WINE ANALYSIS

A PROJECT REPORT BY

Tyag Raj

EDWISOR.COM

Table of Contents

1 Problem Statement

Wine makers always relied on wine experts who use their jargons to rate wine qualities. But what exactly is wine quality based on? What are the criteria? In this project, we look at the Wine Quality dataset and analyze the data with R to explore the relationship between wine quality and it's chemical properties.

2 Data

2.1 Variables for wine data

```
[1] "fixed.acidity"       "volatile.acidity"     "citric.acid"         "residual.sugar"       "chlorides"
 [6] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"             "pH"                  "sulphates"
[11] "alcohol"
```

[12] “quality” is the target variable and others are independent variables for this data analysis.

2.2 Dimension of data

wine data set has 6497 observations and 12 variables.

2.3 Summary of wine

```
     fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides       free.sulfur.dioxide
      Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   Min.   :  1.00
    1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00
  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700   Median : 29.00
    Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603   Mean   : 30.53
    3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00
     Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100   Max.   :289.00
            total.sulfur.dioxide   density           pH          sulphates         alcohol         quality
             Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
           1st Qu.: 77.0         1st Qu.:0.9923   1st Qu:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000
      Median :118.0        Median :0.9949   Median :3.210   Median :0.5100   Median :10.30   Median :6.000
        Mean   :115.7         Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
              3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
           Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
```

2.4 Observations from the summary of wine

1. There is a big range for sulfur. dioxide (both Free and Total) across the observations.
2. The alcohol content varies from 8.00 to 14.90 for the observations in dataset.
3. The quality of the samples ranges from 3 to 9 with 6 being the median.
4. The range for fixed acidity is fairly high with minimum being 3.8 and maximum being 15.9
5. pH value varies from 2.720 to 4.010 with a median being 3.210.

2.5 Structure of dataset

All the independent variables are in numerical forms except quality variable which is in integer form.

```
$ 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 ...
```

3 Understanding Variables

1. Fixed acidity- Most acids involved with wine are fixed or nonvolatile.
2. Volatile acidity -The amount of acetic acid in wine, if it is at very high levels can lead to sour, vinegar taste.
3. Citric acid -Found in small quantities, citric acid is used to add ‘freshness’ and flavor to wines.
4. Residual sugar - The amount of sugar levels remaining after fermentation stops, it’s rare to find wines with sugar levels less than 1 gram/liter and wines with sugar level greater than 45 grams/liter are considered sweet.
5. Chlorides -It’s the amount of salt in the wine. Salty is not a common wine descriptor. It’s also not a positive one probably goes without saying.

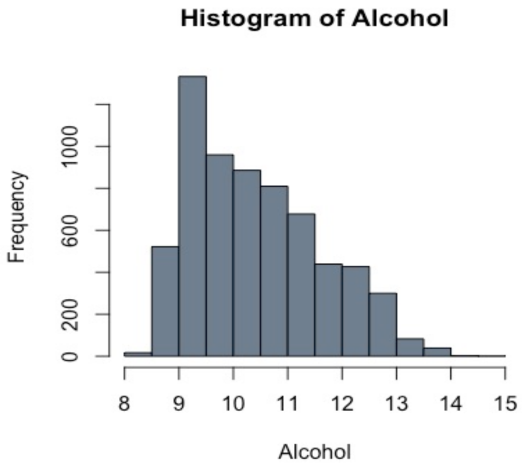
6. Free sulfur dioxide - The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.
7. Total sulfur dioxide - amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.
8. Density - 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 - A wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant.
11. Alcohol - The level of percent of alcohol content in the wine.

