HEART ATTACK ANALYSIS

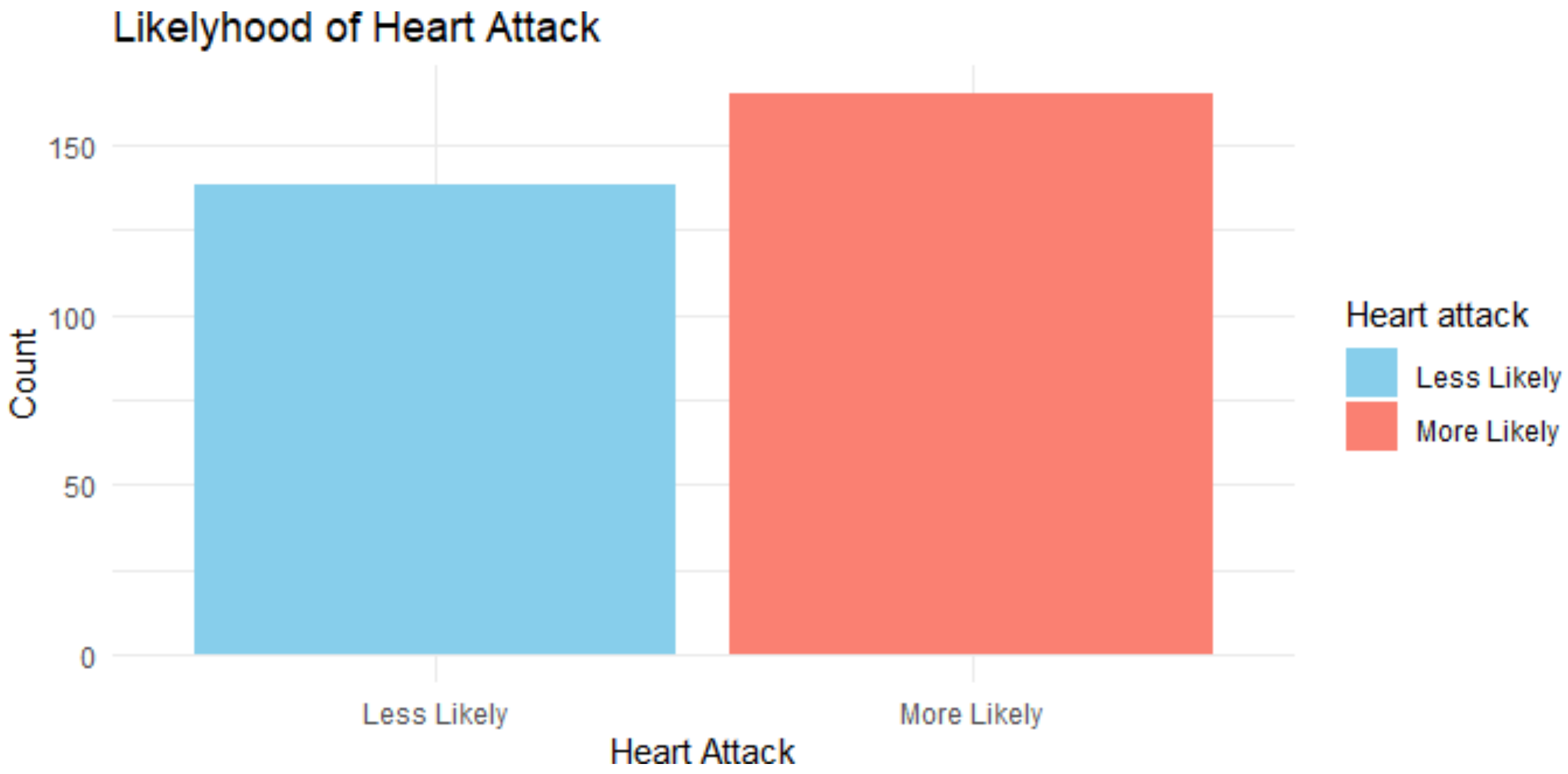
Matteo Montrucchio

INDEX

Exploratory Data Analysis

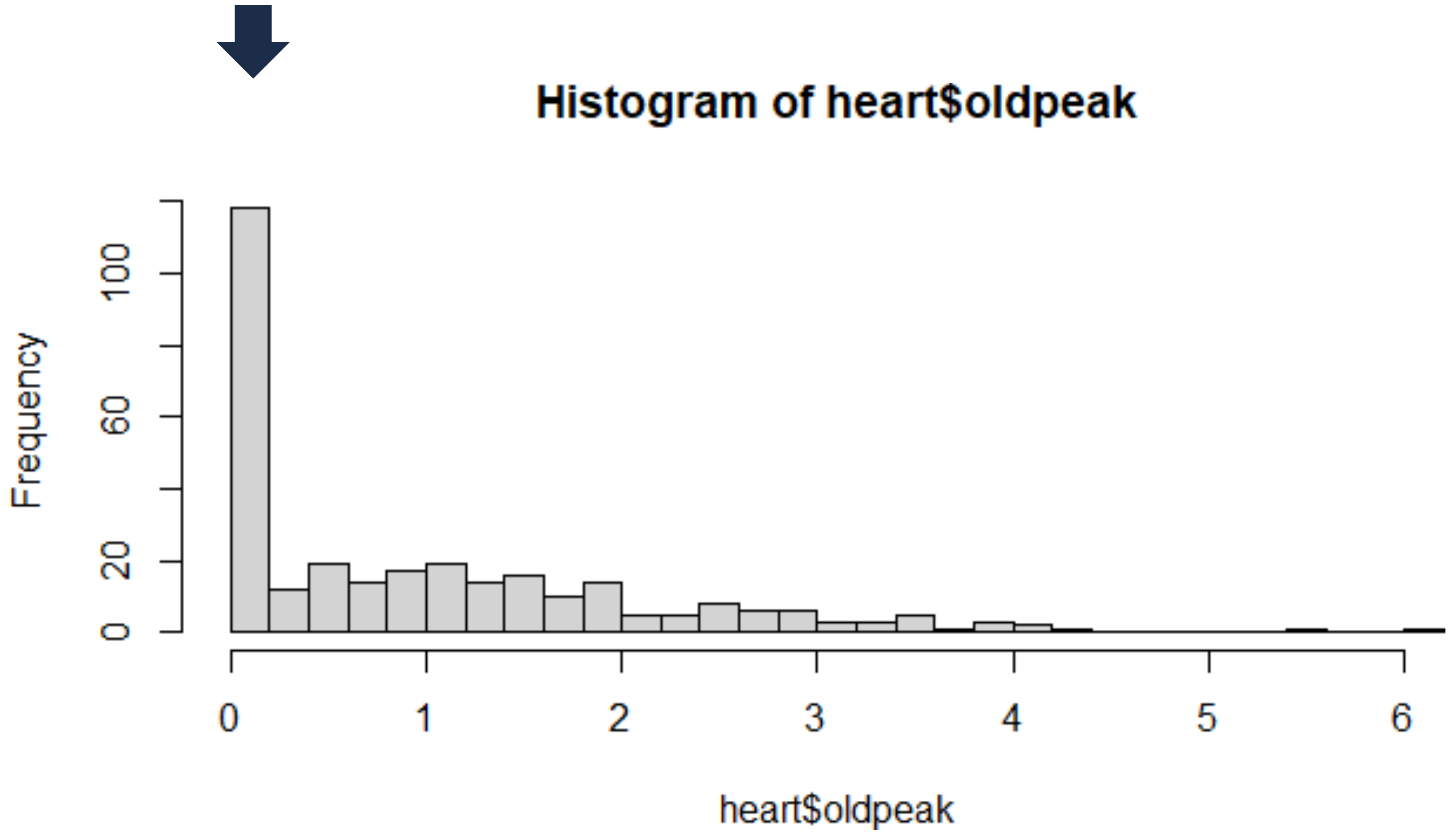
EDA

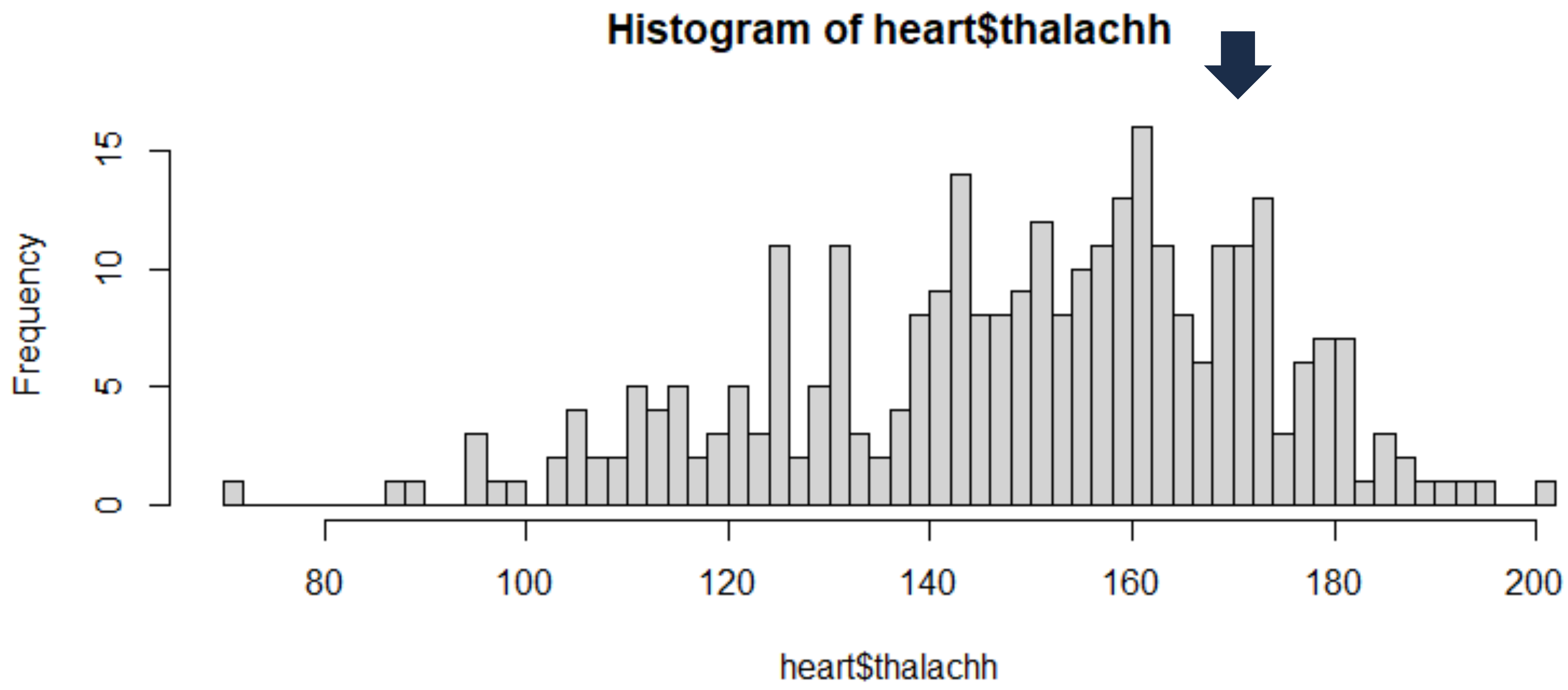
- The dataset is **balanced**
- There are no missing values
- There are no duplicates



Univariate Analysis

- **Histograms** indicate normal distributions for most variables, except for "oldpeak" and "thalachh," which show left and right skewness, respectively

