Project - Analysis of Heart Disease

Introduction

- Analysis of heart disease is very vital according to the upcoming generation. Disease are rising much more and are by threatened by lives. So, the purpose of choosing this dataset is to predict the rate of heart disease based on cholesterol and other factors that are described in details further and also to check whether risk of heart failure depended on any gender or not.

Executive Summary

- Divided by gender that is male and female who are most likely to get heart attack and based on age too. The investigation further explains and solves the latter problem.
- There are various attributes that are used to investigate the above problem. Histogram and boxplot is used to understand better. Various tests are used to investigate more clearly.
- Hypothesis Test is carried out to find the relation between gender and risk of heart failure.
- Linear Regression testing is also carried in this investigation to check the accuracy level of the application
- The purpose of this investigation is mainly depended on the male and female who has most likely to have heart attack or a heart disease according to their age.

Part1 : Data Preparation

Loading Packages

```
library(MASS)
library(MLmetrics)
library(ISLR)
library(readr)
library(dplyr)
library(car)
library(lattice)
library(ggplot2)
library(tidyverse)
library(hexbin)
options(scipen=2)
```

Exploring Dataset

```
Heart_Data <- read.csv("D:/Foram/Others/Illinois Drive Sem III/ProgCyberAnalytics/heart.csv")
#View(Heart_Data)
str(Heart_Data)
```

```
## 'data.frame':    918 obs. of  12 variables:
##  $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
##  $ Sex           : int  0 1 0 1 0 0 1 0 0 1 ...
##  $ ChestPainType : int  2 3 2 4 3 3 2 2 4 2 ...
##  $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
##  $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
##  $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RestingECG    : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
##  $ ExerciseAngina: int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
##  $ ST_Slope      : int  1 2 1 2 1 1 1 1 2 1 ...
##  $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

Summary of Heart Disease Dataset

```
summary(Heart_Data)
```

```
##       Age             Sex         ChestPainType     RestingBP
##  Min.   :28.00   Min.   :0.0000   Min.   :1.000   Min.   :  0.0
##  1st Qu.:47.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0
##  Median :54.00   Median :0.0000   Median :4.000   Median :130.0
##  Mean   :53.51   Mean   :0.2102   Mean   :3.252   Mean   :132.4
##  3rd Qu.:60.00   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0
##   Cholesterol      FastingBS        RestingECG         MaxHR
##  Min.   :  0.0   Min.   :0.0000   Min.   :0.0000   Min.   : 60.0
##  1st Qu.:173.2   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:120.0
##  Median :223.0   Median :0.0000   Median :0.0000   Median :138.0
##  Mean   :198.8   Mean   :0.2331   Mean   :0.6035   Mean   :136.8
##  3rd Qu.:267.0   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:156.0
##  Max.   :603.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##  ExerciseAngina      Oldpeak          ST_Slope      HeartDisease
##  Min.   :0.0000   Min.   :-2.6000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median : 0.6000   Median :2.000   Median :1.0000
##  Mean   :0.4041   Mean   : 0.8874   Mean   :1.638   Mean   :0.5534
##  3rd Qu.:1.0000   3rd Qu.: 1.5000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   : 6.2000   Max.   :3.000   Max.   :1.0000
```

Data observations and variables

- There are total 918 observations with 12 variables. Each of them are explained below.
  - Age - Age of the people
  - Sex - Gender (male indicated by 0 and female indicated by 1)
  - Chest Pain Type :
    - Value 1: Typical Angina (TA)
    - Value 2: Atypical Angina (ATA)
    - Value 3: Non-Anginal pain (NAP)
    - Value 4: Asymptomatic (ASY)
  - RestingBP - Blood pressure of the patient while entering hospital
  - Cholesterol - Serum cholesterol in mg/dl

- FastingBS - Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- RestingECG - Resting electrocardiographic results
  - Value 1: Normal
  - Value 2: Abmornal
  - Value 3: Probable or Definite Left Ventricular Hypertrophy
- MaxHR - Maximum Heart Rate
- ExerciseAngina - Exercise Angina 0 for FALSE and 1 for TRUE
- Oldpeak - ST (Stress Test) Depression by exercise wrt rest
- ST_Slope - ST Slope of peak 1 for up sloping 2 for flat 3 for down sloping
- HeartDisease - Final target 0 for FALSE 1 for TRUE

Missing values

