



@wesportfr.com

Rapport du projet de Collecte et Stockage de données

Prédiction du salaire des joueurs de NBA en accord
avec leurs performances et statistiques sportives

Louis BERTHIER – Massiwa CHABBI – Thomas DUHAMEL – Malek FARHAT

TABLE DES MATIERES

I. Définition du sujet	3
1. Origines du projet	3
2. Problématique et intérêts du projet	3
II. Notre plan d'expériences et notre hypothèse	3
1. Notre hypothèse	3
2. Notre démarche afin de réaliser la prédiction	4
3. Notre plan d'expériences	4
4. Les limites de notre plan d'expériences	5
III. La collecte de données	5
1. Les sites utilisés	5
2. Les problèmes rencontrés	6
a. La connexion	6
b. Les observations	7
IV. Le stockage des données	8
1. A travers un CSV	8
2. A travers une BDD	9

I. Définition du sujet

1. Origines du projet

Le basketball, inventé par James Naismith en 1891 devient rapidement de plus en plus populaire en Amérique du Nord. C'est en 1949 que naît la National Basketball Association (NBA), qui aujourd'hui constitue **l'une des quatre ligues majeures** du sport américain. Nous avons tous les 4 un réel intérêt pour le sport et plus particulièrement pour la **NBA**. C'est pourquoi nous avons décidé **d'étudier les salaires des joueurs en fonction de leurs performances sportives**.

En voyant les salaires astronomiques de nombreux joueurs, nous nous demandons donc : **dans quelles mesures les performances sportives des joueurs affectent-elles le salaire de ces derniers ?**

2. Problématique et intérêts du projet

Malheureusement, à la suite de la crise sanitaire, les saisons ont été **perturbées**, on a donc décidé de se ramener à une **année dite « classique »**. C'est pourquoi nous étudierons **la saison régulière de l'année 2018-2019**.

Les données relatives à la NBA sont **libres d'accès** et retrouvables sur de nombreux sites internet. L'objectif à travers ce projet est de pouvoir **prédire le salaire d'un joueur de NBA en fonction de ses performances** au cours de la saison. Ainsi, un responsable marketing au sein d'une équipe peut essayer de **maximiser les gains et la rentabilité des échanges** de joueurs au cours de chaque nouvelle saison.

En parallèle, cela permettrait par exemple d'adapter les salaires des nouveaux joueurs qui font une première saison fulgurante, que l'on qualifie de « **Rookies** » et dont la valeur « marchande » peut **exploser** au cours de la saison. En parallèle, cela permet de vérifier la valeur des anciens joueurs mondialement connus, que l'on qualifie de « **Stars** » et dont les performances peuvent **chuter** d'une saison à l'autre.

II. Notre plan d'expériences et notre hypothèse

1. Notre hypothèse

L'objectif est de valider l'hypothèse suivante : « **De bonnes performances au cours d'une saison valorisent le salaire du joueur en question** »

2. Notre démarche afin de réaliser la prédiction

Par suite de notre pré-traitement des données et de la création de notre base de données, nous avons pour objectif de **prédire le salaire des joueurs**. Pour réaliser cela, différentes analyses peuvent être menées :

- **Une première approche générale** : on étudie les statistiques **dans leur globalité** notamment à travers le minimum, le maximum, les écarts-types, la moyenne, la médiane... On profite également de cette première approche pour **visualiser nos données** avec les graphiques associés ou des graphiques plus visuels comme les boîtes à moustaches pour **prendre réellement connaissance** de nos données.
- **L'apprentissage automatique avec apprentissage dit supervisé** : nous avons des **données labélisées** et un **problème de régression** (estimation du salaire). On peut notamment utiliser la **régression linéaire** comme modèle d'apprentissage.

