PRACTICA 2

Tipologia i cicle de vida de les dades.

Núria Pont Vilà
i
Ricard Clemente Martí

Es pot trobar tot el contingut de la pràctica a:
https://github.com/ricardclemente/UOC_M2951_Practica2

Índex

Pràctica 2

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset serà el que vam generar a la pràctica 1. Consisteix en un conjunt de dades provinents del rànquing d'atletisme espanyol dels últims cinc anys. Així doncs conté el conjunt d'atletes i resultats d'aquests en els últims anys. Creiem que les dades tenen utilitats molt diverses, però ens hem volgut centrar en utilitzar les dades per esbrinar la progressió dels atletes, així doncs, ens agradaria respondre a la pregunta següent; Podem saber quina serà la marca d'un atleta X en una prova Y la següent temporada? O més fàcil, millorarà el atleta X la seva marca la temporada següent?
Som conscients que en les dades del rànquing hi poden faltar molts atributs que poden ser decisius en aquesta pregunta (Lesions, condicions d'entrenament, disponibilitat de material...), però ens agradaria poder fer-nos una idea general.

2. Integració i selecció de les dades d'interès a analitzar.

En principi, per el què fa a la integració no afegirem més dades al dataset. Tot i això, creiem que segons les necessitats de l'estudi es podrien afegir més anys de dades (en el ránking hi tenim fins als últims 15 anys) o més atletes ja que en el datset només es mostren les 1000 millors marques i per a proves amb molt participants en podríem trobar bastantes més.

En canvi la selecció creiem que si que és important en el nostre cas ja que en les dades hi trobem bastants atributs que no són rellevants en l'estudi que es vol dur a terme. A continuació llistem els atributs i indicarem els que ens semblen rellevants:

- **Tipus de competició o pista**: Pista Coberta, Aire Lliure o Ruta. Aquesta és una dada que no té molta rellevància ja que la majoria de proves són diferents segons el tipus de pista, per exemple les curses de 5km, 10km, 15km… només es donen en la ruta i si es fan en pista les condicions són tant diferents que no es considera la mateixa prova. En canvi els 100m per posar un altre exemple només es fan al Aire lliure, ja que en pista coberta no es poden fer perquè la pista té unes mides diferents i en ruta no té sentit. Així doncs aquest atribut creiem que no és necessari i l'eliminarem, ja que amb la prova en tenim prou.
- **Sexe**. Indica el sexe del atleta. És rellevant, però passa semblant que amb el camp anterior, ja que en principi les proves masculina i femenina són diferents. Crec que es pot prescindir d'aquest camp.
- **Prova**: Si ens fixem en el camp de prova, es pot veure que ja conté els dos camps previs (per ex: «100m MASC. AL»). Aquest camp és indispensable.
- **Posició en el rànquing**: Aquest camp indica la posició del atleta en la temporada concreta. Aquesta pot ser una dada rellevant ja que pot mostrar una relació amb els atletes que tenen més recursos, ja que els primers del rànquing acostumen a tenir algun ajut (econòmic, accés a instal·lacions d'alt rendiment, ajut mèdic...)
- **Marca**: Indica la millor marca del atleta en la temporada. Aquest és un altre dels camps clau.
- **Vent**: Indica el registre del vent en la prova que es va fer la marca. No creiem que sigui un camp rellevant ja que depèn de la competició (no es pot preedir) i no es registra en totes les proves.
- **Atleta**: indica la persona que va fer la marca, és indispensable com a identificador.
- **Data de naixement**: Ens indica l'edat del atleta, és important per saber amb quants anys va fer la marca
- **Categoria**: No és molt rellevant ja que va relacionat amb l'edat de l'atleta

