Importing the Libraries :-

```
library(dplyr)
library(tidyverse)
library(tidyr)
library(skimr)
library(ggplot2)
library(corrplot)
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, 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, 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:

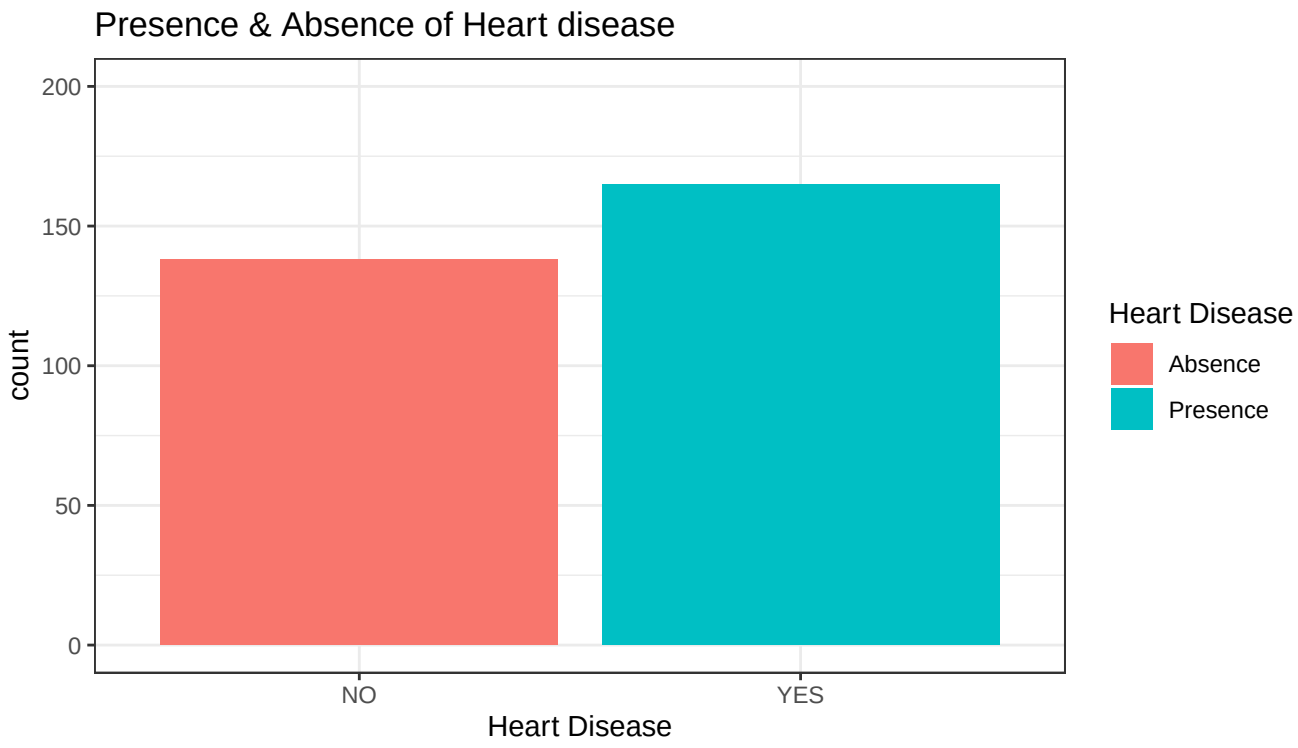
```
 target        sex         fbs      exang                     cp
 NO :138   FEMALE: 96   <=120:258   NO :204   ASYMPTOMATIC    : 23
