Project on Heart Diseases

Heart Disease Analysis and Predictions

AUTHOR
Akshatha

PUBLISHED
January 15, 2023

ABSTRACT
This report presents an analysis of heart disease data using R Studio. The data set includes demographic information and various medical measurements for a group of patients, and the goal of the analysis is to identify risk factors for heart disease, indicate whether chest results to heart disease and the kind of chest pain patients experience and predict whether a person's age and gender affect their risk of heart disease.Exploratory data analysis techniques such as visualizations were used to examine the relationship between different variables and the presence of heart disease. Predicted whether a person has heart disease using the Logistics regression model. Results suggest that risk factors for heart disease include Chest pain, gender, and ca(number of major vessels), while factors such as exercise, regular check-ups, and others may have a protective effect.

Introduction

For this project I aim to analyse the heart disease dataset and predict whether a persons age and gender affect their risk of causing heart disease. The dataset we are using was downloaded from Kaggle websites that are set dates from 1988 till 2019 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them.

To achieve this, making use of R Studio and required libraries to visualise, analyse and predict the dataset.

Questions

1) How do a person's age and gender affect their risk of heart disease? (Prediction)

2) What are the other risk factors causing the heart disease?

3) Are patients experiencing chest pain likely to have heart disease?

4)IS chest pain an indication of heart disease and What kind of chest pain do most heart patients experience?

Data

Dataset has 14 variables with 1025 observations.

Attribute Description

• *age*: The person's age in years

• *sex*: The person's sex (1 = male, 0 = female)

• *cp*: chest pain type

– Value 0: asymptomatic

– Value 1: atypical angina

– Value 2: non-anginal pain

– Value 3: typical angina

• *trestbps*: The person's resting blood pressure (mm Hg on admission to the hospital)

• *chol*: The person's cholesterol measurement in mg/dl

• *fbs*: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)

• *restecg*: resting electrocardiographic results

– Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria

– Value 1: normal

– Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

• *thalach*: The person's maximum heart rate achieved

• *exang*: Exercise induced angina (1 = yes; 0 = no)

• *oldpeak*: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.)

• *slope*: the slope of the peak exercise ST segment

0: downsloping; 1: flat; 2: upsloping

• *ca*: The number of major vessels (0-4)

• *thal* Results of the blood flow observed via the radioactive dye.

– Value 0: NULL (dropped from the dataset previously)

– Value 1: fixed defect (no blood flow in some part of the heart)

– Value 2: normal blood flow

– Value 3: reversible defect (a blood flow is observed but it is not normal)

• *target*: Heart disease (0 = no, 1 = yes)

▶ Code

```
[1] 0
```

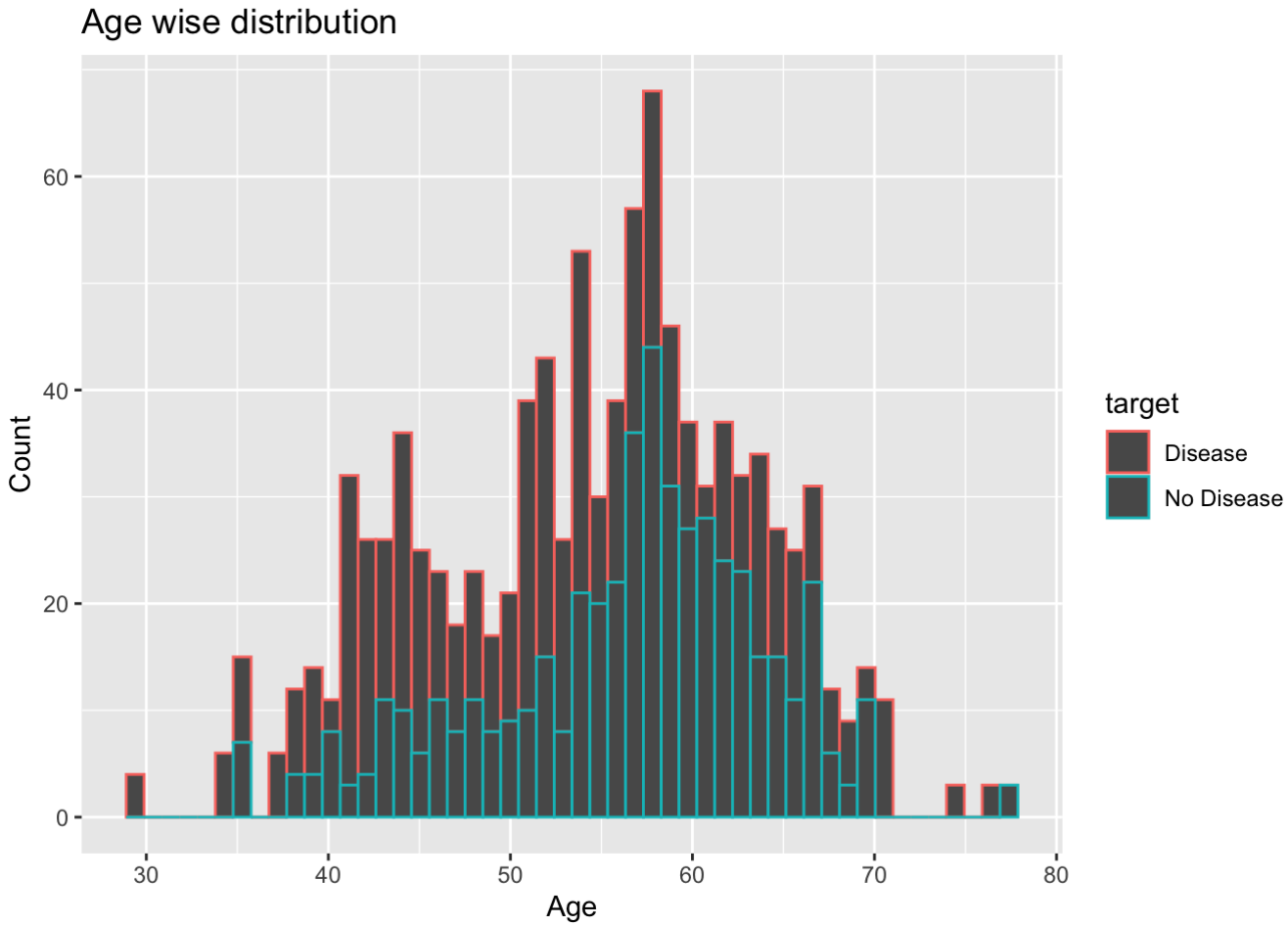
Analysis

▶ Code

▶ Code

1) How do a person's age and gender affect their risk of heart disease?

