Pirlouit Dumez

Hakim Duparcq

Destiné Makani

Guillaume Prévost

Projet Python Scripting

Analyse des données de la NBA

Sommaire

Introduction

Dans le cadre du module Python Scripting proposé en première année de master aux étudiants, nous sommes amenés à réaliser un projet en groupe. Ce projet consiste en l'analyse des données de la NBA (basketball américain), ainsi que leur récupération préalable grâce à un algorithme de web scraping.

Notre but dans ce projet est de récupérer des données sur le site de la NBA, d'analyser ces données et de réaliser des graphiques qui puissent mettre en avant des informations utiles et intéressantes. Il est également proposé aux étudiants d'implémenter un algorithme de prédiction de leur choix.

Pour commencer, nous résumerons les fonctionnalités de notre programme et expliquerons comment l'utiliser. Dans un second temps, nous mettrons en avant l'algorithme de web scraping et son fonctionnement. Enfin, nous procèderons à une analyse approfondie des graphiques et de l'algorithme de prédiction.

Fonctionnalités

Notre programme implémente les fonctionnalités suivantes :

- Scraping des données sur le site officiel de la NBA
- Analyse et processing des données
- Visualisations de ces données
- Prédictions simples

Nous avons également créé un dashboard interactif avec ces graphiques et nous l'avons mis en ligne à l'adresse http://guillaume.pythonanywhere.com

L'ensemble de notre projet comprend cinq fichiers python (.py), ainsi qu'un notebook python (.ipynb). Notre projet est disponible sur GitHub a l'adresse : https://github.com/PirlouitDumez/NBA_Visualization

. Voici un aperçu de l'utilité et du fonctionnement de chaque fichier.

- NBA_official_data_scraping_stats.py : Ce fichier permet de scraper les statistiques des joueurs et des équipes sur le site de la NBA, et les télécharge au format csv. Pour scraper les données, nous utilisons Selenium. Il est facile de scraper une nouvelle table de statistiques, en créant simplement le dictionnaire correspondant (qui contient l'url, le nombre de pages de la table, etc.). Nous avons scrapé six tables de statistiques comme la suivante : https://www.nba.com/stats/players/traditional/?sort=PTS&dir=-1
- NBA_official_data_scraping_matchs.py : Ce fichier permet de scraper le nom des équipes et leurs scores pour tous les matchs qui se sont déroulés depuis le premier janvier 2009. Les données sont enregistrées au format csv. Pour cela nous utilisions l'url https://www.nba.com/games?date=2009-04-15 auquel nous modifions la fin de la chaine de caractère pour passer au jour suivant.
- NBA_official_data_procecing_match.py : Ce fichier permet a partir du dataset des scores des match des equipes d'optenir le pourcentage moyen de victoires par équipes par années. Ce fichier supprime les valeurs manquantes ainsi que les équipes n'étant pas de vraie club (Team Curry Stephen)
- NBA_official_data_processing.py : Ce fichier permet de traiter les datasets scrapés par exemple en retirant les valeurs manquantes (il y en a très peu), ou en mergeant par le nom des joueurs les différentes tables de statistiques téléchargées. Il enregistre au format csv le nouveau dataset résultant.
- NBA_official_visualization.py : Ce fichier contient les codes pour créer tous les graphiques grâce à la librairie Plotly. S'il est lancé directement, il ouvrira chaque graphique dans un onglet du navigateur de l'utilisateur.
- dashboard.py : Ce fichier contient le code nécessaire à la création du dashboard, créé avec la librairie Dash. Pour voir ce dashboard, l'utilisateur peut exécuter le fichier et se rendre sur l'adresse http://127.0.0.1:8050/, ou bien se rendre directement sur le site http://guillaume.pythonanywhere.com/ disponible en ligne.
- PlayersRolePrediction.ipynb : Ce notebook contient l'ensemble du code (de l'exploration des données au développement du modèle) réalisé pour concevoir un modèle capable de prédire le poste d'un joueur sur le terrain en fonction de ses statistiques.

Nous n'avons pas inclus dans le rendu de notre projet les datasets intermédiaires (chaque table de statistique non traitée) dans un souci de clarté. Ils ne sont en effet qu'une étape pour obtenir le dataset nettoyé et mergé qui servira à l'analyse des données et leur visualisation.

Web Scraping

Pour réaliser notre projet et proposer des graphiques de qualité, nous avons besoin de beaucoup de données. Afin d'assouvir ce besoin, nous nous sommes référés au site officiel des statistiques de la NBA : https://www.nba.com/stats/

Ce site contient beaucoup de tables de statistiques tel que celle-ci, qui contient les statistiques traditionnelles de chaque joueur sur la saison : https://www.nba.com/stats/players/traditional/.

Toutes les tables sont présentées au même format, nous avons donc implémenté un algorithme de scraping général permettant de récupérer les informations de n'importe quelle table sur le site. Il suffit de créer un dictionnaire avec quelques informations spécifiques à la page (url, nom du fichier, nombre de pages à scraper, etc.) pour scrapper une nouvelle table.

Pour réaliser notre algorithme, nous avons utilisé Selenium. Le programme récupère les informations de la table en bouclant sur chaque ligne et chaque page de la table (pour les joueurs, il y a 10 pages de 50 lignes) et les enregistre ensuite au format csv. Les données seront ensuite nettoyées et traitées.

