Exploratory_HF

Biswajit Chowdhury

25/08/2019

1. Load the required library

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------- tidyverse 1.3.0 -
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------------------------------- tidyverse_conflicts() -
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
theme_set(theme_pubr())
theme_set(theme_bw())
```

2. load the dataset and overall summary

```r
heart_data<-read.csv("heart.csv")
names(heart_data) <- c("age", "sex", "chest_pain", "resting_bp","cholestrol",
                       "fasting_sugar", "resting_ECG", "max_heart_rate",
                       "exercise_agina", "oldpeak", "slope", "number_major_vessels",
                       "thal", "target")

head(heart_data,3)
```

```
##   age sex chest_pain resting_bp cholestrol fasting_sugar resting_ECG
## 1  63   1          3        145        233             1           0
## 2  37   1          2        130        250             0           1
## 3  41   0          1        130        204             0           0
##   max_heart_rate exercise_agina oldpeak slope number_major_vessels thal target
## 1            150              0     2.3     0                    0    1      1
## 2            187              0     3.5     0                    0    2      1
## 3            172              0     1.4     2                    0    2      1
```

```r
dim(heart_data)
```

```
## [1] 303  14
```

```r
str(heart_data)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age                 : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex                 : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ chest_pain          : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ resting_bp          : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ cholestrol          : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fasting_sugar       : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ resting_ECG         : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ max_heart_rate      : int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exercise_agina      : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak             : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slope               : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ number_major_vessels: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thal                : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ target              : int  1 1 1 1 1 1 1 1 1 1 ...
```

```r
summary(heart_data)
```

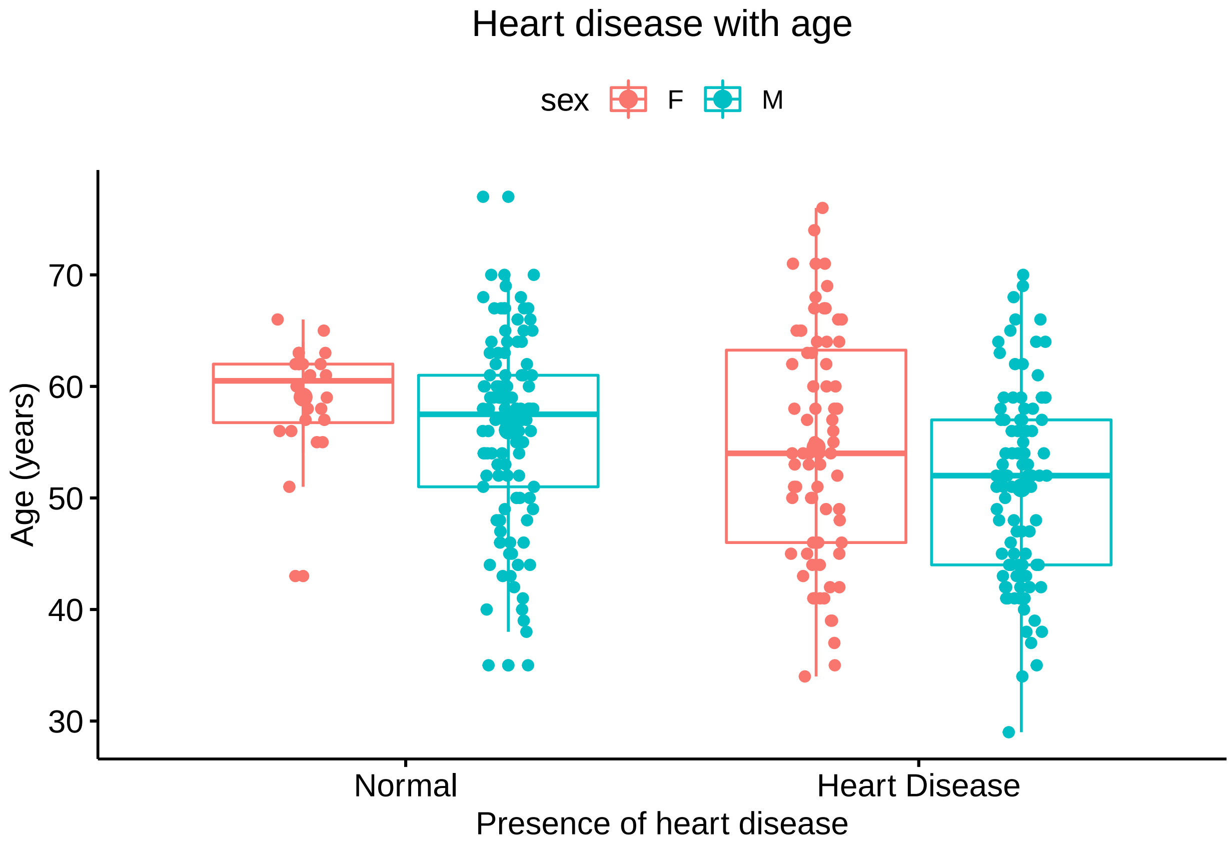
```
##       age             sex           chest_pain      resting_bp   
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
##    cholestrol    fasting_sugar     resting_ECG     max_heart_rate 
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##  exercise_agina      oldpeak         slope       number_major_vessels
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000
##       thal           target
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000
##  Median :2.000   Median :1.0000
##  Mean   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

3. Preprocessing the data for exploratory analysis

```
# variable sex is coded 0 and 1
# we want to attach value labels 0=F, 1=M

heart_data$target <- factor(heart_data$target,
      levels = c(0,1),
      labels = c("Normal", "Heart Disease"))

heart_data$sex <- factor(heart_data$sex,
      levels = c(0,1),
      labels = c("F", "M"))

heart_data$slope <- factor(heart_data$slope,
      levels = c(1,2,3),
      labels = c("Upsloping", "Flat", "Douwnsloping"))

heart_data$chest_pain <- factor(heart_data$chest_pain,
      levels = c(1,2,3, 0),
      labels = c("Typical angina ", "Atypical angina ", "Non-anginal pain ", "Asymptomatic"))

heart_data$fasting_sugar <- factor(heart_data$fasting_sugar,
      levels = c(0,1),
      labels = c("False", "TRUE"))

heart_data$resting_ECG <- factor(heart_data$resting_ECG,
      levels = c(0,1, 2),
      labels = c("Normal", "Mild", "Severe"))

heart_data$exercise_agina <- factor(heart_data$exercise_agina,
      levels = c(0,1),
      labels = c("No", "Yes"))

heart_data$number_major_vessels <- factor(heart_data$number_major_vessels,
      levels = c(0,1, 2, 3, 4),
      labels = c("None", "One", "Two", "Three", "Four"))

heart_data$thal <- factor(heart_data$thal,
      levels = c(1,2,3),
      labels = c("Normal", "Fixed defect", "Reversable defect"))
```

