# PROJET DE PROGRAMMATION R

## MACHINE LEARNING EN NBA

ZACHARIE DESJARDIN
M1 SIAD CLASSIQUE GROUPE 2

# Table des matières

Ce projet nous permettra de mettre en œuvre les méthodes statistiques et de machine learning du langage de programmation R. Il aura pour thème, la NBA.

# I. PRÉPARATION DU JEU DE DONNÉES

On trouve les données sur le site suivant :

https://www.kaggle.com/jacobbaruch/basketball-players-stats-per-season-49-leagues?select=players_stats_by_season_full_details.csv

https://www.basketball-reference.com/awards/all_league.html

## 1. Dataset 1 :

Ce dataset est composé des statistiques individuelles des joueurs de basketball à travers le monde par saison. Il contient les données de 49 ligues entre 1999 et 2020 et comporte 53 949 observations pour plus de 11 000 joueurs. Les 34 variables décrivent à la fois les caractéristiques du joueur comme son nom ou sa taille et les statistiques de la saison comme le nombre de points marqués, le nombre de lancers-francs ou encore le nombre de minutes joués.

Ci-dessous le détail des variables :

Caractéristiques du joueur :

*League Name* : nom de la ligue

*Season* : saison

*Stage* : indique si les données concernent les matchs internationaux, la saison régulière ou les playoffs

*Player* : nom du joueur

*Team* : nom de l'équipe

*Birth_year* : année de naissance

*Birth_month* : mois de naissance

*Birth_date* : date de naissance

*Height* : taille en ft

*Height_cm* : taille en cm

*Weight* : poids en pounds

*Weight*_kg : poids en kg

*Nationality* : nationalité

*High_school :* université du joueur

*Draft_round* : tour de draft

*Draft_pick* : position de draft

*Draft_team* : équipe de draft

Statistiques de la saison :

*GP* : nombre de match joués *(Game Played)*

*MIN* : nombre de minutes jouées (*Minutes*)

*FGM* : nombre de tirs réussis *(Field Goals Made)*

*FGA* : nombre de tirs tentés *(Field Goals Attempts)*

*3PM* : nombre de tirs à 3 points réussis *(3 Points Made)*

*3PA* : nombre de tirs à 3 points tentés *(3 Points Attempts)*

*FTM* : nombre de lancers-francs réussis *(Free Throws Made)*

*FTA* : nombre de lancers-francs tentés *(Free Throws Attempts)*

*TOV* : nombre de ballons perdus *(Turnovers)*

*PF* : nombre de fautes *(Personal Fouls)*

*ORB* : nombre de rebonds offensifs (*Offensive Rebounds)*

*DRB* : nombre de rebonds défensifs (*Defensive Rebounds)*

*REB* : nombre total de rebonds *(Rebounds)*

*AST* : nombre de passes décisives (*Assists)*

*STL* : nombre d'interceptions *(Steals)*

*BLK* : nombre de contres *(Blocks)*

*PTS* : nombre de points marqués (*Points*)

## 2. Dataset 2 : All-NBA Team

https://www.basketball-reference.com/awards/all_league.html

Ce dataset regroupe les joueurs qui font partie de l'équipe de l'année de chaque saison. Chaque année, à l'issue de la saison, 15 joueurs sont répartis en 3 équipes de 5 pour les récompensé de leur saison. Les 5 joueurs de la 1st team sont les 5 meilleurs, ceux de la 2nd team les 5 suivants et ceux de la 3rd les 5 suivants. Dans le cadre de notre étude, nous ne ferons pas de distinctions entre la 1st, la 2nd et la 3rd team, on souhaite juste connaitre les 15 meilleurs joueurs de la saison.

Le dataset comporte 8 variables et 198 observations.

*Season* : saison du joueur

*Lg* : ligue, ici la NBA

*Tm* : Team, 1st, 2nd ou 3rd.

*X* : nom du 1er joueur

*X.1* : nom du 2ème joueur

*X.2* : nom du 3ème joueur

*X.3* : nom du 4ème joueur

*X.4* : nom du 5ème joueur

On conserve uniquement la variable *Season* et le nom du joueur que l'on regroupe en 1 colonne avec la fonction *Pivot_longer* du package *Tidyverse*. De plus on supprime les données antérieures à la saison 2010 - 2011. On a maintenant 150 observations et 2 variables : la saison et le nom du joueur.

## 3. Fusion des 2 datasets

Afin de fusionner les 2 datasets on créer un id composé du nom du joueur et de la saison pour chaque dataset. Cet identifiant est unique (aucun joueur n'a le même nom). On fait ensuite un *merge* entre les 2 datasets.

Au final, on a les statistiques du joueur, ainsi qu'une variable qui indique s'il a fait partie de la *All-NBA Team* (Yes / No).

## 4. Nettoyage du jeu de données

Pour éviter les problèmes liés à la casse, on utilise la fonction *str_to_title* du package *stringr* qui transforme la 1ère lettre de chaque mot en majuscule et le reste en minuscule. Par exemple « LeBron James » devient « Lebron James ».

On divise ces statistiques par le nombre de match joués afin d'obtenir la moyenne par match.

De plus, on transforme les statistiques réussis/tentés en un pourcentage.

Par exemple, voici les 4 premières statistiques de Damian Lillard en 2019-2020 :

| Season | Player | Team | GP | MIN | FGM | FGA |
|---|---|---|---|---|---|---|
| 2019-2020 | Damian Lillard | POR | 66 | 2473.7 | 624 | 1349 |

Désormais, ces statistiques se présenteront ainsi[1] :

| Season | Player | Team | GP | MIN | FGM | %FGA |
|---|---|---|---|---|---|---|

[1] 2473.7/66 = 37.5  624/66 = 9.5  624/1349 = 46.3 %

| | | | | | | |
|---|---|---|---|---|---|---|
| 2019-2020 | Damian Lillard | POR | 66 | 37.5 | 9.5 | 46.3 |

On supprime les joueurs dont le nombre de minutes moyen jouées par match est inférieur à 28 et dont le nombre de match joué est inférieur à 40. Cela évite d'avoir des joueurs qui ont peu joués et qui sont donc peu intéressant.

Ainsi, on conserve les variables suivantes :

### Caractéristiques du joueur :

*Season* : saison

*Player* : nom du joueur

*Team* : nom de l'équipe

*Height* : taille en cm

*Weight* : poids en kg

*All_NBA_Team* : sélection dans la All-NBA Team

### Statistiques de la saison :

*GP* : nombre de match joués *(Game Played)*

*MIN* : nombre de minutes jouées par match

*FGM* : nombre de tirs réussis par match *(Field Goals Made)*

*PFG* : pourcentage de tirs réussis *(Field Goals Attempts)*

*X3P* : nombre de tirs à 3 points réussis par match *(3 Points Made)*

*PX3P* : pourcentage de tirs à 3 points réussis *(3 Points Attempts)*

*FTM* : nombre de lancers-francs réussis par match *(Free Throws Made)*

*PFT* : pourcentage de lancers-francs réussis *(Free Throws Attempts)*

*TOV* : nombre de ballons perdus par match *(Turnovers)*

*PF* : nombre de fautes par match *(Personal Fouls)*

*ORB* : nombre de rebonds offensifs par match (*Offensive Rebounds)*

*DRB* : nombre de rebonds défensifs par match (*Defensive Rebounds)*

*REB* : nombre total de rebonds par match *(Rebounds)*

*AST* : nombre de passes décisives par match (*Assists)*

*STL* : nombre d'interceptions par match *(Steals)*

*BLK* : nombre de contres par match *(Blocks)*

*PTS* : nombre de points marqués par match

## 5. Données manquantes

On vérifie qu'il n'y a pas de données manquantes, cela viendrait fausser nos résultats.

On commence vérifier qu'il n'y a pas de statistiques manquantes dans le jeu de données. On utilise la fonction *vis_miss* du package *visdat*.


On ne peut pas vérifier que tous les joueurs sont bien présents, mais on peut au moins s'assurer que les joueurs de la All-NBA Team sont présents.

En faisant la somme du nombre de joueurs All-NBA Team on obtient 149. Puisqu'il y'a 15 joueurs par saisons on devrait avoir 150 joueurs.

On utilise la fonction *count* pour savoir dans quelle saison il manque un joueur : 2011 - 2012.

On s'aperçoit qu'il manque Dwight Howard lors de cette saison. On regarde donc les statistiques de ce joueur sur les 10 saisons :

| id | Season | Player | Team | GP | MIN | FG | PFG | X3P | PX3P | FT | PFT | TOV | PF | ORB | DRB | REB | AST | STL | BLK | PTS | Height | Weight | All_NBA_Team |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 55 | 2010 - 2011 | Dwight Howard | ORL | 78 | 37.63 | 7.94 | 59.29 | 0.00 | 0.00 | 7.00 | 59.61 | 3.58 | 3.31 | 3.96 | 10.12 | 14.08 | 1.37 | 1.37 | 2.38 | 22.87 | 211 | 120 | Yes |
| 479 | 2012 - 2013 | Dwight Howard | LAL | 76 | 35.81 | 6.18 | 57.81 | 0.01 | 16.67 | 4.67 | 49.24 | 2.96 | 3.83 | 3.30 | 9.13 | 12.43 | 1.42 | 1.11 | 2.45 | 17.05 | 211 | 120 | Yes |
| 703 | 2013 - 2014 | Dwight Howard | HOU | 71 | 33.75 | 6.66 | 59.13 | 0.03 | 28.57 | 4.92 | 54.70 | 3.23 | 3.38 | 3.25 | 8.94 | 12.20 | 1.85 | 0.85 | 1.80 | 18.27 | 211 | 120 | Yes |
| 1114 | 2015 - 2016 | Dwight Howard | HOU | 71 | 32.11 | 5.24 | 62.00 | 0.00 | 0.00 | 3.27 | 48.95 | 2.34 | 3.08 | 3.35 | 8.41 | 11.76 | 1.38 | 0.97 | 1.59 | 13.75 | 211 | 120 | No |
| 1336 | 2016 - 2017 | Dwight Howard | ATL | 74 | 29.71 | 5.24 | 63.30 | 0.00 | 0.00 | 3.05 | 53.30 | 2.30 | 2.74 | 4.00 | 8.70 | 12.70 | 1.41 | 0.86 | 1.24 | 13.54 | 211 | 120 | No |
| 1621 | 2017 - 2018 | Dwight Howard | CHA | 81 | 30.41 | 6.25 | 55.54 | 0.01 | 14.29 | 4.12 | 57.39 | 2.57 | 3.07 | 3.15 | 9.35 | 12.49 | 1.30 | 0.59 | 1.62 | 16.63 | 211 | 120 | No |

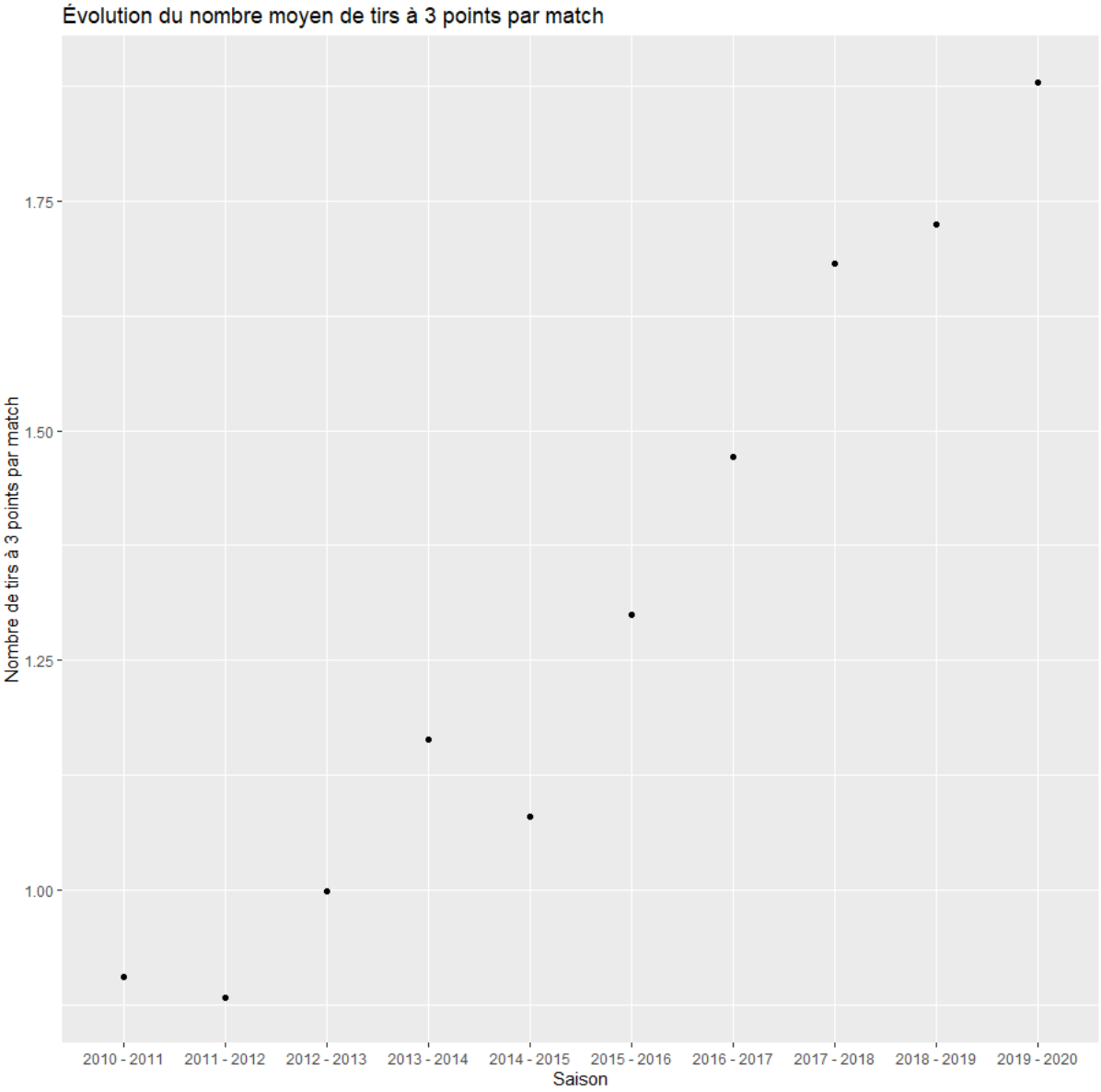
En effet, il manque plusieurs saisons dont la saison 2011 - 2012. On va donc copier les statistiques de la saison 2010 - 2011 pour la saison 2011 - 2012. Puisque Dwight Howard a été All-NBA Team en 2010 - 2011, on suppose qu'avec les mêmes statistiques, il serait All-NBA Team en 2011 - 2012.

