BML MUNJAL UNIVERSITY™

Department of Computer Science and Engineering

School of Engineering and Technology

May 2023

Mathematics for Engineers - II

Project Report

(MTH1711)

Heart Disease Analysis

Group Members:

Vedansh Kumawat
220489

Anusha Singh
220641

Shivanshi Goel
220457

ACKNOWLEDGEMENT

First and foremost, we are very much thankful to our lecturer Dr. Ranjib Banerjee for his guidance, encouragement, help and useful suggestions throughout the completion of the project. His untiring and painstaking efforts, individual help made it possible for our group to complete this work on time. We are glad that we got an opportunity to know more things about the concepts. We would like to appreciate ourselves for the cooperation, team spirit we had. We would also like to appreciate the technology which made our work simpler and easier. We would like to appreciate the web, resources, and our textbook where we got sufficient information to complete our project.

CONTENTS

ABSTRACT

Heart disease kills millions of people worldwide and is the leading cause of death for both men and women. According to WHO, cardiovascular disease is the leading cause of death worldwide, with an estimated 17.9 million people dying from it, with heart attacks and stroke accounting for 85% of deaths. Numerous health factors, such as inherited blood disorders, resting heart rate, cholesterol, and others, all contribute to the prediction of a heart.

Heart disease kills millions of people worldwide and is the leading cause of death for both men and women. According to WHO, cardiovascular disease is the leading cause of death worldwide, with an estimated 17.9 million people dying from it, with heart attacks and stroke accounting for 85% of deaths. Numerous health factors, such as inherited blood disorders, resting heart rate, cholesterol, and others, all contribute to the prediction of a heart attack.

INTRODUCTION

The heart is one of our bodies' most important organs. Heart disease, also known as cardiac disease, is a form of heart dysfunction.

A variety of factors can increase one's risk of developing heart disease. While some of these factors are beyond our control, many can be avoided by living a healthy lifestyle. Uncontrollable factors include gender, age, family history, and heart shape. High blood pressure, high cholesterol, obesity, smoking, and diabetes are all preventable risk factors.

In the United States, heart disease is the leading cause of death. Every 36 seconds, someone in the United States dies from cardiovascular disease. Heart disease kills approximately 655,000 Americans each year, accounting for one out of every four deaths. In this research. In this analysis, we will use a heart disease dataset to investigate the most important factors that contribute to heart disease.

PROBLEM STATEMENT

- □ Do men or women have a higher risk of having a heart attack? When is the most common age for a heart attack? The comparison of gender and age in this study will explain the likelihood of a heart attack.
- □ In this study, the male and female ages, gender, and target variable are used. Following the presentation of the summary statistics for male and female, a histogram and bar plot are plotted to present the data in a more comprehensive manner.
- □ A Chi square test of association between gender and risk of heart attack is also used to determine whether there is an association between gender and risk of heart attack. Following the testing of assumptions, two sample t-tests are used to determine whether there is a difference in the mean age of male and female patients with a high risk of heart attack.
- □ This investigation has established and tested the following hypotheses:
- There is a link between gender and an individual's risk of having a heart attack.
- The age of males and females who are at high risk of having a heart attack.

DATA

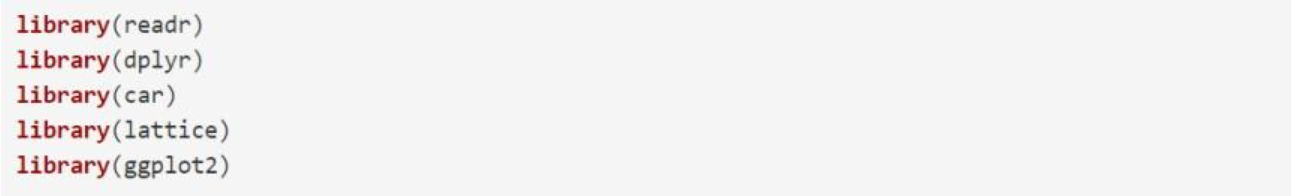
- The dataset used contains 14 variables and 303 observations of various individuals from the United States of America, Cleveland. The dataset used in this task was generated from:

  https://www.kaggle.com/madhav000/attack-prediction-accuracy-morethan-80.
- The observation of the data was drawn from:

  https://archive.ics.uci.edu/ml/datasets/Heart+Disease. The dataset contains 76 variables from different countries such as Hungary and Switzerland but have been subset for easier analysis and the Cleveland database is the one that has been selected for this task.
- Age: Age of individual in years.
- Sex: Gender of the individual. 0 for Female and 1 for Male. Sex variable has been renamed to Gender.
- cp: Chest pain type from 1 to 4. Does not have any levels.
- trestbps: Resting blood pressure. (in mm Hg on admission to the hospital)
- Chol: Serum cholesterol measured in mg/dl.
- fbs: Fasting blood > 120 mg/dl. 0 for False, 1 for True.

- restecg: Resting electrocardiographic results. 0 for normal, 1 for having abnormality, and 2 for probable or definite hypertrophy.
- thalach: Maximum heart rate achieved measured in beats per minutes.
- exang: Exercise induced angina. 0 for False, 1 for True.
- oldpeak: ST depression induced by exercise relative to rest. Does not have any levels.
- slope: Slope of peak exercise ST segment. 1 for upsloping, 2 for flat, 3 for downsloping.
- thal: Thalassemia a inherited blood disorder, 0 for normal, 1 for fixed defect and 2 for reversable defect.
- target: Likelihood of heart attack. 0 for less chance of heart attack and 1 for more chance of heart attack.