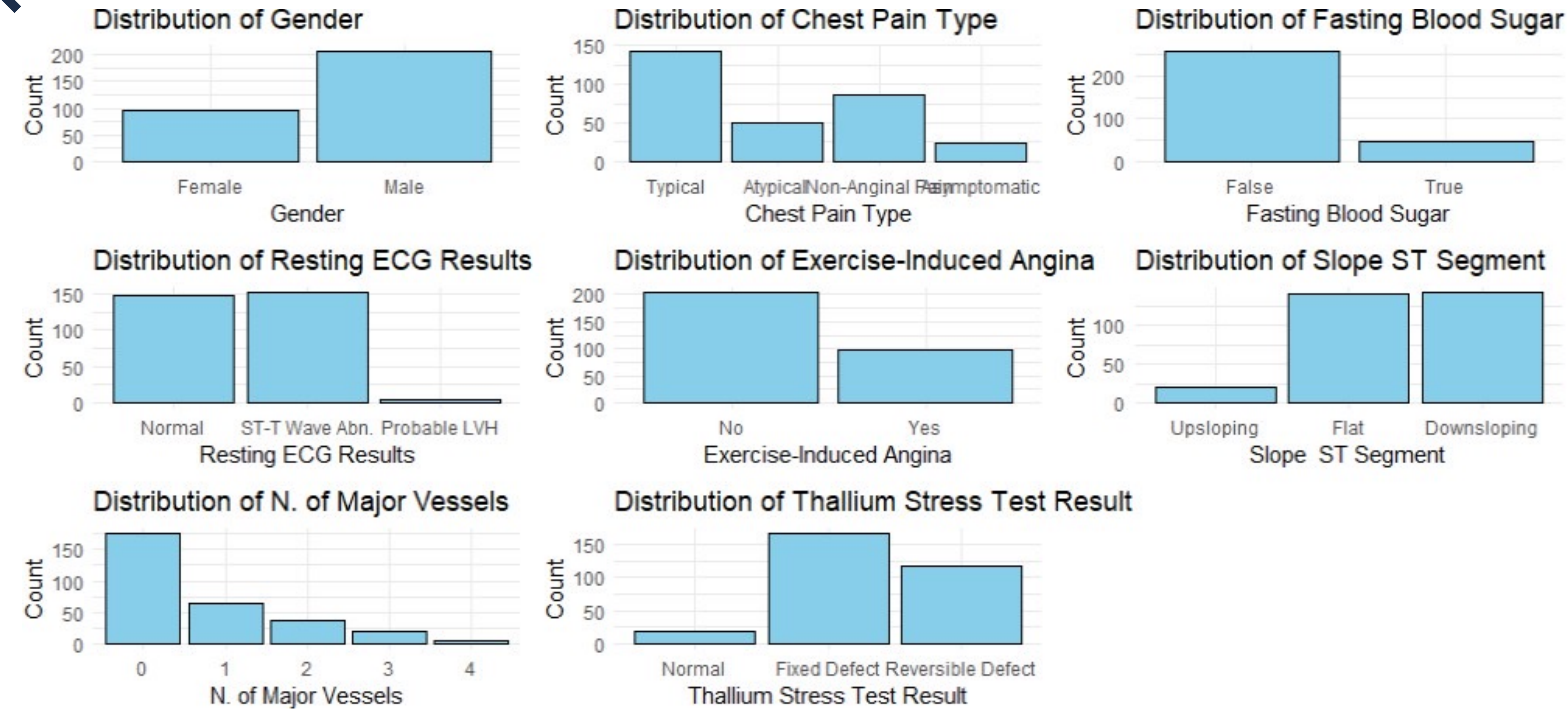




Univariate Analysis

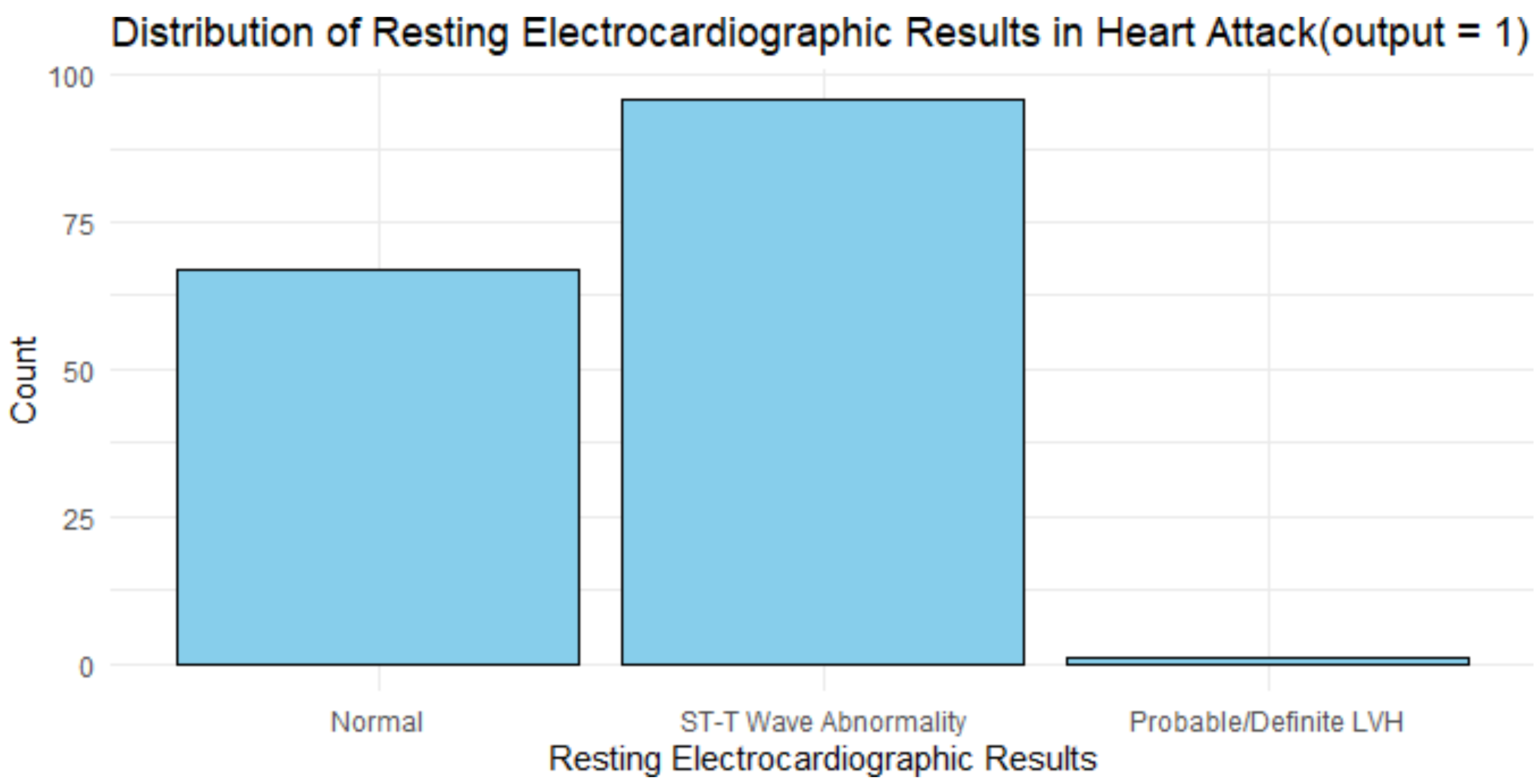
Analyzing **bar charts** of categorical attributes:

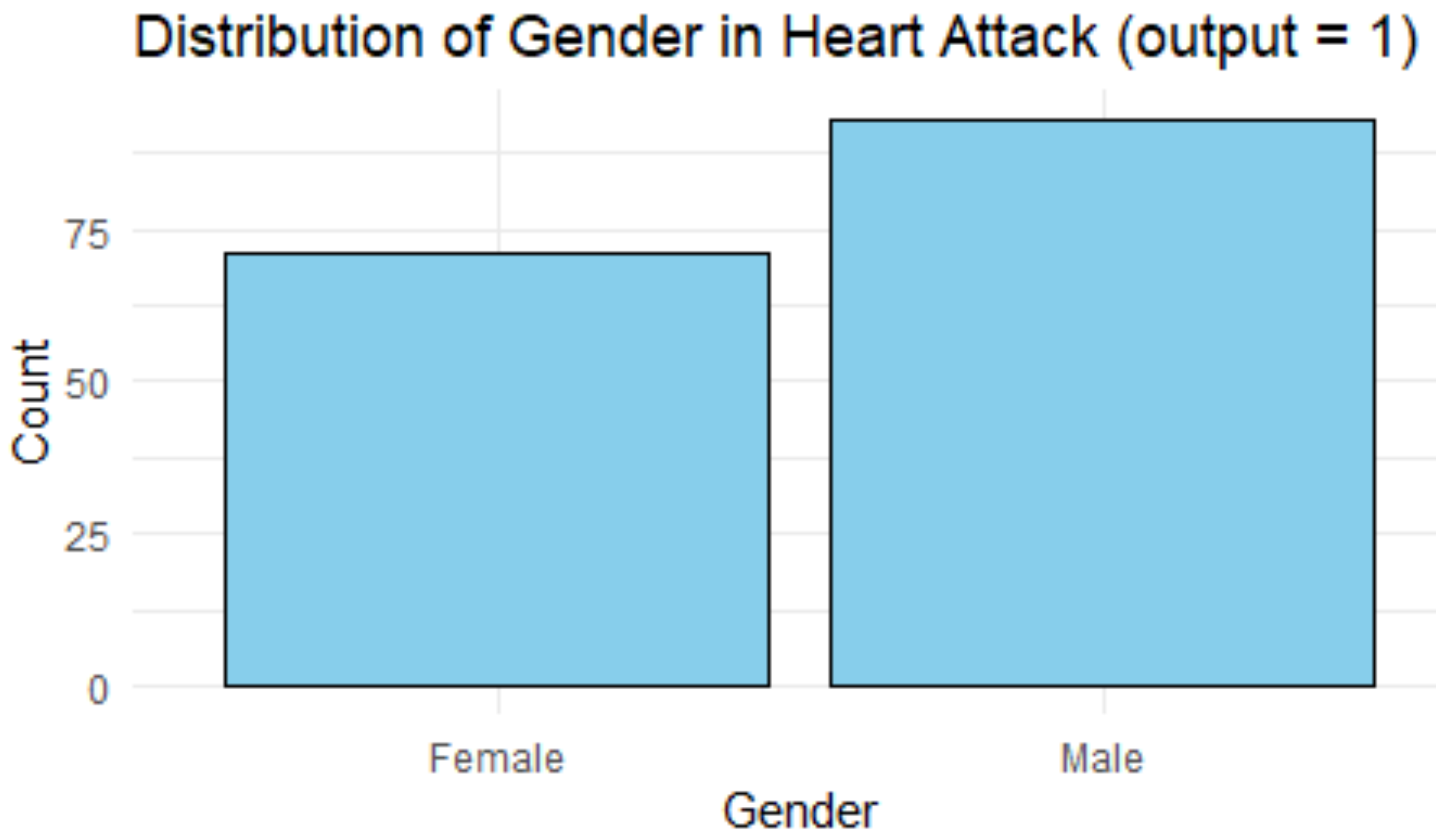
- Gender imbalance with more males
- Resting ecg results show ST-T wave abnormality (type 1) as the most prevalent.

