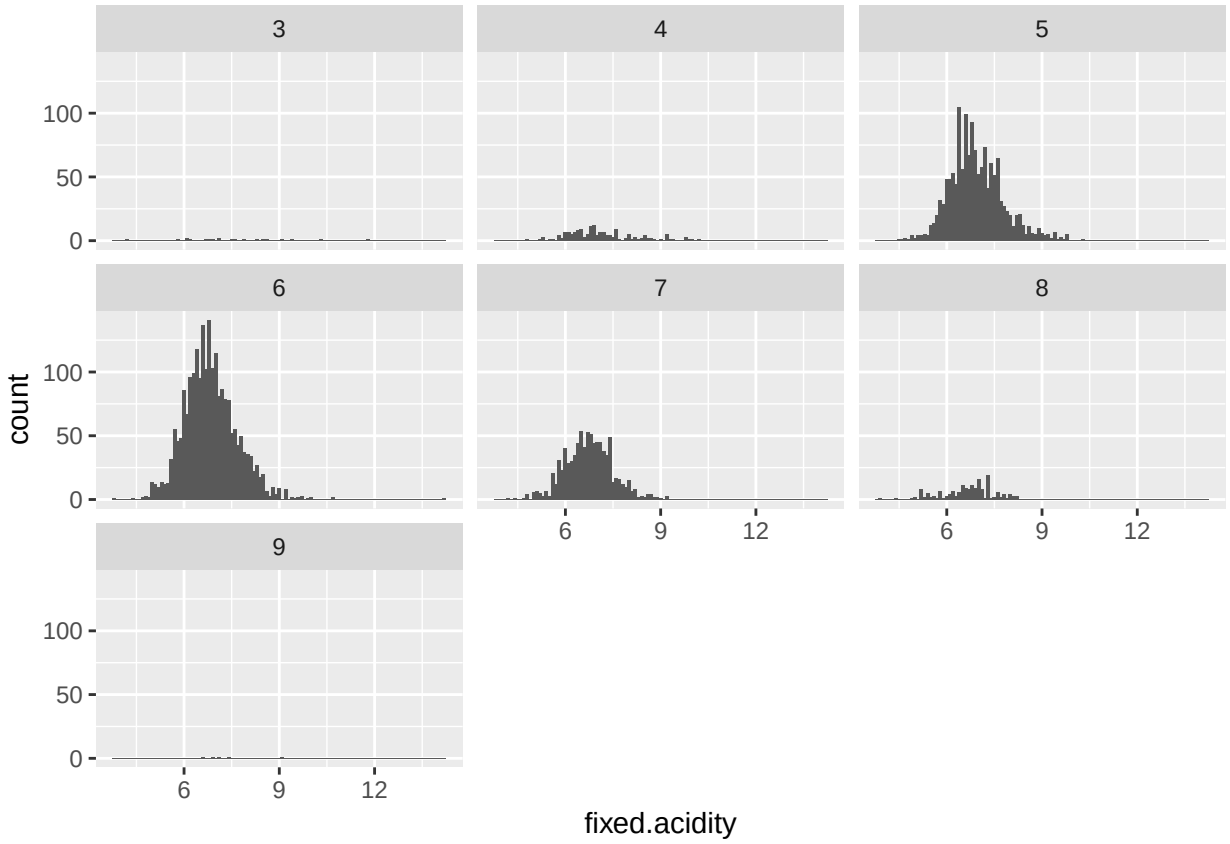
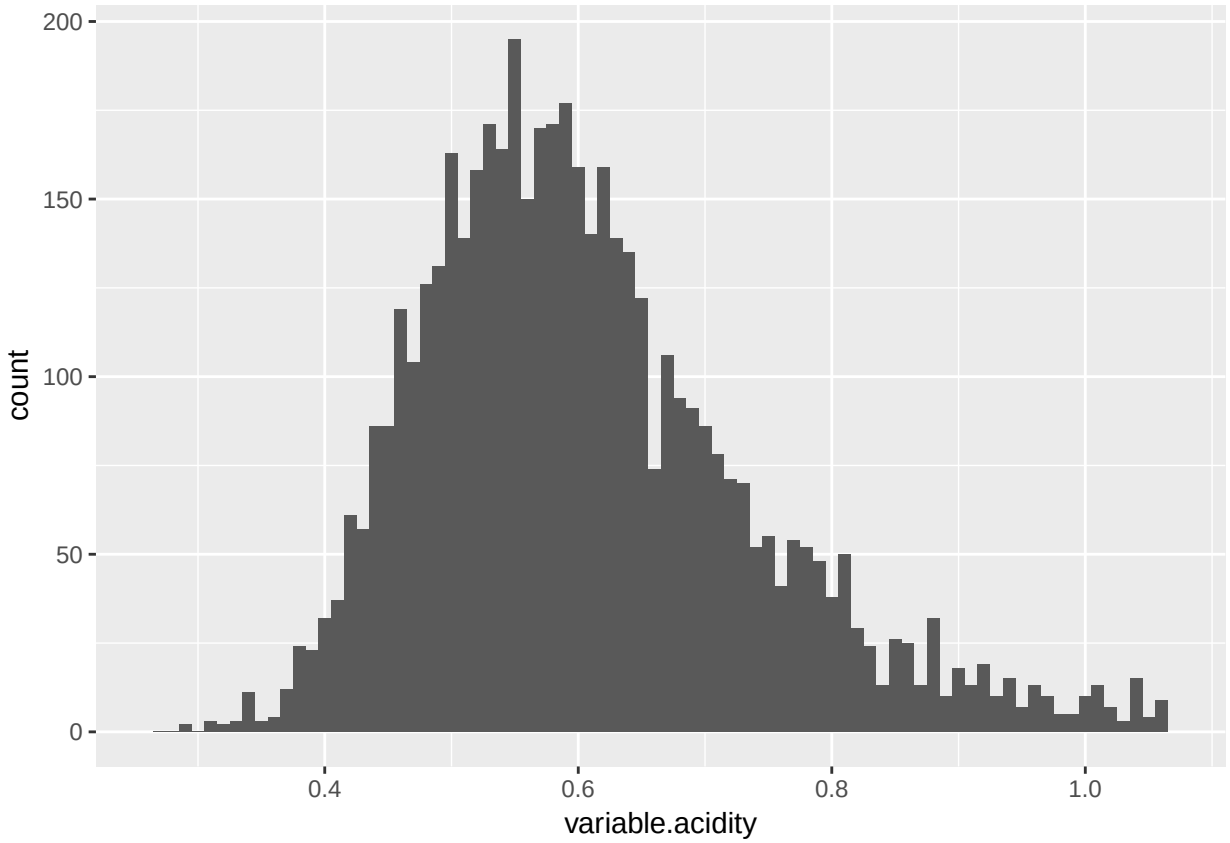
Properties Influencing White Wine by Abhishek Singh

