Wine Quality Data Analysis

AIT580 Final Project report

**Shravan Chintha**

**G01064991**

**Introduction:**

The dataset contains data related to the red variant of Portuguese “Vinho Verde” wine. The dataset is obtained from **UC Irvine Machine Learning Repository**. The UCI Machine Learning Repository is a collection of databases, domain theories and data generators which are used for analysing machine learning algorithms. This archive was created in 1987 by a few graduate students at UC Irvine. Since then, the archive has been used by students, educators, researchers etc., all over the world as a primary source of machine learning data sets.

The UC Irvine Machine Learning repository is hosted by the Center for Machine Learning and intelligent systems at UC Irvine. They maintain data as a service to the machine learning community. The archive is open to public and anyone can obtain a dataset from their website. The archive is simple webpage which displays all the datasets with their descriptions and links to download. A user can navigate to their website and use their searchable interface to get the desired datasets for building their machine learning models.

**About data:**

The data set is created using various samples of Red Vinho Verde wine. The Red Vinho Verde is one of the Portuguese variant of wine which is intense red in colour with vinous aroma, specially of berries goes well with food. The dataset contains 12 attributes with 1599 instances on a whole. Of the 12 attributes present, 11 of the attributes are obtained based on physiochemical test on the wine as input variables and the other attribute is output variable to obtained based on the sensory data based on the input variables.

Attributes:

Input variables:

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulphur dioxide

7 - total sulphur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output variable:

12 - quality (score between 0 and 10)

Where the quality score 0 denotes very bad quality and 10 denotes the very high quality red wine. The dataset is sufficiently large with 1599 instances for data analysis and also complex dataset with a variety of data types i.e., Categorical, numerical and nominal variables present in the data.

The dataset is collected to predict the wine quality as a factor of various chemical compositions or properties involved in making the particular wine. Based on the results, the companies can concentrate on the important factors that increase the quality of wine and thereby implement those methods to achieve better results.

By analysing this data, the following questions can be answered:

1. What factors affect the quality of wine?
2. Which is the most effective model to predict the quality of wine based on their chemical properties?
3. Which properties are least involved in assessing the quality of wine?

As the data set is available for public research, there are no privacy issues with the dataset and can be used by anyone for their research on machine learning algorithms.

**Requirements/resources needed:**

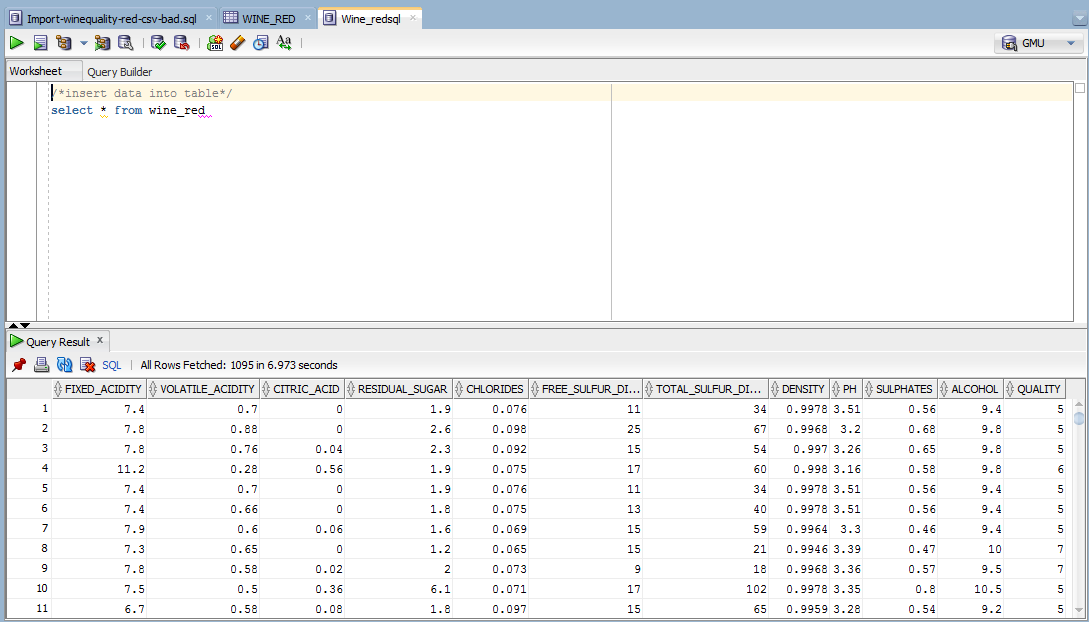
The analysis of data set is done using SQL, R and for few of the visualizations, Tableau is used. These are some of the machine learning tools that are used to analyse data. SQL is used to do exploratory analysis of the data such as finding metadata, knowing the data types etc. R is used to apply different models to the dataset to achieve desired result. And Tableau is used to get few visualizations which helps in analysing the data better and also easy.

