Pingping Chen

> **Tip**: You will see quoted sections like this throughout the template to help you construct your report. Make sure that you remove these notes before you finish and submit your project!

> **Tip**: One of the requirements of this project is that your code follows good formatting techniques, including limiting your lines to 80 characters or less. If you're using RStudio, go into Preferences > Code > Display to set up a margin line to help you keep track of this guideline!

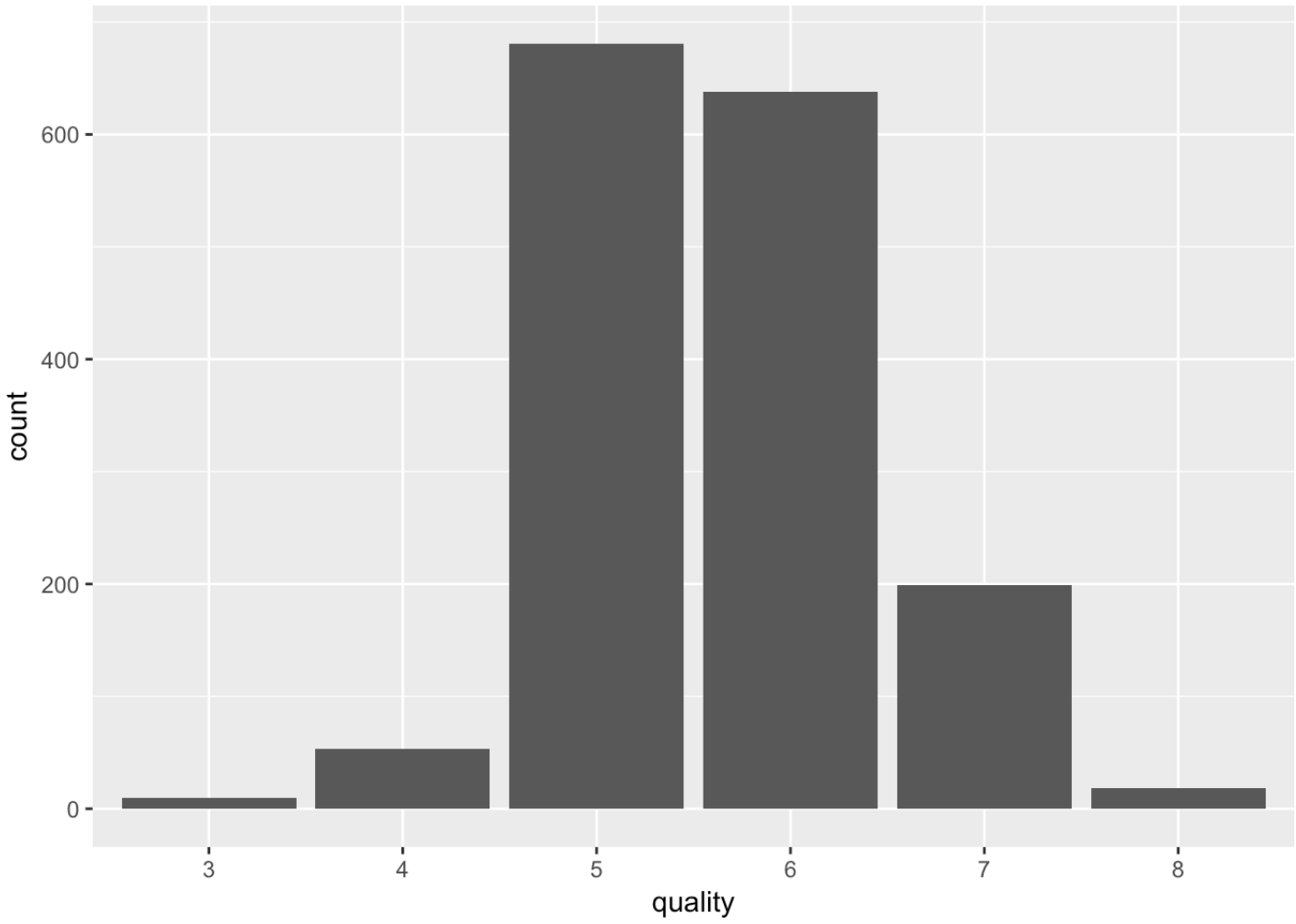
> This report explores Red wine quality with 13 attributes.

Univariate Plots Section

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.07
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 ...
```

```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

The dataset contains 13 variables, with 1599 observations. I changed the quality to factor. The most of Red Wine have the quality 5, 6, and 7.

