Importing the Libraries :-

```
library(dplyr)
library(tidyverse)
library(tidyr)
library(ggplot2)
library(ggpubr)
library(ggcorrplot)
library(caret)
```

Loading the Data :-

```
data <- read.csv("C:/Users/Abhigyan/Downloads/heart.csv")
```

Short Description of Our Data :-

```
glimpse(data)
```

```
Rows: 303
Columns: 14
$ age      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~
$ sex      <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~
$ cp       <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0~
$ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~
$ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
$ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
$ restecg  <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1~
$ thalach  <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
$ exang    <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
$ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~
$ slope    <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~
$ ca       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0~
$ thal     <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
$ target   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

Data Pre-processing :-

```
Heartrate <- data %>%
        mutate(sex = if_else(sex == 1,"MALE","FEMALE"),
               fbs = if_else (fbs == 1 ,">120", "<=120"),
               exang = if_else (exang == 1 ,"YES", "NO"),
                cp = if_else (cp == 0, "TYPICAL", if_else(cp == 1, "ATYPICAL ANGINA",
```

```
                            if_else(cp == 2, "NON-ANGINAL PAIN", "ASYMPTOMATIC"))),
               restecg = if_else(restecg == 0, "NORMAL",
                                 if_else(restecg == 1, "ABNORMALITY", "PROBABLE OR DEFINITE
               slope = as.factor(slope),
               ca = as.factor(ca),
               thal = as.factor(thal),
               target = if_else(target ==1 ,"YES","NO")) %>%
    mutate_if(is.character, as.factor)%>%
    dplyr::select(target,sex,fbs,exang,cp,restecg,slope,ca,thal,everything())
head(Heartrate)
```

```
  target    sex   fbs exang               cp     restecg slope ca thal age
1    YES   MALE  >120    NO     ASYMPTOMATIC      NORMAL     0  0    1  63
2    YES   MALE <=120    NO NON-ANGINAL PAIN ABNORMALITY     0  0    2  37
3    YES FEMALE <=120    NO  ATYPICAL ANGINA      NORMAL     2  0    2  41
4    YES   MALE <=120    NO  ATYPICAL ANGINA ABNORMALITY     2  0    2  56
5    YES FEMALE <=120   YES          TYPICAL ABNORMALITY     2  0    2  57
6    YES   MALE <=120    NO          TYPICAL ABNORMALITY     1  0    1  57
  trestbps chol thalach oldpeak
1      145  233     150     2.3
2      130  250     187     3.5
3      130  204     172     1.4
4      120  236     178     0.8
5      120  354     163     0.6
6      140  192     148     0.4
```

Exploring the Dataset :-

```
glimpse(Heartrate)
```

```
Rows: 303
Columns: 14
$ target   <fct> YES, YES, YES, YES, YES, YES, YES, YES, YES, YES, YES, YES, Y~
$ sex      <fct> MALE, MALE, FEMALE, MALE, FEMALE, MALE, FEMALE, MALE, MALE, M~
$ fbs      <fct> >120, <=120, <=120, <=120, <=120, <=120, <=120, <=120, >120, ~
$ exang    <fct> NO, NO, NO, NO, YES, NO, NO, NO, NO, NO, NO, NO, NO, YES, NO,~
$ cp       <fct> ASYMPTOMATIC, NON-ANGINAL PAIN, ATYPICAL ANGINA, ATYPICAL ANG~
$ restecg  <fct> NORMAL, ABNORMALITY, NORMAL, ABNORMALITY, ABNORMALITY, ABNORM~
$ slope    <fct> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~
$ ca       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0~
$ thal     <fct> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
$ age      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~
$ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~
$ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
$ thalach  <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
$ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~
```

Summary of the Dataset :-

So, using *is.na()* function, we get the number of missing values in our data set is:

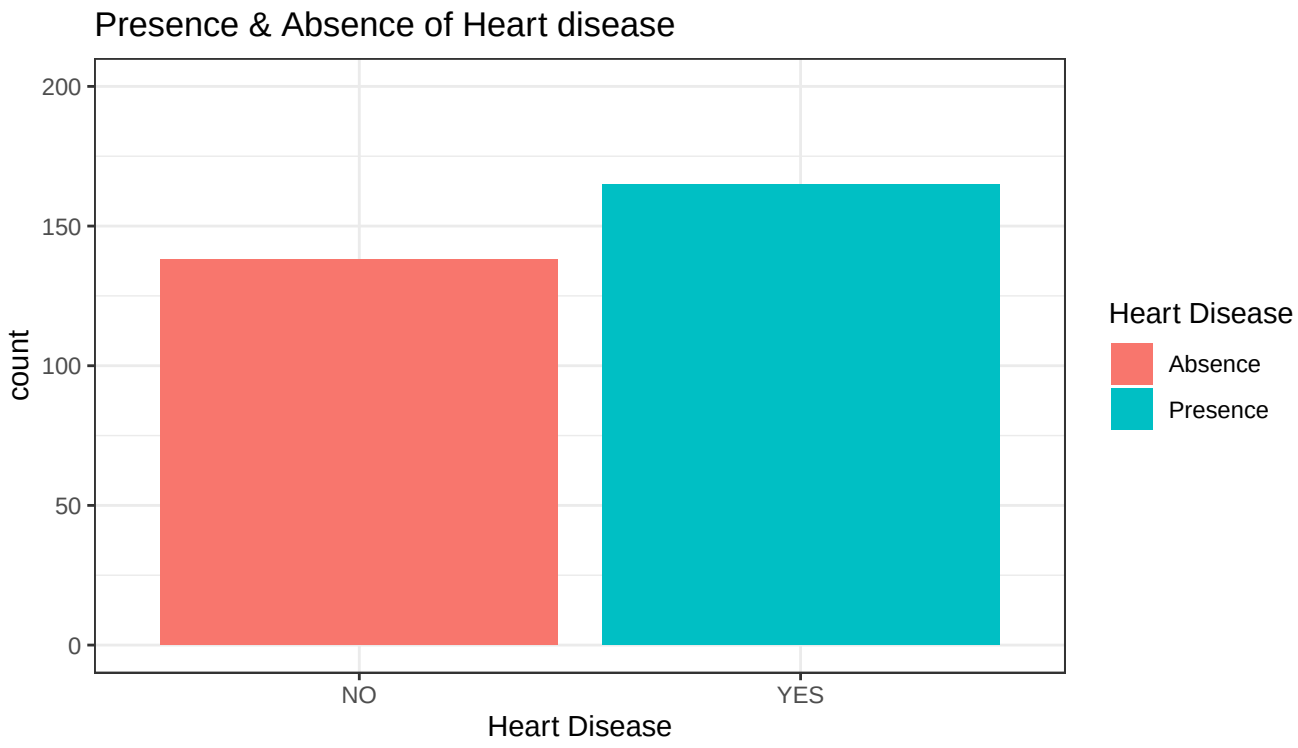
```
 target        sex          fbs       exang                     cp
 NO :138   FEMALE: 96   <=120:258   NO :204   ASYMPTOMATIC    : 23
 YES:165   MALE  :207   >120 : 45   YES: 99   ATYPICAL ANGINA : 50
                                               NON-ANGINAL PAIN: 87
                                               TYPICAL         :143