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

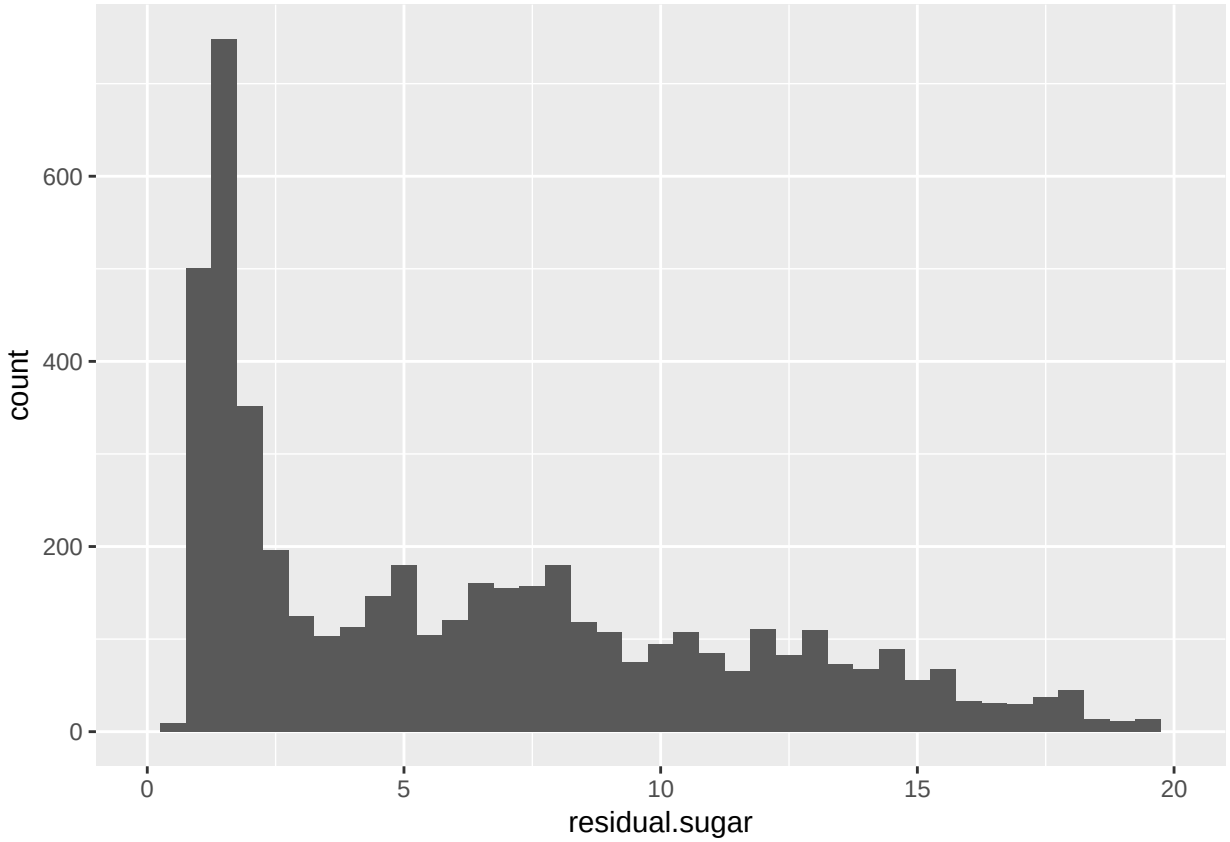
Univariate Plots Section

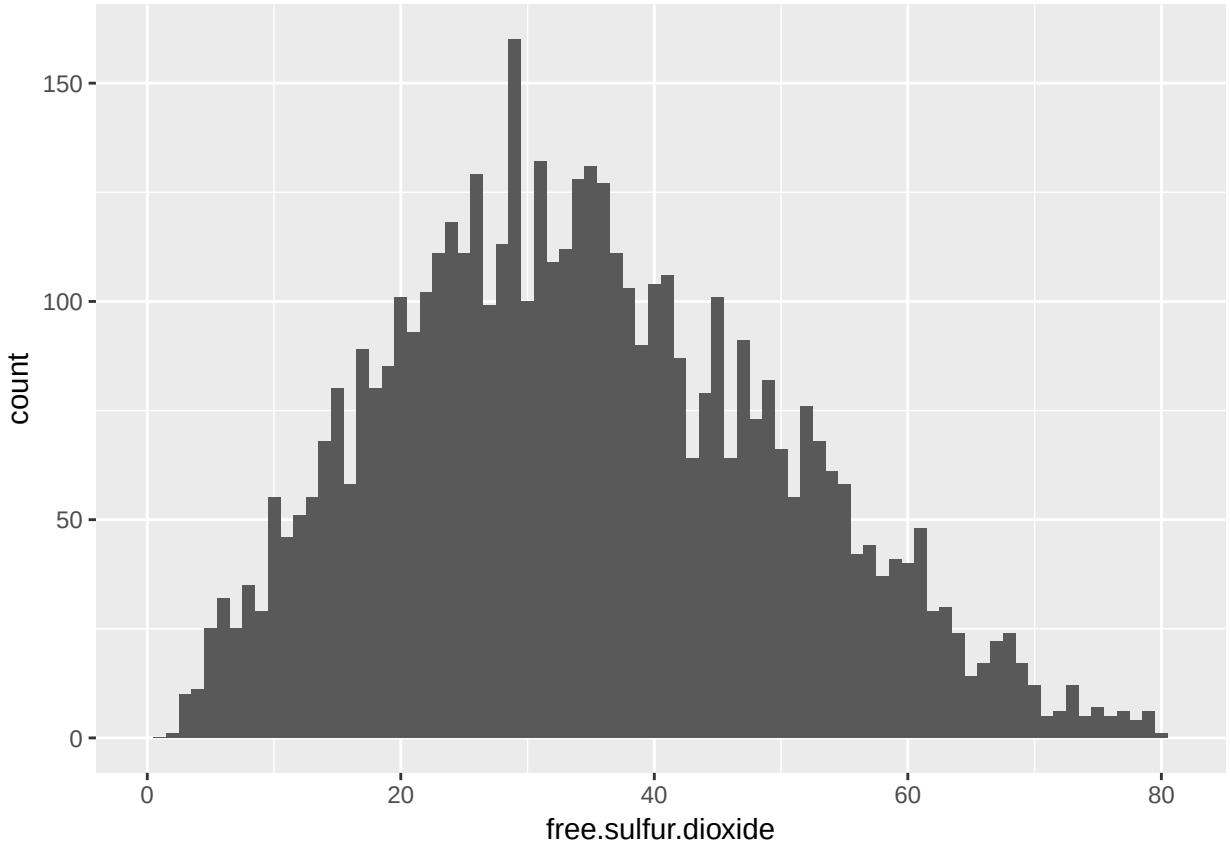


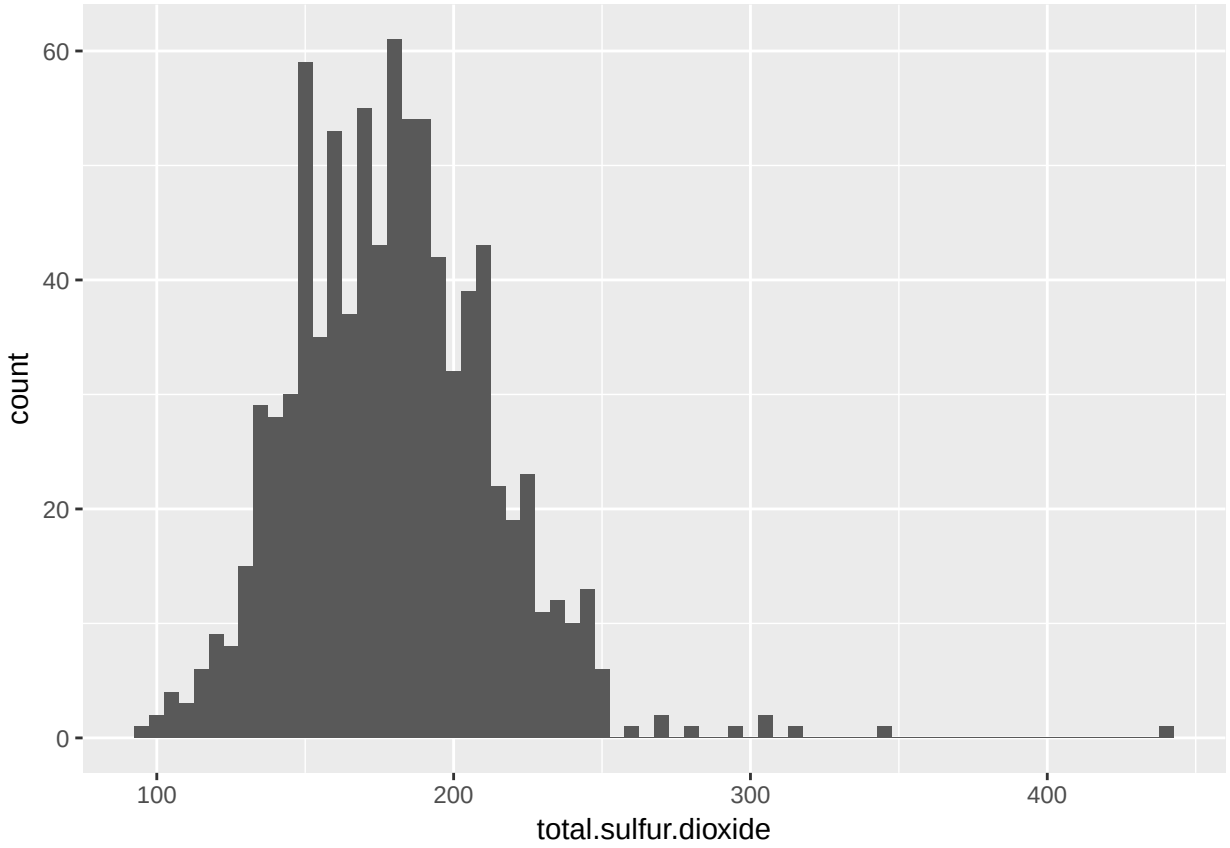

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

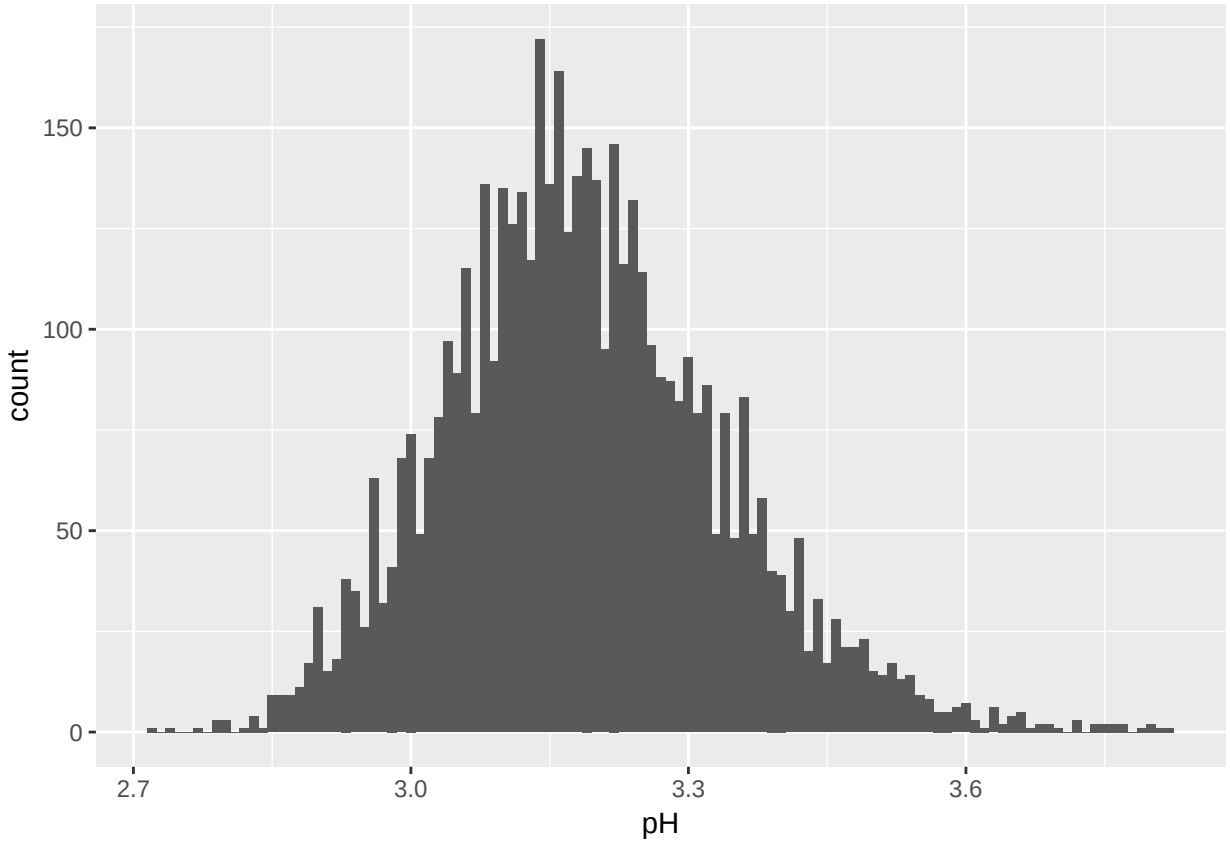


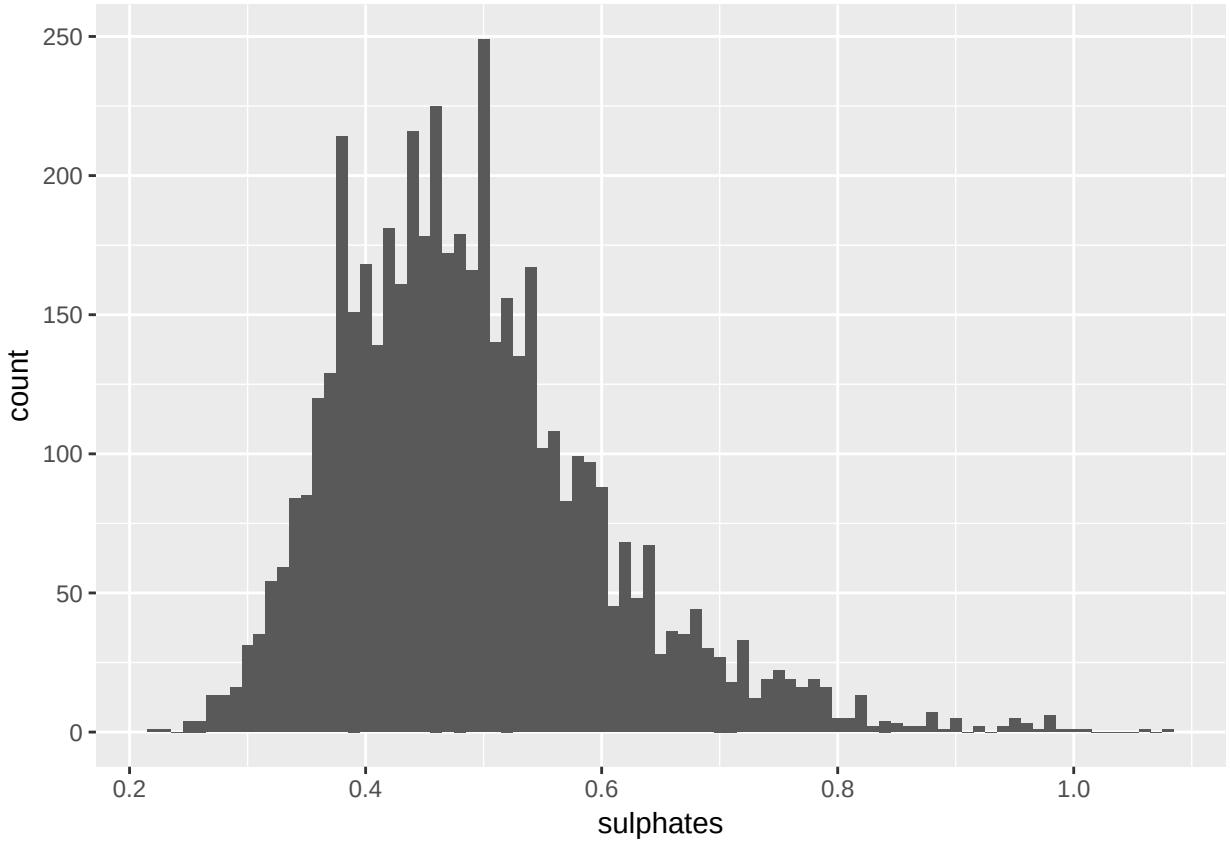

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2600  0.5150  0.5900  0.6124  0.6800  1.8600
```

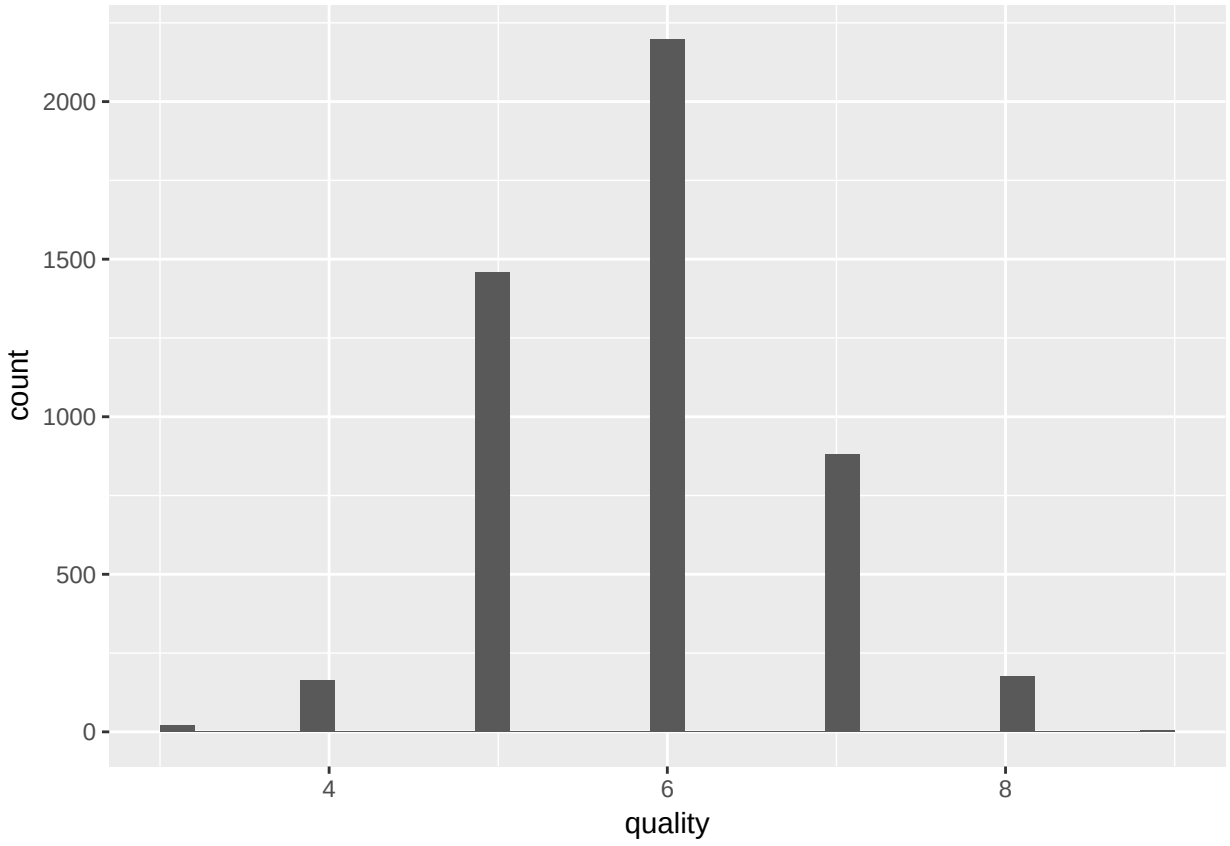

600
400
200
0
count
0
5
10
15
20
residual.sugar


150
100
50
0
count
0
20
40
60
80
free.sulfur.dioxide


60
40
20
0
count
100
200
300
400
total.sulfur.dioxide


150
100
50
0
count
2.7
3.0
3.3
3.6
pH


250
200
150
100
50
0
count
0.2
0.4
0.6
0.8
1.0
sulphates



Univariate Analysis

