QUALITY ANALYSIS OF WINE

A CASE STUDY REPORT

Mayank Aggarwal

mayank953@gmail.com

GITHUB

TABLE OF CONTENTS

QUALITY ANALYSIS OF WINE

1. PROBLEM STATEMENT

Build a classification model which will classify the quality of wine depending on multiple factors and chemical composition of it.

2. INTRODUCTION TO PROJECT

A) CASE STUDY EXPLANATION

This report is an analysis on the Quality of Red and White Wine. Each year, thousands of dollars are spent to review the taste and quality of wine, which is known as **"Wine Tasting"**. We have data of Red and White Wine with us containing various chemical composition and physical features. Here, we use the data and build a classification model which classifies the quality based on other independent variables.

The aim of the project is to reduce the resources spent on volunteers for Wine Tasting before the Launch of Wine using the power of Analytics.

B) PAIN AND GAIN ANALYSIS

The Pain and Gain analysis is really a tricky question here. As we are dealing with taste, which is a subjective opinion and not objective. As it is more depend on the person consuming, reaching a fair accuracy is a pain in the neck.
Wine Tasting (human based) gives us the assurance that the quality of wine is meeting right standards. As Experts are involved for Tasting, they can also give good feedback about Aroma, Aftertaste etc.., which are used to correct the wine. But this is time consuming and expensive.

In New York City, the average wine taster makes **$89,000 per year.**

Using Analytics for Wine quality ratings gives us a scientific approach for Tasting. This approach saves us lot of time and resources, it saves money and effort. But using Analytics for Tasting has few shortcomings, we don't get quality feedbacks from experts to improve the wine further as in traditional methods.

C) DOMAIN KNOWLEDGE

Wine is an alcoholic beverage made from fermented grapes. Grapes are fermented without the addition of sugars, acids, enzymes, water, or other nutrients. Yeast consumes sugar in grapes and convert it to ethanol and CO_2. This is the basic outline of wine production whereas there are many other ingredients and methods used in commercial line. For instance, Sulfur dioxide is used to inhibit microbial spoilage (added before fermentation) and Pasteurization used remove microorganisms by heat.
Wines are basically Red and White wines based on fermenting of red and white grapes respectively.

Quality of Wine is based on many things from density to alcohol content. There are few physical parameters and chemical compositions considered as foundation for Quality Wine.

3. EXPLORATORY DATA ANALYSIS

We have 2 Data sets for Red and White wine individually.

We merge 2 data sets into one to build our model. Red wine data set contains 1599 Observations and 12 Variables and White wine data set contains 4898 Observations and 12 Variables each.

A) UNDERSTANDING OF DATA

We merge both red.csv and white.csv data sets and made a master data set named 'wine'. While combining both data sets row-wise, a new variable 'type' is added to specify red and white wine.

Wine Master data set has 6497 Observations and 13 variables. The variables are discussed below:

```
> dim(wine)
[1] 6497   13
```

```
> str(wine)
'data.frame':   6497 obs. of  13 variables:
 $ 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             : Factor w/ 7 levels "3","4","5","6",..: 3 3 3 4 3 3 3 5 5 3 ...
 $ type                : Factor w/ 2 levels "red","white": 1 1 1 1 1 1 1 1 1 1 ...
```

As seen above wine has variables like

Input variables (based on physicochemical tests):

1- fixed acidity (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)
12- type

Output variable (based on sensory data):

13 - quality (score between 0 and 10)

We have 11 numeric variables and 2 categorical variable (both type and quality), out of which, 12 are independent variables and 1 is categorical variable (i.e., quality).

i) UNDERSTANDING ON DOMAIN KNOWLEDGE BASIS:

We discuss each variable here.

1. **Fixed acidity**: It is the non-volatile acid content in the wine. During research, we found that the fixed acidity is combination of tartaric acid, malic acid and citric acid. In our data, we have a variable for citric acid, so given fixed acidity must be combination tartaric acid and malic acid. The usual expected range for **tartaric acid varies from 1,000 to 4,000 mg/L** and **malic acid varies from 0 to 8,000 mg/L.** So, In our Data, **1-12 g/L is acceptable**.

   This is **Numeric Variable** in our data ranging from **3.8-15.9 g/L**, with **7.215** being the mean.
   source: http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity

2. **Volatile acidity**: Volatile acidity refers to the steam distillable acids present in wine, primarily acetic acid but also lactic, formic, butyric, and propionic acids. The average level of acetic acid in a new dry table wine is less than 400 mg/L, though levels may range from undetectable up to 3g/L.
   U.S. legal limits of Volatile Acidity:
   Red Table Wine 1.4 g/L
   White Table Wine 1.2 g/L
   While acetic acid is generally considered a spoilage product, but winemakers seek a lowel detectible level of acetic acid to add to the perceived complexity of a wine.
   The aroma threshold for acetic acid in red wine varies from 600 mg/L and 900 mg/L, depending on the variety and style.
   Our Data contains **Numeric variable**, Volatile acidity as ranging from : **0.0800 to 1.5800 g/L**, with **0.3397** being the mean, but acceptible upper limit is **1.2 g/L** on average.

   Source: **http://extension.psu.edu/food/enology/wine-production/wine-made-easy-fact-sheets/volatile-acidity-in-wine**
   http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity

3. **Citric acid**: Citric acid is a weak organic acid, which is often used as a natural preservative or additive to wine and adds a sour taste and freshness to it. It is used less frequently than tartaric and malic due to the aggressive citric flavors it can add to the wine. In general, **0 to 500 mg/L citric acid is its composition** in wine. Citric acid is often added to wines to increase acidity, complement a specific flavor or prevent ferric hazes. So, it is an important attribute for Quality of wine. In Our Data, Citric acid ranges is a **Numeric variable** ranging from **0 to 1.6 g/L** which is way more than the acceptable limits. **0.3186** is the mean of citric acid.
   Source: **http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity**

4. **Residual Sugar**: Residual sugars are the left over natural sugars present in grapes after fermentation process. It happens due to interruption of fermentation for different reasons. This adds Sweetness to wine. So, this is one of the most important attributes . This varies with wine type ranging from 0.2- 0.3 g/L for Dry wines to 5-15 g/L for Sweet dessert wines. So, the **acceptable range is from 0.2- 15 g/L** . Our Data has a **Numeric variable**, residual sugar varying from **0.6 g/L to 65.8 g/L** with **3** being the median as it is skewed data(see histograms below)
   Source: https://winemakermag.com/501-measuring-residual-sugar-techniques

