Using logistic Regression Model to Predict Heart Disease Staglog

Kaiwen Fu

2023-04-10

Introduction

Heart disease, also known as cardiovascular disease, is a prevalent and serious health condition that affects the heart and blood vessels, encompassing various conditions such as coronary artery disease, heart attack, heart failure, and stroke. According to WHO(2021), cardiovascular diseases are the leading cause of death globally, accounting for an estimated 17.9 million deaths annually as of 2019. In order to avoid the heart disease, we need more researches and do some predictions.

Numerous risk factors contribute to the development of heart disease, including lifestyle choices, genetics, and medical conditions such as high blood pressure, high cholesterol, obesity, smoking, and diabetes, etc. These risk factors can increase an individual's likelihood of developing heart disease, highlighting the importance of understanding and managing them.

In this paper, we will explore the regression model to recognize which are the most important factors and using the factors we have to predict the likelihood of heart disease.

The rest of paper is organized as follows. The data will be presented in Section Data Description. The analysis will be shown in the Section Data Analysis. The conclusion will be obtained in the Section Conclusion and Discussion.

Data Description

```
hd<- read.csv("HD.csv")
hd$sex<-factor(hd$sex)
hd$cp<-factor(hd$cp,levels = c(0,1,2,3),c("typical angina",
                                          "atypical angina",
                                          "non-anginal pain","asymptomatic"))
hd$fbs <- factor(hd$fbs)
hd$restecg<-factor(hd$restecg)
hd$exang <- factor(hd$exang)
```

```
hd$slope <- factor(hd$slope)
hd$ca <- factor(hd$ca)
hd$thal <- factor(hd$thal)
hd$target<-factor(hd$target, levels =c(0,1), labels = c("No","Yes"))
# transfer all the catagorical variables into factors
str(hd)
```

```
## 'data.frame':    270 obs. of  14 variables:
##  $ age     : int  70 67 57 64 74 65 56 59 60 63 ...
##  $ sex     : Factor w/ 2 levels "0","1": 2 1 2 2 1 2 2 2 2 1 ...
##  $ cp      : Factor w/ 4 levels "typical angina",..: 4 3 2 4 2 4 3 4 4 4 ...
##  $ trestbps: int  130 115 124 128 120 120 130 110 140 150 ...
##  $ chol    : int  322 564 261 263 269 177 256 239 293 407 ...
##  $ fbs     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ restecg : Factor w/ 3 levels "0","1","2": 3 3 1 1 3 1 3 3 3 3 ...
##  $ thalach : int  109 160 141 105 121 140 142 142 170 154 ...
##  $ exang   : Factor w/ 2 levels "0","1": 1 1 1 2 2 1 2 2 1 1 ...
##  $ oldpeak : num  2.4 1.6 0.3 0.2 0.2 0.4 0.6 1.2 1.2 4 ...
##  $ slope   : Factor w/ 3 levels "0","1","2": 2 2 1 2 1 1 2 2 2 2 ...
##  $ ca      : Factor w/ 4 levels "0","1","2","3": 4 1 1 2 2 1 2 2 3 4 ...
##  $ thal    : Factor w/ 3 levels "1","2","3": 1 3 3 3 1 3 2 3 3 3 ...
##  $ target  : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 2 2 2 ...
```

```
summary(hd)
```

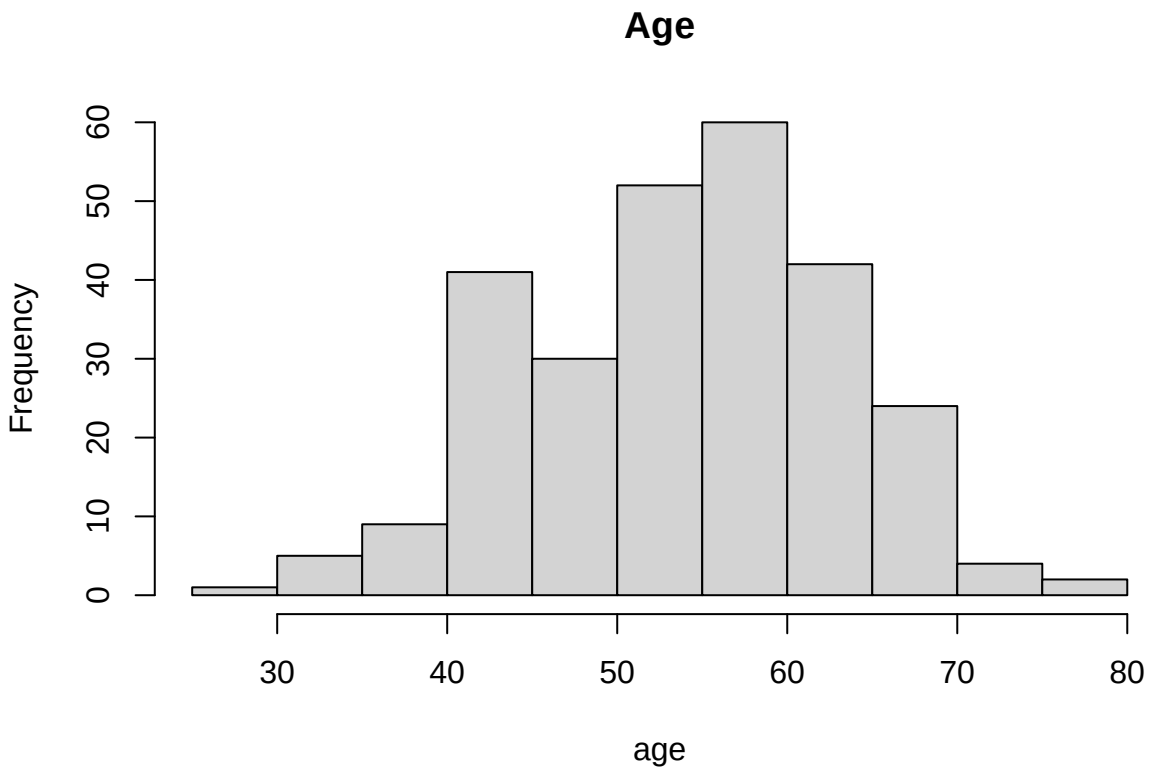
```
##       age        sex                     cp         trestbps          chol
##  Min.   :29.00   0: 87   typical angina  : 20   Min.   : 94.0   Min.   :126.0
##  1st Qu.:48.00   1:183   atypical angina : 42   1st Qu.:120.0   1st Qu.:213.0
##  Median :55.00           non-anginal pain: 79   Median :130.0   Median :245.0
##  Mean   :54.43           asymptomatic    :129   Mean   :131.3   Mean   :249.7
##  3rd Qu.:61.00                                  3rd Qu.:140.0   3rd Qu.:280.0
##  Max.   :77.00                                  Max.   :200.0   Max.   :564.0
##  fbs     restecg    thalach      exang      oldpeak     slope   ca      thal
##  0:230   0:131   Min.   : 71.0   0:181   Min.   :0.00   0:130   0:160   1:152
##  1: 40   1:  2   1st Qu.:133.0   1: 89   1st Qu.:0.00   1:122   1: 58   2: 14
##          2:137   Median :153.5           Median :0.80   2: 18   2: 33   3:104
##                  Mean   :149.7           Mean   :1.05           3: 19
##                  3rd Qu.:166.0           3rd Qu.:1.60
##                  Max.   :202.0           Max.   :6.20
##  target
##  No :150
##  Yes:120
##
```

```
##
##
##
```

This dataset was collected by Dua, D. and Graff, C. (2019) at University of California, Irvine. In this dataset, we total have 270 samples with 14 variables. The detailed explanation and introduction will be presented in the following.