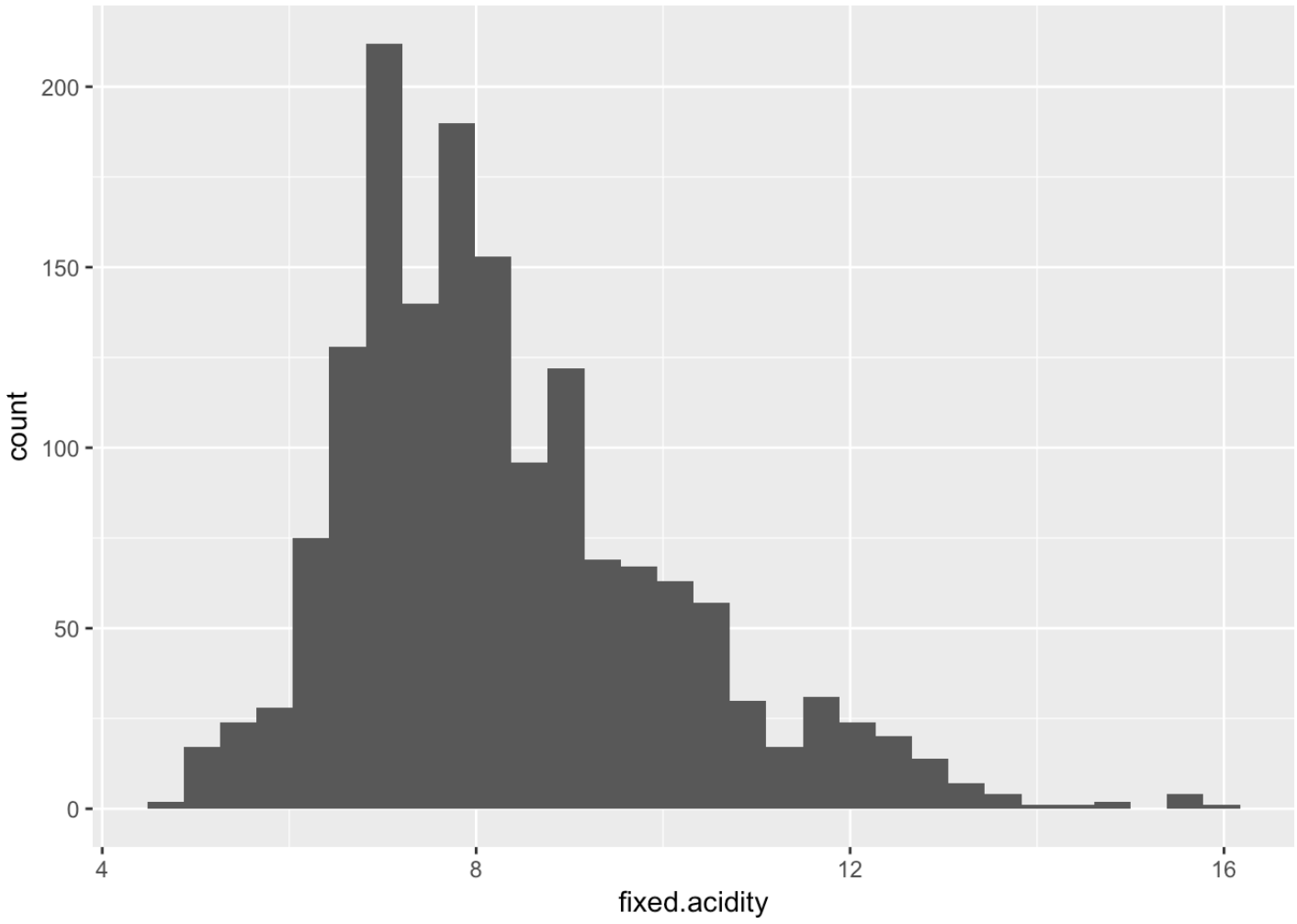


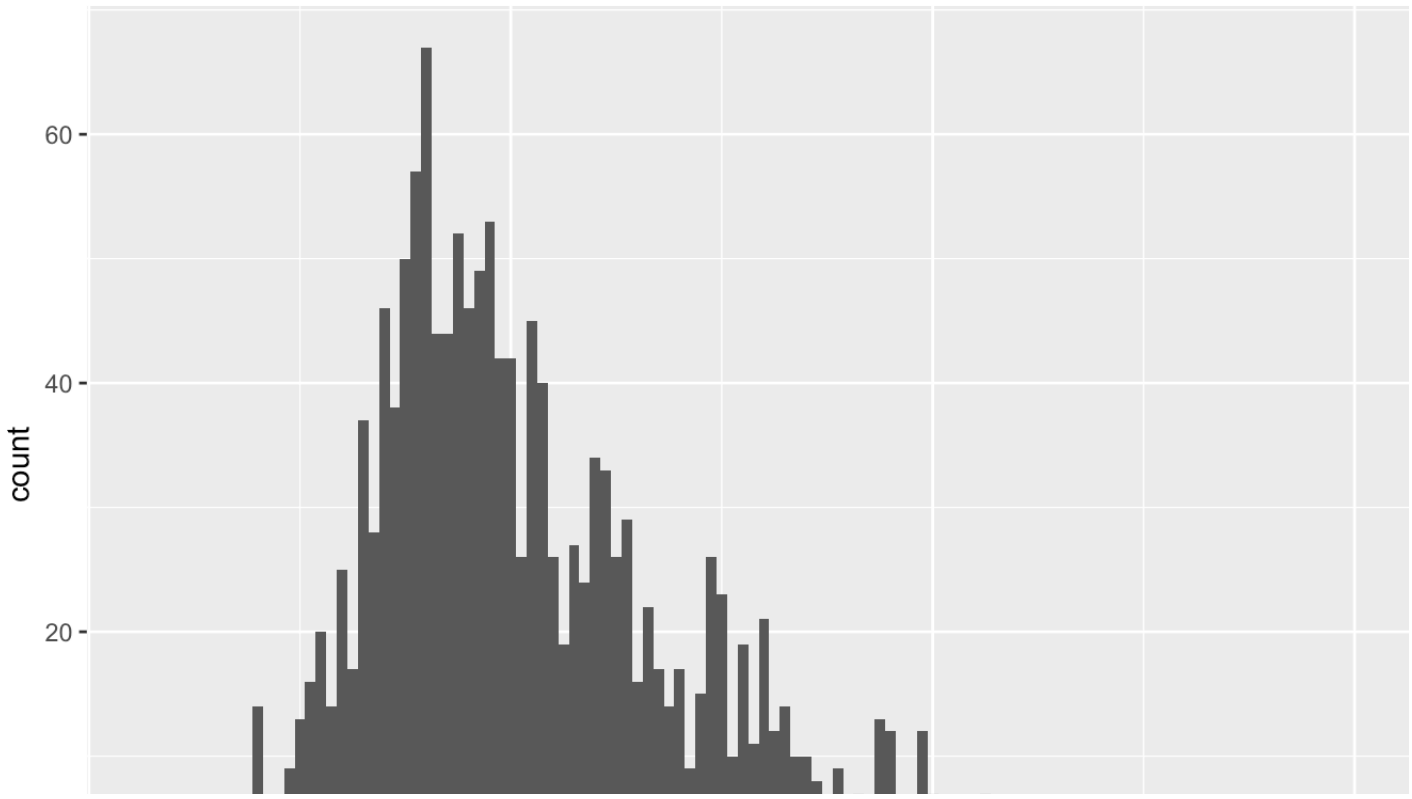

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

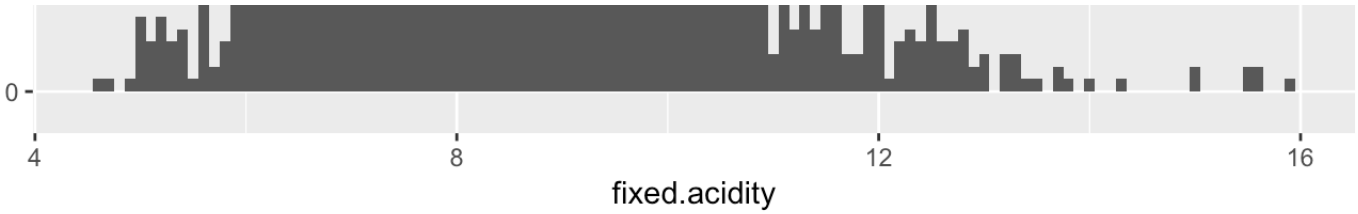
```
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

The distributions of fixed.acidity and volatile.acidity are skewed with right tail. Most Red wine have fixed.acidity between 6 and 9. The volatile.acidity of many red wine falls between 0.2 to 0.8. For the atribute of citric.acid of red wine, it has two peaks, one is at zero and another is 0.49.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

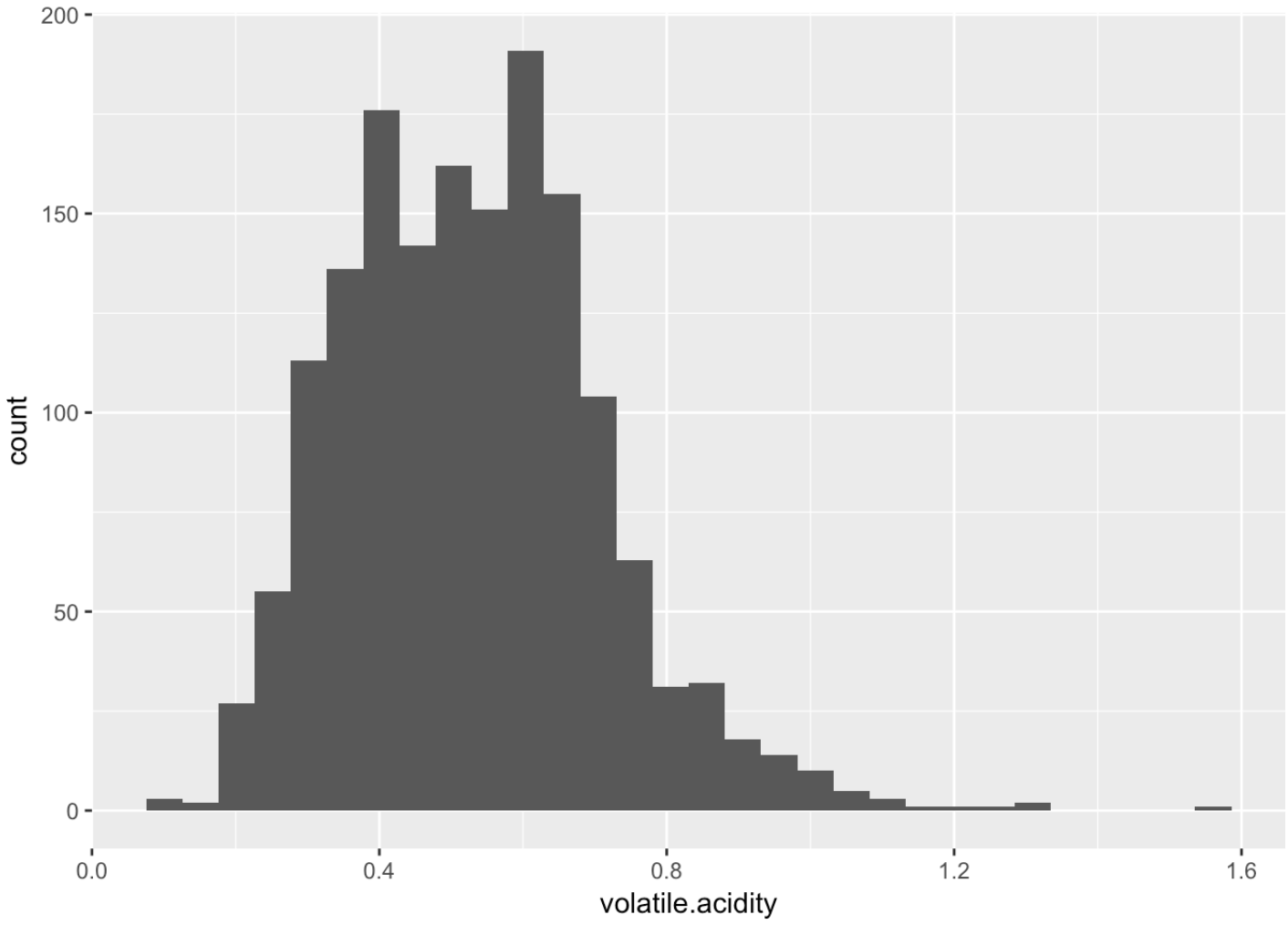

200
150
100
50
0
count
4
8
12
16
fixed.acidity

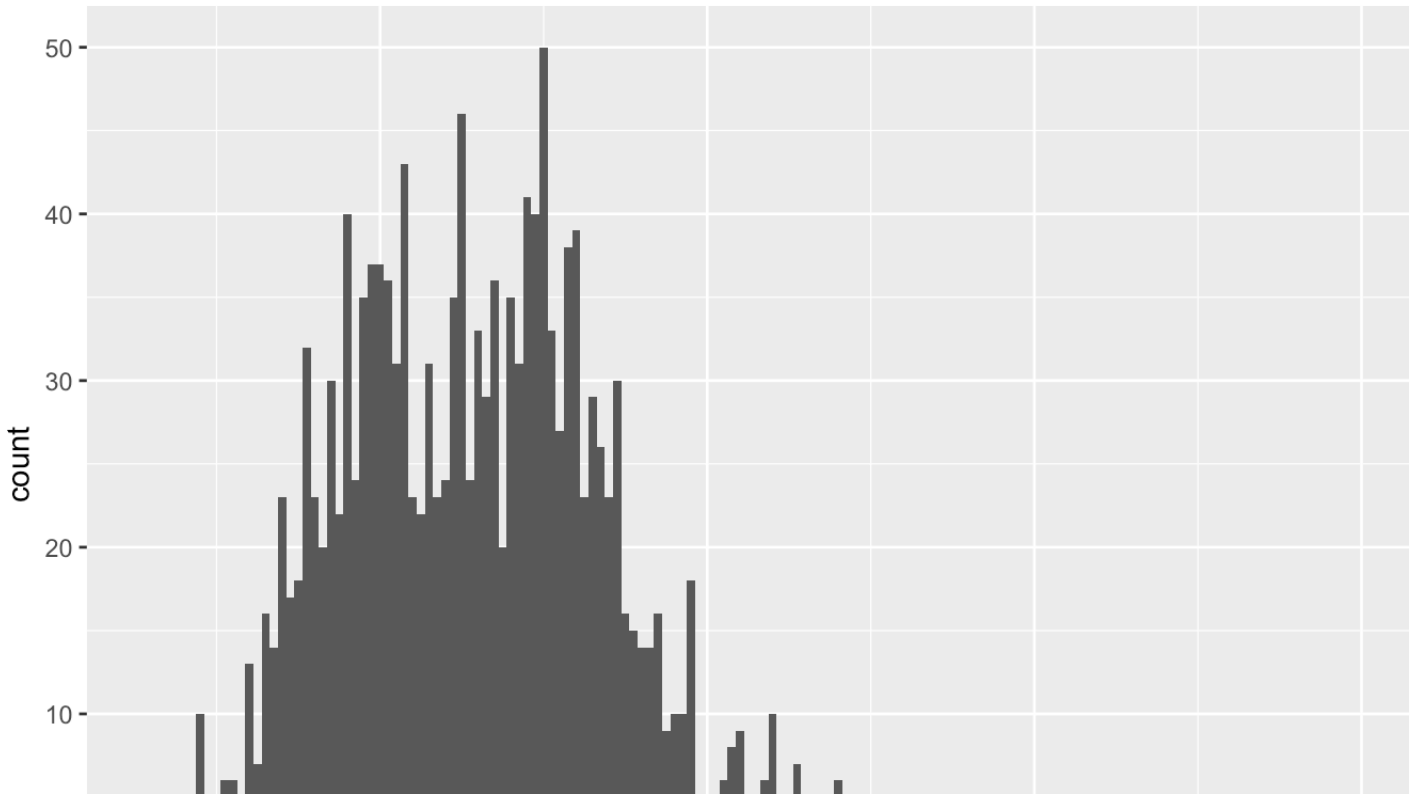

60
40
20
count

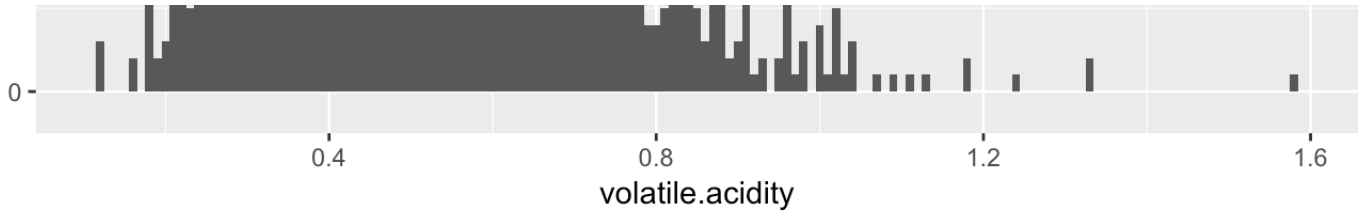



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

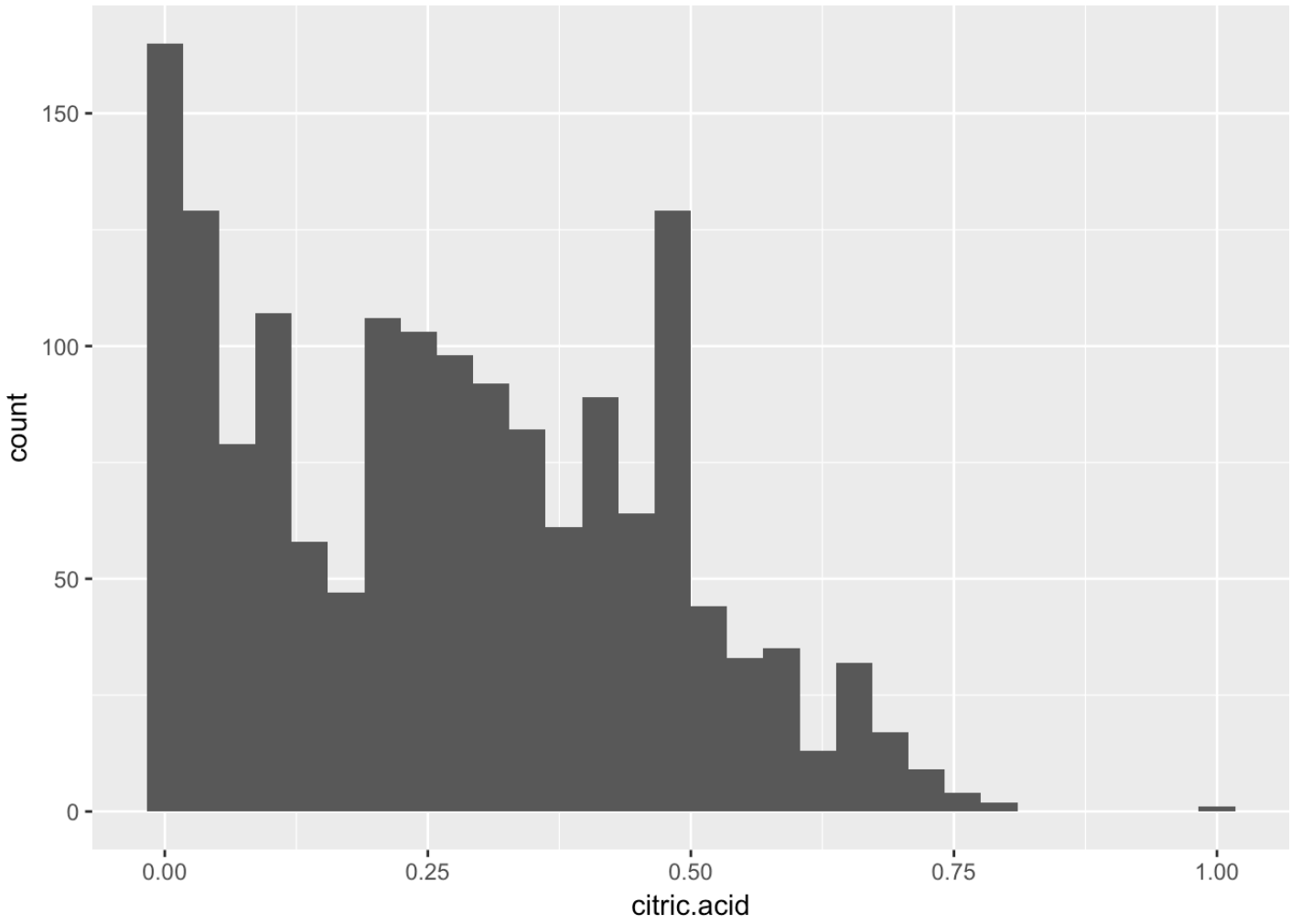

200
150
100
50
0
count
0.0
0.4
0.8
1.2
1.6
volatile.acidity

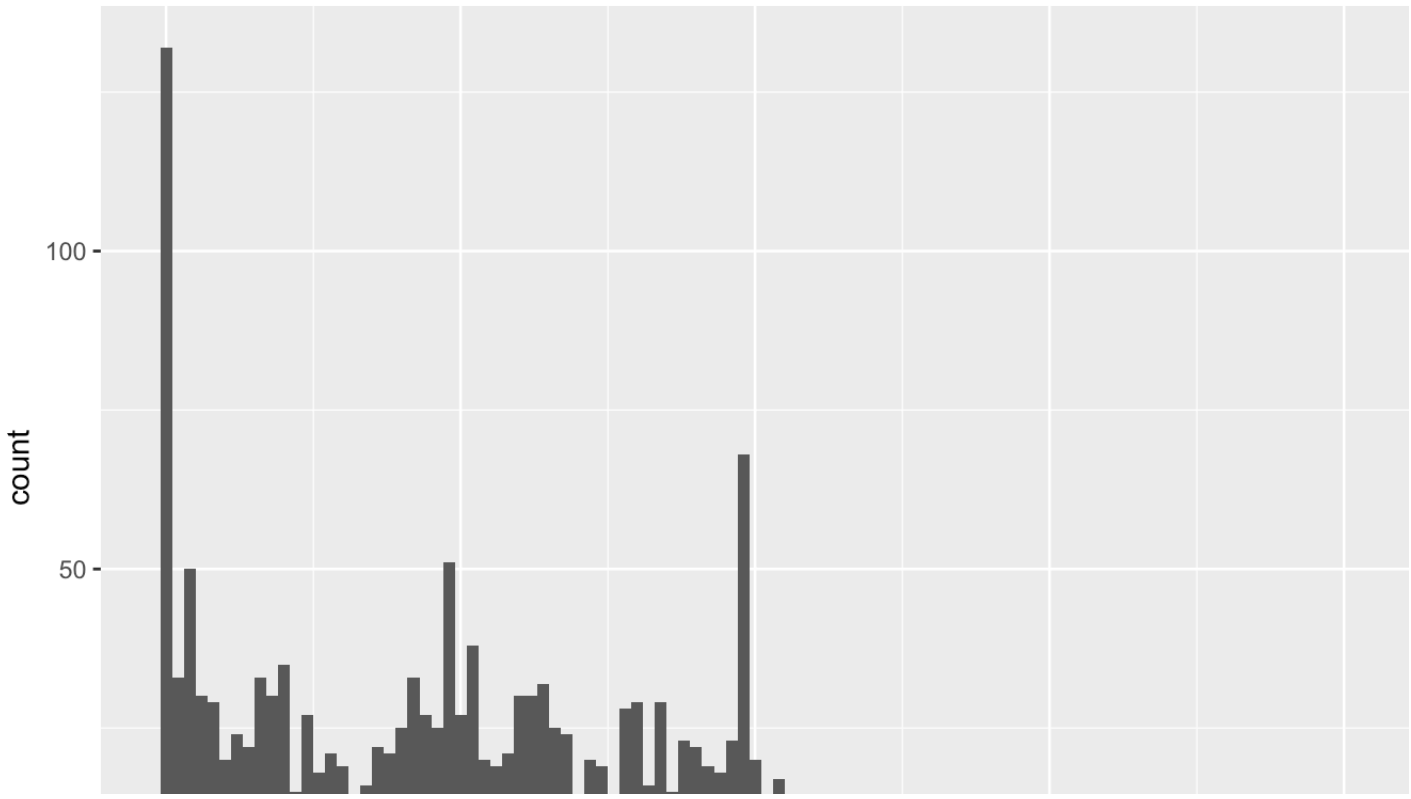

50
40
30
20
10
count

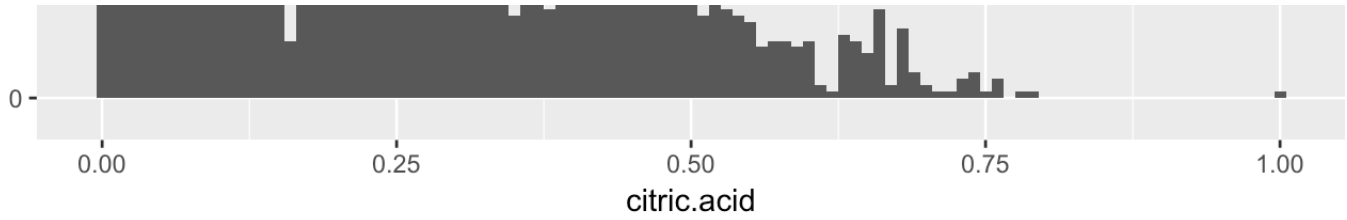



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

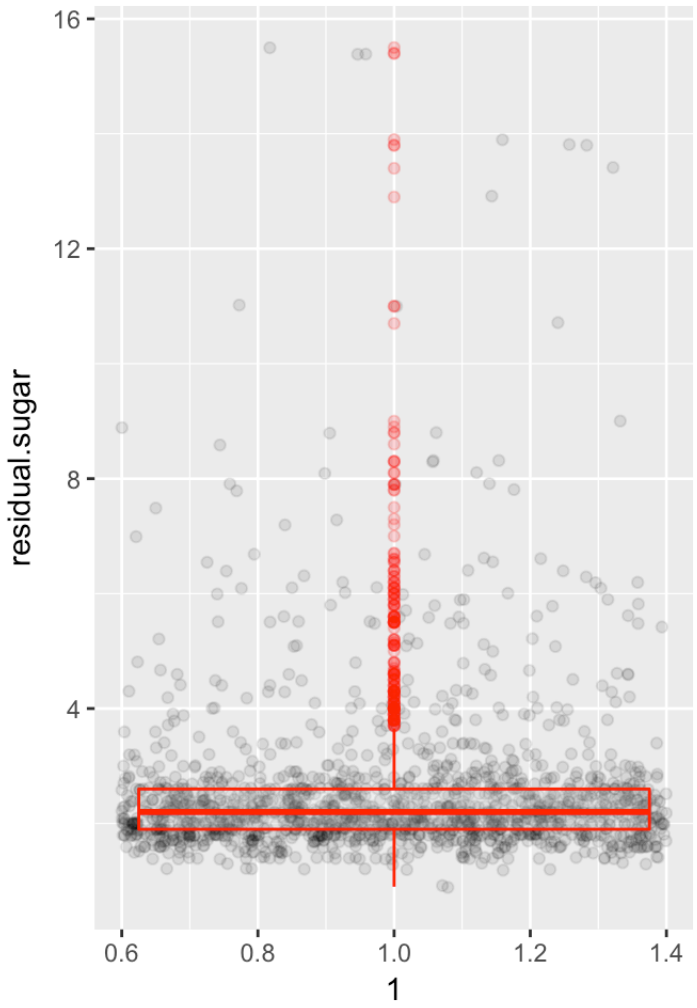
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

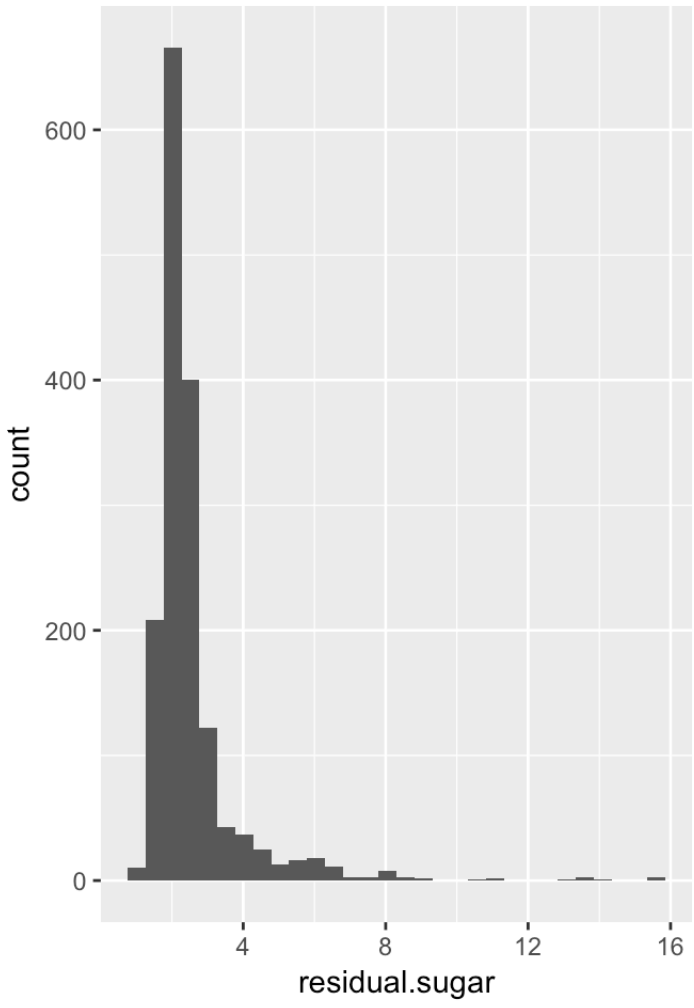

150
100
50
0
count
0.00
0.25
0.50
0.75
1.00
citric.acid


100
50
count

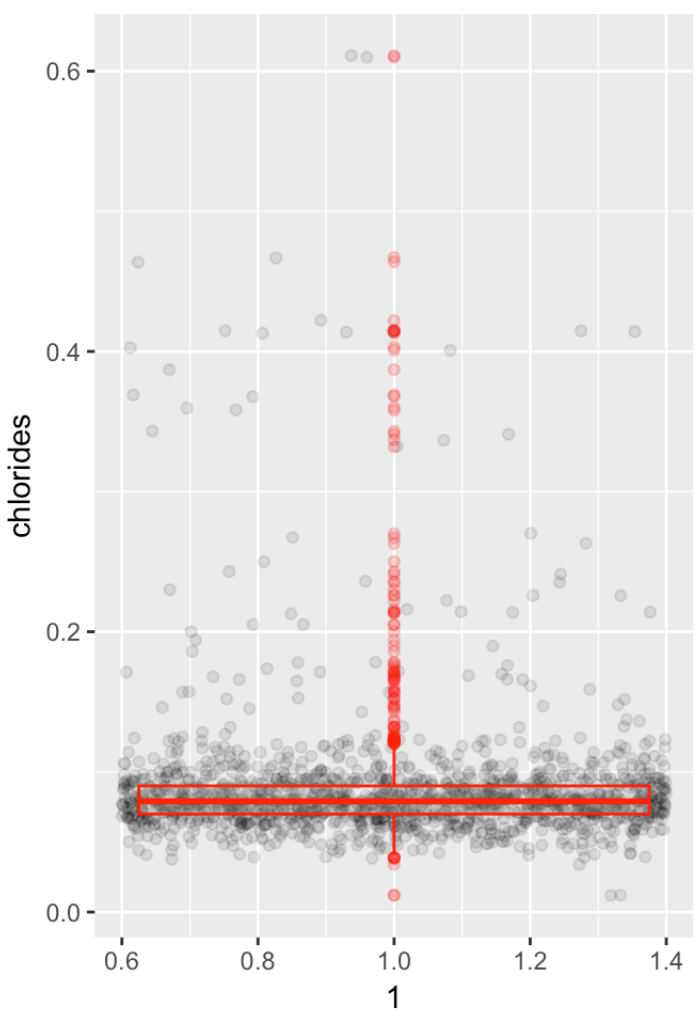


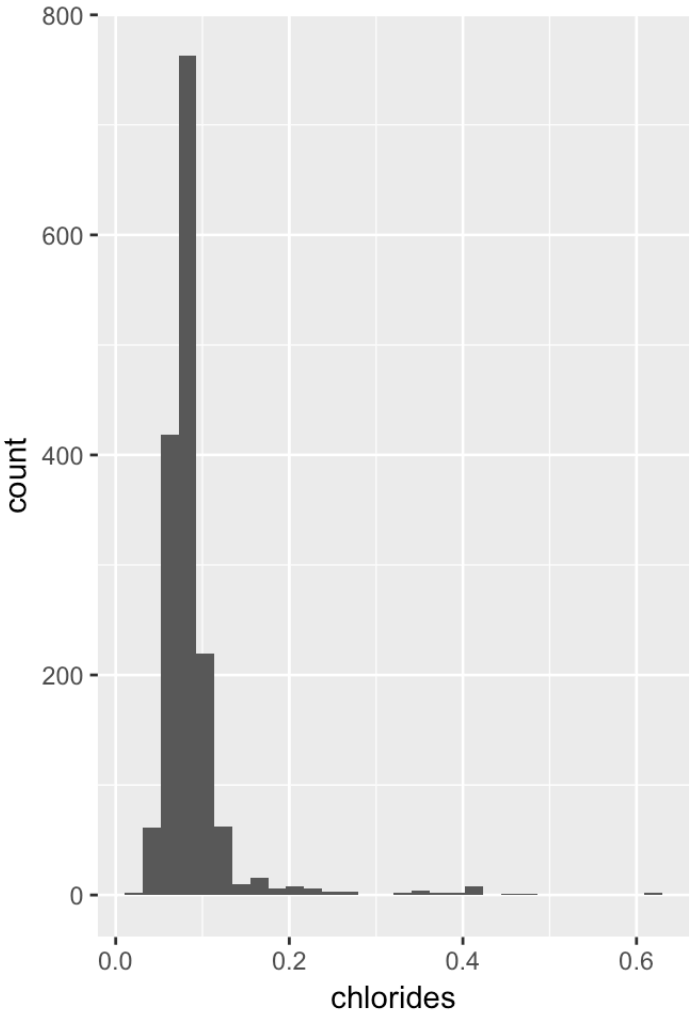

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```



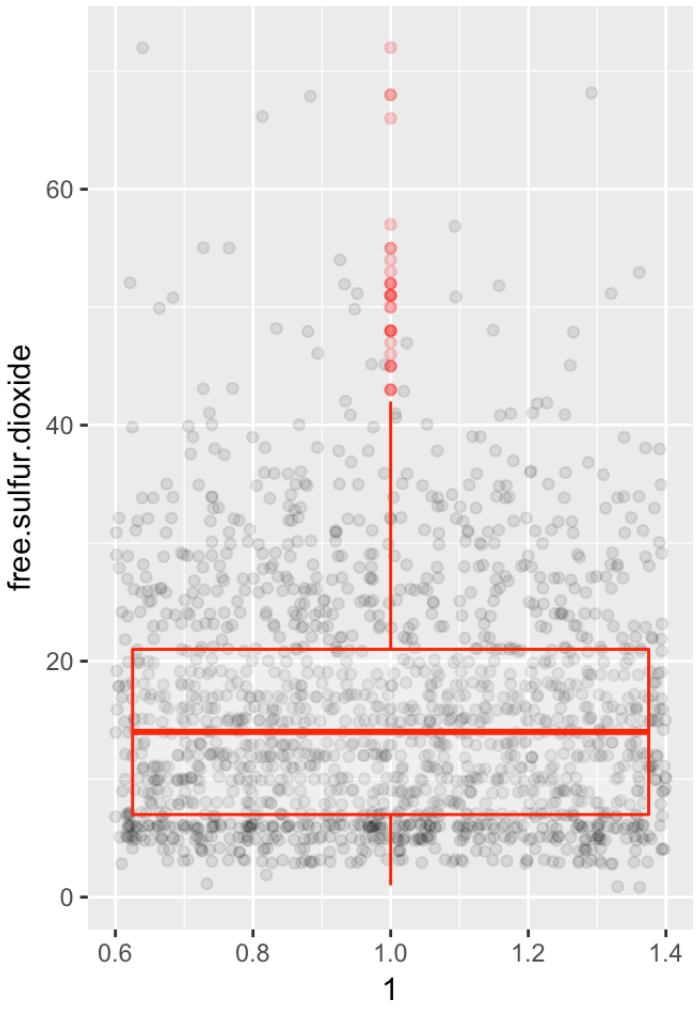


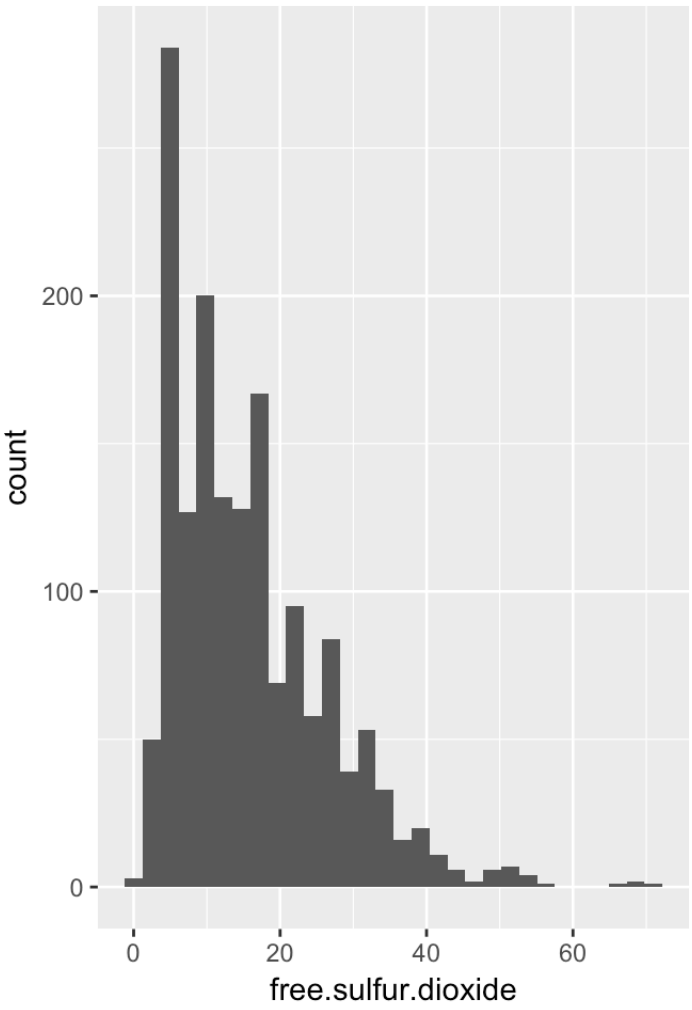

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```



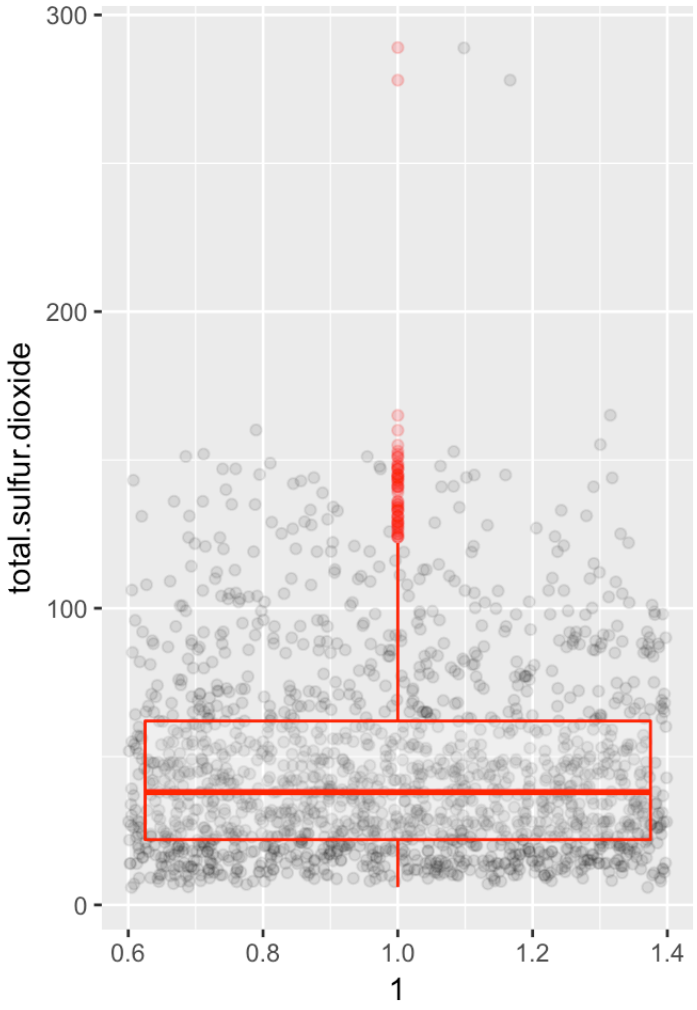


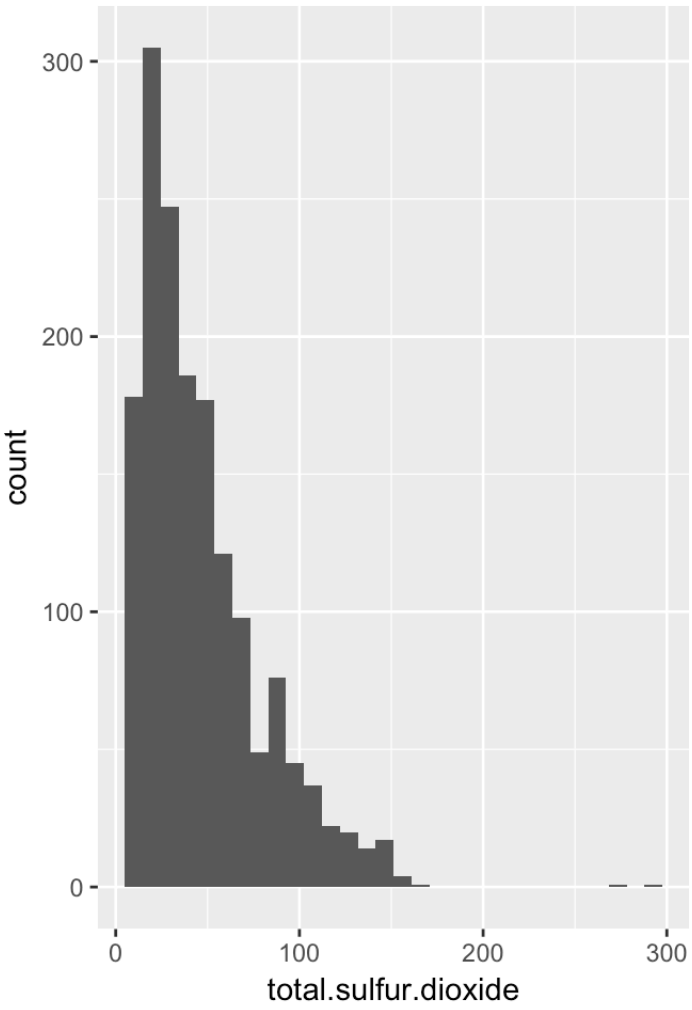

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```



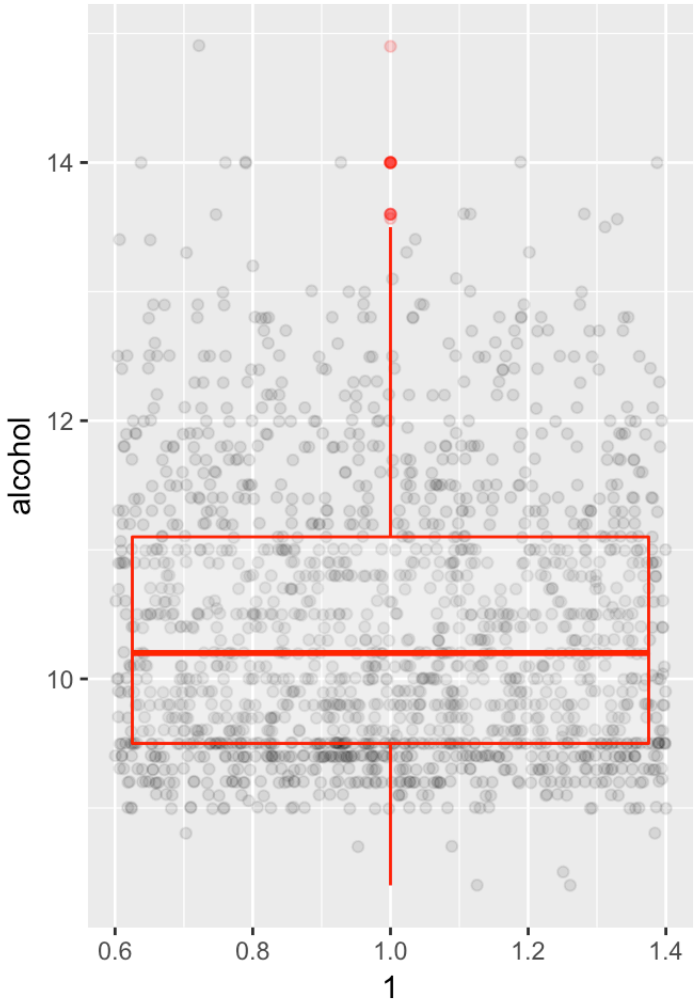


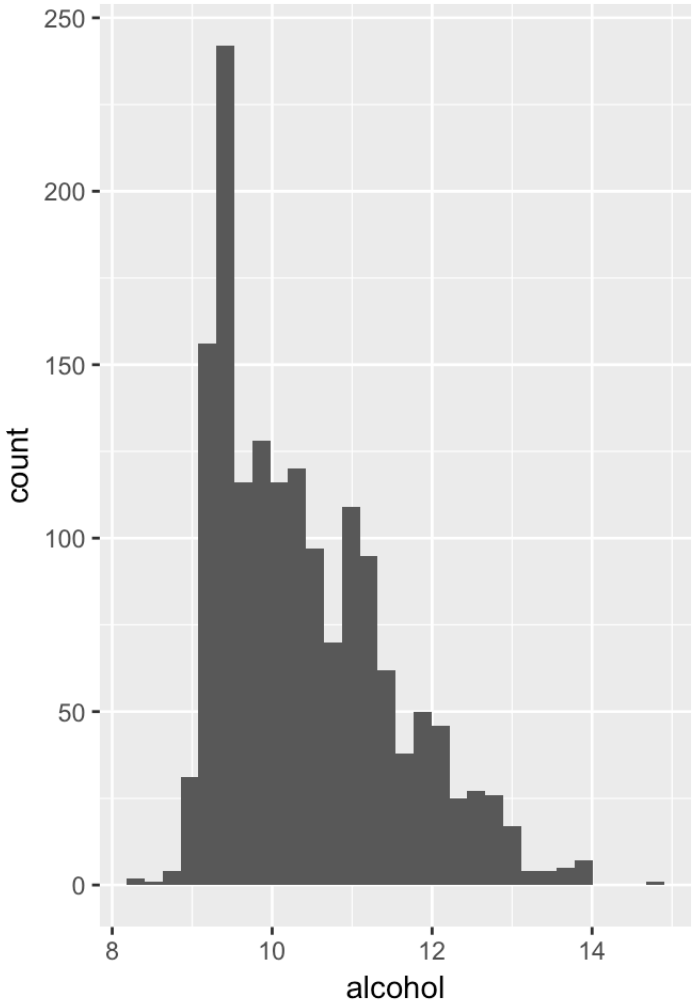

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```





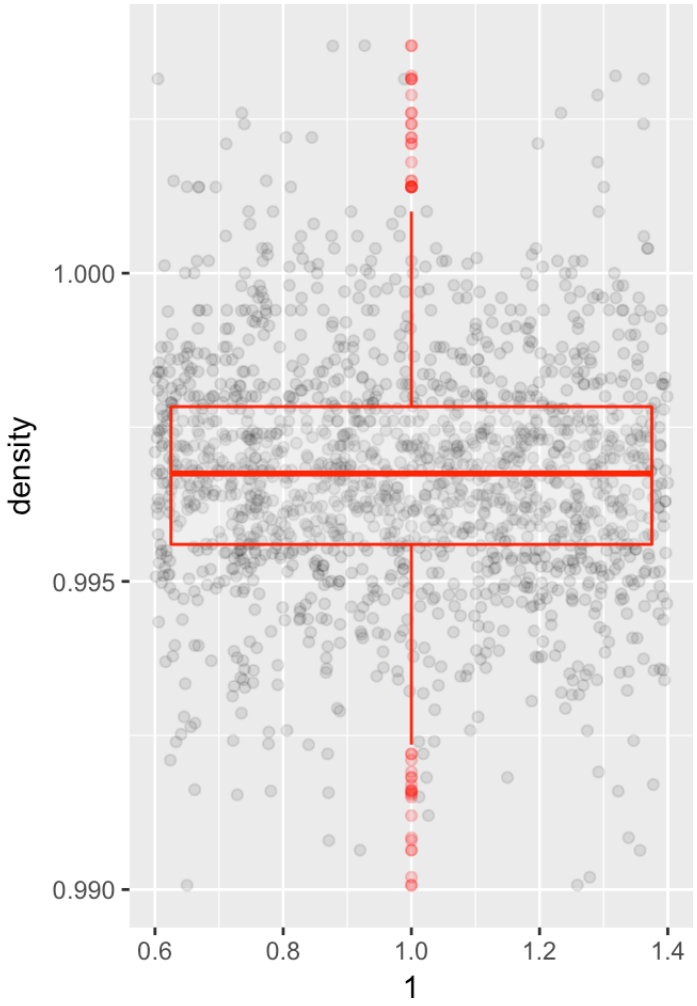

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