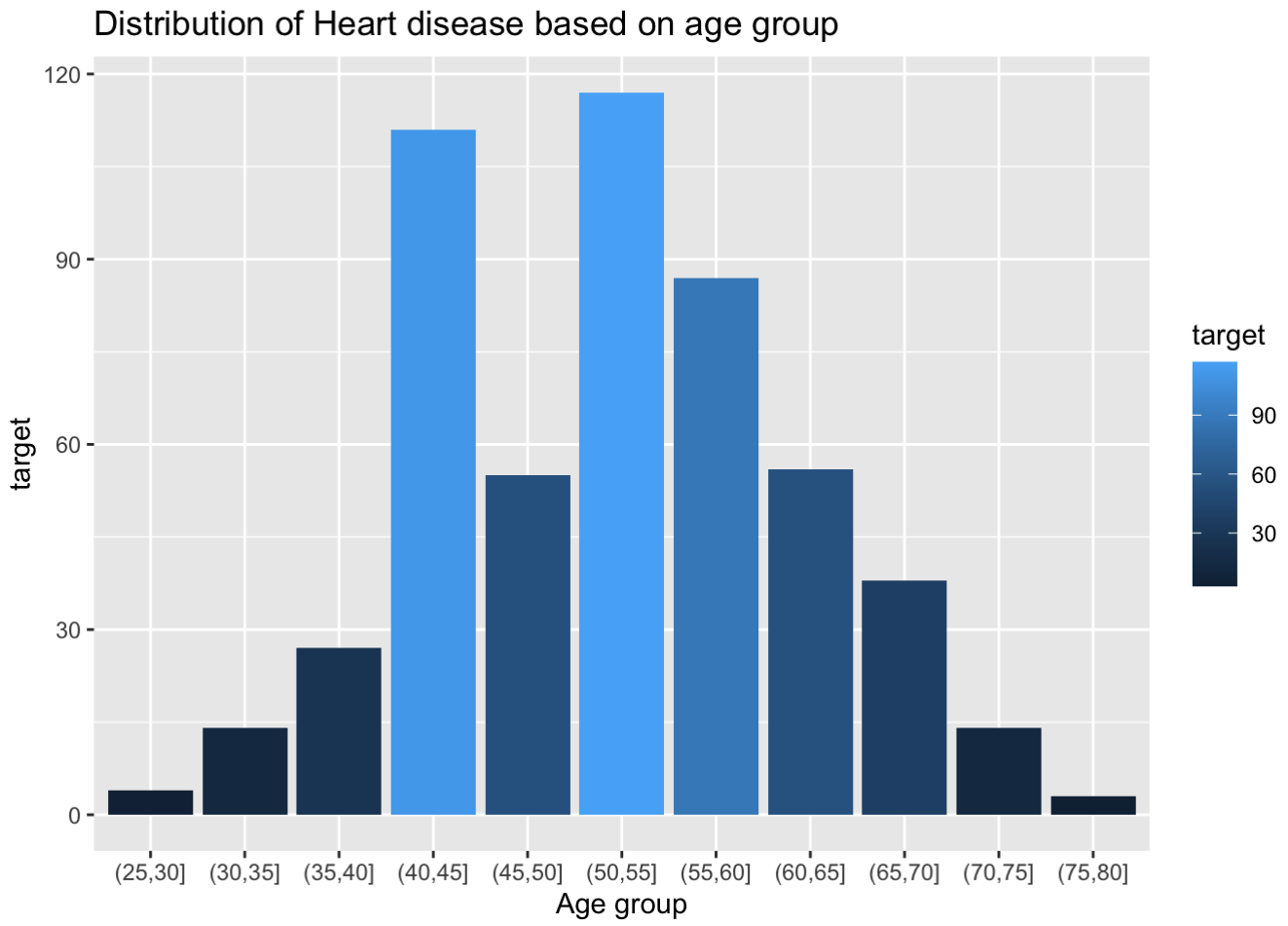
▶ Code



▶ Code

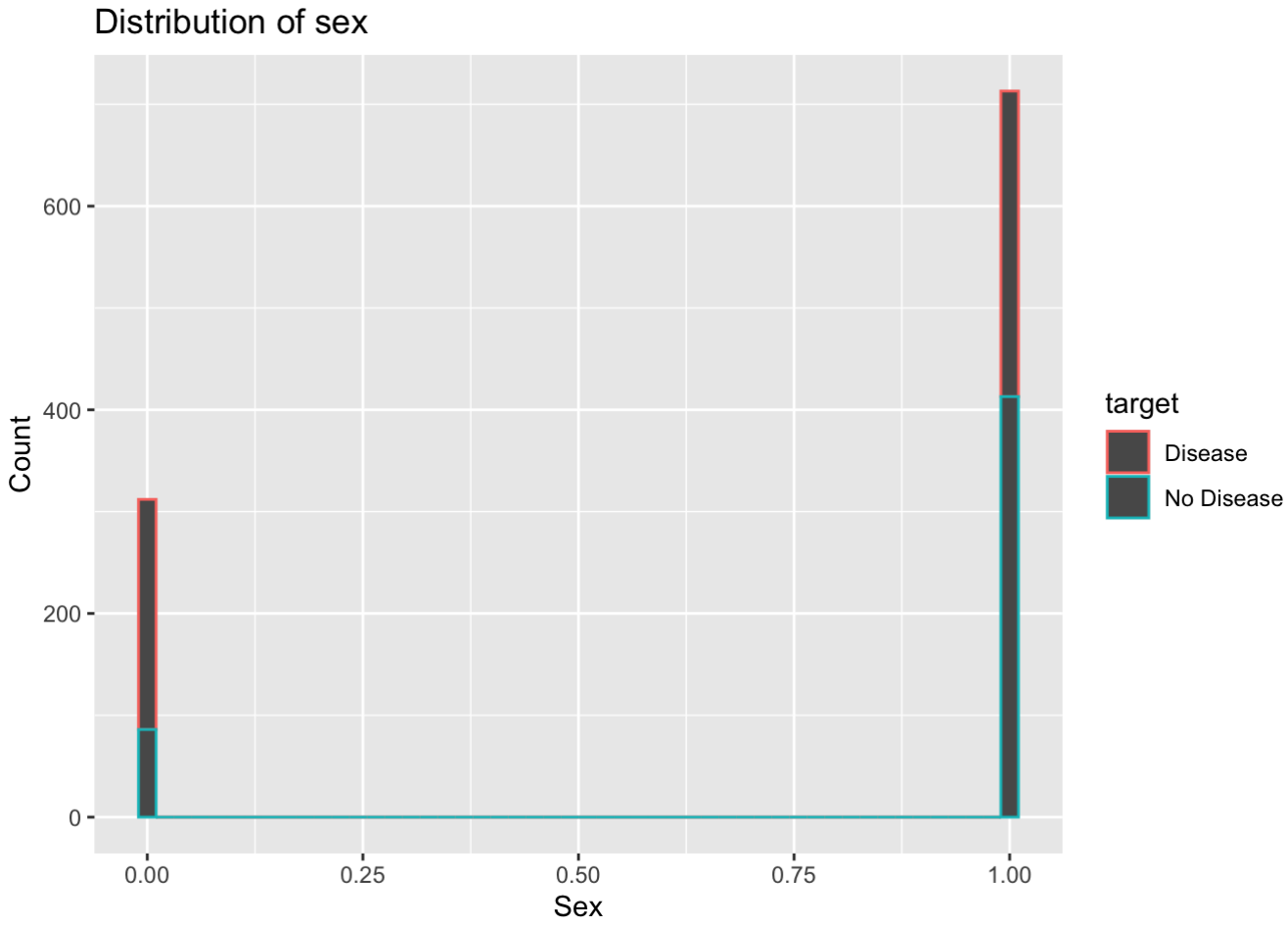
```
# A tibble: 11 × 2
   age_group target
   <fct>      <int>
 1 (25,30]        4
 2 (30,35]       14
 3 (35,40]       27
 4 (40,45]      111
 5 (45,50]       55
 6 (50,55]      117
 7 (55,60]       87
 8 (60,65]       56
 9 (65,70]       38
10 (70,75]       14
11 (75,80]        3
```

▶ Code



From the graph above, the age group 40-45 years and 50-55years are more likely at risk of heart disease.

▶ Code



The dataset revealed that men have a higher count than females and males have a higher percentage of heart disease than females.

▶ Code

```
Female   Male
   0.3    0.7
```

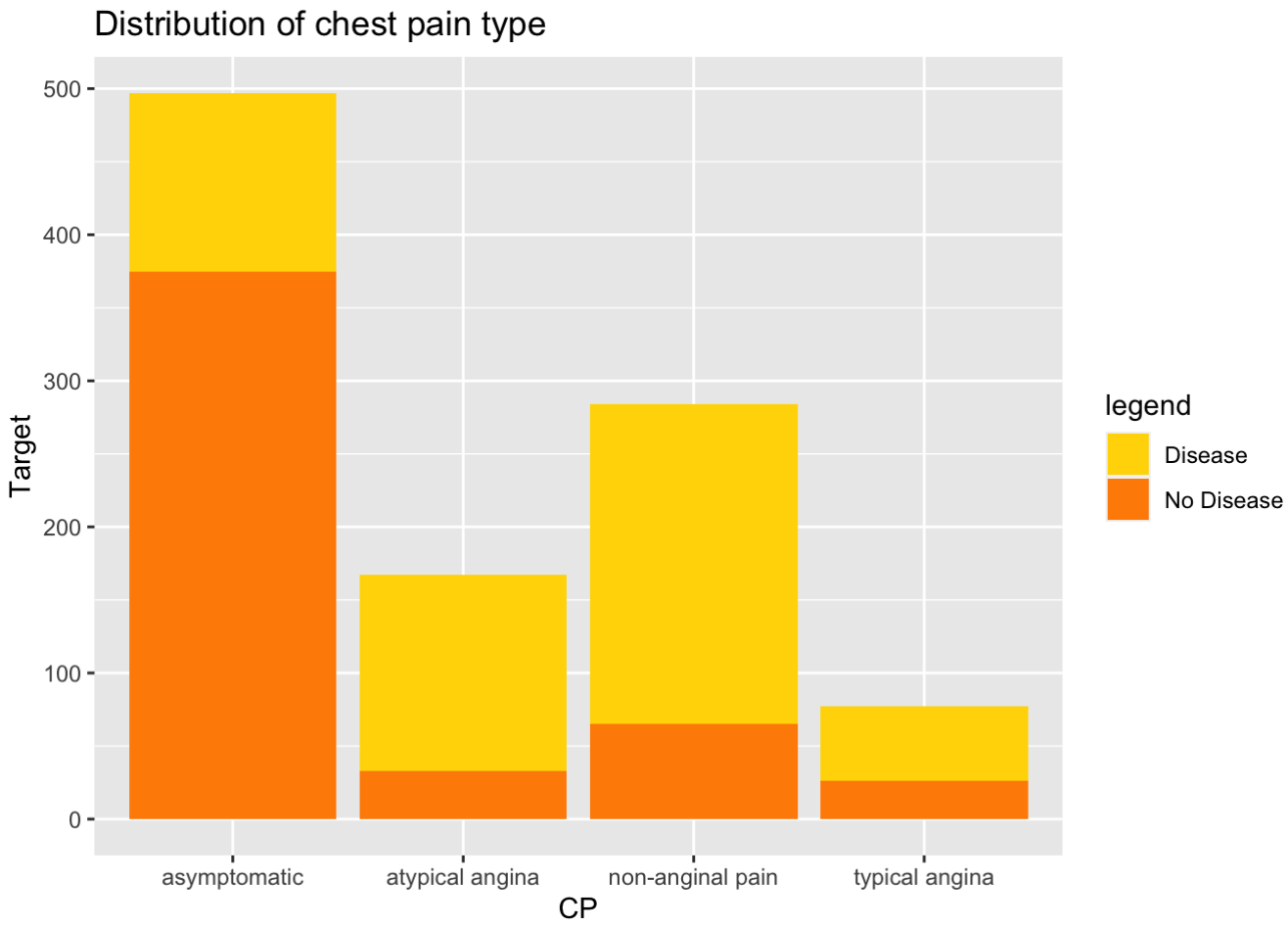
▶ Code

```
        Disease No Disease
Female     0.22       0.08
Male       0.29       0.40
```

From the dataset, there are 70% Males and 30% Females,29% of the Males and 22% of the Females were dignosed with heart disease.From the observation, Males are more likely to get heart disease than Females.

