Red Wine Quality Analysis

Li Cai

March 13, 2017

In this project, I will explore the dataset about red wind quality to find which chemical properties influence it.

Summary Statistics

```
## '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 0.0
6 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 0.0
75 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 0.9
98 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.
51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0
.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9
.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 .
..
```

The dataset has 1599 ovservations of 13 variables, and varicable X can be simply seen as index, so I'd like to remove this column.

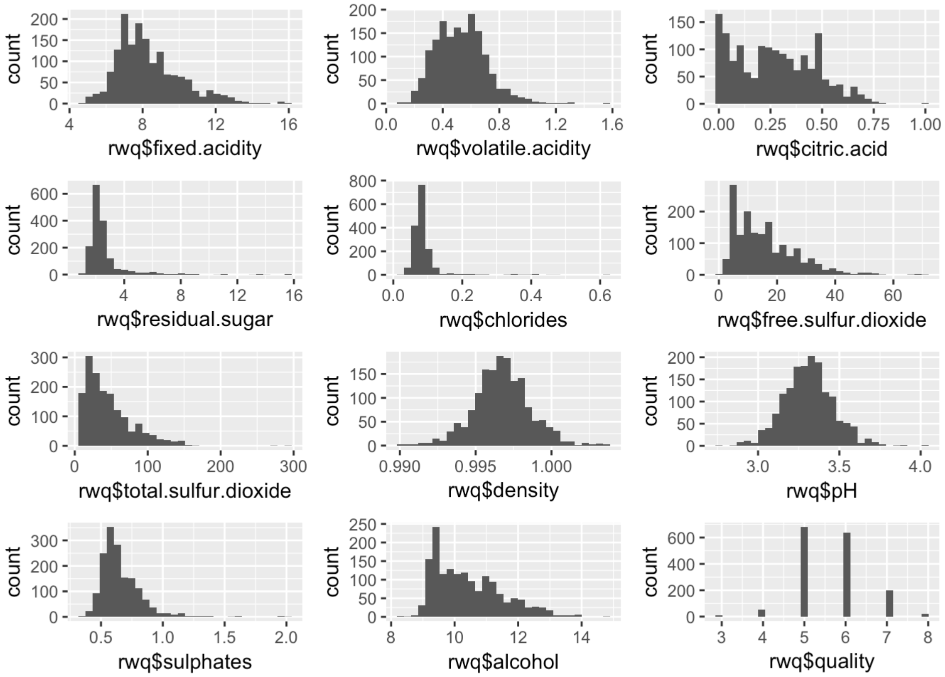
Now, I have 12 variables. Since the most important part is quality, it would be necessary to see the basic statistics on it.

```
##  fixed.acidity   volatile.acidity  citric.acid    r
esidual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   M
in.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1
st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   M
edian : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   M
ean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3
rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   M
ax.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur
.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.
00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.
00
##  Median :0.07900   Median :14.00       Median : 38.
00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.
47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.
00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.
00
##     density             pH          sulphates
alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300
Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200
Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581
Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000
Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

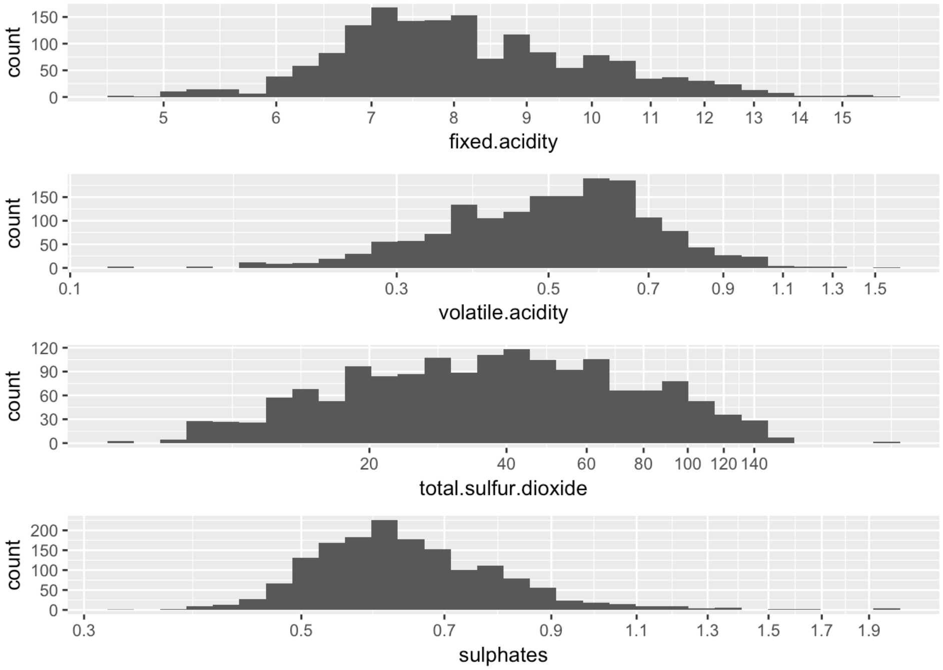
Since the most important part is quality, it would be necessary to see the basic statistics on it. From the dataset documentation, the wine quality is rated on 0-10, actually, the quality values on this data set has the lowest rate 3 and highest rate 8, with a median of 6 and a mean of 5.6.

Univariate Plots Section

At first, I will plot 12 simple historams of all 12 variables to see how they distribute.



These 12 histograms reveal that density and PH are normal disributed but fixed.acidity, volatile.acidity, total.sulfur.dioxide, and sulphates seem to be long-tailed.



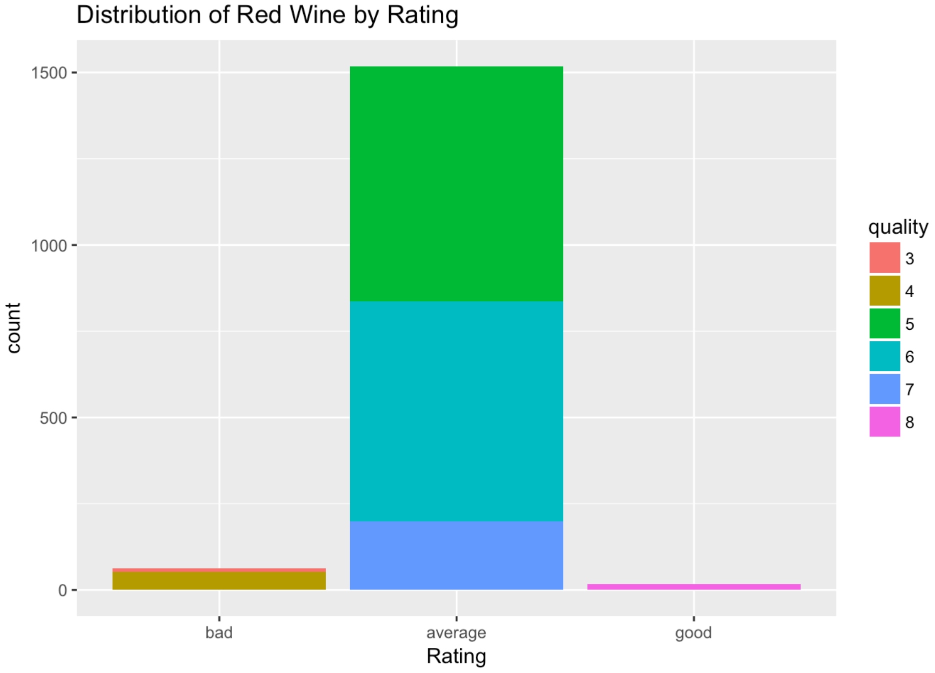
On log10 scale, fixed.acidity, volatile.acidity, total.sulfur.dioxide, and sulphates appear to be normal distributed, although they are still with some outliers.