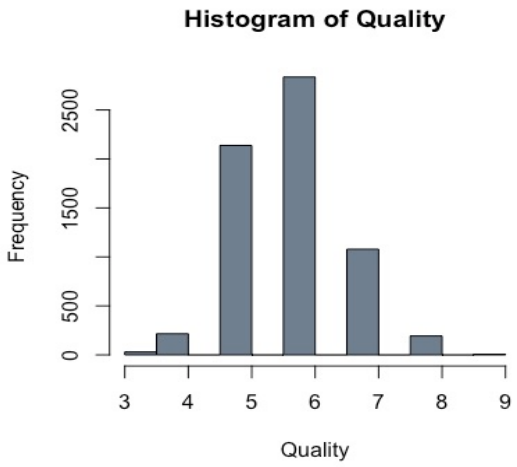
4 Exploratory Data Analysis

4.1 Missing Value Analysis

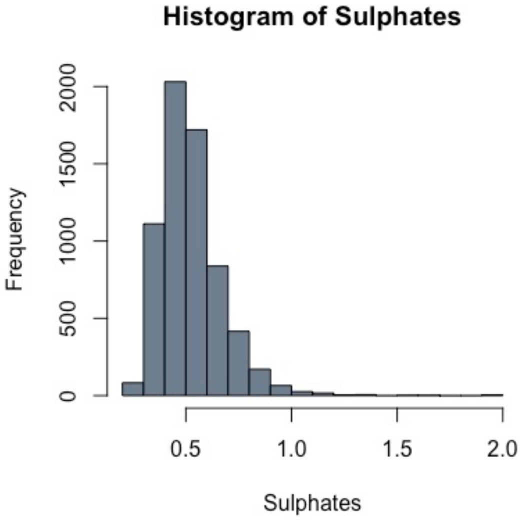
No Na's found in missing value analysis in R.

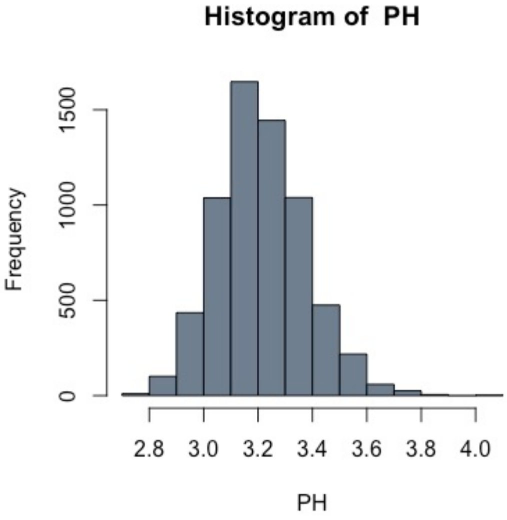
4.2 Distribution of Data

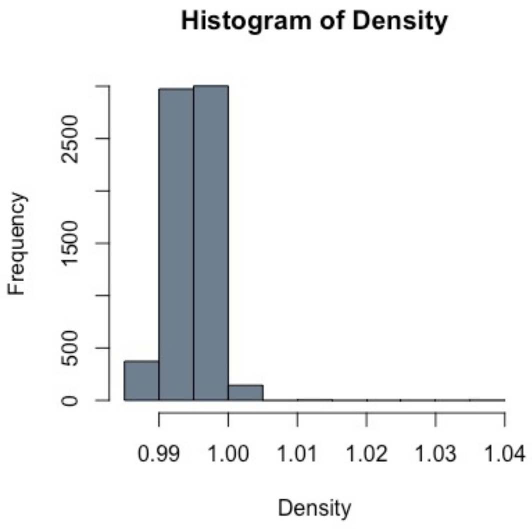


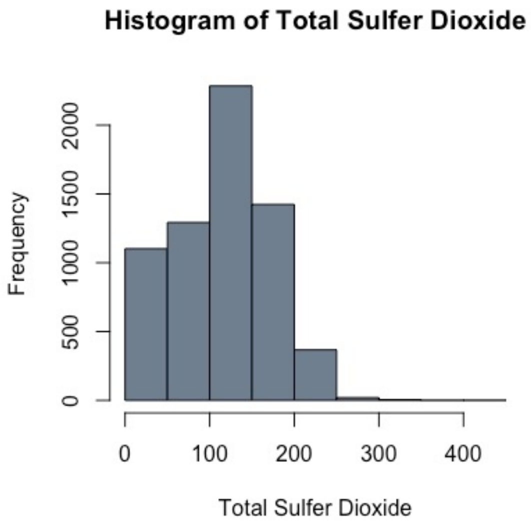


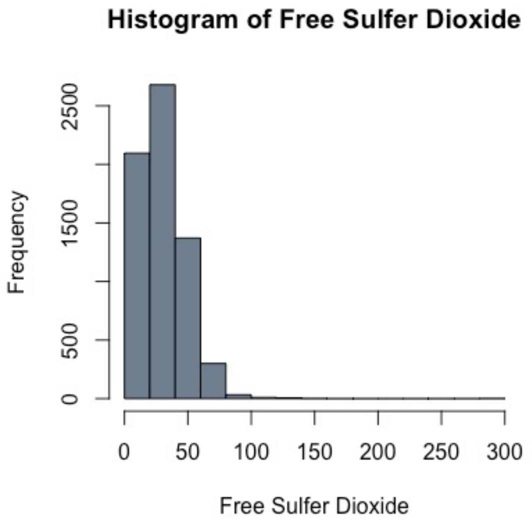
The above left histogram shows alcohol distribution. Although it is not strictly unimodal, it does exhibit some trend as the alcohol level increases, the count decreases. On the right side, histogram for quality shows, most of the values are distributed among 5,6 & 7. Distribution varies from 3 to 9.