```
sum(is.na(Heart_Data))
```

```
## [1] 0
```

> There are no missing values

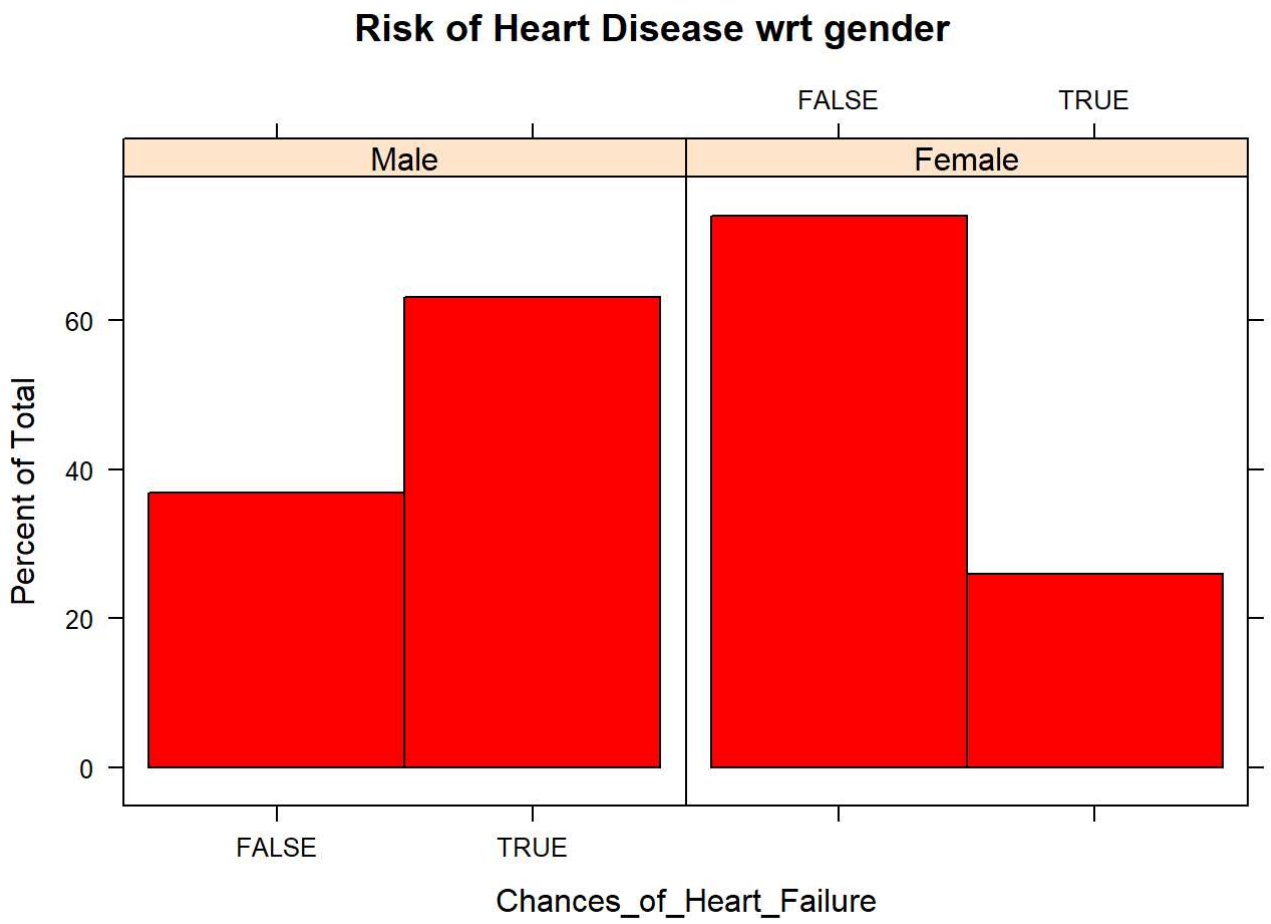
Part 2: Data Exploration

Data Visualization

- To check the risk of heart failure according to gender and age, we created a histogram to see which gender of that age is most likely to have the heart failure.

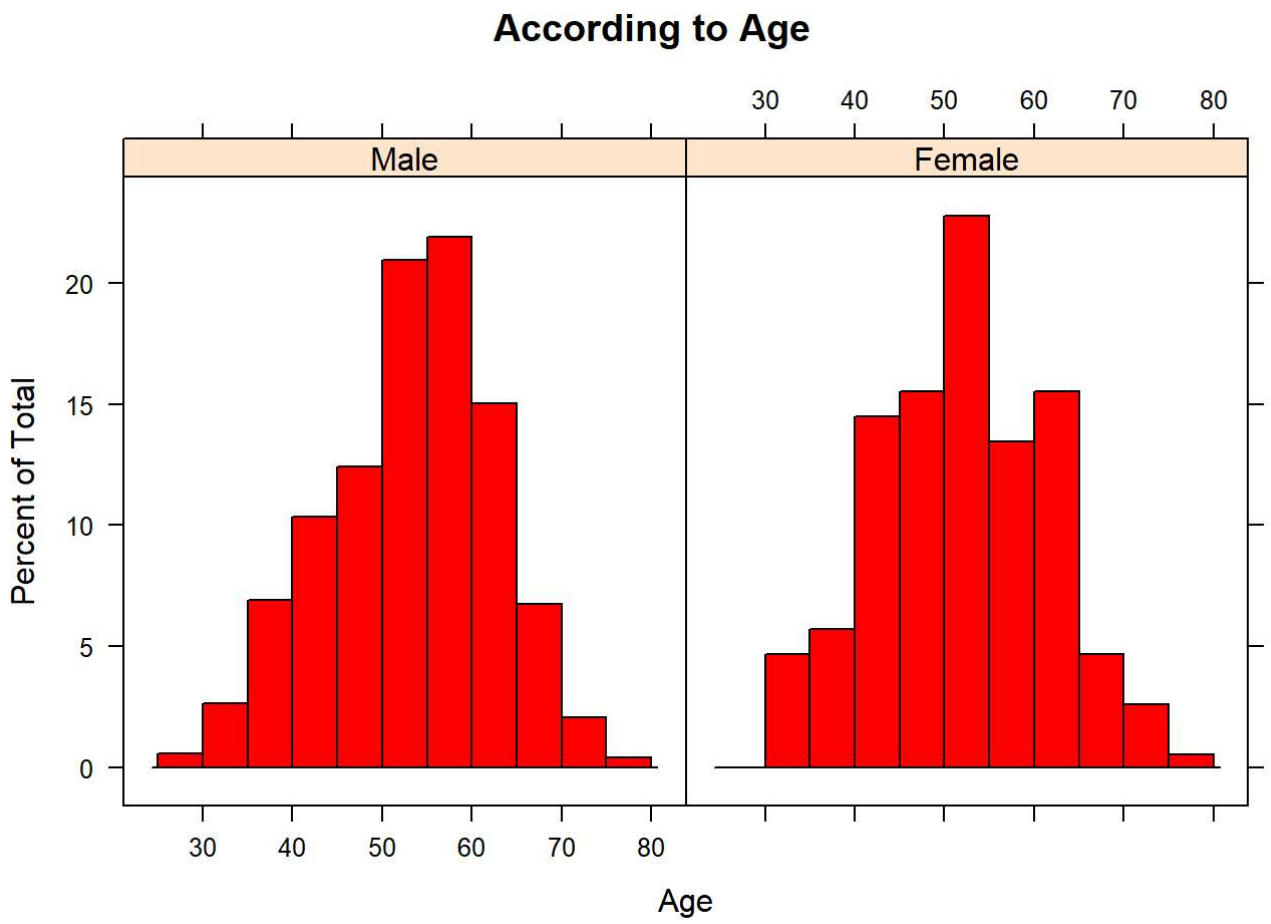
```
#Renaming Labels
Gender <- Heart_Data$Sex %>% factor(levels=c(0,1),
                                    labels=c("Male","Female"))
```

