Exploratory Data Analysis of RedWineQuality

bangshen

2017/3/31

Before start: This report is my homework of udacity data analysis degree, introduction of this project presents here (https://classroom.udacity.com/nanodegrees/nd002/parts/0021345407/modules/316518875375461/lessons/31651887532398 The data set is about red wine downloaded from this link (https://s3.cn-north-1.amazonaws.com.cn/static-documents/nd002/DADataSetOptionsNanodegree_zh.pdf), containing the content of red wine and the quality of each wine assessed by at least 3 Sommeliers, my goal is to analysis which content of red wine will have significant impact on its quality though carrying out this project. Since it is my first time to exploring a dataset, I've viewed other exploratory data analysis based on other dataset. One of them is this report created by Chris Saden (https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html) on the basis of diamonds dataset, which is also recommended by udacity as an demonstration case for students finishing this project easily, the structure of this report imitate the case.

Libararies used in this report

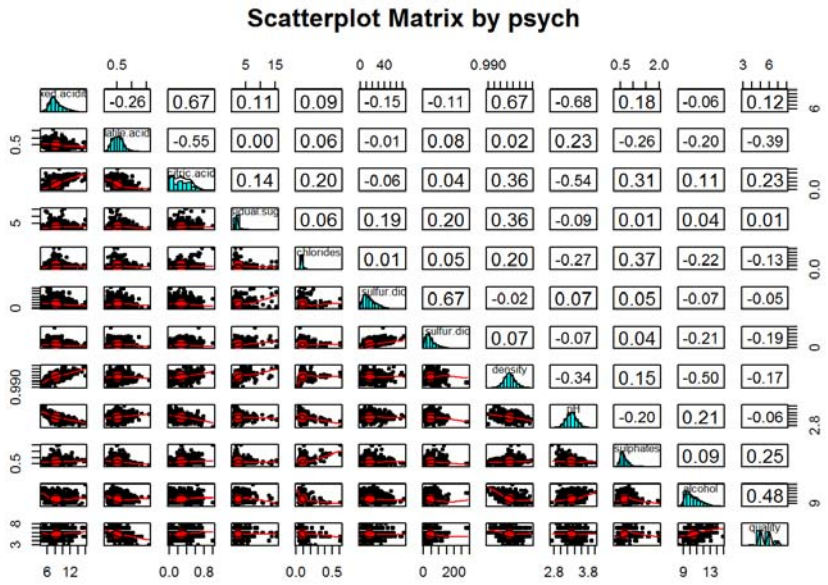
```
library('psych')
library('ggplot2')
library('reshape2')
library('gridExtra')
```

Dataset Overview

```
## '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 residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

There are 1599 observations and 13 variables in our dataset, the data types of variables are num and int. Among 13 variables, the first variable is x which is merely used as ID, the last variable is quality of redwine, which is a dependent variable depend on other 11 independent variables, that is variables from the 2nd column to the 12th column.(I have not been sure if those independent variables do really affect the quality of redwine so far, let's call them like that temporarily.)

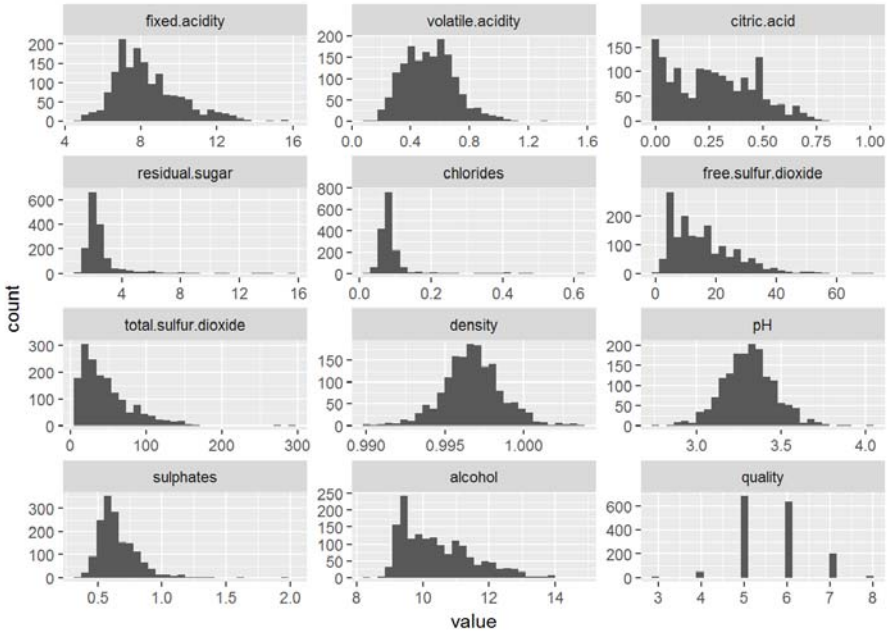


R and its packages is very powerful.This scatterplot matrix of 12 variables can be plotted in less than 1 min, that's really amazing. For the scater plots in down left part of this graph is fuzzy and crowded, hence I gonna analysis correlation coefficients in up right. Just have a glance at those correlation coefficients here, it seems no very strong relationships occur between those variables. some ones with coefficient more than 0.5 are:
+**citric.adic** vs **volatile acidity**(-0.55)
-**citric.adic** vs **fixed.acidity**(0.67)
-**total sulfur dioxide** vs **free sulfur dioxide**(0.67)
-**density** vs **fixed acidity**(0.67)
-**fixed acidity** vs **pH**(-0.68)
-**pH** vs **citric acid**(-0.54)
-**density** vs **alcohol**(-0.5)
And it is notable that no strong coorelationship between quality and other variables except alcohol(0.48) and volatile acidity(-0.39)

Univariate Plots Section

Distribution of all variables

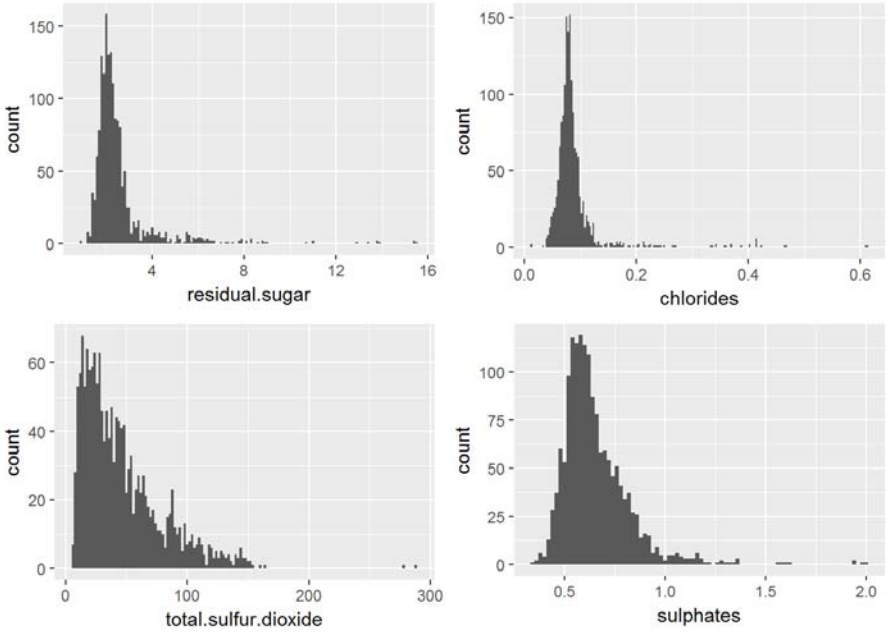
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.    : 4.60   Min.    :0.1200   Min.    :0.000   Min.    : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean    : 8.32   Mean    :0.5278   Mean    :0.271   Mean    : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.    :15.90   Max.    :1.5800   Max.    :1.000   Max.    :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
```

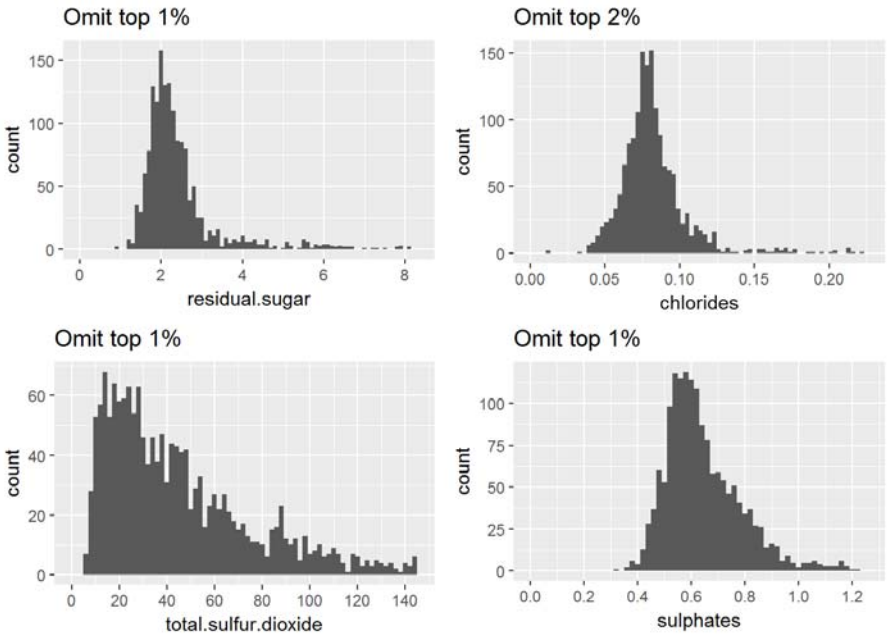


Findings:some are sort of like normal distribution, while others are right skewed distribution with outliers

The distribution of all variables are shown above, among those variables, quality data type is integer that's why its histogram is discrete. Distribution of fixed.acidity, volatile.acidity,density,pH seems kind of like normal distribution, however, distribution of residual.sugar,chlorides,free sulfur dioxide,sulphates,total.sulfur.dioxide and alcohol are with a long tail,some are may be result from outliers.
let's take a look at some variables with obvious outliers.

