LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer’s Name: Dr James Connolly

Assessment Title: Heart Attack Analysis

Work to be submitted to: Blackboard

Date for submission of work: 15th May, 2022

Place and time for submitting work: Letterkenny, 5:00 pm

To be completed by the Student

Student’s Name: Sheetal Padale

Class: MSc in Big Data Analytics

Subject/Module: Data Science

Word Count (where applicable):

I confirm that the work submitted has been produced solely through my own efforts.

Student’s signature: Sheetal Padale Date: 15/05/2022

|  |
| --- |
| **Notes**  **Penalties:** The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute’s Assessment Regulations.]  **Plagiarism:** Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8).  **Cheating:** The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8).  **Continuous Assessment:** For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination. |

**Heart Attack Analysis in R**

1. **ABSTRACT**

Heart Attach is very commonly observed in people of higher age group. The major cause is the unhealthy eating and bad lifestyle. Various factors causing heart attack are cholesterol, sugar level, high blood pressure and many more. This project is the study of relationship between different factors associated to Heart Attack and with the help of R script to visualize the data.

The dataset is in the csv file format and is imported in the script, followed by various analysis conducted on the dataset. The heart attack causes are explored with the help of univariate analysis and is also visualization with the help of box plot, histogram and density plots on factors as sex, age, cholesterol and resting electrocardiogram.

The Exploratory Data Analysis (EDA) is performed for changing the column names of the dataset for better understanding while analyzing the data. Few hypotheses are set and successfully achieve them. The normality test to see whether data has normal distribution or not. If the P-value less than 0.05 it is not normally distributed and if the P-value more than 0.05 it is normally distributed.

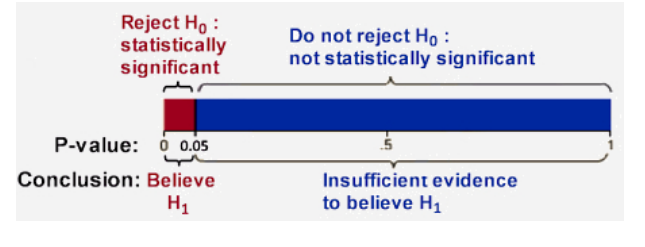


Figure: p-value significance.

Finally, to compare the difference between 2 variables (dependent and independent variables) and answers to the research questions, few tests are used such Mann-Whitney test, Wilcoxon signed rank test and Kruskal-Wallis test for analysis of data.

1. **RESEARCH QUESTIONS**

|  |
| --- |
| The research questions are as follows: |
|  |
|  |  |  |

1. Does cholesterol have more effect for chances of heart attack?
2. Does the old peak have relation with heart attack?
3. Does maximum heart rate (thallachh) has an impact on sex?
4. Does resting blood pressure (trtbps) have any impact on fasting blood sugar (fbs)?
5. Does chest Pain have a relationship with heart attack?

1. **DATA PREPARATION**

The dataset for the analysis is downloaded from Black Board. It is a collection of 303 records of patients which were divided into two types of risk for having heart attacks such namely high and low risk. This dataset has 14 variables, those are age, sex, cp, trtbps, chol, fbs, restecg, thalachh, exng, oldpeak, slp, caa, thall and output.

Following are the Column name present in heart attack analysis.

* age: It represents age of the patient.
* sex: It represents gender of the patient.
* cp: It represents the type of chest pain the patient have. It’s of 4 type
  + Value 1: typical angina
  + Value 2: atypical angina
  + Value 3: non-anginal pain
  + Value 4: asymptomatic
* trtbps: It represents the resting blood pressure (in mm Hg) in blood.
* Chol: It represents the cholesterol in mg/dl fetched via BMI sensor.
* fbs: It represents the (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false).
* resting: It represents the resting electrocardiographic results having 3 values.
  + Value 0: Normal
  + Value 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV).
  + Value 2: Showing probable or definite left ventricular hypertrophy by Estes'

Criteria.

* thalachh: It represents the maximum heart rate achieved.
* exng: It represents the exercise induced angina (if 1 = yes; if 0 = no).
* oldpeak: It represents the ST depression induced by exercise relative to rest (in mm, achieved by subtracting the lowest ST segment points during exercise and rest).
* Slp: It is the slope of the peak exercise. It represents ST segment and are considered to be a crucial for indication of ischemia presence.
  + Value 1: upsloping
  + Value 2: flat
  + Value 3: down sloping
* caa: It represents the number of major vessels (0-3).
* thall: It represents the Thallium level.
* target: 0= less chance of heart attack.

1= more chance of heart attack.

Initially the dataset in the csv format is loaded and stored into dataframe with the name heart\_record. Followed by using

1) Structure function to know the data structure using command sum(is.na(heart\_record))

2) Class to know the dataframe using command class(heart\_record)

3) Sum to check whether any NA values is present or not in the dataset using command

sum(is.na(heart\_record)

4) Knowing whether there are any non-numeric values in the dataset using

is.na(heart\_record)

5) Knowing the first and last 6 records of the file using head(heart\_record) and

tail(heart\_record)

5) Knowing the name of the rows and columns using nrow(heart\_record) and

ncol(heart\_record)

6) To know the summary of the dataset wrt minimum, maximum, mean, median values and many more using summary(heart\_record)

7) Using pair plot to examine the correlation between the variables followed by

knowing the positive correlation between age and sex variables.

8) Univariate analysis.

9) Exploratory Data analysis (EDA).

10) Hypothesis.

Packages and libraries such as ggplot2 and viridis needs to be installed and called for diagrammatic representation of the data for analysis purpose.

The output of the pair plot to know the correlations between various variables is as shown in figure below. Here in this plot the relation of each variable with another is clearly seen in the graphical format for a better understanding. However for more clear idea for correlation, the pairs.pane plot is used.

