Exploration of White Wine Quality Dataset

```
## [1] "data.frame"
```

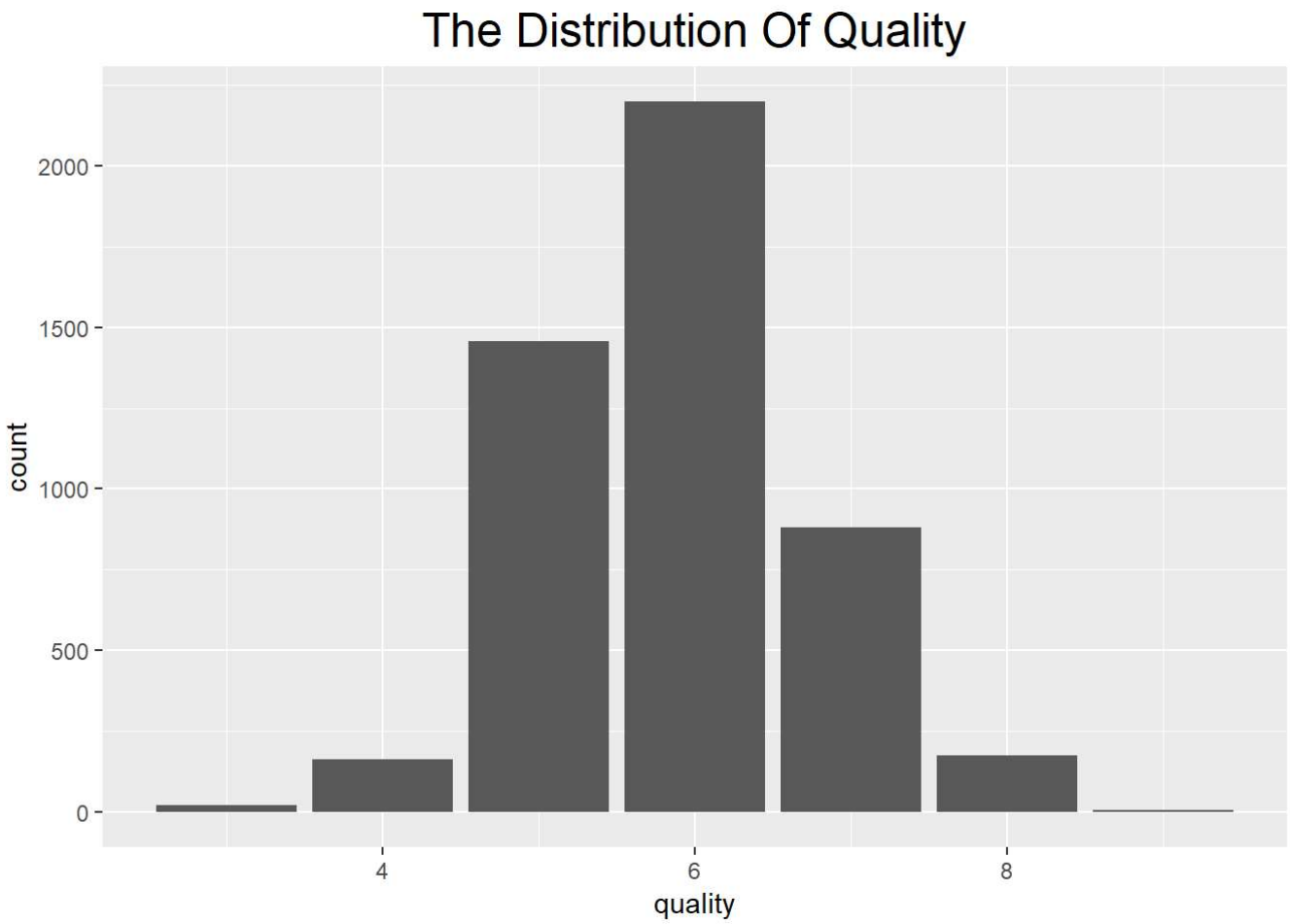
```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol          quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

This dataset consists of 12 variables(X will not be discussed), with 4898 observations.

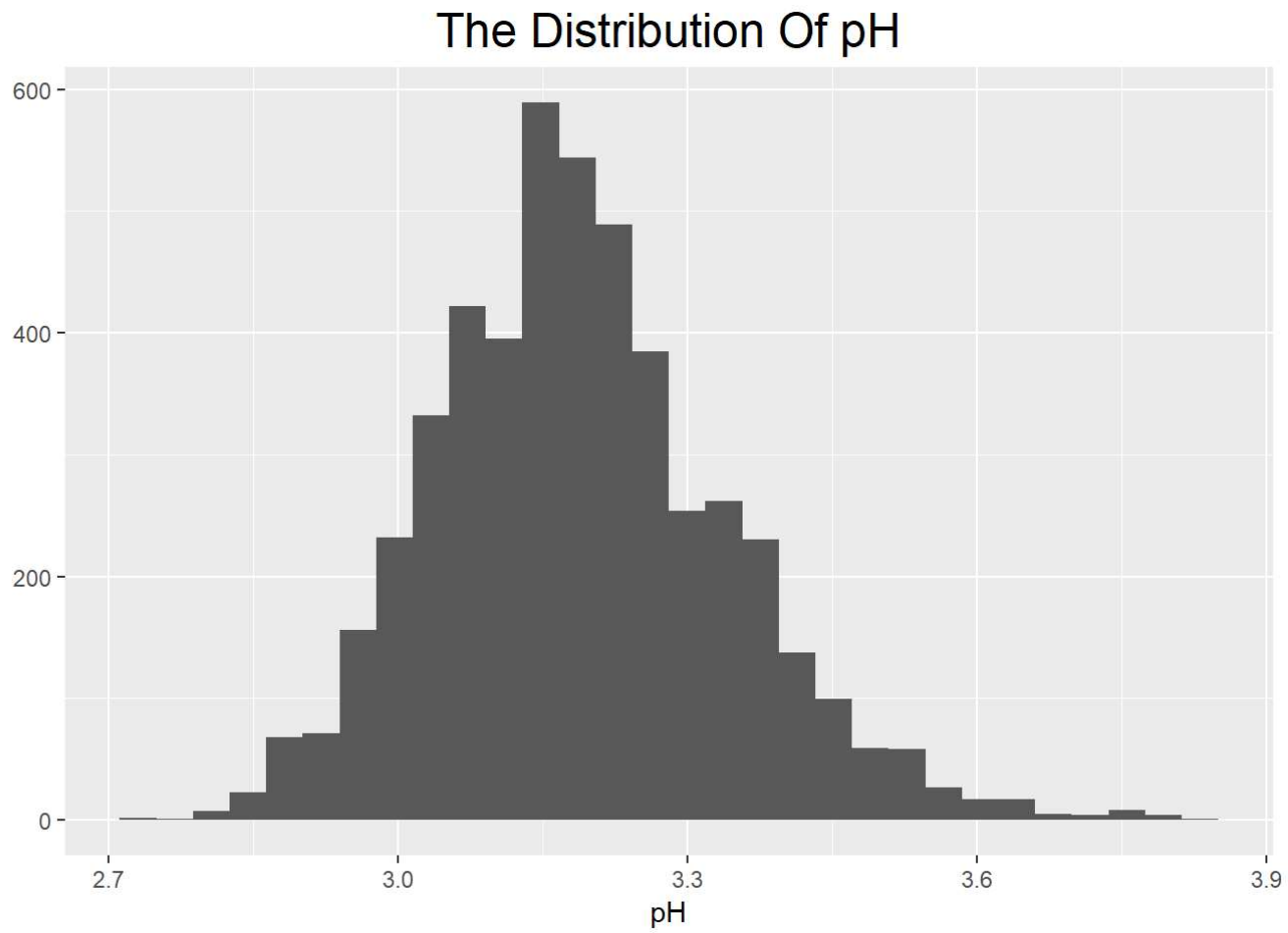
Univariate Plots Section




```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

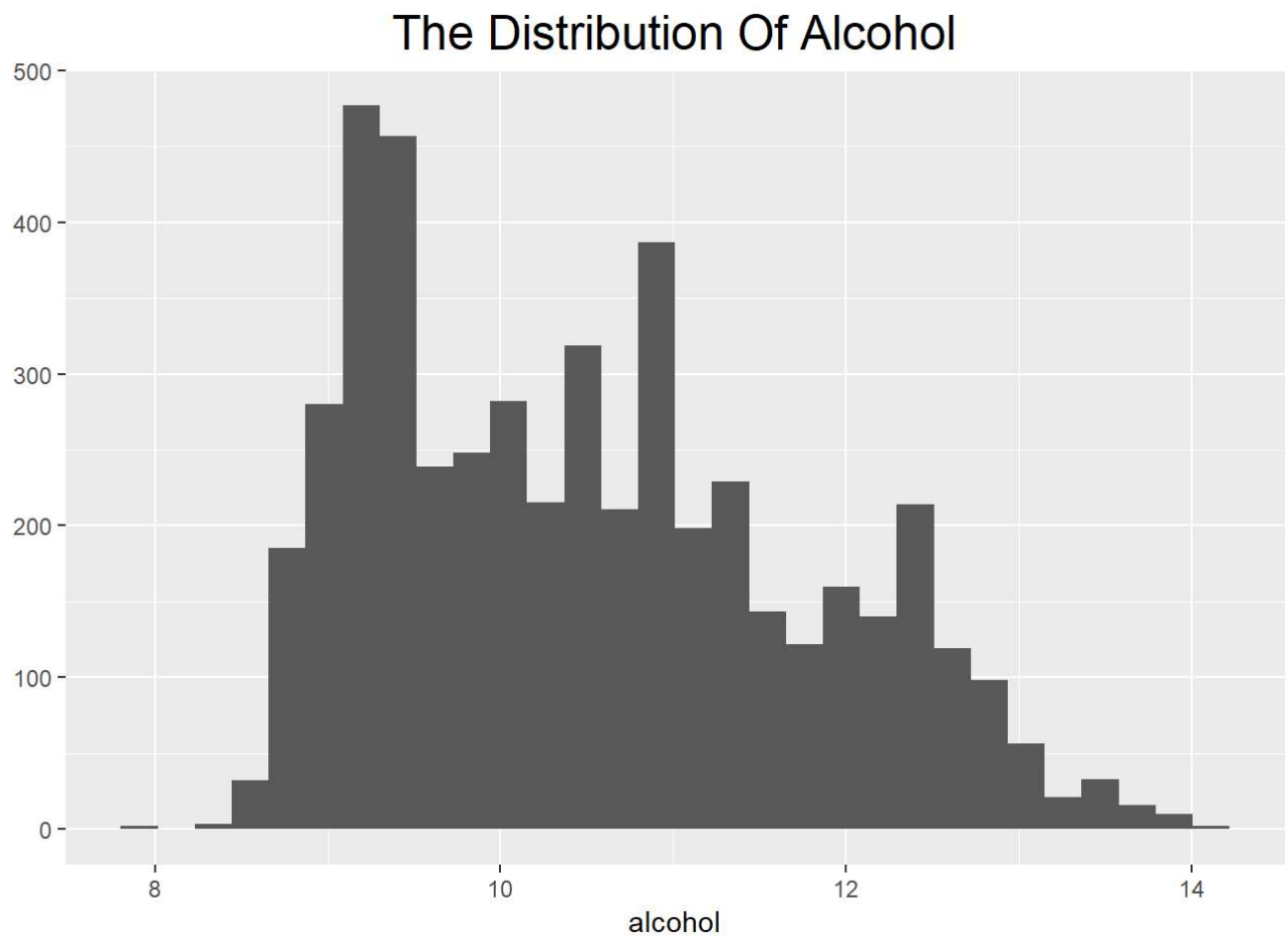
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000
```

The distribution of white wine quality is more frequently in the middle.



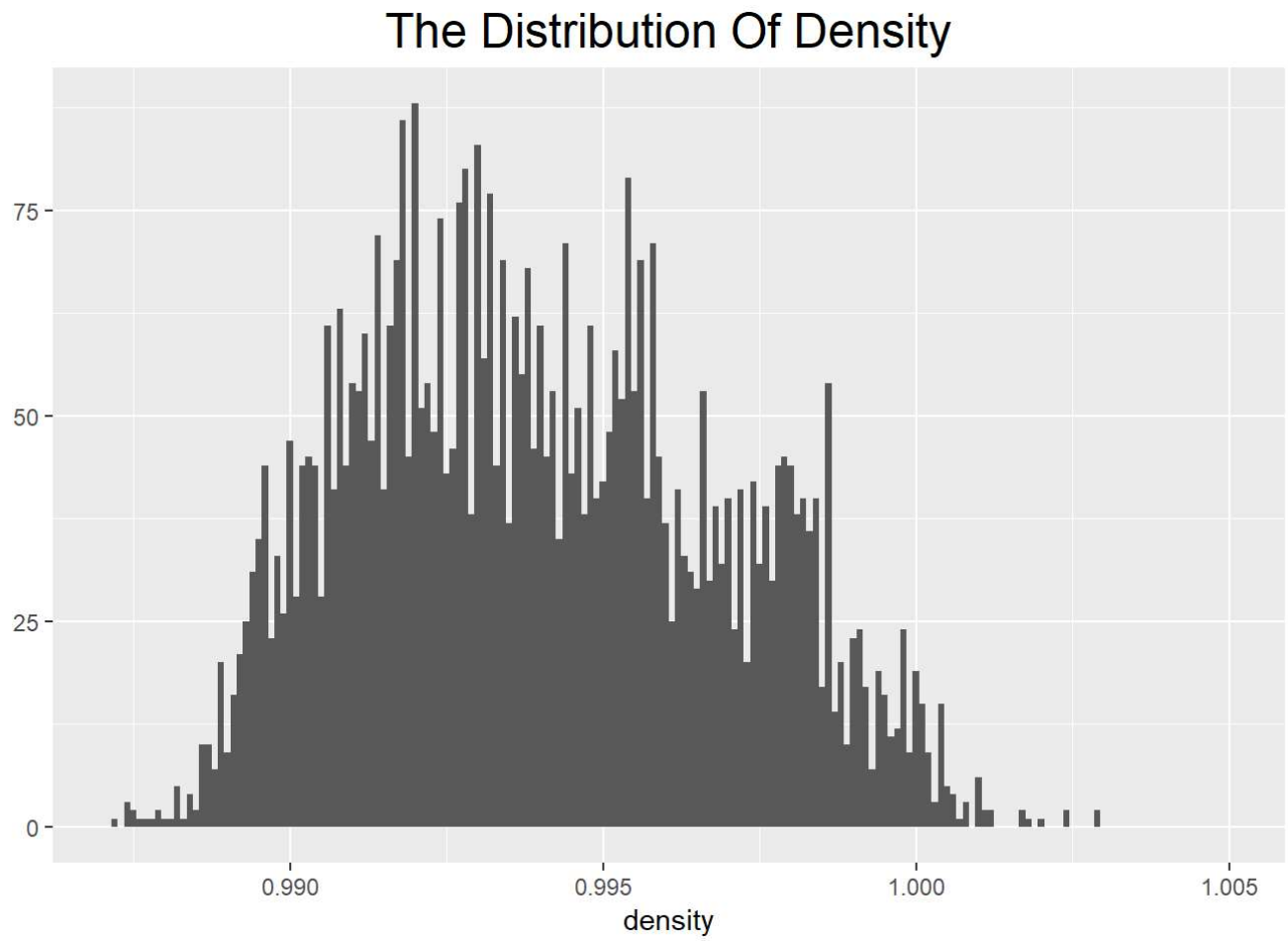

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.720   3.090   3.180   3.188   3.280   3.820
```

The distribution of pH is also more frequently in the middle, the mean value of pH is 3.188, the median value of pH is 3.180.



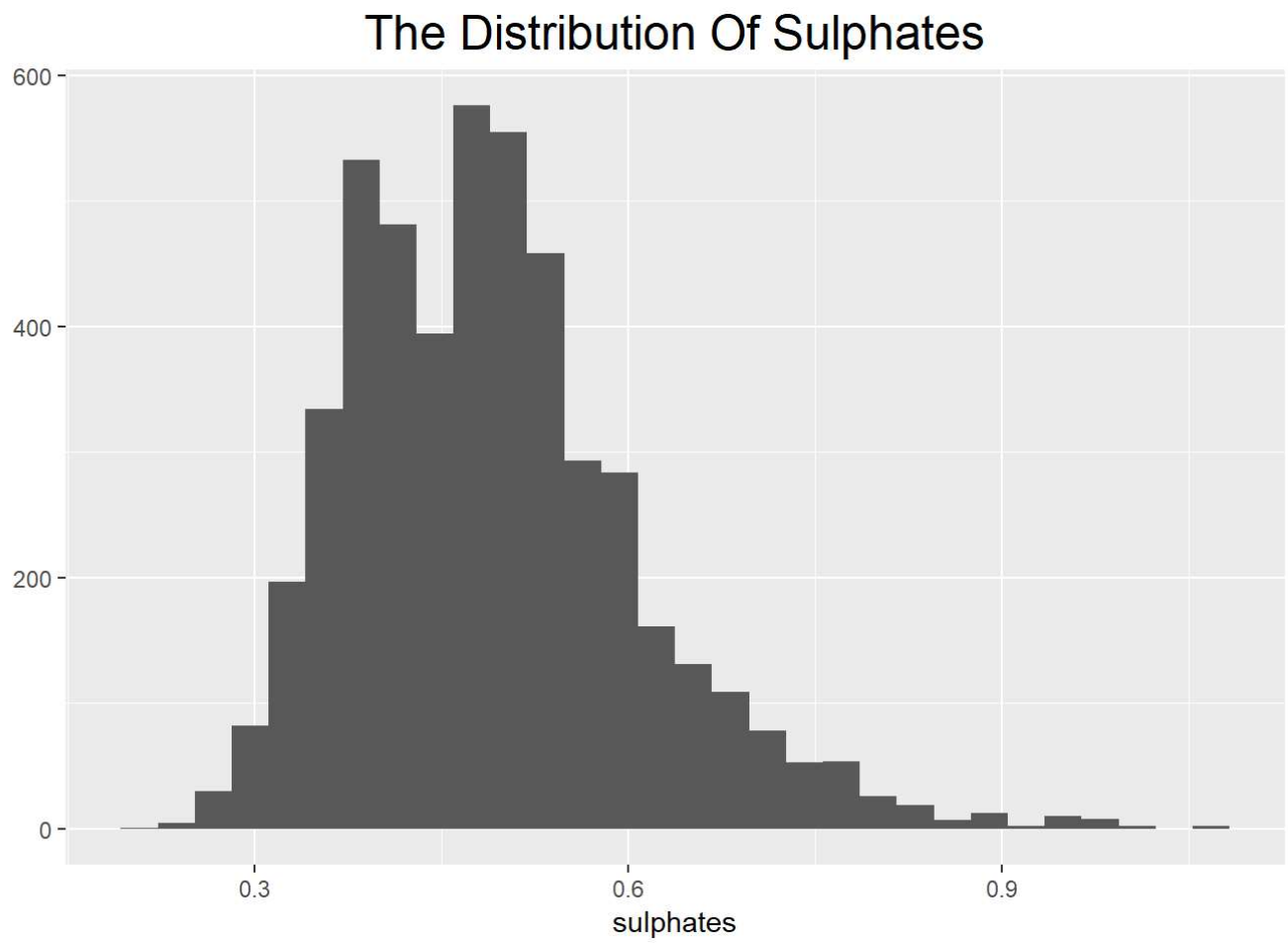

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

More white wine has an alcohol percentage between 9% to 12%, mean 10.51% and median 10.40%.



