Project 4 : Explore and Summarize Data

Xiaomin Xu

Oct 15th, 2015 ==========

1. Introduction

The datasets I used to do this project is *White Wine Quality*, which contains 4,898 white wines with 11 variables on quantifying the chemical properties of each wine. At least 3 wine experts rated the quality of each wine, providing a rating between 0 (very bad) and 10 (very excellent).

This project amis to investigate:

- Which chemical properties influence the quality of wines?

2. Univariate Plots Section

2.1 Dataset Structure Variables descriptions provided by the course web are shown as below:

- input variables:

(1) fixed acidity(tartaric acid - g / dmˆ3): most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

(2) volatile acidity(acetic acid - g / dmˆ3): the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

(3) citric acid(g / dmˆ3): found in small quantities, citric acid can add 'freshness' and flavor to wines

(4) residual sugar(g / dmˆ3): the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

(5) chlorides(sodium chloride - g / dmˆ3): the amount of salt in the wine

(6) free sulfur dioxide (mg / dmˆ3): the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

(7) total sulfur dioxide(mg / dmˆ3): amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

(8) density(g / cmˆ3): the density of water is close to that of water depending on the percent alcohol and sugar content

(9) 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

(10) sulphates(potassium sulphate - g / dm3): a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

(11) alcohol (% by volume): the percent alcohol content of the wine

- Output variable (based on sensory data):

(12) quality (score between 0 and 10)

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

We can see that the datasets loaded from the file contain the row index. The next step is to do some simple data cleaning including removing the row index, categorical variable transformation and standarize variables (if needed).

2.2 Factorize The Outcome Variable "quality" The type of the variable *quality* is now numeric. As we know from the description of the data set, the quality of the wine is shown as a list of score, but the difference between each score, for example, difference between score 3 and 4, could not be quantitativly interpreted as the score refers to a certian rank instead of a specific number. Hence it would be more reasonable to transform the outcome variable into a ordinal variable for future analysis

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ 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             : Factor w/ 7 levels "3","4","5","6",..: 4 4 4 4 4 4 4 4 4 4 ...
```

2.3 Check extreme values in predictors The simplest way to check extrem values is to compare the mean and median of the variable. If thes two differ from each a lot, then it is very possible that the variable carries outliers. Also, the range of the variable should be checked at the same time.

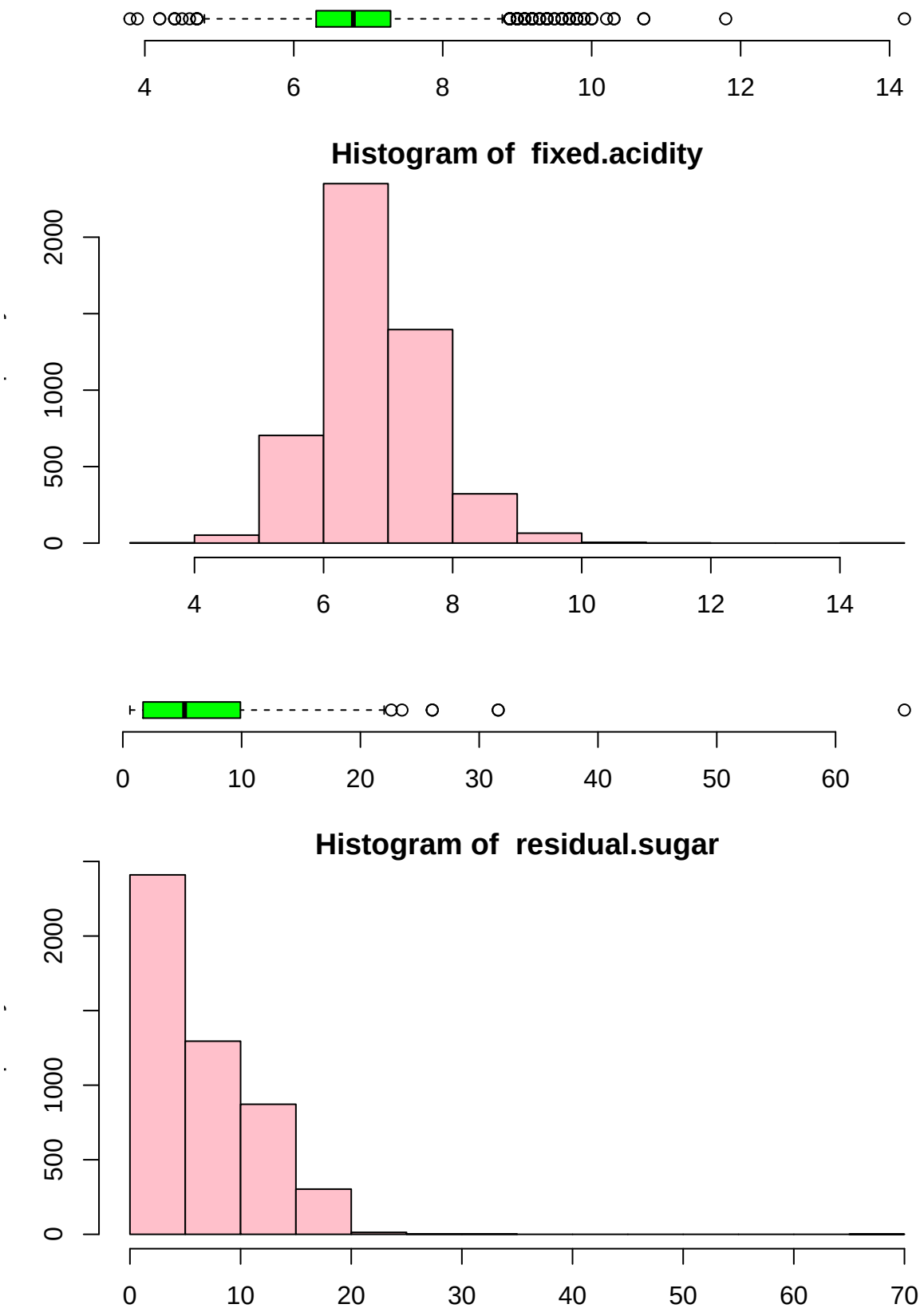
```
##        fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## Median         6.800           0.2600      0.3200          5.200   0.04300
## Mean           6.855           0.2782      0.3342          6.391   0.04577
##        free.sulfur.dioxide total.sulfur.dioxide density    pH sulphates
## Median               34.00                134.0  0.9937 3.180    0.4700
## Mean                 35.31                138.4  0.9940 3.188    0.4898
##        alcohol
```

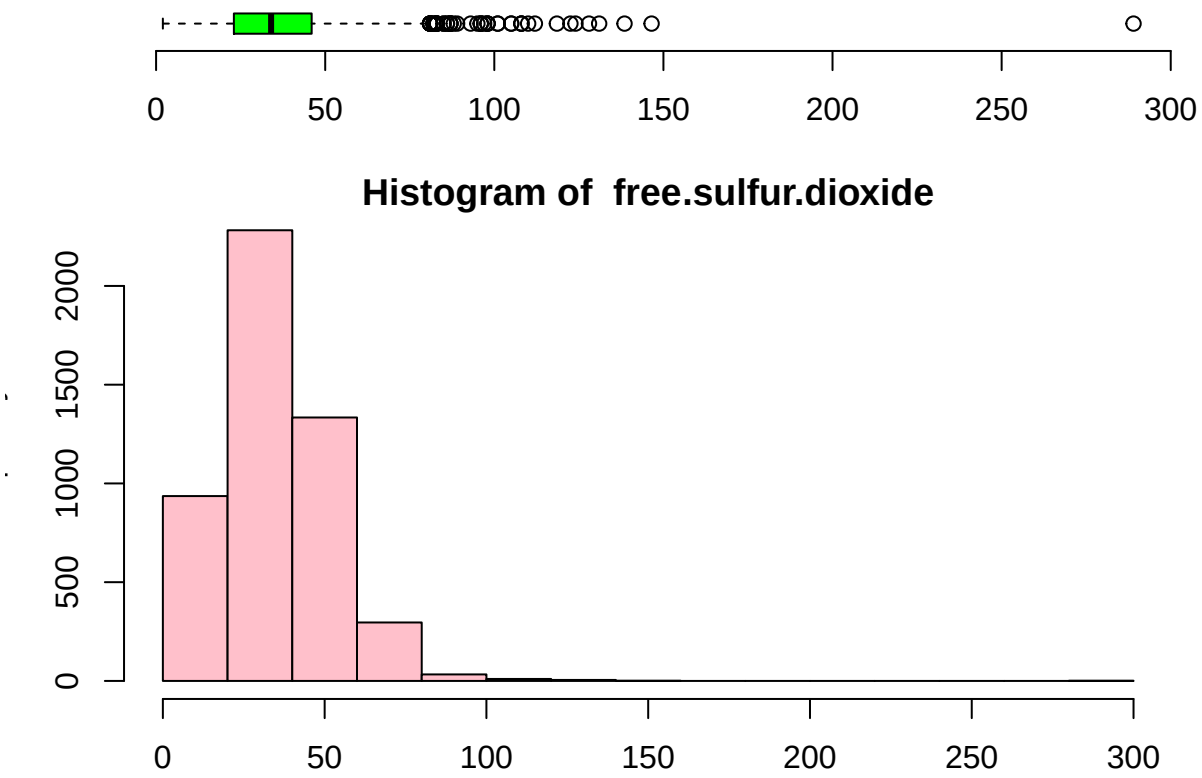
```
## Median   10.40
## Mean     10.51
```

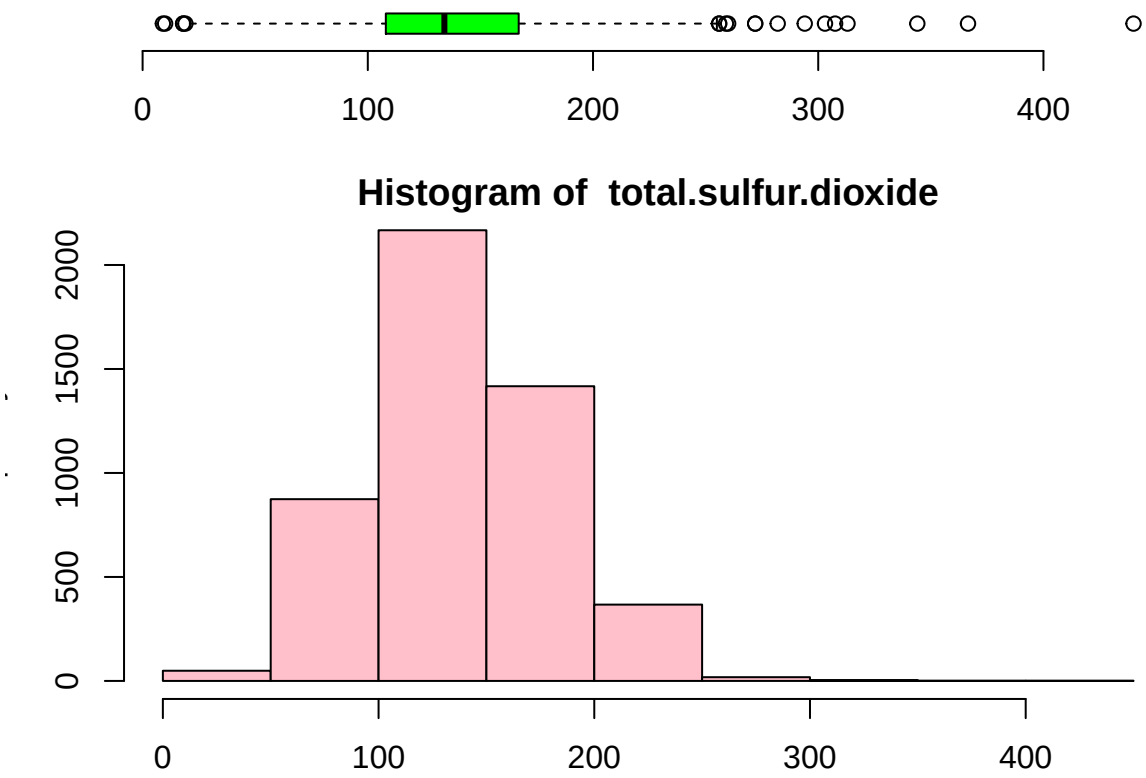
```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## Min.           3.8             0.08        0.00            0.6     0.009
## Max.          14.2             1.10        1.66           65.8     0.346
##      free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
## Min.                   2                    9  0.9871 2.72      0.22
## Max.                 289                  440  1.0390 3.82      1.08
##      alcohol
## Min.     8.0
## Max.    14.2
```

From the table above, it seems no problem exist in the difference between mean and median for every variable. However, there are 4 variables we need to dig further becuase of their ranges: fixed.acidity, residual.sugar, free.sulfur.dioxide and total.sulfur.dioxide.

We check the histograms and boxplots of these variables.


4
6
8
10
12
14
Histogram of fixed.acidity
0
500
1000
2000
4
6
8
10
12
14
0
10
20
30
40
50
60
Histogram of residual.sugar
0
500
1000
2000
0
10
20
30
40
50
60
70


Histogram of free.sulfur.dioxide
0
50
100
150
200
250
300
0
500
1000
1500
2000



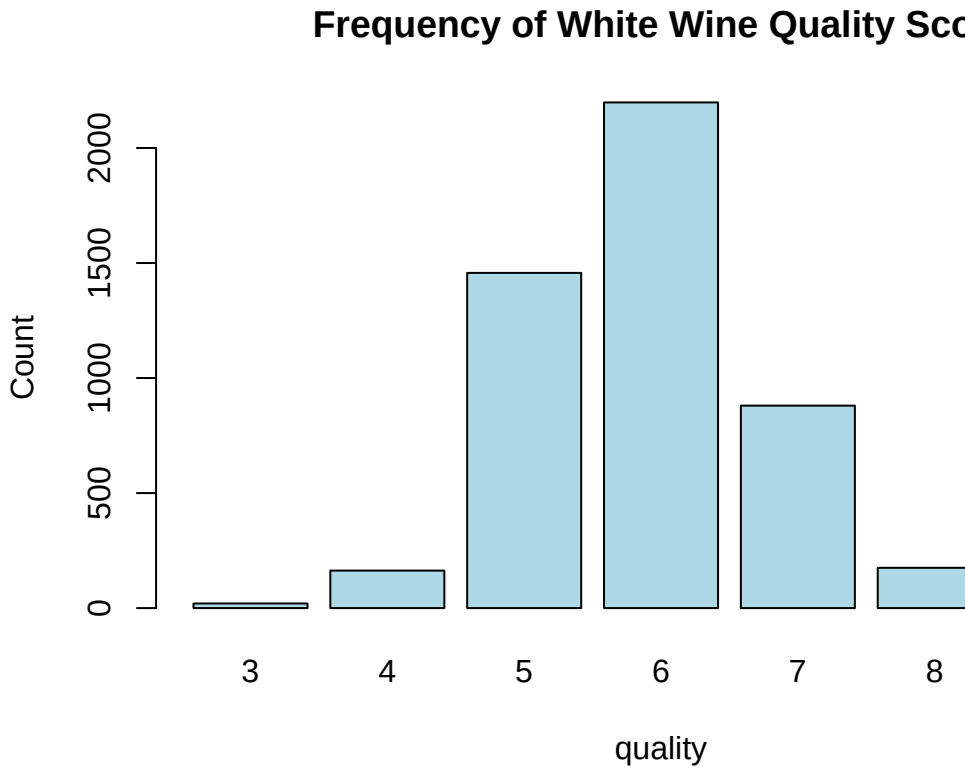
From the histograms above, we could see these variables contain some extreme values. In next step, we calcuate these outliers with IRQ.

