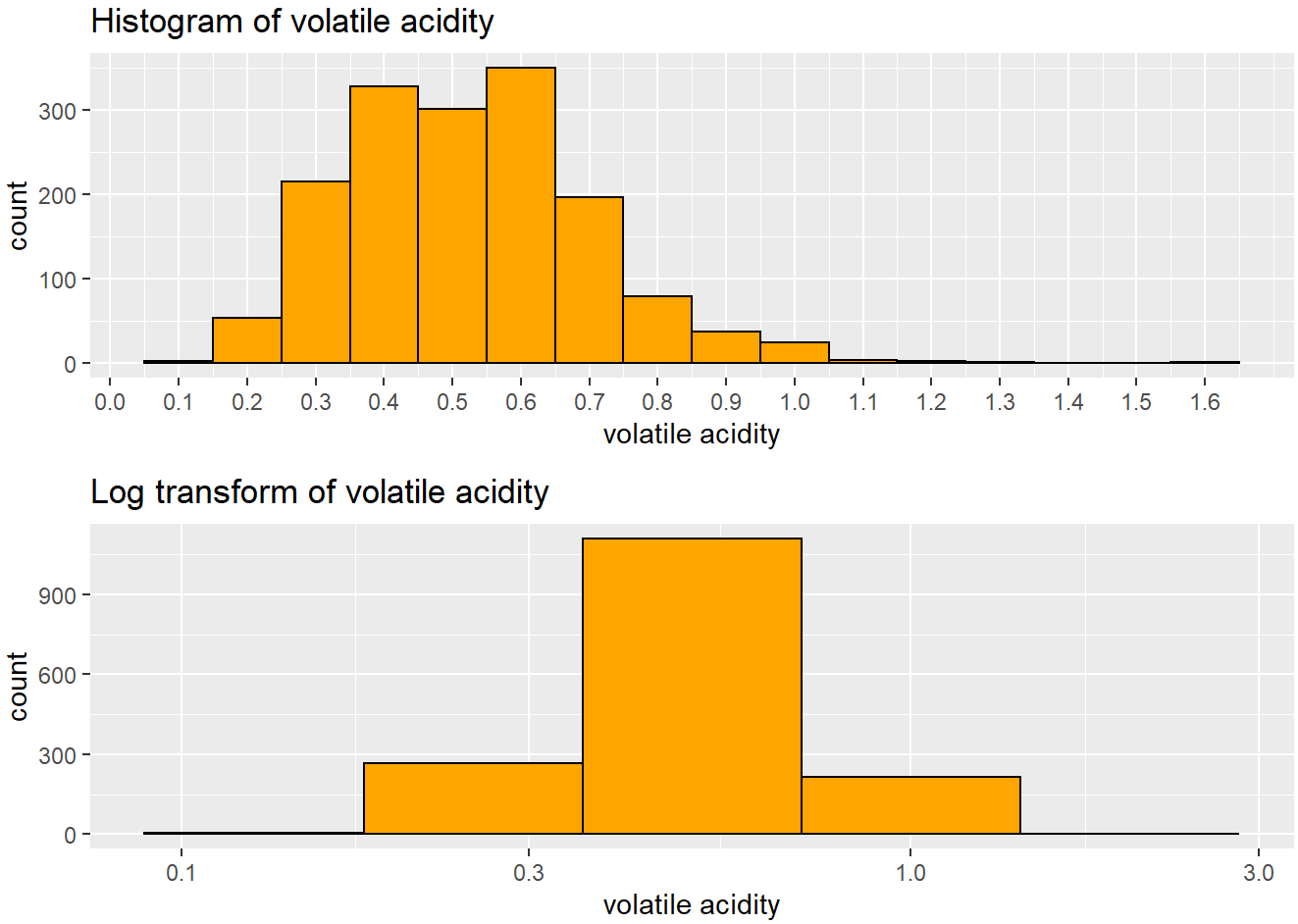
Fixed acidity barchart



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

The distribution of the histogram regarding fixed acidity is somewhat skewed to the right. It has a peak at 7.0 and a mean of 8.3. The minimum value is 4.6 while the maximum value is 15.9.

Volatile acidity comparison with log transform

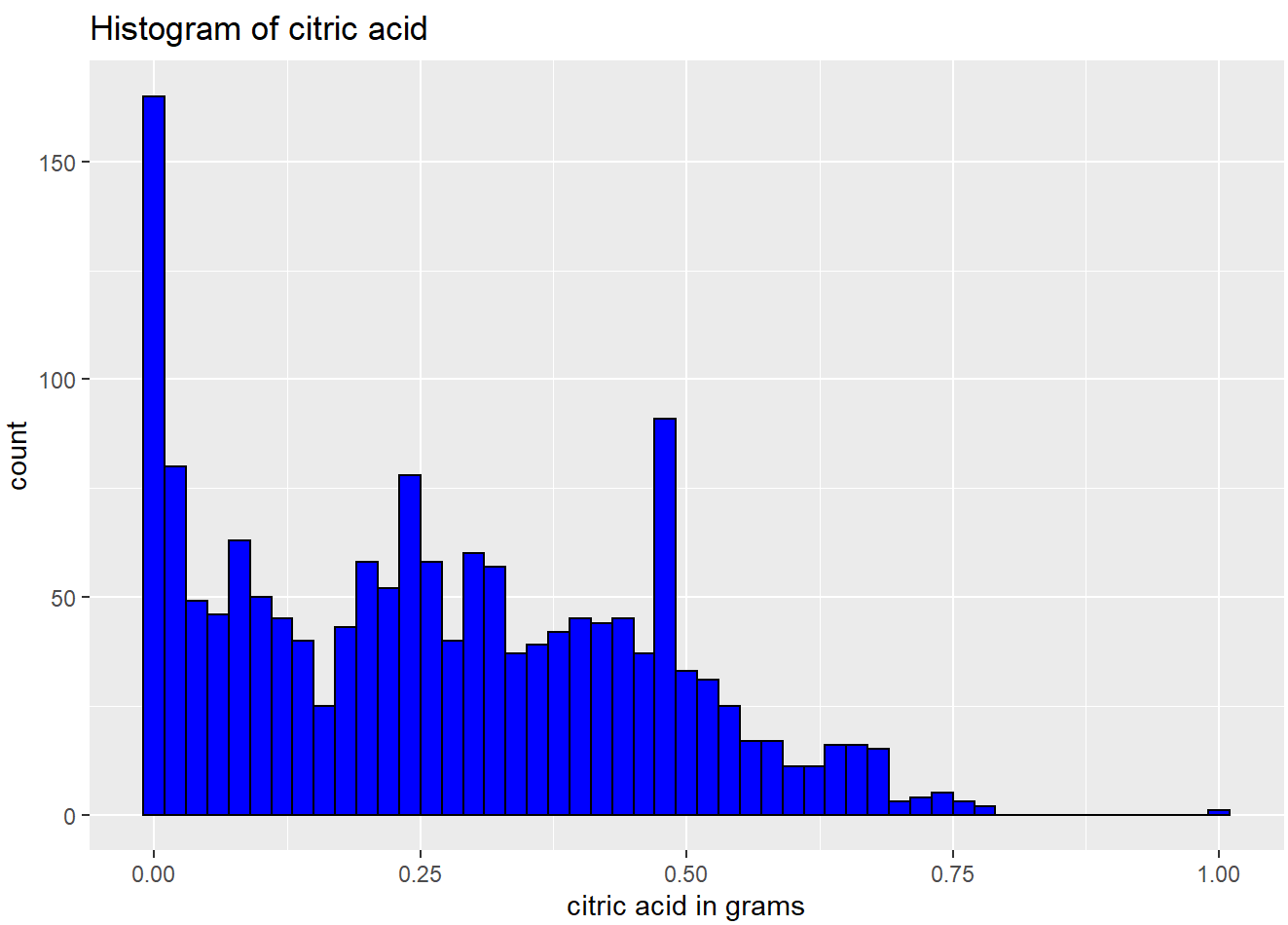



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

plot 1 – The histogram's distribution of volatile acidity appears to be bimodal. There are two peaks, which are at 0.4 and 0.6. The average amount of acetic acid in the wines is about 0.5. The lowest amount of acetic acid found is 0.1, while the highest is 1.6. The wine with a volatile acidity value of 1.6 has about three times the amount of acetic acid than the average wine.

plot 2 – Using a log transformation (specifically scale_x_log10()), the histogram's distribution changes from bimodal to skewed left. The peak is still in between 0.3 and 1.0.

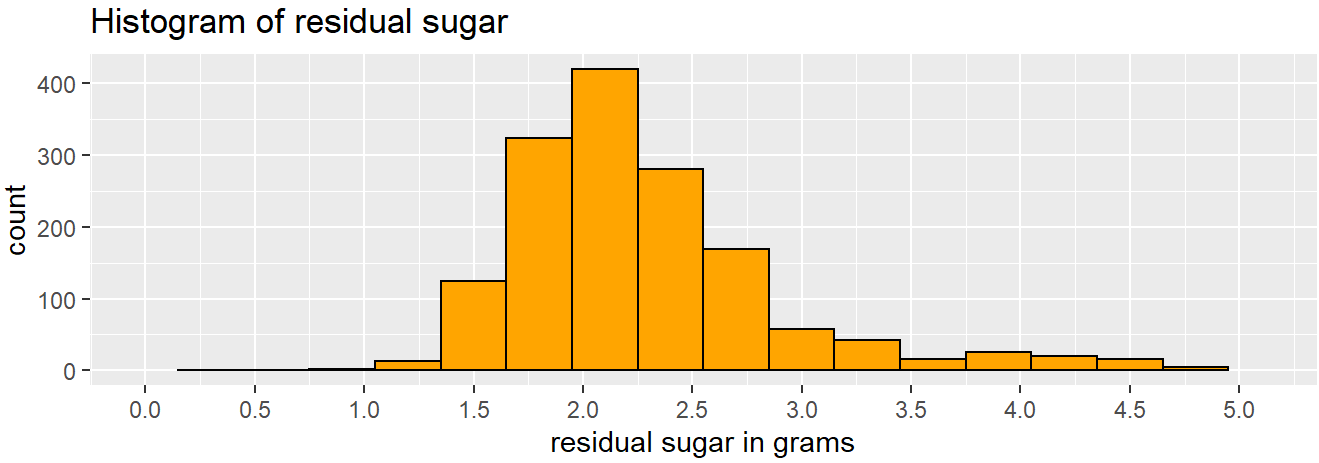
Citric acid histogram

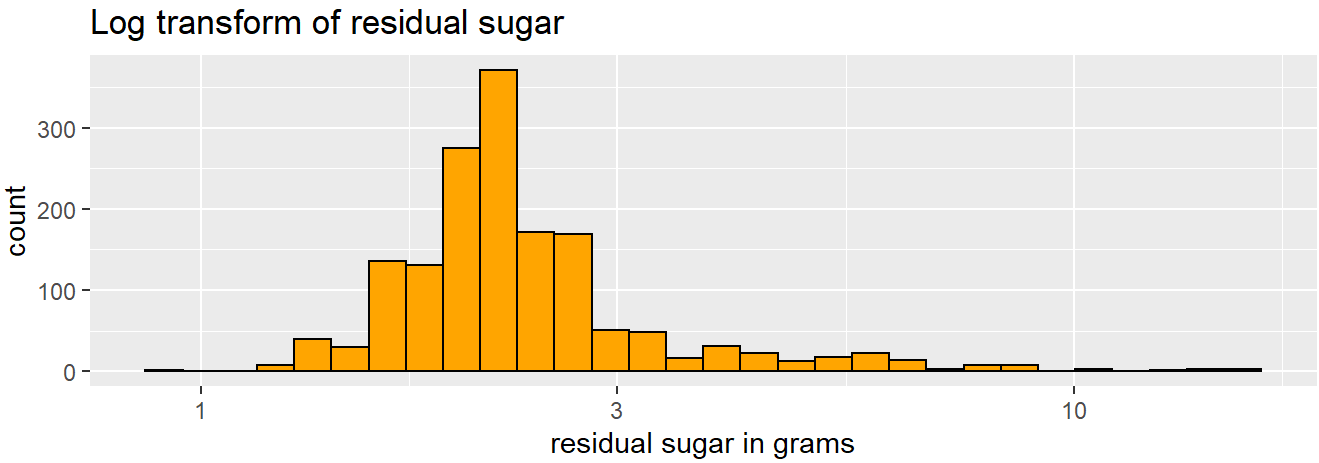



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

The histogram's distribution regarding citric acid is skewed to the right. The average amount of citric acid found in wines is about 0.27 and the lowest and highest amount found is 0.00 and 1.00 respectively.

Residual sugar comparison with log transform

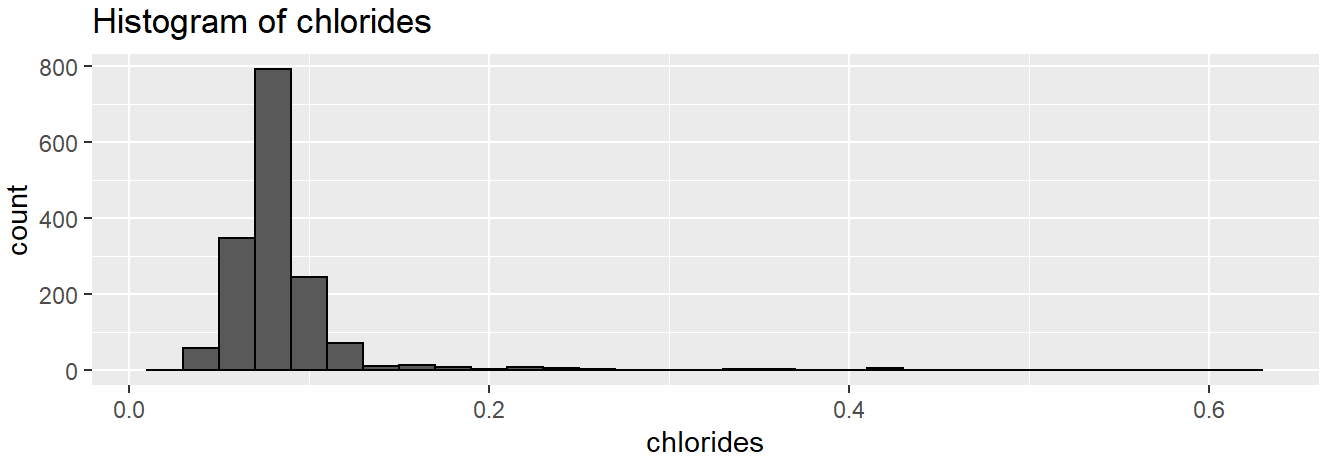


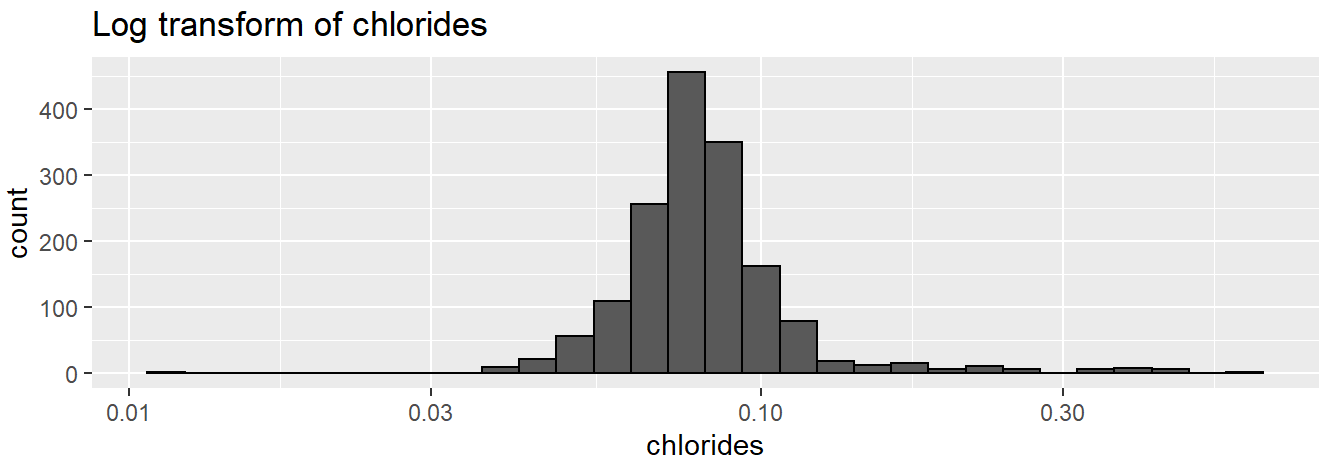



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

Looking at the histogram, we can see that the residual sugar in most wines fall in between 1.0 and 3.0. It is worth noting that I did not include the top 5% of residual sugar in the graph. The minimum amount of sugar found in the wines is 0.9, while the maximum is 15.5. The average amount of sugar in all of the wines is 2.5. To put this into perspective, the wine that has a residual sugar value of 15.5 is six times sweeter than the average wine.

