**R PROJECT**

**TOPIC**

**WINE QUALITY ANALYSIS**

**TEAM MEMBERS**:

Giridhara Vijay.R.A (71762133013)

Povearasu.L (71762133036)

**ABSTRACT**:

The quality of wine is varied based on its chemical composition. There are two varities of wines based on its colour as Red and White wines, some of the important factors decides the quality of wine. We will analyse the quality factor and categorize each, the suitable range of factors for better quality. We will find the best level of product’s composition in the given data as per the analysis using charts, scatterplot matrice, correlation, log, summary, mean and some required functions and conclude the result.

OBJECTIVE:

1. The composition of chemicals and alcholic contents are analysed.
2. The quality of wine is filtered based on the analysis.
3. Both the red and white varities are compared with its count and quality.
4. Checking effect of the sugar residual in wine quality.
5. Checking effect of acids in quality of wines produced.
6. Which wine have the better quality based on its alcoholic content.
7. Summary of the report.

DATASET SOURCE:

<https://www.kaggle.com>

METHODS USED:

1. Correlation
2. Bi-variate analysis
3. Multi-variate analysis
4. P-chart
5. Np-chart

CODE:

library(ggplot2)

library(dplyr)

library(gridExtra)

library(GGally)

library(memisc)

library(pander)

library(corrplot)

library(qicharts)

library(qcc)

#Correlation table

#Putting a Cor test together

simple\_cor\_test <- function(x, y) {

return(cor.test(x, as.numeric(y))$estimate)

}

correlations <- c(

simple\_cor\_test(wine$fixed.acidity, wine$quality),

simple\_cor\_test(wine$volatile.acidity, wine$quality),

simple\_cor\_test(wine$citric.acid, wine$quality),

simple\_cor\_test(log10(wine$residual.sugar), wine$quality),

simple\_cor\_test(log10(wine$chlorides), wine$quality),

simple\_cor\_test(wine$free.sulfur.dioxide, wine$quality),

simple\_cor\_test(wine$total.sulfur.dioxide, wine$quality),

simple\_cor\_test(wine$density, wine$quality),

simple\_cor\_test(wine$pH, wine$quality),

simple\_cor\_test(log10(wine$sulphates), wine$quality),

simple\_cor\_test(wine$alcohol, wine$quality))

barplot(correlations , names.arg =c('fixed.acidity', 'volatile.acidity', 'citric.acid',

'log10.residual.sugar',

'log10.chlordies', 'free.sulfur.dioxide',

'total.sulfur.dioxide', 'density', 'pH',

'log10.sulphates', 'alcohol') , main = "Correlation of Wine Quality")

names(correlations) <- c('fixed.acidity', 'volatile.acidity', 'citric.acid',

'log10.residual.sugar',

'log10.chlordies', 'free.sulfur.dioxide',

'total.sulfur.dioxide', 'density', 'pH',

'log10.sulphates', 'alcohol')

print("Correlation data : ")

print(correlations)

#Multivariate

#With constant alcohol density does not seem to have much effect, confirming our old suspicion

v1 <- ggplot(data = wine,

aes(y = density, x = alcohol,

color = quality)) +

geom\_point() +

scale\_color\_brewer()

# Seems like for wines with higher alcohol content, having higher sulphate makes better wines

v2 <- ggplot(data = wine,

aes(y = sulphates, x = alcohol,

color = quality)) +

geom\_point() +

scale\_y\_continuous(limits=c(0.3,1.5)) +

facet\_wrap(~rating) +

scale\_color\_brewer()

#If volatile acidity os less, better wine

v3 <- ggplot(data = wine,

aes(y = volatile.acidity, x = alcohol,

color = quality)) +

geom\_point() +

facet\_wrap(~rating) +

scale\_color\_brewer()

#Low pH and higher alcohol percent produces better wines

v4 <- ggplot(data = wine,

aes(y = pH, x = alcohol,

color = quality)) +

geom\_point() +

facet\_wrap(~rating) +

scale\_color\_brewer()

#Lower residual sugar produces better wine

v5 <- ggplot(data = wine,

aes(y = residual.sugar, x = alcohol,

color = quality)) +

geom\_point() +

facet\_wrap(~rating) +

scale\_color\_brewer()

#In general lower SO2 produces better wine even though some high outliers for better wine with high SO2

v6 <- ggplot(data = wine,

aes(y = total.sulfur.dioxide, x = alcohol,

color = quality)) +

geom\_point() +

facet\_wrap(~rating) +

scale\_color\_brewer()

