# Heart Disease Analysis

Tyler Campbell

2025-12-30

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

```
library(ca)
library(CCA)
```

```
## Loading required package: fda

## Loading required package: splines

## Loading required package: fds

## Loading required package: rainbow

## Loading required package: pcaPP

## Loading required package: RCurl
```

```
## Loading required package: deSolve

##
## Attaching package: 'fda'

## The following object is masked from 'package:graphics':
##
##     matplot

## Loading required package: fields

## Loading required package: spam

## Spam version 2.11-1 (2025-01-20) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.

##
## Attaching package: 'spam'

## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve

## Loading required package: viridisLite

## Loading required package: RColorBrewer

##
## Try help(fields) to get started.
```

```
library(sem)

library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(MVA)
```

```
## Loading required package: HSAUR2

## Loading required package: tools
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

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

```r
# Loading Data
data <- read.csv("C:/Users/Ty Campbell/Desktop/data/Heart_Disease_Prediction.csv")
head(data)
```

```
##   Age Sex Chest.pain.type  BP Cholesterol FBS.over.120 EKG.results Max.HR
## 1  70   1               4 130         322            0           2    109
## 2  67   0               3 115         564            0           2    160
## 3  57   1               2 124         261            0           0    141
## 4  64   1               4 128         263            0           0    105
## 5  74   0               2 120         269            0           2    121
## 6  65   1               4 120         177            0           0    140
##   Exercise.angina ST.depression Slope.of.ST Number.of.vessels.fluro Thallium
## 1               0           2.4           2                      3        3
## 2               0           1.6           2                      0        7
## 3               0           0.3           1                      0        7
## 4               1           0.2           2                      1        7
## 5               1           0.2           1                      1        3
## 6               0           0.4           1                      0        7
##   Heart.Disease
## 1      Presence
## 2       Absence
## 3      Presence
## 4       Absence
## 5       Absence
## 6       Absence
```

```r
# checking data types and null values
dim(data)
```

```
## [1] 270  14
```

```r
colSums(is.na(data))
```

```
##                     Age                     Sex          Chest.pain.type
##                       0                       0                        0
##                      BP             Cholesterol             FBS.over.120
##                       0                       0                        0
##             EKG.results                  Max.HR          Exercise.angina
##                       0                       0                        0
##           ST.depression             Slope.of.ST Number.of.vessels.fluro
##                       0                       0                        0
##                Thallium           Heart.Disease
##                       0                       0
```

```r
str(data)
```

```
## 'data.frame':    270 obs. of  14 variables:
##  $ Age                    : int  70 67 57 64 74 65 56 59 60 63 ...
##  $ Sex                    : int  1 0 1 1 0 1 1 1 1 0 ...
##  $ Chest.pain.type        : int  4 3 2 4 2 4 3 4 4 4 ...
##  $ BP                     : int  130 115 124 128 120 120 130 110 140 150 ...
```

```
##  $ Cholesterol            : int  322 564 261 263 269 177 256 239 293 407 ...
##  $ FBS.over.120           : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ EKG.results            : int  2 2 0 0 2 0 2 2 2 2 ...
##  $ Max.HR                 : int  109 160 141 105 121 140 142 142 170 154 ...
##  $ Exercise.angina        : int  0 0 0 1 1 0 1 1 0 0 ...
##  $ ST.depression          : num  2.4 1.6 0.3 0.2 0.2 0.4 0.6 1.2 1.2 4 ...
##  $ Slope.of.ST            : int  2 2 1 2 1 1 2 2 2 2 ...
##  $ Number.of.vessels.fluro: int  3 0 0 1 1 0 1 1 2 3 ...
##  $ Thallium               : int  3 7 7 7 3 7 6 7 7 7 ...
##  $ Heart.Disease          : chr  "Presence" "Absence" "Presence" "Absence" ...
```

```
# preprocessing and standardization
data$Heart.Disease <- ifelse(data$Heart.Disease == "Presence", 1, 0)
```

```
# standardize
df <- data %>%
  mutate(across(-Heart.Disease, scale))

head(df)
```

```
##         Age        Sex Chest.pain.type          BP Cholesterol FBS.over.120
## 1 1.7089201  0.6882217       0.8693133 -0.07527007   1.3996132   -0.4162558
## 2 1.3795779 -1.4476387      -0.1832185 -0.91506006   6.0817107   -0.4162558
## 3 0.2817705  0.6882217      -1.2357503 -0.41118607   0.2194151   -0.4162558
## 4 1.0502357  0.6882217       0.8693133 -0.18724207   0.2581101   -0.4162558
## 5 2.1480430 -1.4476387      -1.2357503 -0.63513007   0.3741952   -0.4162558
## 6 1.1600164  0.6882217       0.8693133 -0.63513007  -1.4057758   -0.4162558
##   EKG.results     Max.HR Exercise.angina ST.depression Slope.of.ST
## 1   0.9798441 -1.7559473      -0.6999225     1.1788233   0.6751655
## 2   0.9798441  0.4455818      -0.6999225     0.4802613   0.6751655
## 3  -1.0243824 -0.3745957      -0.6999225    -0.6549018  -0.9524656
## 4  -1.0243824 -1.9286162       1.4234380    -0.7422221   0.6751655
## 5   0.9798441 -1.2379404       1.4234380    -0.7422221  -0.9524656
## 6  -1.0243824 -0.4177629      -0.6999225    -0.5675816  -0.9524656
##   Number.of.vessels.fluro   Thallium Heart.Disease
## 1               2.4680989 -0.8740826             1
## 2              -0.7102161  1.1870729             0
## 3              -0.7102161  1.1870729             1
## 4               0.3492223  1.1870729             0
## 5               0.3492223 -0.8740826             0
## 6              -0.7102161  1.1870729             0
```

```
# data partitioning

n <- nrow(df)
idx <- sample(1:n, size = .7 * n)
train <- df[idx,]
valid <- df[-idx,]
dim(train)
```

```
## [1] 189  14
```

```r
dim(valid)
```

```
## [1] 81 14
```

```r
# logit model
logit_model <- glm(Heart.Disease ~ ., data = train, family = "binomial")
logit_predict <- predict(logit_model, newdata = valid, type = "response")
logit_class <- ifelse(logit_predict > .5, 1, 0)
table(Predicted = logit_class, Actual = valid$Heart.Disease)
```

```
##          Actual
## Predicted  0  1
##         0 41  9
##         1  6 25
```

```r
mean(logit_class == valid$Heart.Disease)
```

```
## [1] 0.8148148
```

```r
# logit performance
TP <- 21
TN <- 41
FP <- 10
FN <- 9

accuracy  <- (TP + TN) / (TP + TN + FP + FN)
precision <- TP / (TP + FP)
recall    <- TP / (TP + FN)
specificity <- TN / (TN + FP)

accuracy
```

```
## [1] 0.7654321
```

```r
precision
```

```
## [1] 0.6774194
```

```r
recall
```

```
## [1] 0.7
```

```r
specificity
```

```
## [1] 0.8039216
```

```r
# install.packages("xgboost")
library(xgboost)
```

```
## Warning: package 'xgboost' was built under R version 4.4.3
```

```r
# Assume train and valid have Heart.Disease as 0/1
y_train <- train$Heart.Disease
y_valid <- valid$Heart.Disease

X_train <- model.matrix(Heart.Disease ~ . - 1, data = train)
X_valid <- model.matrix(Heart.Disease ~ . - 1, data = valid)

dtrain <- xgb.DMatrix(data = X_train, label = y_train)
dvalid <- xgb.DMatrix(data = X_valid, label = y_valid)

params <- list(
  objective = "binary:logistic",
  eval_metric = "logloss",
  eta = 0.05,
  max_depth = 4,
  subsample = 0.8,
  colsample_bytree = 0.8
)

watch <- list(train = dtrain, valid = dvalid)

xgb_fit <- xgb.train(
  params = params,
  data = dtrain,
  nrounds = 500,
  watchlist = watch,
  early_stopping_rounds = 25,
  verbose = 0
)
```

