Explore and Summarize Data

Wenke Huang

August 11, 2017

0. Introduction

In this project, I would like to explore and summarize the dataset `Red Wine Quality`.

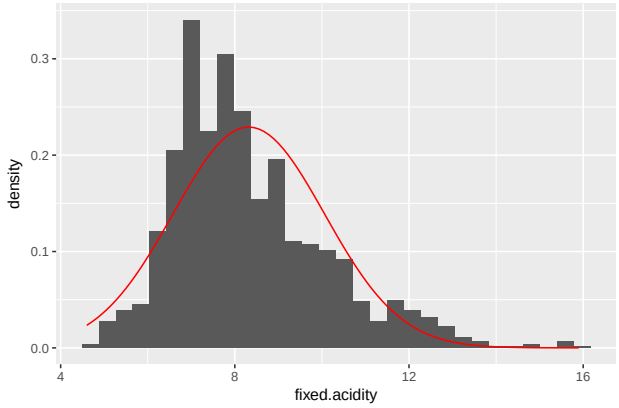
1. A stream-of-consciousness analysis and exploration

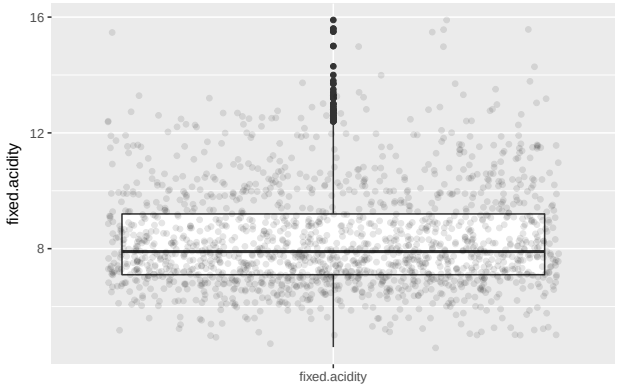
- **Question 1:** What are the distributions of our variables? Can I present them quantitatively and/or visually?

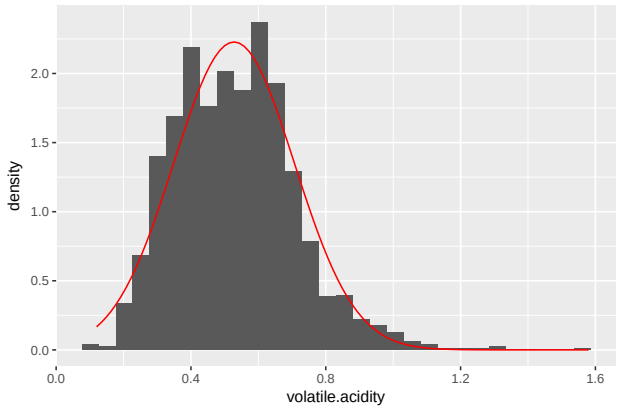
```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar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

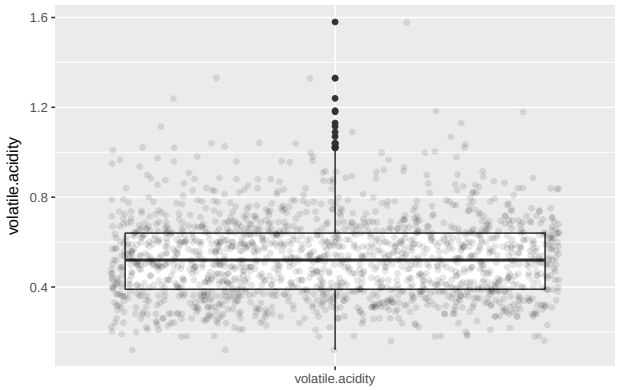
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :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
```

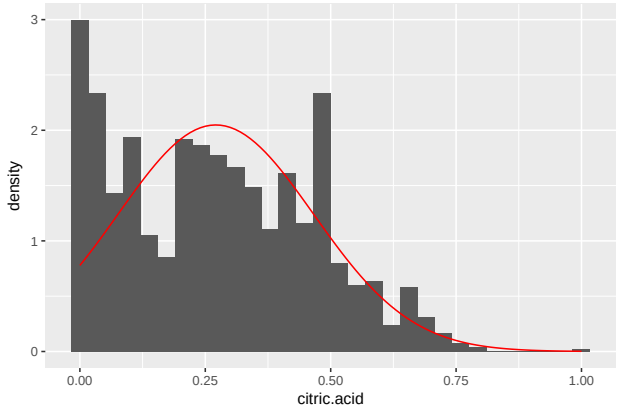
```
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

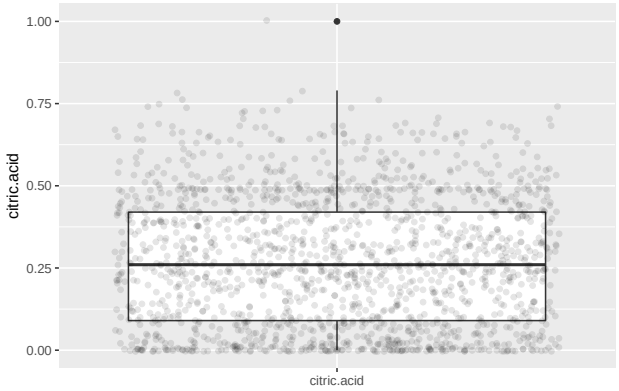


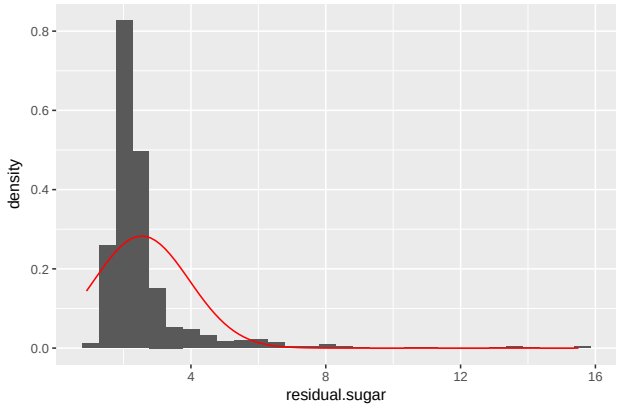


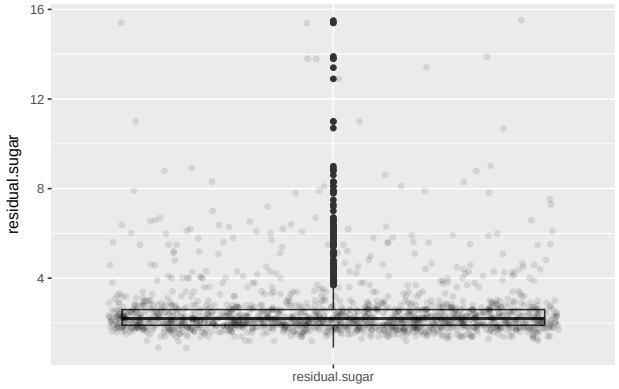


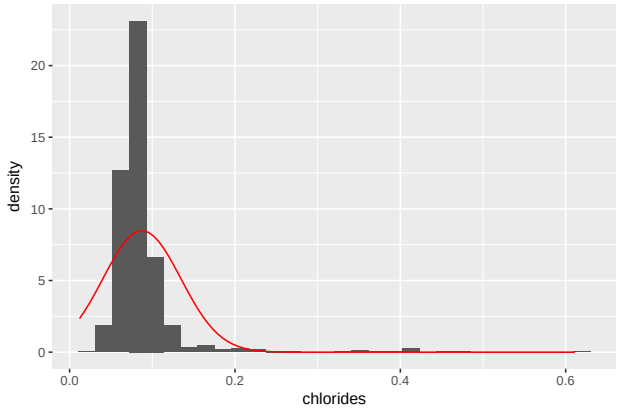


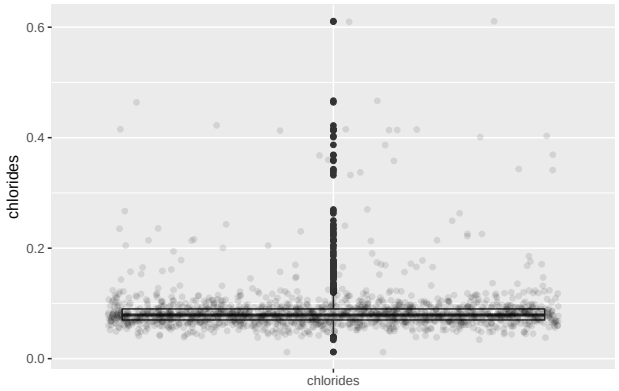

density
citric.acid
0.00
0.25
0.50
0.75
1.00
0
1
2
3

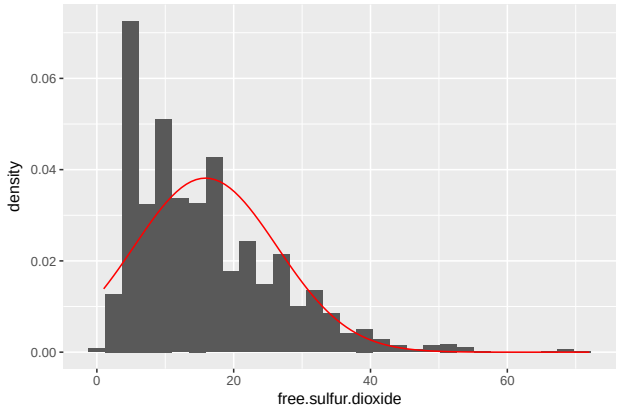

citric.acid
citric.acid
0.00
0.25
0.50
0.75
1.00

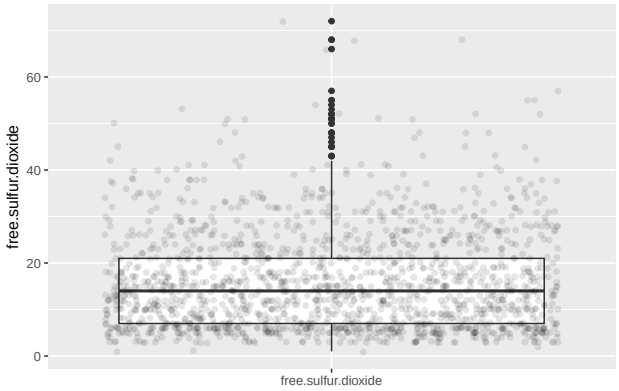

density
residual.sugar
0.0
0.2
0.4
0.6
0.8
4
8
12
16

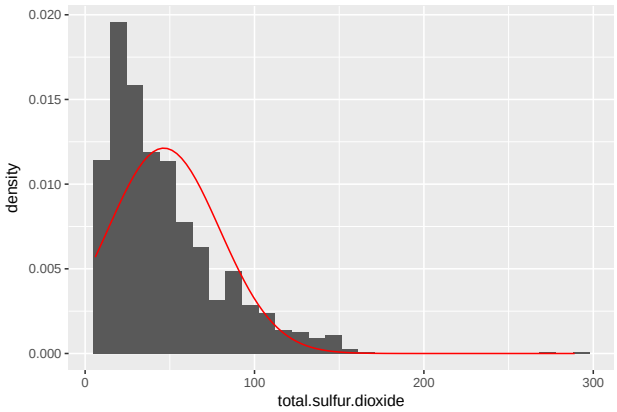

residual.sugar
residual.sugar
4
8
12
16

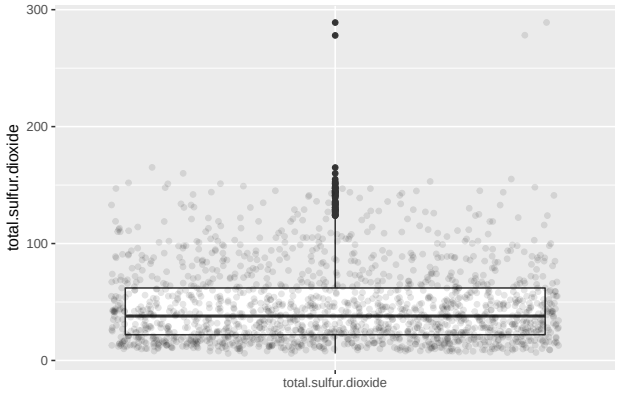

density
chlorides
0
5
10
15
20
0.0
0.2
0.4
0.6

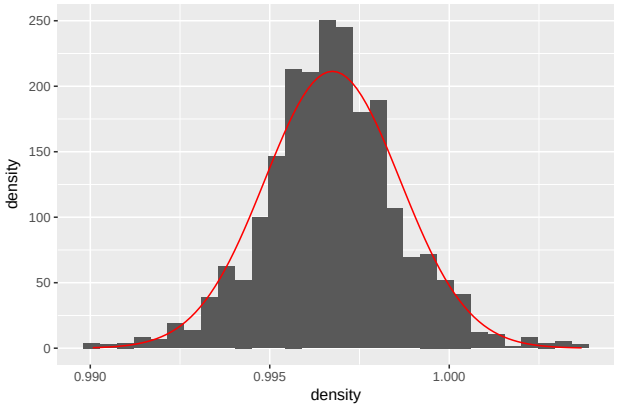

chlorides
chlorides
0.0
0.2
0.4
