Analyzing red wine quality

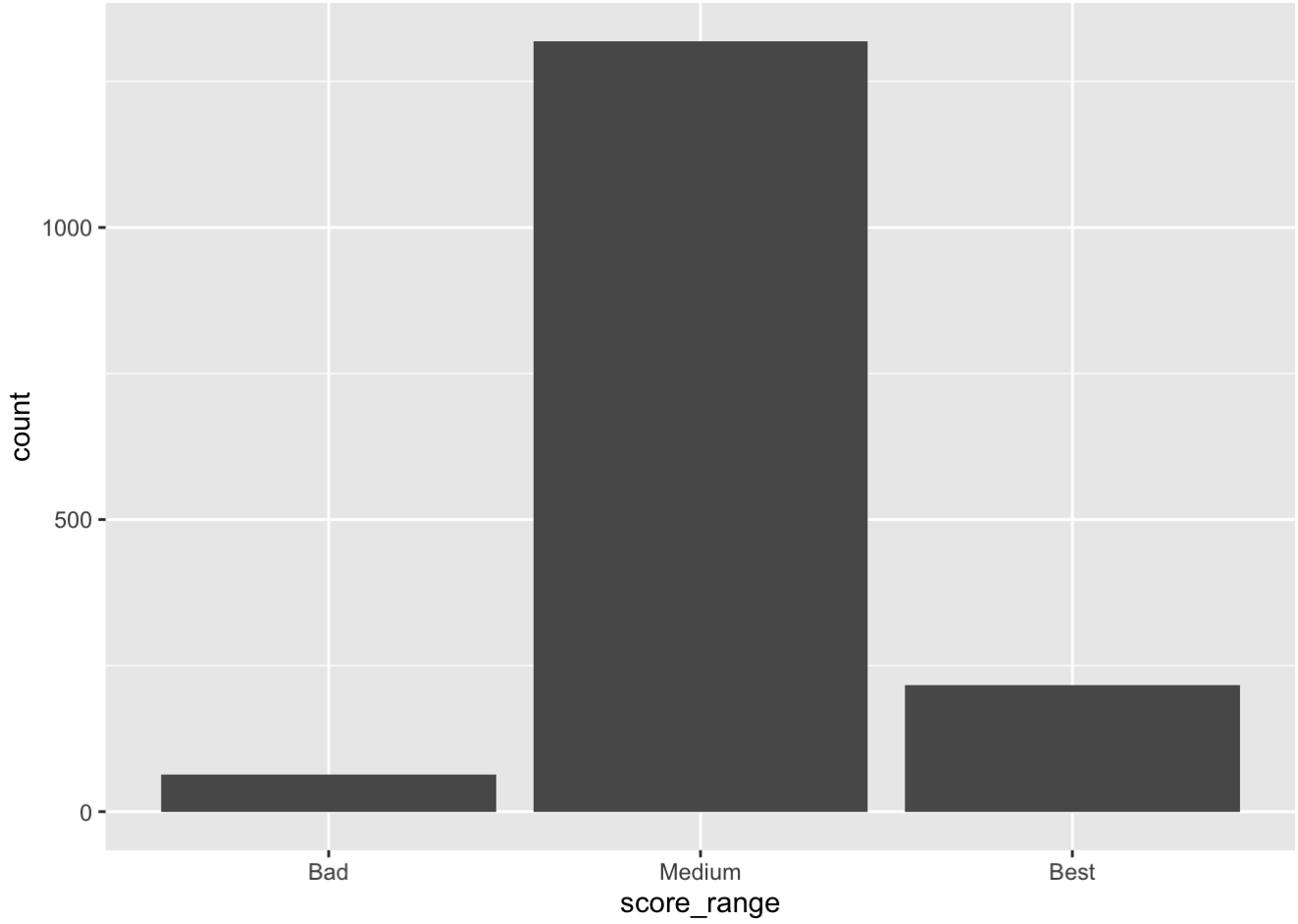
Loading the data

I categorized the quality variable into score_rating consisting of Bad, Medium, Best categories according to the group of quality values it falls within. Therefore:

- 0-5 : Bad
- 5-7 : Medium
- 7-10 : Best

Using the new variable, I created a histogram to see which category the observations fall mostly.

```
##  [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"
```




```
##  Ord.factor w/ 3 levels "Bad"<"Medium"<..: 2 2 2 2 2 2 2 3 3 2 ...
```

Descriptive statistics of the data:

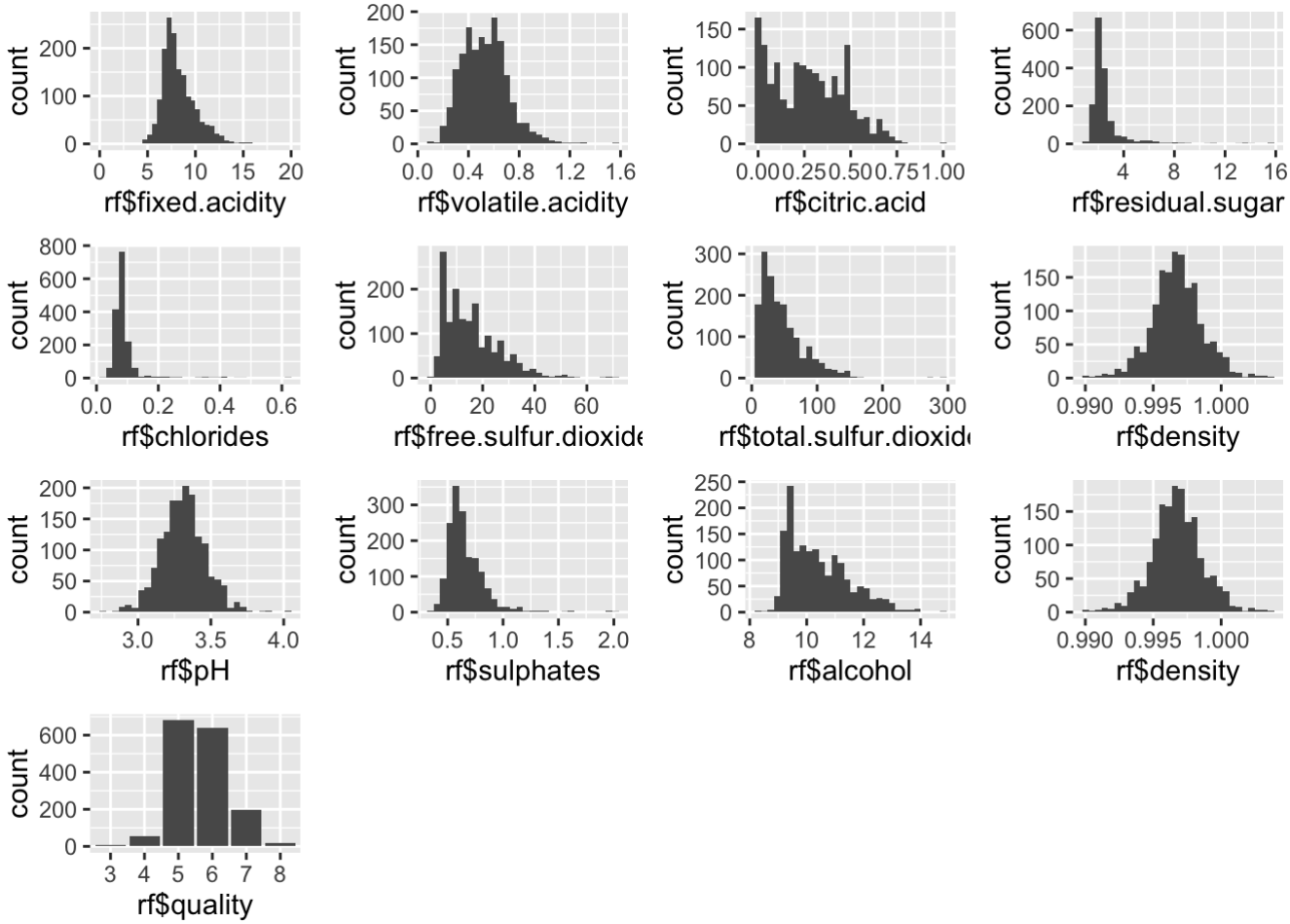
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean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

All the variables consists of outliers but no surprising patterns are seen in the mean, median and quartiles of the data.

Univariate Plots Section:

In order to see the distribution of data and the frequency of the levels of features included in the wine, I plotted histograms for each feature. Checking how they are distributed, I can find out the features that are responsible for a better quality wine.

```
##  [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"             "score_range"
```




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

On plotting the data into histograms, I found:

- Fixed acidity: Positively skewed with a lot of outliers and long tailed

  ```
  - Mean = 8.32
  - Median = 7.90
  - 1st quartile = 7.10
  - 3rd quartile = 9.20
  ```

- Volatile acidity: Almost bimodal and long tailed with outliers

  ```
  - Mean = 0.5728
  - Median = 0.52
  - 1st quartile = 0.39
  - 3rd quartile = 0.64
  ```

- Citric acid: Have large number of 0 values

```
- Mean = 0.271
- Median =0.26
- 1st quartile = 0.09
- 3nd quartile = 0.42
```

- Residual sugar: Have extreme outliers

```
- Mean = 2.539
- Median = 2.2
- 1st quartile = 1.9
- 3rd quartile = 2.6
```

- Chlorides: Have extreme outliers

```
- Mean = 0.08747
- Median = 0.079
- 1st quartile = 0.07
- 3rd quartile = 0.09
```

- Free, total sulfur dioxide, sulphates, alcohol:

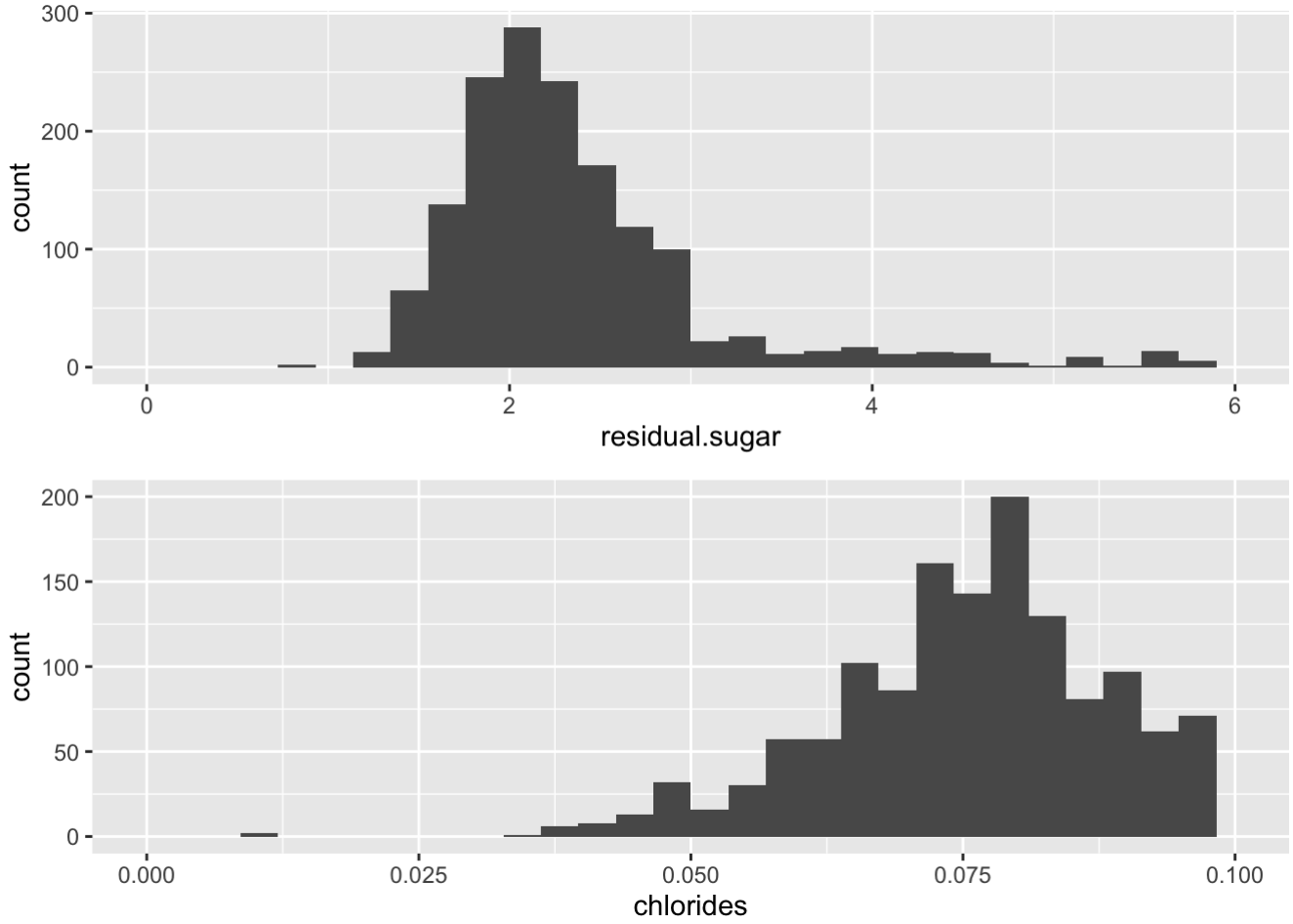
```
Positively skewed and long tailed

- Mean of free.sulfur.dioxide = 15.87
- Median of free.sulfur.dioxide = 14
- 1st quartile = 7
- 3rd quartile = 21

- Mean of total.sulfur.dioxide = 46.47
- Median = 38
- 1st quartile = 22
- 3rd quartile = 62

- Mean of sulphates = 0.6581
- Median = 0.62
- 1st quartile = 0.55
- 3rd quartile = 0.73

- Mean of alcohol = 10.42
- Median = 10.20
- 1st quartile = 9.5
- 3rd quartile = 11.1
```

Density and pH are normally distributed, with few outliers

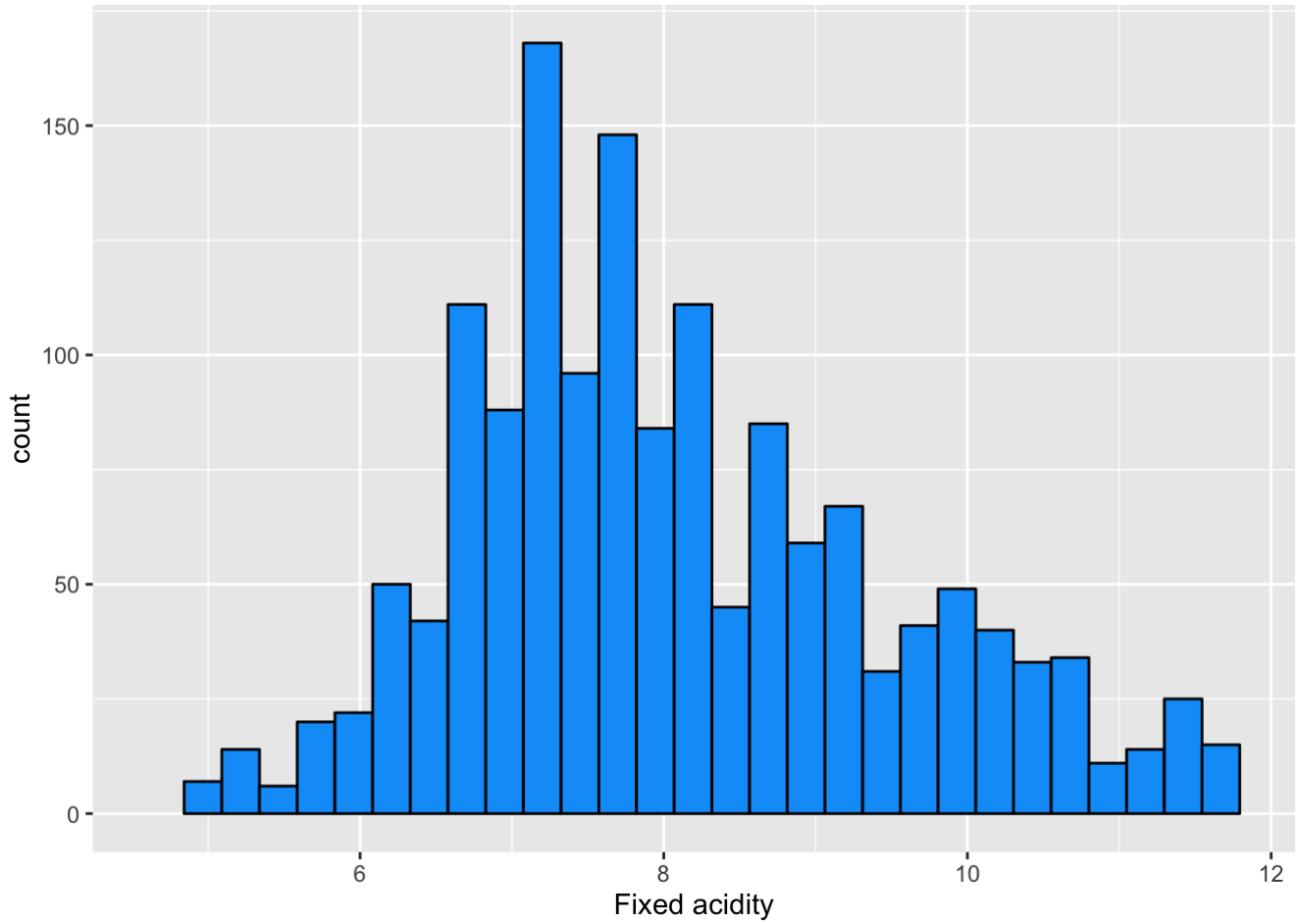
```
- Mean for density = 0.9967
- Median = 0.9968
- 1st quartile = 0.9956
- 3rd quartile = 0.9978

- Mean for pH = 3.311
- Median = 3.310
- 1st quartile = 3.210
- 3rd quartile = 3.4
```