Désormais, on a bien 150 joueurs All-NBA Team.

# II. Analyse de données et datavisualisation

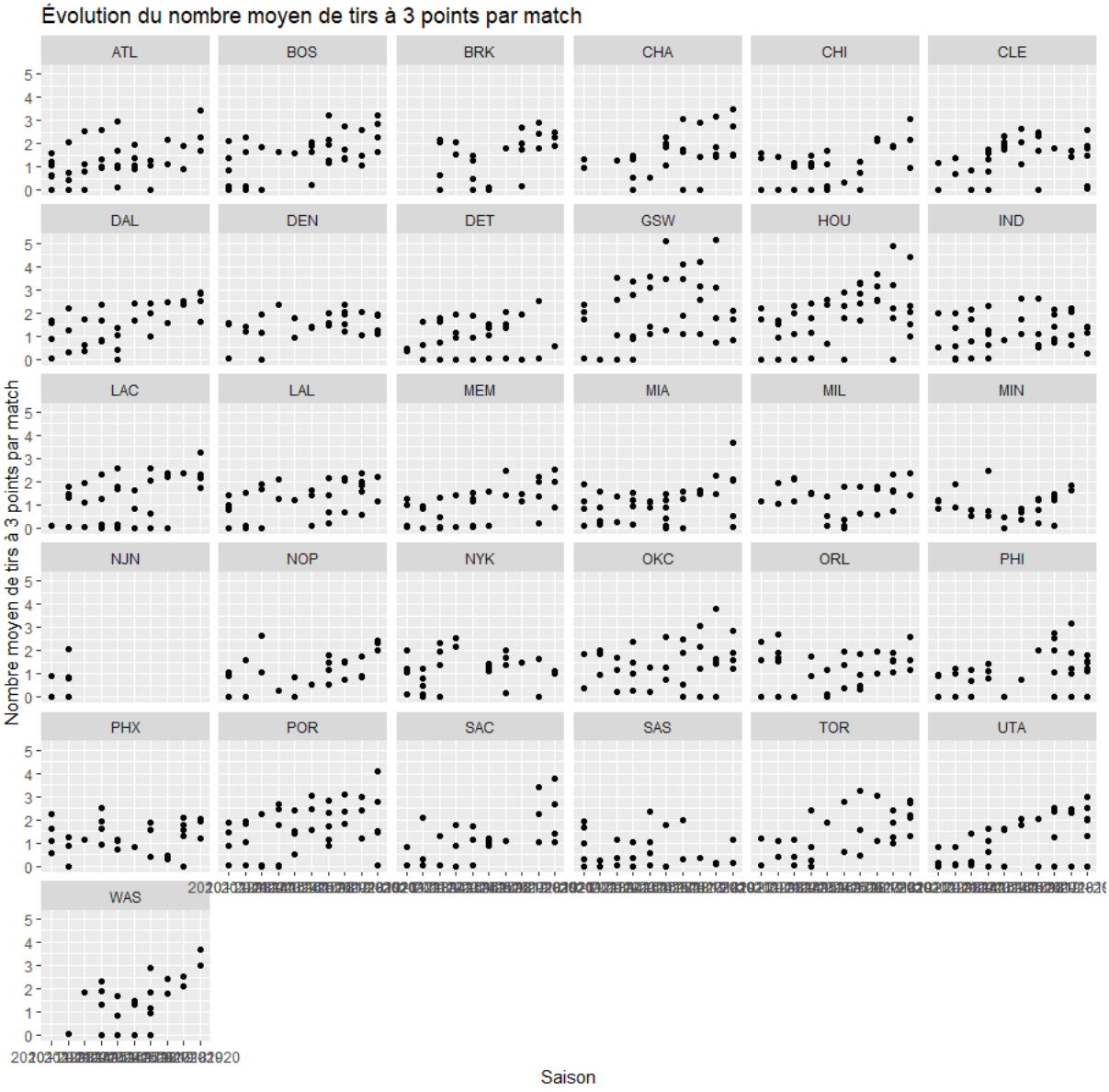
Dans cette partie, on va explorer notre jeu de données, regarder quels sont les records, l'évolution du jeu, notamment sur les tirs à 3 points, les différences entre équipe et enfin s'intéresser à la carrière de LeBron James.

Commençons par regarder comment à évoluer le jeu à 3 points depuis 2010 :




On peut voir qu'entre 2010 et 2020 le nombre moyen de tirs à 3 points réussi par joueur a presque doublé.

On s'intéresse à cette évolution au niveau des équipes :




Certaines équipes comme Golden State (*GSW*) ou Portland (*POR*) ont vu leur nombre de tirs à 3 points augmenté. Par contre certaines équipes comme Minnesota (*MIN*) ou San Antonio (*SAS*) ont vu ce nombre stagné.

Globalement, les équipes ont augmenté le nombre de tirs à 3 points par match depuis 2010.

Cette évolution se voit aussi pour Lebron James, la saison où il a effectué le moins de tirs à 3 points par match est 2011 - 2012, alors que celle où il en a fait le plus est 2019 - 2020.[2]

[2] Sur ces 10 dernières années

Intéressons-nous plus en détail à la carrière de LeBron James :

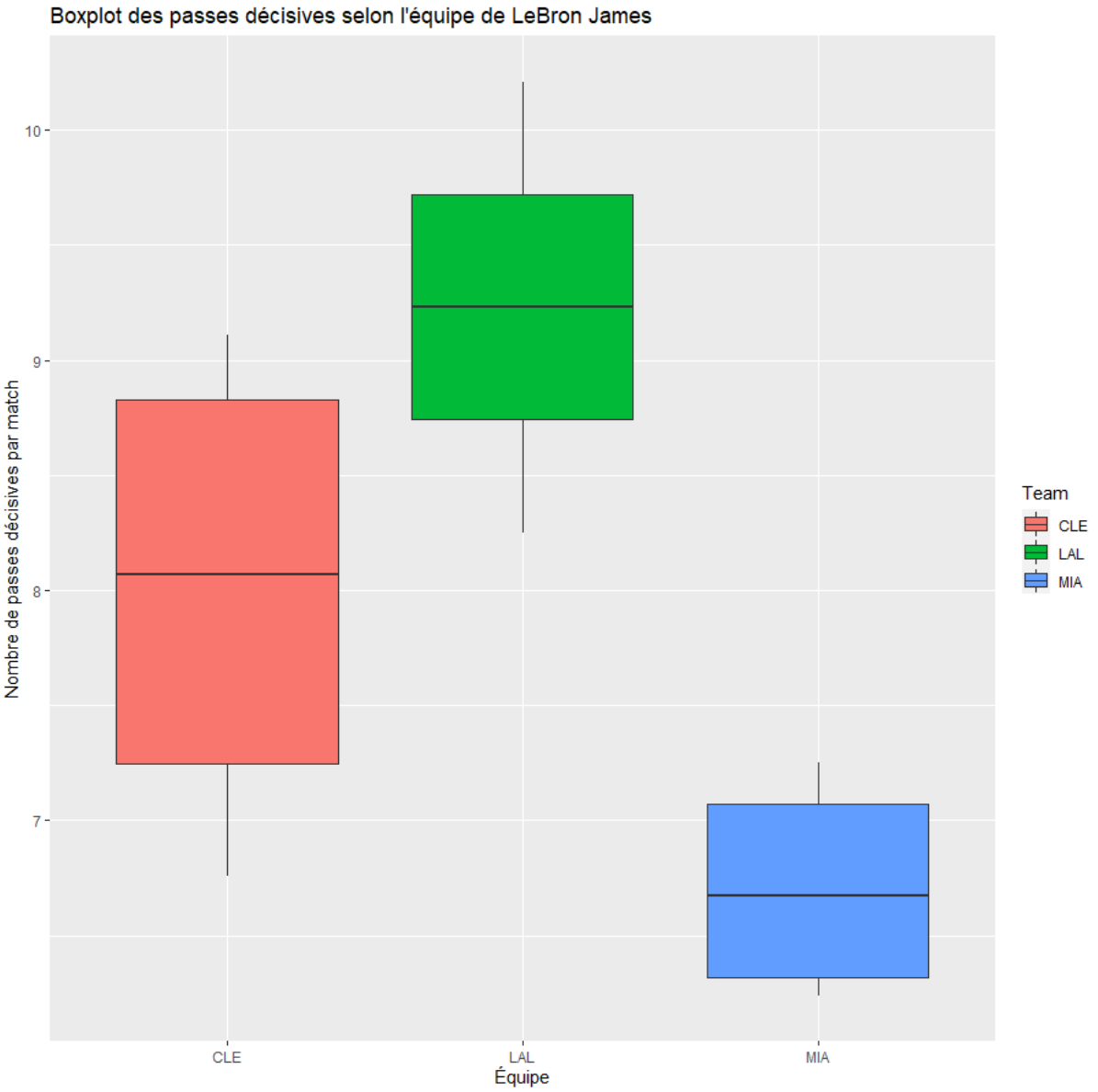
On utilise *apply* pour calculer la moyenne de ses statistiques sur ses 10 dernières saisons, ainsi que celles de la moyenne de l'ensemble des joueurs.

| | GP | MIN | FG | PFG | X3P | PX3P | FT | PFT | TOV | PF | ORB | DRB | REB | AST | STL | BLK | PTS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LeBron James | 71,7 | 36,81 | 9,853 | 52,8 | 1,551 | 35,6 | 5,232 | 72,4 | 3,653 | 1,75 | 1,16 | 6,58 | 7,74 | 7,73 | 1,48 | 0,66 | 26,5 |
| Joueur moyen | 72,1585624 | 32,55 | 5,999 | 46,4 | 1,321 | 31,7 | 3,0085 | 78,2 | 2,0764 | 2,33 | 1,31 | 4,55 | 5,86 | 3,68 | 1,09 | 0,62 | 16,3 |

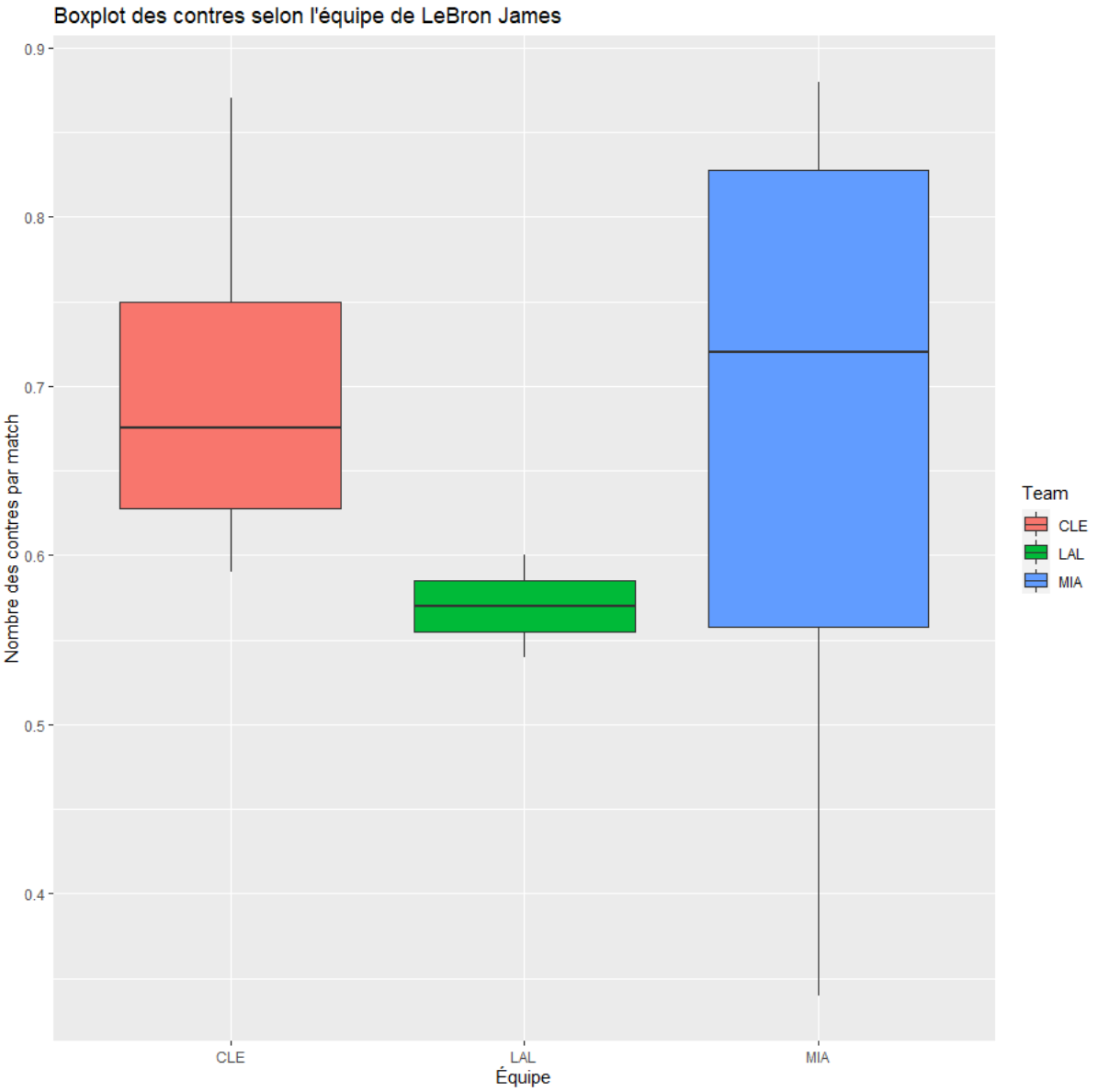
Il n'y a pas débat, LeBron James domine dans presque toutes les catégories sur la dernière décennie.

