

Pràctica 2: Com realitzar la neteja i l'anàlisi de dades?

Maria Magdalena Dicheva

13/1/2023

1. Descripció del dataset

Per aquesta pràctica, s'ha escollit treballar amb el dataset proposat per identificar els factors de risc que influeixen en els atacs cardíacs. Aquest estudi és important per concienciar a la població i que els individus siguin capaços d'evitar conductes de risc i de detectar símptomes que indiquin quan un atac és imminent o ja s'està produint. Així, poden demanar ajuda i reduir la gravetat i les conseqüències al mínim.

Abans del processament del dataset, aquest conté 14 columnes:

- Age: edat del pacient mesurada en anys
- Sex: sexe del pacient (1 = home, 0 = dona)
- exang: angina induïda per exercici (1 = si, 0 = no)
- ca: nombre de vasos sanguinis (0-3)
- cp: tipus de dolor al pit (4 valors possibles)
- trtbps: pressió sanguínia en repòs mesurada en mm/Hg
- chol: nivell de colesterol mesurat en mg/dl
- fbs: nivell de sucre en sang en dejú > 120mg/dl (1 = true, 0 = false)
- rest_ecg: resultats electrocardiogràfics en repòs (2 valors possibles)
- thalach: freqüència cardíaca màxima a la que s'ha arribat
- target: 0 = probabilitat menor d'atac cardíac, 1 = probabilitat major d'atac cardíac
- oldpeak: depressió del segment ST induïda per exercici relatiu
- slp: tipus d'inclinació del segment ST durant el pic màxim d'exercici (3 valors possibles)

Les dades es poden trobar a Kaggle: <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>.

2. Integració i selecció

```
library(readr)
```

```
#Carreguem les dades i les visualitzem
```

```
data <- read_csv("heart.csv", show_col_types = FALSE)
```

```
head(data, 10)
```

```
## # A tibble: 10 x 14
```

```
##   age  sex  cp trtbps  chol  fbs restecg thalachh  exng oldpeak  slp
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1   63    1    3   145   233    1     0     150     0     2.3    0
## 2   37    1    2   130   250    0     1     187     0     3.5    0
## 3   41    0    1   130   204    0     0     172     0     1.4    2
## 4   56    1    1   120   236    0     1     178     0     0.8    2
## 5   57    0    0   120   354    0     1     163     1     0.6    2
## 6   57    1    0   140   192    0     1     148     0     0.4    1
```

```
## 7      56      0      1     140     294      0      0      153      0      1.3      1
## 8      44      1      1     120     263      0      1      173      0      0      2
## 9      52      1      2     172     199      1      1      162      0      0.5      2
## 10     57      1      2     150     168      0      1      174      0      1.6      2
## # ... with 3 more variables: caa <dbl>, thall <dbl>, output <dbl>
```

```
dim(data)
```

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

```
str(data)
```

```
## spec_tbl_df [303 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ age      : num [1:303] 63 37 41 56 57 57 56 44 52 57 ...
## $ sex      : num [1:303] 1 1 0 1 0 1 0 1 1 1 ...
## $ cp       : num [1:303] 3 2 1 1 0 0 1 1 2 2 ...
## $ trtbps   : num [1:303] 145 130 130 120 120 120 140 120 172 150 ...
## $ chol     : num [1:303] 233 250 204 236 354 192 294 263 199 168 ...
## $ fbs      : num [1:303] 1 0 0 0 0 0 0 0 1 0 ...
## $ restecg  : num [1:303] 0 1 0 1 1 1 0 1 1 1 ...
## $ thalachh : num [1:303] 150 187 172 178 163 148 153 173 162 174 ...
## $ exng     : num [1:303] 0 0 0 0 1 0 0 0 0 0 ...
## $ oldpeak  : num [1:303] 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
## $ slp      : num [1:303] 0 0 2 2 2 1 1 2 2 2 ...
## $ caa      : num [1:303] 0 0 0 0 0 0 0 0 0 0 ...
## $ thall    : num [1:303] 1 2 2 2 2 1 2 3 3 2 ...
## $ output   : num [1:303] 1 1 1 1 1 1 1 1 1 1 ...
## - attr(*, "spec")=
## .. cols(
## ..   age = col_double(),
## ..   sex = col_double(),
## ..   cp = col_double(),
## ..   trtbps = col_double(),
## ..   chol = col_double(),
## ..   fbs = col_double(),
## ..   restecg = col_double(),
## ..   thalachh = col_double(),
## ..   exng = col_double(),
## ..   oldpeak = col_double(),
## ..   slp = col_double(),
## ..   caa = col_double(),
## ..   thall = col_double(),
## ..   output = col_double()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
summary(data)
```