**Findings:**

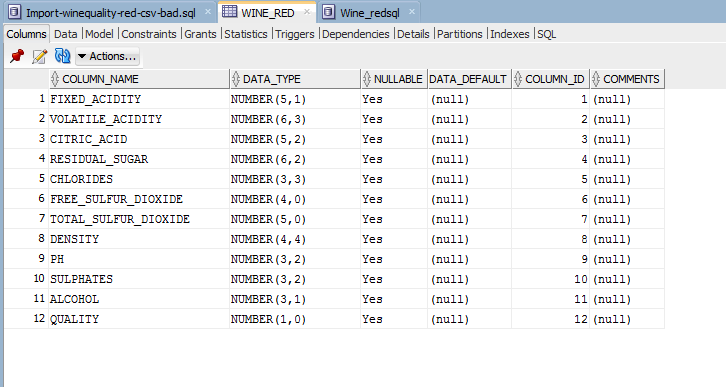
**Exploring the data:**

Using SQL, relevant metadata of variables is found. Inserting dataset into SQL using default import function in SQL would fetch the data types of the variables.

Importing dataset into SQL:

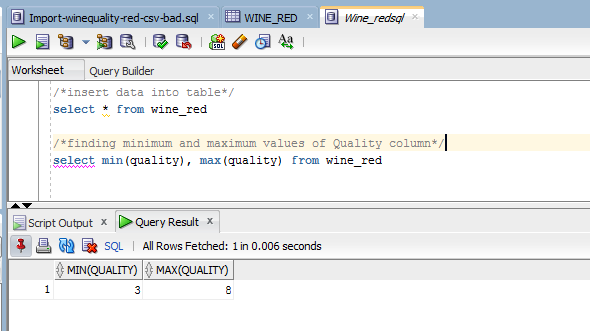


After inserting data by using import function, the SQL db defines the data types of the data inserted. Data types of the data inserted:



All the data types of the data present are “**Number**” data type.

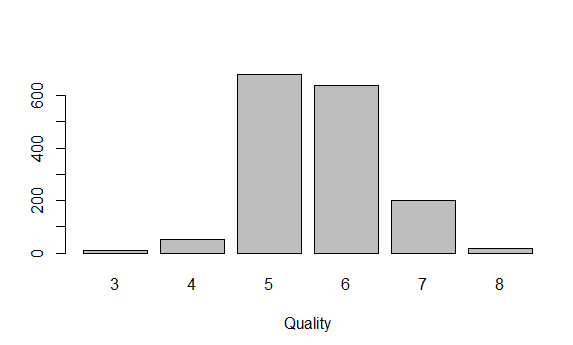
Using SQL commands, Maximum and Minimum values of output variable “Quality” can be determined as below:



Minimum value of Quality is 3 and maximum is 8 i.e., quality is ranging from 3 to 8.

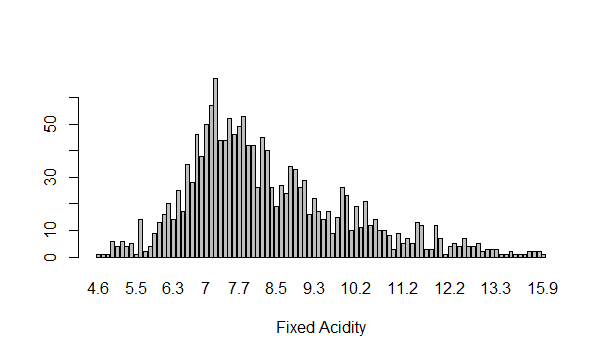
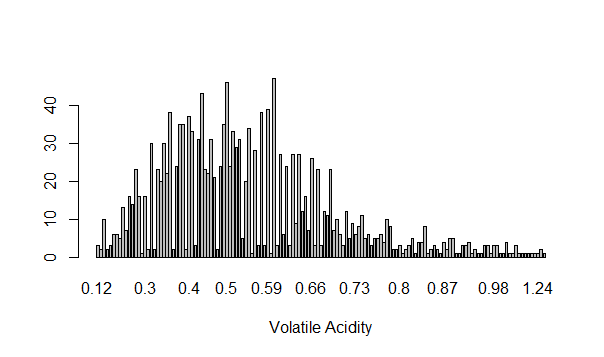
To get deep understanding of data and what methods to apply to get desired results, data needs to be explored and patterns are to be studied.

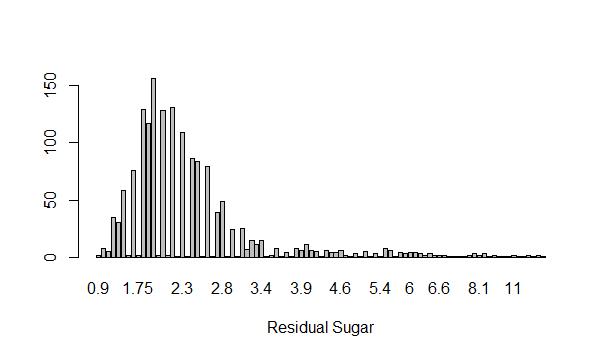
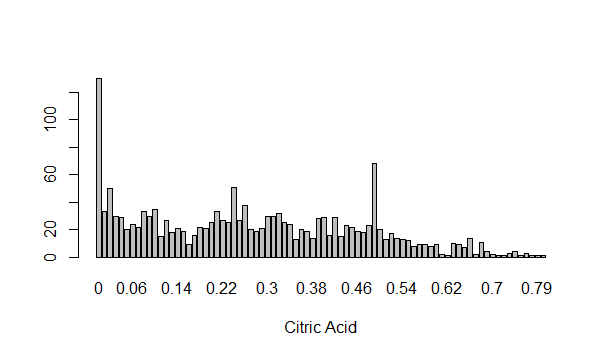
To get knowledge on the spread of the data of different input variables and output variable, they are plotted separately using bar plots in R, as shown below:

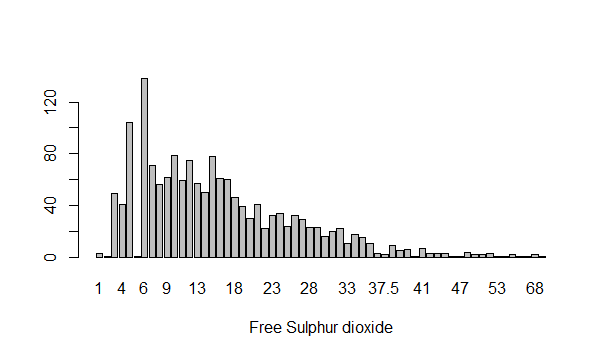
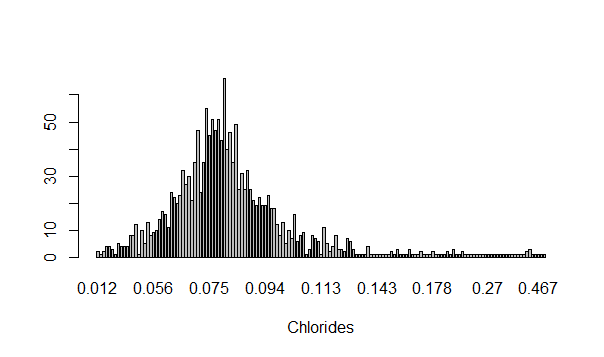


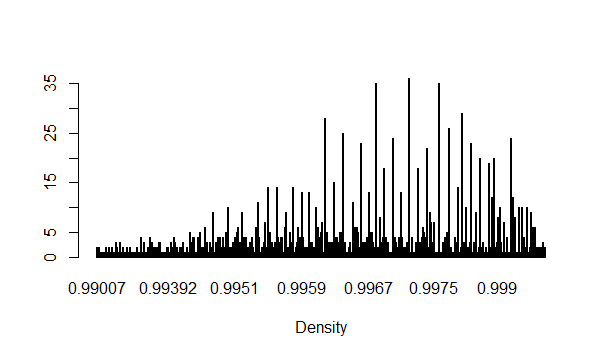
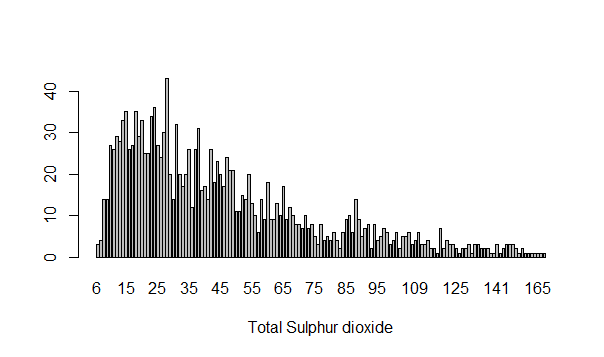
By observing the above plot, it can be known that the spread of the quality is mostly among 5 and 6.

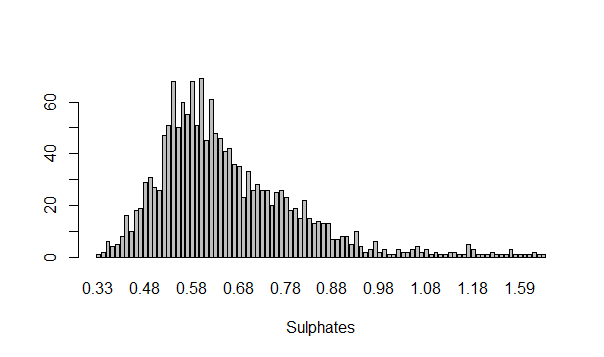
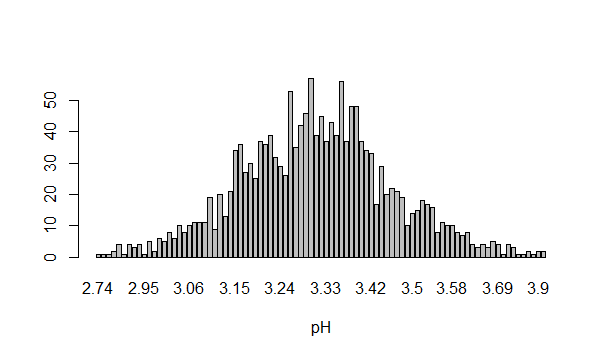
Plots for other variables:

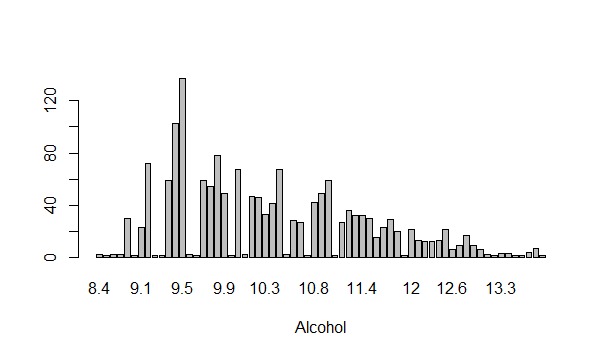
 











All the above plots show the spread of the data with respect to the values in their columns.

