

PRA2. Tipologia i cicle de vida de les dades

Autor: Daniel Rodríguez Morente

Maig 2023

Contents

1	Presentació del projecte i objectiu de l'anàlisi	1
2	Consideracions referents al dataset	1
3	Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?	2
3.1	Descripció de les variables	2
4	Integració i selecció de les dades d'interès a analitzar.	3
5	Neteja de les dades. Les dades contenen zeros o elements buits?	12
6	Anàlisi de les dades	13
7	Conclusions	14

1 Presentació del projecte i objectiu de l'anàlisi

El projecte que es desenvolupa a continuació consisteix en l'estudi de les causes que determinen la possibilitat una cardiopatia. En concret, es vol determinar si els diferents indicadors estudiats tenen una incidència diferent pels homes i les dones.

2 Consideracions referents al dataset

El dataset utilitzat conté informació de diferents indicadors mèdics de persones que han patit o no una cardiopatia.

Les dades han estat publicades per Rashik Rahman sota llicència CC0: Public Domain a [www.kaggle.com](https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset) i es pot accedir a les mateixes a través del següent enllaç: <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>

No s'han realitzat modificacions prèvies al conjunt de dades original.

3 Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

Carreguem el conjunt de dades i fem una revisió del contingut de les diferents variables

```
path = 'heart.csv'
dades <- read.csv(path, sep = ",")
str(dades)
```

```
## '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 ...
## $ cp       : int  3 2 1 1 0 0 1 1 2 2 ...
## $ trtbps   : int  145 130 130 120 120 120 140 120 172 150 ...
## $ chol     : int  233 250 204 236 354 192 294 263 199 168 ...
## $ fbs      : int  1 0 0 0 0 0 0 0 1 0 ...
## $ restecg  : int  0 1 0 1 1 1 0 1 1 1 ...
## $ thalachh : int  150 187 172 178 163 148 153 173 162 174 ...
## $ exng     : 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 ...
## $ slp      : int  0 0 2 2 2 1 1 2 2 2 ...
## $ caa      : int  0 0 0 0 0 0 0 0 0 0 ...
## $ thall    : int  1 2 2 2 2 1 2 3 3 2 ...
## $ output   : int  1 1 1 1 1 1 1 1 1 1 ...
```

Podem observar que es tracta d'un dataset amb 303 observacions i 14 variables, totes elles amb números enters excepte la variable oldpeak que conté dades decimals.

3.1 Descripció de les variables

- **age**. Edat de la persona
- **sex**. Sexe de la persona (1 = home; 0 = dona)
- **cp**. chest pain type Value (1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic)
- **trtbps**. Pressió arterial en repòr (en mm/Hg)
- **chol**. cholestoral in mg/dl fetched via BMI sensor
- **fbs**. (fasting blood sugar > 120 mg/dl) (1 = cert; 0 = fals)
- **restecg**. resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)
- **thalachh**. maximum heart rate achieved
- **exng**. exercise induced angina (1 = sí; 0 = no)
- **oldpeak**. Previous peak
- **slp**. Slope (0 = unsloping 1 = flat 2 = downsloping)
- **caa**. number of major vessels (0-4)
- **thall**. Thal rate (0 = null; 1 = fixed defect; 2 = normal; 3 = reversable defect)

- **output.** Target variable (0 = less chance of heart attack (< 50% diameter narrowing. less chance of heart disease); 1= more chance of heart attack (> 50% diameter narrowing. more chance of heart disease))

Extra (incorporar) - **Medical Definitions 1- Angina: chest pain due to reduced blood flow to the heart muscles. There're 3 types of angina: stable angina, unstable angina, and variant angina. To know more about angina click here: <https://www.nhs.uk/conditions/angina/#:~:text=Angina%20is%20chest%20pain%20caused,of%20these%20>

2- Cholesterol: a waxy substance found in the body cells and it belongs to a group of organic molecules called lipids. There are 3 types of cholesterol; high-density lipoprotein (HDL) and it's known as the "good cholesterol", low-density lipoprotein (LDL) known as the "bad cholesterol", and very-low-density lipoproteins (VLDL) and as the name implies, they're low dense particles that carry triglycerides in the blood.

3- ECG: short for electrocardiogram, it's a routine test usually done to check the heart's electrical activity.

