Heart Attack Analysis_AS

Apurva Sharma & Rezwan Mazumdar

5/23/2021

```
library(readr)
library(tinytex)
library(tidyverse)
library(ggvis)
library(caTools)
library(rpart)
library(class)
library(rpart.plot)
library(ggplot2)
library(ggplot2)
library(ggcorrplot)
library(tibble)
library(purrr)
library(knitr)
library(tidyverse, warn.conflict=F)
library(ggdendro)
```

OBJECTIVE

- This report shall be representative of anlaysis of various predictors for heart attack
- We have secured the Data set from Kaggle(https://www.kaggle.com/johnsmith88/heart-disease-dataset) and the Data has the below attributes:

(i) 1025 patients have been studied for the primary symptoms which lead to heart attack/s
(ii) 13 predictors - Age,sex, chest pain, Resting ECG,Exercise induced Angina, cholesterol, Resting Blood Pressure, Fasting blood sugar, presence of major blood vessels, maximum heart rate achieved.
(iii) Description of variables:

- Age : Age of the patient
- Sex : Sex of the patient
- exang: exercise induced angina (1 = yes; 0 = no)
- ca: number of major vessels (0-3)
- cp : Chest Pain type chest pain type
- Value 1: typical angina
- Value 2: atypical angina

- Value 3: non-anginal pain
- Value 4: asymptomatic
- trtbps : resting blood pressure (in mm Hg)
- chol : cholestoral in mg/dl fetched via BMI sensor
- fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- rest_ecg : resting electrocardiographic results
- thal : normal(0), fixed defect(1) and reversible defect(2) Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach : maximum heart rate achieved
- target : 0 = less chance of heart attack 1 = more chance of heart attack
- The project includes data reading, data exploration & manipulation, 4 classification prediction models:

(i) KNN - Nearest Neighbor
(ii) Linear Regression Model
(iii) Decision Tree
(iv) XG Boost

Converting the data into dataframe and dropping all the NA's

```
heart = read_csv("/Users/apurvasharma/Downloads/heart (1).csv")
heart = na.omit(heart)
```

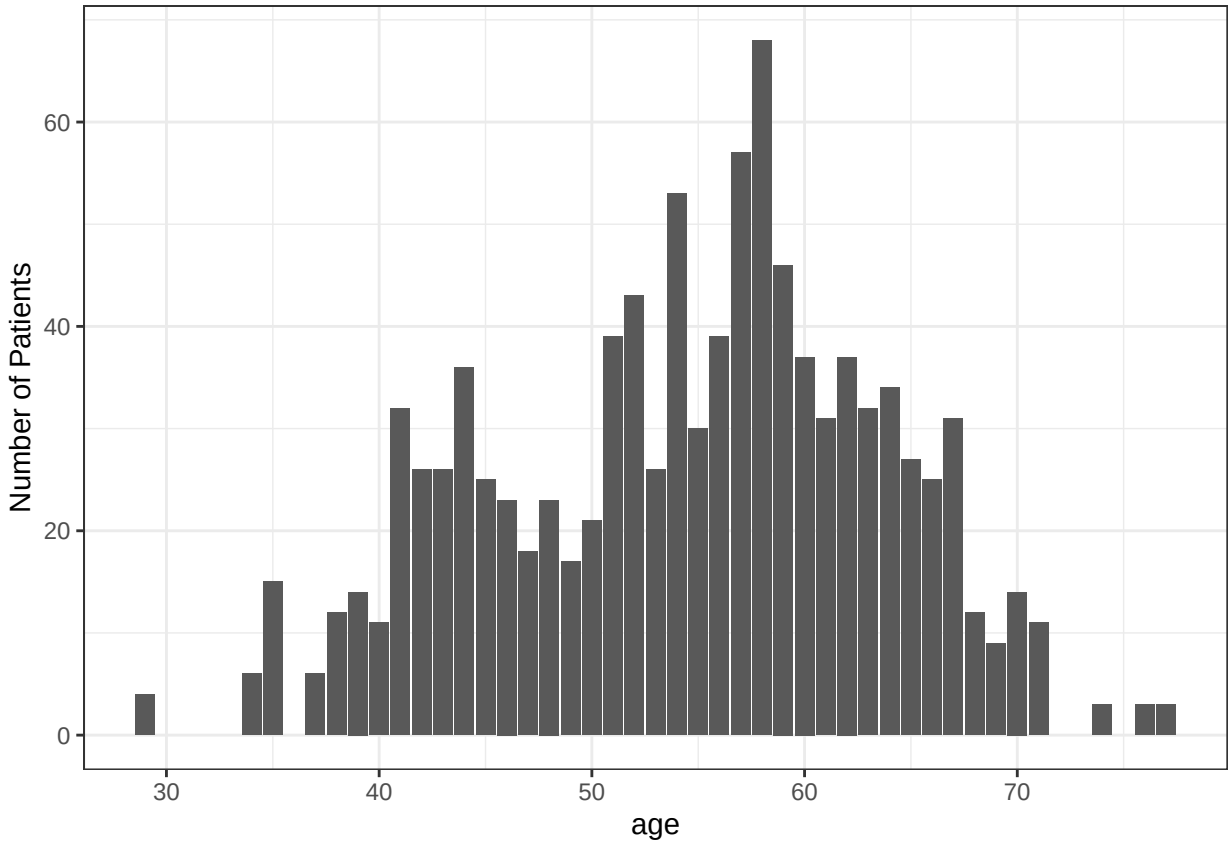
Creating duplicate datasets and changing columns with factorial

```
heart_a = heart
heart_a$disease = factor(heart_a$target,
                         levels=c(0,1),
                         labels=c("No","Yes"))
heart_a$sex = factor(heart_a$sex,
                     levels=c(0,1),
                     labels=c("Female","Male"))
```

Data Exploration

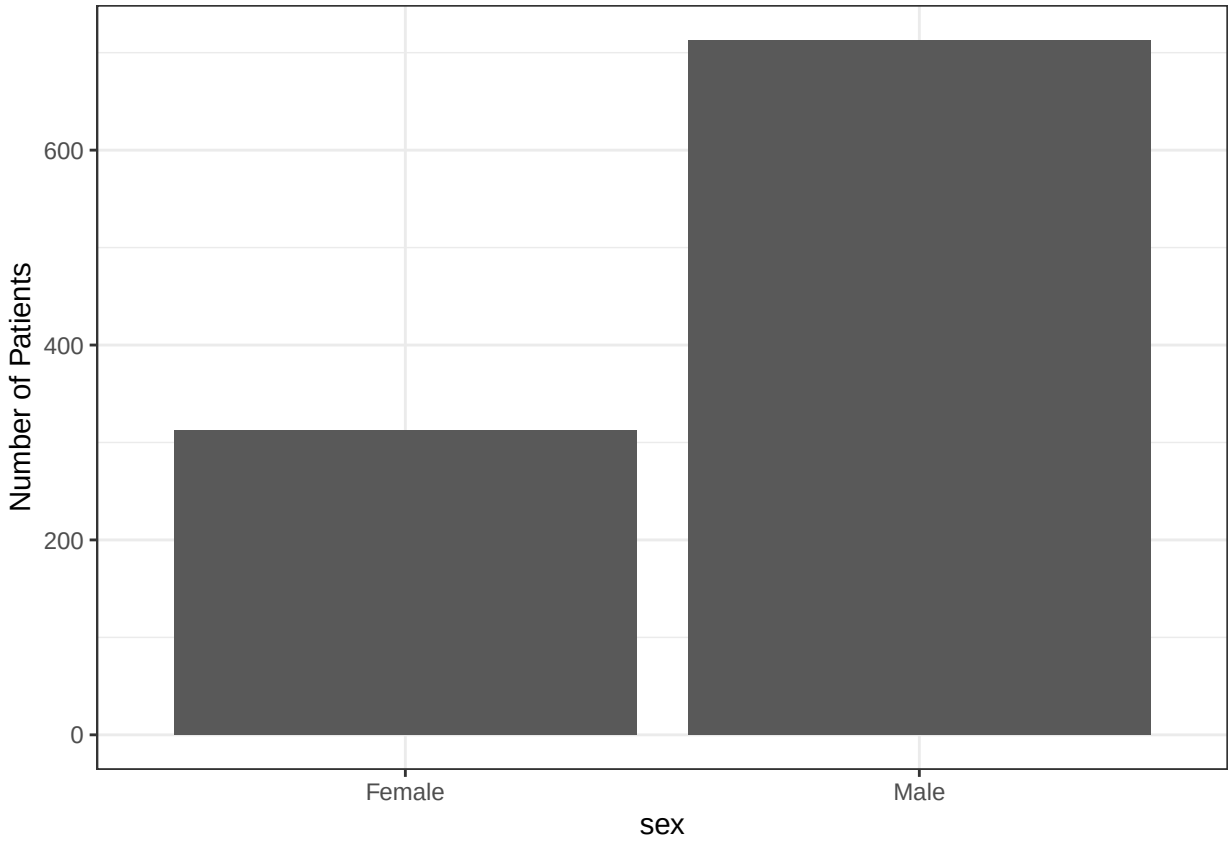
Age Group distribution

```
ggplot(heart_a, aes(x=age)) + theme_bw() + geom_bar()+ labs(y = 'Number of Patients')
```



