Assignment 2

Harold Choi 3866530

9/28/2020

Introduction

- Heart disease claims millions of lives globally, and is a leading cause of death for men and women alike. According to WHO, cardiovascular disease is the number one cause of death globally and an estimated 17.9 million people have died from cardiovascular disease of which 85% are due to heart attack and stroke. There are numerous health factors that contribute to the prediction of a heart attack such as inherited blood disorders, resting heart rate, cholesterol many more.

Problem Statement

- Do male or female have a higher likelihood of a heart attack? At what age is it most common to have a heart attack? This investigation will explain the likelihood of a heart attack through the comparison of gender and age.
- The Male and Female age, gender and target variable are used in this investigation. After giving the summary statistics for male and female separately, a histogram and barplot is plotted to present the data in a more comprehensive manner. In addition, a Chi square test of association between the gender and risk of heart attack is used to test if there is an association between the gender and risk of heart attack. Furthermore, a normality test is used to ensure that the data is is drawn from a normal population distribution. After both assumptions are tested, a two sample t tests is used to evaluate if there is a difference in the mean age of Male and Female who have a high chance of heart attack.
- The following hypothesis have been set out and tested through this investigation:
- There is association between gender and risk of heart attack of an individual.
- The age of Male and Female whom have a high chance of a heart attack are the same.