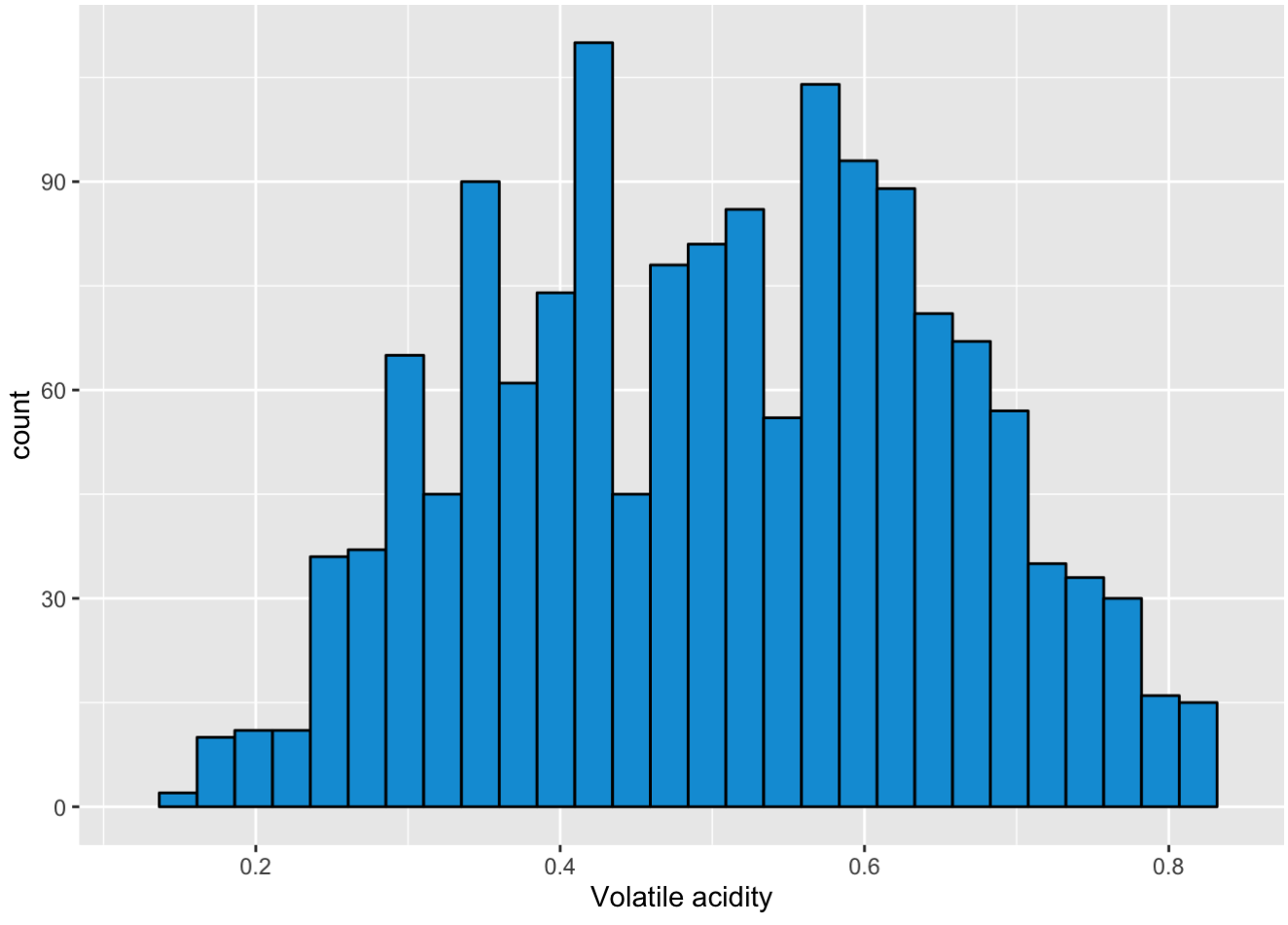
The mean and median of density and pH are almost equal which also implies they have a normal distribution.

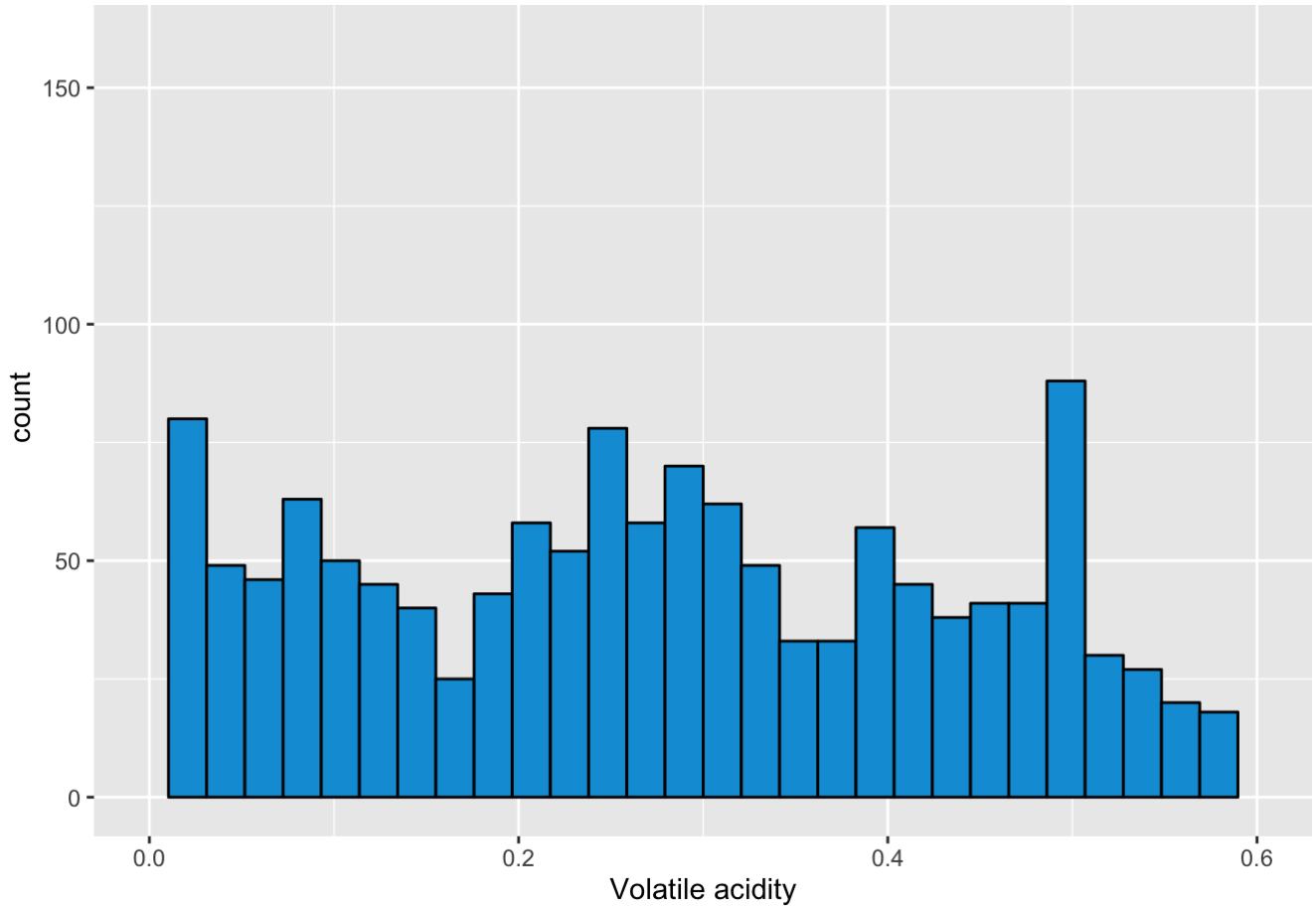
Since residual sugar and chlorides contained a large number of outliers, I adjusted the axes to remove them:

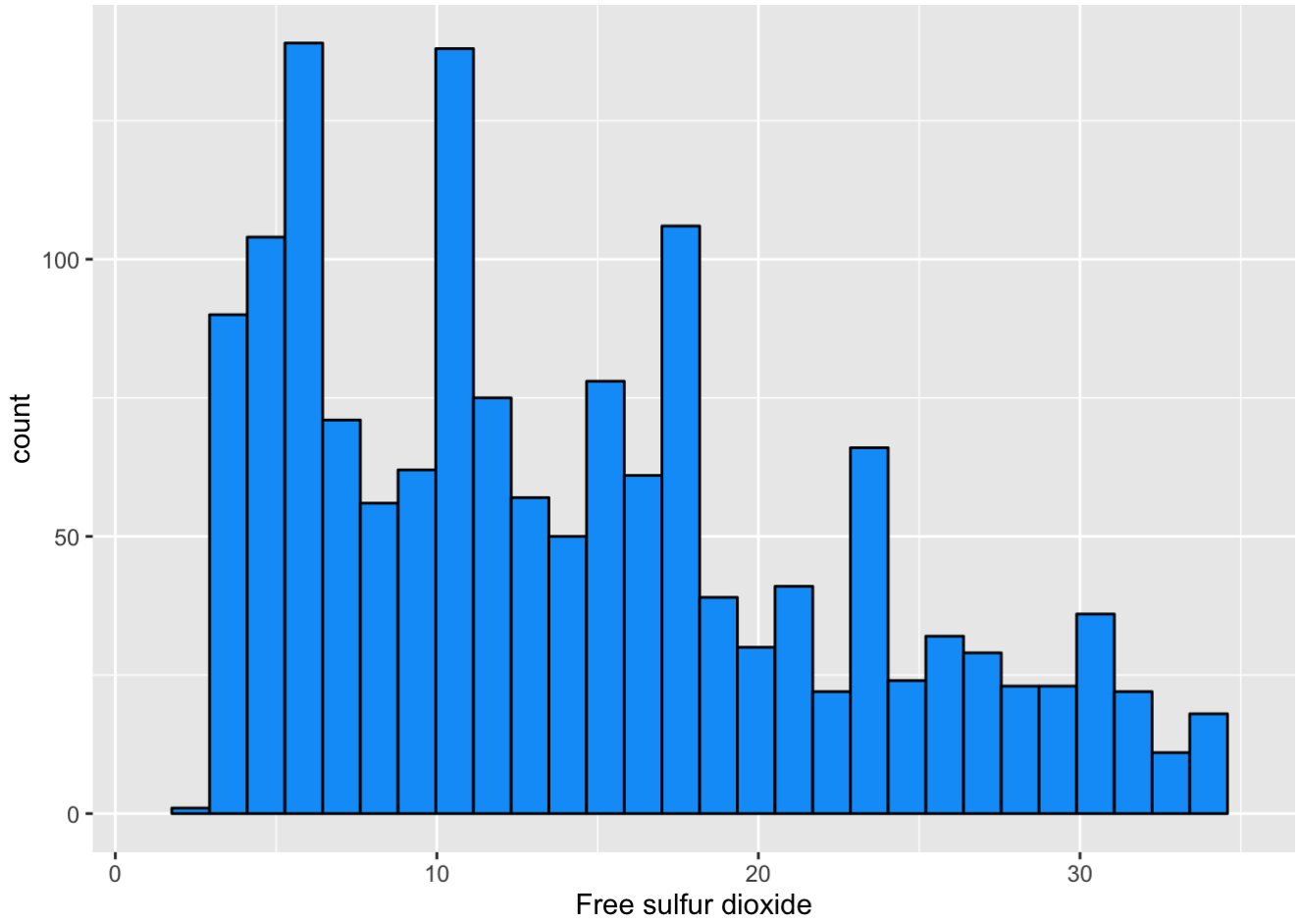


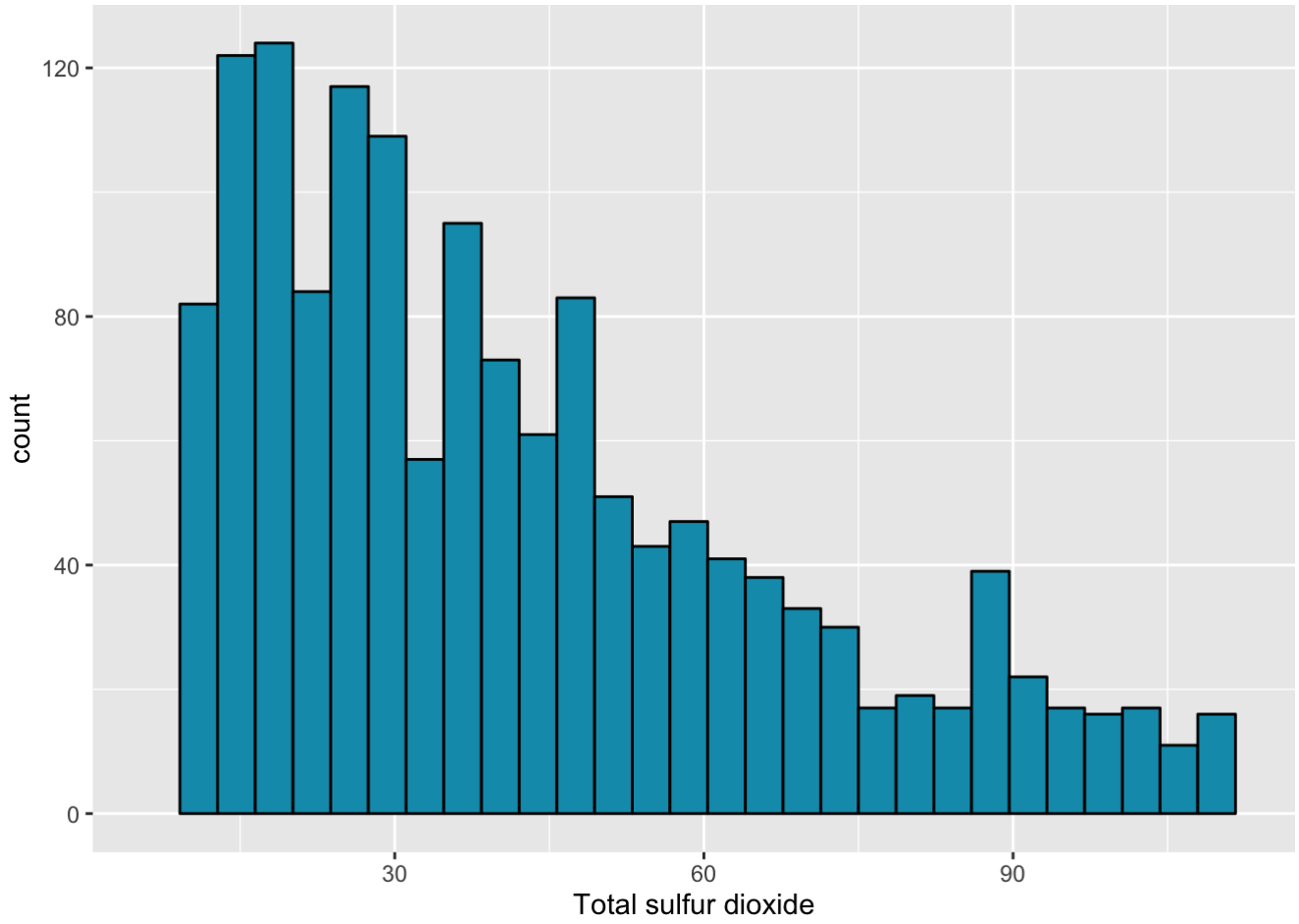
To see the long tailed distributions with few outliers clearly, I used log10 transformations to remove the outliers as well:

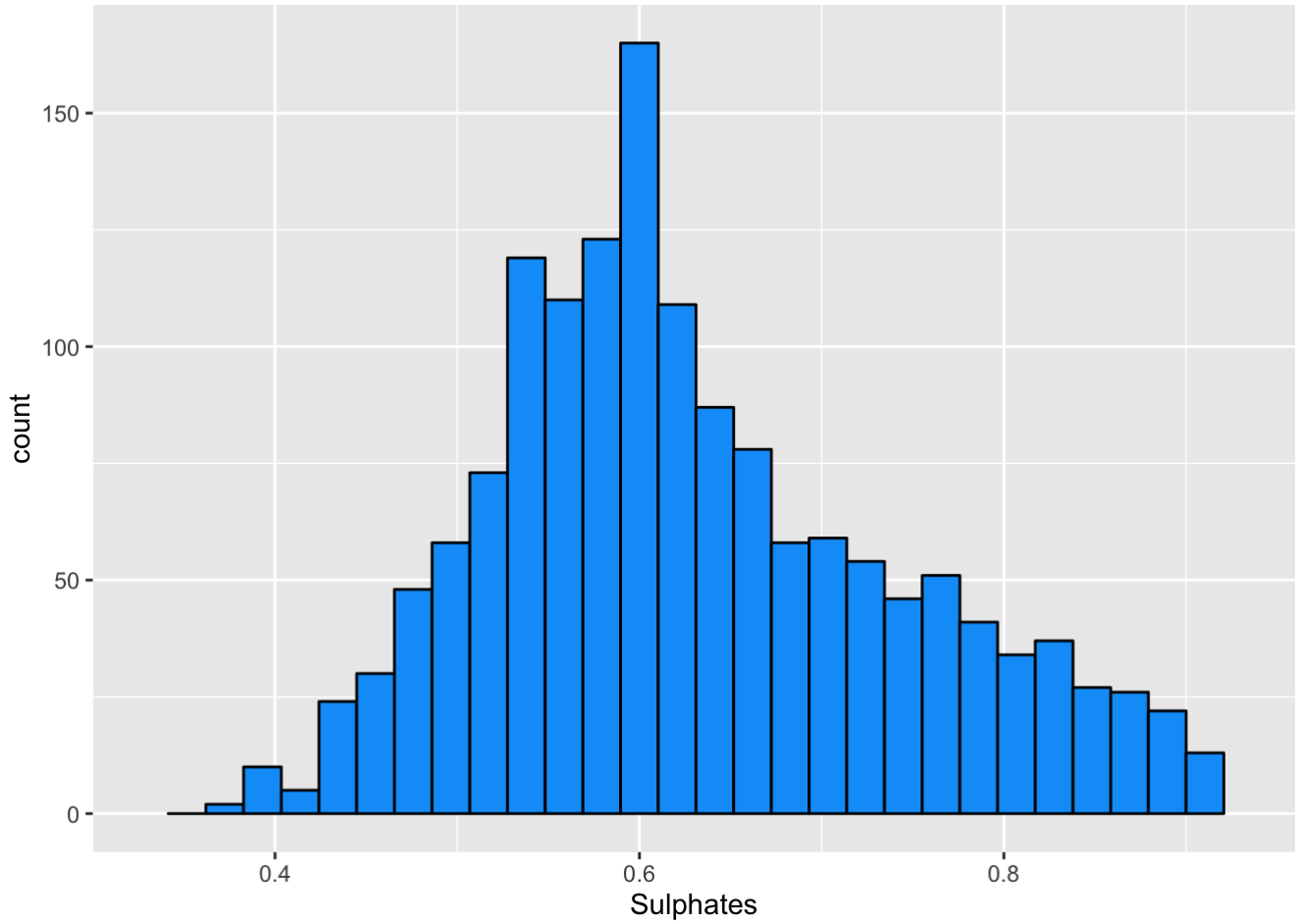

150
100
50
0
count
6
8
10
12
Fixed acidity

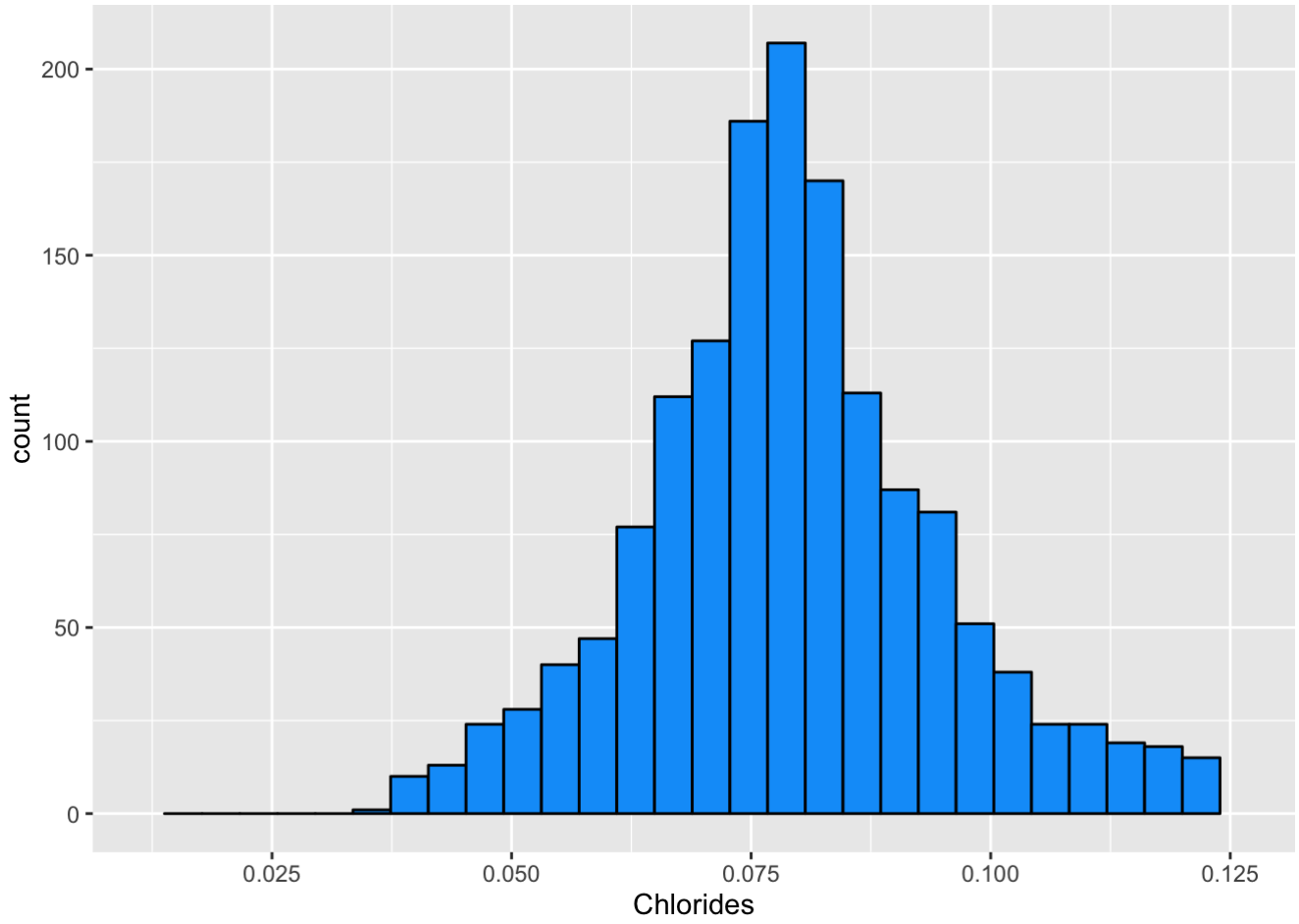

90
60
30
0
count
0.2
0.4
0.6
0.8
Volatile acidity

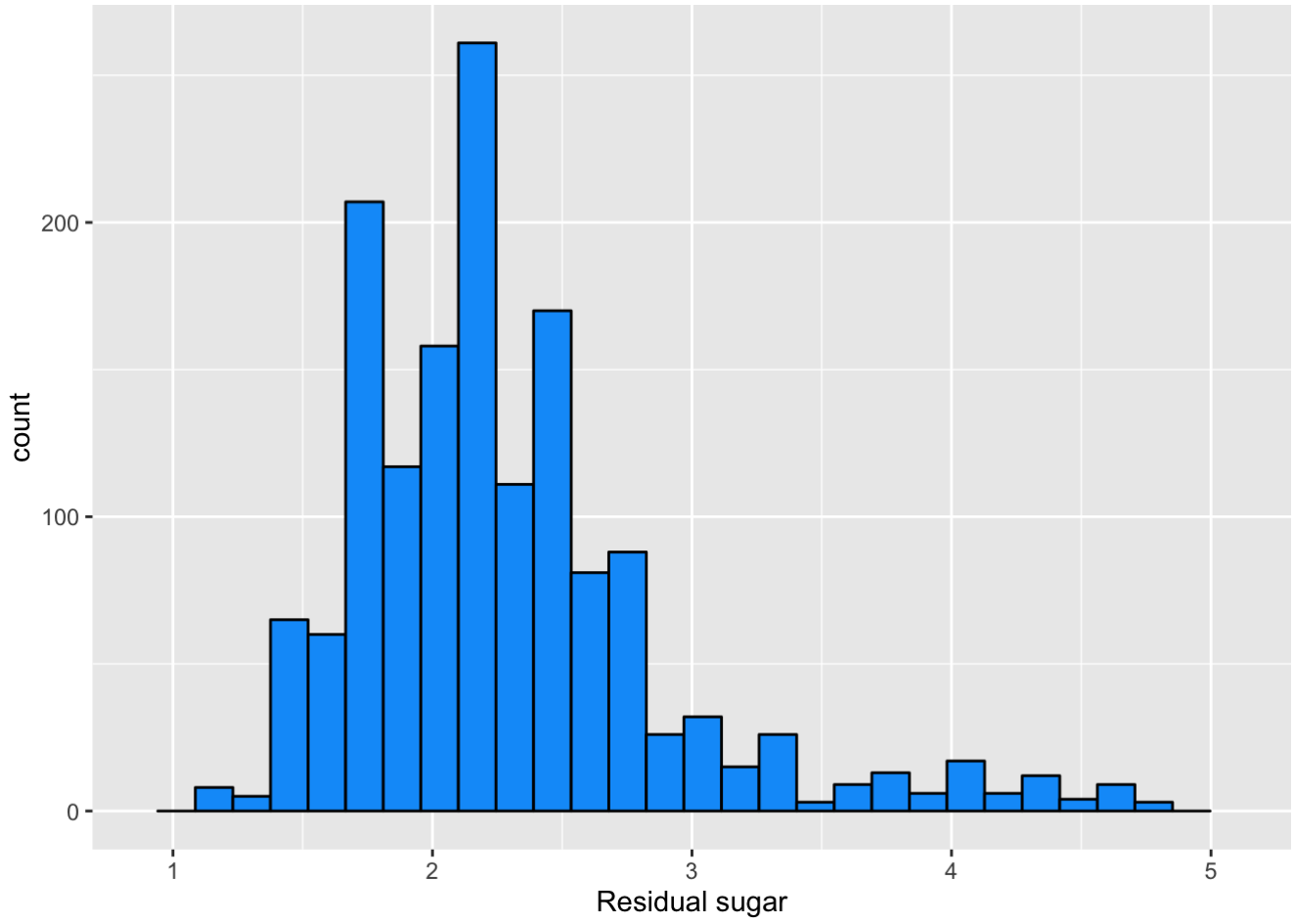

150
100
50
0
count
0.0
0.2
0.4
0.6
Volatile acidity

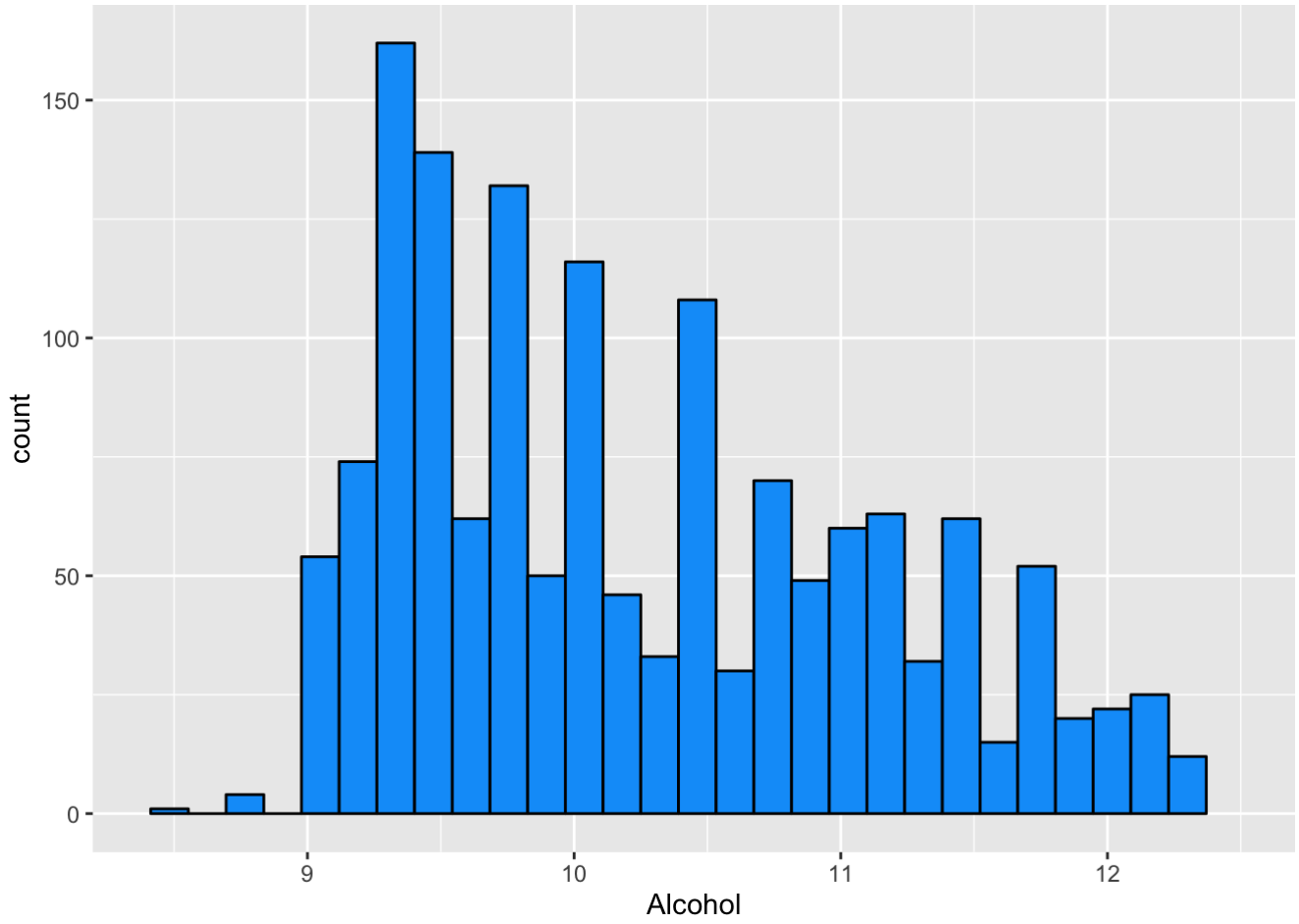

100
50
0
count
0
10
20
30
Free sulfur dioxide

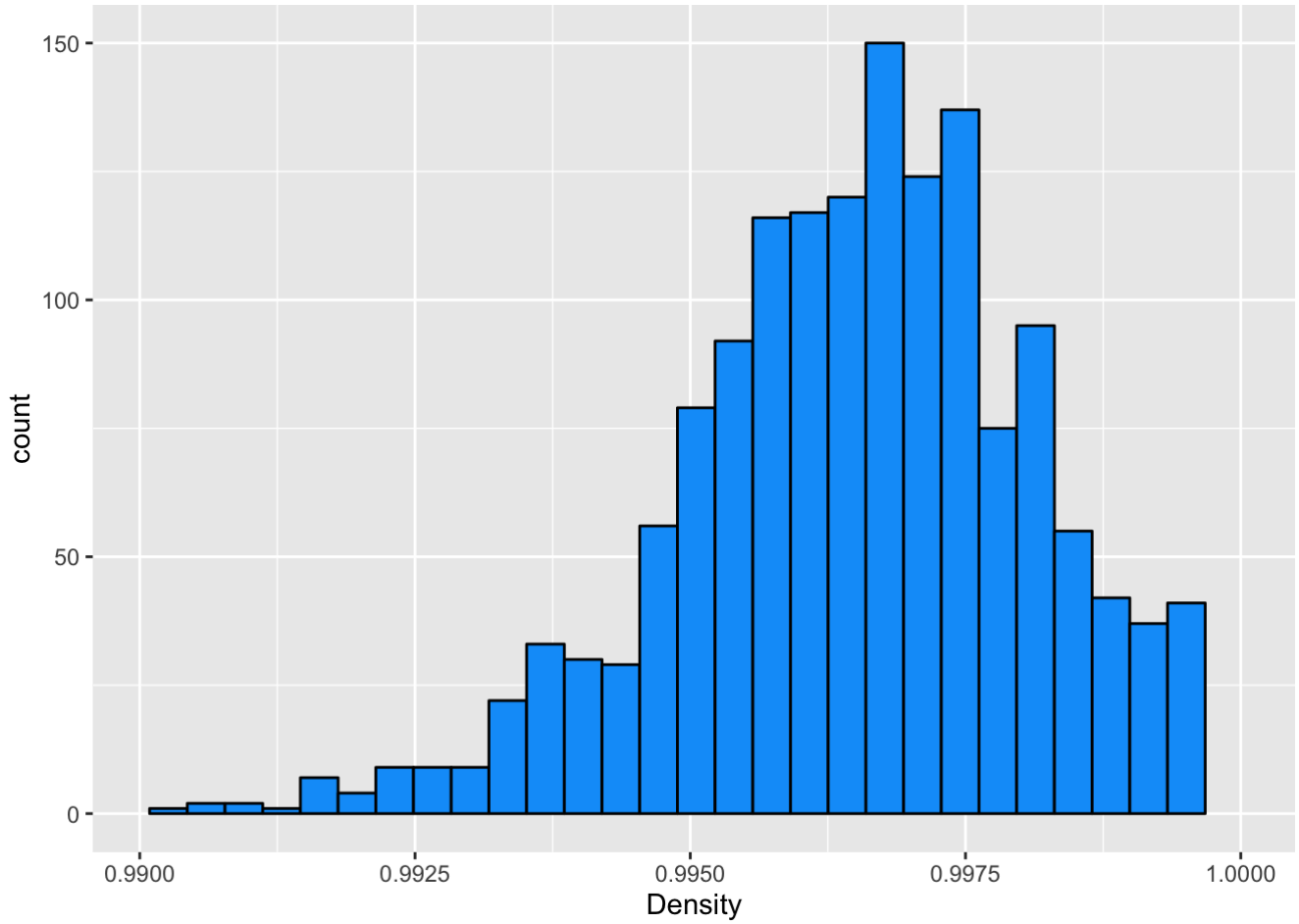

120
80
40
0
count
30
60
90
Total sulfur dioxide


150
100
50
0
count
0.4
0.6
0.8
Sulphates


200
150
100
50
0
count
0.025
0.050
0.075
0.100
0.125
Chlorides


200
100
0
count
1
2
3
4
5
Residual sugar


150
100
50
0
count
9
10
11
12
Alcohol


150
100
50
0
count
0.9900
0.9925
0.9950
0.9975
1.0000
Density

***

On transforming the plots using log10, the fixed, volatile acidity, residual sugar, alcohol and other variables appear to be normally distributed with the exception of citric acid. Citric acid still was not normally distributed.

Univariate Analysis:

What is the structure of your dataset?

There are 1599 observations in the dataset with 13 variables (X, fixed.acidity,volatile.acidity,citric.acid,residual.sugar,chlorides, free.sulfur.dioxide, total.sulfur.dioxide,density, pH, sulphates,alcohol,quality). The variable quality is an ordered factor variables with the following levels.

(worst) ——————-> (best) quality: Worst, medium, Best

Other observations:

- Quality seems to be a factor, discrete variable. It was unordered initially but then I converted into an ordered factor.
- Most of the quality scores were between 3-8 although it was on a 0-10 range.
- Fixed.acidity, volatile.acidity, free and total sulfur dioxides, sulphates, and alcohol are long-tailed.
- residual.sugar and chlorides have outliers
- ph and density almost have normal distibution with a few outliers
- citric.acid appears to be slightly bimodal and have a large number of 0 values
- quality has the maximum counts for 5, 6 and 7

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

The main features in the data set are alcohol, acidity and quality. I'd like to determine which features are best for predicting the quality of a red wine.

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

Among other variables, checking the levels of alcohol and acidity will also help in determining the quality. Acidity and pH are correlated, fixed and volatile acdity seems to be correlated.

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

I created a variable for the quality scoring as worst, medium and best.