Certaines données nécessitent un scraping particulier. C'est par exemple le cas de la vitesse moyenne des joueurs (https://www.nba.com/stats/players/speed-distance/). Les sélecteurs css de cette page sont légèrement différents, nous avons donc adapté notre algorithme pour cette page.

D'autres données comme par exemple la taille, le poids ou le pays des joueurs proviennent du site Beinsport. Effectivement, le lien https://nba.com renvoie à la section nba du site beinsport. Nous avons commencé à scrapper ce site avant de trouver le site officiel des statistiques.

Pour récupérer les données sur les matchs (https://www.nba.com/games) nous avons bouclé sur chaque jour et récupéré les scores des matchs de la journée. Les données sont également enregistrées au format csv.

Data Science

Dashboard

Nous avons déployé un dashboard interactif disponible au lien suivant : http://guillaume.pythonanywhere.com/

Ce dashboard contient quelques visualisations et offre à l'utilisateur la possibilité d'interagir avec ces graphiques. Par exemple, il peut comparer les joueurs de son choix ou obtenir des informations supplémentaires en mettant la souris sur les données des graphiques.

Visualisations

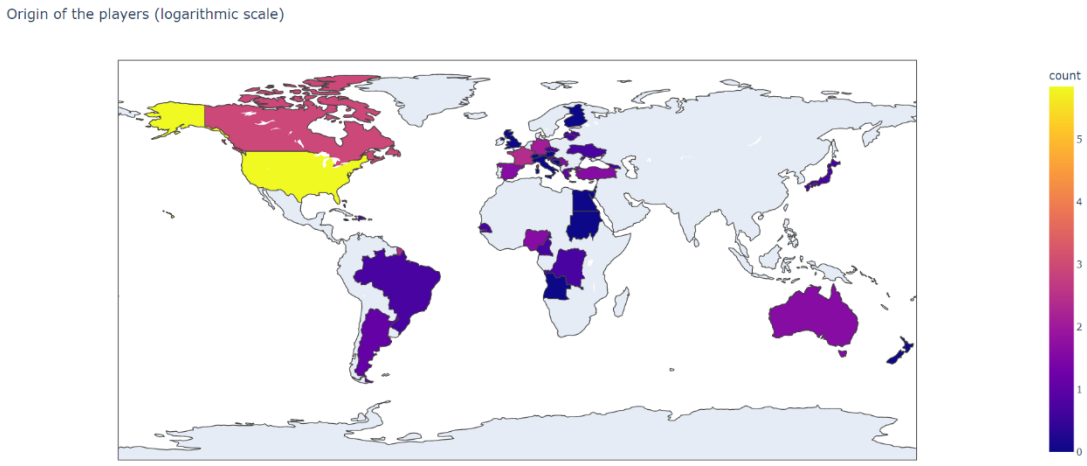
Pour cette partie, nous avons orienté les visualisations sur plusieurs axes, afin de comprendre mieux certains aspects de la NBA.

Les graphiques sont créés grâce à la librairie Plotly afin de fournir une meilleure interaction avec l'utilisateur.

Généralités

Les graphiques ci-dessous ont pour but de fournir un aperçu global des joueurs de la NBA.

D'où viennent les joueurs de la NBA ?

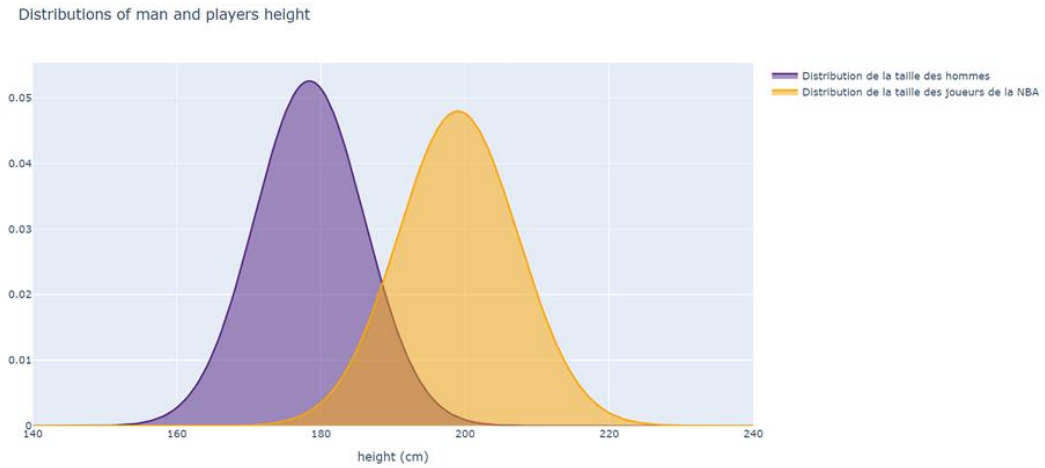


Sans surprise, la grande majorité des joueurs de la NBA viennent des Etats-Unis et du Canada, mais il y a également un pôle important en Europe.

L'échelle des couleurs est logarithmique afin que les nuances de couleurs soient plus visibles.

Analysons maintenant la taille des joueurs de la NBA.

