St. Xavier's College (Autonomous), Kolkata

Analysis of Wine Data

Session : 2019 - 2022

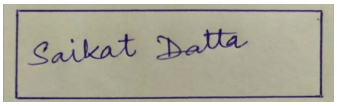
Name - Saikat Datta

Roll No. - 416

Department - Statistics

Supervisor - Dr. Ayan Chandra

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.



Signature of the student

Contents

1 Introduction :

The fundamental aim of statistics is to draw a conclusion from data and to meet this purpose the data must be analysed thoroughly. Again, in order to get a concrete idea about the data, the process of data analysis further needs to follow various steps and techniques. Such a step of data analysis is exploratory data analysis (EDA) which is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
In this project we will perform EDA on the given data and also see the goodness of a fitted model.

2 Objective :

The main objective of this project is to extract maximum knowledge from the given data in different ways and to check the goodness of the regression model fitted to this data by using the method of generalized linear model.

3 Analysis :

Description of the data -

In this project we are working with a wine dataset. This dataset is related to red variants of the Portuguese ”Vinho Verde” wine and is collected from UCI Machine learning repository (Link to the source of the data). The data consist of twelve columns. In order to study the dataset and fit a regression model, we take the categorical variable ‘quality’ of wine as our target attribute and rest of the eleven continuous variables as predictors.
Thus, we have,

Predictors -

1. Fixed acidity - The levels of most acids involved with wine like tartaric, malic, citric, and succinic acids

2. Volatile acidity - The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3. Citric acid - Found in small quantities, citric acid can add ’freshness’ and flavor to wines

4. Residual sugar - The amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1

gram/liter

5. Chlorides - The amount of salt in the wine
6. Free sulfur dioxide - The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion
7. Total sulfur dioxide - Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine
8. Density - The density of water is close to that of water depending on the percent alcohol and sugar content
9. pH - Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)
10. Sulphates - A wine additive which can contribute to sulfur dioxide gas (S02) levels
11. Alcohol - The percent alcohol content of the wine

Response -

Quality of wine - Output variable (based on sensory data, score between 0 and 10)

3.1 Checking for missing values -

Let us now plot the data in the following graph to check for any missing observations.

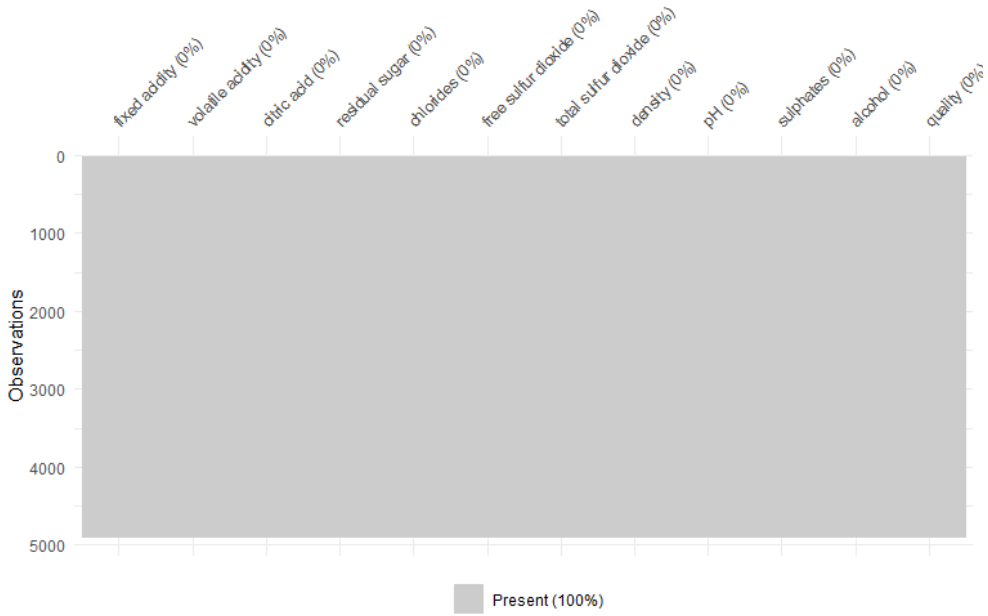


Figure 1: Checking for missing observations in the data

This plot provides a specific visualiation of the amount of missing data, showing in black the location of missing values, and also providing information on the per-

centage of missing values on the overall data, and in each variable (in the legend). From Figure 1, it is evident that none of the columns contain any missing observation. Therefore, we do not need to change any part of the dataset.