0.6

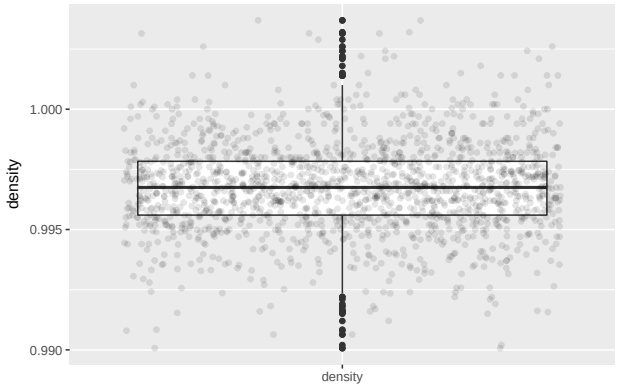

density
free.sulfur.dioxide
0.00
0.02
0.04
0.06
0
20
40
60

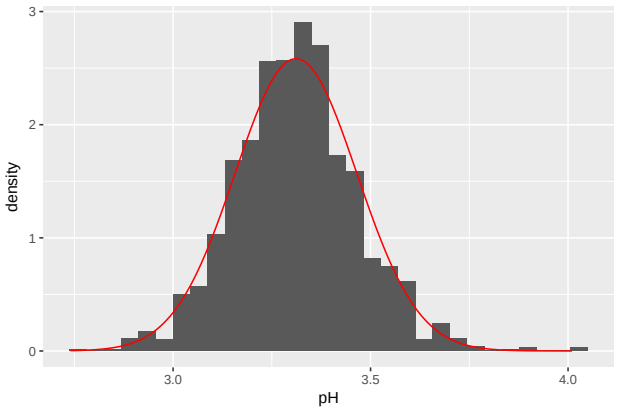

free.sulfur.dioxide
free.sulfur.dioxide
0
20
40
60

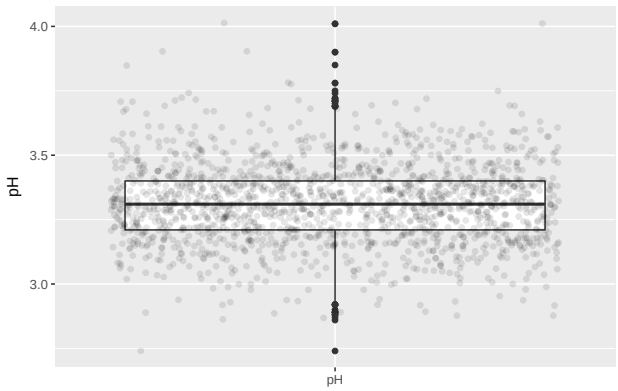

density
0.020
0.015
0.010
0.005
0.000
0
100
200
300
total.sulfur.dioxide

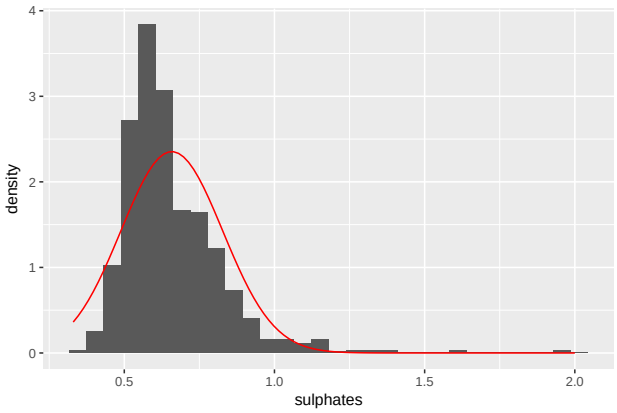

total.sulfur.dioxide
300
200
100
0
total.sulfur.dioxide

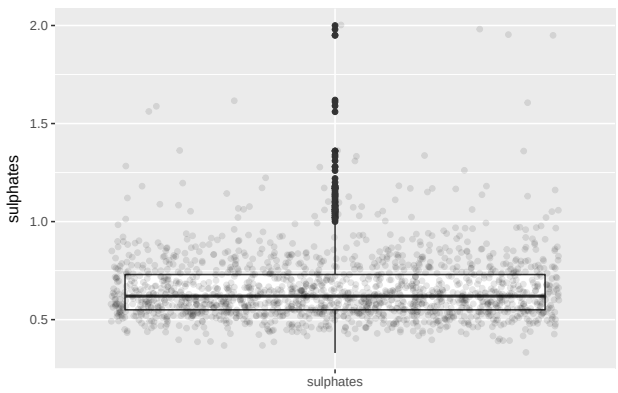

density
250
200
150
100
50
0
0.990
0.995
1.000
density

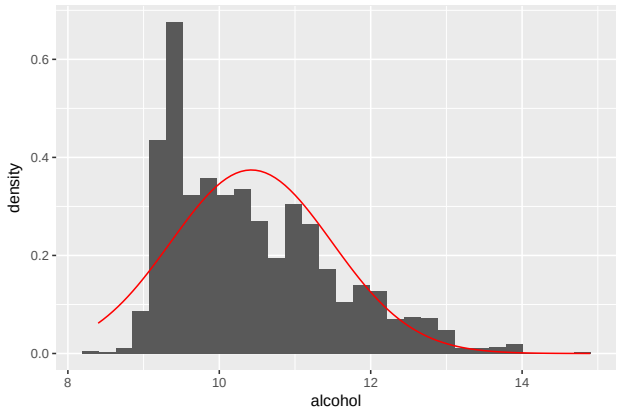

density
1.000
0.995
0.990
density

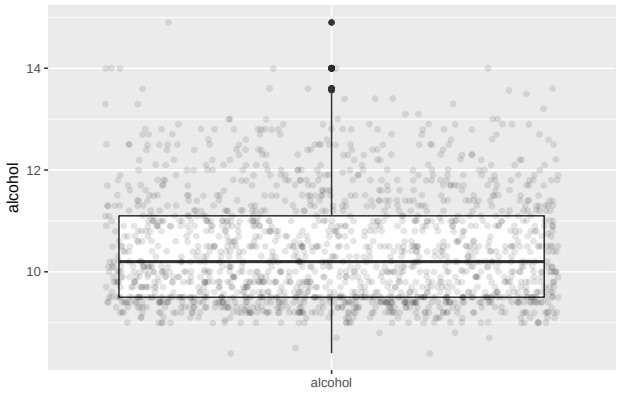

density
3
2
1
0
3.0
3.5
4.0
pH

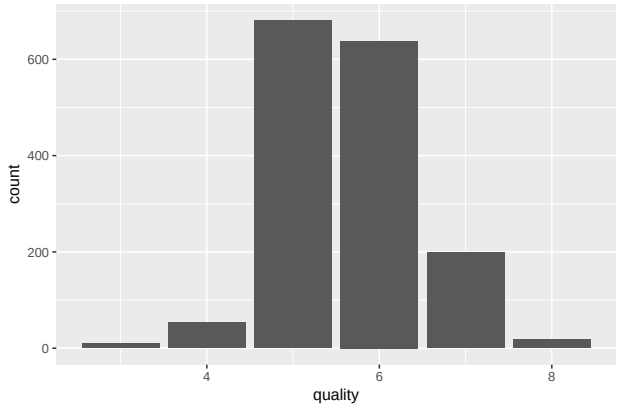

pH
4.0
3.5
3.0
pH


density
4
3
2
1
0
0.5
1.0
1.5
2.0
sulphates


sulphates
2.0
1.5
1.0
0.5
sulphates



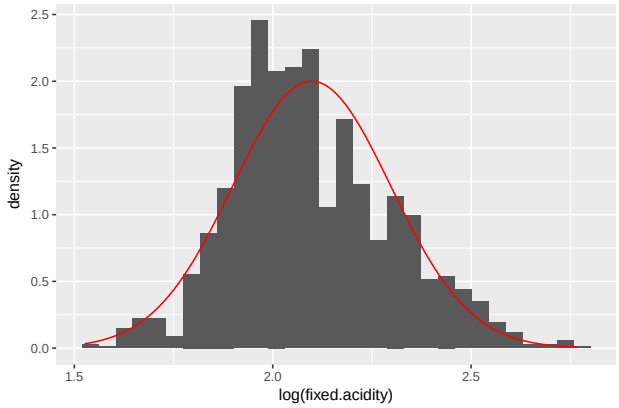


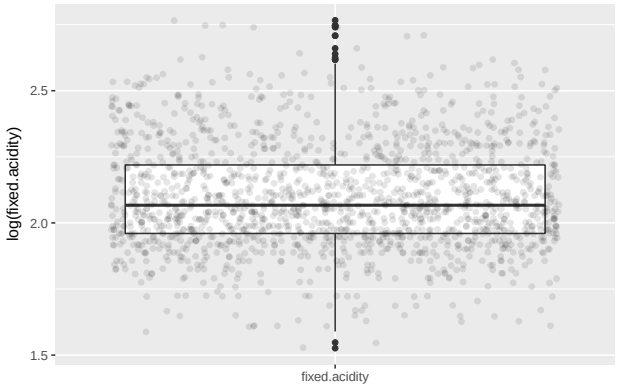


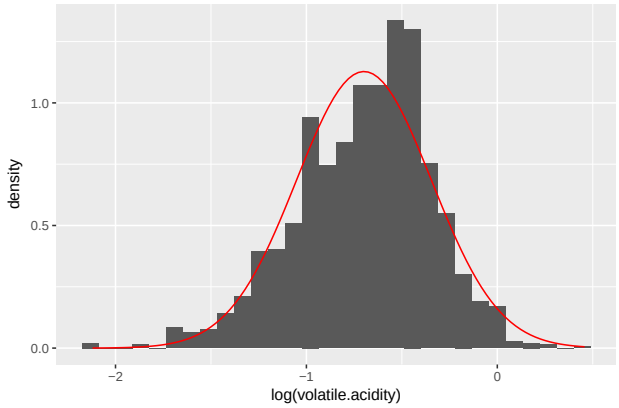
The five number summary statistics and mean values shows that many variables, such as `fixed.acidity` and `chlorides`, have positively skewed distribution. Additionally, We can find that the histograms of those variables that have positive skewness don't fit with their coresponding red normal distribution curves. We will perform logarithm transformation to see if they follow log-normal distribution apart from quality.

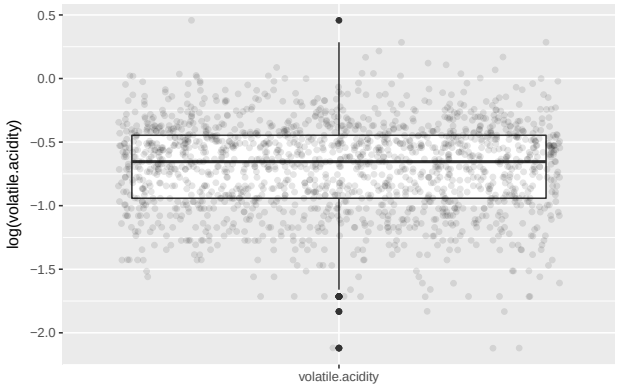
At the same time, variable `volatile.acidity`, `density` and `pH` fit with their normal distribution curves pretty well.