In this dataset, quality rate is from 3-8, and the large majority of the red wines are rated from 5-7. It seems to make sense that create a new rateing including bad(0-4), avreage(5-7) and good(8-10).

```
##      bad average     good
##       63     1518       18
```

The rating result is that more than 90% of red wine are ‘average’.

Plot it.



It’s more straight to see taht most wine are in average rating.

Univariate Analysis

What is the structure of your dataset?

Now, there are 1599 obsevations of 13 variables in the data set, short discriptions of 13 variables are as follows:

- fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- volatile acidity: the amount of acetic acid in wine
- citric acid: add ‘freshness’ and flavor to wines
- residual sugar: the amount of sugar(gram/liter) remaining after fermentation stops
- chlorides: the amount of salt in the wine
- free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion
- total sulfur dioxide: amount of free and bound forms of S02
- density: the density of water is close to that of water depending on the percent alcohol and sugar content
- 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
- sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
- alcohol: the percent alcohol content of the wine
- quality: score between 0 and 10
- rating: rated on bad(0-4), average(5-7), good(8-10)

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

Quality of the red wine is the main featur in the dataset and I will also take a look at how other variables would influence the quality of the wine.

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

The Variable residual.sugar seems to be an intersting one, and I'll explore other variables like alcohol and PH.

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

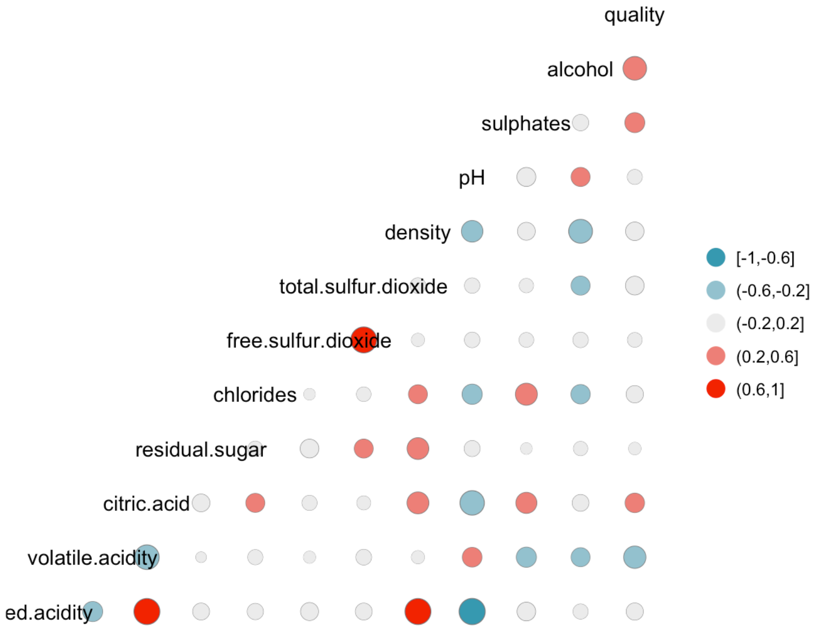
Yes, rating is the variable I created.

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?

As I meantioned above, density and PH are normal disributed but fixed.acidity, volatile.acidity, total.sulfur.dioxide, and sulphates seem to be long-tailed. Also, I remove X from orininal dataset, because it just the index.

Bivariate Plots Section

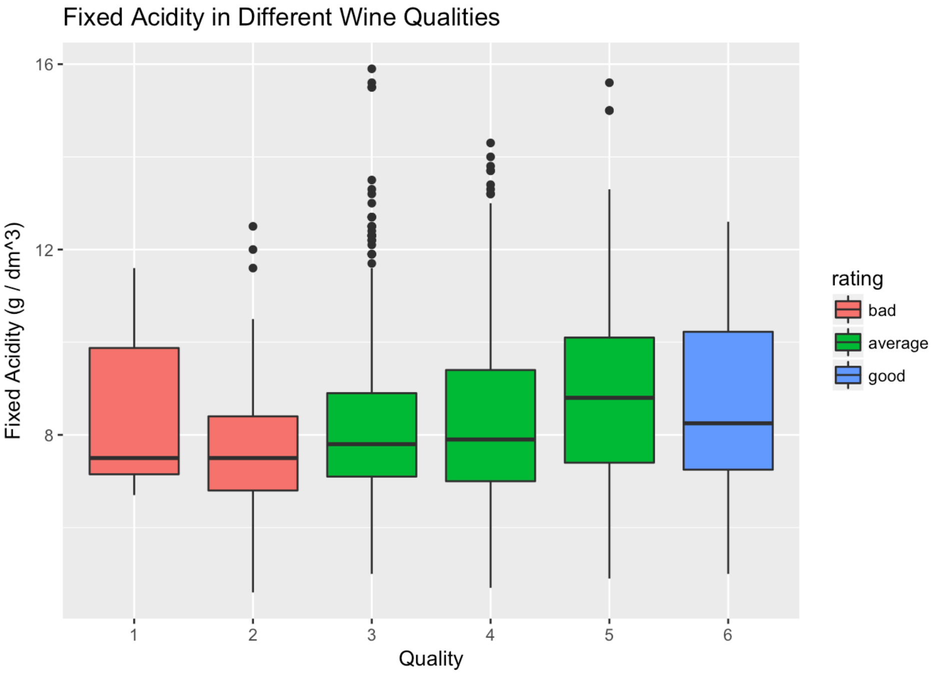
First, I'd like to plot the correlation of all varianles against each other.



