HEART DISEASE PREDICTION

MEMBER

1. Krittin Srithong 62070503402
2. Chayaphon Bunyakan 62070503412
3. Pirada Amornprapawat 62070503437

INTRODUCTION

What is heart disease

Heart disease refers to any condition affecting the heart. It often caused by a fat clot or a blood clot to clog more and more arteries in the heart, making it unable to function effectively various organs receive inadequate blood supply, such as the accumulation of plaque in arteries or arteries can damage your blood vessels and heart. Reduced blood flow can lead to a heart attack, chest pain (angina), or a stroke.

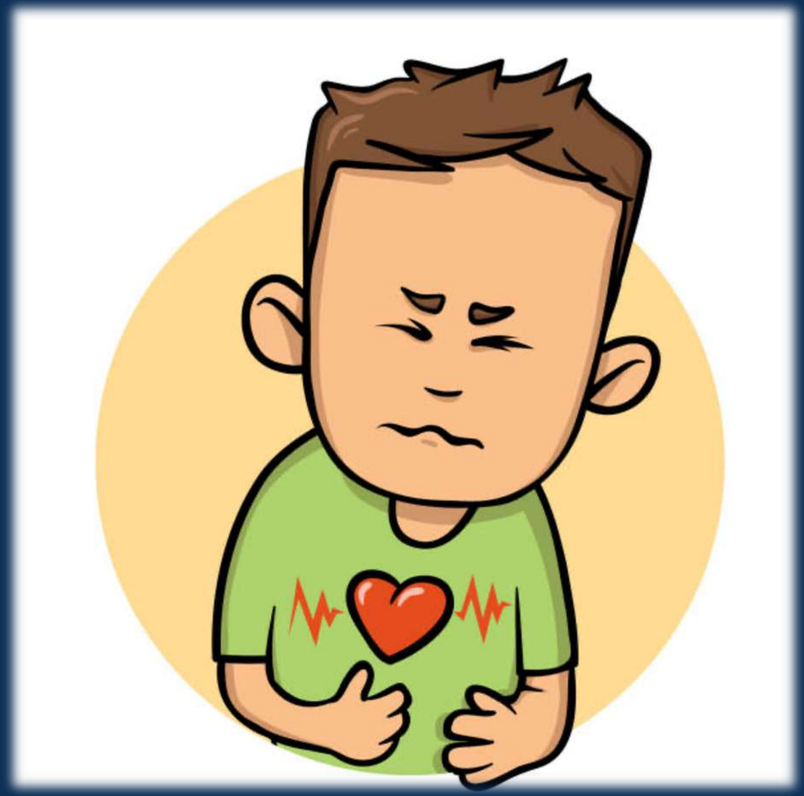
Why prediction is importance

To help prevent heart disease from happen or check whether the patient have it or not. To help doctor or patient do basic diagnosis.

INTRODUCTION

Dataset : Heart Disease Dataset
Author : David Lapp (in Kaggle)

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 14 attributes that can be use for prediction.

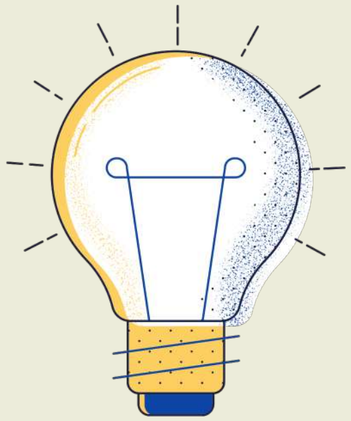


ANALYTIC OBJECTIVE

Have Heart disease

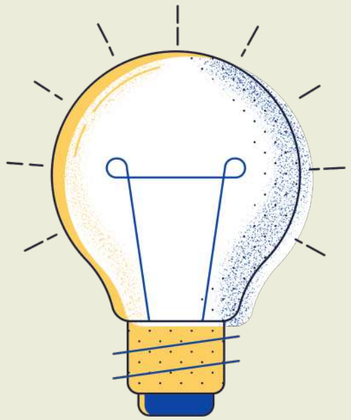
Don't Have Heart disease

To predict that people in the test subject have heart disease or not by using their medical Test result.



DATA DESCRIPTION

- Age
- Chest pain type (cp)
- Sex
- Fasting blood sugar (fbs)
- Serum cholestoral in mg/dl (chol)
- Resting blood pressure (trestbps)
- Resting electrocardiographic results (restecg)
- Maximum Heart Rate Achieved (thalach)



DATA DESCRIPTION

ST segment slope (slope)

Number of Major Vessels (ca)

Thalassemia (thal)

Heart disease Yes / No (target)

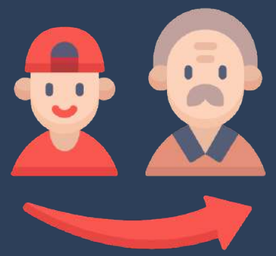
Depression of ST segment

Exercise induced angina (exang)

DATA DESCRIPTION

```r
library(readr)
library(tidyverse)

# importing the data
heartData <- read_csv("heartData.csv")
View(heartData)

# see the structure and the over all of the data
summary(heartData)
```

```
      age             sex               cp             trestbps         chol            fbs              restecg      
 Min.   :29.00   Min.   :0.0000   Min.   :0.0000   Min.   : 94.0   Min.   :126   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:120.0   1st Qu.:211   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :56.00   Median :1.0000   Median :1.0000   Median :130.0   Median :240   Median :0.0000   Median :1.0000  
 Mean   :54.43   Mean   :0.6956   Mean   :0.9424   Mean   :131.6   Mean   :246   Mean   :0.1493   Mean   :0.5298  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.0000   3rd Qu.:140.0   3rd Qu.:275   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :77.00   Max.   :1.0000   Max.   :3.0000   Max.   :200.0   Max.   :564   Max.   :1.0000   Max.   :2.0000  
    thalach          exang           oldpeak          slope             ca              thal           target      
 Min.   : 71.0   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000  
 1st Qu.:132.0   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000  
 Median :152.0   Median :0.0000   Median :0.800   Median :1.000   Median :0.0000   Median :2.000   Median :1.0000  
 Mean   :149.1   Mean   :0.3366   Mean   :1.072   Mean   :1.385   Mean   :0.7541   Mean   :2.324   Mean   :0.5132  
 3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.800   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :202.0   Max.   :1.0000   Max.   :6.200   Max.   :2.000   Max.   :4.0000   Max.   :3.000   Max.   :1.0000  
```



DATA DESCRIPTION

Age

- Age of people in the dataset

Sex

- Male = 1 and Female = 0

Chest pain type (cp)

- There are 4 type of chest pain
- Typical Angina (1), Atypical angina (2), Non-anginal Pain (3) and Asymptomatic (0)

DATA DESCRIPTION

Resting blood pressure (trestbps)

- Resting blood pressure of the people (in mm Hg)

Serum cholestoral in mg/dl (chol)

- The amount of cholesterol (in mg/dl)

Fasting blood sugar (fbs)

- (1) is > 120 and (0) <= 120 mg/dl

DATA DESCRIPTION

Resting electrocardiographic results (restecg)

- 0 = normal
- 1 = having ST-T wave abnormality
- 2 = showing probable or definite left ventricular hypertrophy

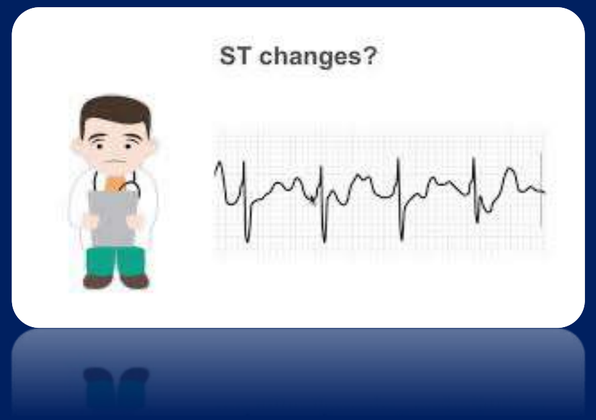
Maximum Heart Rate Achieved (thalach)

- Maximum Heart Rate that the people have

Exercise induced angina (exang)

- Yes = 1 and No = 0

DATA DESCRIPTION

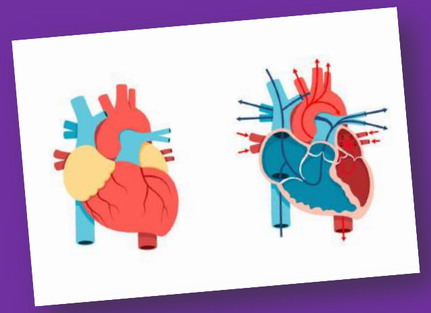


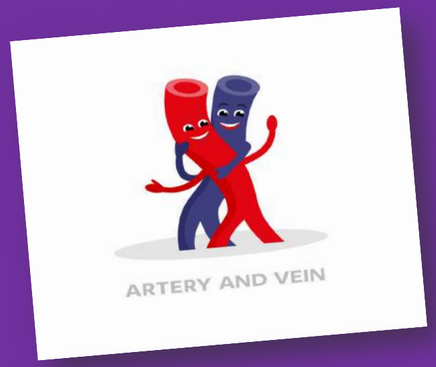
Depression of ST segment(oldpeak)

- The ST segment is a part of the electrocardiogram of a heartbeat that is usually found at a certain level in a normal heartbeat.

ST segment slope (slope)

- The slope of the peak exercise ST segment
- 1 : upsloping, 2 : flat, 3 : down sloping





DATA DESCRIPTION

Thalassemia (thal)

- Results of the blood flow observed via the radioactive dye.
- 1 = normal, 2 = fixed defect, 3 = reversable defect

Heart disease Yes / No (target)

- Have disease = 1 ,don’t have = 0

Number of Major Vessels (ca)

- Number of Major Vessels (1-3) Colored by Fluoroscopy

DATA PREPARATION

```r
# 1.check for Null value
colSums(is.na(heartData))

# 2.tranfrom data for future visualization and modeling
heartData %>%
  mutate(sex = as.factor(ifelse(sex == 1,'Male','Female')),
         cp = as.factor(ifelse(cp == 1,'Typical Angina',
                               ifelse(cp == 2,'Atypical angina',
                                                    ifelse(cp == 3,'Non-anginal Pain','Asymptomatic')))),
         fbs = as.factor(ifelse(fbs == 1,'>120','<=120')),
         restecg = as.factor((ifelse(restecg == 0,'Normal',
                                     ifelse(restecg == 1,'ST-T wave Abnormality','Probable or definite')))),
         exang = as.factor(ifelse(exang == 1,'Yes','No')),
         slope = as.factor(ifelse(slope == 1,'Upsloping',ifelse(slope == 2,'Flat','Downsloping'))),
         ca = as.factor(ca),
         thal = as.factor(ifelse(thal == 1,'Normal',ifelse(thal == 2,'Fixed defect','Reversable defect'))),
         target = as.factor(ifelse(target == 1,'Yes','No')))-> heartData_Modify

summary(heartData_Modify)
```

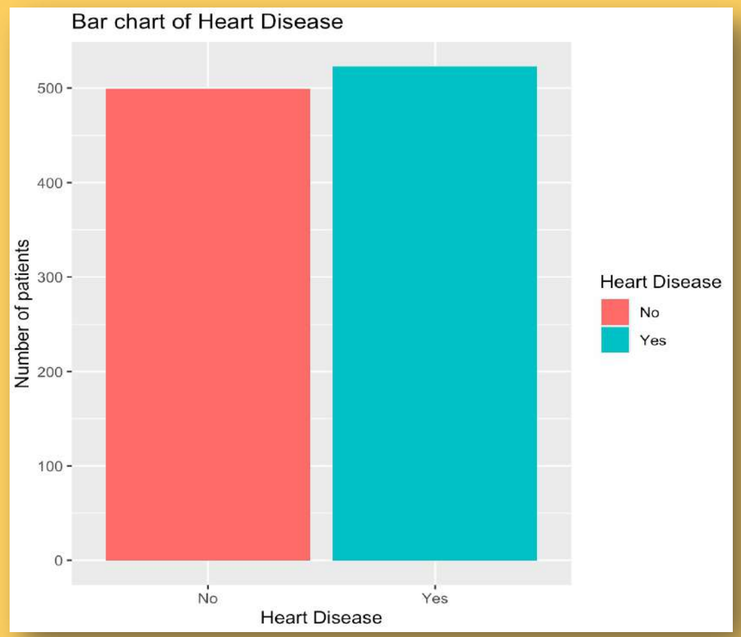
Check for Null and Change some variable to categorical variable