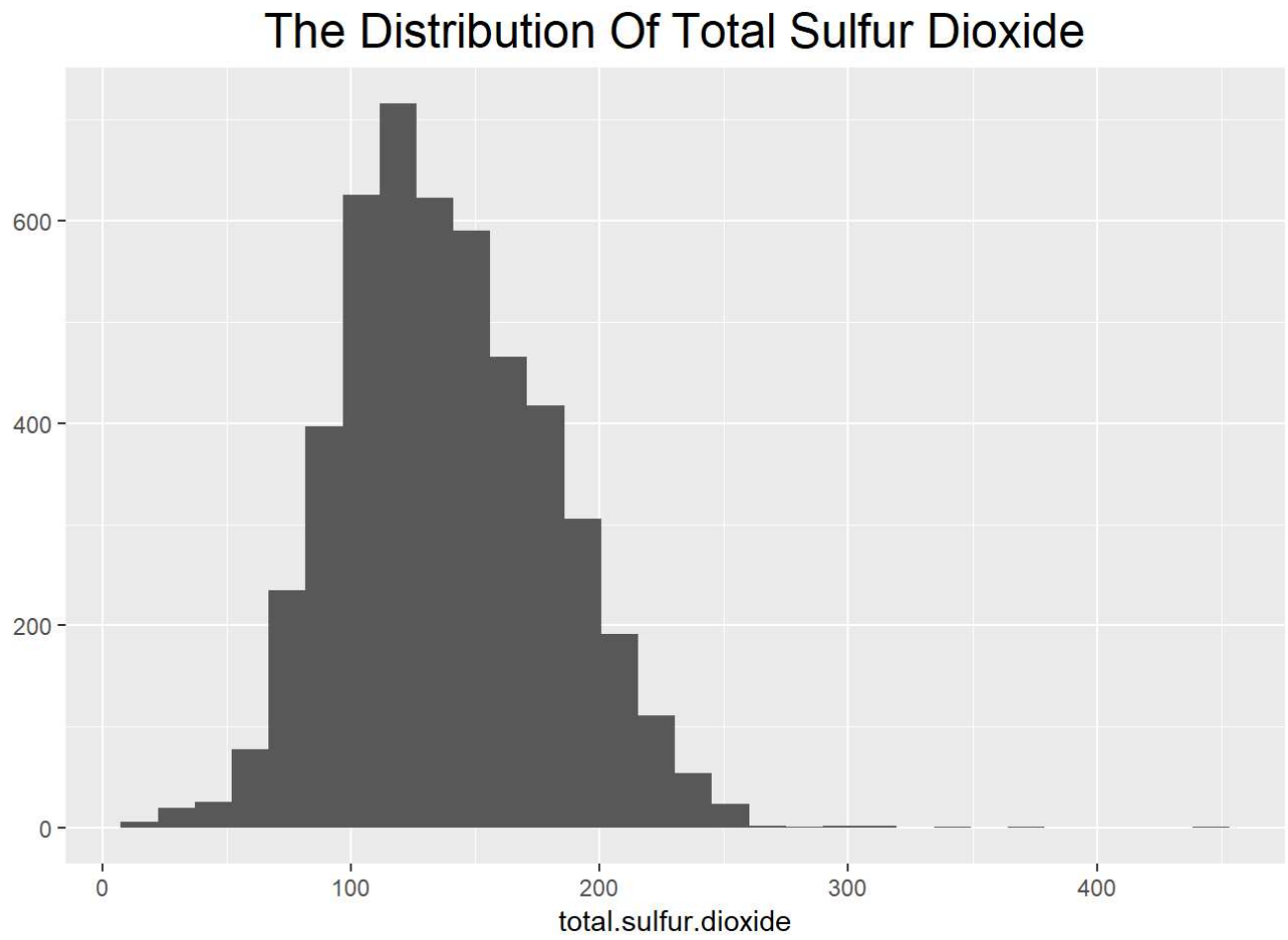

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

Density has a long tail, adjust the X axis range, we can find that most of the density is distributed between 0.990 and 1.000.



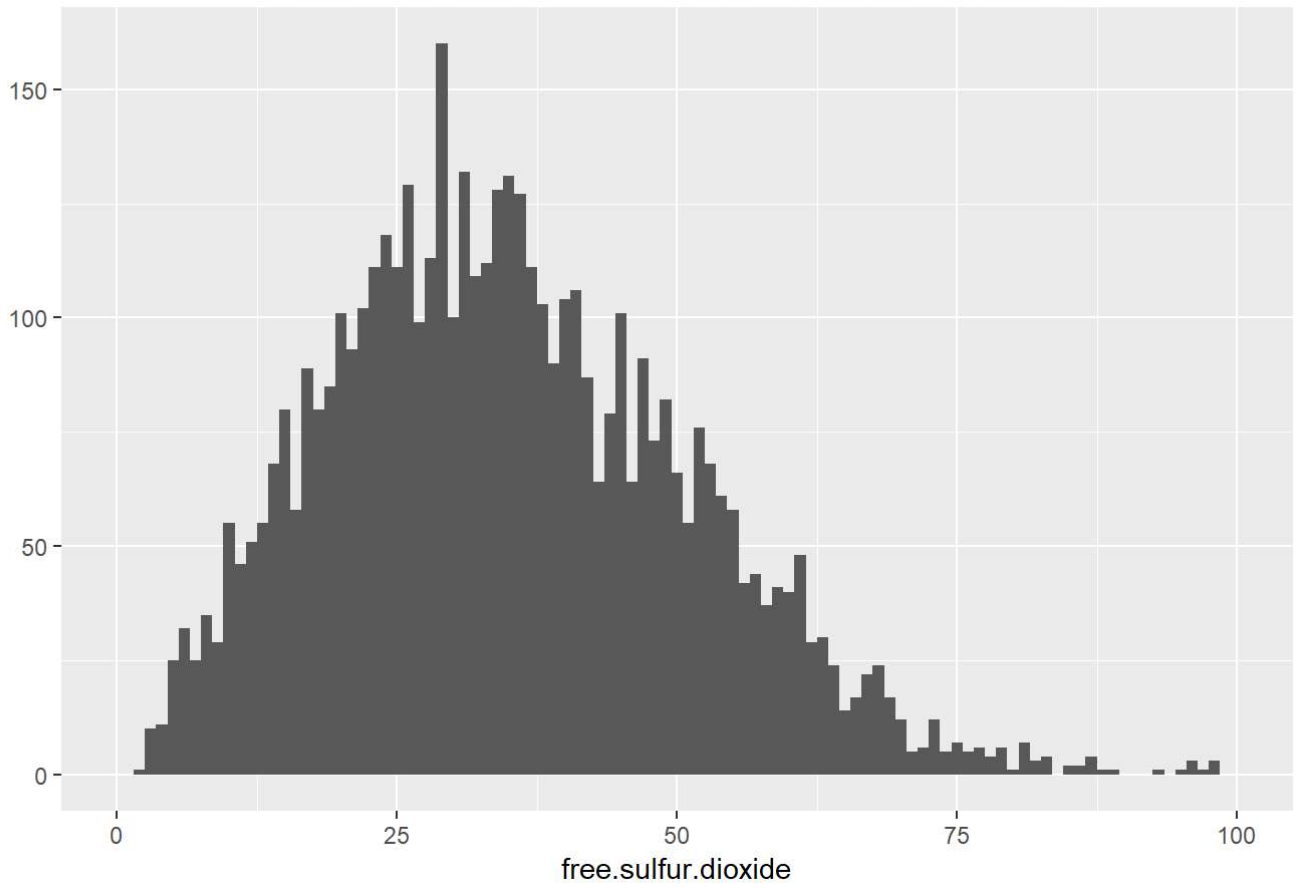

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2200  0.4100  0.4700  0.4898  0.5500  1.0800
```

The sulphates of most white wine is between 0.3 and 0.7. The mean sulphates is 0.4898 and the median is 0.4700.



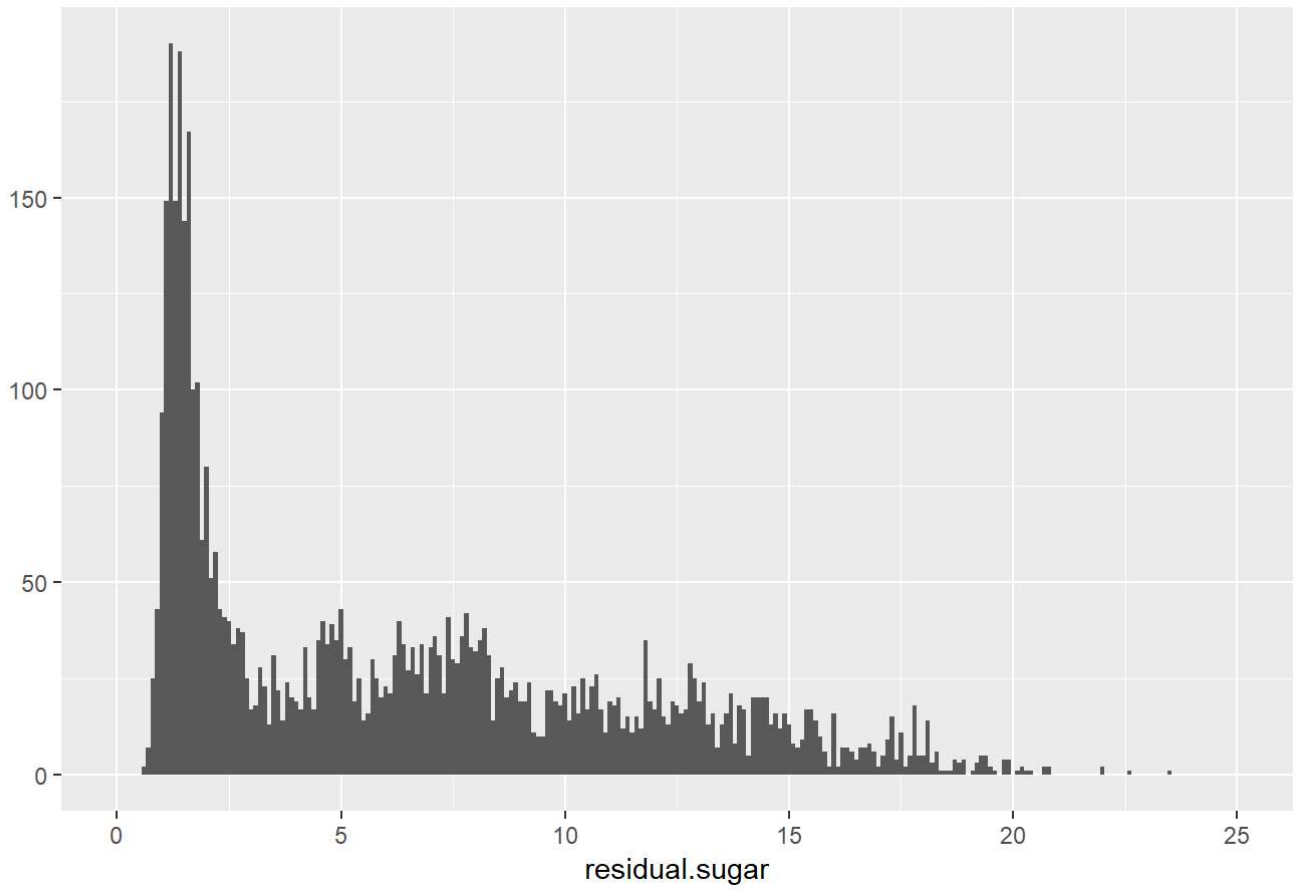

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   108.0   134.0   138.4   167.0   440.0
```

Total sulfur dioxide has a long tail, most is distributed between 50 and 250, the mean total sulfur dioxide is 138.4 and the median is 134.0.



