# Equipe projet

* Fatiha LEHAS
* Stéphane MBOKO
* Sarah HIMEUR

# Description du projet

Les sports américains sont très friands de statistiques, et la NBA (National Basketball Association) ne fait pas exception à la règle.

Le développement constant des nouvelles technologies et des outils numériques, permet désormais de suivre en temps réel les déplacements de tous les joueurs sur un terrain de basket. Les données recueillies sont ainsi très nombreuses et riches.

Le but de ce projet est de :

1. Comparer les tirs (fréquence et efficacité au tir par situation de jeu et par localisation sur le terrain) de 20 des meilleurs joueurs de NBA du 21ème siècle.
2. Pour chacun de ces 20 joueurs encore actifs aujourd’hui (de LeBron James à Giannis Antetokounmpo), estimer à l’aide d’un modèle la probabilité qu’a leur tir de rentrer dans le panier, en fonction de différentes métriques.

# Problématique

Nous sommes face à un problème de **classification**, nous devons déterminer si le tir rentre ou non dans le panier en nous basant sur les données de localisation du tir, les performances du joueur et de son équipe, la situation de jeu.

# Datasets

* + dataset des tirs NBA entre 1997 et 2019 :

<https://www.kaggle.com/jonathangmwl/nba-shot-locations>

* dataset des actions de chaque match entre 2000 et 2020 :

<https://sports-statistics.com/sports-data/nba-basketball-datasets-csv-files/>

* dataset des bilans d’équipe entre 2014 et 2018 :

<https://www.kaggle.com/nathanlauga/nba-games?select=ranking.csv>

* dataset des joueurs de NBA depuis 1950 :

<https://www.kaggle.com/drgilermo/nba-players-stats?select=Players.csv>

**+** [**NBA.com**](https://www.nba.com/stats)

L’ensemble de ces données peuvent également être récupérées via des API qui scrappent NBA.com, le site officiel.

Notamment <https://github.com/swar/nba_api>, grâce aux [contributeurs](https://github.com/swar/nba_api/graphs/contributors) que nous pourrons utiliser afin d’obtenir les données actualisées depuis le site NBA.com.

L’API fournit de nombreuses méthodes permettant de récupérer les joueurs et leur stats, l’historique des matches et le play by play pour chacun d’eux.

Nous pouvons trouver sur le site NBA la liste des meilleurs joueurs de tous les temps, filtrée sur les joueurs encore actifs : <https://www.nba.com/stats/alltime-leaders?ActiveFlag=Yes>

# Exploration des données

## Play by Play

Les [fichiers csv](https://sports-statistics.com/sports-data/nba-basketball-datasets-csv-files/) recensent pour chaque année entre 2000 et 2020 l’intégralité des actions de jeu, de l’entre-deux jusqu’au buzzer final. Ils contiennent chacun entre 500k et 600k lignes.

Le format des fichiers de 2000 à 2019 est le même puis il change pour la saison 2019-2020. Il conviendra d’uniformiser les données lors du preprocessing.

### EVENTMSGTYPE

Le champ **EVENTMSGTYPE** donne le type d’action : panier mis, panier raté, lancer franc, rebond, faute, entre-deux, etc.

Nous nous intéresserons aux catégories 1 (FIELD\_GOAL\_MADE), 2 (FIELD\_GOAL\_MISSED), 3 (FREE THROW). A graph with different colored bars

Description automatically generated

La catégorie des lancers francs contient tous les lancers francs. Il convient de les recatégoriser en réussis/ratés. L’information se trouve dans les champs HOMEDESCRIPTION ou VISITORDESCRIPTION qui contiennent 'MISS' lorsque le tir est raté.

### EVENTMSGACTIONTYPE

Lorsque EVENTMSGTYPE est égal à 1 ou 2, le champ **EVENTMSGACTIONTYPE** indique le type de tir :

A graph with different colored bars

Description automatically generated

A graph of a number of data

Description automatically generated with medium confidence

Les types de tir pourront être regroupés en catégories (par exemple 3-pt jump shot, 2-pt layup, 2-pt jump shot, 2-pt dunk, 2-pt hook shot, free throw) afin de réduire le nombre de catégories. Il s’agit des types de tir qui ont été conservés lors du changement de format en 2019-2020.