                 restecg    slope   ca      thal         age
 ABNORMALITY         :152   0: 21   0:175   0:  2   Min.   :29.00
 NORMAL              :147   1:140   1: 65   1: 18   1st Qu.:47.50
 PROBABLE OR DEFINITE:  4   2:142   2: 38   2:166   Median :55.00
                                    3: 20   3:117   Mean   :54.37
                                    4:  5           3rd Qu.:61.00
                                                    Max.   :77.00
    trestbps          chol          thalach         oldpeak
 Min.   : 94.0   Min.   :126.0   Min.   : 71.0   Min.   :0.00
 1st Qu.:120.0   1st Qu.:211.0   1st Qu.:133.5   1st Qu.:0.00
 Median :130.0   Median :240.0   Median :153.0   Median :0.80
 Mean   :131.6   Mean   :246.3   Mean   :149.6   Mean   :1.04
 3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:166.0   3rd Qu.:1.60
 Max.   :200.0   Max.   :564.0   Max.   :202.0   Max.   :6.20
```

Visualize the Shape of the Distribution :-

- ***For Categorical Variables***

Bar Chart for Target :-

```
ggplot(Heartrate, aes(x = target, fill = target)) +
    geom_bar()+
    labs(x = "Heart Disease",
         y = "count",
         title = "Presence & Absence of Heart disease") +
    coord_cartesian(ylim = c(0, 200)) +
    scale_fill_discrete(name= 'Heart Disease', labels =c("Absence", "Presence")) +
    theme_bw()
```



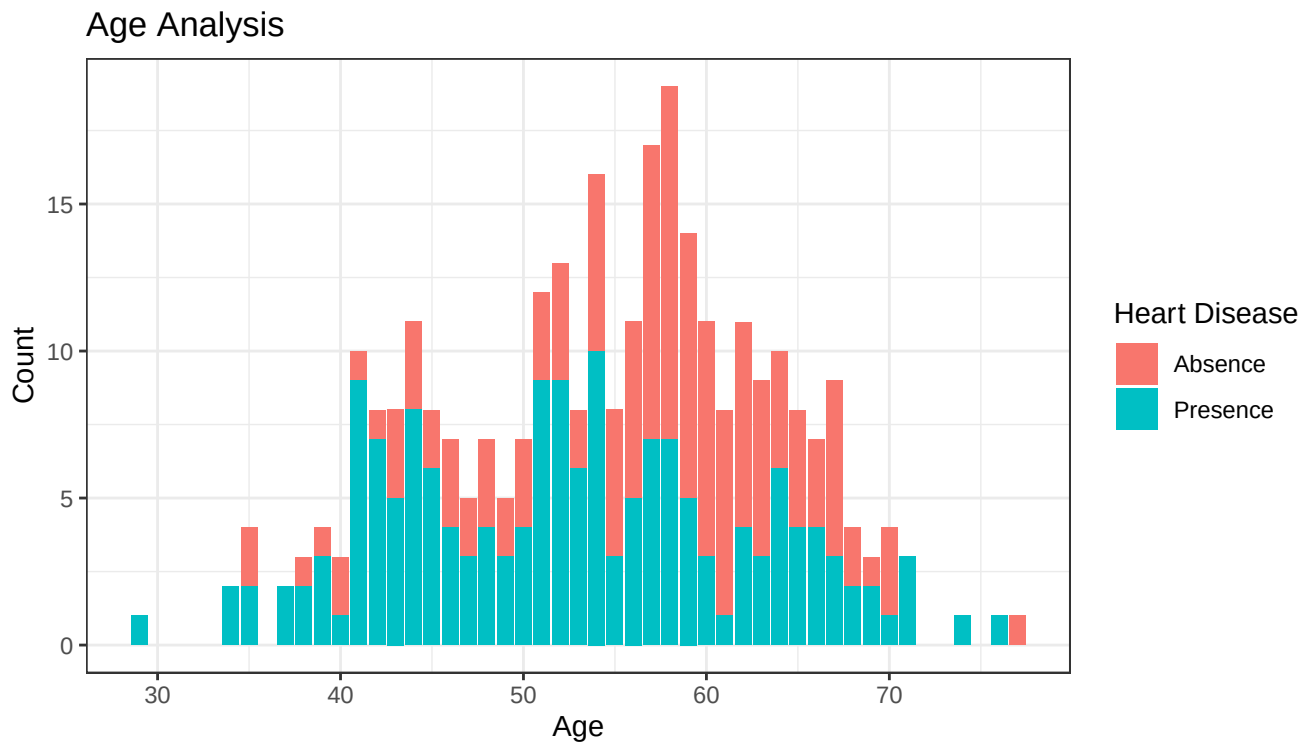
Proportion :-