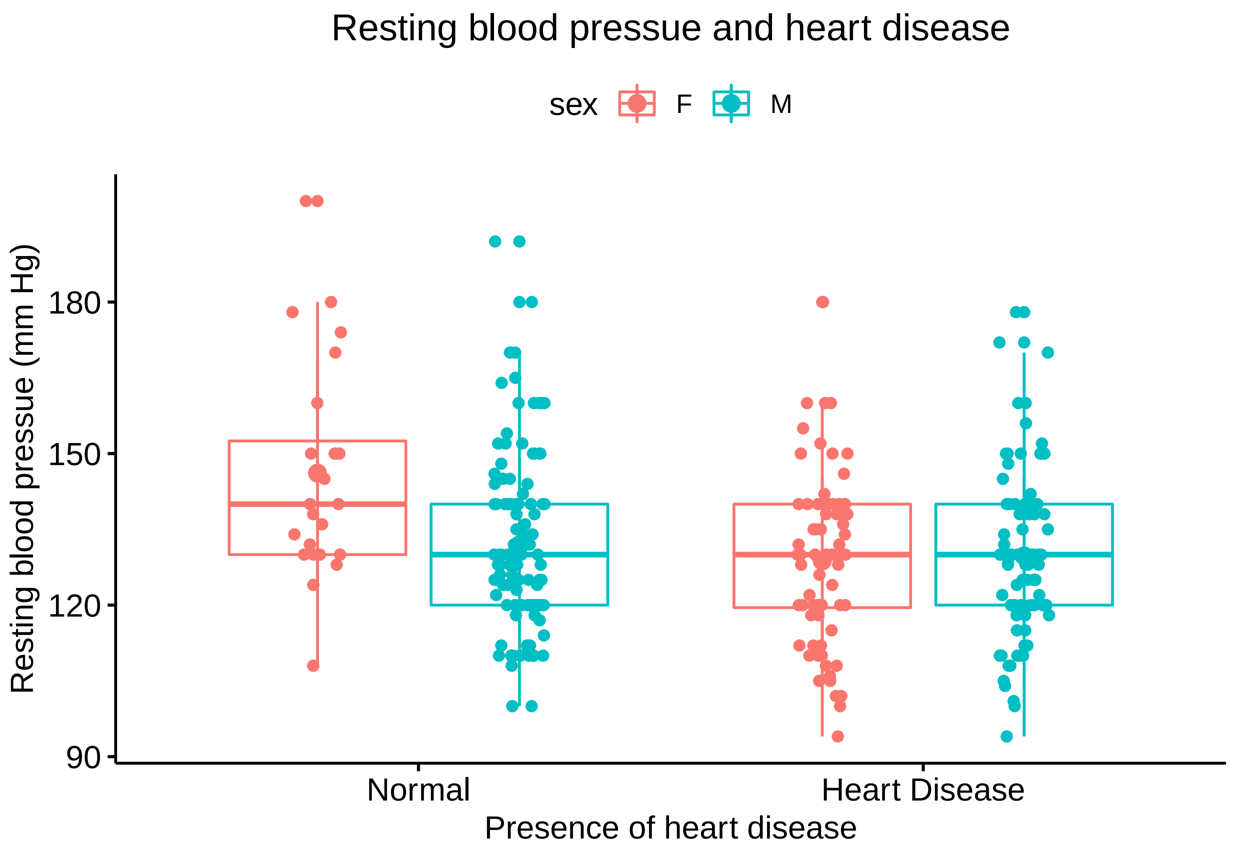
4. Explore the qunatitative vriables

```
# Presence of heart disease among males and females due to age

ggboxplot(heart_data, x = "target", y = "age",
          xlab = "Presence of heart disease", ylab = "Age (years)",
          title = "Heart disease with age", # title of the graph
          color = "sex", # color by gender
          add = c("jitter", "mean_sd"))+ # show the distribution of the values
          theme(plot.title = element_text(hjust = 0.5))
```



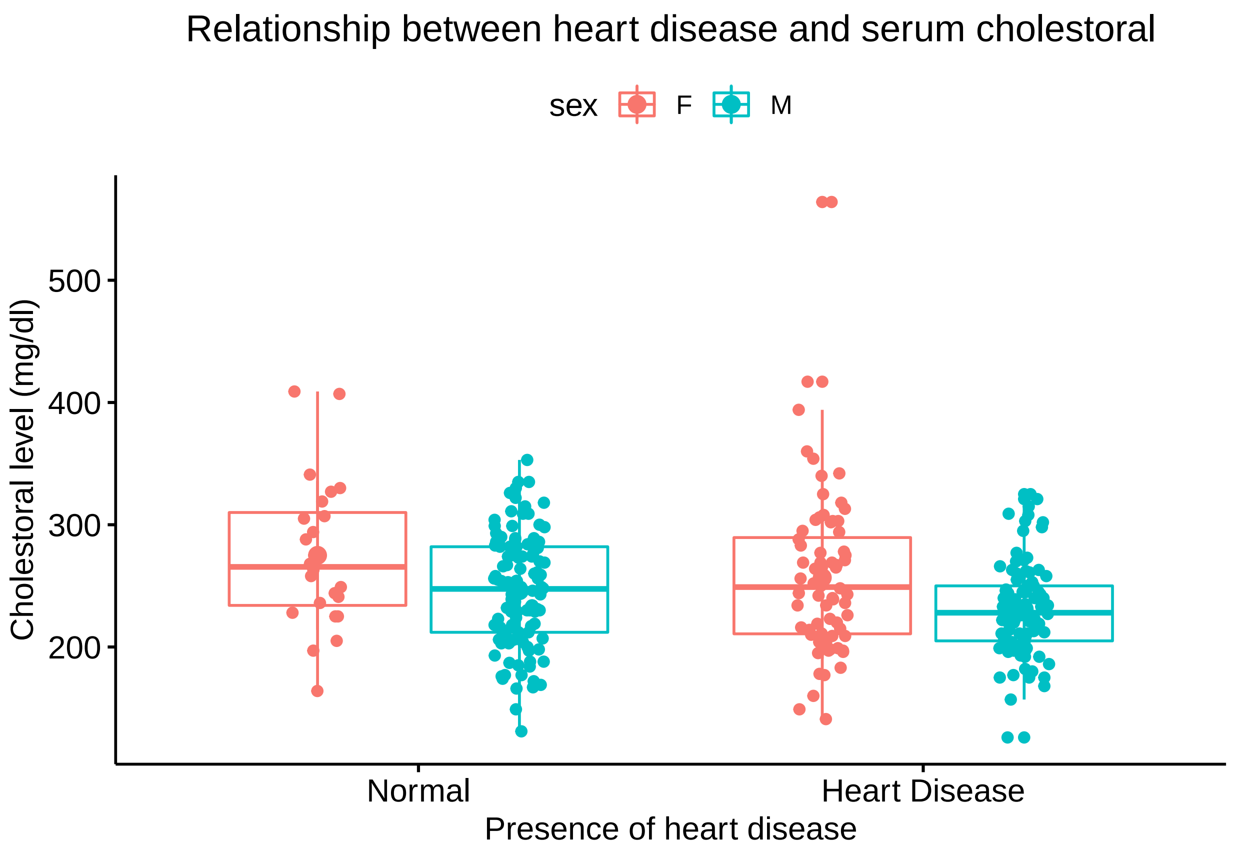
Adult females are more prone to get heart disease compared with men.

```r
# Presence of heart disease among males and females due to resting blood pressue

ggboxplot(heart_data, x = "target", y = "resting_bp",
          xlab = "Presence of heart disease", ylab = "Resting blood pressure (mm Hg)",
          title = "Resting blood pressue and heart disease",
          color = "sex", add = c("jitter", "mean_sd")) +
          theme(plot.title = element_text(hjust = 0.5))
```