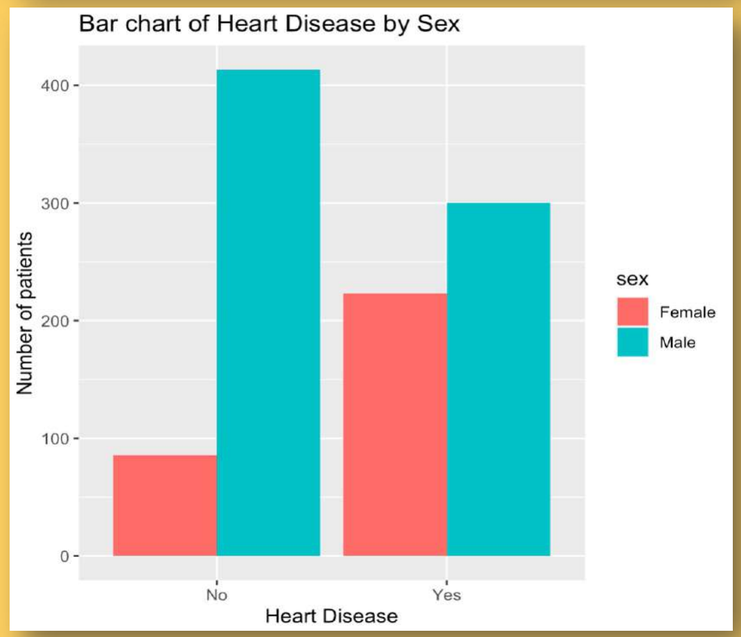
DATA PREPARATION

```
> colSums(is.na(heartData))
     age      sex       cp trestbps     chol      fbs  restecg  thalach    exang  oldpeak    slope       ca     thal   target 
       0        0        0        0        0        0        0        0        0        0        0        0        0        0 
```

```
      age             sex                      cp          trestbps          chol          fbs                         restecg      thalach       exang    
 Min.   :29.00   Female:312   Asymptomatic    :497   Min.   : 94.0   Min.   :126   <=120:872   Normal                 :497   Min.   : 71.0   No :680  
 1st Qu.:48.00   Male  :713   Atypical angina :284   1st Qu.:120.0   1st Qu.:211   >120 :153   Probable or definite : 15   1st Qu.:132.0   Yes:345  
 Median :56.00                Non-anginal Pain: 77   Median :130.0   Median :240               ST-T wave Abnormality:513   Median :152.0            
 Mean   :54.43                Typical Angina  :167   Mean   :131.6   Mean   :246                                           Mean   :149.1            
 3rd Qu.:61.00                                       3rd Qu.:140.0   3rd Qu.:275                                           3rd Qu.:166.0            
 Max.   :77.00                                       Max.   :200.0   Max.   :564                                           Max.   :202.0            
    oldpeak             slope      ca                      thal      target    
 Min.   :0.000   Downsloping: 74   0:578   Fixed defect     :544   No :499  
 1st Qu.:0.000   Flat       :469   1:226   Normal           : 64   Yes:526  
 Median :0.800   Upsloping  :482   2:134   Reversable defect:417            
 Mean   :1.072                     3: 69                                    
 3rd Qu.:1.800                     4: 18                                    
 Max.   :6.200                                                              
```

DATA EXPLORATION AND VISUALIZATION





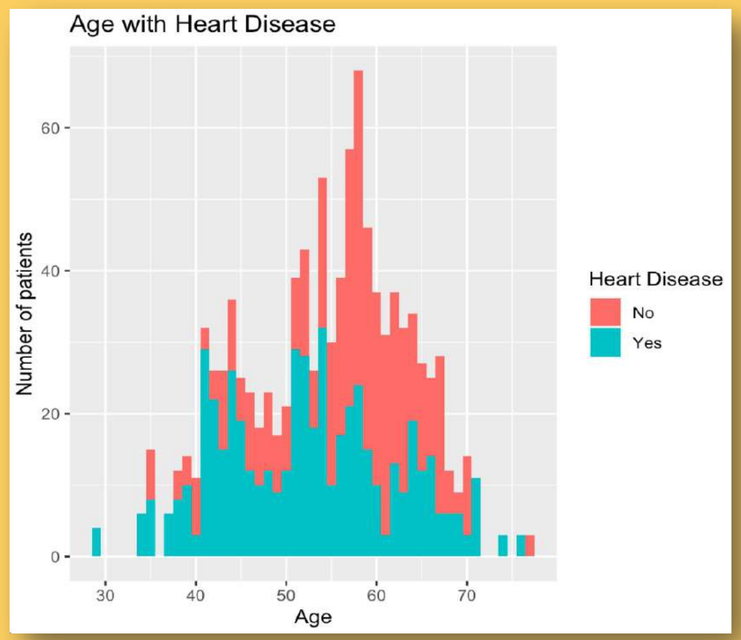
Taget

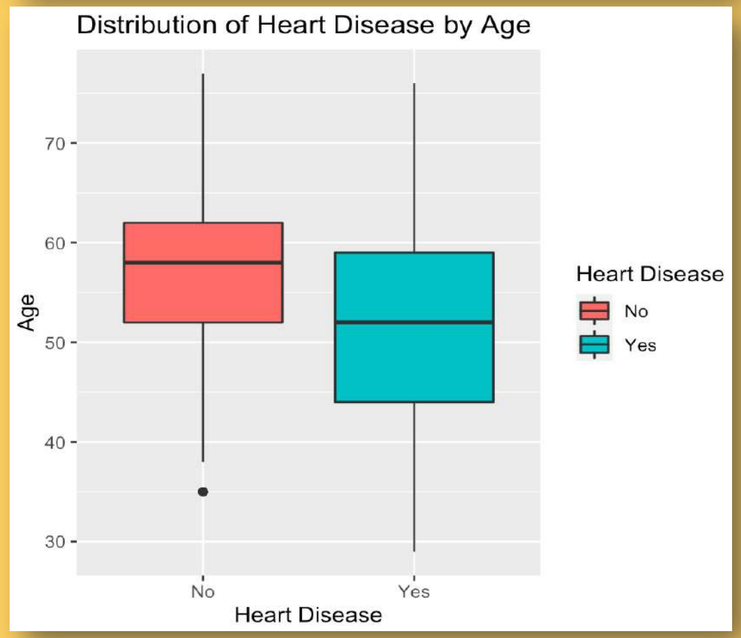
1st graph

- The number people that have Heart Disease is slightly higher than not having Heart Disease.

2nd graph

- The number of male that that don't have Heart Disease is higher than male that have Heart Disease.
- The number of female that that don't have Heart Disease is lower than female that have Heart Disease.





Age

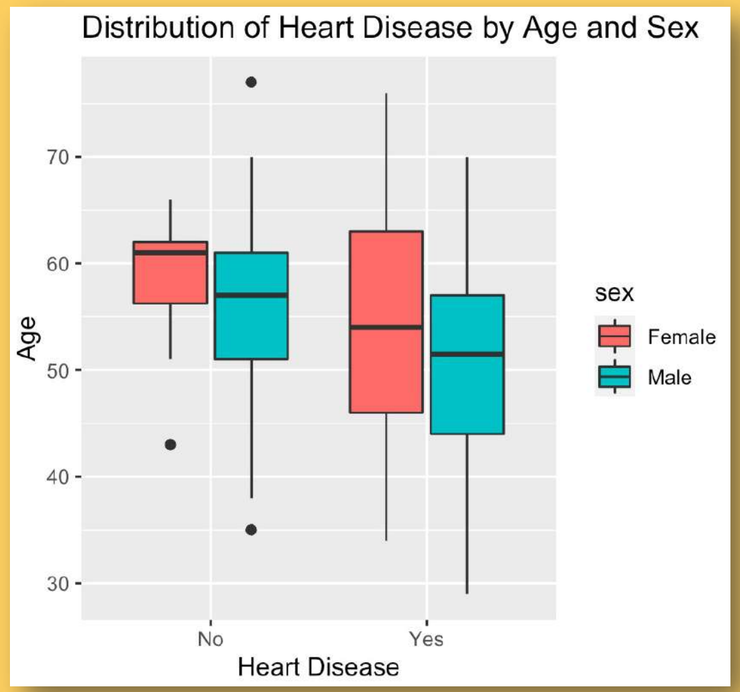
1st graph

- This graph show the distribution of age in the data set that is categorize by the disease.

2nd graph

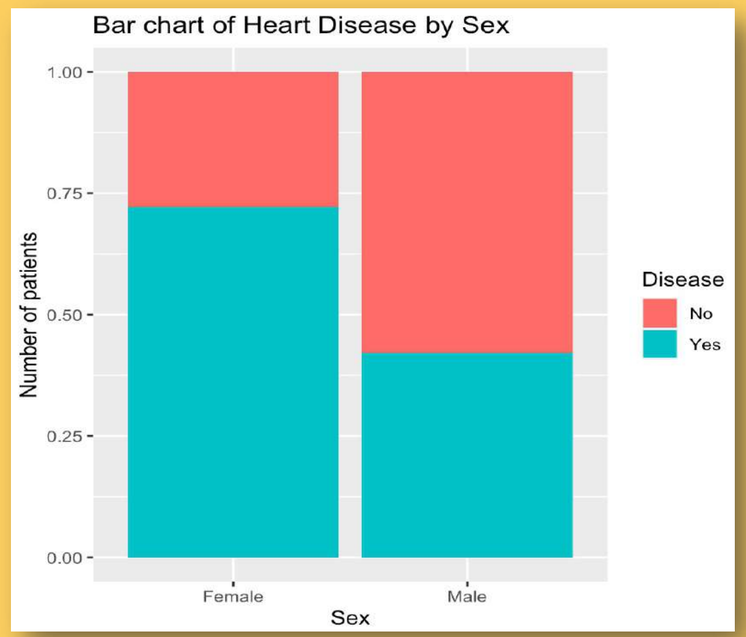
- The popularity of age that don't have heart disease is around 52-62 year and the median is around 58 year.
- The popularity of age that have heart disease is around 44-59 year and the median is around 52 year.

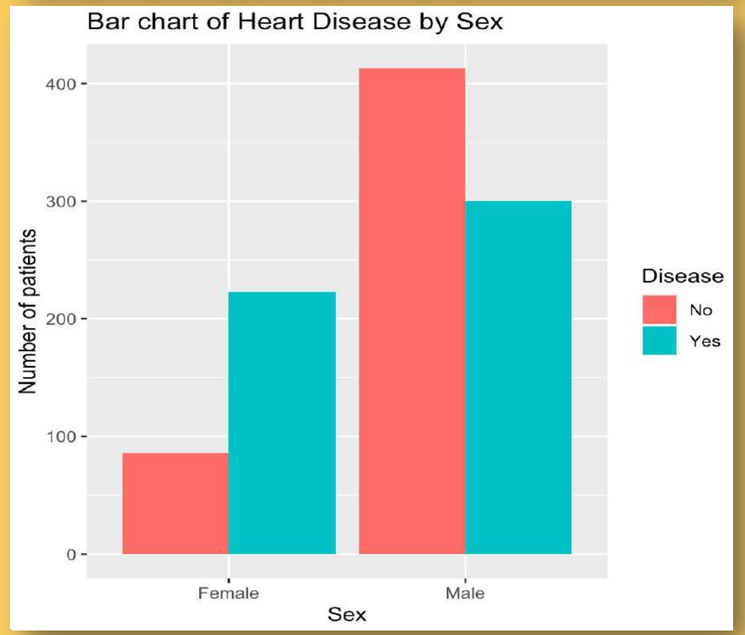
Age



3rd graph

- This graph show the distribution of age in the data set that is categorize by the disease and sex.
- The result show the popularity of age for the people that have or don't have the disease by sex type.

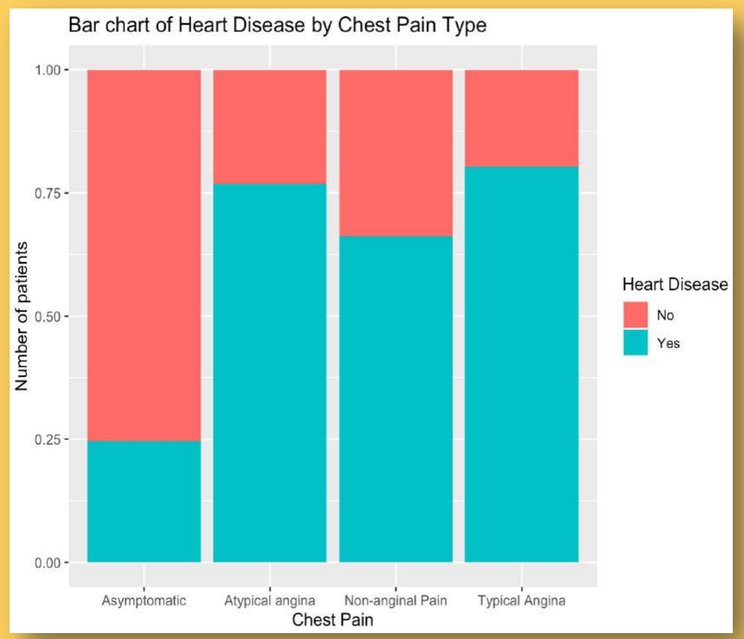


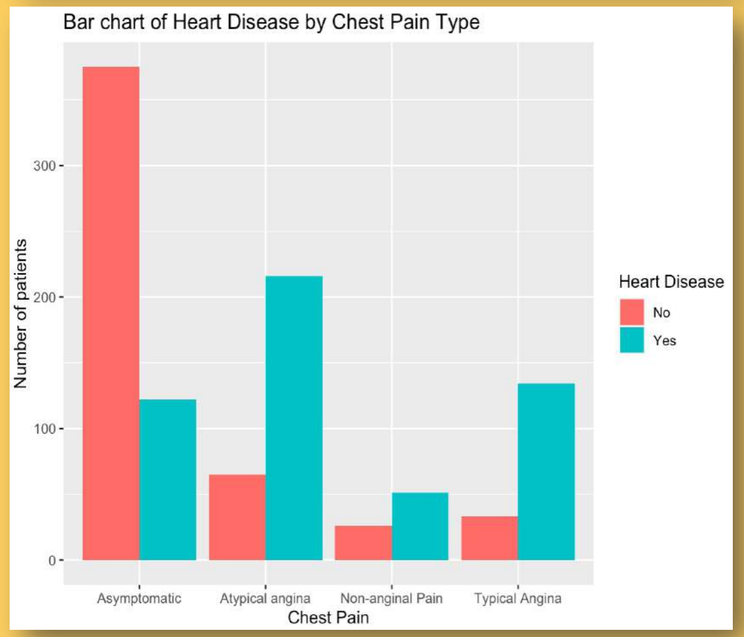


Sex

1st and 2nd graph

- This graph shows the population of people by sex
- Compare between number of people for each sex and categorize by disease.
- Female in the dataset are less likely to have heart disease, male in the data set are more likely to have heart disease.