What is the structure of your dataset?

Looking at the variables for acidity it can be concluded that all three are spread normally with their specific means,however after faceting by quality it seems most of our data is for quality 5 to 7 which is confirmed by looking at the distribution of quality. Distribution for sulfur dioxide for both free and total forms,pH and sulphates is also normal.

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

The main features are acidity,sulfates & residual sugar and we will further investigate how they affect quality of white wine.

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

Other features like density and alcohol content which seem to be related to the residual sugar content may affect the quality. The pH levels may affect quality, however this seems very unlikely.

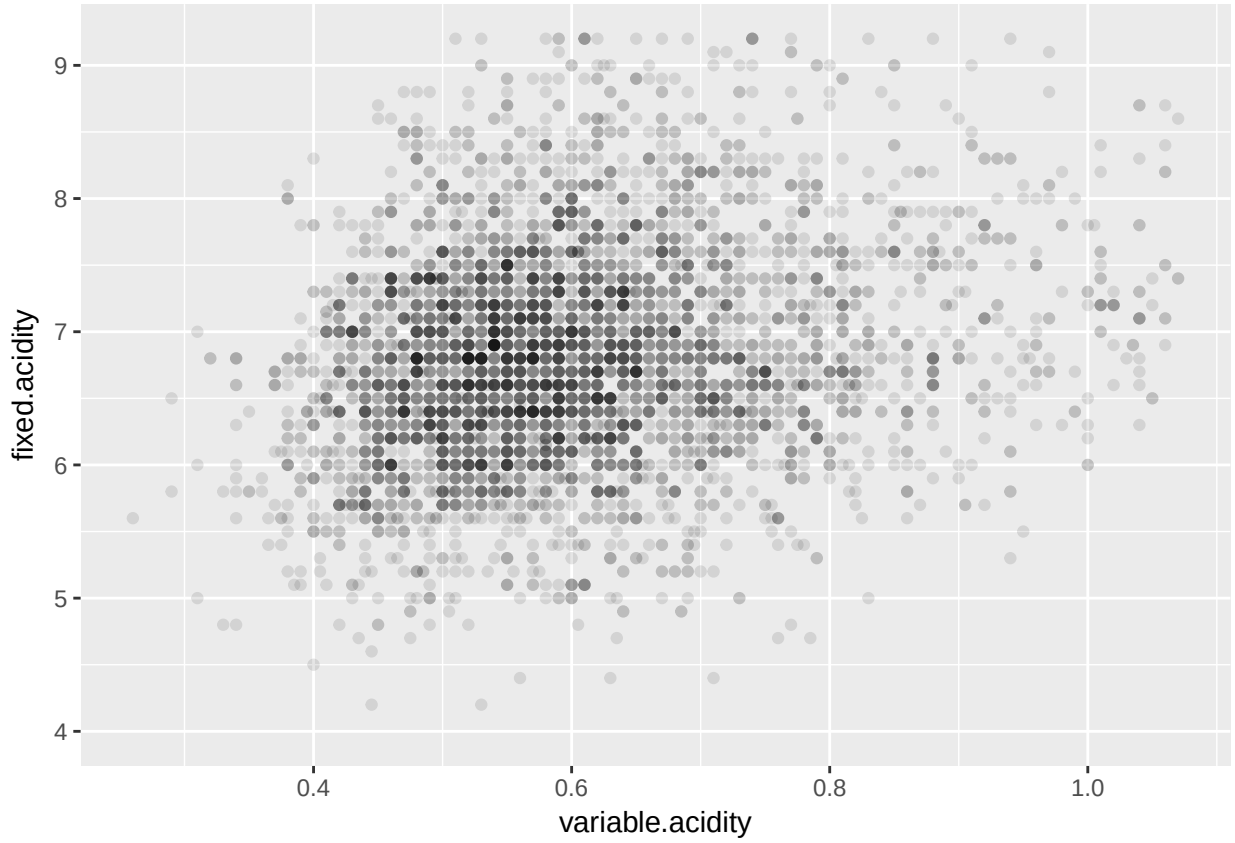
Did you create any new variables from existing variables in the dataset?

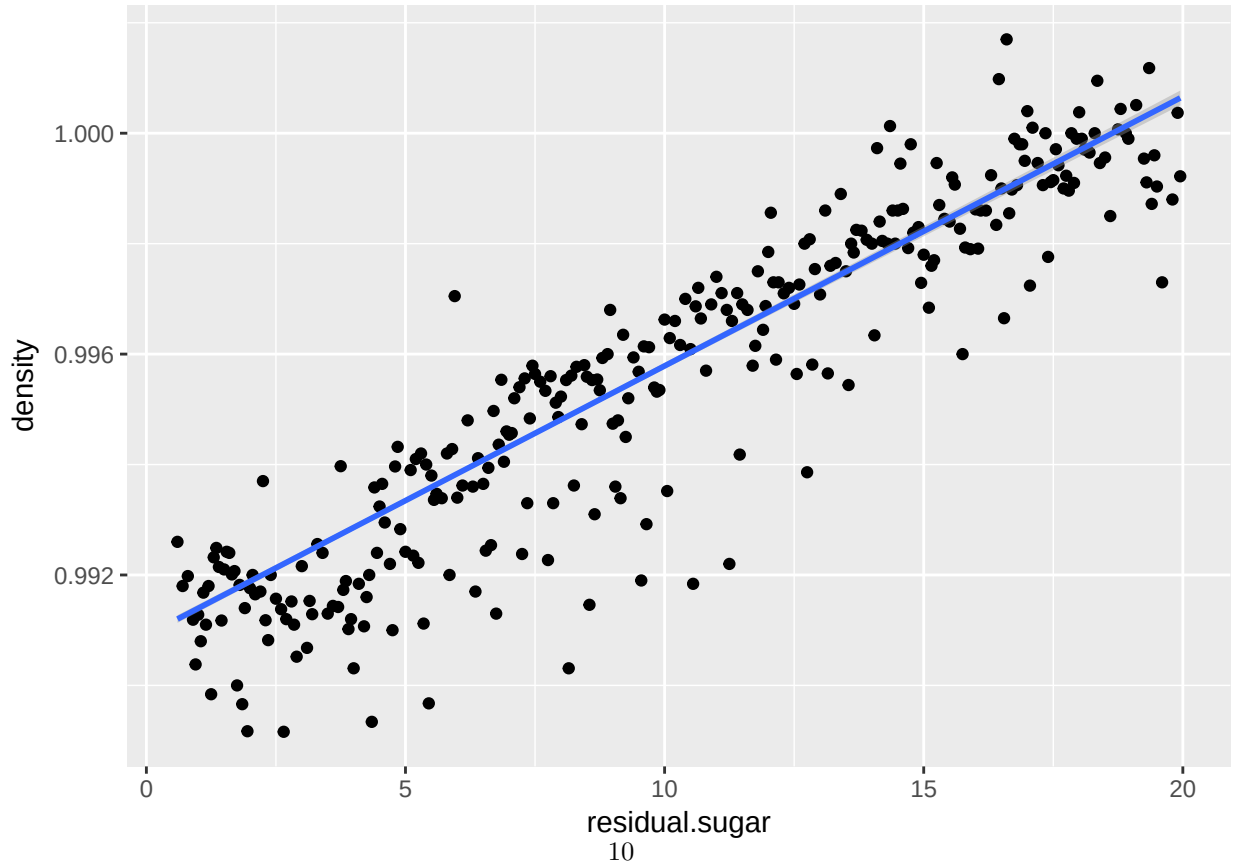
The two variables volatile acidity and citric acidity have similar trends and distributions and hence they can be combined to form a new variable 'variable.acidity'. Moreover these variables can also be combined because both account for a similar factor i.e. both should be present in small amounts only for a good wine.

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?

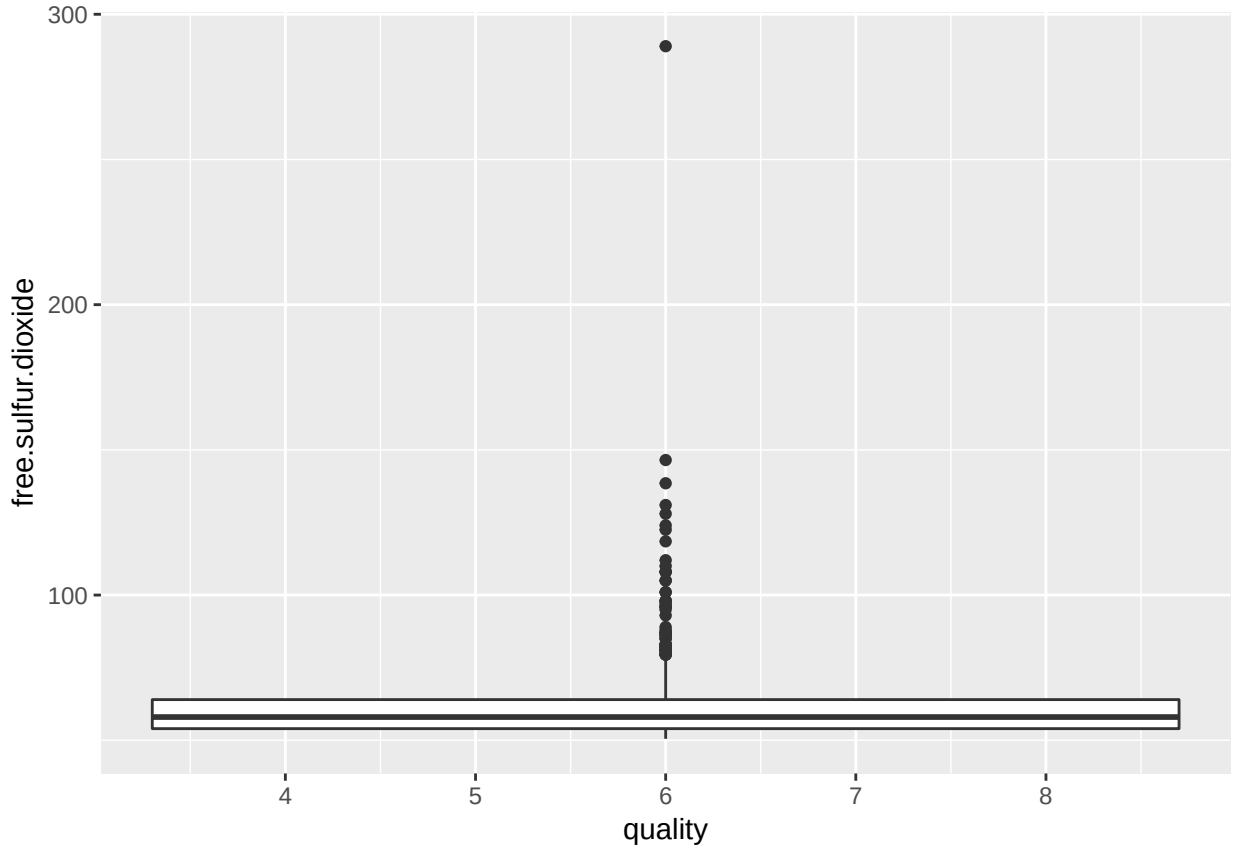
In the residual sugar variable, the bulk of the data seems to be in the region for values less than 20,hence this variable is limited to 