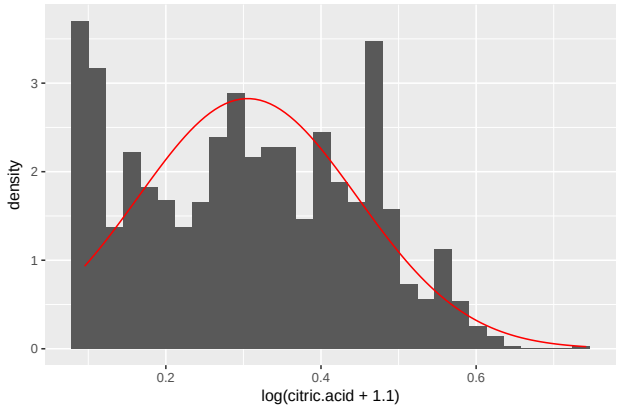
In additiona to histograms, box plots can help us understand how our variables distribute, and at the same time detect the outliers. From the boxplots above, they clearly tell us that many variables tend to be positively skewed, like `residual.sugar`, `chlorides`, etc.

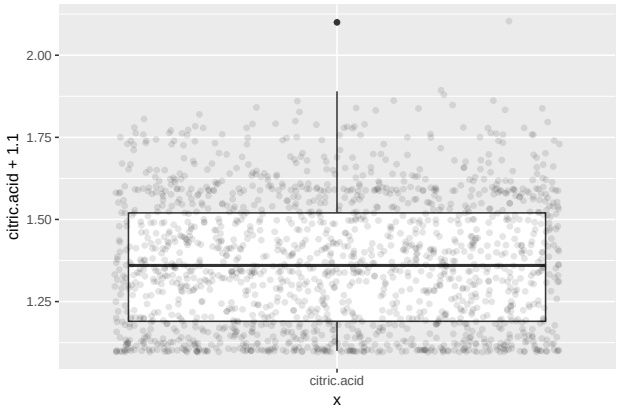

density
log(fixed.acidity)

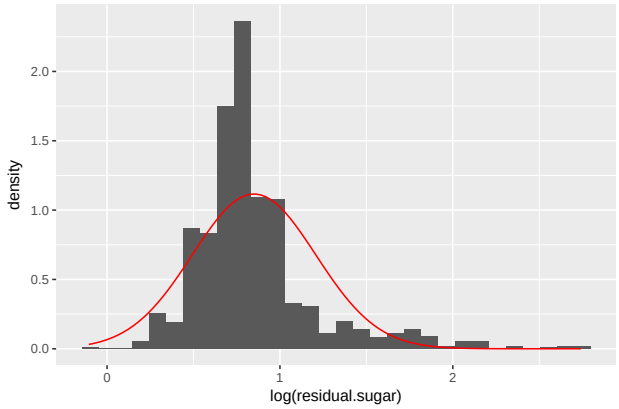

log(fixed.acidity)
fixed.acidity

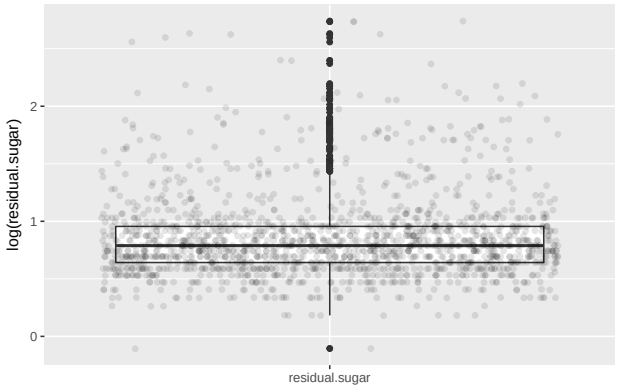

density
log(volatile.acidity)

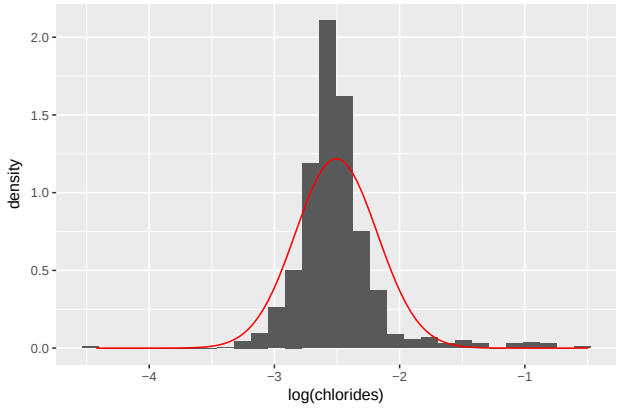

log(volatile.acidity)
volatile.acidity

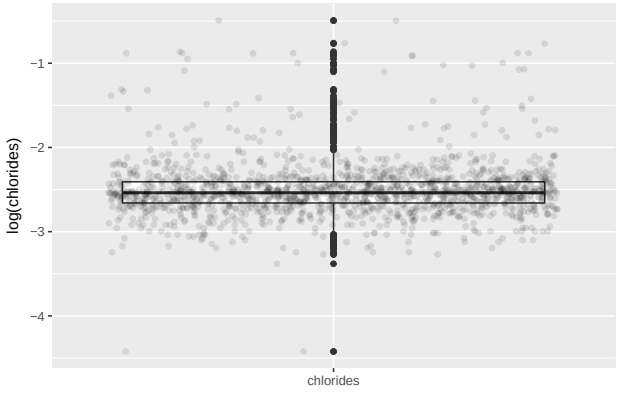

density
log(citric.acid + 1.1)

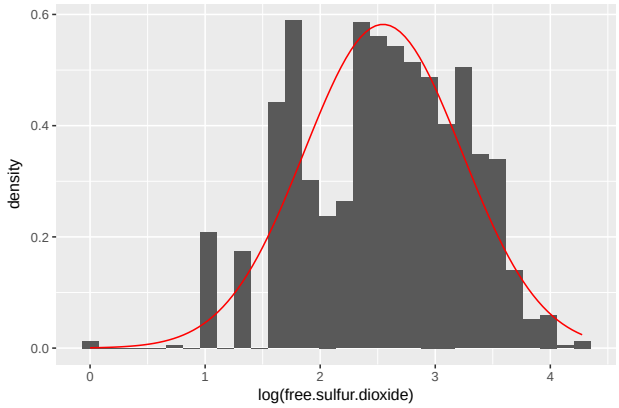

citric.acid + 1.1
citric.acid
x

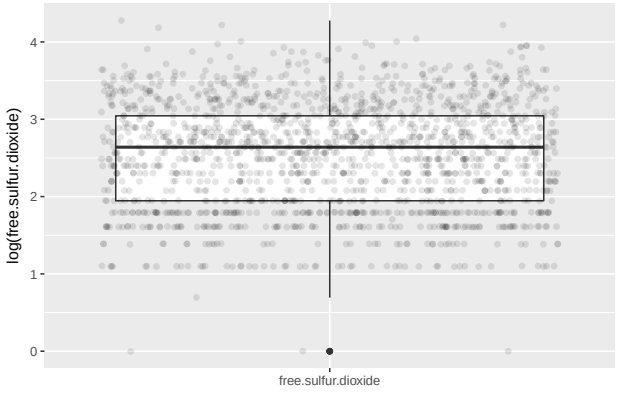

density
log(residual.sugar)

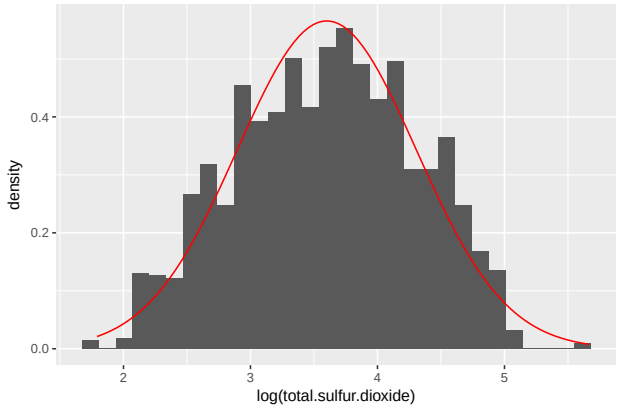

log(residual.sugar)
residual.sugar

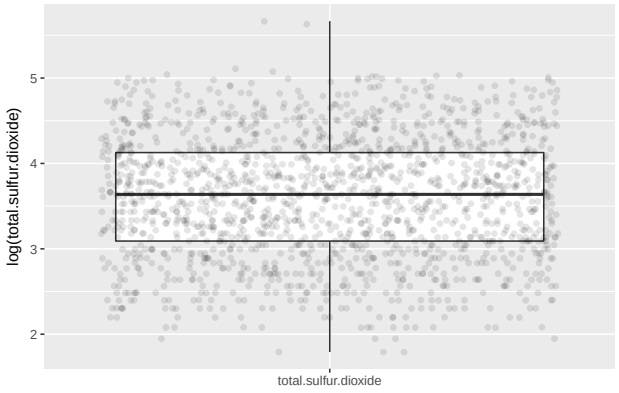

density
log(chlorides)

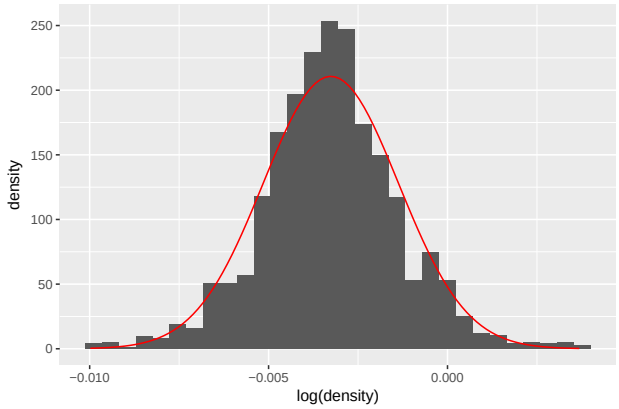

log(chlorides)
chlorides

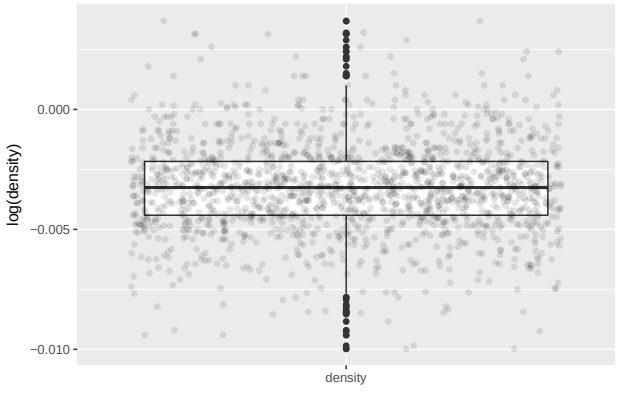

density
log(free.sulfur.dioxide)

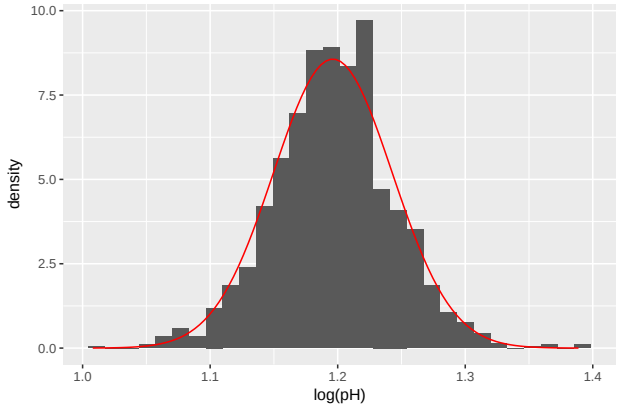

log(free.sulfur.dioxide)
free.sulfur.dioxide

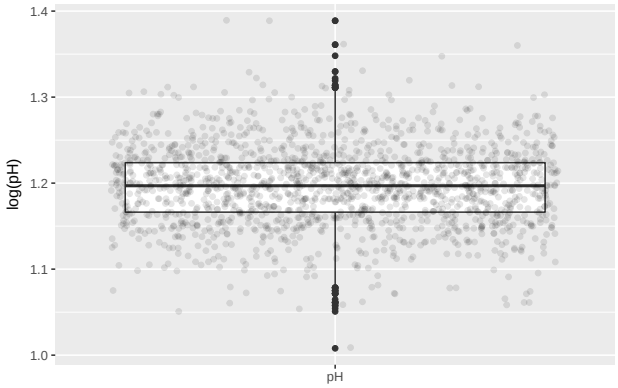

density
log(total.sulfur.dioxide)

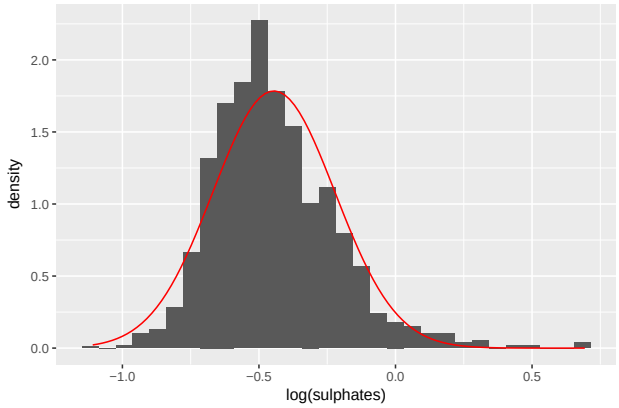

log(total.sulfur.dioxide)
total.sulfur.dioxide

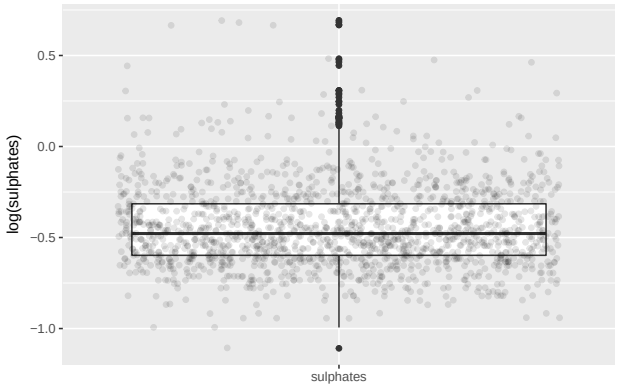

density
log(density)

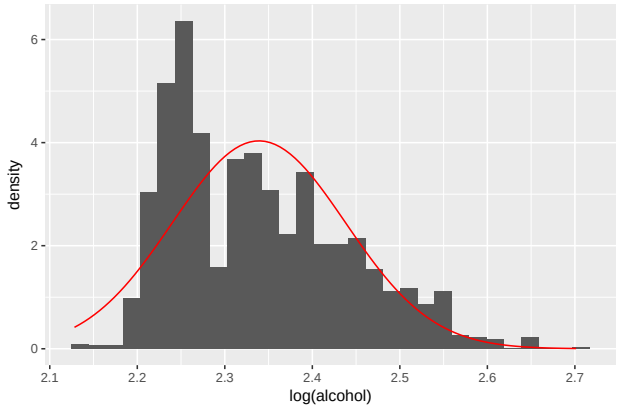

log(density)
density

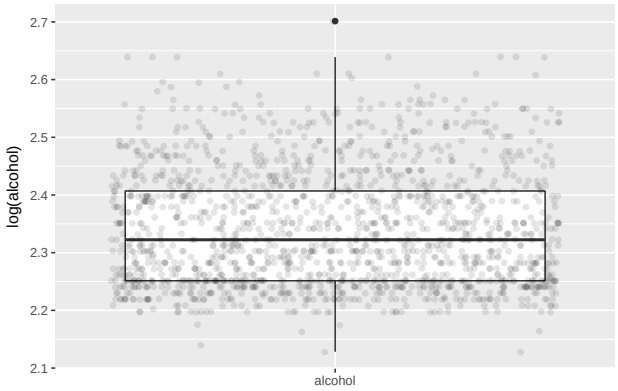


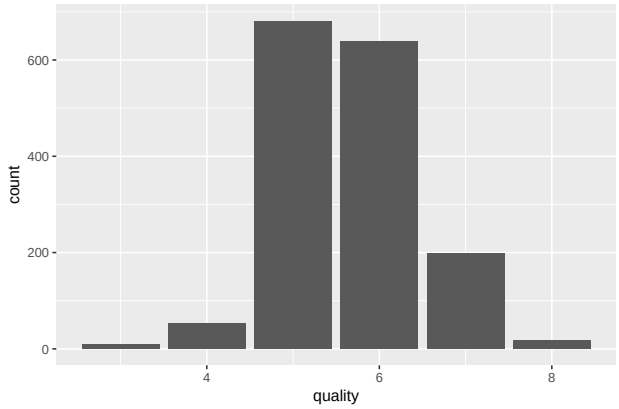












After transformation , some of the transformed variables do have a distribution that we expect, like

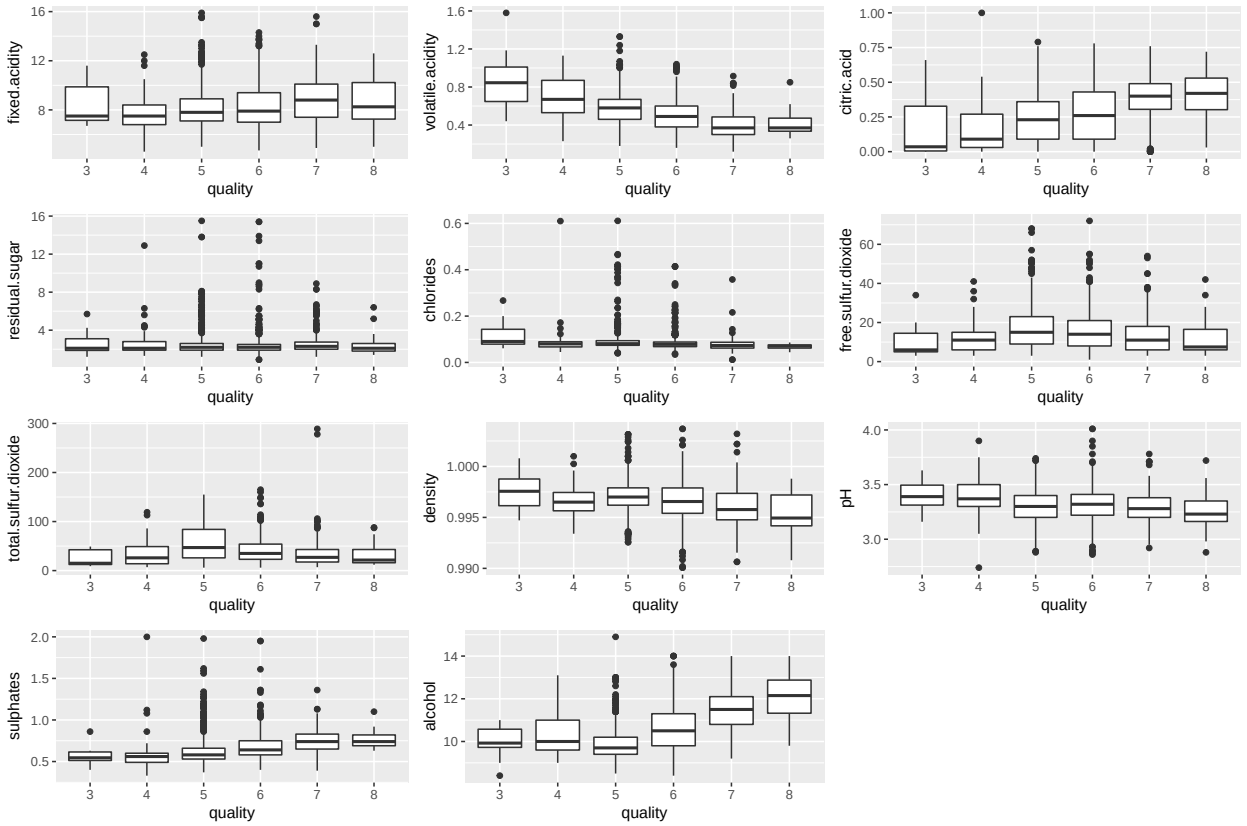
`fixed.acidity`, `free.sulfur.dioxide`, `total.sulfur.dioxide` and `sulphates`. Hence, the log-normal transformation of these three variables will be considered later.

- **Question 2:** Are these variables correlated with each other? What methods can help us find their correlationships?

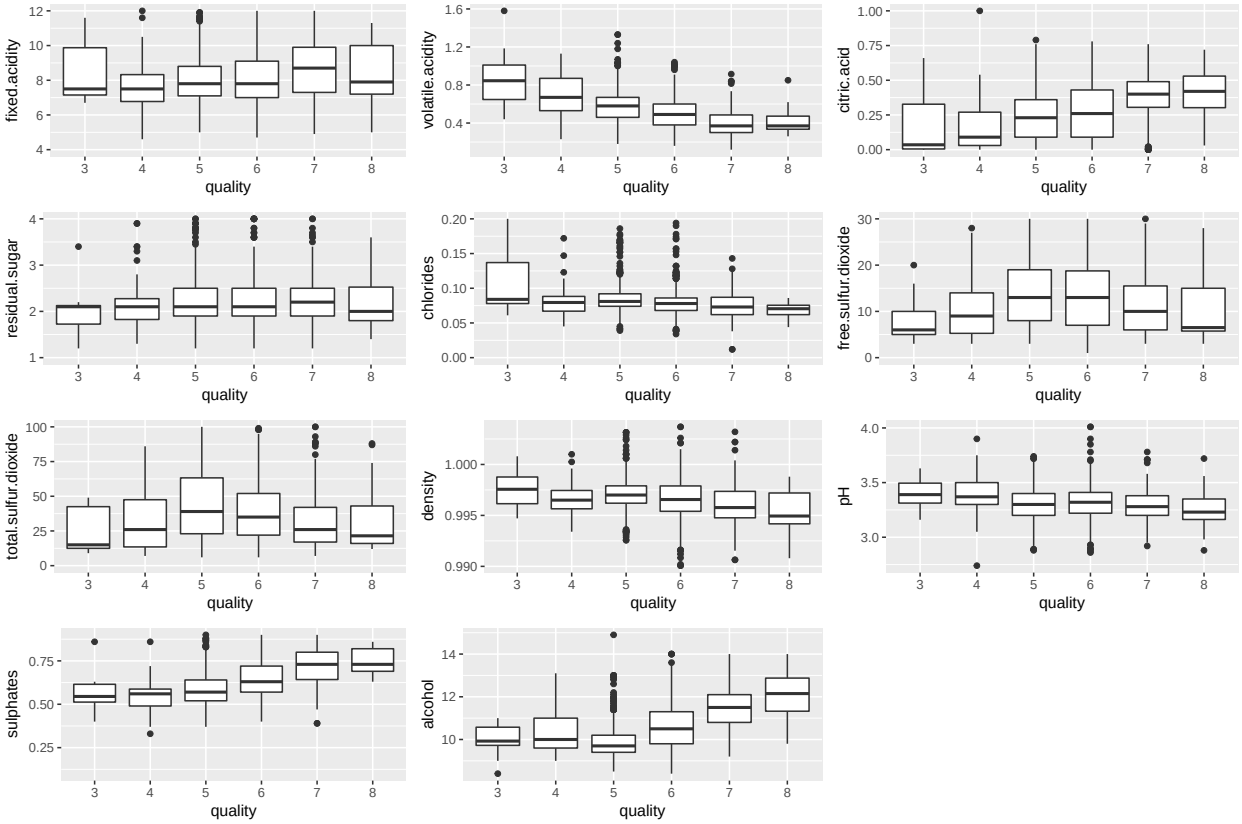
```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity                 1.000           -0.256       0.672