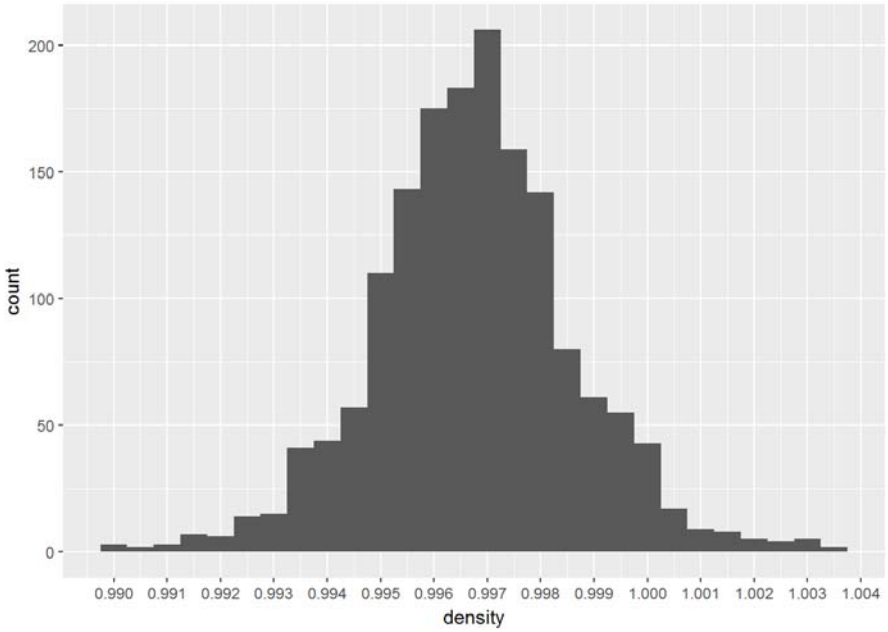
residual.sugar chlorides total.sulfur.dioxide sulphates

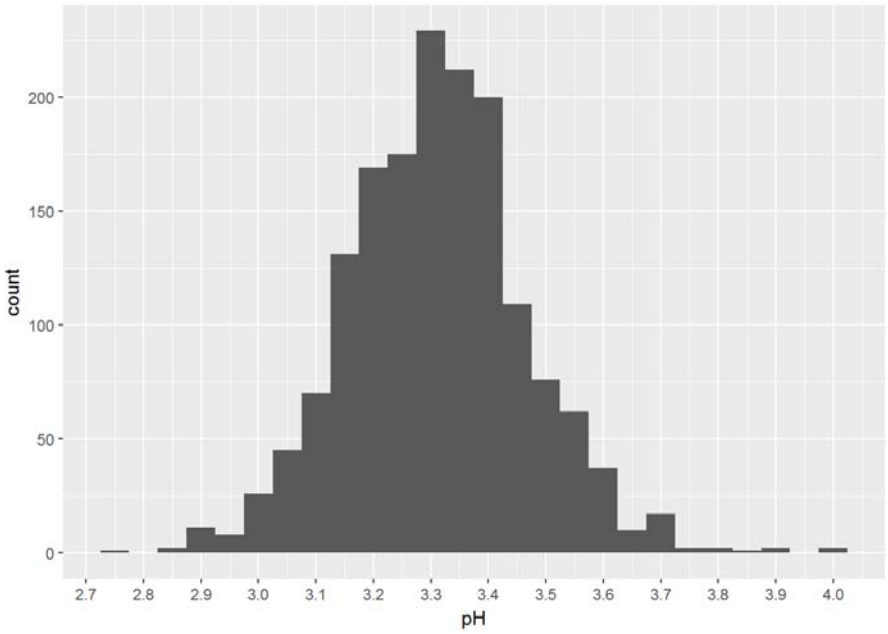




Comparison of two graph, the distribution of those variable can be more normal by omitting top 1% or 2% data:
* To residual.sugar: most are in the range of [1:3]
* For chlorides: most are in the range of [0.05:0.1]
* For sulphates: [0.4:0.8] seems to be the most frequent range
* For total.sulfur.dioxide: more discrete than others

density and pH






```
## [1] "summary(pH)"
```

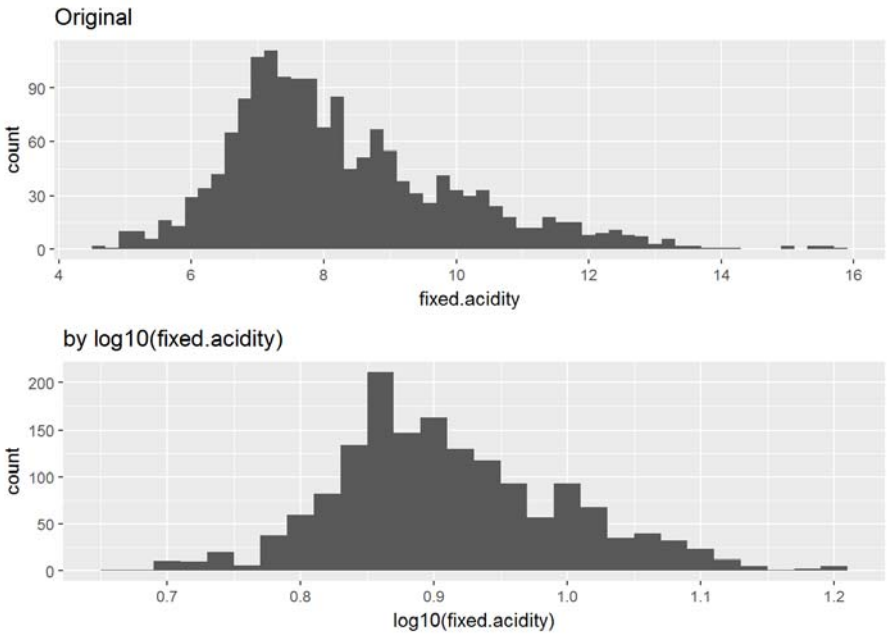
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

```
## [1] "summary(density)"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

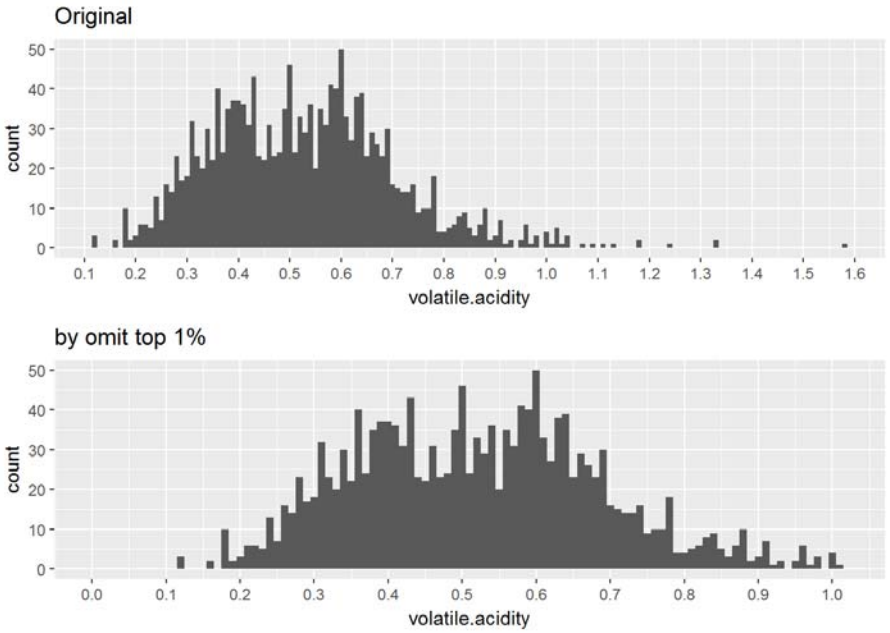
The distribution of density and pH is normal, the max range of density is 0.0139 g/dm^3. No more than, it is reasonable since they all are reawine with predominant content of water.Most wines are located in the range of 3.1-3.5.

fixed.acidity,volatile.acidity,citric.acid

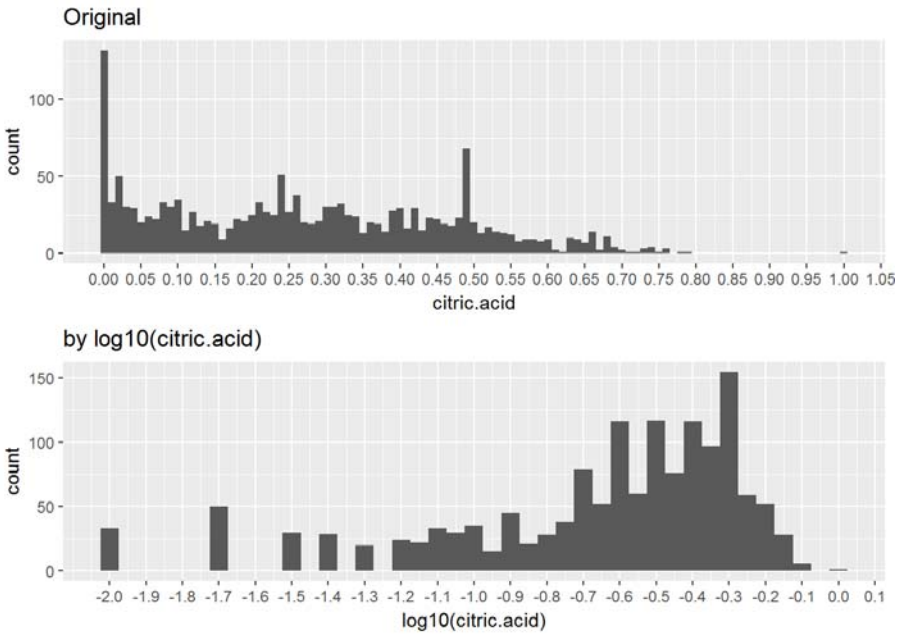


Distribution of fixed.acidity is kind of skewed, but can be fixed by log10()

