#Red Wine Exploratory Data Analysis

By Siye Yu

Today, I will explore the redwine dataset to find out factors which make great quality wine. Everyone may has his or her unique taste, whether its sweetness or bitterness. Let's first take a look at the dataset data summary to get an better idea before we start.

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

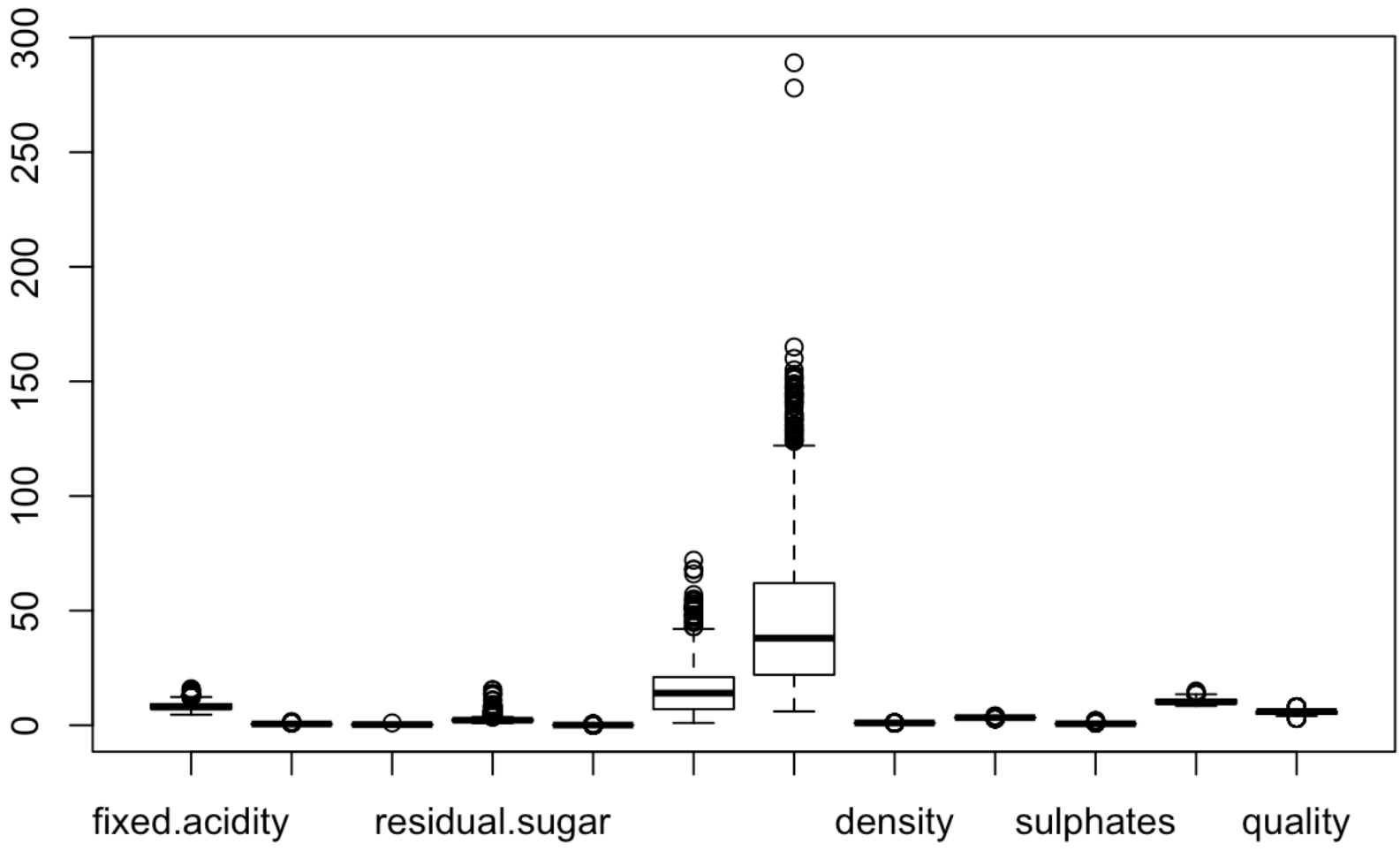
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 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
```

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```
##  [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"
```

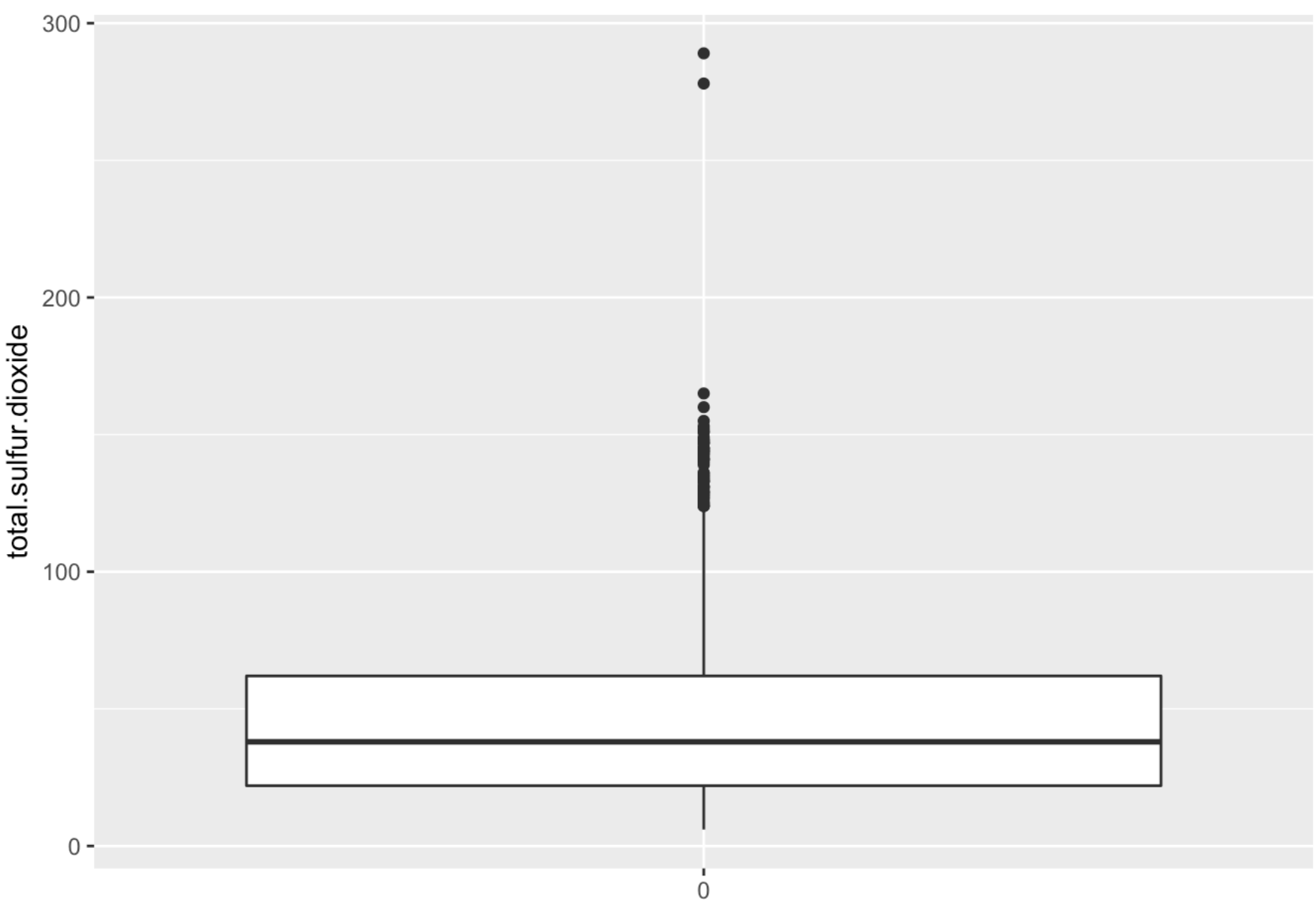
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

