Red Wines Quality Exploration by Renqin Yang

I picked up this red wine quality dataset since my father is a big fan of red wine. And he also taught me some personal experiences on how to rate a red wine. Therefore, I want to test his little theory in this project, with the key variable – quality.

This tidy data set contains 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

Univariate Plots Section

```
## [1] 1599   13
```

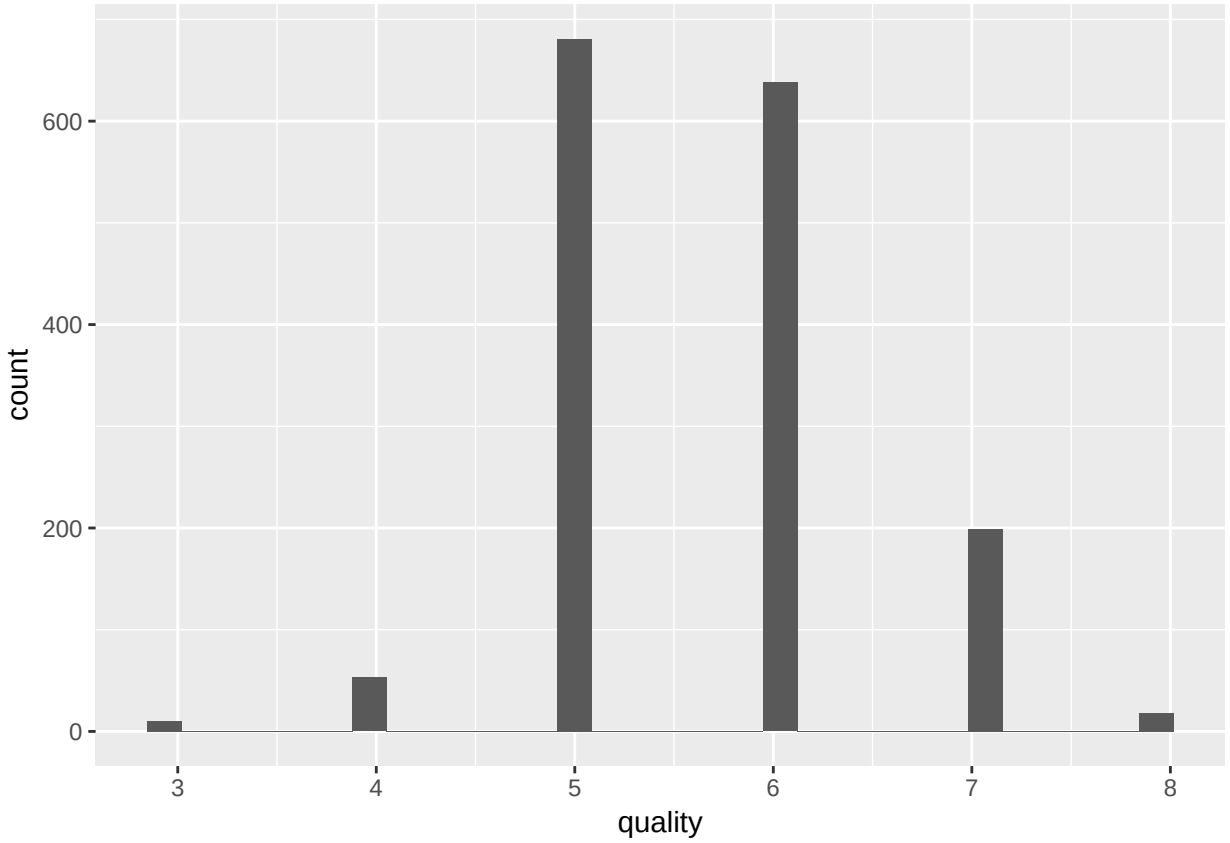
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  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 0.073 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 ...
```

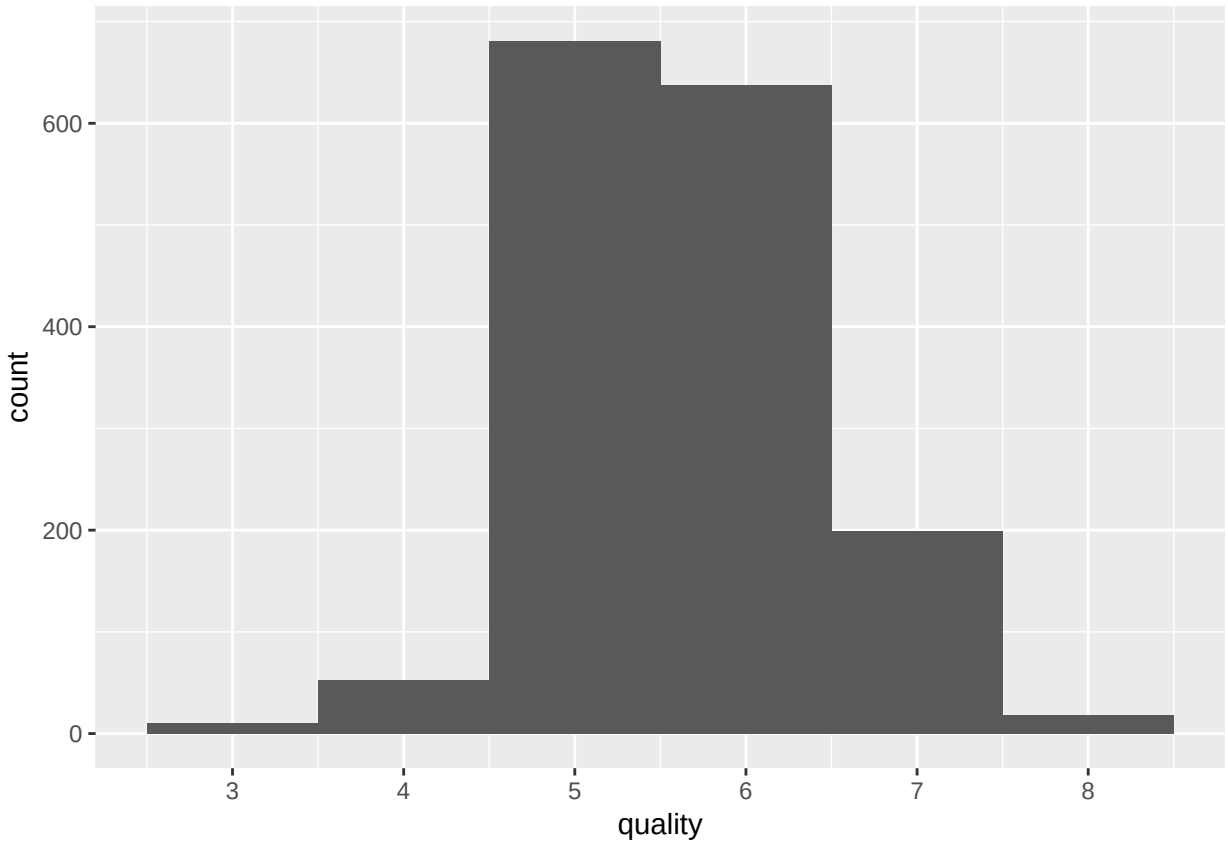
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean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

```
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

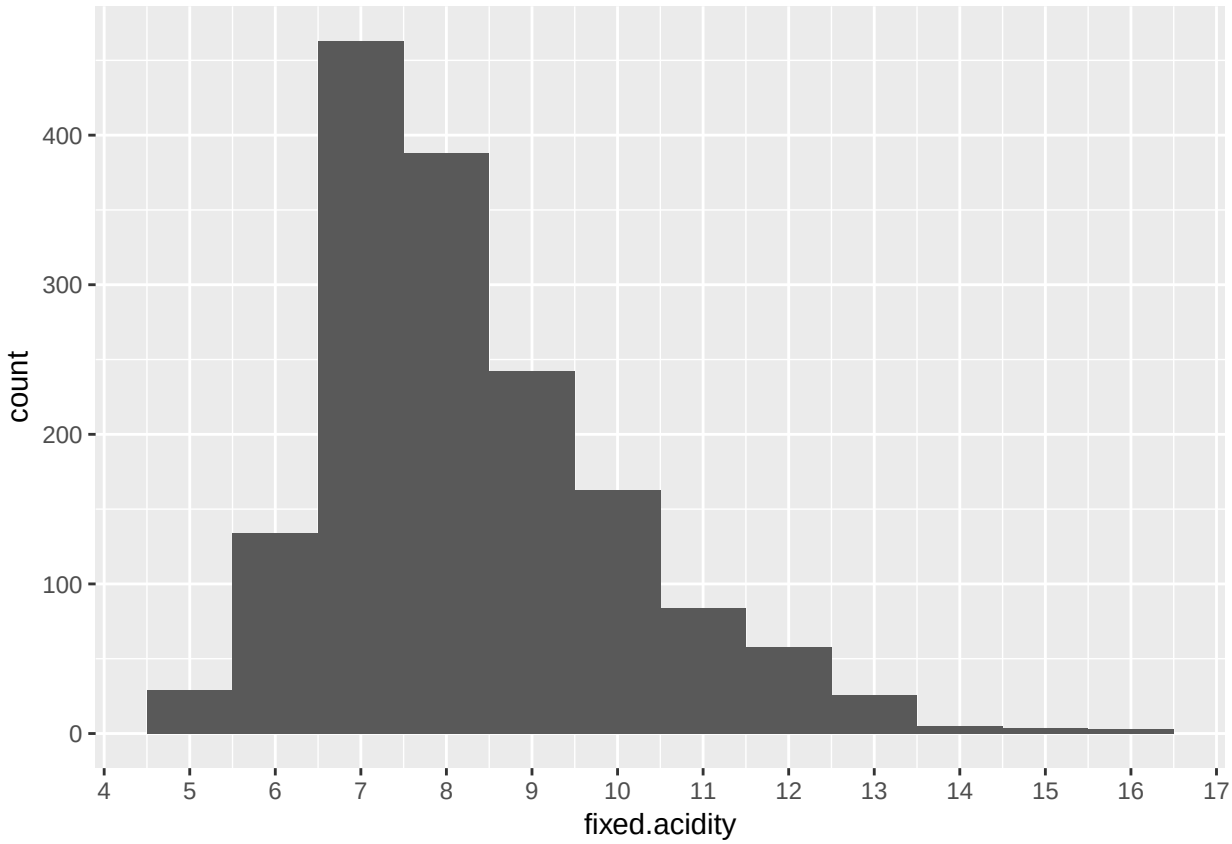
Our dataset consists of 13 variables, with almost 1599 observations.

And the first variable 'X' is an 'id' variable, just used to identify each individual ovservation.

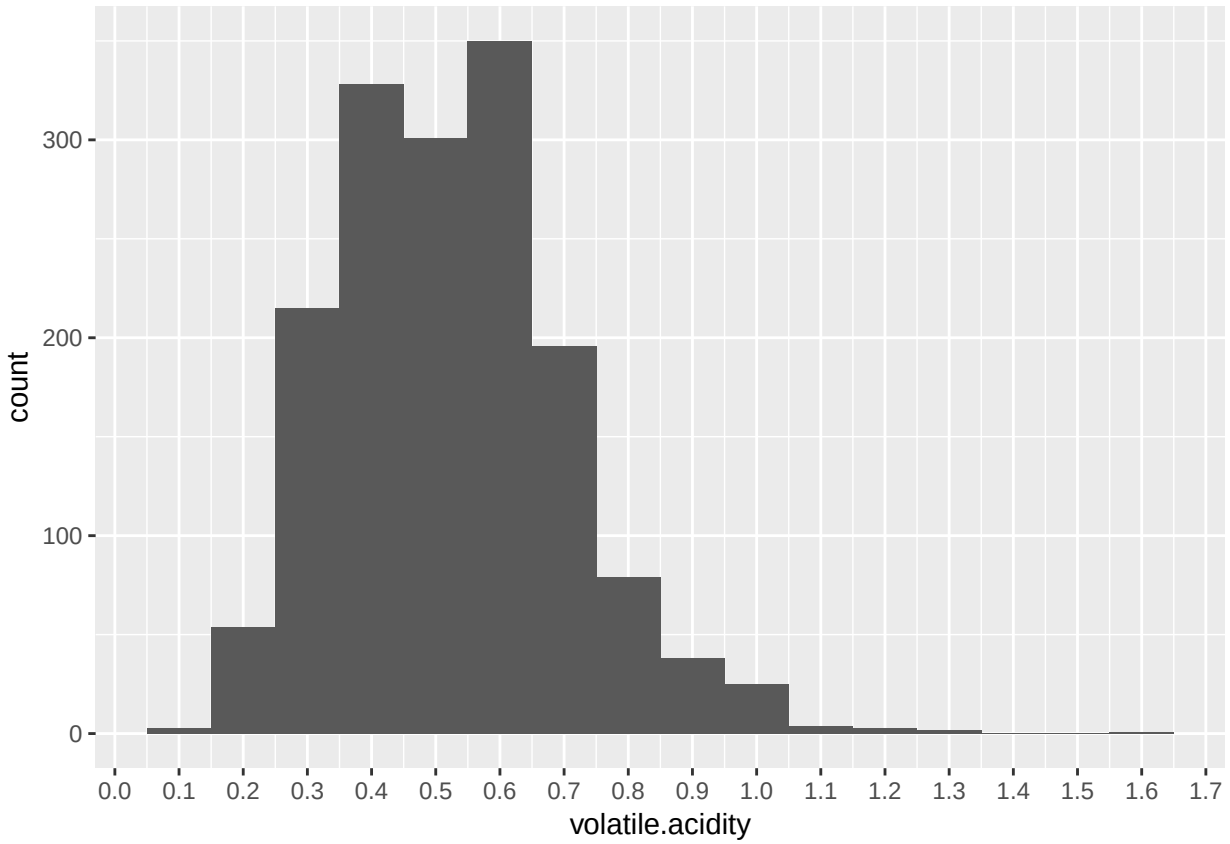




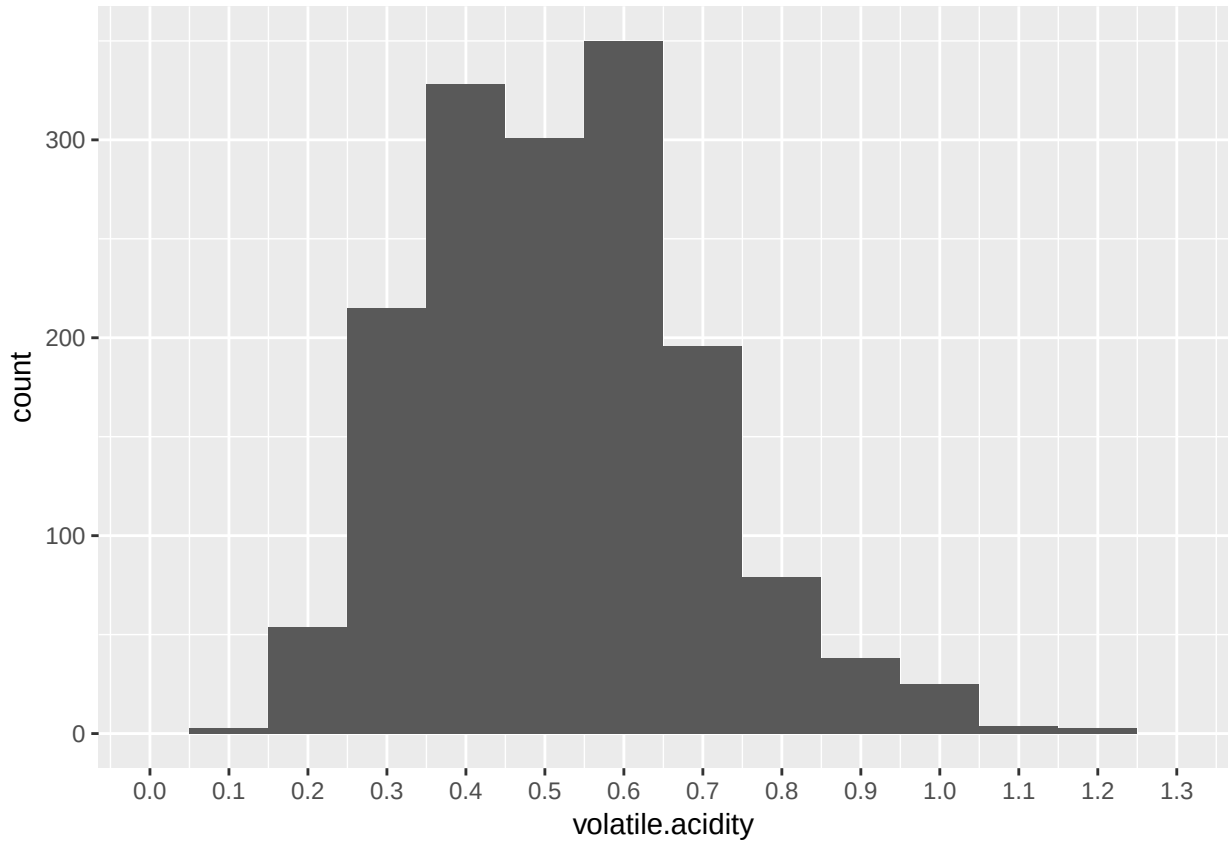
Transformed the very discreted data in a better format by adjusting binwidth and breaks the x coordinates. We can see that most of wines are rated as '5' or '6', a median rate. There is no wine's rate is lower than '3', or higher than '8', and the wines which are rated as '3' or '8' are also very limited.Therefore, the distribution of quality of wine is well, there is most no outliers. Next, I'll exmain the 11 chemical properties of the wine, trying to find one or more, or a combination of chemical properties most likely affect the rate of wine.