### PCTIMESTRING

Temps restant dans le quart-temps. De 12:00 à 0:00 pour les 4 premiers. De 5:00 à 0:00 en cas de prolongations.

### PERIOD

Le champ période indique le quart-temps en court. De 1 à 4 par défaut. En cas de d'égalité on ajoute des prolongations de 5 minutes. On constate que la réussite au tir baisse au fil du jeu :

A graph with blue lines

Description automatically generated

### PLAYER1\_ID, PLAYER1\_NAME, PLAYER1\_TEAM

Ces champs nous indique quel joueur tire et son camp. Ci-dessous le nombre de paniers recensée pour les 20 meilleurs joueurs sur la saison 2000-2001 :

A graph of a number of data

Description automatically generated with medium confidence

### SHOTDISTANCE

La distance est extraite des champs HOMEDESCRIPTION et VISITORDESCRIPTION. On remarque que les tirs réussis sont en moyenne plus proches du panier. Les tirs ratés montrent un grand nombre d’outliers qui peuvent correspondre aux tentatives de tir très éloignées du panier dans les dernières secondes de jeu.

A diagram of a graph

Description automatically generated with medium confidence

# Prochaines étapes :

* Preprocessing :
  + Retirer les colonnes et lignes inutiles et catégoriser les données.
  + Sélectionner les 20 joueurs à observer et supprimer les autres données
  + Pour chaque tir, joindre les informations sur le joueur et la localisation du shoot
  + Concaténer les données de 2000 à 2020
  + Etudier les multi colinéarités
  + Réduire les dimensions
* Modélisation

## NBA shot locations

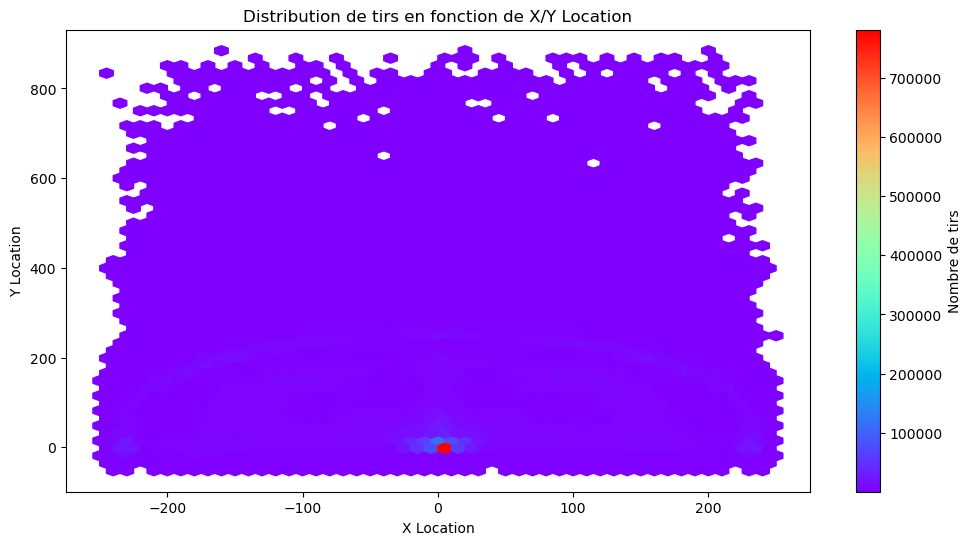
Le jeu de données en question dans cette partie montre les emplacements des tirs de l’année 1997 jusqu’à 2020. Plus précisément, les tirs ont été recensés en fonction des types des saisons, match et leurs dates, événements correspondantes, joueurs impliqués et leurs équipes, périodes du match, secondes et minutes restantes, types d’actions effectuées, type de tir, zone spécifique sur le terrain et zone basique d'où le tir a été effectué, distance exacte et portée de la zone de tir ainsi que les coordonnées X et Y de la position du tir sur le terrain.