The summary of these data spread can be obtained using R,

**Descriptive statistics:**

summary(winequality\_red)

fixed acidity volatile acidity citric acid residual sugar

Min. : 4.600 Min. :0.1200 Min. :0.0000 Min. : 0.900

1st Qu.: 7.100 1st Qu.:0.3900 1st Qu.:0.0900 1st Qu.: 1.900

Median : 7.900 Median :0.5200 Median :0.2600 Median : 2.200

Mean : 8.322 Mean :0.5277 Mean :0.2713 Mean : 2.537

3rd Qu.: 9.200 3rd Qu.:0.6400 3rd Qu.:0.4200 3rd Qu.: 2.600

Max. :15.900 Max. :1.5800 Max. :1.0000 Max. :15.500

chlorides free sulfur dioxide total sulfur dioxide density

Min. :0.01200 Min. : 1.00 Min. : 6.00 Min. :0.9901

1st Qu.:0.07000 1st Qu.: 7.00 1st Qu.: 22.00 1st Qu.:0.9956

Median :0.07900 Median :14.00 Median : 38.00 Median :0.9968

Mean :0.08746 Mean :15.83 Mean : 46.43 Mean :0.9967

3rd Qu.:0.09000 3rd Qu.:21.00 3rd Qu.: 62.00 3rd Qu.:0.9978

Max. :0.61100 Max. :72.00 Max. :289.00 Max. :1.0037

pH sulphates alcohol quality

Min. :2.740 Min. :0.3300 Min. : 8.40 Min. :3.000

1st Qu.:3.210 1st Qu.:0.5500 1st Qu.: 9.50 1st Qu.:5.000

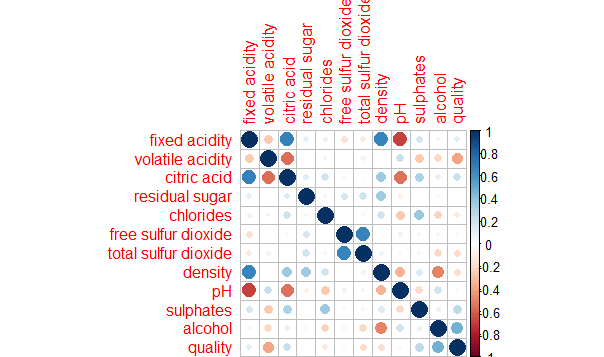
Median :3.310 Median :0.6200 Median :10.20 Median :6.000

Mean :3.311 Mean :0.6584 Mean :10.42 Mean :5.637

3rd Qu.:3.400 3rd Qu.:0.7300 3rd Qu.:11.10 3rd Qu.:6.000

Max. :4.010 Max. :2.0000 Max. :14.90 Max. :8.000

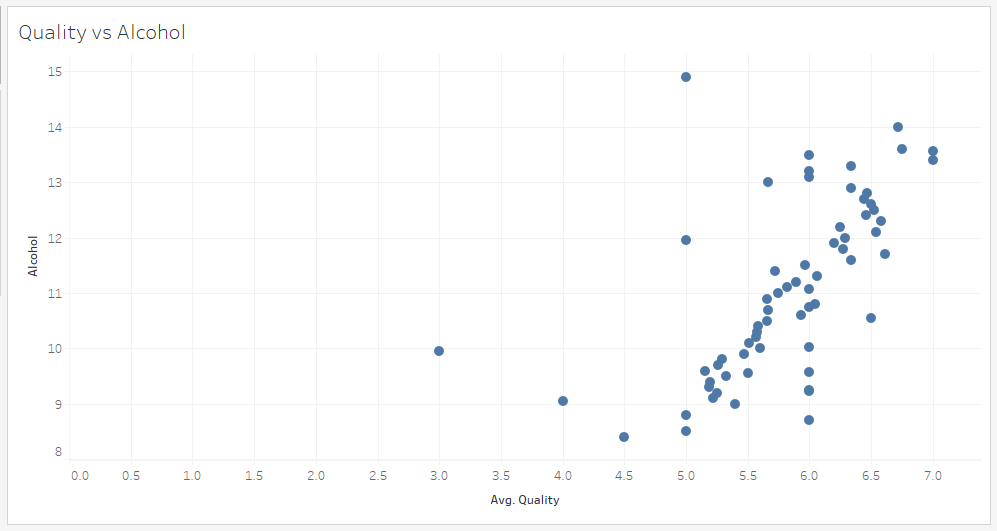
All the 12 attributes may not be the predictor variables used to predict the quality of wine. There may be few variables in the dataset that might not or have very low relation with the output variable. Analysing the variables that are very less related to the output variable may not yield the desired result. Hence to know the relation between all the variables, a correlation plot is plotted in R as shown below:



The above correlation chart shows the relation between all the variables. The blue and red colours indicate positive and negative correlation between the variables respectively. The higher the radius of the circle in the box, the higher is the correlation between the two variables. Based on the above chart obtained, it can be observed that, alcohol is highly positively related to the quality and volatile acidity is the highly negatively related variable with quality. Among all other variables, only Citric acid and Sulphates are considerably related to the output variable quality. All the other variables are not/less correlated.