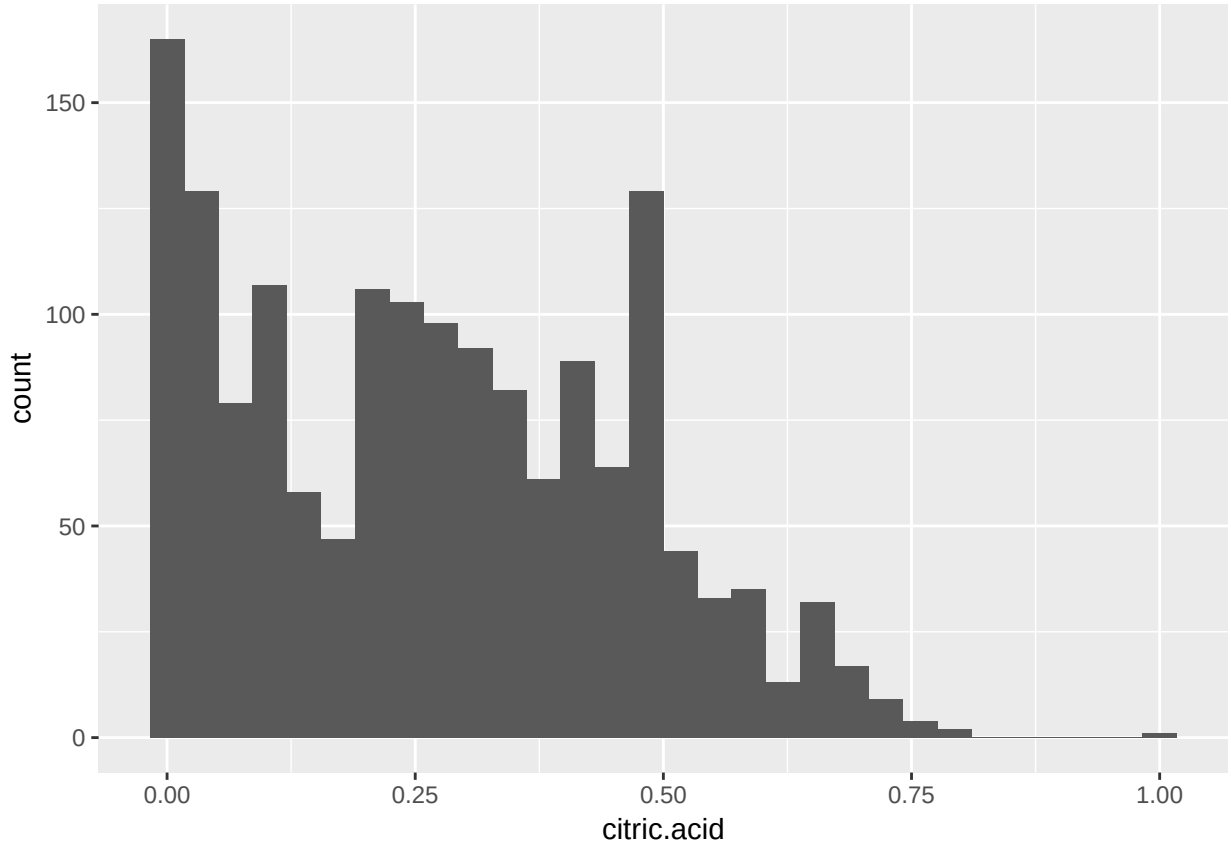


With binwidth equal to 1, the distribution of fixed.acidity is very good, almost a normal distribution, without any rescale on axis or limits on original data.



Even though it rougly distribute well, we can still see some extreme values or outliers. So I decided to add some limits on original data.

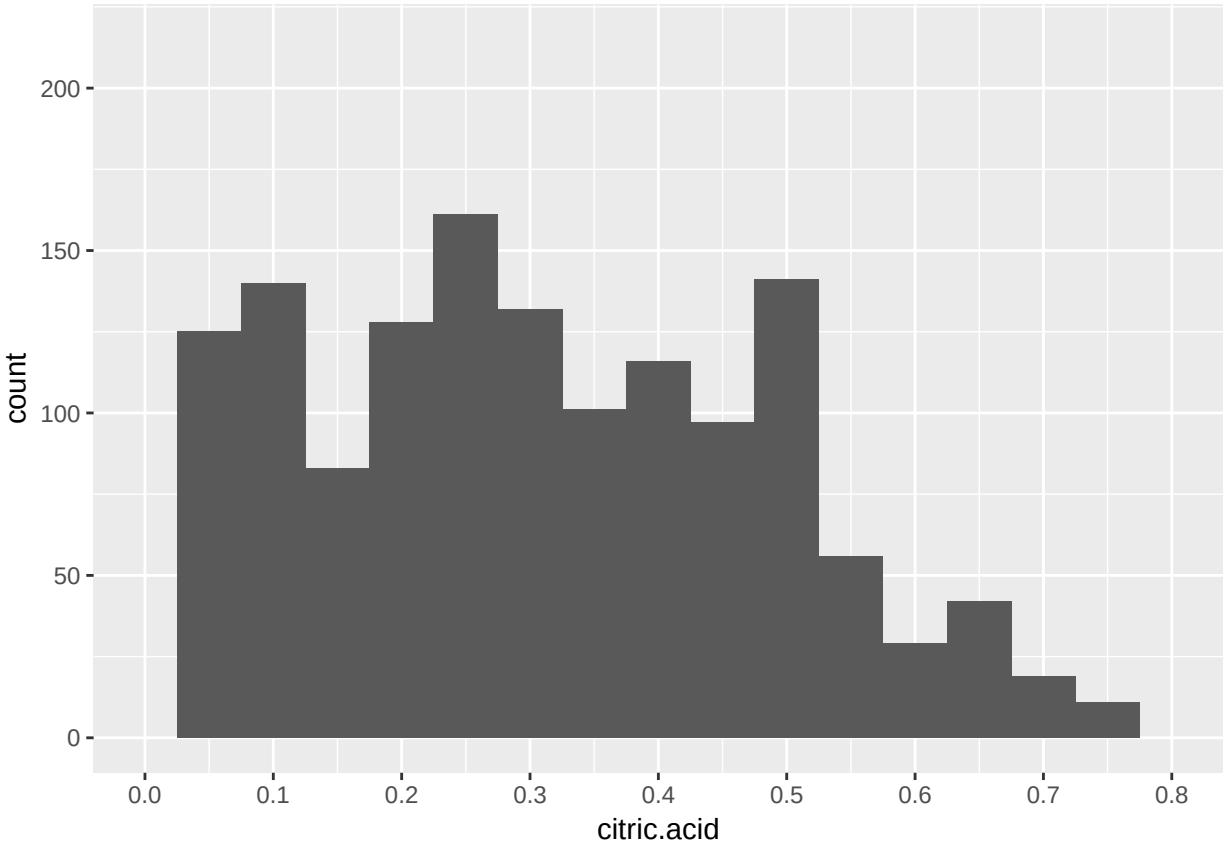

count
300
200
100
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
volatile.acidity

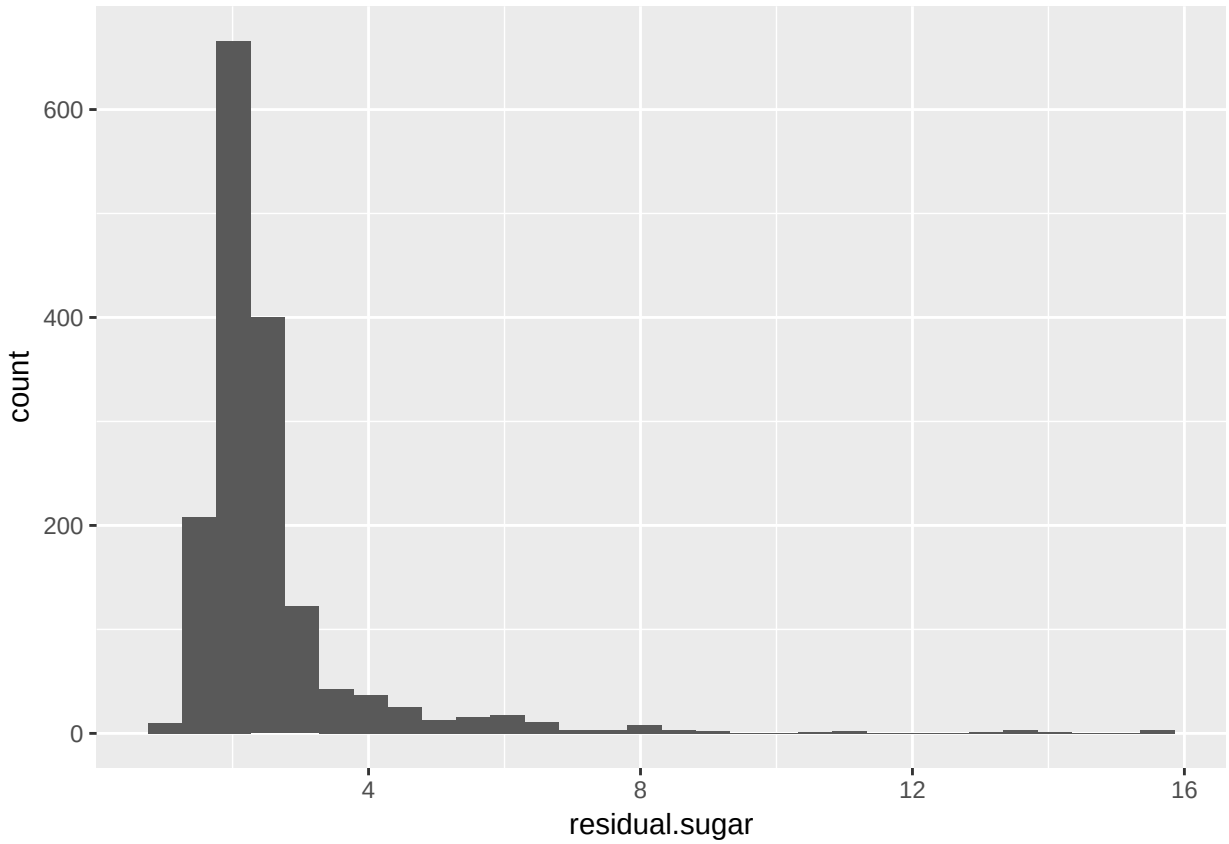

count
150
100
50
0
0.00
0.25
0.50
0.75
1.00
citric.acid

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

```
##
##    0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09  0.1 0.11 0.12 0.13 0.14
##  132   33   50   30   29   20   24   22   33   30   35   15   27   18   21
## 0.15 0.16 0.17 0.18 0.19  0.2 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29
##   19    9   16   22   21   25   33   27   25   51   27   38   20   19   21
##  0.3 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39  0.4 0.41 0.42 0.43 0.44
##   30   30   32   25   24   13   20   19   14   28   29   16   29   15   23
## 0.45 0.46 0.47 0.48 0.49  0.5 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59
##   22   19   18   23   68   20   13   17   14   13   12    8    9    9    8
##  0.6 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69  0.7 0.71 0.72 0.73 0.74
##    9    2    1   10    9    7   14    2   11    4    2    1    1    3    4
## 0.75 0.76 0.78 0.79    1
##    1    3    1    1    1
```

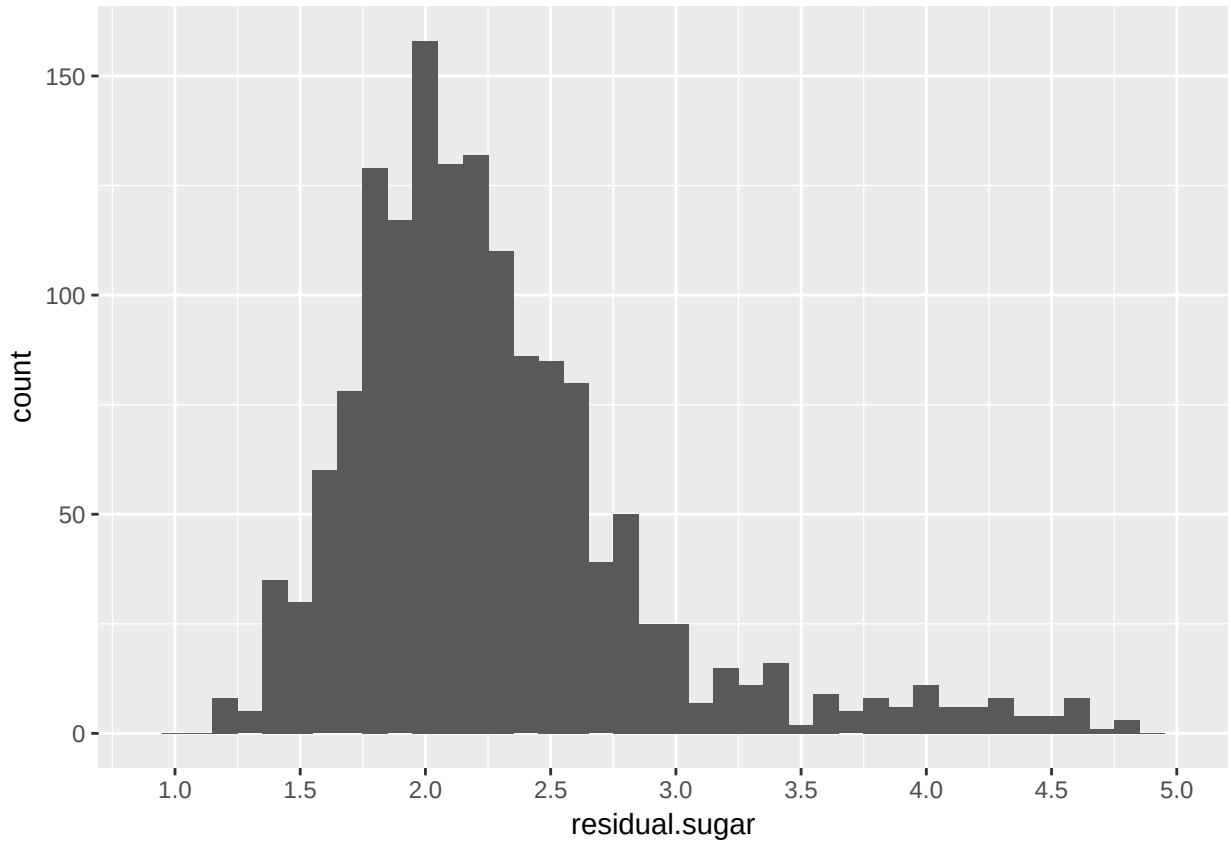
It's clear to see that the outliers on the rightest of the graph, and the distribution of citric.acid is left skewed. I print out the table of citric.acid, we can see that the main body of this data are from 0 to around 0.7. Thereforem, I tranformed this graph by changing the binwidth and the breaks and limits on x coordinaetes.

