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

Carlos Martínez Torró (cmtorro@uoc.edu) i Xavier Roca Canals (xrocaca@uoc.edu)

04/01/2023

Índex de la pràctica

Exercici 1: Descripció del dataset

El nostre dataset recopila desenes de mètriques estadístiques de més de 70 temporades de la NBA i de 25 anys de la WNBA, és a dir, des dels inicis respectius de cada lliga. Les dades es van recollir del web Basketball Reference (propietat del grup Sports Reference) i estan classificades per equip i per temporada, amb un total de 1900 observacions i 53 variables. És a dir, a cada fila trobarem com li ha anat a un equip, des del punt de vista mètric, en una temporada en concret. L'ampla disponibilitat de variables ens permet anar des dels anàlisis més superficials (victòries, derrotes, punts a favor...) a altres més complexos (ritme de joc, eficiència dels llançaments...).

En una era com la nostra en la que les dades s'han infiltrat arreu, ens preguntem si també han arribat a la lliga de bàsquet per excel · lència. L'objectiu que perseguim amb la creació i l'actual anàlisi d'aquest dataset és ambiciós: es pot explicar l'evolució de la NBA a través de les estadístiques? Hi ha algun patró que segueixin aquells equips més exitosos? I estan dirigint aquests equips l'evolució de l'esport?

Exercici 2: Integració i selecció de les dades

A la Pràctica 1 ja vam realitzar una integració de diferents datasets creats a partir del *web scraping*, ja que estàvem interessats en diferents taules i vam optar per fer un dataframe per cadascuna d'elles i després unir els dataframes resultants per les columnes comunes (amb la funció `merge` de la llibreria `pandas`). Per tant, el dataset lliurat a la PRA1 és el que carregarem:

```
# Carreguem les dades amb read.csv:
df <- read.csv("../dataset/NBA_WNBA_statistical_evolution.csv")
```

Com posar la sortida de les funcions head, summary o str pot allargar molt el document i resultar improductiu degut a l'elevat nombre de variables que tenim, podem crear una taula per veure la informació bàsica del dataset: quantes observacions tenim, quantes variables, quines són numèriques, etc.

```
# Carreguem kableExtra per fer la taula amb la informació bàsica. També
# carreguem dplyr per ajudar-nos amb el filtratge de les dades:
library(kableExtra)
library(dplyr)

# Traiem les dades bàsiques del dataset:
cols <- ncol(df)
rows <- nrow(df)

# Informació del tipus de variables:
catCols <- length(names(df)[sapply(df, is.character)])
numCols <- length(names(df)[sapply(df, is.numeric)])

# I informació respecte als valors nuls, primer per cada fila:
compObs <- nrow(df %>% filter(complete.cases(.)))
compObsPct <- paste(compObs, '/', rows, ' = ',
                    round(((compObs/rows) * 100), 2), '%')

# I per les variables:
comVars <- df %>% select_if(~ all(!is.na(.))) %>% length()
comVarsPct <- paste(comVars, '/', cols, ' = ',
                    round(((comVars/cols) * 100), 2), '%')
```

```r
# Definim el que hi haurà al dataframe:
mets <- c("Nombre d'observacions", "Nombre de variables", "Variables numèriques",
          "Variables categòriques", "Casos complets (%) (observacions sense NAs)",
          "Variables completes (%) (variables sense NAs)")
vals <- c(rows, cols, numCols, catCols, compObsPct, comVarsPct)

# Ho passem tot a un dataframe i fem la taula:
resum <- data.frame(mets, vals)
kable(resum, booktabs = TRUE, caption = "Mètriques bàsiques del dataset",
      col.names = c("Mètrica", "Valor")) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

Table 1: Mètriques bàsiques del dataset

Mètrica	Valor
Nombre d'observacions	1900
Nombre de variables	53
Variables numèriques	48
Variables categòriques	5
Casos complets (%) (observacions sense NAs)	1522 / 1900 = 80.11 %
Variables completes (%) (variables sense NAs)	31 / 53 = 58.49 %

Exercici 3: Neteja de les dades

3.1: Les dades contenen zeros o elements buits?

3.2: Identifica i gestiona els valors extrems.

Exercici 4: Anàlisi de les dades

4.1: Selecció dels grups de dades que es volen analitzar/comparar.

4.2: Comprovació de la normalitat i homogeneïtat de la variància.

4.3: Aplicació de proves estadístiques per comparar els grups de dades. Aplicar almenys tres mètodes d'anàlisi diferents.

Exercici 5: Representació dels resultats.

Exercici 6: Resolució del problema.

Exercici 7: Codi.

Exercici 8: Vídeo.