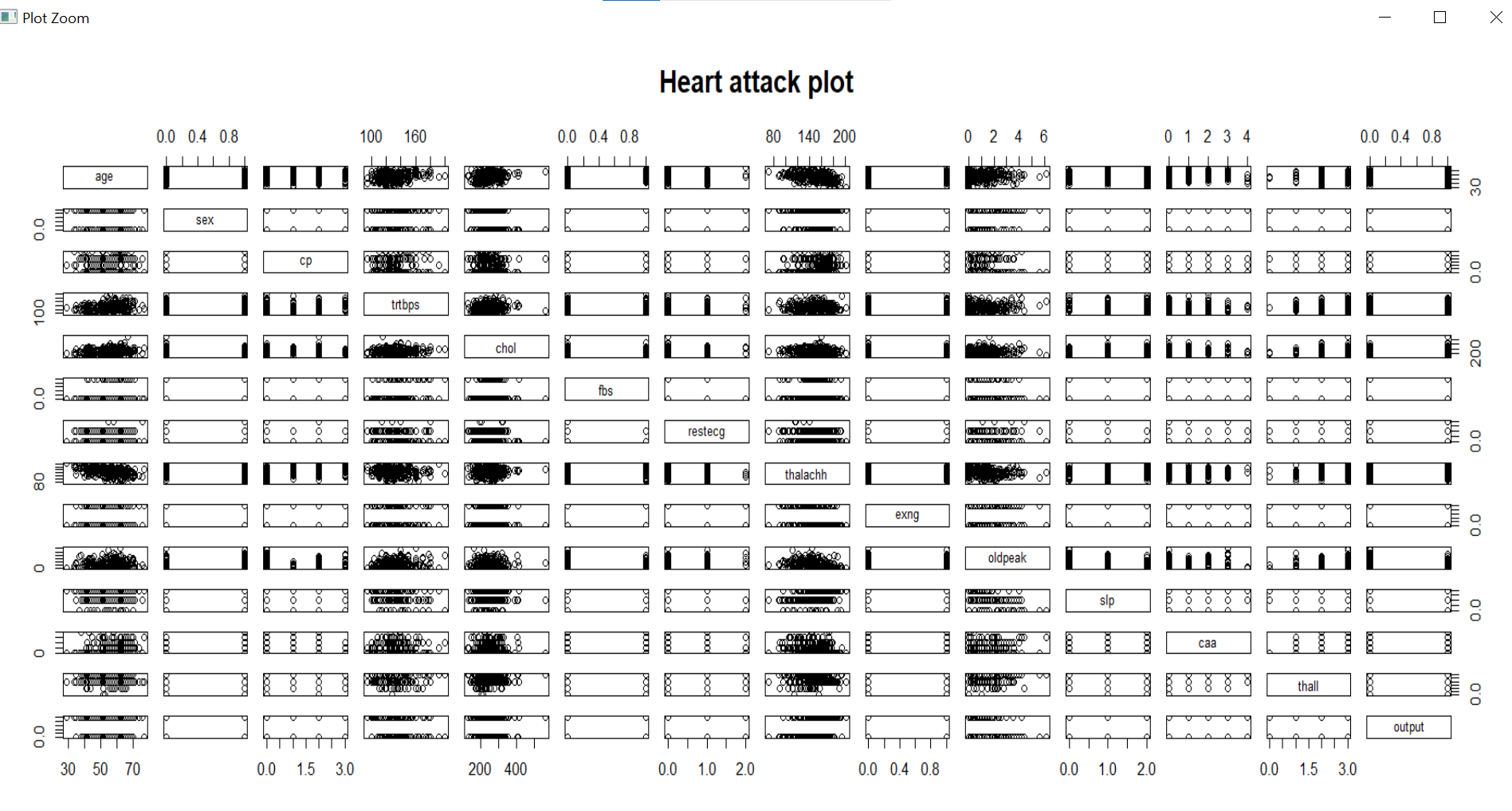


Figure: Correlations between variables of heart\_record file.

The clear and positive correlation between age and sex variables is displayed with the pairs.panel function as shown below in the figure.

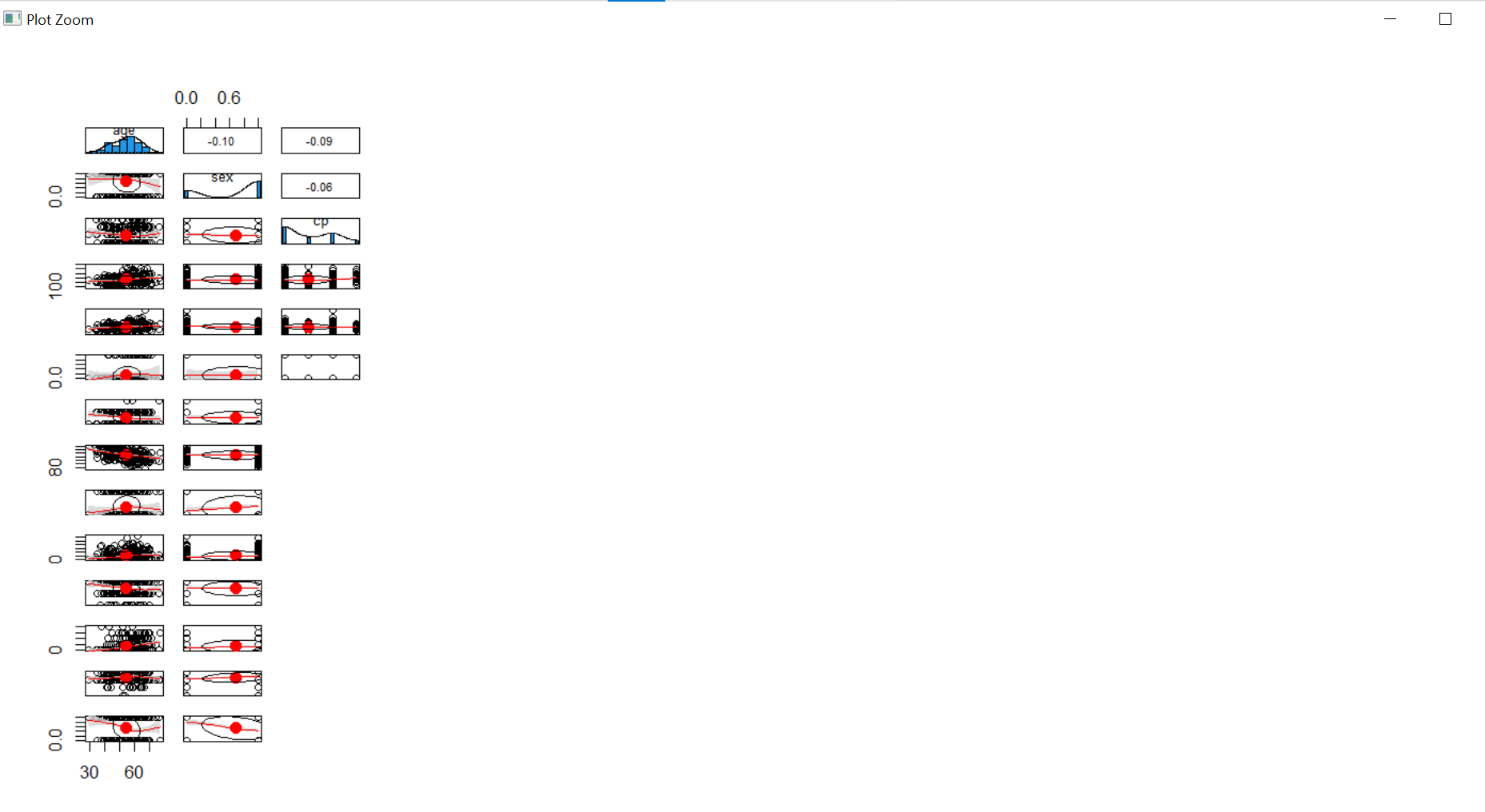


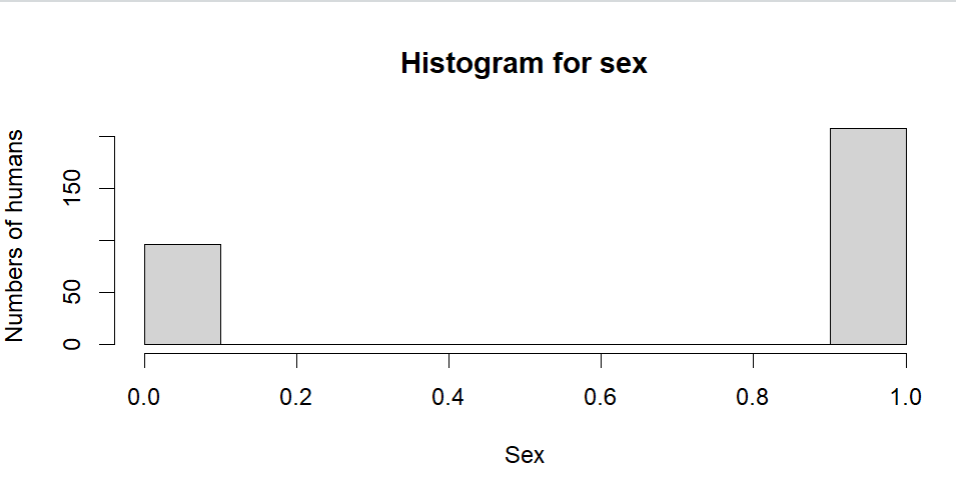
Figure: Positive correlation between age and sex variable.