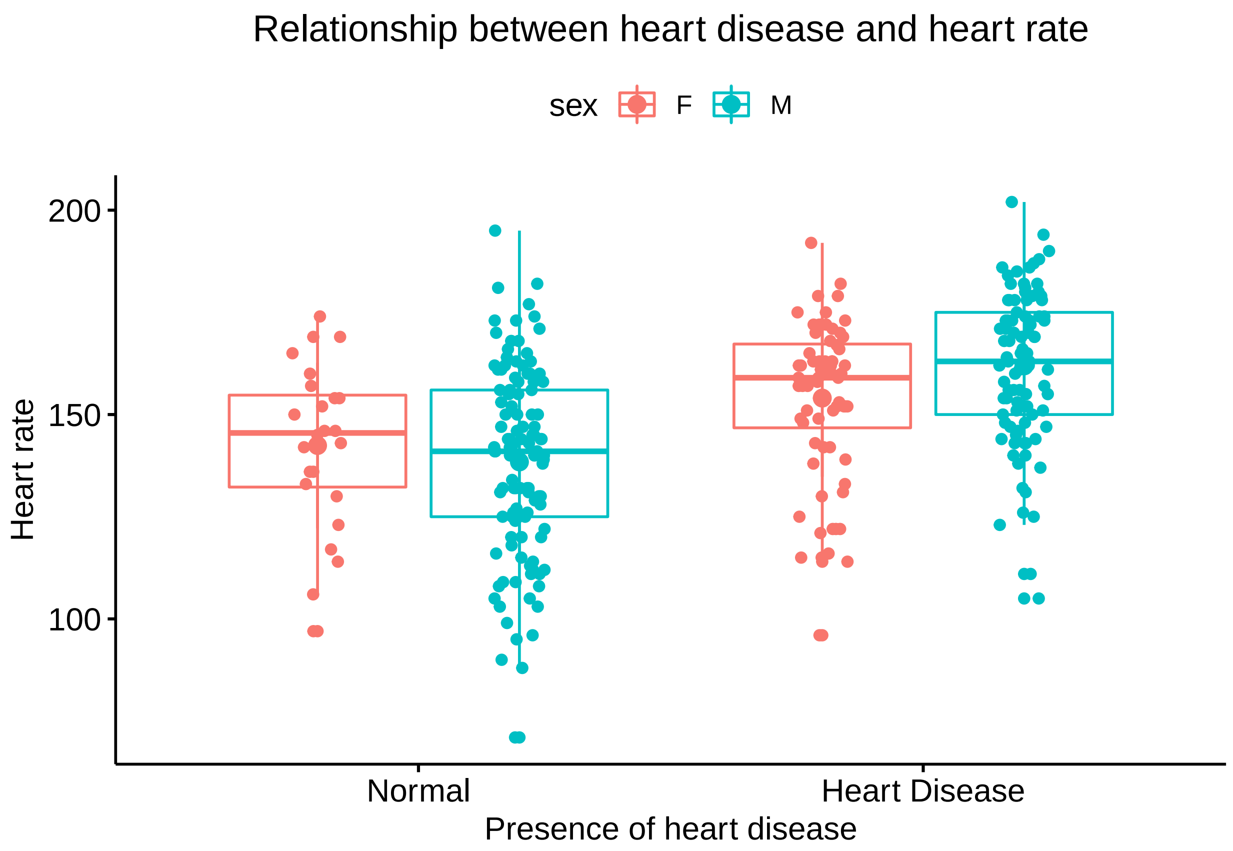
Men and women with heart disease had almost same resting blood pressure.

```r
# Changes of chlestreol level between heart disease and normal people

ggboxplot(heart_data, x = "target", y = "cholestrol",
          xlab = "Presence of heart disease", ylab = "Cholestoral level (mg/dl)",
          title = "Relationship between heart disease and serum cholestoral",
          color = "sex", add = c("jitter", "mean_sd")) +
          theme(plot.title = element_text(hjust = 0.5))
```



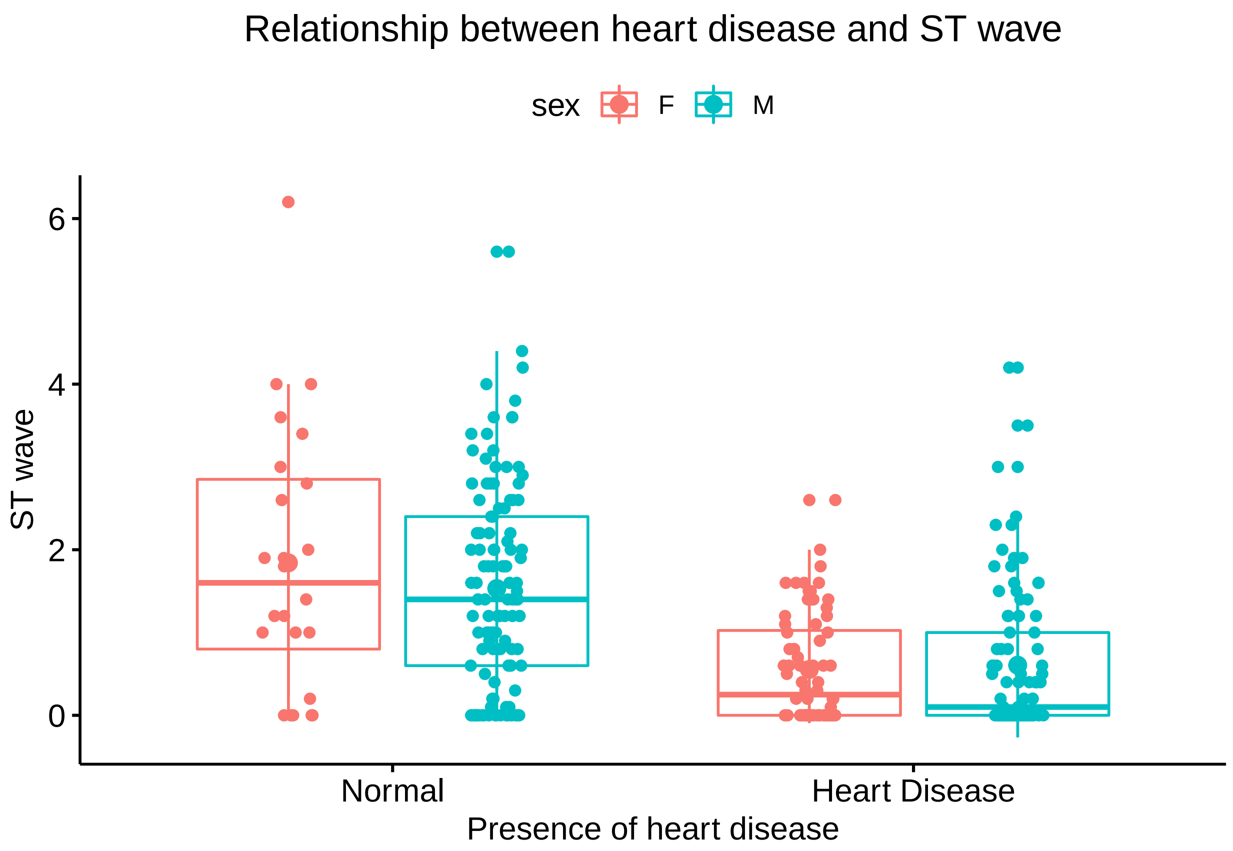
No change in cholestreol level between normal and heart disease.

```
# Chnages of heart rate between normal and disease

ggboxplot(heart_data, x = "target", y = "max_heart_rate",
          xlab = "Presence of heart disease", ylab = "Heart rate",
          title = "Relationship between heart disease and heart rate",
          color = "sex", add = c("jitter", "mean_sd")) +
          theme(plot.title = element_text(hjust = 0.5))
```



Men with heart disease had increased heart rate.

```
# Changes of old peak among the peopple with heart disease and healthy

ggboxplot(heart_data, x = "target", y = "oldpeak",
          xlab = "Presence of heart disease", ylab = "ST wave",
          title = "Relationship between heart disease and ST wave",
          color = "sex", add = c("jitter", "mean_sd")) +
          theme(plot.title = element_text(hjust = 0.5))
```



Presence of reduced ST wave in people with heart disease.