3. Notre plan d'expériences

Pour notre plan d'expériences, voici les différentes composantes essentielles à sa compréhension :

- Notre population étudiée : **Les joueurs de NBA**
- L'échantillon prélevée : **Les 530 joueurs de NBA de la saison régulière 2018-2019**
- Nos caractéristiques d'entrée : **Le nom du joueur (en tant qu'index), le nombre de matchs joués, le ratio de victoires, le pourcentage de tirs réussis, le ratio de lancers francs réussis, le nombre de rebonds, le nombre de passes décisives, le nombre d'interceptions, le nombre de blocs et le nombre de points**

Attention : toutes ces statistiques sont établies en moyenne par match

- Notre réponse en sortie : **Le salaire annuel en dollars (\$)**

Dans le cadre de notre étude, nous avons décidé **d'enlever certaines caractéristiques** que l'on ne trouvait **pas pertinentes** ou **peu représentatives** vis-à-vis du salaire à savoir :

- L'ID du joueur
- L'ID de l'équipe
- L'abréviation du nom de l'équipe
- L'âge
- Le ratio de réussite des tirs à 3 points (étant donné que cette statistique est incluse dans le ratio des tirs réussis)
- Le nombre de pertes de balle
- Le nombre de fautes personnelles

4. Les limites de notre plan d'expériences

Notre étude est largement simplifiée et ne prend pas en compte de nombreux éléments. En effet nous avons :

- Tout d'abord un **nombre relativement faible de données** avec 530 observations.
- Également nous n'avons pas pris en compte la temporalité puisque nous nous focalisons uniquement sur **une seule saison**. Le salaire dépend aussi du salaire et des performances **des saisons précédentes ainsi que des performances hors saison régulière**.
- Enfin nous n'avons pas étudié **la publicité et la médiatisation** autour des joueurs. Le salaire dépend aussi de la façon dont le joueur est représenté à travers les médias, de son image de marque et de ses sponsors.

III. La collecte de données

1. Les sites utilisés

Nous avons récolté nos données par **web scraping**. Nos données ont été extraites des sites internet suivant :

- Pour les statistiques des joueurs :
<https://www.nba.com/stats/players/traditional/?sort=PTS&dir=-1&Season=2018-19&SeasonType=Regular%20Season>