**I. Discussing the univariate analysis and plotting of each variable**

We have performed following functions and their outputs for performing various univariate analysis such as

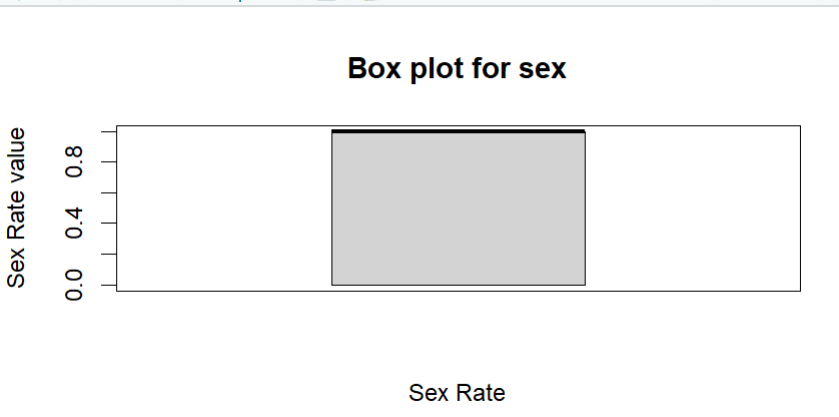
A) **sex variables** in the dataset.

* + Unique sex variables which are either 1 or 0, where 1= male and 0= female.
  + Knowing and displaying the sex variable segregated by male and female category.
  + Installation of packages and libraries such as ggplot2 and viridis for diagrammatic representation of the data for analysis purpose.
  + Histogram plot for the sex is as shown below.



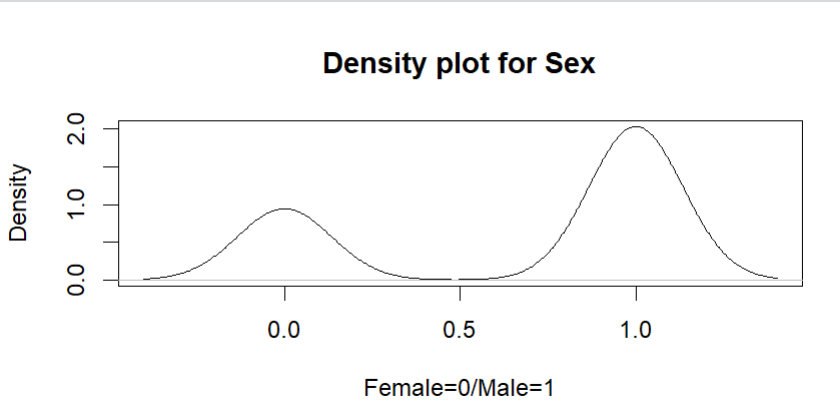
Out of 303 records of human present in the dataset, it can be seen from the histogram that majority of records are male.

* + Visualization of Sex using box plot.



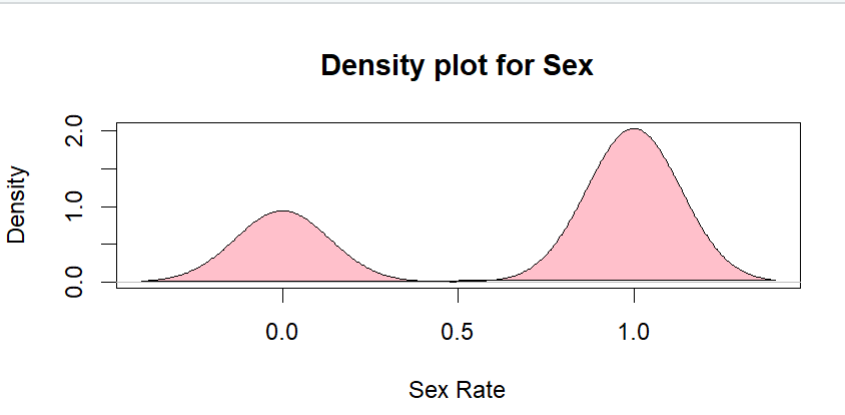
From the above box plot it is seen that the sex rate falls between 0 and 1 only.

* + Viewing the density of sex variable in the heart\_record file using density plot.

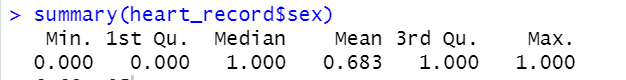


From the above density plot it is evident that the sex rate of female is less as compare to more.

* + Viewing density plot in pink color.



* + For knowing the summary such as mean = 0, 1st quantile = 0, median = 1, mean = 0.683, 3rd quantile = 1, maximum = 1 of the sex vairable. It is shown in the figure.



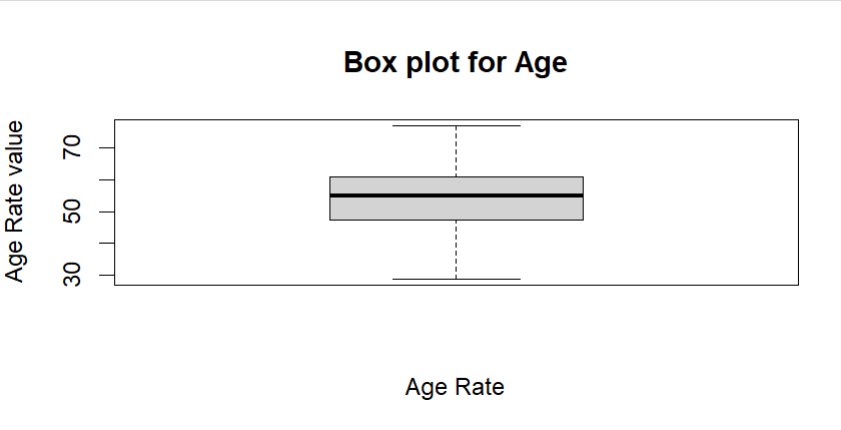
B) To know maximum and minimum **age** present in the file.

* + Unique age variables which are varies from age 29 to 77.
  + Knowing and displaying the maximum age = 77and minimum age = 27 accordingly with the help of max\_age and min\_age variables.
  + Histogram plot for the age is as shown below.



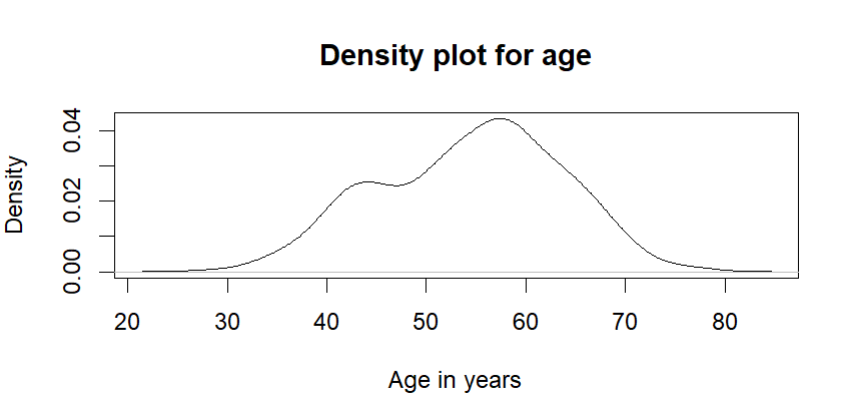
From the histogram, it can be observed that maximum age exceeds 70 and minimum age is somewhere around in the range below 30 years.