```
#Renaming Labels
Chances_of_Heart_Failure <- Heart_Data$HeartDisease %>% factor(levels=c(0,1),
                                        labels=c(FALSE,TRUE))
histogram(~Chances_of_Heart_Failure | Gender, data= Heart_Data, main = "Risk of Heart Disease
wrt gender" ,col="red")
```



Result : So as per our analysis, chances of heart failure in male is more than female as seen in the above graph of Heart dataset

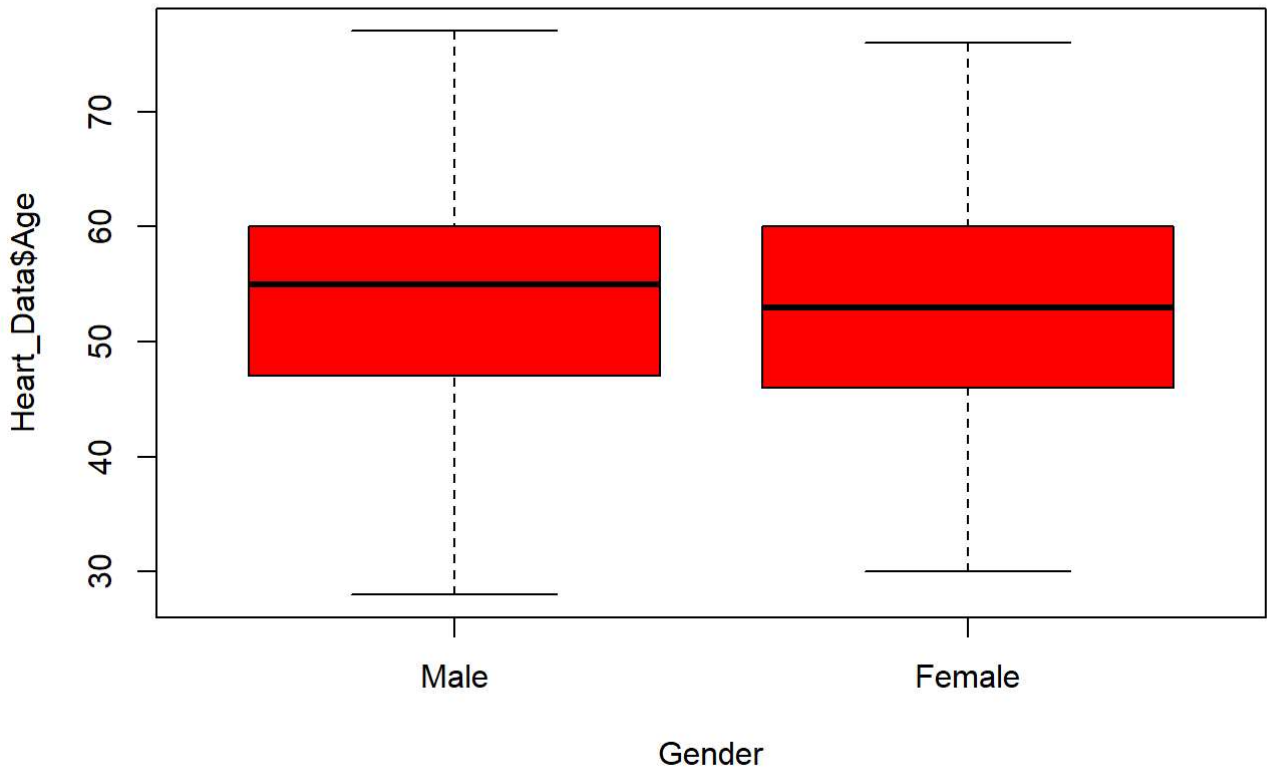
```
histogram(~Age | Gender, data= Heart_Data, main = "According to Age", breaks=10, col="red")
```



Result : So as per our analysis, chances of heart failure in male and female most likely to happen between age 50 and 60.

- To check the risk of heart failure according to gender and age, we created a boxplot to easily identify the relation between age and gender that is more elaborated from the above histogram

