Red Wine Exploration by Serdar Ozsoy

This dataset is related with variants of the Portuguese “Vinho Verde” red wine. There are 1599 instances with 11 attributes and the output is based on sensory data (median of at least 3 evaluations made by wine experts).Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

Dataset is provided from here (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityWhites.csv).

Description can be read from here (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt).

Univariate Plots Section

```
## '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 0.073 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 ...
```

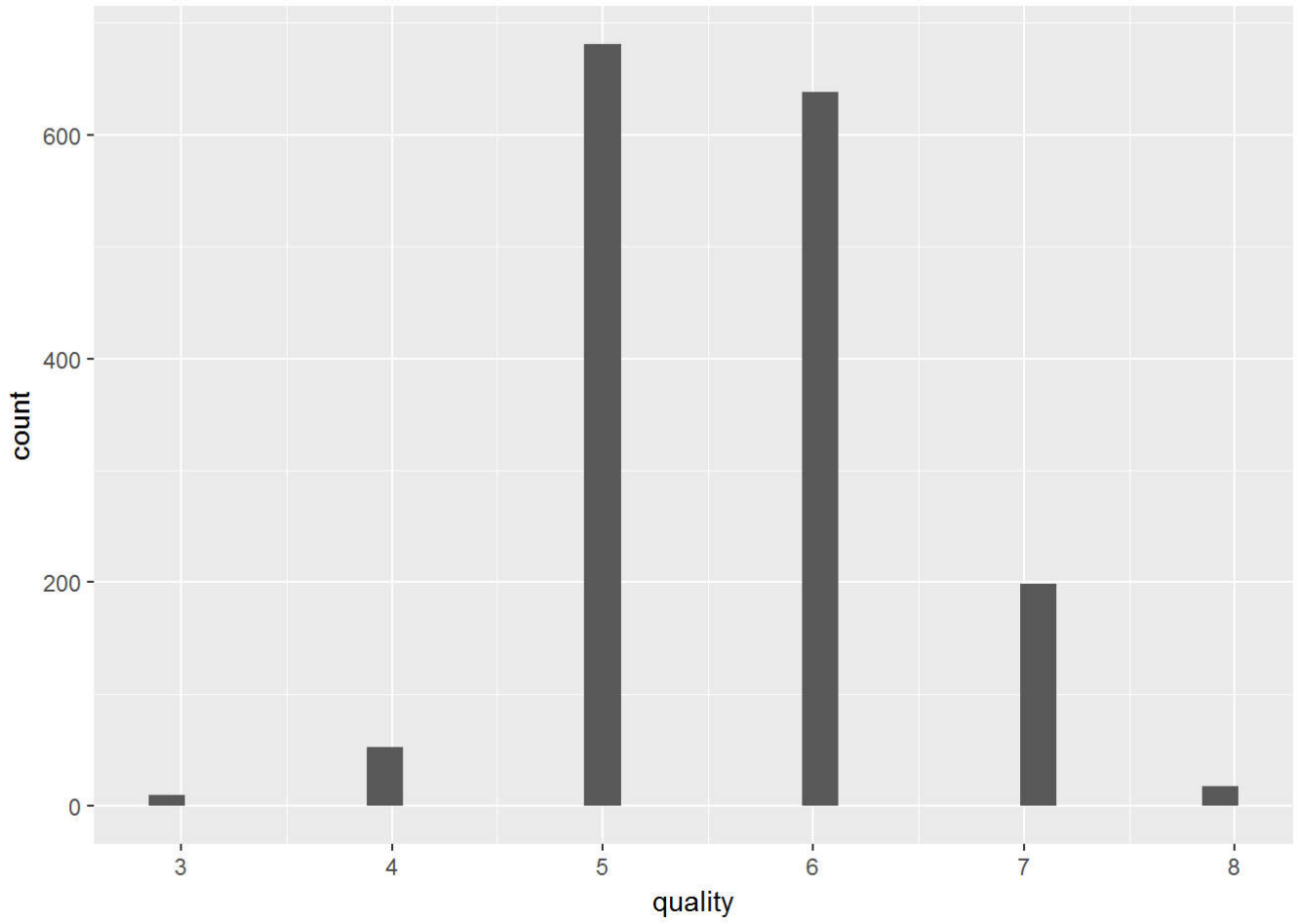
There are 1599 observations of 12 variables. Except quality, type of attributes are numerical. Quality variable is integer type.

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

When you look at whole, means and medians are usually so close for each. It gives us clue for normal distributions but not guareantee. Most suspicious variables which have not normal distribution are residual.sugar, chlorides, total.sulfur.dioxide and quality at first sight.

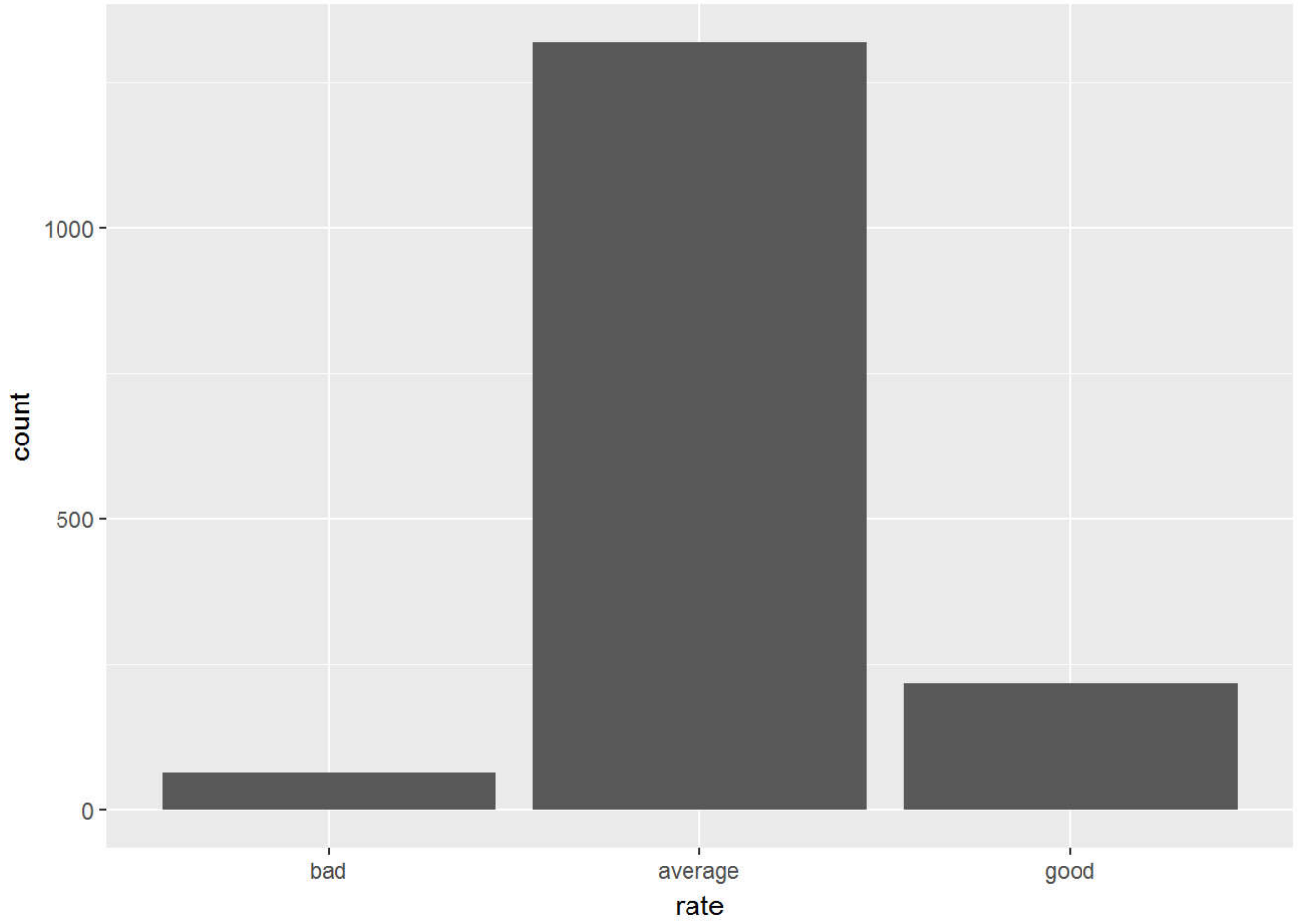
Quality

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



Most wines has 5 and 6 points, peaks at these points. Values are spread between 3 and 8, although scores can be between 1 and 10. It is not balanced and other scores than peak values may behave like outliers.

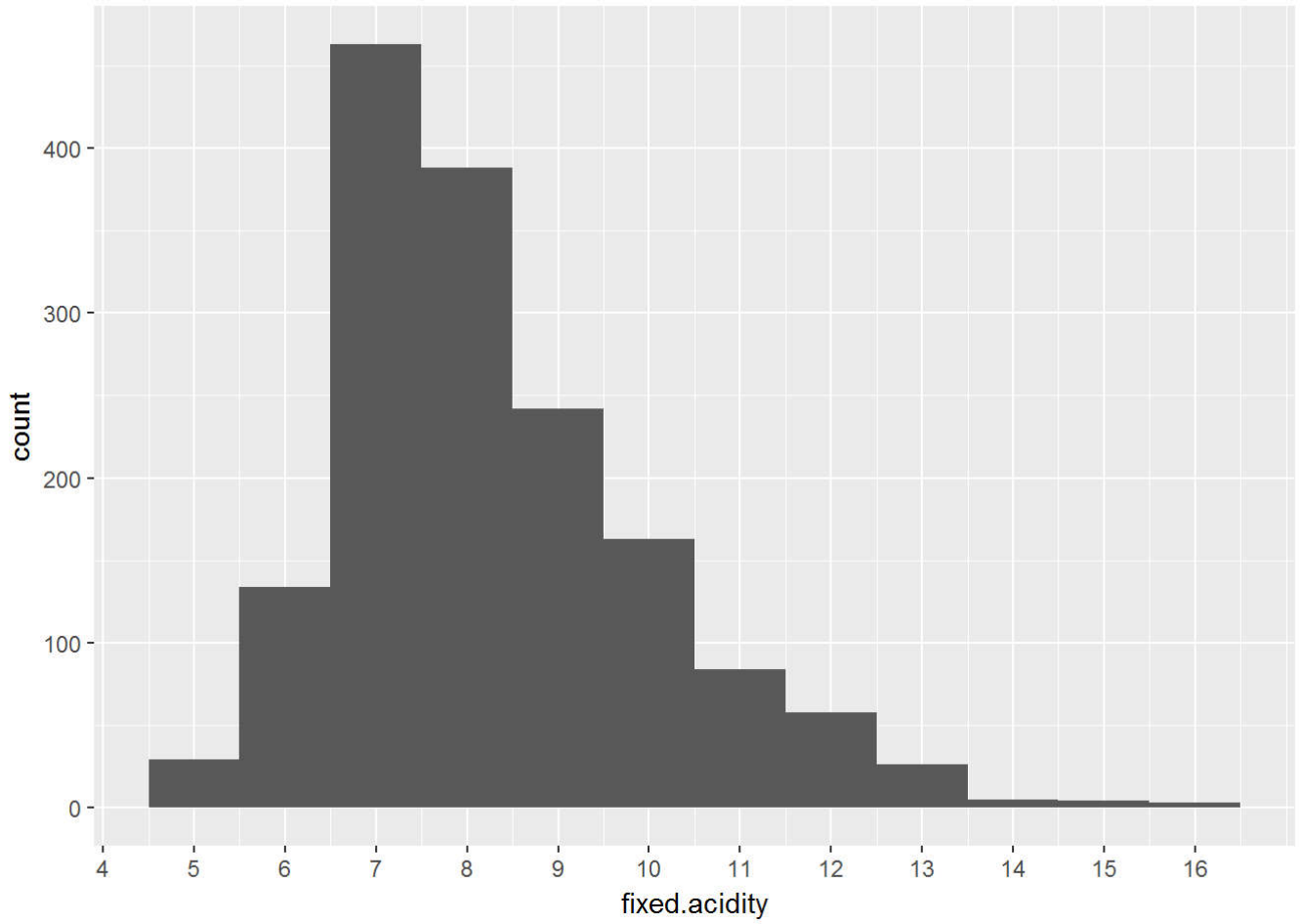
At this point, we can create new variable as “rate”. Rate includes categorical variables “bad” for 4 points and below, “normal” for 5 and 6 points, “good” for 7 points and above.



This variable will help us in bivariate analysis.

Fixed acidity

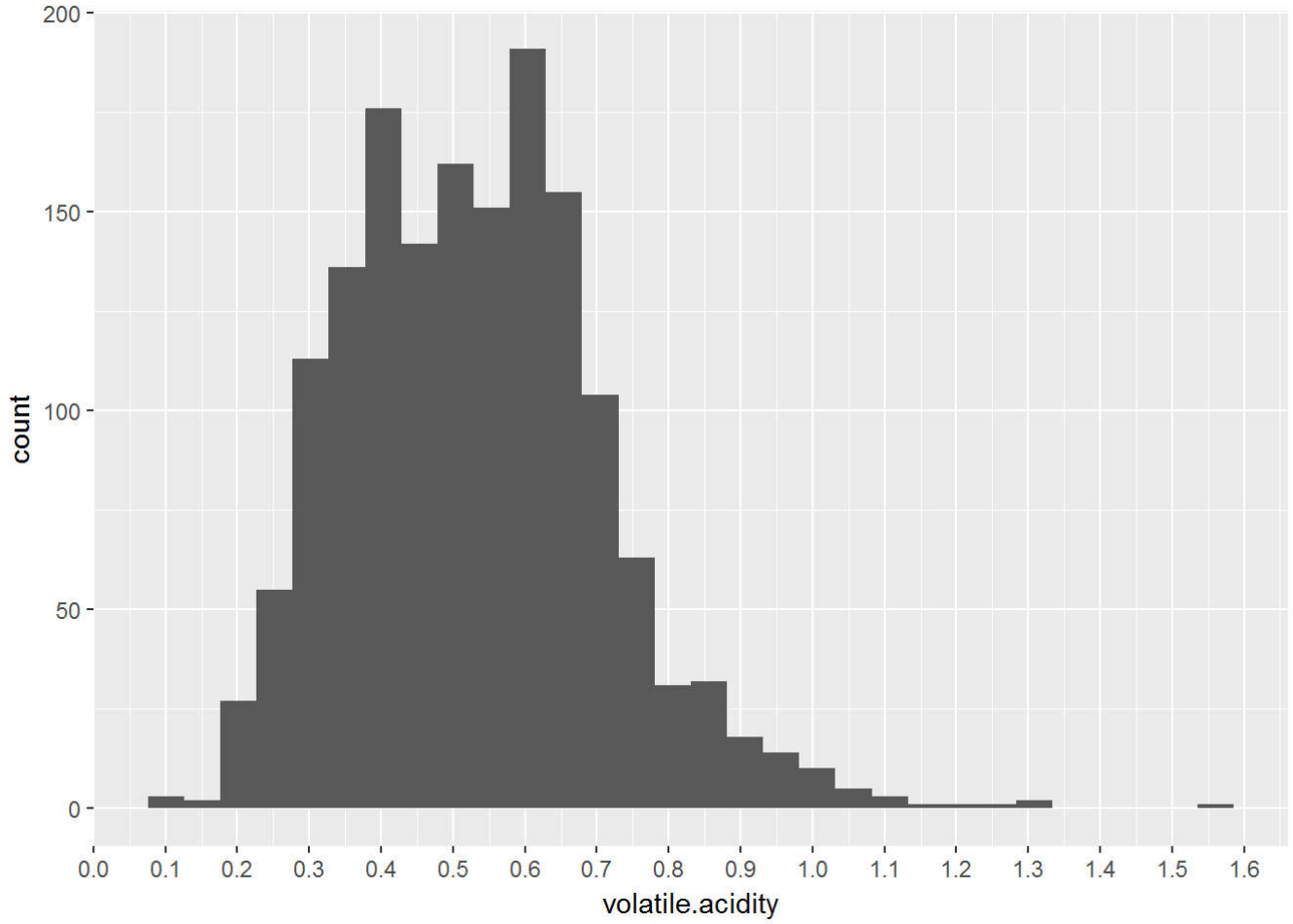
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```



Peak stands at 7. Most values are spread between 5 and 13.Distribution is lightly positive skewed.