* + Visualization of Age using box plot.



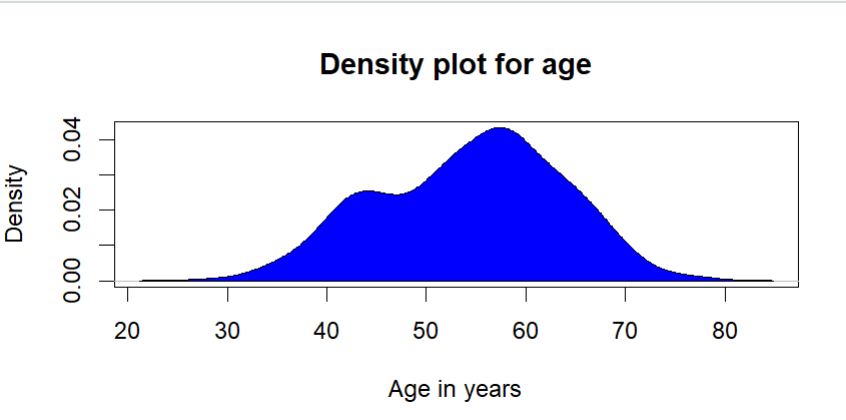
From the above box plot it is seen that the majority of the records in the dataset fall in the range between 40 to 70 years.

* + Viewing the density of age variable in the heart\_record file using density plot.

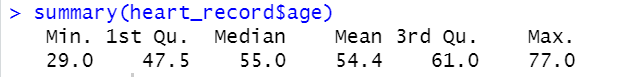


From the above density plot it is evident that the maximum and minimum ages fall int eh range of 40 and 70 years respectively.

* + Viewing density plot in blue color.

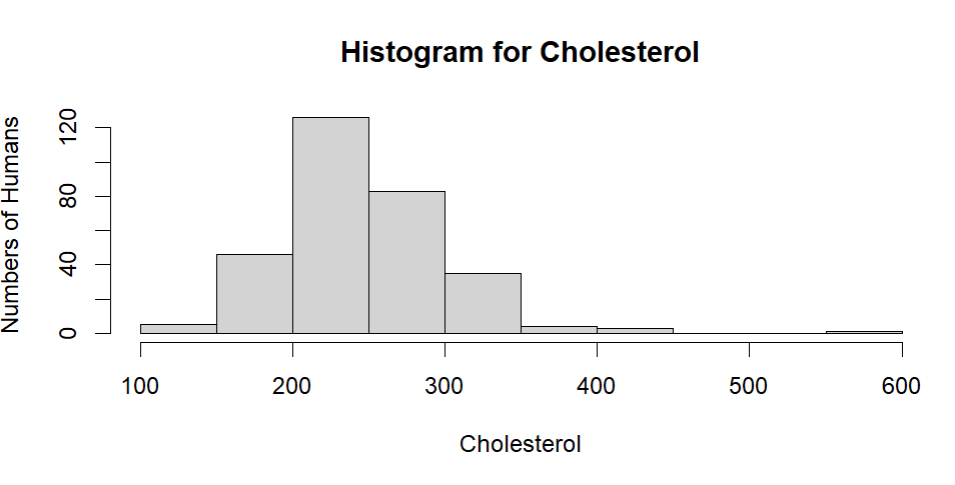


* + For knowing the summary such as mean = 29, 1st quantile = 47.5, median = 55, mean = 54.4, 3rd quantile = 61, maximum = 77 of the age variable. It is shown in the figure.



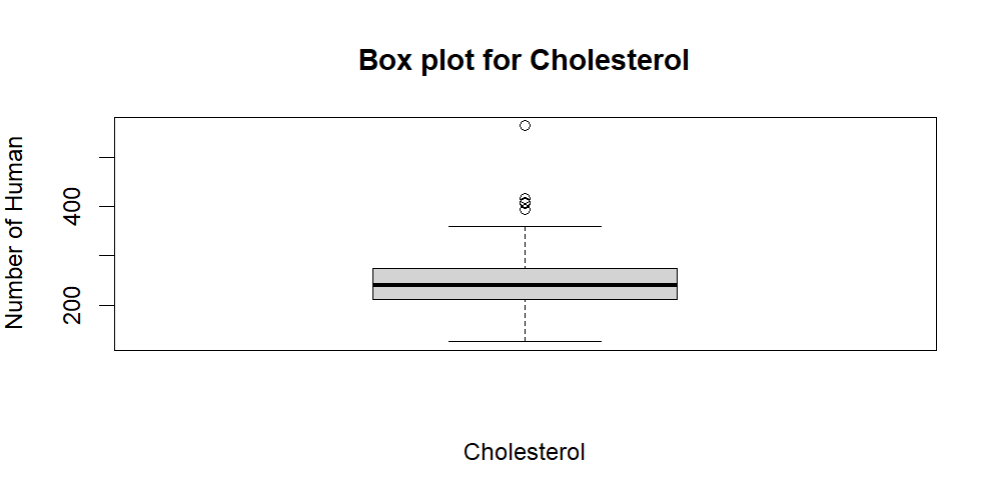
C) To know maximum and minimum **chol**(**cholesterol)** present in the file.

* + Unique cholesterol variables which are varies from 126 to 564.
  + Knowing and displaying the maximum cholesterol= 126 and minimum cholesterol= 564 accordingly with the help of max\_cholestrol and min\_cholestrol variables.
  + Histogram of cholestrol is as shown below.



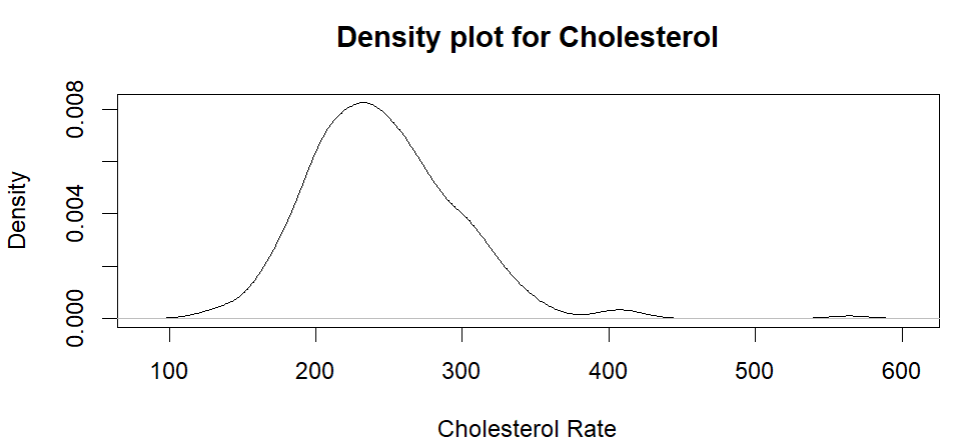
From the Histogram of cholesterol, it is observed that the maximum cholesterol occurs in the range if 200 mg/dl to 250 mg/dl in 120 number of humans.

* + Visualization of Cholesterol using box plot.