Age is a continuous variables with mean 55.43 and median 55

```
hist(hd$age, main = "Age", xlab = "age")
```




```
# we don't need to consider whether it is normally distributed or not
# because it is not important in this paper.
```

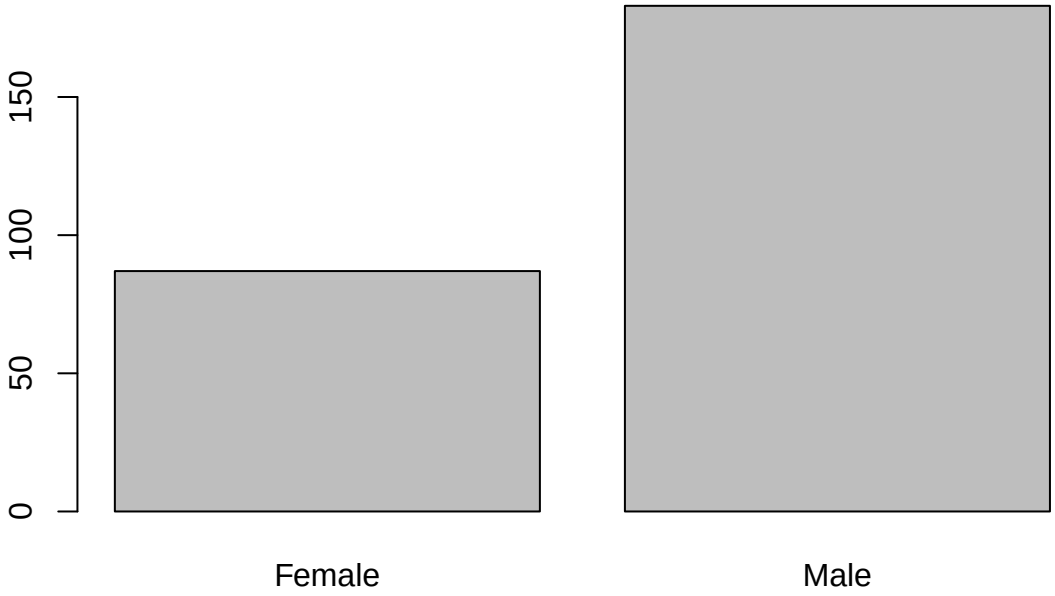
From the histogram, we can say that most of patients' age is at the range 40-70, and only few patients' age are less than 40 and greater than 70.

Sex is a Nominal variables where 1 represents Males and 0 represents Females

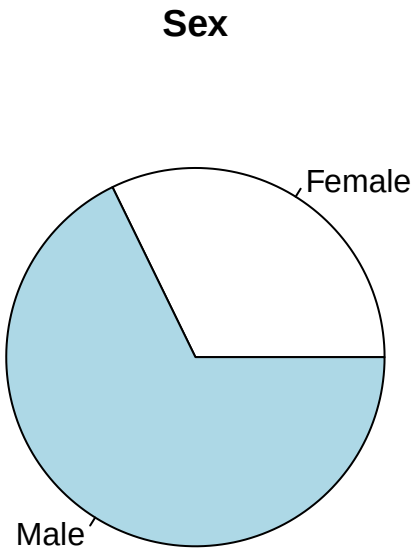
```
table(hd$sex)
```

```
##
##   0   1
##  87 183
```

```
barplot(table(hd$sex), names.arg = c("Female", "Male"))
```




```
pie(table(hd$sex), labels = c("Female", "Male"), main = "Sex", cex= 1)
```

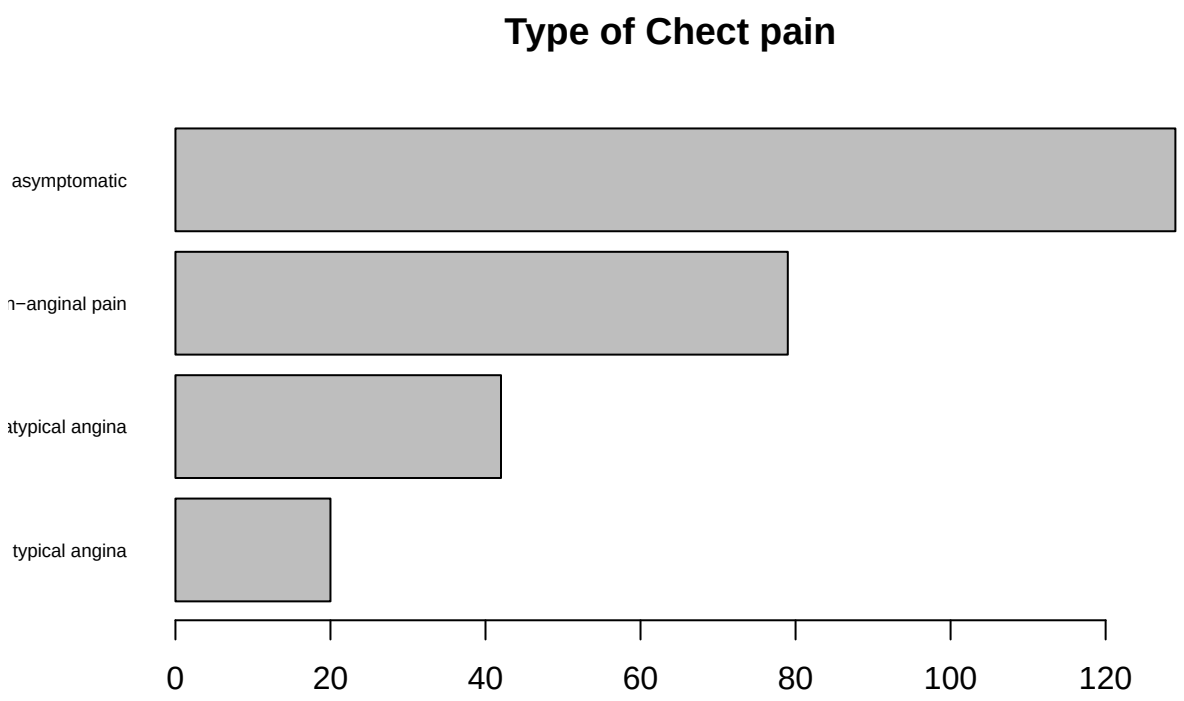


We have 87 Females and 183 Males in this dataset, and from the barplot and the pie chart, we can say that the most of patients in this dataset are Male.

cp: Type of chest pain experienced by patient where 0 means typical angina, 1 represents atypical angina, 2 represents non-anginal pain and 3 represents asymptomatic. cp is a nominal variable.

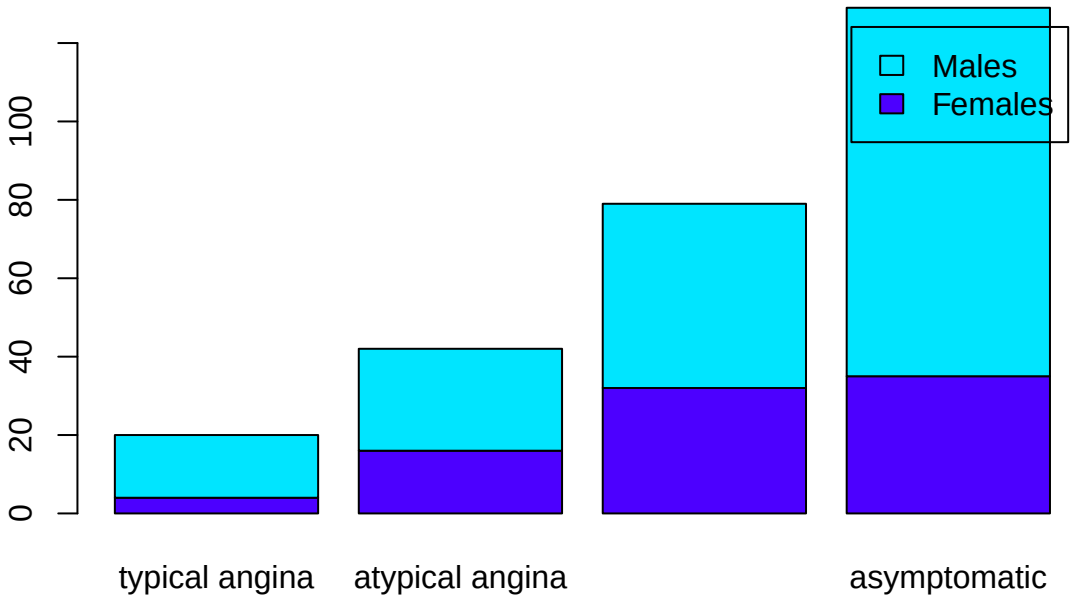
```
par(las=1)
barplot(table(hd$cp), main = "Type of Chect pain" , horiz=TRUE,cex.names=0.6,
```

```
        names.arg = c("typical angina",
                                    "atypical angina",
                                    "non-anginal pain","asymptomatic"))
```



Most of patients feel asymptomatic and only few of them experienced typical angina.

