IUT de Paris - Rives de Seine
Université Paris Cité

RENDU TD/ TP

DOMAINE APPLICATION

Projet : Etude des déterminants de la performance en Basket Ball

JUIN 2025

Hajar CHAKIR
Kinaza ALI
Nadia ZIDAN

BUT 2 EMS 21 FI

Table des matières

AVANT-PROPOS

Dans un contexte où la performance sportive est au cœur des enjeux de sélection, d'entraînement et de carrière des joueurs, comprendre les facteurs qui influencent les performances en basket-ball devient essentiel. Ce rapport s'inscrit dans cette démarche d'analyse, en s'appuyant sur une base de données riche issue de la NBA, couvrant plusieurs décennies de carrière de milliers de joueurs.

À travers une approche rigoureuse mêlant exploration statistique, modélisation et visualisation, notre objectif est de mettre en lumière les **principaux déterminants de la performance individuelle**. Nous avons choisi de focaliser notre attention sur des indicateurs objectifs comme le nombre de points marqués, les rebonds, les passes décisives, ou encore les minutes jouées, tout en étudiant leurs liens avec des variables explicatives telles que l'âge, la taille, le poste ou la durée de carrière.

Le rapport débute par une prise en main des données et un nettoyage des variables afin de garantir la fiabilité des résultats. Nous avons ensuite entrepris une série d'analyses, notamment :

- une **analyse descriptive** des performances des joueurs,
- une **étude de la relation entre la performance et l'âge**,
- une **comparaison selon les postes** occupés,
- ainsi qu'une **analyse factorielle** pour synthétiser les dimensions principales expliquant la performance.

Au-delà des chiffres, ce travail vise à proposer des clés de lecture concrètes sur l'évolution de la performance sportive au fil de la carrière, en posant des constats objectivés et interprétables.

Partie technique, analyse exploratoire des données

PARTIE TECHNIQUE, TRAITEMENT ET NETTOYAGE DES DONNEES

Dans cette première partie, l'aspect technique de notre ouvrage est abordé (Cf. question 1 consigne). En effet, au préalable, il a été nécessaire, lors de la collecte de la base de données, de l'exploiter de manière conforme afin de pouvoir en exploiter son potentiel et répondre ainsi à la problématique.

D'abord, l'objectif est d'entreprendre une analyse exploratoire des données afin de comprendre la nature des données que nous avons a notre disposition. Ces donnes proviennent de l'Insep et regroupent des informations sur des joueurs de la NBA et ceux dans l'objectifs de retracer leur performance. La base de données à notre disposition est composée de 35 variables et 17 405 lignes. Cette base de données compare les résultats entre l'année n et l'année n-1 de joueur de la NBA et ceux entre 1949 et 2010.

35 variables

17405 individus

L'objectif ici est donc de classifier les 35 variables à notre disposition en deux catégories qui sont des variables qui caractérisent la performance des joueurs ou alors d'hypothétiques déterminants.

20 variables de qualitatives de performance

field_goal_made	def-rebond
field_goal_attempted	assists
field_goal_percentage	steals
three_points_made	blocks
three_points_attempted	turnover
three_points_percentage	points
free_throws_made	minutes_played
free_throws_attempted	fouls
free_throws_percentage	off_rebond
GS	duree_carriere

Games Started (GS) → Matchs commencés : Nombre de fois où un joueur a été titulaire (dans le cinq de départ) lors d'un match.

games : Nombre de matchs joués sur la saison étudiée

Tirs de champ (Field Goals) :

- **Made (FGM) → Tirs réussis** : Nombre total de tirs réussis (2 ou 3 pts).
- **Attempted (FGA) → Tirs tentés** : Nombre total de tirs tentés (2 et 3 pts).
- **Percentage (FG%) → Pourcentage de réussite aux tirs** : FGM/FGA

Tirs à trois points (Three-Point Shots) :

- **Made (3PM) → Tirs à trois points réussis** : Nombre de tirs marqués à trois points.
- Attempted (3PA) → Tirs à trois points tentés : Nombre de tirs à trois points tentés.
- Percentage (3P%) → Pourcentage de réussite à trois points

Lancers francs (Free Throws) :

- Free Throws Made (FTM) → Lancers francs réussis : Nombre de lancers francs marqués.
- Free Throws Attempted (FTA) → Lancers francs tentés : Nombre total de lancers francs tirés.
- Free Throws Percentage (FT%) → Pourcentage de réussite aux lancers francs : F

Rebonds :

- Offensive Rebounds (OREB) → Rebonds offensifs : Nombre de rebonds récupérés par un joueur en attaque (après un tir manqué de son équipe).
- Defensive Rebounds (DREB) → Rebonds défensifs : Nombre de rebonds captés en défense (après un tir manqué de l'adversaire).

Autres statistiques :

- Assists (AST) → Passes décisives : Nombre de passes qui mènent directement à un panier marqué.
- Steals (STL) → Interceptions : Nombre de ballons volés à l'adversaire.
- Blocks (BLK) → Contres : Nombre de tirs adverses bloqués.
- Turnover (TO) → Balles perdues : Nombre de fois où un joueur perd la possession du ballon (passe ratée, violation, interception…).
- Fouls (PF) → Fautes personnelles : Nombre de fautes commises par un joueur.
- points : nombre de points marqués sur la saison
- minutes_played : nombre de minutes jouées sur la saison
- **duree_carriere :** Durée de la carrière du joueur en NBA

8 variables de déterminants.

age	taille
first_match	poste
last_matcg	age_last_match
age_first_match	games

age : age du joueur lors de la saison étudiée
first_match & last match : années du premier et dernier match du joueur en NBA
age_first_match & age_last_match : age du joueur pour son premier match et son dernier match en NBA
Poids : Poids en kg
Taille : Taille en cm
Poste : poste du joueur sur la saison étudiée
games : Nombre de matchs joués sur la saison étudiée

Les 6 autres variables sont des variables de caractérisation des individus et donc n'ont aucun impact direct ou indirect avec la performance.

ii- Analyse descriptive des variables de performance

L'ensemble des variables caractérisant la performance sont de type numérique, ainsi une boucle est faite sur le logiciel R afin de faire des graphiques pour étudier la dispersion et variation.

	skim_variable	n_missing	complete_rate	mean	sd
1	games	0	1	57.0	24.8
2	GS	8254	0.526	36.2	29.9
3	field_goal_made	194	0.989	229.	198.
4	field_goal_attempted	44	0.997	498.	416.
5	field_goal_percentage	194	0.989	0.439	0.0780
6	three_points_made	9645	0.446	35.1	43.5
7	three_points_attempted	6977	0.599	75.2	107.
8	three_points_percentage	9645	0.446	0.319	0.128
9	free_throws_made	497	0.971	124.	121.
10	free_throws_attempted	401	0.977	165.	155.
11	free_throws_percentage	497	0.971	0.725	0.121
12	off_rebond	3603	0.793	75.8	71.5
13	def.rebond	3397	0.805	170.	153.
14	assists	412	0.976	136.	145.
15	steals	3773	0.783	48.3	40.9
16	blocks	4381	0.748	31.0	41.4
17	turnover	4419	0.746	88.0	70.7
18	points	132	0.992	593.	516.
19	minutes_played	361	0.979	1383.	959.
20	fouls	0	1	132.	87.4

	p0	p25	p50	p75	p100	hist
1	1	41	66	78	88	
2	1	7	29	68	83	
3	1	64	182	345	1597	
4	1	146	405	756	3159	
5	0.056	0.4	0.443	0.482	1	
6	1	3	15	54	269	
7	1	4	22	109	678	
8	0.036	0.25	0.327	0.376	1	
9	1	31	88	182	840	
10	1	44	121	241	1363	
11	0.125	0.667	0.743	0.805	1	
12	1	21	53	111	587	
13	1	50	133	242	1111	
14	1	31	90	191	1164	
15	1	16	39	70	301	
16	1	6	16	38	456	
17	1	28	73	133	366	
18	1	163	468	895	4029	
19	1	496.	1330	2189	3882	
20	0	54	134	201	386	