#Comparing the acids

#Higher citric acid and low volatile acidity produces better wines

v7 <- ggplot(data = wine,

aes(y = citric.acid, x = volatile.acidity,

color = quality)) +

geom\_point() +

facet\_wrap(~rating) +

scale\_color\_brewer()

#citric acid and fixed acidity may be correlated. But the quality does not seem to be dependent here

v8 <- ggplot(data = wine,

aes(y = citric.acid, x = fixed.acidity,

color = quality)) +

geom\_point() +

facet\_wrap(~rating) +

scale\_color\_brewer()

#Cannot really distinguish Average from Good wine based on these two factors

v9 <- ggplot(data = wine,

aes(y = fixed.acidity, x = volatile.acidity,

color = quality)) +

geom\_point() +

facet\_wrap(~rating) +

scale\_color\_brewer()

grid.arrange(v1,v2,v3,v4,v5 , v6 ,v7,v8,v9,ncol = 3 , top = "Multi Variate Analysis")

wine <- read.csv('Winedataset.csv')

#Converting Wine quality into a ordered factor

wine$quality <- factor(wine$quality, ordered = T)

#Creating a new 'rating' variable into the dataframe for different quality range

wine$rating <- ifelse(wine$quality < 5, 'bad', ifelse(

wine$quality < 7, 'average', 'good'))

wine$rating <- ordered(wine$rating,

levels = c('bad', 'average', 'good'))

#Summary of the dataframe

print(summary(wine))

#Quality and rating

q1<-ggplot(data = wine, aes(x = quality)) +

stat\_count(width = 1, color = 'black',fill = I('orange')) +ggtitle("Quality")

q2<-ggplot(data = wine, aes(x = rating)) +

stat\_count(width = 1, color = 'black',fill = I('blue'))+ggtitle("Rating")

grid.arrange(q1,q2, ncol =2)

#Bivariate analysis

#Fixed acidity : Doesn't seem to have much effect

a1<-ggplot(data = wine, aes(x = quality, y = fixed.acidity)) +

geom\_boxplot() +ggtitle("Fixed Acidty")

#Volatile Acidity : Seems to have negative effect. With increase, quality seems to go down

a2<-ggplot(data=wine, aes(x = quality, y = volatile.acidity)) +

geom\_boxplot()

#Citric acid (Better wines tend to have higher citric acid)

a3<-ggplot(data=wine, aes(x=quality, y=citric.acid)) +

geom\_boxplot()

#Residual Sugar(Almost has no effect to quality. This is contrary to previous assumption)

a4<-ggplot(data=wine, aes(x=quality, y=residual.sugar)) +

geom\_boxplot()

#Chlorides

a5<-ggplot(data=wine, aes(x=quality, y=chlorides)) +

geom\_boxplot()

#Free SO2(We see too little and we get a poor wine and too much : we get an average wine)

a6<-ggplot(data=wine, aes(x=quality, y=free.sulfur.dioxide)) +

geom\_boxplot()

#Total SO2(Just like free SO2)

a7<-ggplot(data=wine, aes(x=quality, y=total.sulfur.dioxide)) +

geom\_boxplot()

#Density(Better wines tend to have lower densities but is it due to alcohol content?)

a8<-ggplot(data=wine, aes(x=quality, y=density)) +

geom\_boxplot()

#pH(Better wines seems to be more acidic. Now let's see contribution of each acid on pH)

a9<-ggplot(data=wine, aes(x=quality, y=pH)) +

geom\_boxplot()

grid.arrange(a1,a2,a3,a4,a5,a6,a7,a8,a9,ncol =3 , top = "BiVariate Analysis")

#Contribution of each acid to pH(We see all of them has negative correlation on pH except

#volatile acidity. But how's that possible! Is it possible that there is a Simson's effect?)

c\_1<-ggplot(data = wine, aes(x = fixed.acidity, y = pH)) +

geom\_point() +

scale\_x\_log10(breaks=seq(5,15,1)) +

xlab("log10(fixed.acidity)") +

geom\_smooth(method="lm")

c\_2<-ggplot(data = wine, aes(x = volatile.acidity, y = pH)) +

geom\_point() +

scale\_x\_log10(breaks=seq(.1,1,.1)) +

xlab("log10(volatile.acidity)") +

geom\_smooth(method="lm")

c\_3<-ggplot(data = subset(wine, citric.acid > 0), aes(x = citric.acid, y = pH)) +

geom\_point() +

scale\_x\_log10() +

xlab("log10(citric.acid)") +

geom\_smooth(method="lm")

grid.arrange(c\_1,c\_2,c\_3 , ncol = 3 , top = "Correlatio on Ph Effect")