NBA

NBA 75

Games

Schedule

News

Watch

Stats

Standings

Teams

Players

Fantasy

NBAbet

League Pass

Store

Tickets

Sign In

Stats

Home

Players

Teams

Leaders

Stats 101

Tools

Quick Links

Players General Traditional

SEASON

2018-19

SEASON TYPE

Regular Season

PER MODE

Per Game

SEASON SEGMENT

All Games

Advanced Filters

2018-19

0 RECENT FILTERS

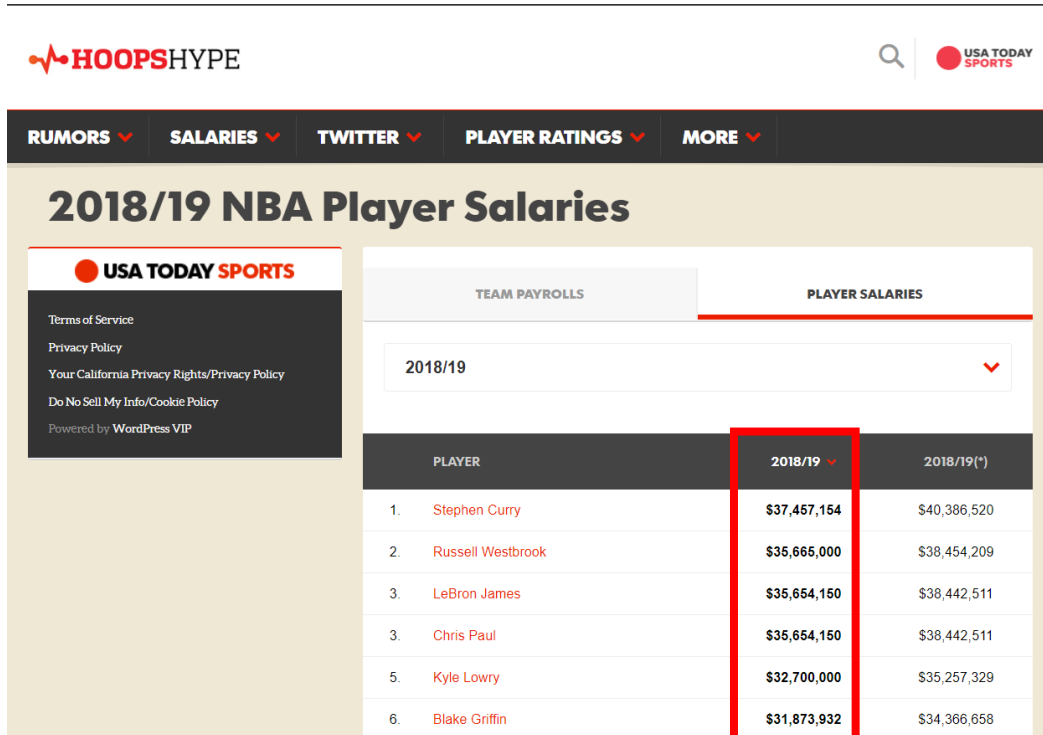
GLOSSARY

SHARE

530 Rows | Page 1 of 11

PLAYER	TEAM	AGE	GP	W	L	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	FP	DD2	TD3	+/-
1 James Harden	HOU	29	78	51	27	36.8	36.1	10.8	24.5	44.2	4.8	13.2	36.8	9.7	11.0	87.9	0.8	5.8	6.6	7.5	5.0	2.0	0.7	3.1	58.7	34	7	4.6
2 Paul George	OKC	29	77	46	31	36.9	28.0	9.2	21.0	43.8	3.8	9.8	38.6	5.9	7.0	83.9	1.4	6.8	8.2	4.1	2.7	2.2	0.4	2.8	49.3	24	1	6.4
3 Giannis Antetokounmpo	MIL	24	72	56	16	32.8	27.7	10.0	17.3	57.8	0.7	2.8	25.6	6.9	9.5	72.9	2.2	10.3	12.5	5.9	3.7	1.3	1.5	3.2	56.2	54	5	9.1
4 Joel Embiid	PHI	25	64	43	21	33.7	27.5	9.1	18.7	48.4	1.2	4.1	30.0	8.2	10.1	80.4	2.5	11.1	13.6	3.7	3.5	0.7	1.9	3.3	53.7	58	2	5.8
5 LeBron James	LAL	34	55	28	27	35.2	27.4	10.1	19.9	51.0	2.0	5.9	33.9	5.1	7.6	66.5	1.0	7.4	8.5	8.3	3.6	1.3	0.6	1.7	52.0	32	8	2.1
6 Stephen Curry	GSW	31	69	52	17	33.8	27.3	9.2	19.4	47.2	5.1	11.7	43.7	3.8	4.2	91.6	0.7	4.7	5.3	5.2	2.8	1.3	0.4	2.4	43.8	3	0	10.0
7 Kawhi Leonard	TOR	28	60	41	19	34.0	26.6	9.3	18.8	49.6	1.9	5.0	37.1	6.1	7.1	85.4	1.3	6.0	7.3	3.3	2.0	1.8	0.4	1.5	44.8	14	0	5.9
8 Devin Booker	PHX	22	64	15	49	35.0	26.6	9.2	19.6	46.7	2.1	6.5	32.6	6.1	7.1	86.6	0.6	3.5	4.1	6.8	4.1	0.9	0.2	3.1	40.8	12	0	-5.2
9 Kevin Durant	GSW	30	78	54	24	34.6	26.0	9.2	17.7	52.1	1.8	5.0	35.3	5.7	6.5	88.5	0.4	5.9	6.4	5.9	2.9	0.7	1.1	2.0	45.0	16	2	8.4
10 Anthony Davis	NOP	26	56	25	31	33.0	25.9	9.5	18.3	51.7	0.9	2.6	33.1	6.1	7.7	79.4	3.1	8.9	12.0	3.9	2.0	1.6	2.4	2.4	56.1	37	0	2.4

- Pour le salaire des joueurs :
<https://hoopshype.com/salaries/players/2018-2019/>