```
## $fixed.acidity
## [1] 10.7 10.7 14.2 11.8
##
## $residual.sugar
## [1] 65.8
##
## $free.sulfur.dioxide
## [1] 131.0 122.5 118.5 146.5 128.0 138.5 124.0 289.0
##
## $total.sulfur.dioxide
## [1] 366.5 440.0
```

We could see according to IQR, these variables all had some extreme values. But we cannot just remove them as outliers as these values may be important criteria to score the wine. Let us see the relation between extrem values and quality of wine.

```
##
## 3 4 5 6 7 8 9
## 6 2 2 4 0 0 0
```

There are 14 extreme values in total and we could see wines scored to "low"(score less than 4) cover most of the extreme cases, which means we should keep these values in the dataset. It is very possible that they are not outliers or errors, but important criteria leading to low score to a wine.



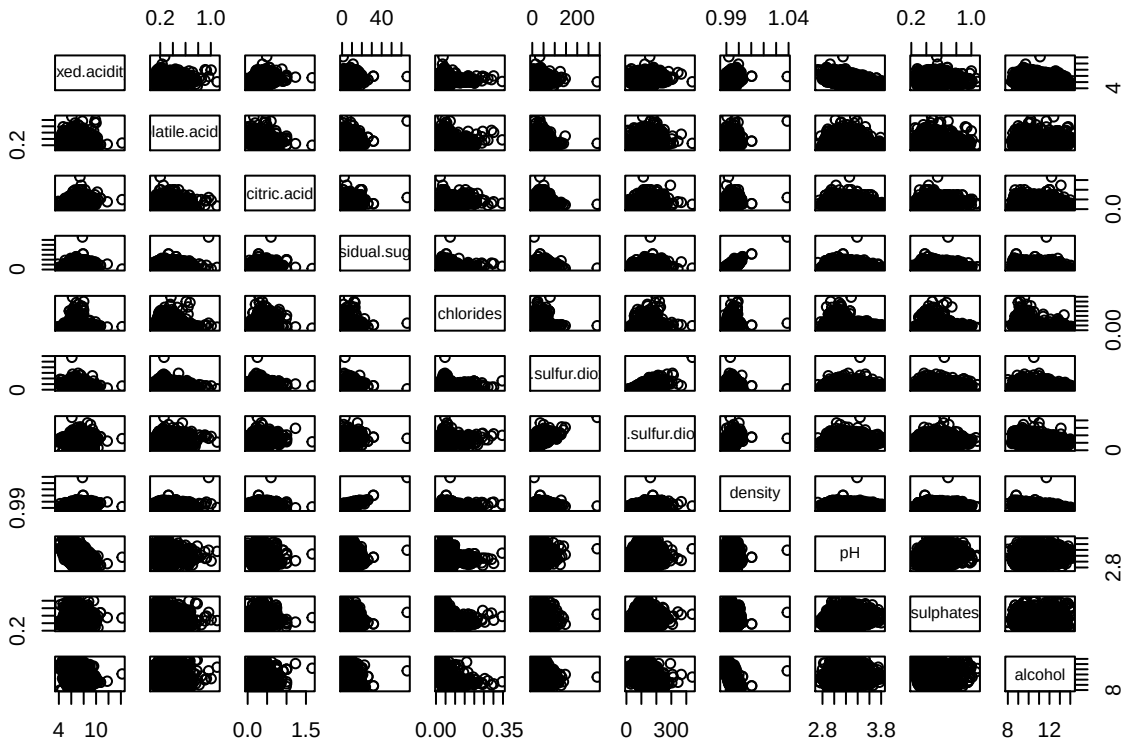
2.4 Analyze the Outcome Variable

From the chart of quality, we know there are 7 levels in the quality measurement. Most wines are scored to the medium level, which ranged from 5 to 7. Very few wines get excellent score(score greater than 8) or evaluated as "bad"(score less than 4).

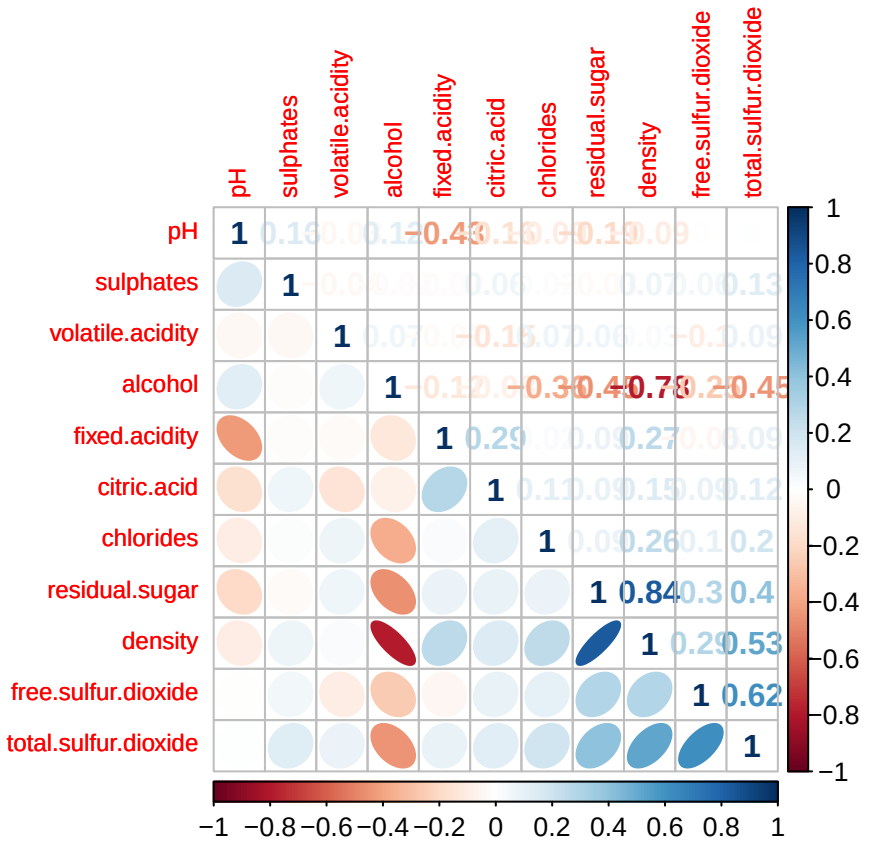
3. Bivariate Analysis

This part is to check relationship between variables. The first part we check the how the predictors related and the second part we check how they related to the outcome variable.

