Tipologia i cicle de vida de les dades

Pràctica 1 - Alejandro Santamarta Martinez

L'objectiu d'aquest script d'scraping és aconseguir recopilar estadístiques de rendiment individual de jugadors de la NBA la web Basketball Reference.

1. Context.

Les dades que es pretén aconseguir es refereixen a les estadístiques individuals dels jugadors de la NBA a la temporada 2019-2020. L'objectiu amb aquestes dades és poder nutrir més endavant un model que pugui predir aproximadament el rendiment d'un jugador cara al final de la temporada.

Les dades es poden trobar a la web oficial de la NBA (un partit d'exemple: https://es.global.nba.com/boxscore/#!/0021900751), però l'abast de les estadístiques és limitat, ja que hi ha moltes mètriques que no apareixen. Cercant un poc per la xarxa se poden trobar alternatives de datasets ja montats, però tots estan referits a estadístiques agregades anualment, però no a la evolució partit a partit dels jugadors (exemples: https://www.kaggle.com/drgilermo/nba-players-stats#Seasons_Stats.csv, https://data.world/jgrosz99/nba-player-data-1978-2016), a més, la major part d'aquests datasets no estan actualitzats.

Finalment, per obtenir les dades es pot utilitzar el portal estadístic Basketall Reference (https://www.basketball-reference.com/). A aquest fil de reddit s'exposa que històricament havia una eina d'exportació d'una taula concreta en csv (https://www.reddit.com/r/sportsbook/comments/59hsno/best_website_to_pull_nba_statistics_into_excel/) però sembla no haver un lloc a on estigui fàcilment agregada la informació partit a partit de tota una temporada.

2. Definir un títol pel dataset.

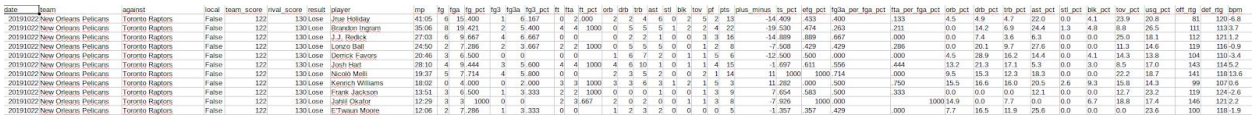
NBA player stats per game for season 2019-2020

3. Descripció del dataset.

Com s'ha explicat al primer punt el dataset consisteix en les estadístiques (bàsiques i avançades) individuals dels jugadors de la NBA durant la temporada 2019-2020.

4. Representació gràfica.

Imatge de la capçalera del dataset sencera



Detall de les columnes

date	team	against	local	team_score	rival_score	result	player	mp
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	Jrue Holiday	41:05
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	Brandon Ingram	35:06
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	J.J. Redick	27:03
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	Lonzo Ball	24:50
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	Derrick Favors	20:46
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	Josh Hart	28:10
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	Nicolò Melli	19:37
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	Kenrich Williams	18:02
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	Frank Jackson	13:51
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	Jahlil Okafor	12:29
20191022	New Orleans Pelicans	Toronto Raptors	False	122	130	Lose	E'Twaun Moore	12:06

fg	fga	fg_pct	fg3	fg3a	fg3_pct	ft	fta	ft_pct	orb	drb	trb	ast	stl	blk	tov	pf	pts	plus_minus	ts_pct	efg_pct	fg3a_per_fga_pct	fta_per_fga_pct
6	15	.400	1	6	.167	0	2	.000	2	2	4	6	0	2	5	2	13	-14	.409	.433	.400	.133
8	19	.421	2	5	.400	4	4	1000	0	5	5	5	1	2	2	4	22	-19	.530	.474	.263	.211
6	9	.667	4	6	.667	0	0		0	2	2	1	0	0	3	3	16	-14	.889	.889	.667	.000
2	7	.286	2	3	.667	2	2	1000	0	5	5	5	0	0	1	2	8	-7	.508	.429	.429	.286
3	6	.500	0	0		0	0		1	6	7	2	0	1	1	5	6	-12	.500	.500	.000	.000
4	9	.444	3	5	.600	4	4	1000	4	6	10	1	0	1	1	4	15	-1	.697	.611	.556	.444
5	7	.714	4	5	.800	0	0		2	3	5	2	0	0	2	1	14	11	1000	1000	.714	.000
0	4	.000	0	2	.000	3	3	1000	3	3	6	3	1	2	1	5	3	11	.282	.000	.500	.750
3	6	.500	1	3	.333	2	2	1000	0	0	0	1	0	0	1	3	9	7	.654	.583	.500	.333
3	3	1000	0	0		2	3	.667	2	0	2	0	0	1	1	3	8	-7	.926	1000	.000	1000
2	7	.286	1	3	.333	0	0		1	2	3	2	0	0	0	0	5	-1	.357	.357	.429	.000