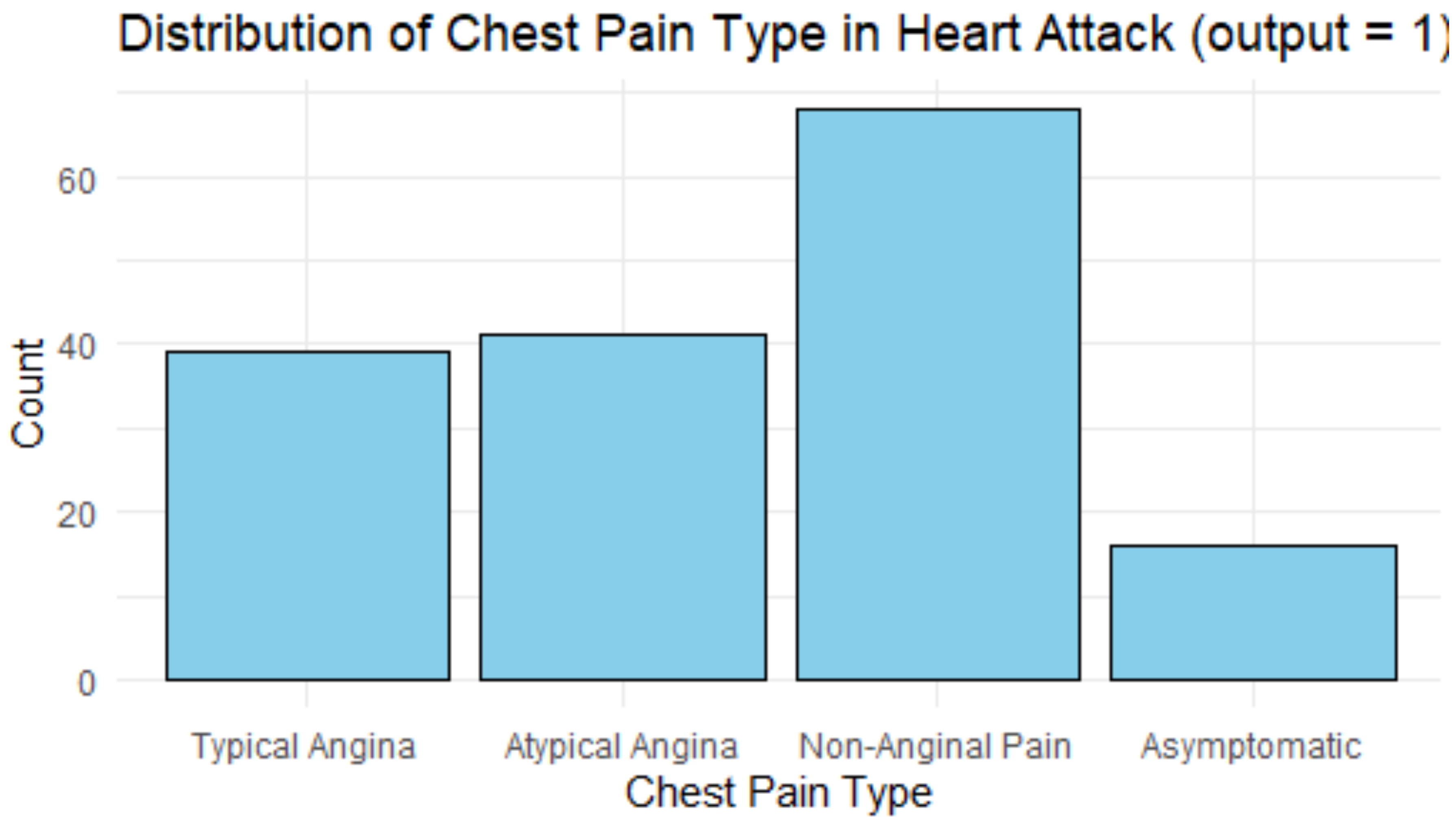

Distribution of Gender
Count
200
150
100
50
0
Female
Male
Gender
Distribution of Chest Pain Type
Count
150
100
50
0
Typical
Atypical
Non-Anginal Pain
Asymptomatic
Chest Pain Type
Distribution of Fasting Blood Sugar
Count
200
100
0
False
True
Fasting Blood Sugar
Distribution of Resting ECG Results
Count
150
100
50
0
Normal
ST-T Wave Abn.
Probable LVH
Resting ECG Results
Distribution of Exercise-Induced Angina
Count
200
150
100
50
0
No
Yes
Exercise-Induced Angina
Distribution of Slope ST Segment
Count
100
50
0
Upsloping
Flat
Downsloping
Slope ST Segment
Distribution of N. of Major Vessels
Count
150
100
50
0
0
1
2
3
4
N. of Major Vessels
Distribution of Thallium Stress Test Result
Count
150
100
50
0
Normal
Fixed Defect
Reversible Defect
Thallium Stress Test Result

Bivariate Analysis

- Heart attack is more prevalent in males
- Individuals with a chest pain type (cp) equal to 3 (non-anginal) are more likely to have a heart attack
- Those with a rest ECG value of 1, indicating non-normal heartbeats, have a higher chance of heart attack


Distribution of Resting Electrocardiographic Results in Heart Attack(output = 1)
100
75
50
25
0
Count
Normal
ST-T Wave Abnormality
Probable/Definite LVH
Resting Electrocardiographic Results

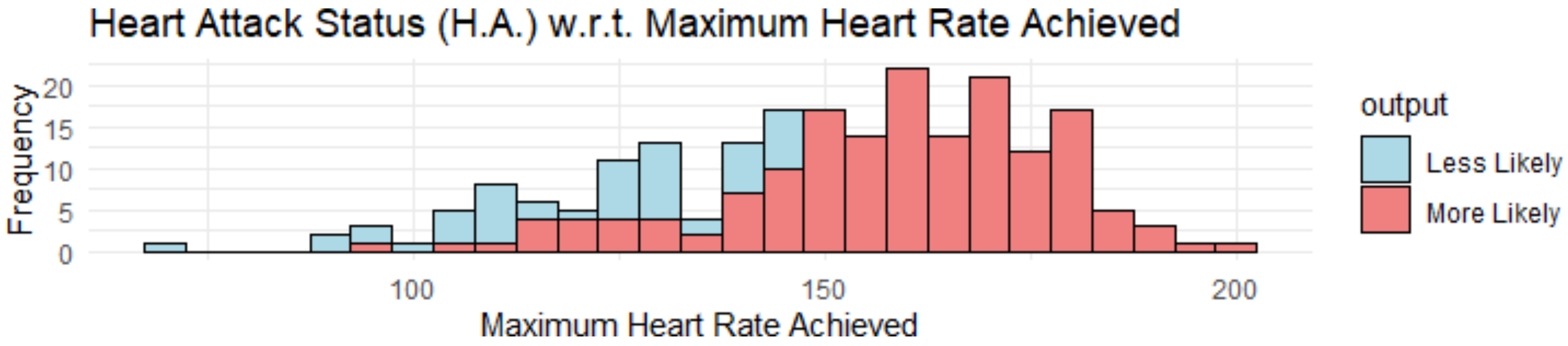

Distribution of Gender in Heart Attack (output = 1)
75
50
25
0
Count
Female
Male
Gender

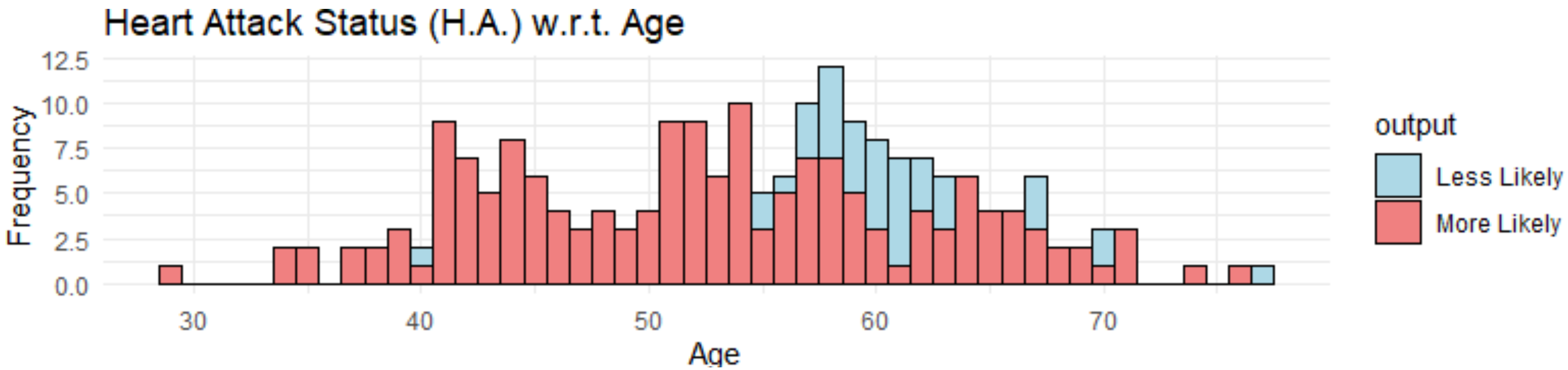

Distribution of Chest Pain Type in Heart Attack (output = 1)
60
40
20
0
Count
Typical Angina
Atypical Angina
Non-Anginal Pain
Asymptomatic
Chest Pain Type