Volatile Acidity

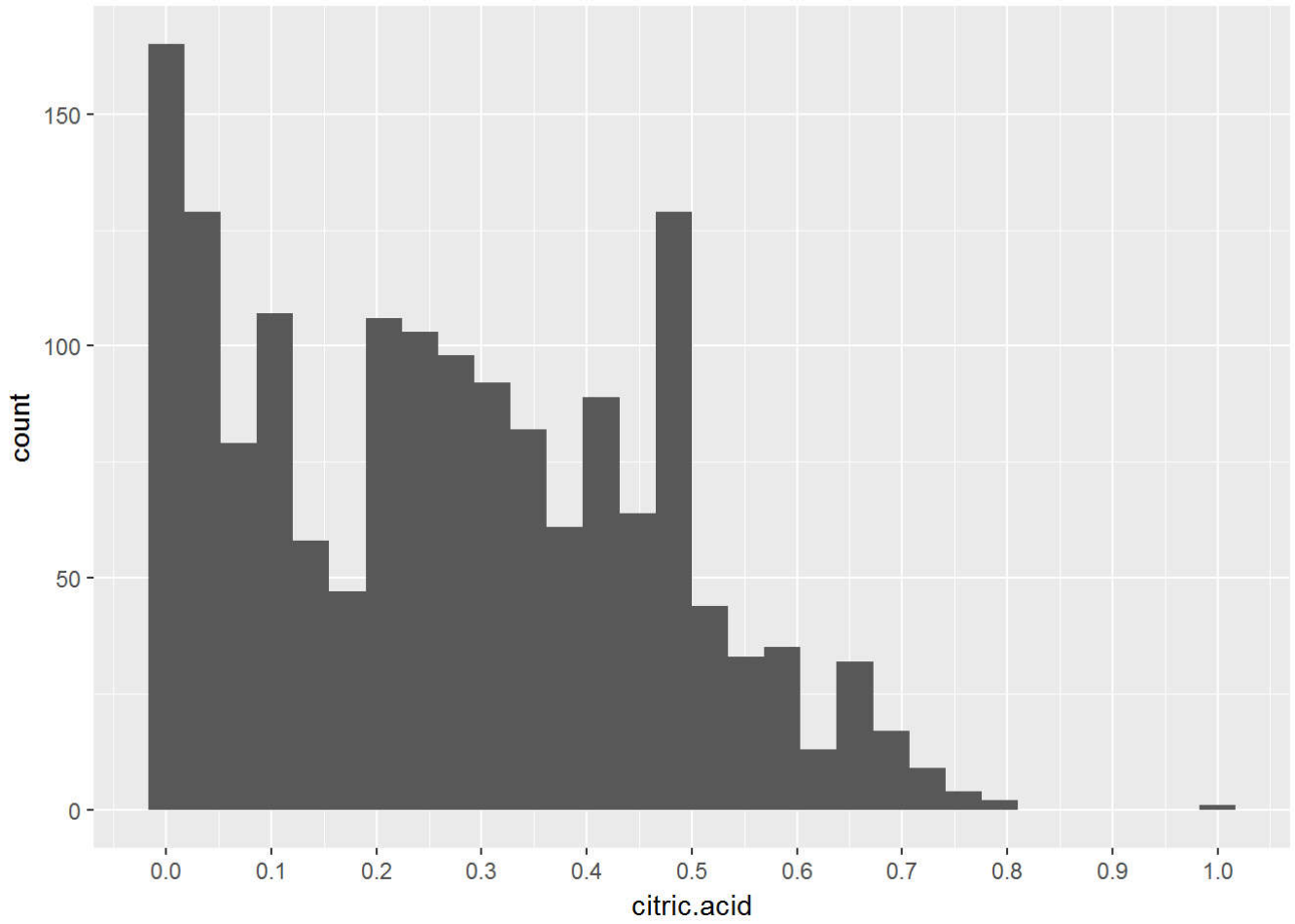
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```



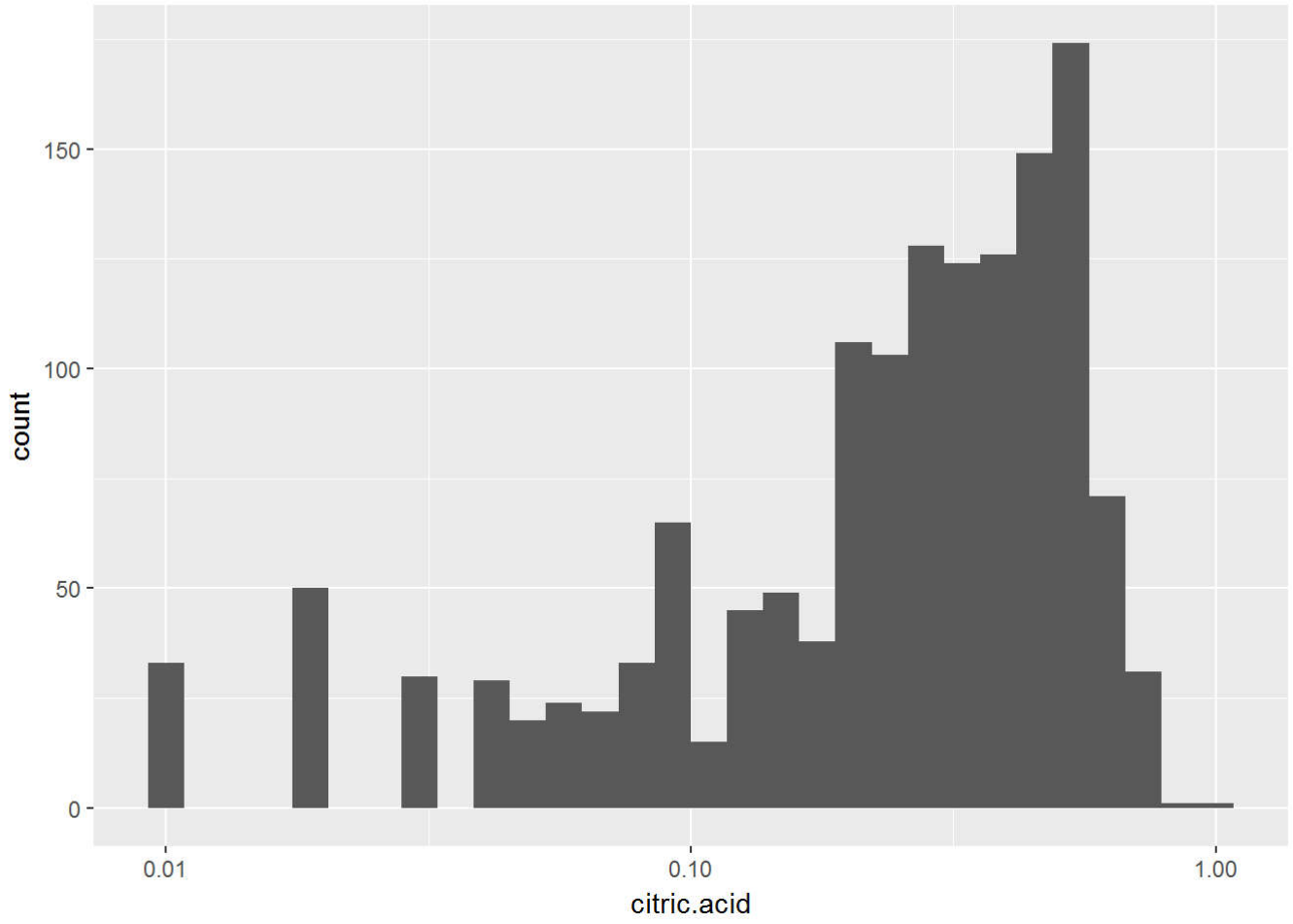
There are two peaks at 0.4 and 0.6. Values are spread between 0.2 and 0.9.

Citric acid

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



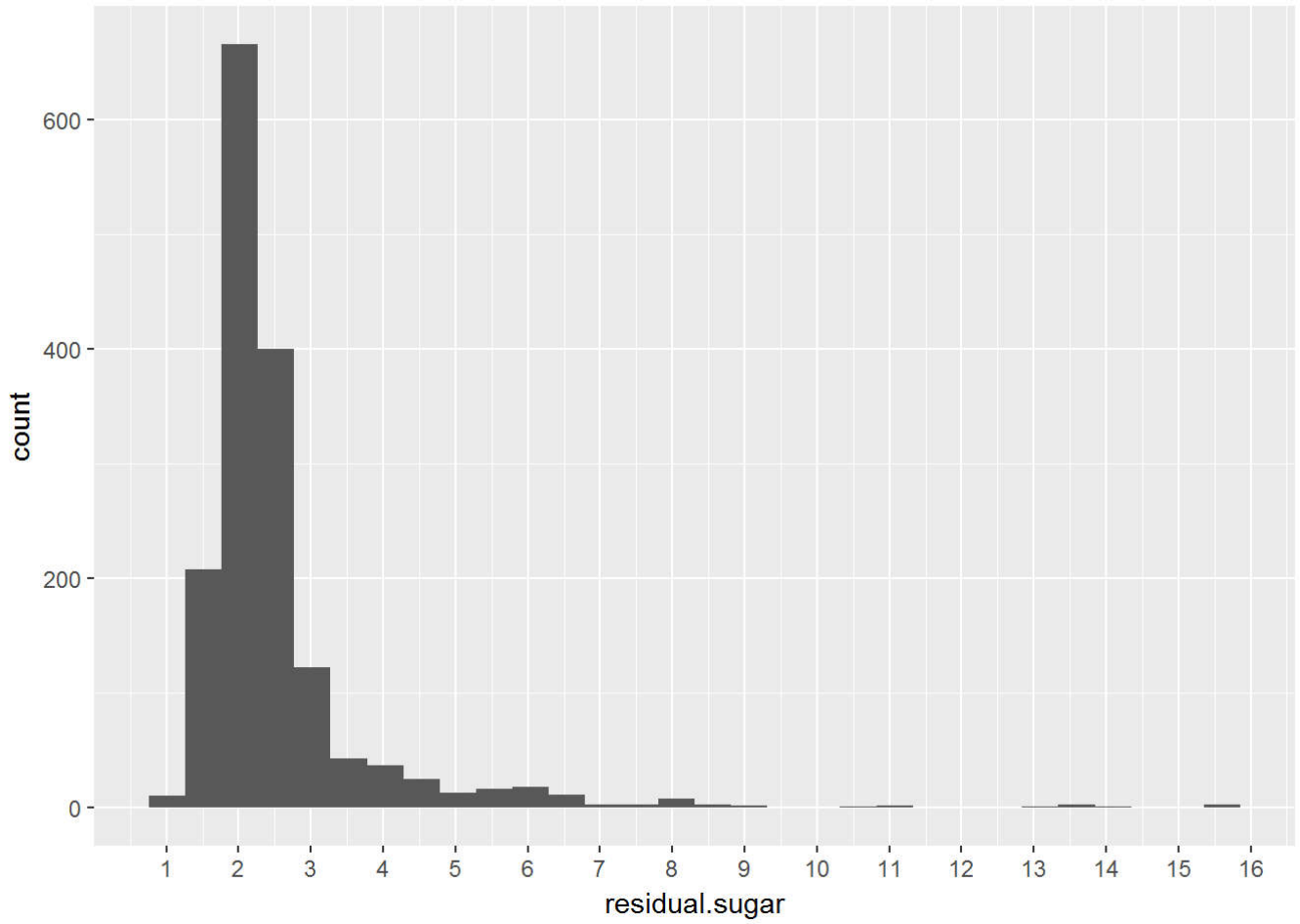
Data peaks at three values: 0, 0.25 and 0.5 Most values spread within 1 and 4. It has multiple peaks so called multimodal data.



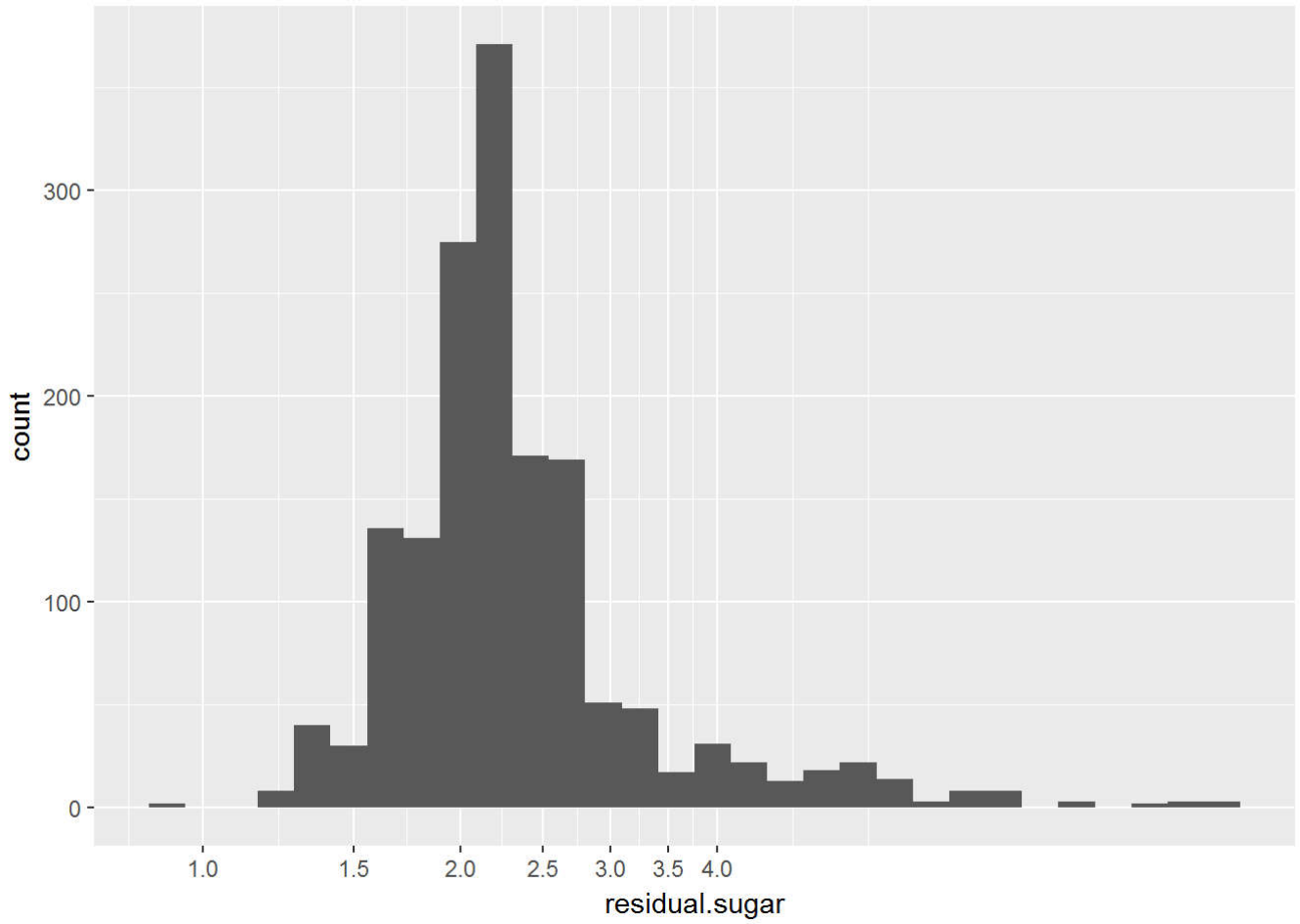
We performed log10 transformation but again it has not normal distribution.