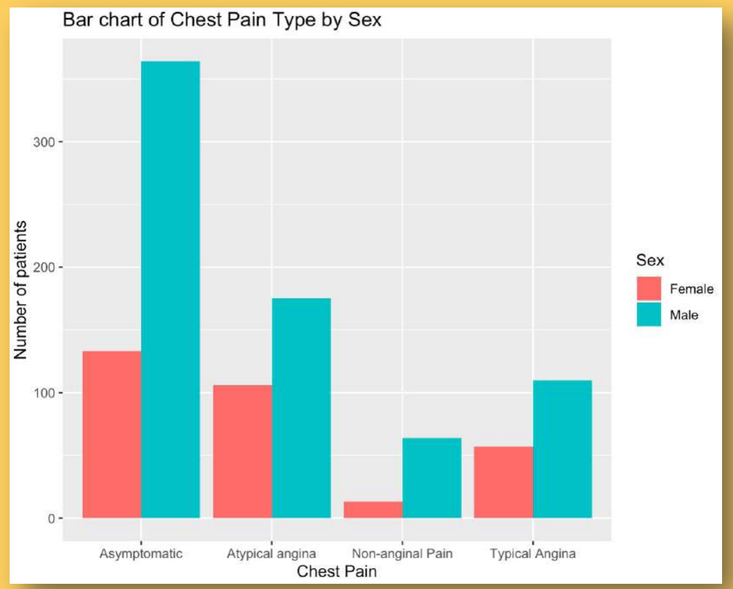


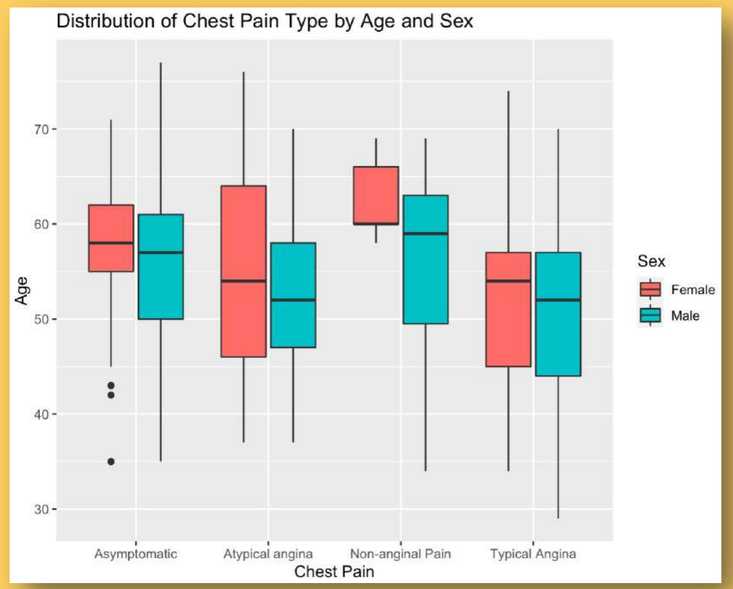


Chest pain type (cp)

1st and 2nd graph

- This graph shows and count the number of people that have chest pain for each type and categorize by the disease.
- These to graph can answer the question about the number of people that have each type of chest pain and see that they have heart disease or not.
- For example, Alypical angina type is the highest type for the people that have disease.





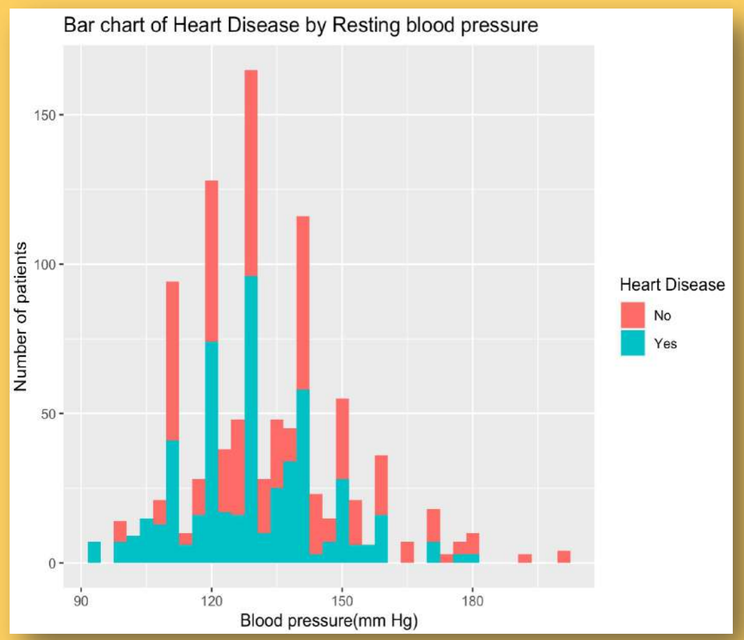
Chest pain type (cp)

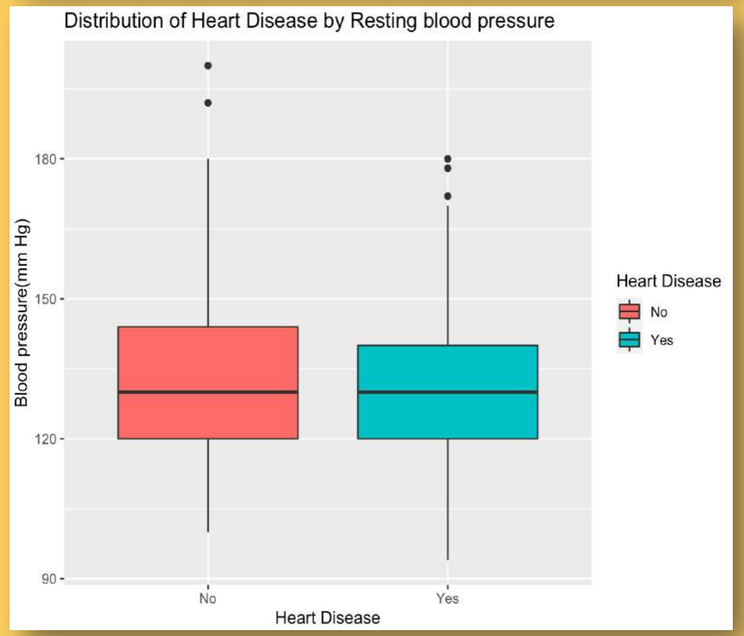
3rd graph

- This graph compare the number of chest pain for each type by sex type.
- For example , Asmptomatic type is have the highest case for male.

4th graph

- This graph show the distribution of each chest pain type by age.





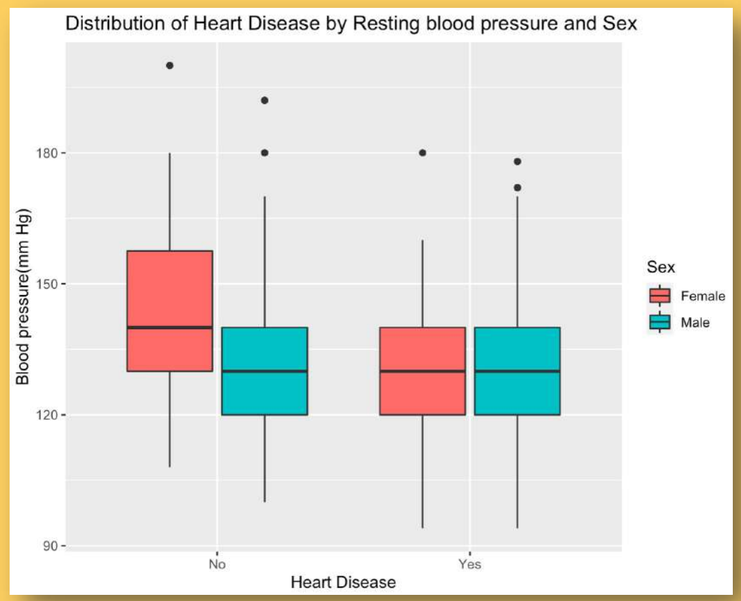
Resting blood pressure (trestbps)

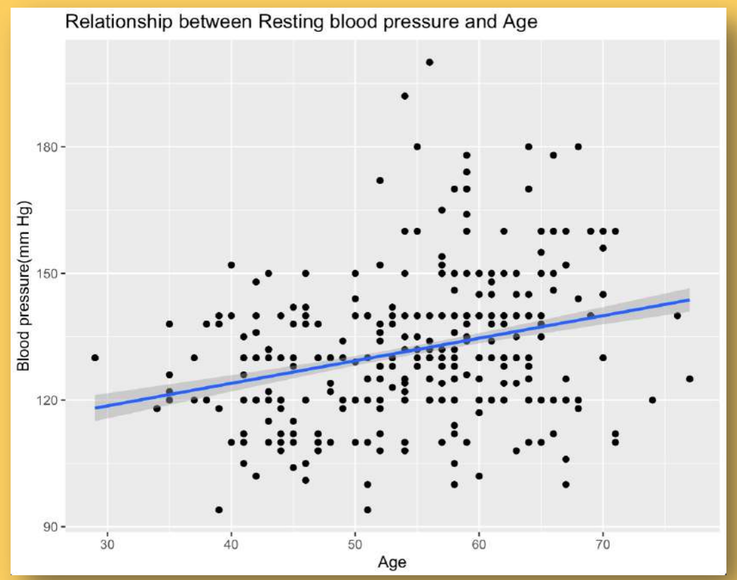
1st graph

- This graph show the distribution of resting blood pressure in the dataset that is categorize by the disease.

2nd graph

- From this graph, the resting blood pressure have similar trend between people that have or don’t have disease.





Resting blood pressure (trestbps)

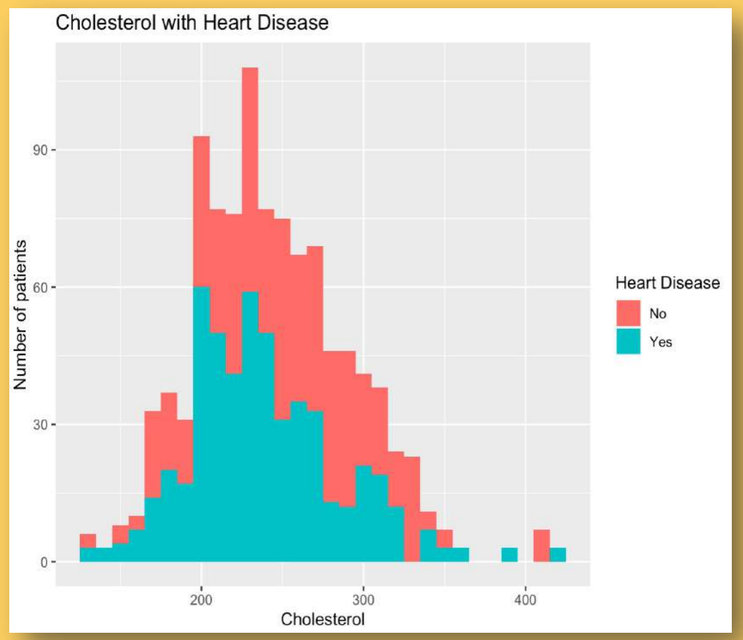
3rd graph

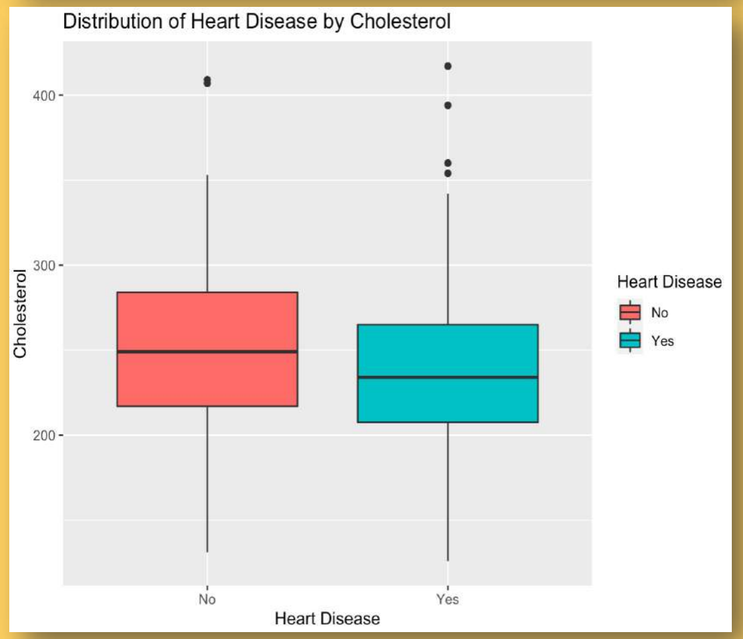
- Same as the second graph but are categorize by sex type.