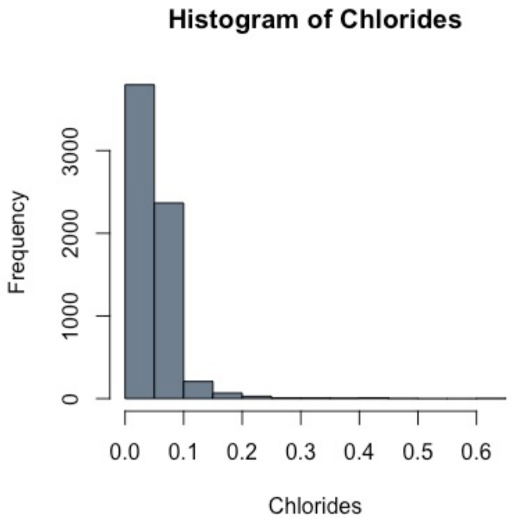




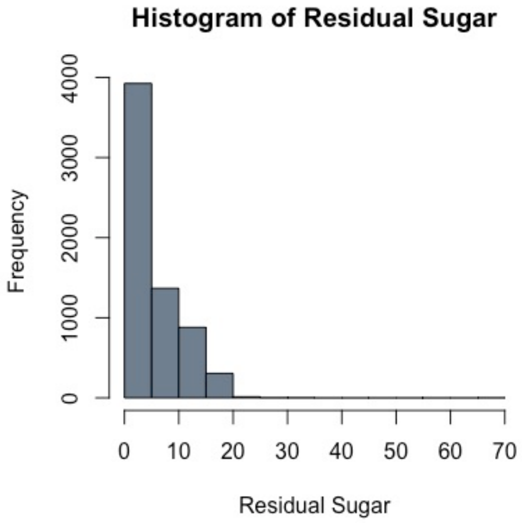


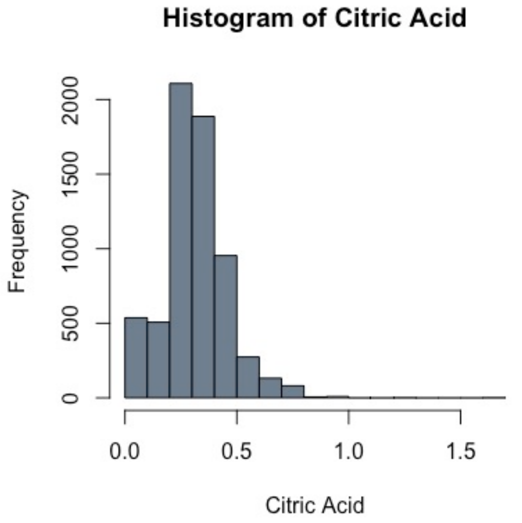




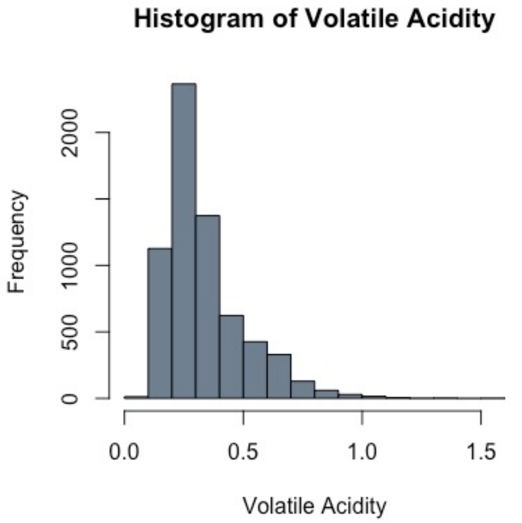


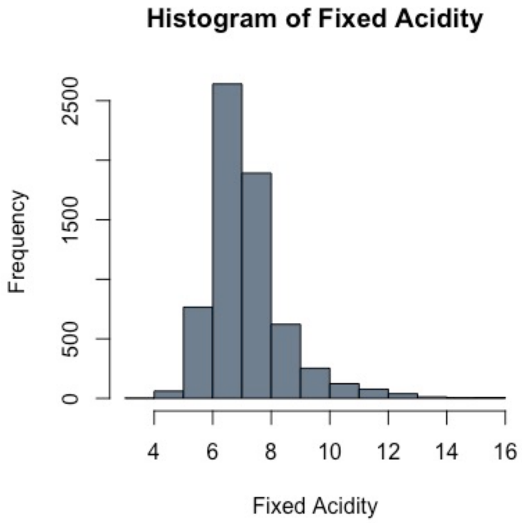
The above plots are for sulphates, pH, density, total sulfur dioxide, free sulfur dioxide and chlorides. Chlorides are very concentrated at lower levels and some outliers are present in the higher spectrum. This also may be the distinct factor between different quality levels of wine. Free and total sulfur dioxide present similar patterns of distribution, peaking at lower levels, reducing in count at higher levels. Sulphates levels are right skewed, with some outlier at the higher end. When it comes to pH and density distribution, an increasing normal pattern is visible.


Histogram of Residual Sugar


Histogram of Citric Acid

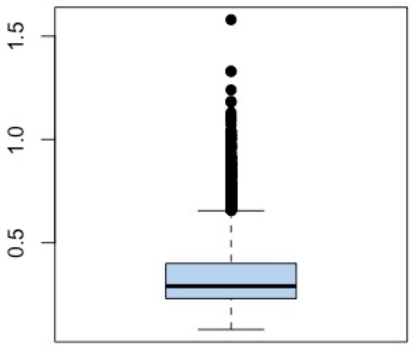
The histogram of residual sugar. As it shows, the distribution is unimodal, nearly normal and right skewed. It seems that there are outliers in the higher end, i.e. high residual sugar levels. This may potentially be the wines that have higher quality or otherwise. Citric acid is more uniform with a peak at the lower end.


Histogram of Volatile Acidity

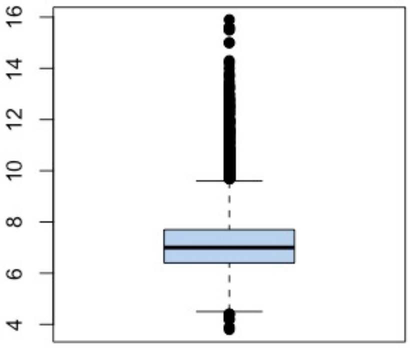

Histogram of Fixed Acidity

Both fixed and volatile acidity show somewhat normal distribution.

