# Red Wine Quality Prediction

by Jie Hu, Email: jie.hu.ds@gmail.com (mailto:jie.hu.ds@gmail.com)

This markdown will use explorsive data analysis to figure out which attributes affect quality of red wine significantly. To do this, I use the dataset (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityReds.csv) including the quality rate by at least 3 experts and the chemical properties of the wine. This dataset might indicate how current experts, representing the test nowadays, think what a good red wine is.

# Univariate Plots Section

## Explore part

To begin with, let's summarise the data:

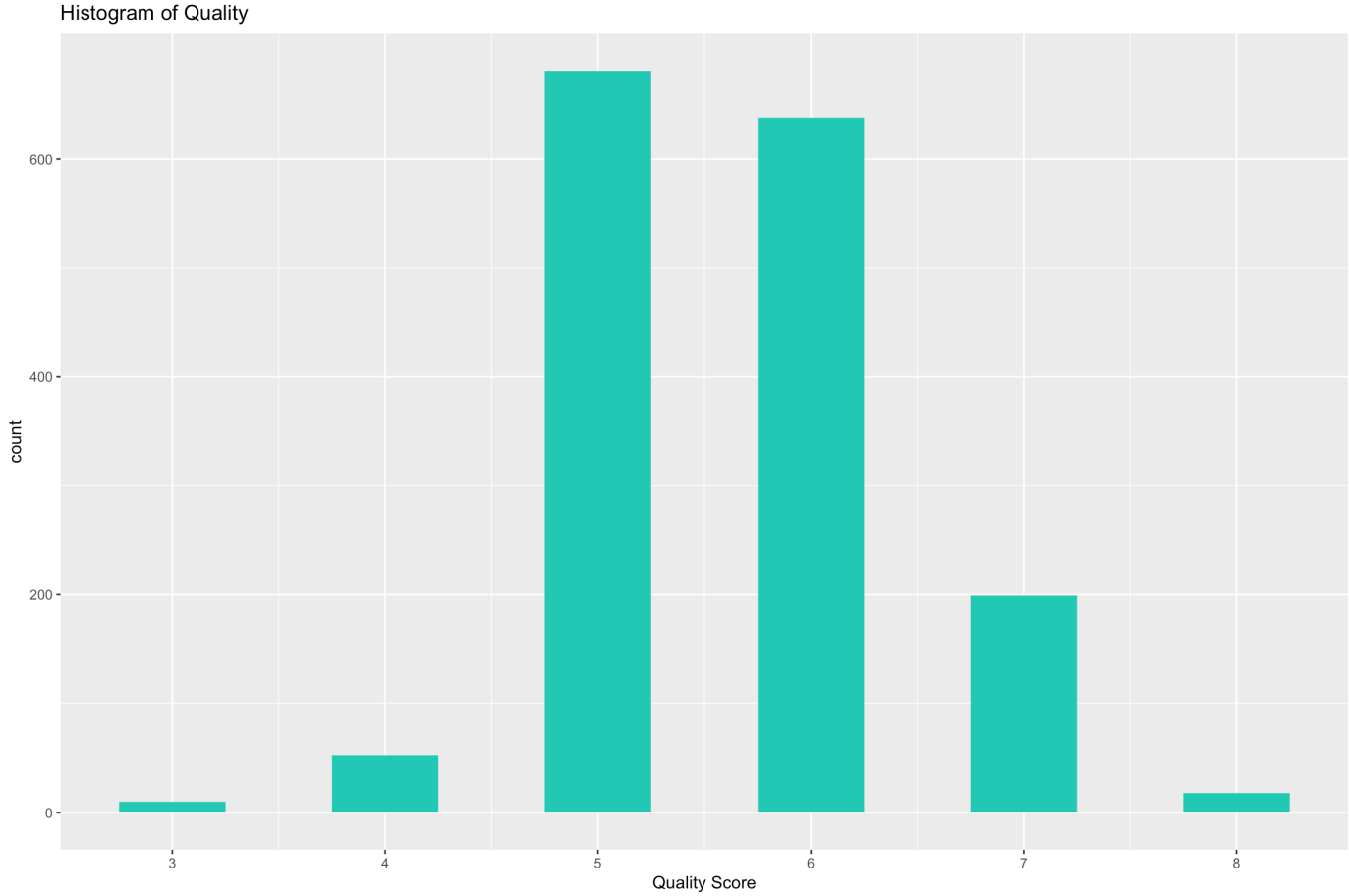
```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : 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 0.07
3 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 ...
```

The dataset includes 1599 observations with 12 variables.

Then let's explot the variables one by one. Because all the variables are numeric, I will mainly use histogram to explore and figure out if there're something interesting worth further steps.

**- quality -**

Quality is what this report concerns with, the rate here represent average rate from at least 3 experts. First let's see how the quality of 1599 wine distributed.



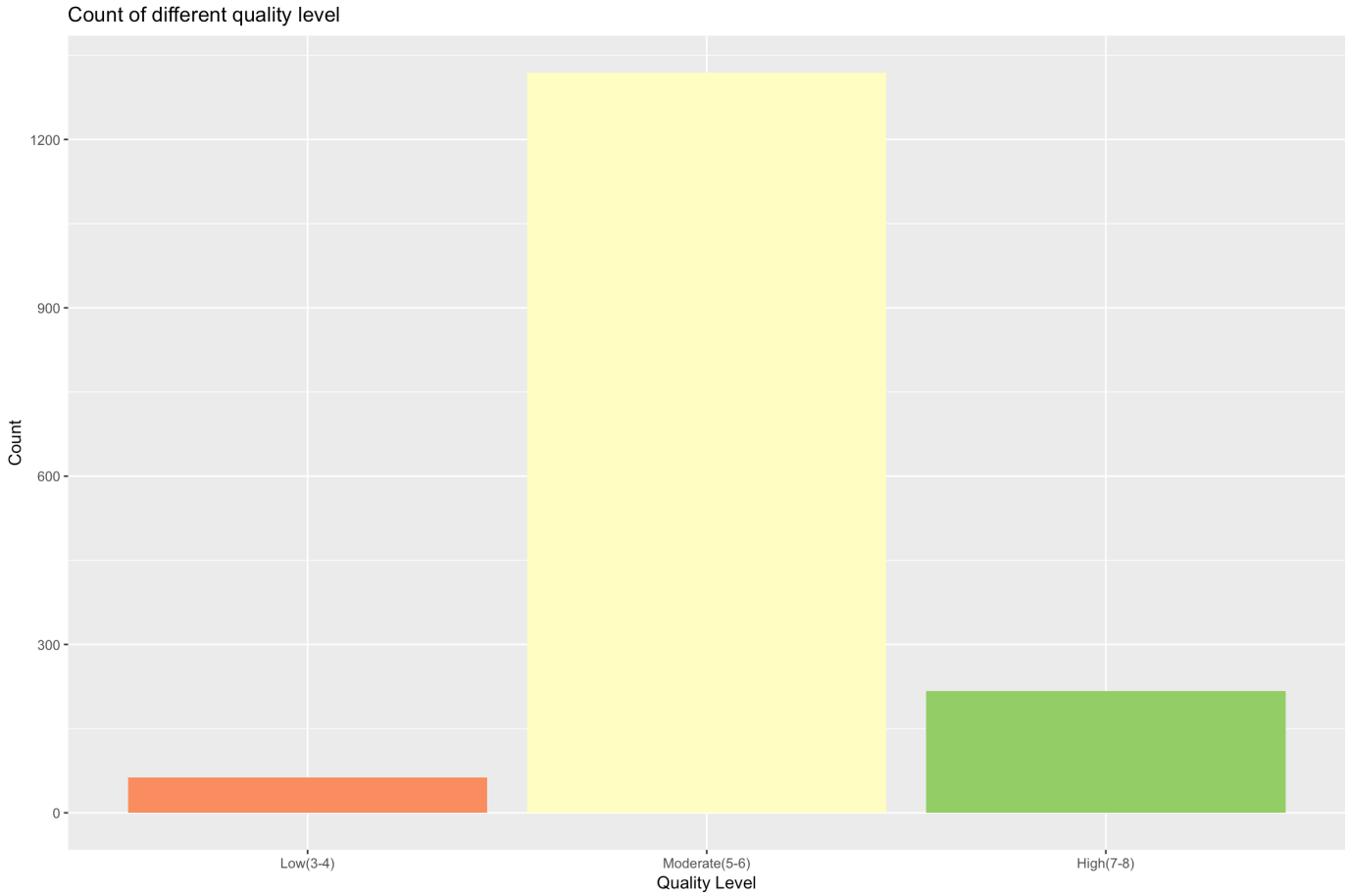

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

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

Quality, ranging from 3-8, is integer type data. About 82.5% observations get 5-6 ratings, while only 14.2% (227 counts) got 3,7 or 8 scores on quality rating. Because the score were average made by 3 or more experts.

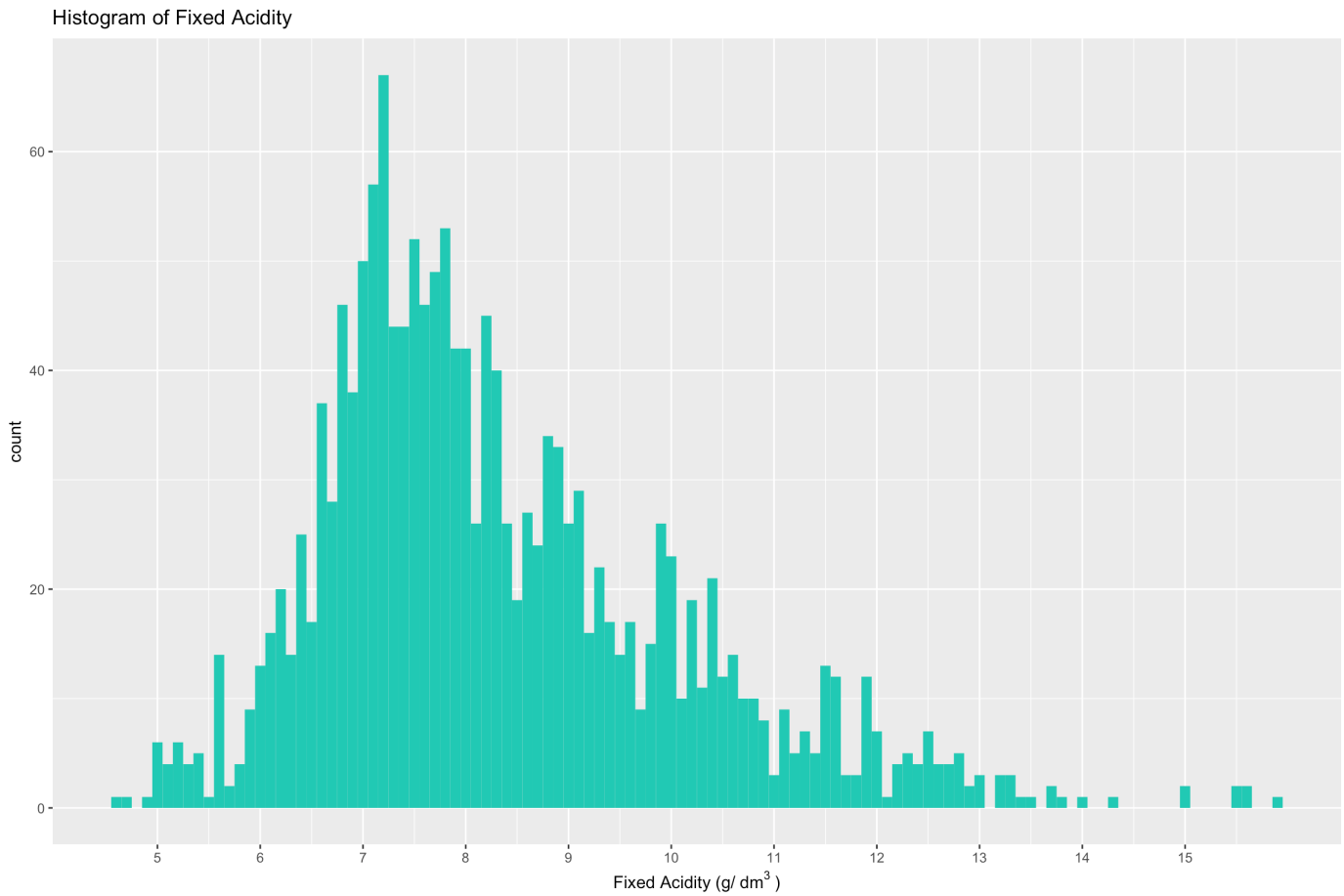
Now let's check quality by dividing them into 3 groups: Low"(3-4), "Moderate"(5-6), "High"(7-8)

```
##
##      Low(3-4) Moderate(5-6)     High(7-8)
##            63          1319           217
```




The moderate quality wines takes up most part and we have much less data on low/high quality wines, which might produce some bias in this analysis. But I will assume the data is reliable and trustworthy.

### -fixed.acidity-



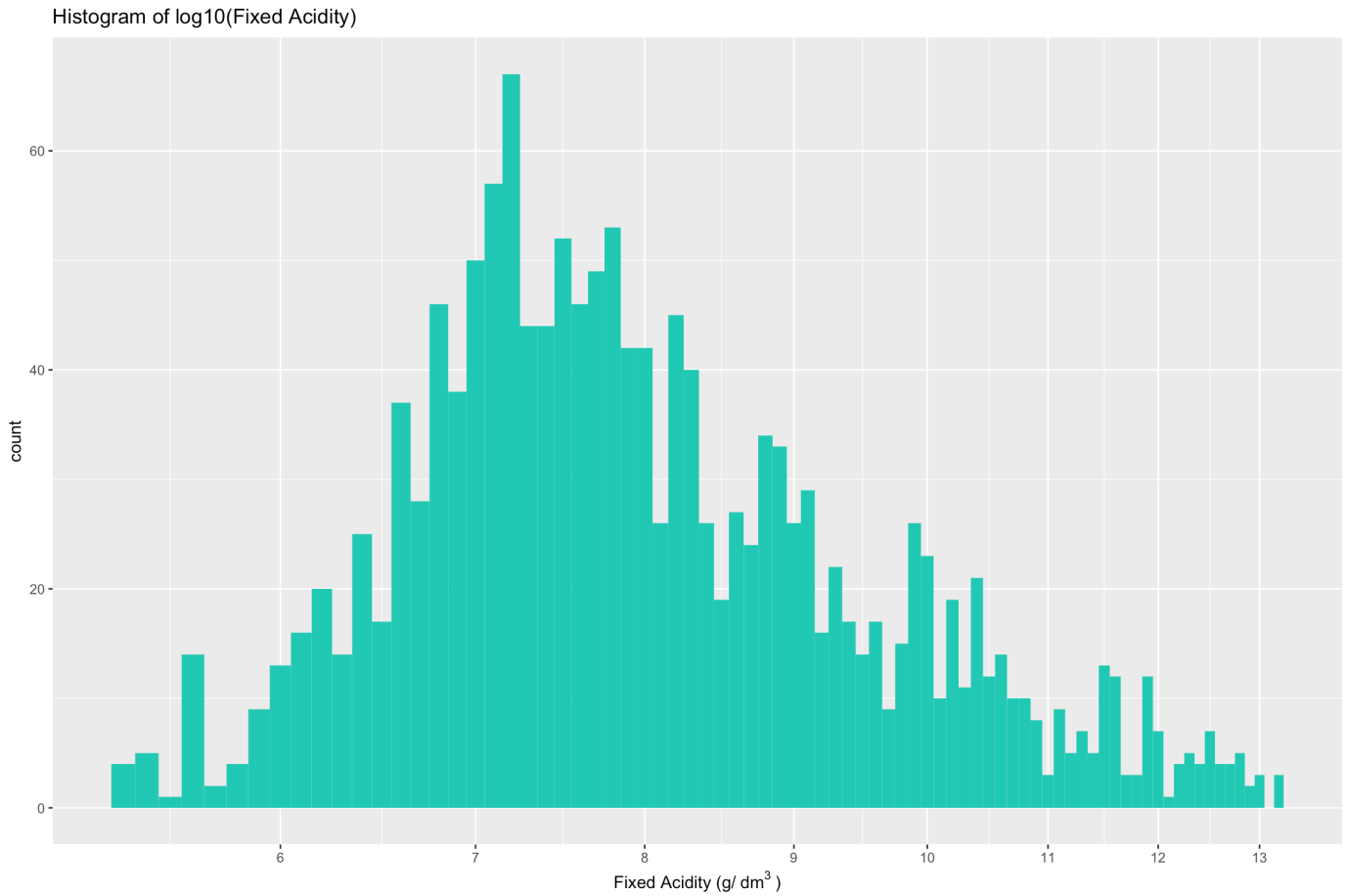
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

“fixed.acidity” is a measure of inside liquid concentration. The histogram a little right-skewed distributed with some outliers located at right side. The most frequent values are between 7-8. IQR is 2.1. The maximum is a little far from 3rd quantile compared with IQR.

To improve the plot, I removed the top 1% and bottom 1% data and use log transformation:

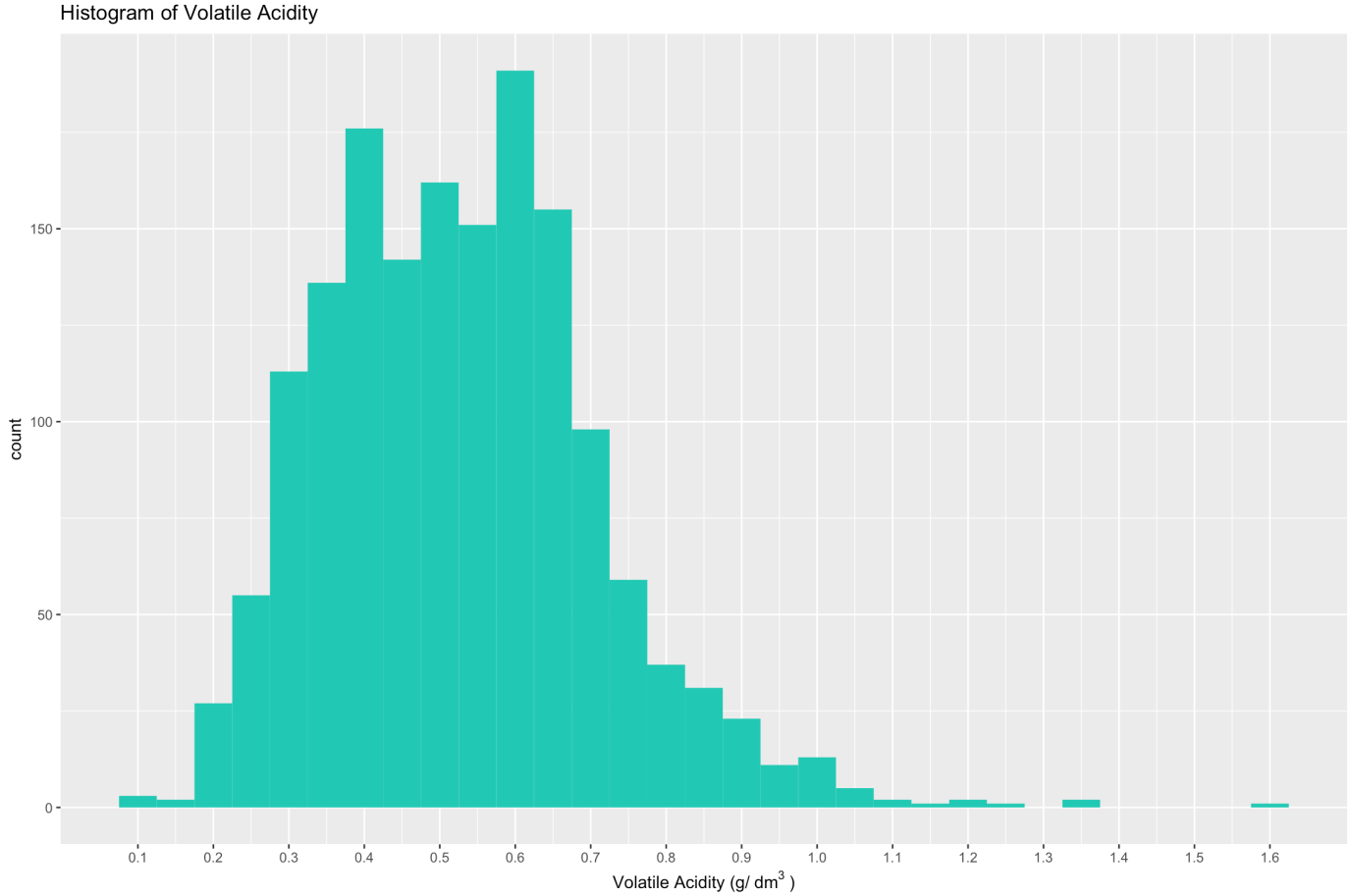
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.300   7.100   7.900   8.294   9.200  13.200
```

Now the 3rd quantile is not far from max value, and data gathers more around center.




It looks much better, symmetric without too much attention on outliers. We can see majority fixed acidity gathering in middle part.

**-volatile.acidity-**



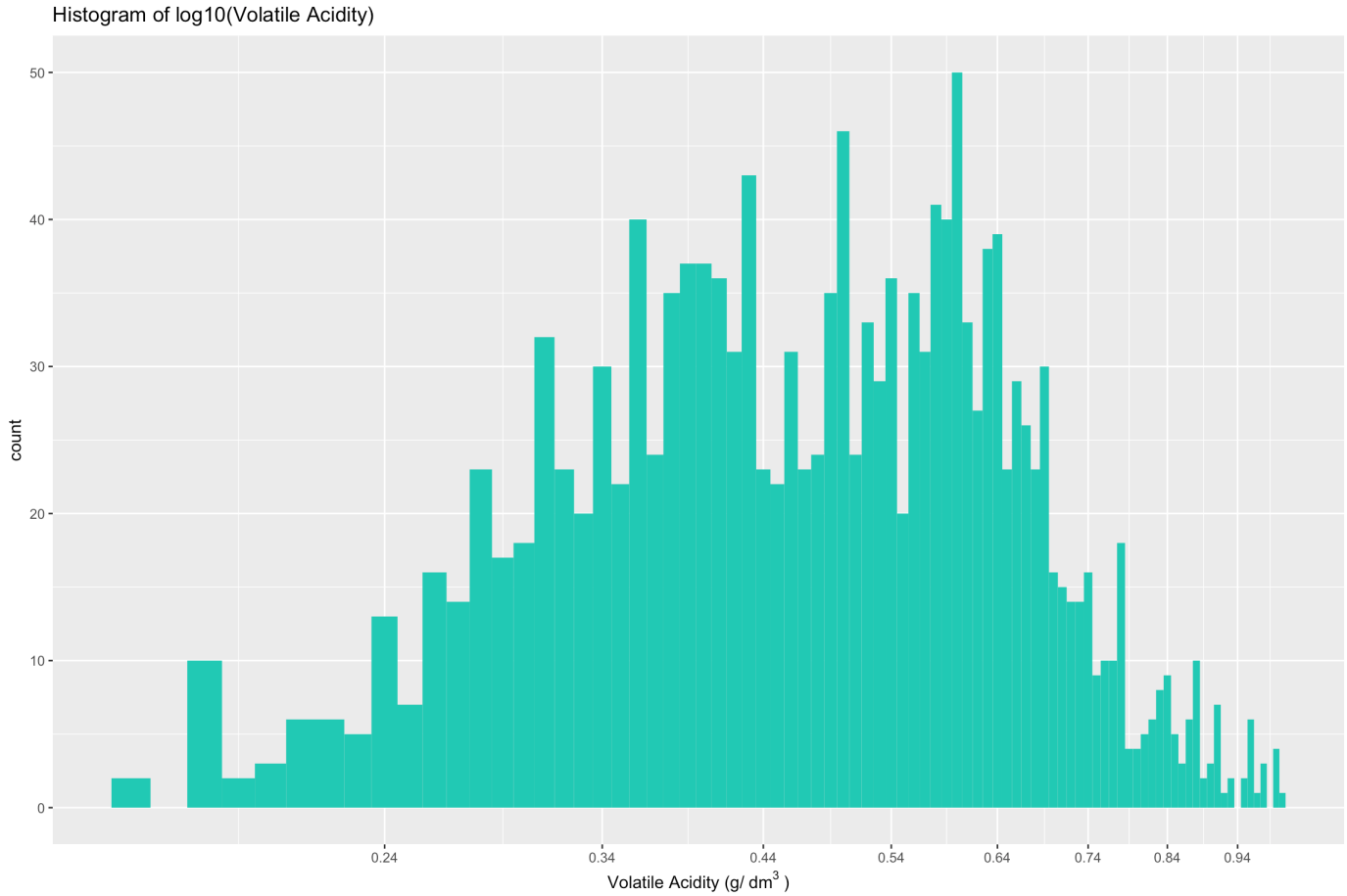

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

“volatile.acidity” is measure of acidity above-surface of liquid. The histogram is right-skewed distributed with some outliers located at right side. The most frequent values are between 0.4-0.6. IQR is 0.25.

To improve the plot, I removed the top 1% outliers and use log transformation:

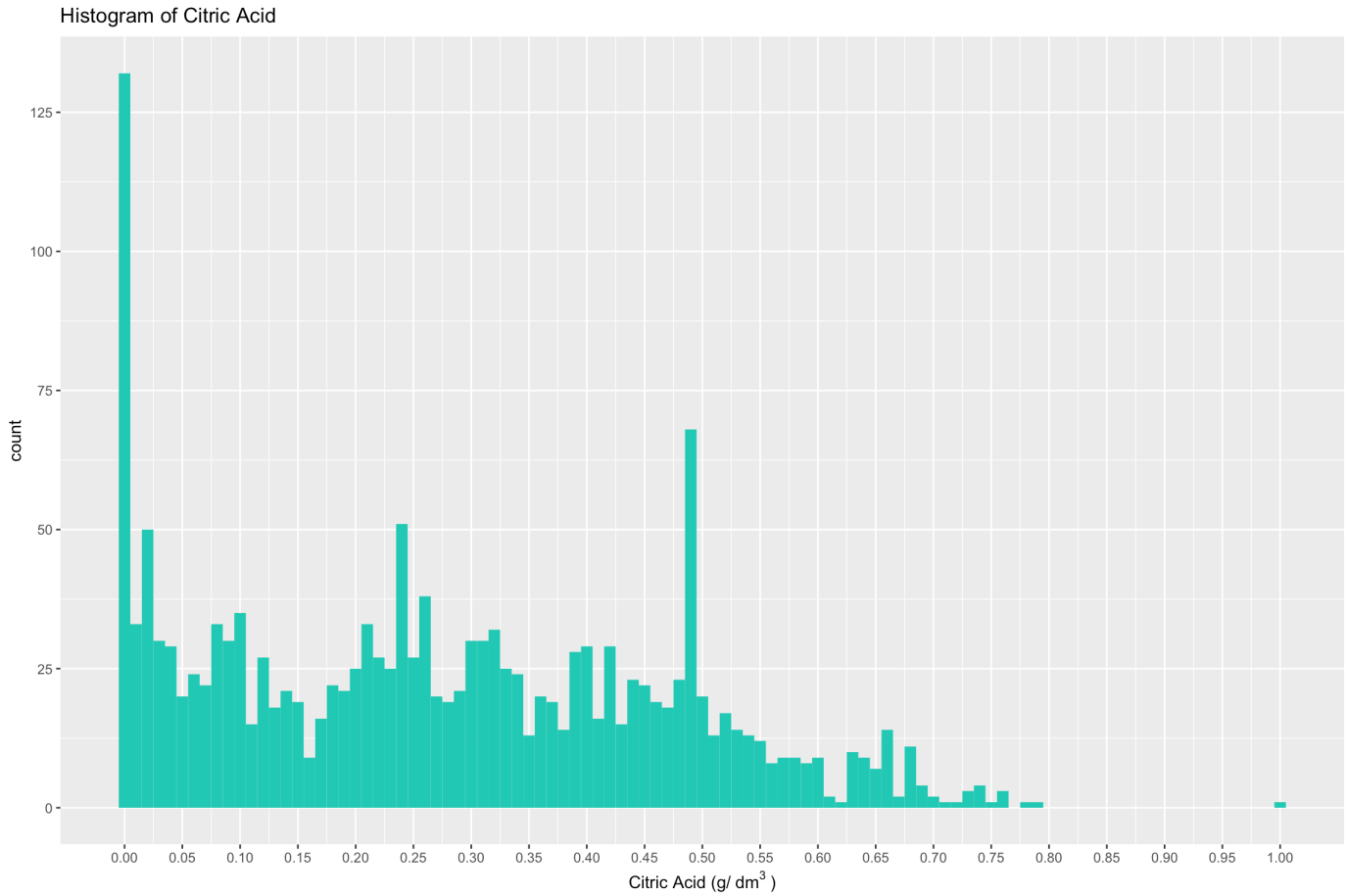
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1600  0.3900  0.5200  0.5213  0.6300  1.0100
```