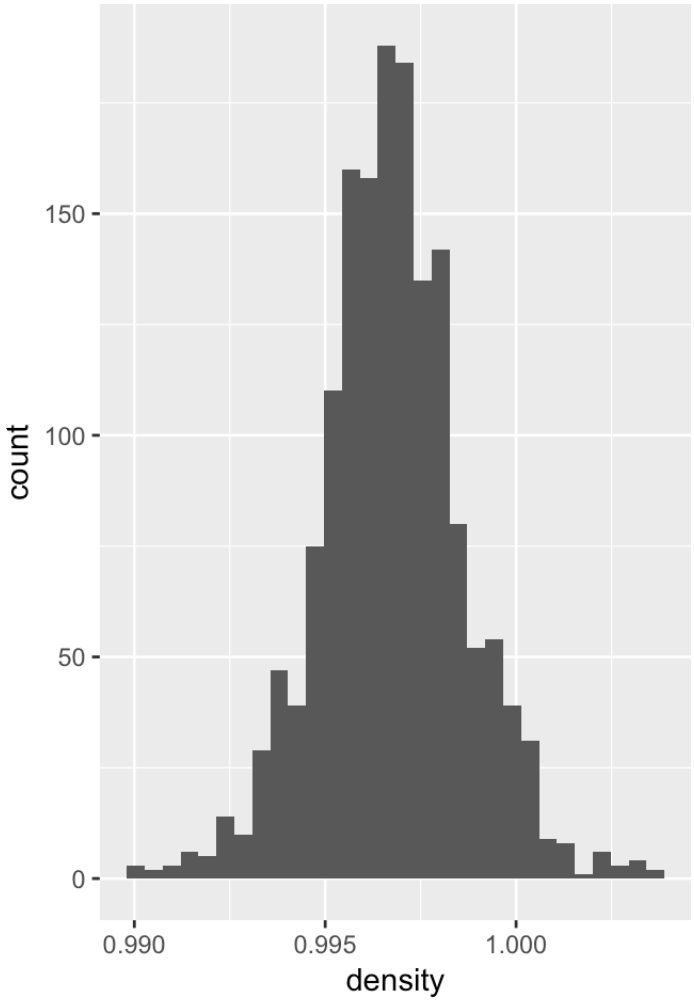




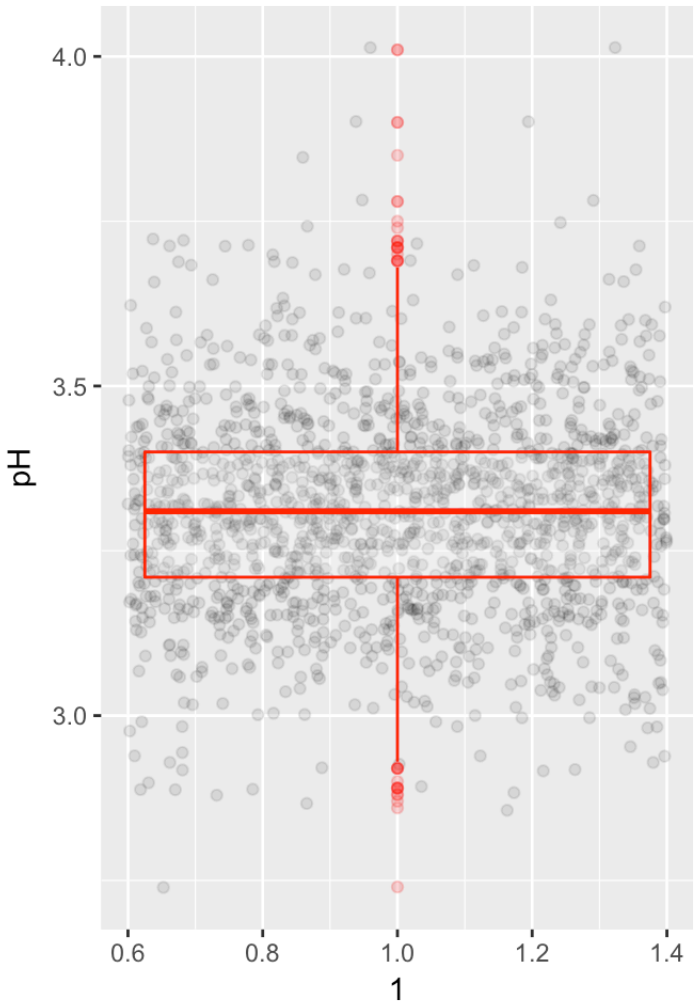

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

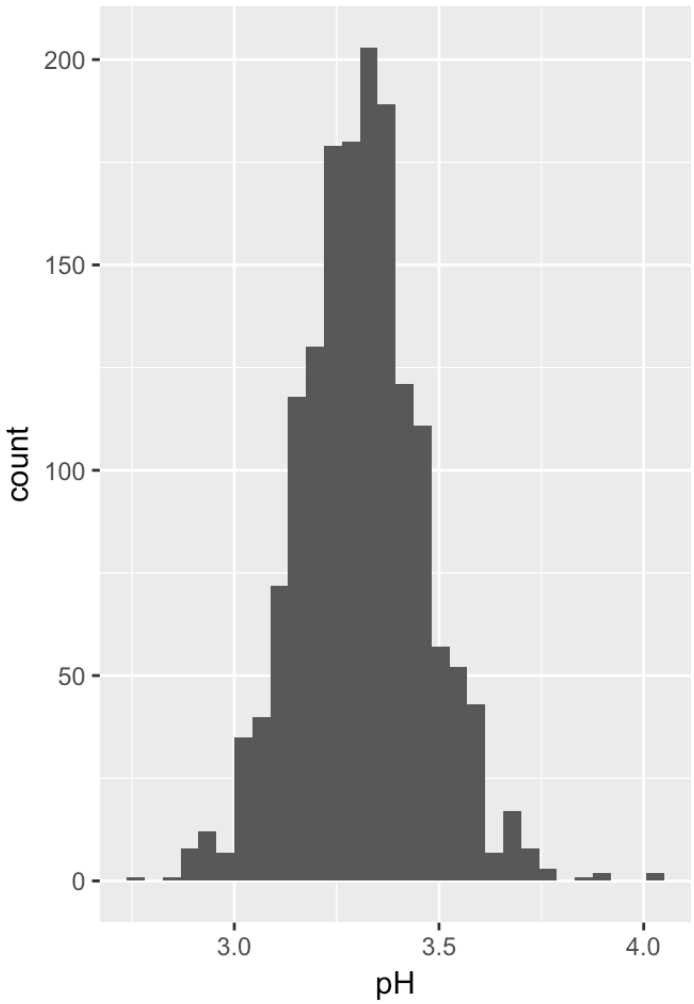
The distribution of residual.sugar centered at some values and with a long right tail. And most red wines have residual.sugar between 1.4 and 3. The third quantile is 2.6, but the maximum of residual.sugar is 15.5 which indicates the distribution is highly right skewed. The distribution of chlorides has the similar shape with residual.sugar's and majority of red wines have chlorides less than 0.09. The third quantile of free.sulfur.dioxide is 21, but the maximun is 72. Total.sulfur.dioxide has the same situation. The third quantile of total.sulfur.dioxide is 62 and the maximun is 289, the minmum is 6. The variance is huge in the total.sulfur.dioxide. The variation in the alcohol is relative smaller than other variables that we have mentioned, but there are still some outliers. In the modeling section, we could remove the outlier of alcohol is greater than 14.



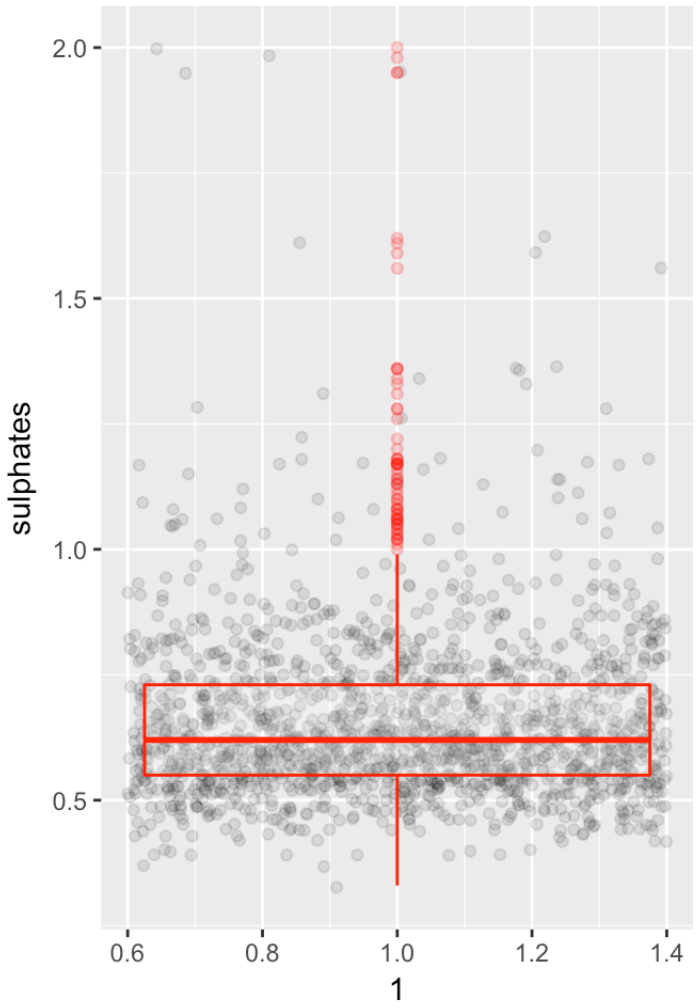


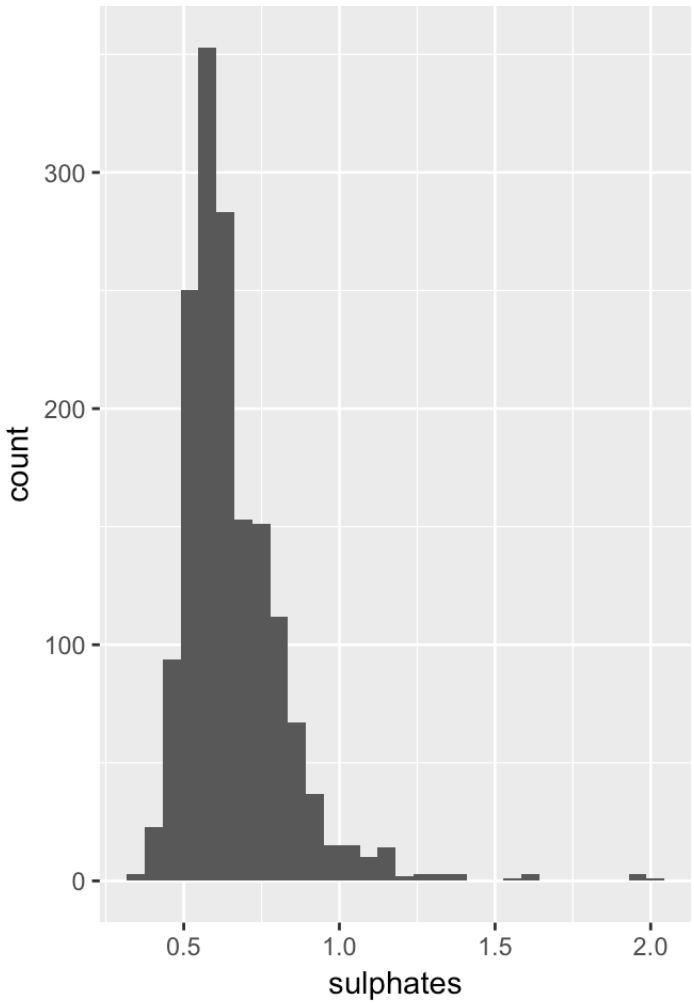

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```






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






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

The distributions of density and PH look like look symmetrically distributed but with outliers.

Univariate Analysis

What is the structure of your dataset?

There are 1599 red wine observations in the dataset with 12 features(fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality). All variables are numerical variables.

Important observations: . Most rea wine have quality 5, 6, and 7. . The median quality for red wine 6 and the max quality is 8. . The histograms of density and PH are approximately noraml distributed. . The distributions of free.sulfur.dioxide, total.sulfur.dioxide and alcohol are highly right skewed which have long right tail.