Seul son nombre de rebonds offensifs, de fautes et son pourcentage aux lancers-francs sont moins bon que la moyenne.

Il serait également intéressant de voir s'il existe des différences dans son jeu selon l'équipe dans laquelle il était.




Ici on voit que LeBron James fait plus de passes décisives en moyenne maintenant qu'il joue au Los Angeles Lakers que quand il jouait à Cleveland ou à Miami.




Cependant, lorsqu'il jouait à Miami, LeBron James faisait plus de contres par match (autour de 0.7 par match) qu'actuellement à Los Angeles (autour de 0.6).

On peut donc dire que le style de jeu de LeBron James dépend de l'équipe dans laquelle il joue.

Enfin, intéressons-nous à un record, existe-t-il sur cette décennie un joueur qui est en triple-double en moyenne sur toute une saison.[3]

Russell Westbrook est en effet le seul joueur à avoir fait cette performance, de plus il l'a faites 3 saisons d'affilé. Il a bien mérité son surnom « Mr Triple-double ».

[3] Être en triple-double signifie marquer 10 points, faire 10 passes décisives et faire 10 rebonds. Cette performance est rare en match, et quasiment impossible à faire sur une saison entière (en moyenne).

| | PTS | AST | REB |
|---|---|---|---|
| 2016 - 2017 | 31.58 | 10.37 | 10.67 |
| 2017 - 2018 | 25.35 | 10.25 | 10.05 |
| 2018 - 2019 | 22.95 | 10.74 | 11.05 |

## III. Apprentissage non-supervisé : clustering

Les joueurs de la NBA ont tous des styles de jeu différent, par exemple il y'a les *Two-way player* (joueurs bons en attaque et en défense), les *non-scorer* (joueurs bons partout mais qui marque peu de points), les *defensive tower* (joueurs très physiques et très défensifs), etc. Il serait intéressant de faire une typologie des différents joueurs.

Pour cela on effectuera un clustering en utilisant l'algorithme de K-means. Le but est de créer des classes qui contiennent les joueurs les plus semblables possible, tout en maximisant les différences entre les classes.

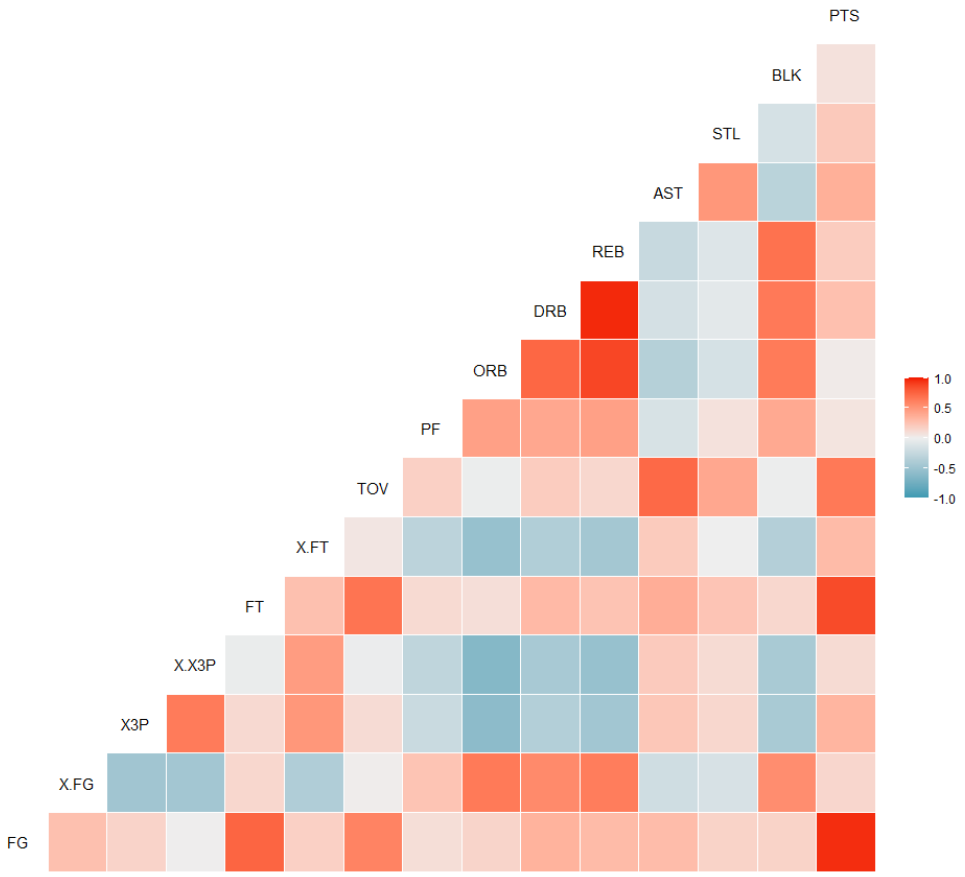
Le clustering se fera sur l'ensemble des joueurs depuis 2010, un joueur peut donc être dans un cluster une année, puis dans un autre l'année suivante (cela signifie que son style de jeu a évolué).

On commence par sélectionner les variables.

Tout d'abord, on retire les variables qui ne sont pas intéressantes dans le cadre de la classification, c'est le cas des variables liées aux caractéristiques du joueur (*Season, Player, Team, Height, Weight & All_NBA_Team*). C'est aussi le cas pour les variables suivantes *: GP, MIN, TOV* et *PF*. En effet ces variables ne sont pas intéressantes pour classer les joueurs, on ne souhaite pas que le nombre de minutes jouées ou le nombre de fautes influent sur la construction des groupes.

On regarde ensuite les corrélations entres variables :

Matrice de corrélation des variables du jeu de données NBA

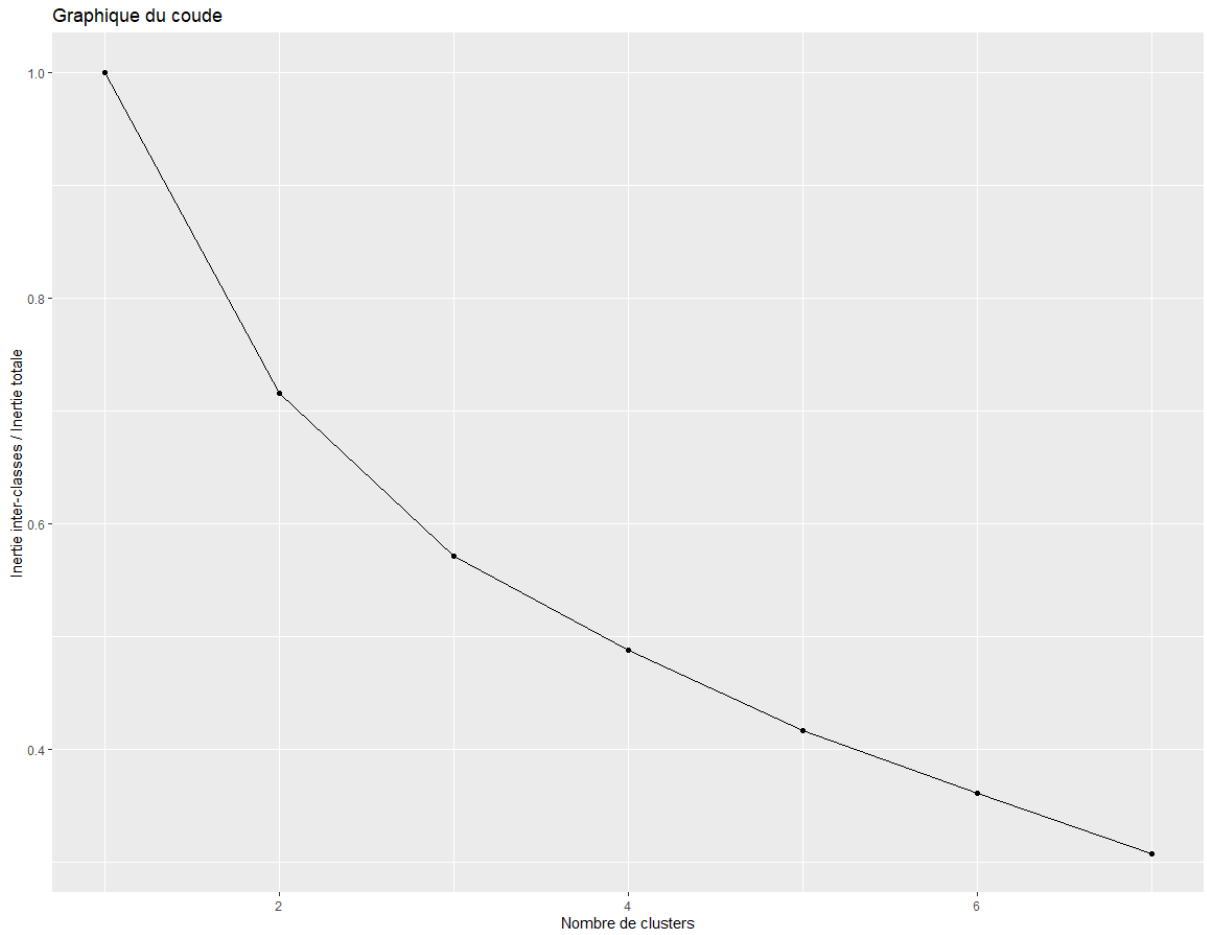



*PTS* est très corrélé à *FT* et *FG*. De plus REB est très corrélé à DRB et ORB (en effet, REB = ORB + DRB).

L'indice de corrélation entre *PTS* et *FG* est de 0.96, autrement dit ces 2 variables sont très corrélées (ce qui est logique, le nombre de points et le nombre de tirs réussis sont liés).

On va ensuite centrer et réduire nos variables afin d'éviter que 2 variables n'ayant pas la même unité ai un poids différent lors du clustering.

Pour choisir le nombre de classes adéquat on regarde le rapport de l'inertie interclasses sur l'inertie totale pour chaque cluster, lorsque la chute d'un cluster à l'autre devient trop peu significative, on considère que la perte d'information est trop importante et l'ajout d'un cluster n'est pas assez intéressant.




Le nombre de classes idéale est de 2 ou 3. On commence donc par un clustering à 3 classes. Toutefois dans cette classification, un nombre élevé de classes serait plus intéressant, on fera donc un 2ème clustering avec cette fois-ci 10 classes.

## 1. Clustering à 3 classes

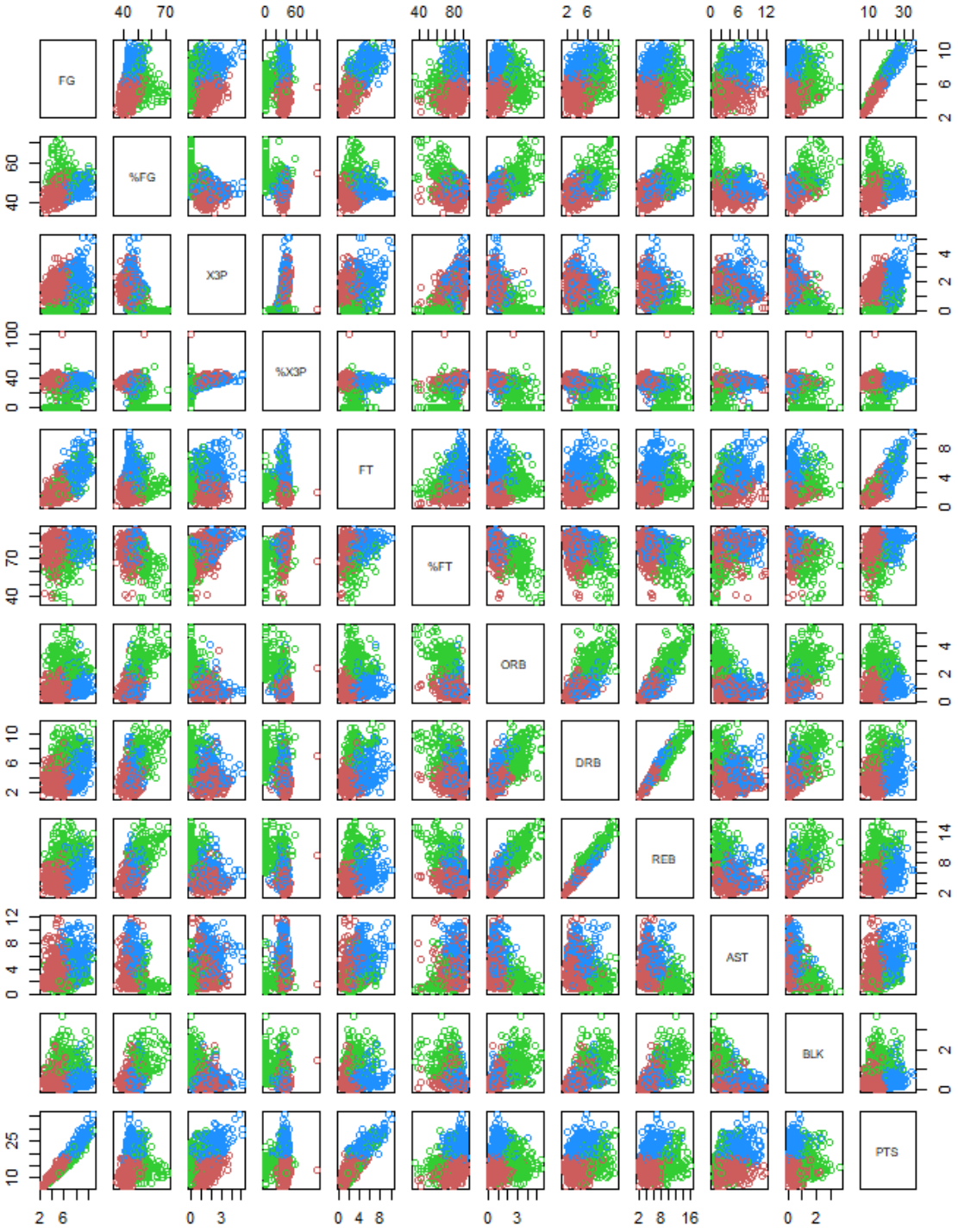
On passe maintenant au paramétrage du k-means : on créer 3 clusters avec 20 itérations et un maximum de 50 itérations.