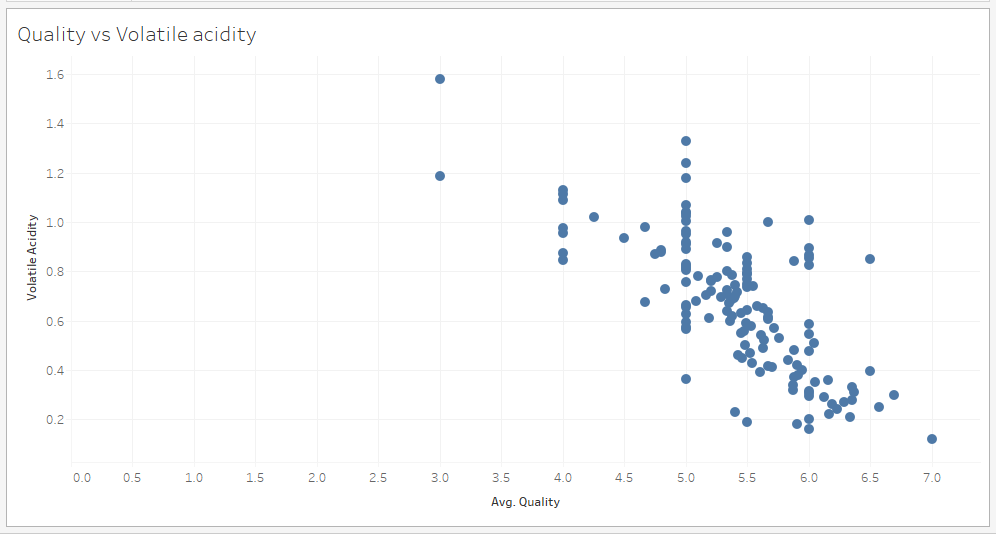
Plotting these variables against average of output variable “Quality” using Tableau, yielded the below plots:

Quality Vs Alcohol:

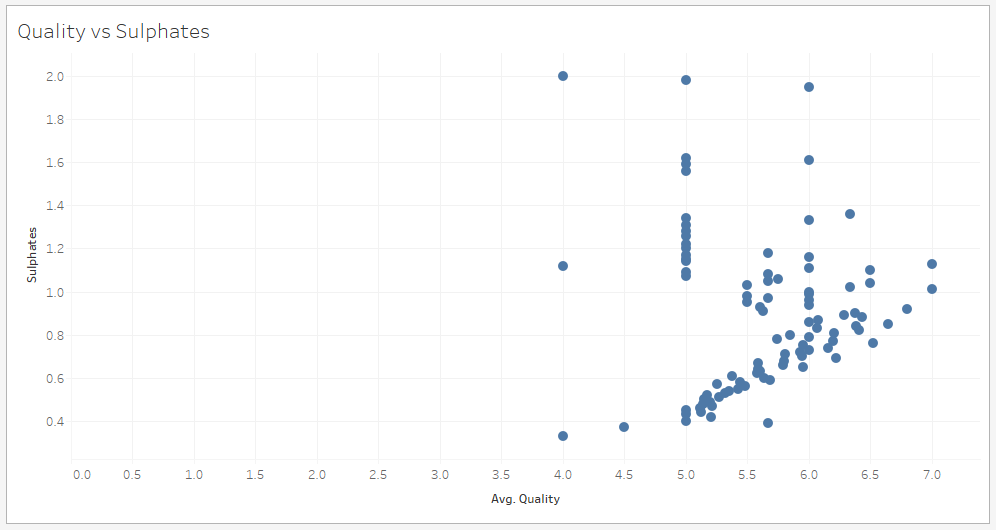


From the above plot, it can be observed that, with most of the data points, with the increase in Alcohol in the wine, the average of quality is getting increased. Positive correlation between Alcohol and Quality can be explained by this plot.

Quality Vs Volatile acidity:

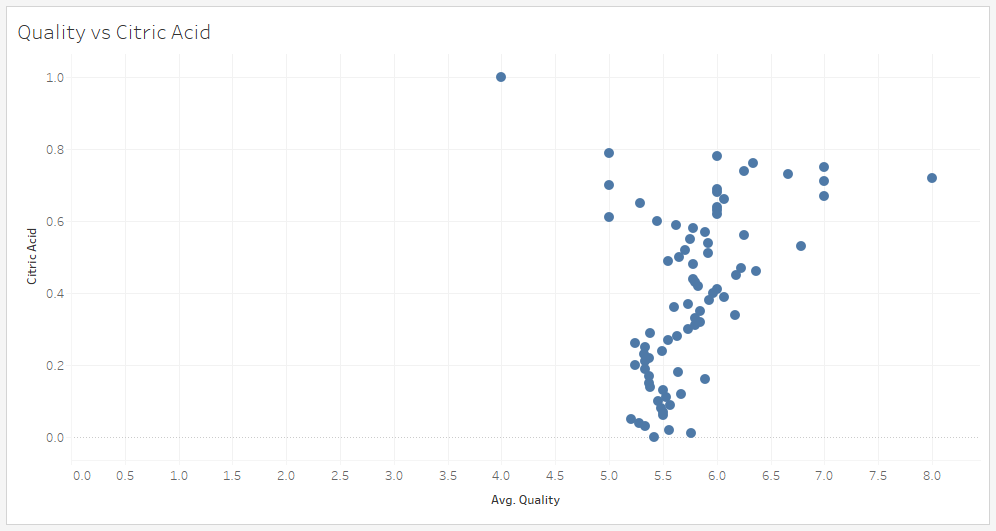
From the above plot, it is understood that with the decrease in volatile acidity in the wine, the average quality of the wine is getting increased. This explains the negative correlation of Volatile acidity with Quality of the wine.