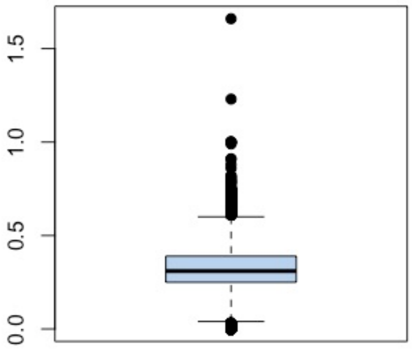
4.3 Outliers Analysis



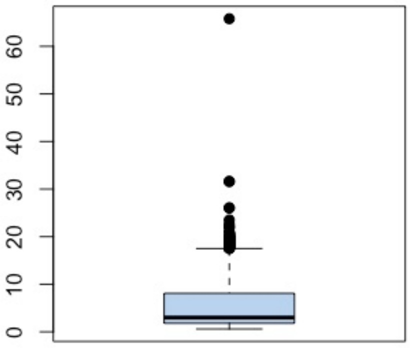
Volatile Acidity



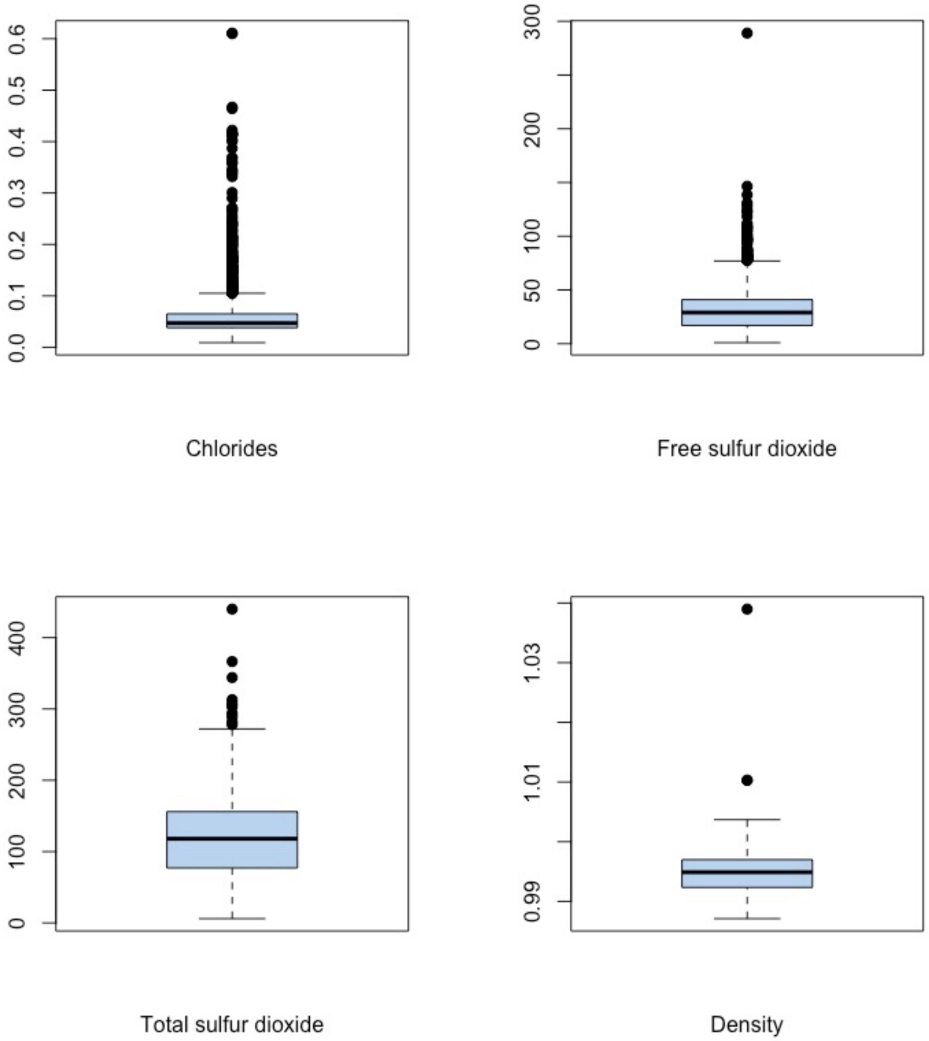
Fixed Acidity

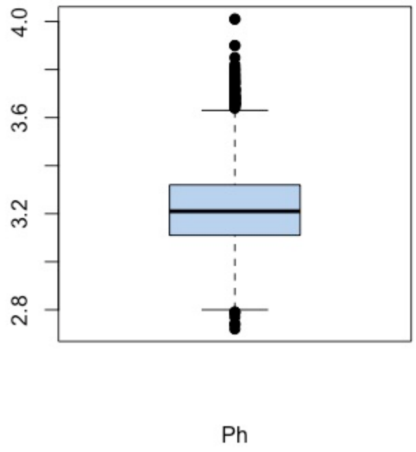


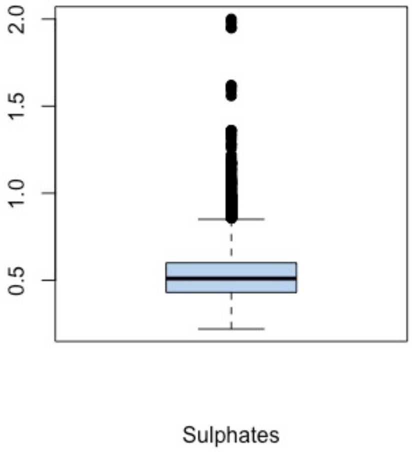
Citric Acid

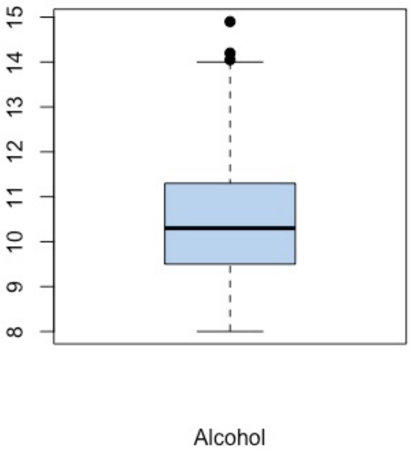


Residual Sugar


0.0
0.1
0.2
0.3
0.4
0.5
0.6
Chlorides
0
50
100
200
300
Free sulfur dioxide
0
100
200
300
400
Total sulfur dioxide
0.99
1.01
1.03
Density







4.4 Observations based on boxplots of variables

1. All variables have outliers.
2. Fixed acidity, volatile acidity and citric acid have outliers. If those outliers are eliminated distribution of the variables may be taken to be symmetric.
3. Residual sugar has a positively skewed distribution; even after eliminating the outliers' distribution will remain skewed.
4. Some of the variables, e.g. free sulphur dioxide, density, have a few outliers but these are very different from the rest.
5. Mostly outliers are on the larger side.
6. Alcohol has an irregular shaped distribution but it does not have pronounced outliers.