 YES:165   MALE  :207   >120 : 45   YES: 99   ATYPICAL ANGINA : 50
                                              NON-ANGINAL PAIN: 87
                              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



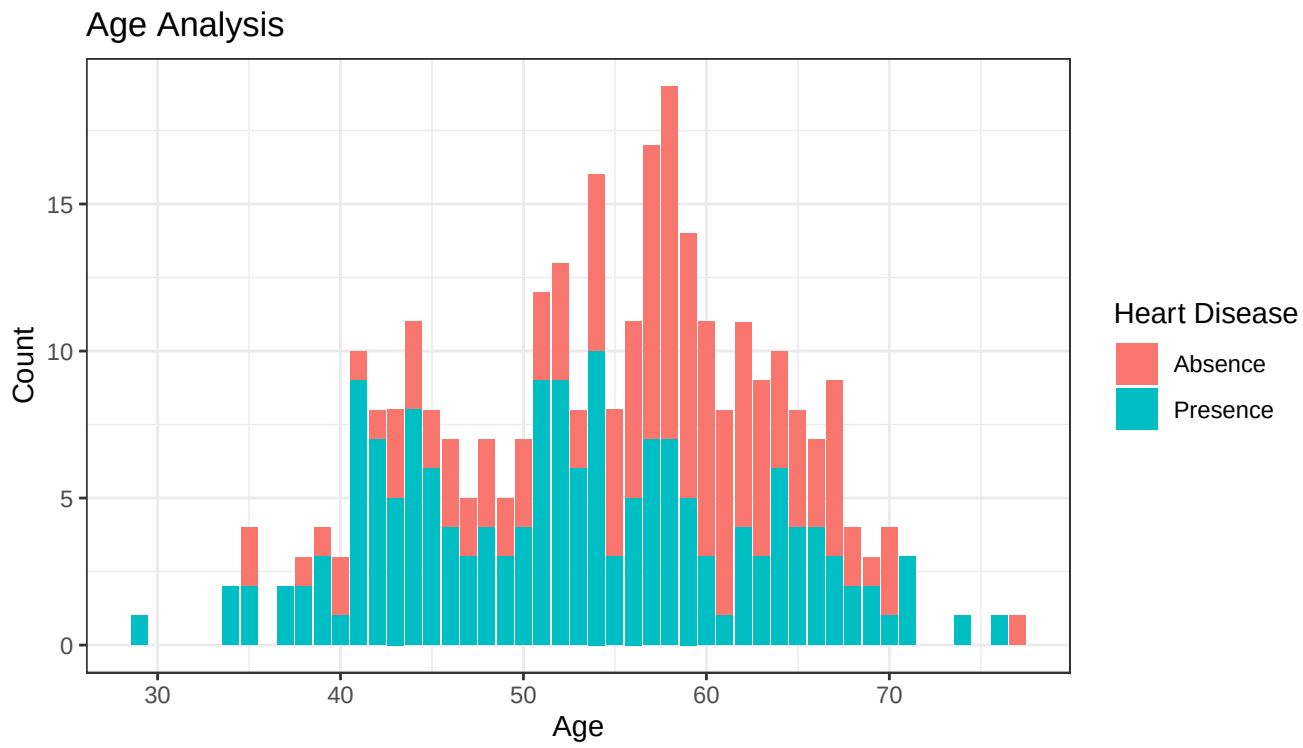
Proportion :-

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

```
   NO   YES
0.455 0.545
```

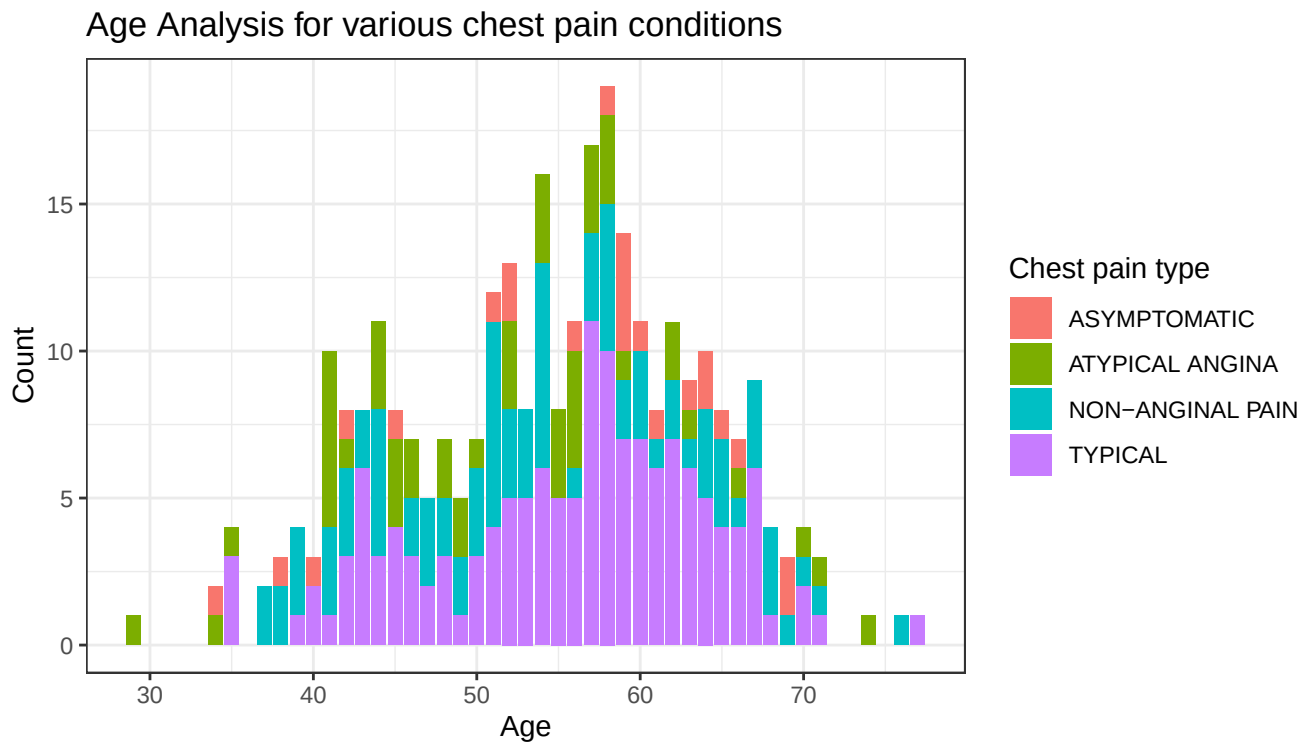
Age Analysis :-

```
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()