## volatile.acidity             -0.256            1.000      -0.552
## citric.acid                   0.672           -0.552       1.000
## residual.sugar                0.115            0.002       0.144
## chlorides                     0.094            0.061       0.204
## free.sulfur.dioxide          -0.154           -0.011      -0.061
## total.sulfur.dioxide         -0.113            0.076       0.036
## density                       0.668            0.022       0.365
## pH                           -0.683            0.235      -0.542
## sulphates                     0.183           -0.261       0.313
## alcohol                      -0.062           -0.202       0.110
##                       residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity                  0.115     0.094              -0.154
## volatile.acidity               0.002     0.061              -0.011
## citric.acid                    0.144     0.204              -0.061
## residual.sugar                 1.000     0.056               0.187
## chlorides                      0.056     1.000               0.006
## free.sulfur.dioxide            0.187     0.006               1.000
## total.sulfur.dioxide           0.203     0.047               0.668
## density                        0.355     0.201              -0.022
## pH                            -0.086    -0.265               0.070
## sulphates                      0.006     0.371               0.052
## alcohol                        0.042    -0.221              -0.069
##                       total.sulfur.dioxide density     pH sulphates alcohol
## fixed.acidity                       -0.113   0.668 -0.683     0.183  -0.062
## volatile.acidity                     0.076   0.022  0.235    -0.261  -0.202
## citric.acid                          0.036   0.365 -0.542     0.313   0.110
## residual.sugar                       0.203   0.355 -0.086     0.006   0.042
## chlorides                            0.047   0.201 -0.265     0.371  -0.221
## free.sulfur.dioxide                  0.668  -0.022  0.070     0.052  -0.069
## total.sulfur.dioxide                 1.000   0.071 -0.066     0.043  -0.206
## density                              0.071   1.000 -0.342     0.149  -0.496
## pH                                  -0.066  -0.342  1.000    -0.197   0.206
## sulphates                            0.043   0.149 -0.197     1.000   0.094
## alcohol                             -0.206  -0.496  0.206     0.094   1.000
```

```
## [1] "citric.acid & fixed.acidity"
## [1] "total.sulfur.dioxide & free.sulfur.dioxide"
## [1] "density & fixed.acidity"
## [1] "pH & fixed.acidity"
```

Based on the pearson correlation coefficient matrix, we can extract these four pairs of variables that have the absolute value of correlation coefficient not less than 0.6, which indicates that they are either strongly positively or negatively correlated.