4- ST depression: a type of ST-segment abnormality. the ST segment is the flat, isoelectric part of the ECG and it represents the interval between ventricular depolarization and repolarization. For more details check this link: <https://litfl.com/st-segment-ecg-library/>.

5- Thalassemia: it's a genetic blood disorder that is characterized by a lower rate of hemoglobin than normal.

4 Integració i selecció de les dades d'interès a analitzar.

Revisem la distribució de les diferents variables

```
summary(dades)
```

```
##          age          sex          cp          trtbps
## 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
##          chol          fbs          restecg          thalachh
## 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
##          exng          oldpeak          slp          caa
## 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
##          thall          output
## 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
```

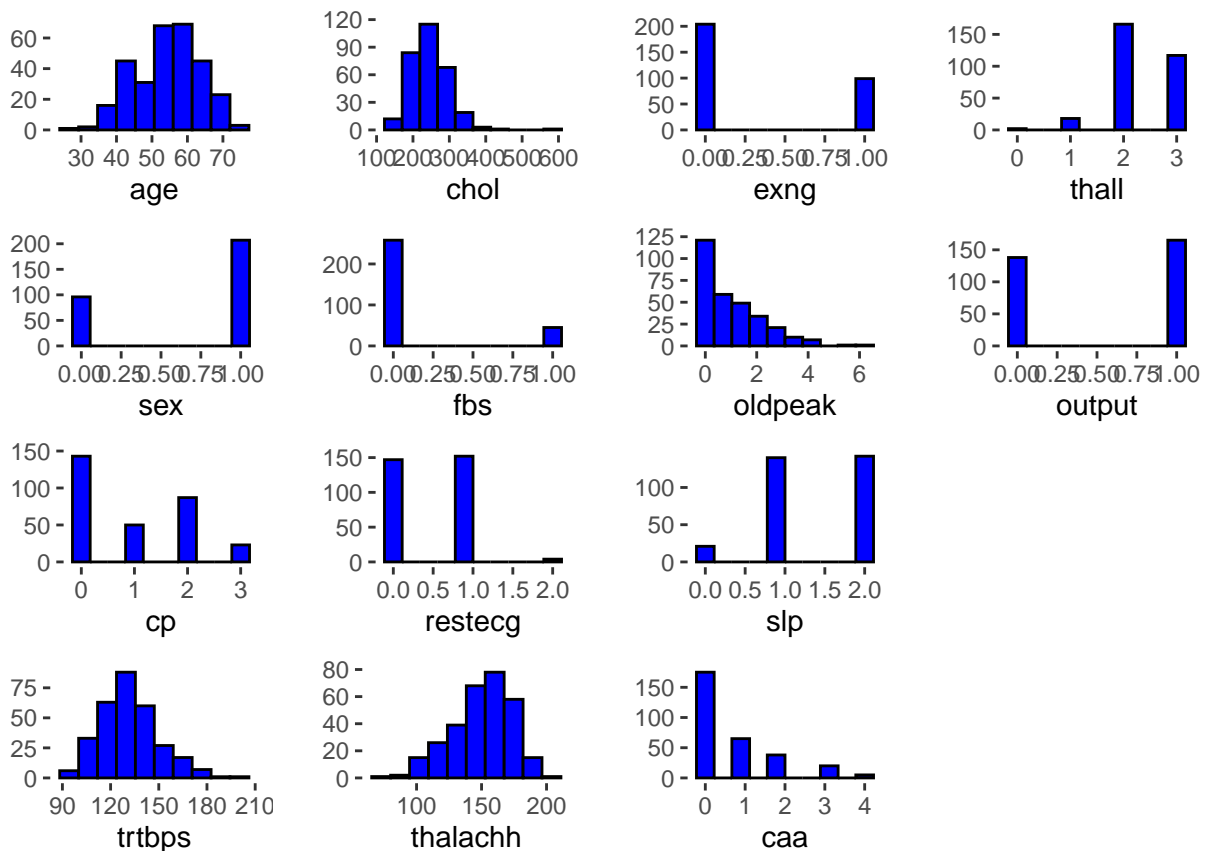
A priori, no observem que hi hagi valor perduts, però més endavant farem una comprovació adhoc.