```
## Warning in throw_err_or_depr_msg("Parameter '", match_old, "' has been renamed
## to '", : Parameter 'watchlist' has been renamed to 'evals'. This warning will
## become an error in a future version.
```

```r
p_valid <- predict(xgb_fit, dvalid)          # probabilities
pred_valid <- ifelse(p_valid > 0.5, 1, 0)    # class

table(Predicted = pred_valid, Actual = y_valid)
```

```
##          Actual
## Predicted  0  1
##         0 40  9
##         1  7 25
```

```r
mean(pred_valid == y_valid)
```

```
## [1] 0.8024691
```

```r
# random forest model
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.4.3
```

```
## randomForest 4.7-1.2

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
rf_model <- randomForest(as.factor(Heart.Disease) ~ ., data = train, ntree = 1000, mtry = 3, importance
rf_predict <- predict(rf_model, newdata = valid)

table(Predicted = rf_predict, Actual = valid$Heart.Disease)
```

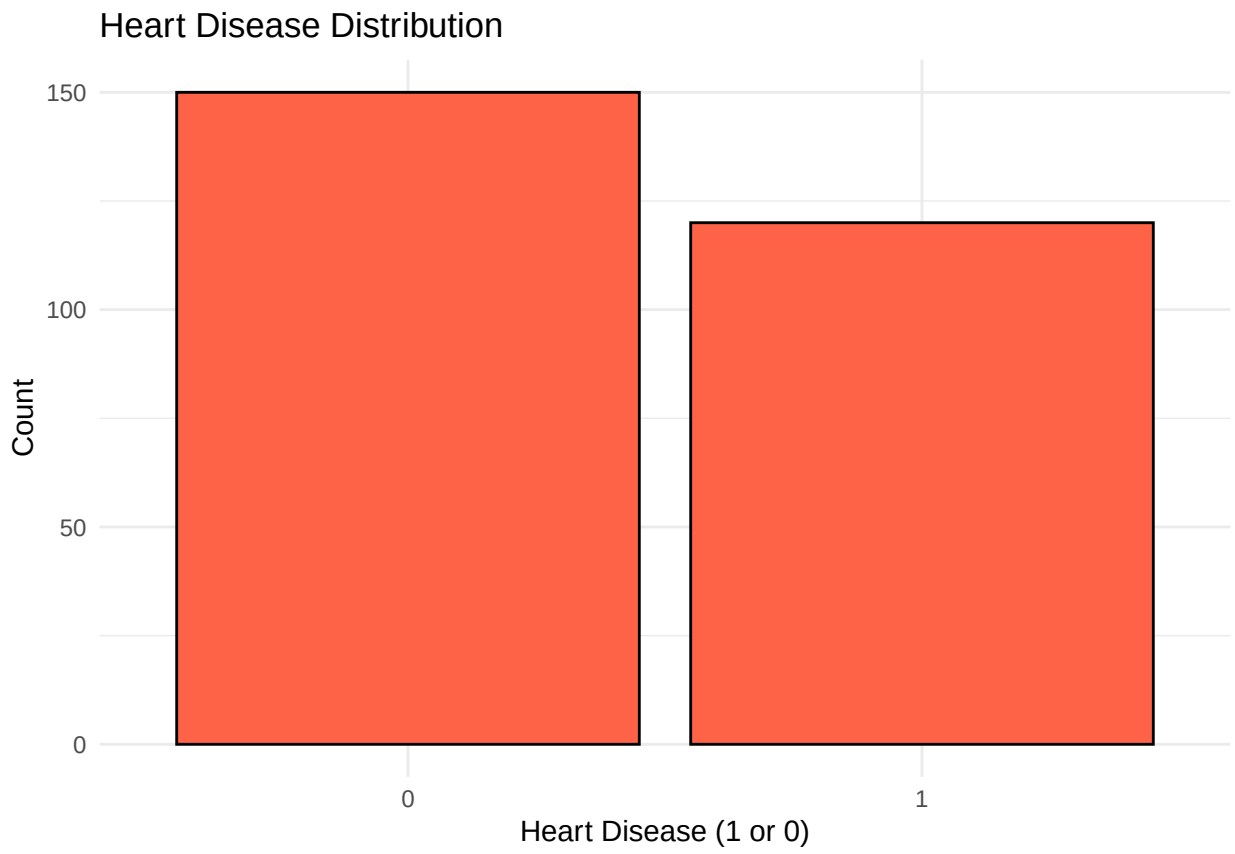
```
##          Actual
## Predicted  0  1
##         0 44  9
##         1  3 25
```