5. Explore the Categorical vriables

Calculate the frequency of categorical variables

```
# Counts for gender categories
table(heart_data$sex)
```

```
##
##   F   M
##  96 207
```

```
# Cross classification counts for gender by heart failure

table(heart_data$target, heart_data$sex)
```

```
##
##                   F   M
##   Normal         24 114
##   Heart Disease  72  93
```

Assess the count of people by heart disease, gender, and fasting sugar

```
# Multidimensional tables based on three or more categorical variables.

table1 <- table(heart_data$target, heart_data$sex,
                heart_data$fasting_sugar)

ftable(table1) # print the results more attractively
```

```
##                    False TRUE
##
## Normal        F       18    6
##               M       98   16
## Heart Disease F       66    6
##               M       76   17
```

Assess the count of people by heart failure, gender, chest pain

```
table1 <- table(heart_data$target, heart_data$sex,
                heart_data$chest_pain)

ftable(table1)
```

```
##                    Typical angina  Atypical angina  Non-anginal pain  Asymptomatic
##
## Normal        F                 2                1                 0            21
##               M                 7               17                 7            83
## Heart Disease F                16               34                 4            18
##               M                25               35                12            21
```

Assess the count of people by heart disease, gender, and resting ECG

```
table1 <- table(heart_data$target, heart_data$sex,
                heart_data$resting_ECG)

ftable(table1)
```

```
##                    Normal Mild Severe
##
## Normal        F        13    9      2
##               M        66   47      1
## Heart Disease F        31   40      1
##               M        37   56      0
```

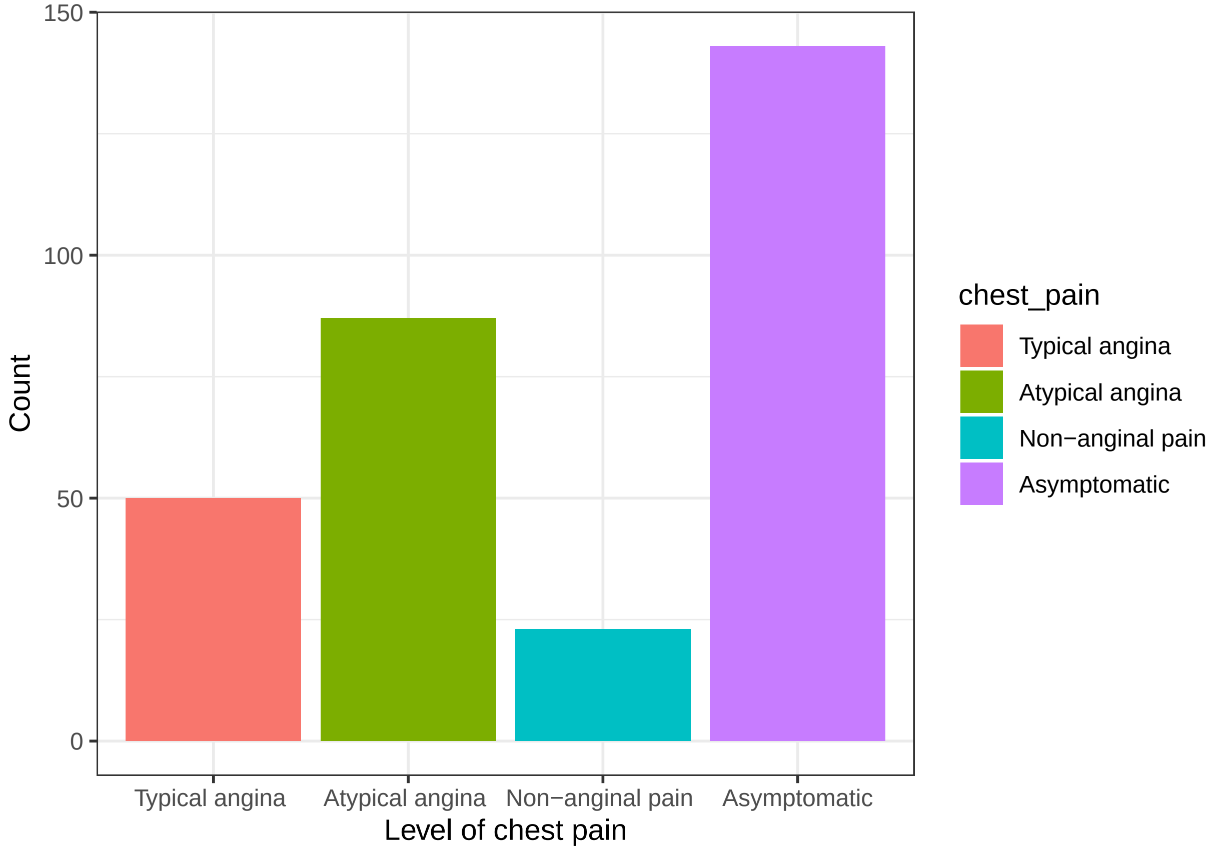
Assess the count of people by heart disease, gender, exercise agina, and resting ECG

```
table1 <- table(heart_data$target, heart_data$sex,
                 heart_data$exercise_agina, heart_data$resting_ECG)

ftable(table1)
```

```
##                      Normal Mild Severe
##
## Normal        F No        8    2      0
##                 Yes       5    7      2
##               M No       29   22      1
##                 Yes      37   25      0
## Heart Disease F No       25   38      1
##                 Yes       6    2      0
##               M No       31   47      0
##                 Yes       6    9      0
```