```
boxplot(Heart_Data$Age ~ Gender, data = Heart_Data, col="red")
```



Result : So as per our analysis, it shows the same result in a different way to elaborate histogram

Part 3 : Data Analysis

(A) Hypothesis Testing

- To check where there is association between gender and risk of heart failure we need to do hypothtesis testing

```
summary(Heart_Data$Sex)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.2102  0.0000  1.0000
```

```
summary(Heart_Data$HeartDisease)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.5534  1.0000  1.0000
```

```
var(Heart_Data$Sex)
```

```
## [1] 0.16622
```

```
var(Heart_Data$HeartDisease)
```

```
## [1] 0.2474204
```

- H0: Gender and Heart failure are related
- H1: Gender and Heart failure aren't related

```
vartestResult <- var.test(Heart_Data$Sex,Heart_Data$HeartDisease)
vartestResult
```

```
##
##  F test to compare two variances
##
## data:  Heart_Data$Sex and Heart_Data$HeartDisease
## F = 0.67181, num df = 917, denom df = 917, p-value = 1.956e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5901936 0.7647175
## sample estimates:
## ratio of variances
##           0.671812
```

Result : As pvalue 1.956^{-09} < 0.05. We should reject H0. That is, risk of heart failure and gender are not related to each other

- Based on the above result, we need to check if there is evidence of difference between the gender and risk of heart failure
    - H0: Means of Gender = Heart failure
    - H1: Means of Gender ≠ Heart failure

```
ttestResult <-t.test(Heart_Data$Sex,Heart_Data$HeartDisease)
ttestResult
```

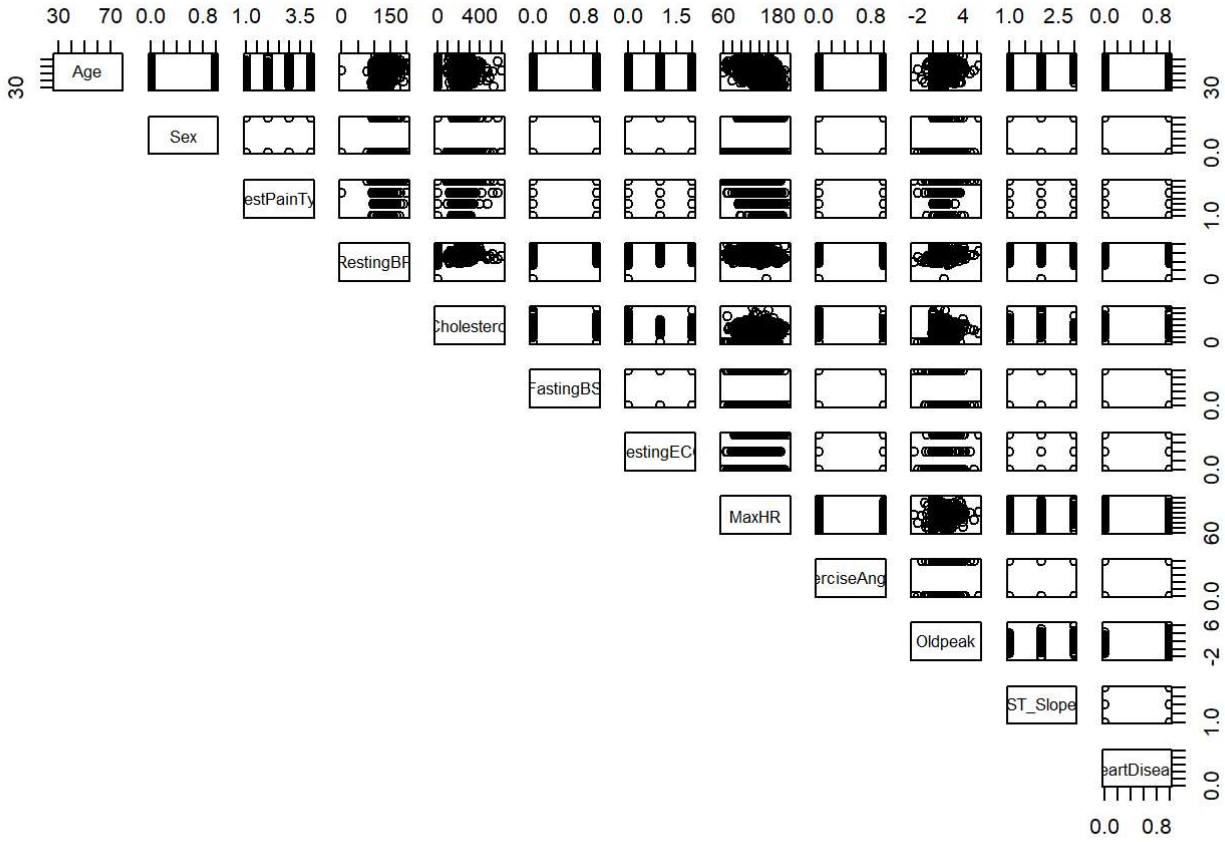
```
##
##  Welch Two Sample t-test
##
## data:  Heart_Data$Sex and Heart_Data$HeartDisease
## t = -16.165, df = 1765.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3847701 -0.3015044
## sample estimates:
## mean of x mean of y
## 0.2102397 0.5533769
```

Result: As pvalue 2.2^-16 < 0.05. We should reject H0. There is no evidence of a significant relation between Risk of failure and gender

(B) Linear Regression Model

Exploratory Data Analysis

```
pairs(Heart_Data, lower.panel = NULL) #checking correlation between variables
```