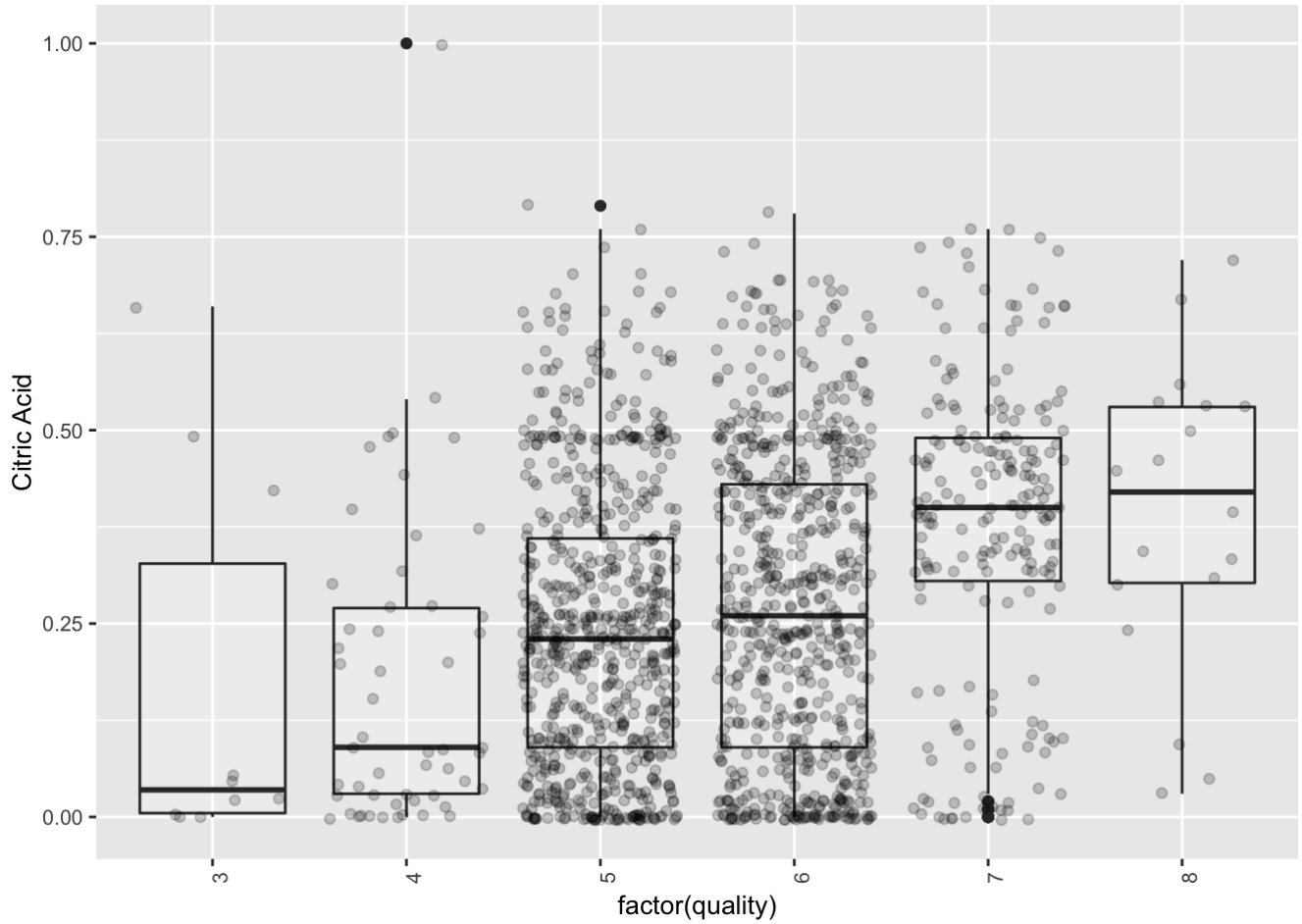
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

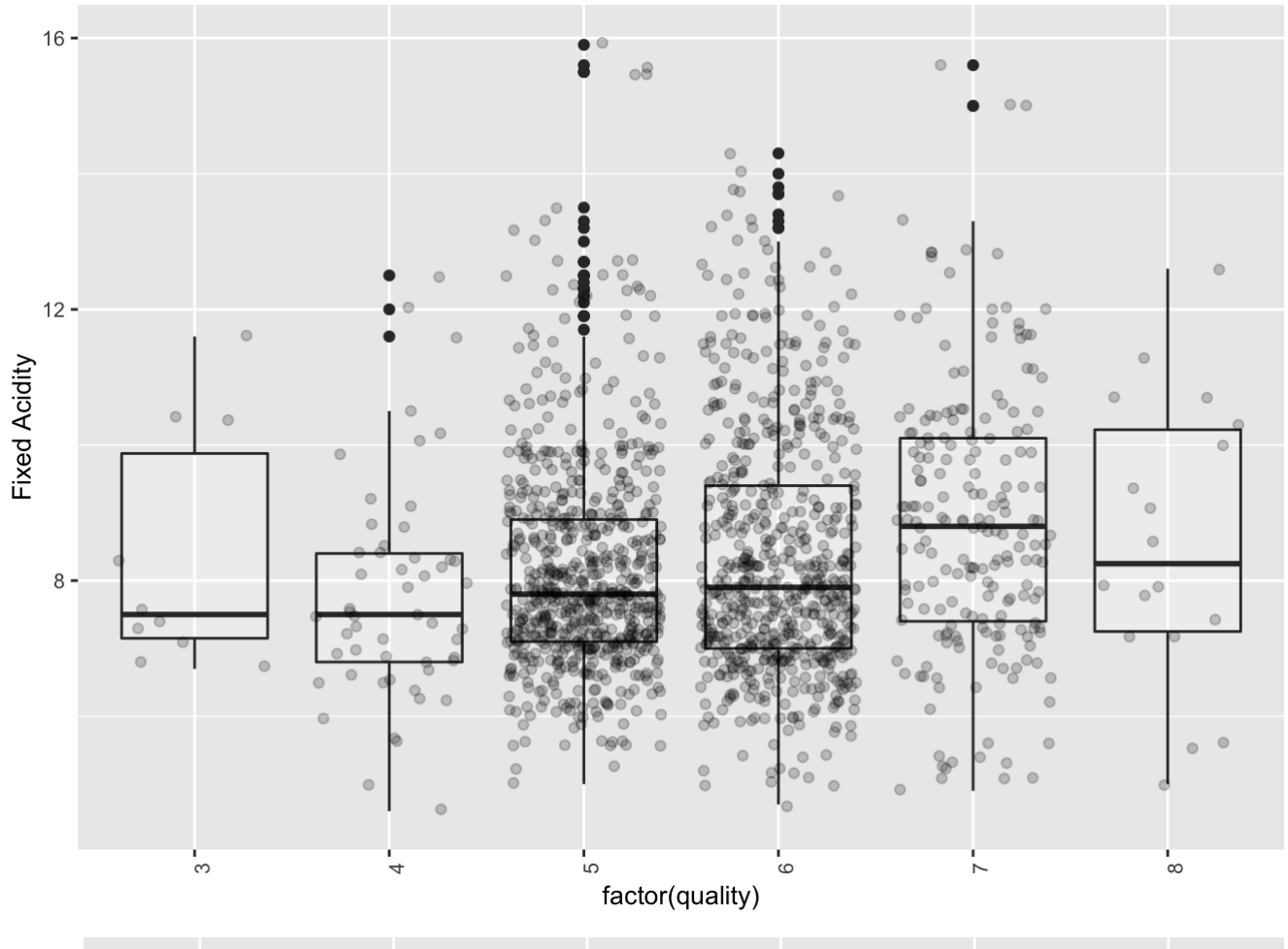
I log-transformed the long tailed data as to convert them into normal distributions.On transforming all the long tailed plots into plots using log10, I found volatile.acitdity and fixed.acidity are normally distributed. citric.acid and free.sulfur.dioxide donot have normal distibutions even after transforming the data. To avoid looking at the long tailed data, I have adjusted the axes. The data is tidy data.

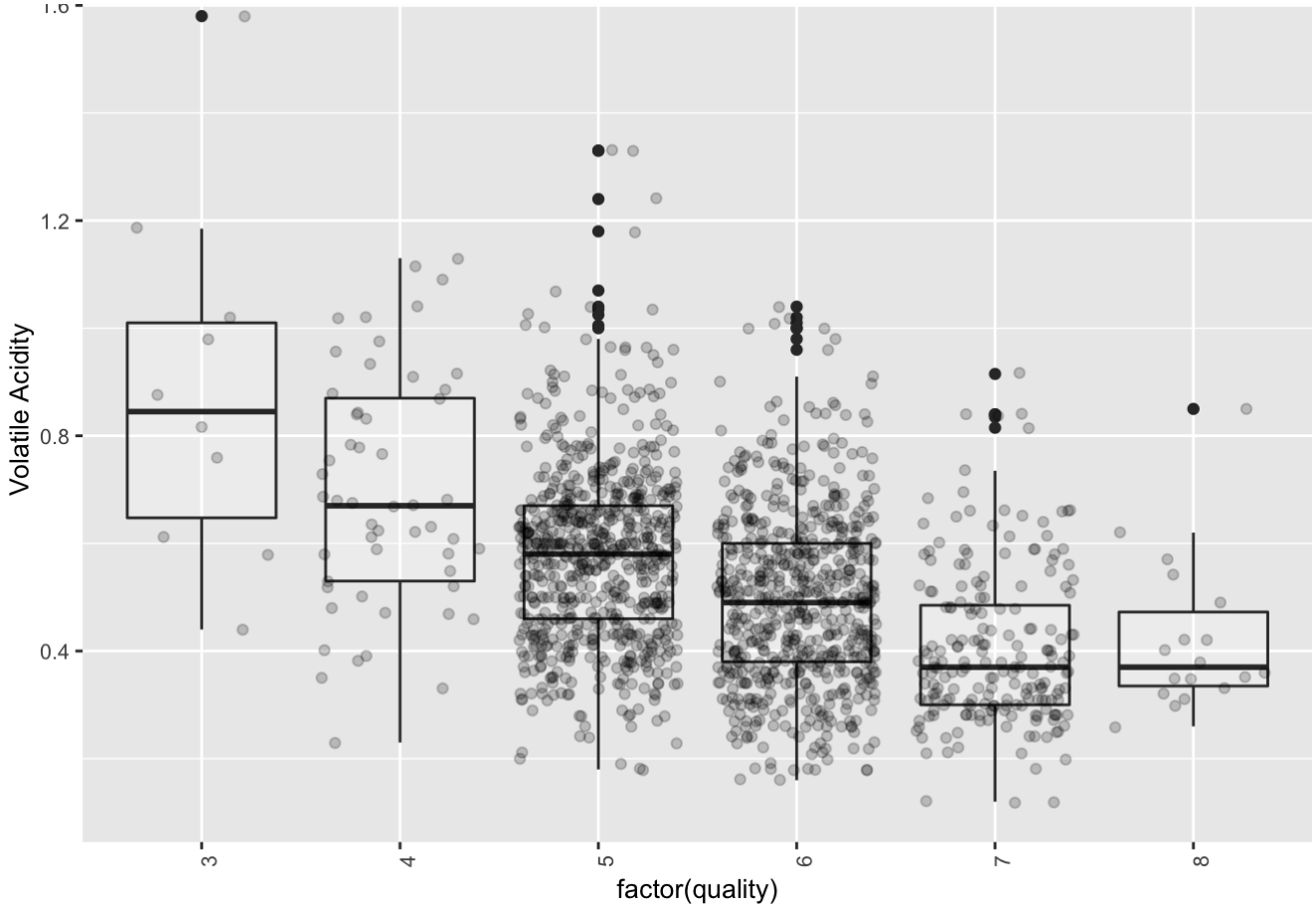
Bivariate plot section

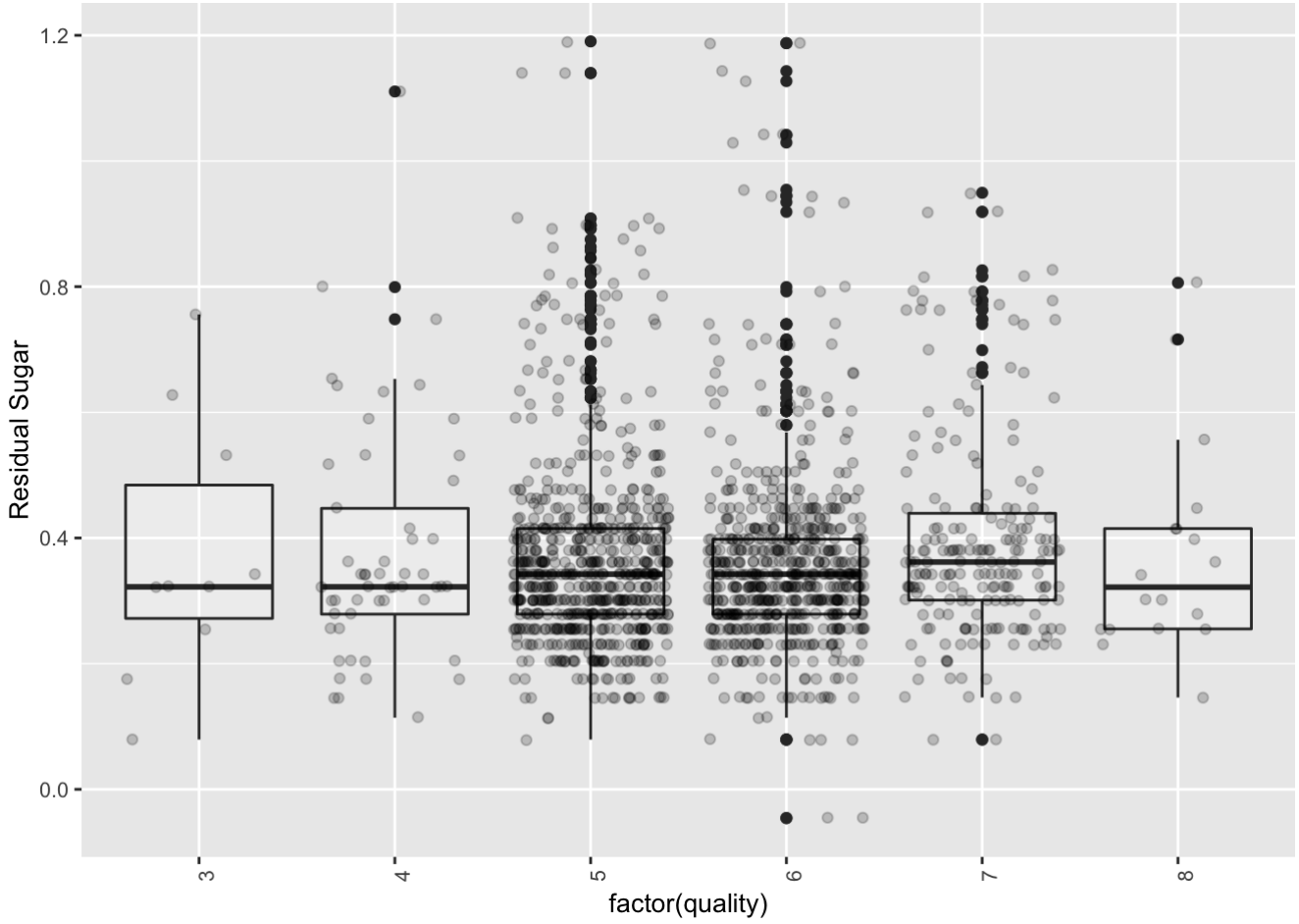
To get an idea of how the variables affect the quality of the red wine, I decided to create boxplots for all the features. This will also help me to identify the main variables that lead to a better quality wine along with the outliers in the dataset. I used boxplots as bivariate boxplots, with x-axis as rating or quality, will be more interesting in showing trends with wine quality.

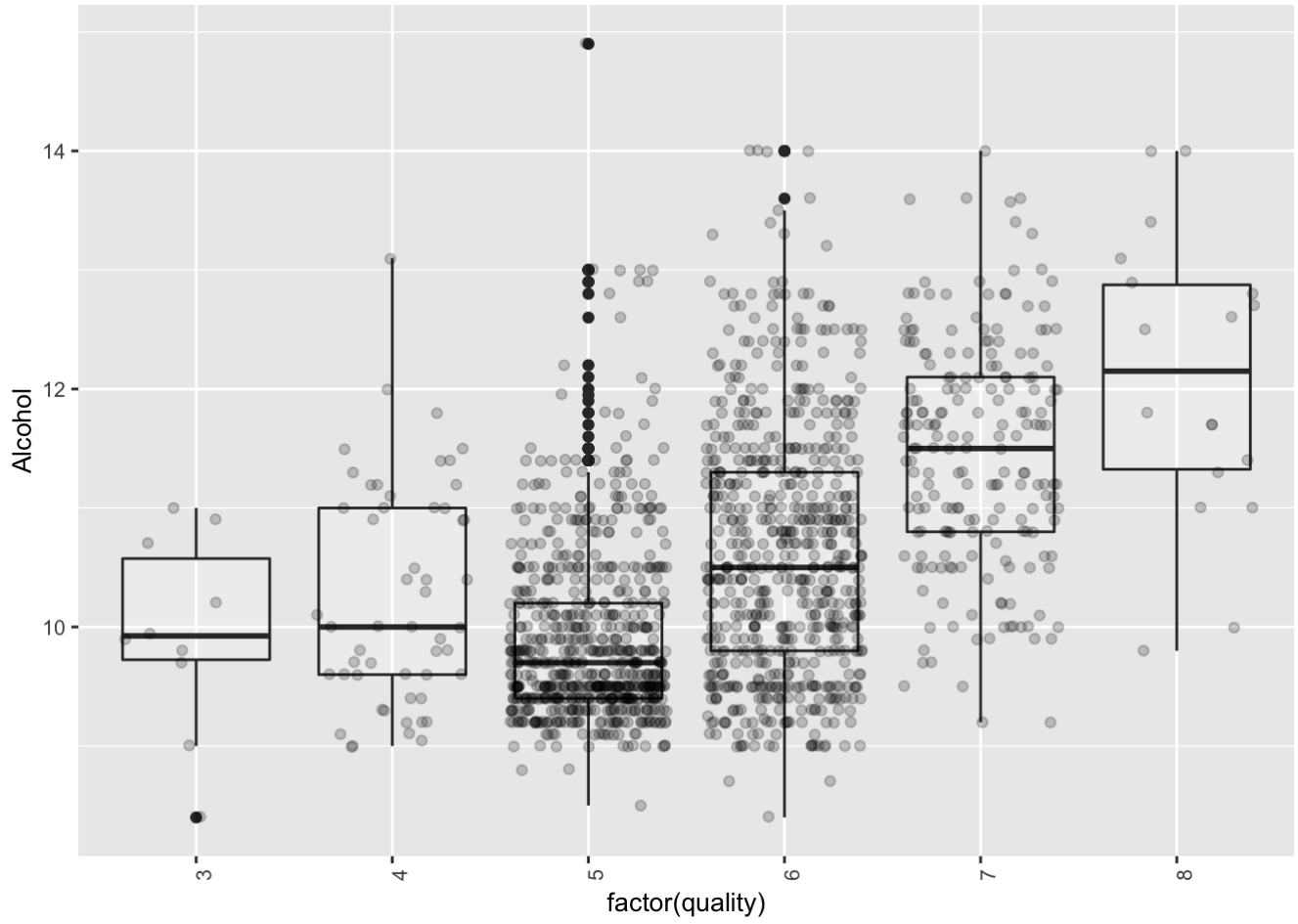
I assumed high quantity of all acids(fixed, volatile and citric acids) lead to better quality wine.