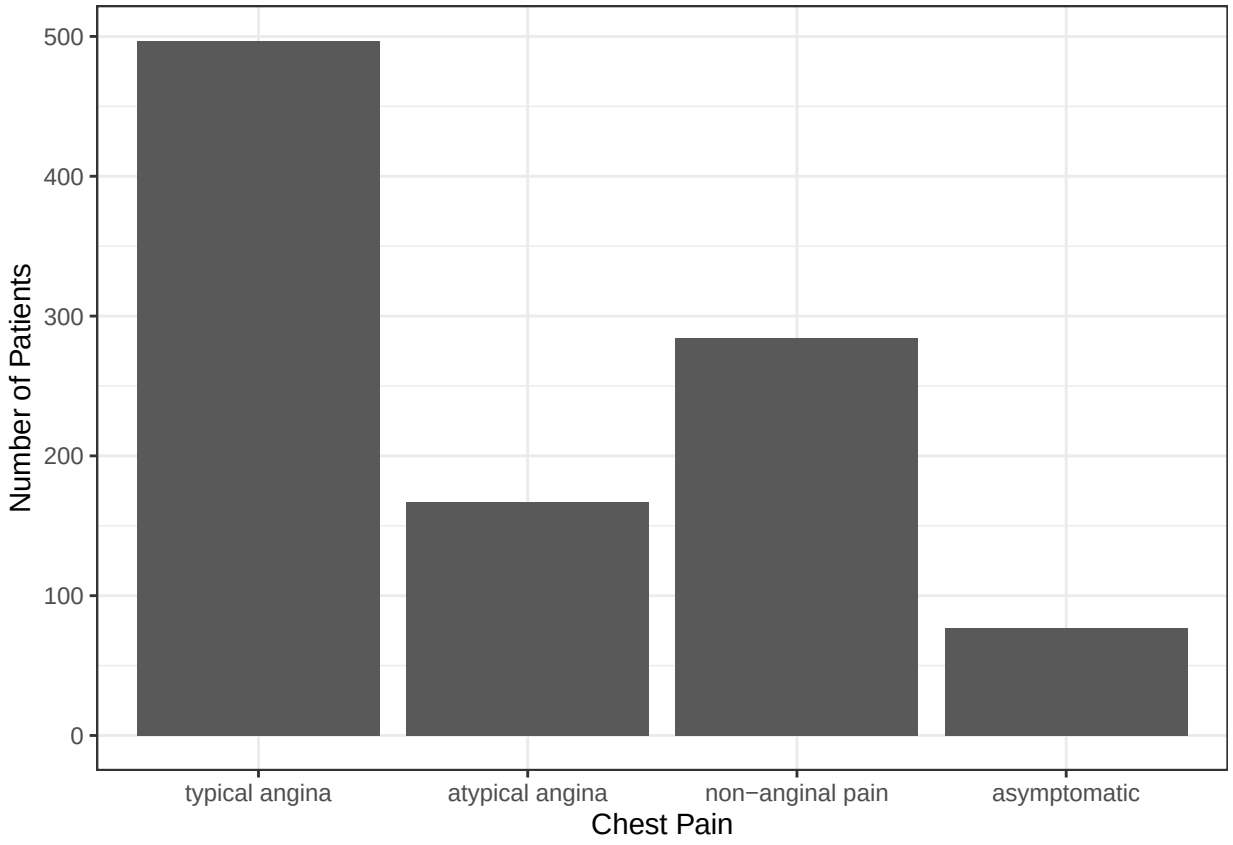
Gender Distribution

```
ggplot(heart_a, aes(x=sex)) + theme_bw() + geom_bar() + labs(y = 'Number of Patients')
```



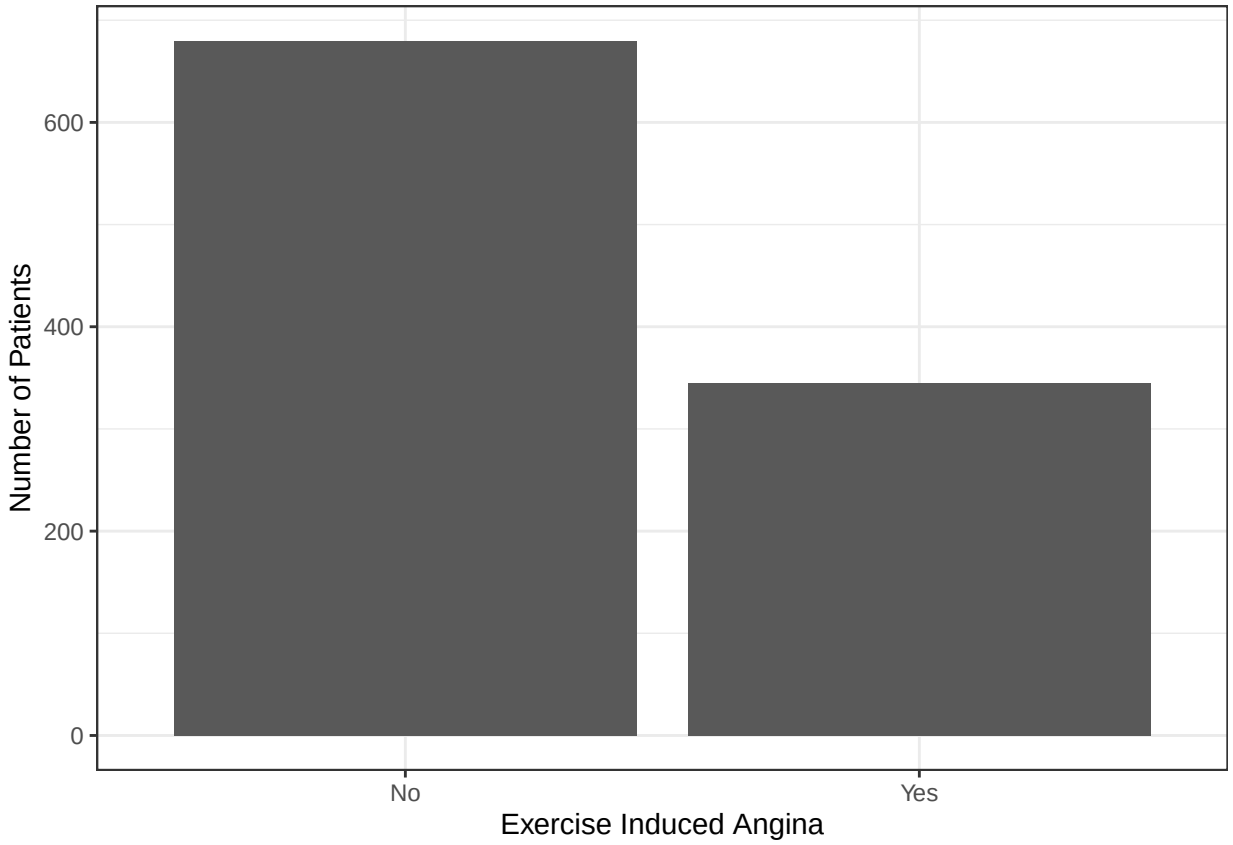
Chest Pain

```
heart_a$cp = factor(heart_a$cp,
                    levels=c(0,1,2,3),
                    labels=c("typical angina","atypical angina","non-anginal pain","asymptomatic"))
ggplot(heart_a, aes(x=cp)) + theme_bw() + geom_bar() + labs(y = 'Number of Patients', x = 'Chest Pain')
```



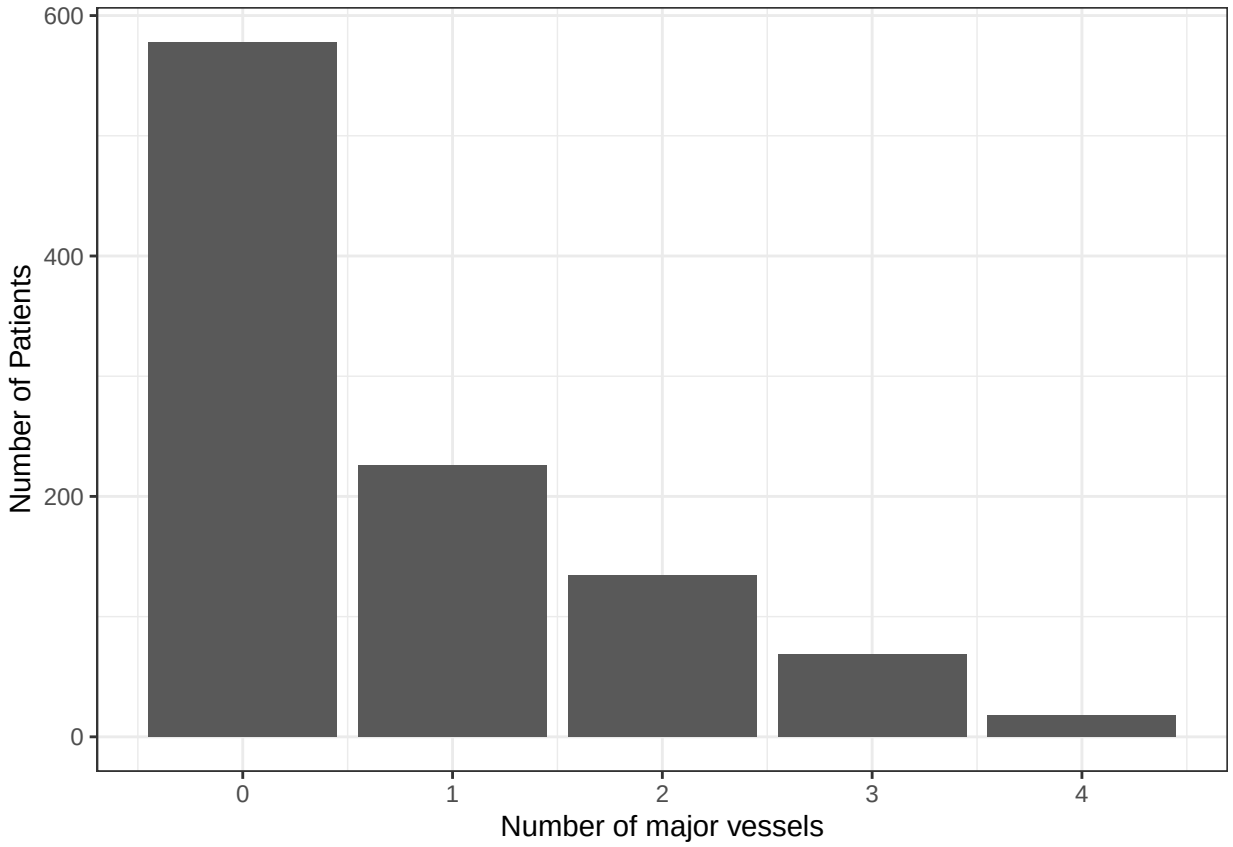
Exercise Induced Angina

```
heart_a$exang = factor(heart_a$exang,
                       levels=c(0,1),
                       labels=c("No","Yes"))
ggplot(heart_a, aes(x=exang)) + theme_bw() + geom_bar() + labs(y = 'Number of Patients', x = 'Exercise
```



