Heart diseases feature selection and exploratory data analysis

Wangjun Shen

# Introduction

The objective of this assignment is to prepare a data set that can be used to predict heart disease, a common and serious health issue that affects a significant portion of the population. The data set we will be working on is a subset of a larger real-world data set collected by multiple healthcare institutions. It contains various attributes related to patients’ health, such as age, gender, blood pressure, and other clinical measurements. These attributes can be leveraged to predict the presence of heart disease.

Data mining plays a critical role in addressing challenges related to predicting disease spread and similar healthcare problems. By analyzing large and complex data sets, we can identify patterns and relationships that may not be immediately apparent. This, in turn, allows us to develop more accurate predictive models. In the context of heart disease, data mining techniques can help us extract insights and patterns from patient data. These insights can aid in identifying risk factors, predicting disease outcomes, and developing effective treatment strategies.

For example, by using data mining techniques on the given data set, we can identify the most important predictors of heart disease. This knowledge can then be used to develop a classification model that accurately predicts the presence of the disease in patients. Moreover, data mining can also help us identify subpopulations that are more susceptible to the disease. We can then develop tailored prevention and treatment strategies for these subpopulations.

Overall, this assignment provides an opportunity to apply data mining techniques to a real-world data set. It also allows us to gain hands-on experience with feature engineering, data exploration, and predictive modeling. The insights gained from this exploration can inform our work in subsequent assignments. This, in turn, can help us develop more accurate and effective predictive models for heart disease and other similar health challenges.

# Related Work

# Data exploration

## Features Selection

The details of the original data set are as follows.

| Variable | Description | Type |
| --- | --- | --- |
| id | A unique ID that identifies a participant in the study | Numerical |
| age | Age in years | Numerical |
| sex | Male and Female were recorded | Categorical |
| cp | Chest Pain type: typical angina; atypical angina; non-anginal pain; and asymptomatic | Categorical |
| trestbps | Resting blood pressure (in mm Hg on admission to the hospital) | Numerical |
| chol | Serum Cholestoral in mg/dl | Numerical |
| fbs | Fasting blood sugar > 120 mg/dl (True or False) | Boolean |
| restecg | Resting electrocardiographic results: normal; having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) or showing probable or definite left ventricular hypertrophy by Estes’ criteria | Categorical |
| thalach | Maximum heart rate achieved | Numerical |
| exang | Exercise induced angina (True/False) | Boolean |
| oldpeak | ST depression induced by exercise relative to rest | Numerical |
| slope | The slope of the peak exercise ST segment: upsloping; flat; downsloping | Categorical |
| major\_vessels | Number of major vessels (0-3) colored by flourosopy | Numerical |
| restwm | Rest wall motion abnormality: none; mild or moderate; moderate or severe; akinesis or dyskmem | Categorical |
| target | Heart disease diagnosed (disease/no disease) | Categorical |

Import the data set and view.

# load data set first  
heart.full <- read.csv("heart.csv")  
  
# then check the data set  
str(heart.full)

## 'data.frame': 1025 obs. of 15 variables:  
## $ id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ age : int 52 53 70 61 62 58 58 55 46 54 ...  
## $ sex : chr "male" "male" "male" "male" ...  
## $ cp : chr "typical angina" "typical angina" "typical angina" "typical angina" ...  
## $ trestbps : int 125 140 145 148 138 100 114 160 120 122 ...  
## $ chol : int 212 203 174 203 294 248 318 289 249 286 ...  
## $ fbs : logi FALSE TRUE FALSE FALSE TRUE FALSE ...  
## $ restecg : chr "ST-T wave abnormality" "normal" "ST-T wave abnormality" "ST-T wave abnormality" ...  
## $ thalach : int 168 155 125 161 106 122 140 145 144 116 ...  
## $ exang : logi FALSE TRUE TRUE FALSE FALSE FALSE ...  
## $ oldpeak : num 1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...  
## $ slope : chr "downsloping" "upsloping" "upsloping" "downsloping" ...  
## $ major\_vessels: int 2 0 0 1 3 0 3 1 0 2 ...  
## $ restwm : chr "akinesis or dyskmem" "akinesis or dyskmem" "akinesis or dyskmem" "akinesis or dyskmem" ...  
## $ target : chr "no disease" "no disease" "no disease" "no disease" ...