3.2 Analysis of Response -

The barplot of the 'quality' of wine is shown below -

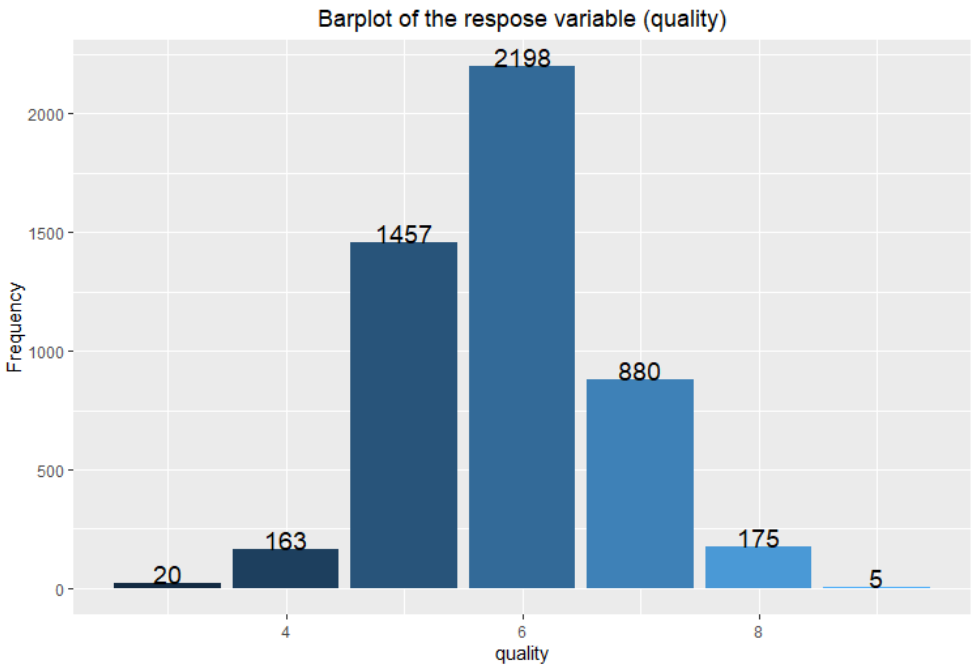


Figure 2: Barplot of Quality of wine

- The "quality" of wine is a categorical variable that ranges from 3 to 9.
- From Figure 2 we observe that the number of data points having qualities

6, 5 and 7 is very high and that of the points having qualities 3, 4, 8 and 9 is very low.

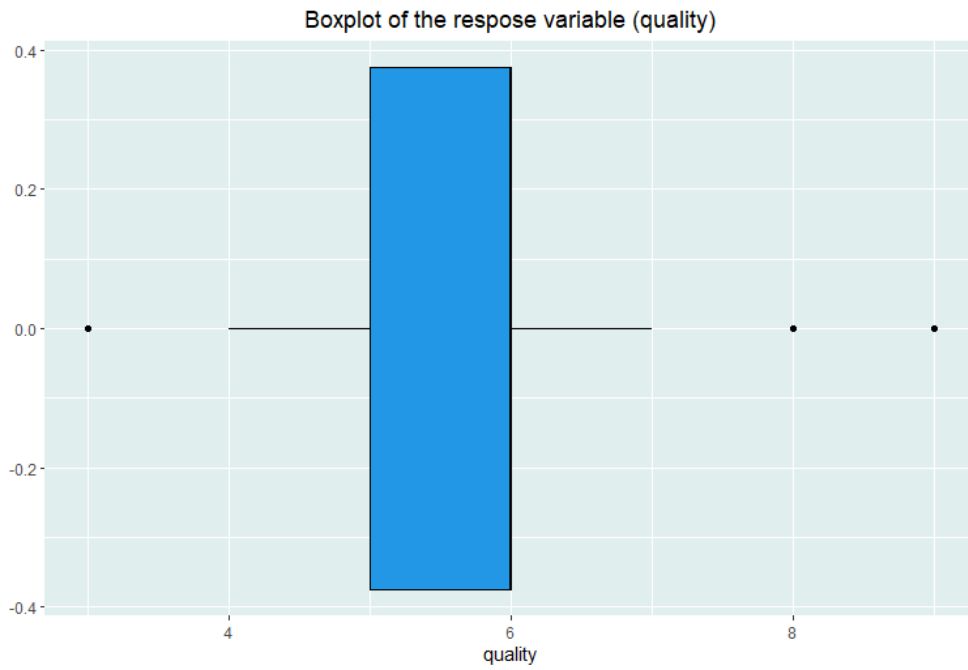


Figure 3: Boxplot of Quality of wine

Now, the boxplot of the response variable shows that the median of the quality lies at 6.