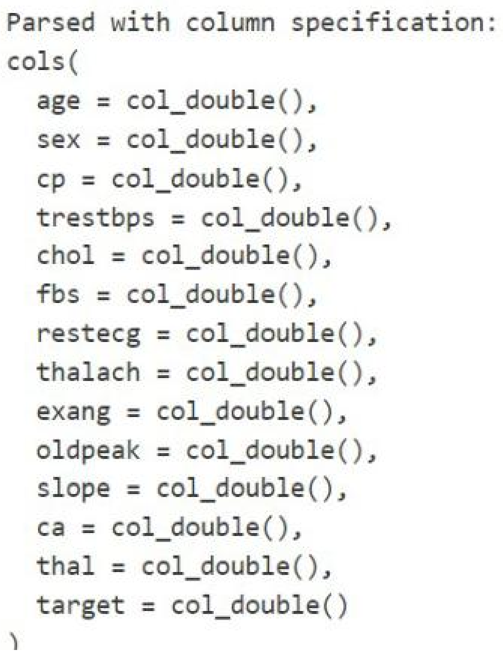
PREPROCESSING OF DATA



```
library(readr)
library(dplyr)
library(car)
library(lattice)
library(ggplot2)
```

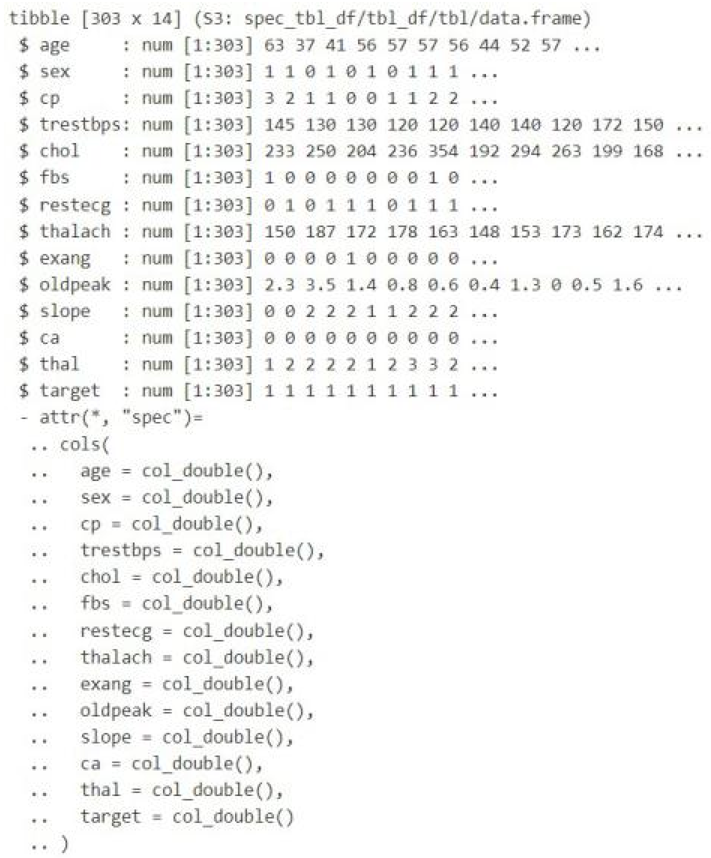
• Read_csv is used to import the file, and factor is used to order the sex and target variables. Furthermore, the sex variable has been renamed gender.
• The summary statistics of the age of the individuals in this study, broken down by male and female, are shown below. The summarise function generates a new data frame with the minimum, median, maximum, and mean age, first and third quartile, standard deviation, number of observations, and missing observations.

```
heart <- read_csv("heart.csv")
```



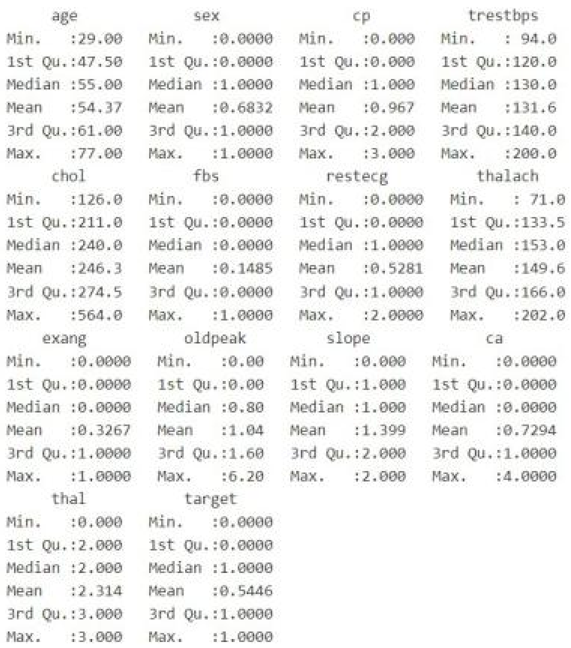
```
Parsed with column specification:
cols(
  age = col_double(),
  sex = col_double(),
  cp = col_double(),
  trestbps = col_double(),
  chol = col_double(),
  fbs = col_double(),
  restecg = col_double(),
  thalach = col_double(),
  exang = col_double(),
  oldpeak = col_double(),
  slope = col_double(),
  ca = col_double(),
  thal = col_double(),
  target = col_double()
)
```

```
str(heart)
```



```
tibble [303 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ age     : num [1:303] 63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : num [1:303] 1 1 0 1 0 1 0 1 1 1 ...
 $ cp      : num [1:303] 3 2 1 1 0 0 1 1 2 2 ...
 $ trestbps: num [1:303] 145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : num [1:303] 233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : num [1:303] 1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : num [1:303] 0 1 0 1 1 1 0 1 1 1 ...
 $ thalach : num [1:303] 150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : num [1:303] 0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num [1:303] 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : num [1:303] 0 0 2 2 2 1 1 2 2 2 ...
 $ ca      : num [1:303] 0 0 0 0 0 0 0 0 0 0 ...
 $ thal    : num [1:303] 1 2 2 2 2 1 2 3 3 2 ...
 $ target  : num [1:303] 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "spec")=
  .. cols(
  ..   age = col_double(),
  ..   sex = col_double(),
  ..   cp = col_double(),
  ..   trestbps = col_double(),
  ..   chol = col_double(),
  ..   fbs = col_double(),
  ..   restecg = col_double(),
  ..   thalach = col_double(),
  ..   exang = col_double(),
  ..   oldpeak = col_double(),
  ..   slope = col_double(),
  ..   ca = col_double(),
  ..   thal = col_double(),
  ..   target = col_double()
  .. )
```