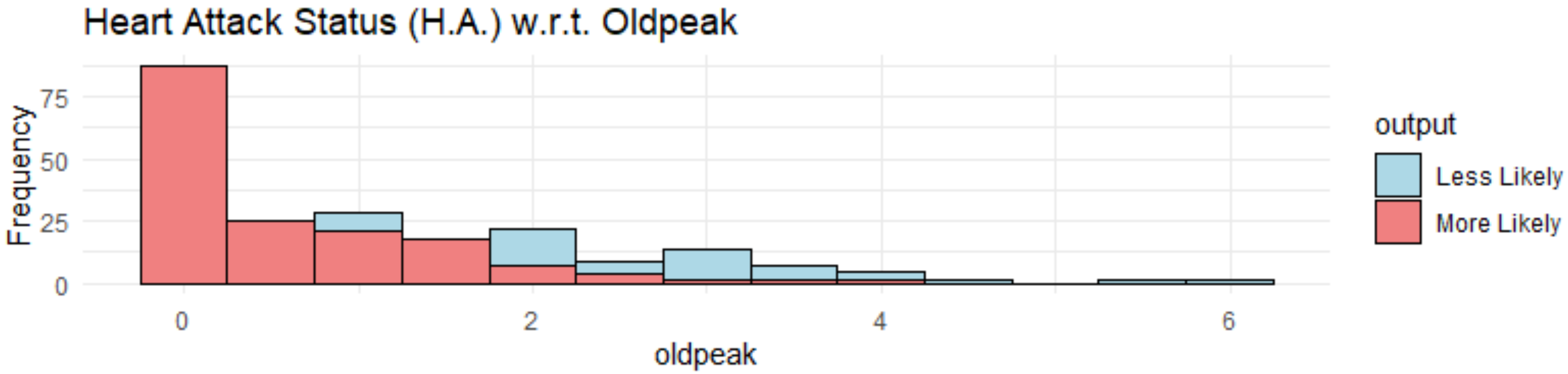
Bivariate Analysis

- Individuals in the middle age range (40 to 60 years) exhibit a higher likelihood of experiencing a heart attack.

- Previous peak (oldpeak) exhibits a negative correlation with the chances of experiencing a heart attack.


Heart Attack Status (H.A.) w.r.t. Maximum Heart Rate Achieved
Frequency
0
5
10
15
20
100
150
200
Maximum Heart Rate Achieved
output
Less Likely
More Likely


Heart Attack Status (H.A.) w.r.t. Age
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
30
40
50
60
70
Age
output
Less Likely
More Likely


Heart Attack Status (H.A.) w.r.t. Oldpeak
Frequency
0
25
50
75
0
2
4
6
oldpeak
output
Less Likely
More Likely

Correlation

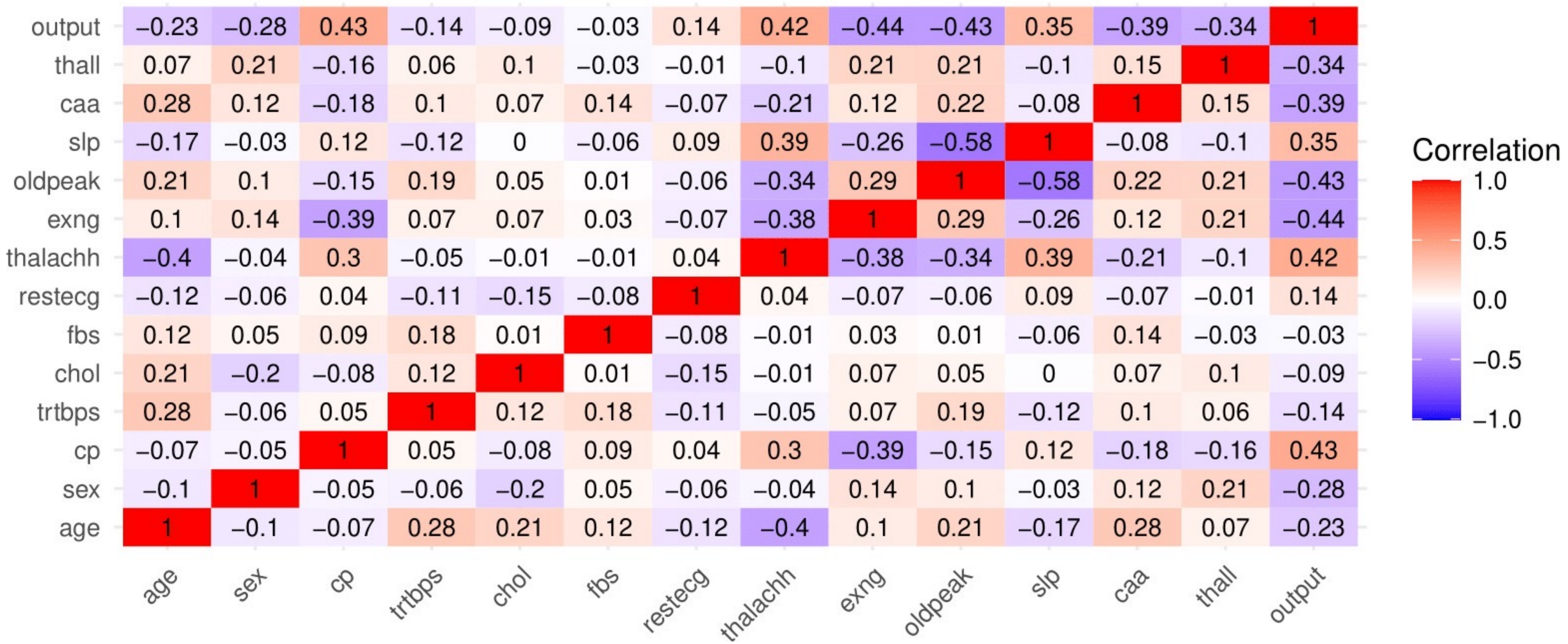
Correlation

- **Positive Correlations with 'output' (Heart Attack)**:
- cp (Chest Pain Type): Strong positive correlation. Severity increase →Higher heart attack likelihood.

- **Negative Correlations with 'output'**:
- exng (Exercise-Induced Angina): Negative correlation. Absence →Higher heart attack likelihood.

- **Other Correlations**:
- slp with oldpeak : Negative correlations. Certain ST segment patterns →Lower ST depression

