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

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

10/01/2023

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

Com bé hem comentat en l'anterior apartat, trobem temporades en les quals ens falta informació sobre estadístiques bàsiques sobre el joc. D'aquesta forma, ens dificulta l'anàlisi sobre aquestes temporades, ja que no disposem dels elements bàsics per entendre com era l'estil o funcionament de joc de cada equip. En conseqüència, s'ha decidit descartar aquestes temporades.

3.1: Les dades contenen zeros o elements buits?

Així i tot, trobem temporades més antigues a l'aparició del tir de tres punts que si disposen d'aquestes estadístiques. Observant el conjunt de dades, trobem que a partir de la temporada 1973-74 disposem de tota la informació necessària. Així i tot, analitzarem quins valors buits disposem en el nostre conjunt de dades a partir d'aquesta temporada.

```
## fg3a_per_fga_pct             fg3             fg3a          fg3_pct
##              102             102              102              102
##           Season          League            Team             wins
##                0               0                0                0
##           losses    win_loss_pct               gb       pts_per_g
##                0               0                0                0
##    opp_pts_per_g        Playoffs              age       wins_pyth
##                0               0                0                0
##      losses_pyth             mov              sos              srs
##                0               0                0                0
##          off_rtg         def_rtg          net_rtg             pace
##                0               0                0                0
##  fta_per_fga_pct          ts_pct          efg_pct          tov_pct
##                0               0                0                0
##          orb_pct         ft_rate      opp_efg_pct      opp_tov_pct
##                0               0                0                0
##          drb_pct     opp_ft_rate                g               mp
##                0               0                0                0
##               fg              fga           fg_pct              fg2
##                0               0                0                0
##             fg2a         fg2_pct               ft              fta
##                0               0                0                0
```

```
##          ft_pct             orb             drb             trb
##               0               0               0               0
##             ast             stl             blk             tov
##               0               0               0               0
##              pf
##               0
```

Com era d'esperar, les úniques variables que observem amb valors buits són les que fan referència als tirs de tres punts. El que farem, és omplir aquests valors buits amb el valor 0, ja que s'ha decidit que és el valor que realment representa aquests camps. Com no existia el tir de tres no es van realitzar cap tir.

Un altre fet a tenir en compte, és el camp `gb`. Aquesta variable ens mostra quants partits té cada equip per darrere del rival que ocupa el primer lloc. En alguns casos, aquesta informació ve representada amb el valor -. Entenem que aquest valor, fa referència al fet que no té cap partit per darrere dels rivals (és a dir, és l'equip que va primer de la seva divisió) i el seu valor real és 0.

3.2: Identifica i gestiona els valors extrems.

En aquest cas, a partir de la funció 'summary' podem observar els mínims i màxims valors de cada variable. No s'exemplificarà en el document, ja que el gran volum de variables dificultaria la lectura del document de la pràctica. Així i tot, un cop comprovat aquest, sí que podem trobar valors molt petits en comparació a la mitjana, com pot ser el cas de victòries que ha assolit un equip en una temporada.

Ho podem veure en aquest petit exemple:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   22.00   36.00   35.41   47.00   73.00
```

És normal, que ens podem trobar aquests casos en múltiples variables, ja que en una mateixa temporada, els equips guanyadors presentaran valors màxims en victòries i els perdedors mínims en aquestes. En conseqüència, també es veurà reflectit en les estadístiques d'aquests equips.

El que s'ha decidit és no realitzar cap modificació, ja que aquests valors són correctes perquè es basen en dades reals sobre les temporades i les estadístiques de joc de cada un dels equips. Per tant, hem d'assumir que serà possible trobar valors que siguin *outliers*, però que haurem de tractar com un valor més.

