Red Wine Quality by Rama

Which chemical properties influence the quality of red wines?

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

The original dataset has 1599 instances with 13 variables. A variable(quality_cat) is added to the dataset to categorize the quality of the wine into Bad,Average and Good corresponding to the values - upto 4, 5 to 7 and 8 and above respectively.Another variable(wine_quality) is added to store the factor of quality variable.

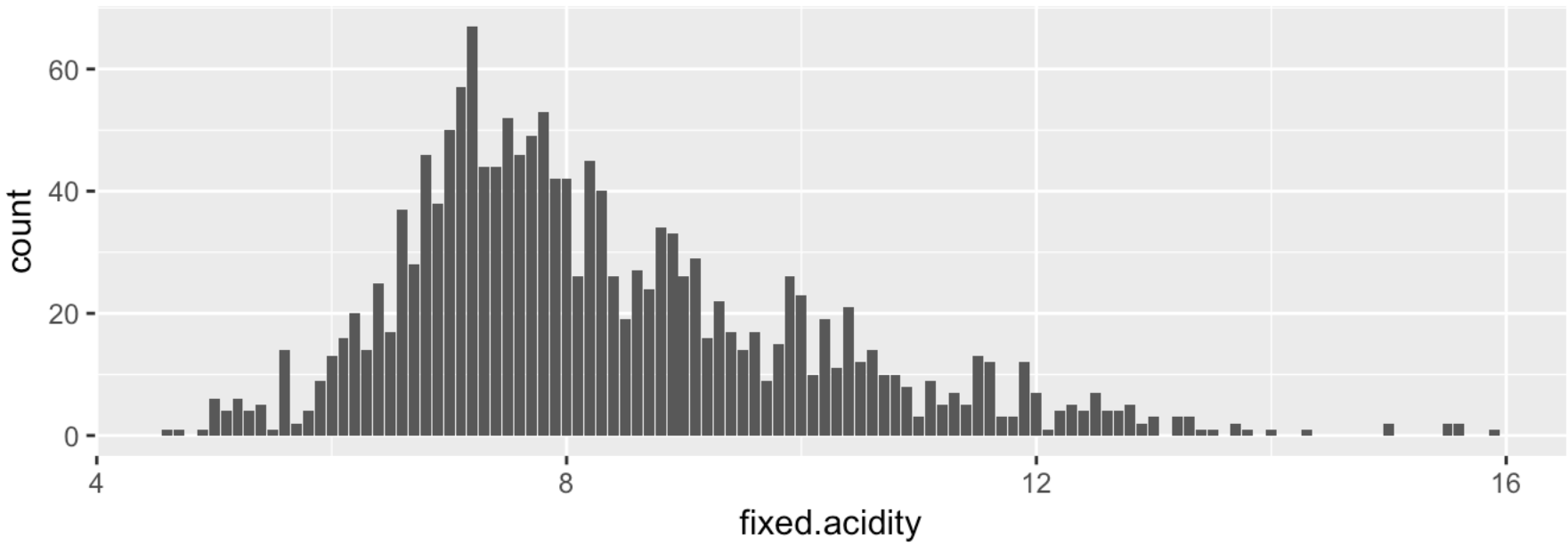
Univariate Plots Section

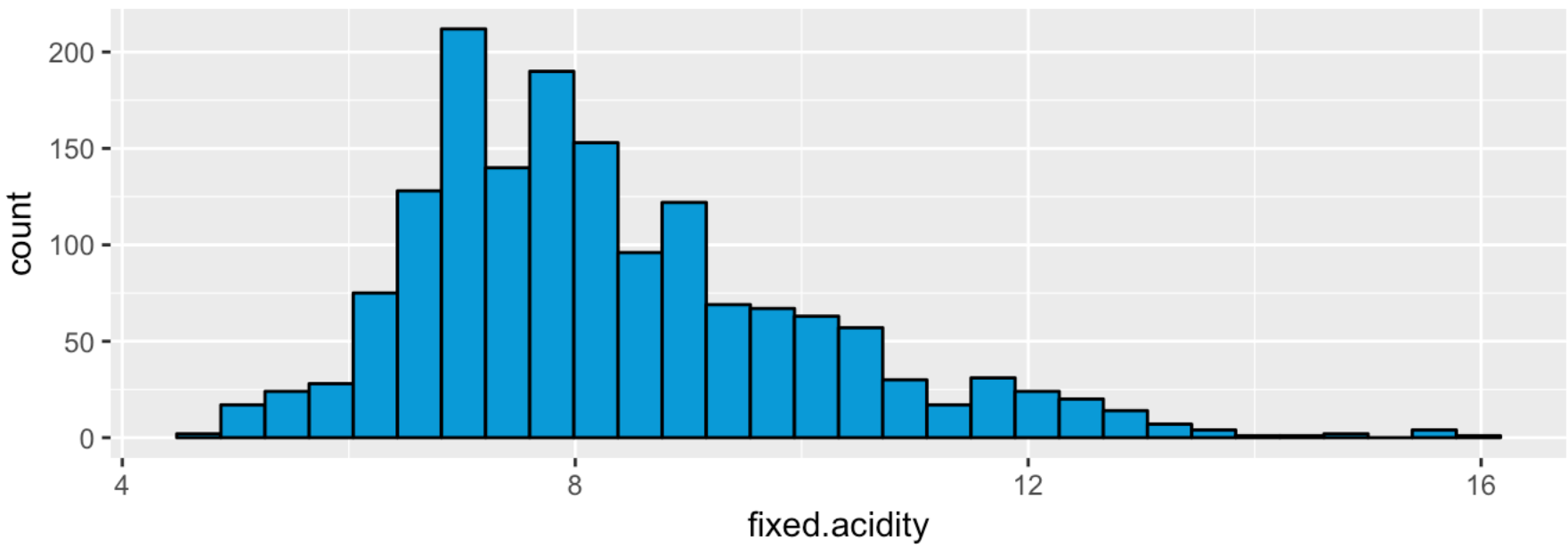
```
## [1] 1599   15
```

```
## 'data.frame':    1599 obs. of  15 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 ...
##  $ quality_cat         : Factor w/ 3 levels "Bad","Average",..: 2 2 2 2 2 2 2 2 2
2 ...
##  $ wine_quality        : Factor w/ 6 levels "3","4","5","6",..: 3 3 3 4 3 3 3 5 5
3 ...
```

```
##        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       quality_cat   wine_quality
##  Min.   : 8.40   Min.   :3.000   Bad    :  63   3:  10
##  1st Qu.: 9.50   1st Qu.:5.000   Average:1518   4:  53
##  Median :10.20   Median :6.000   good   :  18   5:681
##  Mean   :10.42   Mean   :5.636                  6:638
##  3rd Qu.:11.10   3rd Qu.:6.000                  7:199
##  Max.   :14.90   Max.   :8.000                  8:  18
```

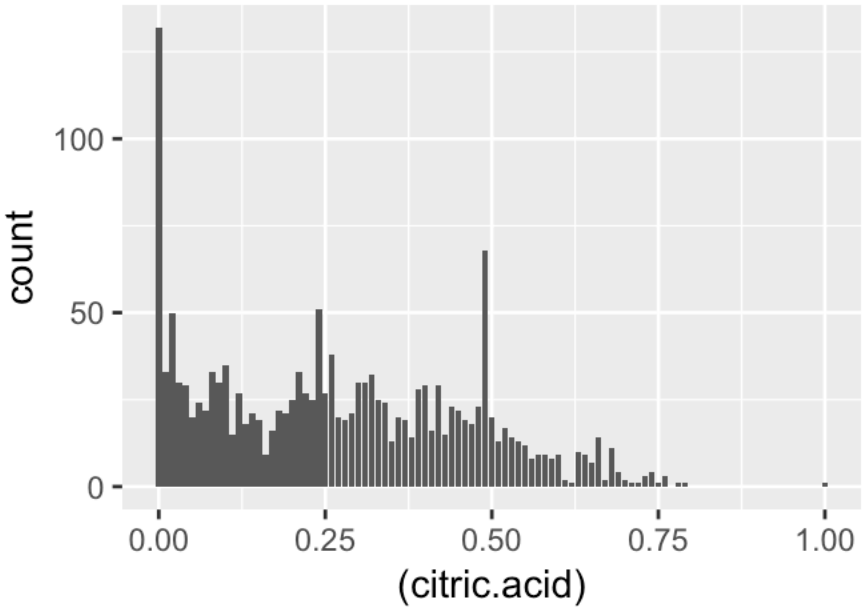
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

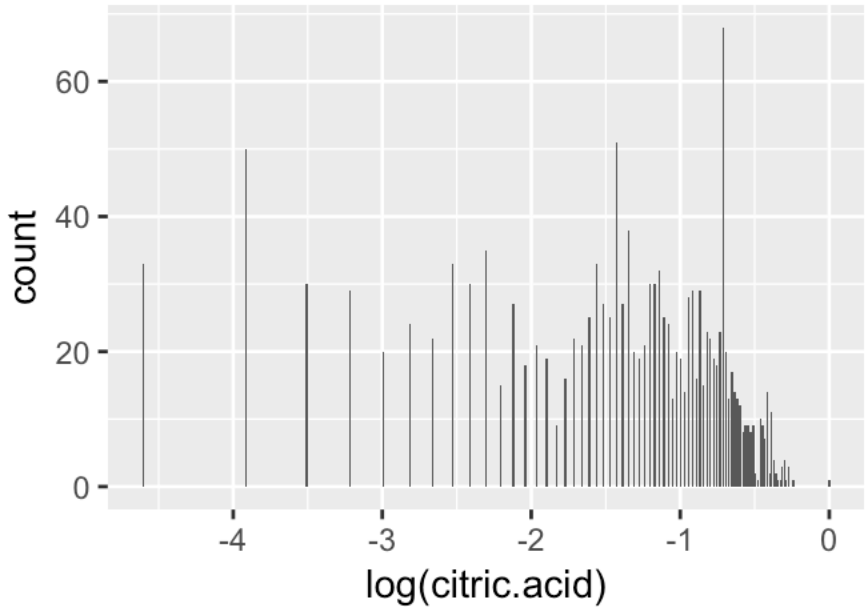


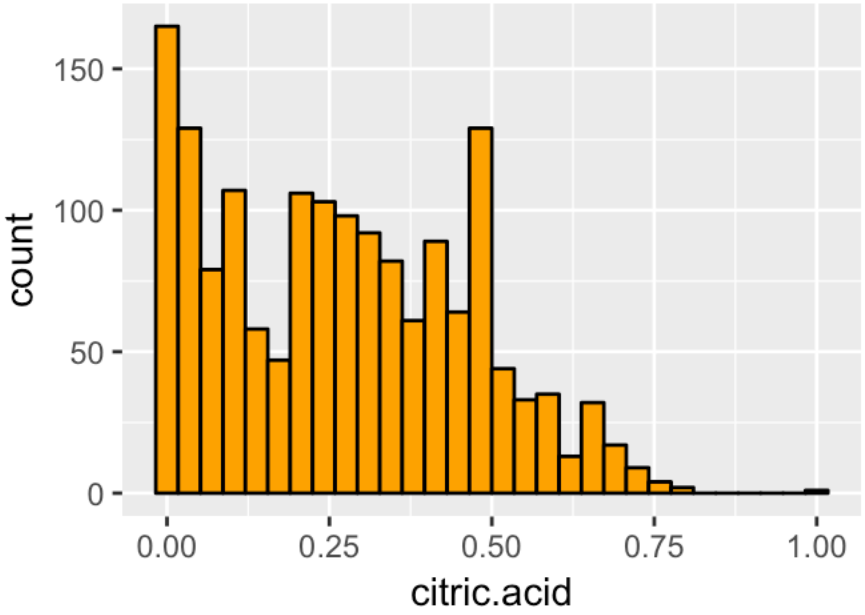


In the plot of fixed.acidity ,there is a peak around 7. when we change the binwidth in histogram, the shape of the distribution changes.

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

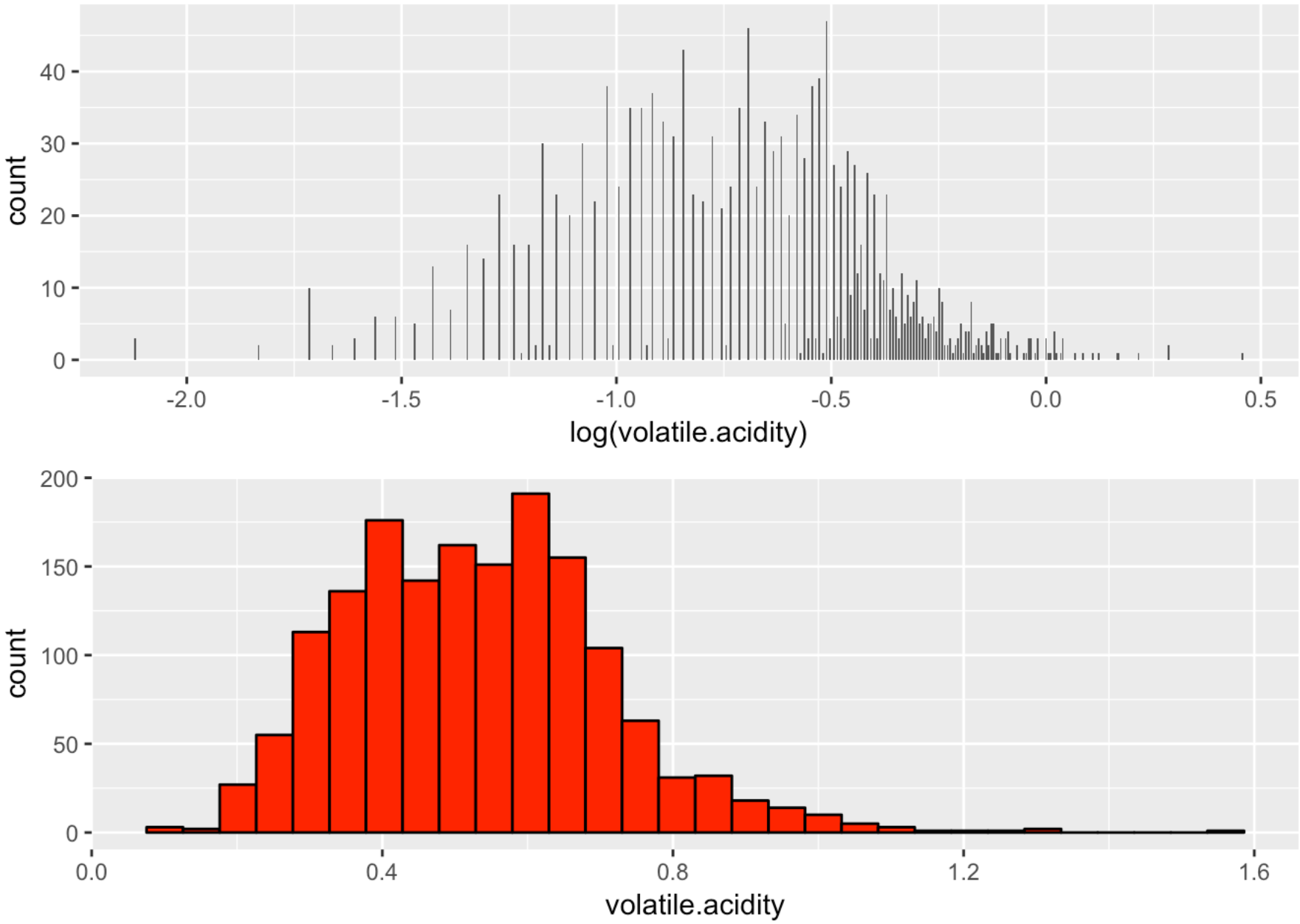






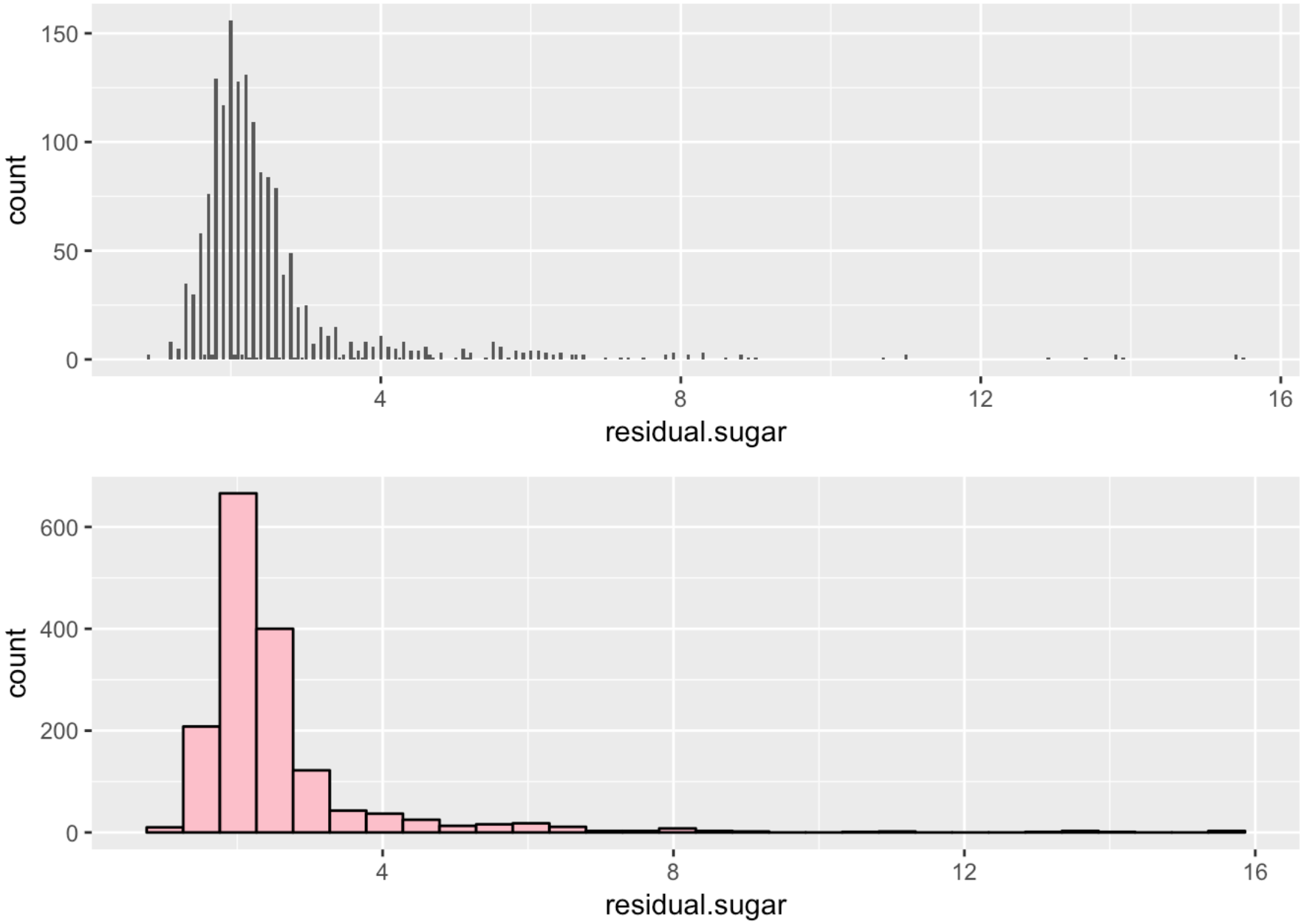
The value of citric acid ranges from 0 to 1. Most of the wines have zero citric acid .The second peak is at 0.48/.49 .

The median is at 0.260 ,the 3rd quartile is at 0.420 and the max is at 1.000 which indicates the presence of the outliers. Did you notice the plot at 1 ?. yes, it’s an outlier.



