MCSD1113 STATISTIC FOR DATA SCIENCE

PROJECT

LECTURER'S NAME:
Dr. Mohamad Shukor Talib

STUDENT NAME	MATRIC NO.
MOHD NOR BIN MOHIDIN	MCS231008
ZUHAYR ARIF BIN ZAKARIA	MCS231002
NUR AZIMAH BINTI MOHD SALLEH	MCS231011

TABLE OF CONTENTS

HEART DISEASE PREDICTION

1.0 INTRODUCTION OF THE PROJECT

In this project, we will perform the exploratory data analysis and inference statistical analysis from the Heart Disease University of California Irvine's (UCI) dataset to understand and interpret the connection between variables. The Heart Disease UCI dataset provides valuable information on various clinical and demographic factors associated with the presence or absence of heart disease. Below is the expected insight that we want to find from this project.

i) Risk Factors Identification: Identifying known risk factors for heart disease such as age, sex, cholesterol levels and blood pressure to discover what are the factors that influence and cause someone to suffer from heart disease.

ii) Correlation Analysis: We anticipate finding correlations whether positive or negative correlations between different variables in the dataset.

iii) Visualization of Relationships: Visually explore the relationships between variables through bar plot, histograms, boxplots, etc to identify trends, outliers, and patterns.

iv) Statistical Inference: We can test hypotheses and make inferences about the population based on sample data by performing inferential statistical analysis.

Therefore, we aim to gain a deeper understanding of the complex interplay between various factors and their association with heart disease, ultimately improving our knowledge and skills in data science statistics.

1.1 Dataset

This dataset is hosted on Kaggle (https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction/data), and it was from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Heart+Disease. A dataset named "Heart_Disease_Prediction.csv" consists of 14 attributes and 270 rows. The "Heart.Disease" field represents the absence or presence of the heart disease. Below is the attribute information in the data set:

Table 1.1: The Attributes of the Heart Disease Prediction

No.	Attribute	Type of Data	Description
1.	Age	Integer/ Ratio	The person's age in years
2.	Sex	Categorical/ Nominal	The person's sex (1 = male, 0 = female)
3.	Chest.pain.type	Categorical /Ordinal	The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4.	BP	Integer/Ratio	The person's resting blood pressure (mm Hg on admission to the hospital)
5.	Cholesterol	Integer/Ratio	The person's cholesterol measurement in mg/dl
6.	FBS.over.120	Categorical/ Nominal	The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
7.	EKG.results	Categorical/ Ordinal	Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
8.	Max.HR	Integer/ Ratio	The person's maximum heart rate achieved
9.	Exercise.angina	Categorical/ Nominal	Exercise induced angina (1 = yes; 0 = no)
10.	ST.depression	Integer/ Ratio	ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.)
11.	Slope.of.ST	Categorical/ Ordinal	The slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
12.	Number.of.vessels.fluro	Categorical /Ordinal	The number of major vessels (0-3)
13.	Thallium	Categorical/ Ordinal	A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
14.	Heart.Disease	Categorical/ Nominal	Heart disease (Absence, Presence)

1.2 Data Exploratory

1.2.1 Data Pre-Processing

Based on the ncol() and nrow() functions, it can be identified that the dataset consists of 14 columns and 270 rows. For the preliminary understanding of the data, we employed the head() function, which enabled us to observe the overview of the dataset. As data cleaning processes, we examined the presence of missing values with sum(is.na()) and colSums(is.na()) functions, and fortunately, we found that the dataset was complete, with no missing values in any of the variables. Subsequently, to know the structure of the dataset, we used the str() function for a concise overview, providing insights into the variable types and presenting a subset of observations from each column. This provided us with an initial understanding of the data types we are dealing with. For the detailed type of data and level of data, we have already identified first as presented in Table 1.1.

1.2.2 Change Data Type

Based on Figure 1.3 in the appendix, the data types are mostly integers and numeric. In the case of categorical variables, to visualize these data on plots, we would have to change the data types into factors, which are special vectors that represent categorical data. While factors look (and often behave) like character vectors, they are actually integers with labels. Since the values for categorical data are mostly 0,1,2,3 to represent their details as mentioned in Table 1.1, we decided to give a better label name for these factor values for easier to read and analyse. Figure 1.4 in the appendix shows the result after change the values for particular variable into factors.

2.0 DATA DESCRIPTIVE

We have performed data descriptive which can give insight into the relationship, distribution, patterns, and tendencies of the variables associated with the heart disease variable.

2.1 Bar chart

We created bar plots to identify the distribution of heart disease and also to identify the relationship between heart disease variable with other variables. As shown in Figure 2.1 in the appendix, the distribution for absence and presence of the heart disease is quite balanced.

2.1.1 Relationship Between Other Variables with Heart Disease Variables

To discover the relationship between other variables with heart disease variables, we have chosen the eight (8) variables which are Sex, Chest.pain.type, FBS.over.120, EKG.results, Exercise.angina, Slope.of.ST, Number.of.vessels.fluro, and Thallium to provide descriptive analysis regarding the possibility of factors that causing the heart disease by using the bar chart. Based on the Figure 2.2 in the appendix, below is the finding for the eight variables analysis of the heart disease variable: -

i) Sex vs heart disease: There are approximately half the observation of women than men. This indicate, men are more likely to have a heart disease than women.

ii) Chest.pain.type vs heart disease: The data description does not specify how the pain classification was determined. However, it appears that identifying heart disease in patients solely based on their symptoms can be challenging.

iii) FBS.over.120 vs heart disease: Determining the sugar level in the blood alone doesn't provide much insight into whether a patient has heart disease or not, as it's not a definitive indicator.

iv) EKG.results vs heart disease: It indicates that certain abnormalities in the rest EKG may serve as strong indicators of the presence of heart disease.

v) Exercise.angina vs heart disease: This specific characteristic is deemed to be a reliable indicator of the presence of heart disease. However, it is worth noting that identifying whether or not a patient is experiencing angina is not always straightforward as it can be confused with other types of pain as atypical angina.

vi) Slope.of.ST vs heart disease: Plot shows if the slope is downsloping, the depression of the ST segment can help to determine if the patient has a heart disease.

vii) Number.of.vessels.fluro vs heart disease: This feature indicates the higher value of this feature are more likely to have heart disease.

viii) Thallium vs heart disease: From the plot, it can be seen that reversable defects are more likely to cause heart disease.

2.2 Pie Chart

We use pie chart to display the proportion of variable for categorical data. However, we found that this pie chart does not describe any useful information regarding relationship of the data among each other. Therefore we just choose two variables that have more than three categories to illustrate the numerical proportion of the variables which are Chest.pain.type and Number.of.vessel.fluro as presented in Figure 2.3 and 2.4 in the appendix.

2.3 Stem & Leaf

For Stem & leaf plots, we selected three variables which are BP, Cholesterol, and Max.HR to show the distribution of data since these three variables fall into continuous data with ratio level as presented in Figure 2.5, 2.6 and 2.7 in the appendix.

2.4 Histogram

For histogram plot, we chosen five continuous variables which are Age, BP, Cholesterol, Max.HR and ST.depression to show the frequency of data, the shape and spread of the data and also to discover the relationship between these five variables with heart disease variable. Based on the Figure 2.8 in the appendix, below is the finding for the five variables analysis of the heart disease variable: -

i) Age vs heart disease: It can be seen that the age is a risk factor where the higher the age, the more likely that the patient has a heart disease.

ii) BP vs heart disease: By the different peaks, looks like most people tend to have a normal blood pressure inside certain groups. It also looks like very high pressures can indicate that there is a heart disease.

iii) Cholesterol vs heart disease: The histogram shows that the majority of people in the dataset have high levels of cholesterol. It also indicates that up to a certain level, the presence of heart disease is slightly higher in those with higher levels of cholesterol.

iv) Max.HR vs heart disease: The histogram may seem odd since it shows that higher heart rates are associated with a lower risk of heart disease. This can be explained by the fact that the maximum healthy heart rate is dependent on one's age (220 minus age). Younger individuals often have higher heart rates.

v) ST.depression vs heart disease: As can be seen in the plot, a significant displacement of this segment could indicate the presence of a heart disease.

2.5 Box Plot

By creating boxplots for each continuous variable which are Age, BP, Cholesterol, Max.HR and St.depression, we visually examined the distributions and identified any potential outliers. These boxplots revealed variations in the range, median, and interquartile range of each variable. In the appendix, Figure 2.9 shows that there are outliers in all variables except for Age. These outliers were observed in patients' blood pressure, cholesterol levels, maximum heart rate and electrocardiogram (ECG/EKG) readings. It is important to consider these outliers in the analysis rather than ignoring them.

2.6 Descriptive Analysis

We computed summary statistics for the continuous variables, including minimum, maximum, mean, median, and quartiles. The summary statistics as shown in Figure 2.10 in the appendix provided a concise overview of the central tendencies and spread of the data, enabling us to quickly assess the dataset's characteristics. Based on these summaries, we can see that the average of age is 54 where average blood pressure is 131 with average of cholesterol and maximum heart rate is 249.7 and 149.7 respectively.