```
round(prop.table(table(Heartrate$target)), 3)
```

```
   NO   YES
0.455 0.545
```

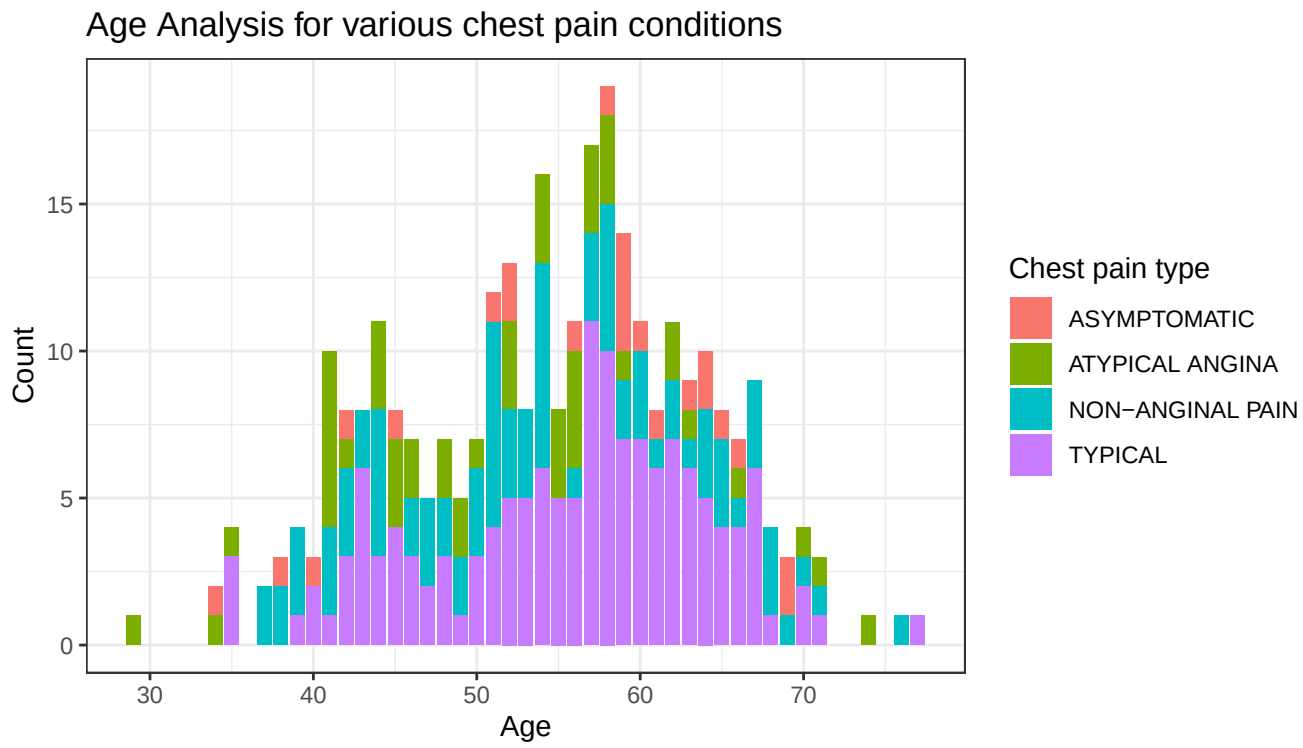
Age Analysis :-

```
age_target <- Heartrate %>%
  group_by(age) %>%
  count(target)

ggplot(data = Heartrate, mapping = aes(x = age, fill = target)) +
  geom_bar() +
  labs(x = "Age",
       y = "Count",
       fill = "Target",
       title = "Age Analysis") +
  scale_fill_discrete(name= 'Heart Disease', labels =c("Absence", "Presence")) +
  theme_bw()
```



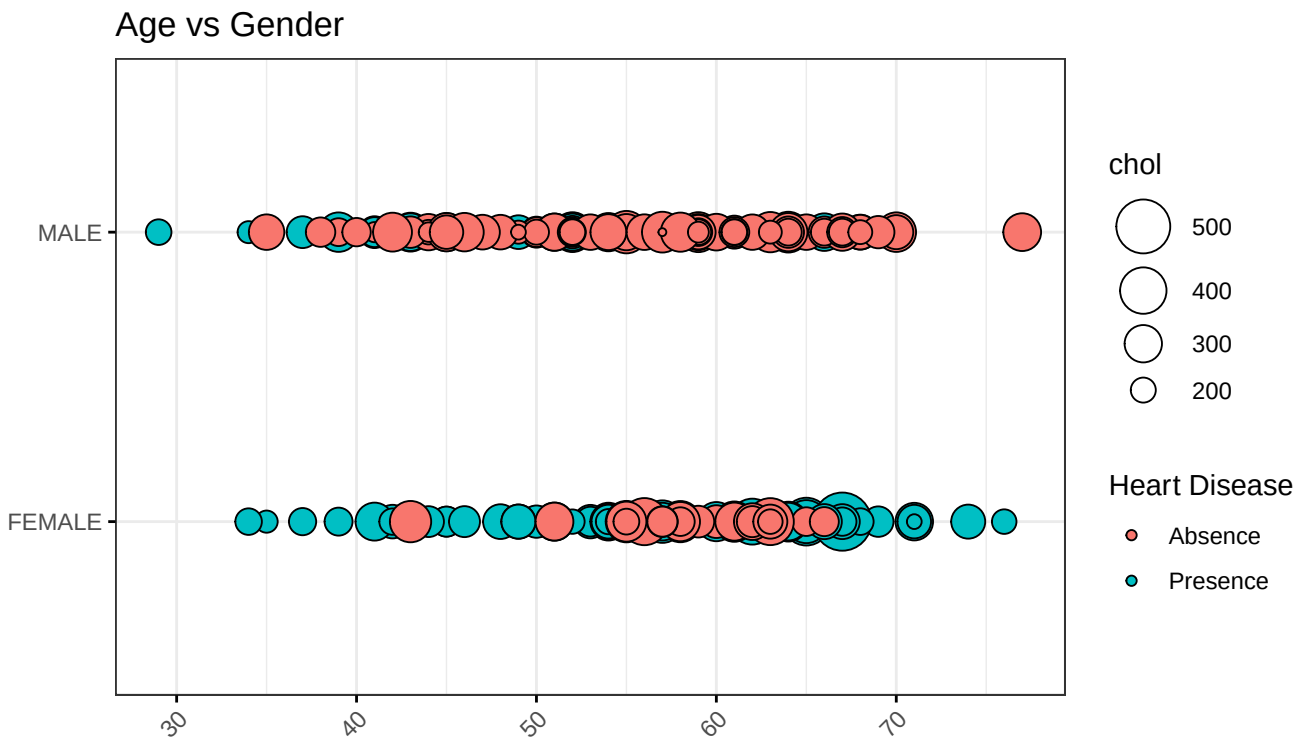
Age Analysis for various chest pain conditions :-