```



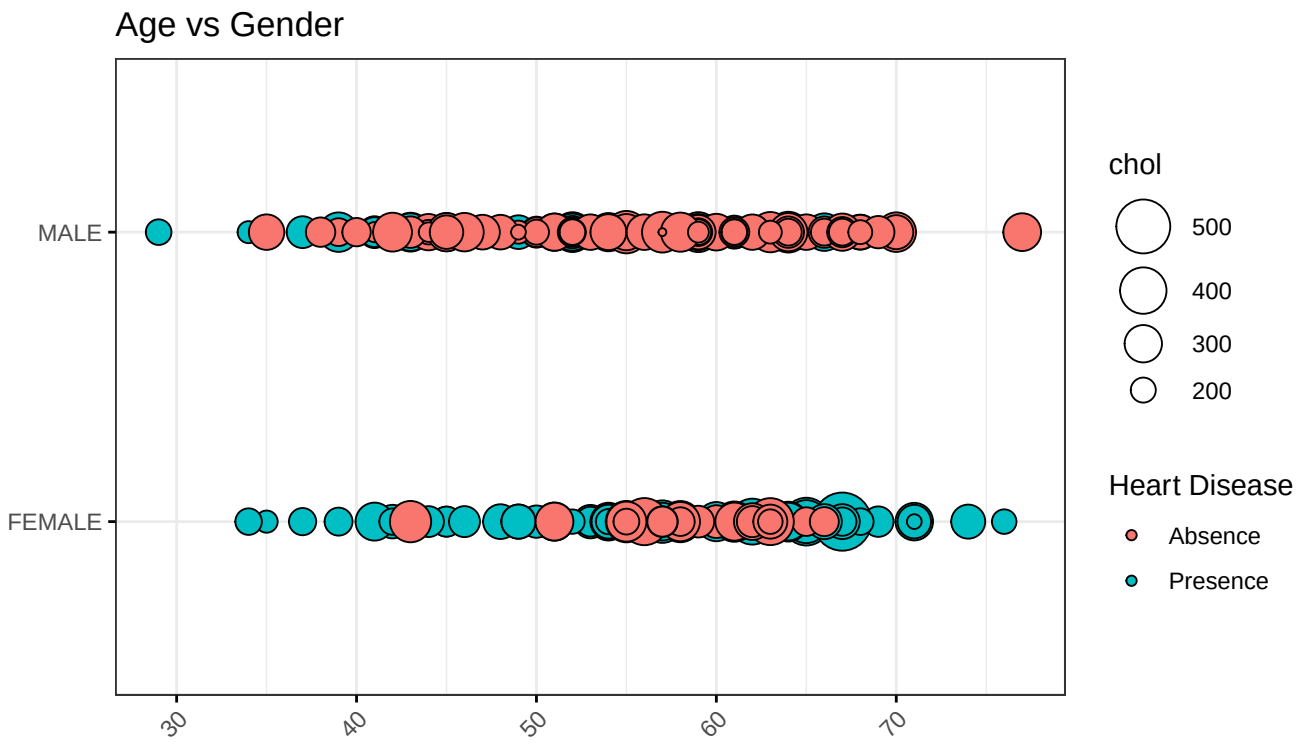
Age Analysis for various chest pain conditions :-

```
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 :-

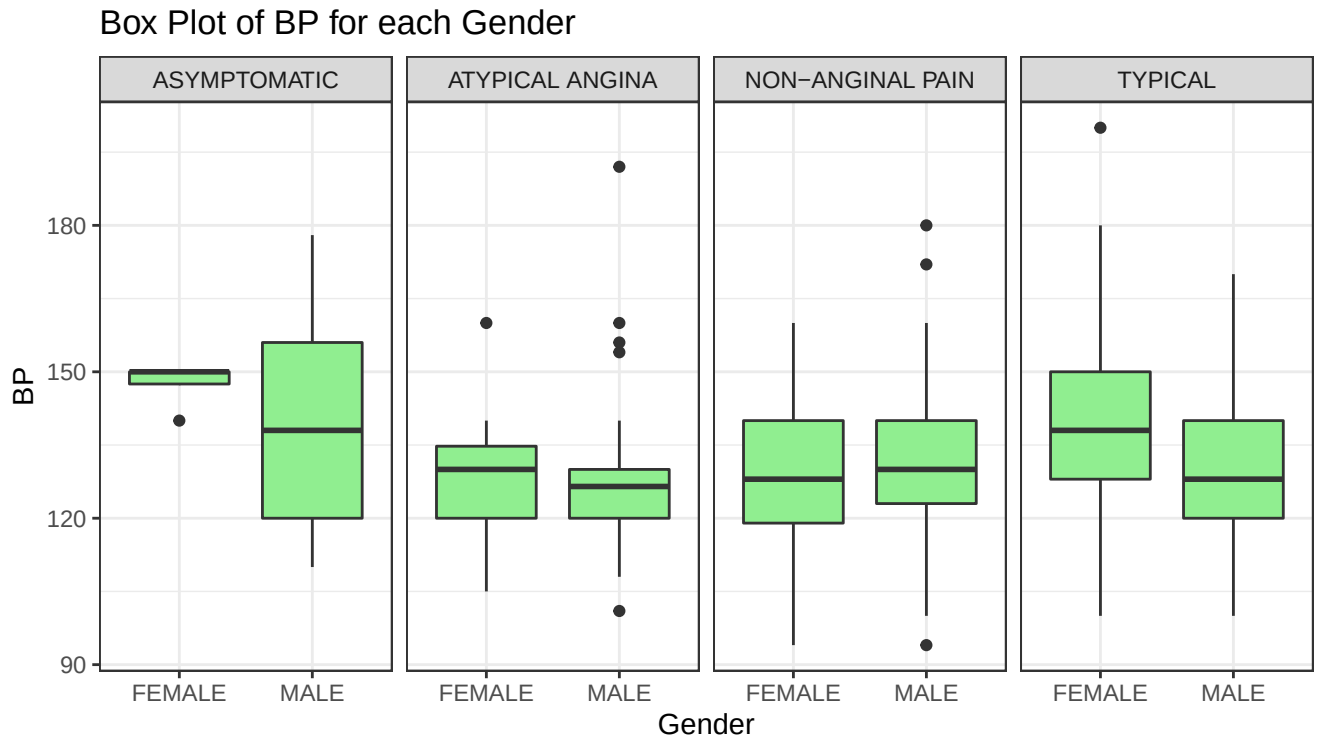
```
options(repr.plot.width = 20, repr.plot.height = 8)

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 :-

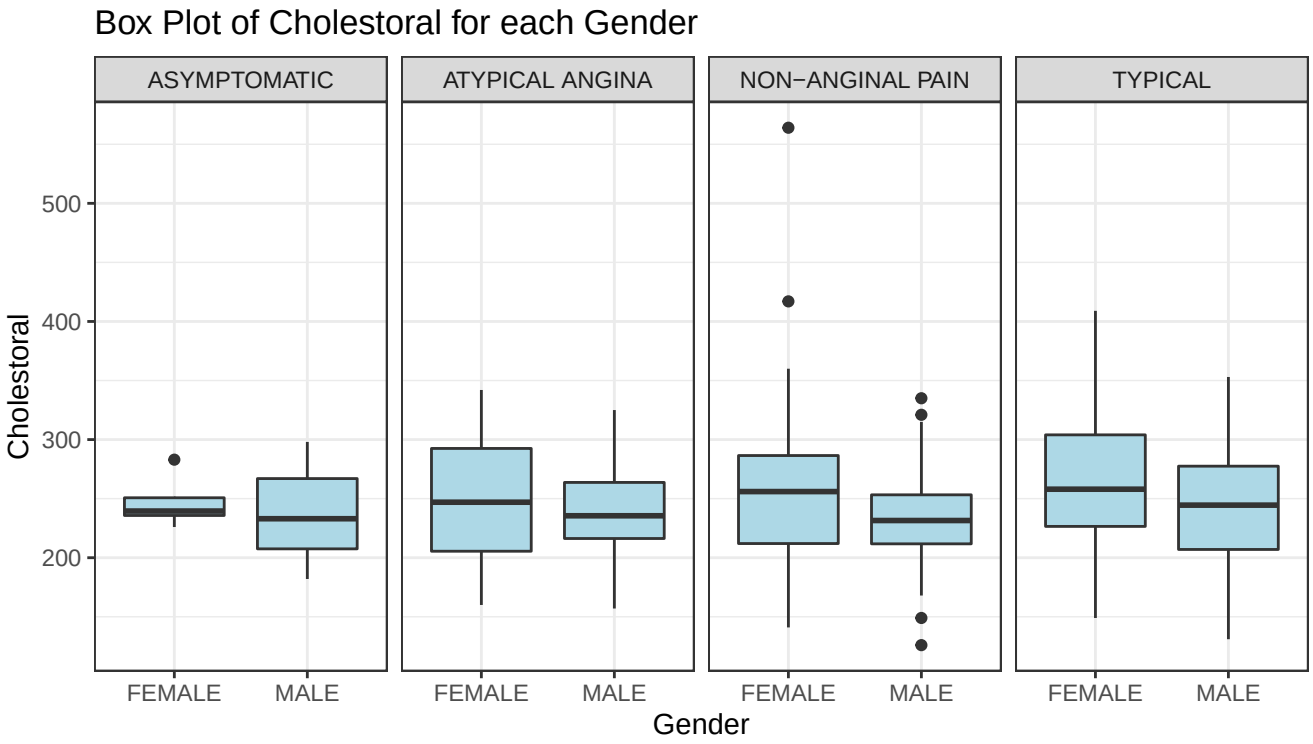
Box Plot of BP for each Gender :-

```
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 :-