3.0 INFERENTIAL ANALYSIS

In this section, we would test the hypothesis and make inferences about the population based on sample data. These includes hypothesis testing 1-sample and 2-sample, goodness of fit test, Chi Square test of independence, correlation, regression, and Analysis of Variance (ANOVA).

3.1 Hypothesis Testing 1-Sample.

In this experiment, we conducted one-sample hypothesis tests to determine if the average Age in the dataset is significantly different from age of 54 as shown in Figure 3.1 in the appendix.

Hypothesis:
H_0: The mean age of heart disease equal to 54. ($\mu = 54$)
H_1: The mean age of heart disease is not equal to 54 ($\mu \neq 54$)

The mean age in the dataset is calculated to be approximately 54.43 year. The obtained p-value (3.795e-14), is less than the conventional significance level of 0.05 suggesting that we reject the null hypothesis. Therefore, there is strong evidence to reject the null hypothesis and conclude that the true mean age in the population is not equal to 54.

3.2 Hypothesis Testing 2-Sample.

For hypothesis testing 2 sample, we compared the means of two independent samples when the variances are assumed to be unequal or unknown. This statistical testing performed on the continuous variable Age, comparing individuals with heart disease (‘Presence’) to those without heart disease (‘Absence’) as shown in Figure 3.2 in the appendix.

Hypothesis:

H_0: There is no significant difference in the mean age between individuals with heart disease and those without. ($\mu 1 = \mu 2$)

H_1: There is a significant difference in the mean age between the two groups. ($\mu 1 \neq \mu 2$)

The mean age in the dataset is calculated to be approximately 56.59 year with heart disease and 52.71 without heart disease. The obtained p-value (0.0003526) is less than the conventional significance level of 0.05 suggesting that we reject the null hypothesis. Therefore, there is strong evidence to reject the null hypothesis and conclude that there is a statistically significant difference in mean ages between individuals with heart disease and those without heart disease.

3.3 Goodness of Fit Test

The chi-square goodness-of-fit test was conducted to assess whether the observed frequencies of categories in the variable Chest.pain.type aligns with the expected probabilities. The categories under consideration are "Typical angina," "Atypical angina," "No angina," and "Asymptomatic".

Hypotheses:

H_0: The observed frequencies match the expected frequencies based on the specified probabilities.

H_1: The observed frequencies do not match the expected frequencies based on the specified probabilities.

Based on the Figure 3.3 in the appendix, the observed frequencies for the "Typical angina," "Atypical angina," "No angina," and "Asymptomatic" are 20, 42, 79 and 129 respectively. We defined the expected probabilities for each type are 0.10, 0.20, 0.30 and 0.40. The chi-squared test yielded a test statistic of 8.6142 with 3 degrees of freedom and a p-value of 0.03489. Consequently, since the p-value (0.03489) is less than the conventional significance level of 0.05, we reject the null hypothesis and conclude there is sufficient evidence to suggest that the observed frequency of chest pain types do not match the expected frequencies.

3.4 Chi Square Test of Independence

This test is to investigate the potential association between gender (Sex) and the presence of exercise-induced angina (Exercise.angina). Figure 3.4 in the appendix shows the result for chi square test between these two variables.

Hypotheses:

H_0: There is no association between gender and exercise-induced angina.

H_1: There is an association between gender and exercise-induced angina.

The obtained p-value of 0.004809 is less than the conventional significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is evidence to support the presence of an association between gender and exercise-induced angina.

3.5 Correlation

We examined the correlation between Blood Pressure and Cholesterol, as represented in the scatter plot below: -

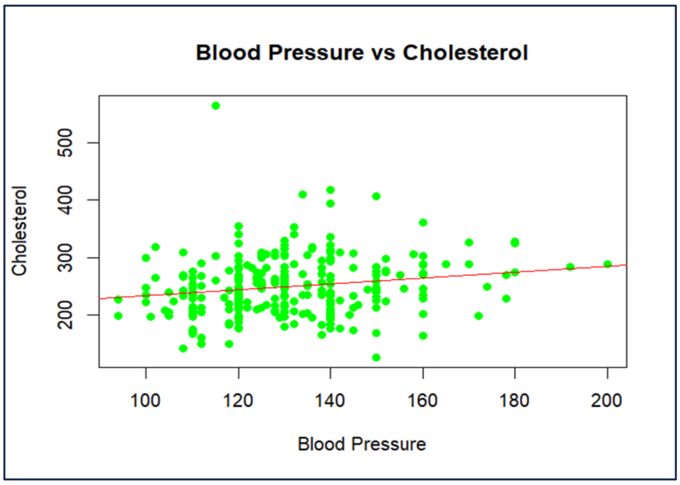


Figure 3.5: Correlation between Blood Pressure and Cholesterol

Based on the scatter plot and Figure 3.5 in the appendix, it can be seen that the Pearson correlation coefficient is approximately 0.173. Since the correlation coefficient is positive, it indicates a positive linear relationship between blood pressure and cholesterol levels. This means that as one variable increases, the other tends to increase as well, and vice versa. However, it's important to note that while the correlation is statistically significant, the strength of the correlation is relatively weak (closer to 0), suggesting that the relationship between these variables is not particularly strong.

3.6 Regression

For a linear regression model in R, we examined the relationship between the independent variable Age and the dependent variable Max.HR as represented in the scatter plot below: -

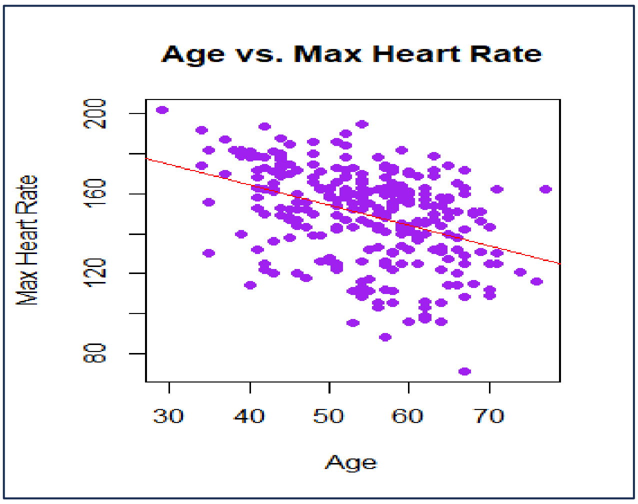


Figure 3.6: Linear Regression between Age and Max Heart Rate

Based on the scatter plot and Figure 3.6 in the appendix, it shows that the intercept which represent the value of the Max.HR when the Age is zero is approximately 205.3573. Meanwhile the coefficient of Age represents the change in the Max.HR for a one-unit change in the Age is approximately -1.0229. Therefore, the equation that represents the relationship between Age and Max.HR is:

Max.HR = 205.3573 - 1.0229 * Age

This equation indicates that for every one-year increase in Age, the Max.HR is expected to decrease by approximately 1.0229 beats per minute.

Figure 3.7 also shows the F-statistic tests for the significance of the regression model. In this output, the F-statistic is 51.72 with a very small p-value (6.386e-12), indicating that the overall regression model is highly significant. Therefore, it can be concluded that that there is a significant relationship between Age and Max.HR in this dataset. Specifically, as Age increases, Max.HR tends to decrease, with each one-year increase in age associated with a decrease of approximately 1.0229 units in Max.HR.

3.7 Analysis of Variance (ANOVA)

In this dataset, we use ANOVA to test whether there are significant differences in the means of the Max.HR variables across different levels of a Chest Pain Type variables.

Hypotheses:
H_0: The mean Max.HR is the same across different chest pain types.
H_1: The mean Max.HR differs among different chest pain types.

Based on the Figure 3.7 in the appendix, the F value is 13.27 with d.f.n is 3 and d.f.d is 266. Since the obtained p-value (4.22e-08) is extremely small and less than the conventional significance level of 0.05, we reject the null hypothesis and conclude that there are significant differences in Max.HR across different chest pain types.

4.0 CONCLUSION

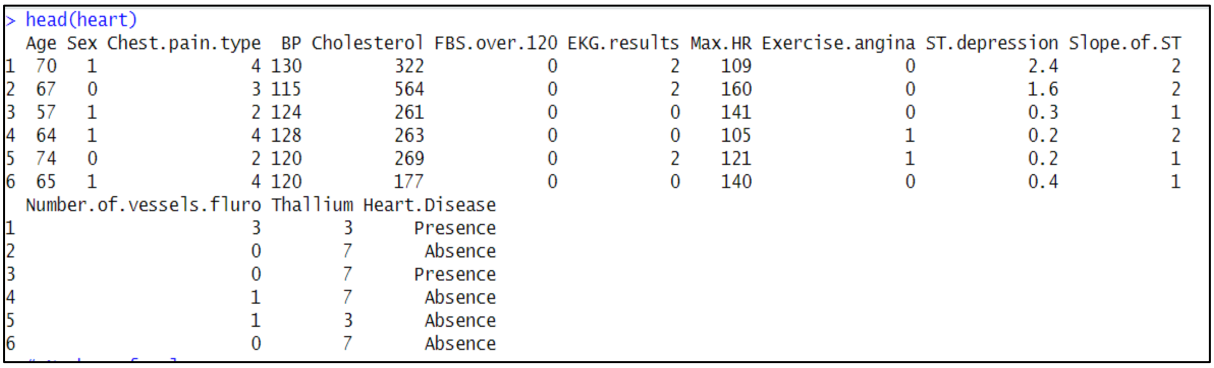
In conclusion, this project uncovered significant insights about the attributes of heart disease in the dataset. From the descriptive analysis, we discovered that men are more likely to have a heart disease than women, patients with a reversible defect on thalassemia are more likely to have heart disease, and the likelihood of a patient having heart disease increases with age.