```
mean(rf_predict == valid$Heart.Disease)
```

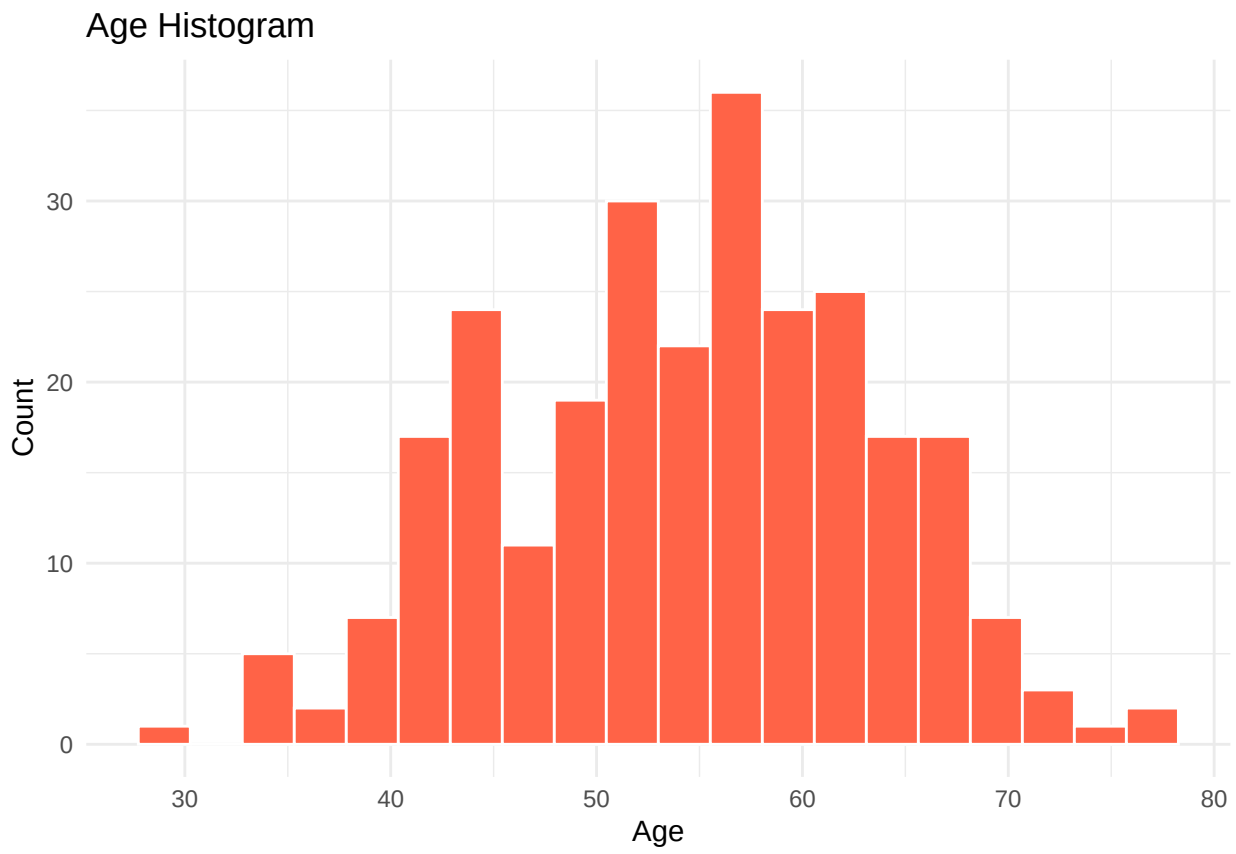
```
## [1] 0.8518519
```

## Visualizations

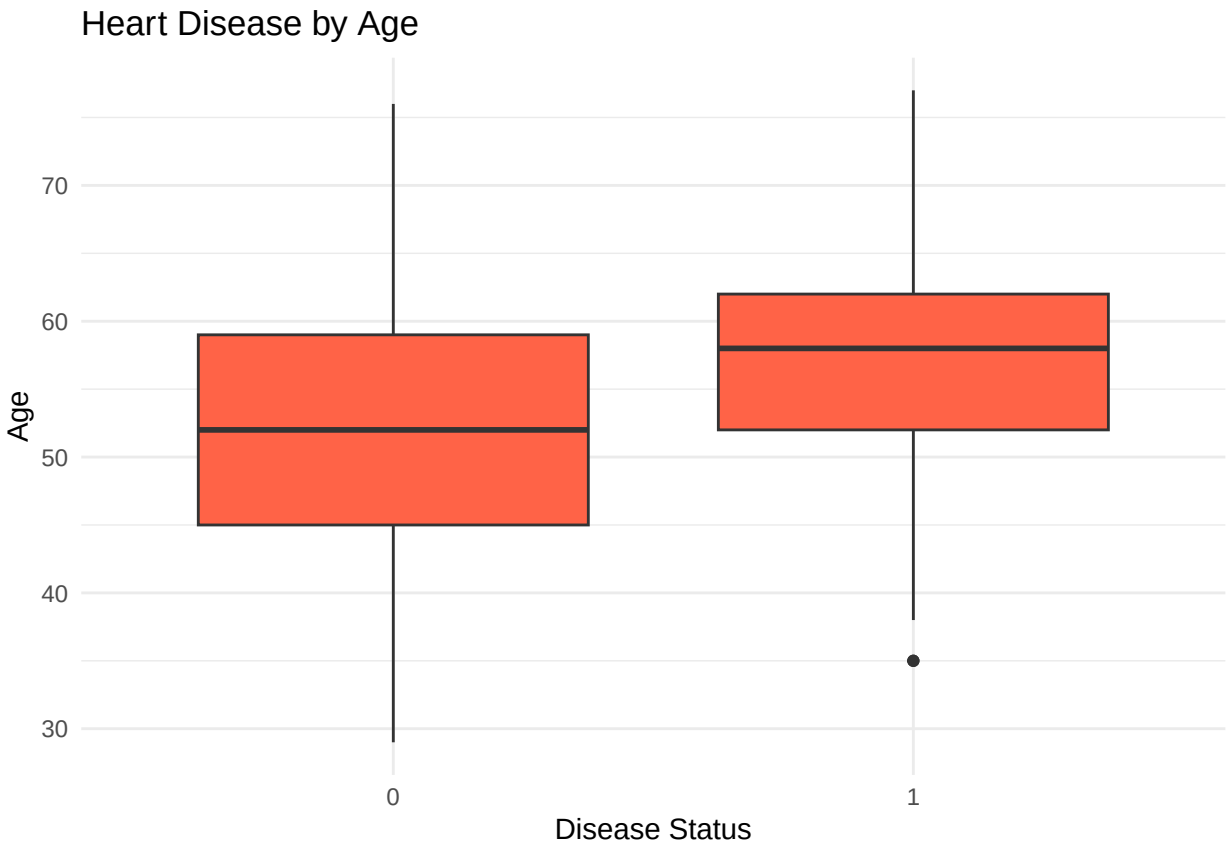
```
# heart disease boxplot
ggplot(data, aes(x = factor(Heart.Disease))) +
         geom_bar(fill = "tomato", col = "black") +
         labs(
           title = "Heart Disease Distribution",
           x = "Heart Disease (1 or 0)",
           y = "Count"
         ) +
         theme_minimal() +
         theme(legend.position = "none")
```



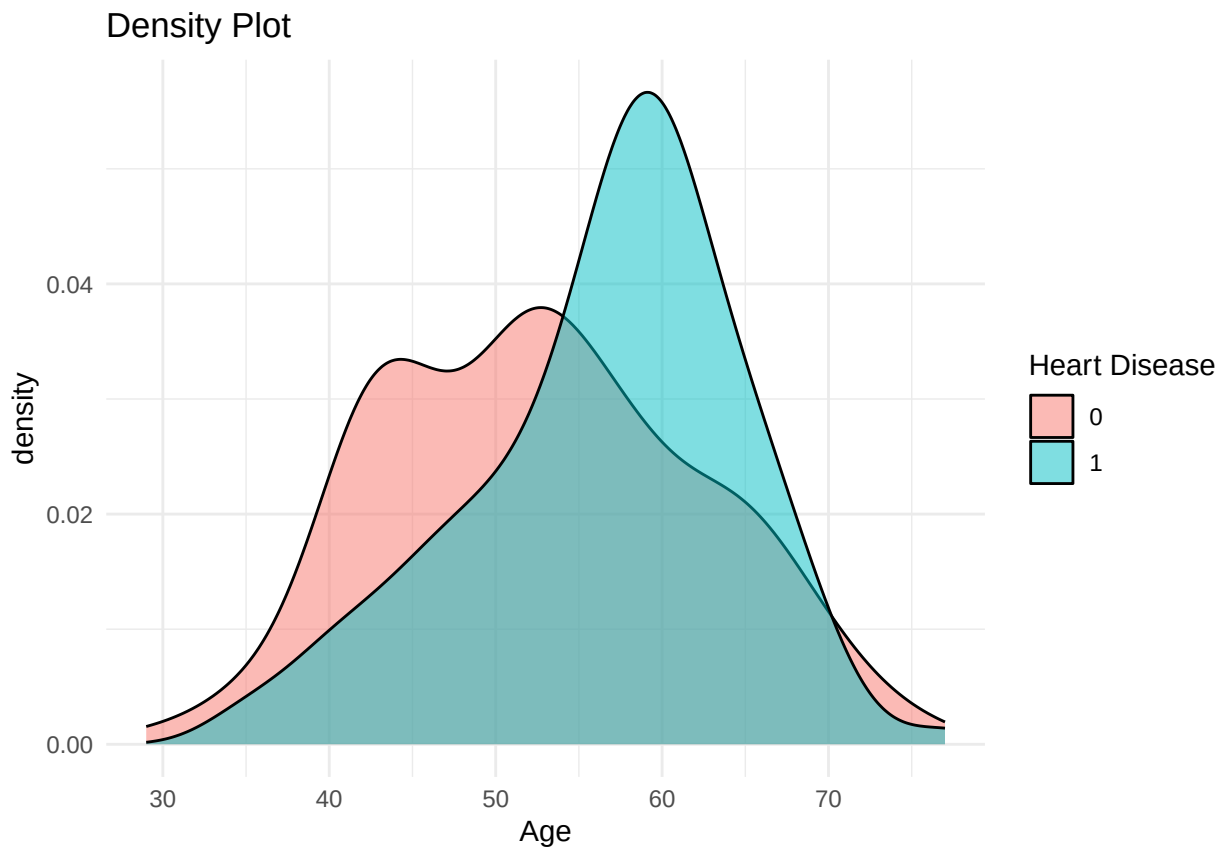