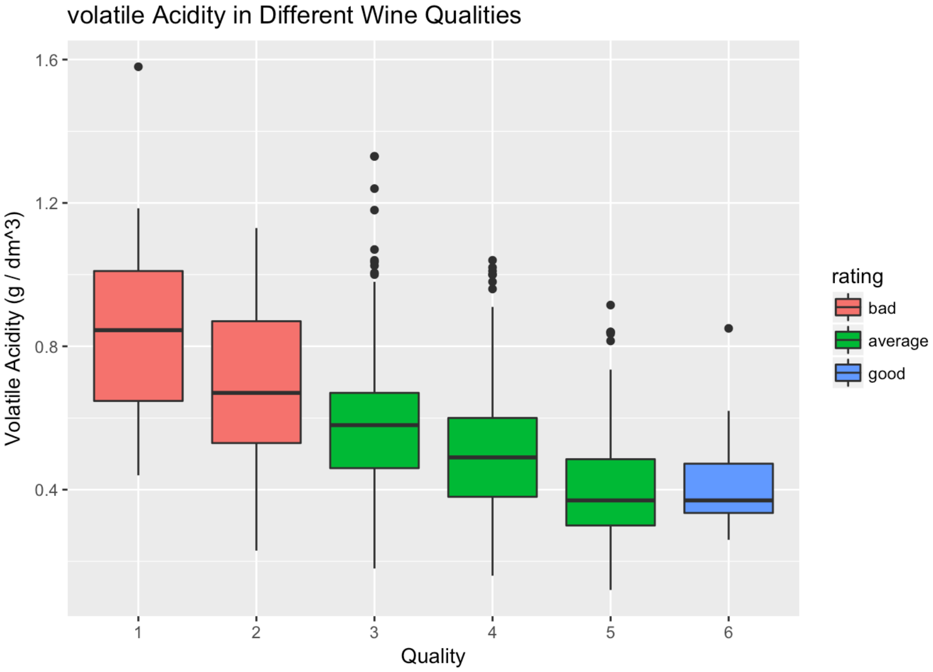
Clearly, there are some strong correlations between some variables such as total.sulfur.dioxide and free.sulfur.dioxide, volatile.acidity and fixed.acidity, total.sulfur.dioxide and fixed.acidity, fixed.acidity and PH.

Also, this plots tells that quality has higher corelations with alcohol, sulphates, density, total.sulfur.dioxide, chlorides, citric.acid, volatile.acidity and fixed.acidity than other variables.

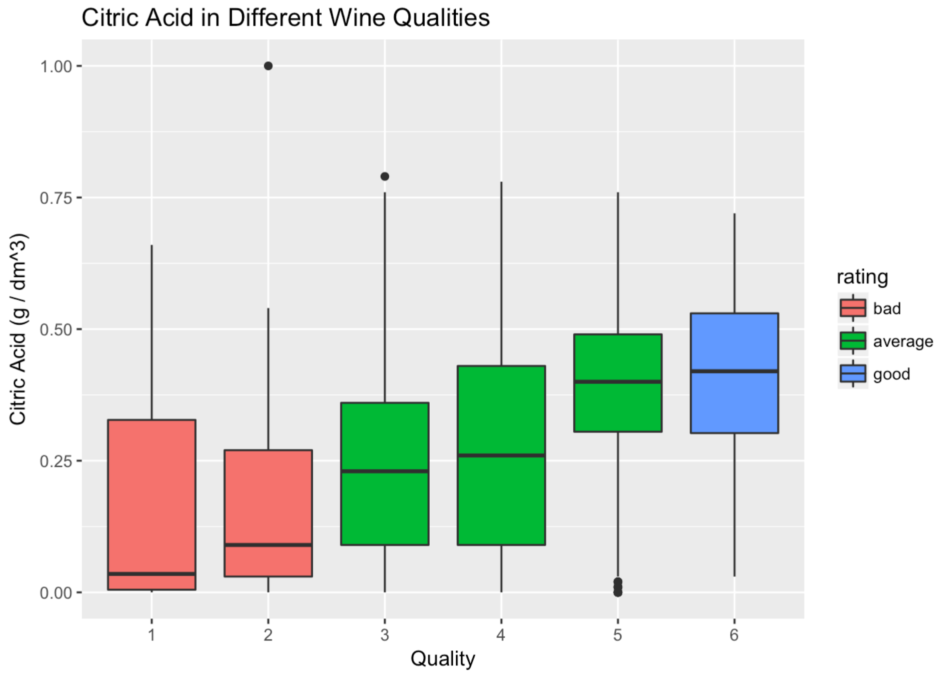
Create boxplots of these variables vs. quality can see how they affect the quality of the red wine.



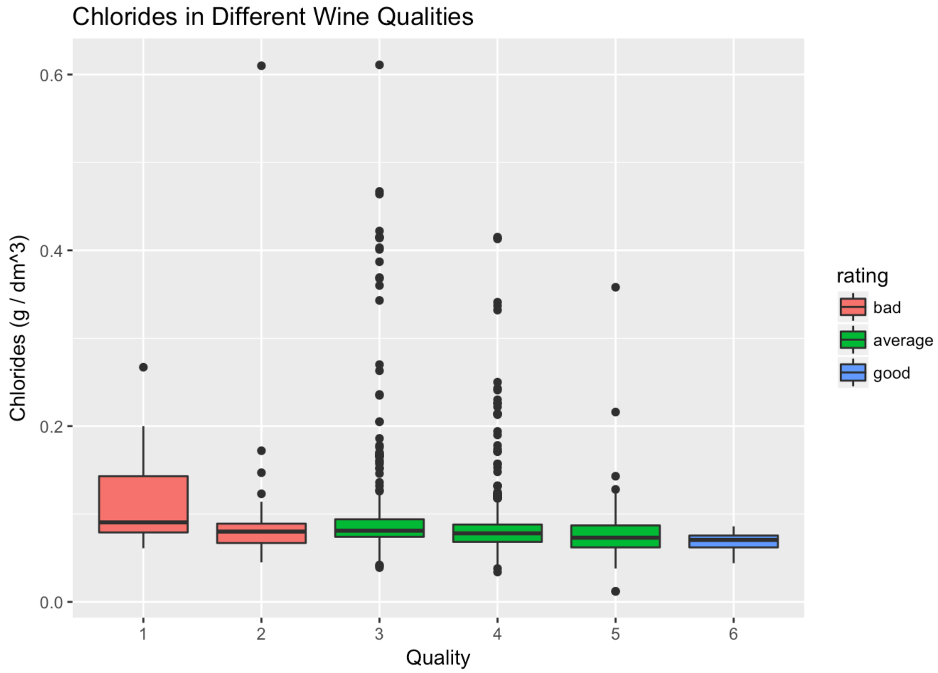
It seems like that higher quality red wine has higher fixed.acidity, but it's not obvious.



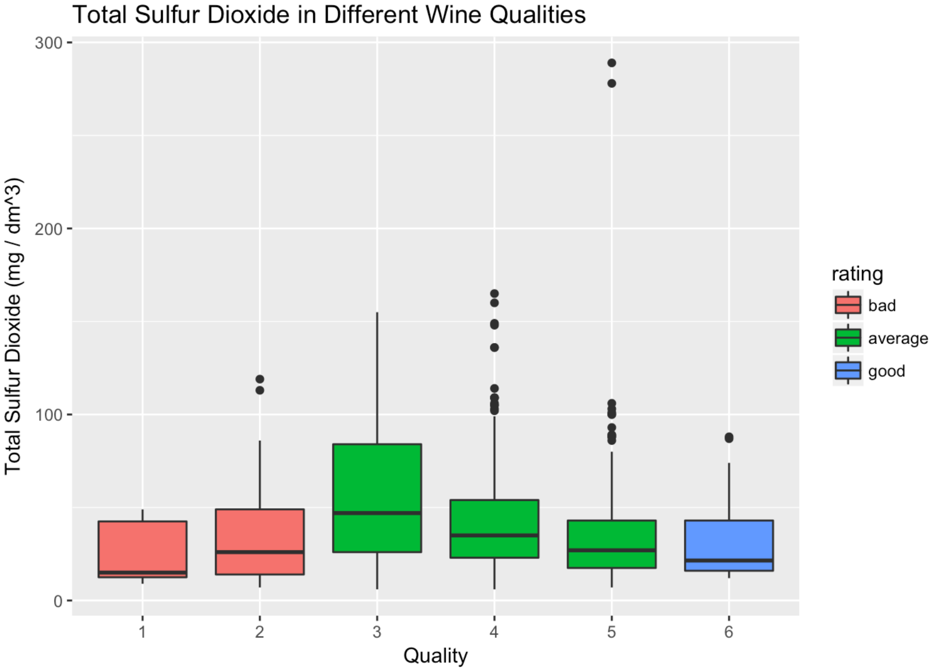
It's clear that higher quality red wine has lower volatile.acidity.