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

I have not found which one feature is the most important. But free.sulfur.dioxide and sulphates are included in total.sulfur.dioxide.

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

Fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, total.sulfur.dioxide, and alcohol likely contribute to the quality of red wine.

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

I didnot create any new variable because I have not seen there is any relationship between those 12 variables.

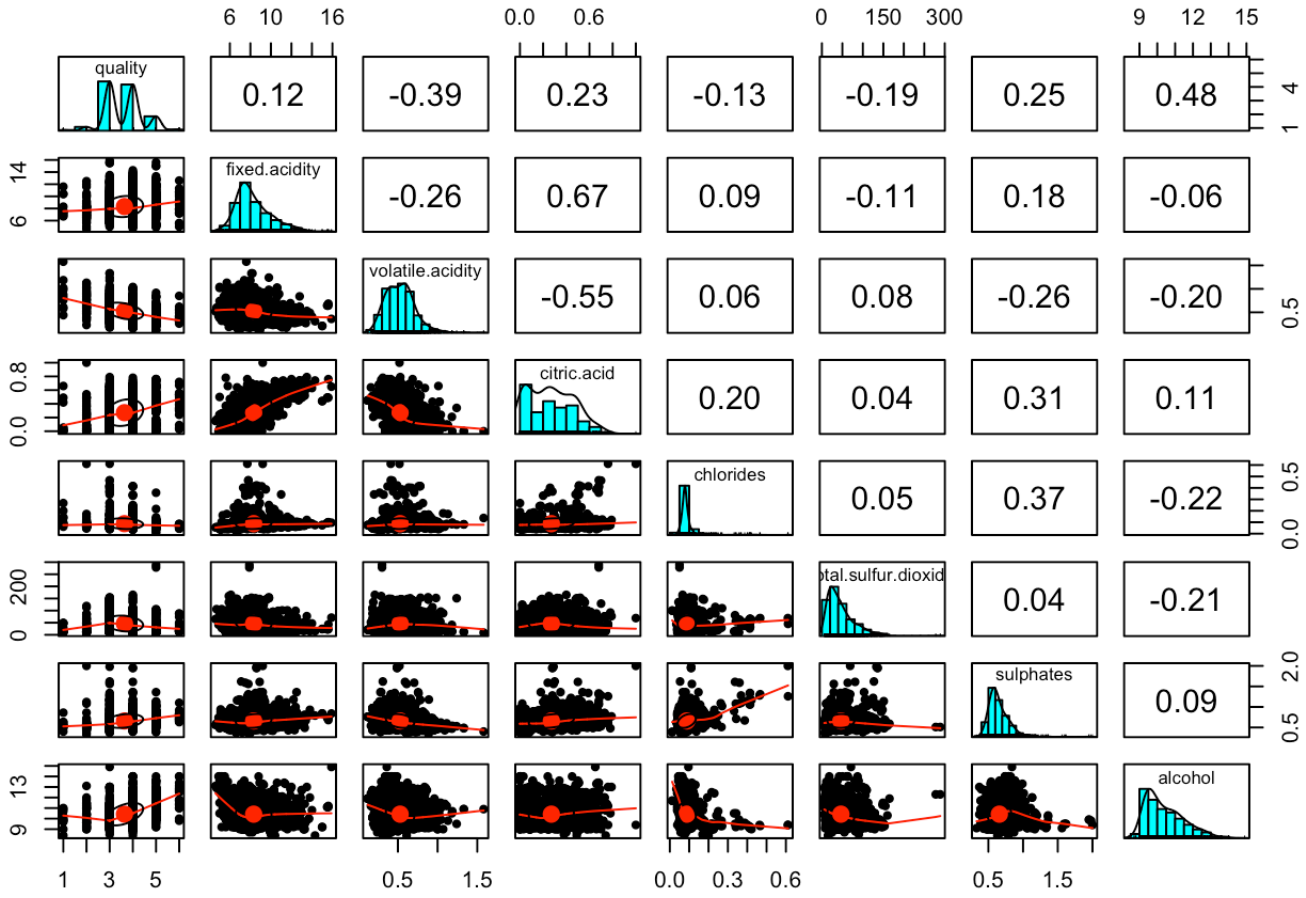
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or

change the form
of the data? If so, why did you do this?

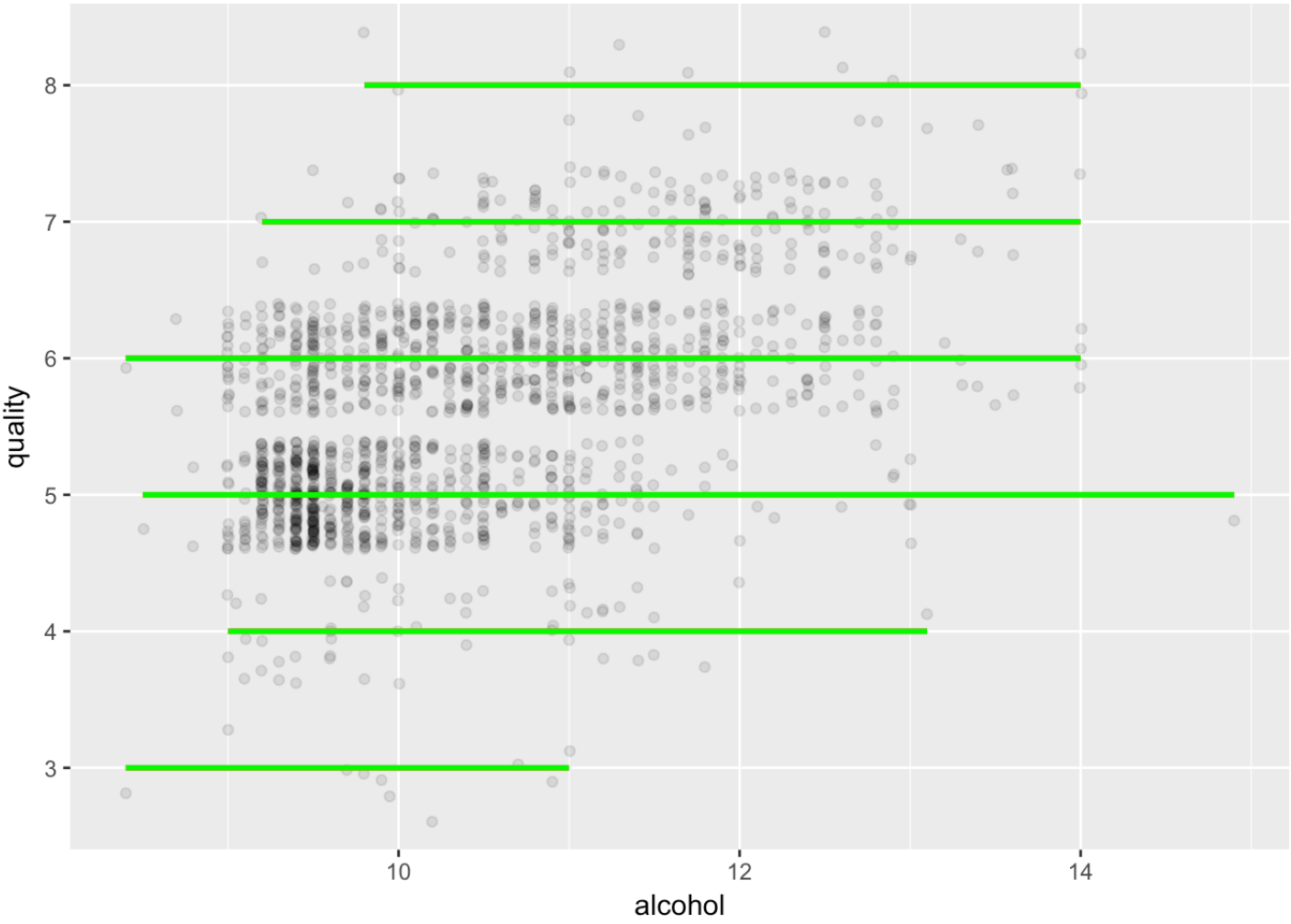
Since the distributions of free.sulfur.dioxide and total.sulfur.dioxide are highly right skewed. I made log-transformation on those two variables. The transformed distribution of total.sulfur.dioxide is seemed approximately normal. Though the transformed distribution of free.sulfur.dioxide is not seemed approximatedly normal, some values of free.sulfur.dioxide with no count have been revealed.

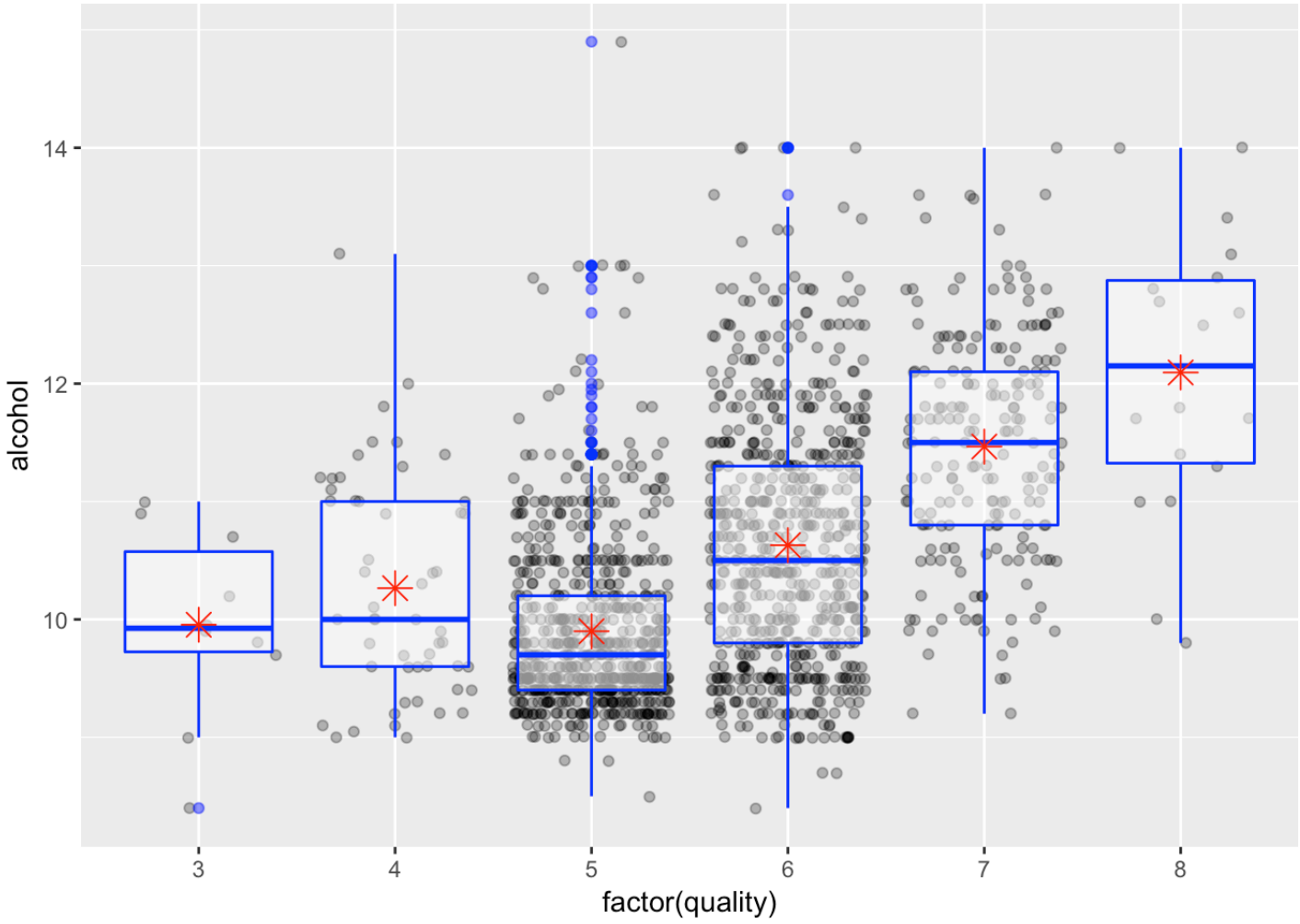
Bivariate Plots Section

From the correlation matrix, we can see that the quality are highly correlated with volatile.acidity and alcohol. Also we could see that fixed.acidity, volatile.acidity, citric.acidity, density, and pH are correlated with each other.



From the correlation plots, fixed.acidity, citric.acid, chlorides, total.sulfur.dioxide, sulphates do not seem to have strong correlations with quality. But citric.acid is correlated with volatile.acidity.


quality
8
7
6
5
4
3
10
12
14
alcohol

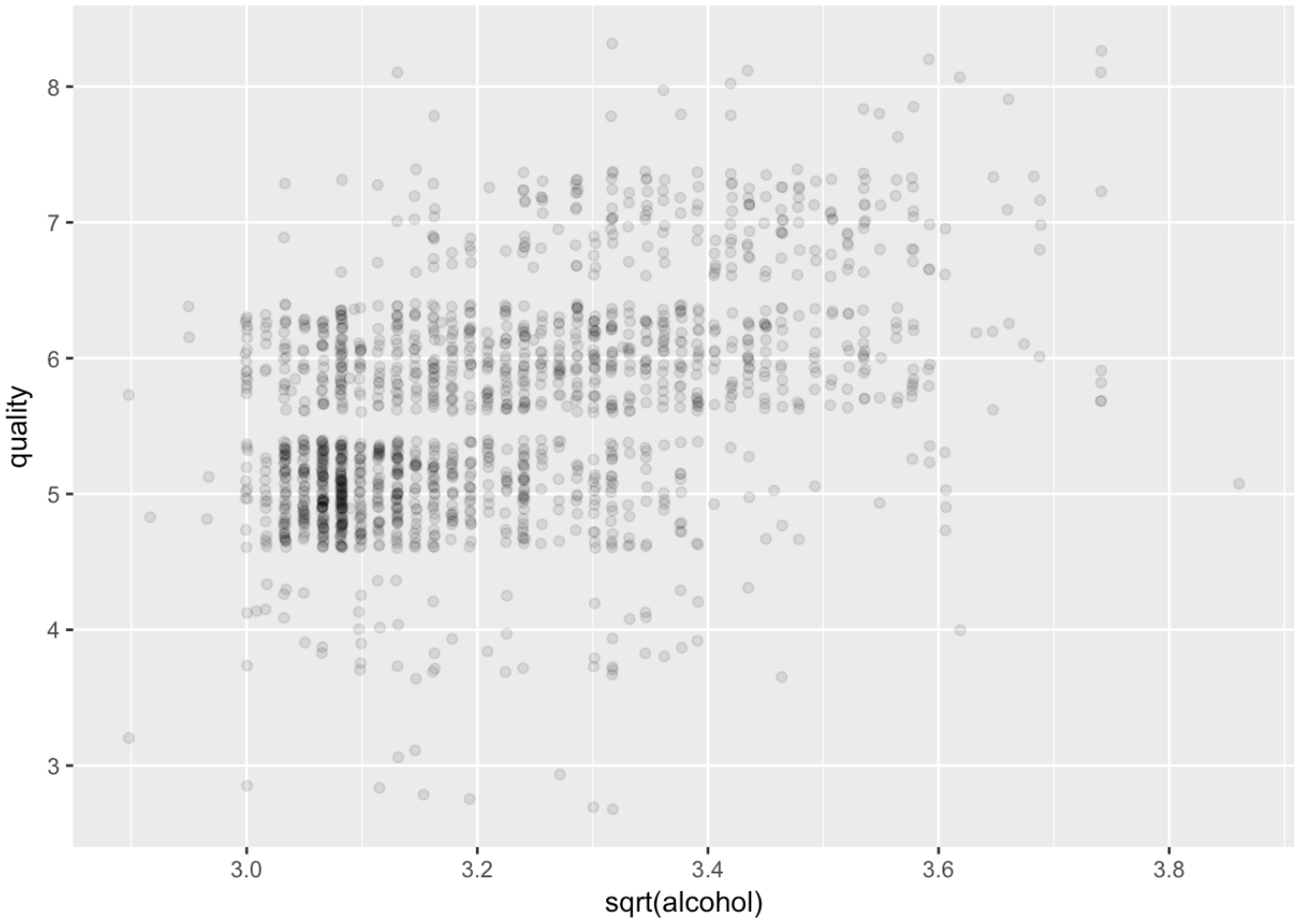


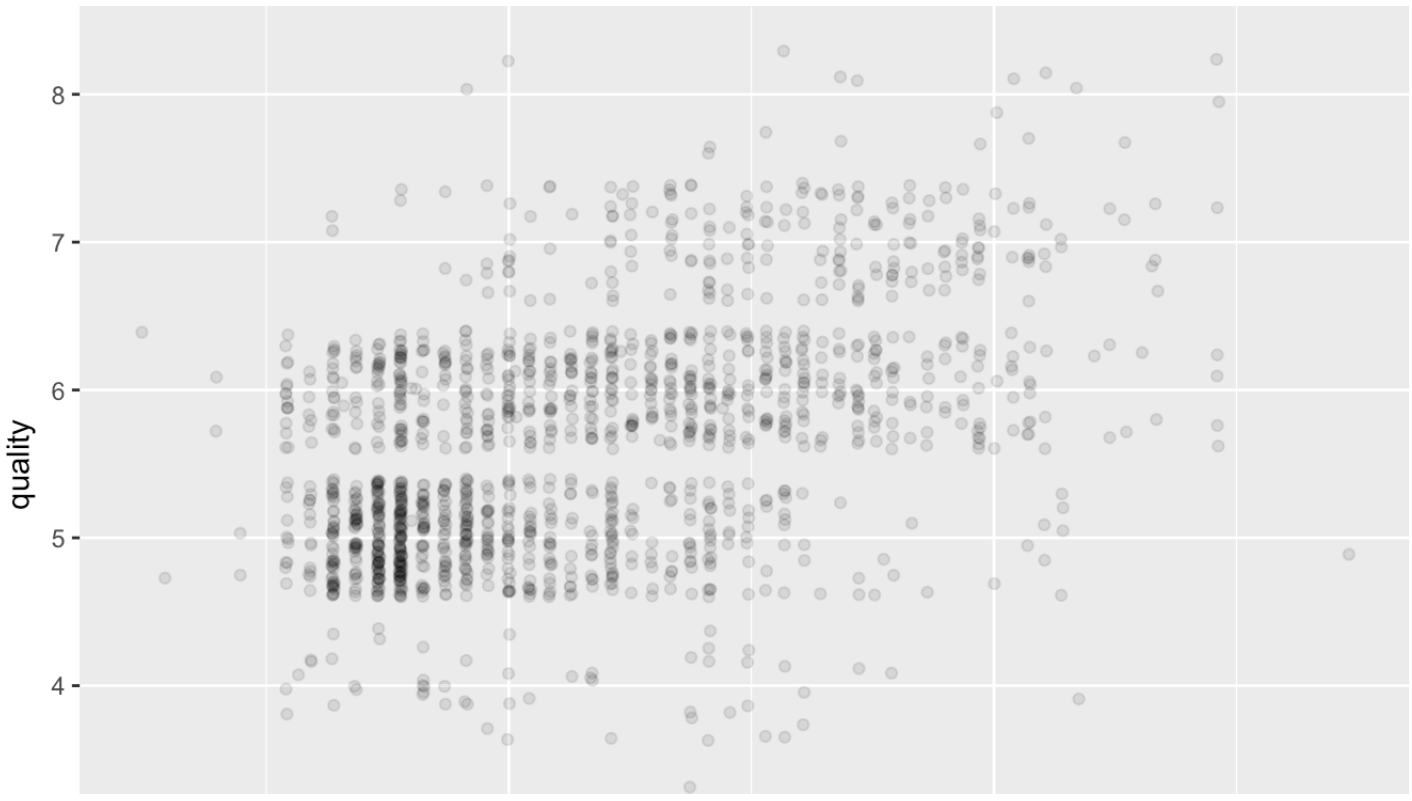
This plot clearly shows that the quality increases as the mean of alcohol increases.

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored
```

