Heart

Pablo Parra

29/12/2021

Contents

1 Summary

1.1 Introduction / Overview

This is the last project of the Professional Certificate Program of Data Science, organized by Harvard University in the platform edX. There are not restrictions about the dataset selected and the methods applied.

The dataset selected contains... Credits to "fedesoriano" and hospitals...

The main goal is...

The methods utilized in the proyect are ...

1.2 Data structure

These are the 12 columns/features contained in the data:

- **Age**: Age of the patient
- **Sex**: Sex of the patient (M: Male, F: Female)
- **ChestPainType**: Type of chest pain (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)
- **RestingBP**:
- **Cholesterol**:
- **FastingBS**:
- **RestingECG**:
- **MaxHR**:
- **ExerciseAngina**:
- **Oldpeak**:
- **ST_Slope**:
- **HeartDisease**:

The first step is to load the data. It could be downloaded **here**. And then, after locating the file in the active directory, it could be loaded in R by the following line:

```
# Loading data
data <- read_csv("heart.csv")
```

This is the structure of the data:

```
# Data structure
str(data)
```

```
## spec_tbl_df [918 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Age           : num [1:918] 40 49 37 48 54 39 45 54 37 48 ...
##  $ Sex           : chr [1:918] "M" "F" "M" "F" ...
##  $ ChestPainType : chr [1:918] "ATA" "NAP" "ATA" "ASY" ...
##  $ RestingBP     : num [1:918] 140 160 130 138 150 120 130 110 140 120 ...
##  $ Cholesterol   : num [1:918] 289 180 283 214 195 339 237 208 207 284 ...
##  $ FastingBS     : num [1:918] 0 0 0 0 0 0 0 0 0 0 ...
##  $ RestingECG    : chr [1:918] "Normal" "Normal" "ST" "Normal" ...
##  $ MaxHR         : num [1:918] 172 156 98 108 122 170 170 142 130 120 ...
##  $ ExerciseAngina: chr [1:918] "N" "N" "N" "Y" ...
```

