**FINAL PROJECT**

**AIT 580-P04**

**By**

**REDDIVARI SAI SARAN**

**G01142501**

**WINE DATA SET ANALYSIS**

A picture containing clipart

Description generated with high confidence

**INTRODUCTION**

Two Datasets considered are red and white wine variants of Portuguese “Vinho Verde” wine [1]. The dataset has a set of physicochemical and sensory variables are available due to individual privacy and logistic issues. The datasets which are chosen are viewed as a set of classification and regressions tasks. However, the data is related to the wine it does not include the grape type, wine band, and wine selling price.

**WHO**

Who collected the data?

F. Almeida, P.Cortez, A.Cerdeira, T.Matos, J.Reis are the people involved in collecting the data to understand how the quality and the pH value varies for the different types of white and red wine.

**NEED**

Why did they collect this data?

The primary purpose of collecting the data is for modeling the wine preferences by using data mining from different physicochemical properties. Several methods were computed to identify and understand which methods tend to achieve the best results.

Is there any privacy, quality, or other issues with this data?

**Quality issues**

Data quality issues are one of the critical reason where there could be a possibility of experiencing noise, abnormality and missing values of data which could be removed by imputing mean, using a filter to remove the missing values that could produce a correct outcome. I think the data set does not have any privacy issues since it is a data that can be accessible for everyone.

What could potential questions be answered by studying this data?

* Average pH value and quality of red wine and white wine (by using SQL)
* Which factors profoundly influence the quality of wine
* Are Free Sulphur dioxide and total Sulphur dioxide value of wine data related?
* What is the correlation of quality with other predictors?

**REQUIREMENTS and RESOURCES NEEDED**

HARDWARE REQUIREMENTS AND SOFTWARE REQUIREMENTS

HARDWARE REQUIREMENT :

The complete project analysis is done using the laptop of following configurations that includes

* Processor Name - Intel® Core™ i7-8550U CPU
* System Type - 64-bit operating system, x64-based processor
* RAM - 16.0 GB
* Processor speed - 1.80GHz 1.99GHz

SOFTWARE REQUIREMENT :

* R studio is used mainly for visualization and modeling.
* Oracle SQL Developer is used to create the database and for analyzation.

**DATA DESCRIPTION**

Size of the dataset- The dataset of 381.23 KB consists of 13 attributes along with data types:

* Type - Nominal
* Fixed Acidity - Ratio
* Volatile Acidity - Ratio
* Citric Acid - Ratio
* Residual Sugar - Ratio
* Chlorides - Ratio
* Free Sulfur Oxide - Ratio
* Total Sulfur Dioxide - Ratio
* Density - Ratio
* pH - Interval
* Sulphates - Ratio
* Alcohol - Ratio
* Quality - Ratio

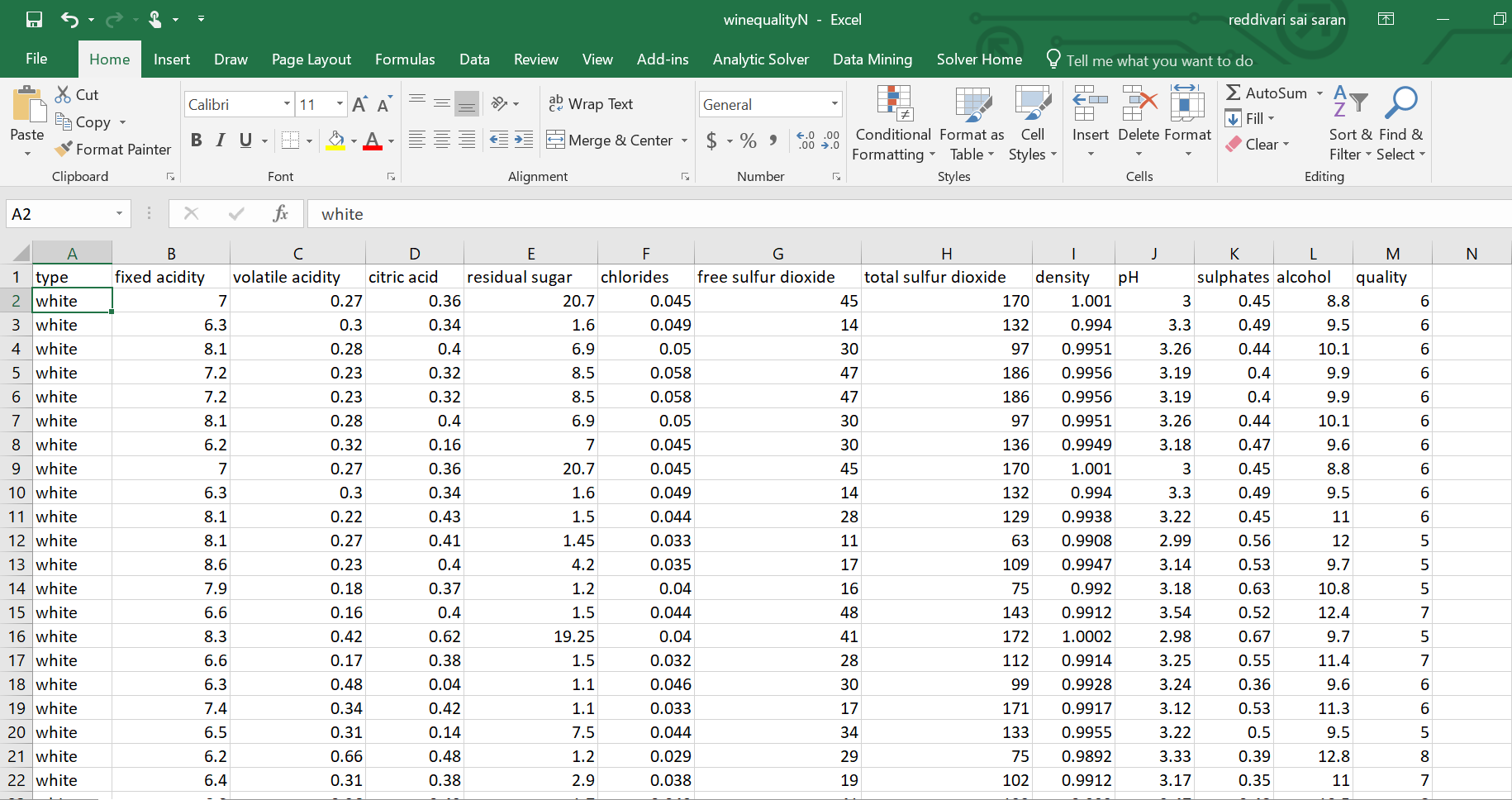
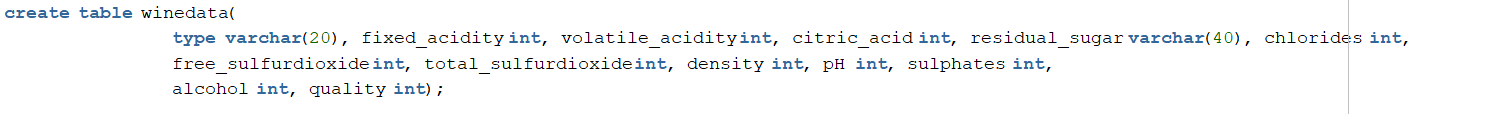


