**Rapport de projet**

**Analyse des tirs de joueurs NBA**

**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 (selon ESPN).
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.

La difficulté de ce projet réside dans l’élaboration d’un modèle fiable, ainsi que dans le fait que de nombreux autres paramètres, parfois difficilement quantifiables ou difficiles à trouver, peuvent avoir une influence importante sur l’issue d’un tir :

la pression défensive exercée sur le tireur, la force du contact avec un défenseur, l’orientation des appuis du tireur face au panier, la bonne tenue du ballon en main, la qualité de la dernière passe avant le tir, l’état physique du joueur à ce moment, etc.

**Données :**

* + 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/>

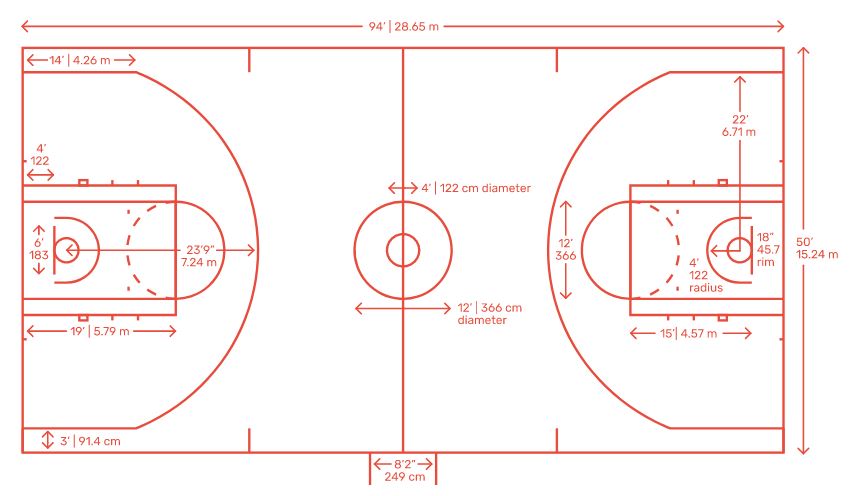
* + dataset des bilans d’équipe entre 2014 et 2018 :

<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>

**Dimensions d’un terrain de basket :**

****

**Exploration et Data visualisation**

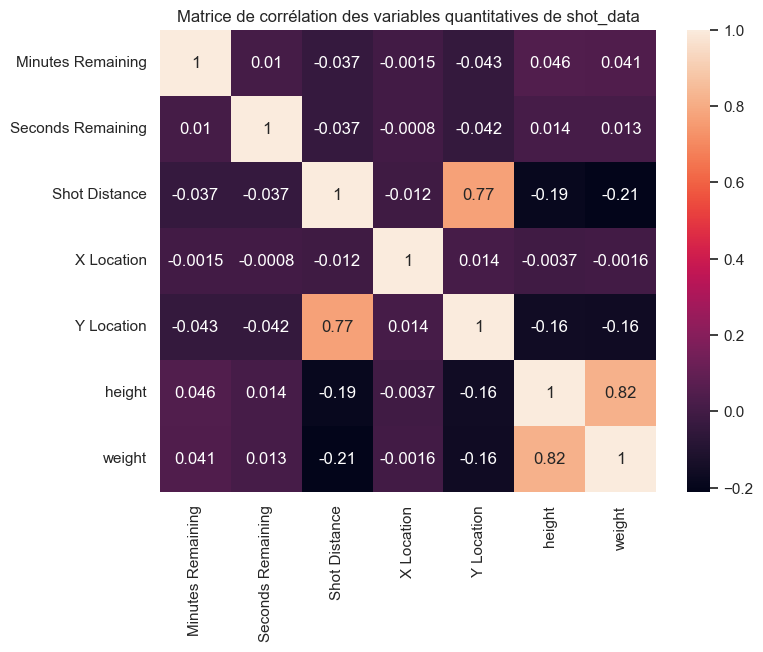
L’analyse des différentes variable de chaque dataset est présentée et résumée dans le document suivant :