```
## Warning: Removed 15 rows containing non-finite values (stat_bin).
```

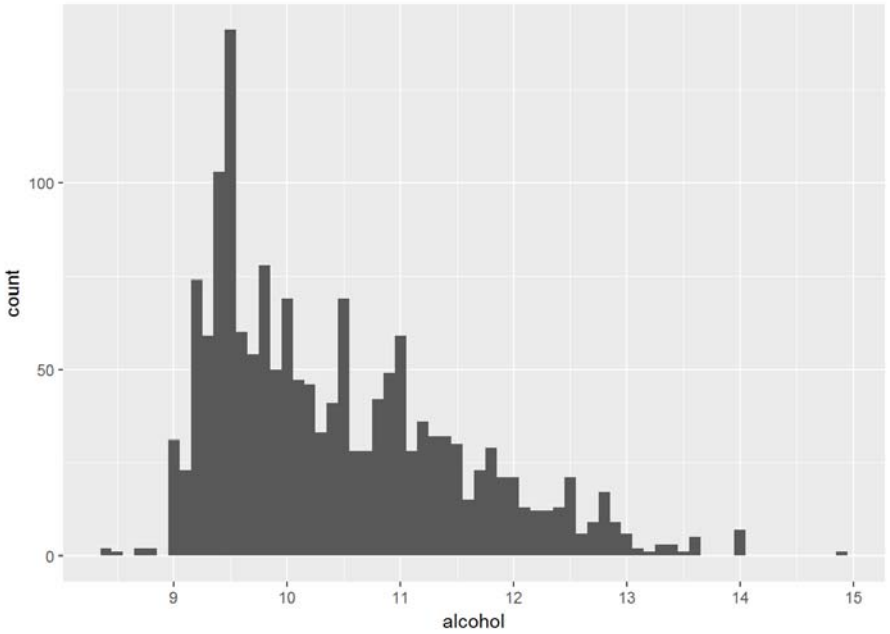


There are some obvious peaks here in the histogram of volatile.acidity, the range can be efficiently fixed by omitting top 1% data,



For citric acid, its distribution are relatively flat, and two main peaks are shown at 0 and 0.5 g/dm^3, I am more interested in two peaks, I also tried to transfer its x-axis,but it is no use to transfer this data by log10().

alcohol

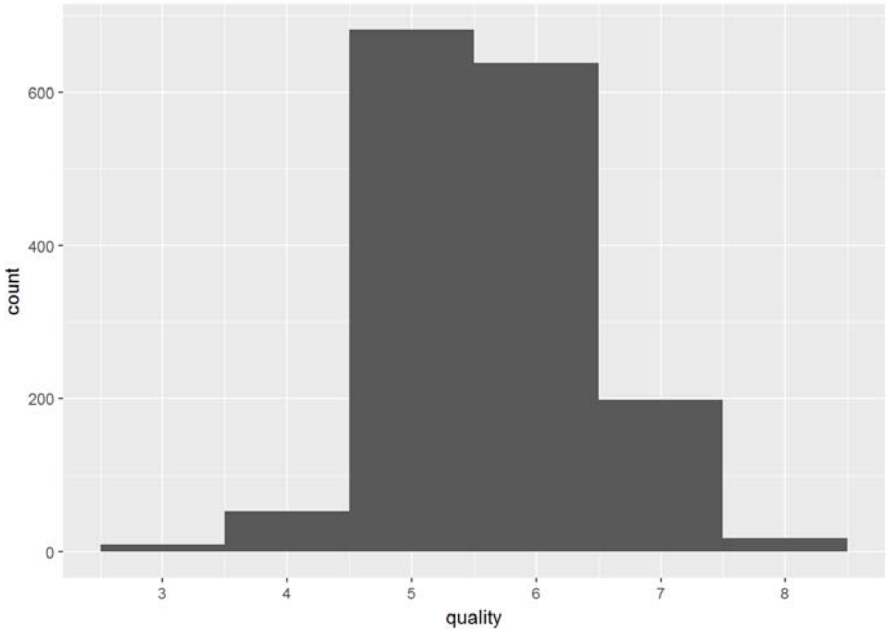



```
## [1] "summary(alcohol)"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

The alcohol of redwines are most in the range of 9:11, however, there still have some wines are with very high alcohol content up to around 15%

quality




```
## [1] "summary(quality)"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

```
## [1] "number of each quality"
```

```
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

Maximun of quality is 8 and the minmum is 3, something interested is that most of redwines are been graded 5 or 6, which means Moderate level. redwines with quality of 3 or 4 and 7 or 8 can be grouped as low quality and good quality, respectively. In the latter part, I will investigate the features of three groups.

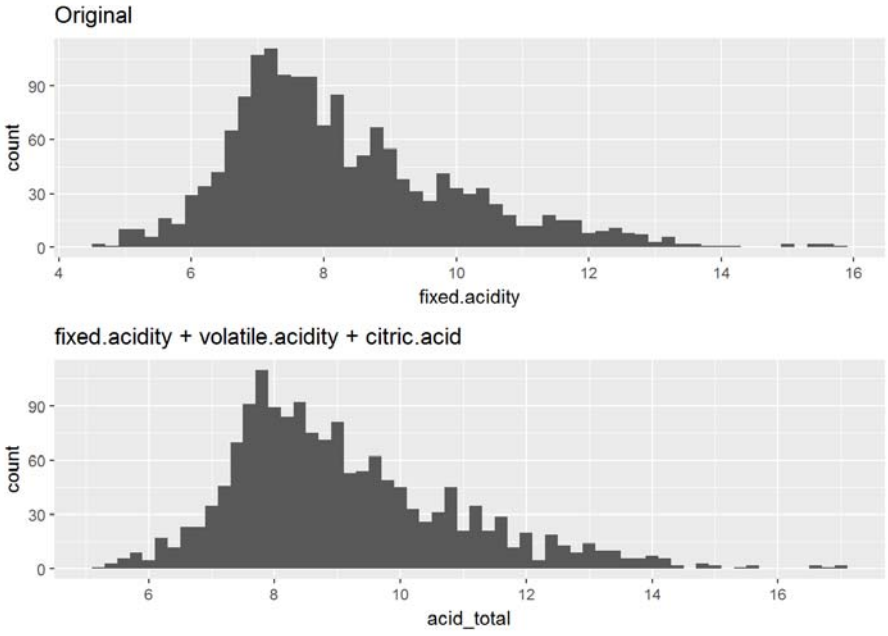
New variables

new variables: acid_total

Because there are three type of acid in this data set, I wanna add them together and labled as `acid_total` , code is

```
wine <- transform(wine, acid_total = fixed acidity + volatile acidity + citric acid)
```

The distribution of acid_total is shown below:

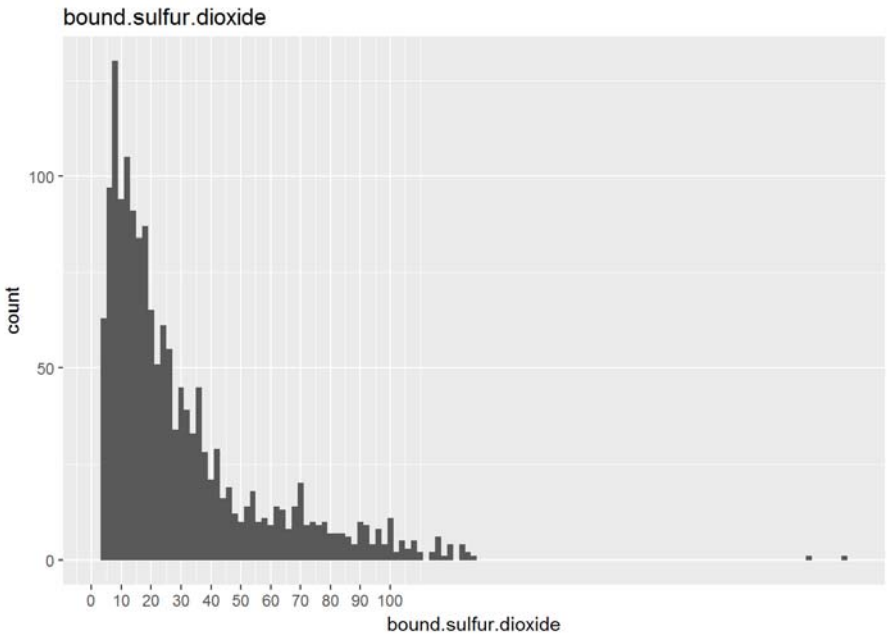


Since the acidity of volatile acid and citric acid is much lower than fixed acidity, hence this combination of three acids seems no significant changes compared with fixed.acidity