#Sulphates(better wines seems to have higher sulphates. Although medium wines have many outliers)

print(ggplot(data=wine, aes(x=quality, y=sulphates)) +

geom\_boxplot()+ggtitle("Sulphates quality analysis"))

#Alcohol(Better wines have higher alcohol)

print(ggplot(data=wine, aes(x=quality, y=alcohol)) +

geom\_boxplot()+ggtitle("Alcohol quality analysis"))

p <- qcc(wine$free.sulfur.dioxide,wine$total.sulfur.dioxide,type = "p" ,data.name = "Sulfur Content")

np\_chart <- qcc(wine$free.sulfur.dioxide,wine$total.sulfur.dioxide, type = "np", data.name = "Sulfur Content")

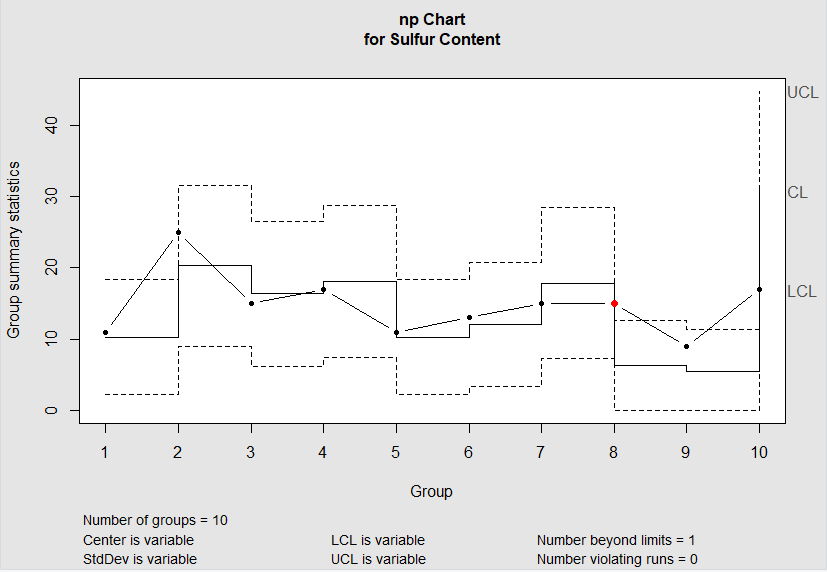
print("P CHART : ")

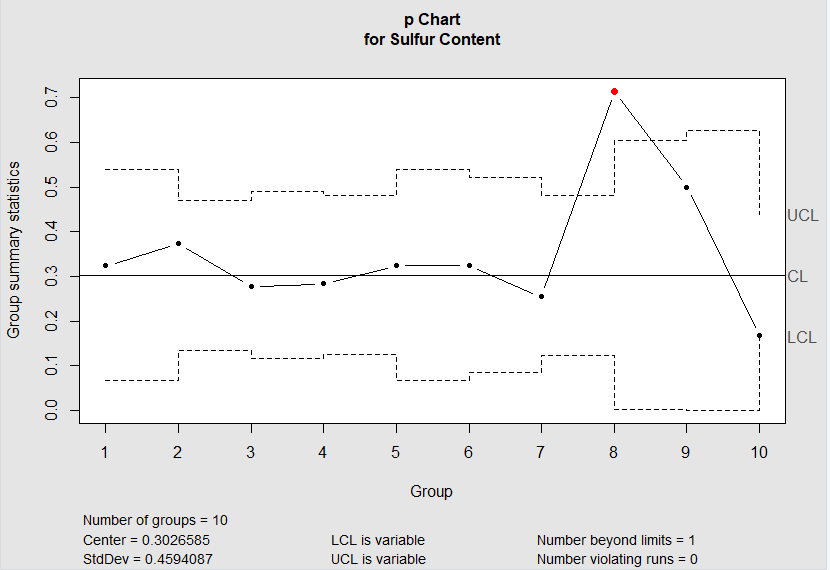
summary(p)