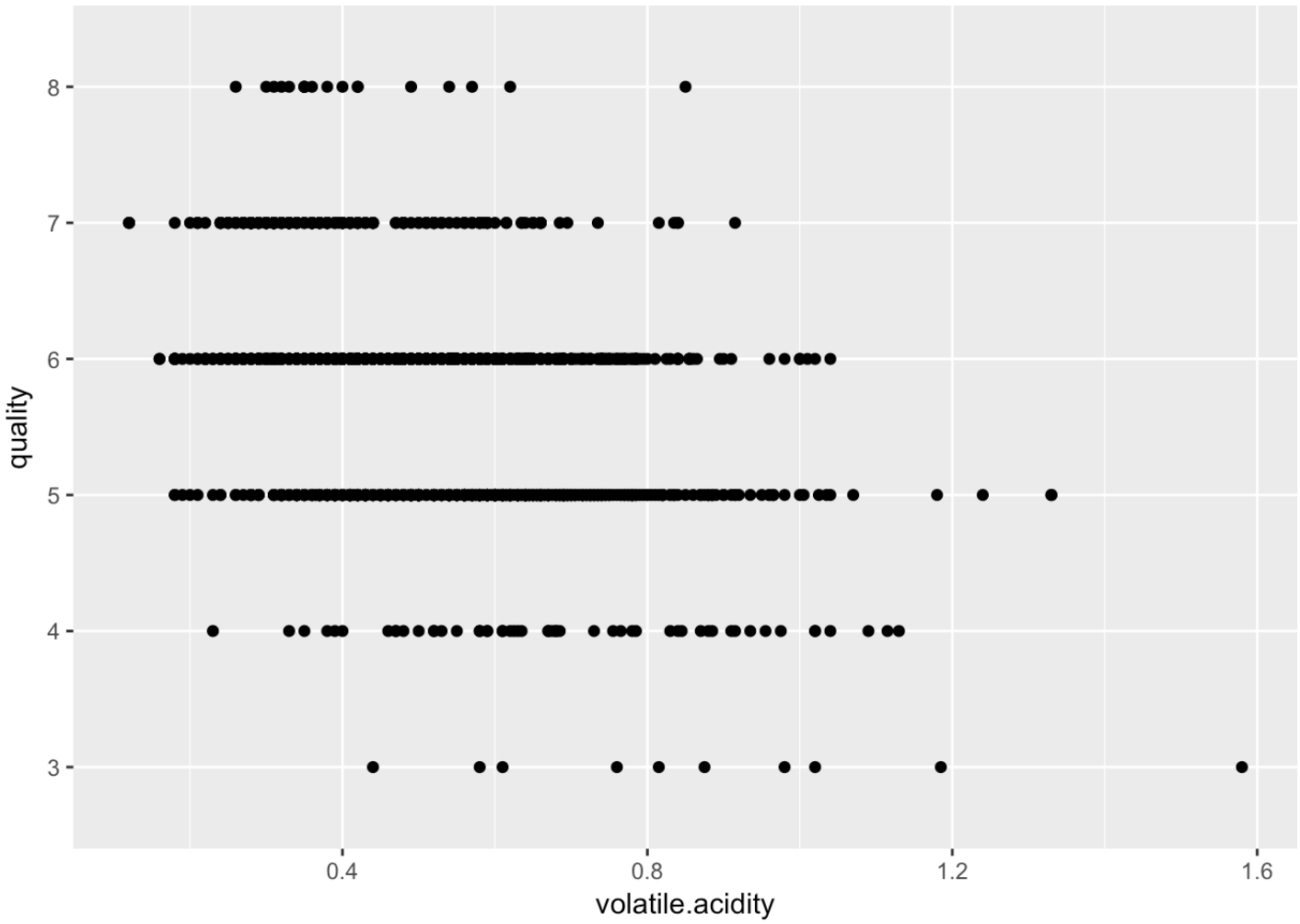
```
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

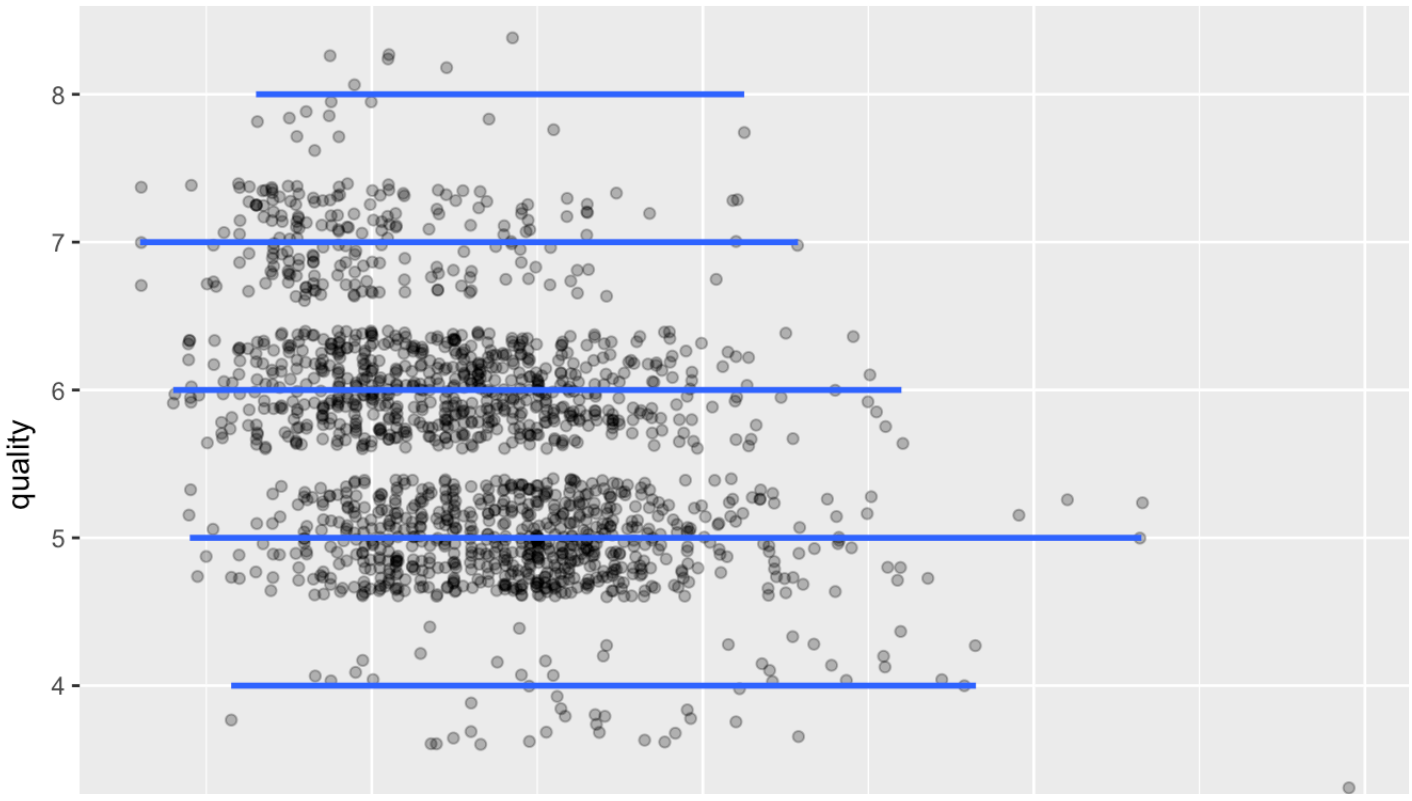
The correlation coefficient between quality and alcohol is 0.48. Though there are many overplotting, I used two different methods to smooth. The green is linear regression line, the red is loess line. The Adjusted R-squared of simple linear regression is 0.2263 which means only 22.63% variation of red wine quality can be explained by alcohol. On the other hand the simple linear regression isn't a good algorithm for red wine dataset.

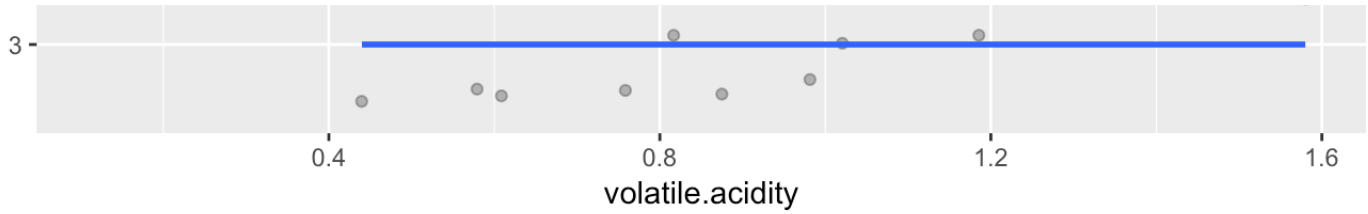

quality
8
7
6
5
4
3
3.0
3.2
3.4
3.6
3.8
sqrt(alcohol)


quality
8
7
6
5
4

Tried to fix overplotting problem, two transformations on alcohol was made. These two plots show that the overplotting cannot be easily fix only through transformations.


quality
8
7
6
5
4
3
0.4
0.8
1.2
1.6
volatile.acidity

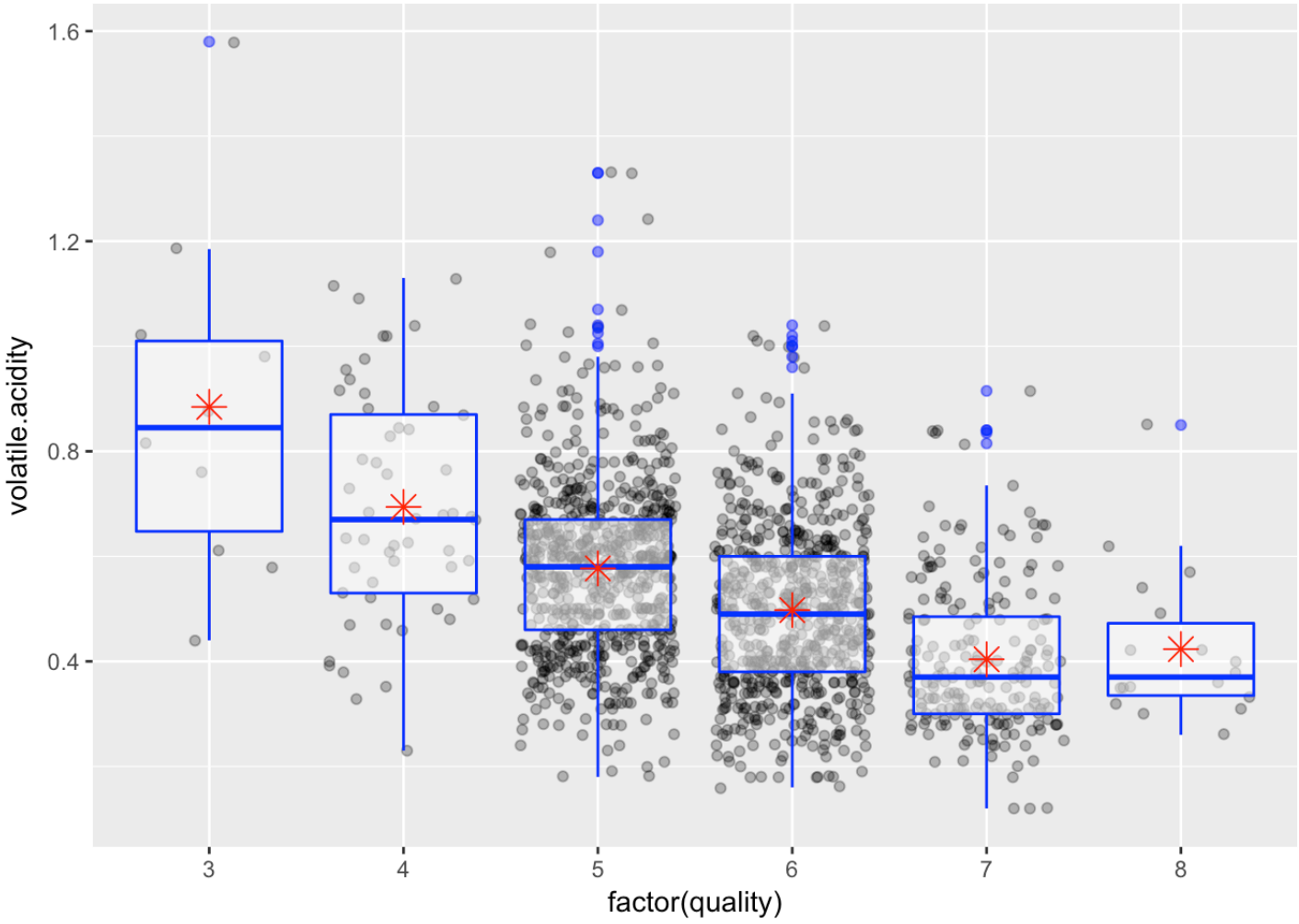

quality
8
7
6
5
4



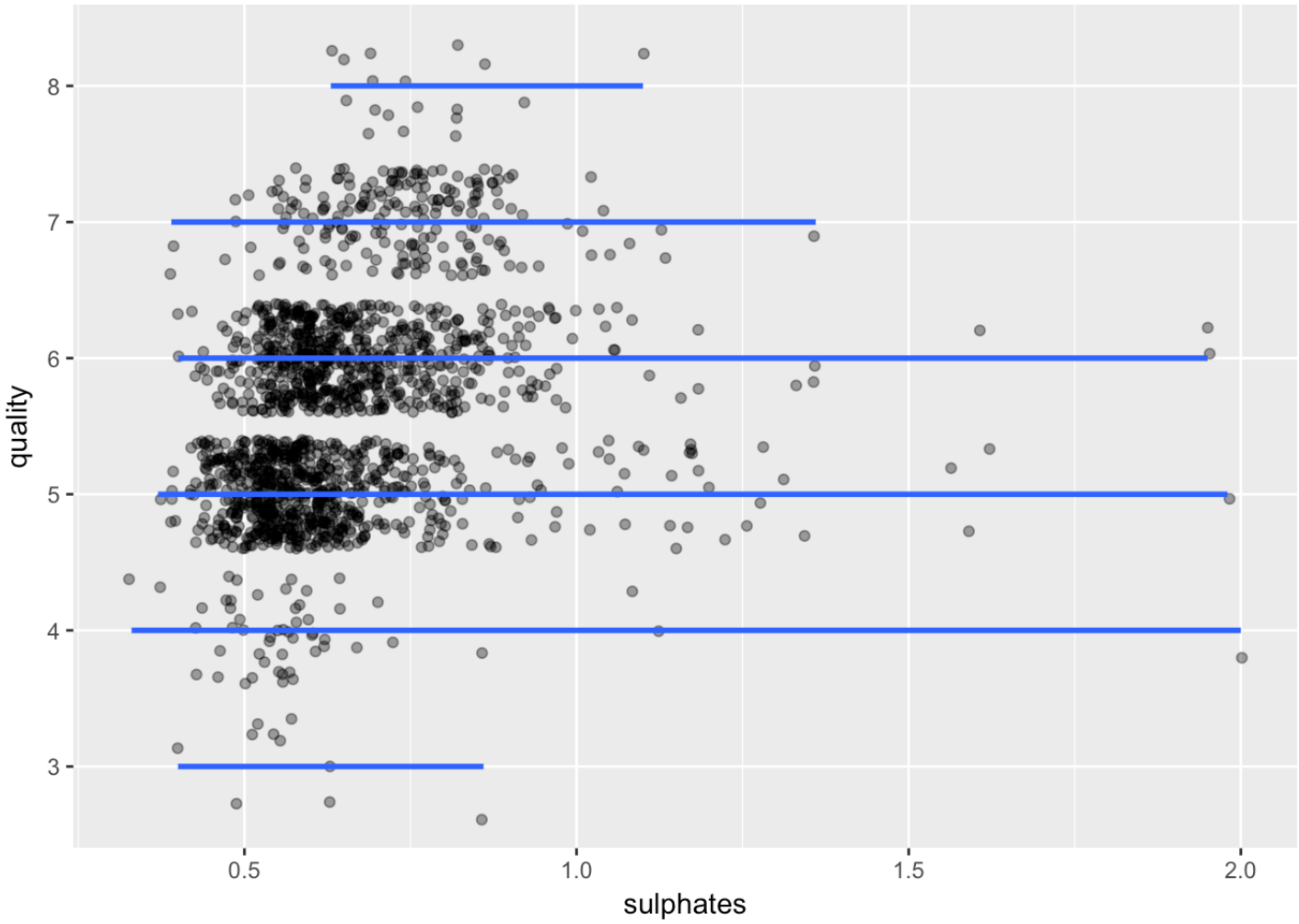

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored
```

```
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

Comparing volatile.acidity to quality, the first plot suffers from overplotting. Most red wines have a volatile.acidity between 0.3 and 0.7. After adding jitter, transparency, let us see the slight negative correlation between volatile.acidity and quality. The correlation coefficient between these two variable is -0.39.



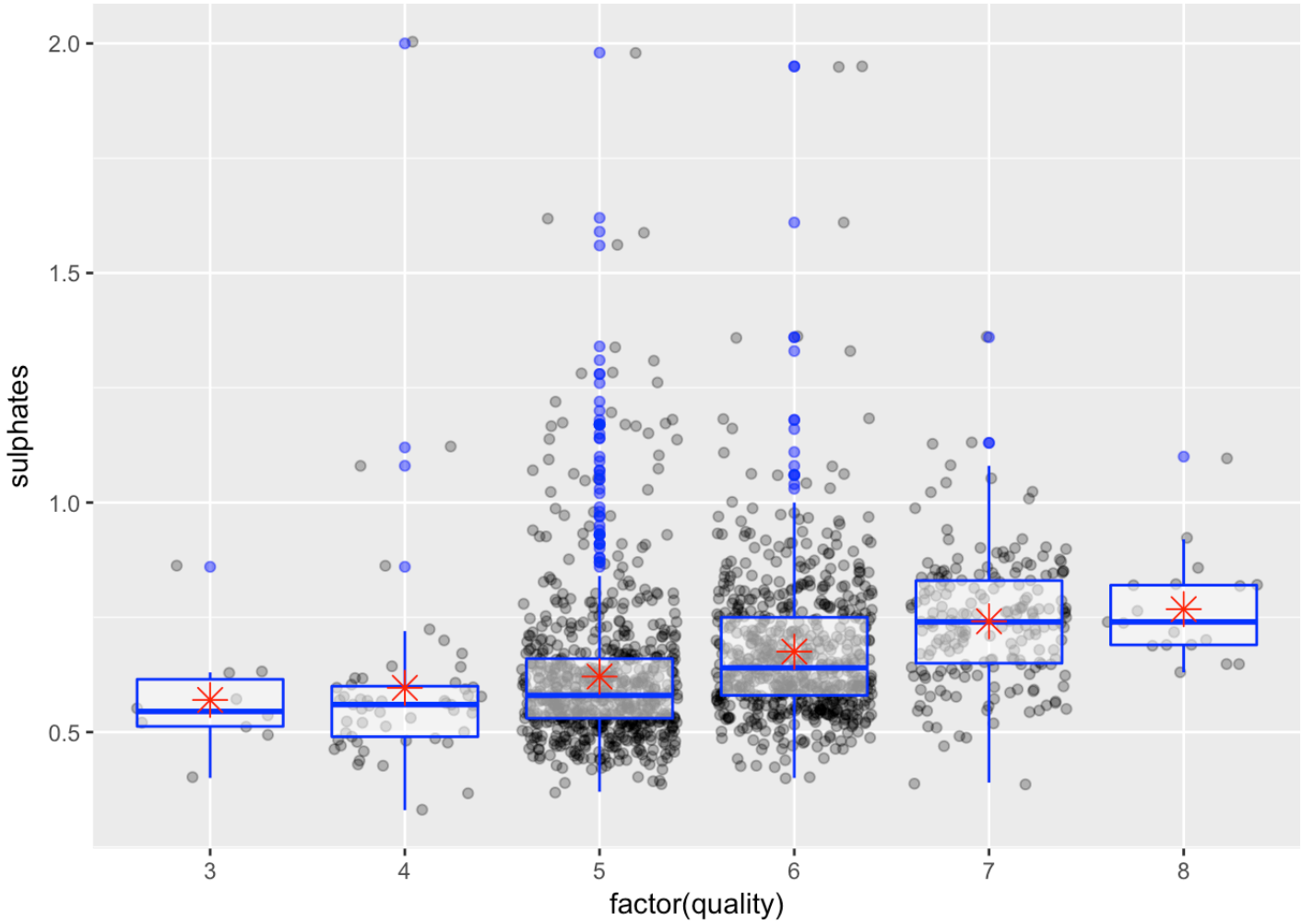
This plot shows that the relationship between quality and volatile.acidity is negative. As the mean of volatile.acidity increases, the red wine quality increases.