5. **Chlorides**: Chlorides are part of grape salts. Salty taste is a turn-off for a Wine lover. So, this should be as minimum as possible. Wine contains 2 to 4 g/L of salts of mineral acids. In which, concentration of chlorides vary from country to country with preferences and laws. As per Brazillian law, it can be maximum of 0.2 g/L. However, it varies based on country. For instance, in Australia, the maximum level of chloride allowed is 0.607 g/L. **Maximum limit** for chloride concerntration can be considered as **0.156**(mean of Median percentages of various countries like South Africa, Argentina, Australia, Chile, USA, France).
   In the Data, **Numerical variable** chlorides are ranging from **0.009-0.6110 g/L** with **0.05603** being the mean and with **0.047** being the median.
   Source: http://www.scielo.br/pdf/cta/v35n1/0101-2061-cta-35-1-95.pdf

6. **Free Sulfur Dioxide**: Sulphur dioxide (SO2) is used as an antioxidant and preservative and has become widely used in winemaking. 'Free' SO2 is that which is unbound to compounds in the wine and is therefore able to exert an antioxidant/preservative action.
   Free SO2 in wine ranges from about **40% to 75% of the total SO2**, so that putting maximum limit of **112.5 mg/L for red wine and 150 mg/L for white wine**. The 40% level for wines that are turbid or sweet and the 75% level for clean, dry wines. Wine makers try their best to increase Free SO2 to 50% of Total SO2.
   Our Data contains, **Numeric variable**, free SO2 ranging from **1 to 289 mg/L** with **29** being the median and **30.53** as mean.
   Source: https://www.practicalwinery.com/janfeb09/page5.htm
   https://www.campdenbri.co.uk/services/free-sulphur-dioxide.php

7. **Total Sulfur Dioxide**: As there are Free SO2 and 'Bound' SO2 is that which has already been complexed to other compounds in the wine (such as sugars) and has essentially been quenched such that it no longer has antioxidant/preservative activity. Total SO2, is the sum of both of these forms. Too much Sulfur Dioxide cause health implications like allergic reactions to significant hangovers(considered dangerous for asthmatics).
   The EU has set a legal limit for total SO2 of 150 mg/litre in red wines and 200 mg/litre in white wines. Making an average limit as **175 mg/L** of Total SO2.
   The Data contains **Numerical variable** representing Total SO2 composition varying from **6- 440 mg/L** with **115.7** being the mean and **118** being median.
   Source: http://www.morethanorganic.com/sulphur-in-the-bottle

8. **Density**: Density of wine is one of more important attribute to take care while wine making. Wine go through methods like filtering to take out the wine fine. This is also important in giving “finish” (aftertaste) experience for the wine lover. It is important to remember that dissolved sugars increase density where an increase in alcohol decreases density.
   We have **Numeric variable** Density in our data, ranging from **0.9871 to 1.0390 g/cm^3**, with **0.9947** as mean.
   Source: http://www.grapeheaven.com/learn/Wine-Quality-Control-Mechanisms-i156.aspx

9. **pH**: 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. The pH of a wine is critical not only to its flavor but to nearly every aspect of the wine. According to wine maker Alison Crowe of Winemaker Magazine “**pH is the backbone of a wine**”. Difference in pH can make your wine the best or break it to worst.
   In Technology of Winemaking the following pH ranges are recommended for wine musts:
   White Wines < 3.3
   Red Wines < 3.4

Our Data has a **Numeric variable**, **pH** ranging from **2.720-4.10** with **3.219** as mean.
Source: http://winemakersacademy.com/importance-ph-wine-making/

10. **Sulphates**: This is a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant. They are added to wine due to $CuSO_4$ addition.
    The Data contains **Numeric variable Sulphates** ranging from **0.22- 2.0 g/L** with **0.5313** as mean.
    Source: http://www.wineadds.com/faq/copper

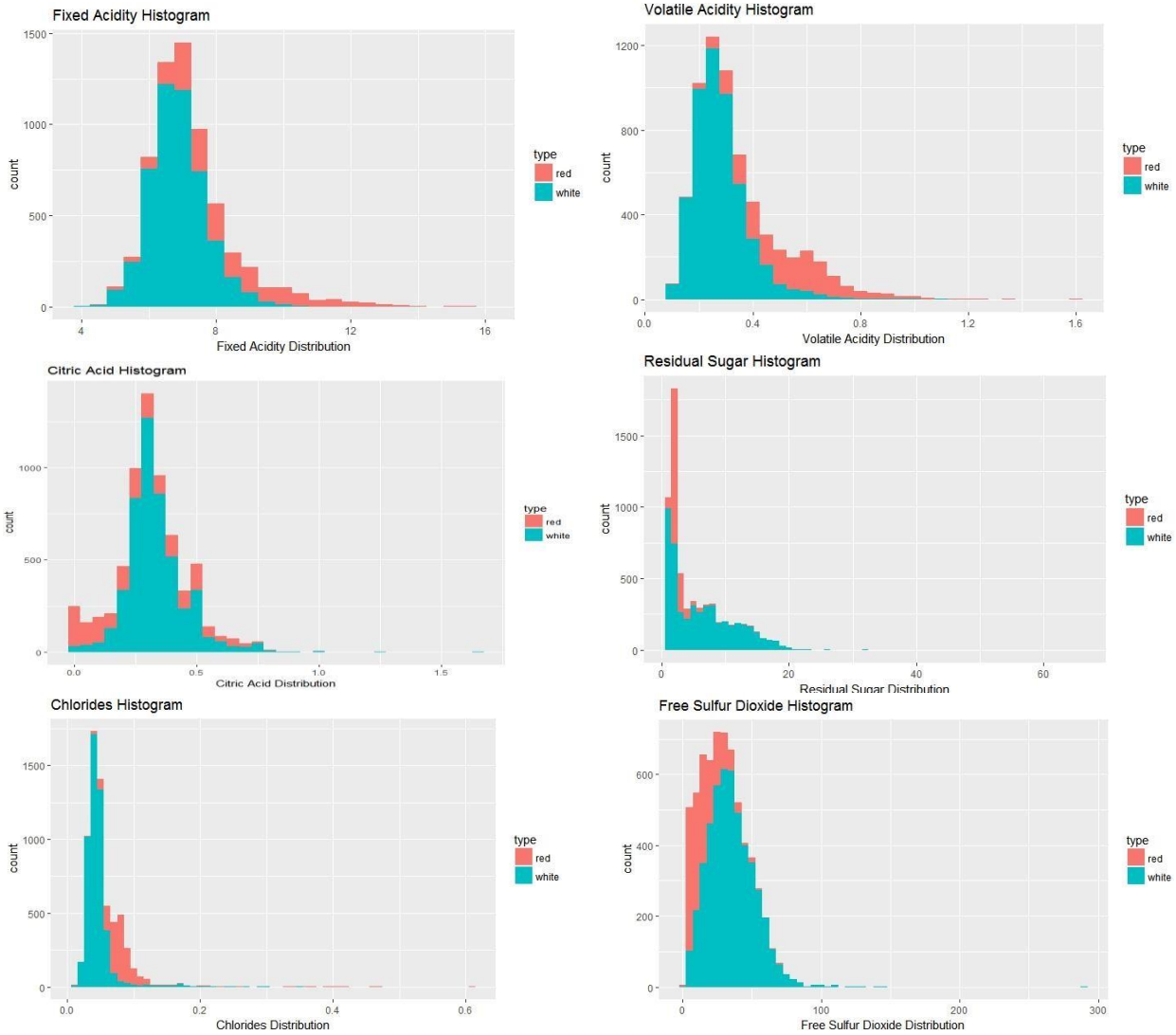
11. **Alcohol**: Alcohol is produced in the process of fermentation of sugars by yeast. Technically, the name is ethyl alcohol. Different concentrations of alcohol in the human body have different effects on a person. Alcohol content in wine is usually **9%–16%** (most often 12.5%–14.5%). More than this is considered as dangerous.
    Drinking enough to cause a blood alcohol concentration (BAC) of 0.03%-0.12% typically causes an overall improvement in mood and possible euphoria, increased self-confidence and sociability, decreased anxiety. Whereas, BAC from 0.35% to 0.80% causes a coma (unconsciousness), life-threatening respiratory depression and possibly fatal alcohol poisoning.