Quality Vs Sulphates:



From the above plot, it can be observed that, for many of the points, with the increase in Sulphates, average value of quality is getting increased. As this not fully correlated, this pattern is happening for only these data points.

Quality Vs Citric acid:



From the above plot, it can be observed that the increase in Citric acid component in wine increases the average quality of wine but not as much as other attributes do.

**Modelling:**

**Multiple Linear regression:**

Applying multiple linear regression model on the output variables and required predictor input variables, obtained the below results:

> fit <- lm(quality ~ alcohol+volatileacidity+citricacid+sulphates, data=wine)

> summary(fit)

Call:

lm(formula = quality ~ alcohol + volatileacidity + citricacid +

sulphates, data = wine)

Residuals:

Min 1Q Median 3Q Max

-2.71402 -0.38607 -0.06302 0.46631 2.20416

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.64780 0.20138 13.149 < 2e-16 \*\*\*

alcohol 0.30899 0.01582 19.529 < 2e-16 \*\*\*

volatileacidity -1.26544 0.11275 -11.224 < 2e-16 \*\*\*

citricacid -0.07997 0.10395 -0.769 0.442

sulphates 0.69487 0.10321 6.732 2.32e-11 \*\*\*

---

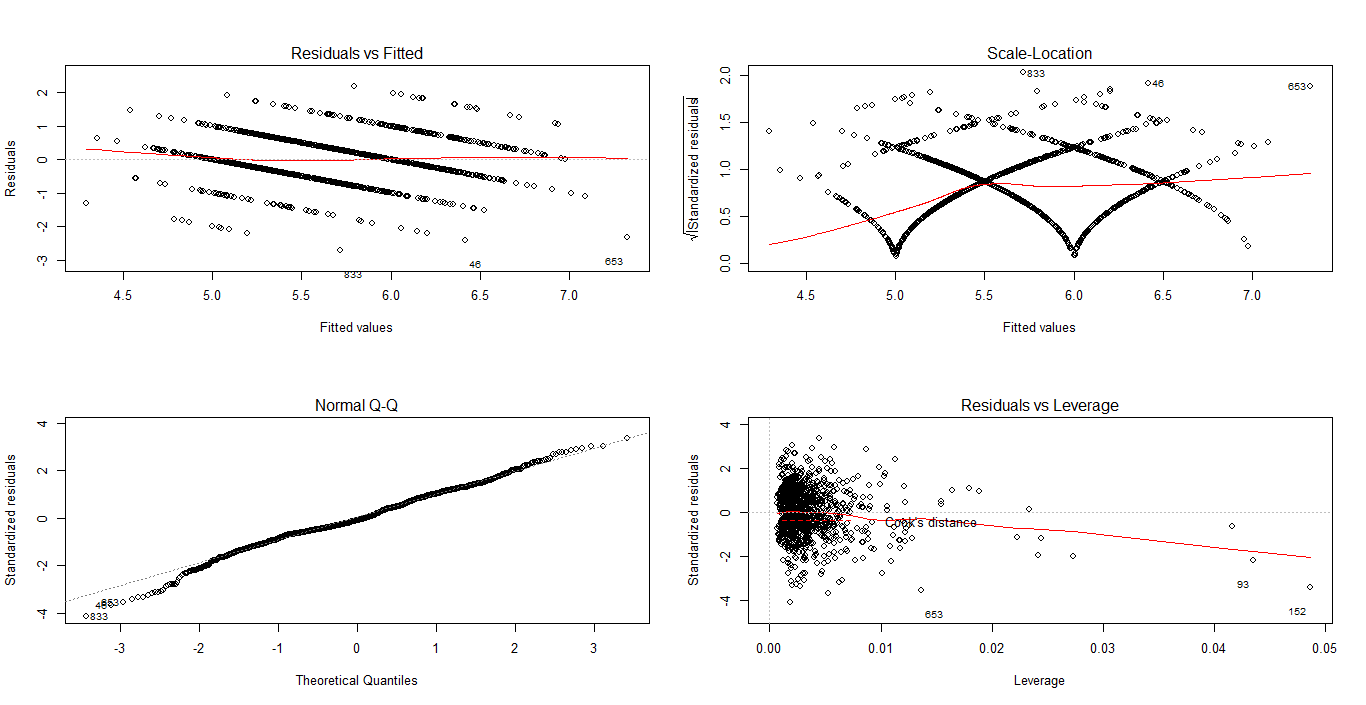
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6592 on 1592 degrees of freedom

Multiple R-squared: 0.3356, Adjusted R-squared: 0.334

F-statistic: 201.1 on 4 and 1592 DF, p-value: < 2.2e-16

And the residual plots obtained as below:



The R-squared value obtained is **0.3356** which is not closer to 1. So this model may not be ideal to use to predict the output variable.

