Exploring Red Wine Quality

Shane Kao

Wednesday, February 18, 2015

Goal

Which chemical properties influence the quality of red wines?

Data Overview

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).

```
setwd("C:/Users/asus/Downloads")
data=read.csv("wineQualityReds.csv",stringsAsFactors=FALSE)
str(data)
```

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

```
summary(data)
```

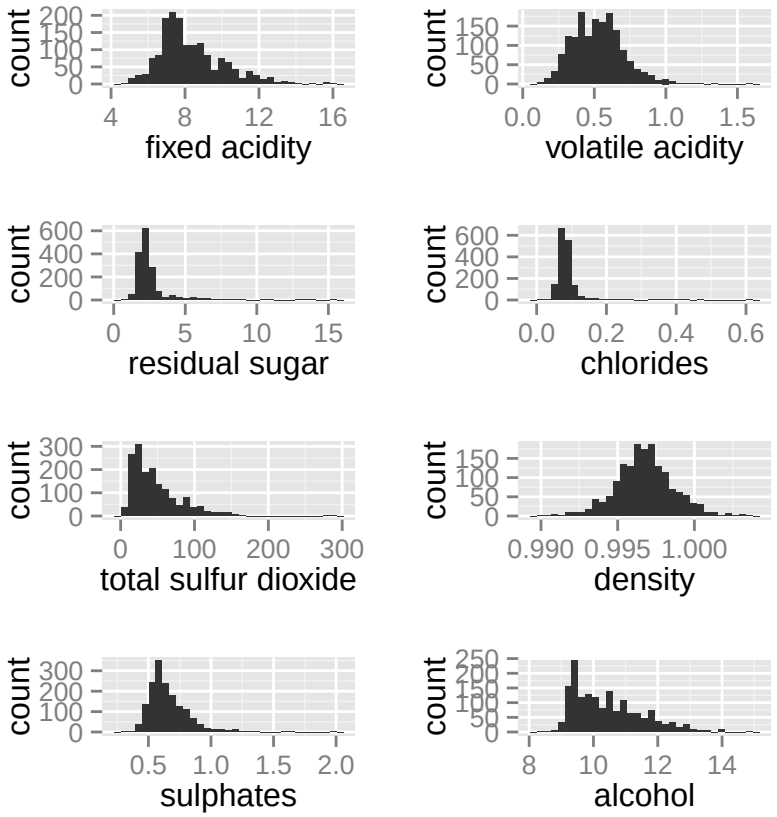
```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 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
```

- The feature `X` is row index of data, it may provide no further information.
- The feature `quality` is an ordered, categorical, discrete variable.
- From the variable descriptions, {`fixed.acidity`,`volatile.acidity`,`citric.acid`} and {`free.sulfur.dioxide`,`total.` may strongly correlated.

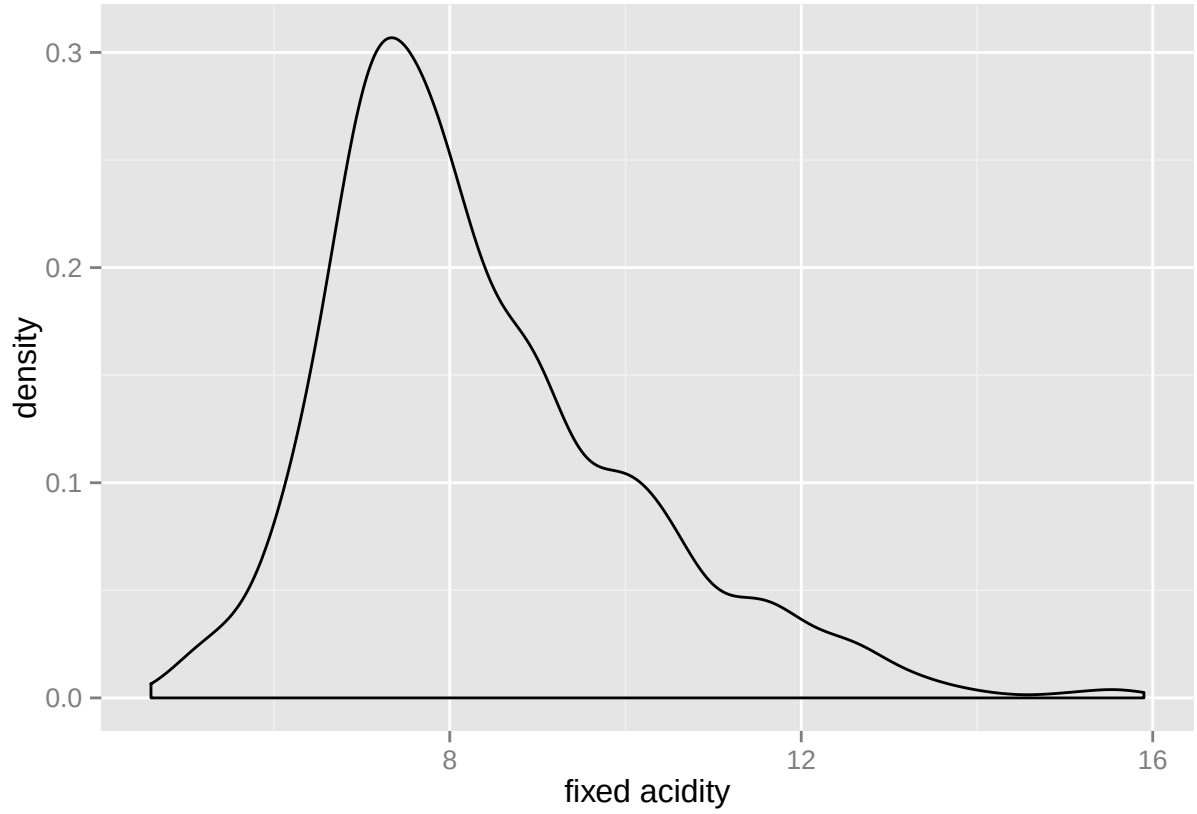
Univariate Plots Section

Univariate Analysis

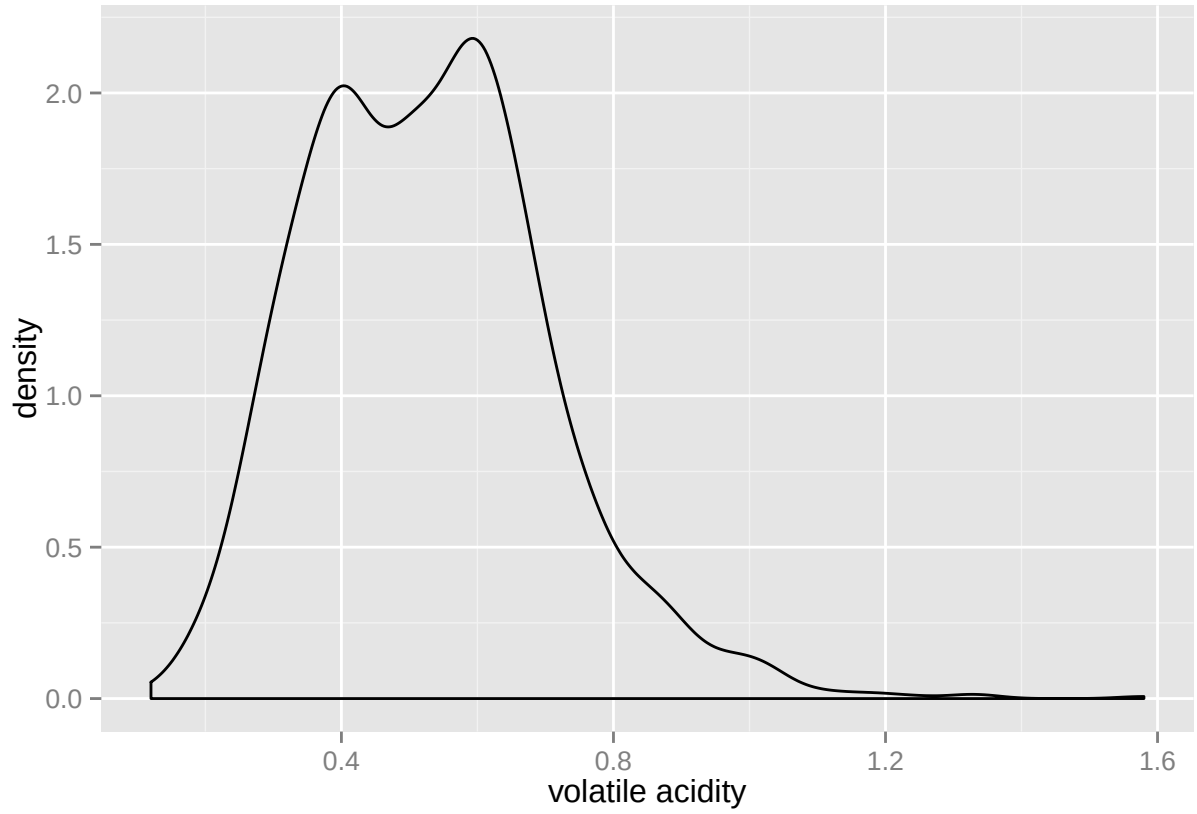


The following figures shows distributions for all features:

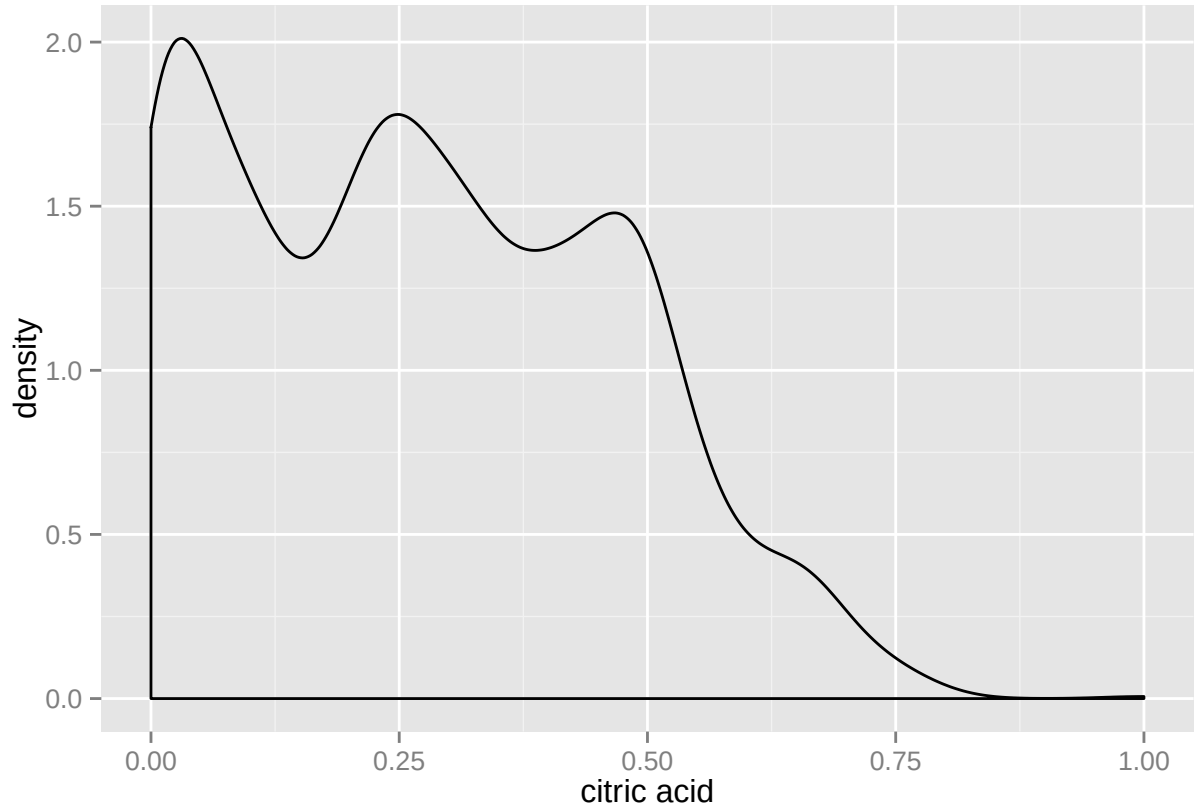
Now, we display density estimation for each features