```
age_cp <- Heartrate %>%
  filter(target == "YES") %>%
  group_by(age) %>%
  count(cp)

  ggplot(data = Heartrate, mapping = aes(x= age, fill = cp)) +
     geom_bar() +
     labs(x = "Age",
          y = "Count",
          title = "Age Analysis for various chest pain conditions") +
     scale_fill_discrete(name = "Chest pain type") +
     theme_bw()
```



Age vs Gender :-

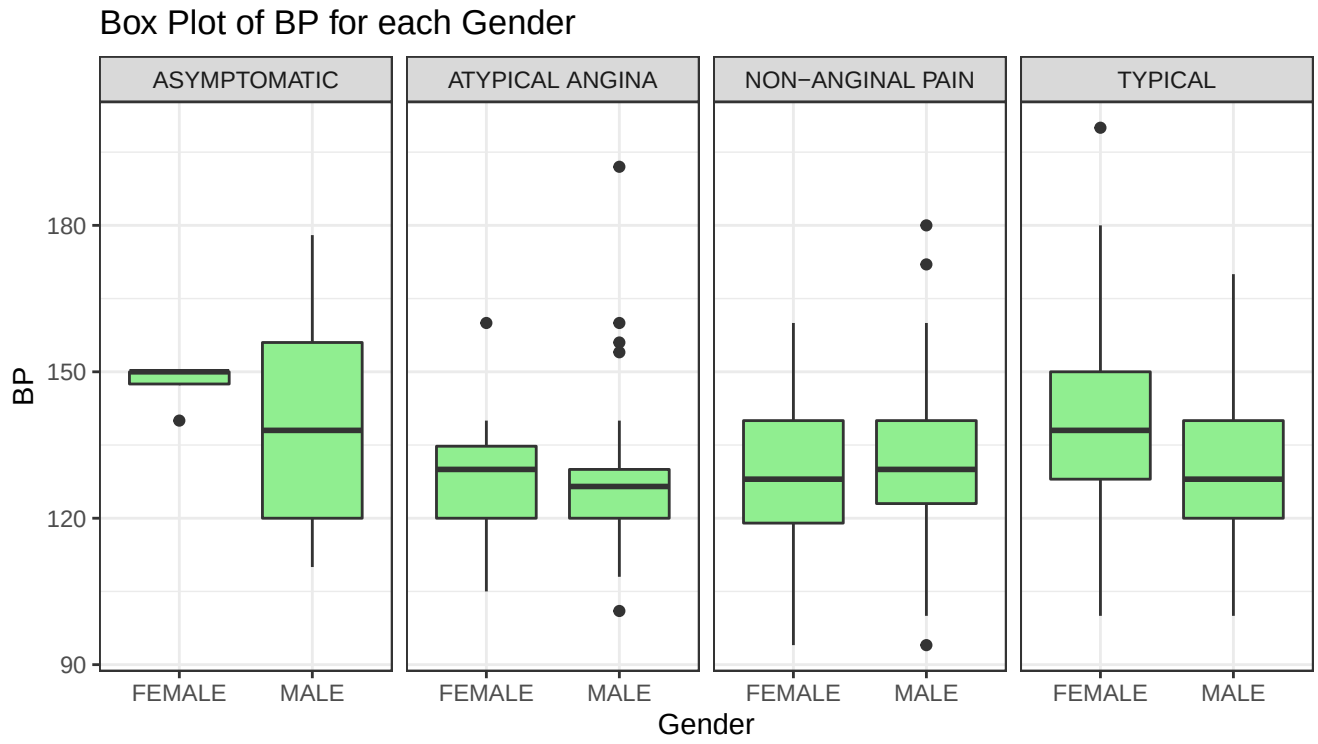
```
options(repr.plot.width = 20, repr.plot.height = 8)

