

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

# Renombrem les variables perquè s'entenguin millor
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
## filter, lag

## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union

data = rename(data, c("chest_pain"="cp", "cholesterol"="chol",
                      "exercise_induced_angina"="exng", "vessels_numbers"="caa",
                      "blood_pressure"="trtbps", "fasting_blood_sugar"="fbs",
                      "max_heart_rate_achieved"="thalachh", "stress_test"="thall"))

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 ...
## $ chest_pain : num [1:303] 3 2 1 1 0 0 1 1 2 2 ...
## $ blood_pressure : num [1:303] 145 130 130 120 120 140 140 120 172 150 ...
## $ cholesterol : num [1:303] 233 250 204 236 354 192 294 263 199 168 ...
## $ fasting_blood_sugar : 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 ...
## $ max_heart_rate_achieved: num [1:303] 150 187 172 178 163 148 153 173 162 174 ...
## $ exercise_induced_angina: 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 ...
## $ vessels_numbers : num [1:303] 0 0 0 0 0 0 0 0 0 0 ...
## $ stress_test : 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      chest_pain      blood_pressure
## 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
## cholesterol  fasting_blood_sugar  restecg      max_heart_rate_achieved
## 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
## exercise_induced_angina  oldpeak      slp      vessels_numbers
## 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
## stress_test      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, de les 14 variables originals s'han escollit 7 que podrien jugar un paper important en el que es vol avaluar:

- age
- sex
- blood_pressure
- cholesterol
- fasting_blood_sugar
- exercise_induced_angina
- max_heart_rate_achieved

```
# Creem el nou dataset amb les variables mencionades
```

```
data1 <- data.frame(data$age, data$sex, data$blood_pressure, data$cholesterol, data$fasting_blood_sugar
```

```
# Visualitzem el nou dataset
```

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

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

```
##      age      sex chest_pain blood_pressure cholesterol fasting_blood_sugar restecg
```

```
##      <dbl> <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1      63      1          3          145      233          1          0
## 2      37      1          2          130      250          0          1
## 3      41      0          1          130      204          0          0
## 4      56      1          1          120      236          0          1
## 5      57      0          0          120      354          0          1
## 6      57      1          0          140      192          0          1
## 7      56      0          1          140      294          0          0
## 8      44      1          1          120      263          0          1
## 9      52      1          2          172      199          1          1
## 10     57      1          2          150      168          0          1
## # ... with 7 more variables: max_heart_rate_achieved <dbl>,
## #   exercise_induced_angina <dbl>, oldpeak <dbl>, slp <dbl>,
## #   vessels_numbers <dbl>, stress_test <dbl>, output <dbl>
```

3. Neteja de les dades

3.1 Zeros i elements buits

```
# Comprovem si les dades contenen elements nuls
any(is.null(data1))
```

```
## [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(data1))
```

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

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

Comprovem també si el dataset conté dades duplicades mitjançant l'extracció de dades úniques:

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5      v purrr 0.3.4
## v tibble 3.1.6       v stringr 1.4.1
## v tidyr 1.2.0        v forcats 0.5.1
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