Data

- The dataset used contains 14 variables and 303 observations of various individuals from the United States of America, Cleverland. The dataset used in this task was generated from: https://www.kaggle.com/madhav000/attack-prediction-accuracy-more-than-80 (https://www.kaggle.com/madhav000/attack-prediction-accuracy-more-than-80).
- The observation of the data was drawn from https://archive.ics.uci.edu/ml/datasets/Heart+Disease (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). The dataset contain 76 variables from different countries such as Hungary and Switzerland but have been subset for easier analysis and the Cleverland database is the one that have 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 have 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 masured 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.
- ca: Number of major vessels colored by flouroscopy, measured from 0 to 3.
- 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)
```

- The file is imported using read_csv and the sex and target variables are ordered using factor. In addition I have renamed the sex variable to gender.
- The summary statistics group by Male and Female of the age of the individuals in this study is shown below. The summarise function is used to create a new data frame that contains 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	Min	Q1	Median	Q3	Max	Mean	SD	n	Missing
<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>

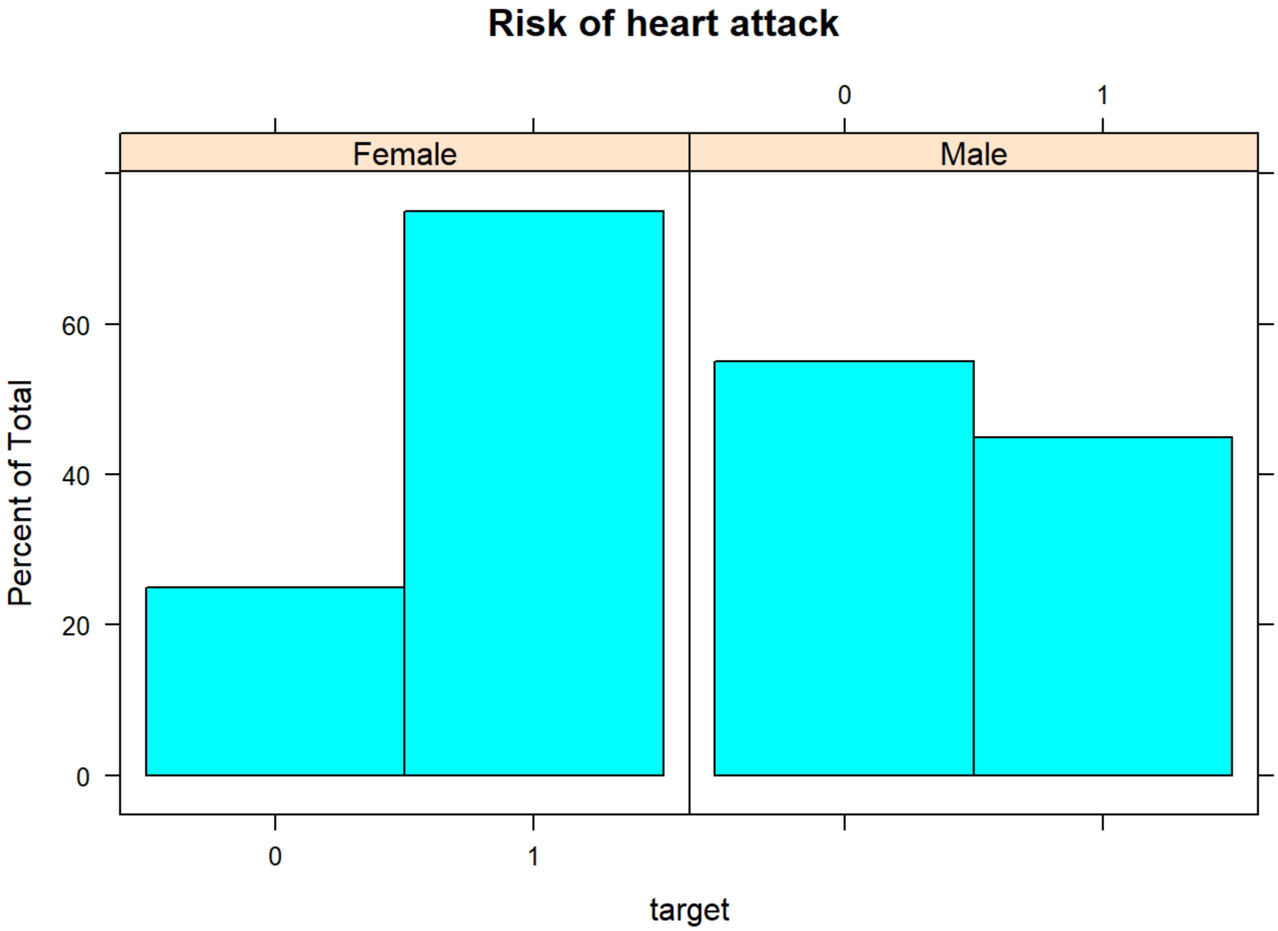
gender <fctr>	Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	n <int>	Missing <int>
Female	34	49.75	57	63.0	76	55.67708	9.409396	96	0
Male	29	47.00	54	59.5	77	53.75845	8.883803	207	0

2 rows

Descriptive Statistics and Visualisation

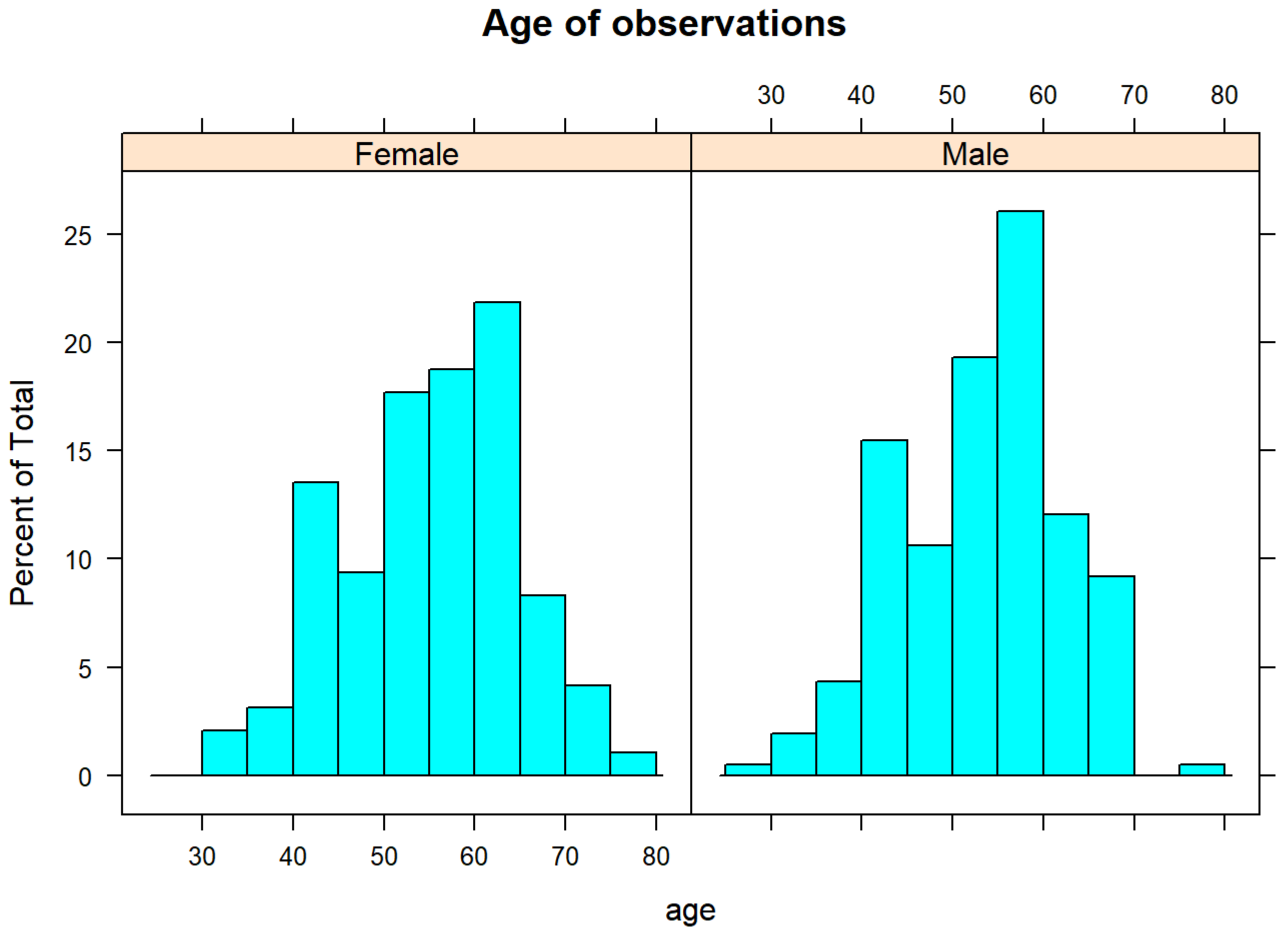
- The mean age of gender is very close, the Male age being 53.76 and the Female age being 55.68.
- In terms of standard deviation, the Female have a higher standard deviation and there a higher standard error than the Male.
- By observing the histogram, Female appears to have a higher likelihood of a heart attack as the proportion is much bigger compared to the Male whose likelihood of heart attack looks 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.

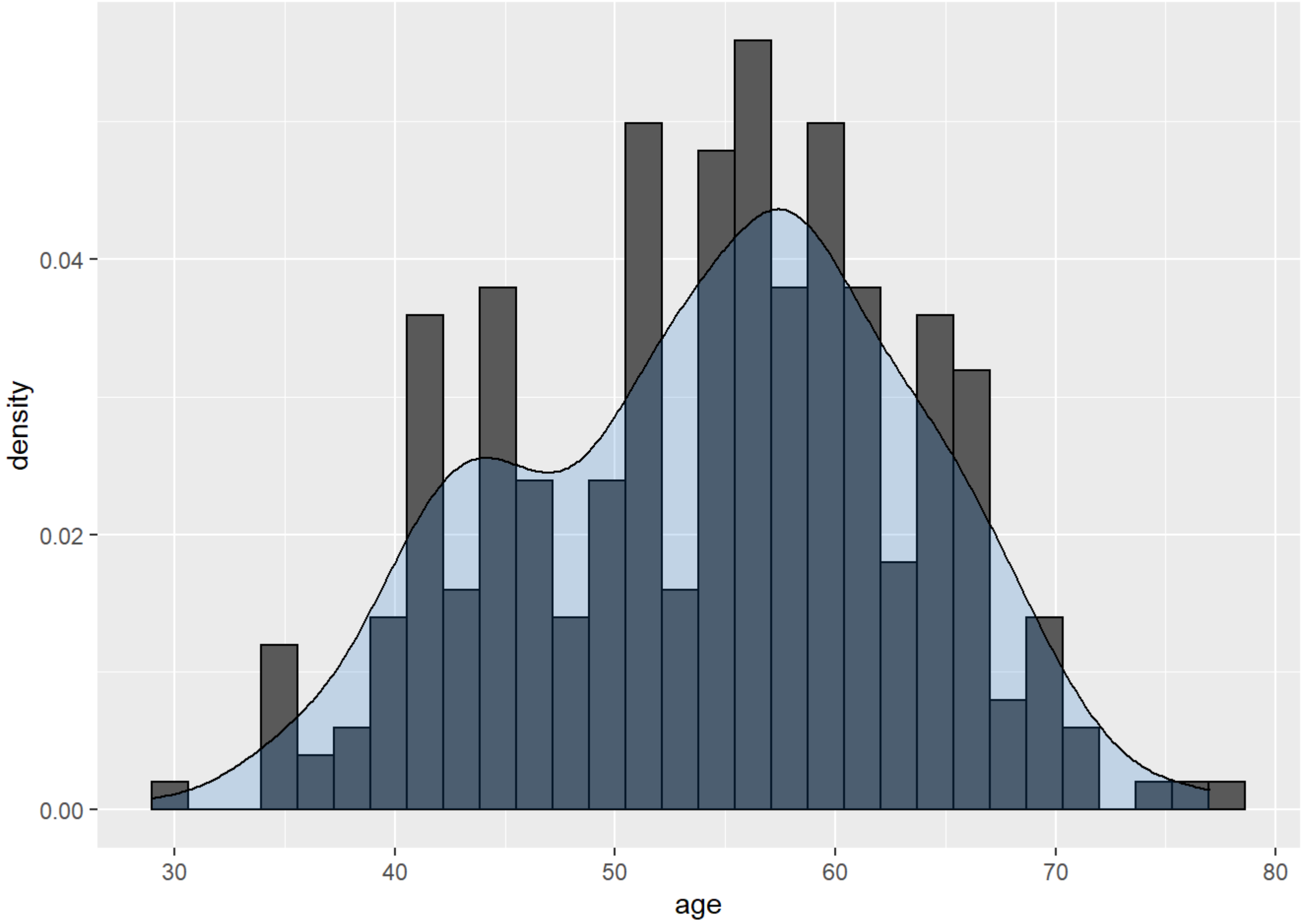
```
heart %>% histogram(~age | gender, data= ., main = "Age of observations", breaks=10)
```



- 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="black")+
          geom_density(alpha=.2, fill="dodgerblue3")
```