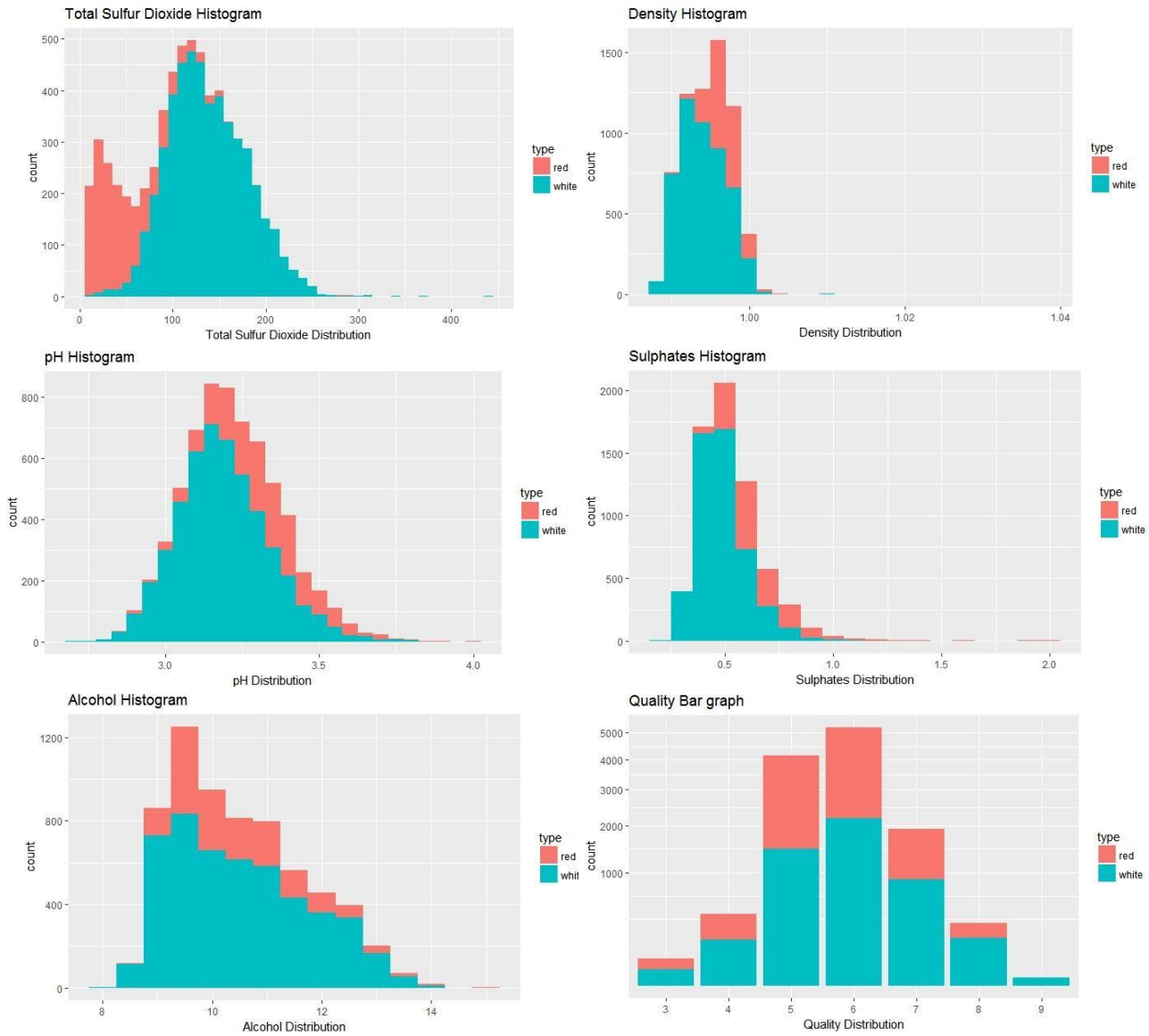
    In our data, Alcohol content is a **Numeric variable** ranging from **8.0- 14.90** (% by volume) with **10.30** being median value.

    Source: https://en.wikipedia.org/wiki/Alcohol_by_volume
    https://en.wikipedia.org/wiki/Wine
12. **Quality**: Quality variable is our dependent/target variable. After Wine Tasting, the experts give the ratings for the wine. This is determining how good the wine is, in terms of Taste, Texture, Feel, Aftertaste and Aroma.
    Our data contains quality variable has levels: **3, 4, 5, 6, 7, 8, 9** and it is a categorical variable.
13. **Type**: Type variable helps us determine whether it is red wine or white wine. It is added by us while analysis.
    Type is categorical variables with levels “Red” and “White”.

ii) UNDERSTANDING ON DATA BASIS:

To explain the Distribution of each variable, we are using Histogram and bar graphical representation:





KNOWN INSIGHTS OR FIRST LEVEL OF INSIGHTS:

1. Data distribution is not normal for all the variables. They all are mostly positive skewed variables.
2. Acidity(both fixed and volatile acidity combined) for Red wine is more than White wine.
3. Only pH variable and Total sulfur dioxide for white wine are normally distributed.
4. Residual sugar and chlorides are the most postive skewed variable in our data.
5. Free Sulfur dioxide and Total Sulfur dioxide shows drastic differences in data between Red and White wine. White being more Sulfur dioxide content than Red wine.
6. From Density Histogram, we can infere Red being more denser than White wines, but most dense wine being a white wine.
7. Among our variables, pH, Fixed acidity and Quality(target) variables are the only Normally distributed data with perfect bell shape. There must be some important correlation between them which is a good news.
8. Quality variable has highest value at 5 and 6 for both red and white wines. White wines are more high rated Red wines on every rating level.