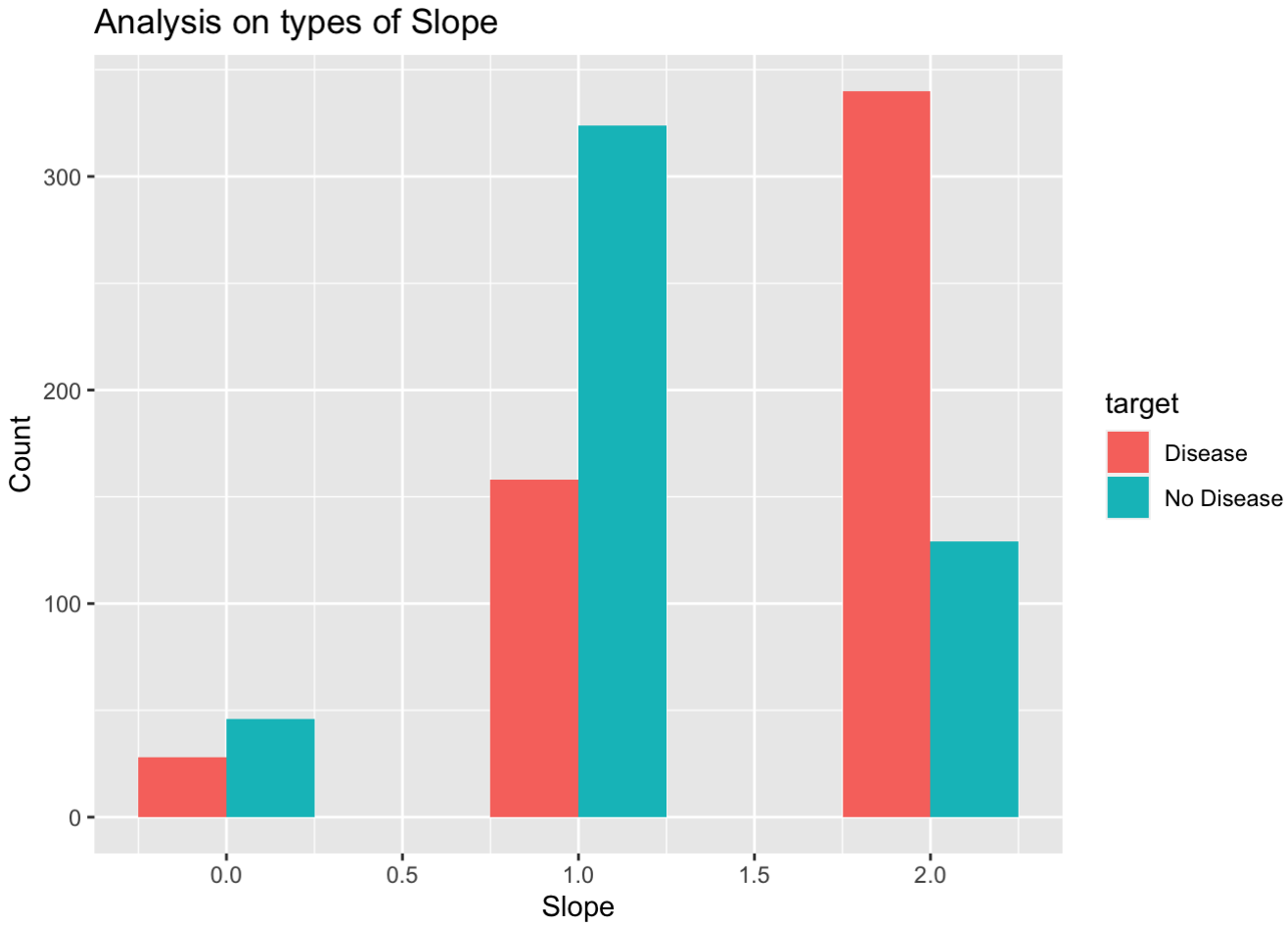
2) What are the other risk factors causing the heart disease?

▶ Code



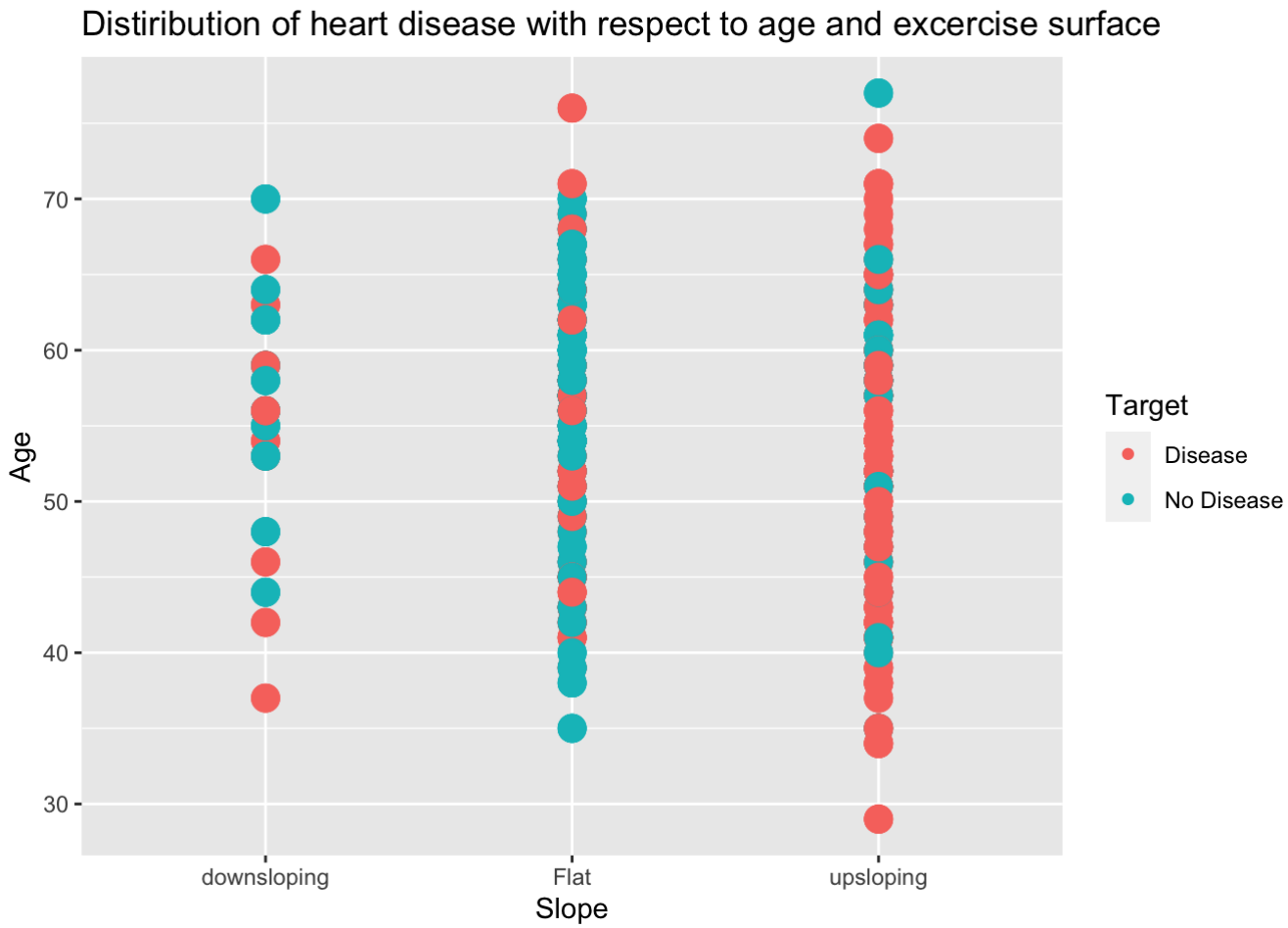
From the above plot, chest pain seems to appear more with heart disease patients whereas, with no heart disease, patients are comparatively less to feel chest pain.