3.1 Relation between Predictors First let us see the scatter plot to have some basic idea how they related.



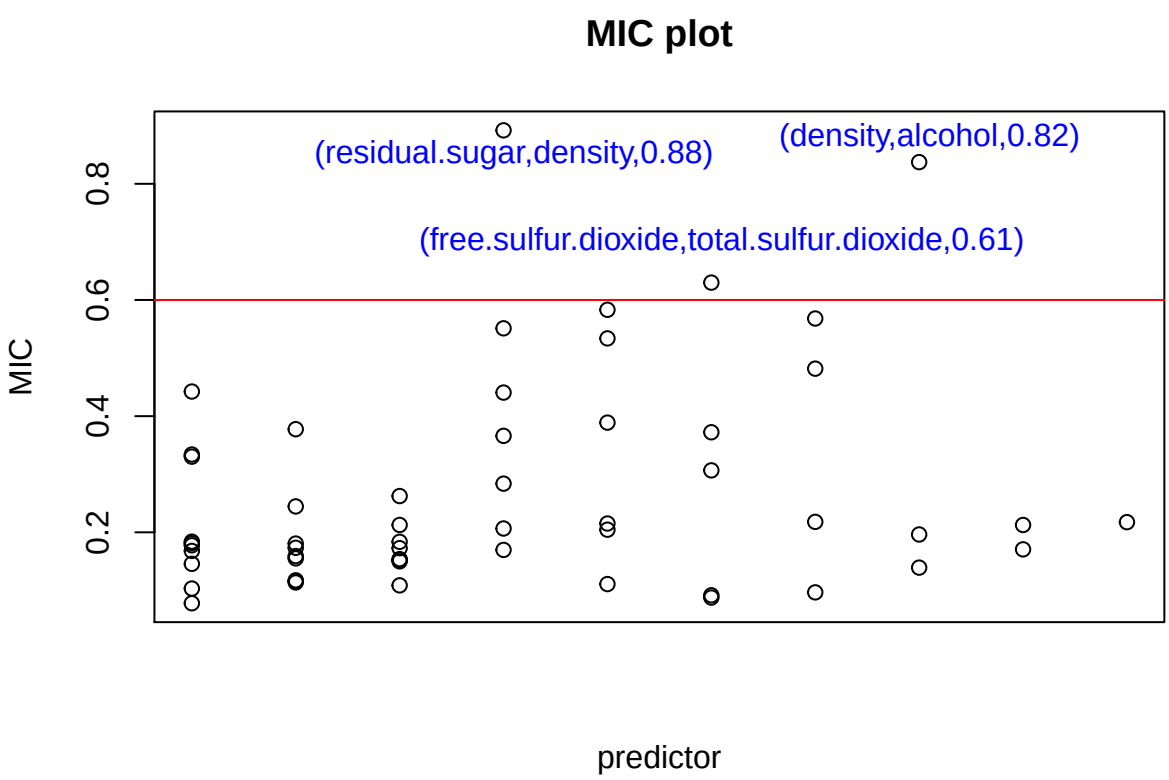
We could see most predictors are not linearly correlated. Let us see their correlation plot to confirm this.



Linear relationship is not obvious among the predictors, except two pairs, *residual.sugar* and *density*, *density* and *alcohol*. The absolute values of correlation coeffitions for these two pairs are higher than 0.7, pretty high, which refer to strongly correlation.

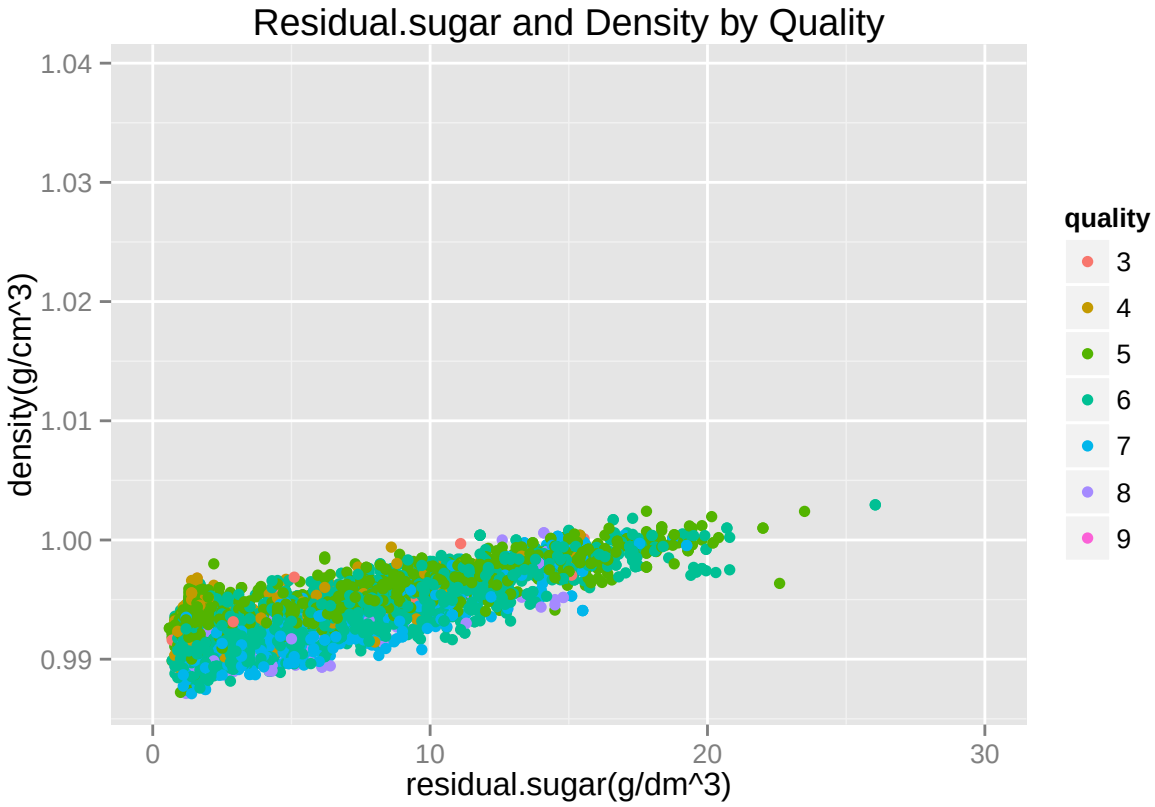
Now we use **Maximal Information Coefficient(MIC)** to check furthur. MIC could test not only linear, but non-linear relation between variables.

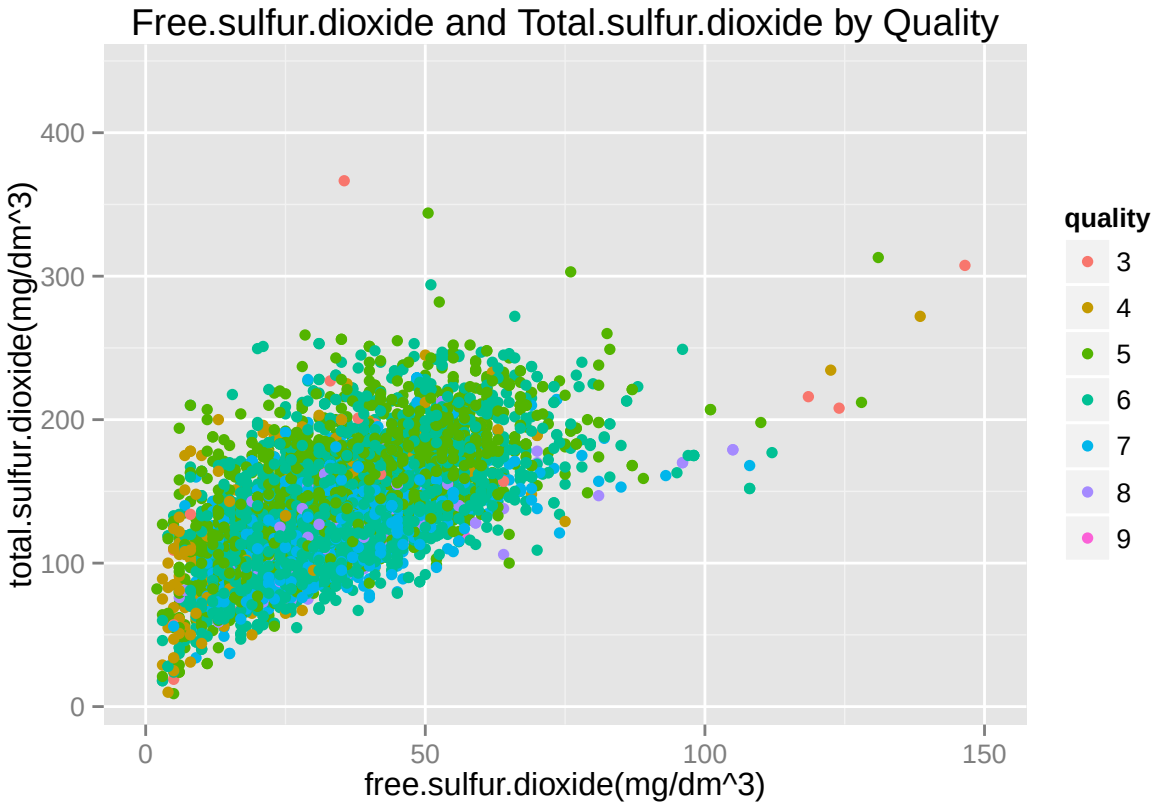
```
##          first variable      second variable       MIC
## 31       residual.sugar              density 0.8920264
## 52              density              alcohol 0.8373402
## 41  free.sulfur.dioxide total.sulfur.dioxide 0.6299096
## 40            chlorides              alcohol 0.5831443
## 46 total.sulfur.dioxide              density 0.5680982
## 34       residual.sugar              alcohol 0.5512067
```

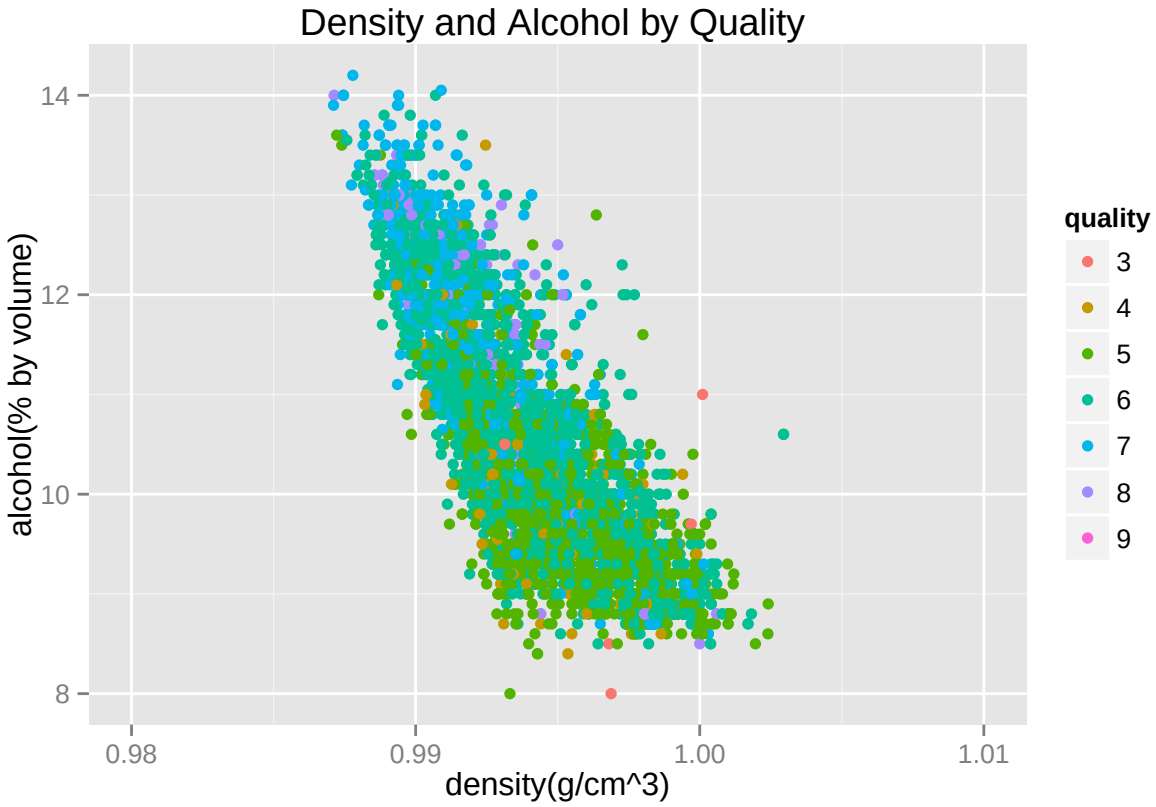


Now we could see, moderate relations exist among several pairs of predictors, **free.sulfur.dioxide** and **total.sulfur.dioxide**, **chlorides** and **alcohol** or **total.sulfur.dioxide** and **density**, etc.

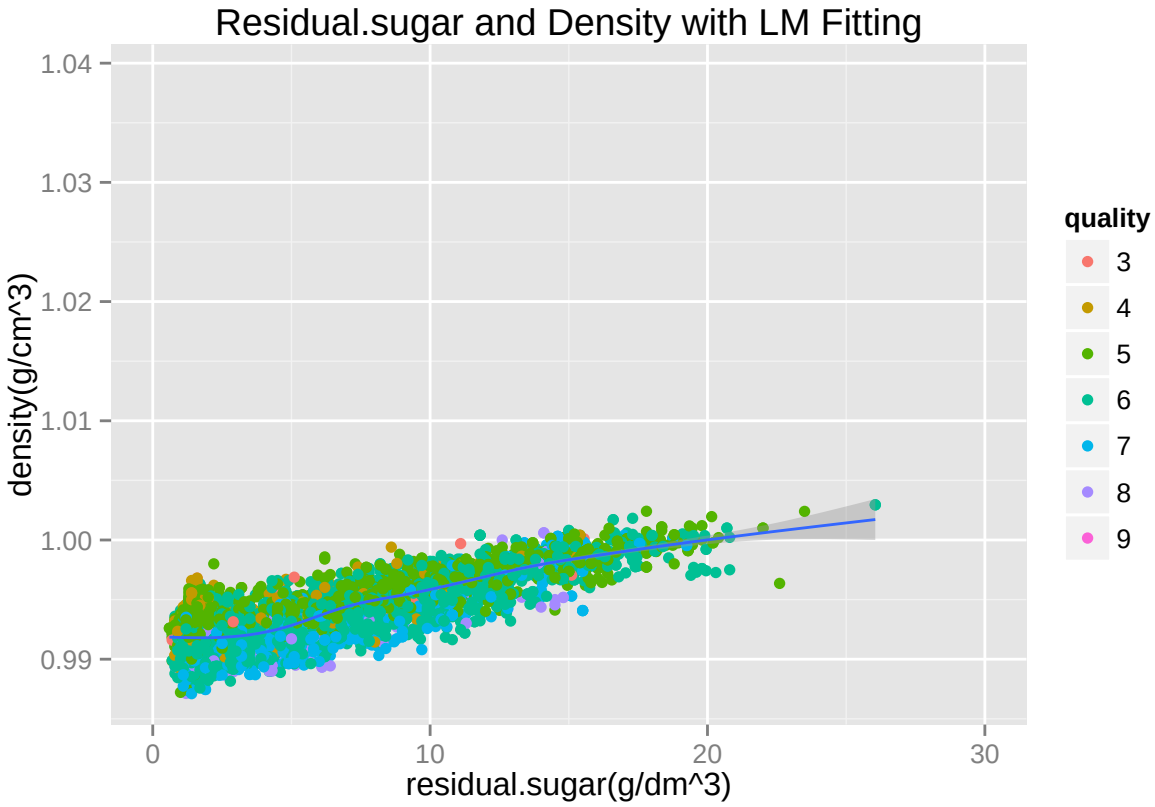