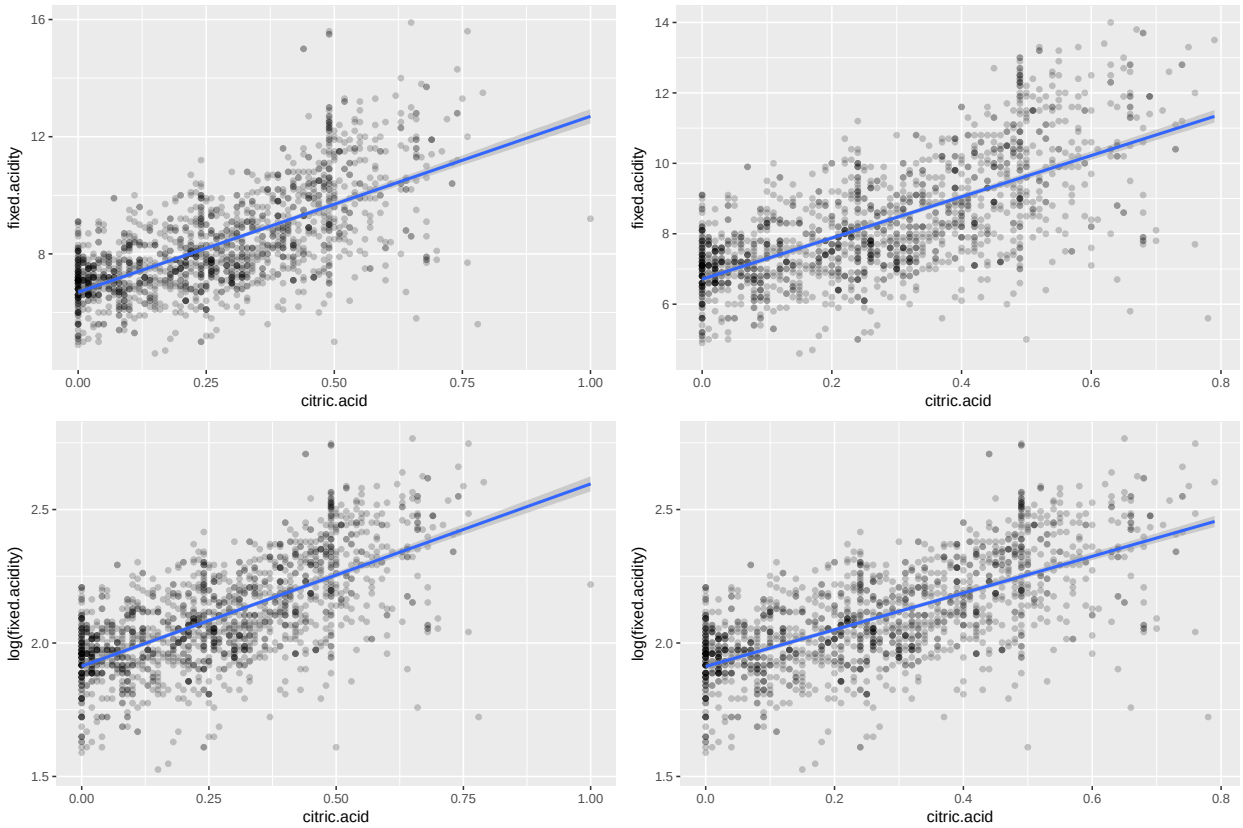


In addition to the correlationships among continuous variables, I am also interested in the correlationship between different continuous variables and variable `quality`. The boxplot is used here. Due to the outliers that we have mentioned above, it is a little harder to read some of these plots above, like `residual.sugar` and `chlorides`. We will restrict the y-axis targeting those variables and then re-draw the new boxplots.



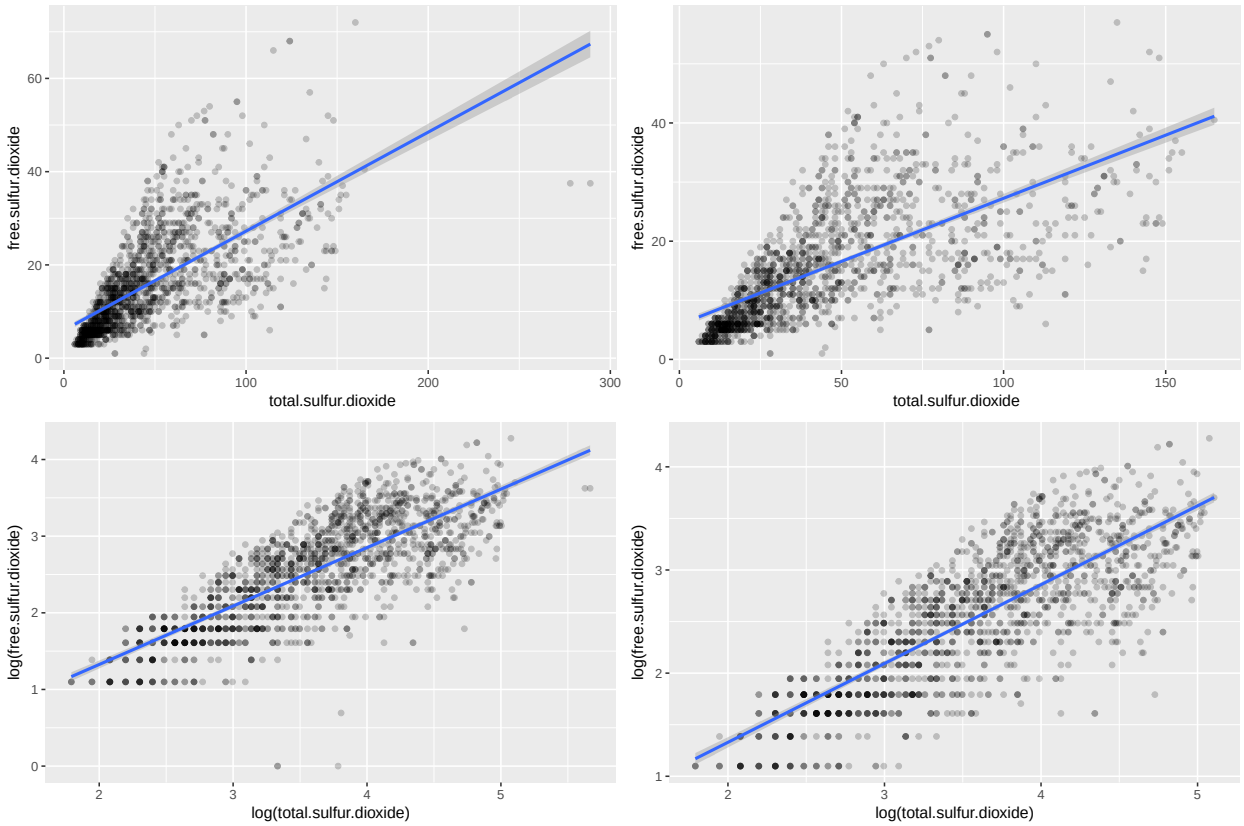
For variables including `volatile.acidity`, `sulphates` and `alcohol`, it is clear that as the quality goes up, their corresponding values tend to go up monotonically. For the rest variables, it is a littl bit hard to read their trends. Additionally statistical tests like ANVOA can be considered later.

- **Question 3:** Let's go back to those four pairs of continuous variables. Can I visualize their correlationship on a 2-D plot? Is there any obvious linear regression trend? Will removing outliers change their linear regression trend? Is logarithm transformation helpful (for those that may follow log-normal distribution)?



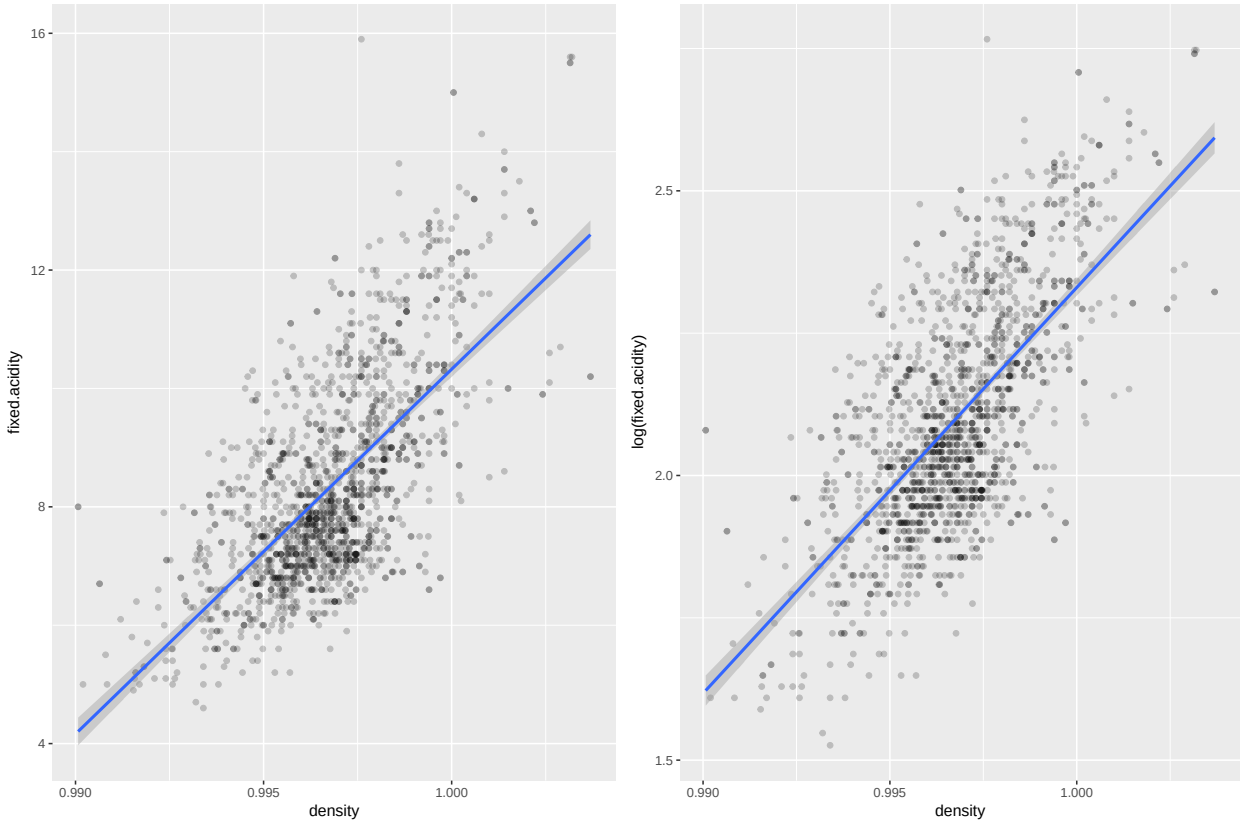
The top-left, top-right, bottom-left and bottom-right are scatter plots with their linear regression line of original variables, original variables without outliers, variables after logarithm transformation and variables after logarithm transformation without outliers.