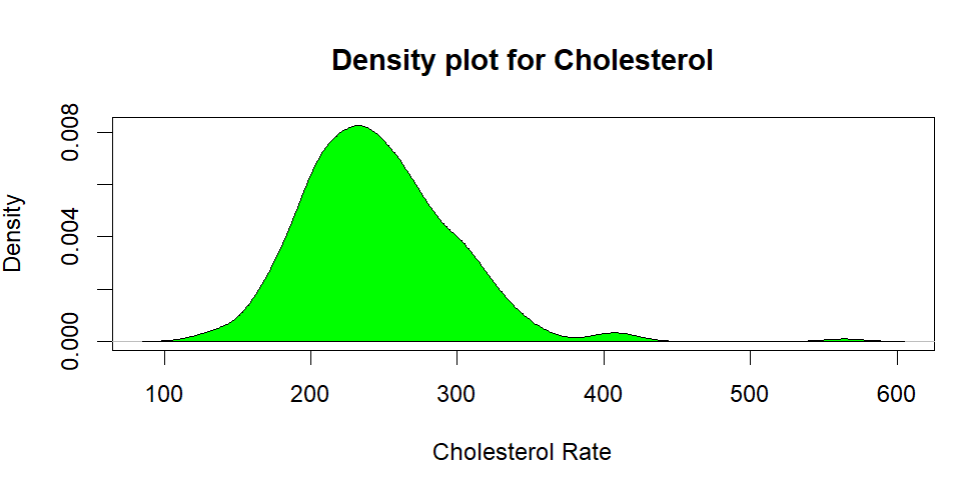


From the above plot, it can be observed that there are some outliers in the representation which can be ignored, but the maximum number of cholesterols occurs in the range of 200 to 250 humans.

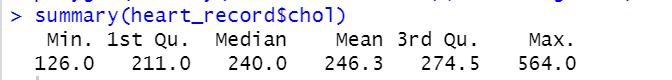
* + Viewing the density of cholesterol variable in the heart\_record file using density plot.



* + Viewing density plot in green color.

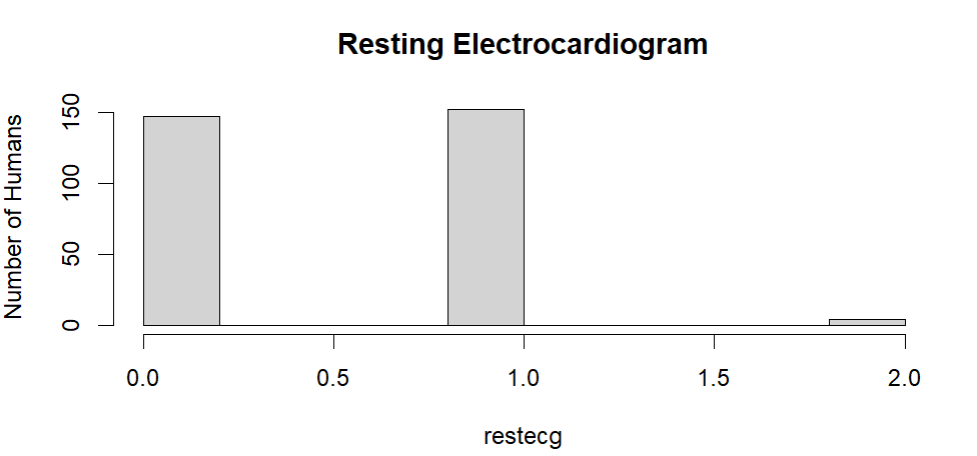


* + For knowing the summary such as mean = 126, 1st quantile = 211, median = 240, mean = 246.3, 3rd quantile = 274.5, maximum = 564 of the chol variable. It is shown in the figure.



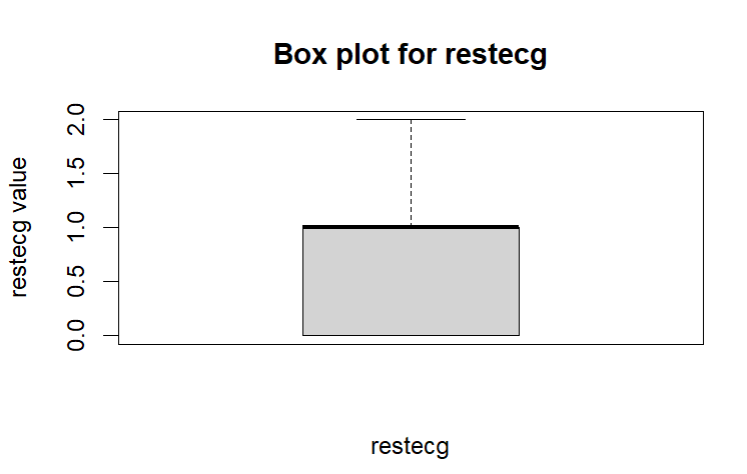
D) To know maximum and minimum **restecg** (**Resting electrocardiogram)** present in the file.

* + Unique **restecg** variables which are ranges from 0, 1 and 2.
  + Knowing and displaying the number of humans present in 0, 1 and 2 ranges of **restecg.**
  + Knowing and displaying the min\_restecg = 0 and max\_restecg = 2 accordingly with the help of min\_restecg and max\_restecg variables.
  + Histogram plot for the Resting electrocardiogram is as shown below.



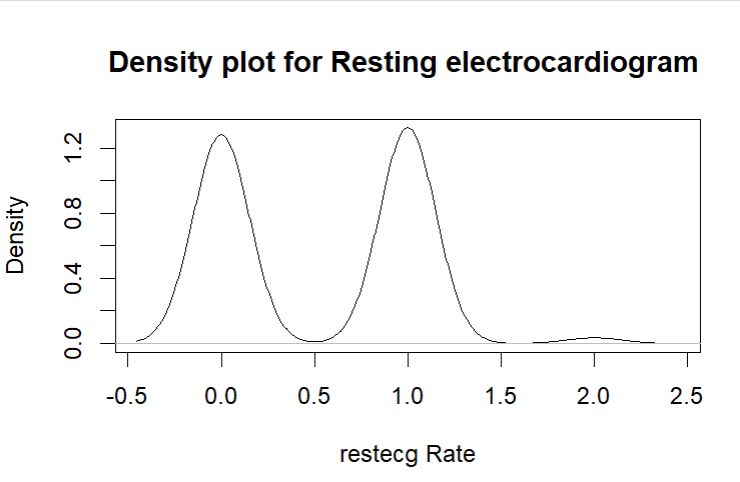
From the Histogram of restecg, it is observed that the maximum restecg occurs in the range if 0 in 150 humans, 1 in 150 humans and 2 in 3 humans.

* + Visualization of **restecg** using box plot.

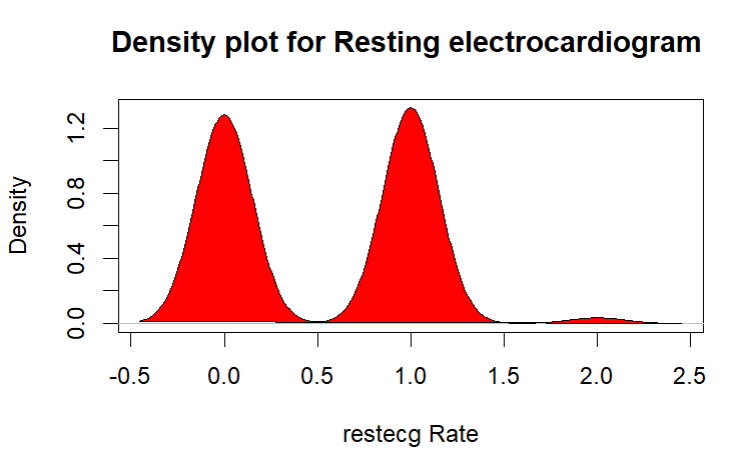


From the above plot we can view the resting electrocardiogram.

