# CASE STUDY ON WINE DATA

- By Sanjay VS

# OBJECTIVE

The objective of this case study is to analyse the given wine data and to reduce man power which wine company hire to taste the quality of wine. We are going to build a classification model which will classify the quality of wine depending on multiple factors

# EXPLORATORY DATA ANALYSIS

There are 12 variables in this dataset. They are:

- FIXED ACIDITY: Predominant fixed acids found in wines are tartaric, malic, citric and succinic. They are non-volatile(do not evaporate readily)
- VOLATILE ACIDITY: Volatile acid refers to the steam distillable acids present in wine, primarily acetic acid but also lactic, formic, butyric and propionic acids. Presence of this leads to production of sometimes unpleasant aroma compounds.
- CITRIC ACID: Found in small quantities, citric acid adds to the freshness and flavour of wines.
- RESIDUAL SUGAR: It is usually measured in grams of sugar per litre of wine (g/l or g/L). Typically refers to the sugar remaining after fermentation stops but can also result from addition of unfermented must or ordinary table sugar.
- CHLORIDES : The amount of salt in the wine
- FREE $SO_2$: Sulphur dioxide plays a very important role in preventing oxidization. It also kills unwanted yeast or bacteria and maintains freshness of the wine.
- TOTAL $SO_2$: This refers to the amount of free and bound forms of $SO_2$ present in the wine. In concentrations over 50ppm $SO_2$ becomes evident in the nose and affects taste of the wine.
- DENSITY: This is the density of the wine. It mainly depends on the percent alcohol and sugar content.
- PH: Describes how acidic or basic the wine is on a scale of 0-14.
- SULPHATES: A wine additive which contributes to the presence of $SO_2$ in wine.
- ALCOHOL: The percentage of alcohol present in the wine.

- QUALITY: Output variable which gives the quality of the wine. Higher the number better the wine. Lies between 0-10

# DATA PRE-PROCESSING

The data given is in two different datasets as white wine data and red wine data. We see that there are 1599 observations for red wine and 4898 observations for white wine. Once the dataset is loaded into the environment a new variable "Type" is created for both the datasets which represent what kind of data it is. This will help us later after merging to classify between Red and White wine. Then the type variable is converted into factor.

## *Removing duplicate elements:*

If we closely look at the data we are given by sorting with respect to different columns, we see that there are many duplicate entries in the dataset. This will deviate measures like mean and overall distribution of the data. It may also cause problems when we create a predictive model on the data as values in the train data and test data may end up being the same thus giving the illusion of increased accuracy of your model.

So removing the duplicated data is necessary. The size of the data reduces to 1359 observations of red wine and 3961 observations of white wine. As we see the number of duplicates in this dataset is 1177. This could have affected our analysis significantly.