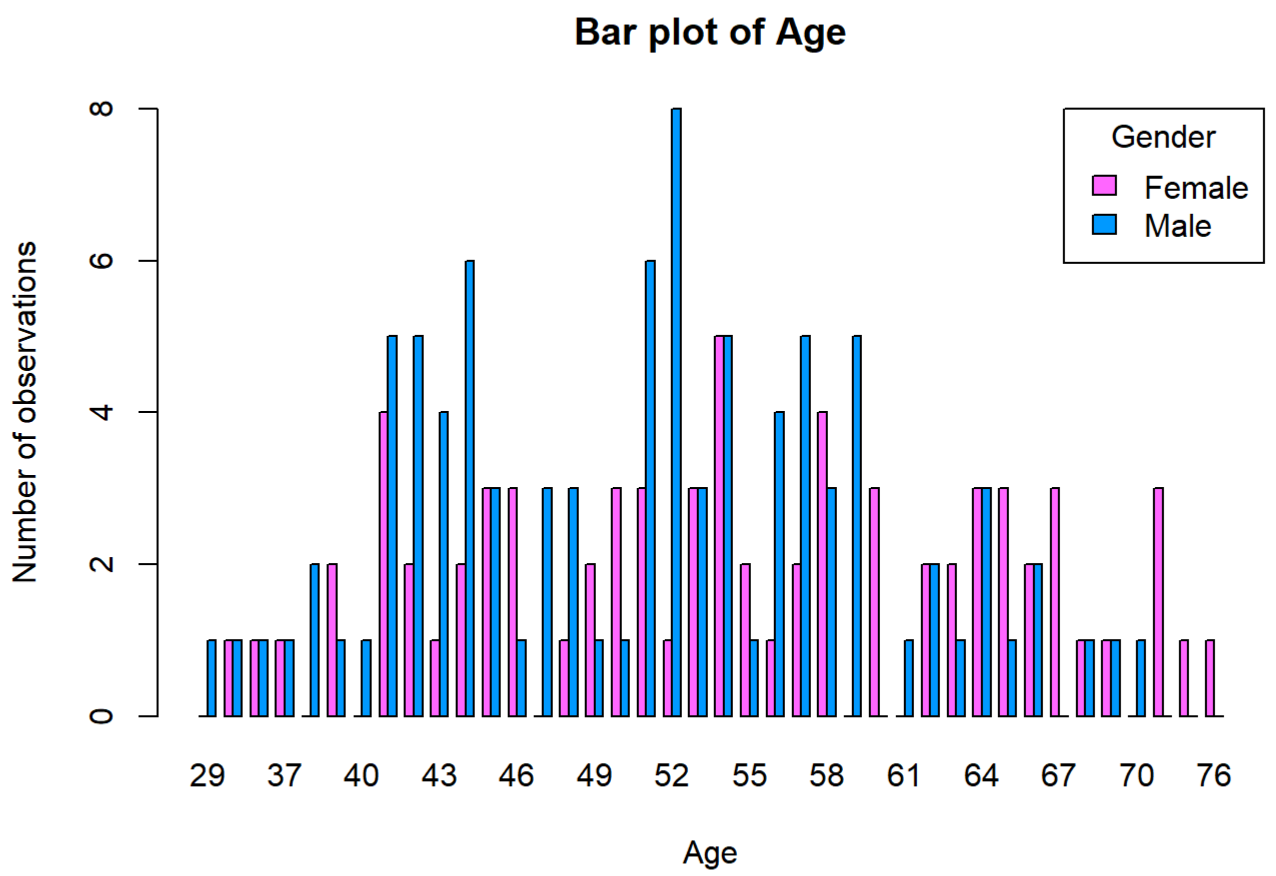
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- The heart dataset is filter for individuals who have a high likelihood of heart attack.
- To compare the age of these individuals, the plot below shows the barplot side by side. By observing we can see that there are greater number of observations of ages for the Male compared to the Female. However just observing the plot is inconclusive of whether Male and Female have the same age of a high likelihood of a heart attack.