```
cor(Heart_Data) # produces correlation between variables
```

```
##                      Age          Sex ChestPainType    RestingBP Cholesterol
## Age            1.00000000 -0.055750099    0.16589586  0.254399356 -0.09528177
## Sex           -0.05575010  1.000000000   -0.16825414 -0.005132708  0.20009232
## ChestPainType  0.16589586 -0.168254135    1.00000000  0.022168346 -0.13613950
## RestingBP      0.25439936 -0.005132708    0.02216835  1.000000000  0.10089294
## Cholesterol   -0.09528177  0.200092323   -0.13613950  0.100892942  1.00000000
## FastingBS      0.19803907 -0.120075988    0.11670254  0.070193336 -0.26097433
## RestingECG     0.21315196  0.018343366    0.03138321  0.097661436  0.11209457
## MaxHR         -0.38204468  0.189185764   -0.34365368 -0.112134997  0.23579240
## ExerciseAngina 0.21579269 -0.190664102    0.41662480  0.155101089 -0.03416587
## Oldpeak        0.25861154 -0.105733537    0.24502682  0.164803043  0.05014811
## ST_Slope       0.26826399 -0.150692544    0.31747954  0.075162174 -0.11147054
## HeartDisease   0.28203851 -0.305444916    0.47135450  0.107588980 -0.23274064
##                 FastingBS RestingECG       MaxHR ExerciseAngina     Oldpeak
## Age            0.19803907 0.21315196 -0.38204468     0.21579269  0.25861154
## Sex           -0.12007599 0.01834337  0.18918576    -0.19066410 -0.10573354
## ChestPainType  0.11670254 0.03138321 -0.34365368     0.41662480  0.24502682
## RestingBP      0.07019334 0.09766144 -0.11213500     0.15510109  0.16480304
## Cholesterol   -0.26097433 0.11209457  0.23579240    -0.03416587  0.05014811
## FastingBS      1.00000000 0.05070670 -0.13143849     0.06045067  0.05269786
## RestingECG     0.05070670 1.00000000  0.04855228     0.03611881  0.11442795
## MaxHR         -0.13143849 0.04855228  1.00000000    -0.37042487 -0.16069055
## ExerciseAngina 0.06045067 0.03611881 -0.37042487     1.00000000  0.40875250
## Oldpeak        0.05269786 0.11442795 -0.16069055     0.40875250  1.00000000
## ST_Slope       0.17577434 0.07880669 -0.34341944     0.42870594  0.50192127
## HeartDisease   0.26729119 0.06101109 -0.40042077     0.49428199  0.40395072
##                  ST_Slope HeartDisease
## Age            0.26826399   0.28203851
## Sex           -0.15069254  -0.30544492
## ChestPainType  0.31747954   0.47135450
## RestingBP      0.07516217   0.10758898
## Cholesterol   -0.11147054  -0.23274064
## FastingBS      0.17577434   0.26729119
## RestingECG     0.07880669   0.06101109
## MaxHR         -0.34341944  -0.40042077
## ExerciseAngina 0.42870594   0.49428199
## Oldpeak        0.50192127   0.40395072
## ST_Slope       1.00000000   0.55877071
## HeartDisease   0.55877071   1.00000000
```

- Include the package and data; Let's split the data set into two parts HeartTraining and HeartTesting

```
i <- sample(2, nrow(Heart_Data), replace = TRUE, prob = c(0.8, 0.2) )
HeartTraining <- Heart_Data [i==1,]
HeartTesting <- Heart_Data [i==2,]
```

Multiple Linear Regression

```
#Using all Attributes to construct a linear model.
f_lm <- lm(Cholesterol~., , data=HeartTraining)
summary(f_lm)
```

```
##
## Call:
## lm(formula = Cholesterol ~ ., data = HeartTraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -276.38  -48.53   11.44   64.41  374.82
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     39.4692    51.3610   0.768 0.442469
## Age             -0.4526     0.4656  -0.972 0.331250
## Sex             37.6050     9.4052   3.998 7.06e-05 ***
## ChestPainType   -4.6418     4.7705  -0.973 0.330872
## RestingBP        0.6737     0.2106   3.200 0.001438 **
## FastingBS      -55.5399     9.6210  -5.773 1.17e-08 ***
## RestingECG      16.1392     4.7972   3.364 0.000809 ***
## MaxHR            0.7428     0.1799   4.128 4.10e-05 ***
## ExerciseAngina  29.2917     9.4882   3.087 0.002100 **
## Oldpeak          6.7891     4.2318   1.604 0.109093
## ST_Slope        -1.4402     8.0660  -0.179 0.858339
## HeartDisease   -25.4131    10.4650  -2.428 0.015416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.8 on 701 degrees of freedom
## Multiple R-squared:  0.1918, Adjusted R-squared:  0.1791
## F-statistic: 15.12 on 11 and 701 DF,  p-value: < 2.2e-16
```

There are 12 observations and Age,ChestPainType, ExerciseAngina,ST_Slope are all insignificant attributes.

- So, to remove all the insignificant attributes, we did this.