Cela signifie que R répète 20 fois l'algorithme avec des initialisations différentes à chaque fois (l'initialisation est aléatoire) et choisis celui où l'inertie intra-classe est la plus faible (on cherche à avoir les classes les plus différentes possibles).

On obtient 3 clusters contenant respectivement 495, 209 et 242 joueurs.

Le ratio entre l'inertie expliquée par le clustering et l'inertie totale est de 41.6 %, ce qui est satisfaisant.

On peut maintenant visualiser la relation entre chaque cluster et les variables utilisées pour construire ce cluster :




| | |
|---|---|
| | Cluster 1 |
| | Cluster 2 |
| | Cluster 3 |

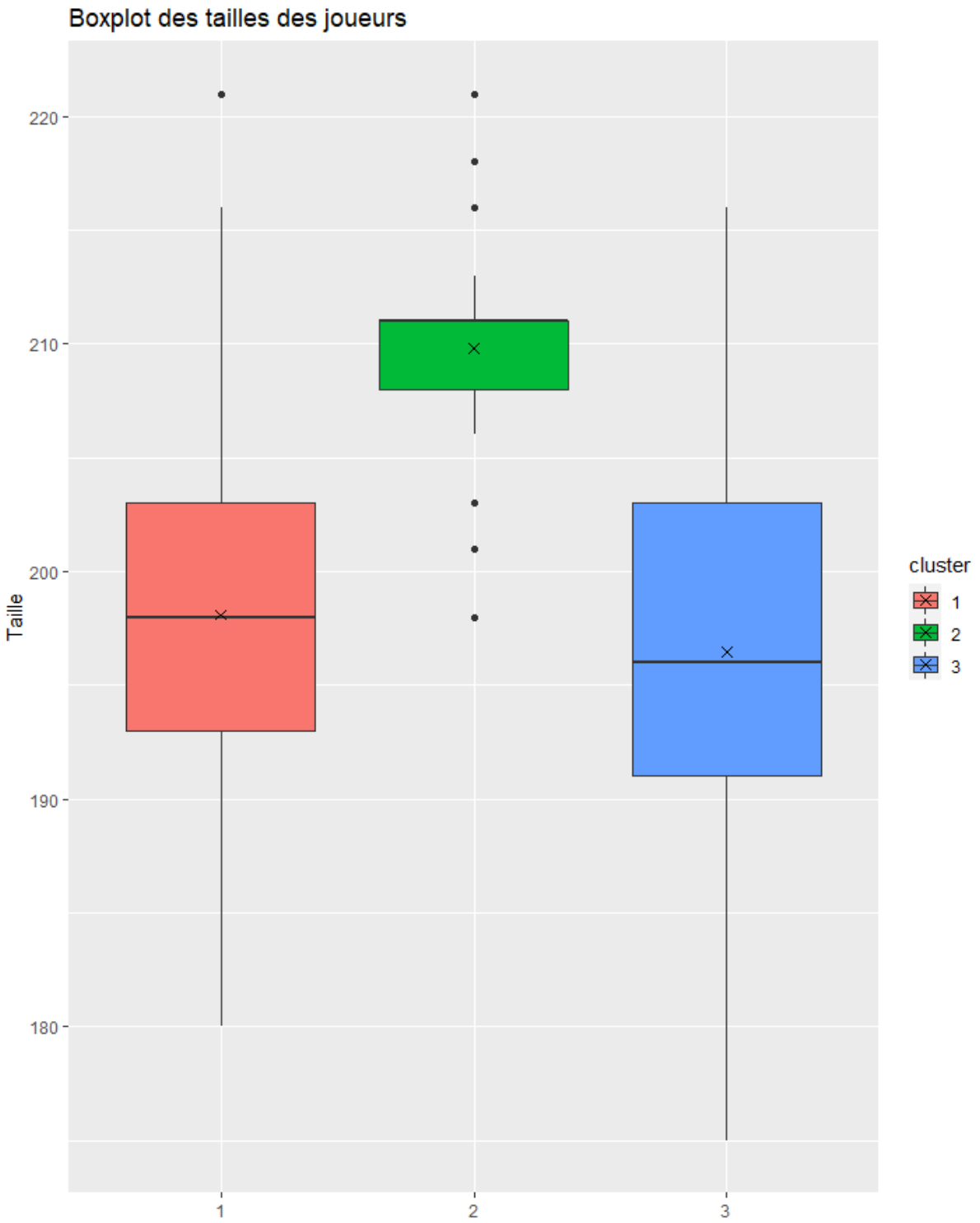
Ce moyen de visualisation nous offre une première idée de la répartition des clusters. Mais en raison du nombre élevé de variables, d'autre visualisations semblent nécessaires.

Commençons par nous intéressés aux 2 variables très importantes que sont les points et les rebonds :

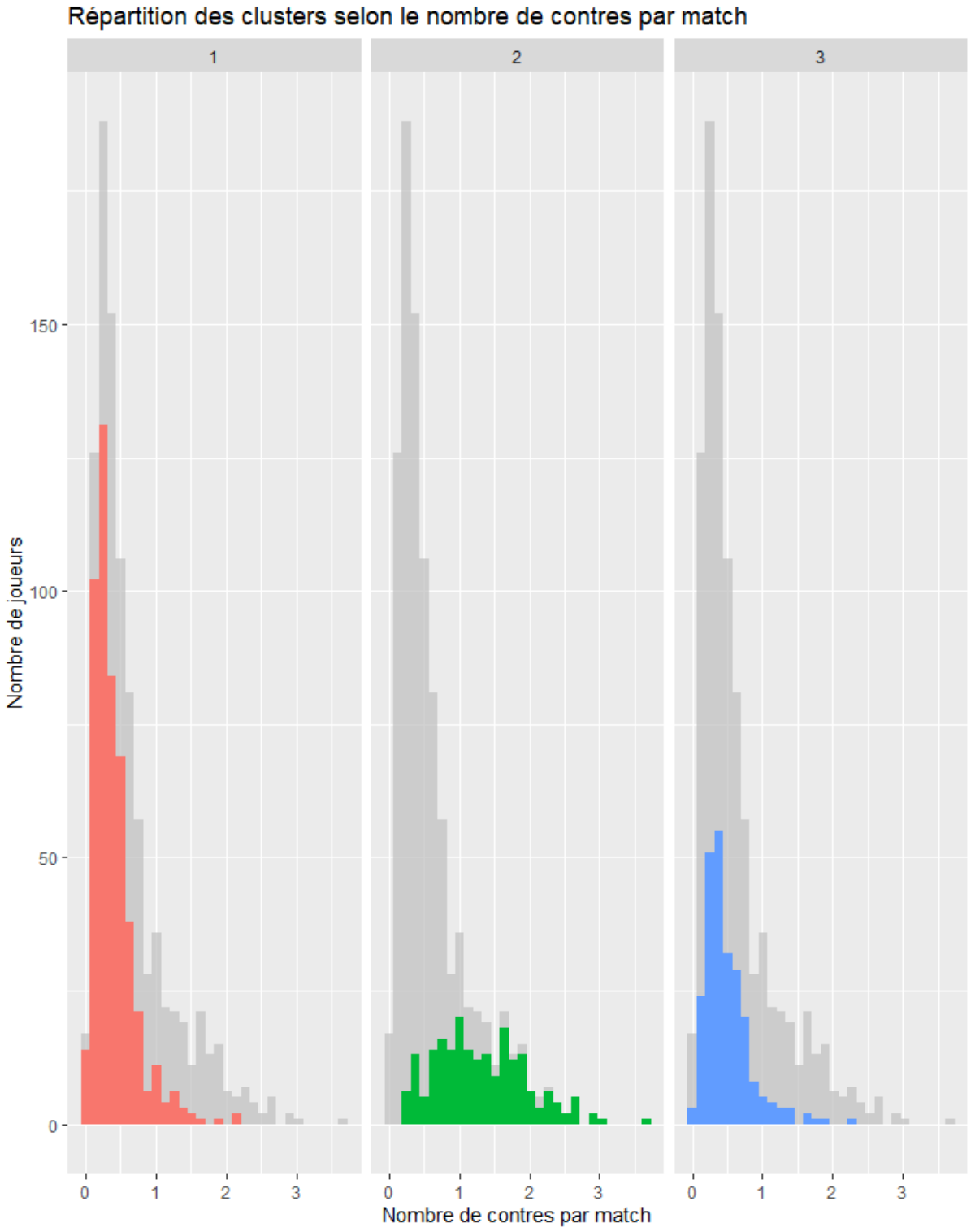

En observant le graphique, on voit tout de suite que le 1er cluster contient des joueurs qui marquent peu de points et font peu de rebonds. Le cluster 2 est plus composé de joueurs faisant beaucoup de rebonds et le cluster 3 de joueurs marquant beaucoup de points.

Le cluster 2 semble donc regrouper les joueurs plutôt défensifs et le cluster 3 les joueurs plutôt offensifs. Le cluster 1 regroupe le reste des joueurs, qui sont plutôt moyen.

Pour vérifier ces suppositions on peut regarder le nombre de contres par match et la taille qui seront caractéristiques des joueurs défensifs. On peut également regarder le nombre de passes décisives qui est plus caractéristiques des joueurs offensifs.

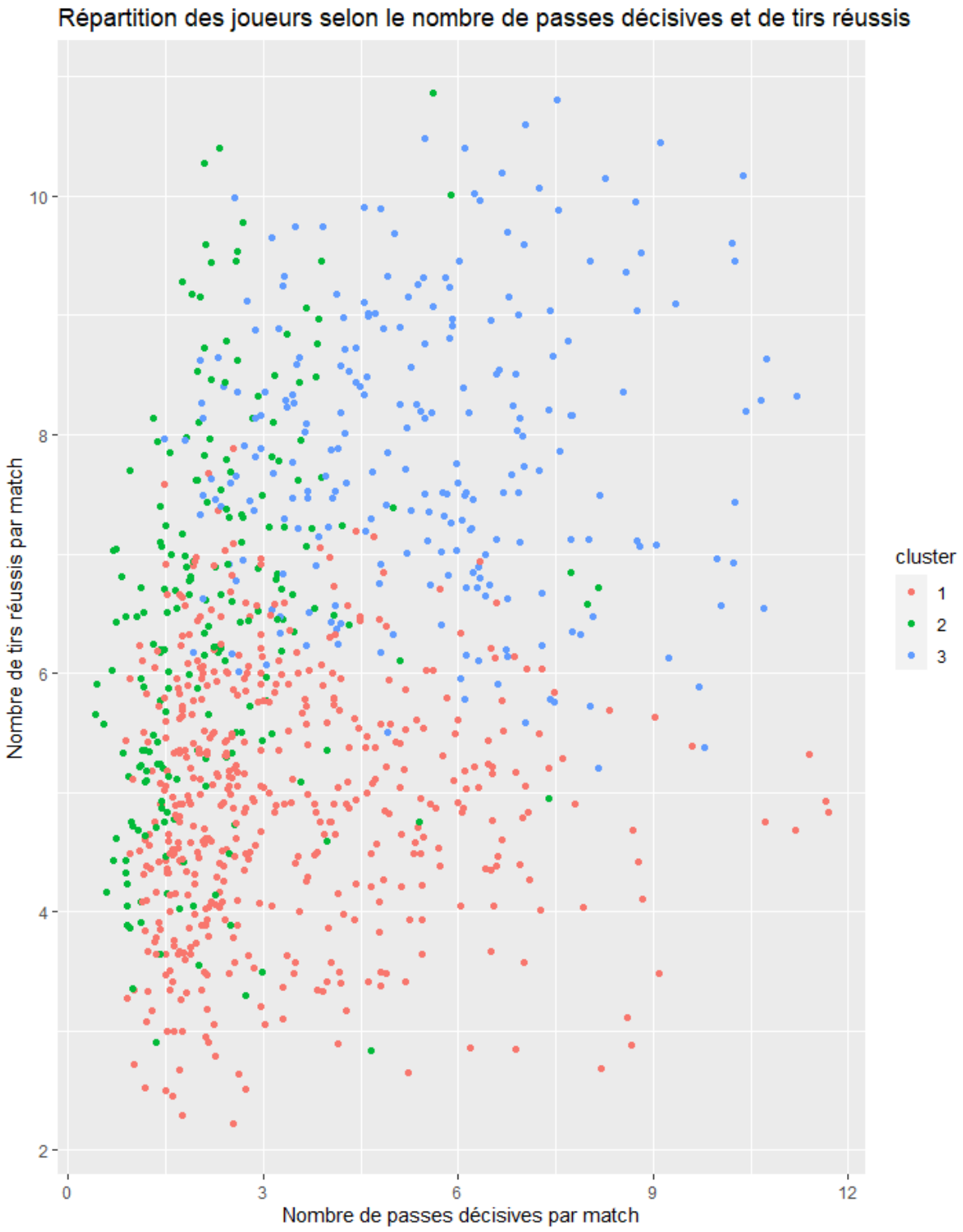


On voit ici que les joueurs du cluster 2 sont bien plus grands que les autres[4], malgré que ce cluster soit celui qui contient le plus d'outliers. La moyenne des joueurs du cluster 2 est de 2m10 contre 1m97 et 1m96 pour les clusters 1 et 3.