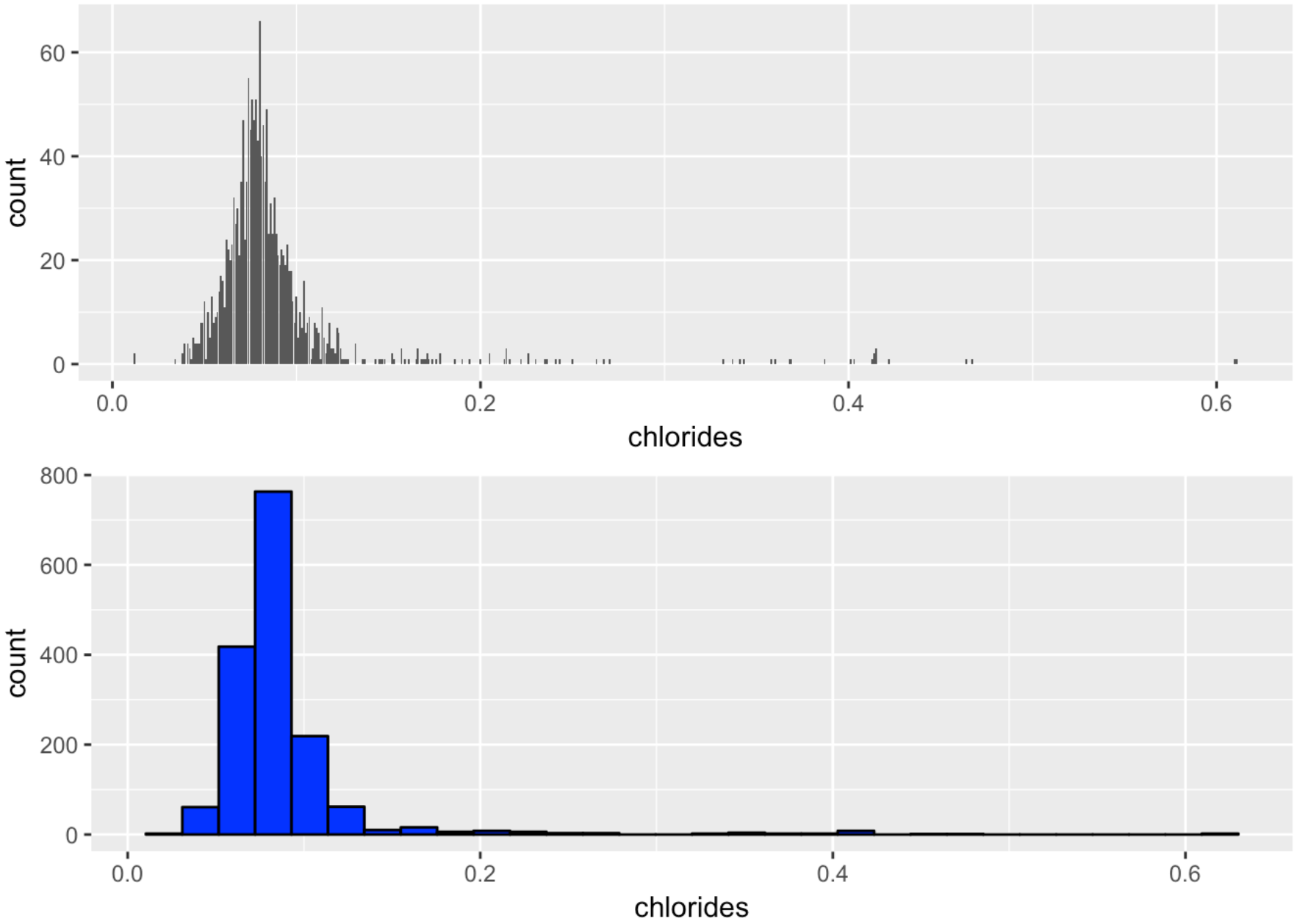

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

The distribution of the volatile.acidity shows there are two peaks at 0.6 and 0.4 in the histogram plot. summary results shows the presence of outliers.



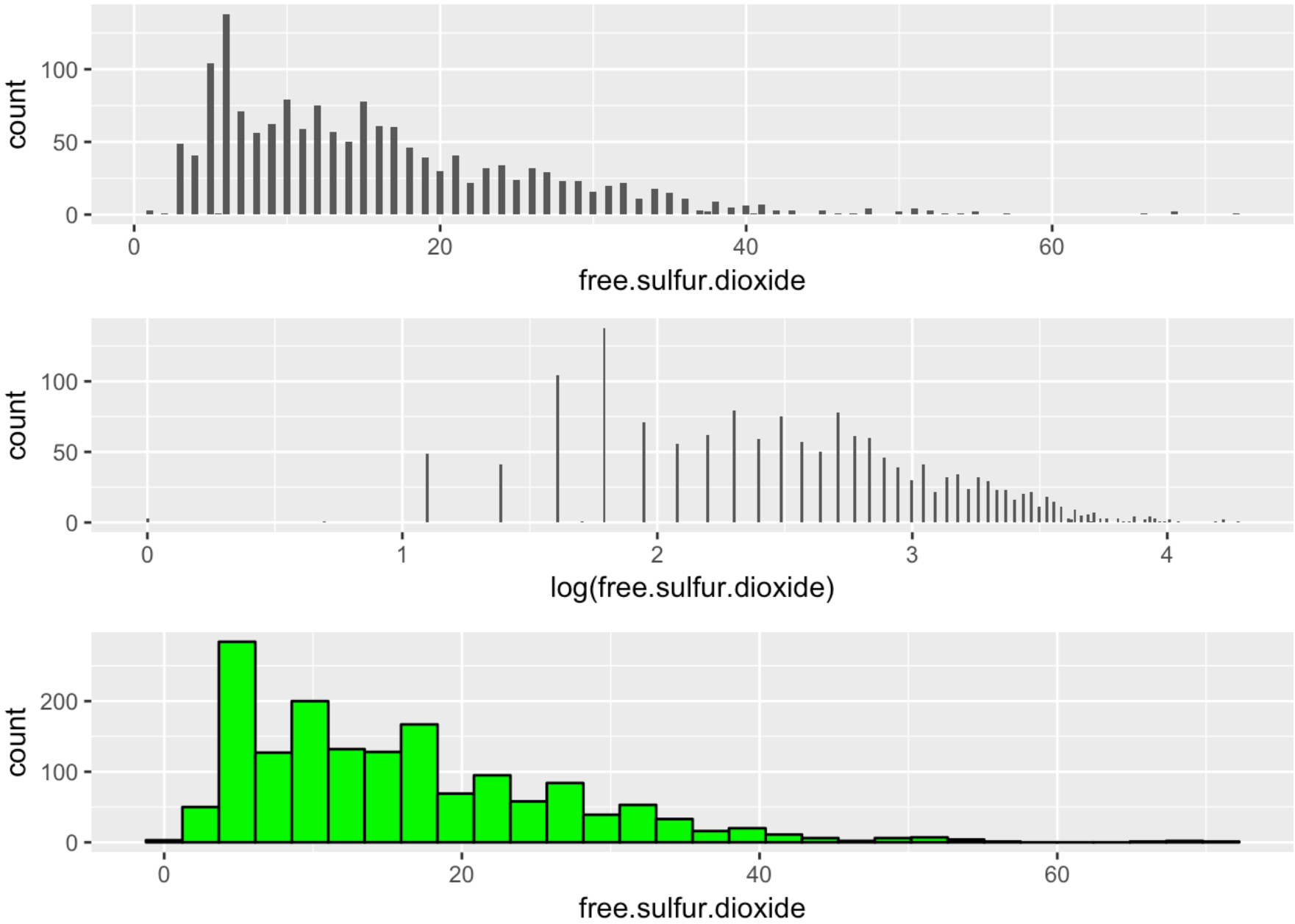

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

The distribution of the residual.sugar has a highest peak at 2. The max is at 15.50 , Median is at 2.200 and 3rd quartile is at 2.600 , these values indicates the presence of outliers.



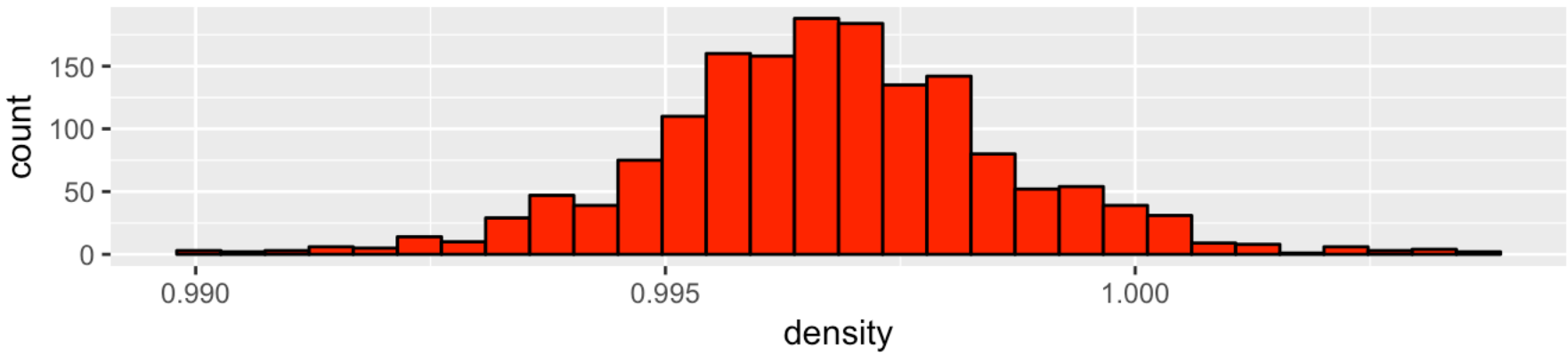

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

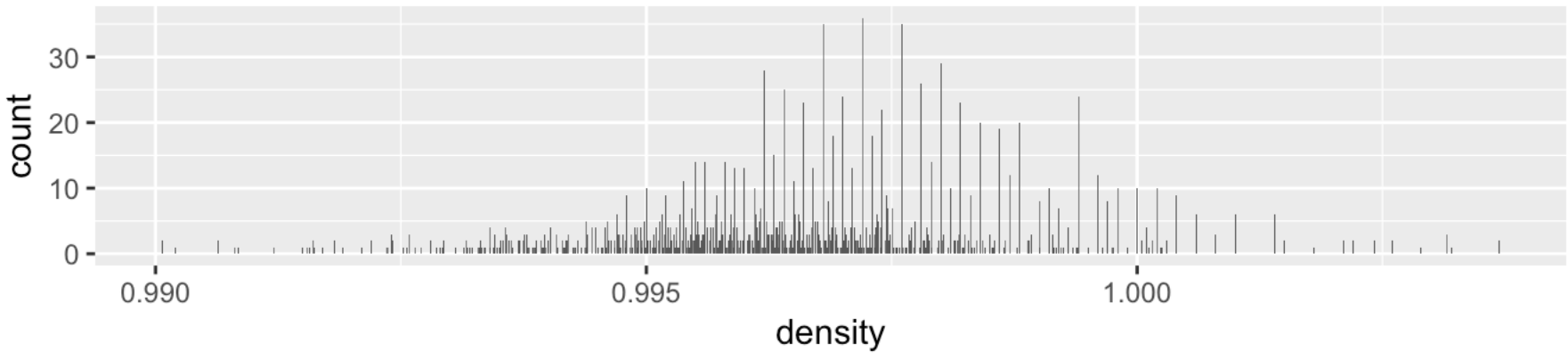
The distribution of chlorides has a peak at 0.08/0.09 . The value of chlorides ranges from 0.01 to 0.6. The result summaries shows the presence of outliers.

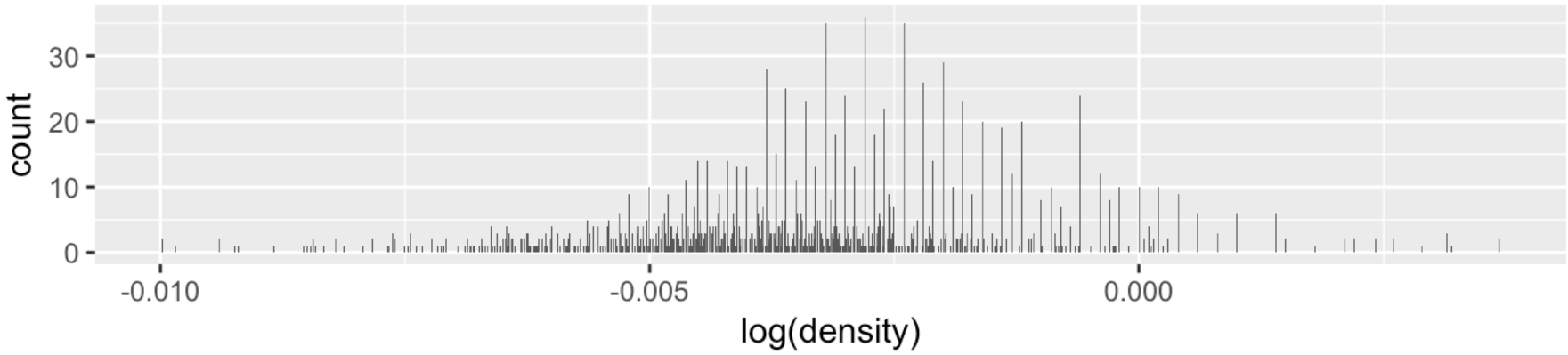



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

The distribution of free.sulfur.dioxide has a peak at 5 in the histogram plot. There are values even beyond 60. The summary result indicates the presence of outliers.

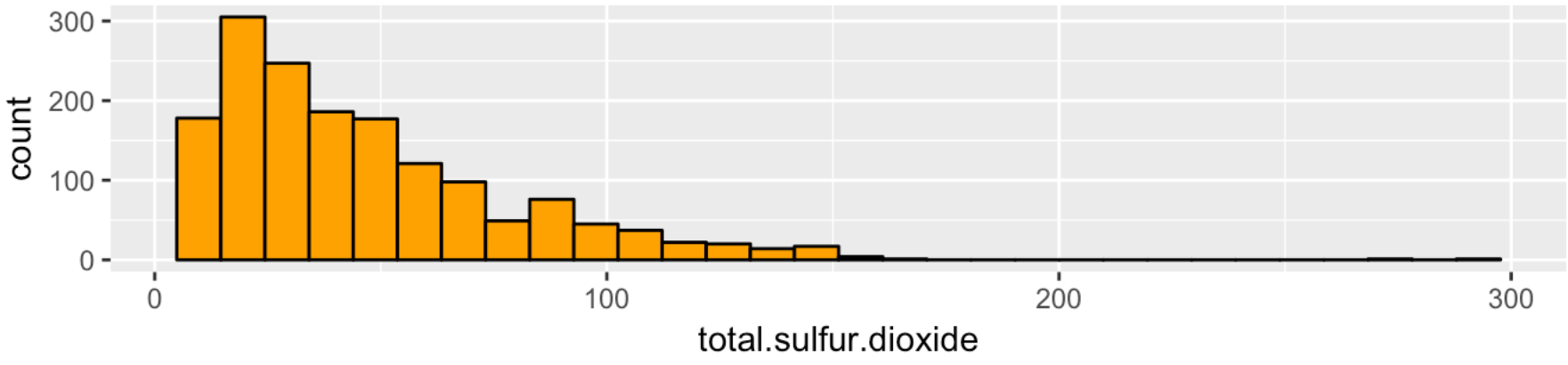


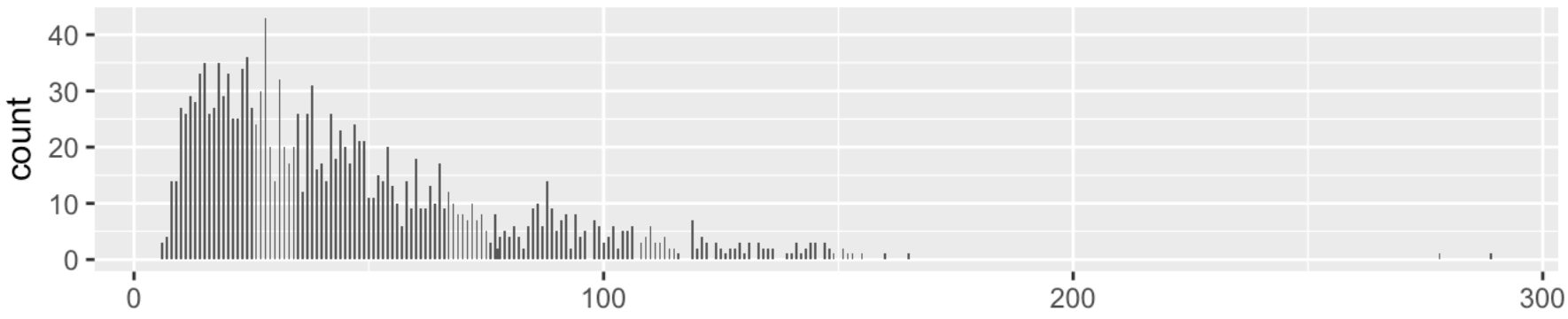


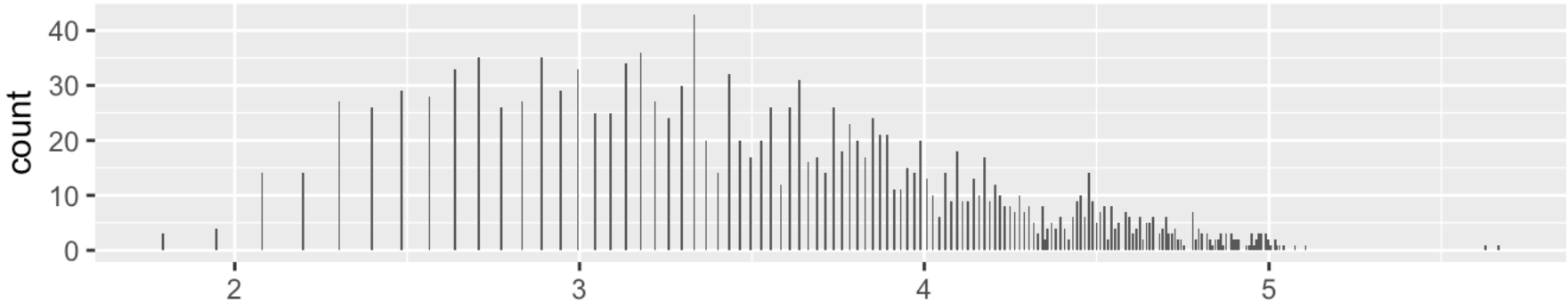



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

The distribution of density in the bar plot has different shades of lines. When there are more datapoints at particular x value ,the line has a darker shade. When there are few datapoints at a particular x value, the line has a lighter shade. Also, each scale in x-axis is very small.



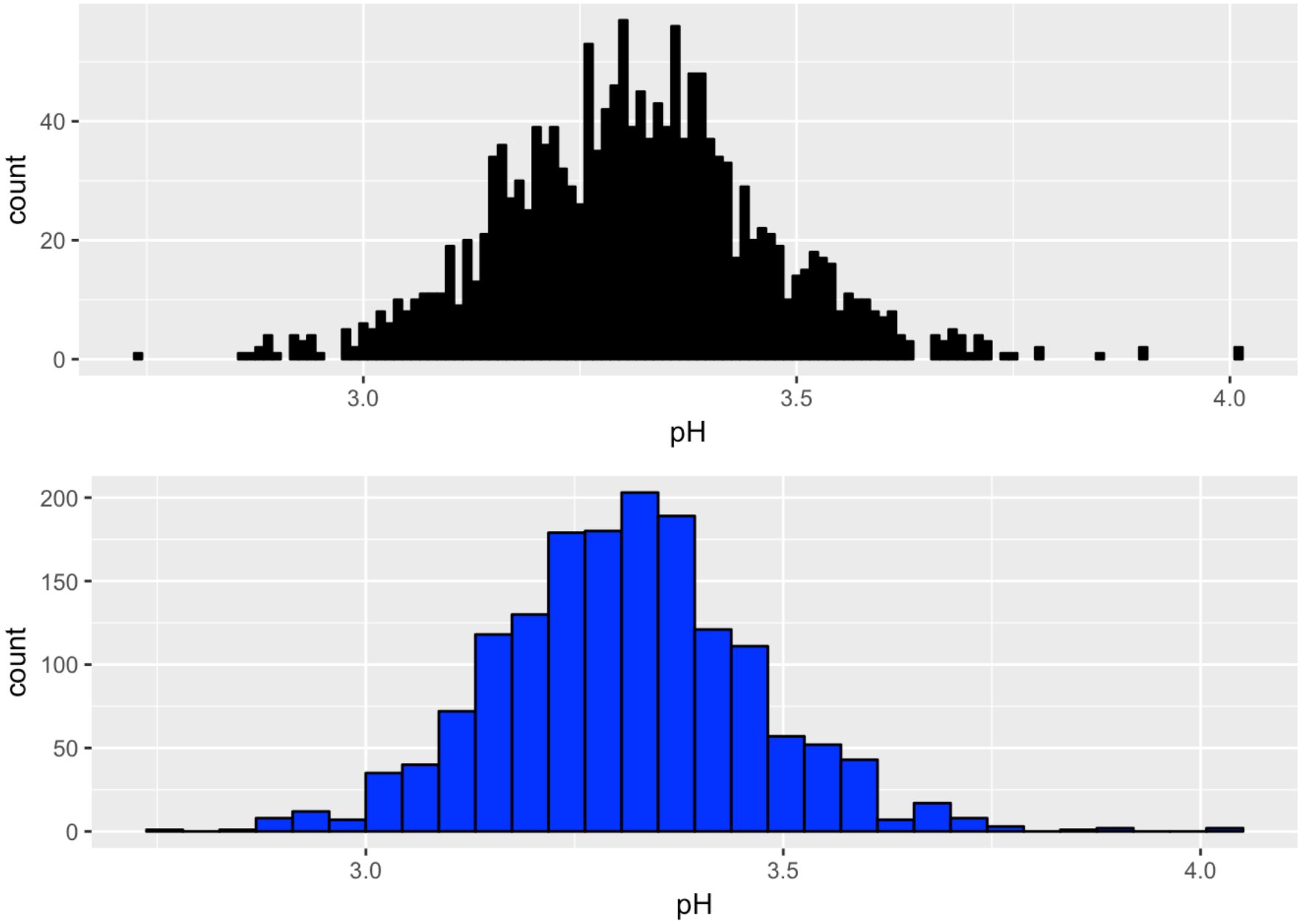





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

The distribution of total.sulfur.dioxide in the histogram plot has a peak value around 20. The median,mean ,3rd quartile and the max in summary result shows the presence of outliers.