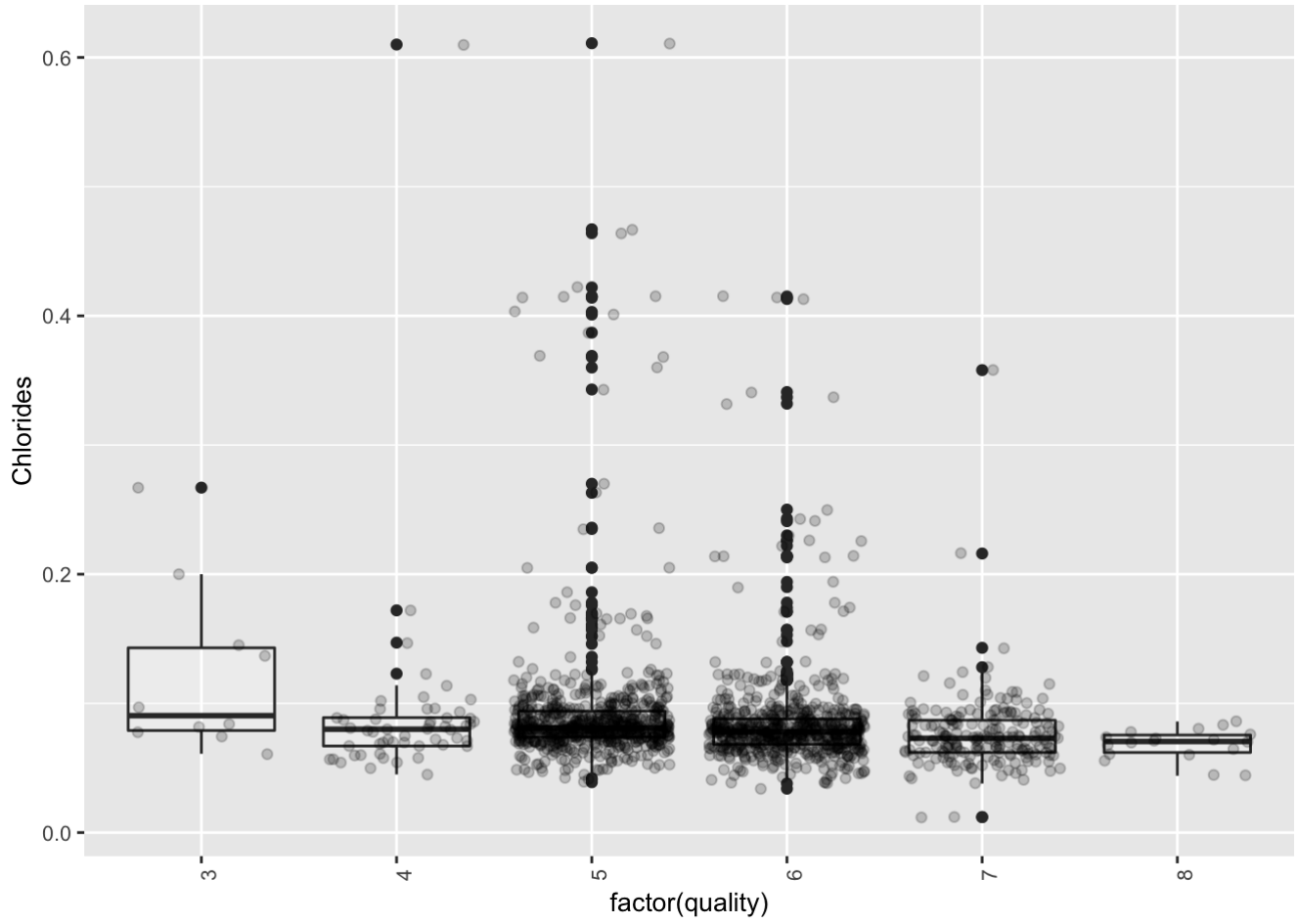

1.00
0.75
0.50
0.25
0.00
Citric Acid
3
4
5
6
7
8
factor(quality)

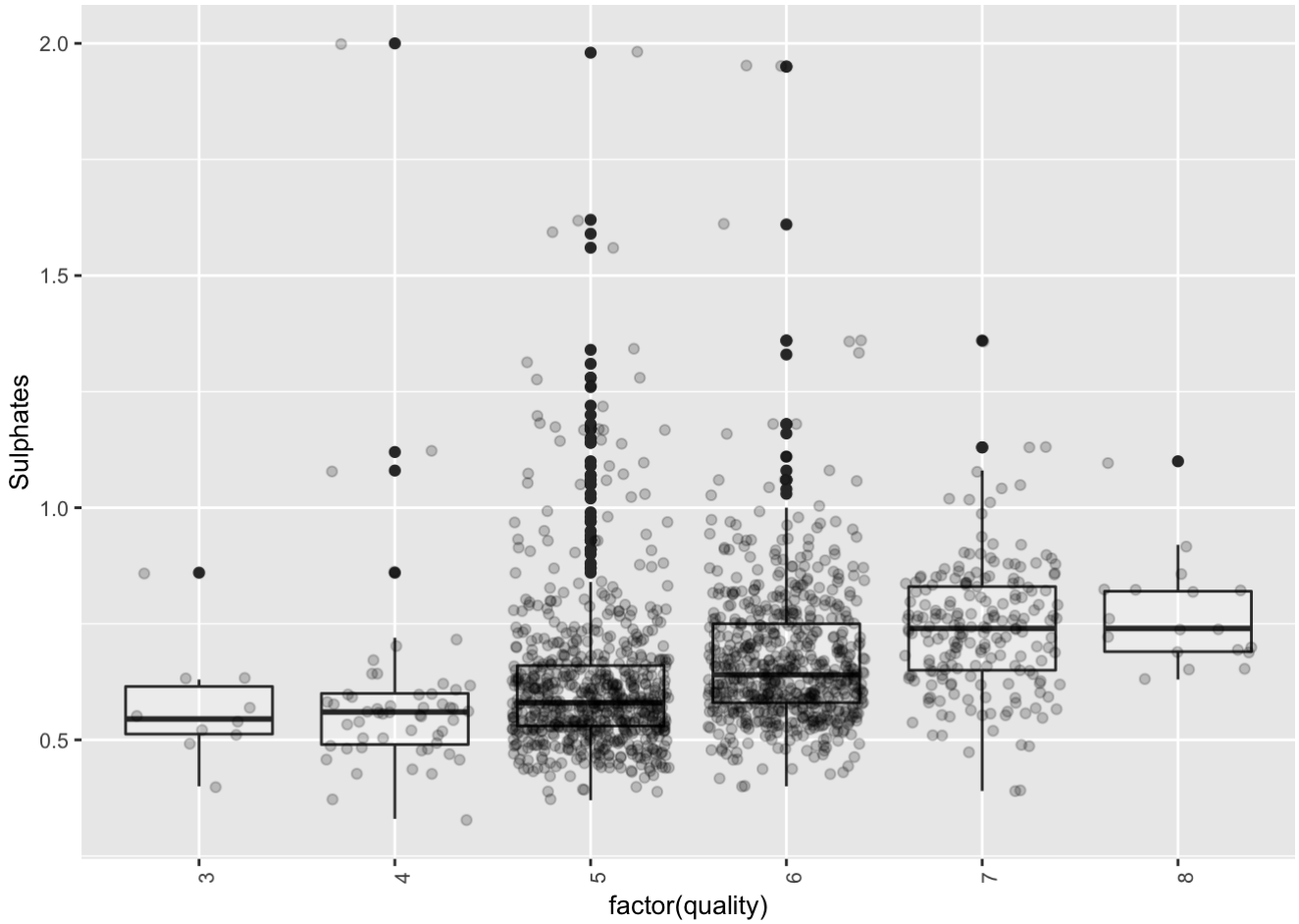

16
12
8
Fixed Acidity
3
4
5
6
7
8
factor(quality)

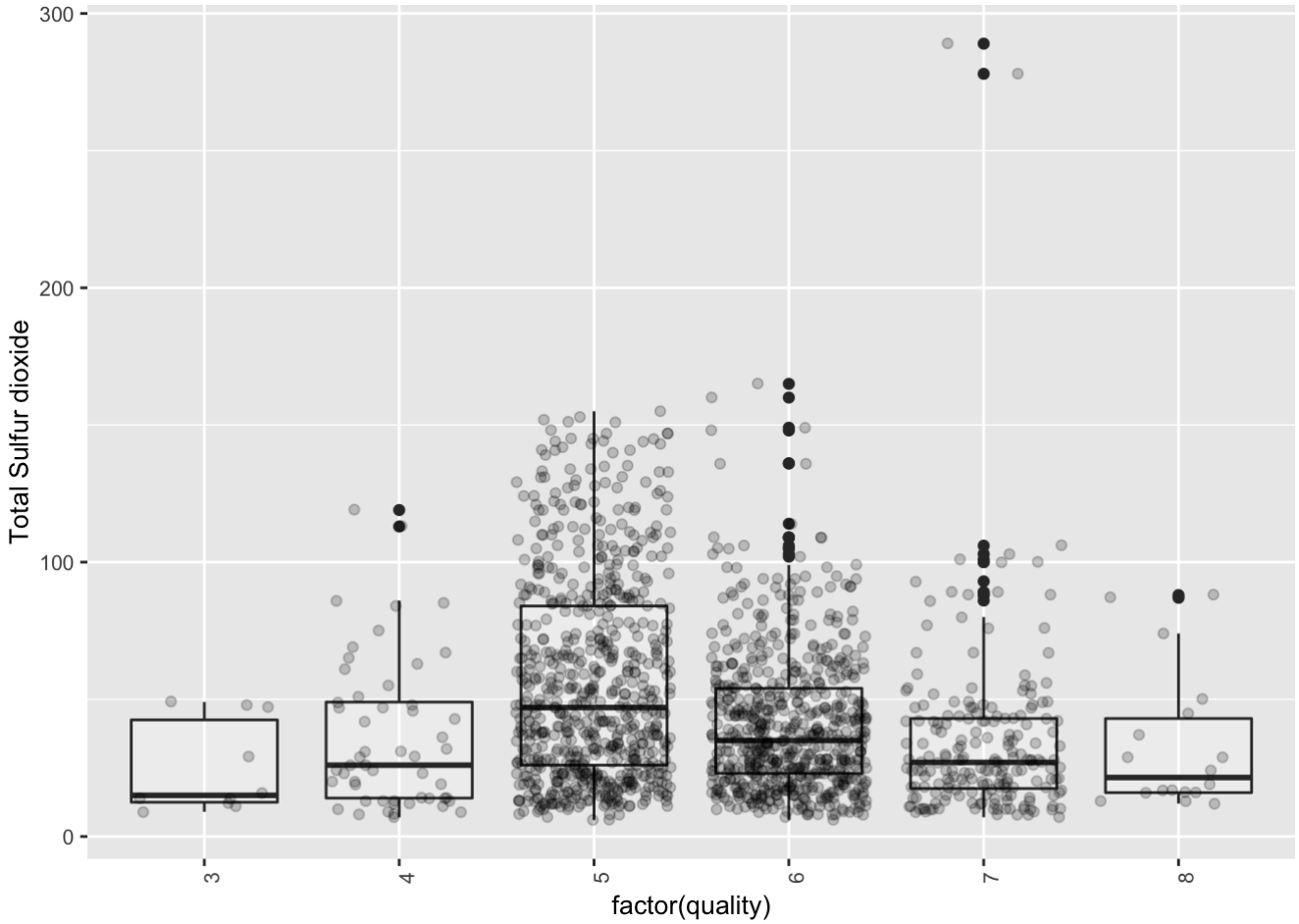

Volatile Acidity
1.2
0.8
0.4
3
4
5
6
7
8
factor(quality)

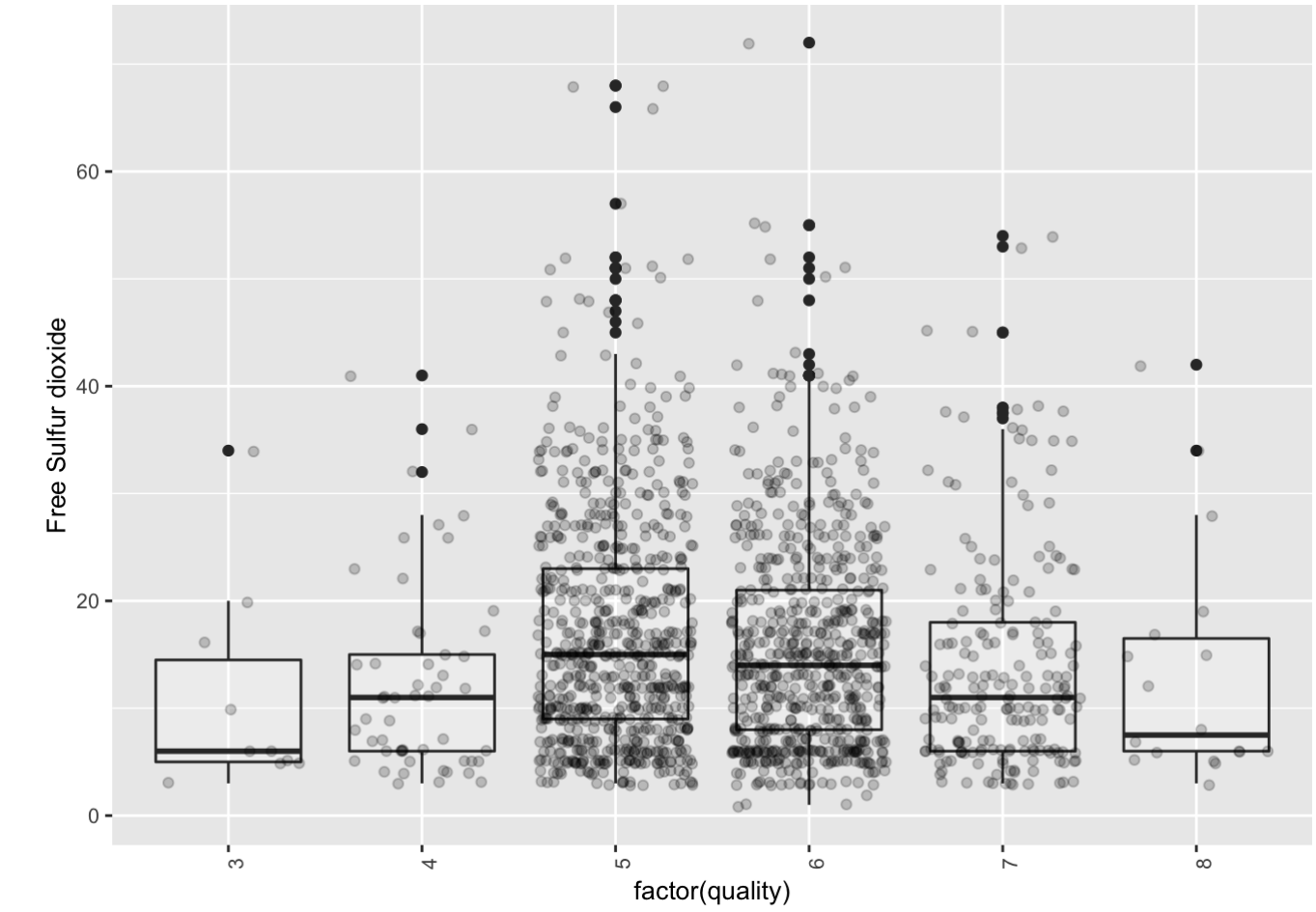

Residual Sugar
1.2
0.8
0.4
0.0
3
4
5
6
7
8
factor(quality)

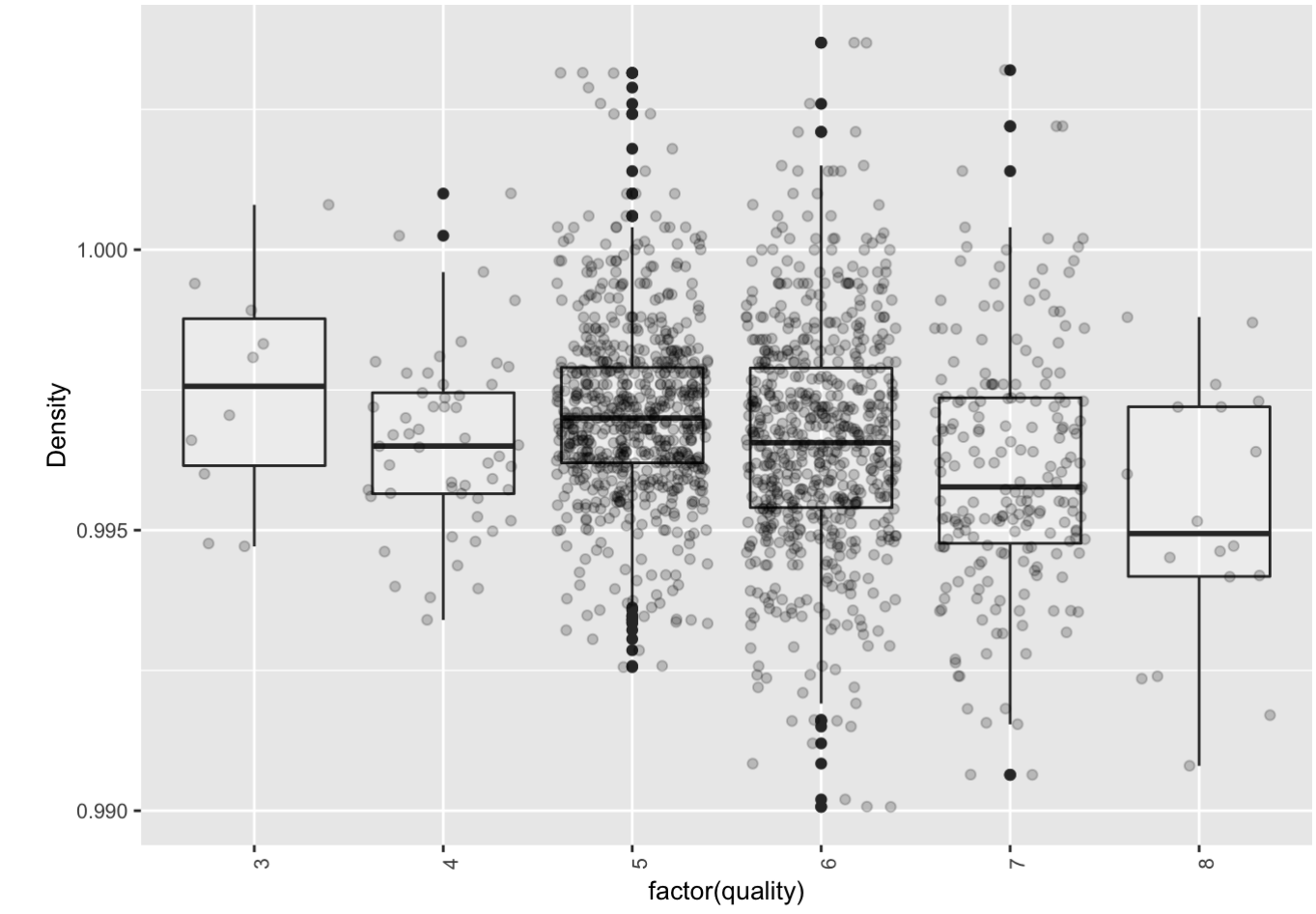

Alcohol
14
12
10
3
4
5
6
7
8
factor(quality)