Now the IQR is 0.11, data concentrates more on center.




It looks much better, symmetric and bell shaped without extreme outliers. We can see majority volatile acidity gathering in middle part. The most common log10(volatile.acidity + 1) values are between 0.19 and 0.21

**-citric.acid-**



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

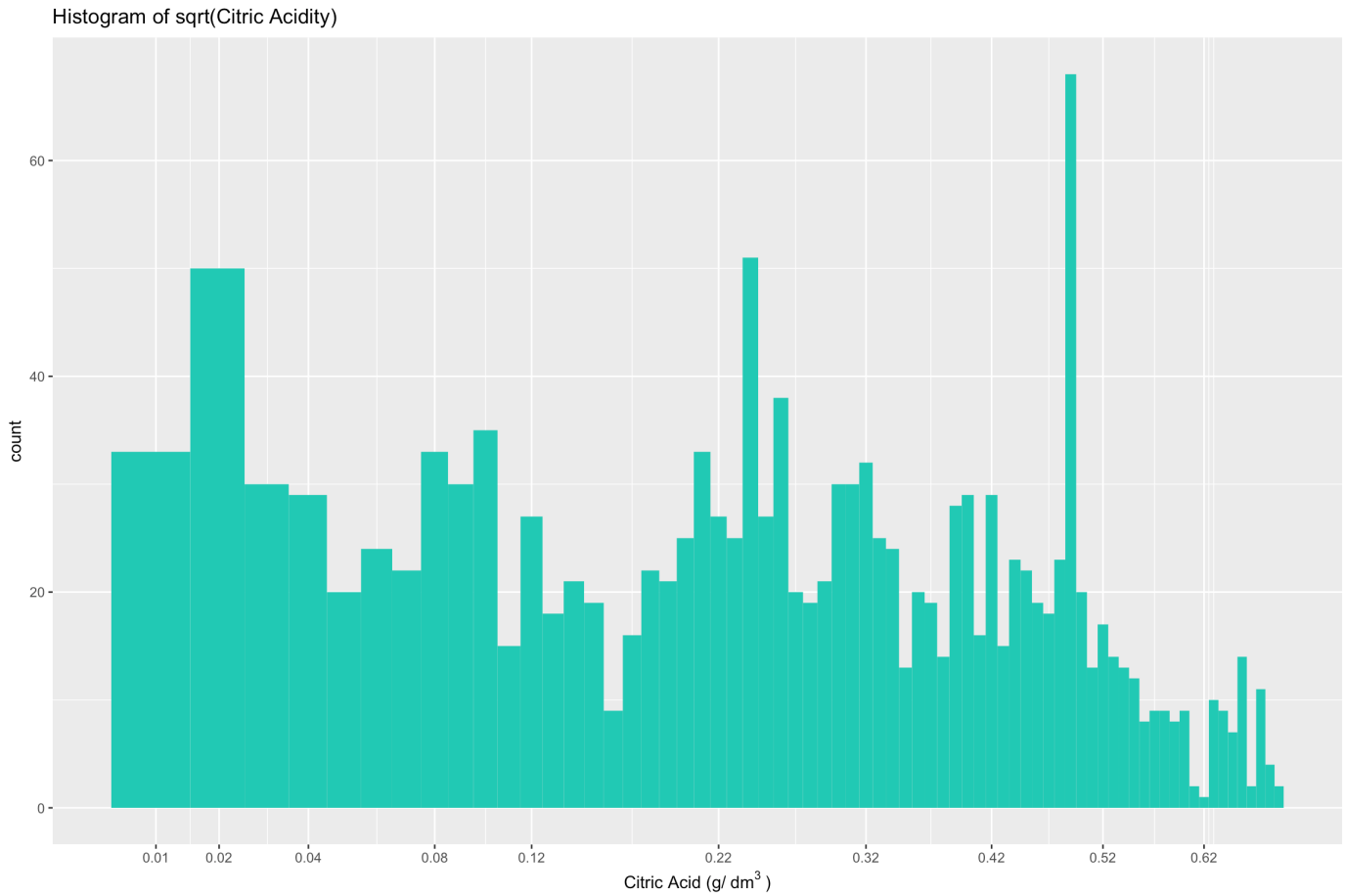
“citric.acid” is right-skewed distributed with some outliers located at very right side. The most frequent values 0. It’s also interesting a lot of wine have citric.acid = 0, IQR is 0.33.

To improve the plot, I just removed the top 1% outliers and use log transformation (the downside has a lot of value without citric acid so it’s better to keep them):

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.1300  0.2800  0.2902  0.4400  0.7000
```

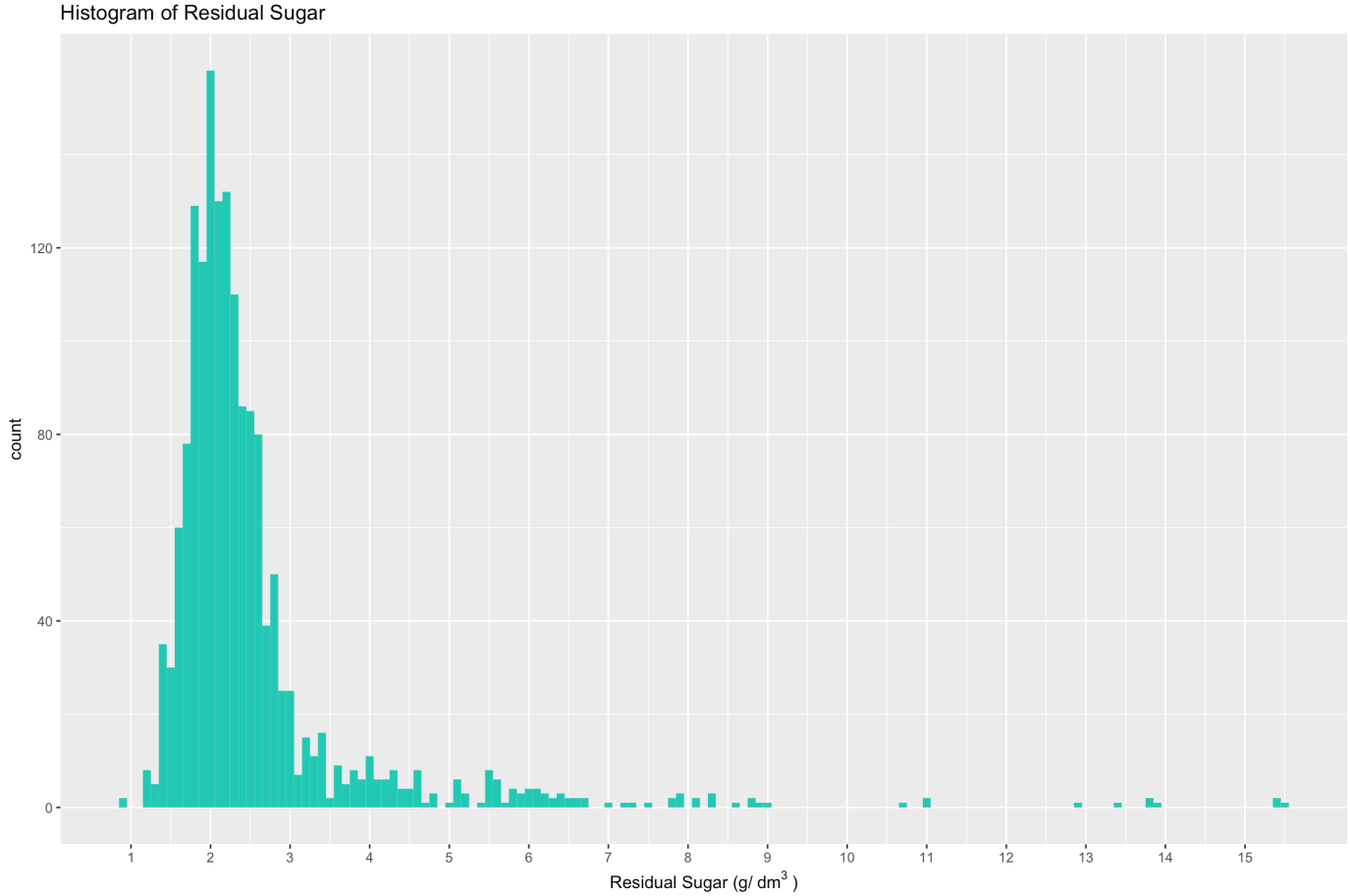
Now the IQR is 0.31, data gathers more in center.

This time, I use square root instead of log10, because log10 plot has binwidth varies significantly.




It looks much better. But notice the scale of citric acid is not evenly distributed, instead, the scale is exponentially decreased. This plot can let us more focus on majority citric acid gathering in middle part.
The most common citric.acid values are around 0.49. Before it was 0.1, by using higher accurate, we can see it has been changed.

**-residual.sugar-**



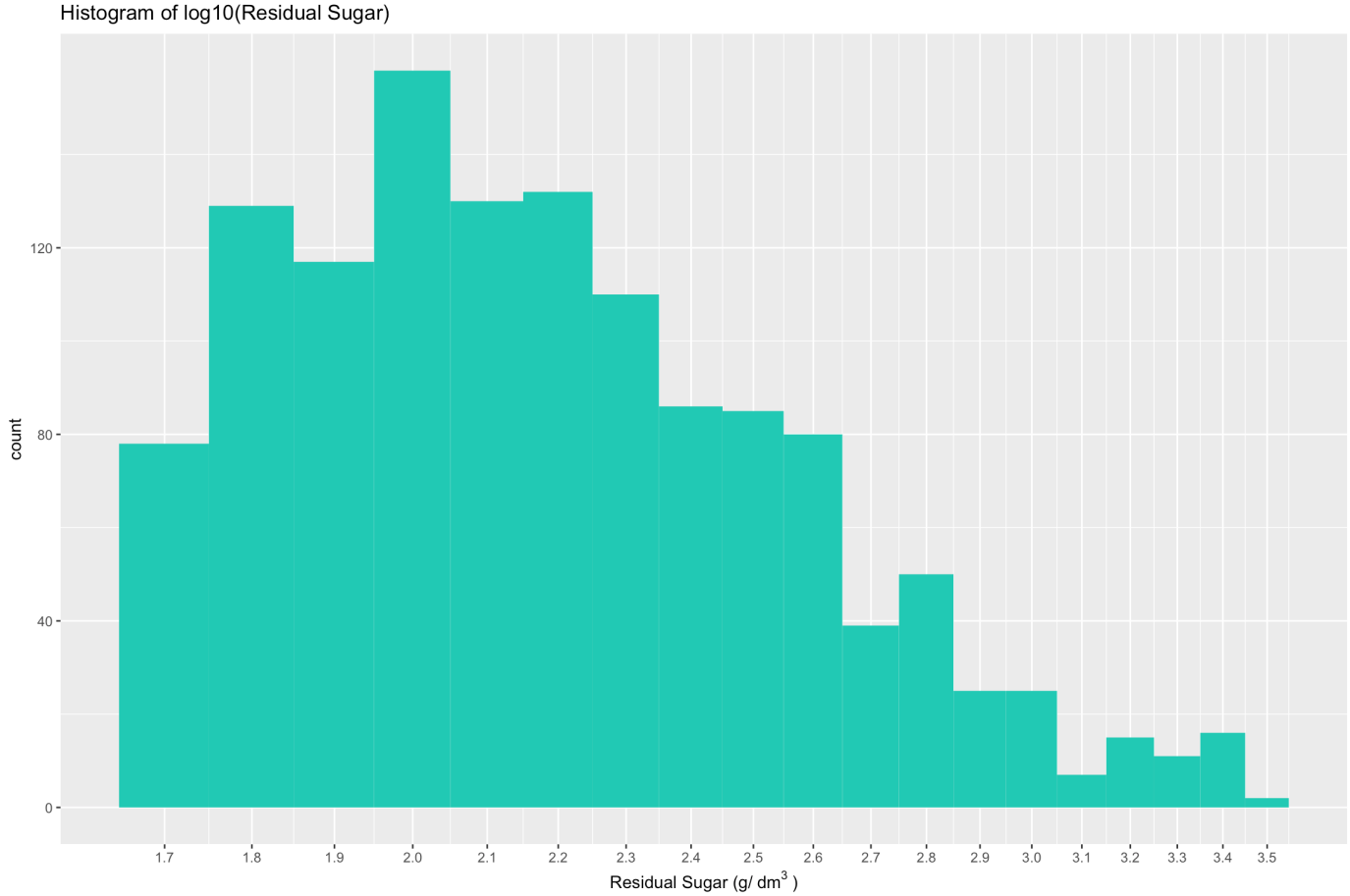

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

“residual.sugar” is right-skewed distributed with a lot of outliers located at right side and a little bit at left side. The most frequent values are between 1.9-2.4. IQR is 1.7. However, the maximum is locate too far away from 3rd quantile.

To improve the plot, I just removed the top and bottom 10% outliers and use log transformation, because this time we have too many outliers with extreme value:

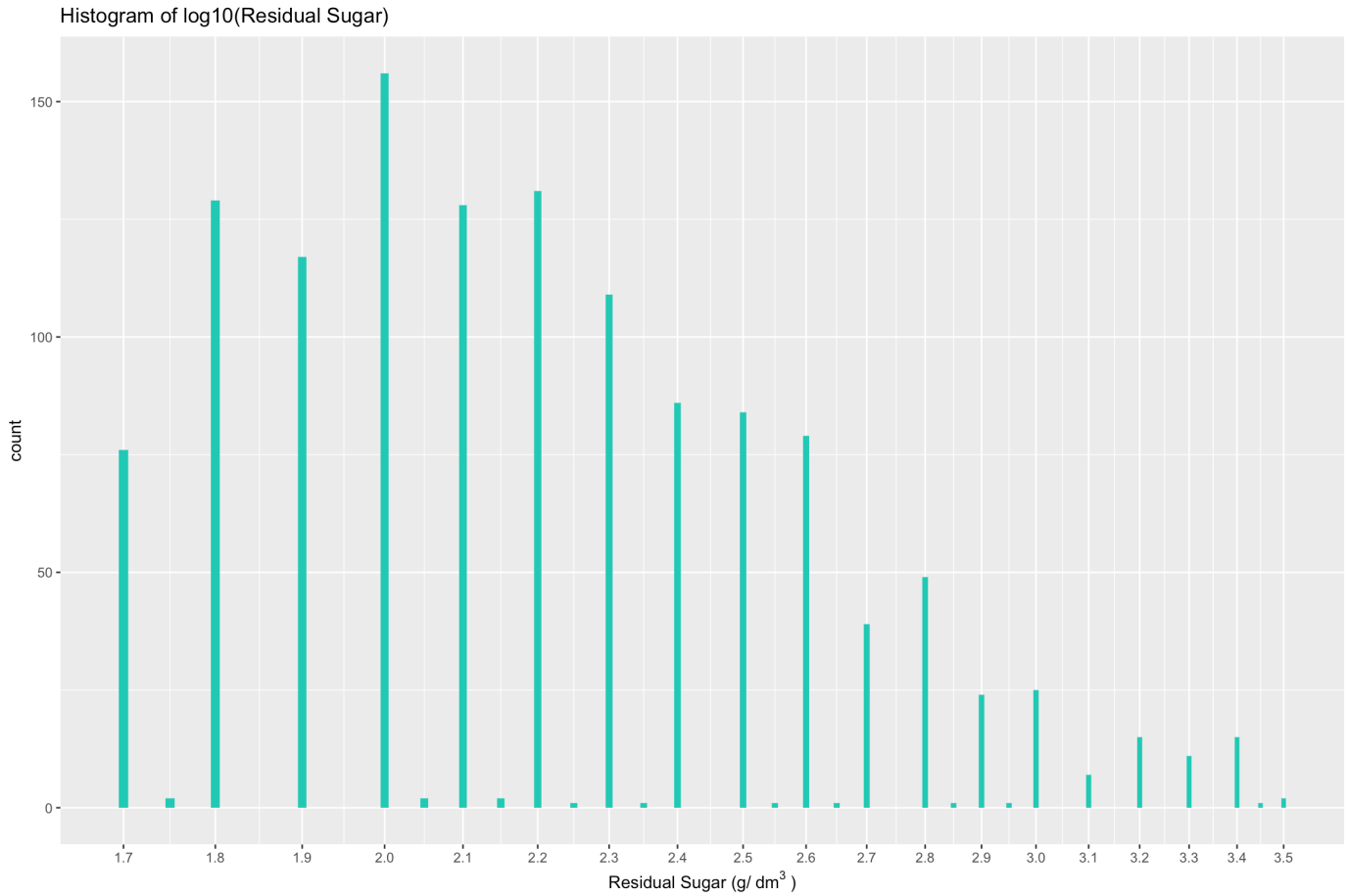
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.700   1.950   2.200   2.245   2.500   3.500
```

Now the IQR is 0.55, data gathers more in center.