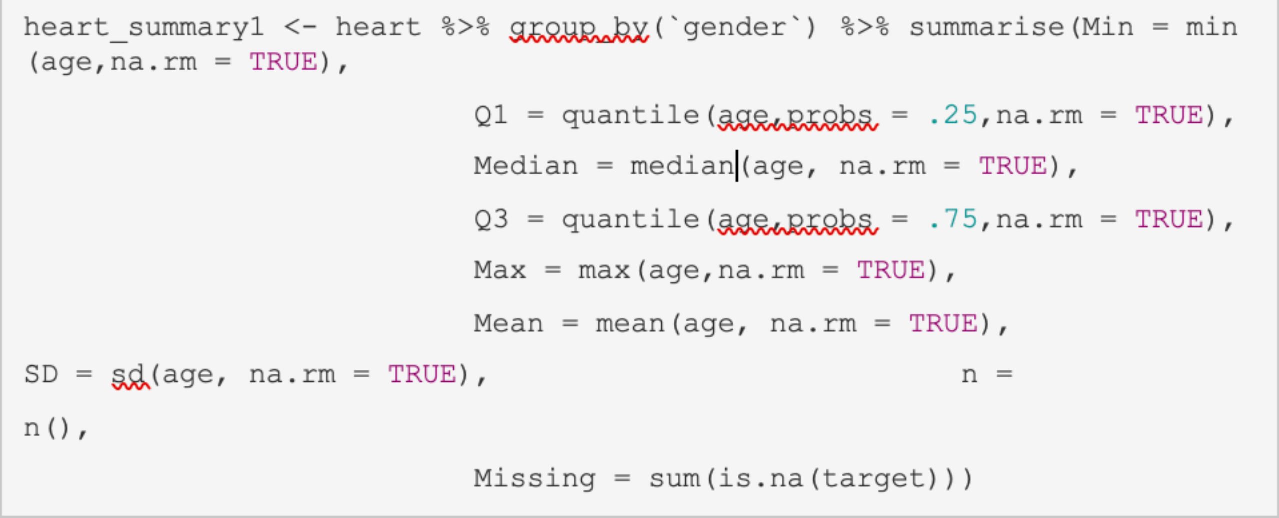
summary(heart)



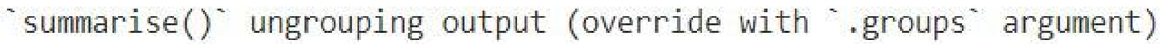
```
      age             sex               cp           trestbps    
 Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
 Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
 Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
 Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
      chol            fbs             restecg          thalach     
 Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
 1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5  
 Median :240.0   Median :0.0000   Median :1.0000   Median :153.0  
 Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6  
 3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
 Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
     exang            oldpeak         slope             ca        
 Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
 Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
 3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
      thal           target      
 Min.   :0.000   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:0.0000  
 Median :2.000   Median :1.0000  
 Mean   :2.314   Mean   :0.5446  
 3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :3.000   Max.   :1.0000  
```

```
heart$sex <- heart$sex %>% factor(levels=c(0,1),
                                   labels=c("Female","Male"))
heart$target <- heart$target %>% factor(levels=c(0,1),
labels=c(0,1))


heart <- heart %>% rename(gender = sex)


str(heart)
```

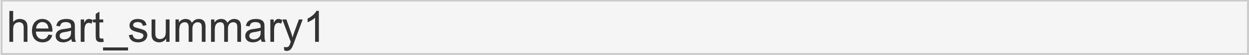
```
tibble [303 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ age     : num [1:303] 63 37 41 56 57 57 56 44 52 57 ...
 $ gender  : Factor w/ 2 levels "Female","Male": 2 2 1 2 1 2 1 2 2 2 ...
 $ cp      : num [1:303] 3 2 1 1 0 0 1 1 2 2 ...
 $ trestbps: num [1:303] 145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : num [1:303] 233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : num [1:303] 1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : num [1:303] 0 1 0 1 1 1 0 1 1 1 ...
 $ thalach : num [1:303] 150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : num [1:303] 0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num [1:303] 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : num [1:303] 0 0 2 2 2 1 1 2 2 2 ...
 $ ca      : num [1:303] 0 0 0 0 0 0 0 0 0 0 ...
 $ thal    : num [1:303] 1 2 2 2 2 1 2 3 3 2 ...
 $ target  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 - attr(*, "spec")=
  .. cols(
  ..   age = col_double(),
  ..   sex = col_double(),
  ..   cp = col_double(),
  ..   trestbps = col_double(),
  ..   chol = col_double(),
  ..   fbs = col_double(),
  ..   restecg = col_double(),
  ..   thalach = col_double(),
  ..   exang = col_double(),
  ..   oldpeak = col_double(),
  ..   slope = col_double(),
  ..   ca = col_double(),
  ..   thal = col_double(),
  ..   target = col_double()
  .. )
```



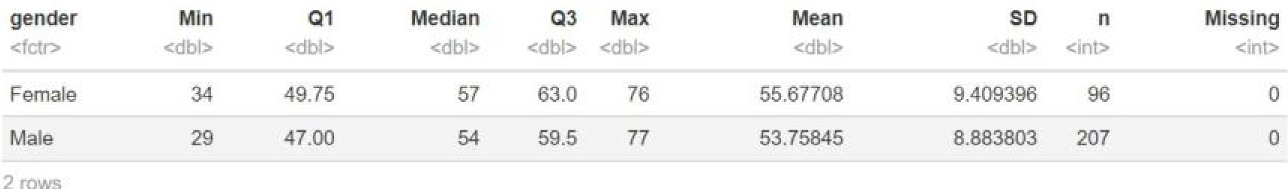
```
heart_summary1 <- heart %>% group_by(`gender`) %>% summarise(Min = min
(age,na.rm = TRUE),
                                Q1 = quantile(age,probs = .25,na.rm = TRUE),
                                Median = median(age, na.rm = TRUE),
                                Q3 = quantile(age,probs = .75,na.rm = TRUE),
                                Max = max(age,na.rm = TRUE),
                                Mean = mean(age, na.rm = TRUE),
SD = sd(age, na.rm = TRUE),                                   n =
n(),
                                Missing = sum(is.na(target)))
```



```
`summarise()` ungrouping output (override with `.groups` argument)
```



```
heart_summary1
```



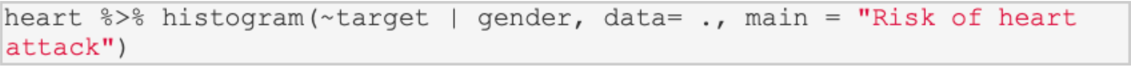
gender <fctr>	Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	n <int>	Missing <int>
Female	34	49.75	57	63.0	76	55.67708	9.409396	96	0
Male	29	47.00	54	59.5	77	53.75845	8.883803	207	0