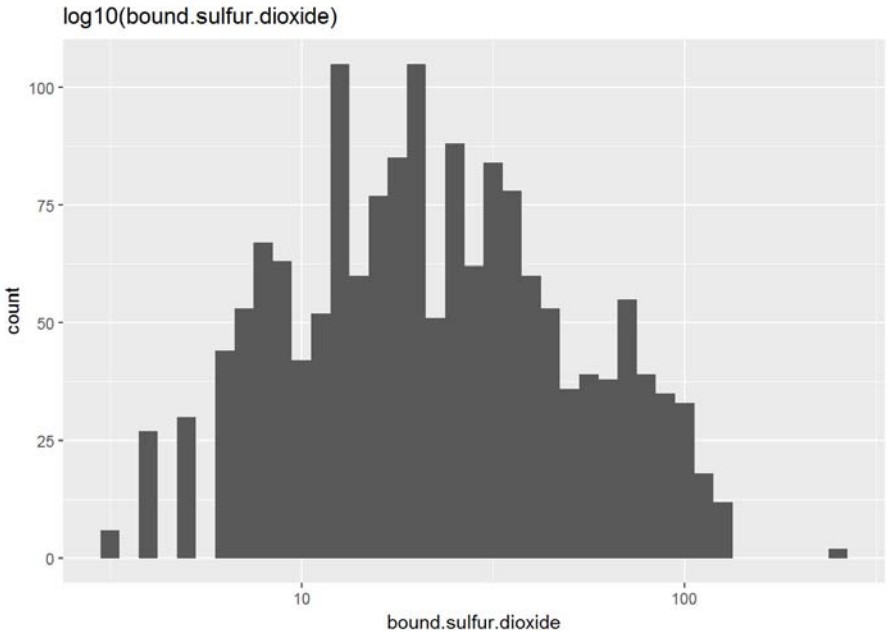
new variables:bound.sulfur.dioxide

A total sulfur dioxide is the amount of free and bound forms of S02, we have free sulfur dioxide data here in our data, I just want to know if bound form of so2 will affect the redwine quality or not, code for calculating this is:

```
wine <- transform(wine, bound.sulfur.dioxide = total.sulfur.dioxide - free.sulfur.dioxide)
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   12.00   21.00   30.59   39.00  251.50
```



new variable: quality_factor

codes are

```
wine <- transform(wine, quality_factor = factor(wine$quality))
```

new variable:quality.bucket

I created a variable named `quality.bucket` which will divide quality into three groups with three lables. For quanlity of 3 and 4 will be labeled as *Bad*, 5 and 6 will be labeled as *Moderate*, 7 and 8 will be labeled as 'Good', will will really reflect the quality of wine, code is

```
wine$quality.bucket <- cut(wine$quality, c(2, 4, 6, 8), labels = c('Bad','Moderate','Good'))
```

new variable:alcohol.bucket

The factor type of alcohol is created by code:

```
wine$alcohol.bucket <- cut(wine$alcohol, c(8, 10, 12, 15), labels = c('low_alcohol','Middle_alcohol','high_alcohol'))
```

All new variables are shown below:

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality acid_total bound.sulfur.dioxide quality_factor quality.bucket
## 1       5       8.10                   23              5       Moderate
## 2       5       8.68                   42              5       Moderate
## 3       5       8.60                   39              5       Moderate
## 4       6      12.04                   43              6       Moderate
## 5       5       8.10                   23              5       Moderate
## 6       5       8.06                   27              5       Moderate
##   alcohol.bucket
## 1    low_alcohol
## 2    low_alcohol
## 3    low_alcohol
## 4    low_alcohol
## 5    low_alcohol
## 6    low_alcohol
```

Univariate Analysis

What is the structure of your dataset?

there are 1599 observtions and 13 variables in this data set, the first variable in the first column is ID, other 12 variables are numbers with the type of `int` and `num`. units of each variables are:

1 - fixed acidity (tartaric acid - g / dm^3)
2 - volatile acidity (acetic acid - g / dm^3)
3 - citric acid (g / dm^3)
4 - residual sugar (g / dm^3)
5 - chlorides (sodium chloride - g / dm^3
6 - free sulfur dioxide (mg / dm^3)
7 - total sulfur dioxide (mg / dm^3)
8 - density (g / cm^3)
9 - pH
10 - sulphates (potassium sulphate - g / dm3)
11 - alcohol (% by volume)
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Other observations:
1. citric acid and volatile acidity are much low than fixed acidity
2. quality are mostly evaluated as 5 or 6
3. citric acid distribution is flat, pH and density distribution are more normal, and other variables are distrubuted with a long tail

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

The main features are quality,ph,density, I'd also like to determine will factor will affect quanlity obviously

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

acidity(fixed acidity, volatile acidity,citric acid),pH,density,and alcohol would have a obvious impact on the quality of redwine.

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

I created four new variables here. one is the `acid_total` eaquals to sum of three acids(fixed acidity, volatile acidity,citric acid), but found its distribution is just similar with fixed acidity. Another is `bound.sulfur.dioxide`, meaning bound form of SO2 in the wine. Third is a `quality_factor` which is transformed from quality by factor function.Forth is `quality.bucket` which will grade evaluate the quality of redwine by *Bad*,*Moderate* and *Good*. Last is `alcohol.bucket`, by this alcohol of 8 to 10 will be grouped into low_alcohol, 10 to 12 is middle_alcohol, and alcohol more than 12 will be reckoned as high_alcohol.

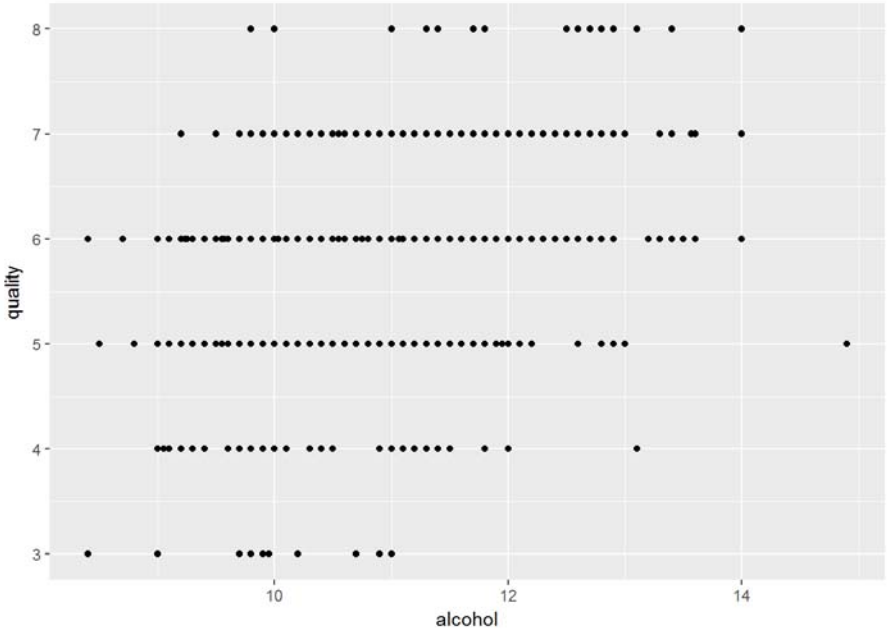
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

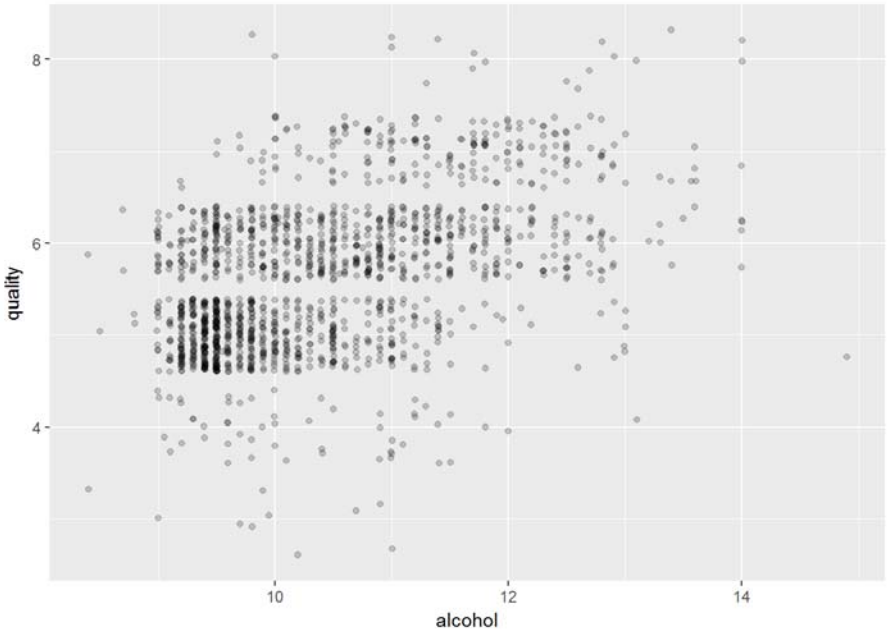
This data is so tidy that I do not need extra cleaning, some variables are skewed distributed, and can be transformed by scaling x axis.

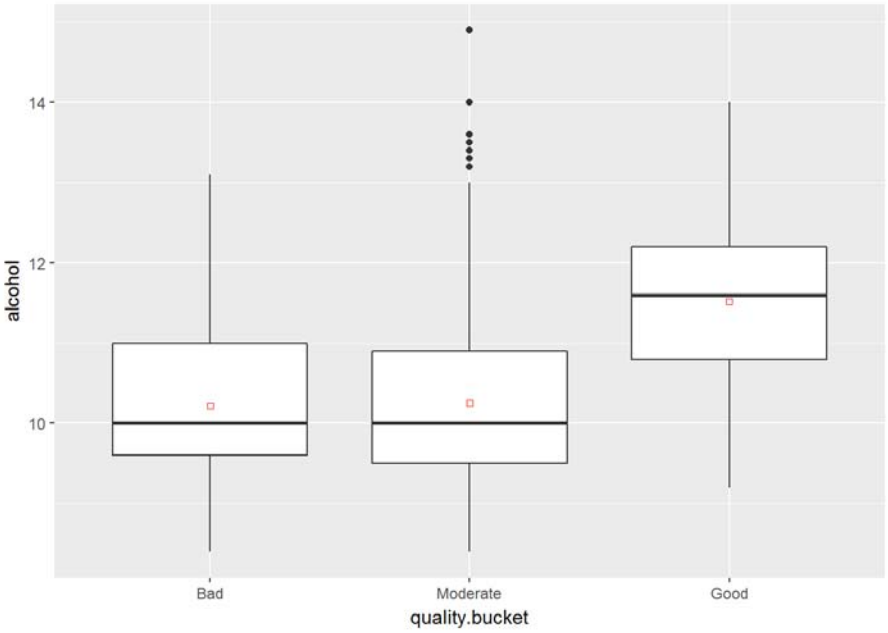
Bivariate Plots Section

main interested variable:quality

From the scatter plot mentioned before, it is found that alcohol have positive correlated with quality. And here I wanna to investigate relations between quality and other variables like alcohol,pH,density.

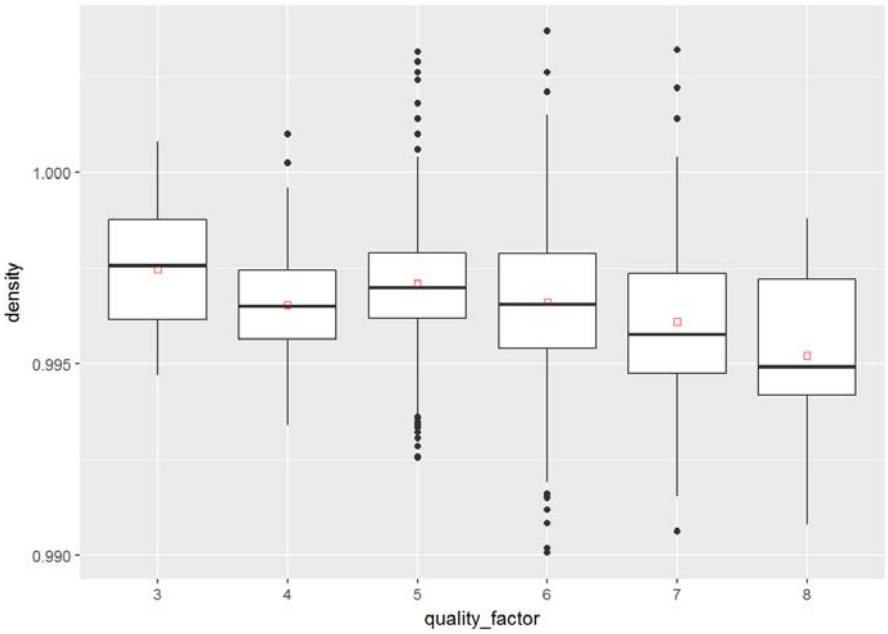


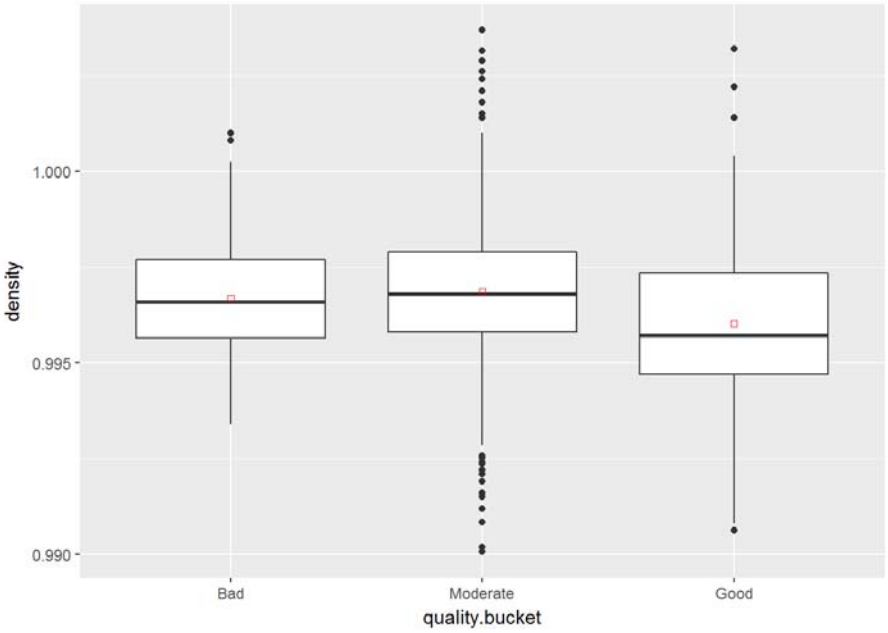





```
## wine$quality.bucket: Bad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.60   10.00   10.22   11.00   13.10
## ------------------------------------------------------------
## wine$quality.bucket: Moderate
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.00   10.25   10.90   14.90
## ------------------------------------------------------------
## wine$quality.bucket: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.60   11.52   12.20   14.00
```

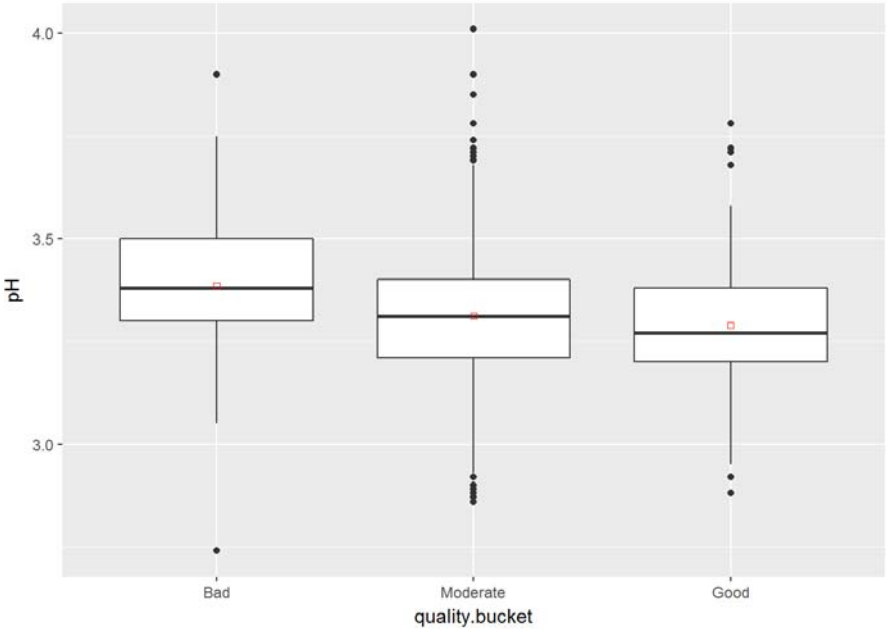
Quality values are integers, but by adding jitter and transparency, it looks better. It seems that mean of alcohol for each quality level is increased with the increasing of quality(from 10.22 to 11.52) indicating the positive correlationship betweent quanlity and alcohol,although this correlationship is weak.
Bad and Moderate are similar according to the boxplot, nevertheless, for good quality redwines have high alcohol, that's really interesting.