Also, we can see the categories 3, 8 and 9 are being regarded as outliers. This is because there are very few observations that have qualities 3, 8 or 9.

3.3 Analysis of Predictors -

In our data all the eleven predictors are continuous variables. To understand our data properly, we observe the distributions and find various descriptive measures for each of them.

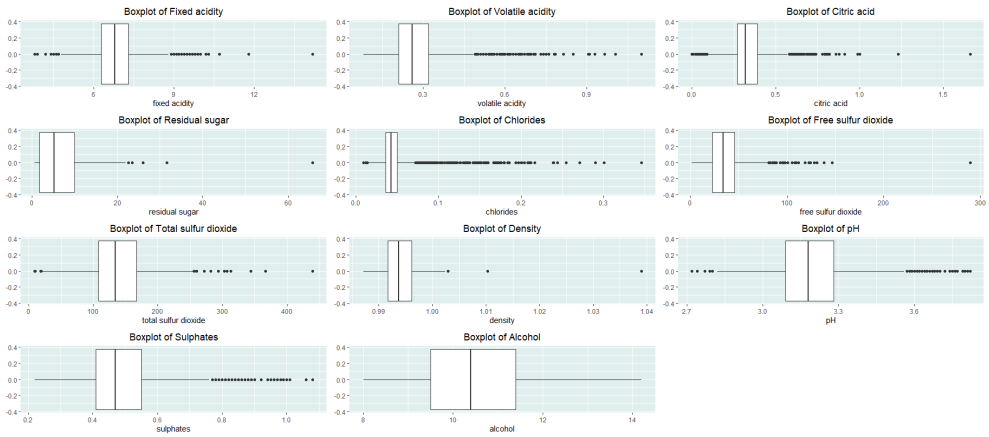


Figure 4: Boxplot of the predictors

As it is observed, the predictors have different ranges. There are outliers present in each variables except for the predictor ‘Alcohol’. The number of outliers present in each predictor is quite high apart from ‘Density’ and ‘Alcohol’.
Now, we summarise the values of descriptive measures for the predictors in the following table -

Predictor	Mean	Median	Skewness (γ_1)	Kurtosis (β_2)
fixed acidity	6.855	6.8	0.6476	5.1687
volatile acidity	0.2782	0.26	1.5765	8.0852
citric acid	0.3342	0.32	1.2815	9.167
residual sugar	6.391	5.2	1.0768	6.4651
chlorides	0.04577	0.043	5.0218	40.525
free sulfur dioxide	35.31	34	1.4063	14.4534
total sulfur dioxide	138.4	134	0.3906	3.57
density	0.994	0.9937	0.9775	12.7826
pH	3.188	3.18	0.4576	3.529
sulphates	0.4898	0.47	0.9769	4.5881
alcohol	10.51	10.4	0.4872	2.3011

Observe that,

- The value of γ_1 is greater than 0 for all the variables. Hence, each of their distribution is positively skewed. Also, the skewness of 'chlorides' is much higher than any other predictors.
- The values of β_2 for 'total sulfur dioxide' and 'pH' are respectively 3.57 and 3.529, indicating that their distributions are more or less mesokurtic.
- The distribution of 'alcohol' is clearly platykurtic and the distributions of rest of the variables are leptokurtic.