4th graph

- The graph show the relationship between age and Resting blood pressure.
- Weak Positive relationship

```
> cor(heartData_Modify$trestbps, heartData_Modify$age)
[1] 0.2761243
```





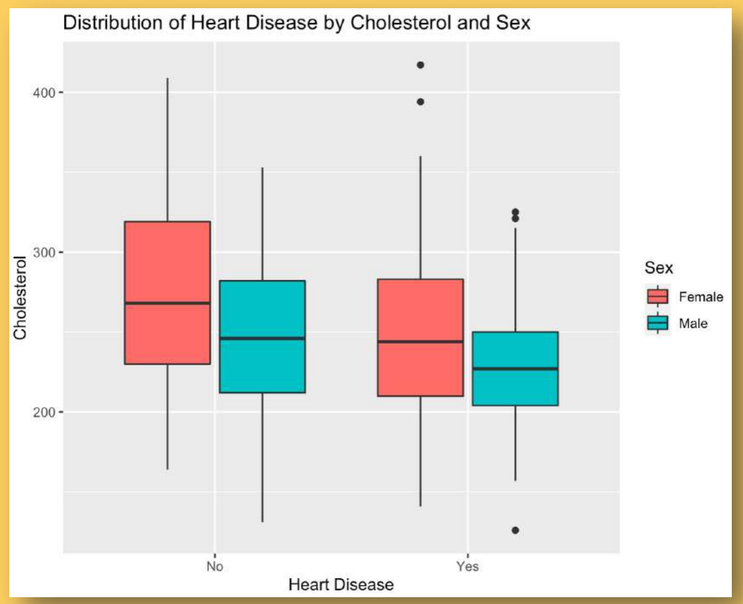
Serum cholesteral in mg/dl (chol)

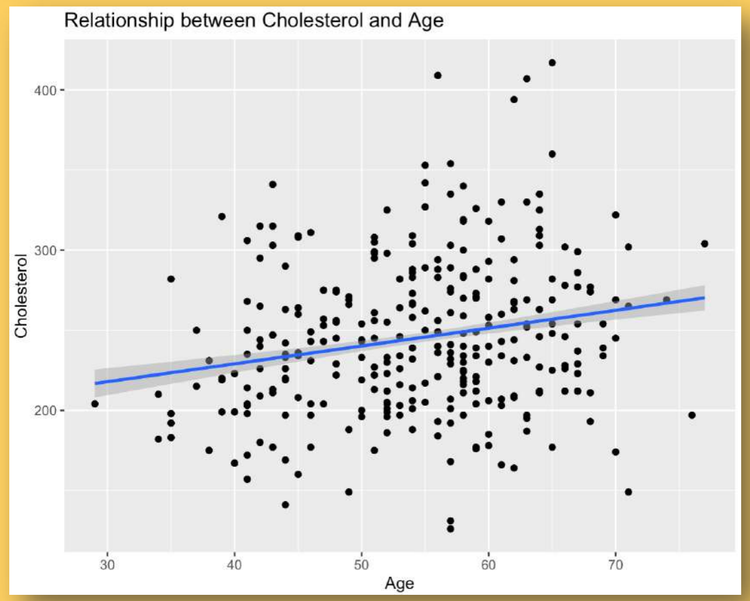
1st graph

- This graph show the distribution of Serum cholesterol in the dataset that is categorize by the disease.

2nd graph

- From this graph, the Serum cholesterol have similar trend between people that have or don’t have disease.





Serum cholesteral in mg/dl (chol)

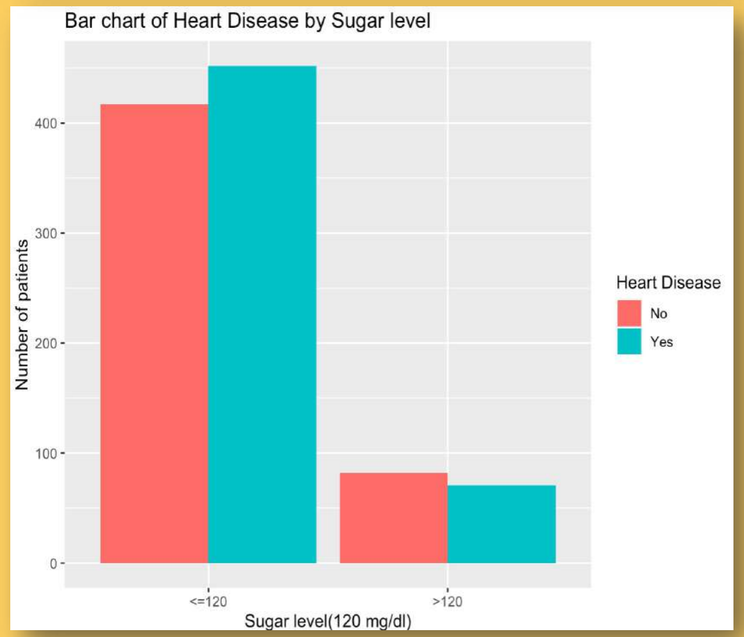
3rd graph

- Same as the second graph but are categorize by sex type.

4th graph

- The graph show the relationship between age and Serum cholesterol.
- Weak Positive relationship

```
> cor(heartData_Modify$chol, heartData_Modify$age)
[1] 0.2071956
```



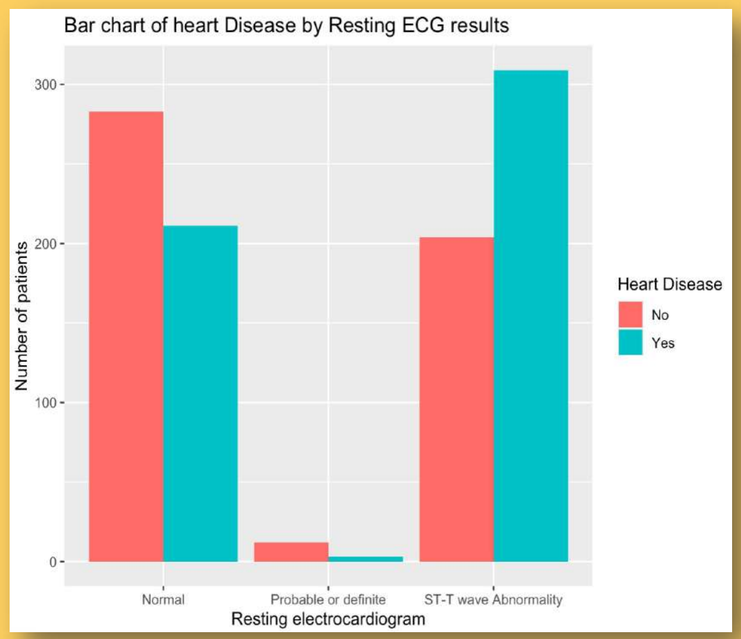
Fasting blood sugar (fbs)

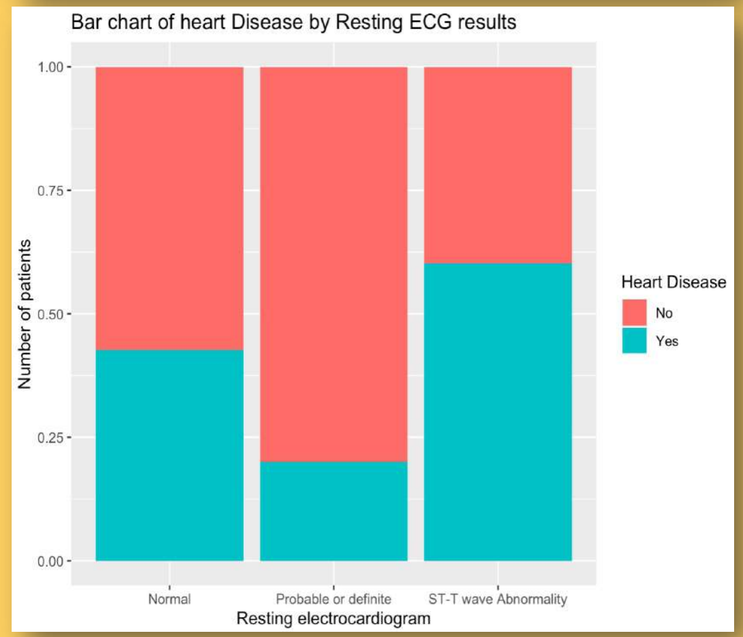
1st graph

- This graph shows the distribution of Fasting blood sugar in the dataset that is categorize by the Type of blood sugar and disease.

2nd graph

- From this graph count the number people that have blood sugar in each type
- Most of the people have less than or equal 120 of sugar level and there is slightly different between people that have or don’t have disease.





Resting electrocardiographic results (restecg)

1st and 2nd graph

- This graph shows and count the number people that have each type of resting electrocardiographic and categorize by the disease.
- These to graph can answer the question about the number of people that have each type of resting electrocardiographic and see that they have heart disease or not.
- For example, People that have heart disease tend to have ST-T Abnormality type the most follow by Normal type.

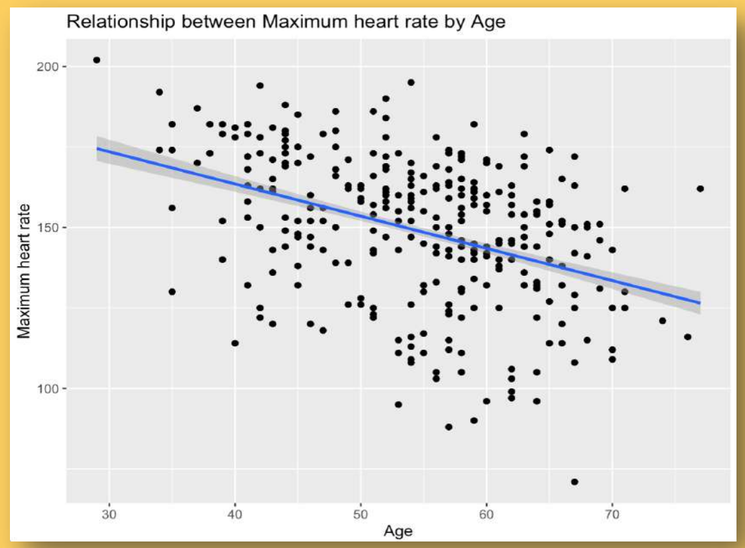
Maximum Heart Rate Achieved (thalach)

1st graph

- This graph show the distribution of Maximum Heart Rate Achieved in the dataset that is categorize by the disease.

2nd graph

- From this graph, the Maximum Heart Rate Achieved for each type of people are less overlap.
- People that have disease, The range is around 150 – 175, median is around 162.
- People that have don't disease, The range is around 125 to 155, median is around 145.



Maximum Heart Rate Achieved (thalach)

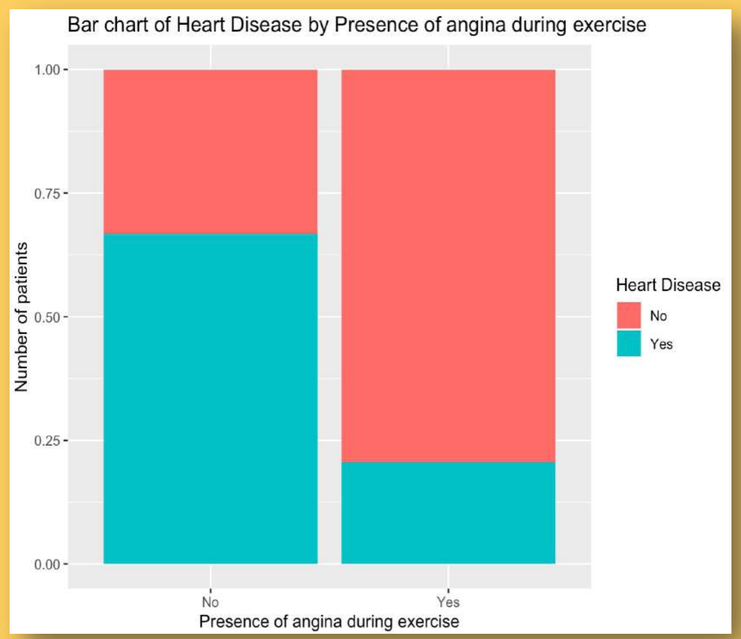
3rd graph

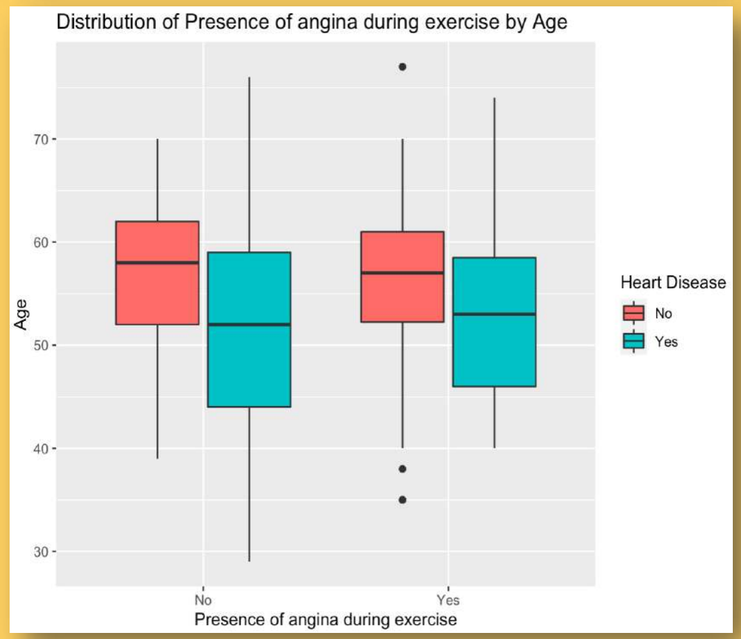
- Same as the second graph but are categorize by sex type.