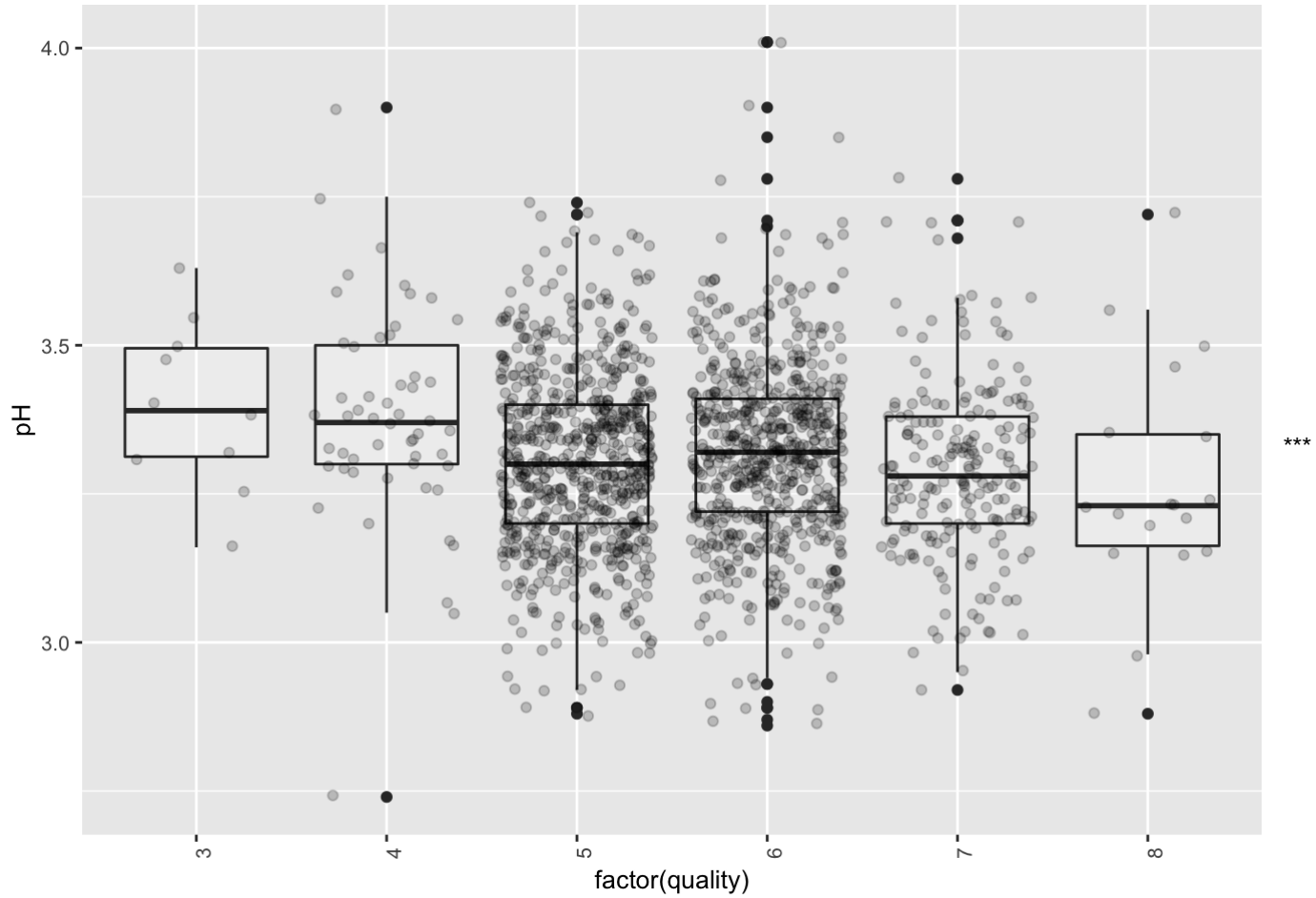

Chlorides
0.6
0.4
0.2
0.0
3
4
5
6
7
8
factor(quality)


Sulphates
2.0
1.5
1.0
0.5
3
4
5
6
7
8
factor(quality)


Total Sulfur dioxide
300
200
100
0
3
4
5
6
7
8
factor(quality)


Free Sulfur dioxide
60
40
20
0
3
4
5
6
7
8
factor(quality)


Density
1.000
0.995
0.990
3
4
5
6
7
8
factor(quality)



On plotting the boxplots, I could see that the chlorides, residual sugar and the sulfur dioxides do not have significant impact on the quality.

Between citric acid and quality, the points are not closely clustered but the gradient is positive. But the quanity of citric acid is more for quality 5 and 6(medium score). There are quite a few outliers as well in the plot.

Similar to citric acid, the fixed acids also showed a similar trend as the points are more clustered together for the quality 5 and 6.

The volatile acidity showed a negative gradient with the levels decreasing with the increase in quality which also opposed to my assumption that higher quantity of all acids lead to better wine quality.

The alcohol also showed a positive gradient with the most number of points clustered in quality 5.

The density showed a negative gradient as the quality increased. Therefore, I concluded that a good quality red wine consists of the following features:

High fixed acidity and citric acid, Low volatile acidity Low pH High alcohol Low density

To further materialize my conclusion, I decided to check the strength of the relations of the features with the quality via calculating the correlations.

Correlation between the wine features and quality

Fixed acidity and quality

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

Volatile acidity and quality

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

Suphates and quality

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

Citric acid and quality

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

Alcohol and quality

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

Density and quality

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

Residual sugar and quality

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

Chlorides and quality

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

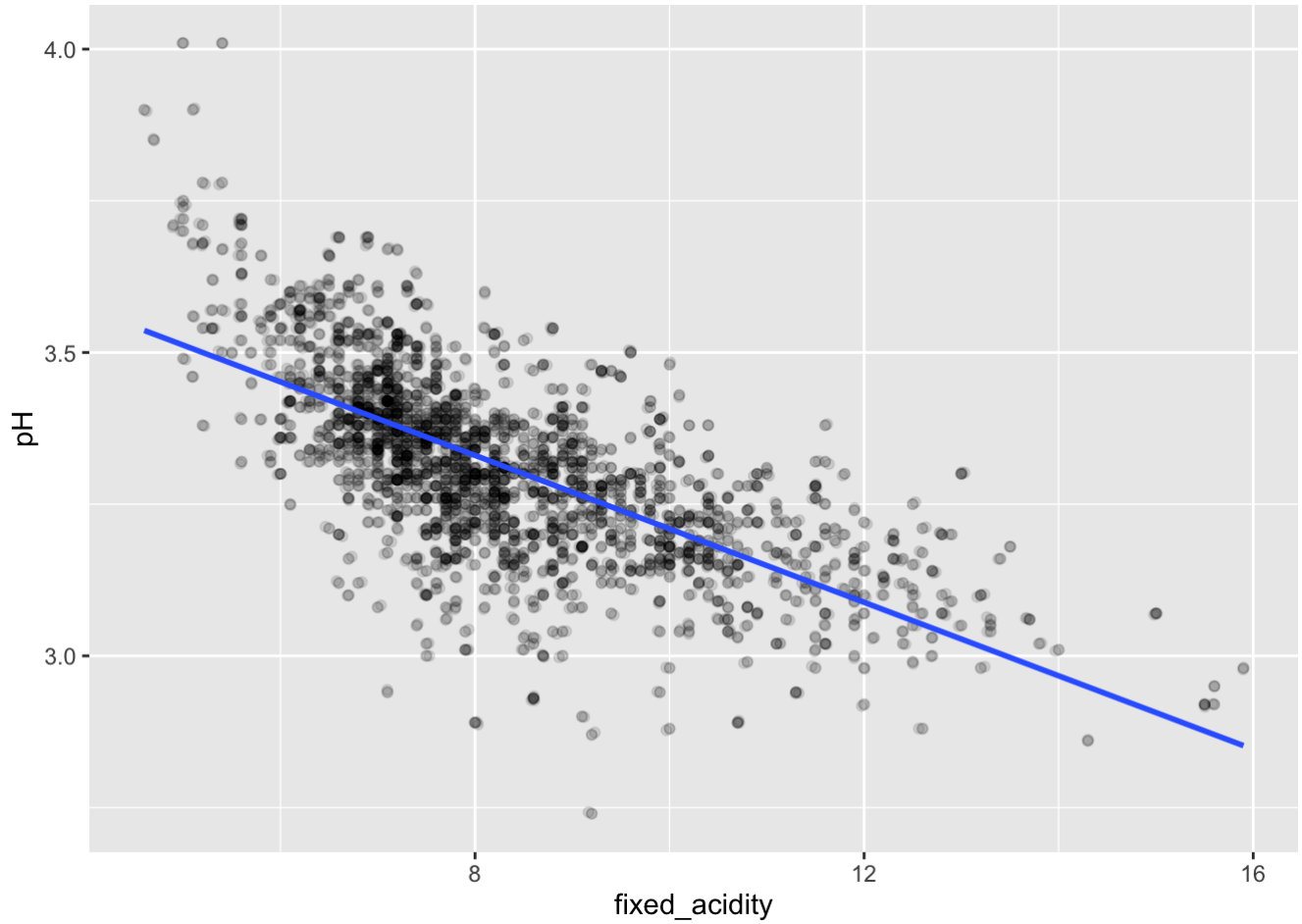
pH and quality

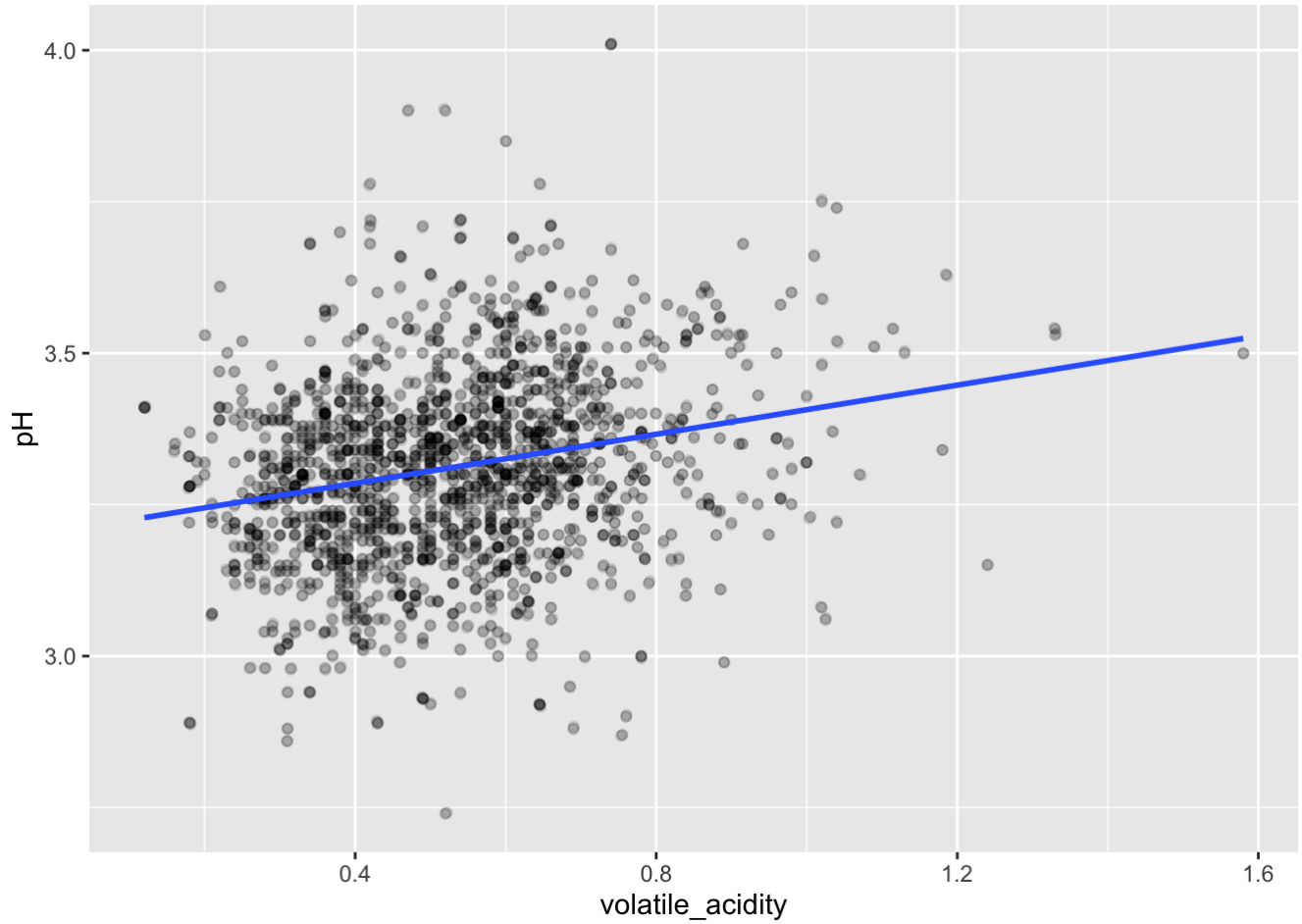
```
## [1] -0.05773139
```

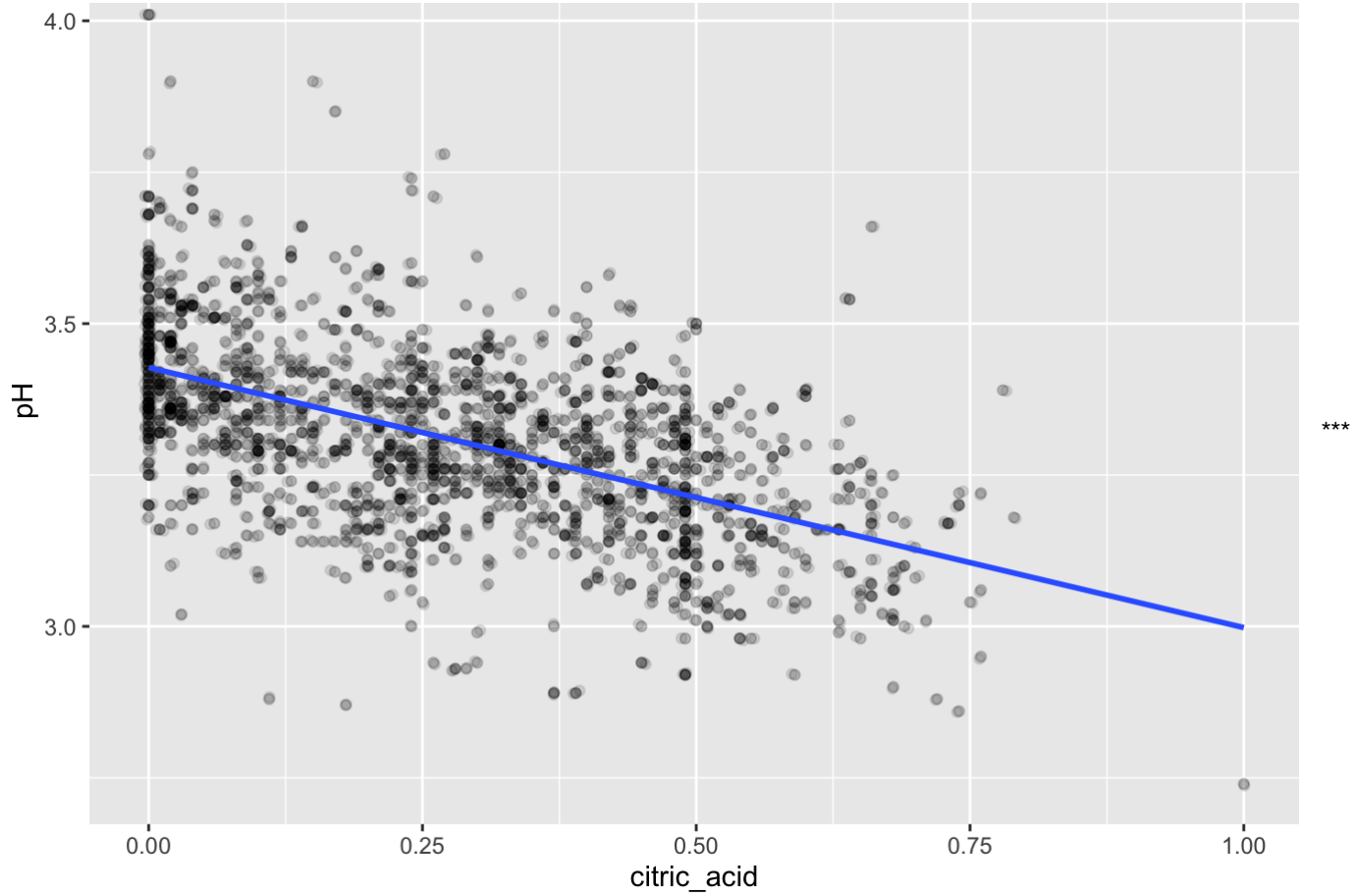
Therefore, the following features had the highest relation with quality:

alcohol sulphates volatile acidity citric acid

pH decreased with increase in quality. Therefore, I decided to check for the relation between pH and the acids via scatterplots


4.0
3.5
3.0
pH
8
12
16
fixed_acidity

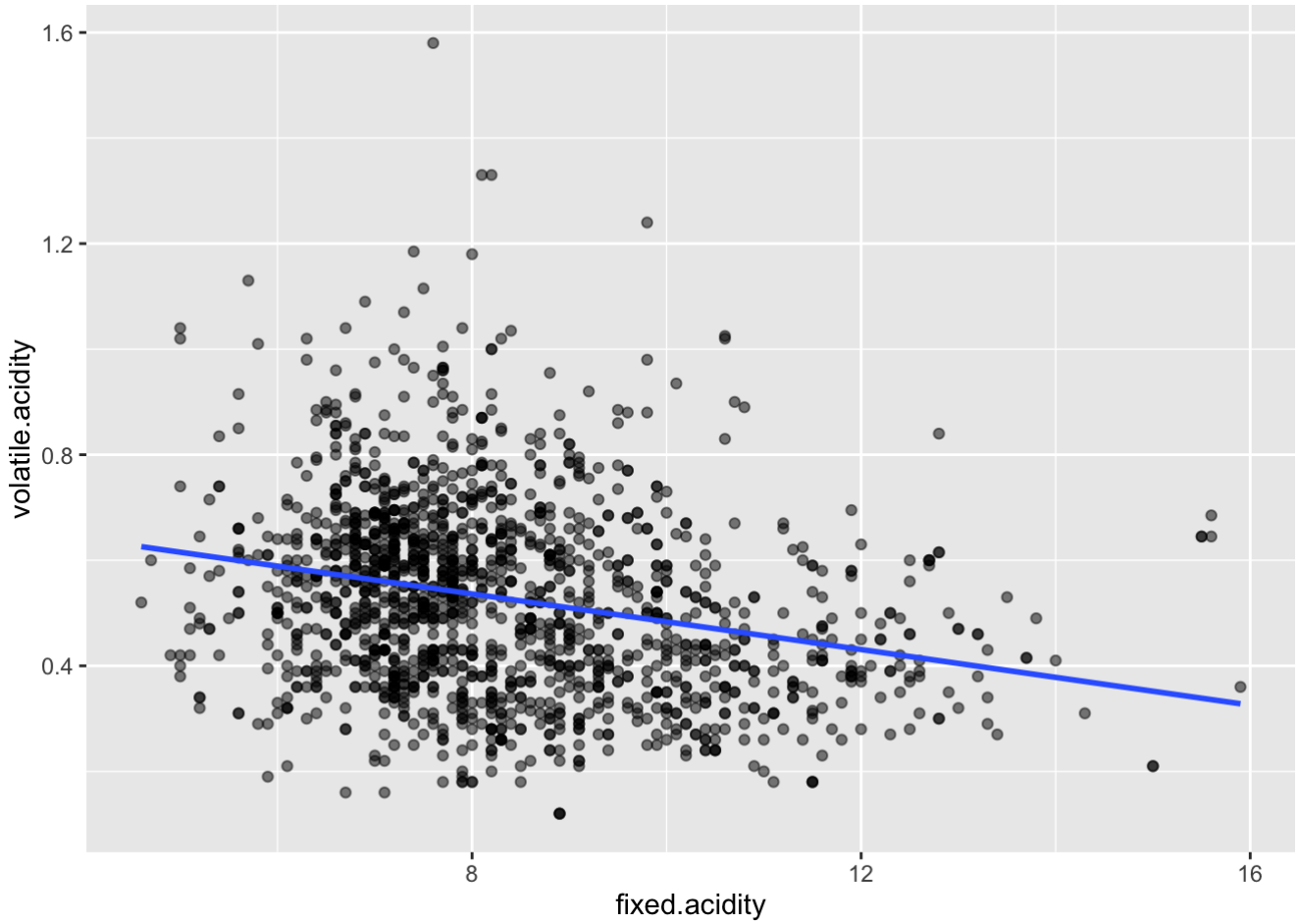

4.0
3.5
3.0
pH
0.4
0.8
1.2
1.6
volatile_acidity



On plotting the scatterplot, the pH level decreases with increase in acidity. This is quite clear, since less pH level leads to more acidic level.