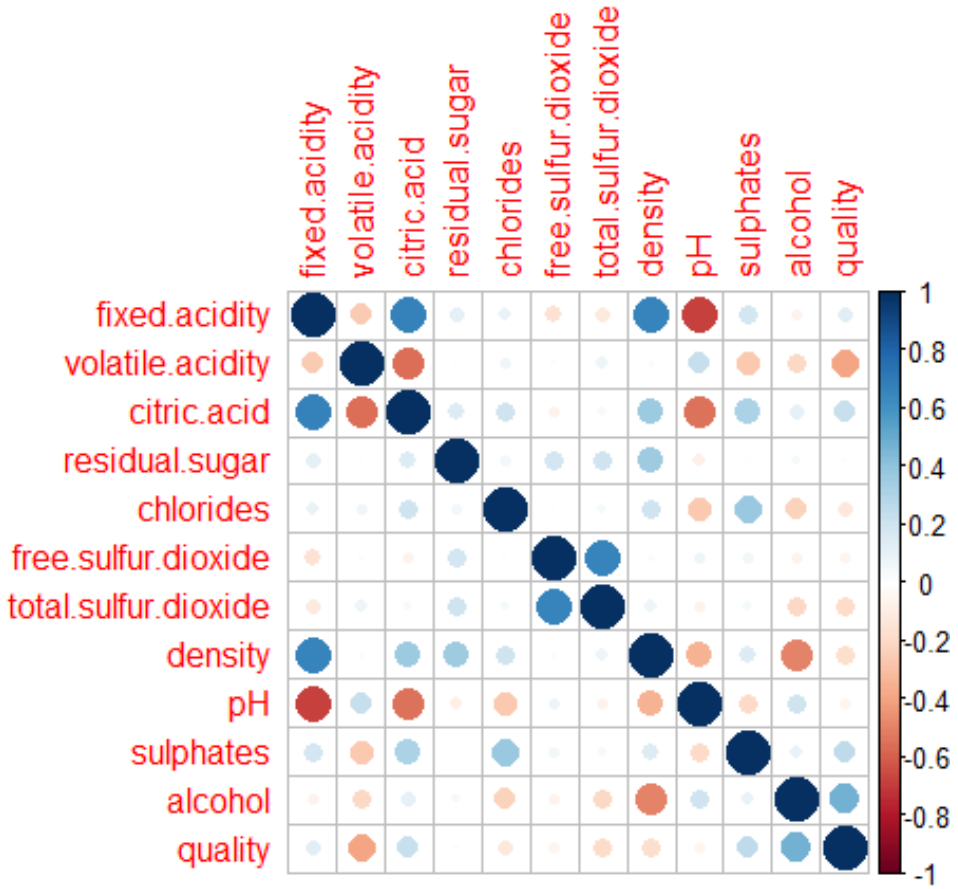
### *Merging the data:*

Once duplicate elements are removed we go on to merge the ttwo datasets to create master data. Order of the columns is also changed to take quality to be the last column for convenience.
We then change the datatype of the variable 'type' to factor.

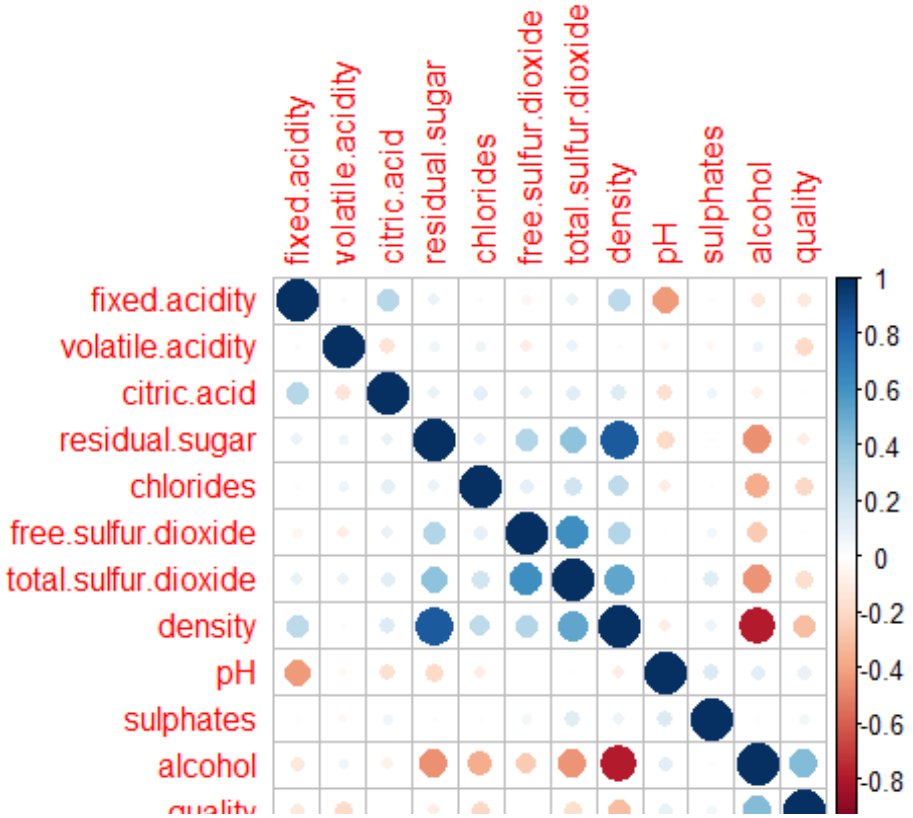
# EXPLORATORY DATA ANALYSIS

Since we have 12 variables we need to know which of these variables actually have a direct impact on the quality. Also we need to know the relation between different variables and how change in one variable is going to effect the other. For this we build a correlation plot which gives us an idea of the relationship between each variable.