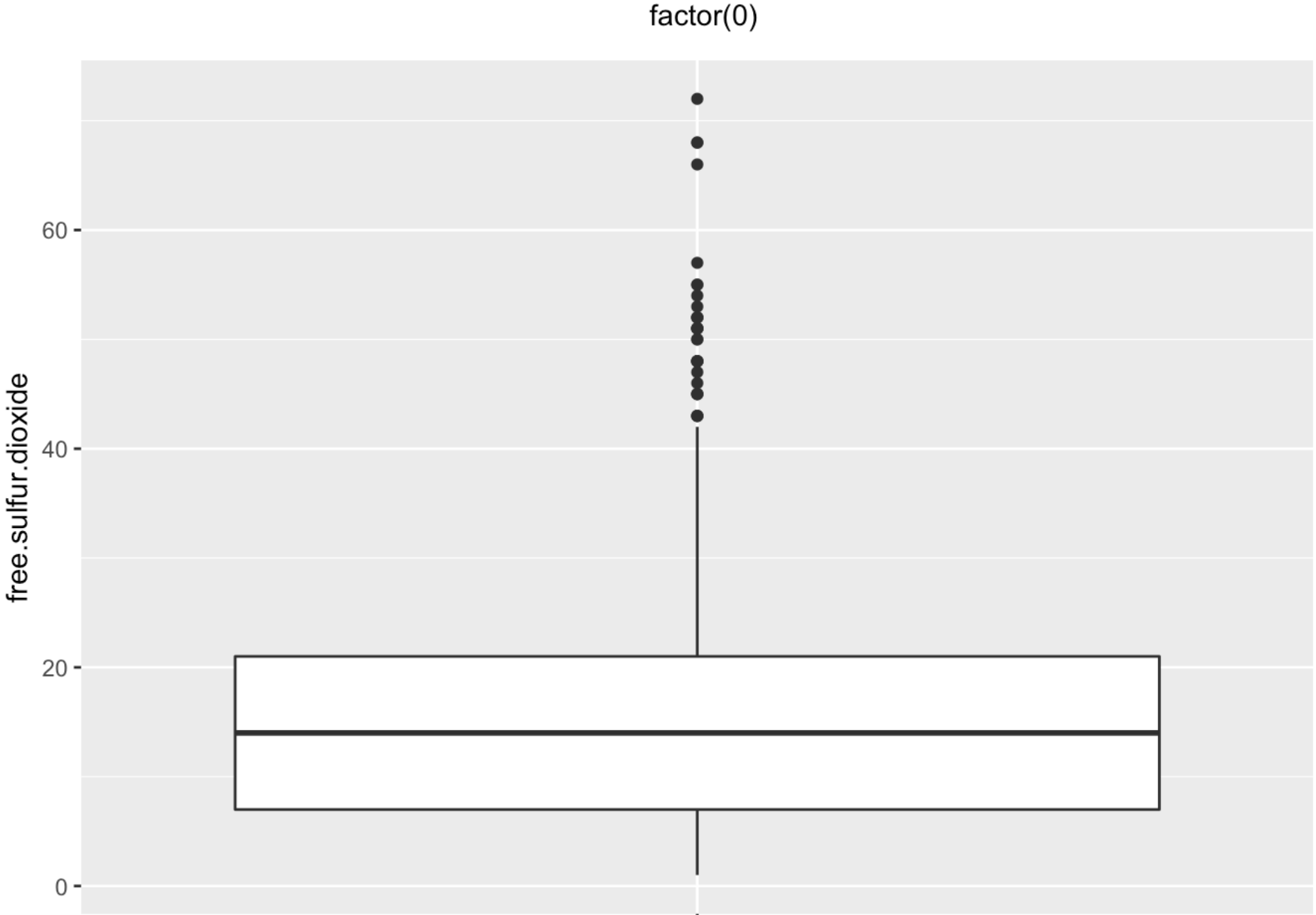
Univariate Plots Section

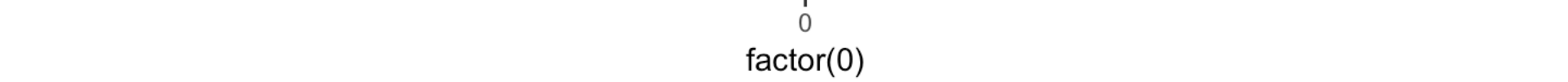


Plot shows outlier for two columns, free.sulfur.dioxide and total.sulfur.dioxide

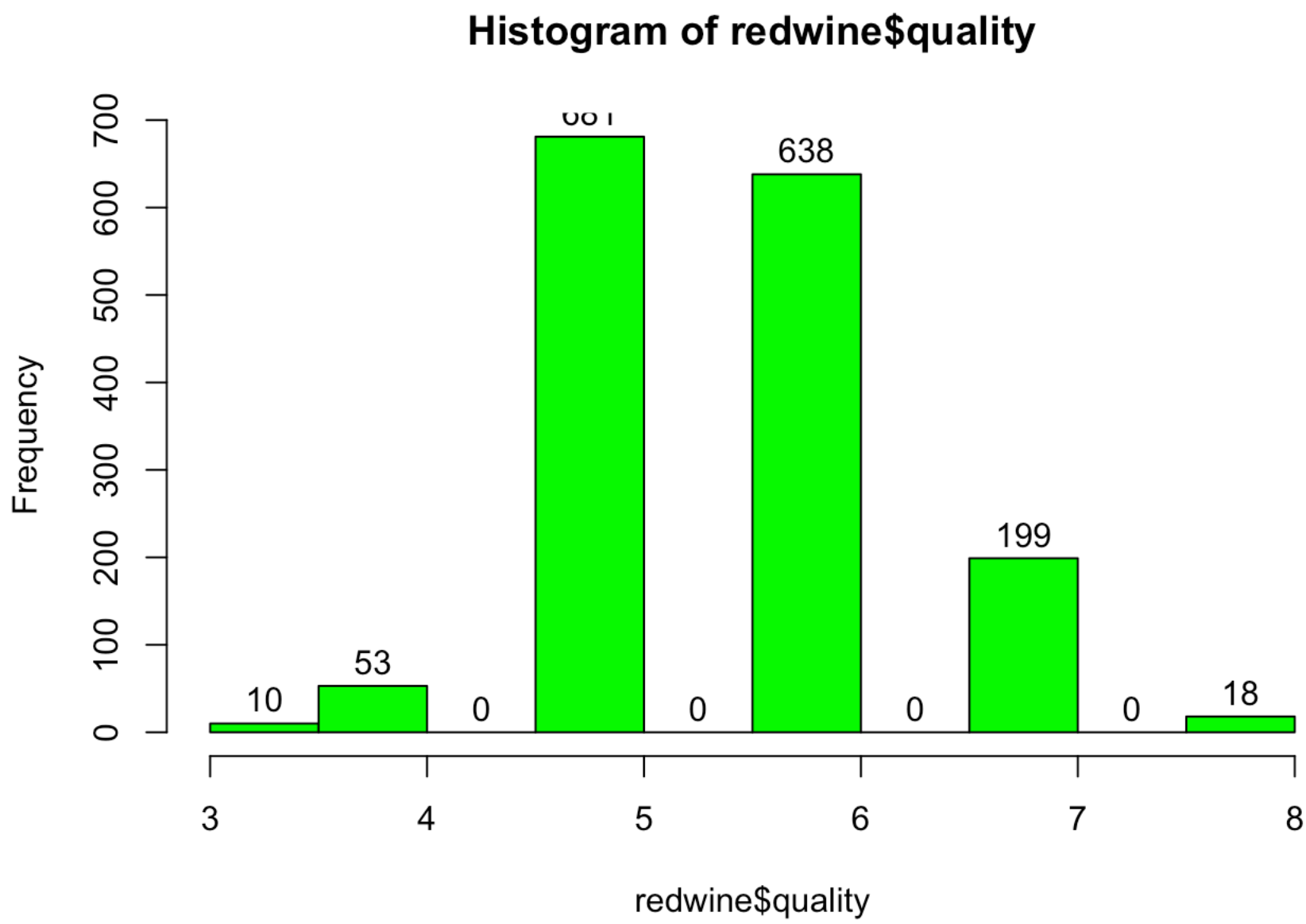
To get a better picture, I'll create two boxplots for the columns.


total.sulfur.dioxide
300
200
100
0
0

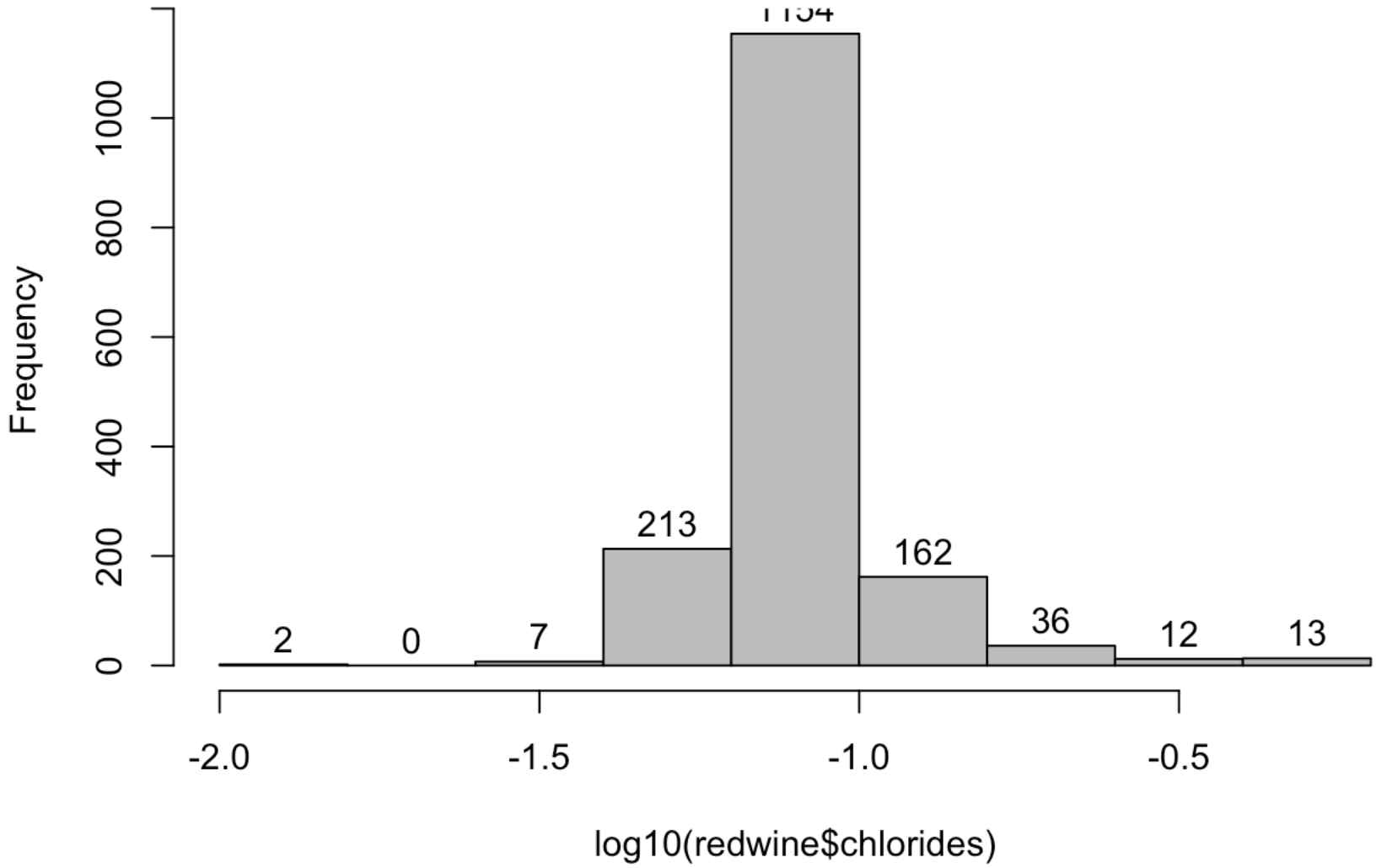

factor(0)
free.sulfur.dioxide
60
40
20
0

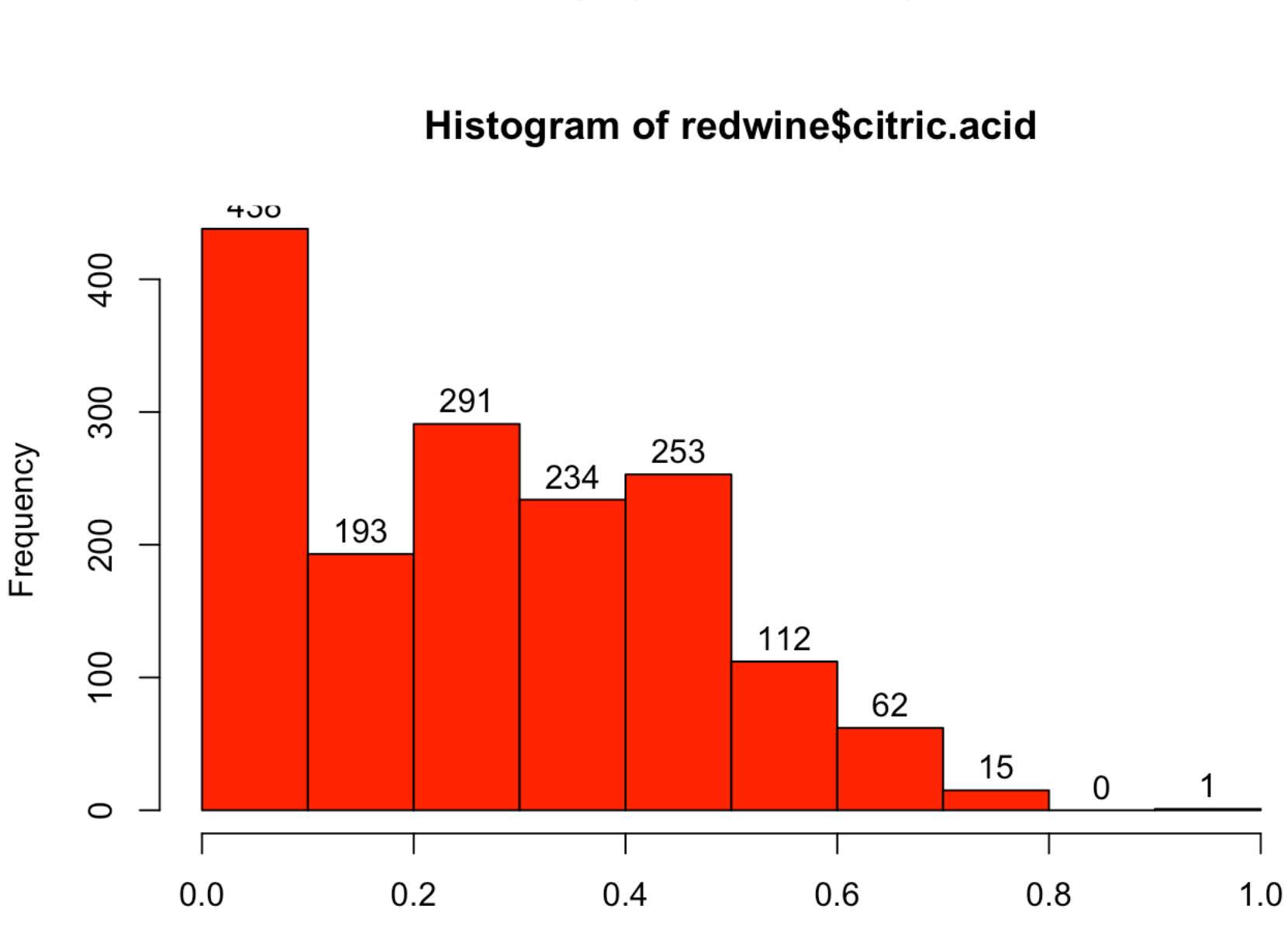


Distribution of quality histogram, which shows a normal distrubtin.