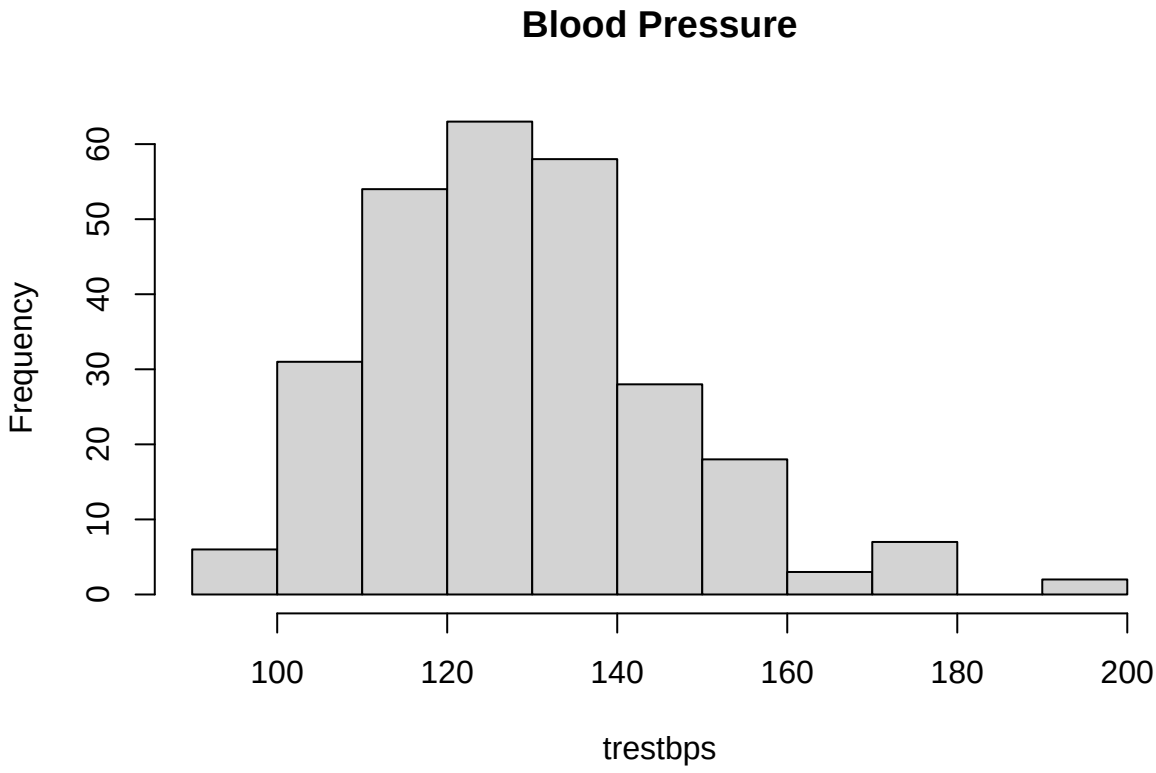
```
barplot((table(hd$sex,hd$cp)), legend=c("Females","Males"),
         names.arg = c("typical angina",
                                    "atypical angina",
                                    "non-anginal pain","asymptomatic"),
        col = topo.colors(4))
```



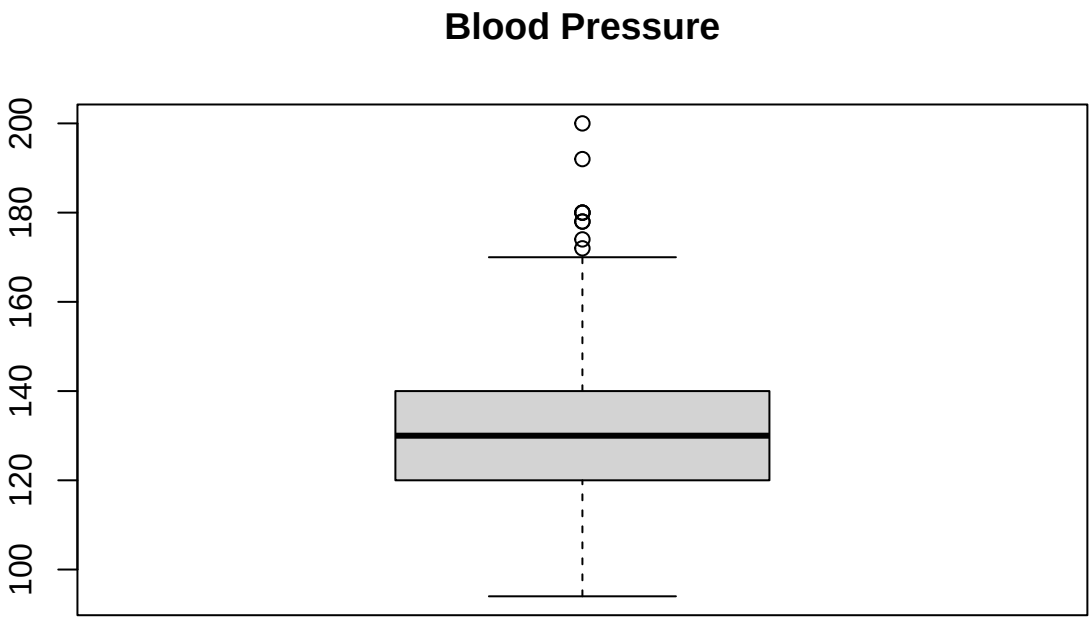
The patients who feel asymptomatic most are Males and the patients who experienced typical angina are also Males.

trestbps: This is a continuous variable which means the patients' level of blood pressure at resting mode in mm/HG.

```
hist(hd$trestbps, xlab = "trestbps", main = c("Blood Pressure"))
```



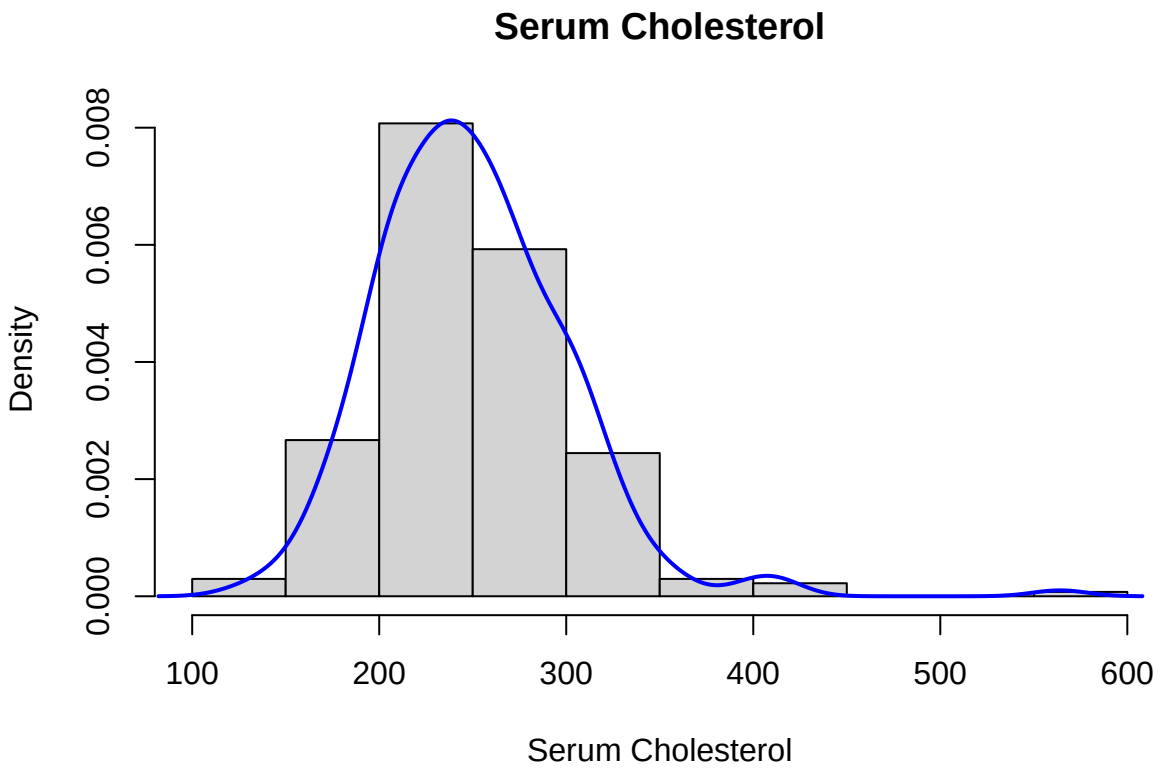
```
boxplot(hd$trestbps,main = "Blood Pressure")
```