```
#Interpreting the coefficients of all the attributes. And identifying insignificant attributes.
f_lm1 <- lm(Cholesterol~.,+Sex+RestingBP+FastingBS+RestingECG+MaxHR+ExerciseAngina+Oldpeak+HeartDisease, data=HeartTraining)
summary(f_lm1)
```

```
##
## Call:
## lm(formula = Cholesterol ~ ., data = HeartTraining, subset = +Sex +
##     RestingBP + FastingBS + RestingECG + MaxHR + ExerciseAngina +
##     Oldpeak + HeartDisease)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -112.30  -32.29   -8.85   15.17  466.24
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    111.85163   37.29387   2.999  0.00280 **
## Age             -1.54150    0.36684  -4.202 2.99e-05 ***
## Sex             91.67323    9.44678   9.704  < 2e-16 ***
## ChestPainType   -8.70694    4.32723  -2.012  0.04459 *
## RestingBP        0.60803    0.14035   4.332 1.69e-05 ***
## FastingBS      -17.07844    6.59606  -2.589  0.00982 **
## RestingECG       5.15695    4.49390   1.148  0.25155
## MaxHR            0.06707    0.12889   0.520  0.60298
## ExerciseAngina   6.90578    6.42487   1.075  0.28281
## Oldpeak          6.79553    2.67487   2.541  0.01128 *
## ST_Slope       -12.43898    4.21212  -2.953  0.00325 **
## HeartDisease   -50.69756    8.68266  -5.839 8.04e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.03 on 701 degrees of freedom
## Multiple R-squared:  0.3597, Adjusted R-squared:  0.3497
## F-statistic:  35.8 on 11 and 701 DF,  p-value: < 2.2e-16
```

There is improvement in Residual standard error. But there is a improvement in R-Squared value too.

```
#Predict Cholestrol in HeartTesting.
ypred <- predict(object = f_lm, data = HeartTesting)
summary(ypred)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.78  171.10  202.65  199.39  231.59  323.59
```

```
MAE (y_pred = ypred, y_true = HeartTesting$Cholesterol)
```

```
## [1] 94.54886
```

```
MSE (y_pred = ypred, y_true = HeartTesting$Cholesterol)
```

```
## [1] 15104.56
```

Subset Selection Linear Regression Model

Forward Stepwise

```
intercept_only <- lm(Sex ~ 1, data = HeartTraining)
all <- lm(Sex ~. ,data = HeartTraining)
forward <- stepAIC(intercept_only, direction="forward", scope = formula(all))
```

```
## Start:  AIC=-1232.2
## Sex ~ 1
##
##                  Df Sum of Sq    RSS     AIC
## + HeartDisease    1   13.2935 112.98 -1309.5
## + Cholesterol     1    6.3533 119.92 -1267.0
## + ExerciseAngina  1    5.0548 121.22 -1259.3
## + MaxHR           1    4.9941 121.28 -1259.0
## + ChestPainType   1    4.6535 121.62 -1257.0
## + FastingBS       1    2.8680 123.41 -1246.6
## + ST_Slope        1    2.6025 123.67 -1245.0
## + Oldpeak         1    2.2977 123.98 -1243.3
## + Age             1    0.5871 125.69 -1233.5
## <none>                        126.28 -1232.2
## + RestingBP       1    0.1464 126.13 -1231.0
## + RestingECG      1    0.1204 126.16 -1230.9
##
## Step:  AIC=-1309.51
## Sex ~ HeartDisease
##
##                  Df Sum of Sq    RSS     AIC
## + Cholesterol     1   3.15541 109.83 -1327.7
## + MaxHR           1   0.80733 112.18 -1312.6
## + FastingBS       1   0.44290 112.54 -1310.3
## + RestingECG      1   0.40460 112.58 -1310.1
## + ExerciseAngina  1   0.33087 112.65 -1309.6
## <none>                        112.98 -1309.5
## + ChestPainType   1   0.22967 112.75 -1309.0
## + ST_Slope        1   0.18667 112.80 -1308.7
## + Age             1   0.05840 112.93 -1307.9
## + RestingBP       1   0.00734 112.98 -1307.6
## + Oldpeak         1   0.00651 112.98 -1307.5
##
## Step:  AIC=-1327.71
## Sex ~ HeartDisease + Cholesterol
##
##                  Df Sum of Sq    RSS     AIC
## + ExerciseAngina  1   0.56710 109.26 -1329.4
## + MaxHR           1   0.35622 109.47 -1328.0
## <none>                        109.83 -1327.7
## + ST_Slope        1   0.17050 109.66 -1326.8
## + ChestPainType   1   0.16168 109.67 -1326.8
## + RestingECG      1   0.15465 109.67 -1326.7
## + Age             1   0.10145 109.73 -1326.4
## + RestingBP       1   0.09358 109.73 -1326.3
## + Oldpeak         1   0.08577 109.74 -1326.3
## + FastingBS       1   0.07055 109.76 -1326.2
##
## Step:  AIC=-1329.4
## Sex ~ HeartDisease + Cholesterol + ExerciseAngina
##
##                 Df Sum of Sq    RSS     AIC
## + ST_Slope       1   0.35392 108.91 -1329.7
## <none>                       109.26 -1329.4
## + MaxHR          1   0.16750 109.09 -1328.5
```

```
## + Age            1   0.16236 109.10 -1328.5
## + RestingECG     1   0.14018 109.12 -1328.3
## + FastingBS      1   0.08915 109.17 -1328.0
## + RestingBP      1   0.04800 109.21 -1327.7
## + ChestPainType  1   0.04408 109.22 -1327.7
## + Oldpeak        1   0.01618 109.25 -1327.5
##
## Step:  AIC=-1329.71
## Sex ~ HeartDisease + Cholesterol + ExerciseAngina + ST_Slope
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       108.91 -1329.7
## + MaxHR          1  0.217826 108.69 -1329.1
## + Oldpeak        1  0.122295 108.78 -1328.5
## + FastingBS      1  0.122100 108.79 -1328.5
## + RestingECG     1  0.112068 108.80 -1328.5
## + Age            1  0.111587 108.80 -1328.4
## + ChestPainType  1  0.052265 108.86 -1328.0
## + RestingBP      1  0.041499 108.87 -1328.0
```