Through hypothesis testing, we have discovered that there is a significant difference in mean ages between individuals with heart disease and those without. Additionally, we have identified an association between gender and exercise-induced angina. In terms of correlation, we have found a weak positive linear relationship between blood pressure and cholesterol. Furthermore, the results of an ANOVA test comparing means of Max.HR across levels of Chest Pain Type indicates that there are significant differences in Max.HR across different chest pain types. All of these findings provide a comprehensive analysis and enhance our understanding of the various factors that contribute to heart disease.

APPENDIX

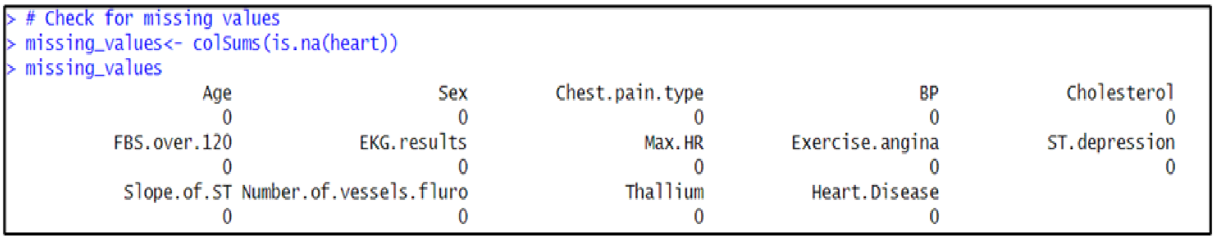
- FIGURES

FIGURE 1.1 – 1.4



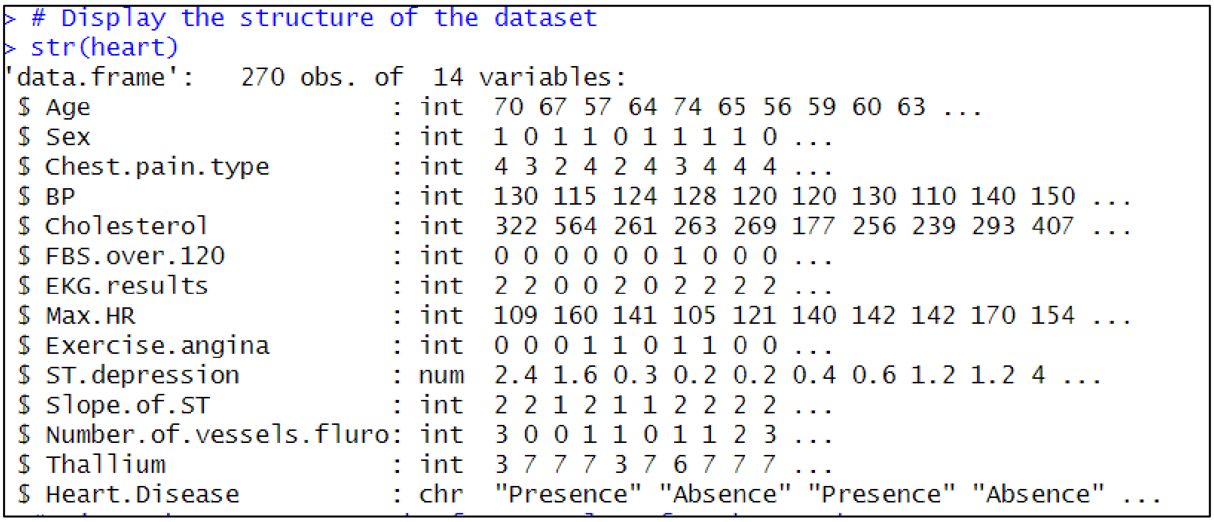
```
> head(heart)
  Age Sex Chest.pain.type  BP Cholesterol FBS.over.120 EKG.results Max.HR Exercise.angina ST.depression Slope.of.ST
1  70   1               4 130         322            0           2    109               0           2.4           2
2  67   0               3 115         564            0           2    160               0           1.6           2
3  57   1               2 124         261            0           0    141               0           0.3           1
4  64   1               4 128         263            0           0    105               1           0.2           2
5  74   0               2 120         269            0           2    121               1           0.2           1
6  65   1               4 120         177            0           0    140               0           0.4           1
  Number.of.vessels.fluro Thallium Heart.Disease
1                       3        3      Presence
2                       0        7       Absence
3                       0        7      Presence
4                       1        7       Absence
5                       1        3       Absence
6                       0        7       Absence
```

Figure 1.1: Overview of the Dataset



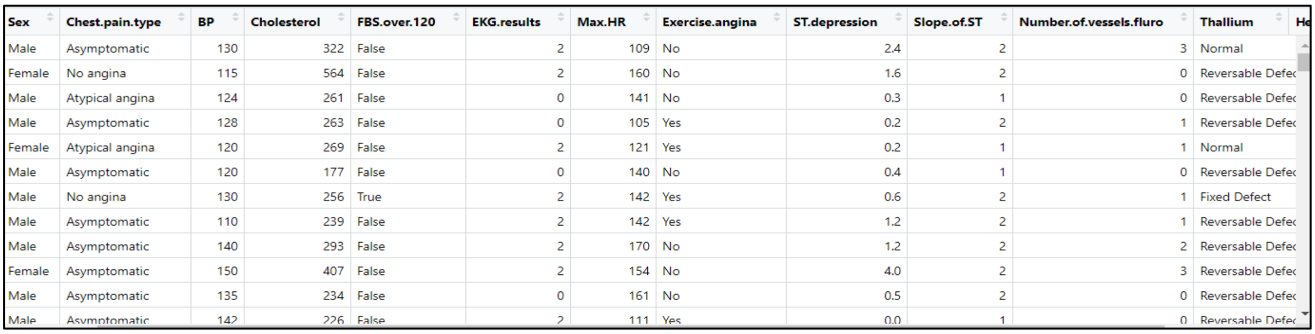
```
> # Check for missing values
> missing_values<- colSums(is.na(heart))
> missing_values
                    Age                     Sex         Chest.pain.type                      BP             Cholesterol
                      0                       0                       0                       0                       0
           FBS.over.120             EKG.results                  Max.HR         Exercise.angina           ST.depression
                      0                       0                       0                       0                       0
            Slope.of.ST Number.of.vessels.fluro                Thallium           Heart.Disease
                      0                       0                       0                       0
```

Figure 1.2: No Missing Values in Dataset



```
> # Display the structure of the dataset
> str(heart)
'data.frame':	270 obs. of  14 variables:
 $ Age                    : int  70 67 57 64 74 65 56 59 60 63 ...
 $ Sex                    : int  1 0 1 1 0 1 1 1 1 0 ...
 $ Chest.pain.type        : int  4 3 2 4 2 4 3 4 4 4 ...
 $ BP                     : int  130 115 124 128 120 120 130 110 140 150 ...
 $ Cholesterol            : int  322 564 261 263 269 177 256 239 293 407 ...
 $ FBS.over.120           : int  0 0 0 0 0 0 1 0 0 0 ...
 $ EKG.results            : int  2 2 0 0 2 0 2 2 2 2 ...
 $ Max.HR                 : int  109 160 141 105 121 140 142 142 170 154 ...
 $ Exercise.angina        : int  0 0 0 1 1 0 1 1 0 0 ...
 $ ST.depression          : num  2.4 1.6 0.3 0.2 0.2 0.4 0.6 1.2 1.2 4 ...
 $ Slope.of.ST            : int  2 2 1 2 1 1 2 2 2 2 ...
 $ Number.of.vessels.fluro: int  3 0 0 1 1 0 1 1 2 3 ...
 $ Thallium               : int  3 7 7 7 3 7 6 7 7 7 ...
 $ Heart.Disease          : chr  "Presence" "Absence" "Presence" "Absence" ...
```

Figure 1.3: The Structure of the Dataset.