4th graph

- The graph show the relationship between age and Maximum Heart Rate Achieved.
- Weak negative relationship

```
> cor(heartData_Modify$thalach, heartData_Modify$age)
[1] -0.3933922
```





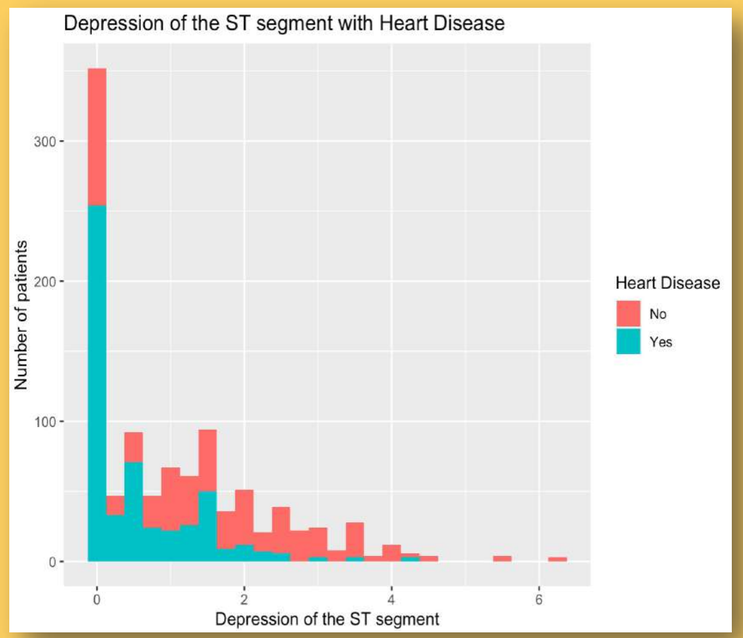
Exercise induced angina (exang)

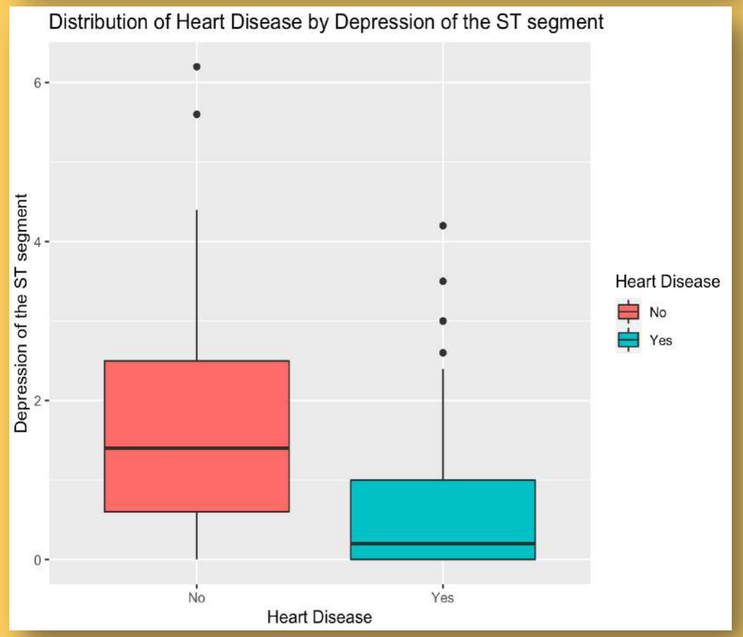
1st graph

- From this graph count the number people that have blood sugar in each type.
- Most of the people that have heart disease will not have Presence of angina.

2nd graph

- This graph shows the distribution of age in the dataset that is categorize by Presence of angina and disease.





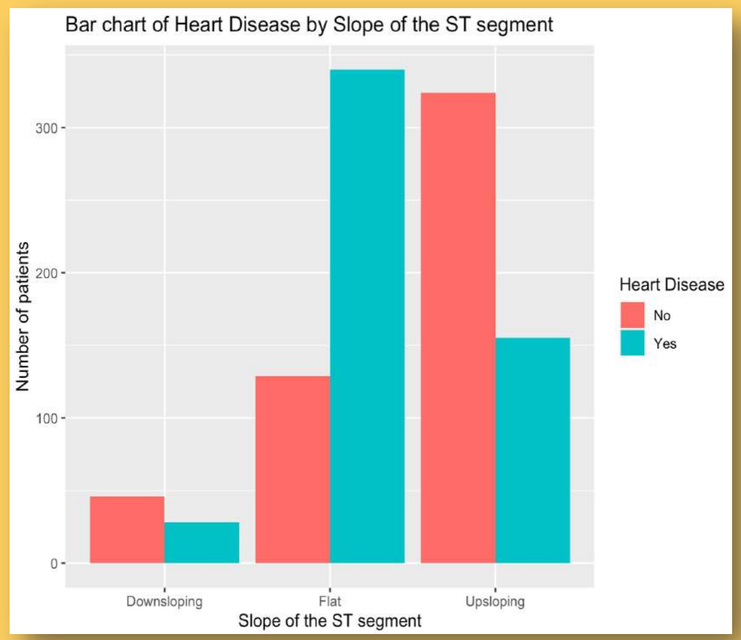
Depression of ST segment (oldpeak)

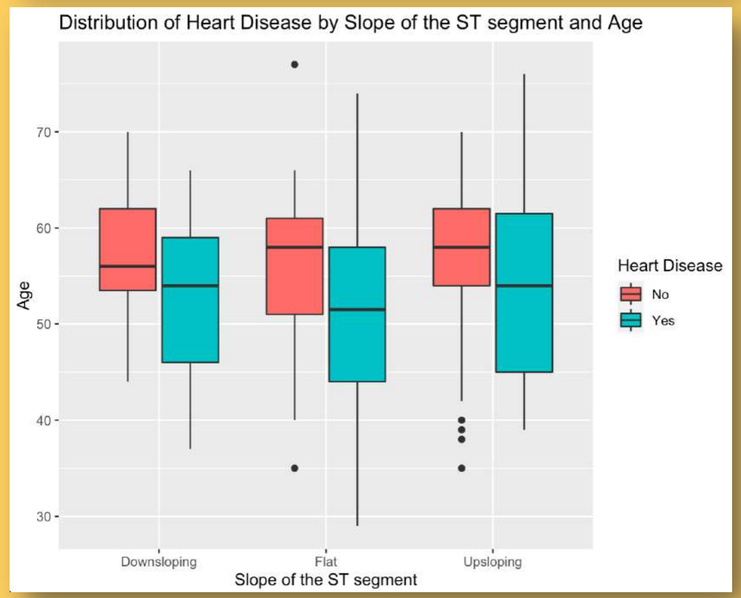
1st graph

- This graph show the distribution of Depression of ST segment in the dataset that is categorize by the disease.

2nd graph

- From this graph, Depression of ST segment for each type of people are less overlap and have noticeable different.
- People that have disease, The range is around 0 – 1, median is around 0.2.
- People that have don’t disease, The range is around 0.7 to 2.25, median is around 1.4.





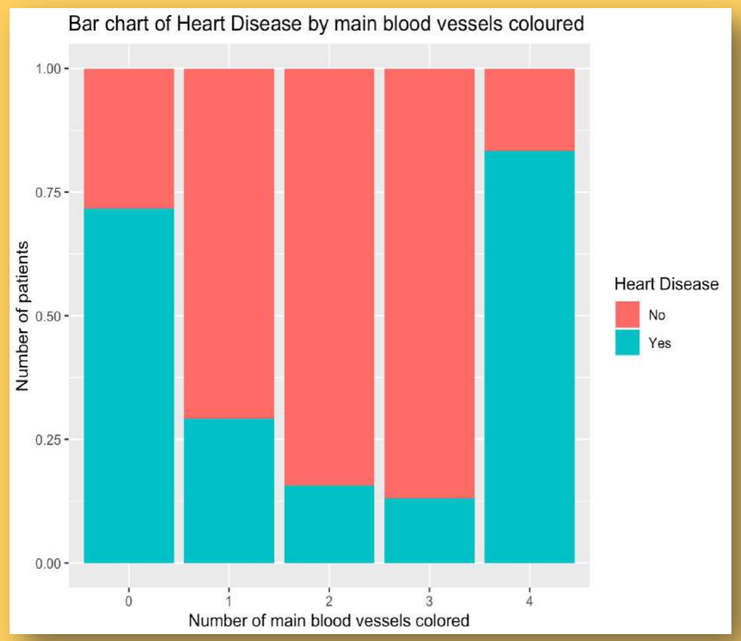
ST segment slope (slope)

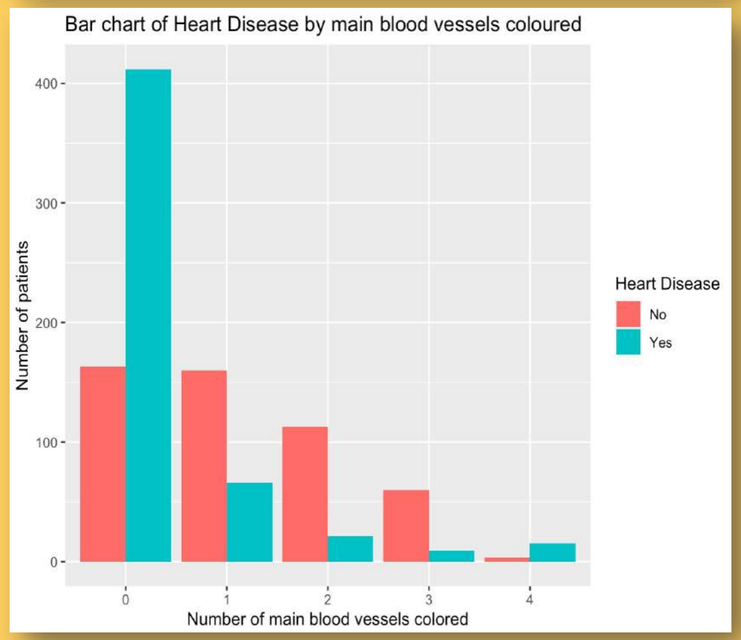
1st graph

- This graph count the number of patients that have different type of ST segment slope and categorize by the disease.
- The most popular type for heart disease patient is upsloping type.
- The most popular type for non heart disease patient is flat type.

2nd graph

- This graph show the distribution between ST segment slope and age in the dataset that is categorize by the disease.

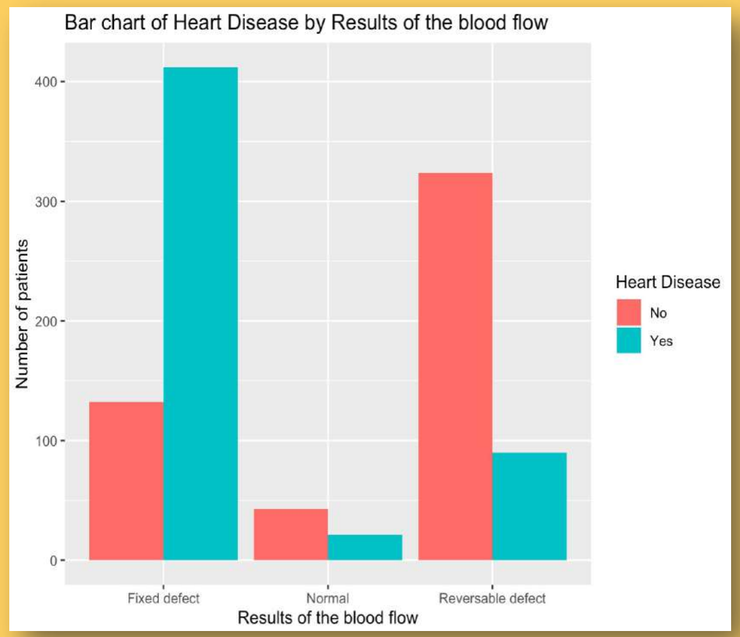


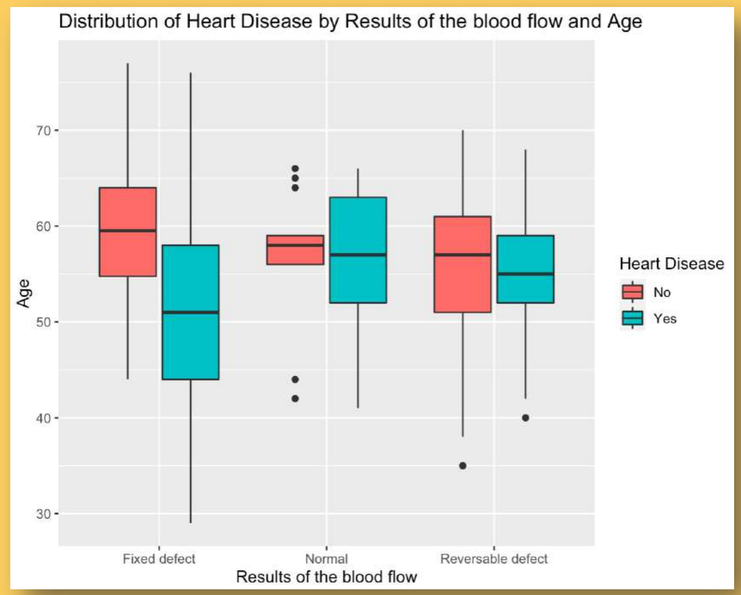


Number of Major Vessels (ca)

1st and 2nd graph

- This graph shows and count the number people that have Major Vessels for each type and categorize by the disease.
- These to graph can answer the question about the number of people that have each type of Number of Major Vessels (1-4) and see that they have heart disease or not.
- For example, People that don’t have heart disease tend to have type 0.





Thalassemia (thal)

1st graph

- This graph shows and count the number people that have each type of Thalassemia and categorize by the disease.
- Most of the people that have heart disease will have Reversable defect type of Thalassemia the most.
- Most of the people that have don't heart disease will have Fixed defect type of Thalassemia the most

2nd graph

- This graph shows the distribution of age in the dataset that is categorize by Thalassemia type and disease.