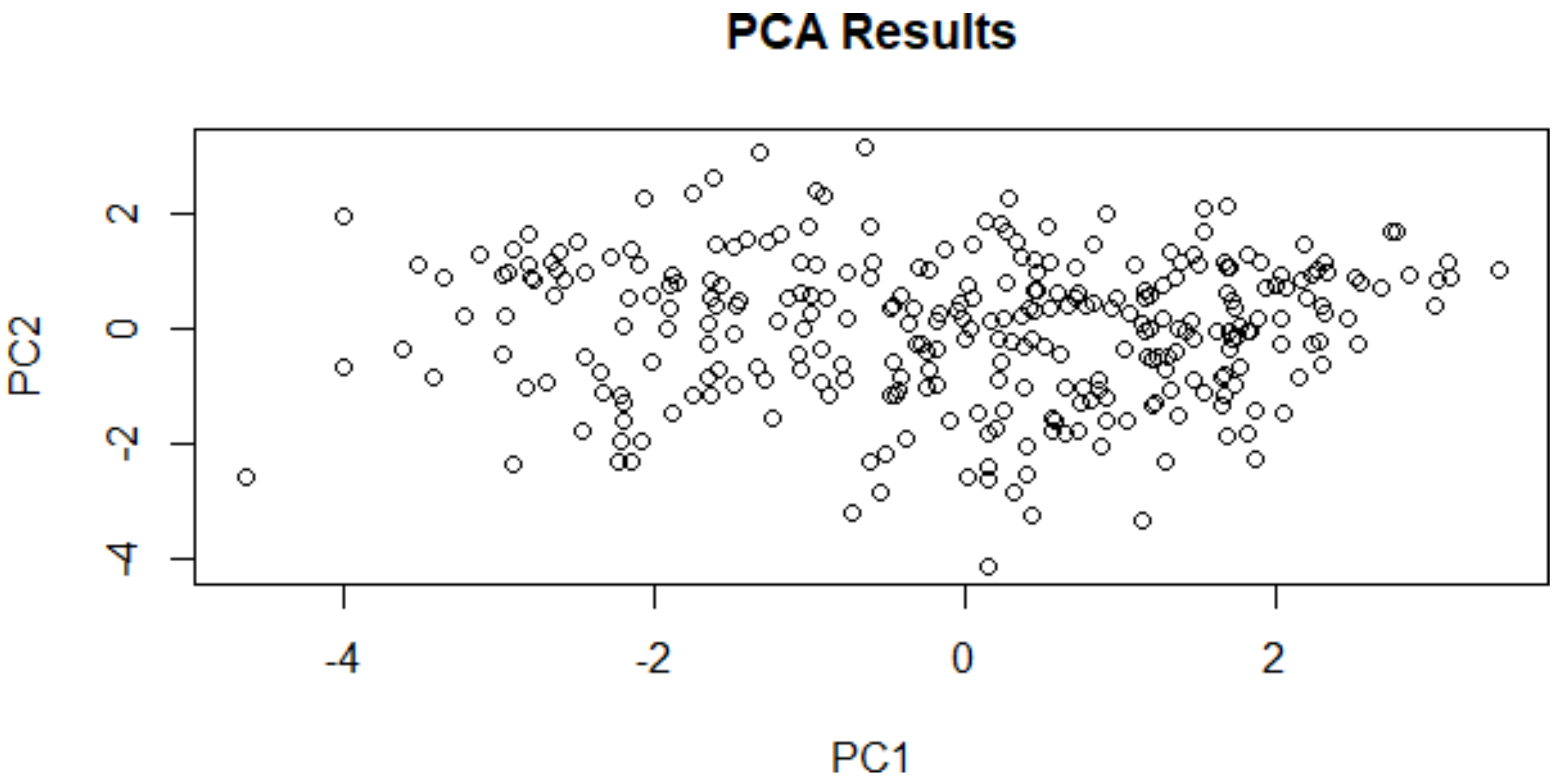
Correlation Matrix



Principal Components Analisys

PCA

- The first 8 components (PC), capture 79.50% of the total variance in the original dataset.
- PC1: represents the most influential pattern in the data contributing to 21.25% of the total variance

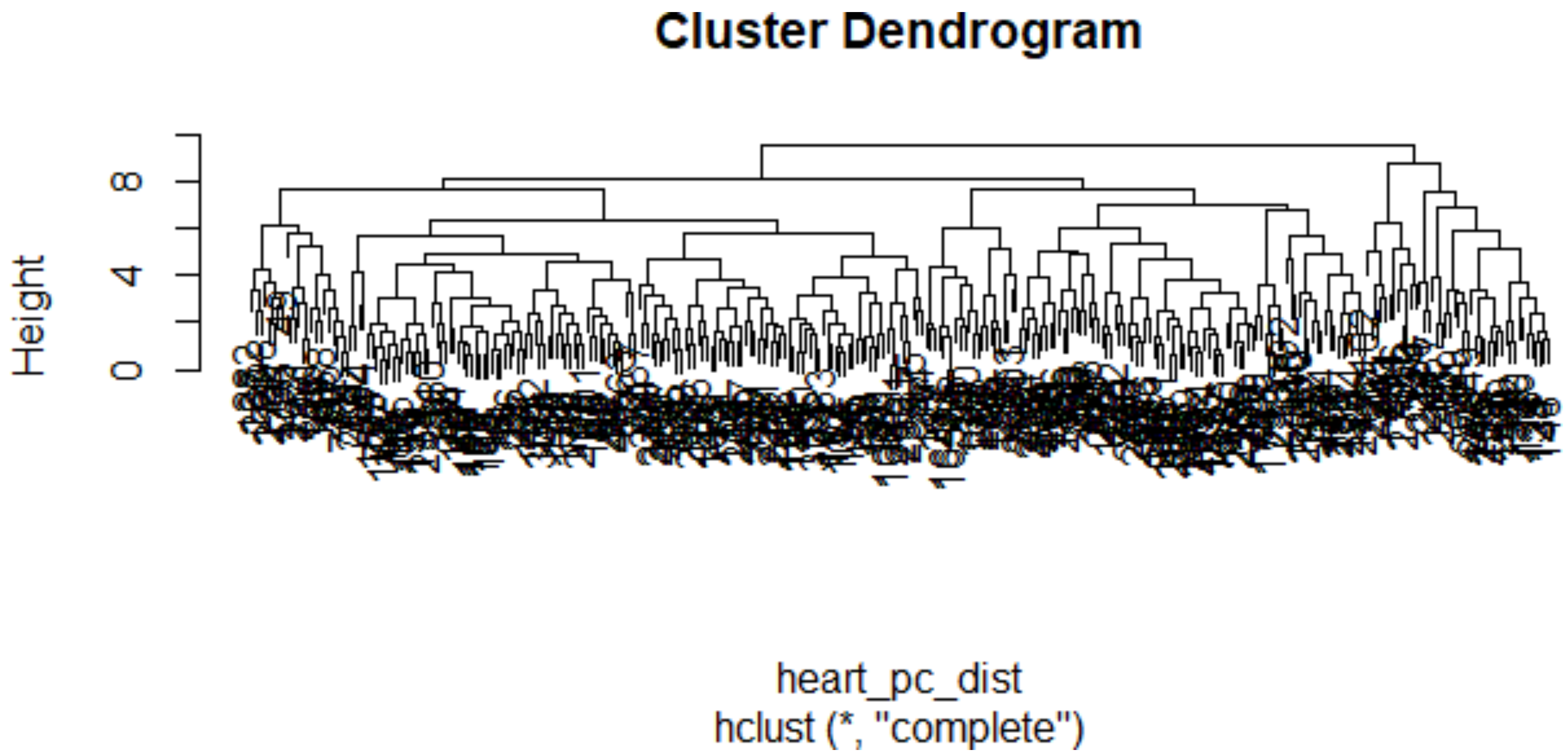



```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     1.6622 1.2396 1.10582 1.08681 1.01092
Proportion of Variance 0.2125 0.1182 0.09406 0.09086 0.07861
Cumulative Proportion  0.2125 0.3307 0.42481 0.51567 0.59428
                           PC6     PC7     PC8    PC9    PC10
Standard deviation     0.98489 0.92885 0.88088 0.8479 0.78840
Proportion of Variance 0.07462 0.06637 0.05969 0.0553 0.04781
Cumulative Proportion  0.66890 0.73527 0.79495 0.8503 0.89807
                          PC11    PC12   PC13
Standard deviation     0.72808 0.65049 0.6098
Proportion of Variance 0.04078 0.03255 0.0286
Cumulative Proportion  0.93885 0.97140 1.0000
```