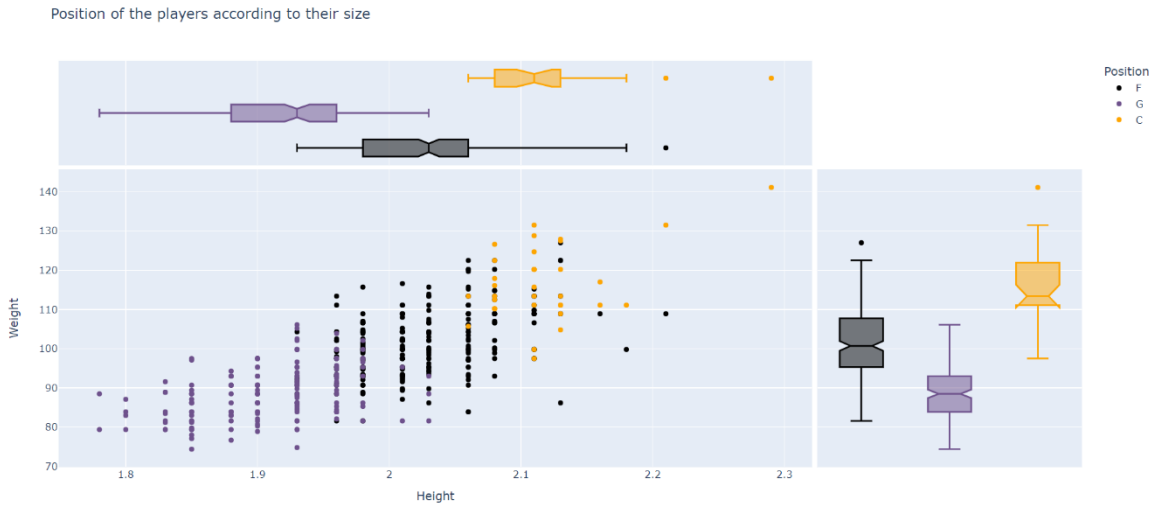
Quelle est la taille des joueurs de basketball comparé à la taille des hommes ?



Rien de surprenant ici, les joueurs de la NBA (en moyenne 199.3cm) sont significativement plus grands que les hommes dans le monde (en moyenne 178.7cm). Cohérant car la taille est un atout majeur pour pouvoir contrôler le terrain, la technique de dunk notamment demande une grande taille.

Postes des joueurs

<u>La position des joueurs a-t-elle un rapport avec les tailles des joueurs ?</u>



Cette section a pour but de clarifier les postes des joueurs. Les joueurs de la NBA peuvent avoir trois postes : C, F et G. A quoi cela correspond-il ?

Au vu de cette information, nous pourrions faire une supposition sur les postes :

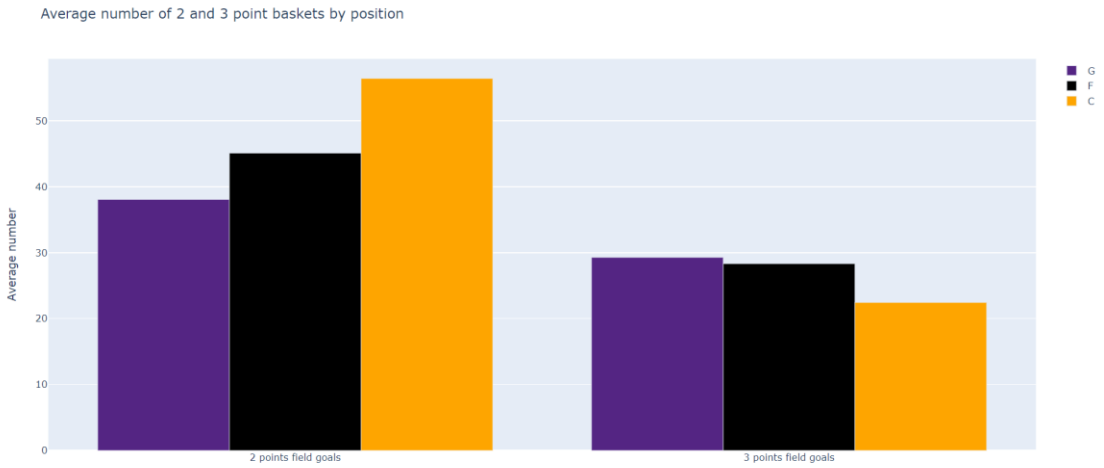
- G : Les joueurs petits, qui mettent les paniers de loin (3pts)

- F : Les joueurs polyvalents, qui mettent des paniers à 3pts et 2pts

- C : Les grands joueurs le plus proche du panier, qui mettent les paniers à 2pts (dunks)

Il est intéressant de relever la corrélation importante entre le poste et le gabarit des joueurs. Il y a une corrélation Pearson de 0.79 entre la poste et la taille.

Il est également intéressant de relever les gabarits particuliers comme par exemple les joueurs grands et légers ou petits et lourds. Ci-dessous Zion Williamson et Aleksej Pokusevski ,deux exemples trouvés grâce au graphique.