B) Preprocessing and Cleaning

A) MISSING VALUE ANALYSIS

Our Data contains no missing values in it.

B) OUTLIER ANALYSIS

Our Data has 11 independent variables. Let us see

```
> summary(wine)
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900
 1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800
 Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700
 Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603
 3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500
 Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100

 free.sulfur.dioxide total.sulfur.dioxide    density             pH
 Min.   :  1.00      Min.   :  6.0        Min.   :0.9871   Min.   :2.720
 1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110
 Median : 29.00      Median :118.0        Median :0.9949   Median :3.210
 Mean   : 30.53      Mean   :115.7        Mean   :0.9947   Mean   :3.219
 3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320
 Max.   :289.00      Max.   :440.0        Max.   :1.0390   Max.   :4.010

   sulphates         alcohol          type       quality
 Min.   :0.2200   Min.   : 8.00   red  :1599   3:  30
 1st Qu.:0.4300   1st Qu.: 9.50   white:4898   4: 216
 Median :0.5100   Median :10.30                5:2138
 Mean   :0.5313   Mean   :10.49                6:2836
 3rd Qu.:0.6000   3rd Qu.:11.30                7:1079
 Max.   :2.0000   Max.   :14.90                8: 193
                                               9:   5
```

By analyzing the summary of wine, we can say there seems to be lot of outliers in our data set. Mainly variables like fixed acidity, volatile acidity and citric acid have lot of outliers as there is significant distance between Q_3 and Maximum value. Ranges of each variable in our wine data set are given below:

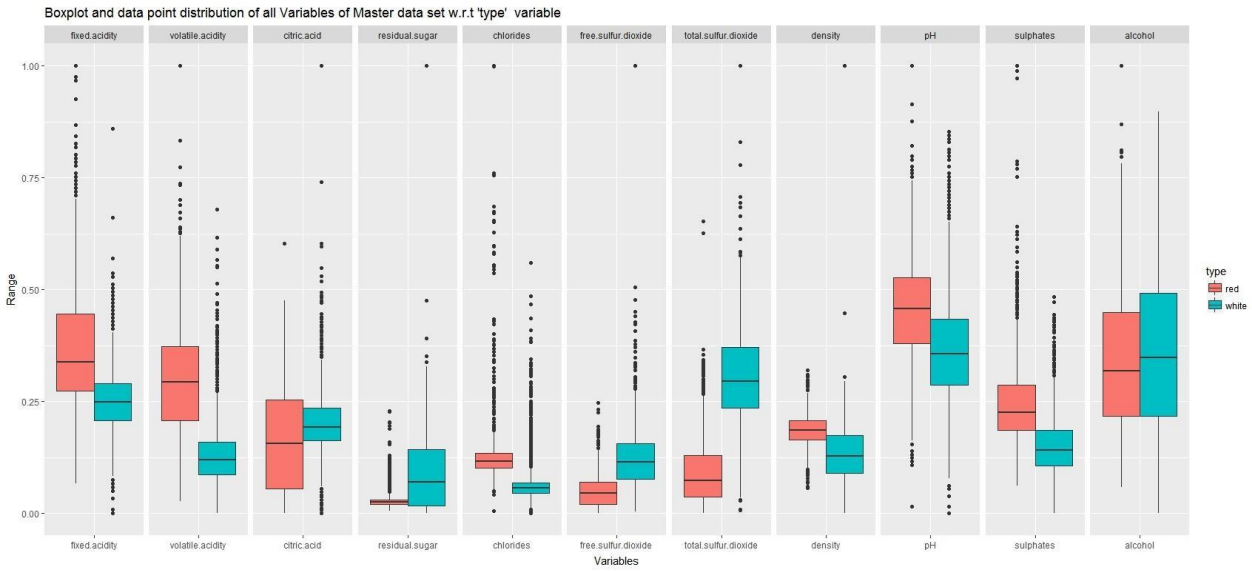
Independent variables or predictors

Fixed acidity	: 3.8000- 15.900
Volatile acidity	: 0.0800- 1.5800
Citric acid	: 0.0000- 1.6600
Residual acid	: 0.6000- 65.800
Chlorides	: 0.0090- 0.6110
Free Sulfur dioxide	: 1.0000- 289.00
Total Sulfur dioxide	: 6.0000- 440.00
Density	: 0.9871- 1.0390
pH	: 2.7200- 4.0100
Sulphates	: 0.2200- 2.0000
Alcohol	: 8.0000- 14.900
Type	: “Red” and “White”

Target Variable

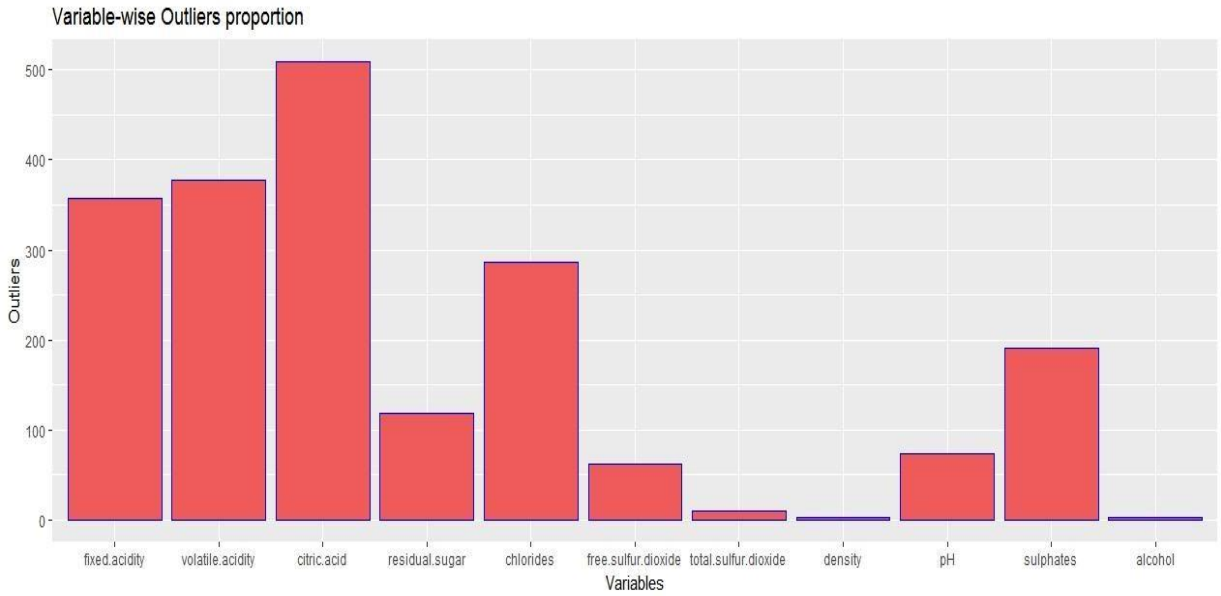
Quality (levels)	: 3, 4, 5, 6, 7, 8, and 9.