- **Número de llicència**: És un identificador del atleta, el problema que pot tenir és que hi ha alguns atletes que canvien de número de llicència (perquè canvien de comunitat autònoma o perquè un any no renoven la fitxa). Creiem que és millor el nom ja que és menys habitual que es canviï
- **Federació**: Indica la comunitat autònoma del atleta. No tenim clar que sigui significatiu en la marca del atleta. Tot i que podria ser un indicador de on hi ha pistes més bones i millors entrenadors, la realitat es que els atletes competeixen en diverses comunitats al llarg del any i entrenen en comunitats diferents a la de la seva federació. Així que no és molt rellevant.
- **Club**: Indica el club de l'atleta. En principi pel que fa a la marca no hauria d'influir, ja que en l'atletisme el club no dona indicis de l'entrenador o la ubicació del atleta (sobretot en edat senior)
- **Posició** (en la prova que es va fer la marca): Aquest atribut en podem prescindir totalment ja que la posició depèn de la competició i no està relacionat amb la marca. Amb la mateixa marca pots quedar primer a una cursa popular, i últim en una carrera professional.
- **Lloc on es va fer la marca**. És un altre atribut prescindible, ja que el lloc no és molt significatiu respecte a la marca, perquè un atleta pot anar a competir on vulgui.
- **Data de la marca**. Creiem que aquest camp si que és clau ja que ens indica la temporada i l'edat de l'atleta quan va fer la marca.

Així doncs ens quedarem amb els següents atributs:
Prova, Marca, Atleta i Anys (aquest últim l'obtindrem ajuntant les dades de data de naixement i la data de la marca). També ens interessarà saber la marca anterior de l'atleta i la millora respecte la marca de la temporada anterior.
La marca anterior l'obtindrem buscant els registres del mateix atleta i ordenant-los segons la data de la marca. La millora l'obtindrem de restar a la marca actual la marca anterior.

Apart de la selecció d'atributs pensem que podem reduir la quantitat, el què farem serà centrar-nos en una prova en concret ja que hi ha moltes proves que no tenen a veure l'una amb l'altre (valors dels resultats, millora segons l'edat, tipus de dades del resultat molt diversos), i per tant creiem que l'estudi s'ha de fer prova a prova.

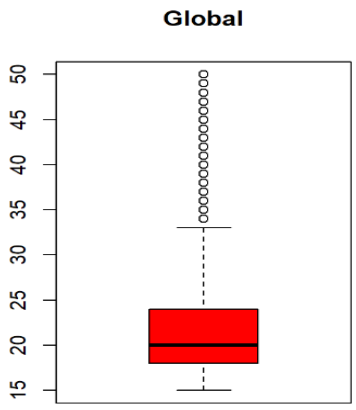
3. Neteja de les dades.

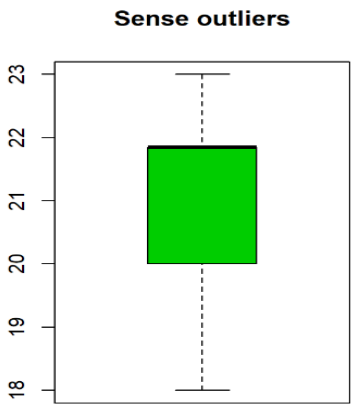
3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Les dades s'han hagut d'arreglar bastant, principalment perquè hem trobat incongruències o formats erronis, deguts a errors en la pujada de dades. Per exemple hi ha unes files que no tenen els atributs ben distribuïts dins el csv, perquè estaven en un format diferent.
També hem vist que en l'atribut de Marca hi ha dades que contenen una «A» al final, és degut a que en el rànquing es posa una A al final quan la marca s'ha fet en Altitud (més de 1500m), tot i que és una fet que es senyalitza la marca es conta com a vàlida, així que el que hem fet ha estat eliminar les A's.