```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored
```

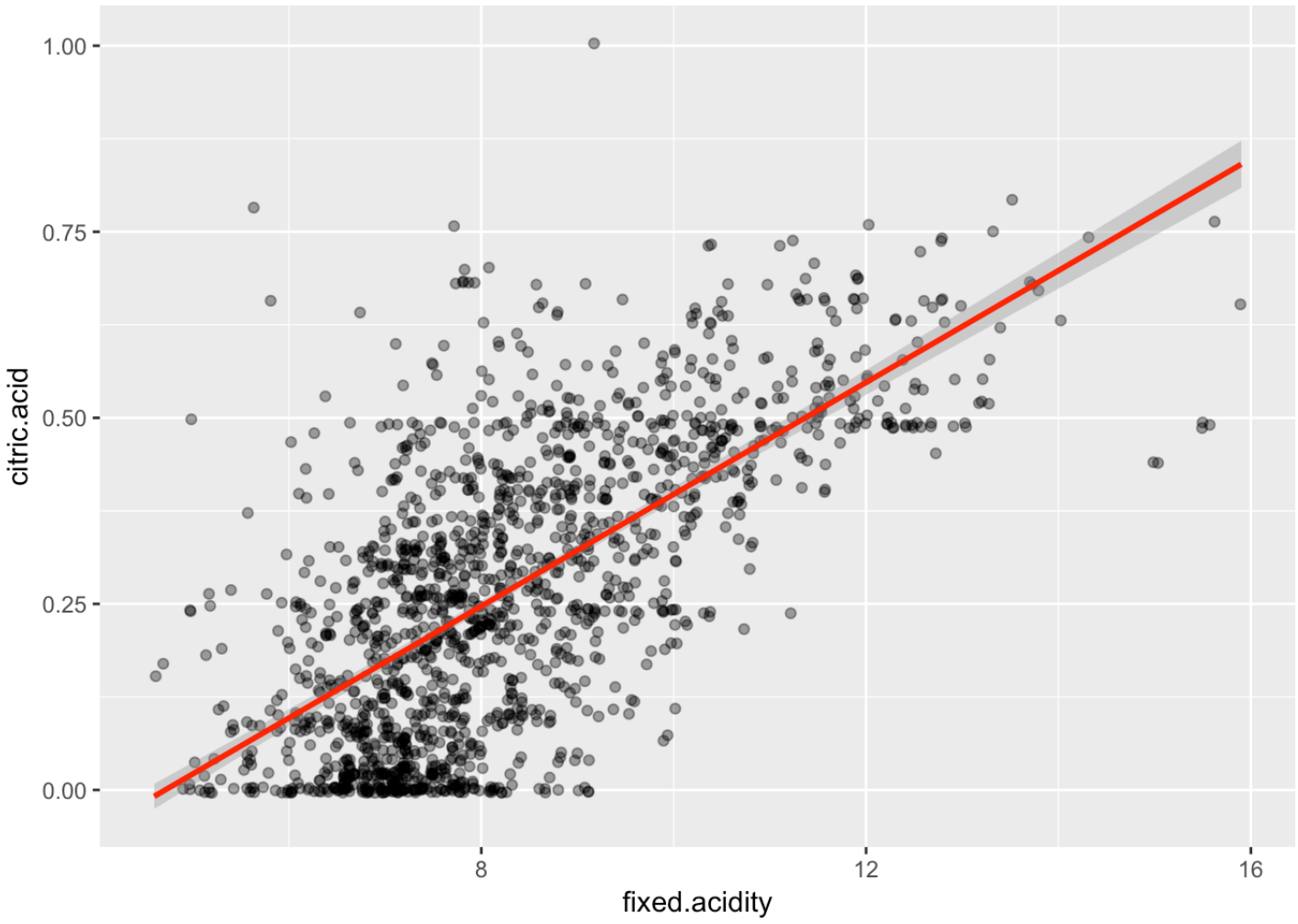
```
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