Checking for normality of the predictors -

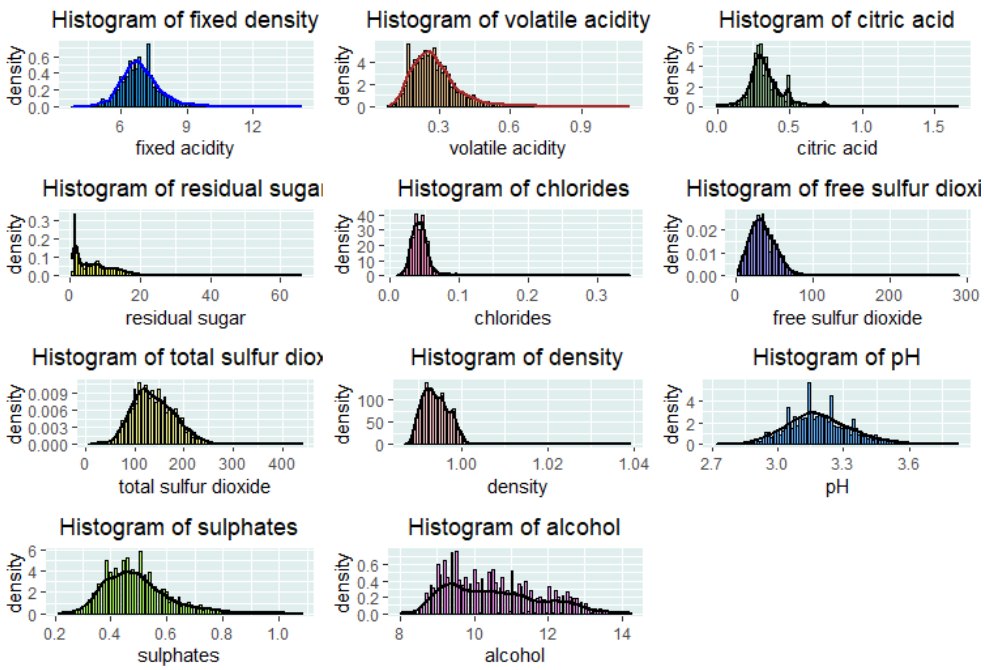


Figure 5: Histogram of the predictors

From Figure 5 it seems that the predictors are not normally distributed.

We conduct **Shapiro-Wilk test** to check for normality of the population distribution for each predictor.

An overview of the test

The Shapiro–Wilk test tests the null hypothesis that a sample $X_1, ..., X_n$ came from a normally distributed population.

The test statistic is

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where,

- $x_{(i)}$ is the ith order statistic,
- $\bar{x} = \frac{x_1 + ... + x_n}{n}$

The coefficients a_i's are given by, $(a_1, ..., a_n) = \dfrac{m^T V^{-1}}{C}$
where C is a vector norm:

$$C = ||V^{-1} m|| = (m^T V^{-1} V^{-1} m)^{1/2}$$

and a vector m, $m = (m_1, ..., m_n)^T$ is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally, V is the covariance matrix of those normal order statistics.
There is no name for the distribution of W. The cutoff values for the statistics are calculated through Monte Carlo simulations.

Test procedure :

Here, to test

$$H_0 : \text{The population distribution of the predictor is normal.}$$

against

$$H_1 : \text{The population distribution of the predictor is not normal.}$$

The results of the test are given in the following table -

Predictor	w-observed	p-value
Fixed acidity	0.9765615	$1.150151x10^{-27}$
Volatile acidity	0.9045497	$4.586797x10^{-48}$
Citric acid	0.9222473	$1.013179x10^{-44}$
Residual sugar	0.8845686	$2.820710x10^{-51}$
Chlorides	0.5908084	$2.140584x10^{-75}$
Free sulfur dioxide	0.9420691	$3.857845x10^{-40}$
Total sulfur dioxide	0.9890146	$4.383453x10^{-19}$
Density	0.9548048	$1.780895x10^{-36}$
pH	0.9880965	$6.505521x10^{-20}$
Sulphates	0.9516094	$1.821979x10^{-37}$
Alcohol	0.9553024	$2.569014x10^{-36}$

Interpretation

We can see that the p-value is lower than the level of significance (0.05) for each of the tests. So, in each case the null hypothesis is getting rejected.

Thus, in the light of the data one can say that the population distributions of all the predictors differ significantly from normality.

3.4 Correlation among the Predictors -

The pairplot showing the values of total correlation and the scatterplot among the predictors is given below -