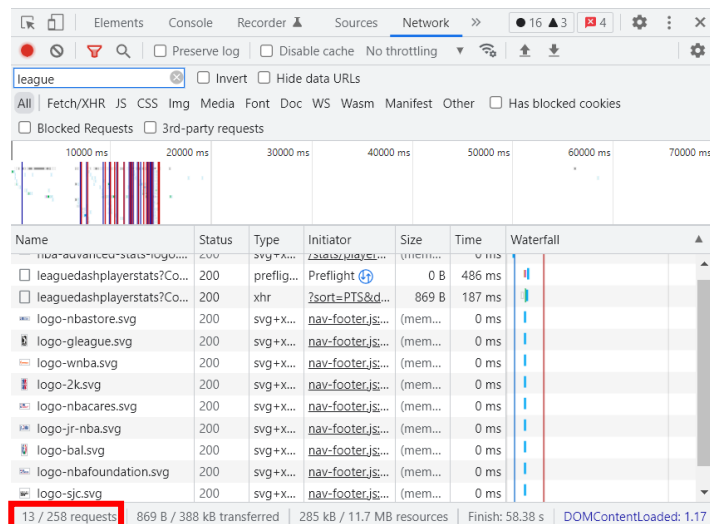
PLAYER	2018/19	2018/19(*)
1. Stephen Curry	\$37,457,154	\$40,386,520
2. Russell Westbrook	\$35,665,000	\$38,454,209
3. LeBron James	\$35,654,150	\$38,442,511
3. Chris Paul	\$35,654,150	\$38,442,511
5. Kyle Lowry	\$32,700,000	\$35,257,329
6. Blake Griffin	\$31,873,932	\$34,366,658

Pour le salaire, on considère le **salaire de gauche** à savoir celui qui n'a **pas été ajusté** pour faire face à l'inflation.

2. Les problèmes rencontrés

a. La connexion

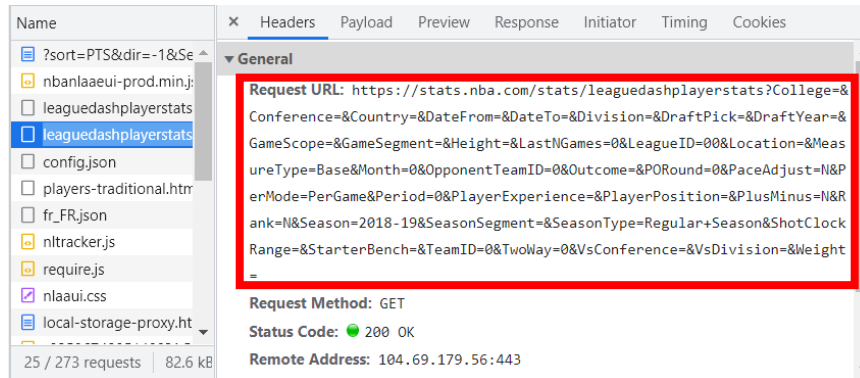
Au cours de notre récolte, nous avons été confrontés à des problèmes de connexions. Tout d'abord nous devons nous **connecter** afin de **récupérer un fichier JSON** afin de **récolter les données**. Cependant, nous avons passé un peu de temps à le trouver :