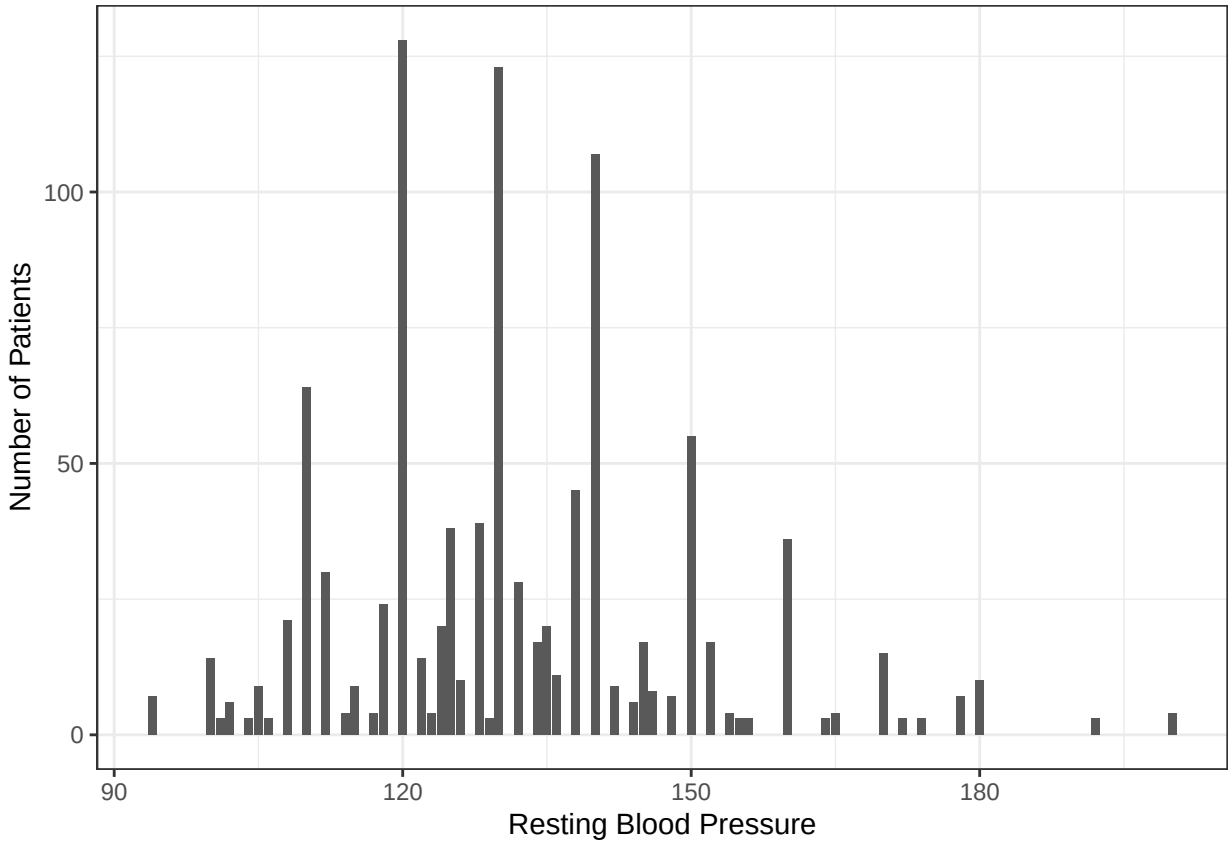
Number of major vessels

```
ggplot(heart_a, aes(x=ca)) + theme_bw() + geom_bar() + labs(y = 'Number of Patients', x = 'Number of ma
```



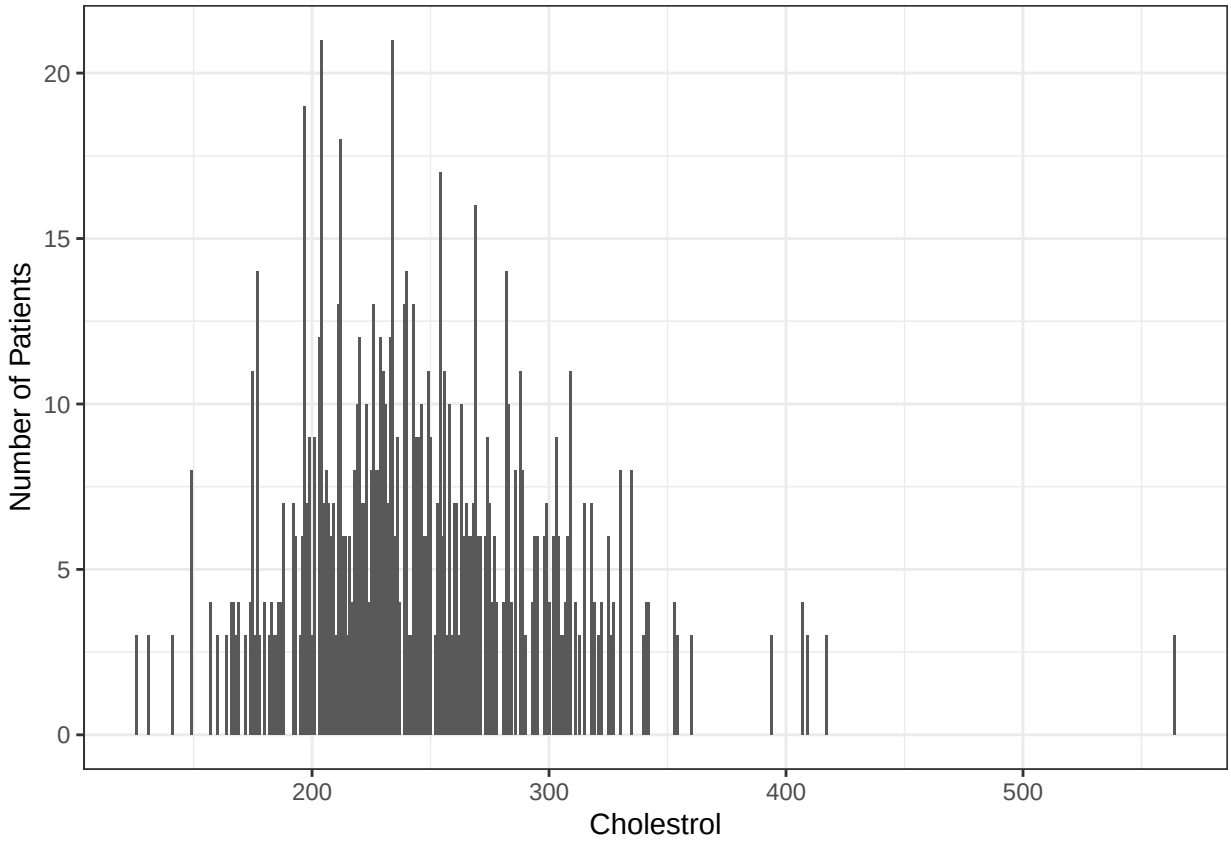
Resting Blood Pressure

```
ggplot(heart_a, aes(x = trestbps)) + theme_bw() + geom_bar() + labs(y = 'Number of Patients', x = 'Rest
```



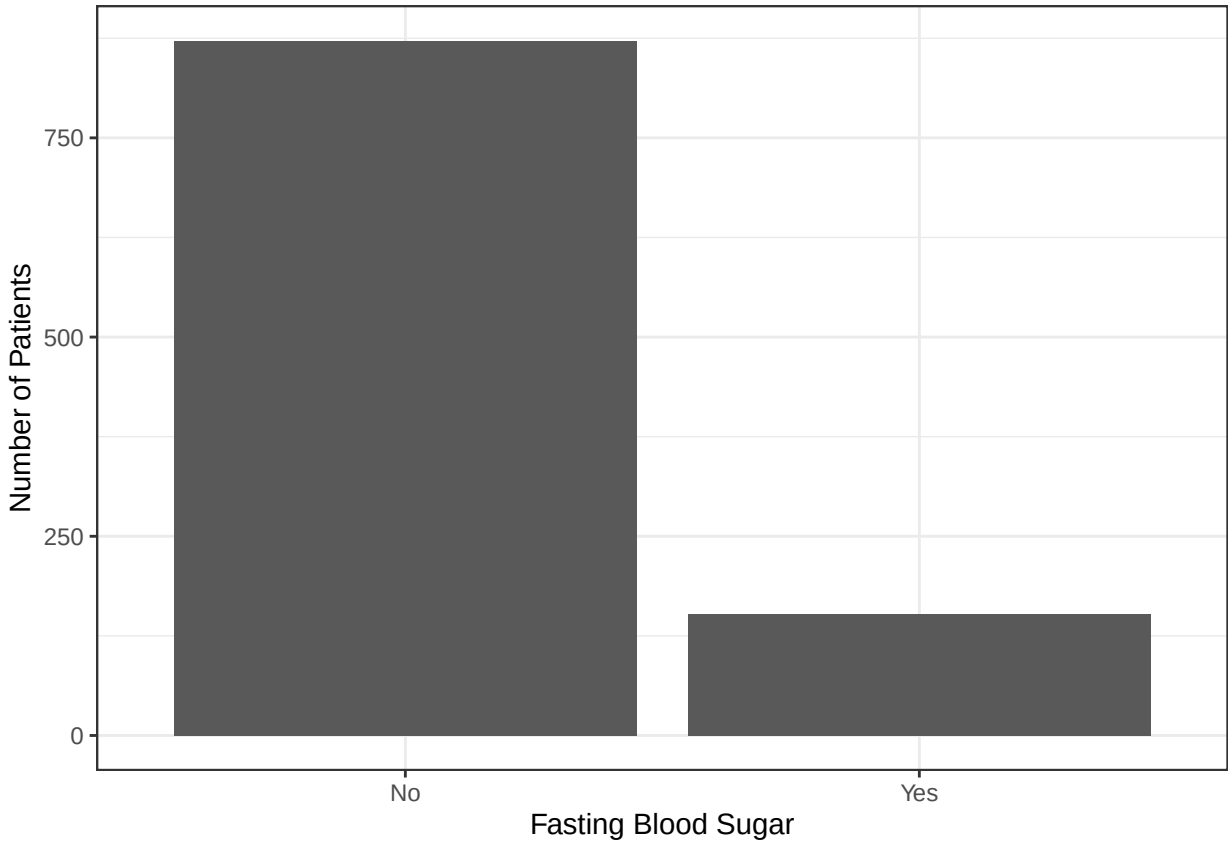
Cholestrol Levels

```
ggplot(heart_a, aes(x=chol),binwidth = 5) + theme_bw() + geom_bar() + labs(y = 'Number of Patients', x =
```