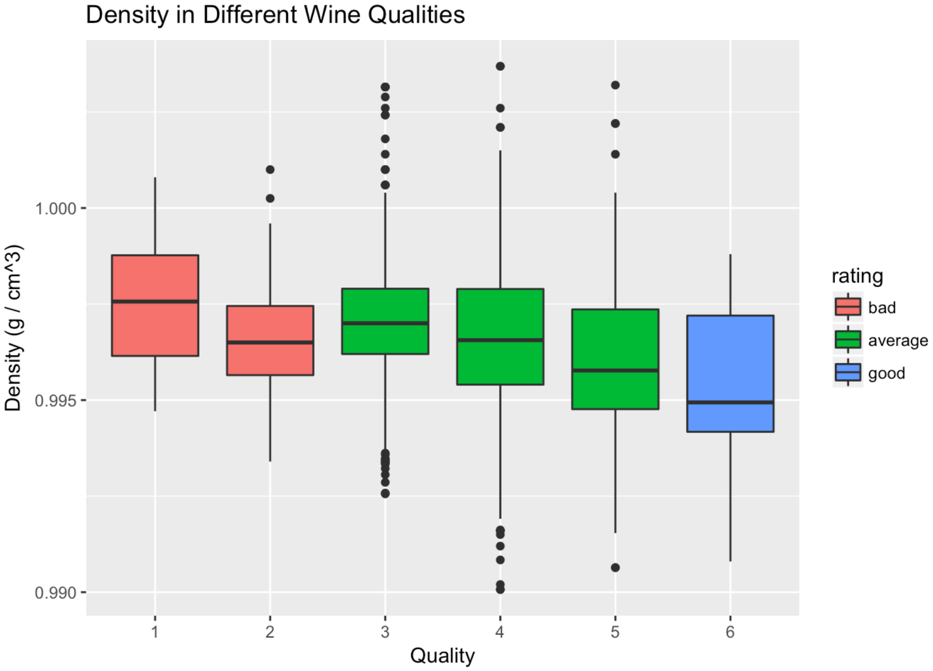
It’s clear that higher quality red wine has higher citric.acid.



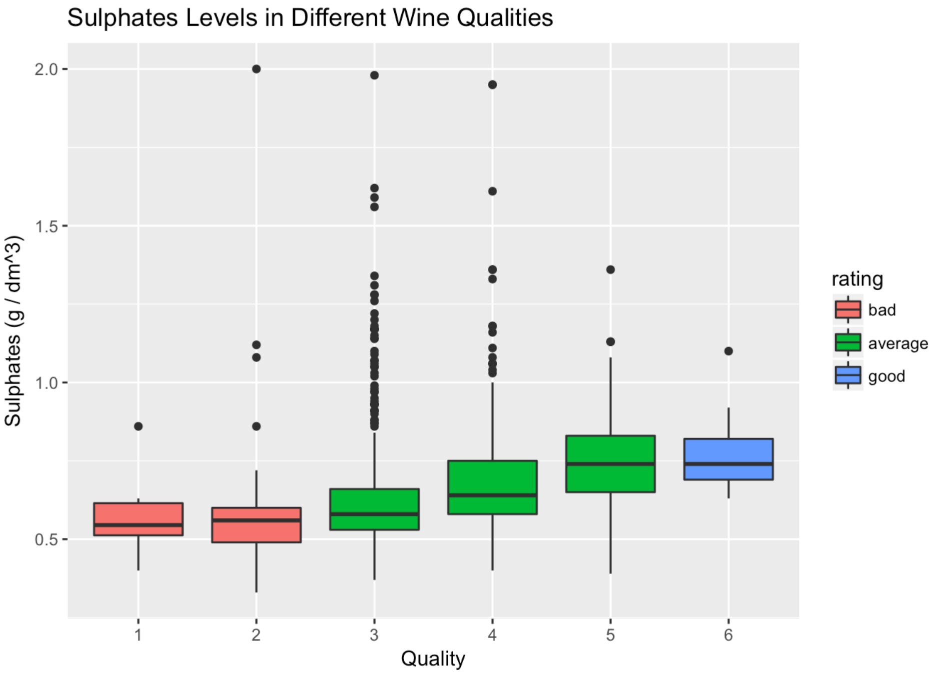
It seems like that higher quality red wine has lower chlorides.



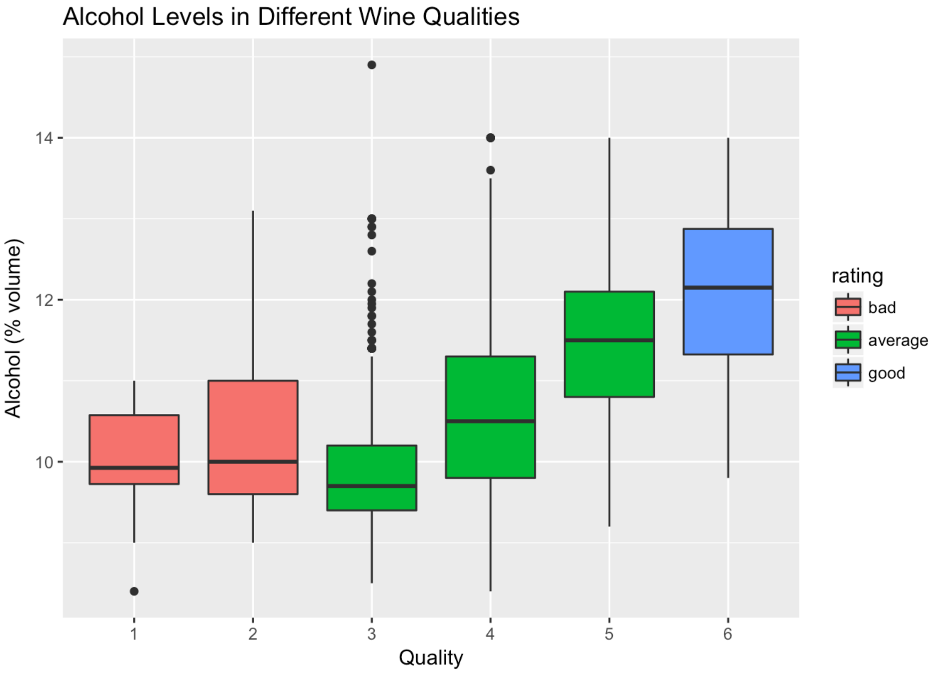
Cannot see specific relationship between total.sulfur.dioxide and quality.