[**rapport d'exploration des données**](https://docs.google.com/spreadsheets/d/10T5VdpQMjAsnyxNMDMqfEV9x0by2_Xxe/edit?usp=share_link&ouid=115159239274882355549&rtpof=true&sd=true)

Parmi les différentes variables présentes dans les datasets, la variable *Shot Made Flag* est la variable cible puisqu’elle indique si un tir est réussi (valeur = 1) ou non (valeur = 0). Les variables *X Location* et *Y Location* indiquent les coordonnées du tir par rapport au panier comme illustré sur la figure suivante :



Pour avoir un premier aperçu des liens entre les différentes variables quantitatives présentes dans les données des tirs, on peut commencer par s’intéresser à la matrice de corrélation :



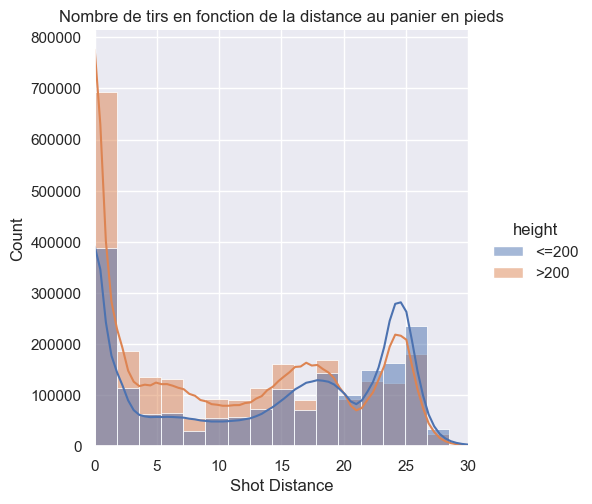
On remarque des corrélations attendues :

* le poids et la taille d’un joueur sont fortement corrélées
* la variable *Y Location*, qui représente la coordonnée sur la longueur du terrain,est corrélée avec *Shot Distance*

On note également une anti-corrélation faible entre la taille d’un joueur, *height*, et la distance du tir *Shot Distance*. On peut vérifier visuellement cette relation en discrétisant la variable *height* en 2 catégories :

1. Joueurs de plus de 2 m
2. Joueurs de moins de 2m

On peut alors s’intéresser au nombre de tirs effectués en fonction de la distance pour ces deux catégories :