Let us Analyze Outliers using Boxplot. We normalize Data to bring variables into same range. Boxplot is given below.



A Data frame 'wine_out' is created to explain the proportions of outliers by each variable.

```
> wine_out
                Variables Outliers Percentage
1          fixed.acidity      357       5.49
2       volatile.acidity      377       5.80
3            citric.acid      509       7.83
4         residual.sugar      118       1.82
5              chlorides      286       4.40
6    free.sulfur.dioxide       62       0.95
7   total.sulfur.dioxide       10       0.15
8                density        3       0.05
9                     pH       73       1.12
10             sulphates      191       2.94
11               alcohol        3       0.05
>
```



As we can see, outliers are at most less than 7% of a variable. We decide to impute outliers.

First we will replace outliers with NAs, thereafter we will impute them.
Here, we have to consider domain knowledge and use it carefully. We will replace only those outliers who are not within normal ranges of wine composition. So, not all outliers are replaced with NAs.
Only **816** NAs are introduced among initial **1989** Outliers.

Mean and Median methods are checked for imputation but they are proved to be futile. KNN Imputation is much more appropriate for the data. '**wine_imputed**' is the new data set after imputing.

C) NORMALIZATION AND STANDARDIZATION

Among 11 Numeric variables we have, only 2 variables (pH and fixed acidity) are normally distributed and all others are mostly positively skewed. So, we use Normalization on Data as Standardization requires an assumption of normally distributed variables.
We use 'clustersim' package for normalizing data and create '**wine_norm**'. Quality and Type variables are removed and then added as they are categorical variables.

D) VARIABLE IMPORTANCE AND FEATURE ENGINEERING

Our Master data set has 6497 Observations and 13 Variables including dependent variable (quality) and variables (type) added during analysis. Feature selection or Variable importance is a crucial step in building our model. **Not all variables carry equal information to explain our Target variable**. So, we go through feature engineering and take those variables which explains Target variable's variance more clearly.

Here we have analyzed variable importance using correlation matrix, plots using random forest, corrplot& ggcorrplot and Variance Inflation Factor (VIF) .

i) CORRELATION MATRIX:

We are building correlation matrix using 'spearman' technique. We are specifically using 'spearman ' other than 'pearson' because 'spearman' correlation technique doesn't require any assumptions like normally distributed data which suits our condition.

The 'spearman' correlation matrix is shown below:

```
> cor(wine_norm, use = "complete.obs", method= "spearman")
              f.acidity   v.acidity citric.acid residual.sugar   chlorides     free.SO2
f.acidity      1.00000000  0.20028066  0.26861452    -0.03159386  0.36501073 -0.260131744
v.acidity      0.20028066  1.00000000 -0.30668996    -0.06593183  0.41923405 -0.366101849
citric.acid    0.26861452 -0.30668996  1.00000000     0.06829653 -0.09333414  0.127457738
residual.sugar -0.03159386 -0.06593183  0.06829653     1.00000000 -0.03698971  0.389309769
chlorides      0.36501073  0.41923405 -0.09333414    -0.03698971  1.00000000 -0.269594925
free.SO2      -0.26013174 -0.36610185  0.12745774     0.38930977 -0.26959493  1.000000000
Total.SO2     -0.23473467 -0.34422203  0.15595672     0.45514359 -0.27889390  0.740409478
density        0.43401999  0.26116839  0.05177960     0.52536845  0.59720490  0.006624505
pH            -0.24311239  0.19793490 -0.27473453    -0.22932764  0.17625895 -0.167386916
sulphates      0.21546589  0.25152172  0.02762634    -0.13135577  0.37288312 -0.215460162
alcohol       -0.11071588 -0.02392883  0.03400209    -0.33050063 -0.39460123 -0.186362386
              Total.SO2      density           pH     sulphates      alcohol
f.acidity     -0.23473467  0.434019992 -0.24311239  0.215465892 -0.110715881
v.acidity     -0.34422203  0.261168393  0.19793490  0.251521724 -0.023928832
citric.acid    0.15595672  0.051779597 -0.27473453  0.027626340  0.034002093
residual.sugar 0.45514359  0.525368451 -0.22932764 -0.131355771 -0.330500627
chlorides     -0.27889390  0.597204900  0.17625895  0.372883117 -0.394601234
free.SO2       0.74040948  0.006624505 -0.16738692 -0.215460162 -0.186362386
Total.SO2      1.00000000  0.062390565 -0.24597024 -0.252930560 -0.310751604
density        0.06239056  1.000000000  0.01258605  0.274795937 -0.699674983
pH            -0.24597024  0.012586054  1.00000000  0.266600396  0.139359788
sulphates     -0.25293056  0.274795937  0.26660040  1.000000000  0.005974672
alcohol       -0.31075160 -0.699674983  0.13935979  0.005974672  1.000000000
>
```

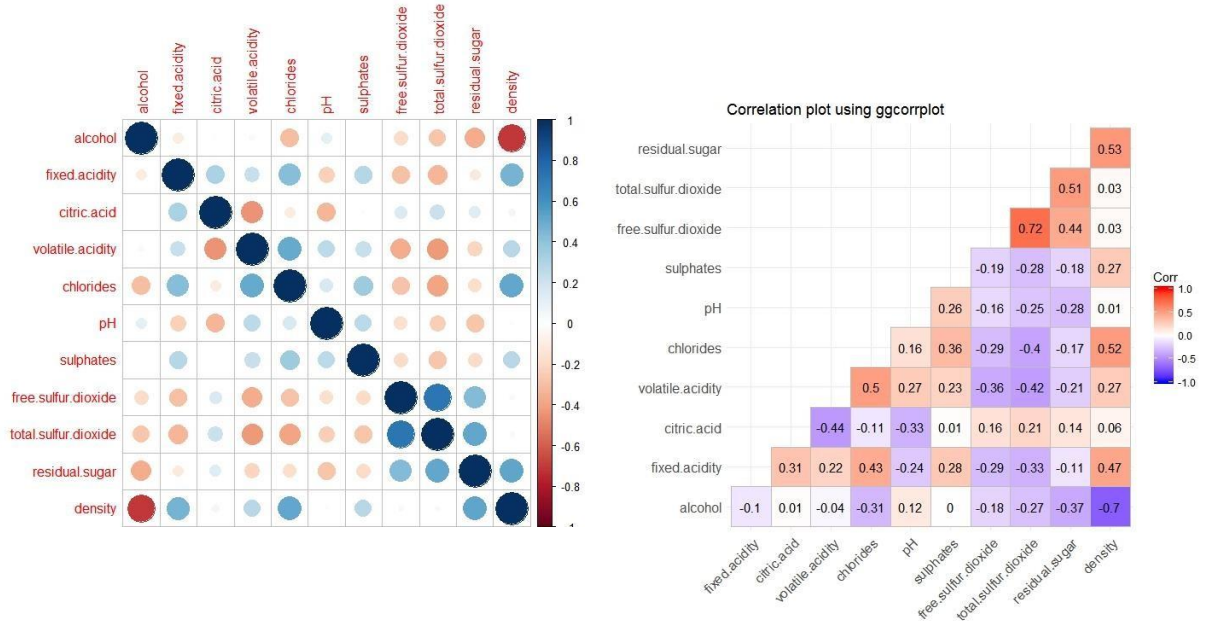
By analysing at the above Correlation matrix, we can come to know that there is clear correlation between Total Sulfur dioxide and Free Sulfur dioxide; and Alcohol and Density variables.