We can say that most of people have the blood pressure at the range 100 to 160, and few of them have the blood pressure over 160. And from the boxplot, we have 6 outliers.

chol is a continuous variable which means the serum cholesterol in mg/dl.

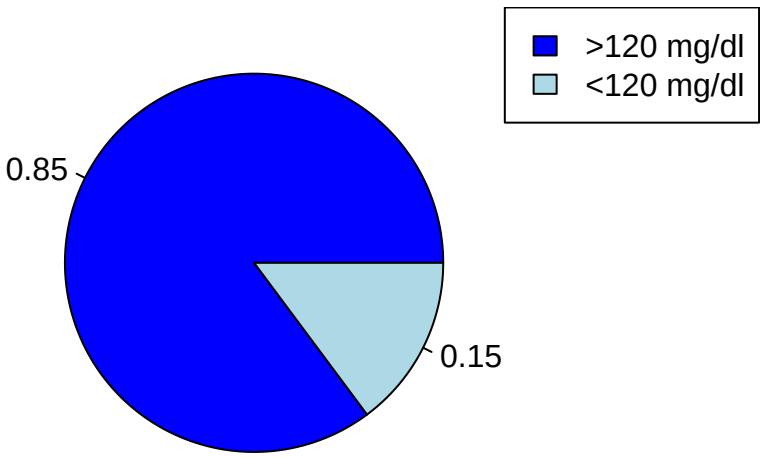
```
hist(hd$chol,freq = FALSE, main = "Serum Cholesterol", xlab = "Serum Cholesterol")
lines(density(hd$chol), col="blue", lwd=2)
```



We can say that the most patients' level of serum cholesterol are between 150 to 300 and few of them are not in this range.

fbs is a Nominal variable which represent the blood sugar fasting levels. we assign the 1 to the patients whoes blood sugar levels on fasting greater than 120 mg/dl, otherwise 0.

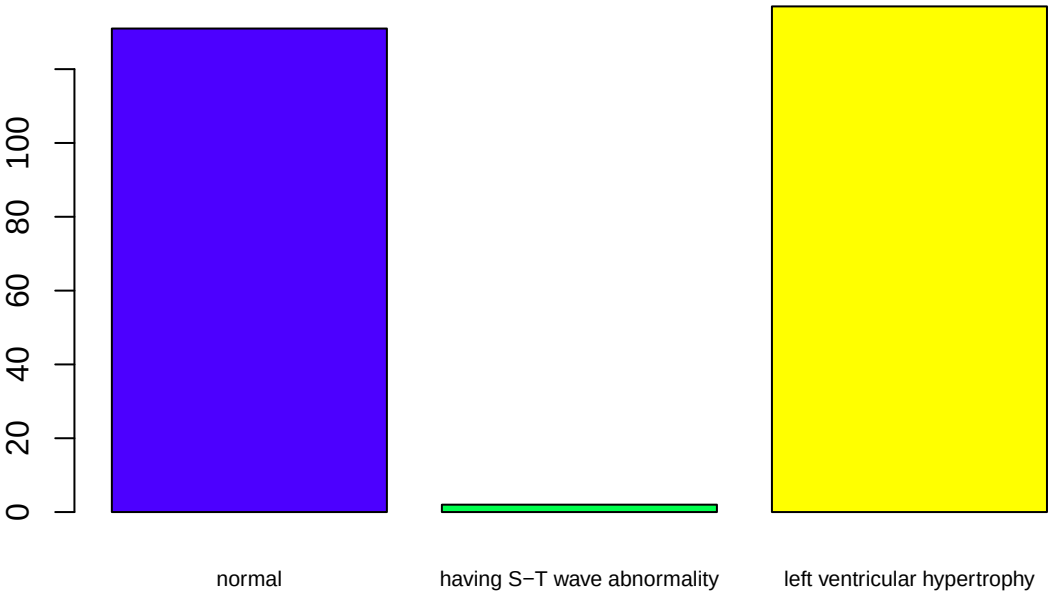
```
hd$fbs<-factor(hd$fbs)
pie(table(hd$fbs), labels=round(table(hd$fbs)/270, digit = 2),
    col = c("blue", "lightblue"),
      main=" blood sugar fasting levels")
legend("topright", c(">120 mg/dl", "<120 mg/dl"), fill =c("blue", "lightblue") )
```



From above pir chart, we can say that 85% patients have blood sugar fasting levels less than 120 mg/dl, and 15% patients' blood sugar fasting levels are greater than 120md/dl.

Restecg is a Nominal variable with 3 distinct values. 0 means normal, 1 means having S-T wave abnormality, 2 means probable or definite left ventricular hypertrophy by Estes' criteria.

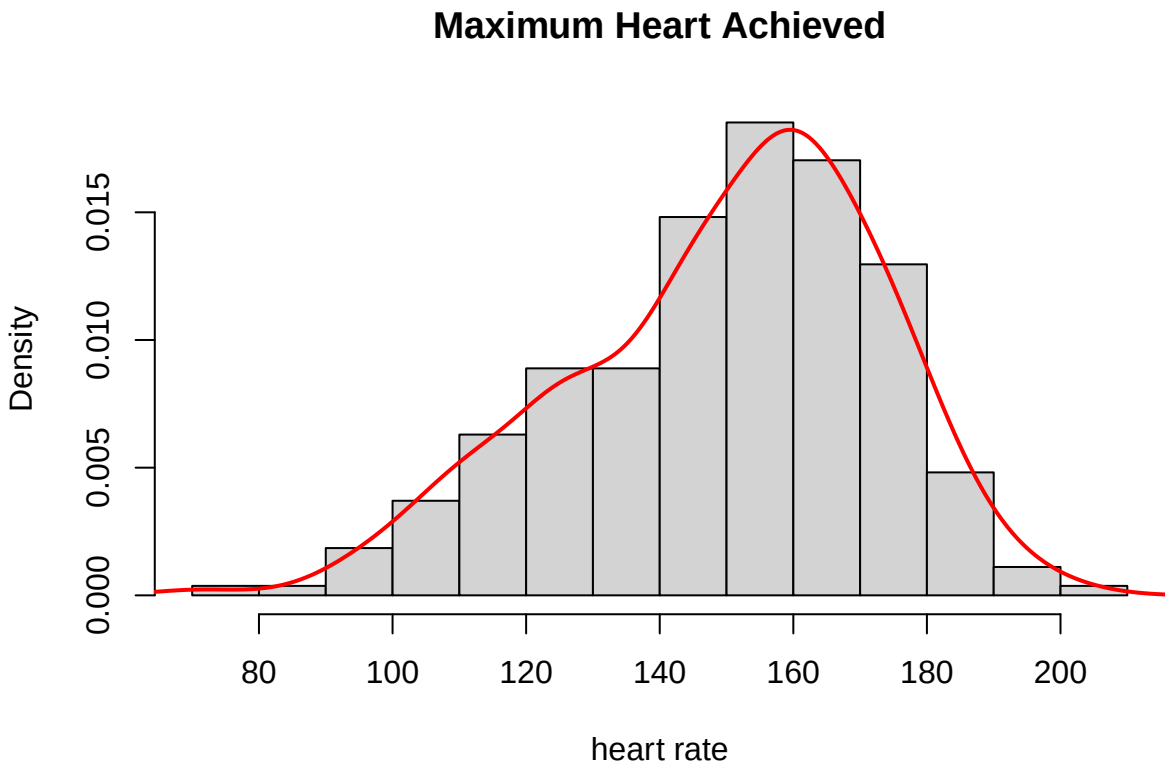
```
hd$restecg<-factor(hd$restecg)
barplot(table(hd$restecg), col = topo.colors(3), names.arg = c("normal","having S-T wave
```