20 to avoid outliers. Also a residual sugar value of 20+ makes the wine very sweet. For sulfur dioxide content we are only looking at those wines which have an prominent smell or taste of sulfur dioxide and this happens when the free sulphur dioxide exceeds 50ppm.This explains the subsetting done.

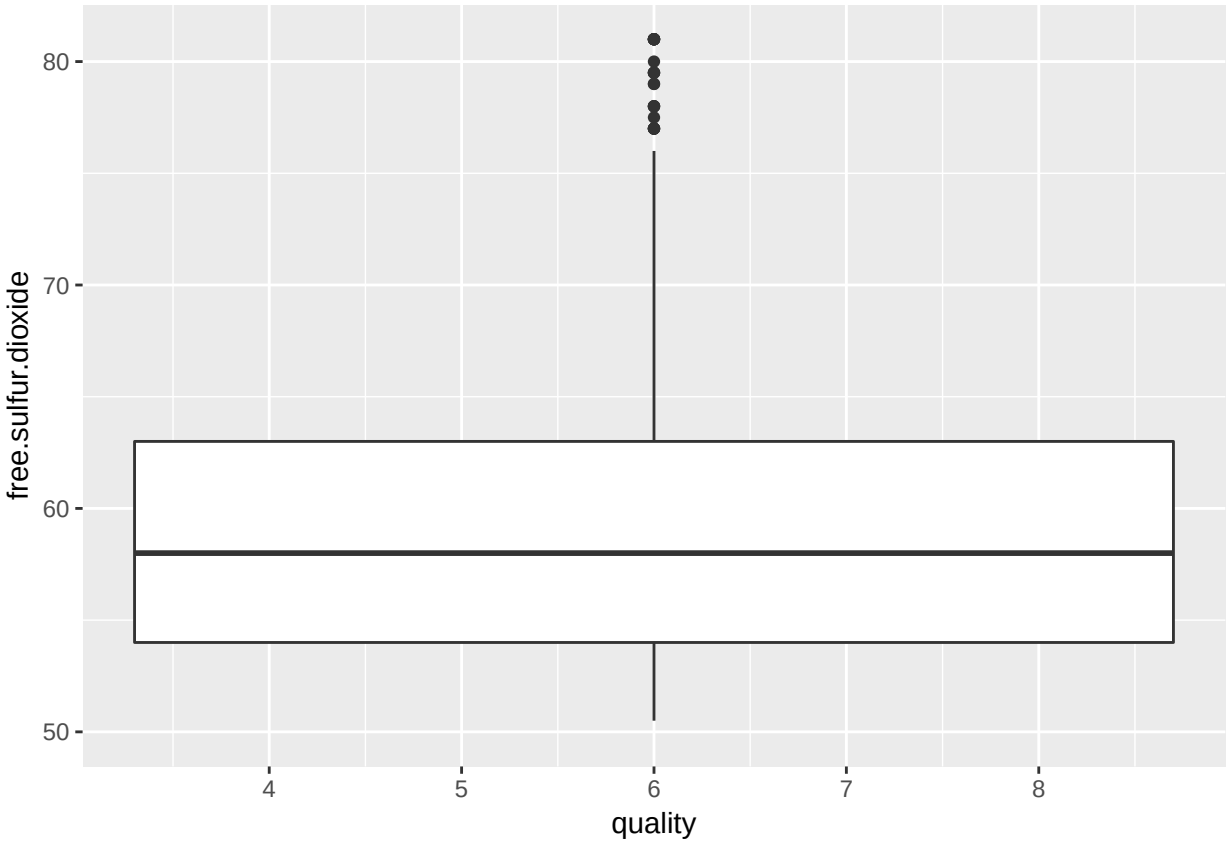
Bivariate Plots Section

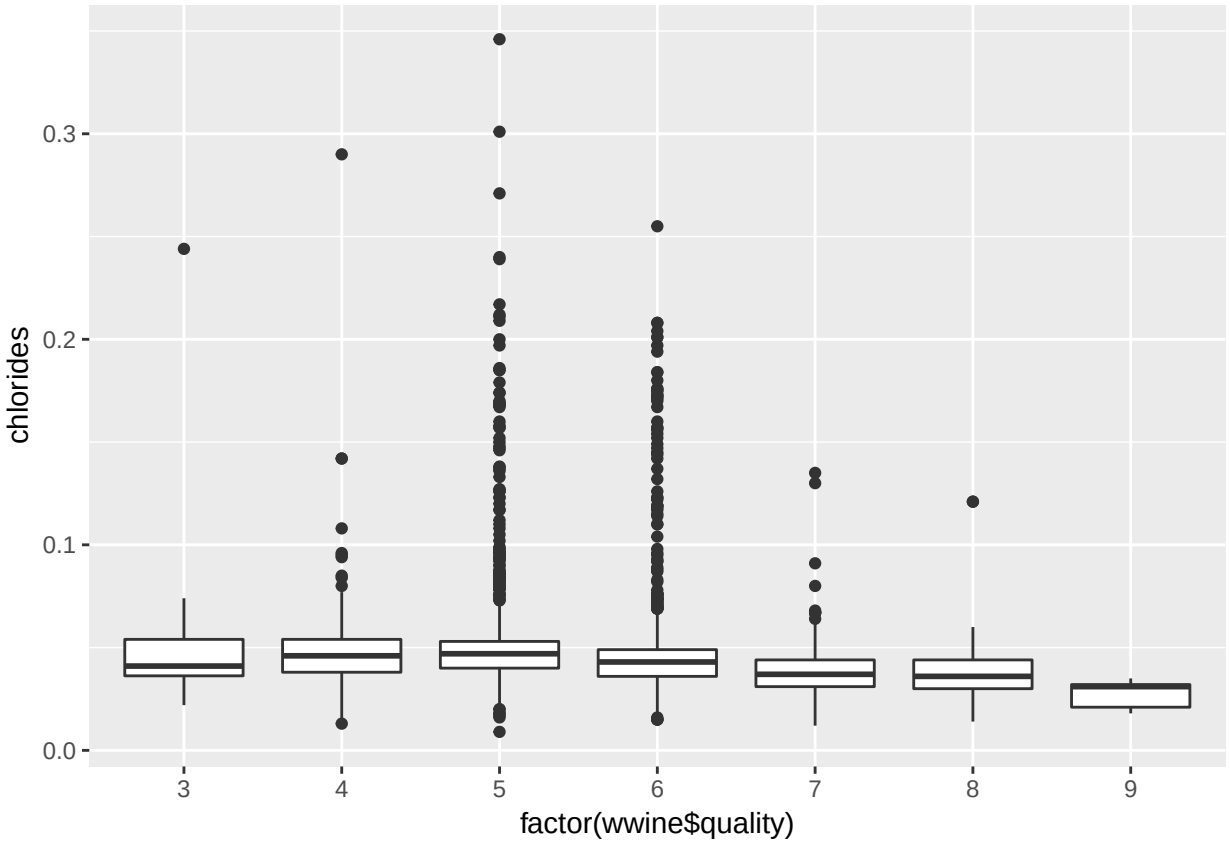


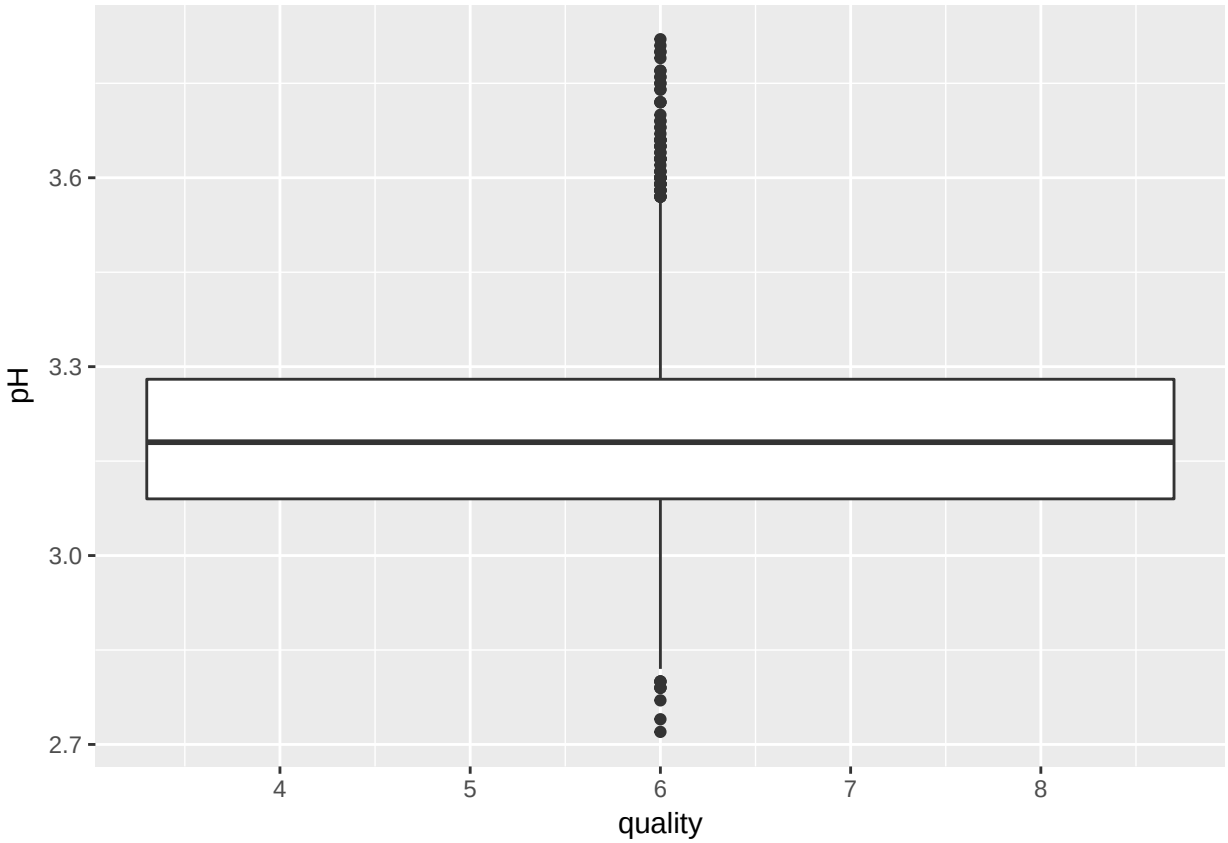


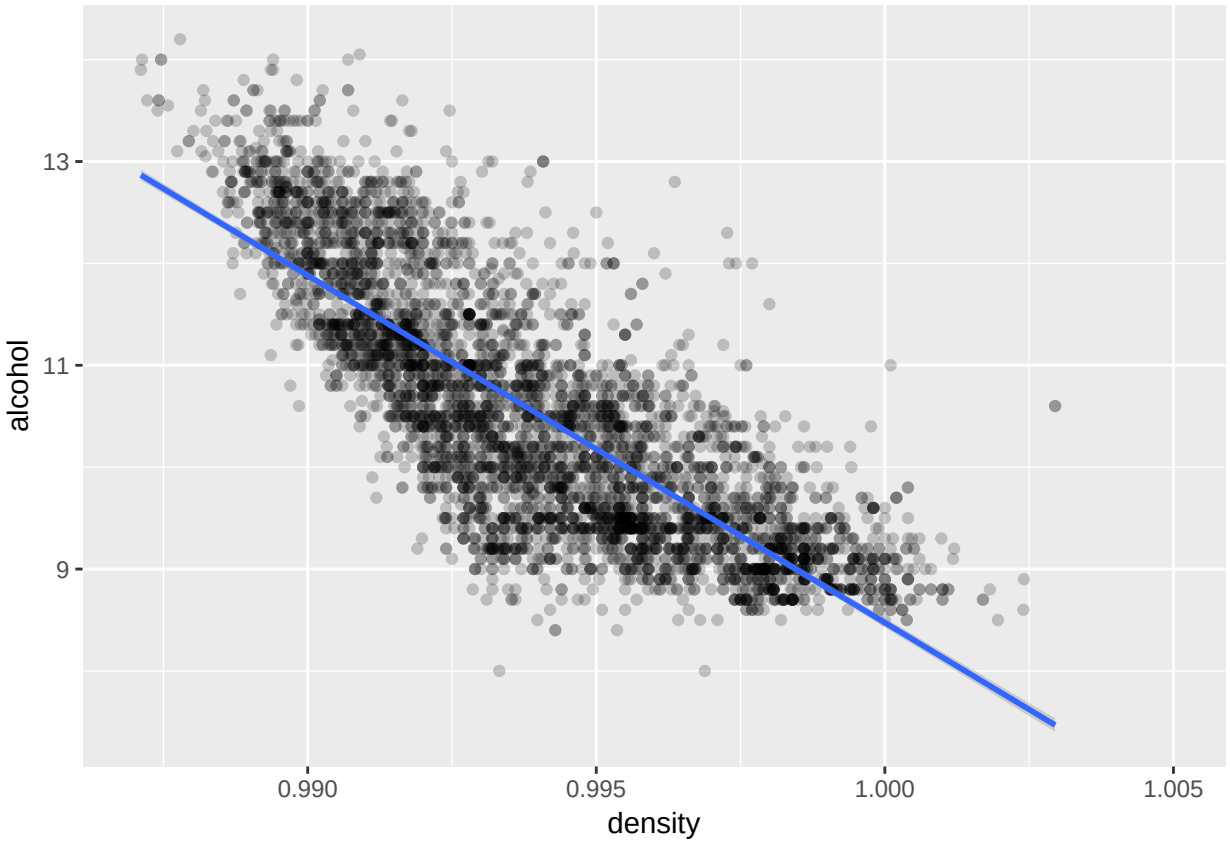
```
##
##  Pearson's product-moment correlation
##
## data:  wwine$residual.sugar and wwine$density
## t = 107.87, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8304732 0.8470698
## sample estimates:
##       cor
##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