Residual sugar

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



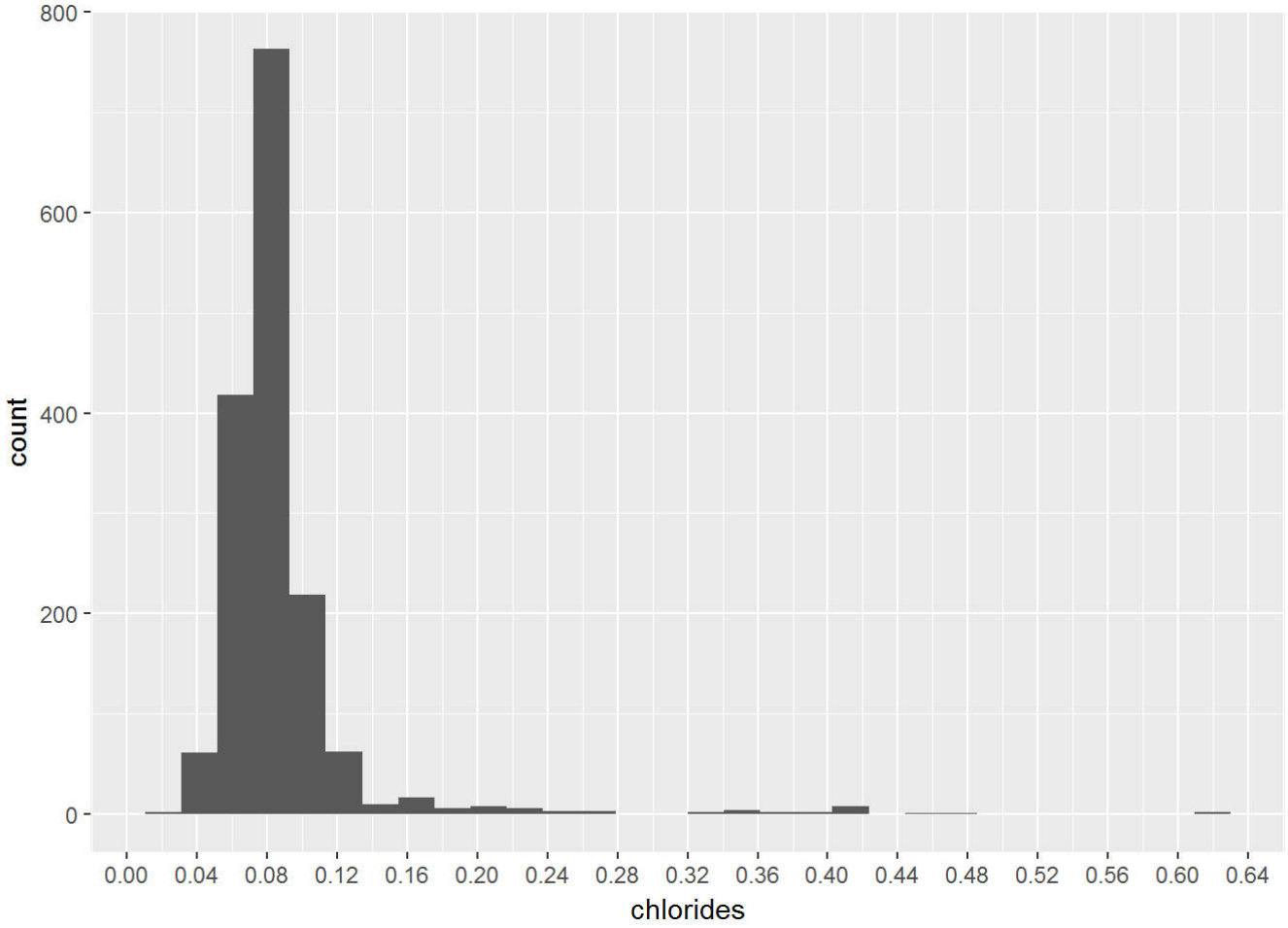
Data peaks at 2. Most values spread within 1 and 4. There are outliers so distinct. We can perform log10 transformation for this variable.



Again data is not normally distributed, positively skewed. Data peaks at 2.2. Most values spread within 1.5 and 3.

Chlorides

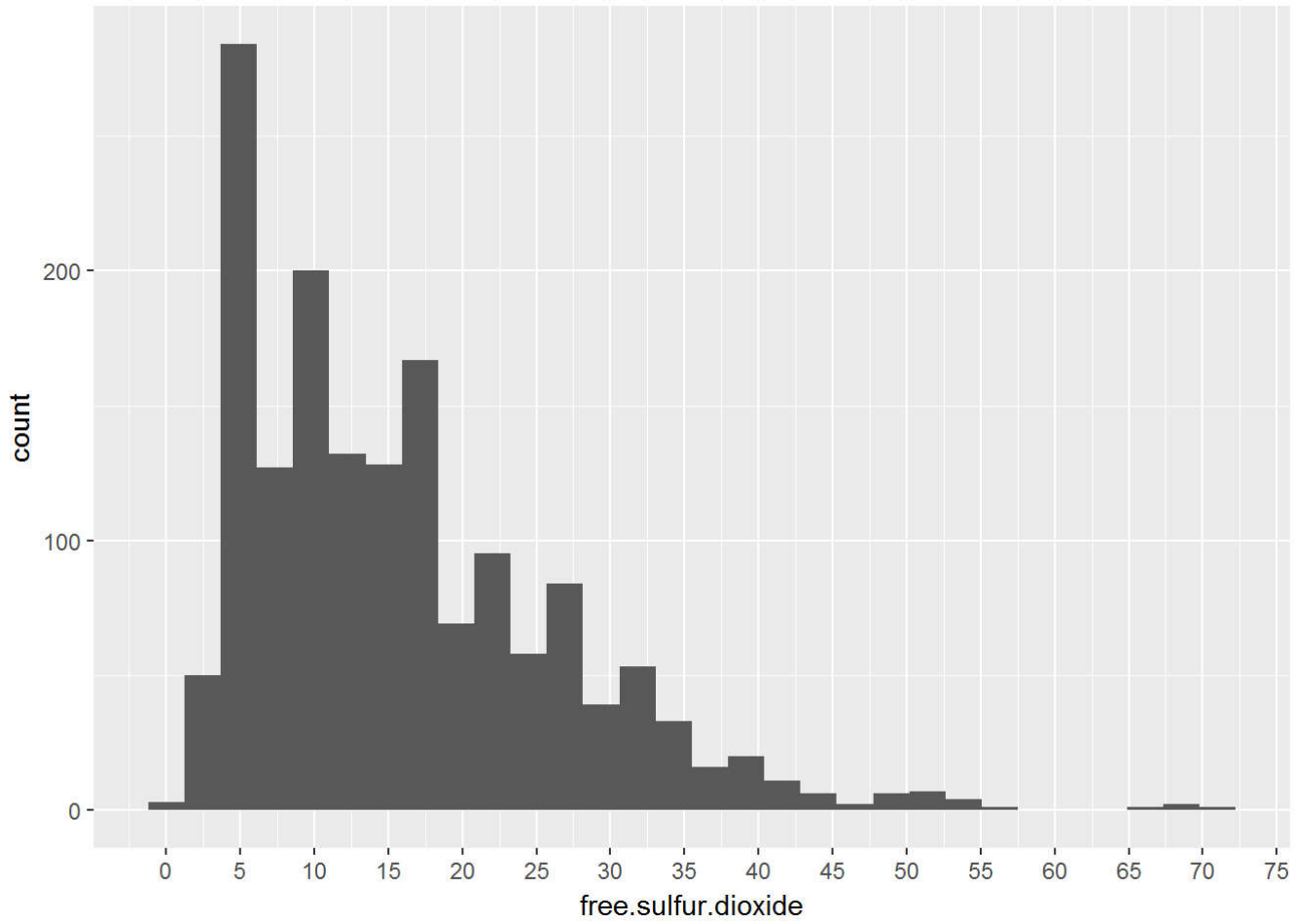
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```



Data peaks at 0.08. Most values spread within 0.04 and 1.2. Distribution is positively skewed.

Free sulfur dioxide

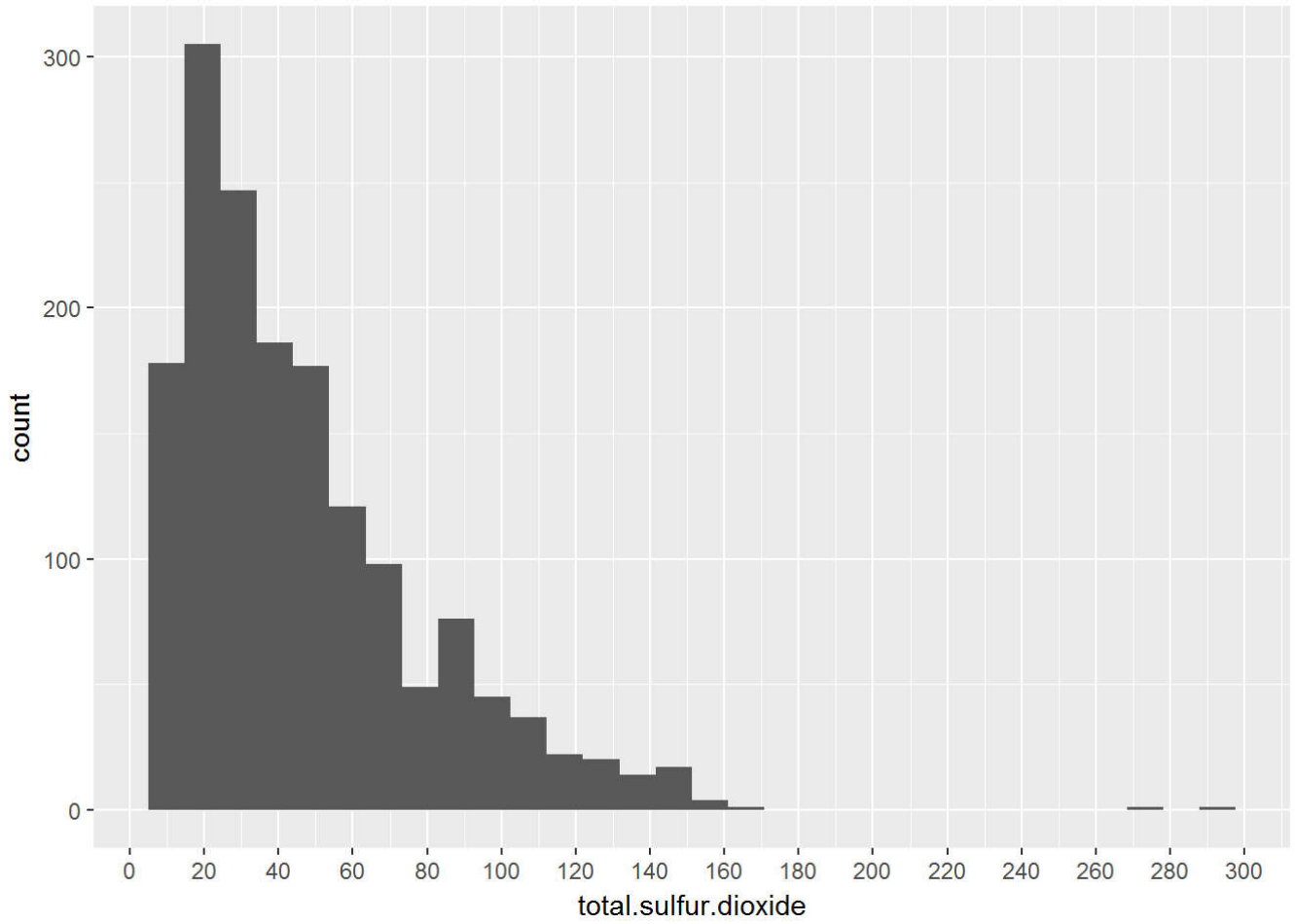
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```



Peak stands at 5. Values mostly spread between 2 and 35. Distribution is lightly positive skewed.

Total sulfur dioxide

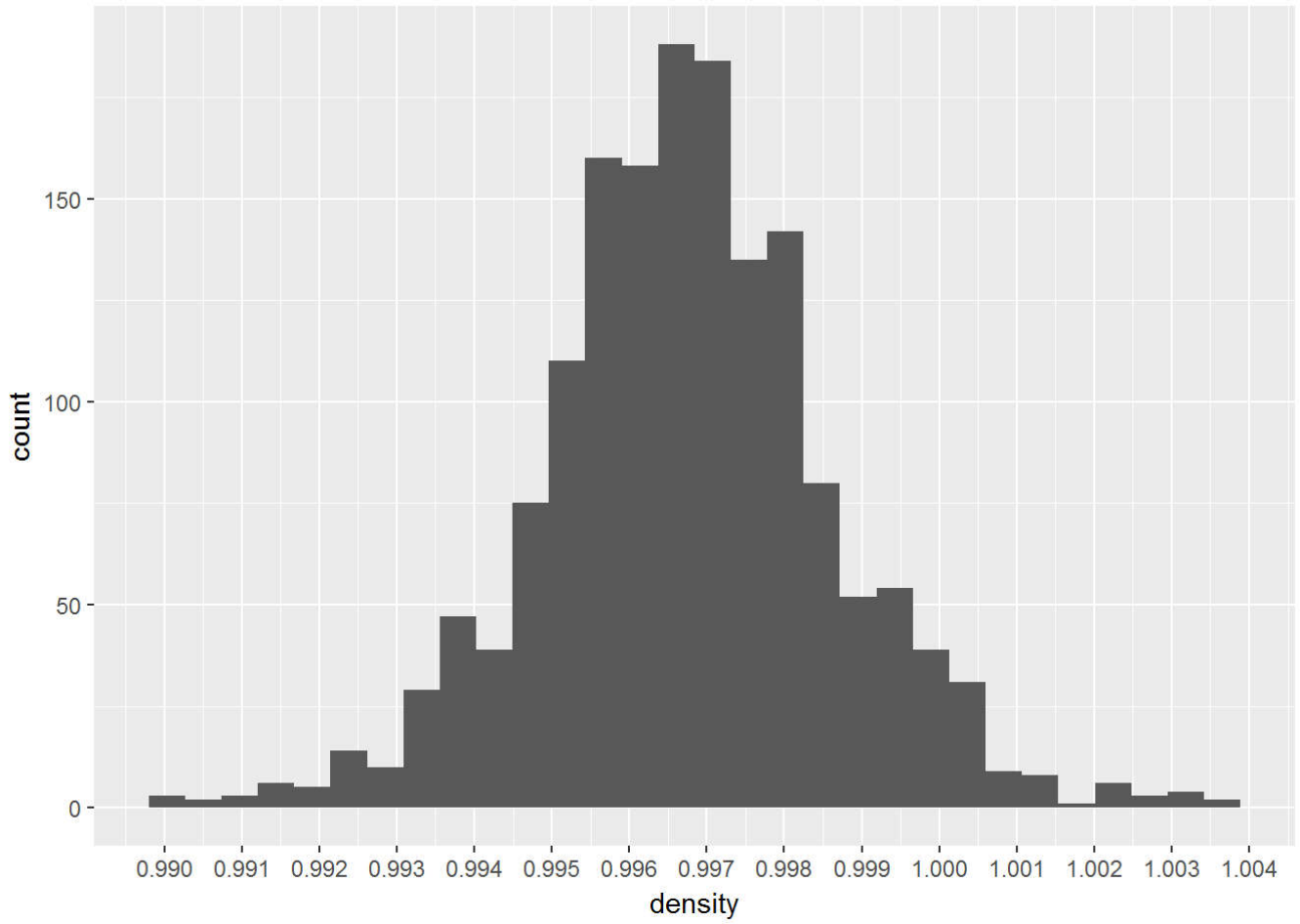
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```



Data peaks at near 20. Most Values spread between 6 and 160. Distribution is positively skewed.