Chlorides comparison with log transform



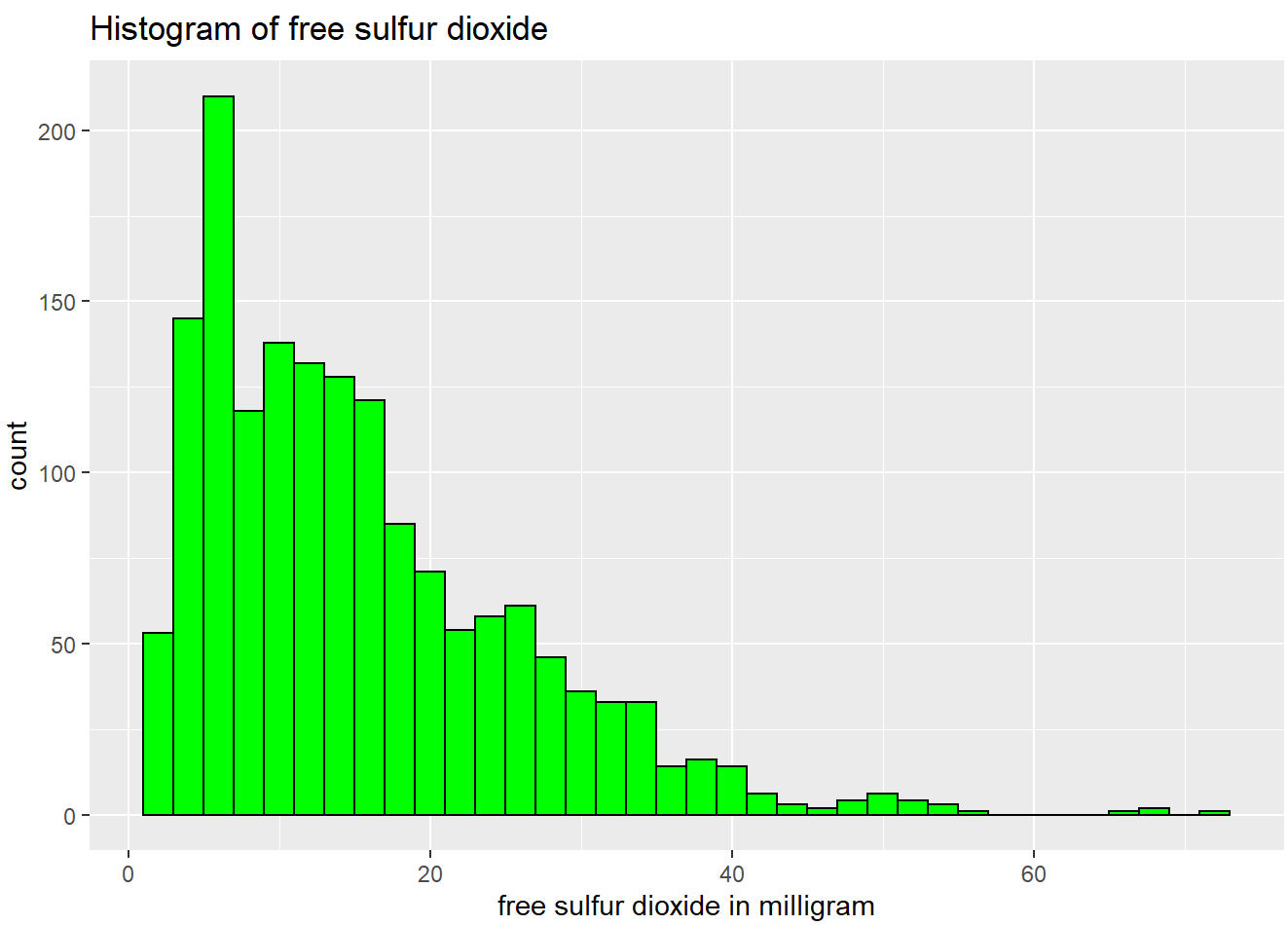



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

plot1 - The histogram's distribution regarding chlorides is symmetrical. The highest amount of salt found in the wine is 0.61 while the lowest amount is 0.01. The average amount of salt in the wines is 0.08. Most wines in the data set have a chloride value between 0.03 and 0.10.

plot2 – Using a log transformation (specifically scale_x_log10()), the histogram's distribution did not change and stayed symmetrical. We can now clearly see the values in which most wines fall in between regarding chlorides.

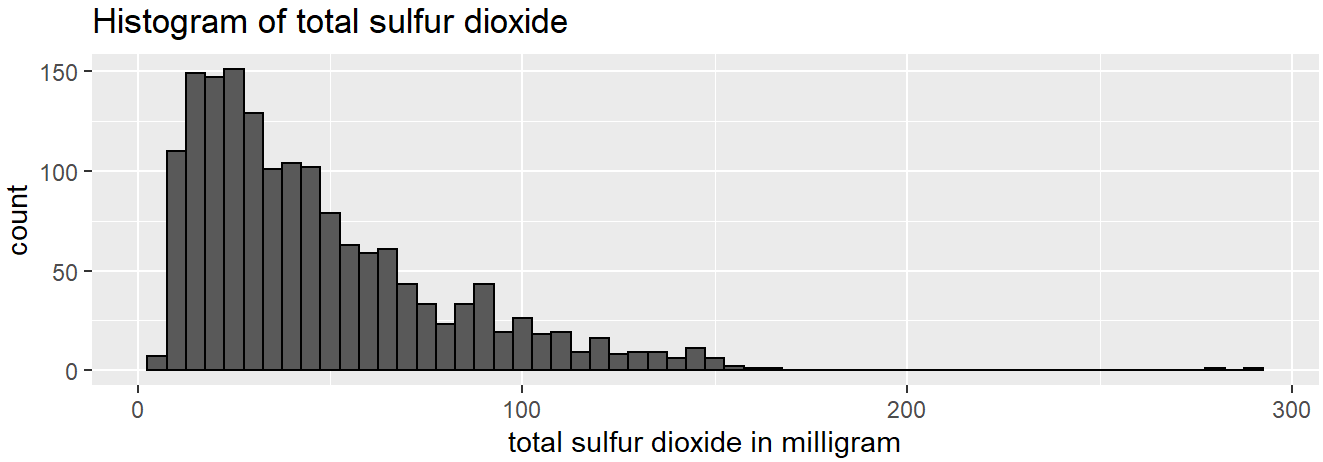
Free sulfur dioxide histogram

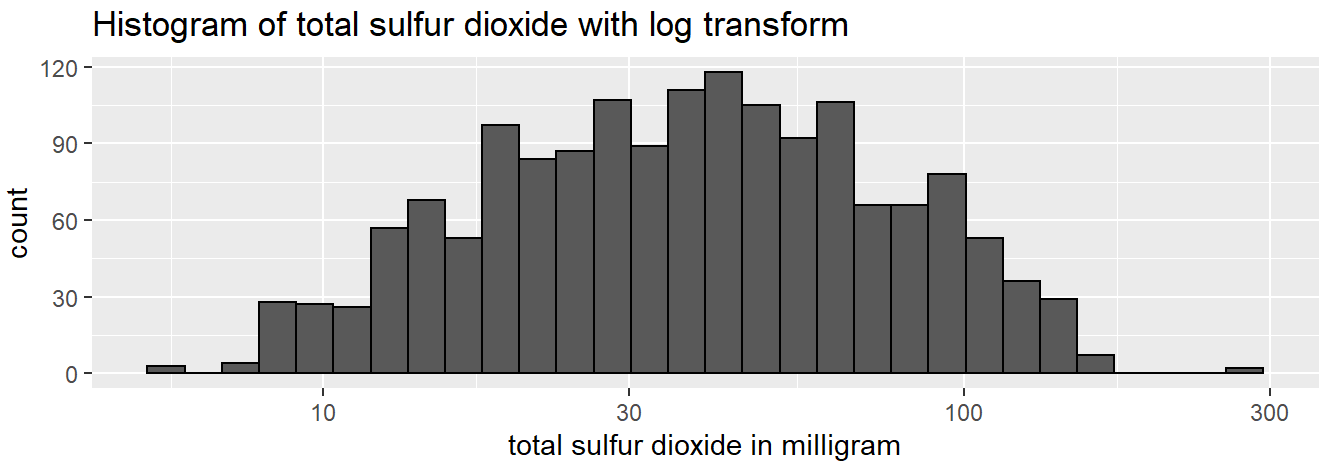



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

The histogram's distribution regarding free sulfur dioxide is skewed to the right. The average amount of free sulfur dioxide found in the wines is 15.8. The minimum and maximum amount is 1.0 and 72.0 respectively. It is interesting to note that the wine with the highest amount of free sulfur dioxide is four times larger than the average wine.

Total sulfur dioxide comparison with log transform



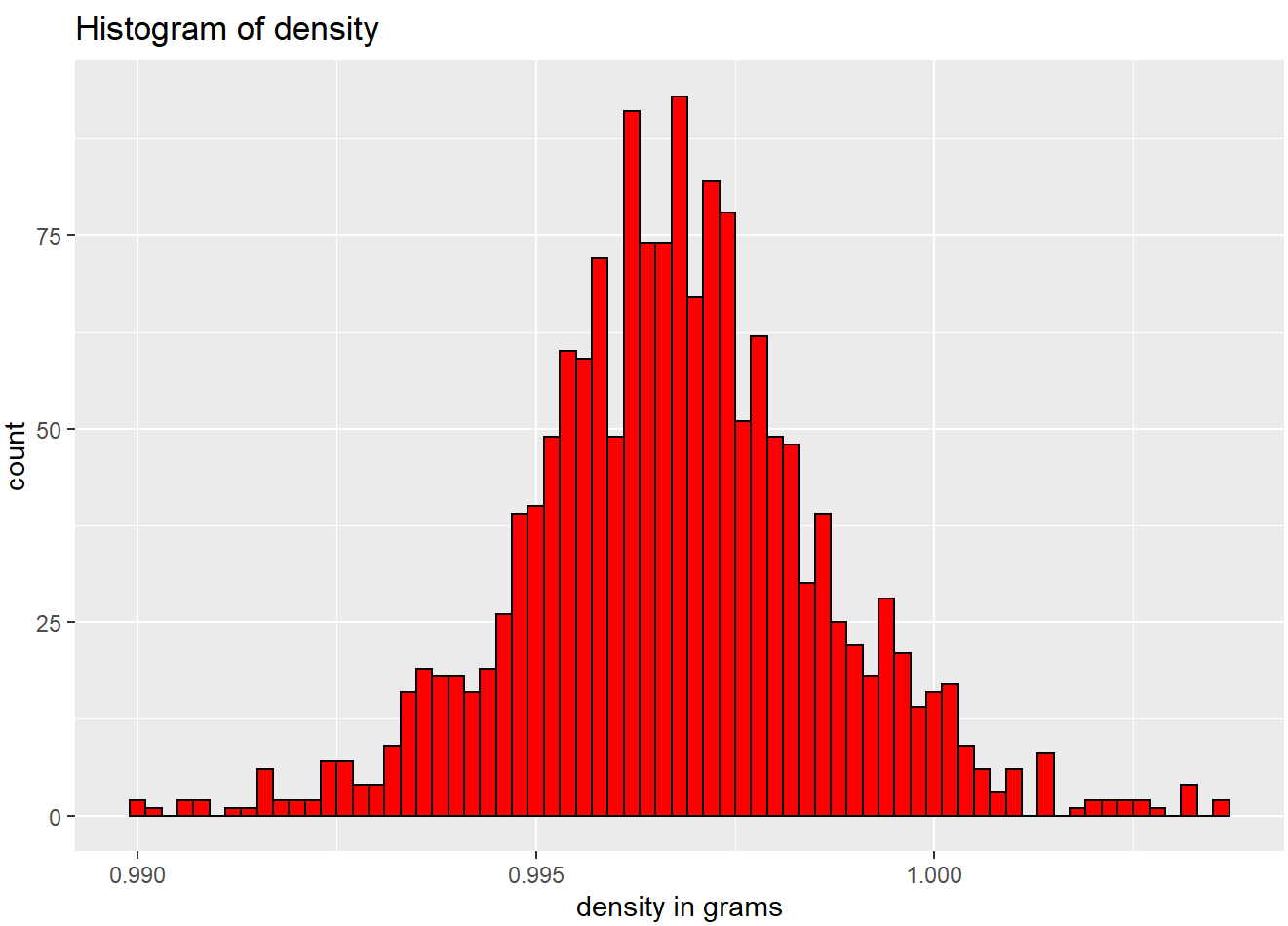



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

plot1 – The histogram's distribution regarding total sulfur dioxide is skewed to the right. The average total sulfur dioxide in the wines is 46.4. The maximum and the minimum value is 289.0 and 6.0 respectively.

plot2 - Using a log transformation (specifically scale_x_log10()), the histogram's distribution changed to symmetrical. We can clearly see that the total sulfur dioxide value for the majority of wines fall in between 10 and 100.

