Tipologia i cicle de vida de les dades

PAC 1 - *Web scraping*

Màster en ciència de dades

Oscar Galera i Alfaro

11 de novembre de 2019

Qüestions

Membres del grup: Òscar Galera i Alfaro

1. **Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.**

   La informació s'ha extret del web *Basketball Reference*[1], que compta amb les dades estadístiques de tots els jugadors que han competit en la lliga nord-americanta de *basketball NBA*.

   S'ha decidit treballar amb aquest web per la seva simple estructuració, la freqüent actualització de les dades (aquestes s'actualitzen cada dia que hi ha partit) i la seva completesa.

2. **Definir un títol pel dataset. Triar un títol que sigui descriptiu.**

   Un possible títol per les dades extretes seria:

   Estadístiques dels jugadors de la *NBA* al llarg de la història

3. **Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol**

   Les dades extretes s'han agrupat en tres *datasets*:

   - **1.** `data/info_players.csv` on s'emmagatzema la informació de tots els jugadors, tal com: nom, alçada, pes, universitat d'origen...

     Està format per **4748 registres** i **8 variables**

   - **2.** `data/data_players.csv` on s'emmagatzemen les dades referents a les estadístiques de tots els jugadors relatives a cada temporada.

     Està format per **28.753 registres** i **12 variables**

   - **3.** `data/totals.csv` on s'emmagatzemen les dades globals dels jugadors.

     Està format per **4748 registres** i **9 variables**

4. **Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment**

[1] https://www.basketball-reference.com/

En el següent diagrama relacional s'il·lustren els *datasets* comentats en l'apartat anterior, així com les seves relacions.

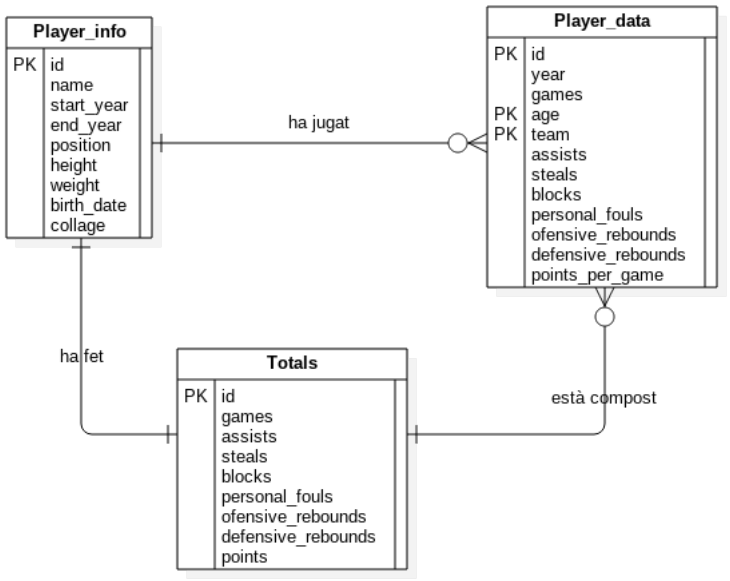

Figura 1: Diagrama relacional

5. **Contingut. Explicar els camps que inclou el dataset**

- *Dataset* `data/info_players.csv`:
  - *id*: Identificador del jugador.
  - *start_year*: Any d'inici a la lliga.
  - *end_year*: Any de final a la lliga.
  - *position*: Posició.
  - *height*: Alçada.
  - *weight*: Pes.
  - *birth_date*: Data de naixement.
  - *collage*: Universitat d'origen.
- *Dataset* `data/data_players.csv`:
  - *id*: Identificador del jugador.
  - *year*: Temporada.
  - *games*: Nombre de partits.
  - *age*: Edat del jugador.
  - *team*: Inicials de l'equip.
  - *assist*: Percentatge d'assistències.
  - *steals*: Percentatge de robaments de pilota.
  - *blocks*: Percentatge de taps.

  - *personal_fouls*: Percentatge de faltes personals.
  - *ofensive_rebounds*: Percentatge de rebots ofensius.
  - *defensive_rebounds*: Percentatge de rebots defensius.
  - *points_per_game*: Percentatge de punts.
- *Dataset* `data/totals.csv`:
  - *id*: Identificador del jugador.
  - *games*: Nombre de partits.
  - *assist*: Nombre total d'assistències.
  - *steals*: Nombre total de robaments de pilota.
  - *blocks*: Nombre total de taps.
  - *personal_fouls*: Nombre total de faltes personals.
  - *ofensive_rebounds*: Nombre total de rebots ofensius.
  - *defensive_rebounds*: Nombre total de rebots defensius.
  - *points*: Nombre total de punts.

El període de temps de les dades

Aquestes dades tenen una alta volatilitat perquè s'actualitzen després de cada partit que es juga durant la fase regular o *playoffs*. El calendari de partits sol tenir una durada de 9 mesos i va des de principis d'Octubre a finals de Juny. [2]

I com s'ha recollit.