3.2. Identificació i tractament de valors extrems.

Com diu la teoria els valors atípics poden ser indicatius de dades que pertanyen a una població diferent de la resta de les mostres establertes.
Si mirem els següents gràfics de boxplot podem observar el següent, en el primer gràfic hi ha tots els valors que tractem i que tinguin millores o no, es desestimem les dades que aquest resultat és NA, conseqüència de que només tenim un registre de l'atleta i per tant no podem tenir una referència anterior per contrastar la millora.
En el segon boxplot eliminem els valors que sobresurten, podem identificar clarament per aquesta prova en concret de '100m' on les dades sense outliers son les que van en el rang de 18 a 23.






```
result_final$stats
```

```
##      [,1]
## [1,]   15
## [2,]   18
## [3,]   20
## [4,]   24
## [5,]   33
```

```
result_final$conf
```

```
##          [,1]
## [1,] 19.78749
## [2,] 20.21251
```

```
result_final$n
```

```
## [1] 1990
```

```
result_final$out
```

```
##   [1] 36 40 34 49 40 36 40 38 39 47 46 45 38 37 41 50 49 39 38 41 50 48 38
##  [24] 44 46 45 41 40 35 34 41 40 39 50 49 48 36 38 37 36 35 44 34 50 49 48
##  [47] 35 34 37 44 43 42 41 49 48 47 46 34 36 35 46 45 44 37 36 37 36 35 34
##  [70] 39 36 48 47 46 45 44 38 37 35 49 37 36 44 43 42 42 36 43 42 41 35 43
##  [93] 39 38 37 42 36 39 44 43 42 39 38 40 39 38 38 37 36 40 39 38
```

```
result_imputed$stats
```

```
##          [,1]
## [1,] 18.00000
## [2,] 20.00000
## [3,] 21.84623
## [4,] 21.84623
## [5,] 23.00000
```

```
result_imputed$conf
```

```
##          [,1]
## [1,] 21.78084
## [2,] 21.91162
```

```
result_imputed$n
```

```
## [1] 1990
```

```
result_imputed$out
```

```
## numeric(0)
```

4. Anàlisi de les dades.

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

anàlisis a aplicar).

Hem decidit que l'estudi el farem sobre les dades de cent metres masculí. I per cada registre calcularem varis camps nous, ja explicats en el punt 2:
- Millora: millora respecte la marca anterior (si no té una marca anterior el registre s'elimina)
- Marca anterior: Marca anterior feta per el mateix atleta en una temporada prèvia.

L'objectiu principal serà buscar si hi ha relació entre la millora i l'edat de l'atleta o la marca anterior. En principi s'intueix que hauríem de veure que la gent més jove millora més, així com la gent amb marques més dolentes també millora més.

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

Per analitzar la normalitat utilitzem el test de shapiro-wilk i obtenim els següents resultats a R:

```
> shapiro.test(dades$millora)

      Shapiro-Wilk normality test

data:  dades$millora
W = 0.99635, p-value = 0.0001007


> shapiro.test(dades$anys)

      Shapiro-Wilk normality test

data:  dades$anys
W = 0.80618, p-value < 2.2e-16


> shapiro.test(dades$marcaAnt)

      Shapiro-Wilk normality test

data:  dades$marcaAnt
W = 0.98117, p-value = 1.541e-15
```

es pot veure que tots els p-values són menors que 0,05 i arribem a la conclusió que les dades no segueixen una distribució normal.

Per veure la homogeneïtat de la variància utilitzem el test de Fligner-Killeen ja que no compleixen amb la hipòtesi de normalitat.

```
> fligner.test(millora ~ anys, data = dades)

      Fligner-Killeen test of homogeneity of variances

data:  millora by anys
Fligner-Killeen:med chi-squared = 29.85, df = 35, p-value = 0.7149


> fligner.test(millora ~ marcaAnt, data = dades)

      Fligner-Killeen test of homogeneity of variances

data:  millora by marcaAnt
Fligner-Killeen:med chi-squared = 187.31, df = 174, p-value = 0.2322
```

p-value és major que 0,05 en ambdós casos així que podem dir que tant 'Anys' com 'marcaAnt' tenen variàncies similars a 'millora'

4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Utilitzem el test de kruskal:

```
> kruskal.test(millora ~ anys, data = dades)
```

```
	Kruskal-Wallis rank sum test

data:  millora by anys
Kruskal-Wallis chi-squared = 271.72, df = 35, p-value < 2.2e-16
```

```
> kruskal.test(millora ~ marcaAnt, data = dades)

	Kruskal-Wallis rank sum test

data:  millora by marcaAnt
Kruskal-Wallis chi-squared = 487.32, df = 174, p-value < 2.2e-16
```

Atès que el p-valor obtingut és menor que el nivell de significació, es pot concloure que la millora mostra diferències segons l'edat i segons la marca anterior.