Ce jeu de données contient autour de 4.7 millions de lignes sans aucune valeur manquante ni doublon.

La colonne “Game Date” est du type int64, on doit la convertir au bon format date pour pouvoir l’exploiter

### X/Y Location

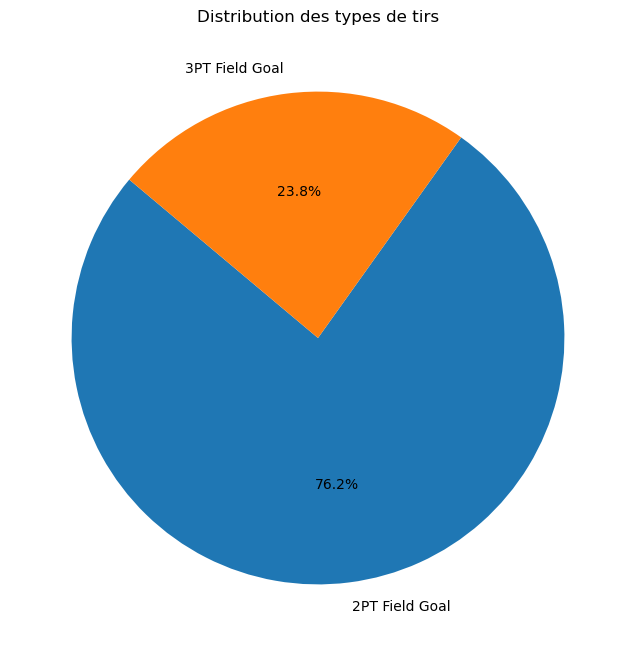
La figure représente la distribution des tirs en fonction de la localisation X-Y. La distribution des tirs n'est pas uniforme. Il y a des zones de forte concentration de tirs, notamment dans la partie centrale X,Y = (0,0) c’est qui logique puisque le panier se situe à ce point . La zone la plus éloignée du centre semble avoir une densité de tirs plus faible. Finalement, la faible densité de tirs dans la zone la plus éloignée du centre pourrait indiquer que les tireurs n'avaient pas la capacité d'atteindre cette zone ou qu'ils ne considéraient pas cette zone pour une action de tir.



### 

### Shot Type

Cette figure représente la distribution des types de tirs. Les tirs pour marquer 2 points sont dominants avec un pourcentage de 76%, ce qui suggère qu'il s'agit d'une stratégie de tir plus courante ou plus efficace. Pour les tirs de 3 points qui représentent 23% du total des tirs. ils sont utilisés effectués moins fréquemment, ce qui pourrait indiquer un taux de réussite plus faible ou une préférence pour les tirs à plus courte portée.



### 

### 

### Shot Zone Basic

Le graphique à barres montre le nombre de tirs par zone de tir de base. Il existe une tendance générale à la baisse du nombre de tirs dans les coins droit et gauche. La zone restricted area compte le plus grand nombre de tirs, suivie par Mid-Range, Above the break et In The Paint (Non-RA). Les joueurs peuvent effectuer un grand nombre de tirs dans restricted area ce qui est raisonnable vu que c’est proche du panier. Le nombre de tirs dans ‘Mid-Range’ est plus important que dans la zone ‘in the paint’ malgré que ce dernier est plus proche du panier. Cela est dû à plusieurs facteurs: le tactique adapté dans le jeu, l’intensité de la défense dans cette zone, la capacité du joueur.