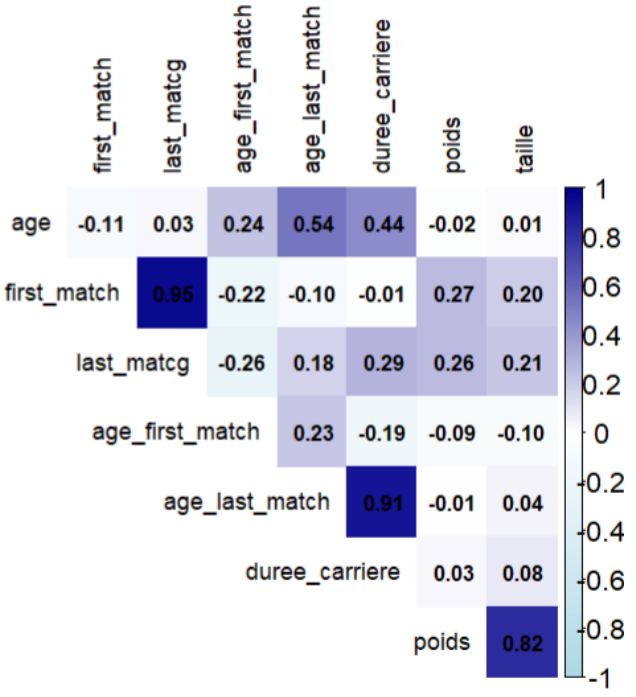
Une vue d'ensemble sur les indicateurs de tendance centrale est faite ainsi que sur les indicateurs de dispersion. En moyenne, sur toute la base de données 73% des lancés francs sont réussi.

Une information sur la présence des données par variable est aussi présente. On constate que la variable Games n'a aucune valeurs manquantes alors que la variable du taux de trois points réussis indique avoir plus de la moitié de ses données qui manquent.

La dispersion est également étudiée ainsi on peut pour chaque variable de performance avoir un ordre d'idée global des résultats pour tous joueurs confondus et toutes saisons confondues.

iii- Analyse descriptive des variables de déterminants.

```
      age          first_match     last_matcg    age_first_match
 Min.   :18.00   Min.   :1949   Min.   :1949   Min.   :18.00
 1st Qu.:24.00   1st Qu.:1974   1st Qu.:1981   1st Qu.:22.00
 Median :26.00   Median :1986   Median :1996   Median :22.00
 Mean   :26.61   Mean   :1984   Mean   :1992   Mean   :22.59
 3rd Qu.:29.00   3rd Qu.:1996   3rd Qu.:2006   3rd Qu.:23.00
 Max.   :44.00   Max.   :2010   Max.   :2010   Max.   :63.00
 NA's   :10      NA's   :15     NA's   :15     NA's   :32
 age_last_match  duree_carriere      poids            taille
 Min.   :18.00   Min.   : 1.000   Min.   : 60.33   Min.   :160.0
 1st Qu.:27.00   1st Qu.: 6.000   1st Qu.: 86.18   1st Qu.:193.0
 Median :31.00   Median :10.000   Median : 95.25   Median :200.7
 Mean   :30.82   Mean   : 9.233   Mean   : 95.53   Mean   :199.4
 3rd Qu.:34.00   3rd Qu.:13.000   3rd Qu.:102.97   3rd Qu.:205.7
 Max.   :68.00   Max.   :44.000   Max.   :149.69   Max.   :231.1
 NA's   :32      NA's   :31       NA's   :6
```



Matrice de corrélation des variables numérique de

poste	n	pourcentage
Arri\xe8re	7312	42.0
Interieur	3089	17.7
Meneur	7004	40.2

Pour les variables de déterminant, seule deux d'entres elles sont de type qualitatif, ainsi un tableau de fréquence est réalisé pour la variable poste, elle regroupe 3 postes différents ainsi que le pourcentage de joueurs occupant ce poste. Une lie pourra donc être fait entre le poste occupé et les résultats du joueur par exemple.

La deuxième variable est la variable team et cette dernière regroupe 66 équipes.

Ainsi, cette première analyse descriptive qui divise la base de données en 2 grands axe nous permet d'avoir un ordre d'idées sur les axes de recherche du reste du TP et d'explorer les donnes .

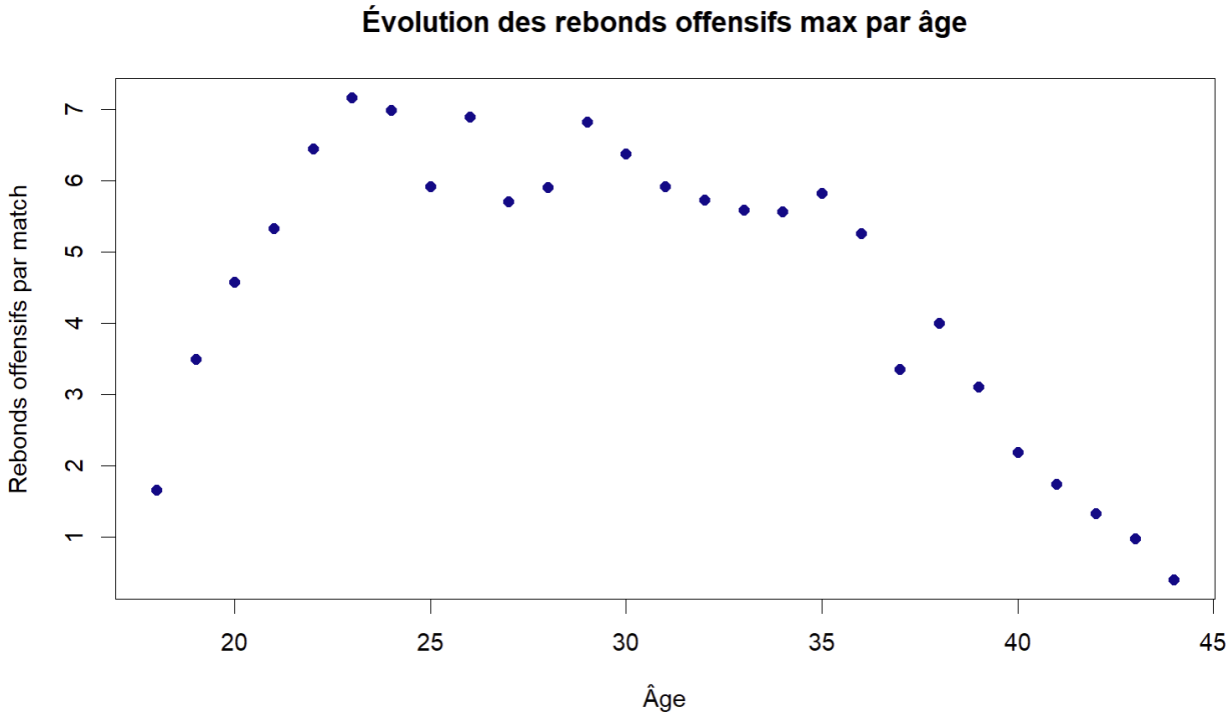
Analyse de la performance en fonction de l'âge

Analyse de la performance en fonction de l'âge

Dans cette section, nous cherchons à étudier la manière dont la performance des joueurs évolue selon leur âge. Parmi les différentes statistiques disponibles, nous avons choisi d'analyser le nombre de rebonds offensifs par match, une mesure pertinente de l'implication physique et stratégique dans le jeu.

Choix de l'indicateur : rebonds offensifs par match

Pour obtenir cette variable, nous avons calculé le ratio entre le nombre total de rebonds offensifs et le nombre de matchs joués par saison. Afin de rendre les comparaisons cohérentes, nous avons ensuite extrait le nombre maximal de rebonds offensifs par âge, pour observer les meilleures performances atteintes à chaque tranche d'âge.