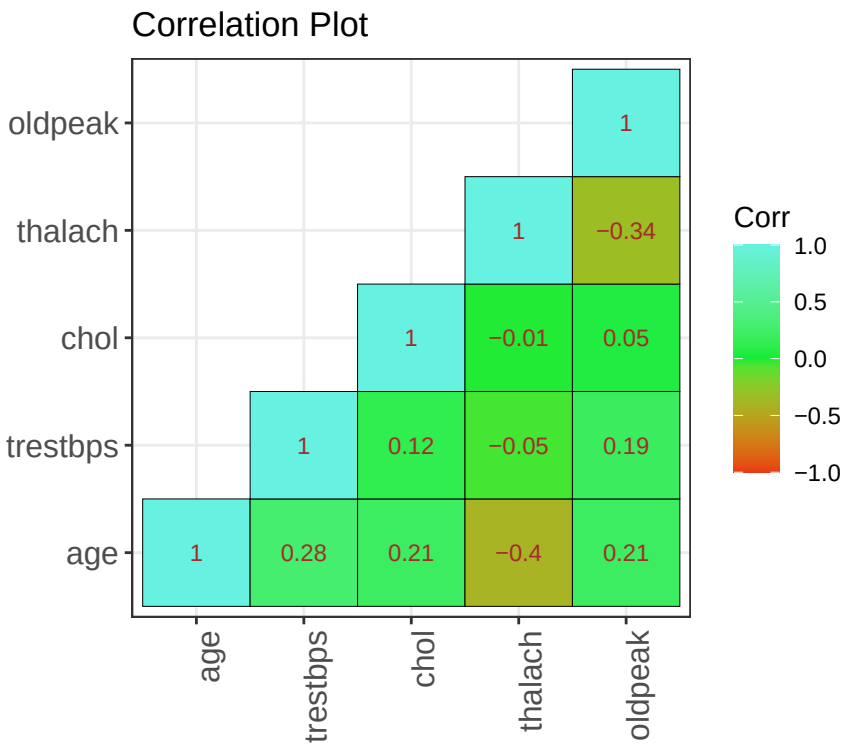
```
ggplot(data = Heartrate, mapping  = aes(x = sex, y = chol)) +
    geom_boxplot(fill= 'lightblue') +
    facet_grid(~ cp) +
    labs(x = "Gender",
         y = "Cholestoral",
         title = "Box Plot of Cholestoral for each Gender") +
    theme_bw()
```



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
-3.0430  -0.2727   0.0755   0.4773   3.2282

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.764e+01  3.956e+03   0.004 0.996442
sexMALE                   -1.768e+00  6.930e-01  -2.551 0.010730 *
fbs>120                    3.231e-01  7.381e-01   0.438 0.661538
exangYES                  -5.683e-01  5.629e-01  -1.010 0.312697
```

```
cpATYPICAL ANGINA           -2.377e+00  9.720e-01  -2.446 0.014449 *
cpNON-ANGINAL PAIN          -4.453e-01  8.322e-01  -0.535 0.592607
cpTYPICAL                   -2.905e+00  8.825e-01  -3.291 0.000997 ***
restecgNORMAL               -8.807e-01  4.835e-01  -1.821 0.068539 .