```
##     Season             League              Team               wins
##  Length:1900        Length:1900        Length:1900        Min.   : 2.00
##  Class :character   Class :character   Class :character   1st Qu.:24.00
##  Mode  :character   Mode  :character   Mode  :character   Median :37.00
##                                                           Mean   :35.85
##                                                           3rd Qu.:47.00
##                                                           Max.   :73.00
##
##      losses       win_loss_pct         gb             pts_per_g
##  Min.   : 3.00   Min.   :0.0880   Length:1900        Min.   : 56.90
##  1st Qu.:25.00   1st Qu.:0.3890   Class :character   1st Qu.: 92.50
##  Median :36.00   Median :0.5120   Mode  :character   Median :101.15
##  Mean   :35.85   Mean   :0.4999                      Mean   : 98.37
##  3rd Qu.:46.00   3rd Qu.:0.6100                      3rd Qu.:108.30
##  Max.   :73.00   Max.   :0.9000                      Max.   :126.50
##
##  opp_pts_per_g      Playoffs              age          wins_pyth
##  Min.   : 55.90   Length:1900        Min.   :22.40   Min.   : 1.00
```

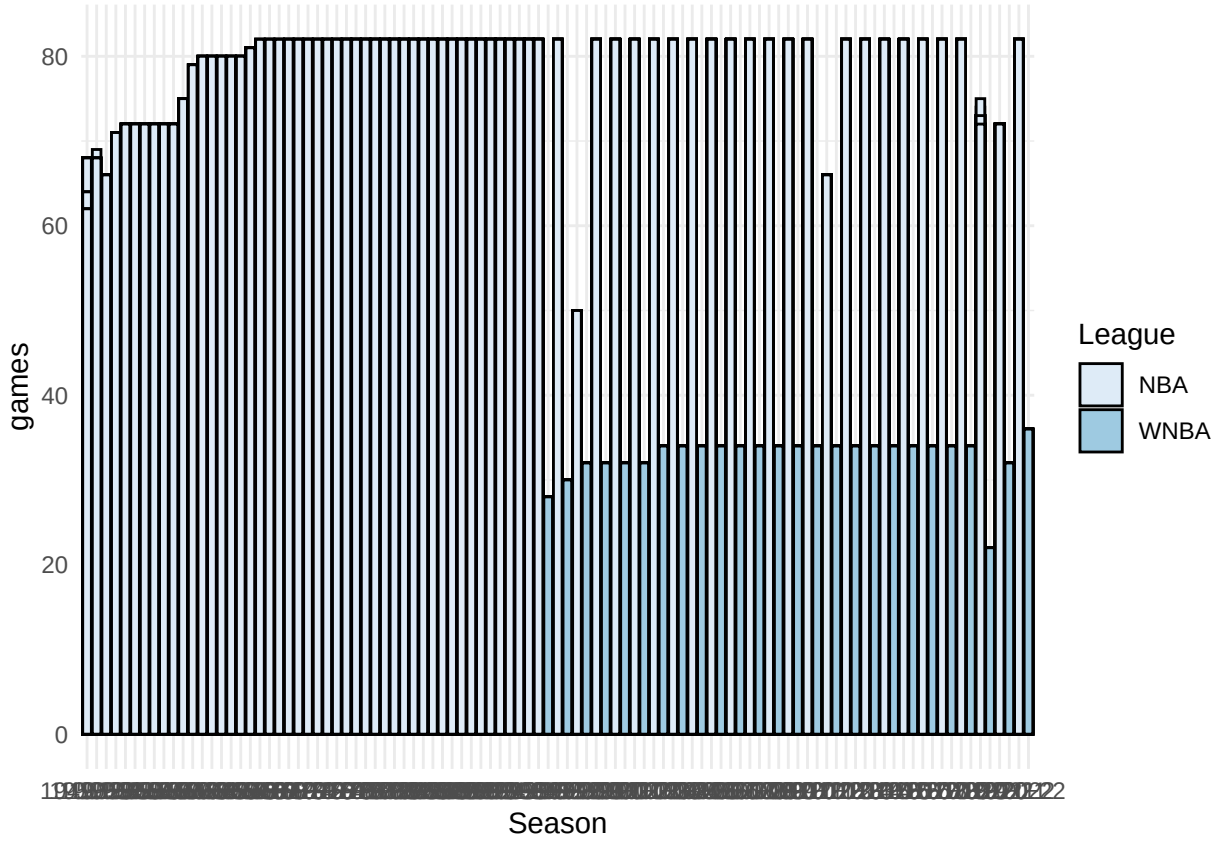
```
##  1st Qu.: 92.17   Class :character   1st Qu.:25.70   1st Qu.:24.00
##  Median :101.30   Mode  :character   Median :26.60   Median :37.00
##  Mean   : 98.38                      Mean   :26.79   Mean   :35.92
##  3rd Qu.:107.92                      3rd Qu.:27.70   3rd Qu.:47.00
##  Max.   :130.80                      Max.   :32.00   Max.   :70.00
##                                      NA's   :28
##   losses_pyth         mov                 sos                 srs
##  Min.   : 2.00   Min.   :-15.370000   Min.   :-1.9800000   Min.   :-14.680000
##  1st Qu.:25.00   1st Qu.: -3.040000   1st Qu.:-0.2900000   1st Qu.: -2.897500
##  Median :36.00   Median :  0.205000   Median :-0.0100000   Median :  0.165000
##  Mean   :35.79   Mean   : -0.002774   Mean   :-0.0003684   Mean   : -0.004642
##  3rd Qu.:46.00   3rd Qu.:  3.210000   3rd Qu.: 0.2900000   3rd Qu.:  2.927500
##  Max.   :72.00   Max.   : 12.780000   Max.   : 1.7900000   Max.   : 12.740000
##
##     off_rtg         def_rtg         net_rtg               pace
##  Min.   : 80.4   Min.   : 79.3   Min.   :-19.900000   Min.   : 62.30
##  1st Qu.: 99.1   1st Qu.: 99.4   1st Qu.: -3.200000   1st Qu.: 89.60
##  Median :104.5   Median :104.6   Median :  0.200000   Median : 94.70
##  Mean   :103.4   Mean   :103.4   Mean   :  0.001275   Mean   : 94.64
##  3rd Qu.:108.2   3rd Qu.:108.1   3rd Qu.:  3.400000   3rd Qu.:101.20
##  Max.   :118.3   Max.   :117.6   Max.   : 18.100000   Max.   :136.30
##  NA's   :17      NA's   :17      NA's   :17           NA's   :17
##  fta_per_fga_pct  fg3a_per_fga_pct     ts_pct          efg_pct
##  Min.   :0.1940   Min.   :0.0110   Min.   :0.3770   Min.   :0.3100
##  1st Qu.:0.2790   1st Qu.:0.1280   1st Qu.:0.5070   1st Qu.:0.4610
##  Median :0.3090   Median :0.2030   Median :0.5270   Median :0.4820
##  Mean   :0.3116   Mean   :0.1988   Mean   :0.5236   Mean   :0.4778
##  3rd Qu.:0.3400   3rd Qu.:0.2650   3rd Qu.:0.5440   3rd Qu.:0.5010
##  Max.   :0.5540   Max.   :0.5190   Max.   :0.6100   Max.   :0.5750
##                   NA's   :378
##     tov_pct         orb_pct         ft_rate        opp_efg_pct
##  Min.   : 9.9   Min.   :17.30   Min.   :0.1430   Min.   :0.4150
##  1st Qu.:13.3   1st Qu.:25.80   1st Qu.:0.2100   1st Qu.:0.4690
##  Median :14.3   Median :29.30   Median :0.2320   Median :0.4860
##  Mean   :14.7   Mean   :28.91   Mean   :0.2347   Mean   :0.4867
##  3rd Qu.:15.8   3rd Qu.:32.00   3rd Qu.:0.2570   3rd Qu.:0.5030
##  Max.   :23.0   Max.   :39.10   Max.   :0.4120   Max.   :0.5640
##  NA's   :259    NA's   :259                      NA's   :208
##   opp_tov_pct      drb_pct       opp_ft_rate           g              mp
##  Min.   :10.3   Min.   :61.50   Min.   :0.1390   Min.   :22.0   Min.   :194.1
##  1st Qu.:13.3   1st Qu.:68.00   1st Qu.:0.2070   1st Qu.:72.0   1st Qu.:240.6
##  Median :14.3   Median :70.80   Median :0.2270   Median :82.0   Median :241.2
##  Mean   :14.7   Mean   :71.07   Mean   :0.2298   Mean   :71.7   Mean   :234.2
##  3rd Qu.:15.8   3rd Qu.:73.90   3rd Qu.:0.2510   3rd Qu.:82.0   3rd Qu.:241.9
##  Max.   :24.2   Max.   :82.10   Max.   :0.3470   Max.   :82.0   Max.   :244.9
##  NA's   :259    NA's   :259     NA's   :208                     NA's   :140
##        fg              fga             fg_pct           fg3
##  Min.   :20.2   Min.   : 51.30   Min.   :0.310   Min.   : 0.100
##  1st Qu.:34.5   1st Qu.: 78.30   1st Qu.:0.435   1st Qu.: 3.300
##  Median :38.2   Median : 83.75   Median :0.453   Median : 5.200
##  Mean   :37.3   Mean   : 82.69   Mean   :0.450   Mean   : 5.458
##  3rd Qu.:42.0   3rd Qu.: 89.00   3rd Qu.:0.470   3rd Qu.: 7.175
##  Max.   :49.9   Max.   :119.60   Max.   :0.545   Max.   :16.700
##                                                  NA's   :378
```