2 rows

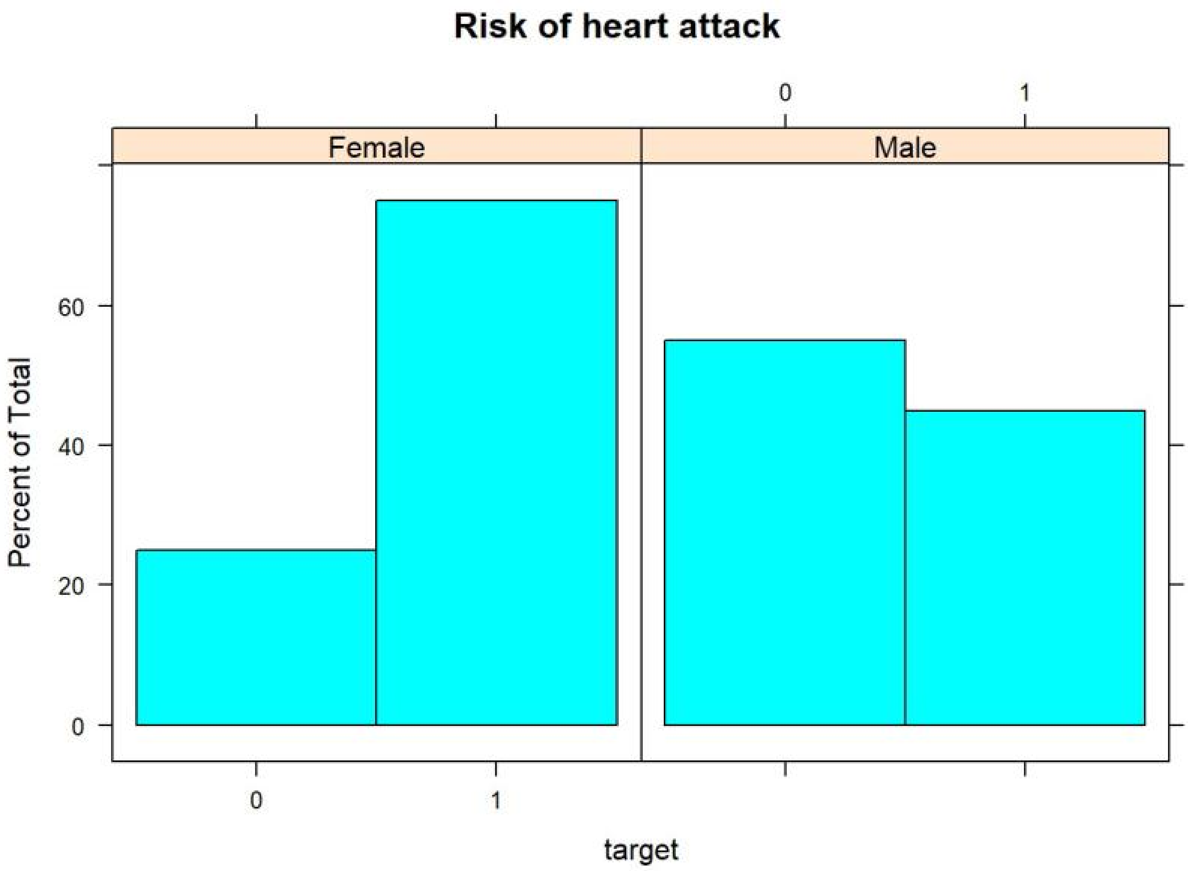
DESCRIPTIVE STATISTICS AND VISUALISATION

• The male and female mean ages are very close, with the male age being 53.76 and the female age being 55.68.

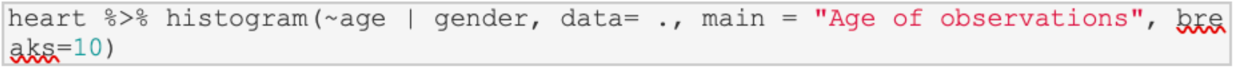
• Female has a higher standard deviation and a higher standard error than male. • Based on the histogram, female appears to have a higher likelihood of a heart attack because the proportion is much larger than male, whose likelihood of a heart attack appears to be fairly close.



```
heart %>% histogram(~target | gender, data= ., main = "Risk of heart
attack")
```

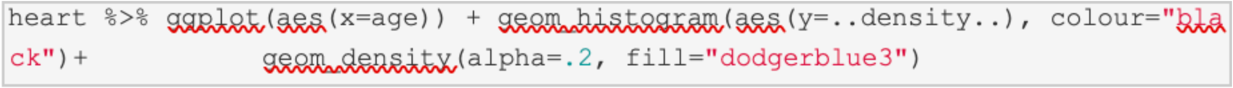


- In terms of the age of individuals, the Female sample appears to have the highest percent of total in the age range of 55 to 60 years followed by 50 to 55 years. While the Male sample appears to have the highest percent of total in the age range of 60 to 65 years.



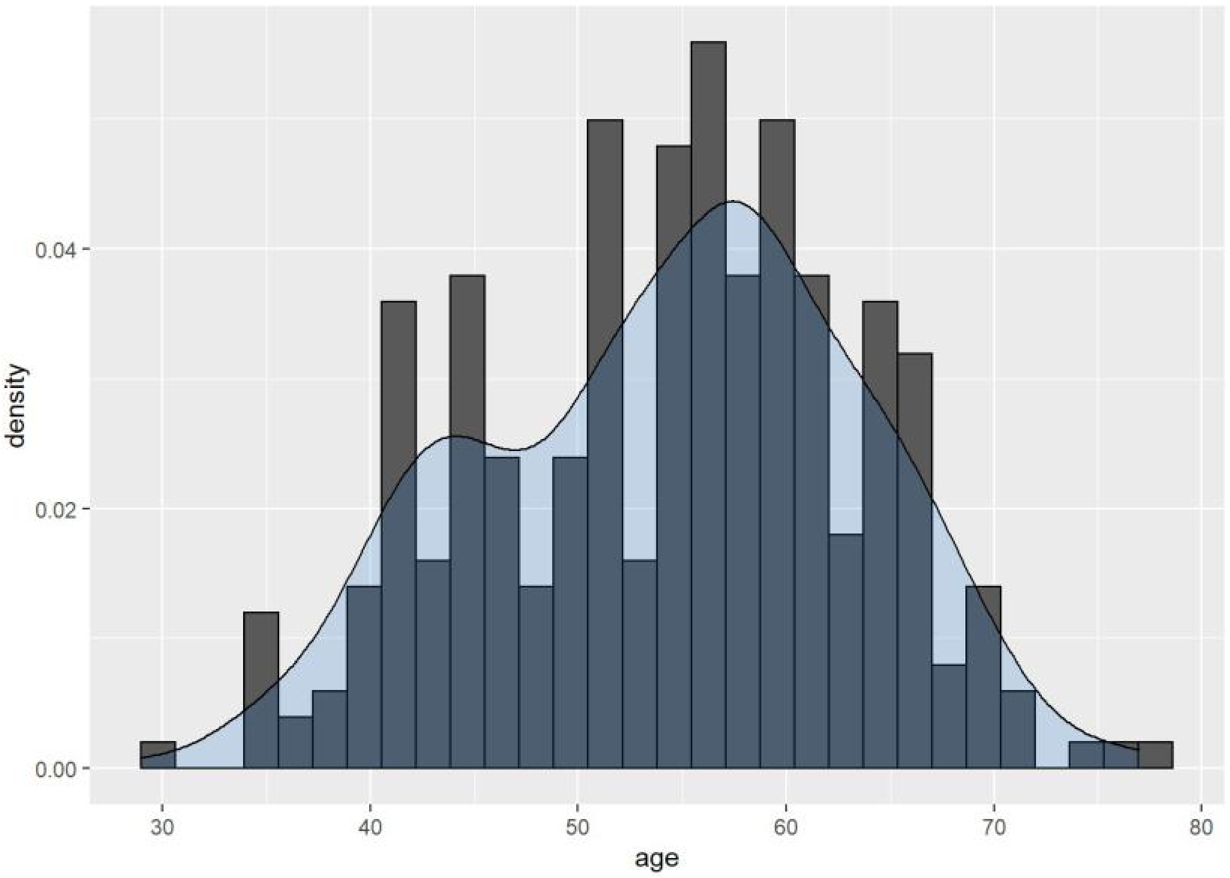
```
heart %>% histogram(~age | gender, data= ., main = "Age of observations", bre
aks=10)
```

