# AIT-580 DATA ANALYSIS PROJECT REPORT

## Deliverable 1:

The data set selected is Wine Quality data set [1], which is related to business domain.

The size of data set is 343 KB, but the storage required is 1 MB.

The input variables are based on physicochemical tests.

1. Fixed acidity: acids present in wine can be fixed or nonvolatile. (Ratio data type)

2. Volatile acidity: the higher amount of acetic acid gives vinegar taste. (Ratio data type)

3. Citric acid: this adds freshness and flavor to wine and found in small quantities. (Ratio data type)

4. Residual sugar: this is the amount of sugar left out after fermentation. (Ratio data type)

5. Chlorides: salt in wine. (Ratio data type)

6. Free Sulphur dioxide: this prevents in growth of microbes and wine oxidation. (Interval data type)

7. Total Sulphur dioxide: its quantity decides taste of wine. (Interval data type)

8. Density: water density depending on alcohol and sugar content. (Ratio data type)

9. pH: describes acidic and basic of wine. (Ratio data type)

10. Sulphates: this acts as antioxidant and antimicrobial. (Ratio data type)

11. Alcohol: it is the amount of alcohol present in wine. (Ratio data type)

Output variable based on sensory data

12. Quality: it is based on score between 0 and 10. (Interval data type)

The characteristic of data set is multivariate. Total number of attributes in it are 12 and number of instances present in it are 4898. There are no missing values. Characteristic of attribute is real.

Paulo Cortez owns the data set. He is an Associate Professor (with Habilitation) at the Department of Information Systems and Coordinator of the Information Systems and Technologies (IST) research group of Algoritmi Research Center, University of Minho.

Two datasets are included, related to red and white “Vinho Verde” wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests.

Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

“These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So, it could be interesting to test feature selection methods.” [2]

### Specific questions:

1. Which wine quality is good?

2. What makes wine quality good?

3. How wine is categorized? (poor, normal, excellent)

4. What makes wine price costlier?

I need Tableau and R Studio for my analysis.

The input variables are irrelevant.

## Deliverable 2:

### Data Preparation

There are two csv data files of wine quality with respect to “Red” and “White” wines. Both the data files are merged manually to a single csv file with comma (,) delimiter for the better analysis and visualizations. To differentiate between characteristics of red and white wines, a new column “wine type” is added to new csv file. Another new column “rating” is added for the quality scoring between 0 – 10, which is categorized as poor, normal and excellent.

- wine type – it represents red and white wines. (Nominal data)
- rating – it characterizes quality of wines as poor, normal and excellent. (Ordinal data)
  poor – if quality is less than 5
  normal – if quality is less than 7
  excellent – if quality greater or equal to 7

The data analysis is made using R, Tableau and SQL.

## Part 1 – A

### Section 1:

Following are the libraries used in R script for the data analysis.

library(tidyverse)

library(corrplot)

library(ggplot2)

The combined csv file “MixedWine” is imported into the R script with the below script.

```
rwwine <- read.csv("MixedWine.csv", header = TRUE)
```

```
names(rwwine)
```

```
##  [1] "ï..wine.type"         "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
```

```
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"              "rating"
```

Tukey's five number summary of the attributes is given as below:

```
summary(rwwine)
```

```
##  ï..wine.type fixed.acidity    volatile.acidity  citric.acid
##  Red  :1599   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  White:4898   1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500
##               Median : 7.000   Median :0.2900   Median :0.3100
##               Mean   : 7.215   Mean   :0.3397   Mean   :0.3186
##               3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900
##               Max.   :15.900   Max.   :1.5800   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  1.00
##  1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00
##  Median : 3.000   Median :0.04700   Median : 29.00
##  Mean   : 5.443   Mean   :0.05603   Mean   : 30.53
##  3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00
##  Max.   :65.800   Max.   :0.61100   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300
##  Median :118.0        Median :0.9949   Median :3.210   Median :0.5100
##  Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313
##  3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000
##  Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000
##     alcohol         quality            rating
##  Min.   : 8.00   Min.   :3.000   excellent:1277
##  1st Qu.: 9.50   1st Qu.:5.000   normal   :4974
##  Median :10.30   Median :6.000   poor     : 246
##  Mean   :10.49   Mean   :5.818
##  3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :14.90   Max.   :9.000
```

There were no missing values in the data. So, it made data analysis simpler.

```
sum(is.na(rwwine))
```

```
## [1] 0
```

To find a better correlation between attributes, correlation plot is plotted.

```
w <- rwwine[c(2:13)]
crw = cor(w)
corrplot(crw, method = "number")
```

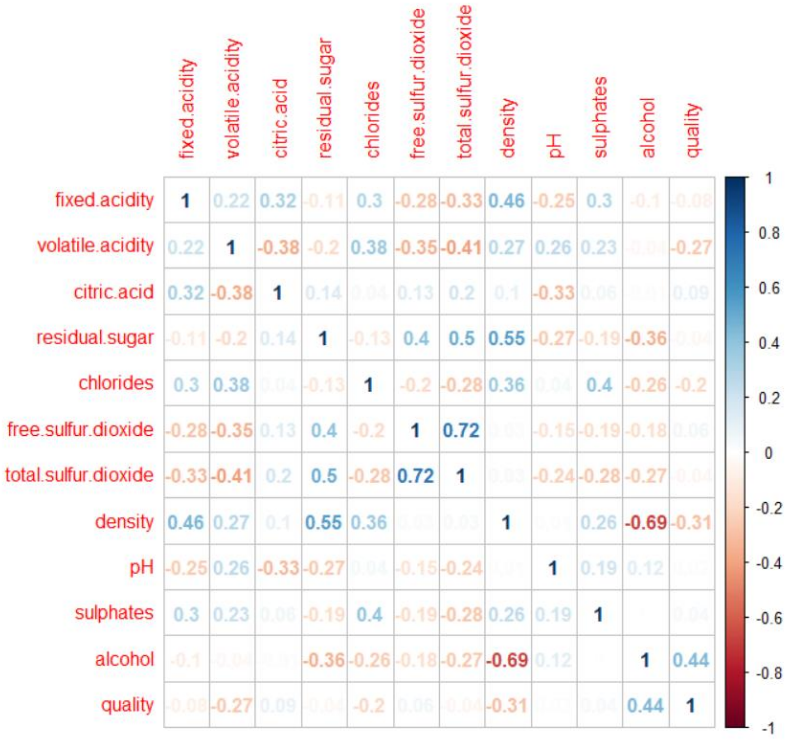



*Figure 1: Correlation matrix*

Scatter plots are plotted between the attributes that are showing a better or strong correlation.

```
plot(rwwine$pH, rwwine$fixed.acidity)
```

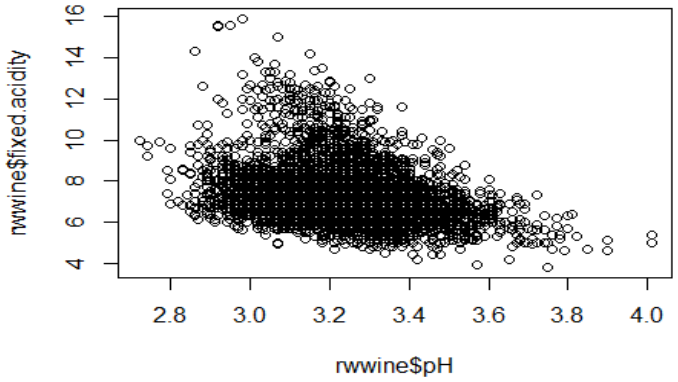



*Figure 2: Scatter plot between pH and fixed acidity*

```
plot(rwwine$total.sulfur.dioxide, rwwine$free.sulfur.dioxide)
```

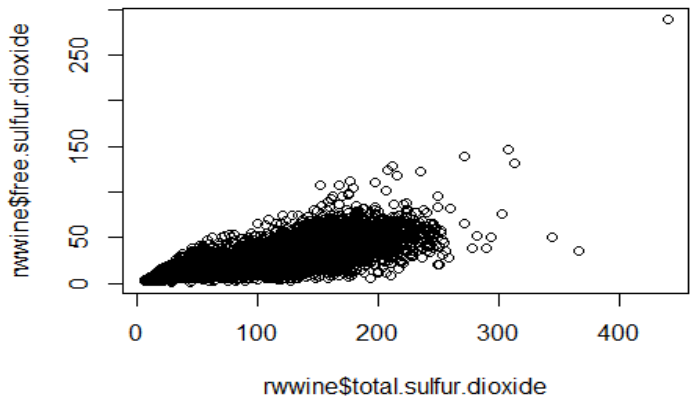



*Figure 3: Scatter plot between free sulfur dioxide and total sulfur dioxide*

```
plot(rwwine$alcohol, rwwine$density)
```

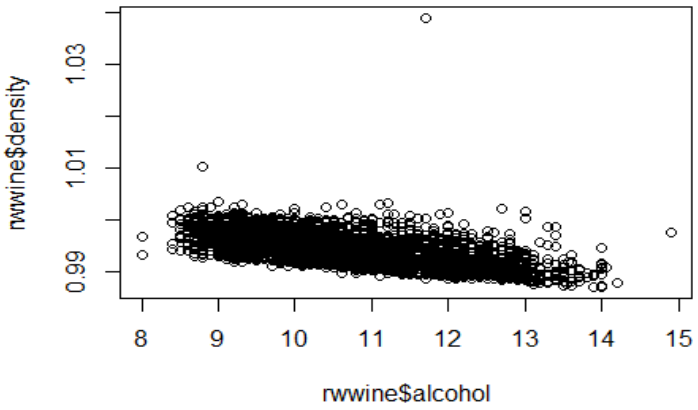



*Figure 4: Scatter plot between alcohol and density*

```
plot(rwwine$quality, rwwine$alcohol)
```

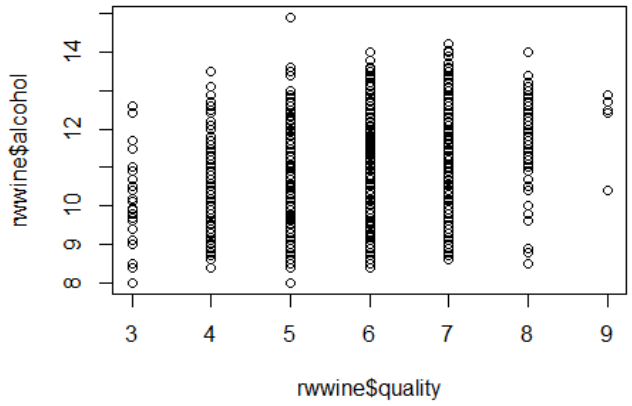



*Figure 5: Scatter plot between quality and alcohol*

```
plot(rwwine$quality, rwwine$citric.acid)
```