80
70
60
50
free.sulfur.dioxide
4
5
6
7
8
quality


chlorides
0.3
0.2
0.1
0.0
3
4
5
6
7
8
9
factor(wwine$quality)


3.6
3.3
3.0
2.7
pH
4
5
6
7
8
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?

The boxplot between quality and sulfates(both free and total) show that there are many outliers present in the dataset and hence we consider only till the 99 percentile value. The vast difference can be seen from the two plots before and after transformation. The same can be said by looking at boxplot for chloride for each quality,there is a vast difference between the median and the max value. For the plot of acidities after limiting the axis it can be seen that there is a cluster of points for variable acidity between 0.4 to 0.8 and fixed acidity of 5.5 to 8,thus most of our data seems to lie in this range.Also the alpha parameter and jitter is used to avoid over-plotting and to get a better view.

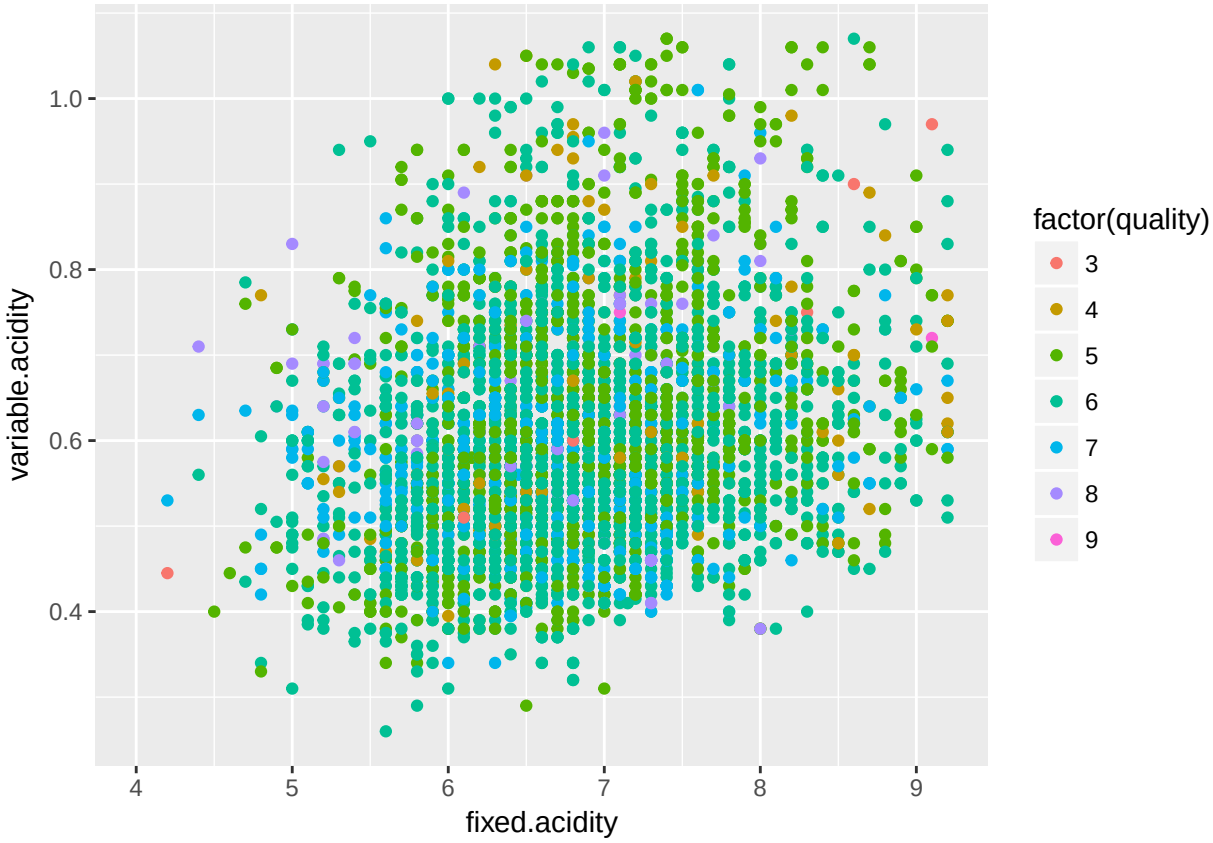
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

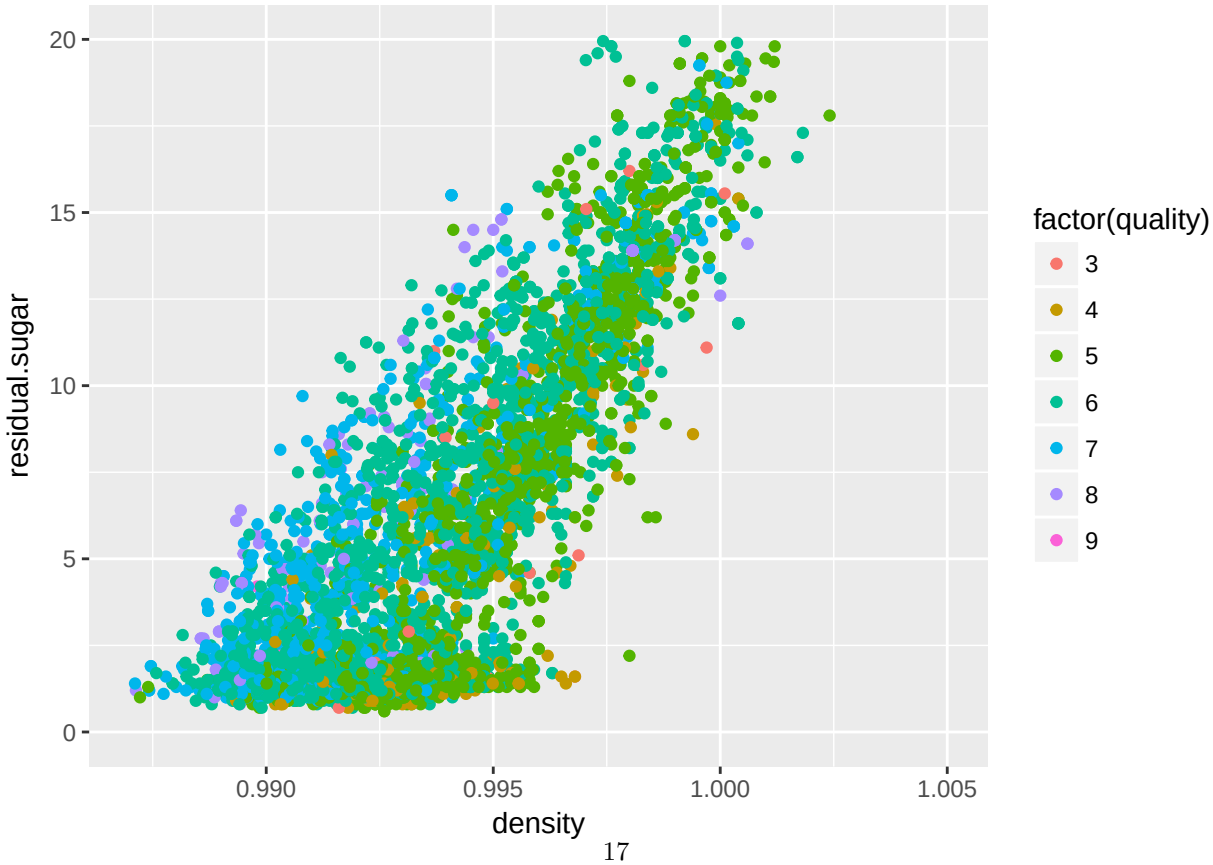
A peculiar linear relationship can be observed between the variables density and residual sugar.This is explained by the positive correlation between the two variables and evident by fitting a linear model. Also the plot between denisty and alcohol,shows that with increasing density the alcohol content decreases.