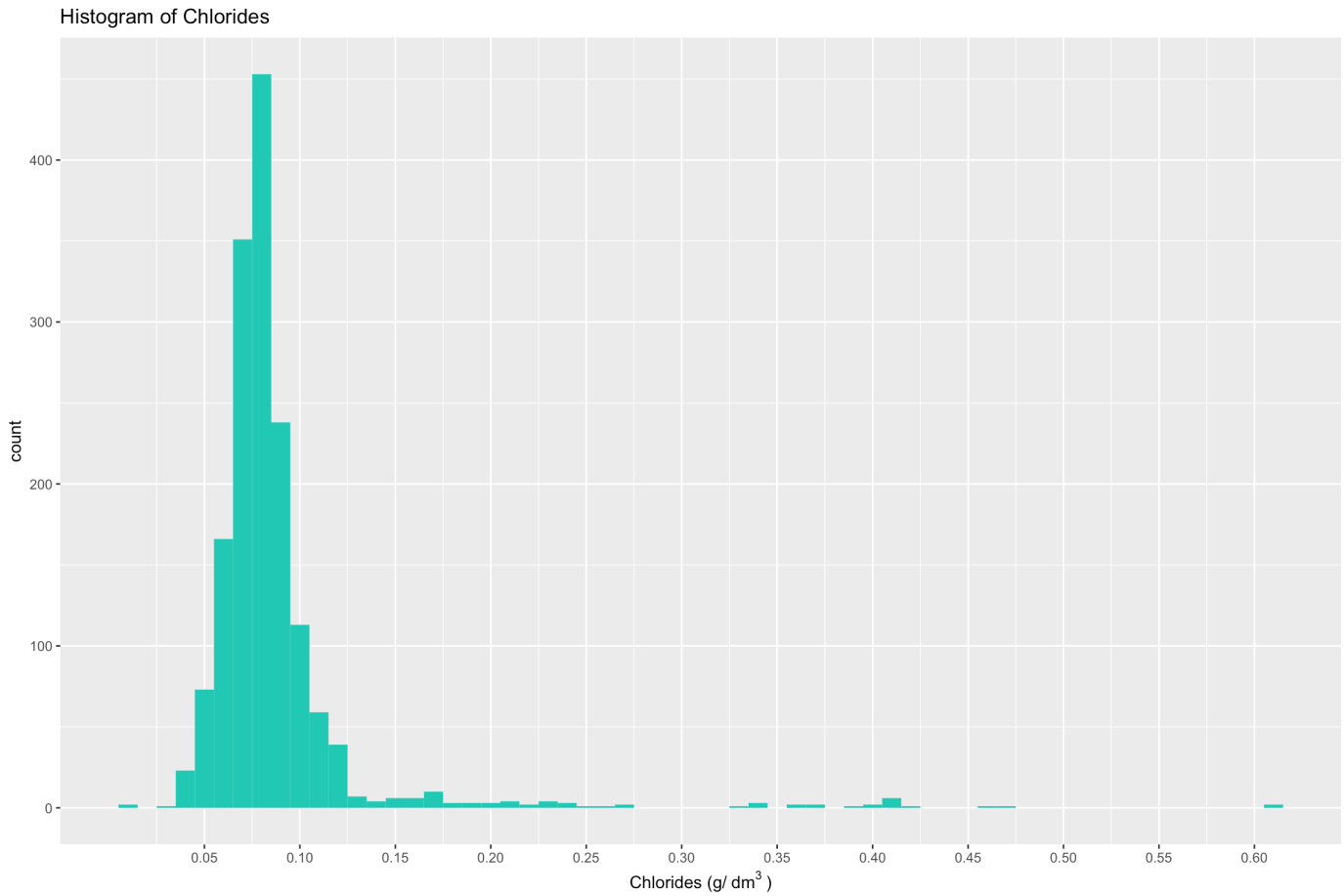

It looks much better, more symmetric without extreme outliers. From this plot, we can see more details, for example, the most common value is 2.0!

However, it might be possible 2.0 include all values from 1.95 to 2.05. So here let's use smaller binwidth:




And it's clearn the residual sugar was measured by discrete values with 0.05 as maximum accuracy.

**-chlorides-**



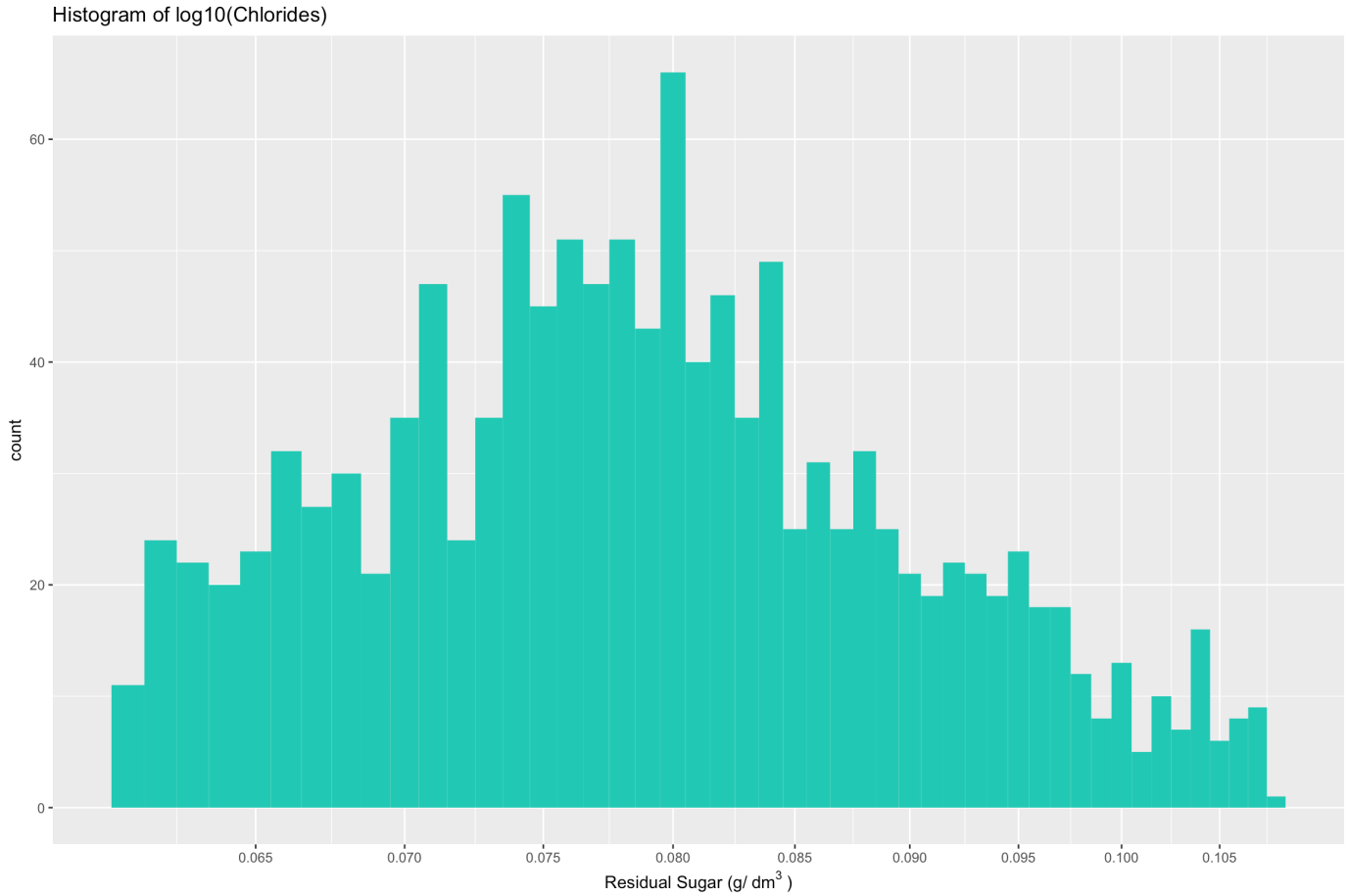

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

“chlorides” is right-skewed distributed with a lot of outliers located at right side. The most frequent values are between 0.062-0.112. IQR is 0.02.

To improve the plot, I just removed the top and bottom 10% outliers and use log transformation, because we have too many outliers with extreme value:

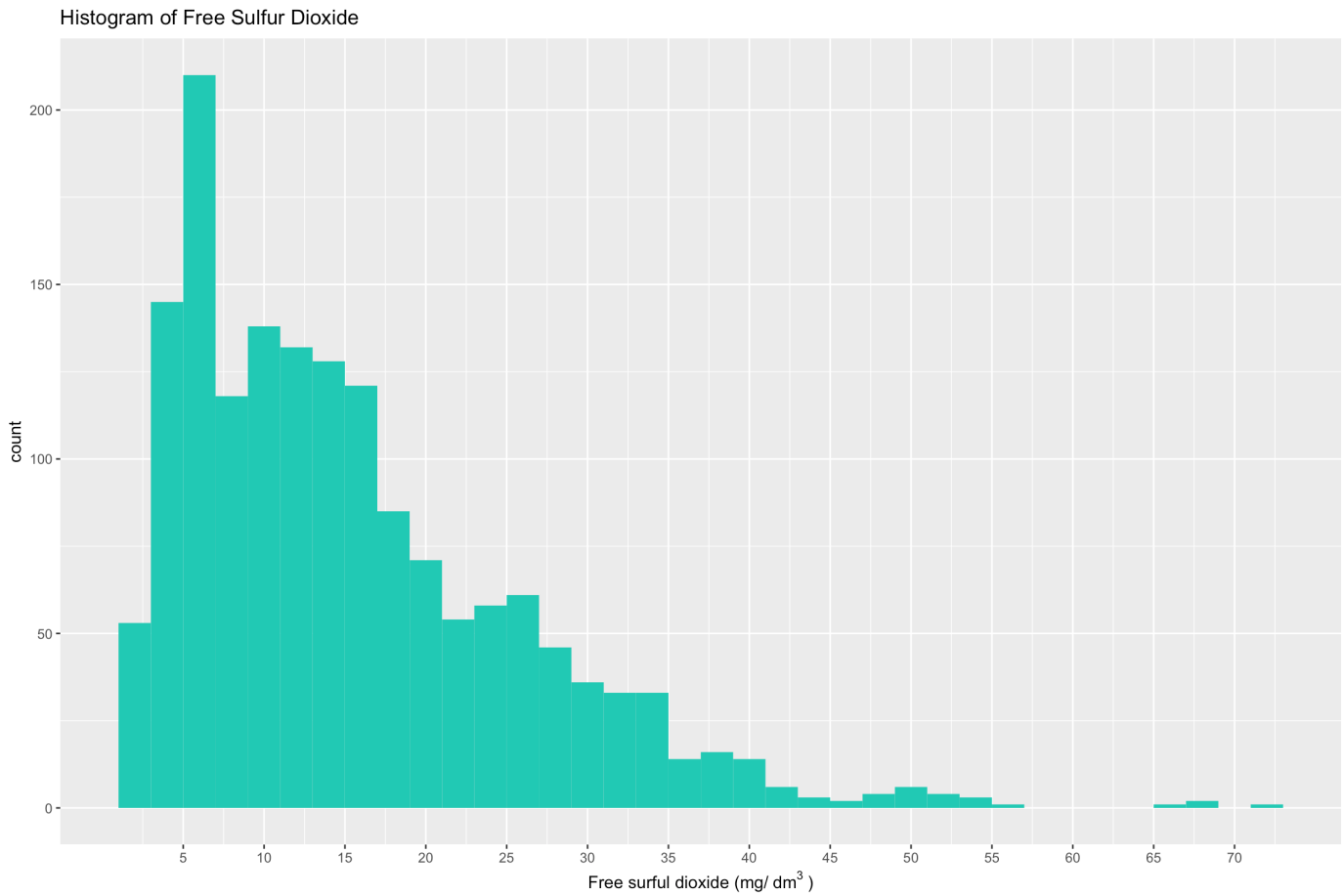
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06100 0.07300 0.07900 0.08029 0.08700 0.10800
```

Now the IQR is 0.026, 3rd quantile is not far from max value, and data gathers more in center.




It looks much better, more symmetric and bell shaped without extreme outliers. We can clearly see the most common value is 0.08.

### -free.sulfur.dioxide-



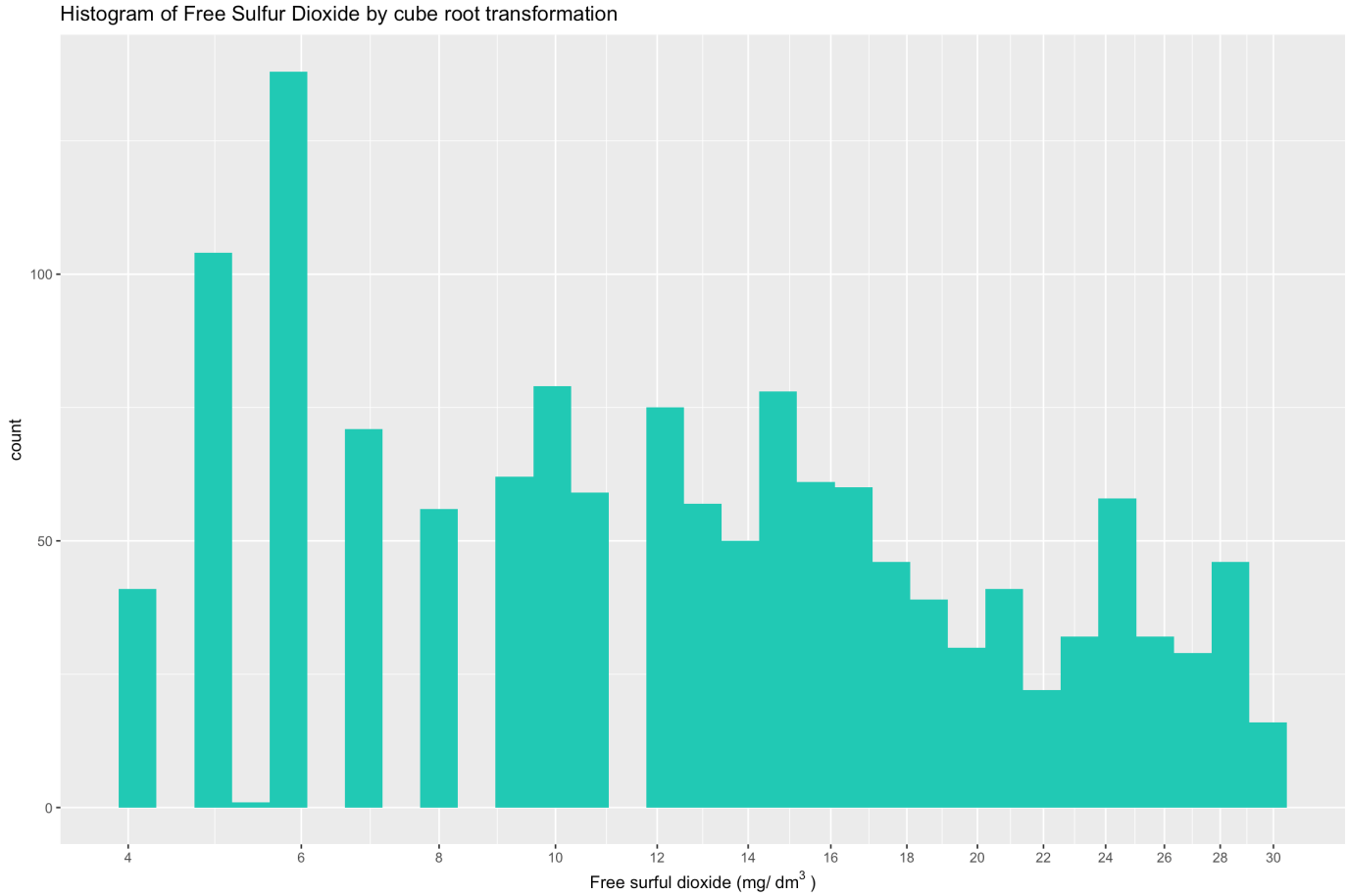
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

“free.sulfur.dioxide” is right-skewed distributed with a lot of outliers located at right side. The most frequent values are between 5-8. IQR is 14. Notice the number of free sulfur is larger than other ingredients like acidity, it’s because of different unit is applied. Sulfur is using $g/dm^3$, while acidity variables are using $mg/dm^3$. Actually, wine contains much less sulfur (free or total) than other ingredients.

To improve the histogram, I just removed the top and bottom 10% outliers and use log transformation:

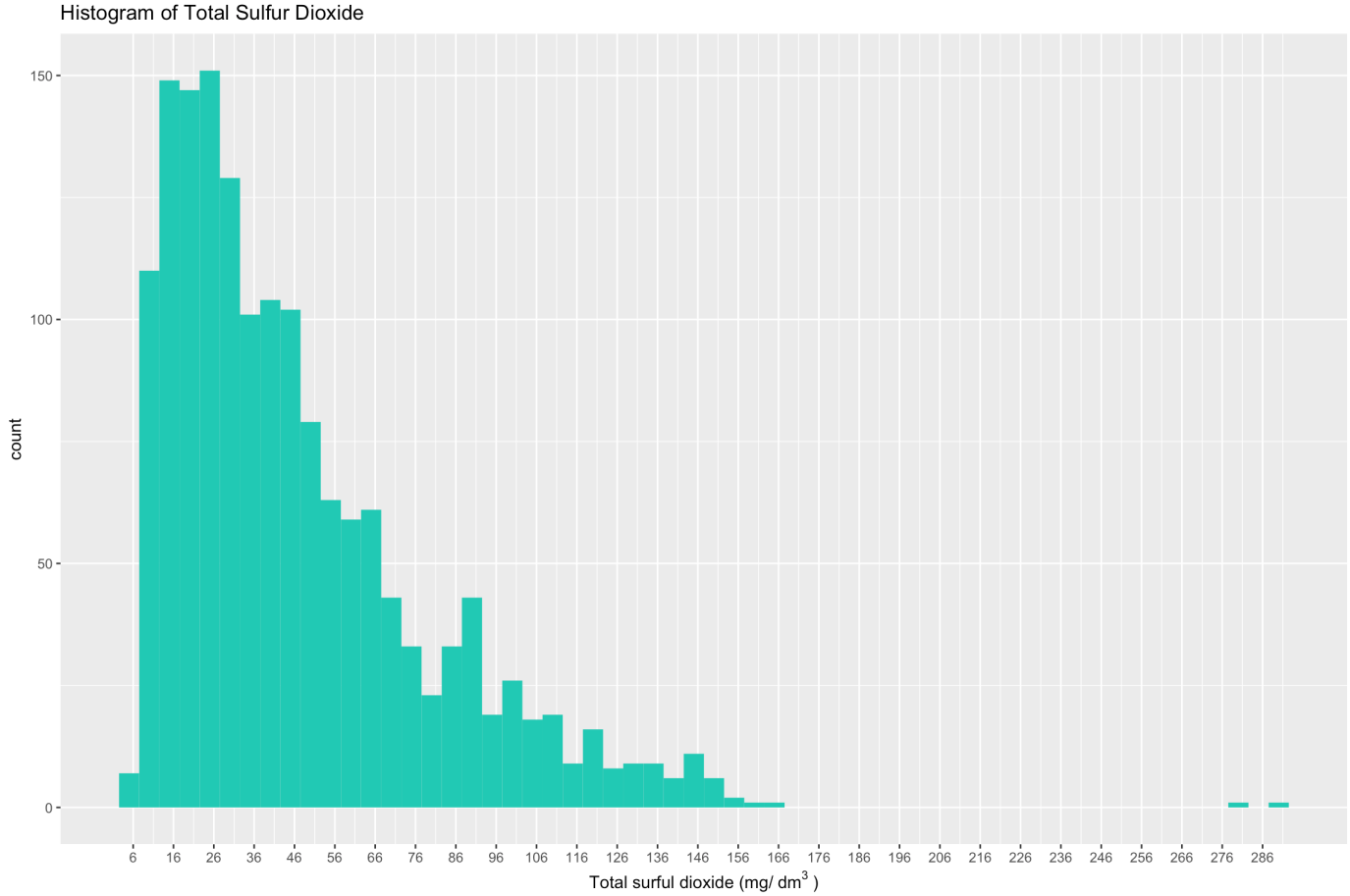
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    7.00   13.00   13.77   18.00   30.00
```

Now the IQR is 11, 3rd quantile is not far from max value, and data gathers more in center. However, because we do a big skewness, I will apply cube root transformation this time.




It looks much better without many extreme outliers. And zoom in by this plot we can see the most common value is 6.

**-total.sulfur.dioxide-**




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

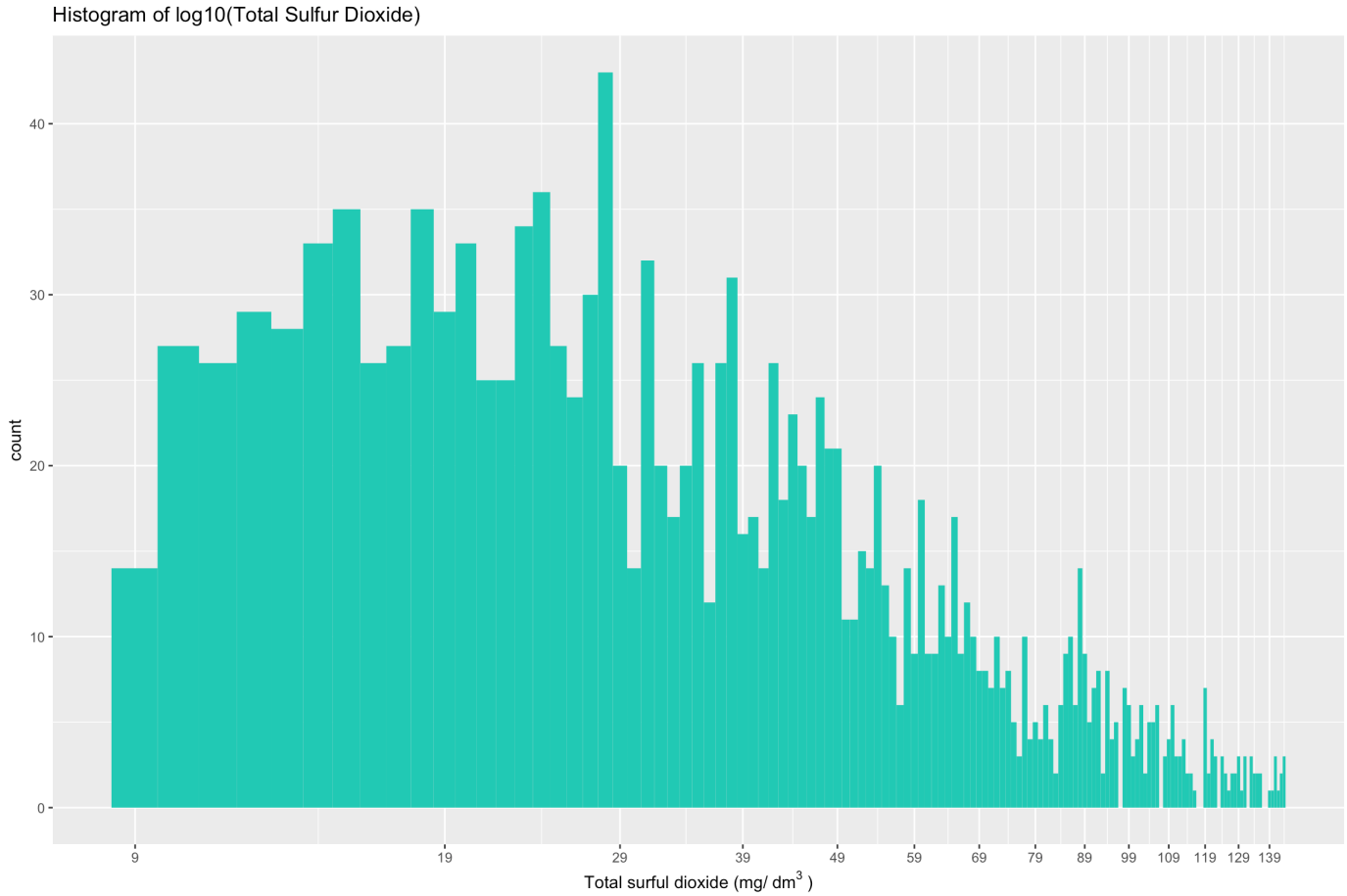
“total.sulfur.dioxide” is right-skewed distributed with some outliers located at right side. The most frequent values are between 15-25. IQR is 40, and the maximum is pretty far away from 3rd quantile, almost 6 times!