▶ Code



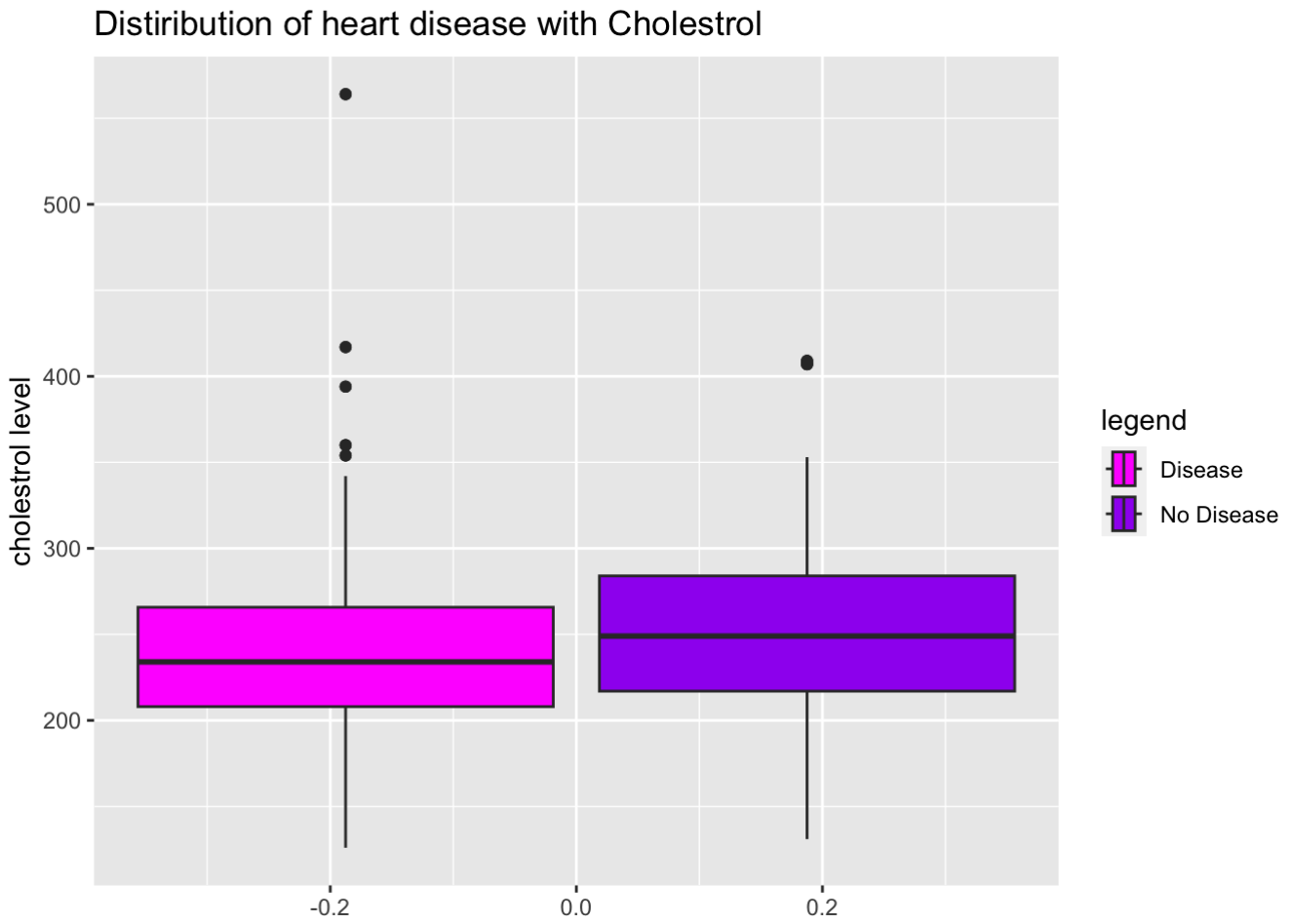
We can observe that those who exercise on an upslope have a higher risk of developing heart disease than those who exercise on a flat or downward slope. Let's look at the age range that participates in the upsloping exercise.

▶ Code



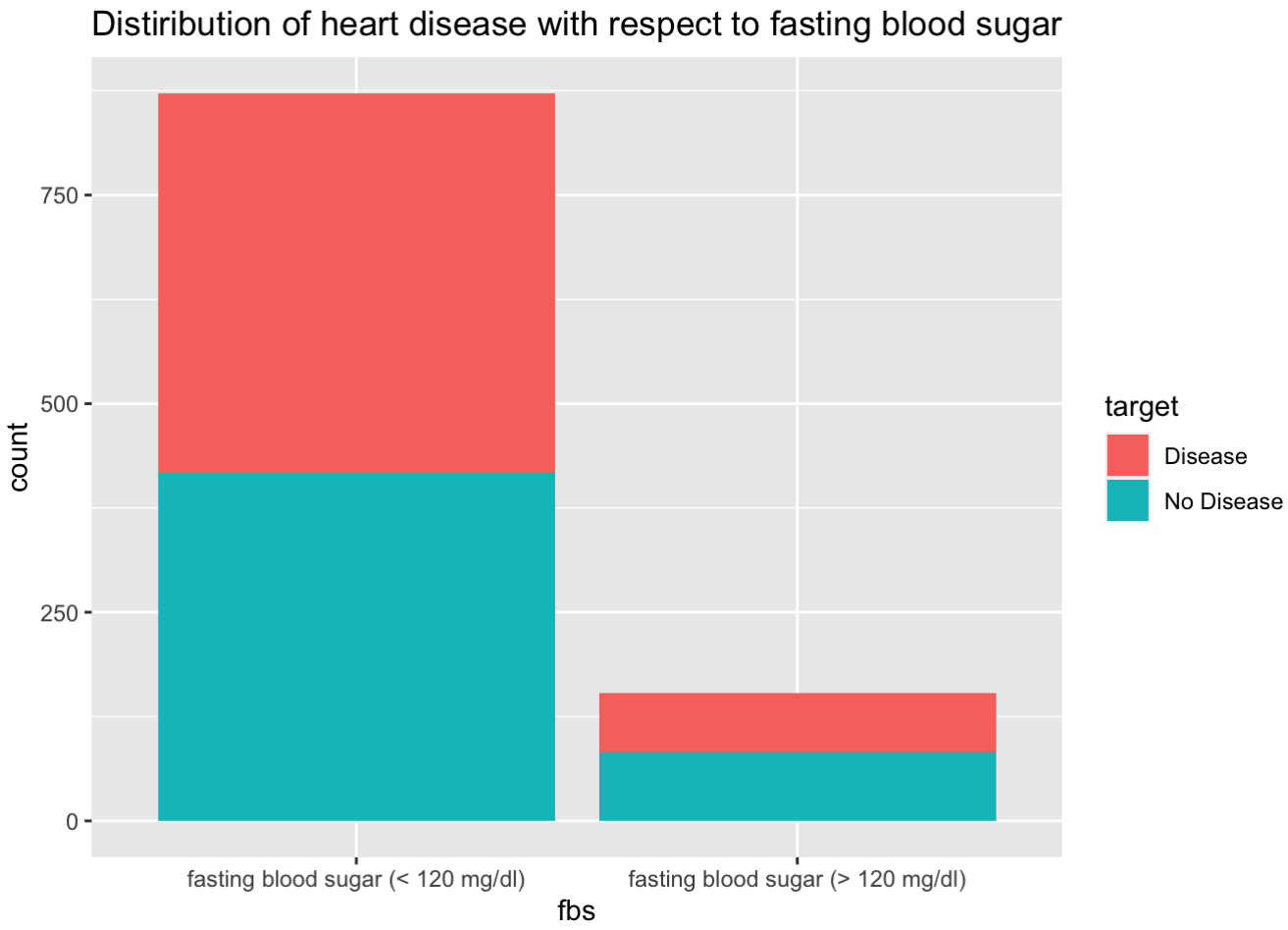
Most people including the aged are found to exercise in flat and upsloping areas.

▶ Code



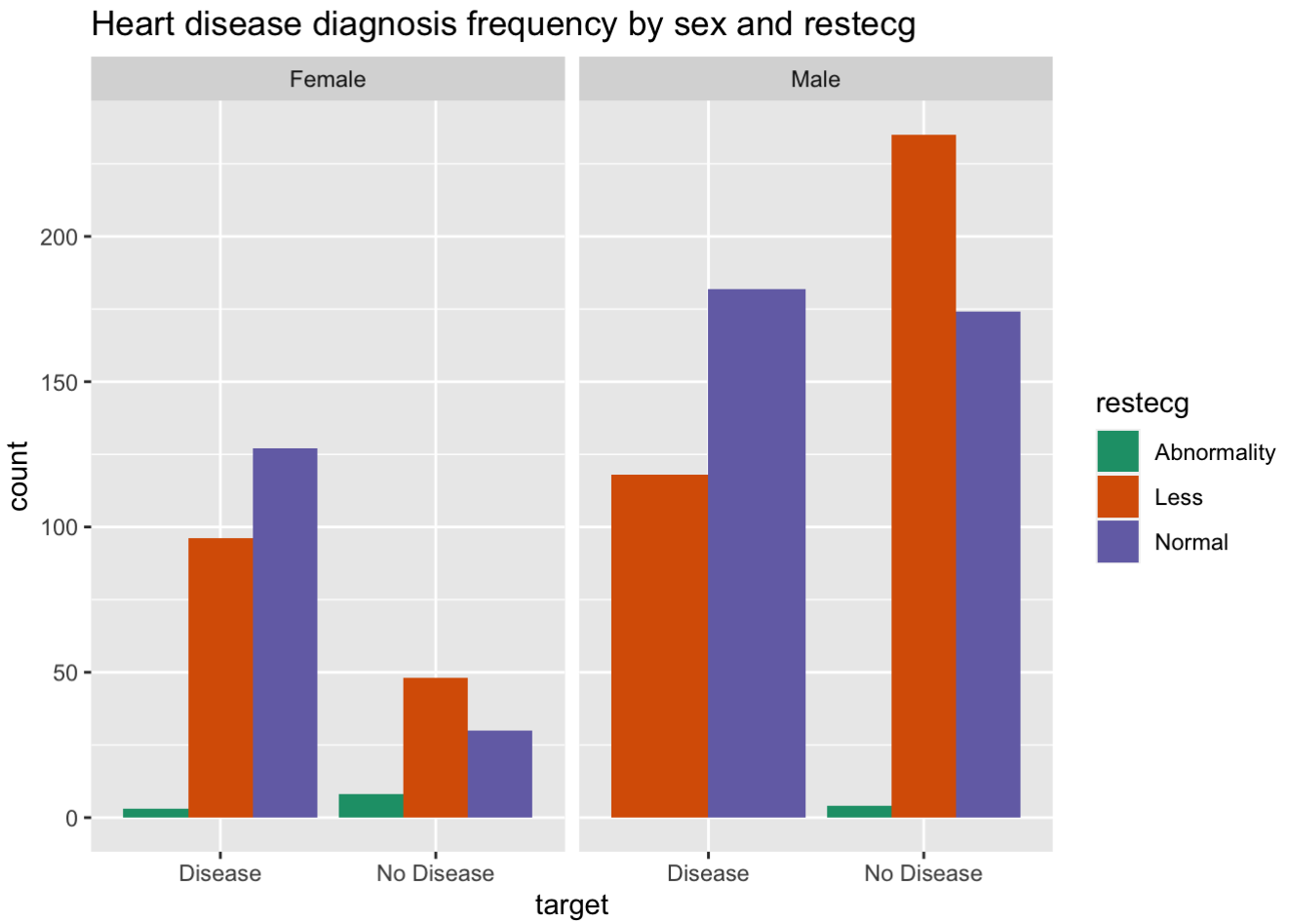
According to the website (https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/lipid-panel), a good cholesterol level for an adult is less than 200mg/dL. According to the plot below, it starts at 200 for both persons with heart disease and those without it. There are some outliers as well. Heart disease is not primarily caused by cholesterol.