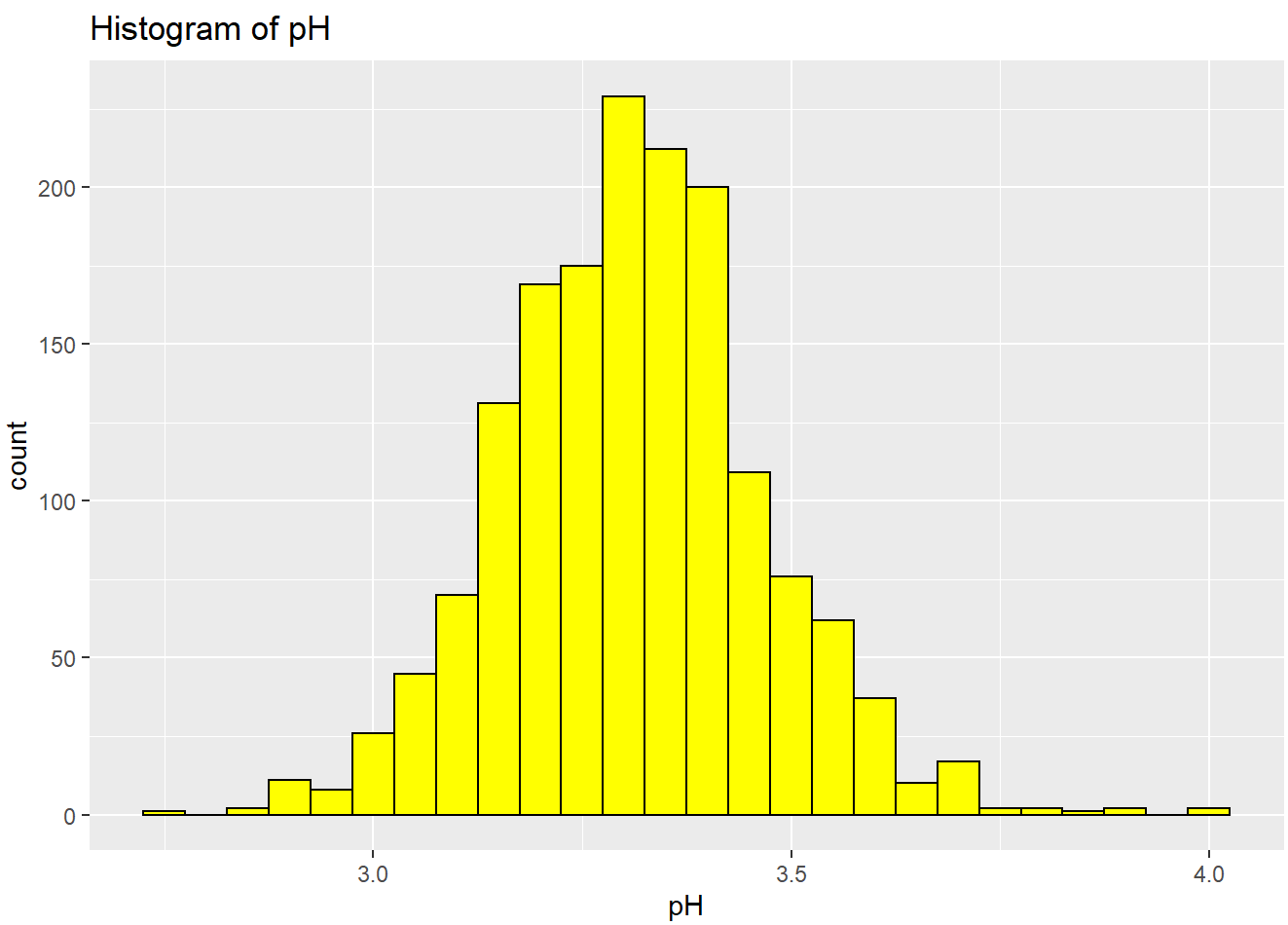
Density histogram




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

The histogram's distribution regarding density is symmetrical. The minimum value is 0.99 while the maximum value is 1.0037. The average density of the wines is 0.9967. Calculating the standard deviation of the variable 'density,' we can see that it is small with a value of 0.0018.

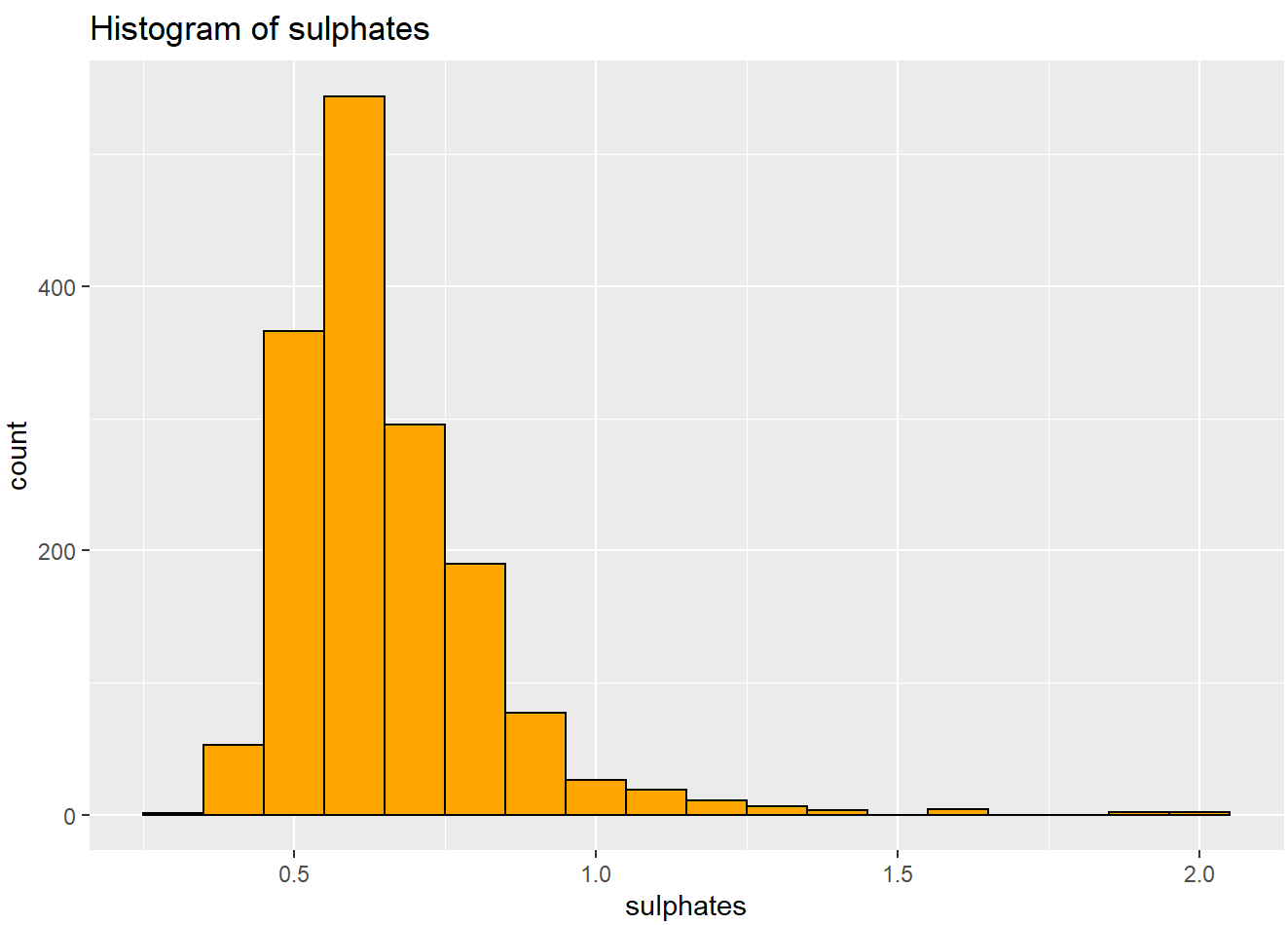
pH histogram




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

The histogram's distribution regarding pH is symmetrical or unimodal. Most wines' pH falls in between 3.0 and 3.6. The wine that is closest to 0 (very acidic) has a value of 2.7, while the wine that is closest to 14 (very basic) has a value of 4.0. The wines in this dataset have an average pH value of 3.3. This means that they are acidic like sodas.

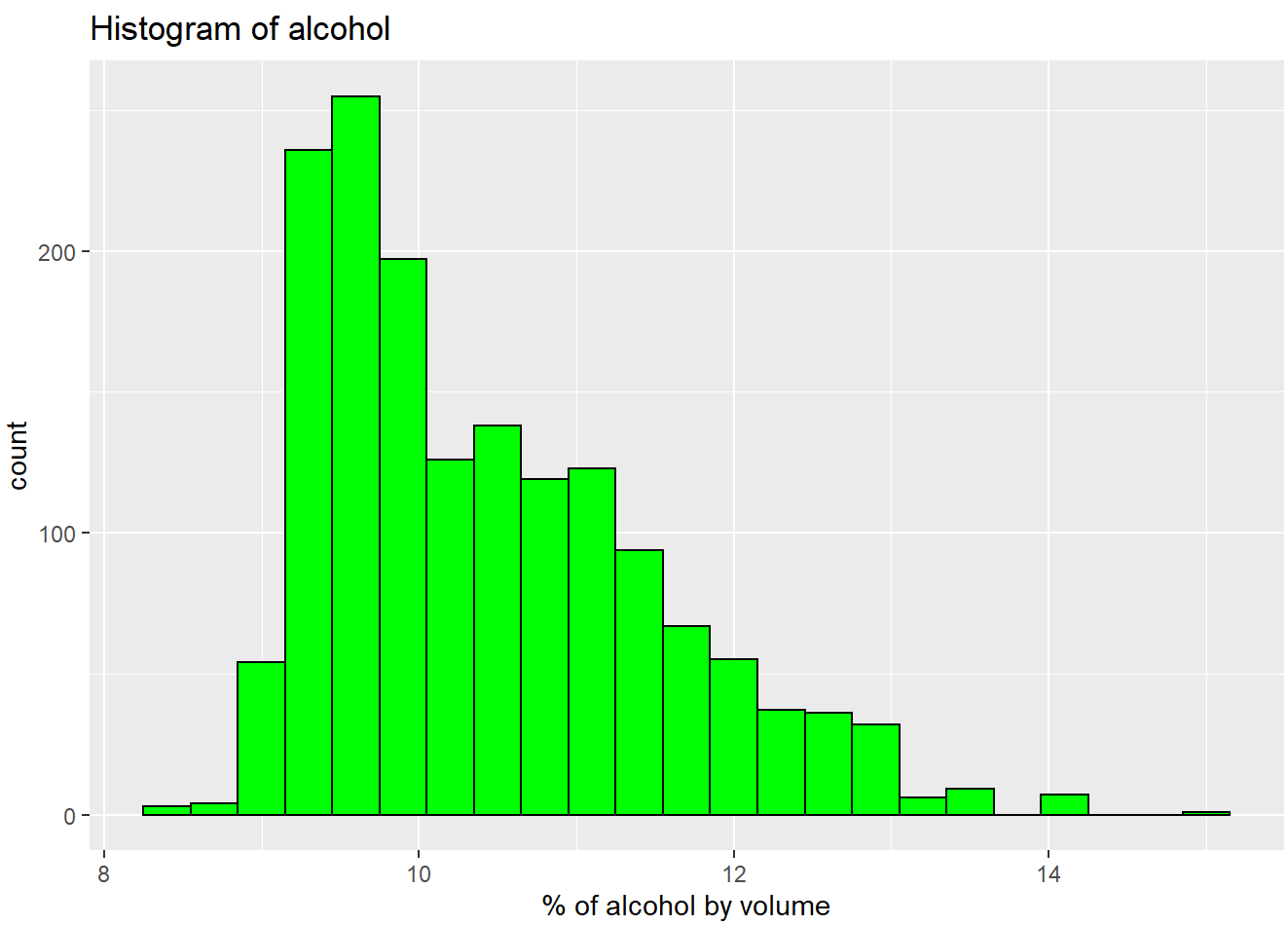
Sulphates histogram




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

The histogram's distribution regarding sulphates is skewed to the right. This wine additive has a minimum value of 0.22 and a maximum value of 2.00. The mean of wines with sulphates is 0.65.

Alcohol barchart




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

The distribution of the histogram regarding alcohol is skewed to the right with a mean of 10.4. This means that the average percent alcohol content of all the wines in the dataset is 10.4 The lowest percent alcohol content found is 8.4, while the highest is 14.9.

Univariate Analysis

What is the structure of your dataset?

The Red Wine dataset had 1599 rows with 13 variables initially. The number of variables became 14 after adding the ‘rating’ variable. The variable ‘quality’ is the categorical variable, and the rest of the variables have a numerical data type.

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

The main feature of interest in the dataset is the how a variable influences the quality of wine such as alcohol.

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

I think most variables such as acidity(fixed, volatile) will help support my investigation into my main feature of interest. I also think pH will play a huge role on quality.

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

Yes, I changed the variable ‘quality’ into an ordered factor and created a ’rating ’variable which is as follows: wine qualities of 3-4 are rated C, 5-6 are rated B, and 7-8 are rated A.

Of the features you investigated, were there any unusual distributions?

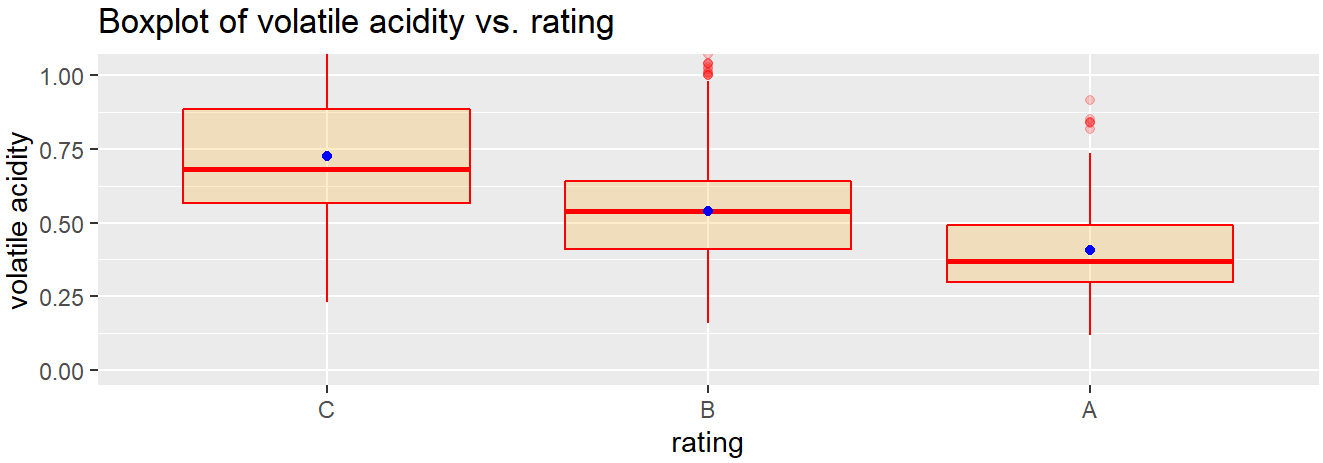
There were certainly some unusual distributions. As we saw with the graphs, there were only few symmetrical and most were skewed. Also, I was not expecting to have some extreme outliers such as from residual sugar and chloride. For some of the variables, I have taken the 99 and 95 quantiles to have a better view of the data.