It shows that higher quality red wine has lower density.



It shows that higher quality red wine has higher sulphates.



It shows that higher quality red wine has higher alcohols.

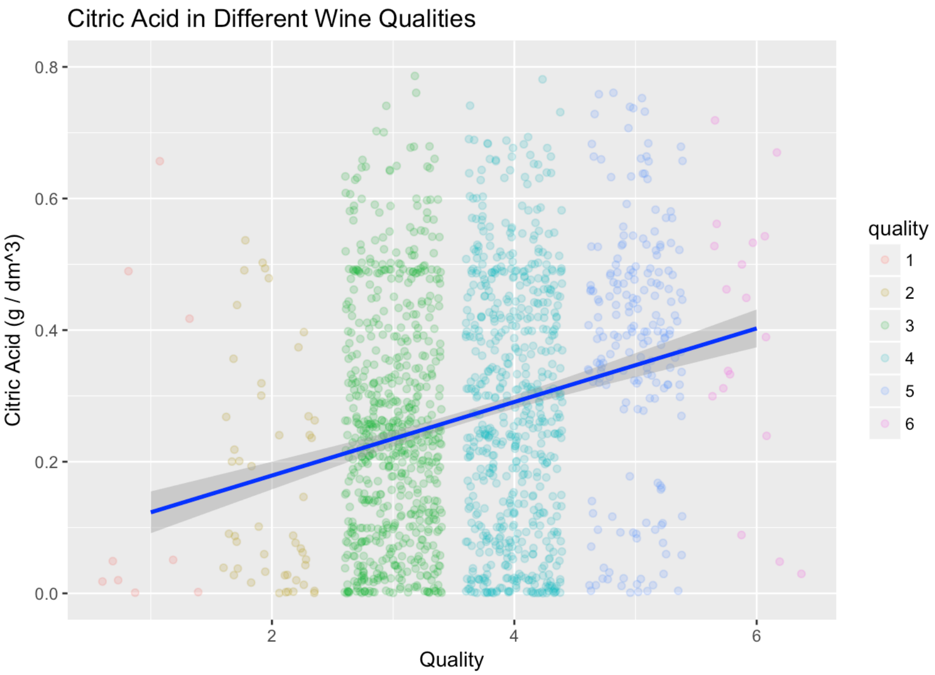
After explored all boxplots, there are no obvious relayionship between total.sulfur.dioxide and quality and I found that a good wine seems has the following characteristic:

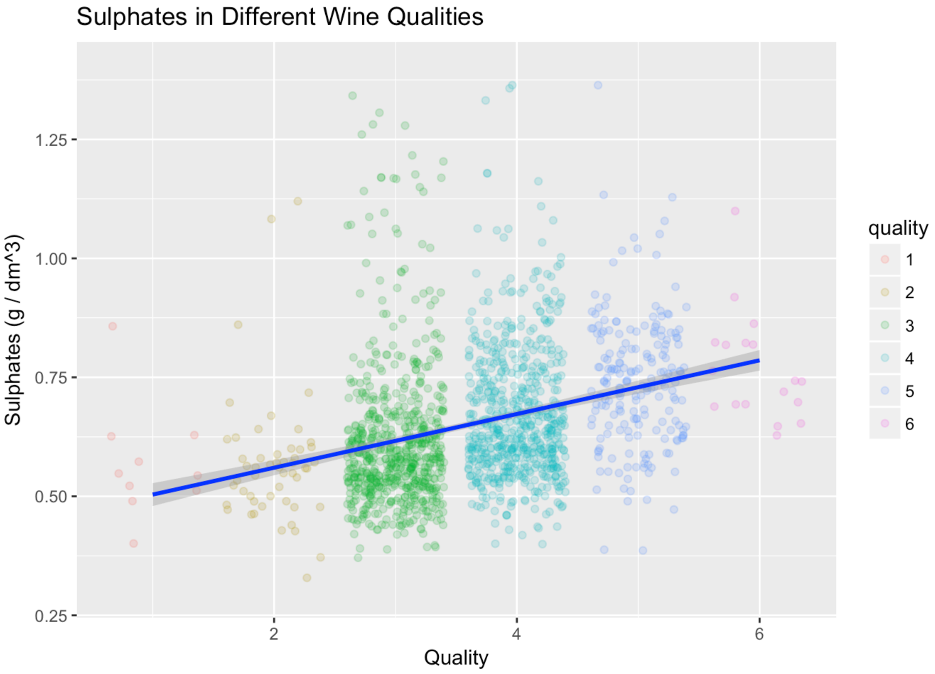
- higer fixed.acidity, citric.acid, sulphates, and alcohol
- lower volatile.acidity, chlorides and density

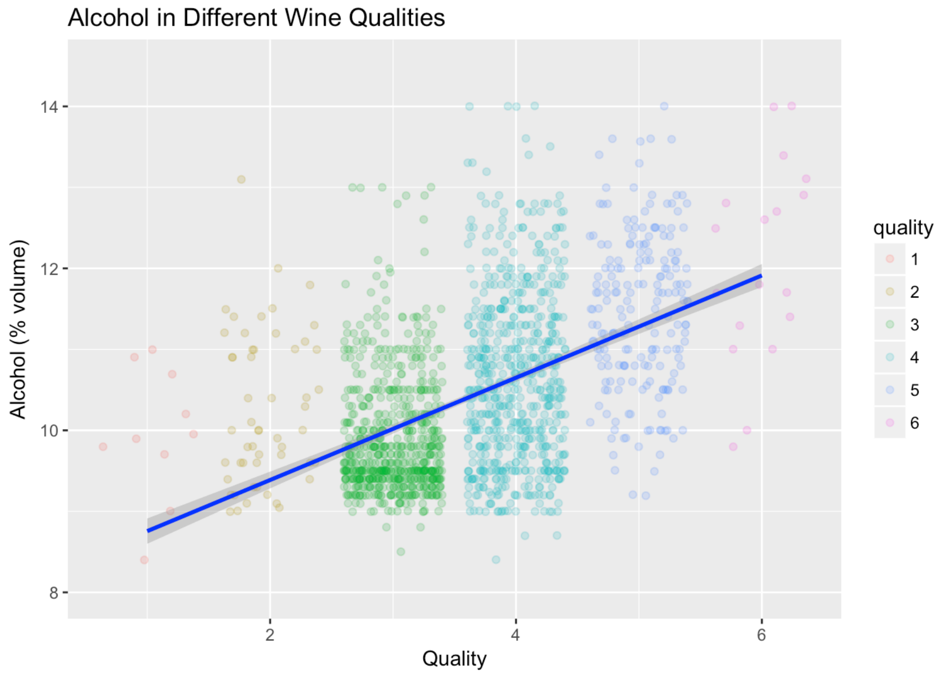
And then I will calculate the correlation for each of these seven variable against quality.