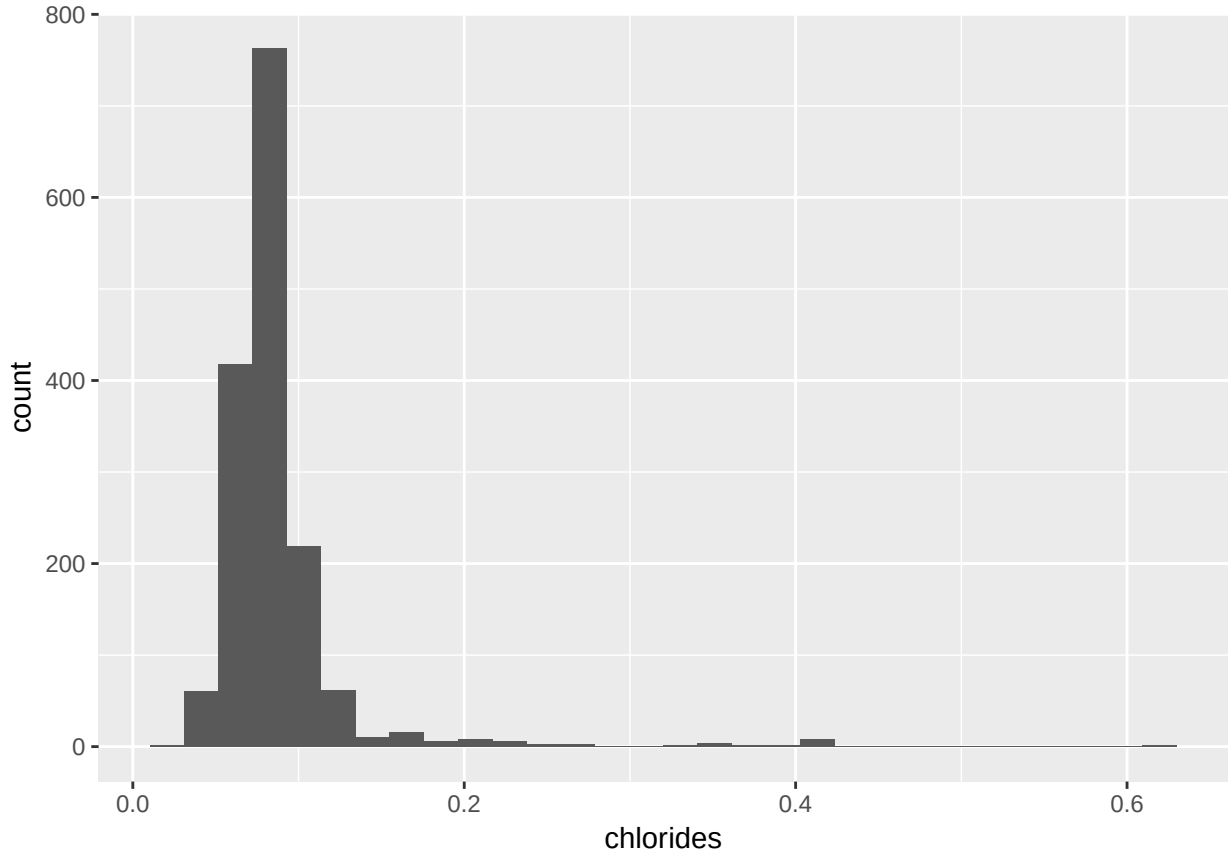




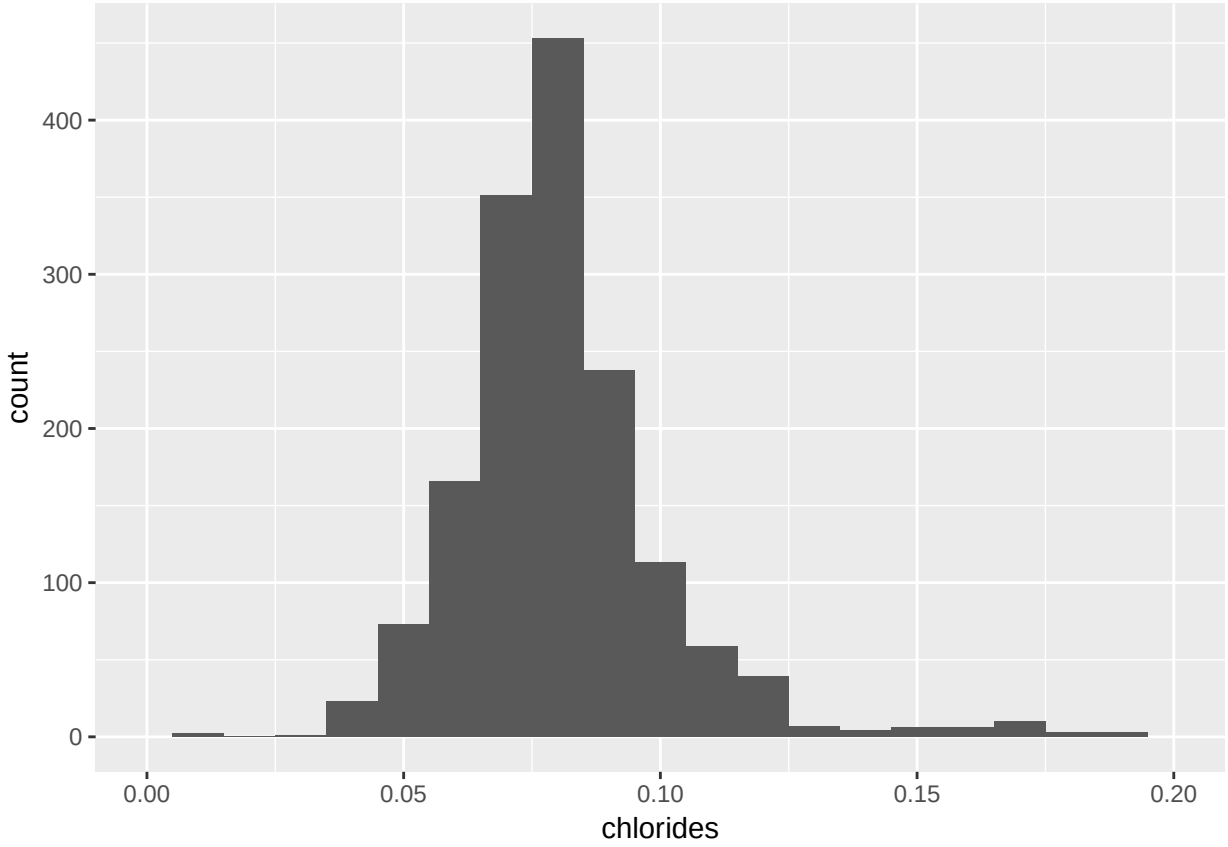

```
##
##  0.9  1.2  1.3  1.4  1.5  1.6 1.65  1.7 1.75  1.8  1.9    2 2.05  2.1 2.15
##    2    8    5   35   30   58    2   76    2  129  117  156    2  128    2
##  2.2 2.25  2.3 2.35  2.4  2.5 2.55  2.6 2.65  2.7  2.8 2.85  2.9 2.95    3
##  131    1  109    1   86   84    1   79    1   39   49    1   24    1   25
##  3.1  3.2  3.3  3.4 3.45  3.5  3.6 3.65  3.7 3.75  3.8  3.9    4  4.1  4.2
##    7   15   11   15    1    2    8    1    4    1    8    6   11    6    5
## 4.25  4.3  4.4  4.5  4.6 4.65  4.7  4.8    5  5.1 5.15  5.2  5.4  5.5  5.6
##    1    8    4    4    6    2    1    3    1    5    1    3    1    8    6
##  5.7  5.8  5.9    6  6.1  6.2  6.3  6.4 6.55  6.6  6.7    7  7.2  7.3  7.5
##    1    4    3    4    4    3    2    3    2    2    2    1    1    1    1
##  7.8  7.9  8.1  8.3  8.6  8.8  8.9    9 10.7   11 12.9 13.4 13.8 13.9 15.4
##    2    3    2    3    1    2    1    1    1    2    1    1    2    1    2
## 15.5
##    1
```

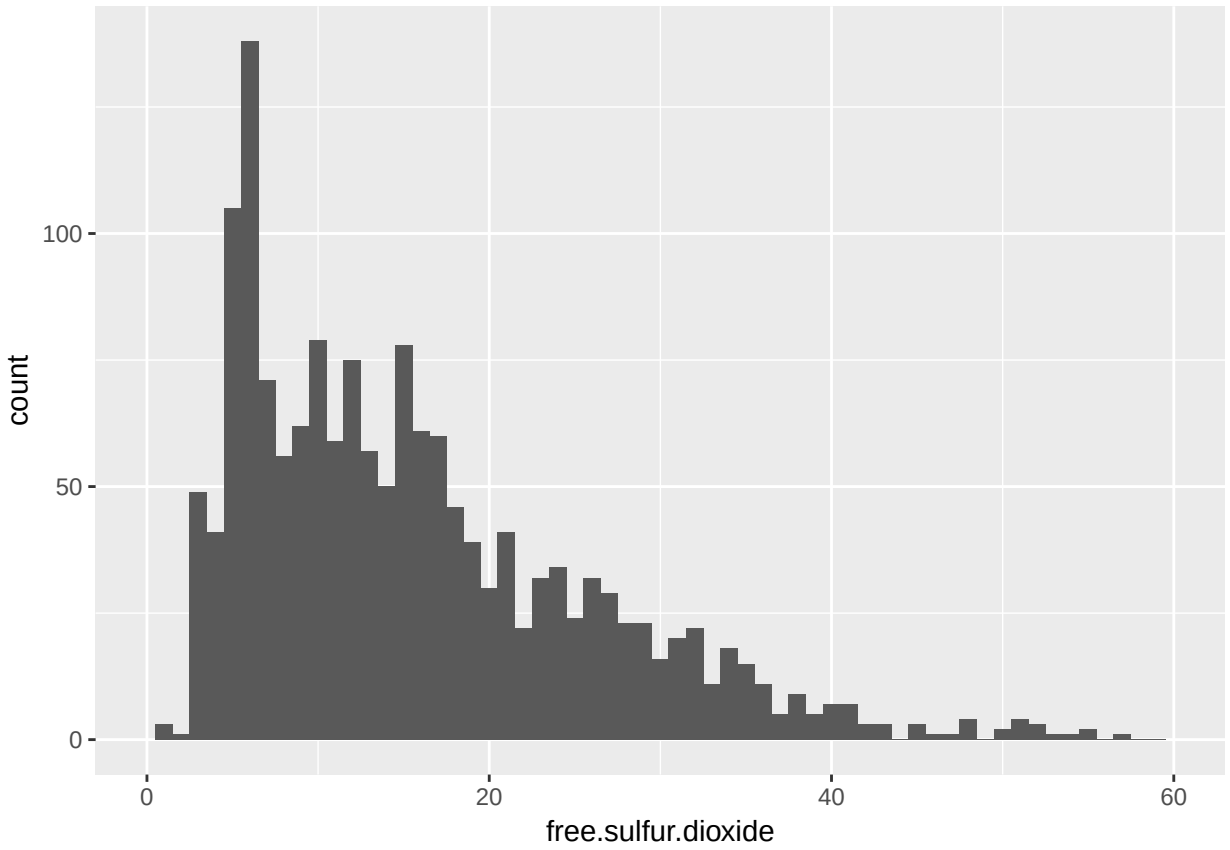
Again, I just adjusted the binwidth and limits on x coordinates to tranform the original graph.


150
100
50
0
count
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
residual.sugar

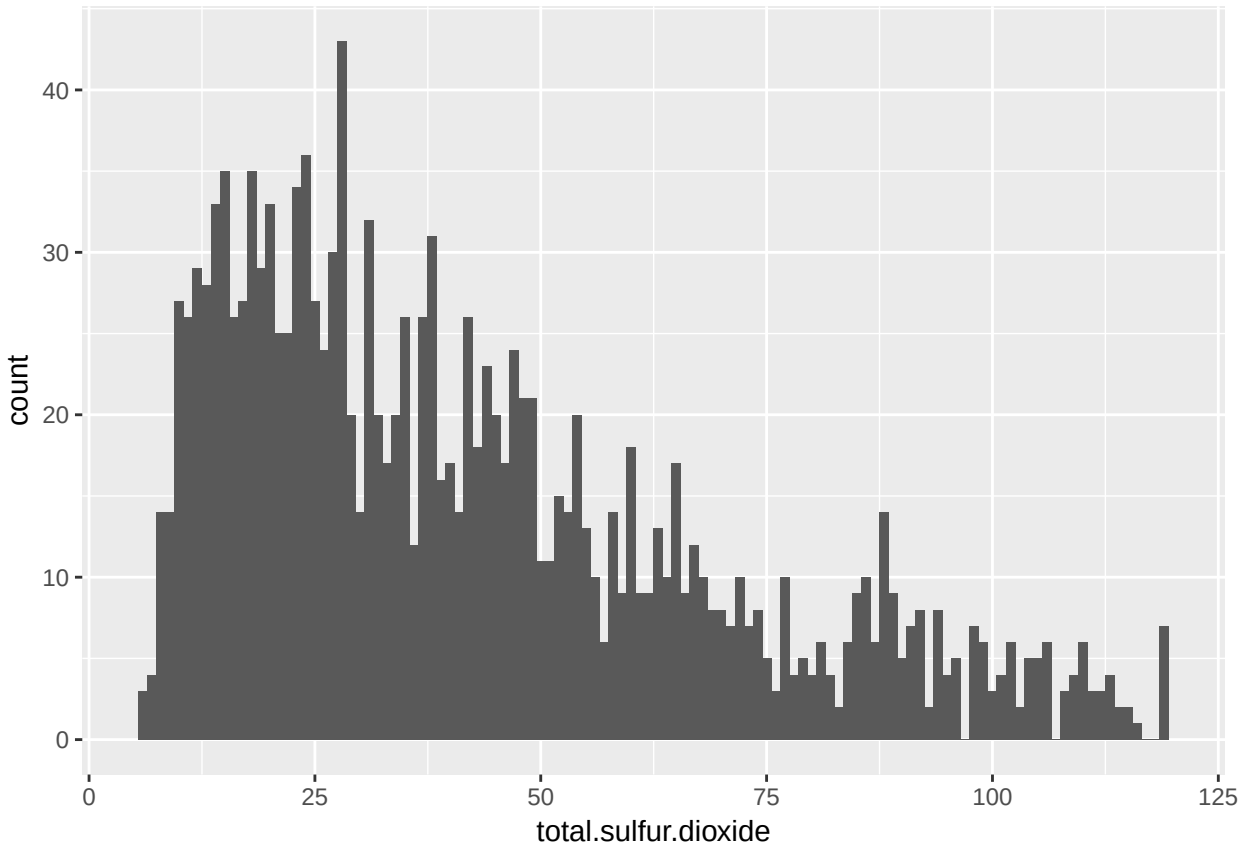

800
600
400
200
0
count
0.0
0.2
0.4
0.6
chlorides

Clearly, the outliers and extreme values contaminate the distribution of chlorides, I would adjust this by adding limits on data.

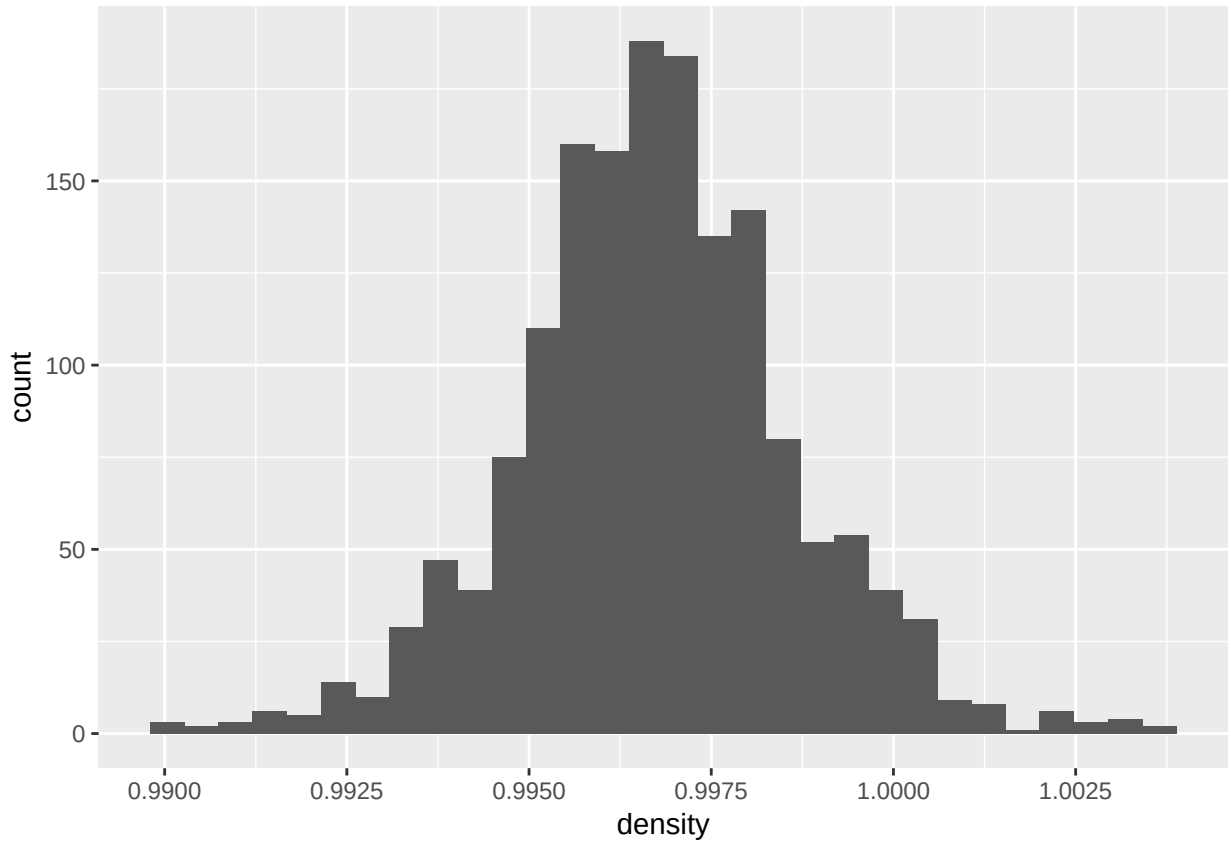


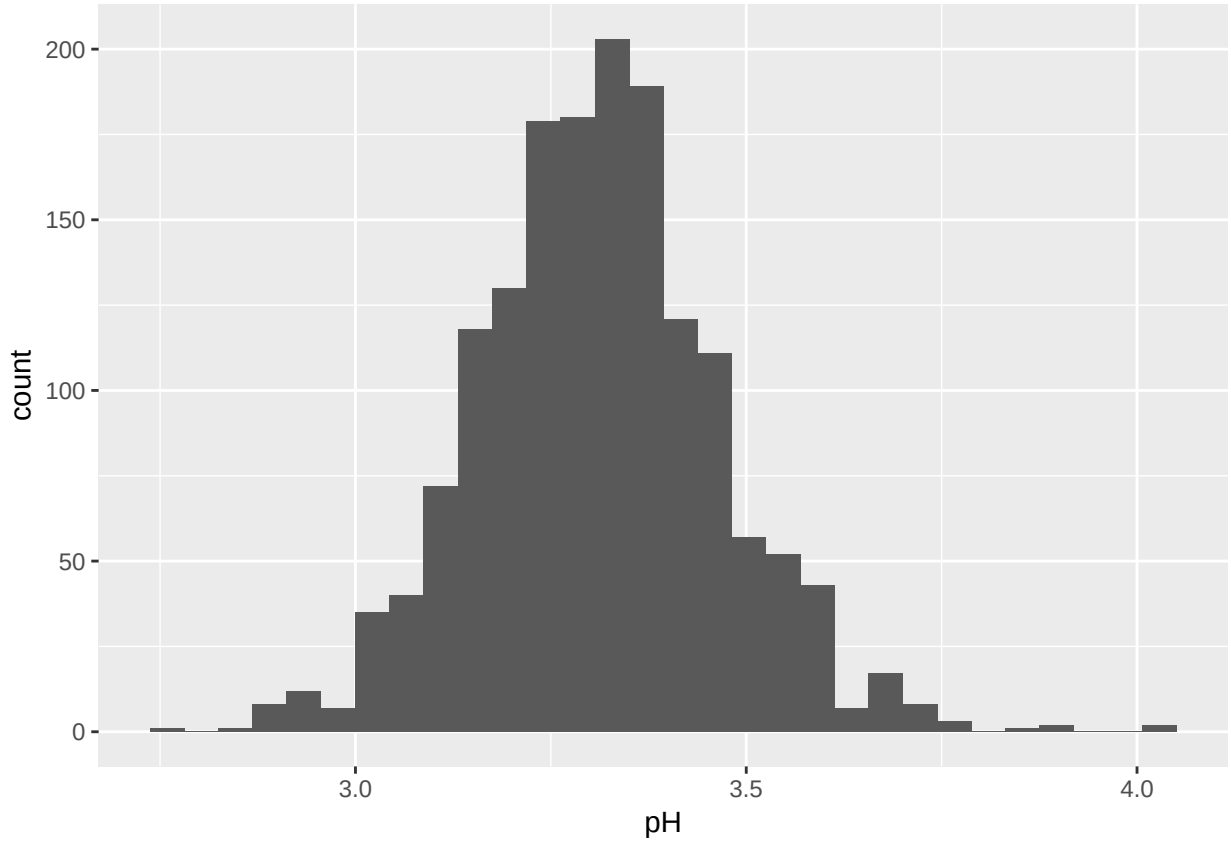


After refining and adding limits, the free.sulfur.dioxide seems distribute well, most wines have a concentration of free.sulfur.dioxide below 20 (mg / dmˆ3).