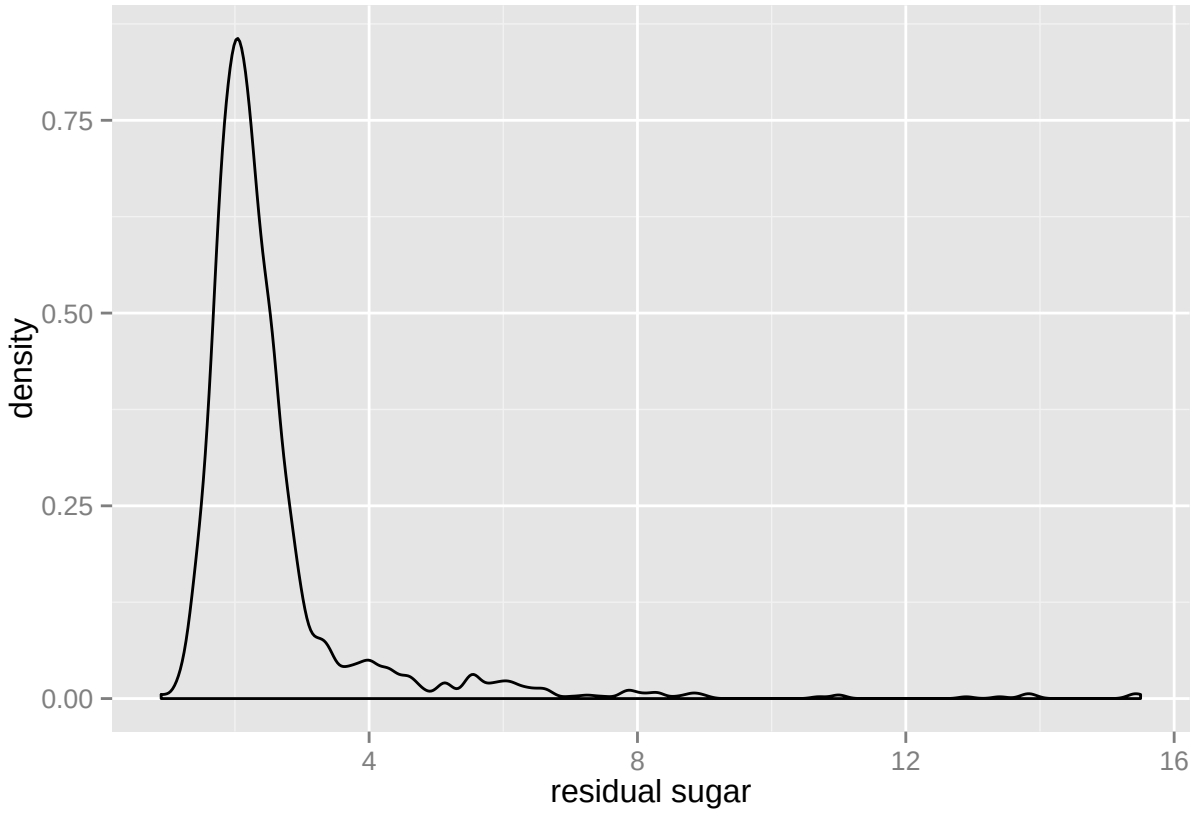
The sample average of `fixed.acidity` is about 8



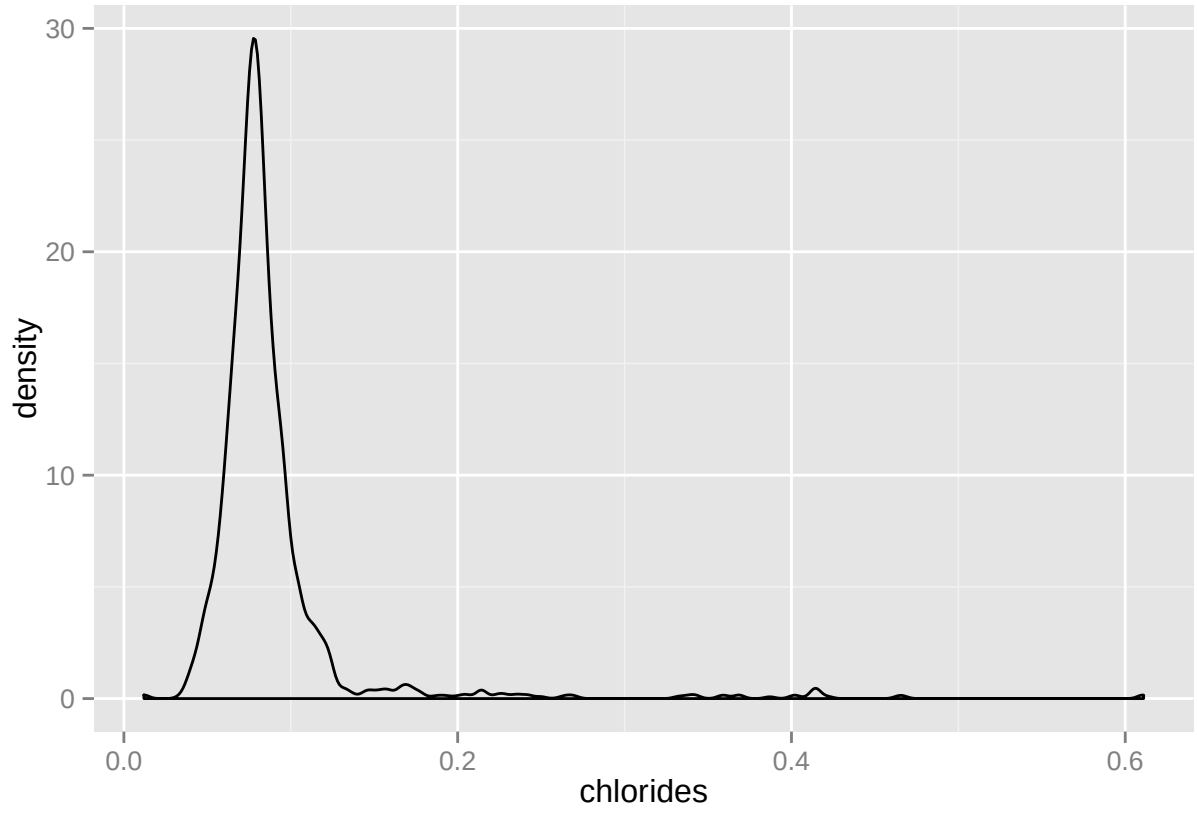
The distribution of `volatile.acidity` reveal bimodal



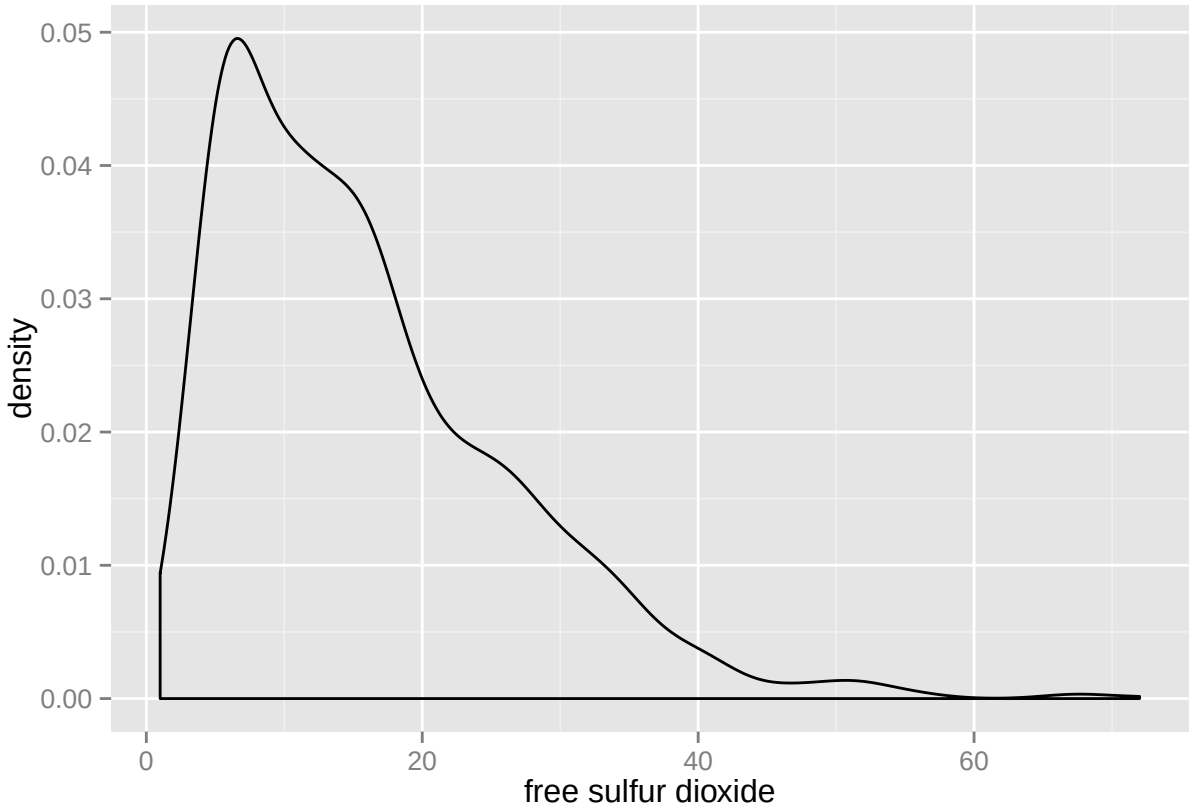
The distribution of `citric.acid` reveal trimodal



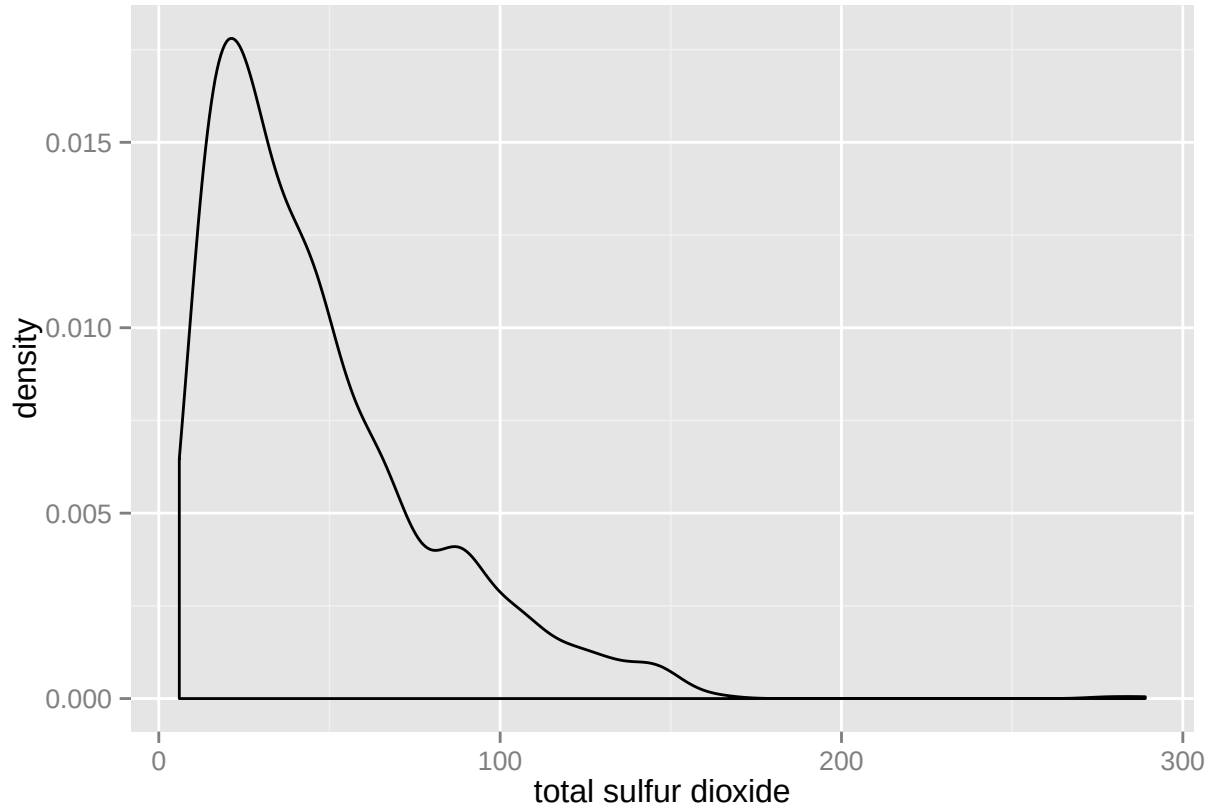
Most of the `residual.sugar` is smaller than 4