Heartrate %>%
  ggballoonplot(x = "age", y = "sex",size = "chol", size.range = c(0, 10),
                fill = "target", show.label = FALSE, ggtheme = theme_bw()) +
              scale_fill_viridis_d(option = "E") +
              scale_fill_discrete(name= 'Heart Disease', labels =c("Absence", "Presence")) +
              labs(title = "Age vs Gender")
```



Detecting the outliers :-

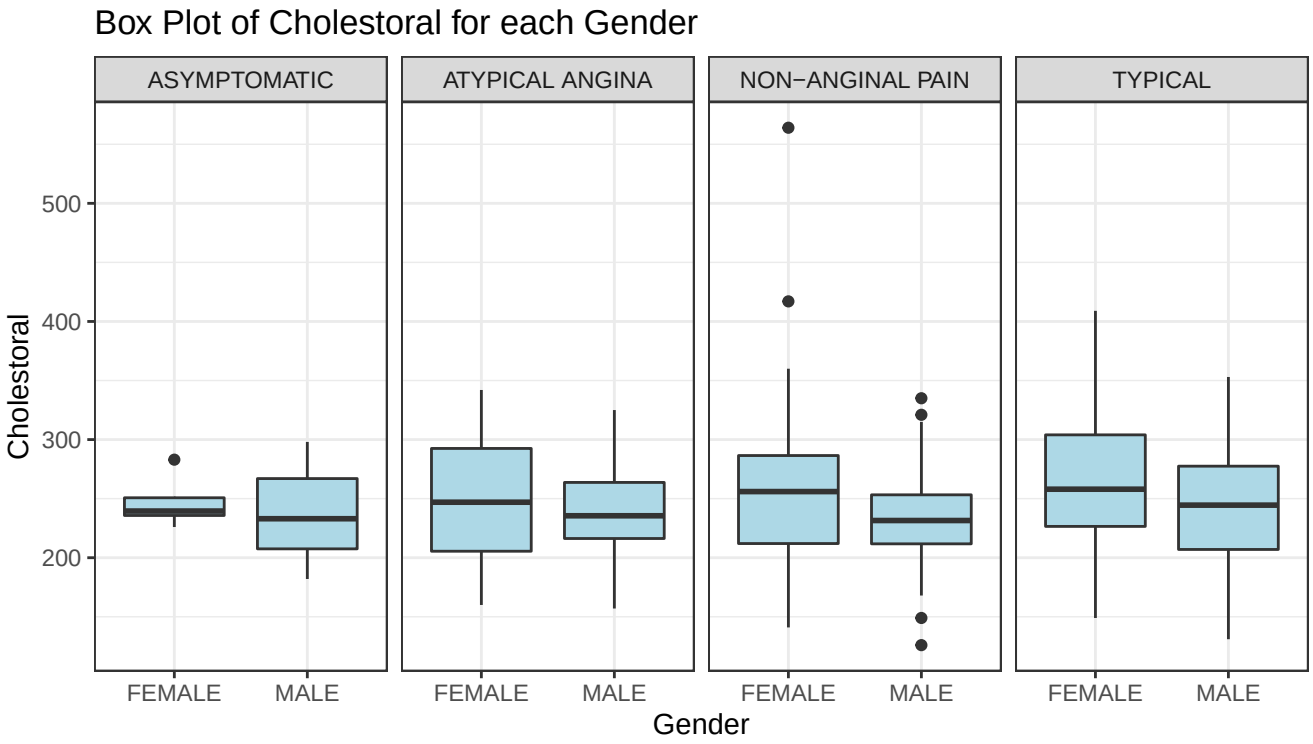
Box Plot of BP for each Gender :-

```
ggplot(data = Heartrate, mapping  = aes(x = sex, y = trestbps)) +
    geom_boxplot(fill= 'lightgreen') +
    facet_grid(~ cp) +
    labs(x = "Gender",
         y = "BP",
         title = "Box Plot of BP for each Gender") +
    theme_bw()
```



Box Plot of Cholestoral for each Gender :-

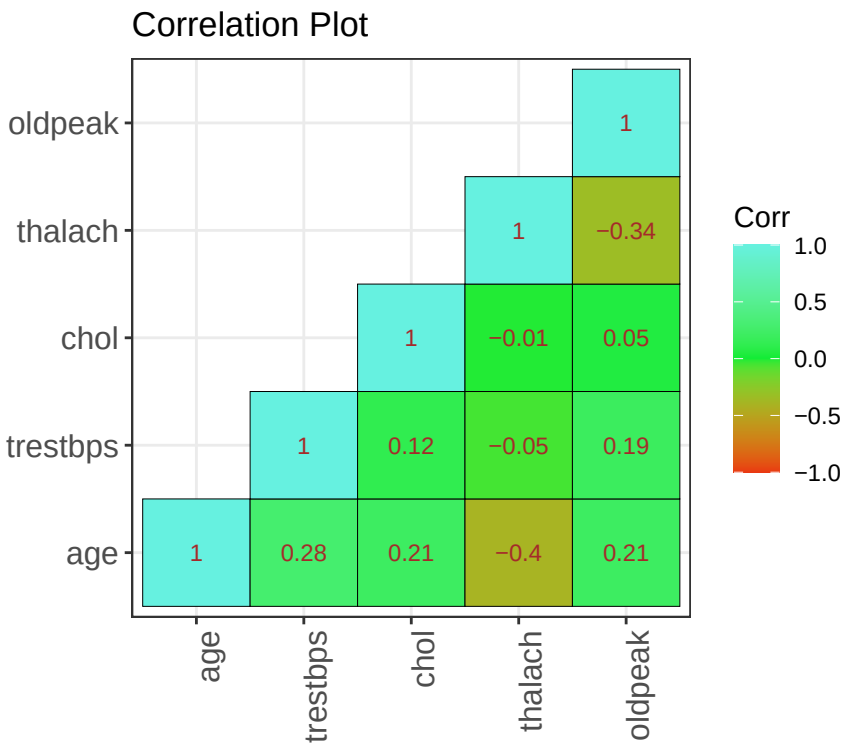
```
ggplot(data = Heartrate, mapping  = aes(x = sex, y = chol)) +
    geom_boxplot(fill= 'lightblue') +
    facet_grid(~ cp) +
    labs(x = "Gender",
         y = "Cholestoral",
         title = "Box Plot of Cholestoral for each Gender") +
    theme_bw()
```



Correlation Plot :-

```
cor.heart <- cor(Heartrate[,10:14])
cor.heart

ggcorrplot(corr.heart, method = "square", type = "lower", ggtheme = ggplot2::theme_bw,   t
  hc.order = FALSE, hc.method = "complete", lab = TRUE,
  lab_col = "brown", lab_size = 3, tl.cex = 12, tl.col = "black", tl.srt = 90,
  digits = 2)
```