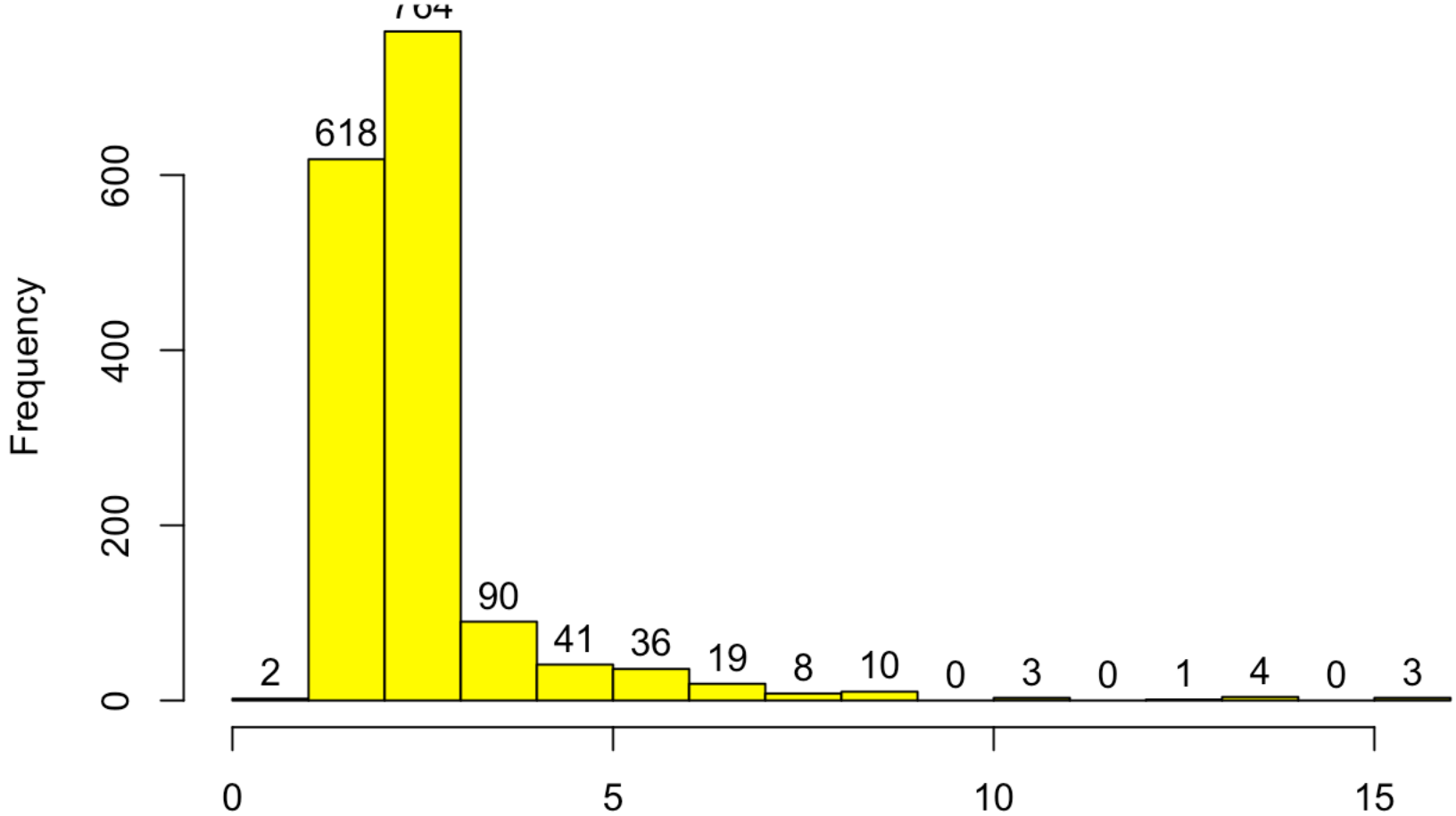


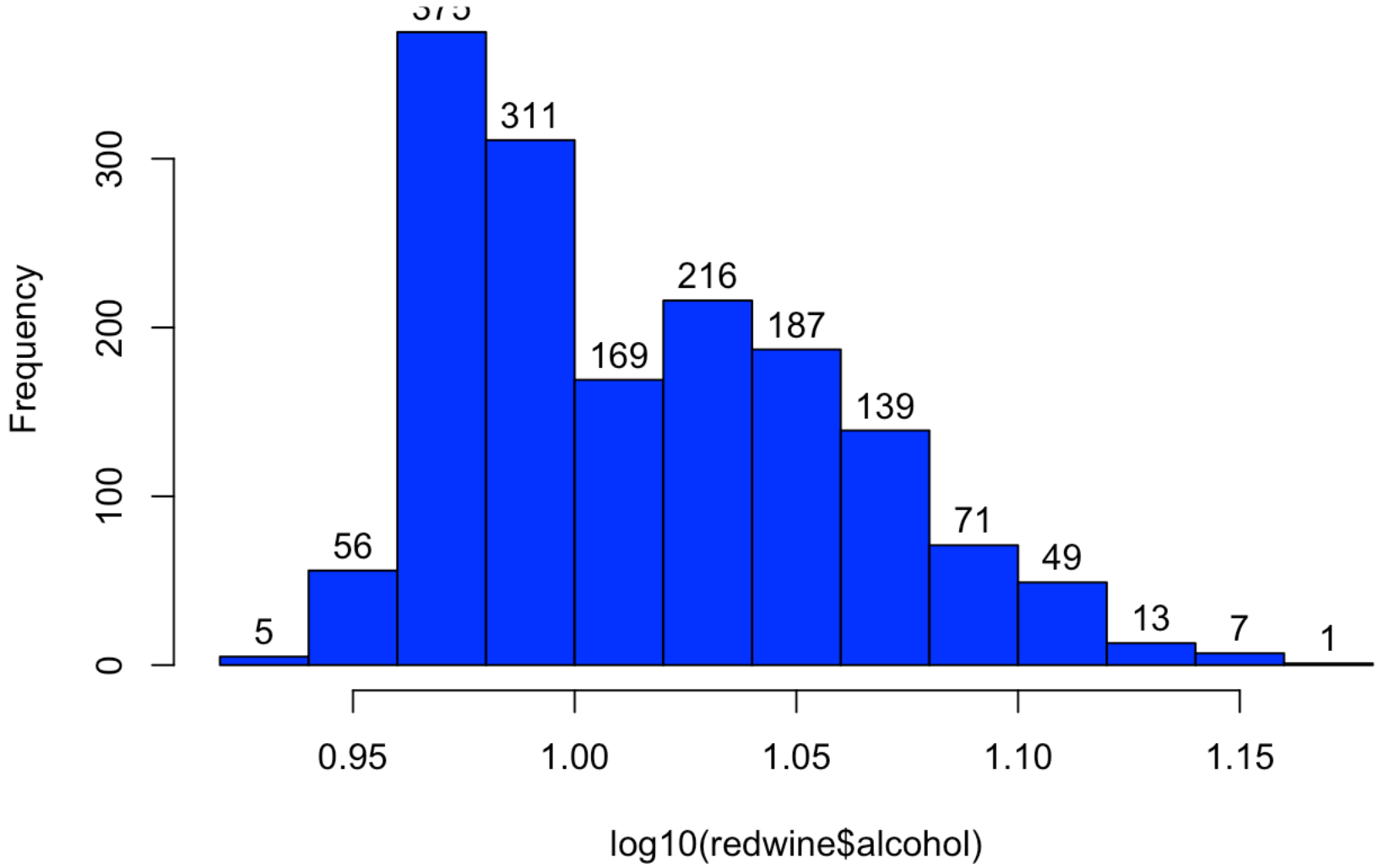
Histogram of count for chlorides, citric acid, residual sugar and alcohol variable to show distrubution


Histogram of log10(redwine$chlorides)
Frequency
0
200
400
600
800
1000
2
0
7
213
1154
162
36
12
13
-2.0
-1.5
-1.0
-0.5
log10(redwine$chlorides)


Histogram of redwine$citric.acid
Frequency
0
100
200
300
400
438
193
291
234
253
112
62
15
0
1
0.0
0.2
0.4
0.6
0.8
1.0

redwine$citric.acid





Univariate Analysis

What is the structure of your dataset?

it has 1599 obs. of 13 varaiables

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

whether critic.acid(freshness), sugar(sweatness), chlorides(saltyness) has any impact on the quality of wine.

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

the alochol percentage may help me to answer few question since it has the medium strength of correlation on both density and quality.

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

No new variable was created

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?

No, the data is already clean and easy to plot. It doesn't appear to be useful when analysing the data.

Bivariate Plots Section

```
## redwine$quality: 3
## [1] 9.955
## ------------------------------------------------------------
## redwine$quality: 4
## [1] 10.26509
## ------------------------------------------------------------
## redwine$quality: 5
## [1] 9.899706
## ------------------------------------------------------------
## redwine$quality: 6
## [1] 10.62952
## ------------------------------------------------------------
## redwine$quality: 7
## [1] 11.46591
## ------------------------------------------------------------
## redwine$quality: 8
## [1] 12.09444
```

```
## redwine$quality: 3
## [1] 0.171
## ------------------------------------------------------------
## redwine$quality: 4
## [1] 0.1741509
## ------------------------------------------------------------
## redwine$quality: 5
## [1] 0.2436858
## ------------------------------------------------------------
## redwine$quality: 6
## [1] 0.2738245
## ------------------------------------------------------------
## redwine$quality: 7
## [1] 0.3751759
## ------------------------------------------------------------
## redwine$quality: 8
## [1] 0.3911111
```

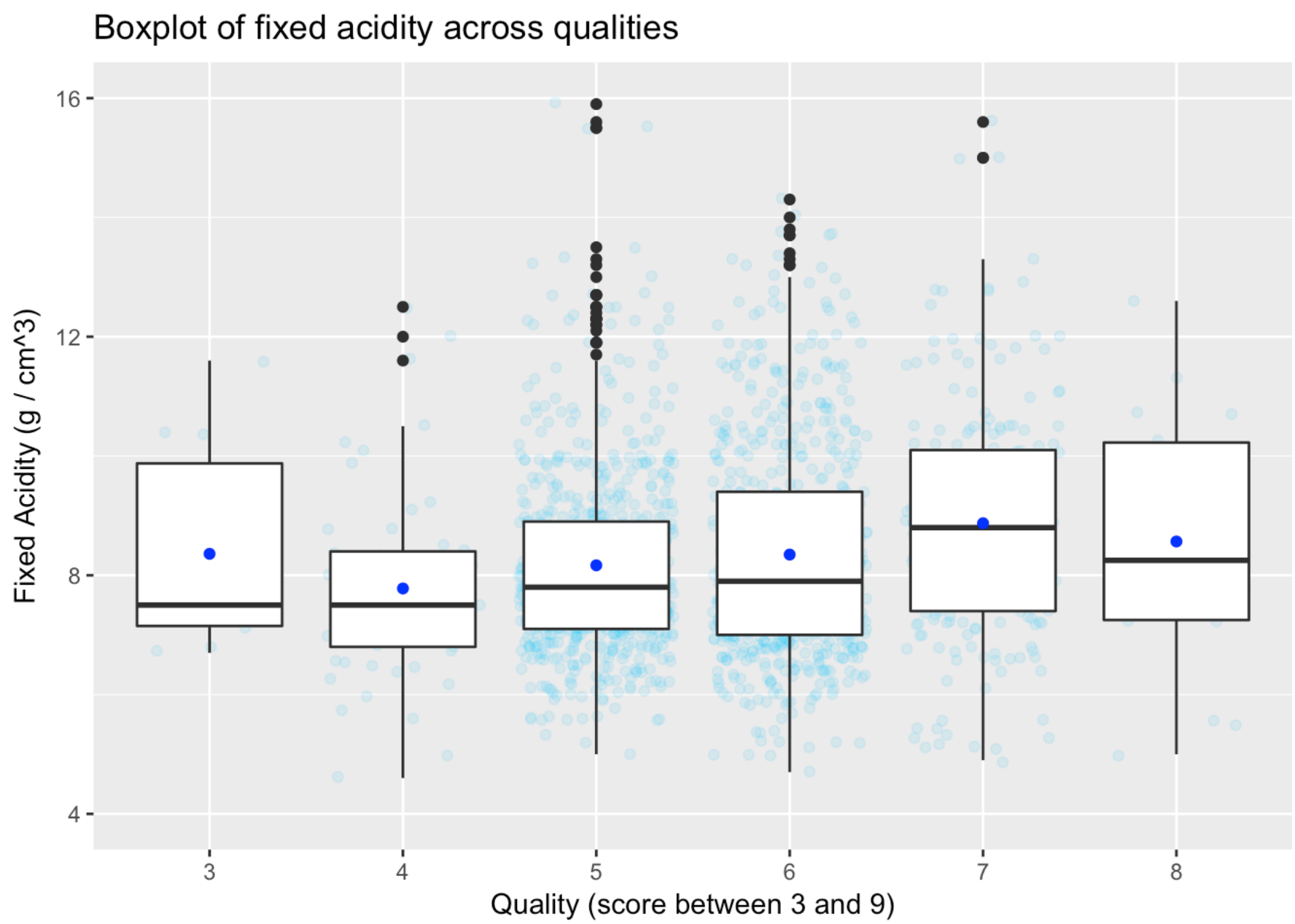
```
## redwine$quality: 3
## [1] 2.635
## ------------------------------------------------------------
## redwine$quality: 4
## [1] 2.69434
## ------------------------------------------------------------
## redwine$quality: 5
## [1] 2.528855
## ------------------------------------------------------------
## redwine$quality: 6
## [1] 2.477194
## ------------------------------------------------------------
## redwine$quality: 7
## [1] 2.720603
## ------------------------------------------------------------
## redwine$quality: 8
## [1] 2.577778
```

```
## redwine$quality: 3
## [1] 0.1225
## ------------------------------------------------------------
## redwine$quality: 4
## [1] 0.09067925
## ------------------------------------------------------------
## redwine$quality: 5
## [1] 0.09273568
## ------------------------------------------------------------
## redwine$quality: 6
## [1] 0.08495611
## ------------------------------------------------------------
## redwine$quality: 7
## [1] 0.07658794
## ------------------------------------------------------------
## redwine$quality: 8
## [1] 0.06844444
```

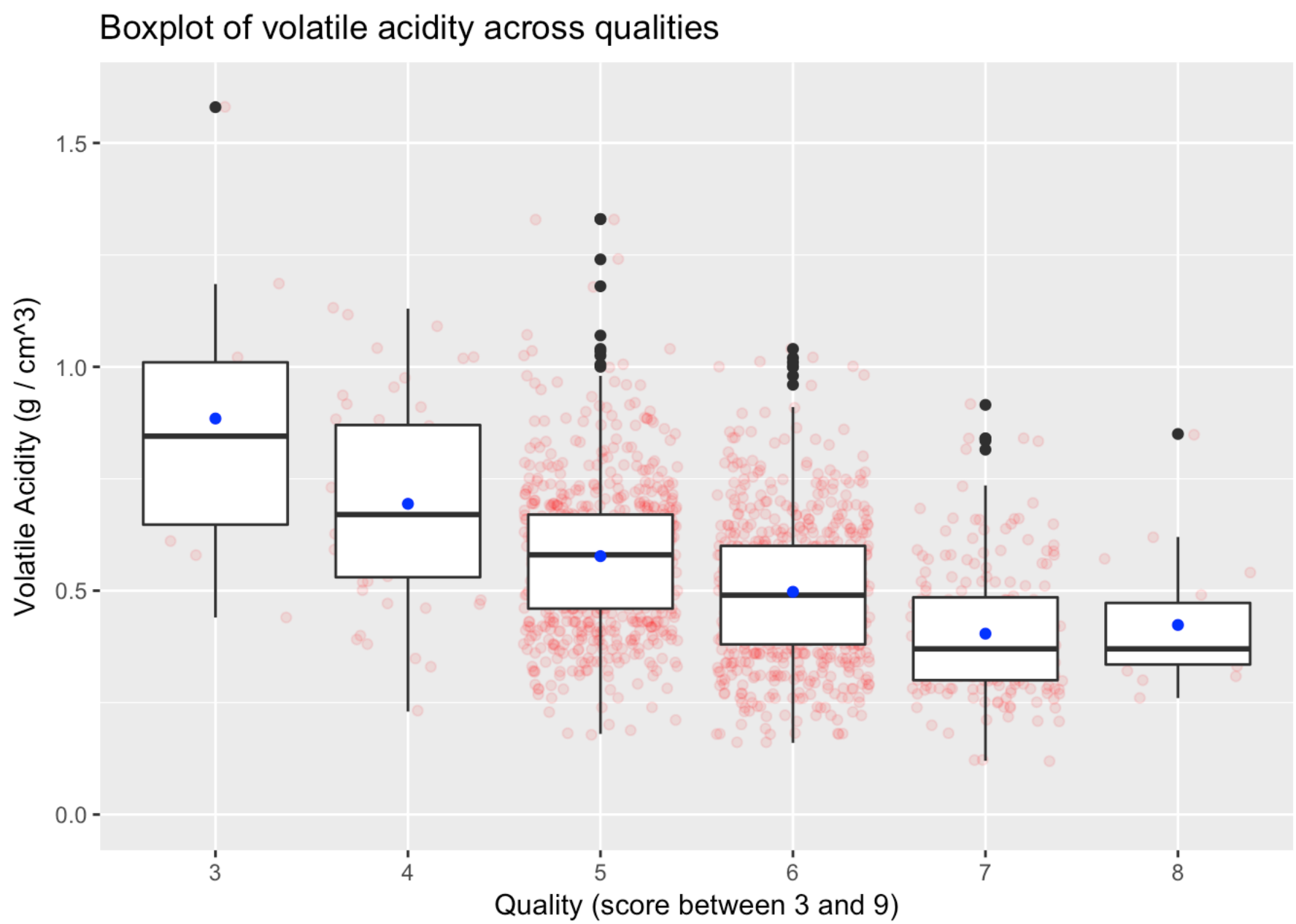
Looks like the lower chlorides, the better quality, we will investigate further to make sure the finding