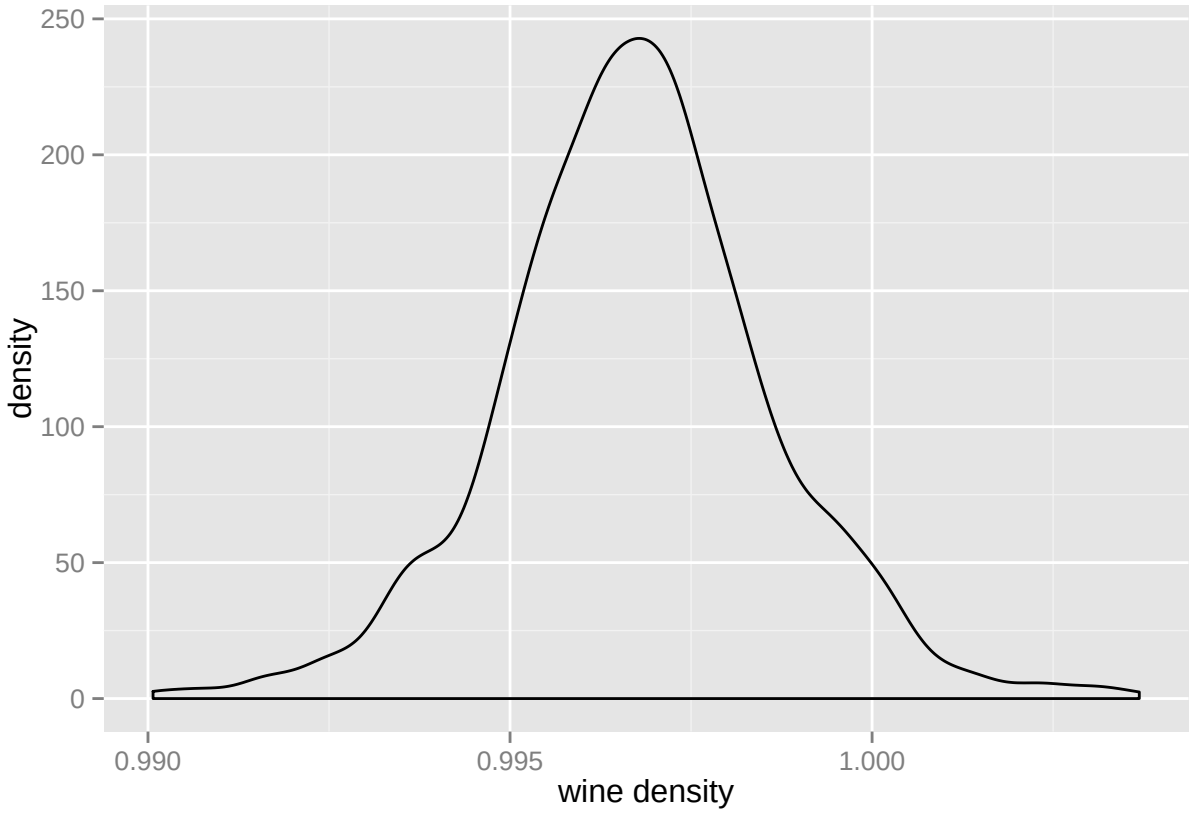
Most of the `chlorides` is smaller than 0.1



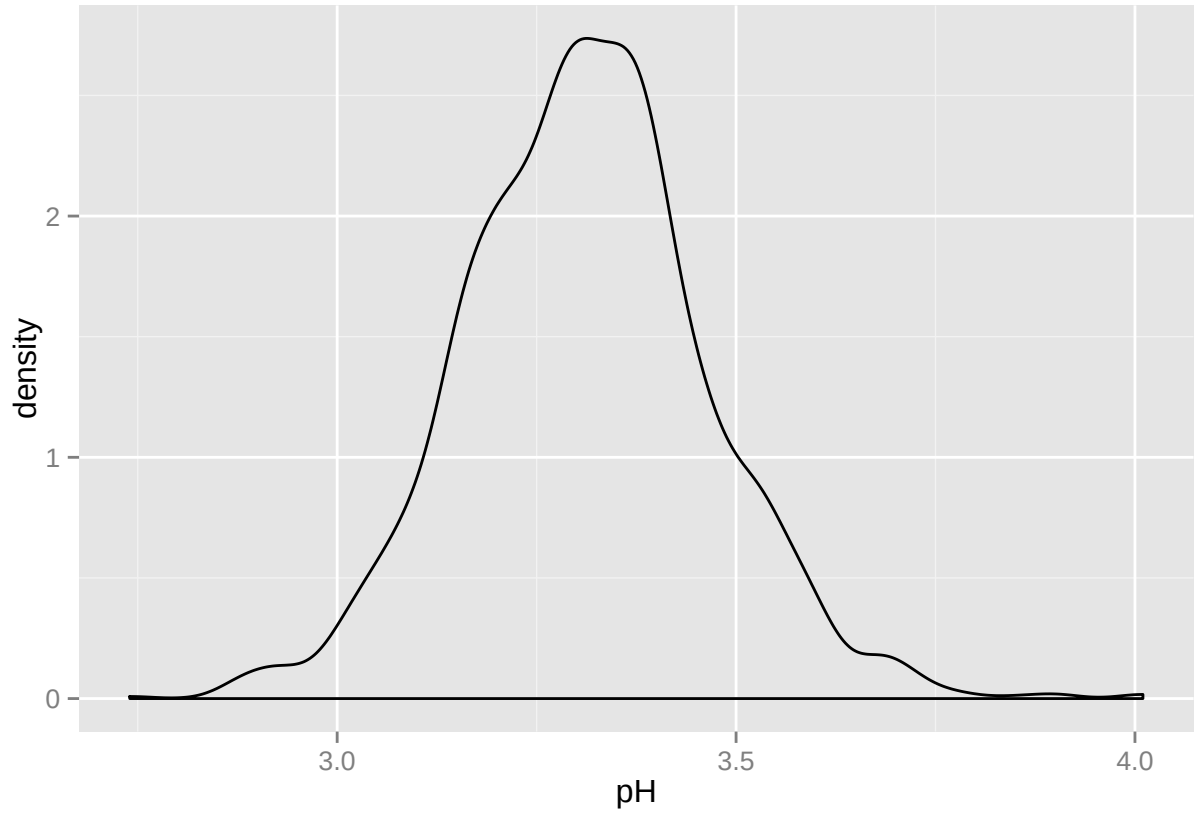
The mode of `free.sulfur.dioxide` is about 5



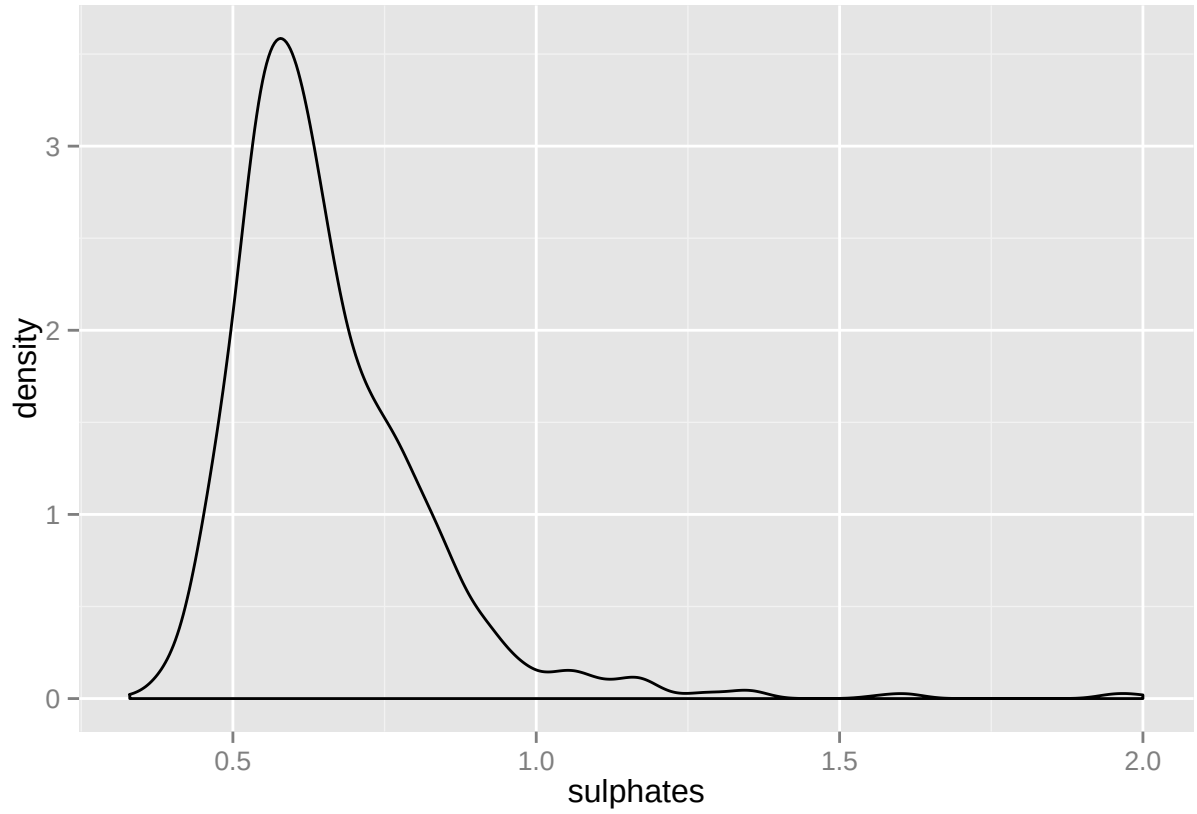
The shape of distribution for `total.sulfur.dioxide` and `free.sulfur.dioxide` are similar



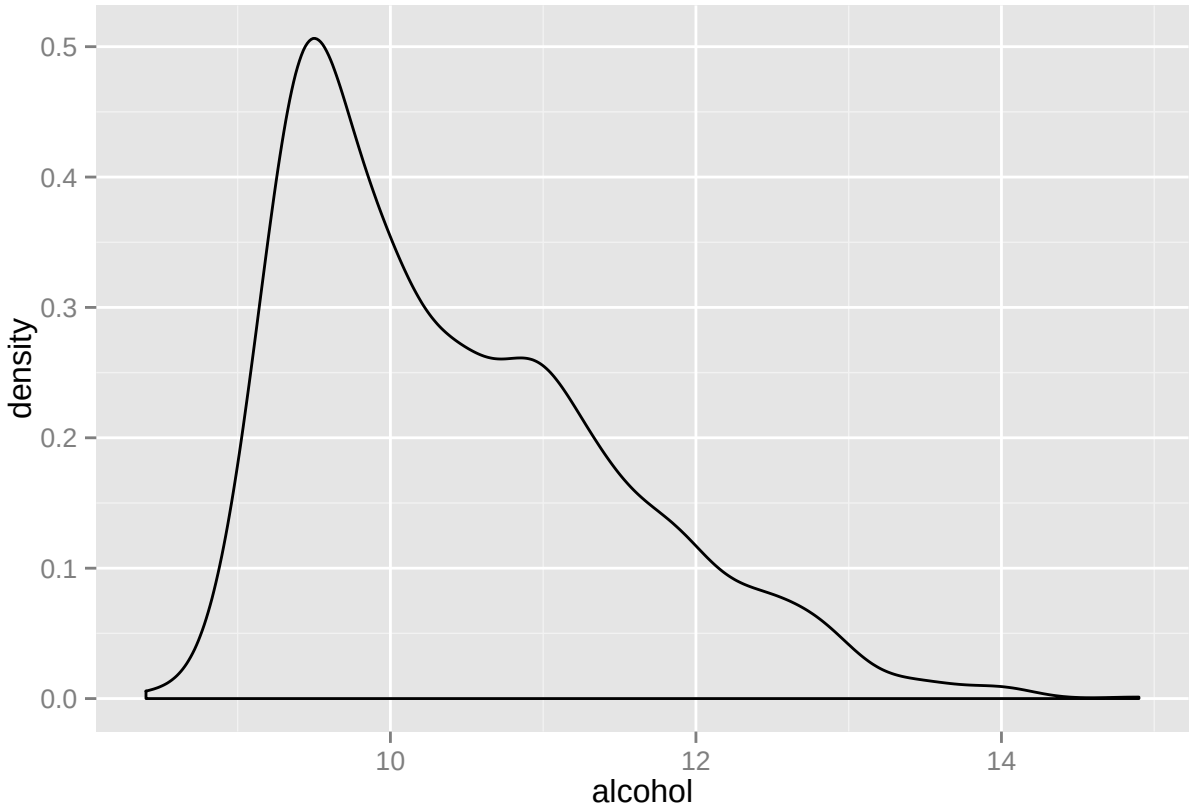
`density` appear to be normally-distributed.