Fig 1: Wine dataset

It contains 13 attributes and the wine data set include 4870 white wine type and 1592 red wine type.

**SQL SCHEMA:**

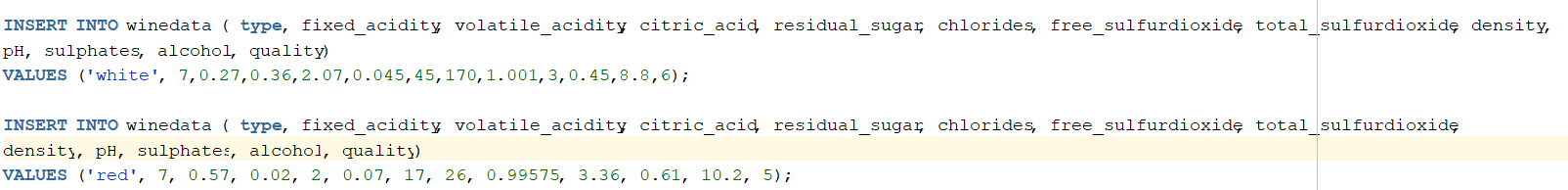
**CREATE TABLE :**

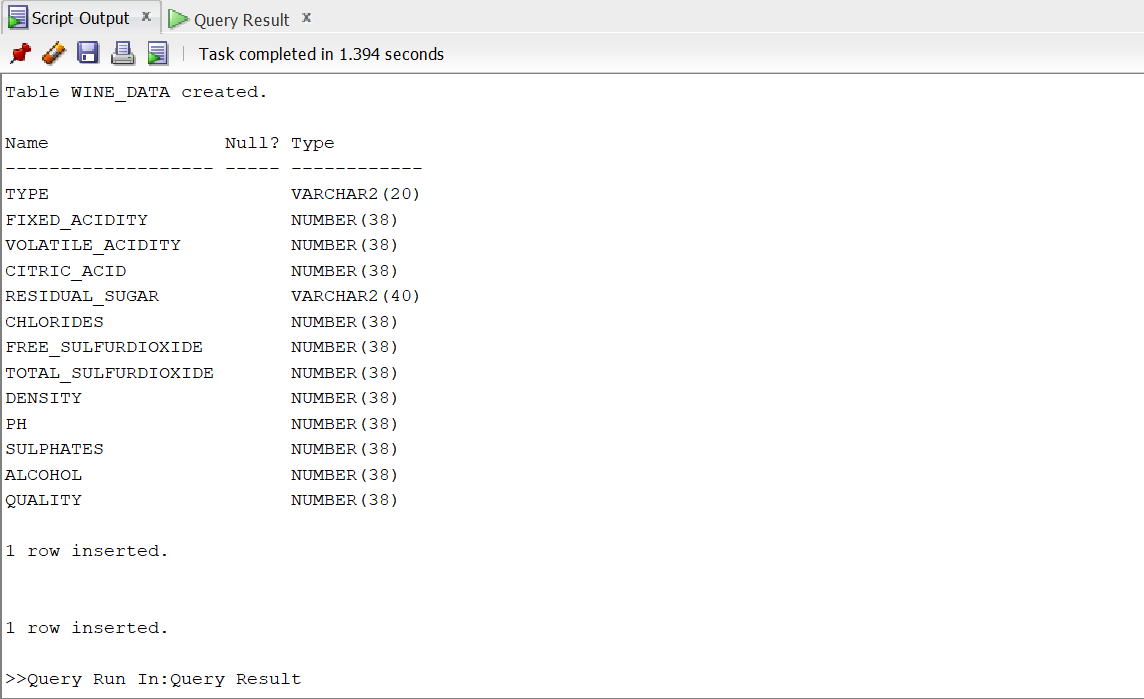


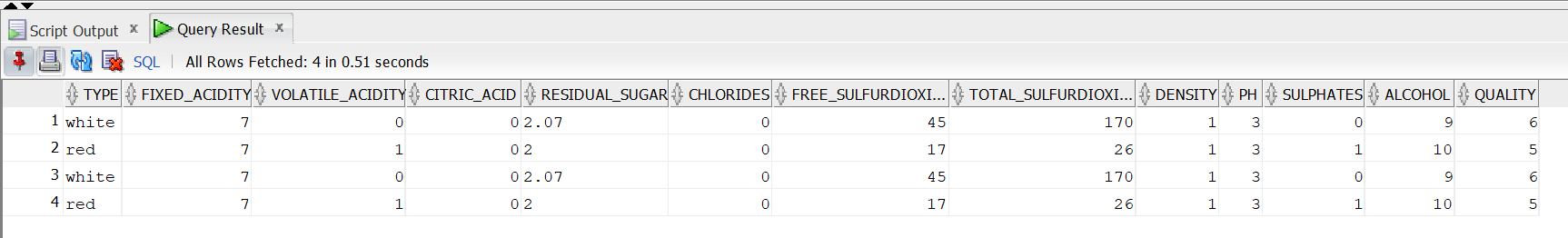