```
# age histogram
ggplot(data, aes(x = Age)) +
  geom_histogram(
    bins = 20,
    fill = "tomato",
    col = "white"
  ) +
  labs(
    title = "Age Histogram",
    x = "Age",
    y = "Count"
  ) +
  theme_minimal()
```



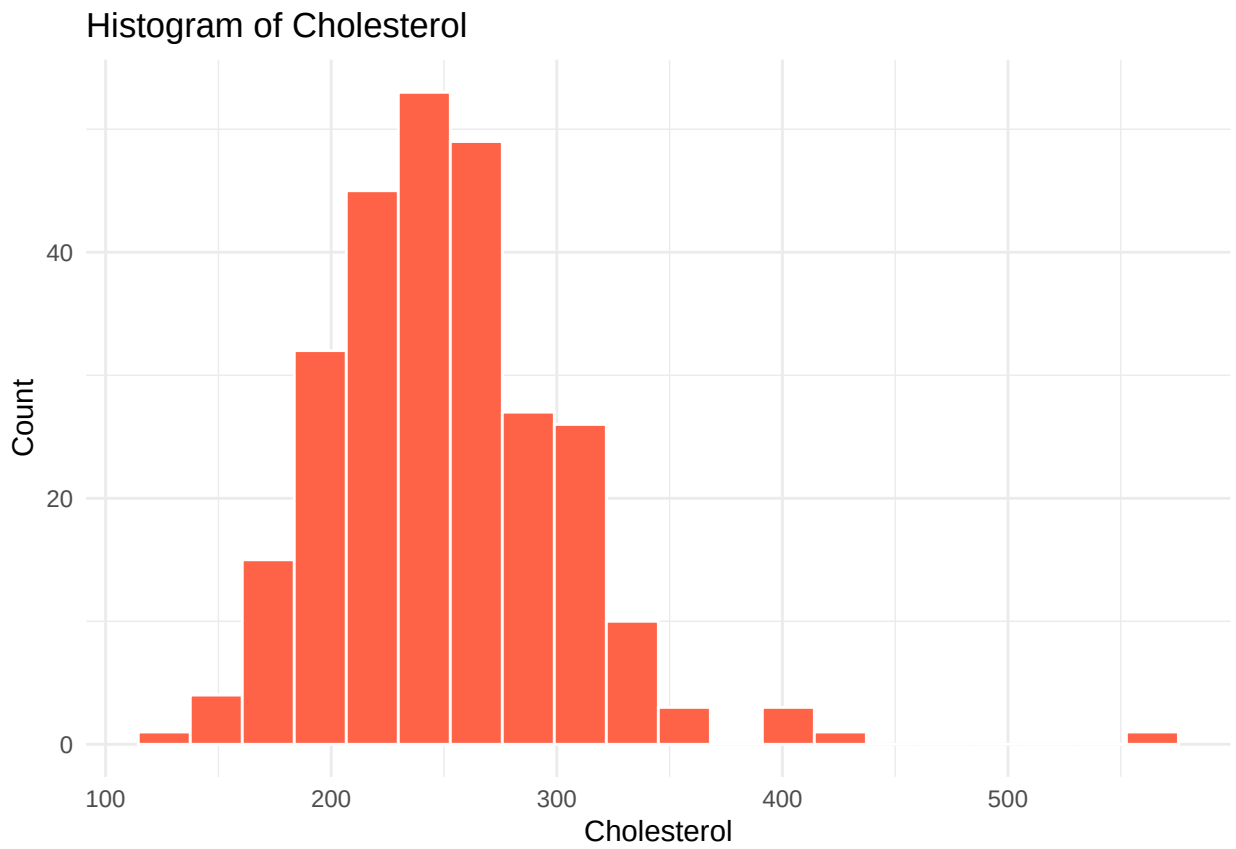

```
# boxplot of heart disease by age
ggplot(data, aes(x = factor(Heart.Disease), y = Age)) +
  geom_boxplot(
    fill = "tomato"
  ) +
  labs(
    title = "Heart Disease by Age",
    x = "Disease Status",
    y = "Age"
  ) +
  theme_minimal()
```



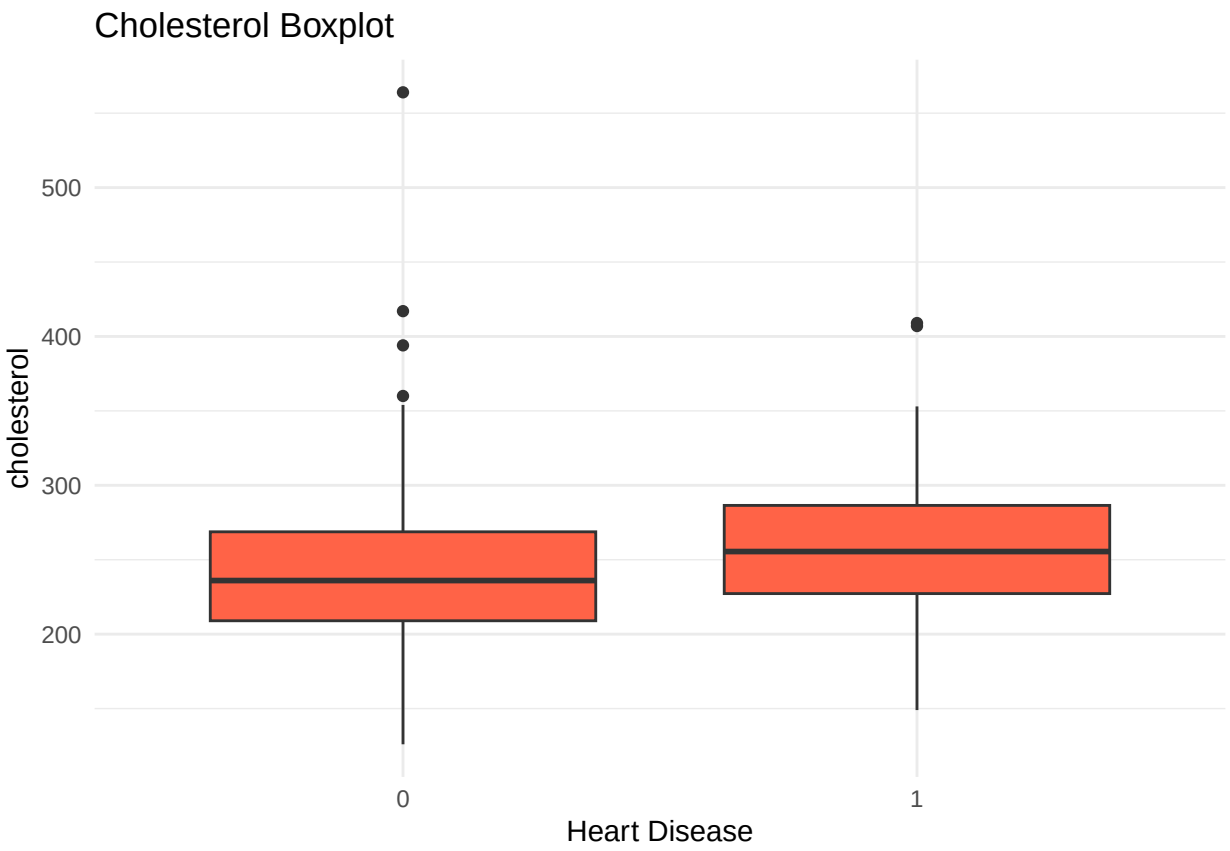