From the above barplot, we can say that only few patients have S-T wave abnormality, and the number of patient between who are normal and who have left ventricular hypertrophy have the relatively same amount.

thalach is a contunous varible which represents the maximum heart rate achived.

```
hist(hd$thalach, main =  "Maximum Heart Achieved", xlab =
       "heart rate", freq = FALSE)
lines(density(hd$thalach), col = c("red"), lwd= 2)
```



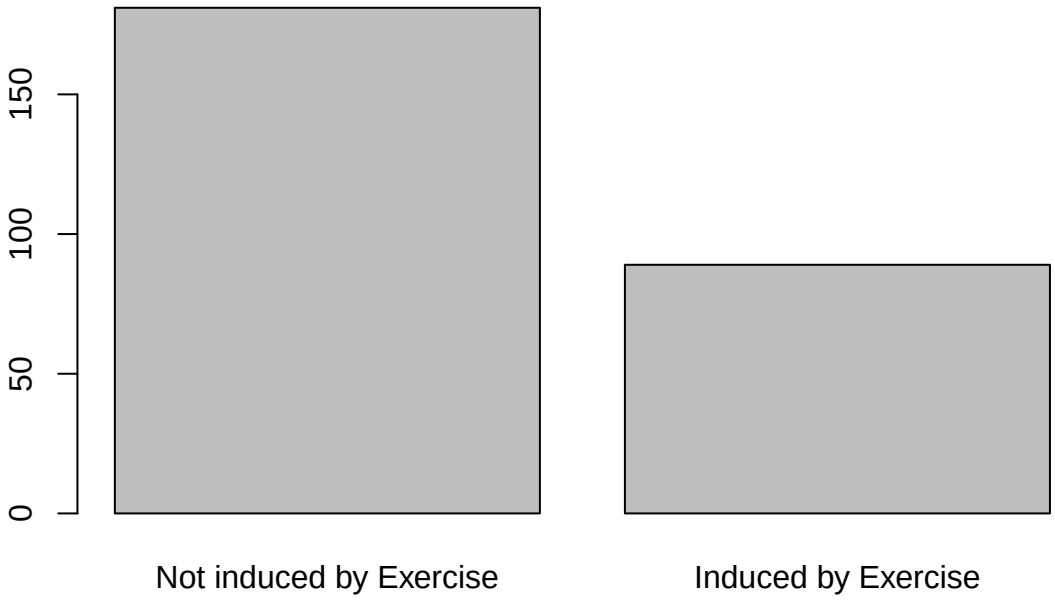
From the above figure, we can say that the trend skewed right and most of patients' maximux heart achieved located at the range of 140 to 180.

exang is a nominal variable which only contain 0 and 1. 1 represents the angina induced by exercise("Yes"), and 0 represents "No".

```
hd$exang<-factor(hd$exang)
table(hd$exang)
```

```
##
##   0   1
## 181  89
```

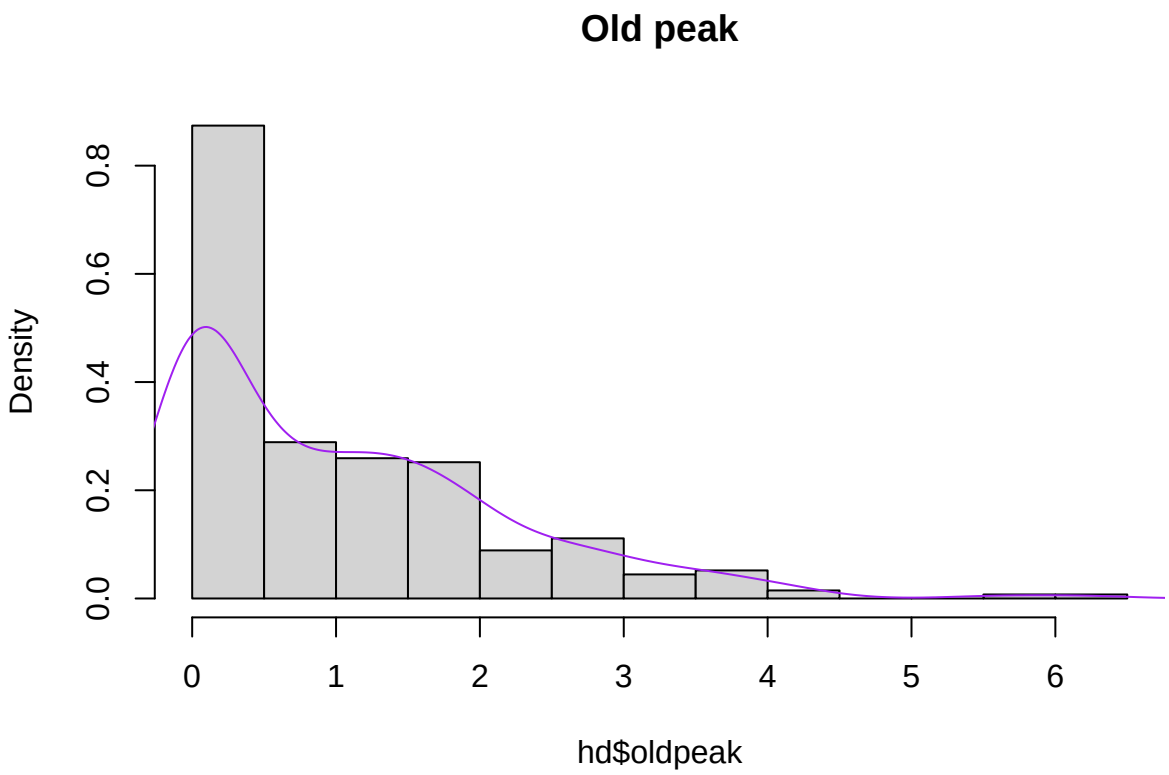
```
barplot(table(hd$exang), names.arg = c("Not induced by Exercise","Induced by Exercise"))
```



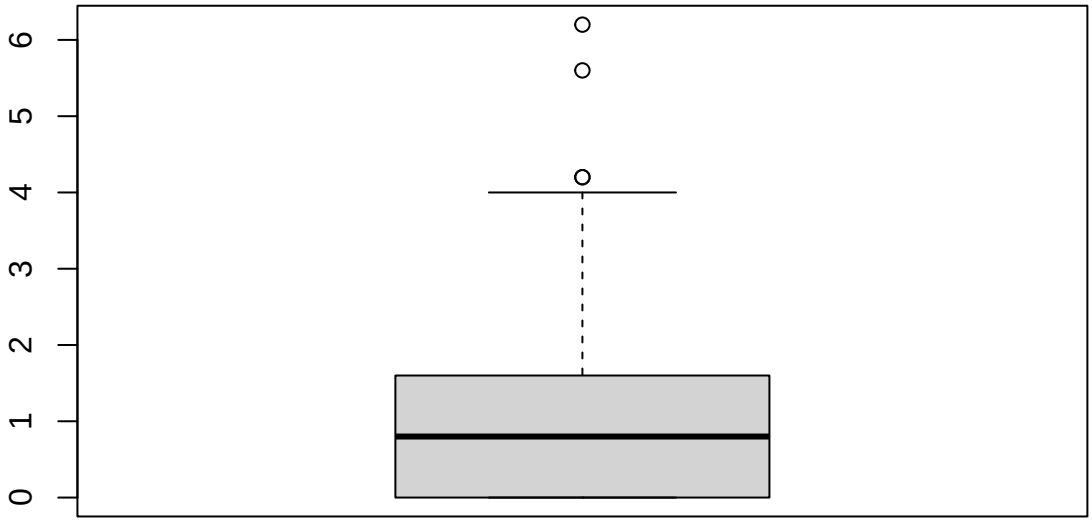
We can say that most of patients suffering angina which is not induced by exercise.