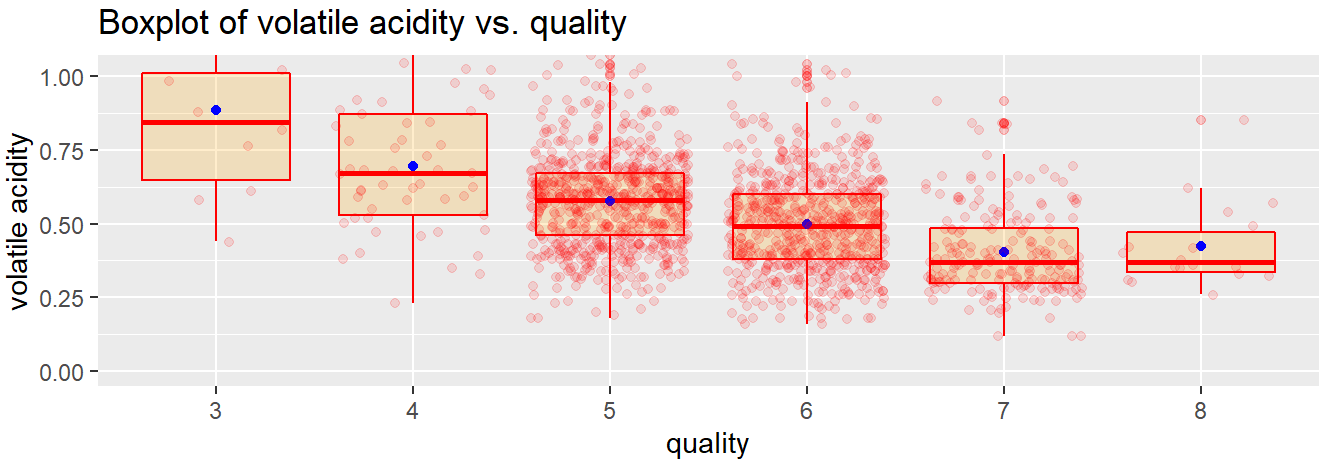
Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I did not perform any operations on the data set. The data set has already been cleaned and prepared for data analysis.

Bivariate Plots Section

Boxplot of volatile acidity vs. rating and quality





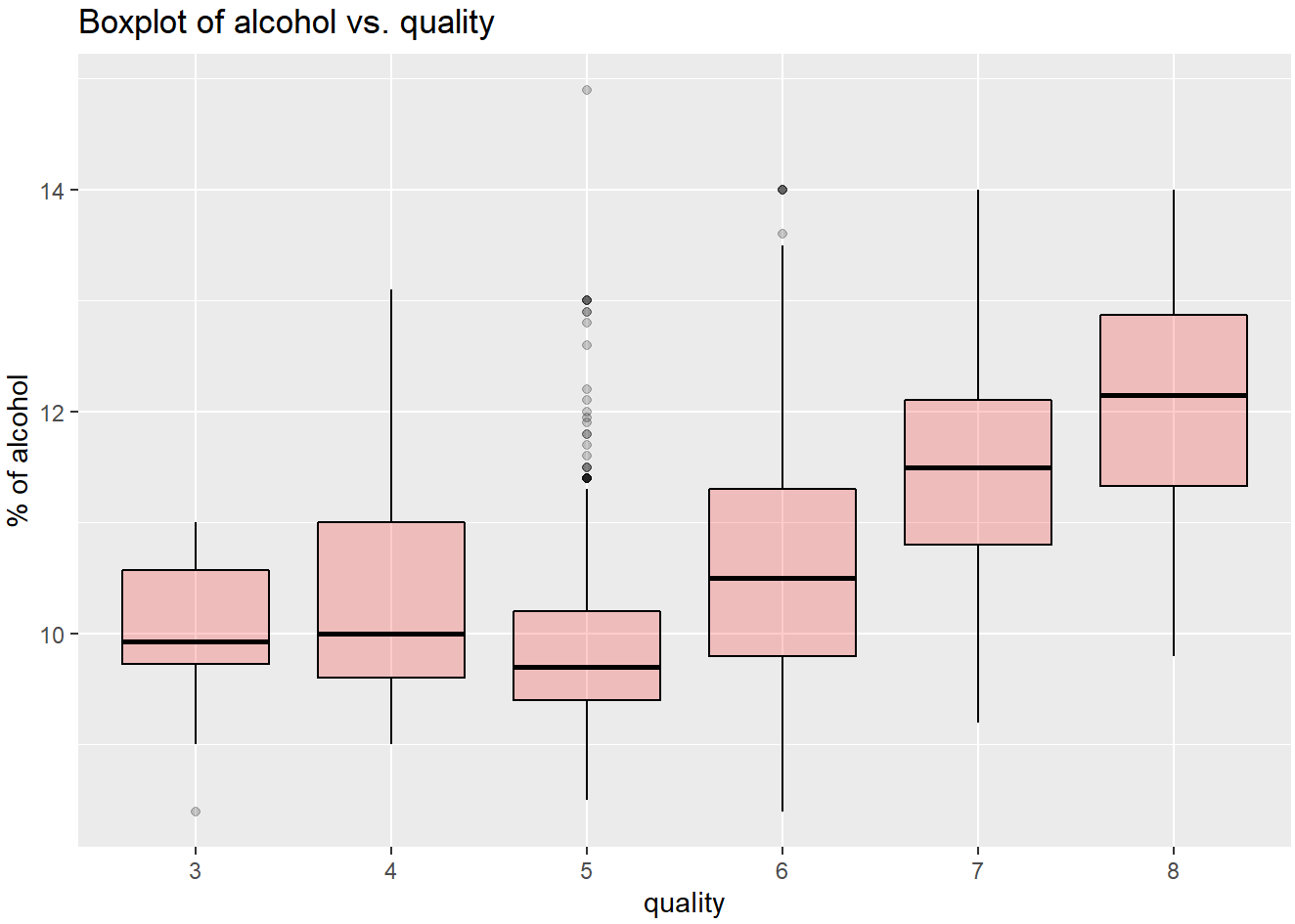
plot1 – For this boxplot, we used the newly created variable ‘rating’ as our independent variable and ‘volatile acidity’ as dependent variable. The blue dot is the mean of each rating. As we can see, rated A wines on average have a lesser amount of acetic acid than both B and C rated wines.

plot2 – Substituting ‘rating’ with ‘quality’, we can see that higher quality wines definitely have lesser amount of acetic acid. Also, we can see how the majority of wines in the data has a quality of either 5 or 6 by the number of red dots in the graph.

```
##
##  Kendall's rank correlation tau
##
## data:  yo$volatile.acidity and yo$quality
## z = -15.498, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.3007787
```

Using the correlation test on both volatile acidity and quality, there is a weak negative relationship between the two variables with a value of -0.3.

Boxplot of quality vs. alcohol

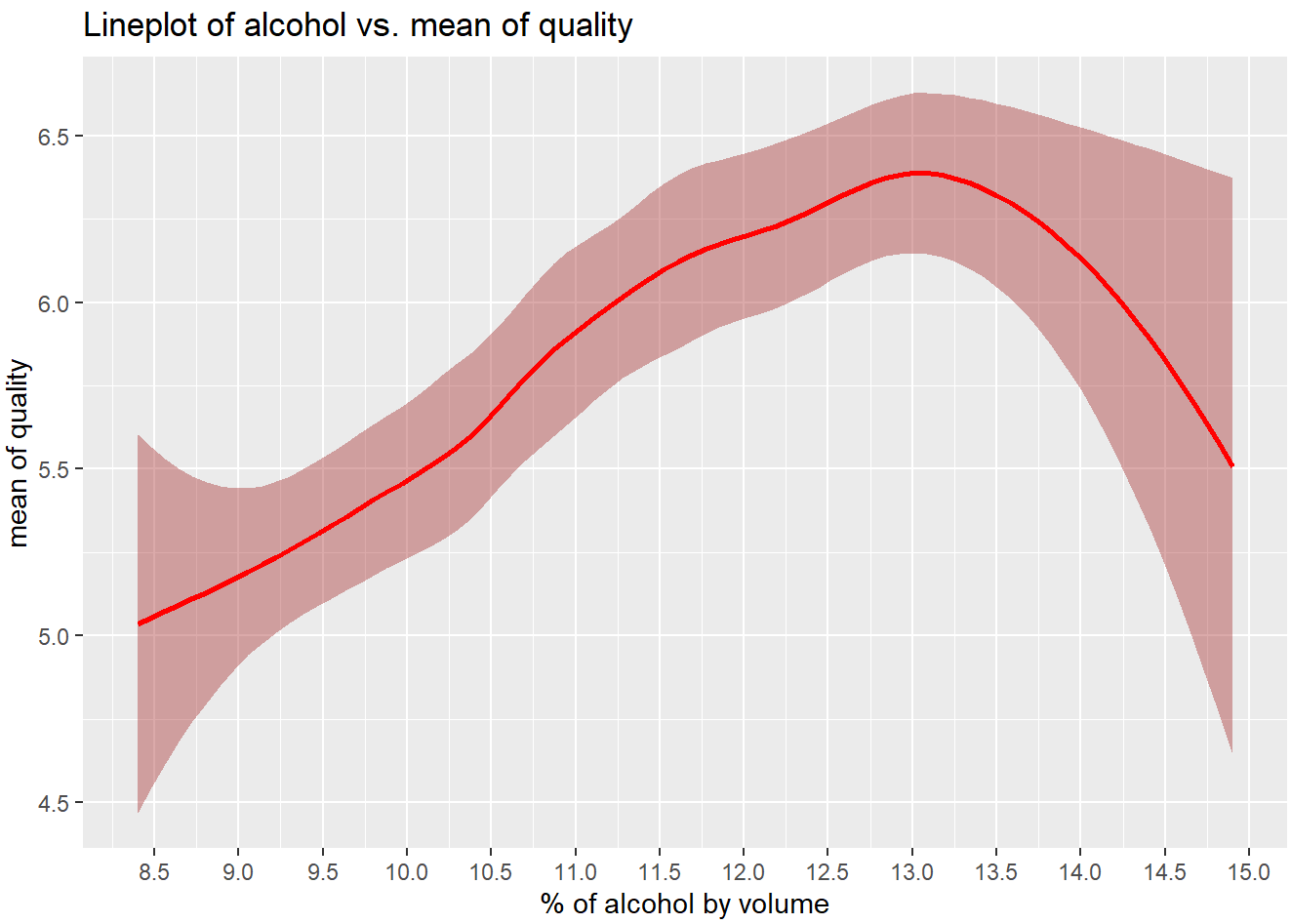


For this boxplot, we used the variable ‘quality’ as our independent variable and ‘alcohol’ as dependent variable. As we can see, higher quality wines have higher alcohol content.

```
##
##  Pearson's product-moment correlation
##
## data:  yo$alcohol and yo$quality
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

Using the correlation test on both alcohol and quality, there is a moderate positive relationship between the two variables with a value of 0.47

Does that mean that more alcohol gives us a better wine?

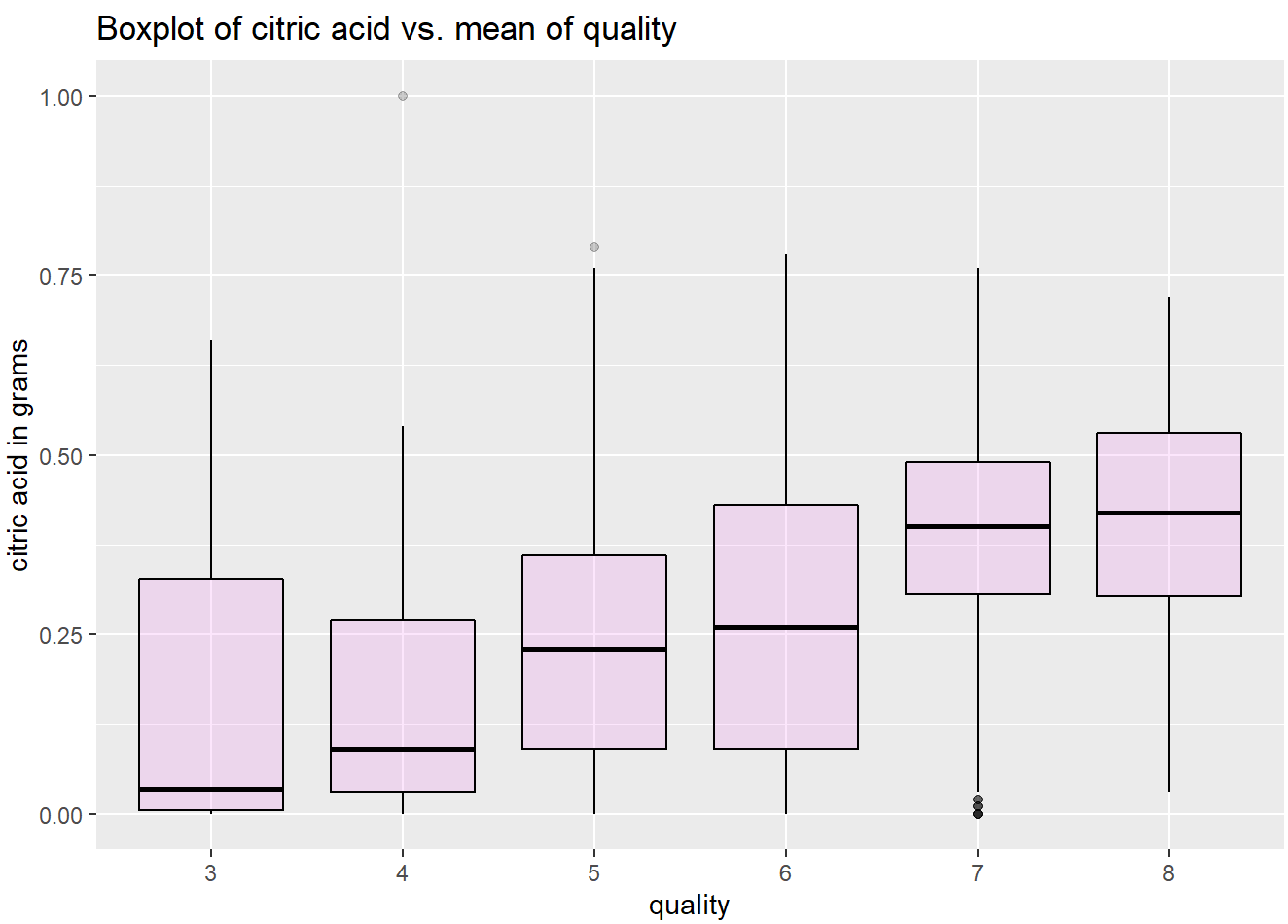


For this plot, we can see that the quality of the wine increases as alcohol increases. However, once the percentage of alcohol hits the 13.0 mark, the quality of the wine starts to decrease.