▶ Code



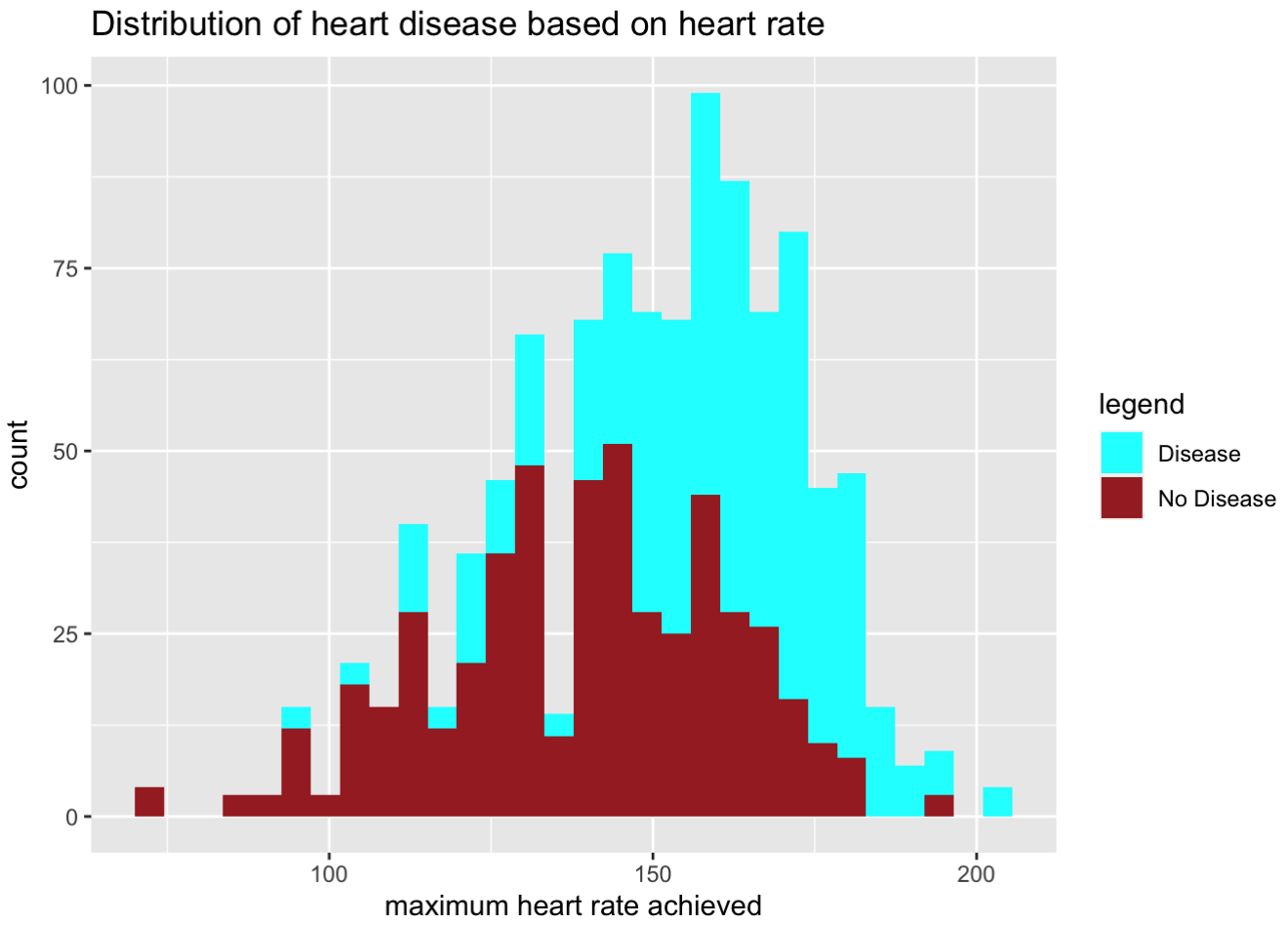
The likelihood of developing heart disease is not affected by blood sugar levels less than or larger than 120 mg/dl. Heart disease affects persons with normal blood sugar levels as well.

▶ Code



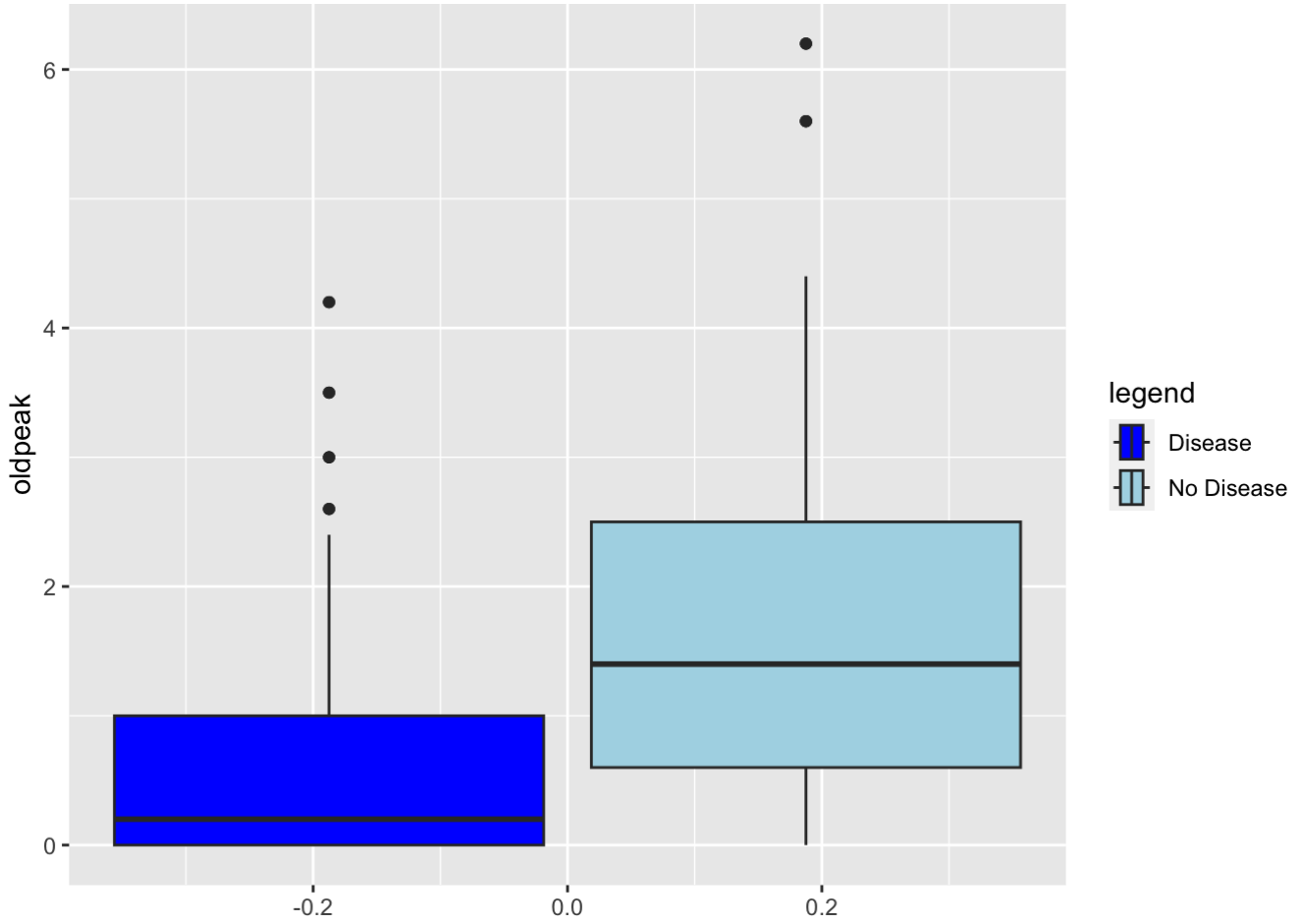
Heart disease is not significantly impacted by abnormalities found in ECG readings. Since Less ECGs are normally distributed and more are normal, we can conclude that ECG frequency is not a risk factor for developing heart disease.