Notice wine contains a log of total sulfur dioxide being compared with other ingredients, even more than free sulfur. It’s reasonable total sulfur should be more than free sulfur because conceptually, free sulfur is part of total surful.

To improve the plot, I just removed the top 1% outliers and use log transformation:

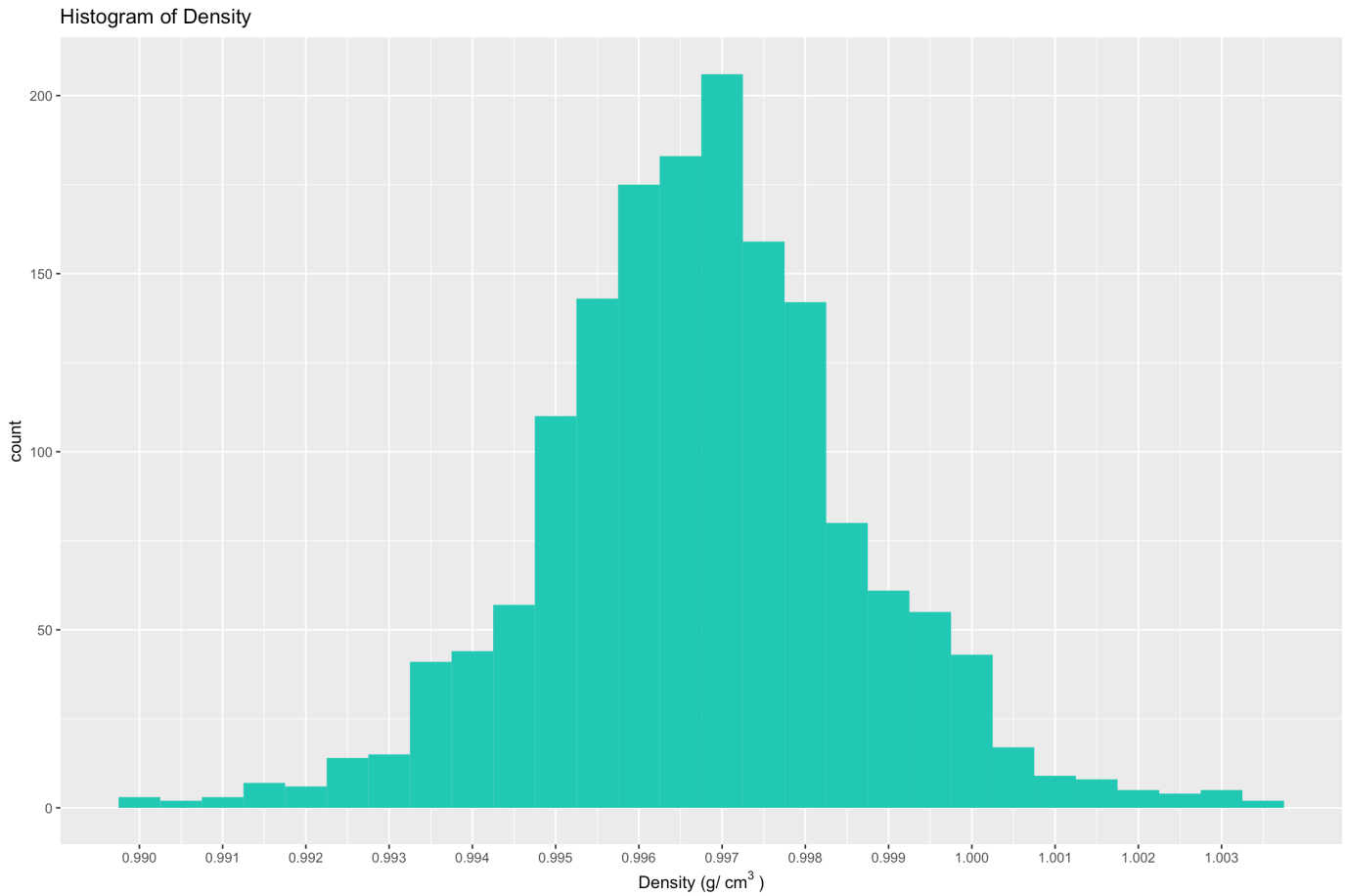
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   22.00   38.00   45.62   61.00  144.00
```

Now the IQR is 39, 3rd quantile is not very far from max value, and data gathers more in center.




It looks much better without extreme outliers. We can easily figure out 28 is the most common value.

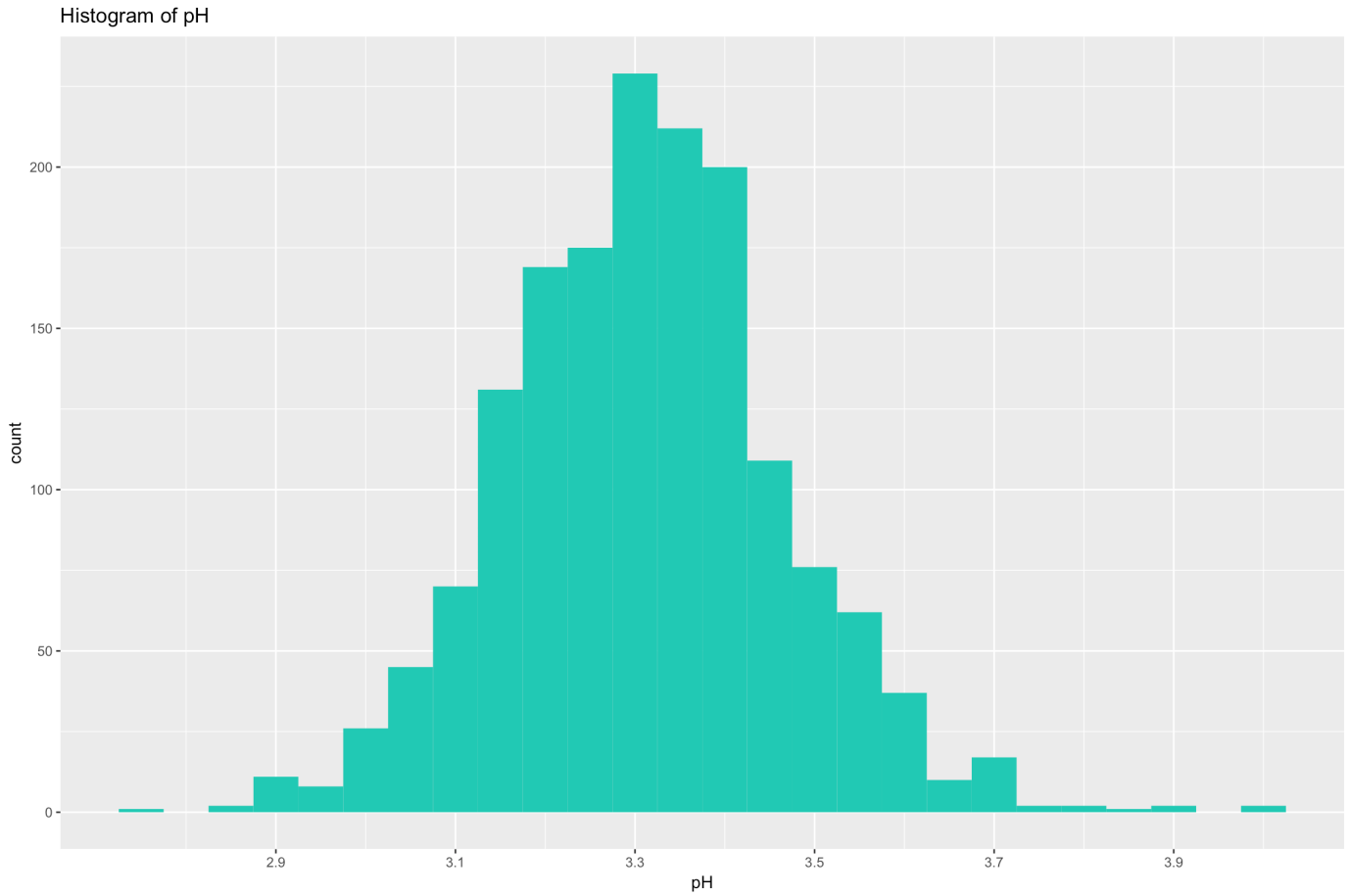
**-density-**



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

“density” is approximately symmetric, and it’s surprising the difference among different wines, though they might test significantly different, are not that big. So it’s not necessary to do transformation for density.

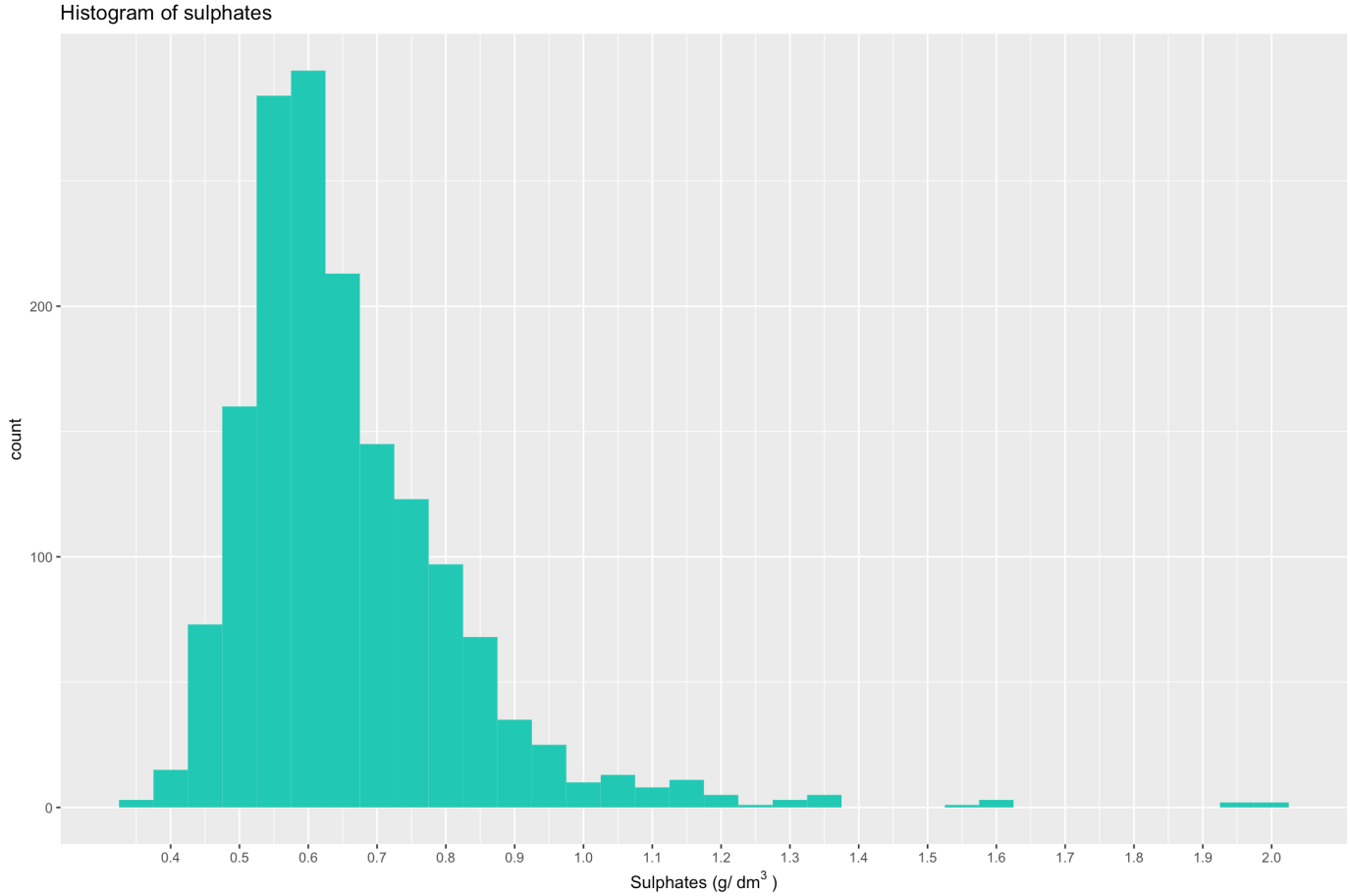
**-pH-**




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

“pH” is almost symmetric. The most frequent value is between 3.24 and 3.44, IQR is 0.19. It’s not a big difference. So it’s also not necessary to do transformation for pH.

**-sulphates-**



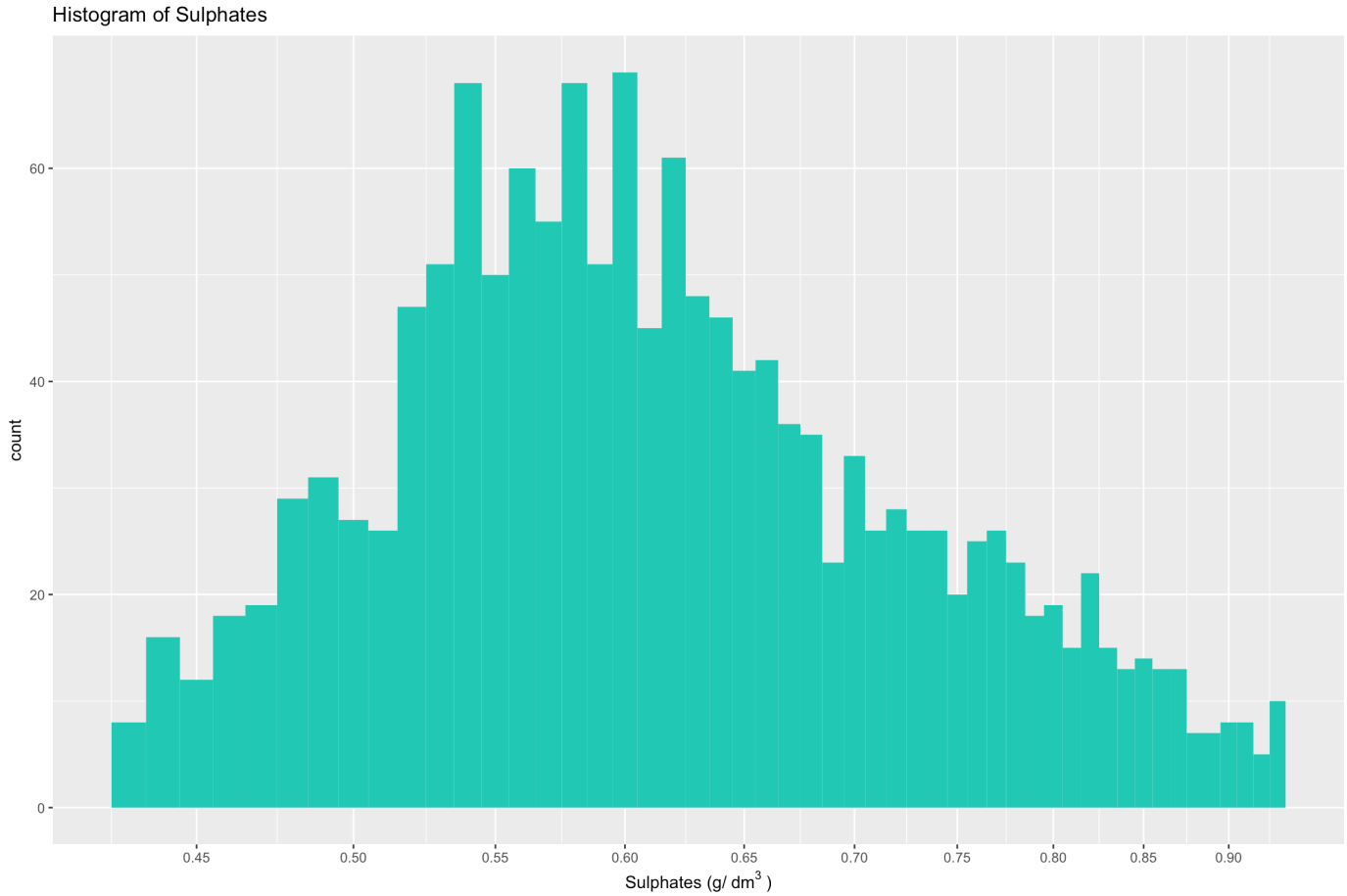

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

“sulphates” is a little bit right-skewed distributed with some outliers located at right side. The most frequent values are between 0.5-0.7. IQR is 0.18.

To improve the plot, I just removed the top 5% outliers:

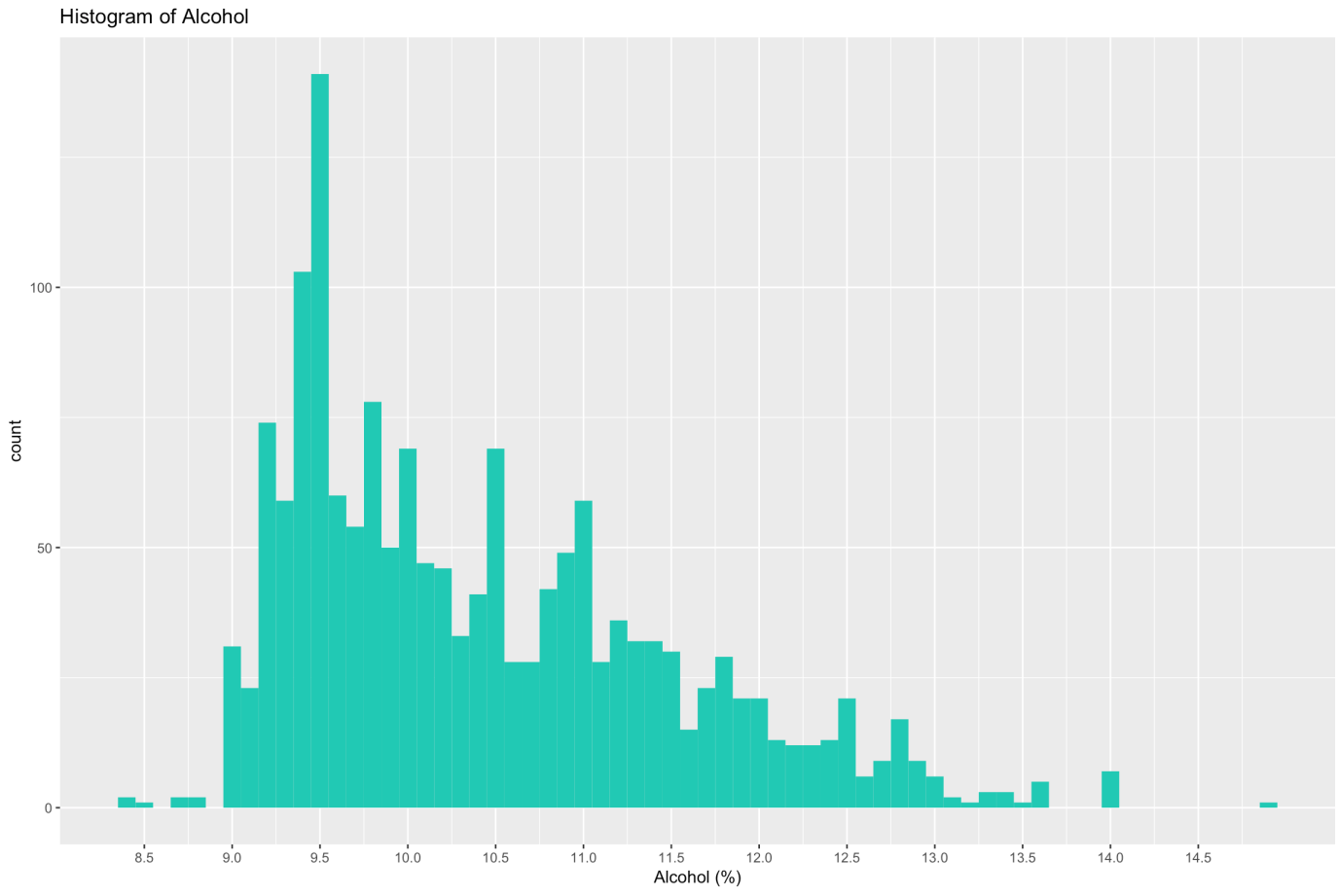
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4300  0.5500  0.6200  0.6349  0.7100  0.9300
```

Now the IQR is still 0.18, 3rd quantile is not very far from max value, and data gathers more in center. Then I apply log transformation to zoom in and check more details.




Now most of the data meets at center and we have 0.6 as the most common value.