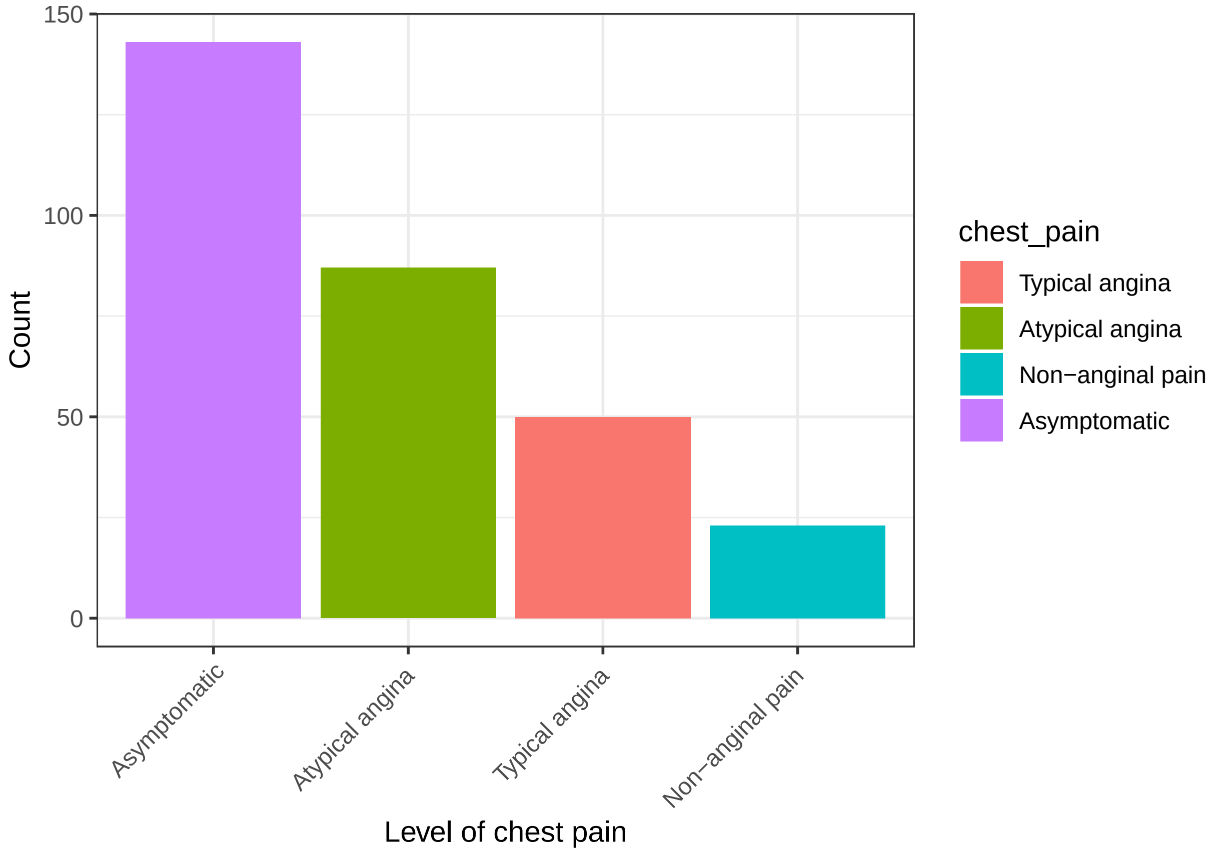
Visualize the categorical variables

```
# Here we can assess the number of people diagnosed with chest pain
ggplot(heart_data, aes(x = chest_pain, fill=chest_pain)) +
        geom_bar() +  xlab('Level of chest pain') + ylab("Count")
```



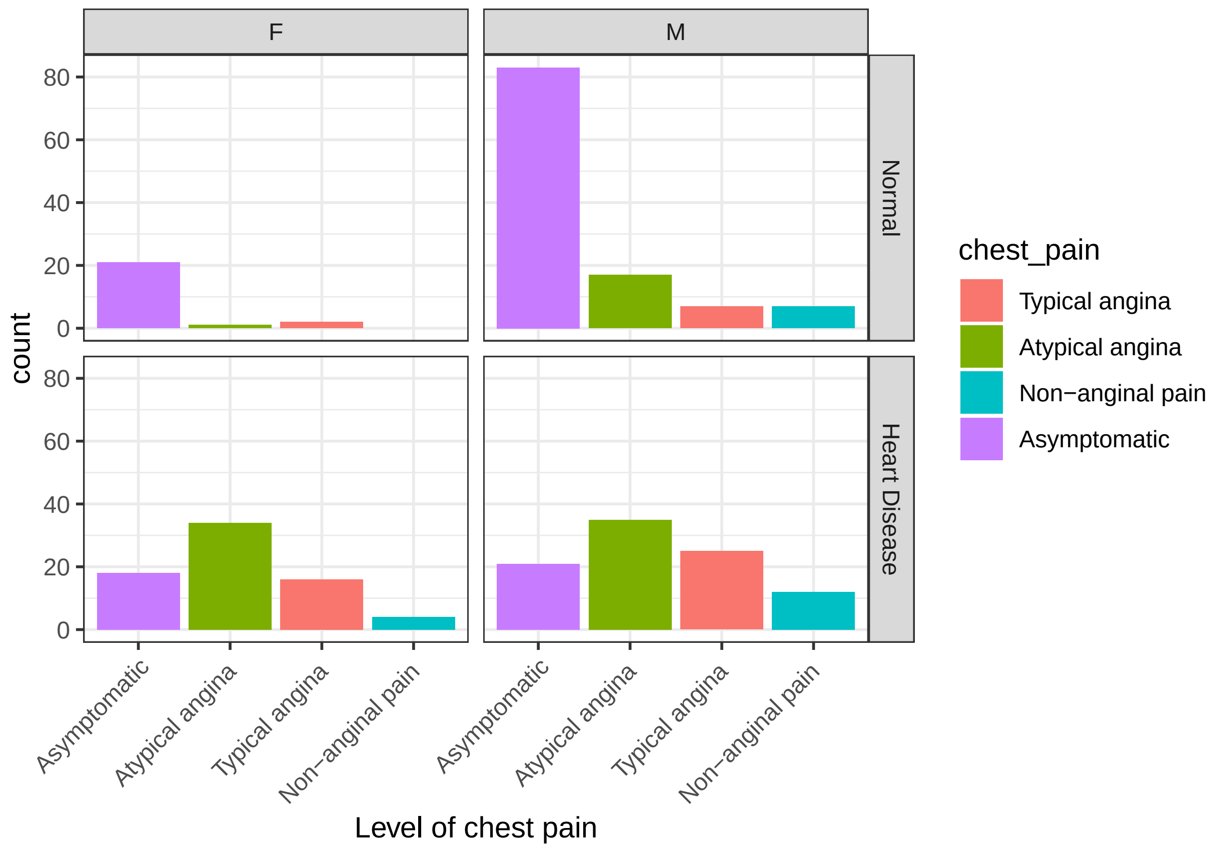
To make it easy to understand, we can make it in desending order. For this we will make a functions that sort the variables crosponding their total counts

```r
# re-order levels
reorder_size <- function(x) {
        factor(x, levels = names(sort(table(x), decreasing = TRUE)))
}

ggplot(heart_data, aes(x = reorder_size(chest_pain), fill=chest_pain)) +
        geom_bar() +
  xlab('Level of chest pain') + ylab("Count") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



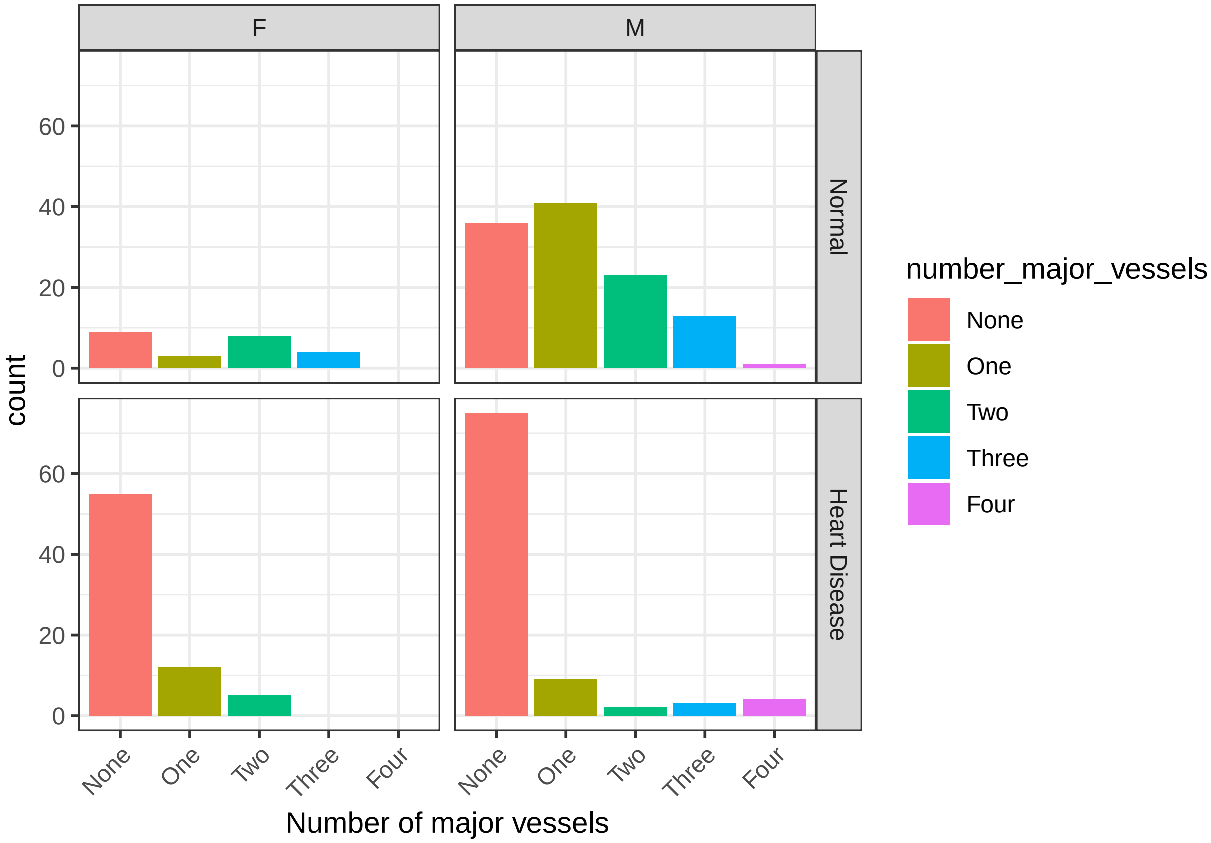
Changes of chest pain with sex and heart disease