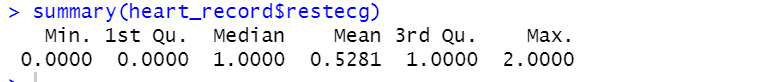
* + Viewing the density of resting electrocardiogram variable in the heart\_record file using density plot.



* + Viewing density plot in red color.



* + For knowing the summary such as mean = 0, 1st quantile = 0, median = 1, mean = 0.5281, 3rd quantile = 1, maximum = 2 of the restecg variable. It is shown in the figure.



Further the column names are change of cp to chestpain, chol to cholesterol , fbs to fasting\_blood\_sugar , restecg to resting\_electrocardio , caa to major\_vessels and trtbps toresting\_blood\_pressure for the better understanding of the columns used for analysis purpose.

**II. Hypothesis set:**

Q1. Does cholesterol have more effect for chances of heart attack?

H0: cholesterol does not have any effect heart attack (Null Hypothesis).

H1: cholesterol has effect on heart attack (Alternate Hypothesis).

The Dependent variable = cholesterol and the

Independent variable = heart attack (output)

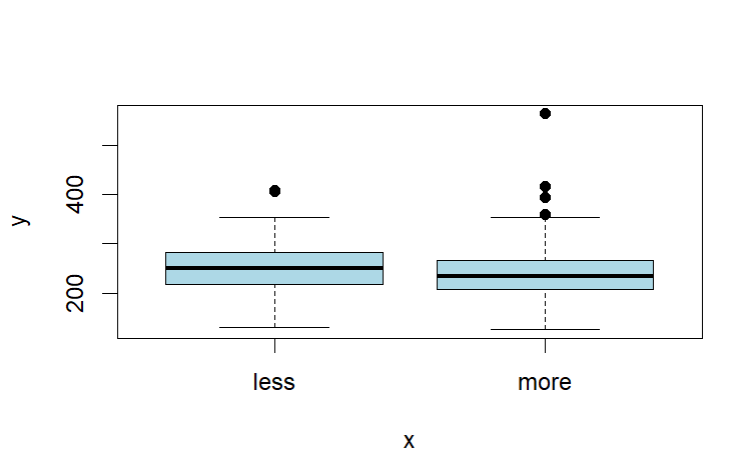
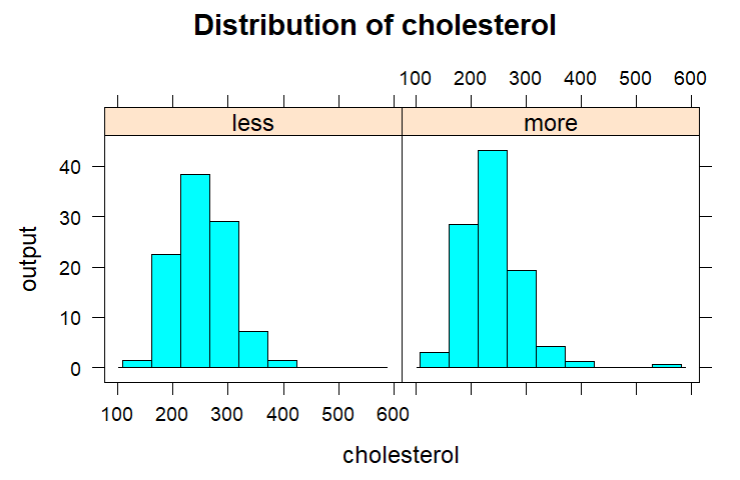
* Box Plotting for better visualization of cholesterol. 

Figure: Plotting of cholesterol for better visualization using plot function.

* Visualization of cholesterol using histogram



Q2. Does the old peak have relation with heart attack?

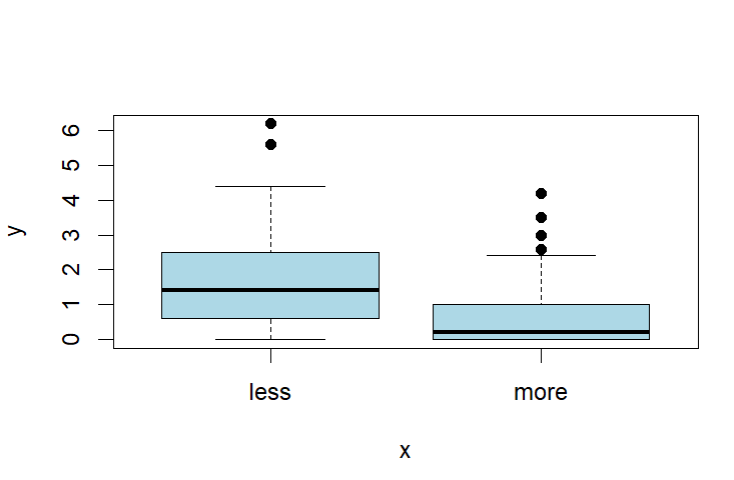
H0: old peak has lower chance of heart attack (Null Hypothesis).

H1: old peak has higher chance of heart attack (Alternate Hypothesis).

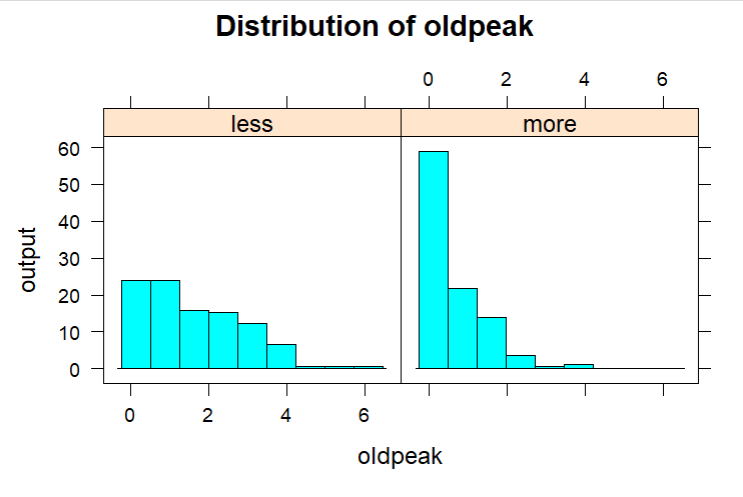
The Dependent variable = oldpeak and the

Independent variable = heart attack (output)

* Box Plotting for better visualization of old peak



* Visualization of old peak
* using histogram



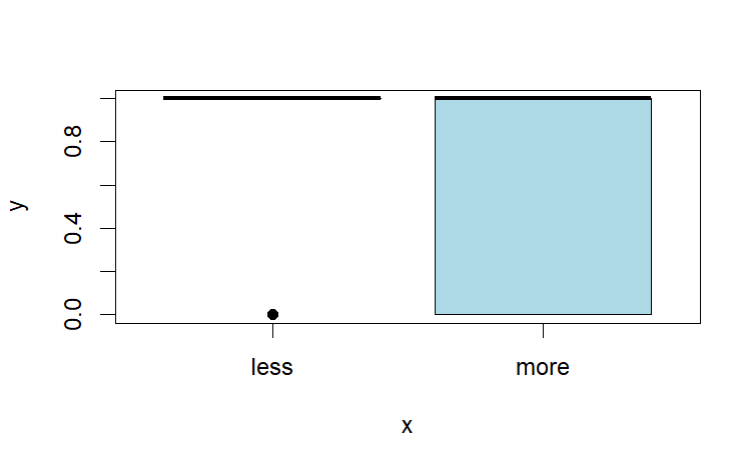
Q3. Does maximum heart rate (thallachh) has an impact on sex?