[4] Attention, la taille n'a pas été utilisé pour faire le clustering, on l'utilise juste pour caractériser nos classes.

Les joueurs du cluster 2 représente presque l'ensemble des joueurs à plus de 1 contre par match, on voit que globalement ces joueurs font plus de contres que ceux des autres clusters.




En observant ce graphique, on récupère plusieurs informations :

- Il n'y a pas de grande différence entre le nombre de passes décisives faites par les joueurs du groupe 1 et 3, alors que les joueurs du cluster 2 font peu de passes.
- Les joueurs du cluster 1 réussissent peu de tirs par match (entre 2 et 8), contrairement à ceux du cluster 3 (entre 5 et 11). Sur cette statistique, les joueurs du cluster 2 ont un nombre de tirs réussis variés.

Ce clustering nous a donc permis de séparer les en 3 groupes :

- Les joueurs défensifs, plus grands, faisant beaucoup de contres et de rebonds
- Les joueurs offensifs, marquant beaucoup de points et faisant beaucoup de passes décisives
- Les joueurs moyens qui ne tirent pas beaucoup, mais qui font tout de même des passes décisives

## 2. Clustering à 10 classes

En faisant un clustering à 10 classes, les résultats seront plus intéressants et permettront une plus grande liberté au niveau de l'interprétation.

On recommence le paramétrage du k-means : on crée cette fois-ci 10 clusters, toujours avec 20 itérations et un maximum de 50 itérations.

Les clusters sont composés respectivement de 200, 99, 73, 62, 77, 59, 92, 37, 153 et 94 individus.

Le ratio entre l'inertie expliquée par le clustering et l'inertie totale est de 63.3 %.

Commençons par représenter les clusters selon les points et les rebonds :


On distingue déjà quelques groupes comme le cluster 4, des joueurs marquant beaucoup de points, ou le cluster 8 des joueurs faisant beaucoup de rebonds.

On peut désormais caractériser les classes en étudiant la moyenne de chaque statistique par cluster, on met en surbrillance chaque statistique selon sa valeur par rapport aux autres classes :

| Cluster | FG | %FG | X3P | %X3P | FT | %FT | ORB | DRB | REB | AST | STL | BLK | PTS | Height | Weight | Effectif |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5,51 | 44,40 | 1,87 | **37,75** | 2,38 | 83,02 | 0,73 | 3,36 | 4,10 | 2,92 | 0,88 | 0,31 | 15,26 | 198,00 | 95,71 | 200 |
| 2 | 5,83 | 48,61 | 0,71 | 33,82 | 2,39 | 74,47 | 1,80 | 5,40 | 7,21 | 2,43 | 0,96 | 1,04 | 14,76 | 207,22 | 107,15 | 99 |
| 3 | 5,78 | 50,75 | 0,01 | 4,15 | 2,62 | 72,16 | 2,71 | 6,09 | 8,80 | 2,19 | 0,86 | 1,09 | 14,18 | 209,22 | 114,12 | 73 |
| 4 | **9,23** | 47,51 | **2,43** | 36,41 | **6,47** | 83,86 | 1,05 | 5,55 | 6,60 | 6,39 | 1,42 | 0,58 | **27,35** | 198,16 | 100,44 | 62 |
| 5 | 6,92 | 45,32 | 1,33 | 33,07 | 4,11 | 80,75 | 0,92 | 3,95 | 4,87 | 6,41 | **1,72** | 0,46 | 19,29 | 192,26 | 91,87 | 77 |
| 6 | 8,34 | 50,44 | 0,60 | 29,15 | 4,72 | 78,09 | 2,65 | 7,89 | 10,54 | 3,24 | 0,94 | 1,33 | 22,00 | 210,66 | 115,12 | 59 |
| 7 | 4,83 | 43,52 | 1,09 | 33,30 | 2,42 | 78,05 | 0,70 | 3,26 | 3,96 | **6,49** | 1,41 | 0,31 | 13,16 | 191,18 | 89,41 | 92 |
| 8 | 5,67 | **61,02** | 0,01 | 6,58 | 3,08 | 55,88 | **3,99** | **8,71** | **12,70** | 1,33 | 0,93 | **1,76** | 14,42 | **212,19** | **118,43** | **37** |
| 9 | 3,95 | 42,92 | 1,44 | 35,83 | 1,40 | 75,70 | 0,86 | 3,55 | 4,41 | 2,40 | 1,07 | 0,41 | 10,73 | 199,67 | 98,16 | 153 |
| 10 | 7,66 | 45,59 | 2,09 | 37,40 | 4,20 | **84,06** | 0,81 | 4,01 | 4,82 | 4,29 | 1,04 | 0,42 | 21,61 | 197,40 | 95,81 | 94 |

Cela va nous permettre de caractériser chaque classe. De plus, on relève les individus les plus proches de la moyenne de chaque cluster.[5]

Cluster 1 : *Swingman low defensive*

Ces joueurs ont globalement un très bon pourcentage aux tirs à 3 points et aux lancers-francs mais font peu de rebonds, d'interceptions ou de contres. Ils ont une moyenne de 15.26 points.

Ce sont donc des joueurs offensifs, jouant arrière ou ailier, adroits mais très peu actif en défense.

Joueurs types :

- Wesley Matthews (2013 - 2014)
- Rodney Hood (2015 - 2016)
- Evan Fournier (2018 - 2019)

Cluster 2 : *Versatile power forward*

Ces joueurs sont assez polyvalents, faisant pas mal de rebonds et plutôt athlétique (taille et poids). C'est également le cluster composé du plus grand nombre de joueurs : 200.

Ce sont donc des ailiers forts, athlétique et très polyvalent.

Joueurs types :

- Paul Millsap (2012 - 2013)
- Chris Bosh (2013 - 2014)
- David West (2013 - 2014)

Cluster 3 : *Athletic & defensive power forward*

[5] Pour cela, on utilise une boucle qui calcule la distance euclidienne entre les moyennes du cluster et les statistiques de chaque joueur.

Ces joueurs sont similaires à ceux du cluster 2 mais prenant plus de rebonds et plus athlétique, en revanche ils ne prennent pas de tirs à 3 points et font peu d'interceptions, ils sont donc moins polyvalents.

Ce sont donc des ailiers forts, axés sur la défense et très athlétique.

Joueurs types :

- Elton Brand (2010 - 2011)
- Derrick Favors (2013 - 2014)
- Greg Monroe (2015 - 2016)

Cluster 4 : *Offensive weapon*

Ces joueurs ont globalement de très bonnes statistiques, ils ont une moyenne de 27.35 points (la plus élevée) et sont ceux qui prennent le plus de tirs, tirs à 3 points et de lancer-francs par match. Ils font également beaucoup de passes décisives et ne sont pas mauvais aux rebonds.

Ce cluster est donc composé des meilleurs joueurs de la NBA, ils sont bons partout mais excellent en attaque et sont les leaders offensifs de leur équipe.

Joueurs types :

- Kobe Bryant (2012 - 2013)
- Bradley Beal (2018 - 2019)
- Kawhi Leonard (2019 - 2020)

Cluster 5 : *Offensive shooting guard*

Ces joueurs sont assez petits, marquent 19.29 points de moyenne, font beaucoup de passes décisives et peu de rebonds.

Ce sont des meneurs de jeu portés sur l'attaque

Joueurs types :

- Kemba Walker (2012 - 2013)
- Monta Ellis (2013 - 2014)
- Eric Bledsoe (2014 - 2015)

Cluster 6 : *Two-way center*

Ces joueurs sont très grands (2m10 de moyenne), marquent beaucoup de points (22 points de moyenne), prennent 10.54 rebonds en moyenne, etc. Ils sont polyvalents sauf aux tirs à 3 points.

Ce sont des pivots aussi bien à l'aise en attaque qu'en défense.

Joueurs types :

- Lamarcus Aldridge (2014 - 2015)
- Joel Embiid (2017 - 2018)
- Joel Embiid (2019 - 2020)

Cluster 7 : *Passer & non scorer guard*

Ce cluster est similaire au cluster 5, les joueurs sont petits, font beaucoup de passes décisives et très peu de rebonds, mais ils marquent peu de points.

Ce sont des meneurs de jeu portés sur l'attaque mais qui marquent peu, ils servent leurs coéquipiers en faisant des passes décisives.

Joueurs types :

- Jrue Holiday (2010 - 2011)
- Kyle Lowry (2010 - 2011)
- Jeremy Lin (2012 - 2013)

Cluster 8 : *Defensive tower*

Les joueurs de ce cluster sont adroits, sont les meilleurs en rebonds, en contres et sont les plus grands.

Ce sont des joueurs très athlétiques, portés sur la défense en faisant beaucoup de rebond et de contres.

Ce cluster est le plus restreint, ces joueurs sont rares et donc très recherchés par les équipes.

Joueurs types :

- Dwight Howard (2015 - 2016)
- Dwight Howard (2016 - 2017)
- Clint Capela (2018 - 2019)

Cluster 9 : *Average player*

Ce cluster est composé de joueurs moyens voir mauvais, sans caractéristiques particulières, si ce n'est leur bon pourcentage de tirs réussis à 3 points.

Joueurs types :

- James Anderson (2013 - 2014)
- George Hill (2015 - 2016)
- Dillon Brooks (2017 - 2018)

Cluster 10 : *Offensive player*

Ce cluster est similaire au cluster 1 mais en meilleur. Les joueurs ont des bons pourcentages aux 3 points et aux lancers-francs et ont 21.61 points de moyenne.

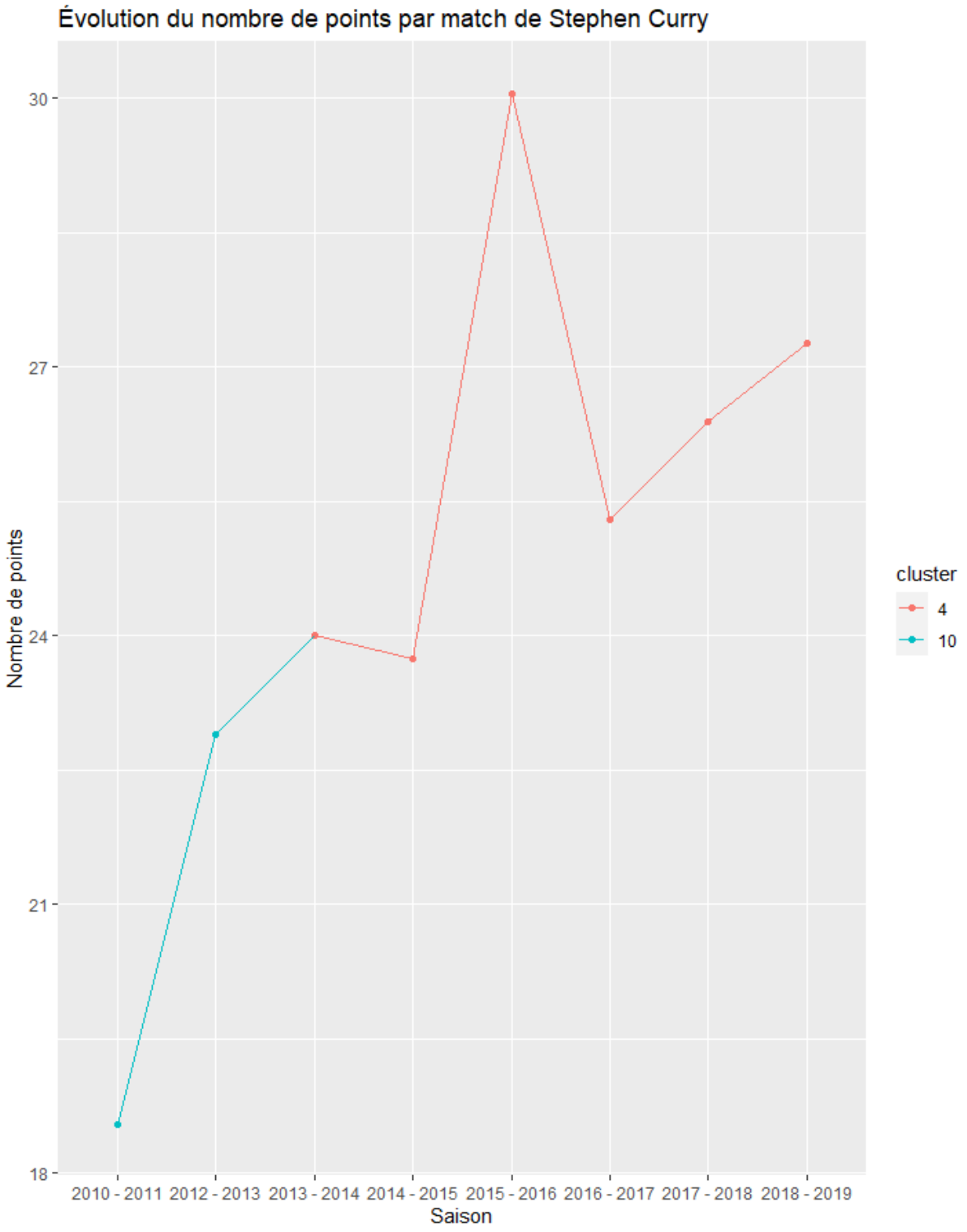
Ces joueurs sont de bons attaquants (meilleur que ceux du cluster 1 mais moins bon que ceux du cluster 4).