```
## wine$quality.bucket: Bad
## [1] 0.9966887
## ------------------------------------------------------------
## wine$quality.bucket: Moderate
## [1] 0.9968673
## ------------------------------------------------------------
## wine$quality.bucket: Good
## [1] 0.9960303
```

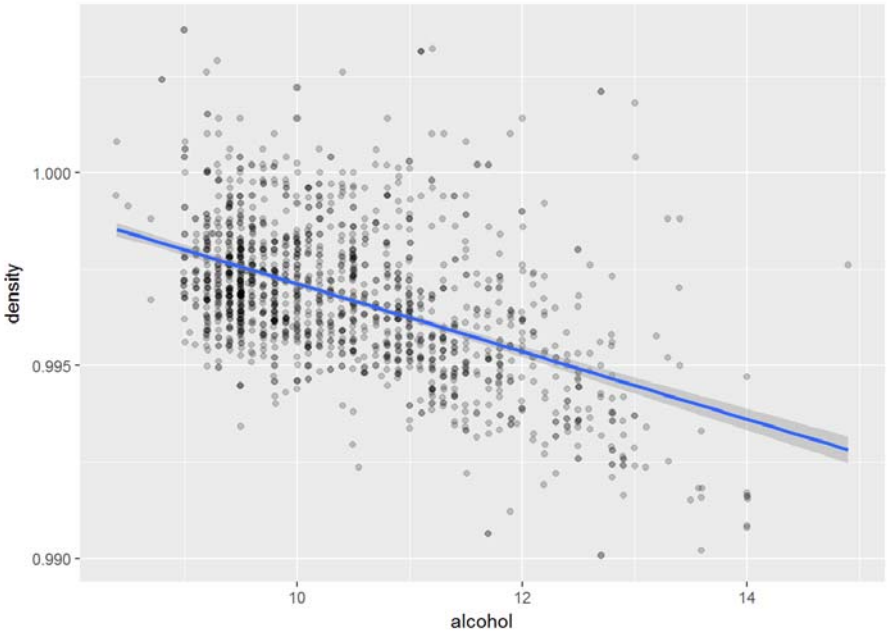
No obvious trend has been found, but one thing notable is the good redwine always with lower density than bad and Moderate.This can be also demostrated by analysising the relationship of density and alcohol, since high alcohol wine always be assessed as better quality and hence the density will decrease.




```
## wine$quality.bucket: Bad
## [1] 3.384127
## ------------------------------------------------------------
## wine$quality.bucket: Moderate
## [1] 3.311296
## ------------------------------------------------------------
## wine$quality.bucket: Good
## [1] 3.288802
```

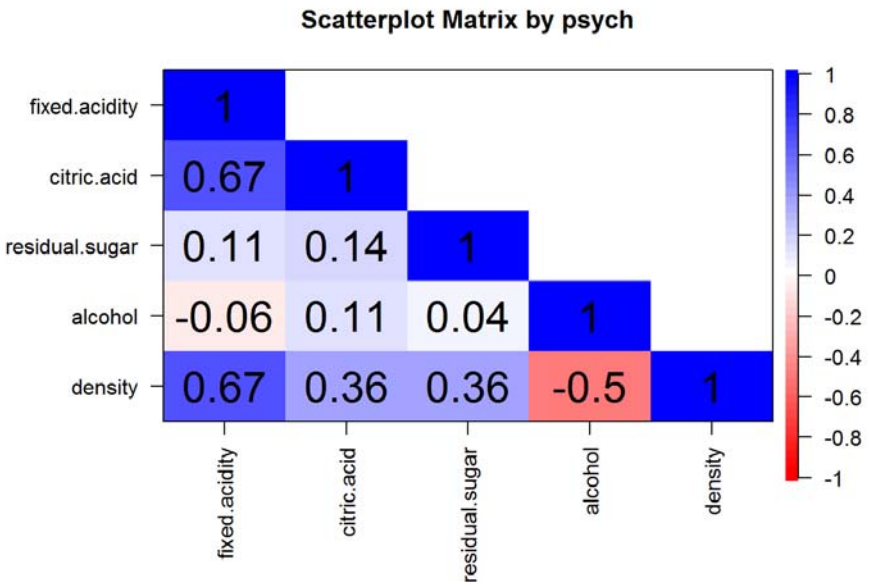
Good quality redwines seems with lower pH values.

Minor interested variable:Density

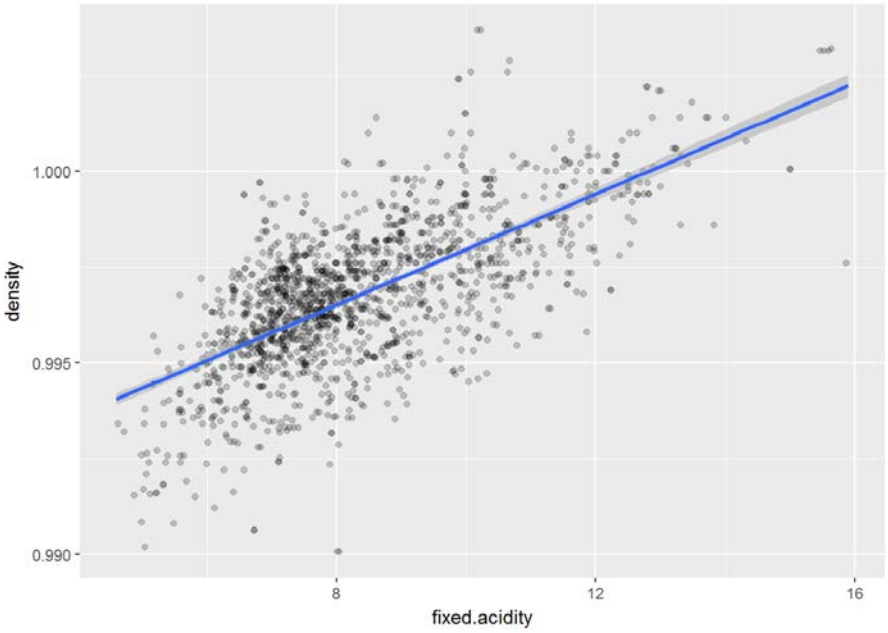



```
##
##  Pearson's product-moment correlation
##
## data:  wine$alcohol and wine$density
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798
```

Density decreases with the increasing of alcohol which makes sense because alcohol density is lower than water. Apart from alcohol, we note that fixed.acidity, citric.acid,residual.sugar also have slight correlationship with density from the scatterplot matrix.

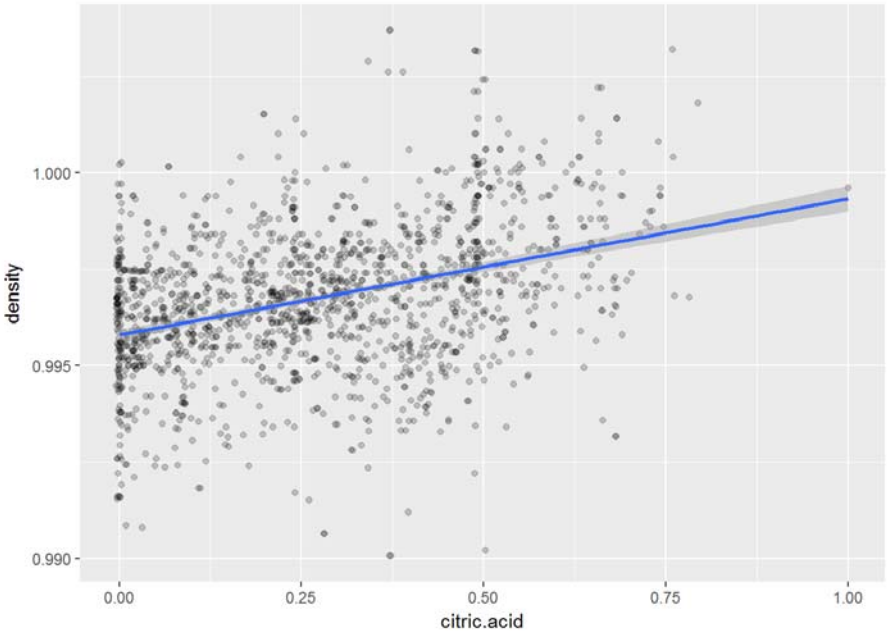


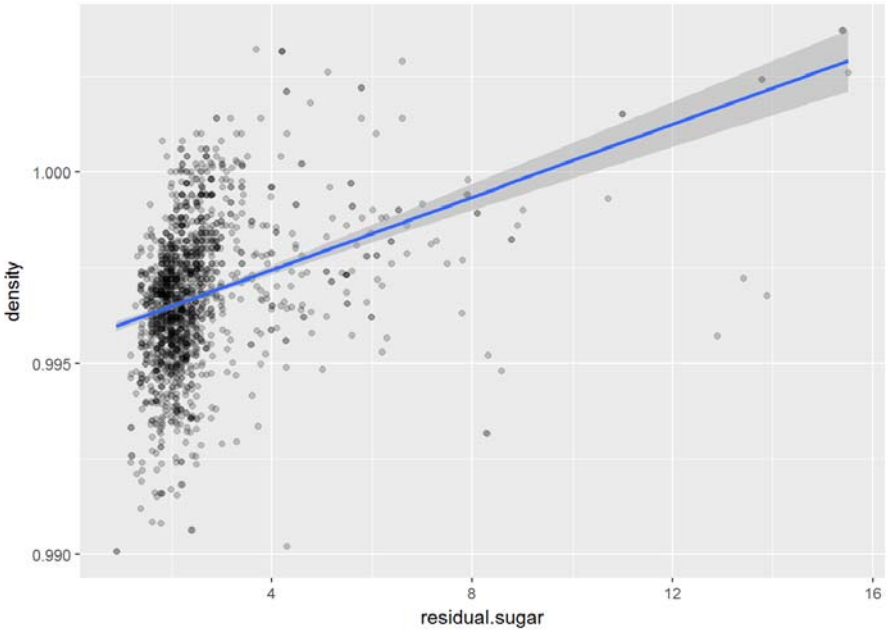
No surprise here, acid and sugar should have effect on density, especially fixed.acidity, the corrletionship coefficient is 0.67. let's look at the scatter plot of fixed.acidity and density