La recopilació de les dades s'ha fet a través d'un programa escrit amb *Python* (veure apartat 9) sobre el web `https://www.basketball-reference.com/`, utilitza les llibreries:

- *urllib3*: S'usa per descarregar el codi *HTML*.
- *Beautifulsoup*: S'usa per aplicar tècniques de *web scraping* i així extreure les dades d'interès.

6. **Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).**

   El propietari de les dades és l'organització *Sports Reference LLC* i té com a principal objectiu proporcionar dades estadístiques que ajudin als aficionats a respondre possibles preguntes que es plantegin. El web oficial del patrocinador de les dades és:

   `https://www.sports-reference.com/`

   No s'han trobat cites ni anàlisis anteriors que siguin públics i que treballin amb dades similars a les aquí presents.

[2] La temporada 2019/2020 va del 22 d'Octubre al 4 de Juny, per a més informació consultar: `https://www.nbamaniacs.com/calendario-nba/`

7. **Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre**

   Ús de les dades

   Aquestes dades podrien ser utilitzades en àmbits molt diversos, com ara, en les comparacions estadístiques entre jugadors, equips o fins i tot conferències, per així obtenir un criteri objectiu en base al rendiment dels jugadors.

   També podria utilitzar-se com a font de dades per notícies referents a nous records obtinguts.

   En cas d'obtenir permís d'ús comercial, també es podrien usar com a base estadística en les cases d'apostes.

   Consultes:

   En el projecte `usa-map`[3] s'han plantejat una sèrie de preguntes i s'han resolt usant les dades extretes en aquesta pràctica, les preguntes són:

   - **1. Quants jugadors han jugat a cada equip en una temporada i com es distribueixen pels diferents estats?**

     S'ha de fer una agrupació dels equips en que ha estat cada jugar (usant el dataset `data/data_players.csv`) i fer un recompte.

     Solució: temporada 2018-19
     - *GA:* 16
     - *MA:* 22
     - *NY:* 39
     - *IL:* 14
     - *NC:* 17
     - *OH:* 20
     - *TX:* 56
     - *CO:* 20
     - *MI:* 20
     - *CA:* 72
     - *IN:* 16
     - *TN:* 19
     - *FL:* 36
     - *WI:* 20

[3]Aquest subprojecte està disponible en la carpeta `usa-map` del repositori *github*

- *MN:* 25
- *LA:* 21
- *PA:* 25
- *AZ:* 23
- *OR:* 18
- *UT:* 22
- *DC:* 18

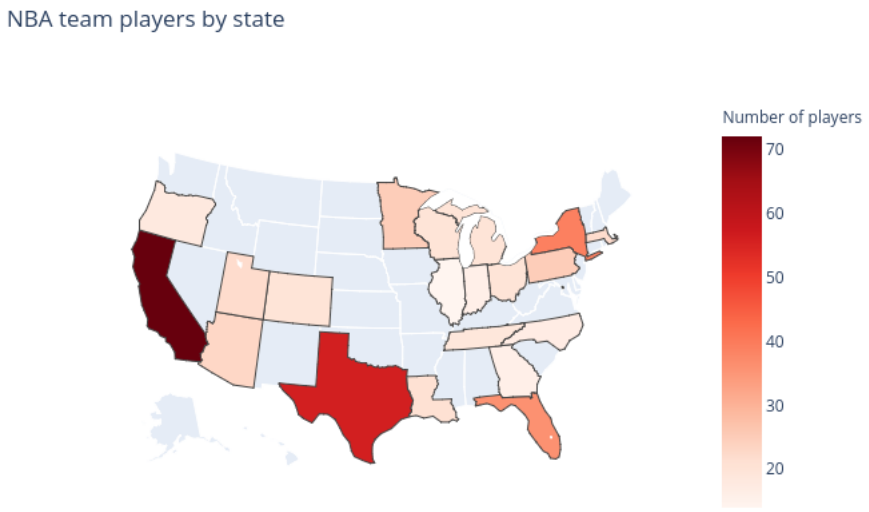

Figura 2: Resposta consulta 1

- **2. Quin és el jugador amb major nombre de punts per partit en tota l'història?**

  Cal buscar el jugador amb valor màxim en el camp *points* (utilitzant el *dataset* `data/totals.csv`) i llavors buscar la informació del jugador (amb el *dataset* `data/info_players.csv` usant del camp *id*.)

  Solució:

  - ***Name:*** *Kareem Abdul-Jabbar*
  - ***Start year:*** *1970*
  - ***End year:*** *1989*
  - ***Points:*** *38391*

- **3. Quins són els jugadors més joves que van debutar a l'*NBA*?**

  Cal buscar els jugadors amb valor mínim en el camp *age* (*dataset* `data/data_`

`players.csv`) i llavors buscar i mostrar la informació dels jugadors (amb el dataset `data/info_players.csv` i el camp *id*).

Solució:

- ***Id:*** *331*
- ***Name:*** *Andris Biedri*
- ***Age:*** *18*
- ***Start year:*** *2005*
- ***End year:*** *2014*
- ***Position:*** *C*

- ***Id:*** *547*
- ***Name:*** *Stan Brown*
- ***Age:*** *18*
- ***Start year:*** *1948*
- ***End year:*** *1952*
- ***Position:*** *F*

- ***Id:*** *561*
- ***Name:*** *Kobe Bryant*
- ***Age:*** *18*
- ***Start year:*** *1997*
- ***End year:*** *2016*
- ***Position:*** *G-F*

...