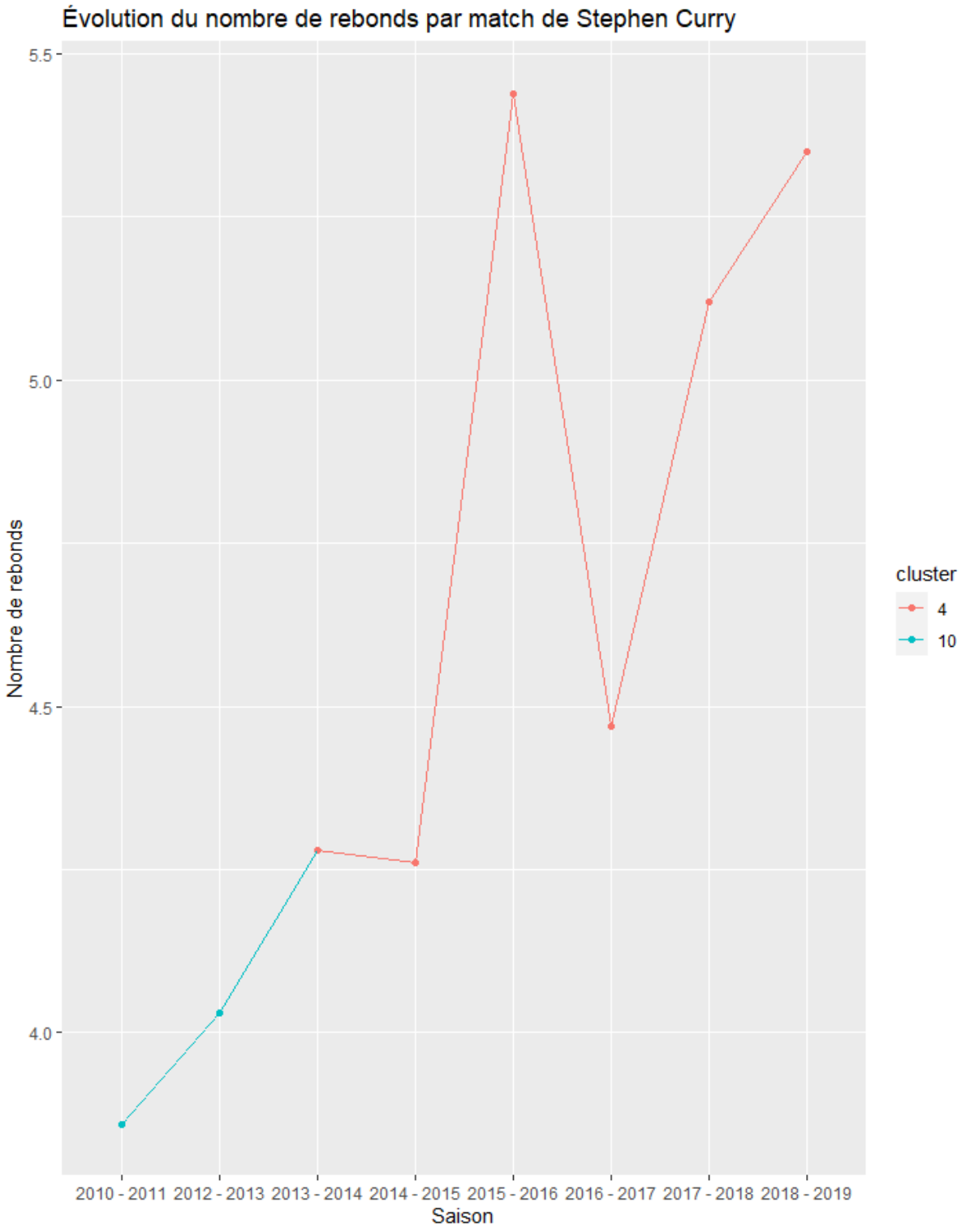
Joueurs types :

- Gordon Hayward (2016 - 2017)
- Bradley Beal (2017 - 2018)
- Zach Lavine (2018 - 2019)

On peut s'intéresser à quelques joueurs qui ont changé de clusters au cours des saisons, pour voir quel à était leur évolution.

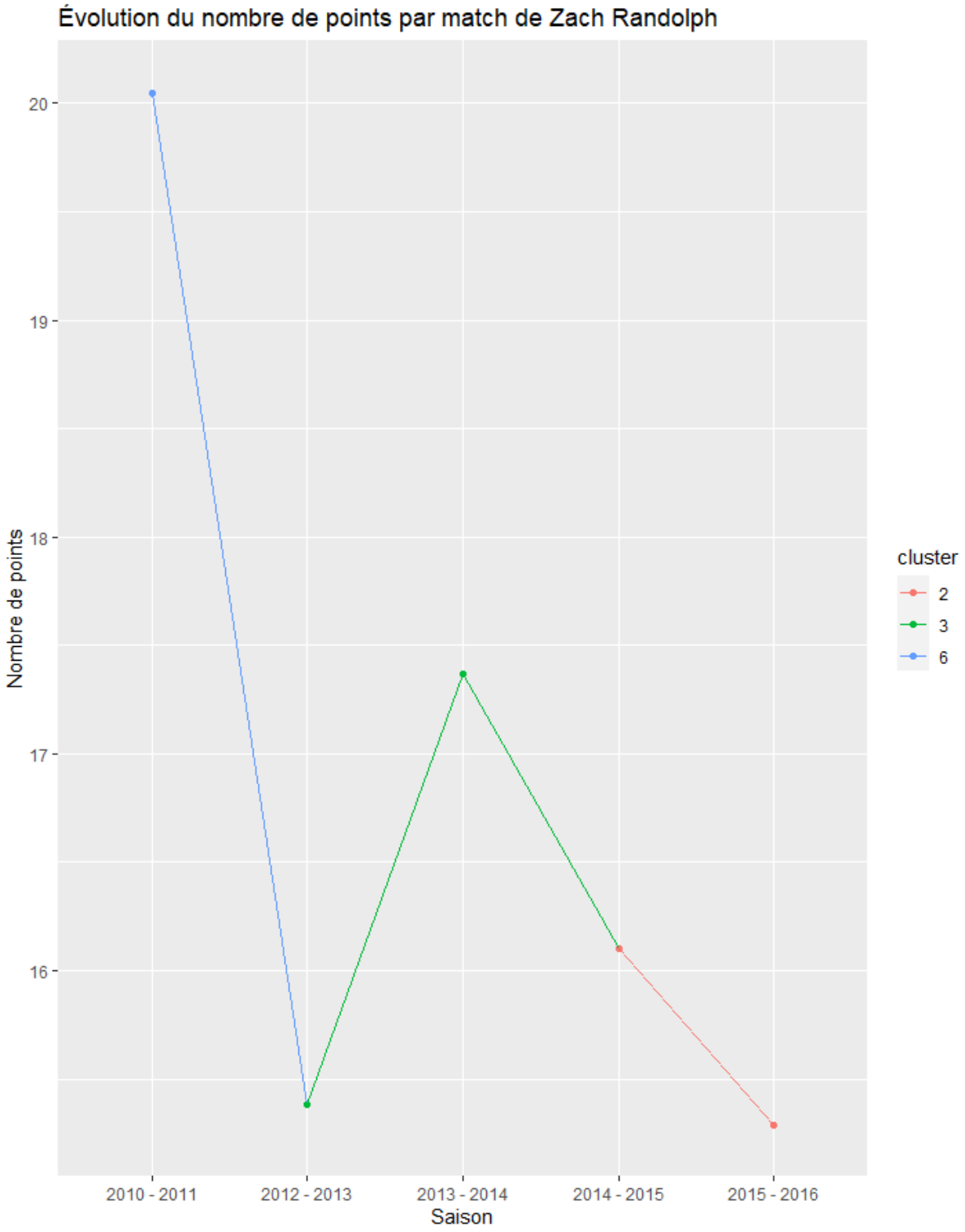
Évolution de Stephen Curry :

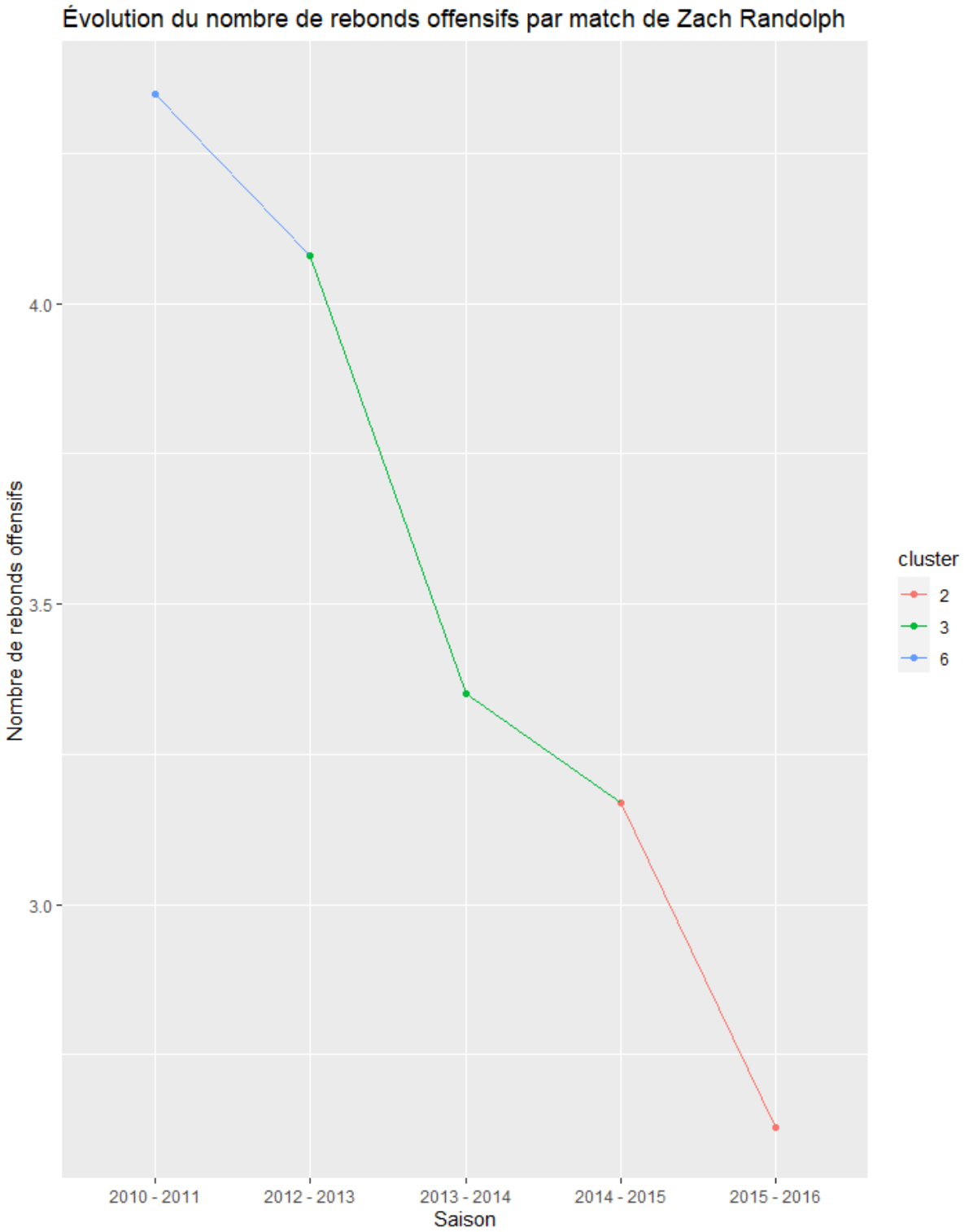

Évolution du nombre de points par match de Stephen Curry
Nombre de points
30
27
24
21
18
2010 - 2011
2012 - 2013
2013 - 2014
2014 - 2015
2015 - 2016
2016 - 2017
2017 - 2018
2018 - 2019
Saison
cluster
4
10




Stephen Curry est passé en 2013-2014 dans le cluster 4, ce qui signifie qu'à partir de cette saison il est devenu meilleur en attaque et plus polyvalent en général. Cela est confirmer avec les graphiques, où l'on voit l'augmentation du nombre de points et de rebonds par match. Stephen Curry a donc eu une évolution positive au cours de sa carrière.

Zach Randolph :






Zach Randolph est passé des *Two way players* (bon en attaque et défense) au *Athletic & defensive power forward* puis au *Versatile power forward*. Cette « chute » de niveau se reflète par l'évolution de son nombre de points et de rebonds offensifs par match.

# IV. Apprentissage supervisé : random forest

Dans la course à l'information que l'on connait aujourd'hui, les médias souhaitent connaitre quels joueurs feront partie de la All-NBA Team afin de préparer leurs articles à l'avance et de pouvoir le mettre en ligne le plus rapidement possible. Pour cela il peut être intéressant de faire une prédiction en utilisant l'algorithme de forêts aléatoires.

Ce dernier fonctionne en faisant plusieurs arbres de décisions sur des sous-ensembles aléatoires de notre jeu de donnée. La prédiction est ensuite faite en sélectionnant la variable apparu le plus fréquemment.

Commençons par séparer notre jeu de données en 2 sous-datasets : *train* et *test*. Le dataset d'entrainement servira à entrainer notre modèle et le dataset de test servira à vérifier que notre modèle fonctionne sur d'autres données que celles qu'il connait.

On peut maintenant paramétrer notre algorithme, on choisit par défaut de faire 500 arbres et de tester 3 variables à chaque division de nœuds. On peut désormais lancer notre algorithme sur le dataset *train*.

## 1. Entrainement du modèle

On obtient un taux d'erreur *Out Of Bag* de 8.98 %. Cela nous sert à mesurer la performance de notre modèle, plus il est proche de 0, meilleur est le modèle. Ici, 91.02 % des estimations faites par notre modèle sont corrects (c'est l'accuracy).

| Matrice de confusion | | | | |
|---|---|---|---|---|
| | | Valeur effective | | |
| Valeur prédite | | No | Yes | Total |
| | No | 614 | 23 | 637 |
| | Yes | 45 | 75 | 120 |
| | Total | 659 | 98 | 757 |

Notre modèle à correctement fait les prédictions suivantes :

- 614 joueurs ne feront pas partie de la All-NBA Team, ce sont les vrais négatifs.
- 75 joueurs feront partie de la All-NBA Team, ce sont les vrais positifs.

Ces 689 joueurs représentent bien 91 % du total.

Mais notre modèle à également fait les erreurs suivantes :

- 23 joueurs n'ont pas étés prédit dans la All-NBA Team, alors qu'ils le sont, ce sont les faux négatifs.