Fem una representació de les diferents variables per tal de facilitar la revisió prèvia del dataset

```
histogrames_num <- list()
variables_num <- names(dades)
dades_num <- dades %>% select(all_of(variables_num))

for(i in 1:ncol(dades_num)){
  var <- names(dades_num)[i]
  grafic <- ggplot(dades_num, aes_string(x = var)) +
    geom_histogram(bins = 10, fill = "blue", color = "black") +
    labs(y = "") +
    theme(panel.grid = element_blank(), panel.background = element_blank())
  histogrames_num[[i]] <- grafic
}

multiplot(plotlist = histogrames_num, cols = 4)
```



Apreciem que hi ha més informació d'homes que de dones i que el número de registres amb output igual a 1 és lleugerament superior al valor 0.

Revisem si tenim registres amb idèntics valors a totes les variables per tal de valorar si tenim registres duplicats

```
dim(unique(dades))
```

```
## [1] 302 14
```

Comprovem que hi ha 302 registres diferents, per la qual cosa, donat el nivell d'especificitat de les dades, considerem que hi ha un registre repetit.

Eliminem el registre repetit i conservem la resta donat que tenim un número de registres perfectament gestionable i, per tant no és necessari plantejar agrupacions que facilitin l'ús del dataset

```
dades <- unique(dades)
```

Donat que l'estudi es base en la diferència entre homes i dones, ens interessa comprobar quina informació tenim per cada grup

```
print('Distribució entre homes i dones en valors absoluts:')
```

```
## [1] "Distribució entre homes i dones en valors absoluts:"
```

```
print(addmargins(table(dades$sex, dades$output)))
```

```
##
##      0    1 Sum
## 0    24   72  96
## 1   114   92 206
## Sum 138 164 302
```

```
print('Pes relatiu de cada sexe dins el valor de la variable output:')
```

```
## [1] "Pes relatiu de cada sexe dins el valor de la variable output:"
```

```
print(round(prop.table(table(dades$sex, dades$output), 2), 2))
```

```
##
##      0    1
## 0 0.17 0.44
## 1 0.83 0.56
```

```
print('Pes relatiu de la variable output dins de cada sexe:')
```

```
## [1] "Pes relatiu de la variable output dins de cada sexe:"
```

```
print(round(prop.table(table(dades$sex, dades$output), 1), 2))
```

```
##
##      0    1
## 0 0.25 0.75
## 1 0.55 0.45
```

Observem que tenim 207 homes i 96 dones i que la distribució del camp output dins de cada grup és diferent, tenint més pes en el valor 1 en dones que en homes.

O POTSER NO. VARLORAR SI APORTA ALGUNA COSA

Aplicarem una anàlisi de components principals per tal valorar si podem treballar amb menys variables

```
dades_acp <- prcomp(dades, center = TRUE, scale = TRUE)
summary(dades_acp)
```

```
## Importance of components:
##              PC1    PC2    PC3    PC4    PC5    PC6    PC7
## Standard deviation  1.8192 1.2557 1.1068 1.10095 1.01153 0.98461 0.93027
## Proportion of Variance 0.2364 0.1126 0.0875 0.08658 0.07308 0.06925 0.06181
## Cumulative Proportion 0.2364 0.3490 0.4365 0.52310 0.59619 0.66543 0.72725
##              PC8    PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation  0.88297 0.84761 0.78999 0.72720 0.65579 0.61065 0.60382
## Proportion of Variance 0.05569 0.05132 0.04458 0.03777 0.03072 0.02664 0.02604
## Cumulative Proportion 0.78294 0.83425 0.87883 0.91660 0.94732 0.97396 1.00000
```

Observem que, tot i que hi ha dues components principals que expliquen juntes un 34,9% per la variància, aquesta està molt repartida i necessitem 13 dels 14 components per explicar el 95% de la variància.

4.0.1 Normalització de les dades

Tenim quatre variables numèriques que ens pot interessar normalitzar per tal que siguin comparables en el nostre estudi. Primer de tot, comprovarem si la distribució de les variables trtbps, chol, thalachh i oldpeak és o no normal aplicant el test de Shapiro

```
shapiro_trtbps <- shapiro.test(dades$trtbps)
print(shapiro_trtbps)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades$trtbps
## W = 0.96573, p-value = 1.419e-06
```

```
shapiro_chol <- shapiro.test(dades$chol)
print(shapiro_chol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades$chol
## W = 0.94658, p-value = 5.196e-09
```

```
shapiro_thalachh <- shapiro.test(dades$thalachh)
print(shapiro_thalachh)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades$thalachh
## W = 0.97679, p-value = 8.268e-05
```

```
shapiro_oldpeak <- shapiro.test(dades$oldpeak)
print(shapiro_oldpeak)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades$oldpeak
## W = 0.84522, p-value < 2.2e-16
```

En els quatre casos, observant el valor de p podem dir que es rebutja la hipòtesi nul·la i, per tant, no es distribueixen com una normal.