▶ Code



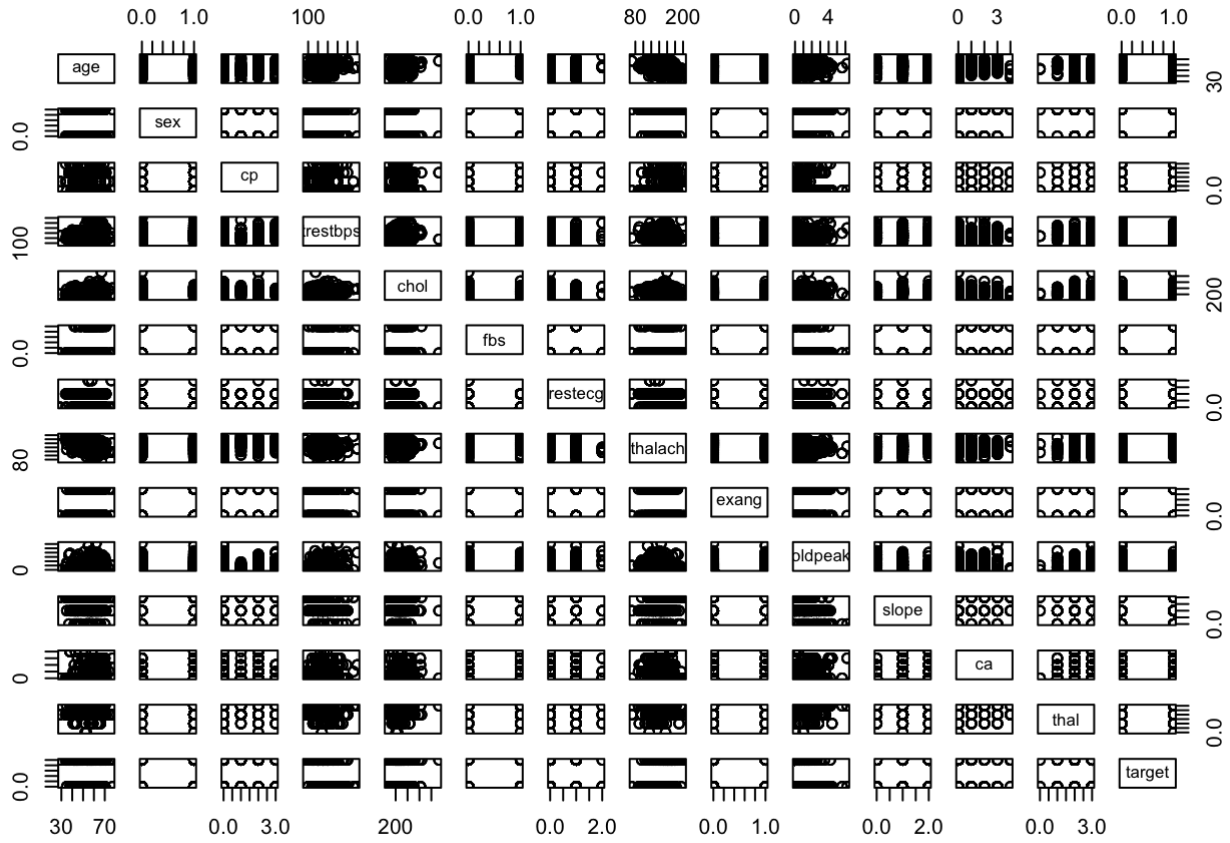
Adults typically have a resting heart rate between 60 and 100 beats per minute. The healthier the person is, the lower their heart rate. As the heart rate rises, more people are prone to get heart disease, as we observe from the plot.

▶ Code



Exercise-induced ST depression as compared to rest (the term “ST” refers to places on the ECG plot). People with heart disease have an old peak value below 1, while those without heart disease have a ST depression value above 2. Also, there are outliers.