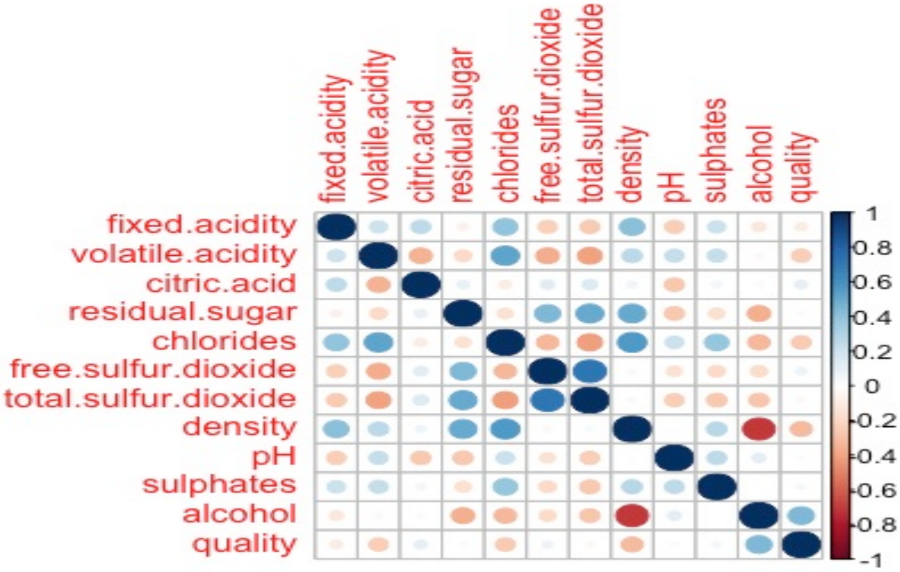
4.5 Outlier treatment

Boxplot method is used to identify and replace outliers with NA's and Mice package is used for imputing the Na's. MICE assumes that the missing data are missing at random (MAR), which means that the probability that a value is missing depends only on observed value and can be predicted using them. It imputes data on a variable by variable basis by specifying an imputation model per variable.

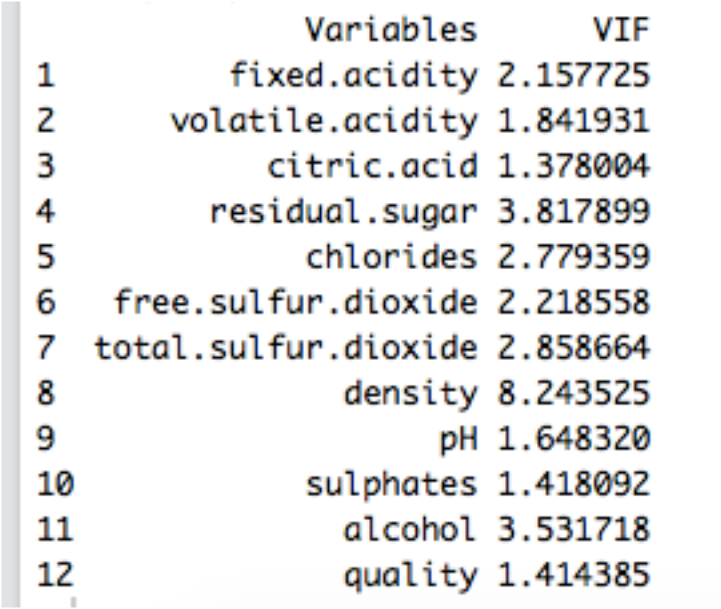
After treating outliers, there are still some outliers for variables like sulphates, citric acid and residual sugar, no further treatment is done at this stage.

4.6 Correlation Matrix

With the help of correlation matrix plot, it's clear that free sulfur dioxide and total sulfur dioxide are highly positively correlated and will not bring significant variance to target variable. The same could be said for alcohol and density variables, which are negatively highly correlated.



below results indicate high VIF for density and it's removed to build regression and other models.



