Prediction of Heart Disease Using Regression

Final Project

Student Details

Prerit Miglani (S3815838) Anirudhda Pardhi (S3807109)

Table of content:

TABLE OF CONTENTS

1.INTRODUCTION :

As, the number of heart diseases are witnessed to increase by humongous numbers these days , which corresponds to many factors such as stress, lifestyle, pollution etc. Also,heart diseases leading to deaths among the young population has also been significant these days. So,in this project we are taking a clinical dataset of heart diseases and will be dealing with some clinical associations with the disease. The dataset has been sourced from Kaggle.com(https://www.kaggle.com/ronitf/heart-disease-uci (https://www.kaggle.com/ronitf/heart-disease-uci)) .This dataset contains 14 variables and 303 records of both male & female of various age groups having different types of clinical illness which will be analysed in respect to causing the heart disease. So, we will be analysing all the variables and correlate it leading to some meaningful associations with the clinical results and also we would model the relationship between Heart attribute(response variable) against the most significant attributes(predictor variables) using logistic regression.

2.Dataset and description:

Following are the Variables in the dataset with their description:

1. Age: The person’s age in years

2. Gender: The person's gender (1 = male, 0 = female)
3. CP: The chest pain experienced (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
4. TBps: The person's resting blood pressure (mm Hg when admitted)
5. Chol: The person's cholesterol measurement in mg/dl
6. Fbs: The person's fasting blood sugar (if > 120 mg/dl, 1 = True; 0 = False)
7. Recg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy criteria)
8. Thalach: The person's maximum heart rate
9. Exang: Exercise induced angina (1 = yes; 0 = no)
10. Op: ST depression induced by exercise ('ST' relates to positions on the ECG plot)
11. Slope: the slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
12. CA: The number of major vessels (0-3)
13. Thal: A blood disorder called thalassemia (1 = normal; 2 = fixed defect; 3 = reversable defect)
14. Heart: Heart disease (0 = no, 1 = yes)

3.Goals & Objective:

The model explain the reasons to develop a relationship between the heart disease of Male & Female for different parameters like Chest Pain, Resting Blood Pressure, Age , Cholesterol , Fasting blood sugar , Resting electrocardiographic measurement, Maximum Heart Rate, Blood disorder respectively.

Load Packages

```
# This is a chunk where you can load the necessary packages required to reproduce the report

library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(colourpicker)
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```
library(score)
```

```
## Loading required package: msm
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(leaps)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

4. DATA PREPOCESSING

```
## Reading the heart.csv

heart<-read.csv("/Users/preritmiglani/Desktop/heart.csv")

## Changing the Column names:

colnames(heart)<- c("Age","Gender","CP","TBps",
                    "Chol","Fbs","Recg","Thalach","Exang","Op",
                    "Slope","Ca","Thal","Heart")

## Checking for the null values:

sum(is.na(heart))
```

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

```
class(heart)
```

```
## [1] "data.frame"
```

```
## Checking for the structure
str(heart)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ Age    : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ Gender : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ CP     : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ TBps   : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ Chol   : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ Fbs    : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ Recg   : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ Thalach: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ Exang  : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Op     : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ Slope  : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ Ca     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Thal   : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ Heart  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
#dimensions of the dataframe i.e. 303 records and 14 attributes
dim(heart)
```

```
## [1] 303  14
```

```
#summary of the dataframe
summary(heart)
```

```
##       Age            Gender             CP             TBps      
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
##       Chol            Fbs              Recg           Thalach     
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5  
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0  
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6  
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
##      Exang             Op           Slope             Ca        
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
##       Thal           Heart       
##  Min.   :0.000   Min.   :0.0000  
##  1st Qu.:2.000   1st Qu.:0.0000  
##  Median :2.000   Median :1.0000  
##  Mean   :2.314   Mean   :0.5446  
##  3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :3.000   Max.   :1.0000
```

```
#factorizing the categorical columns

#sex(1 = male; 0 = female)
heart$Gender = factor(heart$Gender,levels = c("0","1"), labels= c("female", "male"))

#fbs(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
heart$Fbs = factor(heart$Fbs, levels =c("0", "1"), labels = c("false", "true"))

#exang exercise induced angina (1 = yes; 0 = no)
heart$Exang = factor(heart$Exang,levels  = c("0", "1"), labels = c("no", "yes"))

# heart disease likely to exist (0 = no ; 1= yes)

heart$Heart = factor(heart$Heart,levels= c("0","1"),labels = c("no","yes"))

# cp: chest pain type
#-- Value 1: typical angina
#-- Value 2: atypical angina
#-- Value 3: non-anginal pain
#-- Value 4: asymptomatic
heart$CP = factor(heart$CP, levels=  c("0", "1","2","3"), labels = c("typical angina",
                                                                     "atypical angina",
                                                                     "non-anginal pain",
                                                                     "asymptomatic"))
#slope: the slope of the peak exercise ST segment
#-- Value 1: upsloping
#-- Value 2: flat
#-- Value 3: downsloping
heart$Slope = factor(heart$Slope, levels = c("0", "1","2"), labels = c("upsloping",
                                                                       "flat",
                                                                       "downsloping"))

#restecg: resting electrocardiographic results
#-- Value 0: normal
#-- Value 1: having ST-T wave abnormality
#-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
heart$Recg = factor(heart$Recg, levels = c("0", "1","2"), labels = c("normal",
                                                                     "ST-T wave abnormality",
                                                                     "left ventricular hypertroph
y"))

## Again, checking for the structure
str(heart)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ Age    : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ Gender : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 1 2 2 2 ...
##  $ CP     : Factor w/ 4 levels "typical angina",..: 4 3 2 2 1 1 2 2 3 3 ...
##  $ TBps   : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ Chol   : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ Fbs    : Factor w/ 2 levels "false","true": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Recg   : Factor w/ 3 levels "normal","ST-T wave abnormality",..: 1 2 1 2 2 2 1 2 2 2 ...
##  $ Thalach: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ Exang  : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
##  $ Op     : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ Slope  : Factor w/ 3 levels "upsloping","flat",..: 1 1 3 3 3 2 2 3 3 3 ...
##  $ Ca     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Thal   : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ Heart  : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
```

```
head(heart)
```

	...	Gender	CP	T...	Chol	Fbs	Recg	Thalach	Exang
	<int>	<fct>	<fct>	<int>	<int>	<fct>	<fct>	<int>	<fct>
1	63	male	asymptomatic	145	233	true	normal	150	no
2	37	male	non-anginal pain	130	250	false	ST-T wave abnormality	187	no
3	41	female	atypical angina	130	204	false	normal	172	no
4	56	male	atypical angina	120	236	false	ST-T wave abnormality	178	no
5	57	female	typical angina	120	354	false	ST-T wave abnormality	163	yes
6	57	male	typical angina	140	192	false	ST-T wave abnormality	148	no

6 rows | 1-10 of 15 columns

5. DATA VISUALISATION & EXPLORATION