Let us see how these variables related for detail by their scatter plots.

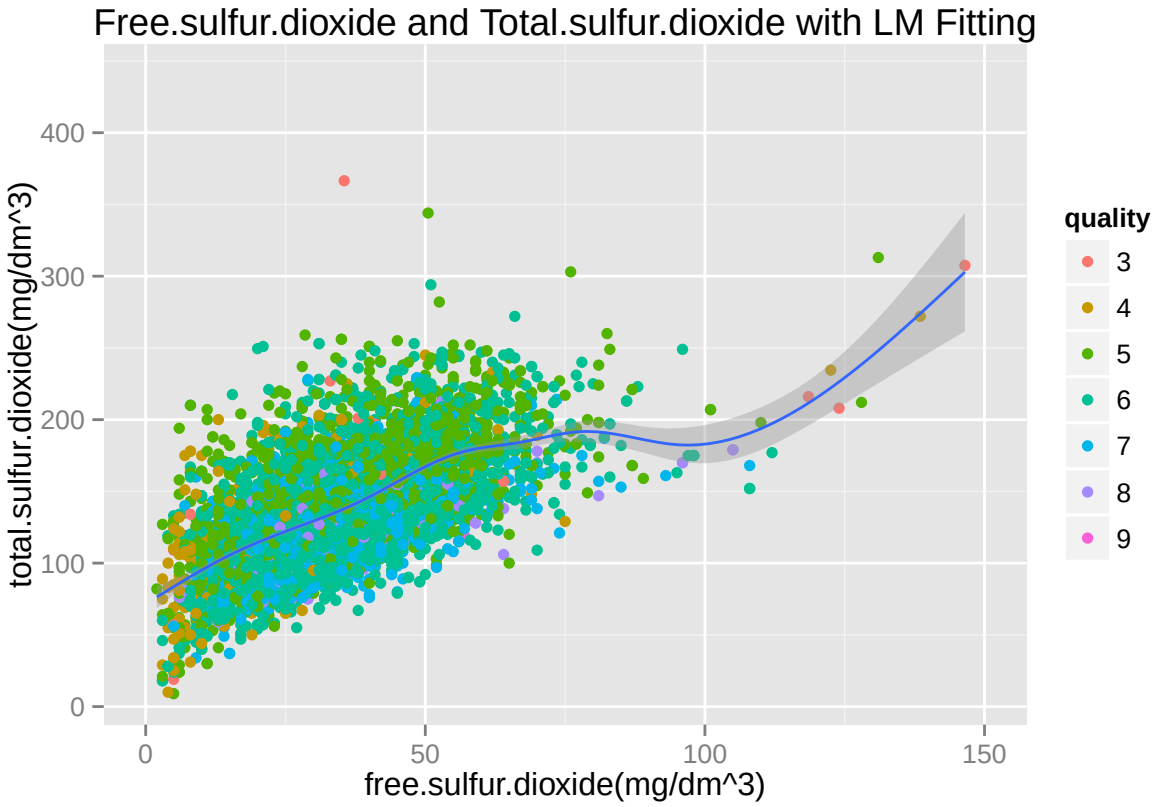

Residual.sugar and Density by Quality
density(g/cm^3)
1.04
1.03
1.02
1.01
1.00
0.99
0
10
20
30
residual.sugar(g/dm^3)
quality
3
4
5
6
7
8
9

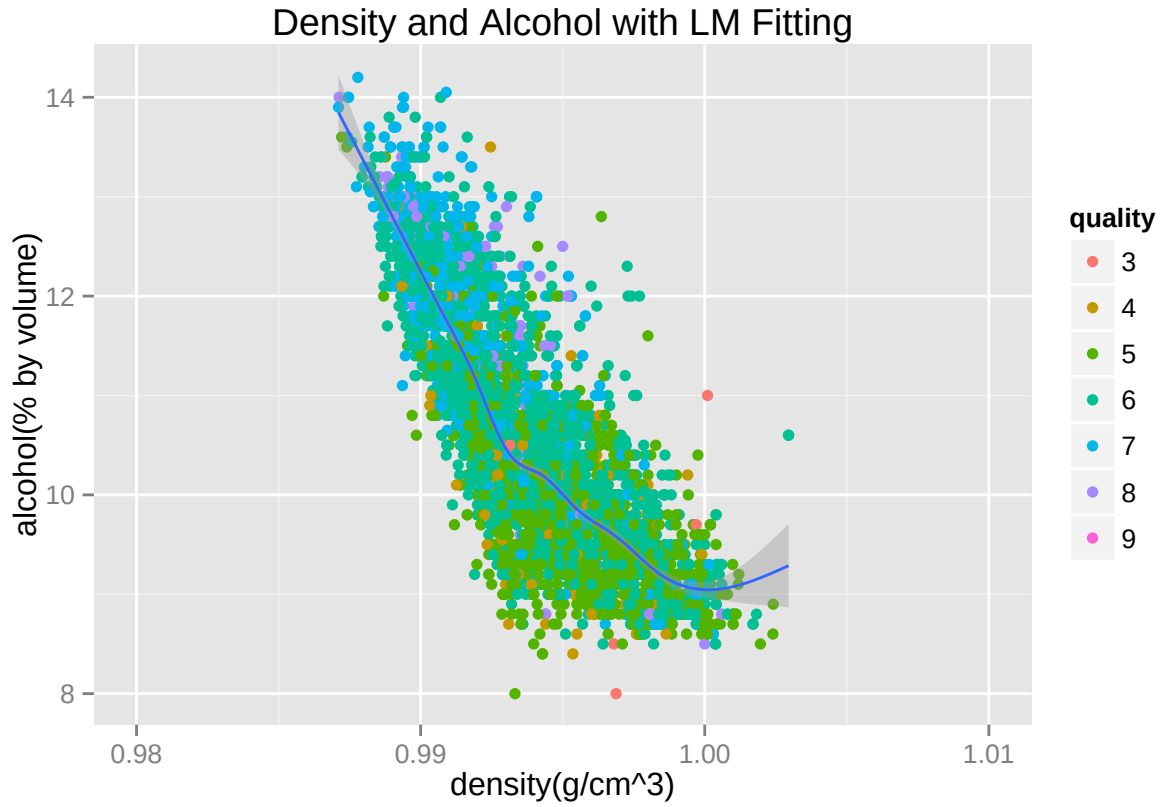

Free.sulfur.dioxide and Total.sulfur.dioxide by Quality
total.sulfur.dioxide(mg/dm^3)
400
300
200
100
0
0
50
100
150
free.sulfur.dioxide(mg/dm^3)
quality
3
4
5
6
7
8
9



The Scatter Plots above show that the first two are positively related and the last pair is negatively related. Let us confirm this by adding a linear regression fitting line into the graphs.


Residual.sugar and Density with LM Fitting
density(g/cm^3)
1.04
1.03
1.02
1.01
1.00
0.99
0
10
20
30
residual.sugar(g/dm^3)
quality
3
4
5
6
7
8
9


Free.sulfur.dioxide and Total.sulfur.dioxide with LM Fitting
total.sulfur.dioxide(mg/dm^3)
400
300
200
100
0
0
50
100
150
free.sulfur.dioxide(mg/dm^3)
quality
3
4
5
6
7
8
9



The first two plots confims our hypothesis in the positive linear relationship within the first two paris. The final plot gives great information that **density** and **alcohol** are linearly negatively related when **density** smaller than 1,after that, they positively related, but this only applies to very few cases. Their relationship is like a quadratic relationship.

3.2 Relation between Predictors and Outcome First use Spearman Ratio Correlation Test.