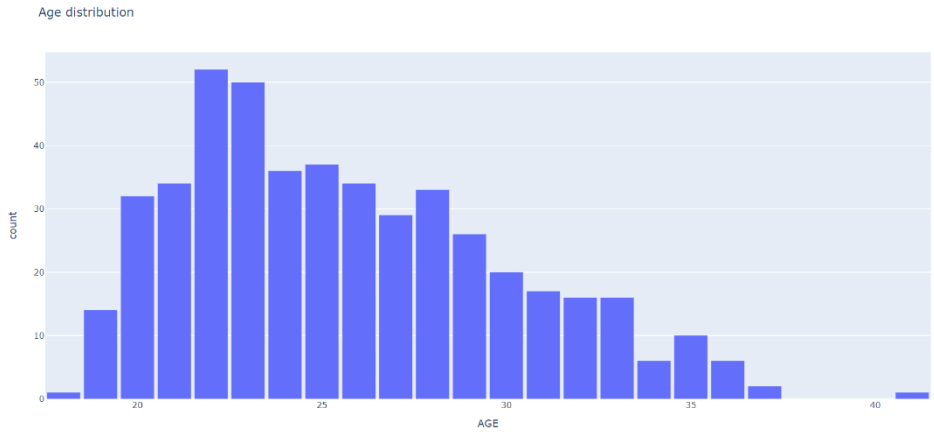


Ce graphique montre que les joueurs au poste G marque en proportion plus de paniers à 3 points que de paniers à 2 points, à l'inverse des joueurs de poste C. Cela confirme notre suggestion précédente sur les postes. Une recherche sur internet (source : Wikipédia) corrobore nos visualisations en définissant les postes ainsi :

- **G (Guard)** : La plupart des joueurs G sont prolifiques à partir de la plage à trois points
- **F (Forward)** : Le joueur Forward est souvent considéré comme le plus polyvalent des postes de basketball
- **C (Center)** : Le joueur centre joue généralement près de la ligne du fond, ou du panier.

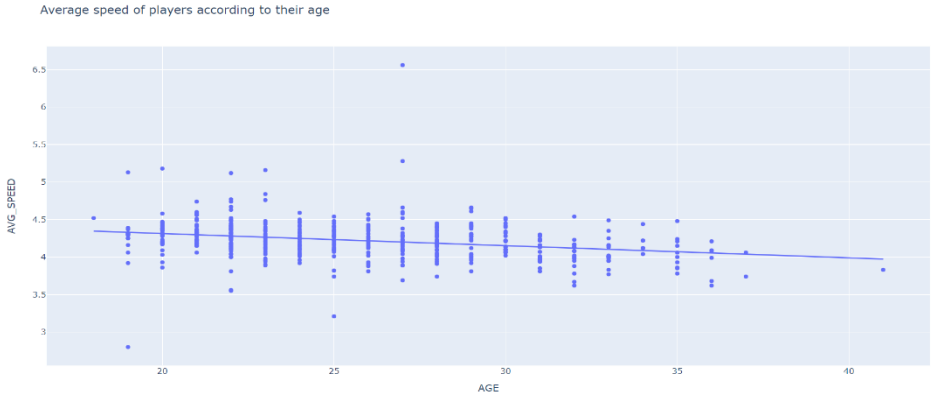
Performance en fonction de l'âge

Le graphique ci-dessous mets en évidence la répartition des âges au sein de la NBA.



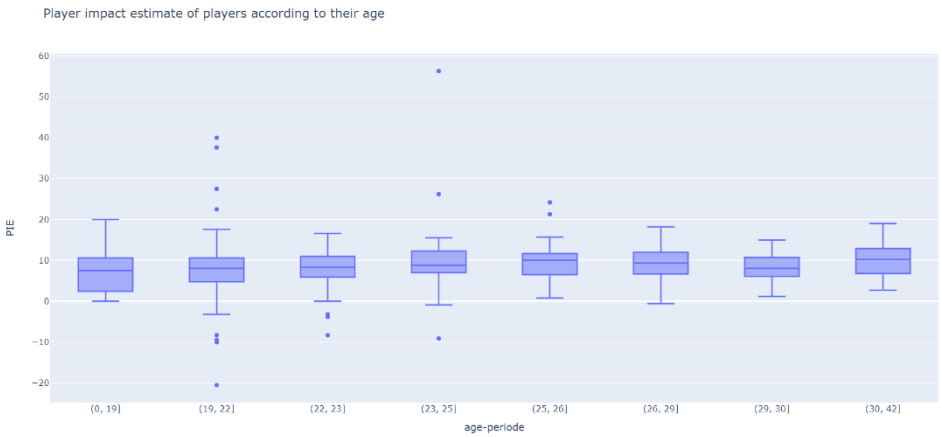
La grande majorité des joueurs de NBA est composé de joueurs dont l'âge est inférieur à la trentaine. L'âge moyen au cours de la saison 2021-2022 est estimé à 25.7 ans.

Evaluons maintenant la vitesse moyenne en fonction de l'âge, cela est présenté sur le graphique ci-dessous.



La droite de régression linéaire n'est pas horizontale, on constate une diminution de la vitesse moyenne en fonction de l'âge.