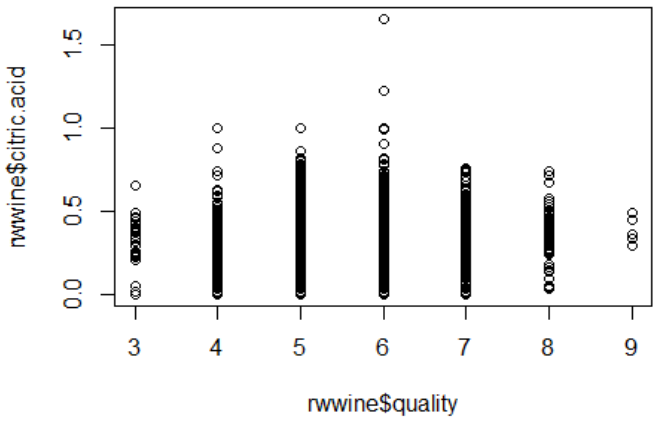



*Figure 6: Scatter plot between quality and citric acid*

Simple data analysis and visualizations of data made using bar plot and histogram.

```
ggplot(rwwine, aes(x = quality)) + geom_bar(color="black", fill="green") +
  labs(x = "Red and White wine quality", y = "Total number of smaples", title
= "Bar plot for total count of quality of wine")
```

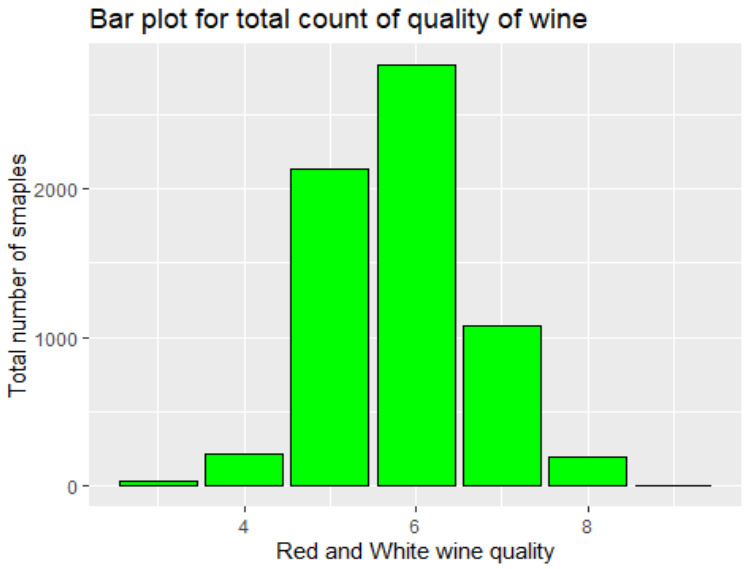



*Figure 7: Bar plot shows total count of quality of wines*

```
ggplot(rwwine, aes(x = rwwine$rating)) + geom_bar(color="black", fill="orange
") +
  labs(x = "Ratings of wine", y = "Total count of each rating for wine", titl
e = "Bar plot for Ratings")
```

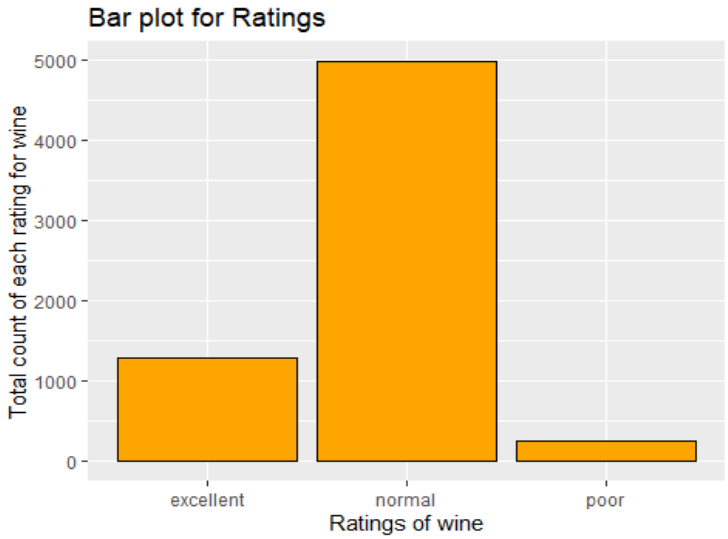



*Figure 8: Bar plot shows count of each ratings of wines*

```
ggplot(data = rwwine, aes(x=rwwine$quality)) + geom_histogram(fill="blue",bin
width = 1) +
  labs(x = "Quality of wines", y = "Total count of each type of quality", tit
le = "Histogram for the quality of wines")
```

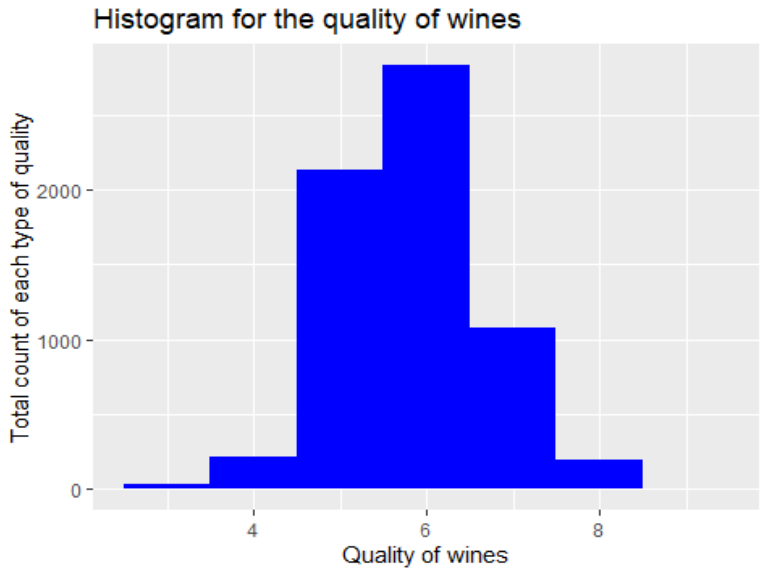



*Figure 9: Histogram shows count of each quality of wines*

```
ggplot(data = rwwine) + geom_bar(aes(x=rwwine$ï..wine.type, fill = as.factor(
quality))) +
  labs(x = "Types of wines", y = "Total count of each type of quality", title
= "Stacked bar plot for types of quality of wines")
```

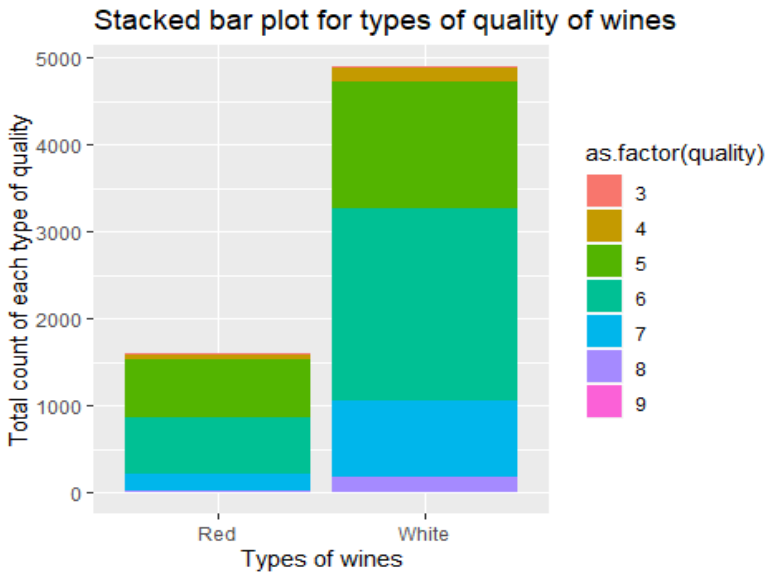



*Figure 10: Stacked bar plot shows count of quality of red and white wine*

**Section 2:**

From the below two graphs, red and white wine quality analysis is made using bar plot.

```
ggplot(rwwine[rwwine$ï..wine.type == "Red",], aes(x = quality)) + geom_bar(fill
=              "Red",              color              =              "black")+
```

```
  labs(x = "Quality of Red wine", y = "Total count of each quality",title =
"Bar plot shows total count of red wine")
```

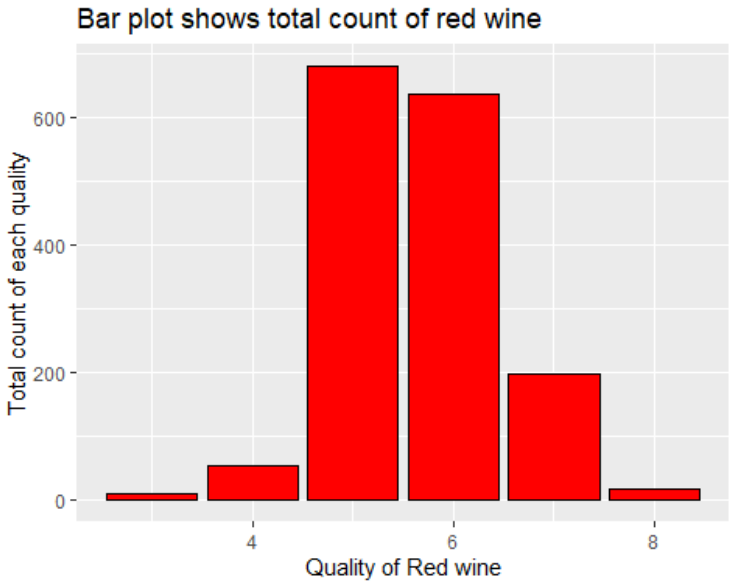



*Figure 11: Bar plot shows count of quality of red wine*

From Figure 11, it is observed that most of the red wine samples quality is 5.

```
ggplot(rwwine[rwwine$ï..wine.type == "White",], aes(x = quality)) + geom_bar(
fill = "Yellow", color = "black")+
  labs(x = "Quality of White wine", y = "Total count of each quality",title =
"Bar plot shows total count of white wine")
```