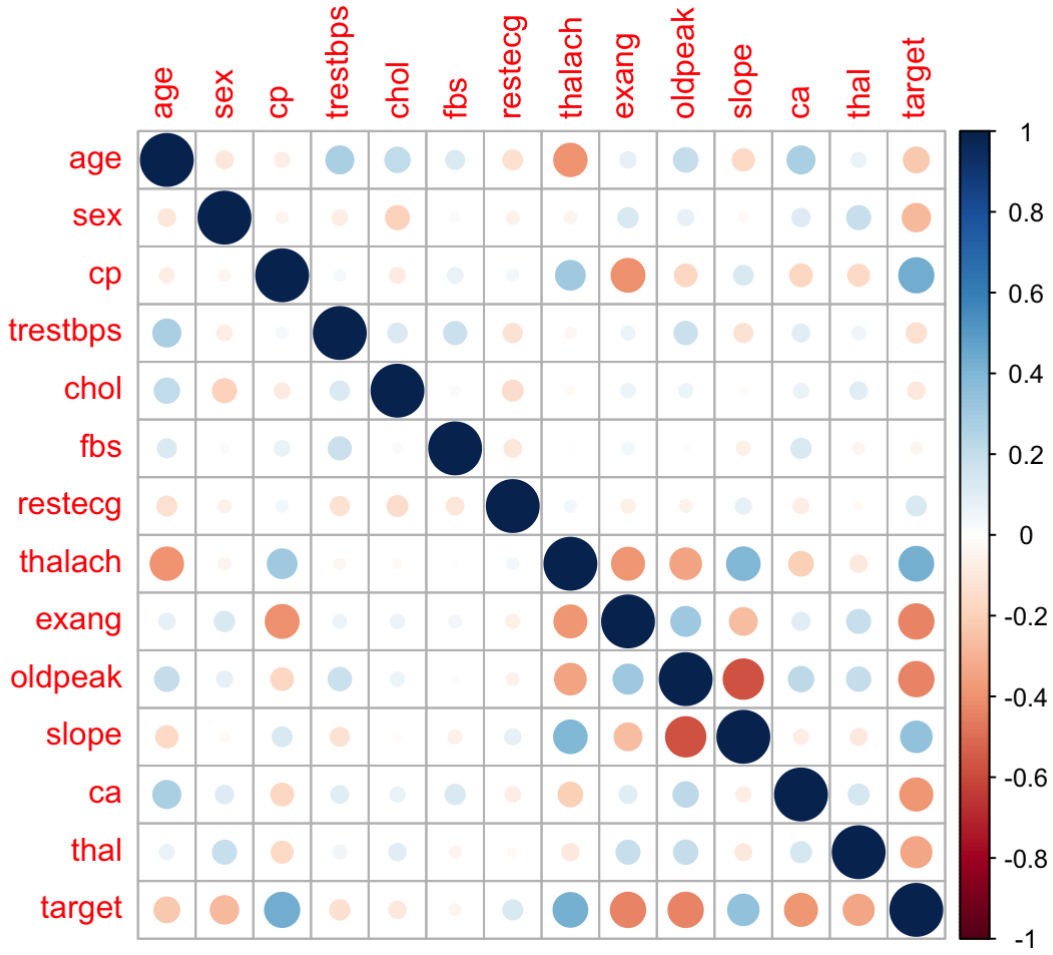
▶ Code



Correlation Plot

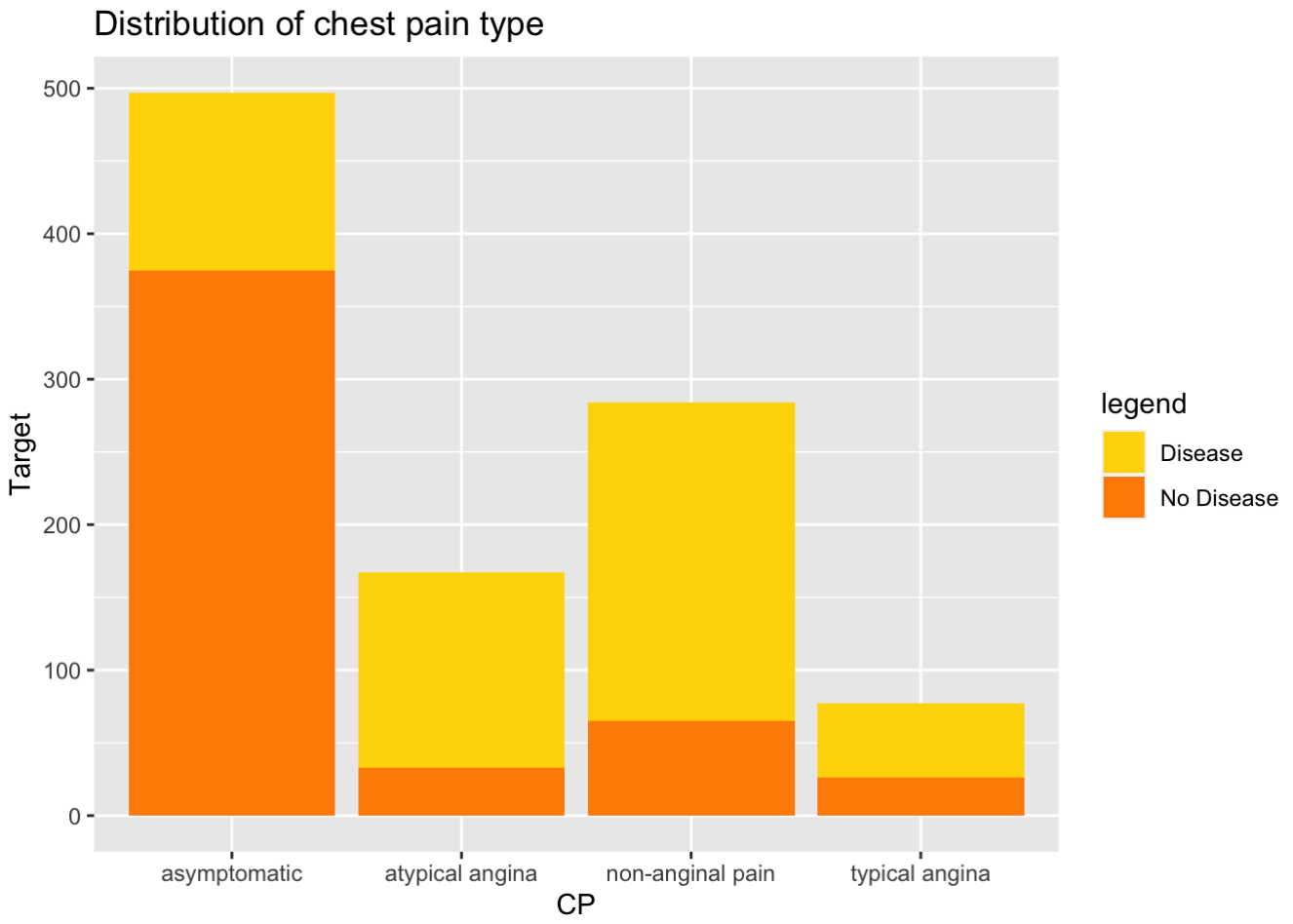
▶ Code

▶ Code

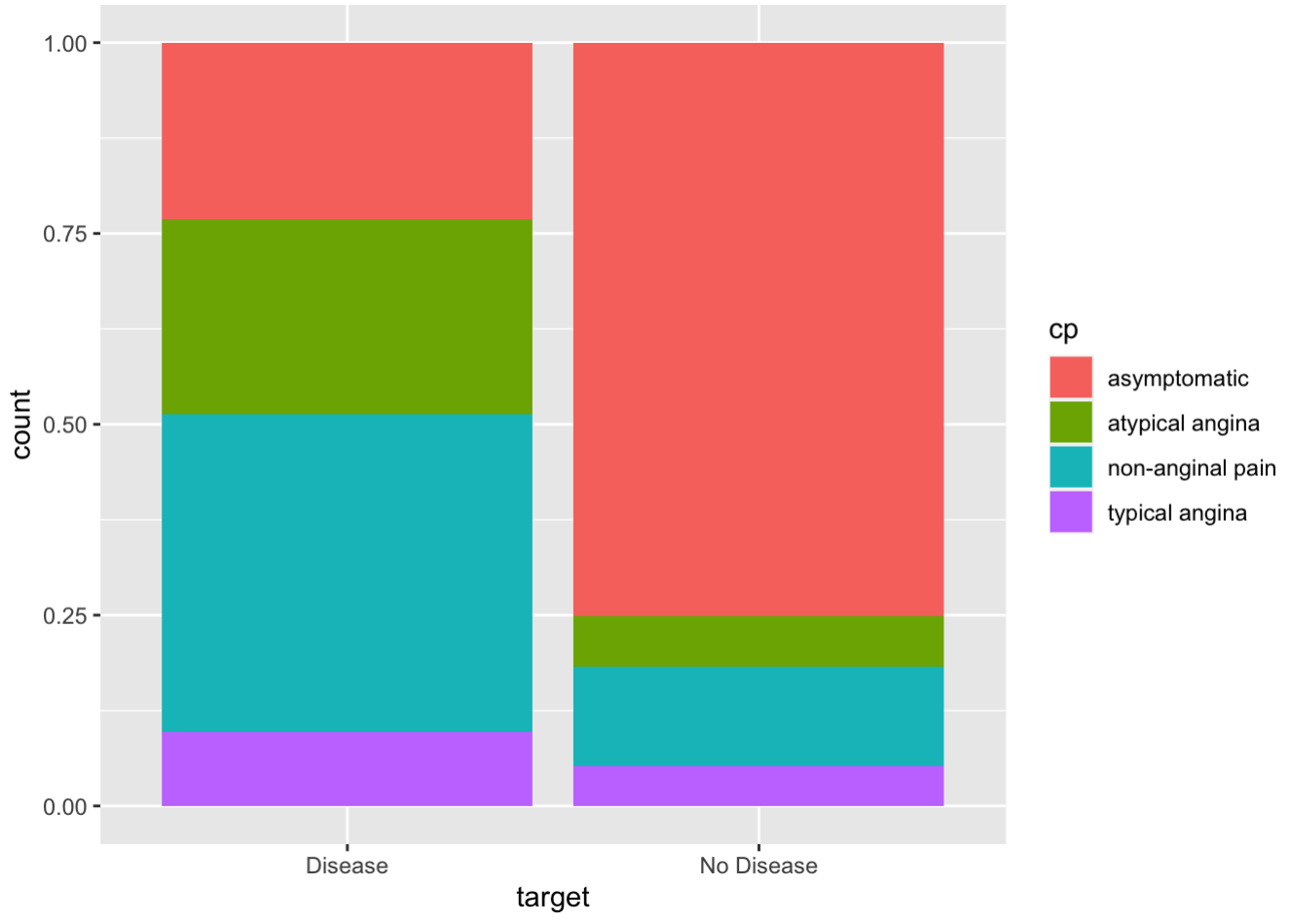


From the confusion matrix, we can observe that the major factors affecting heart disease irrespective of age and gender are chest pain, thalach, and slope. exang(Exercise-induced angina), old peak (ST depression induced by exercise relative to rest) and ca(The number of major vessels), Thal (Blood flow), sex, and age are insignificant in causing heart disease.

▶ Code


Distribution of chest pain type
Target
500
400
300
200
100
0
asymptomatic
atypical angina
non-anginal pain
typical angina
CP
legend
Disease
No Disease

▶ Code


count
1.00
0.75
0.50
0.25
0.00
Disease
No Disease
target
cp
asymptomatic
atypical angina
non-anginal pain
typical angina

3. Are patients experiencing chest pain likely to have heart disease?

Yes, patients with heart disease are more likely to experience chest pain.However, it cannot be a sign of heart disease because chest pain is common in healthy persons as well.

4. IS chest pain an indication of heart disease and What kind of chest pain do most heart patients experience?

We cannot firmly agree that experiencing chest pain necessarily implies developing heart disease. Although there are other possible causes of chest pain, those who have heart disease are more prone to experience it. We can deduce from the plot that the majority of patients with heart disease experience non-anginal pain.

Do a person's age and gender affect their risk of heart disease? (Prediction)

Logistic Regression

▶ Code

▶ Code

```
Call:
glm(formula = target ~ ., family = "binomial", data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4963  -0.3992   0.1160   0.5816   2.7044

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.149946   1.418938   2.925  0.00345 **
age         -0.009333   0.012659  -0.737  0.46097
sex         -1.830358   0.256692  -7.131 1.00e-12 ***
cp           0.834424   0.100624   8.293  < 2e-16 ***
trestbps    -0.016628   0.005672  -2.932  0.00337 **
chol        -0.007967   0.002384  -3.342  0.00083 ***
fbs         -0.125336   0.287788  -0.436  0.66319
restecg      0.414037   0.188750   2.194  0.02827 *
thalach      0.023558   0.005674   4.152 3.30e-05 ***
exang       -0.972007   0.224869  -4.323 1.54e-05 ***
oldpeak     -0.581932   0.116653  -4.989 6.08e-07 ***
```

```
slope        0.529522   0.188912   2.803  0.00506 **
ca          -0.755847   0.104149  -7.257 3.95e-13 ***
thal        -0.900804   0.156419  -5.759 8.47e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1402.16  on 1011  degrees of freedom
Residual deviance:  713.02  on  998  degrees of freedom
AIC: 741.02

Number of Fisher Scoring iterations: 6
```

From the summary of the logistic regression model, there is a strong association between chest pain and heart disease diagnosis having a p-value less than 2e-16. There is also less significant variable like FBS, restecg, exang, and slope. Let's remove these variables as it causes overfitting of the model.

Age does not increase the risk of heart disease, according to the p-value of age (0.46) compared to the alpha value (0.05), and the presence of heart disease is indicated by the p-value of sex (1.0e-12), which is less than the alpha value (0.05).