- The plot below shows the distribution in the ages of the individual in this investigation. The curve appears to be more negatively skewed.

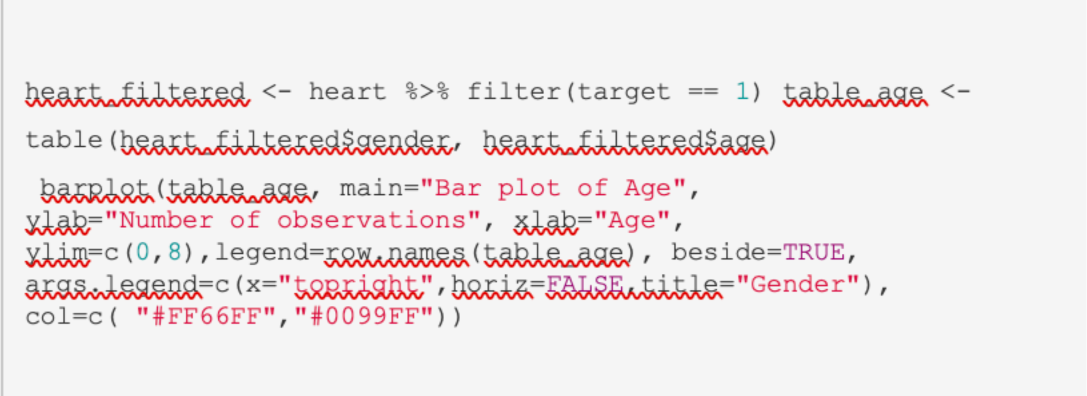


```
heart %>% ggplot(aes(x=age)) + geom_histogram(aes(y=..density..), colour="bla
ck")+            geom_density(alpha=.2, fill="dodgerblue3")
```

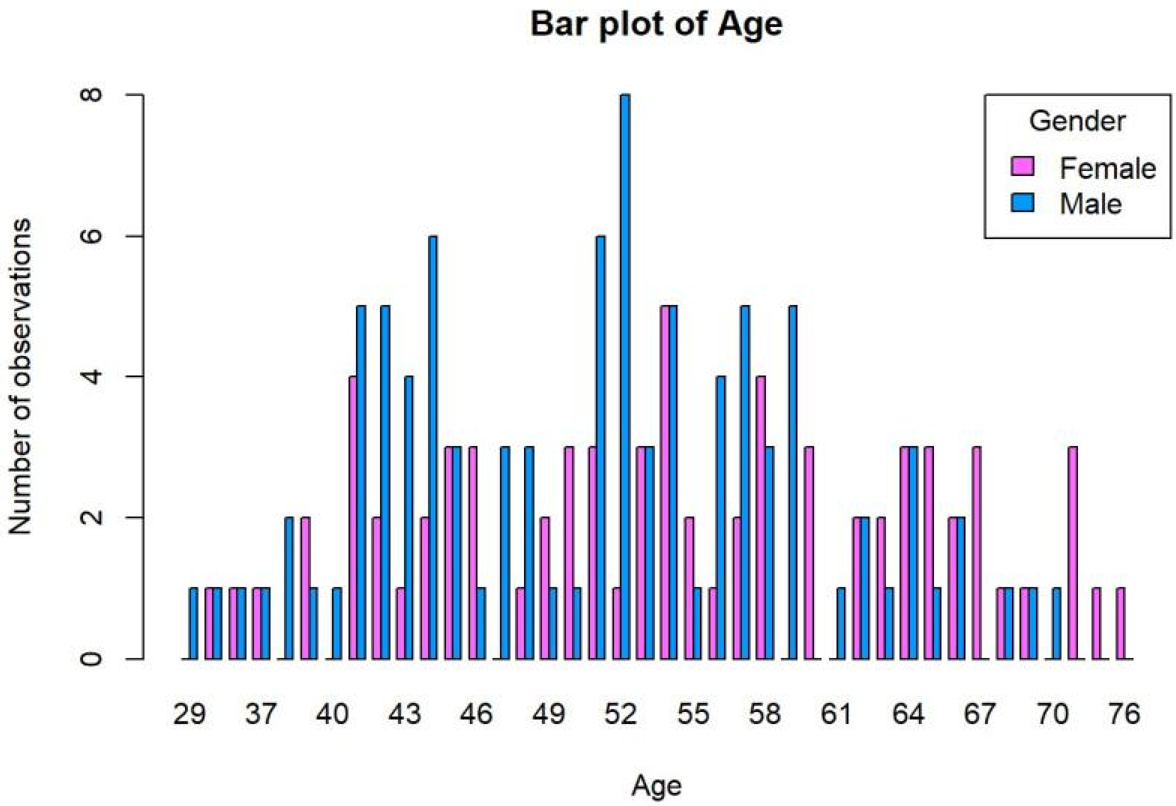
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



• The heart dataset is filtered for people who are at high risk of having a heart attack.The plot below compares the ages of these individuals by displaying the barplot side by side. By observing, we can see that the Male has a greater number of observations of ages than the Female. Observing the plot, however, is inconclusive as to whether Male and Female have the same age of a high likelihood of a heart attack.