##	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
```

Per arribar a l'objectiu proposat, s'han escollit...

3. Neteja de les dades

3.1 Zeros i elements buits

```
# Comprovem si les dades contenen elements nuls
any(is.null(data))
```

```
## [1] FALSE
```

Les dades no contenen cap dada nul·la, pel que no s'ha de fer res.

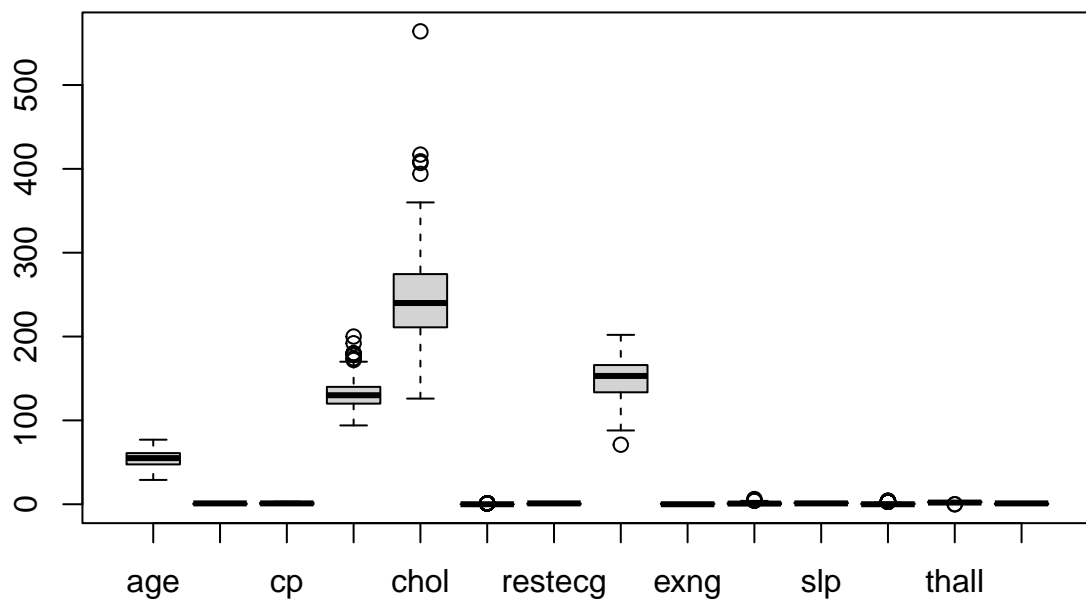
```
# Comprovem si contenen element buits
any(is.na(data))
```

```
## [1] FALSE
```

Les dades no contenen cap element buit, pel que es deixa com està.

3.2 Valors extrems

```
# Detectem valors extrems amb boxplots
data.bp <- boxplot(data)
```



```
data.bp$out
```

```
## [1] 172.0 178.0 180.0 180.0 200.0 174.0 192.0 178.0 180.0 417.0 564.0 394.0
## [13] 407.0 409.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
## [25] 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
## [37] 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
## [49] 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 71.0
## [61] 4.2 6.2 5.6 4.2 4.4 3.0 4.0 3.0 3.0 4.0 4.0 4.0
## [73] 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0
## [85] 3.0 4.0 3.0 3.0 3.0 3.0 0.0 0.0
```