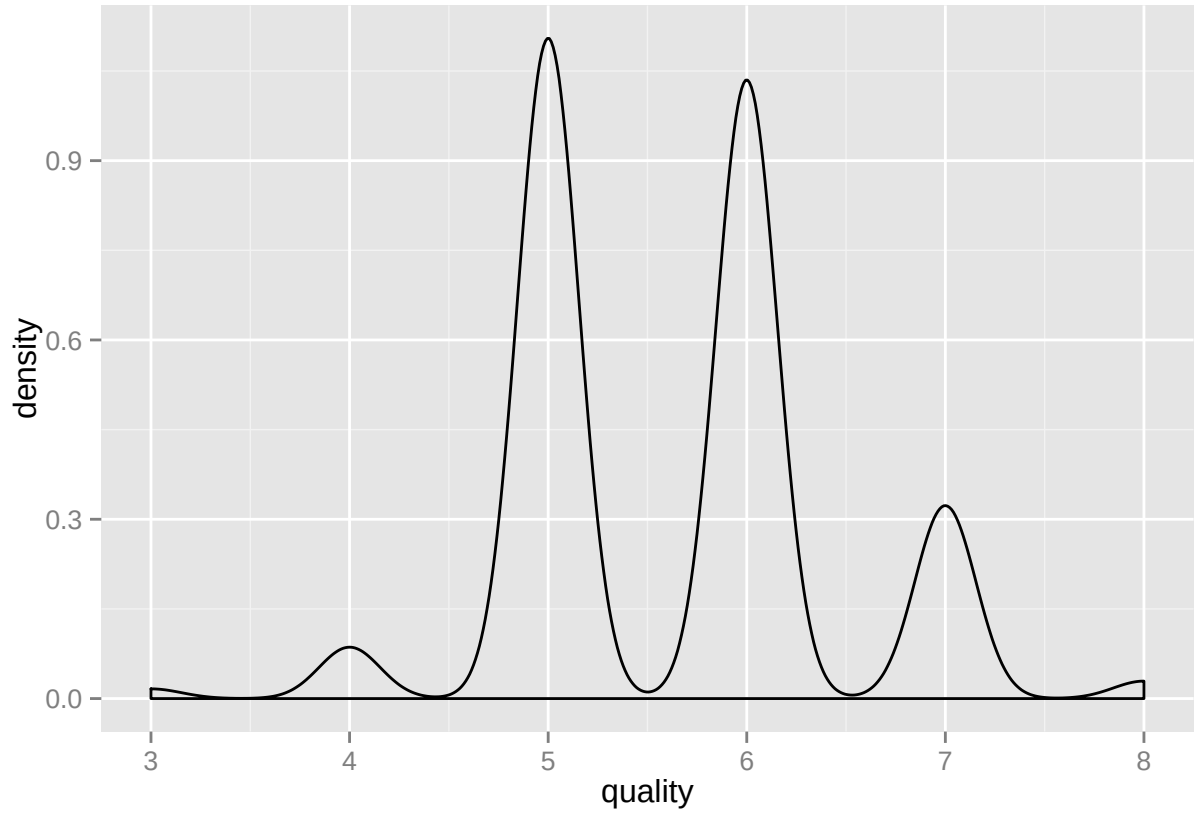
- `pH` appear to be normally-distributed.



The mode of `sulphates` is about 0.6



The mode of `alcohol` is about 9



rating 5 and rating 6 has more points than other rating.

Short questions

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

- It is convenient to interpret the result by creating variable `rating` , classifying each wine as low, medium and high, assign `quality` 3 and 4 to low level, 5 and 6 to medium and 7 and 8 to high.

```
data$rating<-rep("",dim(data)[1])
data$rating[which(data$quality%in%3:4)]<-"low"
data$rating[which(data$quality%in%5:6)]<-"medium"
data$rating[which(data$quality%in%7:8)]<-"high"
table(data$rating)
```

```
##
##   high    low medium
##    217     63   1319
```

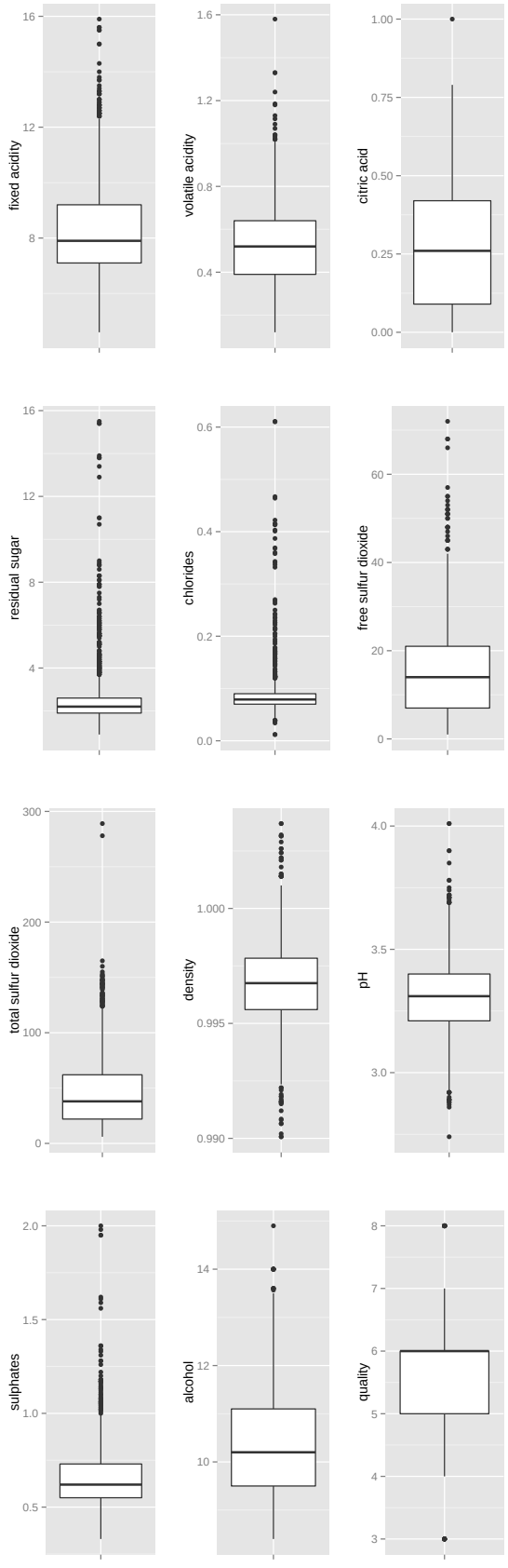
- I create a combined variable, `acid`, taking average of `fixed.acidity`, `volatile.acidity` and `citric.acid` after standardization.

```
data$acid<-((data$fixed.acidity-mean(data$fixed.acidity))/
                    sd(data$fixed.acidity)+
            (data$volatile.acidity-mean(data$volatile.acidity))/
                    sd(data$volatile.acidity)+
            (data$citric.acid-mean(data$citric.acid))/
                    sd(data$citric.acid))/3
```

- I create a combined variable, `dioxide`, taking average of `free.sulfur.dioxide` and `total.sulfur.dioxide` after standardization.

```
data$dioxide<-((data$free.sulfur.dioxide-mean(data$free.sulfur.dioxide))/
                       sd(data$free.sulfur.dioxide)+
                (data$total.sulfur.dioxide-mean(data$total.sulfur.dioxide))/
                       sd(data$total.sulfur.dioxide))/2
```

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?

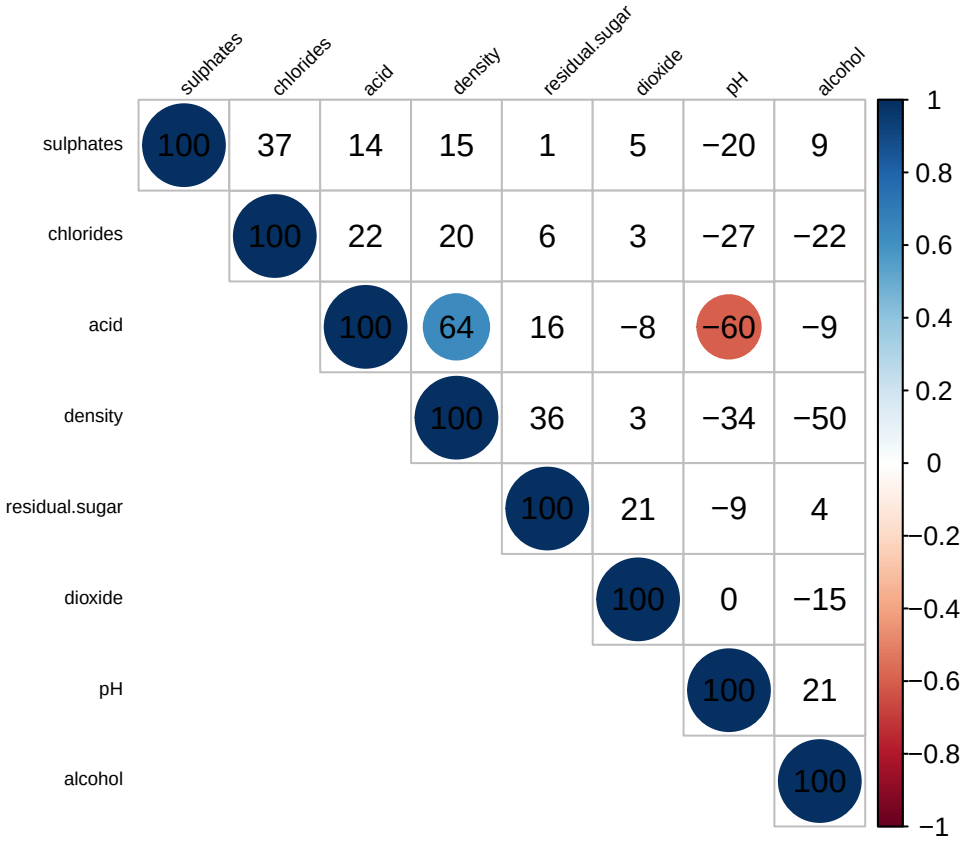


- `residual.sugar` and `chlorides` have extreme outliers.
- I don't tidy or adjust any data.