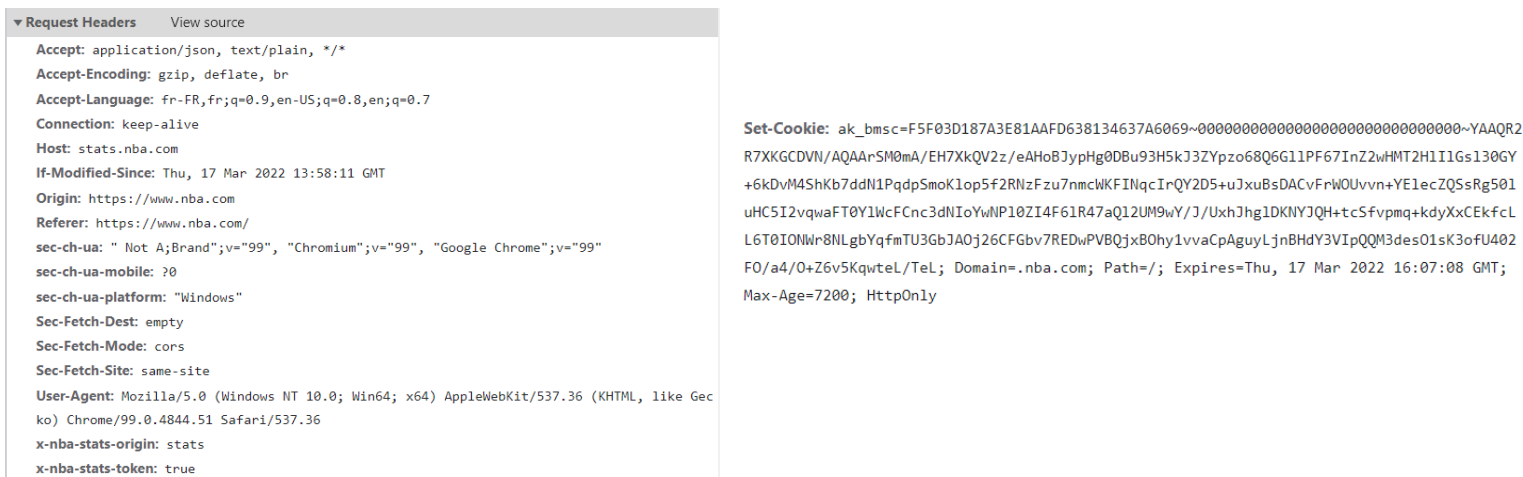
Name	Status	Type	Initiator	Size	Time	Waterfall
logo-sjc.svg	200	svg+x...	nav-footer.js...	(mem...)	0 ms	

13 / 258 requests | 869 B / 388 kB transferred | 285 kB / 11.7 MB resources | Finish: 58.38 s | DOMContentLoaded: 1.17 s

Une fois le fichier correspondant identifié, nous voulions nous connecter mais nous avons été **confrontés à un échec**.



En effet, récupérer simplement l'adresse n'était pas suffisant, nous avons donc **intégré l'ensemble des headers** ainsi que **le cookie** à notre requête :



b. Les observations

Nous n'avons pas un nombre de joueurs identiques entre l'accès aux statistiques et l'accès au salaire. En effet, sur la saison 2018-2019 nous avons **530 joueurs pour les statistiques** alors que pour **le salaire nous en avons 576**. On a donc **restreint l'étude** à l'ensemble des joueurs **présents sur les statistiques** de la saison 2018-2019 puisqu'il s'agit de l'ensemble **le plus petit** des deux. Cette différence peut s'expliquer par le fait que certains joueurs ont **arrêté de jouer** cette saison ou avant, cependant **ils possèdent un contrat encore effectif** et donc un salaire sur cette année (la durée des contrats est en général de 4 années).

Remarque :

Au début nous n'arrivions pas à **gérer l'accès aux données** statistiques par l'API. Nous avons donc opté pour **un autre site** où l'accès était beaucoup **plus simple**. Toutefois nous sommes revenus sur le site présenté au-dessus pour nous entraîner à **la récolte des données via la gestion des API et des cookies**.

Voici le site sur lequel nous nous étions rabattus :