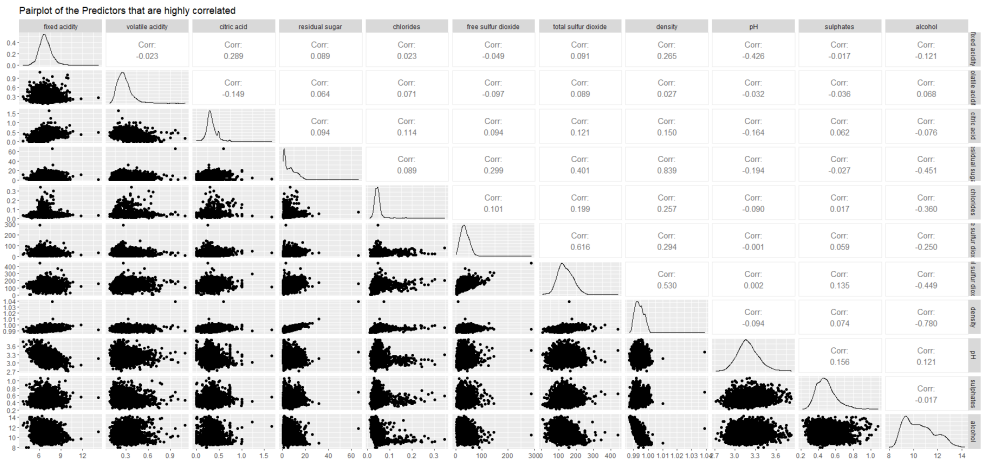


Figure 6: Pairplot of the Predictors

As we can observe from the pairplot -

- The correlation between “density” and “residual sugar” is 0.84.
- The correlation between “alcohol” and “density” is -0.78.
- The correlation between “total sulfur dioxide” and “free sulfur dioxide” is 0.62.

Therefore, multicollinearity is present between the predictors in each of these pairs mentioned above. There may be other variables that have high partial correlation among them but have a low value of total correlation coefficient. So, in order to visualize the effect of one predictor on the other by omitting the effects of rest of the variables, we plot the following partial correlation heatmap -

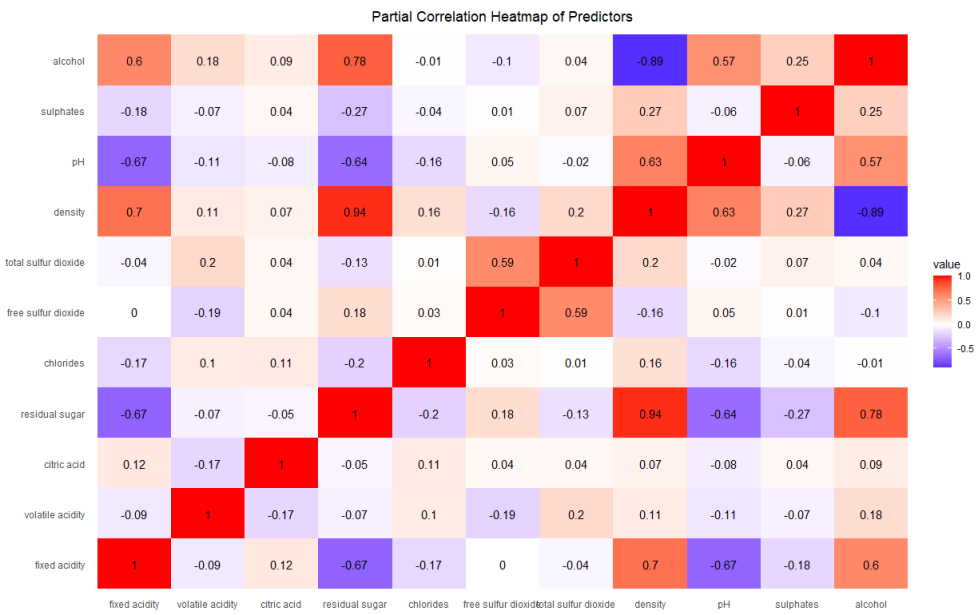


Figure 7: Partial Correlation heatmap of the predictors

Figure 7 shows that the previously mentioned predictors also have high partial correlation between them and there are some other pairs of predictors with high partial correlation as well. These pairs are given by -

- Fixed acidity and Residual sugar with partial correlation coefficient -0.67

- The partial correlation between Alcohol and Residual sugar is 0.78
- The partial correlation coefficient between Density and Fixed acidity is 0.7
- The partial correlation coefficient between pH and Fixed acidity is -0.67
- The partial correlation coefficient between pH and Density is 0.63

4 Building a Regression Model :

4.1 An Overview of Generalized Regression Model -

Generalized linear model (GLM) is a flexible generalization of ordinary linear regression. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.
In a generalized linear model, each outcome Y of the dependent variables is assumed to be generated from a particular distribution in an exponential family, a large class of probability distributions that includes the normal, binomial, Poisson and gamma distributions, among others.
The mean, μ , of the distribution depends on the independent variables, X, through:

$$E(Y|X) = \mu = g^{-1}X\beta$$

where $E(Y|X)$ is the expected value of Y conditional on X; Xβ is the linear predictor, a linear combination of unknown parameters β; g is the link function.

In this framework, the variance is typically a function, V, of the mean:

$$Var(Y|X) = V(g^{-1}X\beta)$$

It is convenient if V follows from an exponential family of distributions, but it may simply be that the variance is a function of the predicted value.