From the output results, it can be found that the data set contains 15 variables and 1025 observation. The details of each variable are consistent with the description.

Then check the distribution of missing values in the data set.

sum(is.na(heart.full))

## [1] 0

There are not missing values in this data.

In the following section, we will discuss the importance of features, specifically the correlation between variables. To measure this correlation, we will use random forest to calculate Gini Importance or Mean Decrease in Impurity (MDI). This method determines the importance of each feature by summing the number of splits (across all trees) that include the feature, proportionally to the number of samples it splits.

The first is the correlation coefficient between the variables. The data set is undergoing a transformation from its original data types to appropriate data types required for analysis. Originally, the variables age, trestbps, chol, thalach, and oldpeak were imported as character vectors representing age, resting blood pressure, serum cholesterol, maximum heart rate achieved, and ST depression induced by exercise relative to rest, respectively. These variables have been converted to a numeric data type since they represent numerical measurements. Similarly, the variables sex, cp, restecg, slope, restwm, and target were originally imported as character vectors representing sex, chest pain type, resting electrocardiographic results, slope of the peak exercise ST segment, presence of a major vessels colored by fluoroscopy, and heart disease status, respectively. These variables have been converted to factor data type since they represent categorical variables. This transformation allows for easier data manipulation and analysis, especially when exploring relationships between variables.

# convert age, trestbps, chol, thalach, and oldpeak to numeric  
heart.full$age <- as.numeric(heart.full$age)  
heart.full$trestbps <- as.numeric(heart.full$trestbps)  
heart.full$chol <- as.numeric(heart.full$chol)  
heart.full$thalach <- as.numeric(heart.full$thalach)  
heart.full$oldpeak <- as.numeric(heart.full$oldpeak)  
  
  
#   
# heart.full$sex <- as.numeric(heart.full$sex)  
# heart.full$cp <- as.numeric(heart.full$cp)  
# heart.full$restecg <- as.numeric(heart.full$restecg)  
# heart.full$slope <- as.numeric(heart.full$slope)  
# heart.full$restwm <- as.numeric(heart.full$restwm)  
# heart.full$target <- as.numeric(heart.full$target)  
  
#   
# convert sex, cp, restecg, slope, restwm, and target to factor  
heart.full$sex <- as.factor(heart.full$sex)  
heart.full$cp <- as.factor(heart.full$cp)  
heart.full$restecg <- as.factor(heart.full$restecg)  
heart.full$slope <- as.factor(heart.full$slope)  
heart.full$restwm <- as.factor(heart.full$restwm)  
heart.full$target <- as.factor(heart.full$target)

Next we calculate the correlation between the variables.

# calculate the correlation matrix  
# if convert to factor  
cor.matrix <- cor(heart.full[, c("age", "trestbps", "chol", "thalach", "oldpeak", "major\_vessels")])  
cor.matrix

## age trestbps chol thalach oldpeak  
## age 1.0000000 0.27112141 0.21982253 -0.39022708 0.20813668  
## trestbps 0.2711214 1.00000000 0.12797743 -0.03926407 0.18743411  
## chol 0.2198225 0.12797743 1.00000000 -0.02177209 0.06488031  
## thalach -0.3902271 -0.03926407 -0.02177209 1.00000000 -0.34979616  
## oldpeak 0.2081367 0.18743411 0.06488031 -0.34979616 1.00000000  
## major\_vessels 0.2715505 0.10455372 0.07425934 -0.20788842 0.22181603  
## major\_vessels  
## age 0.27155053  
## trestbps 0.10455372  
## chol 0.07425934  
## thalach -0.20788842  
## oldpeak 0.22181603  
## major\_vessels 1.00000000

# if convert to numeric  
# cor.matrix.numeric <- cor(heart.full)  
# cor.matrix.numeric

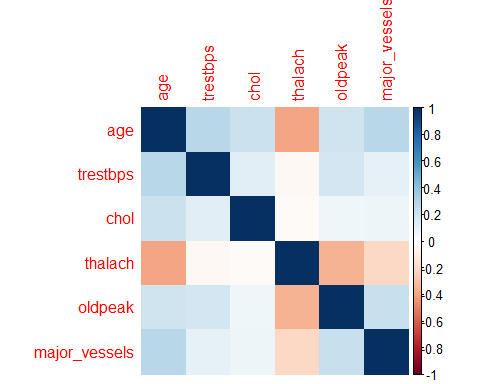
For ease of observation, we visualize the correlation.