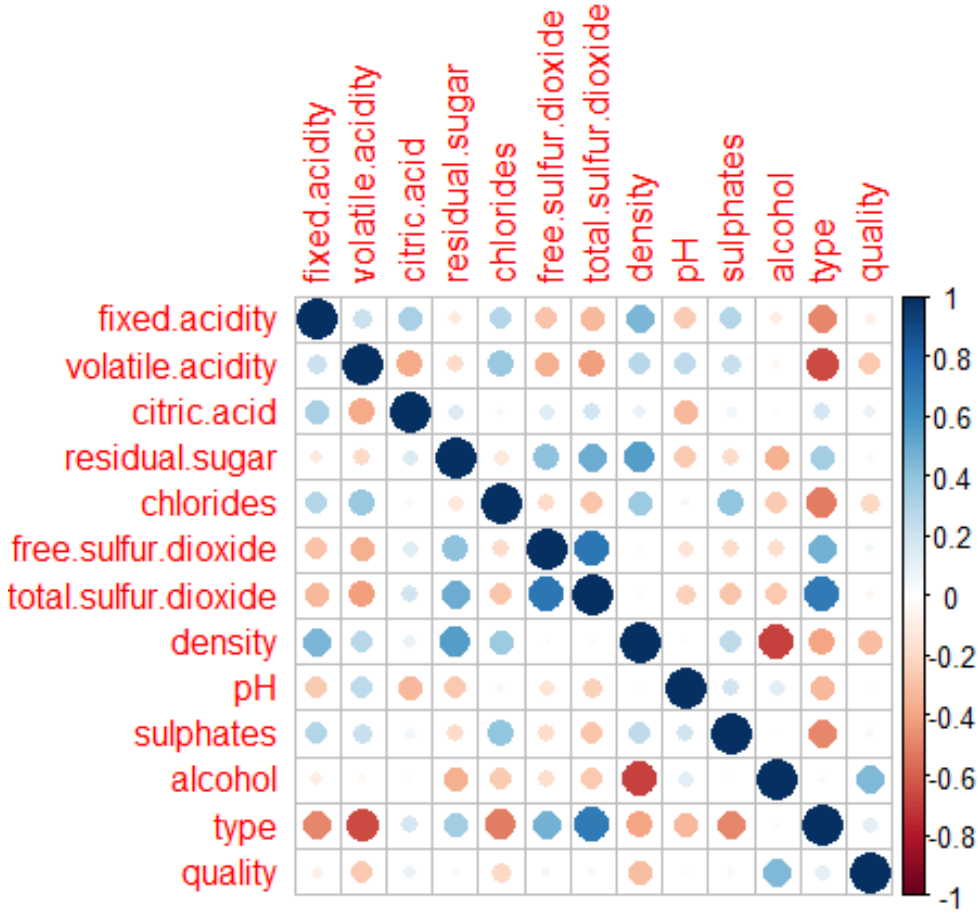
RED WINE
CORRELATION PLOT



WHITE WINE
CORRELATION PLOT




Wine Correlation Data
(Merged data)



As we see the correlation plots give the following inputs:

*Free $SO_2$*: Noticeable positive correlation with Total $SO_2$ and Residual sugar Negative correlation with volatile acidity and fixed acidity

*Total $SO_2$*: Positive correlation between free $SO_2$ and residual sugar Negative correlation with pH, Sulphates and Alcohol

*pH*: Positive correlation with Sulphates ,Alcohol and Volatile Acidity Negative correlation with Total and Free $SO_2$,Residual sugar ,citric acid, acidity(volatile and Fixed)

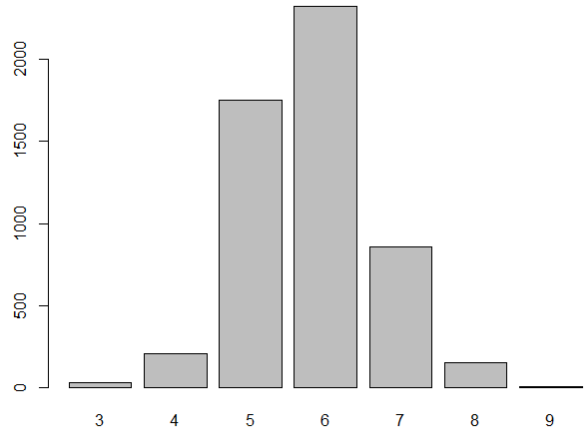
*Alcohol*: Positive correlation with pH and quality NEGATIVE Correlation with density, total and free $SO_2$, chlorides

*Quality*: positive correlation with alcohol negative correlation with density, chlorides, volatile acidity