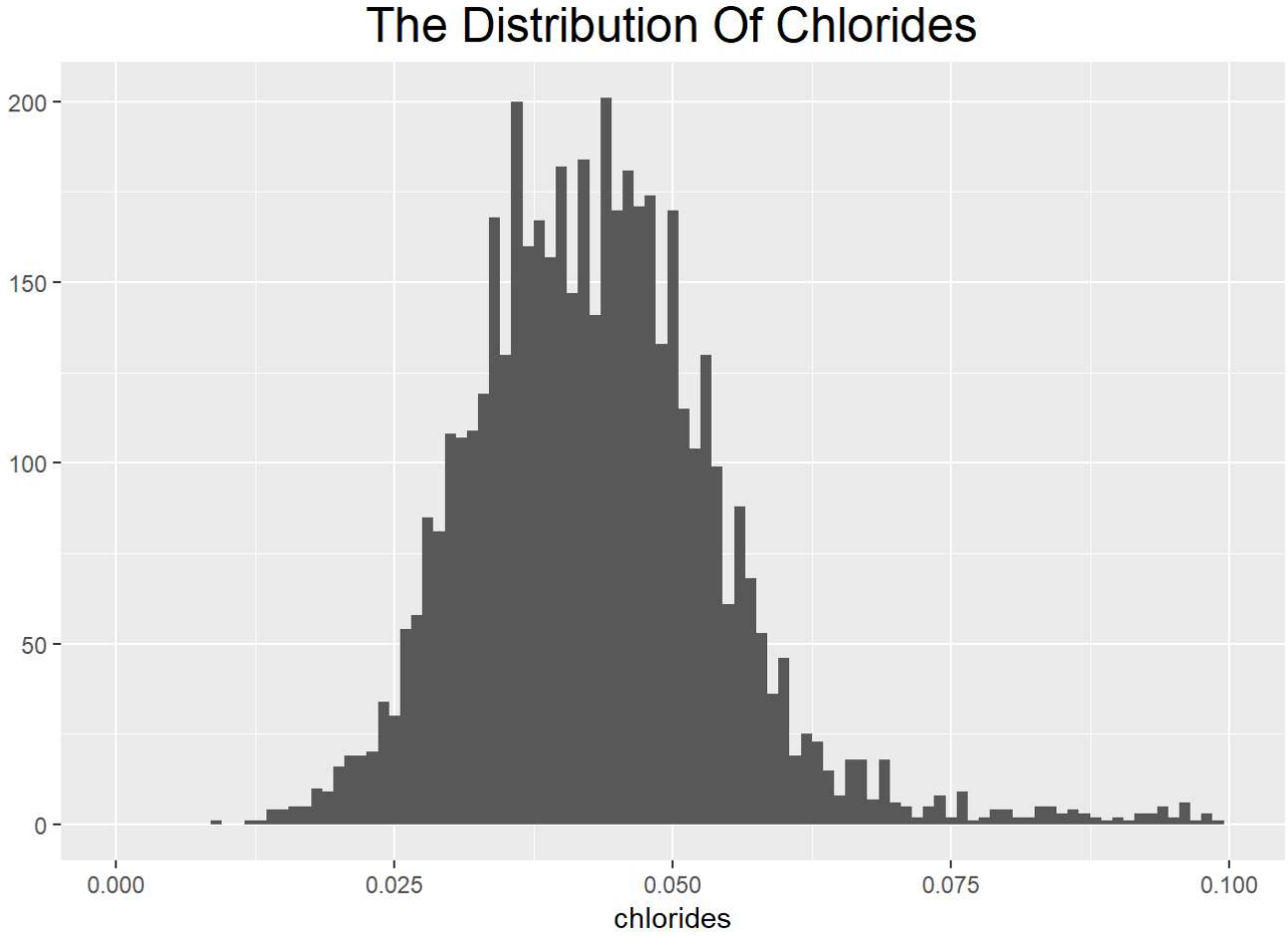

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   23.00   34.00   35.31   46.00  289.00
```

Free sulfur dioxide has a long tail, adjust the X axis range, we can find that most is distributed between 0 and 75, the mean free sulfur dioxide is 35.31 and the median is 34.00.



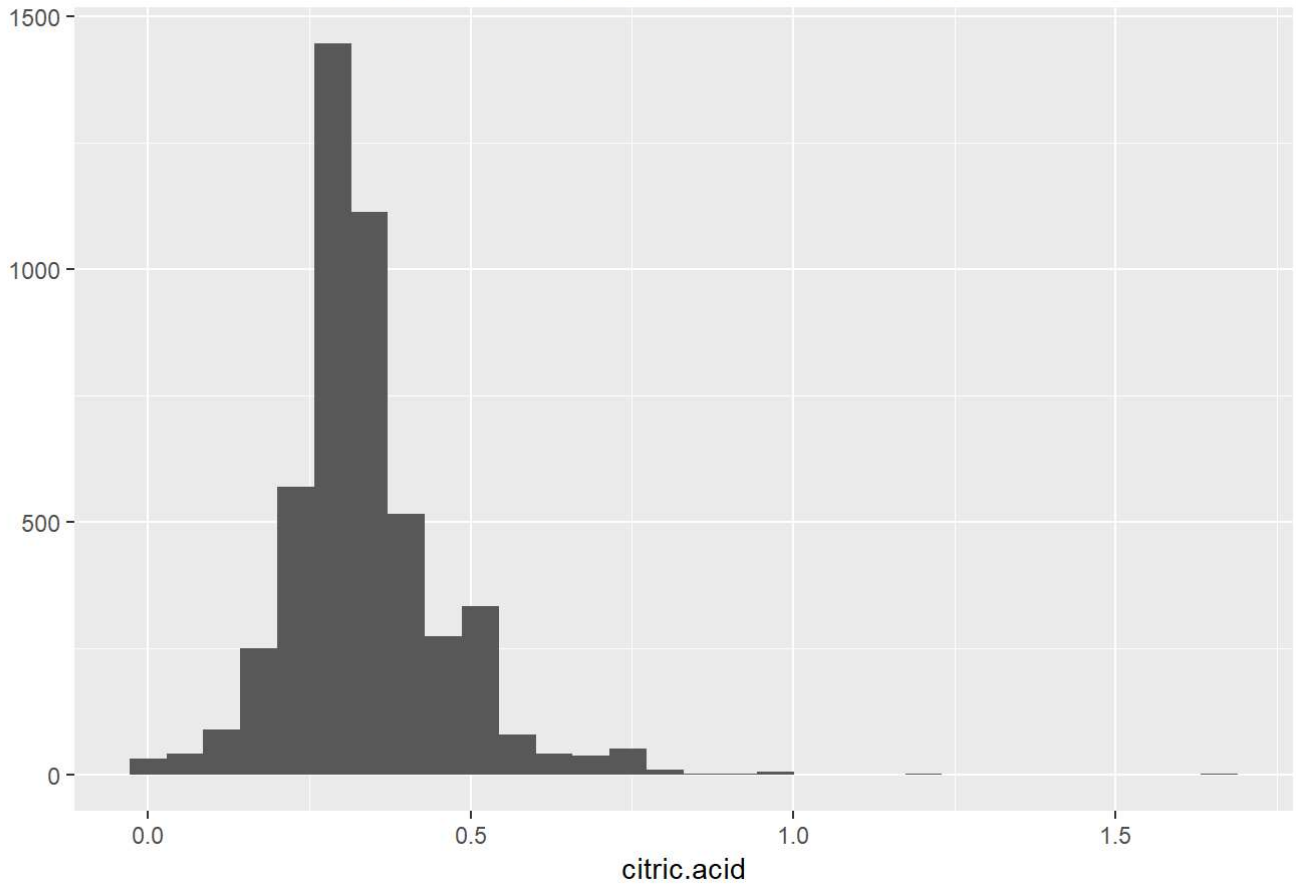

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

Residual sugar has a long tail, most is distributed between 0 and 20. Transformed the long tail data, the transformed peaking residual sugar peaking around 2. The mean is 6.391 and the median is 5.200.



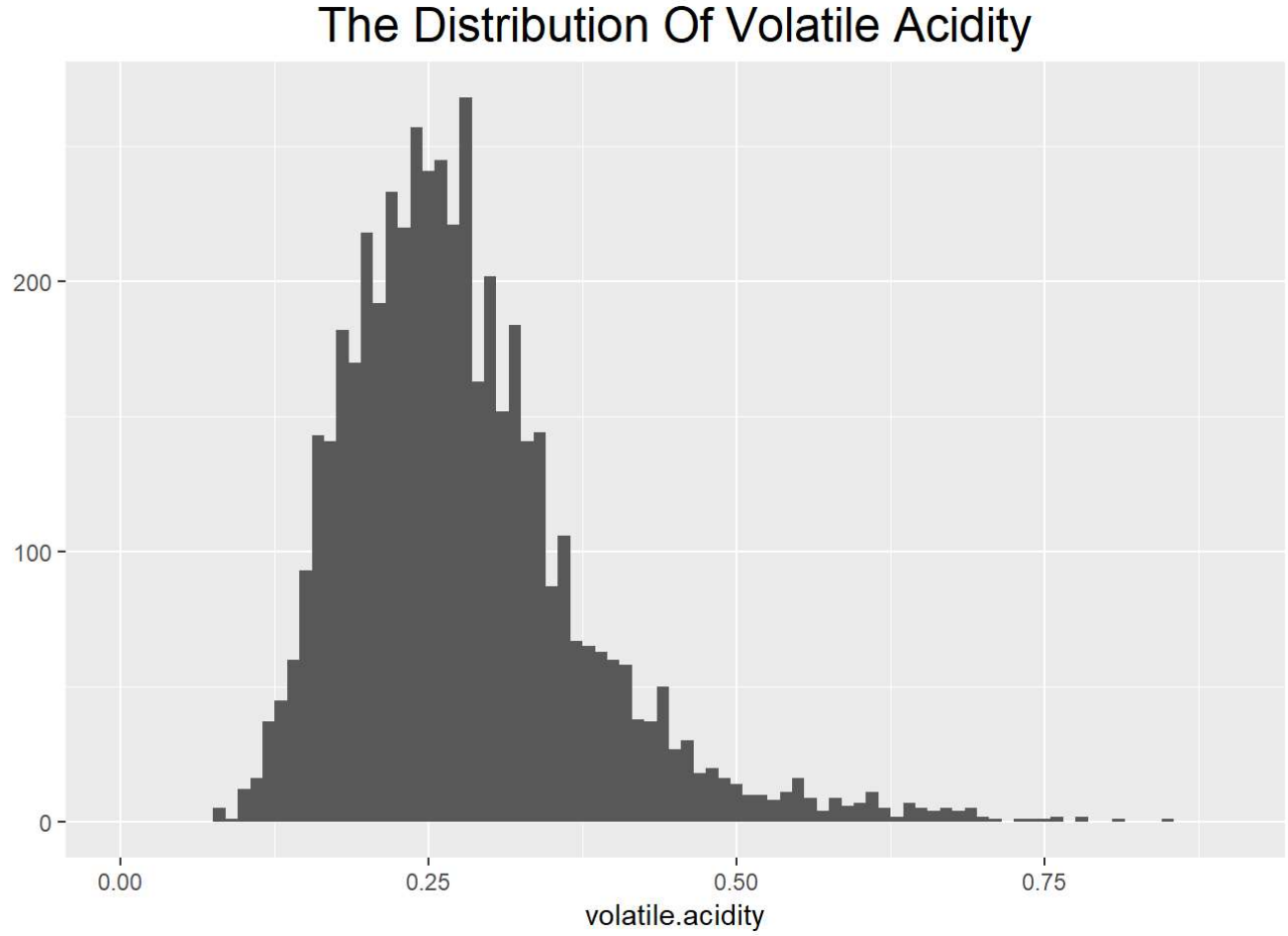

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00900 0.03600 0.04300 0.04577 0.05000 0.34600
```

Chlorides has a long tail. Adjust the X range and binwidth, most is distributed between 0.025 and 0.075. The mean is 0.04577 and the median is 0.043.



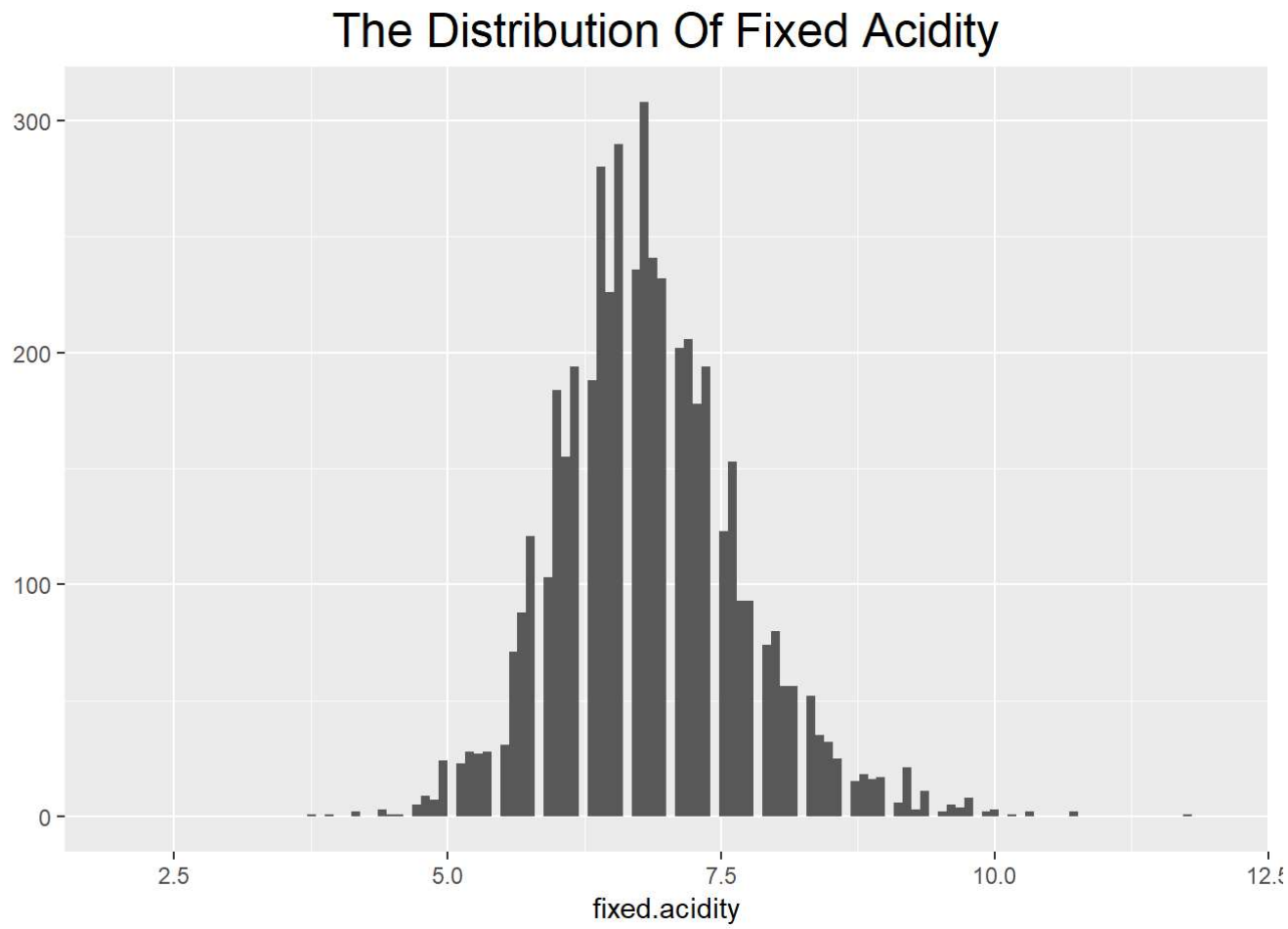

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

Citric acid has a long tail. The mean is 0.3342 and the median is 0.3200.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

Volatile acidity has a long tail. Adjust the X range and binwidth, most is distributed between 0.1 and 0.5. The mean is 0.2782 and the median is 0.26.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

Fixed acidity has a long tail. Adjust the X range and binwidth, most is distributed between 5 and 10. The mean is 6.855 and the median is 6.800.

Univariate Analysis

What is the structure of your dataset?

There are 4,898 obeservations in the dataset with 12 features(fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality. The description of each feature is as follows.)

1. fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2. volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3. citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines

4. residual sugar: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5. chlorides: the amount of salt in the wine

6. free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7. total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8. density: the density of water is close to that of water depending on the percent alcohol and sugar content

9. pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3.4 on the pH scale

10. sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

11. alcohol: the percent alcohol content of the wine

Output variable (based on sensory data):

12. quality (score between 0 and 10)

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

The main features in the data set are quality, pH, alcohol and density. I'd like to explore which features are best for predicting the quality of white wine. I suspect pH and alcohol can effect the quality, I'll explore in the next section.

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

Other features like density, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides etc… may also effect quality.

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

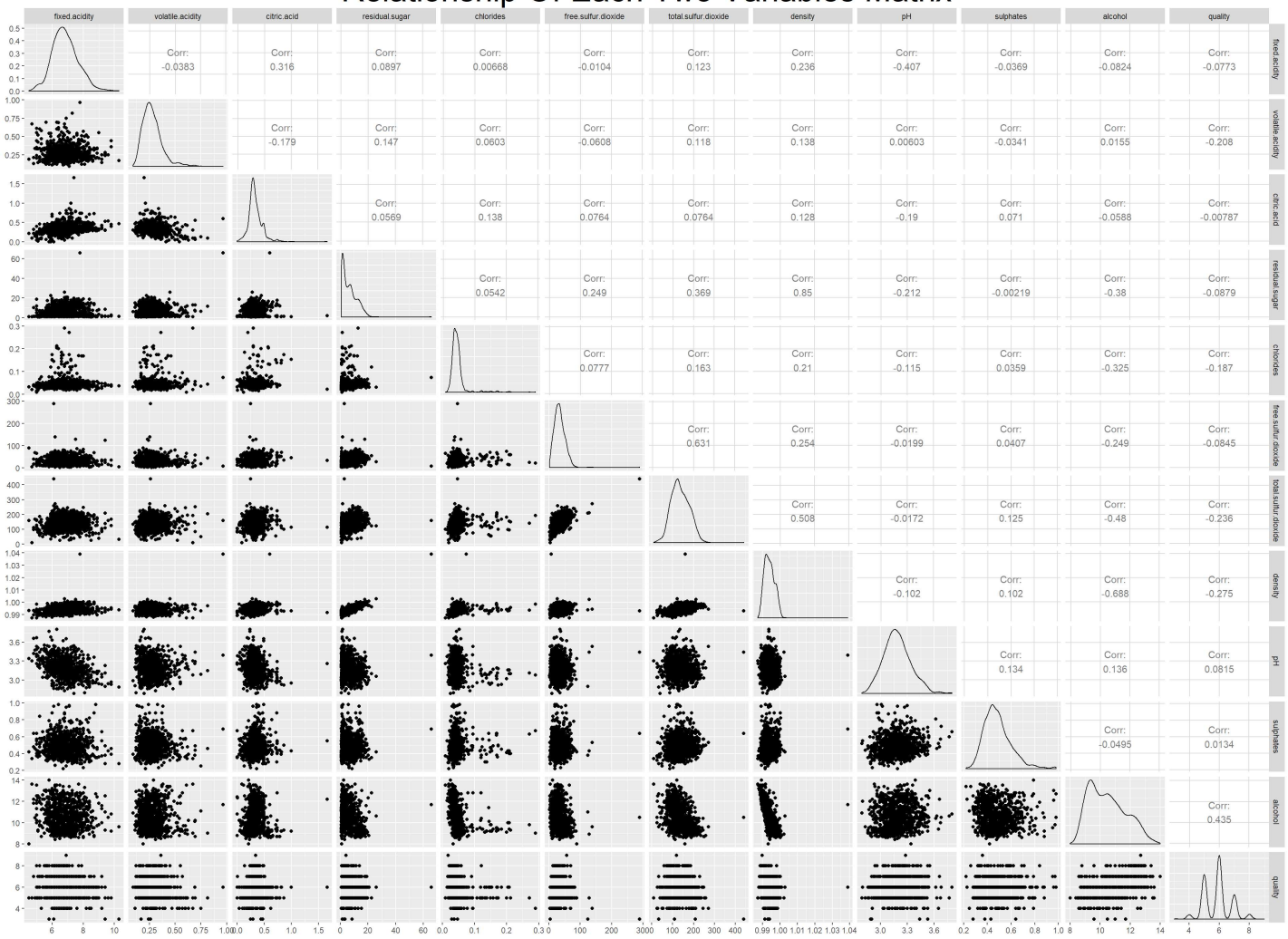
No, I don't creat any new variables.

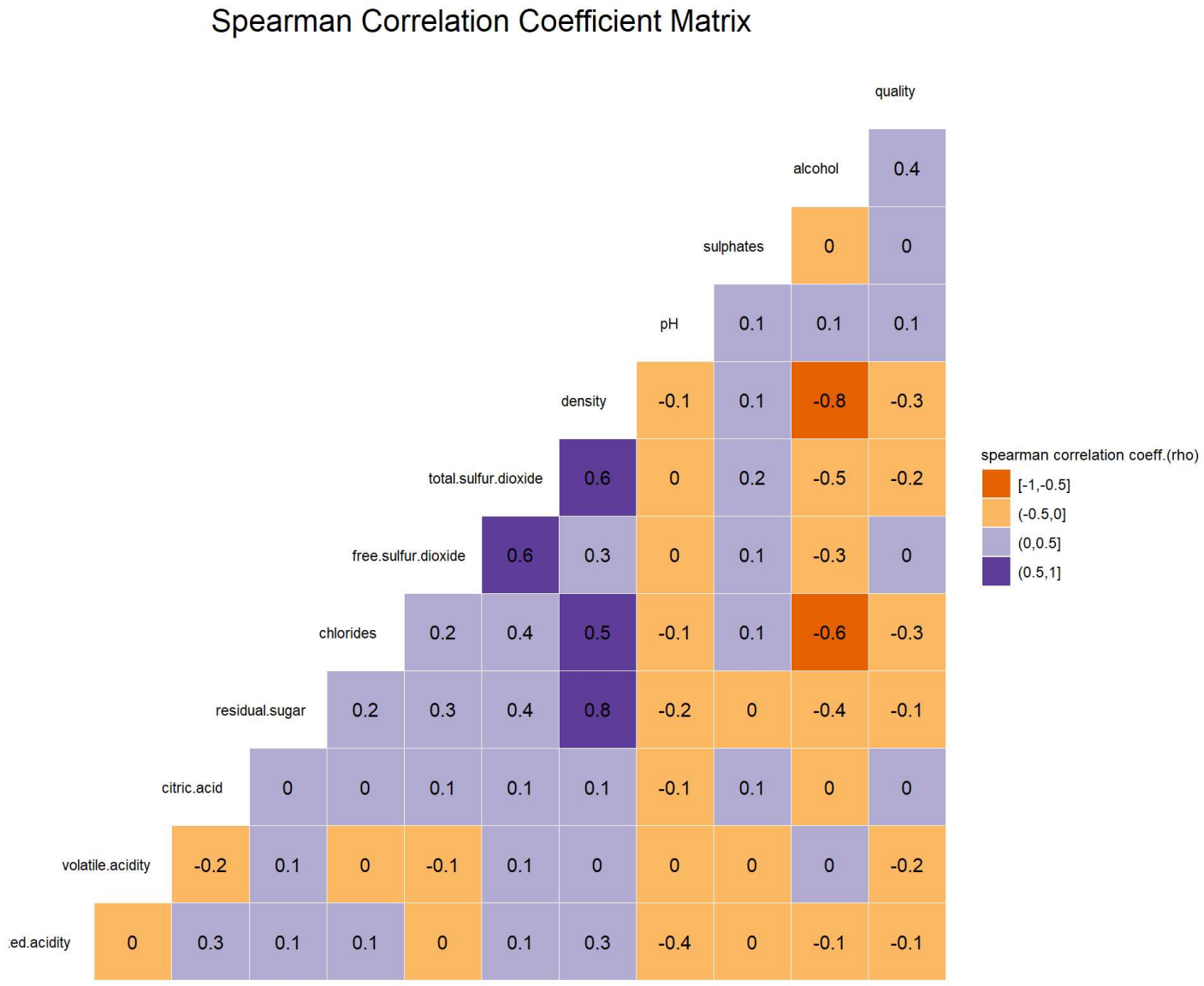
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

No.