### 

### 

### Period

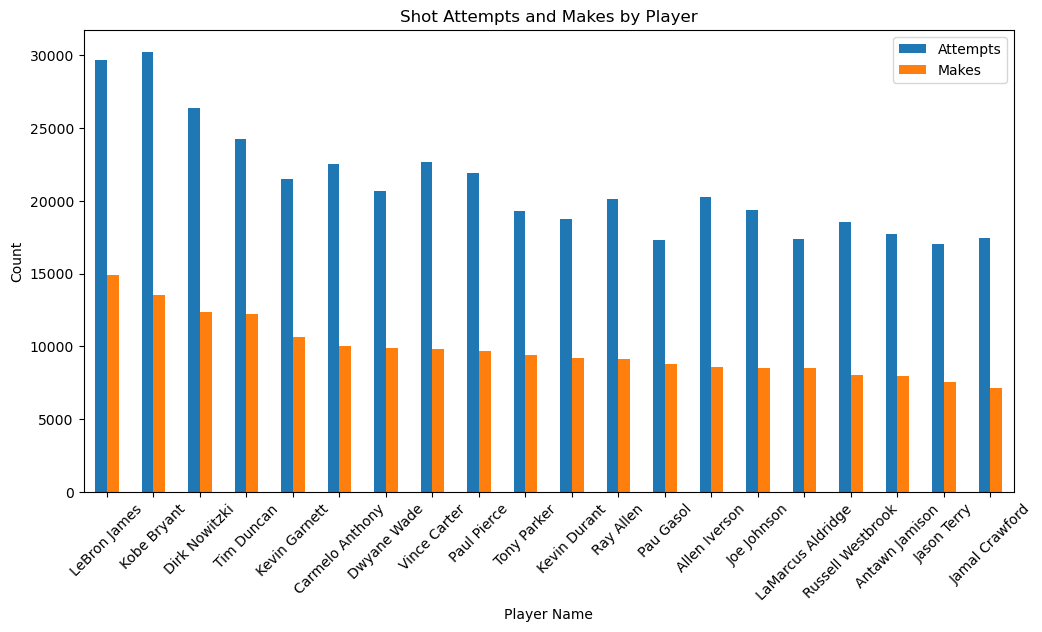
Cette figure représente la distribution des tirs en fonction des périodes de match. Les périodes 1 à 4 sont celles où la majorité des tirs ont été effectués.

### 

### 

### Tirs tentatives/marqués par joueur

La figure ci-dessous montre le nombre de tirs tentés et marqués pour les top 20 joueurs par ordre décroissants de nombre de tirs marqués.



## NBA player stat

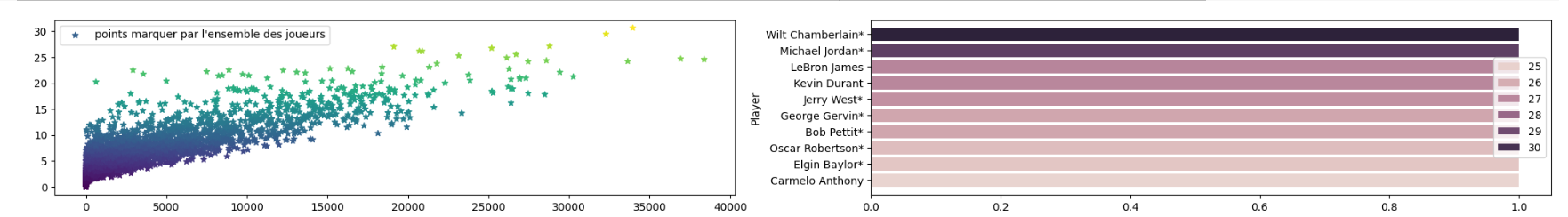
Les fichiers csv recensent pour chaque année les stats des joueurs durant leur carrière ainsi que leur année de naissance, leur université, etc…

Dans le cadre du projet, nous allons avant tout nous intéresser au stats des joueurs.

Les stats données dans les différents csv étant les stats cumulé sur une saison et non pas la moyenne pour chaque match, il a fallu faire quelque modification pour avoir un tableau de chaque joueur nba avec leur stat en moyenne par match au cours de leurs carrière.