```
##  $ Oldpeak       : num [1:918] 0 1 0 1.5 0 0 0 0 1.5 0 ...
##  $ ST_Slope      : chr [1:918] "Up" "Flat" "Up" "Flat" ...
##  $ HeartDisease  : num [1:918] 0 1 0 1 0 0 0 0 1 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Age = col_double(),
##   ..   Sex = col_character(),
##   ..   ChestPainType = col_character(),
##   ..   RestingBP = col_double(),
##   ..   Cholesterol = col_double(),
##   ..   FastingBS = col_double(),
##   ..   RestingECG = col_character(),
##   ..   MaxHR = col_double(),
##   ..   ExerciseAngina = col_character(),
##   ..   Oldpeak = col_double(),
##   ..   ST_Slope = col_character(),
##   ..   HeartDisease = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

2 Methods / Analysis

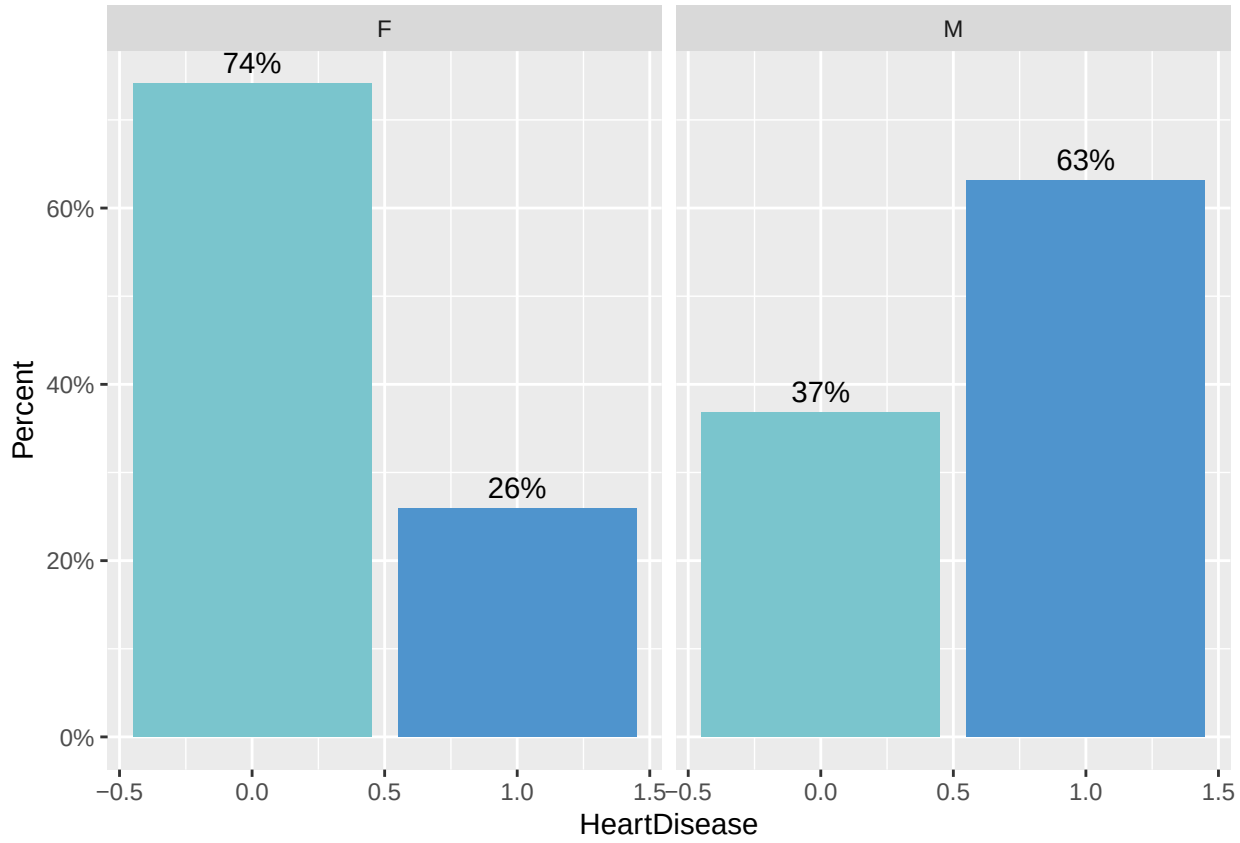
2.1 Data cleaning

Here, data cleaning.

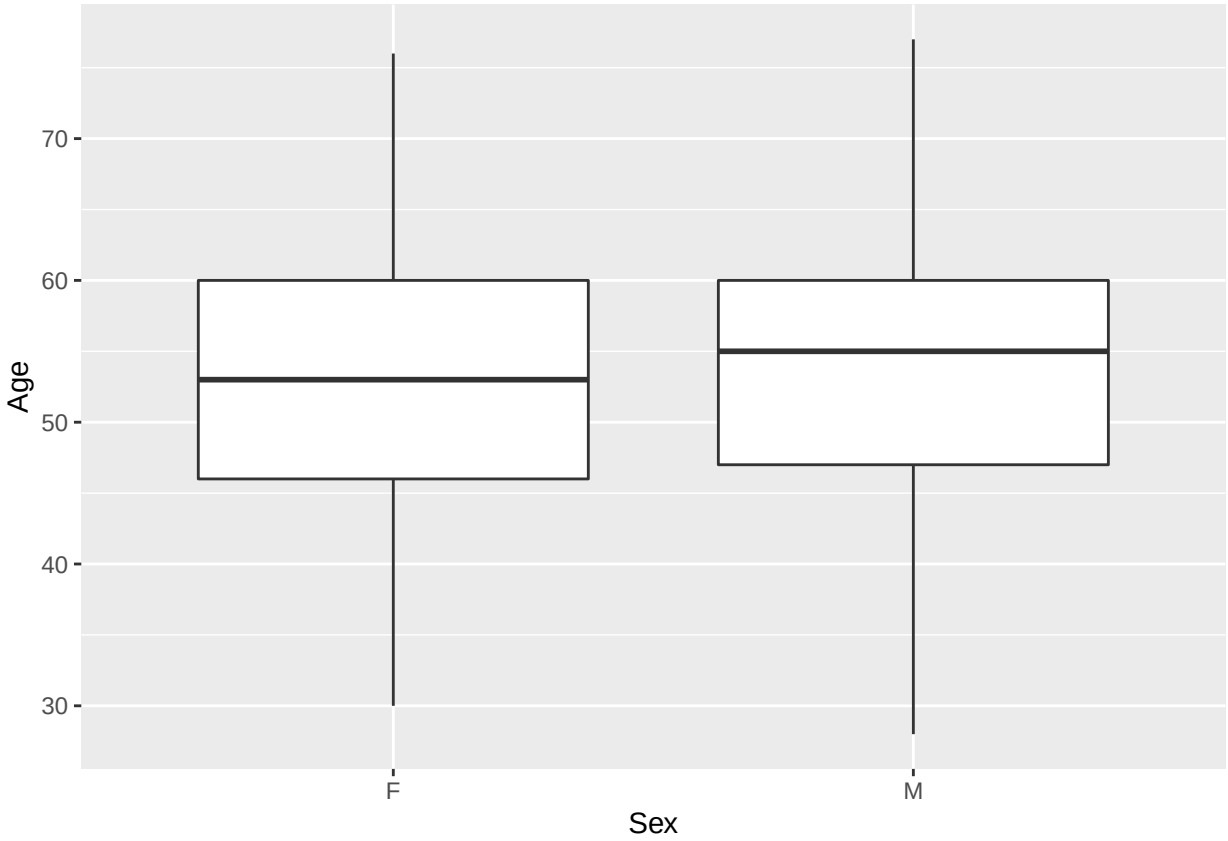
2.2 Data exploration and data visualization

View the data

```
# Gráfico 1
data %>% ggplot(aes(x= HeartDisease,  group=Sex)) +
  geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
  geom_text(aes( label = scales::percent(..prop..),
                 y= ..prop.. ), stat= "count", vjust = -.5) +
  labs(y = "Percent", fill="HeartDisease") +
  facet_grid(~Sex) +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_manual(values=c("cadetblue3", "steelblue3")) +
  theme(legend.position="none")
```




```
# Gráfico 2
data %>% ggplot(aes(Sex, Age, fill = HeartDisease)) + geom_boxplot() +
  scale_fill_manual(values=c("cadetblue3", "steelblue3"))
```



2.3 Modeling approach

These are the models used:

- Model 1
- Model 2
- Model 3

Specific train/test split (e.g. 50/50 vs 90/10)

(Que no se vean los warnings!!)

2.3.1 Model 1

Model 1

2.3.2 Model 2

Model 2

2.3.3 Model 3

Model 3

3 Results

These are the results

4 Conclusion

The conclusion