Next let’s look at some of the supporting variables

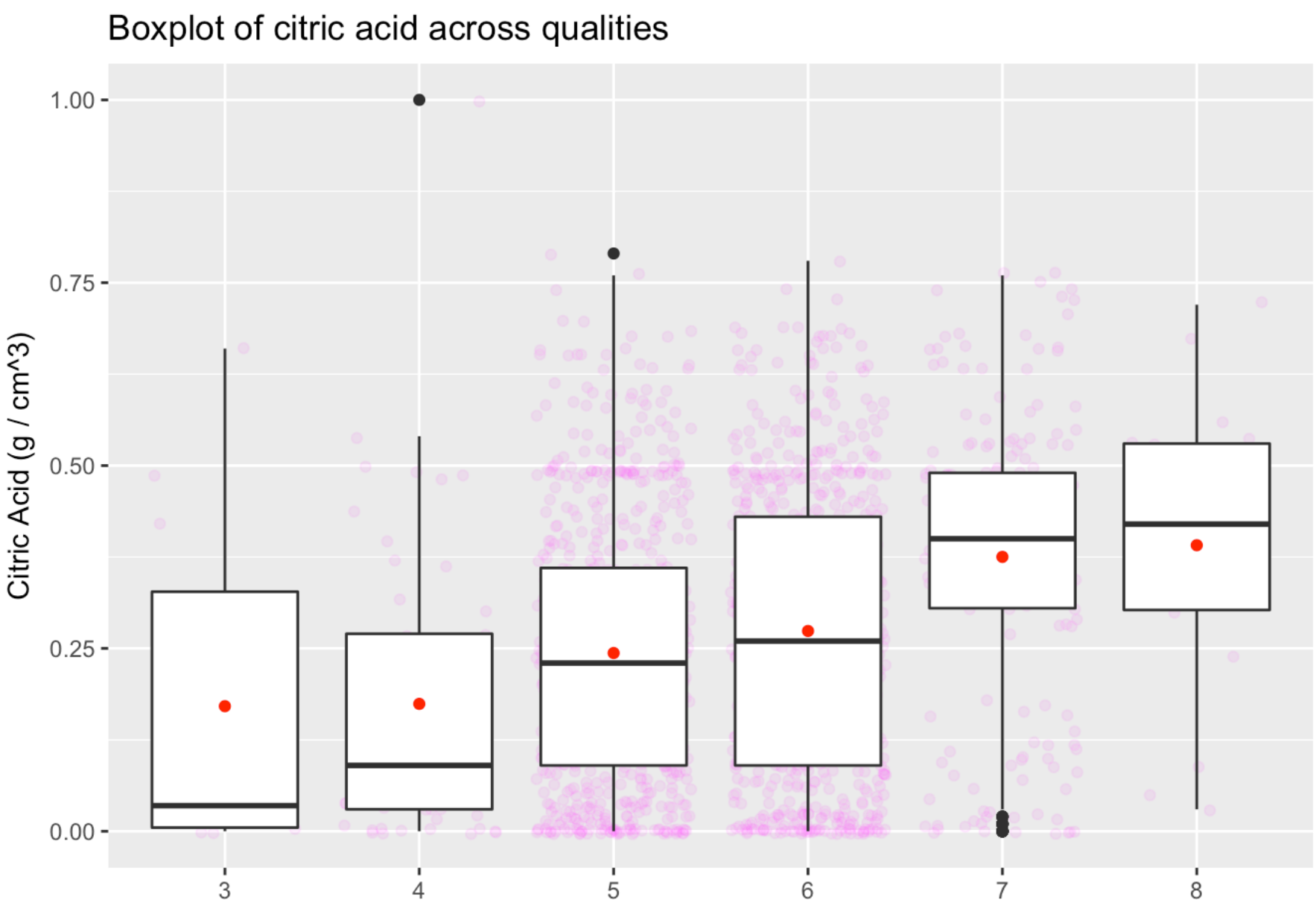
First take a look at the fixed acidity, as the graph shows, there is almost no correlation between quality and fixed acidity



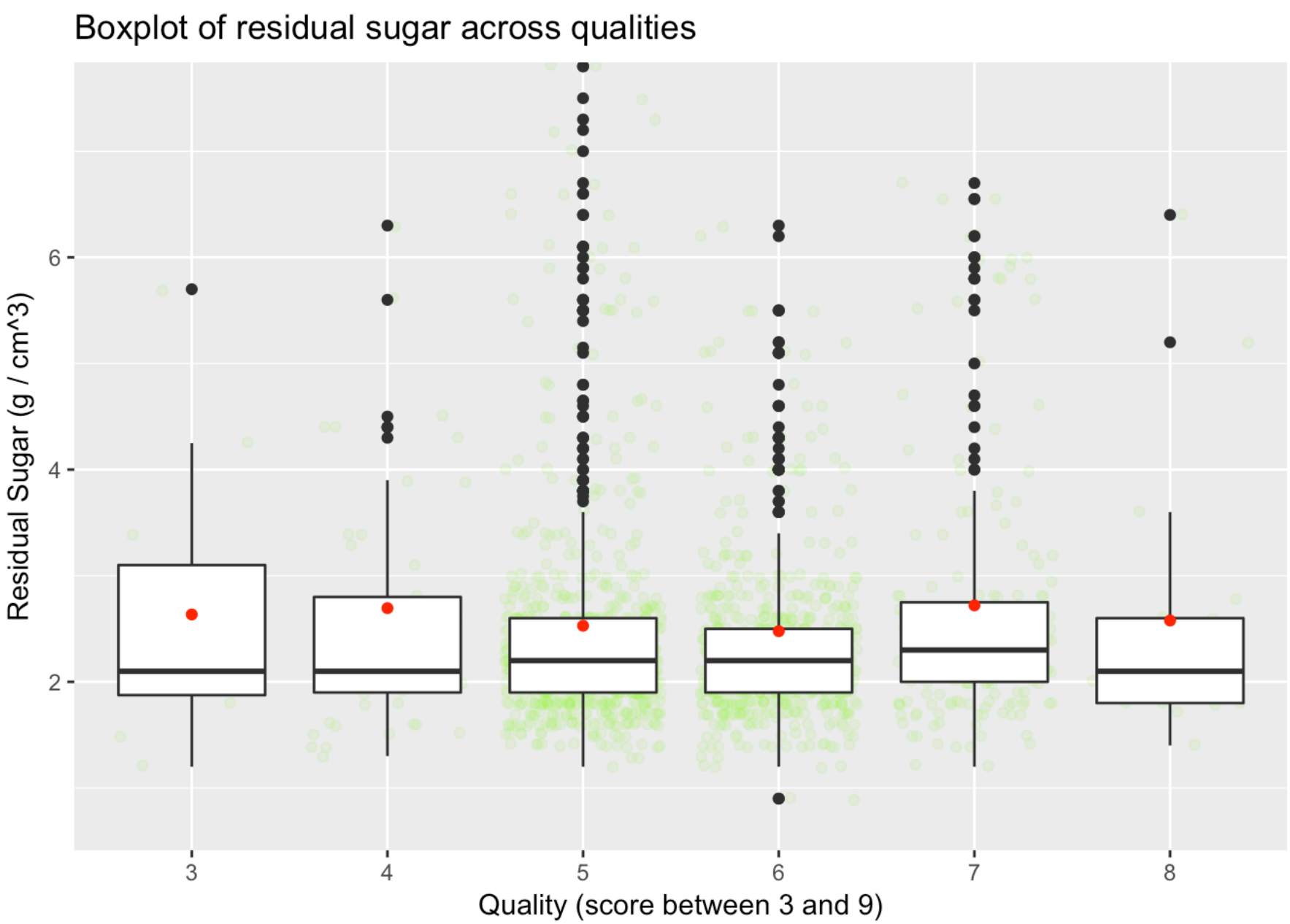
Graph shows volatile acidity has a negative correlation with quality.



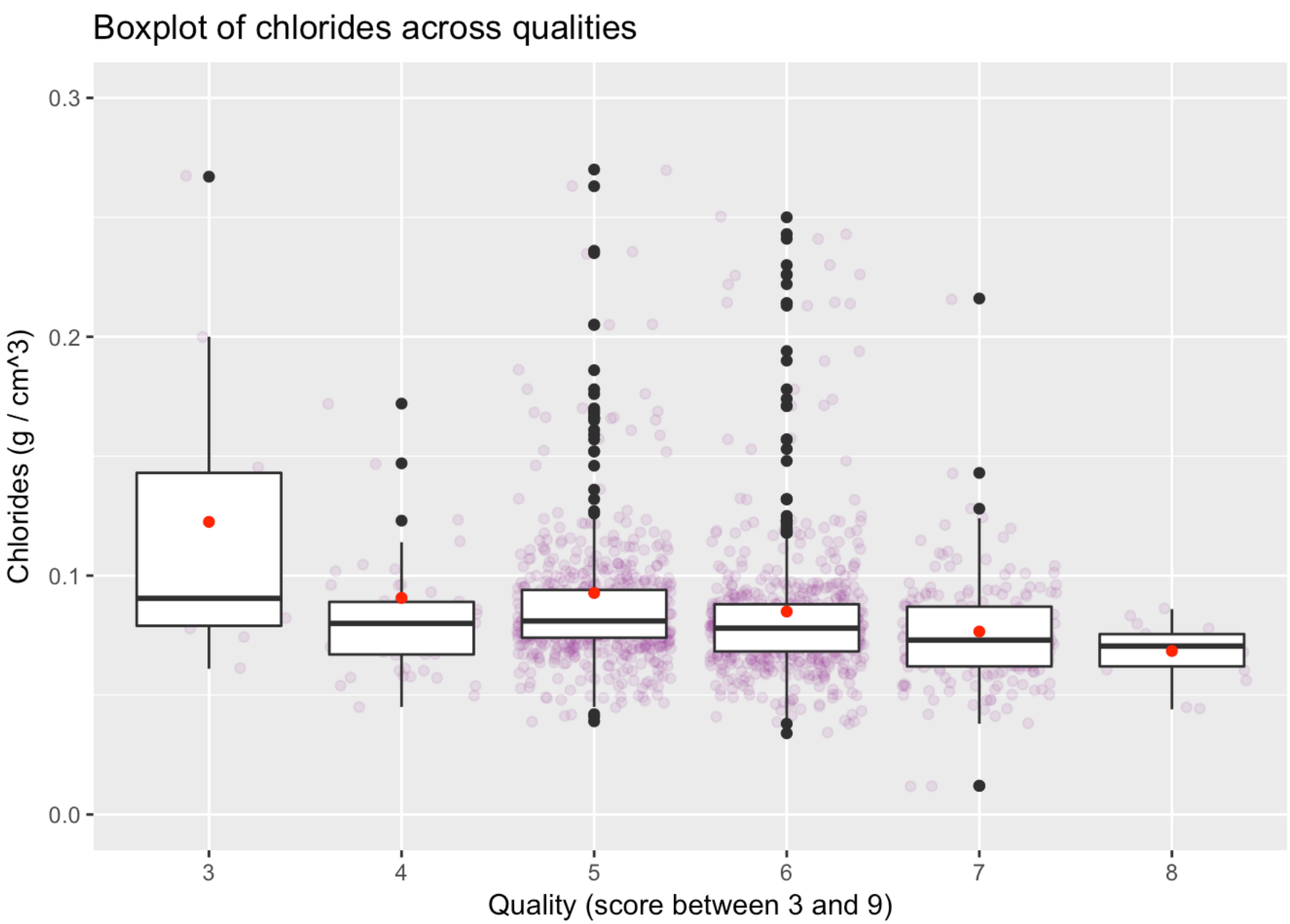
Citric acid has a somewhat postitive relationship with quality, as citric acid increases, the quality also increase