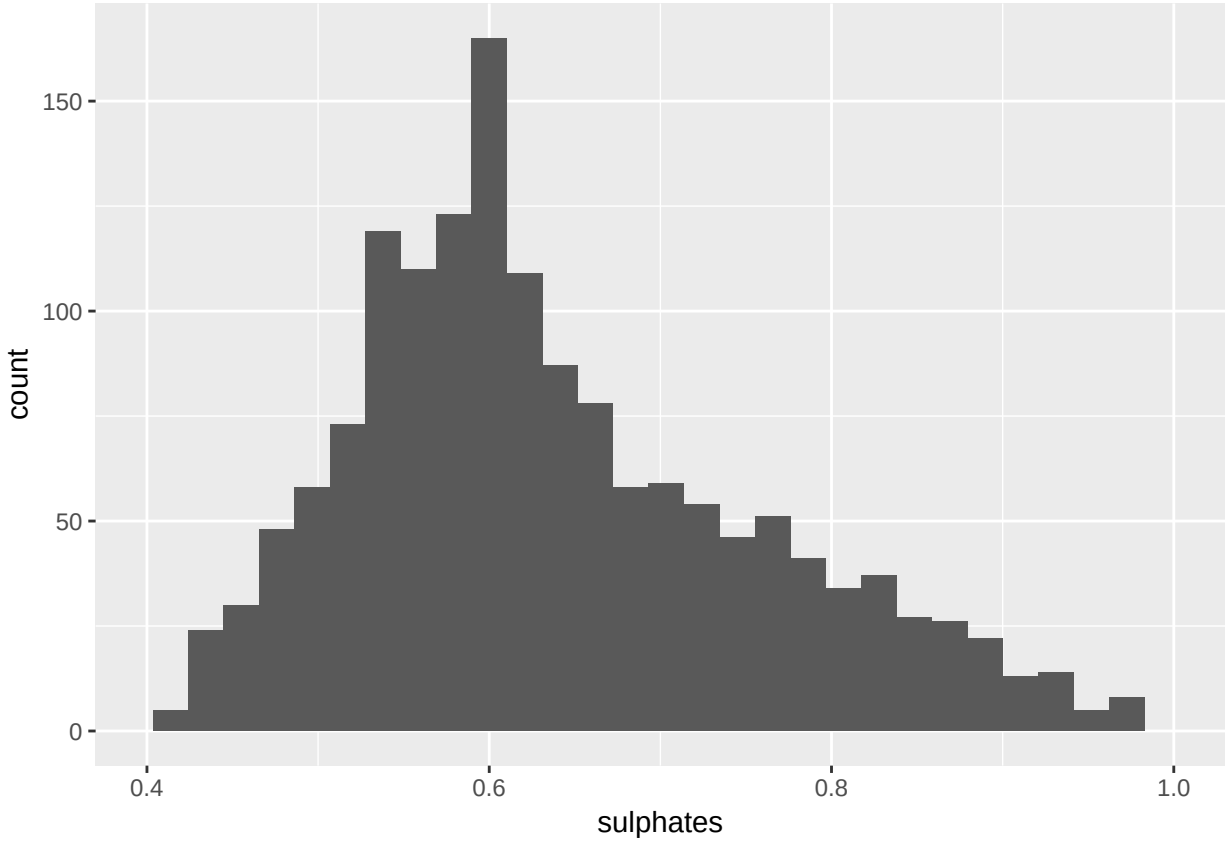


Similar with free.sulfur.dioxide, the distribution of the total mount of them also tends to focus on a relative lower value. Most of wines have a concentration of total.sulfur.dioxide below 65 (mg / dmˆ3).

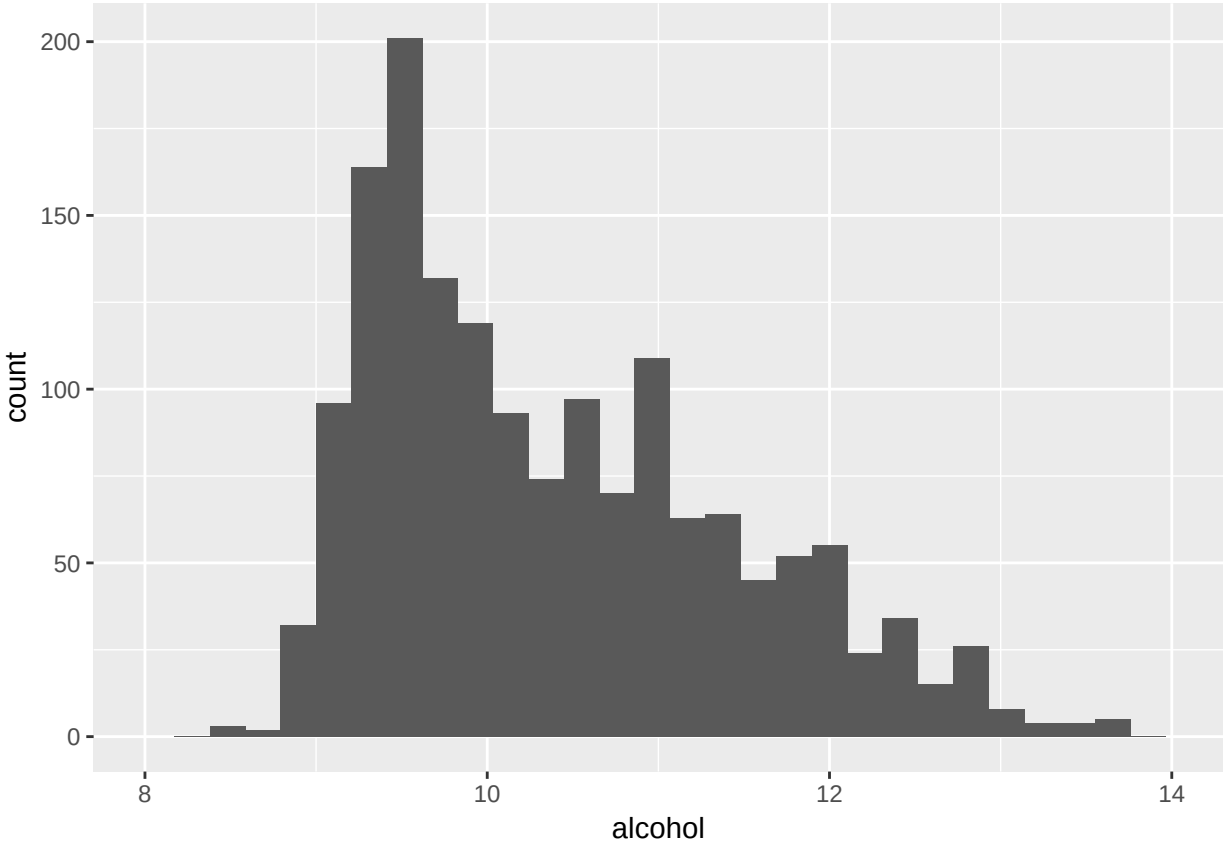

150
100
50
0
count
0.9900
0.9925
0.9950
0.9975
1.0000
1.0025
density


200
150
100
50
0
count
3.0
3.5
4.0
pH

Again, with default binwidth,the distribution of density and pH are almost a perfect normal distribution, without any rescale on axis or limits on original data.



After adjustments, the distribution of sulphates seems a little right skewed. More wines tended to have a concentration of sulphates between 0.5 and 0.7 (potassium sulphate - g / dm3).



Also, without any rescale on axis or limits on original data, the distribution of alcohol seems good.

Univariate Analysis

What is the structure of your dataset?

There are 1599 sample of wines in the dataset with 12 features (fixed.acidity,volatile.acidity,citric.acid,residual.sugar,chlorides,free
There is no facoter variables. Other observations:

- Most wines' quality rated as '5' or '6'.
- There is no wine's rate is lower than '3', or higher than '8', and the wines which are rated as '3' or '8' are also very limited.
- Most wines' pH are between 3 and 3.5, a relative narrow range.
- The distribution of density is almost a normal distribution. About 75% of wines' density are between 0.9956 and 0.9978.
- The median quality for a wine is '6' and the max quality is 8.

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

The main features in the data set are quality. I'd like to determine what kind of features are best for predicting the quality of a wine. At this stage, I suspect volatile.acidity, free.sulfur.dioxide, pH and some combination of the other variables can be used to build a predictive model to the quality of wines.

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

fixed.acidity, citric.acid, residual.sugar, chlorides, density, sulphates and alcohol likely contribute to the quality of wines.

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

No yet, for this stage.

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?

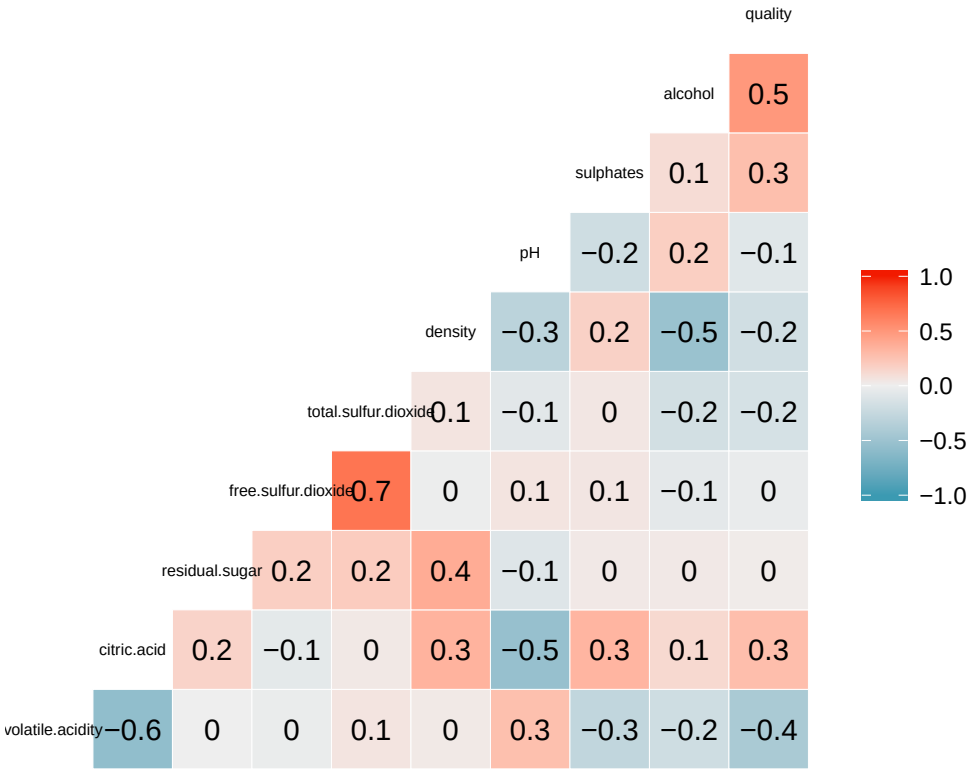
First, in the histogram graph of quality, the whole data seems to be too discreted, which made me hard to tell the partern in it. I transformed the very discreted data in a better format by adjusting binwidth and breaks the x coordinates. Thenm in the histogram of citric.acid, It's clear to see that the outliers on the rightest of the graph, and the distribution of citric.acid is left skewed. I printed out the table of citric.acid, we can see that the main body of this data are from 0 to around 0.7. Thereforem, I tranformed this graph by changing the binwidth and the breaks and limits on x coordinaetes.

Bivariate Plots Section

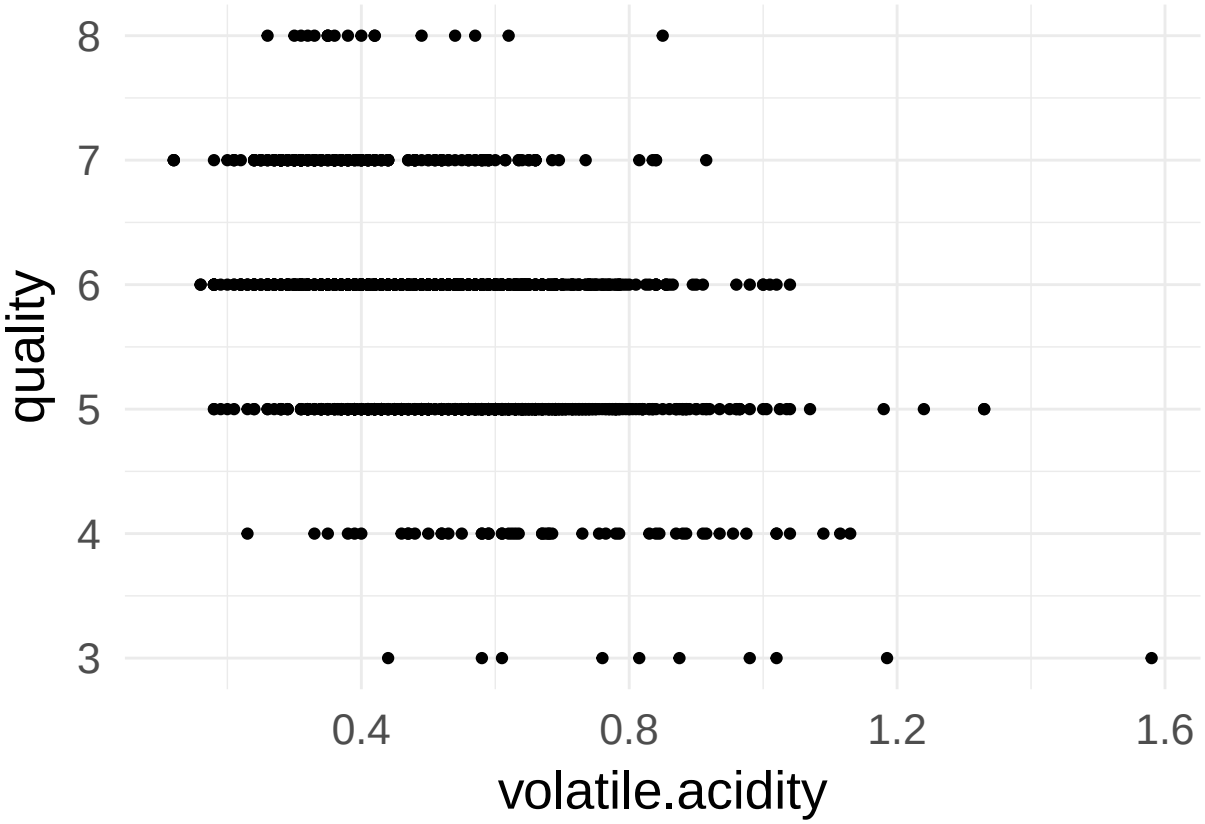
```
##                       volatile.acidity citric.acid residual.sugar
## volatile.acidity           1.000000000 -0.55249568    0.001917882
## citric.acid               -0.552495685  1.00000000    0.143577162
## residual.sugar             0.001917882  0.14357716    1.000000000
## free.sulfur.dioxide       -0.010503827 -0.06097813    0.187048995
## total.sulfur.dioxide       0.076470005  0.03553302    0.203027882
## density                    0.022026232  0.36494718    0.355283371
## pH                         0.234937294 -0.54190414   -0.085652422
## sulphates                 -0.260986685  0.31277004    0.005527121
## alcohol                   -0.202288027  0.10990325    0.042075437
## quality                   -0.390557780  0.22637251    0.013731637
##                       free.sulfur.dioxide total.sulfur.dioxide     density
## volatile.acidity              -0.01050383           0.07647000  0.02202623
## citric.acid                   -0.06097813           0.03553302  0.36494718
## residual.sugar                 0.18704900           0.20302788  0.35528337
## free.sulfur.dioxide            1.00000000           0.66766645 -0.02194583
## total.sulfur.dioxide           0.66766645           1.00000000  0.07126948
## density                       -0.02194583           0.07126948  1.00000000
## pH                             0.07037750          -0.06649456 -0.34169933
## sulphates                      0.05165757           0.04294684  0.14850641
## alcohol                       -0.06940835          -0.20565394 -0.49617977
## quality                       -0.05065606          -0.18510029 -0.17491923
##                                     pH    sulphates     alcohol     quality
## 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
## free.sulfur.dioxide    0.07037750  0.051657572 -0.06940835 -0.05065606
## total.sulfur.dioxide -0.06649456  0.042946836 -0.20565394 -0.18510029
```