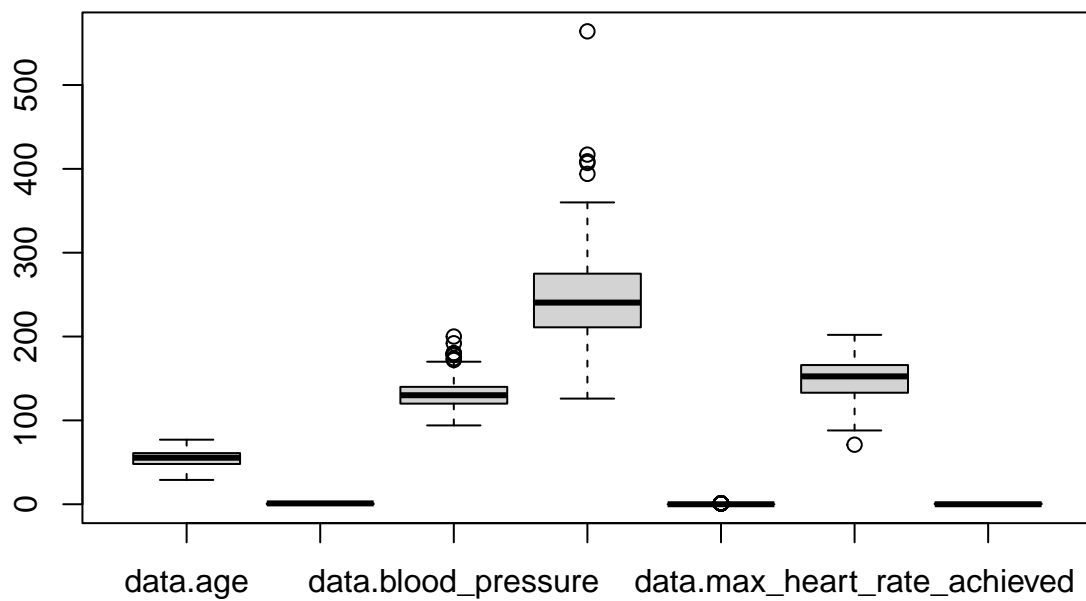
```
data2 <- unique(data1)
dim(data2)
```

```
## [1] 302 7
```

Com es pot veure, les files han passat de ser 303 a 302. Per tant, s'ha eliminat un registre.

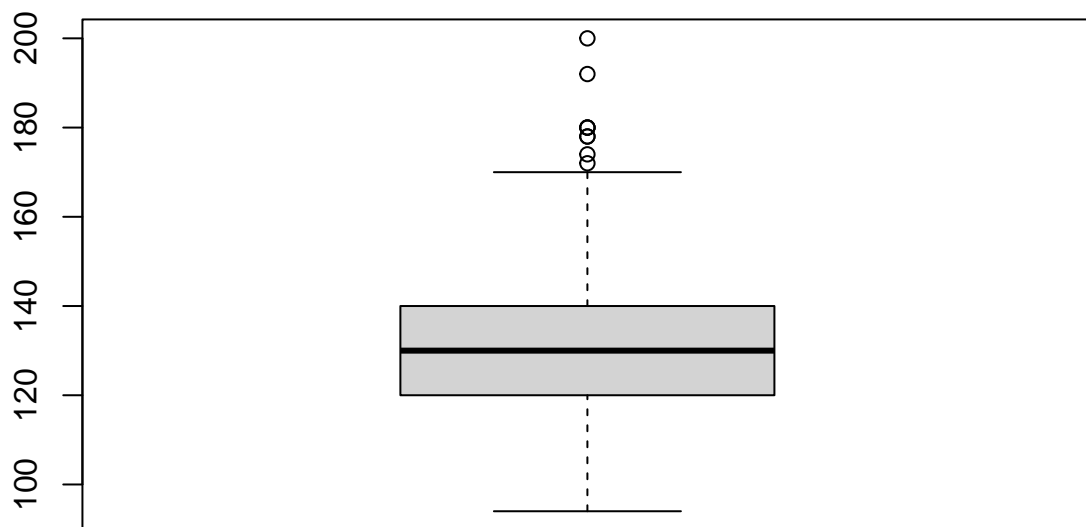
3.2 Valors extrems

```
# Detectem valors extrems amb boxplots
boxplot(data2)
```