```
## 1. Now, we will check for the cp variable and would lookout for the most frequent chest pain re
ported by patients when admitted :

t3<- table(heart$CP)
t3
```

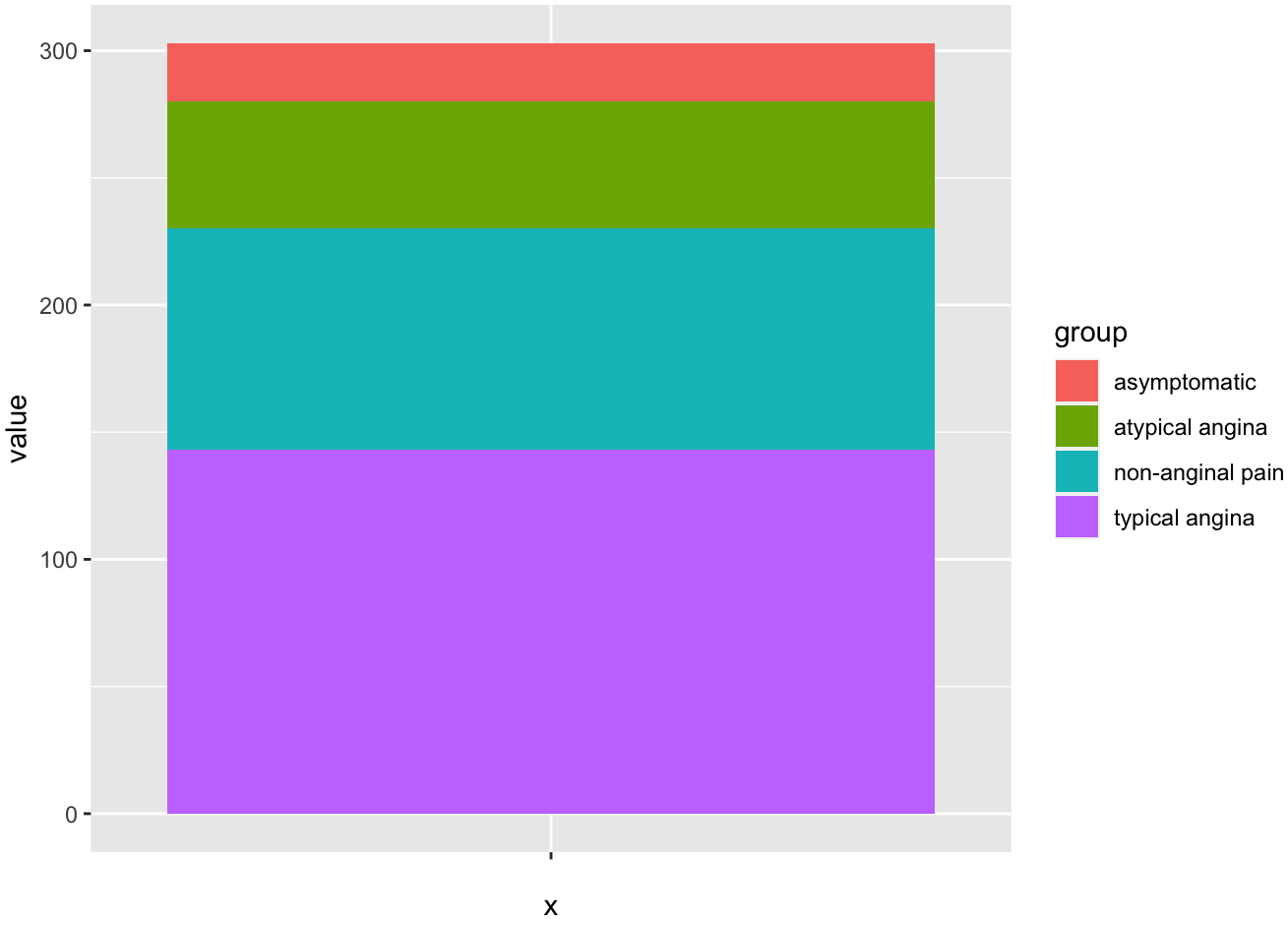
```
##
##   typical angina  atypical angina non-anginal pain     asymptomatic
##              143               50               87               23
```

```
df3<-data.frame(
  group = c("typical angina", "atypical angina", "non-anginal pain","asymptomatic"),
  value = c(143, 50, 87,23)
)
head(df3)
```

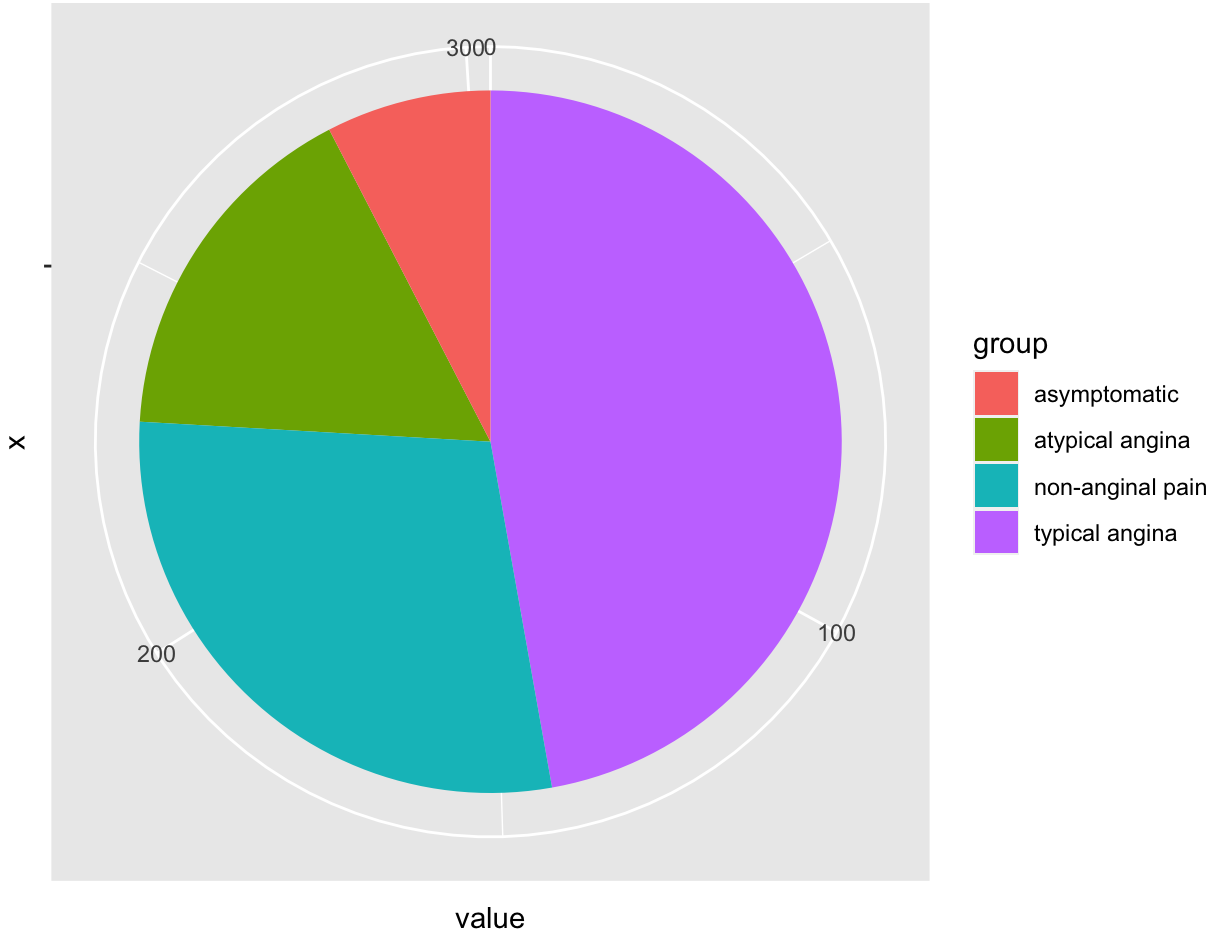
	group <fct>	value <dbl>
1	typical angina	143
2	atypical angina	50
3	non-anginal pain	87
4	asymptomatic	23

4 rows

```
bp2<- ggplot(df3, aes(x="", y=value, fill=group))+
  geom_bar(width = 1, stat = "identity")
bp2
```




```
pie2 <- bp2 + coord_polar("y", start=0)
pie2
```




```
## Typical angina is the most frequent type of chest pain experienced by the patients while non-an
ginal pain comes after that, followed by atypical angina whereas asymptomatic is the least type re
ported.