Let us see graphical representations and other check our **hypothesis** for correlations.

ii) CORRELATION PLOTS:

Two kinds of correlation plots are implemented here. We use basic 'corrplot' package to represent basic **Hclust** correlation plot and 'ggcorrplot' package to represent more detailed correlation plot.

Graphical representations are shown below:



By understanding above correlation plots, we can get evidences to support our hypothesis about our correlations.

1. There is high correlation between Alcohol- Density and Total Sulfur dioxide- Free Sulfur dioxide pairs.
2. Alcohol -Density is negatively correlated pairwith 70% correlation and Total and Free Sulfur dioxide is positively correlated with 72% correlation.

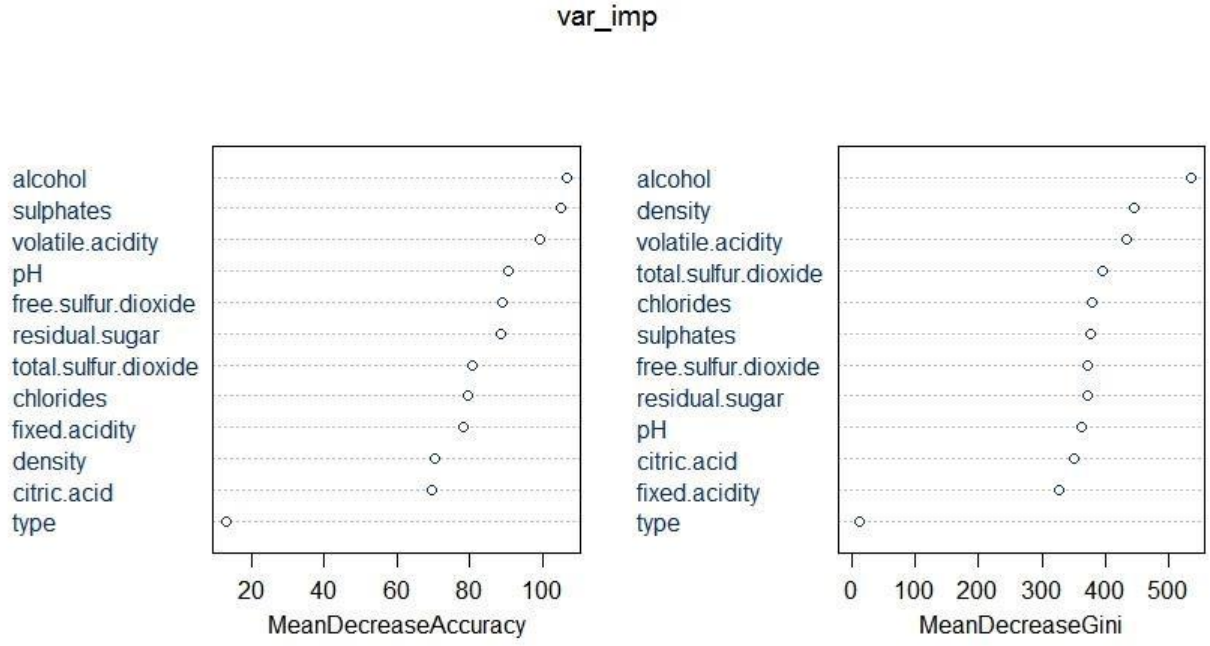
iii) VARIANCE INFLATION FACTOR (VIF): Variance inflation factor is checked for ournormalized wine_norm data and results are shown below:

```
> vif(wine_norm[-c(12,13)])
             Variables       VIF
1        fixed.acidity  4.252169
2     volatile.acidity  2.040306
3          citric.acid  1.708796
4       residual.sugar  6.540126
5            chlorides  2.434762
6  free.sulfur.dioxide  2.173823
7 total.sulfur.dioxide  2.969803
8              density 15.342545
9                   pH  2.218627
10          sulphates  1.482605
11            alcohol  4.543896
>
```

iv) VARIABLE IMPORTANCE PLOT USING RANDOM FOREST:

Variable importance is analyzed for our normalized wine_norm data set using random forest. Variable importance parameter is assigned to **var_imp**.

We analyze both **Mean Decrease In Accuracy** and **Mean Decrese In Gini** using **varImpPlot()** function.



VarImpPlot() graph explains us two things. It presents how much a variable can explain the target variable's variance individually using Mean Decrese Accuracy and how much variance will be affected if a variable is taken out from the analysis.

From the above plot, it is evident that **type is the least significant variable** in our data. Also, **citric acid and fixed acidity are contributing less significant** role in explaining Target variable's variance.
As we have two correlation pairs from our later analysis: i.e., alchol and density; total and free sulfur

dioxide, we can deduce now that, density is less significant than alcohol and free sulfur dioxide is less significant than total sulfur dioxide(using MeanDecreaseGini).

HIDDEN INSIGHTS OR SECOND LEVEL OF INSIGHTS:

1. Our data has two pairs of correlated varaibles.
2. **Type** is the least significant variable and it is followed by **Citric acid** and **Fixed acidity**.
3. Out of first correlated pair, **Density is the less significant then Alcohol**.
4. Out of second correlated pair, **Free Sulfur Dioxide is less significant than Total Sulfur Dioxide.**
5. As correlation between the variables in our pairs are not very high (threshold=80%), we train our models with and without the correlated pairs.
6. **Density** is least significant variable while building model as Variance inflation factor ismore than 15.

C) MODEL BUILDING

We are building our model using Statistical Methods and Machine learning algorithms for this analysis. Logistic regression for ordinal variables, Decision tree algorithm, Random forest and KNN classification algorithms are implemented.

A) LOGISTIC REGRESSION MODEL

Logistic regression is our **Base Model**. Logistic regression model is build using 'MASS' package and plor() function. Plor() function is used as our target categorical variable is ordinal data type.

