Heart Disease Analysis

aurellaa

2025-03-05

1. Introduction

The Heart Disease UCI dataset contains medical records of patients with various risk factors for heart disease. This dataset is commonly used for predictive modeling and medical research. It includes features such as age, sex, chest pain type, cholesterol levels, blood pressure, and more.

This report aims to identify significant predictors of heart disease and assess how well a logistic regression model can classify patients as healthy or unhealthy based on their characteristics.

This project was made alongside Josh Starmer's StatQuest video: "Logistic Regression, Clearly Explained!!!!".

2. Data Inspection and Overview

2.1 Loading Required Libraries

Below is a list of the packages used in this report:

```r
library(dplyr)
library(ggplot2)
library(cowplot)
```

2.2 Inspecting the Dataset

```r
# Loading the dataset
heart_data <- read.csv("heart_disease_uci.csv")

# First looks at the data
head(heart_data)
```

```
##   id age    sex   dataset              cp trestbps chol   fbs        restecg
## 1  1  63   Male Cleveland  typical angina      145  233  TRUE lv hypertrophy
## 2  2  67   Male Cleveland    asymptomatic      160  286 FALSE lv hypertrophy
## 3  3  67   Male Cleveland    asymptomatic      120  229 FALSE lv hypertrophy
## 4  4  37   Male Cleveland     non-anginal      130  250 FALSE         normal
## 5  5  41 Female Cleveland atypical angina      130  204 FALSE lv hypertrophy
## 6  6  56   Male Cleveland atypical angina      120  236 FALSE         normal
##   thalch exang oldpeak       slope ca              thal num
## 1    150 FALSE     2.3 downsloping  0      fixed defect   0
## 2    108  TRUE     1.5        flat  3            normal   2
## 3    129  TRUE     2.6        flat  2 reversable defect   1
## 4    187 FALSE     3.5 downsloping  0            normal   0
## 5    172 FALSE     1.4   upsloping  0            normal   0
## 6    178 FALSE     0.8   upsloping  0            normal   0
```

```
str(heart_data)
```

```
## 'data.frame':    920 obs. of  16 variables:
##  $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age     : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : chr  "Male" "Male" "Male" "Male" ...
##  $ dataset : chr  "Cleveland" "Cleveland" "Cleveland" "Cleveland" ...
##  $ cp      : chr  "typical angina" "asymptomatic" "asymptomatic" "non-anginal" ...
##  $ trestbps: int  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs     : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ restecg : chr  "lv hypertrophy" "lv hypertrophy" "lv hypertrophy" "normal" ...
##  $ thalch  : int  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang   : logi  FALSE TRUE TRUE FALSE FALSE FALSE ...
##  $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope   : chr  "downsloping" "flat" "flat" "downsloping" ...
##  $ ca      : int  0 3 2 0 0 0 2 0 1 0 ...
##  $ thal    : chr  "fixed defect" "normal" "reversable defect" "normal" ...
##  $ num     : int  0 2 1 0 0 0 3 0 2 1 ...
```

Inspecting the dataset reveals that there are 920 observations for 16 variables:

- "id", which gives a unique ID for each patient.
- "age", the age of the patient in years.
- "sex", consisting of male and female.
- "dataset", which shows the location in which the data was taken. In this case, the data was collected from the city Cleveland.
- "cp", which stands for "chest pain". This variable consists of four categories:
  - 1: Typical angina
  - 2: Atypical angina
  - 3: Non-anginal pain
  - 4: Asymptomatic
- "trestbps", which is the resting blood pressure in mmHg.
- "chol", which is serum cholesterol in mg/dL.
- "fbs", which stands for fasting blood sugar. This variable is a binary one, where 1 indicates the value is more than 120mg/dL, and 0 indicates that the value is less than or equal to 120mg/dL.
- "restecg", which is the resting ECG. There are 3 categories:
  - 0: Normal
  - 1: ST-T wave abnormality

  - 2: Left ventricular hypertrophy

- “exang”, which stands for “Exercise Induced Angina”. 1 is true, and 0 is false.
- “oldpeak”, which is the ST depression induced by exercise.
- “ca” which is the number of major vessels (0 to 3) that are visible during fluoroscopy. 0 indicates no visible narrowing, while 3 indicates severe blockage.
- “thal”, short for thalassemia (a genetic blood disorder affecting hemoglobin production). There are three categories:
  - 1: Normal
  - 2: Fixed defect
  - 3: Reversible defect
- “num”, which specifies to diagnosis of heart disease. Values range from 0-4, depending on severity of disease (where 0 is healthy, and 1-4 is unhealthy).

2.3 Recoding the Variables

Some of the variables may be recoded in order to better suit the aim of this analysis.

The “num” variable is renamed to “heart_disease” for better clarity. This variable originally took values 0-4 indicating presence and severity of heart disease. For the purpose of conducting logistic regression, it has been converted into a binary variable where 0 indicates absence of heart disease and 1 indicates presence of heart disease.

Additionally, since logistic regression requires categorical variables to be treated as factors, relevant columns like sex, cp, and restecg have been converted to factors for proper analysis.

```r
heart_data <- heart_data |>
  mutate(sex = as.factor(sex),
         dataset = as.factor(dataset),
         cp = as.factor(cp),
         restecg = as.factor(restecg),
         slope = as.factor(slope),
         thal = as.factor(thal),
         ca = as.factor(ca)) |>
  mutate(num = ifelse(num > 0, 1, 0)) |>
  rename(heart_disease = num)
```

```r
# converting the heart disease column such that 0 is "healthy" and 1 is "unhealthy"

heart_data$heart_disease <- ifelse(test= heart_data$heart_disease == 0,
                                   yes = "Healthy", no = "Unhealthy")
heart_data$heart_disease <- as.factor(heart_data$heart_disease)
```