```
                age    trestbps         chol      thalach     oldpeak
age       1.0000000  0.27935091  0.213677957 -0.398521938  0.21001257
trestbps  0.2793509  1.00000000  0.123174207 -0.046697728  0.19321647
chol      0.2136780  0.12317421  1.000000000 -0.009939839  0.05395192
thalach  -0.3985219 -0.04669773 -0.009939839  1.000000000 -0.34418695
oldpeak   0.2100126  0.19321647  0.053951920 -0.344186948  1.00000000
```



Training Data & Test Data :-

```
a=sample(c(TRUE,FALSE),nrow(Heartrate),replace=T,prob=c(0.7,0.3))
trainData=Heartrate[a==TRUE,]
testData=Heartrate[a==FALSE,]
```

i) Logit Model :-

```
logRegModel <- glm(target ~ ., data=trainData, family = 'binomial')
summary(logRegModel)
```

```

Call:
glm(formula = target ~ ., family = "binomial", data = trainData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.83553  -0.12412   0.07565   0.39347   2.09850

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.750e+00  4.650e+00   0.376  0.70672
sexMALE                   -3.063e+00  9.322e-01  -3.285  0.00102 **
fbs>120                    1.108e+00  8.234e-01   1.346  0.17840
exangYES                  -6.918e-01  7.734e-01  -0.895  0.37103
```

```
cpATYPICAL ANGINA           -3.219e+00  1.285e+00  -2.504  0.01226 *
cpNON-ANGINAL PAIN          -1.810e+00  1.142e+00  -1.585  0.11292
cpTYPICAL                   -3.686e+00  1.205e+00  -3.060  0.00222 **
restecgNORMAL               -5.043e-02  5.801e-01  -0.087  0.93072
restecgPROBABLE OR DEFINITE  9.319e-01  3.390e+00   0.275  0.78339
slope1                      -1.274e+00  1.196e+00  -1.066  0.28659
slope2                       5.583e-01  1.216e+00   0.459  0.64616
ca1                         -2.402e+00  8.613e-01  -2.788  0.00530 **
ca2                         -4.905e+00  1.192e+00  -4.114  3.9e-05 ***
ca3                         -9.097e-01  1.614e+00  -0.564  0.57306
ca4                          1.235e+01  1.693e+03   0.007  0.99418
thal1                        4.293e+00  2.442e+00   1.758  0.07868 .
thal2                        2.107e+00  2.203e+00   0.957  0.33879
thal3                        1.598e-01  2.216e+00   0.072  0.94253
age                          1.464e-02  3.950e-02   0.371  0.71088
trestbps                    -4.354e-02  1.641e-02  -2.652  0.00799 **
chol                        -6.714e-03  5.340e-03  -1.257  0.20863
thalach                      6.648e-02  2.414e-02   2.754  0.00589 **
oldpeak                     -1.774e-01  3.156e-01  -0.562  0.57404
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 253.383  on 186  degrees of freedom
Residual deviance:  93.395  on 164  degrees of freedom
AIC: 139.39

Number of Fisher Scoring iterations: 15
```

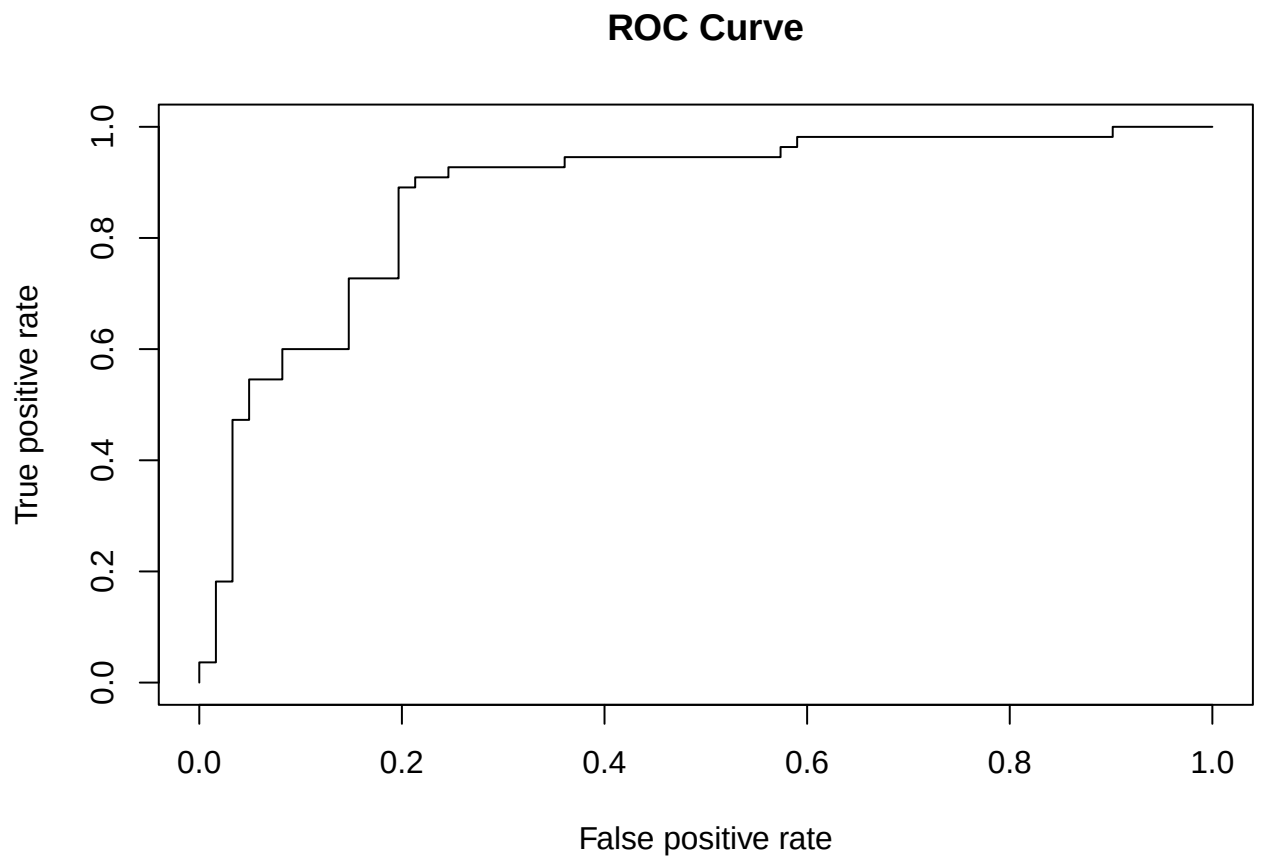
Accuracy of the Model :-