Density

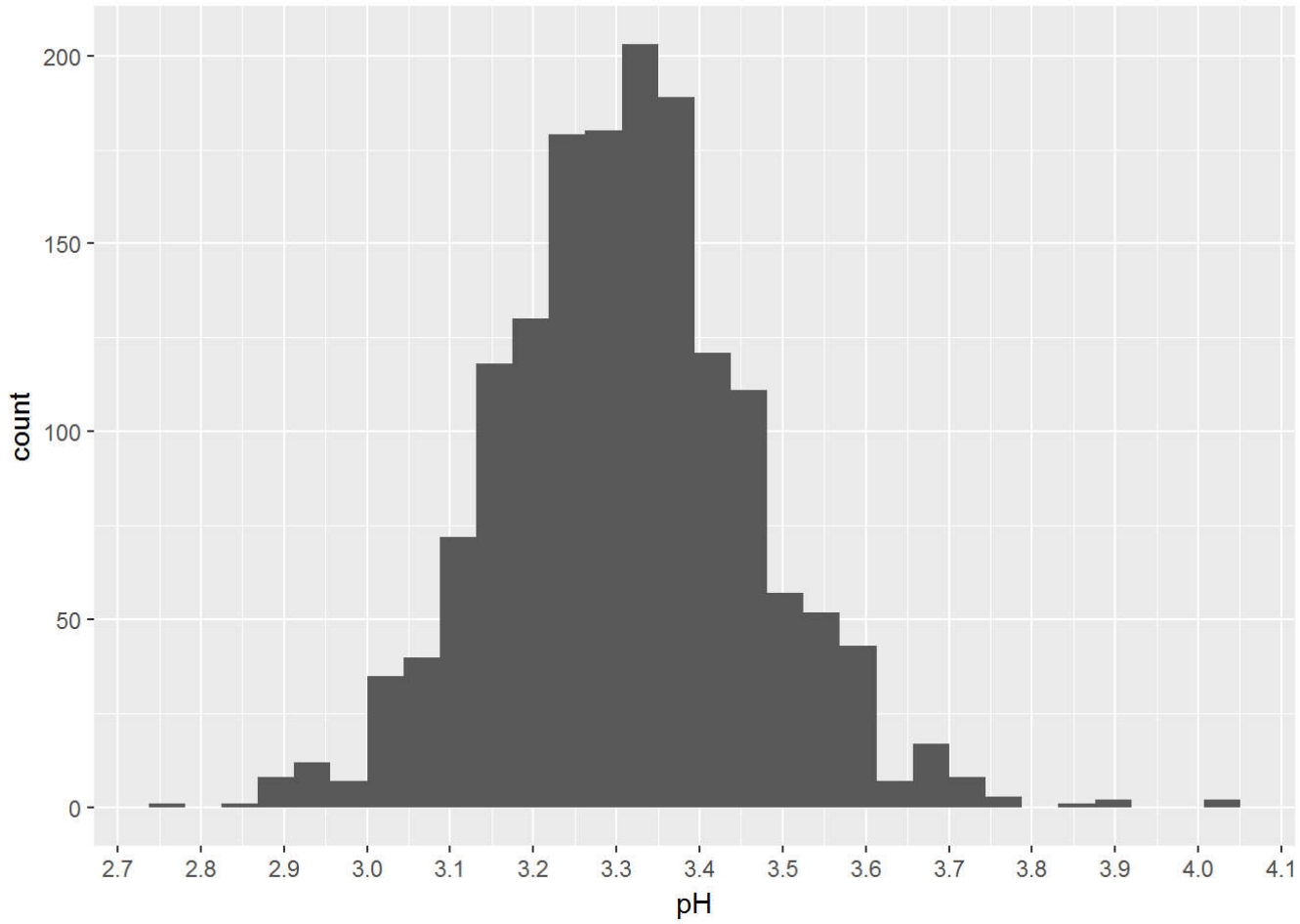
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```



Data peaks at near 0.997 . Most values spread between 0.993 and 1.002. It is so close to normal distribution.

pH

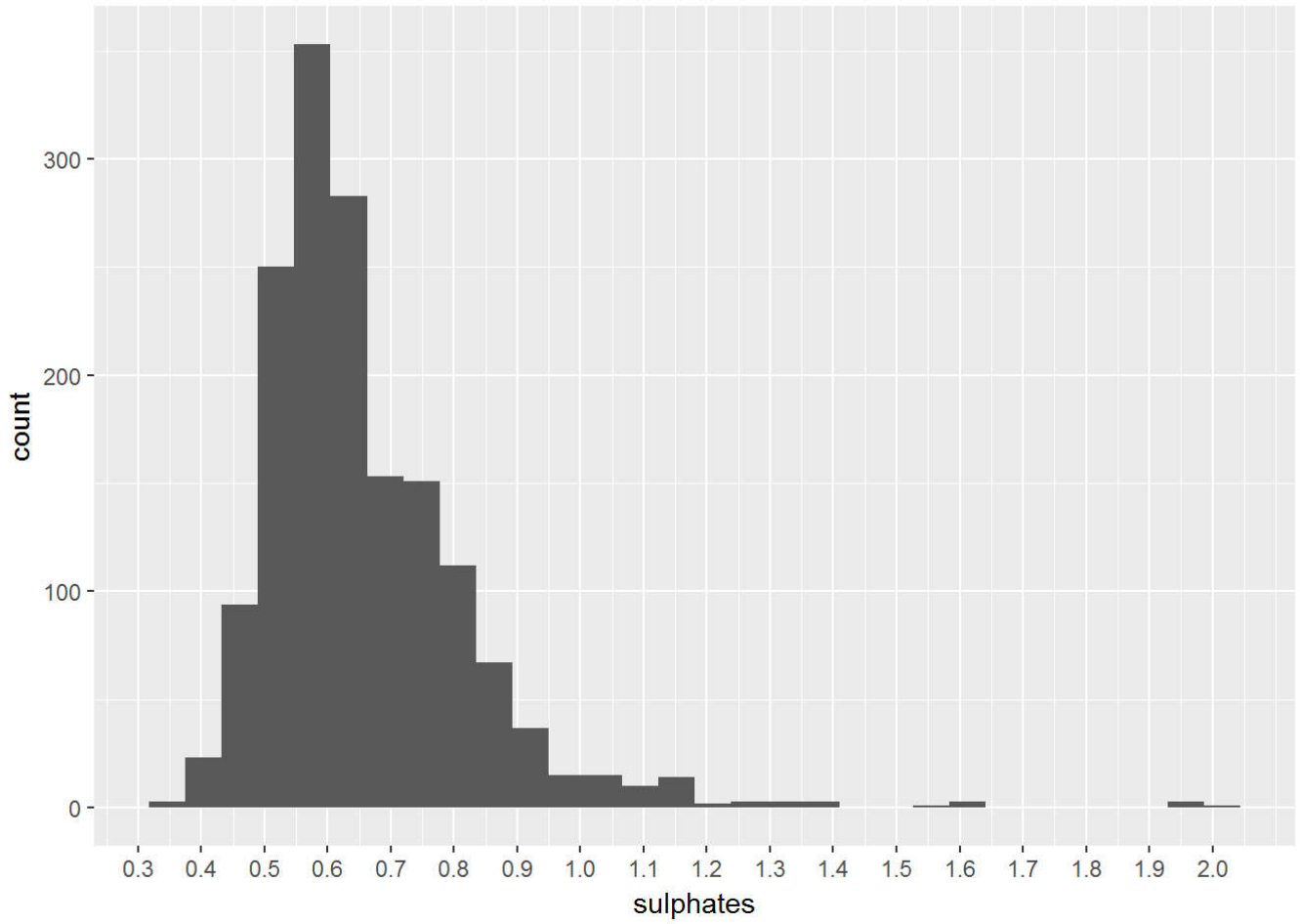
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```



Peak at near 3.3 . Values mostly spread between 3 and 3.6.

Sulphates

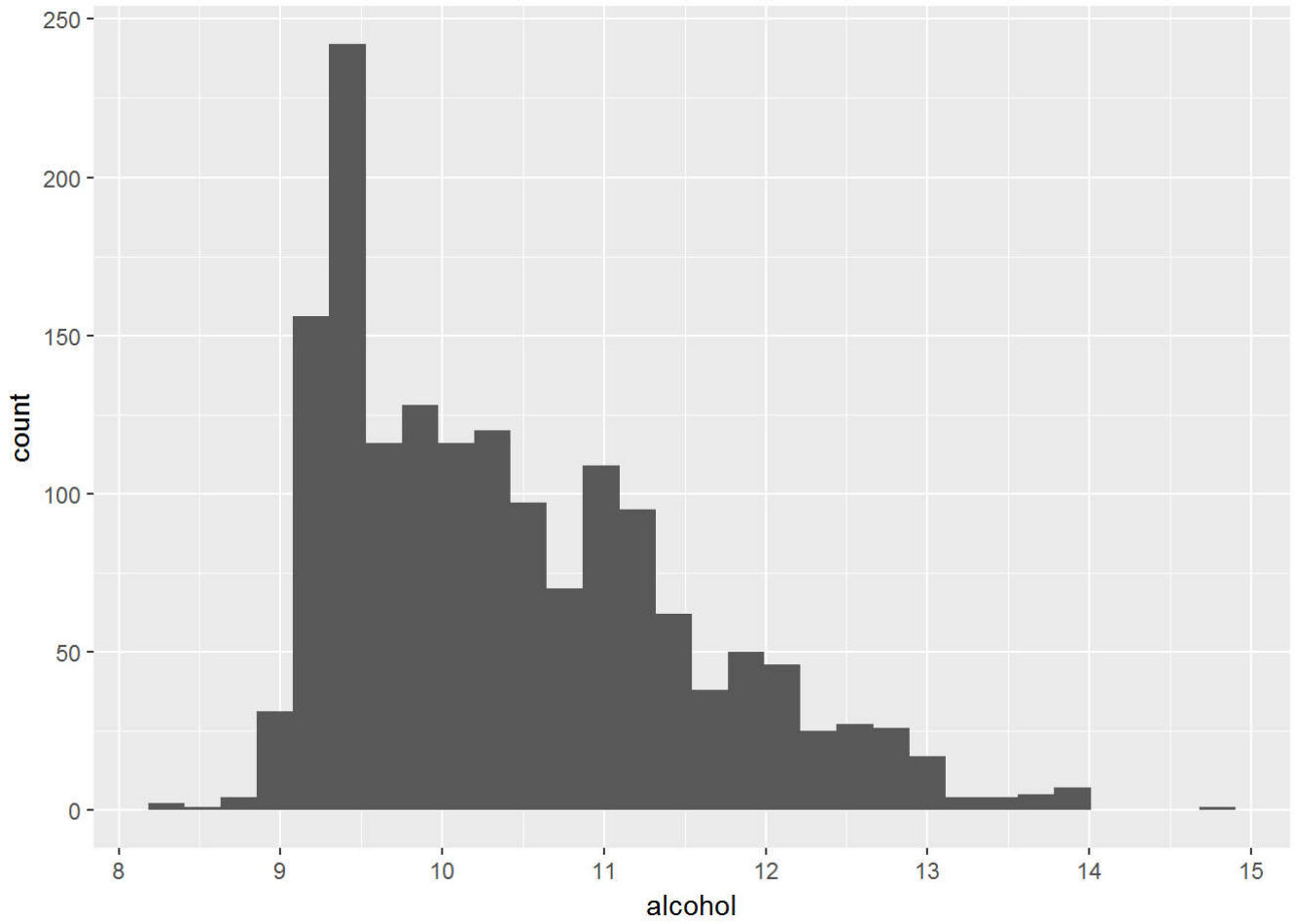
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```



Peak stands at near 0.6. Values mostly spread between 0.4 and 1.

Alcohol

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



Peak at near 9.4. Values mostly spread between 9 and 13. Distribution is positively skewed.

Univariate Analysis

What is the structure of your dataset?

There 1599 observations for each 12 attributes. Variables are positively skewed in general. Variables as pH and density, are behaved like normal distributions. As stated in reference document at introduction, dataset is not balanced,there are much more average wines than good or bad ones.

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

Main feature of interest is quality variable. We would find how the others affect quality variable.

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

I expect pH, residual sugar and total sulfur dioxide will help support to find quality of the wine. However, it is very difficult to estimate at this stage.

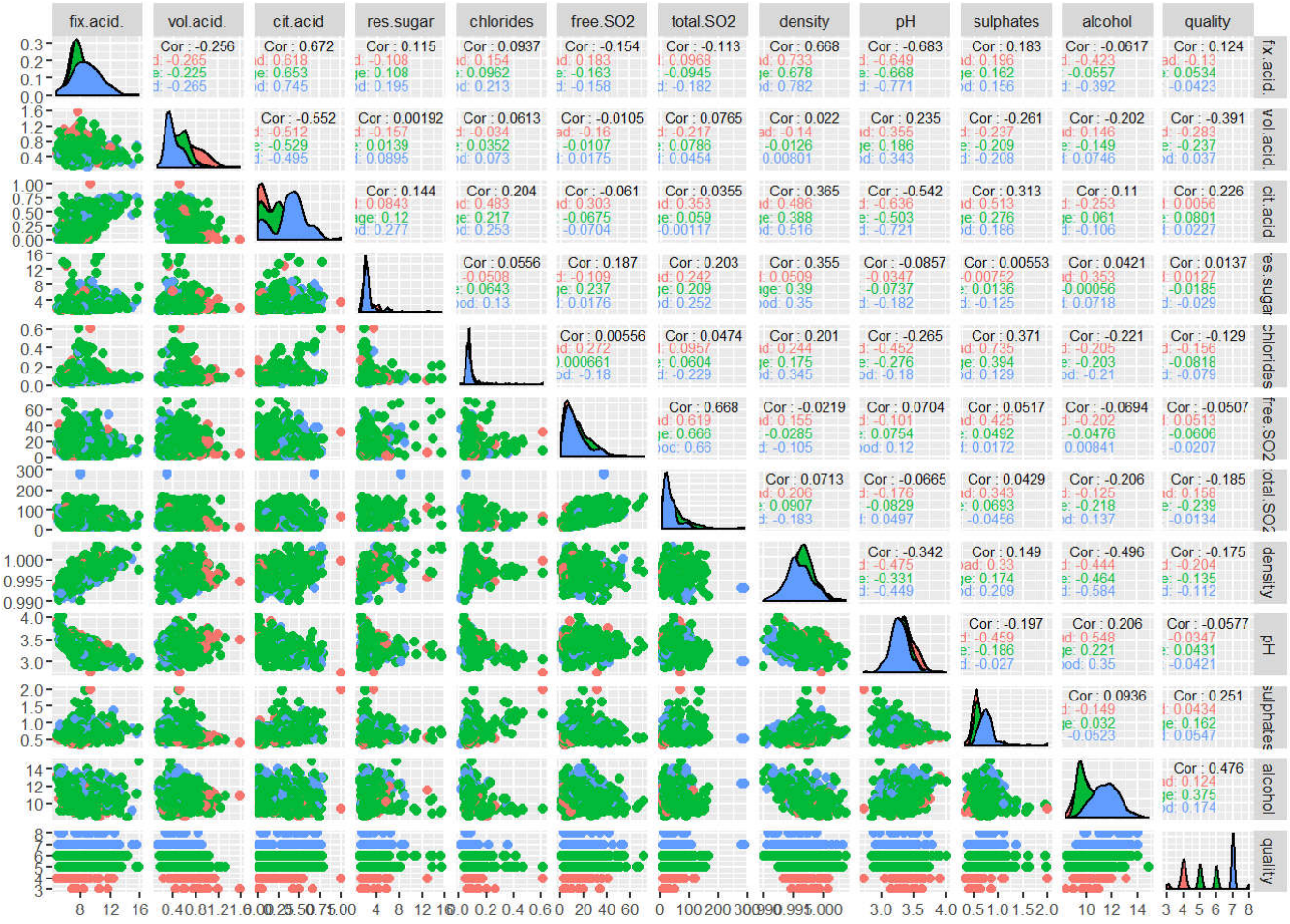
Did you create any new variables from existing variables in the dataset?