The following displays the dataset with the newly added changes:

```r
str(heart_data)
```

```
## 'data.frame':    920 obs. of  16 variables:
##  $ id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age          : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 2 1 1 2 2 ...
##  $ dataset      : Factor w/ 4 levels "Cleveland","Hungary",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ cp           : Factor w/ 4 levels "asymptomatic",..: 4 1 1 3 2 2 1 1 1 1 ...
##  $ trestbps     : int  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol         : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs          : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ restecg      : Factor w/ 4 levels "","lv hypertrophy",..: 2 2 2 3 2 3 2 3 2 2 ...
##  $ thalch       : int  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang        : logi  FALSE TRUE TRUE FALSE FALSE FALSE ...
##  $ oldpeak      : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope        : Factor w/ 4 levels "","downsloping",..: 2 3 3 2 4 4 2 4 3 2 ...
##  $ ca           : Factor w/ 4 levels "0","1","2","3": 1 4 3 1 1 1 3 1 2 1 ...
##  $ thal         : Factor w/ 4 levels "","fixed defect",..: 2 3 4 3 3 3 3 3 4 4 ...
##  $ heart_disease: Factor w/ 2 levels "Healthy","Unhealthy": 1 2 2 1 1 1 2 1 2 2 ...
```

```r
head(heart_data)
```

```
##   id age    sex   dataset              cp trestbps chol   fbs        restecg
## 1  1  63   Male Cleveland  typical angina      145  233  TRUE lv hypertrophy
## 2  2  67   Male Cleveland    asymptomatic      160  286 FALSE lv hypertrophy
## 3  3  67   Male Cleveland    asymptomatic      120  229 FALSE lv hypertrophy
## 4  4  37   Male Cleveland     non-anginal      130  250 FALSE         normal
## 5  5  41 Female Cleveland atypical angina      130  204 FALSE lv hypertrophy
## 6  6  56   Male Cleveland atypical angina      120  236 FALSE         normal
##   thalch exang oldpeak       slope ca              thal heart_disease
## 1    150 FALSE     2.3 downsloping  0      fixed defect       Healthy
## 2    108  TRUE     1.5        flat  3            normal     Unhealthy
## 3    129  TRUE     2.6        flat  2 reversable defect     Unhealthy
## 4    187 FALSE     3.5 downsloping  0            normal       Healthy
## 5    172 FALSE     1.4   upsloping  0            normal       Healthy
## 6    178 FALSE     0.8   upsloping  0            normal       Healthy
```

2.4 Checking for Imbalance of Data

Before fitting a logistic regression model, it is important to check whether the data is imbalanced. A highly imbalanced dataset—where one class (e.g., "Healthy") dominates the other ("Unhealthy")—can lead to biased predictions. If the model sees too few cases of a particular condition, it may not learn to predict that outcome well.

To check for imbalance, we use cross-tabulations (xtabs) to see the distribution of key categorical variables across the two heart disease categories.

```r
# do healthy and diseased samples come from each gender?
xtabs(~ heart_disease + sex, data = heart_data)
```

```
##              sex
## heart_disease Female Male
##     Healthy      144  267
##     Unhealthy     50  459
```

```r
# were all 4 levels of chest pain reported by many patients?
xtabs(~ heart_disease + cp, data = heart_data)
```

```
##              cp
## heart_disease asymptomatic atypical angina non-anginal typical angina
##     Healthy            104             150         131             26
##     Unhealthy          392              24          73             20
```

```r
# were both high and low fasting blood sugar reported by many patients?
xtabs(~ heart_disease + fbs, data = heart_data)
```

```
##              fbs
## heart_disease FALSE TRUE
##     Healthy     353   44
##     Unhealthy   339   94
```

```r
# were all levels of restecg reported by many patients?
xtabs(~ heart_disease + restecg, data = heart_data)
```

```
##              restecg
## heart_disease     lv hypertrophy normal st-t abnormality
##     Healthy     0             82    268               61
##     Unhealthy   2            106    283              118
```

From initial checks, it appears that ST-T wave abnormality has only 5 patients reported (2 healthy, 3 unhealthy). This small sample size might reduce the reliability of this variable as a predictor.

2.5 Handling Missing Values

```r
# check for missing values
colSums(is.na(heart_data))
```

```
##            id           age           sex       dataset            cp
##             0             0             0             0             0
##      trestbps          chol           fbs       restecg        thalch
##            59            30            90             0            55
##         exang       oldpeak         slope            ca          thal
##            55            62             0           611             0
## heart_disease
##             0
```

It should be noted that omission of missing values may impose bias on the results of the statistical analyses.

```r
# removing the missing values
heart_data <- na.omit(heart_data)
```