H0: maximum heart rate (thallachh) does not have an impact on sex (Null Hypothesis).

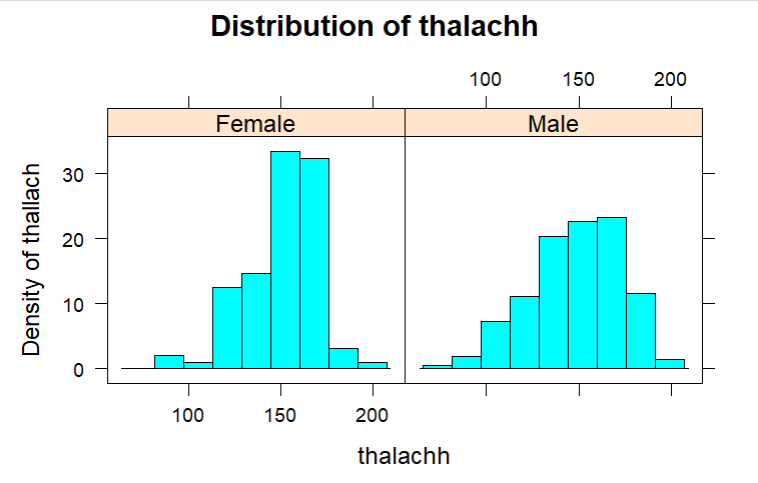
H1: maximum heart rate (thallachh) have an impact on sex (Alternate Hypothesis).

The Dependent variable = sex and the Independent variable = thallachh

* Box Plotting for better visualization of thallachh.



* Visualization of thallachh using histogram



Q4. Does resting blood pressure (trtbps) have any impact on fasting blood sugar (fbs)?

H0: resting blood pressure (trtbps) does not have an impact on fasting blood sugar (fbs)

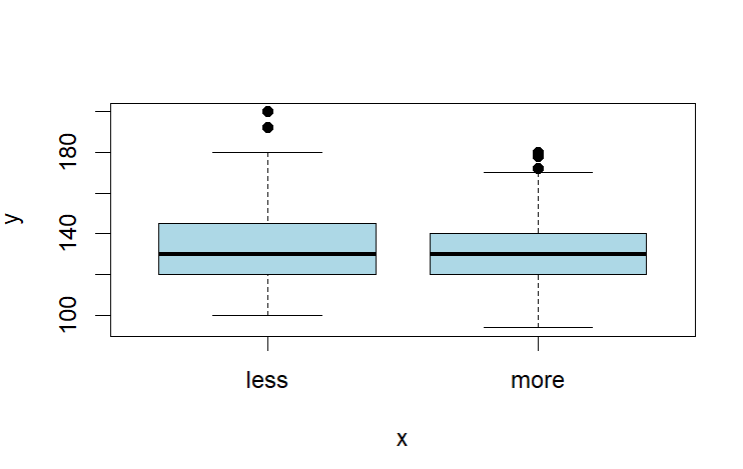
(Null Hypothesis).

H1: resting blood pressure (trtbps) does have an impact on fasting blood sugar (fbs)

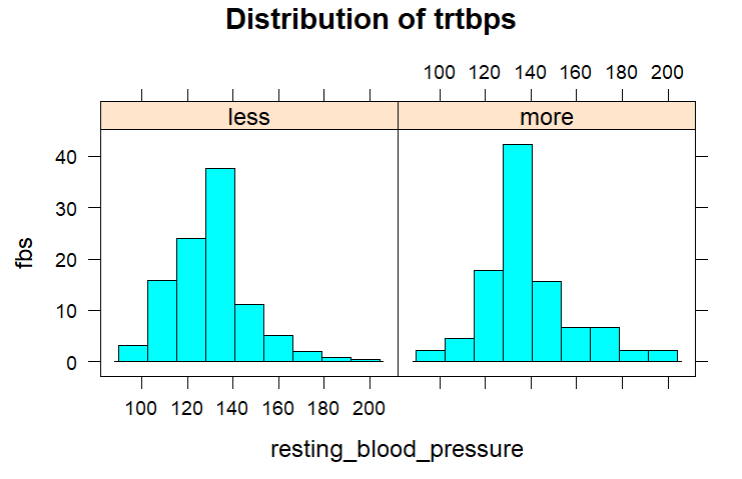
(Null Hypothesis).

The Dependent variable = fbs and the Independent variable = trtbps

* Box Plotting for better visualization of trtbps.



* Visualization of trtbps using histogram

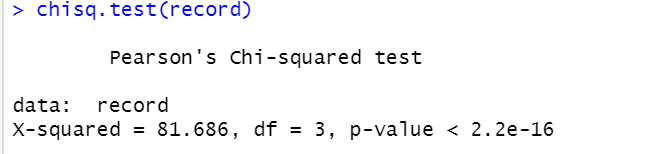


Q5. Does chest Pain have a relationship with heart attack?

H0 : Chest Pain and Heart attack are not related to each other(Null Hypothesis)

H1 : Chest Pain and Heart attack are related to each other (Alternate Hypothesis)

Here both the variables are Dependent variables.



The records are dispayed using the chisq.test(record)

**III. Statistical Methods**

Formal test of normality has been performed on dependent variable as method of statistical tests. In case we get the p-value is less than 0.05, then the data is not normally distributed and if p-value is more than 0.05 then the data is normally distributed.

**For Hypothesis 1:**

For formal test of normality, we used shapiro.test and the p-value is 1.46e-06 which is clearly lower than 0.05 and hence distribution is not normal.

**For Hypothesis 2:**

For formal test of normality, we used shapiro.test and the p-value is 8.18e-17 which is clearly lower than 0.05 and hence distribution is not normal.

**For Hypothesis 3:**

For formal test of normality, we used shapiro.test and the p-value is 6.620819e-05 which is clearly lower than 0.05 and hence distribution is not normal.

**For Hypothesis 4:**

For formal test of normality, we used shapiro.test and the p-value is 1.458097e-06 which is clearly lower than 0.05 and hence distribution is not normal.

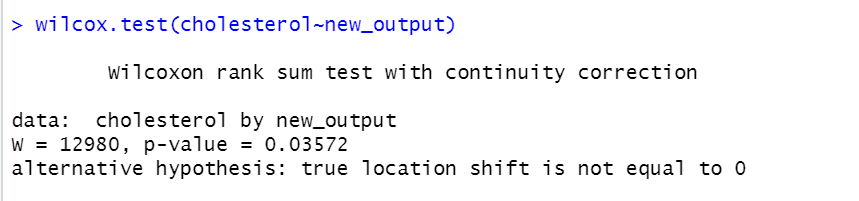
**For Hypothesis 5:**

Since both the variables are dependent variable we perform Pearson's Chi-squared Test for Data.