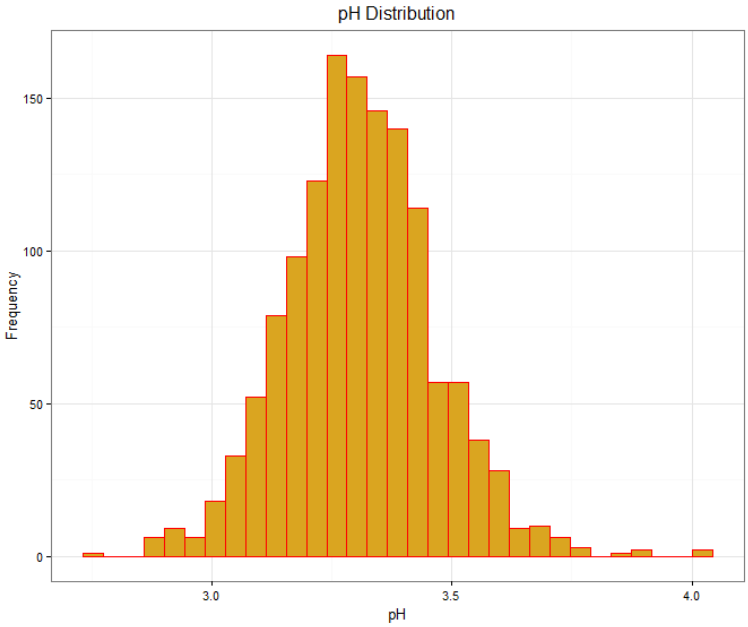
**DISTRIBUTION OF DATA:**

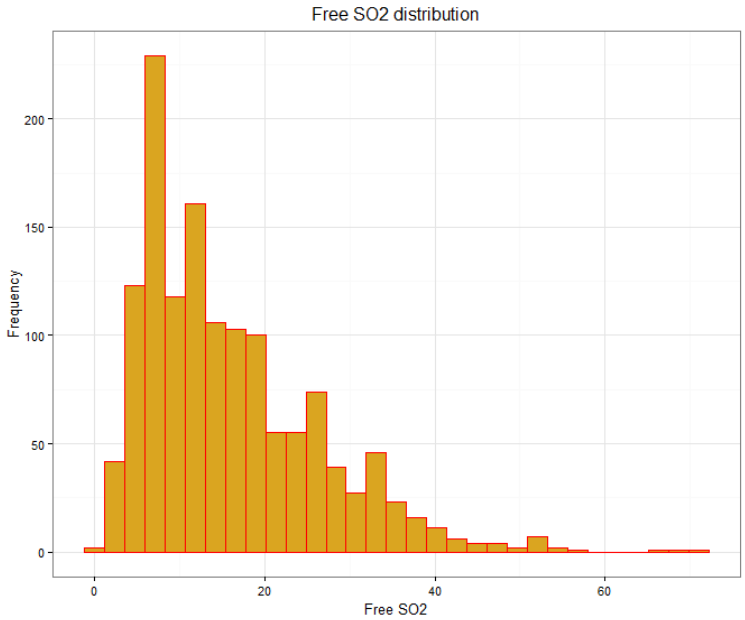
Firstly we will take a look at the distribution of quality in the data. Looking at the following bar plot, we can see that the data is not uniform i.e. here are more values that are average than there are good or bad. We can see that the total number of observations for wines with quality 6 clearly are more than any other quality wine data.