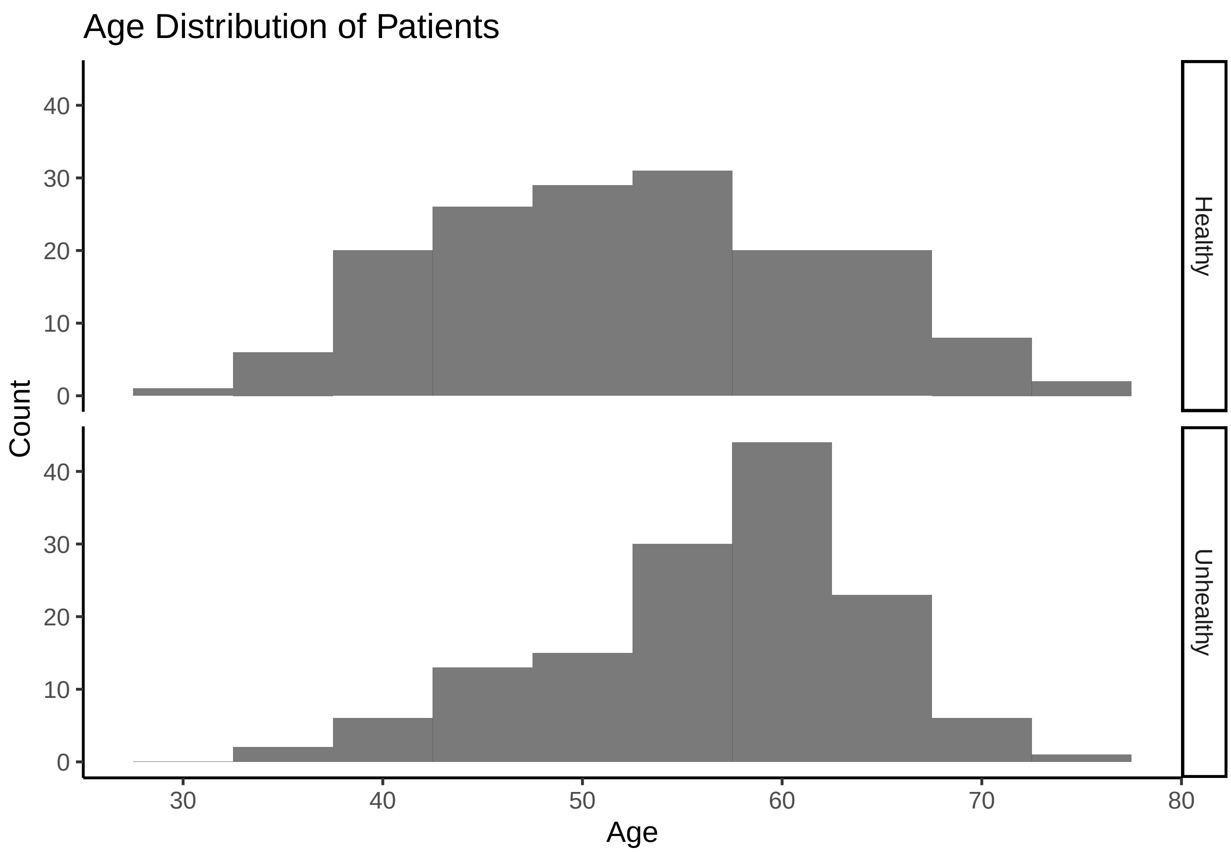
```r
# summary statistics
summary(heart_data)
```

```
##        id             age             sex                dataset
##  Min.   :  1.0   Min.   :29.00   Female: 97   Cleveland    :299
##  1st Qu.: 76.5   1st Qu.:48.00   Male  :206   Hungary      :  2
##  Median :152.0   Median :56.00                Switzerland  :  0
##  Mean   :156.9   Mean   :54.51                VA Long Beach:  2
##  3rd Qu.:229.5   3rd Qu.:61.00
##  Max.   :760.0   Max.   :77.00
##                cp          trestbps          chol          fbs
##  asymptomatic   :146   Min.   : 94.0   Min.   :  0.0   Mode :logical
##  atypical angina: 50   1st Qu.:120.0   1st Qu.:211.0   FALSE:259
##  non-anginal    : 84   Median :130.0   Median :240.0   TRUE :44
##  typical angina : 23   Mean   :131.7   Mean   :245.5
##                        3rd Qu.:140.0   3rd Qu.:275.0
##                        Max.   :200.0   Max.   :564.0
##              restecg        thalch        exang            oldpeak
##                  :  0   Min.   : 71.0   Mode :logical   Min.   :0.000
##  lv hypertrophy  :147   1st Qu.:132.0   FALSE:202       1st Qu.:0.000
##  normal          :151   Median :152.0   TRUE :101       Median :0.800
##  st-t abnormality:  5   Mean   :149.2                   Mean   :1.053
##                         3rd Qu.:165.0                   3rd Qu.:1.600
##                         Max.   :202.0                   Max.   :6.200
##          slope     ca                     thal       heart_disease
##             :  2   0:180                    :  4   Healthy  :163
##  downsloping: 21   1: 65   fixed defect     : 18   Unhealthy:140
##  flat       :140   2: 38   normal           :164
##  upsloping  :140   3: 20   reversable defect:117
##
##
```

3. Explanatory data analysis

3.1 Age Distribution of Healthy and Unhealthy Patients

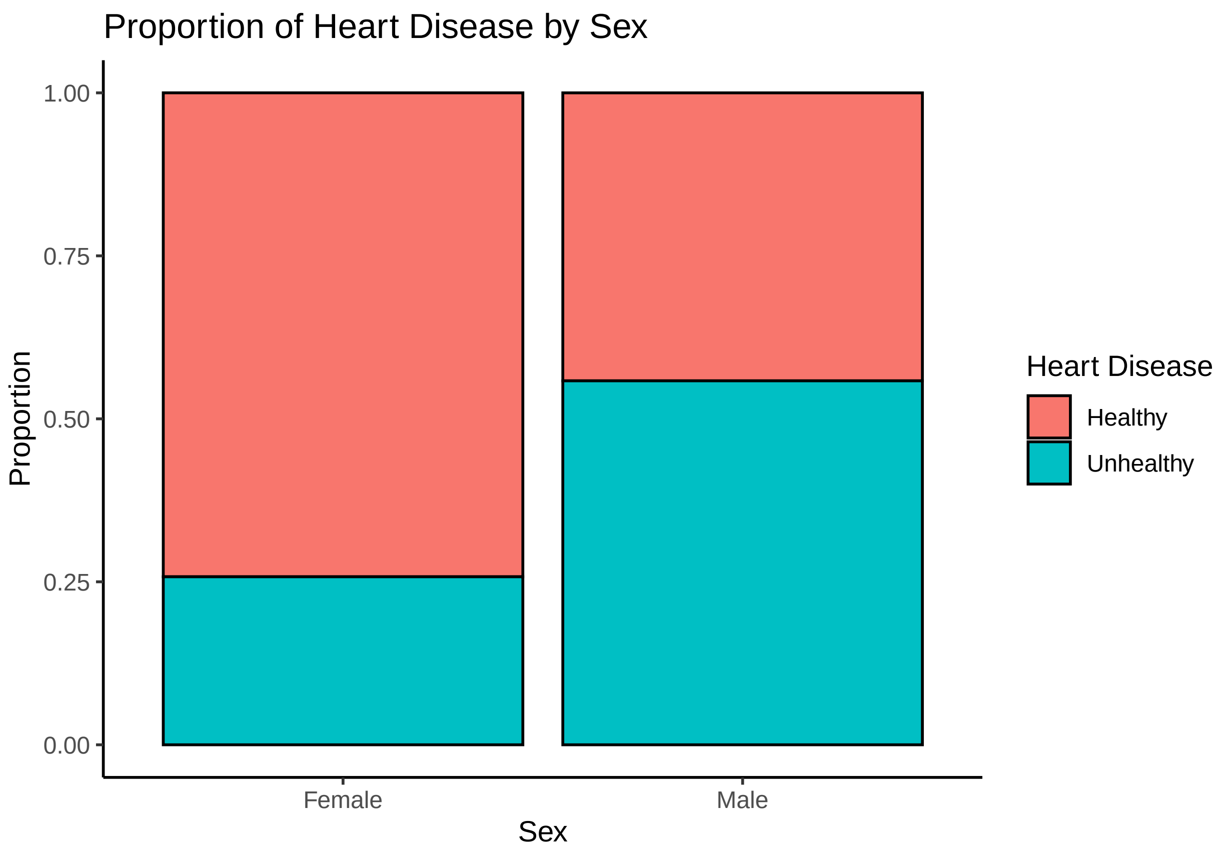
```
# age distribution
ggplot(heart_data, aes(x = age)) +
  geom_histogram(binwidth = 5, alpha = 0.8) +
  facet_grid(rows = vars(heart_disease)) +
  labs(title = "Age Distribution of Patients", x = "Age", y = "Count")
```