Hirerachical Clustering

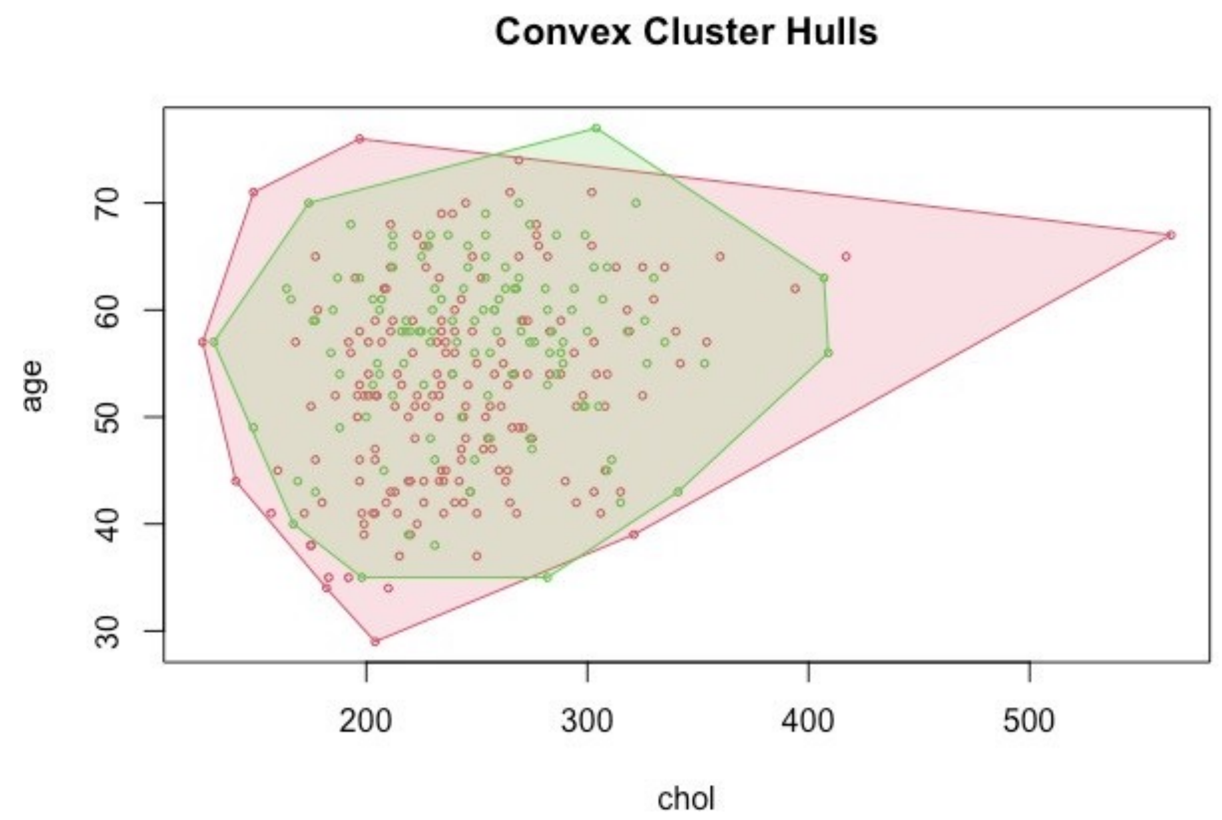
Hierachical Clustering

- There are 260 observations in Cluster 1 and 43 in Cluster 2

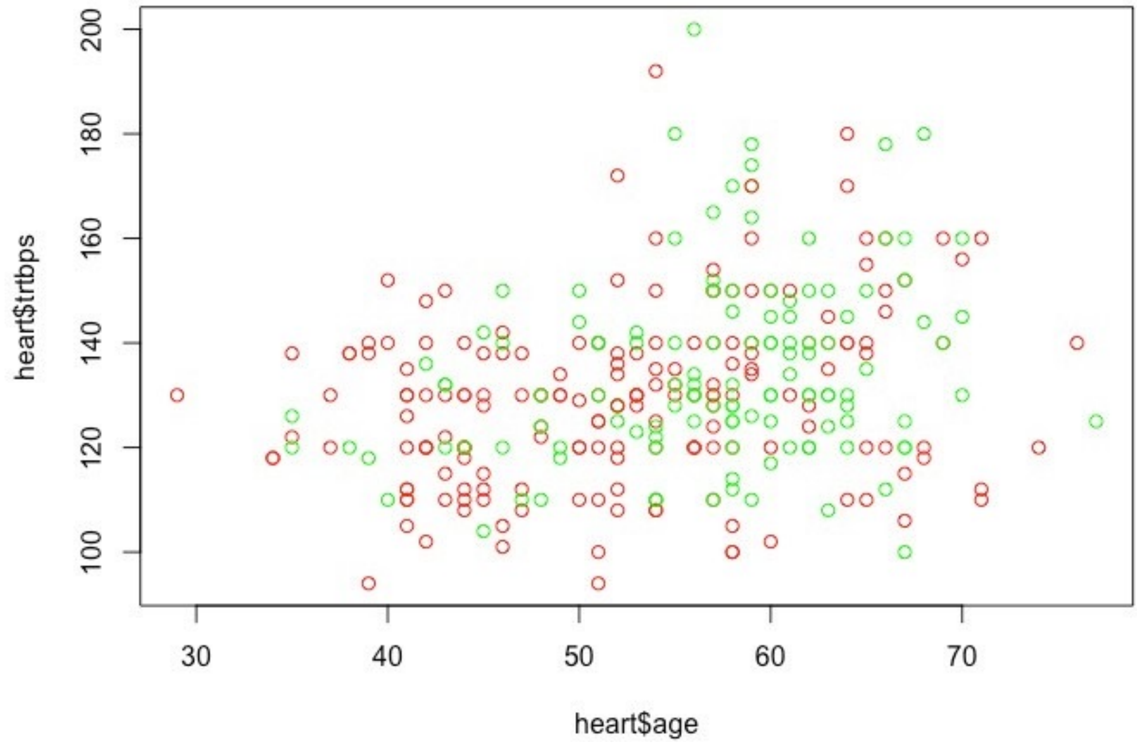


Prototype Clustering

K-Means and Fuzzy C-Means




```
Memberships:
               1         2
 [1,] 0.4999270 0.5000730
 [2,] 0.4994725 0.5005275
 [3,] 0.4977166 0.5022834
 [4,] 0.4968100 0.5031900
 [5,] 0.4993047 0.5006953
 [6,] 0.4988801 0.5011199
 [7,] 0.4988390 0.5011610
 [8,] 0.4976600 0.5023400
 [9,] 0.4992600 0.5007400
[10,] 0.4981491 0.5018509
[11,] 0.4977643 0.5022357
[12,] 0.4972856 0.5027144
[13,] 0.4964204 0.5035796
[14,] 0.5000661 0.4999339
[15,] 0.4989450 0.5010550
```



Regression

Logistic regression

The first model has several variables that were marked as not significant, and therefore it was created a second model

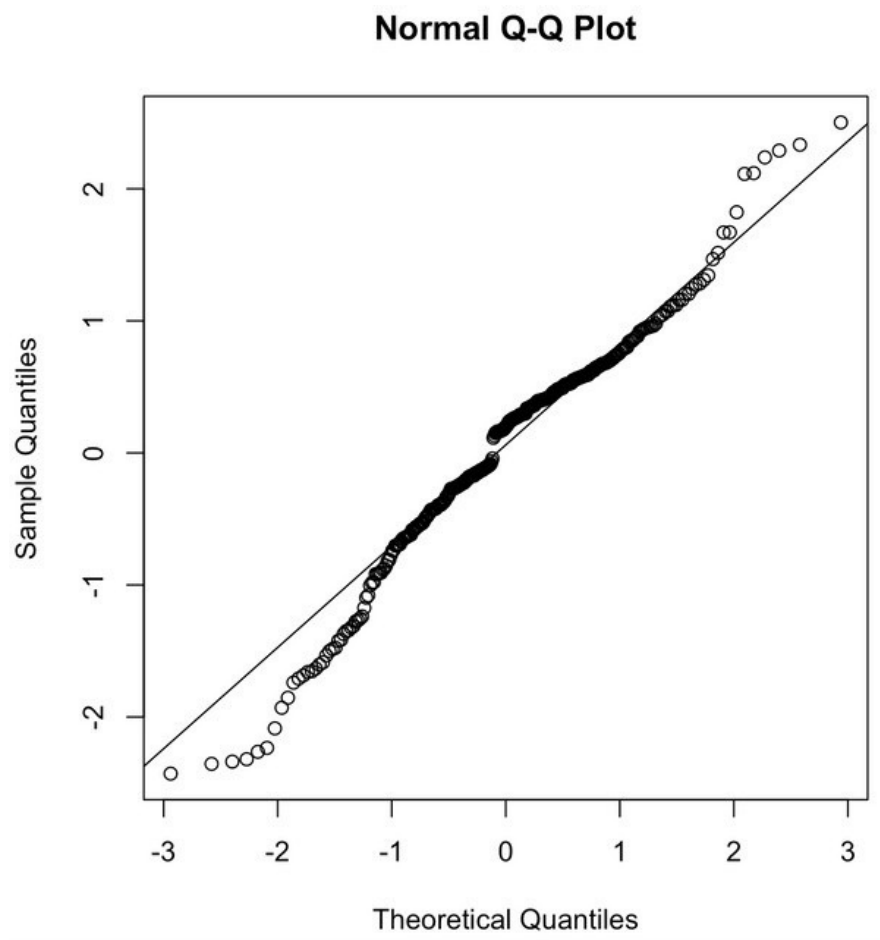
```
glm(formula = output ~ age + sex + cp + trtbps + chol + fbs +
    restecg + thalachh + exng + oldpeak + slp + caa + thall,
    family = binomial, data = heart)

glm(formula = output ~ sex + cp + thalachh + exng + oldpeak +
    caa + thall, family = binomial, data = heart)
```