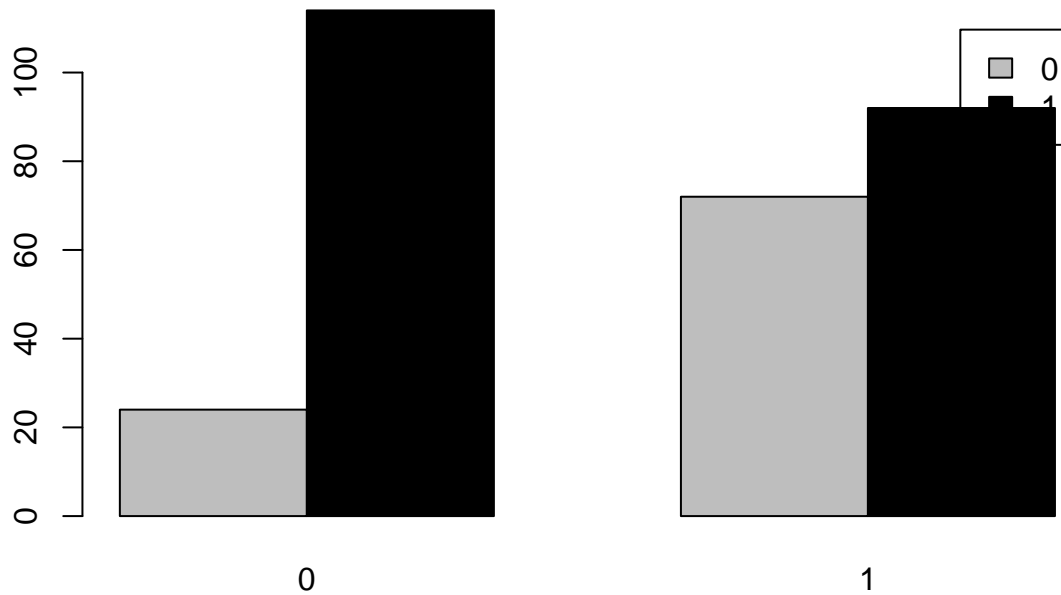
```
print(table(dades$sex, dades$output))
```

```
##
##      0      1
## 0  24  72
## 1 114  92
```

```
print(prop.table(table(dades$sex, dades$output), 2))
```

```
##
##           0           1
## 0 0.1739130 0.4390244
## 1 0.8260870 0.5609756
```

```
barplot(table(dades$sex, dades$output), beside = TRUE, col = c("grey", "black"), legend = rownames(table(dades$sex, dades$output)))
```



```
print(prop.table(table(dades$chol, dades$output)))
```

```
##
##           0           1
## 126 0.000000000 0.003311258
## 131 0.003311258 0.000000000
## 141 0.000000000 0.003311258
## 149 0.003311258 0.003311258
## 157 0.000000000 0.003311258
## 160 0.000000000 0.003311258
## 164 0.003311258 0.000000000
## 166 0.003311258 0.000000000
## 167 0.003311258 0.000000000
## 168 0.000000000 0.003311258
## 169 0.003311258 0.000000000
## 172 0.003311258 0.000000000
## 174 0.003311258 0.000000000
## 175 0.000000000 0.006622517
## 176 0.003311258 0.000000000
## 177 0.006622517 0.006622517
## 178 0.000000000 0.003311258
## 180 0.000000000 0.003311258
## 182 0.000000000 0.003311258
## 183 0.000000000 0.003311258
## 184 0.003311258 0.000000000
```



```

## 185 0.003311258 0.000000000
## 186 0.000000000 0.003311258
## 187 0.003311258 0.000000000
## 188 0.006622517 0.000000000
## 192 0.000000000 0.006622517
## 193 0.003311258 0.003311258
## 195 0.000000000 0.003311258
## 196 0.000000000 0.006622517
## 197 0.006622517 0.013245033
## 198 0.003311258 0.003311258
## 199 0.000000000 0.009933775
## 200 0.003311258 0.000000000
## 201 0.000000000 0.009933775
## 203 0.006622517 0.003311258
## 204 0.006622517 0.013245033
## 205 0.003311258 0.003311258
## 206 0.006622517 0.000000000
## 207 0.003311258 0.003311258
## 208 0.000000000 0.006622517
## 209 0.000000000 0.006622517
## 210 0.000000000 0.003311258
## 211 0.000000000 0.013245033
## 212 0.013245033 0.003311258
## 213 0.000000000 0.006622517
## 214 0.000000000 0.006622517
## 215 0.000000000 0.003311258
## 216 0.003311258 0.003311258
## 217 0.003311258 0.000000000
## 218 0.006622517 0.000000000
## 219 0.003311258 0.006622517
## 220 0.000000000 0.009933775
## 221 0.000000000 0.006622517
## 222 0.000000000 0.006622517
## 223 0.003311258 0.006622517
## 224 0.003311258 0.000000000
## 225 0.006622517 0.000000000
## 226 0.000000000 0.013245033
## 227 0.000000000 0.006622517
## 228 0.003311258 0.003311258
## 229 0.009933775 0.000000000
## 230 0.009933775 0.000000000
## 231 0.006622517 0.003311258
## 232 0.003311258 0.003311258
## 233 0.003311258 0.009933775
## 234 0.006622517 0.013245033
## 235 0.000000000 0.006622517
## 236 0.003311258 0.006622517
## 237 0.003311258 0.000000000
## 239 0.006622517 0.006622517
## 240 0.000000000 0.013245033
## 241 0.003311258 0.000000000
## 242 0.000000000 0.003311258
## 243 0.006622517 0.006622517
## 244 0.003311258 0.006622517