The majority of patients with heart disease are concentrated within the 55–65 age range, suggesting an increased likelihood of heart disease among older individuals. However, it is important to note that both healthy and unhealthy patients are predominantly within the 50–60 age range. Given the similar age distribution in both groups, age alone may have limited predictive power in distinguishing heart disease cases.

3.2 Proportion of Heart Disease by Sex

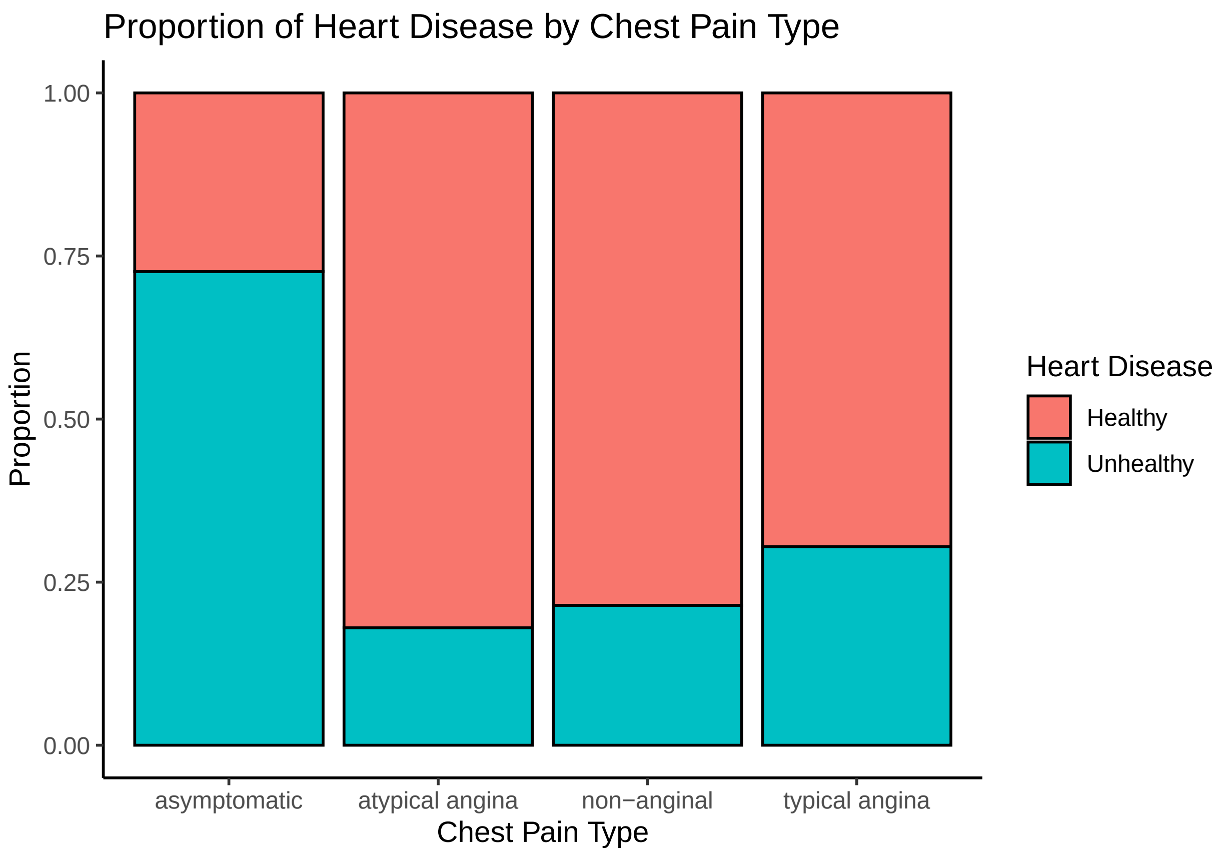
```r
# visualizing the data (heart disease proportion by sex)
ggplot(heart_data, aes(x = as.factor(sex), fill = as.factor(heart_disease))) +
  geom_bar(position = "fill", color = "black") +
  labs(x = "Sex", y = "Proportion", title = "Proportion of Heart Disease by Sex",
       fill = "Heart Disease")
```



Males seem to have a significantly higher proportion of heart disease cases than females. This suggests that males may have higher heart disease risk compared to females.