```r
ggplot(heart_data, aes(x = reorder_size(chest_pain), fill=chest_pain)) +
        geom_bar() +
  xlab('Level of chest pain') +
  facet_grid(target~ sex) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



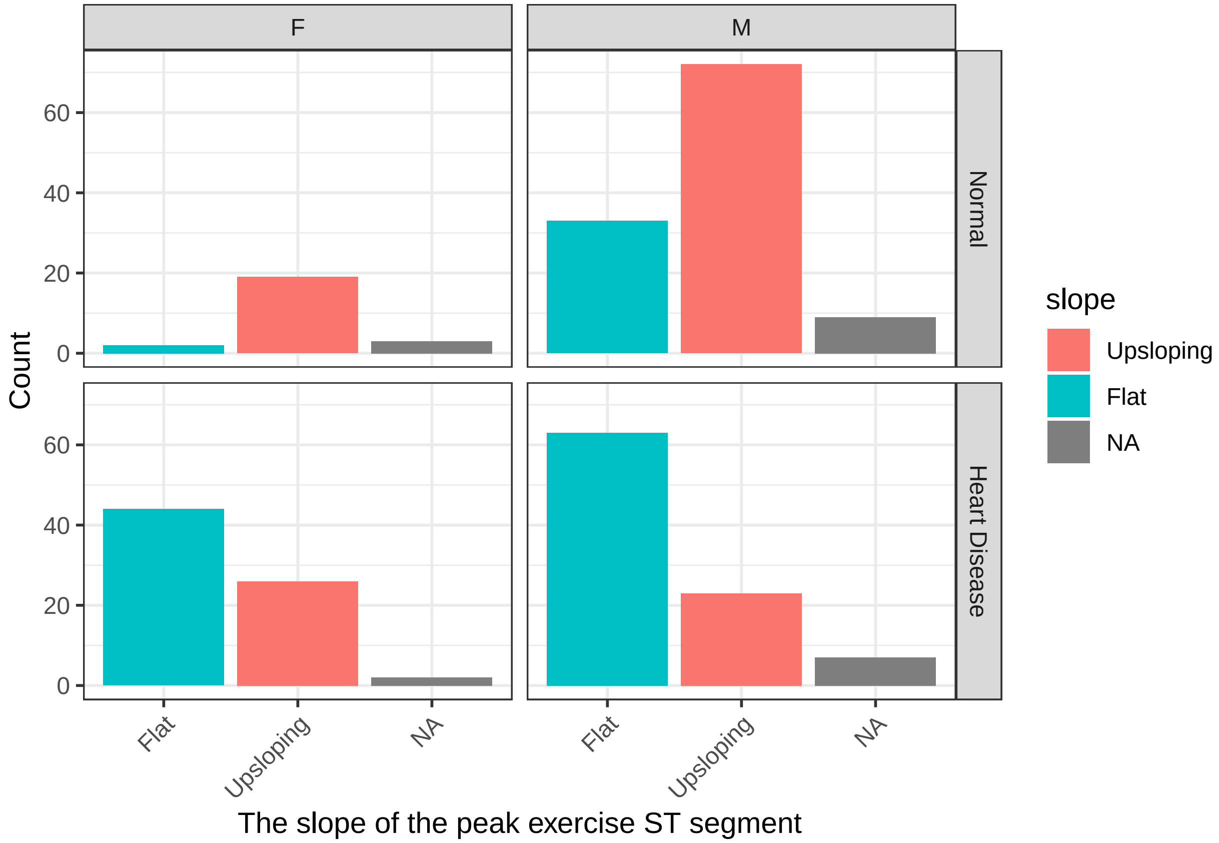
Number of major vessels with gender and heart disease

```
ggplot(heart_data, aes(x = reorder_size(number_major_vessels), fill=number_major_vessels)) +
        geom_bar() +
  xlab('Number of major vessels') +
  facet_grid(target~ sex) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



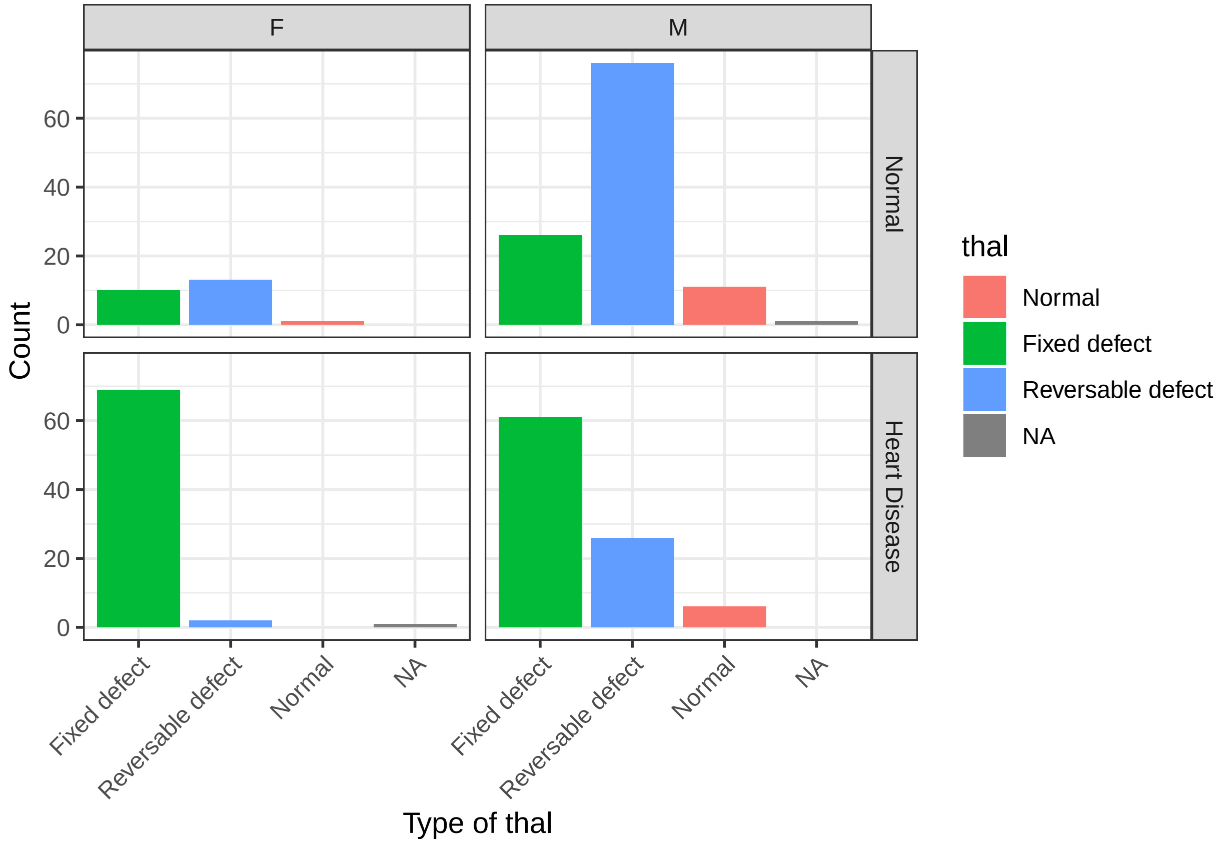
The slope of the peak exercise of ST segment with gender and heart disease