## 2. Now, we will subset two variables and look for some meaningful association. Here, we would o
bserve Gender & Chol levels and find the statistical variation between the two:

heart_subset <- heart %>% dplyr::select(`Chol`,`Gender`)
View(heart_subset)
heart_subset_male<-heart_subset %>% filter(.,heart_subset$Gender=="male")
head(heart_subset_male)
```

	Chol	Gender
	<int>	<fct>
1	233	male
2	250	male
3	236	male
4	192	male
5	263	male
6	199	male

6 rows

```
heart_subset_female<-heart_subset %>% filter(.,heart_subset$Gender=="female")
head(heart_subset_female)
```

	Chol	Gender
	<int>	<fct>
1	204	female
2	354	female
3	294	female
4	275	female
5	283	female
6	219	female

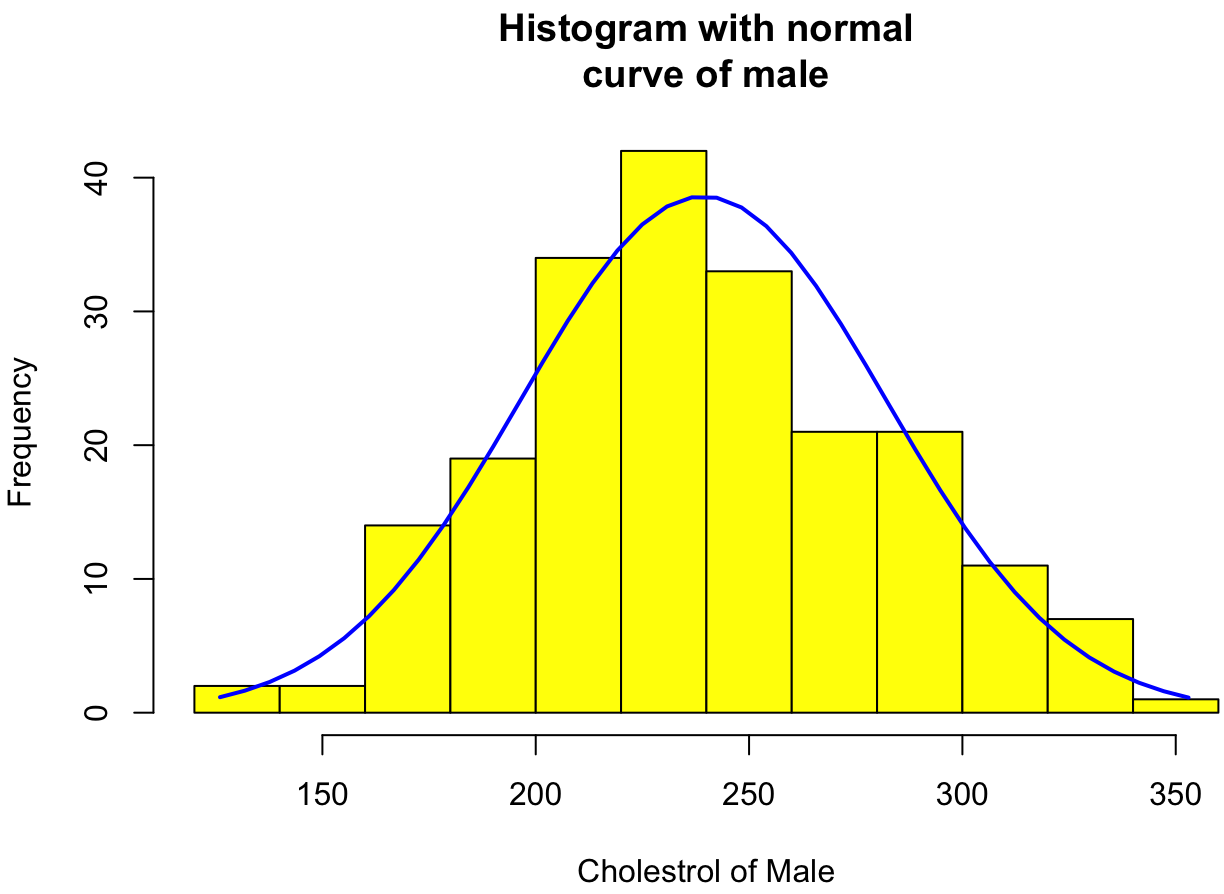
6 rows