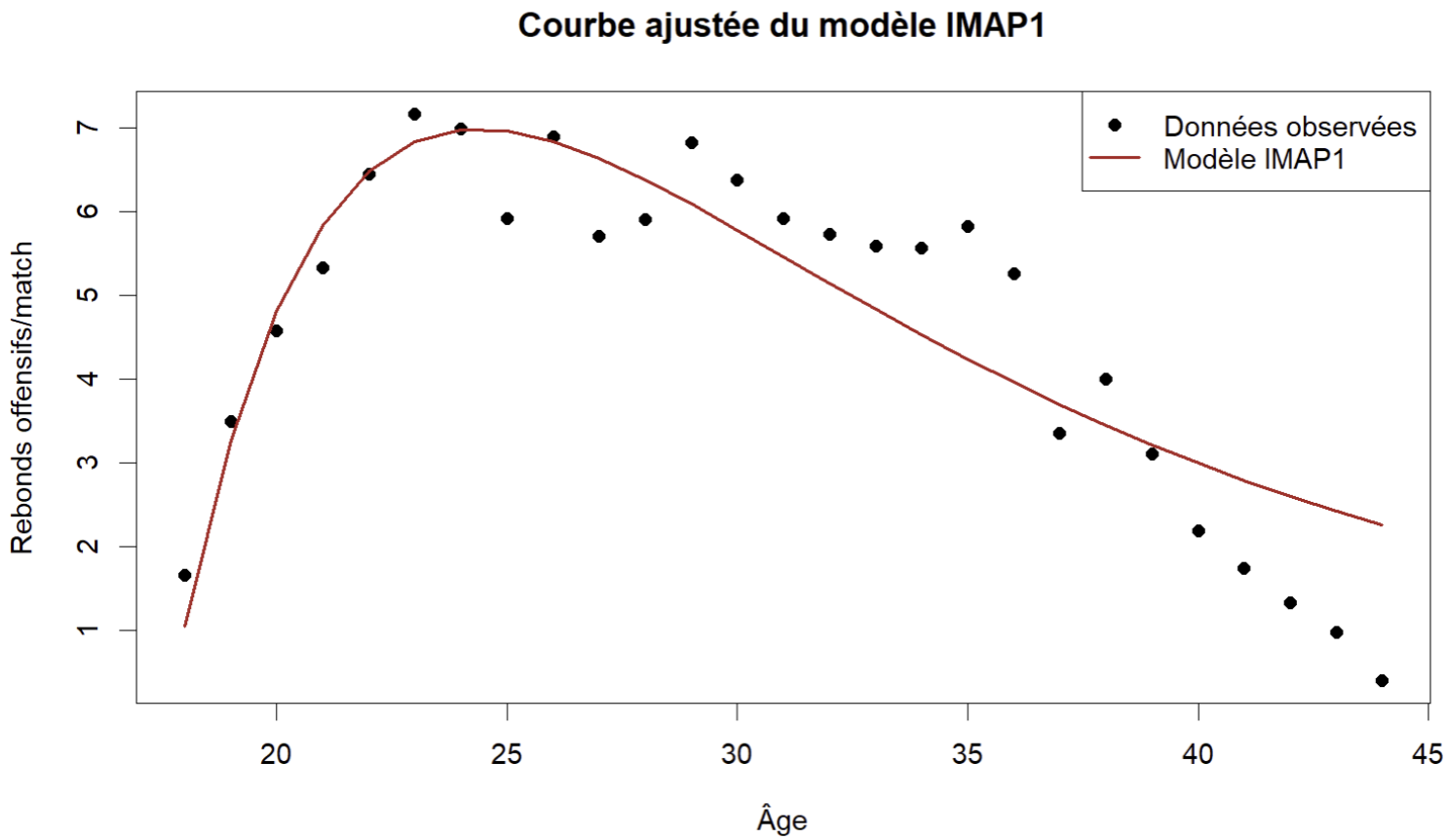


Ce graphique met en évidence une **tendance en cloche** : la performance s'améliore jusqu'à un certain âge avant de décroître progressivement.

On remarque un pic de performance situé autour de la vingtaine, suggérant un âge optimal pour cette statistique.

Ajustement d'un modèle non linéaire – Modèle IMAP1

Pour modéliser cette évolution, nous avons utilisé une fonction non linéaire appelée **IMAP1**, adaptée aux phénomènes biologiques où la performance croît puis décline. Les paramètres du modèle ont été estimés via la méthode MMC (Minimisation de la Somme des Carrés), en recherchant la meilleure configuration parmi 500 essais aléatoires.

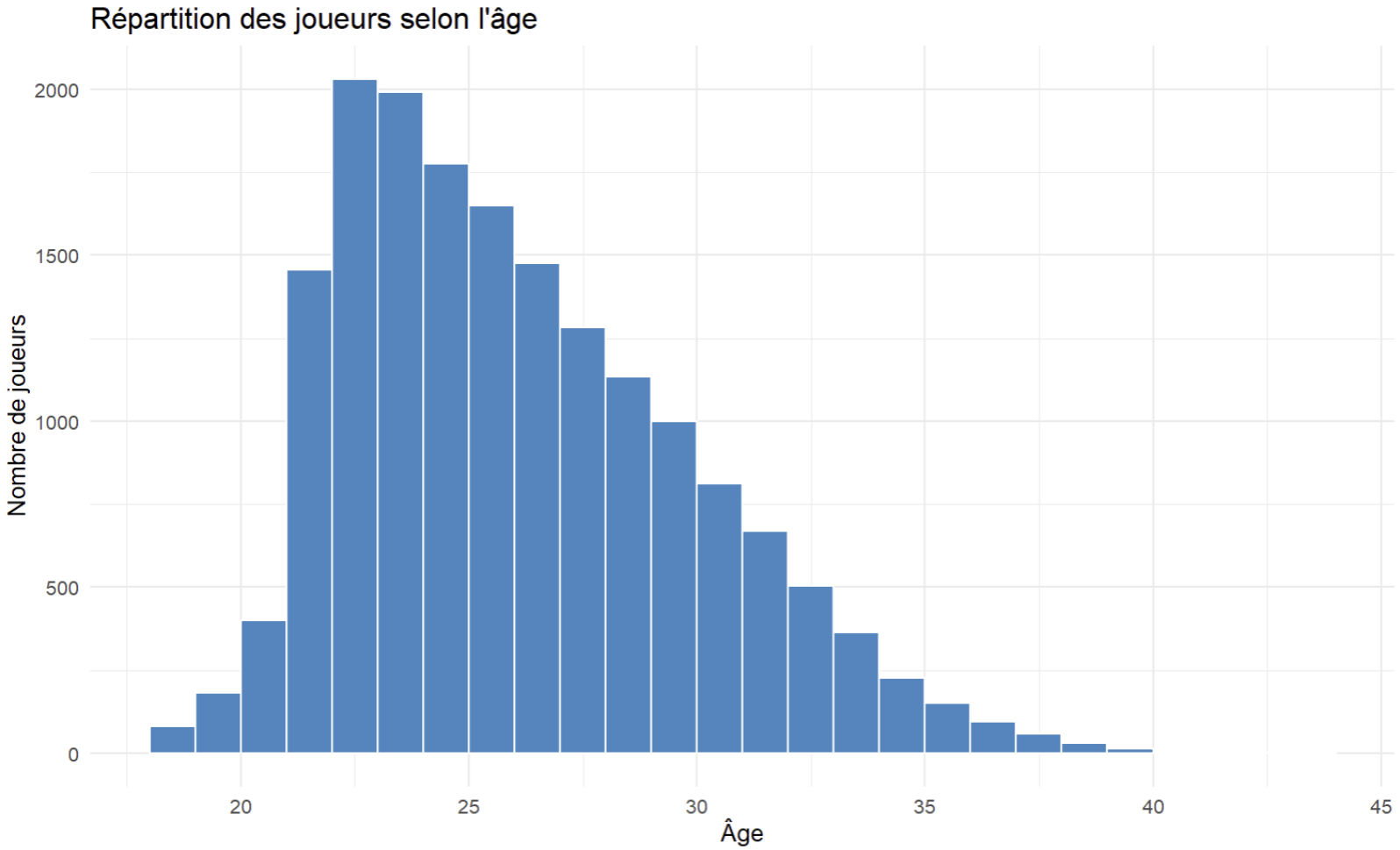


Le modèle IMAP1 s'ajuste très bien aux données observées, comme l'indique un **coefficient de détermination R^2 de 0,822**, ce qui montre une **bonne qualité de prédiction**.

Le **pic de performance** estimé par le modèle est atteint à l'**âge de 24 ans**, ce qui confirme visuellement l'observation du graphique précédent.

Distribution de l'âge des joueurs

Enfin, pour vérifier que l'évolution de la performance n'est pas biaisée par la répartition des joueurs selon leur âge, nous avons examiné l'histogramme de la population. On observe une concentration de joueurs entre **22 et 30 ans**, ce qui correspond à la tranche d'âge la plus représentée. Cela valide la robustesse de notre analyse, car les âges autour du pic de performance sont bien documentés dans l'échantillon.