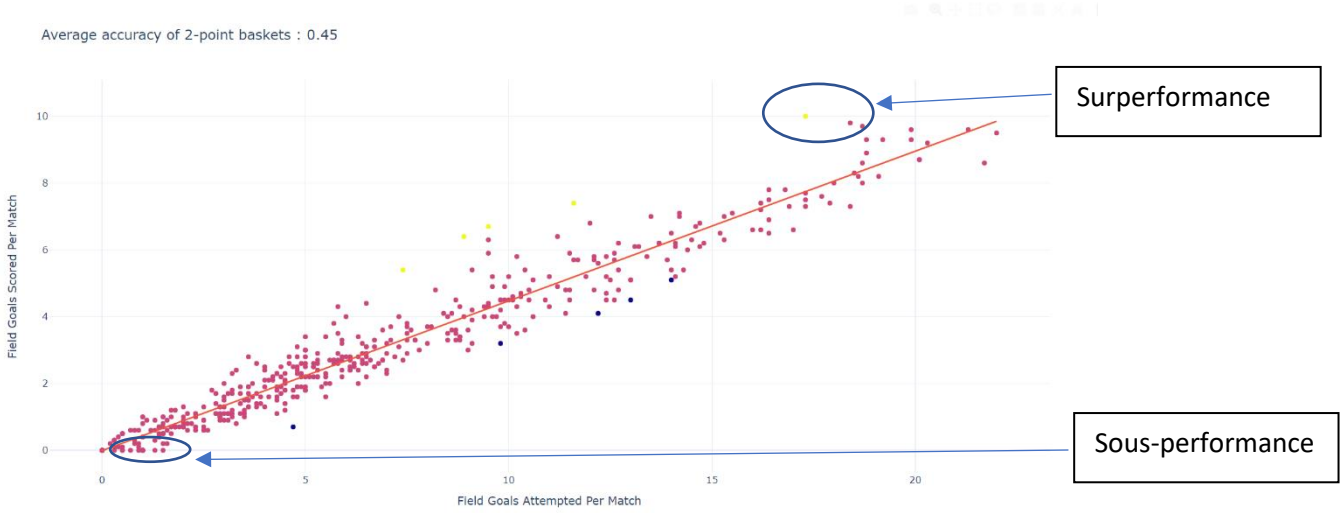
Evaluons maintenons l'impact de l'âge classé en plusieurs catégories (à savoir 0-19,19-22,22-23,23-25,25-26,26-29,29-30 et 30-42) sur le PIE (Player Impact Estimate), c'est un indicateur permettant de définir quel est le pourcentage d'actions positives d'un match effectuées par un joueur lorsqu'il est sur le terrain. Un PIE% supérieur à 10 est souvent l'indicateur d'un joueur de qualité car celui-ci fera en moyenne plus d'actions positives que les neuf autres joueurs sur le parquet.



On constate que la médiane des joueurs de 30-42 ans et le 5e centile est au-dessus de celle des joueurs dont l'âge est inférieur à 30 ans. On peut donc supposer que l'efficacité d'un joueur augmente avec l'âge tandis que ces performances physiques sont en baisse.

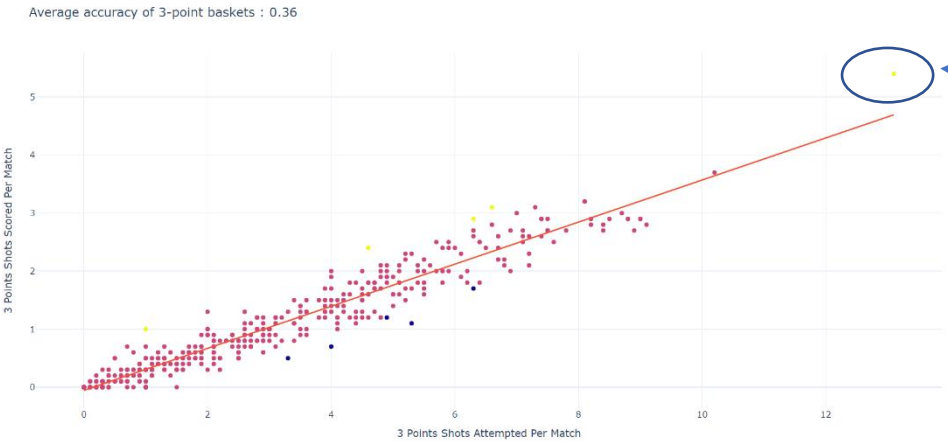
Performances sur les tirs à 2 et 3 points

- Nous allons ici visualiser la performance des joueurs sur les tirs à 2 points, puis à 3 points afin de mettre en évidence les meilleurs tireurs.



On peut voir sur ce graphique le nombre de paniers à 2 points marqués par rapport au nombre de paniers à 2 points tentés. Une régression linéaire a été tracé afin de pouvoir évaluer la performance d'un joueur, pour voir s'il surperforme le modèle, ou au contraire sous-performe. On peut extraire de ce graphique certains joueurs particulièrement efficaces dans cet exercice, comme Nikola Jokic ou Deandre Ayton, qui sont les plus éloignés dans la courbe en termes de sur-performance. A l'inverse, on peut noter des joueurs comme Jalen Suggs ou Saddiq Bay qui sont assez nettement sous la courbe de régression.

Nous avons également observé la précision des tirs à 3 points dans le graphique ci-dessous.



De la même façon que pour les tirs à 2 points, nous pouvons observer la performance des joueurs sur ceux à 3 points, en traçant une régression linéaire pour comparer les performances d'un joueur aux performances moyennes. On peut voir que Stephen Curry se détache fortement des autres joueurs, avec beaucoup plus de tirs tentés et de tirs marqués à 3 points que les autres. En plus d'être capable de se créer beaucoup plus d'opportunités de tirs à 3 points que les autres joueurs de NBA, il est bien plus efficace qu'eux dans cet exercice. En effet, il se situe largement au-dessus de la courbe de régression, de plus ce 15 décembre 2021 Stephen Curry a battu le record de paniers à trois points (2 977) inscrit en saison régulière de NBA.