We can read that the linear trend is obvious. It also seems that removing outliers does not move the position or change the slope of the linear regression line dramatically.

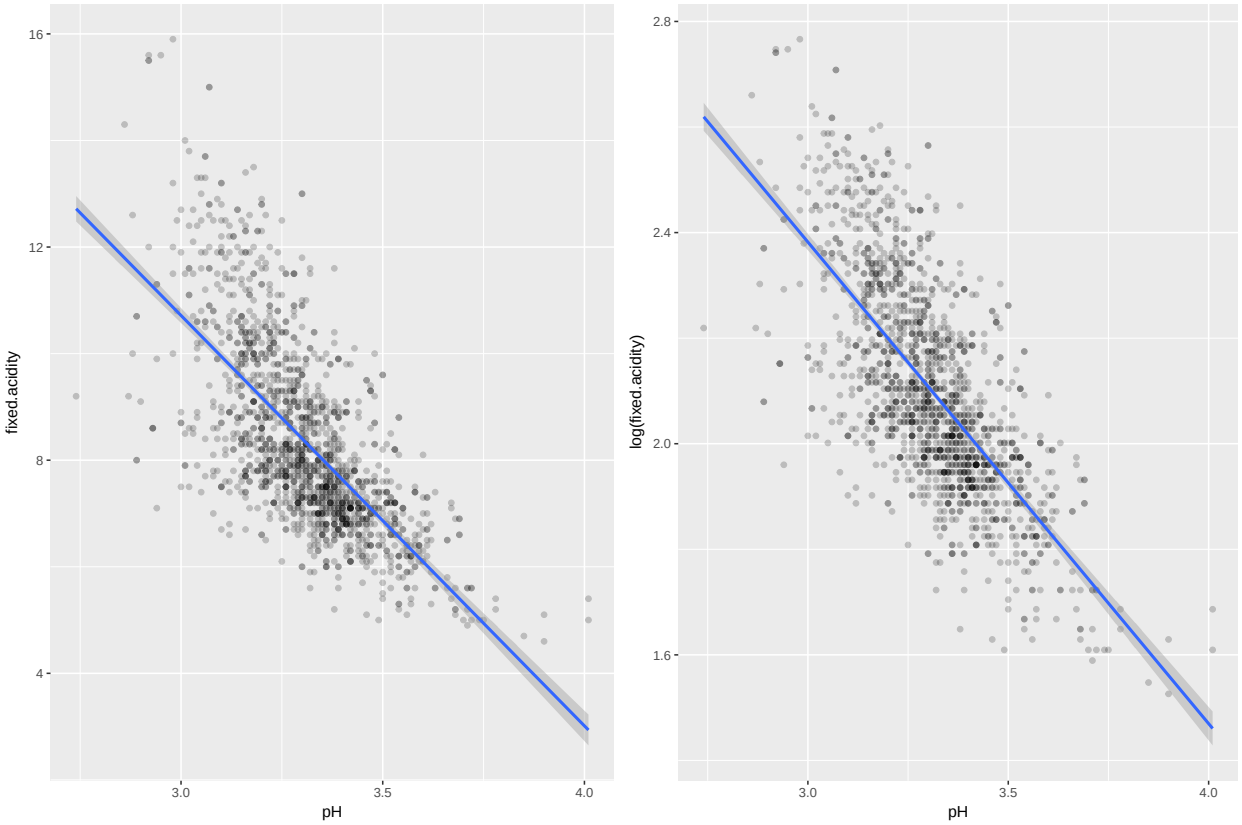


Like the previous one, the top-left, top-right, bottom-left and bottom-right are scatter plots with their linear regression line of original variables, original variables without outliers, both variables after logarithm transformation and variables after logarithm transformation without outliers.

We can read the for the two plots in the first row, the linear trend is not obvious. However, after we perform the logarithm transformation on both y-axis and x-axis, the linear trend becomes much more clear.

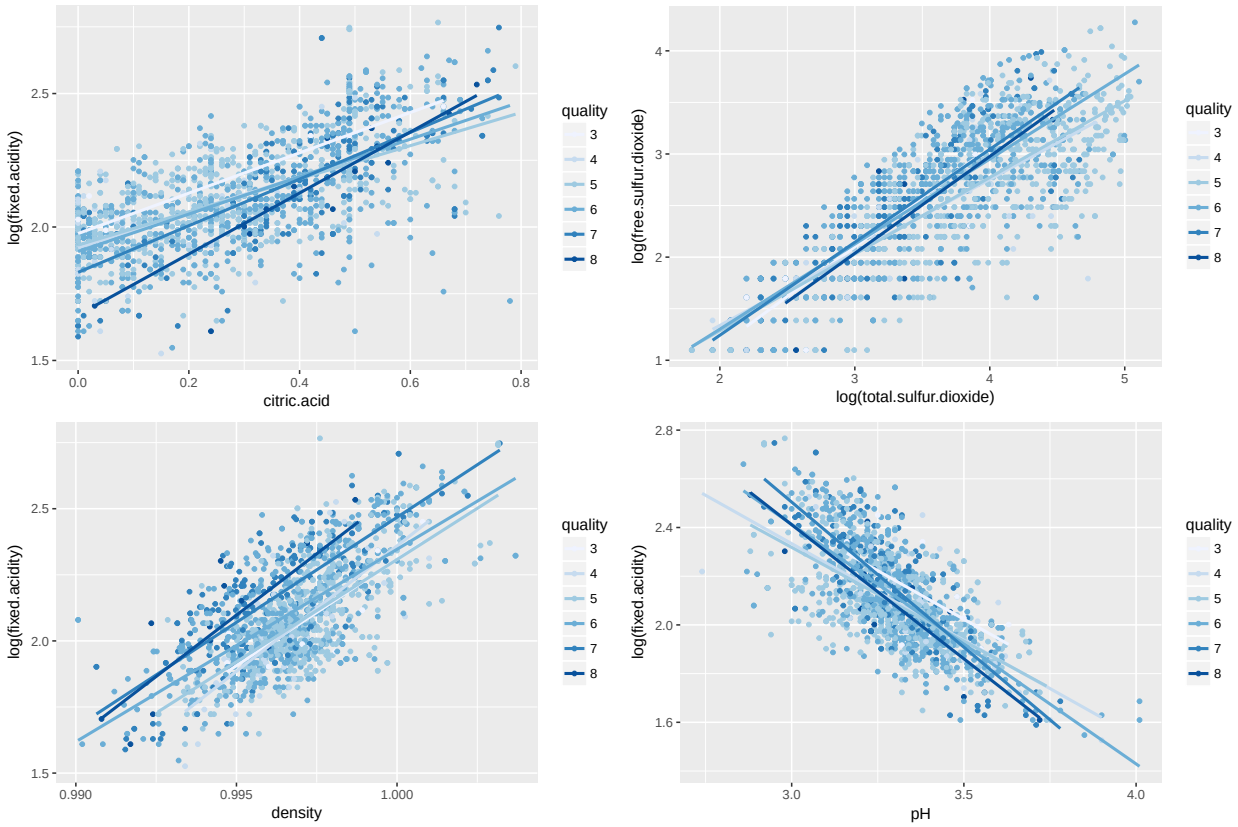


The left and right are scatter plots with their linear regression line of original variables and variables after logarithm transformation. No outlier is removed here. We can still detect the linear trend, but however, it seems that logarithm transformation on variable `fixed.acidity` does not make the trend more obvious.



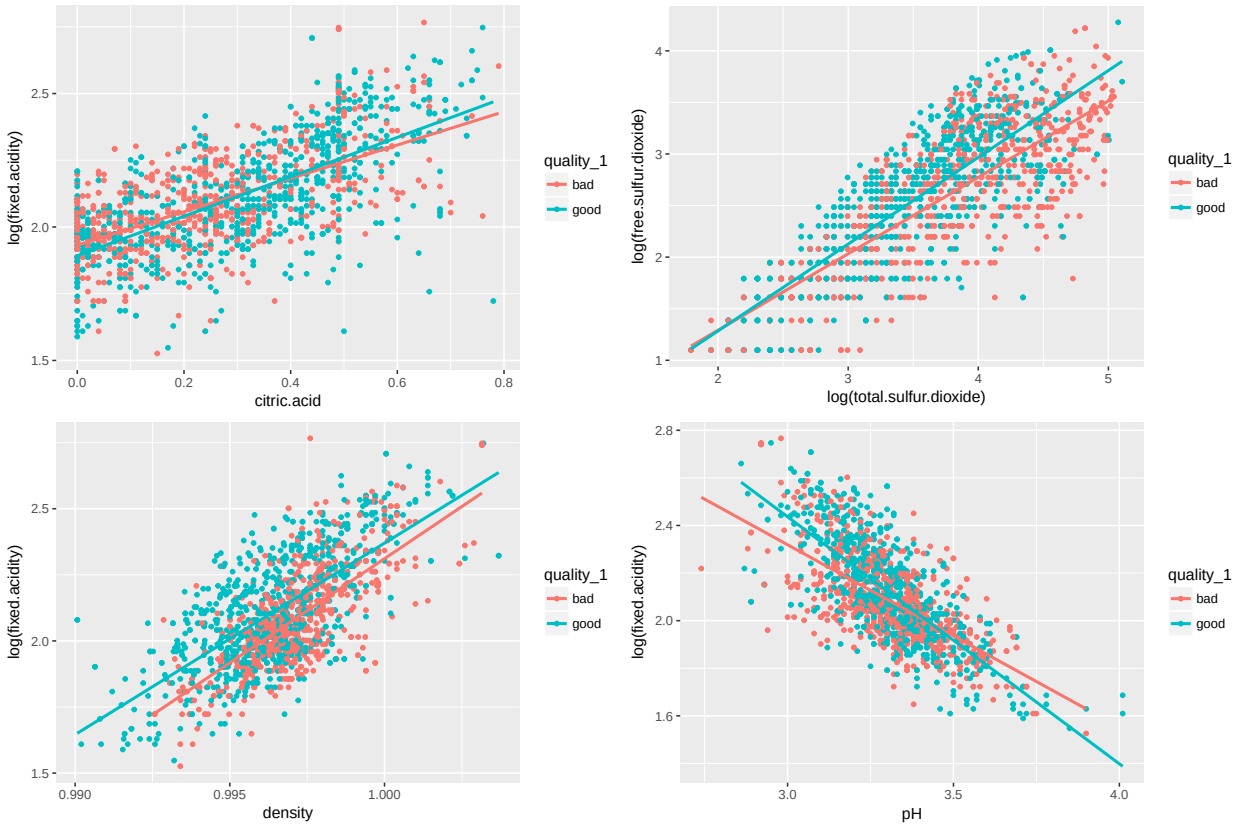
The left and right are scatter plots with their linear regression line of original variables and variables after logarithm transformation like the previous one. No outlier is removed here, either. We can still detect the linear trend, and it seems that logarithm transformation on variable `fixed.acidity` does slightly make this trend more obvious.