```
##
##  Pearson's product-moment correlation
##
## data:  wine$fixed.acidity and wine$density
## t = 35.877, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6399847 0.6943302
## sample estimates:
##       cor
## 0.6680473
```

Both Alcohol and fixed.acidity would influence the density, which is because of the alcohol and fixed.acidity are the predominant content in redwine except water. alcohol is negative correlated whereas fixed.acidity is positive correlated with density





citric.acid and residual.sugar have lower coorelation coefficient compared with alcohol and fixed.acid, that may be the result from low concentration of them in redwine.

new variables and density or quality

```
##                       dnsty qulty acd_t bnd..
## density                1.00
## quality               -0.17  1.00
## acid_total             0.68  0.10  1.00
## bound.sulfur.dioxide  0.10 -0.21 -0.06  1.00
```

New variables's correlational structure is shown above.No strong correlation is observed. It means new variables are meaningless.

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 is correlated with alcohol and pH and density

1. **quality and alcohol** : For moderate and bad quality redwines, alcohol content have littel difference, but alcohol of good quanlity redwine is higher than bad and moderate wine.
2. **quality and density**: the trend is similar with the last, for bad and moderate, there is no obvious difference observed, but when we pay attention on the good wines,its density is allways lower than other quality level. This makes sense as the negative correlationship between alcohol and density.
3. **quality and pH**: The total trend is quality rises with pH values drop, since pH is related to the acid content in wine, hence, basically, the correlationship of quality and pH indicates the correlationship of quality and acid content in redwine.

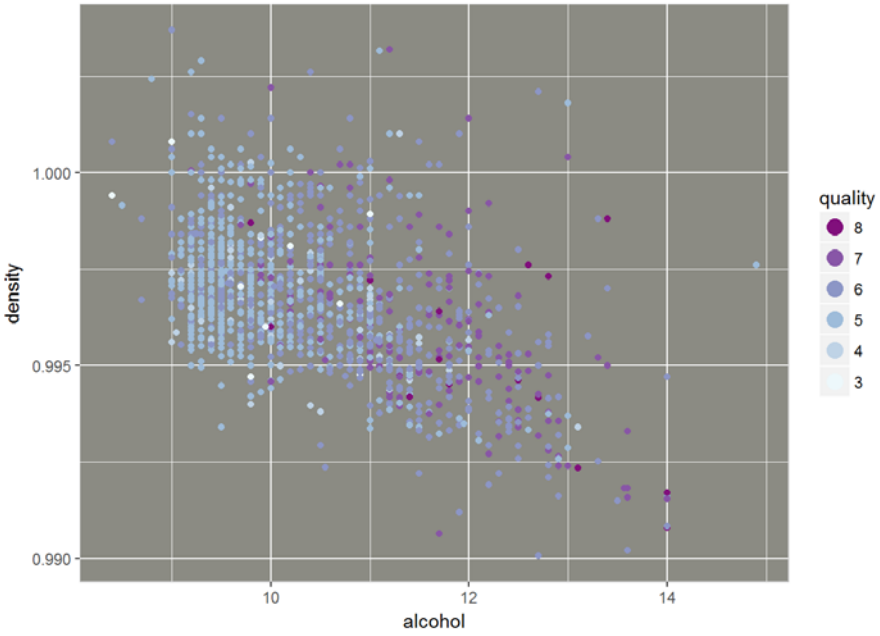
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

I do have notice some other relationships.
As we known,density is determined by the ingredient.In redwines, the predominant content is water, then fixed.acidity and residual.sugar.Density has strong correlationship with alcohol and fixed.acidity, and little correlationship with citricacid and residual.sugar.

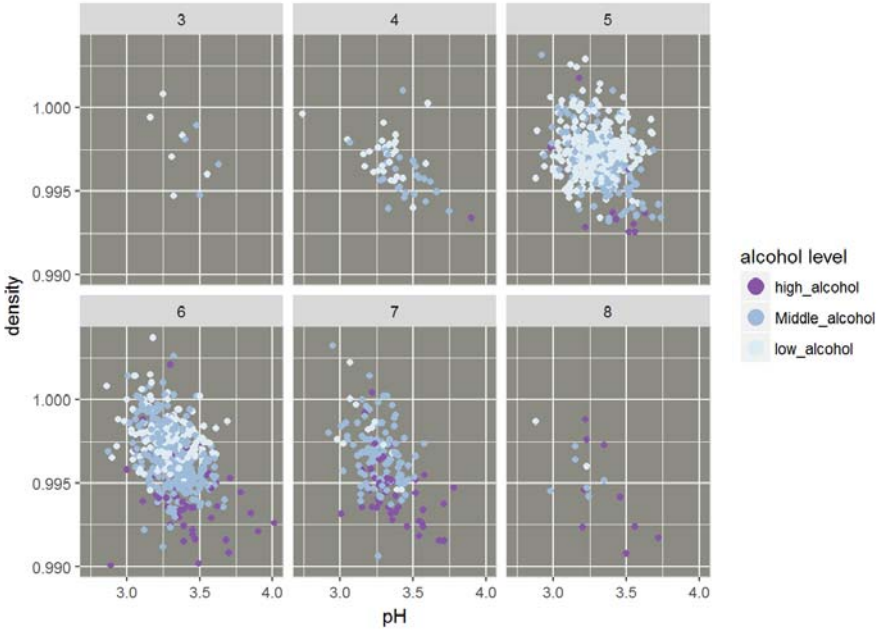
What was the strongest relationship you found?

I did not find very strong relationship(correlation coefficient more than 0.9) in the dataset. The strongest relationship is 0.67 of density and fixed.acidity and 0.67 of citric.acid and fixed.acidity.

Multivariate Plots Section



The number of high quality level(more than 6) is small, but trend is obvious, most high quality level wines are in the right of this plot, indicating high alcohol.



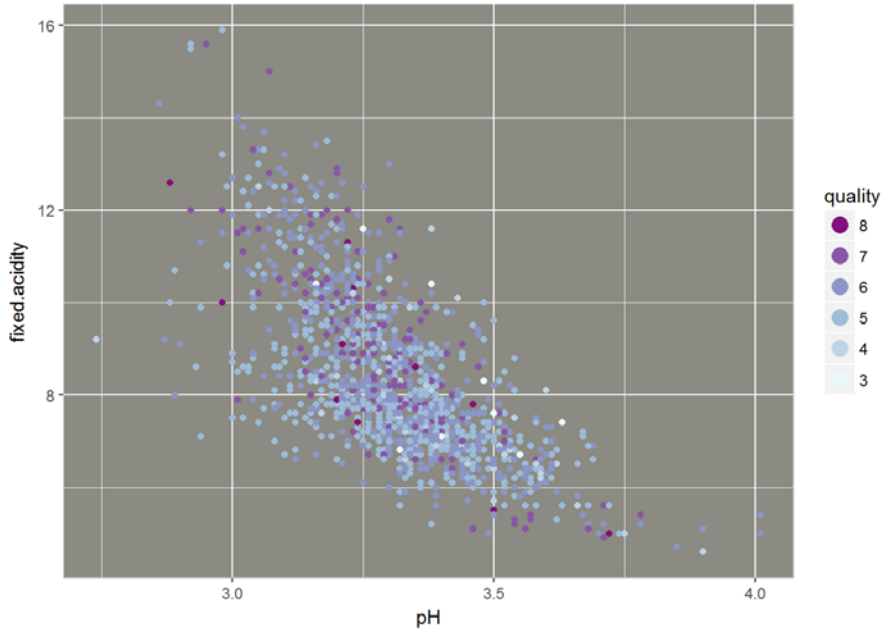

```
## [1] "number of wine.alcohol >=12"
```

```
##
## FALSE  TRUE
##  1437   162
```

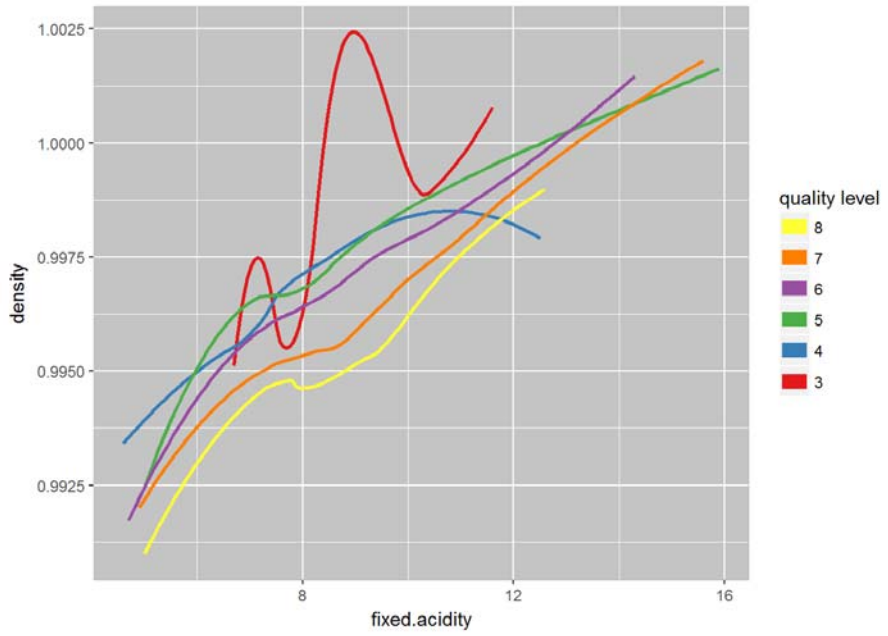
```
## [1] "number of wine.alcohol >=12 & wine.quality >=6"
```