A continuació mirarem si es compleix la regressió:

```
> ml=lm(formula = millora ~ anys+marcaAnt, data = dades)

> summary(ml)

Call:
lm(formula = millora ~ anys + marcaAnt, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74744 -0.14662  0.00098  0.14275  1.16337

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.4755357  0.1547870  15.993   <2e-16 ***
anys         0.0078850  0.0008161   9.662   <2e-16 ***
marcaAnt    -0.2371374  0.0132290 -17.926   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.221 on 1987 degrees of freedom
Multiple R-squared:  0.1954,	Adjusted R-squared:  0.1946
F-statistic: 241.2 on 2 and 1987 DF,  p-value: < 2.2e-16
```

Efectivament veiem que no es pot crear una regressió prou bona per a preedir la millora segons la marca anterior i els anys ja que R^2 és 0,1954, o sigui només s'ajusta a un 20% de les dades.
Per separat obtenim els següents gràfics de regressió:

Calculem la correlació de Spearman ja que no compleix normalitat:

```
> cor.test(dades$millora,dades$anys, method="spearman")

	Spearman's rank correlation rho

data:  dades$millora and dades$anys
S = 872100000, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3360172
```

```
> cor.test(dades$millora,dades$marcaAnt, method="spearman")
```

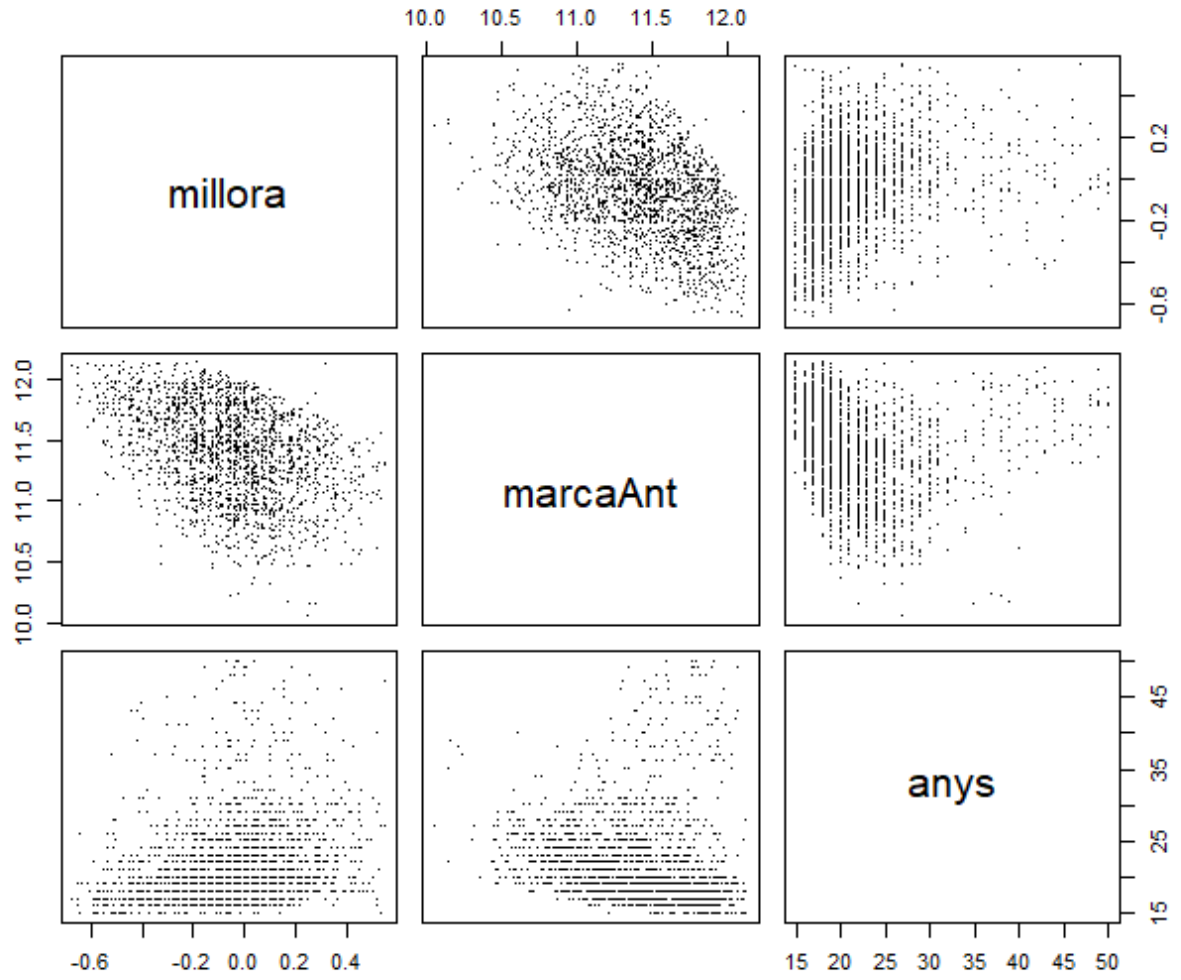
```
      Spearman's rank correlation rho

data:  dades$millora and dades$marcaAnt
S = 1839200000, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.4003257
```