**DESCRIBE TABLE :**



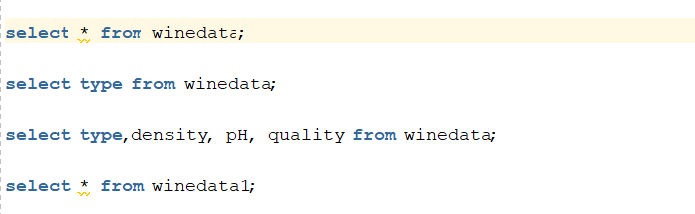
**INSERT VALUES INTO TABLE :**



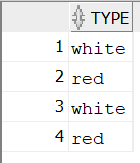
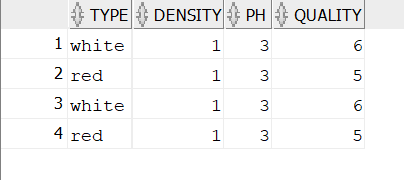
**OUTPUT:** : :



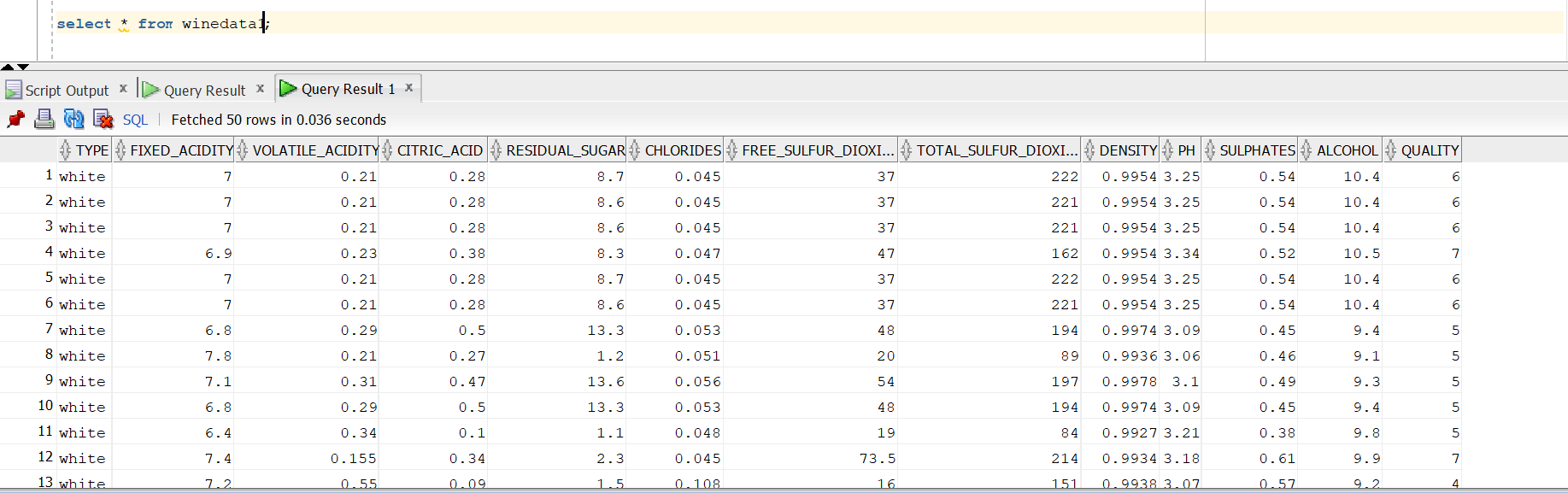
**SELECT STATEMENTS :**



OUTPUT :