Majority red wine have sulphates between 0.45 and 0.75. The Adjusted R_squared is 0.06261 which means that the sulphates only explanes 6.3% about the red wine quality. The simple linear regression is not a good algorithm to apply between quality and sulphates.

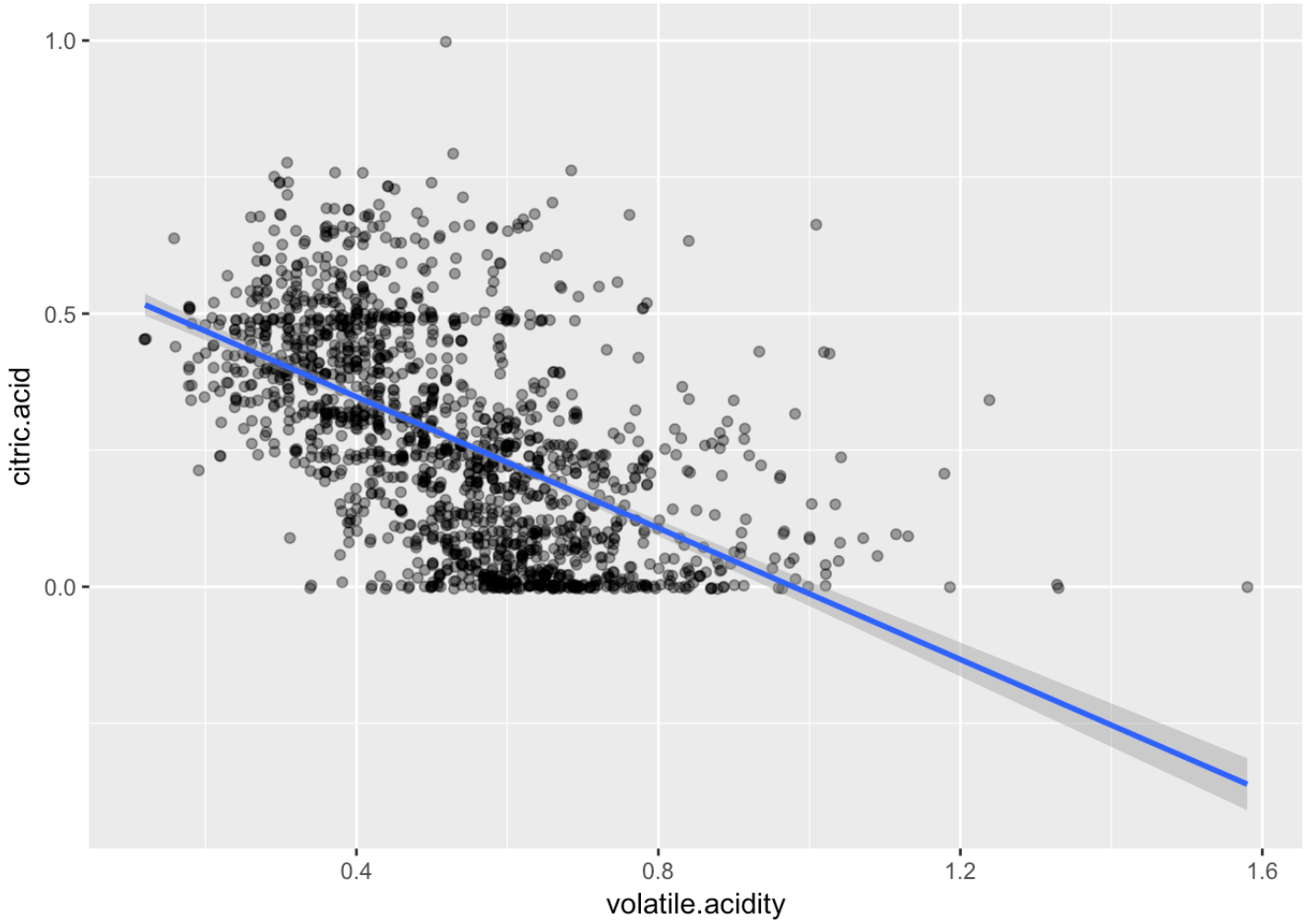


It can be seen that there is slightly positive relationship between quality and sulphates from the above figure.

Next, I'll look at how the relationship between fixed.acidity and citric.acid.



The relationship between fixed.acidity and citric.acid seems like a positive line.



The relationship between volatile.acidity and citric.acid is a slight negative line.

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?

Quality correlates strongly with alcohol and volatile.acidity compring with other variables.

As alcohol increases, red wine seems to have a higher quality tendency, but this tendency looks like special parabola rather than linear since there are some red wines with a low quality for the same alcohol.

Volatile.acidity, another variable seems to have correlation with quality, explains 15.2% variance of quality based on Adjusted R-squared.

The other variable sulphates is incorporated into the model, but it only explains 6.2 percent variance of quality based on Adjusted R-squared. In this way, we could eliminate it from the linear regression model.

In this red wine quality case, the simple linear regression algorithm could not fit the data very well. We might consider other classification algorithm such as Linear Discriminant Analysis.

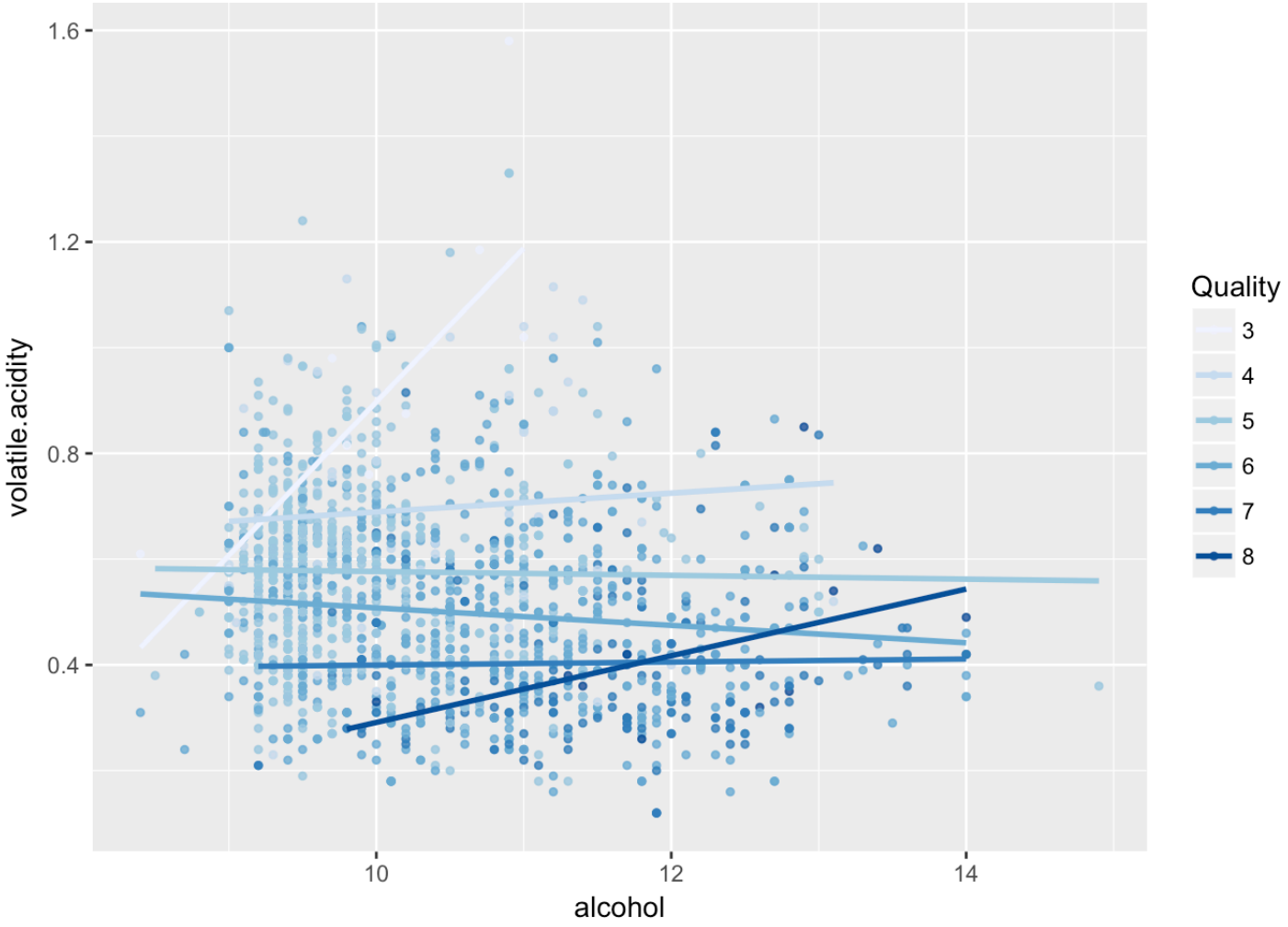
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

From the correlation matrix, there are some variables correlate to each other except quality. The correlation coefficient between fixed.acidity and citric.acid is 0.67. And the correlation coefficient between volatile.acidify and citric.acid is -0.55. We can not include all these correlated variables into the simple linear regression since there are highly correlated.

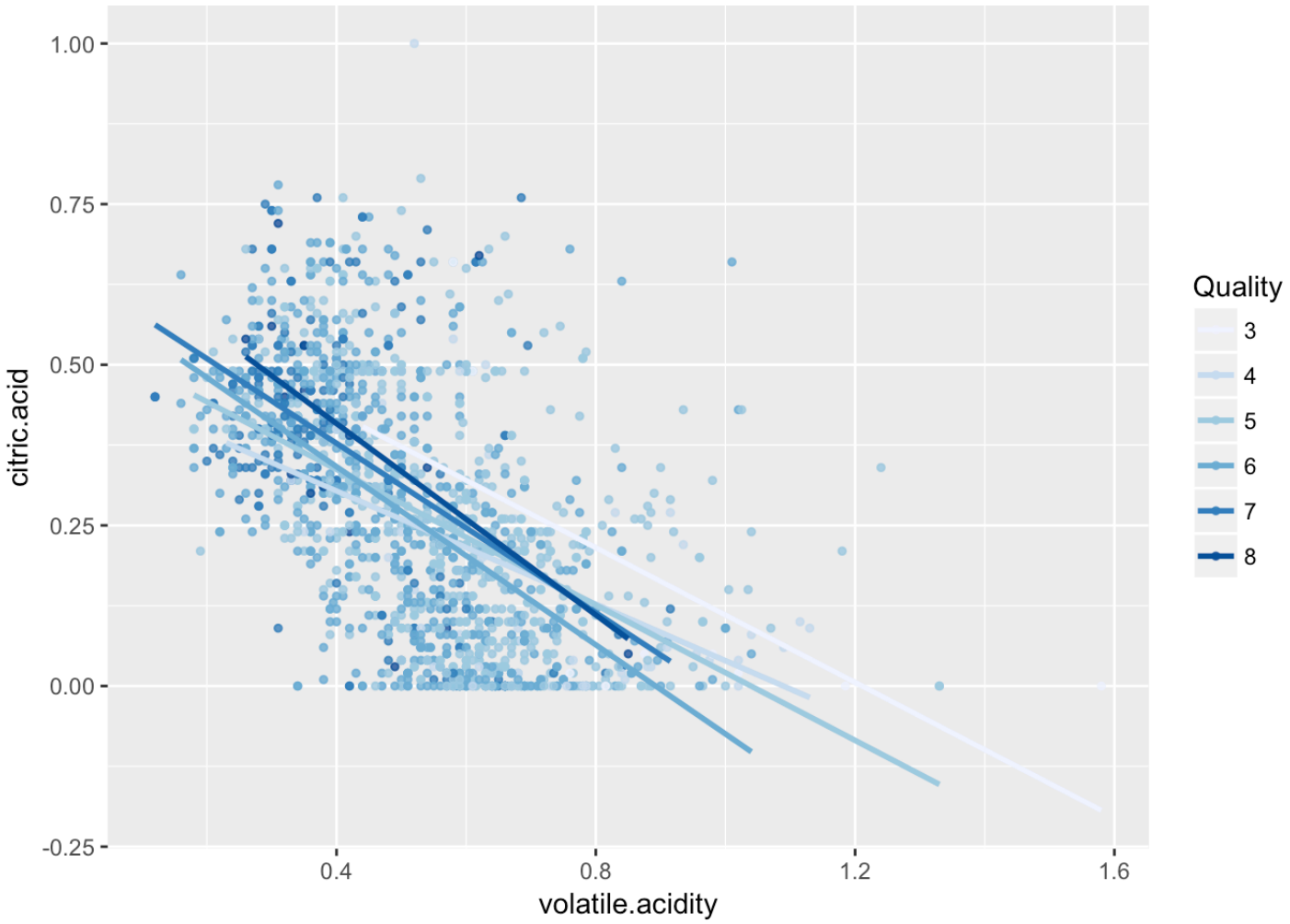
What was the strongest relationship you found?

The red wine quality is positively correlated with alcohol but the relationship seems not to be linear. Also, the read wine quality is negatively correlated with volatile.acidity.

Multivariate Plots Section



This coloful plot shows us that red wine tends to have higher quality as alcohol increases in a special range of volatile.acidity. For example, the red wine quality increases as the alcohol increases from 8 to 14 for volatile.acidity in the range of 0.4 to 0.6. The plot shows that the variations of alcohol and volatile.acidity are large for quality is 8. For quality is 3, the variations of alcohol and volatile.acidity are the greatest than other qualities.



This plot shows us that as volatile.acidity decreases, citric.acid decreases.