```
##       fg3a          fg3_pct            fg2             fg2a
##  Min.   : 0.90   Min.   :0.1040   Min.   :15.50   Min.   : 36.90
##  1st Qu.: 9.90   1st Qu.:0.3220   1st Qu.:28.10   1st Qu.: 57.30
##  Median :15.20   Median :0.3450   Median :31.30   Median : 66.10
##  Mean   :15.63   Mean   :0.3348   Mean   :32.93   Mean   : 70.18
##  3rd Qu.:19.90   3rd Qu.:0.3620   3rd Qu.:40.40   3rd Qu.: 84.60
##  Max.   :45.40   Max.   :0.4280   Max.   :49.90   Max.   :119.60
##  NA's   :378     NA's   :378
##     fg2_pct             ft             fta            ft_pct
##  Min.   :0.3100   Min.   :10.10   Min.   :13.40   Min.   :0.6350
##  1st Qu.:0.4530   1st Qu.:16.80   1st Qu.:22.00   1st Qu.:0.7340
##  Median :0.4760   Median :19.00   Median :25.30   Median :0.7550
##  Mean   :0.4717   Mean   :19.40   Mean   :25.77   Mean   :0.7547
##  3rd Qu.:0.4950   3rd Qu.:21.73   3rd Qu.:28.80   3rd Qu.:0.7760
##  Max.   :0.5750   Max.   :31.90   Max.   :42.40   Max.   :0.8750
##
##       orb             drb             trb             ast
##  Min.   : 5.30   Min.   :16.80   Min.   :24.80   Min.   :12.20
##  1st Qu.:10.20   1st Qu.:27.60   1st Qu.:40.30   1st Qu.:20.10
##  Median :11.70   Median :29.80   Median :42.80   Median :22.20
##  Mean   :11.89   Mean   :29.24   Mean   :43.54   Mean   :22.05
##  3rd Qu.:13.50   3rd Qu.:31.70   3rd Qu.:45.40   3rd Qu.:24.50
##  Max.   :18.50   Max.   :42.20   Max.   :80.20   Max.   :31.40
##  NA's   :259     NA's   :259     NA's   :17
##       stl              blk             tov              pf
##  Min.   : 4.800   Min.   :2.000   Min.   :10.90   Min.   :14.20
##  1st Qu.: 7.200   1st Qu.:4.100   1st Qu.:14.10   1st Qu.:20.10
##  Median : 8.000   Median :4.700   Median :15.20   Median :22.10
##  Mean   : 8.061   Mean   :4.779   Mean   :15.66   Mean   :22.25
##  3rd Qu.: 8.800   3rd Qu.:5.400   3rd Qu.:16.80   3rd Qu.:24.50
##  Max.   :12.900   Max.   :8.700   Max.   :24.50   Max.   :32.10
##  NA's   :259      NA's   :259     NA's   :259
```

Exercici 4: Anàlisi de les dades

Abans de començar amb l'anàlisi, haurem de tenir en compte que els valors absoluts ens poden portar a error i que haurem d'optar, en general, pels relatius (percentatges o mètriques ajustades per partit o possessions). Això és degut a que la NBA i la WNBA tenen un nombre de partits totals diferents, però també perquè la pròpia NBA ha evolucionat en aquest aspecte: en el seu inici es jugaven molts menys partits. Veiem-ho amb un gràfic:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    22.0    72.0    82.0    71.7    82.0    82.0
```



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.