```
##
##  Pearson's product-moment correlation
##
## data:  up13$quality and up13$alcohol
## t = -0.39861, df = 21, p-value = 0.6942
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4816540  0.3376049
## sample estimates:
##         cor
## -0.08665653
```

I made a variable called up13 that consists of wines that have an alcohol content of above 13. I used this variable to test the correlation between ‘up13quality’ and ‘up13alcohol.’ Seeing the result, there is a negative relationship between the two variables with a value of -0.08. It is interesting that once we hit the 13.0 alcohol content, the correlation between alcohol and quality declines, as we can see from the result of the correlation test.

Boxplot of quality vs. citric acid

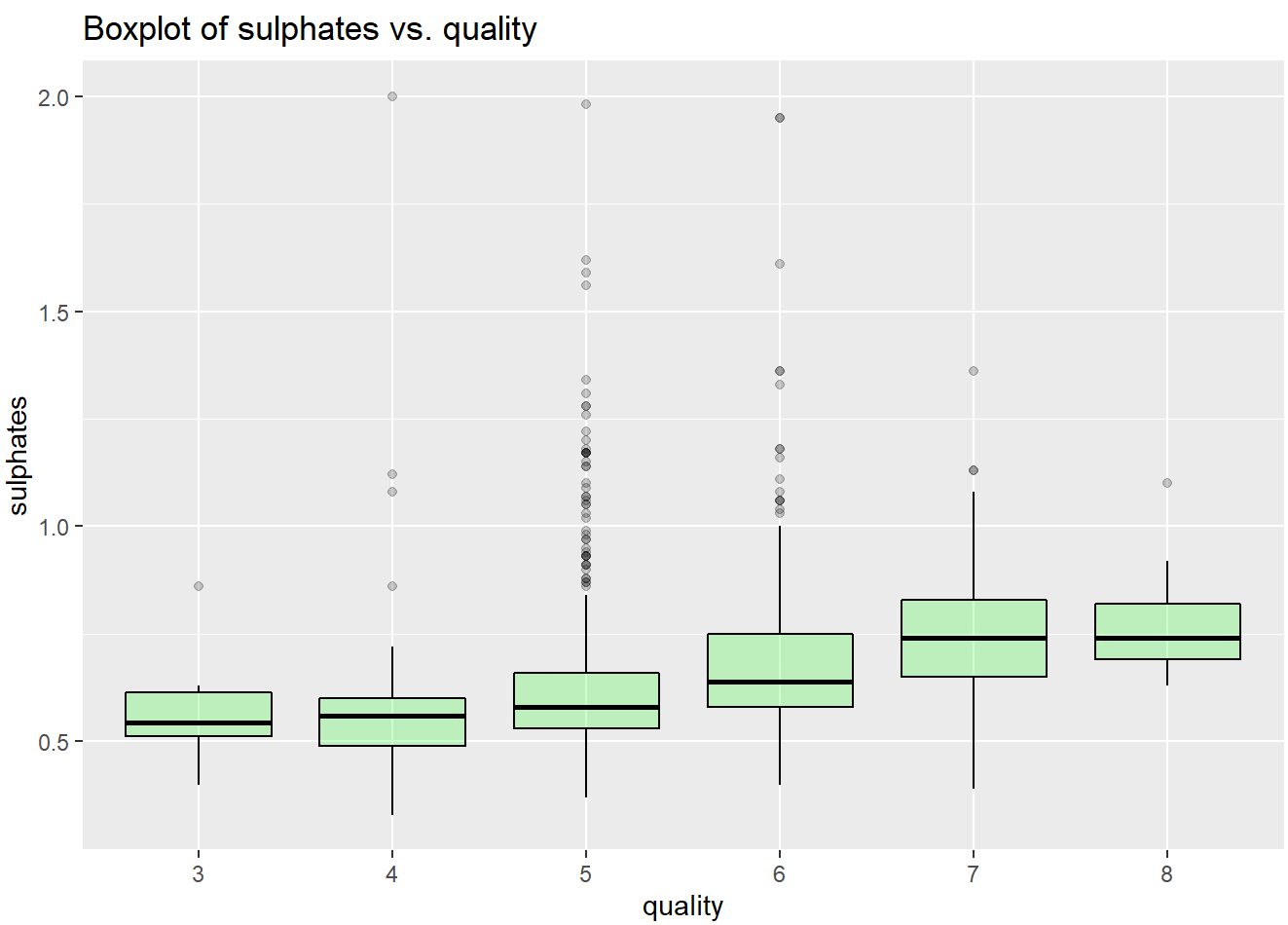


For this boxplot, we used the variable ‘quality’ as our independent variable and ‘citric acid’ as dependent variable. As we can see, higher quality wines have higher values of citric acid.

```
##
##  Pearson's product-moment correlation
##
## data:  yo$citric.acid and yo$quality
## t = 9.2875, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1793415 0.2723711
## sample estimates:
##       cor
## 0.2263725
```

Using the correlation test on both citric and quality, there is a weak positive relationship between the two variables with a value of 0.22.

Boxplot of sulphates vs. quality



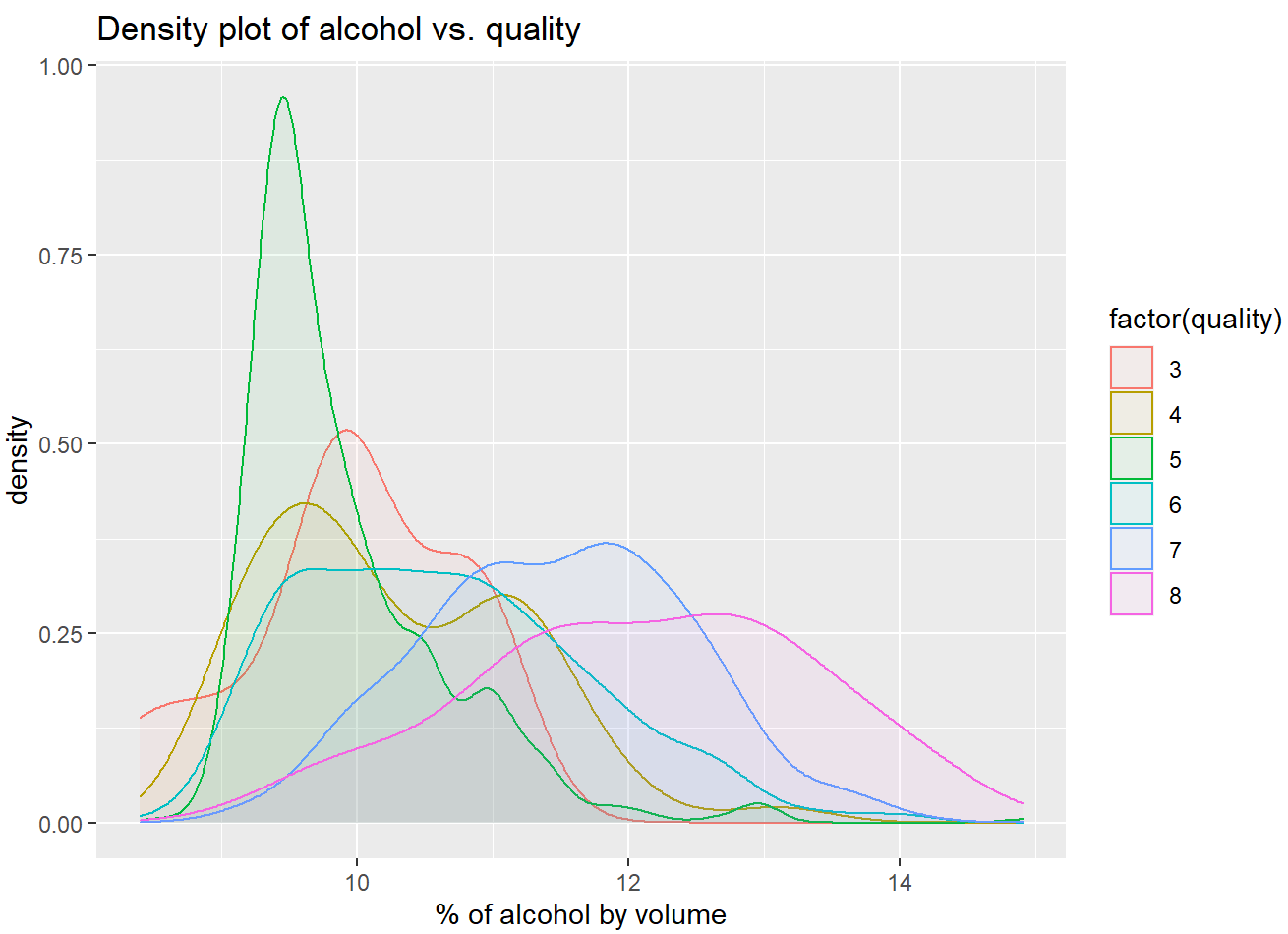
For this boxplot, we used the variable ‘quality’ as our independent variable and ‘citric acid’ as dependent variable. As we can see, higher quality wines have higher values of sulphates, though the difference is not that significant.