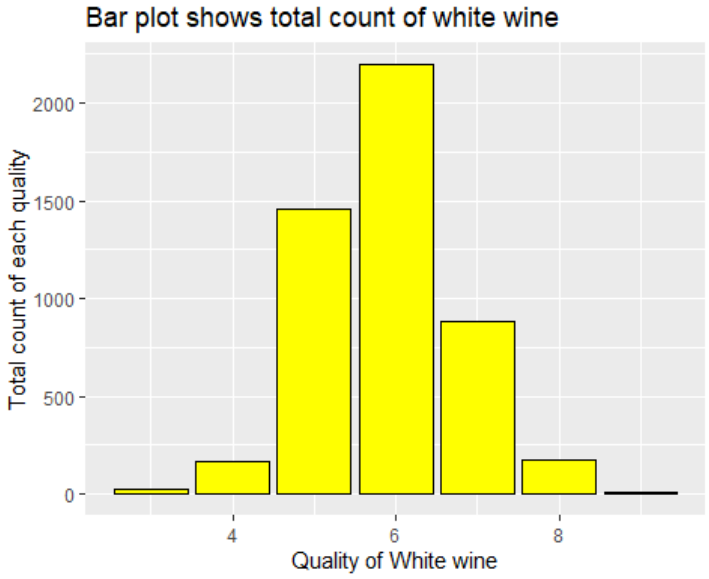



*Figure 12: Bar plot shows count of quality of white wine*

From Figure 12, it is observed that most of the white wine samples quality is 6.

```
ggplot(rwwine, aes(x = factor(rwwine$quality, levels = c(0,1,2,3,4,5,6,7,8,9,10)), y = rwwine$alcohol)) + geom_boxplot(color="black", fill="green") + labs(x="Red & White wine quality", y="Alcohol amount", title = "Quality vs Alcohol")
```

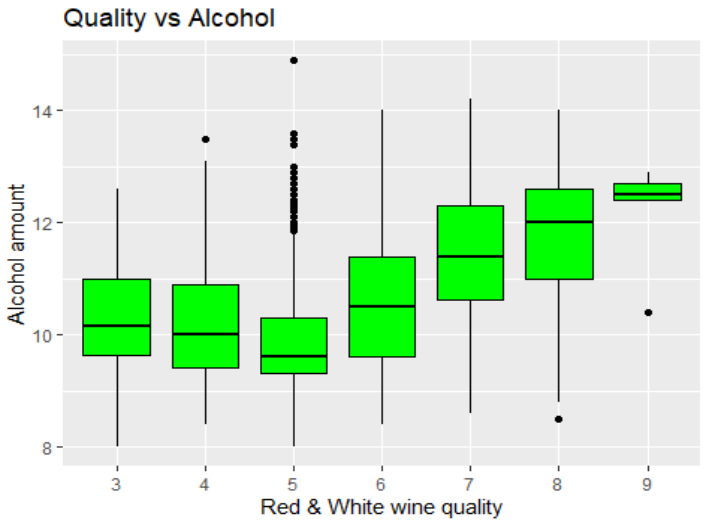



*Figure 13: Box plot for quality and alcohol*

Figure 13 depicts that as the alcohol content in both red and white wine increases, its quality also increases.

```
ggplot(rwwine, aes(x = factor(rwwine$quality, levels = c(0,1,2,3,4,5,6,7,8,9,10)), y = rwwine$citric.acid)) + geom_boxplot(color="black", fill="blue") + labs(x="Red & White wine quality", y="Citric acid", title = "Quality vs Citric acid")
```

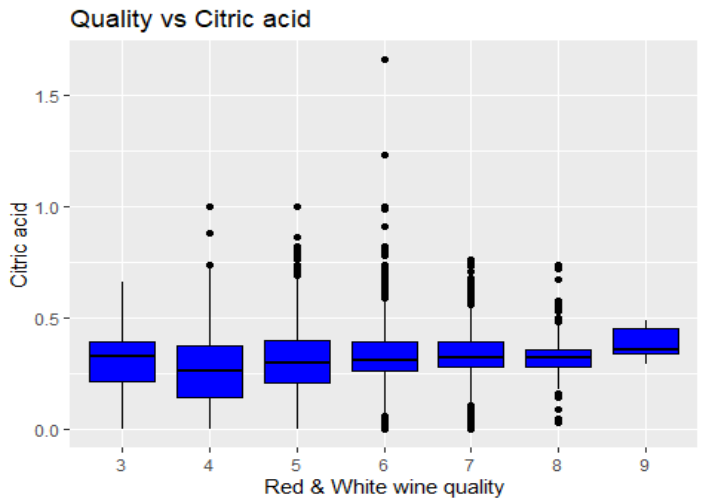



*Figure 14: Box plot for quality and citric acid*

From Figure 14, it is observed that quality of both red and white wine also increases with the increase in citric acid. Higher quality wines contain more citric acid.

```
ggplot(rwwine, aes(x = factor(rwwine$quality, levels = c(0,1,2,3,4,5,6,7,8,9,10)), y = rwwine$residual.sugar)) +
  geom_boxplot(color="black", fill="Pink") + labs(x="Red & White wine quality", y="Residual Sugar", title = "Quality vs Residual Sugar")
```

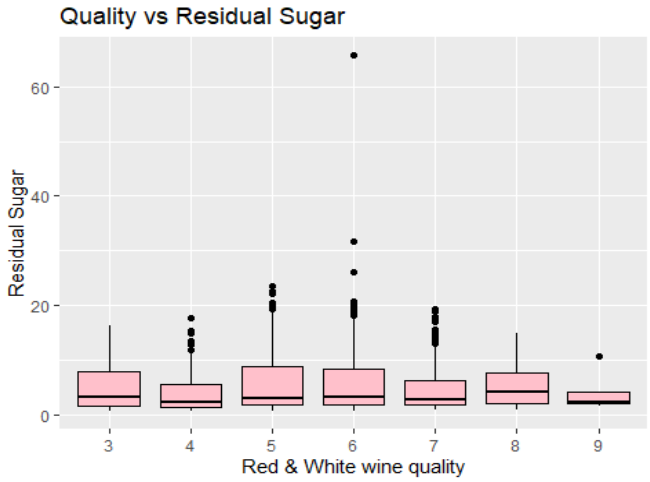



*Figure 15: Box plot for quality and residual sugar*

Figure 15 shows residual sugar is less in higher quality wines than lower quality ones.

```
ggplot(data =rwwine[rwwine$ï..wine.type == "White",], aes(x = factor(quality, levels = c(0,1,2,3,4,5,6,7,8,9,10)), y = total.sulfur.dioxide)) +
  geom_boxplot(color="black", fill="yellow") + labs(x="White wine quality", y="Total sulphur dioxide", title = "Quality vs Total Sulphur dioxide")
```

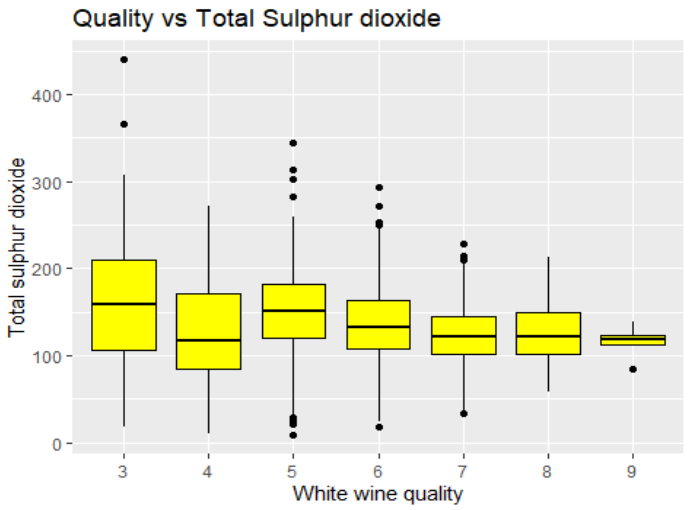



*Figure 16: Box plot for quality and total sulfur dioxide of white wine*

```
ggplot(data =rwwine[rwwine$ï..wine.type == "Red",], aes(x = factor(quality, levels = c(0,1,2,3,4,5,6,7,8,9,10)), y = total.sulfur.dioxide)) +
  geom_boxplot(color="black", fill="Yellow") + labs(x="Red wine quality", y="Total sulphur dioxide", title = "Quality vs Total Sulphur dioxide")
```

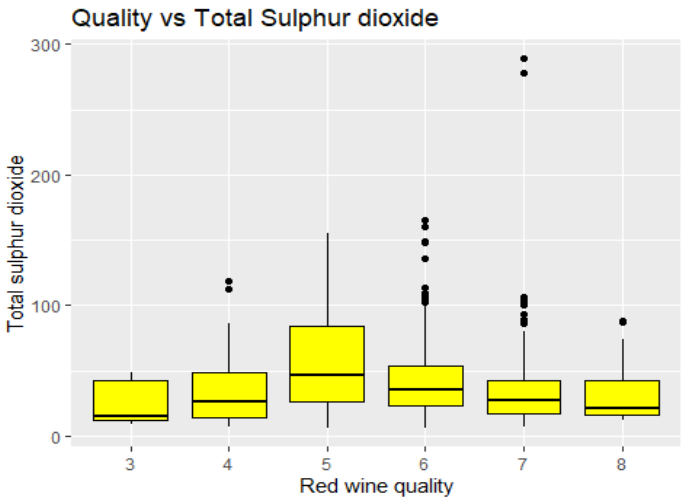



*Figure 17: Box plot for quality and total sulfur dioxide in red wine*

From both the Figure 16 and Figure 17, it is observable that total sulfur dioxide is less in high quality wines of both red and white wine. Lower quality wines have high total sulfur dioxide.

```
ggplot(data =rwwine[rwwine$ï..wine.type == "White",], aes(x = factor(quality, levels = c(0,1,2,3,4,5,6,7,8,9,10)), y = chlorides)) + geom_boxplot(color="black", fill="Purple") + labs(x="White wine quality", y="Chlorides", title = "Quality vs Chlorides")
```

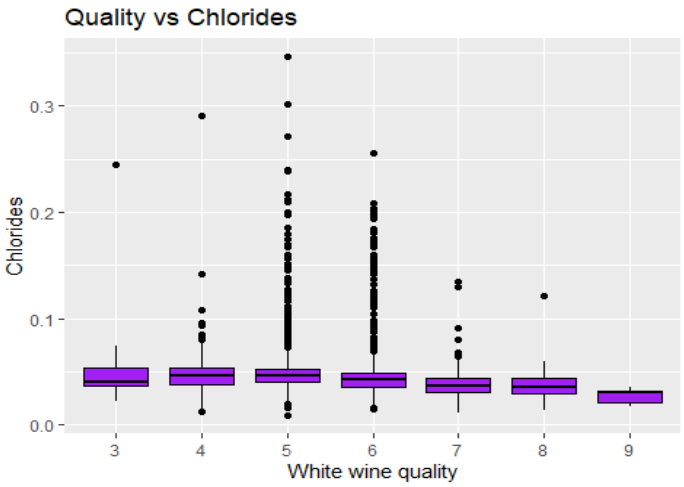



*Figure 18: Box plot for quality and chlorides of white wine*

```
ggplot(data =rwwine[rwwine$ï..wine.type == "Red",], aes(x = factor(quality, levels = c(0,1,2,3,4,5,6,7,8,9,10)), y = chlorides)) + geom_boxplot(color="black", fill="Purple") + labs(x="Red wine quality factor", y="Chlorides", title = "Quality vs Chlorides")
```

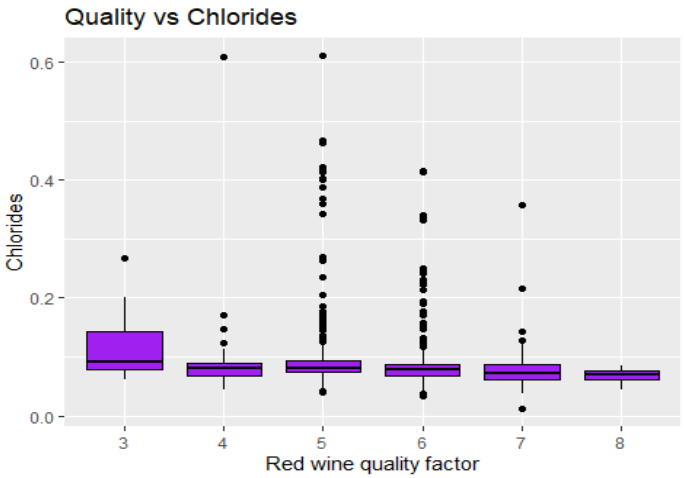



*Figure 19: Box plot for quality and chlorides of red wine*

Figure 18 and Figure 19 show that higher quality of red and white wines have lesser amount of chlorides than lower quality wines.