3.3 Proportion of Heart Disease by Chest Pain Type

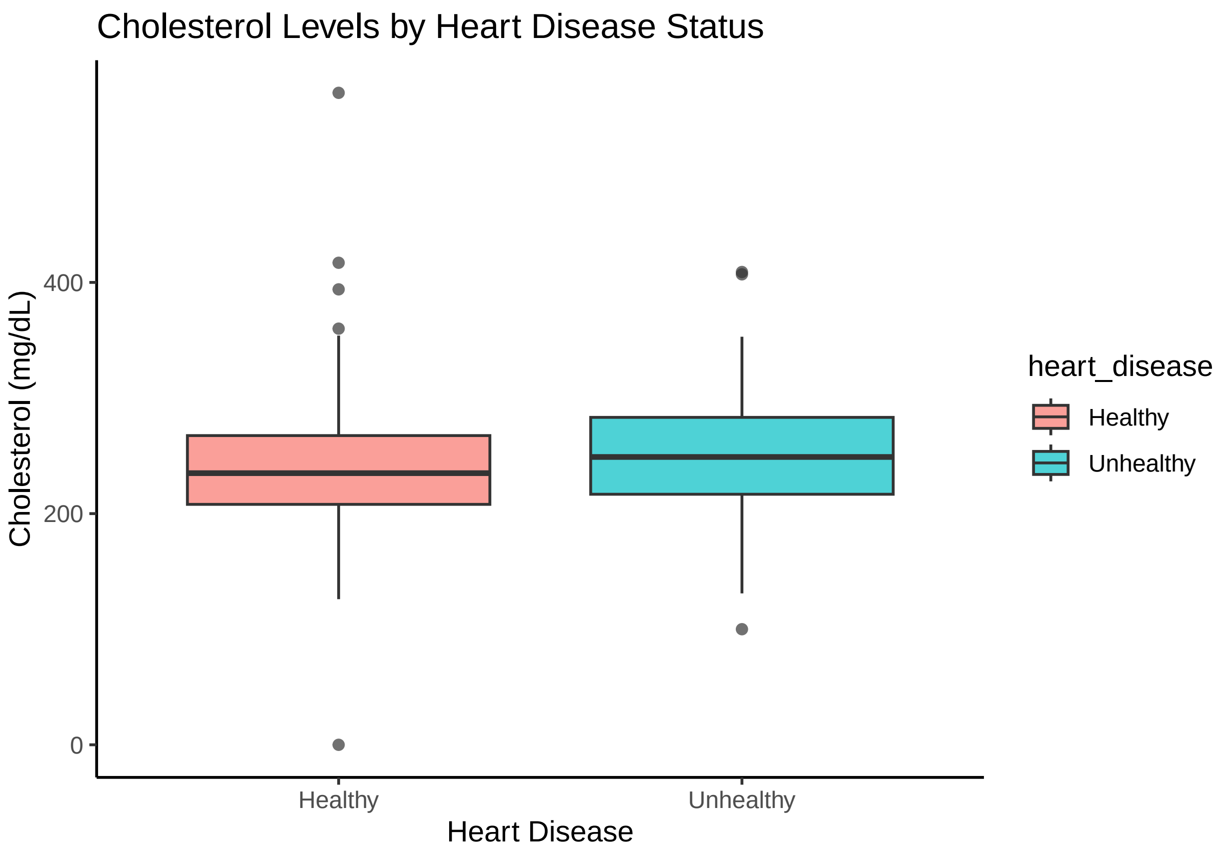
```r
# visualizing the data (heart disease proportion by chest pain type)
ggplot(heart_data, aes(x = cp, fill = heart_disease)) +
  geom_bar(position = "fill", color = "black") +
  labs(title = "Proportion of Heart Disease by Chest Pain Type",
       x = "Chest Pain Type", y = "Proportion", fill = "Heart Disease")
```



Patients who experienced asymptomatic chest pain have the highest proportion of unhealthy cases, suggesting that the absence of typical chest pain symptoms does not indicate a lower risk of heart disease.

3.4 Cholesterol Levels by Heart Disease Status

```r
# visualizing the data (heart disease proportion by cholesterol)
ggplot(heart_data, aes(x = heart_disease, y = chol, fill = heart_disease)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Cholesterol Levels by Heart Disease Status",
       x = "Heart Disease", y = "Cholesterol (mg/dL)")
```