D'autres joueurs comme Patty Mils ou Ruddy Gay sont également très efficaces dans cet exercice. On peut repérer des joueurs comme Paul Watson ou Michael Porter Jr qui se situent bien en dessous du modèle, ce qui nous permet ici de mettre en évidence leur manque d'efficacité dans les tirs à 3 points.

Comparaisons

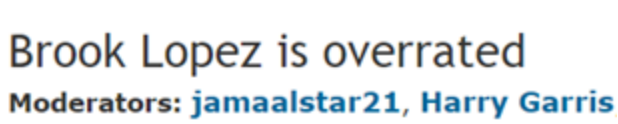
Ce graphique nous permet de comparer les statistiques de deux joueurs afin de déterminer les forces et faiblesses relatives de l'un et l'autre en un coup d'œil. Ce graphique est interactif sur le dashboard : l'utilisateur peut sélectionner les joueurs qu'il souhaite comparer. Par défaut, les joueurs sont Kevin Durant et Stephen Curry.

Ici, on remarque clairement que Stephen Curry se démarque de Kevin Durant sur ne nombre de paniers à 3 points marqués (3PM) et sur le nombre de vols (STL). D'ailleurs, Stephen Curry est récemment devenu le joueur ayant marqué le plus de paniers à trois points de l'histoire de la NBA.

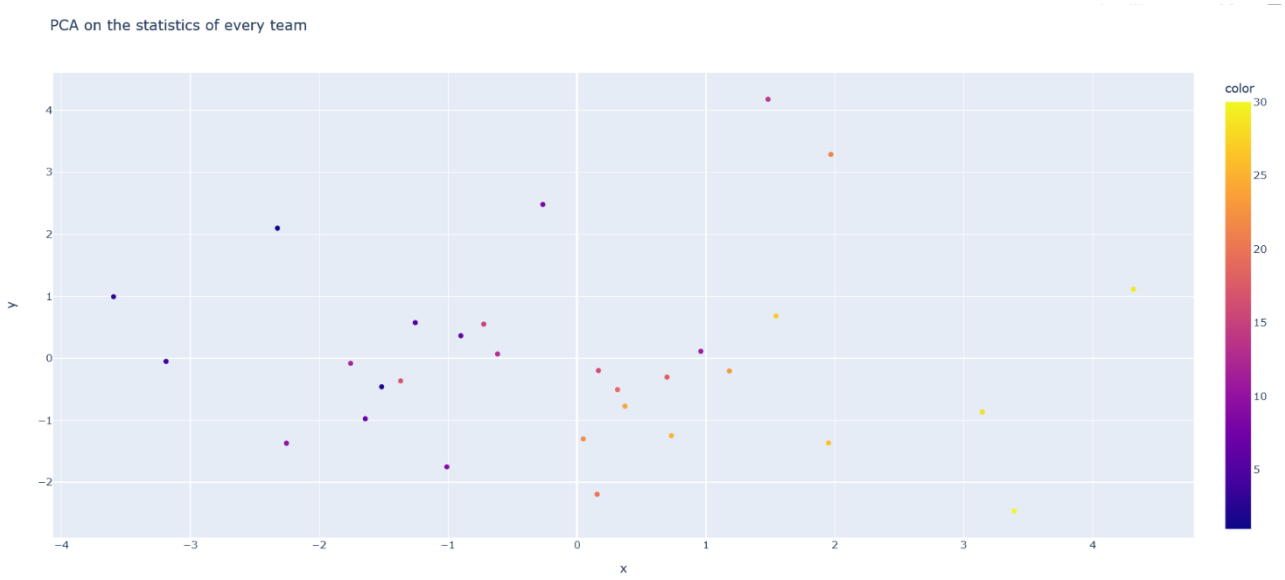
PCA

Nous avons implémenté un PCA (principal components analysis) afin d'afficher les joueurs en fonction de certaines statistiques, et ce en deux dimensions. La couleur de chaque joueur dépend de sa place dans le classement. Le but ici est de déterminer s'il y a des joueurs qui sont plutôt surclassés un point bleu entouré de points jaunes) ou sous classés (un point jaune entouré de points bleus).

Une recherche sur internet nous permet de corroborer les suppositions quant au classement de certains joueurs repérés grâce à ce graphique. Ci-dessous deux titres d'articles publiés sur internet.