Bivariate Plots Section

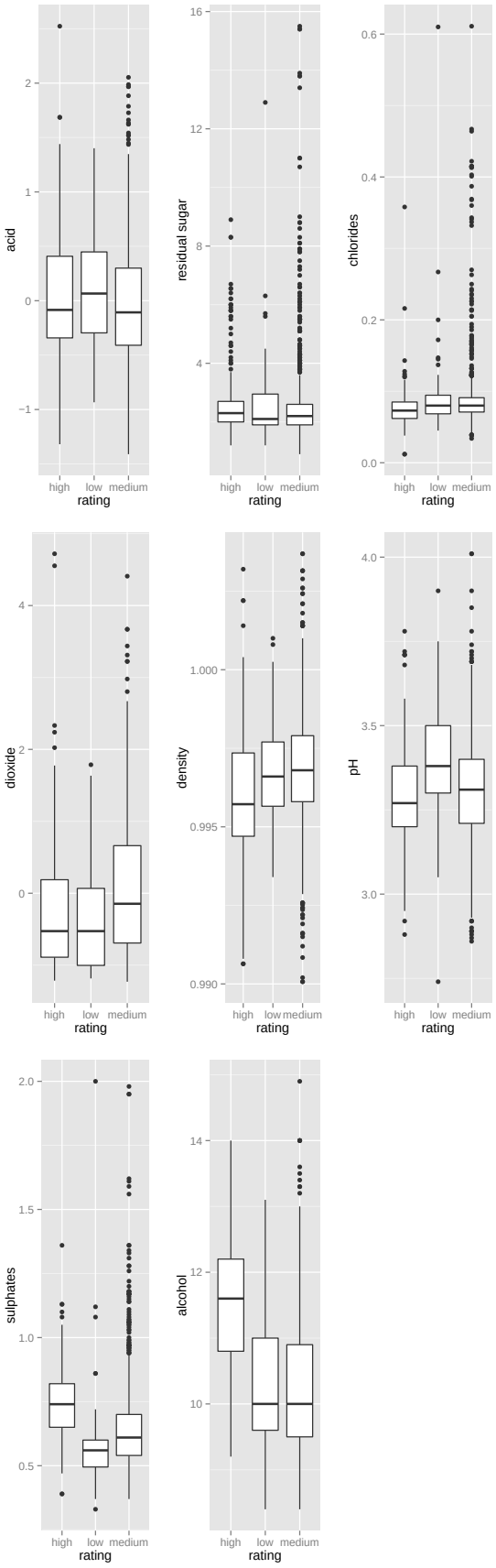
Bivariate Analysis

We use the following figure to check the correlation between different features



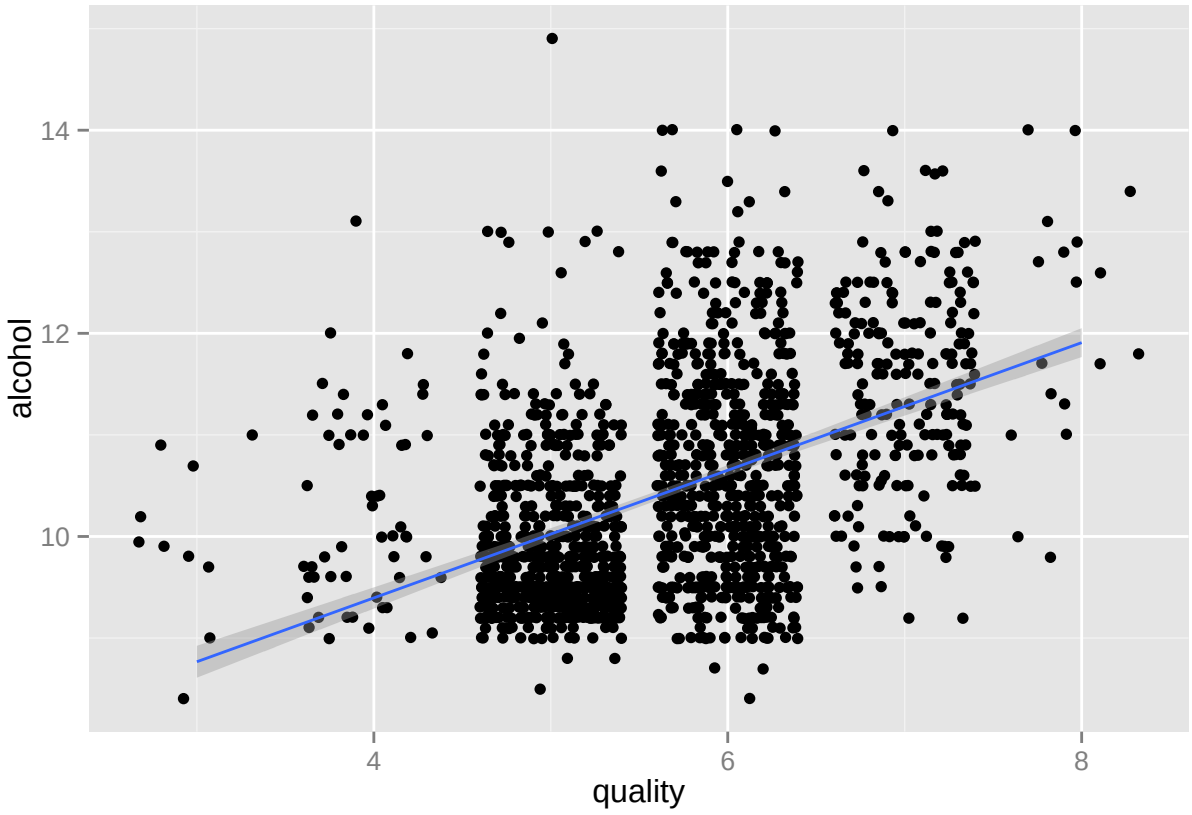
Most of features are un-correlated, but `acid` is correlated with `pH` and `density`. Note that there are only two points colored off the main diagonal (acid vs. density and acid vs ph) because corresponding correlation coefficient is regarded as significant.

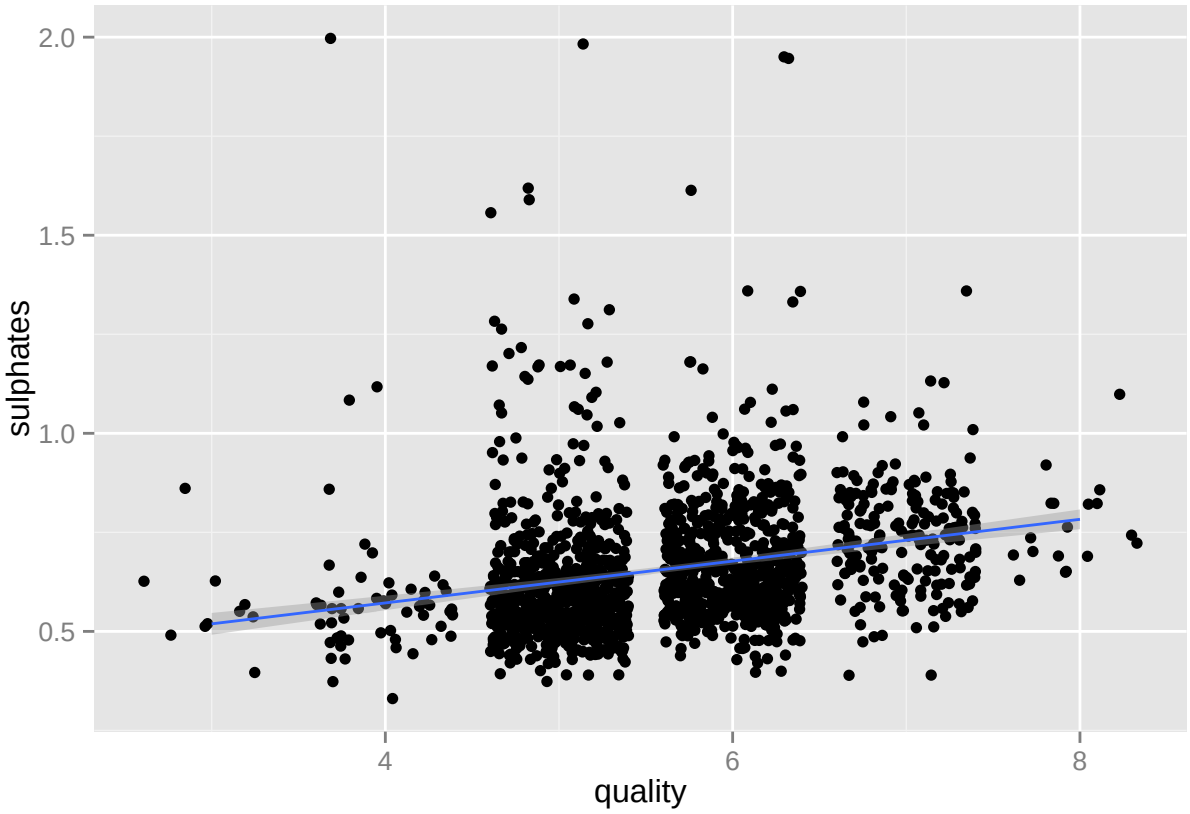
The following boxplot shows how the features affect `rating`:

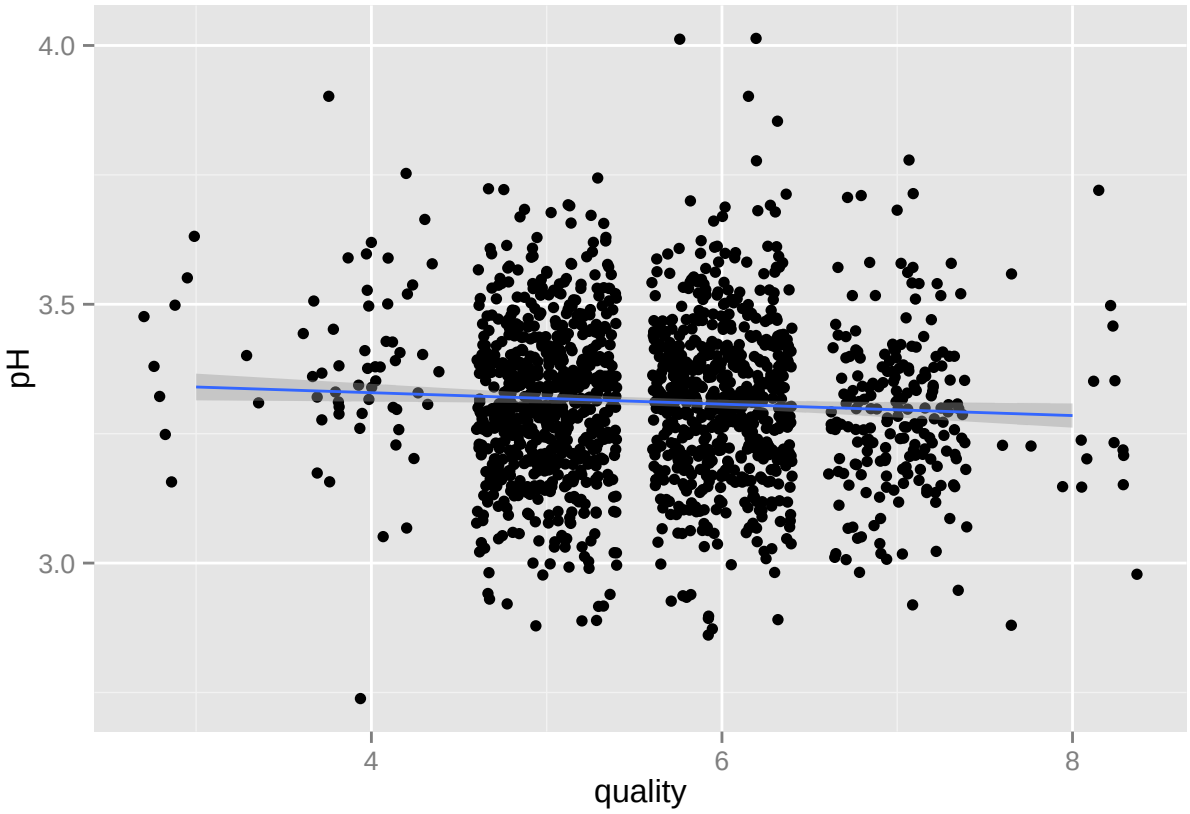

acid
residual sugar
chlorides
dioxide
density
pH
sulphates
alcohol
high
low
medium
rating

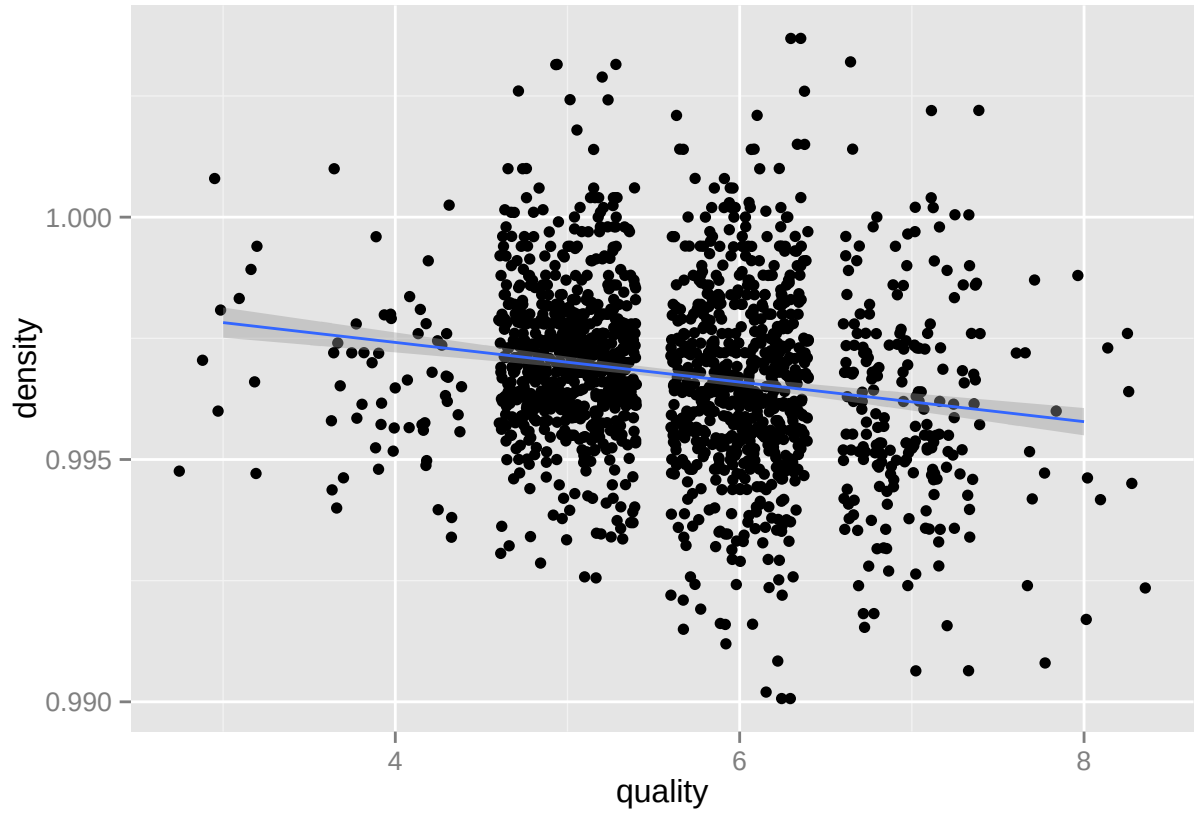
- Higher `alcohol`, higher `rating`.
- Higher `sulphates`, higher `rating`.
- Lower `pH`, higher `rating`.
- Lower `density`, higher `rating`.