Oldpeak is a continuous varible which represents exercise induced ST-depression in relative with the state of rest.

```
hist(hd$oldpeak,main = "Old peak",freq = FALSE)
lines(density(hd$oldpeak), col =c("purple"))
```



```
boxplot(hd$oldpeak)
```



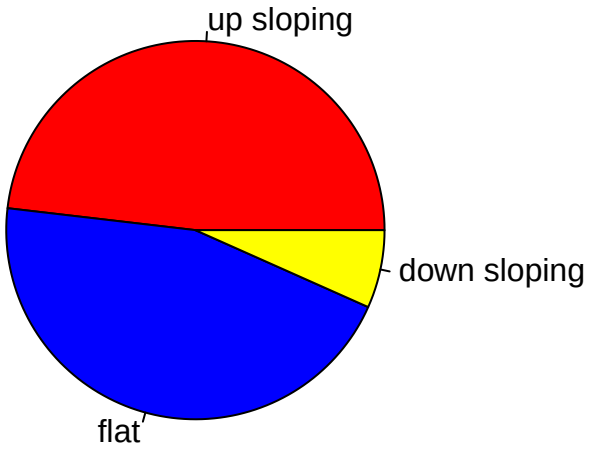
From the above histogram, we can get the information that this variable skews left and has decreasing pattern as x axis goes large.

Slope is a nominal factor which represents that ST segment measured in terms of slope during peak exercise.

```
table(hd$slope)
```

```
##
##   0   1   2
## 130 122  18
```

```
pie(table(hd$slope),
    labels = c("up sloping", "flat", "down sloping"),
               col = c("red", "blue","yellow"))
```



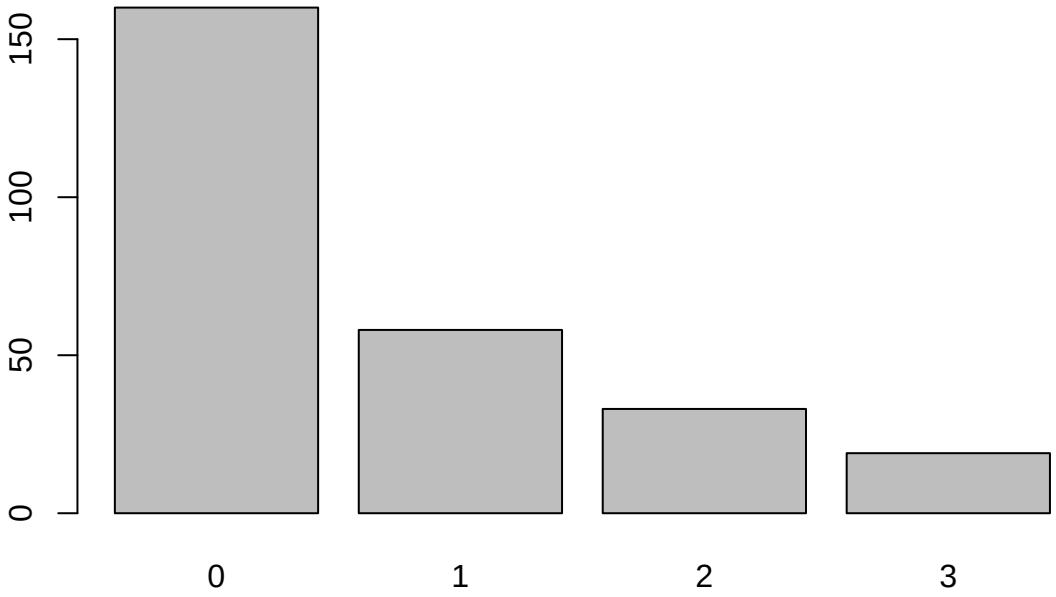
From the above pie chart, we can say that half of the patients' ST segment measured up sloping.

ca is a nominal variable which represents number of major vessels colored by flourosopy

```
table(hd$ca)
```

```
##
##   0   1   2   3
## 160  58  33  19
```

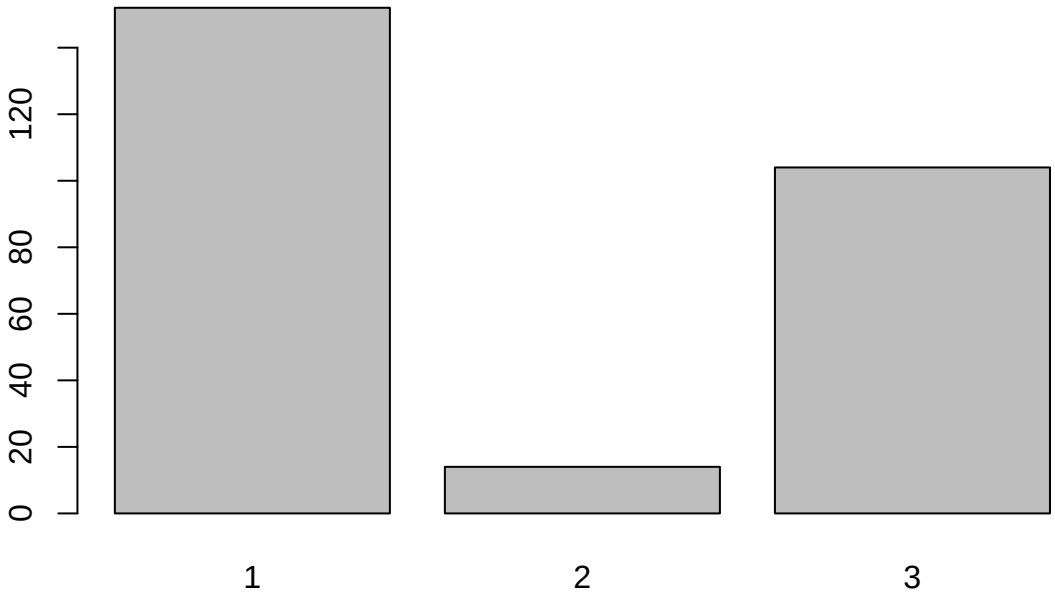
```
barplot(table(hd$ca))
```



We can say that over 150 patients have 0 major vessels colored by flourosopy. And only few patients have 3 vessels colored by flourosopy.

thal is the nominal variable which represents the degree of patients suffering from the blood disorder. 1 represents the normal blood flow, 2 represents no blood flow in some part, 3 represents reversible defect.

```
barplot(table(hd$thal))
```

