¿Cómo realizar la limpieza y análisis de datos?

Autores: Eduardo Mora González y Diego Sánchez De La Fuente

Enero 2023

Contents

body { text-align: justify}

Instalamos y cargamos las librerías necesarias.

```
if (!require('readr')) install.packages('readr'); library('readr')
```

CARGA DEL FICHERO DE DATOS

```
datos <- read_csv("./fichero_original_datos.csv")
```

Ahora vamos a ver las estructura del juego de datos

```
str(datos)
```

```
## 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>
```

Vamos ahora a sacar estadísticas básicas

```
summary(datos)
```

```
##       Age            Sex            ChestPainType        RestingBP
##  Min.   :28.00   Length:918         Length:918         Min.   :  0.0
##  1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
##  Median :54.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :53.51                                         Mean   :132.4
##  3rd Qu.:60.00                                         3rd Qu.:140.0
##  Max.   :77.00                                         Max.   :200.0
##   Cholesterol      FastingBS       RestingECG           MaxHR
##  Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
##  1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0
##  Median :223.0   Median :0.0000   Mode  :character   Median :138.0
##  Mean   :198.8   Mean   :0.2331                      Mean   :136.8
##  3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
##  Max.   :603.0   Max.   :1.0000                      Max.   :202.0
##  ExerciseAngina        Oldpeak          ST_Slope          HeartDisease
##  Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000
##  Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
##  Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000
##                     Mean   : 0.8874                      Mean   :0.5534
##                     3rd Qu.: 1.5000                      3rd Qu.:1.0000
##                     Max.   : 6.2000                      Max.   :1.0000
```

Preprocesado y gestión de características

Valores nulos del conjunto de los datos

De tipo numérico

```
colSums(is.na(datos))
```

```
##           Age           Sex  ChestPainType      RestingBP    Cholesterol
##             0             0             0             0             0
```

```
##       FastingBS     RestingECG          MaxHR ExerciseAngina        Oldpeak
##               0              0              0              0              0
##       ST_Slope   HeartDisease
##               0              0
```

De tipo cadena

```
colSums(datos=="")
```

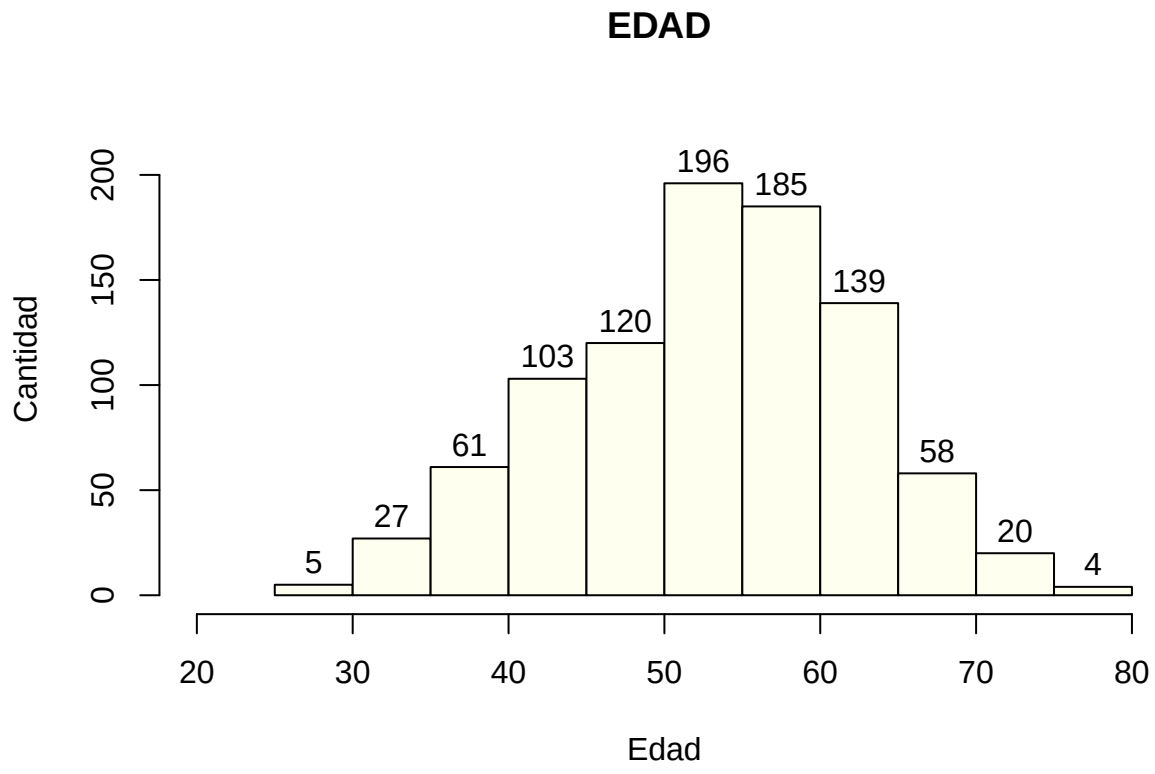
```
##             Age             Sex  ChestPainType      RestingBP    Cholesterol
##               0               0               0              0              0
##       FastingBS     RestingECG          MaxHR ExerciseAngina        Oldpeak
##               0               0               0              0              0
##       ST_Slope   HeartDisease
##               0               0
```

Como se puede comprobar, tenemos la "suerte" de no tener ningún valor nulo o vacío en los dos juegos de datos.

Normalización del conjunto de los datos

- **EDAD**

```
#Histograma de la característica edad del primer conjunto de datos
h1 <- hist(datos$Age, xlab="Edad", col="ivory", ylab="Cantidad", main="EDAD ", ylim = c(0, 225), xlim =
text(h1$mids,h1$counts,labels=h1$counts, adj=c(0.5, -0.5))
```



Como se puede observar, la franja de entre los 50 y 60 años son donde más datos existen, mientras que los extremos donde menos datos.

- **SEXO**

Normalizamos para tenerlo de tipo numérico todas la variables

```
#Cambiamos las letras por los números
datos$Sex [datos$Sex == "M"] <- 1
datos$Sex [datos$Sex == "F"] <- 0

#Pasamos de carácter a numérico
datos$Sex <- as.numeric(datos$Sex)
```

Una vez normalizada la característica , analizamos el conjunto de los datos contemplados en esta.

```
h1 <- hist(datos$Sex, xlab="Sexo", col=c("ivory", "lightcyan"), ylab="Cantidad", main="SEXO", breaks = 1
text(h1$mids,h1$counts,labels=h1$counts, adj=c(0.5, -0.5))
axis(1, at =c(0.25, 0.75), cex.axis=1, labels = c("Mujeres","Hombres" ))
axis(2)
```