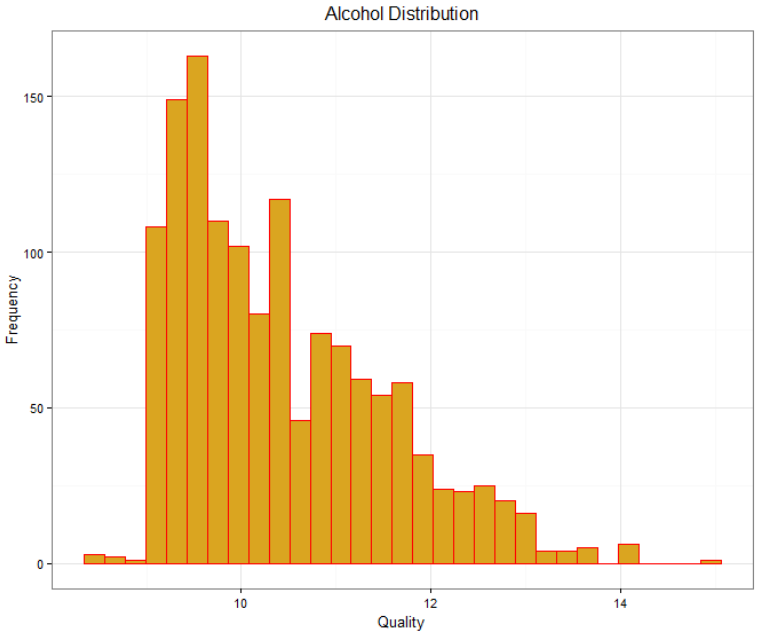
QUALITY DISTRIBUTION

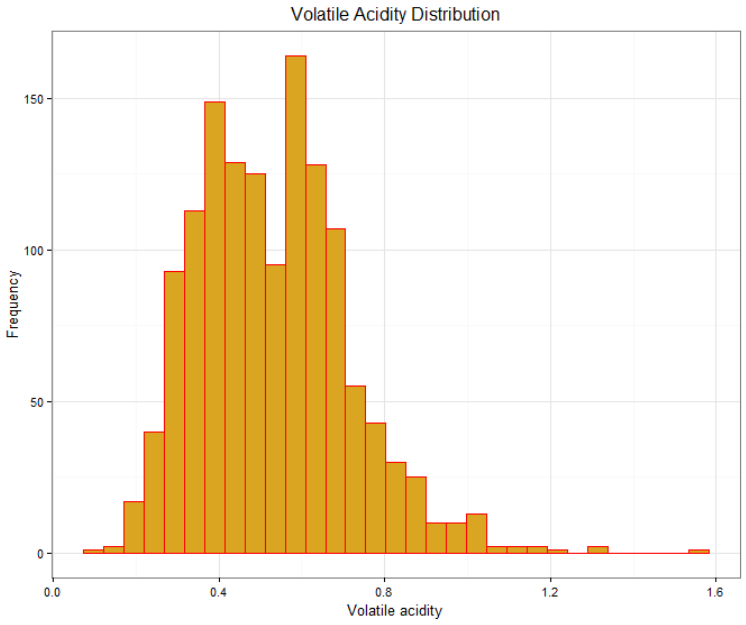


A distribution of the key parameters of pH, Free $SO_2$, Alcohol and Volatile Acidity is given for the Red wine dataset.