I created “rate” variable by cutting quality variable as categorical variable. Quality values are changed to categorical variables “bad” for 4 points and below, “average” for 5 and 6 points, “good” for 7 points and above. It will also behaves like our main feature.

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?

Distributions of residual.sugar and citric.acids are so different from normal distributions. So, I performed log10 transformations to inspect variables in normal distribution format. However I could not obtain normal distributions after transformations.

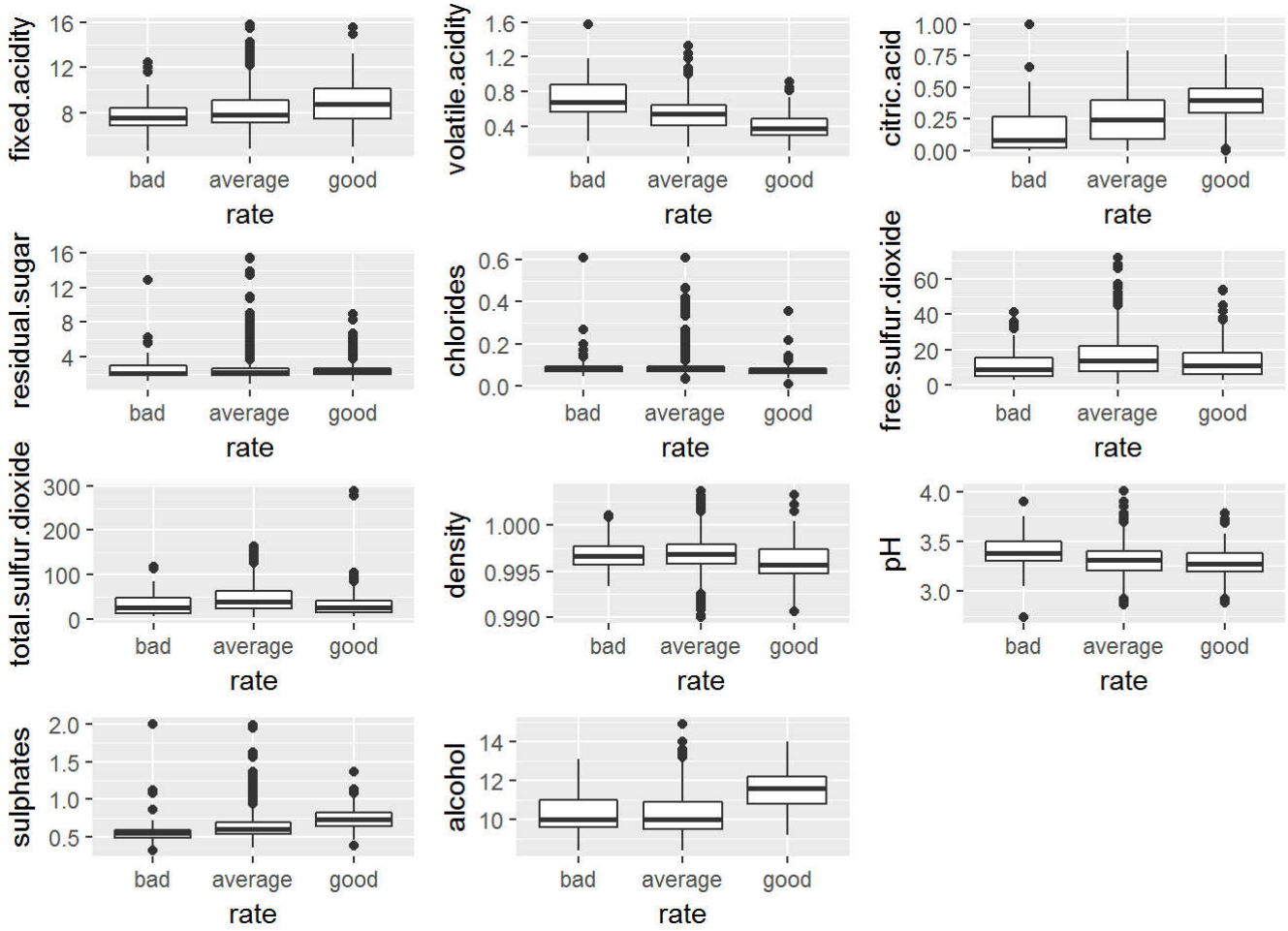
Bivariate Plots Section



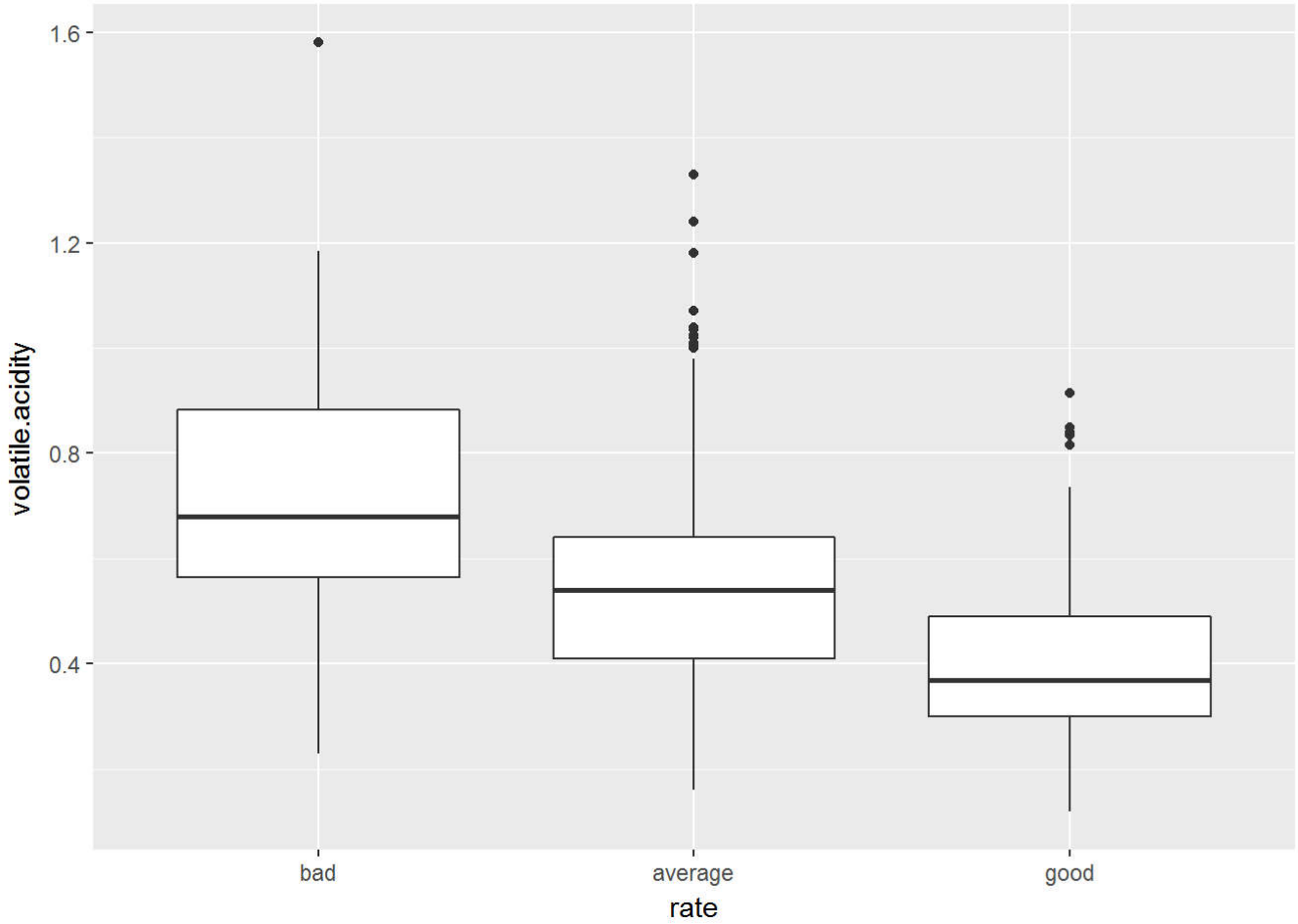
When we look at the correlation values, we detect some high corralated variables:

- citric.acid - fixed.acidity
- citric.acid - Volatile.acidity
- density - fixed.acidity
- pH- citric.acid
- pH- fixed.acidity
- free.sulfur.dioxide - total.sulfur.dioxide
- alcohol - density
- alcohol - quality

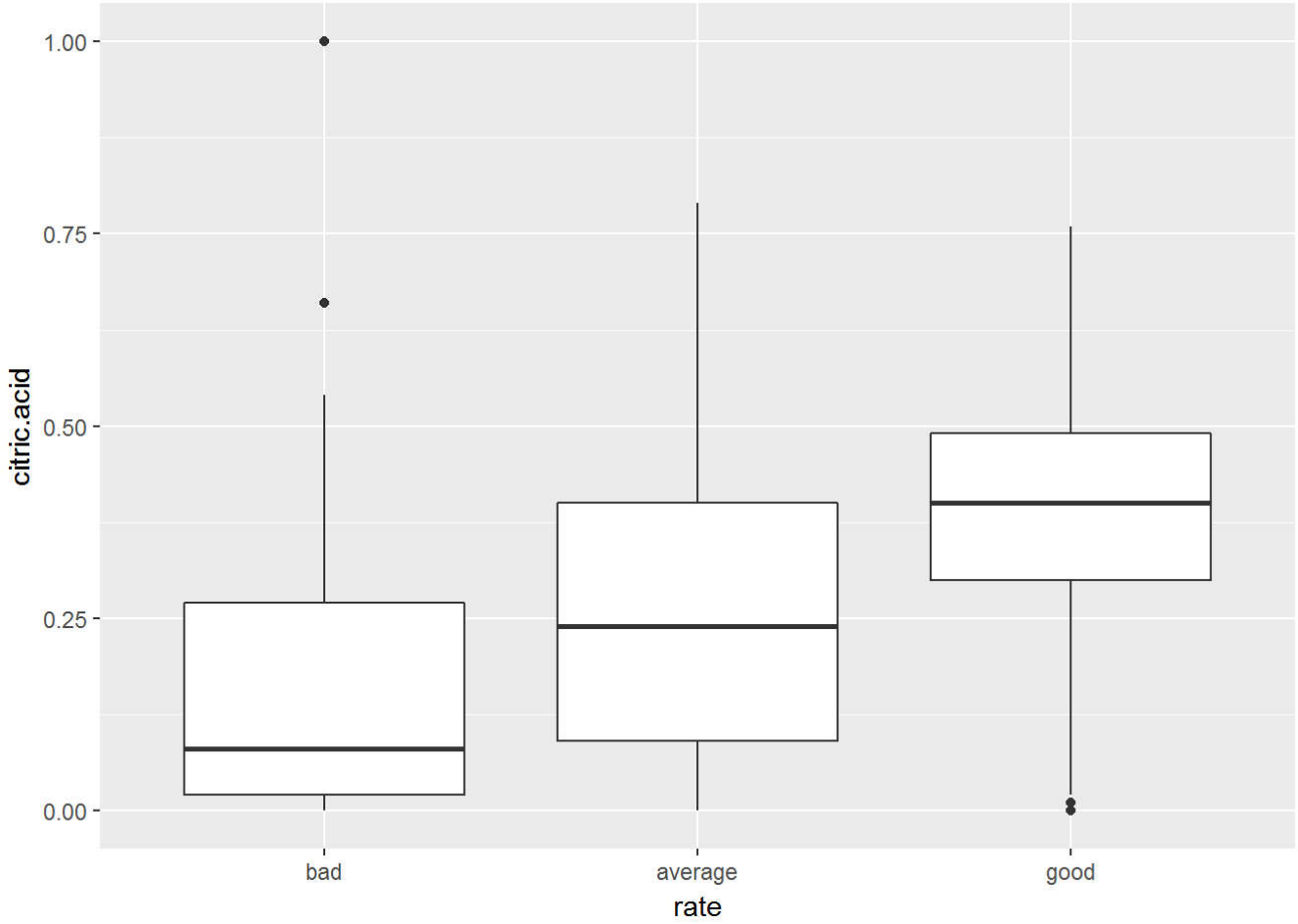
At this point the only correlated value with quality(main feature) seems as alcohol.In addition, we can also investigate correlation by using our new variable “rate”. Rate is categorical, so we will use boxplot.



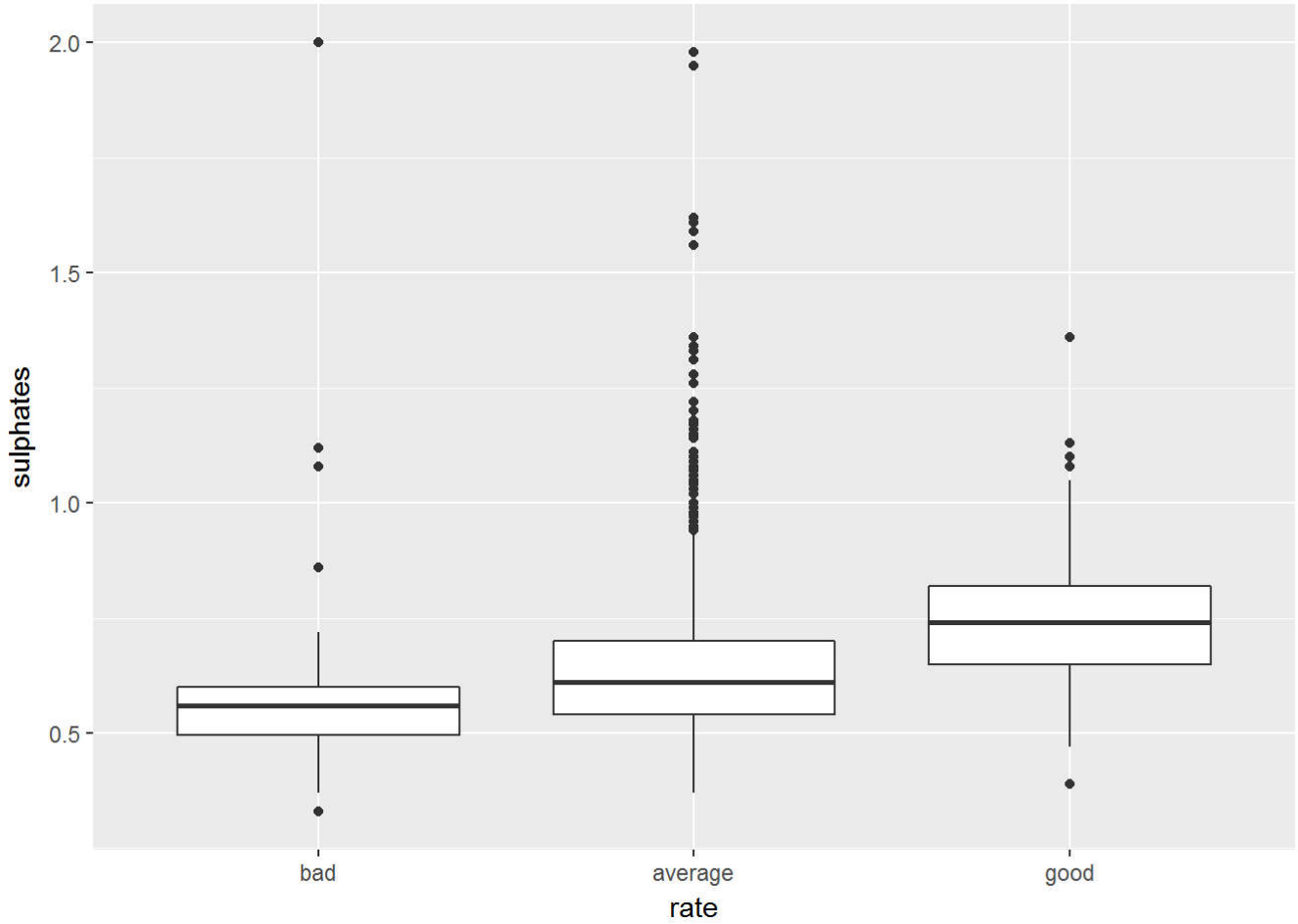
We can see from plots,there is not obvious distiction with average vs. good and average vs. bad in most of the comparisons.However there are distinctions between bad and good. It is easily seen for volatile acidity, citric acid,sulphates and alcohol. Due to number of outliers, some of them is not clear. We can investigate boxplots with more details.