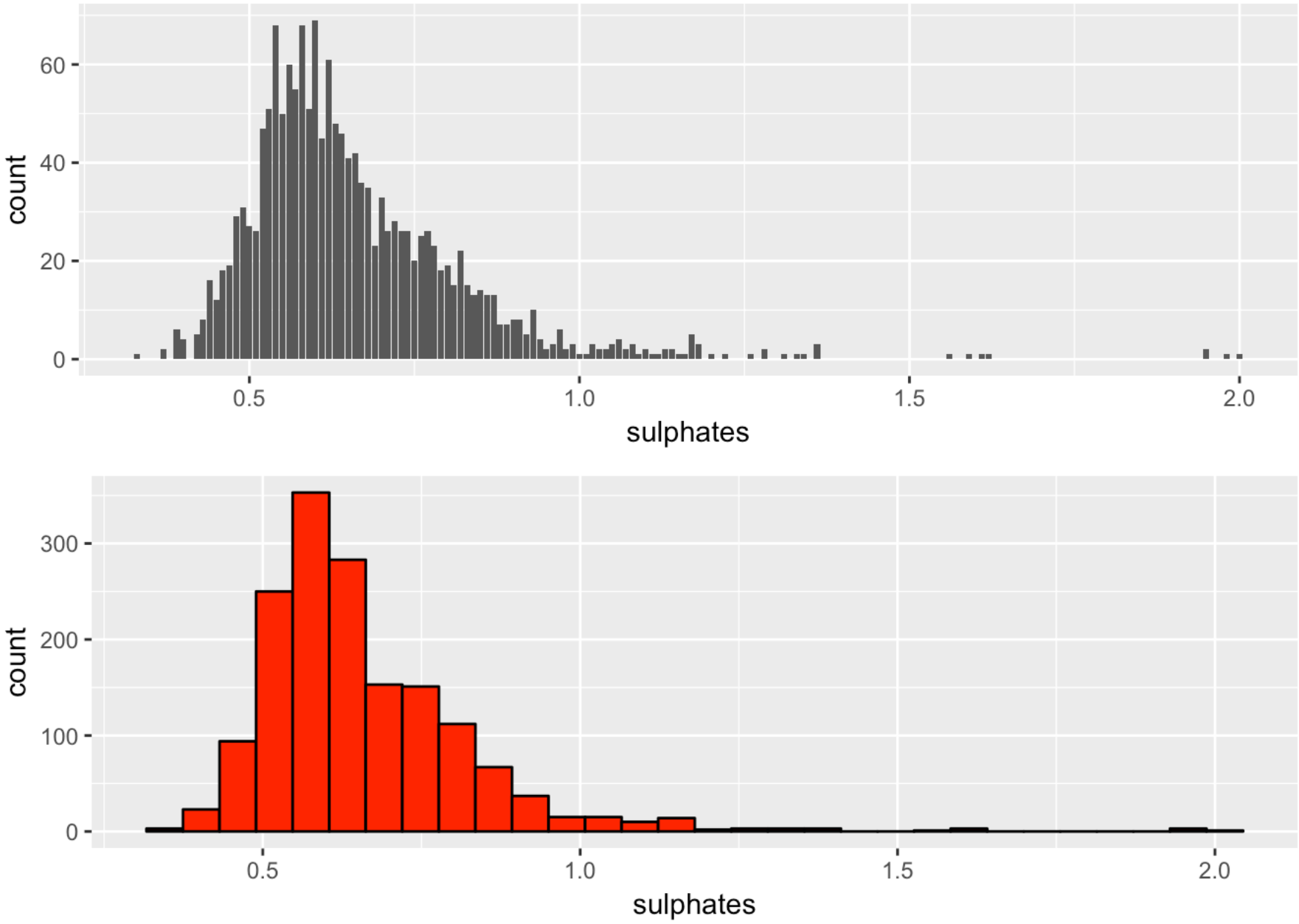
I got curious to know the difference between the total.sulfur.dioxide and free.sulfur.dioxide . free.sulfur.dioxide is the amount of free form of so2 ,whereas total.sulfur.dioxide is the amount of free and bound forms of so2.




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

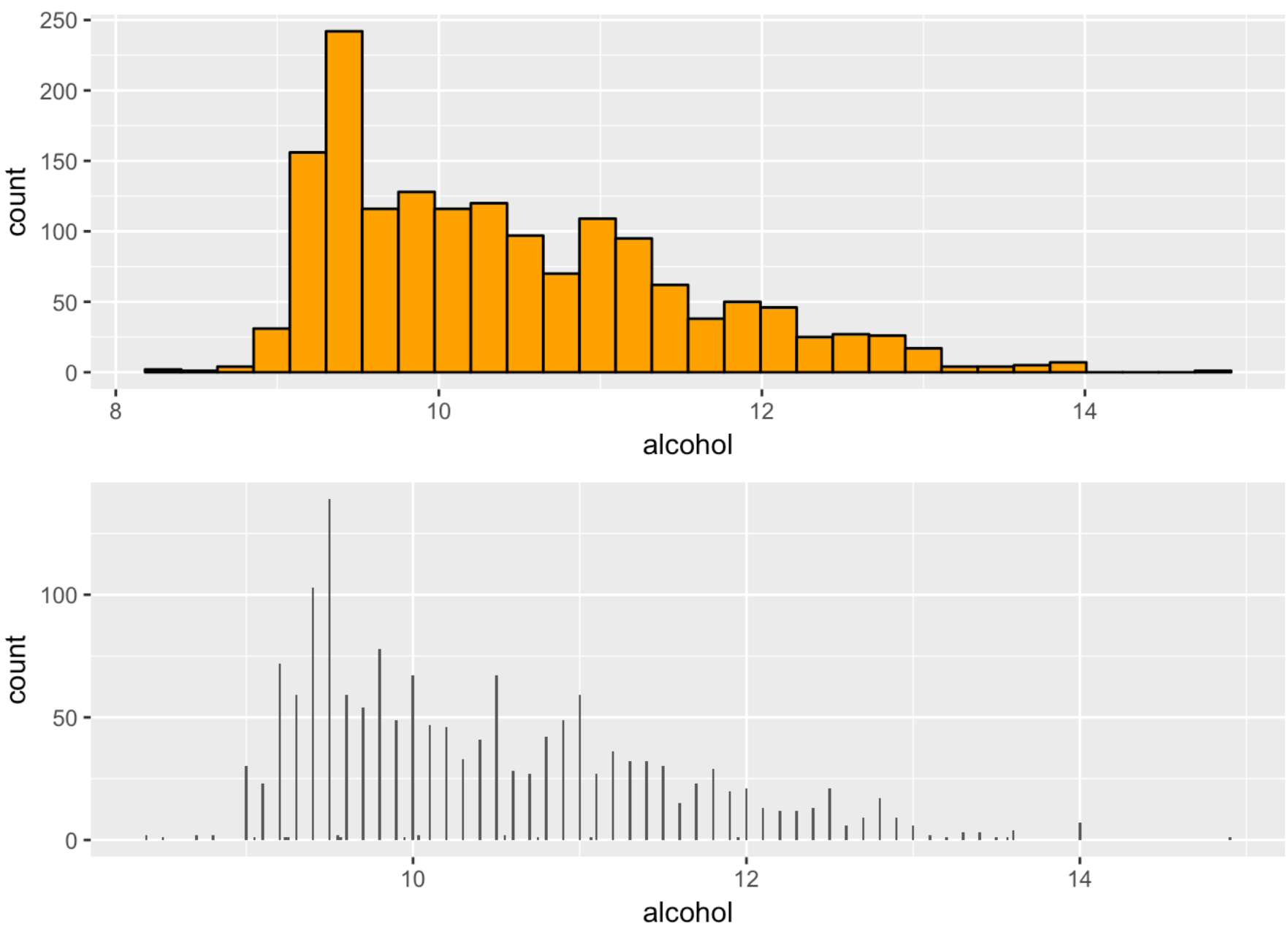
Most of the pH values falls between 3.0 and 3.5. The pH value ranges from 2.74 to 4.01 . The pH describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic).

The summary results shows that the mean and median values are pretty close which means we have a normal distribution .



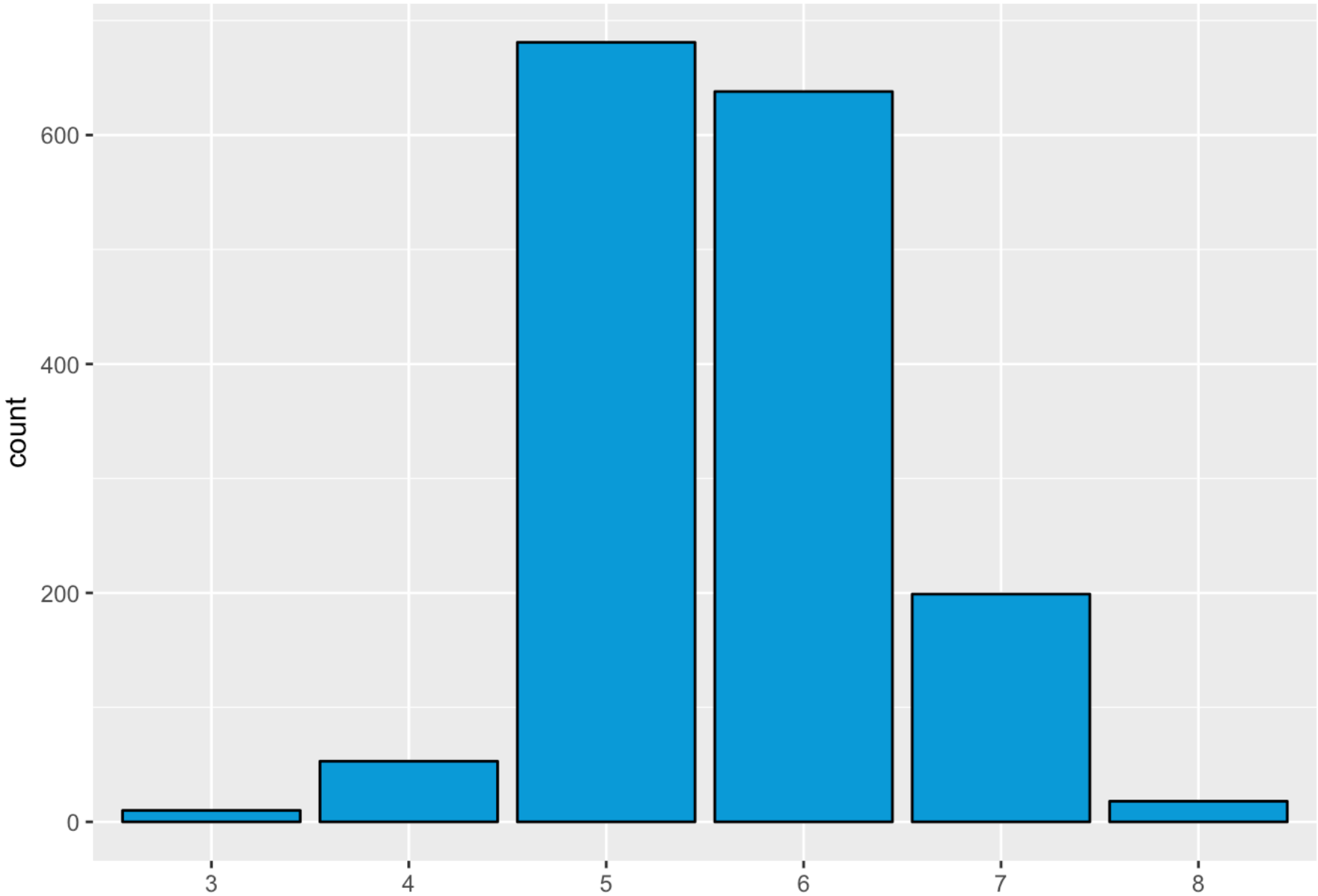

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

The distribution of sulphates has a peak at 0.6 . There are visible plots even beyond 2.0 . The summary results indicates the presence of outliers.



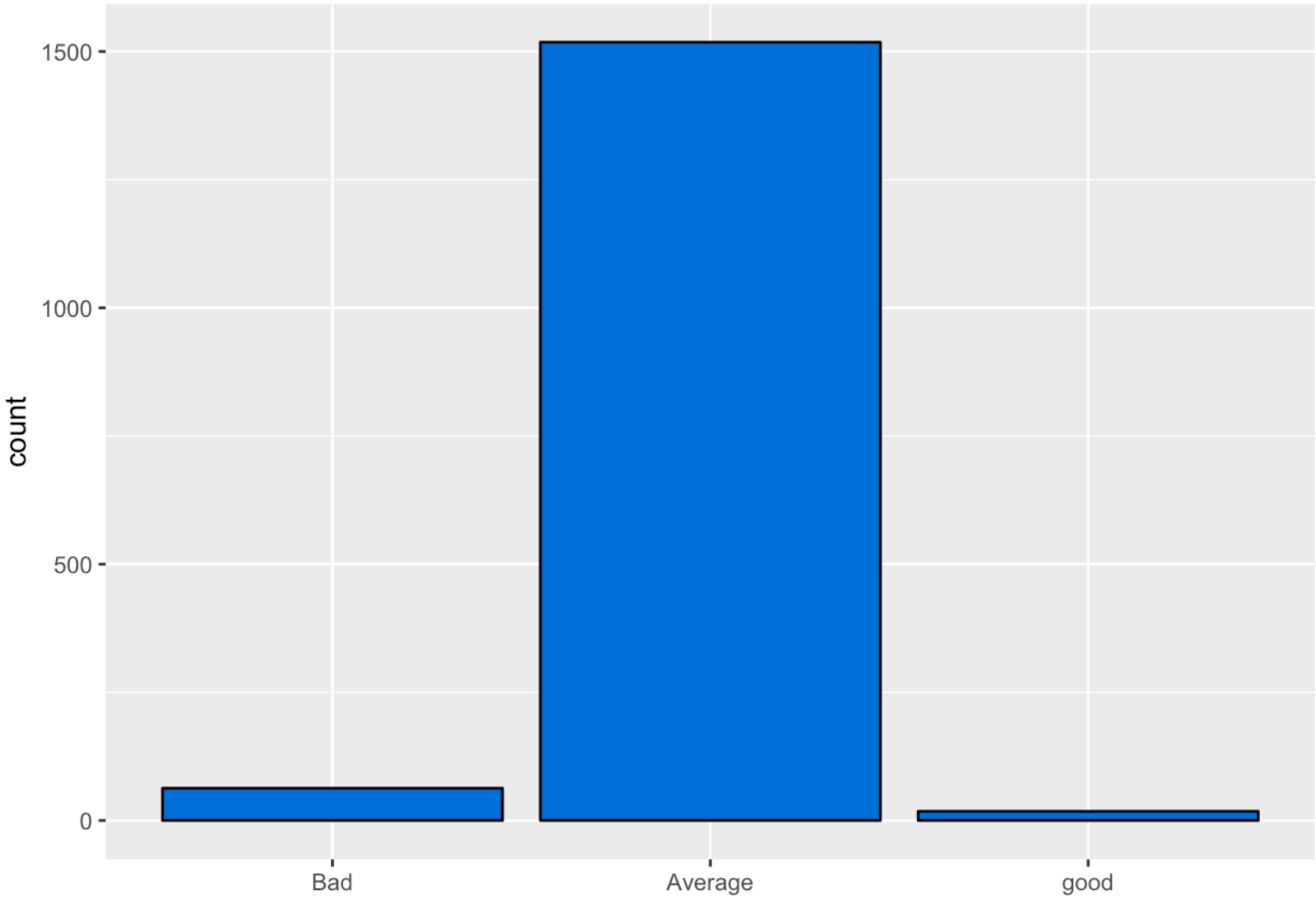

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

The distribution of alcohol has a peak at 9.4 . The value of alcohol ranges from 8.40 to 14.90 . The summary results indicates the presence of outliers.




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

The distribution of quality has the highest peak at 5 and the second highest at 6. The quality ranges from 3(bad) to 8 (very good) . quality is the output variable.




```
##      Bad Average    good
##       63    1518      18
```

A variable(quality_cat) is added to categorize the quality of the wine into Bad,Average and Good. A wine is categorized into Bad,when its value is 4 or less. A wine is categorized into Average,when its value falls in the range of 5 to 7. A wine is categorized into good,when its value is 8 and above.

The distribution of quality_cat shows that most of the wines falls under the average wine quality. The summary clearly shows the observations in each category. Above 90% of the wines falls in average category of wine.

Univariate Analysis

What is the structure of your dataset?

The red wine dataset consists of 1599 observations with 13 variables. The variables are id, fixed.acidity, volatile.acidity , citric.acid ,residual.sugar ,chlorides ,free.sulfur.dioxide , total.sulfur.dioxide,density, pH,sulphates, alcohol and quality .

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

I would like to determine “Which chemical properties influence the quality of red wines?”.The main features in the data are Alcohol and pH. I think, - alcohol decides the quality and taste of the wine - pH describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic). so a right pH will influence the quality.

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

The other features like citric acid ,density ,Fixed .acidity, volatile.acidity ,sulphates ,free.sufur.dioxide, total .sufur.dioxide,chlorides and sulphates might support the investigation. These variables can be used to build a predictive model to determine the influence on the quality of the red wines.

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

A variable(quality_cat) is added to categorize the quality of the wine into Bad,Average and Good corresponding to the values - upto 4, 5 to 7 and 8 to 10 respectively. Another variable (wine_quality) is added to store the factor of quality variable.

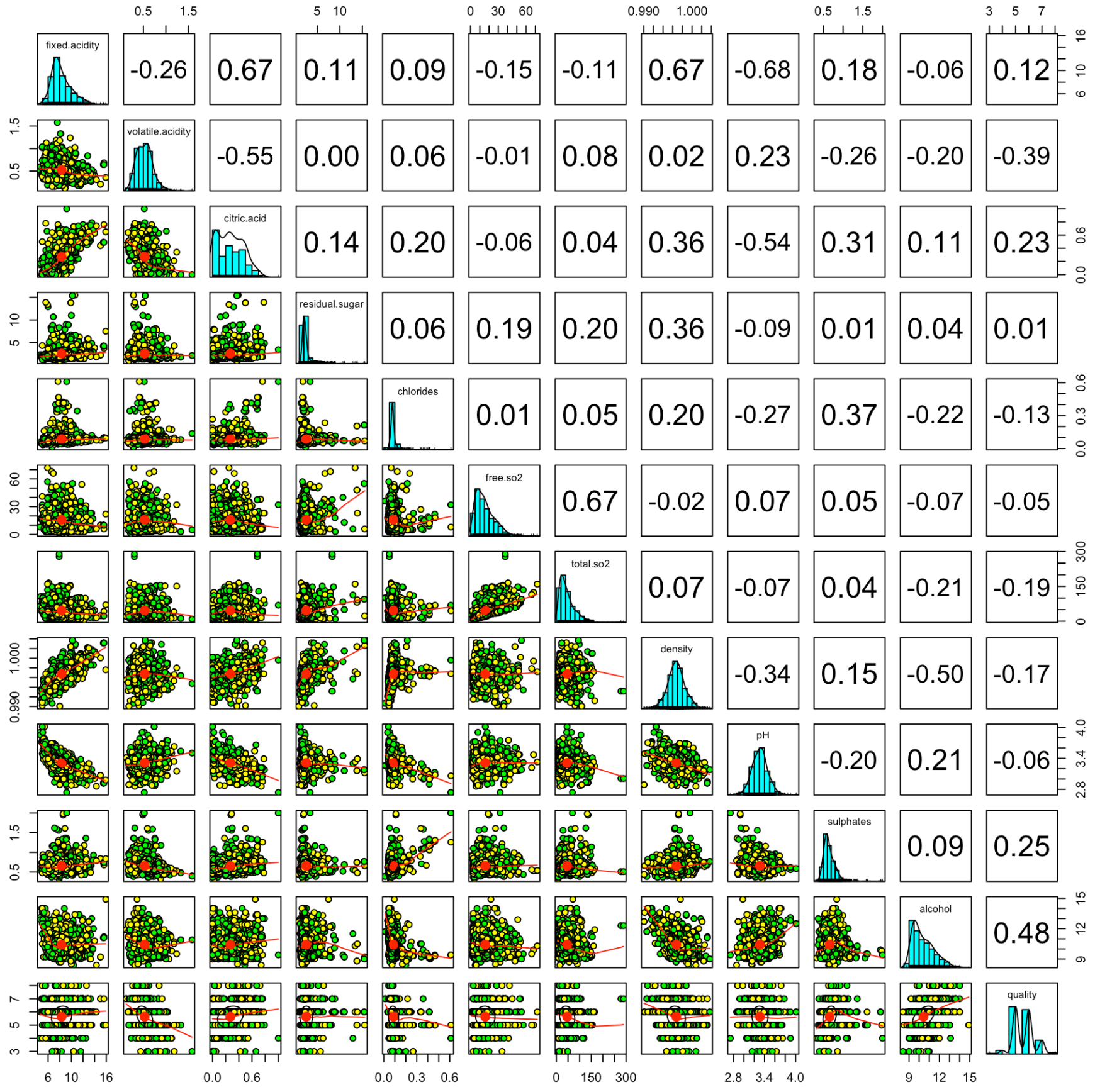
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

The chosen dataset is in tidy format.so there wasn’t any need to change the form of the data. pH had a normal distribution which is different from other variables distribution. The log transformations is calculated on some variables.