We use `quality` to support above result




2.0
1.5
1.0
0.5
sulphates
4
6
8
quality

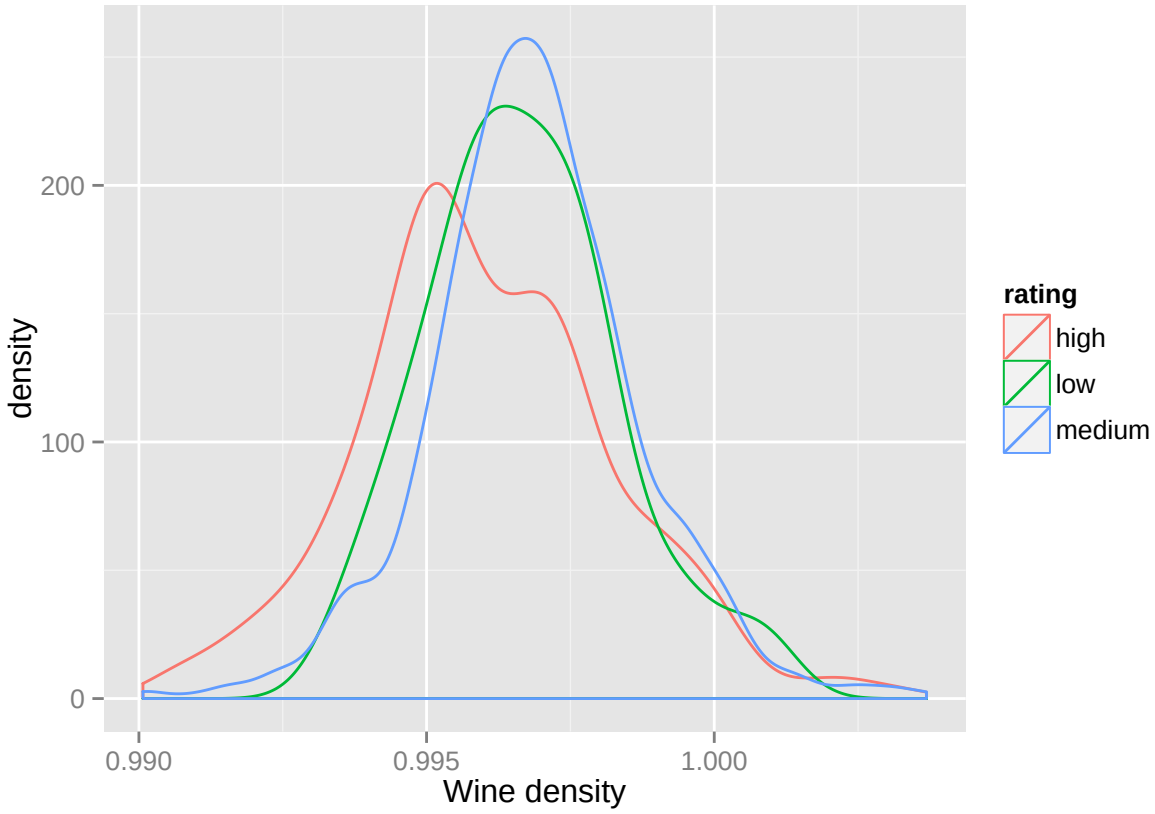

4.0
3.5
3.0
pH
4
6
8
quality

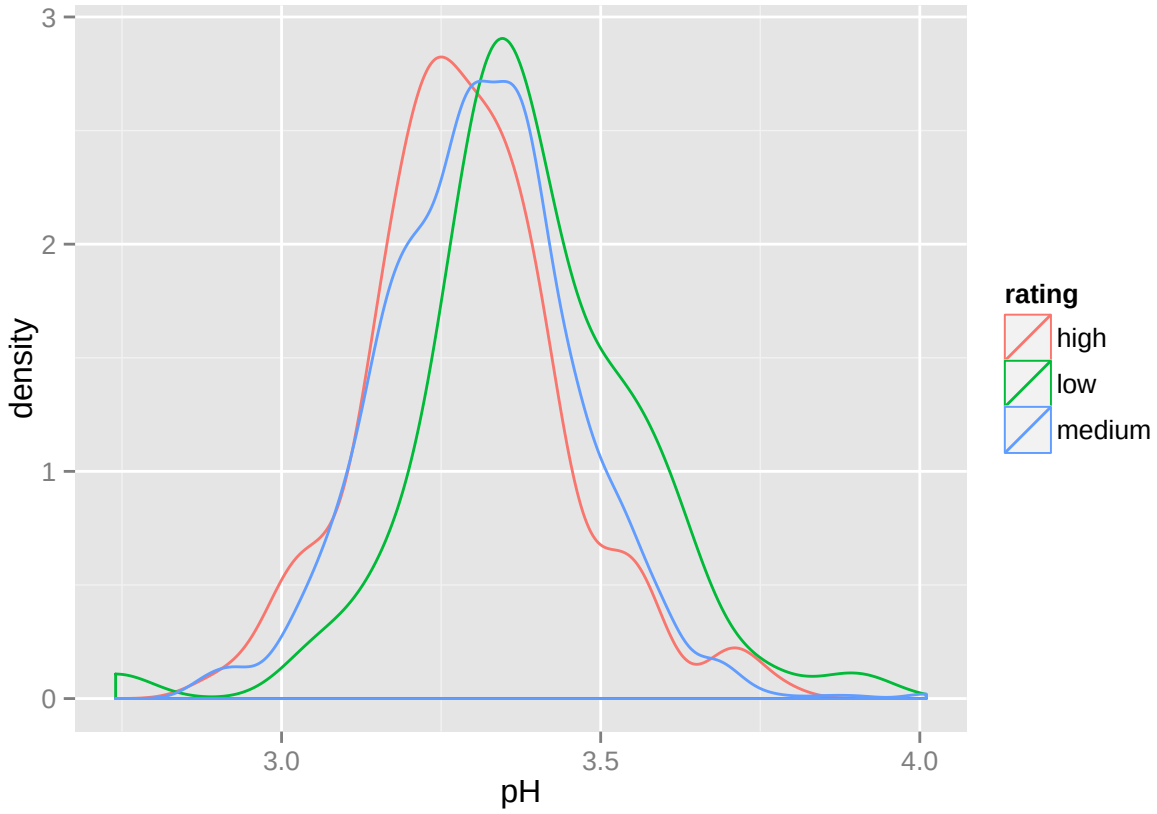


After using `quality`, we can see `alcohol` affect `quality` strongest.

- Higher `alcohol`, higher `rating`.
- Higher `sulphates`, higher `rating`.
- Lower `pH`, higher `rating`.
- Lower `density`, higher `rating`.

In Univariate Plots Section, we conclude that `density` and `pH` appear to be normally-distributed. Now we use the following plots to support the result


density
200
100
0
0.990
0.995
1.000
Wine density
rating
high
low
medium



It seems like wine with medium level appear to normally-distributed. The result is not suprised because medium level have larger sample size.

Short questions

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?

We find out `alcohol`, `sulphates`, `density` and `pH` affect `rating` or `quality`.

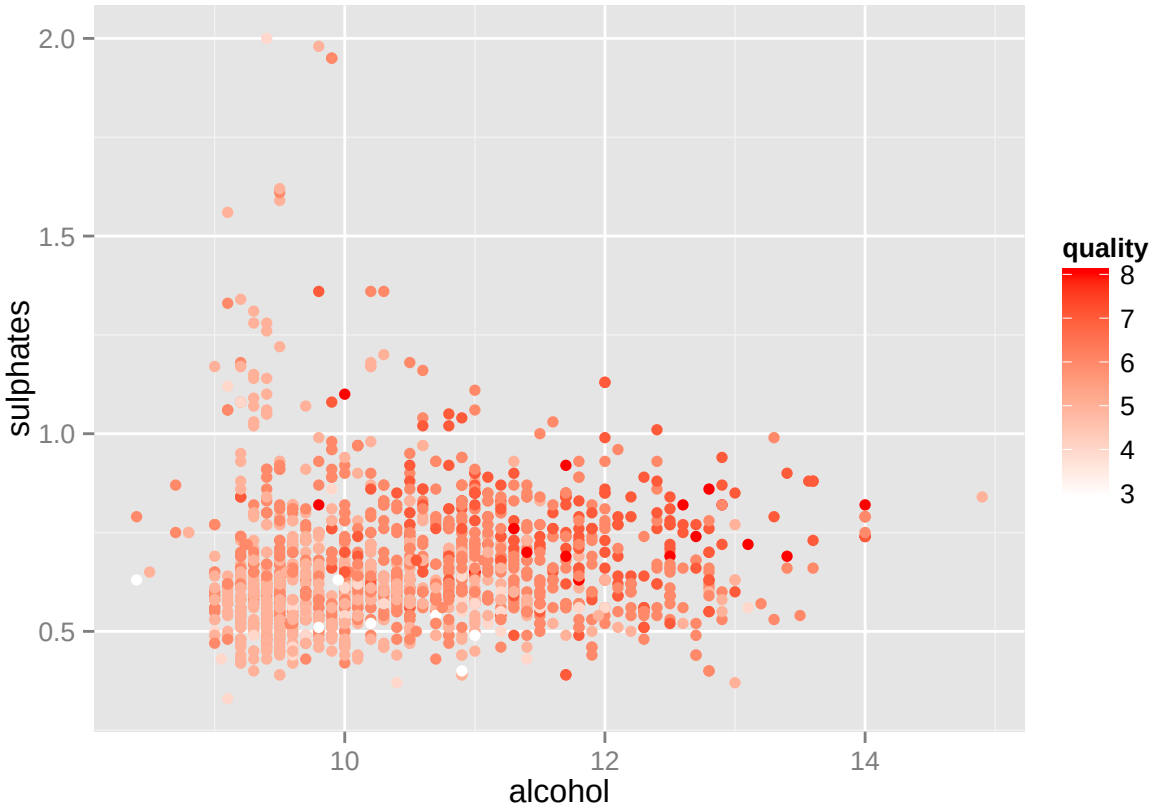
What was the strongest relationship you found?

Higher `alcohol`, higher `rating`.