**-alcohol-**



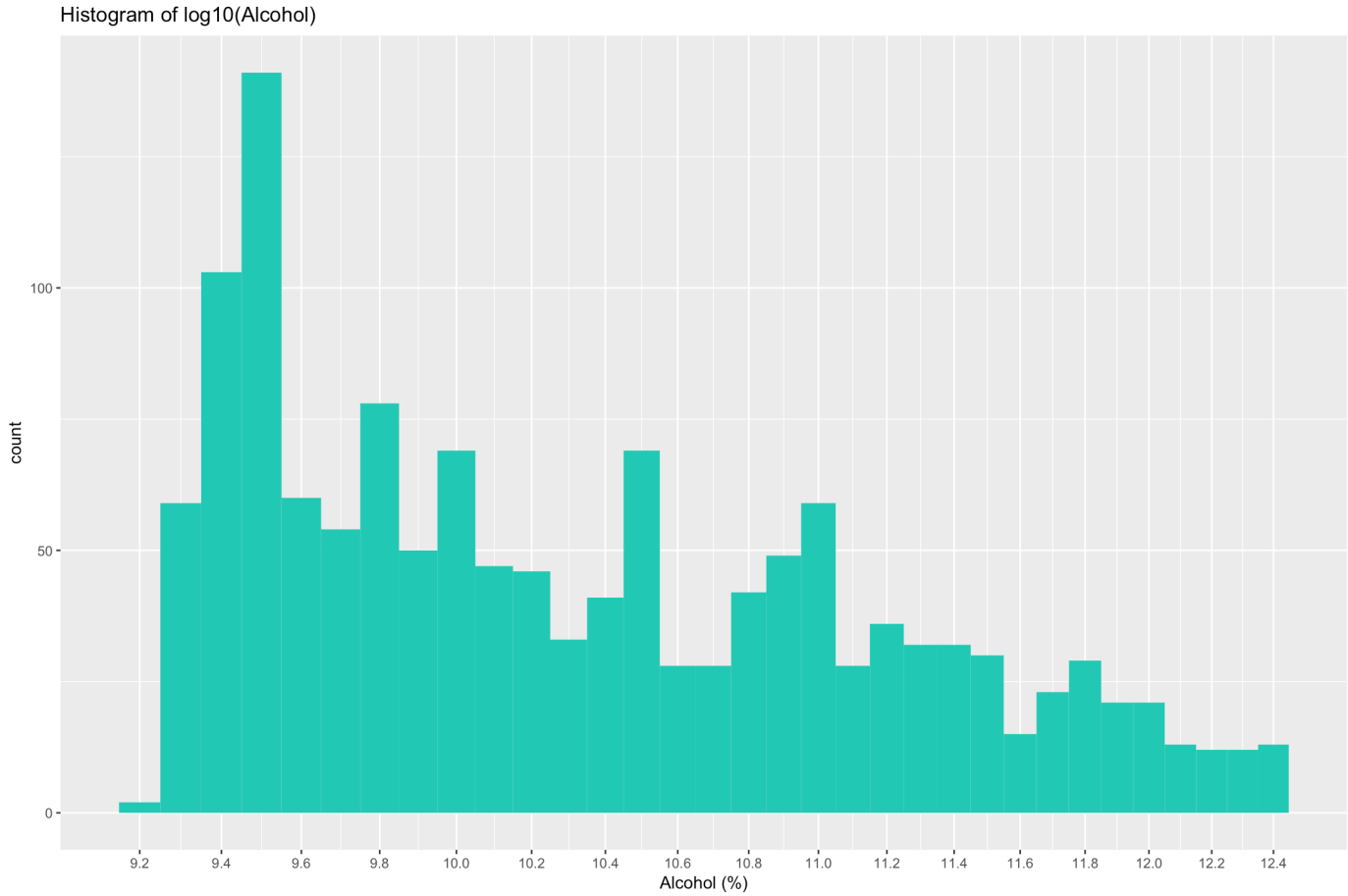
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

“alcohol” is right-skewed distributed with some outliers located at right side. The most frequent values are between 9.4-9.6. IQR is 1.6.

To improve the plot, I just removed the top and bottom 5% outliers and use log transformation:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   9.233   9.600  10.200  10.380  11.000  12.400
```

Now the IQR is 1.4, 3rd quantile is not very far from max value, and data gathers more in center.




Now we can clearly see the most common value is 9.5. The majority data gather at left side more than right side.

# Question answer part

## What is the structure of your dataset?

The red wine quality dataset include 1599 observations and 12 variables. All attributes are numeric, 11 of them are continuous test result and 1, the quality, is rating of integers ranging from 3 to 8. It’s pretty tidy and none of attributes have NA values.

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

After take a look at the attributes of the dataset, I found the variables like pH, alcohol, sulphates etc. most interesting to me, because I leared some red wine quality determinant before and do hope to explore how these variables distributed and how they related to wine quality.

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

By the above correlation matrix, volatile.acidity, sulphates and alcohol are the attributes most coorelated with quality of wine. Thus, these 3 attributes are most attractive to me.

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

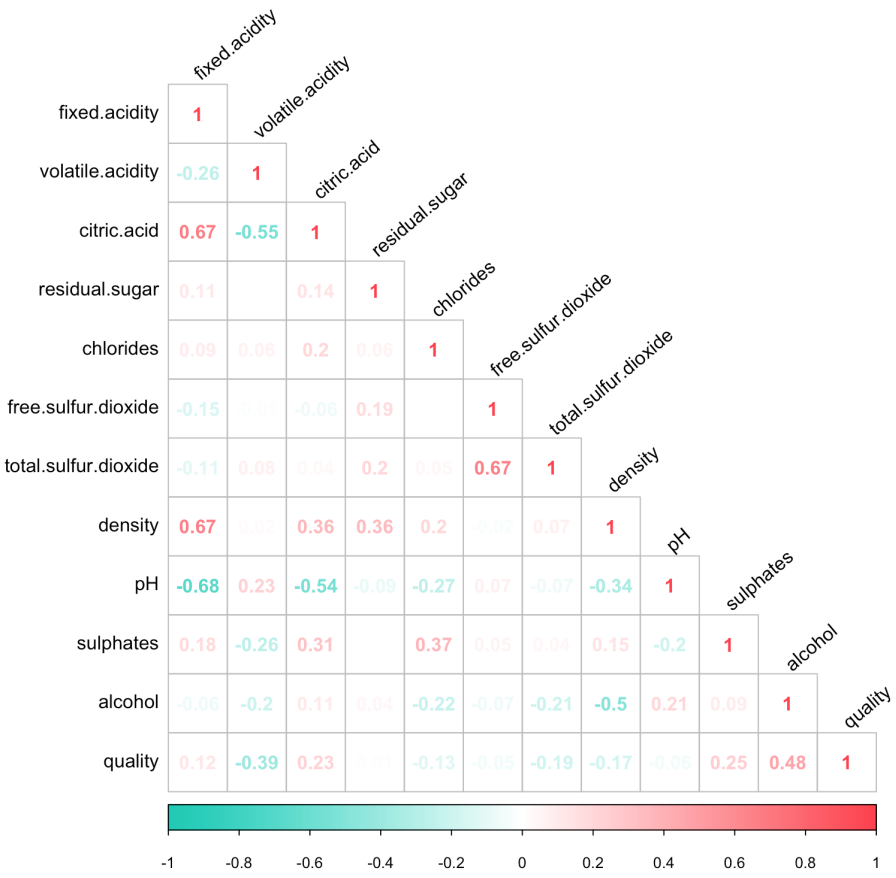
I haven't create any new features so far. But I will create in below analysis.

## 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?

All attributes have outliers with extreme value. I removed the outliers and use log, cube root to reduce skewness. There are several reasons I do transformation: - remove outliers to keep attention on trend of majority data points - transformation can help to zoom in and easily pick out most common value of the distribution - use transformation to make distribution more symmetric so we can apply further model to explore the data, as discussed below

# Bivariate Plots Section

Then I create correlation matrix to figure out which attributes are worth further exploring.

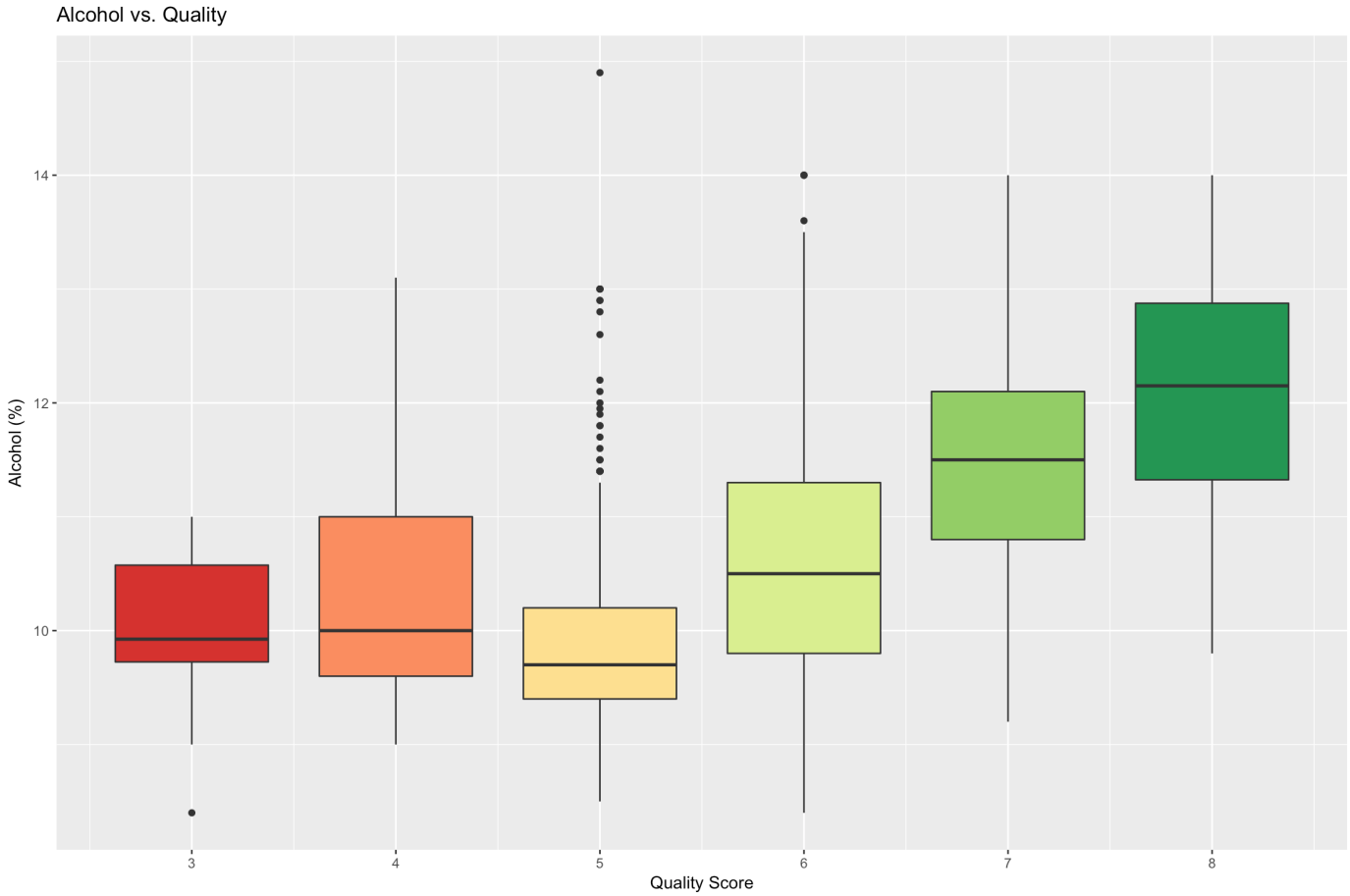


From this plot, we can see some pair of attributes associate with each other. For example, citric.acidity associates positively with fixed.acidity, while pH negatively associates with fixed.acidity. However, there seems no attributes have strong correlation with quality.

Here, alcohol, volatile.acidity, sulphates are top3 attributes that associated with quality. So let's explore further on quality and these 3 variables.

Now let's explore how these 3 attributes interact with quality.

**-quality vs. alcohol-**

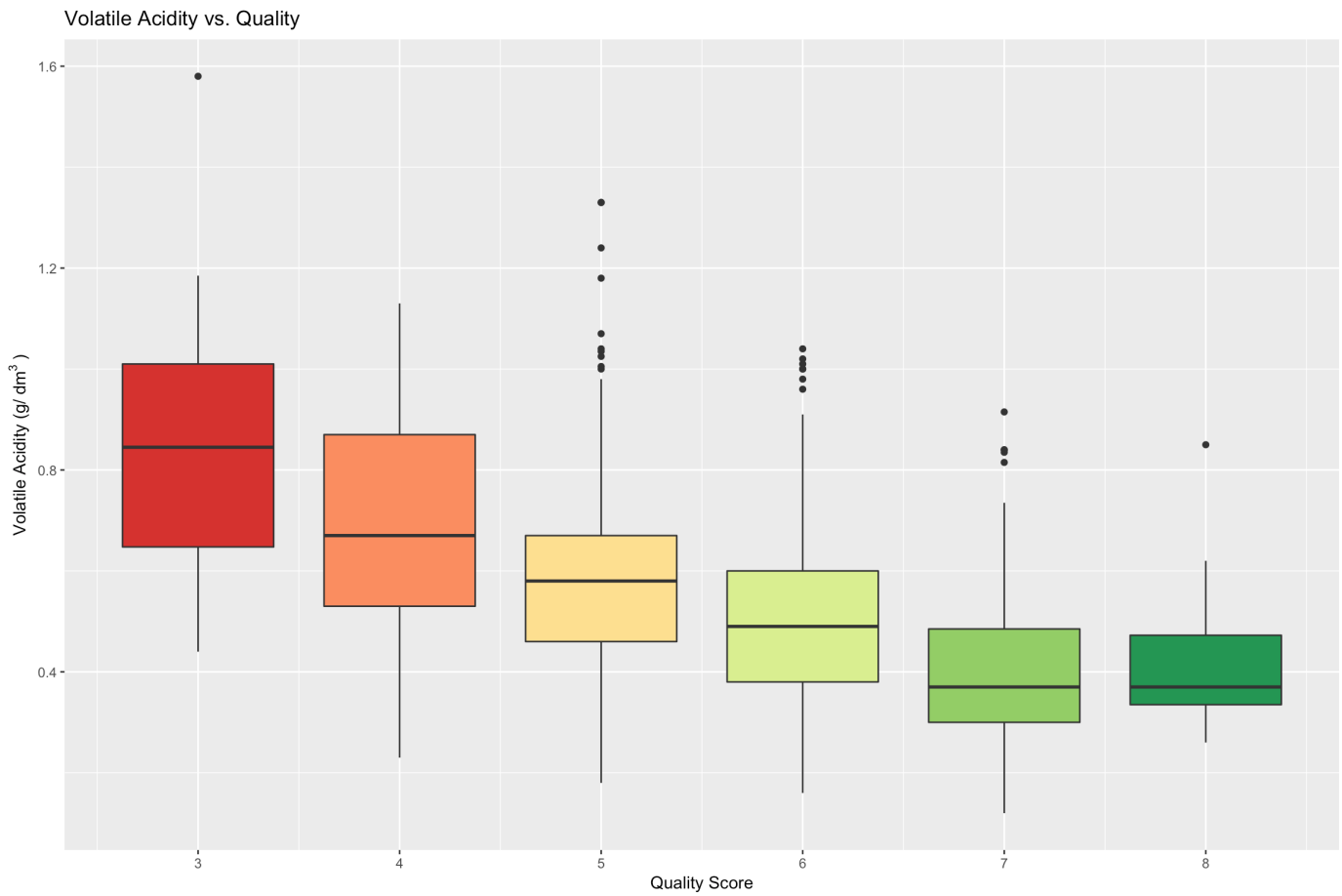



It seems there's positive relationship between alcohol and quality. High quality wines are more likely to have high percentage of alcohol. The corelation coeffecient $R^2 = 0.2263$, which means alcohol can explain only 22.63% the variation of quality.

```
##
## Call:
## lm(formula = alcohol ~ quality, data = wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2517 -0.6233 -0.2233  0.5483  4.8767
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.88160    0.16532   41.62   <2e-16 ***
## quality      0.62835    0.02904   21.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9374 on 1597 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2263
## F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

**-quality vs. volatile.acidity-**

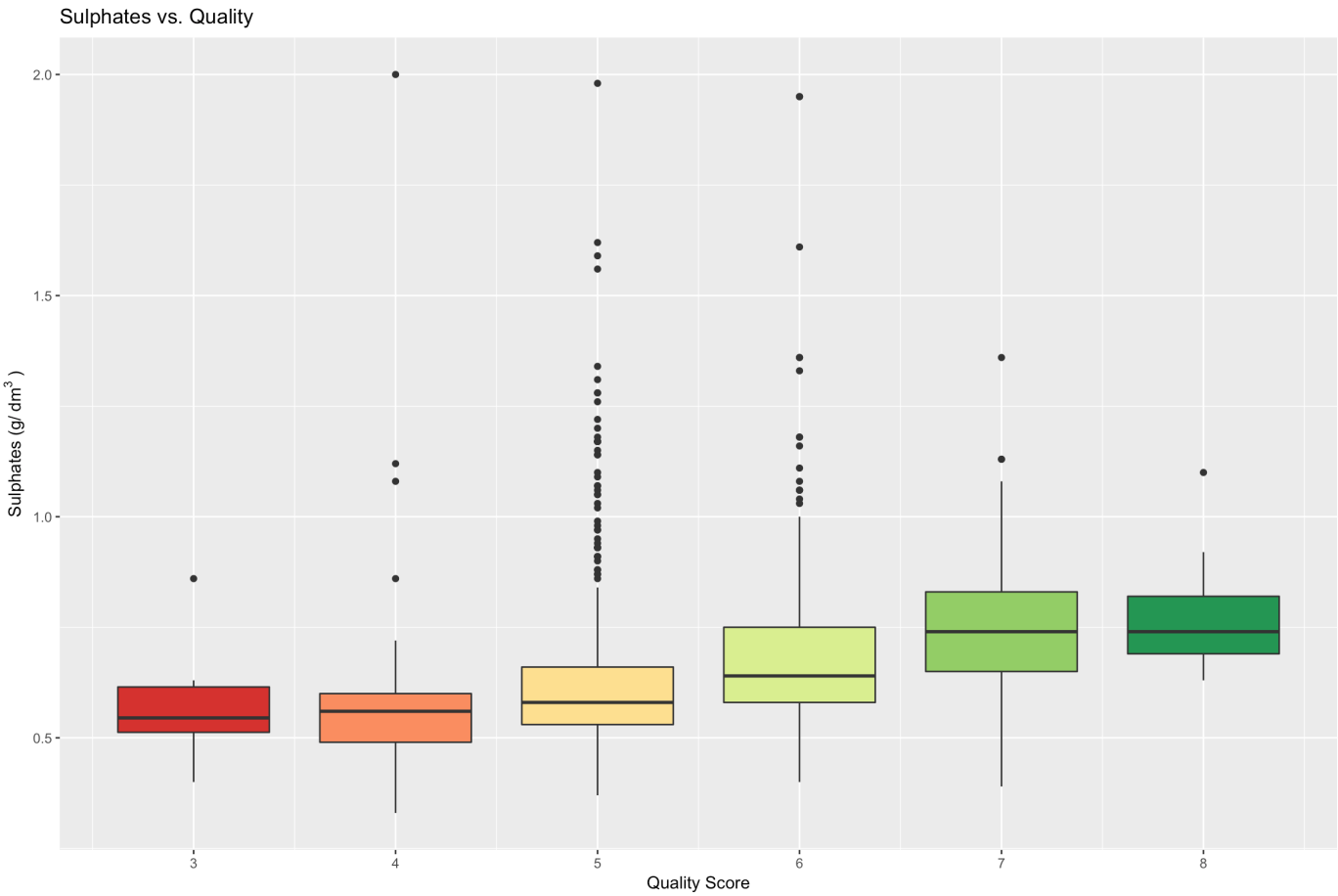
To avoid overplotting, I add transparency in this plot.




Now we can see a negative association between these two attributes. While alcohol is increasing with quality, volatile acidity is negatively associated with quality. The corelation coeffecient $R^2 = 0.152$, which means volatile.acidity can explain only 15.2% the variation of quality.

```
##
## Call:
## lm(formula = quality ~ volatile.acidity, data = wine_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.79071 -0.54411 -0.00687  0.47350  2.93148
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.56575    0.05791  113.39   <2e-16 ***
## volatile.acidity -1.76144    0.10389  -16.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7437 on 1597 degrees of freedom
## Multiple R-squared:  0.1525, Adjusted R-squared:  0.152
## F-statistic: 287.4 on 1 and 1597 DF,  p-value: < 2.2e-16
```

**-sulphates vs. quality-**

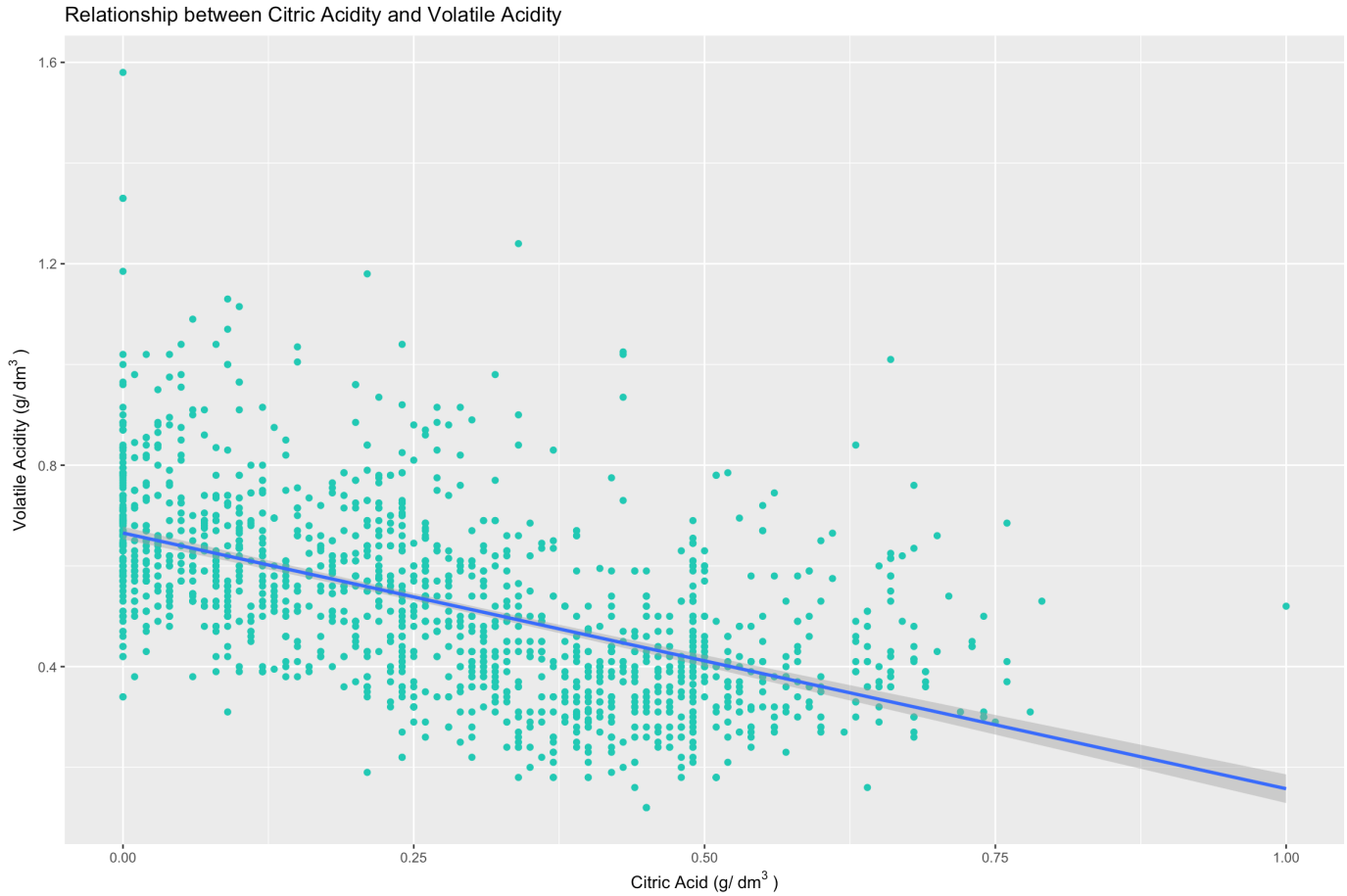



We can see a positive association between these two attributes. The corelation coeffecient $R^2 = 0.06261$, which means sulphates can explain only 6.26% the variation of quality.