Conclusion

L'analyse du nombre de rebonds offensifs par match en fonction de l'âge révèle une relation non linéaire forte, avec une performance maximale atteinte à 24 ans, puis une diminution progressive. Le modèle IMAP1 permet de modéliser avec précision cette évolution, et la bonne répartition des âges dans la base garantit la fiabilité des résultats.

Cette approche pourrait être étendue à d'autres indicateurs de performance (points, passes, temps de jeu) pour une vision plus globale du cycle de performance des joueurs professionnels.

Influence des variables biométriques

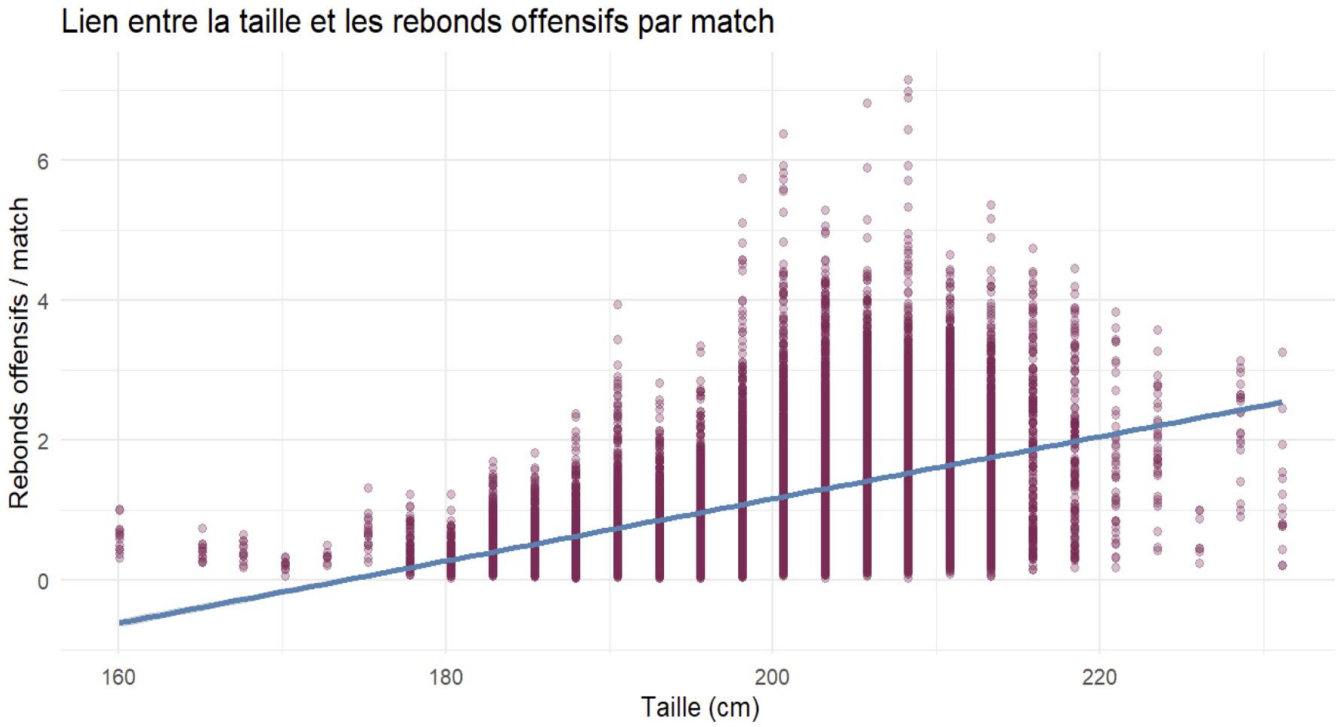
L'Influence des variables biométriques sur la performance

Dans cette partie, nous avons cherché à évaluer dans quelle mesure les caractéristiques biométriques des joueurs, notamment **la taille** et le **poids**, influencent une dimension clé de la performance physique : le **nombre de rebonds offensifs par match**.

Pour cela, nous avons utilisé deux modèles de régression linéaire simples, ajustés respectivement sur la taille et le poids des joueurs, en excluant les valeurs manquantes et aberrantes.

Taille et rebonds offensifs

Le premier graphique ci-dessous montre la relation entre la **taille des joueurs** (en cm) et leur **moyenne de rebonds offensifs par match** :



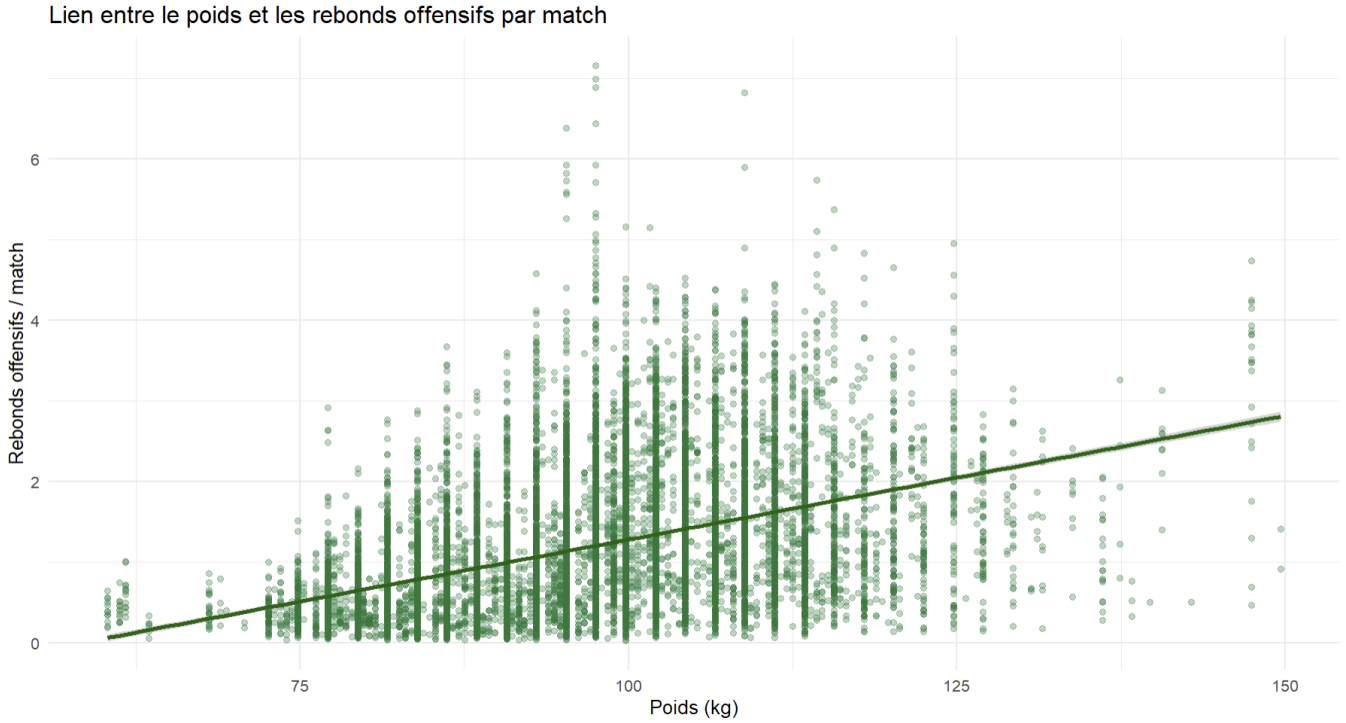
On observe une tendance globale **croissante** : les joueurs plus grands réalisent en moyenne plus de rebonds offensifs. Cette relation est **attendue** car la taille favorise naturellement la capacité à récupérer des ballons sous le panier.

Cependant, on note également une forte **dispersion des données** : certains joueurs de petite taille peuvent obtenir des scores similaires à ceux de grands joueurs. Cela suggère que d'autres facteurs, comme la détente, la technique ou le placement,

L'Influence des variables biométriques sur la performance

Poids et rebonds offensifs

Le second graphique analyse le **poids des joueurs** (en kg) en fonction du nombre de rebonds offensifs par match :



Ici aussi, la **tendance est globalement croissante** : les joueurs plus lourds ont tendance à capter davantage de rebonds offensifs. Cela peut s'expliquer par une meilleure stabilité ou un avantage physique dans les duels sous le panier.