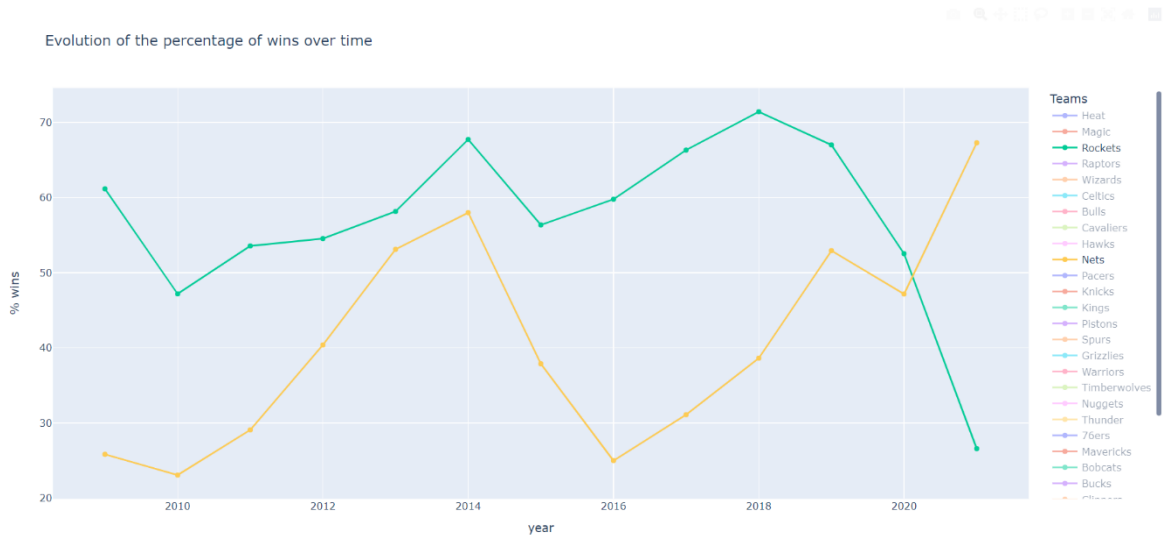
De même que pour les joueurs, nous avons appliqué un algorithme de PCA sur les statistiques des équipes, que l'on a affiché dans un scatterplot. Le but est d'avoir un aperçu visuel de la performance des équipes, et d'ainsi pouvoir observer quelles équipes ont des statistiques semblables, pour en dégager des groupes d'équipes, ou pour voir si certaines équipes sont en sur performance, c'est-à-dire si elles sont bien classées mais ont des statistiques moyennes, ou au contraire en sous performance, avec des statistiques bien au-delà de leurs classements.



Ce graphique permet tout d'abord de mettre en lumière certaines équipes marginales qui se détachent du groupe. On voit également que les équipes les mieux classées (en bleu), ont des statistiques relativement proches. On observe que par exemple l'équipe de San Antonio Spurs, pourtant classée 17ème, possède des statistiques semblables à celle d'Utah Jazz (2ème), ou encore de Chicago Bulls (5ème). On peut ainsi penser que les Spurs ont un classement qui ne reflète pas leur véritable niveau, et on peut espérer que cette équipe remonte la pente et finisse mieux classée en fin de saison avec un peu plus de réussite. A l'inverse, l'équipe de Sacramento Kings est classée 11ème avec des statistiques comparables à celles des Boston Celtics (18ème) ou de Dallas Mavericks (23ème). Dans ce cas, on peut penser qu'ils sur performent et qu'un retour à la normale est prévisible en fin de saison, avec une potentielle chute au classement à prévoir.

Analyse des matchs

Ci-dessous un graphique montrant l'évolution du taux de victoires des équipes en fonction du temps.



Analysons la courbe de pourcentage de victoire des NETS

On observe une croissance de 2010 à 2014, puis une décroissance jusque 2016, puis croissance jusqu'à aujourd'hui. Comme le montre le bilan par saison des Nets (Bilan-Par-Saison-Nets) on peut voir plus

de participation au premier tour entre 2012 et 2014 ainsi qu'entre 2018-2020. Le pourcentage de victoire a donc bien un lien avec l'avancement dans le tournoi NBA.

En cherchant plus profondément dans le transfert de joueur et le Salary cap des Nets_de_Brooklyn, On pourra en déduire quelles décisions ont permis de faire remonter au classement les nets et lesquelles les ont fait descendre.

Saison	Équipe	Conférence		Division		Saison régulière			Playoffs	Entraîneur principal	Réf.
						V	D	PCT			
2009-10	2009-10	Est	15e	Atlantique	5e	12	70	14,6%	-	Lawrence Frank Tom Barrise Kiki Vandeweghe	43
2010-11	2010-11	Est	12e	Atlantique	4e	24	58	29,3%	-	Avery Johnson	44
2011-12	2011-12	Est	12e	Atlantique	5e	22	44	33,3%	-	Avery Johnson	45
Brooklyn Nets											
2012-13	2012-13	Est	4e	Atlantique	2e	49	33	59,8%	**Défaite** au premier tour (Chicago, 3-4)	Avery Johnson P. J. Carlesimo	46
2013-14	2013-14	Est	6e	Atlantique	2e	44	38	53,7%	**Victoire** au premier tour (Toronto, 4-3) **Défaite** en demi-finale de conférence (Miami, 1-4)	Jason Kidd	47
2014-15	2014-15	Est	8e	Atlantique	3e	38	44	46,3%	**Défaite** au premier tour (Atlanta, 2-4)	Lionel Hollins	48
2015-16	2015-16	Est	14e	Atlantique	4e	21	61	25,6%	-	Lionel Hollins Tony Brown	49
2016-17	2016-17	Est	15e	Atlantique	5e	20	62	24,4%	-	Kenny Atkinson	50
2017-18	2017-18	Est	12e	Atlantique	5e	28	54	34,1%	-	Kenny Atkinson	51
2018-19	2018-19	Est	6e	Atlantique	4e	42	40	51,2%	**Défaite** au premier tour (Philadelphie, 1-4)	Kenny Atkinson	52
2019-20	2019-20	Est	7e	Atlantique	4e	35	37	48,6%	**Défaite** au premier tour (Toronto, 0-4)	Kenny Atkinson Jacque Vaughn	53
2020-21	2020-21	Est	2e	Atlantique	2e	48	24	66,7%	**Victoire** au premier tour (Boston, 4-1) **Défaite** en demi-finale de conférence (Milwaukee, 3-4)	Steve Nash	54

On observe une corrélation entre les équipes Rockets et Raptors, ces équipes doivent avoir un niveau assez similaire.