```
ggplot(data, aes(x = Age, fill = factor(Heart.Disease))) +
  geom_density(
    alpha = 0.5
    ) +
  labs(
    title = "Density Plot",
    x = "Age",
    fill = "Heart Disease"
  ) +
  theme_minimal()
```



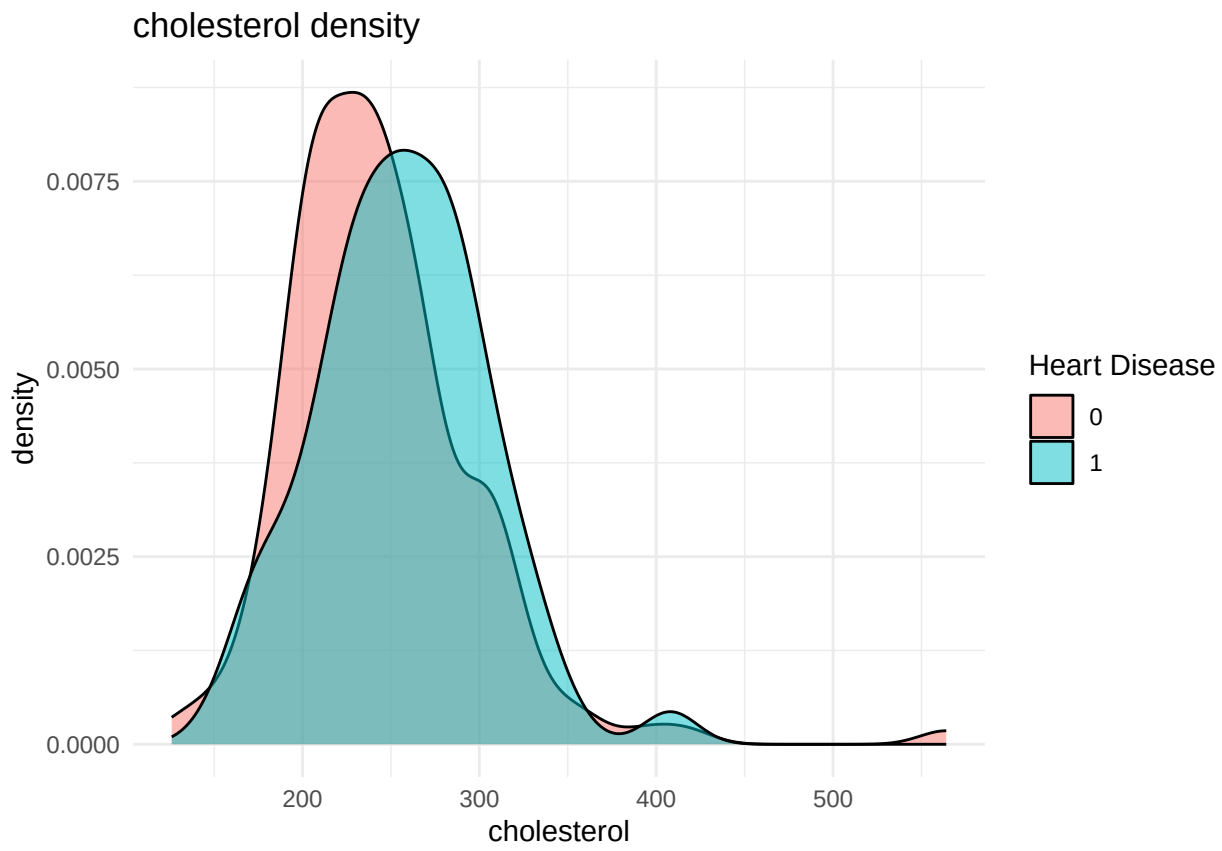

```r
# cholesterol histogram
ggplot(data, aes(x = Cholesterol)) +
  geom_histogram(
    bins = 20,
    fill = "tomato",
    col = "white"
  ) +
  labs(
    title = "Histogram of Cholesterol",
    x = "Cholesterol",
    y = "Count"
  ) +
  theme_minimal()
```



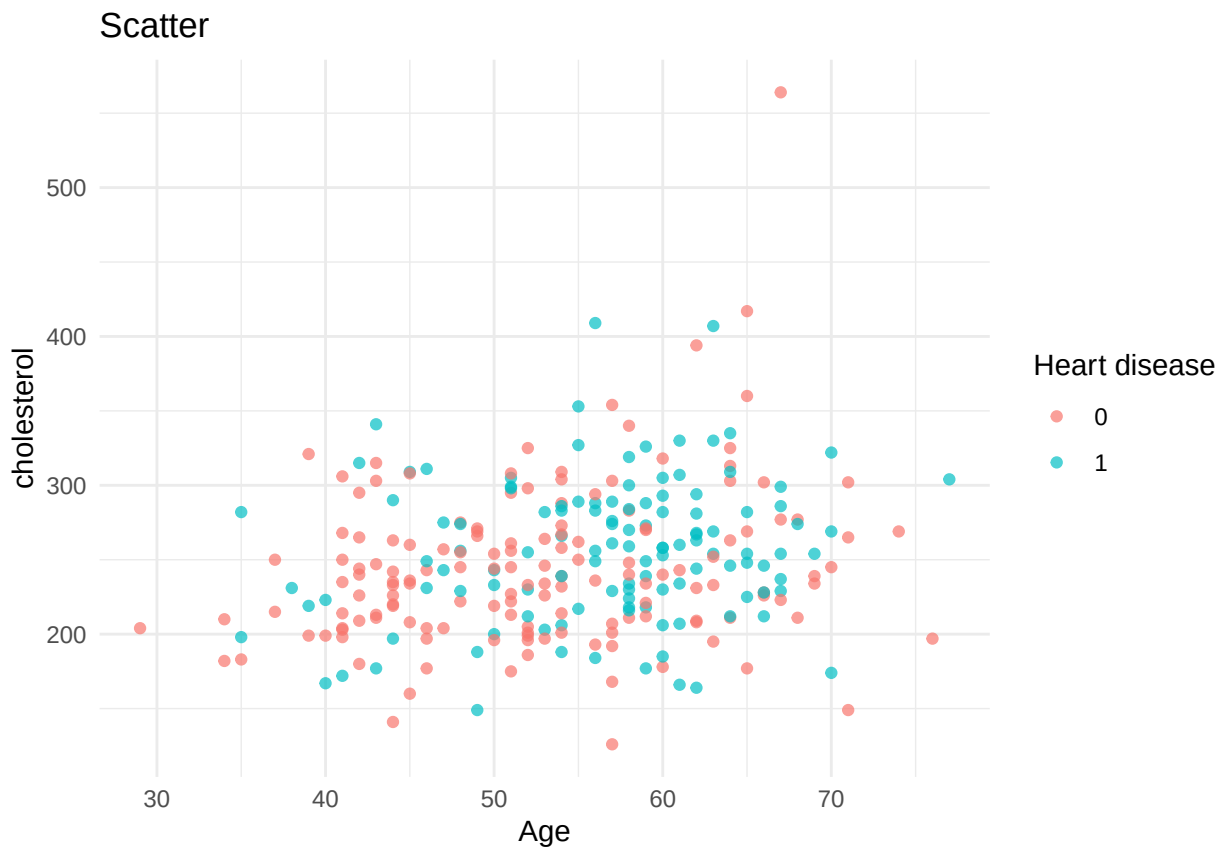

```
ggplot(data, aes(x = factor(Heart.Disease), y = Cholesterol)) +
  geom_boxplot(
    fill = "tomato"
  ) +
  labs(
    title = "Cholesterol Boxplot",
    x = "Heart Disease",
    y = "cholesterol"
  ) +
  theme_minimal()
```