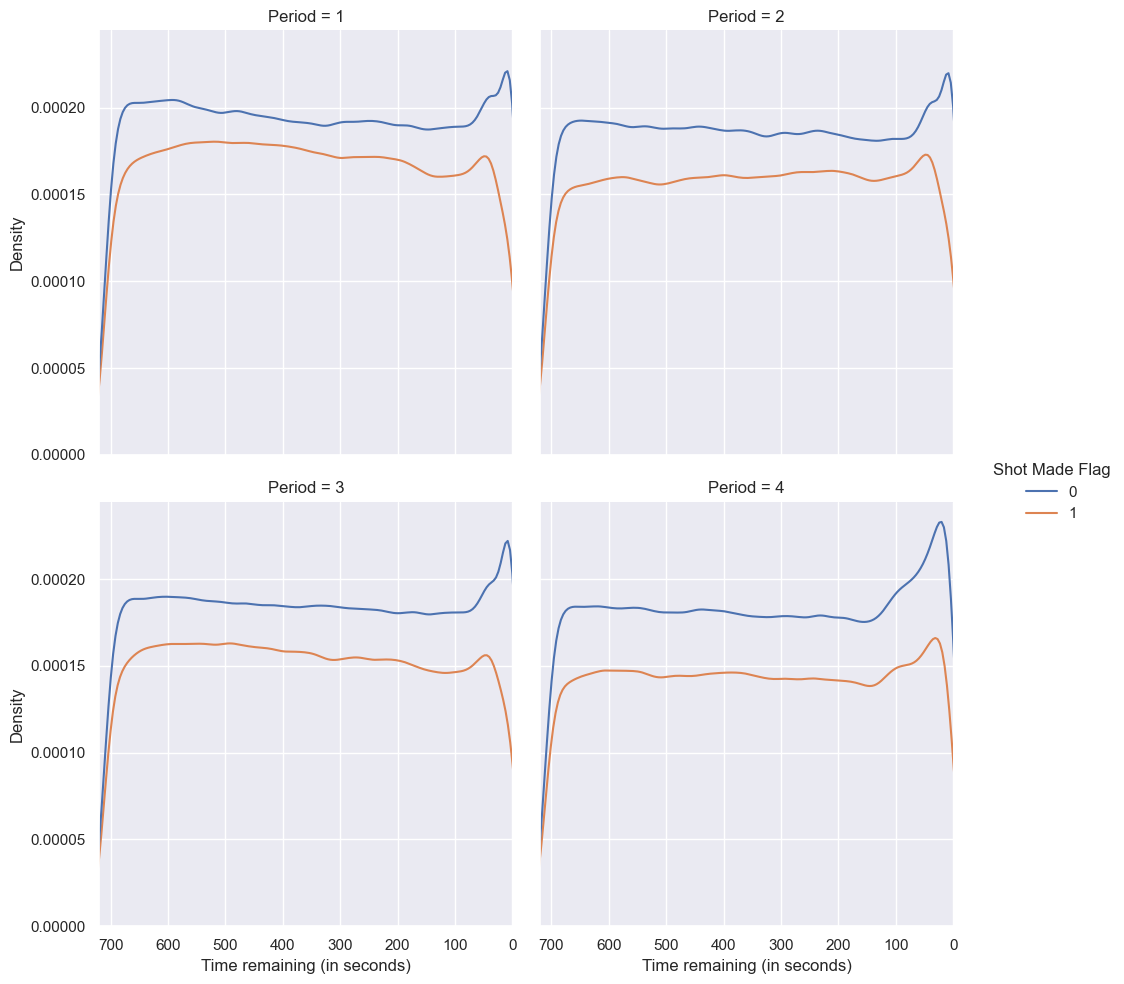
On remarque effectivement que les joueurs de petite taille (<2m) tentent plus de paniers à 3 points que les grands. Le test de corrélation de Pearson entre *height* et *Shot Distance* renvoie :

*Coefficient de correlation = -0.18563341147377913*

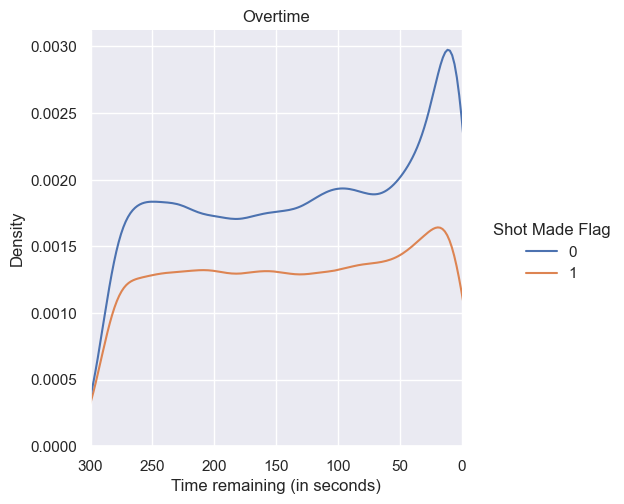
*p\_value = 0.0*

Les deux variables ne sont donc pas indépendantes.

Proportions de tirs ratés ou réussis en fonction du temps restant dans les différents quart-temps (*Period*) :



Même figure mais uniquement pour les prolongations :



On peut également s’intéresser aux performances au tir des 20 meilleurs scoreurs encore actifs par rapport à celles de l’ensemble des joueurs. En sélectionnant uniquement ces 20 joueurs, il est possible d’introduire des biais dans la distribution des variables explicatives. On peut par exemple se demander si les tailles et les poids de ces 20 joueurs sont bien distribuées :