restecgPROBABLE OR DEFINITE -1.513e+01  2.751e+03  -0.005 0.995614
slope1                      -1.173e+00  1.340e+00  -0.876 0.381234
slope2                       2.596e-01  1.448e+00   0.179 0.857736
ca1                         -2.236e+00  6.078e-01  -3.679 0.000234 ***
ca2                         -3.294e+00  9.561e-01  -3.446 0.000570 ***
ca3                         -2.712e+00  1.170e+00  -2.318 0.020426 *
ca4                          1.663e+01  1.784e+03   0.009 0.992565
thal1                       -1.275e+01  3.956e+03  -0.003 0.997429
thal2                       -1.343e+01  3.956e+03  -0.003 0.997292
thal3                       -1.498e+01  3.956e+03  -0.004 0.996979
age                          3.821e-02  3.084e-02   1.239 0.215245
trestbps                    -2.805e-02  1.408e-02  -1.992 0.046415 *
chol                        -5.022e-03  4.944e-03  -1.016 0.309750
thalach                      3.163e-02  1.408e-02   2.246 0.024683 *
oldpeak                     -3.699e-01  3.224e-01  -1.148 0.251149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296.37  on 214  degrees of freedom
Residual deviance: 127.44  on 192  degrees of freedom
AIC: 173.44

Number of Fisher Scoring iterations: 16
```

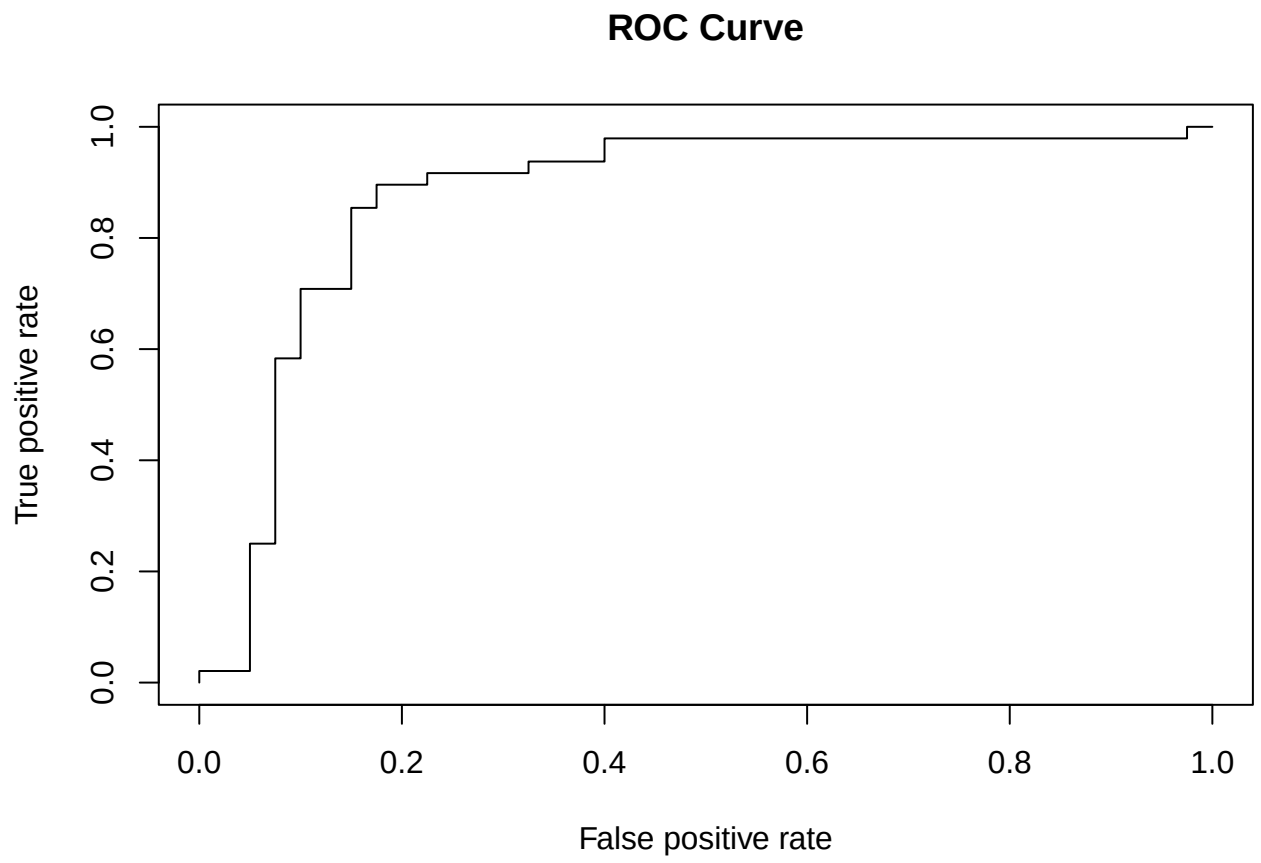
Accuracy of the Model :-

```
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
0.8636364
```

ROC Curve :-

```
options=-1
library("ROCR")
pred <- predict(logRegModel, 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(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 40 controls (testData$target NO) < 48 cases (testData$target YES).