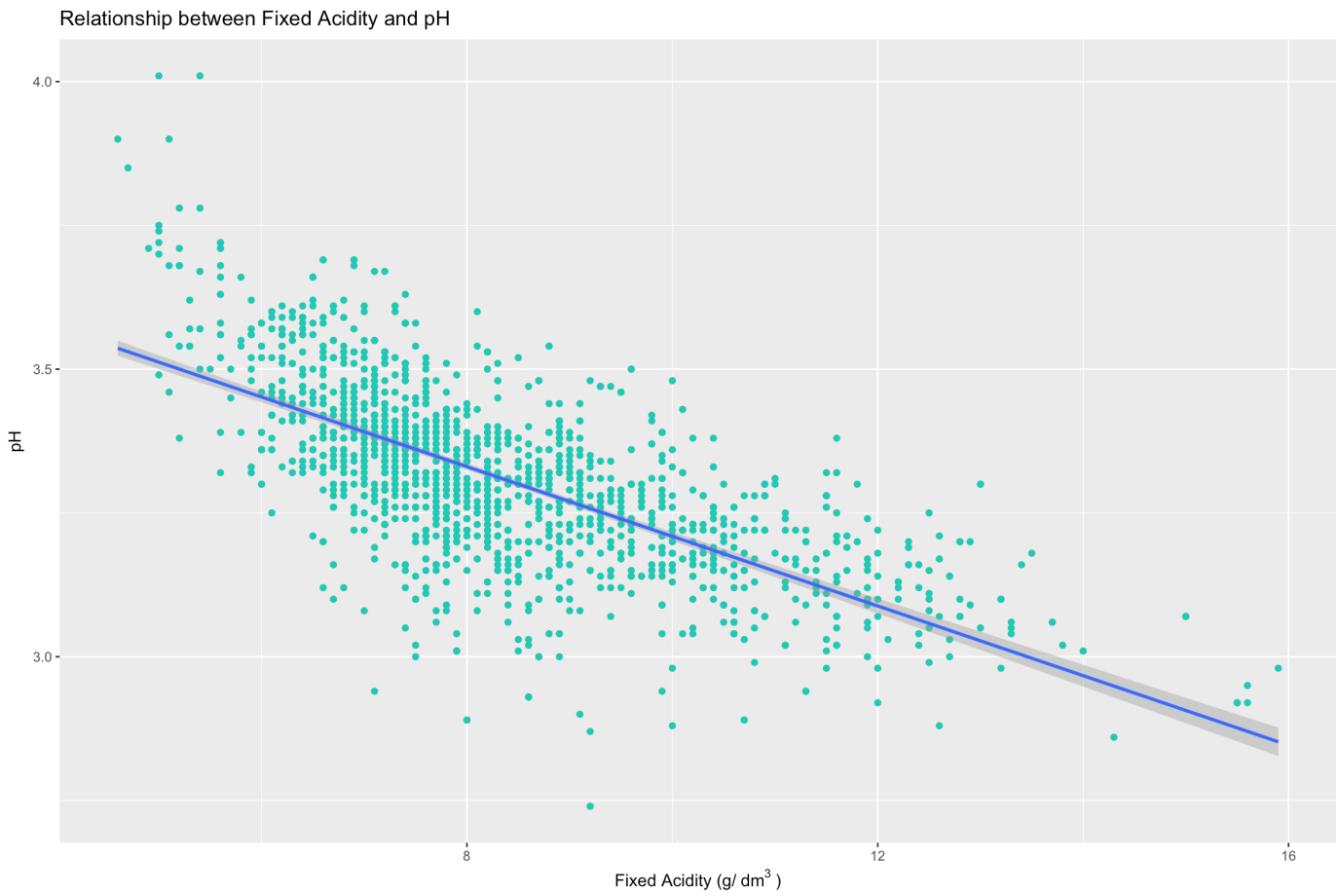
```
##
## Call:
## lm(formula = quality ~ sulphates, data = wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2432 -0.5424  0.1102  0.4456  2.3977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.84775    0.07842   61.82   <2e-16 ***
## sulphates    1.19771    0.11539   10.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7819 on 1597 degrees of freedom
## Multiple R-squared:  0.0632, Adjusted R-squared:  0.06261
## F-statistic: 107.7 on 1 and 1597 DF,  p-value: < 2.2e-16
```

From the matrix, we can also see there're corelated attributes:


This positive association can be due to the fact that citric acid provides solid support to fixed acidity.




Volatile acid, however, is negatively associated with citric acid because citric acid rarely volatilize. You can test it but can hardly smell.

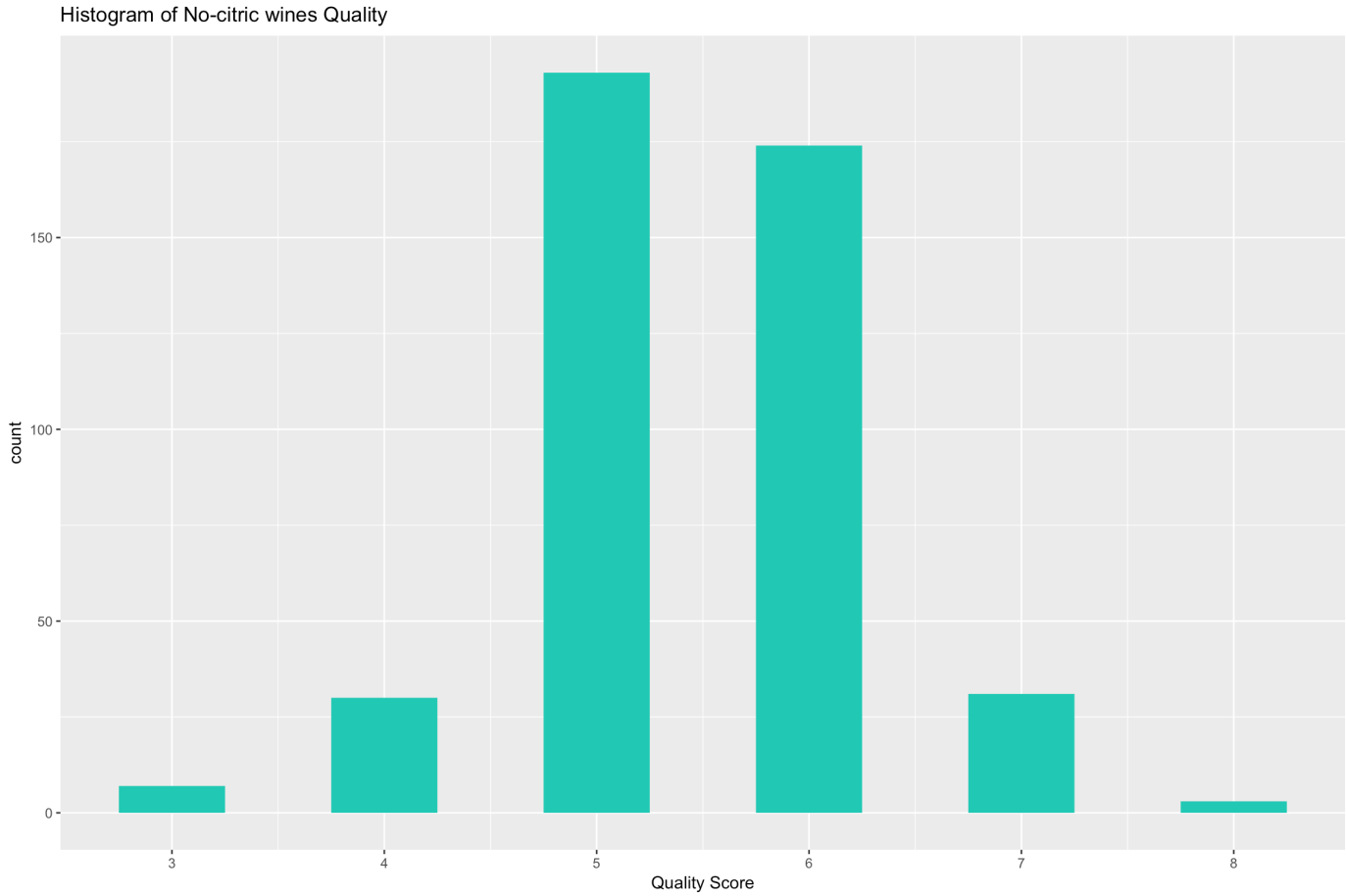



Acidity inside liquid, the wine, will certainly decrease pH. The smaller the value of pH is, the liquid becomes more acid.

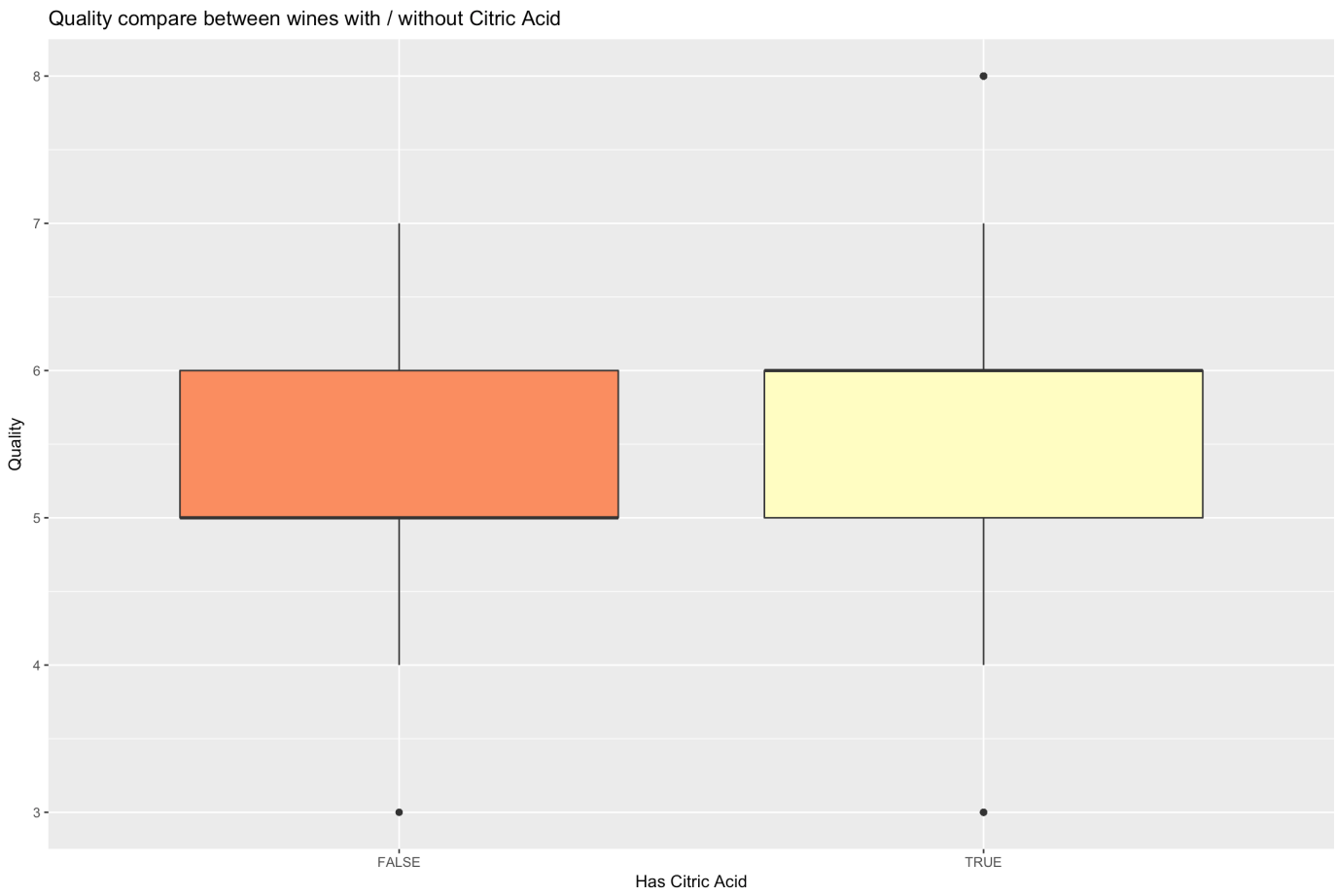

Acidity is provided by acid mocules with heavier weight than water and alcohol because these acid moleculars are resolved as ions wondering among the space of water moleculars.

Next, there're some questions I'm interested in.

1. How's quality of wine with no citric.acid Recall that we find there's over 150 observations with citric.acid value equal to 0, we have adequate data to explore:

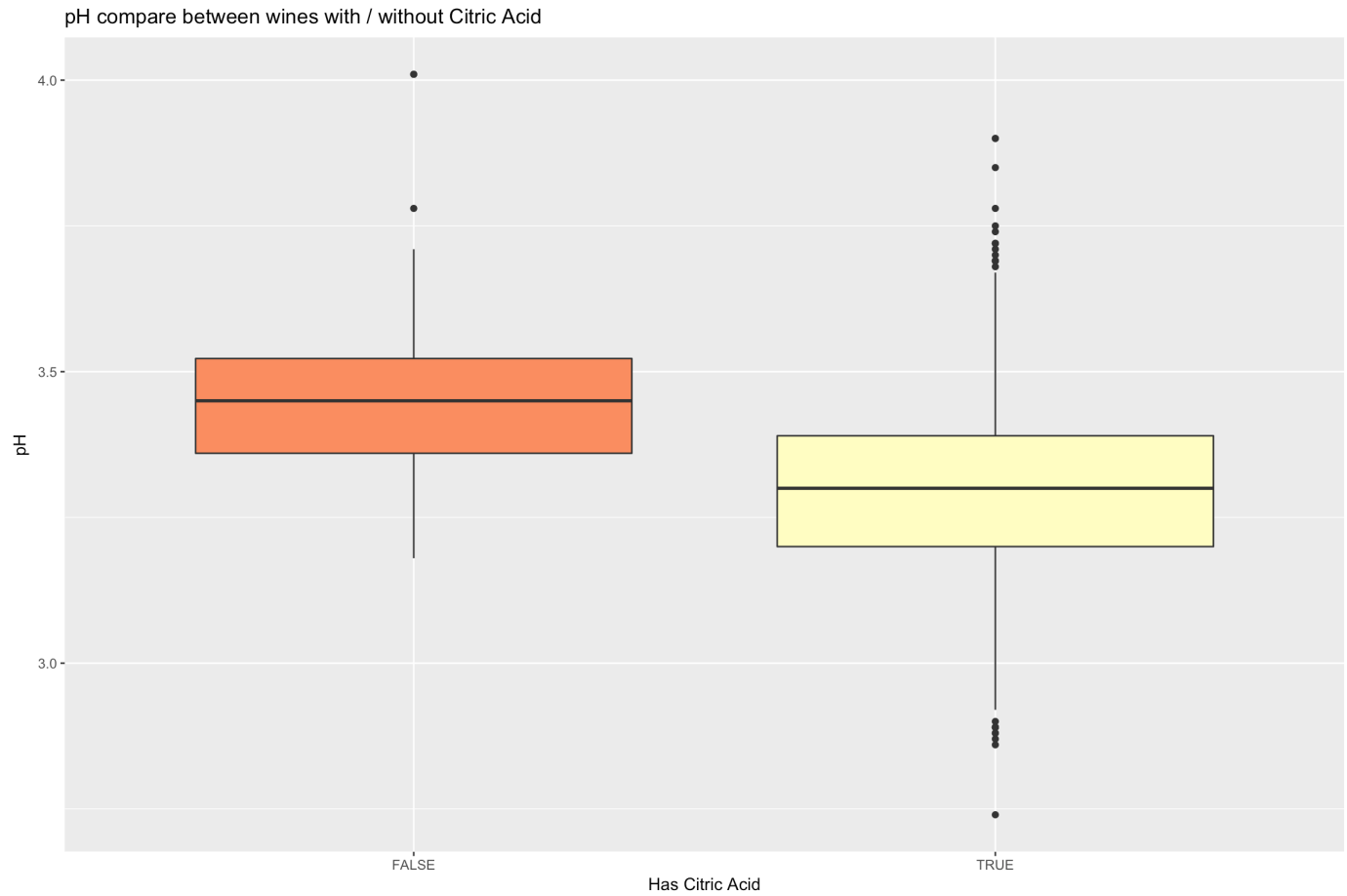



The distribution is quite similar to overall wine data. Now let's make boxplots to compare the quality of two groups: with/without citric acid.




Though the median is different, the plots looks quite similar. Whether having citric acid seems do not affect quality significantly.

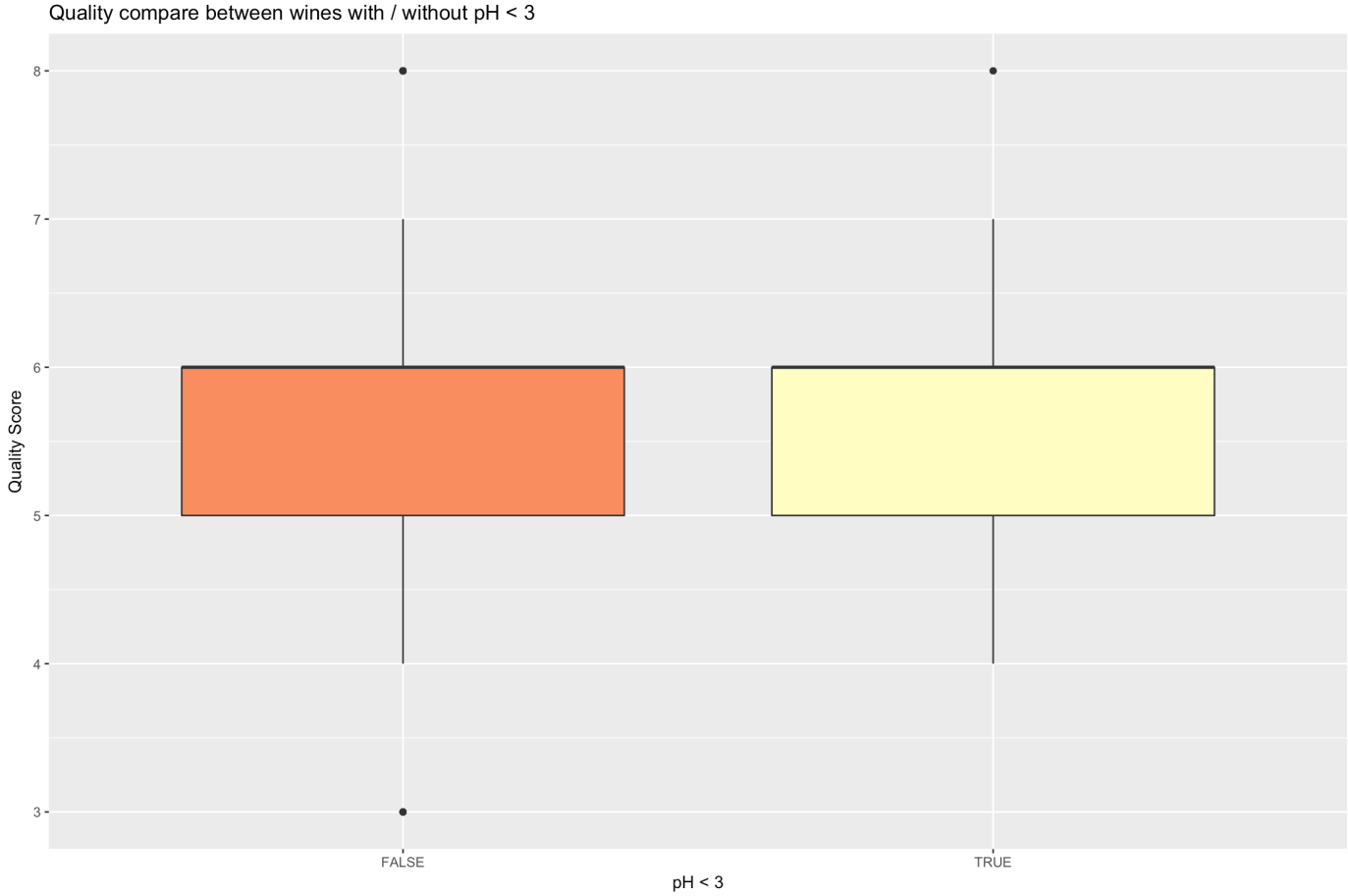
2. How pH without or without citric acid? And how's quality of wines with the most extremely acid pH?




We can see the wines with citric acid have lower pH. But we don't know if extreme pH will affect quality. If yes, citric acid can be a indirect factor to improve quality of wine, because from the above correlation matrix, we can see, beside fixed acid, citric is the biggest factor associates with pH.

```
##
## FALSE  TRUE
##  1570    29
```

There're 29 wines with pH less than 3 (A strong acid level!).