Mais comme pour la taille, la **variabilité reste importante**. Des joueurs relativement légers peuvent aussi avoir de très bonnes performances en rebond offensif.

Pour conclure, les deux modèles ajustés montrent une **relation positive mais modérée** entre les variables biométriques et la performance. Cela signifie que ces facteurs **influencent partiellement** la capacité à faire des rebonds offensifs, mais ne sont pas les seuls à l'expliquer.

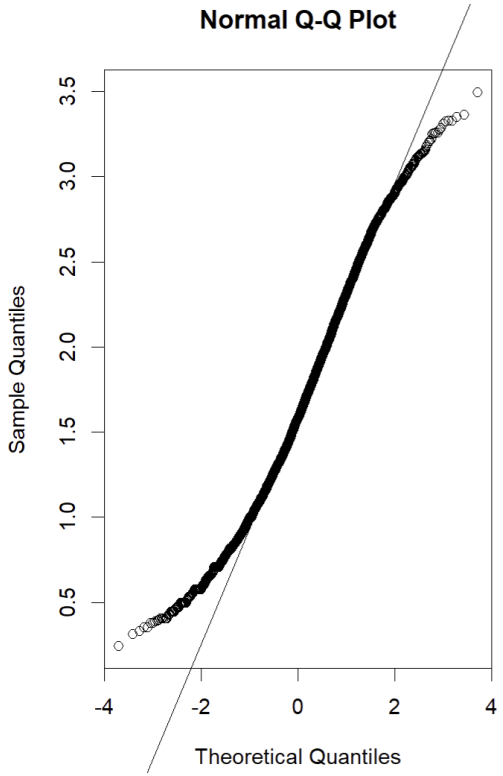
La taille semble avoir un **effet plus clair** que le poids. Toutefois, l'ajustement reste peu précis (faible R^2), ce qui est courant en sciences sociales et du sport, où les comportements sont multifactoriels.

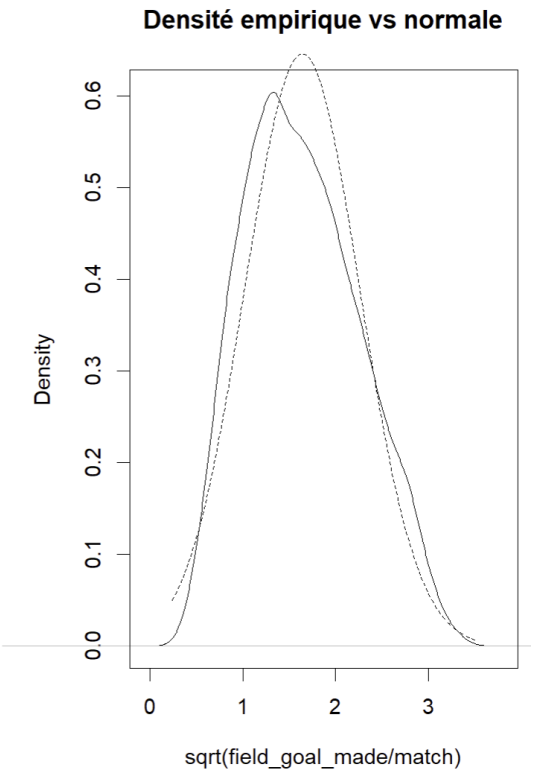
Différence entre les postes

Différences entre les postes

L'objectif de cette section est d'analyser les spécificités des différents postes au basket-ball (Arrière, Intérieur, Meneur) à partir de deux variables clés : le nombre de tirs réussis par match et les rebonds offensifs. Pour cela, deux approches complémentaires ont été mobilisées, une analyse de la variance (ANOVA) pour une variable quantitative, et un test du Khi² pour une variable catégorisée.

1.Comparaison du nombre de tirs réussis par match





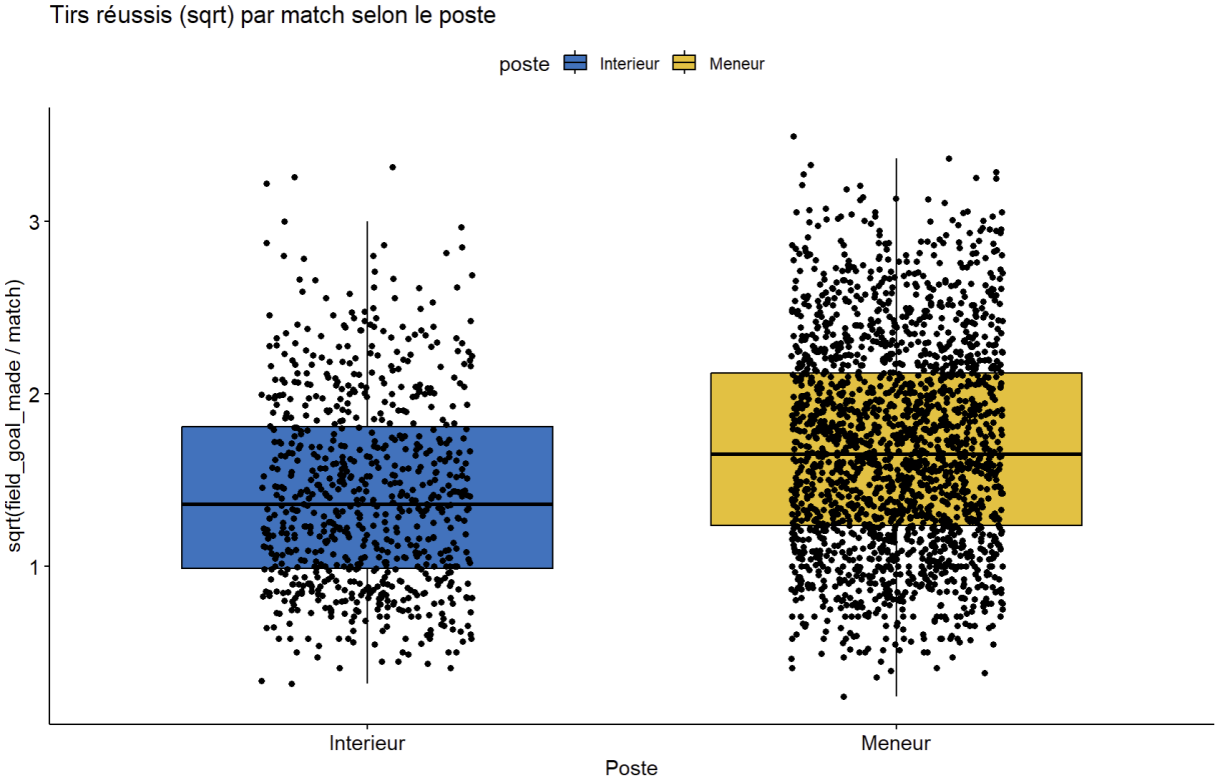
Afin d'évaluer la performance offensive des joueurs selon leur poste, la variable *field_goal_made per match* a été calculée, puis transformée par racine carrée afin d'améliorer la normalité de sa distribution.

Cette transformation a été validée par le **graphique "Normal Q-Q Plot"** et par le **graphique "Densité empirique vs normale"**, montrant une distribution globalement symétrique, ce qui permet l'utilisation d'une ANOVA.

Le **graphique "Tirs réussis (sqrt) par match selon le poste"** permet ensuite de comparer la distribution des tirs réussis (après transformation) selon les postes. On y observe une moyenne plus élevée chez les meneurs que chez les intérieurs, avec une dispersion relativement comparable. Les arrières ont été exclus de cette analyse en raison d'un effectif insuffisant.