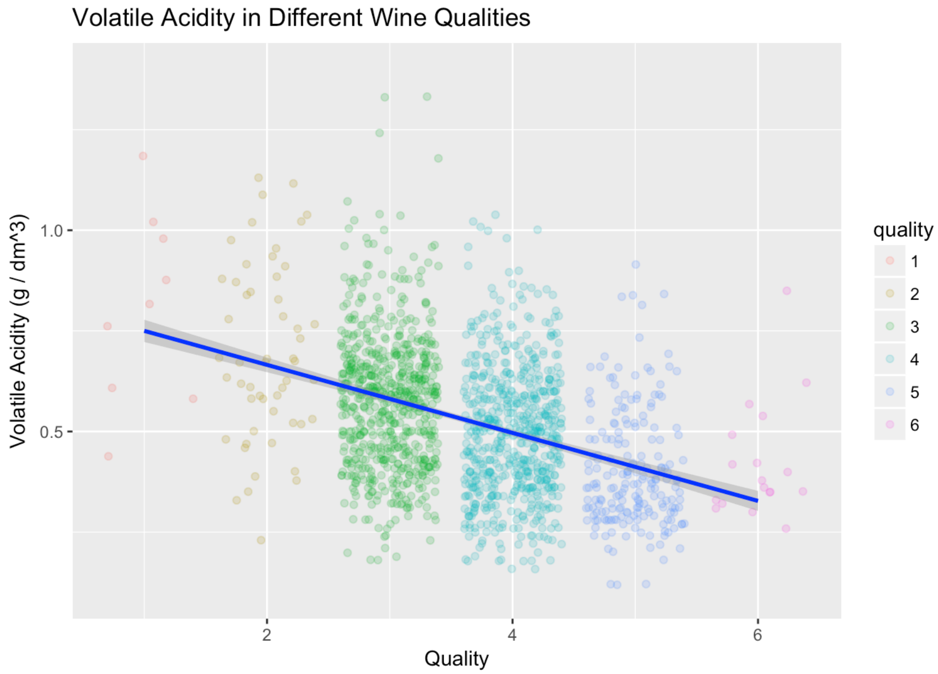
```
##    fixed.acidity      citric.acid        sulphates
alcohol
##        0.1240516        0.2263725        0.2513971
0.4761663
## volatile.acidity        chlorides          density
##       -0.3905578       -0.1289066       -0.1749192
```

The result shows that the correlations between quality and fixed.acidity, chlorides, density are lower than 0.2, it’s pretty samll. And then I’ll plot the rest four variables vs. quality with removing some outliers.


Citric Acid in Different Wine Qualities

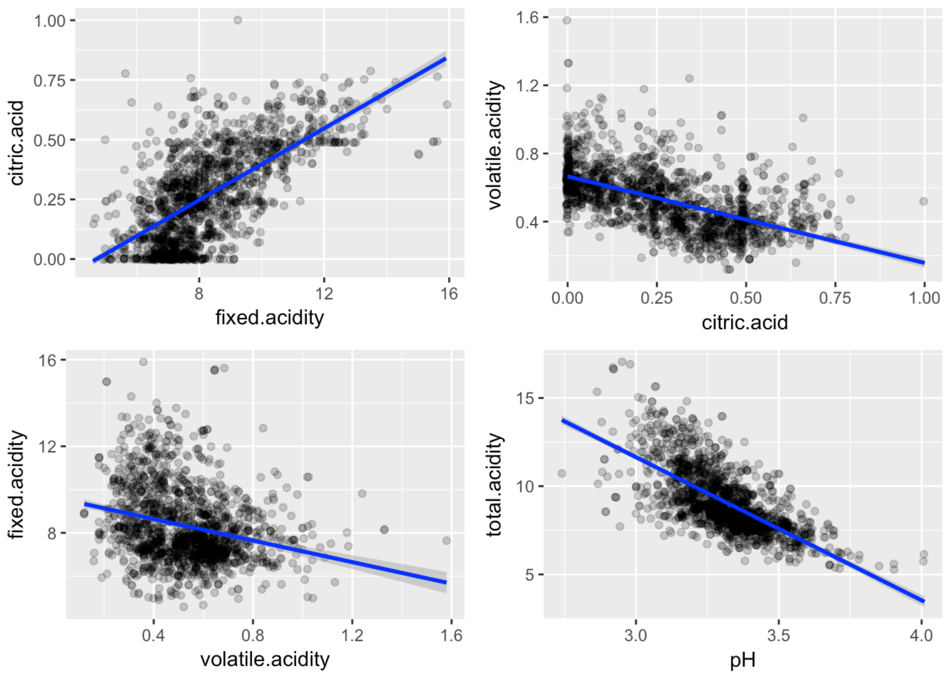

Sulphates in Different Wine Qualities





It also will be interesting to see the correlations between variables about acid, and acidity against PH. First, calculate correlations and then create the scattor plots.

```
##      fixed.acidity_citric.acid   citric.acid_volati
le.acidity
##                      0.6717034
-0.5524957
## volatile.acidity_fixed.acidity             PH_tot
al.acidity
##                     -0.2561309
-0.6834838
```



Obviously, these variables has high correaltions. citric.acid and fixed.acidity, fixed.acidity and volatile.acidity have highly positive correlations, citric.acid and volatile.acidity, total.acidity and PH have highly negative correlations.

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? From the boxplot, I found some trends that a good red wine has. After the caluculation of eight variables against the quality of red wine and create scatterplot, four rariables were removed, and the result is a good wine seems has the following characteristic:

- higer fixed.acidity, citric.acid, sulphates, and alcohol
- lower volatile.acidity, chlorides and density

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

I observed the correlations between acid variables, and ph vs. total.acidity, results are as follows:

- fixed.acidity & citric.acid : 0.6717034
- citric.acid & volatile.acidity: -0.5524957
- volatile.acidity & fixed.acidity: -0.2561309
- PH & total.acidity: -0.6834838

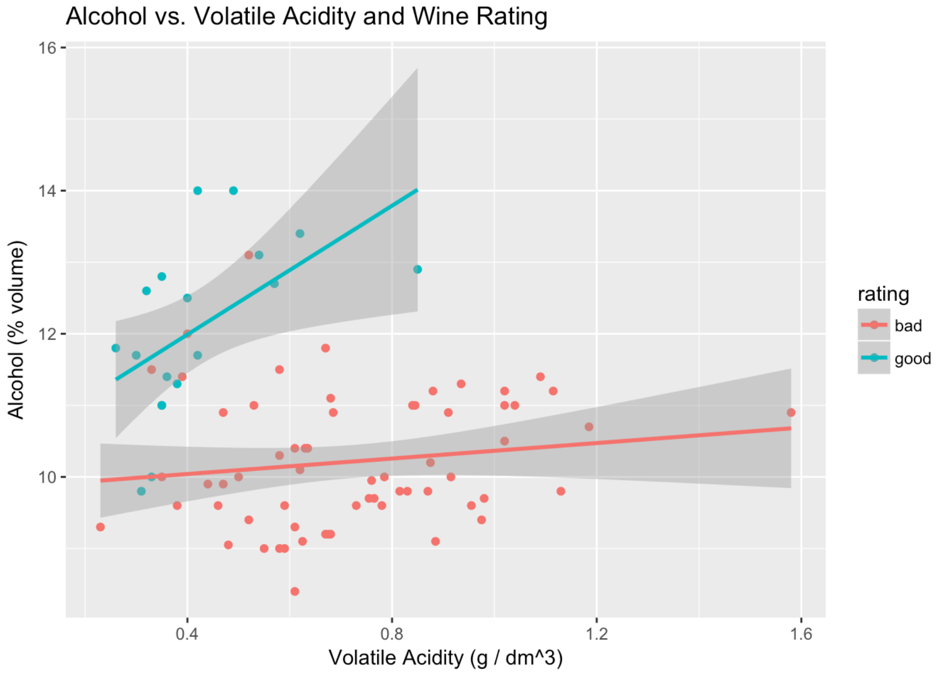
The most important relationship is between quality and alcohol, alcohol have the highest correlation(0.4761663) with quality of red wine.