```
##               predictor spearman correlation
## 1         fixed.acidity              0.00000
## 2      volatile.acidity              0.00000
## 3           citric.acid              0.19956
## 4        residual.sugar              0.00000
## 5             chlorides              0.00000
## 6   free.sulfur.dioxide              0.09703
## 7  total.sulfur.dioxide              0.00000
## 8               density              0.00000
## 9                    pH              0.00000
## 10            sulphates              0.01971
## 11              alcohol              0.00000
```

The p value of the test shows that except **citric.acid** and **sulphates**, other variables are statistically significantly related to the outcome.

Next use Kruskal-Wallis test to see if the same result would appear.

```
##               predictor kruskal.wallis
```

```
## 1         fixed.acidity        0.00000
## 2      volatile.acidity        0.00000
## 3           citric.acid        0.04117
## 4        residual.sugar        0.00000
## 5             chlorides        0.00000
## 6   free.sulfur.dioxide        0.00000
## 7  total.sulfur.dioxide        0.00000
## 8               density        0.00000
## 9                    pH        0.00000
## 10            sulphates        0.03219
## 11              alcohol        0.00000
```

We get the same result from these two test! In the futhur analysis, we will check how these feactures effect the quality of a wine and order of their importance.

4.Multivariate Analysis

In this part, I used 3 different algorithms to test the importances of the predictors, sort them by the order and record the order as a *rank score* every time. Finally, each feature would have 4 rank score and sum their score to get a final score as the final importance order of the feature. The lower the score, the more effects it has on the outcome.

4.1 Learning vector quantization(LVQ) The first model is LVQ, a prototype-based supervised classification algorithm.

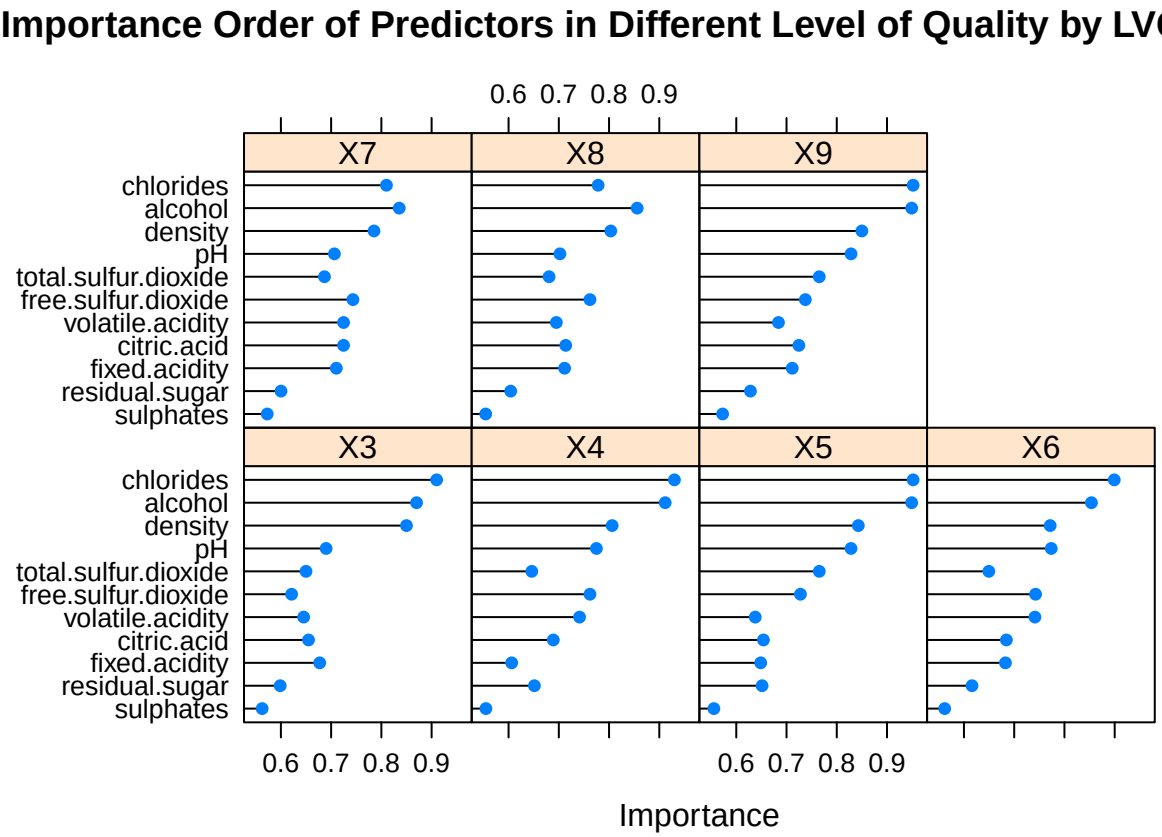
```
## ROC curve variable importance
##
##   variables are sorted by maximum importance across the classes
##                           X3     X4     X5     X6     X7     X8     X9
## chlorides             0.9100 0.9301 0.9520 0.8991 0.8101 0.7783 0.9520
## alcohol               0.8700 0.9117 0.9491 0.8537 0.8355 0.8560 0.9491
## density               0.8500 0.8061 0.8429 0.7716 0.7854 0.8034 0.8500
## pH                    0.6900 0.7748 0.8284 0.7737 0.7065 0.7023 0.8284
## total.sulfur.dioxide  0.6500 0.6462 0.7652 0.6495 0.6867 0.6806 0.7652
## free.sulfur.dioxide   0.6213 0.7620 0.7277 0.7425 0.7434 0.7620 0.7374
## volatile.acidity      0.6453 0.7413 0.6377 0.7413 0.7247 0.6951 0.6840
## citric.acid           0.6550 0.6890 0.6540 0.6844 0.7247 0.7137 0.7247
## fixed.acidity         0.6773 0.6061 0.6487 0.6826 0.7103 0.7114 0.7114
## residual.sugar        0.5986 0.6514 0.6514 0.6160 0.6002 0.6042 0.6282
## sulphates             0.5627 0.5549 0.5555 0.5616 0.5728 0.5544 0.5728