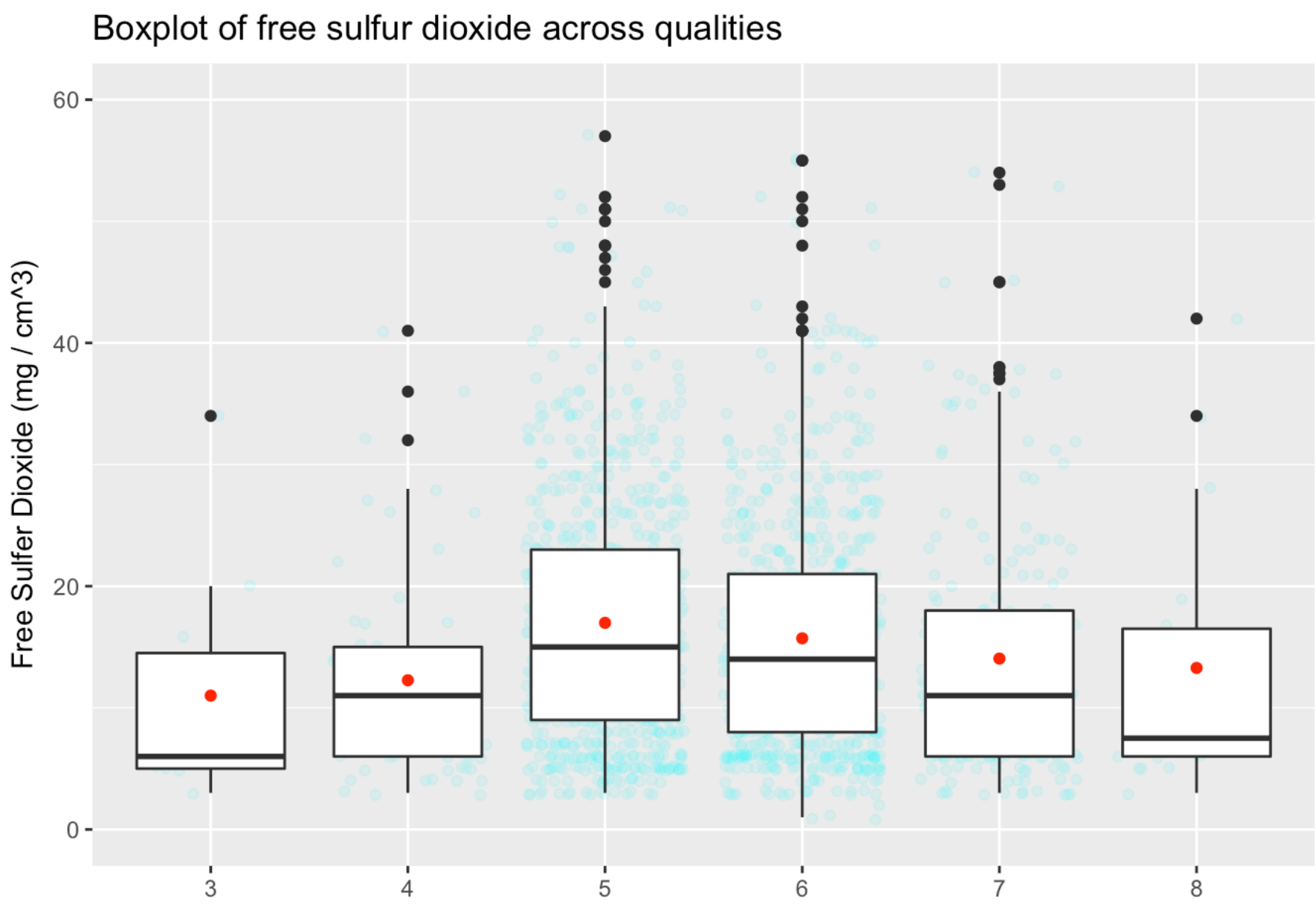
Now let’s take a look at the sweetness factor(residual sugar), there is almost no correlation between residual sugar and quality.



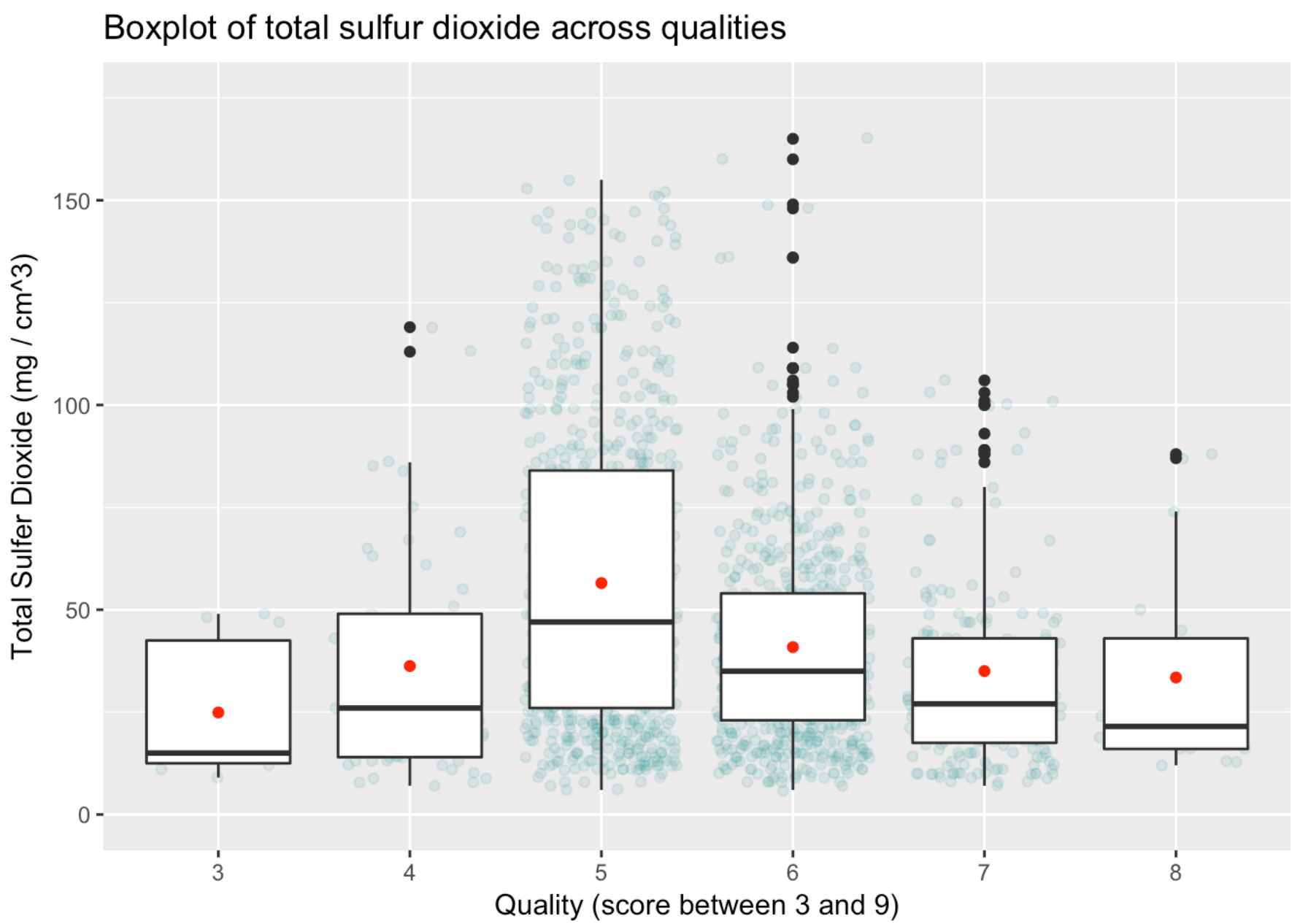
What about chlorides, the graph shows a very weak negative correlation with quality



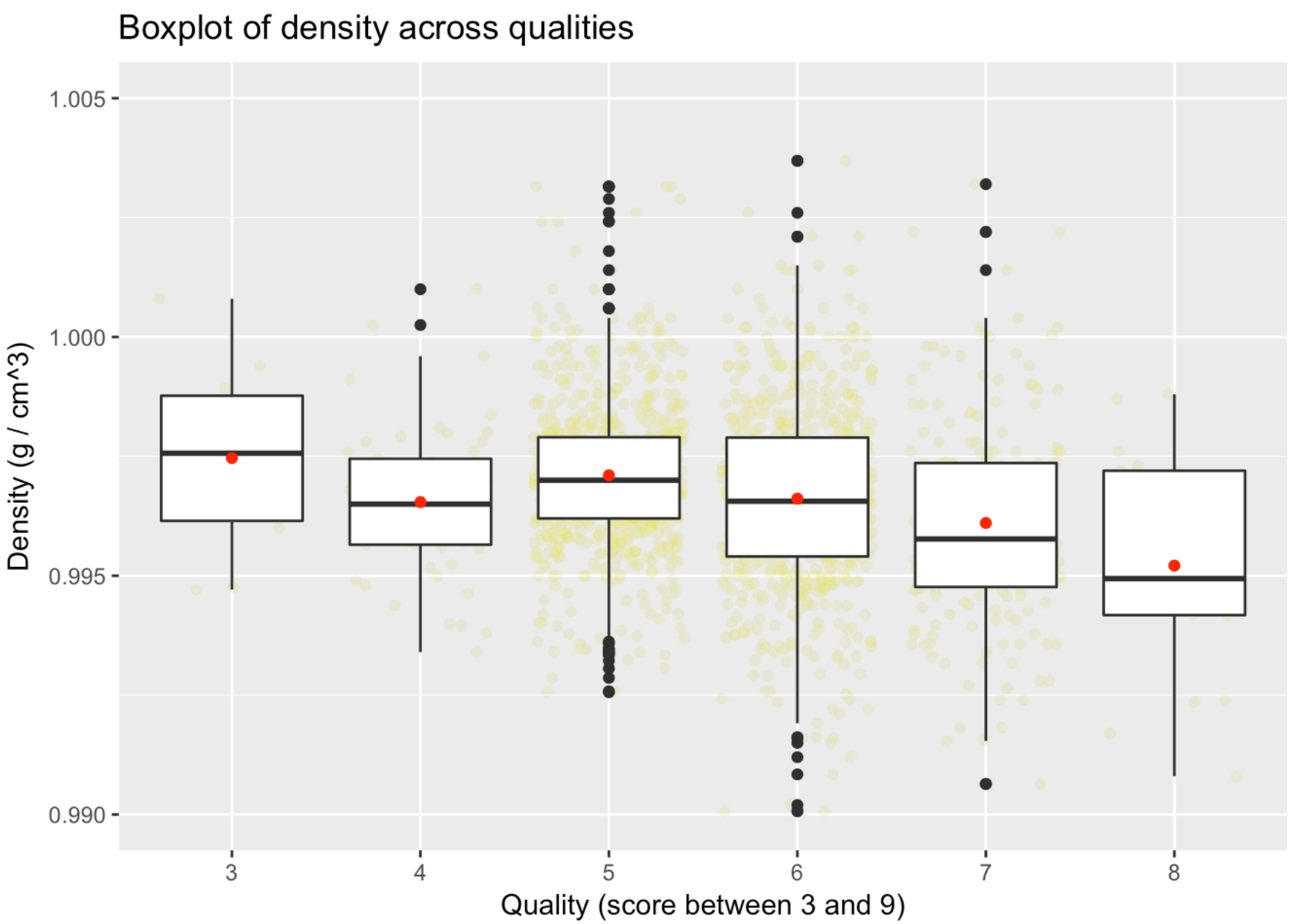
Free sulfur dioxide doesn’t have much correlation with quality, as the graph shows, it’s almost random without much effect on quality



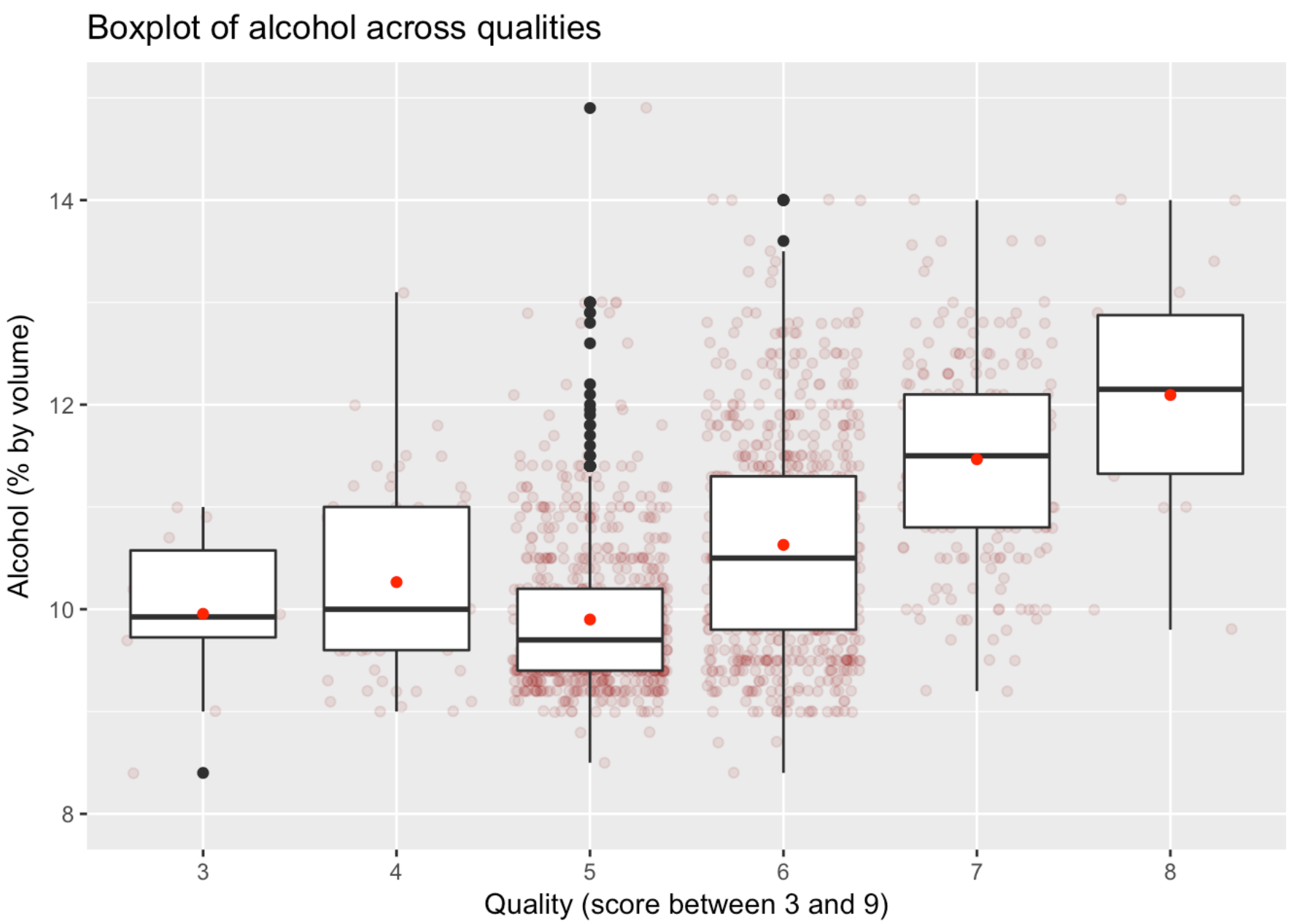
Total sulfur dioxide has almost the same correlation as free sulfur dioxide with quality, can not infer any impact on quality