orb_pct	drb_pct	trb_pct	ast_pct	stl_pct	blk_pct	tov_pct	usg_pct	off_rtg	def_rtg	bpm
4.5	4.9	4.7	22.0	0.0	4.1	23.9	20.8	81	120	-6.8
0.0	14.2	6.9	24.4	1.3	4.8	8.8	26.5	111	113	3.7
0.0	7.4	3.6	6.3	0.0	0.0	25.0	18.1	112	121	1.2
0.0	20.1	9.7	27.6	0.0	0.0	11.3	14.6	119	116	-0.9
4.5	28.9	16.2	14.4	0.0	4.1	14.3	13.8	104	110	-3.4
13.2	21.3	17.1	5.3	0.0	3.0	8.5	17.0	143	114	5.2
9.5	15.3	12.3	18.3	0.0	0.0	22.2	18.7	141	118	13.6
15.5	16.6	16.0	20.5	2.6	9.3	15.8	14.3	99	107	0.6
0.0	0.0	0.0	12.1	0.0	0.0	12.7	23.2	119	124	-2.6
14.9	0.0	7.7	0.0	0.0	6.7	18.8	17.4	146	121	2.2
7.7	16.5	11.9	25.6	0.0	0.0	0.0	23.6	100	118	-1.9

És important notar que el que pareixen anomalies de valors de 1000 als percentatges és per la interpretació que fa LibreOffice dels valors de 1.000 com a mil unitats. Al csv no existeix aquesta anomalia.

5. Contingut.

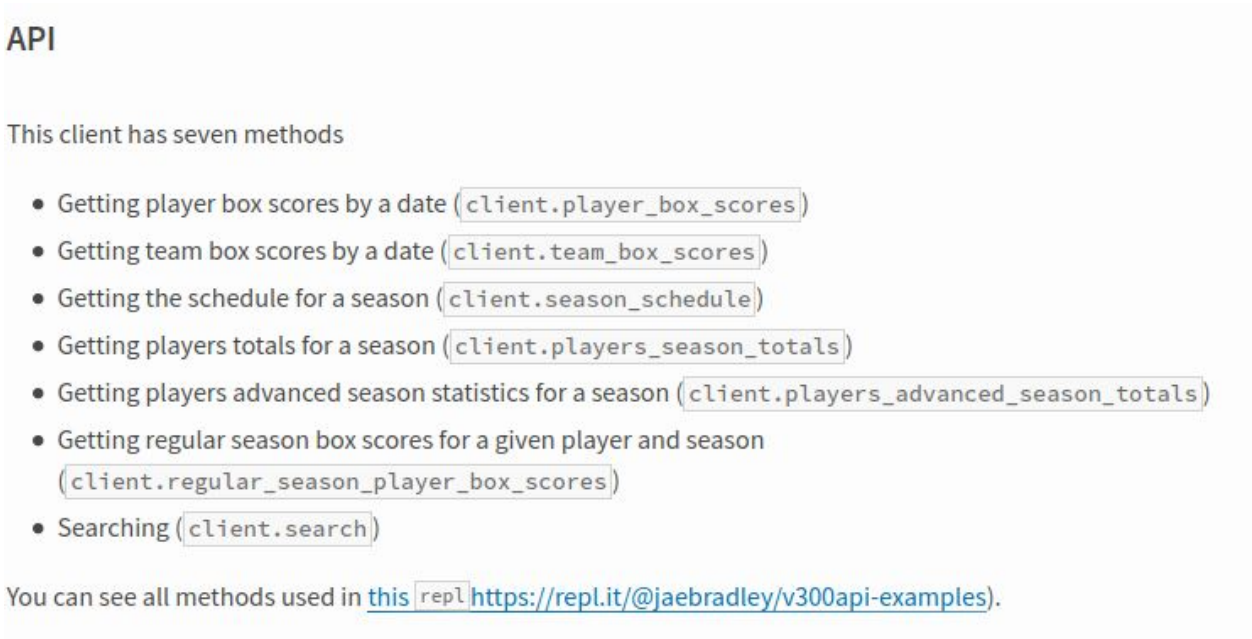
El dataset consisteix en els següents camps:

- date: data de la realització del partit a on s'han recopilat aquestes estadístiques en format *YYYYMMDD*
- team: equip per a qui juga aquest jugador en aquest partit
- against: equip rival a aquest partit
- local: si l'equip del jugador juga com a local (True) o visitant (False)
- team_score: puntuació final de l'equip del jugador
- rival_score: puntuació final de l'equip rival
- result: resultat final, victòria per l'equip del jugador (True) o derrota (False)
- player: Nom del jugador
- mp: minuts disputats del jugador al partit en format *minuts:segons*
- fg (Field Goals): cistelles anotades al partit
- fga (Field Goals Attempts): cistelles intentades al partit
- fg_pct (Field Goals Percentage): la divisió de fga/fg
- fg3 (Field Goals 3pt): triples encertats pel jugador al partit
- fg3a (Field Goals 3pt Attempted): triples intentats pel jugador
- fg3_pct (Field Goals 3pt Percentage): la divisió de fg3a/fg3
- ft (Free Throws): tirs lliures encertats pel jugador al partit
- fta (Free Throws Attempted): tirs lliures intentats pel jugador al partit
- ft_pct (Free Throws Percentage): la divisió de fta/f
- orb (Offensive Rebounds): rebots ofensius (davall la cistella rival) aconseguits pel jugador al partit
- drb (Defensive Rebounds): rebots defensius (davall la pròpia cistella) aconseguits pel jugador al partit

- trb (Total Rebounds): la suma de orb + drb
- ast (Assistències): passades del jugador que han acabat amb cistella per part d'un company immediatament després aconseguides pel jugador al partit
- stl (Steals): pilotes robades pel jugador a un jugador de l'equip rival durant el partit.
- blk (Blocks): "taps" (https://ca.wikipedia.org/wiki/Tap_(b%C3%A0squet)) aconseguits del jugador al partit
- tov (Turnovers): pèrdues de possessió, pilotes que el jugador ha "perdut" i ha significat que la possessió ha passat al rival durant el partit
- pf (Personal Fouls): faltes personals del jugador durant el partit
- pts (Points): punts aconseguits pel jugador durant el partit
- plus_minus (plus-minus): diferència de punts aconseguits per l'equip del jugador vs l'equip rival el temps que el jugador ha estat jugant (https://en.wikipedia.org/wiki/Plus%E2%80%93minus)
- ts_pct (True Shooting Percentage): % d'efectivitat combinada de tirs de 2, triples i tirs lliures.
- efg_pct (Effective Field Goald Percentage): % reajustat tenint en compte els punts que val un tir de 2 i un tir de 3 (https://en.wikipedia.org/wiki/Effective_field_goal_percentage)
- fg3a_per_fga_pct (3PAr o Three Point Attempt Rate): percentatge de triples intentats sobre el total de tirs del jugador al partit.
- fta_per_fga_pct (FTr o Free Throw Attempt Rate): tirs lliures intentats pel jugador respecte de tirs de camp.
- orb_pct (Offensive Rebound Percentage): percentatge (estimat) de rebots ofensius que un jugador ha aconseguit del total de rebots ofensius "disponibles" pel jugador.
- drb_pct (Defensive Rebound Percentage): percentatge (estimat) de rebots defensius que un jugador ha aconseguit del total de rebots ofensius "disponibles" pel jugador.
- trb_pct (Total Rebound Percentage): percentatge (estimat) total de rebots que un jugador ha aconseguit del total de rebots "disponibles" pel jugador.
- ast_pct (Assist Percentage): percentatge de tirs aconseguits pels compnyas assistits pel jugador mentre ha estat a pista
- stl_pct (Steal Percentage): percentatge de pilotes robades pel jugador sobre el total de possessions de pilota de l'equip rival mentre el jugador ha estat a pista.
- blk_pct (Block Percentage): percentatge de jugades taponades pel judador del total de tirs de camp intentats per l'equip rival mentre el jugador ha estat a pista.
- tov_pct (Turnover Percentage): percentatge (estimat) de pilotes perdudes per un jugador per cada 100 jugades on ha participat.
- usg_pct (Usage Percentage): percentatge de jugades de l'equip on el jugador ha participat.
- off_rtg (Offensive Rating): punts produïts (punts del jugador + punts aconseguits per companys després d'assistències del jugador) per cada 100 possessions de l'equip
- def_rtg (Deffensive Rating): punts encaixats per l'equip per cada 100 possessions.
- bpm (Box Plus-Minus): comparació del rati plus-minus amb la mitja de la temporada de la resta de jugadors (https://www.basketball-reference.com/about/bpm2.html)