```
forward$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Sex ~ 1
##
## Final Model:
## Sex ~ HeartDisease + Cholesterol + ExerciseAngina + ST_Slope
##
##
##               Step Df   Deviance Resid. Df Resid. Dev       AIC
## 1                                      712   126.2777 -1232.202
## 2   + HeartDisease  1 13.2934811       711   112.9842 -1309.512
## 3    + Cholesterol  1  3.1554102       710   109.8288 -1327.708
## 4 + ExerciseAngina  1  0.5671033       709   109.2617 -1329.399
## 5       + ST_Slope  1  0.3539236       708   108.9078 -1329.713
```

```
summary(forward)
```

```
##
## Call:
## lm(formula = Sex ~ HeartDisease + Cholesterol + ExerciseAngina +
##     ST_Slope, data = HeartTraining)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50068 -0.31045 -0.12303  0.03778  1.00509
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.1865020  0.0530933   3.513 0.000472 ***
## HeartDisease   -0.2364420  0.0379828  -6.225 8.25e-10 ***
## Cholesterol     0.0006537  0.0001387   4.712 2.95e-06 ***
## ExerciseAngina -0.0775382  0.0351030  -2.209 0.027503 *
## ST_Slope        0.0448509  0.0295685   1.517 0.129752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3922 on 708 degrees of freedom
## Multiple R-squared:  0.1376, Adjusted R-squared:  0.1327
## F-statistic: 28.23 on 4 and 708 DF,  p-value: < 2.2e-16
```

```
library(MLmetrics)
ypre_for <- predict(forward, HeartTesting)
MAE(y_pred =ypre_for, y_true = HeartTesting$Sex )
```

```
## [1] 0.2825996
```

> There is a major improvement in mean absolute error

```
MSE(y_pred = ypre_for, y_true = HeartTesting$Sex)
```

```
## [1] 0.1296843
```

> There is a major improvement in mean squared error

Backward Stepwise

```
backward <- stepAIC(all, direction='backward')
```

```
## Start:  AIC=-1322.4
## Sex ~ Age + ChestPainType + RestingBP + Cholesterol + FastingBS +
##     RestingECG + MaxHR + ExerciseAngina + Oldpeak + ST_Slope +
##     HeartDisease
##
##                  Df Sum of Sq    RSS     AIC
## - ChestPainType   1    0.0340 107.92 -1324.2
## - RestingECG      1    0.0562 107.95 -1324.0
## - RestingBP       1    0.0587 107.95 -1324.0
## - FastingBS       1    0.2093 108.10 -1323.0
## - Oldpeak         1    0.2285 108.12 -1322.9
## <none>                        107.89 -1322.4
## - Age             1    0.3286 108.22 -1322.2
## - ExerciseAngina  1    0.3323 108.22 -1322.2
## - MaxHR           1    0.3416 108.23 -1322.2
## - ST_Slope        1    0.5129 108.40 -1321.0
## - Cholesterol     1    2.4605 110.35 -1308.3
## - HeartDisease    1    4.2599 112.15 -1296.8
##
## Step:  AIC=-1324.18
## Sex ~ Age + RestingBP + Cholesterol + FastingBS + RestingECG +
##     MaxHR + ExerciseAngina + Oldpeak + ST_Slope + HeartDisease
##
##                  Df Sum of Sq    RSS     AIC
## - RestingBP       1    0.0516 107.98 -1325.8
## - RestingECG      1    0.0540 107.98 -1325.8
## - FastingBS       1    0.2051 108.13 -1324.8
## - Oldpeak         1    0.2338 108.16 -1324.6
## <none>                        107.92 -1324.2
## - Age             1    0.3287 108.25 -1324.0
## - MaxHR           1    0.3829 108.31 -1323.7
## - ExerciseAngina  1    0.3976 108.32 -1323.6
## - ST_Slope        1    0.5119 108.44 -1322.8
## - Cholesterol     1    2.4875 110.41 -1309.9
## - HeartDisease    1    4.7782 112.70 -1295.3
##
## Step:  AIC=-1325.84
## Sex ~ Age + Cholesterol + FastingBS + RestingECG + MaxHR + ExerciseAngina +
##     Oldpeak + ST_Slope + HeartDisease
##
##                  Df Sum of Sq    RSS     AIC
## - RestingECG      1    0.0519 108.03 -1327.5
## - FastingBS       1    0.2151 108.19 -1326.4
## - Oldpeak         1    0.2581 108.23 -1326.1
## - Age             1    0.2910 108.27 -1325.9
## <none>                        107.98 -1325.8
## - MaxHR           1    0.3900 108.37 -1325.3
## - ExerciseAngina  1    0.4252 108.40 -1325.0
## - ST_Slope        1    0.5459 108.52 -1324.2
## - Cholesterol     1    2.4371 110.41 -1311.9
## - HeartDisease    1    4.7720 112.75 -1297.0
##
## Step:  AIC=-1327.49
## Sex ~ Age + Cholesterol + FastingBS + MaxHR + ExerciseAngina +
##     Oldpeak + ST_Slope + HeartDisease
```