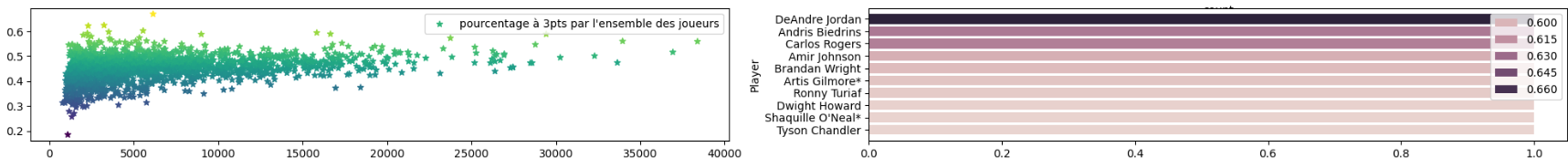
### Meilleur marqueur de l’histoire

On peut voir dans le graphe ci-dessous les meilleur marqueur de l’histoire au nombre de points marqueur mais aussi à la moyenne par match.



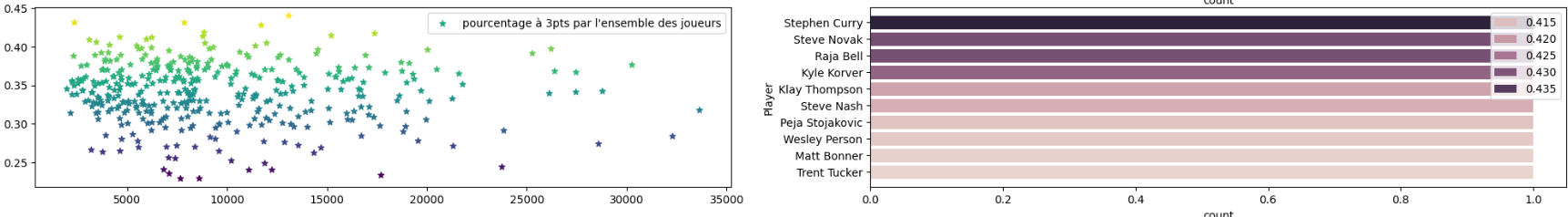
### Meilleur Shooter à 2pts

Ici nous avons tout les shoots pris après la ligne à 3pts, ainsi que les dunk. résultats les meilleurs moyenne par match sont forcément tenu par des pivot qui vont dans la plupart des cas ne faire que des dunk.



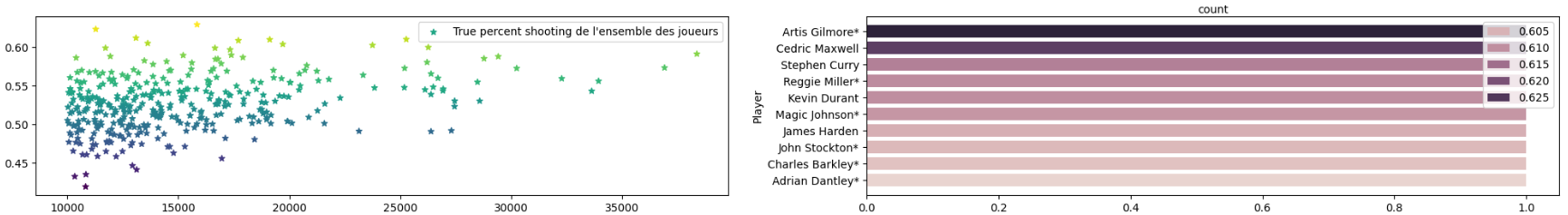
### Meilleur Shooter à 3pts

Nous avons là fait le choix de ne pas prendre les joueurs ayant tenté moins de 1000 shoot à 3Pts dans l’équation car le volume devient trop faible pour ce faire une réelle idée



### Meilleur TS

Le true shooting percentage permet d’évaluer l’efficacité d’un joueurs lors d’une phase offensive en pondérant la valeur d’un shoot (2pts et 3pts) et en prenant en compte le nombre de lancers francs mis



# la suite

regarder l’efficacité d’un joueur en fonction de l’équipe qu’il rencontre. Ceci serait à fairedans un premier temps avec 20 des meilleurs joueur du 21ème siècle selon espn, avant de potentiellement le faire pour l’ensemble des 3 meilleur marqueurs de chaque équipes