Heart

Pablo Parra

29/01/2022

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**. MODIFICAR And then, after locating the file in the active directory, it could be loaded in R by the following line:

```
# Loading data - Different options
# If you only need to read the data online;
data <- read_csv("https://raw.githubusercontent.com/p-parra/Heart-Project/main/heart.csv")

# If you prefer to download it in your PC:
# url <- "https://raw.githubusercontent.com/p-parra/Heart-Project/main/heart.csv"
# download.file(url, destfile = "heart.csv")
# data <- read_csv("heart.csv")

# If it is already downloaded:
# data <- read_csv("heart.csv")
```

This is the structure of the data and the first 6 rows 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>
```

```
# First rows of data
head(data)
```

```
## # A tibble: 6 x 12
##     Age Sex   ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
##   <dbl> <chr> <chr>             <dbl>       <dbl>     <dbl> <chr>      <dbl>
## 1    40 M     ATA                 140         289         0 Normal       172
## 2    49 F     NAP                 160         180         0 Normal       156
## 3    37 M     ATA                 130         283         0 ST            98
## 4    48 F     ASY                 138         214         0 Normal       108
## 5    54 M     NAP                 150         195         0 Normal       122
## 6    39 M     NAP                 120         339         0 Normal       170
## # ... with 4 more variables: ExerciseAngina <chr>, Oldpeak <dbl>,
## #   ST_Slope <chr>, HeartDisease <dbl>
```

2 Methods / Analysis

2.1 Data cleaning

In order to prevent process data with missing values, it is necessary observe is the is any NA.

```
is.na(data) %>% sum()
```

```
## [1] 0
```

As could be seen, there is not any NA or missing value.

Anyway, if 'Cholesterol' feature is observed in detail, there are several '0's that must be taken into account.

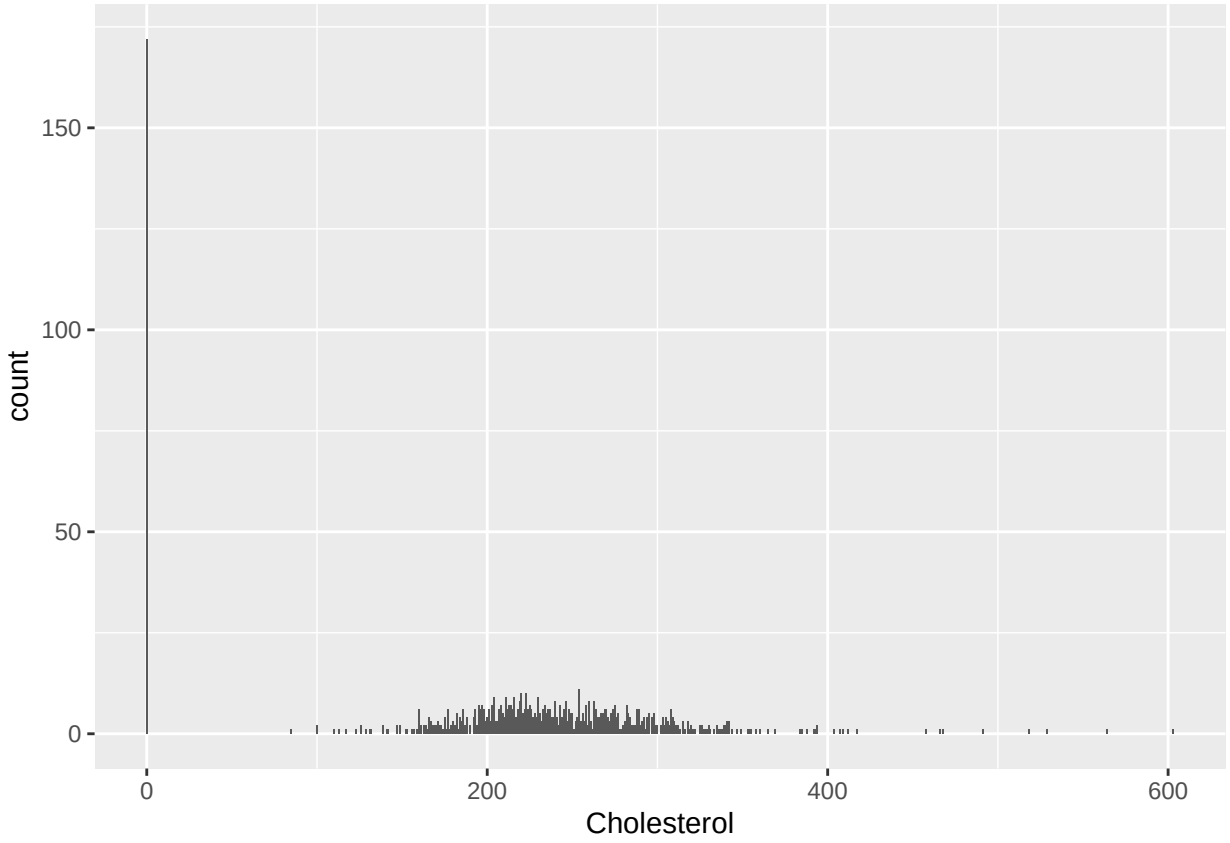
```
# Searching '0' values:
table(data$RestingBP)
```