As the graph shows, density has a relative weak negative correlation with quality

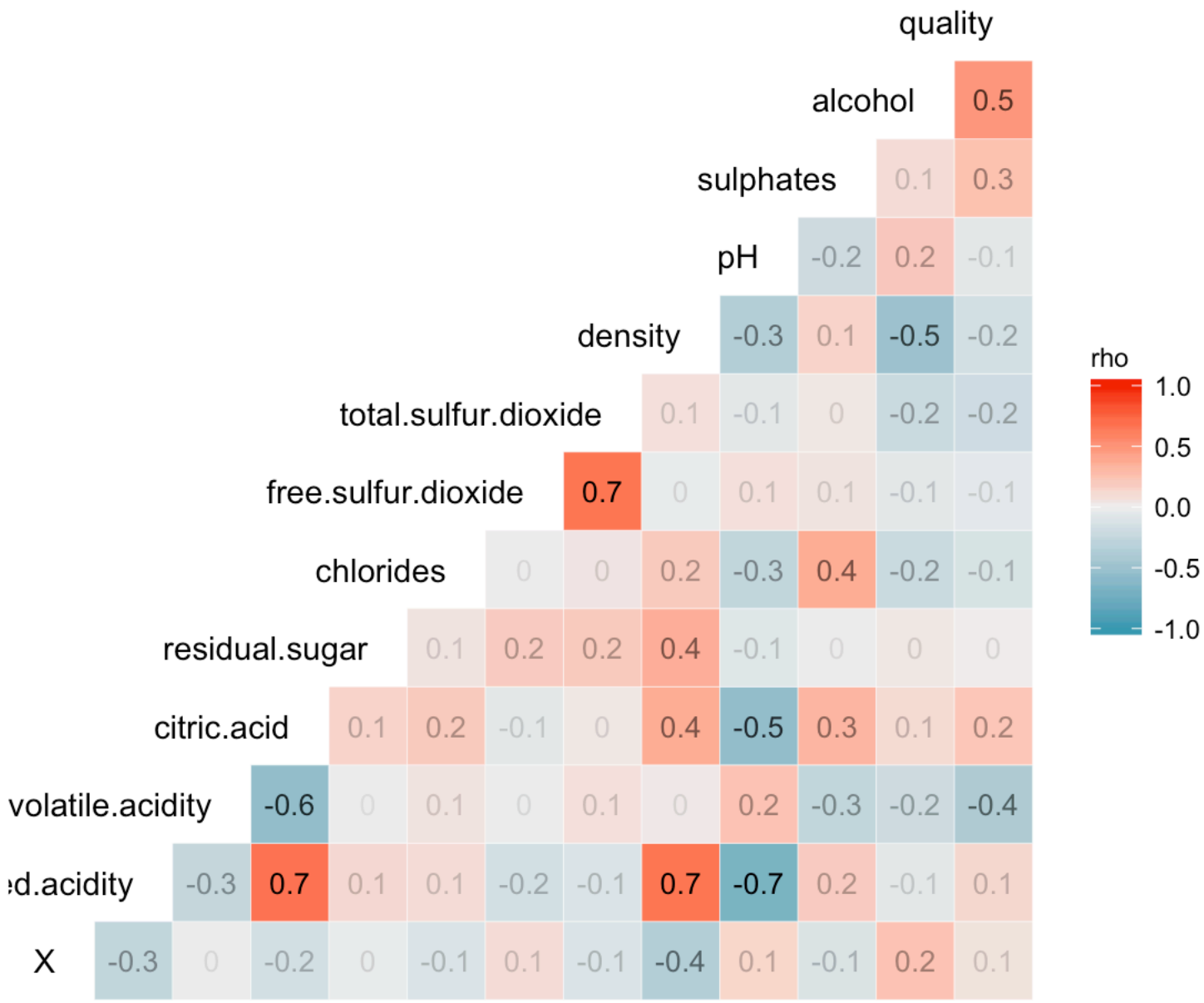


Last but not the least, alcohol shows somewhat promising positive correlation, which I'll invetigate further in the next section



Based on the graph, there does not appear to be a strong relationship between quality and total/free.sulfur.dioxide. SO2 is not evident in most of the wines, regardless of quality.

Let’s take a look at the overall correlation between variables.




```
##                                X fixed.acidity volatile.acidity
## X                    1.000000000   -0.26848392     -0.008815099
## fixed.acidity       -0.268483920    1.00000000     -0.256130895
## volatile.acidity    -0.008815099   -0.25613089      1.000000000
## citric.acid         -0.153551355    0.67170343     -0.552495685
## residual.sugar      -0.031260835    0.11477672      0.001917882
## chlorides           -0.119868519    0.09370519      0.061297772
## free.sulfur.dioxide  0.090479643   -0.15379419     -0.010503827
## total.sulfur.dioxide -0.117849669   -0.11318144      0.076470005
## density             -0.368372087    0.66804729      0.022026232
## pH                   0.136005328   -0.68297819      0.234937294
## sulphates           -0.125306999    0.18300566     -0.260986685
## alcohol              0.245122841   -0.06166827     -0.202288027
## quality              0.066452608    0.12405165     -0.390557780
##                     citric.acid residual.sugar    chlorides
## X                   -0.15355136   -0.031260835 -0.119868519
## fixed.acidity        0.67170343    0.114776724  0.093705186
## volatile.acidity    -0.55249568    0.001917882  0.061297772
## citric.acid          1.00000000    0.143577162  0.203822914
## residual.sugar       0.14357716    1.000000000  0.055609535
## chlorides            0.20382291    0.055609535  1.000000000
## free.sulfur.dioxide -0.06097813    0.187048995  0.005562147
```

```
## total.sulfur.dioxide  0.03553302    0.203027882  0.047400468
## density                0.36494718    0.355283371  0.200632327
## pH                    -0.54190414   -0.085652422 -0.265026131
## sulphates              0.31277004    0.005527121  0.371260481
## alcohol                0.10990325    0.042075437 -0.221140545
## quality                0.22637251    0.013731637 -0.128906560
##                      free.sulfur.dioxide total.sulfur.dioxide     density
## X                         0.090479643          -0.11784967 -0.36837209
## fixed.acidity            -0.153794193          -0.11318144  0.66804729
## volatile.acidity         -0.010503827           0.07647000  0.02202623
## citric.acid              -0.060978129           0.03553302  0.36494718
## residual.sugar            0.187048995           0.20302788  0.35528337
## chlorides                 0.005562147           0.04740047  0.20063233
## free.sulfur.dioxide       1.000000000           0.66766645 -0.02194583
## total.sulfur.dioxide      0.667666450           1.00000000  0.07126948
## density                  -0.021945831           0.07126948  1.00000000
## pH                        0.070377499          -0.06649456 -0.34169933
## sulphates                 0.051657572           0.04294684  0.14850641
## alcohol                  -0.069408354          -0.20565394 -0.49617977
## quality                  -0.050656057          -0.18510029 -0.17491923
##                                pH    sulphates     alcohol     quality
## X                       0.13600533 -0.125306999  0.24512284  0.06645261
## fixed.acidity          -0.68297819  0.183005664 -0.06166827  0.12405165
## volatile.acidity        0.23493729 -0.260986685 -0.20228803 -0.39055778
## citric.acid            -0.54190414  0.312770044  0.10990325  0.22637251
## residual.sugar         -0.08565242  0.005527121  0.04207544  0.01373164
## chlorides              -0.26502613  0.371260481 -0.22114054 -0.12890656
## free.sulfur.dioxide     0.07037750  0.051657572 -0.06940835 -0.05065606
## total.sulfur.dioxide   -0.06649456  0.042946836 -0.20565394 -0.18510029
## density                -0.34169933  0.148506412 -0.49617977 -0.17491923
## pH                      1.00000000 -0.196647602  0.20563251 -0.05773139
## sulphates              -0.19664760  1.000000000  0.09359475  0.25139708
## alcohol                 0.20563251  0.093594750  1.00000000  0.47616632
## quality                -0.05773139  0.251397079  0.47616632  1.00000000
```

Correlation Coefficient: based on the correlation coefficient, only few variables has large strength of association (> 0.5 or < -0.5). citric.acid - fixed.acidity density - fixed.acidity pH - fixed.acidity critic.acid - volatile.acidity pH - ctic.acid total.sulfur.dioxide - free.sulfur.dioxide Additionally, none of the variable has strong corrleation with quality, only “alcohol” comes closest at 0.47.

Bivariate Analysis

Summary: based on the correlation coefficient and graph, there are weak correlation between the citric.acid(freshness), residual.sugar(sweatness), chlorides(saltyness) and quality. Though the preliminary grouping results showed a possible correlation.

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?

All three variables(citric.acid, residual.sugar, cholrides) has weak or none correlation with other variables, except critic.acid and pH.

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

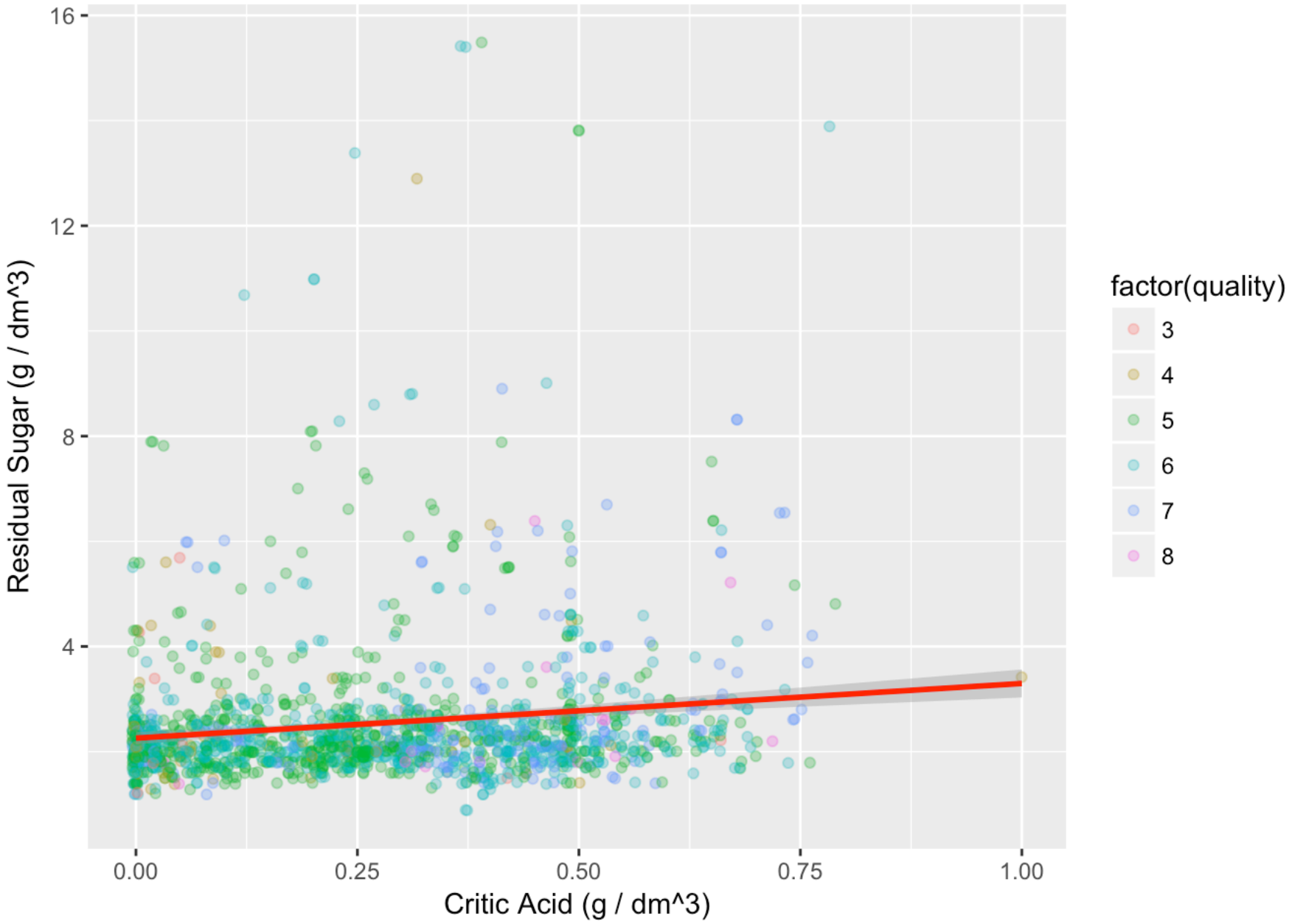
There is strong correlation between fixed.acidity with density, citric.acid, and pH. This means fixed.acidity is a strong component of red wine. Although, this is kind of expected due to the acid nature. But in the same vein, citric.acid and volatile.acidity does not have strong correlation with those same variables. Based on these findings, I suspect the rating can be subjective, instead of objective. Which we will investigate further.

What was the strongest relationship you found?