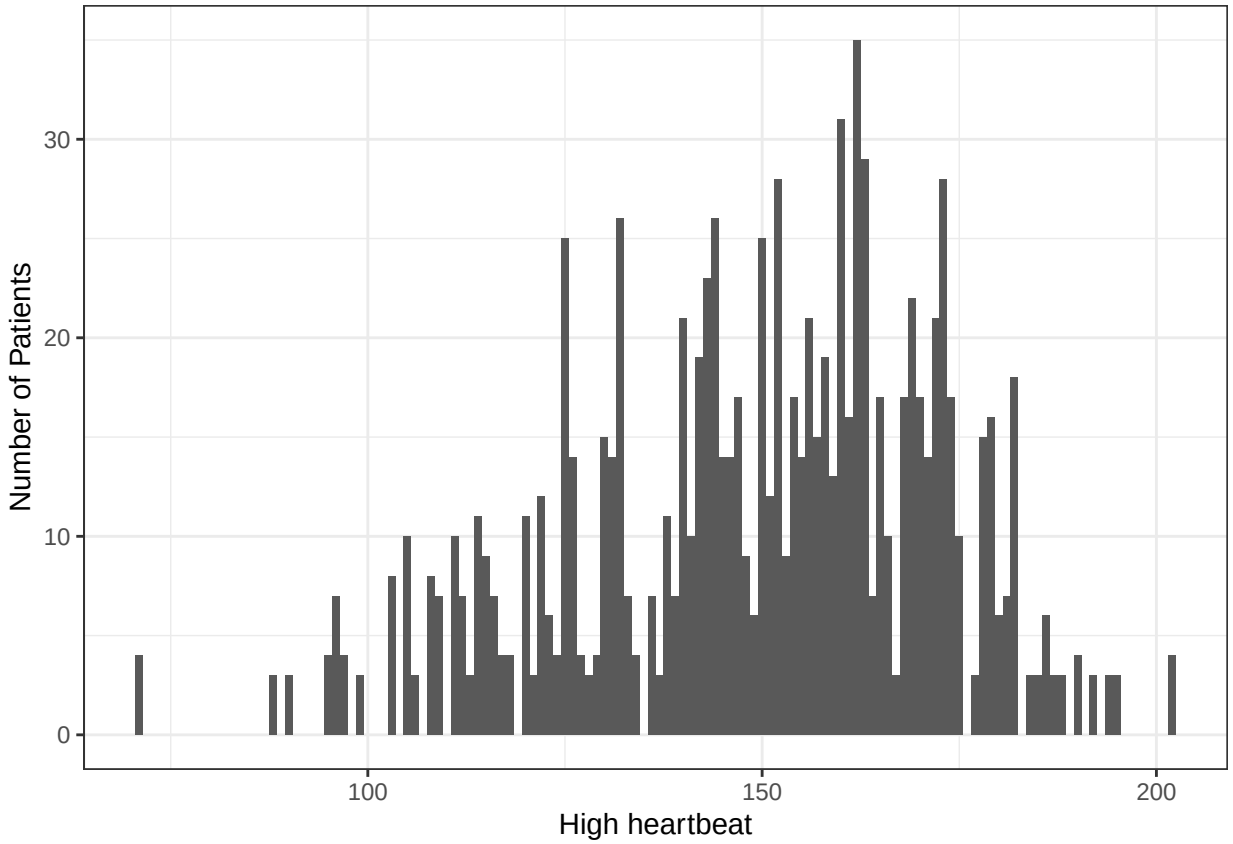
Fasting Blood Sugar

```
heart_a$fbs = factor(heart_a$fbs,
                     levels=c(0,1),
                     labels=c("No","Yes"))
ggplot(heart_a, aes(x=fbs)) + theme_bw() + geom_bar()+ labs(y = 'Number of Patients', x = 'Fasting Blood
```



Resting ECG

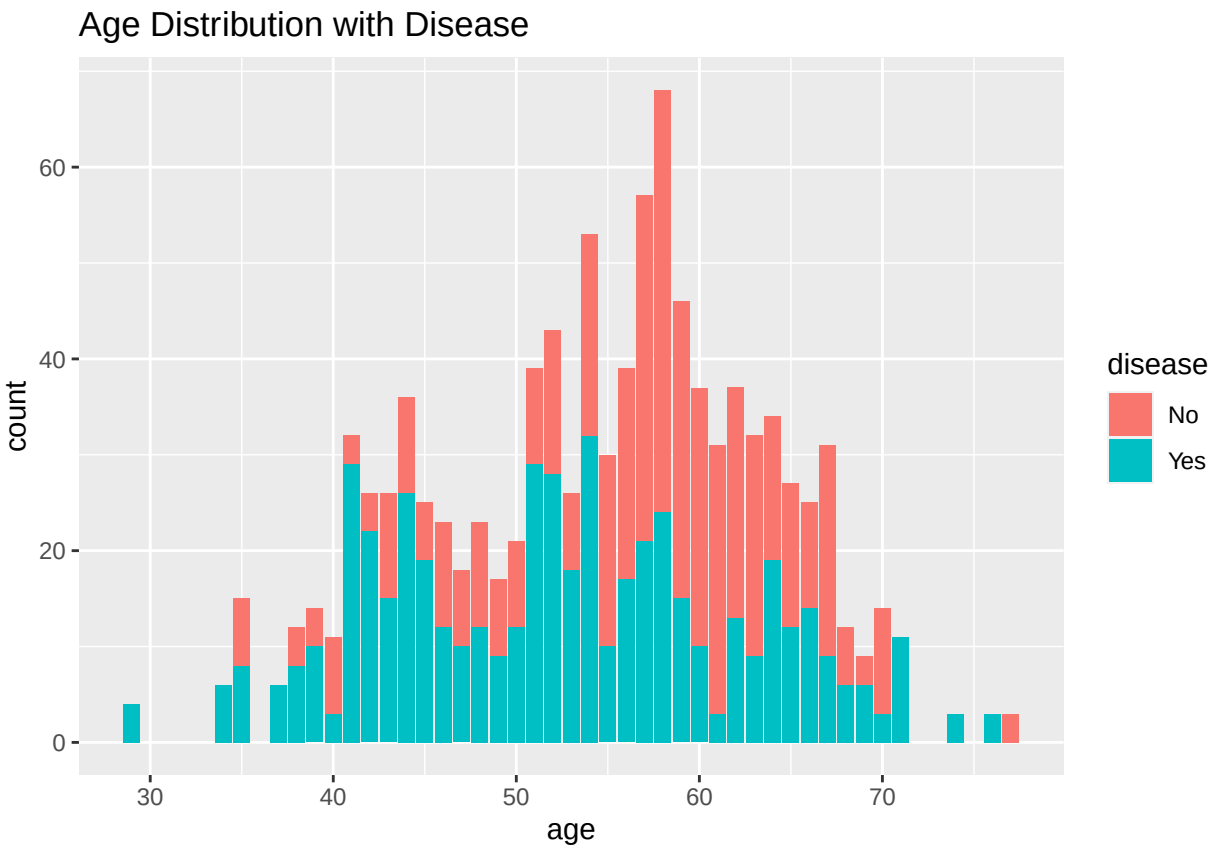
```
ggplot(heart_a, aes(x=thalach)) + theme_bw() + geom_bar()+ labs(y = 'Number of Patients', x = 'High hea
```



Exploratory Data Analysis

Age Distribution

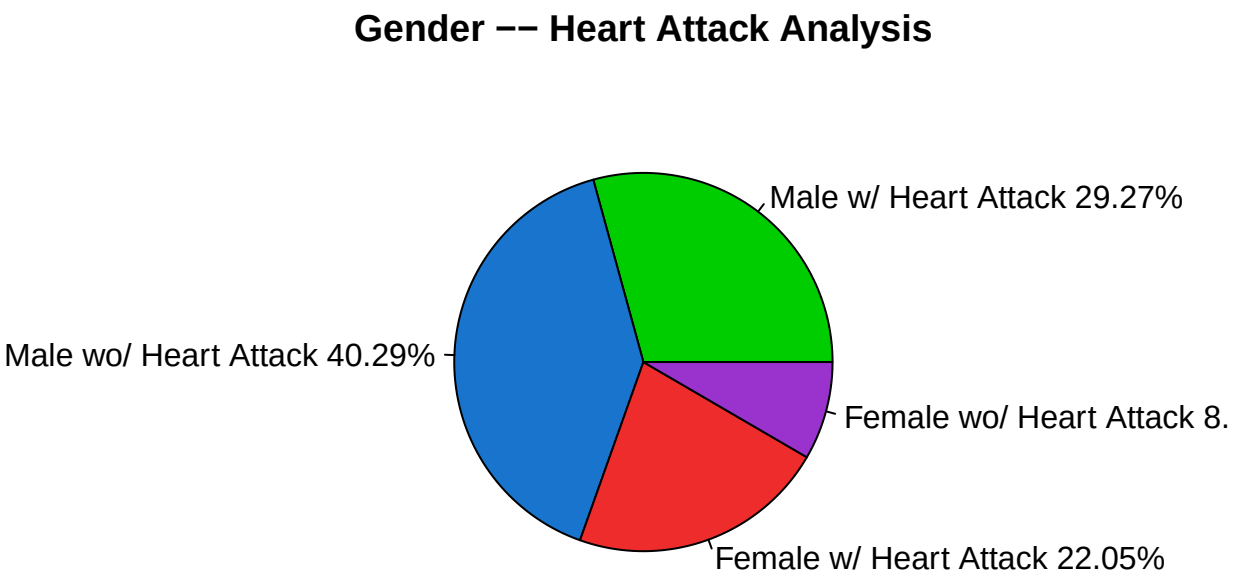
```
ggplot(heart_a) + geom_bar(aes(age, fill=disease), position = 'stack')+ggtitle("Age Distribution with D
```



- The Age group between 40-50 has been observed to be more susceptible to heart attack/s

Rate of Heart attacks for females and males

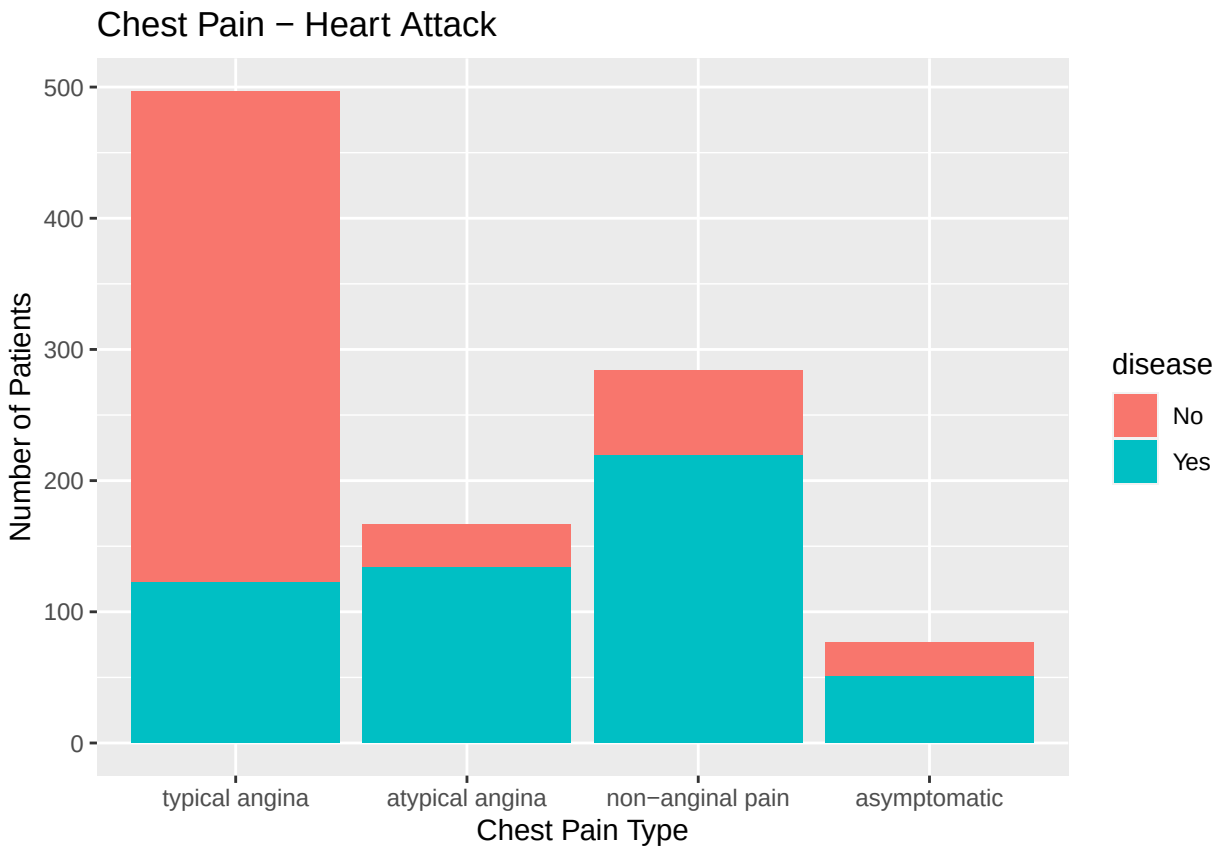
```
male = heart_a[heart_a$sex == "Male",]
female = heart_a[heart_a$sex == "Female",]

male_wDisease = male[male$disease == "Yes",]
male_woDisease = male[male$disease == "No",]
female_wDisease = female[female$disease == "Yes",]
female_woDisease = female[female$disease == "No",]

pie = data.frame(
  group = c("Male w/ Heart Attack", "Male wo/ Heart Attack", "Female w/ Heart Attack", "Female wo/ Hear
  value = c(NROW(male_wDisease$disease), NROW(male_woDisease$disease), NROW(female_wDisease$disease), NR
  )

pie = data.frame(
  group = c("Male w/ Heart Attack", "Male wo/ Heart Attack", "Female w/ Heart Attack", "Female wo/ Hear
  value = c(NROW(male_wDisease$disease), NROW(male_woDisease$disease), NROW(female_wDisease$disease), NR
  per = c((NROW(male_wDisease$disease)/sum(pie$value)), NROW(male_woDisease$disease)/sum(pie$value),NROW
)


lbl = paste(pie$group, round(pie$per, 2))
lbl = paste(lbl,"%", sep = "")

pie(pie$per, labels = lbl, col=c("green3", "dodgerblue3", "firebrick2", "darkorchid3"), main = "Gender -
```