```
##
##   0  80  92  94  95  96  98 100 101 102 104 105 106 108 110 112 113 114 115 116
##   1   1   1   2   6   1   1  15   1   3   3   9   3   7  58  14   1   2  19   2
## 117 118 120 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138
##   1  10 132  12   2  12  29   7   1  18   1 118   4  17   6  11  20  13   5  17
## 139 140 141 142 143 144 145 146 148 150 152 154 155 156 158 160 164 165 170 172
##   5 107   3  11   2   8  18   4   2  55   7   3   8   2   4  50   1   2  14   2
## 174 178 180 185 190 192 200
##   1   3  12   1   2   1   4
```

```
table(data$Cholesterol)
```

```
##
##   0  85 100 110 113 117 123 126 129 131 132 139 141 142 147 149 152 153 156 157
## 172   1   2   1   1   1   1   2   1   1   1   2   1   1   2   2   1   1   1   1
## 159 160 161 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179
##   1   6   2   2   2   1   4   3   2   2   2   3   2   2   1   4   1   6   1   2
## 180 181 182 183 184 185 186 187 188 190 192 193 194 195 196 197 198 199 200 201
##   3   2   5   1   4   3   6   2   4   2   4   6   2   7   6   7   6   3   4   6
## 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221
##   3   7   9   3   3   6   7   5   4   9   6   7   7   6   9   4   6   8  10   5
## 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241
##   6  10   6   7   6   4   5   4   9   5   3   6   7   5   6   6   4   4   8   4
## 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261
##   2   7   4   6   8   3   6   5   5   1   3   4  11   3   5   3   7   2   8   3
## 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281
##   1   8   6   4   4   5   5   6   6   4   3   5   6   7   4   5   1   1   2   3
## 282 283 284 285 286 287 288 289 290 291 292 293 294 295 297 298 299 300 302 303
##   7   5   4   2   2   2   6   6   2   3   4   1   4   5   4   5   2   2   2   4
## 304 305 306 307 308 309 310 311 312 313 315 316 318 319 320 321 322 325 326 327
##   2   4   3   2   6   4   3   2   2   1   3   1   3   1   2   1   1   2   2   1
## 328 329 330 331 333 335 336 337 338 339 340 341 342 344 347 349 353 354 355 358
##   1   1   2   1   1   2   1   1   1   2   2   3   3   1   1   1   1   1   1   1
## 360 365 369 384 385 388 392 393 394 404 407 409 412 417 458 466 468 491 518 529
##   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1   1   1   1   1   1
## 564 603
##   1   1
```

```
data %>% ggplot(aes(Cholesterol)) + geom_bar()
```



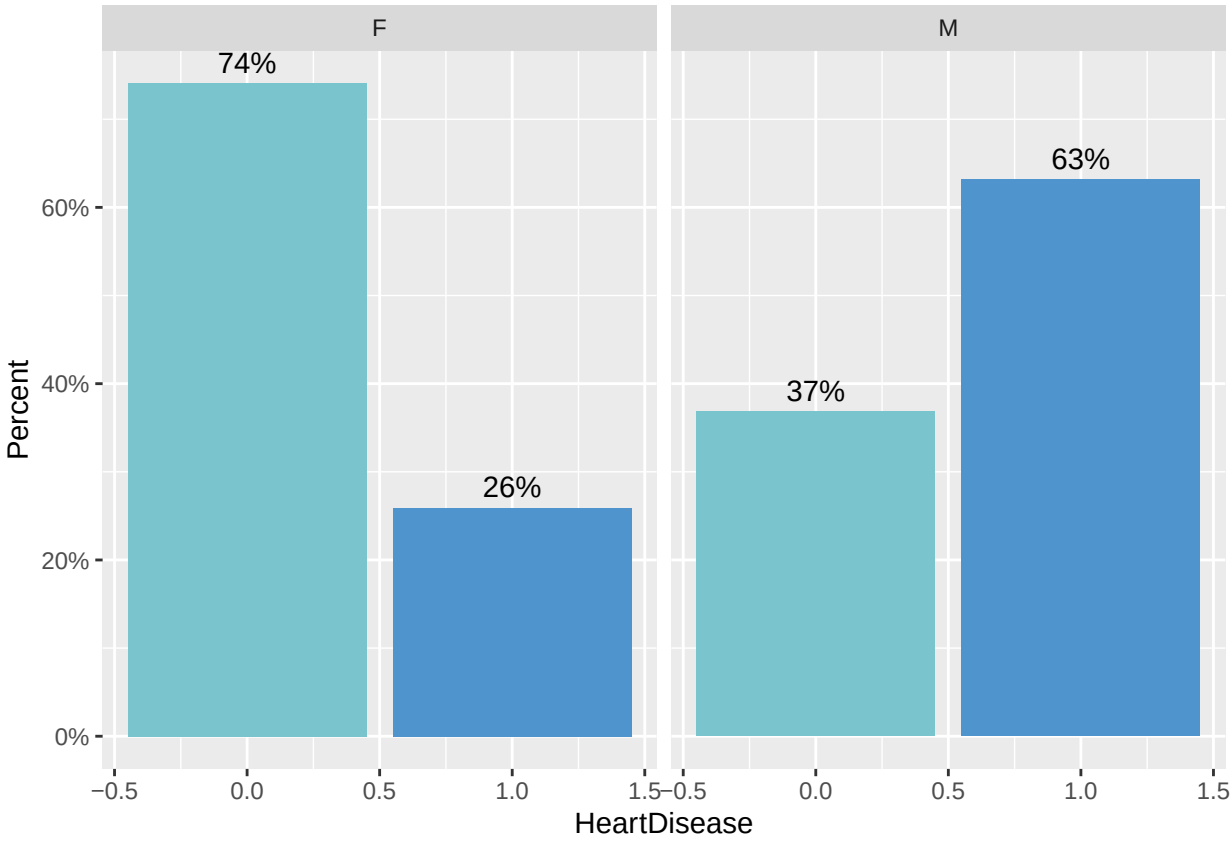