```
# cholesterol density plot
ggplot(data, aes(x = Cholesterol, fill = factor(Heart.Disease))) +
  geom_density(
    alpha = .5
  ) +
  labs(
    title = "cholesterol density",
    x = "cholesterol",
    y = "density",
    fill = "Heart Disease"
  ) +
  theme_minimal()
```

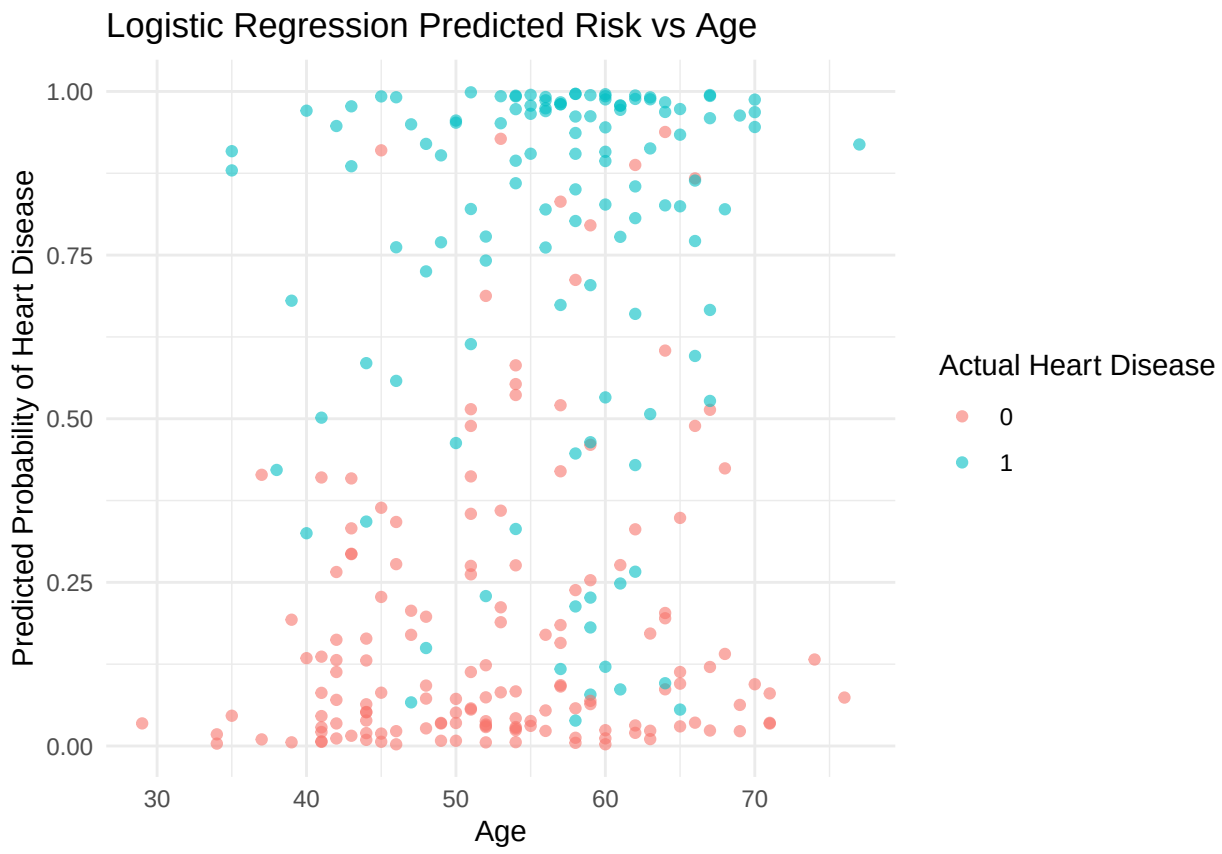



```r
# scatter plot
ggplot(data, aes(x = Age, y = Cholesterol, color = factor(Heart.Disease))) +
  geom_point(alpha = 0.7) +
  labs(
    title = "Scatter",
    x = "Age",
    y = "cholesterol",
    color = "Heart disease"
  ) +
  theme_minimal()
```




```r
# logistic variable
data$logit_prob <- predict(logit_model, newdata = df, type = "response")
```

```r
# logistic regression plot
ggplot(data, aes(x = Age, y = logit_prob, color = factor(Heart.Disease))) +
  geom_point(alpha = 0.6) +
  labs(
    title = "Logistic Regression Predicted Risk vs Age",
    x = "Age",
    y = "Predicted Probability of Heart Disease",
    color = "Actual Heart Disease"
  ) +
  theme_minimal()
```




```
library(dplyr)
multi_var <- data %>%
  dplyr::select(-Heart.Disease, -Sex)
```

```
multi_var <- scale(multi_var)
multi_var <- as.data.frame(multi_var)
head(multi_var)
```

```
##         Age Chest.pain.type          BP Cholesterol FBS.over.120 EKG.results
## 1 1.7089201       0.8693133 -0.07527007   1.3996132   -0.4162558   0.9798441
## 2 1.3795779      -0.1832185 -0.91506006   6.0817107   -0.4162558   0.9798441
## 3 0.2817705      -1.2357503 -0.41118607   0.2194151   -0.4162558  -1.0243824
## 4 1.0502357       0.8693133 -0.18724207   0.2581101   -0.4162558  -1.0243824
## 5 2.1480430      -1.2357503 -0.63513007   0.3741952   -0.4162558   0.9798441
## 6 1.1600164       0.8693133 -0.63513007  -1.4057758   -0.4162558  -1.0243824
##       Max.HR Exercise.angina ST.depression Slope.of.ST Number.of.vessels.fluro
## 1 -1.7559473      -0.6999225     1.1788233   0.6751655               2.4680989
## 2  0.4455818      -0.6999225     0.4802613   0.6751655              -0.7102161
## 3 -0.3745957      -0.6999225    -0.6549018  -0.9524656              -0.7102161
## 4 -1.9286162       1.4234380    -0.7422221   0.6751655               0.3492223
## 5 -1.2379404       1.4234380    -0.7422221  -0.9524656               0.3492223
## 6 -0.4177629      -0.6999225    -0.5675816  -0.9524656              -0.7102161
##     Thallium logit_prob
## 1 -0.8740826  1.4334779
## 2  1.1870729  0.1935783
```

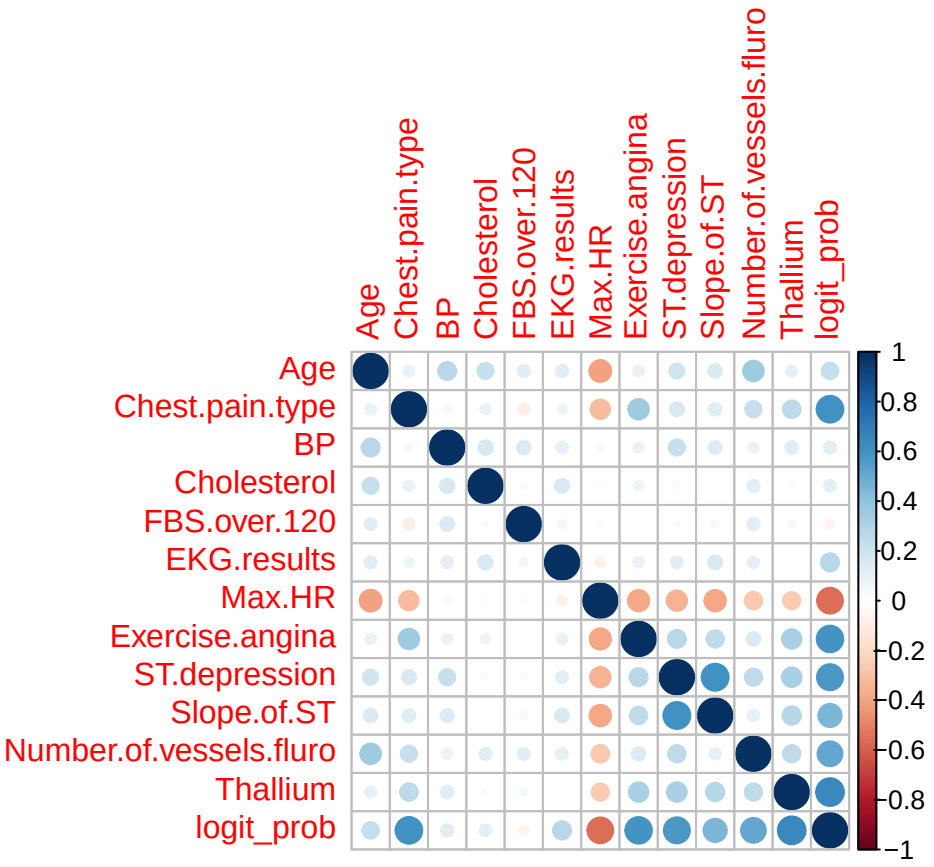
```
## 3  1.1870729 -0.8428601
## 4  1.1870729  1.3041743
## 5 -0.8740826 -0.8047743
## 6  1.1870729 -0.2386462
```