Here we are using 2 methods namely 'logistic' and 'probit' to build our Logistic regression.

i) Logistic method:

We get **0.5477** accuracy with all variables and highest **0.55** accuracy with removing the least significant variables type, density and fixed acidity in both normal and normalized data sets i.e., **wine_imputed and wine_norm**. Confusion matrix for 0.55 accuracy is given below.

```
Confusion Matrix and Statistics

          
pred        3    4    5    6    7    8    9
  3         0    0    0    0    0    0    0
  4         0    0    0    0    0    0    0
  5         3   23  258  114   14    5    0
  6         3   20  168  420  165   22    1
  7         0    0    2   32   37   12    0
  8         0    0    0    1    0    0    0
  9         0    0    0    0    0    0    0

Overall Statistics

               Accuracy : 0.55
                 95% CI : (0.5225, 0.5773)
    No Information Rate : 0.4362
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2691
 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  0.00000   0.6028   0.7407  0.17130 0.0000000 0.0000000
Specificity          1.000000  1.00000   0.8177   0.4829  0.95756 0.9992070 1.0000000
Pos Pred Value             NaN      NaN   0.6187   0.5257  0.44578 0.0000000       NaN
Neg Pred Value       0.995385  0.96692   0.8075   0.7066  0.85292 0.9699769 0.9992308
Prevalence           0.004615  0.03308   0.3292   0.4362  0.16615 0.0300000 0.0007692
Detection Rate       0.000000  0.00000   0.1985   0.3231  0.02846 0.0000000 0.0000000
Detection Prevalence 0.000000  0.00000   0.3208   0.6146  0.06385 0.0007692 0.0000000
Balanced Accuracy    0.500000  0.50000   0.7102   0.6118  0.56443 0.4996035 0.5000000
```

ii) Probit method:

We get **0.5485** accuracy with all variables and highest **0.55** accuracy with removing the least significant variables type, density, fixed acidity and citric acid in both normal and normalized data sets i.e., **wine_imputed and wine_norm** and respective confusion matrix is shown below.

```
Confusion Matrix and Statistics


pred    3    4    5    6    7    8    9
   3    0    0    0    0    0    0    0
   4    0    0    0    0    0    0    0
   5    3   22  244  106   11    2    0
   6    3   21  184  443  179   29    1
   7    0    0    0   18   26    8    0
   8    0    0    0    0    0    0    0
   9    0    0    0    0    0    0    0

Overall Statistics

               Accuracy : 0.5485
                 95% CI : (0.5209, 0.5758)
    No Information Rate : 0.4362
    P-Value [Acc > NIR] : 2.878e-16

                  Kappa : 0.2556
 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  0.00000   0.5701   0.7813   0.1204     0.00 0.0000000
Specificity          1.000000  1.00000   0.8349   0.4311   0.9760     1.00 1.0000000
Pos Pred Value            NaN      NaN   0.6289   0.5151   0.5000      NaN       NaN
Neg Pred Value       0.995385  0.96692   0.7982   0.7182   0.8478     0.97 0.9992308
Prevalence           0.004615  0.03308   0.3292   0.4362   0.1662     0.03 0.0007692
Detection Rate       0.000000  0.00000   0.1877   0.3408   0.0200     0.00 0.0000000
Detection Prevalence 0.000000  0.00000   0.2985   0.6615   0.0400     0.00 0.0000000
Balanced Accuracy    0.500000  0.50000   0.7025   0.6062   0.5482     0.50 0.5000000
```

B) DECISION TREE MODEL

We have implemented decision trees with its default values on our training and test data sets. The highest accuracy that we get is **0.5323**, confusion matrix is given below.

```
Confusion Matrix and Statistics


pred   3   4   5   6   7   8   9
   3   0   0   0   0   0   0   0
   4   0   0   0   0   0   0   0
   5   3  24 274 149  13   0   0
   6   3  19 154 418 203  39   1
   7   0   0   0   0   0   0   0
   8   0   0   0   0   0   0   0
   9   0   0   0   0   0   0   0

Overall Statistics

              Accuracy : 0.5323
                95% CI : (0.5048, 0.5597)
    No Information Rate : 0.4362
    P-Value [Acc > NIR] : 2.133e-12

                 Kappa : 0.223
 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  0.00000   0.6402   0.7372   0.0000     0.00 0.0000000
Specificity          1.000000  1.00000   0.7833   0.4284   1.0000     1.00 1.0000000
Pos Pred Value            NaN      NaN   0.5918   0.4994      NaN      NaN       NaN
Neg Pred Value       0.995385  0.96692   0.8160   0.6782   0.8338     0.97 0.9992308
Prevalence           0.004615  0.03308   0.3292   0.4362   0.1662     0.03 0.0007692
Detection Rate       0.000000  0.00000   0.2108   0.3215   0.0000     0.00 0.0000000
Detection Prevalence 0.000000  0.00000   0.3562   0.6438   0.0000     0.00 0.0000000
Balanced Accuracy    0.500000  0.50000   0.7117   0.5828   0.5000     0.50 0.5000000
```

The default tree model is not influencing with or without correlation pairs. So, we chose to control the decision tree parametric values like minimum splits.

When Minimum splits are kept at 2 and , there is an increase in accuracy. With all the variables, we have achieved 0.5946 and with cutting off variables like type and fixed acidity we reached our highest accuracy with decision trees. i.e., 0.6962 with normalization and 0.6962 without normalization.

Confusion matrix for **0.6962** accuracy is given below:

```
Confusion Matrix and Statistics


pred   3   4   5   6   7   8   9
   3   0   0   3   0   0   0   0
   4   3  10   7  10   5   0   0
   5   1  16 296 111  13   2   0
   6   2  11  97 358  66  11   1
   7   0   6  21  76 119   7   0
   8   0   0   3  12  13  19   0
   9   0   0   1   0   0   0   0

Overall Statistics

               Accuracy : 0.6169
                 95% CI : (0.5899, 0.6434)
    No Information Rate : 0.4362
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.432
 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 0.232558   0.6916   0.6314  0.55093  0.48718 0.0000000
Specificity          0.997682 0.980111   0.8360   0.7435  0.89852  0.97780 0.9992302
Pos Pred Value       0.000000 0.285714   0.6743   0.6557  0.51965  0.40426 0.0000000
Neg Pred Value       0.995374 0.973913   0.8467   0.7228  0.90943  0.98404 0.9992302
Prevalence           0.004615 0.033077   0.3292   0.4362  0.16615  0.03000 0.0007692
Detection Rate       0.000000 0.007692   0.2277   0.2754  0.09154  0.01462 0.0000000
Detection Prevalence 0.002308 0.026923   0.3377   0.4200  0.17615  0.03615 0.0007692
Balanced Accuracy    0.498841 0.606335   0.7638   0.6875  0.72472  0.73249 0.4996151
```

C) RANDOM FOREST MODEL

Coming to our Ensemble model, we use Random forest 500 trees. Using ntree value other than 500 trees is resulting less efficient model than 500 trees.

For Normal data wine_imputed: With all variables and using 500 trees, our random forest model gave an accuracy of **0.6992** which is quite higher than the prediction of regression and decision tree models.