##             chlorides               alcohol               density
##              6.231508              6.225171              5.709413
##                    pH   free.sulfur.dioxide      volatile.acidity
##              5.304142              5.096365              4.869363
##           citric.acid  total.sulfur.dioxide         fixed.acidity
##              4.845445              4.843412              4.747855
##        residual.sugar             sulphates
##              4.350044              3.934724
```

The first importance table gives the order of the importance of each feature in every level quality. The second one sums the importance acorss all levels of quality for each feature to see the importance order of each

feature without classifying the level of quality. The result given by **LVQ** shows that **chlorides**,**alcohol** and **density** are top 3 imporant feature in scoring a white wine, **sulphates** is the least important one, which matches the result given by two testings we did in the previous part.

LVQ also visualizes the importances level of each feature in differenct wine quality level, plot is shown below.



The top 2 features are much more important in every level of quality than others obviously in the plot.

4.2 Boruta The second model is Boruta. Below is a paragrpha of short introduction to this algorithm from Wikipedia.

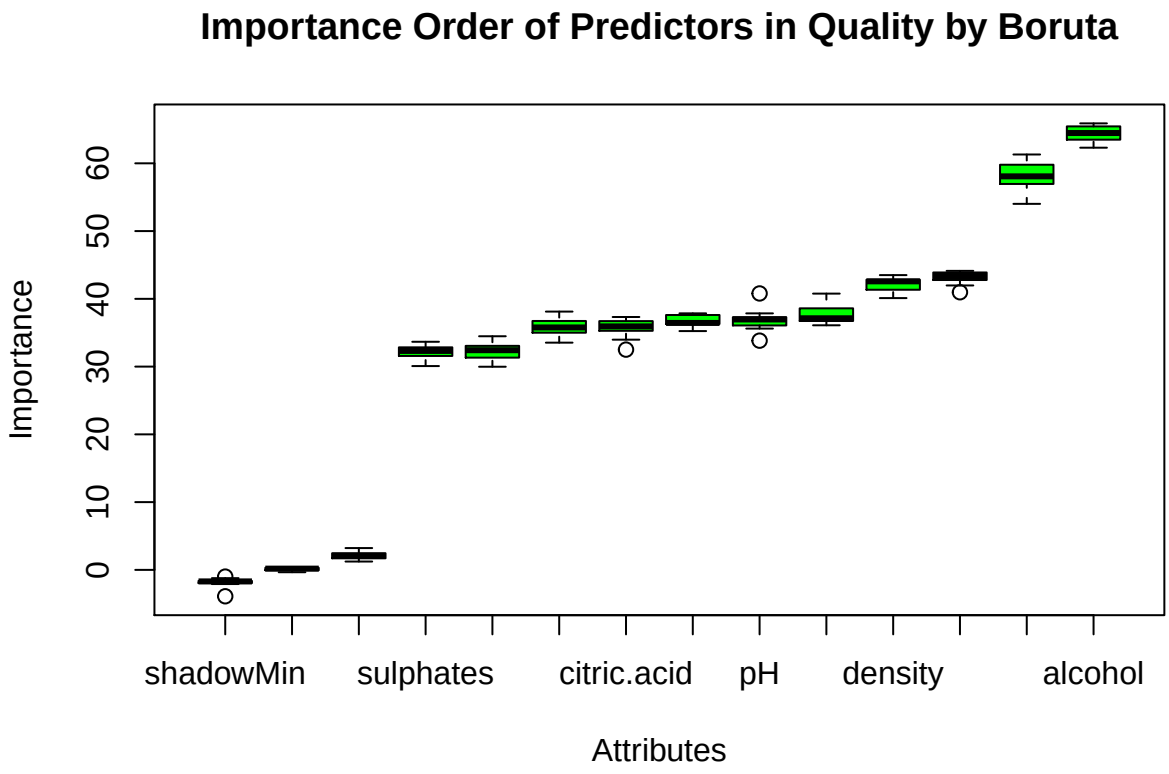
Boruta is an algorithm in the field of machine-learning, and more specifically, a feature-selection algorithm. The aim of the algorithm as presented in the original paper describing it is to find all relevant features (compare with minimal-optimal features set). The Boruta algorithm is not a stand-alone algorithm, but is implemented as a wrapper algorithm around the random-forest classification algorithm.

```
##              alcohol     volatile.acidity  free.sulfur.dioxide
##             707.9869             641.1375             474.5285
##              density            chlorides                   pH
##             463.3990             414.4095             405.4343
##       residual.sugar total.sulfur.dioxide          citric.acid
##             404.0347             394.3448             392.9291
##        fixed.acidity            sulphates
##             355.5135             353.8145
```

The result given by **Boruta** is **volatile.acidity**,**alcohol** and **free.sulfur.dioxide** are top 3 imporant feature in scoring a white wine, **sulphates** is still the least important one, which matches the result given by two

testings we did in the previous part. **density** ranks No.4. The dramatic difference between the importance order given by these two algorithms appears among **volatile.acidity** and **chlorides**.

Boruta visulize the result of importance order in the form of boxplot,which is shown below.



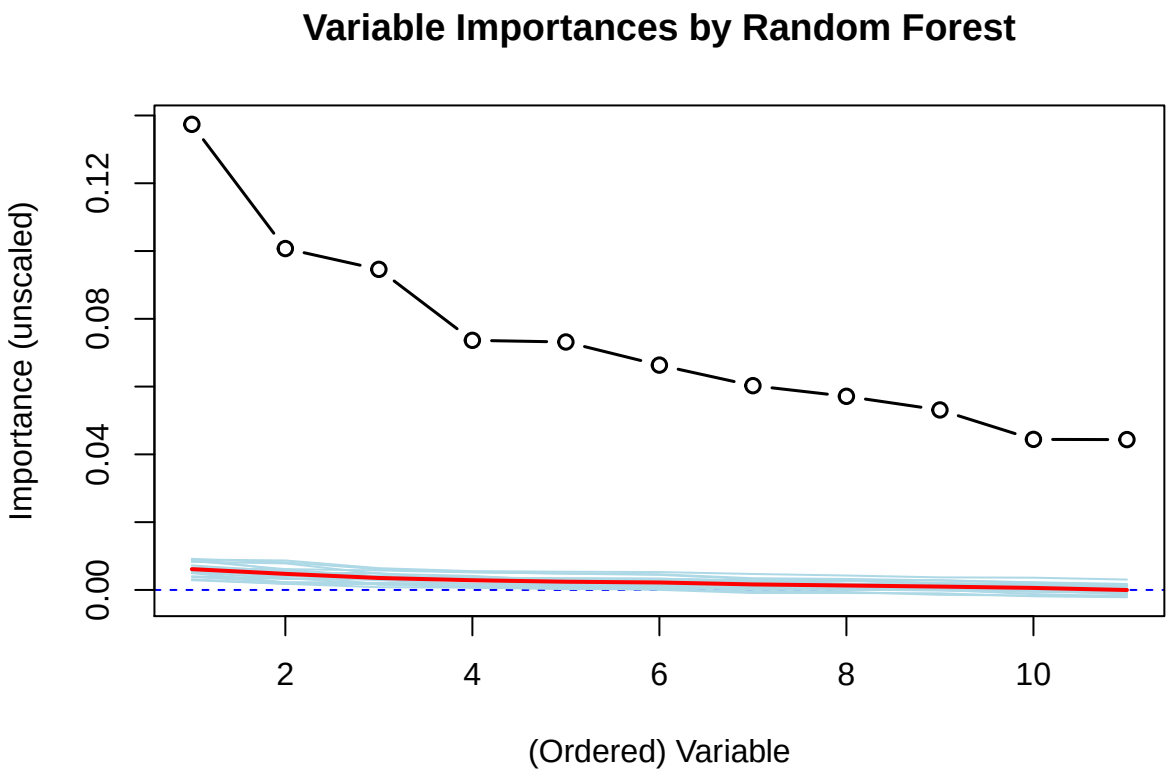
Alcohol and **Volatile.acidity** are more important than others a lot in the scoring of white wine quality, as chosen by Boruta. **Free.sulfur.dioxide** outshines **density** slightly, following the top 2.

4.3 Random Forest The third model is Random Forest.

```
FALSE  [1] "density"              "alcohol"              "residual.sugar"
FALSE  [4] "total.sulfur.dioxide" "free.sulfur.dioxide"  "fixed.acidity"
FALSE  [7] "chlorides"            "pH"                   "volatile.acidity"
FALSE [10] "citric.acid"          "sulphates"
```

The result given by **Random Forest** is **alcohol**,**density** and **residual.sugar** are top 3 imporant feature in scoring a white wine, **sulphates** and **volatile.acidity** are the least important ones.

Random Forest provides the plot showing importance level and number of variables.



We can see importance decreased dramatically in the fourth variable, which means the top 3 ones are far more important than others.

4.4 Final rank table Now we sum the rank score given by the 3 algorithms and see the final result.

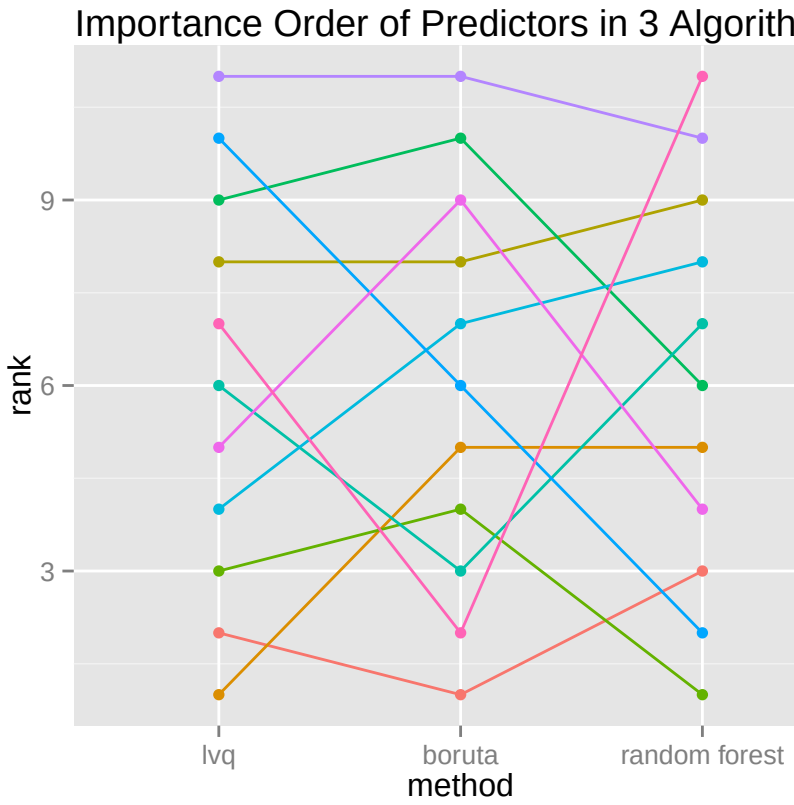
```
##    colnames.predictors. lvq boruta rf final.rank.score
## 1         fixed.acidity   9     10  6              25
## 2      volatile.acidity   7      2 11              20
## 3           citric.acid   8      8  9              25
## 4        residual.sugar  10      6  2              18
## 5             chlorides   1      5  5              11
## 6   free.sulfur.dioxide   6      3  7              16
## 7  total.sulfur.dioxide   5      9  4              18
## 8               density   3      4  1               8
## 9                    pH   4      7  8              19
## 10            sulphates  11     11 10              32
## 11              alcohol   2      1  3               6