```
ggplot(heart_data, aes(x = reorder_size(slope), fill=slope)) +
        geom_bar() +
  xlab('The slope of the peak exercise ST segment') + ylab(" Count")+
  facet_grid(target~ sex) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Changes of thal with sex and heart disease

```
ggplot(heart_data, aes(x = reorder_size(thal), fill=thal)) +
        geom_bar() +
  xlab('Type of thal') + ylab(" Count")+
  facet_grid(target~ sex) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



6. Proportion/Percentage

To do this we simply use the frequency tables produced by table() to the prop.table() function

```
# percentages of gender categories
table1<- table(heart_data$sex)
prop.table(table1)
```

```
##
##         F         M
## 0.3168317 0.6831683
```

```
# percentage of cross classication counts for gender by heart disease

table2<- table(heart_data$target, heart_data$sex)
prop.table(table2)
```

```
##
##                          F          M
##   Normal        0.07920792 0.37623762
##   Heart Disease 0.23762376 0.30693069
```

```
round(prop.table(table2), 3)*100
```

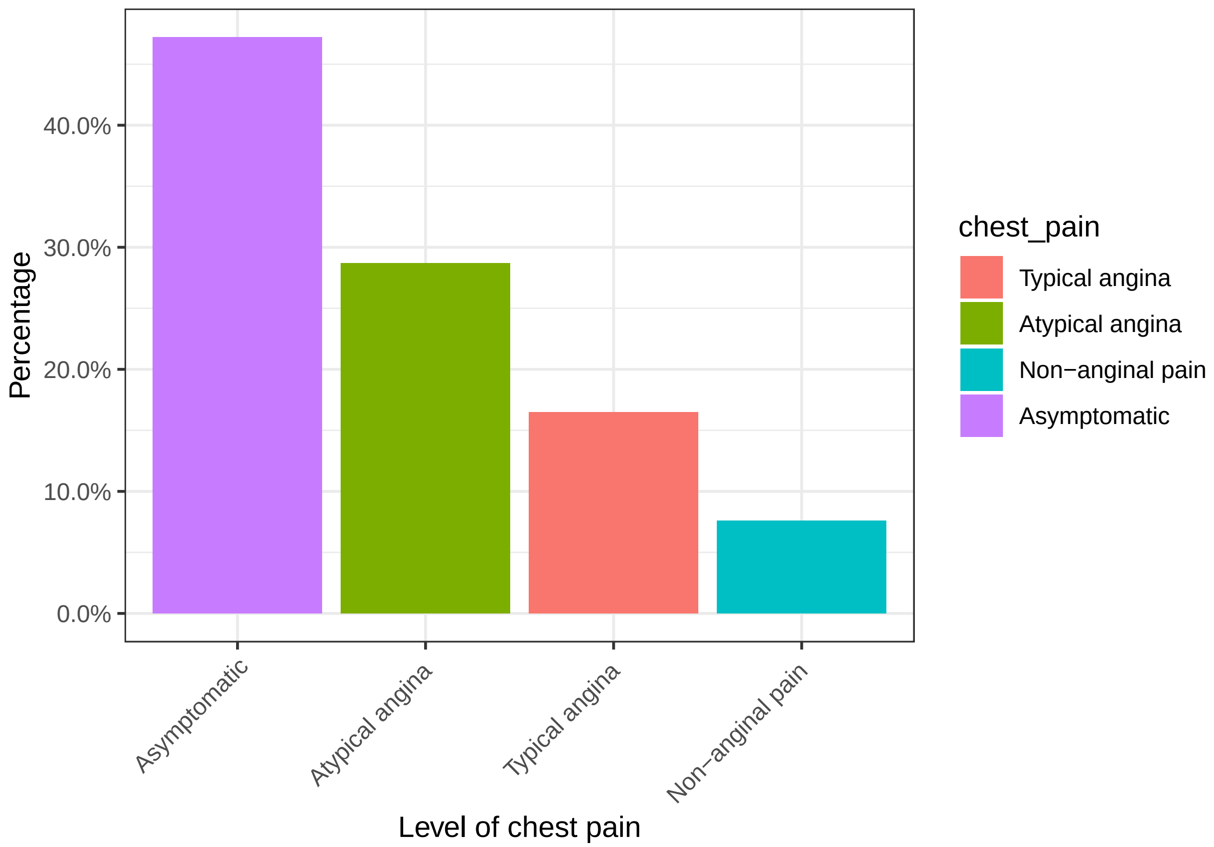
```
##
##                      F    M
##   Normal          7.9 37.6
##   Heart Disease 23.8 30.7
```