```
heart_subset %>% group_by(`Gender`)%>% summarise(Mean=mean(`Chol`,na.rm=TRUE),
                                                 Min = min(`Chol`,na.rm = TRUE),
                                                 Q1 = quantile(`Chol`,probs = .25,na.rm = TRUE),
                                                 Median = median(`Chol`,na.rm = TRUE),
                                                 Q3 = quantile(`Chol`,probs = .75,na.rm = TRUE),
                                                 Max= max(`Chol`,na.rm = TRUE),
                                                 Mean = mean(`Chol`,na.rm = TRUE),
                                                 SD = sd(`Chol`,na.rm = TRUE),
                                                 IQR = (Q3-Q1))
```

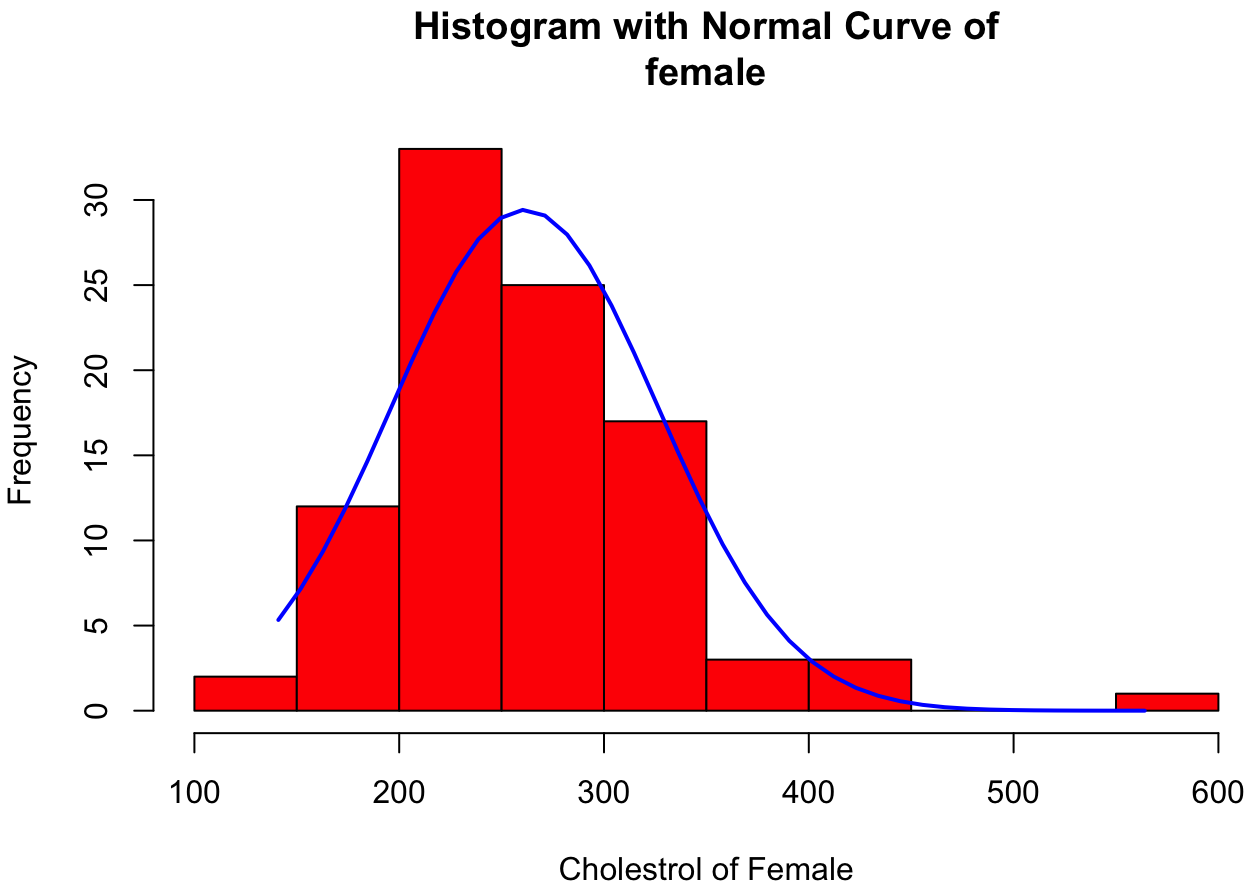
Gender	Mean	Min	Q1	Median	Q3	Max	SD	IQR
<fct>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>
female	261.3021	141	214.75	253	296.75	564	65.08895	82
male	239.2899	126	208.00	235	268.00	353	42.78239	60

2 rows

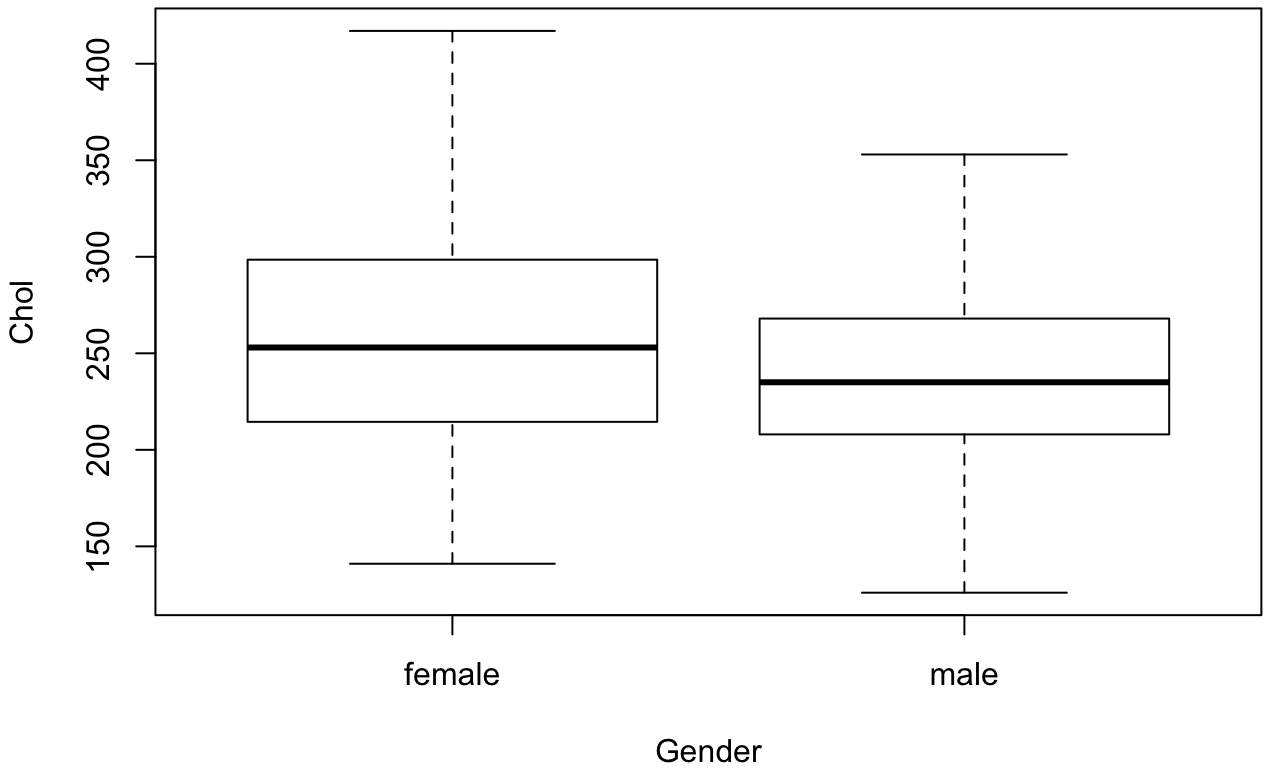
```
## We could see that the mean value of Chol is higher in females as compared to males .Now, we wi
ll visualise this


x<-heart_subset_male$Chol
h<-hist(x,breaks=10 ,col="yellow",xlab = "Cholestrol of Male",main="Histogram with normal
curve of male")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit<-yfit*diff(h$mids[1:2])*length(x)
lines(xfit,yfit,col="blue",lwd=2)
```



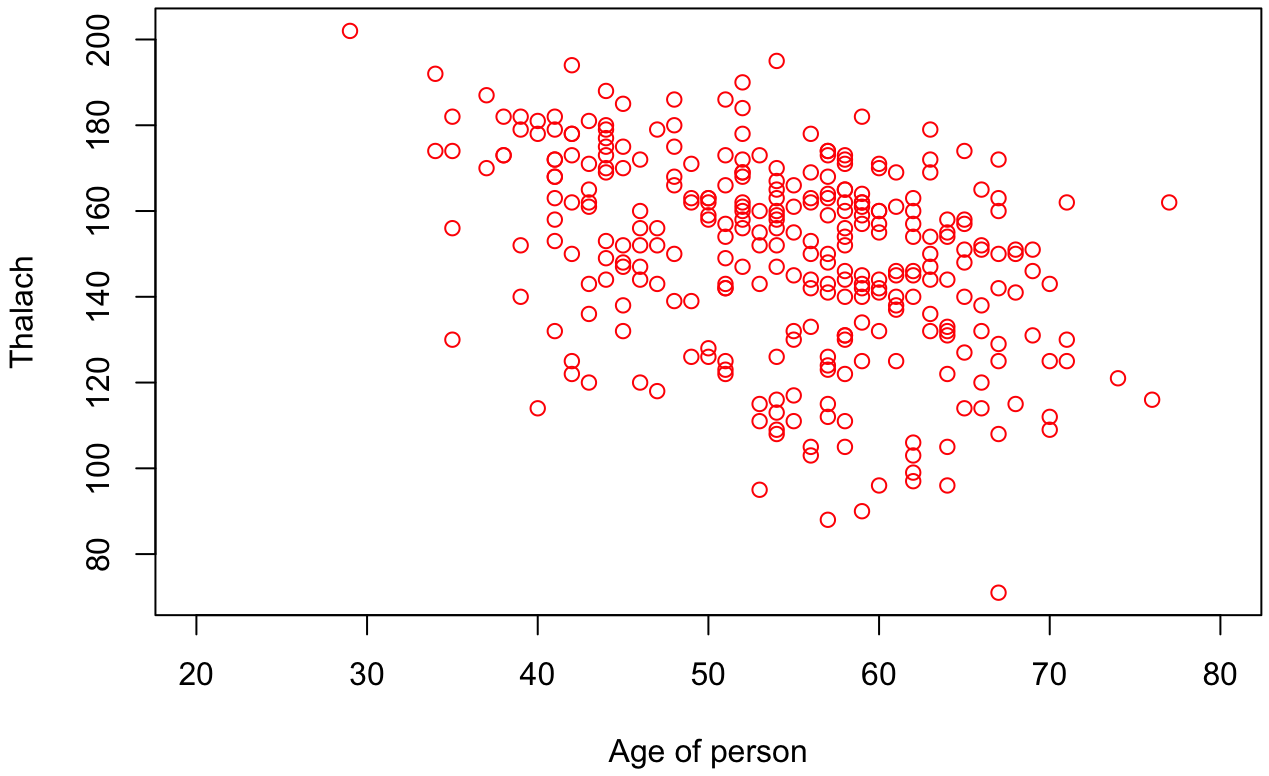