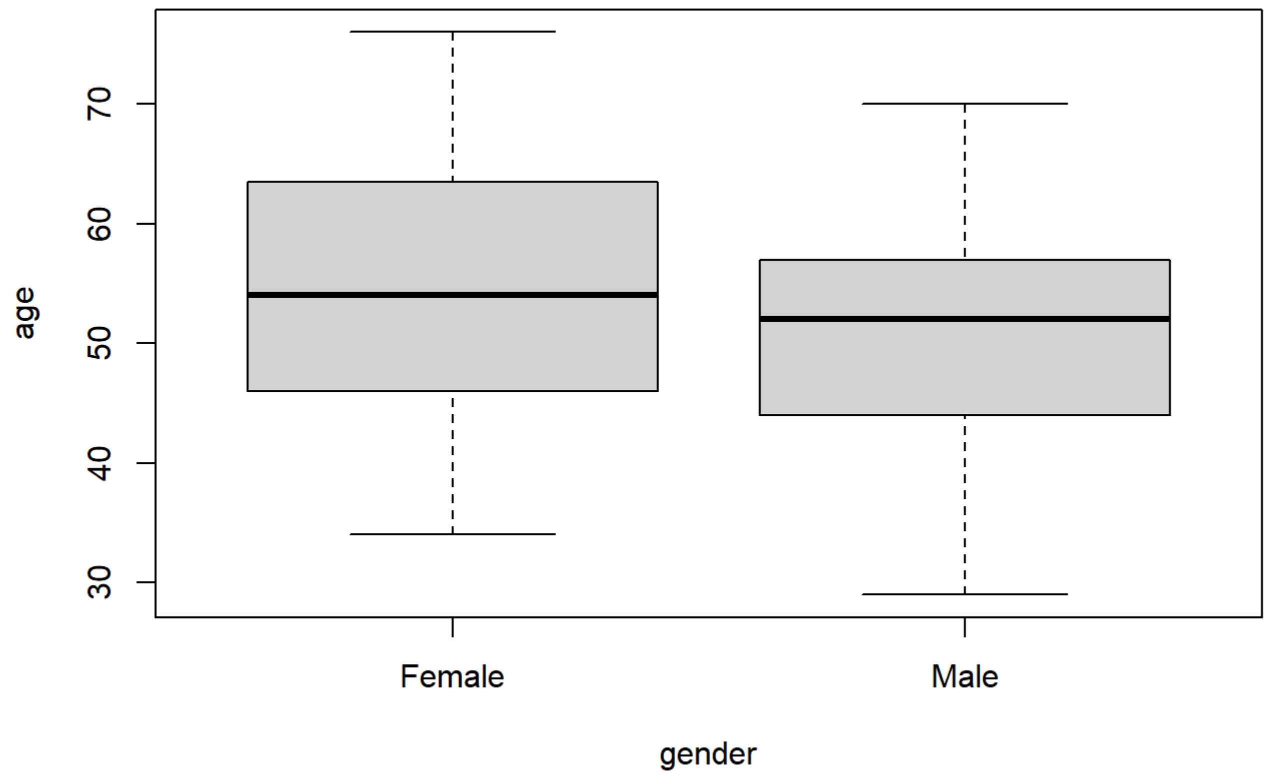
```
heart_filtered <- heart %>% filter(target == 1)
table_age <- table(heart_filtered$gender, heart_filtered$age)

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 shows that there are no outliers in the data that requires to be dealt with. The interquartile range of the Female is greater than the Male which indicates there is a greater variability in the age. Based on the summary statistics, the Female have a higher mean age of 54.56 compared to Male at 50.90. In addition, the standard deviation of Female is 10.27, which is higher than that of Male at 8.68.

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