- The pie chart indicates that out of the 1028 patients studied for heart attacks , 29% men had experienced heart attacks whereas Females with a history was heart attack was only 8.39% stating that men are at a higher risk of heart diseases and failure

Chest Pain as an indicator for Heart Disease

```
ggplot(heart_a)+geom_bar(aes(x= cp, fill = disease), position = "stack")+ggtitle("Chest Pain - Heart At
```



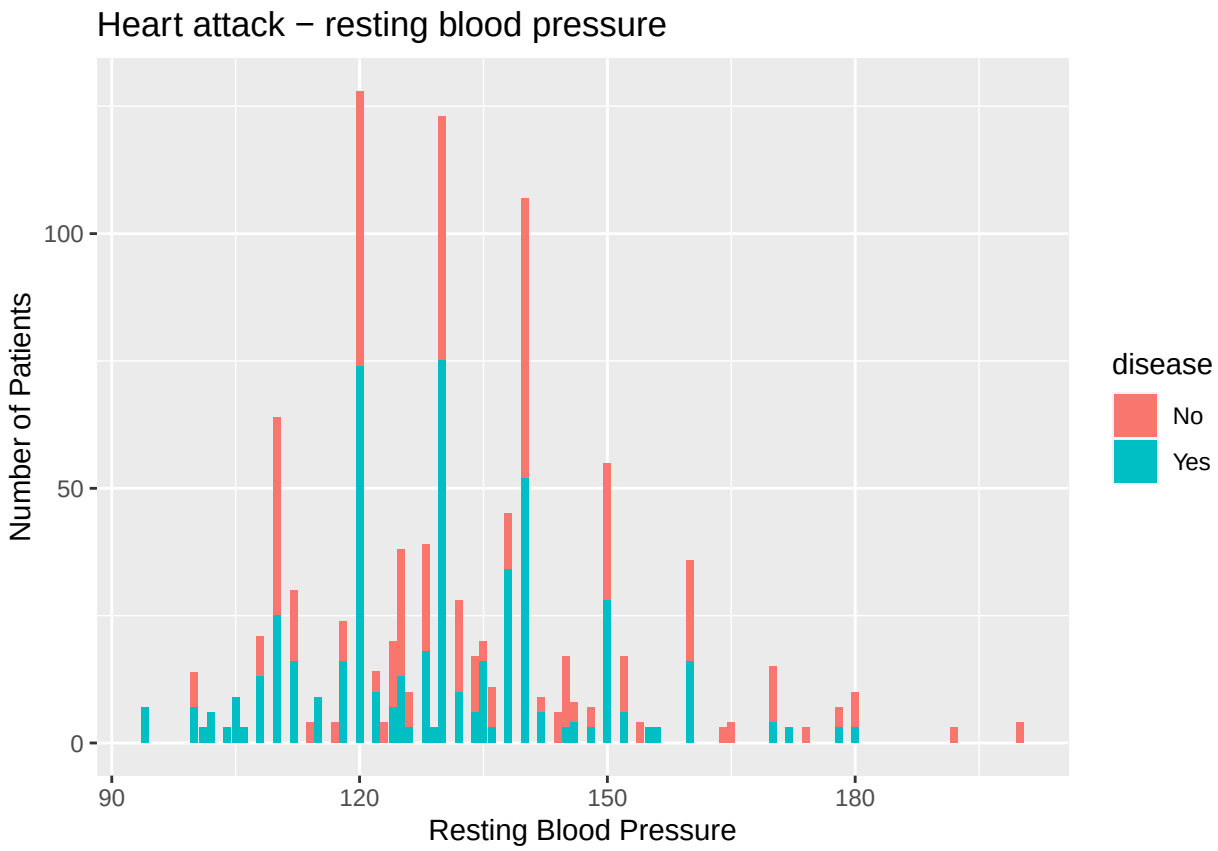
- Legends in the graph :

(i) asymptomatic (ii)typical angina
(ii) atypical angina
(iii) non-anginal pain

- Out of the 1028 patients studied approximately 110/500 patients who experienced heart attack had asymptomatic heart pain
- Approximately 130/160 with typical angina experienced heart attack/s
- Approximately 220/290 with atypical angina experienced heart attack/s
- Approximately 50/75 with non-anginal pain experienced heart attack/s, making chest pain an important factor to be considered while examining/diagnosing a heart attack

Resting Blood Pressure

```
ggplot(heart_a)+geom_bar(aes(x= trestbps, fill = disease), position = "stack")+ggtitle("Heart attack - r
```

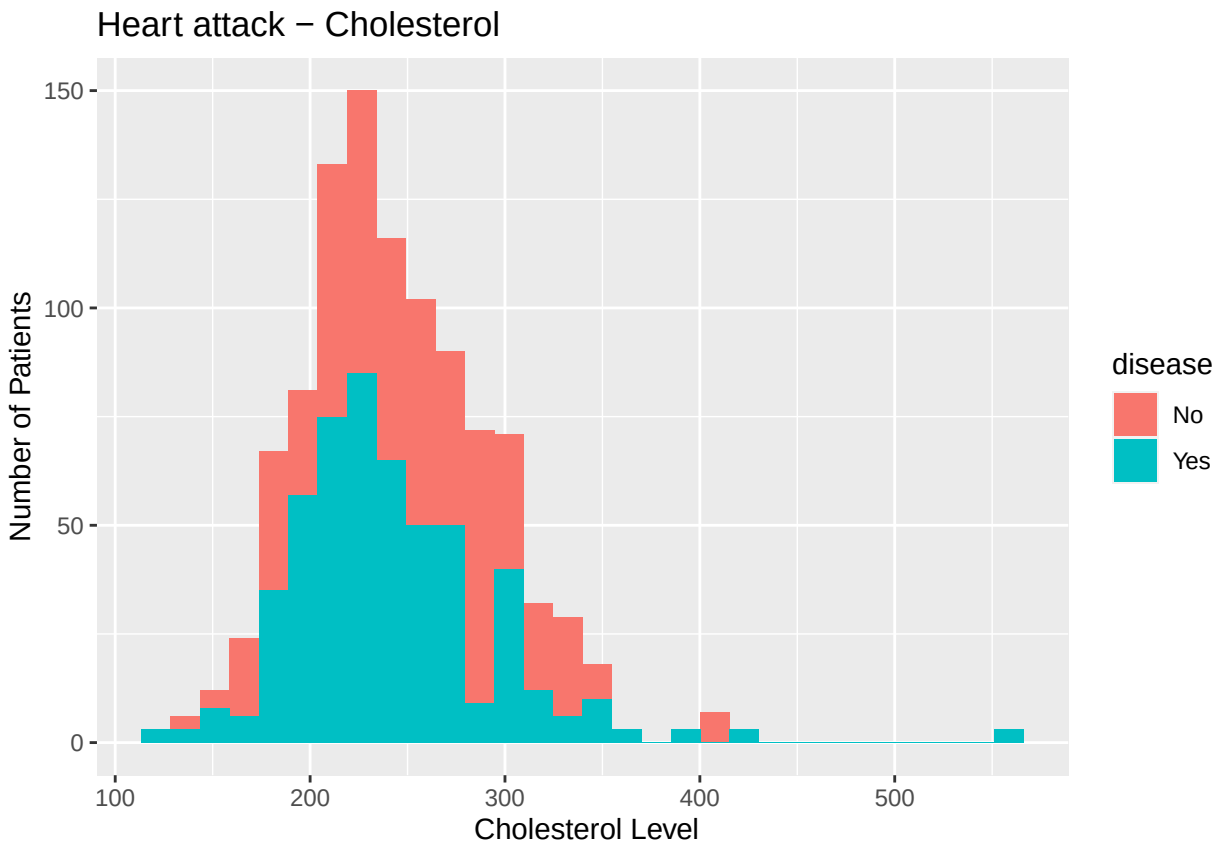


- Patients with a Blood Pressure in the range of 130-150 are at a higher risk of experiencing a heart attack

Cholesterol and heart attack

```
ggplot(heart_a) + geom_histogram(aes(chol, fill=disease), position = 'stack')+ggtitle("Heart attack - Cl
```