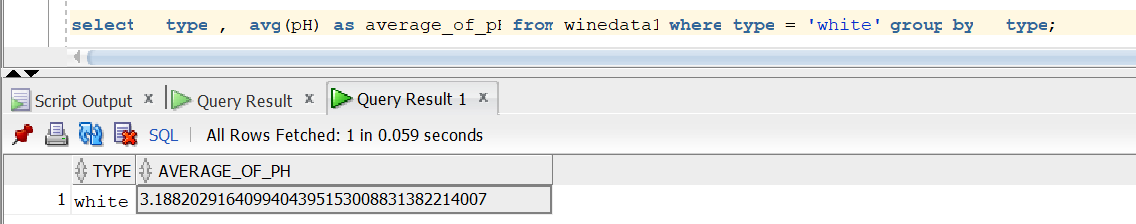
 

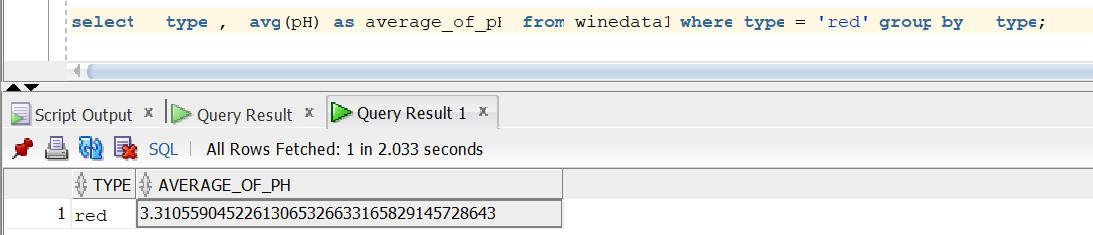
**RESULTS**



**Average of pH value:**

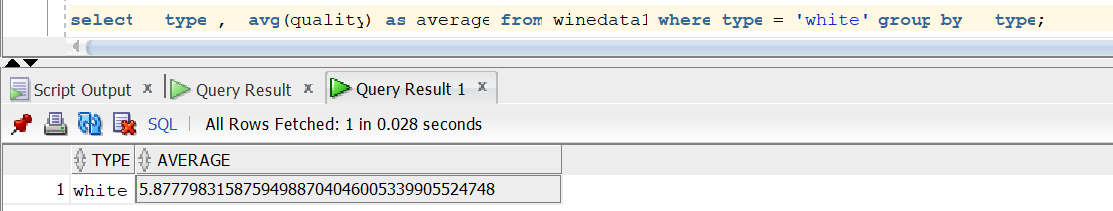
Average of pH value is mainly used to determine whether the alcohol is hard or soft.

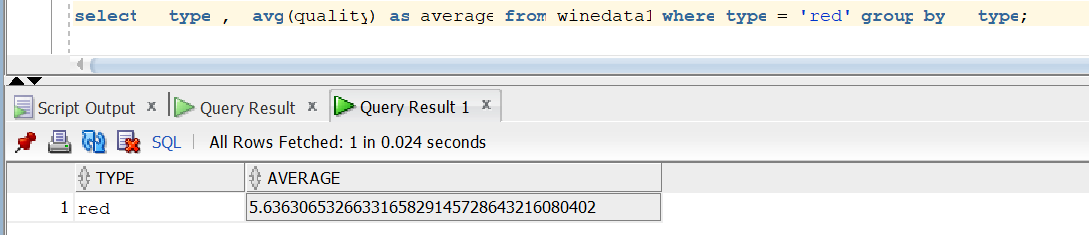




**Average of Quality:**

Average of Quality is used primarily to assess the quality of white wine and red wine.