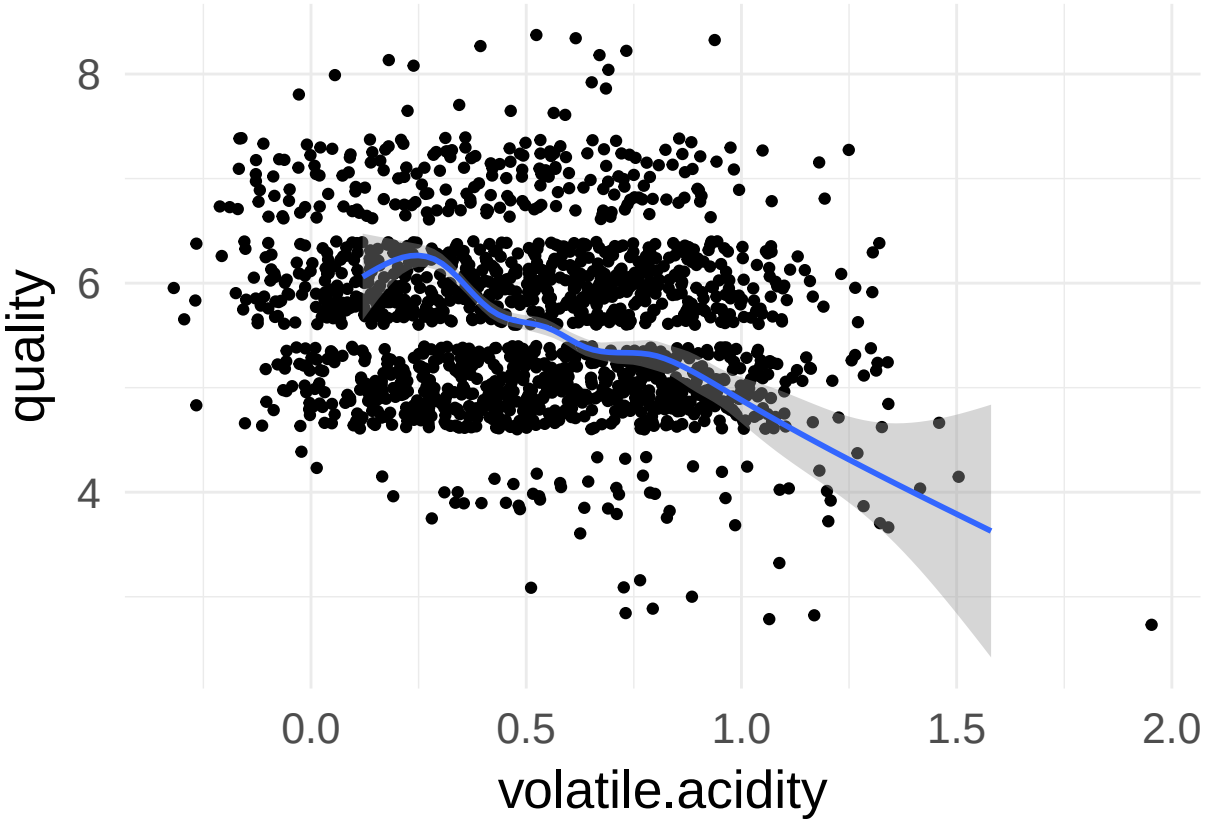
```
## density              -0.34169933  0.148506412 -0.49617977 -0.17491923
## pH                    1.00000000 -0.196647602  0.20563251 -0.05773139
## sulphates            -0.19664760  1.000000000  0.09359475  0.25139708
## alcohol               0.20563251  0.093594750  1.00000000  0.47616632
## quality              -0.05773139  0.251397079  0.47616632  1.00000000
```

I used a subset of 10 variables to make the correlation matrix and scatterplot matrix. I didn't include the variables X(id variable), fixed.acidity, chlorides.



From the subset of data and its correlation matrix and scatterplot matrix, residual.sugar, free.sulfur.dioxide, pH do not seem to have strong correlations with quality, but total.sulfur.dioxide and density are moderately correlated with alcohol, the most correlated variable with quality. I want to look closer at scatter plots involving quality and some other variables like volatile.acidity, citric.acid,total.sulfur.dioxide,density,sulphates and alcohol.

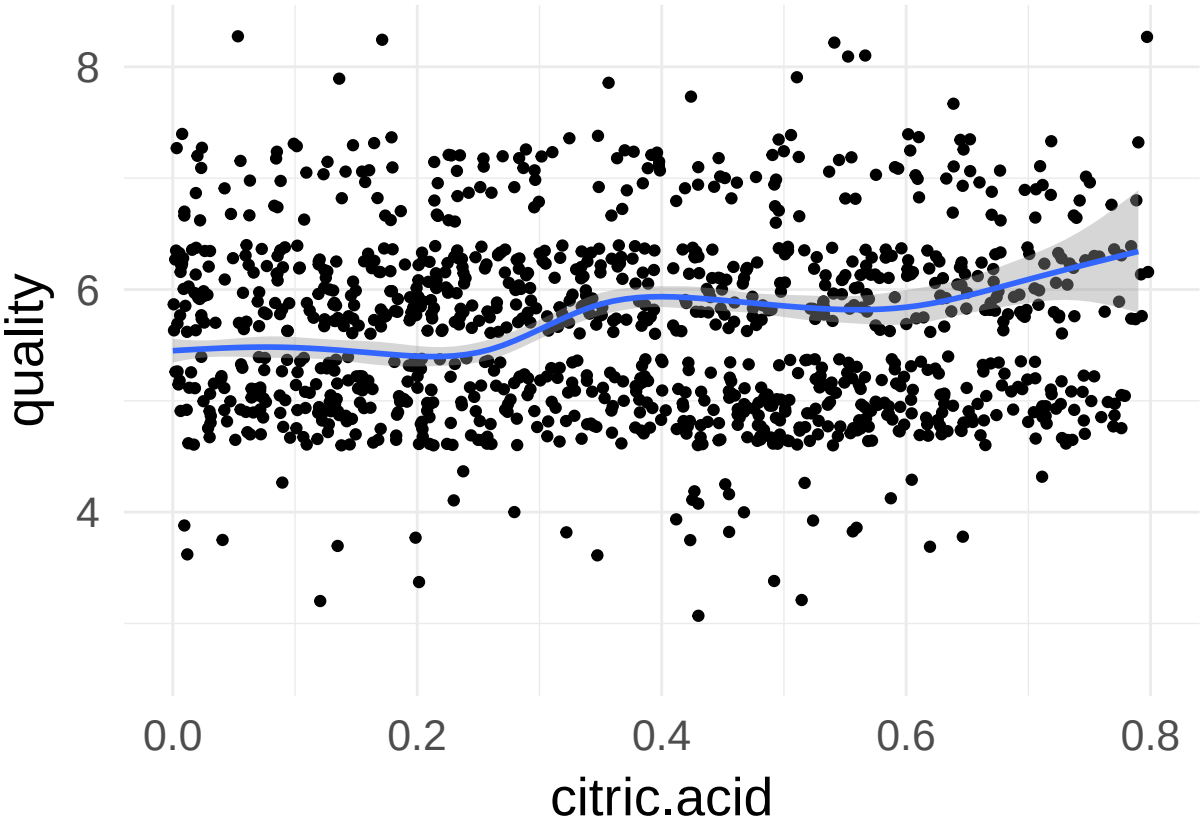

quality
8
7
6
5
4
3
0.4
0.8
1.2
1.6
volatile.acidity


quality
8
6
4
0.0
0.5
1.0
1.5
2.0
volatile.acidity

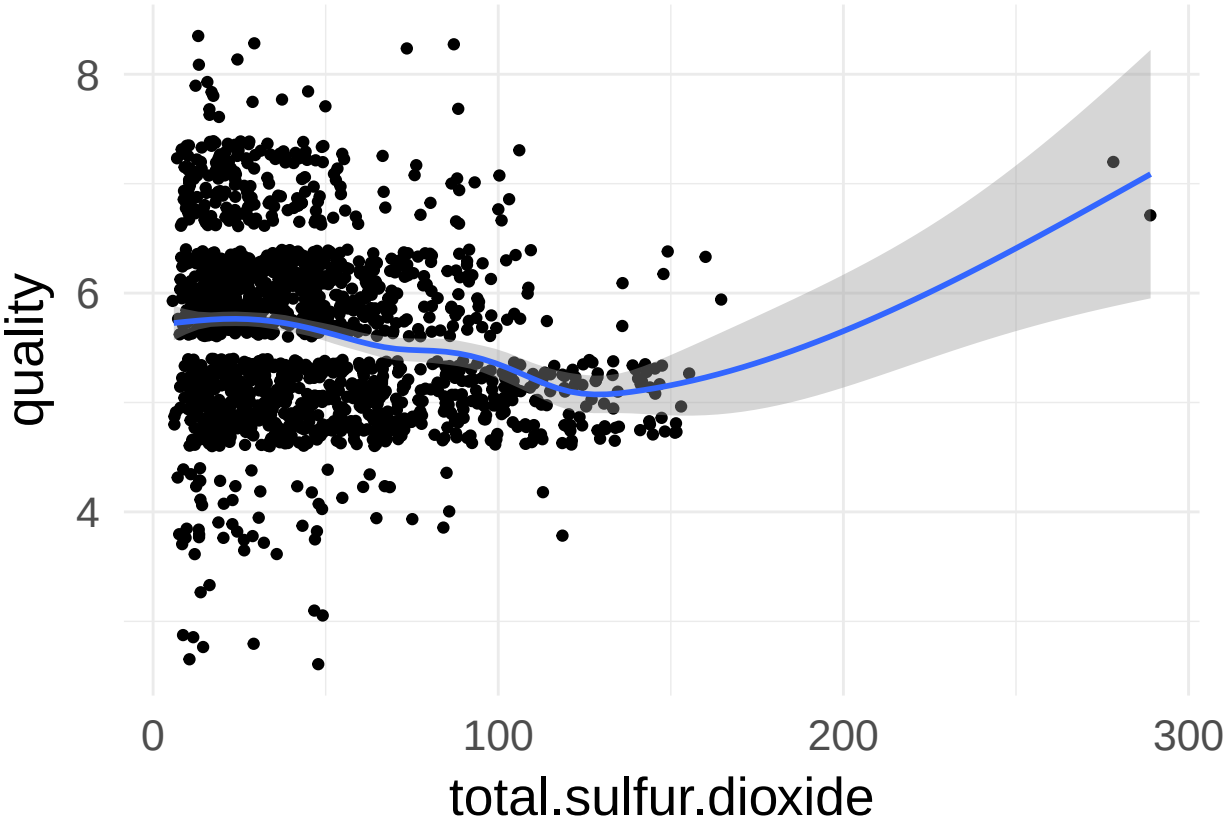
##

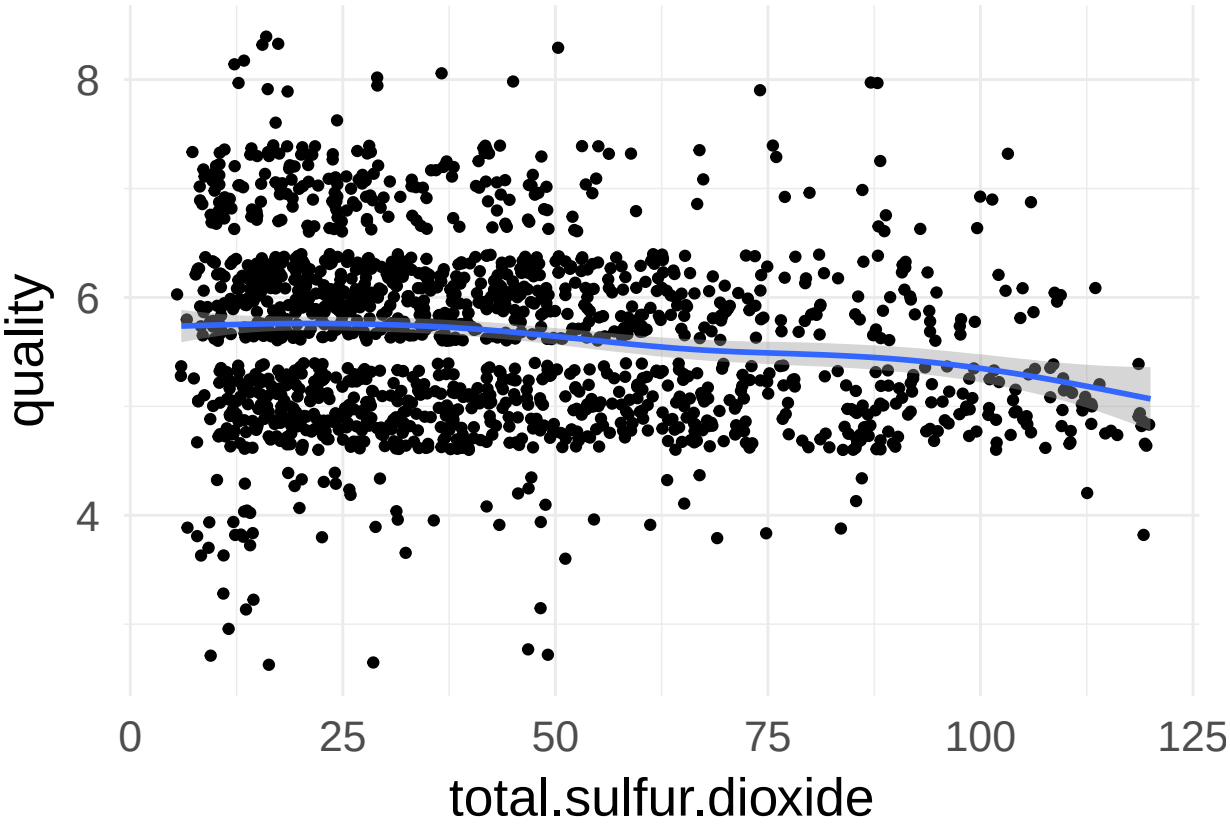
```
## Call:
## lm(formula = quality ~ volatile.acidity, data = red)
##
## Coefficients:
##      (Intercept)  volatile.acidity
##            6.566            -1.761
```

Comparing volatile.acidity to quality. The first plot suffers from quality values being integers and some overplotting. I added jitter, changed its width and height and added smoother to the graph. Finally, we can see the correlation bwtween volatile.acidity and quality, even thuogh it's not very strong.

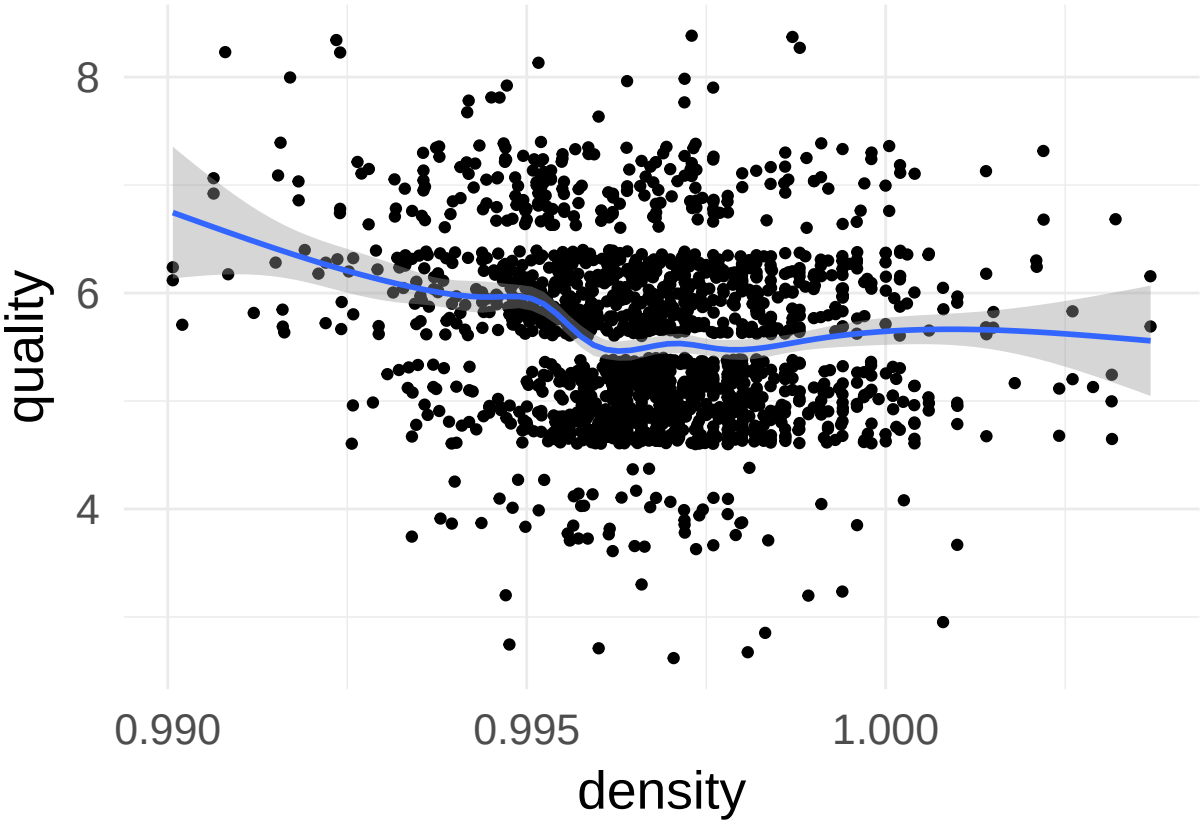


From the graph, it seems the correlation between citric.acid and quality is not so strong.