What was the strongest relationship you found?

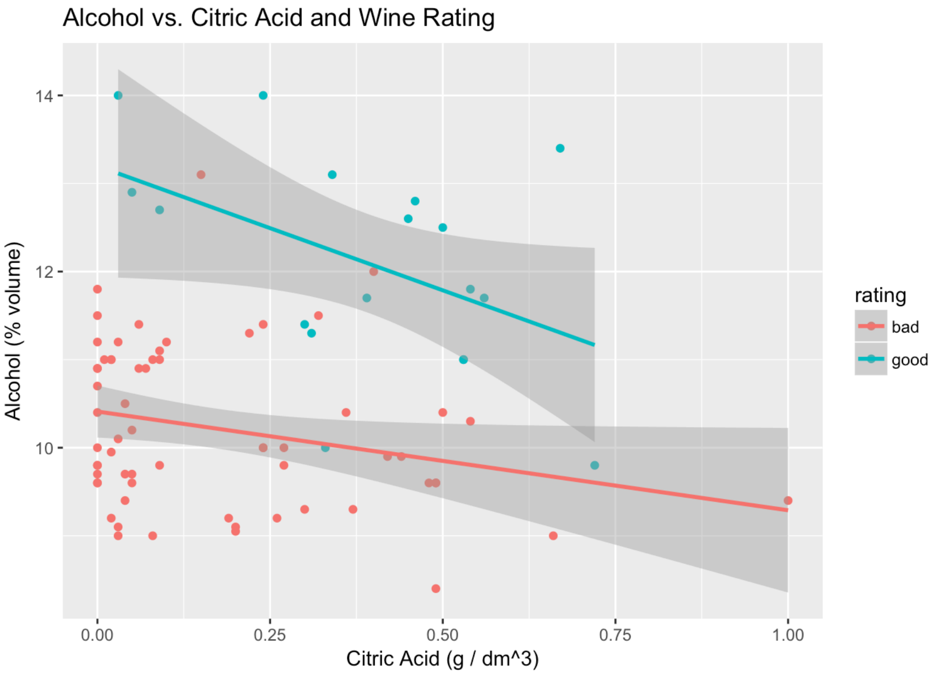
The strongest relationship I found is PH vs.total.acidity, which is more than -0.68, and the second is correlation between fixed.acidity and citric.acid(0.6717034).

Multivariate Plots Section

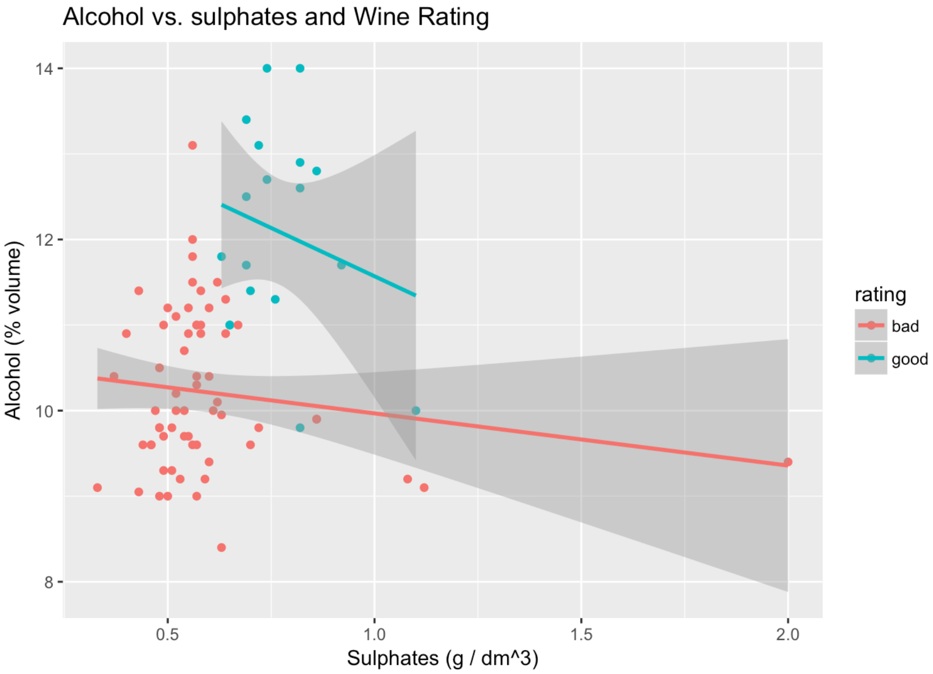
As mentioned above, alcohol has the highest correlation with the quality of re wines, so I will create multivariate plots that are alohol vs. other three variables and rating. For the rating, I will remove average rating before plot because average has the most quantities, remove it can make the plot more obvious to find the differences between 'good' and 'bad'.



For a 'good' red wine, it has higher alcohol and lower volatile.acidity, and it has more positive correlation between alcohol and volatile.acidity.



For a ‘good’ red wine, it has higher alcohol but is not clear for volatile.acidity, and it has more negative correlation between alcohol and citric.acid.



For a ‘good’ red wine, it has higher alcohol and higer sulphates, and it has more negative correlation between alcohol and sulphates.

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?