But the pH with volatile acids show positive gradient. On doing some research, the best quality red wine always have less level of volatile acid also known as acetic acid as increase in levels lead to wine faults and turns the wine taste to more of a vinegar taste.

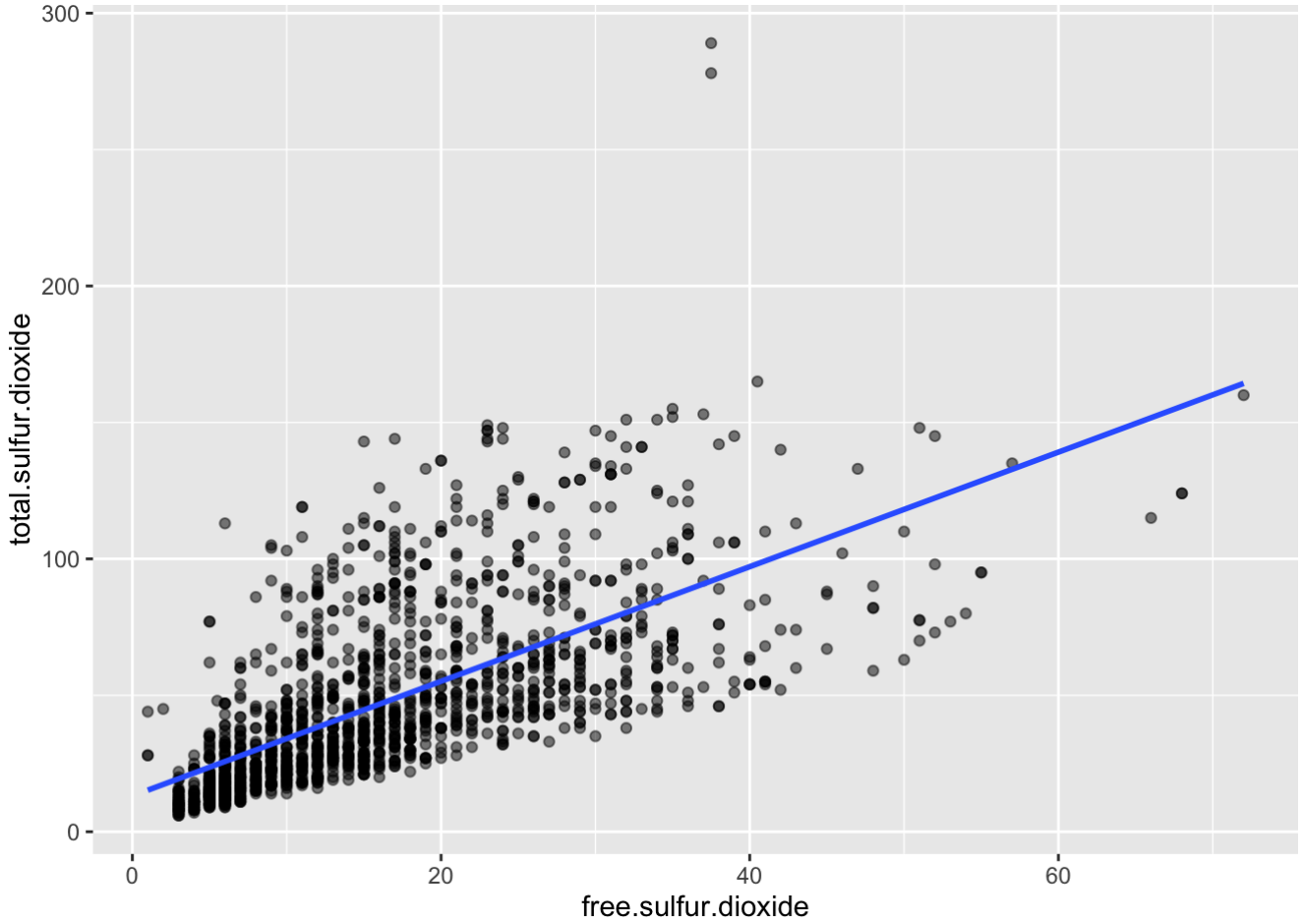
Again I assumed that the fixed and volatile acidity maybe correlated to each other. Therefore, to check the dependency and correlation, I plotted the scatterplot and calculated the correlation between them:




```
##
##  Pearson's product-moment correlation
##
## data:  rf$fixed.acidity and rf$volatile.acidity
## t = -10.589, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3013681 -0.2097433
## sample estimates:
##        cor
## -0.2561309
```

They seem to be quite correlated to each other although I woudn't say its a strong correlation.

Likewise, I decided to explore the correlation between total and free sulfur dioxide:



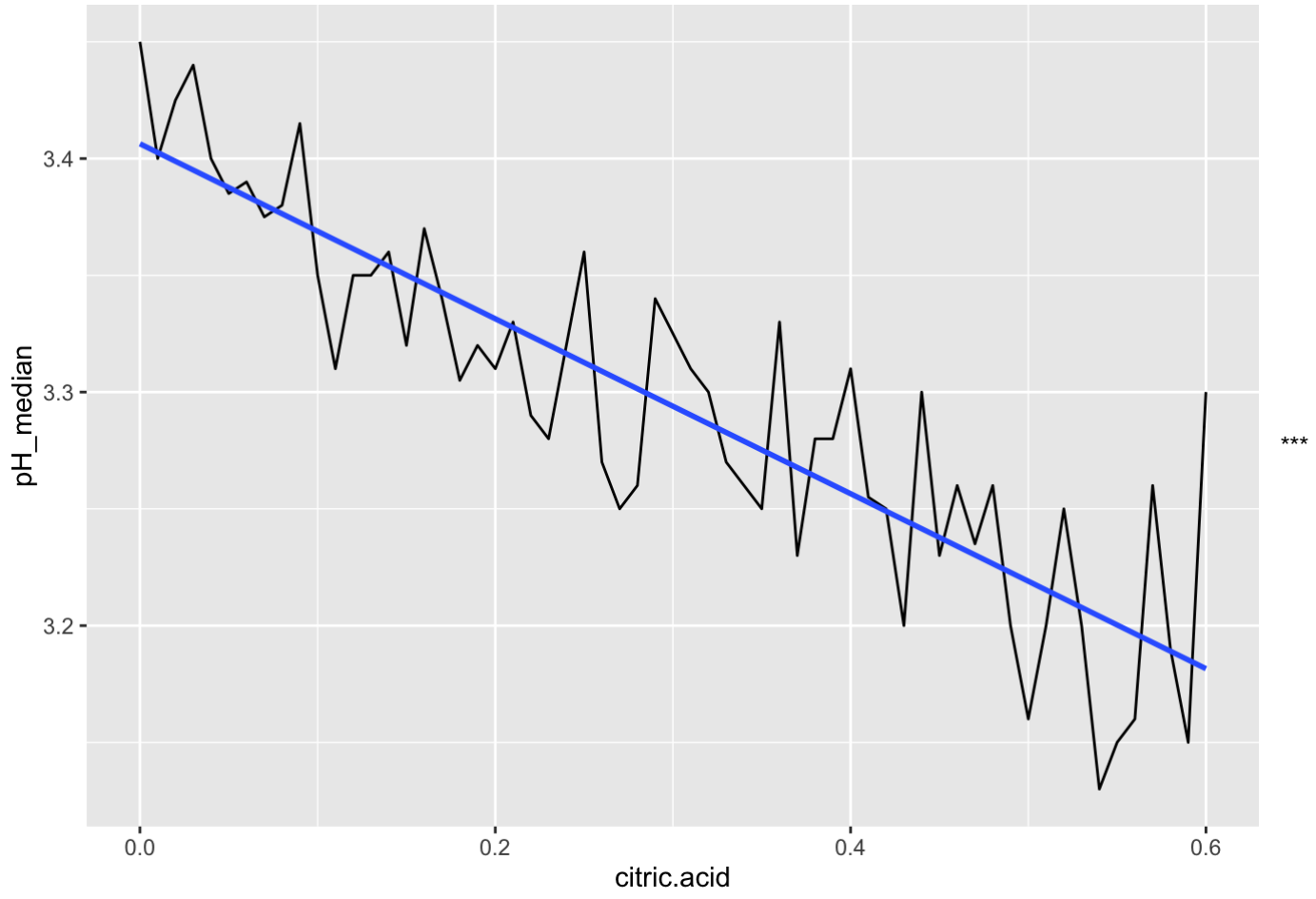

```
##
##  Pearson's product-moment correlation
##
## data:  rf$free.sulfur.dioxide and rf$total.sulfur.dioxide
## t = 35.84, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6395786 0.6939740
## sample estimates:
## 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

This showed a strong relationship between free and total sulfur dioxide, and thus I concluded that free and total sulfur dioxide are dependent on each other. ***

Conditional summaries to calculate the average pH level for the citric acids and plotting them

While checking the relationship between pH and acids, I found that the pH level with citric acid has the strongest relation. Therefore, I decided to calculate the conditional summaries as to how much average pH is present in the citric acid observations and plotted the pH median in a line graph:

```
## Source: local data frame [20 x 4]
##
##    citric.acid  pH_mean pH_median     n
##          <dbl>    <dbl>     <dbl> <int>
## 1         0.00 3.456212     3.450   132
## 2         0.01 3.427576     3.400    33
## 3         0.02 3.429400     3.425    50
## 4         0.03 3.427333     3.440    30
## 5         0.04 3.430000     3.400    29
## 6         0.05 3.389000     3.385    20
## 7         0.06 3.425000     3.390    24
## 8         0.07 3.370000     3.375    22
## 9         0.08 3.376364     3.380    33
## 10        0.09 3.412000     3.415    30
## 11        0.10 3.362571     3.350    35
## 12        0.11 3.306000     3.310    15
## 13        0.12 3.344444     3.350    27
## 14        0.13 3.373889     3.350    18
## 15        0.14 3.376667     3.360    21
## 16        0.15 3.338421     3.320    19
## 17        0.16 3.340000     3.370     9
## 18        0.17 3.367500     3.340    16
## 19        0.18 3.313182     3.305    22
## 20        0.19 3.314762     3.320    21
```



I found that the relationship is almost linear and clearly the pH decreases with the increase in acidity. But 132 number of observations have a citric acid value of 0. The mean and median of pH is constant between 3-3.5 and doesnot show any surprising patterns

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

- The fixed acidity is strongly correlated to pH whereas th volatile acidity weakly correlates to pH.
- Again volatile acidity has a stronger correlation with alcohol than fixed
- Citric acid has a strong correlation with pH but again weak correlation with alcohol
- As the amount of suphates increases, the variance in quality also increases.
- volatile.acidity, sulphates, citric.acid, alcohol have strong correlations with quality
- residual.sugar, chlorides and sulfur dioxides have no impact on the quality.
- The correlation between pH & citric acid and pH & fixed acidity is high which makes sense.

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

- Density and quality are weakly correlated to each other and better quality wine has less density
- Again density and alcohol are strongly related to each other

What was the strongest relationship you found?

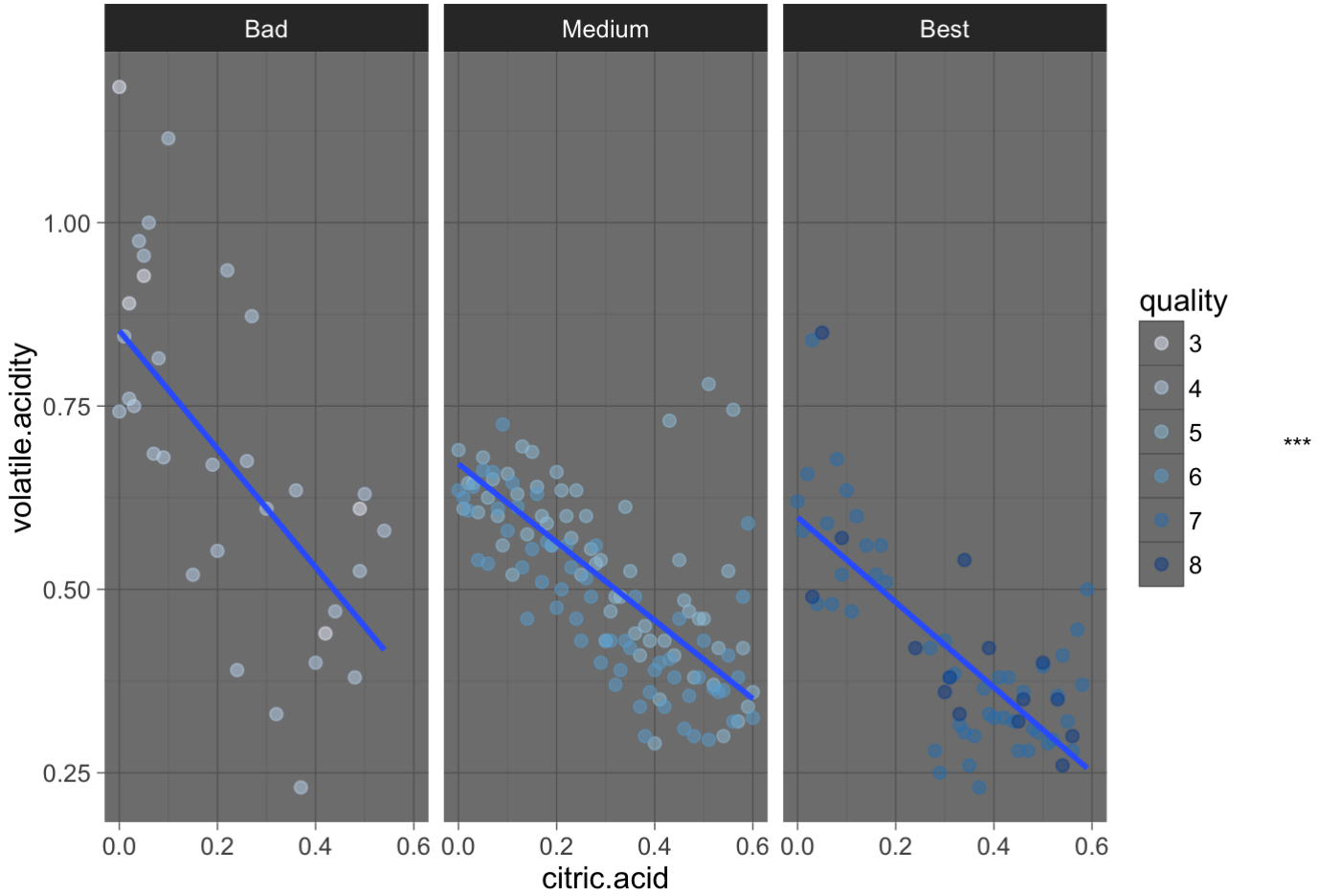
- The strongest relationship is between fixed.acidity and citric.acid

Multivariate plot section:

Analyzing the pattern between the various features that are of central importance and third qualitative variable score_range:

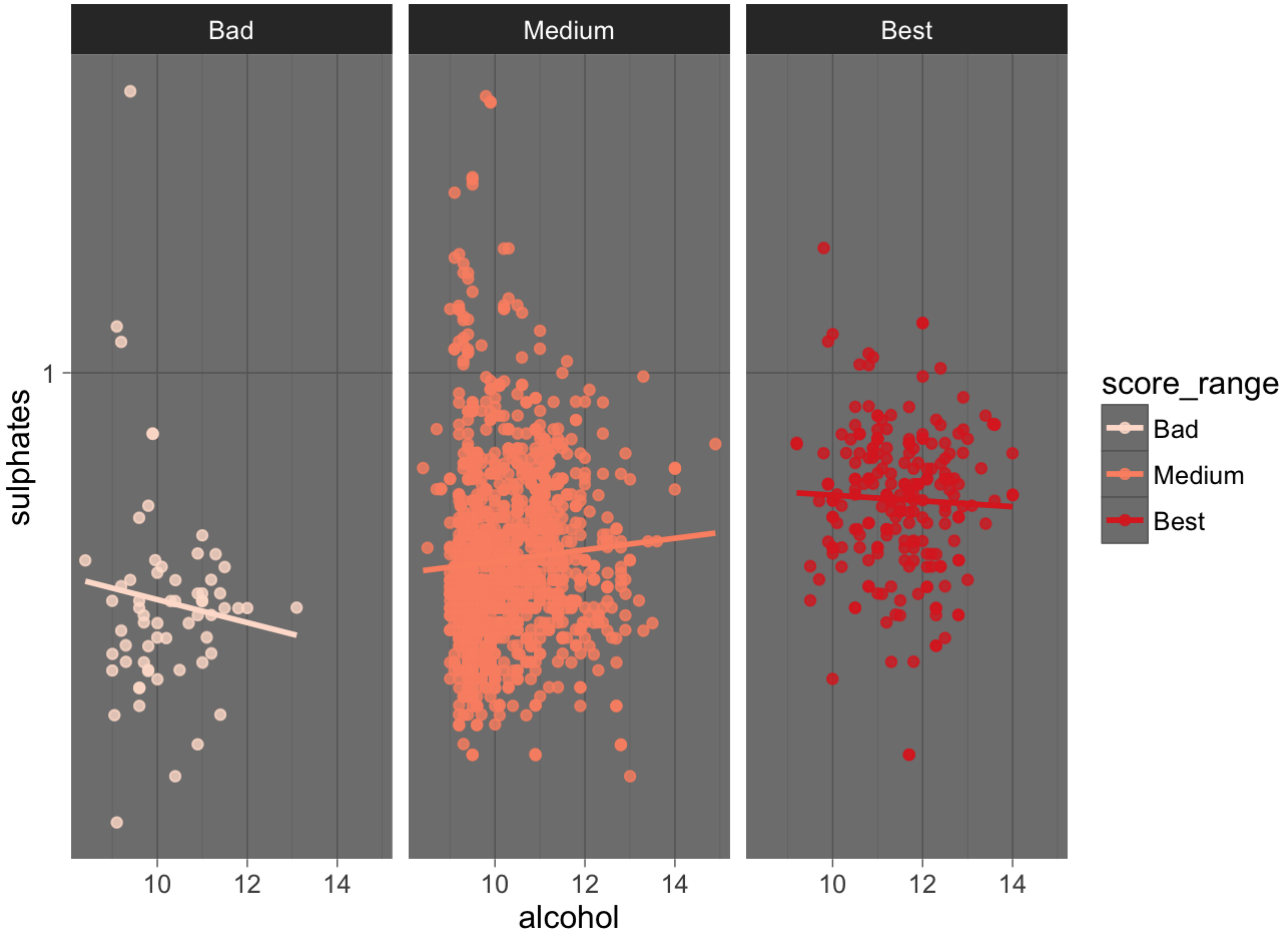
I decided to plot the data to check the relationship between the variables which are of importance like acids, alcohol, pH and sulphates to determine the quality of red wine using quality as a third variable faceted by score_range .

Acids with quality

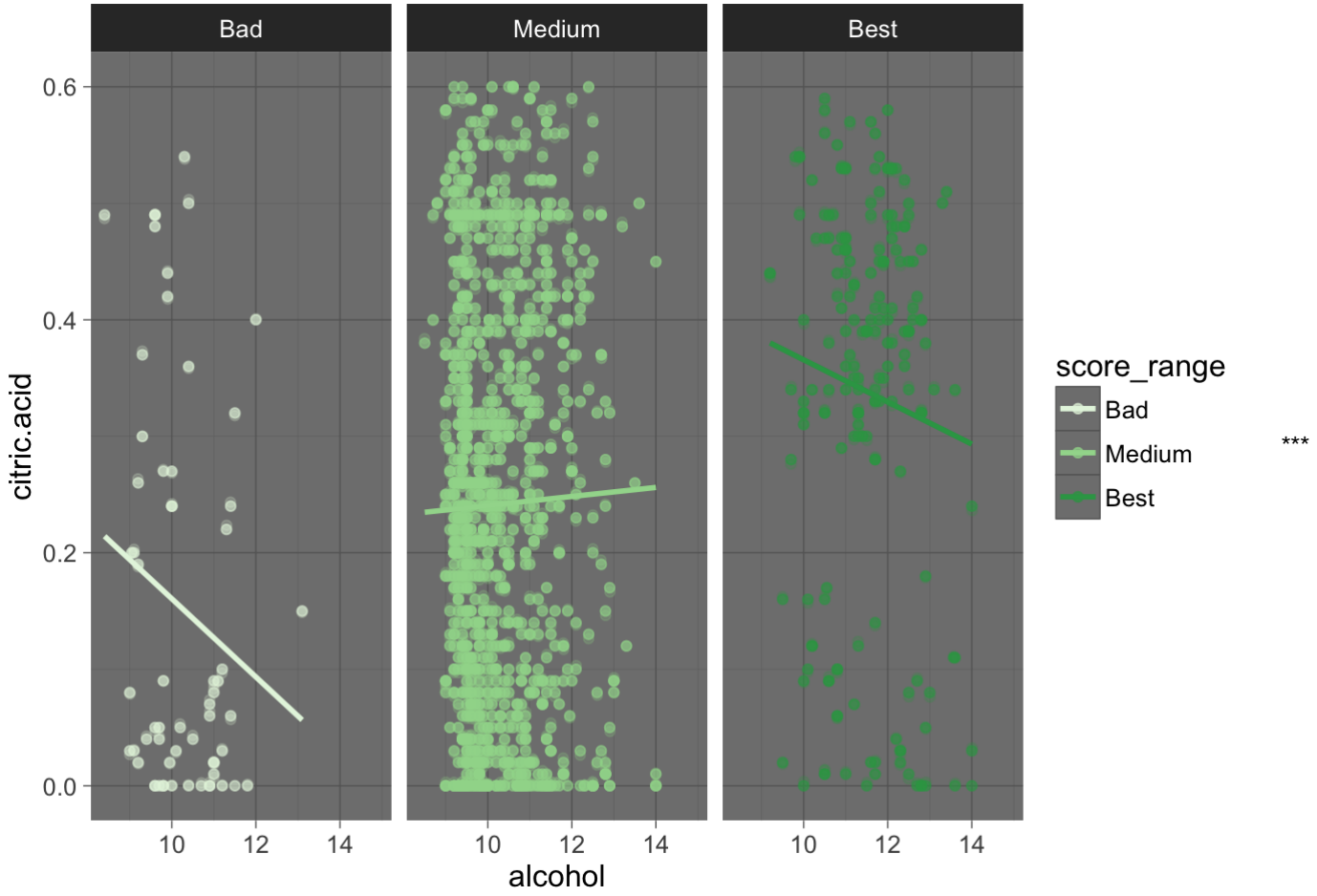


I used scatterplots to see the distribution and since the points were too much scattered, so I faceted it by score_range to have a clear visualization. Since 132 observations of citric.acid have 0 values, so I omitted the initial 5% of citric acid.

Alcohol and sulphates with quality

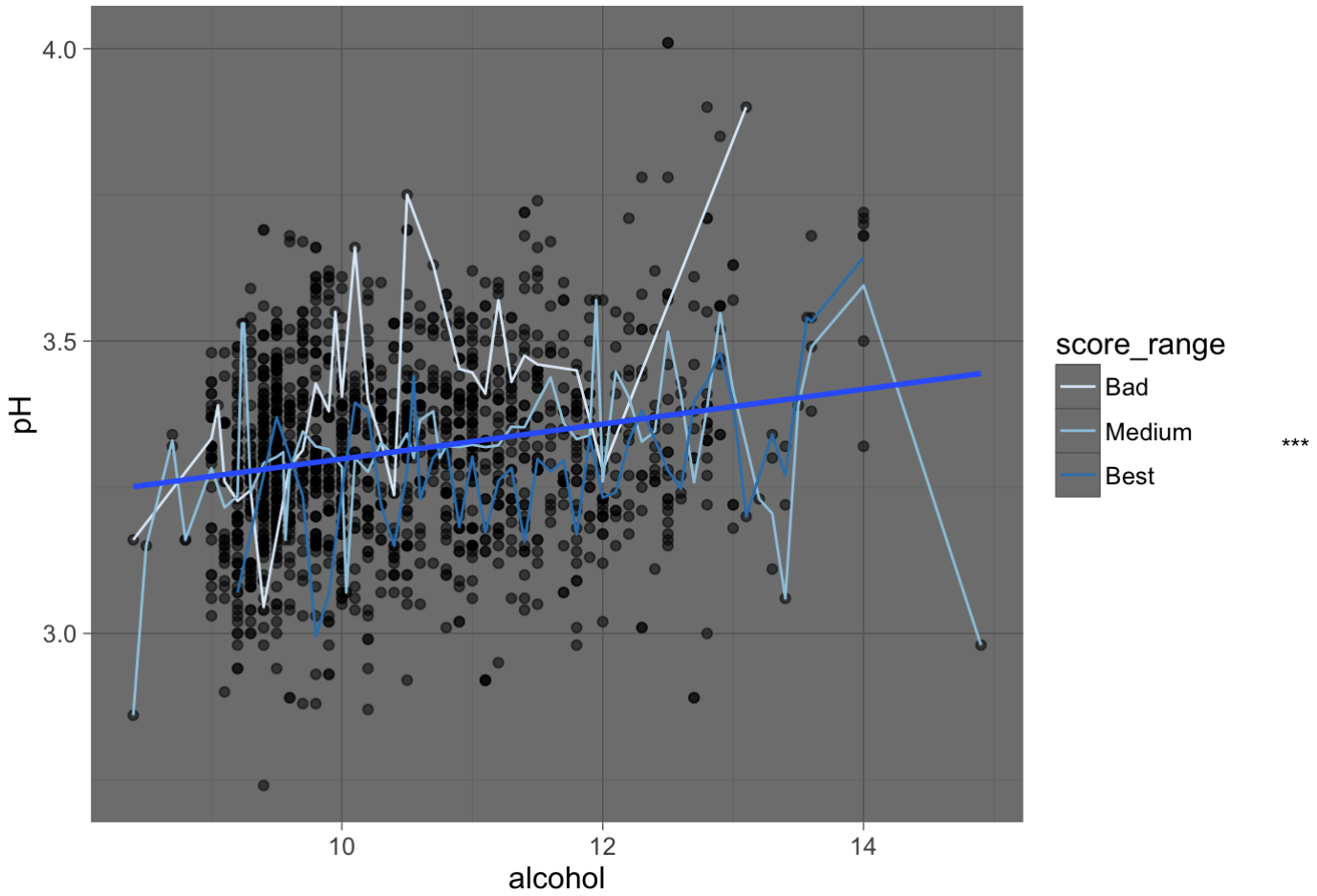


Alcohol and citric acid



Alcohol and pH with quality

Here, I used line plots as I found the line plots to be more precise in showing the changes in trends as it connects the points. Therefore, an overall pattern can be seen. I overlayed the median summary of the data over the plot to see the quality by mean pH level in the alcohol



pH level has very less impact on the quaity of the wine.

Observations: On plotting the features against quality/score_range as a third variable, I see that sulphates do not have a dramatic impact on the quality, therefore I decided to consider acids and alcohol as the primary factors that lead to better wine quality. From the plots it was clear that higher citric acid and lower volatile acids lead to a better wine quality.

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?

The most important features that lead to a good quality wine are: -Higher levels of citric acid, fixed acidity, suphates, alcohol and lower volatile acidity

Therefore, I observed the relation of each of these variables with each other and quality as a third qualitative variable

Were there any interesting or surprising interactions between features?

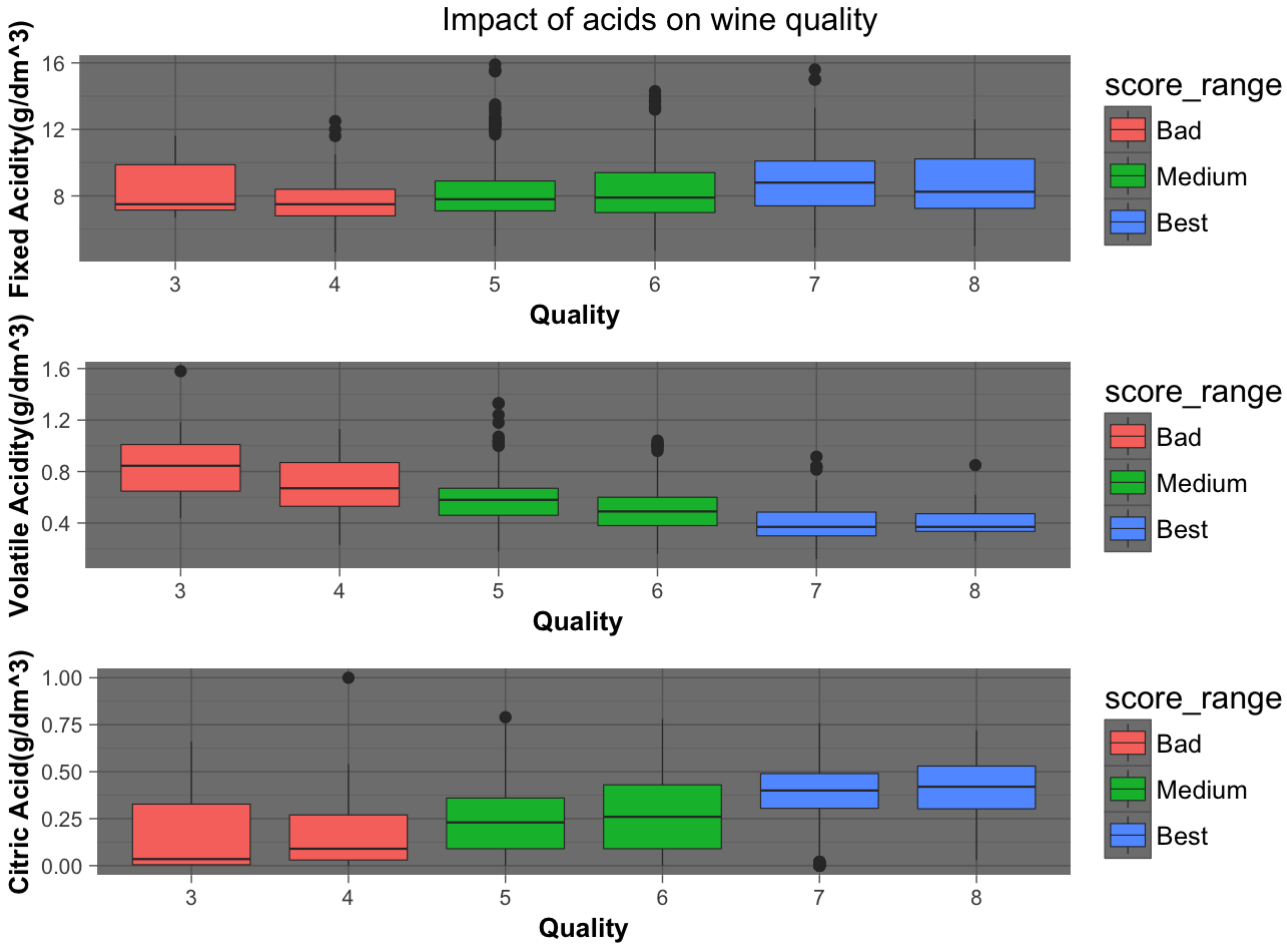
Yes, I found that density and quality of wine are not strongly correlated while density and alcohol are strongly correlated, therefore, I plotted to analyze the pattern with these three variables in my plot3.

Final Plots and Summary:

Since quality is a categorical variable, it is better to use boxplots and barplots to view the data

Plot 1:

Plotting the acidity and quality variables to find out the impact acidity has on the quality of wine



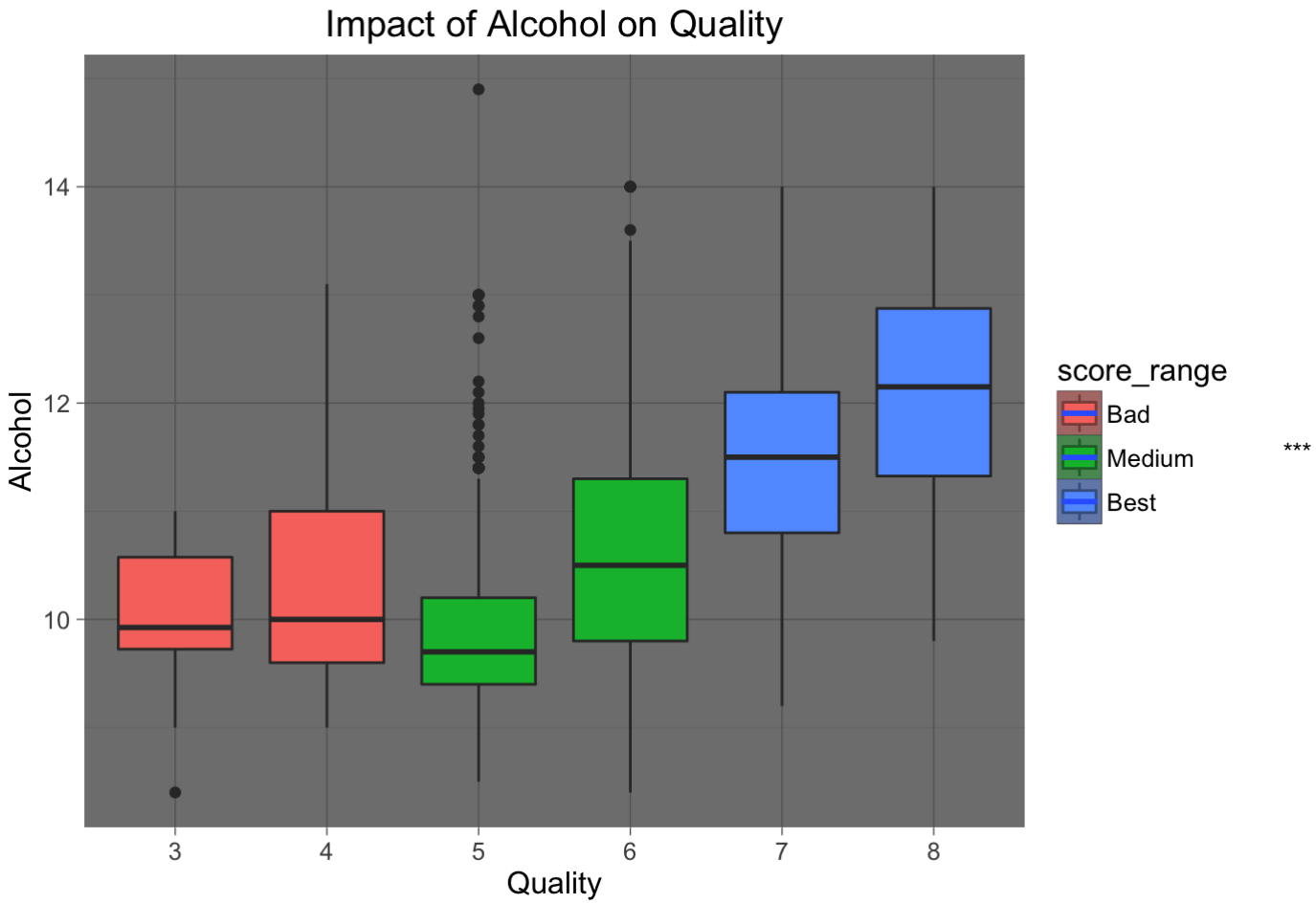
***

Description of the plot:

I created these plots to check the impact of acidity on the quality of wine. I found out these patterns in the plots: - Higher levels of citric and fixed acids led to better quality of wine - Lower levels of volatile acids led to better quality of wine

Plot 2:

Plotting the alcohol and quality variables to find out the impact acidity has on the quality of wine



Description:

-Higher levels of alcohol led to higher wine quality -The alcohol also has a strong correlation with the quality -The median of the alcohol is 10.20 for the medium quality -The alcohol and quality also has a strong correlation

Plot3:

We found out density and quality are weakly related and alcohol and density are strongly related. So, I plotted it to find out the patterns:


```
## NULL
```

Description:

- The best quality wine has more alcohol levels and less density(<1)
- Density cannot be taken as a confident factor for determining the quality as according to the plot even some of the best quality observations has more density and less alcohol level
- But on seeing this plot, it can be concluded that the better quality wine has less density

Reflection

By performing this exploratory data analysis on the data, I was able to find out patterns in the data that lead to good quality red wine.I faced difficulty in analyzing the citric.acid variable as most of the observations had 0 counts. So in further analysis, I took the 95% quantile in the data and omitted those observations with 0 values. I assumed fixed and volatile acidity to be strongly correlated but on further analysis my assumption did not yield the correct results. However, most of the data consists of medium quality observations. I would like to analyze that does always the bad quality have low density and other factors. Moreover, I would also like to analyze as to why a large number of observations have a 0 value in citric acid.