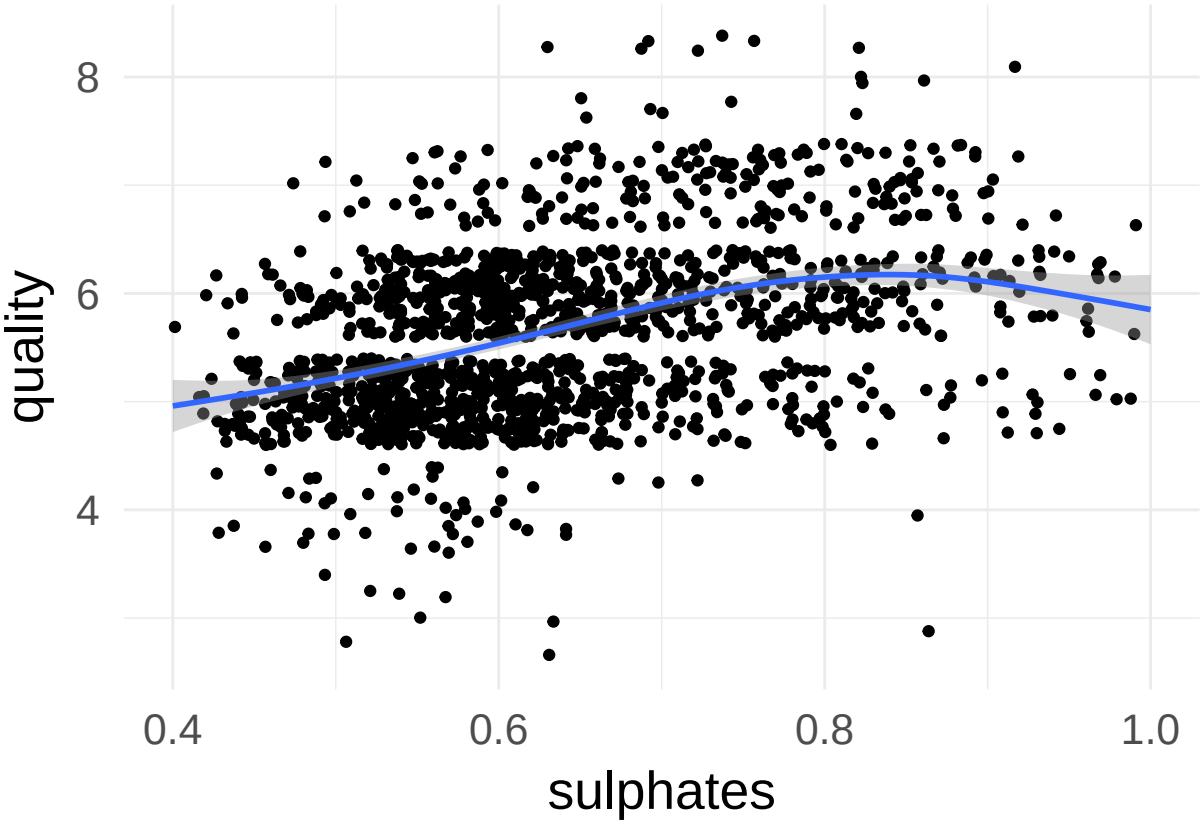




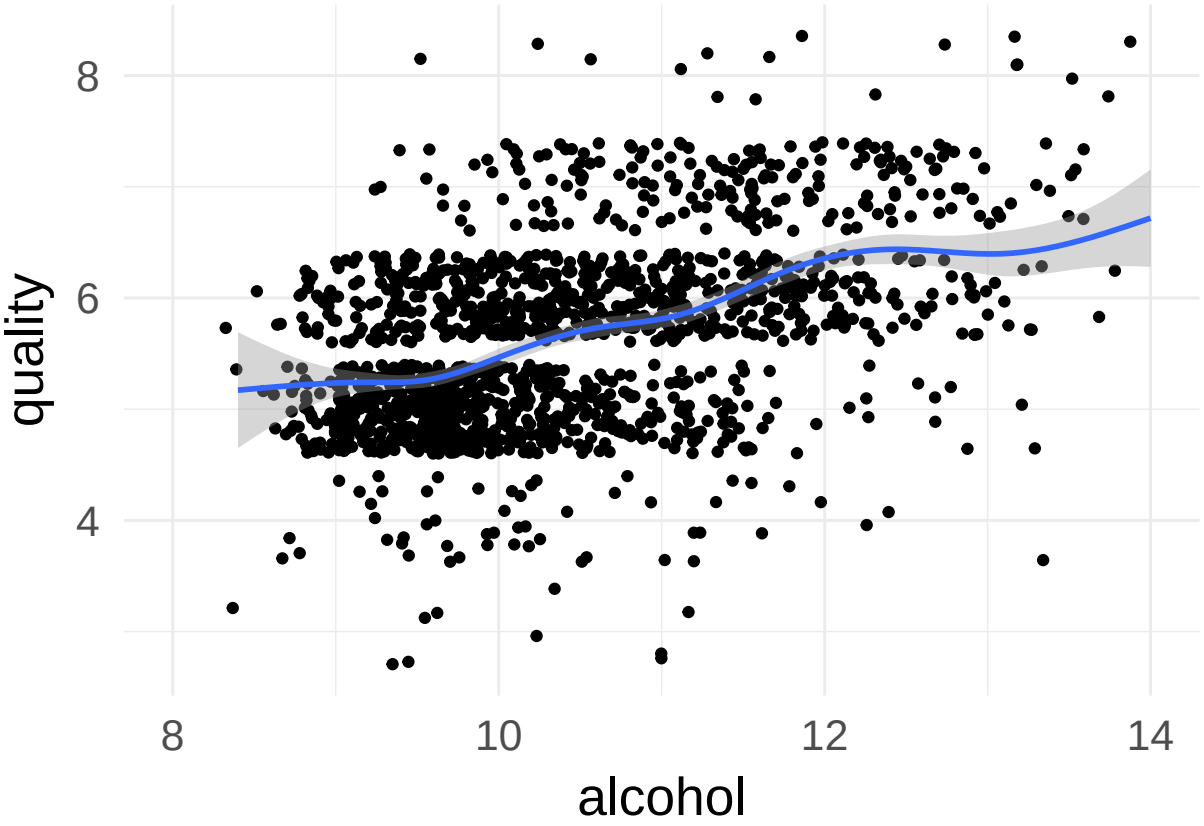
The first graph suffers from outliers, therefore I need to refine this graph by adding limitations on x coordinates.



From the graph, it can tell that density actually has a relative poor correlation with quality. However, this result is consist with its relative low correlation coefficient.

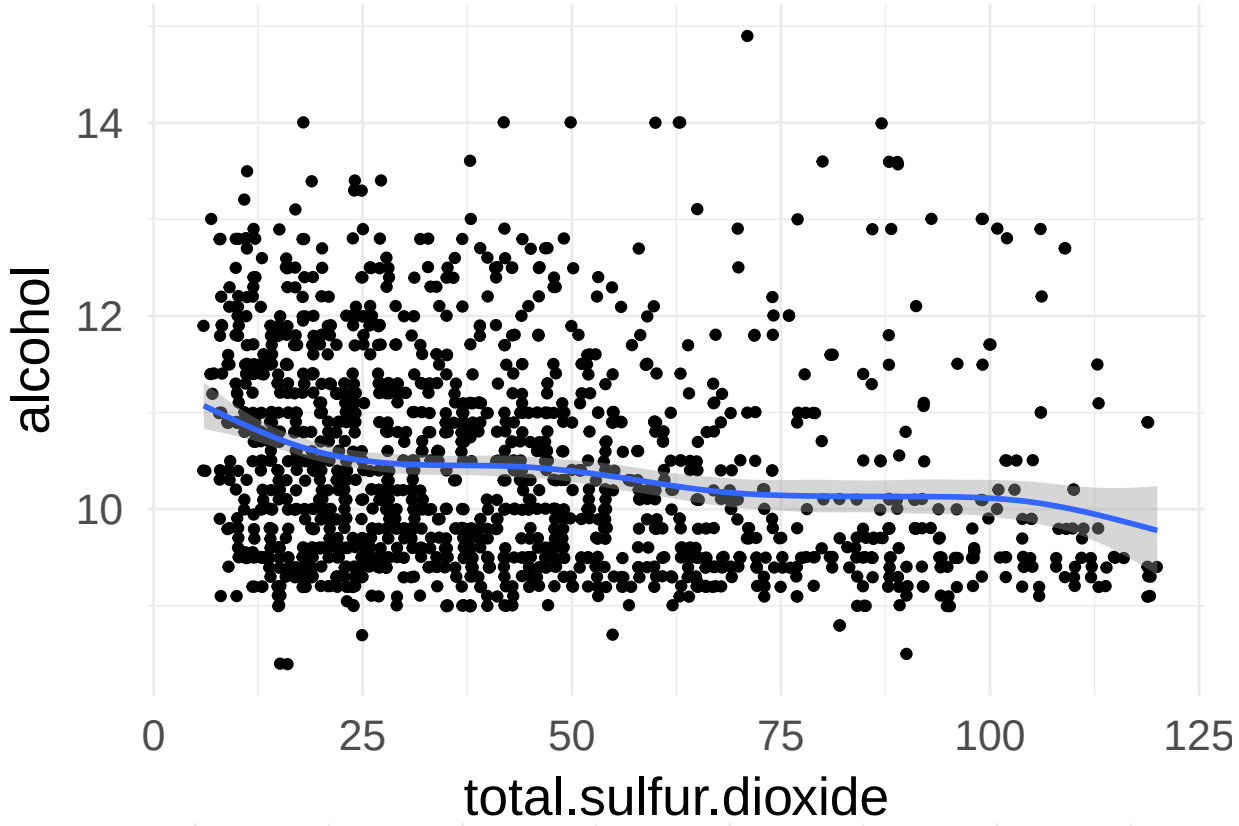


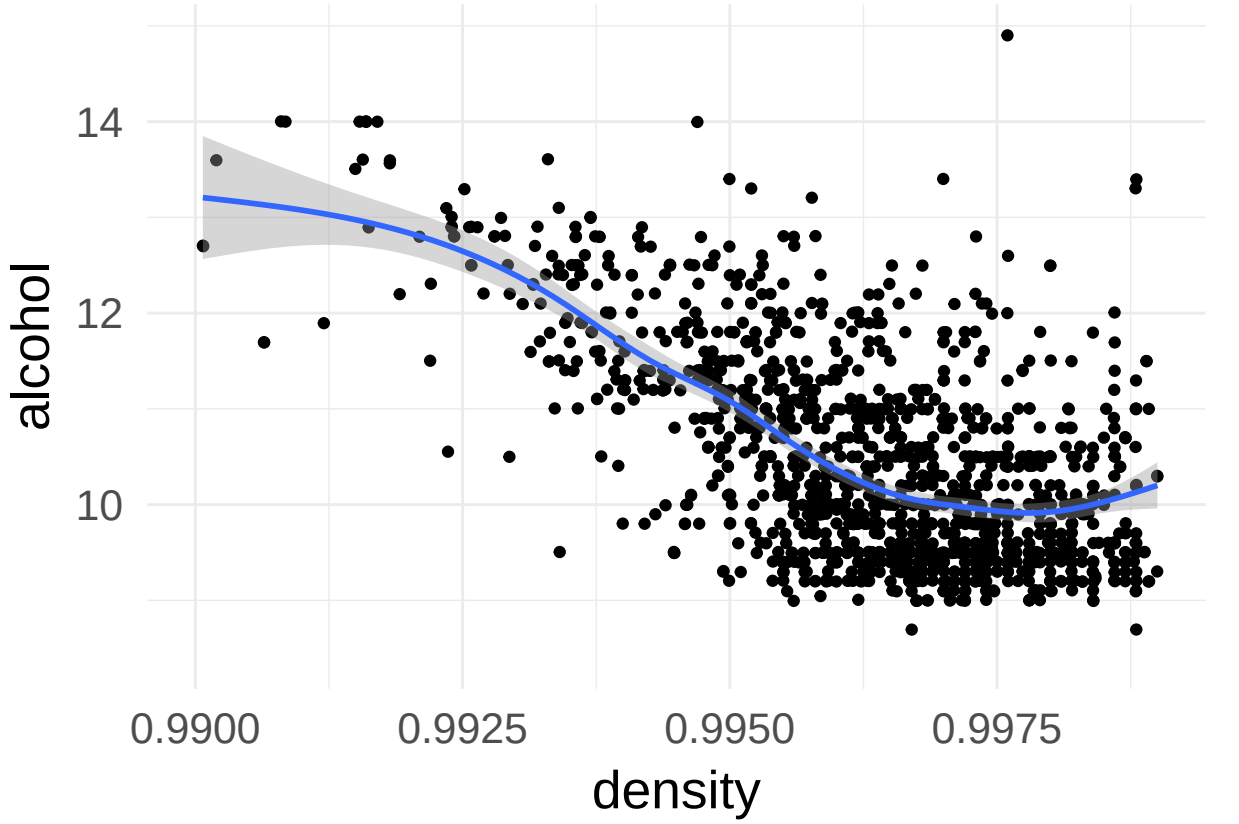

```
##
## Call:
## lm(formula = quality ~ sulphates, data = red)
##
## Coefficients:
## (Intercept)    sulphates
##       4.848        1.198
```



From these plots above, it's clearly to see and verify that the concentration of alcohol has the most correlated relationship with the wines quality.

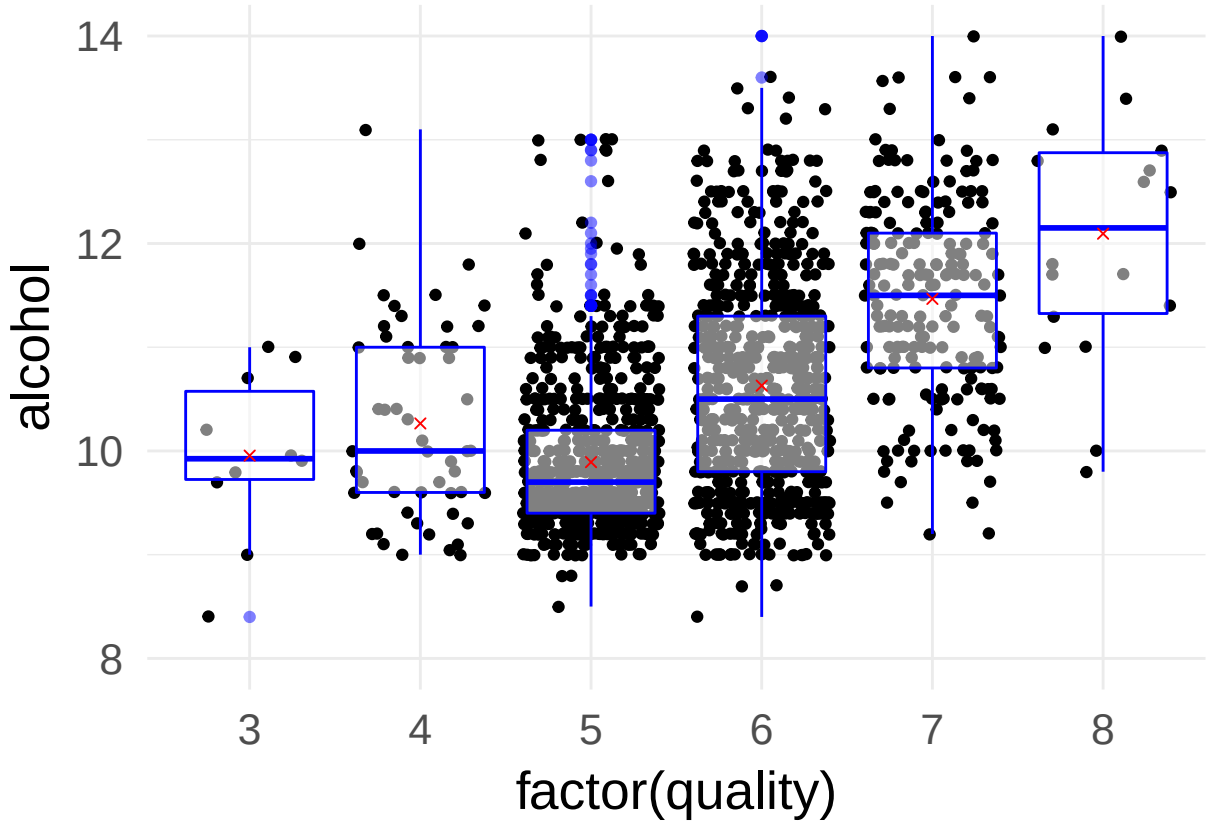
Starting explore variables(total.sulfur.dioxide and density) having potential correlation with alcohol, the most correlated variable with quality.



