1. Data Exploration

```
1   getwd()
2   setwd("C:\\Users\\HP\\Downloads")
3
4   # Load the Data set
5   df = read.csv("cleveland.csv")
6   df
7
8   # Check whether there are any null values or missing values
9   is.null(df)
10
11  # Replace the missing values
12  df[df == '?' | df == '-'] = NA
13  df
14
15  # Save the modified data set as a CSV file
16  write.csv(df, file = 'modified_cleveland_dataset.csv', row.names = FALSE)
```

Removed all '?' values and NULL values if there are any in the given 'Cleveland' dataset.

## 2. Data Analysis

```
> summary(df)
      age             sex              cp           trestbps          chol            fbs            restecg          thalach          exang           oldpeak
 Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00
 Median :56.00   Median :1.0000   Median :3.000   Median :130.0   Median :241.0   Median :0.0000   Median :1.0000   Median :153.0   Median :0.0000   Median :0.80
 Mean   :54.44   Mean   :0.6799   Mean   :3.158   Mean   :131.7   Mean   :246.7   Mean   :0.1485   Mean   :0.9901   Mean   :149.6   Mean   :0.3267   Mean   :1.04
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0   3rd Qu.:275.0   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60
 Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0   Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20
     slope             ca             thal              num
 Min.   :1.000   Length:303       Length:303        Min.   :0.0000
 1st Qu.:1.000   Class :character Class :character  1st Qu.:0.0000
 Median :2.000   Mode  :character Mode  :character  Median :0.0000
 Mean   :1.601                                      Mean   :0.9373
 3rd Qu.:2.000                                      3rd Qu.:2.0000
 Max.   :3.000                                      Max.   :4.0000
>
```
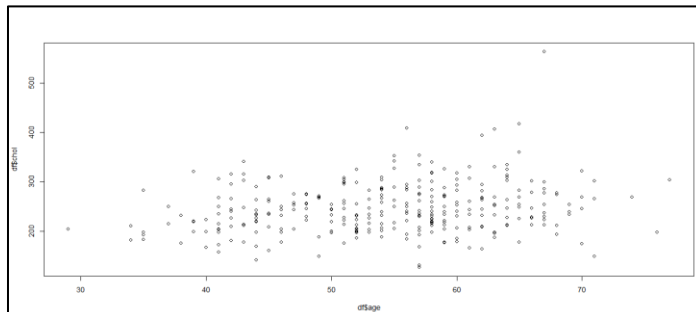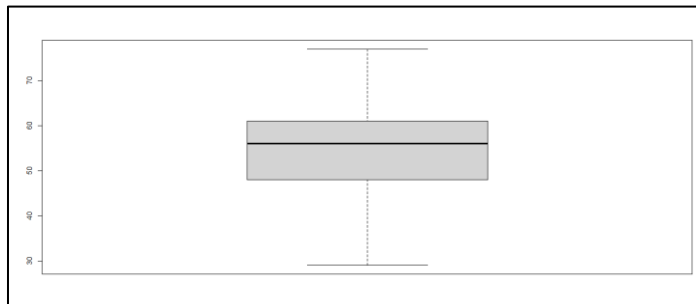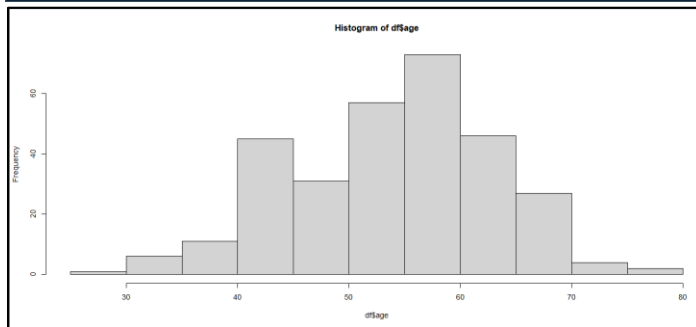
3. **Data Modeling**

   Yes, heart disease can be predicted using logistic regression. A prominent statistical technique for modeling binary outcomes, such as the presence or absence of a disease, is logistic regression. In this technique, the numbers 1 - 4 indicate the existence of heart disease at various levels, while 0 indicates the lack of heart disease.

   You would require a dataset including predictors (or independent variables) like age, sex, blood pressure, cholesterol levels, and other pertinent parameters that are known to relate to heart disease to perform logistic regression for heart disease prediction.

   One of the various predictive modeling methods that can be used to forecast cardiac disease is logistic regression. Depending on the type and complexity of the data, other techniques including decision trees, random forests, and neural networks may also be useful.

   Furthermore, since the training data used to train logistic regression models is only as good as the data itself, it is critical to make sure that the dataset is representative and free of biases to produce reliable results.

4. **Report & Summary**

   In this assignment I did the part, when a user enters all the patient details (age, sex, cp, trestbps, chol, fbs, restecg, thalch, exang, oldpeak, slope, ca, thal), then it will predict whether the patient has a defective heart with defective levels (1/2/3/4) or a healthy heart.

   First, I downloaded the "processed.cleveland.data" dataset from the given site and performed the data exploration part by importing the dataset into R Studio. Then I checked whether there are any NULL values or '?' value fields in the dataset and replaced them with NA.

   After that step I did the data analysis part by performing some analyses such as summary(), histogram, boxplot, and scatterplot. In the .ipynb file, I have also included a summary test, heatmap, histogram, boxplot, and a plot. After getting a proper understanding by the analysis of data started doing the prediction model.

5. **GitHub Repository Link**

https://github.com/KavinduKasthurirathne/Heart-Disease-Prediction-Model.git