```
logRegPrediction <- predict(logRegModel, testData,type="response")
class_pred=if_else(logRegPrediction>0.5, "YES", "NO")
logRegConfMat <- confusionMatrix(as.factor(class_pred),testData[,"target"])
logRegConfMat$overall['Accuracy']

 Accuracy
0.7931034
```

ROC Curve :-

```
options=-1
library("ROCR")
pred <- predict(logRegModel, newdata = testData, type = "response")
```

```
pred_y <- prediction(pred, testData$target)
per = performance(pred_y, "tpr", "fpr")
plot(per, main="ROC Curve")

library("pROC")
glm_res = predict(logRegModel, testData, type="response")
AUC = roc(response = testData$target, predictor = glm_res)
AUC


Call:
roc.default(response = testData$target, predictor = glm_res)

Data: glm_res in 61 controls (testData$target NO) < 55 cases (testData$target YES).
Area under the curve: 0.8766
```



ii) Probit Model:

```
probit_model <- glm(target ~ ., data = trainData, family = binomial(link="probit"))
summary(probit_model)
```

```
Call:
glm(formula = target ~ ., family = binomial(link = "probit"),
    data = trainData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.83834  -0.08144   0.02572   0.39383   2.01221

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.864339   2.521507   0.343 0.731759
sexMALE                     -1.776778   0.498460  -3.565 0.000365 ***
fbs>120                      0.674448   0.453897   1.486 0.137304
exangYES                    -0.456252   0.430202  -1.061 0.288893
cpATYPICAL ANGINA           -1.780848   0.707524  -2.517 0.011835 *
cpNON-ANGINAL PAIN          -1.092501   0.627231  -1.742 0.081546 .
cpTYPICAL                   -2.033242   0.655893  -3.100 0.001935 **
restecgNORMAL               -0.010520   0.322901  -0.033 0.974011
restecgPROBABLE OR DEFINITE  0.585022   1.579840   0.370 0.711155
slope1                      -0.640265   0.662271  -0.967 0.333657
slope2                       0.363434   0.687501   0.529 0.597062
ca1                         -1.329598   0.467350  -2.845 0.004441 **
ca2                         -2.755678   0.639812  -4.307 1.65e-05 ***
ca3                         -0.647248   0.922042  -0.702 0.482697
ca4                          2.970272 264.870410   0.011 0.991053
thal1                        2.409370   1.287656   1.871 0.061327 .
thal2                        1.132882   1.137257   0.996 0.319176
thal3                        0.083029   1.149954   0.072 0.942441
age                          0.007355   0.021851   0.337 0.736418
trestbps                    -0.024184   0.009147  -2.644 0.008197 **
chol                        -0.004066   0.002935  -1.386 0.165890
thalach                      0.038580   0.013022   2.963 0.003050 **
oldpeak                     -0.085585   0.178947  -0.478 0.632458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 253.38  on 186  degrees of freedom
Residual deviance:  93.40  on 164  degrees of freedom
AIC: 139.4

Number of Fisher Scoring iterations: 14
```

Accuracy of the model:

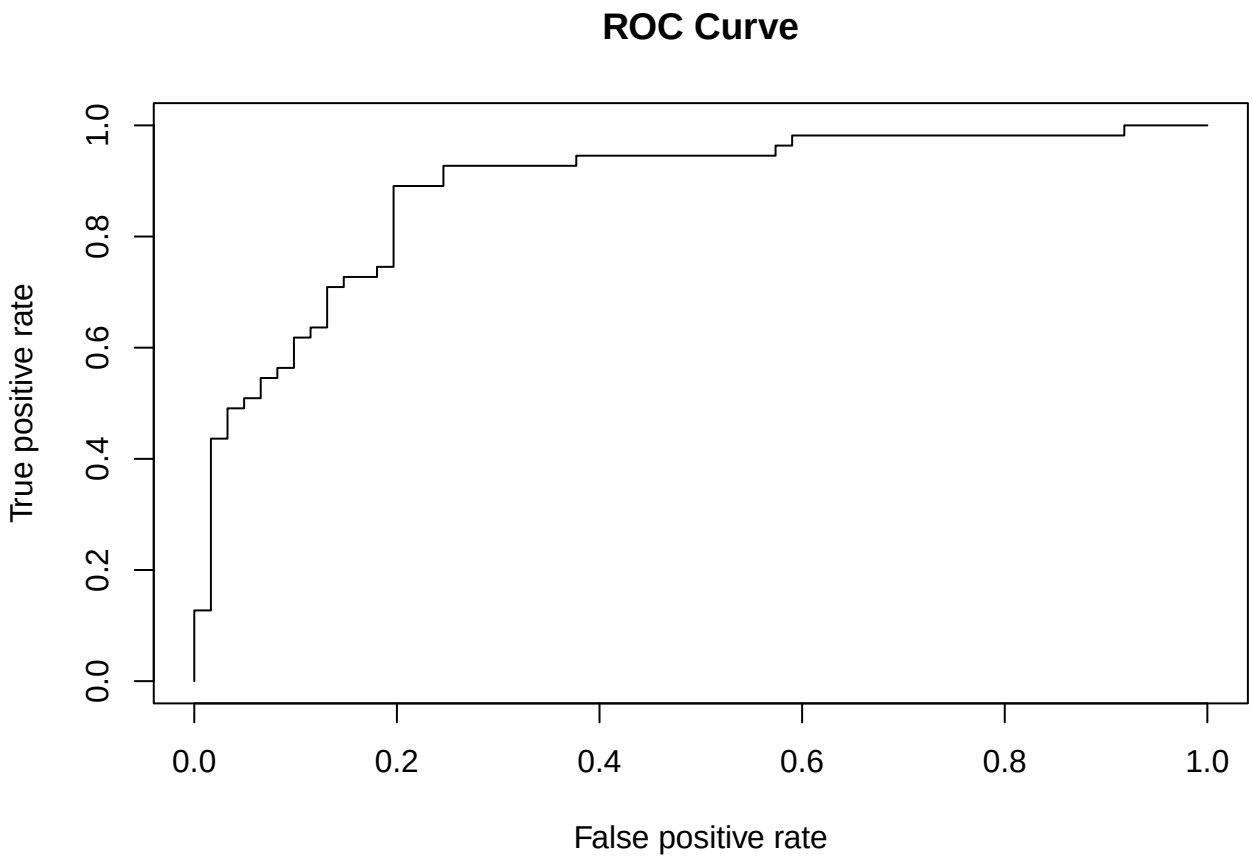
```
pred <- predict(probit_model, newdata = testData, type = "response")
pred_ty <- if_else(pred > 0.5, 1, 0)
test_target <- if_else(testData[,"target"] == "YES", 1, 0)
ConfMat <- confusionMatrix(as.factor(pred_ty), as.factor(test_target))
ConfMat$overall['Accuracy']
```