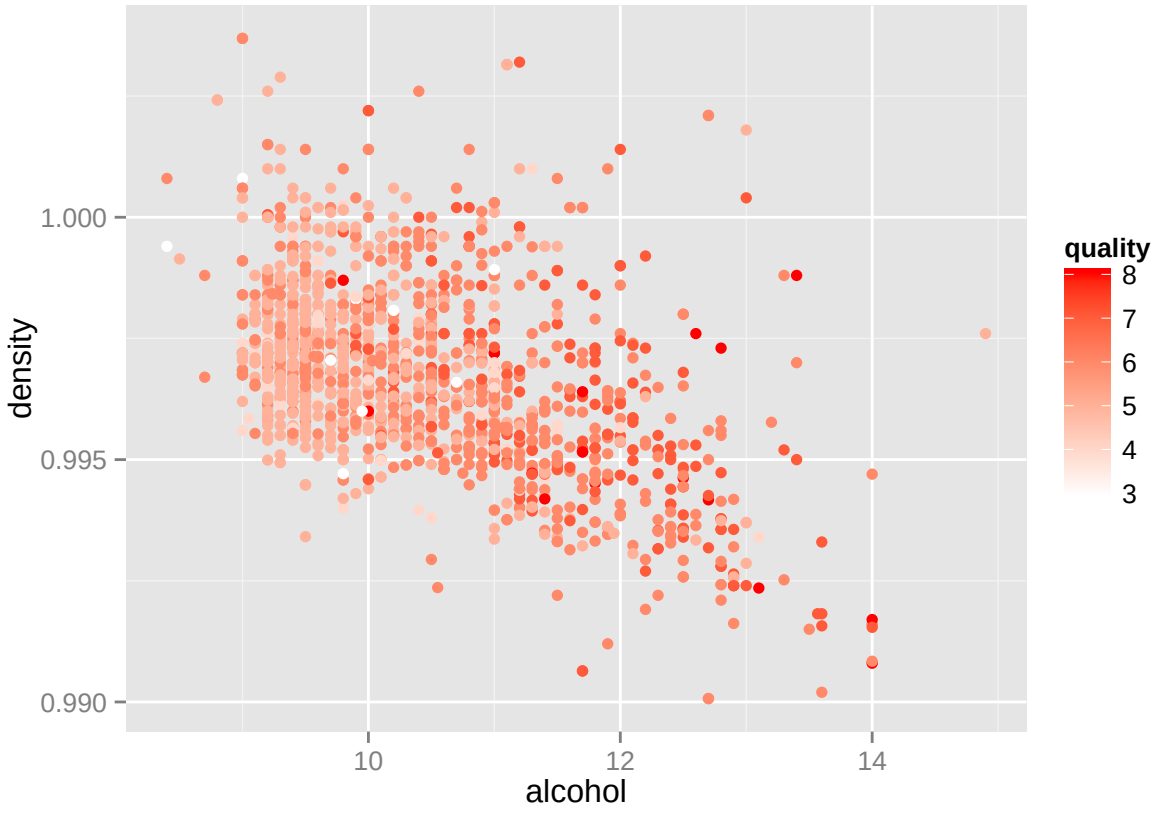
Multivariate Plots Section

Multivariate Analysis

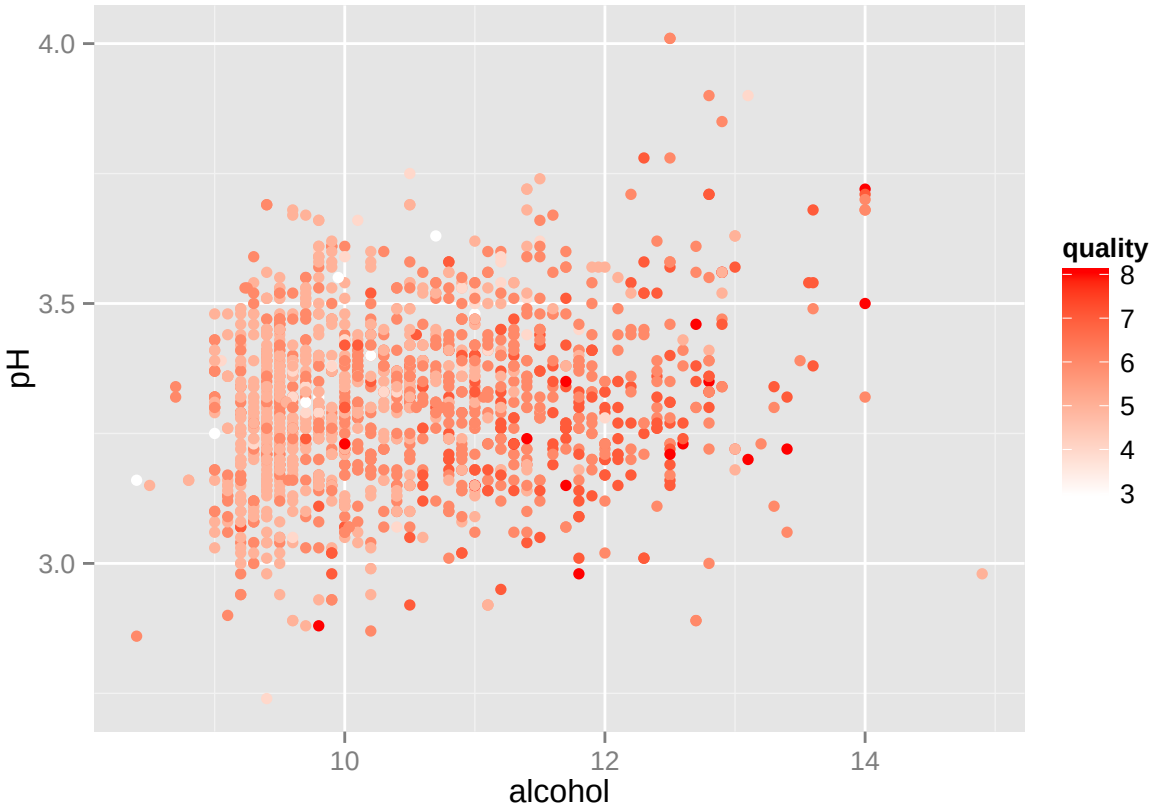
I examined the scatter plot of all the pair of features containing `alcohol`, `sulphates`, `density` and `pH` between different `quality`.



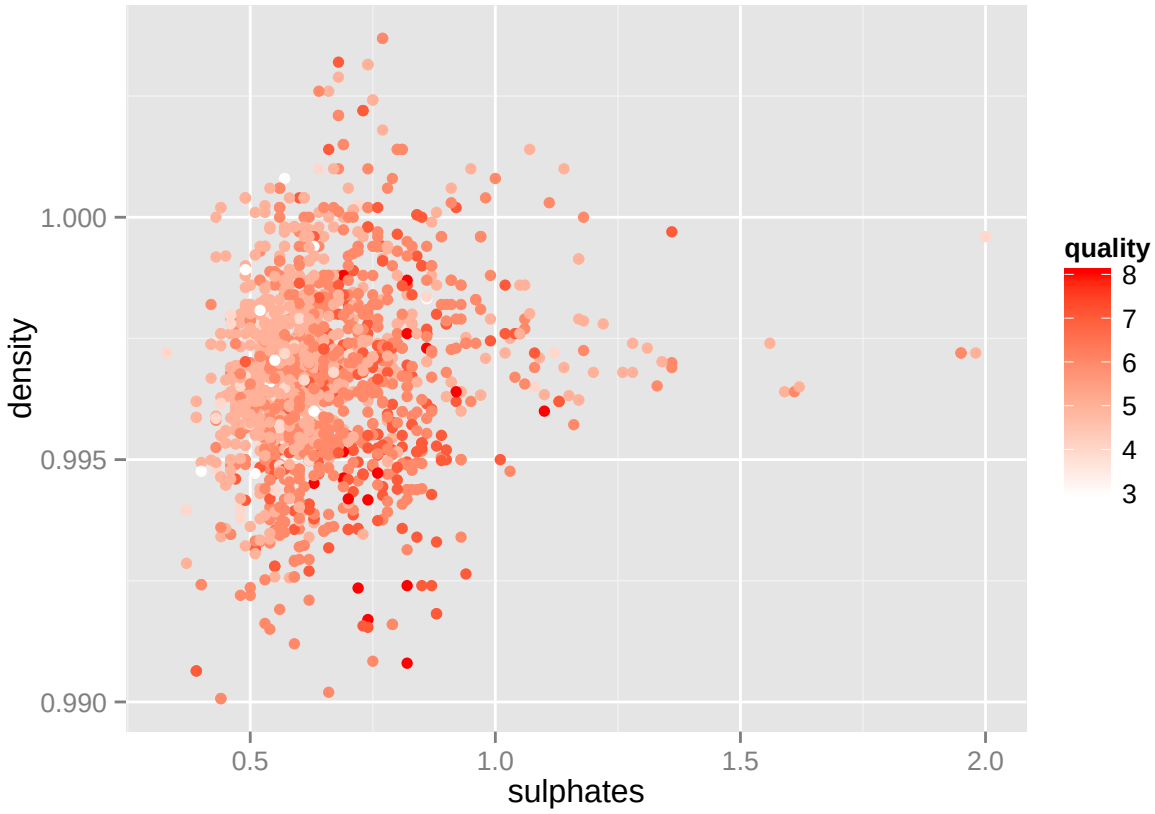
Lower `sulphates` and lower `alcohol`, lower `quality`.



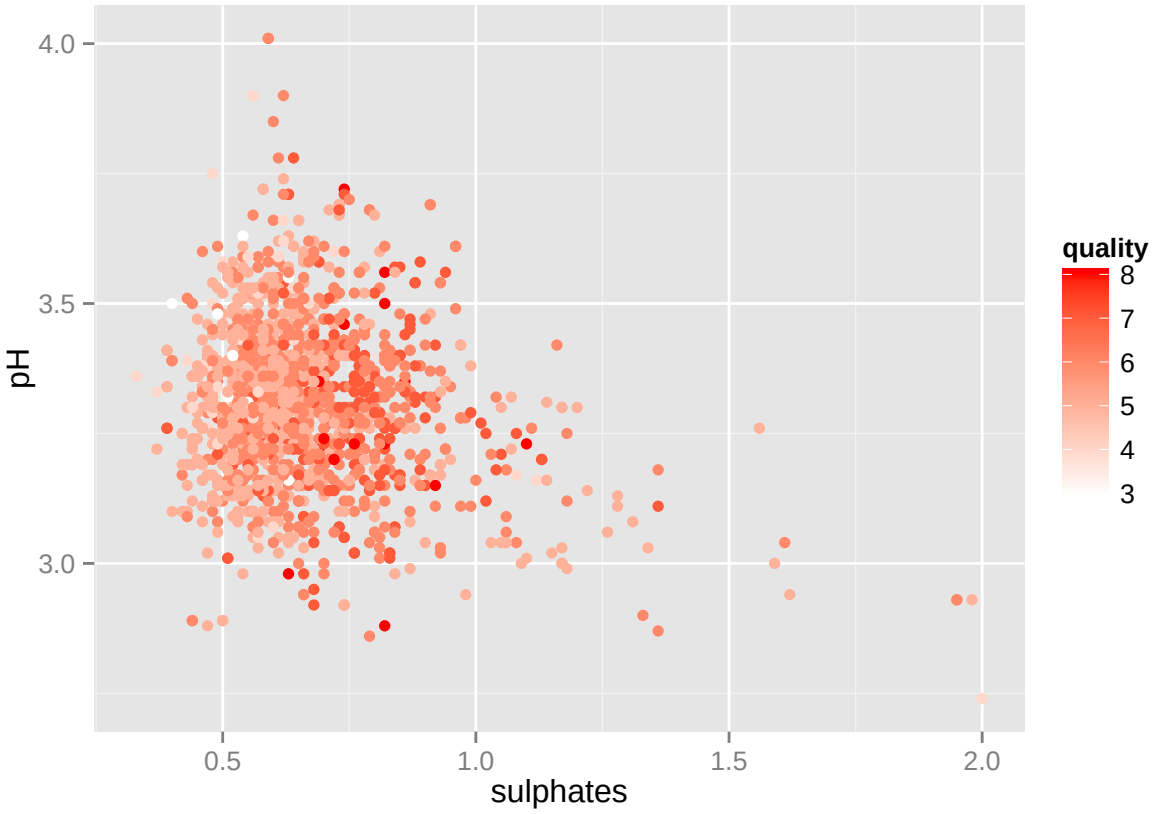
Higher `density` and lower `alcohol`, lower `quality`.