**CORRELATION ANALYSIS :**

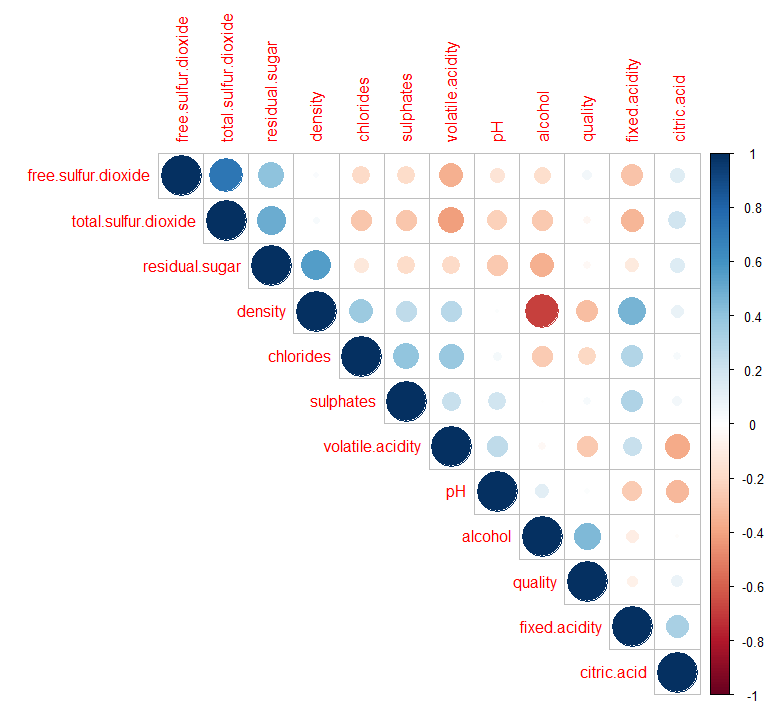


Fig 2: Correlation Analysis

The above correlation matrix clearly states the correlation of one variable with the other variable, i.e., it depicts how the there exists an interdependence between the variable quantities. Correlation coefficient value drawn from the above correlation matrix is -1 to 1. From the above correlation plot, it is evident that the quality of wine has is highly correlated with the percentage of alcohol content in it. It is understood from the graph is that the quality of wine has a negative correlation with both density and volatile density.

**BOX PLOT ANALYSIS :**

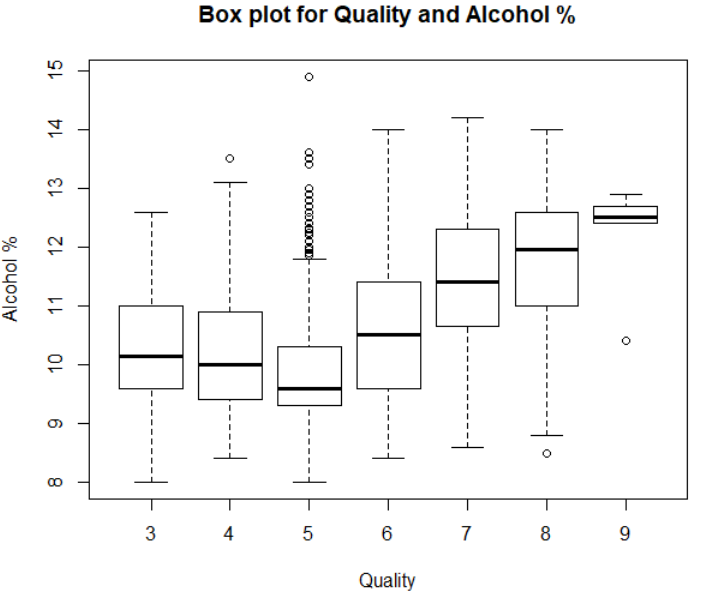
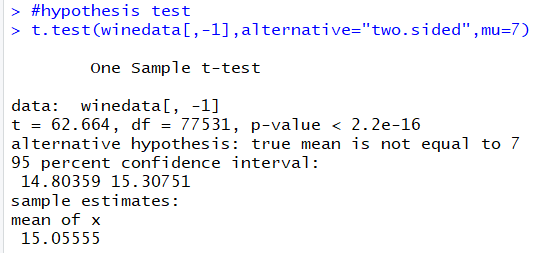


Fig 3: Box plot for Alcohol and Quality

From the above boxplot, it is understood that the box plotted is for the attributes quality and Alcohol %. It is evident from the above boxplot that the quality of the wine increased with an increase in alcohol in percentage. The quality range between 8-9 tends to have less number of outliers and has certainly good quality depending on the alcohol % in it.

**HYPOTHESIS TEST:**



Hypothesis Testing- From the above results, it is understood that as p-value is low, the results are truly significant and we can reject the null hypothesis at 95% confidence interval.