6. Agraïments.

Les dades ha estat extretes de Basketball Reference, propietat de Sports Reference (https://www.sports-reference.com/about.html).
Com ja s'ha comentat al primer punt hi ha altres fonts de dades, més orientades al rendiment de un jugador al llarg d'una temporada sencera que al partit a partit.
Sobre Basketball Reference en sí ja hi ha un parell d'alternatives, inclús figura com exemple a posts d'scraping (https://towardsdatascience.com/web-scraping-nba-stats-4b4f8c525994), però novament orientat a l'extracció de les estadístiques anuals del jugadors.
Finalment, hi ha un paquet madur a pypi que fa scraping de la web de Basketball Reference (https://pypi.org/project/basketball-reference-web-scraper/), però està centrat novament en extraccions no orientades al proposat a l'scraping realitzat per a la pràctica.



API

This client has seven methods

- Getting player box scores by a date (`client.player_box_scores`)
- Getting team box scores by a date (`client.team_box_scores`)
- Getting the schedule for a season (`client.season_schedule`)
- Getting players totals for a season (`client.players_season_totals`)
- Getting players advanced season statistics for a season (`client.players_advanced_season_totals`)
- Getting regular season box scores for a given player and season (`client.regular_season_player_box_scores`)
- Searching (`client.search`)

You can see all methods used in this `repl` https://repl.it/@jaebradley/v300api-examples).

7. Inspiració.

Com se va comentar al primer punt, l'objectiu potencial d'aquest joc de dades és poder ser utilitzat per generar un model que predigui el rendiment generañ d'un jugador a un proper partit de la temporada.
Aquest model podria ser utilitzat tant per apostes esportives com per, de forma més innòcua, millorar el rendiment a l'hora de participar a webs de "Fantasy" de NBA (https://www.espn.com/fantasy/basketball/).
L'script s'ha fet suficientment parametritzable com per poder obtenir la resta de temporades si escau per fer una anàlisi partit a partit de vàries temporades consecutives.

8. Llicència.

S'ha seleccionat Unknown License, ja que els propietaris de les dades no són clars al respecte de l'ús que se li pot donar a les seves dades.

9. Codi.

https://github.com/Darksneer/NBAReferenceScrapper

El codi és la selecció de fitxers necessaris (main.py executable, scraper.py de llibreria i csv generat) del venv del projecte de PyCharm utilitzat per generar el csv.

10. Dataset.

https://zenodo.org/record/3744154#.Xo19bvHta00