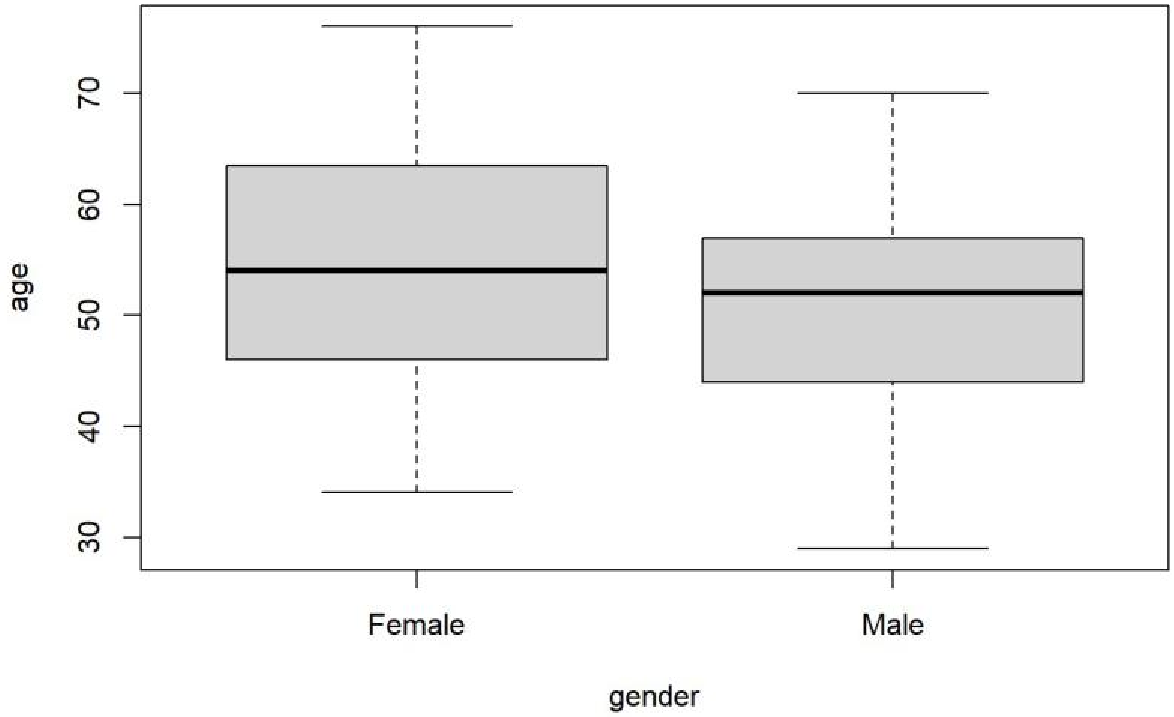


```
heart_filtered <- heart %>% filter(target == 1) table_age <-
table(heart_filtered$gender, heart_filtered$age)
 barplot(table_age, main="Bar plot of Age",
ylab="Number of observations", xlab="Age",
ylim=c(0,8),legend=row.names(table_age) , beside=TRUE,
args.legend=c(x="topright",horiz=FALSE,title="Gender"),
col=c( "#FF66FF","#0099FF"))
```



• The boxplot demonstrates that there are no outliers in the data that must be addressed. The Female has a larger interquartile range than the Male, indicating that there is more variation in age. According to the summary statistics, females have a higher mean age of 54.56 than males, who have a mean age of 50.90. Furthermore, Female has a higher standard deviation of 10.27 than Male, which is 8.68.

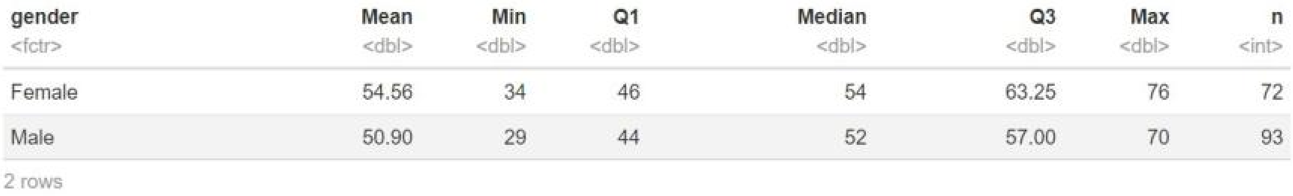
```
boxplot(age ~ gender, data=heart_filtered)
```



```
heart_summary2 <- heart_filtered %>% group_by(gender) %>% summarise(Mean = ro
und(mean(age, na.rm = TRUE),2),
                                                                  Min = min(age,na.rm = TRUE)
,
                                                                  Q1 = quantile(age,probs = .
25,na.rm = TRUE),
                                                                  Median = median(age, na.rm
= TRUE),
                                                                  Q3 = quantile(age,probs = .
75,na.rm = TRUE),
                                                                  Max= max(age,na.rm=TRUE),
n = n())
```

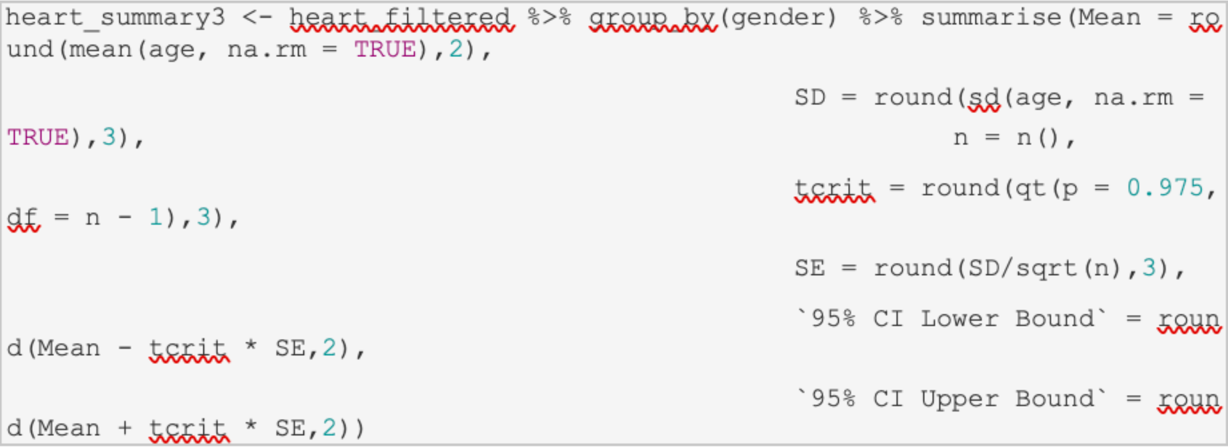
```
`summarise()` ungrouping output (override with `.groups` argument)
```

```
heart_summary2
```