```
##
##  Pearson's product-moment correlation
##
## data:  yo$sulphates and yo$quality
## t = 10.38, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2049011 0.2967610
## sample estimates:
##       cor
## 0.2513971
```

Using the correlation test on both sulphates and quality, there is a weak positive relationship between the two variables with a value of 0.25.

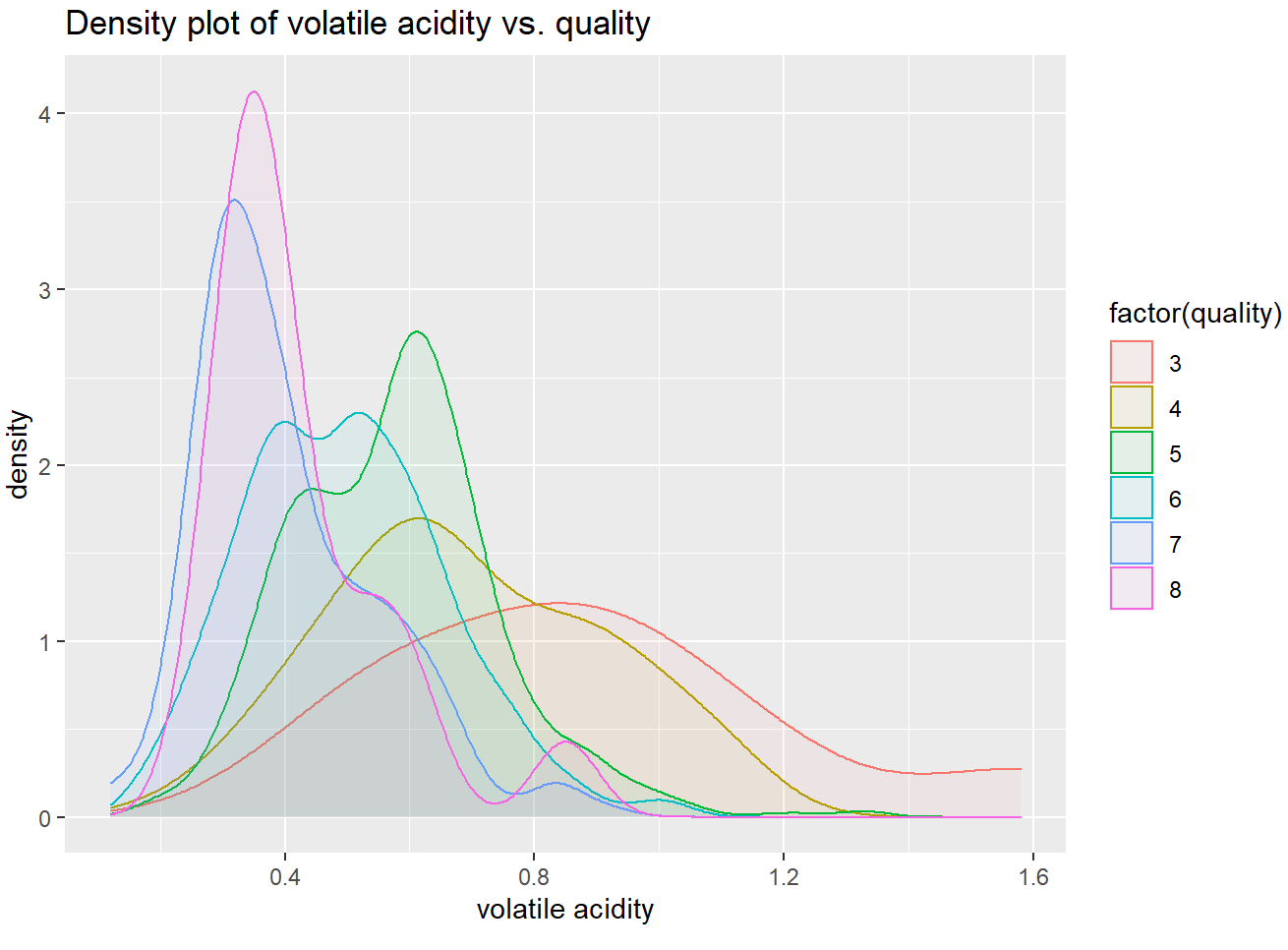
View of quality through density plots

Density plot for alcohol vs. quality



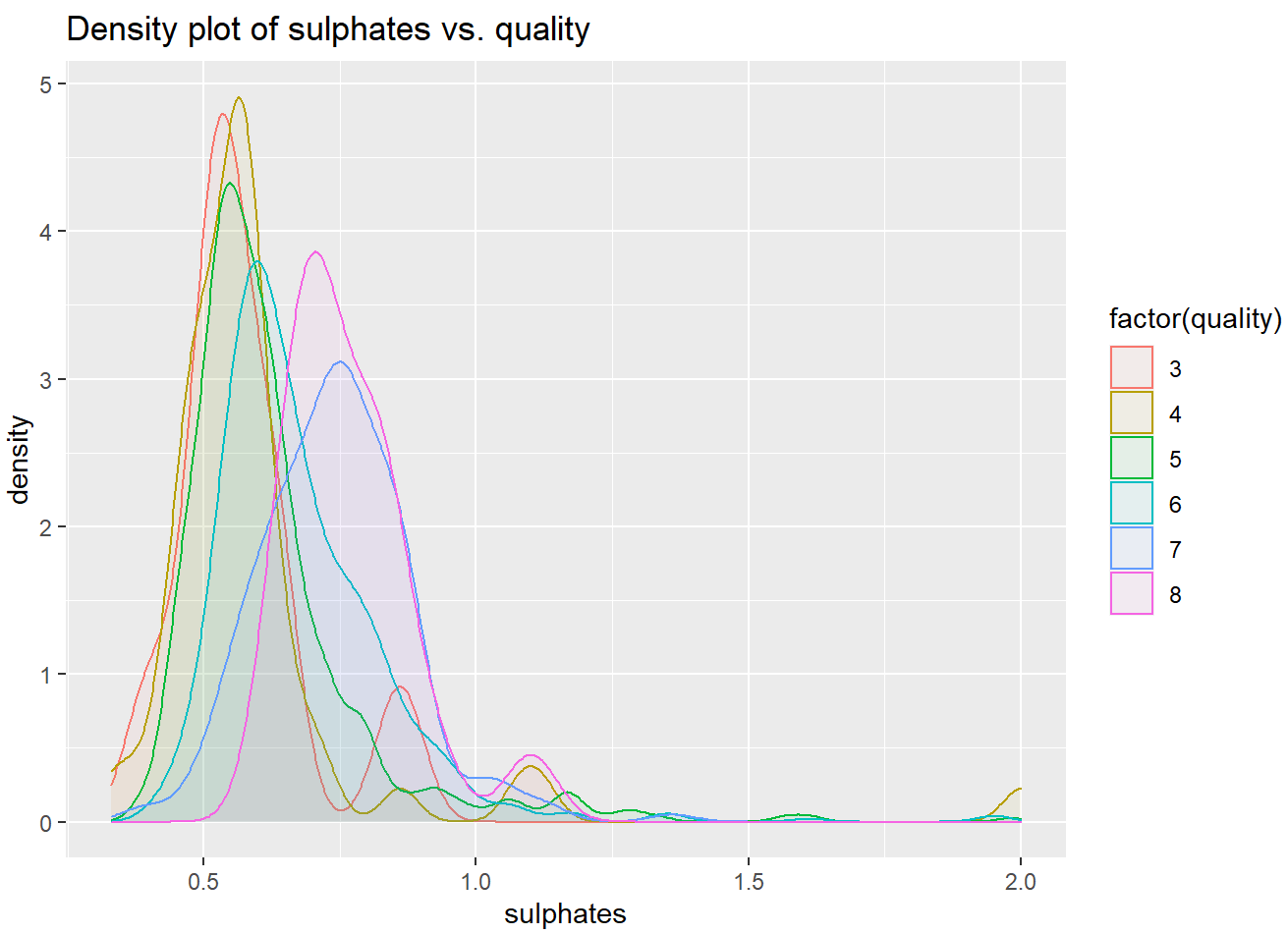
For wines with a quality of 5, we can see a sharp peak at around 8% alcohol. As alcohol increases, higher qualities' peaks move towards right.

Density plot for volatile acidity vs. quality



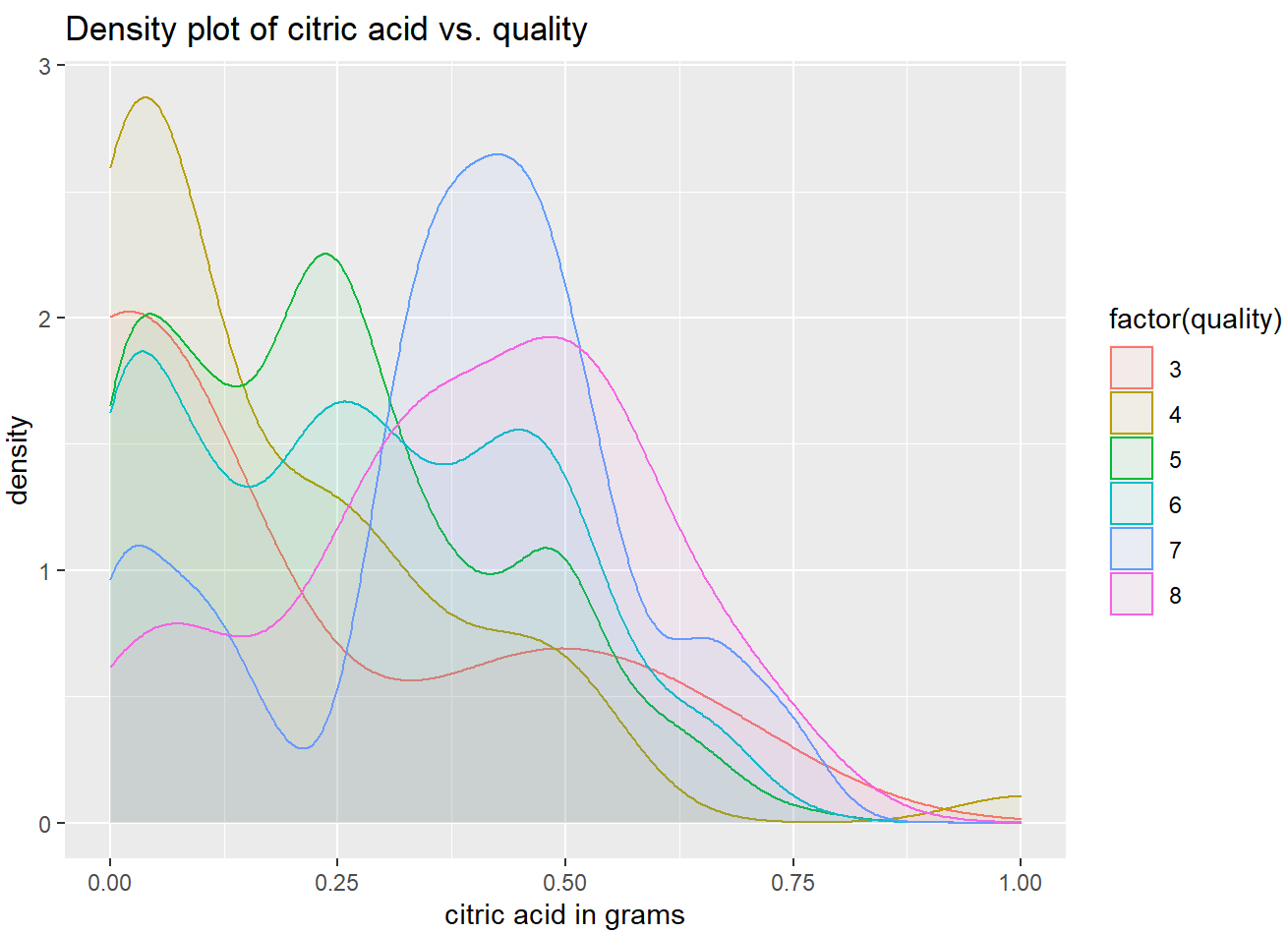
For wines with a quality of 8, we can see a sharp peak at around 0.3 volatile acidity. As volatile acidity increases, lower qualities' peaks move towards right.

Density plot for sulphates vs. quality



Almost all wines in the dataset have sulphates below the 1.0 mark. As sulphates increases, higher qualities' peaks move towards right.

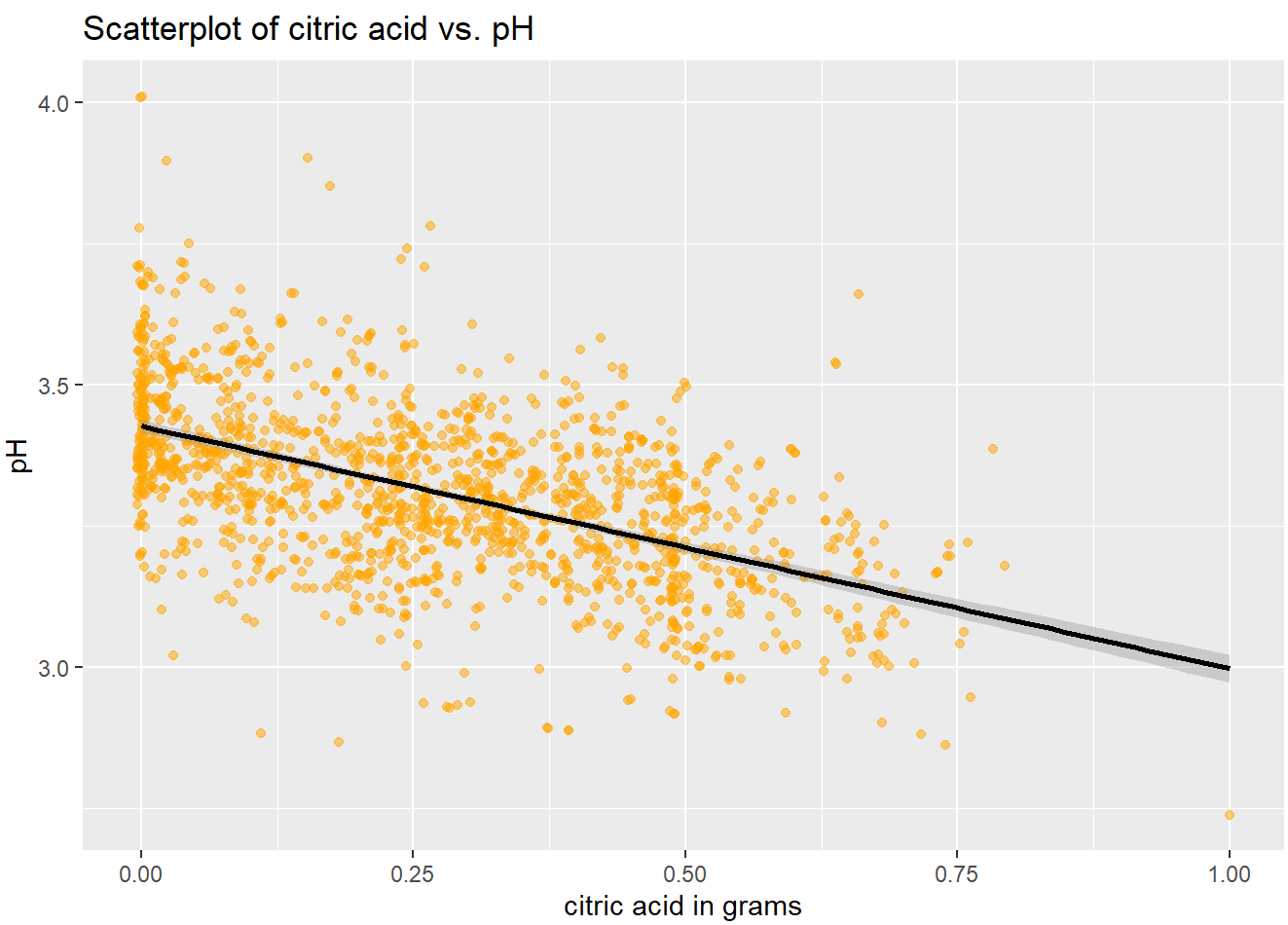
Density plot for citric acid vs. quality



We can see that citric acid increases, the higher qualities' peaks move towards right.

pH relationship with fixed acidity and citric acid

Scatterplot of citric acid vs. pH

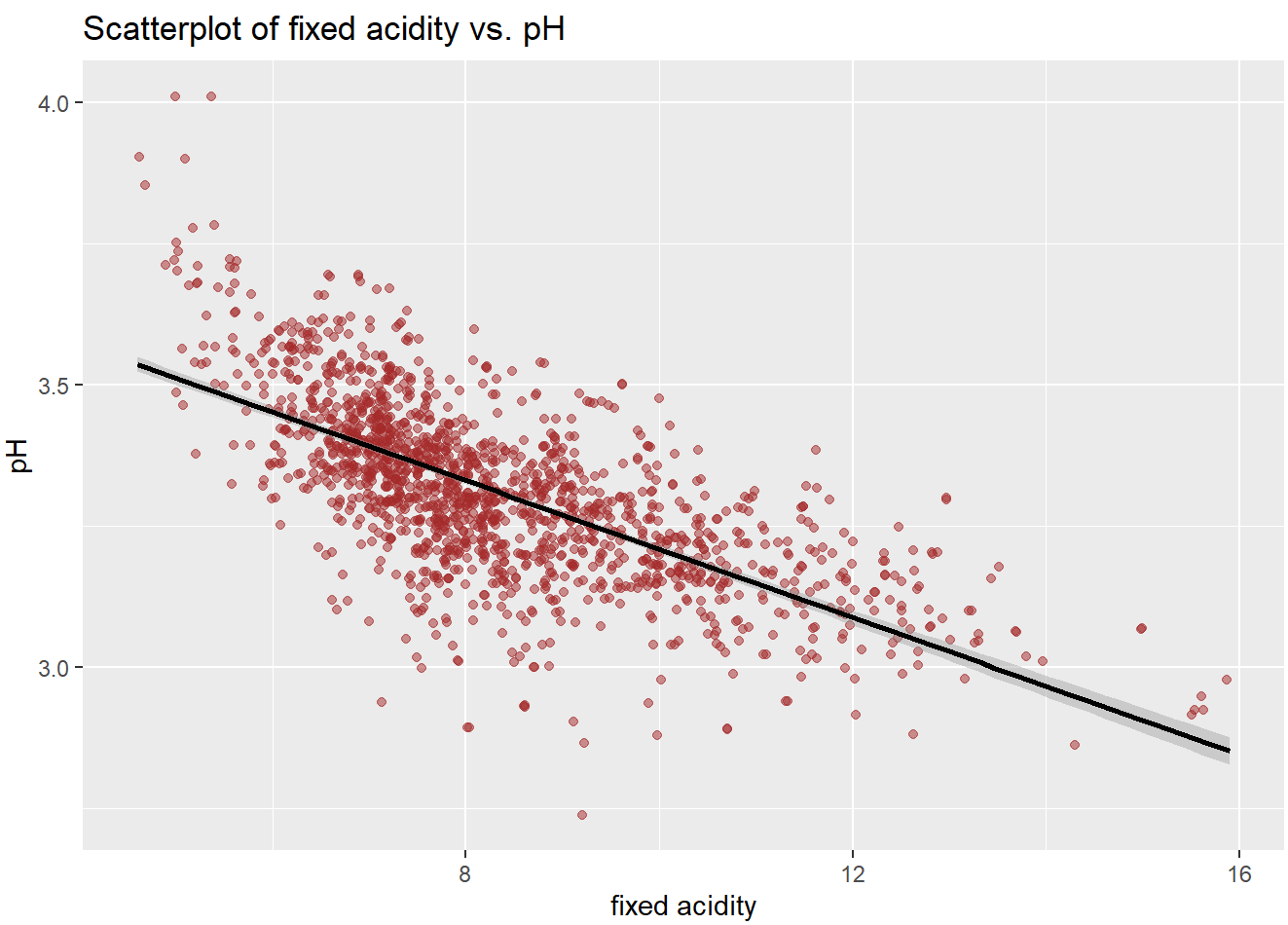


As we can see in the scatterplot, as citric acid increases, pH decreases.

```
##
##  Pearson's product-moment correlation
##
## data:  yo$pH and yo$citric.acid
## t = -25.767, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5756337 -0.5063336
## sample estimates:
##        cor
## -0.5419041
```

Using the correlation test on both pH and citric acid, there is a moderate negative relationship between the two variables with a value of -0.54.

Scatterplot of fixed acidity vs. pH

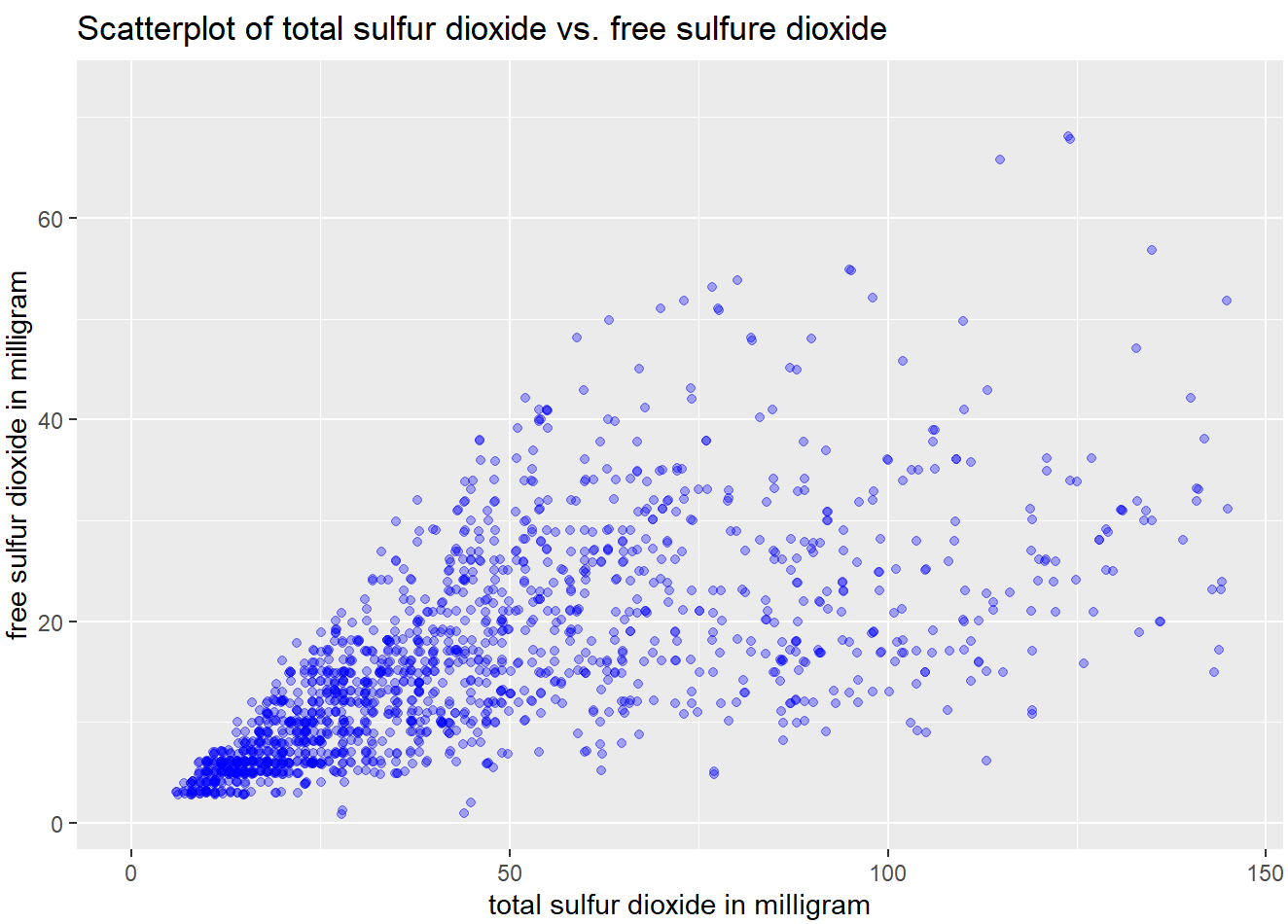


As we can see in the scatterplot, as fixed acidity increases, pH decreases. The same thing also happened when we used citric acid as our independent variable.