```
cor <- cor(multi_var)

round(cor,2)
```

```
##                            Age Chest.pain.type    BP Cholesterol FBS.over.120
## Age                       1.00            0.10  0.27        0.22         0.12
## Chest.pain.type           0.10            1.00 -0.04        0.09        -0.10
## BP                        0.27           -0.04  1.00        0.17         0.16
## Cholesterol               0.22            0.09  0.17        1.00         0.03
## FBS.over.120              0.12           -0.10  0.16        0.03         1.00
## EKG.results               0.13            0.07  0.12        0.17         0.05
## Max.HR                   -0.40           -0.32 -0.04       -0.02         0.02
## Exercise.angina           0.10            0.35  0.08        0.08         0.00
## ST.depression             0.19            0.17  0.22        0.03        -0.03
## Slope.of.ST               0.16            0.14  0.14       -0.01         0.04
## Number.of.vessels.fluro   0.36            0.23  0.09        0.13         0.12
## Thallium                  0.11            0.26  0.13        0.03         0.05
## logit_prob                0.23            0.60  0.13        0.12        -0.07
##                          EKG.results Max.HR Exercise.angina ST.depression
## Age                             0.13  -0.40            0.10          0.19
## Chest.pain.type                 0.07  -0.32            0.35          0.17
## BP                              0.12  -0.04            0.08          0.22
## Cholesterol                     0.17  -0.02            0.08          0.03
## FBS.over.120                    0.05   0.02            0.00         -0.03
## EKG.results                     1.00  -0.07            0.10          0.12
## Max.HR                         -0.07   1.00           -0.38         -0.35
## Exercise.angina                 0.10  -0.38            1.00          0.27
## ST.depression                   0.12  -0.35            0.27          1.00
## Slope.of.ST                     0.16  -0.39            0.26          0.61
## Number.of.vessels.fluro         0.11  -0.27            0.15          0.26
## Thallium                        0.01  -0.25            0.32          0.32
## logit_prob                      0.28  -0.56            0.59          0.57
##                          Slope.of.ST Number.of.vessels.fluro Thallium logit_prob
## Age                             0.16                   0.36     0.11       0.23
## Chest.pain.type                 0.14                   0.23     0.26       0.60
## BP                              0.14                   0.09     0.13       0.13
## Cholesterol                    -0.01                   0.13     0.03       0.12
## FBS.over.120                    0.04                   0.12     0.05      -0.07
## EKG.results                     0.16                   0.11     0.01       0.28
## Max.HR                         -0.39                  -0.27    -0.25      -0.56
## Exercise.angina                 0.26                   0.15     0.32       0.59
## ST.depression                   0.61                   0.26     0.32       0.57
## Slope.of.ST                     1.00                   0.11     0.28       0.45
## Number.of.vessels.fluro         0.11                   1.00     0.26       0.52
## Thallium                        0.28                   0.26     1.00       0.65
## logit_prob                      0.45                   0.52     0.65       1.00
```

```
library(corrplot)
corrplot(cor)
```




```
table(data$Heart.Disease)
```

```
##
##   0   1
## 150 120
```

```
pca <- prcomp(multi_var, .scale = TRUE)
```

```
## Warning: In prcomp.default(multi_var, .scale = TRUE) :
##  extra argument '.scale' will be disregarded
```

```
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.955 1.2312 1.10520 1.02386 0.98524 0.94899 0.89983
## Proportion of Variance 0.294 0.1166 0.09396 0.08064 0.07467 0.06928 0.06228
## Cumulative Proportion  0.294 0.4106 0.50457 0.58521 0.65988 0.72915 0.79144
##                           PC8    PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.8440 0.7783 0.75807 0.63644 0.58758 0.26123
## Proportion of Variance 0.0548 0.0466 0.04421 0.03116 0.02656 0.00525
## Cumulative Proportion  0.8462 0.8928 0.93703 0.96819 0.99475 1.00000
```

```r
print(pca$rotation[,1:2])
```

```
##                                PC1          PC2
## Age                     -0.22266906  0.451621477
## Chest.pain.type         -0.27712033 -0.285704464
## BP                      -0.13063745  0.478499061
## Cholesterol             -0.09138260  0.372059767
## FBS.over.120            -0.01533569  0.404902194
## EKG.results             -0.13923185  0.243921666
## Max.HR                   0.34404285  0.062813586
## Exercise.angina         -0.31178970 -0.201893459
## ST.depression           -0.34526262 -0.001049262
## Slope.of.ST             -0.30912082 -0.020069828
## Number.of.vessels.fluro -0.27246869  0.186413384
## Thallium                -0.31298500 -0.139276852
## logit_prob              -0.47544309 -0.155946169
```

```r
# kmeans cluster
km <- kmeans(multi_var, centers = 2, nstart = 25)
table(km$cluster)
```

```
##
##   1   2
## 106 164
```

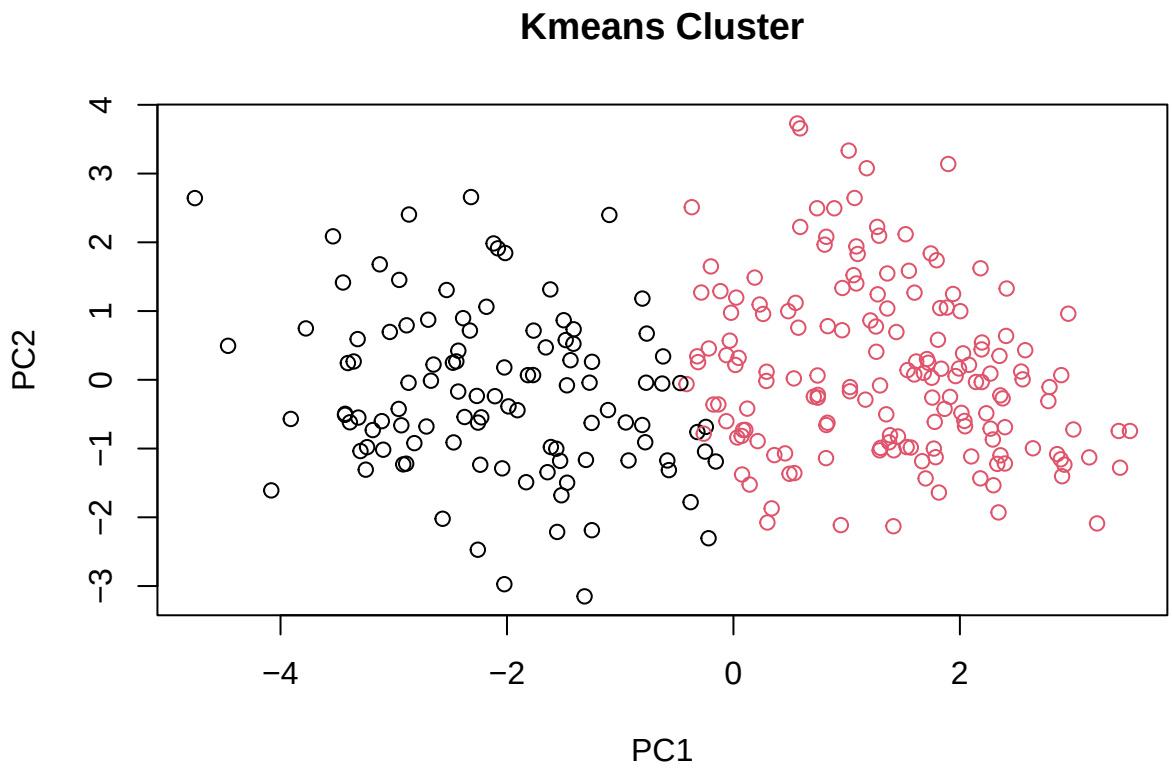
```r
# pca scores
pc_scores <- as.data.frame(pca$x)
head(pc_scores)
```