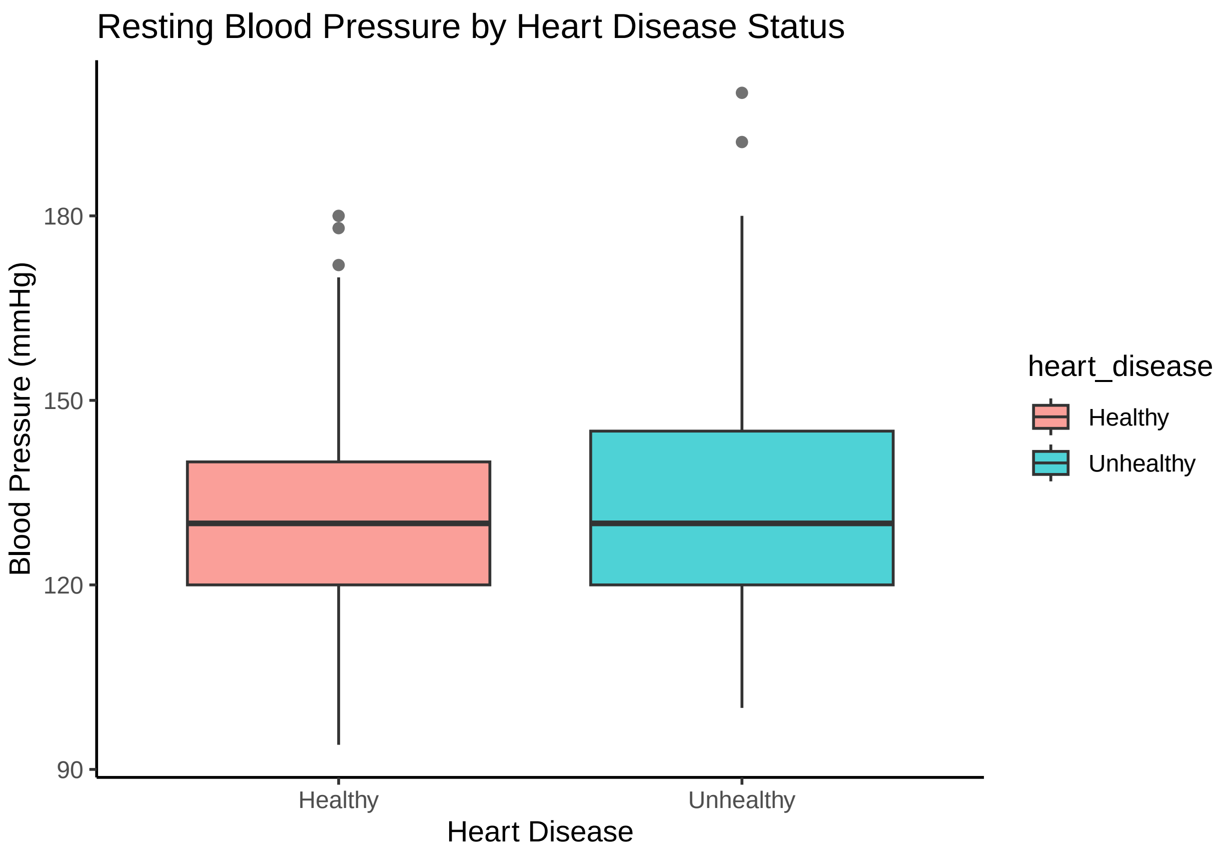
Both groups of patients show a wide range of cholesterol levels.

The median cholesterol level seems slightly higher in the unhealthy group compared to the healthy group.

Both groups have outliers with cholesterol levels higher than 400 mg/dL, but there are more outliers in the healthy group, with some very extreme cholesterol values. Interestingly, there is one outlier with a cholesterol level of nearly 0 mg/dL in the healthy group.

3.5 Resting Blood Pressure by Heart Disease Status

```
# visualizing the data (heart disease proportion by trestbps)
ggplot(heart_data, aes(x = heart_disease, y = trestbps, fill = heart_disease)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Resting Blood Pressure by Heart Disease Status",
       x = "Heart Disease", y = "Blood Pressure (mmHg)")
```

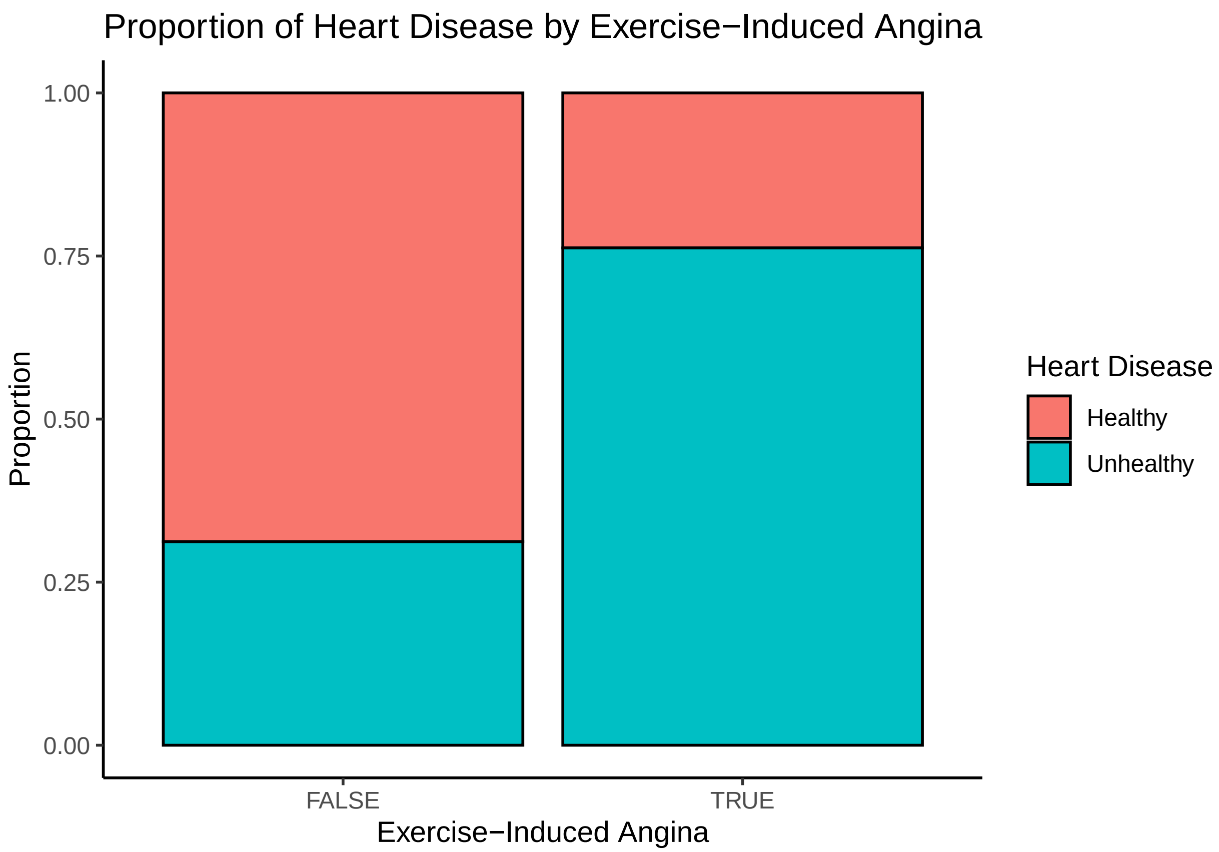


The median resting blood pressure is slightly higher in the unhealthy group than in the healthy group. This suggests that people with heart disease tend to have higher resting blood pressure.