Sex	Chest.pain.type	BP	Cholesterol	FBS.over.120	EKG.results	Max.HR	Exercise.angina	ST.depression	Slope.of.ST	Number.of.vessels.fluro	Thallium	He
Male	Asymptomatic	130	322	False	2	109	No	2.4	2	3	Normal	
Female	No angina	115	564	False	2	160	No	1.6	2	0	Reversable Defec	
Male	Atypical angina	124	261	False	0	141	No	0.3	1	0	Reversable Defec	
Male	Asymptomatic	128	263	False	0	105	Yes	0.2	2	1	Reversable Defec	
Female	Atypical angina	120	269	False	2	121	Yes	0.2	1	1	Normal	
Male	Asymptomatic	120	177	False	0	140	No	0.4	1	0	Reversable Defec	
Male	No angina	130	256	True	2	142	Yes	0.6	2	1	Fixed Defect	
Male	Asymptomatic	110	239	False	2	142	Yes	1.2	2	1	Reversable Defec	
Male	Asymptomatic	140	293	False	2	170	No	1.2	2	2	Reversable Defec	
Female	Asymptomatic	150	407	False	2	154	No	4.0	2	3	Reversable Defec	
Male	Asymptomatic	135	234	False	0	161	No	0.5	2	0	Reversable Defec	
Male	Asymptomatic	142	226	False	2	111	Yes	0.0	1	0	Reversable Defec	

Figure 1.4: Dataset after Change Label Name for Values

FIGURE 2.1 – 2.10

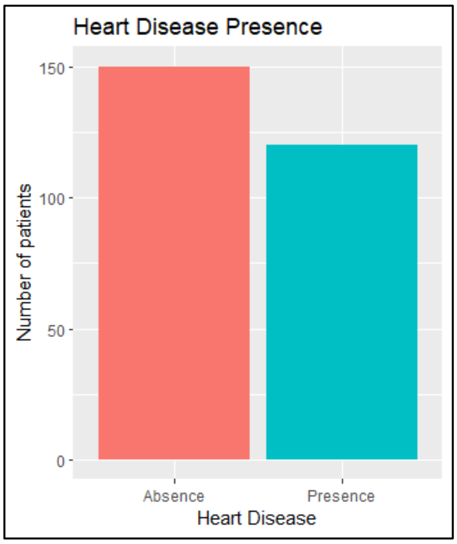


Figure 2.1: Distribution of Heart Disease Variables

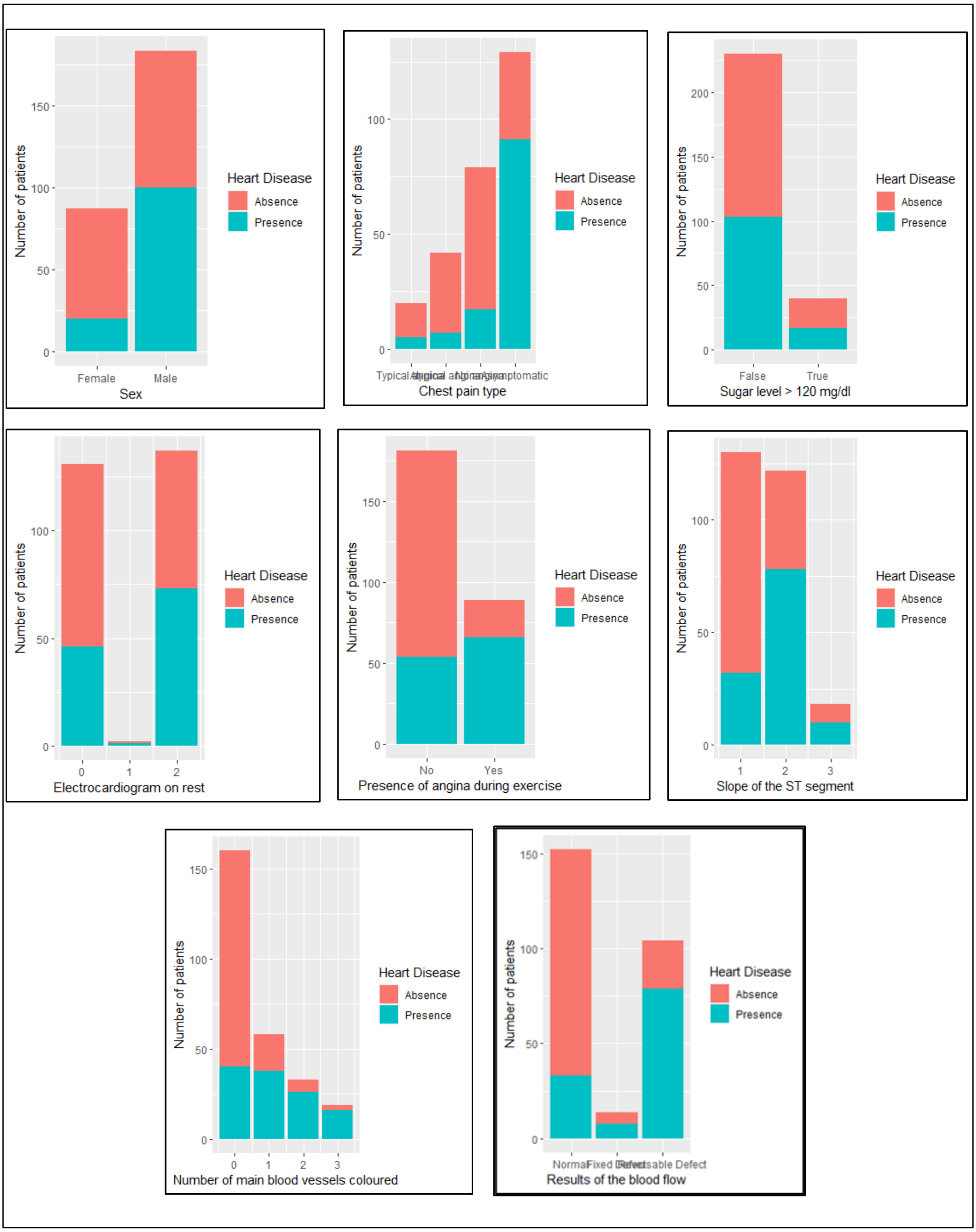


Figure 2.2: Descriptive Analysis for Relationship between Eight Variables with Heart Disease Variable using Bar Plot.

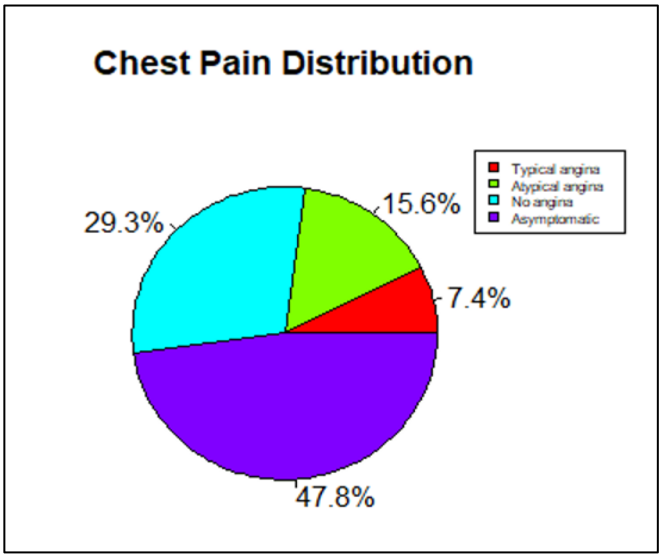


Figure 2.3: Pie Chart of Chest Pain Distribution

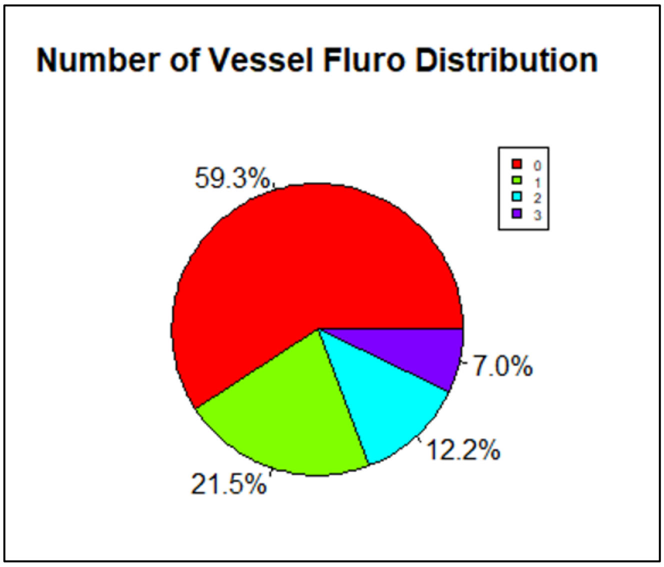
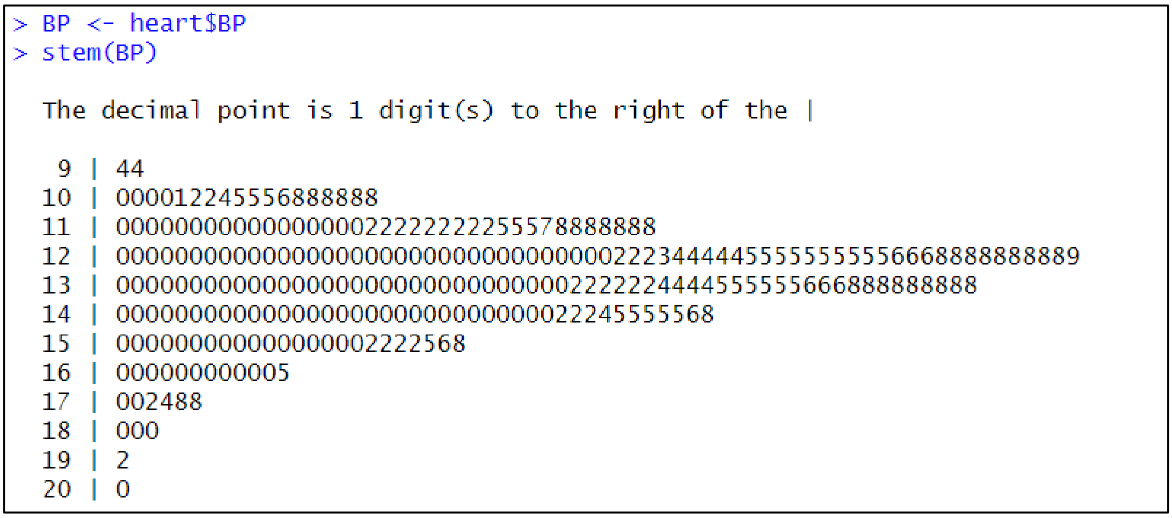