2. Répartition des rebonds offensifs selon le poste

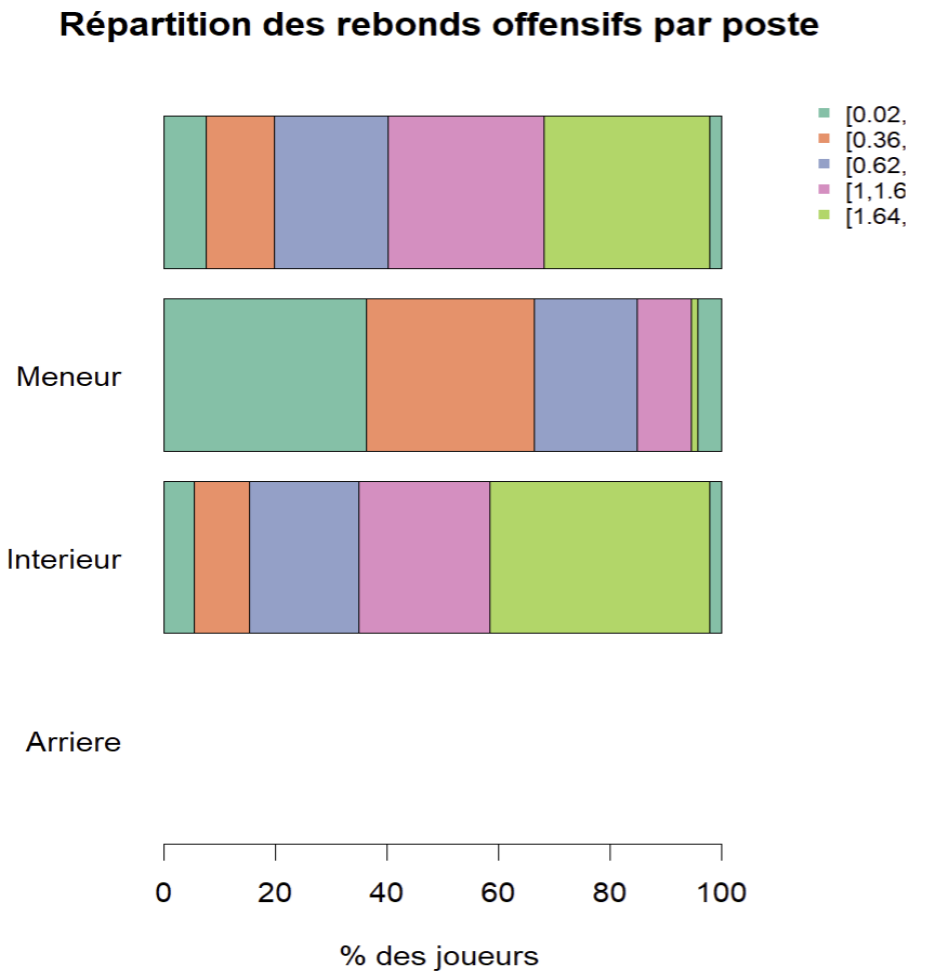
La variable "rebonds offensifs par match" a été discrétisée en cinq classes d'effectifs équivalents à l'aide de la fonction quant.cut. Cette catégorisation permet de comparer les profils de rebond selon les postes sur une base commune. La répartition est représentée dans le graphique **"Répartition des rebonds offensifs par poste"**, qui met en évidence des différences nettes entre les profils selon le poste :



Il ressort une nette spécialisation des postes :

- Les **intérieurs** sont largement représentés dans les classes les plus élevées de rebonds offensifs, ce qui reflète leur présence proche du panier.
- Les **meneurs**, au contraire, sont très présents dans les classes basses, ce qui s'explique par leur éloignement du cercle et leur rôle de gestion du jeu.
- Les **arrières** sont faiblement représentés dans les données, avec peu d'impact sur cette dimension du jeu.

Le test du Khi² (χ^2 = 1045.2, ddl = 4, p < 2.2e-16) révèle une **dépendance statistique significative** entre le poste et le niveau de rebonds offensifs. Autrement dit, le poste d'un joueur influence fortement sa capacité à prendre des rebonds. Cette conclusion est renforcée par la visualisation ci-dessous, où l'on observe des profils clairement différenciés selon le rôle du joueur.



Conclusion

Cette double analyse confirme que **le poste d'un joueur influence significativement ses caractéristiques de jeu**. Les **meneurs** se distinguent par une efficacité supérieure au tir, tandis que les **intérieurs** dominent dans le secteur du rebond offensif. Ces différences reflètent des rôles spécifiques dans l'organisation du jeu et justifient une approche différenciée dans l'analyse des performances sportives.

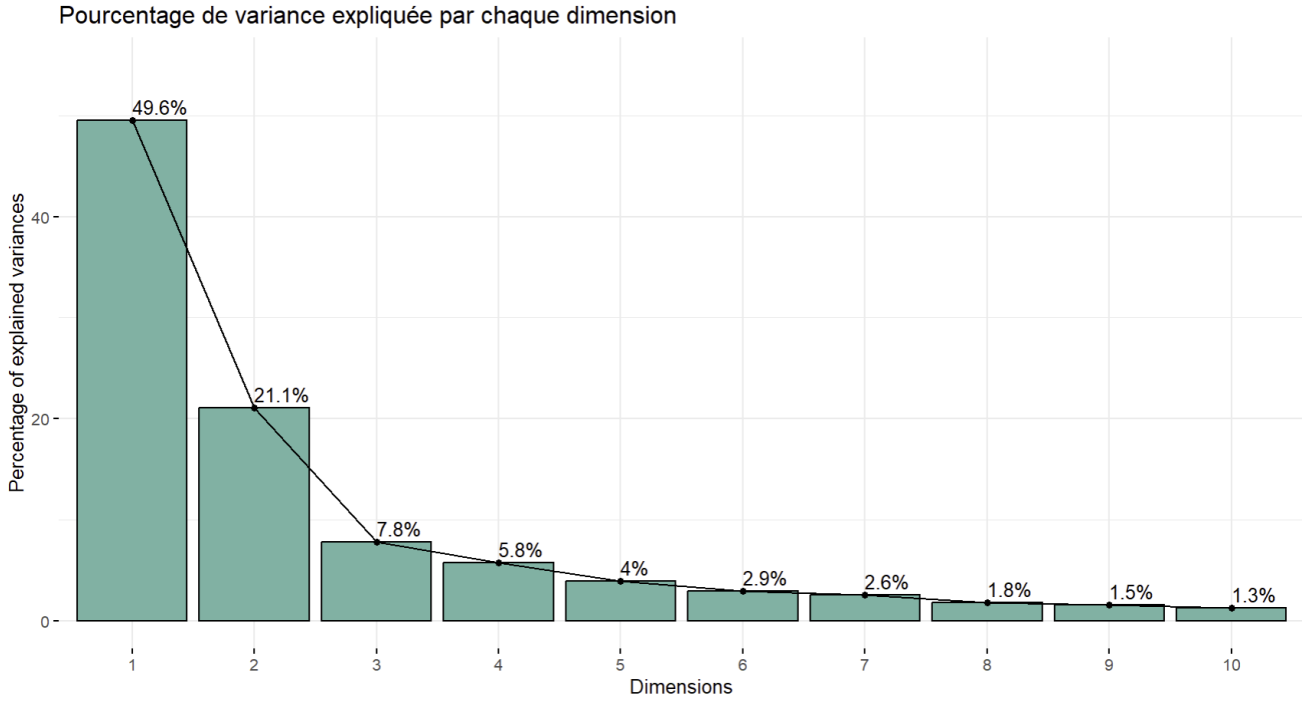
L'analyse reste cependant limitée par l'absence de certaines variables contextuelles (niveau du championnat, âge, temps de jeu), et l'exclusion partielle de certains postes (arrière) pourrait biaiser la généralisation des résultats.

Analyse factorielle

Analyse factorielle

L'objectif de cette section est d'obtenir une vue d'ensemble des profils de joueurs à partir de toutes les variables disponibles, à travers une **analyse en composantes principales (ACP)**. Cette méthode permet de réduire la dimensionnalité du jeu de données tout en conservant l'essentiel de l'information, et d'identifier les principales structures sous-jacentes aux variables.