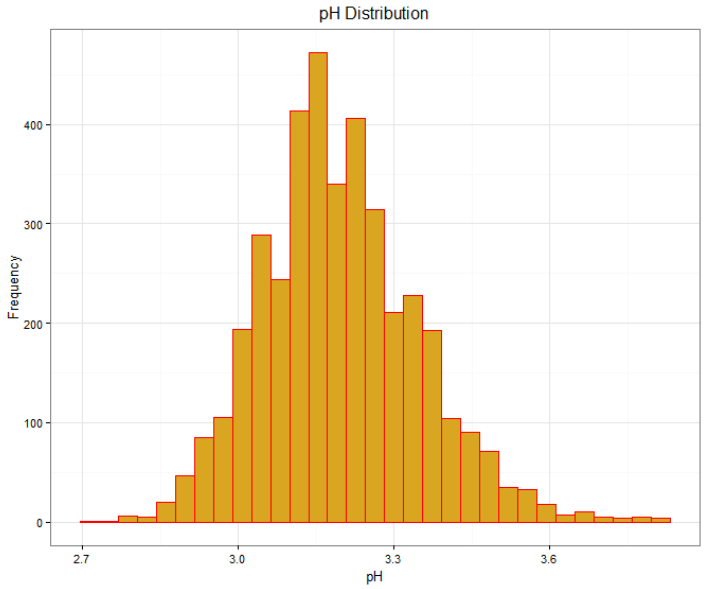



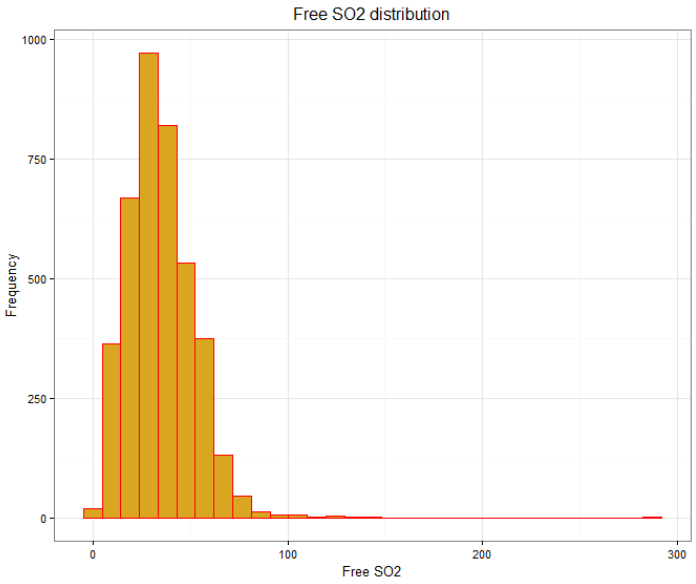

The following observations are obtained from the above plot

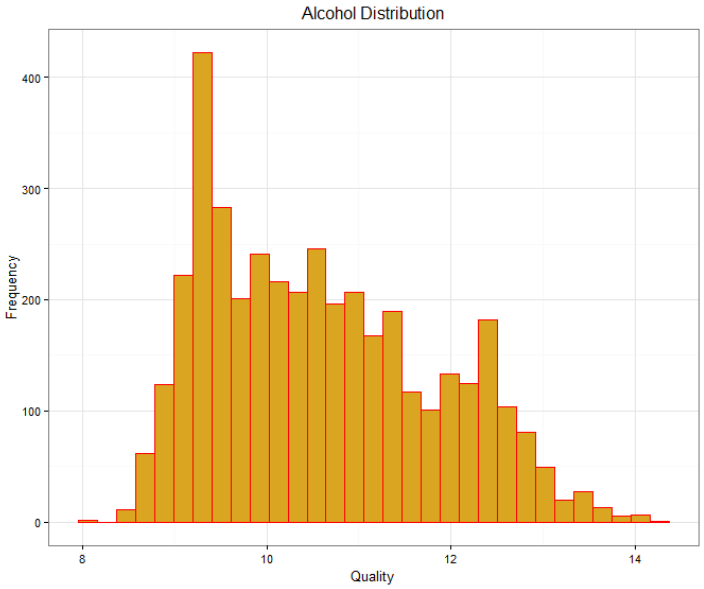
1. pH value shows a normal distribution with most of the samples having values between 3 and 3.5.
2. Free $SO_2$ shows a left skewed distribution as we can see that most of the samples lie towards the left of the graph.

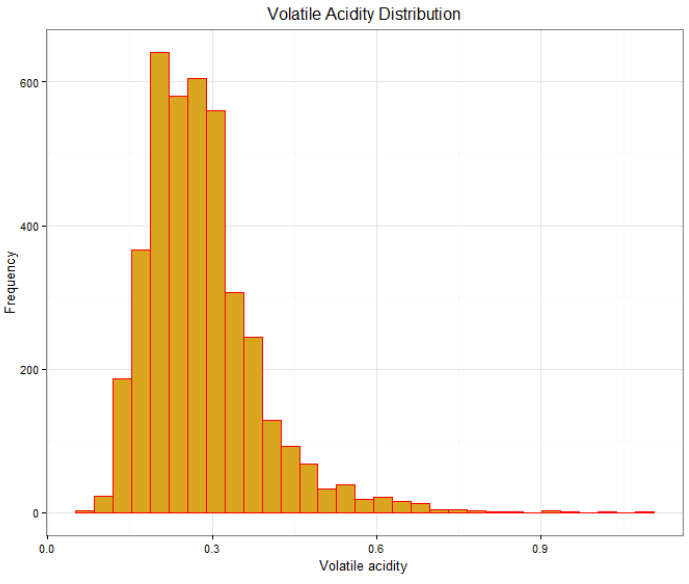
3. Alcohol content varies from 8 to 14 with most of the samples having values between 9 to 10.
4. Volatile acidity also shows kind of a normal curve with most of the values coming in the range 0.5 to 0.7

We do a similar analysis for the white wine data:



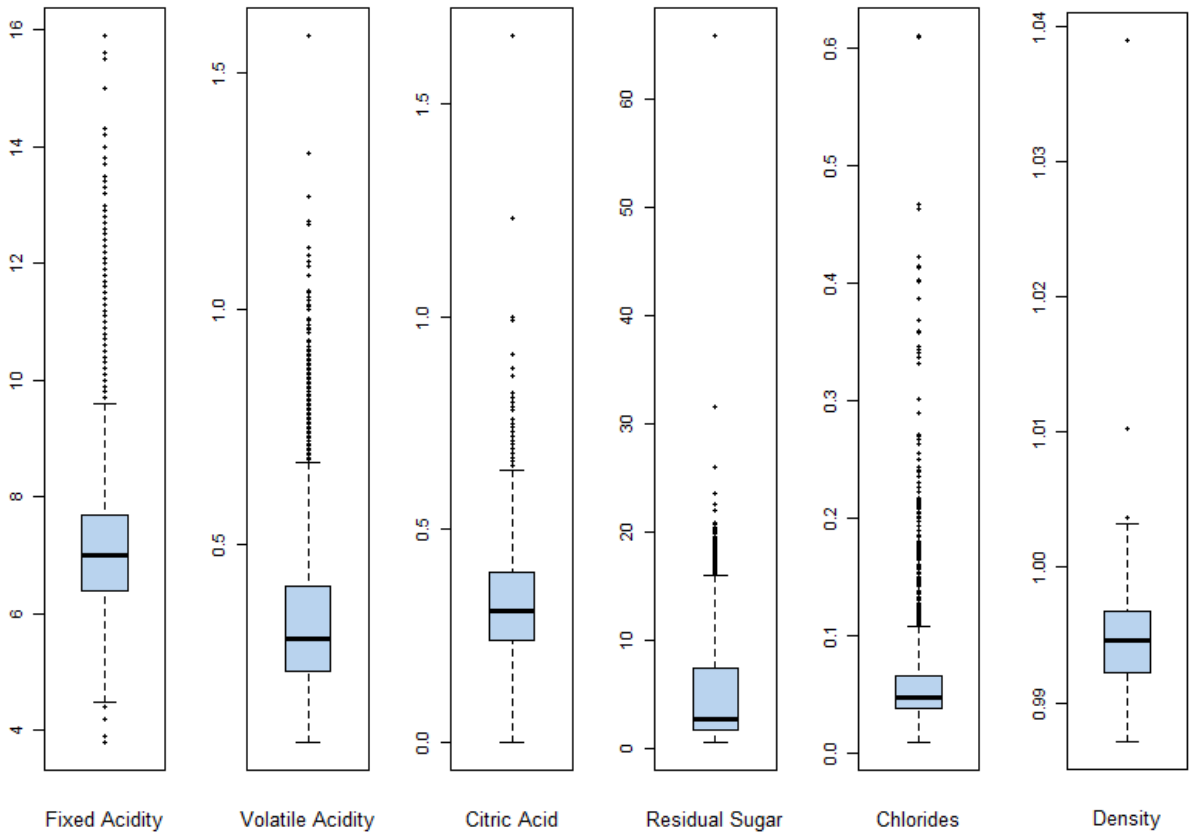



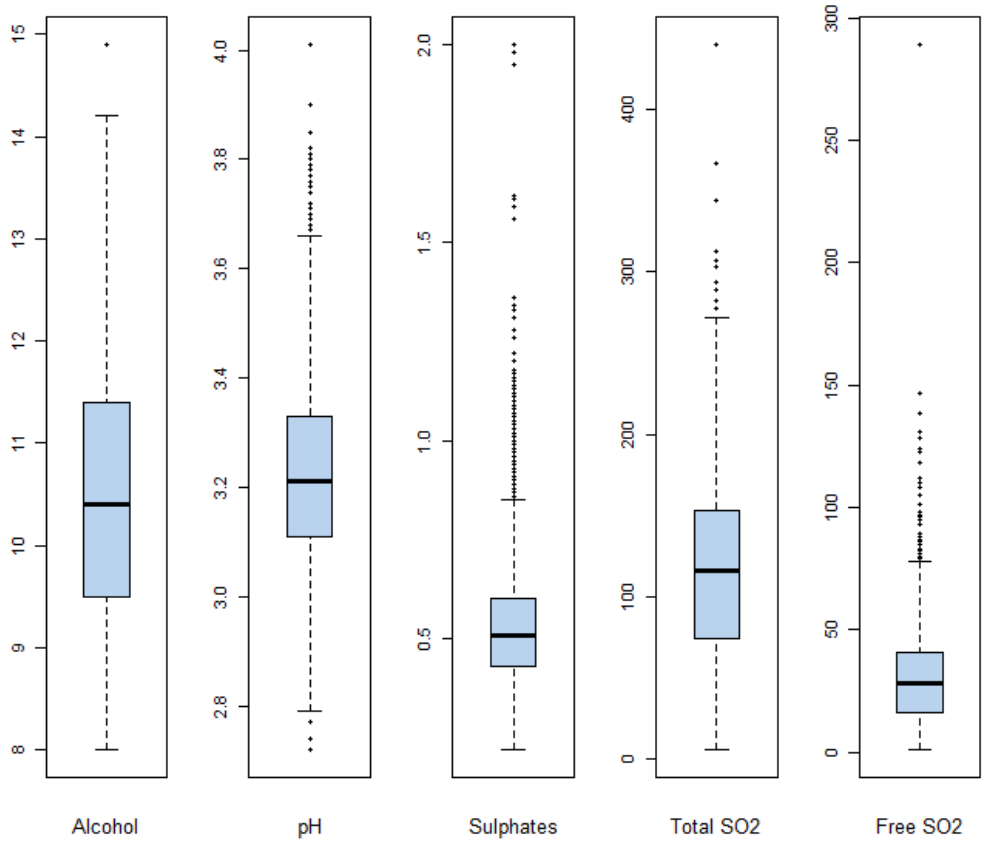
The observations made from these plots are :