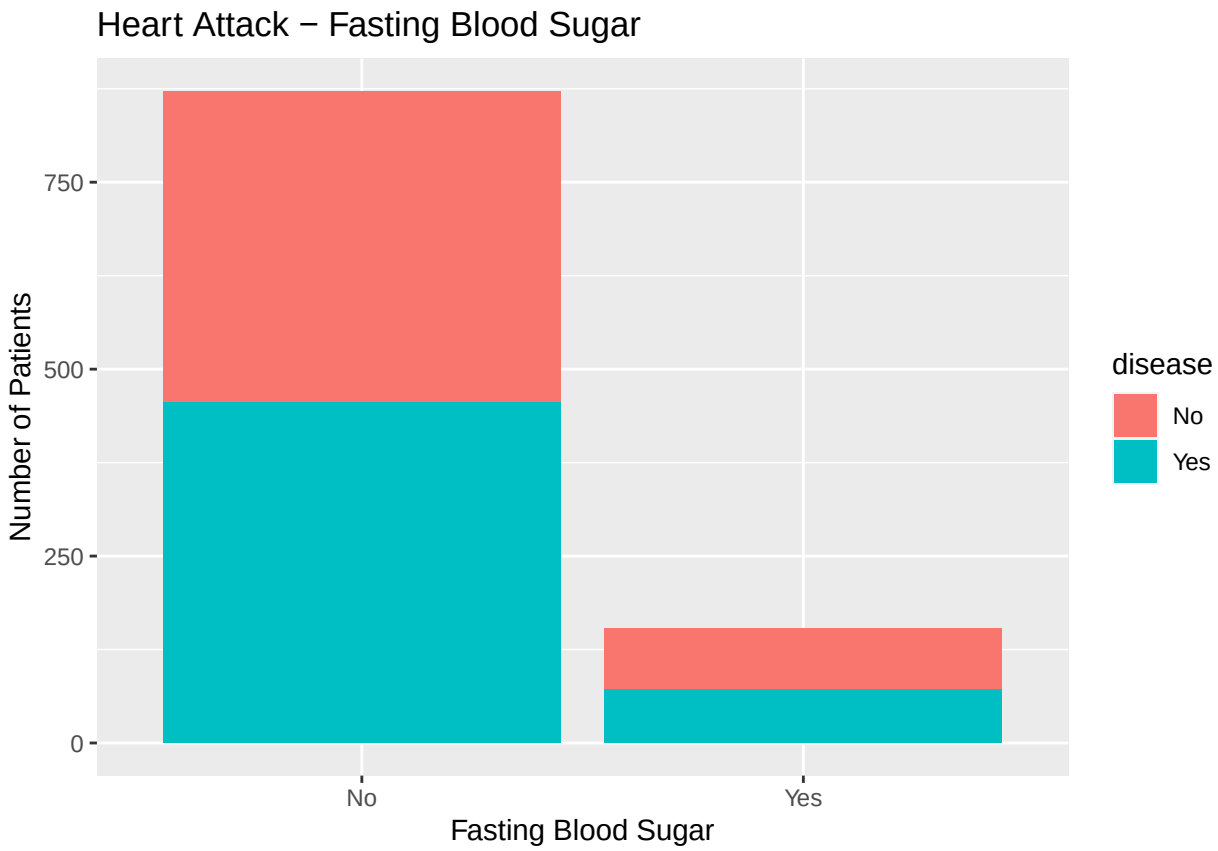
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- Patients with an observed Cholesterol from 200 mg/d and above showed highest share from the patients studied of experiencing a heart attack
- While all the patients with cholesterol with 350 mg/d and above were experiencing a heart attack with an exception of about 8 patients not experiencing one
- The above analysis makes high cholestrol levels as a key indicator of heart diseases

Fasting Blood Sugar

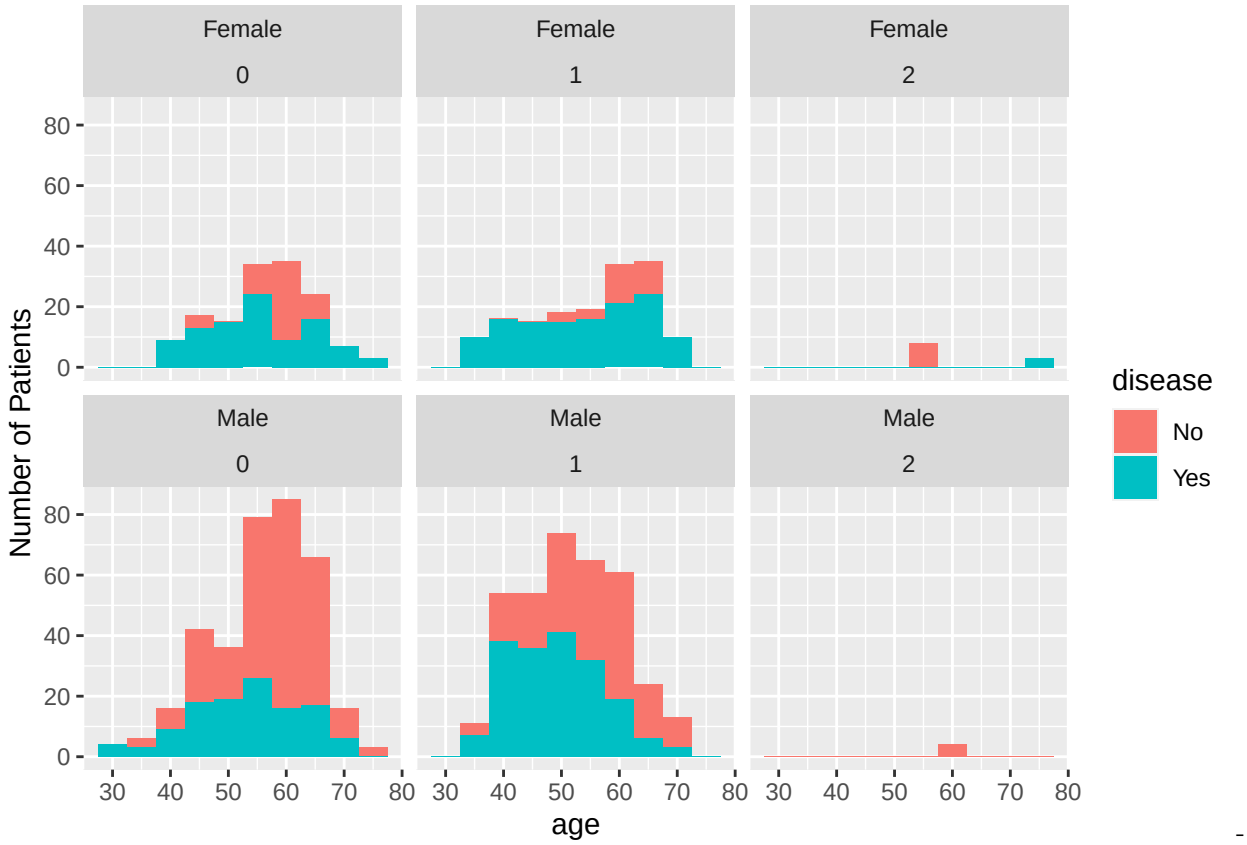
```
ggplot(heart_a)+geom_bar(aes(x= fbs, fill = disease), position = "stack")+ggtitle("Heart Attack - Fasti
```



- 400/800 Patients with Fasting blood sugar < 120 mg/dl experienced a heart attack i.e 50% of patients with low fbs were susceptible to heart diseases
- Approximately 60/140 patients experienced a heart attack whose Fasting blood sugar > 120 mg/dl
- The above analysis indicates that Fasting blood sugar is a moderate indciator of heart diseases

Resting ECG

```
ggplot(heart_a)+geom_histogram(aes(x= age, fill = disease), position = "stack",binwidth = 5) + facet_wra
```

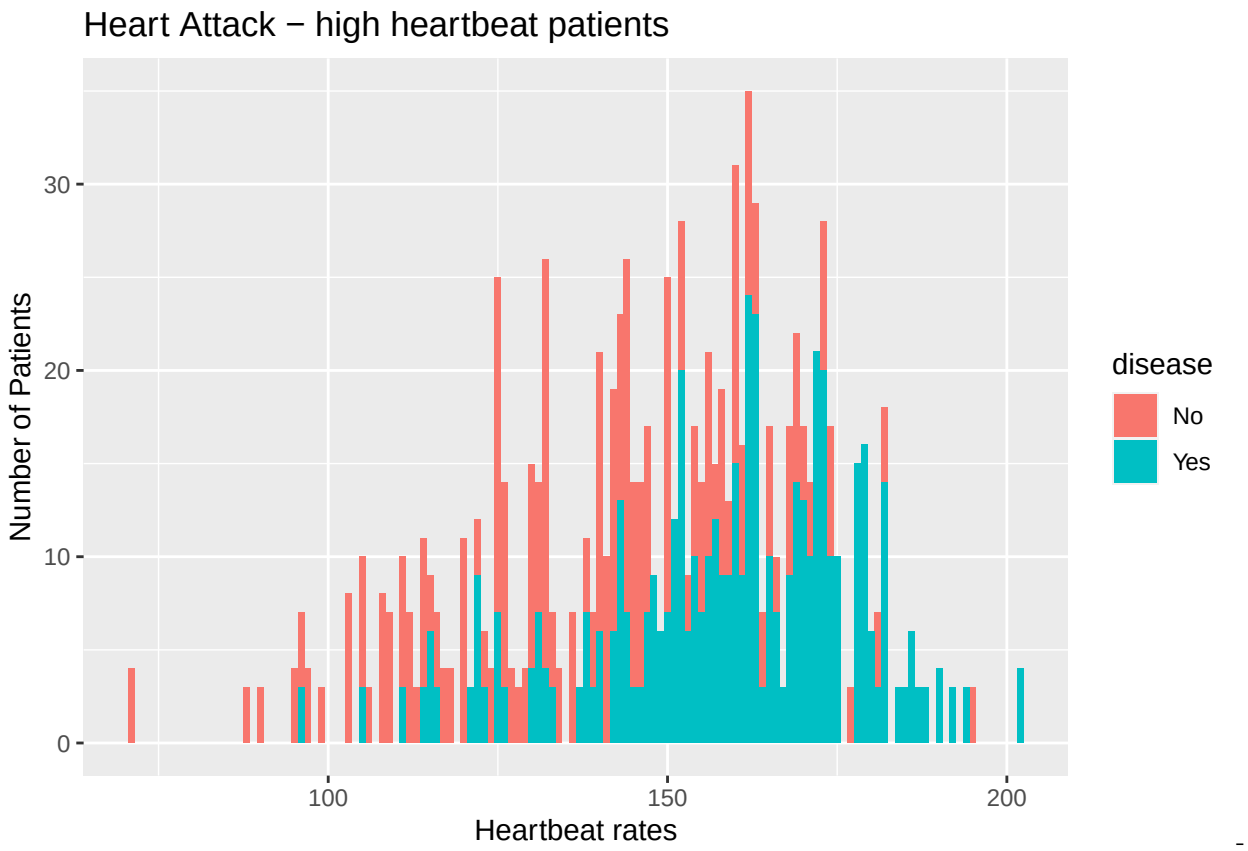


Females invariably with a normal resting ECG have been observed to experience heart attack/s - About 85% Females having ST-T wave abnormality experienced heart attack/s - However, males show a different pattern fewer experienced heart attack/s on a resting ECG - Males in the age group of 35- 45 having ST-T wave abnormality experienced more heart attack/s

Legends as below: Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

High heartbeat rate as an indicator for Heart Disease

```
ggplot(heart_a)+geom_bar(aes(x= thalach, fill = disease), position = "stack")+ggtitle("Heart Attack - hi
```