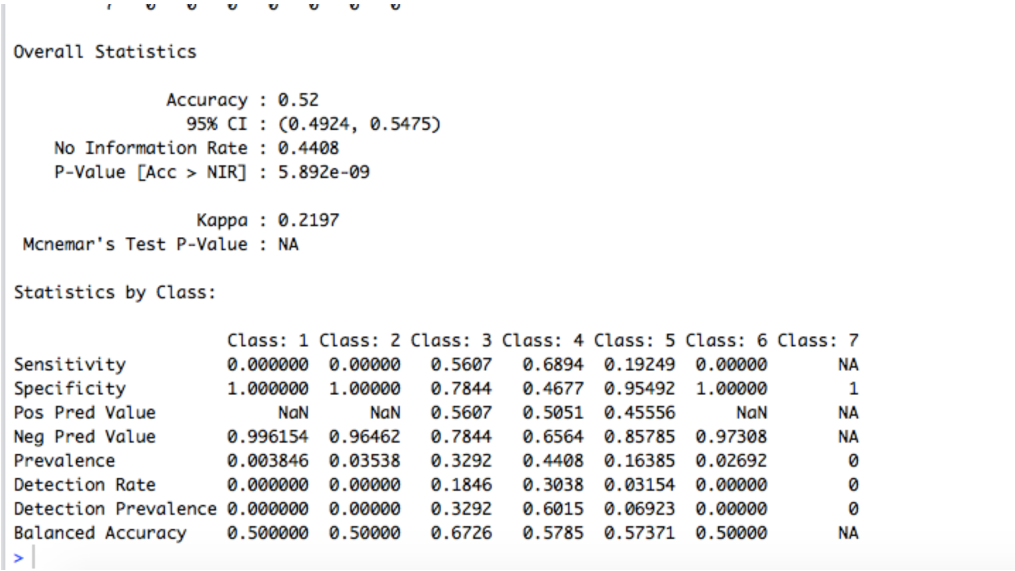
```
              Variables      VIF
1         fixed.acidity 2.157725
2      volatile.acidity 1.841931
3           citric.acid 1.378004
4        residual.sugar 3.817899
5             chlorides 2.779359
6   free.sulfur.dioxide 2.218558
7  total.sulfur.dioxide 2.858664
8               density 8.243525
9                    pH 1.648320
10            sulphates 1.418092
11              alcohol 3.531718
12              quality 1.414385
```

5 Model development

following are the models which will classify the quality of wine depending on multiple factors.

5.1 Model 1 Logistic Regression Model

The data is divided into training and testing set in the proportion of 80:20. Model was built on training data and tested on test data. The model is 52 % accurate. In classification, one of the key attributes that we consider is the test error rate. In this case, since we are not dealing with binary classification, it is not possible to use the ROC curve or AUC as a criterion to assess our performance.

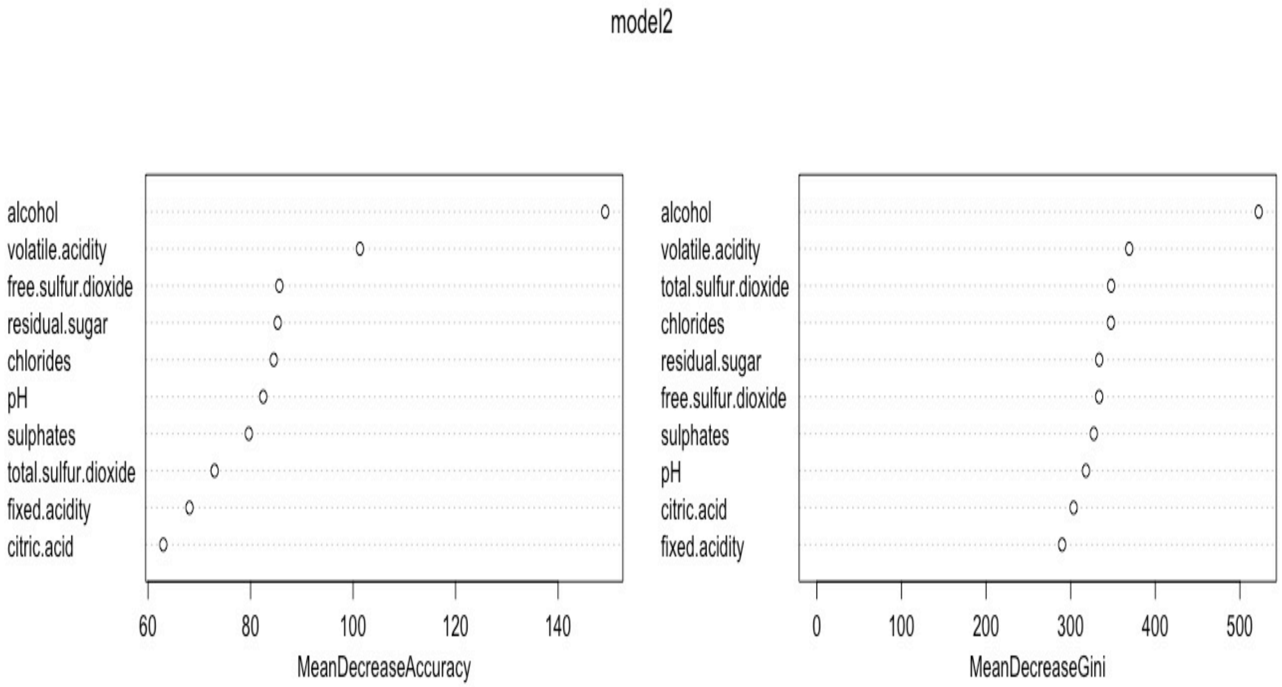


```
         7    0    0    0    0    0    0    0

Overall Statistics

               Accuracy : 0.52
                 95% CI : (0.4924, 0.5475)
    No Information Rate : 0.4408
    P-Value [Acc > NIR] : 5.892e-09

                  Kappa : 0.2197
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7
Sensitivity          0.000000  0.00000   0.5607   0.6894  0.19249  0.00000       NA
Specificity          1.000000  1.00000   0.7844   0.4677  0.95492  1.00000        1
Pos Pred Value            NaN      NaN   0.5607   0.5051  0.45556      NaN       NA
Neg Pred Value       0.996154  0.96462   0.7844   0.6564  0.85785  0.97308       NA
Prevalence           0.003846  0.03538   0.3292   0.4408  0.16385  0.02692        0
Detection Rate       0.000000  0.00000   0.1846   0.3038  0.03154  0.00000        0
Detection Prevalence 0.000000  0.00000   0.3292   0.6015  0.06923  0.00000        0
Balanced Accuracy    0.500000  0.50000   0.6726   0.5785  0.57371  0.50000       NA
>
```