1. For white wine too, pH distribution gives a normal curve with most of the values having a pH of between 3 and 3.5.
2. Free $SO_2$ has most of its samples around 50 and the range is from 0 to 120.
3. Alcohol has almost a uniform distribution but has most its values in the 9.5 to 10 range.
4. Volatile acidity shows almost a normal distribution with some extremities after 0.6.

# OUTLIER ANALYSIS

Box plots for all the variables is given below :






We observe from the following plots that all the variables have outliers. But some of these variables have so many outliers that if we end up removing all of them we might end up losing a lot of data.

The data we have with us is already a non-uniform distribution in terms of our target variable quality. As seen from the bar plot on Quality distribution we see that after removing the duplicate values we have :

| Quality | No. of observations |
|---|---|
| 3 | 30 |
| 4 | 206 |
| 5 | 1752 |
| 6 | 2323 |
| 7 | 856 |
| 8 | 148 |
| 9 | 5 |

What see is that the number of observations with quality value greater than 6 is much less as compared to the others. We also see that the number of observations with quality value 5 and 6 are more. So it would make sense to remove only the outliers with quality values as 5 or 6 as it would make the dataset more balanced.

After removing the outliers of qualities 5 and 6 the data we end up with 1406 observations of quality 5 and 1680 observations of quality 6. This not only solves our problem of sampling but we don't even lose data from the values with quality 3,4,7 and 8 as we have less observations for them anyway.

# RANDOM FOREST

The classification model I have used to predict the taste is random forest.

The problem statement asks us to build a model which would classify quality of wine. The aim of the project is to reduce the man power which wine company hired to taste the quality of wine before launching to the market. We know that the response variable quality is assumed to be ordinal and not continuous. Number of observations of Categories of quality too low(4 or less) or too high(more than 8) is very small. Therefore the wines are classified into 3 different categories by combining 3,4 and 5 to form one category(Bad taste), 6 into another category(normal taste) and finally 7 and above into the last category(good taste).

We create this new variable 'taste' and merge it to the dataset.

Next we classify the data into train data and test data. Then unsupervised Random forest method on training data is run with the number of trees = 1000. It leads to the following importance of predictors

| | Mean Decrease Accuracy |
|---|---|
| Fixed acidity | 31.45567 |
| Volatile acidity | 64.53439 |
| Citric acid | 37.23535 |
| Residual sugar | 33.08047 |
| Chlorides | 49.41673 |
| Free sulphur dioxide | 42.18247 |
| Total sulphur dioxide | 36.25573 |
| Density | 43.62715 |
| PH | 30.82379 |
| Sulphates | 38.21215 |
| Alcohol | 99.21511 |

As we can see alcohol is very important in determining the quality of the wine.

When we run our model on the test data we create the following confusion matrix

| PREDICTIONS | BAD | GOOD | AVERAGE |
|---|---|---|---|
| BAD | 423 | 25 | 124 |
| GOOD | 9 | 175 | 90 |
| AVERAGE | 134 | 131 | 362 |

The accuracy we get for this is 65.17%.

This is much higher compared to the accuracy we get by building a Naive Bayes model on top of the data which results in 55% accuracy.

Sensitivity for bad class is 74%, good class is 52.8% and average class is 62.8%

Specificity for bad class is 83.5%, good class is 91.33% and normal class is 62.85%.

As we can see the number of values where a good class is rated as bad and a bad class rated as good is less comparatively.

# CONCLUSION

What we find is that the main factors affecting the quality of the wine is Alcohol. Volatile acidity seems to be a determinant for faulty wines.

A model with decent accuracy can be built with the amount of data given. Another approach where we consider the majority output of Random forest, knn and naïve bayes is also worth a try in such a scenario but it gives similar results.