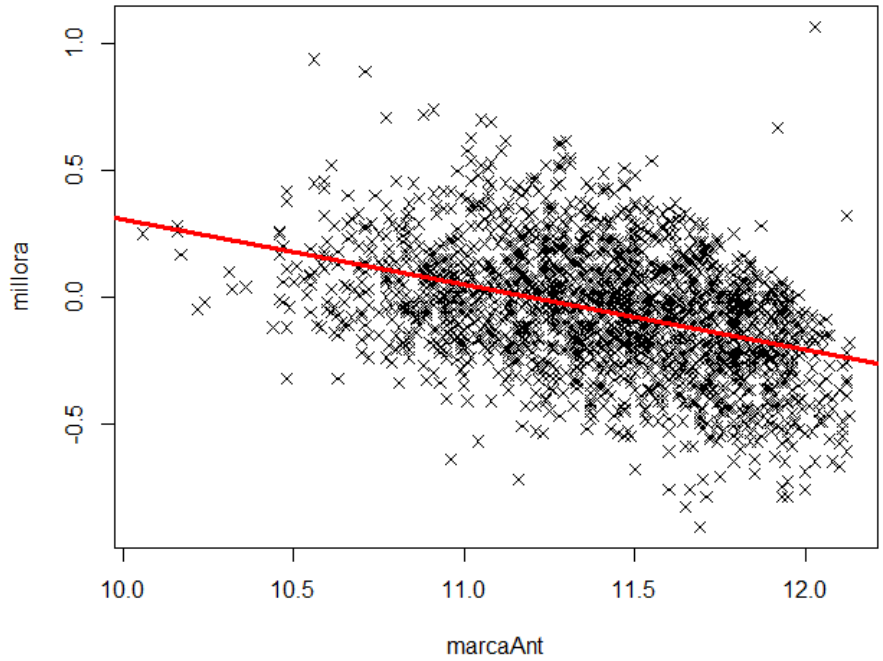
veiem que en cap dels dos casos el coeficient de correlació és major a 0,57. Per tant podem dir que no tenen prou correlació.