**Section 3:**

Fixed acidity gradually decreases from lower to higher quality of red and white wines, as shown in Figure 20.

```
ggplot(data =rwwine, aes(x = factor(quality, levels = c(0,1,2,3,4,5,6,7,8,9,10)), y = fixed.acidity)) + geom_boxplot(color="black", fill="Green") + labs(x="Red & White wine quality", y="Fixed Acidity", title = "Quality vs Fixed Acidity")
```

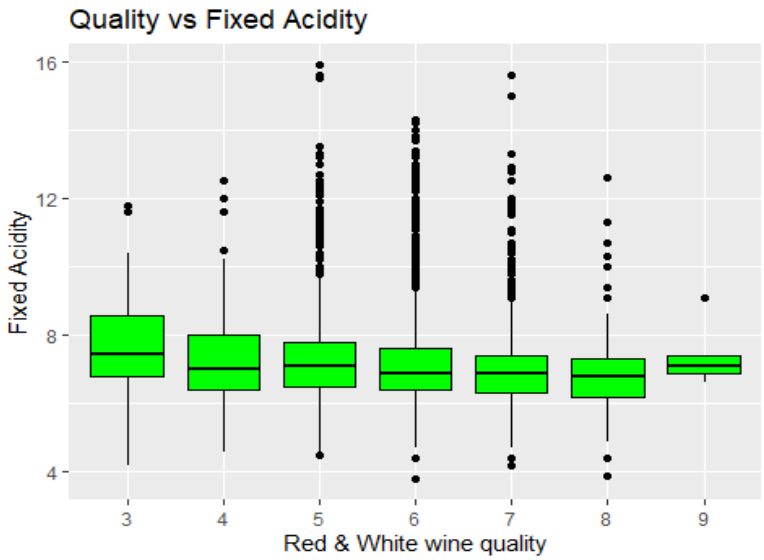



*Figure 20: Box plot for quality and fixed acidity of white wine*

**Section 4:**

**1**. A linear regression model is built with the alcohol as target variable and quality as predictor variable. The p value is less than 2.2e-16, which is good. Multiple R squared value is 0.1974, which means that 19.74% of the variability of target variable is explained by the predictor variable. However, the value closer to 100%, the better the correlation is explained by the model.

```
plot(x= rwwine$quality, y= rwwine$alcohol, main = "Linear Regression Model")
mod <- lm(rwwine$alcohol ~ rwwine$quality)
summary(mod)

##
## Call:
## lm(formula = rwwine$alcohol ~ rwwine$quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3157 -0.7952 -0.1952  0.7048  4.9048
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.96086    0.08933   77.92   <2e-16 ***
## rwwine$quality  0.60686    0.01518   39.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.069 on 6495 degrees of freedom
## Multiple R-squared:  0.1974, Adjusted R-squared:  0.1973
## F-statistic:  1598 on 1 and 6495 DF,  p-value: < 2.2e-16

abline(mod, col = "Blue", lwd = 3)
```

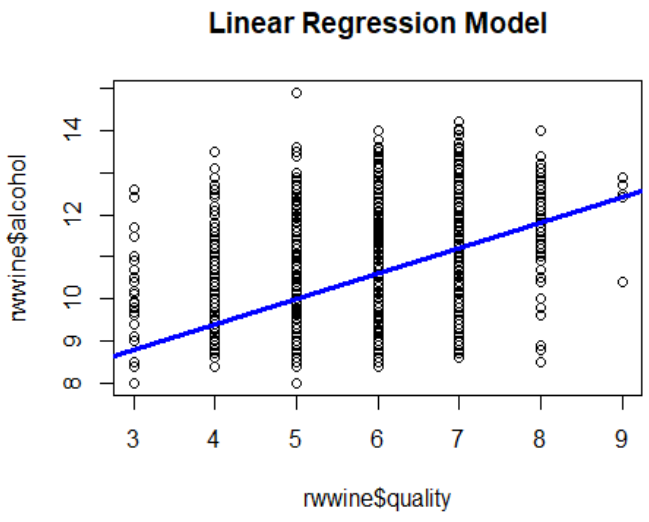
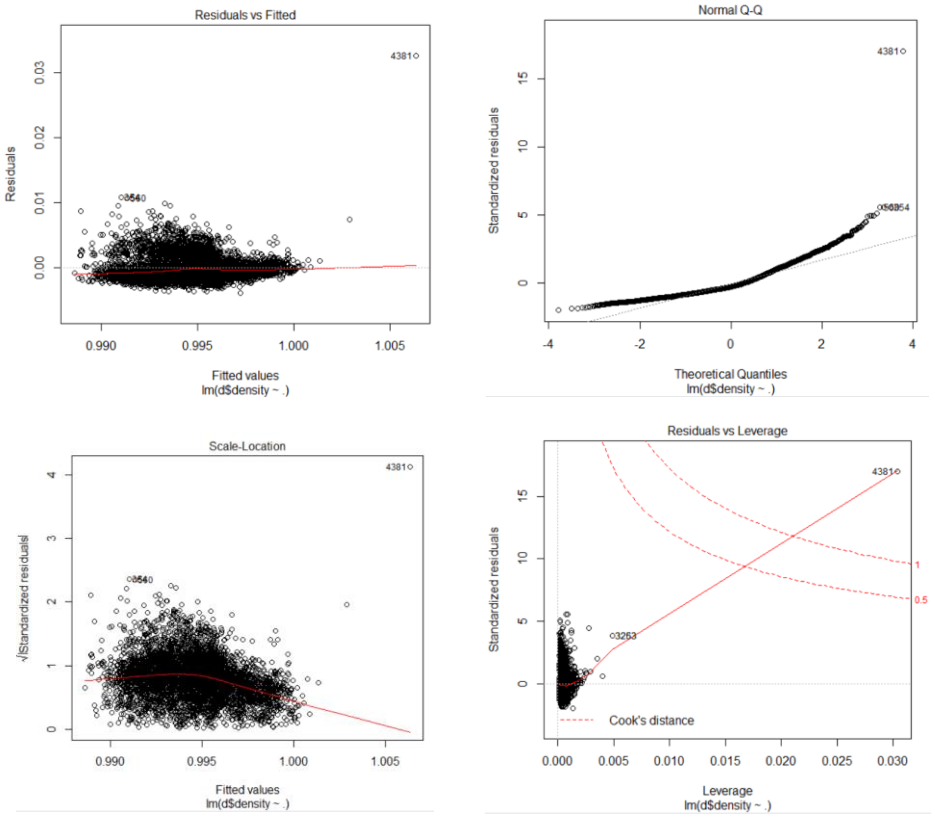



*Figure 21: Scatter plot showing linear regression for quality and alcohol of both the wines*

**2**. Another linear regression model built with density as target variable and alcohol, residual sugar as predictor variables. P value obtained is less than 2.2e-16. And multiple R squared value is 0.5789, which means target variable can be estimated with predictors of an accuracy of 57.89%.

```
d<-rwwine[c(5,9,12)]
r<-lm(d$density~., data =d)
plot(r)
```




```
summary(r)
```

```
Call:
lm(formula = d$density ~ ., data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.003894 -0.001343 -0.000543  0.000973  0.032632

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.008e+00  2.411e-04 4182.63   <2e-16 ***
residual.sugar  2.212e-04  5.439e-06   40.68   <2e-16 ***
alcohol        -1.409e-03  2.170e-05  -64.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001946 on 6494 degrees of freedom
Multiple R-squared:  0.5789,   Adjusted R-squared:  0.5788
F-statistic:  4464 on 2 and 6494 DF,  p-value: < 2.2e-16
```

**Section 5:**

An alternative hypothesis test is conducted for the variable alcohol with one-sided alternative. It is assumed that the mean value of alcohol content in wines is less than 11. Confidence interval is 95% in this case. The respective alpha (a) value would be 0.05. From obtained results of t.test(), it is noticed that p value (2.2e-16) is less than alpha (0.05). So, there is enough evidence to reject the null hypothesis. In this case, the null hypothesis is rejected. But the alternative hypothesis is failed to reject, because we assumed that the mean value (11) would be less than the true mean value (10.4918), which is true.

```
boxplot(rwwine$alcohol)
```

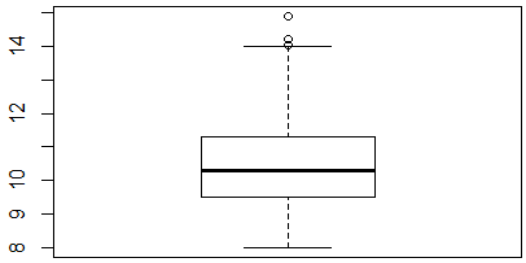



*Figure 22: Box plot for alcohol content in wine*

```
t.test(rwwine$alcohol, alternative = "less", mu = 11)

##
##  One Sample t-test
##
## data:  rwwine$alcohol
## t = -34.344, df = 6496, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 11
## 95 percent confidence interval:
##      -Inf 10.51614
## sample estimates:
## mean of x
##   10.4918
```

## Basic SQL queries of the data:

SQL queries are executed using PostgreSQL. In order to import the data set in PostgreSQL, first we need to create the schema. Schema is created using the below SQL query.

*create schema myschema;*

Once the schema is created, we need to create the table inside the created schema.