Bivariate Plots Section

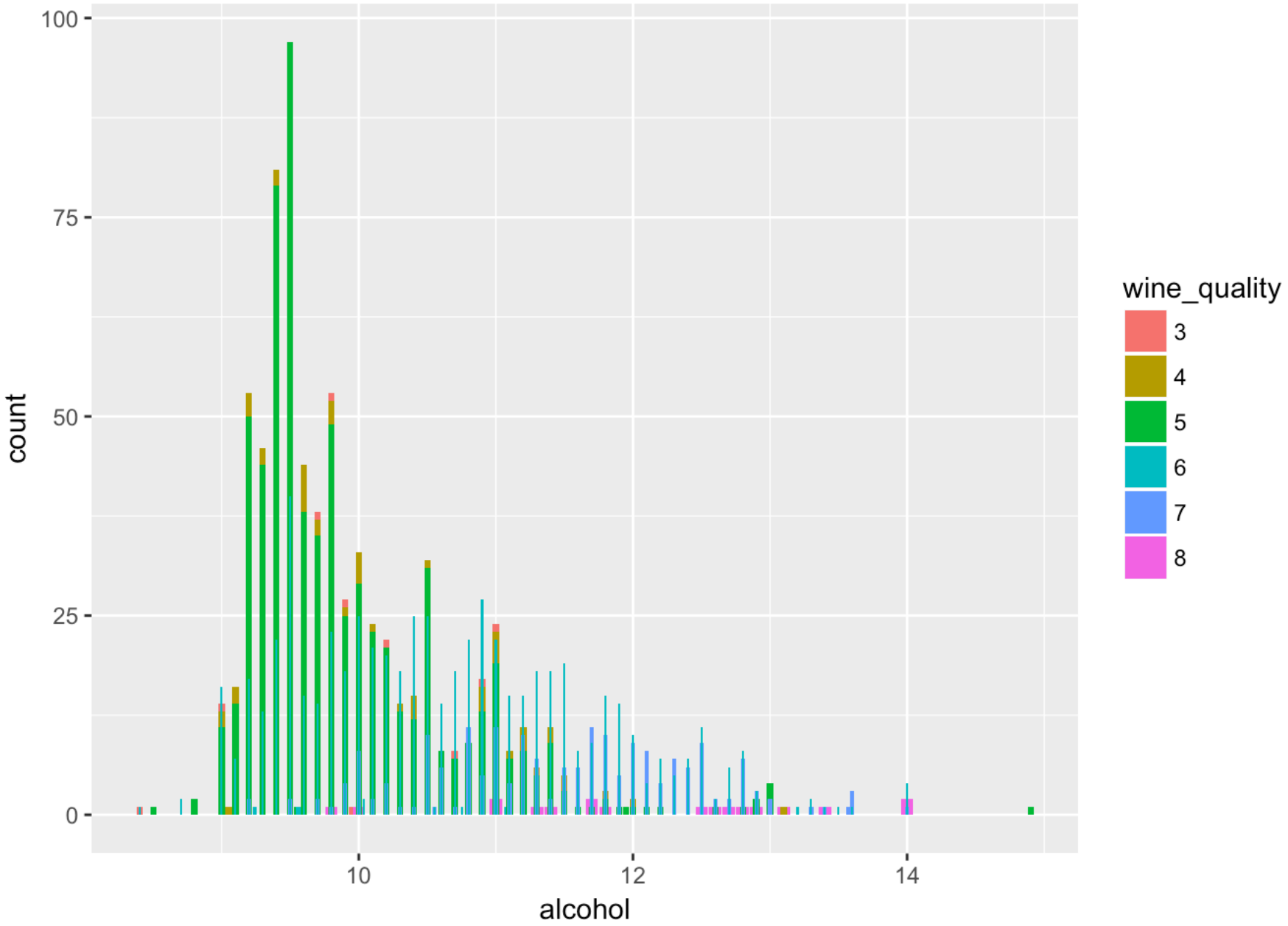
```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```

```
##                fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity      1.00000000     -0.256130895  0.67170343    0.114776724
## volatile.acidity  -0.25613089      1.000000000 -0.55249568    0.001917882
## citric.acid        0.67170343     -0.552495685  1.00000000    0.143577162
## residual.sugar     0.11477672      0.001917882  0.14357716    1.000000000
## chlorides          0.09370519      0.061297772  0.20382291    0.055609535
## free.so2          -0.15379419     -0.010503827 -0.06097813    0.187048995
## total.so2         -0.11318144      0.076470005  0.03553302    0.203027882
## density            0.66804729      0.022026232  0.36494718    0.355283371
## pH                -0.68297819      0.234937294 -0.54190414   -0.085652422
## sulphates          0.18300566     -0.260986685  0.31277004    0.005527121
## alcohol           -0.06166827     -0.202288027  0.10990325    0.042075437
## quality            0.12405165     -0.390557780  0.22637251    0.013731637
##                    chlorides     free.so2   total.so2     density
## fixed.acidity     0.093705186 -0.153794193 -0.11318144  0.66804729
## volatile.acidity  0.061297772 -0.010503827  0.07647000  0.02202623
## citric.acid       0.203822914 -0.060978129  0.03553302  0.36494718
## residual.sugar    0.055609535  0.187048995  0.20302788  0.35528337
## chlorides         1.000000000  0.005562147  0.04740047  0.20063233
## free.so2          0.005562147  1.000000000  0.66766645 -0.02194583
## total.so2         0.047400468  0.667666450  1.00000000  0.07126948
## density           0.200632327 -0.021945831  0.07126948  1.00000000
## pH               -0.265026131  0.070377499 -0.06649456 -0.34169933
## sulphates         0.371260481  0.051657572  0.04294684  0.14850641
## alcohol          -0.221140545 -0.069408354 -0.20565394 -0.49617977
## quality          -0.128906560 -0.050656057 -0.18510029 -0.17491923
##                           pH    sulphates     alcohol     quality
## 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.so2          0.07037750  0.051657572 -0.06940835 -0.05065606
## total.so2        -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
```




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

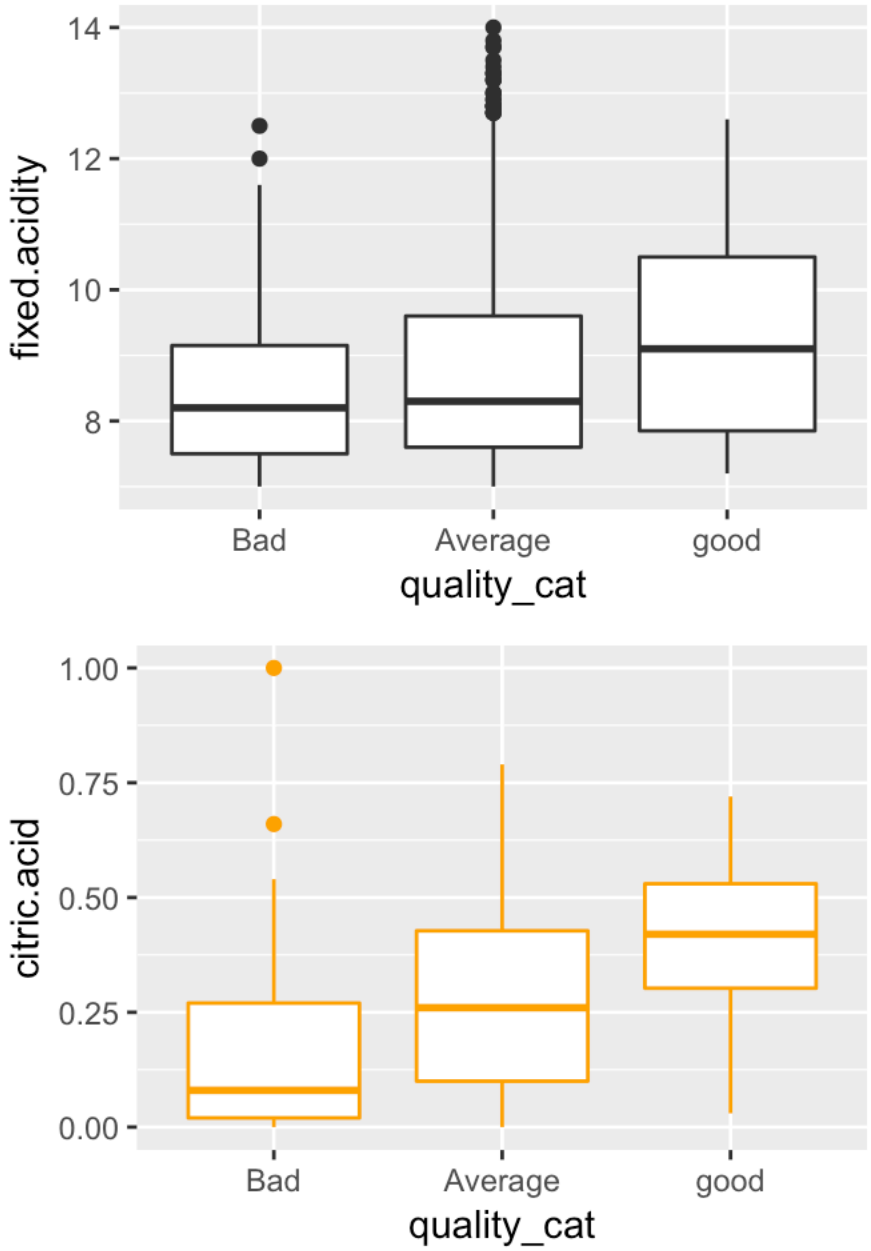
I used pairs.panels from psych package to create a correlation plot.The markings on the top or bottom of each column are the value ranges of the variable described in that particular column. for example, density has a minimum value of 0.9901 and maximum value of 1.0040 in the summary and you can notice , density has a marking of 0.9900 to 1.00 in the plot. setting fig.width and fig.height in the r-chunks greatly helped in the visualization of the plot.

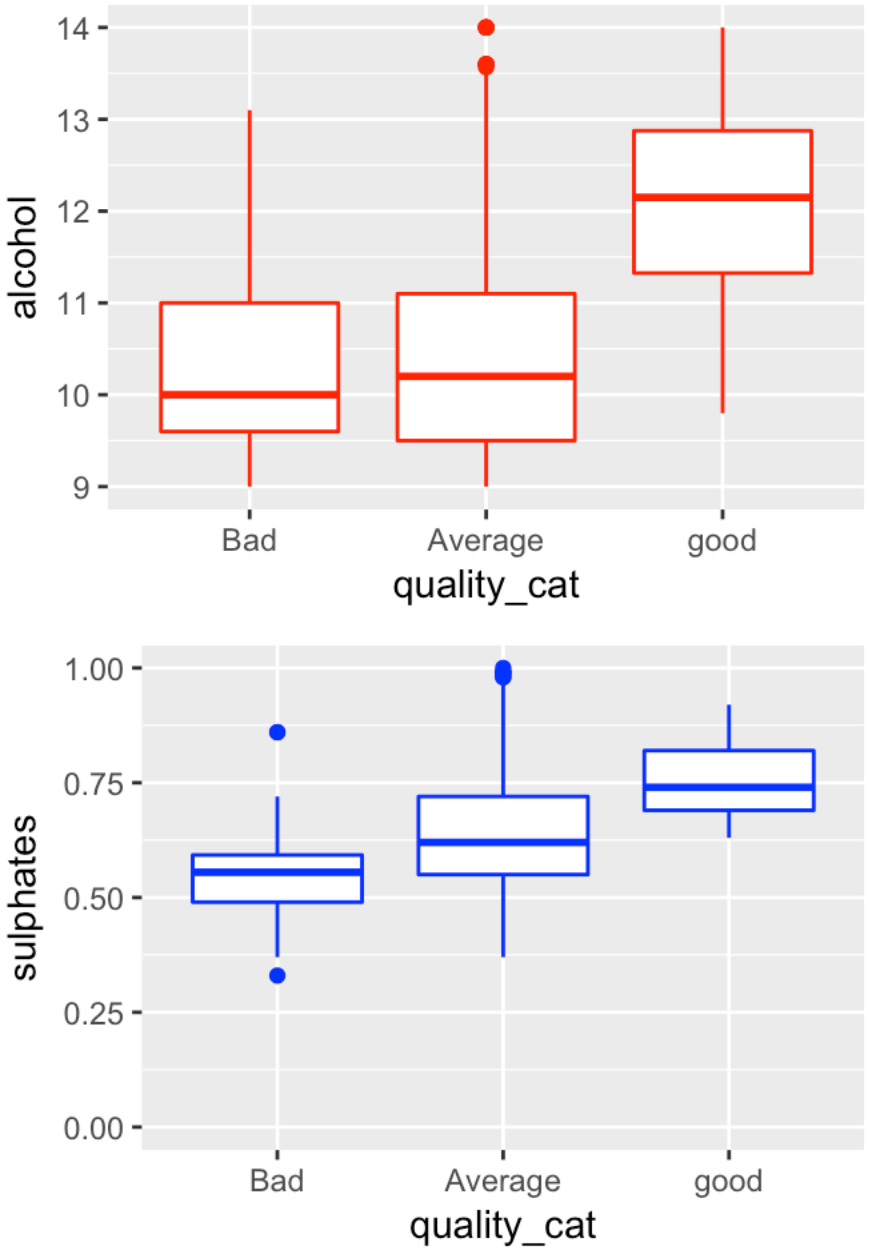

100
75
50
25
0
count
10
12
14
alcohol
wine_quality
3
4
5
6
7
8

```
## red_wine$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.580  11.000
## ------------------------------------------------------------
## red_wine$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## ------------------------------------------------------------
## red_wine$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## ------------------------------------------------------------
## red_wine$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## ------------------------------------------------------------
## red_wine$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## ------------------------------------------------------------
## red_wine$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

In the bar plot,most of the wines falls into the average quality . It clearly demonstrates the predominant distribution of wine_quality values of 5,6 and 7. The Alcohol vs wine quality have a positive correlation than quality - any other variable combination.

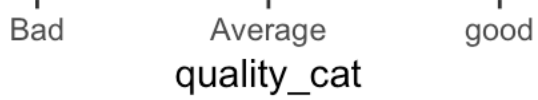
The summary results shows that the wine quality 8 have highestmedian value of 12.15.





The variables like citric.acid,fixed.acidity,alcohol and sulphates show positive correlation with the quality_cat. As the value of these variables increases the quality_cat also increases.


volatile.acidity
1.6
1.2
0.8
0.4
4
6
8
quality
volatile.acidity
1.6
1.2
0.8
0.4
Bad
Average
good
quality_cat
chlorides
0.15
0.10
0.05
0.00
Bad
Average
good
quality_cat
pH
4.0
3.5
3.0

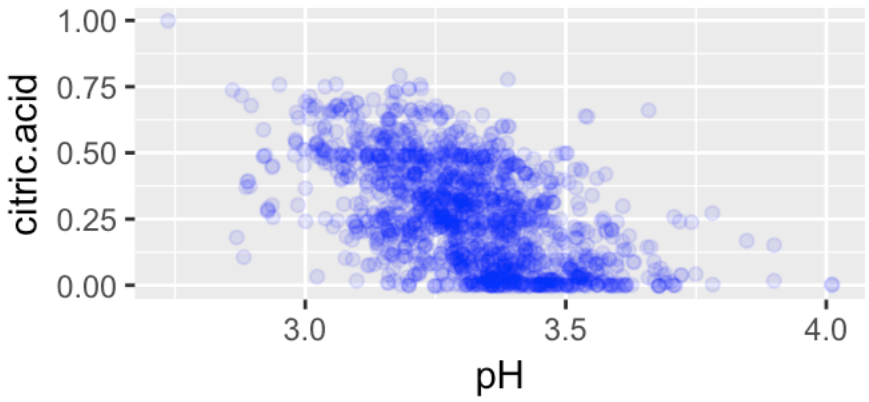


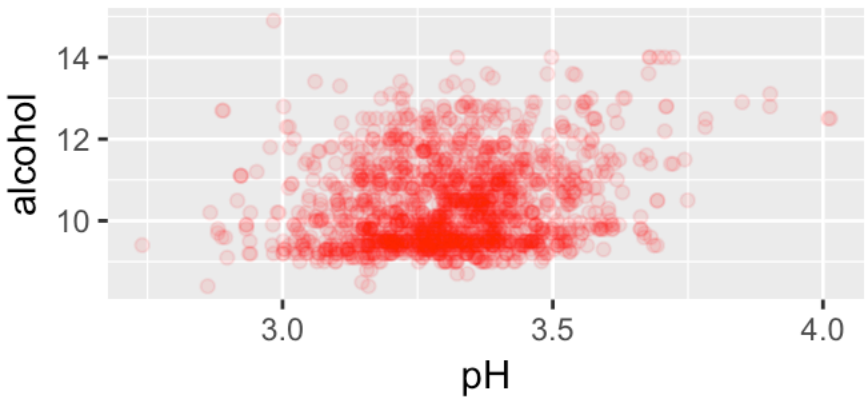

```
## red_wine$quality_cat: Bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04500 0.06850 0.08000 0.09573 0.09450 0.61000
## ------------------------------------------------------------
## red_wine$quality_cat: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08735 0.09000 0.61100
## ------------------------------------------------------------
## red_wine$quality_cat: good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04400 0.06200 0.07050 0.06844 0.07550 0.08600
```

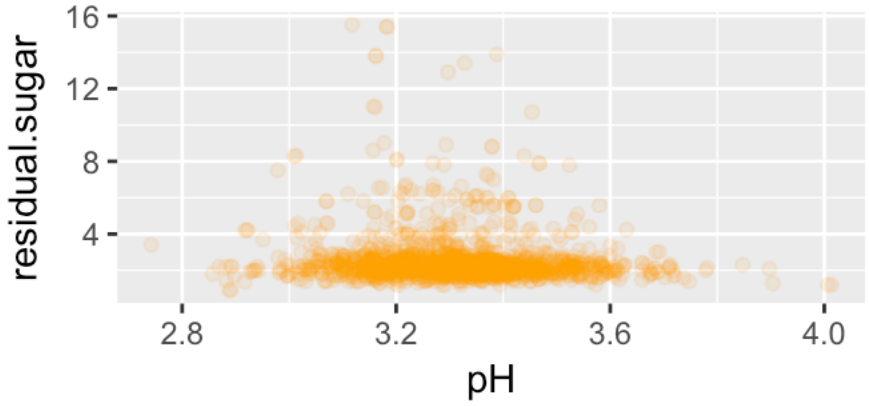
The quality of the wine decreases when the variables like pH,volatile.acidity and chlorides increases.In the chlorides-quality_cat boxplots, the median of Bad and Average categories are pretty close. The summary of chlorides with quality_cat shows their median values.

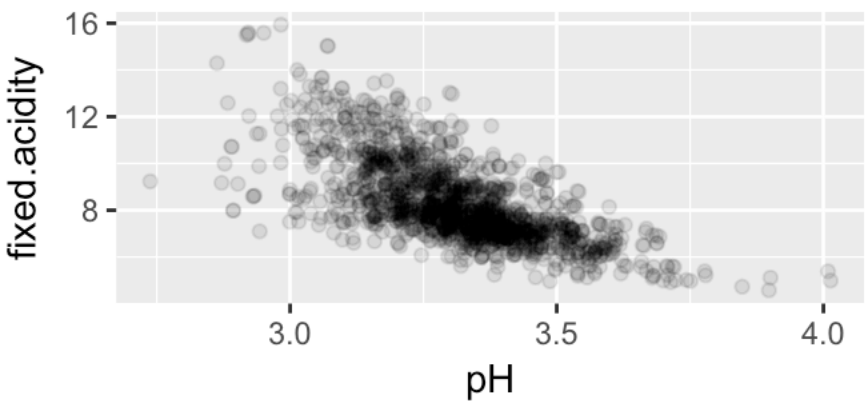
```
## red_wine$quality_cat: Bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9934  0.9957  0.9966  0.9967  0.9977  1.0010
## ------------------------------------------------------------
## red_wine$quality_cat: Average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9968  0.9979  1.0040
## ------------------------------------------------------------
## red_wine$quality_cat: good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9908  0.9942  0.9949  0.9952  0.9972  0.9988
```

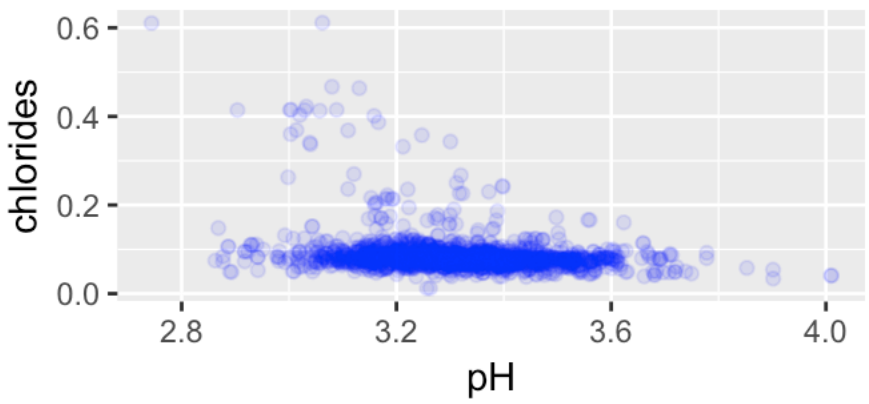
There is a mixed trend of quality_cat with the variables like density,residual.sugar, free.sulfur.dioxide and total.sulfur.oxide. The median values improves from bad to average and then drops from average to good.