```

245 0.000000000 0.009933775
246 0.006622517 0.003311258
247 0.003311258 0.003311258
248 0.003311258 0.003311258
249 0.009933775 0.000000000
250 0.000000000 0.009933775
252 0.000000000 0.003311258
253 0.003311258 0.003311258
254 0.013245033 0.003311258
255 0.003311258 0.003311258
256 0.006622517 0.003311258
257 0.000000000 0.003311258
258 0.006622517 0.003311258
259 0.003311258 0.000000000
260 0.003311258 0.003311258
261 0.003311258 0.003311258
262 0.000000000 0.003311258
263 0.003311258 0.006622517
264 0.003311258 0.003311258
265 0.000000000 0.006622517
266 0.003311258 0.003311258
267 0.003311258 0.003311258
268 0.003311258 0.003311258
269 0.006622517 0.009933775
270 0.003311258 0.003311258
271 0.000000000 0.006622517
273 0.003311258 0.003311258
274 0.009933775 0.000000000
275 0.003311258 0.003311258
276 0.003311258 0.000000000
277 0.000000000 0.006622517
278 0.000000000 0.003311258
281 0.003311258 0.000000000
282 0.013245033 0.000000000
283 0.006622517 0.003311258
284 0.003311258 0.000000000
286 0.006622517 0.000000000
288 0.006622517 0.003311258
289 0.006622517 0.000000000
290 0.003311258 0.000000000
293 0.003311258 0.000000000
294 0.003311258 0.003311258
295 0.000000000 0.006622517
298 0.003311258 0.003311258
299 0.006622517 0.000000000
300 0.003311258 0.000000000
302 0.000000000 0.006622517
303 0.000000000 0.009933775
304 0.003311258 0.003311258
305 0.003311258 0.000000000
306 0.000000000 0.003311258
307 0.003311258 0.000000000
308 0.000000000 0.006622517
309 0.006622517 0.003311258

```
## 311 0.003311258 0.000000000
## 313 0.000000000 0.003311258
## 315 0.003311258 0.003311258
## 318 0.003311258 0.003311258
## 319 0.003311258 0.000000000
## 321 0.000000000 0.003311258
## 322 0.003311258 0.000000000
## 325 0.000000000 0.006622517
## 326 0.003311258 0.000000000
## 327 0.003311258 0.000000000
## 330 0.006622517 0.000000000
## 335 0.006622517 0.000000000
## 340 0.000000000 0.003311258
## 341 0.003311258 0.000000000
## 342 0.000000000 0.003311258
## 353 0.003311258 0.000000000
## 354 0.000000000 0.003311258
## 360 0.000000000 0.003311258
## 394 0.000000000 0.003311258
## 407 0.003311258 0.000000000
## 409 0.003311258 0.000000000
## 417 0.000000000 0.003311258
## 564 0.000000000 0.003311258
```