*create table myschema.wine(*

*winetype varchar(50), fixedacidity numeric, volatileacidity numeric, citricacid numeric,*

*residualsugar numeric, chlorides numeric, freesulfurdioxide numeric,*

*totalsulfurdioxide numeric, density numeric, pH numeric, sulphates numeric,*

*alcohol numeric, quality numeric, rating varchar(50));*

The above SQL query creates the table “wine” in schema “myschema” with the attributes mentioned in it. Once table is created csv file can be imported into the table with the delimiter comma (,) from import option.

Below query prints the entire content of the table.

*select * from myschema.wine;*

Output pane

Data Output | Explain | Messages | History

| | winetype character varying(50) | fixedacidity numeric | volatileacidity numeric | citricacid numeric | residualsugar numeric | chlorides numeric | freesulfurdioxide numeric | totalsulfurdioxide numeric | density numeric | ph numeric | sulphates numeric | alcohol numeric | quality numeric | rating character varying(50) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Red | 7.4 | 0.7 | 0 | 1.9 | 0.076 | 11 | 34 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 | normal |
| 2 | Red | 7.8 | 0.88 | 0 | 2.6 | 0.098 | 25 | 67 | 0.9968 | 3.2 | 0.68 | 9.8 | 5 | normal |
| 3 | Red | 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15 | 54 | 0.997 | 3.26 | 0.65 | 9.8 | 5 | normal |
| 4 | Red | 11.2 | 0.28 | 0.56 | 1.9 | 0.075 | 17 | 60 | 0.998 | 3.16 | 0.58 | 9.8 | 6 | normal |
| 5 | Red | 7.4 | 0.7 | 0 | 1.9 | 0.076 | 11 | 34 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 | normal |
| 6 | Red | 7.4 | 0.66 | 0 | 1.8 | 0.075 | 13 | 40 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 | normal |
| 7 | Red | 7.9 | 0.6 | 0.06 | 1.6 | 0.069 | 15 | 59 | 0.9964 | 3.3 | 0.46 | 9.4 | 5 | normal |
| 8 | Red | 7.3 | 0.65 | 0 | 1.2 | 0.065 | 15 | 21 | 0.9946 | 3.39 | 0.47 | 10 | 7 | excellent |
| 9 | Red | 7.8 | 0.58 | 0.02 | 2 | 0.073 | 9 | 18 | 0.9968 | 3.36 | 0.57 | 9.5 | 7 | excellent |
| 10 | Red | 7.5 | 0.5 | 0.36 | 6.1 | 0.071 | 17 | 102 | 0.9978 | 3.35 | 0.8 | 10.5 | 5 | normal |
| 11 | Red | 6.7 | 0.58 | 0.08 | 1.8 | 0.097 | 15 | 65 | 0.9959 | 3.28 | 0.54 | 9.2 | 5 | normal |
| 12 | Red | 7.5 | 0.5 | 0.36 | 6.1 | 0.071 | 17 | 102 | 0.9978 | 3.35 | 0.8 | 10.5 | 5 | normal |
| 13 | Red | 5.6 | 0.615 | 0 | 1.6 | 0.089 | 16 | 59 | 0.9943 | 3.58 | 0.52 | 9.9 | 5 | normal |
| 14 | Red | 7.8 | 0.61 | 0.29 | 1.6 | 0.114 | 9 | 29 | 0.9974 | 3.26 | 1.56 | 9.1 | 5 | normal |
| 15 | Red | 8.9 | 0.62 | 0.18 | 3.8 | 0.176 | 52 | 145 | 0.9986 | 3.16 | 0.88 | 9.2 | 5 | normal |
| 16 | Red | 8.9 | 0.62 | 0.19 | 3.9 | 0.17 | 51 | 148 | 0.9986 | 3.17 | 0.93 | 9.2 | 5 | normal |
| 17 | Red | 8.5 | 0.28 | 0.56 | 1.8 | 0.092 | 35 | 103 | 0.9969 | 3.3 | 0.75 | 10.5 | 7 | excellent |
| 18 | Red | 8.1 | 0.56 | 0.28 | 1.7 | 0.368 | 16 | 56 | 0.9968 | 3.11 | 1.28 | 9.3 | 5 | normal |
| 19 | Red | 7.4 | 0.59 | 0.08 | 4.4 | 0.086 | 6 | 29 | 0.9974 | 3.38 | 0.5 | 9 | 4 | poor |
| 20 | Red | 7.9 | 0.32 | 0.51 | 1.8 | 0.341 | 17 | 56 | 0.9969 | 3.04 | 1.08 | 9.2 | 6 | normal |
| 21 | Red | 8.9 | 0.22 | 0.48 | 1.8 | 0.077 | 29 | 60 | 0.9968 | 3.39 | 0.53 | 9.4 | 6 | normal |
| 22 | Red | 7.6 | 0.39 | 0.31 | 2.3 | 0.082 | 23 | 71 | 0.9982 | 3.52 | 0.65 | 9.7 | 5 | normal |
| 23 | Red | 7.9 | 0.43 | 0.21 | 1.6 | 0.106 | 10 | 37 | 0.9966 | 3.17 | 0.91 | 9.5 | 5 | normal |
| 24 | Red | 8.5 | 0.49 | 0.11 | 2.3 | 0.084 | 9 | 67 | 0.9968 | 3.17 | 0.53 | 9.4 | 5 | normal |
| 25 | Red | 6.9 | 0.4 | 0.14 | 2.4 | 0.085 | 21 | 40 | 0.9968 | 3.43 | 0.63 | 9.7 | 6 | normal |
| 26 | Red | 6.3 | 0.39 | 0.16 | 1.4 | 0.08 | 11 | 23 | 0.9955 | 3.34 | 0.56 | 9.3 | 5 | normal |
| 27 | Red | 7.6 | 0.41 | 0.24 | 1.8 | 0.08 | 4 | 11 | 0.9962 | 3.28 | 0.59 | 9.5 | 5 | normal |
| 28 | Red | 7.9 | 0.43 | 0.21 | 1.6 | 0.106 | 10 | 37 | 0.9966 | 3.17 | 0.91 | 9.5 | 5 | normal |

OK. Unix Ln

Below query gives the average value of alcohol.

*select avg(alcohol) from myschema.wine;*

```
sulphates numeric,
alcohol numeric,
quality numeric,
rating varchar(50));

select * from myschema.wine;

select avg(alcohol) from myschema.wine;

select min(freesulfurdioxide) from myschema.wine;
```

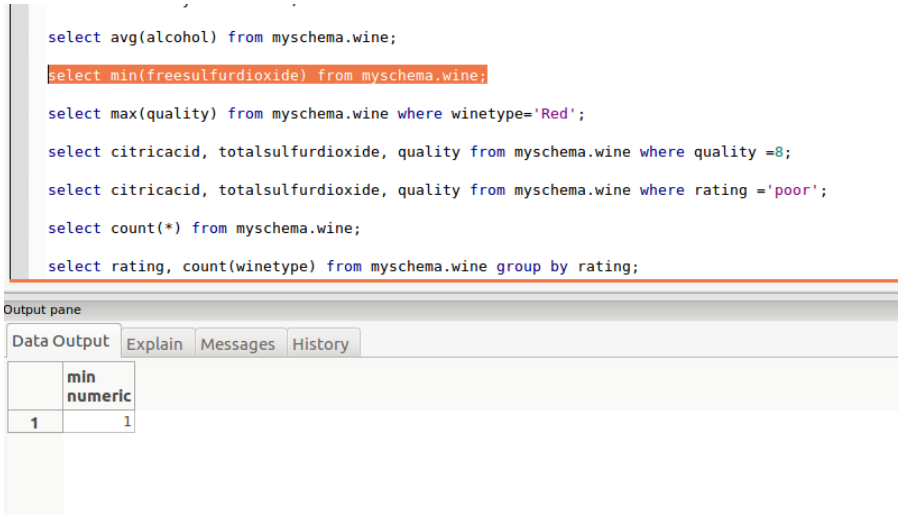
Output pane

Data Output | Explain | Messages | History

| | avg numeric |
|---|---|
| 1 | 10.4918008311494536 |

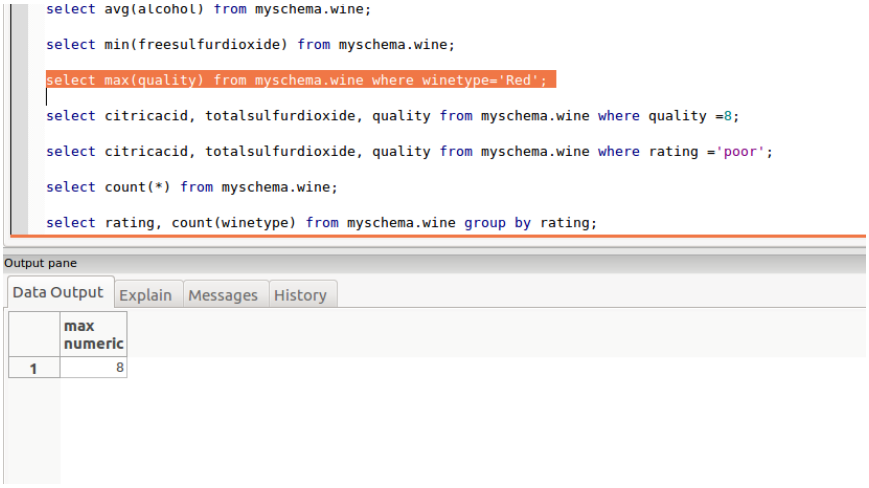
Below query gives minimum value of free sulfur dioxide.

*select min(freesulfurdioxide) from myschema.wine;*



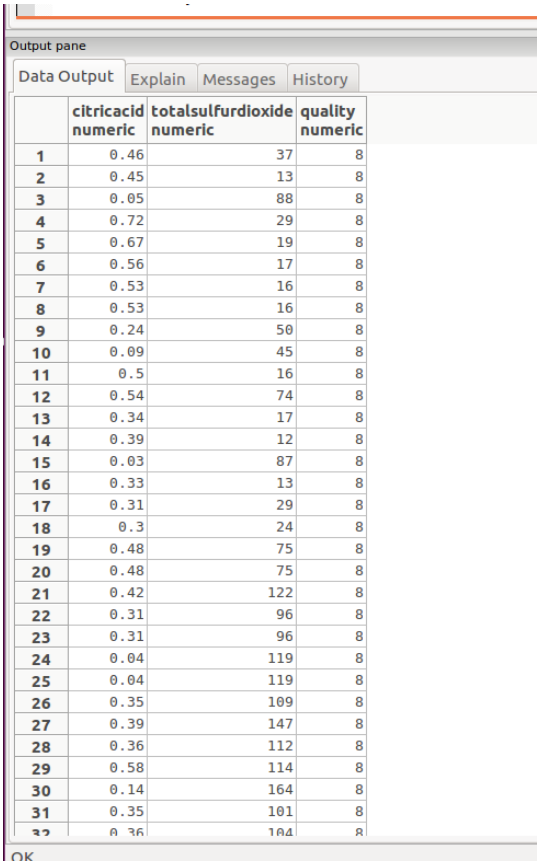

Below query prints the maximum value of quality of red wine.