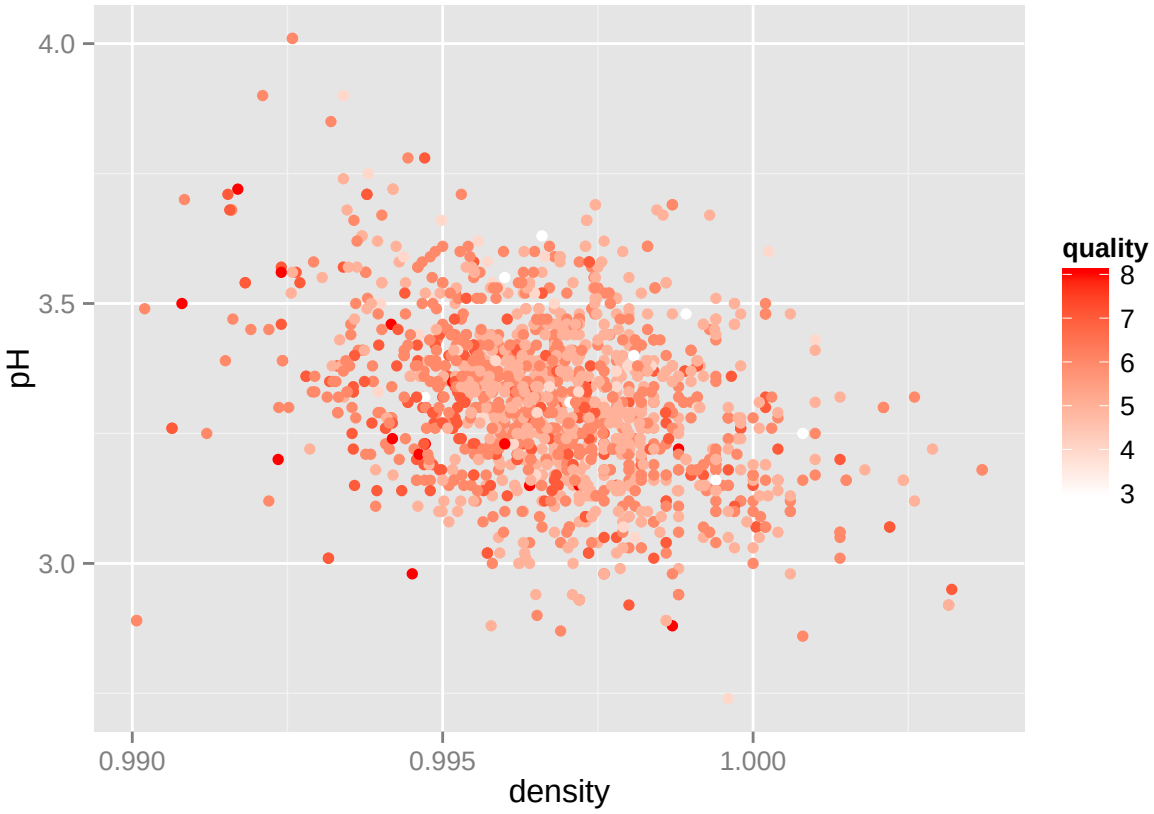
Higher `pH` and lower `alcohol`, lower `quality`.



Lower `sulphates`, lower `quality`.



Lower `sulphates`, lower `quality`.



Higher `pH` and higher `density`, lower `quality`.

The above plots support the result in Bivariate Plots Section.

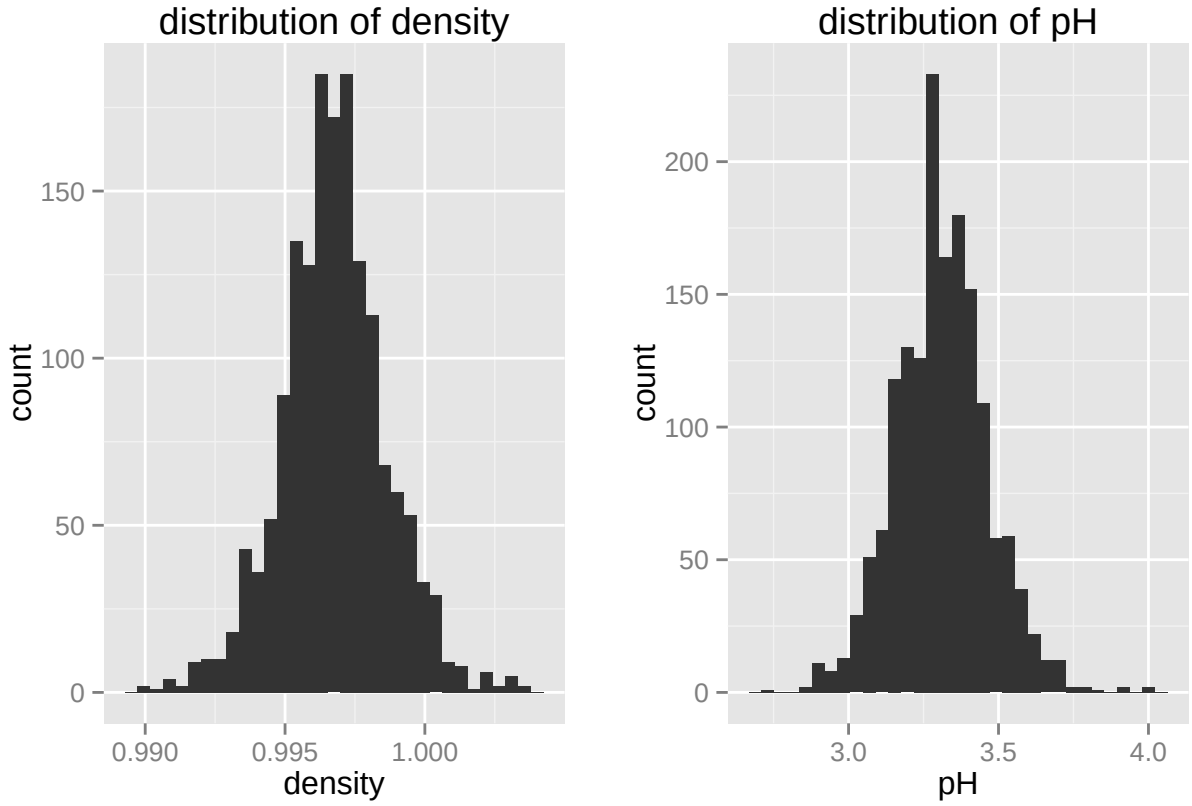
Short questions

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?

Higher `density` and lower `alcohol`, lower `quality`.

Final Plots and Summary

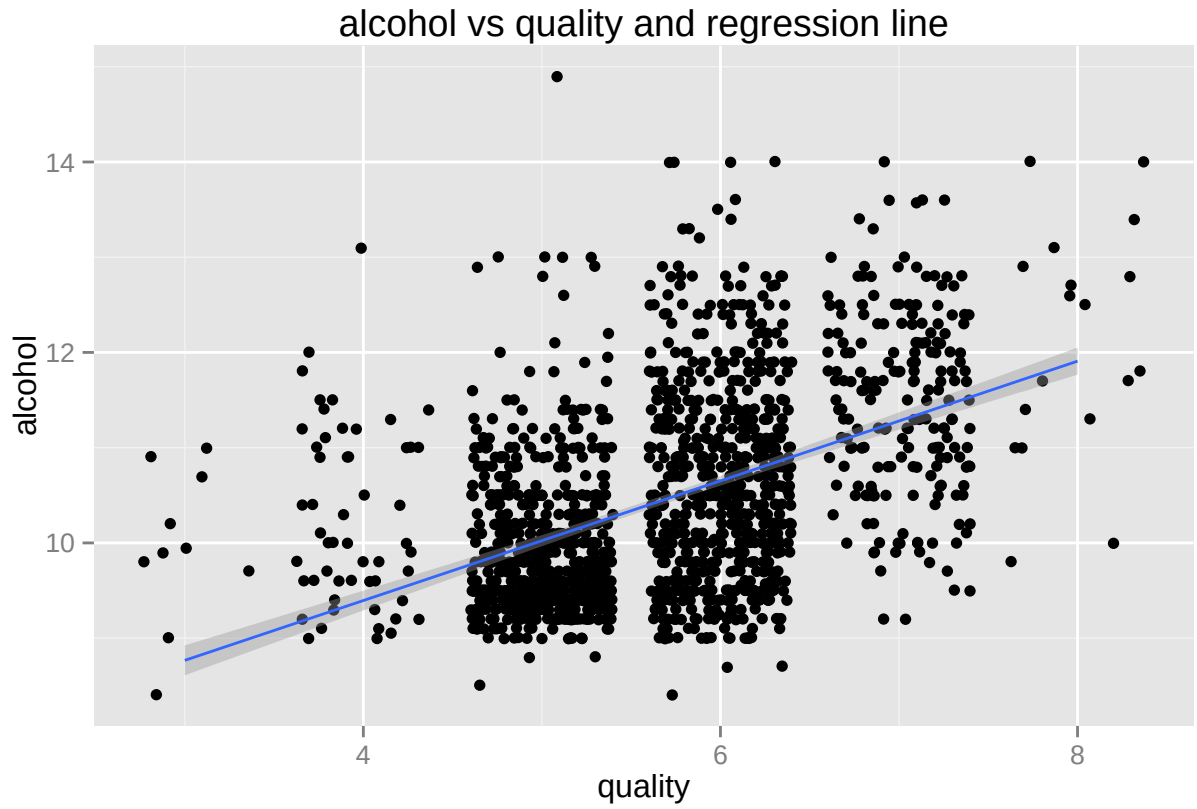
Plot One



Description One

`density` and `pH` appear to be normally-distributed.

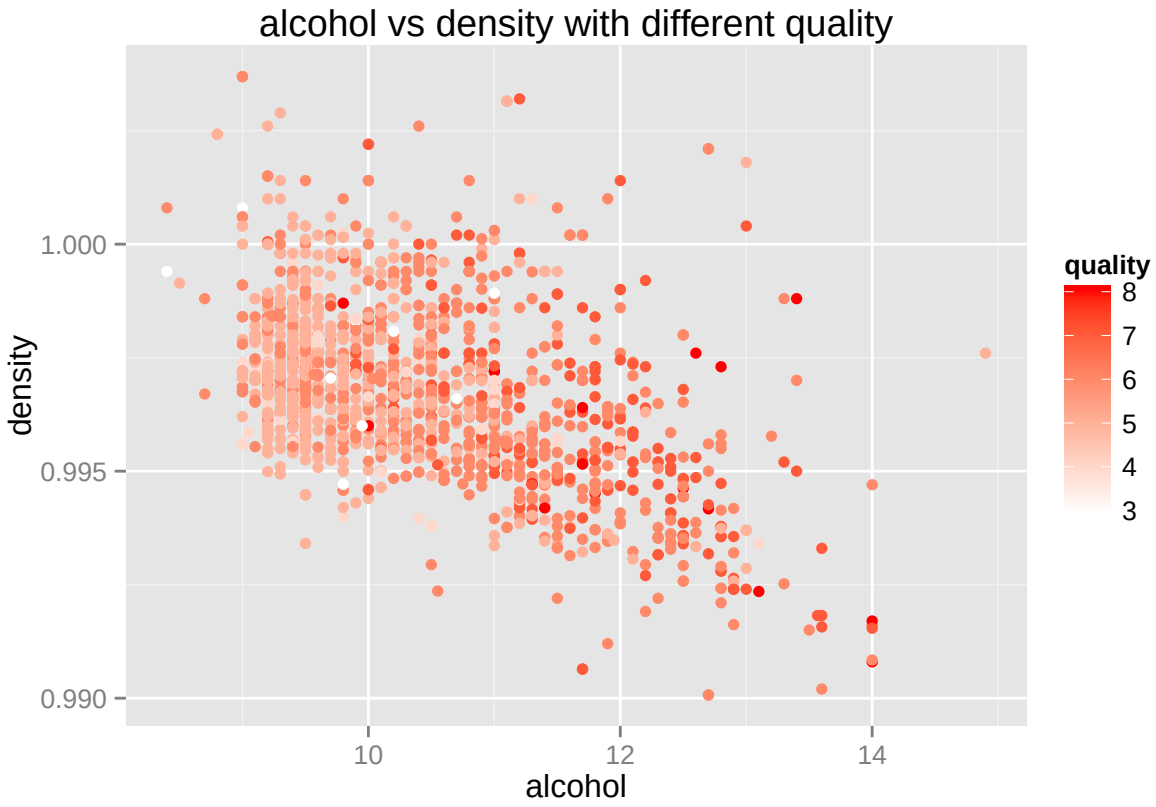
Plot Two



Description Two

Higher `alcohol`, higher `quality`.

Plot Three



Description Three

Higher `density` and lower `alcohol`, lower `quality`.

Reflection

Through this exploratory data analysis, I think feature `alcohol` influence the quality of red wines, however, wine experts give many 5 and 6 score of measure of wine quality, maybe just use the data of quality score {3,4} compare to {7,8} will show clearly patterns. I think data visualiztion technique is hard to make decision, it provides too many information, sometimes we just want to know what is the quality of this wine, further study with machine learning could be done to predict the wine quality.