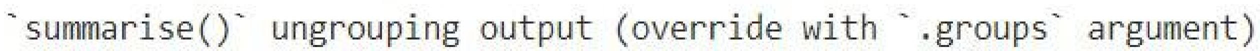


gender <fctr>	Mean <dbl>	Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	n <int>
Female	54.56	34	46	54	63.25	76	72
Male	50.90	29	44	52	57.00	70	93

2 rows

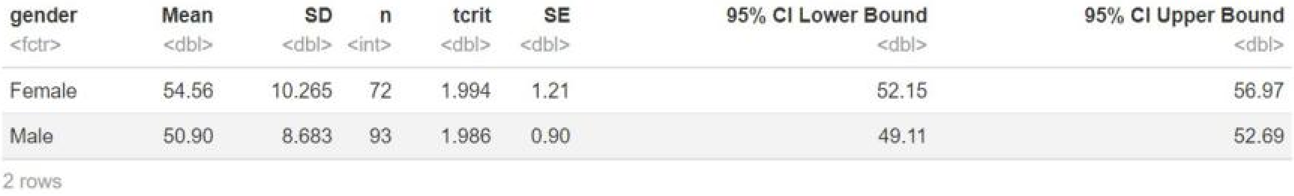


```
heart_summary3 <- heart_filtered %>% group_by(gender) %>% summarise(Mean = ro
und(mean(age, na.rm = TRUE),2),

                                                  SD = round(sd(age, na.rm =
TRUE),3),
                                                            n = n(),

                                                  tcrit = round(qt(p = 0.975,
df = n - 1),3),

                                                  SE = round(SD/sqrt(n),3),

                                                  `95% CI Lower Bound` = roun
d(Mean - tcrit * SE,2),

                                                  `95% CI Upper Bound` = roun
d(Mean + tcrit * SE,2))
```



```
`summarise()` ungrouping output (override with `.groups` argument)
```

heart_summary3



gender <fctr>	Mean <dbl>	SD <dbl>	n <int>	tcrit <dbl>	SE <dbl>	95% CI Lower Bound <dbl>	95% CI Upper Bound <dbl>
Female	54.56	10.265	72	1.994	1.21	52.15	56.97
Male	50.90	8.683	93	1.986	0.90	49.11	52.69

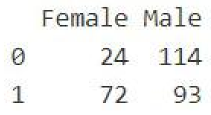
2 rows

HYPOTHESIS TESTING

Chi square test of association

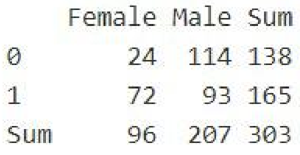
- Specify the null and alternate hypotheses.
- H0: Probability that a heart attack and gender are related.
- H1: It is unlikely that a heart attack and gender are related.
- It is assumed that no more than 25% of expected counts are less than 5 and that all individual counts are 1 or greater.
- Table_heart2 shows that males have a 0.44 chance of having a heart attack, while females have a 0.75 chance of having a heart attack.

```
table_heart <- table(heart$target, heart$gender)

table_heart
```



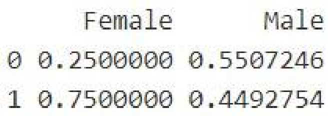
```
   Female Male
 0     24  114
 1     72   93
```

```
table_heart %>% addmargins()
```



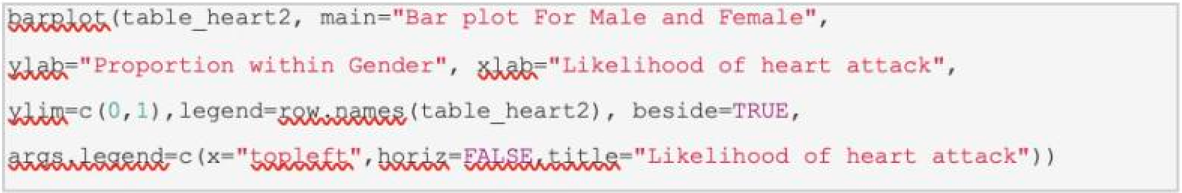
```
      Female Male Sum
 0        24  114 138
 1        72   93 165
 Sum      96  207 303
```

```
table_heart2 <- table_heart %>% prop.table(margin=2)

table_heart2
```

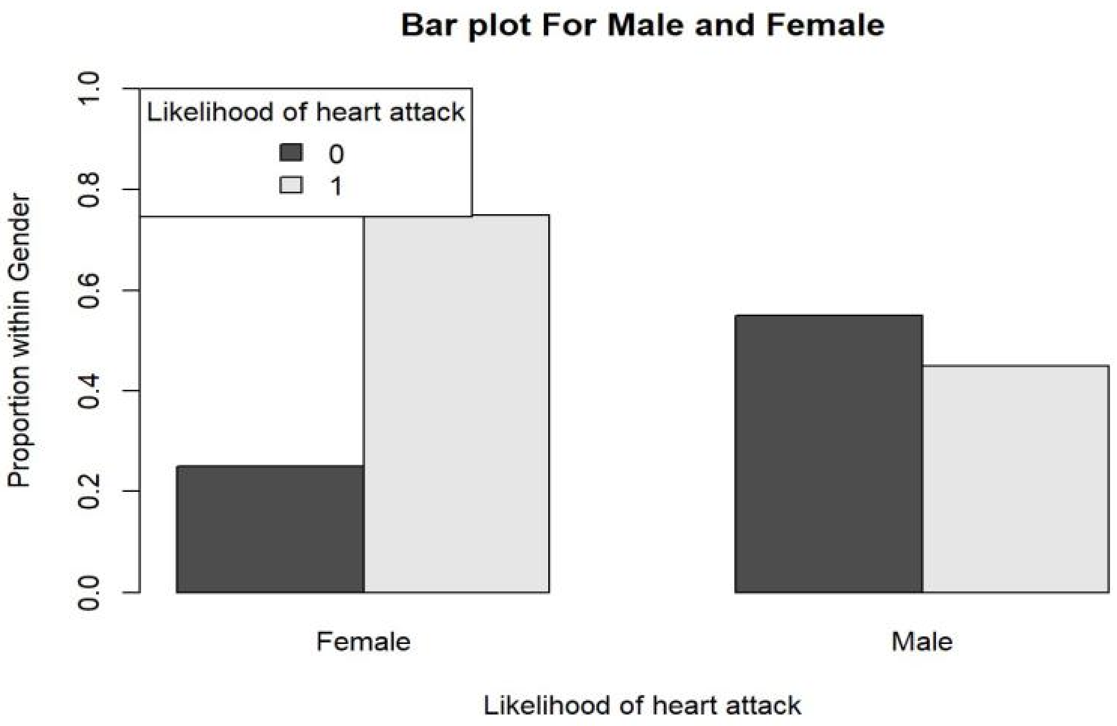


```
     Female      Male
0 0.2500000 0.5507246
1 0.7500000 0.4492754
```