```
heart_summary2 <- heart_filtered %>% group_by(gender) %>% summarise(Mean = round(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	Mean	Min	Q1	Median	Q3	Max	n
<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<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 = round(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` = round(Mean - tcrit * S
E,2),
                                                  `95% CI Upper Bound` = round(Mean + tcrit * S
E,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

- State the null and alternate hypothesis
- H0: Likelihood that heart attack and gender are related.
- H1: Likelihood that heart attack and gender are not related.
- Assumption is that no more than 25% of expected counts are less than 5 and all individual counts are 1 or greater.
- Table_heart2 shows that Male have a 0.44 chance of high likelihood of heart attack while Female have a 0.75 chance of high likelihood of heart attack.

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

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

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

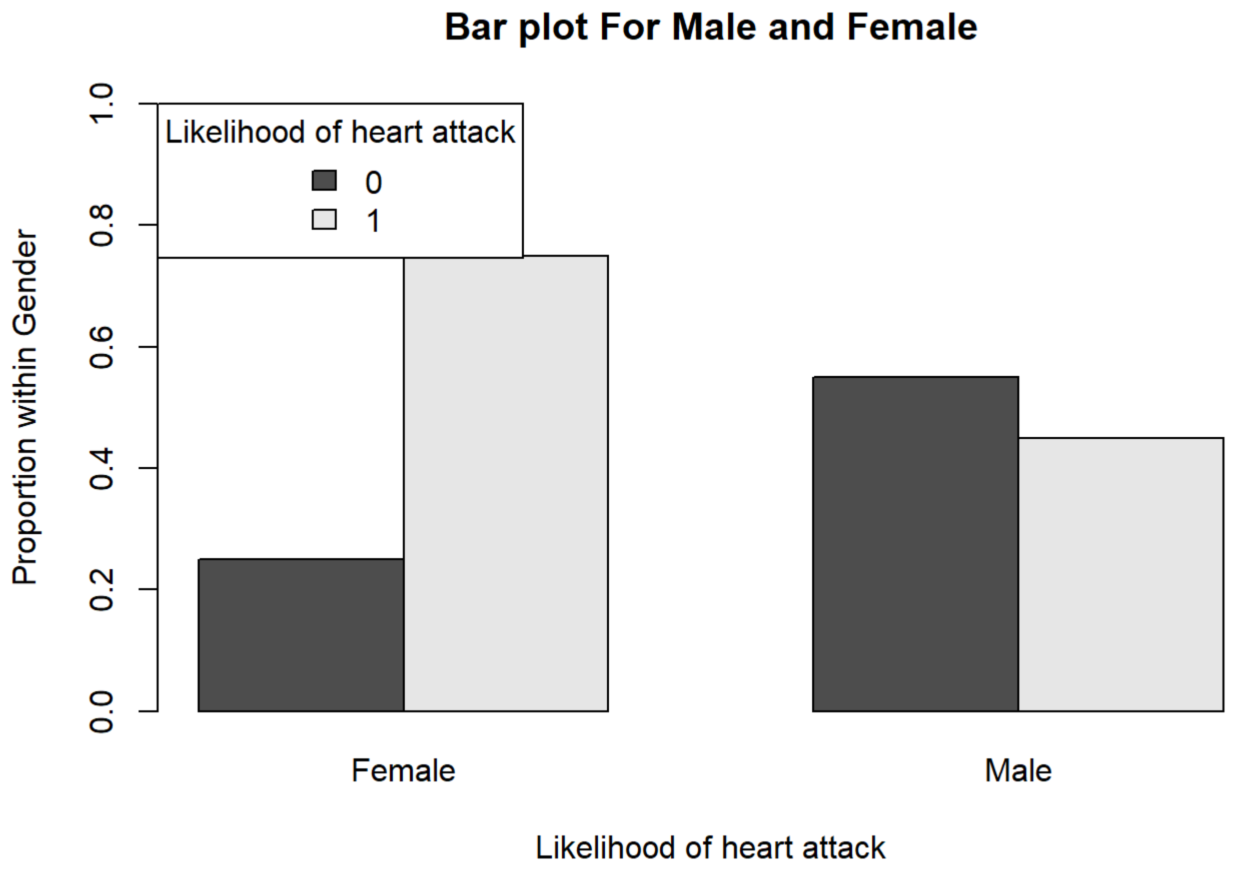
```
##         
##          Female Male Sum
##   0          24  114 138
##   1          72   93 165
##   Sum        96  207 303
```

```
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 was used to test for a statistically between the gender and likelihood of heart attack. The result of the test found a statistically significant association, X^2 = 22.717, p-value = 1.877e-06 < 0.001. The results suggest that there is no evidence of an association between the likelihood of a heart attack and the gender of the individual.Therefore the likelihood of a heart attack is independent of the gender of the individual.