```
## We can see that cholesterol level in males follow a normal distribution.

y<-heart_subset_female$Chol
h<-hist(y,breaks = 10,col="red",xlab="Cholestrol of Female",main="Histogram with Normal Curve of
female")
xfit<-seq(min(y),max(y),length=40)
yfit<-dnorm(xfit,mean=mean(y),sd=sd(y))
yfit<-yfit*diff(h$mids[1:2])*length(y)
lines(xfit,yfit,col="blue",lwd=2)
```




```
## In females, the cholesterol levels are not normal distributed.

## Now, creating a boxplot between the two variables to visualise the trends:

boxplot(`Chol`~`Gender`,data=heart,outline=FALSE)
```

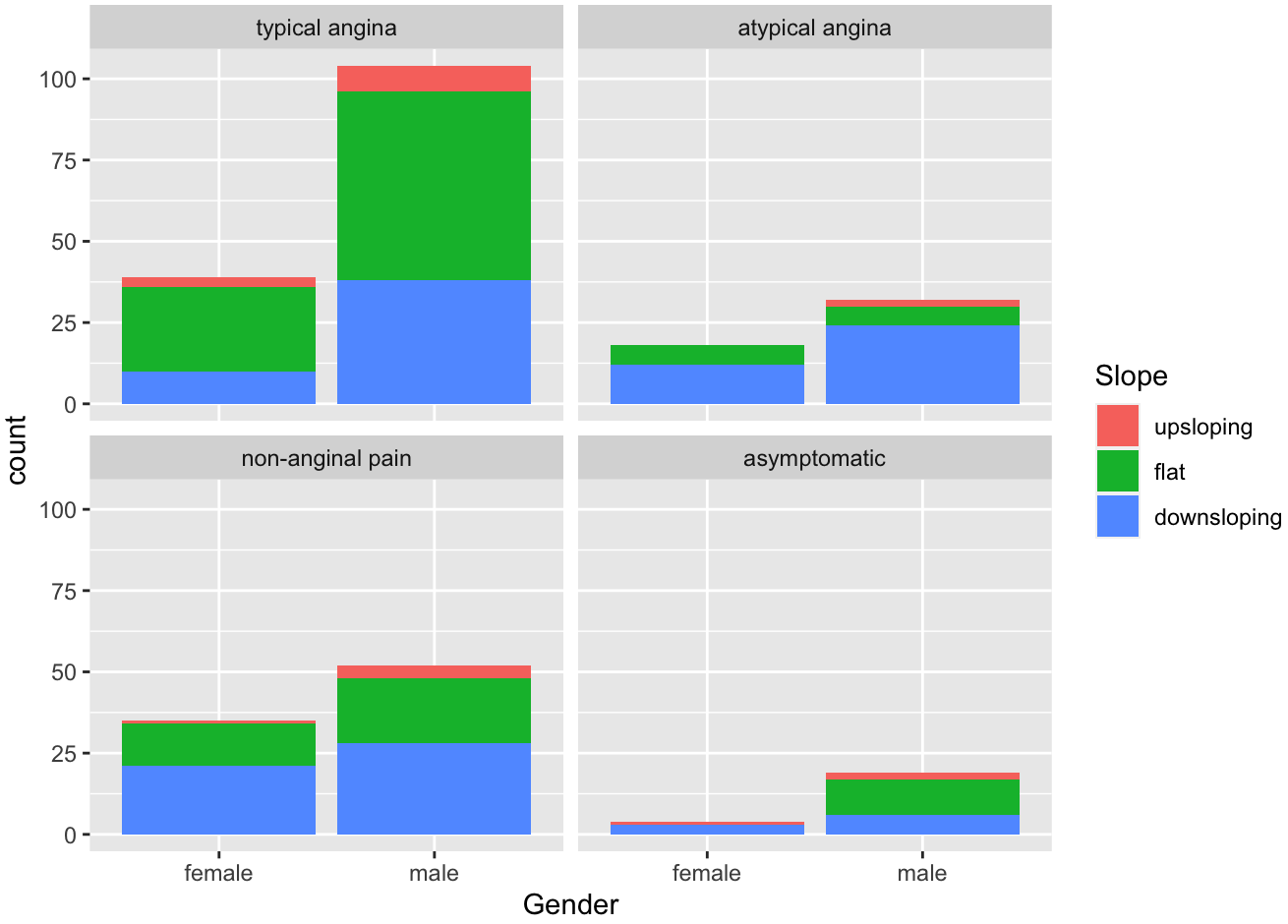



```
## We, could now clearly see that females have higher cholesterol levels than males.



## 3. Scatterplot between age and maximum heartrate achieved(Thalach)
plot(heart$Age,heart$Thalach,xlab = "Age of person",ylab = "Thalach",main = "Relationship between
     age of a person and maximum heart rate acheived",xlim=c(20,80), col='red')
```




```
## We, could observe clearly that higher the age , higher are the rate of maximum heart rate achie
ved as more density in the region between 50 to 70 years




## 4.Density plot between Thalach and heart disease occurence
p1<-ggplot(heart, aes(x = Thalach, fill=Heart)) +
  geom_density(alpha=0.15) +
  xlab("Maximum hear rate") +
  ylab("Count") +
  ggtitle("Relation of heart rate with heart disease") +
  scale_fill_discrete(name = "Heart disease", labels = c("No", "Yes"))
p1
```


```
## We could clearly observe that maximum heart achieved is more likely to increase the chances of
 having a heart disease.
## Earlier, we have seen people having higher age (50-70) are having higher heart rate and thus ar
e more prone to develop a heart disease.



## 5.Creating a clustered bar chart between CP,gender and slope
p <- ggplot(heart, aes(Gender, fill = Slope)) + facet_wrap(~CP)
p + geom_bar()
```




```
## By looking at the clustered bar chart, we can see that typical angina type of chest pain is mos
t frequent and is more common in males as compared to females and also which contains flat slope E
CG value as the highest in males


## Changing to numeric for further classification:

heart$Slope<-as.numeric(heart$Slope)
heart$CP<-as.numeric(heart$CP)
heart$Recg<-as.numeric(heart$Recg)
heart$Fbs<-as.numeric(heart$Fbs)
heart$Exang<-as.numeric(heart$Exang)
```

6. STATISTICAL MODELLING Using Logistic Regression:

We will use logistic regression model to infer the effect of explanotary variables against the response variable and would hence deduce the factors linked to heart diseases.

Logistic Regression 1 : Taking Heart as response variables and Tbps, Chol, Op & Thalach as predictors:

```
mod.fit1<- glm(formula = Heart ~ TBps + Chol + Op + Thalach , family  = binomial, data = heart)
summary(mod.fit1)
```

```
##
## Call:
## glm(formula = Heart ~ TBps + Chol + Op + Thalach, family = binomial,
##     data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1750  -0.8147   0.4943   0.8584   2.4062
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.026686   1.512901  -1.340    0.180
## TBps        -0.011285   0.008084  -1.396    0.163
## Chol        -0.003128   0.002638  -1.186    0.236
## Op          -0.713544   0.141671  -5.037 4.74e-07 ***
## Thalach      0.034751   0.006807   5.105 3.31e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 322.08  on 298  degrees of freedom
## AIC: 332.08
##
## Number of Fisher Scoring iterations: 4
```

```
## We could see that two predictors Thalach and Op including the intercept are statistically signi
ficant whereas Chol and TBps are statistically insignificant.
## Also, the |z|-statistics of the Thalach and Op are greater than 2, which proves to be statistic
ally significant.