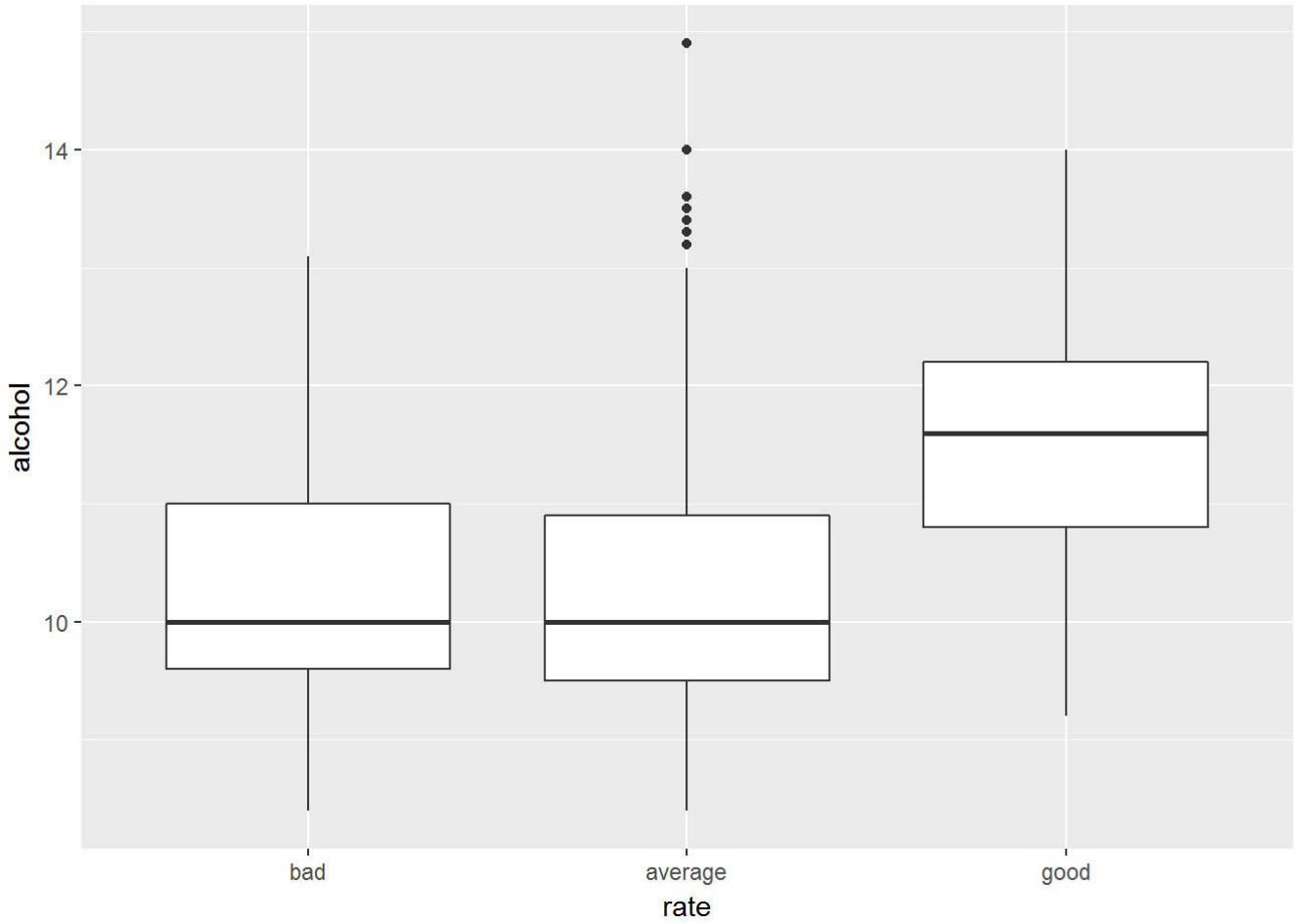
Wines labeled as good have lower volatile acidity value.



Wines labeled as good have higher citric acid value.

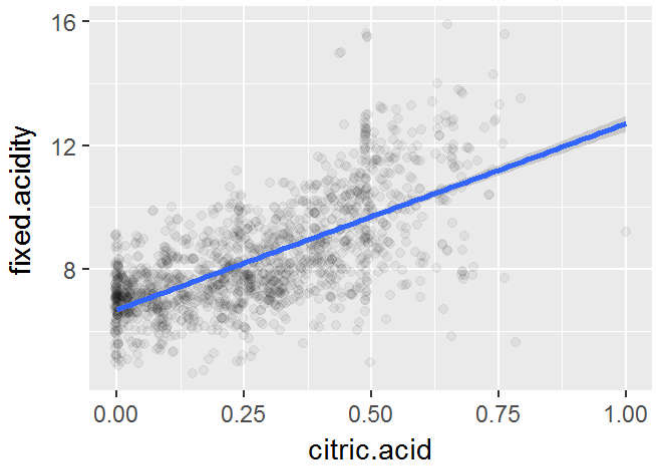


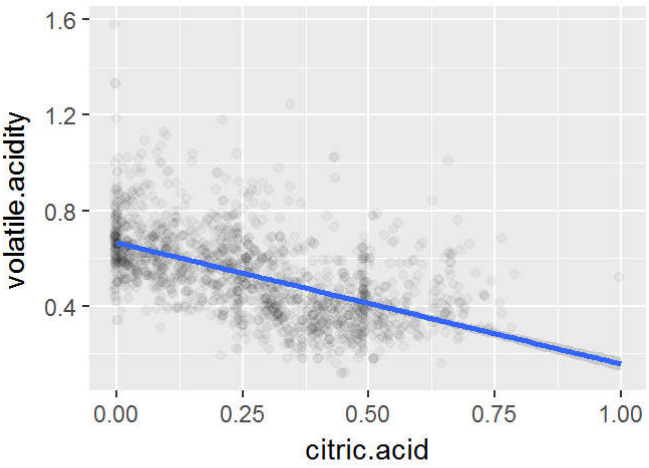
Wines labeled as good have higher sulphates amount.

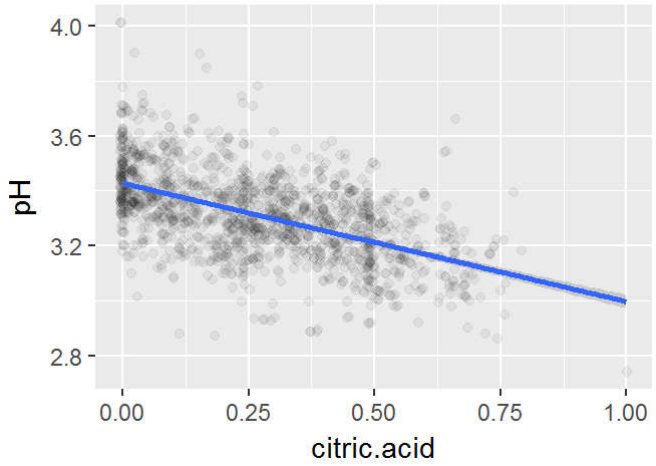


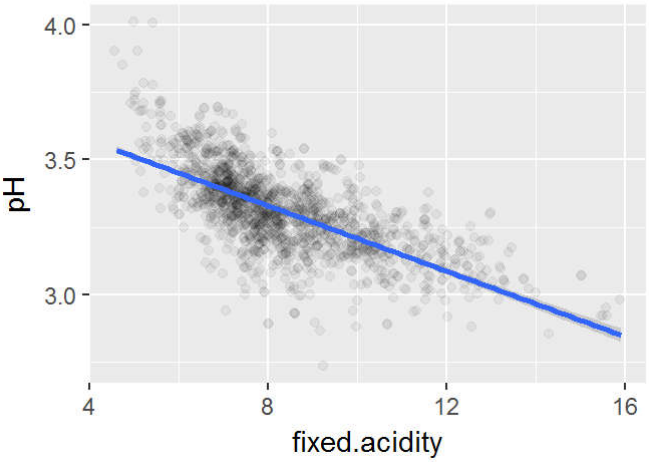
Wines labeled as good have higher alcohol level.

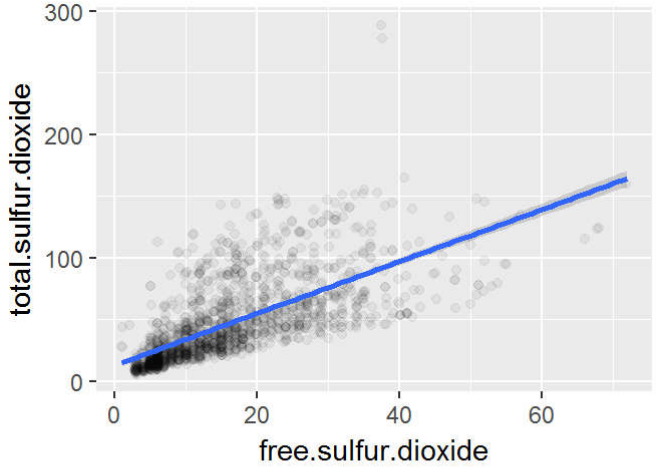
When finding correlation between other variables,so we can use scatterplot due to fact that they are numerical.

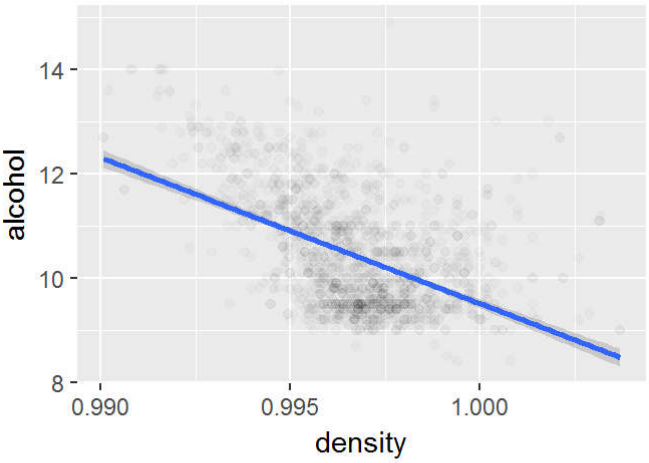


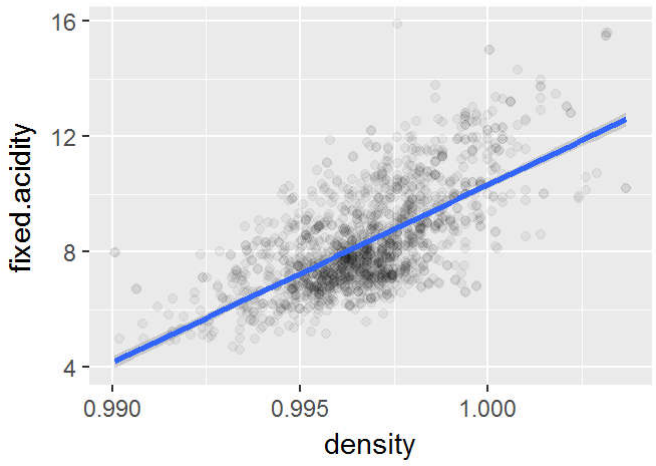












Relation of these variables is summarized as:

- citric.acid and fixed.acidity are correlated with correlation factor 0.672
- citric.acid - volatile.acidity are correlated with correlation factor with correlation -0.552
- pH and citric.acid are correlated with correlation factor -0.542
- pH and fixed.acidity are correlated with correlation factor -0.683
- free.sulfur.dioxide and total.sulfur.dioxide are correlated with correlation factor 0.668
- alcohol and density are correlated with correlation factor -0.496
- density and fixed.acidity with correlation 0.668

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?

Wines are keen to labeled as “good” quality with:

- lower volatile acidity value.
- higher citric acid value.
- higher sulphates amount.
- higher alcohol level.

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

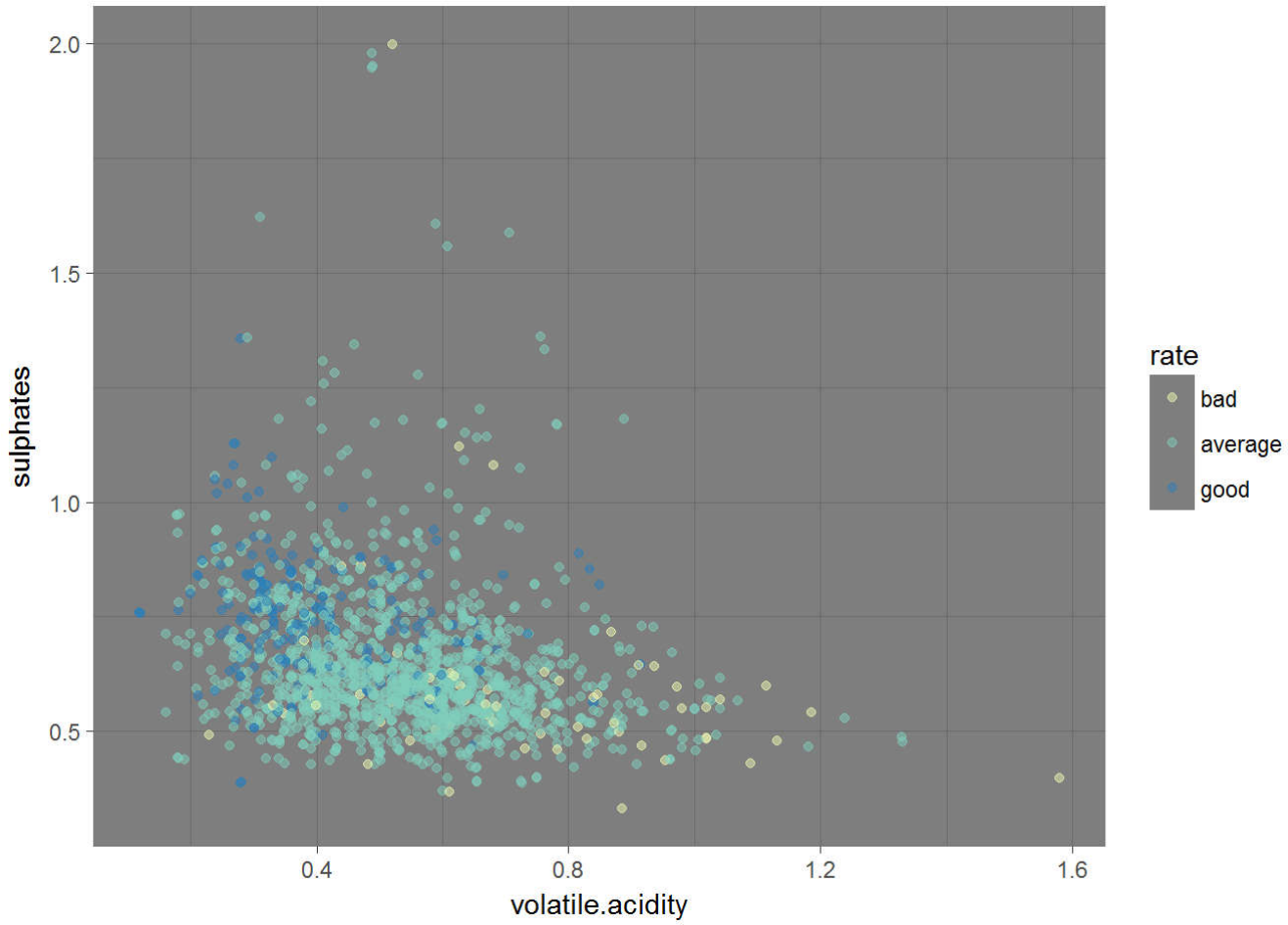
Both of them has name acidity, however Citric acid and Volatile acidity have negative correlation between each other. Also relationship density with fixed acidity and relationship density with alcohol is also interesting for me.