*select max(quality) from myschema.wine where winetype='Red';*




Below query prints citric acid, total sulfur dioxide and quality from table whose quality is 8.

*select citricacid, totalsulfurdioxide, quality from myschema.wine where quality =8;*



Output pane

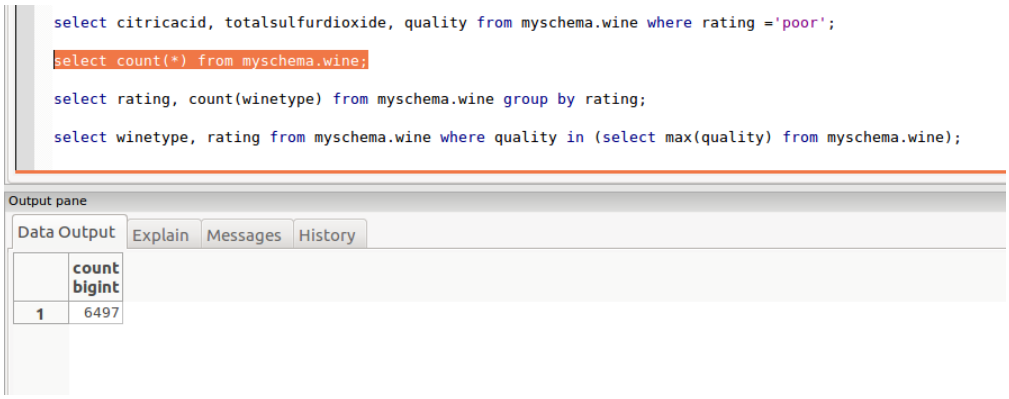
Data Output | Explain | Messages | History

| | citricacid numeric | totalsulfurdioxide numeric | quality numeric |
|---|---|---|---|
| 1 | 0.46 | 37 | 8 |
| 2 | 0.45 | 13 | 8 |
| 3 | 0.05 | 88 | 8 |
| 4 | 0.72 | 29 | 8 |
| 5 | 0.67 | 19 | 8 |
| 6 | 0.56 | 17 | 8 |
| 7 | 0.53 | 16 | 8 |
| 8 | 0.53 | 16 | 8 |
| 9 | 0.24 | 50 | 8 |
| 10 | 0.09 | 45 | 8 |
| 11 | 0.5 | 16 | 8 |
| 12 | 0.54 | 74 | 8 |
| 13 | 0.34 | 17 | 8 |
| 14 | 0.39 | 12 | 8 |
| 15 | 0.03 | 87 | 8 |
| 16 | 0.33 | 13 | 8 |
| 17 | 0.31 | 29 | 8 |
| 18 | 0.3 | 24 | 8 |
| 19 | 0.48 | 75 | 8 |
| 20 | 0.48 | 75 | 8 |
| 21 | 0.42 | 122 | 8 |
| 22 | 0.31 | 96 | 8 |
| 23 | 0.31 | 96 | 8 |
| 24 | 0.04 | 119 | 8 |
| 25 | 0.04 | 119 | 8 |
| 26 | 0.35 | 109 | 8 |
| 27 | 0.39 | 147 | 8 |
| 28 | 0.36 | 112 | 8 |
| 29 | 0.58 | 114 | 8 |
| 30 | 0.14 | 164 | 8 |
| 31 | 0.35 | 101 | 8 |
| 32 | 0.36 | 104 | 8 |

OK

Below query prints the total count of observations of the table.

*select count(*) from myschema.wine;*



```
select citricacid, totalsulfurdioxide, quality from myschema.wine where rating ='poor';

select count(*) from myschema.wine;

select rating, count(winetype) from myschema.wine group by rating;

select winetype, rating from myschema.wine where quality in (select max(quality) from myschema.wine);
```

Output pane

Data Output | Explain | Messages | History

| | count bigint |
|---|---|
| 1 | 6497 |

Below query prints citric acid, total sulfur dioxide and quality from the table whose rating is poor.

*select citricacid, totalsulfurdioxide, quality from myschema.wine where rating ='poor';*

Output pane

Data Output | Explain | Messages | History

| | citricacid numeric | totalsulfurdioxide numeric | quality numeric |
|---|---|---|---|
| 1 | 0.08 | 29 | 4 |
| 2 | 0.09 | 19 | 4 |
| 3 | 0.3 | 46 | 4 |
| 4 | 0.15 | 65 | 4 |
| 5 | 0.26 | 43 | 4 |
| 6 | 0.2 | 119 | 4 |
| 7 | 0.04 | 85 | 4 |
| 8 | 1 | 69 | 4 |
| 9 | 0.02 | 20 | 4 |
| 10 | 0.03 | 42 | 4 |
| 11 | 0.03 | 8 | 4 |
| 12 | 0.06 | 31 | 4 |
| 13 | 0.36 | 55 | 4 |
| 14 | 0.04 | 67 | 4 |
| 15 | 0 | 61 | 4 |
| 16 | 0.49 | 49 | 4 |
| 17 | 0.66 | 47 | 3 |
| 18 | 0.49 | 16 | 3 |
| 19 | 0.49 | 47 | 4 |
| 20 | 0.24 | 14 | 4 |
| 21 | 0.27 | 23 | 4 |
| 22 | 0.22 | 86 | 4 |
| 23 | 0.01 | 14 | 4 |
| 24 | 0.02 | 13 | 4 |
| 25 | 0 | 14 | 3 |
| 26 | 0.48 | 84 | 4 |
| 27 | 0.04 | 14 | 4 |
| 28 | 0.1 | 12 | 4 |
| 29 | 0.24 | 7 | 4 |
| 30 | 0.07 | 9 | 4 |
| 31 | 0.42 | 48 | 3 |
| 32 | 0.44 | 51 | 4 |

Below query prints rating and total count of each rating.

*select rating, count(winetype) from myschema.wine group by rating;*

```
select citricacid, totalsulfurdioxide, quality from myschema.wine where rating ='poor';

select count(*) from myschema.wine;

select rating, count(winetype) from myschema.wine group by rating;

select winetype, rating from myschema.wine where quality in (select max(quality) from myschema.wine);
```

Output pane

Data Output | Explain | Messages | History

| | rating character varying(50) | count bigint |
|---|---|---|
| 1 | poor | 246 |
| 2 | normal | 4974 |
| 3 | excellent | 1277 |

Below query prints wine type and its rating whose quality is maximum.

*select winetype, rating from myschema.wine where quality in (select max(quality) from myschema.wine);*

```
select count(*) from myschema.wine;

select rating, count(winetype) from myschema.wine group by rating;

select winetype, rating from myschema.wine where quality in (select max(quality) from myschema.wine);
```

Output pane

Data Output | Explain | Messages | History

| | winetype character varying(50) | rating character varying(50) |
|---|---|---|
| 1 | White | excellent |
| 2 | White | excellent |
| 3 | White | excellent |
| 4 | White | excellent |
| 5 | White | excellent |

## Part 1 – B

The wine data set visualizations are made with the help of Tableau software.