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = rw)
## m2: lm(formula = quality ~ alcohol + volatile.acidity, data = rw)
## m3: lm(formula = quality ~ alcohol + volatile.acidity + sulphates,
##     data = rw)
## m4: lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
##     citric.acid, data = rw)
##
## ============================================================
##                         m1         m2         m3         m4
## ------------------------------------------------------------
##   (Intercept)       -0.125      1.095***   0.611**    0.646**
##                     (0.175)    (0.184)    (0.196)    (0.201)
##   alcohol            0.361***   0.314***   0.309***   0.309***
##                     (0.017)    (0.016)    (0.016)    (0.016)
##   volatile.acidity             -1.384***  -1.221***  -1.265***
##                                (0.095)    (0.097)    (0.113)
##   sulphates                                0.679***   0.696***
##                                           (0.101)    (0.103)
##   citric.acid                                        -0.079
##                                                      (0.104)
## ------------------------------------------------------------
##   R-squared             0.2        0.3        0.3        0.3
##   adj. R-squared        0.2        0.3        0.3        0.3
##   sigma                 0.7        0.7        0.7        0.7
##   F                   468.3      370.4      268.9      201.8
##   p                     0.0        0.0        0.0        0.0
##   Log-likelihood    -1721.1    -1621.8    -1599.4    -1599.1
##   Deviance            805.9      711.8      692.1      691.9
##   AIC                3448.1     3251.6     3208.8     3210.2
##   BIC                3464.2     3273.1     3235.7     3242.4
##   N                  1599       1599       1599       1599
## ============================================================
```

Applied linear regression on this dataset. The Adjusted R-squared is not higher than 0.3, so the linear regression would not be a good model. Next, I am going to explore this dataset by using classification algorithms such as KNN.

```
##   1   2   3   4   5   6
##   0   0  65 316   8   0
```

```
##
```

```
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  389
##
##
##                | rw_test_pred
## rw_test_labels |         3 |         4 |         5 | Row Total |
## ---------------|-----------|-----------|-----------|-----------|
##              1 |         3 |         2 |         0 |         5 |
##                |     0.600 |     0.400 |     0.000 |     0.013 |
##                |     0.046 |     0.006 |     0.000 |           |
##                |     0.008 |     0.005 |     0.000 |           |
## ---------------|-----------|-----------|-----------|-----------|
##              2 |         8 |         9 |         1 |        18 |
##                |     0.444 |     0.500 |     0.056 |     0.046 |
##                |     0.123 |     0.028 |     0.125 |           |
##                |     0.021 |     0.023 |     0.003 |           |
## ---------------|-----------|-----------|-----------|-----------|
##              3 |        40 |       126 |         0 |       166 |
##                |     0.241 |     0.759 |     0.000 |     0.427 |
##                |     0.615 |     0.399 |     0.000 |           |
##                |     0.103 |     0.324 |     0.000 |           |
## ---------------|-----------|-----------|-----------|-----------|
##              4 |        14 |       152 |         7 |       173 |
##                |     0.081 |     0.879 |     0.040 |     0.445 |
##                |     0.215 |     0.481 |     0.875 |           |
##                |     0.036 |     0.391 |     0.018 |           |
## ---------------|-----------|-----------|-----------|-----------|
##              5 |         0 |        23 |         0 |        23 |
##                |     0.000 |     1.000 |     0.000 |     0.059 |
##                |     0.000 |     0.073 |     0.000 |           |
##                |     0.000 |     0.059 |     0.000 |           |
## ---------------|-----------|-----------|-----------|-----------|
##              6 |         0 |         4 |         0 |         4 |
##                |     0.000 |     1.000 |     0.000 |     0.010 |
##                |     0.000 |     0.013 |     0.000 |           |
##                |     0.000 |     0.010 |     0.000 |           |
## ---------------|-----------|-----------|-----------|-----------|
##   Column Total |        65 |       316 |         8 |       389 |
```

```
##                |     0.167 |     0.812 |     0.021 |           |
## ---------------|-----------|-----------|-----------|-----------|
##
##
```

After I had applied KNN to classify the red wine quality based on training dataset, I got a crosstable of prediction and test dataset. KNN had a good probability of right predictions at quality equals to 6. The performances of KNN were terrible on all other quality levels.

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?

In a specifical range of volatile.acidity, the red wine quality will increase as the alcohol increasing.

Were there any interesting or surprising interactions between features?

Citric.acid are monotonically decreasing as volatile.acidity decreasing.

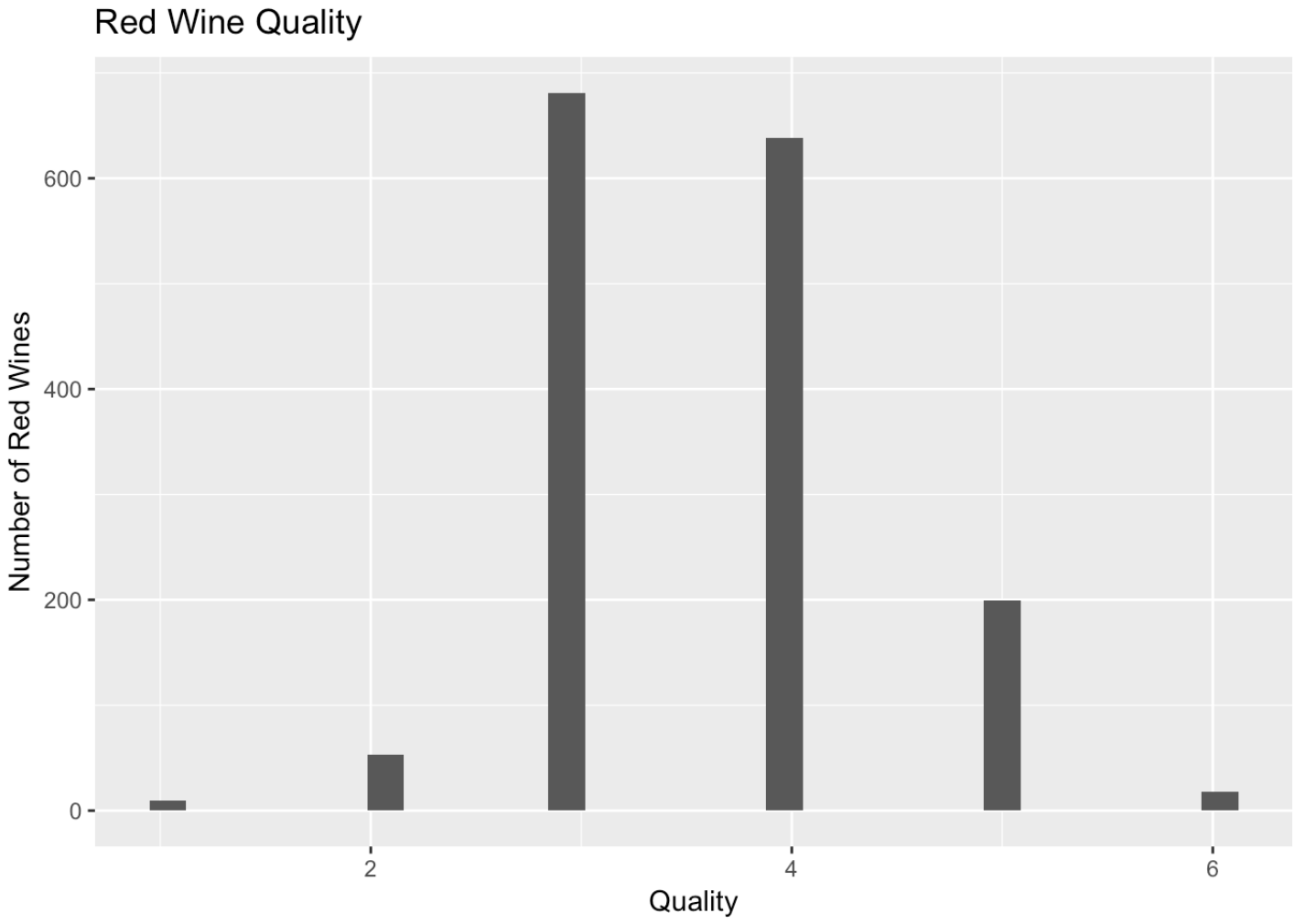
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Yes, I created two models with red wine dataset. One of them is linear regression model with one variable, two variables, three variables and four variables. The other model is K-nearest algorithm for classification with k = 24. Based on Adjusted R-squared, the multiple linear regression isn't a good model for red wine dataset. For the KNN model, the predicion on the test dataset is also not a good model. But I think it is better than multiple regression.

Final Plots and Summary

Plot One

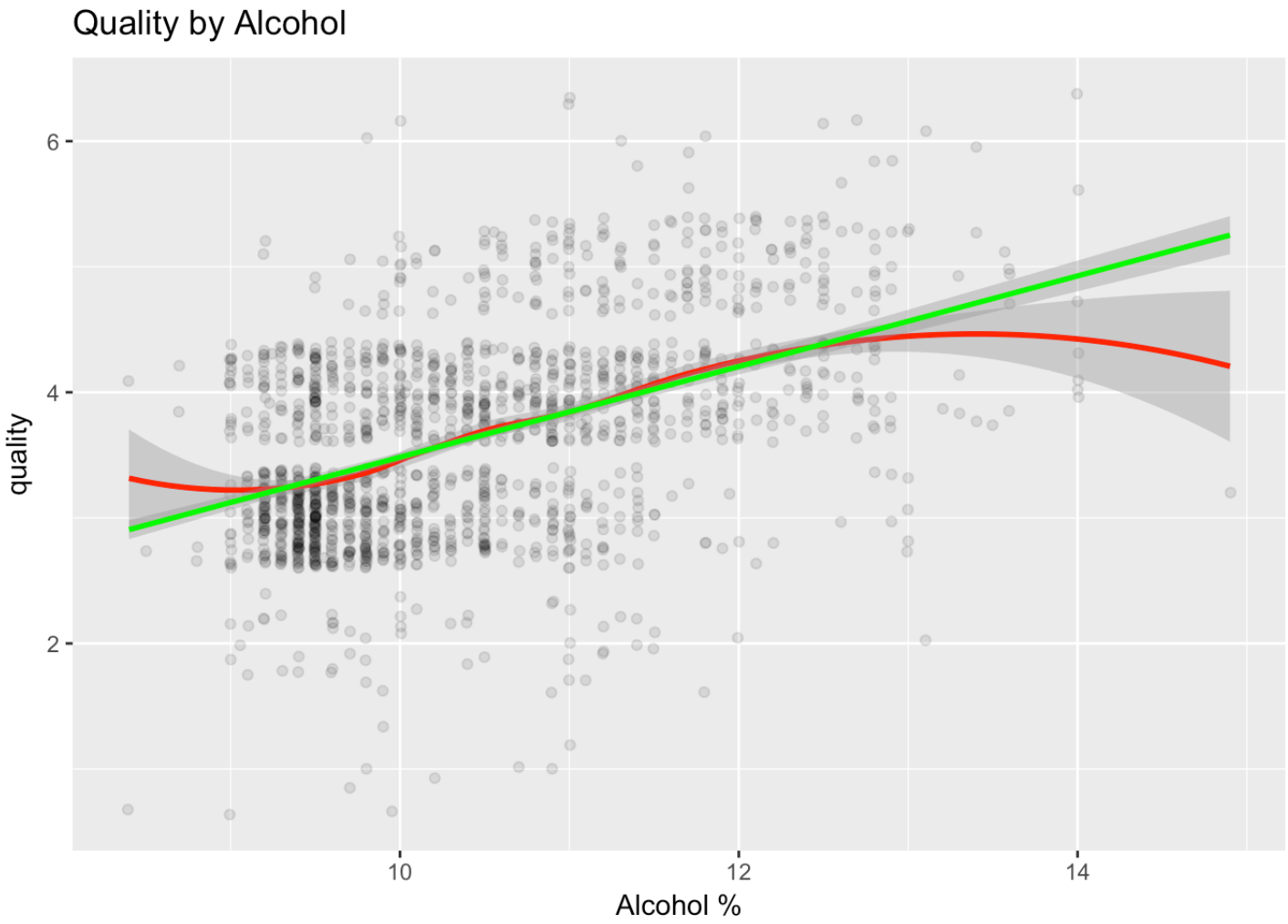
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Description One

There are almost 690 red wines have quality 5 and more than 600 red wines have quality 6. 200 red wines have quality 7. Majority red wine qualities are 5 and 6.

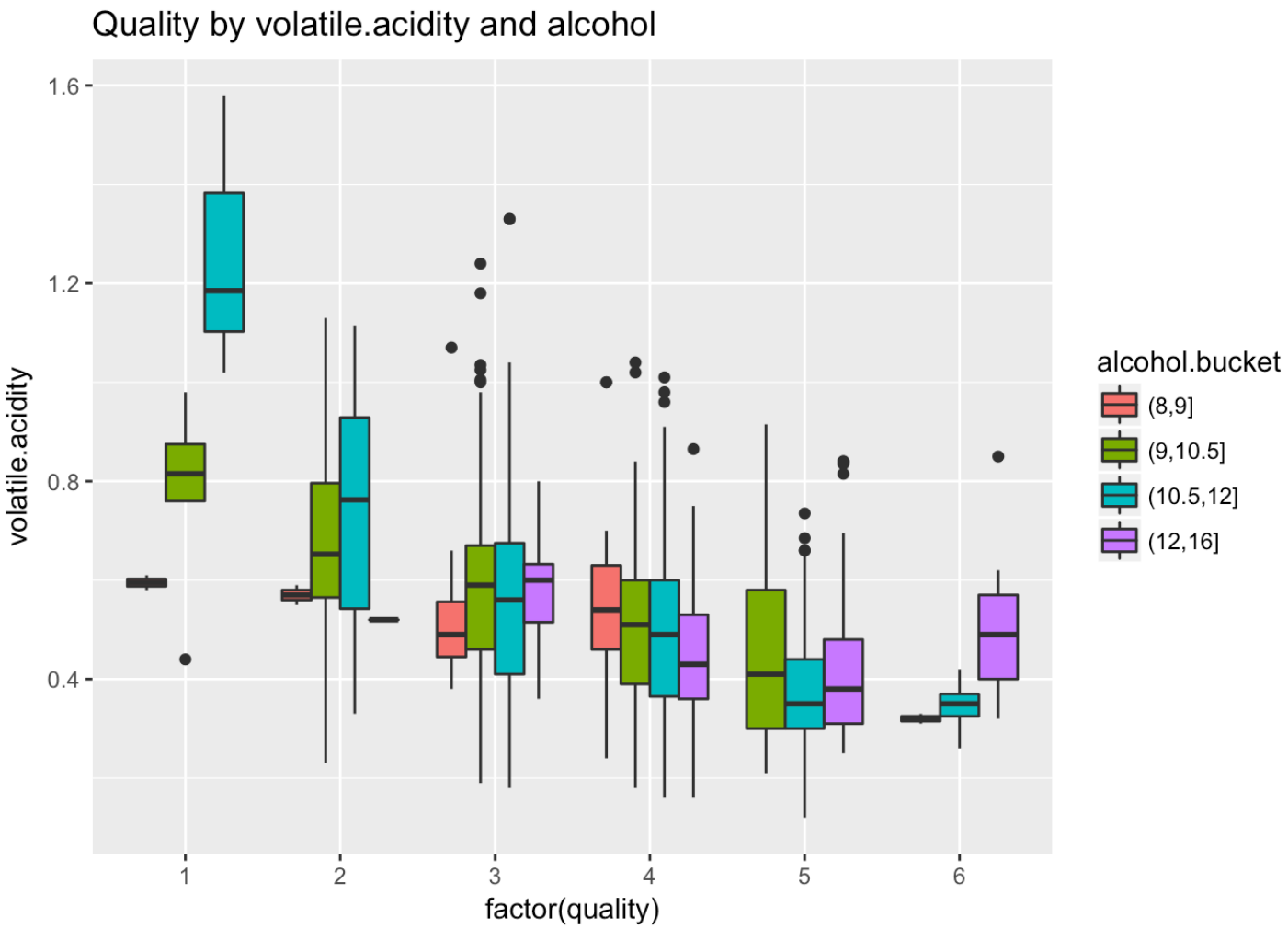
Plot Two



Description Two

The plot shows that the relationship between quality and alcohol is positive. As alcohol of red wine increases, the quality also increases. There are two methods to smooth this dataset. One is using linear regression, another is using loess. But there are a lot of overplotting, linear regression isn't a good model for red wine dataset.

Plot Three



Description Three

The boxplots clearly show that quality tends to be higher as volatile.acidity decreases and the alcohol increases. It indicates two opposite directions relationship between quality and volatile.acidity and between quality and alcohol. The correlation coefficient between quality and volatile.acidity is negative. The correlation coefficient between quality and alcohol is positive. Multiple regression can be used. We can interpret the relationship between quality with either volatile.acidity or alcohol by fixing another one.

Reflection

The red wine dataset contains 1599 observations across 12 variables. I started by exploring individual variables in the dataset. I plotted histograms of all variables, some variables have approximately normal distribuion such as density and pH. The distribuions of free.sulfur.dioxide, total.sulfur.dioxide and alcohol are highly right skewed which have long right tail. Then I explored relationship between quality with all other variables and tried to find an appropriate model.

Though there was a postively trend between quality and alcohol, the dataset suffers from overplotting. In this case, classification algorithm might be more appropriate to classify the red wine quailty. But I sitll applied linear regression for qulity, as we known, the modeling results are not very well based on the Adjusted R-squared which are not higher than 30%.

After I had applied the K-Nearest Neighbors, I got the crosstable on the test dataset. For the quality equals to 6, this classification algorithm can get 88% right predictions of total number of quality is 6. But it doesn’t get a reasonable probability of right predictions among all other quality levels.

After I had applied linear regression and KNN, I known that those two method are not appropriate for the red wine dataset. A more advanced classification algorithm should be used such as Random Forest.