Area under the curve: 0.8734
```



ii) Probit Model:

```r
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
-3.0834  -0.3279   0.0455   0.5037   3.2159

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  5.545779 605.104472   0.009 0.992687
sexMALE                     -0.904080   0.358828  -2.520 0.011751 *
fbs>120                      0.191943   0.382923   0.501 0.616189
exangYES                    -0.256731   0.310036  -0.828 0.407631
cpATYPICAL ANGINA           -1.122825   0.523773  -2.144 0.032055 *
cpNON-ANGINAL PAIN          -0.251480   0.461172  -0.545 0.585543
cpTYPICAL                   -1.518287   0.467170  -3.250 0.001154 **
restecgNORMAL               -0.529695   0.263450  -2.011 0.044367 *
restecgPROBABLE OR DEFINITE -4.105002 424.742111  -0.010 0.992289
slope1                      -0.555449   0.671567  -0.827 0.408184
slope2                       0.095189   0.733541   0.130 0.896752
ca1                         -1.085757   0.323167  -3.360 0.000780 ***
ca2                         -1.714677   0.503589  -3.405 0.000662 ***
ca3                         -1.429167   0.599225  -2.385 0.017078 *
ca4                          5.229608 270.532749   0.019 0.984577
thal1                       -3.109729 605.102791  -0.005 0.995900
thal2                       -3.349344 605.102371  -0.006 0.995584
thal3                       -4.183826 605.102461  -0.007 0.994483
age                          0.020033   0.016730   1.197 0.231138
trestbps                    -0.014474   0.007655  -1.891 0.058662 .
chol                        -0.002648   0.002675  -0.990 0.322359
thalach                      0.016517   0.007498   2.203 0.027616 *
oldpeak                     -0.202328   0.170629  -1.186 0.235709
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296.37  on 214  degrees of freedom
Residual deviance: 130.66  on 192  degrees of freedom
AIC: 176.66

Number of Fisher Scoring iterations: 15
```

Accuracy of the model:

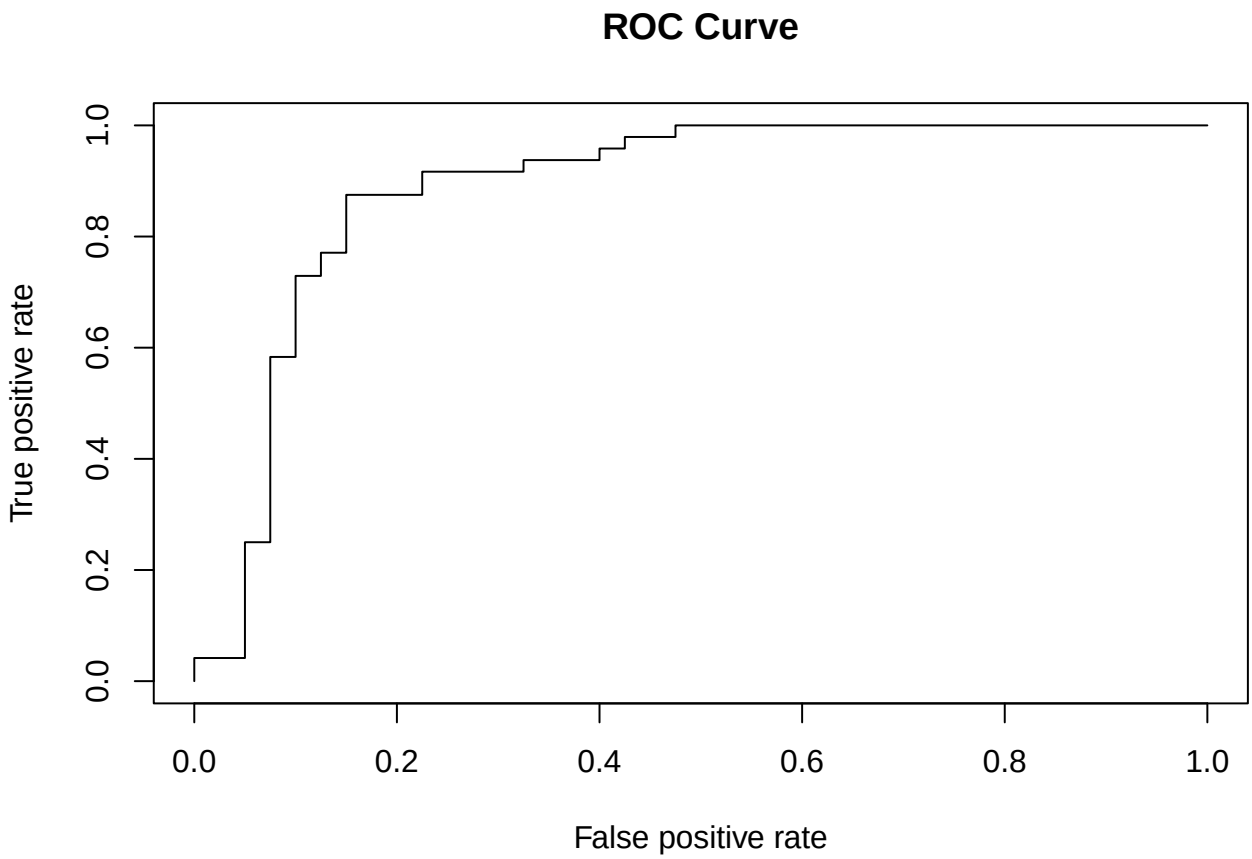
```
pred <- predict(probit_model, newdata = testData, type = "response")
pred_ty <- if_else(pred > 0.5, 1, 0)
test_target <- if_else(testData[,"target"] == "YES", 1, 0)
ConfMat <- confusionMatrix(as.factor(pred_ty), as.factor(test_target))
ConfMat$overall['Accuracy']
```