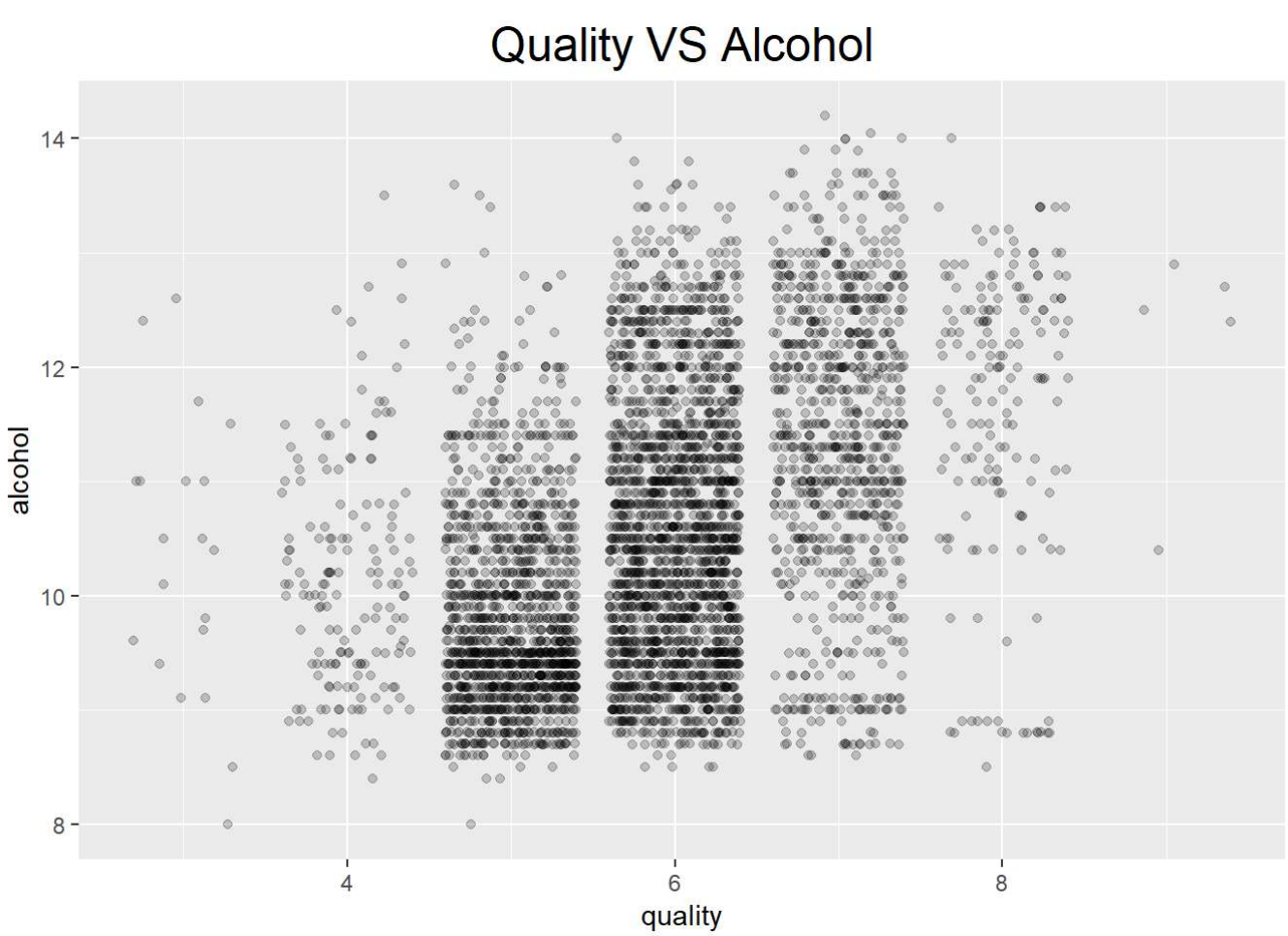
Bivariate Plots Section

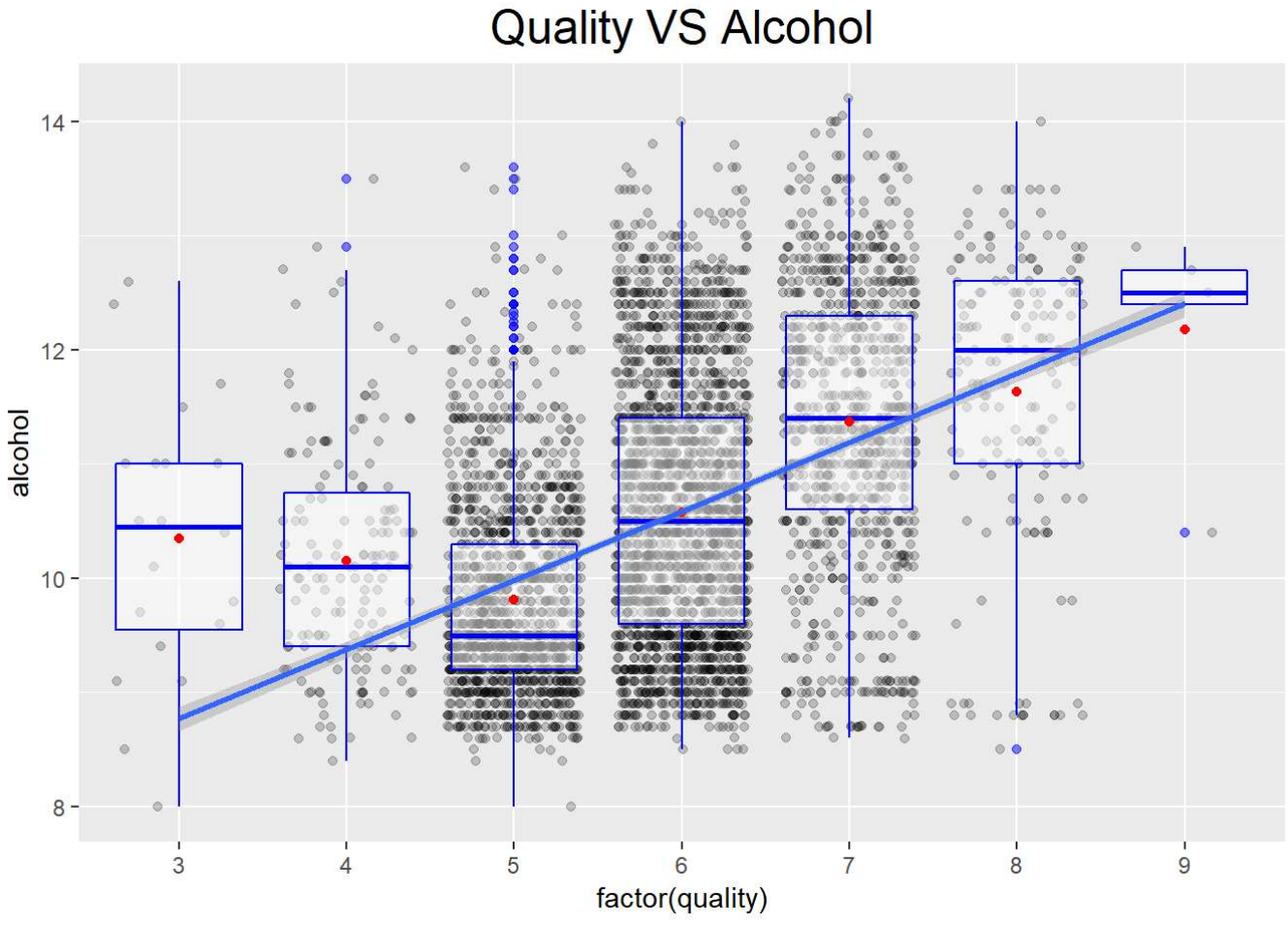
Relationship Of Each Two Variables Matrix





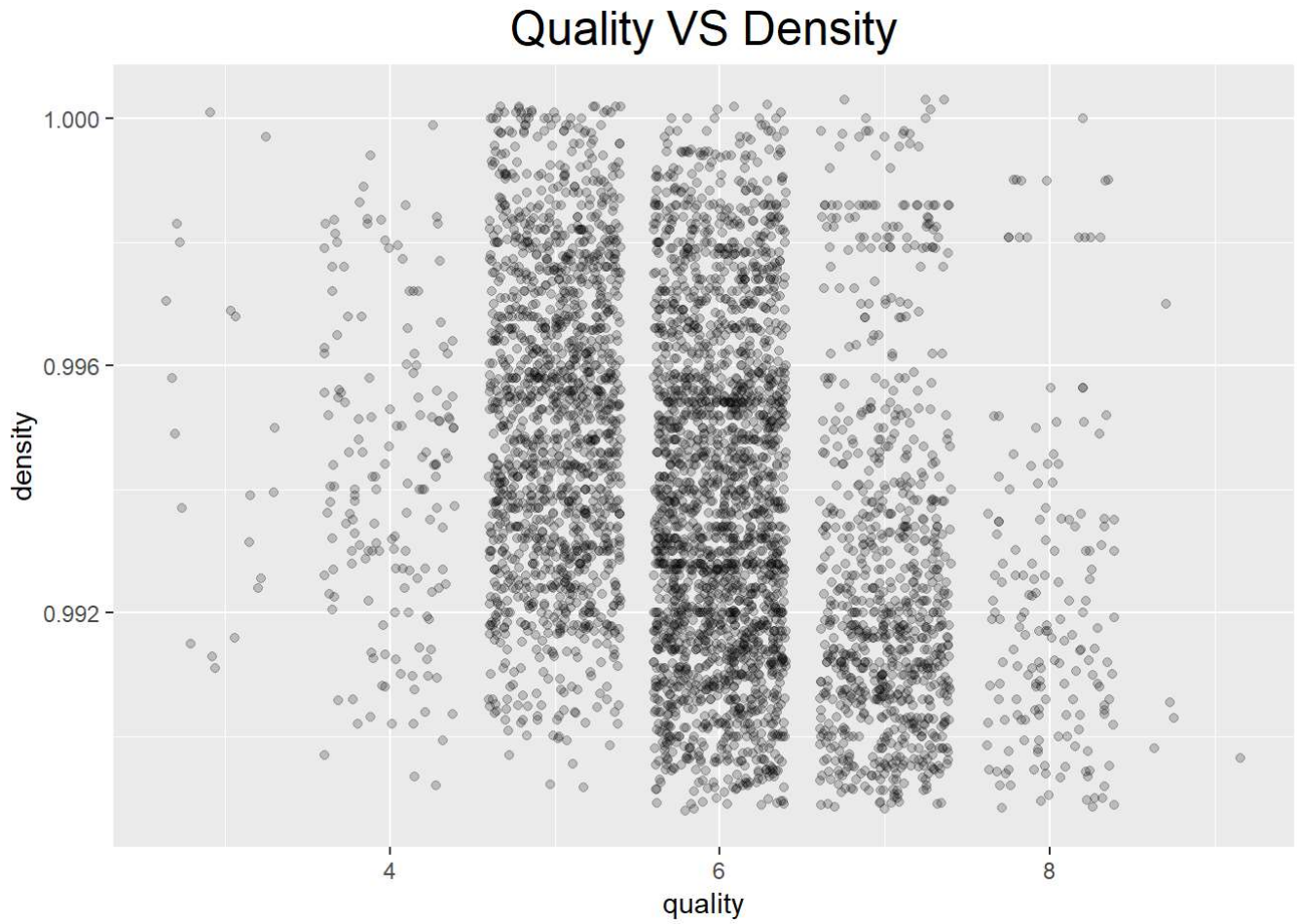
Use and ggparis() and ggcorr(), we can find that fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide and density have negative influence on quality, while free.sulfur.dioxide, pH, sulphates and alcohol have positive influence on quality. It seems that density and alcohol have moderate correlations with quality, there isn't any variable has strong correlation with quality. Density has strong correlation with residual sugar and alcohol, while moderately correlated with free sulfur dioxide and total sulfur dioxide. Alcohol has strong correlation with density, while moderately correlated with residual sugar, chlorides and total sulfur dioxide. Next, I will explore these variables, alcohol, denssity, residual sugar, chlorides, free sulfur dioxide and total sulfur dioxide.

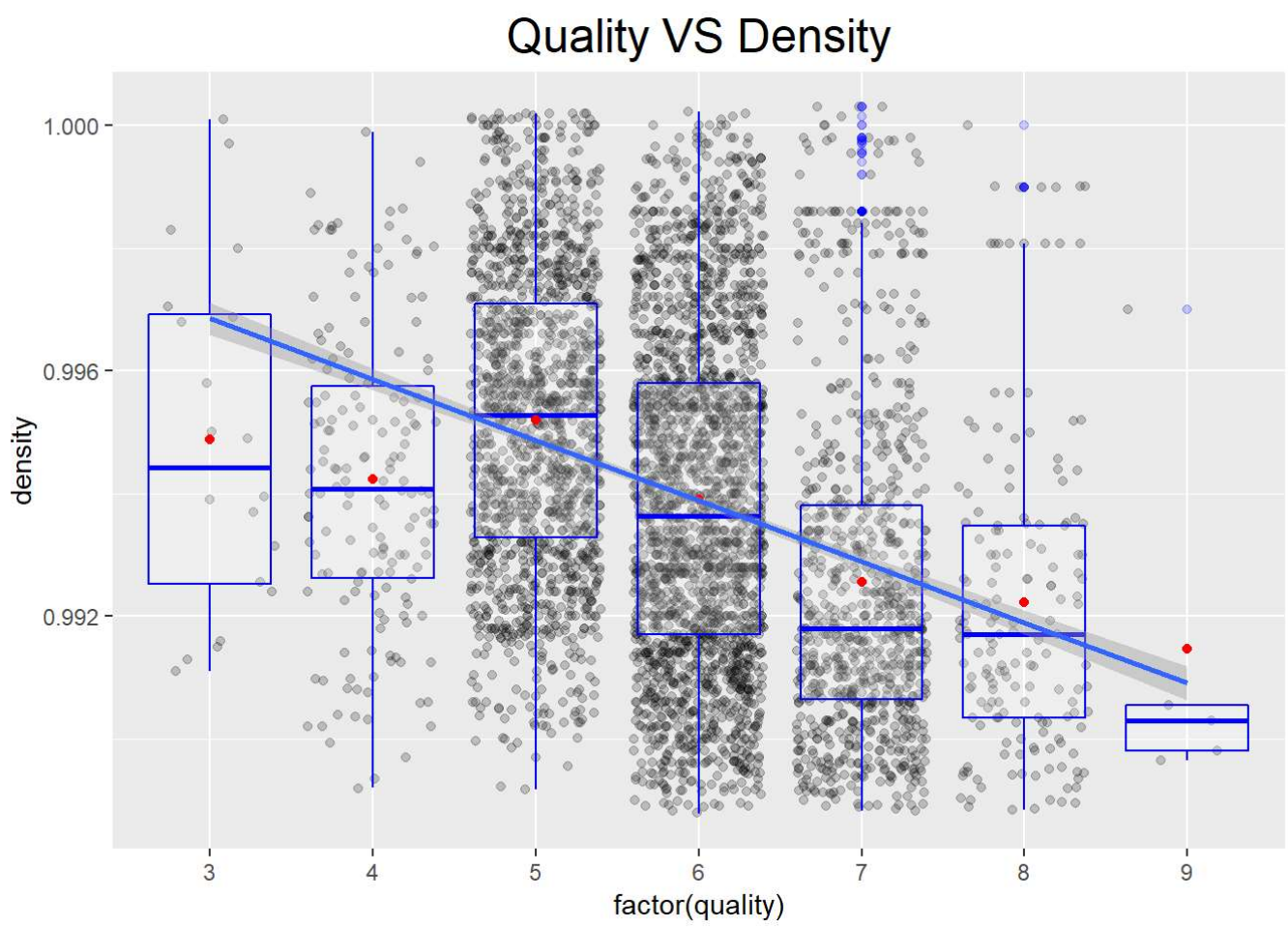

Quality VS Alcohol
alcohol
14
12
10
8
4
6
8
quality


Quality VS Alcohol
alcohol
14
12
10
8
3
4
5
6
7
8
9
factor(quality)

```
##
##  Pearson's product-moment correlation
##
## data:  wq$quality and wq$alcohol
## t = 33.858, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4126015 0.4579941
## sample estimates:
##       cor
## 0.4355747
```