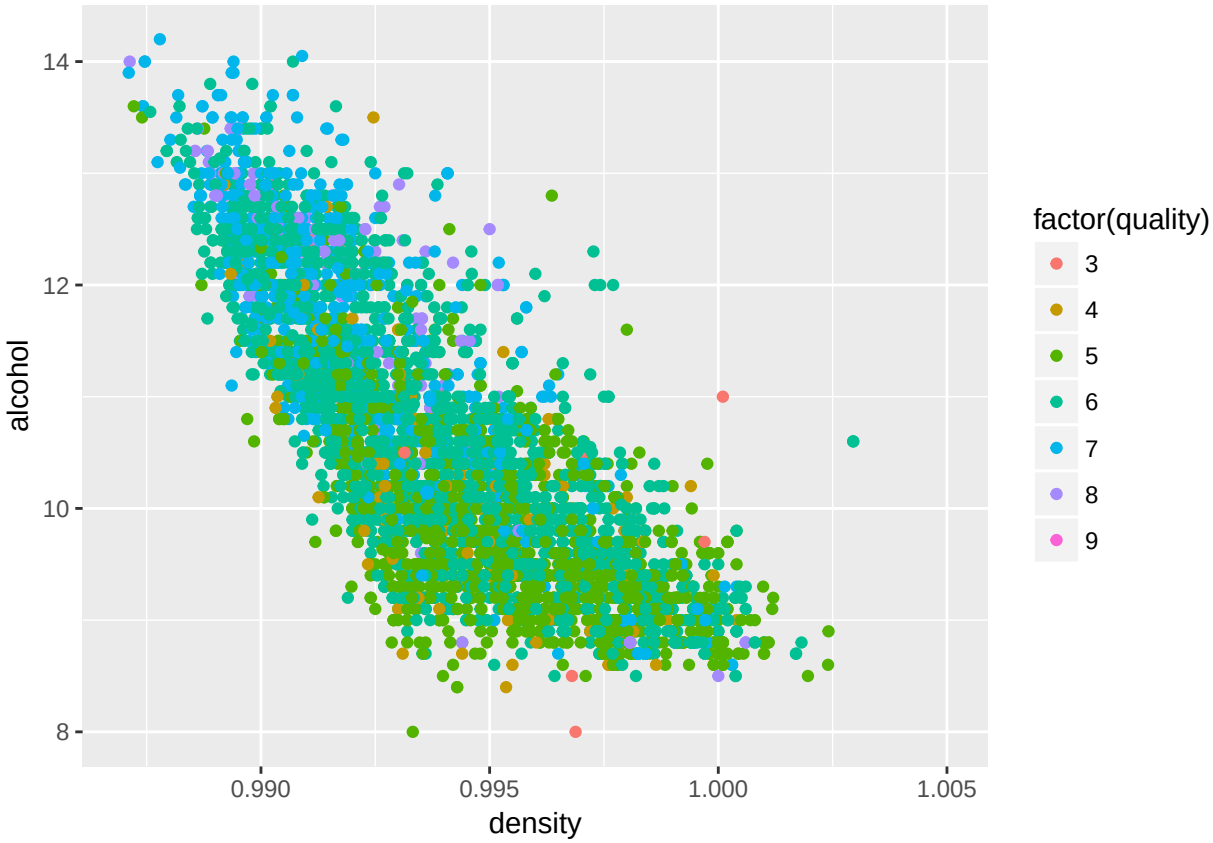
What was the strongest relationship you found?

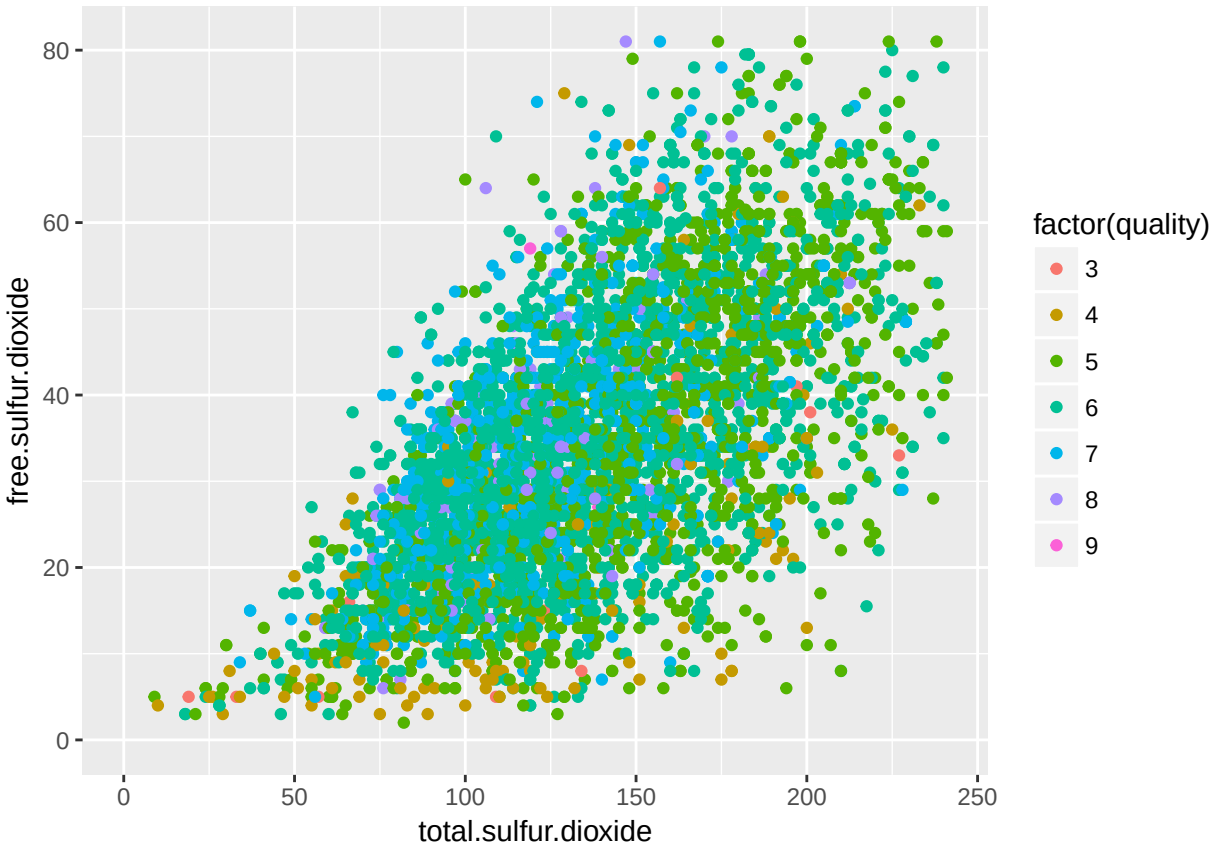
There seems a possibility of interconnectivity between the three variables density,alcohol and residual sugar which can be further explored in multivariate analysis.

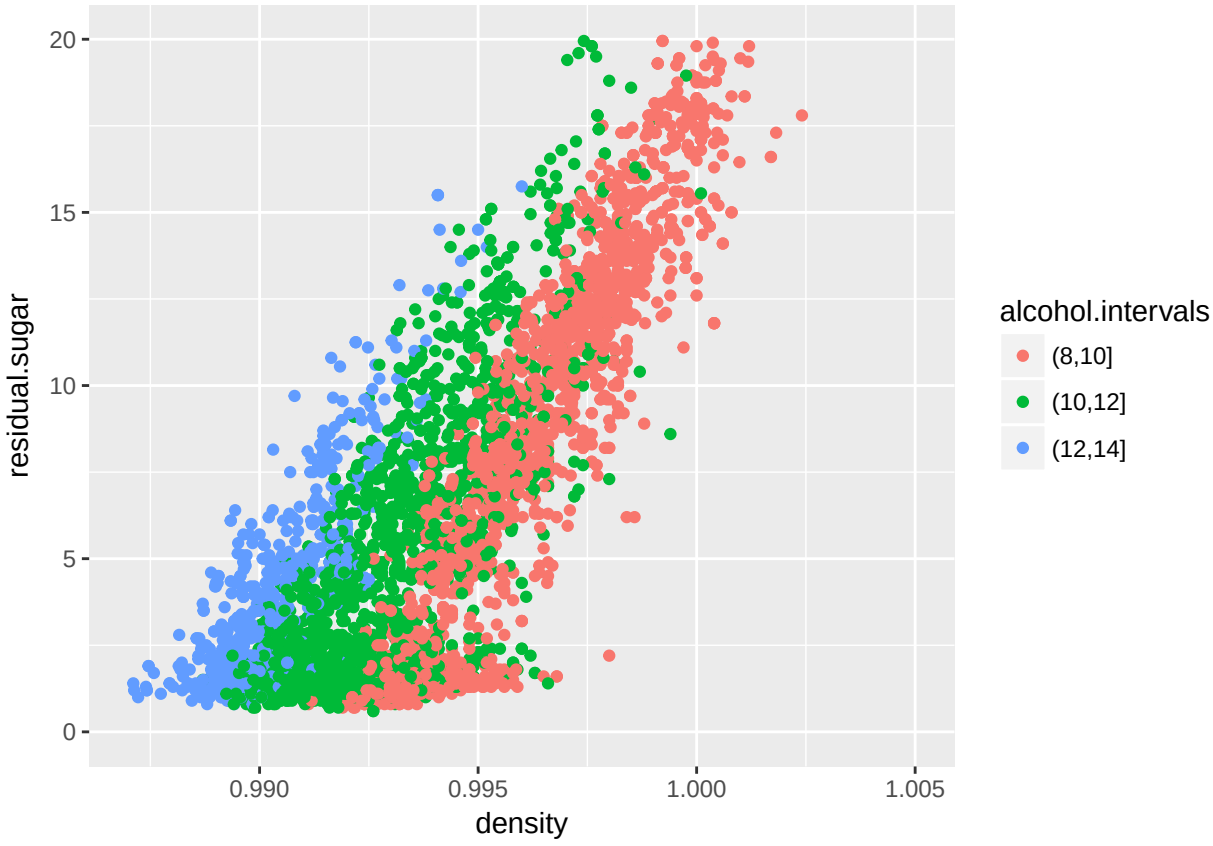
Multivariate Plots Section






14
12
10
8
alcohol
0.990
0.995
1.000
1.005
density
factor(quality)
3
4
5
6
7
8
9


free.sulfur.dioxide
total.sulfur.dioxide
0
20
40
60
80
0
50
100
150
200
250
factor(quality)
3
4
5
6
7
8
9



Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

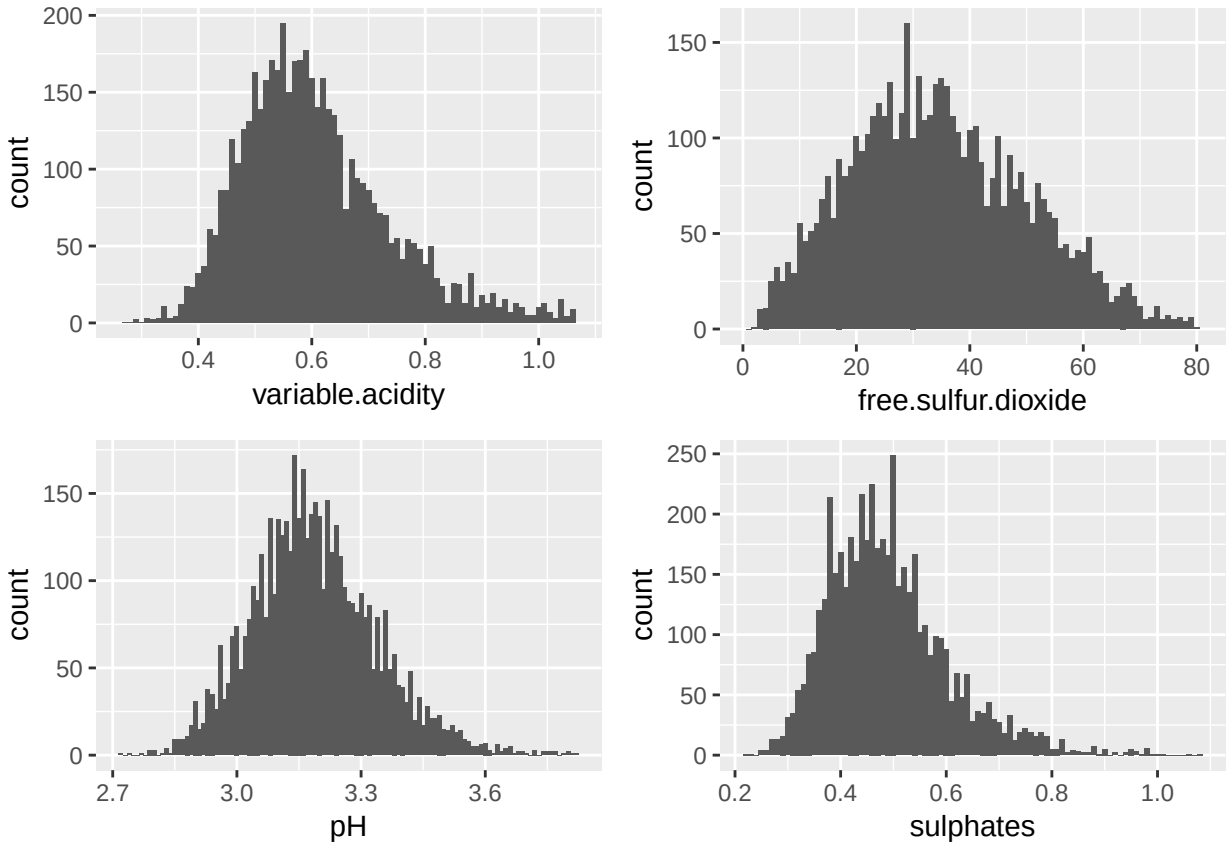
After colouring the points according to quality,it can be concluded that there is no visible way in which quality affects acidity. It can be seen that relation between density and residual sugar follows a different linear relation for a different quality. From the plot between alcohol and density we can see that wines of higher quality are less denser and have comparatively higher alcohol content than wines of lower quality. As seen previously there were many outliers in free and total sulphates and hence in an attempt to avoid these we first limit out scale to the 99th percentile.After this we look how quality affects this relationship.

Were there any interesting or surprising interactions between features?

On further investigating our initial surmise of the relationship between alcohol,density and residual sugar, we first divide the alcohol variable into three different ranges using the cut variable to get a better look.Seeing the effects of these alcohol ranges on the relationship between density and residual sugar we see that for each range there is a different slope for the curve.Thus there is a connection between the three variables.

Final Plots and Summary

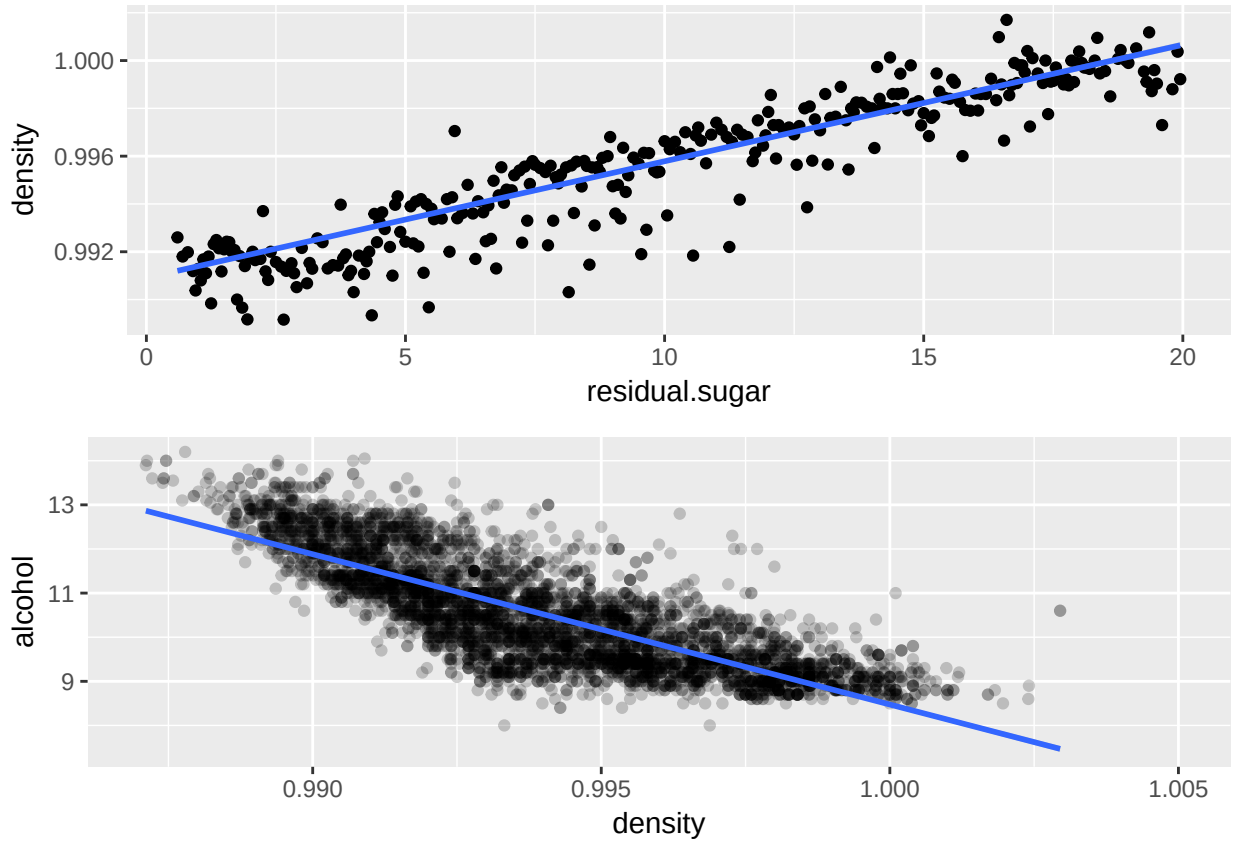
Plot One



Description One

From univariate analysis we've seen most of our variables have a normal distribution but with outliers also present,some have been plotted here together.

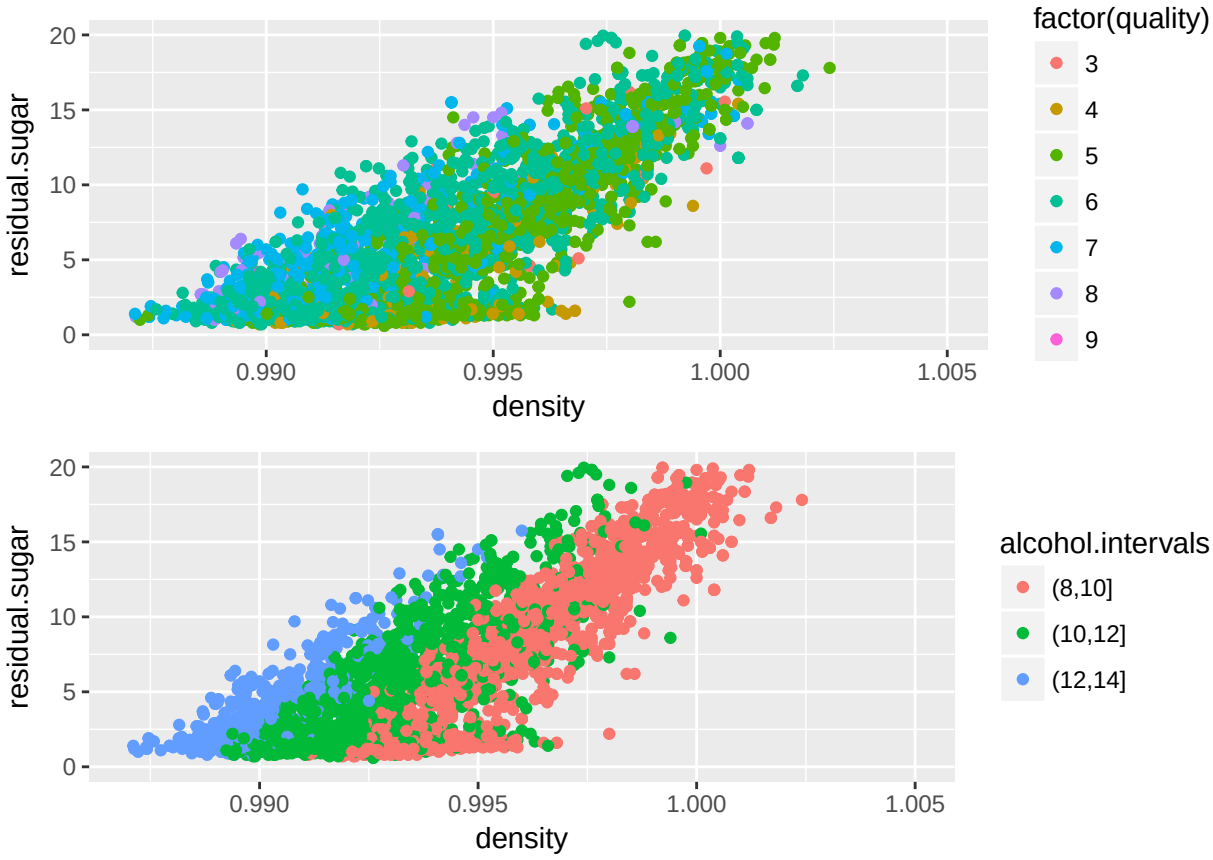
Plot Two



Description Two

The two strong relationships observed during bivariate analysis for the variables density,residual sugar and alcohol,hinting an interconnectivity between the three.

Plot Three



Description Three

The results of further exploration showing how quality affects the relationships and finally confirming the interconnectivity present.

Reflection

Thus we have explored the different variables of the white wine dataset and seen the relationships between variables and how these changes determine the quality of wine.However speaking about quality of wine can be a very subjective topic as different people may prefer different prevalent tastes(eg.acidic,sweet,etc) which make the wine perfect according for them.Though an attempt made above may help in generalizing the preferences.