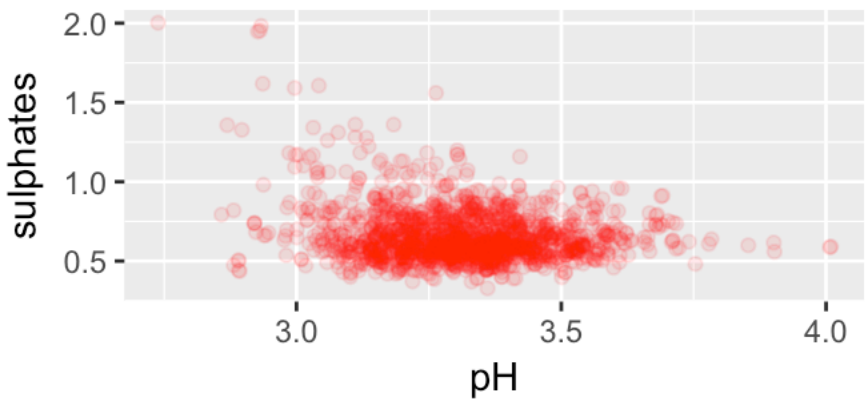

citric.acid
1.00
0.75
0.50
0.25
0.00
3.0
3.5
4.0
pH

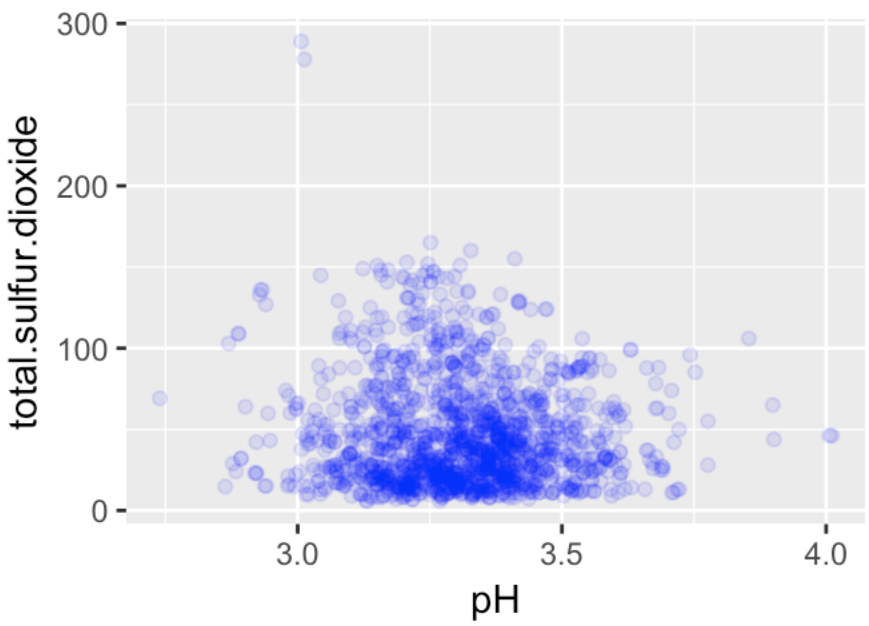

alcohol
14
12
10
3.0
3.5
4.0
pH

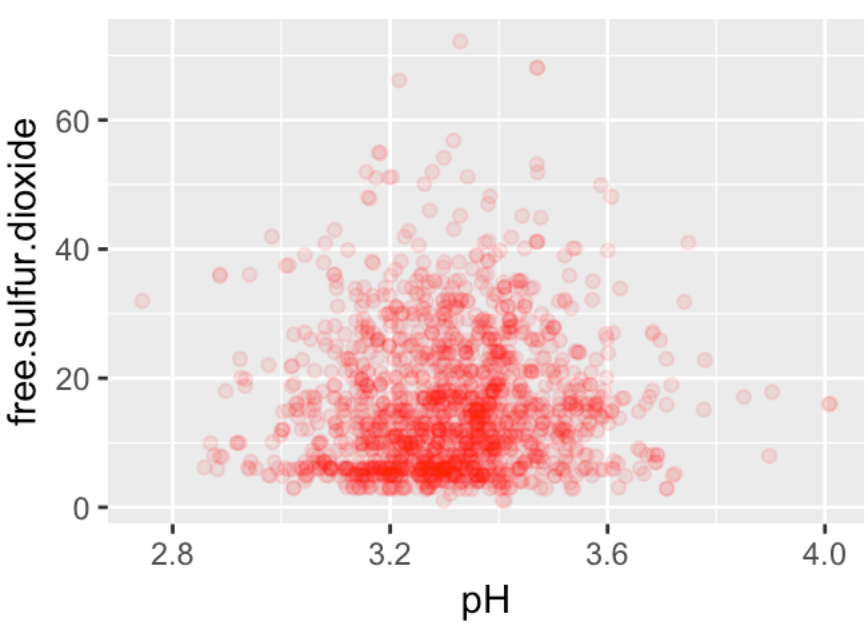

residual.sugar
16
12
8
4
2.8
3.2
3.6
4.0
pH


fixed.acidity
16
12
8
3.0
3.5
4.0
pH


chlorides
0.6
0.4
0.2
0.0
2.8
3.2
3.6
4.0
pH


sulphates
2.0
1.5
1.0
0.5
3.0
3.5
4.0
pH


total.sulfur.dioxide
300
200
100
0
3.0
3.5
4.0
pH

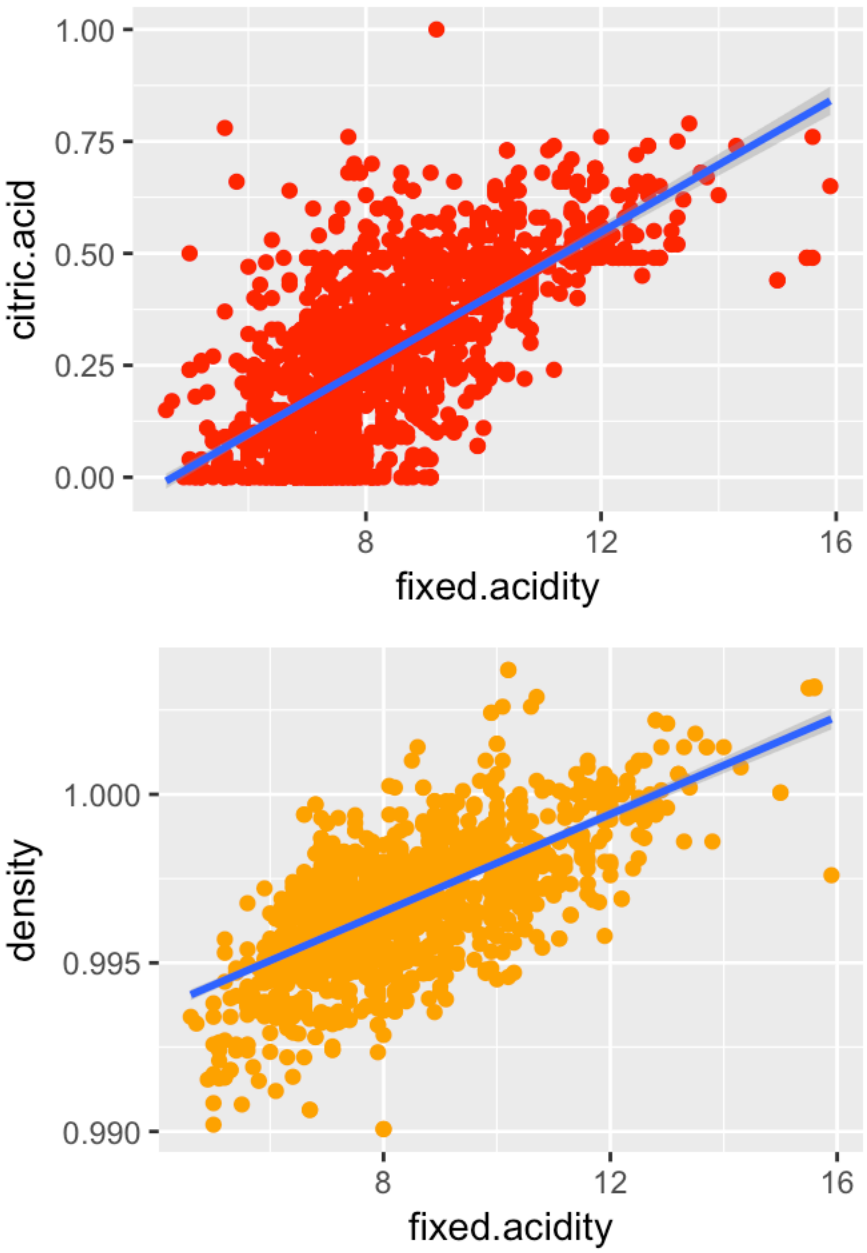

free.sulfur.dioxide
60
40
20
0
2.8
3.2
3.6
4.0
pH

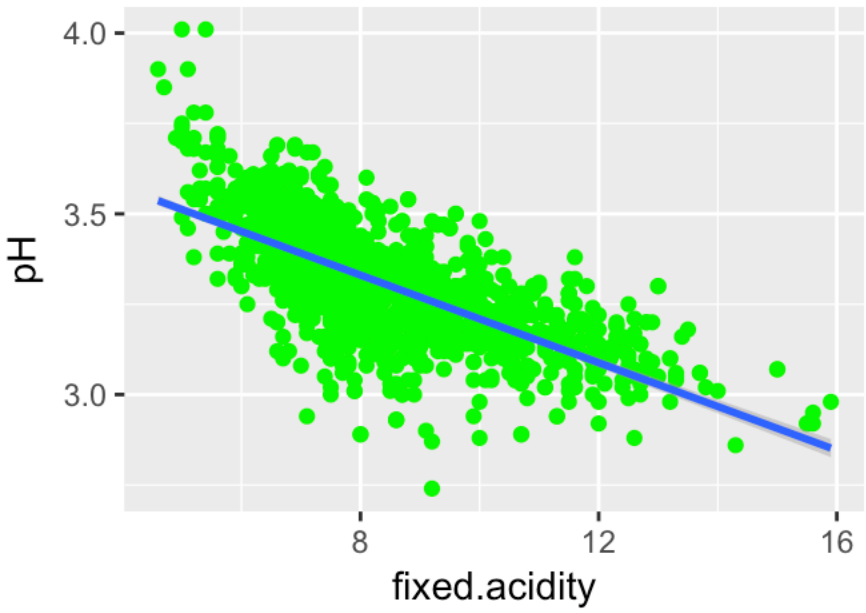
volatile.acidity
1.6
1.2
0.8
0.4

density
1.000
0.995
0.990

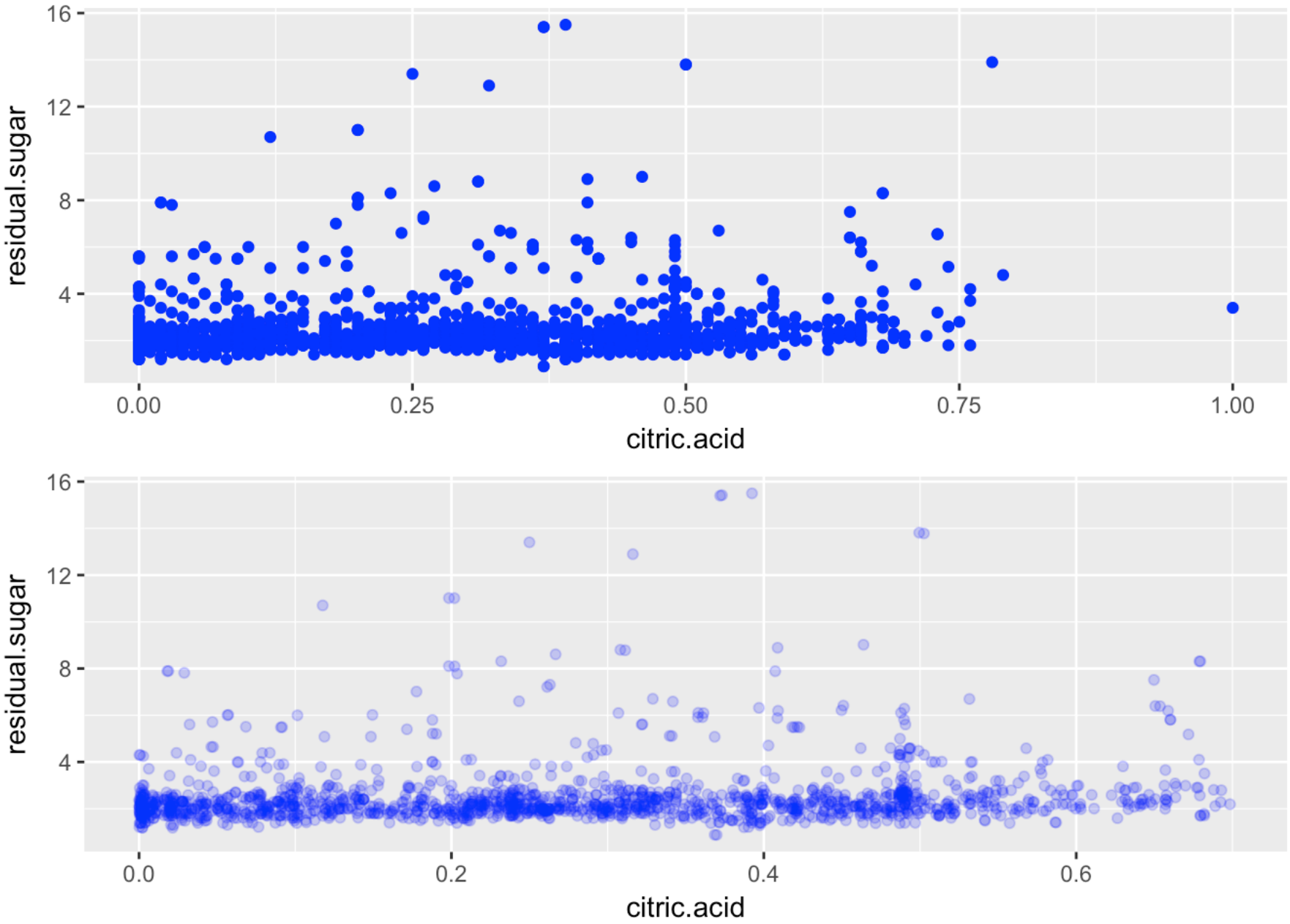
The scatterplots of citric.acid vs pH ,fixed.acidity vs pH and density vs pH shows a linear association. However, there is still quite a bit of scatter around the pattern. A positive correlation value indicates a positive linear association and negative value indicates a negative linear association Consequently, a correlation values of -0.54, 0.66, 0.67 and 0.68 are reasonable to strong. It is common for a correlation to decrease as sample size increases.

```
##                fixed.acidity citric.acid    density         pH
## fixed.acidity      1.0000000   0.6717034  0.6680473 -0.6829782
## citric.acid        0.6717034   1.0000000  0.3649472 -0.5419041
## density            0.6680473   0.3649472  1.0000000 -0.3416993
## pH                -0.6829782  -0.5419041 -0.3416993  1.0000000
```

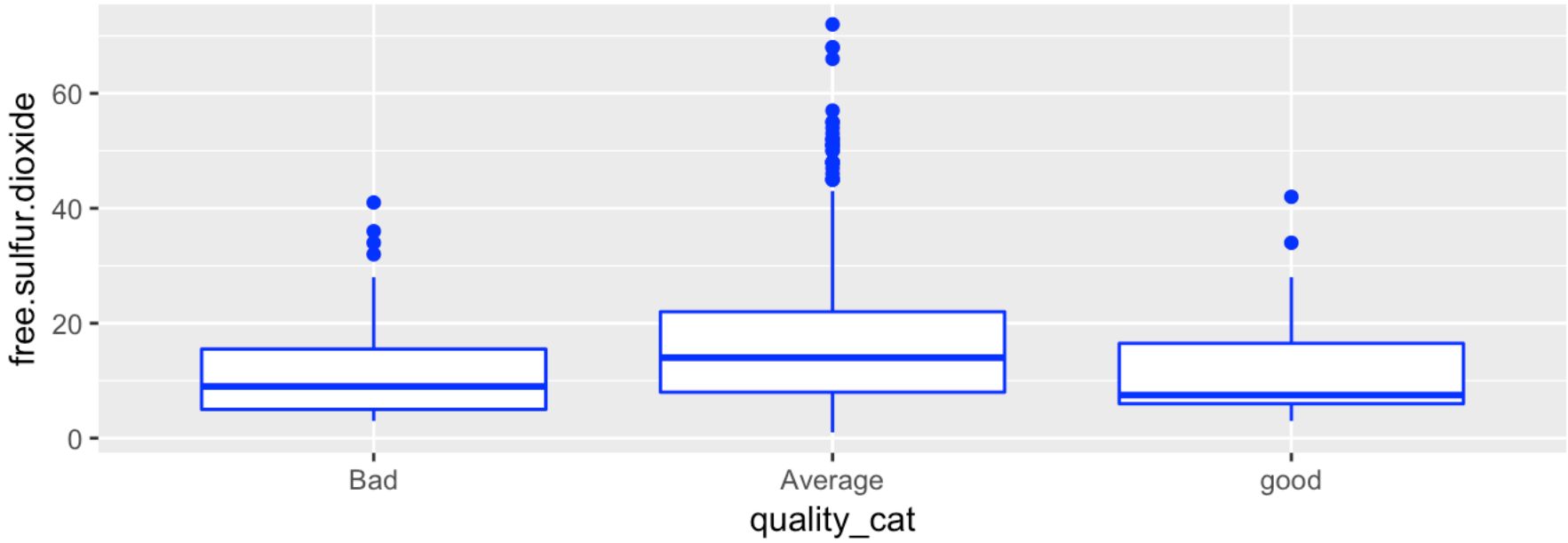


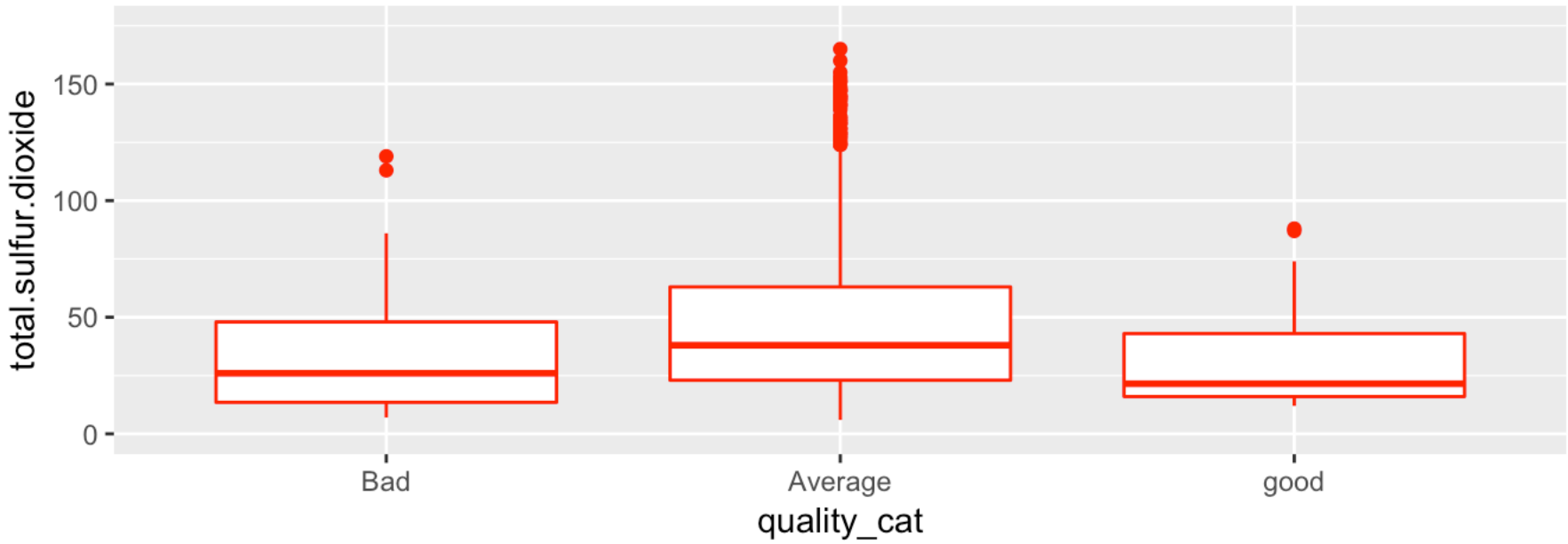


fixed.acidity has a linear association with density,citric.acid and pH .The scatterplots of these variables are shown above.