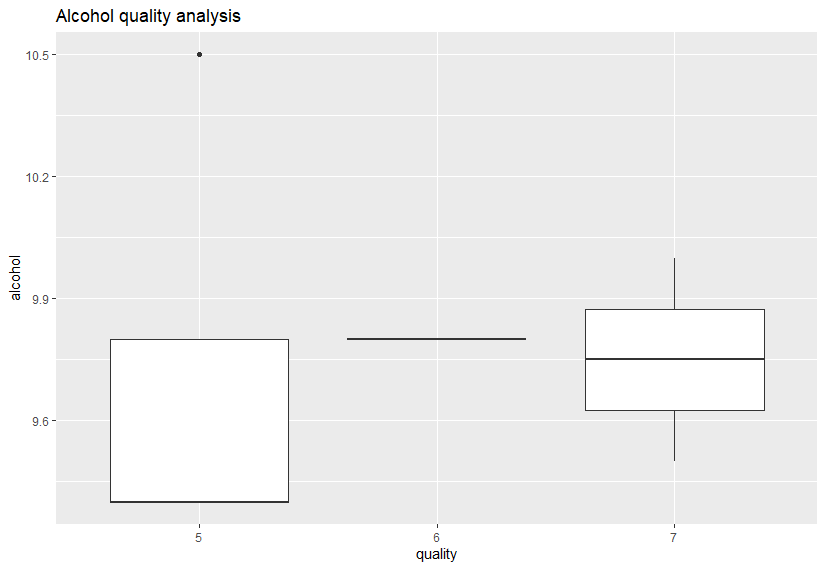
print("NP CHART : ")

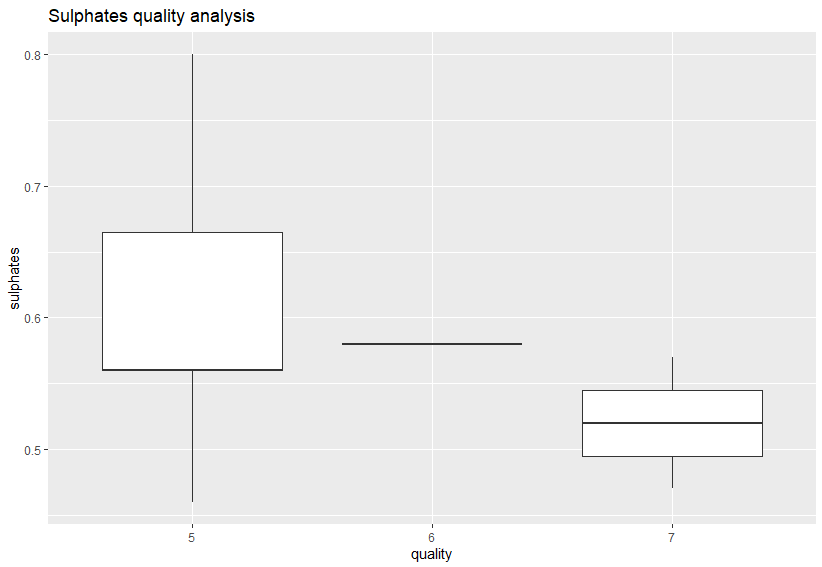
summary(np\_chart)

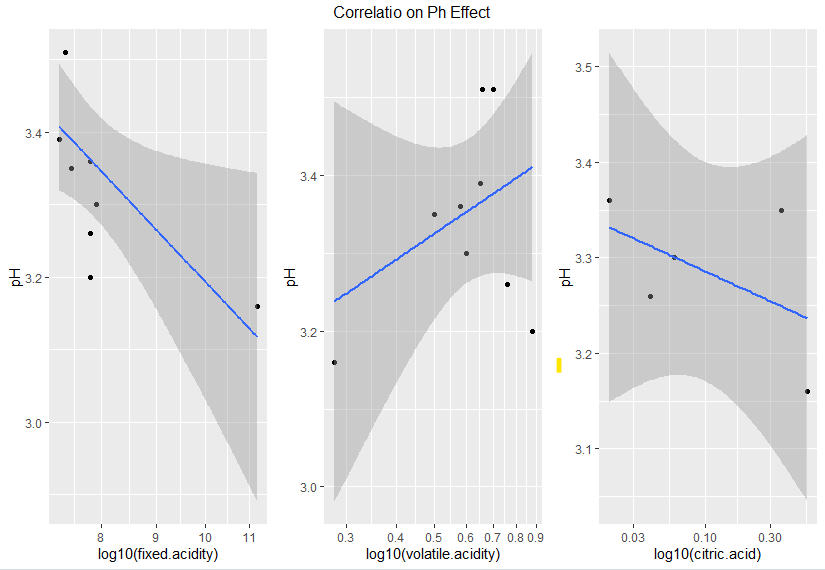
OUTPUT:

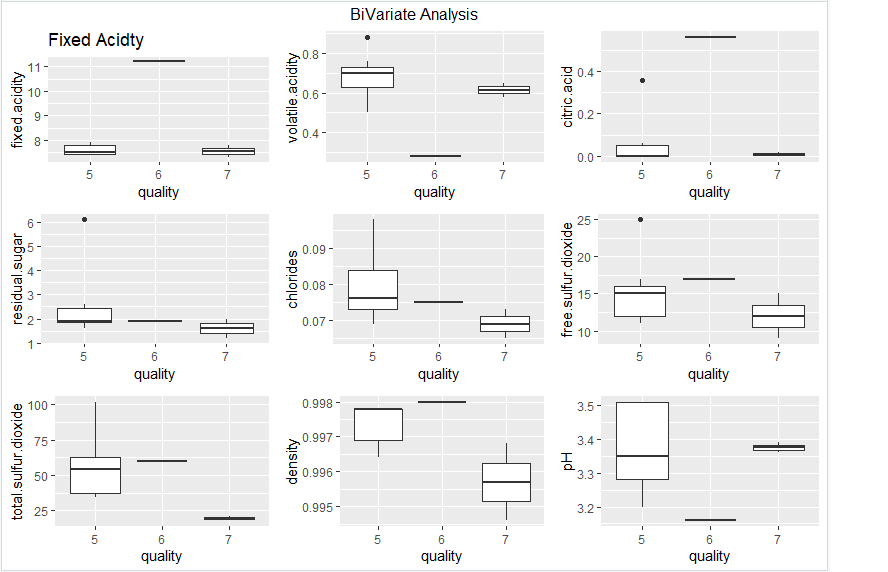


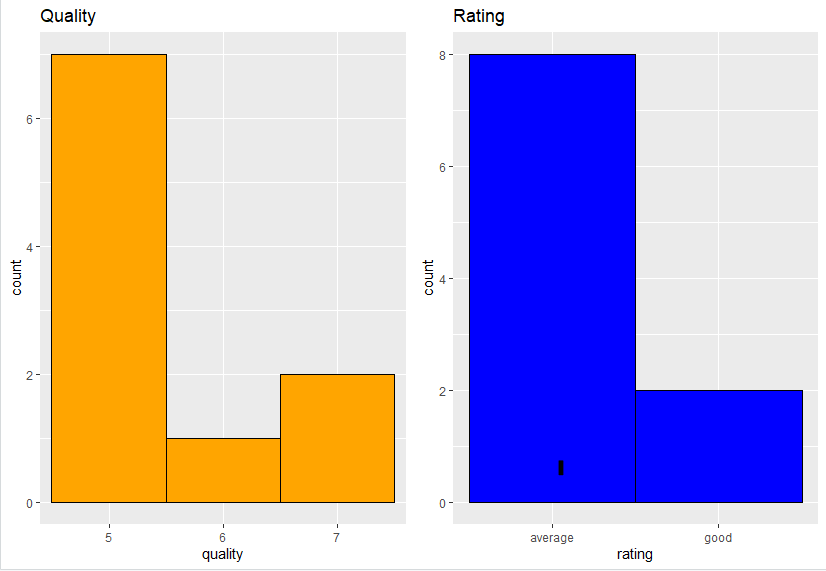


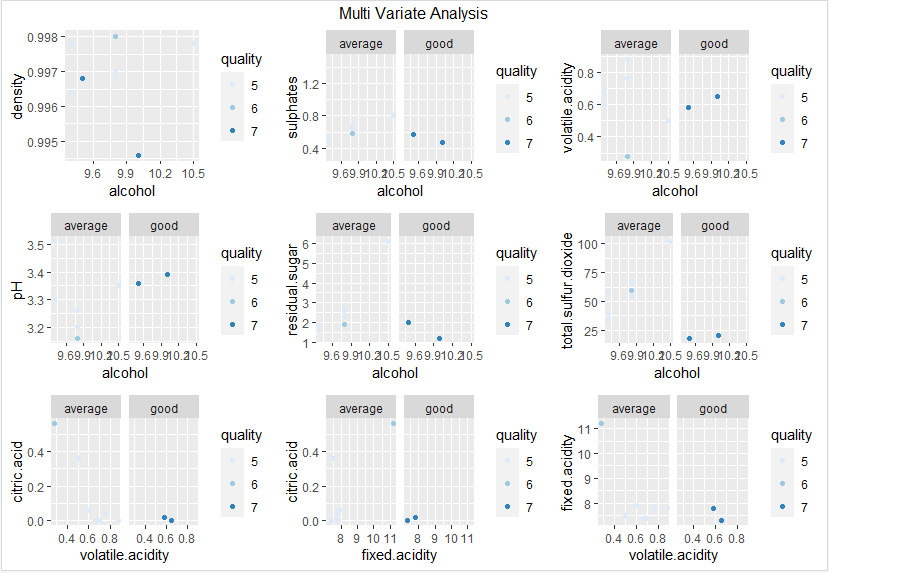














REFERENCE LINKS:

<https://www.statology.org>

<https://web.stanford.edu>

<http://www.sthda.com>

<https://towardsdatascience.com>