Outliers are present in both groups, however the unhealthy group has more extreme outliers, possibly indicating more individuals with dangerously high blood pressures.

3.6 Proportion of Heart Disease by Exercise-Induced Angina

```
# visualizing the data (heart disease proportion by exang)
ggplot(heart_data, aes(x = exang, fill = heart_disease)) +
  geom_bar(position = "fill", color = "black") +
  labs(title = "Proportion of Heart Disease by Exercise-Induced Angina",
       x = "Exercise-Induced Angina", y = "Proportion", fill = "Heart Disease")
```



There is a larger proportion of unhealthy patients in those that have exercise-induced angina, suggesting that the presence of exercise-induced angina increases the likelihood of having heart disease.

4. Simple Logistic Regression

In this section, a logistic regression model is fit to predict heart disease status (heart_disease) using a single predictor: sex.

The logistic regression model is fit using the glm() function:

```r
logistic1 <- glm(heart_disease ~ sex, data = heart_data, family = "binomial")
```

Examining the results:

```r
summary(residuals(logistic1, type = "deviance"))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.27829 -1.27829 -0.77207 -0.02169  1.07976  1.64671
```

```r
summary(logistic1)
```

```
##
## Call:
## glm(formula = heart_disease ~ sex, family = "binomial", data = heart_data)
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0578     0.2321  -4.557 5.20e-06 ***
## sexMale       1.2919     0.2712   4.763 1.91e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 418.30  on 302  degrees of freedom
## Residual deviance: 393.48  on 301  degrees of freedom
## AIC: 397.48
##
## Number of Fisher Scoring iterations: 4
```

Interpretations:

- The deviance residuals appear roughly symmetrical and centered around 0, indicating that the model is fitting the data reasonably well.
- The coefficients correspond to the following model:

heart disease = -1.0578 + 1.2919 x the patient is male

- 1.2919 is the log(odds ratio) of the odds that a male will have heart disease over the odds that a female will. In other words, the odds of heart disease in males are $^{1.2919}$ 3.64 times higher than in females.
- The p-values are < 0.05, and therefore the log(odds) and log(odds ratios) are both statistically significant.
- The number of Fisher Scoring iterations tell us how quickly the glm() function converged on the maximum likelihood estimates for the coefficients. If the model takes too many iterations, it may indicate convergence issues, but here, the model converged efficiently.

From this simple logistic regression, it can be concluded that males have significantly higher odds of developing heart disease compared to females. However, since heart disease is influenced by multiple factors, this analysis will be extended by including additional predictors in a multiple logistic regression model.

5. Multiple Logistic Regression

5.1 Fitting the Logistic Regression Model

```r
logistic2 <- glm(heart_disease ~ ., data = heart_data, family = "binomial")
summary(logistic2)
```

```
##
## Call:
## glm(formula = heart_disease ~ ., family = "binomial", data = heart_data)
##
## Coefficients:
```

```
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -3.396e+01  2.150e+03  -0.016 0.987398
## id                        2.802e-03  2.341e-03   1.197 0.231339
## age                      -2.424e-02  2.532e-02  -0.957 0.338368
## sexMale                   1.775e+00  5.602e-01   3.169 0.001527 **
## datasetHungary            1.377e+01  4.855e+03   0.003 0.997737
## datasetVA Long Beach      1.224e+01  1.541e+03   0.008 0.993660
## cpatypical angina        -9.381e-01  5.774e-01  -1.625 0.104210
## cpnon-anginal            -1.977e+00  5.221e-01  -3.786 0.000153 ***
## cptypical angina         -2.510e+00  7.304e-01  -3.436 0.000590 ***
## trestbps                  2.772e-02  1.194e-02   2.322 0.020214 *
## chol                      4.899e-03  4.110e-03   1.192 0.233238
## fbsTRUE                  -5.124e-01  5.883e-01  -0.871 0.383772
## restecgnormal            -5.156e-01  3.985e-01  -1.294 0.195760
## restecgst-t abnormality  2.569e-01  2.774e+00   0.093 0.926196
## thalch                   -1.782e-02  1.175e-02  -1.517 0.129303
## exangTRUE                 6.859e-01  4.517e-01   1.518 0.128891
## oldpeak                   4.376e-01  2.432e-01   1.799 0.072042 .
## slopedownsloping          3.262e+01  2.150e+03   0.015 0.987894
## slopeflat                 3.328e+01  2.150e+03   0.015 0.987649
## slopeupsloping            3.197e+01  2.150e+03   0.015 0.988138
## ca1                       2.252e+00  5.162e-01   4.363 1.28e-05 ***
## ca2                       3.276e+00  7.921e-01   4.136 3.54e-05 ***
## ca3                       2.152e+00  9.137e-01   2.356 0.018494 *
## thalfixed defect         -2.562e+00  2.761e+00  -0.928 0.353383
## thalnormal               -2.299e+00  2.673e+00  -0.860 0.389804
## thalreversable defect    -8.833e-01  2.677e+00  -0.330 0.741462
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 418.3  on 302  degrees of freedom
## Residual deviance: 182.1  on 277  degrees of freedom
## AIC: 234.1
##
## Number of Fisher Scoring iterations: 15
```

```
contrasts(heart_data$cp)
```

```
##                 atypical angina non-anginal typical angina
## asymptomatic                  0           0              0
## atypical angina               1           0              0
## non-anginal                   0           1              0
## typical angina                0           0              1
```

Interpretations:

- Sex (Male) ($p = 0.0015$, coef = 1.775)
  - $p = 0.0015$ is < 0.05 , meaning that sex is a useful predictor of heart disease
  - The coefficients correspond to the following model:

$$\log\left(\frac{1-P(\text{heart disease})}{P(\text{heart disease})}\right) = -33.96 + 1.775 \times (\text{Male})$$