- **Question 4:** I'm interested in adding the variable `quality` into these scatter plots. I would like to see if wine with different qualities may distribute in particular areas in the scatter plots.



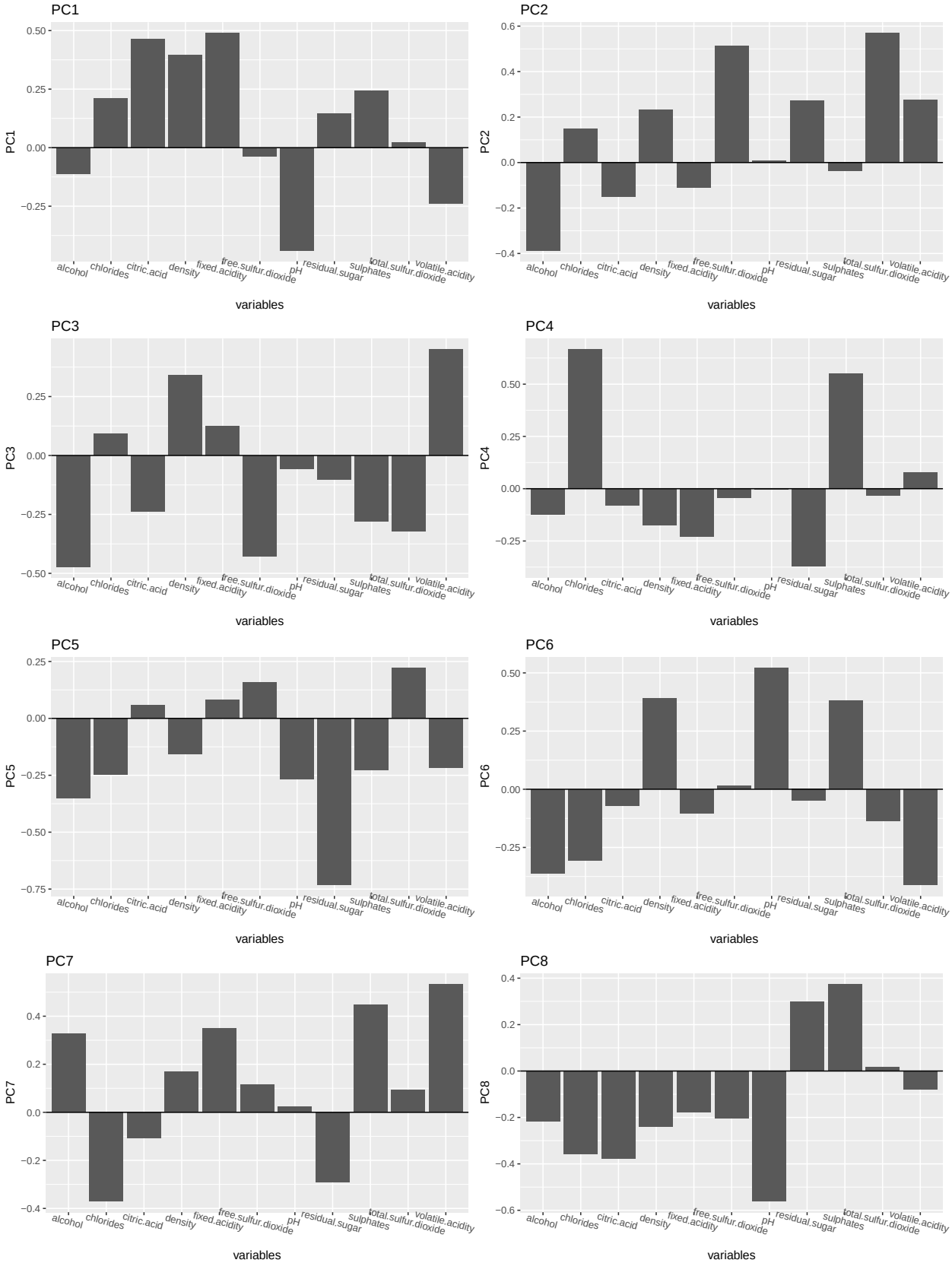
We would like to color our plots based on their corresponding quality, to see if wine with different qualities may tend to distribute in paricular areas of our scatter plots (the four variable pairs above). Here, we only focus the scatter plot with logarithm transformation and without outliers if we performed and removed, respectively.

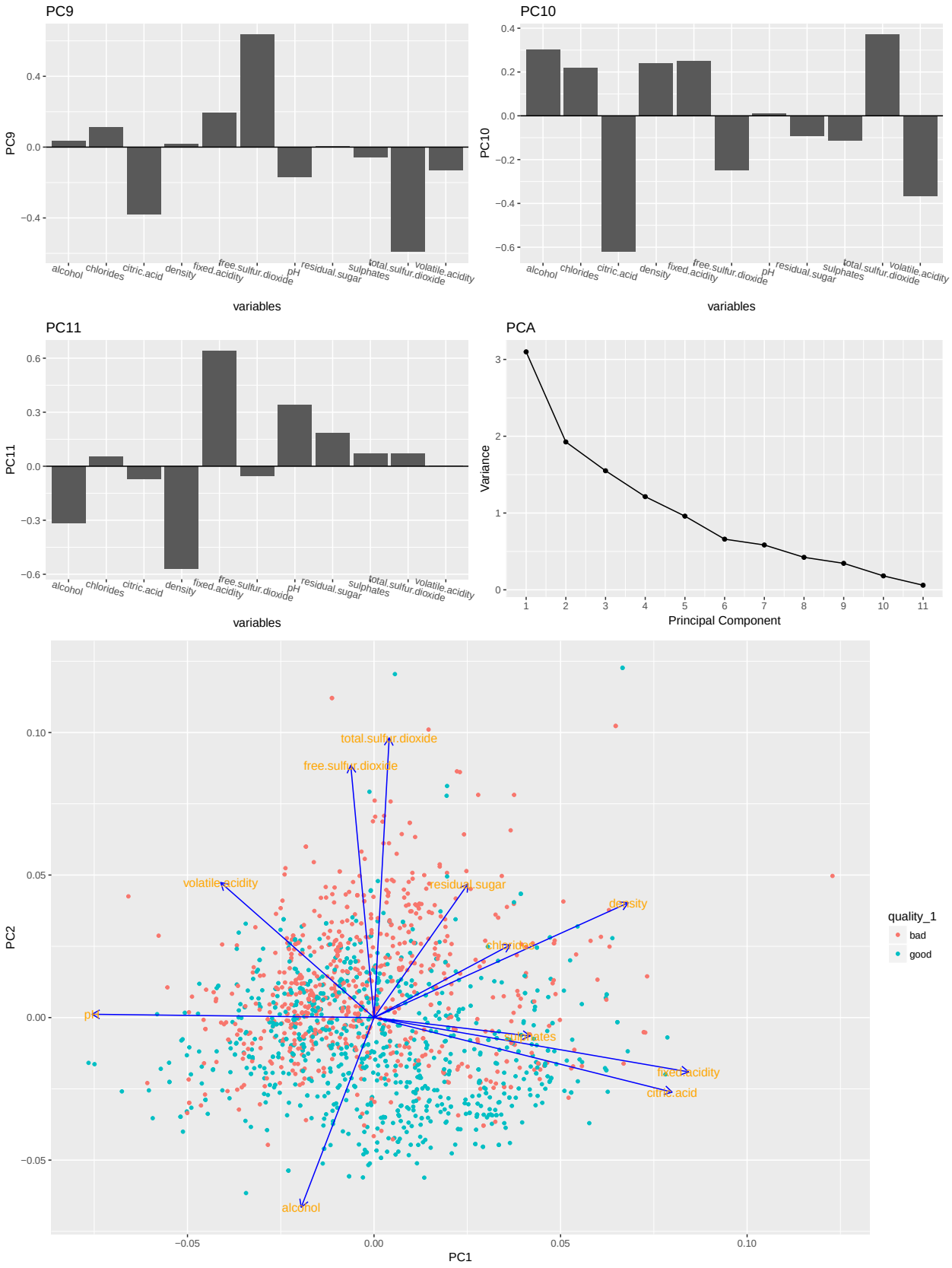
The darker coler indicates higher wine quality while the lighter color indicates the lower wine quality. We can read from these four plots that basically it is impossible to extract very useful information from them. Hence, we would like combine several qualities scores together. If the score is lower than or equal to 5, we set it to be "bad"; otherwise, it is "good".



I am interested in the scatter plot of `log(fixed.acidity)` vs `density`. It seems that there exists a border between good and bad wine. For the rest three plots, it is still hard to see if wine with different qualites tend to distribute in particular areas.

- **Question 5:** Would PCA be helpful to find the 'jointed' distributions of other possible variables and `quality`? I would like to re-find the solutions of **Question 4** by using the scatter plot of PC1 vs PC2 this time.


PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
variables
alcohol
chlorides
citric.acid
density
fixed.acidity
free.sulfur.dioxide
pH
residual.sugar
sulphates
total.sulfur.dioxide
volatile.acidity

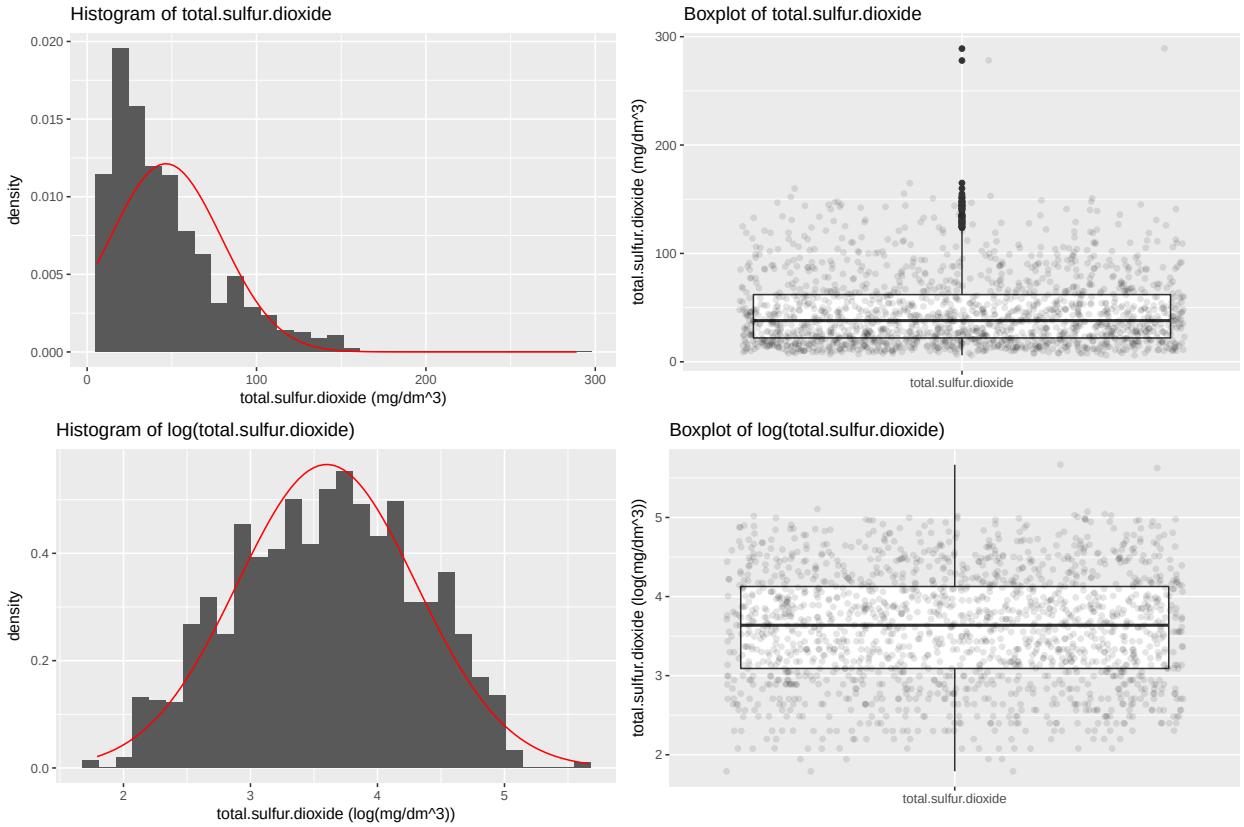


Here, we find out that the first two principal components only explain part of the total variance, which is not

what usually want. This may be the reason why our scatter plot does not offer an output that I want even though we can still read the trend that good quality wine tends to have negative PC2.