##  [1] alcohol              density              chlorides
##  [4] free.sulfur.dioxide  residual.sugar       total.sulfur.dioxide
##  [7] pH                   volatile.acidity     fixed.acidity
## [10] citric.acid          sulphates
## 11 Levels: alcohol chlorides citric.acid density ... volatile.acidity
```

From the rank table, we could see **alcohol** and **density** are far more important than others in the effect on **quality**, followed by **chlorides**, **free.sulfur.dioxide** and **volatile.acidity**. These are top 5 features effect

the quality of white wine.

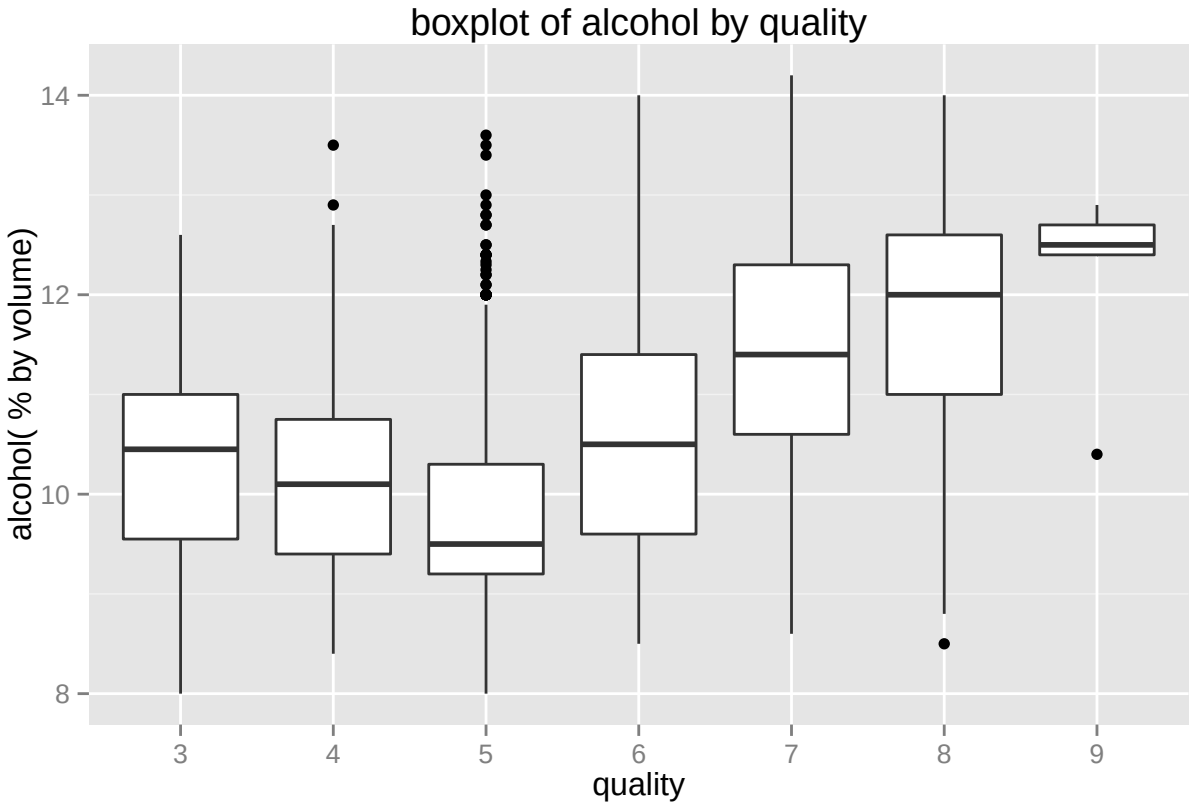
In terms of the least important one, the last 3 are **sulphates**, **citric.acid** and **fixed.acidity**. The first two are not suprising as we already get this from the Spearman Test and Kruskal-Wallis test.

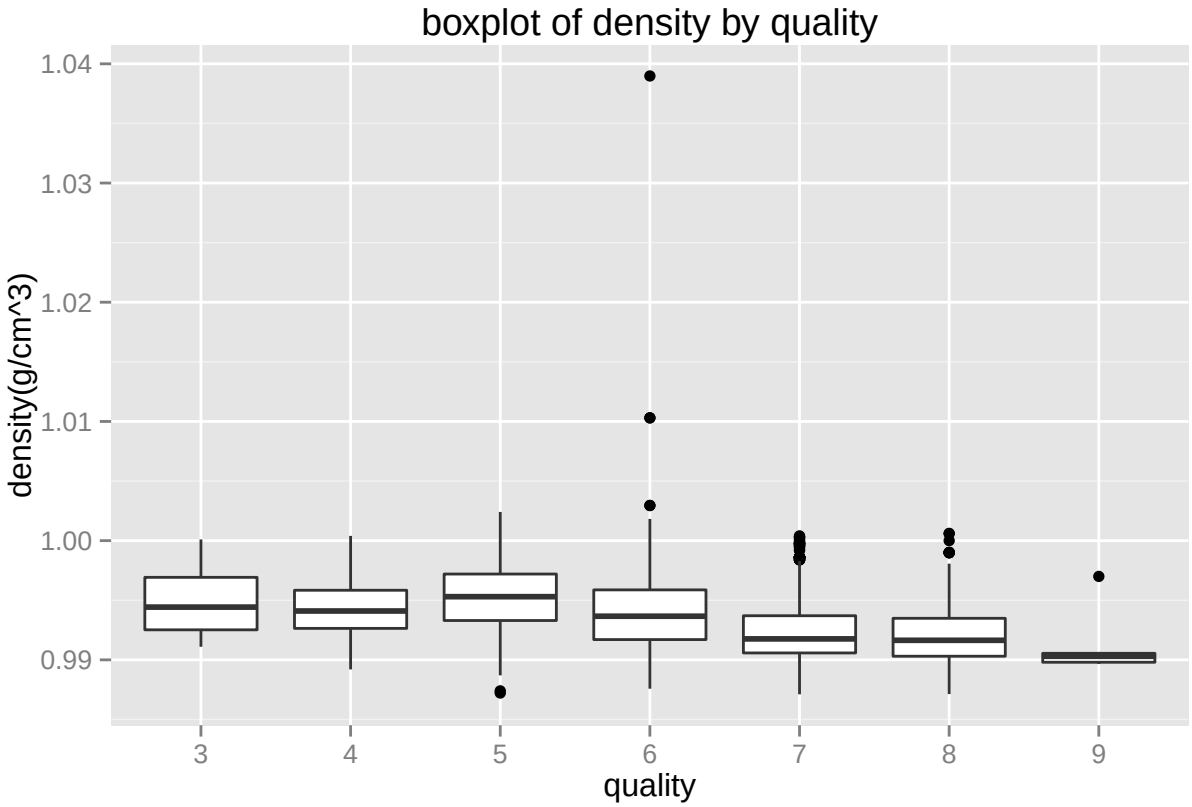


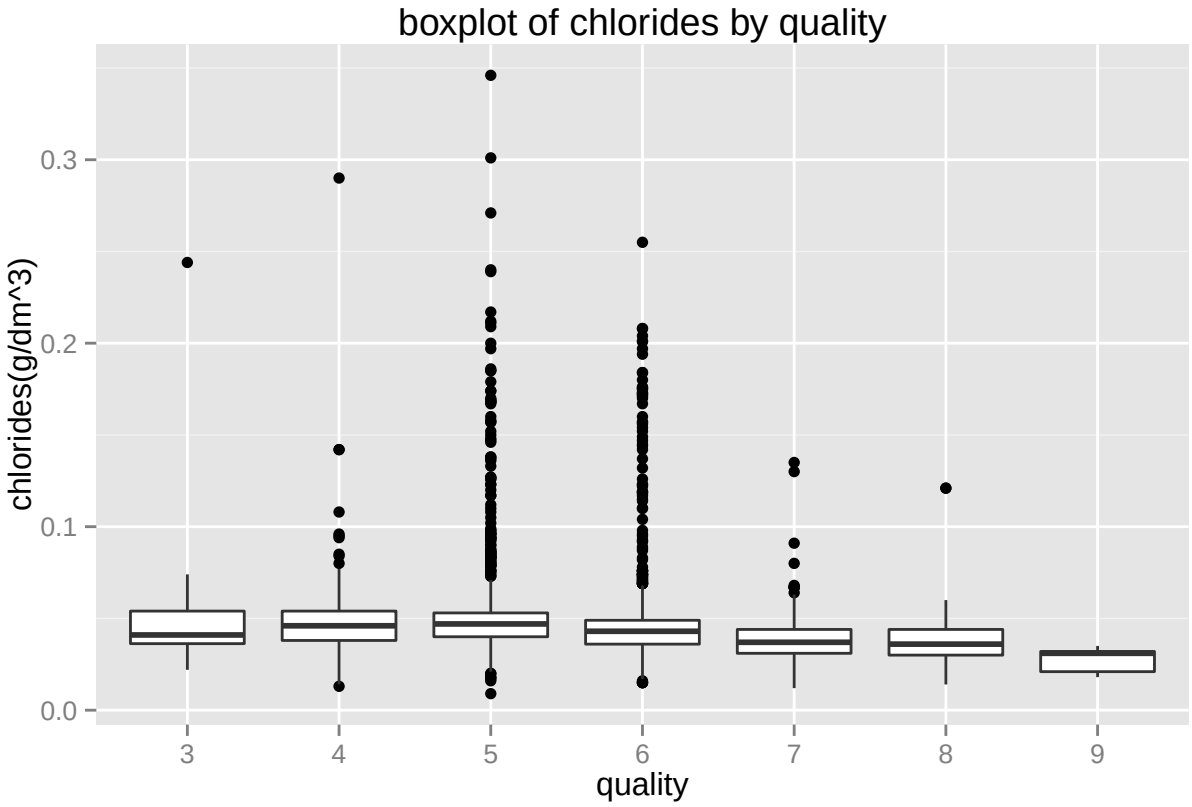
Let us visualize the order of importance of each feature.

The most stable one is **alcohol**,followed by **density**, which means these two are most important in the scoring of quality of white wine,which could be confirmed in all 3 algorithms.**Chlorides** is good in the **lvq**, but less important as selected by the last two algorithm. Even though, it still more important than the left.

In the end of this part, we would like to see how thers 3 top features related to **quality** in visualization.

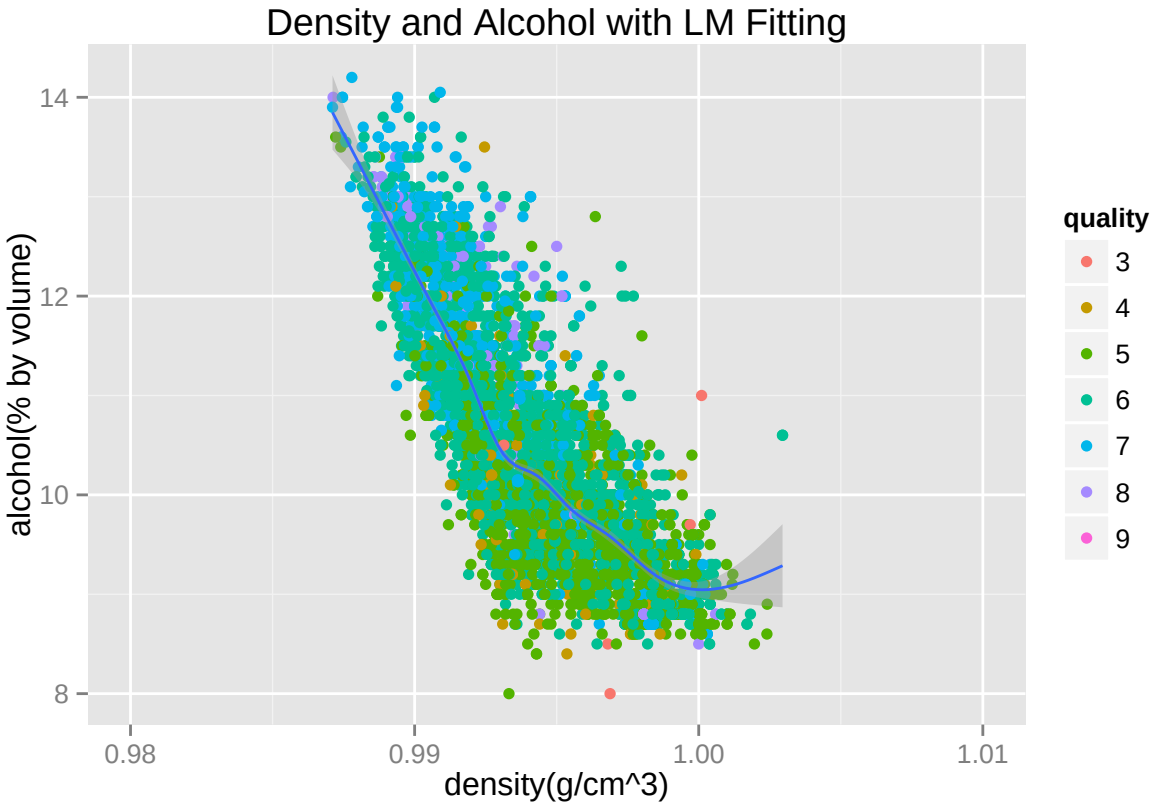

boxplot of alcohol by quality
alcohol(% by volume)
14
12
10
8
3
4
5
6
7
8
9
quality


boxplot of density by quality
density(g/cm^3)
1.04
1.03
1.02
1.01
1.00
0.99
3
4
5
6
7
8
9
quality



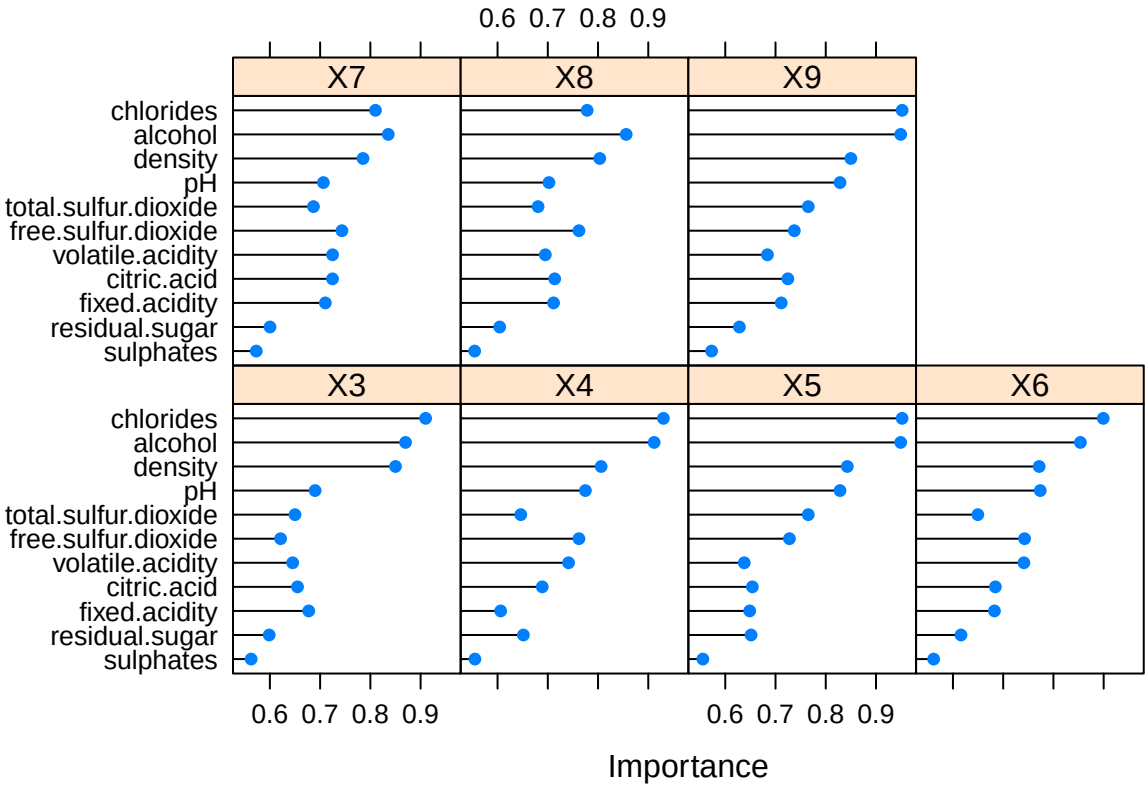
Alcohol is extremely high in white wine with good quality and is lowest in the medium level wine. In contrast, **density** and **chlorides** are low in good quality wine.

5.Final Plot



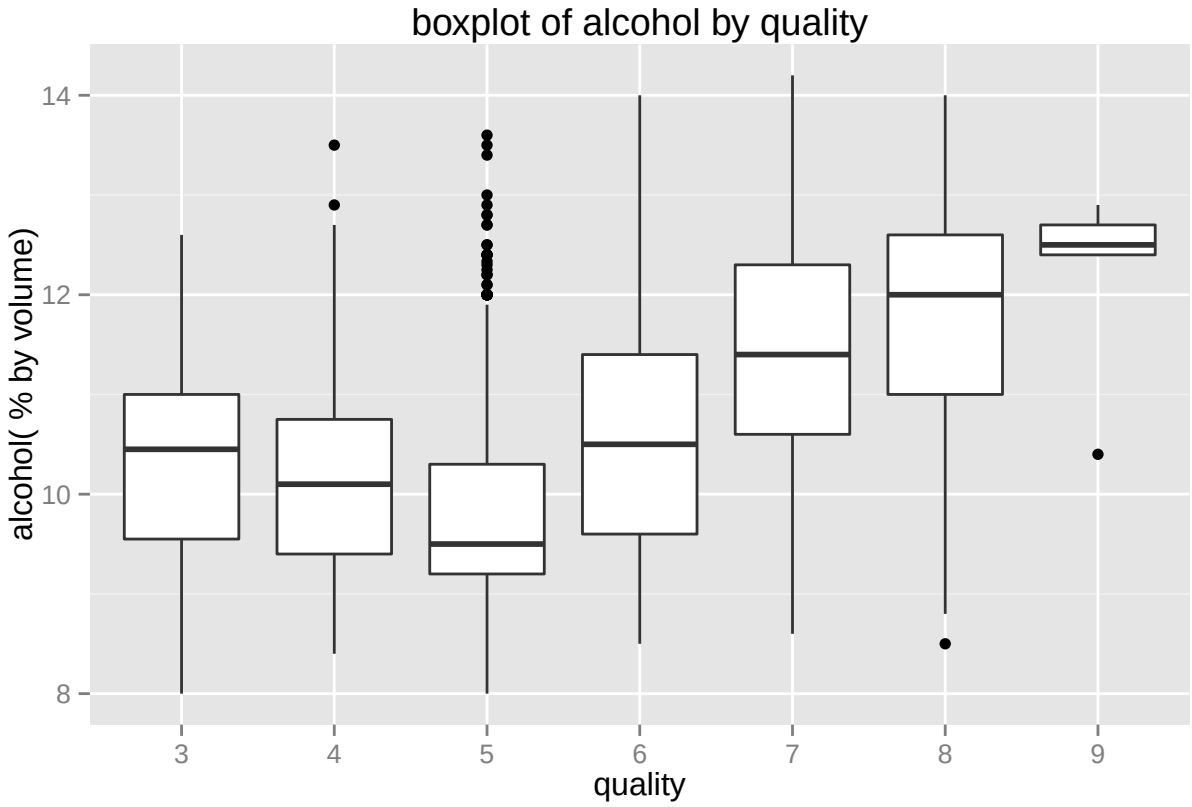
5.1 Plot 1

- Description 1: In white wine, **alcohol** and **density** are quadratic related. **Alcohol** decreases with the growth of **density** until **density** reaches to 1.01 g/cmˆ3. After that, **alcohol** increases with **density**. This is interpretable because **alcohol** is extremely high in both bad white wine and good white wine, while **density** in good white wine is respetively lower than in other levels. In most cases, these two featues are negatively correlated.



5.2 Plot 2

- Description 2: This is the importance table of features to every level of quality given by **lvq**. We can see the importance of the top 3 features **chlorides**,**density** and **alcohol** significantly outshine others in scoring wines with low score and high score. For those socred to medium, whose quality score ranged from 5 to 7, despite the low effects of the last two features **sulphate** and **residual.sugar**, other features did not differe from each other a lot.