Figure 2.4: Pie Chart of Number of Vesel Fluro Distribution

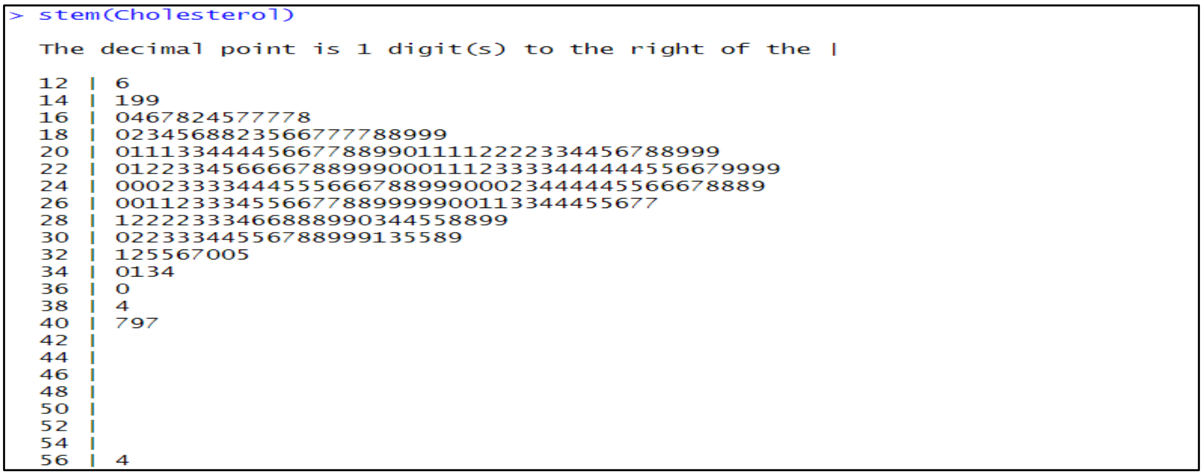


```
> BP <- heart$BP
> stem(BP)

  The decimal point is 1 digit(s) to the right of the |

   9 | 44
  10 | 000012245556888888
  11 | 0000000000000000022222222255578888888
  12 | 00000000000000000000000000000000002223444445555555555666888888888 9
  13 | 0000000000000000000000000000000222222444455555566688888888 8
  14 | 000000000000000000000000000000222455555 68
  15 | 000000000000000002222568
  16 | 000000000005
  17 | 002488
  18 | 000
  19 | 2
  20 | 0
```

Figure 2.5: Stem and Leaf plots for BP Distribution

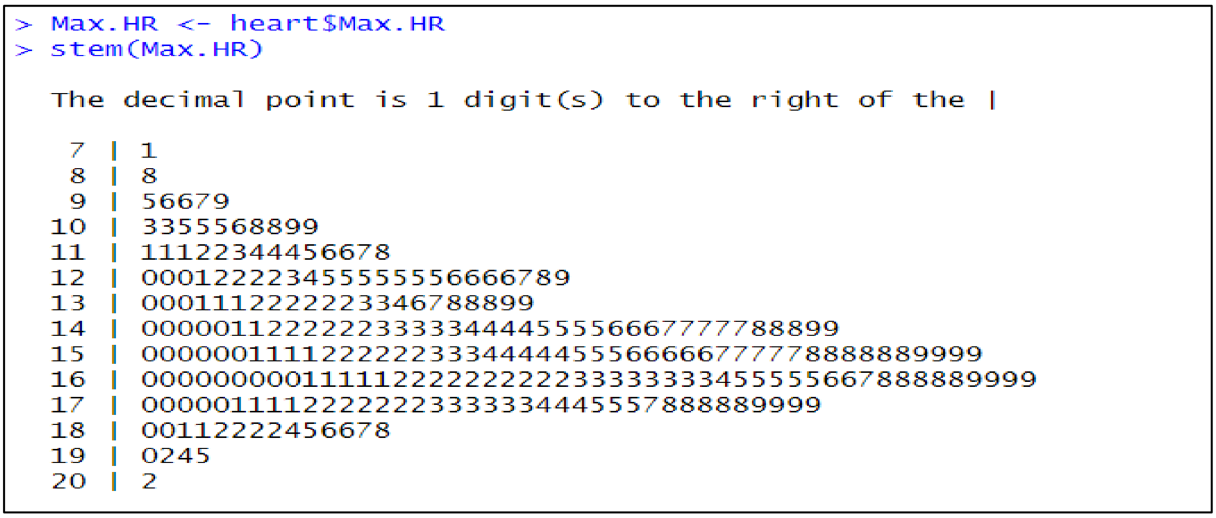


```
> stem(Cholesterol)

  The decimal point is 1 digit(s) to the right of the |

  12 | 6
  14 | 199
  16 | 0467824577778
  18 | 0234568823566777788999
  20 | 0111334444566778899011112222334456788999
  22 | 01223345666678899900011123333444444556679999
  24 | 00023333444555666788999000234444445566678889
  26 | 001123334556677889999900113344455677
  28 | 12222333466888990344558899
  30 | 02233344556788999135589
  32 | 125567005
  34 | 0134
  36 | 0
  38 | 4
  40 | 797
  42 |
  44 |
  46 |
  48 |
  50 |
  52 |
  54 |
  56 | 4
```

Figure 2.6: Stem and Leaf plots for Cholesterol Distribution



```
> Max.HR <- heart$Max.HR
> stem(Max.HR)

  The decimal point is 1 digit(s) to the right of the |

   7 | 1
   8 | 8
   9 | 56679
  10 | 3355568899
  11 | 11122344456678
  12 | 000122223455555556666789
  13 | 0001112222223346788899
  14 | 00000112222223333344444555556667777788899
  15 | 000000111122222233344444555666667777778888889999
  16 | 00000000011111222222222223333333345555566788889999
  17 | 00000111122222223333334445557888889999
  18 | 00112222456678
  19 | 0245
  20 | 2
```