https://www.basketball-reference.com/leagues/NBA_2019_per_game.html

2018-19 NBA Season

Standings

Schedule and Results

Leaders

Coaches

Player Stats

Other

2019 Playoffs Summary

Back to top

Totals

Per Game

Per 36 Min

Per 100 Poss

Advanced

Play-by-Play

Shooting

Adjusted Shooting

Player Per Game

Share & Export

☒ When table is sorted, hide non-qualifiers for rate stats

Glossary

Hide Partial Rows

Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	Alex Abrines	SG	25	OKC	31	2	19.0	1.8	5.1	.357	1.3	4.1	.323	0.5	1.0	.500	.487	0.4	0.4	.923	0.2	1.4	1.5	0.6	0.5	0.2	0.5	1.7	5.3
2	Quincy Acy	PF	28	PHO	10	0	12.3	0.4	1.8	.222	0.2	1.5	.133	0.2	0.3	.667	.278	0.7	1.0	.700	0.3	2.2	2.5	0.8	0.1	0.4	0.4	2.4	1.7
3	Jaylen Adams	PG	22	ATL	34	1	12.6	1.1	3.2	.345	0.7	2.2	.338	0.4	1.1	.361	.459	0.2	0.3	.778	0.3	1.4	1.8	1.9	0.4	0.1	0.8	1.3	3.2
4	Steven Adams	C	25	OKC	80	80	33.4	6.0	10.1	.595	0.0	0.0	.000	6.0	10.1	.596	.595	1.8	3.7	.500	4.9	4.6	9.5	1.6	1.5	1.0	1.7	2.6	13.9
5	Bam Adebayo	C	21	MIA	82	28	23.3	3.4	5.9	.576	0.0	0.2	.200	3.4	5.7	.588	.579	2.0	2.8	.735	2.0	5.3	7.3	2.2	0.9	0.8	1.5	2.5	8.9
6	Deng Adel	SF	21	CLE	19	3	10.2	0.6	1.9	.306	0.3	1.2	.261	0.3	0.7	.385	.389	0.2	0.2	1.000	0.2	0.8	1.0	0.3	0.1	0.2	0.3	0.7	1.7
7	DeVaughn Akoon-Purcell	SG	25	DEN	7	0	3.1	0.4	1.4	.300	0.0	0.6	.000	0.4	0.9	.500	.300	0.1	0.3	.500	0.1	0.4	0.6	0.9	0.3	0.0	0.3	0.6	1.0
8	LaMarcus Aldridge	C	33	SAS	81	81	33.2	8.4	16.3	.519	0.1	0.5	.238	8.3	15.8	.528	.522	4.3	5.1	.847	3.1	6.1	9.2	2.4	0.5	1.3	1.8	2.2	21.3
9	Rawle Alkins	SG	21	CHI	10	1	12.0	1.3	3.9	.333	0.3	1.2	.250	1.0	2.7	.370	.372	0.8	1.2	.667	1.1	1.5	2.6	1.3	0.1	0.0	0.8	0.7	3.7
10	Grayson Allen	SG	23	UTA	38	2	10.9	1.8	4.7	.376	0.8	2.6	.323	0.9	2.1	.443	.466	1.2	1.6	.750	0.1	0.5	0.6	0.7	0.2	0.2	0.9	1.2	5.6
11	Jarrett Allen	C	20	BRK	80	80	26.2	4.2	7.1	.590	0.1	0.6	.133	4.1	6.5	.629	.595	2.5	3.5	.709	2.4	6.0	8.4	1.4	0.5	1.5	1.3	2.3	10.9
12	Kadeem Allen	PG	26	NYK	19	1	21.9	3.4	7.4	.461	0.9	1.9	.472	2.5	5.5	.457	.521	2.2	2.8	.778	0.4	2.3	2.7	4.0	0.8	0.2	1.4	2.4	9.9
13	Al-Farouq Aminu	PF	28	POR	81	81	28.3	3.2	7.3	.433	1.2	3.5	.343	2.0	3.9	.514	.514	1.9	2.1	.867	1.4	6.1	7.5	1.3	0.8	0.4	0.9	1.8	9.4
14	Justin Anderson	SF	25	ATL	48	4	9.6	1.3	3.3	.408	0.5	1.6	.312	0.8	1.7	.500	.484	0.5	0.7	.743	0.5	1.3	1.8	0.5	0.5	0.3	0.5	1.0	3.7

IV. Le stockage des données

1. A travers un CSV

Après avoir eu accès à l'ensemble de nos données à la suite de la récolte, la **question du stockage** s'en suivait. Dans notre cas, nous avons tout d'abord exporté les données au **format CSV** afin de vérifier que la table créée était bien en accord avec notre dataframe et avec nos attentes. Voici l'excel obtenu :