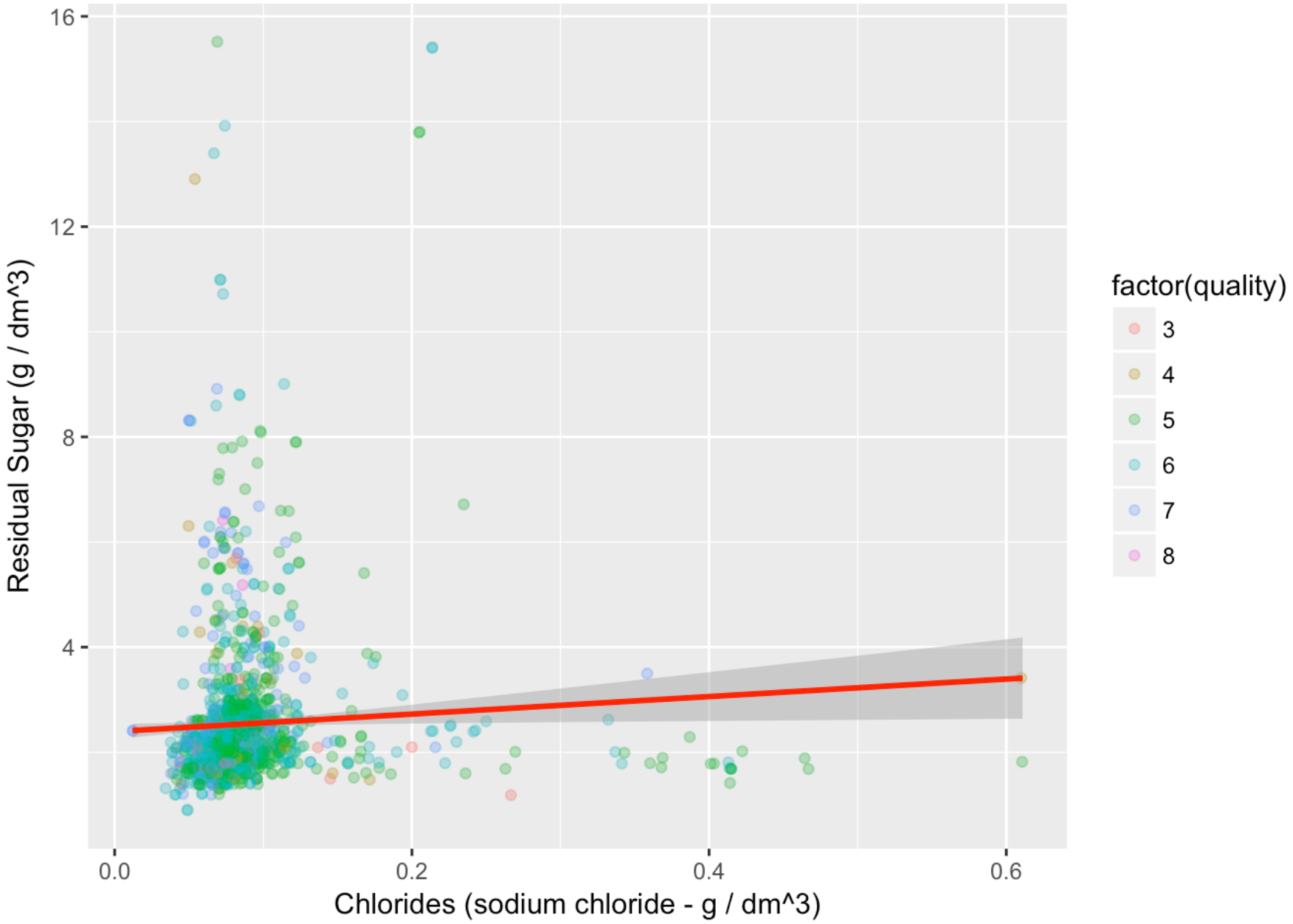
Between fixed.acidity with density, citric.acid, and pH.

Multivariate Plots Section

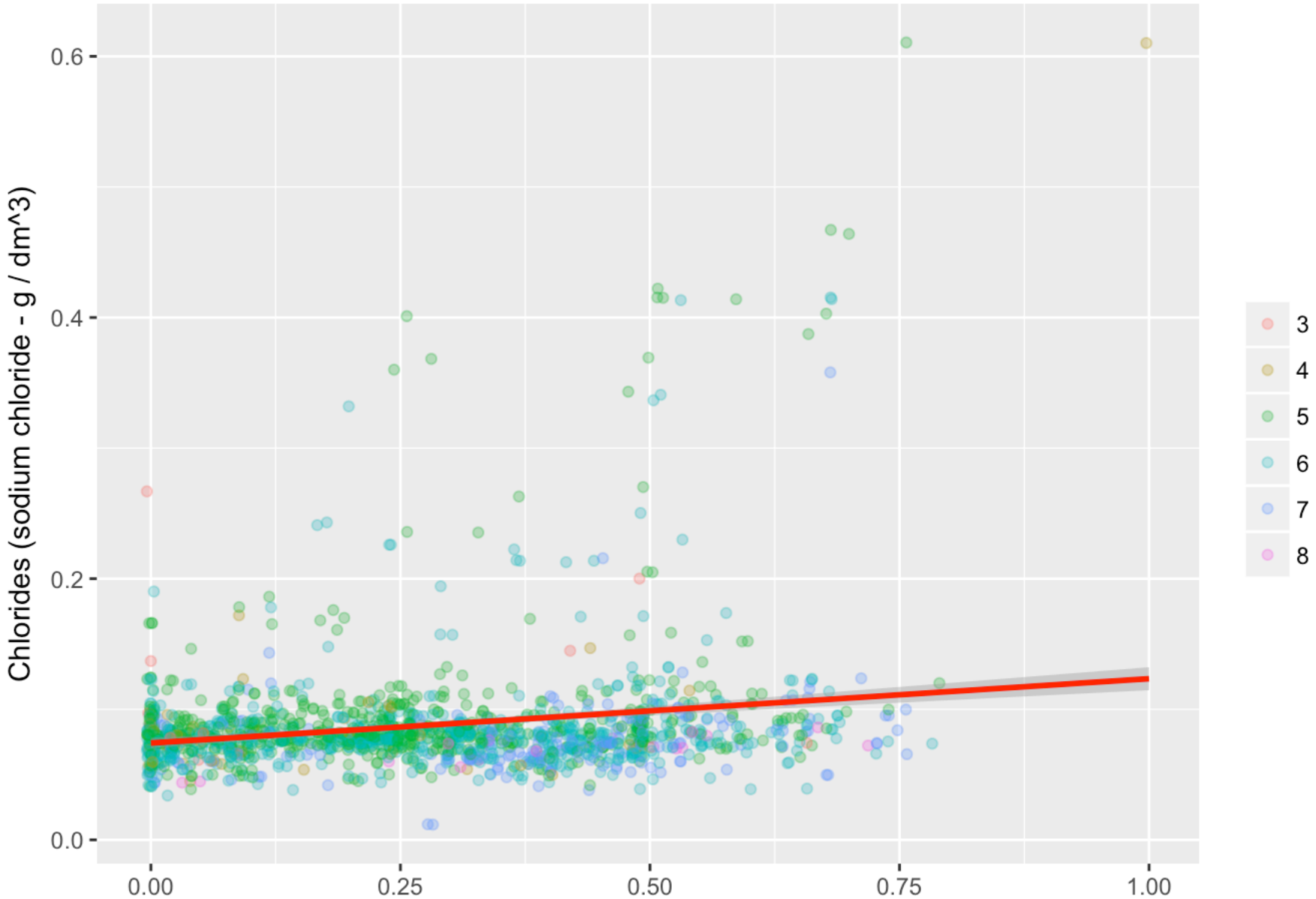
Critric acid does not seems to have a strong relationship with residual suger in term of quality



Also is the relationship between chlorides and residual suger

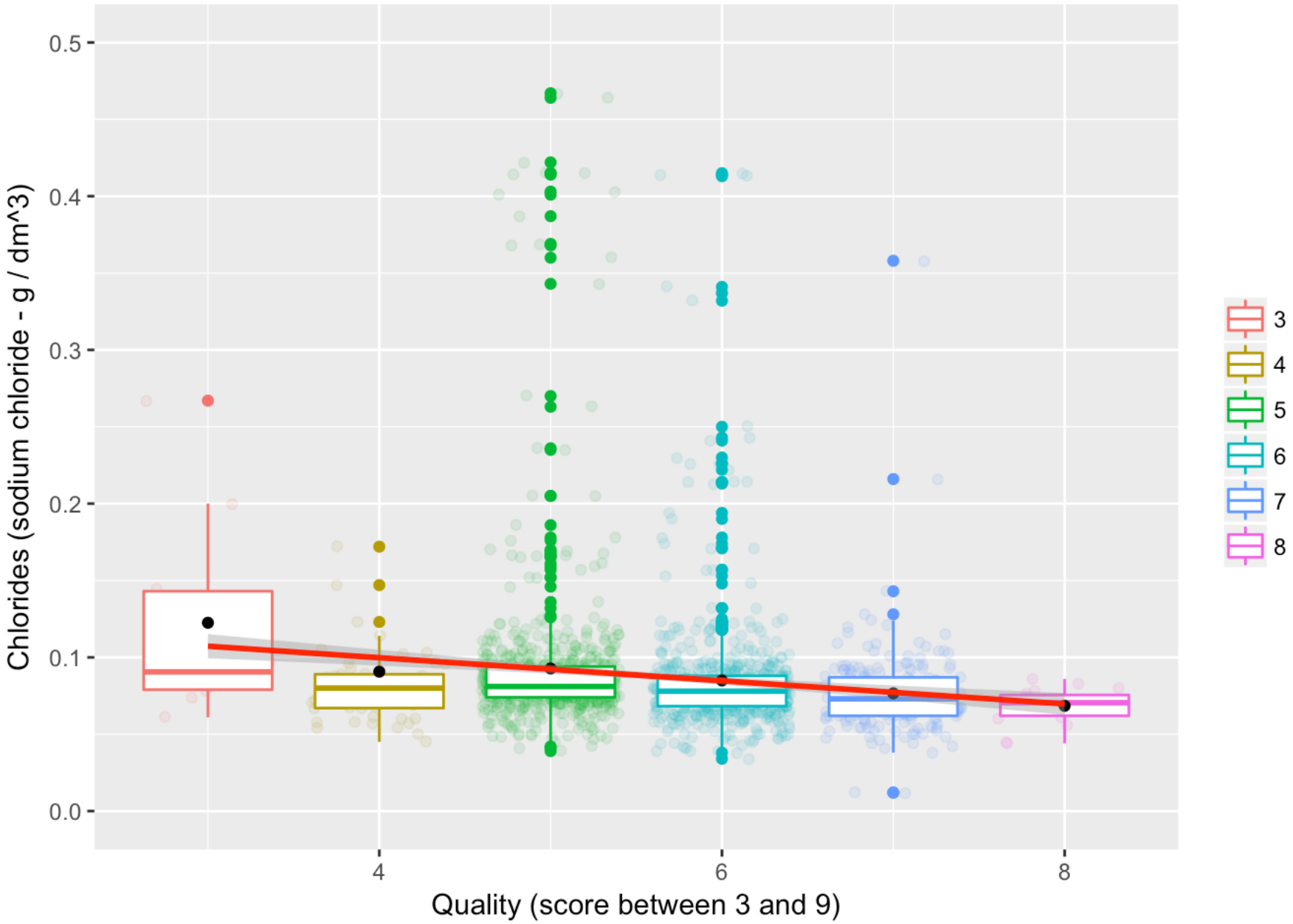


So is between citric acid and chlorides



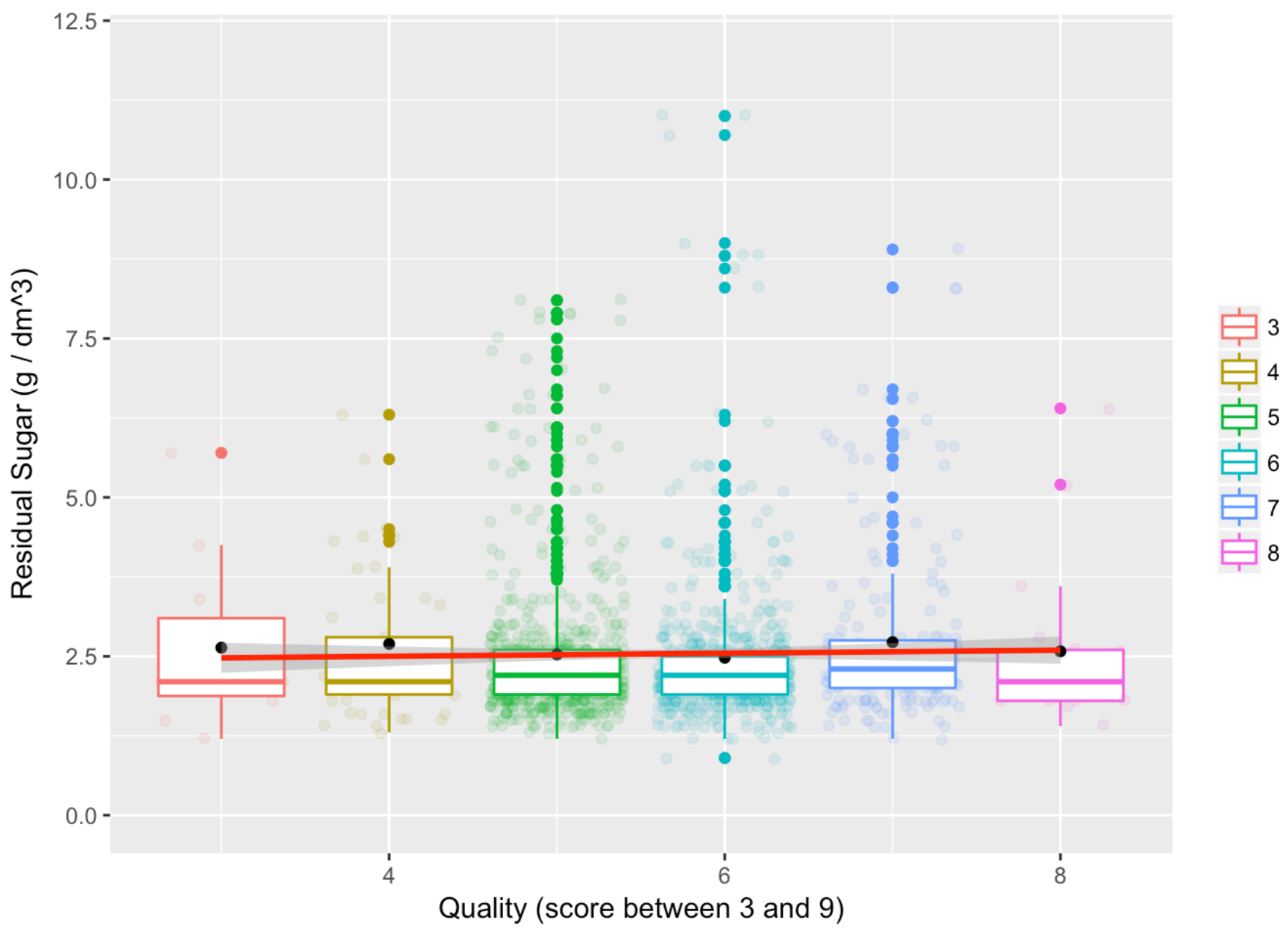
Citric Acid (g / dm^3)

Now with quality, let's take a look at some of the variables. Hope it will clue us in. First stop: chlorides.