Figure 2.7: Stem and Leaf plots for Max.HR Distribution

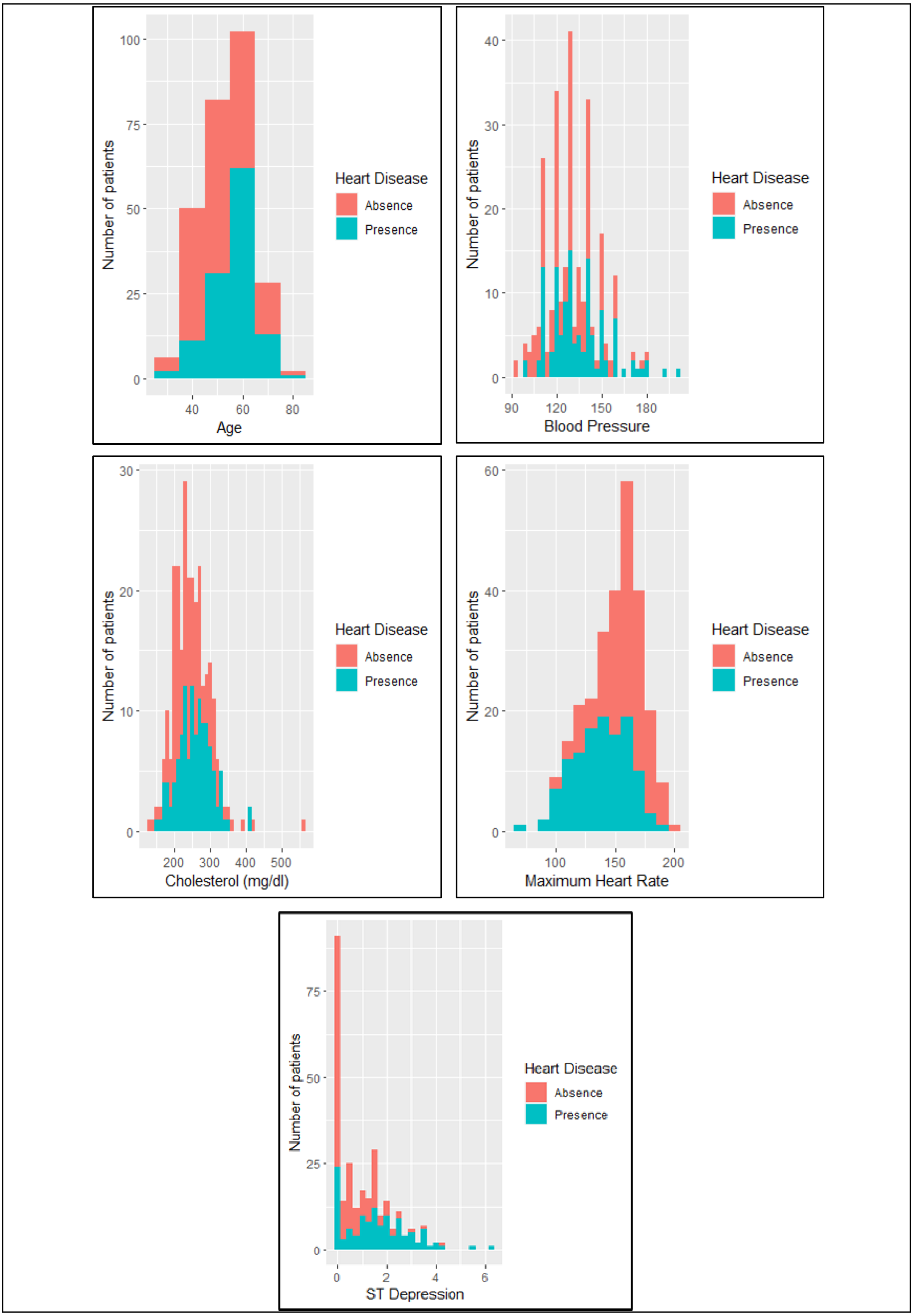


Figure 2.8: Descriptive Analysis for the Relationship between Five Variables with Heart Disease Variable using Histogram.

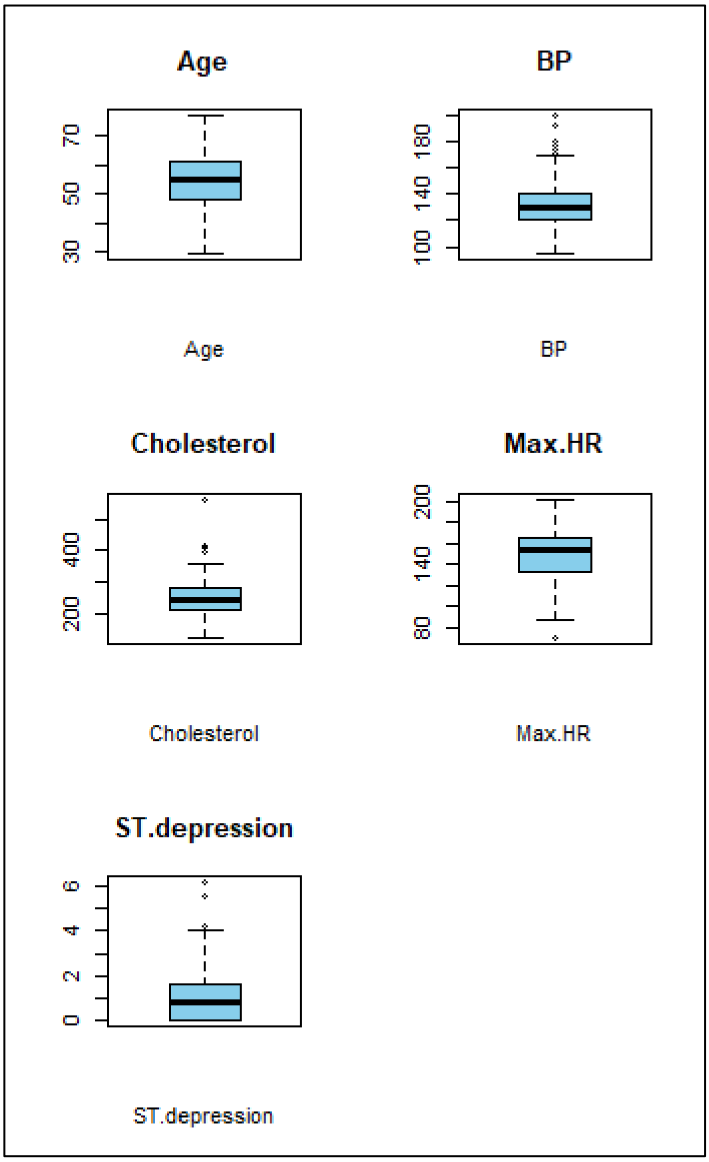
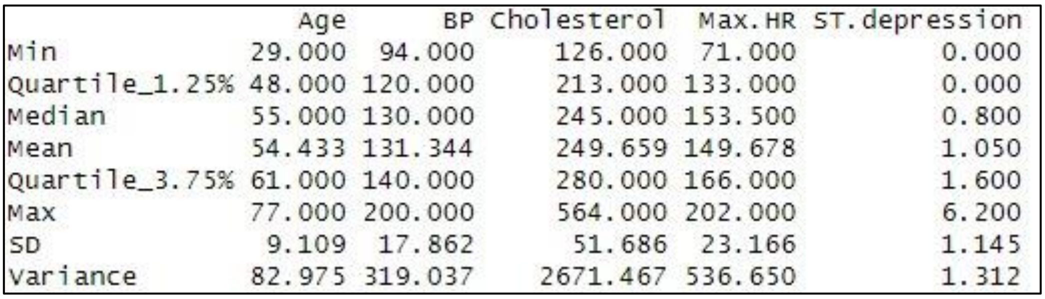


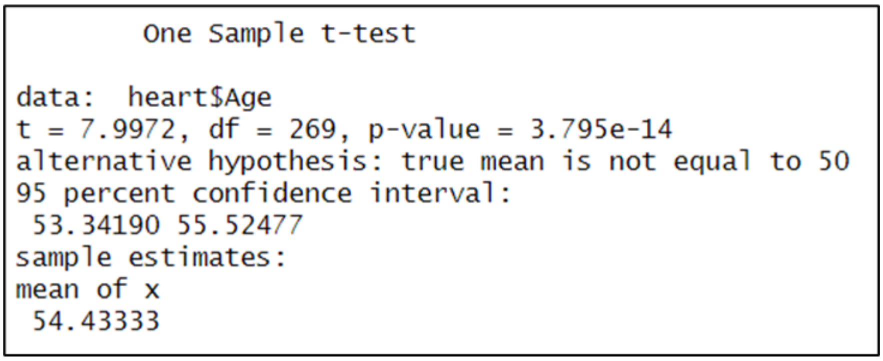
Figure 2.9: Box Plot for Continuous Variable



```
                   Age      BP Cholesterol  Max.HR ST.depression
Min             29.000  94.000     126.000  71.000         0.000
Quartile_1.25% 48.000 120.000     213.000 133.000         0.000
Median          55.000 130.000     245.000 153.500         0.800
Mean            54.433 131.344     249.659 149.678         1.050
Quartile_3.75% 61.000 140.000     280.000 166.000         1.600
Max             77.000 200.000     564.000 202.000         6.200
SD               9.109  17.862      51.686  23.166         1.145
Variance        82.975 319.037    2671.467 536.650         1.312
```

Figure 2.10: The Summary Statistics for Continuous Variables

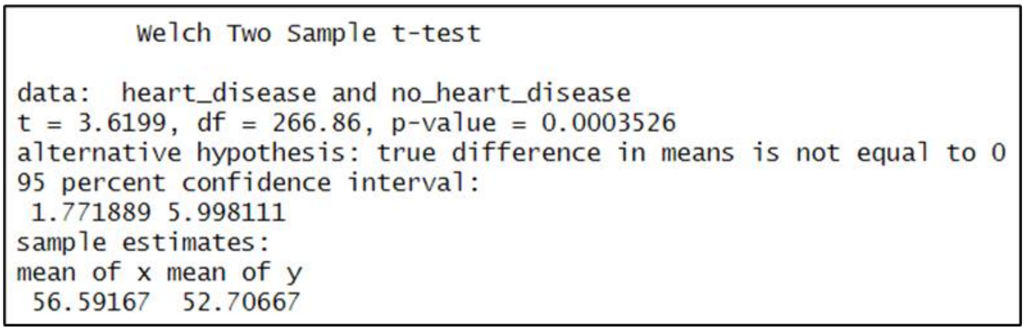
FIGURE 3.1 – 3.6



```
        One Sample t-test

data:  heart$Age
t = 7.9972, df = 269, p-value = 3.795e-14
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 53.34190 55.52477
sample estimates:
mean of x 
 54.43333 
```