The unknown parameters, β, are typically estimated with maximum likelihood or other suitable estimation techniques.

4.2 Fitting A Binomial Logistic Regression -

In order to fit a logistic regression model to data, we first convert the response variable into a binary response variable, say Y ,where

$$Y = \begin{cases} 0, & \text{if } 3 < quality \leq 5 \\ 1, & \text{if } 5 < quality \leq 9 \end{cases}$$

Here, 0 indicates a wine of poor quality and 1 indicates a good wine.

Now, using R we fit the mentioned model to this data. The results of this fit is given below -

term	estimate	std.error	statistic	p.value	significance code
Intercept	258.2369	70.98588	3.637863	0.000275	***
fixed acidity	0.036481	0.071777	0.50826	0.611271	
volatile acidity	-6.45896	0.412817	-15.6461	3.53E-55	***
citric acid	0.115819	0.302932	0.382327	0.702219	
residual sugar	0.170066	0.027035	6.290582	3.16E-10	***
chlorides	0.885178	1.671362	0.529615	0.596379	
free sulfur dioxide	0.009601	0.002782	3.450522	0.00056	***
total sulfur dioxide	-0.00133	0.001211	-1.1008	0.270982	
density	-270.874	71.9519	-3.76466	0.000167	***
pH	1.089958	0.361796	3.01263	0.00259	**
sulphates	1.797398	0.359513	4.999533	5.75E-07	***
alcohol	0.742941	0.09361	7.936536	2.08E-15	***

‘***’ lies between p-values [0, 0.001]

‘**’ lies between p-values [0.001, 0.01]

‘*’ lies between p-values [0.01, 0.05]

‘ ’ lies between p-values [0.1, 1]

Null deviance: 6245.4 on 4897 degrees of freedom

Residual deviance: 4932.6 on 4886 degrees of freedom

Goodness of the fit and interpretation of the results -

To test,

$$H_0 : \text{The model is correctly specified.}$$

against

$$H_1 : \text{The model is not correctly specified.}$$

The test statistic is given by, D = $2(l_{max} - l)$ where, l is the log-likelihood of the model of interest and l_{max} is that of saturated model.

Now, for logistic regression model, the log-likelihood of the saturated model is 0.

Thus, the original expression of deviance, D = $2(l_{max} - l)$ reduces to D = -2l.

The test statistic D then becomes the Residual deviance.

Under H_0, D follows a χ^2_{4886} distribution.

We reject H_0 if and only if $D > \chi^2_{4886}$.

Observe that, Residual deviance of our model is 4932.6 which is less than χ^2_{4886} (5049.73).

Hence, we accept the null hypothesis.

Therefore, in the light of the data, one can say that the log-likelihood of our model is close to that of the saturated model and the model is correctly specified.

Again, the null deviance is high which indicates that the Full model is significantly deviated from Null model, i.e, it makes sense to use more than a single parameter for fitting the model.

Also, from the results we see that ‘Fixed acidity’, ‘Citric acid’, ‘Chlorides’ and

‘Total sulfur dioxide’ become insignificant in predicting whether the quality of wine is good or not under the fitted binomial logistic regression model.

Moreover, we get,

$$R^2 = 1 - (\frac{Residual\ \ devience}{Null\ \ devience})$$
$$= 0.2102047$$

Therefore, only 21.02047% of the total deviance is explained by the fitted regression model.

5 Acknowledgement :

I would like to express my thanks of gratitude to the Rector and Principal of my college, Rev. Dr. Dominic Savio, S.J. for giving me the opportunity to work on this project. I would also like to thank the Vice Principal of Arts and Science department, Professor Betram Da'Silva and the Dean of Science Dr. Tripati Dutta. I am also very grateful to the Head of the Department of Statistics Dr. Durba Bhattacharya for guiding me with the guidelines of the project.

I would like to express my special thanks to my project supervisor Professor Ayan Chandra. He constantly supported me and guided me with his valuable advice whenever I was stuck with my project.

Finally, I would like to thank my family and my friends. They helped me with the relevant knowledge they had and kept me motivated to work constantly on my project. My work was made easier by them.