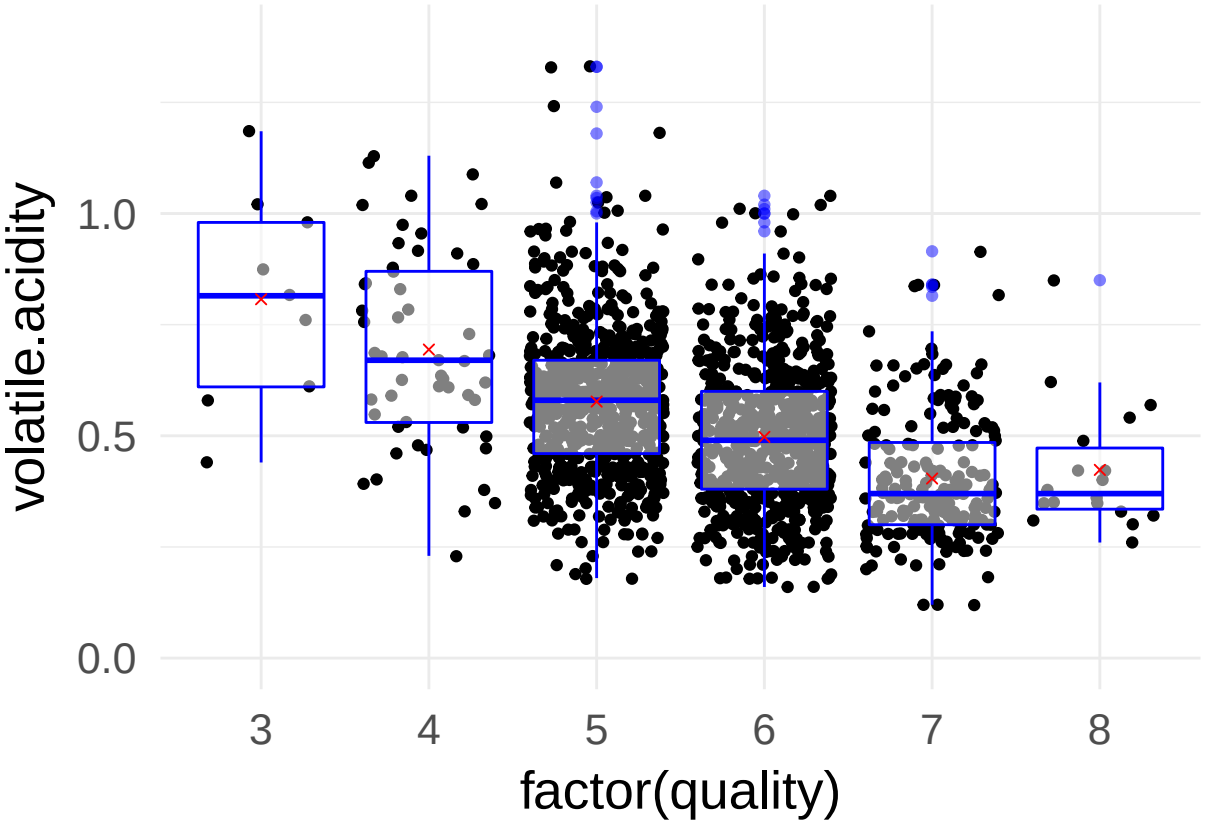


From these two plots, we can find that total.sulfur.dioxide and density both have some correlations with the concentration of alcohol. And obviously density have a much stronger correlation with alcohol. Next,Let's

concentrate on alcohol and volatile.acidity with quality.




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




```
## factor(red$quality): 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4400  0.6475  0.8450  0.8845  1.0100  1.5800
## ------------------------------------------------------------
## factor(red$quality): 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.230   0.530   0.670   0.694   0.870   1.130
## ------------------------------------------------------------
## factor(red$quality): 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.180   0.460   0.580   0.577   0.670   1.330
## ------------------------------------------------------------
## factor(red$quality): 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1600  0.3800  0.4900  0.4975  0.6000  1.0400
## ------------------------------------------------------------
## factor(red$quality): 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3000  0.3700  0.4039  0.4850  0.9150
## ------------------------------------------------------------
## factor(red$quality): 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2600  0.3350  0.3700  0.4233  0.4725  0.8500
```

The boxplots and summary matrixs above showed and demonstrated two most strong correlation with quality among all the 11 chemical components. While volatile.acidity negatively correlated with quality, alcohol showed a strong and positive correlation 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?

Quality correlates strongly with volatile.acidity and alcohol.

Generally, as the concentration of volatile.acidity in the wine increases, the quality of the wine in price increases. However, alcohol concentration has a more strong, and conversely, positive correlation with quality of the wine.

And, the quality of wines also tends to be positively correlated with citric.acid and sulphates. Although, they has relative poor correlation with the quality, we still can tell the positive correlated pattern from their graph.

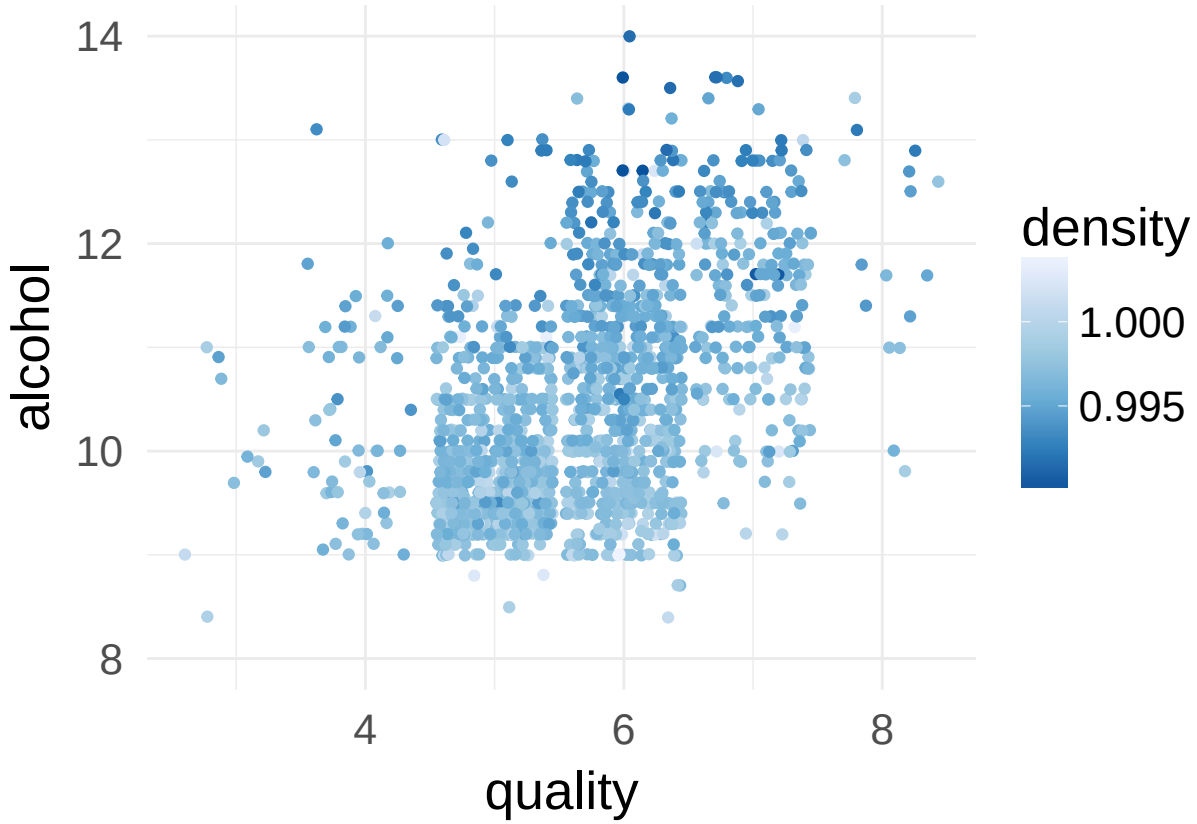
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

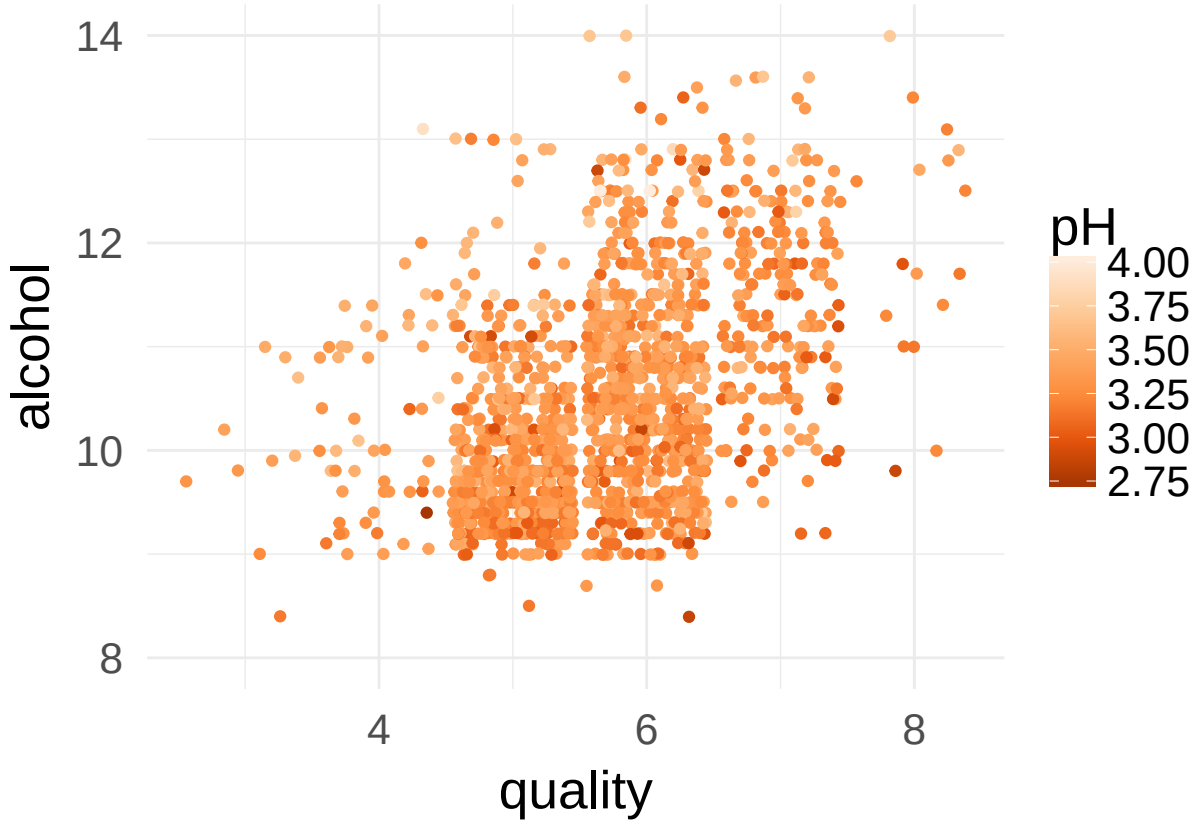
The concentration of alcohol tends to correlated with density and total.sulfur.dioxide, which are both positive correlated with alcohol. And it seems that density has a more strong correlation.

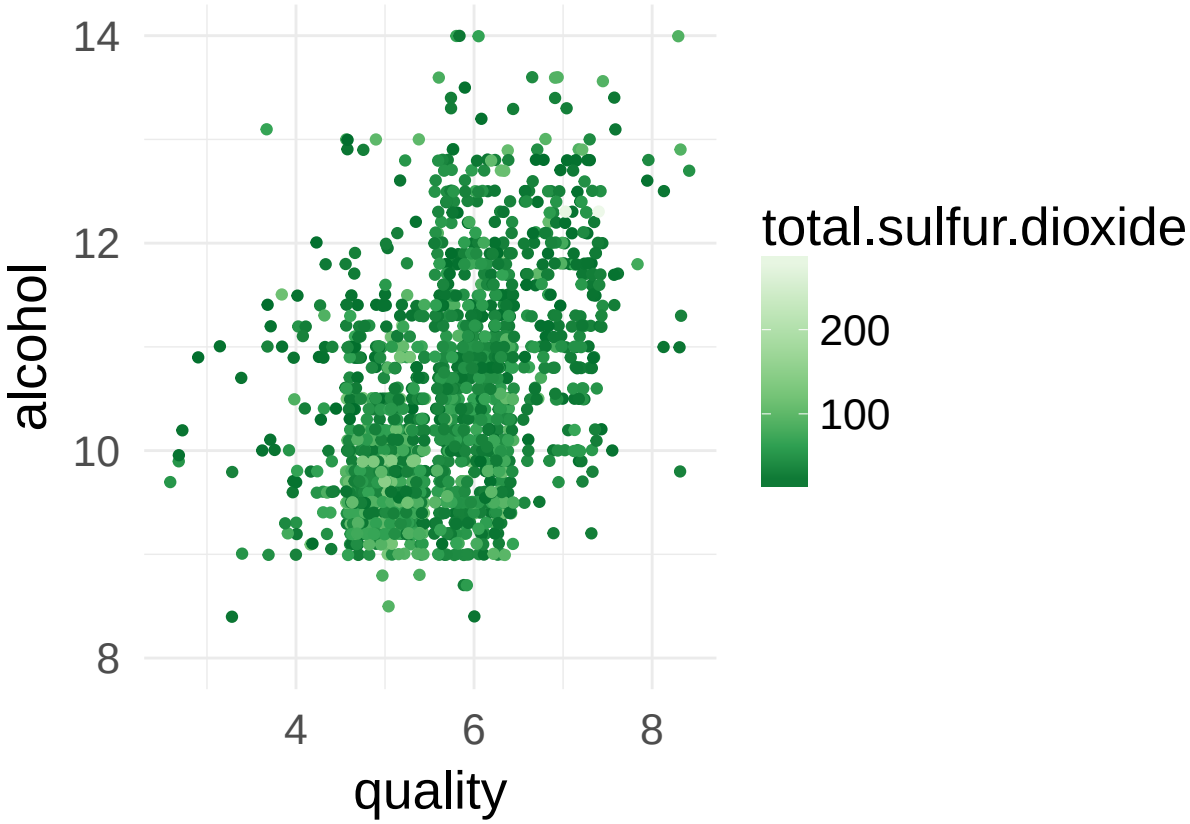
What was the strongest relationship you found?

The quality of wines is positively and strongly correlated with the alcohol concentration of wines. The variables volatile.acidity also correlate with the price, but less strongly than alcohol, and it's a negative correlation. Either alcohol or volatile.acidity could be used in a model to predict the quality of wines.