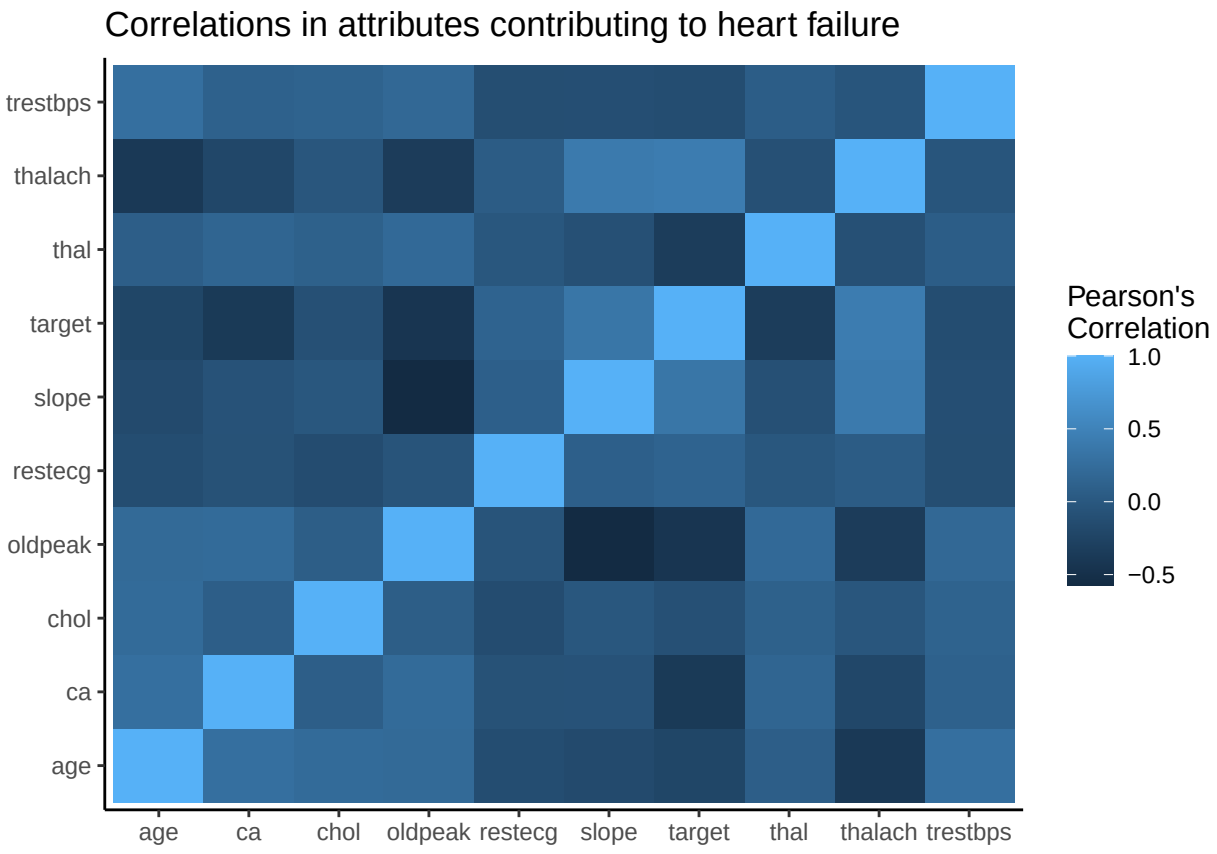


The patients with 150- 200 accounted for the highest heart attack/s amongst the patients studied. - Few ranges show a 100% share of the heart attack like ~ 125,160 - 200 - A little variation is noticed for the patients with higher heart beat not experiencing heart attck contributing to become the outliers in the dataset

Data Analysis & Modelling

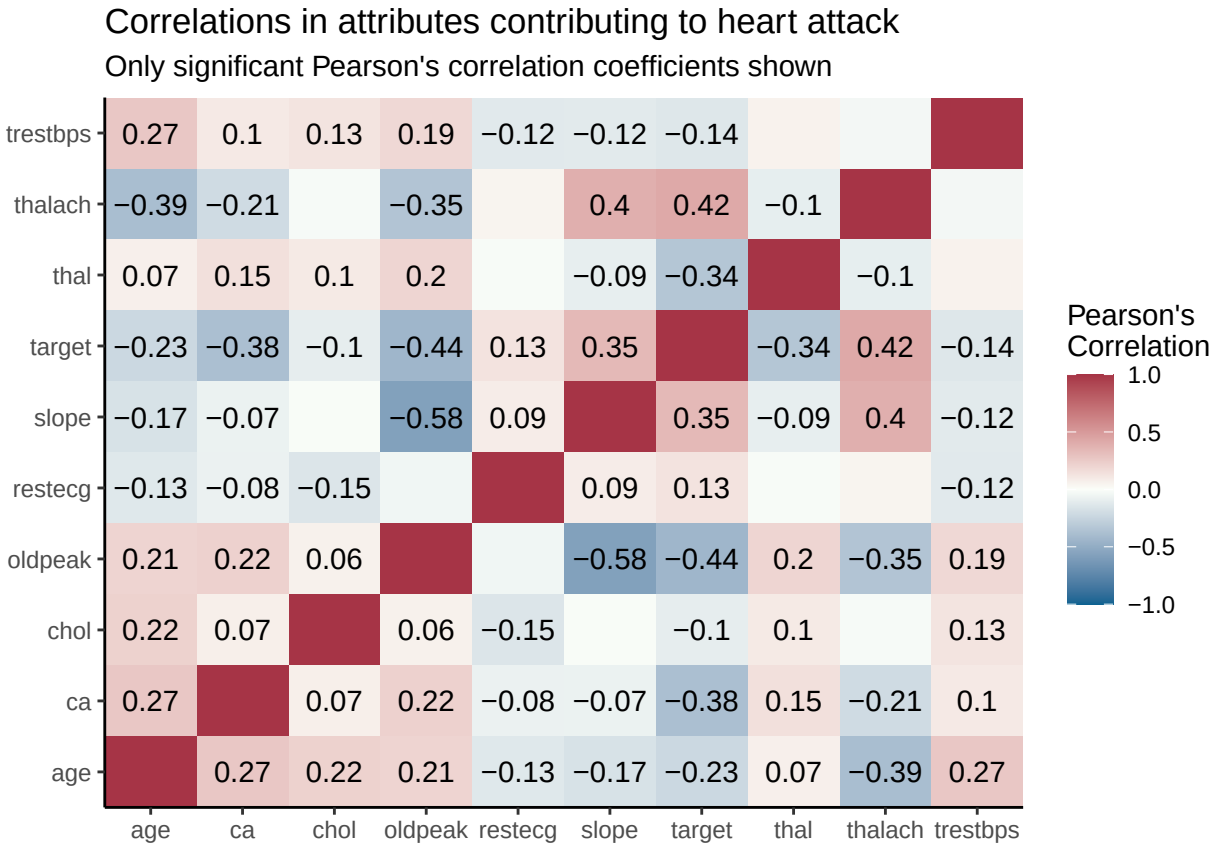
How strong is the correlation of the predictors with the heart disease

```
formatted_cors(heart_a) %>%ggplot(aes(x = measure1, y = measure2, fill = r)) +geom_tile() +labs(x = NULL
```




```
formatted_cors(heart_a) %>% ggplot(aes(measure1, measure2, fill=r, label=round(r_if_sig,2))) +geom_tile
```

```
## Warning: Removed 24 rows containing missing values (geom_text).
```



- The correlation plot signifies the correlation between the 13 variables and the target(heart attack)
- The plot has been customized to show the significant correlation coefficients where the p- values are less than 0.05
- Chest pain and maximum heart rate achieved had the strongest positive correlation with heart attack of 42% and 43% respectively
- Slope defined as The ST segment shift relative to exercise-induced increments in heart rate, the ST/heart rate slope (ST/HR slope), has been proposed as a more accurate ECG criterion for diagnosing significant coronary artery disease (CAD), shows a moderate correlation with a positive correlation coefficient of 35%
- Resting ECG shows a positive correlation with the target(heart attack) with a correlation coefficient of 13%
- The number of blood vessels have a negative correlation of 38% with heart attack/s claiming that higher the number of blood vessels reduce the chances of heart attack
- According to the data lesser age group has experienced more heart attack/s

Machine Learning

Creating sensitivity, specificity, and accuracy

```
sensitivity = function(cm) {
  return(cm[1,1]/(cm[1,1]+cm[1,2]))
}

specificity = function(cm) {
  return(cm[2,2]/(cm[2,1]+cm[2,2]))
}
accuracy = function(cm) {
  return((cm[1,1]+cm[2,2])/(cm[1,1]+cm[1,2]+cm[2,1]+cm[2,2]))
}
```

Creating the Training and Test Data Sets

```
ind = sample(2, nrow(heart), replace=TRUE, prob=c(0.70, 0.30))
heart$disease = factor(heart$target,
                       levels=c(0,1),
                       labels=c("No","Yes"))
ind = sample(2, nrow(heart), replace=TRUE, prob=c(0.70, 0.30))
heart.training = heart[ind == 1, 1:13]
heart.test = heart[ind == 2, 1:13]
heart.trainLabels = heart[ind ==1, 15]
heart.testLabels = heart[ind == 2, 15]
```

K-nearest neighbors algorithm

```
prediction = knn(train = heart.training,
                 test = heart.test,
                 cl = heart.trainLabels$disease,
                 k= 3)
result = cbind(prediction, heart.testLabels)
(confusionMatrix= table(actual_value=result$disease, Predicted_value= result$prediction))
```

```
##             Predicted_value
## actual_value  No Yes
##          No  101  37
##          Yes  29 115
```

- The performance of this method, on the same test data set as before, can be calculated with the same functions we had defined earlier: The sensitivity is 0.7318841, the specificity is 0.7986111, and the overall accuracy is 0.7659574.
- The nearest neighbour model has nearly more false positives than the false negatives making it unreliable for an accurate prediction of heart attack. We shall need more data points to arrive at a robust model

Logistic Regression

```
trainingWithLabel = heart.training
trainingWithLabel$disease = heart.trainLabels$disease

logisticModel = glm(disease ~ age + sex + cp + trestbps + chol + fbs +restecg + thalach + exang +
                      oldpeak + slope + ca + thal, data = trainingWithLabel,
                    family = "binomial")
```

Predicting the test set:

```
prediction = predict(logisticModel, heart.test, type='response')
heart.test$predicted = ifelse(prediction >.7, 1,0)
```

Result of the confusion matrix:

```
result = cbind(heart.testLabels, prediction > .7 )
(confusionMatrix = table(actual_value=result$disease, Predicted_value=result$`prediction > 0.7` ))
```

```
##             Predicted_value
## actual_value FALSE TRUE
##          No    118   20
##          Yes    30  114
```

- The LRR also has more false positives than false negatives making it an unrelaible model to predict a heart attack
- The performance of this method, on the same test data set as before, can be calculated with the same functions we had defined earlier: The sensitivity is 0.8550725, the specificity is 0.7916667, and the overall accuracy is 0.822695.

Decision Tree

Decision on Chest pain, Maxiumum heart rate achieved, and Blood vessels.