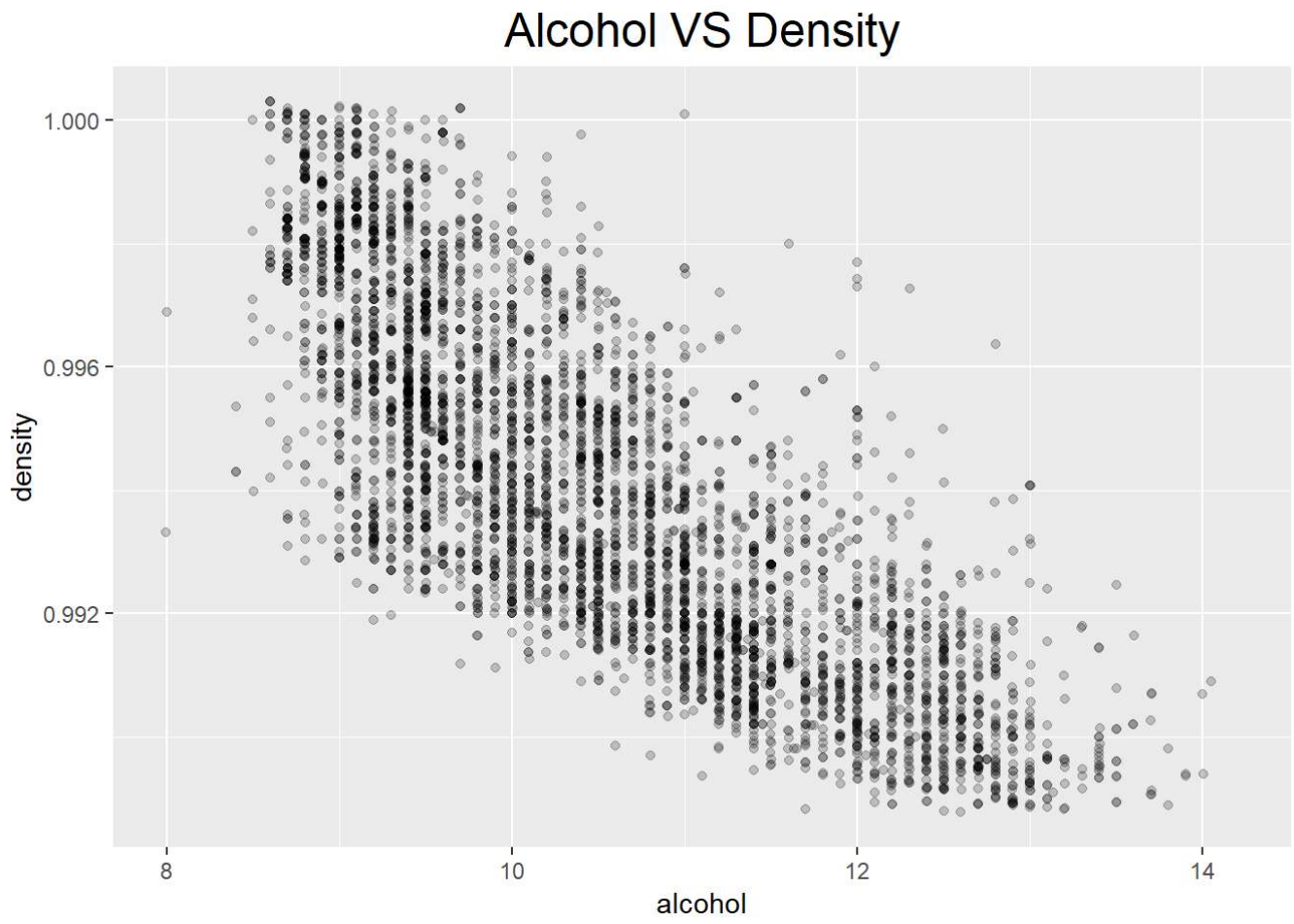
With the increase of alcohol, quality increases.

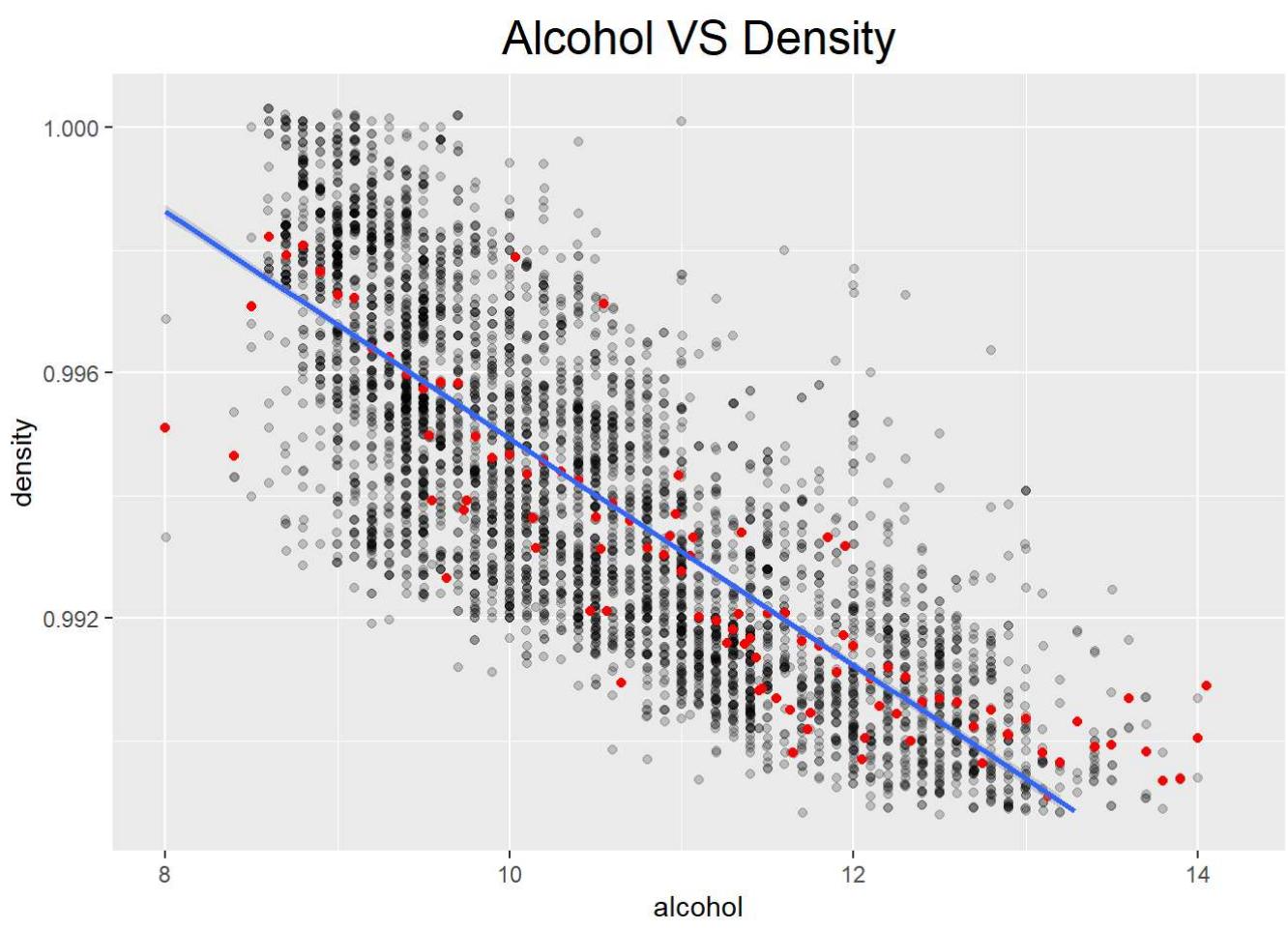

Quality VS Density
1.000
0.996
0.992
density
4
6
8
quality


Quality VS Density
1.000
0.996
0.992
density
3
4
5
6
7
8
9
factor(quality)

```
##
##  Pearson's product-moment correlation
##
## data:  wq$quality and wq$density
## t = -22.581, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3322718 -0.2815385
## sample estimates:
##        cor
## -0.3071233
```

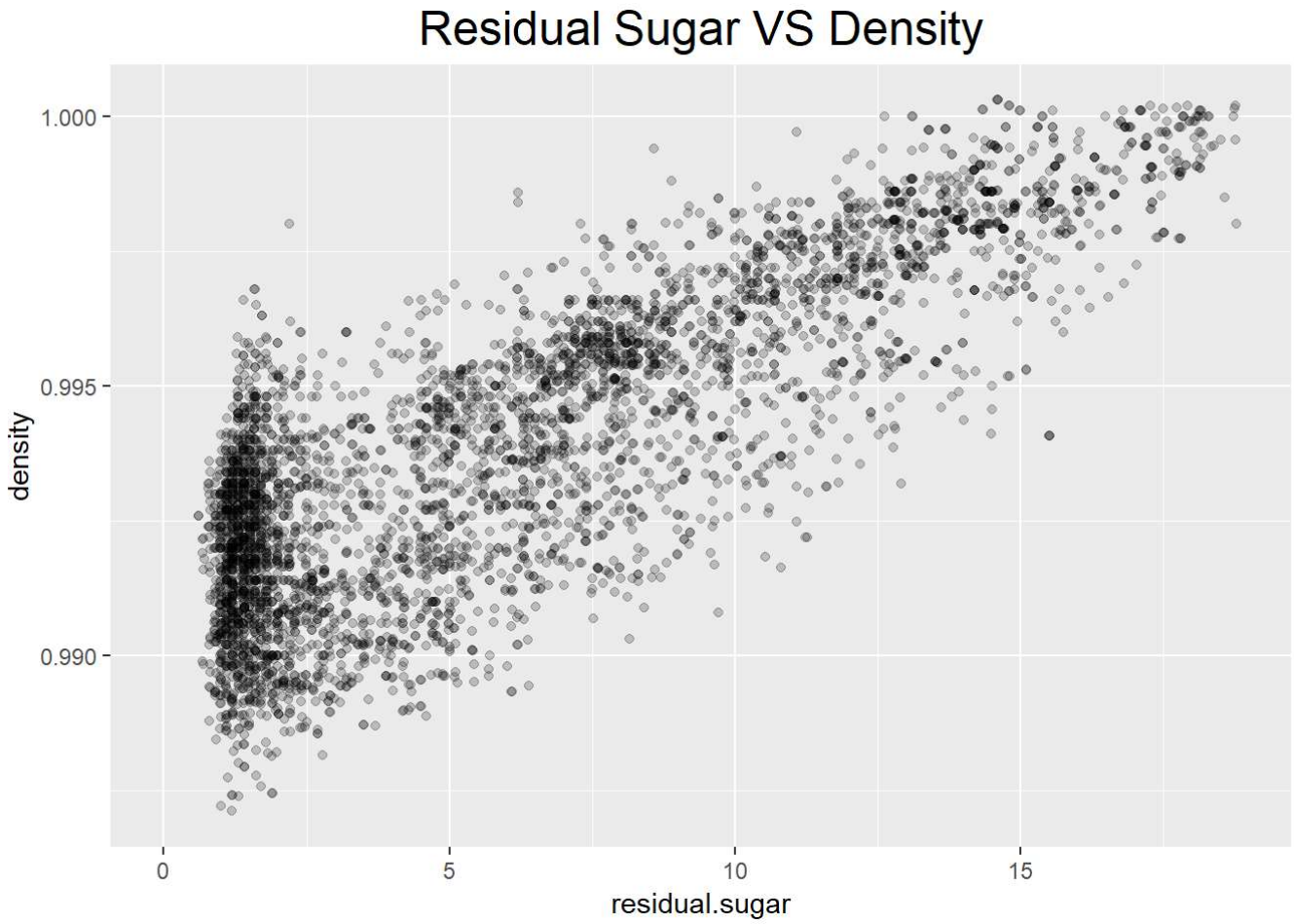
At a certain level, with the density decreasing, quality increases.

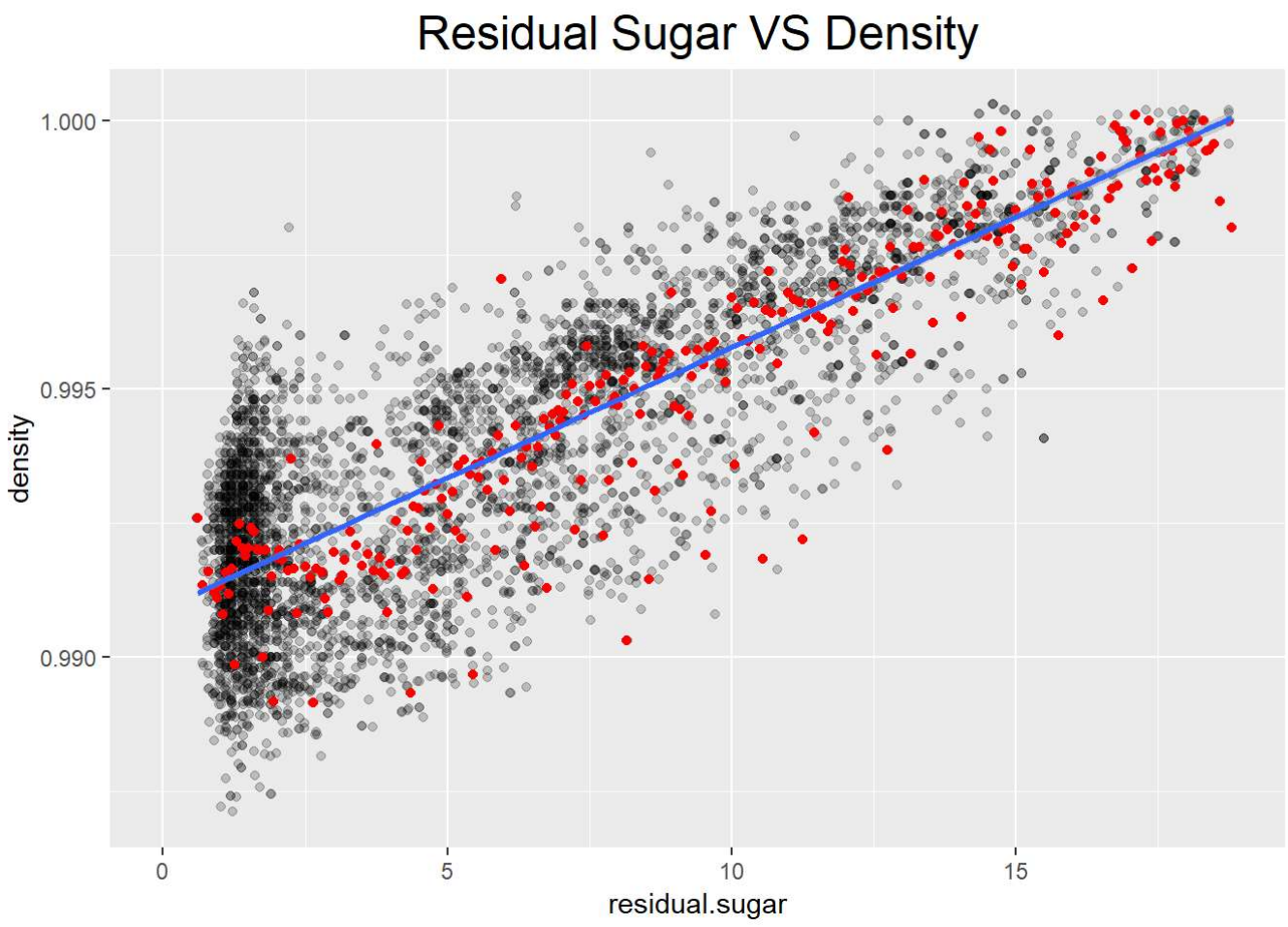

Alcohol VS Density
1.000
0.996
0.992
density
8
10
12
14
alcohol


Alcohol VS Density
1.000
0.996
0.992
density
8
10
12
14
alcohol

```
##
##  Pearson's product-moment correlation
##
## data:  wq$alcohol and wq$density
## t = -87.255, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7908646 -0.7689315
## sample estimates:
##       cor
## -0.7801376
```

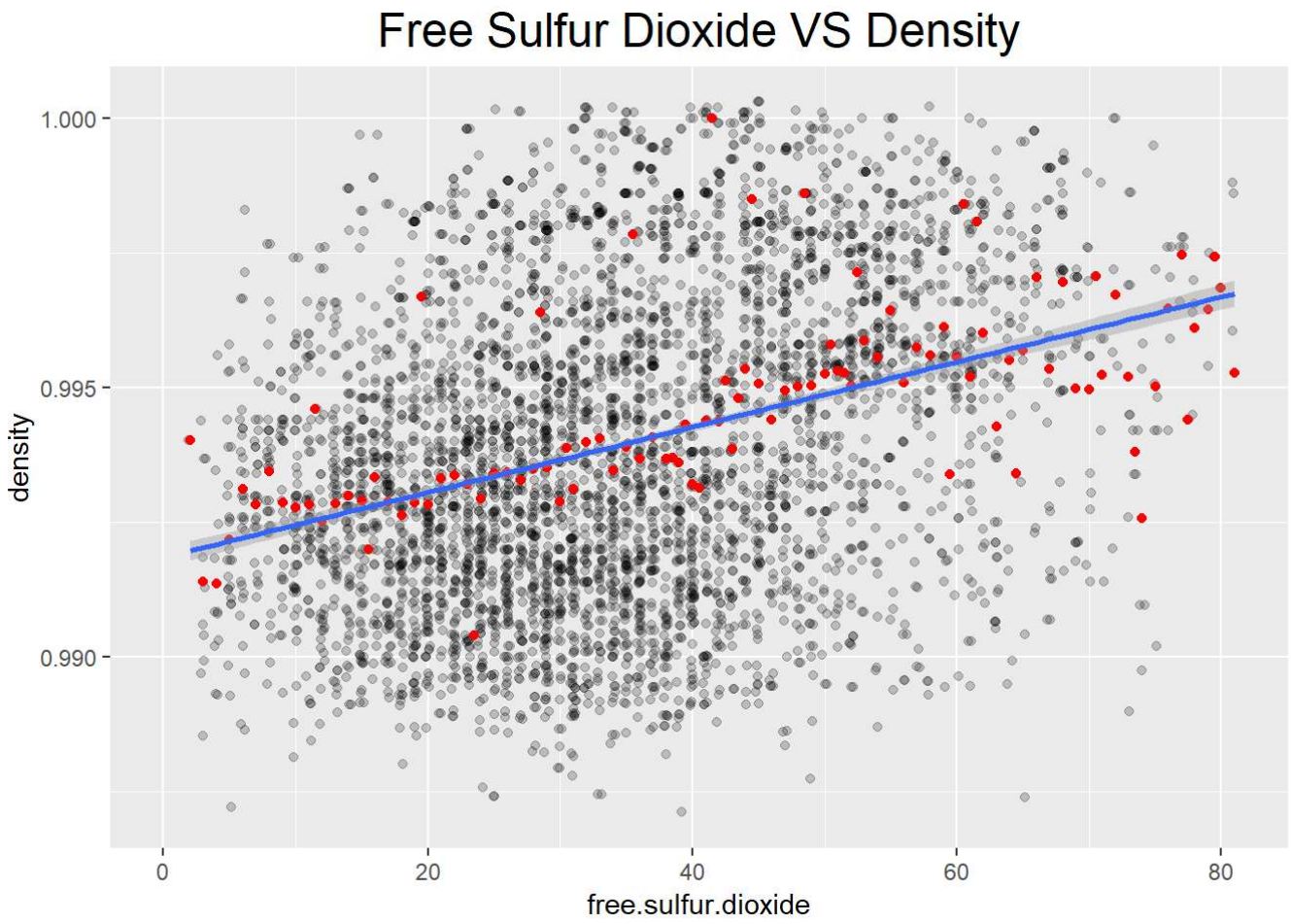
Density has a strong negative correlation with alcohol.