Prédiction

Nous avons réalisé plusieurs modèles de prédiction tout au long de notre projet.

Tout d'abord, nous avons implémenté des modèles de régression linéaire sur nos différents graphiques, afin de visualiser plus facilement certains aspects de ces graphiques. Par exemple, sur les performances des joueurs sur les tirs à 2 et 3 points, nous avons décidé d'ajouter une régression linéaire pour facilement se rendre compte des joueurs au-dessus du lot, en calculant la distance entre le joueur et la courbe de régression.

Ensuite, nous avons voulu prédire le poste qu'un joueur occupe sur le terrain, en fonction de ses statistiques. Pour ce faire, nous avons utilisé les données que nous avions récolté au préalable sur les joueurs, et nous les avons traitées pour les fournir à différents modèles de classification. Le défi ici est que nous ne disposons pas d'énormément de données sur les joueurs, car le nombre de joueurs évoluant en NBA est assez limité, avec environ 500 joueurs, et seulement une cinquantaine évoluant au poste de "Center". Au final, nous sommes parvenus à concevoir un modèle capable d'obtenir une précision de 95% sur ses prédictions, ce qui est un score élevé, et dont nous nous satisfaisons. Nous aurions probablement pu améliorer ce score en optimisant encore plus les hyperparamètres de notre modèle, et en disposant de plus de données.

Le détail du code pour réaliser ce modèle se trouve sur le notebook "PlayersRolePrediction.ipynb" disponible dans le dossier soumis.

Points faibles et améliorations possibles

Tout d'abord, nous aurions pu effectuer des tests statistiques (par exemple des tests ANOVA ou chi-deux) sur les données récoltées afin de déterminer si les différentes observations que nous avons faites tout au long de notre étude sont véritablement fondées ou si celles-ci peuvent relever du hasard.

De plus, n'avons pas pu réaliser toutes les observations et prédictions que nous aurions désirées, certaines informations n'étaient tout simplement pas disponibles sur le site officiel de la NBA. Par exemple, la prédiction du pourcentage de chance qu'un tir soit marqué en fonction de la position du joueur sur le terrain, et la réalisation d'une heatmap indiquant les zones préférentielles de certains joueurs au cours d'un match étaient deux aspects que nous aurions voulu développer, mais nous n'avions pas accès à suffisamment de données pertinentes sur ces domaines pour pouvoir réaliser ces études. Ainsi, en récupérant des informations sur d'autres sites que le site de la NBA, nous aurions eu la possibilité de pousser certains aspects plus loin que ce que nous avons été en mesure de faire.

Difficultés rencontrées

La réalisation de ce projet s'est vu ralentir par plusieurs difficultés techniques.

Le premier challenge fut le scraping des données : le lien https://www.nba.com renvoie vers la section nba du site beinsport, dont le scraping s'est avéré fastidieux. Nous avons fini par trouver le site officiel - plus facile à scraper - avec les statistiques, qui se trouve à l'adresse https://www.nba.com/stats.

Le deuxième obstacle que nous avons rencontré fut la compréhension des données spécifiques au basketball. Effectivement, les noms colonnes de nos datasets correspondent à des abréviations de statistiques spécifiques au basketball tel que 3PA%, OREB, FT%, etc. Un glossaire nous a aidé dans la compréhension de ces statistiques.

Une meilleure compréhension de ces données fut vitale dans le choix des graphiques à réaliser. Effectivement, le dataset étant très large (89 colonnes pour le dataset des statistiques des joueurs), il est nécessaire de faire le tri dans les informations récoltées pour en extraire le plus important. Il nous a ensuite fallu faire preuve de créativité pour choisir et créer des graphiques qui apportent de l'information utile à l'utilisateur.

Créer ces graphiques fut aussi une difficulté dans le sens où nous avons dû passer par une longue phase d'apprentissage et de familiarisation avec la librairie Plotly. Nous avons fait le choix d'utiliser Plotly car cette librairie permet de créer des graphiques interactifs et facilement intégrables sur un dashboard avec Dash.

Le déploiement du dashboard sur internet fut le dernier obstacle à la réalisation du projet. N'ayant jamais fait ça auparavant, il nous a fallu trouver un hébergeur et apprendre à l'utiliser.

Conclusion

Pour conclure, ce projet fut pour nous l'opportunité d'approfondir notre maitrise des outils de scraping, de traitement et de visualisation des données, ainsi que du développement de modèles de prédiction. Outre l'aspect technique, nous avons également renforcé nos compétences organisationnelles et de travail en équipe. De plus, nous avons découvert de nombreux aspects du basket et de la NBA que nous ignorions, et que nous avons trouvé très intéressants.

Nous avons cherché à faire des graphiques les plus pertinents possibles, en essayant de mieux comprendre certaines particularités de la NBA, que ce soit sur les joueurs ou sur les équipes. Ainsi, nous espérons que les graphiques proposés font honneur à notre travail et apportent des informations intéressantes à celui qui les visionne, sans qu'il n'ait besoin de connaissance préalables sur le basketball.