```
##
##                  Df Sum of Sq    RSS     AIC
## - FastingBS       1    0.2073 108.24 -1328.1
## - Oldpeak         1    0.2495 108.28 -1327.8
## <none>                        108.03 -1327.5
## - Age             1    0.3765 108.41 -1327.0
## - MaxHR           1    0.4300 108.46 -1326.7
## - ExerciseAngina  1    0.4358 108.46 -1326.6
## - ST_Slope        1    0.5565 108.58 -1325.8
## - Cholesterol     1    2.5789 110.61 -1312.7
## - HeartDisease    1    4.7373 112.77 -1298.9
##
## Step:  AIC=-1328.13
## Sex ~ Age + Cholesterol + MaxHR + ExerciseAngina + Oldpeak +
##     ST_Slope + HeartDisease
##
##                  Df Sum of Sq    RSS     AIC
## - Oldpeak         1    0.2222 108.46 -1328.7
## - Age             1    0.2977 108.53 -1328.2
## <none>                        108.24 -1328.1
## - ExerciseAngina  1    0.4056 108.64 -1327.5
## - MaxHR           1    0.4081 108.64 -1327.4
## - ST_Slope        1    0.5042 108.74 -1326.8
## - Cholesterol     1    3.0425 111.28 -1310.4
## - HeartDisease    1    5.2559 113.49 -1296.3
##
## Step:  AIC=-1328.66
## Sex ~ Age + Cholesterol + MaxHR + ExerciseAngina + ST_Slope +
##     HeartDisease
##
##                  Df Sum of Sq    RSS     AIC
## - Age             1    0.2321 108.69 -1329.1
## <none>                        108.46 -1328.7
## - MaxHR           1    0.3383 108.80 -1328.4
## - ST_Slope        1    0.3430 108.80 -1328.4
## - ExerciseAngina  1    0.5340 108.99 -1327.2
## - Cholesterol     1    2.9224 111.38 -1311.7
## - HeartDisease    1    5.6480 114.11 -1294.5
##
## Step:  AIC=-1329.14
## Sex ~ Cholesterol + MaxHR + ExerciseAngina + ST_Slope + HeartDisease
##
##                  Df Sum of Sq    RSS     AIC
## - MaxHR           1    0.2178 108.91 -1329.7
## <none>                        108.69 -1329.1
## - ST_Slope        1    0.4042 109.09 -1328.5
## - ExerciseAngina  1    0.5263 109.22 -1327.7
## - Cholesterol     1    2.9314 111.62 -1312.2
## - HeartDisease    1    5.5033 114.19 -1295.9
##
## Step:  AIC=-1329.71
## Sex ~ Cholesterol + ExerciseAngina + ST_Slope + HeartDisease
##
##                  Df Sum of Sq    RSS     AIC
## <none>                        108.91 -1329.7
## - ST_Slope        1    0.3539 109.26 -1329.4
```

```
## - ExerciseAngina  1    0.7505 109.66 -1326.8
## - Cholesterol     1    3.4160 112.32 -1309.7
## - HeartDisease    1    5.9607 114.87 -1293.7
```

```
backward$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Sex ~ Age + ChestPainType + RestingBP + Cholesterol + FastingBS +
##     RestingECG + MaxHR + ExerciseAngina + Oldpeak + ST_Slope +
##     HeartDisease
##
## Final Model:
## Sex ~ Cholesterol + ExerciseAngina + ST_Slope + HeartDisease
##
##
##              Step Df   Deviance Resid. Df Resid. Dev       AIC
## 1                                    701   107.8909 -1322.402
## 2 - ChestPainType  1 0.03397173       702   107.9248 -1324.177
## 3     - RestingBP  1 0.05164945       703   107.9765 -1325.836
## 4    - RestingECG  1 0.05193934       704   108.0284 -1327.493
## 5     - FastingBS  1 0.20728999       705   108.2357 -1328.126
## 6       - Oldpeak  1 0.22219451       706   108.4579 -1328.664
## 7           - Age  1 0.23205048       707   108.6900 -1329.140
## 8         - MaxHR  1 0.21782613       708   108.9078 -1329.713
```

```
summary(backward)
```

```
##
## Call:
## lm(formula = Sex ~ Cholesterol + ExerciseAngina + ST_Slope +
##     HeartDisease, data = HeartTraining)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50068 -0.31045 -0.12303  0.03778  1.00509
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.1865020  0.0530933   3.513 0.000472 ***
## Cholesterol     0.0006537  0.0001387   4.712 2.95e-06 ***
## ExerciseAngina -0.0775382  0.0351030  -2.209 0.027503 *
## ST_Slope        0.0448509  0.0295685   1.517 0.129752
## HeartDisease   -0.2364420  0.0379828  -6.225 8.25e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3922 on 708 degrees of freedom
## Multiple R-squared:  0.1376, Adjusted R-squared:  0.1327
## F-statistic: 28.23 on 4 and 708 DF,  p-value: < 2.2e-16
```

```
ypre_back <- predict(object = backward, newdata = HeartTesting)
MAE (y_pred = ypre_back, y_true = HeartTesting$Sex )
```

```
## [1] 0.2825996
```

```
MSE (y_pred = ypre_back, y_true = HeartTesting$Sex )
```

```
## [1] 0.1296843
```

> Result: MAE and MSE for both forward and backward are similar. That means the application will run as intended with very minute error with better accuracy.

Conclusion

- As per our investigation people of age 50-60 irrespective of the gender are most likely to have the risk of heart failure. And, there is no evidence in the fact that gender is related to risk of heart failure