Residual Sugar VS Density
1.000
0.995
0.990
density
0
5
10
15
residual.sugar


Residual Sugar VS Density
1.000
0.995
0.990
density
0
5
10
15
residual.sugar

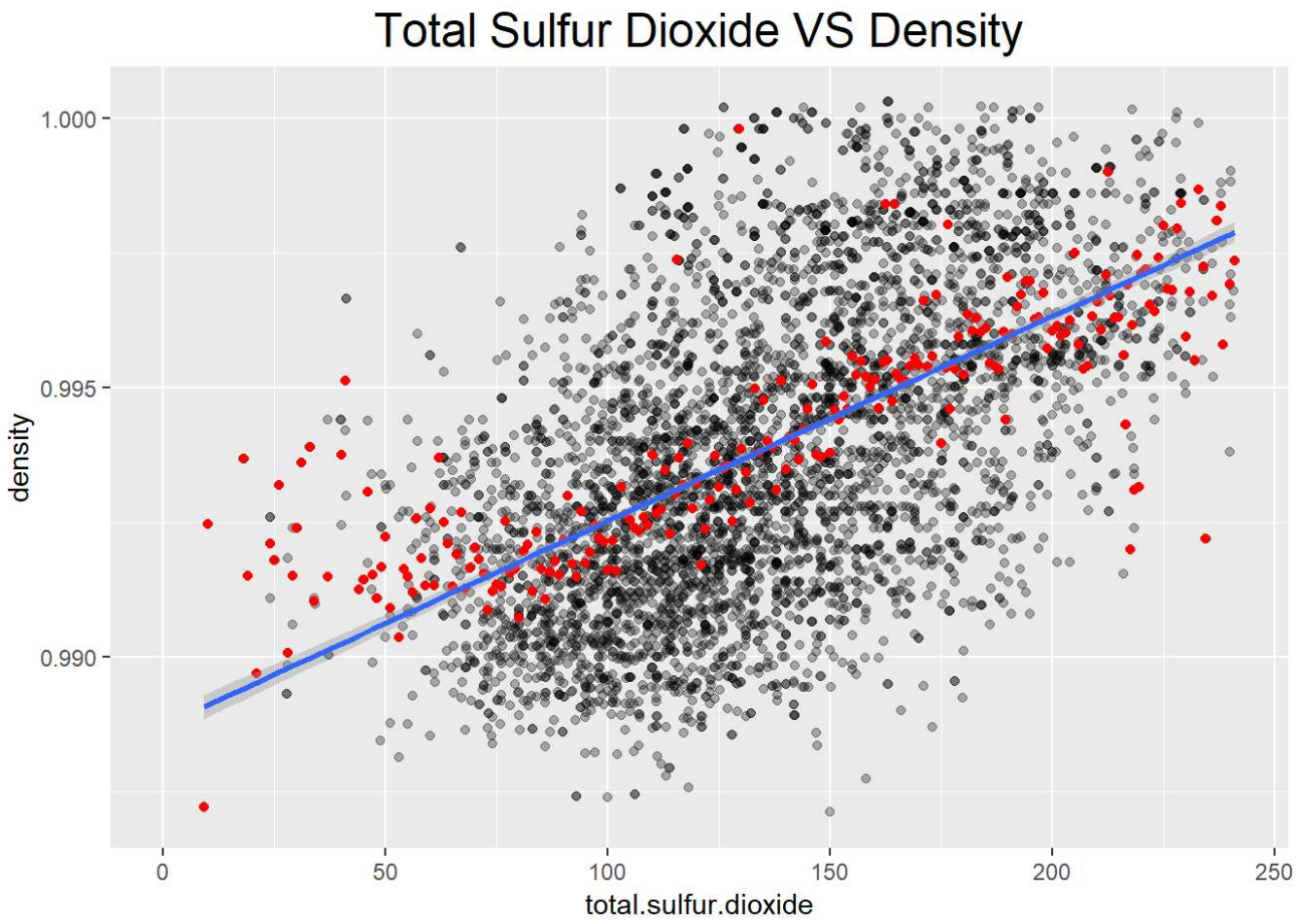
```
##
##  Pearson's product-moment correlation
##
## data:  wq$residual.sugar and wq$density
## t = 107.87, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8304732 0.8470698
## sample estimates:
##       cor
## 0.8389665
```

Density has a strong positive correlation with residual sugar, more residual sugar, higer density.



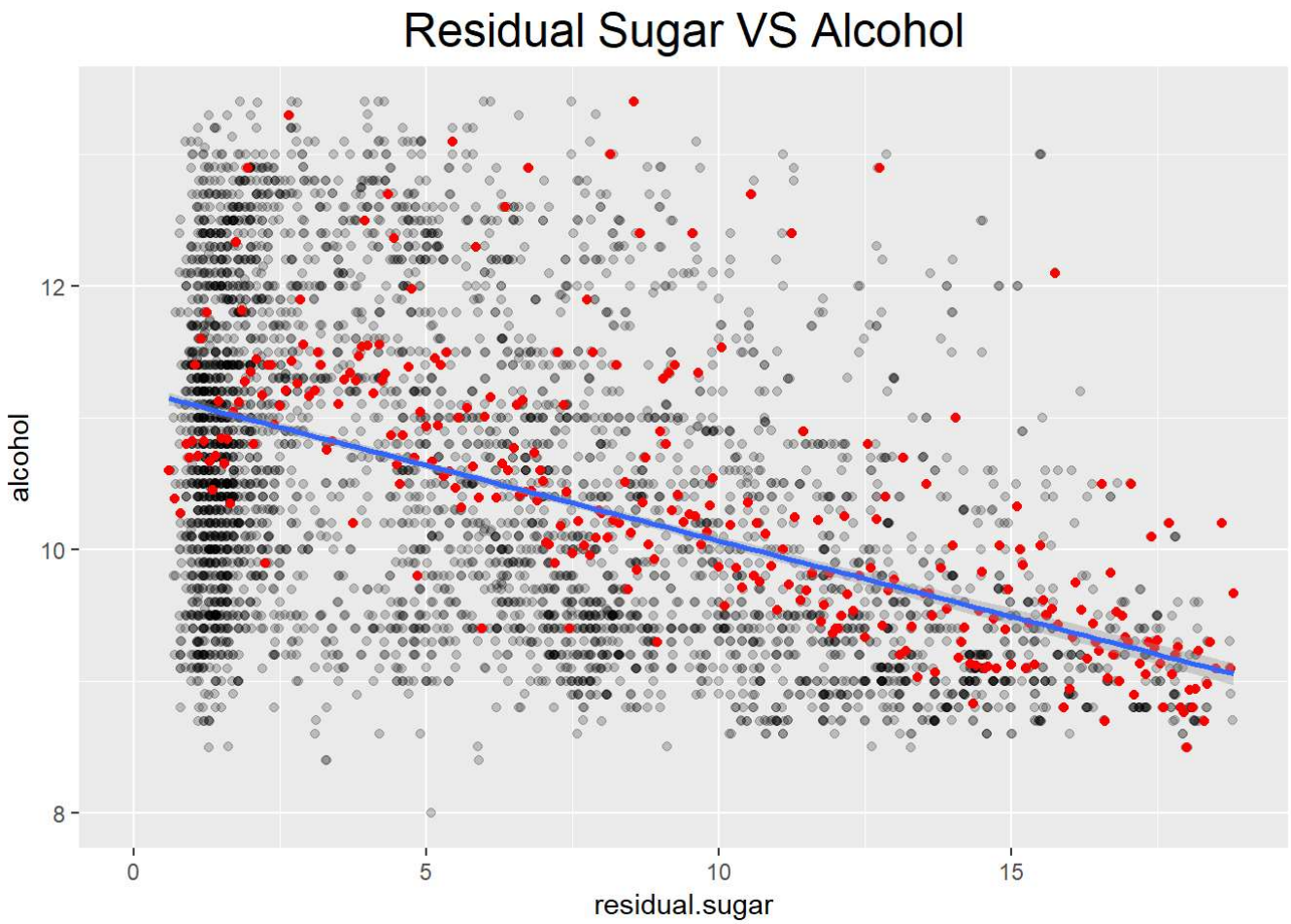
```
##
##  Pearson's product-moment correlation
##
## data:  wq$free.sulfur.dioxide and wq$density
## t = 21.54, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2684156 0.3195836
## sample estimates:
##       cor
## 0.2942104
```

Density has a moderate positive correlation with free sulfur dioxide, more free sulfur dioxide, higer density.



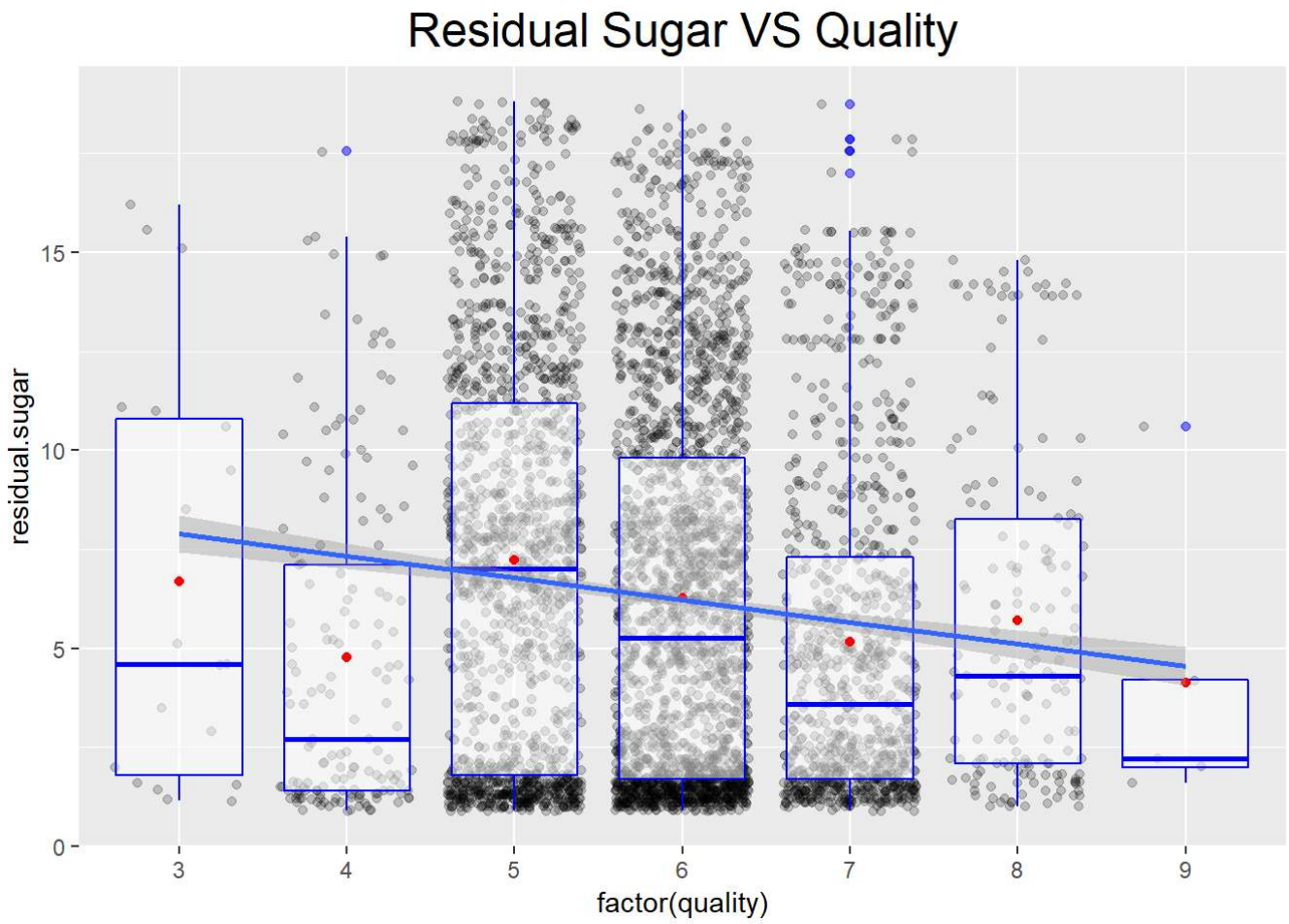
```
##
##  Pearson's product-moment correlation
##
## data:  wq$total.sulfur.dioxide and wq$density
## t = 43.719, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5094349 0.5497297
## sample estimates:
##       cor
## 0.5298813
```

Density has a moderate positive correlation with total sulfur dioxide, more total sulfur dioxide, higer density.



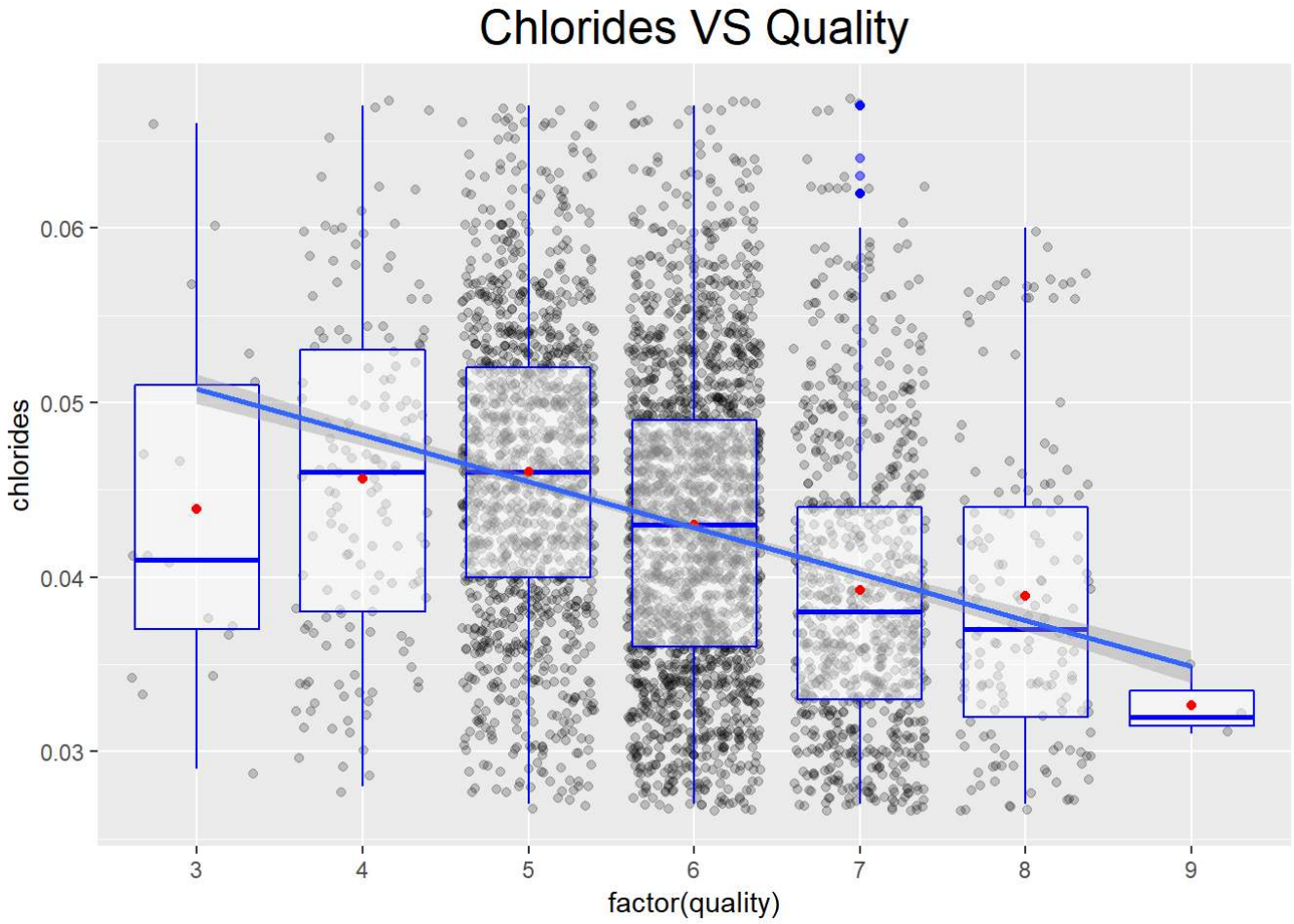
```
##
##  Pearson's product-moment correlation
##
## data:  wq$residual.sugar and wq$alcohol
## t = -35.321, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4726723 -0.4280267
## sample estimates:
##       cor
## -0.4506312
```

Alcohol has a moderate positive correlation with redisual sugar.