**HISTOGRAM ANALYSIS :**

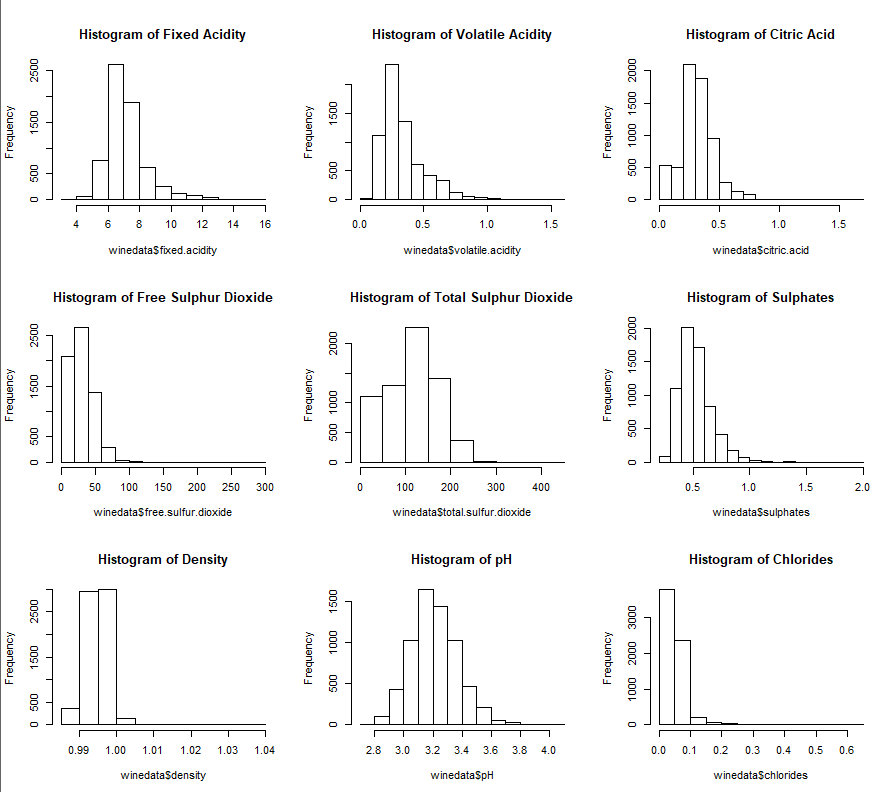


Fig 4: Histogram for all the attributes

The above histograms are plotted using some of the predictor's variables with the frequency which helps in determining the quality of the wine. The histogram for frequency and one predictor variable, i.e., winedata$chlorides is highly skewed to the right from the graph plotted above. Most of the histograms plotted above using the predictor variables and frequency are observed to be right-skewed.

**SCATTER PLOT MATRIX ANALYSIS:**

The below plot is a scatter plot matrix which is constructed using two variables. The scatter plot could be one of the best ways, of visualizing every possible combination, this could also be helpful in distinguishing different observations in each plot. The red portions of the plot indicate red wine and the blue part reports white wine from the plots plotted