1. Pourcentage de variance expliquée par chaque dimension



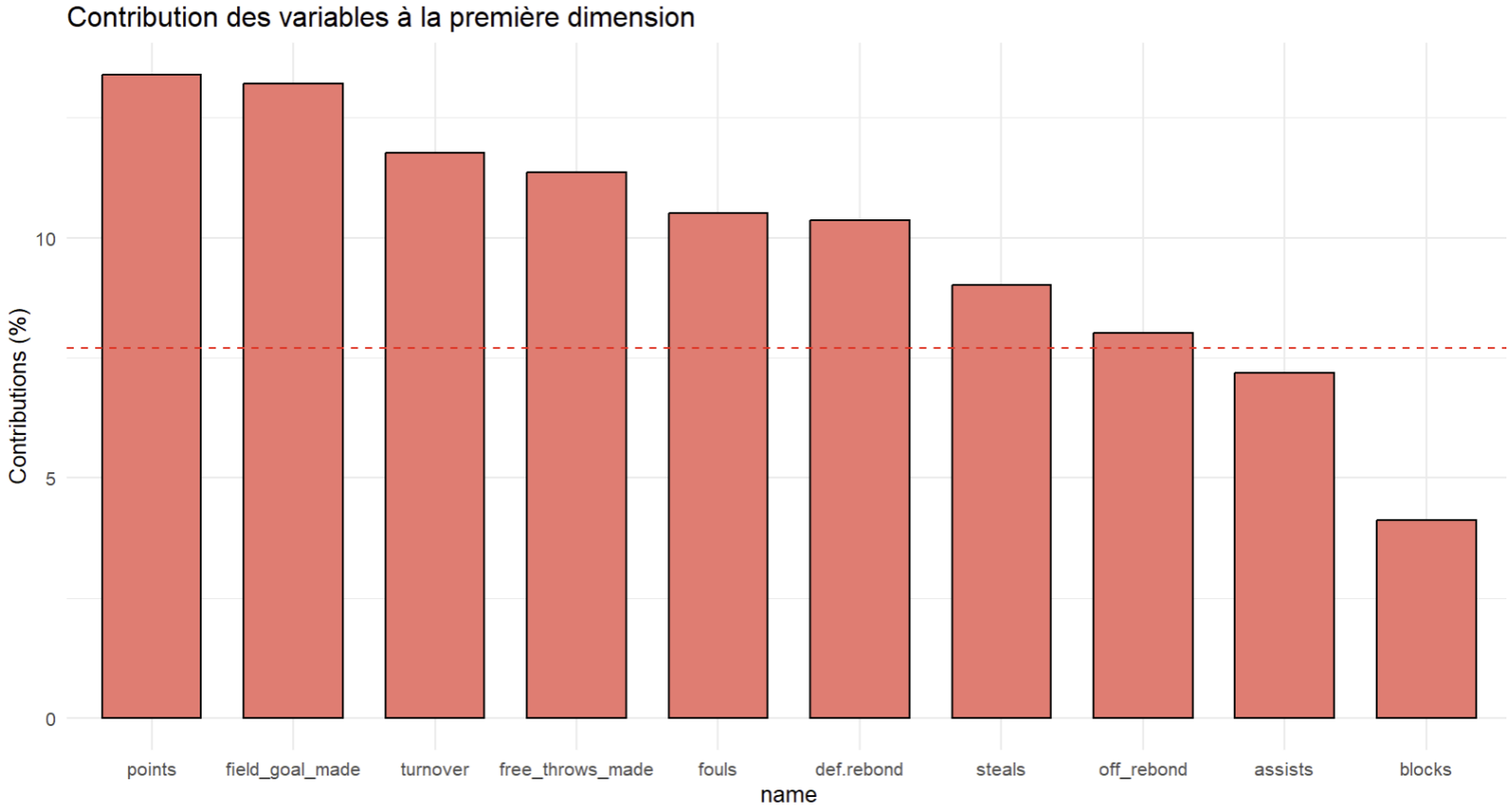
Le **graphique du pourcentage de variance expliquée** permet d'évaluer combien d'information est capturée par chaque axe principal.

La **première dimension (Dim 1)** explique environ **49,6 %** de l'inertie totale, tandis que la **deuxième dimension (Dim 2)** en explique **21,1 %**, soit **70,7 % cumulés sur les deux premières dimensions**.

Cela signifie qu'une grande partie de la variabilité du jeu de données peut être interprétée à l'aide de ces deux axes.

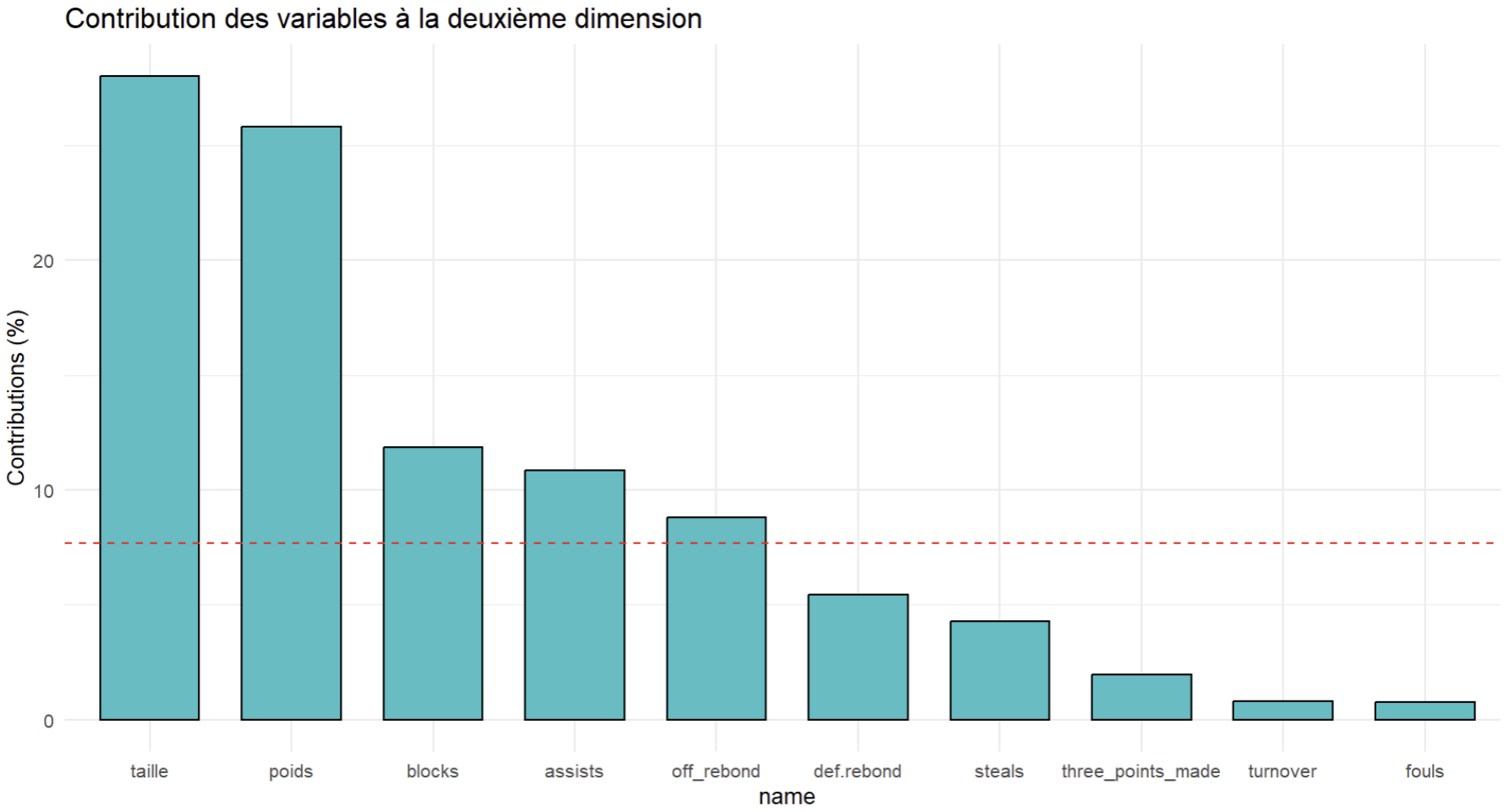
2. Contribution des variables à la première dimension

Le graphique des contributions à la première dimension met en évidence les variables les plus discriminantes. On observe que les variables points, field_goal_made, turnover, free_throws_made et fouls sont les plus contributives. Ces variables sont liées à l'efficacité offensive et à la participation active dans le jeu, ce qui suggère que la Dim 1 reflète un axe de performance individuelle offensive.