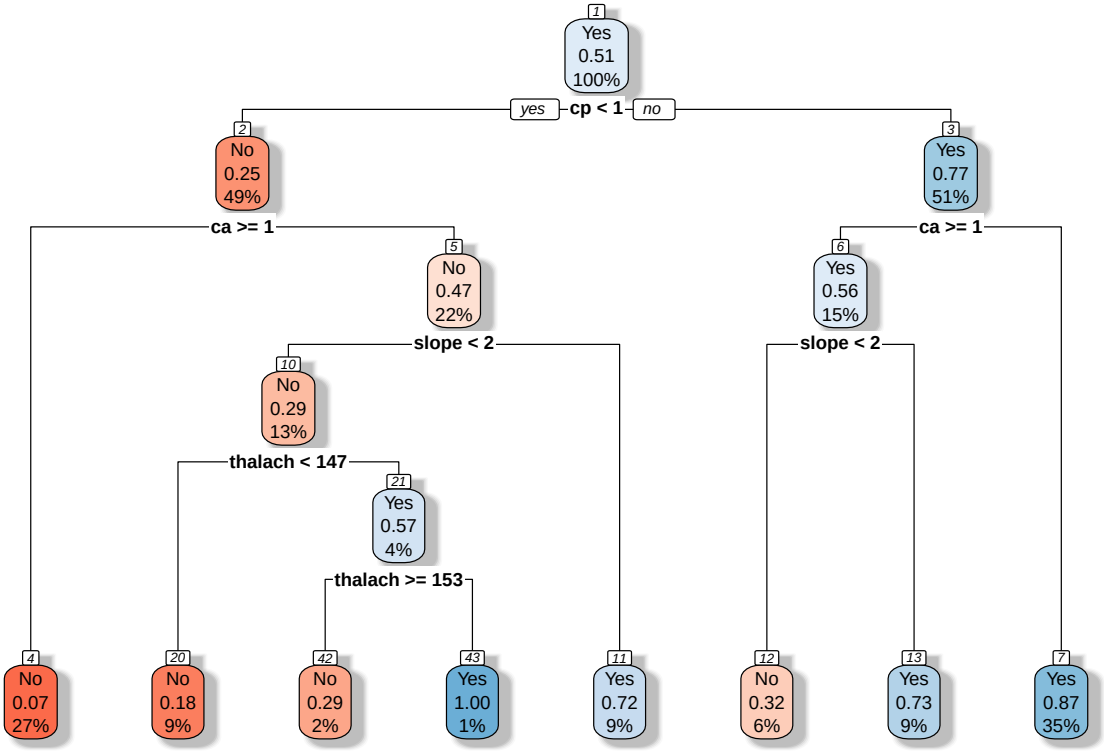
```
model = rpart(disease ~ cp  + thalach+ ca + slope, data = trainingWithLabel,
              method = "class")

prediction = predict(model, heart.test, type='class')
(confusionMatrix = table(Actul_Value = heart.testLabels$disease,
                         Predicted_value = prediction))
```

```
##            Predicted_value
## Actul_Value  No Yes
##         No  111  27
##         Yes  18 126
```

- Based on the confusion matrix the DT has more false negatives which are 36 which shall alert a health care physician even in case of no heart attack than the false positives(16) making it more reliable a model to predict a heart attack
- The performance of this method, on the same test data set as before, can be calculated with the same functions we had defined earlier: The sensitivity is 0.8043478, the specificity is 0.875, and the overall accuracy is 0.8404255.

Decision Tree

```
rpart.plot(model, box.palette="RdBu", shadow.col="gray", nn=TRUE)
```



- Based on the decision tree having asymptomatic chest pain there is 48% chance of not having a heart attack with having one or more critical blood vessels it increases chance of having a heart attack by 20%. Furthermore, having maximum heart rate less than 147 can further decrease the chance of heart attack by 10% but if the slope is less than 2 then it slightly increases the chance of the heart attack.

- Based on the decision tree having symptomatic chest pain can lead to higher chance of heart attack by 52%. Additionally, having more than one critical blood vessels patients will likely to experience higher chance of heart attack by 17% but if the slope is less then 2 then it decreases the chance of heart attack by 7%.

XG Boost

```
require(caTools)
new_heart = heart
library(caret)
```

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

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

```
## The following objects are masked _by_ '.GlobalEnv':
##
##     sensitivity, specificity
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(80)

idx = createDataPartition(new_heart$target, p=.75, list = FALSE)

train = new_heart[idx,]
test = new_heart[-idx,]

drop = c('target', 'disease')

x_train = train[ , !(names(train) %in% drop)]
y_train = train$target

x_test = test[ , !(names(test) %in% drop)]
y_test = test$target

negative_cases = sum(y_train==0)
positive_cases = sum(y_train == 1)

## Model training
library(xgboost)
```

```
##
## Attaching package: 'xgboost'
```

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

```
dtrain = xgb.DMatrix(data = as.matrix(x_train), label= y_train)
dtest <- xgb.DMatrix(data = as.matrix(x_test), label= y_test)
```

```r
set.seed(80)

xgb_model <- xgboost(data = dtrain,
                     max.depth = 4,
                     nround = 90,
                     early_stopping_rounds = 3,
                     objective = "binary:logistic",
                     scale_pos_weight = negative_cases/positive_cases,
                     gamma = 1
)
```

```
## [18:15:48] WARNING: amalgamation/../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evalu
## [1]  train-logloss:0.527564
## Will train until train_logloss hasn't improved in 3 rounds.
##
## [2]  train-logloss:0.433062
## [3]  train-logloss:0.358886
## [4]  train-logloss:0.305452
## [5]  train-logloss:0.266815
## [6]  train-logloss:0.232068
## [7]  train-logloss:0.206535
## [8]  train-logloss:0.186842
## [9]  train-logloss:0.170951
## [10] train-logloss:0.153929
## [11] train-logloss:0.143346
## [12] train-logloss:0.133980
## [13] train-logloss:0.122438
## [14] train-logloss:0.113129
## [15] train-logloss:0.108896
## [16] train-logloss:0.102793
## [17] train-logloss:0.097343
## [18] train-logloss:0.092580
## [19] train-logloss:0.088015
## [20] train-logloss:0.083153
## [21] train-logloss:0.078813
## [22] train-logloss:0.075219
## [23] train-logloss:0.073081
## [24] train-logloss:0.070087
## [25] train-logloss:0.066209
## [26] train-logloss:0.064196
## [27] train-logloss:0.063196
## [28] train-logloss:0.059210
## [29] train-logloss:0.059210
## [30] train-logloss:0.059210
## [31] train-logloss:0.059210
## Stopping. Best iteration:
## [28] train-logloss:0.059210
```

```r
pred = predict(xgb_model, dtest)
xgbpred = ifelse(pred > 0.50, 1, 0)

error = mean(xgbpred != y_test)
print(paste("Test erorr was", error))
```

```
## [1] "Test erorr was 0.03515625"
```

```
##Testing

library(caret)
confusionMatrix(as.factor(xgbpred), as.factor(y_test))
```

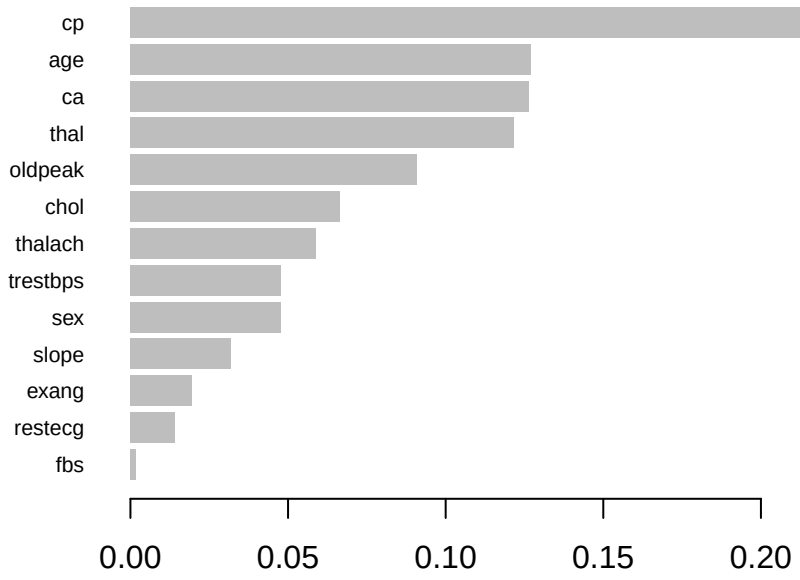
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 126   4
##          1   5 121
##
##                Accuracy : 0.9648
##                  95% CI : (0.9343, 0.9838)
##     No Information Rate : 0.5117
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9297
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9618
##             Specificity : 0.9680
##          Pos Pred Value : 0.9692
##          Neg Pred Value : 0.9603
##              Prevalence : 0.5117
##          Detection Rate : 0.4922
##    Detection Prevalence : 0.5078
##       Balanced Accuracy : 0.9649
##
##        'Positive' Class : 0
##
```

Feature importance

```
mat = xgb.importance(feature_names = colnames(y_train), model = xgb_model)
mat
```

```
##       Feature        Gain       Cover   Frequency
##  1:        cp 0.246037046 0.140917142 0.078498294
##  2:       age 0.127049650 0.152092771 0.197952218
##  3:        ca 0.126369518 0.125008948 0.075085324
##  4:      thal 0.121723629 0.095927814 0.044368601
##  5:   oldpeak 0.090812517 0.118829707 0.112627986
##  6:      chol 0.066383101 0.074420750 0.129692833
##  7:   thalach 0.058958776 0.088173367 0.112627986
##  8: trestbps 0.047810309 0.066743712 0.095563140
##  9:       sex 0.047720431 0.059446776 0.051194539
## 10:     slope 0.031946281 0.031585176 0.034129693
## 11:     exang 0.019494634 0.025443530 0.027303754
## 12:   restecg 0.014008373 0.018442499 0.034129693
## 13:       fbs 0.001685735 0.002967806 0.006825939
```

```
xgb.plot.importance(importance_matrix = mat)
```



- XG Boost has served as the best classification model of the 4 model we have worked on with highest Accuracy of 96%
- The false positives and false negatives in the prediction dataset are as small as 5 and 4 respectively stabilizing the false outcomes with a precision of 96.8%(True Positive/ (False Positive + True Positive))

FINAL REMARKS

- Chest pain is considered the most important factor predicting the risk of heart attack/s which can be observed in all the models - Pearson Correlation and XG Boost
- Age is an important predictor where patients between the age group of 40-60 are the highest risk of heart attack/s
- The number of critical blood vessels stand third in the importance predictor predicting heart attack/s
- Maximum heart rate achieved and old peak carry a moderately strong correlation with the risk of heart attack/s
- Males are prone to more heart attack/s than females
- Ironically Fasting Blood Sugar (Diabetes) is not a strong predictor for a heart attack/s