## Checking the Model Adequacy:
## Finding the p-value for the goodness of fit test
deviance(mod.fit1)
```

```
## [1] 322.0776
```

```
pchisq(mod.fit1$deviance, df=mod.fit1$df.residual, lower.tail=FALSE)
```

```
## [1] 0.1615728
```

```
### The chi-square test statistic of 322.07 with 298 degrees of freedom gives a p-value of 0.161,
 indicating that the null hypothesis is plausible, and we can conclude that logistic model is adeq
uate for the prediction of the Heart disease


## Also,Comparing mod.fit1 with the null(intercept only) model:
Null<-glm(formula= Heart~1,family=binomial,data=heart)

anova(Null,mod.fit1, test="Chisq")
```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	302	417.6381	*NA*	*NA*	*NA*
2	298	322.0776	4	95.56043	8.660847e-20

2 rows

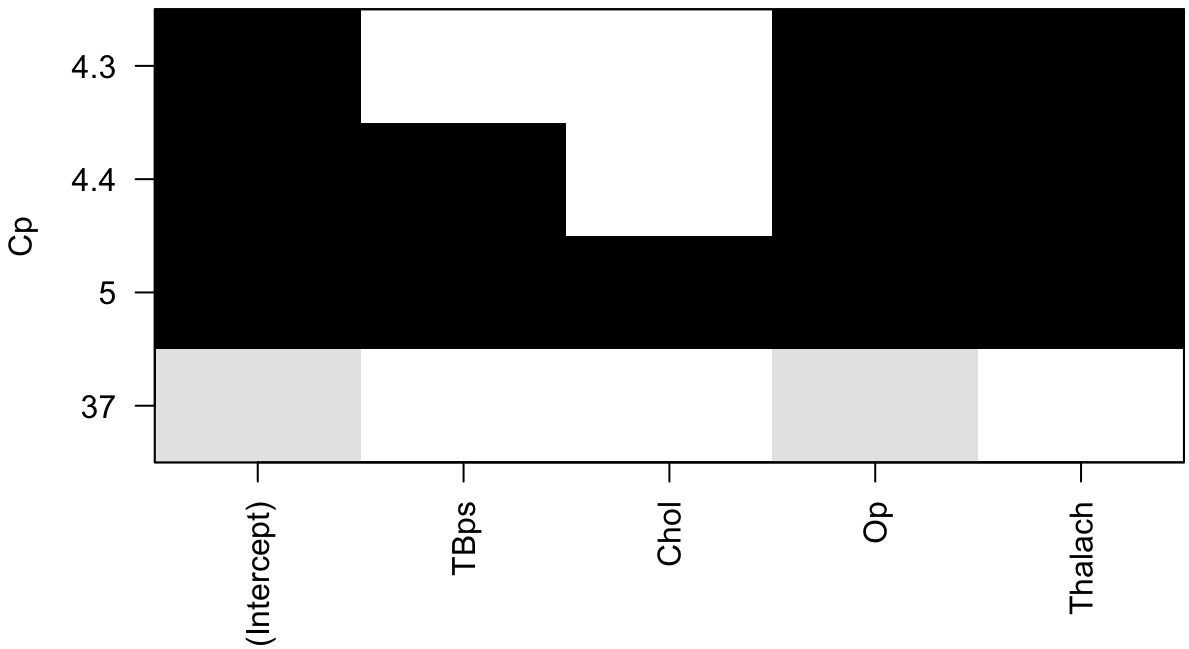
```
### Now, by comparing both models, we could say that the model is significant therefore rejecting
 Null model as we can also see that the Residual Deviation is improved by adding the predictors to
the Null model


## Checking the multicollinearity:
vif(mod.fit1)
```

```
##     TBps     Chol       Op  Thalach
## 1.017541 1.018546 1.031425 1.038719
```

```
## By looking at the vif values, we could say that multicollinearity is not present in the model f
it


## Finding best regression subsets for the model:
r<-leaps::regsubsets(Heart ~ TBps + Chol + Op + Thalach, data=heart)
summary(r)
```

```
## Subset selection object
## Call: regsubsets.formula(Heart ~ TBps + Chol + Op + Thalach, data = heart)
## 4 Variables  (and intercept)
##         Forced in Forced out
## TBps        FALSE      FALSE
## Chol        FALSE      FALSE
## Op          FALSE      FALSE
## Thalach     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          TBps Chol Op  Thalach
## 1  ( 1 ) " "  " "  "*" " "
## 2  ( 1 ) " "  " "  "*" "*"
## 3  ( 1 ) "*"  " "  "*" "*"
## 4  ( 1 ) "*"  "*"  "*" "*"
```

```
par(mfrow=c(1,1))
plot(r, scale="Cp")
```




```
## By looking at the plot, we can deduce that model with Op and Thalach present the best possible
 subset for this regression model in the prediction of heart disease

## BEST MODEL:
best_model1 = glm(Heart ~ Op + Thalach, family  = binomial, data=heart)
summary(best_model1)
```

```
##
## Call:
## glm(formula = Heart ~ Op + Thalach, family = binomial, data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2611  -0.8266   0.5198   0.8535   2.1924
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.130000   1.045559  -3.950 7.81e-05 ***
## Op          -0.735820   0.139896  -5.260 1.44e-07 ***
## Thalach      0.033875   0.006723   5.039 4.69e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 325.88  on 300  degrees of freedom
## AIC: 331.88
##
## Number of Fisher Scoring iterations: 4
```

```
## Confidence Interval
confint.default(best_model1)
```

```
##                  2.5 %      97.5 %
## (Intercept) -6.1792572 -2.08074285
## Op          -1.0100107 -0.46162886
## Thalach      0.0206975  0.04705182
```

```
## Odds ratio for Op
exp(best_model1$coefficients[2])
```

```
##        Op
## 0.4791125
```

```
1/exp(1*best_model1$coefficients[2])
```

```
##       Op
## 2.087192
```

```
## With 1 unit increase in the Op, the estimated odds of success changes by 2.087 times.