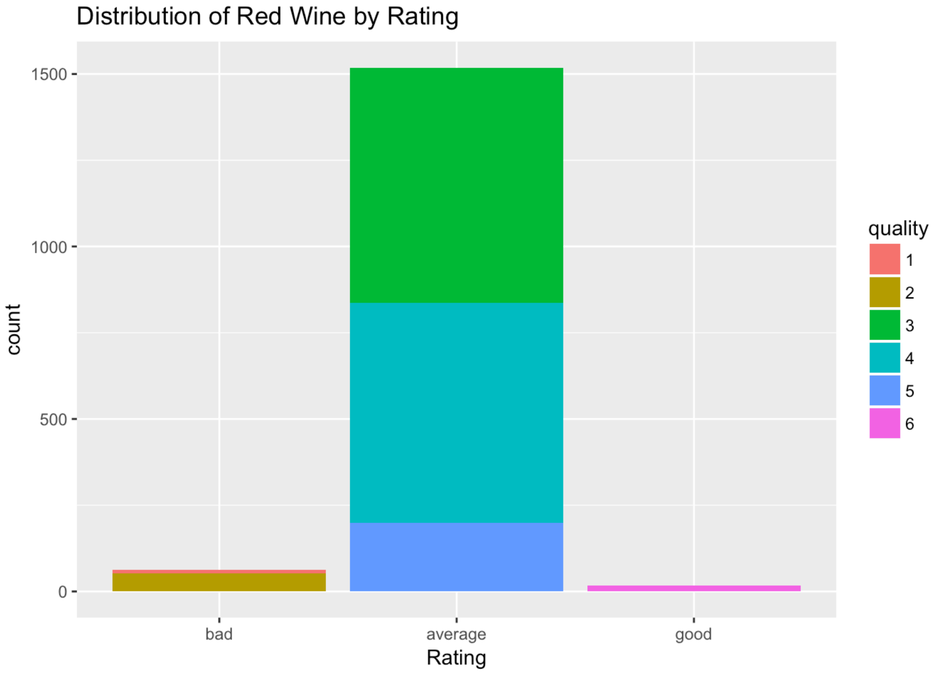
Only one thing can be comfirmed, that is good red wine have higer alcohol. Other three factors have limit influence on quality of red wine.

Were there any interesting or surprising interactions between features?

The cititric.acid for both good or bad red wind doesn't have clear differences.

Final Plots and Summary

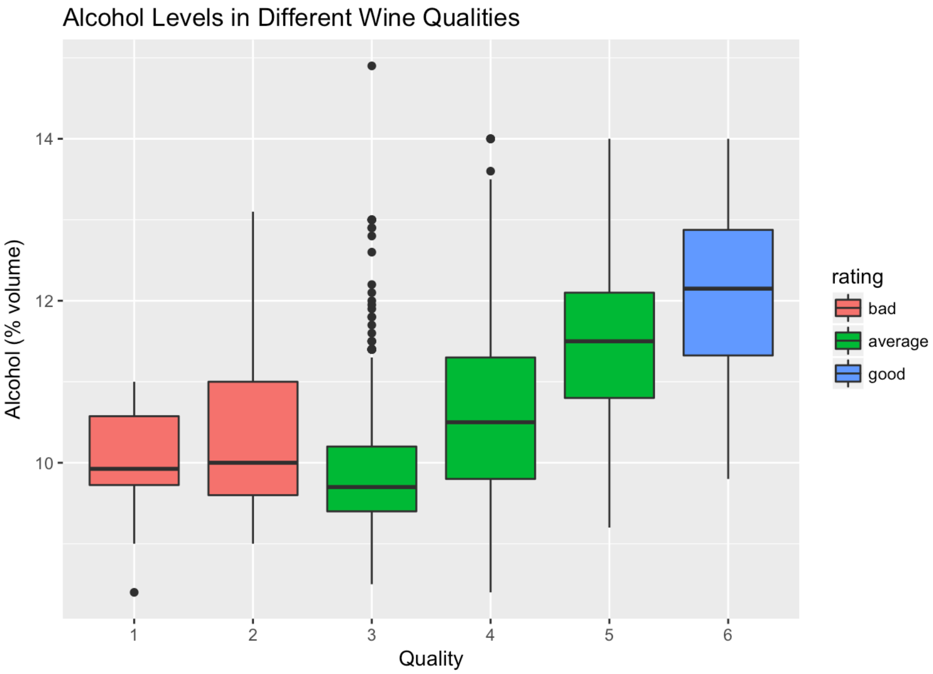
Plot One



Description One

This histogram clearly reveals that most red wind is on average rating, and most of these 'avearge' red wines have quality 5 or 6.

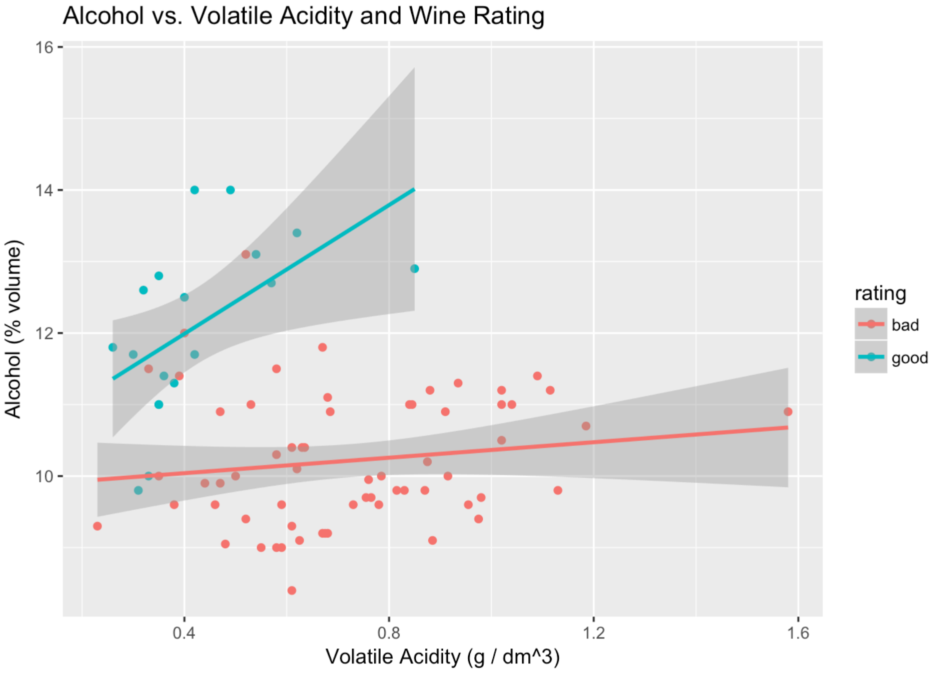
Plot Two



Description Two

These boxplots demonstrate the effect of alcohol on red wine quality, that is, higer quality red wine has higer alcohol, some outliers doesn't show this relationship.

Plot Three



Description Three

After remove the avearge rating wind data, this scatterplot shows the realationship between alcohol and the quality of red wine again. Additionaly, the realationship of vlotile.acidity and quality can be seen from this plot, higer quality red wine has lower volatile.acidity. And the good wine has more positive correlation between alcohol and volatile.acidity.

Reflection

The object of this exploratory data analysis is to find out which chemical properties would afftect the quality of red wines.In order to obverse the quality more directly, I divided the quality into new three rating: bad, average and good. I plotted and calculated the correlations between quality and the variables. However, after all these anaysis, I found none of the correlations were above 0.7. Aalcohol is the most important facor that influence the quality of red wines, the acidity also affect the quality to some extent. I hvae to say the measure of red wine quality is subjective, which means the data analysis is not enough to reveal all factors and to rate a res wine.

In this dataset, more than 80% red wines are ratedas 5 or 6, this limitation makes that there is not enough data to analyze good wine’s factors. In furter analysis, a dataset has more obeervations woll be preferred. However,