```
##
##  Pearson's product-moment correlation
##
## data:  yo$fixed.acidity and yo$pH
## t = -37.366, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7082857 -0.6559174
## sample estimates:
##        cor
## -0.6829782
```

Using the correlation test on both pH and fixed acidity, there is a moderate negative relationship between the two variables with a value of -0.68.

Scatterplot of free sulfur dioxide vs. total sulfur dioxide



As we can see in the scatterplot, as total sulfur dioxide increases, free sulfur dioxide increases.

```
##
##  Pearson's product-moment correlation
##
## data:  yo$free.sulfur.dioxide and yo$total.sulfur.dioxide
## t = 35.84, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6395786 0.6939740
## sample estimates:
##       cor
## 0.6676665
```

Using the correlation test on both free sulfur dioxide and total sulfur dioxide, there is a moderate positive relationship between the two variables with a value of 0.66. This explains why the two variables move in the same direction as the other.

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?

Regarding alcohol, it is interesting that below 13% alcohol concentration, wines with higher qualities have more alcohol content than wines with lower qualities as we saw with the box and density plots. That is not the case though when we go past the 13% alcohol concentration, as the quality of wines starts to decrease when adding more alcohol.

Regarding volatile acidity, when we calculated the correlation between volatile acidity and quality, we got -0.3. This explains why the two variables move in the opposite direction as the other, as we saw in the plots that higher quality wines have lesser volatile acidity.

Regarding citric acid, 0.22 is the correlation between citric acid and quality. Since it is positive, it explains why better wines seem to have more citric acid as we saw in the plots.

Regarding sulphates, better wines seem to have more sulphates as we saw in the plots. Also, the correlation between sulphates and quality is positive.

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

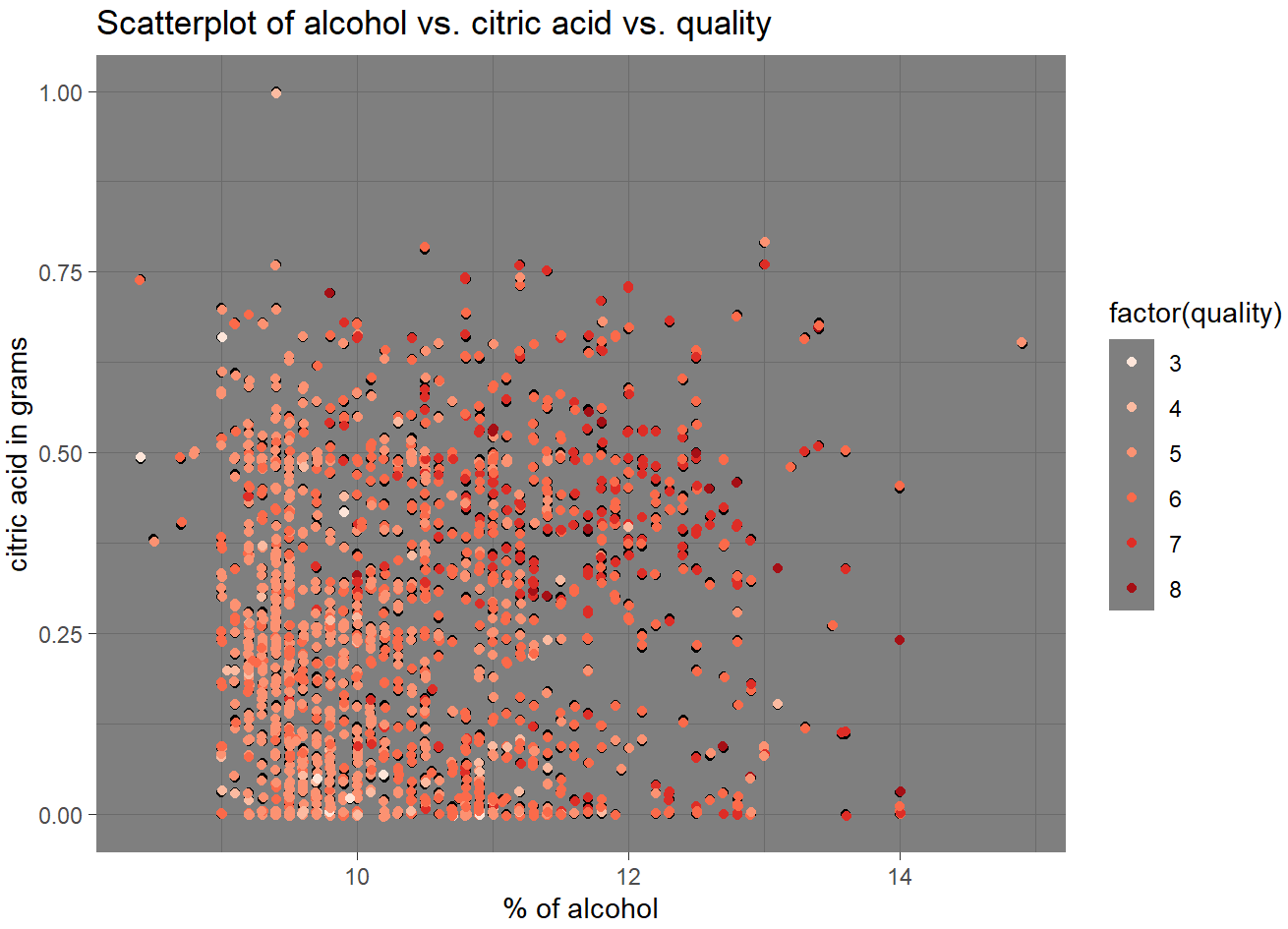
I think it is interesting that the correlation between free sulfur dioxide and total sulfur dioxide is high, which is about 0.66. It is the second highest correlation value that I got throughout this project.

What was the strongest relationship you found?

The relationship between pH and fixed acidity was the strongest relationship I've found. It has a correlation value of -0.68.

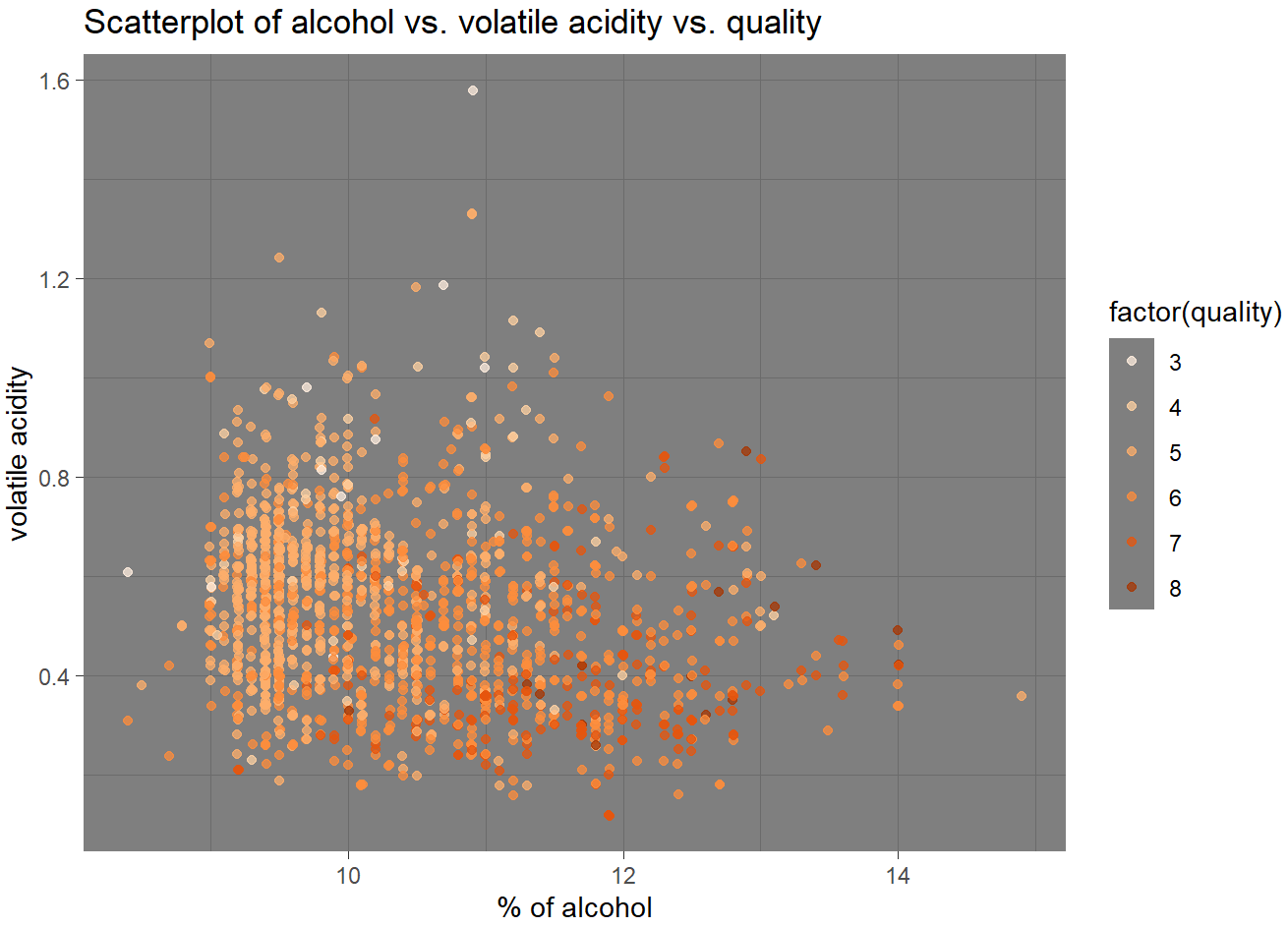
Multivariate Plots Section

Scatterplot of alcohol vs. citric acid vs. quality



In this scatterplot, I used the variable 'alcohol' as our independent variable and 'citric.acid' as dependent variable. The third variable is 'quality.' As we can see, most rated C wines (3-4) have less than 10% alcohol. Most rated A (7-8) and B (5-6) wines, however, have more than 10% alcohol. The difference between A and B wines is that most A wines have higher alcohol concentration than most B wines. Additionally, most A wines have at least 0.25 of citric acid.

Scatterplot of alcohol vs. volatile acidity vs. quality

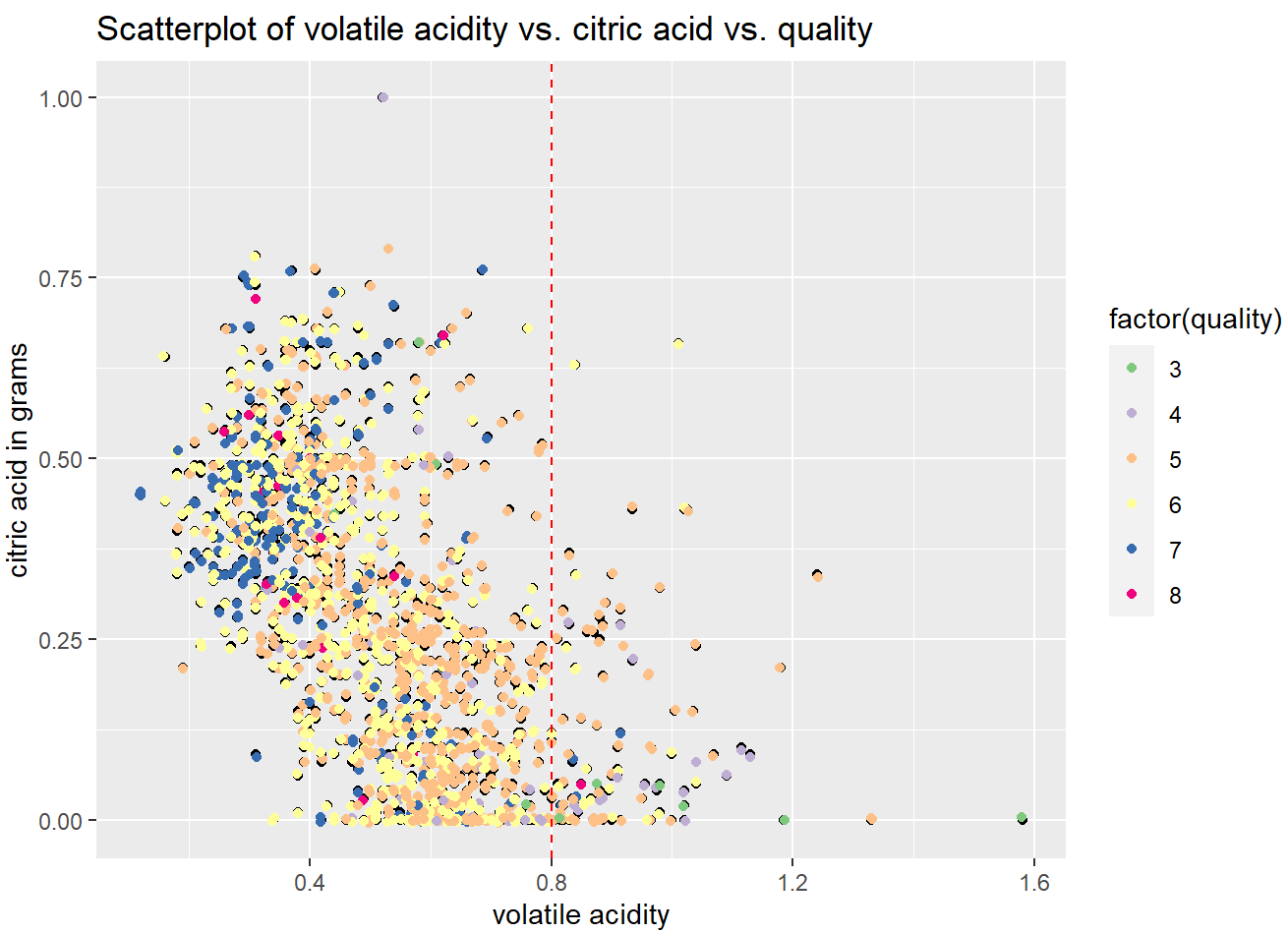


In this scatterplot, I used the variable ‘alcohol’ as our independent variable and ‘volatile.acidity’ as dependent variable. The third variable is ‘quality.’ As we can see, most rated A and B wines have less than 0.4 volatile acidity while most rated C wines have at least 0.4 volatile acidity. As for the alcohol concentration, like we discussed earlier, most A and B wines have more than 10% alcohol while most C wines have less than 10% alcohol.

Acid Talk

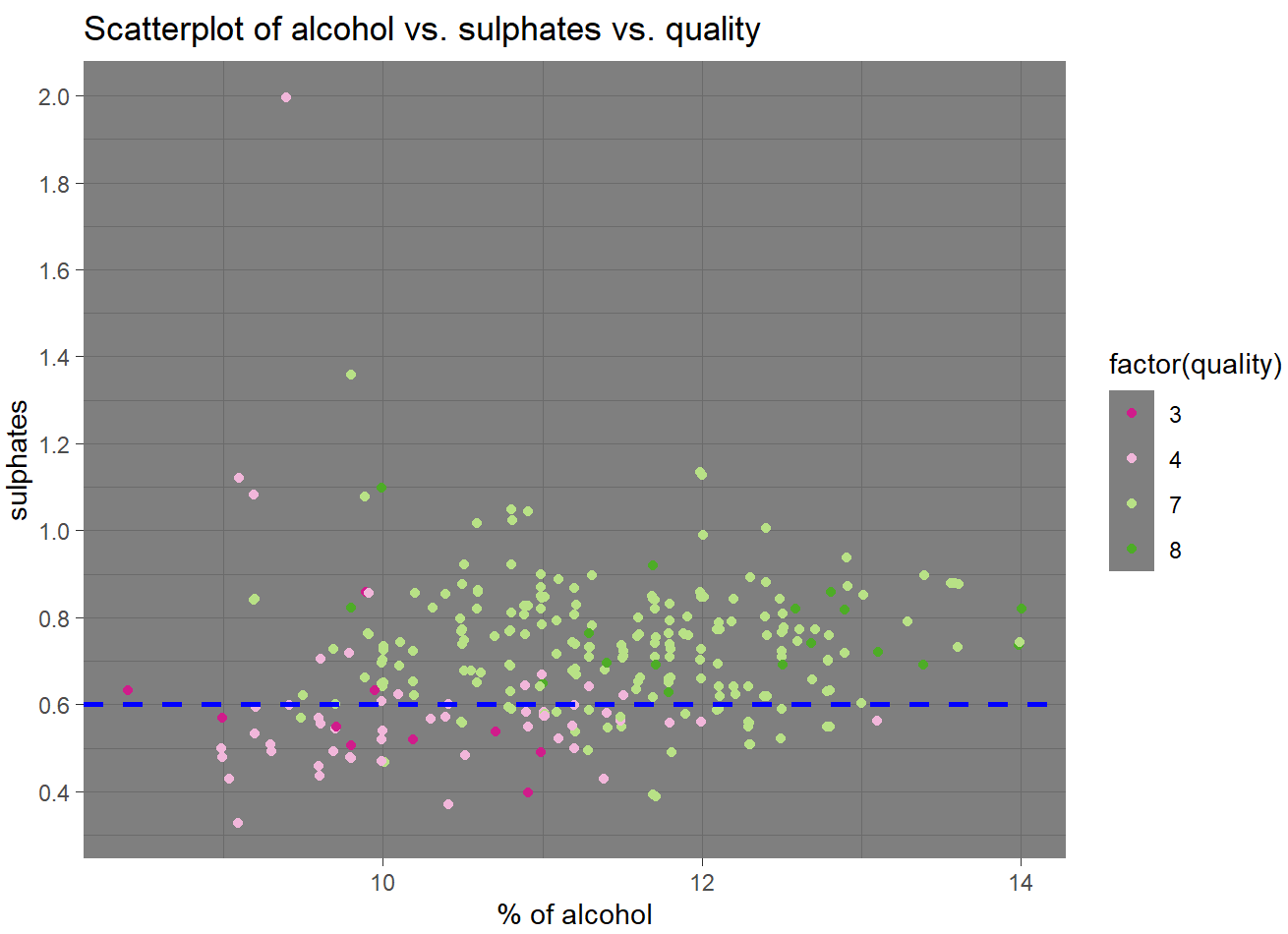
Let’s look at the relationship between volatile acidity and citric acid since they have opposite correlation signs

Scatterplot of volatile acidity vs. citric acid vs. quality



In this scatterplot, I used the variable ‘volatile acidity’ as our independent variable and ‘citric acid’ as dependent variable. The third variable is ‘quality.’ The majority of all the wines in the dataset have a volatile acidity of less than 0.8. Most rated A wines have at least 0.25 citric acid, while most rated B wines have at least 0.0 citric acid. Most C wines are all over the place in the graph. Regarding volatile acidity, as we discussed earlier, most rated A and B wines have less than 0.4 volatile acidity, while some rated C wines have a volatile acidity that is above 0.8.