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.2.3

## corrplot 0.92 loaded

corrplot(cor.matrix, method = "color")



Visualization of correlation matrix

The resulting plot will show the correlations between variables using different colors, where red indicates strong negative correlation, blue indicates strong positive correlation, and white indicates no correlation. The strength of the correlation is determined by the absolute value of the correlation coefficient, with values closer to 1 indicating stronger correlation.

Then use random forest to detect the importance of each feature.In order to understand whether these features differ between genders, the data will firstly be created based on the sub-dataset.

# data set for each gender  
heart.male <- subset(heart.full, sex == "male")  
heart.female <- subset(heart.full, sex == "female")

Then use those two sub datasets to build the random forest:

# Random forest for male subset  
library(randomForest)

## Warning: package 'randomForest' was built under R version 4.2.2

## randomForest 4.7-1.1

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

set.seed(123)  
rf\_model\_male <- randomForest(target ~ ., data = heart.male, importance = TRUE, ntree = 500)  
importance(rf\_model\_male)

## disease no disease MeanDecreaseAccuracy MeanDecreaseGini  
## id 1.606825 3.213858 3.376221 11.154653  
## age 33.928469 40.857084 42.838250 34.448119  
## sex 0.000000 0.000000 0.000000 0.000000  
## cp 38.745483 36.059468 41.121602 47.161819  
## trestbps 33.903759 37.024285 41.751265 25.962476  
## chol 34.122660 35.756098 39.121768 29.972875  
## fbs 16.795603 16.203869 19.636756 3.761825  
## restecg 19.083360 19.416802 21.509917 5.531215  
## thalach 37.581333 37.899592 43.252118 51.572260  
## exang 17.532815 19.401905 20.468760 11.933409  
## oldpeak 32.589020 38.902000 40.817055 37.108368  
## slope 24.318522 24.566740 27.686364 17.111266  
## major\_vessels 37.680805 40.115231 43.835239 42.184090  
## restwm 31.201322 32.398437 34.984151 25.367047

Based on the provided correlation matrix, the following features appear to be the most important in predicting the presence of heart disease:

* cp (chest pain type)
* thalach (maximum heart rate achieved)
* major\_vessels (number of major vessels colored by fluoroscopy)
* oldpeak (ST depression induced by exercise relative to rest)
* trestbps (resting blood pressure)
* age

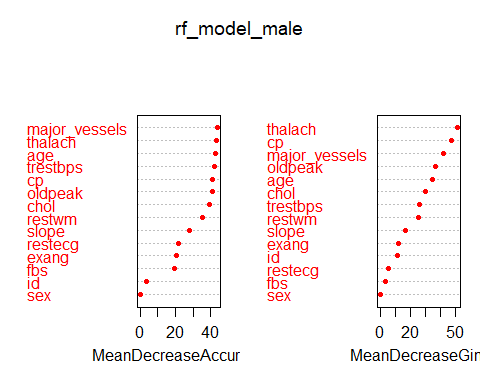
These features have the highest correlation with the target variable (disease/no disease), as well as high values for MeanDecreaseAccuracy and MeanDecreaseGini, indicating that they are important predictors for a machine learning model.

Chest pain type (cp) has the highest correlation with the target variable and high values for both MeanDecreaseAccuracy and MeanDecreaseGini, suggesting that it is a strong predictor for heart disease. Maximum heart rate achieved (thalach) and the number of major vessels colored by fluoroscopy (major\_vessels) also have high correlations and values for MeanDecreaseAccuracy and MeanDecreaseGini, making them strong predictors as well. ST depression induced by exercise relative to rest (oldpeak) and resting blood pressure (trestbps) have lower correlations but still have high values for MeanDecreaseAccuracy and MeanDecreaseGini, indicating that they are important features in predicting the presence of heart disease. Finally, age is also an important feature as it has a moderate correlation with the target variable and a relatively high value for MeanDecreaseAccuracy.