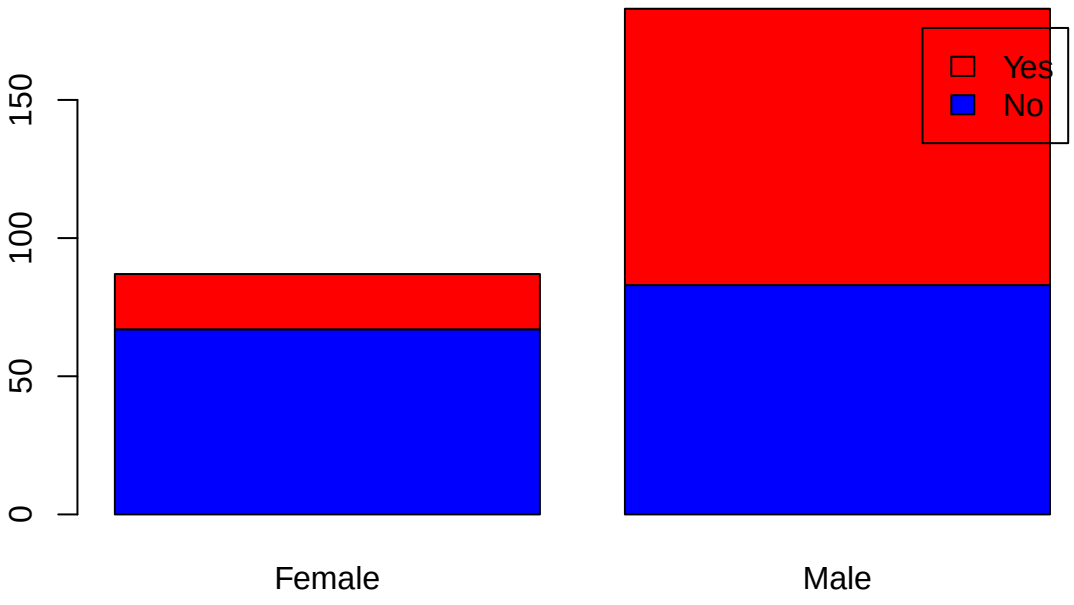


target is out goal to predict, 0 means patients are not suffering from heart disease staglog and 1 means patients are suffering from the heart disease staglog.

```
table(hd$target,hd$sex)
```

```
##
##          0   1
##   No    67  83
##   Yes   20 100
```

```
barplot((table(hd$target,hd$sex)), legend=c("No","Yes"),
        names.arg = c("Female","Male"),
        col = c("blue","red"))
```



We have 150 patients are not suffering but 120 patients suffering from the heart disease. and 100 Male suffering from the disease, and 20 Female.

Data Analysis

Since the "target" in this dataset is binary, 0 and 1. So, we want to conduct the logistic regression to determine which factors affect our regression model. First, we need to check whether at least one of the predictors is associated with target.

H_0 : The parameters that two models under comparison do not have in common are all zero

H_a : The parameters that two models under comparison do not have in common at least are not zero

```
fit.null <- glm(target~1, data=hd, family=binomial())
fit1 <- glm(target~., data = hd, family = binomial())
anova(fit1,fit.null,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach +
##     exang + oldpeak + slope + ca + thal
## Model 2: target ~ 1
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1       249     161.66
## 2       269     370.96 -20   -209.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above test, since the p-value are less than 0.05, we reject the null hypothesis and get the conclusion that at least one predictors are associated with the target. Then, we need to test whether there exists the multicollinearity problem. if multicollinearity exists, it will affect our analysis.

```
library(car)
```

```
## Loading required package: carData
```

```
vif(fit1)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.481227  1        1.217057
## sex      1.684318  1        1.297813
## cp       1.922381  3        1.115081
## trestbps 1.313574  1        1.146113
## chol     1.357859  1        1.165272
## fbs      1.188371  1        1.090124
## restecg  1.175801  2        1.041318
## thalach  1.546327  1        1.243514
## exang    1.143408  1        1.069303
## oldpeak  1.620909  1        1.273149
## slope    1.842725  2        1.165105
## ca       1.903984  3        1.113296
## thal     1.449503  2        1.097248
```

From the result, the VIF results are all less than 5. Thus, we can say that there does not exist multicollinearity problem, and we don't need to drop the highly correclated factors in this dataset.

Then, we have the logistic regression model as the form

$$\ln(\frac{\pi_i}{1-\pi_i}) = \eta_0 + \eta_1 X_1 + \cdots + \eta_{13} X_{13},$$

where η_0 represents the intercept and η_i, $i \in [1, 13]$ represents the coefficients for each variable. Specifically, X_1 represents age, X_2 represents Sex X_3 represents type of chest pain, X_4 represents level of blood pressure, X_5 represents Serum cholesterol, X_6 represents blood sugar level, X_7 represents result of electrocardiogram, X_8 represents Maximum heart rate achieved, X_9 represents whether angina induced by exercise, X_{10} represents the variable oldpeak, X_{11} represents the slope, X_{12} represents the number of major vessels and X_{13} represents the blood disorder.Also, π_i is the probability that patient i will suffer from the heart disease. We have 5 assumptions for logistic regression: 1. the outcome is binary 2.There is no multicollinearity between variables. 3. There is large sample size. 4. Each variable are independent. 5. There is a linear relationship between independent variables and log-odds.

```
fit1 <- glm(target ~. ,data =hd, family = binomial())
summary(fit1)
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(), data = hd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9284  -0.4382  -0.1170   0.2969   2.9516
```

```
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -7.686960   3.287776  -2.338 0.019385 *
## age                -0.025110   0.026610  -0.944 0.345362
## sex1                1.898998   0.606029   3.134 0.001727 **
## cpatypical angina   1.741171   0.945881   1.841 0.065652 .
## cpnon-anginal pain  0.784877   0.792325   0.991 0.321881
## cpasymptomatic      2.748658   0.812980   3.381 0.000722 ***
## trestbps            0.031110   0.012799   2.431 0.015074 *
## chol                0.006557   0.004300   1.525 0.127297
## fbs1               -0.376047   0.606920  -0.620 0.535522
## restecg1            0.803613   3.561836   0.226 0.821499
## restecg2            0.676840   0.425719   1.590 0.111863
## thalach            -0.020480   0.012182  -1.681 0.092736 .
## exang1              0.534717   0.467726   1.143 0.252944
## oldpeak             0.476061   0.252840   1.883 0.059720 .
## slope1              1.113087   0.514352   2.164 0.030460 *
## slope2              0.128387   1.061333   0.121 0.903716
## ca1                 2.152088   0.559653   3.845 0.000120 ***
## ca2                 3.100493   0.807546   3.839 0.000123 ***
## ca3                 2.164689   0.926071   2.337 0.019413 *
## thal2              -0.318858   0.865164  -0.369 0.712461
## thal3               1.468745   0.465422   3.156 0.001601 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 370.96  on 269  degrees of freedom
## Residual deviance: 161.66  on 249  degrees of freedom
## AIC: 203.66
##
## Number of Fisher Scoring iterations: 6
```

Since we have some significant variables, then, we are going to reduce the model making this model fitter. So, we are going to find the model which has the smallest AIC. In order to get the model, we can use the variable stepwise selection.

```
reduced <- step(fit1, direction="backward")
```

```
## Start:  AIC=203.66
## target ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach +
##     exang + oldpeak + slope + ca + thal
```

```
##
##            Df Deviance    AIC
## - fbs       1   162.05 202.05
## - restecg   2   164.25 202.25
## - age       1   162.56 202.56
## - exang     1   162.96 202.96
## <none>          161.66 203.66
## - chol      1   163.94 203.94
## - thalach   1   164.61 204.61
## - slope     2   166.96 204.96
## - oldpeak   1   165.41 205.41
## - trestbps  1   168.01 208.01
## - thal      2   173.98 211.98
## - sex       1   172.65 212.65
## - cp        3   182.64 218.64
## - ca        3   191.78 227.78
##
## Step:  AIC=202.05
## target ~ age + sex + cp + trestbps + chol + restecg + thalach +
##     exang + oldpeak + slope + ca + thal
##
##            Df Deviance    AIC
## - restecg   2   164.63 200.63
## - age       1   162.98 200.98
## - exang     1   163.28 201.28
## <none>          162.05 202.05
## - chol      1   164.31 202.31
## - thalach   1   165.10 203.10
## - slope     2   167.20 203.20
## - oldpeak   1   166.18 204.18
## - trestbps  1   168.08 206.08
## - thal      2   174.74 210.74
## - sex       1   172.89 210.89
## - cp        3   185.00 219.00
## - ca        3   191.87 225.87
##
## Step:  AIC=200.63
## target ~ age + sex + cp + trestbps + chol + thalach + exang +
##     oldpeak + slope + ca + thal
##
##            Df Deviance    AIC
## - age       1   165.45 199.45
## - exang     1   165.88 199.88
## <none>          164.63 200.63
## - thalach   1   167.79 201.79
```