Scatterplot of alcohol vs. sulphates vs. quality



In this scatterplot, I used the variable ‘alcohol’ as our independent variable and ‘sulphates’ as dependent variable. The third variable is ‘quality.’ Rated B has been removed to clearly see the difference between good and bad wines. As we can see, most rated A wines have at least 0.7 sulphates, while most C wines have less than 0.6 sulphates. Rated A wines also have more alcohol content than rated C wines with at least 10% alcohol.

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?

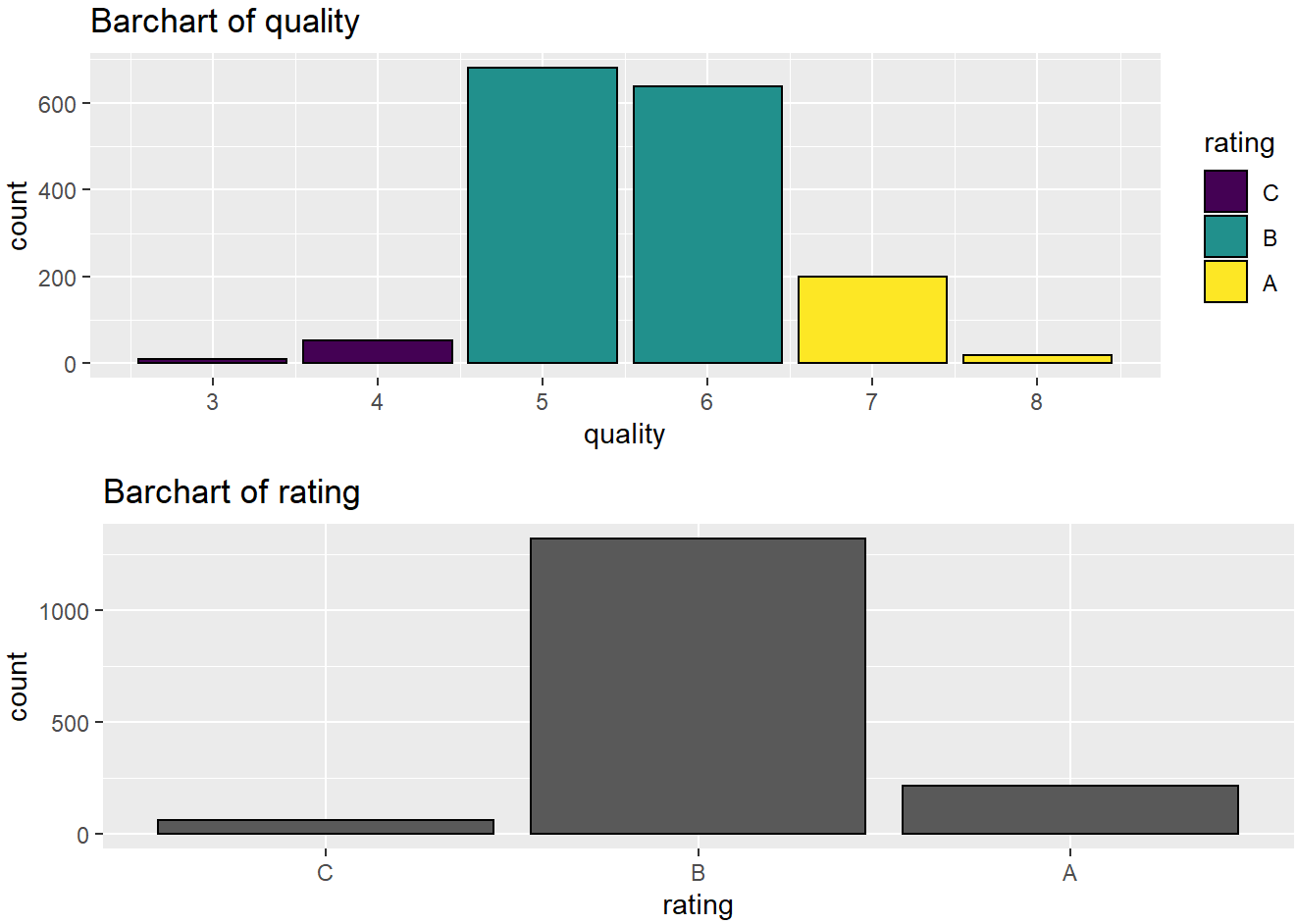
We saw in the graphs that most rated A wines have a higher citric acid, higher alcohol content, and lower volatile acidity. This tells me that these characteristics contribute towards better wines.

Were there any interesting or surprising interactions between features?

The graph relating alcohol and sulphate is interesting because we can clearly see the difference between rated A wines and rated C wines. Also, regardless of alcohol content, it is interesting that nearly all wines have less than 1.0 sulphates.

Final Plots and Summary

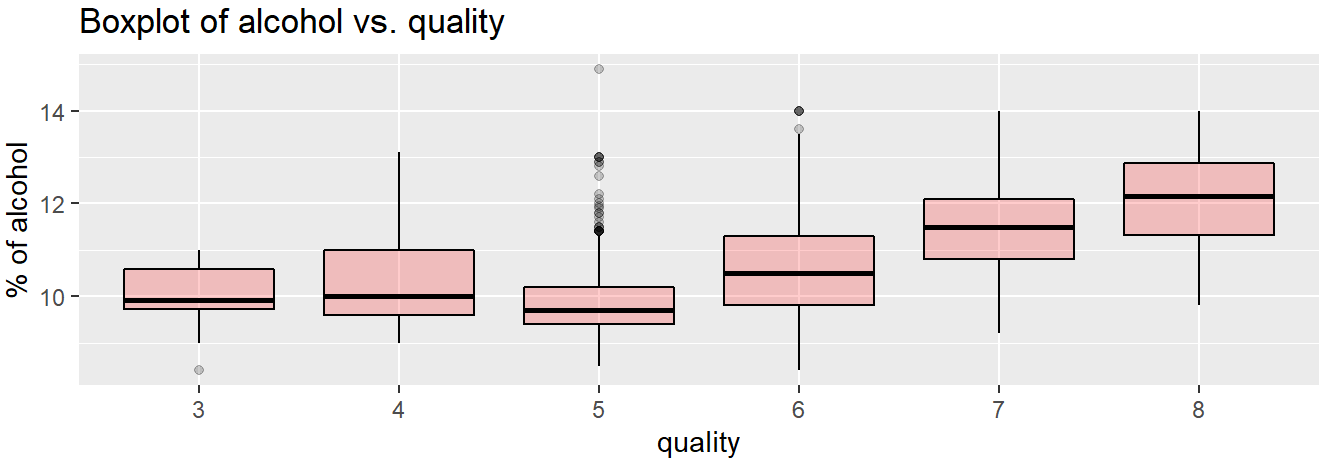
Plot one

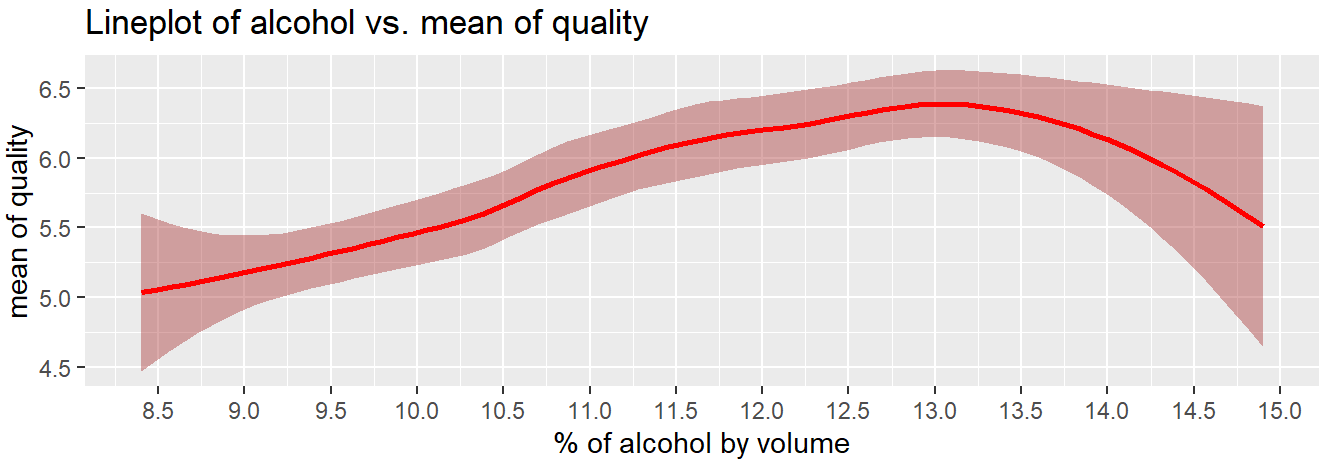


Description one

This bar graph is from the univariate section. This graph is really important because this is where we got the idea of making a ‘rating’ variable, which is dividing the quality levels into three ratings: A, B and C. We used the ‘rating’ variable throughout our analysis on the red wines dataset. Also, this graph gave us the information that most wines are of quality 5 and 6, as they make up 82.5% of the dataset while the rest of the qualities have a combined percentage of 17.5%. Obviously, this dataset is limited, which is why it is difficult to determine the characteristics of a good wine.

Plot two





Description two

The boxplot is from the bivariate section and it shows the relationship between alcohol and quality. As we can see, the alcohol content gets higher as the quality of wines increases. We also did a correlation test between alcohol and quality, and there is a positive relationship between the two. There's definitely a significant difference of alcohol content if we compare quality 3 to quality 8.

This leads to the question, does more alcohol gives us better wine?

Looking at the line plot, we can see that the quality of wines gets higher as alcohol content increases, but the trend stops at 13%. As we go past 13%, the quality of wines gets lower as alcohol content increases. More alcohol doesn't necessarily give us better wine.

Plot three

Description three

This scatterplot is from the multivariate section and it shows the relationship between citric acid and alcohol on quality. As we can see, nearly all wines with qualities of 3 and 4 (Rated C) have alcohol content no more than 10%. The opposite is true for wines with qualities of 5 and above (Rated A and B), as most have alcohol content above 10%. Most rated A wines, however, have more alcohol content than most rated B wines. Regarding citric acid, most A wines have at least 0.25 citric acid, while most B wines have at least 0.0 citric acid. C wines are also like B wines but it is difficult to see in the graph since there are not a lot of them in the dataset.

Reflection

The red wine dataset consisted of 1599 red wines with 12 variables from 2009. I added an extra variable called ‘rating,’ which made the total number of variables to 13. I started my analysis by making a graph for each variable to get a feel of the numbers and to figure out any interesting relationships between different variables. From the univariate analysis, I took advantage of using histograms to see the highs and lows of each variable. Using log transform (log10) also helped me understand each variable by comparing a normal graph to a graph with a logarithmic function. From the bivariate analysis, I used three different plot types: box, scatter and density. These plot types helped me figure out which variables contribute to better wines. Using correlation tests also gave me an understanding of the correlation between two different variables. Finally, from the multivariate analysis, I used scatterplots like in the bivariate analysis but with a third variable, which is ‘quality.’ This gave us a stronger understanding of the data, specifically learning which factors contribute to the quality of wines.