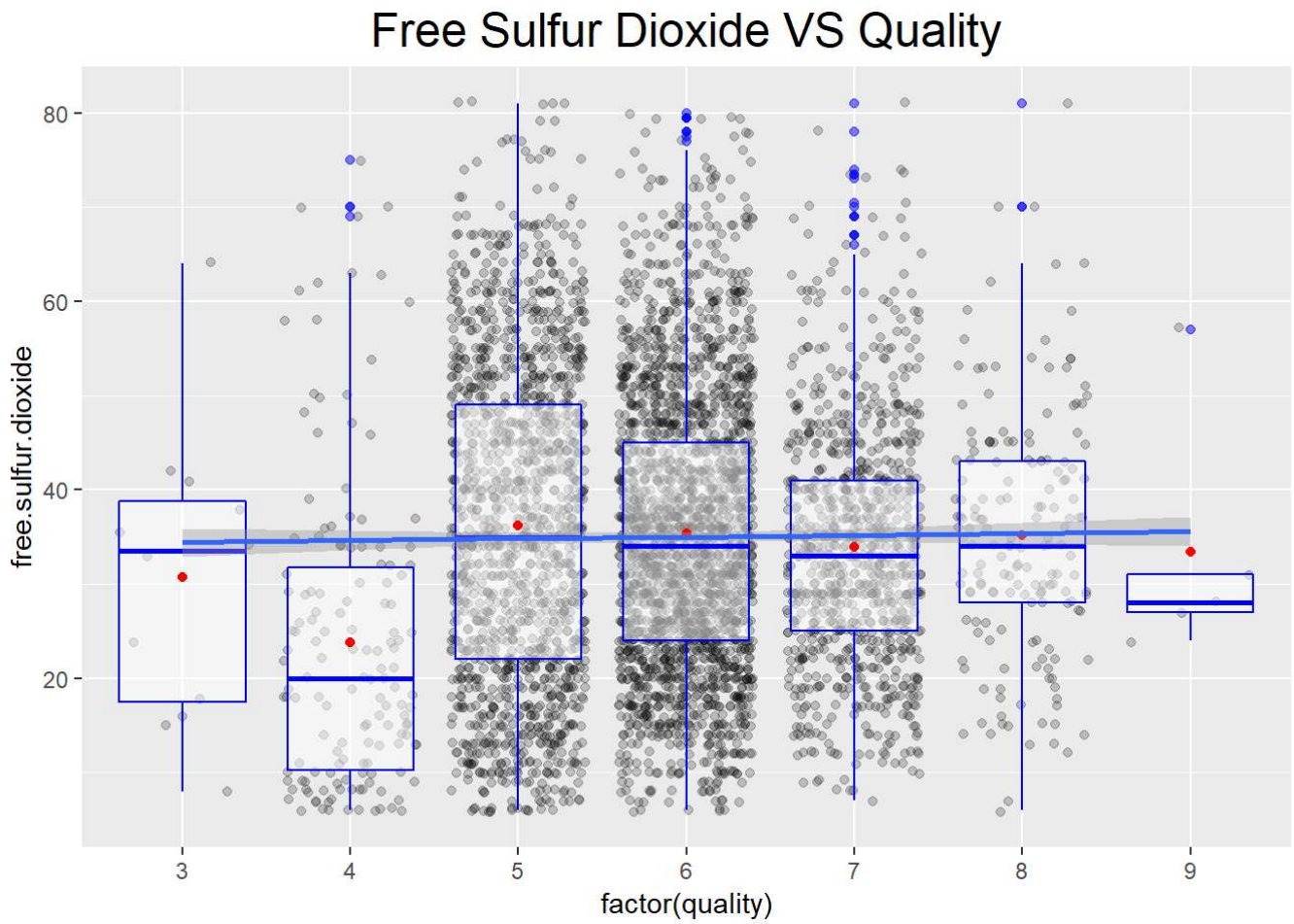

```
##
##  Pearson's product-moment correlation
##
## data:  wq_r$quality and wq_r$residual.sugar
## t = -7.1651, df = 4815, p-value = 8.953e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.13057389 -0.07468736
## sample estimates:
##       cor
## -0.1027117
```

Remove outliers, it is easy to see that quality doesn't have obvious relationship with residual sugar.



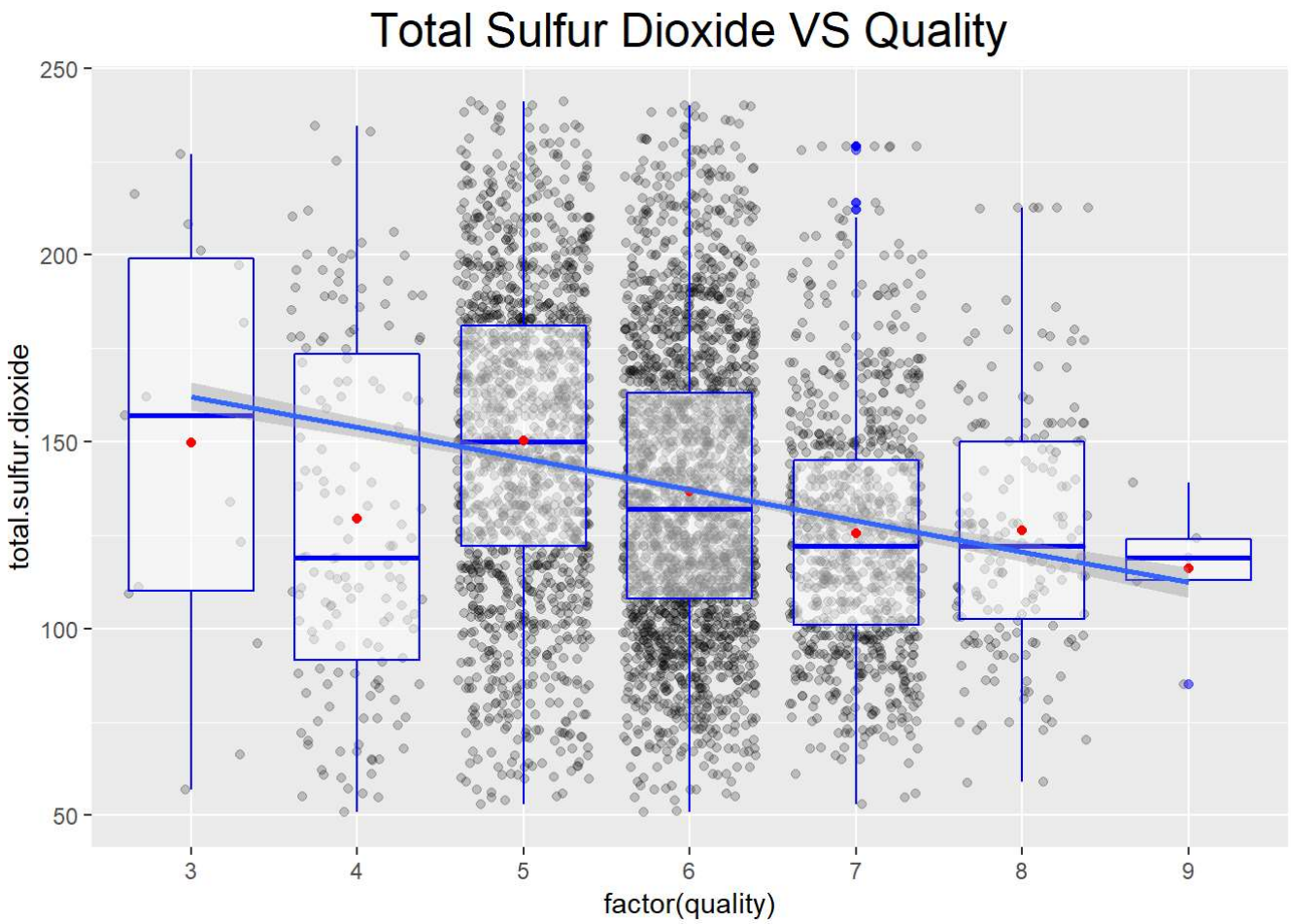

```
##
##  Pearson's product-moment correlation
##
## data:  wq_c$quality and wq_c$chlorides
## t = -18.254, df = 4427, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2917543 -0.2369700
## sample estimates:
##       cor
## -0.2645756
```

Remove top 5% and bottom 5% outliers, it is easy to see that with chlorides increases, quality decreases.




```
##
##  Pearson's product-moment correlation
##
## data:  wq_f$quality and wq_f$free.sulfur.dioxide
## t = 0.75958, df = 4806, p-value = 0.4475
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01731668  0.03921132
## sample estimates:
##        cor
## 0.01095607
```

Free sulfur dioxide doesn't have obvious influence on quality.




```
##
##  Pearson's product-moment correlation
##
## data:  wq_t$quality and wq_t$total.sulfur.dioxide
## t = -13.024, df = 4798, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2119747 -0.1573235
## sample estimates:
##        cor
## -0.1847919
```

Total sulfur dioxide has weak relationship with quality.

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?

Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide and density have negative influence on quality, while free.sulfur.dioxide, pH, sulphates and alcohol have positive influence on quality. It seems that density and alcohol have

moderate correlations with quality, there isn't any variable has strong correlation with quality.

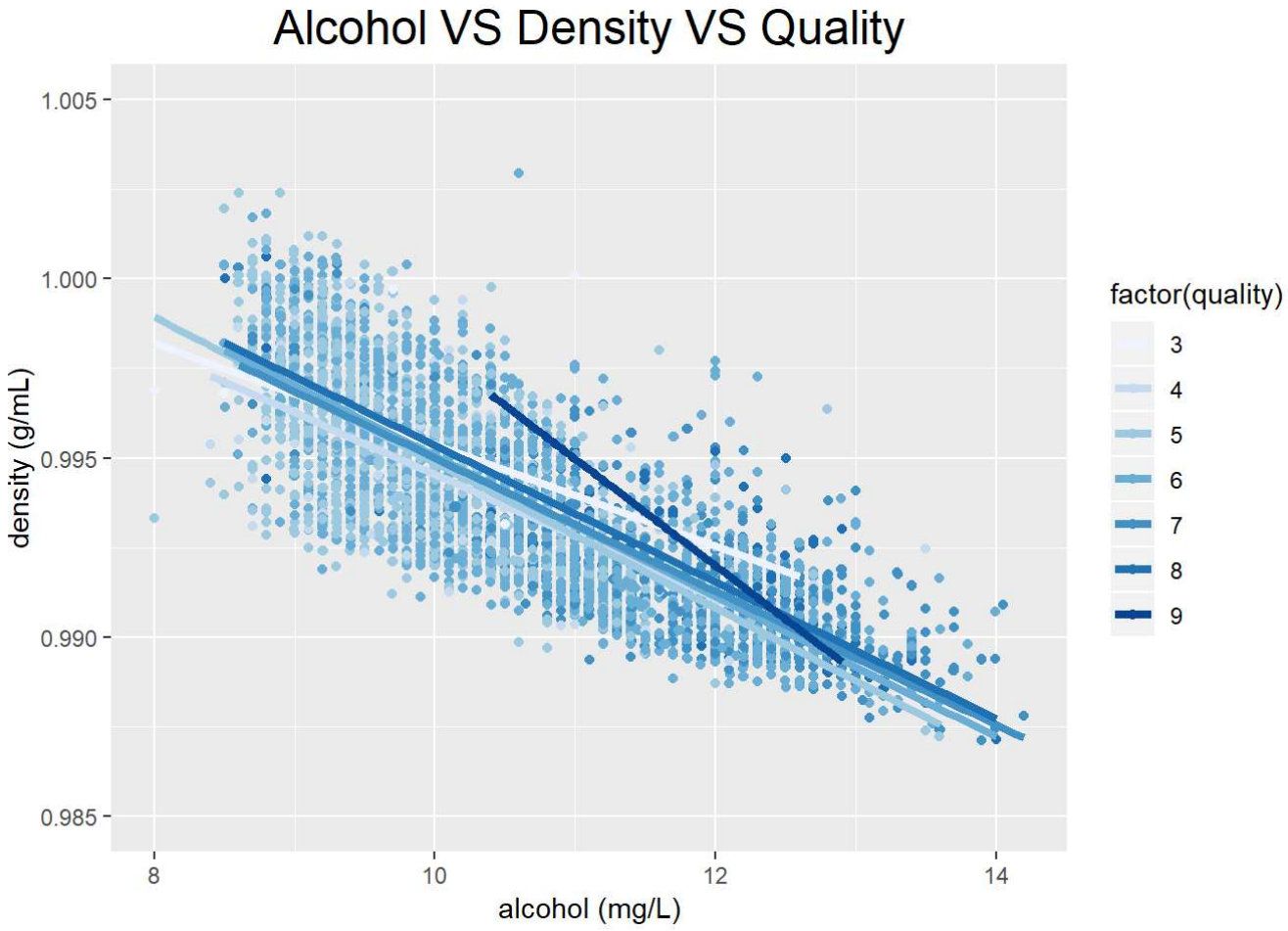
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

Yeah, I obeserved the variables that influence density and alcohol. Density has strong correlation with residual sugar and alcohol, while moderately correlated with free sulfur dioxide and total sulfur dioxide. Alcohol has strong correlation with density, while moderately correlated with residual sugar, chlorides and total sulfur dioxide.

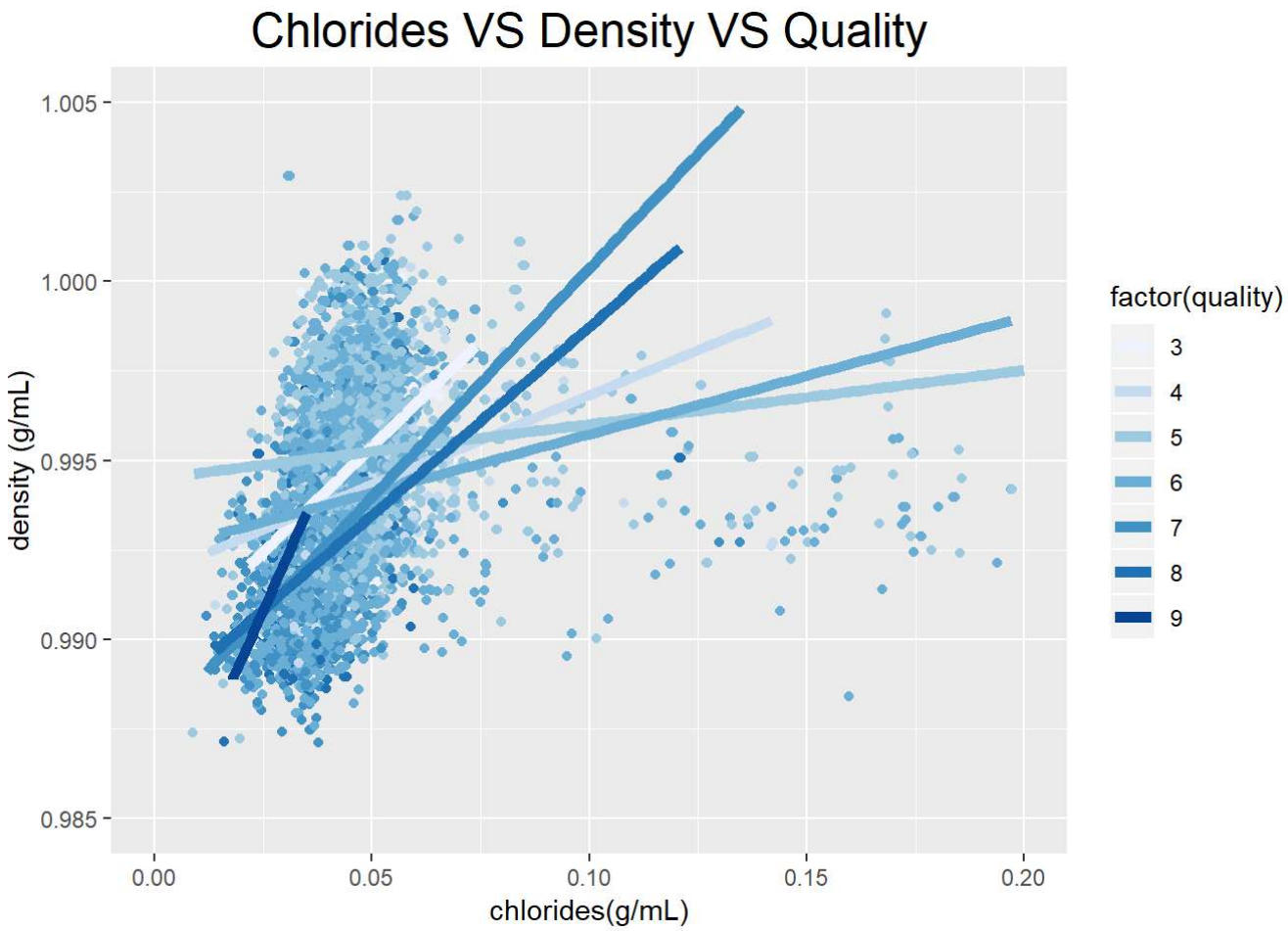
What was the strongest relationship you found?

Alcohol has the greatest influence on quality, alcohol and density have the strongest relationship.