## Odds ratio for Thalach
exp(best_model1$coefficients[3])
```

```
##  Thalach
## 1.034455
```

```
1/exp(10*best_model1$coefficients[3])
```

```
##  Thalach
## 0.712663
```

```
## With 10 unit increase in the Thalach, the estimated odds of success changes by 0.712 times.
```

Thus, we could say that Op and Thalach are statistically significant in the prediction of heart diseases.

Logistic Regression 2: Taking Heart as response variables and factor variables Thal, Recg, Slope, CP & Fbs as predictors:

```
mod.fit2<- glm(formula = Heart ~ Thal + Recg + Slope + CP + Fbs , family  = binomial, data = hear
t)
summary(mod.fit2)
```

```
##
## Call:
## glm(formula = Heart ~ Thal + Recg + Slope + CP + Fbs, family = binomial,
##     data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7227  -0.7914   0.3610   0.7239   1.9352
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.0677     1.0458  -1.977   0.0480 *
## Thal         -1.1807     0.2385  -4.951 7.37e-07 ***
## Recg          0.4757     0.2727   1.745   0.0811 .
## Slope         1.2551     0.2465   5.092 3.55e-07 ***
## CP            0.9838     0.1523   6.461 1.04e-10 ***
## Fbs          -0.5411     0.4289  -1.262   0.2071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 293.18  on 297  degrees of freedom
## AIC: 305.18
##
## Number of Fisher Scoring iterations: 4
```

```
## We could see that predictors Thal, Slope and CP including the intercept are statistically signi
ficant whereas Recg is statistically near 0.05 but is partially insignificant in the prediction an
d Fbs is totally insignificant.
## Also, the |z|-statistics of the Thal, Slope and CP are greater than 2, which proves to be stati
stically significant.



## Again,checking the Model Adequacy:
## Finding the p-value for the goodness of fit test
deviance(mod.fit2)
```

```
## [1] 293.1758
```

```
pchisq(mod.fit2$deviance, df=mod.fit2$df.residual, lower.tail=FALSE)
```

```
## [1] 0.551796
```

```
### The chi-square test statistic of 293.175 with 297 degrees of freedom gives a p-value of 0.551,
indicating that the null hypothesis is plausible and thus we can conclude that logistic model is a
dequate for the prediction of heart disease


## Also,Comparing mod.fit2 with the null(intercept only) model:
Null<-glm(formula= Heart~1,family=binomial,data=heart)

anova(Null,mod.fit2, test="Chisq")
```

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	302	417.6381	*NA*	*NA*	*NA*
2	297	293.1758	5	124.4623	3.557652e-25

2 rows

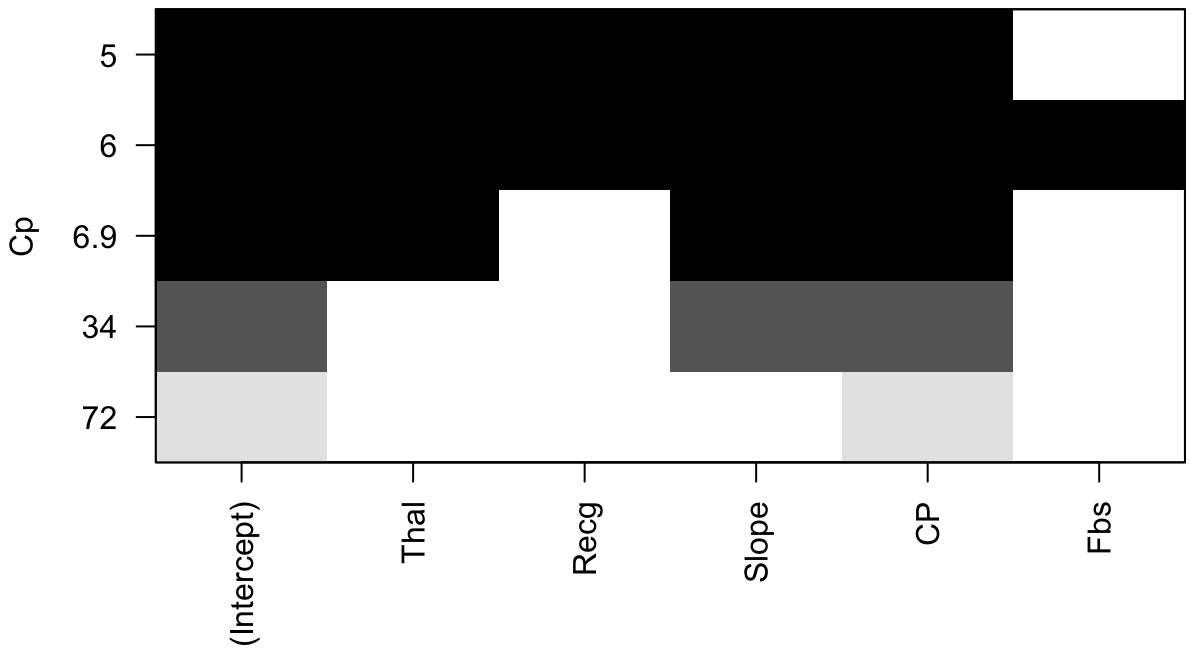
```
### Now, by comparing both models, we could say that the model is significant therefore rejecting
 Null model as we can also see that the Residual Deviation is improved by adding the predictors to
the Null model


## Checking the multicollinearity:
vif(mod.fit2)
```

```
##     Thal     Recg    Slope       CP      Fbs
## 1.028186 1.006827 1.037528 1.070165 1.045679
```

```
## By looking at the vif values, we could say that multicollinearity is not present in the model f
it

## Finding best regression subsets for the model:
r<-leaps::regsubsets(Heart ~ Thal + Recg + Slope + CP + Fbs, data=heart)
summary(r)
```

```
## Subset selection object
## Call: regsubsets.formula(Heart ~ Thal + Recg + Slope + CP + Fbs, data = heart)
## 5 Variables  (and intercept)
##       Forced in Forced out
## Thal      FALSE      FALSE
## Recg      FALSE      FALSE
## Slope     FALSE      FALSE
## CP        FALSE      FALSE
## Fbs       FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          Thal Recg Slope CP  Fbs
## 1  ( 1 ) " "  " "  " "   "*" " "
## 2  ( 1 ) " "  " "  "*"   "*" " "
## 3  ( 1 ) "*"  " "  "*"   "*" " "
## 4  ( 1 ) "*"  "*"  "*"   "*" " "
## 5  ( 1 ) "*"  "*"  "*"   "*" "*"
```

```
par(mfrow=c(1,1))
plot(r, scale="Cp")
```



```
## By looking at the plot, we can deduce that model with predictors Thal, CP , Slope including Rec
g which is partially insignificant is also included in the regression model whereas Fbs is not inc
luded

## BEST MODEL:
best_model2 = glm(Heart ~ Thal + Recg + Slope + CP, family  = binomial, data=heart)
summary(best_model2)
```

```
##
## Call:
## glm(formula = Heart ~ Thal + Recg + Slope + CP, family = binomial,
##     data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6755  -0.7699   0.3796   0.7441   1.9461
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.6938     0.9296  -2.898  0.00376 **
## Thal         -1.1613     0.2380  -4.880 1.06e-06 ***
## Recg          0.4950     0.2715   1.823  0.06827 .
## Slope         1.2503     0.2450   5.104 3.33e-07 ***
## CP            0.9566     0.1497   6.392 1.64e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 294.77  on 298  degrees of freedom
## AIC: 304.77
##
## Number of Fisher Scoring iterations: 4
```

```
## Confidence Interval
confint.default(best_model2)
```

```
##                   2.5 %     97.5 %
## (Intercept) -4.51577239 -0.8717829
## Thal        -1.62777231 -0.6948637
## Recg        -0.03711993  1.0270244
## Slope        0.77016642  1.7304622
## CP           0.66327486  1.2499400
```

```
## We, can see that Recg is the only predictor including zero between its confidence interval show
ing insignificance upto an extent.
```

Thus, we could say that CP, Slope and Thal are highly statistically significant in the prediction of heart diseases whereas Recg is partially significant and Fbs is insignificant

Logistic Regression 3: Taking Gender as response variable and CP, Heart, Chol & TBps as predictors:

```
mod.fit3 <- glm(formula = Gender ~ CP + Heart + Chol + TBps , family  = binomial, data = heart)
summary(mod.fit3)
```

```
##
## Call:
## glm(formula = Gender ~ CP + Heart + Chol + TBps, family = binomial,
##     data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1668  -1.0864   0.5474   0.8832   1.4132
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.573250   1.273173   4.377 1.20e-05 ***
## CP           0.222107   0.150474   1.476  0.13993
## Heartyes    -1.764917   0.335311  -5.264 1.41e-07 ***
## Chol        -0.010106   0.002838  -3.560  0.00037 ***
## TBps        -0.012559   0.007869  -1.596  0.11046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 378.42  on 302  degrees of freedom
## Residual deviance: 332.78  on 298  degrees of freedom
## AIC: 342.78
##
## Number of Fisher Scoring iterations: 4
```

```
## We could see that two predictors Heart and Chol including the intercept are statistically signi
ficant whereas TBps and CP are statistically insignificant in the prediction of Gender class.
## Also, the |z|-statistics of the Heart and Chol are greater than 2, which proves to be statistic
ally significant.