MODEL EXPLANATION

Yes

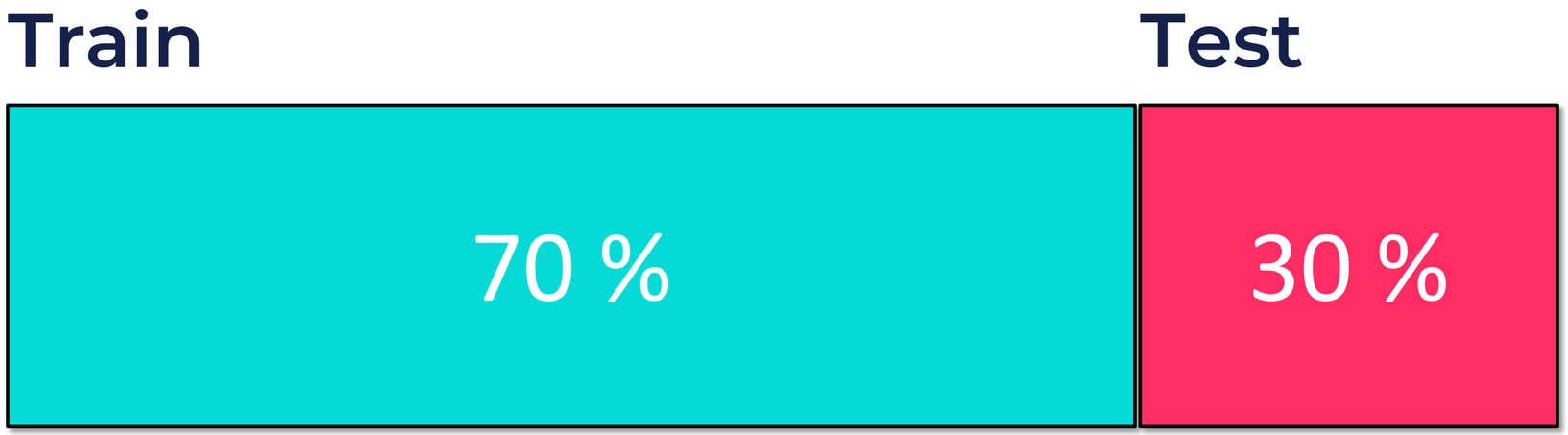
(Have Heart disease)

No

(Don't Have Heart disease)

We will create a model from the dataset to predict the variable **Target** (Yes or No). We will use Logistic Regression and Decision Tree to create a model

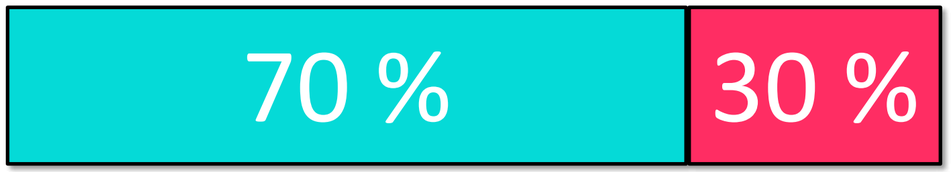
MODEL EXPLANATION



Split the data in to 2 Part, Train for training the model,
Test is for testing the result

MODEL EXPLANATION

```r
# training and testing data
set.seed(555)
test_ind <- sample(nrow(heartData_Modify), 0.3*nrow(heartData_Modify))
heartData_training <- heartData_Modify[-test_ind,]
heartData_testing <- heartData_Modify[test_ind,]

summary(heartData_testing)
summary(heartData_training)
```



Sample the data in to 2 Part

Test (30 %)

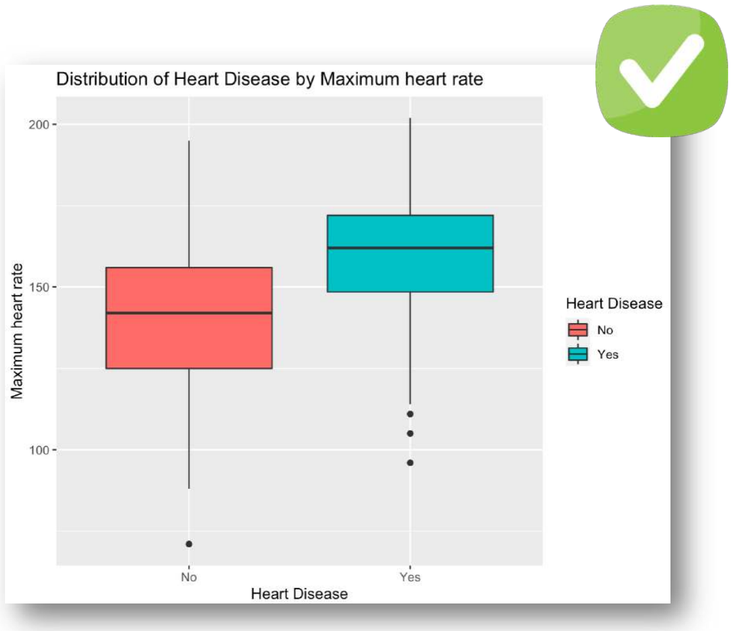
```
> summary(heartData_testing)
      age            sex                      cp          trestbps
 Min.   :29.00   Female: 90   Asymptomatic    :154   Min.   : 94.0
 1st Qu.:48.25   Male  :216   Atypical angina : 85   1st Qu.:120.0
 Median :56.00                Non-anginal Pain: 20   Median :130.0
 Mean   :54.66                Typical Angina  : 47   Mean   :132.1
 3rd Qu.:61.00                                       3rd Qu.:140.0
 Max.   :77.00                                       Max.   :200.0
      chol          fbs                        restecg       thalach
 Min.   :126.0   <=120:263   Normal                 :153   Min.   : 71.0
 1st Qu.:212.0   >120 : 43   Probable or definite :  4   1st Qu.:132.0
 Median :242.0               ST-T wave Abnormality:149   Median :152.0
 Mean   :243.8                                           Mean   :149.1
 3rd Qu.:274.0                                           3rd Qu.:168.0
 Max.   :417.0                                           Max.   :202.0
 exang         oldpeak              slope      ca
 No :202   Min.   :0.000   Downsloping: 32   0:163
 Yes:104   1st Qu.:0.000   Flat       :133   1: 66
           Median :0.800   Upsloping  :141   2: 47
           Mean   :1.084                     3: 22
           3rd Qu.:1.600                     4:  8
           Max.   :6.200
                thal     target
 Fixed defect     :166   No :153
 Normal           : 22   Yes:153
 Reversable defect:118
```

Train (70 %)

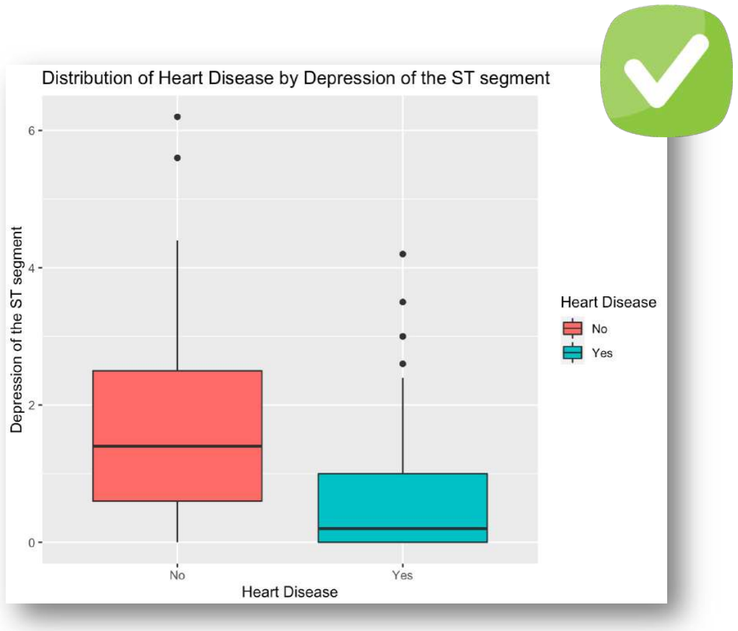
```
> summary(heartData_training)
      age            sex                      cp          trestbps
 Min.   :29.00   Female:219   Asymptomatic    :343   Min.   : 94.0
 1st Qu.:47.00   Male  :497   Atypical angina :196   1st Qu.:120.0
 Median :55.00                Non-anginal Pain: 57   Median :130.0
 Mean   :54.28                Typical Angina  :120   Mean   :131.5
 3rd Qu.:61.00                                       3rd Qu.:140.0
 Max.   :77.00                                       Max.   :200.0
      chol          fbs                        restecg       thalach
 Min.   :126.0   <=120:606   Normal                 :341   Min.   : 71.0
 1st Qu.:210.0   >120 :110   Probable or definite : 11   1st Qu.:133.0
 Median :239.0               ST-T wave Abnormality:364   Median :152.0
 Mean   :245.6                                           Mean   :149.1
 3rd Qu.:275.0                                           3rd Qu.:165.0
 Max.   :417.0                                           Max.   :202.0
 exang         oldpeak              slope      ca
 No :475   Min.   :0.000   Downsloping: 42   0:412
 Yes:241   1st Qu.:0.000   Flat       :336   1:160
           Median :0.800   Upsloping  :338   2: 87
           Mean   :1.064                     3: 47
           3rd Qu.:1.800                     4: 10
           Max.   :6.200
                thal     target
 Fixed defect     :378   No :346
 Normal           : 42   Yes:370
 Reversable defect:296
```

Logistics Regression

Relationship between Numerical Variable and The Target Variable

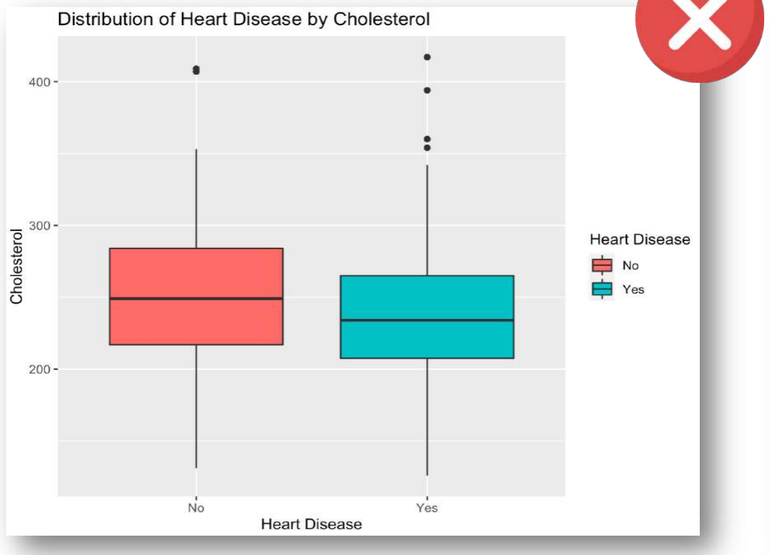


thalach

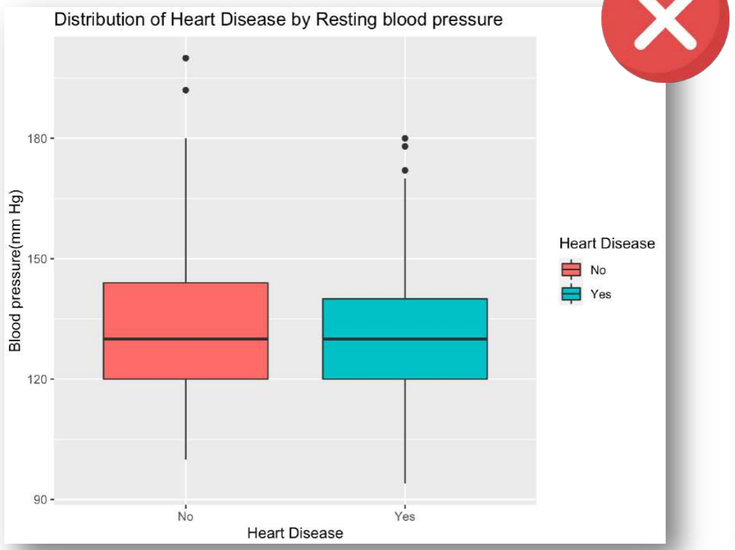


Old peak

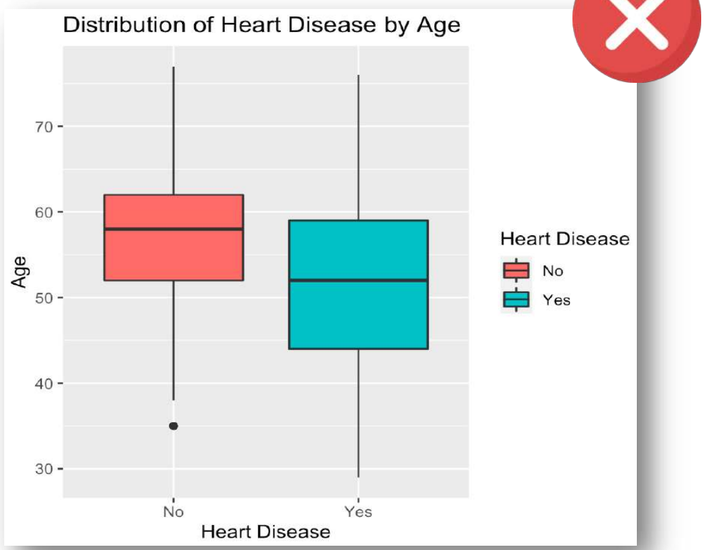
Relationship between Numerical Variable and The Target Variable



Chol



Trestbps



Age

Relationship between Categorical Variable and The Target Variable

```
> chisq.test(table(heartData_Modify$sex,heartData_Modify$target))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(heartData_Modify$sex, heartData_Modify$target)
X-squared = 76.93, df = 1, p-value < 2.2e-16
```

```
> chisq.test(table(heartData_Modify$fbs,heartData_Modify$target))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(heartData_Modify$fbs, heartData_Modify$target)
X-squared = 1.421, df = 1, p-value = 0.2332
```

```
> chisq.test(table(heartData_Modify$slope,heartData_Modify$target))

        Pearson's Chi-squared test

data:  table(heartData_Modify$slope, heartData_Modify$target)
X-squared = 158.46, df = 2, p-value < 2.2e-16
```

```
> chisq.test(table(heartData_Modify$cp,heartData_Modify$target))

        Pearson's Chi-squared test

data:  table(heartData_Modify$cp, heartData_Modify$target)
X-squared = 278.72, df = 3, p-value < 2.2e-16
```

```
> chisq.test(table(heartData_Modify$restecg,heartData_Modify$target))

        Pearson's Chi-squared test

data:  table(heartData_Modify$restecg, heartData_Modify$target)
X-squared = 36.842, df = 2, p-value = 9.997e-09
```

```
> chisq.test(table(heartData_Modify$exang,heartData_Modify$target))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(heartData_Modify$exang, heartData_Modify$target)
X-squared = 193.26, df = 1, p-value < 2.2e-16
```

```
> chisq.test(table(heartData_Modify$ca,heartData_Modify$target))

        Pearson's Chi-squared test

data:  table(heartData_Modify$ca, heartData_Modify$target)
X-squared = 255.36, df = 4, p-value < 2.2e-16
```

```
> chisq.test(table(heartData_Modify$thal,heartData_Modify$target))

        Pearson's Chi-squared test

data:  table(heartData_Modify$thal, heartData_Modify$target)
X-squared = 283.53, df = 2, p-value < 2.2e-16
```