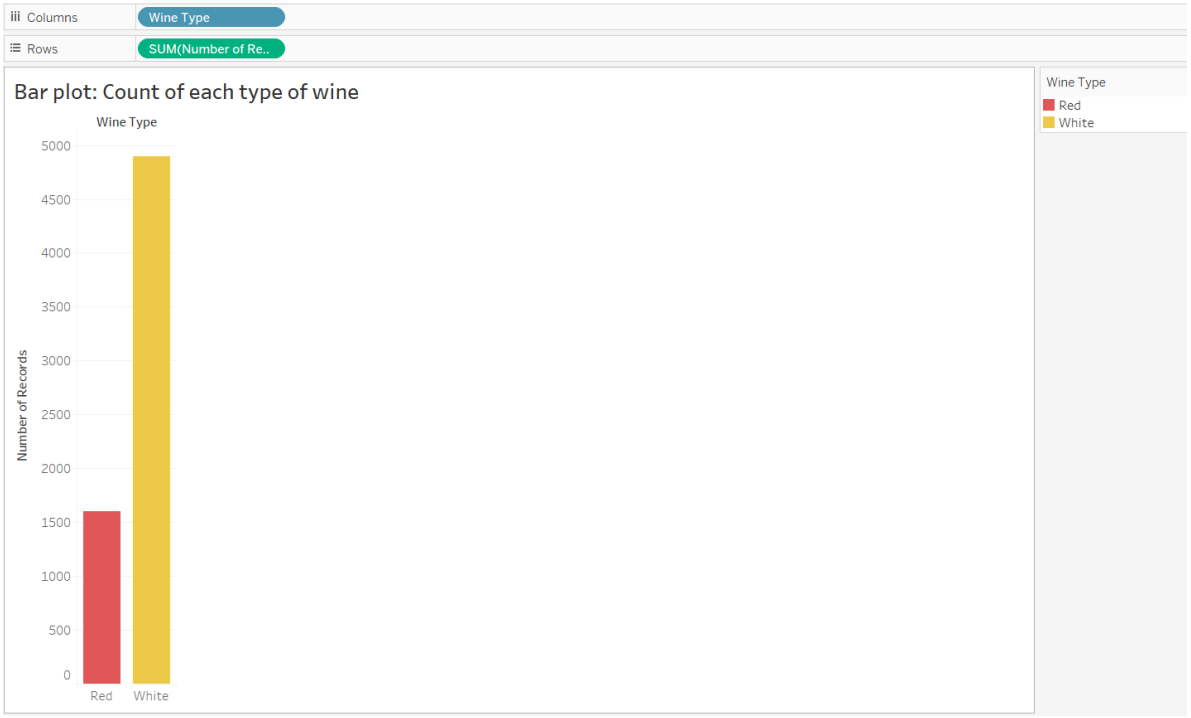



*Figure 23: Bar plot shows count of records of each wine*

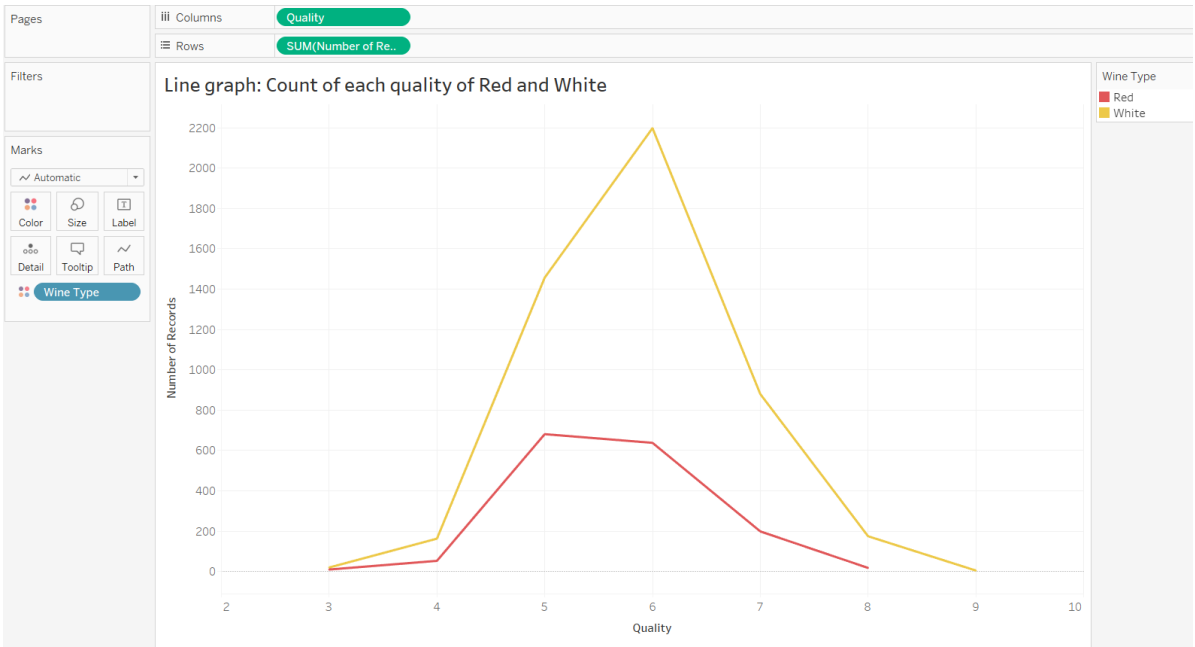



*Figure 24: Line graph shows count of quality of red and white wine*

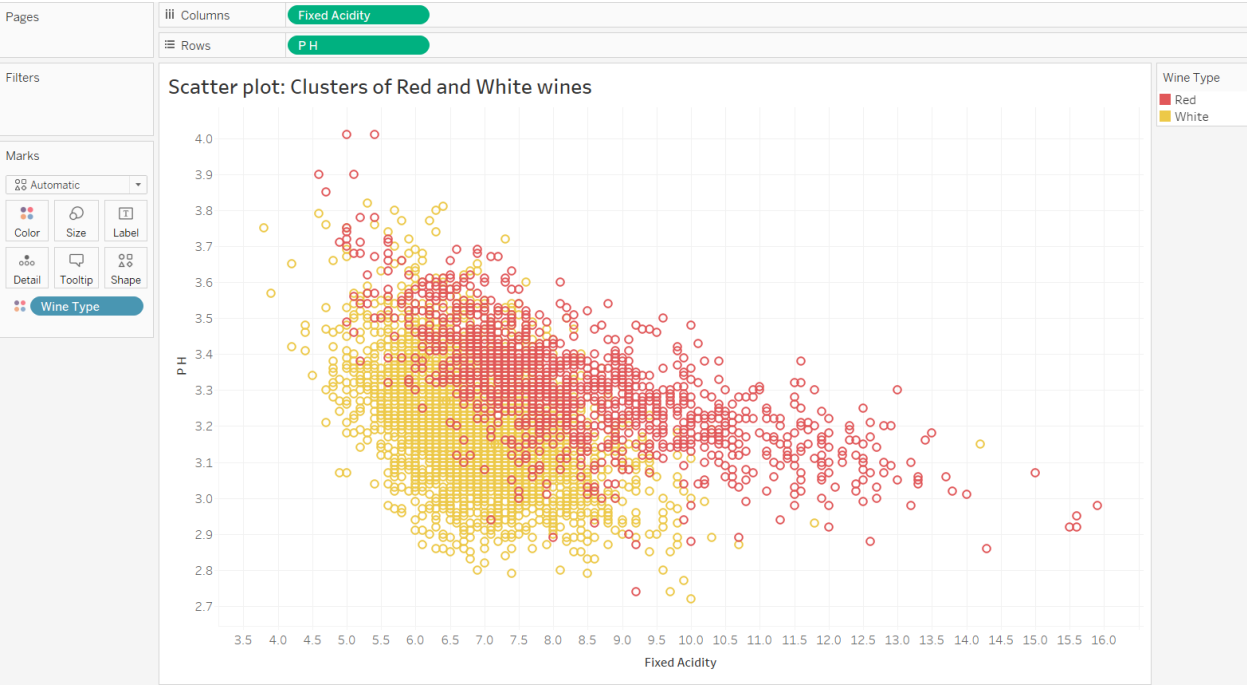



*Figure 25: Scatter plot for fixed acidity and pH*

It is observed from Figure 25 that when a scatter plot is plotted with fixed acidity and pH, clusters of red and white are formed. Red wine has higher pH and fixed acidity than white wine.

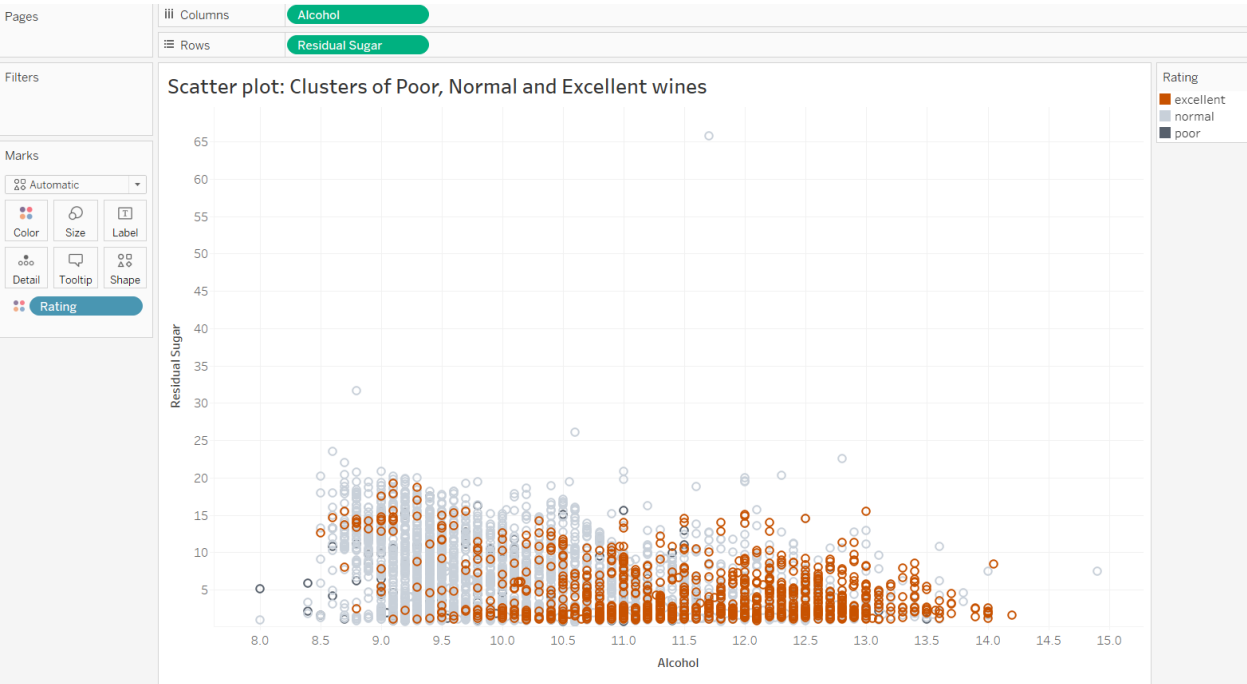



*Figure 26: Scatter plot for alcohol and residual sugar*

Figure 26 depicts different clusters of ratings (poor, normal, excellent) of red and white wines, when scatter plot is plotted between alcohol and residual sugar.

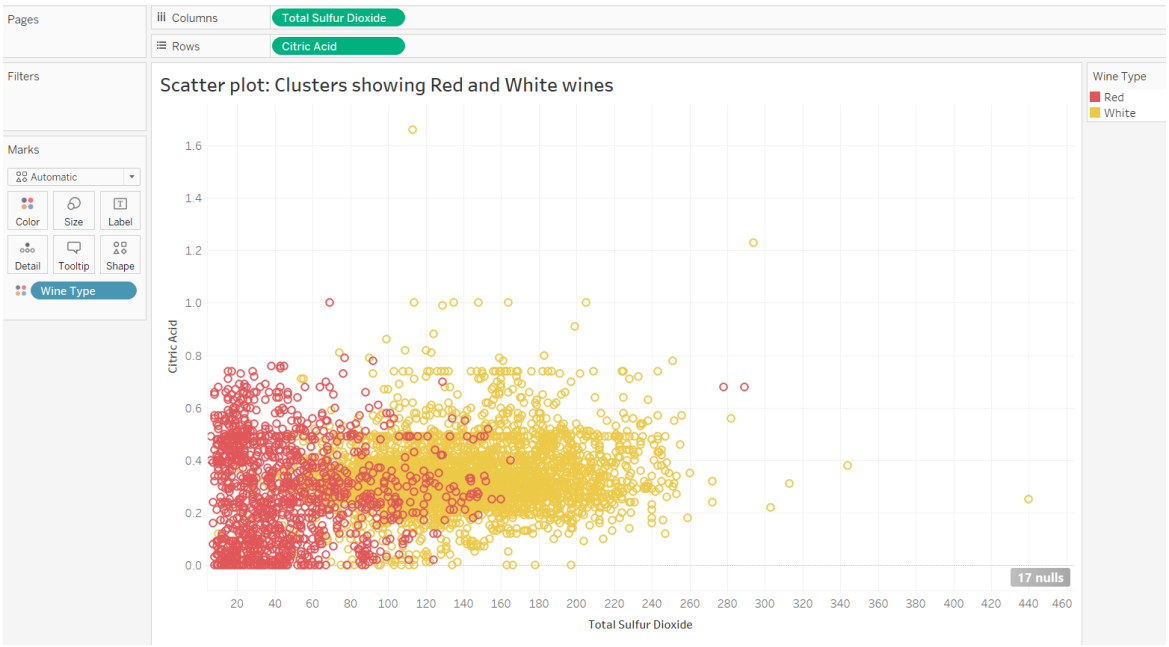



*Figure 27: Scatter plot for total sulfur dioxide and citric acid*

Clustering of red and white wines can be observed from Figure 27, when a scatter plot is plotted between attributes total sulfur dioxide and citric acid.