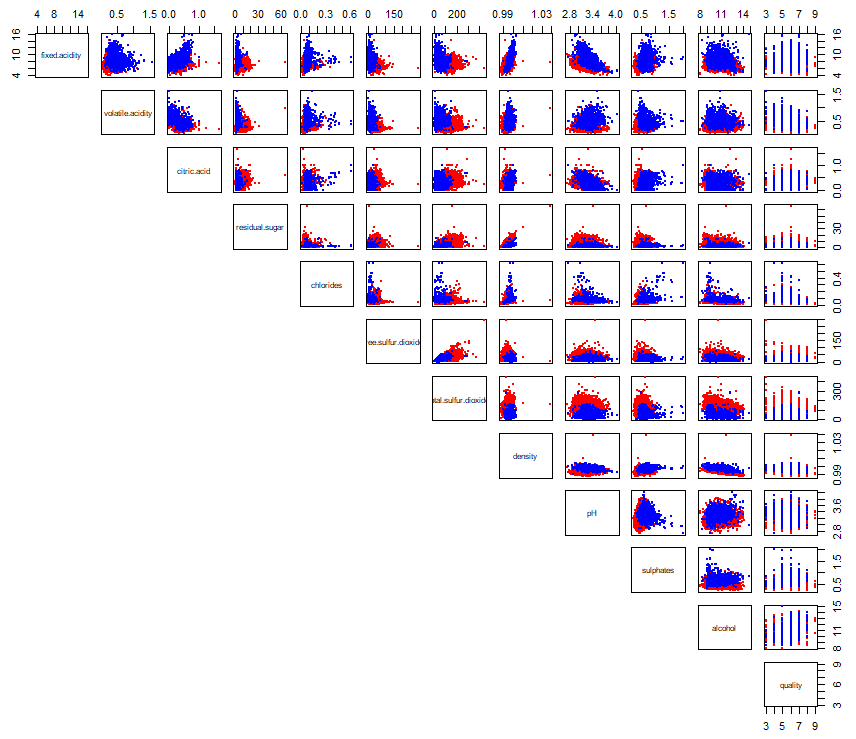


Fig 5: Scatter Plot Matrix

**SCATTER PLOT:**

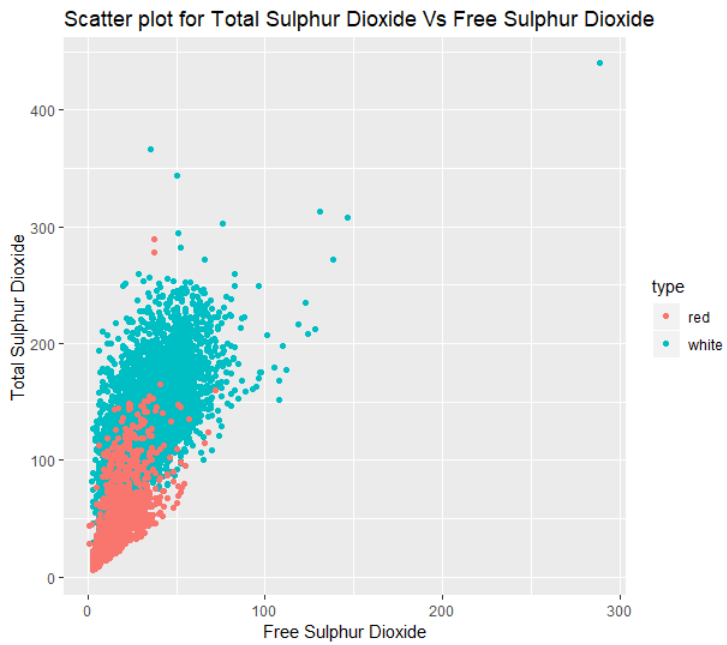
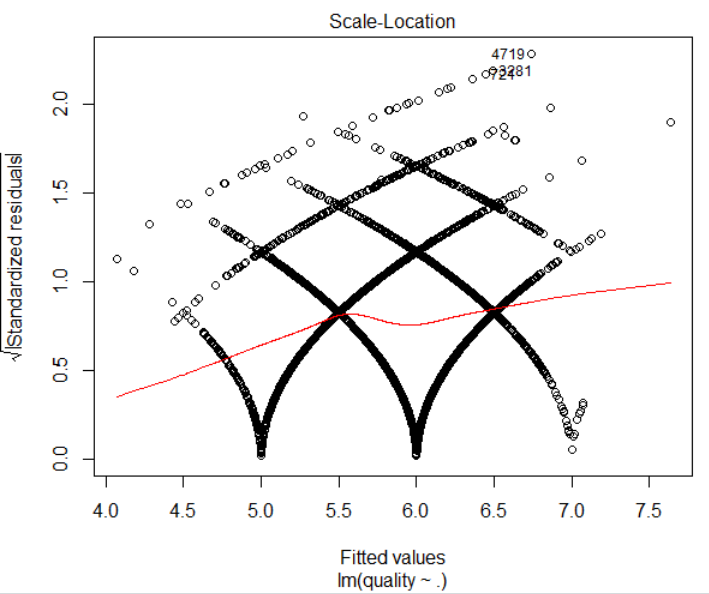
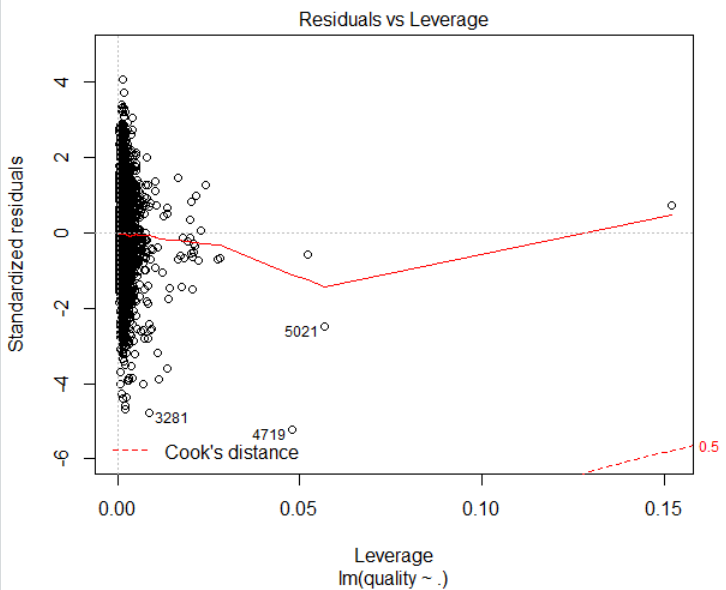


Fig 6: Scatter plot for Total Sulfur Dioxide and Free Sulfur Dioxide

The above scatter plot helps in visualizing the attributes free Sulphur dioxide and total Sulfur dioxide. From the scatter plot above it is evident that white wine has more of total Sulphur dioxide when compared to the red wine which is considerably less than white wine.