5.2 Model 2 Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5   6   7
         1   0   0   0   0   0   0   0
         2   0   6   0   1   0   0   0
         3   2  29 299  89   5   0   0
         4   3  11 128 456  99  11   0
         5   0   0   1  27 108   8   0
         6   0   0   0   0   1  16   0
         7   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.6808
                 95% CI : (0.6547, 0.7061)
    No Information Rate : 0.4408
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4964
```

```
Mcnemar's Test P-Value : NA
```



Importance of predictors are given in the above dotplot. Randomforest method increased the accuracy significantly to 68%.

5.3 Model 3 KNN

Nearest neighbor classifier is used with three levels (Low, Medium, High) of quality. It turned out that for *k* = 5, test data misclassification rate is lowest, when all predictors are being used.

```
      model3
   3   4   5   6   7   8   9
 3   0   0   0   0   0   0   0
 4   1   1   4   0   0   0   0
 5   3  31 246 131  27   0   1
 6   2   6 148 380 109  13   2
 7   1   3  14  63  87  15   0
 8   0   0   0   5   3   4   0
 9   0   0   0   0   0   0   0
```

```
Overall Statistics

               Accuracy : 0.5523
                 95% CI : (0.5248, 0.5796)
    No Information Rate : 0.4454
    P-Value [Acc > NIR] : 6.883e-15

                  Kappa : 0.3027
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 3  Class: 4 Class: 5 Class: 6 Class: 7 Class: 8 Class: 9
Sensitivity          0.000000 0.0243902   0.5971   0.6563  0.38496 0.125000 0.000000
Specificity          1.000000 0.9960286   0.7827   0.6117  0.91061 0.993691 1.000000
Pos Pred Value            NaN 0.1666667   0.5604   0.5758  0.47541 0.333333      NaN
Neg Pred Value       0.994615 0.9690881   0.8072   0.6891  0.87556 0.978261 0.997692
Prevalence           0.005385 0.0315385   0.3169   0.4454  0.17385 0.024615 0.002308
Detection Rate       0.000000 0.0007692   0.1892   0.2923  0.06692 0.003077 0.000000
Detection Prevalence 0.000000 0.0046154   0.3377   0.5077  0.14077 0.009231 0.000000
Balanced Accuracy    0.500000 0.5102094   0.6899   0.6340  0.64779 0.559345 0.500000
```

5.4 Model 4 Rpart tree based model

Accuracy dropped in tree based model. Model was 54% accurate.

6 Conclusion

It does not look like wine quality is well supported by its chemical properties. At each quality level variability of the predictors is high and the groups are not well separated. It will be very difficult to rely merely on the chemical properties to predict the quality with out the support of wine tasters. A mix of Wine data analysis and wine tasters rating could be an ideal solution.