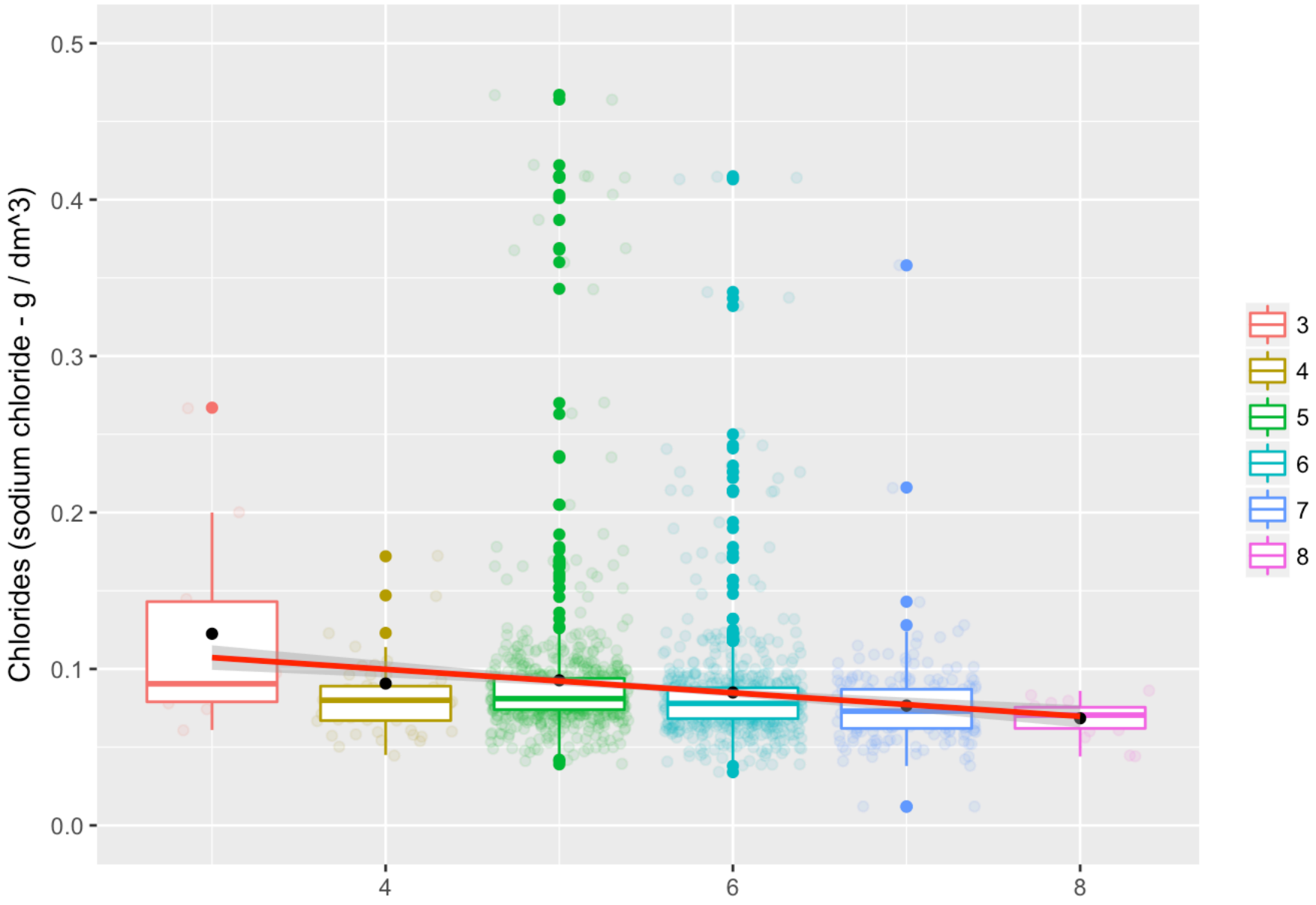


But the graph is not promising, the correlation is weak and negative.

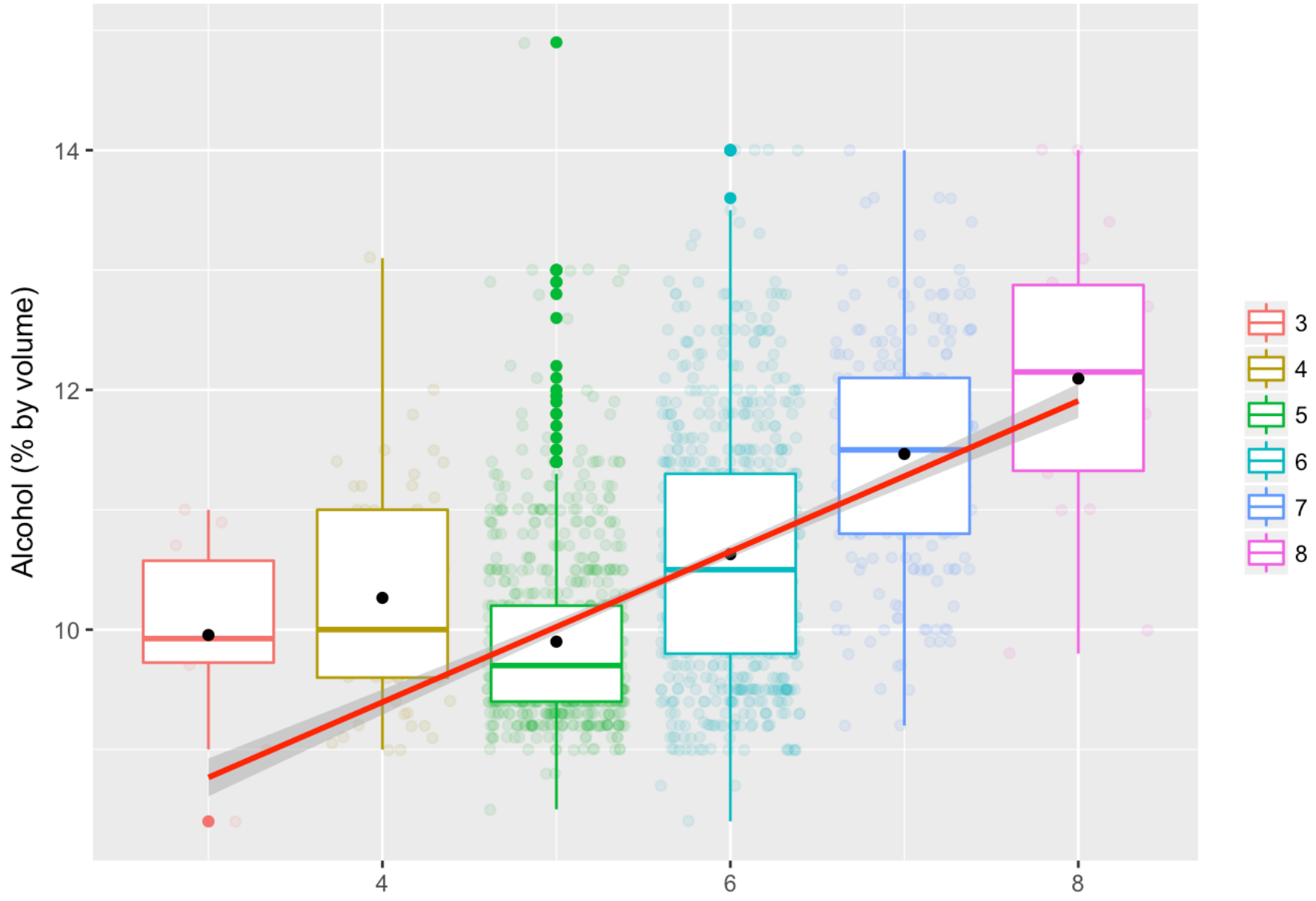
Also, sweetness does not seem to be a factor in term of quality



Overall chlorides has a weak negative correlation with quality



These charts confirmed my suspicion on citric acid which does not seem to play a significant role in quality. Now let’s take a look at



Quality (score between 3 and 9)

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 graphs illustrated what I discovered the previous section. There are weak correlation between the critic.acid, residual.sugar, and cholrides.

Were there any interesting or surprising interactions between features?

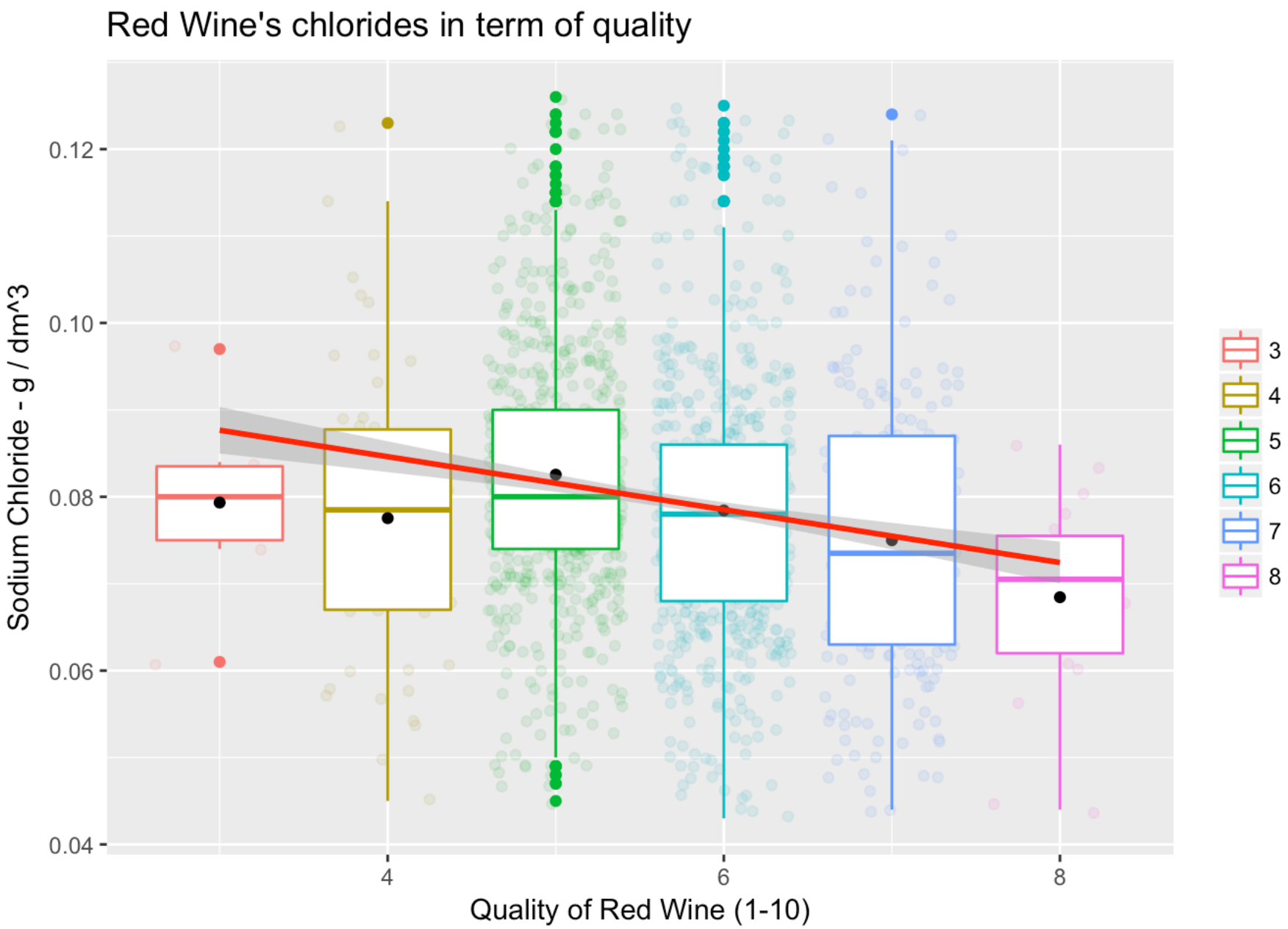
Nothing out of ordinary, except some outliers on many graphs.

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

Due to the data's cleaness, it was fairly easy create graph and not much difficulty was encontered. This may not be the same when apply to other datasets. ——

Final Plots and Summary

Plot One



Description One

There is no significant relationship between the chloride and wine quality. Remove this varilable as a key variable responsible for wine's quality.

Plot Two

Description Two

This graph also shows the lack of relationship between residual sugar and quality of wine. Even though in common sense, the sweetness usually is associated with the quality of wine as personal taste. But the graph does not show correlation between them.

Plot Three

Description Three

Critical acid has strong correlation than the previous two variables in relation of quality. With the data points listed on the garph, it clearly presents the image that the correlation is still not strong enough. Even though people would think that is one of the common criteria for wine quality.

Reflection

Without break down the data, it is easy to get lost in this giant dataset and let conventional wisdom takes over. As I set out at the beginning, the relationship among variables were totally unexpected. As I went through the whole process, it became more and more clear that my expectation was false. And the correlation and graphs proved me wrong. In data analyasis, how to differentiate personal belief and data can be challenge. I need to have an empty mindset to start with. That way, it will be easier to fall into the “trap”.

A further investigation into the dataset can be wether correlation equal causation for this dataset. Even though we find some variables stronger correlation with quality. But was it the correct correlation, or was it by chance.