**Relative Importance:**

This model provides the relative importance of each of the predictor among the 4 predictors that are chosen in above steps. By applying Relative importance for the 4 variables to determine the top variable to predict quality using R, obtained the below results:

library(relaimpo)

calc.relimp(fit,type=c("lmg","last","first","pratt"),

rela=TRUE)

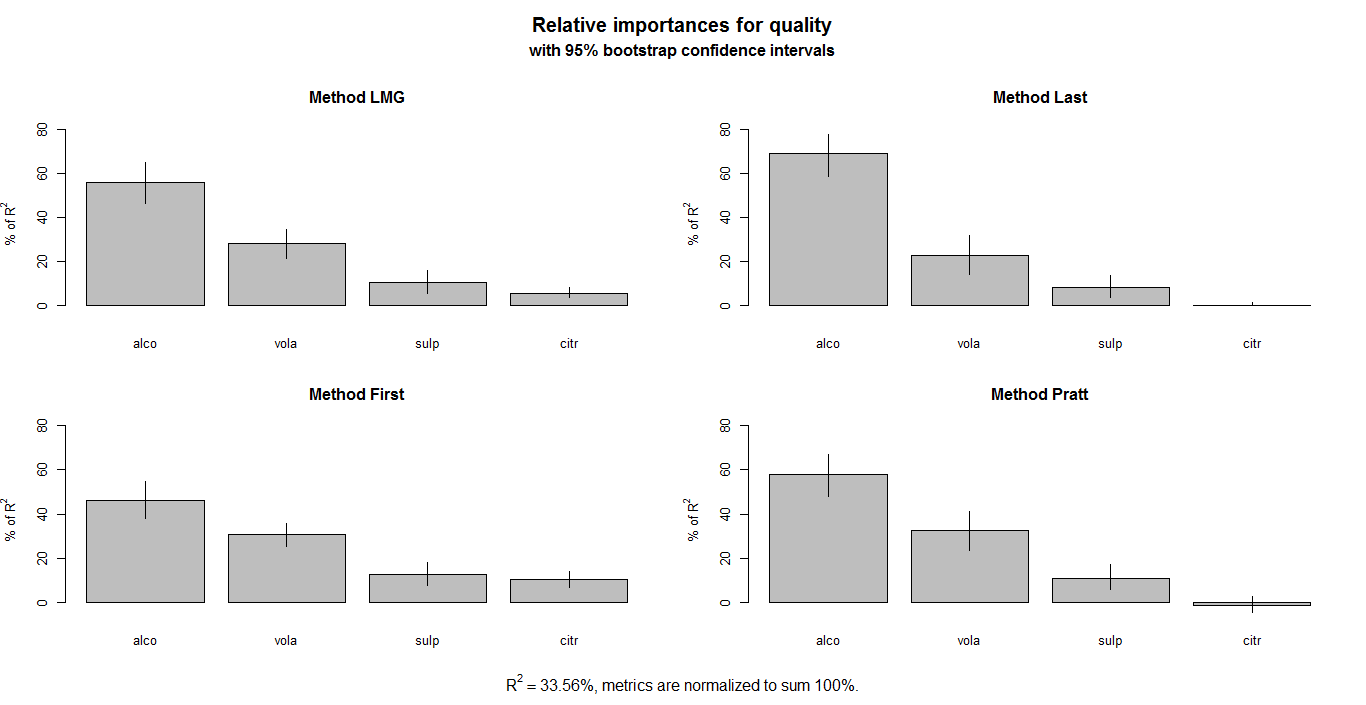
# Bootstrap Measures of Relative Importance (1000 samples)

boot <- boot.relimp(fit, b = 1000, type = c("lmg",

"last", "first", "pratt"), rank = TRUE, diff = TRUE, rela = TRUE)

booteval.relimp(boot) # print result

plot(booteval.relimp(boot,sort=TRUE)) # plot result



The above plot shows Alcohol content is the top predictor to predict the quality of wine among the 4 correlated input variables in the dataset.

**Classification:**

Applying Classification tree model on the data set, obtained the below results:

library(rpart)

> fit <- rpart(quality ~ alcohol+volatileacidity+sulphates+citricacid,

+ method="class", data=wine)

>

> printcp(fit)

Classification tree:

rpart(formula = quality ~ alcohol + volatileacidity + sulphates +

citricacid, data = wine, method = "class")

Variables actually used in tree construction:

[1] alcohol sulphates volatileacidity

Root node error: 918/1597 = 0.57483

n= 1597

CP nsplit rel error xerror xstd

1 0.235294 0 1.00000 1.00000 0.021521

2 0.015251 1 0.76471 0.78431 0.021661

3 0.012527 3 0.73420 0.76362 0.021603

4 0.010000 5 0.70915 0.76362 0.021603

> summary(fit) # detailed summary of splits

Call:

rpart(formula = quality ~ alcohol + volatileacidity + sulphates +

citricacid, data = wine, method = "class")