• The chi-square test of association for gender and risk of heart attack yields a p-value of 1.877e-06, which is less than 0.001. As a result, H0 is rejected as the null hypothesis, and the chi-square test of association is statistically significant. The findings indicate that there is no evidence of a link between a person's gender and their risk of having a heart attack. As a result, the likelihood of having a heart attack is determined by the individual's gender.



```
barplot(table_heart2, main="Bar plot For Male and Female",
ylab="Proportion within Gender", xlab="Likelihood of heart attack",
ylim=c(0,1),legend=row.names(table_heart2), beside=TRUE,
args.legend=c(x="topleft",horiz=FALSE,title="Likelihood of heart attack"))
```



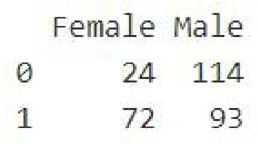
```
chi_heart <- chisq.test(table_heart, p=c(0.5,0.5))

chi_heart
```

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

data:  table_heart
X-squared = 22.717, df = 1, p-value = 1.877e-06
```

```
chi_heart$expected
```

```
     Female      Male
0 43.72277  94.27723
1 52.27723 112.72277
```

```
chi_heart$observed
```



```
   Female Male
 0     24  114
 1     72   93
```

Two sample t test

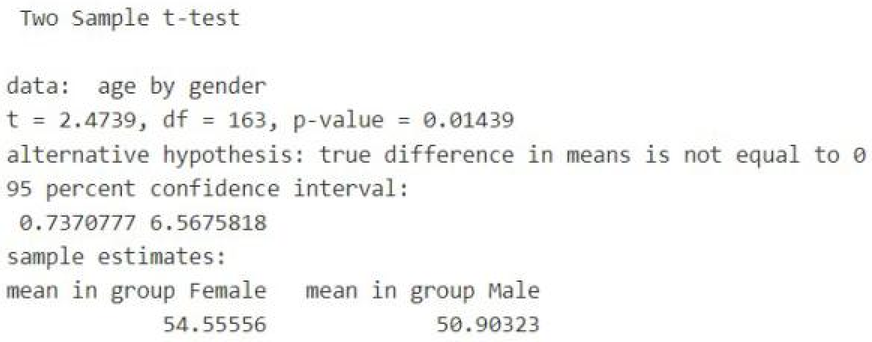
- The next test used will be to understand is there statistical difference in the age of Male and Female of which have a higher chance of heart attack.
- The heart dataset is filtered for individuals who have a high likelihood of heart attack.

```
heart_filtered <- heart %>% filter(target == 1)
```

Student T Test

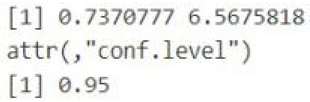
• The t test is used to compare the difference in mean age between the male and female populations. The two sample t tests assume that the populations being compared are independent of one another, and that the data for both the male and female populations are normally distributed and have equal variance. As previously stated, the following assumptions have been verified.

- H0: M1 - M2 = 0
- HA: M1 - M2 =/= 0, where M1 and M2 are the male and female mean ages, respectively.



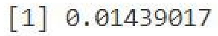
```
 Two Sample t-test

data:  age by gender
t = 2.4739, df = 163, p-value = 0.01439
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7370777 6.5675818
sample estimates:
mean in group Female   mean in group Male 
            54.55556             50.90323 
```

```
result$conf.int
```



```
[1] 0.7370777 6.5675818
attr(,"conf.level")
[1] 0.95
```

```
result$p.value
```



```
[1] 0.01439017
```

- The two sample t test revealed a statistically significant difference in the mean age of male and female patients with a high risk of heart attack. We reject the H0 because p value = 0.014 0.05 and the 95% CI of the estimated population difference [6.57,0.74] does not capture the H0: M1 - M2=0. As a result, the two sample t test was statistically significant, and we can conclude that the age of heart disease in men differs significantly from that in women.

• CONCLUSION

- The first analysis was based on a categorical association to see if gender influences the likelihood of having a heart attack. Based on the visualisation, there appears to be evidence that the female gender is at a higher risk of heart attack, but the chi square yields a different result.
- The chi square test of association between gender and likelihood is statistically significant, implying that a person's gender is unrelated to their risk of having a heart attack. As a result, whether a person has a high or low risk of having a heart attack has nothing to do with their gender.
- As a result of testing both assumptions of the t test above, we can conclude that there is normality and equal variance. The t test is used to identify people who are at high risk of having a heart attack. The t test results show that there is a difference in the mean age of Male and Female. Finally, the average age of men and women who are at high risk of having a heart attack differs.

- At a stricter confidence interval of 0.01, however, the p value of 0.014 is greater than 0.01 and the null hypothesis is not rejected, implying that the mean age of Male and Female is the same.

REFERENCES

- World Health Organization 2017, *Cardiovascular disease (CVDs)* Webpage (HTML Format), World Health Organization, Melbourne, https://www.who.int/en/news-room/fact-sheets/detail/cardiovasculardiseases-(cvds)

- https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-ofdeath

- https://www.cdc.gov/heartdisease/facts.htm

- https://news.mit.edu/2019/machine-learning-shows-no-difference-anginasymptoms-between-men-and-women-1106