```
##
## FALSE  TRUE
##  1451   148
```

Most of points are located in the quality equals to 5,6,7 indicating that those wine are normal, and this also increase the of diffculty of data analysis. Look at quality equals to 3 and 4, bad wines are always low alcohol, whileas good wines are reverse.There are 162 observations with alcohol more than and equal to 12, in where 148 wines are evaluated more than and equal to 6. The proportion is up to 91%




```
## `geom_smooth()` using method = 'loess'
```



Again, the conclusion is same with the analysis of bivariate analysis, high concentration of fixed.acidity always with high density

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?

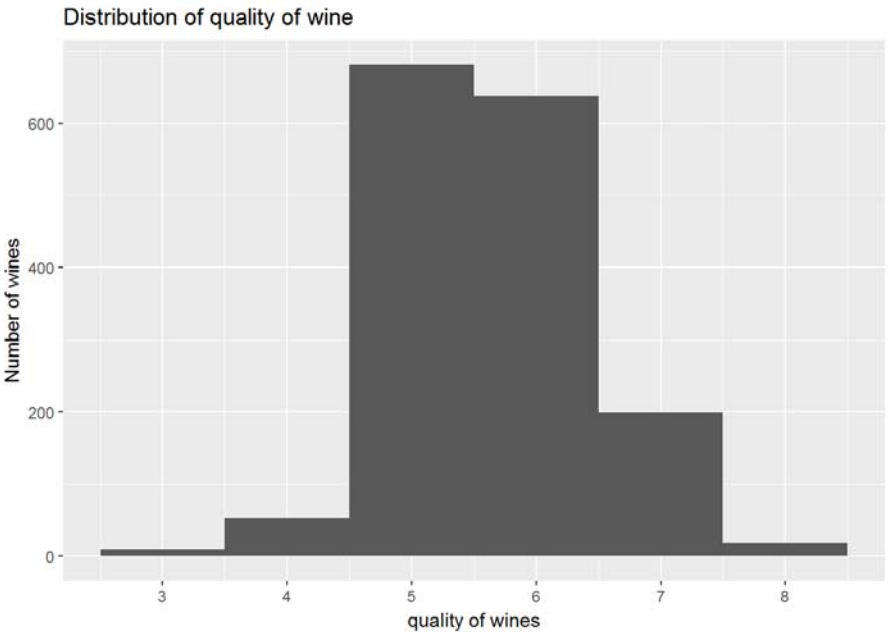
wines with high alcohol and low density always be assessed excellent.
High concentration of fixed.acidity always with high density

Were there any interesting or surprising interactions between features?

91% high alcohol level(alcohol >= 12) wines are evaluated as good wine(quality >=6), this is so interesting.

Final Plots and Summary

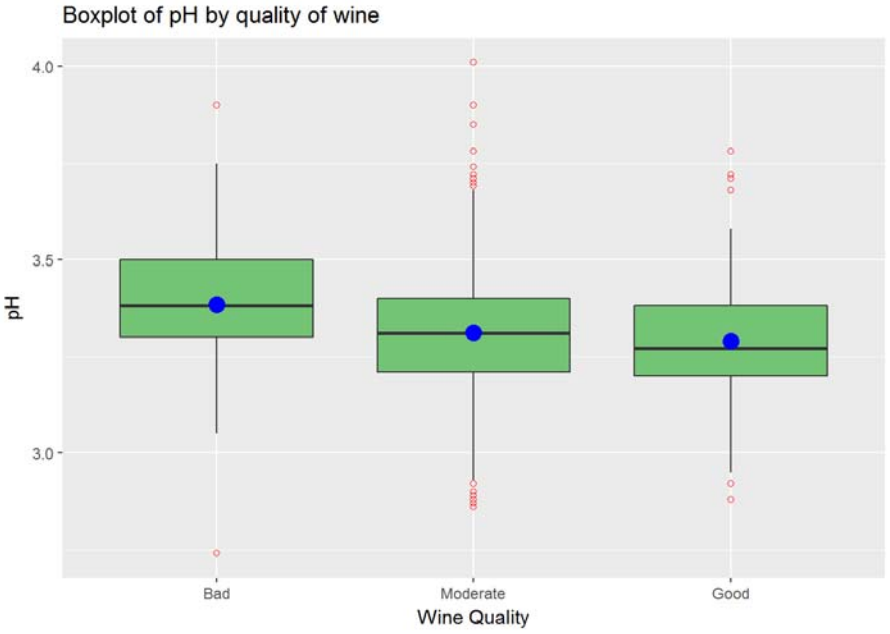
Plot One



Description One

The distribution of wine quality in this data set seems normal, and most observation are evaluated as 5 or 6. Range of quality is [3,8], there are very little wine being assessed as 3 or 8.

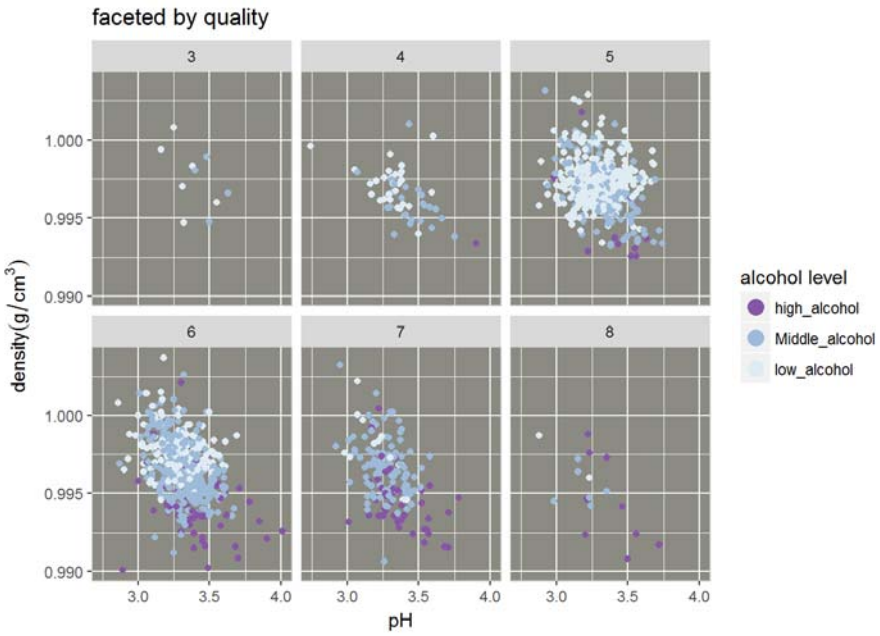
Plot Two



Description Two

When the quality has been grouped into three groups, it is found that good quality redwines seems always with lower pH values.(Blue points mean the mean of pH)

Plot Three



Description Three

Most of points are located in the quality equals to 5,6,7 indicating that those wine are normal. Look at quality equals to 3 and 4(bad wines), bad wines always are low alcohol, whileas good wines are reverse. 91% high alcohol level(alcohol >= 12) wines are evaluated as good wine(quality >=6)

Reflection

This dataset contains information on 1599 kinds of wines with 12 variables, at the begainning I tried to understand the strcture and some backgroud details of this dataset. Then I started to analysize univatiate variables, taking more attention on their distribution.Then investigated bivariate variables, focused on the relationship of variables, and last was the study of multivaribles.Most interested variables is quality, and tried to understand which variables will affect the quality of wine.

Most of variables are normal distribution, and some have outliners, distribution of quality is normal which makes sense. quality is related with alcohol, bad wines(quality = 3 or 4) are low alcohol, however, **91% high alcohol level(alcohol >= 12) wines are evaluated as good wine(quality >=6).** Density,pH also show the correlationship with quality. There is no very strong correlationship is obversed, this is because the quality of wine is determined by many complex factors or some key factors not been included in this data set.The trouble that I runed into during this data analysis process is that quality do not have strong correlatonship with other variables, but I still try to investigate some and try to find some interesting conclusions.I created some new variables, but found that it is useless and meaningless.The combination of variables in the dataset is meaningless. And it is why I did not do further study on multivariate analysis part.

Limitation of this report is the limitation of data size. There are only 1599 record in this dataset, we do not know if there are any factors will interfere variables.Another limitation is I did not give a model to assess quality of wine, since I do not find very strong correlationship between quality and variables.

Future work: one is to collect more wines records, both observations and variales, enormous observations will concluded more precise conclusions, and other variables like grape type and origin, and water type will also have impact on the quality of redwine.