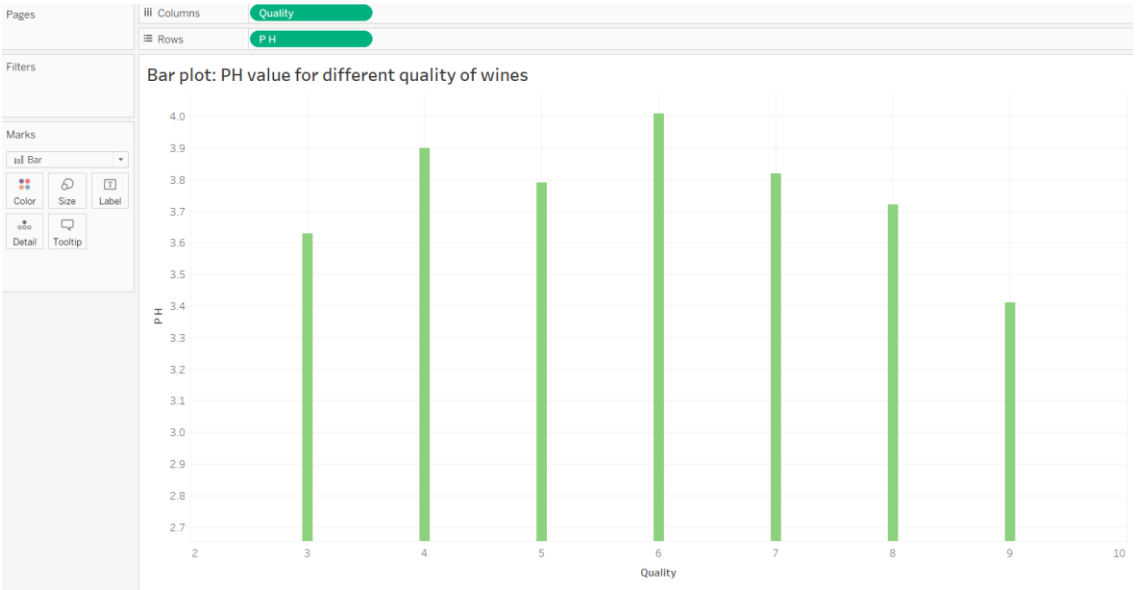



*Figure 28: Bar plot for quality and pH*

When a bar plot is plotted with quality and pH variables, it is observed from Figure 28 that pH value decreases with increase in quality.


*Figure 29: Bar plot for quality and volatile acidity*

Figure 29 shows that volatile acidity of wines decreases with increase in quality.

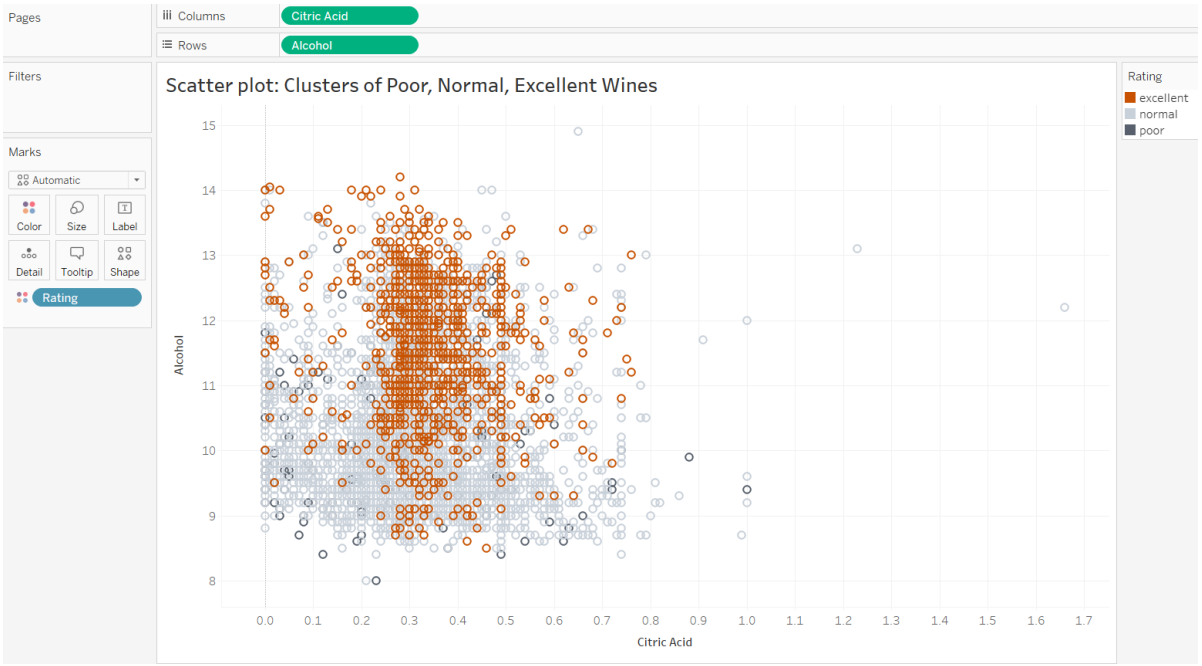



*Figure 30: Scatter plot for citric acid and alcohol*

Clusters of ratings normal, poor and excellent are observed in Figure 30, when a scatter plot is plotted with the attribute alcohol and citric acid.

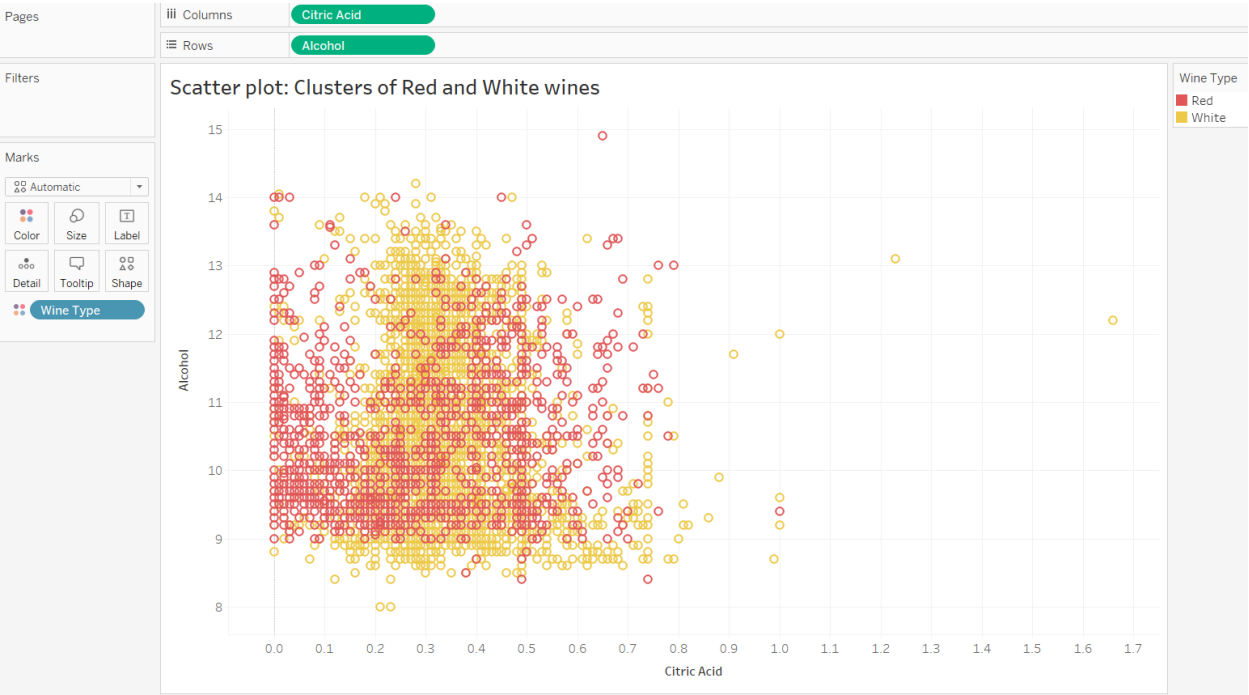



*Figure 31: Scatter plot for citric acid and alcohol*

Clustering of red and white wine type is observable in the scatter plot of Figure 31.

## Part 2

Specific questions:

1. Which wine quality is good?

From Figure 11 and Figure 12, it is observed that most of the red wine quality is scored as 5. Whereas, white wine quality is scored as 6. We can infer from this that white wine quality is much better than red wine quality. However, red wine data has less observations. This inference can be interpreted very well when we have similar observations as white wine data.

2. What makes wine quality good?

We can observe from the Figures 13 and 14 that when alcohol and citric acid content in wine increases the quality score is increasing. Also, from Figures 15,16,17,18 and 19 it is observable that higher quality wines have lesser content of residual sugar, total sulfur dioxide, and chlorides. So, for a better wine quality it should have higher alcohol, citric acid and lesser residual sugar, total sulfur dioxide, chlorides, but in equal proportion.

3. How wine is categorized? (poor, normal, excellent)

Wines can be categorized as poor, if quality score is less than 5, normal if quality score is less than 7 and excellent if quality score is higher or equal to 7. Based on the quality scores, a new column “rating” is added in data set, which contains these categories. The categorization is made by the observations of quality score given. Below R script can be used to add rating column in the data frame. But this column is added manually in the data set, as it is used for visualizations in Tableau.

```
rwwine$rating <- ifelse(rwwine$quality < 5, "poor", ifelse(rwwine$quality < 7, "normal", "excellent"))
```

4. What makes wine price costlier?

Older wines are usually more expensive than younger ones. As time passes, wines start to age. Time gradually alters the flavor of the fruit in the wines. And it also reduces the acidity of the wine [3]. From Figure 20, we can observe that higher quality wines have lesser fixed acidity. So, the fixed acidity can determine the cost of wines. Lesser the fixed acidity, costlier the wine.

R gives greater software environment for the statistical computing and graphics. It offers various libraries, with which analysis of data can be made very easily. Tableau provides powerful data visualization platform. It helps in extracting information from raw data. Whereas, SQL is a structured query language, with which we can insert, delete and update the database records. It optimizes and maintains the database. With these language and tools, wine quality, price, parameters contributing to price and categorizes can be determined.

Terms:

Outlier – it is an observation that lies outside of distribution pattern.

Physiochemical – it relates to physics and chemistry.

Challenges:

Determining the price of wines was challenging. Out of 14 attributes, only 1 attribute that is fixed acidity was able to determine the price. It consumed time in understanding parameters that affect the price of wine. Also, the data records for red wine are comparatively lesser than white wine. Similar data records as white wine could have made better analysis in determining wine quality.

In the future, it is needed to determine the chemical property of the wines that are responsible for the sweetness in higher quality wines.

## References:

[1] P. Cortez, A. Cerdeia, 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, 2009.

[2] Paulo Cortez. (2009). UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[3] Wine Savvy. (2015). Wine Cooler. Retrieved from https://learn.winecoolerdirect.com/why-your-wine-is-expensive/