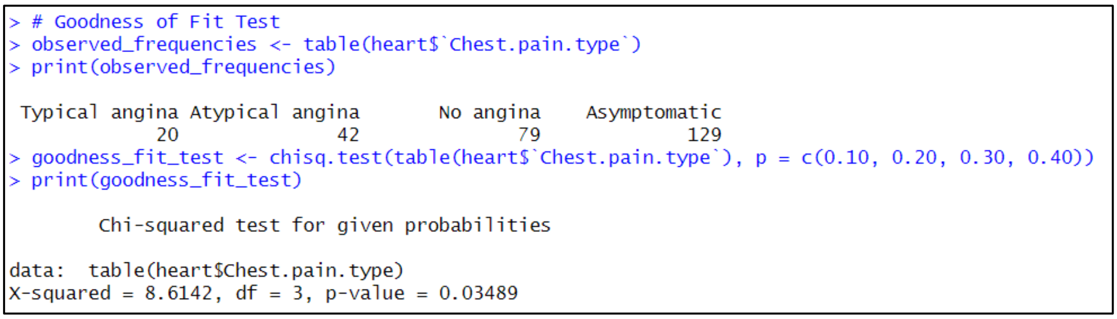
Figure 3.1: One Sample t- Test



```
        Welch Two Sample t-test

data:  heart_disease and no_heart_disease
t = 3.6199, df = 266.86, p-value = 0.0003526
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.771889 5.998111
sample estimates:
mean of x mean of y 
 56.59167  52.70667 
```

Figure 3.2: Two Sample t- Test



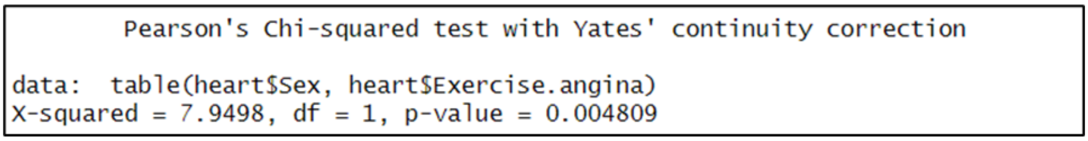
```
> # Goodness of Fit Test
> observed_frequencies <- table(heart$`Chest.pain.type`)
> print(observed_frequencies)

 Typical angina Atypical angina       No angina    Asymptomatic 
             20              42              79             129 
> goodness_fit_test <- chisq.test(table(heart$`Chest.pain.type`), p = c(0.10, 0.20, 0.30, 0.40))
> print(goodness_fit_test)

        Chi-squared test for given probabilities

data:  table(heart$Chest.pain.type)
X-squared = 8.6142, df = 3, p-value = 0.03489
```

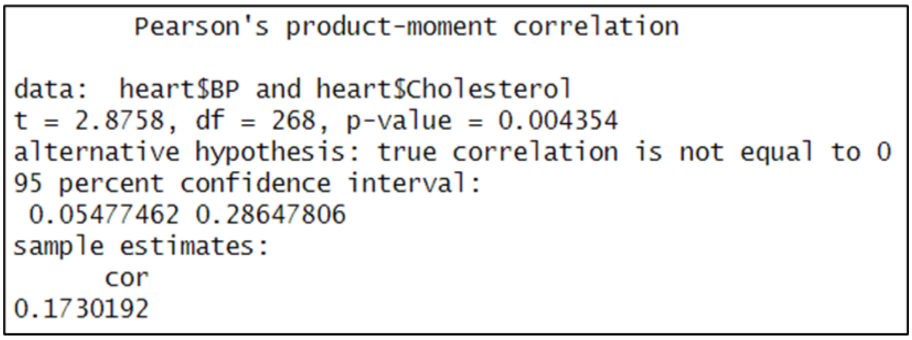
Figure 3.3: Goodness of Fit Test



```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table(heart$Sex, heart$Exercise.angina)
X-squared = 7.9498, df = 1, p-value = 0.004809
```

Figure 3.4: Chi Square test of independence



```
         Pearson's product-moment correlation

data:  heart$BP and heart$Cholesterol
t = 2.8758, df = 268, p-value = 0.004354
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.05477462 0.28647806
sample estimates:
      cor
0.1730192
```

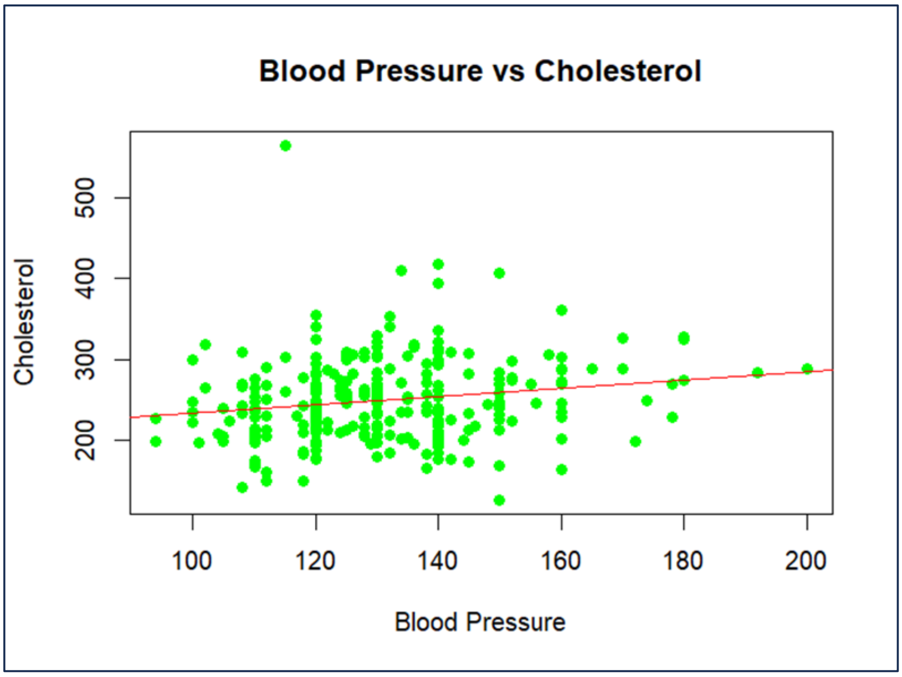
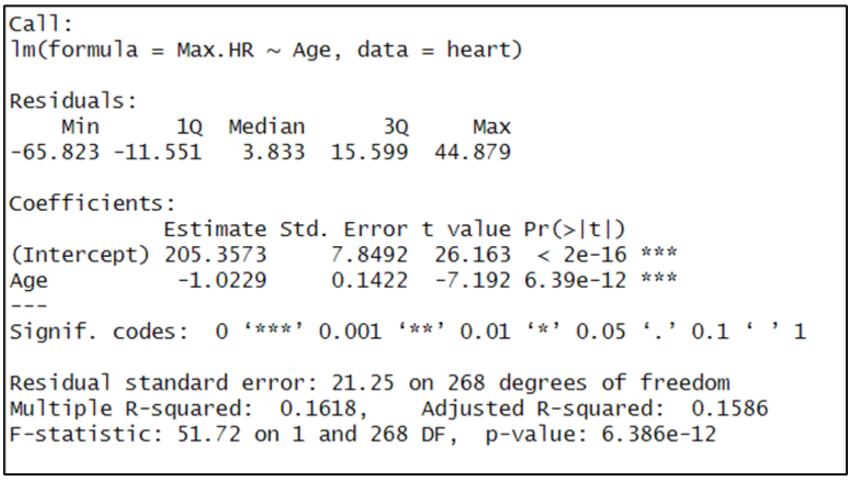


Figure 3.5: Correlation between Blood Pressure and Cholesterol



```
Call:
lm(formula = Max.HR ~ Age, data = heart)

Residuals:
    Min      1Q  Median      3Q     Max 
-65.823 -11.551   3.833  15.599  44.879 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 205.3573     7.8492  26.163  < 2e-16 ***
Age          -1.0229     0.1422  -7.192 6.39e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.25 on 268 degrees of freedom
Multiple R-squared:  0.1618,	Adjusted R-squared:  0.1586 
F-statistic: 51.72 on 1 and 268 DF,  p-value: 6.386e-12
```