6 Appendix :

The R code for this project is given below -

Listing 1: wine

```
rm(list=ls())
library(readxl)
library(dplyr)
library(moments)
library(ggplot2)
library(gridExtra)
library(reshape2)
library(visdat)
library(lattice)
library(caret)
library(magrittr)
library(pROC)
library(broom)
library(GGally)
wine <- read_excel("C:\\Users\\SAIKAT DATTA\\OneDrive\\
   Desktop\\wine dissertation\\wine.xlsx")
View(wine)
#To visualize missing values
sum(is.na(wine))
vis_miss(wine)
```

```r
skewness(wine)
#Distribution of the response
ggplot(data = wine)+
  geom_boxplot(aes(x=quality),color='black',fill='4')+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of the respose variable (quality)')+
  theme(plot.title = element_text(hjust=0.5))

quality_data <- melt(wine$quality, id = c("Type")) %>%
  group_by(wine$quality) %>% summarize(count = n())

ggplot(data = quality_data,aes(x=`wine$quality`,y=count,
                               fill=`wine$quality`))+
  geom_bar(stat = "identity",show.legend = F)+
  labs(title = 'Barplot of the respose variable (quality)',
       x='quality',
       y='Frequency')+
  theme(plot.title = element_text(hjust=0.5))+
  geom_text(aes(label = count), vjust = 0,lwd=5)

#Boxplot of predictors and descriptive measures
colnames(wine)
```

```
b1 <- ggplot(data = wine)+
  geom_boxplot(aes(x=`fixed acidity`,fill=quality))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Fixed acidity')+
  theme(plot.title = element_text(hjust=0.5))
b2 <- ggplot(data = wine)+
  geom_boxplot(aes(x=`volatile acidity`,fill=quality))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Volatile acidity')+
  theme(plot.title = element_text(hjust=0.5))
b3 <- ggplot(data = wine)+
  geom_boxplot(aes(x=`citric acid`,fill=quality))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Citric acid')+
  theme(plot.title = element_text(hjust=0.5))
b4 <- ggplot(data = wine)+
  geom_boxplot(aes(x=`residual sugar`,fill=quality))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Residual sugar')+
  theme(plot.title = element_text(hjust=0.5))
b5 <- ggplot(data = wine)+
  geom_boxplot(aes(x=chlorides,fill=quality))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Chlorides')+
```

```
  theme(plot.title = element_text(hjust=0.5))
b6 <- ggplot(data = wine)+
  geom_boxplot(aes(x=`free sulfur dioxide`,fill=quality)
    )+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Free sulfur dioxide')+
  theme(plot.title = element_text(hjust=0.5))
b7 <- ggplot(data = wine)+
  geom_boxplot(aes(x=`total sulfur dioxide`,fill=quality
    ))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Total sulfur dioxide')+
  theme(plot.title = element_text(hjust=0.5))
b8 <- ggplot(data = wine)+
  geom_boxplot(aes(x=density,fill=quality))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Density')+
  theme(plot.title = element_text(hjust=0.5))
b9 <- ggplot(data = wine)+
  geom_boxplot(aes(x=pH,fill=quality))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of pH')+
  theme(plot.title = element_text(hjust=0.5))
b10 <- ggplot(data = wine)+
```

```
  geom_boxplot(aes(x=sulphates,fill=quality))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Sulphates')+
  theme(plot.title = element_text(hjust=0.5))
b11 <- ggplot(data = wine)+
  geom_boxplot(aes(x=alcohol,fill=quality))+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Boxplot of Alcohol')+
  theme(plot.title = element_text(hjust=0.5))
grid.arrange(b1,b2,b3,b4,b5,b6,b7,b8,b9,b10,
             b11,nrow=4)

#Descriptive measures of the predictors
sw <- summary(wine[,1:11]);sw
skewness(wine[,1:11])
kurtosis(wine[,1:11])

#Histogram of features
h1 <- ggplot(data = wine)+
  geom_histogram(aes(x=`fixed acidity`,
                     y=..density..),col='black',fill=4,
                 bins = 100)+
  geom_density(aes(`fixed acidity`),color='blue',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
```

```
  labs(title = 'Histogram of fixed density')+
  theme(plot.title = element_text(hjust=0.5))
h2 <- ggplot(data = wine)+
  geom_histogram(aes(x=`volatile acidity`,
                     y=..density..),col='black',fill='burlywood',
                 bins = 100)+
  geom_density(aes(`volatile acidity`),color='brown',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Histogram of volatile acidity')+
  theme(plot.title = element_text(hjust=0.5));h2
h3 <- ggplot(data = wine)+
  geom_histogram(aes(x=`citric acid`,
                     y=..density..),col='black',fill='darkseagreen',
