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

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

08/01/2023

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.

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, quantes són numèriques, quantes categòriques, etc.

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 %

Com podem veure a la taula, hi ha un gran percentatge de casos complets (és a dir, equips amb tota la informació de l temporada completa), mentre que aproximadament el 60% de les variables no tenen cap valor NA. Podem fer una ullada a quines són les variables amb més valors perduts del nostre dataset.

Veiem que algunes d'aquestes variables estan relacionades amb el llançament de tres punts. Això és degut a que fins 1979 no existia aquest llançament, així que hi ha gairebé trenta anys de registres on aquestes variables tenen valor nul per definició. Per altra banda, algunes mètriques com els rebots ofensius (variable *orb*) o defensius (variable *drb*) i les pèrdues (com el percentatge del propi equip amb la variable *tov_pct* o del contrari amb *opp_tov_pct*) també van ser estadístiques que no es recollien originalment, així que és normal que hi hagi valors perduts en aquestes. De fet, si filtrem i només agafem les observacions posteriors a la temporada 1978-79 (el llançament de tres va començar a la 1979-80) veurem que el nombre de valors perduts és molt diferent:

```
## Valors perduts després de la temporada 1978-79: 0
```

Table 2: Variables amb més valors perduts

Variable	Valors NA
fg3a_per_fga_pct	378
fg3	378
fg3a	378
fg3_pct	378
tov_pct	259
orb_pct	259
opp_tov_pct	259
drb_pct	259
orb	259
drb	259

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.