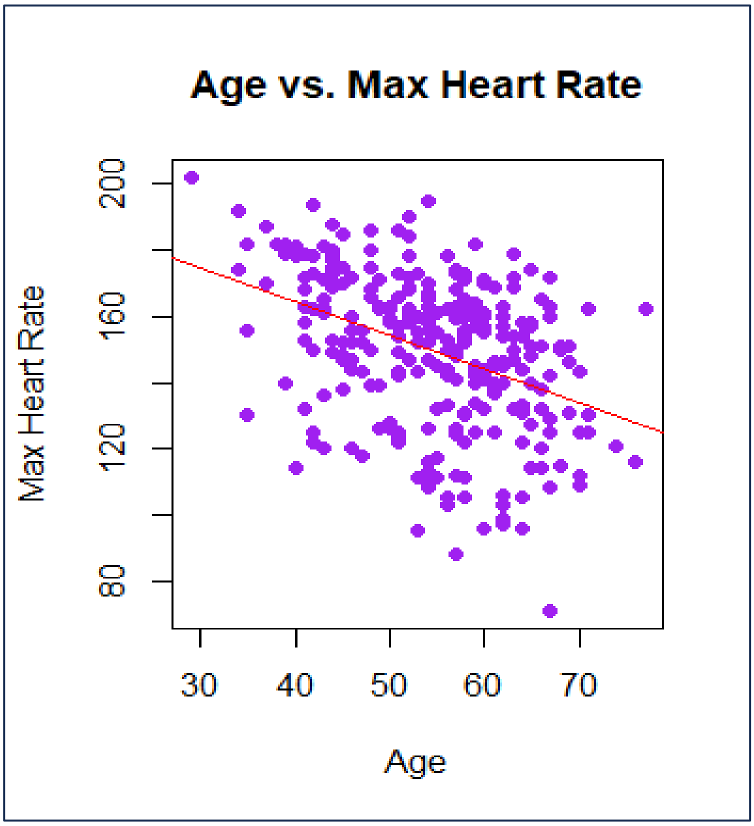


Figure 3.6: Linear Regression between Age and Max Heart Rate

- **R CODES**

```
rm(list = ls())


#Install packages
install.packages("tidyverse")


#Load libraries
library(ggplot2)
library(dplyr)


#read csv


heart <- read.csv("https://raw.githubusercontent.com/mohd-nor/R-PROJECT-MCSD-1113/main/Heart_Disease_Prediction.csv")
View(heart)


#check first rows
head(heart)


# Number of column
ncol(heart)


# Number of row
nrow(heart)


# DATA CLEANING


# Check null value
sum(is.na(heart))
heart
```

```
# Check data type for each column
# Display the structure of the dataset
str(heart)

# Transform categorical variable to R factors
heart$Sex <- as.factor(heart$Sex)
heart$Chest.pain.type <- as.factor(heart$Chest.pain.type)
heart$FBS.over.120<- as.factor(heart$FBS.over.120)
heart$Exercise.angina<- as.factor(heart$Exercise.angina)
heart$Thallium<- as.factor(heart$Thallium)

# Give a better name to the factor values for the graphs
levels(heart$Sex) <- c("Female", "Male")
levels(heart$Chest.pain.type) <- c("Typical angina", "Atypical angina", "No angina",
"Asymptomatic")
levels(heart$FBS.over.120) <- c("False", "True")
levels(heart$Exercise.angina) <- c("No", "Yes")
levels(heart$Thallium) <- c("Normal", "Fixed Defect", "Reversable Defect")

heart

# Check for missing values
missing_values<- colSums(is.na(heart))
missing_values

#check unique values for each colum
checkUniqueValues <- function(data) {
  for (col in names(data)) {
    unique_values <- unique(data[[col]])
    print(paste("Unique values in column", col, ":"))
    print(unique_values)
    print("----------------------------------")
  }
}
checkUniqueValues(heart)

#VISUALISATION

#Barchart
ggplot(heart, aes(Heart.Disease, fill=Heart.Disease)) +
  geom_bar() +
  labs(title = "Heart Disease Presence", x="Heart Disease", y="Number of patients") +
  guides(fill=FALSE)

ggplot(heart, aes(Sex, fill=Heart.Disease)) +
  geom_bar() +
  labs(fill="Heart Disease", x="Sex", y="Number of patients")

ggplot(heart, aes(Chest.pain.type, fill=Heart.Disease)) +
  geom_bar() +
  labs(fill="Heart Disease", x="Chest pain type", y="Number of patients")

ggplot(heart, aes(FBS.over.120, fill=Heart.Disease)) +
  geom_bar() +
  labs(fill="Heart Disease", x="Sugar level > 120 mg/dl", y="Number of patients")

ggplot(heart, aes(EKG.results, fill=Heart.Disease)) +
  geom_bar() +
  labs(fill="Heart Disease", x="Electrocardiogram on rest", y="Number of patients")

ggplot(heart, aes(Exercise.angina, fill=Heart.Disease)) +
  geom_bar() +
  labs(fill="Heart Disease", x="Presence of angina during exercise", y="Number of
patients")
```

```
ggplot(heart, aes(Slope.of.ST, fill=Heart.Disease)) +
  geom_bar() +
  labs(fill="Heart Disease", x="Slope of the ST segment", y="Number of patients")


ggplot(heart, aes(Thallium, fill=Heart.Disease)) +
  geom_bar() +
  labs(fill="Heart Disease", x="Results of the blood flow", y="Number of patients")


ggplot(heart, aes(Number.of.vessels.fluro, fill=Heart.Disease)) +
  geom_bar() +
  labs(fill="Heart Disease", x="Number of main blood vessels coloured", y="Number of
patients")


# Piechart
create_pie_chart <- function(table_data, main_title) {
  pie_percent <- prop.table(table_data) * 100  # Calculate percentages
    # Create the pie chart with formatted percentage labels and matching legend colors
  pie(table_data, labels = sprintf("%.1f%%", pie_percent), main = main_title, col =
rainbow(length(table_data)))
    # Add legend to the right of the chart
  x_legend <- 1  # Adjust as needed
  y_legend <- 1  # Adjust as needed
    # Add legend to the specified position
  legend(x = x_legend, y = y_legend, legend = names(table_data), fill =
rainbow(length(table_data)), cex = 0.5)
}
# Create pie charts with formatted percentage labels and matching legend colors
create_pie_chart(table(heart$Chest.pain.type), "Chest Pain Distribution")
create_pie_chart(table(heart$Number.of.vessels.fluro), "Number of Vessel
Fluro Distribution")


#Stem and Leaf
```

```
BP <- heart$BP
stem(BP)

Cholesterol <- heart$Cholesterol
stem(Cholesterol)

Max.HR <- heart$Max.HR
stem(Max.HR)

#Histogram
ggplot(heart, aes(Age, fill=Heart.Disease)) +
  geom_histogram(binwidth=10) +
  labs(fill="Heart Disease", x="Age", y="Number of patients")

ggplot(heart, aes(BP, fill=Heart.Disease)) +
  geom_histogram(binwidth=3) +
  labs(fill="Heart Disease", x="Blood Pressure", y="Number of patients")

ggplot(heart, aes(Cholesterol, fill=Heart.Disease)) +
  geom_histogram(binwidth=10) +
  labs(fill="Heart Disease", x="Cholesterol (mg/dl)", y="Number of patients")

ggplot(heart, aes(Max.HR, fill=Heart.Disease)) +
  geom_histogram(binwidth=10) +
  labs(fill="Heart Disease", x="Maximum Heart Rate", y="Number of patients")
ggplot(heart, aes(ST.depression, fill=Heart.Disease)) +
  geom_histogram(binwidth=0.25) +
  labs(fill="Heart Disease", x="ST Depression", y="Number of patients")

#Box plot
par(mfrow = c(3, 2))
```

```
selected_variables <- c("Age", "BP", "Cholesterol", "Max.HR", "ST.depression")
for (variable in selected_variables) {
  boxplot(heart[[variable]], col = "skyblue", xlab = variable, main = variable)
}

# Summary statistics of the dataset
summary(heart)

# Selecting only numeric variables
cont_data <- select(heart, Age, BP, Cholesterol, Max.HR, ST.depression)
cont_data <- lapply(cont_data, as.numeric)
# Creating a custom function to calculate the desired summary statistics
custom_summary <- function(x) {
  round(
    c(Min = min(x),
      Quartile_1 = quantile(x, 0.25),
      Median = median(x),
      Mean = mean(x),
      Quartile_3 = quantile(x, 0.75),
      Max = max(x),
      SD = sd(x),
      Variance = var(x)),
    digits = 3)
}
summary_table <- sapply(cont_data, custom_summary)
print(summary_table)

# Hypothesis Testing - 1-Sample T-Test
t_test_1 <- t.test(heart$Age, mu = 54)
print(t_test_1)
```

```
# Hypothesis Testing - 2-Sample T-Test
# Separate data into two groups based on heart disease status
heart_disease <- heart$Age[heart$Heart.Disease == 'Presence']
no_heart_disease <- heart$Age[heart$Heart.Disease == 'Absence']

# Perform 2-sample t-test
t_test_2 <- t.test(heart_disease, no_heart_disease)
print(t_test_2)

# Goodness of Fit Test
observed_frequencies <- table(heart$`Chest.pain.type`)
print(observed_frequencies)

goodness_fit_test <- chisq.test(table(heart$`Chest.pain.type`), p = c(0.10, 0.20, 0.30, 0.40))
print(goodness_fit_test)

# Chi-Square Test of Independence
chisq.test(table(heart$Sex, heart$`Exercise.angina`))

# Correlation
cor_test <- cor.test(heart$BP, heart$Cholesterol)

# Scatter Plot
plot(heart$BP, heart$Cholesterol, main="Blood Pressure vs Cholesterol", xlab="Blood Pressure", ylab="Cholesterol", pch=19, col='green')
model_2 <- lm(heart$Cholesterol ~ heart$BP, data=heart)
abline(model_2, col="red")

# Print Correlation Test Result
print(cor_test)
```

```
# Simple Linear Regression
# Fit a linear regression model
lm_model <- lm(Max.HR ~ Age, data = heart)

# Plotting scatter plot
plot(heart$Age, heart$Max.HR, main = "Age vs. Max Heart Rate",
    xlab = "Age", ylab = "Max Heart Rate", pch = 16, col = "purple")

# Adding regression line to the plot
abline(lm_model, col = "red")

# Display the regression equation and R-squared value
summary(lm_model)

# ANOVA
# Perform ANOVA
anova_result <- aov(Max.HR ~ Chest.pain.type, data = heart)

# Summary of ANOVA
summary(anova_result)
```