Overall, these six features can be considered the most important for predicting the presence of heart disease for male in this dataset.

Visualization.

varImpPlot(rf\_model\_male, col = "red", pch = 20)



# Random forest for female subset  
set.seed(123)  
rf\_model\_female <- randomForest(target ~ ., data = heart.female, importance = TRUE, ntree = 500)  
importance(rf\_model\_female)

## disease no disease MeanDecreaseAccuracy MeanDecreaseGini  
## id 1.932703 -3.148355 -0.5202009 2.590325  
## age 23.376089 24.724758 28.2364781 12.018921  
## sex 0.000000 0.000000 0.0000000 0.000000  
## cp 19.702221 20.271296 22.9715941 12.537614  
## trestbps 18.409296 18.651176 22.1370717 8.429236  
## chol 19.941082 22.898914 25.6314494 8.321454  
## fbs 7.611649 9.737944 10.2944996 1.418907  
## restecg 14.299858 15.552097 17.3754050 2.982791  
## thalach 19.430050 20.969639 24.2566433 8.272732  
## exang 18.303884 21.574578 22.6248431 9.167964  
## oldpeak 20.677704 21.784792 24.3639560 16.486048  
## slope 17.218824 19.973871 21.5111294 8.150365  
## major\_vessels 19.573486 20.510840 22.7582749 10.407631  
## restwm 22.156482 24.907233 26.5256308 22.290165

Based on the provided correlation matrix, the important features that can be selected are:

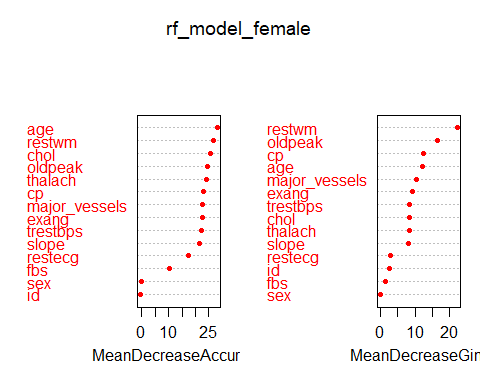
* Age: It has a high correlation with both disease and no disease, and also has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
* cp (Chest pain type): It has a moderate correlation with disease and no disease, and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
* thalach (Maximum heart rate achieved): It has a moderate correlation with disease and no disease, and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
* oldpeak (ST depression induced by exercise relative to rest): It has a moderate correlation with disease and no disease, and has a high value for MeanDecreaseAccuracy.
* major\_vessels (Number of major vessels (0-3) colored by flourosopy): It has a moderate correlation with disease and no disease, and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
* restwm (resting wall motion abnormalities): It has a moderate correlation with disease and no disease, and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
* exang (exercise-induced angina): It has a moderate correlation with both disease and no disease, and has a high value for MeanDecreaseGini.

The features “trestbps”, “chol”, “fbs”, and “restecg” have moderate correlations with disease and no disease, but have relatively lower values for MeanDecreaseAccuracy and MeanDecreaseGini, so they are not included in the list of important features.

Overall, these seven features can be considered the most important for predicting the presence of heart disease for female in this dataset.

Visualization.

varImpPlot(rf\_model\_female, col = "red", pch = 20)



Therefore, comprehensive consideration, the features that will be retained are: age, cp, thalach, oldpeak, major\_vessels, restwm, exang and sex.

Now create a new data set with those selected features.

# Create new dataset with selected features and target variable  
heart\_features\_selected <- heart.full[, c("id", "age", "sex", "cp", "thalach", "exang", "oldpeak", "major\_vessels", "restwm", "target")]  
  
# Save new dataset as CSV file  
write.csv(heart\_features\_selected, "heart\_features\_selected.csv", row.names = FALSE)

## Descriptive Statistics

Import the new data set and check the details of that new data set:

heart.selected <- read.csv("heart\_features\_selected.csv")  
  
str(heart.selected)