Multivariate Plots Section

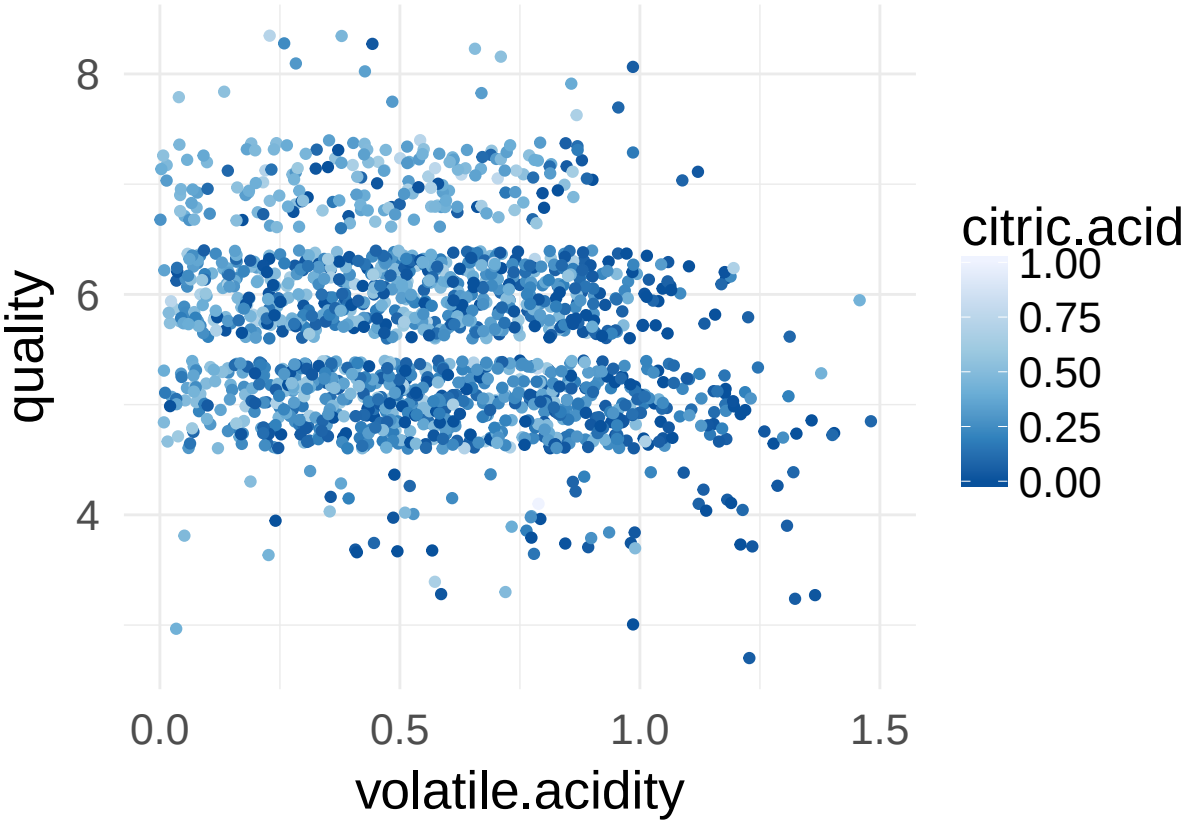


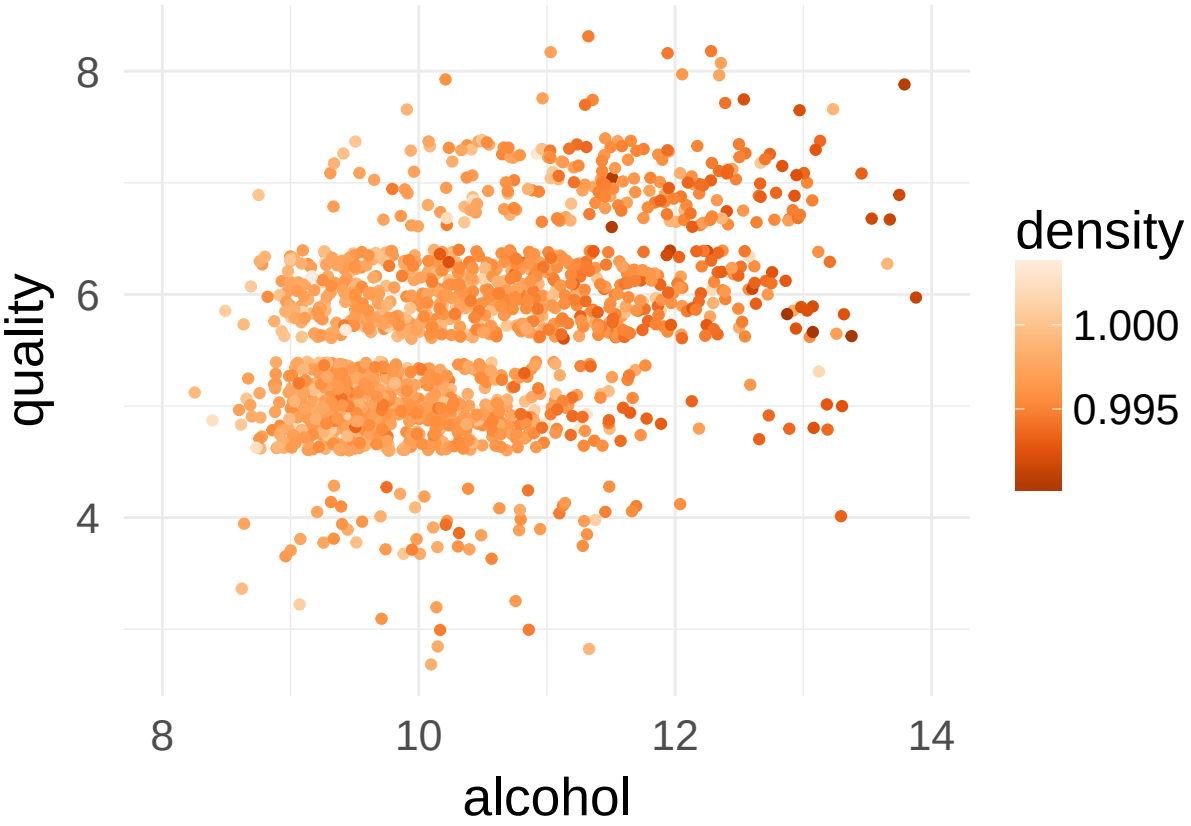




I plotted out the scatterpoint plot of alchohol vs quality, which colored by density,pH,total.sulfur.dioxide respectively. Those three variables are the most correlated ones with alcohol. The first plot, colored by density, is the best plot. We can easily tell that the wines with higher quality tend to have higher alcohol concentration and lower density. Though the later two plots have no such good and clear correlation pattern as the first plot have, Combining three of them, we still can recognize that most the high qualified wines are more likely to have relative higher alcohol concentration, lower density, higher pH and lower total.sulfur.dioxide concentration.

Next, I'll adding some most pormising variables(other chemical components) as 'continuous category' variable into scatterplots, finding their combined correlation with quality, and trying to build a comprehension predict model.





Here, I picked the most promising and correlated with alcohol and volatile.acidity variables respectively, and adding them as 'continuous category' variables into the original scatterplots. According to these plots, the

patterns are relative easy to recognize.

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?

Generally, the wines with high quality contain relative high concentration alcohol and citric.acid, low concentrate volatile.acidity, and have relative low density.

If we account for constant concentration of volatile.acidity, better quality of wine usually contains higher concentration of citric.acid. And the wine with better quality tend to have lower density holding concentration of alcohol constant.

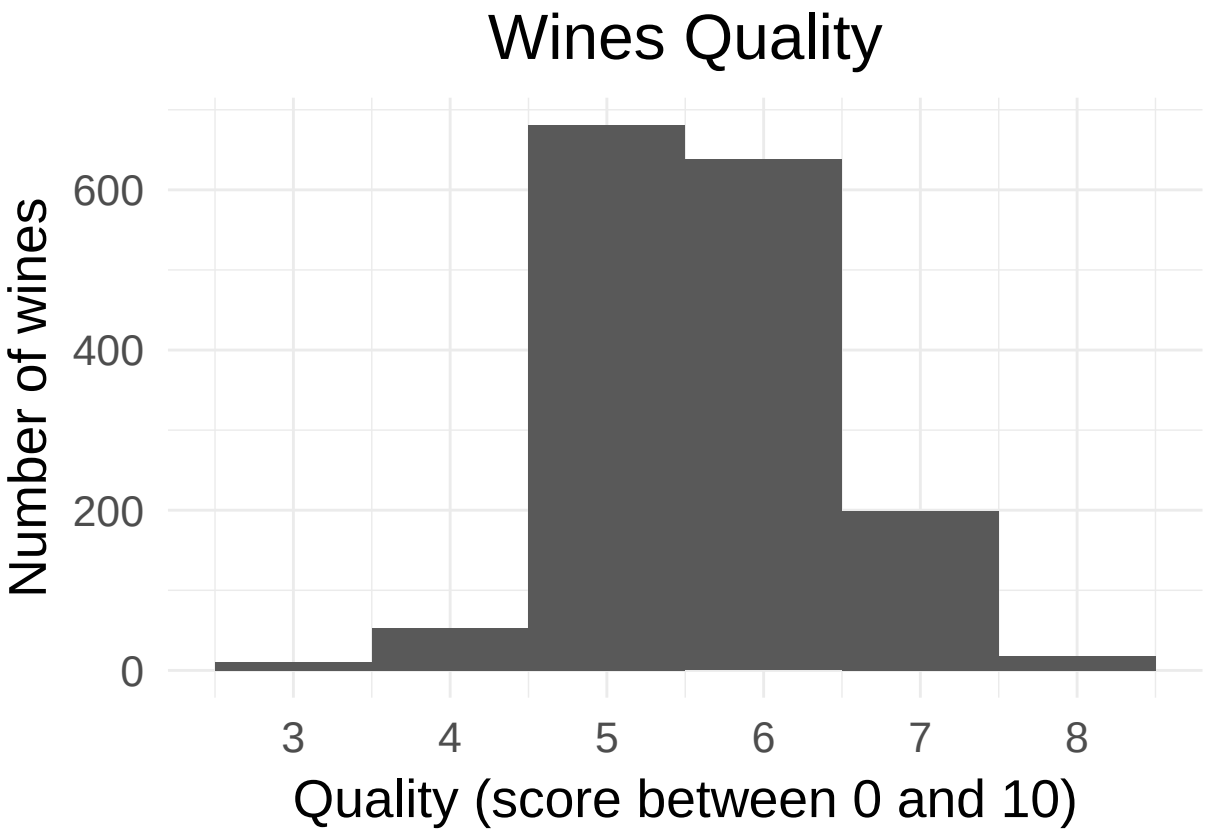
Were there any interesting or surprising interactions between features?

At the Univariate Plots Section, I though that free.sulfur.dioxide and pH would be most correlated variables with quality, according to my father experience. However, after later exploration, it seems that those two variables are relative irrelevant to the predict model of wines' quality. Only pH is somehow correlated with alcohol concentration, the most correlated variable with quality.

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

Final Plots and Summary

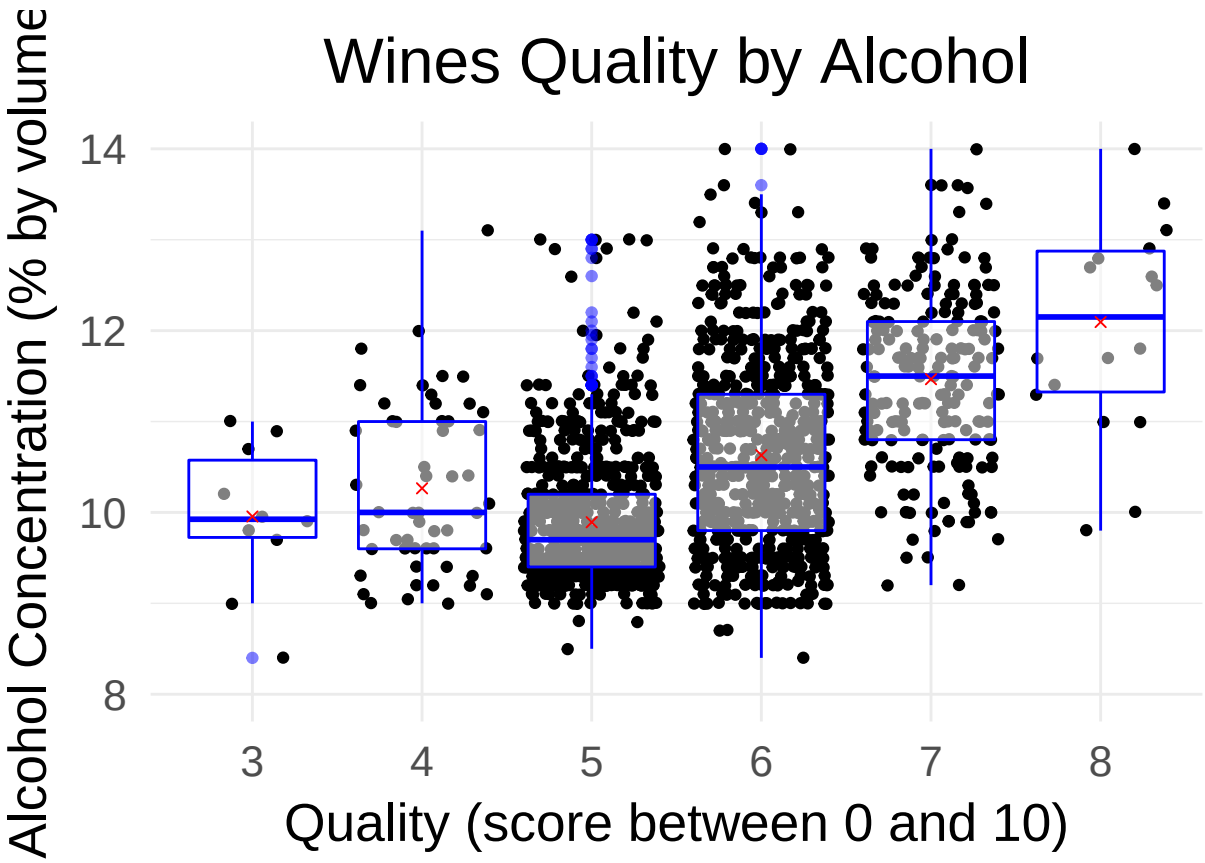
Plot One



Description One

Most of wines are rated as '5' or '6', a median rate. There is no wine's rate is lower than '3', or higher than '8', and the wines which are rated as '3' or '8' are also very limited.

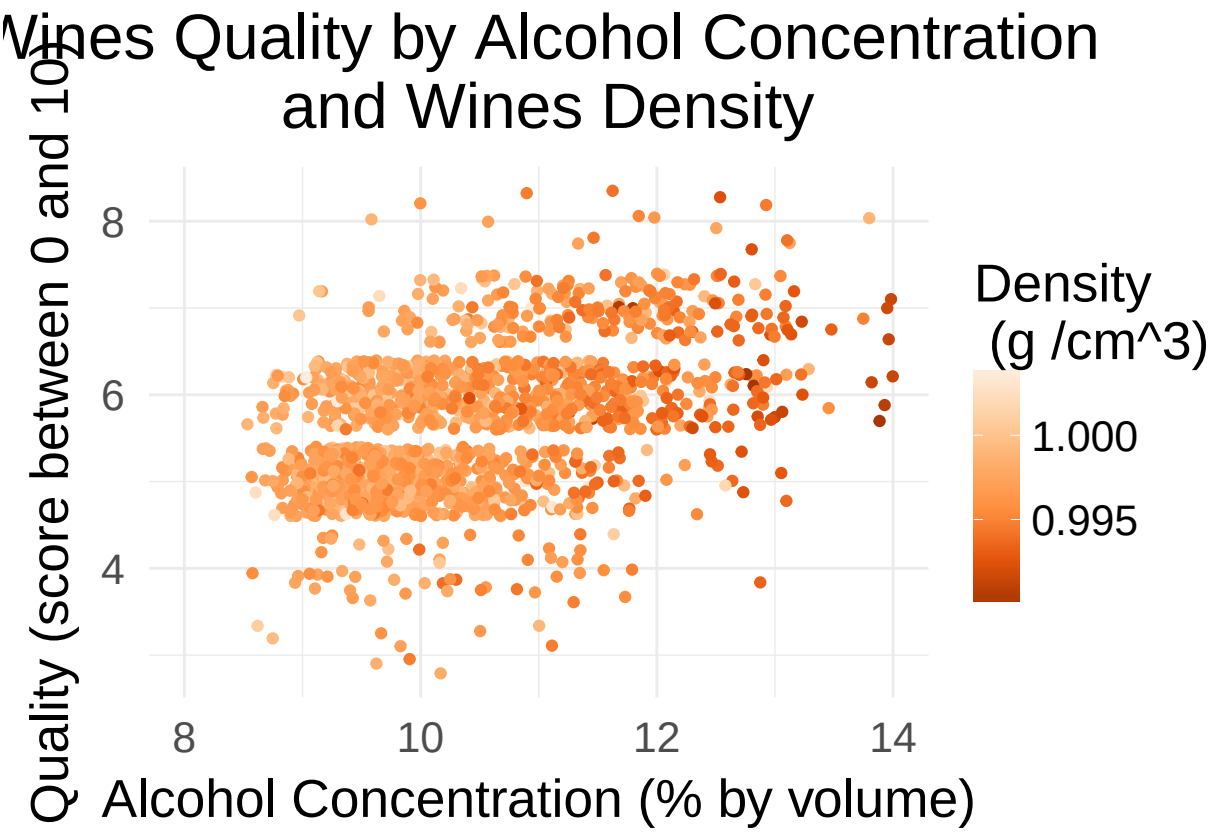
Plot Two



Description Two

The most strong correlation with quality among all the 11 chemical components, Alcohol showed a relatively strong and positive correlation with quality.

Plot Three



Description Three

Generally, the wines with high quality contain relative high concentration alcohol,and have relative low density.

Reflection

The red wine quality data set contains information 1,599 red wines with 11 variables on the chemical properties of the wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent). I started by understanding the individual variables in the data set, and then I explored interesting questions and leads as I continued to make observations on plots. Eventually, I explored the quality of wines across many variables and roughly bulit up a predict model by the two of most correlated variables with quality.

There was a clear trend between the concentration of alcohol and volatile.aciity of wines and its quality. I was surprised that free.sulfur.dioxide and pHdid not have a strong positive correlation with price, but pH is somehow correlated with alcohol concentration, the most correlated variable with quality. I also picked the most promising and correlated with alcohol and volatile.acidity variables respectively, and adding them as 'continuous category' variables into the original scatterplots. The results showed that the wines with high quality contain relative high concentration alcohol and citric.acid, low concentrate volatile.acidity, and have relative low density.

Some limitations of this model include the source of the data. The size of the dataset are not big enough to conduct very persuasive and not ensured updated. To investigate this data further,I would be interested in developing a linear model to predict current red wines quality and to determine to what extent the model is accurate at rating wines.

Personally, I would say making plots with clear relationships between variables and consisted with the standard norms are some of the most challenging parts of this projects. I did spend a lot of time on this aspect. Of course, the analysis of the whole dataset also are somehow tricky, but I could manage it by following my thoughts of flow, as long as I maintain enough concentration, I can make this part get through. Finally, I learnt a lot from completing this project. Problem solving and making the results to be consisted with standard norms are probably the most impressive parts.