**LINEAR REGRESSION:**



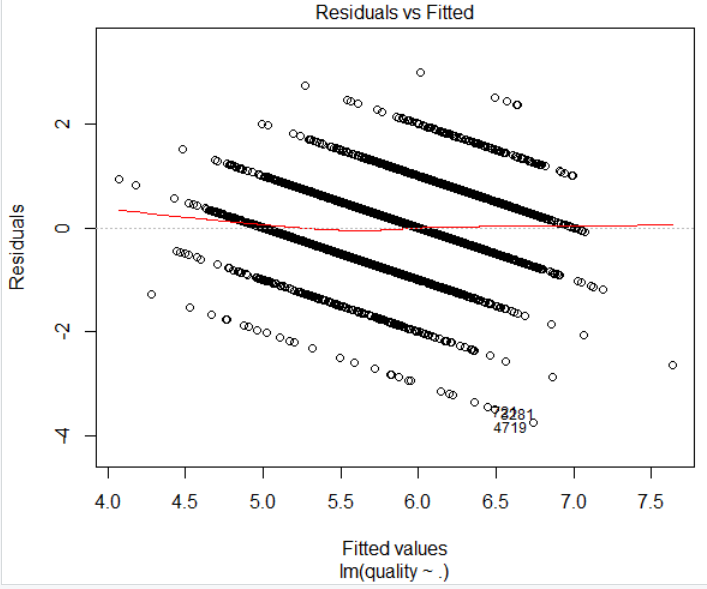
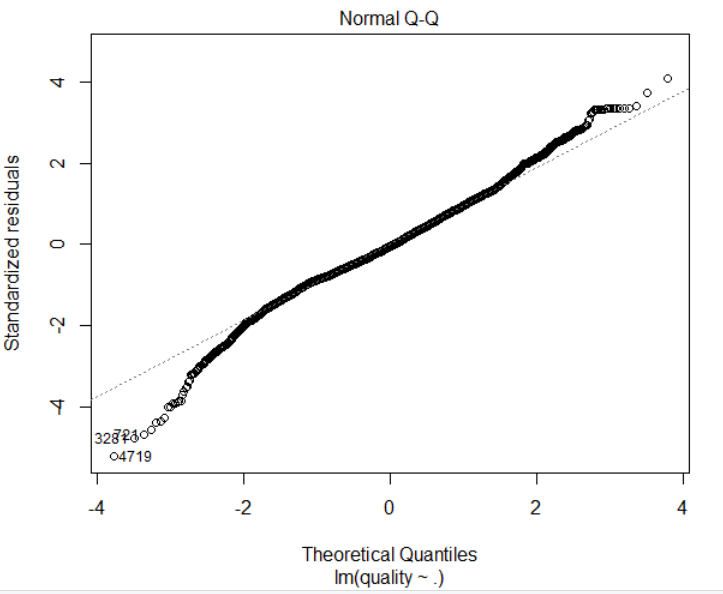
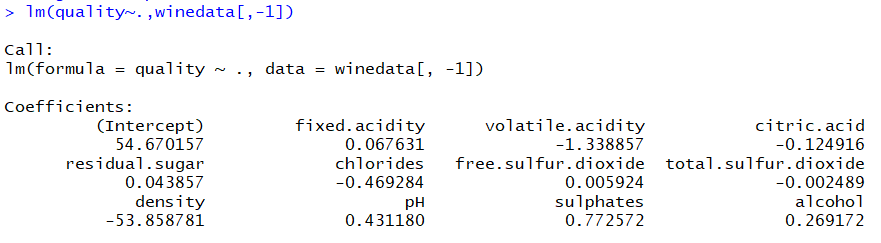


Fig 7: Linear Regression

The above linear regression is stated as a linear approach which could be helpful in modeling the relationship between scalar response and any of the explanatory variables.



**CONCLUSION**

1. Average pH value and quality of red wine and white wine (by using SQL)

The quality for both the red and white wine are calculated by using the Oracle SQL developer. It understood that the white wine has a higher average quality value when compared to the red wine.

The average pH value for both the red and white wine changes where after computing the values for both the red wine and white wine. It understood that white wine has a higher average pH value when compared to red wine.

1. Which factors profoundly influence the quality of the wine?

The factors that tend to influence the quality of wine highly is alcohol, because it is very evident from the correlation matrix that more is the percentage of alcohol content in wine then automatically the quality of fine also is high.

1. Are Free Sulphur dioxide and total Sulphur dioxide value of wine data related?

Free Sulphur dioxide and total Sulphur dioxide value of wine data are related where from the above scatter plot plotted above between the free Sulphur dioxide and total Sulphur dioxide. It understood that with no increase in free Sulphur dioxide the total Sulphur dioxide of white tends to increase to a higher value when compared to the total Sulphur dioxide of red wine.

1. What is the correlation of quality with other predictors?

Correlation of quality with other predictors changes, i.e., from the correlation matrix plotted above the quality attribute is highly correlated with alcohol content in wine in percentages whereas the quality changes, i.e., it tends to have a negative correlation with the other predictors as observed.

**EXPLAIN TERMS**

**R studio :**

R Studio is an open source integrated development environment which is used to organize, clear and analysis the data. The user can quickly view the graphs, data tables, output and R code at the same time [2].

**pH value :**

pH value is used to indicate whether the alcohol is acidic or basic or neutral. If pH value is greater than 7 then it is basic if the pH value is less than 7 then it is acidic if the pH value equal o 7 then it is pure water [3].

**Linear Regression :**

Linear Regression is the simplest form of regression. It derives the relationship between the independent and dependent variable [4].

**Oracle SQL Developer :**

Oracle SQL Developer is an open source which has an integrated development environment that makes the developer simplifies the development and management of the database both in traditional and cloud. The SQL Developer offer complete SQL application, a worksheet for script and queries, report interface, provide migration platform to the 3rd party database [5].

# References

|  |  |
| --- | --- |
| [1] | J. Reis, P. Cortez, F. Almeida, A. Cerderia and T. Matos, "Wine Quality Data Set," 07 October 2013. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Wine+Quality. |
| [2] | K. S. University, "Statistical & Qualitative Data Analysis Software: About R and RStudio," [Online]. Available: https://libguides.library.kent.edu/statconsulting/r. |
| [3] | T. D. Star, "What should be the pH value of drinking water?," [Online]. Available: https://www.thedailystar.net/health/what-should-be-the-ph-value-drinking-water-138382. |
| [4] | R-Bloggers, "Types of Regression," [Online]. Available: https://www.r-bloggers.com/scatterplot-matrices-in-r/. |
| [5] | Oracle, "Oracle SQL Developer," [Online]. Available: https://www.oracle.com/database/technologies/appdev/sql-developer.html. |

**DATASET LINK TO DOWNLOAD:**

kaggle datasets download -d rajyellow46/wine-quality