```
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

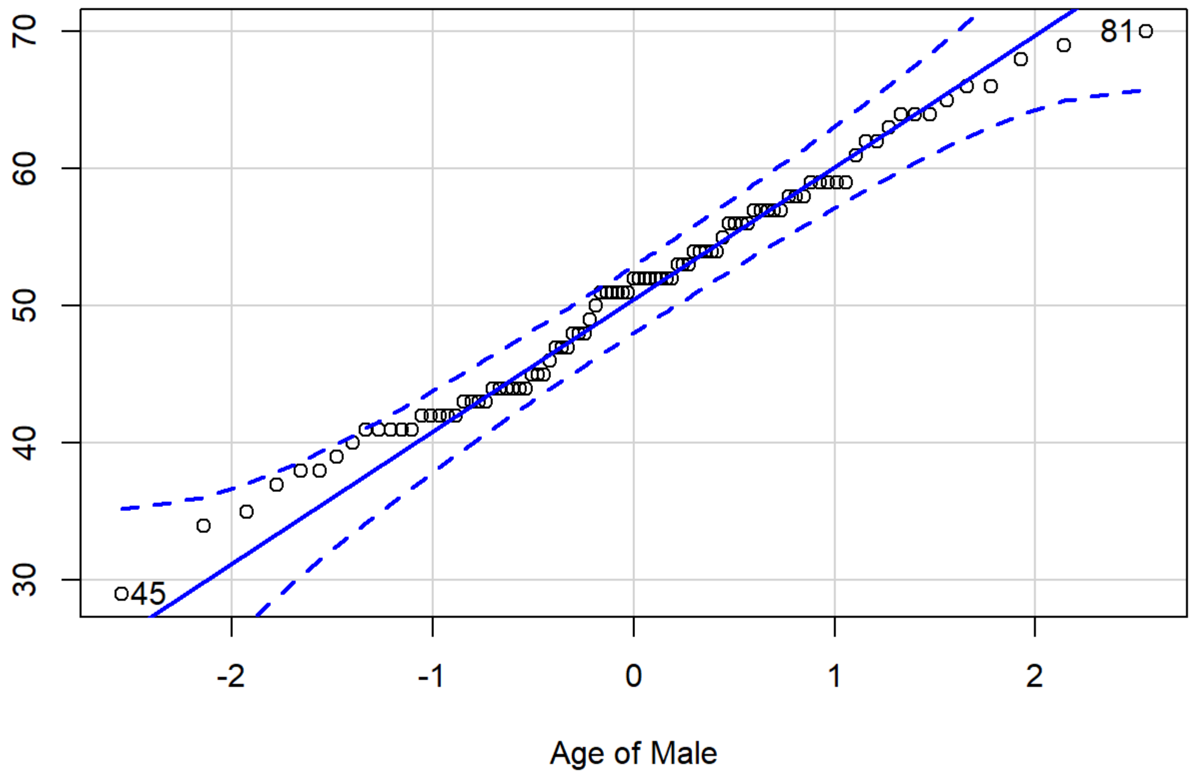
- The next test used will be to understand is there statistical difference in the age of Male and Female of which have a higher chance of heart attack.
- The heart dataset is filtered for individuals who have a high likelihood of heart attack.

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

Testing normality assumption

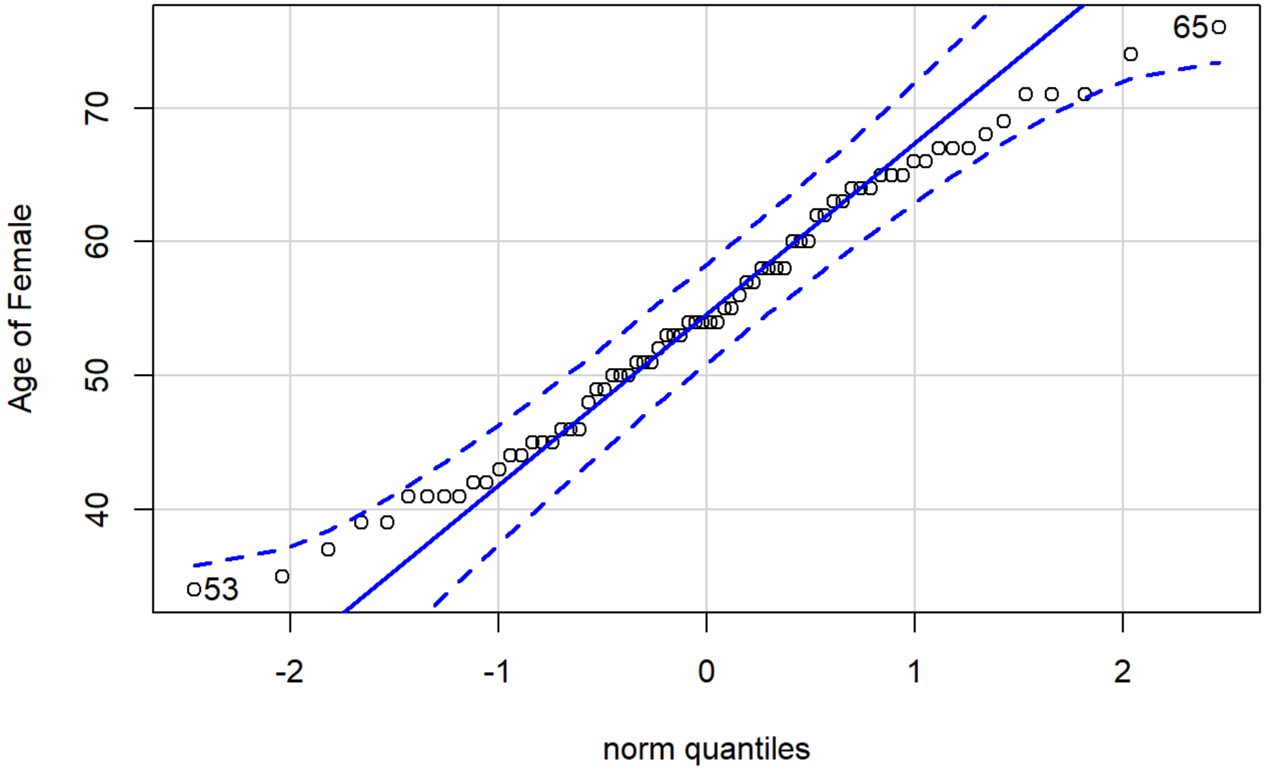
- The data for male and female who have a high chance of heart attack is plotted visually to check the normality.

```
heart_male <- heart_filtered %>% filter(gender=="Male")
heart_male$age %>% qqPlot(dist="norm", xlab = "Age of Male")
```



```
## [1] 45 81
```

```
heart_female <- heart_filtered %>% filter(gender=="Female")
heart_female$age %>% qqPlot(dist="norm", ylab = "Age of Female")
```



```
## [1] 65 53
```