```
 Accuracy
0.7844828
```

ROC Curve :-

```
options=-1
library("ROCR")
pred <- predict(probit_model, newdata = testData, type = "response")
pred_y <- prediction(pred, testData$target)
per = performance(pred_y, "tpr", "fpr")
plot(per, main="ROC Curve")

library("pROC")
glm_res = predict(probit_model, testData, type="response")
AUC = roc(response = testData$target, predictor = glm_res)
AUC
```

```


Call:
roc.default(response = testData$target, predictor = glm_res)

Data: glm_res in 61 controls (testData$target NO) < 55 cases (testData$target YES).
Area under the curve: 0.8832
```



iii) cloglog Model:

```
cloglog_model <- glm(target ~ ., data = trainData, family = binomial(link="cloglog"))
summary(cloglog_model)


Call:
glm(formula = target ~ ., family = binomial(link = "cloglog"),
    data = trainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8436  -0.2493   0.0000   0.4074   1.9082

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.080e-01  2.482e+00   0.124 0.901247
sexMALE                   -2.101e+00  5.228e-01  -4.018 5.86e-05 ***
fbs>120                    8.576e-01  4.844e-01   1.770 0.076646 .
exangYES                  -5.755e-01  4.635e-01  -1.242 0.214340
cpATYPICAL ANGINA         -1.896e+00  7.345e-01  -2.581 0.009842 **
```

```
cpNON-ANGINAL PAIN          -1.323e+00  6.602e-01  -2.004 0.045043 *
cpTYPICAL                   -2.107e+00  6.908e-01  -3.050 0.002286 **
restecgNORMAL                5.615e-02  3.296e-01   0.170 0.864714
restecgPROBABLE OR DEFINITE  8.264e-01  1.404e+00   0.589 0.556090
slope1                      -5.782e-01  7.306e-01  -0.791 0.428683
slope2                       5.481e-01  7.523e-01   0.729 0.466271
ca1                         -1.547e+00  5.025e-01  -3.078 0.002083 **
ca2                         -2.933e+00  6.729e-01  -4.359 1.30e-05 ***
ca3                         -8.079e-01  1.089e+00  -0.742 0.458178
ca4                          1.761e+00  1.807e+04   0.000 0.999922
thal1                        2.555e+00  1.378e+00   1.854 0.063794 .
thal2                        8.642e-01  1.179e+00   0.733 0.463704
thal3                       -1.561e-01  1.218e+00  -0.128 0.898023
age                          3.411e-04  2.165e-02   0.016 0.987434
trestbps                    -2.267e-02  9.779e-03  -2.318 0.020425 *
chol                        -4.308e-03  2.866e-03  -1.503 0.132771
thalach                      4.329e-02  1.284e-02   3.371 0.000748 ***
oldpeak                     -4.093e-02  1.959e-01  -0.209 0.834491
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 253.383  on 186  degrees of freedom
Residual deviance:  97.117  on 164  degrees of freedom
AIC: 143.12

Number of Fisher Scoring iterations: 25
```

Accuracy of the model:

```
pred <- predict(cloglog_model, newdata = testData, type = "response")
pred_ty <- if_else(pred > 0.5, 1, 0)
test_target <- if_else(testData[,"target"] == "YES", 1, 0)
ConfMat <- confusionMatrix(as.factor(pred_ty), as.factor(test_target))
ConfMat$overall['Accuracy']

 Accuracy
0.7931034
```

ROC Curve :-

```
options=-1
library("ROCR")
pred <- predict(cloglog_model, newdata = testData, type = "response")
```

```r
pred_y <- prediction(pred, testData$target)
per = performance(pred_y, "tpr", "fpr")
plot(per, main="ROC Curve")

library("pROC")
glm_res = predict(cloglog_model, testData, type="response")
AUC = roc(response = testData$target, predictor = glm_res)
AUC
```

```
Call:
roc.default(response = testData$target, predictor = glm_res)

Data: glm_res in 61 controls (testData$target NO) < 55 cases (testData$target YES).
Area under the curve: 0.8793
```