PLAYER_NAME	GP	W_PCT	FG_PCT	FT_PCT	REB	AST	STL	BLK	PTS	SALARIES
Aaron Gordon	78	0.513	0.449	0.731	7.4	3.7	0.7	0.7	16.0	21590909
Aaron Holiday	50	0.62	0.401	0.82	1.3	1.7	0.4	0.3	5.9	1914480
Abdel Nader	61	0.623	0.423	0.75	1.9	0.3	0.3	0.2	4.0	1378242
Al Horford	68	0.603	0.535	0.821	6.7	4.2	0.9	1.3	13.6	28928710
Al-Farouq Aminu	81	0.642	0.433	0.867	7.5	1.3	0.8	0.4	9.4	6957105
Alan Williams	5	0.2	0.615	0.5	3.8	0.6	0.2	0.0	3.6	77250
Alec Burks	64	0.297	0.405	0.823	3.7	2.0	0.6	0.3	8.8	11536515
Alex Abrines	31	0.677	0.357	0.923	1.5	0.6	0.5	0.2	5.3	3667645
Alex Caruso	25	0.32	0.445	0.797	2.7	3.1	1.0	0.4	9.2	77250
Alex Len	77	0.364	0.494	0.648	5.5	1.1	0.4	0.9	11.1	4350000
Alex Poythress	21	0.333	0.494	0.621	3.6	0.8	0.2	0.5	5.1	77250
Alfonzo McKinnie	72	0.736	0.487	0.563	3.4	0.4	0.3	0.2	4.7	1349383
Alize Johnson	14	0.643	0.25	0.5	1.4	0.1	0.1	0.2	0.9	838464
Allen Crabbe	43	0.442	0.367	0.732	3.4	1.1	0.5	0.3	9.6	19332500
Allonzo Trier	64	0.203	0.448	0.803	3.1	1.9	0.4	0.2	10.9	3382000
Amile Jefferson	12	0.75	0.625	0.875	1.8	0.3	0.3	0.3	2.3	77250
Amir Johnson	51	0.627	0.503	0.756	2.9	1.2	0.3	0.3	3.9	2393887
Andre Drummond	79	0.506	0.533	0.59	15.6	1.4	1.7	1.7	17.3	25434262
Andre Iguodala	68	0.691	0.5	0.582	3.7	3.2	0.9	0.8	5.7	16000000
Andre Ingram	4	0.25	0.0	0.0	0.5	0.0	0.3	0.0	0.0	76236
Andrew Bogut	11	0.727	0.5	1.0	5.0	1.0	0.3	0.7	3.5	486892
Andrew Harrison	17	0.353	0.302	0.944	1.2	1.4	0.2	0.1	3.2	354500
Andrew Wiggins	73	0.425	0.412	0.699	4.8	2.5	1.0	0.7	18.1	25467250
Anfernee Simons	20	0.85	0.444	0.563	0.7	0.7	0.1	0.0	3.8	1837800
Angel Delgado	2	0.0	0.2	0.5	2.0	0.0	0.5	0.0	1.5	77250
Ante Zizic	59	0.22	0.553	0.705	5.4	0.9	0.2	0.4	7.8	1952760
Anthony Davis	56	0.446	0.517	0.794	12.0	3.9	1.6	2.4	25.9	25434263

2. A travers une BDD

Afin de répondre à la consigne, nous avons également opté pour le système de gestion de base de données relationnelle dit **SQLite** afin de créer **une véritable base** de données et non une simple table **contrairement au CSV**. En effet, nous avons dans le cas de notre projet **peu de données** et **non confidentielles**, ainsi SQLite pourra amplement **supporter cette charge**.

On aurait également pu se servir de **MySQL** qui permet d'avoir une **base protégée** ainsi que des **types de données plus flexibles** contrairement à SQLite qui est un peu plus restreint.

Toutefois, SQLite permet de stocker l'ensemble des informations **dans un fichier**, ce qui facilite la copie alors que pour MySQL on doit **tout condenser** dans un seul fichier pour permettre la copie ou l'export. De plus SQLite est **plus accessible** et ne **requiert que très peu d'assistance**.

Également notre projet est **relativement petit** et ne nécessite **pas beaucoup d'évolutivité** ainsi SQLite remplit parfaitement les conditions afin de stocker nos données.