Logistics Regression

```r
# remove fbs,restecg,age,chol and trestbps
model <- glm(target ~ oldpeak+thalach+sex+cp+exang+slope+ca+thal, data = heartData_training, family = binomial)
summary(model)

res <- predict(model,heartData_testing,type = "response")
res
res_c <- factor(ifelse(res > 0.6,"Yes","No"))
res_c

confusionMatrix(res_c,heartData_testing$target,mode="prec_recall", positive="Yes")
```

Logistics Regression

```r
# remove fbs,restecg,age,chol and trestbps
model <- glm(target ~ oldpeak+thalach+sex+cp+exang+slope+ca+thal, data = heartData_training, family = binomial)
summary(model)

res <- predict(model,heartData_testing,type = "response")
res
res_c <- factor(ifelse(res > 0.6,"Yes","No"))
res_c

confusionMatrix(res_c,heartData_testing$target,mode="prec_recall", positive="Yes")
```

```
glm(formula = target ~ oldpeak + thalach + sex + cp + exang +
    slope + ca + thal, family = binomial, data = heartData_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9534  -0.2739   0.0723   0.4078   3.2424

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.817220   1.305171   0.626  0.53122
oldpeak                  -0.629823   0.156656  -4.020 5.81e-05 ***
thalach                   0.011510   0.007191   1.601  0.10945
sexMale                  -1.731944   0.366854  -4.721 2.35e-06 ***
cpAtypical angina         2.147406   0.354349   6.060 1.36e-09 ***
cpNon-anginal Pain        2.823071   0.474417   5.951 2.67e-09 ***
cpTypical Angina          0.712454   0.355788   2.002  0.04523 *
exangYes                 -0.826619   0.300387  -2.752  0.00593 **
slopeFlat                 1.001507   0.610639   1.640  0.10098
slopeUpsloping           -0.791268   0.570670  -1.387  0.16558
ca1                      -2.350111   0.347104  -6.771 1.28e-11 ***
ca2                      -3.396475   0.554798  -6.122 9.24e-10 ***
ca3                      -1.522765   0.588900  -2.586  0.00972 **
ca4                       2.035045   1.024577   1.986  0.04701 *
thalNormal                0.882154   0.535881   1.646  0.09973 .
thalReversable defect    -1.451335   0.290266  -5.000 5.73e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 991.78  on 715  degrees of freedom
Residual deviance: 420.77  on 700  degrees of freedom
AIC: 452.77

Number of Fisher Scoring iterations: 6
```

Predicting Result

```
> res
          1           2           3           4           5           6           7           8           9
0.9849221370 0.9568024271 0.2535817731 0.6122908199 0.4452328230 0.0706893383 0.0090378057 0.9953497482 0.7973522535
         10          11          12          13          14          15          16          17          18
0.9016846327 0.0186941197 0.9980035368 0.0011721036 0.9972034068 0.7636061470 0.2301074840 0.9829686566 0.8468001657
         19          20          21          22          23          24          25          26          27
0.0305544102 0.2952855763 0.8943075532 0.4987203321 0.0577110853 0.0346118674 0.0467575543 0.0228194101 0.8640814975
         28          29          30          31          32          33          34          35          36
0.3494684439 0.0667404662 0.5826488406 0.0043116688 0.8653220452 0.0122141115 0.0323557675 0.4437135086 0.0038395120
         37          38          39          40          41          42          43          44          45
0.0471977325 0.7657559742 0.0090378057 0.9980931968 0.9865393467 0.2414098322 0.7940072277 0.0242463833 0.0471977325
         46          47          48          49          50          51          52          53          54
0.0160688213 0.4017872503 0.6786020032 0.6122908199 0.4412245541 0.0348368218 0.7605748561 0.0006265594 0.0186941197
         55          56          57          58          59          60          61          62          63
0.0335328380 0.0325620874 0.6929995668 0.9402452903 0.0955199536 0.5833943893 0.6368702414 0.7583071709 0.0294434458
         64          65          66          67          68          69          70          71          72
0.2715260049 0.6786020032 0.6900422193 0.5833943893 0.9872396673 0.5786269696 0.0467575543 0.6197565963 0.9518542733
         73          74          75          76          77          78          79          80          81
0.9427801223 0.0028839439 0.9805378276 0.9518542733 0.0043116688 0.0192372222 0.9310533853 0.9614276632 0.6236100279
         82          83          84          85          86          87          88          89          90
0.0620826328 0.2066197608 0.0255890024 0.0242463833 0.9578801869 0.9891304040 0.5786269696 0.7055976531 0.0061289067
         91          92          93          94          95          96          97          98          99
0.0327460549 0.0124669222 0.5529706105 0.1344888369 0.5529706105 0.3551051618 0.2066197608 0.9148788930 0.0061289067
        100         101         102         103         104         105         106         107         108
0.9975915889 0.2301074840 0.0015396025 0.0004205080 0.9277240632 0.7412059713 0.8640814975 0.9755413850 0.9850220259
        109         110         111         112         113         114         115         116         117
0.4023573272 0.7518841779 0.0094488291 0.9619172186 0.9956793631 0.0011721036 0.0618823513 0.8985808962 0.1079444393
        118         119         120         121         122         123         124         125         126
0.7009693534 0.9980035368 0.6504835375 0.8943075532 0.6056923895 0.7208764645 0.8053251358 0.9945150392 0.3769913670
        127         128         129         130         131         132         133         134         135
0.6948406909 0.7657559742 0.1344888369 0.2535817731 0.1734138917 0.7076890274 0.9872396673 0.6786020032 0.2758197198
        136         137         138         139         140         141         142         143         144
```

```
> res_c
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30
Yes Yes  No Yes  No  No  No Yes Yes Yes  No Yes  No Yes Yes  No Yes Yes  No  No Yes  No  No  No  No  No Yes  No  No  No
 31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
 No Yes  No  No  No  No  No Yes  No Yes Yes  No Yes  No  No  No  No Yes Yes  No  No Yes  No  No  No  No Yes Yes  No  No
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
Yes Yes  No  No Yes Yes  No Yes  No  No Yes Yes Yes  No Yes Yes  No  No Yes Yes Yes  No  No  No  No Yes Yes  No Yes  No
 91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
 No  No  No  No  No  No  No Yes  No Yes  No  No  No Yes Yes Yes Yes Yes  No Yes  No Yes Yes  No  No Yes  No Yes Yes Yes
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150
Yes Yes Yes Yes Yes  No Yes Yes  No  No  No Yes Yes Yes  No Yes  No  No Yes Yes Yes Yes  No Yes  No  No Yes  No Yes  No
151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
 No  No Yes  No  No  No  No  No Yes Yes  No  No Yes Yes Yes Yes Yes Yes Yes Yes  No  No Yes  No  No  No Yes  No Yes  No
181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210
Yes  No  No Yes  No  No  No Yes Yes  No  No  No Yes Yes  No Yes Yes  No Yes  No Yes Yes  No Yes  No  No  No  No Yes Yes
211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240
 No Yes Yes  No  No  No Yes Yes Yes  No  No Yes Yes  No Yes  No Yes  No  No Yes  No Yes Yes Yes Yes  No  No  No  No  No
241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270
Yes  No Yes Yes  No  No  No Yes Yes  No Yes  No  No Yes  No Yes  No  No Yes Yes Yes Yes  No Yes Yes  No Yes  No Yes Yes
271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300
 No  No  No Yes  No Yes  No  No  No  No Yes  No Yes Yes Yes Yes Yes  No  No Yes Yes Yes Yes Yes  No  No Yes  No  No  No
301 302 303 304 305 306
Yes Yes  No  No  No  No
Levels: No Yes
```

Confusion Matrix

Prediction	Reference No	Reference Yes
No	137	24
Yes	16	129

Logistics Regression model

```
> confusionMatrix(res_c,heartData_testing$target,mode="prec_recall", positive="Yes")
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  137  24
       Yes  16 129

               Accuracy : 0.8693
                 95% CI : (0.8263, 0.9049)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7386

 Mcnemar's Test P-Value : 0.2684

              Precision : 0.8897
                 Recall : 0.8431
                     F1 : 0.8658
             Prevalence : 0.5000
         Detection Rate : 0.4216
   Detection Prevalence : 0.4739
      Balanced Accuracy : 0.8693

       'Positive' Class : Yes
```

Confusion Matrix

Prediction	Reference No	Reference Yes
No	137	24
Yes	16	129

Model Evaluation

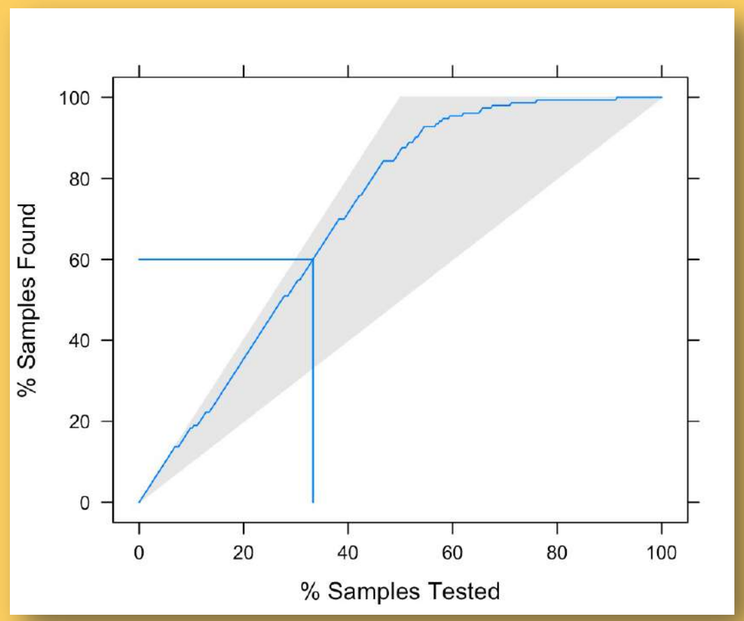
```
                Accuracy : 0.8693
                  95% CI : (0.8263, 0.9049)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                   Kappa : 0.7386

 Mcnemar's Test P-Value : 0.2684
```

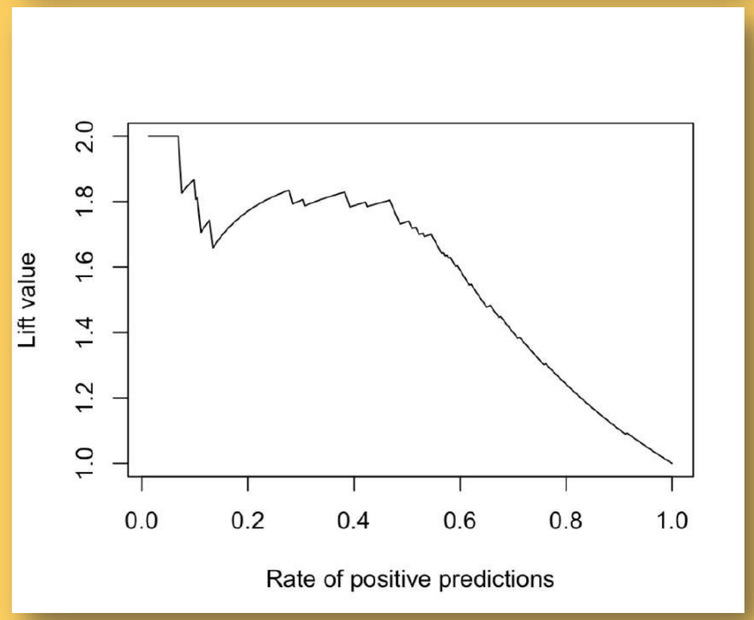
Confusion Matrix

	Reference	
Prediction	No	Yes
No	137	24
Yes	16	129

Model Evaluation

Precision	0.8897
Recall	0.8431
F1	0.8658
Prevalence	0.5000
Detection Rate	0.4216
Detection Prevalence	0.4739
Balanced Accuracy	0.8693
'Positive' Class	Yes





Lift Analysis

```r
# lift
res_dataFrame <- as.data.frame(res)
res_dataFrame %>%
  mutate('actual' = heartData_testing$target) %>%
  rename('prob' = 'res') %>%
  arrange(desc(prob)) -> lift_result
head(lift_result)

lift_obj <- lift(actual ~ prob,data = lift_result,class="Yes")
plot(lift_obj, values = 60)

pred <- prediction(res, heartData_testing$target)
perf_lift <- performance(pred, "lift", "rpp")
plot(perf_lift)

TopDecileLift(res,as.integer(heartData_testing$target)-1)
```

```
> TopDecileLift(res,as.integer(heartData_testing$target)-1)
[1] 1.806
```

Cross validation

```r
# cross-validation
train_control <- trainControl(method="cv", number=10)
model_cv <- train(target ~ oldpeak+thalach+sex+cp+exang+slope+ca+thal, data=heartData_testing,
                  trControl=train_control,
                  method="glm")
model_cv
model_cv$finalModel
model_cv$resample
```

```
> model_cv
Generalized Linear Model

306 samples
  8 predictor
  2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 275, 275, 276, 276, 275, 276, ...
Resampling results:

  Accuracy   Kappa
  0.8798522  0.7599022
```