```
## - chol      1   167.95 201.95
## - slope     2   169.98 201.98
## - oldpeak   1   168.65 202.65
## - trestbps  1   171.03 205.03
## - thal      2   176.15 208.15
## - sex       1   176.19 210.19
## - cp        3   187.61 217.61
## - ca        3   194.90 224.90
##
## Step:  AIC=199.45
## target ~ sex + cp + trestbps + chol + thalach + exang + oldpeak +
##     slope + ca + thal
##
##            Df Deviance    AIC
## - exang     1   166.72 198.72
## <none>          165.45 199.45
## - thalach   1   167.88 199.88
## - chol      1   168.32 200.32
## - slope     2   170.67 200.67
## - oldpeak   1   169.79 201.79
## - trestbps  1   171.12 203.12
## - thal      2   177.24 207.24
## - sex       1   177.60 209.60
## - cp        3   189.11 217.11
## - ca        3   195.73 223.73
##
## Step:  AIC=198.72
## target ~ sex + cp + trestbps + chol + thalach + oldpeak + slope +
##     ca + thal
##
##            Df Deviance    AIC
## <none>          166.72 198.72
## - chol      1   169.72 199.72
## - thalach   1   169.85 199.85
## - slope     2   172.42 200.42
## - oldpeak   1   171.29 201.29
## - trestbps  1   172.74 202.74
## - thal      2   179.57 207.57
## - sex       1   178.88 208.88
## - cp        3   195.42 221.42
## - ca        3   196.98 222.98
```

The reduced model has the AIC value 198.7, which are all less than the AIC of full model 203.66. after reducing the model, we also need to check whether the multicollinearity problem.

```
vif(reduced)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## sex      1.620500  1        1.272988
## cp       1.644880  3        1.086482
## trestbps 1.177683  1        1.085211
## chol     1.237748  1        1.112541
## thalach  1.316071  1        1.147201
## oldpeak  1.581924  1        1.257746
## slope    1.788804  2        1.156487
## ca       1.528960  3        1.073329
## thal     1.352960  2        1.078503
```

Since the value are all less than 5, we can say that there does not exists multicollinearity problem which does not affect our analysis.

From above, we can get the logistic regression model

```
coef(reduced)
```

```
##        (Intercept)                sex1  cpatypical angina cpnon-anginal pain
##       -8.341127590         1.938018753         1.571627943         0.598611211
##     cpasymptomatic            trestbps                chol             thalach
##        2.817274217         0.027539138         0.006962048        -0.019366787
##            oldpeak              slope1              slope2                 ca1
##        0.504231178         1.139405100         0.180205834         2.065998617
##                ca2                 ca3               thal2               thal3
##        2.716824431         1.922220122        -0.454991930         1.399751836
```

$$\ln(\frac{\pi_i}{1-\pi_i}) = -8.341127590 + 1.938018753 \cdot \text{sex1} + 1.571627943 \cdot \text{cp atypical angina}$$

$$+0.598611211 \cdot \text{cp non-anginal pain} + 2.817274217 \cdot \text{cp asymptomatic}$$

$$+0.027539138 \cdot \text{trestbps} + 0.006962048 \cdot \text{chol} - 0.019366787 \cdot \text{thalach}$$

$$+0.504231178 \cdot \text{oldpeak} + 1.139405100 \cdot \text{slope1} + 0.180205834 \cdot \text{slope2}$$

$$+2.065998617 \cdot \text{ca1} + 2.716824431 \cdot \text{ca2} + 1.922220122 \cdot \text{ca3}$$

$$-0.454991930 \cdot \text{thal2} + 1.399751836 \cdot \text{thal3}$$

```
exp(coef(reduced))
```

```
##        (Intercept)               sex1  cpatypical angina cpnon-anginal pain
##       2.385033e-04       6.944978e+00       4.814480e+00       1.819590e+00
##     cpasymptomatic           trestbps               chol            thalach
##       1.673118e+01       1.027922e+00       1.006986e+00       9.808195e-01
##            oldpeak             slope1             slope2                ca1
##       1.655712e+00       3.124909e+00       1.197464e+00       7.893176e+00
##                ca2                ca3               thal2               thal3
##       1.513219e+01       6.836119e+00       6.344531e-01       4.054194e+00
```

One unit increase in tresbps (level of blood pressure at resting mode)multiplies the odds of an heart disease by 1.027922. One unit increase in chol(serum cholesterol in mg/dl) multiplies the odds of an heart disease by 1.006986. One unit increase in thalach(Maximum heart rate) multiplies the odds of an heart disease by 0.9808195. One unit increase in oldpeak(Exercise induced ST-depression at resting mode) multiplies the odds of an heart disease by 1.655712.

From the above coefficient table, we can say that Male have 6 times get heart disease odds of Female if holding other predictors constant. For those who suffering the heart disease, they are more likely experience asymptomatic chest of pain. Patients who has 2 major vessels colored by flourosopy have approximately 2 times odds of patients who have 1 major vessels and 3 major vessels colored by flourosopy. Patients who have reversible defect blood disorder have 10 times odds of patient who have normal disorder.

Conclusion and Discussion

First, we need to check the assumptions about the logistic regression. The output is binary, only including "Yes" and "No", and there is no multicollinearity problem based on the VIF we got from the last section. Then, we have a large sample size(270), each variable is independent, and there is a linear relationship between independent variables and log odds. So, all the assumptions are met and this logistic regression model is appropriate.

There still have several problems in this project. First of all, this dataset only include some variables which are related to patients' information about body health, but there are more other factors, like smoking or not, having more meat or not, etc. Some of them may correlated to the health information, but other factors are still important in some ways. Even though this data already involved so many variables, other factors may play an important way in this model. From last section, we get a relatively better model to predict whether patients are suffering from the heart disease. So, we want to know how good our model is compared with full model.

```
set.seed(1234)



n <- nrow(hd)
train <- sample(n, 0.8*n)
testid <- setdiff(1:nrow(hd), train)


fit2 <- glm(target ~ sex + cp + trestbps + chol + thalach +
              oldpeak + slope + ca + thal,
            data = hd, family = binomial(), subset = train)
testactual <- hd$target[testid]
testpred<-predict(fit2, newdata = hd[testid,], type = "response")
test.pred<-rep("No",length(testid))
test.pred[testpred > 0.5] <- "Yes"
mean(test.pred==testactual)
```

```
## [1] 0.7777778
```

The prediction accuracy for reduced model are approximately 78%.

```
fit3 <-glm(target~.,data = hd, family = binomial(), subset = train)
testactual1 <- hd$target[testid]
testpred1<-predict(fit3, newdata = hd[testid,], type = "response")
test.pred1<-rep("No",length(testid))
test.pred1[testpred1 > 0.5] <- "Yes"
mean(test.pred1==testactual1)
```

```
## [1] 0.7592593
```

And the prediction accuracy for the full model is approximately 76%

After reducing the model, we only get approximately a 2% increase in prediction accuracy. This is the limitation of this logistic regression model. In order to improve the prediction accuracy, we may drop more variables. For example, we may pick the group of people who experience asymptomatic, then we can use logistic regression again to make a more fitted model. Or using another method to predict the result.

Reference

1. WHO(2021)."Cardiovascular diseases (CVDs)",[https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#]

2. Mozaffarian, D., Wilson, P. W., and Kannel, W. B. (2008), "Beyond established and novel risk factors: lifestyle risk factors for cardiovascular disease," Circulation, 117, 3031–3038.

3. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.