Sulfur-dioxide relationships and pH-acids relationships are expected.

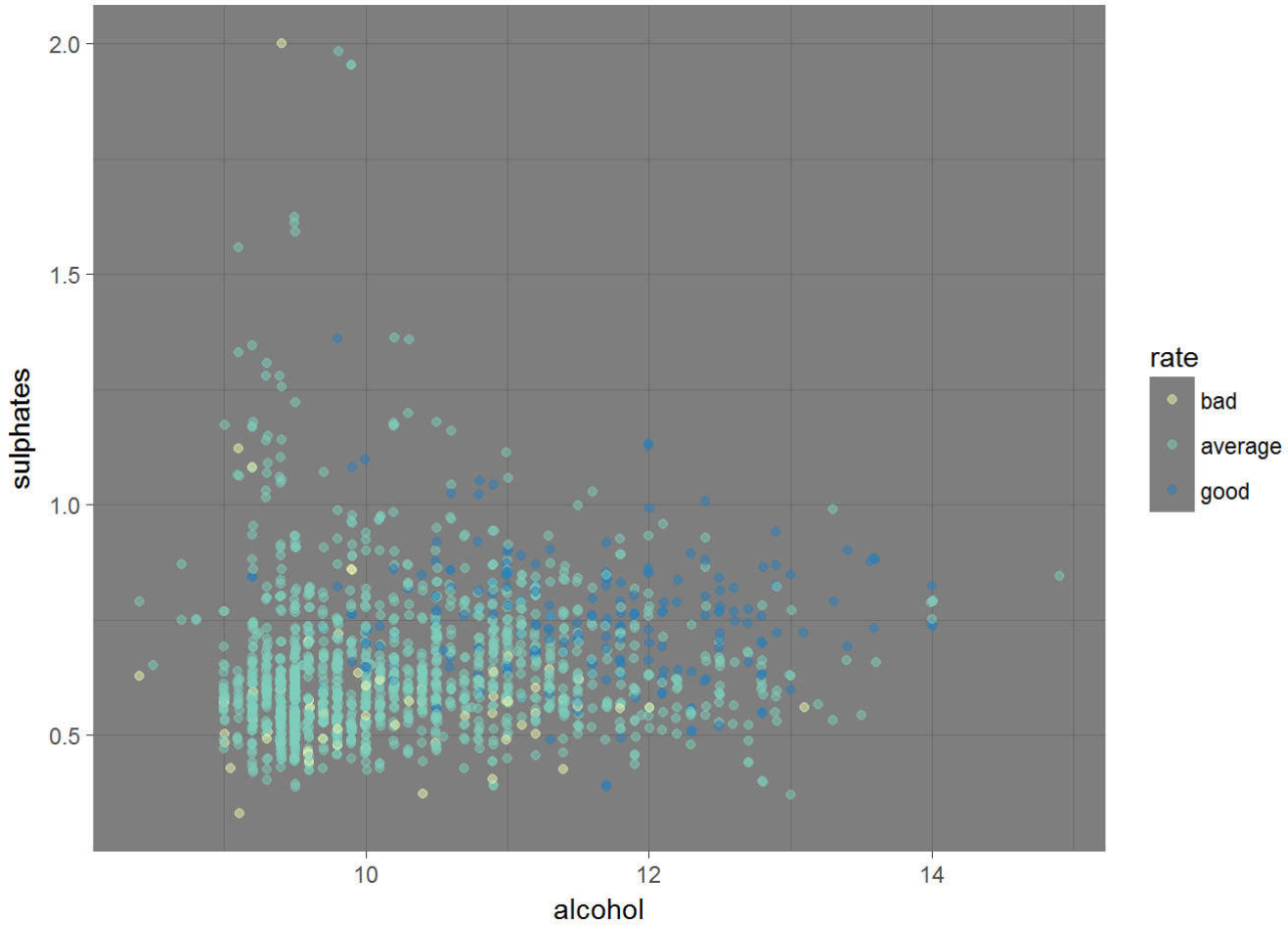
What was the strongest relationship you found?

The strongest relation was found between pH and fixed acidity with the correlation -0.683. It is logical because pH values is related with acidity as chemical definiton.

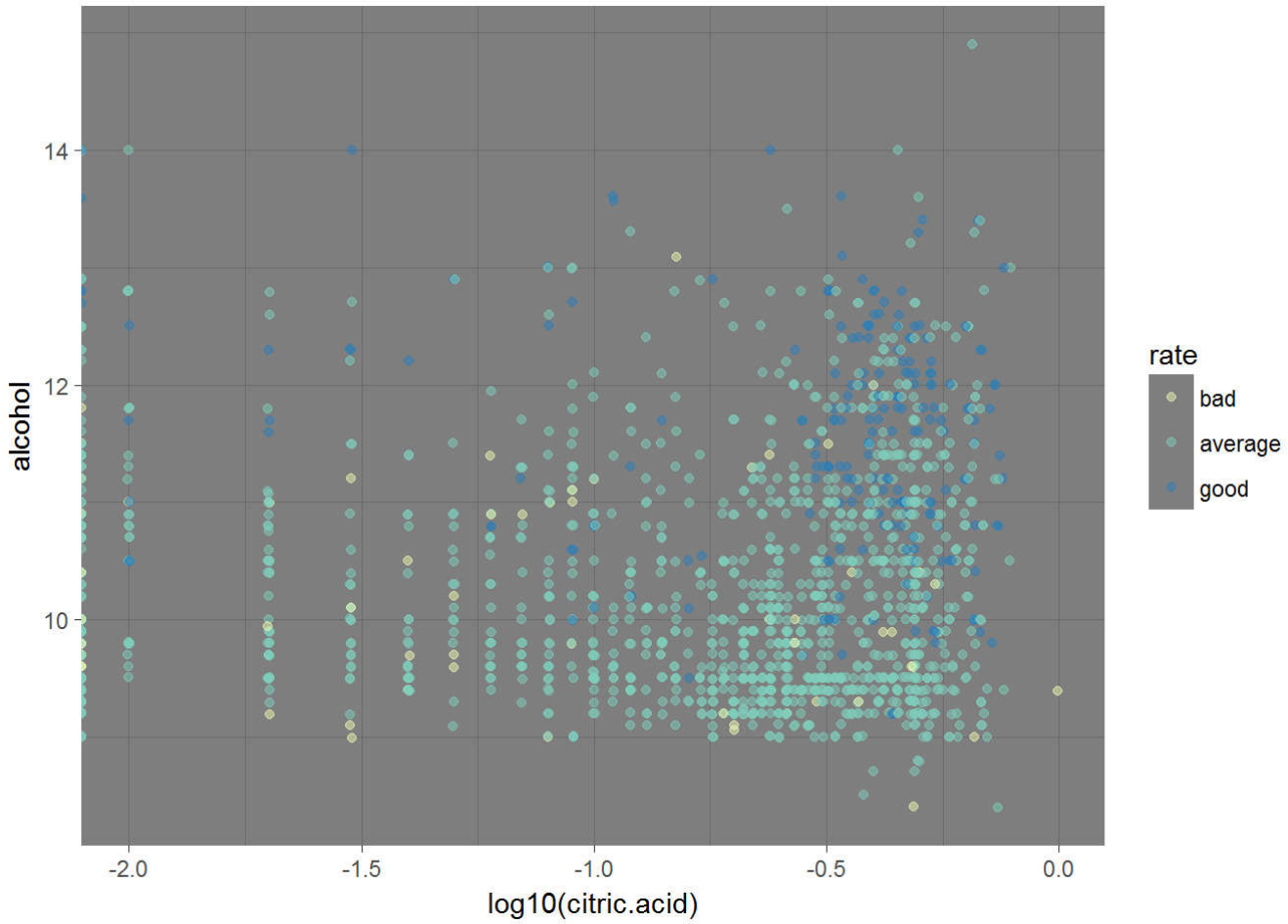
Multivariate Plots Section



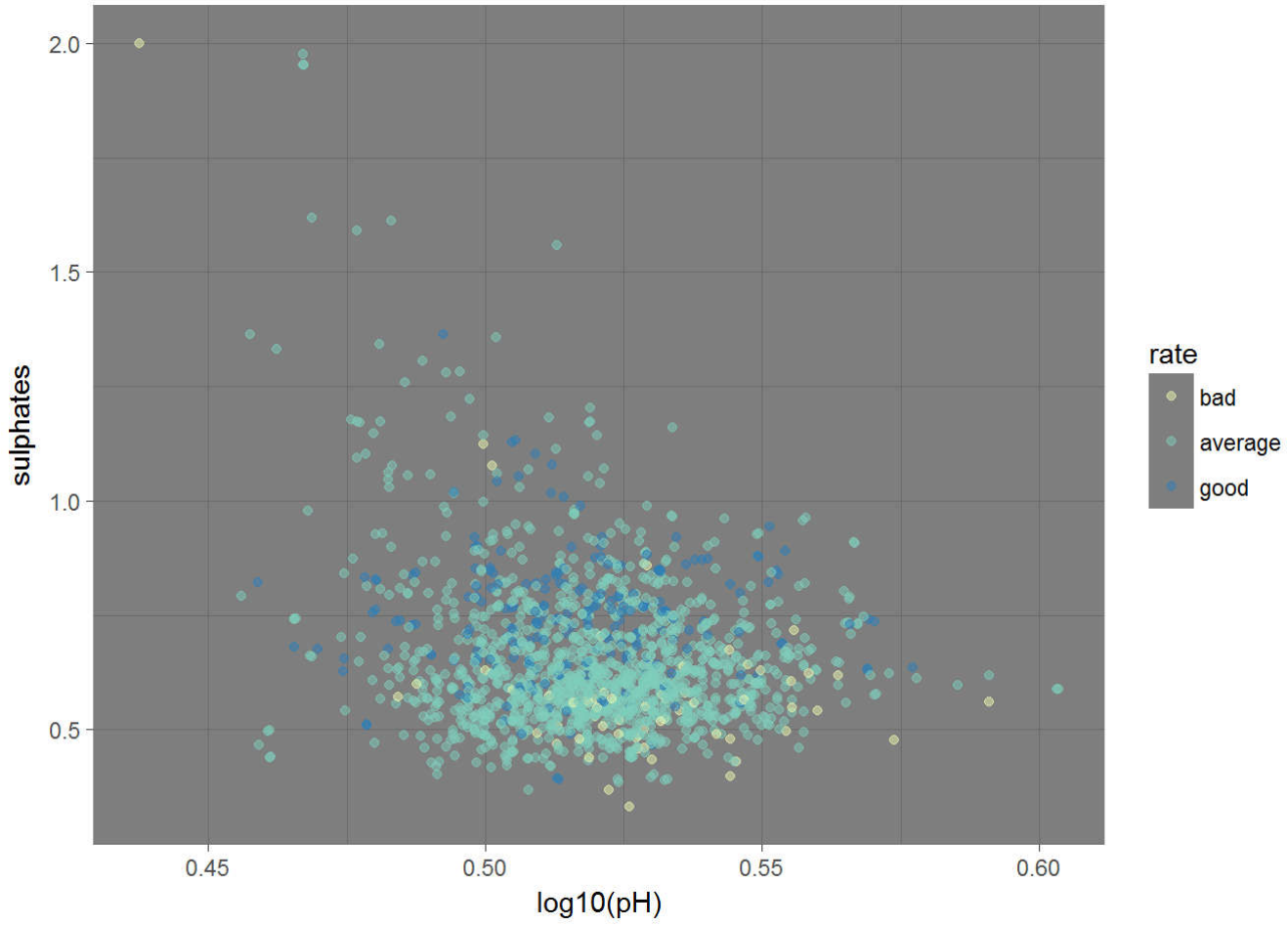
Very obvious finding in here: higher sulphates with lower volatile acidity mostly gives better quality wines.



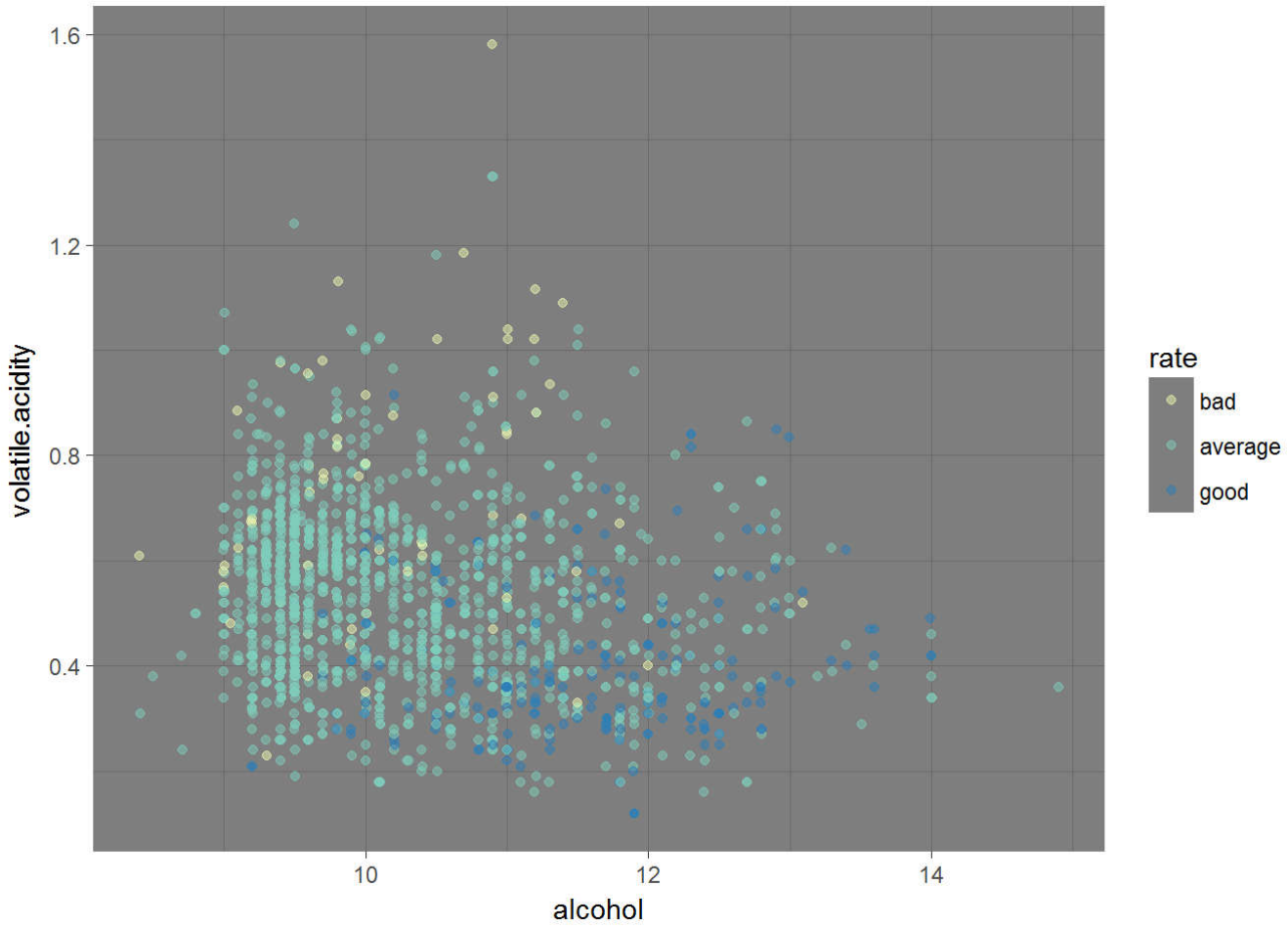
It seems that wines with higher alcohol percentage and higher sulphates amounts are keen to be better quality.