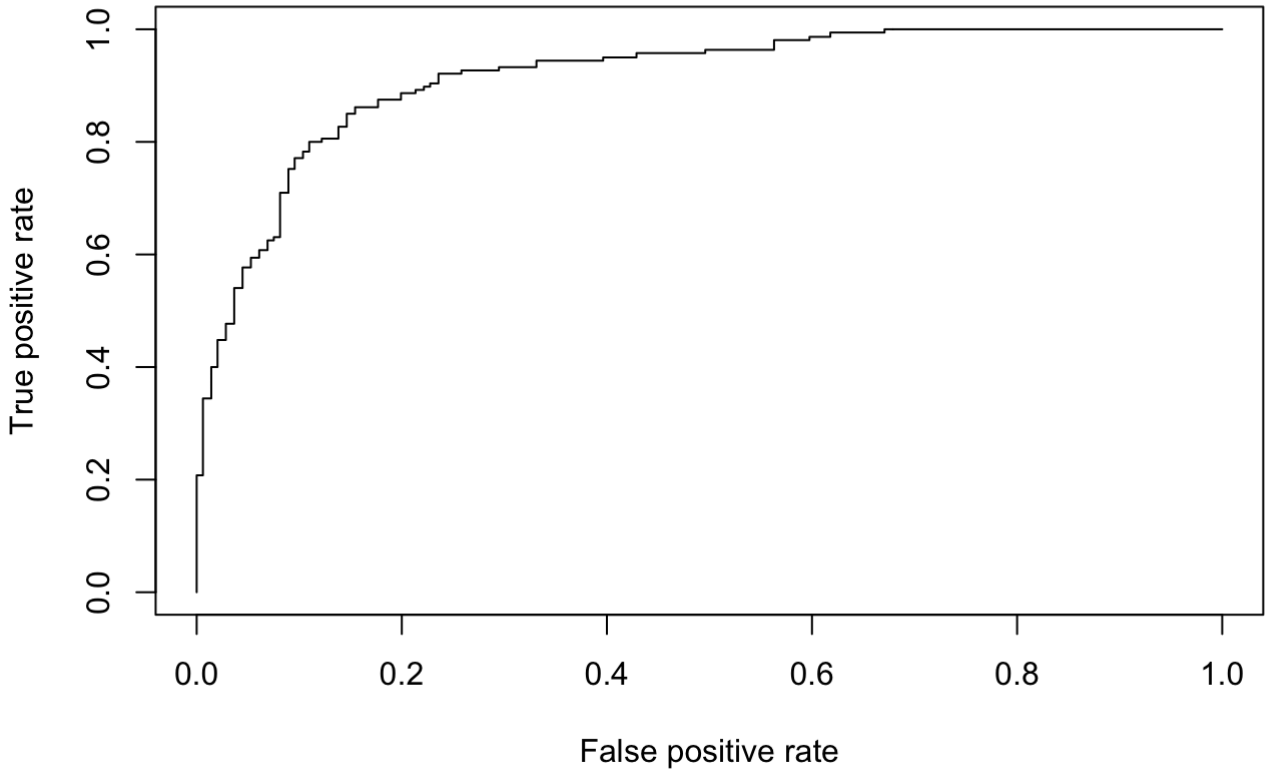
▶ Code

```
Call:
glm(formula = target ~ sex + cp + trestbps + chol + thalach +
    oldpeak + ca + thal, family = "binomial", data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4868  -0.4718   0.1222   0.6062   2.5242

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.447365   1.094560   3.150 0.001635 **
sex         -1.785589   0.238795  -7.477 7.58e-14 ***
cp           0.900756   0.093941   9.589  < 2e-16 ***
trestbps    -0.019727   0.005334  -3.699 0.000217 ***
chol        -0.008822   0.002195  -4.020 5.82e-05 ***
thalach      0.033123   0.004983   6.647 3.00e-11 ***
oldpeak     -0.788245   0.100882  -7.814 5.56e-15 ***
ca          -0.697411   0.097248  -7.171 7.42e-13 ***
thal        -0.860950   0.151259  -5.692 1.26e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1402.16  on 1011  degrees of freedom
Residual deviance:  748.15  on 1003  degrees of freedom
```

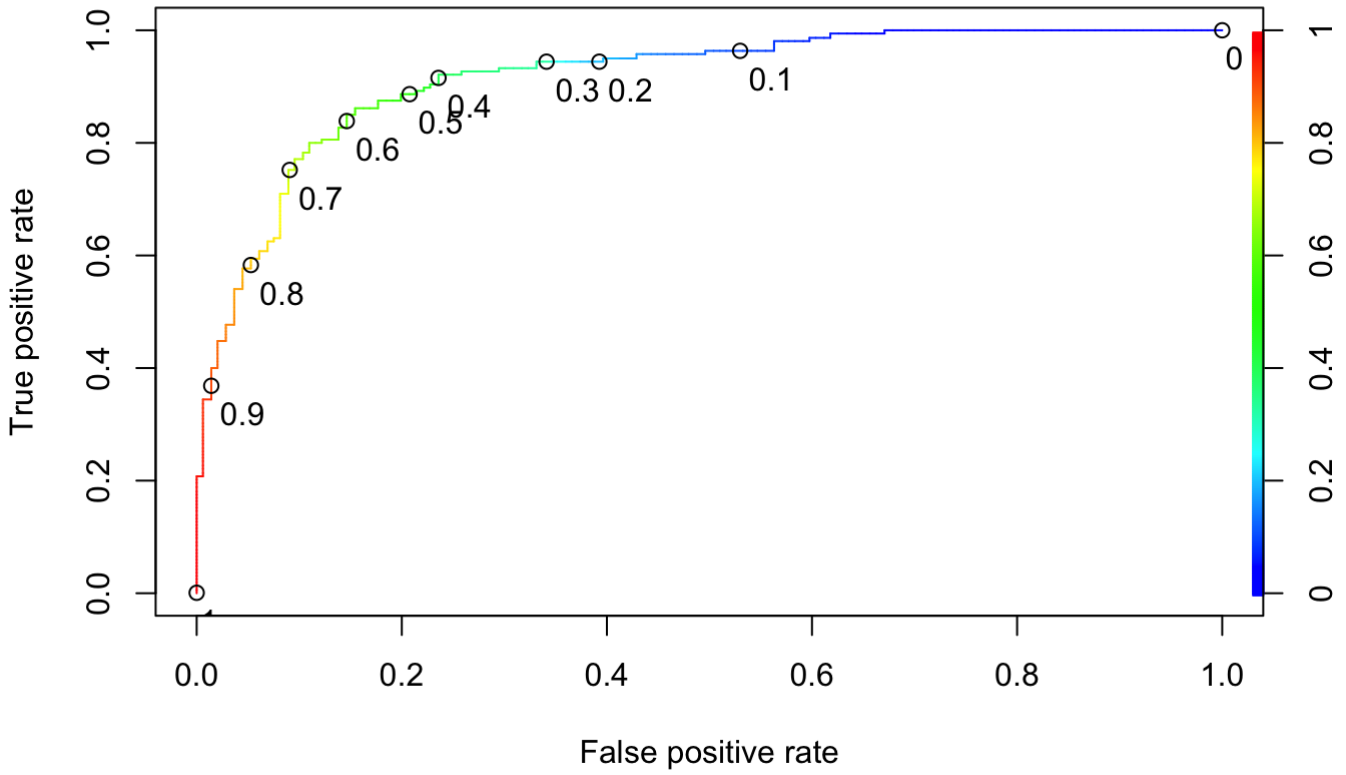
```
AIC: 766.15

Number of Fisher Scoring iterations: 6
```

Plotting ROCR Curve

▶ Code



▶ Code



From ROCR curve threshold of 0.7 seems to be okay so that true positives are maximised such that maximum number of patients with heart disease are not identified as healthy.

Area under curve

▶ Code

```
[1] 0.9154862
```

We can see the value of AUC so Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.The AUC value is 0.91 which means our model is able to distinguish between patients with the disease and no disease with a probability of 0.91,So it is a good value.

Accuracy

▶ Code

```
   FALSE TRUE
0    448   44
1    129  391
```

Accuracy on training data

▶ Code

```
[1] 0.8290514
```

Logistic regression after removing less significant attributes performed best with an accuracy of testing 82%.

Conclusion

In conclusion, analysis of the heart disease dataset has revealed several important insights. We found that factors such as Chest pain, Sex and ca(number of major vessels) are strongly associated with the presence of heart disease. From our plot, we can infer that non-anginal pain(chest pain) is mostly seen in heart disease patients and Males are more vulnerable to being diagnosed with heart disease than females. Based on the score (82%), the model predicted is good. Factors like smoking, obesity, and a history of heart disease in the family were missing in our dataset which could help in predicting the model. Our visualizations also highlighted the importance of regular exercise on a flat surface and downsloping. Additionally, the sample size is relatively small, which may limit our findings.

In future research, it would be beneficial to conduct a larger study with a more diverse population to further validate our results. Additionally, it would be interesting to investigate the potential role of other factors such as stress and sleep in the development of heart disease. Overall, we recommend the importance of maintaining a healthy lifestyle, regular check-ups, early medical attention to chest pain, and public awareness of heart disease(factors) in preventing heart disease.