**IV. Results**

**For Hypothesis 1:**

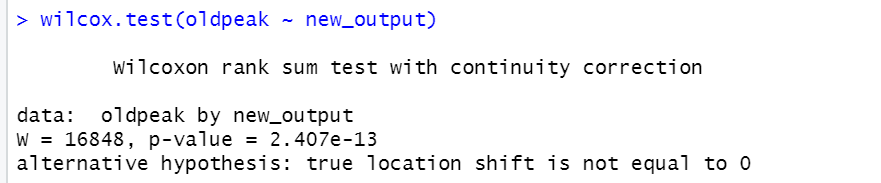
The wilcox.test also known as Mann-Whitney test have been applied to examine the dependent continuous variable (cholesterol) with an independent categorical variable (new\_output). We get the output as follow.

****

**from Mann-Whitney test we get p-value = 0.03572 is** much less than the alpha value = 0.05, we reject the Null hypothesis H0 and accept the Alternate hypothesis H1.

Hence it can be mentioned that cholesterol has effect on heart attack.

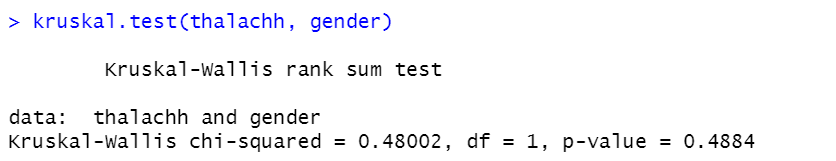
**For Hypothesis 2:**

****

**from Mann-Whitney test we get p-value = 2.407e-13** ismuch less than the alpha value = 0.05, we reject the Null hypothesis H0 and accept the Alternate hypothesis H1.

Hence it can be said that old peak has higher chance of heart attack.

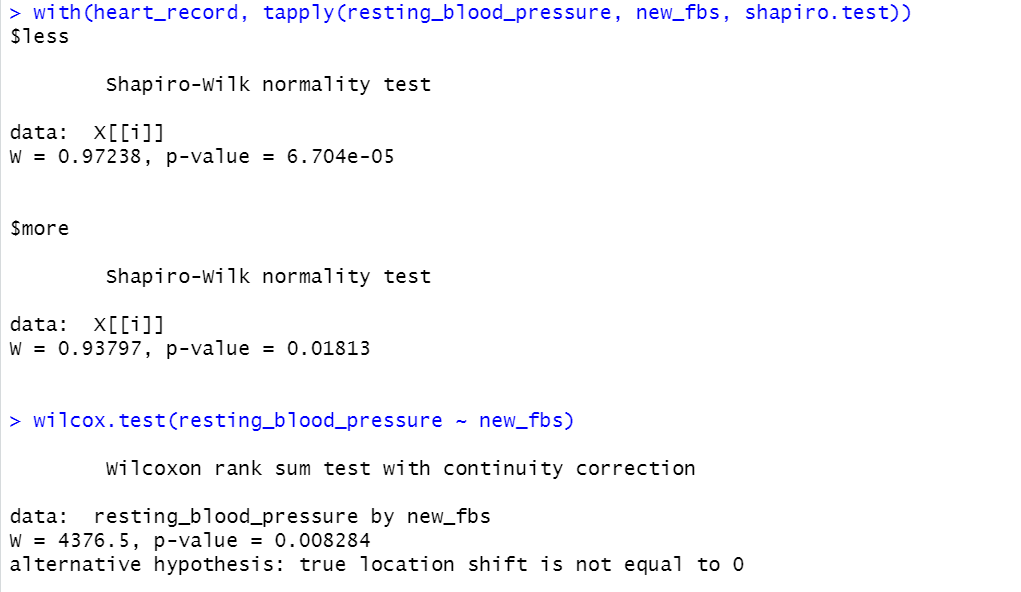
**For Hypothesis 3:**

****

**from kruskal.test we get p-value = 0.4884** and is more than 0.05 the Null hypothesis is excepted and the alternative hypothesis is rejected**.**

Hence it can be said that maximum heart rate (thallachh) does not have an impact on sex (Null Hypothesis).

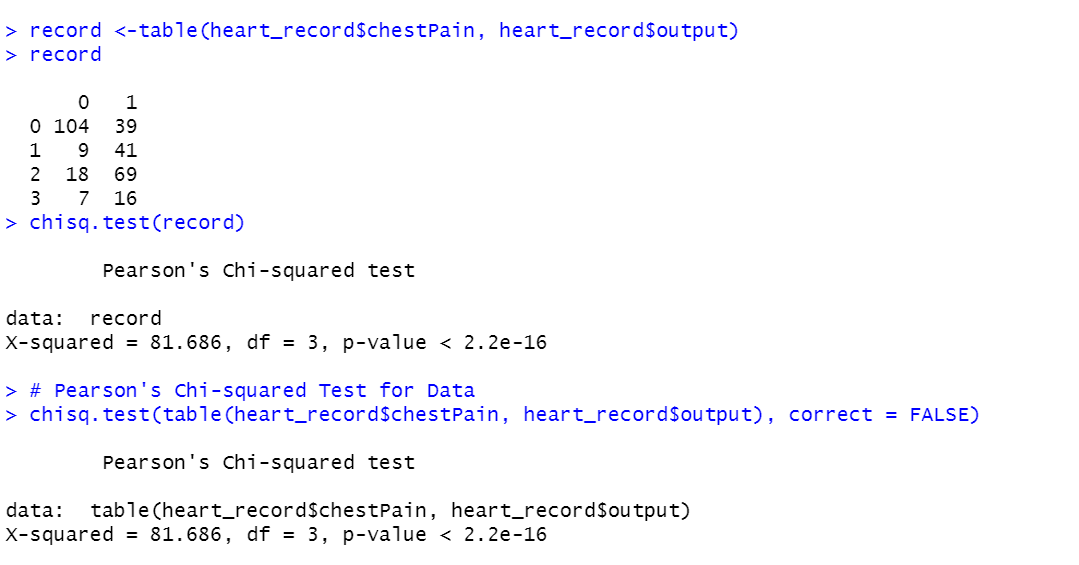
**For Hypothesis 4:**

****

**from Mann-Whitney test we get p-value = 0.008284** ismuch less than the alpha value = 0.05, we reject the Null hypothesis H0 and accept the Alternate hypothesis H1.

Hence it can be said that resting blood pressure (trtbps) does have an impact on fasting blood sugar (fbs)

**For Hypothesis 5:**

****

**For Pearson's Chi-squared Test for data** we get the p value < 2.2e-16, the Null hypothesis is rejected H0 and the alternative hypothesis H1 is excepted**.**

Hence it can be said that Chest Pain and Heart attack are related to each other.

**V. Conclusion**

With the help of analysis performed, we can mention that there is various component in human blood as well as some medical factors contributing to increase the chances of heart attack. Out of them few we taken into analysis and hypothesis testing. Hence to recapitulated it can be said that

* + Cholesterol(chol) has effect on heart attack.
  + oldpeak has higher chance of heart attack.
  + The maximum heart rate (thallachh) does not have an impact on sex.
  + The resting blood pressure (trtbps) does have an impact on fasting blood sugar (fbs)
  + And finally, the chest Pain and Heart attack are related to each other.

With proper exercise and good habits, we can reduce the chances of heart attack.

**VI. Reference**

1. S. Manikandan, "Heart attack prediction system," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 817-820, doi: 10.1109/ICECDS.2017.8389552.
2. M. Nabeel, M. J. Awan, M. Raza, H. Muslih-Ud-Din and S. Majeed, "Heart Attack Disease Data Analytics and Machine Learning," 2021 International Conference on Innovative Computing (ICIC), 2021, pp. 1-6, doi: 10.1109/ICIC53490.2021.9692977.