8. **Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:**

   S'ha decidit que la llicència més adequada per aquestes dades és la ***CC BY-NC-SA 4.0*** degut a les característiques i la finalitat d'aquestes. La llicència permet:

   - **(BY) Reconeixement:** S'ha de reconèixre adequadament l'autoria de les dades, d'aquesta manera queda reflectit qui és el creador, però al mateix temps es permeten millores especificant l'autor d'aquestes.
   - **(NC) No comercial:** D'aquesta manera s'evita que altres parts utilitzin les dades amb finalitats econòmiques, permetent aquest ús exclusiu al creador.
   - **(SA) Compartir igual:** Amb aquesta restricció és facilita la distribució de les dades sempre d'acord amb la ideologia del creador.

9. **Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en *Python* o, alternativament, en *R*.**

   Tant el codi d'aquest projecte com les dades obtingudes després de la seva execució, es poden trobar en el repositori *github* que hi ha en l'adreça:

   `https://github.com/ogalera-dev/web-scraping`

10. **Dataset. Presentar el dataset en format CSV**

    En el següents fragments de dades hi ha una extracció de cada un dels *datasets* obtingut a partir del *web scraping* descrit en aquesta pràctica.

    Mostra del *dataset* `data/info_players.csv`

```
id,year,games,age,team,assists,steals,blocks,personal_fouls,ofensive_rebounds,defensive_rebounds,points_per_game
1,1990-91,43,22,POR,0.3,0.1,0.3,0.9,0.6,1.4,3.1
1,1991-92,71,23,POR,0.4,0.4,0.2,1.9,1.1,2.5,6.1
1,1992-93,75,24,TOT,0.4,0.3,0.3,2.5,1.7,2.8,7.7
1,1992-93,12,24,MIL,0.8,0.5,0.3,2.0,1.0,2.1,5.3
1,1992-93,63,24,BOS,0.3,0.3,0.3,2.6,1.8,3.0,8.2
1,1993-94,13,25,BOS,0.2,0.2,0.2,1.5,0.9,2.6,4.9
1,1994-95,54,26,TOT,0.2,0.3,0.2,1.9,0.7,1.4,4.7
1,1994-95,51,26,SAC,0.3,0.3,0.2,2.0,0.7,1.4,5.0
1,1994-95,3,26,PHI,0.0,0.0,0.0,0.7,1.0,1.7,0.7
2,1968-69,49,22,TOT,0.8,,,2.3,,,7.3
2,1968-69,20,22,CIN,0.2,,,0.9,,,1.9
```

    Mostra del *dataset* `data/data_players.csv`

```
id,name,start_year,end_year,position,height,weight,birth_date,collage
1,Alaa Abdelnaby,1991,1995,F-C,6-10,240,"June 24, 1968",Duke
2,Zaid Abdul-Aziz,1969,1978,C-F,6-9,235,"April 7, 1946",Iowa State
3,Kareem Abdul-Jabbar,1970,1989,C,7-2,225,"April 16, 1947",UCLA
4,Mahmoud Abdul-Rauf,1991,2001,G,6-1,162,"March 9, 1969",LSU
5,Tariq Abdul-Wahad,1998,2003,F,6-6,223,"November 3, 1974",Michigan
6,Shareef Abdur-Rahim,1997,2008,F,6-9,225,"December 11, 1976",Cal
7,Tom Abernethy,1977,1981,F,6-7,220,"May 6, 1954",Indiana
8,Forest Able,1957,1957,G,6-3,180,"July 27, 1932",Western Kentucky
9,John Abramovic,1947,1948,F,6-3,195,"February 9, 1919",Salem International University
```

    Mostra del *dataset* `data/totals.csv`

```
id,games,assists,steals,blocks,personal_fouls,ofensive_rebounds,defensive_rebounds,points
1,385.0,128.5,114.19,95.69,775.1,445.0,847.8,2298.7
2,570.0,649.2,133.89,204.79,1272.2,526.0,1151.2,4976.2
3,1560.0,5659.7,1161.39,3199.70,4644.7,2974.79,9382.4,38390.6
4,586.0,2077.0,474.29,43.3,1103.7,213.4,863.0,8556.9
5,321.0,391.0,265.9,118.0,692.1,432.0,722.3,2661.4
```

```
6,915.0,2288.6,900.9,689.8,2553.1,2074.9,4824.8,16409.8
7,358.0,402.2,197.5,63.1,565.8,392.3,657.49,1843.6
8,1.0,1.0,0.0,0.0,1.0,0.0,0.0,0.0
9,65.0,36.9,0.0,0.0,179.89,0.0,0.0,538.69
10,174.0,89.39,87.0,20.5,295.8,49.1,201.4,924.8
```

Contribucions

Contribucions	**Signa**
Recerca prèvia	OGA
Redacció de les respostes	OGA
Desenvolupament del codi	OGA