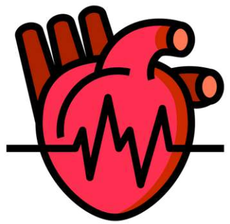
```
> model_cv$finalModel

Call:  NULL

Coefficients:
           (Intercept)                 oldpeak                thalach                  sexMale
            -0.9245408              -0.0978633               0.0202268              -1.3257087
    `cpAtypical angina`      `cpNon-anginal Pain`  `cpTypical Angina`                 exangYes
             2.2246908               0.9834410               1.6384907              -0.6037708
             slopeFlat           slopeUpsloping                     ca1                     ca2
             0.4059217              -0.0004244              -2.2746672              -3.1818647
                   ca3                     ca4             thalNormal  `thalReversable defect`
            -3.8306573               0.8081473              -0.9681917              -1.8014466

Degrees of Freedom: 305 Total (i.e. Null);  290 Residual
Null Deviance:      424.2
Residual Deviance: 192  AIC: 224
```

```
> model_cv$resample
    Accuracy     Kappa Resample
1  0.8709677 0.7416667   Fold01
2  0.8709677 0.7405858   Fold02
3  0.9000000 0.8000000   Fold03
4  0.8333333 0.6666667   Fold04
5  0.7419355 0.4876033   Fold05
6  0.9333333 0.8666667   Fold06
7  0.9666667 0.9333333   Fold07
8  0.9333333 0.8666667   Fold08
9  0.9354839 0.8708333   Fold09
10 0.8125000 0.6250000   Fold10
```

Decision Tree

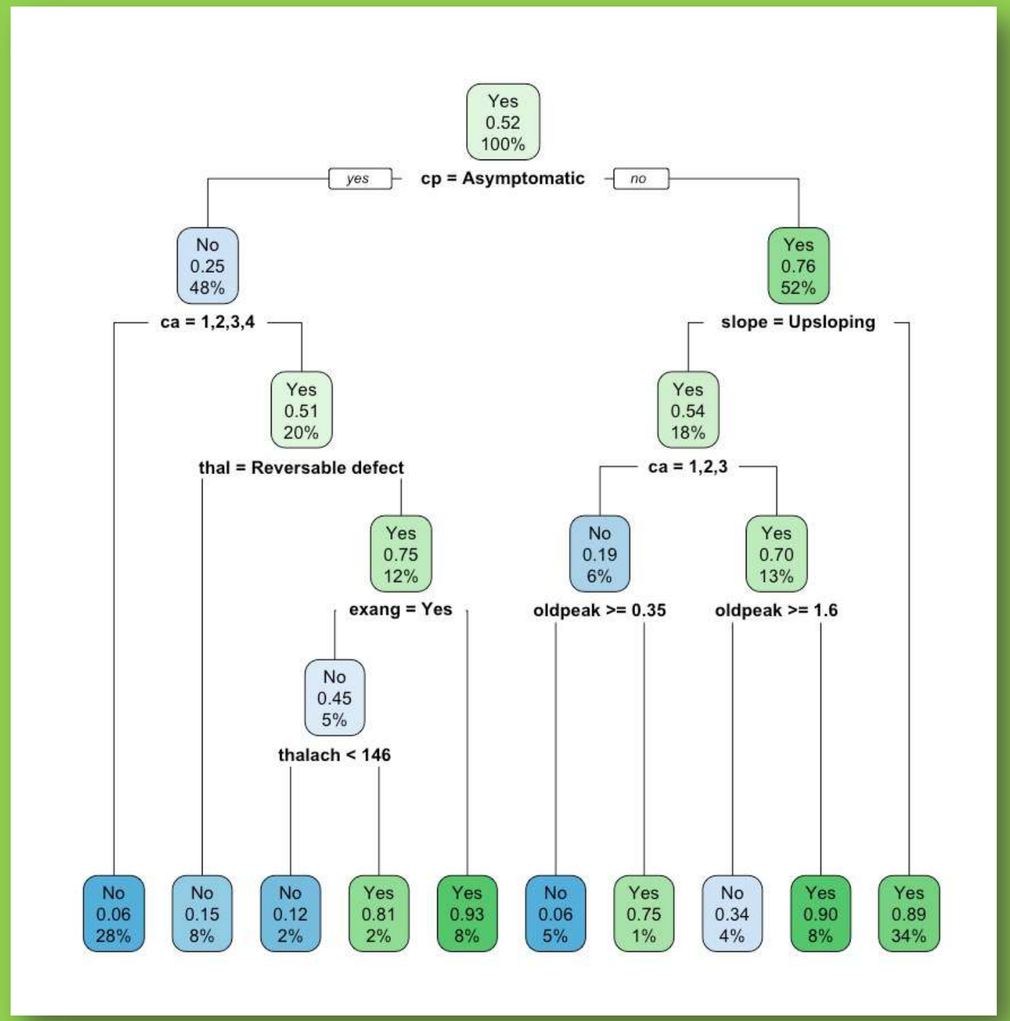
Decision Tree

```r
# Decision tree model
tree <- rpart(target ~ ., data =  heartData_training)
rpart.plot(tree)
tree$variable.importance
summary(tree)

res2 = predict(tree,heartData_testing)
head(res2)
res_t = predict(tree,heartData_testing,type = "class")
res_t

confusionMatrix(res_t,heartData_testing$target, positive="Yes")
```



Decision Tree

```r
# Decision tree model
tree <- rpart(target ~ ., data =  heartData_training)
rpart.plot(tree)
tree$variable.importance
summary(tree)

res2 = predict(tree,heartData_testing)
head(res2)
res_t = predict(tree,heartData_testing,type = "class")
res_t

confusionMatrix(res_t,heartData_testing$target, positive="Yes")
```



Predicting Result

```
> head(res2)
          No       Yes
1 0.11203320 0.8879668
2 0.11203320 0.8879668
3 0.25000000 0.7500000
4 0.07407407 0.9259259
5 0.07407407 0.9259259
6 0.84745763 0.1525424
```

```
> res_t
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26
Yes Yes Yes Yes Yes  No  No Yes Yes Yes  No Yes  No Yes Yes  No Yes Yes  No  No Yes  No  No  No  No  No
 27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52
Yes Yes Yes Yes  No Yes  No  No  No  No  No Yes  No Yes Yes  No Yes  No  No  No  No Yes Yes Yes  No Yes
 53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78
 No  No  No  No Yes Yes  No Yes Yes Yes  No  No Yes Yes Yes Yes Yes  No  No Yes Yes  No Yes Yes  No  No
 79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104
Yes Yes Yes  No  No  No  No Yes Yes Yes Yes  No  No  No Yes  No Yes  No  No Yes  No Yes  No  No  No Yes
105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130
Yes Yes Yes Yes  No Yes  No Yes Yes  No Yes Yes  No Yes Yes Yes Yes Yes Yes Yes  No Yes Yes  No Yes
131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156
 No Yes Yes Yes  No Yes Yes  No Yes Yes Yes Yes Yes Yes  No  No Yes Yes Yes  No  No  No Yes  No  No  No
157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182
 No Yes  No Yes  No  No Yes  No Yes Yes Yes Yes Yes  No  No Yes  No  No  No Yes  No Yes  No  No Yes  No
183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208
 No  No  No  No Yes Yes Yes  No  No  No Yes Yes  No Yes Yes  No Yes  No Yes Yes  No Yes  No  No  No  No
209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234
Yes Yes  No Yes Yes  No  No  No Yes Yes Yes Yes  No Yes Yes  No Yes  No Yes  No Yes Yes  No Yes Yes Yes
235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260
Yes  No  No Yes  No  No Yes  No Yes Yes  No  No Yes Yes Yes  No Yes  No  No Yes  No Yes  No  No Yes Yes
261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286
Yes Yes  No Yes Yes Yes Yes  No Yes  No  No  No Yes Yes Yes Yes  No  No  No  No Yes  No Yes Yes Yes Yes
287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306
Yes  No Yes Yes Yes Yes Yes Yes  No  No Yes  No  No  No Yes Yes  No Yes  No  No
Levels: No Yes
```

Confusion Matrix

```
          Reference
Prediction  No Yes
       No  125  13
       Yes  28 140
```

Decision Tree model

```
> confusionMatrix(res_t,heartData_testing$target, positive="Yes")
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  125  13
       Yes  28 140

               Accuracy : 0.866
                 95% CI : (0.8227, 0.9021)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.732

 Mcnemar's Test P-Value : 0.02878

            Sensitivity : 0.9150
            Specificity : 0.8170
         Pos Pred Value : 0.8333
         Neg Pred Value : 0.9058
             Prevalence : 0.5000
         Detection Rate : 0.4575
   Detection Prevalence : 0.5490
      Balanced Accuracy : 0.8660

       'Positive' Class : Yes
```

Decision Tree model

Confusion Matrix

```
          Reference
Prediction  No Yes
       No  125  13
       Yes  28 140
```

```
               Accuracy : 0.866
                 95% CI : (0.8227, 0.9021)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.732

 Mcnemar's Test P-Value : 0.02878
```

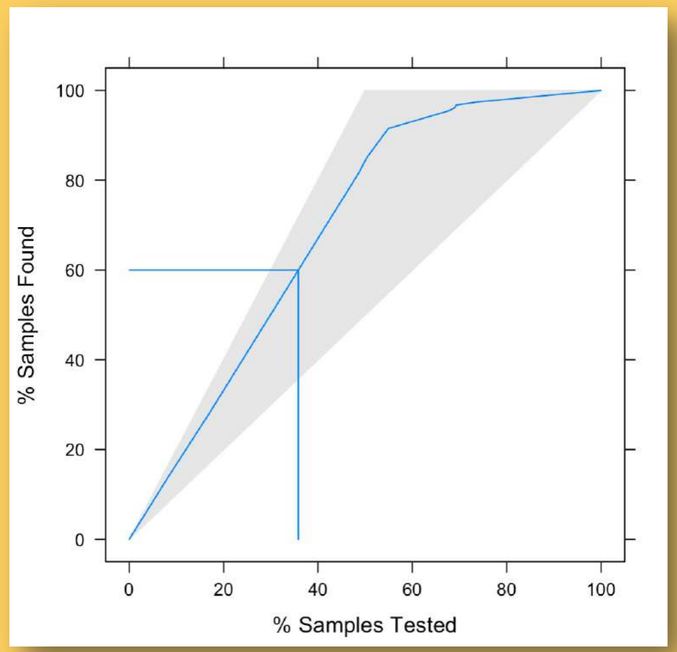
Decision Tree model

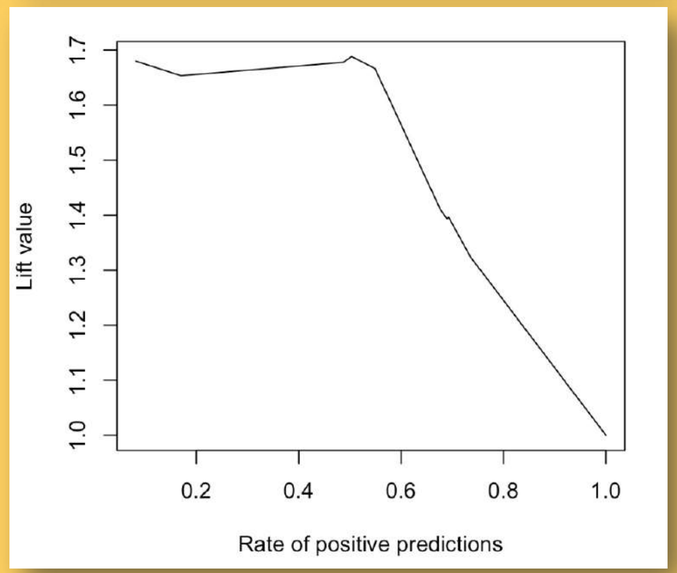
Confusion Matrix

```
          Reference
Prediction  No Yes
       No  125  13
       Yes  28 140
```

```
               Sensitivity : 0.9150
               Specificity : 0.8170
            Pos Pred Value : 0.8333
            Neg Pred Value : 0.9058
                Prevalence : 0.5000
            Detection Rate : 0.4575
      Detection Prevalence : 0.5490
         Balanced Accuracy : 0.8660
```

```
          Precision : 0.8333
             Recall : 0.9150
                 F1 : 0.8723
```





Lift Analysis

```r
# Lift calculation
res.p <- predict(tree,heartData_testing)[,"Yes"]
lift_result <- data.frame(
  prob = res.p,
  y = heartData_testing$target)
lift_obj <- lift(y ~ prob,
                 data = lift_result,
                 class="Yes")
plot(lift_obj, values = 60)

# Lift Chart
pred <- prediction(res.p, heartData_testing$target,
                   label.ordering = c("No","Yes"))
perf_lift <- performance(pred, "lift", "rpp")
plot(perf_lift)

# Lift at 10%
TopDecileLift(res.p,as.integer(heartData_testing$target)-1)
```

```
> TopDecileLift(res.p,as.integer(heartData_testing$target)-1)
[1] 1.677
```

Cross validation

```r
# cv
train_control <- trainControl(method="cv", number=10)
model_cv <- train(target ~ ., data=heartData_testing,
                  trControl=train_control,
                  method="rpart")
model_cv
model_cv$resample
```

```
> model_cv$resample
    Accuracy     Kappa Resample
1  0.7333333 0.4666667   Fold02
2  0.8709677 0.7416667   Fold01
3  0.8064516 0.6141079   Fold03
4  0.8064516 0.6141079   Fold06
5  0.5000000 0.0000000   Fold05
6  0.7419355 0.4854772   Fold04
7  0.8000000 0.6000000   Fold07
8  0.7096774 0.4175365   Fold10
9  0.8333333 0.6666667   Fold09
10 0.6774194 0.3487395   Fold08
```

```
CART

306 samples
 13 predictor
  2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 275, 276, 275, 275, 276, 275, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.03921569  0.7479570  0.49549690
  0.06209150  0.7317204  0.46366005
  0.47058824  0.5362366  0.08792982

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.03921569.
```

THANK YOU