## Again,checking the Model Adequacy:
## Finding the p-value for the goodness of fit test
deviance(mod.fit3)
```

```
## [1] 332.7829
```

```
pchisq(mod.fit3$deviance, df=mod.fit3$df.residual, lower.tail=FALSE)
```

```
## [1] 0.08077479
```

The chi-square test statistic of 332.78 with 298 degrees of freedom gives a p-value of 0.080, indicating that the null hypothesis is plausible and thus we can conclude that logistic model is adequate for the prediction of Gender class

Also,Comparing mod.fit3 with the null(intercept only) model:

```
Null<-glm(formula= Gender~1,family=binomial,data=heart)

anova(Null,mod.fit3, test="Chisq")
```

Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>	Pr(>Chi) <dbl>
1	302	*NA*	*NA*	*NA*
2	298	332.7829	4	45.63874	2.928227e-09

2 rows

Now, by comparing both models, we could say that the model is significant therefore rejecting Null model as we can also see that the Residual Deviation is improved by adding the predictors to the Null model

Checking the multicollinearity:

```
vif(mod.fit3)
```

```
##       CP    Heart     Chol     TBps
## 1.308328 1.401037 1.073332 1.091347
```

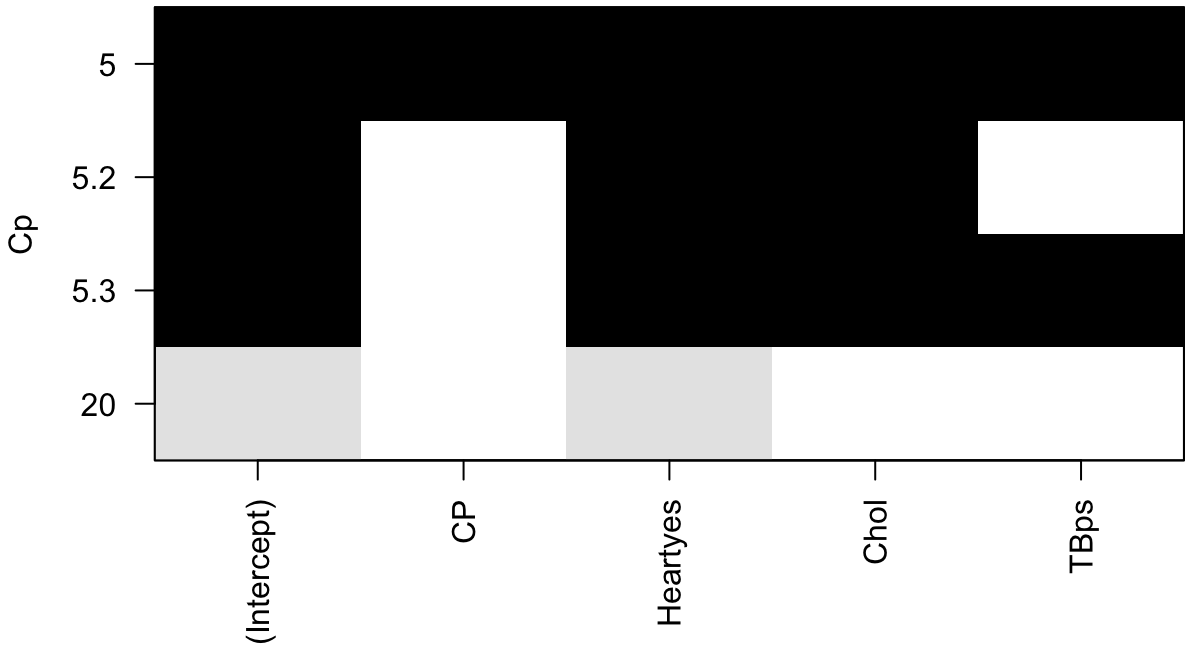
By looking at the vif values of predictors, we could say that multicollinearity is not present in the model fit

Finding best regression subsets for the model:

```
r<-leaps::regsubsets(Gender ~ CP + Heart + Chol + TBps, data=heart)
summary(r)
```

```
## Subset selection object
## Call: regsubsets.formula(Gender ~ CP + Heart + Chol + TBps, data = heart)
## 4 Variables  (and intercept)
##          Forced in Forced out
## CP           FALSE      FALSE
## Heartyes     FALSE      FALSE
## Chol         FALSE      FALSE
## TBps         FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          CP  Heartyes Chol TBps
## 1  ( 1 ) " " "*"      " "  " "
## 2  ( 1 ) " " "*"      "*"  " "
## 3  ( 1 ) " " "*"      "*"  "*"
## 4  ( 1 ) "*" "*"      "*"  "*"
```

```
par(mfrow=c(1,1))
plot(r, scale="Cp")
```




```
## By looking at the plot, we can deduce that model with all predictors including TBps and CP whic
h are insignificant are also included in the regression model but have no significance in predicti
on


## Confidence Interval
confint.default(mod.fit3)
```

```
##                   2.5 %       97.5 %
## (Intercept)  3.07787635  8.068624530
## CP          -0.07281657  0.517029659
## Heartyes    -2.42211381 -1.107720082
## Chol        -0.01566942 -0.004542981
## TBps        -0.02798159  0.002863062
```

```
## We, can see that CP and TBps are the predictors including zero between its confidence interval
 showing insignificance
```

Thus, we could say that Chol and heart disease are statistically significant in the prediction of Gender class and therefore has correlation.

7. CONCLUSION

After working on all the data visualization and three logistic models for the prediction of the heart disease phenomenon, we could deduce the following conclusions:

1. We could observe that maximum heart achieved is more likely to increase the chances of having a heart disease.
2. We can see that typical angina type of chest pain is most frequent and is more common in males as compared to females.
3. We could say that predictors such as Op, Thalach ,CP , Slope, Thal are highly statistically significant in the prediction of heart diseases.
4. Also, predictors such as Recg is partially significant in the analysis of heart disease prediction.
5. And predictors TBps, Chol and Fbs are insignificant in the heart disease prediction.
6. Also, we could see the relationship between gender and heart disease which is also a factor in the analysis and is highly predictable. Also Chol levels in females is higher as compared to males is proved in the third logistic model showing relationship between gender and Chol.

8. REFERENCES

1. (2018). Retrieved 27 September 2020, from https://www.kaggle.com/ronitf/heart-disease-uci/metadata (https://www.kaggle.com/ronitf/heart-disease-uci/metadata)
2. https://rmit.instructure.com/courses/79640/assignments/567741 (https://rmit.instructure.com/courses/79640/assignments/567741)