```
#percentage of heart failure by sex, and fasting sugar
table1 <- table(heart_data$target, heart_data$sex,
                 heart_data$fasting_sugar)

ftable(round(prop.table(table1), 3)*100)
```

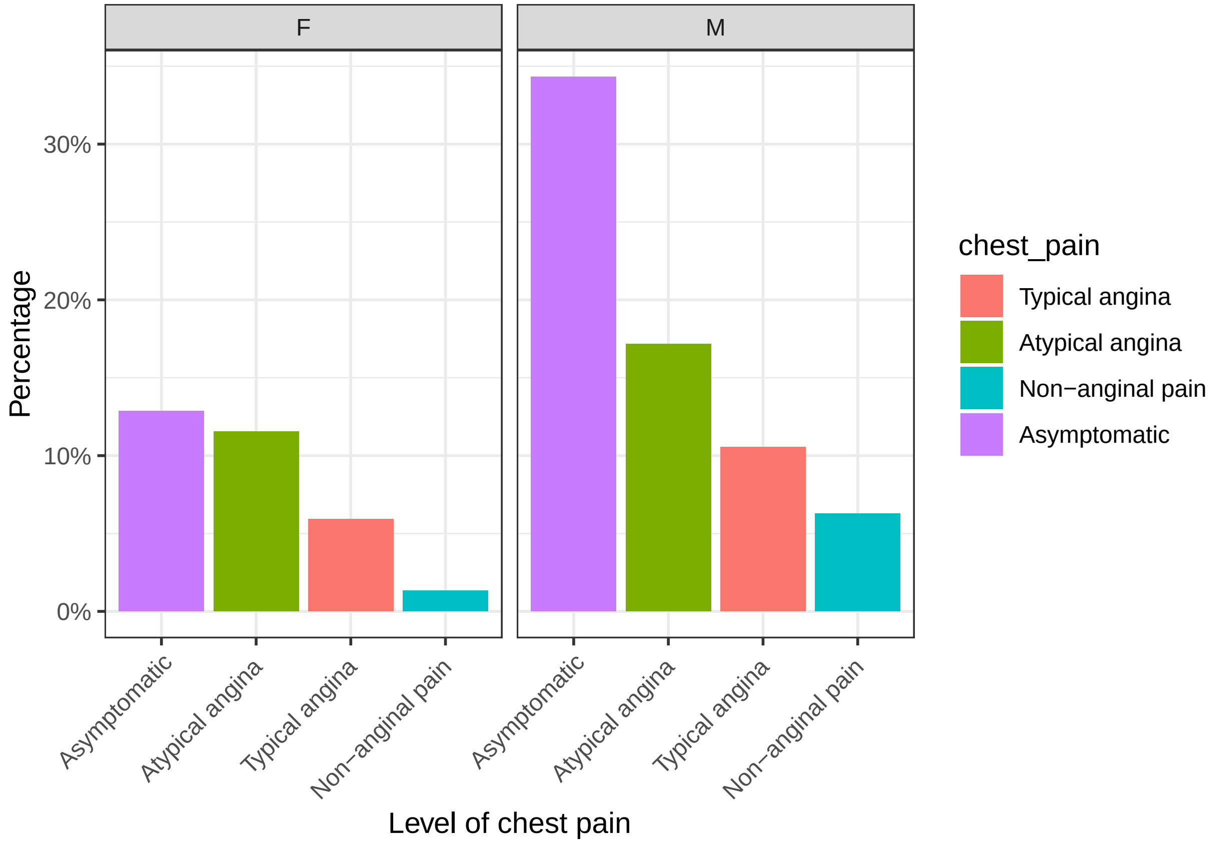
```
##                   False TRUE
##
## Normal        F    5.9  2.0
##               M   32.3  5.3
## Heart Disease F   21.8  2.0
##               M   25.1  5.6
```

Visualize the data

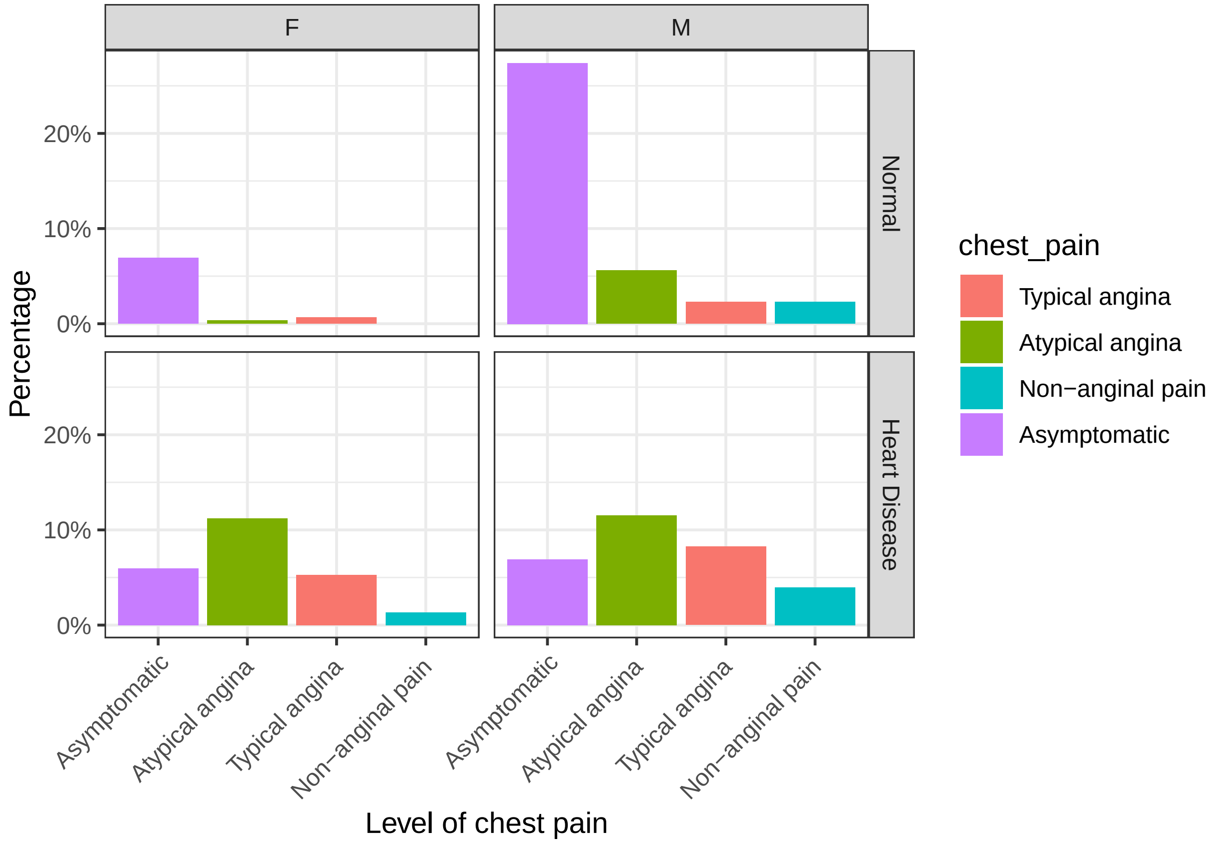
```
ggplot(heart_data, aes(x = reorder_size(chest_pain), fill=chest_pain)) +
        geom_bar(aes(y = (..count..)/sum(..count..))) +
  xlab('Level of chest pain') + ylab("Count") +

  scale_y_continuous(labels = scales::percent, name = "Percentage") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# Chest pain with only one variable heart disease
ggplot(heart_data, aes(x = reorder_size(chest_pain), fill=chest_pain)) +
        geom_bar(aes(y = (..count..)/sum(..count..))) +
  xlab('Level of chest pain') + ylab("Count") +
  scale_y_continuous(labels = scales::percent, name = "Percentage") +
  facet_grid(~ sex) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# chest pain with multiple valriables sex and heart disease

ggplot(heart_data, aes(x = reorder_size(chest_pain), fill=chest_pain)) +
        geom_bar(aes(y = (..count..)/sum(..count..))) +
  xlab('Level of chest pain') +
  scale_y_continuous(labels = scales::percent, name = "Percentage") +
  facet_grid(target~ sex) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


F
M
Normal
Heart Disease
Percentage
20%
10%
0%
Asymptomatic
Atypical angina
Typical angina
Non−anginal pain
Level of chest pain
chest_pain
Typical angina
Atypical angina
Non−anginal pain
Asymptomatic