2. Final Plots and Summary

2.1 Histogram of total.sulfur.dioxide



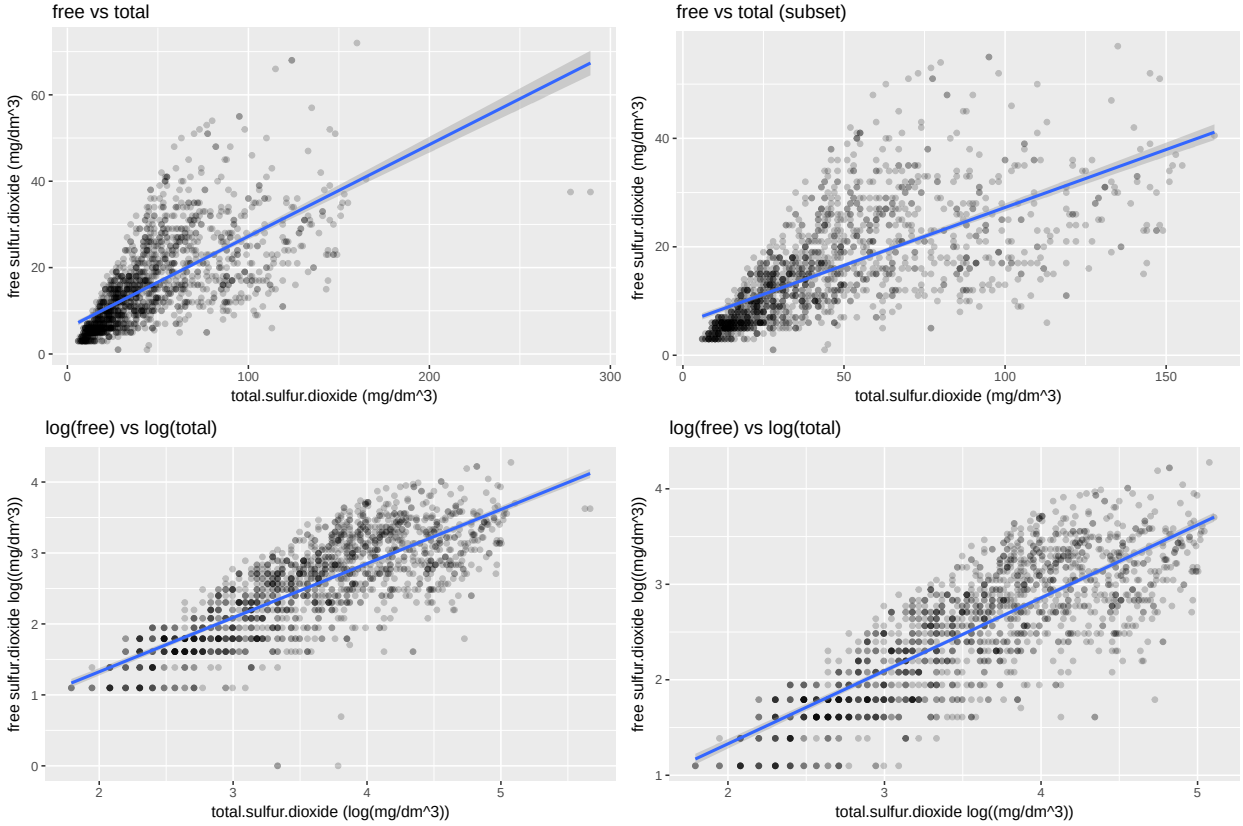
The plot on the left is the histogram of our variable `total.sulfur.dioxide` while the plot on the right is the histogram of the same variable `total.sulfur.dioxide` after logarithm transformation. The red lines in both histograms represent the normal distribution curves with mean and standard deviation being equal to sample mean and standard deviation, respectively.

It is obvious that the variable `total.sulfur.dioxide` does not follow normal distribution based on the left histogram. Because its positively skewed distribution does not fit with the normal distribution curve.

Furtherly, we would like to test if it follows log-normal distribution. The plot on the right shows us that the histogram of `log(total.sulfur.dioxide)` with its normal distribution curve. This one looks much more "normal" than the left one since there is no obvious positive or negative skewness and its distribution basically fits its corresponding normal distribution curve.

Hence, we can conclude that the variable `total.sulfur.dioxide` basically follows the log-normal distribution.

2.2 Scatter plot of free.sulfur.dioxide vs total.sulfur.dioxide



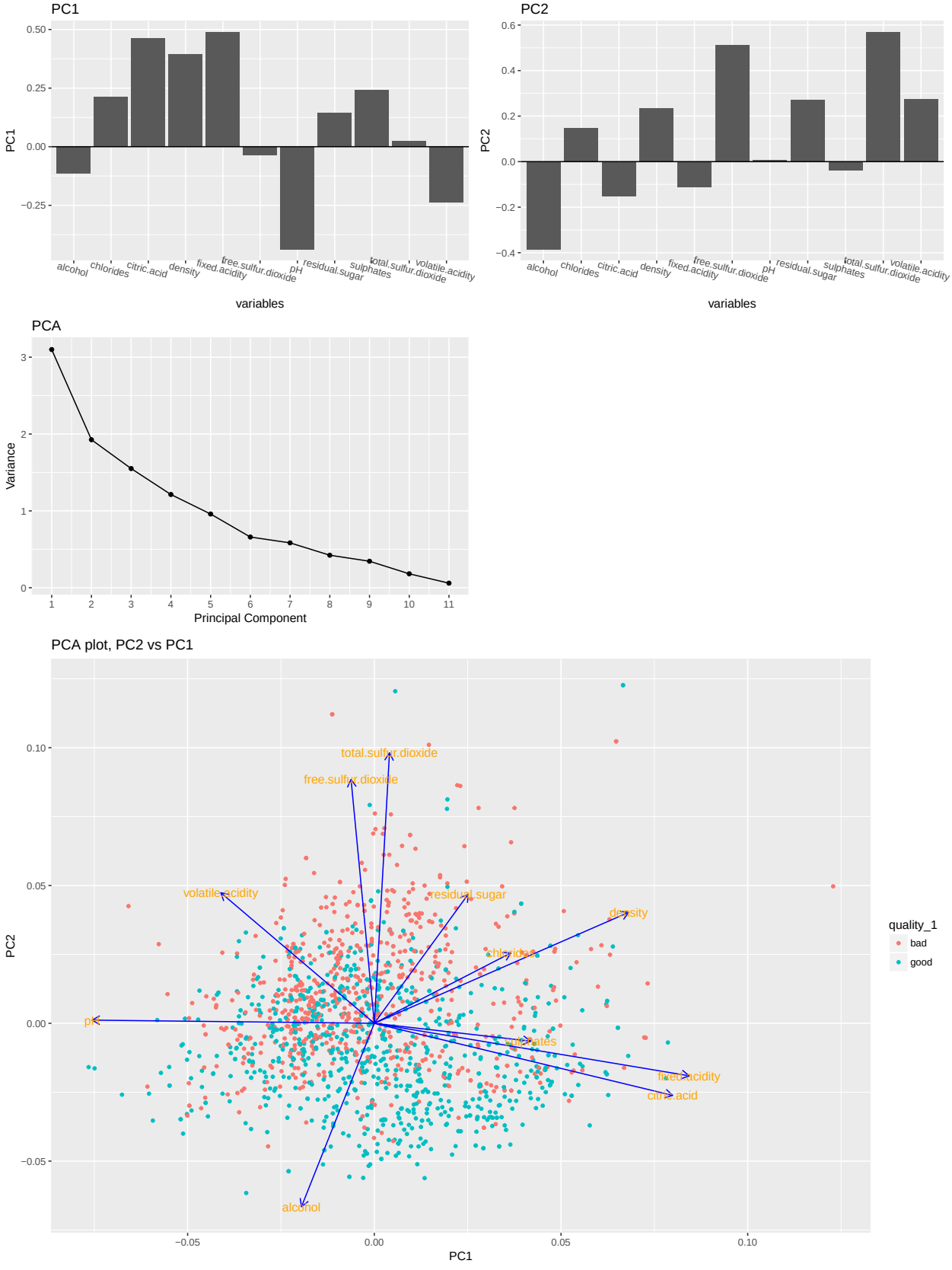
The four plots here are scatter plots related with variable `free.sulfur.dioxide` vs `total.sulfur.dioxide`. The word `free` and `total` represent `free.sulfur.dioxide` and `total.sulfur.dioxide`, respectively. `log()` means that we perform logarithm transformation on the corresponding variable. The blue line in the plot shows the linear regression line.

The top-left plot shows that there may exist potential linear trend between these two variables. The top-right plot shows the scatter plot after we remove those points that are far away from others (AKA outliers) and its corresponding linear regression line. We can see that these two regression lines were moved dramatically.

The bottom-left plot shows that the linear trend is much more obvious after we perform logarithm transformation on both variables. The bottom-right plot is the scatter plot after we remove outliers. The outliers here may not be exactly the same as the top-right plot. We can notice that these two regression lines were basically at the same position.

Hence, we may think that `total.sulfur.dioxide` and `free.sulfur.dioxide` are highly correlated with each other. Especially after we perform logarithm transformation on both variables, there exists an obvious linear trend. Additionally, the pearson correlation coefficient of these two variables and these two after logarithm transformation are 0.6676665 and 0.7846217, respectively, which also prove our opinion.

2.3 Principal Component Analysis plot



Here, we extract the first two Principal Components. We can notice that our 1st PC mainly depends on variable `citric.acid`, `density`, `fixed.acidity` and `pH`. And our 2nd PC mainly depends on variable `alcohol`, `free.sulfur.dioxide` and `total.sulfur.dioxide`. These information can also be read from the arrows of our scatter plot. Because the direction of these arrows are nearly parallel with x-axis and y-axis, respectively.

Here, we drew our scatter plot of PC2 vs PC1. It is still not very easy to clearly define the 'territory' of different qualities, Especially, when PC2 is positive. However, when we have a negative PC2, it tends to be much clearer that most points of this area are pre-defined as a good quality wine.

Based on the plot of PCA we can see that, the first two Principal Components do not explain most of the variances, and additional principal components should be added in case we may face the issue of under-fitting. This may also be the reason that the scatter plot of PC2 vs PC1 may not give us the exact result that we want.

3. Reflection

3.1 My struggle

When I tried to find if we can cluster or stratify the scatter plots obtained from part 2.2 by using variable `quality`, I found its impossible to do that. Because I cannot find any obvious trend which may lead me to the result that I would like to have.

3.2 My success

Most plots have good performances. For example, I successfully found out the distributions of different variables and compared them with the corresponding normal distribution curve. Since some of the distributions look like a log-normal distribution, I also performed logarithm transformation and successfully found several variables which may follow log-normal distribution, such as `total.sulfur.dioxide`.

3.3 Future work

In the future, I would like to mainly focus on the struggle that I had.

- Very few wine products have the score lower than 5 or greater than 7. I may try use the subset of wine products with quality score being 5, 6 and 7 only and reperform PCA.
- After we finished detecting outliers, we may think removing those points which may influence our model seriously. Without outliers, we may be able to get the results that we want.
- We already detected that some variables follow log-normal distribution. We can firstly tranform those variables, then reperform PCA based on our new dataset.

Reference

N/A