- 45 joueurs ont étés prédit dans la All-NBA Team, alors qu'ils ne le sont pas, ce sont les faux positifs.

Ces 68 joueurs représentent bien 8.98 % du total.

On peut calculer la sensibilité et la spécificité du modèle, qui sont respectivement les taux de vrais positifs et de vrais négatifs. Ce sont des indicateurs de l'efficacité de notre modèle :

- Sensibilité = 75/98 = 77 %
- Spécificité = 614/659 = 93 %

On peut également calculer les taux d'erreurs α et β :

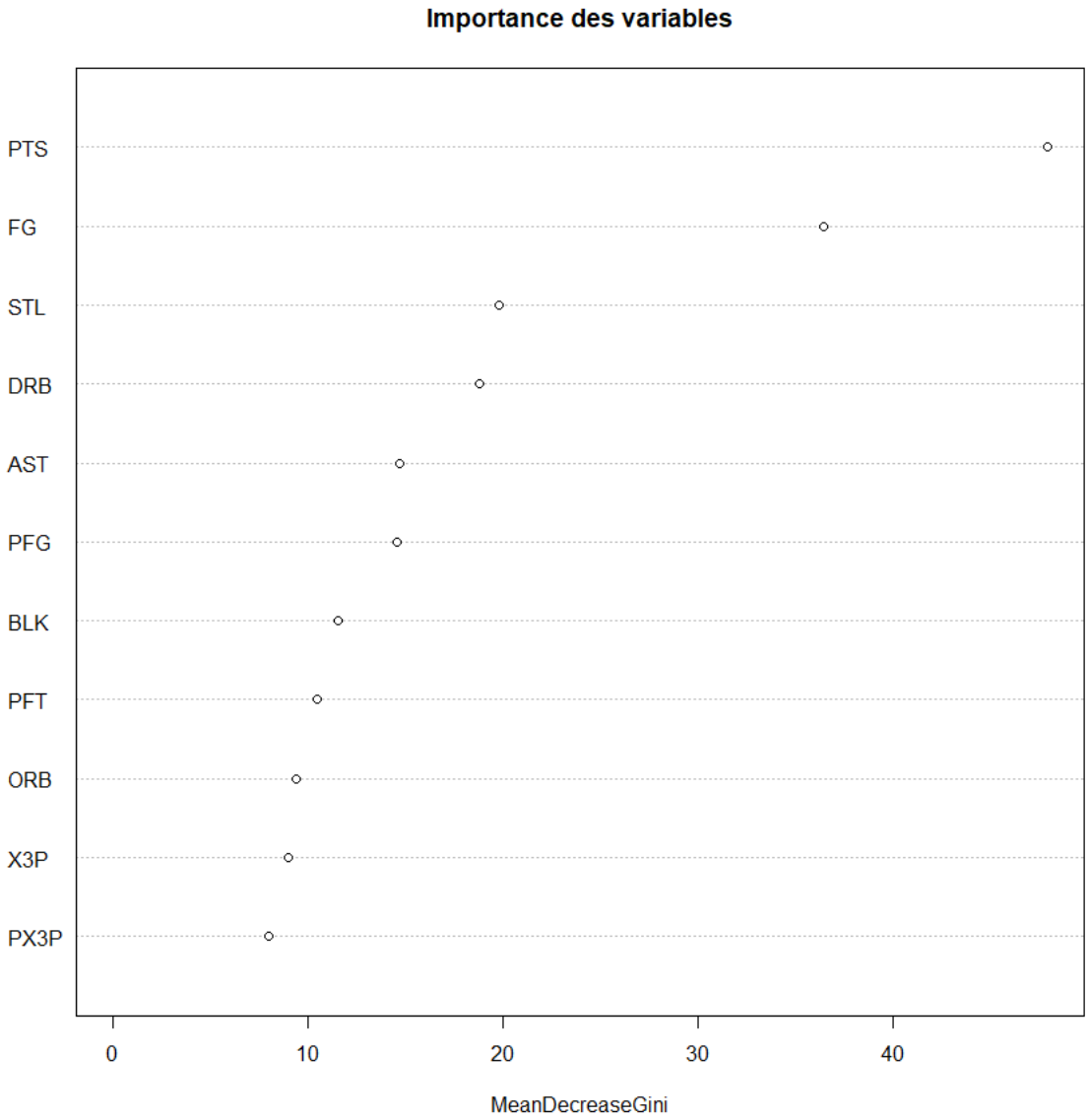
- Taux d'erreur α = 45/659 = 7 %
- Taux d'erreur β = 23/98 = 23 %

Il est donc rare que notre modèle prédise qu'un joueur fait partie de la All-NBA Team, alors qu'il ne l'est pas (7 % des joueurs n'en faisant pas partie). En revanche, parmi les joueurs ayant obtenu cette récompense, 23 % n'ont pas été prédit par le modèle.

Dans notre cas, prédire qu'un joueur fera partit de la All-NBA Team à tort serait moins grave que l'inverse, c'est-à-dire d'en oublier un.[6]

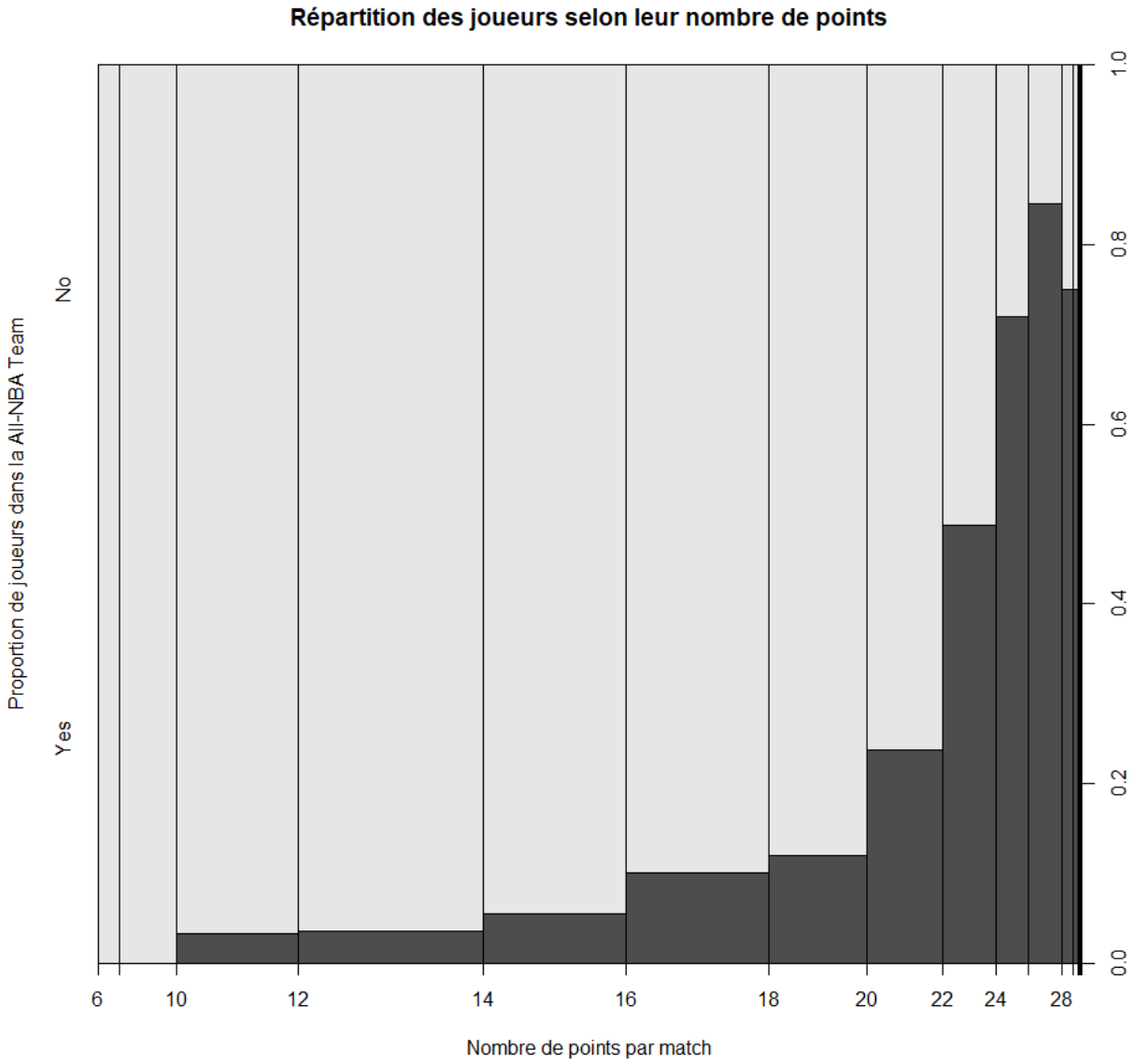
On peut également s'intéresser à l'importance des variables, afin de savoir lesquelles discriminent le plus le fait d'appartenir à la All-NBA Team ou non. Celle-ci est calculé avec la décroissance moyenne de l'indice de Gini.

[6] Evidemment, dans notre cas les erreurs ne sont pas très grave contrairement à des domaines comme le médical par exemple.

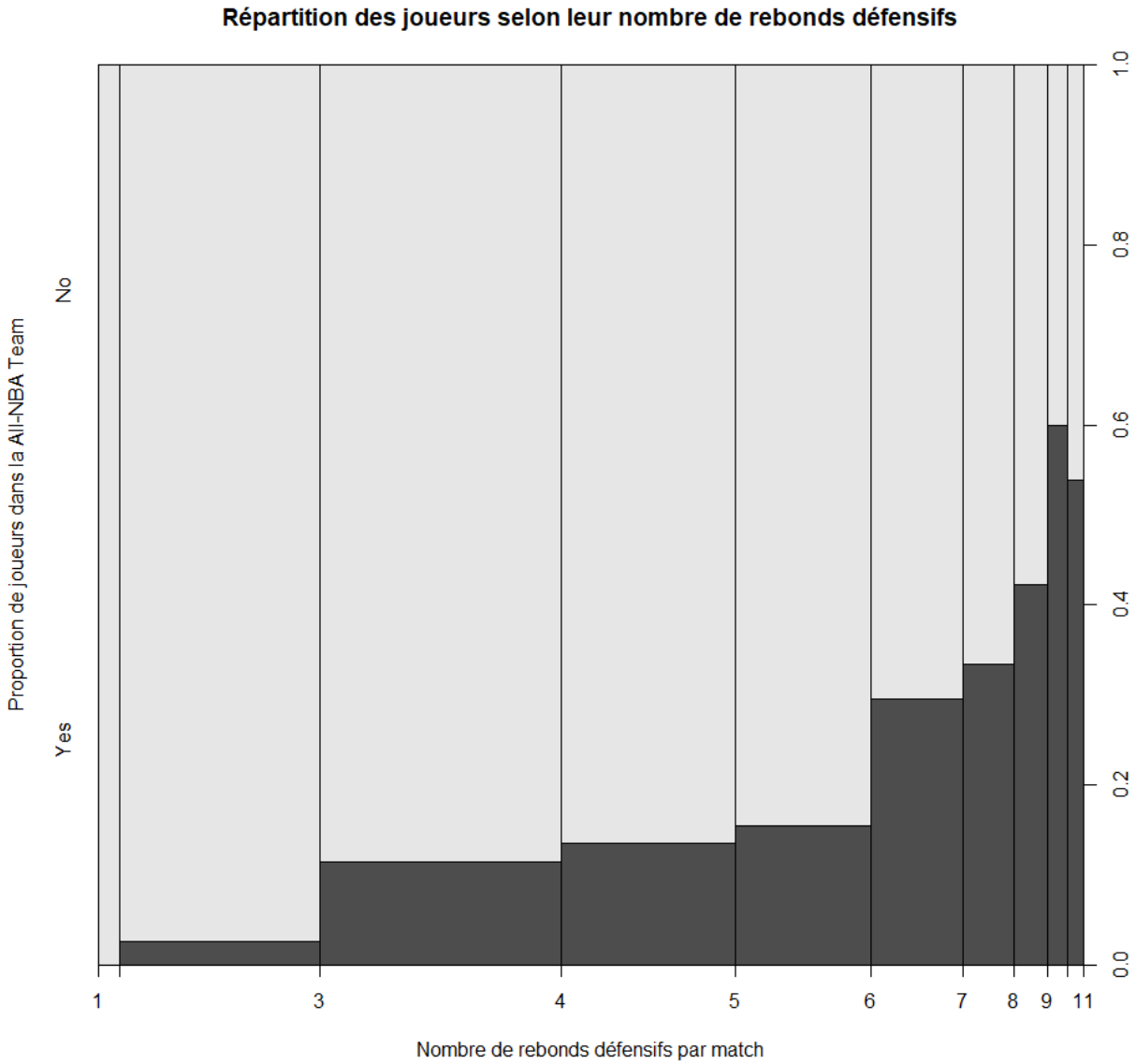



Les variables du nombre de points (PTS) et du nombre de tirs réussis (FG : Fields Goals) sont de loin les plus importante dans notre modèle pour choisir si un joueur va faire partie de la All-NBA Team. Le nombre d'interceptions, de rebonds défensifs et de passes décisives viennent ensuite.