When feature selection is include and with no type and density variables, we get a highest of **0.7015** accuracy.

Both confusion matrices are given below respectively.

```
Confusion Matrix and Statistics


pred   3   4   5   6   7   8   9
   3   0   0   0   0   0   0   0
   4   1   6   0   0   0   0   0
   5   2  22 322  86   6   0   0
   6   3  15 103 453  98  12   1
   7   0   0   3  28 110   9   0
   8   0   0   0   0   2  18   0
   9   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.6992
                 95% CI : (0.6735, 0.7241)
    No Information Rate : 0.4362
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5296
 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 0.139535   0.7523   0.7989  0.50926  0.46154 0.0000000
Specificity          1.000000 0.999204   0.8670   0.6835  0.96310  0.99841 1.0000000
Pos Pred Value            NaN 0.857143   0.7352   0.6613  0.73333  0.90000       NaN
Neg Pred Value       0.995385 0.971384   0.8770   0.8146  0.90783  0.98359 0.9992308
Prevalence           0.004615 0.033077   0.3292   0.4362  0.16615  0.03000 0.0007692
Detection Rate       0.000000 0.004615   0.2477   0.3485  0.08462  0.01385 0.0000000
Detection Prevalence 0.000000 0.005385   0.3369   0.5269  0.11538  0.01538 0.0000000
Balanced Accuracy    0.500000 0.569370   0.8097   0.7412  0.73618  0.72998 0.5000000
```

Above is the confusion matrix of random forest of normal data with all variables

```
Confusion Matrix and Statistics


pred   3   4   5   6   7   8   9
   3   0   0   0   0   0   0   0
   4   1   6   0   1   0   0   0
   5   3  25 326  87   4   0   0
   6   2  12  98 451  99  12   1
   7   0   0   4  28 111   9   0
   8   0   0   0   0   2  18   0
   9   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.7015
                 95% CI : (0.6758, 0.7263)
    No Information Rate : 0.4362
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5342
 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 0.139535   0.7617   0.7954  0.51389  0.46154 0.0000000
Specificity          1.000000 0.998409   0.8635   0.6944  0.96218  0.99841 1.0000000
Pos Pred Value            NaN 0.750000   0.7326   0.6681  0.73026  0.90000       NaN
Neg Pred Value       0.995385 0.971362   0.8807   0.8144  0.90854  0.98359 0.9992308
Prevalence           0.004615 0.033077   0.3292   0.4362  0.16615  0.03000 0.0007692
Detection Rate       0.000000 0.004615   0.2508   0.3469  0.08538  0.01385 0.0000000
Detection Prevalence 0.000000 0.006154   0.3423   0.5192  0.11692  0.01538 0.0000000
Balanced Accuracy    0.500000 0.568972   0.8126   0.7449  0.73803  0.72998 0.5000000
```

Above is the confusion matrix of random forest of normal data with no type and density variables.

For Normalized data wine_norm: With all variables and using 500 trees, our random forest model gave an accuracy of **0.7008** which is quite higher than the prediction of regression and decision tree models.
When feature selection is include and with no type and density variables, we get a highest of **0.7015** accuracy.
Both confusion matrices are given below respectively.

```
Confusion Matrix and Statistics


pred   3   4   5   6   7   8   9
   3   0   0   0   0   0   0   0
   4   1   6   0   1   0   0   0
   5   3  23 322  82   8   0   0
   6   2  14 103 455  97  12   1
   7   0   0   3  29 109   8   0
   8   0   0   0   0   2  19   0
   9   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.7008
                 95% CI : (0.6751, 0.7256)
    No Information Rate : 0.4362
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5323
 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 0.139535   0.7523   0.8025  0.50463  0.48718 0.0000000
Specificity          1.000000 0.998409   0.8670   0.6876  0.96310  0.99841 1.0000000
Pos Pred Value            NaN 0.750000   0.7352   0.6652  0.73154  0.90476       NaN
Neg Pred Value       0.995385 0.971362   0.8770   0.8182  0.90704  0.98436 0.9992308
Prevalence           0.004615 0.033077   0.3292   0.4362  0.16615  0.03000 0.0007692
Detection Rate       0.000000 0.004615   0.2477   0.3500  0.08385  0.01462 0.0000000
Detection Prevalence 0.000000 0.006154   0.3369   0.5262  0.11462  0.01615 0.0000000
Balanced Accuracy    0.500000 0.568972   0.8097   0.7450  0.73386  0.74280 0.5000000
```

Above figure is a confusion matrix for random forest with all variables of normalized wine_norm data

```
Confusion Matrix and Statistics


pred   3   4   5   6   7   8   9
   3   0   0   0   0   0   0   0
   4   1   6   0   1   0   0   0
   5   3  25 327  85   7   0   0
   6   2  12  95 455  94  11   1
   7   0   0   6  26 113  11   0
   8   0   0   0   0   2  17   0
   9   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.7062
                 95% CI : (0.6806, 0.7308)
    No Information Rate : 0.4362
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5419
 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 0.139535   0.7640   0.8025  0.52315  0.43590 0.0000000
Specificity          1.000000 0.998409   0.8624   0.7067  0.96033  0.99841 1.0000000
Pos Pred Value            NaN 0.750000   0.7315   0.6791  0.72436  0.89474       NaN
Neg Pred Value       0.995385 0.971362   0.8816   0.8222  0.90997  0.98283 0.9992308
Prevalence           0.004615 0.033077   0.3292   0.4362  0.16615  0.03000 0.0007692
Detection Rate       0.000000 0.004615   0.2515   0.3500  0.08692  0.01308 0.0000000
Detection Prevalence 0.000000 0.006154   0.3438   0.5154  0.12000  0.01462 0.0000000
Balanced Accuracy    0.500000 0.568972   0.8132   0.7546  0.74174  0.71716 0.5000000
```

This is confusion matrix for random forest of normalized wine_norm data with no type and density is presented above.

ERROR METRIC

Wine which is **good rated but predicted as bad doesn't do any harm** when it reached customer or when it is further refined. But Wine that which is **bad or low quality but predicted as good can impact product's goodwill**. So, error metric chosen here is **False Positive Rate**.

So, the model which gives **least False Positive rate and high Accuracy** will be frozen for Deployment.

D) CONCLUSION

Quality of wine is a **subjective opinion which varies from person to person** depending on their preferences of **Taste & Beauty**. It also depends on many influencing factors like time of Wine Tasting, environmental conditions and many other factors which are humanly not controlled. Yet, this is analysis to show the **Power of Analytics**.

We have started the journey with 0.5477 accuracy to its peak with Random Forest Classifier. Yet there are many other influencing factors involved, with the given data, we have built our model to its peak many techniques right from outlier analysis based on normal ranges of wine chemical composition to Feature selection and Variable importance.