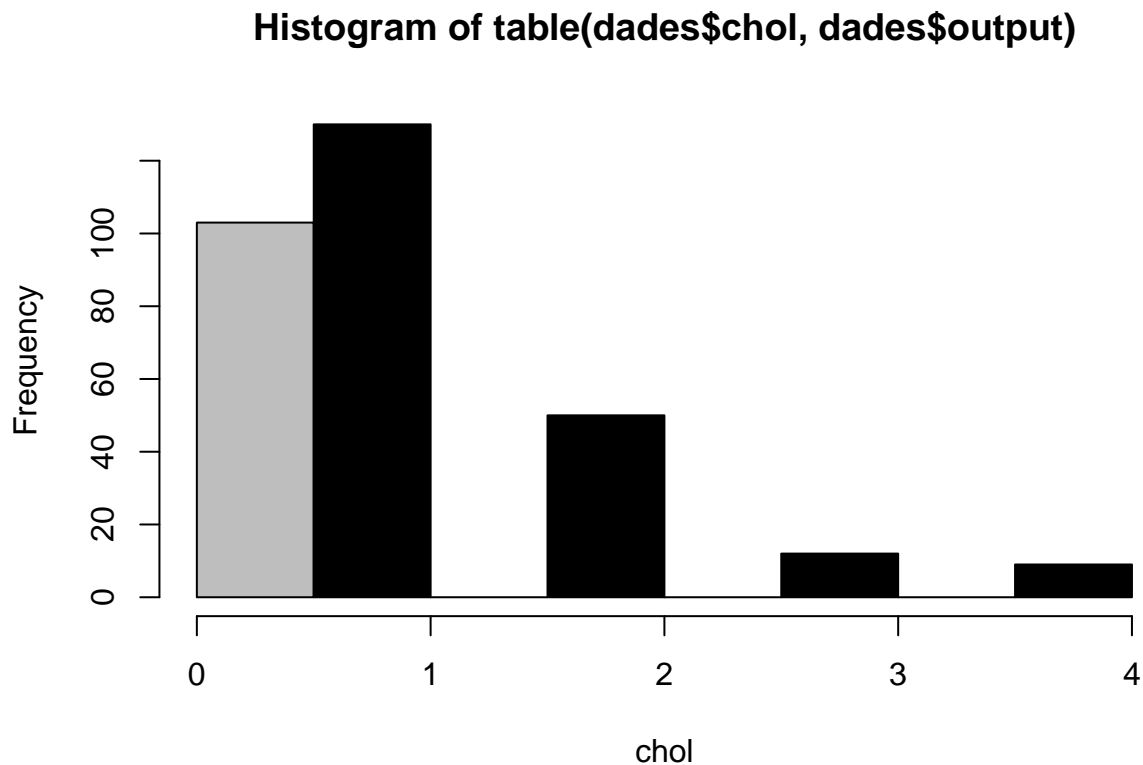
```
hist(table(dades$chol, dades$output), col = c("grey", "black"), legend = rownames(table(dades$chol, dades$output)))
```

```
## Warning in plot.window(xlim, ylim, "", ...): "legend" is not a graphical
## parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "legend" is not a graphical parameter
```

```
## Warning in axis(1, ...): "legend" is not a graphical parameter
```

```
## Warning in axis(2, ...): "legend" is not a graphical parameter
```



5 Neteja de les dades. Les dades contenen zeros o elements buits?

Tot i que amb el resum del dataset no apareixien valors perduts, fem una nova comprovació per evitar errors d'apreciació

```
colSums(is.na(dades))
```

```
##      age      sex      cp  trtbps      chol      fbs  restecg  thalachh
##      0       0       0       0       0       0       0       0
##  exng  oldpeak  slp      caa      thall      output
##      0       0       0       0       0       0
```

```
colSums(dades=="")
```

```
##      age      sex      cp  trtbps      chol      fbs  restecg  thalachh
##      0       0       0       0       0       0       0       0
##  exng  oldpeak  slp      caa      thall      output
##      0       0       0       0       0       0
```

Es confirma que no tenim valors perduts, per tant no em de fer cap modificació al dataset.