5. Representació dels resultats a partir de taules i gràfiques.

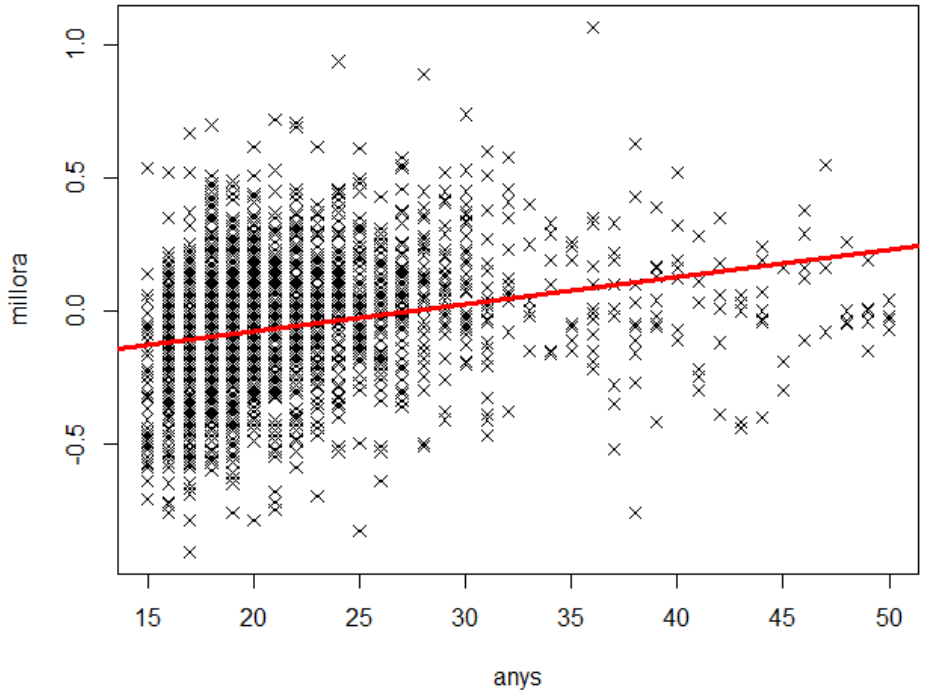
Aquí veiem un plot general de les dades, on es pot observar que no hi ha una relació molt evident entre cap dels camps i la millora.



La imatge següent representa la línia de regressió de la millora segons la marca anterior, i es pot veure que segueix la predicció feta: «com pitjor és la marca anterior més gran és la millora», tot i que no s'ajusta massa bé a les dades com ja hem vist a l'apartat anterior. S'ha de dir que la millora és millor quan més petita ja que vol dir que ha baixat més segons la marca.



En la següent es mostra la millora segons l'edat i la línia de regressió corresponent. També veiem que segueix la predicció feta «a menor edat més gran és la millora» i igual que en el cas anterior no s'ajusta massa bé.



6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Recordem quin era el nostre objectiu:

> *Podem saber quina serà la marca d'un atleta X en una prova Y la següent temporada? O més fàcil, millorarà el atleta X la seva marca la temporada següent?*

Podem concloure que si, podem obtenir crear un model on tracti l'atleta, analitzi les seves dades i pugui predir si millorarà no, i quina seria aquesta. (evidentment com en dit al inici en circumstàncies normals(sense lesions, condicions d'entrenament, disponibilitat de material...)

Com hem vist en l'apartat anterior les principals conclusions son:

- a menor edat més gran és la millora
- com pitjor és la marca anterior més gran és la millora

Per tal de veure el detall dels resultats finals els podem exportar en un arxiu .csv

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

Adjuntem els següents arxius:

- Practica2_Code_R.html
- (execució de la practica i resultats)
- Practica2_Code_R.rmd
  (codi R de la practica)
- Practica2_Code_R.r
  (codi R de la practica)
- resultats_arreglo
  (Arxiu inicial.csv)

Que es poden trobar a:

https://github.com/ricardclemente/UOC_M2951_Practica2