```
values_with_0 <- data %>% filter(Cholesterol == 0) %>% nrow()
values_with_0 / nrow(data)
```

```
## [1] 0.1873638
```

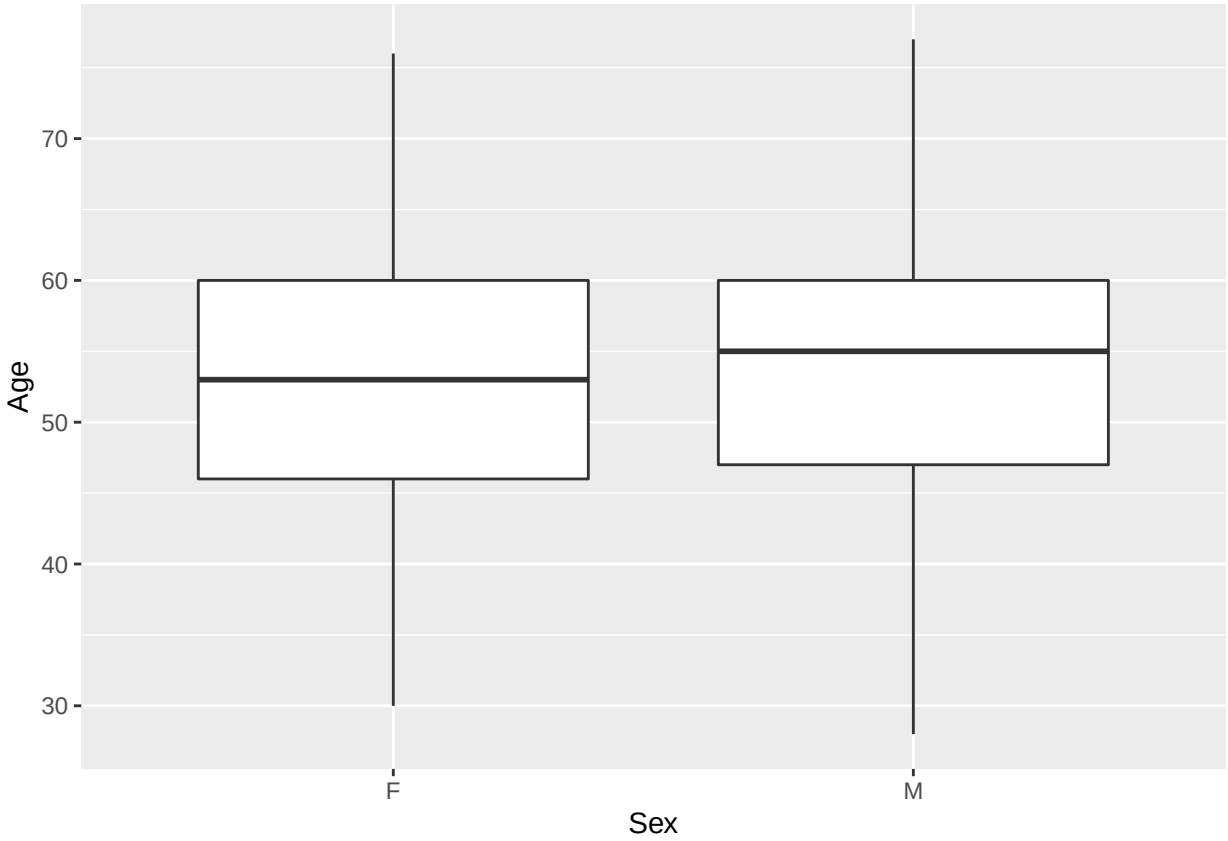
2.2 Data exploration and data visualization

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