n= 1597

CP nsplit rel error xerror xstd

1 0.23529412 0 1.0000000 1.0000000 0.02152093

2 0.01525054 1 0.7647059 0.7843137 0.02166062

3 0.01252723 3 0.7342048 0.7636166 0.02160319

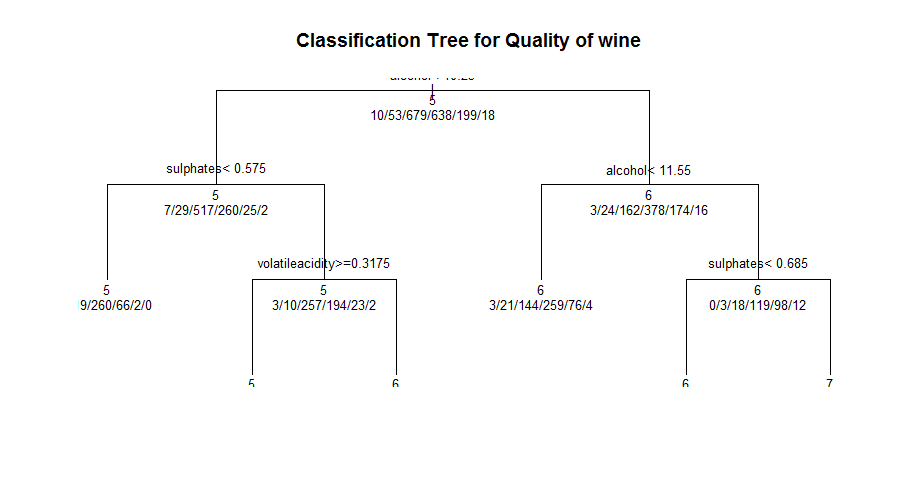
4 0.01000000 5 0.7091503 0.7636166 0.02160319

Variable importance

alcohol sulphates volatileacidity citricacid

53 22 16 9

On plotting tree, obtained the below classification tree.



The above tree shows the relation between other variables and how they are classified in terms of output variable.

**Random Forest:**

Applying Random Forest model to the dataset taking the key predictor variables against the output variable, obtained the below results:

> library(randomForest)

> fit <- randomForest(quality ~ alcohol + volatileacidity + sulphates+citricacid, data=wine)

> print(fit) # view results

Call:

randomForest(formula = quality ~ alcohol + volatileacidity + sulphates + citricacid, data = wine)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 1

Mean of squared residuals: 0.3450516

% Var explained: 47.08

> importance(fit) # importance of each predictor

IncNodePurity

alcohol 286.3116

volatileacidity 220.4439

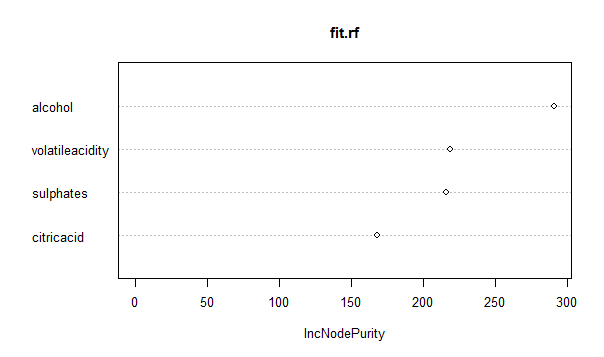
sulphates 217.2853

citricacid 168.2389

Also plotting important variable plot in random forest algorithm, obtained the below plot:

> imp <- importance(fit)

> varImpPlot(fit, cex = 0.8)



The chart shows the importance of variables that are used to predict the quality. Alcohol content stands top among the other variables.

Also, from the above results, we obtained Mean of squared residuals: 0.3450516 and % Var explained: 47.08, mean of squared residuals and %Var explained are used as performance metrics. These metrics suggest that this model is a good model. Comparing this model from above models, Random forest model is a better model as results obtained from this model are close to the desired results.

**Conclusion:**

After performing various modelling algorithms on the dataset, it is observed that Alcohol, Volatile Acidity, Sulphates, Citric Acid are the most important factors that affect the quality of wine. Among these predictors, Alcohol is the top factor that affects the quality of wine made. Hence it can be suggested that, quality of wine can be improved by working on the Alcohol content in the wine.

Among the techniques used above, Random Forest algorithm produced better results of all the other methods used to predict the quality of wine.

The variables that affect least among all the input variables in the dataset are Residual sugar, free sulphur dioxide and pH. It can be suggested that concentrating less on these variables could save the time in producing better quality of red wine.

**Explain/Define terms:**

1. Correlation:

Correlation is a statistical technique that shows whether and how strongly pairs of variables are related. It is a measure of the strength and direction of the linear relationship between two variables that is defined as the covariance of the variables divided by the product of their standard deviations.

1. Multiple Linear regression:

Multiple linear regression is techniques that attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y.

1. Classification:

Classification is a data mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis.

1. Random Forest:

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. Decision trees are individual learners that are combined. They are one of the most popular learning methods commonly used for data exploration.

**References:**

1. Dataset Citation:  
   P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, Modeling wine preferences by data mining from physicochemical properties, In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.
2. Source:

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez

A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal

@2009

1. Cortez et al. 2009. Wine Quality Data Set. Retrieved from UC Irvine machine learning repository site: <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>
2. UCI Machine Learning Repository. About page. Retrieved from UC Irvine machine learning repository site: <http://archive.ics.uci.edu/ml/about.html>
3. Robert I. Kabacoff, Ph.D. Statistics. Multiple Regression. Retrieved from Quick-R: <http://www.statmethods.net/stats/regression.html>
4. Gavin Douglas. 2017, March 7. Random Forest tutorial. Retrieved from GitHub website: <https://github.com/LangilleLab/microbiome_helper/wiki/Random-Forest-Tutorial#assessing-model-fit>
5. Ulrike Grömping. 2006. Relative Importance for Linear Regression in R. Retrieved from Semantic Scholar website: <https://www.semanticscholar.org/paper/Relative-Importance-for-Linear-Regression-in-R-The-Gr%C3%B6mping/1ce789bc60ba0a556c5ae848dde03a0b0e74384a>
6. Andrew Landgraf. 2012, July 19. Random Forest variable importance. Retrieved from R-Bloggers blog: <https://www.r-bloggers.com/random-forest-variable-importance/>