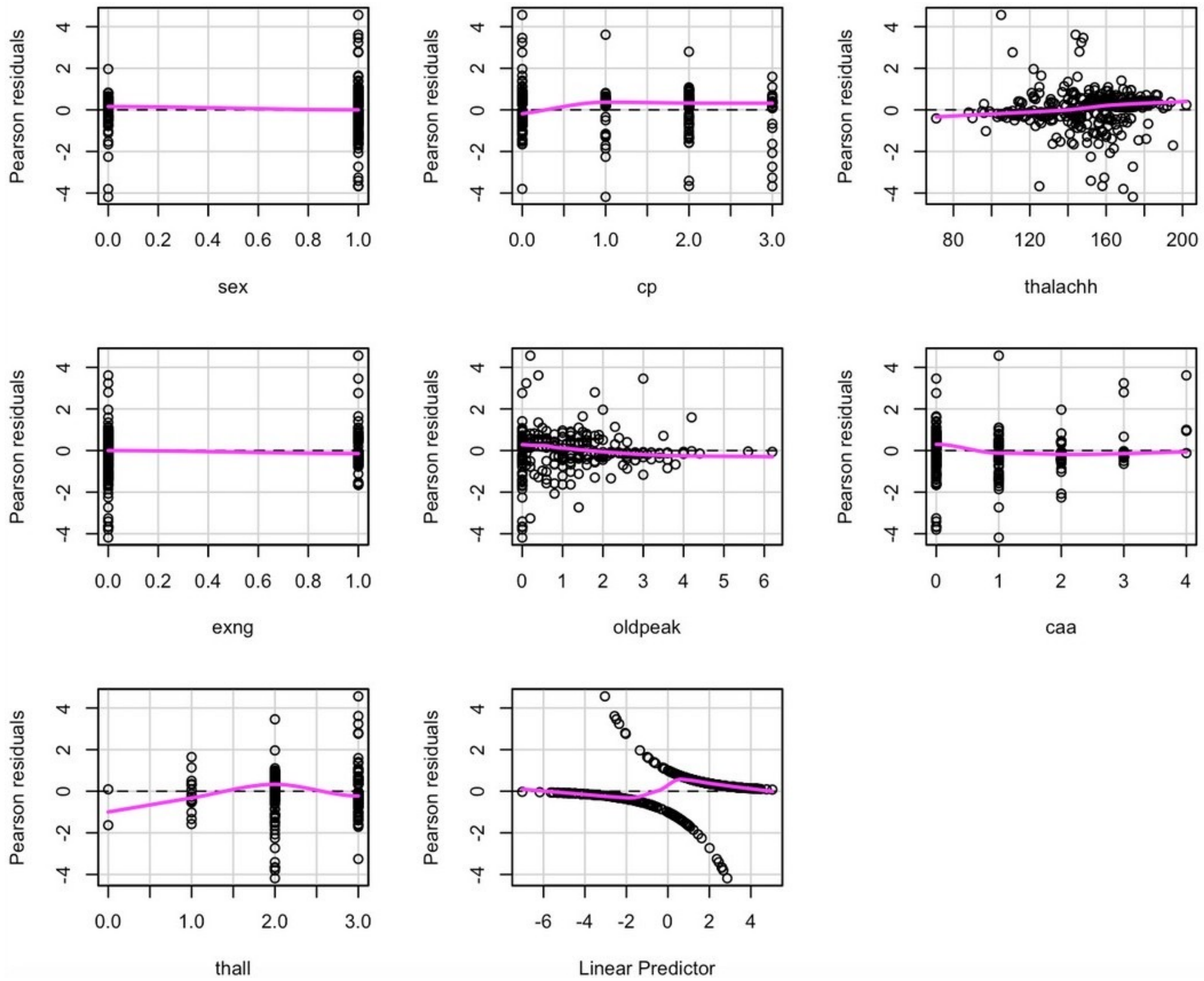
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.463553   1.481531   0.313 0.754366
sex         -1.389604   0.405754  -3.425 0.000615 ***
cp           0.787179   0.174709   4.506 6.62e-06 ***
thalachh     0.023665   0.008813   2.685 0.007248 **
exng        -1.044654   0.388978  -2.686 0.007239 **
oldpeak     -0.740612   0.182361  -4.061 4.88e-05 ***
caa         -0.713347   0.174387  -4.091 4.30e-05 ***
thall       -0.896269   0.274516  -3.265 0.001095 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 223.31  on 295  degrees of freedom
AIC: 239.31
```

```
> vif(logreg2)
     sex       cp thalachh     exng  oldpeak      caa    thall
1.089784 1.152667 1.136840 1.093343 1.110554 1.021568 1.027369
```

Logistic regression

Some other observations can be made about the model





Bayes Classifier

Bayes Classifier

Features: sex, cp, thalachh, exng, oldpeak, caa, thall.

```
   heart.pred4
     0  1
  0 24  9
  1  0 28
```

```
[1] 0.852459
```

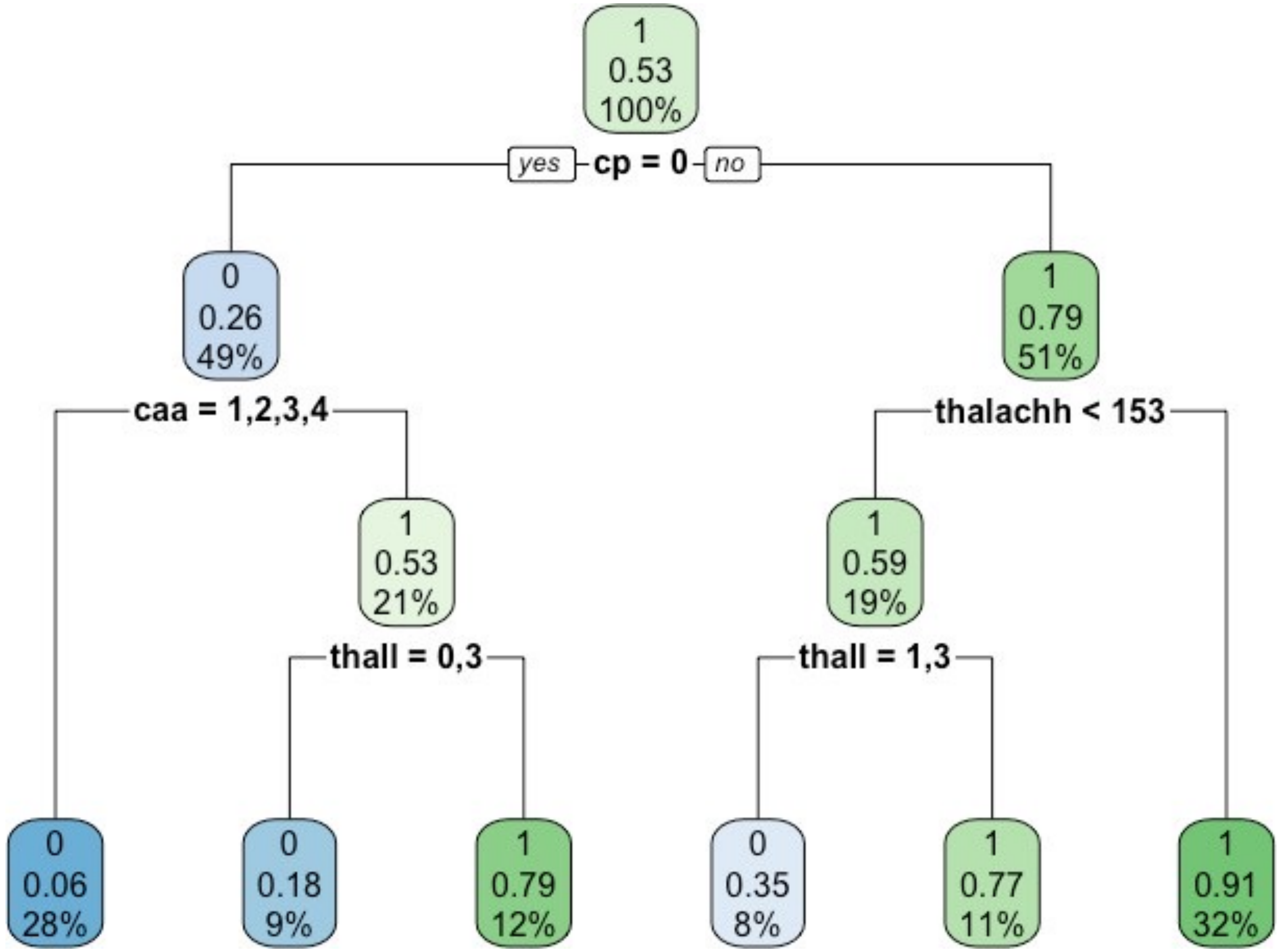
We aimed to predict the occurrence of heart disease based on various factors using a Naive Bayes classifier

To assess the model's generalization performance, the dataset was randomly split into an 80% training set and a 20% testing set.

Decision Tree

Decision Tree

Here on the right, it is reported the decision tree that best represents the data with an overall accuracy of 80%.



THANK YOU!

for your attention