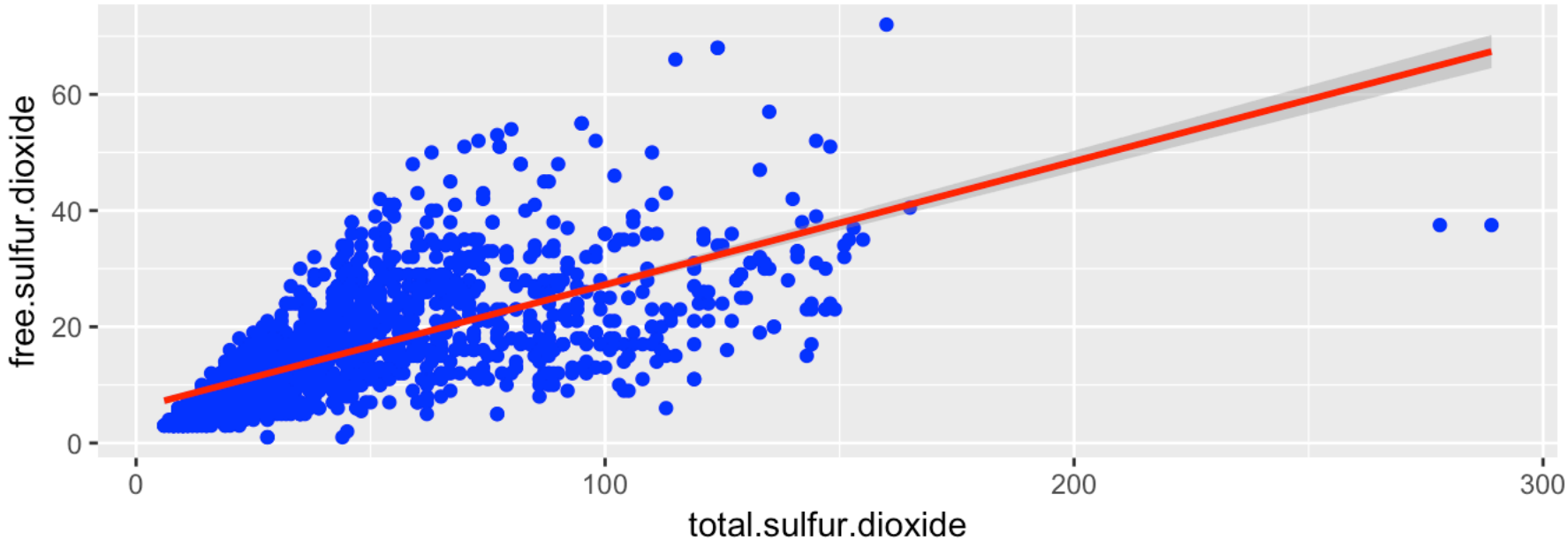


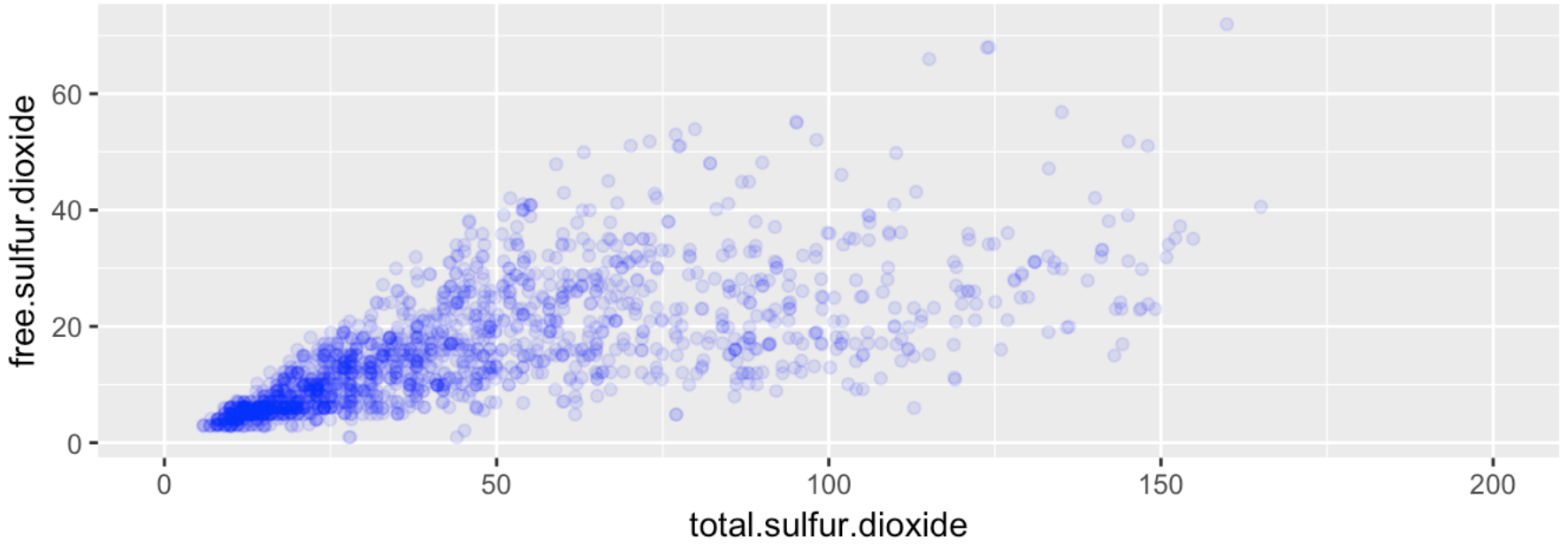
The scatterplot of residual.sugar vs citric.acid is overplotting. when geom_jitter is used,it reduces the overplotting .





The boxplots of free.sulfur.dioxide vs quality_cat and total.sulfur.dioxide vs quality_cat demonstrates same type of distribution but their values varies.





I was curious to see the distribution of Free.sulfur.dioxide vs total.sulfur.dioxide. Free.sulfur.dioxide and total.sulfur.dioxide have a strong relationship with a correlation value of 0.667

Bivariate Analysis

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

The correlation tables demostrates the relationships among various variables.

The variable,quality has a positive relationship with fixed.acidity,citric.acid,alcohol and residual.sugar. The quality has a negative relationship with volatile.acidity,pH and chlorides.

A relationship between two variables are said to be

- strong when the correlation values are between (0.5 and 1) or (-1 and -0.5)
- moderate when the correlation values are between (0.3 and 0.5) or (-0.3 and -0.5)

The correlation value of alcohol - quality has a moderate relationship with the value of 0.476 and the volatile.acidity - quality pair has a moderate relationship with the value of -0.3905 . I was disappointed to know that the quality has 0.476 as its highest correlation value and the variable that is paired with is Alcohol.

A strong relationship is seen in pH- fixed.acidity (-0.6829782) and pH- citric.acid(-0.5419041) plots.

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

A strong relationship is seen in pH- fixed.acidity(-0.6829782) and pH- citric.acid(-0.5419041) plots.

fixed.acidity has a strong relationship with density,citric.acid and pH.

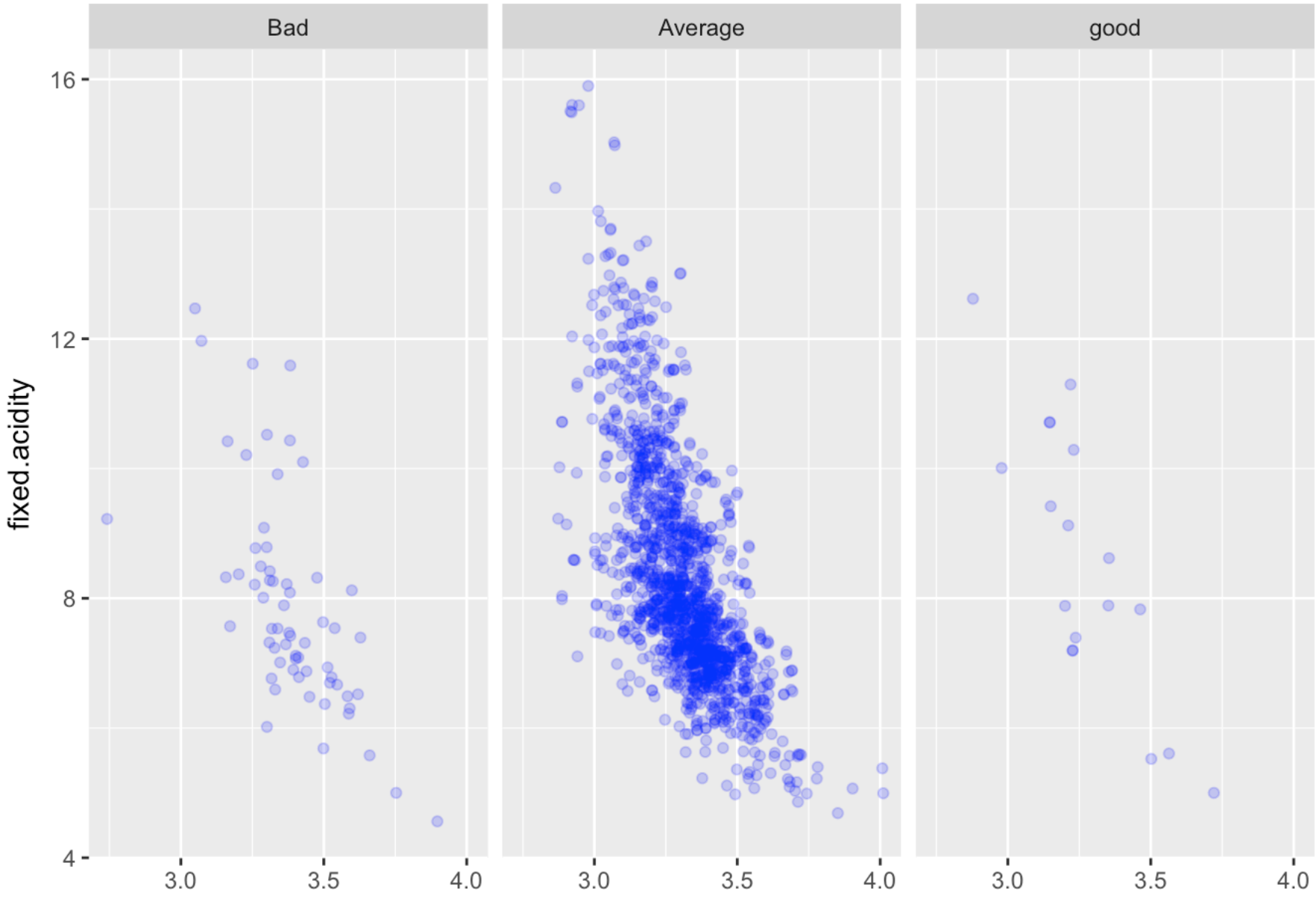
The values are as follows

citric.acid -fixed.acidity 0.6717034
density - fixed.acidity 0.6680473
pH - fixed.acidity -0.6829782

What was the strongest relationship you found?

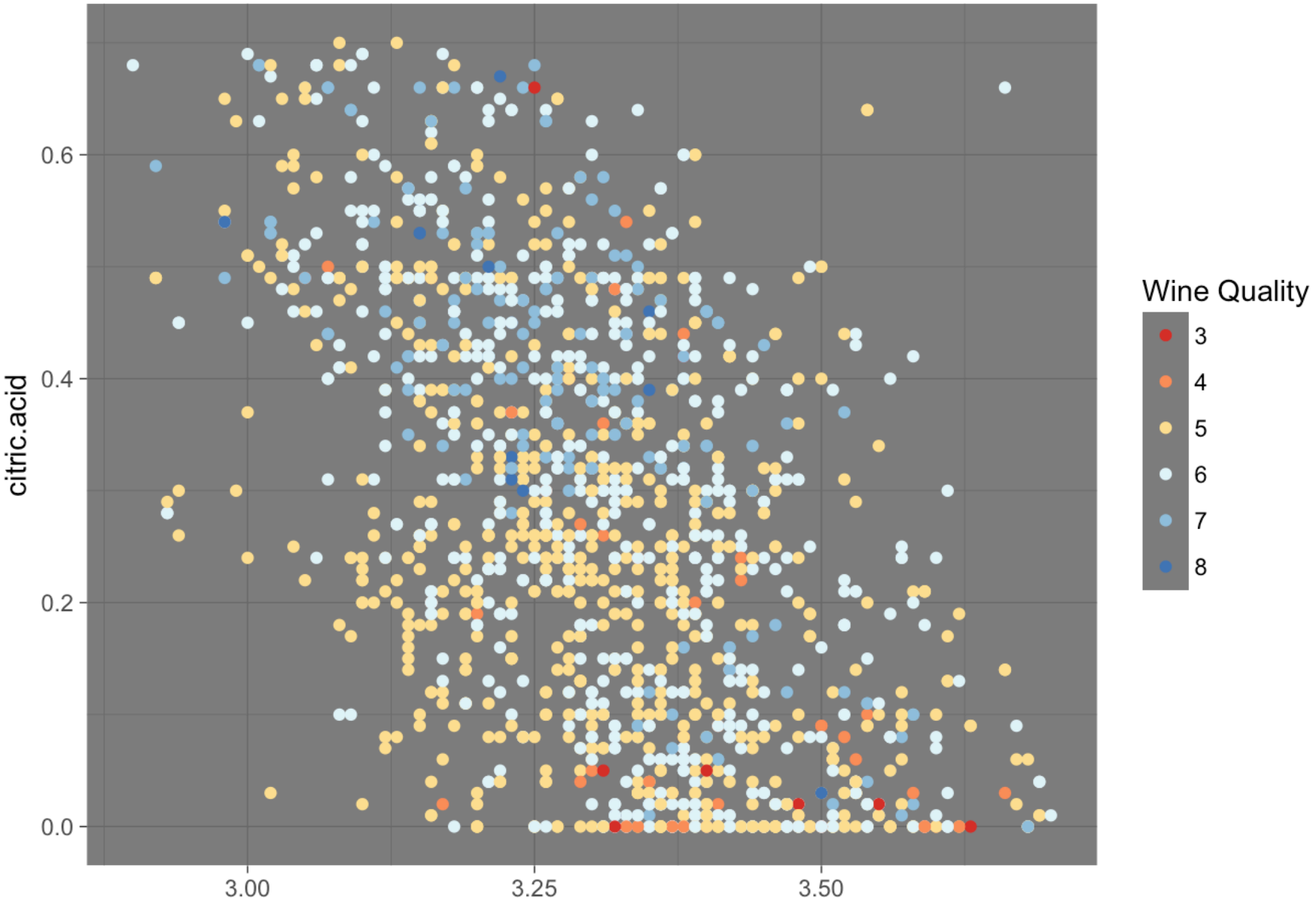
The strongest relationship is pH-Fixed.acidity with a correlation value of -0.683 .

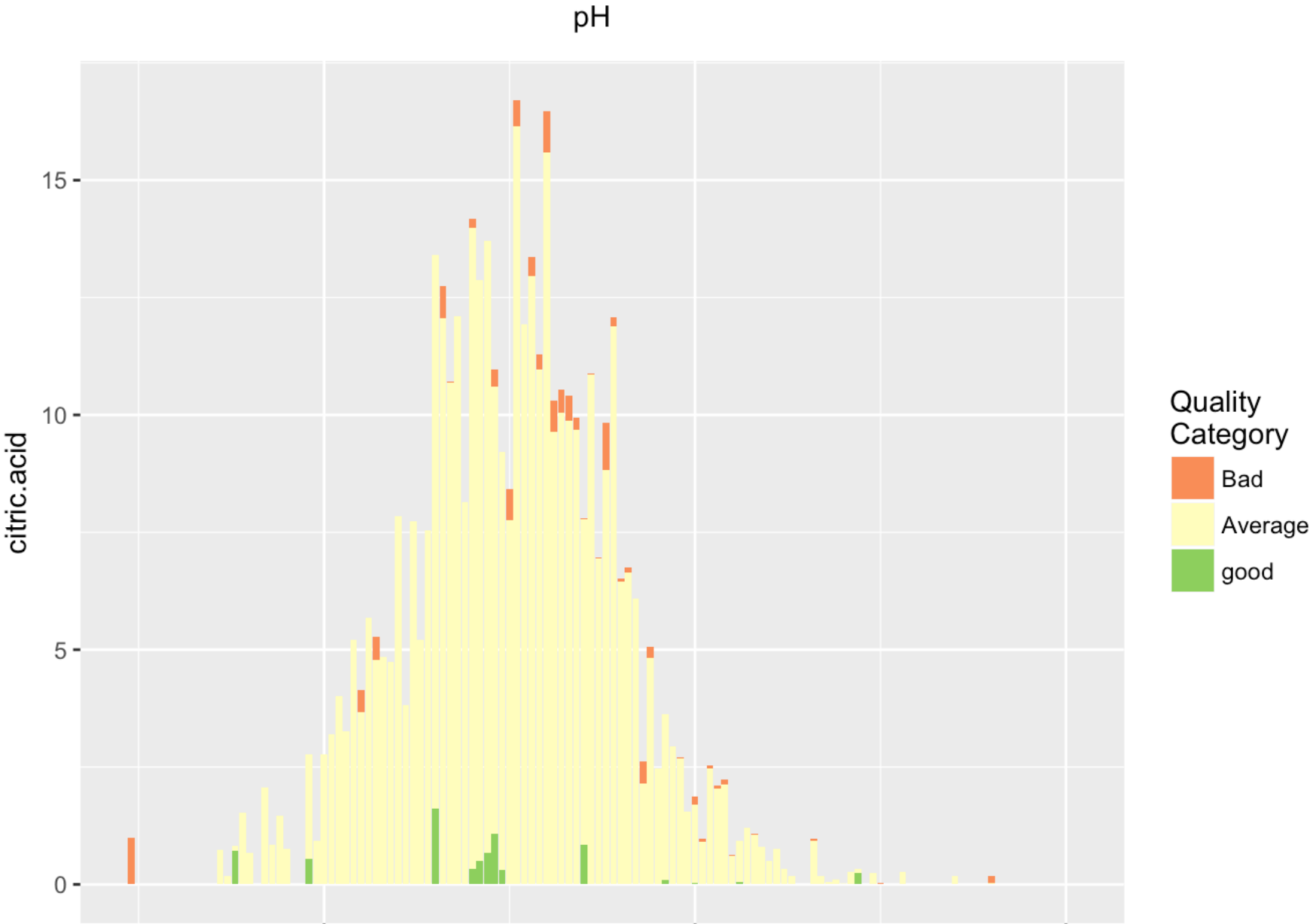
Multivariate Plots Section

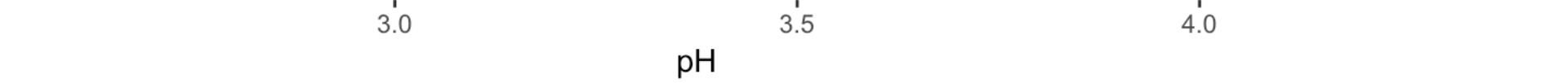

Bad
Average
good
fixed.acidity
16
12
8
4
3.0
3.5
4.0

pH
fixed.acidity
12
10
8
6
Quality Category
Bad
Average
good

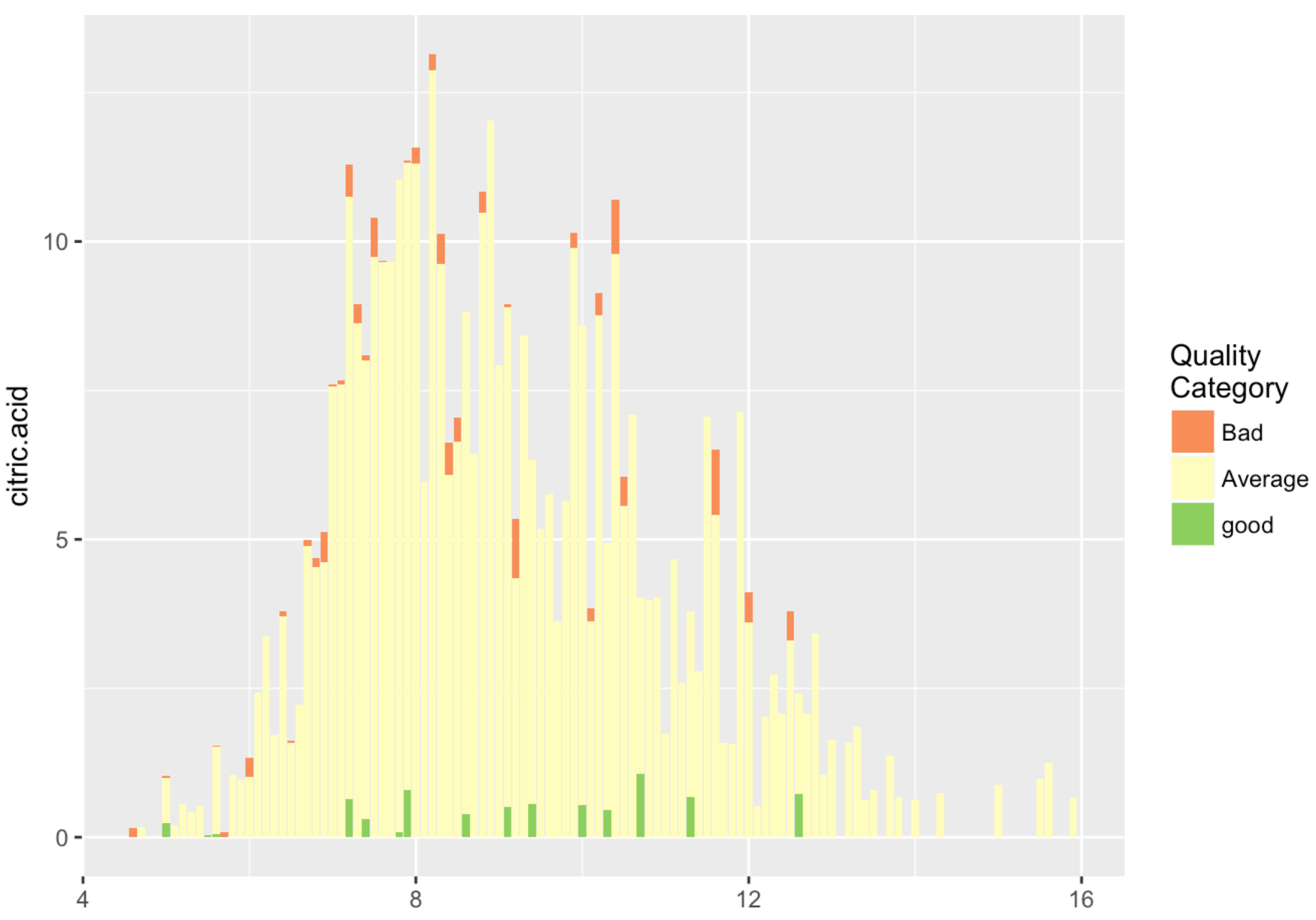
Multivariate plotting starts with the strongest relationship pair pH-fixed.acidity based on the quality_cat. The scatterplot shows that the maximum datapoints are from the average quality of wine.