- By observing the plots, the data points falls within the dashed lines, suggesting that the observation are within the 95% confidence intervals. Therefore we can conclude the normality of both gender as the data points follows a trend of falling inside the dashed lines. In addition, the samples for Male and Female are 93 and 72 respectively and due to Central Limit Theorem we can assume that the distribution of mean will be approximately normally distributed.

Test homogeneity of variance

- Homogeneity of variance is testing using the Levene's test in the car package.
- The following hypothesis is used to for the Levene test:
- H0: Var1 = Var2
- HA: Var1 =/= Var2 ** where Var1 and Var2 refers to the Male and Female population variance respectively.

```
leveneTest(age ~ gender, data = heart)
```

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	1	0.3630374	0.5472778
	301	NA	NA

2 rows

- The p-value for the Levene test of equal variance for age of Male and Female was 0.09271. Since the p-value is greater 0.05 the level of significance, the Levene test is not statistically significant and can assume that population variance of Male and Female are homogeneous and can safely assume equal variance.
- Since both the assumption of normality and homogeneity of variance is tested, the sampling distribution of mean is approximately normal and the variance of Male and Female age has a equal variance. The two sample t test is now applied.

Student t test

- The t test is used to compare the difference between the Male and Female population mean age. The two sample t test assumes the population being compared are independent of each other, the data for both Male and Female population have equal variance and are normally distributed. The following assumption have been checked as mentioned above.

- H0: M1 - M2 = 0

- HA: M1 - M2 =/= 0

where M1 and M2 is the mean age of Male and Female respectively.

```
result <- t.test(age ~ gender, data=heart_filtered,
       var.equal=TRUE, alternative ="two.sided")

result
```

```
##
##  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 result of the two sample t test assuming a equal variance found a statistically significant difference bewteen the mean age of male and female who have high chance of heart attack. A 0.05 level of significance was used. t(df=163)=2.4739, p = 0.014and 95% CI of the estimated population difference [6.57,0.74]. Therefore the two sample t test was is statistically significant and the result of the investigation suggest that the age of heart disease for male is significantly different from female.

Discussion

- The first analysis was based on a categorical assocation if gender plays a role in the likelihood of a heart attack. Based on the visualisation, there looks to be evidence that the Female gender are at a higher risk of heart attack, however using the chi square we see a differing answer.
- Using the chi square test of association between gender and likelihood, the test is statistically significant and we can conclude that a person’s gender is not related to the likelihood of them having a heart attack. Therefore whether a person has a high or low chance of heart attack is not associated with their gender.
- The second investigation was to provide insight on whether male and female have the same age of a high likelihood of a heart attack. The mean age for both the Male and Female exhibited evidence of normality upon inspecting the QQ plot. The central limit theorem also sets out that the distribution of mean of more than 30 samples will convergent to normality. Therefore, t test can be applied due to large sample size in both groups. The Leeve test of homogeneity of variance was applied to check if the variance of the age of male and female are equal. Upon inspecting the p value of 0.09, the investigation suggest that the population variance are homogeneous.
- Therefore, since both assumption of the t test was tested above and there is normality in the age of the individuals and equal variance can be assumed. The t test is used to evaluate individuals with a high likelihood of heart attack. The results of the t test is statistically significant and that there is a evidence of a difference in the mean age of Male and Female. In conclusion, there is a difference in mean age of Male and Female who have a high likelihood of heart attack.

References

World Health Organization 2017, *Cardiovascular disease (CVDs)* Webpage (HTML Format), World Health Organization, Melbourne, viewed 5 September 2020, https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))