- In other words, when the patient is a female, the log(odds of heart disease) are -33.96.
- And when the patient is a male, the log(odds of heart disease) are -32.185.
- Males have higher odds of heart disease compared to females.
- Resting blood pressure (p = 0.02, coef = 0.0277) is a significant predictor of heart disease, where higher blood pressure slightly increases the risk for having heart disease.
- The number of major vessels (ranging from 0 to 3) visible during fluoroscopy, which indicate blockages, is a strong predictor of heart disease. An increase in the number of visible vessels (1, 2, or 3) is associated with higher odds of having heart disease.
- Some types of chest pain are also significant predictors of heart disease. Having non-anginal pain (p = 0.00015, coef = -1.977) and typical angina (p = 0.00059, coef = -2.510) decreases the odds for heart disease compared to asymptomatic chest pain.
- Having atypical angina, however, doesn't seem to be a strong predictor of heart disease in this dataset.
- Odds ratio = exp(1.775) 5.9, meaning the odds of heart disease for males are 5.9 times the odds for females.
- Surprisingly, age is not a significant predictor, with a p-value of 0.338. However, most patients in the dataset (both healthy and unhealthy), were mostly 50-60 year olds, which explains why age might have lost its predictive power
- Cholesterol is not a significant predictor with a p-value of 0.233. Cholesterol alone may not strongly predict heart disease.
- Fasting blood sugar also doesn't significantly predict heart disease, with a p-value of 0.383.
- Resting ECG has no significant impact on heart disease, with a p-value of 0.195-0.926
- Exercise-induced angina is also not a useful predictor with a p-value of 0.128.

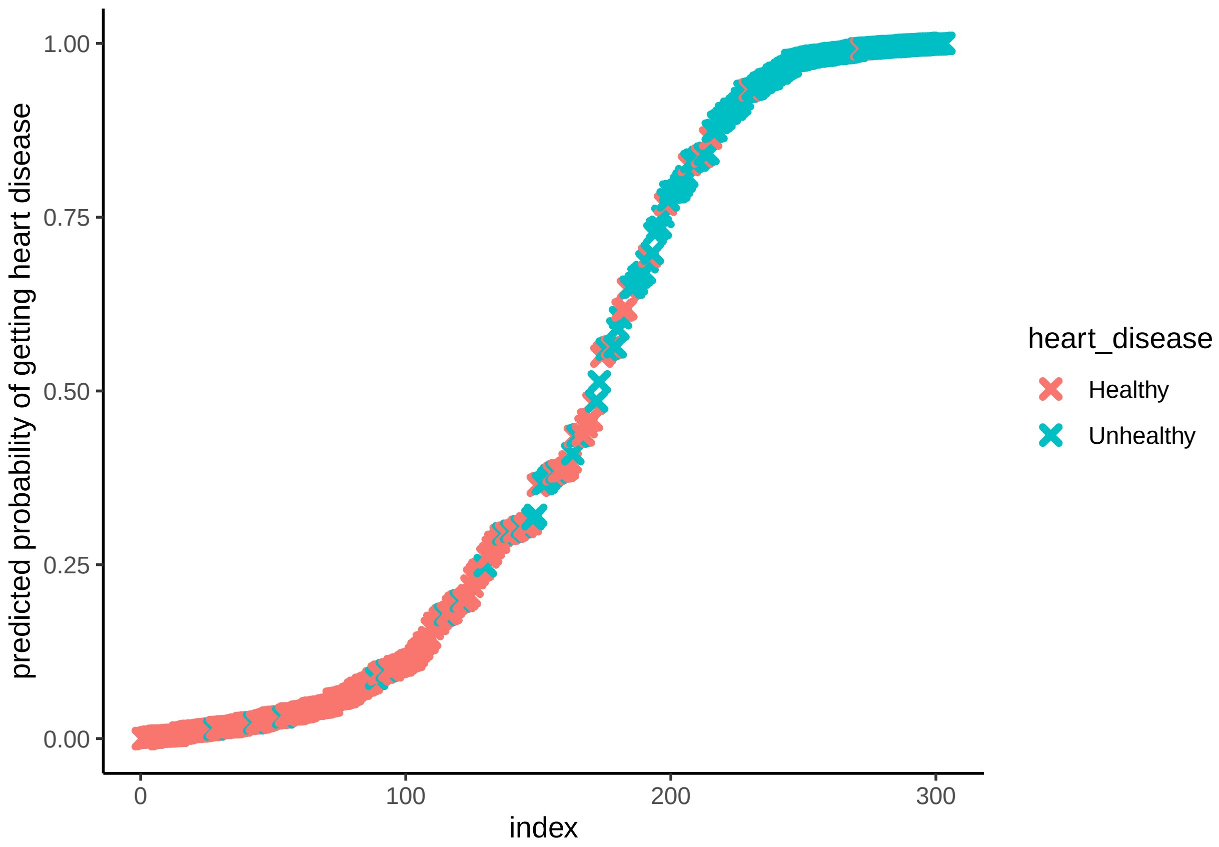
5.2 Visualising the Logistic Regression Model

```r
predicted.data <- data.frame(
  probability.of.hd = logistic2$fitted.values,
  heart_disease = heart_data$heart_disease
)

predicted.data <- predicted.data[
  order(predicted.data$probability.of.hd, decreasing = F),]

predicted.data$rank <- 1:nrow(predicted.data)

ggplot(data = predicted.data, aes(x = rank, y = probability.of.hd)) +
  geom_point(aes(color = heart_disease), alpha = 1, shape = 4, stroke = 2) +
  xlab("index") +
  ylab("predicted probability of getting heart disease")
```



We can see that the model correctly assigned high probabilities to unhealthy individuals and low probabilities to healthy ones. Additionally, the separation between the two groups is fairly strong, meaning the model is capturing useful patterns.