citric.acid
0.6
0.4
0.2
0.0
3.00
3.25
3.50
Wine Quality
3
4
5
6
7
8


pH
citric.acid
15
10
5
0
Quality Category
Bad
Average
good

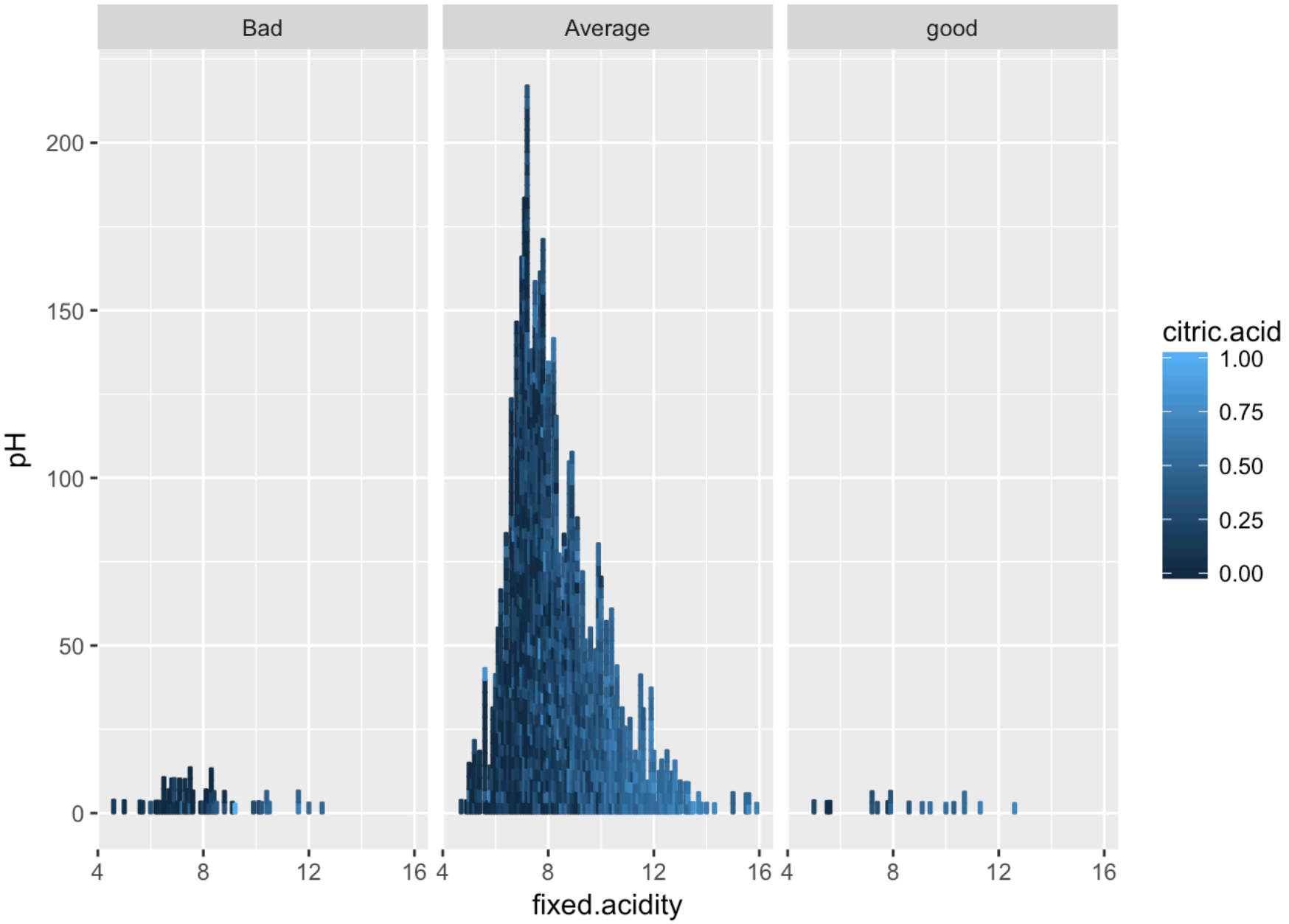


Distribution of Citric acid - pH based on the quality_cat , have most of the values in average quality. citric acid-pH have a correlation value of -0.542

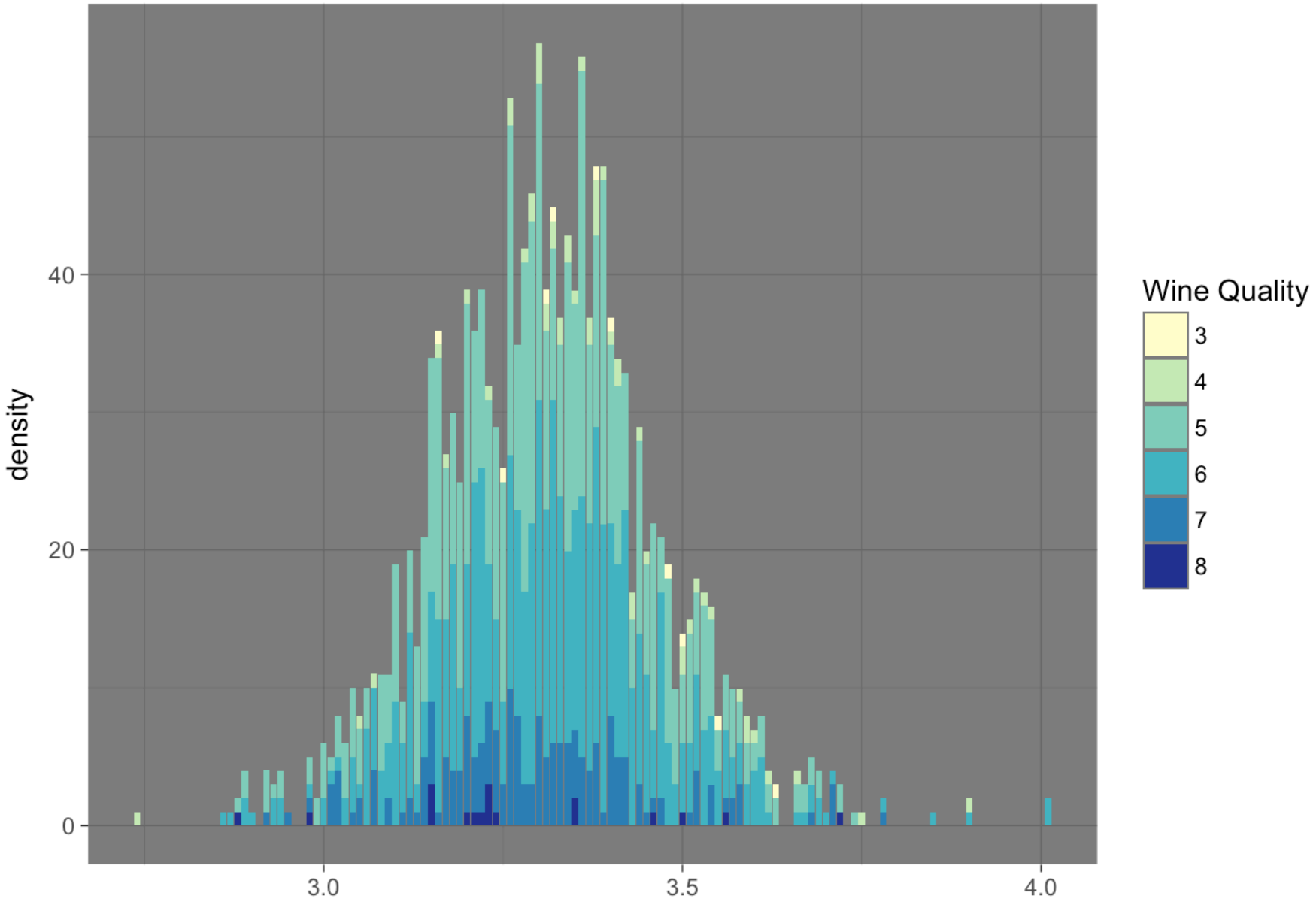

citric.acid
10
5
0
4
8
12
16
Quality
Category
Bad
Average
good


fixed.acidity
citric.acid
0.6
0.4
0.2
0.0
Wine Quality
3
4
5
6
7
8

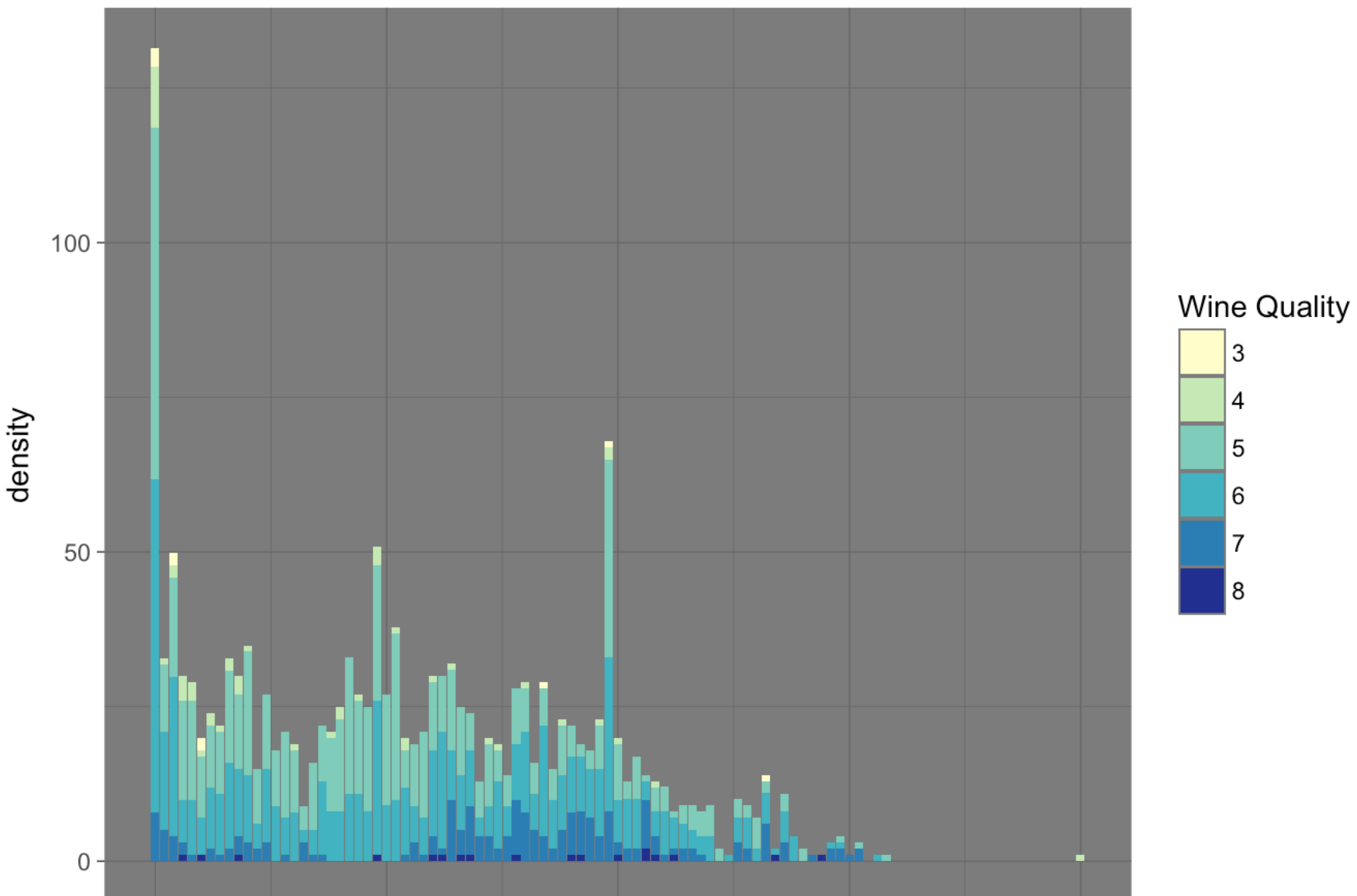
Distribution of Citric.acid - fixed.acidity based on the quality show a strong relationship with a correlation value of 0.67.

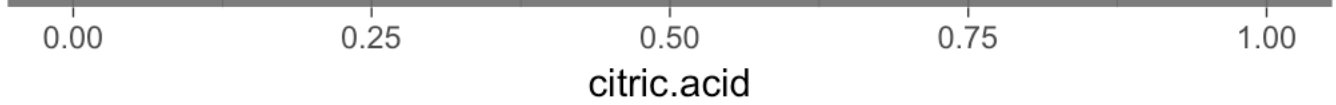


Distribution of pH- Fixed.acidity with citric.acid as color and face_wrap with quality_cat has a four variable distribution.when I used factor of pH values, I had a huge list of values in different colours(created different levels). so, I used pH values without the factor. I was so curious to explore this combo because of the strong relationship between the variables pH, Fixed.acidity and citric.acid .

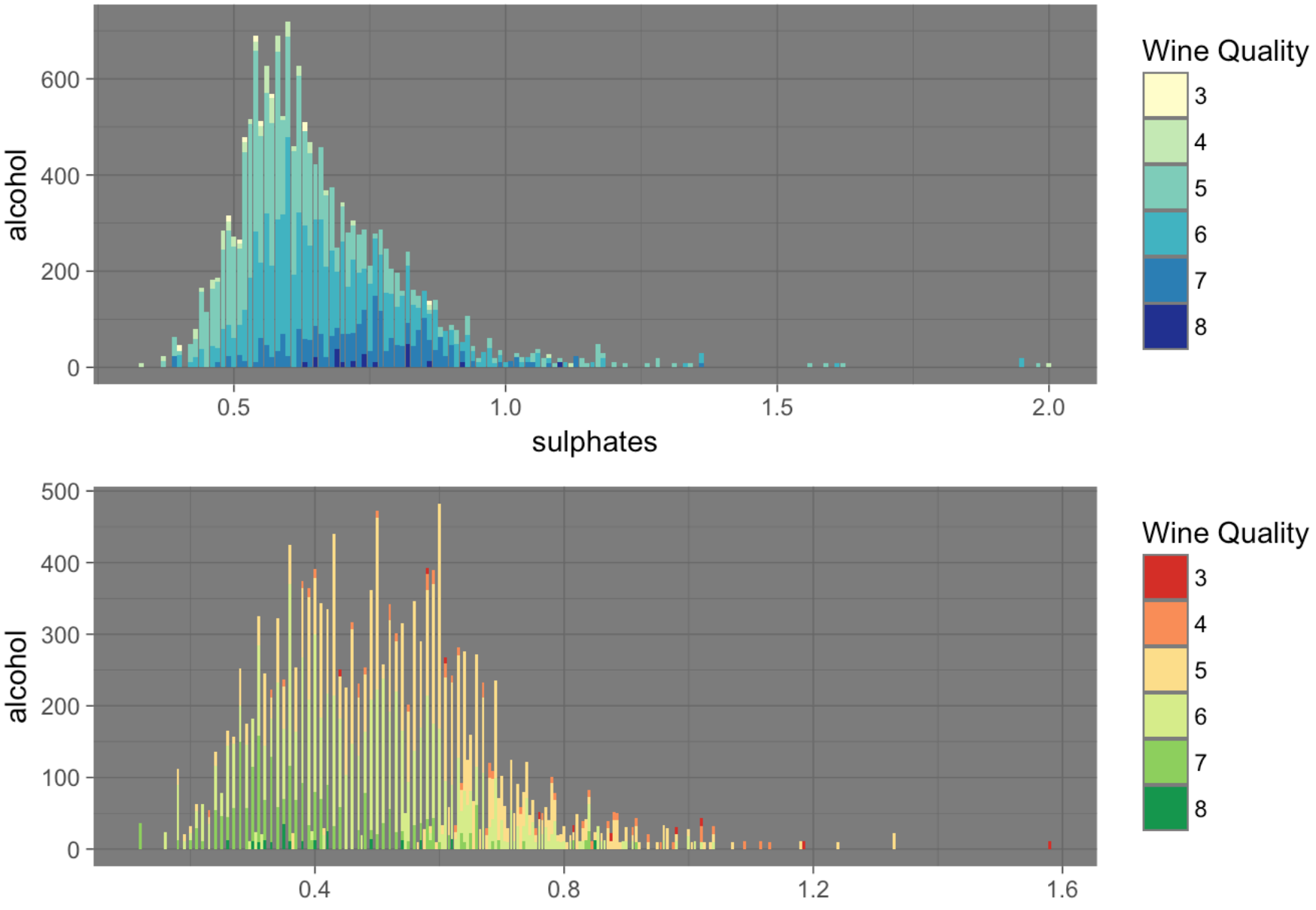

density
40
20
0
3.0
3.5
4.0
Wine Quality
3
4
5
6
7
8

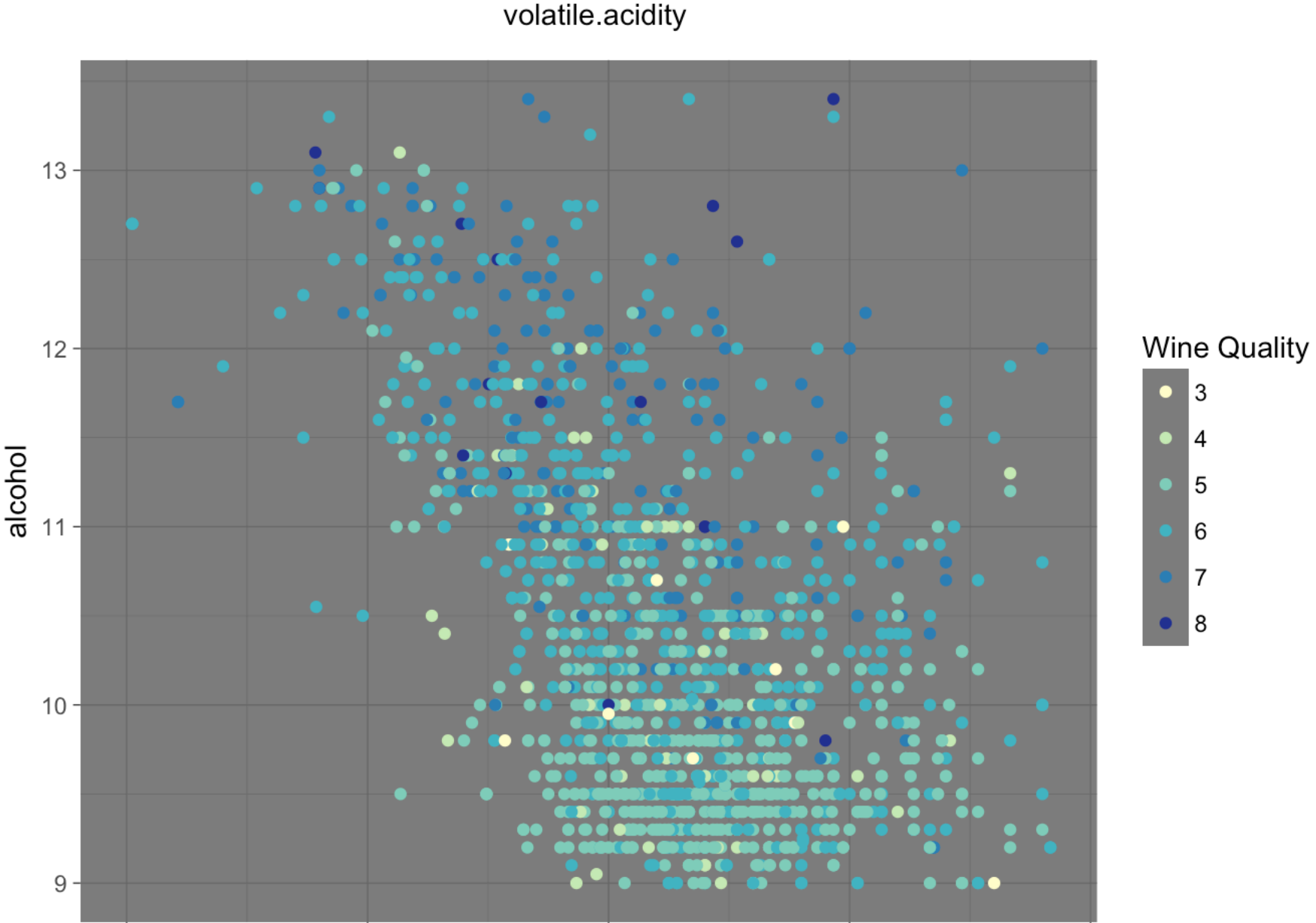
pH

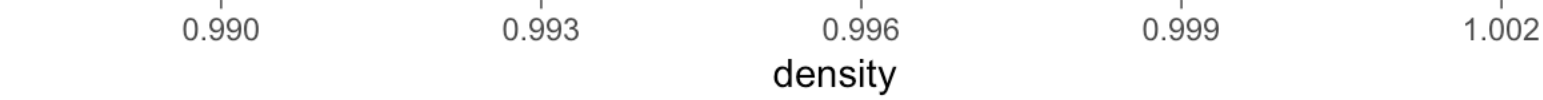

density
100
50
0
Wine Quality
3
4
5
6
7
8



Density-pH and Density - citric.acid based on the wine quality have a moderate relationship.


alcohol
600
400
200
0
0.5
1.0
1.5
2.0
sulphates
Wine Quality
3
4
5
6
7
8
alcohol
500
400
300
200
100
0
0.4
0.8
1.2
1.6
Wine Quality
3
4
5
6
7
8


volatile.acidity
alcohol
13
12
11
10
9
Wine Quality
3
4
5
6
7
8



Quality have a higher relationship with Alcohol than with other variables. so I decided to add one more variable like density,volatile.acidity ,sulphates and check for the interesting plots. most of the datapoints from average wine quality(of values 5,6,7)

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 distribution of pH - Fixed.acidity with citric.acid values as colour and face wrap with Quality_cat. I was so curious to explore this combo because of the strong relationship between the variables pH, Fixed.acidity and citric.acid .