Fem una revisió de la possible existència de valors extrems

```

var_out <- c()
for (i in 1:ncol(dades)){
  outl <- boxplot.stats(dades[,i])$out
  if (!length(outl)==0){var_out <- c(var_out, i)}
}
print(names(dades)[var_out])

```

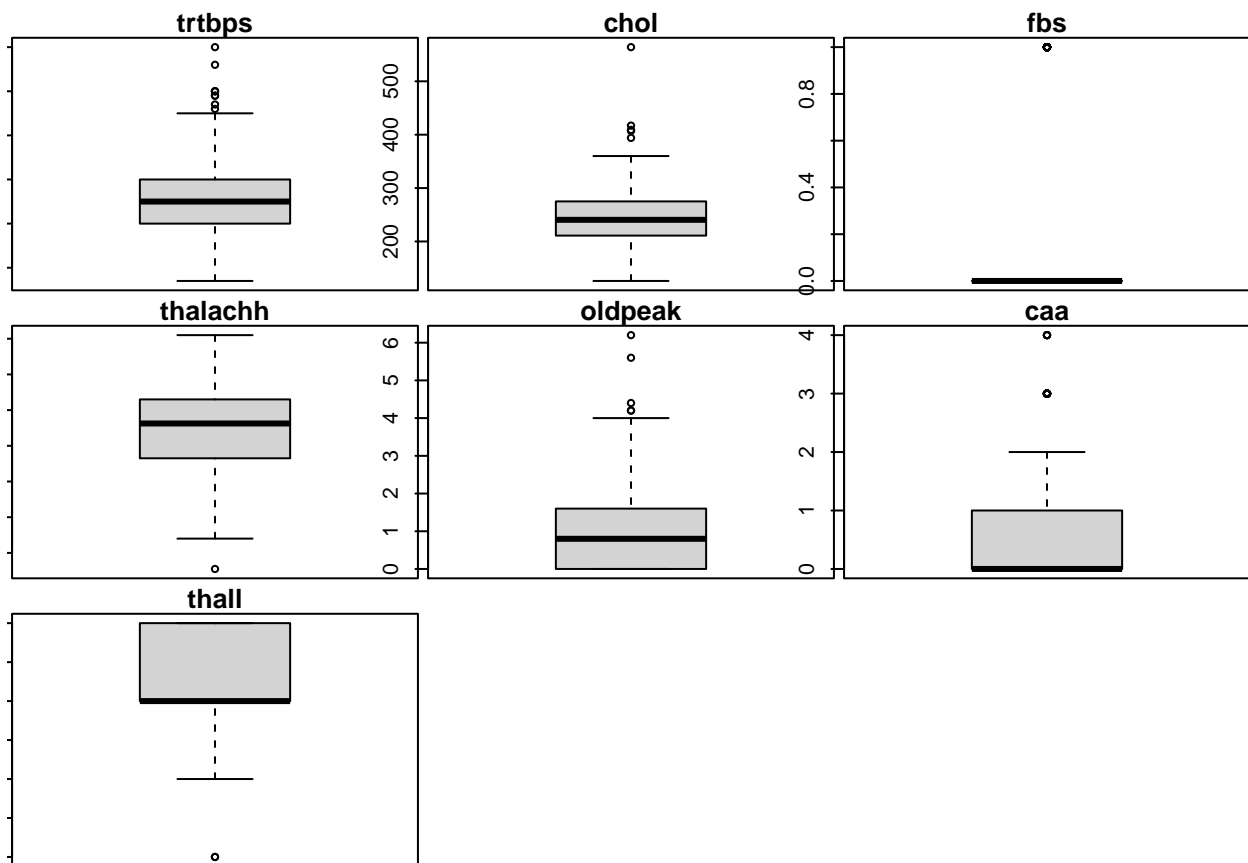
```
## [1] "trtbps"    "chol"      "fbs"       "thalachh"  "oldpeak"   "caa"       "thall"
```

Veiem que tenim 7 variables amb valors extrems

```

par(mfrow = c(3,3), mar = c(0, 0, 1, 0) + 0.2)
for (i in var_out) {
  boxplot(dades[, i], main = colnames(dades)[i])
}

```



6 Anàlisi de les dades

Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?).

Comprovació de la normalitat i homogeneïtat de la variància.

Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

7 Conclusions

A partir dels resultats obtinguts, quines són les conclusions?

Els resultats permeten respondre al problema?