5.3 Plot 3

- Description 3: This plot shows distribution of **alcohol** by different rank of quality. It is very obvious that **alcohol** is quite high in the white wine with high quality score and is lowest in the medium ones. This is reasonable as like global famous white wine **Chardonnay**, the average alchohol is greater than 13%, almost the maximum of the alcohol in our data set.In addition, **alcohol** and **density** are also higher in white wines with lower score, which makes sense because bad white wines may contain Methanol.

6.Reflection

This EDA report is done with a dataset containing 4898 observations of 12 variables, 11 of them ar predicitors, which are elements as criteria to scoring the quality of a white wine and the outcome variable is the socre of a certian white wine, which is an ordinal variable.

From the analysis above, we get to the conclusion that **alcohol**,**density** and **chlorides** are the most important features in scoring the quality of a white wine, while **sulphate** and **citric.acid** are least important criteria in the scoring.

Alochol in White wine with high score are higher than wines in other level, but white wines in low level are also higher in alcohol than those in medium level, which may caused by Methanol.

Most of white wine would be scored as “medium”, which means their score ranged from 5 to 7 and few would be scored as “bad”, but “excellent” is rarer than “bad”.

While doing this report, the first challenge I met is the extreme values found in the predictors. At first I simply took them as outliers and remove from the data set, but after that, I fould the number of bad wine,whose quality score lower then 4, decreases a lot. Therefore, I realize they may be important criteria to define a bad wine and check their distribution with quality and confirm my hypothesis. So it is more

reasonable to keep them in the data set. The second challenge is the importance order given by 3 algorithms are not exactly the same, which makes the result difficult to intepret. After spending a lot of time studying how these 3 algorithms work in feature selection,instead of picking one of them, I produce a rank table to record all information from these 3 models and combine they result as the final conclusion.

The rank table could be improved furthur as we can see from the rank plot, the order of some features varied a lot in 3 algorithms. So maybe in furthur improvement, we could add weights to these 3 models to lower the variances of the orders.