```
##           PC1         PC2         PC3         PC4        PC5        PC6
## 1 -2.95158308  1.45310956 -1.32439727  0.24811734 2.21800588  0.7973748
## 2 -1.09573504  2.39809077 -2.05371634  2.91448692 0.01041309 -1.6438160
## 3  1.29498512 -0.08191726 -0.03470948 -0.84284742 0.30437297 -1.0900887
## 4 -2.47141003 -0.91094329 -0.92604601 -1.06908881 0.63953279 -0.8828207
## 5  0.02490136  1.19603083 -1.29124080  0.02837445 1.87348889  0.5838256
## 6  0.36157154 -1.09556173 -0.43912665 -1.59544396 0.66991104 -0.4626737
##           PC7        PC8        PC9       PC10       PC11       PC12       PC13
## 1 -1.13304147  0.5939057  0.7905200  0.3466217 -0.6594097  0.1278234 -0.2758876
## 2 -0.61549719  4.1059172 -1.2020589 -1.6565643  0.5849665  0.2921236  0.0299745
## 3 -0.19220048  0.2867704 -1.6002607 -0.9825213 -0.2742787  0.4194717  0.2525976
## 4  1.11894013  0.1008178 -0.8172562 -0.4823359 -0.2682254 -0.8630407 -0.2476966
## 5  1.33487139 -0.4835677 -2.0079685  0.8752423  0.5628900  0.1881792  0.5023395
## 6  0.05320517 -0.9311517 -0.2155345 -1.7623136  0.7438717  0.4735880  0.1897508
```

```r
# selecting first two principle compenents
pc_12 <- pc_scores %>%
  dplyr::select(PC1, PC2)
```

```r
# adding to data
multi_var <- multi_var %>%
  bind_cols(pc_12)
```

```r
# kmeans clustering
plot(pca$x[,1], pca$x[,2], col = km$cluster, main = "Kmeans Cluster", xlab = "PC1", ylab = "PC2")
```




```r
table(data$Sex)
```

```
##
##   0   1
##  87 183
```

```r
gender_summary <- data %>%
  group_by(Sex) %>%
  summarise(
    proportion = round(mean(Heart.Disease == 1),2)
  )
gender_summary
```

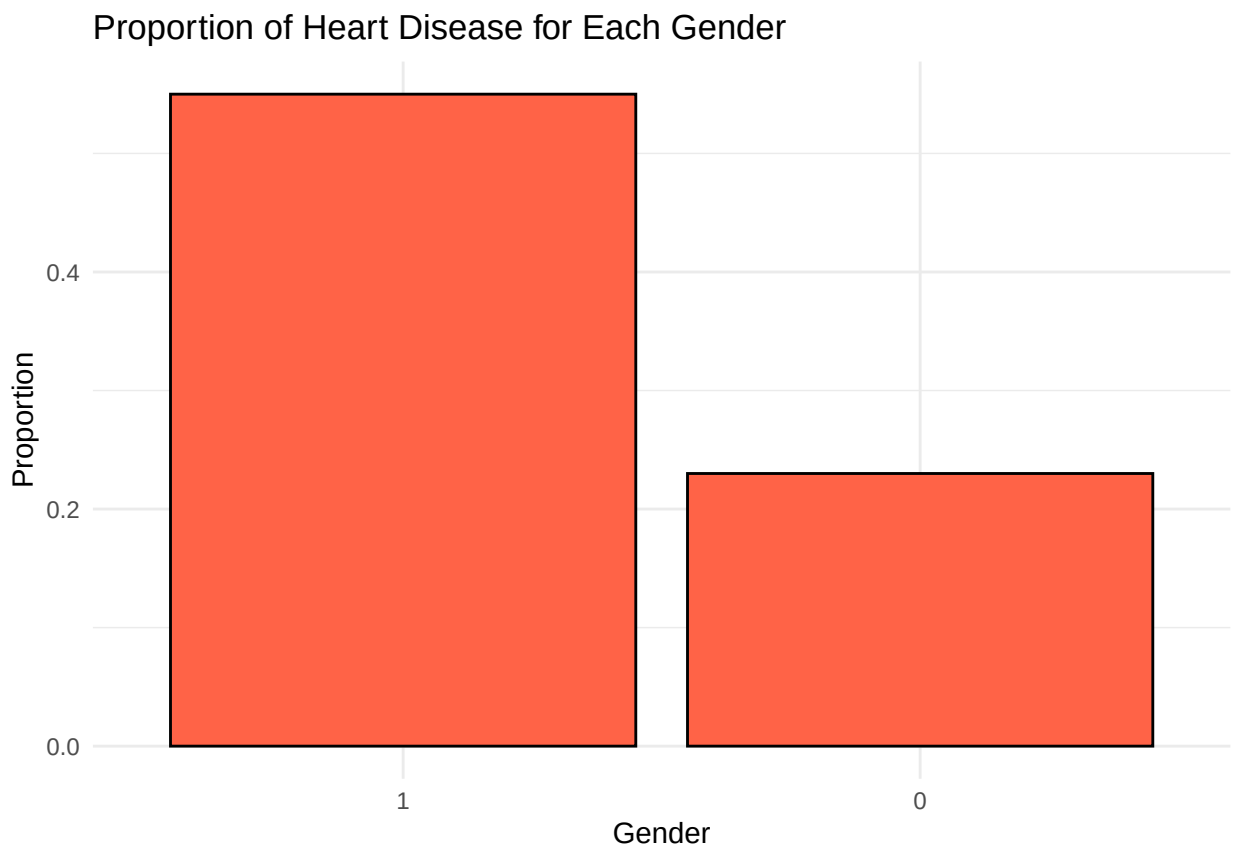
```
## # A tibble: 2 x 2
##     Sex proportion
##   <int>      <dbl>
## 1     0       0.23
## 2     1       0.55
```

```r
kable(gender_summary, caption = "Heart Disease by Gender") %>%
  kable_styling()
```

Table 1: Heart Disease by Gender

| Sex | proportion |
|---|---|
| 0 | 0.23 |
| 1 | 0.55 |

```r
ggplot(gender_summary, aes(x = reorder(Sex, -proportion), y = proportion, fill = sex)) +
  geom_col(fill = "tomato", col = "black") +
  labs(
    title = "Proportion of Heart Disease for Each Gender",
    x = "Gender",
    y = "Proportion"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```




```r
data <- data %>%
  mutate(HR_Cat = factor(ntile(Max.HR, 3), labels = c("Low","Med","High")))

ca_table <- table(data$Heart.Disease, data$HR_Cat)
ca_table
```

```
##
##      Low Med High
##   0   26  53   71
##   1   64  37   19
```

```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.4.3
```

```
ca <- CA(ca_table, graph = FALSE)
mosaicplot(ca_table, main="Heart Disease vs HR Category", color=TRUE)
```