```
 Accuracy
0.8522727
```

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 40 controls (testData$target NO) < 48 cases (testData$target YES).
Area under the curve: 0.8859
```



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
-3.2752  -0.4485   0.0003   0.4612   2.7417

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               2.561e+00  2.750e+04   0.000 0.999926
sexMALE                  -1.157e+00  3.755e-01  -3.082 0.002054 **
fbs>120                   4.063e-01  3.931e-01   1.034 0.301367
exangYES                 -3.444e-01  3.457e-01  -0.996 0.319162
cpATYPICAL ANGINA        -1.008e+00  5.261e-01  -1.916 0.055336 .
```

```
cpNON-ANGINAL PAIN          -3.467e-01  4.648e-01  -0.746 0.455678
cpTYPICAL                   -1.620e+00  4.889e-01  -3.313 0.000923 ***
restecgNORMAL               -4.639e-01  2.792e-01  -1.662 0.096584 .
restecgPROBABLE OR DEFINITE -2.450e+01  2.396e+05   0.000 0.999918
slope1                      -3.180e-01  6.238e-01  -0.510 0.610155
slope2                       4.396e-01  6.633e-01   0.663 0.507540
ca1                         -1.153e+00  3.516e-01  -3.279 0.001042 **
ca2                         -1.857e+00  5.405e-01  -3.435 0.000592 ***
ca3                         -1.557e+00  7.575e-01  -2.055 0.039833 *
ca4                          3.927e+00  1.115e+04   0.000 0.999719
thal1                       -1.272e+00  2.750e+04   0.000 0.999963
thal2                       -1.858e+00  2.750e+04   0.000 0.999946
thal3                       -2.642e+00  2.750e+04   0.000 0.999923
age                          2.127e-02  1.801e-02   1.181 0.237637
trestbps                    -1.379e-02  8.120e-03  -1.698 0.089570 .
chol                        -4.125e-03  2.685e-03  -1.536 0.124494
thalach                      2.324e-02  8.583e-03   2.707 0.006781 **
oldpeak                     -1.243e-01  1.779e-01  -0.699 0.484654
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296.37  on 214  degrees of freedom
Residual deviance: 133.88  on 192  degrees of freedom
AIC: 179.88

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
0.8181818
```

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 40 controls (testData$target NO) < 48 cases (testData$target YES).
Area under the curve: 0.8891
```