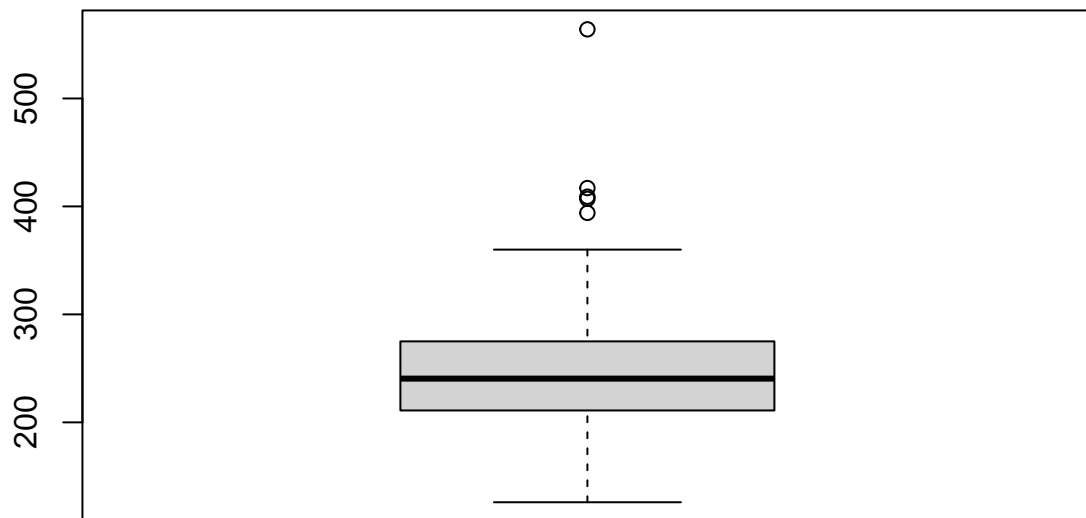


Gràcies als boxplot, es pot veure 4 variables que presenten outliers i quines són. Per tant, les mirem més de prop:

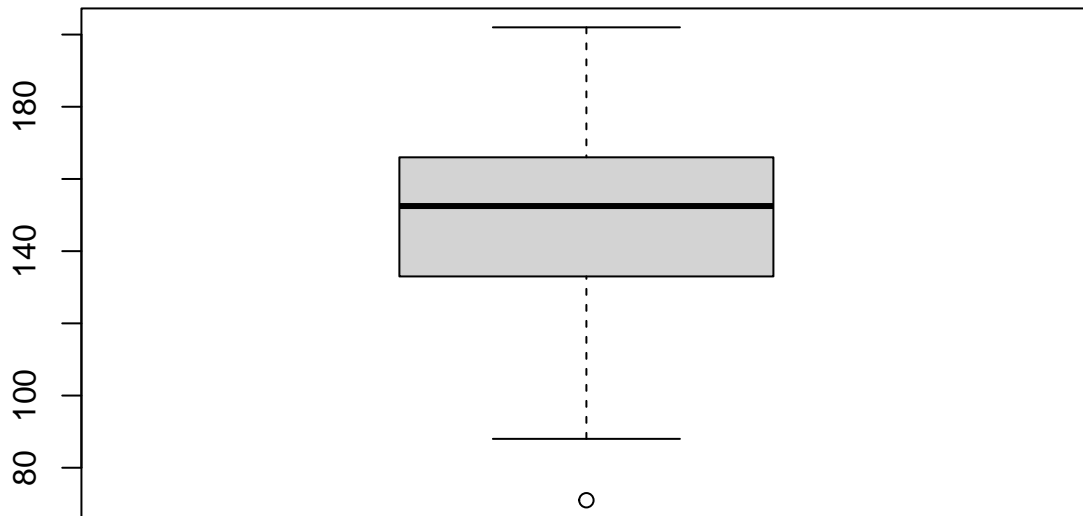
```
blood_pressure_bp <- boxplot(data2$data.blood_pressure)
```



```
cholesterol_bp <- boxplot(data2$data.cholesterol)
```



```
max_heart_rate_achieved_bp <- boxplot(data2$data$max_heart_rate_achieved)
```



```
blood_pressure_outliers <- blood_pressure_bp$out
```

```
cholesterol_outliers <- cholesterol_bp$out
```

```
max_heart_rate_achieved_outliers <- max_heart_rate_achieved_bp$out
```

Com que tenim molts registres i amb les dades que tenim no es podria veure el motiu de perquè aquests valors són extrems, s'ha optat per eliminar-los.

He provat diferents funcions, però no m'he ensortit. Estaré atenta a la publicació de la solució.

Extracció de les dades a un fitxer .csv

Exportem les dades a un fitxer .csv que es trobarà a github:

```
write.csv(data2, file = "C:/Users/adril/Desktop/UOC/Semestre 3/Tipologia i cicle de les dades/PRA2/dades.csv")
```