Quality have a better linear association with Alcohol than with any other variables. Even though it is better ,but it isn't strong. Quality have a second better relationship with volatile.acidity .

citric.acid- fixed.acidity and pH-fixed.acidity also have strong relationships.

Were there any interesting or surprising interactions between features?

Alcohol- density-quality plot have a surprising relationship with a correlation value of -0.496 . Alcohol have a better relationship with density than with any other variables.

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

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = red_wine)
## m2: lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar,
##     data = red_wine)
## m3: lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     fixed.acidity + chlorides, data = red_wine)
## m4: lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     fixed.acidity + chlorides + sulphates + pH, data = red_wine)
## m5: lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     fixed.acidity + chlorides + sulphates + pH + citric.acid +
##     density, data = red_wine)
## m6: lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     fixed.acidity + chlorides + sulphates + pH + citric.acid +
##     density + free.sulfur.dioxide + total.sulfur.dioxide, data = red_wine)
```

```
##
## ================================================================================
==========
##                              m1          m2          m3          m4           m5
m6
## --------------------------------------------------------------------------------
----------
##   (Intercept)            1.875***    3.099***    2.748***    3.656***    27.461       2
1.965
##                         (0.175)     (0.186)     (0.225)     (0.594)     (21.247)     (2
1.195)
##   alcohol                0.361***    0.314***    0.317***    0.305***     0.290***
0.276***
##                         (0.017)     (0.016)     (0.016)     (0.017)      (0.026)      (
0.026)
##   volatile.acidity                  -1.383***   -1.279***   -1.061***    -1.195***    -
1.084***
##                                     (0.095)     (0.099)     (0.101)      (0.119)      (
0.121)
##   residual.sugar                    -0.002      -0.006      -0.003        0.010
0.016
##                                     (0.012)     (0.012)     (0.012)      (0.015)      (
0.015)
##   fixed.acidity                                  0.038***    0.011        0.054*
0.025
##                                                (0.010)     (0.013)      (0.025)      (
0.026)
##   chlorides                                    -0.445      -1.882***    -1.586***    -
1.874***
##                                                (0.365)     (0.404)      (0.417)      (
0.419)
##   sulphates                                                 0.839***     0.885***
0.916***
##                                                           (0.111)      (0.114)      (
0.114)
##   pH                                                      -0.335*      -0.250       -
0.414*
##                                                           (0.155)      (0.189)      (
0.192)
##   citric.acid                                                          -0.361*      -
0.183
##                                                                        (0.143)      (
0.147)
##   density                                                            -24.284      -1
7.881
##                                                                      (21.678)     (2
1.633)
##   free.sulfur.dioxide
0.004*
##                                                                                   (
```

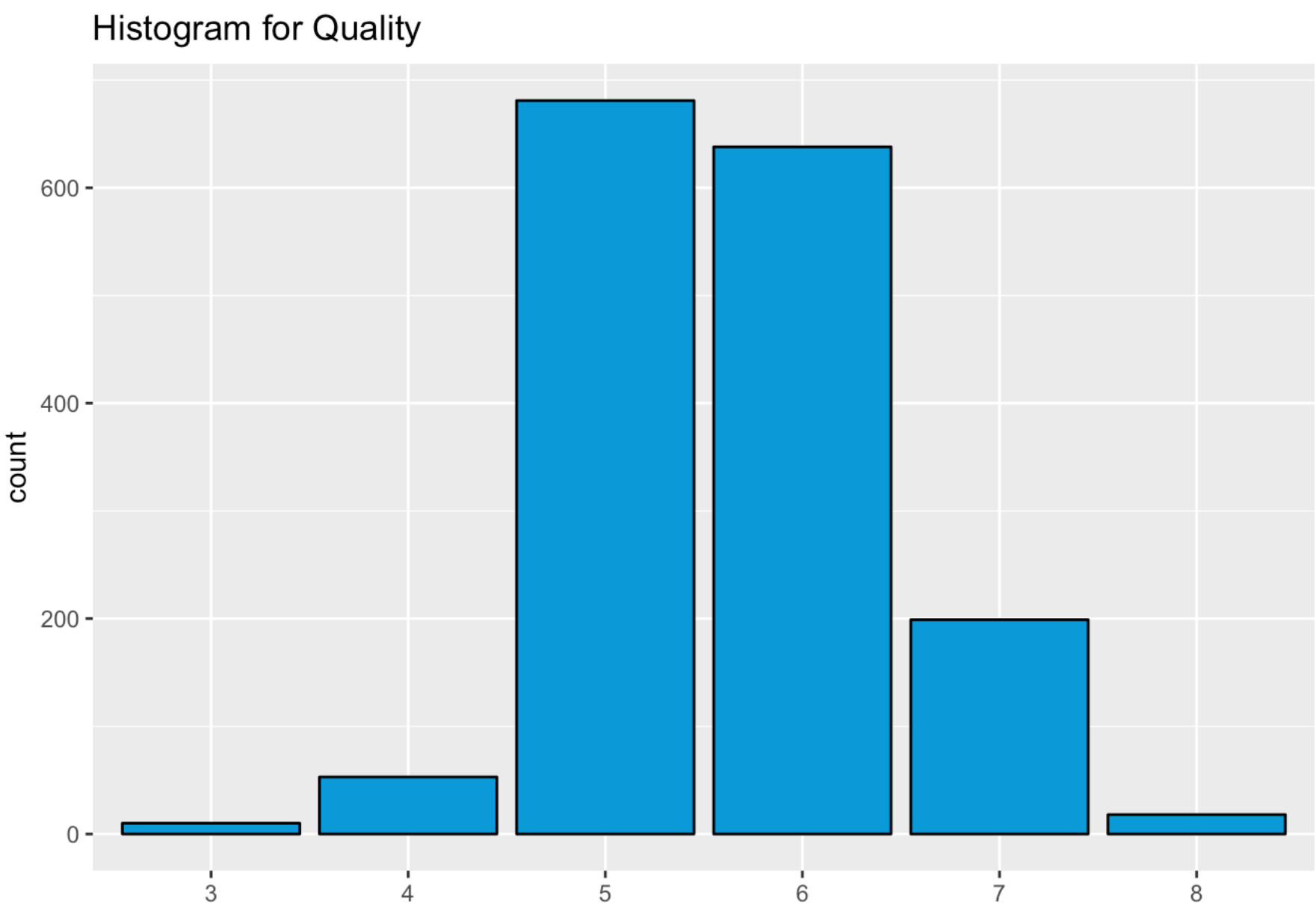
```
0.002)
##   total.sulfur.dioxide                                                    -
0.003***
##                                                                           (
0.001)
## ------------------------------------------------------------------------
----------
##   R-squared                0.2        0.3        0.3        0.3        0.4
0.4
##   adj. R-squared           0.2        0.3        0.3        0.3        0.3
0.4
##   sigma                    0.7        0.7        0.7        0.7        0.7
0.6
##   F                      468.3      246.8      152.2      121.6       95.8
81.3
##   p                        0.0        0.0        0.0        0.0        0.0
0.0
##   Log-likelihood       -1721.1    -1621.8    -1614.4    -1583.9    -1580.0    -
1569.1
##   Deviance               805.9      711.8      705.3      678.9      675.5
666.4
##   AIC                   3448.1     3253.6     3242.9     3185.8     3182.0
3164.3
##   BIC                   3464.2     3280.5     3280.5     3234.2     3241.2
3234.2
##   N                       1599       1599       1599       1599       1599
1599
## ========================================================================
==========
```

I built a linear model for quality using all variables describing chemical properties of the wine. R -squared values are 0.4(max) and 0.2(min). R squared values helps in drawing conclusions about how changes in the predictor values are associated with changes in the response value.

This model has a limitation. The red wine dataset does not include the wine quality over 8 and winequality under 3 . when the data with winequality under 3 and wine quality over 8 are added , this will improve the linear model's outcome.

Final Plots and Summary

Plot One




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

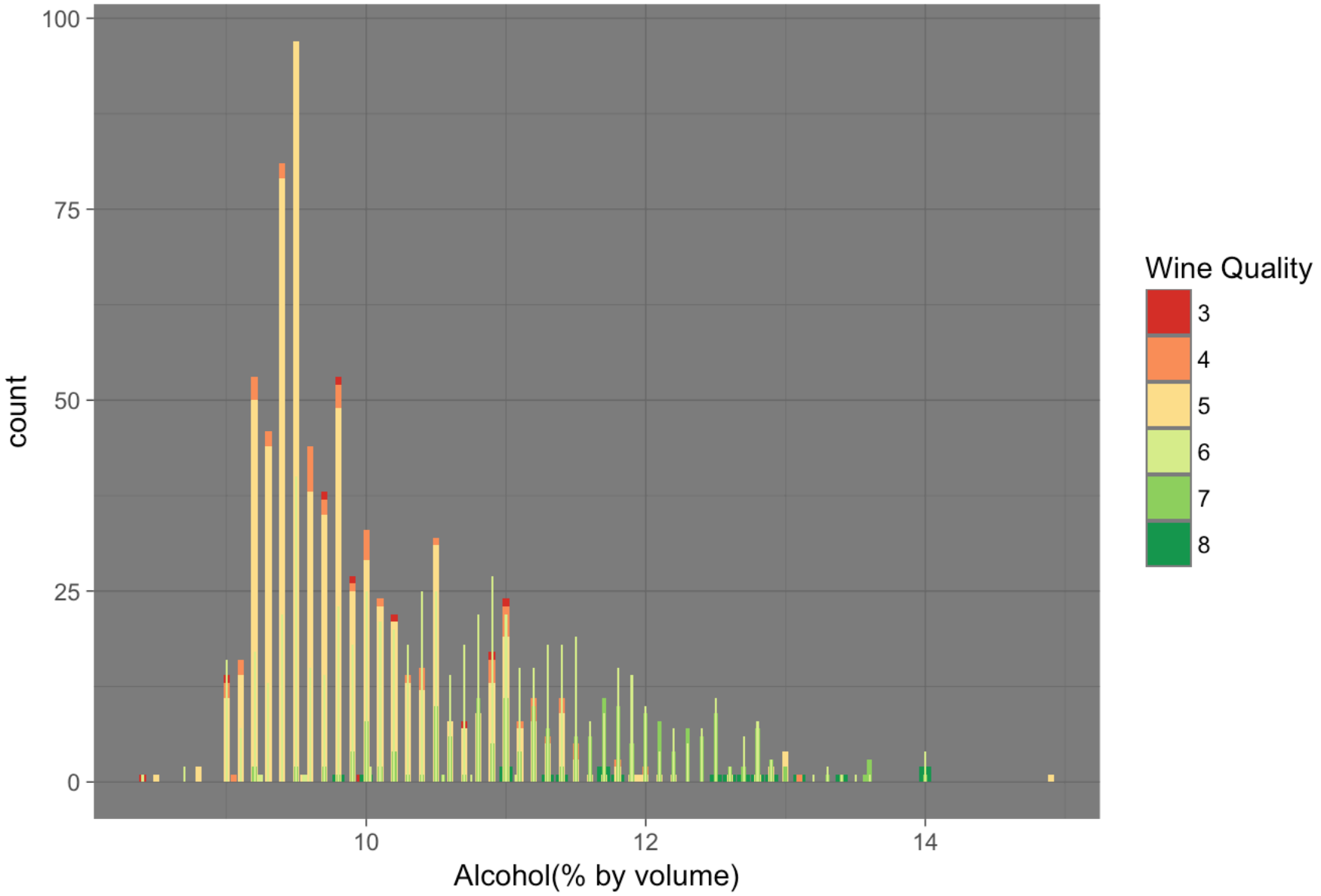
Description One

The distribution of quality has the highest peak at 5 and the second highest at 6. The quality ranges from 3(bad) to 8 (very good) . quality is the output variable.

The summary shows that the distribution of 1599 observations in the wine quality. 681 observations falls in the wine quality of 5 and 638 observations falls in the wine quality of 6.

Plot Two

Distribution of Alcohol percentage based on the wine quality



```
## red_wine$wine_quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.580  11.000
## ------------------------------------------------------------
## red_wine$wine_quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## ------------------------------------------------------------
## red_wine$wine_quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## ------------------------------------------------------------
## red_wine$wine_quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## ------------------------------------------------------------
## red_wine$wine_quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## ------------------------------------------------------------
## red_wine$wine_quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

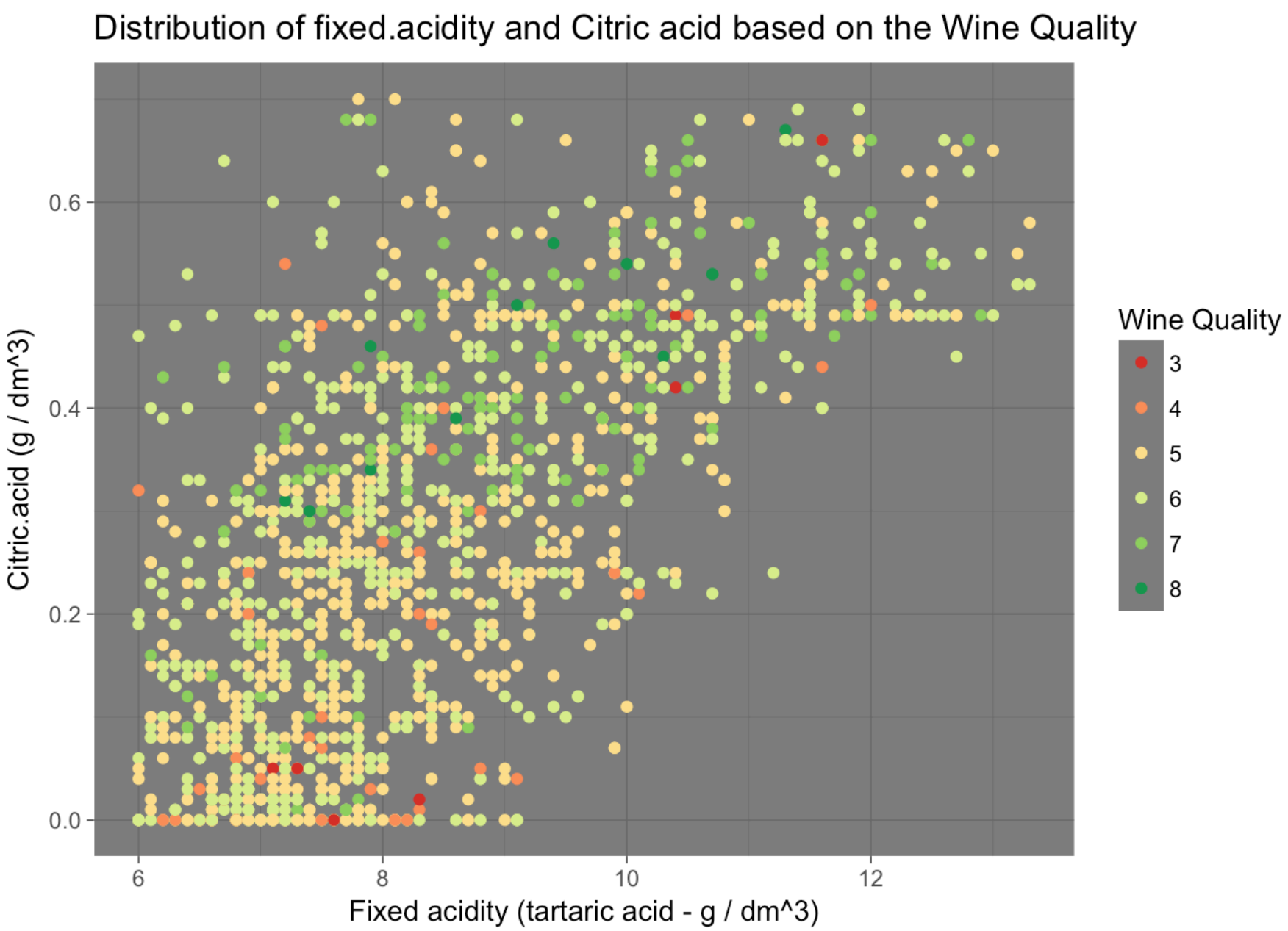
Description Two

In the boxplot for alcohol percentage based on the wine quality plot, the wine of good quality have a higher percentage of alcohol content(over 11.3). The medians of bad,average and good quality wines are gradually increasing. But , bad quality wines and average quality wines share almost same alcohol ranges. As you can see,the average quality wine has a lower first quartile value than the bad quality wine. This plot gives the impression that alcohol plays an important role in deciding the wine quality.

In the bar plot,most of the wines falls into the average quality . It clearly demonstrates the predominant distribution of wine_quality values of 5,6 and 7. The Alcohol vs wine quality have a positive correlation than quality - any other variable combination.

The summary results of Alcohol vs quality has a highest median value of 12.15 for the wine quality value of 8. The wine quality of 5 has the lowest median value of 9.7 .

Plot Three




```
##                fixed.acidity citric.acid   quality
## fixed.acidity      1.0000000   0.6717034 0.1240516
## citric.acid        0.6717034   1.0000000 0.2263725
## quality            0.1240516   0.2263725 1.0000000
```

Description Three

This scatterplot shows the distribution of fixed acidity and citric.acid based on the wine quality.citric.acid can add ‘freshness’ and flavor to wines.most acids involved with wine do not evaporate readily. Chemically the acids influence titrable acidity which affects the taste of the wine. As you can see, there is a positive correlation between citric.acid and Fixed.acidity.when these two variables increases ,it affects the taste and quality of the wine.

The correlation value of these variables shows that - fixed.acidity- citric.acid have a correlation value of 0.672 - quality has a better correlation value with citric.acid than the fixed.acidity as it can add freshness and flavor to wines.Thus improving the quality. - when fixed.acidity and citric.acid increases ,it increases the titrable acidity thus affecting the taste and quality of the wine.

Reflection

I had trouble in the correlation matrix or plot section. I used ggpairs to create the correlation plot. The plot was fine but the correlation values aren't clearly /fully visible in some cells. Also the variables like free.sulfur.dioxide and total.sulfur.dioxide didn't fit within the column cell.I tried different ways to solve this issue. But had no success.I submitted the project without solving that issue.

In the review, I was suggested to use fig.width and fig.height for this issue.I applied this, in the correlation plot.It did help .The correlation values are readable and of same size now but I had problem with plot. Nearly 30 % of the matrix was not visible .so I decided to search for other options.

I found a chart.Correlation from the package("PerformanceAnalytics"), I got a pretty matrix but had a problem with correlation value text in different sizes. The strong correlation values are in appropriate font size whereas weak correlation values were in small font size .

Then I noticed the pretty correlation plot (using psych package) given in the example project of the project description section. I renamed the variables like free.sulfur.dioxide and total.sulfur.dioxide as free.so2 and total.so2 respectively. fig.width and fig.height greatly helped in restoring its size and the plot looks pretty now.

The dataset lacks the wine quality values under 3 and over 8. These values would have impacted the outcome of the output variable. when a dataset includes the wine quality values ranging from 0 to 10, the outcome would be more precise than the current one.

There is no data about grape types, wine brand, wine selling price, etc. Each grape type vary with the sweetness,chemical properties,etc from other grape types.This would have helped in understanding the chemical properties better and hence the influence on the wine quality.

Each wine brand has a unique way of making wines or following the standard way of making wine with interesting twist on some/all processes. So ,any chemical property change because of the wine making process might have influenced the wine quality.

weather also plays an important reason for the end result of wine quality. A good weathered year produces better tasting,healthy grapes,thus influencing on the wine quality.

Moreover, more sophisticated prediction models should be able to provide more accurate predictions for the quality of wine based on its chemical characteristics.