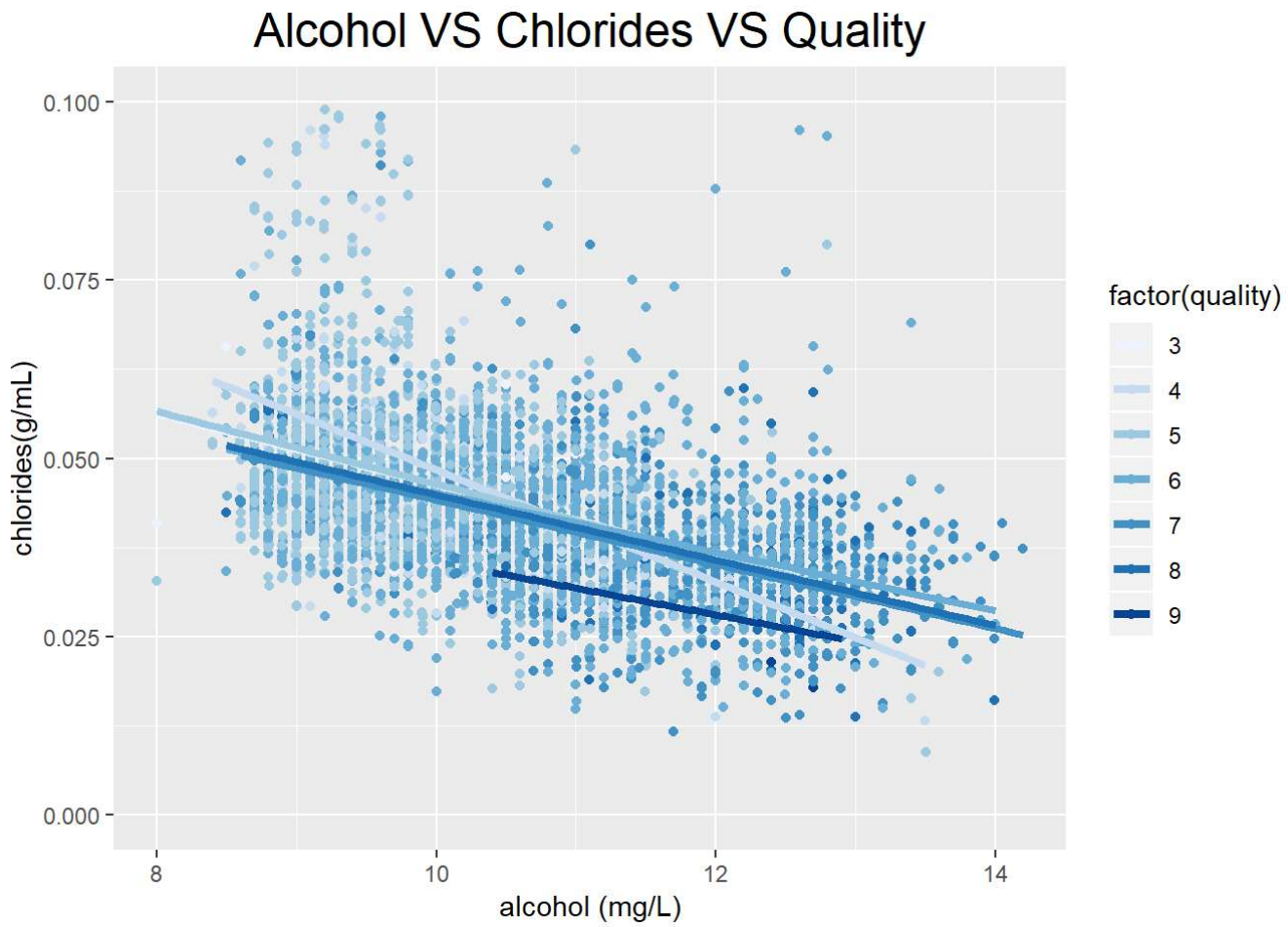
Multivariate Plots Section



With the same density, more alcohol, higher quality. If density is influenced by alcohol more, quality will be higher.



A lower chlorides and a higher density will have a higher quality.



A higher alcohol and a lower chlorides will have a higher quality.

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?

With the same density, more alcohol, higher quality. If density is influenced by alcohol more, quality will be higher. A lower chlorides and a higher density will have a higher quality.

Were there any interesting or surprising interactions between features?

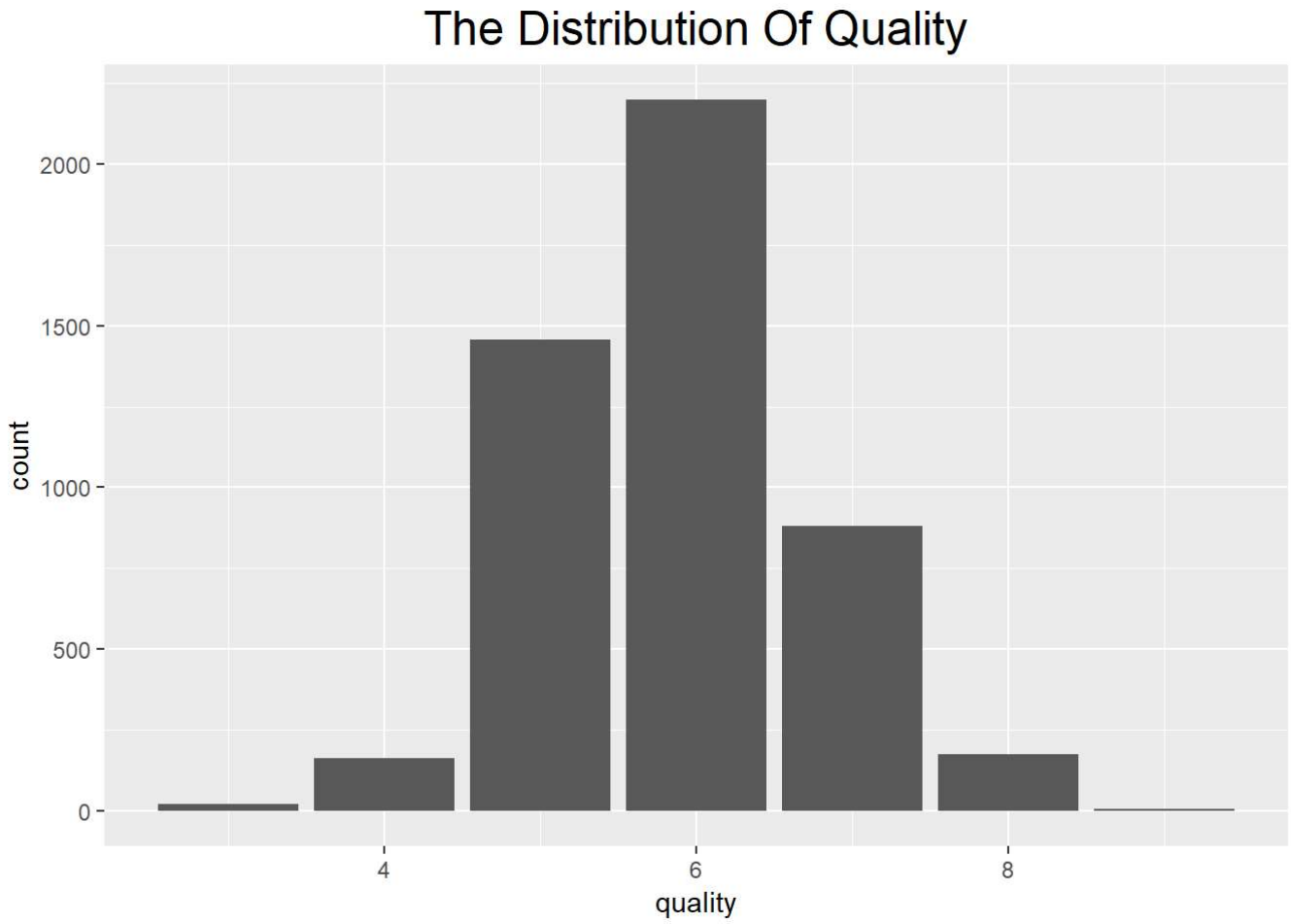
pH and residual sugar have weak relationship with quality.

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

No.

Final Plots and Summary

Plot One



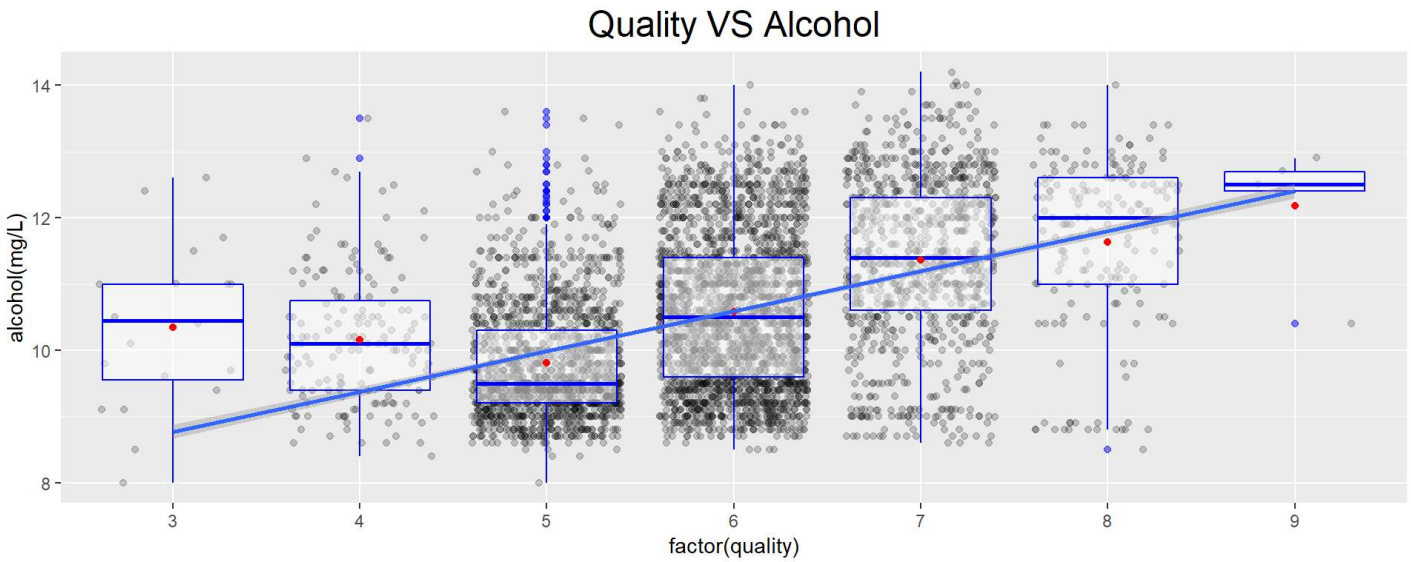

```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

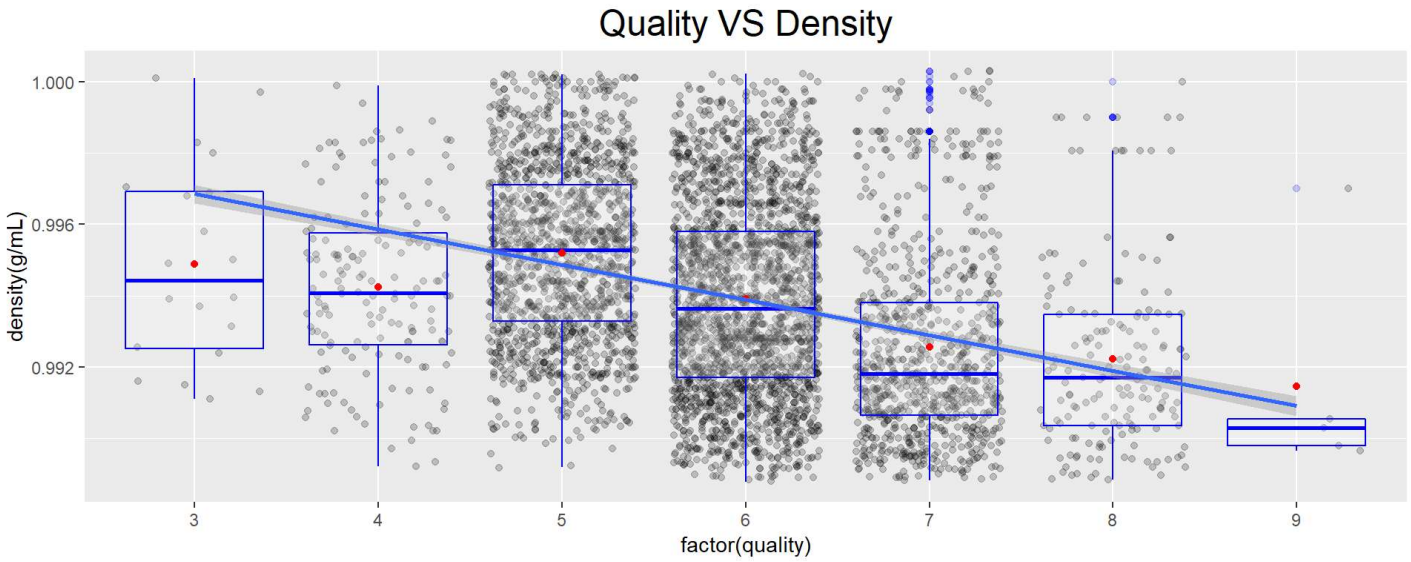
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000
```

Description One

The distribution of quality is concentrated in the middle value, median value is 6, mean value is 5.878. Most wine's quality is 6.

Plot Two

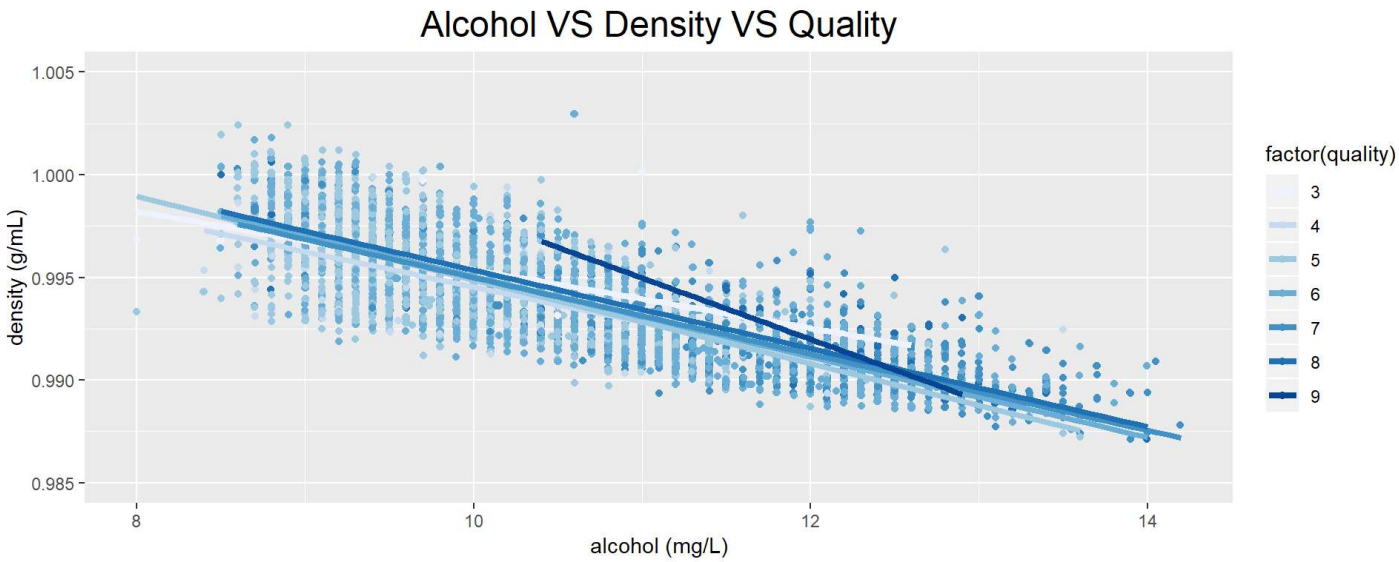


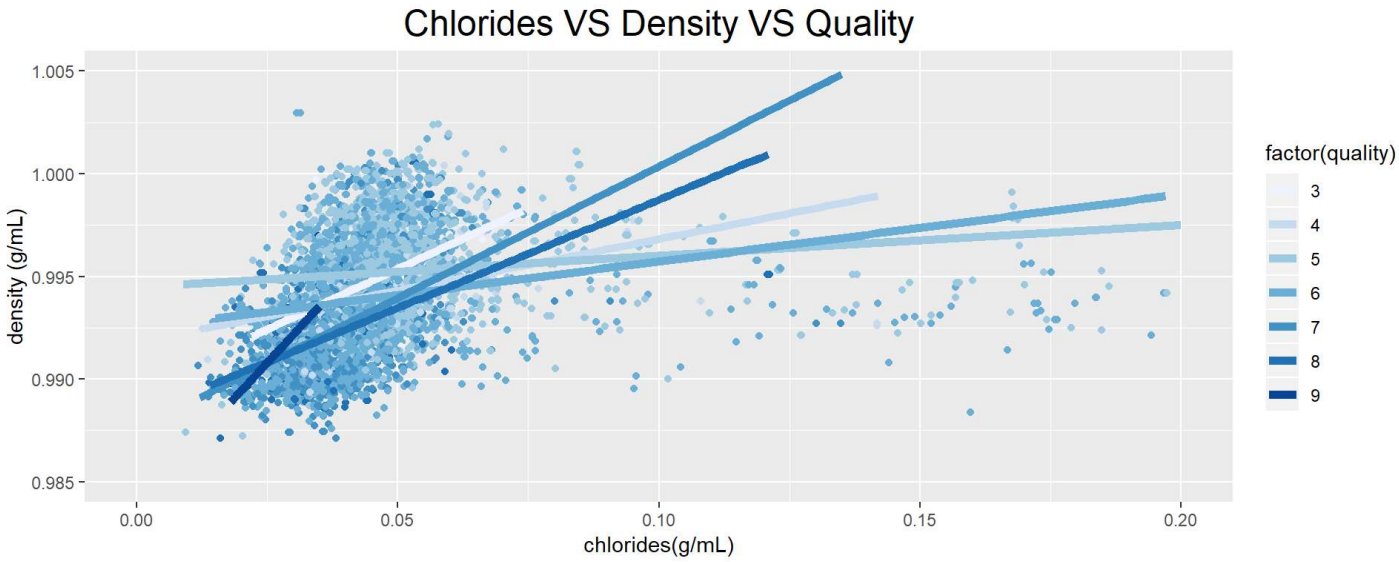


Description Two

With the increase of alcohol, quality increases. With the increase of density, quality decreases.

Plot Three





Description Three

With a low chlorides, if density is influenced by alcohol more, quality will be higher.

Reflection

There are 4,898 obeservations in this white wine quality dataset with 12 features. I need to find these variables which influence wine quality more.

Firstly, I observed distributions of each variable. Secondly, I used ggpairs and ggcorr to explore the relationship between variables. There isn't any variable have strong relationship with quality, it made me a little disappointed. Fortunately, density and alcohol have moderate relationship with quality, and they have strong relationship with some other variables.

Then, I decided to explore the variables that may affect quality, for example, those variables have strong correlation with density and alcohol.

Finally, through these bivariate and multivariate analysis, I found some conclusions.

However， these are only parts of the secrets that is hidden in this dataset.

I should practice more and use more complex visualization to explore datasets.