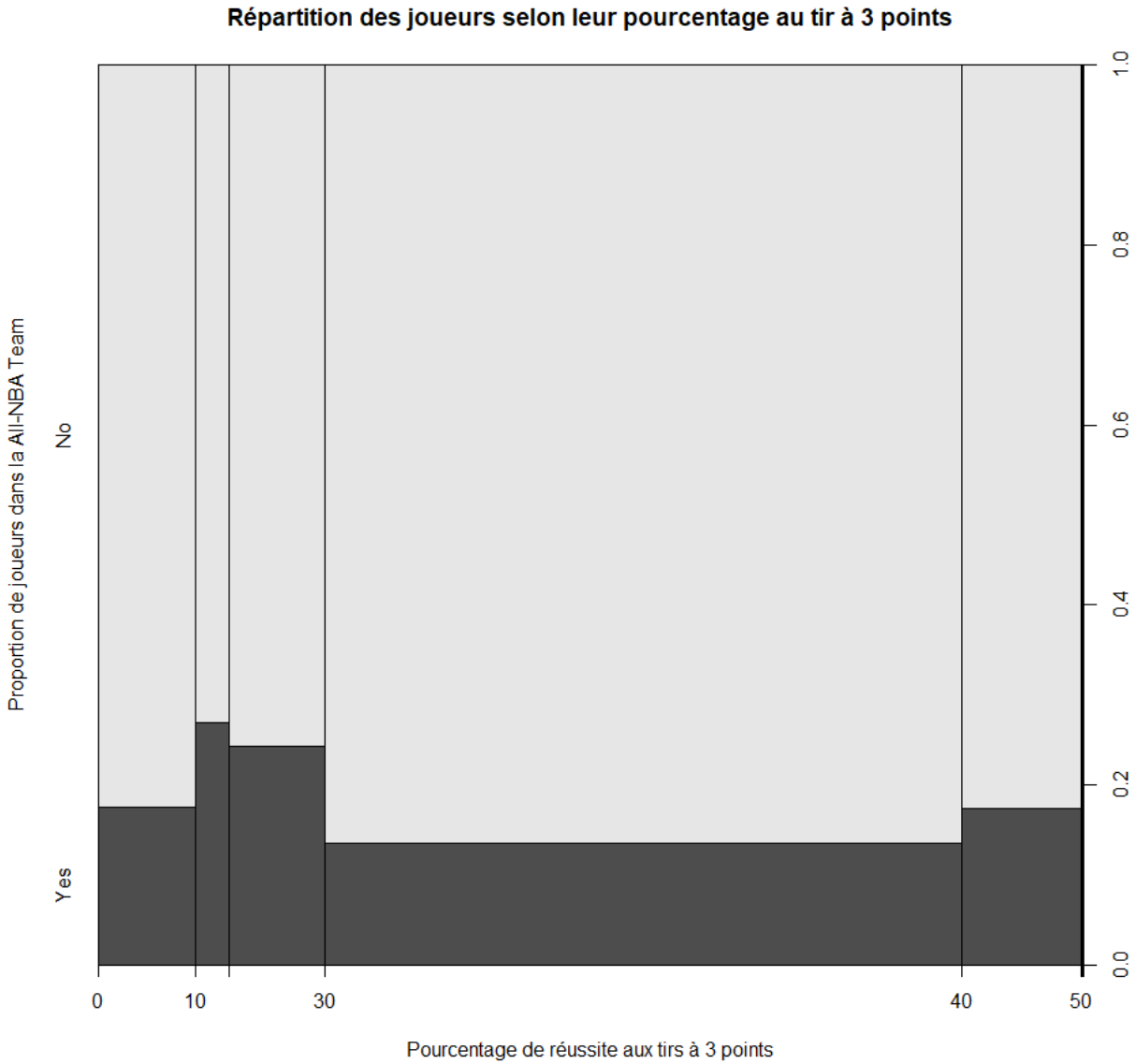
On peut tracer quelques graphiques pour vérifier cela, ci-dessous la répartition des joueurs selon certaines statistiques, avec en foncé la proportion de joueur faisant partie de la All-NBA Team :




On voit bien que le nombre de points influe grandement sur le fait d'être dans la All-NBA Team. Aucun joueur ayant moins de 10 points ne l'est, moins de 20 % des joueurs ayant moins de 20 points ne le sont. En revanche, à partir de 24 points de moyenne, les chances augmentent à 75 % environ.




On suit le même raisonnement pour le nombre de rebonds défensifs, bien que cette variable soit moins importante que les points. Les joueurs faisant beaucoup de rebonds défensifs sont plus souvent dans la All-NBA Team que les autres.

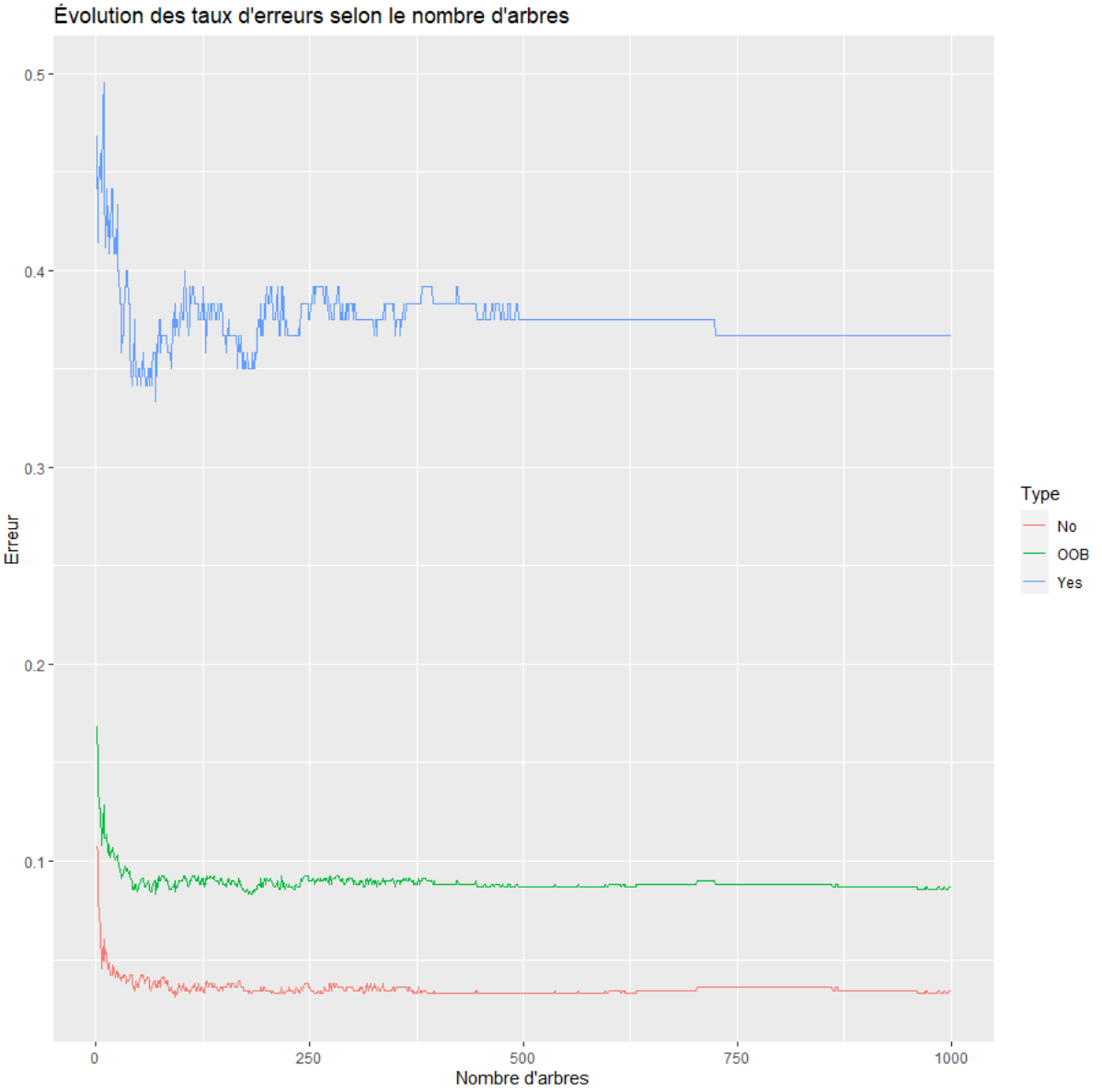



En étudiant l'importance des variables, on a vu que le taux de réussite de tirs à 3 points influe peu le fait d'être dans la All-NBA Team. On le voit bien ici où la proportion de joueurs de la All-NBA Team est la même pour les joueurs très adroits aux 3 points (50 % de réussite) que pour les joueurs très peu adroits (0 % de réussite). Ce critère est donc peu ou pas important.

Enfin, on peut regarder le pourcentage de chances de chaque joueur d'être dans la All-NBA Team d'après notre modèle. Par exemple, Russell Westbrook en 2019 - 2020 avait 90 % de chances de recevoir la récompense, Spencer Dinwiddie, en avait seulement 6 %

On revient maintenant sur le paramétrage de notre algorithme, modifier les paramètres va-t-il influencer les taux d'erreur obtenus ?

Pour cela on refait un random forest, avec cette fois 1000 arbres, puis on trace l'évolution des taux d'erreurs selon le nombre d'arbres utilisé dans l'algorithme.




Au-delà de 500 arbres, les taux d'erreurs convergent, on décide qu'il n'est pas nécessaire de changer le nombre de 500.

## 2. Test du modèle

On peut maintenant effectuer notre algorithme sur notre échantillon test, pour vérifier que le modèle fonctionne sur des nouvelles données qu'il ne connait pas.

Voici la matrice de confusion :

| Matrice de confusion | | | | |
|---|---|---|---|---|
| | | Valeur effective | | |
| Valeur prédite | | No | Yes | Total |
| | No | 153 | 10 | 163 |
| | Yes | 6 | 20 | 26 |
| | Total | 159 | 30 | 189 |

Le taux d'erreur globale est similaire, il est de 8.47 %.

La spécificité est de 96 %, ce qui est meilleur que sur notre échantillon *train* mais la sensibilité est de 67 % ce qui est plus faible.

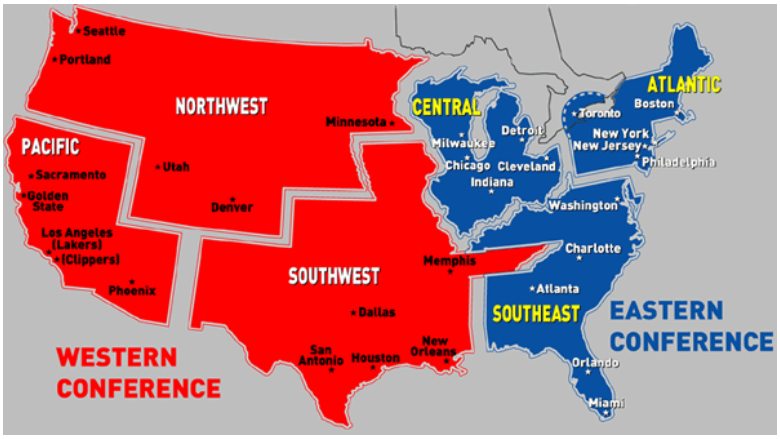
Au final, on valide notre modèle.

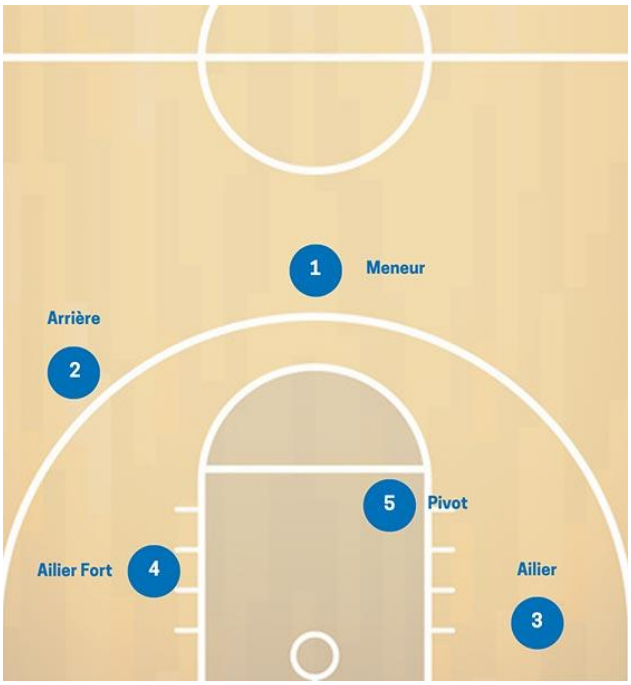
## ANNEXE

Règle de la NBA.

C'est la ligue de basket-ball américaine, Il y a 30 équipes (des franchises), une saison comporte deux parties : la saison régulière (82 matchs) et ensuite les playoffs (séries éliminatoires).

Le championnat est divisé en deux conférences : l'Est et l'Ouest. Ces deux conférences sont-elles même divisées en 3 parties (divisions).






Le meneur de jeu est celui qui mène le jeu, il a la balle au départ des actions, et tente de mettre en place les tactiques imaginées par les coachs grâce à ses passes. Un bon meneur de jeu se juge sur un tas de choses, mais statistiquement, surtout sur son nombre de points et son nombre de passes décisives. Les meneurs sont généralement les plus petits de l'équipe.

Le rôle de l'arrière est de prendre pas mal de tir à 3 points. Mais finalement, aussi souvent de pénétrer dans la raquette pour aller marquer en raison de sa vivacité (plus petite que celle d'un meneur mais plus grande que le reste de l'équipe)

L'ailier, lui, est un peu un intermédiaire. Plus costaud que l'arrière, plus fin et plus vif que les ailiers-fort ou pivot. C'est un poste assez complet en général.

L'ailier fort, est comme son nom l'indique un ailier... mais en plus fort. Plus grand et plus gros en somme. C'est un poste défensif, qui se joue assez près du panier. Ceci dit, les ailiers fort savent aussi apporter beaucoup offensivement en raison de leur taille.

Le pivot, c'est le joueur le plus gros et le plus grand de l'équipe. Pour être un bon pivot il faut être inbougeable. Un roc. Il a donc surtout un rôle défensif dans la raquette, sous le panier. Mais néanmoins un pivot doit pouvoir apporter sur le plan offensif. On lui a donné ce nom parce qu'il se retrouve souvent sous le panier à seulement "pivoter" pour marquer, ou défendre (en raison de sa lenteur aussi).

Source : http://unoeilsurlebasket.over-blog.com/article-les-postes-43677216.html