                 bins = 100)+
  geom_density(aes(`citric acid`),color='black',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Histogram of citric acid')+
  theme(plot.title = element_text(hjust=0.5))
h4 <- ggplot(data = wine)+
  geom_histogram(aes(x=`residual sugar`,
                     y=..density..),col='black',fill='#
```

```
                        F6FC57',
                 bins = 100)+
  geom_density(aes(`residual sugar`),color='black',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Histogram of residual sugar')+
  theme(plot.title = element_text(hjust=0.5))
h5 <- ggplot(data = wine)+
  geom_histogram(aes(x=`chlorides`,
                     y=..density..),col='black',fill='#FF99CC',
                 bins = 100)+
  geom_density(aes(`chlorides`),color='black',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Histogram of chlorides')+
  theme(plot.title = element_text(hjust=0.5))
h6 <- ggplot(data = wine)+
  geom_histogram(aes(x=`free sulfur dioxide`,
                     y=..density..),col='black',fill='#9999FF',
                 bins = 100)+
  geom_density(aes(`free sulfur dioxide`),color='black',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
```

```
  labs(title = 'Histogram of free sulfur dioxide')+
  theme(plot.title = element_text(hjust=0.5))
h7 <- ggplot(data = wine)+
  geom_histogram(aes(x=`total sulfur dioxide`,
                     y=..density..),col='black',fill='#FFFF99',
                 bins = 100)+
  geom_density(aes(`total sulfur dioxide`),color='black',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Histogram of total sulfur dioxide')+
  theme(plot.title = element_text(hjust=0.5))
h8 <- ggplot(data = wine)+
  geom_histogram(aes(x=`density`,
                     y=..density..),col='black',fill='#FFCCCC',
                 bins = 100)+
  geom_density(aes(`density`),color='black',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Histogram of density')+
  theme(plot.title = element_text(hjust=0.5))
h9 <- ggplot(data = wine)+
  geom_histogram(aes(x=`pH`,
                     y=..density..),col='black',fill='
```

```
                        #66B2FF',
                 bins = 100)+
  geom_density(aes(`pH`),color='black',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Histogram of pH')+
  theme(plot.title = element_text(hjust=0.5))
h10 <- ggplot(data = wine)+
  geom_histogram(aes(x=`sulphates`,
                     y=..density..),col='black',fill='#
                        B2FF66',
                 bins = 100)+
  geom_density(aes(`sulphates`),color='black',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Histogram of sulphates')+
  theme(plot.title = element_text(hjust=0.5))
h11 <- ggplot(data = wine)+
  geom_histogram(aes(x=`alcohol`,
                     y=..density..),col='black',fill='#
                        FF99FF',
                 bins = 100)+
  geom_density(aes(`alcohol`),color='black',lwd=1)+
  theme(panel.background = element_rect(fill='azure2'))+
  labs(title = 'Histogram of alcohol')+
  theme(plot.title = element_text(hjust=0.5))
```

```r
grid.arrange(h1,h2,h3,h4,h5,h6,h7,h8,h9,h10,h11,nrow=4)


#Correlation among the features
#Pairplot
pair <- ggpairs(wine[,1:11],
                upper = list(continuous = GGally::wrap(ggally_cor, stars = F)),
                title='Pairplot of the Predictors that are highly correlated')
pair


#Partial correlations
partial.cor_new <-
  corpcor::cor2pcor(cov(wine[,1:11]))
colnames(partial.cor_new)=colnames(wine[,1:11])
rownames(partial.cor_new)=colnames(wine[,1:11])
mel.partial_new = melt(data.matrix(partial.cor_new))
ggplot(mel.partial_new, aes(Var1,Var2))+geom_tile(aes(fill=value)) +
  geom_text(aes(label = round(value, 2)))+
  scale_fill_gradient2(low='blue' ,
                       mid='white',
```

```
                        high='red') +
  labs(title = 'Partial Correlation Heatmap of Predictors')+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    plot.title = element_text(hjust=0.5))

#Fitting Binomial Logistic regression
wine$category[wine$quality <= 5] <- 0
wine$category[wine$quality > 5] <- 1
wine$category <- as.factor(wine$category)
model_glm <- glm(category~.-quality,
                 wine,
                 family = binomial(link = "logit"))
s <- summary(model_glm);s
a <- tidy(model_glm)
write.csv(a,"C:\\Users\\SAIKAT DATTA\\OneDrive\\Desktop\\wine dissertation\\logistic_results.csv", row.names = FALSE)
```

```r

#Goodness of fit
R_sq <- 1-(s$deviance/s$null.deviance);R_sq
```