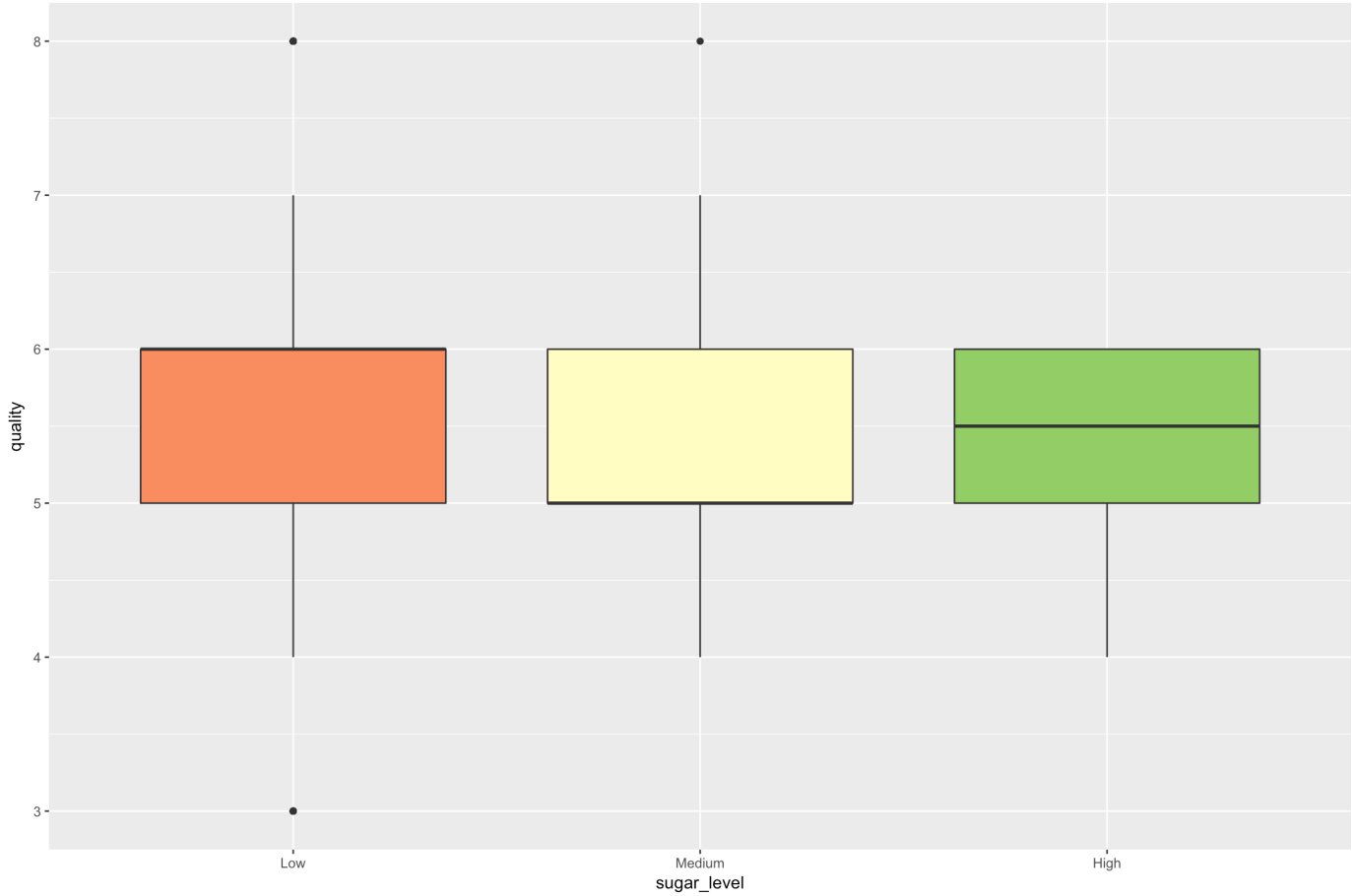

This is really frustrating. :( However, we learn extreme pH doesn't affect quality that much.

3. Are wines with higher residual sugar indicating lower quality?

"Sweetness is happiness!"

According to our experience, food with more sugar will be more attractive. Like cake, cola, and even some of the sweet wines. Does this happen in red wine?

Let's compare quality by cut data into different level of residual sugar, and then plot quality boxcharts in gourp of sugar level:




The plot tells us the relationship between sugarlevel and quality is not very strong. From the correlation, we can draw same conclusion.

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

Let's refresh these interesting discoveries in this section: - Increasing fixed.acidity will lead to decreasing pH because more hydrogen ion appears - Citric.acidity is a kind of acid without volatile, so it's reasonable when citric.acidity increase, fixed acidity will increase. As total acidity is divided into two groups, namely the volatile acids and the nonvolatile or fixed acids. So it's reasonable when Citric.acidity increases, volatile will decrease - when the total amount of acid increases, density will increase because water molecule is much lighter than these acid ion - Whether having citric acid is not a significant indicator of quality - Extreme pH is not an indicator of wine quality - Residual sugar is not a significant indicator, too

# Bivariate Analysis

Before further analysis, it's necessary to remove outliers because these outliers might bias our model. Because data are distributed right-skewed in the dimentions of these 3 attrbutes, I will remove top 1% data.

```
##  volatile.acidity   sulphates         alcohol
##  Min.   :0.1200   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.3900   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.5200   Median :0.6200   Median :10.10
##  Mean   :0.5218   Mean   :0.6493   Mean   :10.39
##  3rd Qu.:0.6350   3rd Qu.:0.7200   3rd Qu.:11.03
##  Max.   :1.0100   Max.   :1.2600   Max.   :13.30
```

After this, I removed 52 observations. Now we have the data which has all maximum of 3 main attributes close to its 3rd quantile. The correlation between quality and one of the main attributors are:

```
##                          [,1]
## volatile.acidity 0.13475395
## sulphates        0.09393676
## alcohol          0.23549438
```

Now, I apply the model:

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = wine_data.improved)
## m2: lm(formula = quality ~ alcohol + volatile.acidity, data = wine_data.improved)
## m3: lm(formula = quality ~ alcohol + volatile.acidity + sulphates,
##     data = wine_data.improved)
##
## ================================================
##                         m1         m2         m3
## ------------------------------------------------
##   (Intercept)        1.721***   2.879***   2.227***
##                     (0.181)    (0.195)    (0.206)
##   alcohol            0.377***   0.331***   0.319***
##                     (0.017)    (0.017)    (0.017)
##   volatile.acidity             -1.298***  -1.066***
##                               (0.102)    (0.104)
##   sulphates                                1.012***
##                                           (0.121)
## ------------------------------------------------
##   R-squared             0.235      0.307      0.337
##   adj. R-squared        0.235      0.307      0.336
##   sigma                 0.691      0.658      0.643
##   F                   475.914    342.767    261.977
##   p                     0.000      0.000      0.000
##   Log-likelihood    -1621.453  -1544.961  -1510.723
##   Deviance            736.898    667.514    638.611
##   AIC                3248.905   3097.922   3031.446
##   BIC                3264.938   3119.299   3058.166
##   N                  1547       1547       1547
## ================================================
```

Even the linear model with all 3 most significant predictors, it can count for roughtly 33.7% the variation on quality.

# 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?

Features like alcohol, volatile.acidity and sulphates have relatively stronger (though below moderate level), while other attributes have quite small or even nearly no relationship with quality.

The pH analysis and citric analysis parts are both frustrating to me, I find both are not significant indicators to wine quality even with extreme values. This is to my great surprise.

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

Yes, I listed 4 pairs of corelated attributes in above analysis: 1. citric.acid ~ fixed.acidity, positively associated 2. citric.acid ~ volatile.acidity, negatively associated 3. fixed.acidity ~ pH, negatively associated 4. density ~ fixed.acidity, positively associated

## What was the strongest relationship you found?

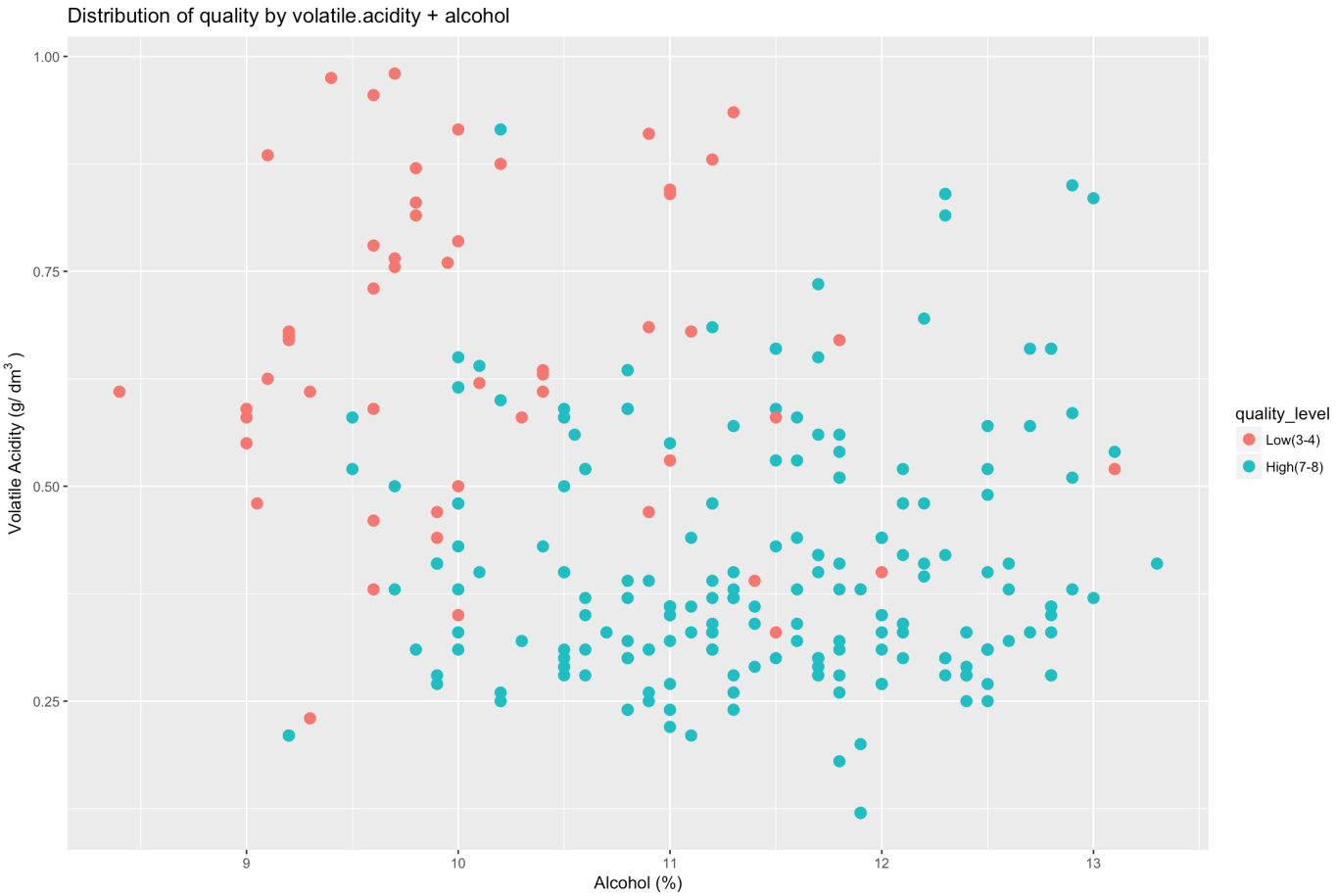
pH and fixed acidity is the pair with strongest relationship ($R = -0.68$) I found.

Besides, 2 pairs have strong relationship with $R = 0.67$: citric.acid ~ fixed.acidity and density ~ fixed.acidity

All these three pairs reach moderate level of association.

# Multivariate Plots Section

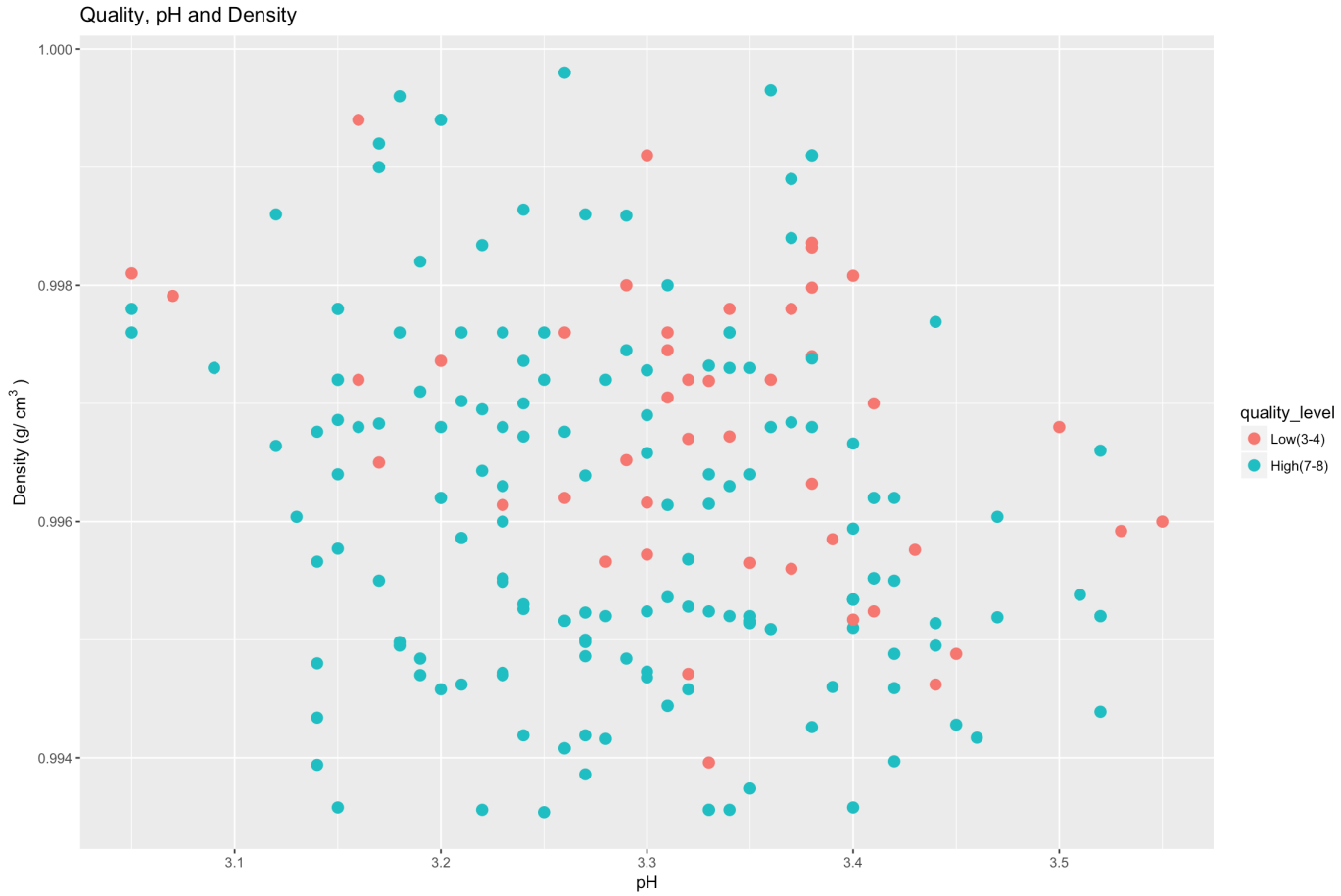
Because the moderate quality has much more variation, here I only consider low/high quality in plots of this section.




There's apparent pattern that high quality wine will more likely to fall into bottom right side of this graph, which means higher alcohol level plus lower volitile acidity will likely to indicate a better red wine.

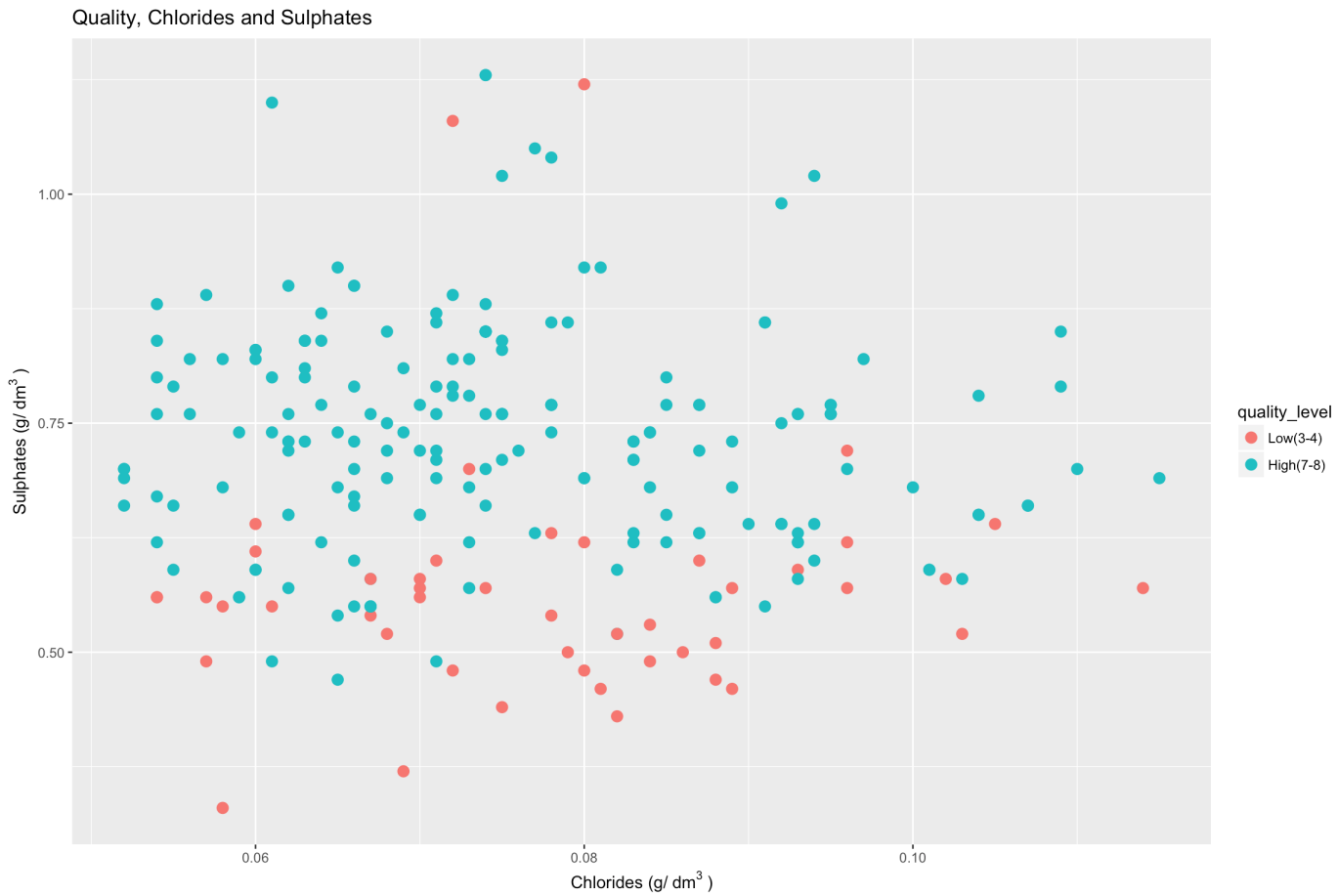
Now I will exam another 4 groups of attributes of great interests (Want to check if anecdotes I heard are right):

- Quality, pH and Density



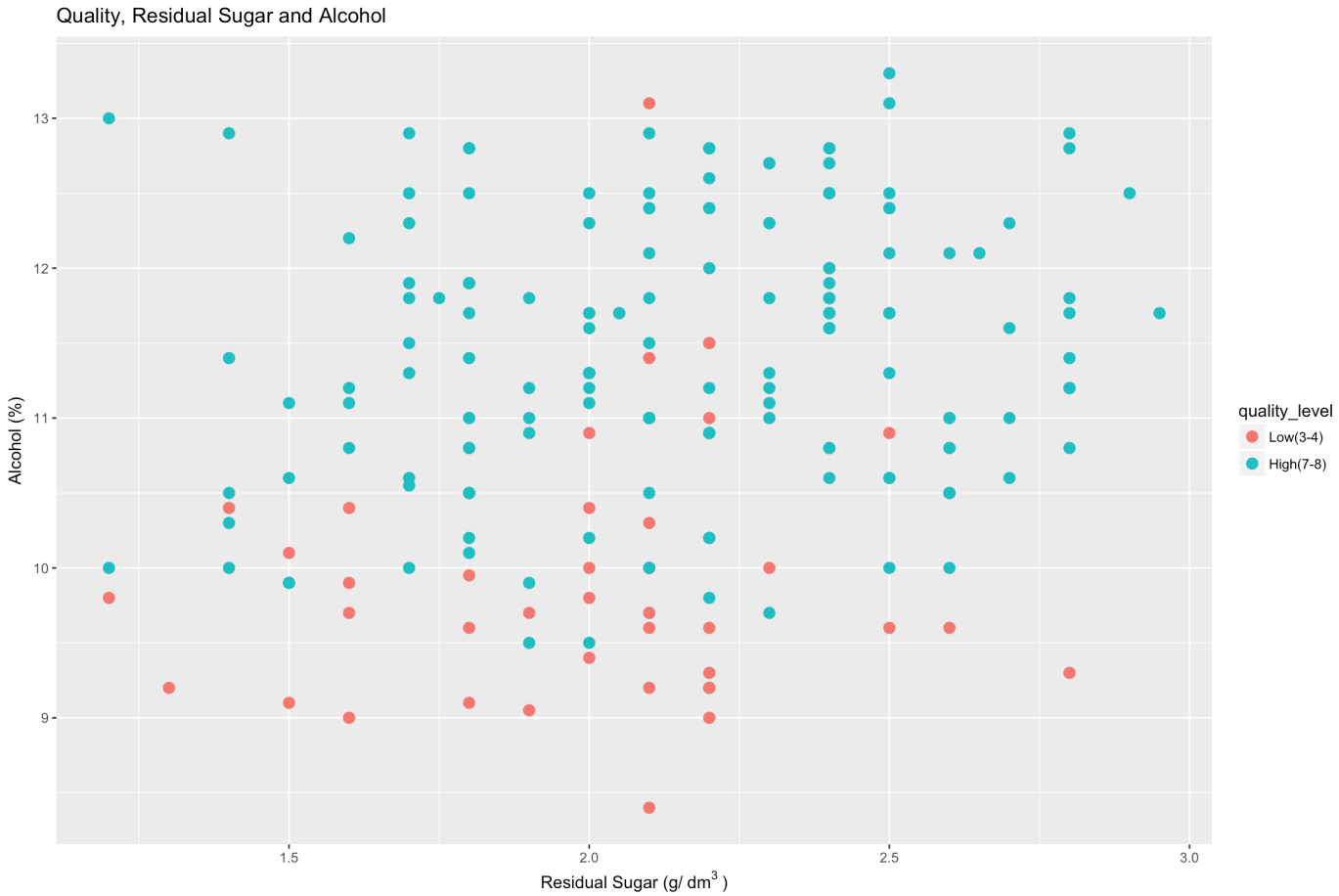

This plot indicate that low quality red wines are a little bit more likely to have higher pH and high density together. However this difference is not very obvious.

- Quality, Chlorides and Sulphates



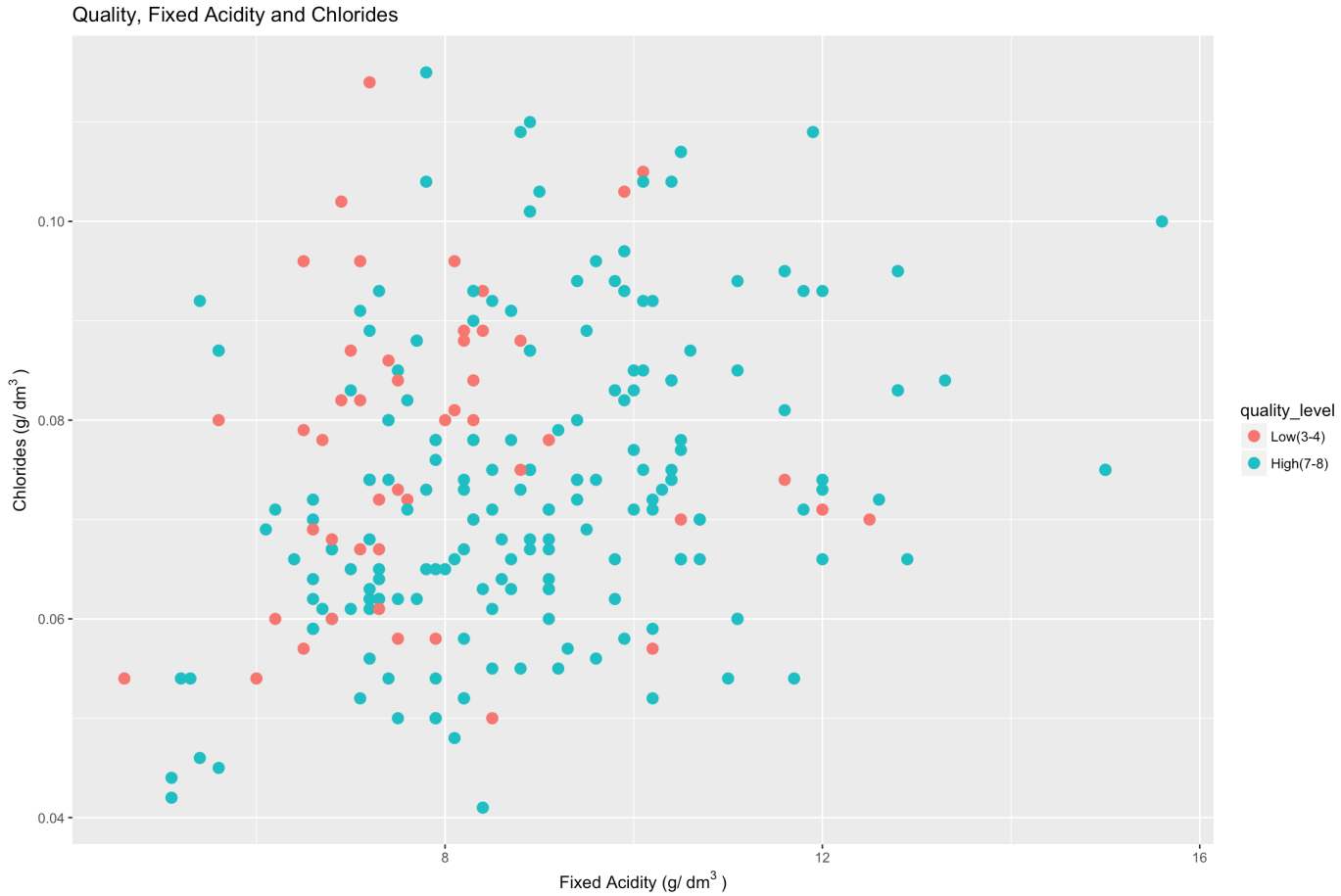
After remove outliers, we can see from the plot that high quality wines are more likely to have high sulphates. The reason could be that sulphates is preservative which can maintain a wine's freshness because of its antioxidant and antibacterial properties.

- Quality, Residual Sugar and Alcohol



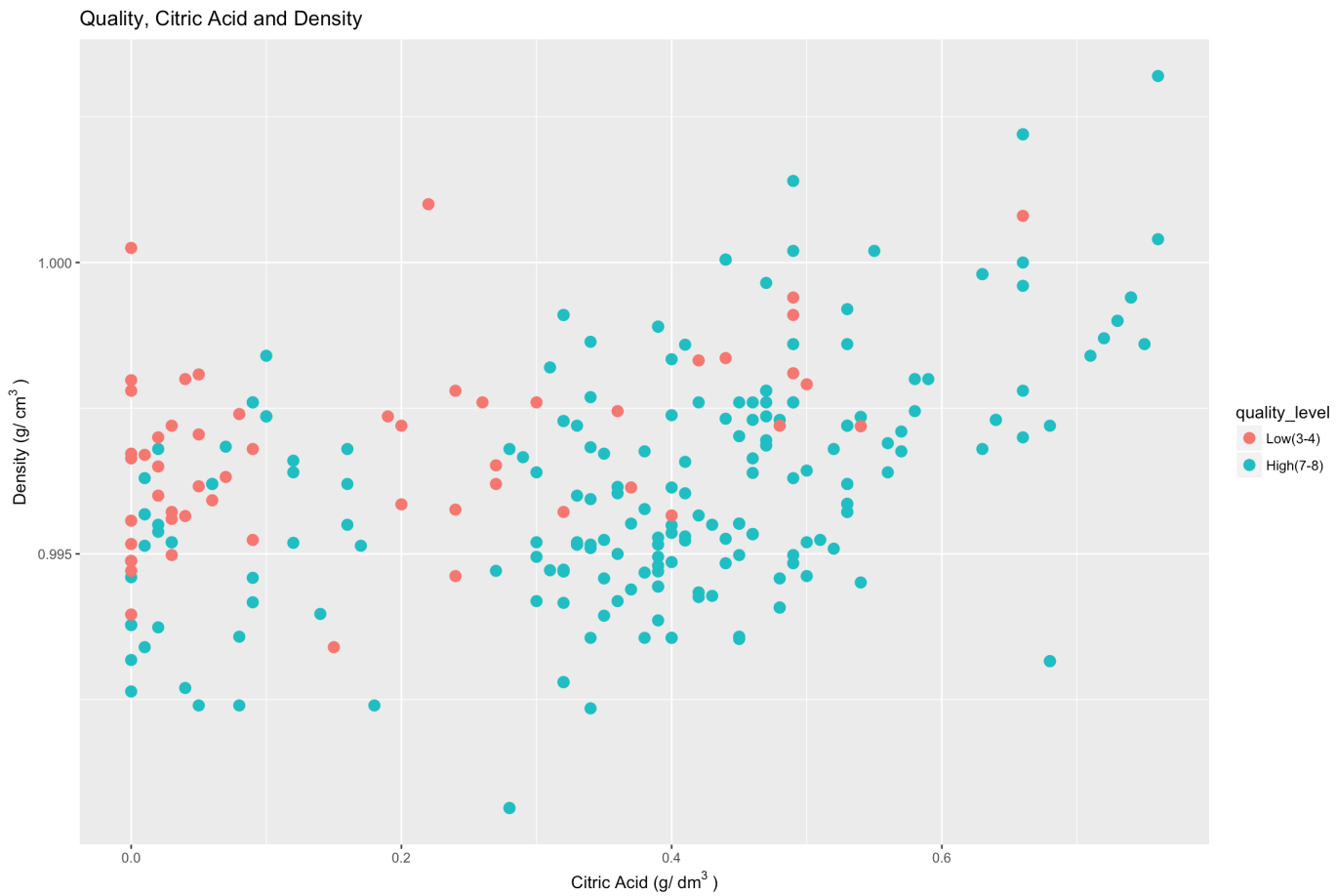

As residual sugar doesn't affect quality, alcohol is main indicator of wine quality - higher-quality wines are more likely having higher percentage alcohol.

- Quality, Fixed Acidity and Chlorides




As there're more red points at top left side, which means low quality wines tend to have less fixed acidity plus more chlorides. However, this pattern is not very strong.

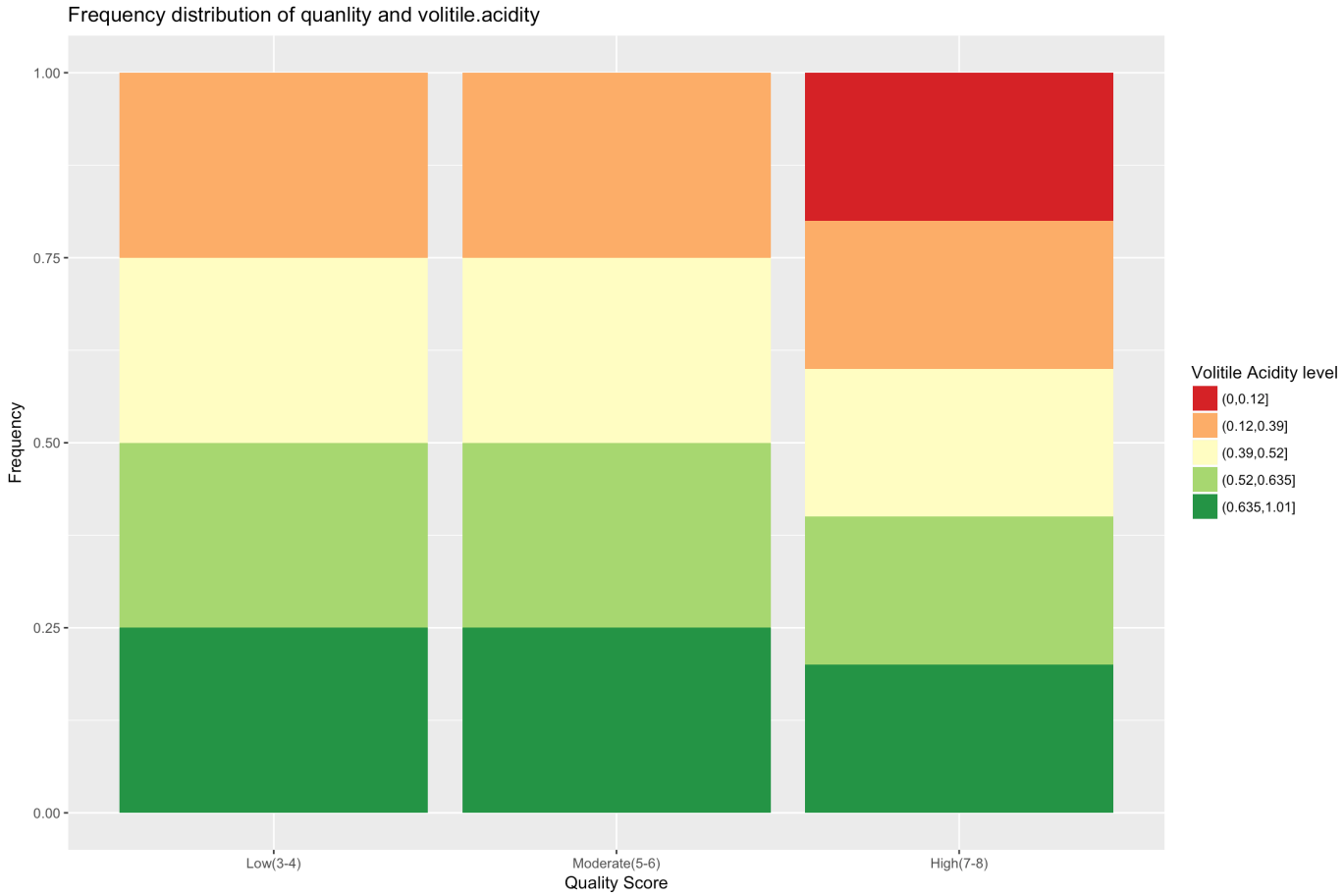
-Quality, Citric Acid and Density-



Citric Acid and Density are the 2 most significant indicators for quality. Here I plot them together. From the chart we can see low quality wines tend to have high density, which is reasonable because more alcohol, as discussed above, can enhance quality but meanwhile reduce density. On the other hand, as majority blue points are lying at right side, high quality wines tend to have higher citric acid.

-How volatile.acidity distributed in wines with different quality-

The plot that quality is associated with volatile.acidity has data dispersed, it might be hard to figure out pattern, but I can draw some statistics conclusion if based on probability.




From such plot, we can see high quality red wine tends to have more probability of low volitile acidity. So if a red wine has strong acid smell, it possibly a low-quality red wine.

# 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?

From these plots, we can see: - Higher pH, higher sulphates, higher alcohol and higher fixed acidity are more likely to indicate red wines with higher quality - it's hard to judge quality by density, chlorides and residual sugar values

## Were there any interesting or surprising interactions between features?

It's surprising these pairs of attributes are almost independent to each other, we can either see horizontal / vertical pattern or the data point massed up.

## OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Yes, in above analysis, I created a linear model to see how the attributes like alcohol, sulphates, and volatile.acidity interact with quality.

The pros of this model include: - it's straightforward and self-explained what kind of wine will be with high quality - easy to plot and predict

However, there're cons: - even with most significant attributes, the model can explain only 33.6% variation of quality, which means over 66.3% variation is out of control, which might make this model unreliable - the data is distributed in a very complicated pattern, applying linear model might be a naive choice to get rid of too much information

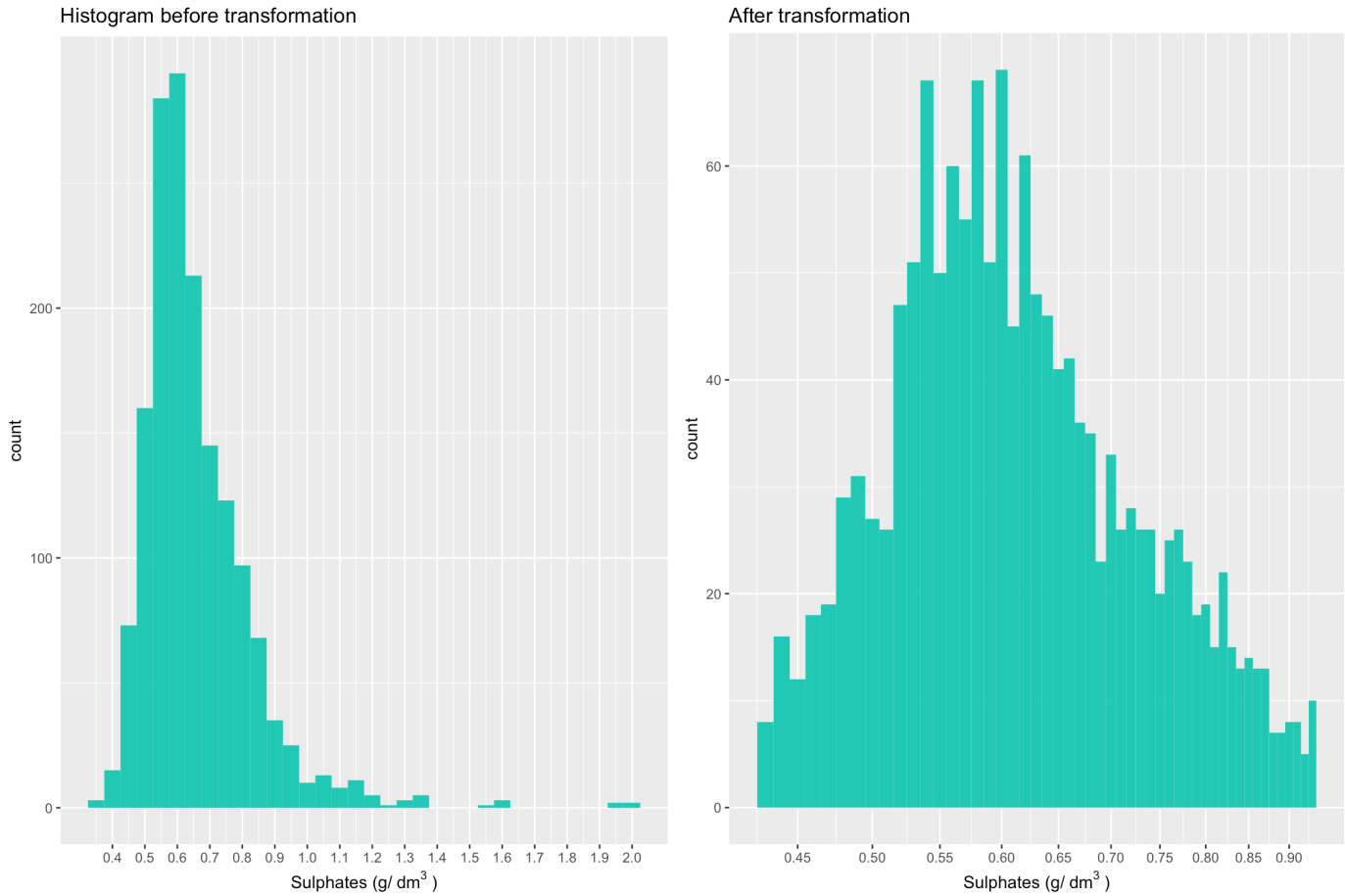
To better improve the result, both advanced model technics and more dimentional data are required, for example, how each score of quality such red wine gets.

---

# Final Plots and Summary

## Plot ONE

Transformation helps a lot to check details of an attribute with skewed distribution.




## Description ONE

From the univariate analysis section, we can see most of the attributes of the red wine data set are numeric, and right skewed with outliers located at upper side. By transformation and removing outliers, we can zoom in and get distribution of majority data. For example, in above plot, we can see the most common value for

sulphates is 0.60.

# Plot TWO

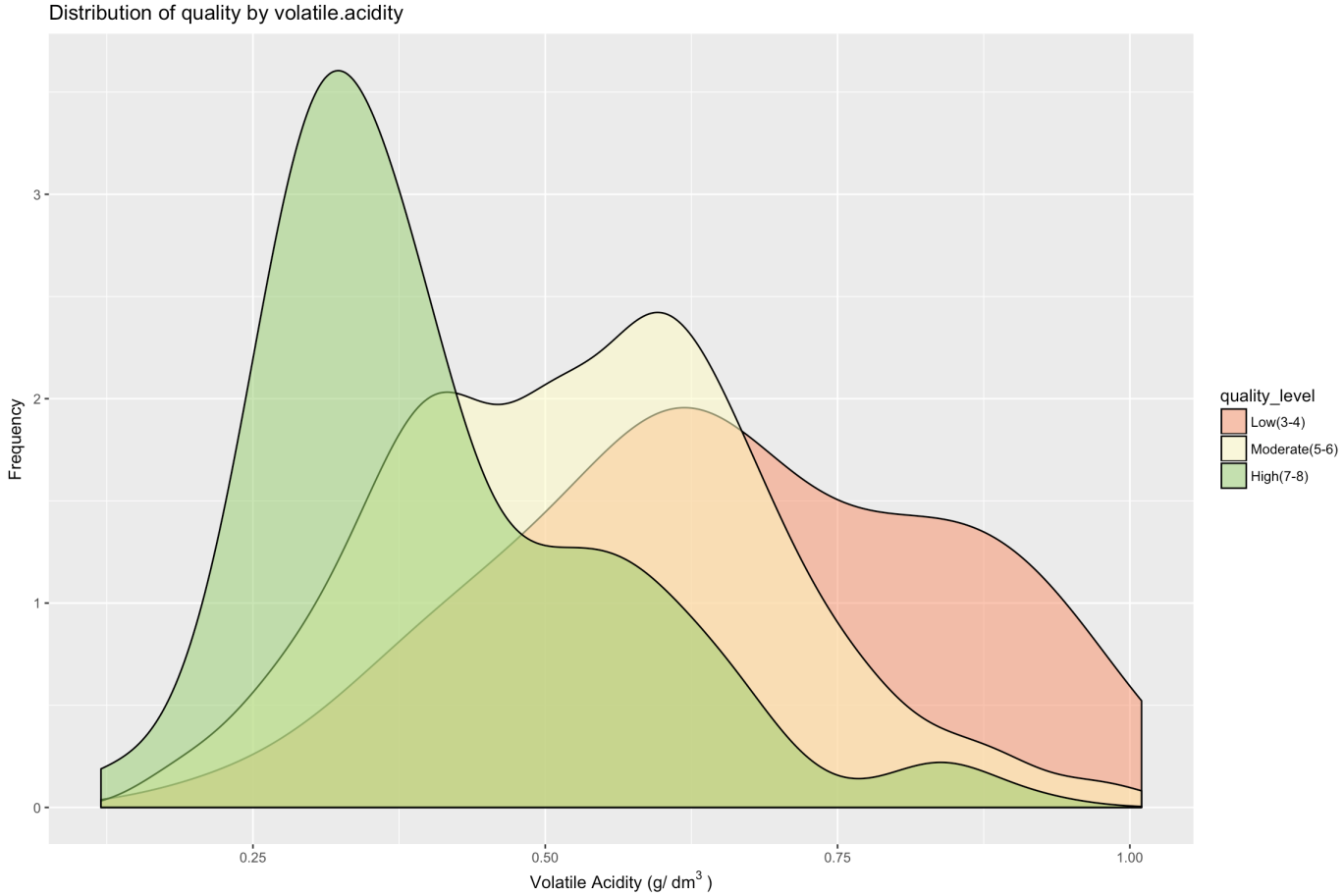
From bivariates analysis section, we can see alcohol is the strongest attribute to predict quality:


# Description TWO

From the plot, we can see though the data is lying everywhere, there's a pattern can be drawn that the quality of red wine will increase with alcohol percentage.

# Plot Three

In the above section, I checked how quality interacts with 2 other attributes, the 2nd strongest relationship are quality vs. volatile.acidity, from below plot we can such correlationship:




## Description Three

By the plot above, we can see the high-quality wine will be more likely to have less volatile acidity, while low quality wine volatiles a lot.

---

# Reflection

In this report, I firstly explore all attributes by their distributions and list the questions I'm interested in. For example, can sweetness, pH and citric acid improve quality?

Then, in bivarate analysis part, I create correlation matrix, select the most 3 significant attributes (alcohol, sulphates, volatile.acidity) associated with quality, and then explore how the data distributes along these 3 attribute dimentions. I find out by histogram and boxplot that data are right skewed distributed along these 3 dimension respectively. To answer my concerns: if sweetness, pH and citric acid can improve quality, I explore the relationship one by one.

Next, I plot how these 3 attributes associate with quality by scatter plot and linear model lines: - in alcohol vs. quality plot and sulphates vs. quality plot, I find both have positive association, thought the association is not very strong - in volatile.acidity, I find negative association I use jitter and set transparency to improve the data visualization. Besides, I use linear model to check how these 3 attributes can be fitted by data and then plot 4 scatter plot with different combination to see the strongest associate between attributes.

Furthermore, I use leveled scatter plot to get idea how quality level distributed in different attributes' dimension scales and reach the assersion: Higher pH, higher sulphates, higher alcohol and higher fixed acidity are more likely to indicate red wines with higher quality.

Finally, I use colored histogram to show that higher quality level wine tends to involve less volatile acidity.

# Forecast

This red wine dataset has 12 attributes messed up. Explorasive data analysis does provide with an efficient way to capture idea. But to improve the accuracy of predicting quality of red wine, we can try more improvements, including: - improve the data, with more data on low / high quality of wines, and more detailed description on how the quality score were given. Better involve more features, like the year of harvest, brew time, location of Vineyard and so on - use machine learning, like SVM / Decision Tree to mine more details of attributes in more advanced dimensional vision

# Reference


- UC Davis, whats-in-wine: http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity (http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity)
- Predicting Red Wine Quality: Exploratory Data Analysis: https://rpubs.com/jeknov/redwine (https://rpubs.com/jeknov/redwine)
- Analysis of Red Wines Dataset by Yohann Lucas: https://rpubs.com/kaltera/UdacityP3 (https://rpubs.com/kaltera/UdacityP3)
- ggplot docs: http://docs.ggplot2.org/0.9.3.1/position_fill.html (http://docs.ggplot2.org/0.9.3.1/position_fill.html)
- Sugar in Wine, The Great Misunderstanding: http://winefolly.com/update/sugar-in-wine-misunderstanding/ (http://winefolly.com/update/sugar-in-wine-misunderstanding/)
- The Truth About Sulfites in Wine & the Myths of Red Wine Headaches: http://www.thekitchn.com/the-truth-about-sulfites-in-wine-myths-of-red-wine-headaches-100878 (http://www.thekitchn.com/the-truth-about-sulfites-in-wine-myths-of-red-wine-headaches-100878)