High alcohol and high log10 of citric acid concentration mostly result with good quality wine.



Higher sulphates and log10 of pH mostly result with good quality wine.



Higher alcohol and lower volatile acidity is tend to produce good quality wine.

When we summarized our findings, we expect to obtain good quality wine with higher alcohol, lower volatile acidity and higher sulphates. We can use these in a mathematical model:

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates,
##     data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7186 -0.3820 -0.0641  0.4746  2.1807
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.61083    0.19569  13.342  < 2e-16 ***
## alcohol            0.30922    0.01580  19.566  < 2e-16 ***
## volatile.acidity -1.22140    0.09701 -12.591  < 2e-16 ***
## sulphates          0.67903    0.10080   6.737 2.26e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6587 on 1595 degrees of freedom
## Multiple R-squared:  0.3359, Adjusted R-squared:  0.3346
## F-statistic: 268.9 on 3 and 1595 DF,  p-value: < 2.2e-16
```

I did not include highly correlated values in the same formula to avoid colinearity problems. Although our great effort, this model explains only %33.46 of quality score of a 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?

Sulphates strengthened effect of volatile acidity and alcohol in terms of quality wine.

Were there any interesting or surprising interactions between features?

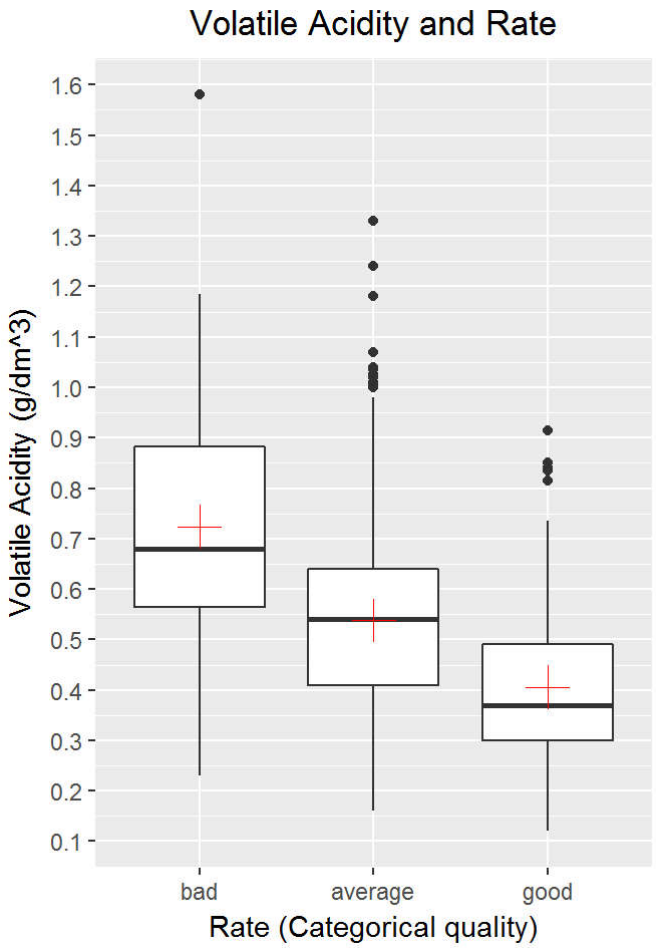
Higher sulphates with lower volatile acidity boosted their effects and help to find better quality wines. Actually only variable correlated with quality variable is alcohol at the first stages of analysis, but sulphates and volatile acidity helps to alcohol to explain quality variable is seem at the later stages of analysis.

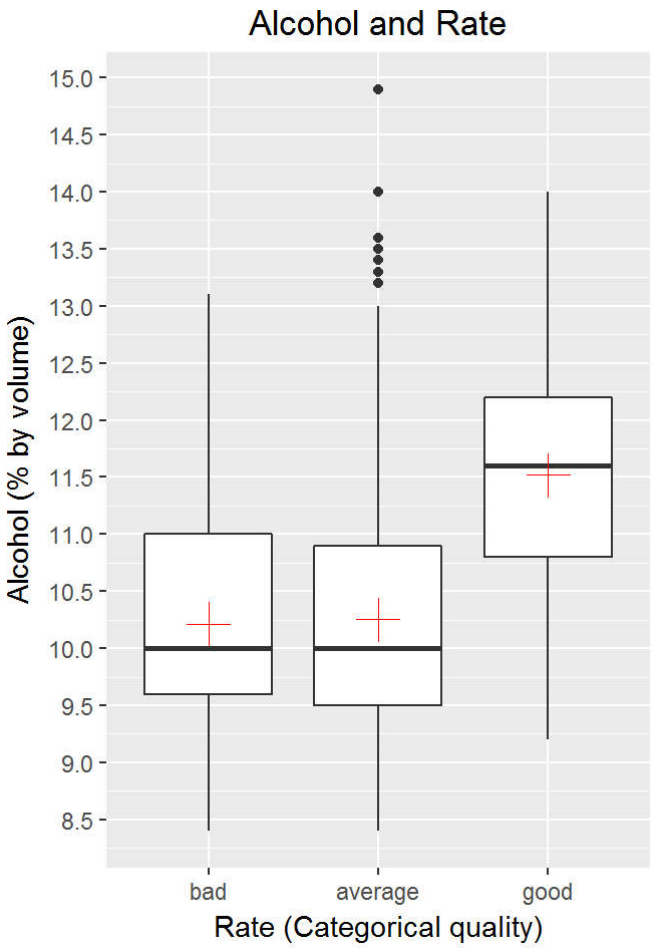
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I created a linear model with the variables alcohol,volatile.acidity and sulphates.Model explained only %33.46 of quality score of a wine. Its strength is being not complex by excluding correlated variables. Limitation is that model is able to explain 1/3 part of changes in quality variable.

Final Plots and Summary

Plot One

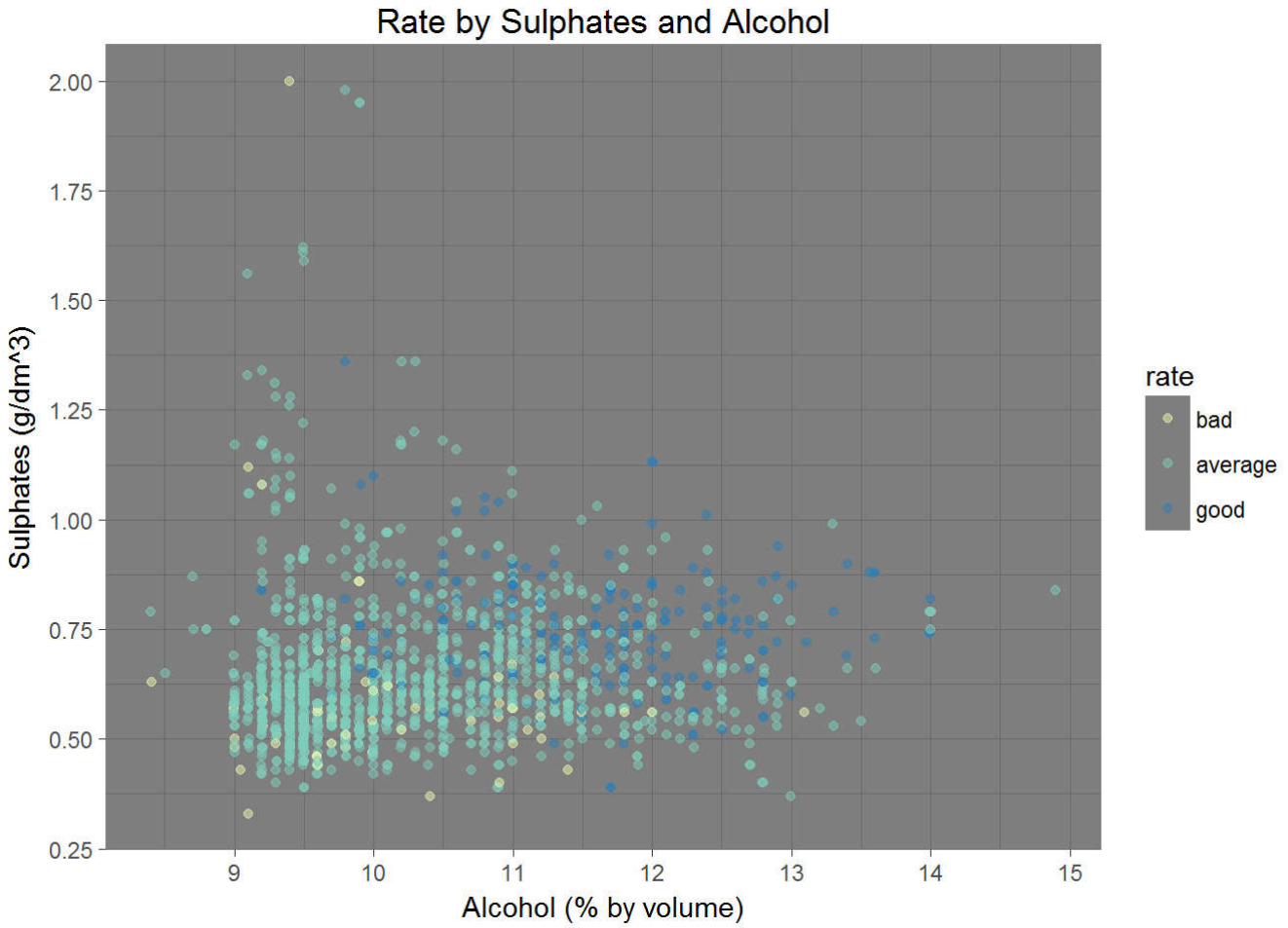




Description One

According the these boxplots, wines labeled as **good** have **lower volatile acidity** value. In addition wines labeled as good have higher alcohol level.In these graphs, there is significant differences between top and bottom lines of boxes.(first and third quantiles). Plus symbols show their means.

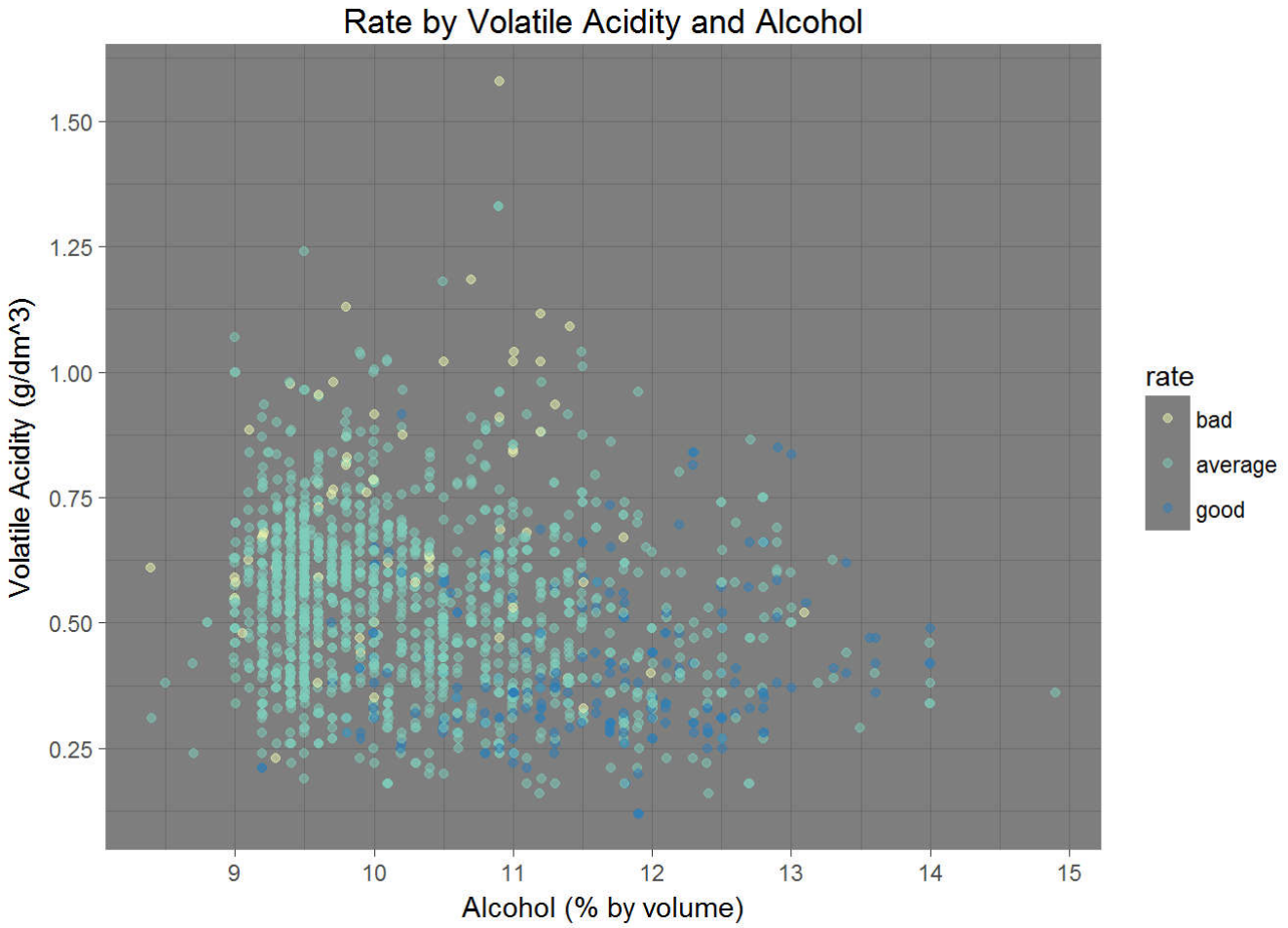
Plot Two



Description Two

According the this plot, It seems that wines with **higher alcohol** percentage and **higher sulphates** amounts are tend to have **better quality**. Wines which have bad quality are placed bottom left of the plot that means wines with lower sulphates amounts and lower alcohol percentage.

Plot Three



Description Three

According the this plot, It seems that wines with **higher alcohol** percentage and **lower volatile acidity** amounts are tend to have **better quality**. These wines are seen at bottom left. Wines which have bad quality are placed top right of the plot that means wines with higher volatile acidity amounts and lower alcohol percentage.

Reflection

This dataset has 11 physiochemical attributes of 1599 red wines and output of each wine is quality evaluated by experts. Firstly, I investigate distributions of all variables, then I tried to find correlation between multiple variables. It takes so much time to get used to variables and finding some correlations.At the last, I had touched all parts of the data.

I used a correlation matrix to see whole parts at first. By using boxplots and scatterplots investigate correlation between variables more detailed. After that, create multivariable plots by using these findings.

Result of all these analysis, wine quality is positively correlated with alcohol, which was not expected at the first stage of analysis. In addition, volatile acidity is negatively correlated with quality.,Lastly, it looks that there is positive correlation betwen sulphates and quality. These three effects explain most parts of quality variable together in less complex way.

Limitation for the dataset is lack of quality variables other than 5 and 6. Also there are most effective attributes of wine which are not included these data set beacuse we could only eplain 1/3 of quality variable. Fermantation duration could be useful data. Also, avarege temperature and rainfall before the harvest would be richen dataset and strength of the possible model.In the next, we can confirm these results by using inferential statistics.