## 'data.frame': 1025 obs. of 10 variables:  
## $ id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ age : int 52 53 70 61 62 58 58 55 46 54 ...  
## $ sex : chr "male" "male" "male" "male" ...  
## $ cp : chr "typical angina" "typical angina" "typical angina" "typical angina" ...  
## $ thalach : int 168 155 125 161 106 122 140 145 144 116 ...  
## $ exang : logi FALSE TRUE TRUE FALSE FALSE FALSE ...  
## $ oldpeak : num 1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...  
## $ major\_vessels: int 2 0 0 1 3 0 3 1 0 2 ...  
## $ restwm : chr "akinesis or dyskmem" "akinesis or dyskmem" "akinesis or dyskmem" "akinesis or dyskmem" ...  
## $ target : chr "no disease" "no disease" "no disease" "no disease" ...

Summarize the selected data set:

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.2

## Warning: package 'readr' was built under R version 4.2.2

## Warning: package 'purrr' was built under R version 4.2.2

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.2

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.1 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::combine() masks randomForest::combine()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ ggplot2::margin() masks randomForest::margin()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

heart.selected$id <- as.numeric(heart.selected$id)  
heart.selected$age <- as.numeric(heart.selected$age)  
heart.selected$sex <- as.factor(heart.selected$sex)  
heart.selected$cp <- as.factor(heart.selected$cp)  
heart.selected$thalach <- as.numeric(heart.selected$thalach)  
heart.selected$oldpeak <- as.numeric(heart.selected$oldpeak)  
heart.selected$major\_vessels <- as.numeric(heart.selected$major\_vessels)  
heart.selected$target <- as.factor(heart.selected$target)  
  
  
summary(heart.selected[, -1])

## age sex cp thalach   
## Min. :29.00 female:312 asymptomatic : 77 Min. : 71.0   
## 1st Qu.:48.00 male :713 atypical angina :167 1st Qu.:132.0   
## Median :56.00 non-anginal pain:284 Median :152.0   
## Mean :54.43 typical angina :497 Mean :149.1   
## 3rd Qu.:61.00 3rd Qu.:166.0   
## Max. :77.00 Max. :202.0   
## exang oldpeak major\_vessels restwm   
## Mode :logical Min. :0.000 Min. :0.0000 Length:1025   
## FALSE:680 1st Qu.:0.000 1st Qu.:0.0000 Class :character   
## TRUE :345 Median :0.800 Median :0.0000 Mode :character   
## Mean :1.072 Mean :0.7541   
## 3rd Qu.:1.800 3rd Qu.:1.0000   
## Max. :6.200 Max. :4.0000   
## target   
## disease :526   
## no disease:499   
##   
##   
##   
##

library(psych)

## Warning: package 'psych' was built under R version 4.2.3

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

## The following object is masked from 'package:randomForest':  
##   
## outlier

describe(heart.selected[, -1])

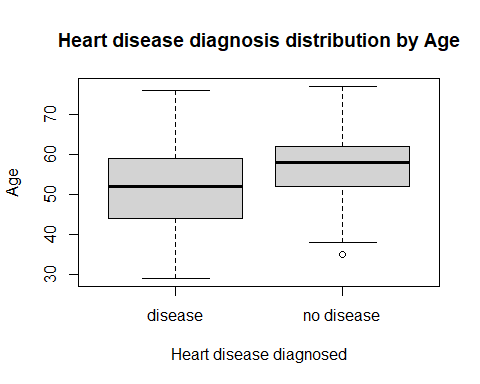
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf

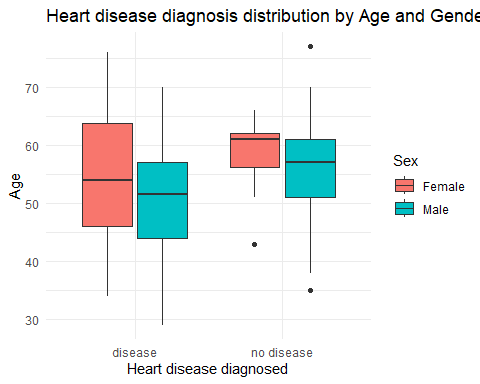
## vars n mean sd median trimmed mad min max range skew  
## age 1 1025 54.43 9.07 56.0 54.66 8.90 29 77.0 48.0 -0.25  
## sex\* 2 1025 1.70 0.46 2.0 1.74 0.00 1 2.0 1.0 -0.85  
## cp\* 3 1025 3.17 0.96 3.0 3.31 1.48 1 4.0 3.0 -0.86  
## thalach 4 1025 149.11 23.01 152.0 150.40 23.72 71 202.0 131.0 -0.51  
## exang 5 1025 NaN NA NA NaN NA Inf -Inf -Inf NA  
## oldpeak 6 1025 1.07 1.18 0.8 0.89 1.19 0 6.2 6.2 1.21  
## major\_vessels 7 1025 0.75 1.03 0.0 0.57 0.00 0 4.0 4.0 1.26  
## restwm\* 8 1025 2.14 0.97 3.0 2.17 0.00 1 4.0 3.0 -0.25  
## target\* 9 1025 1.49 0.50 1.0 1.48 0.00 1 2.0 1.0 0.05  
## kurtosis se  
## age -0.53 0.28  
## sex\* -1.28 0.01  
## cp\* -0.39 0.03  
## thalach -0.10 0.72  
## exang NA NA  
## oldpeak 1.29 0.04  
## major\_vessels 0.68 0.03  
## restwm\* -1.81 0.03  
## target\* -2.00 0.02

**首先对年龄进行探索**

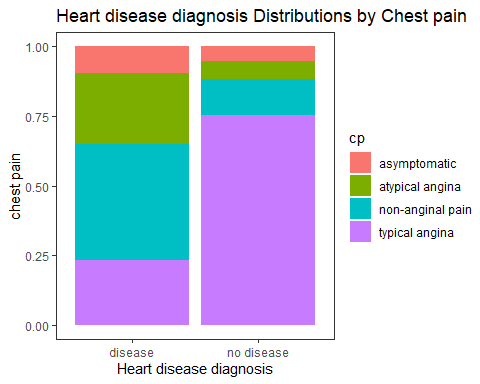
boxplot(heart.selected$age ~ heart.selected$target,  
 main="Heart disease diagnosis distribution by Age",  
 ylab="Age",xlab="Heart disease diagnosed")



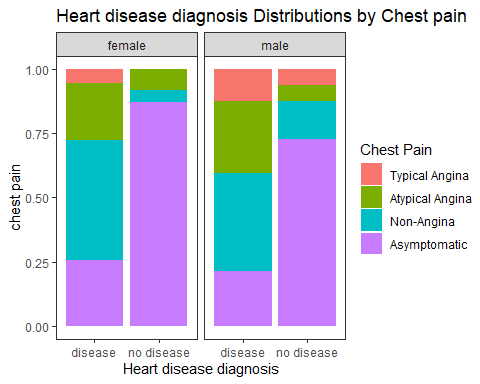
ggplot(heart.selected, aes(x = target, y = age, fill = sex)) +  
 geom\_boxplot() +  
 labs(title = "Heart disease diagnosis distribution by Age and Gender",  
 x = "Heart disease diagnosed",  
 y = "Age") +  
 scale\_fill\_discrete(name = "Sex",  
 labels = c("Female", "Male")) +  
 theme\_minimal()



ggplot(data = heart.selected, aes(x = target, fill = cp)) +   
 geom\_bar(position = "fill") +  
 labs(title = "Heart disease diagnosis Distributions by Chest pain",  
 x = "Heart disease diagnosis",  
 y = "chest pain") +  
 theme\_test()



# Create barplot by chest pain and gender  
ggplot(heart.selected, aes(x = target, fill = cp)) +  
 geom\_bar(position = "fill") +  
 facet\_wrap(~sex) +  
 labs(title = "Heart disease diagnosis Distributions by Chest pain",  
 x = "Heart disease diagnosis",  
 y = "chest pain") +  
 scale\_fill\_discrete(name = "Chest Pain",  
 labels = c("Typical Angina", "Atypical Angina", "Non-Angina", "Asymptomatic")) +  
 theme\_test()



mosaicplot(heart.selected$sex ~ heart.selected$target,  
 main="Heart disease outcome by Gender", shade=FALSE,color=TRUE,  
 xlab="Gender", ylab="Heart disease")