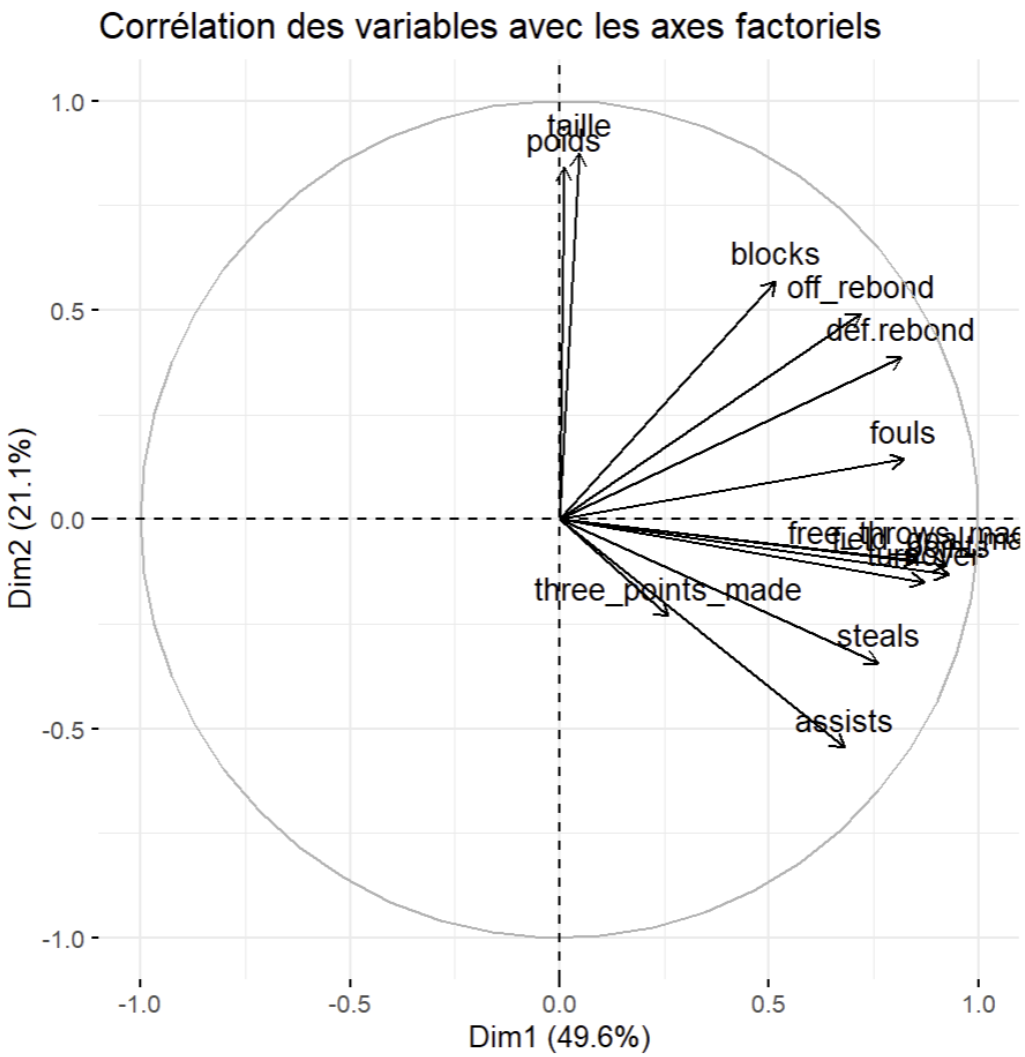


3. Contribution des variables à la deuxième dimension

En ce qui concerne la deuxième dimension, les variables taille et poids sont de loin les plus contributives, suivies de blocks et assists. Cette dimension semble donc refléter davantage des caractéristiques physiques et défensives. La Dim 2 oppose donc des profils de joueurs plus massifs, présents dans la raquette, à des profils plus légers et mobiles.



4. Corrélation des variables avec les axes factoriels



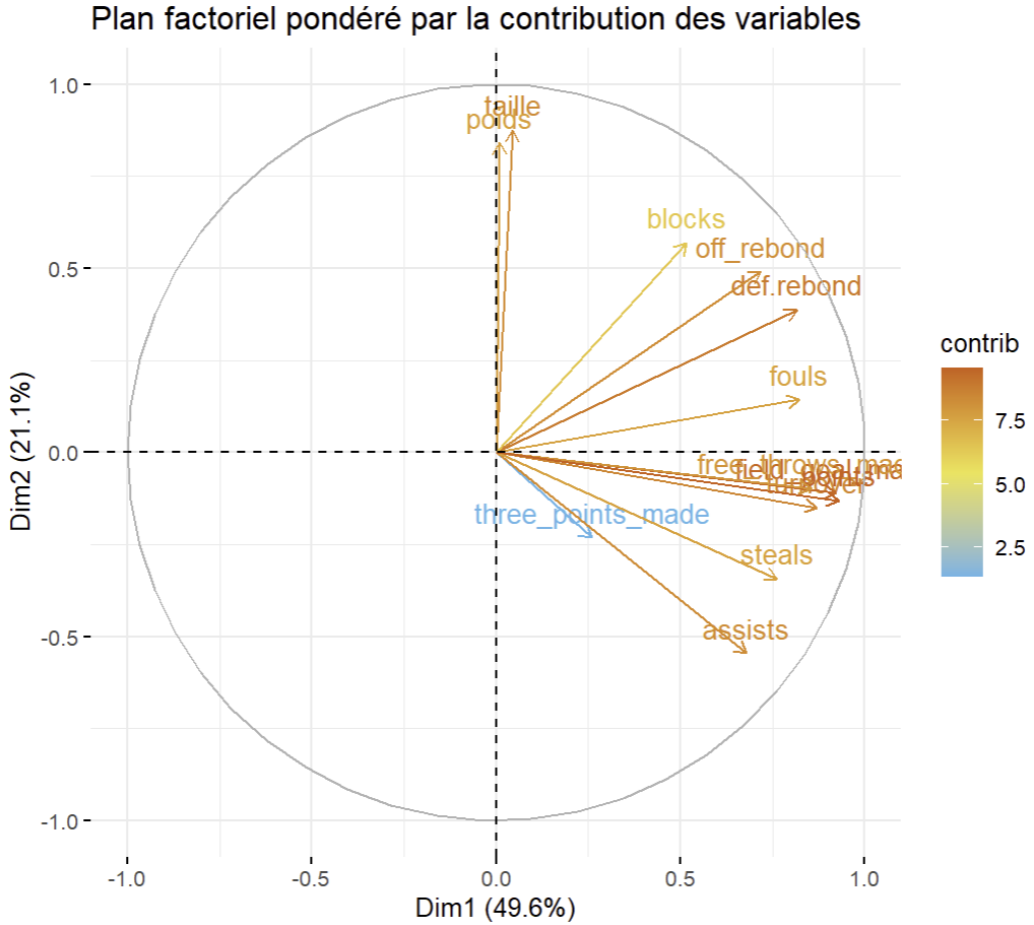
Le **cercle des corrélations** permet de visualiser l'association des variables avec les deux premiers axes. On y observe clairement deux groupes de variables :

- Un groupe **physique et défensif** fortement corrélé à la Dim 2 : **taille, poids, blocks, rebonds** ;
- Un groupe **offensif** corrélé à la Dim 1 : **points, tirs réussis, fautes, passes décisives, pertes de balle**.

La position des flèches proches du cercle et leur alignement sur les axes indiquent une bonne qualité

5. Plan factoriel pondéré par la contribution des variables

Ce graphique pondéré par la contribution des variables confirme les observations précédentes : les variables taille et poids tirent fortement la deuxième dimension, tandis que points, field_goal_made ou free_throws_made influencent davantage la première. L'échelle de couleur met en évidence les variables les plus déterminantes dans la structuration des données.



Pour conclure l'analyse factorielle a permis de dégager deux axes principaux :

- **La première dimension** traduit une **performance offensive**, intégrant les points marqués, la réussite au tir et les fautes subies.
- **La deuxième dimension** correspond à un **profil physique et défensif**, mettant en avant la taille, le poids, et les actions défensives telles que les contres et les rebonds.

Cette représentation globale permet donc d'identifier des profils types de joueurs selon leurs caractéristiques statistiques, et de mieux comprendre les axes qui structurent la diversité des performances.

CONCLUSION GENERALE

Dans ce rapport, nous avons cherché à mieux comprendre ce qui influence la performance des joueurs de basket-ball, à travers différentes étapes d'analyse la description des données, exploration des liens entre variables, modélisation et visualisation, ainssi nous avons pu dégager des tendances claires et tirer des conclusions intéressantes.

Nous avons d'abord pris le temps de bien comprendre les variables à notre disposition, en distinguant celles qui reflètent directement la performance (comme les points, rebonds, passes décisives) de celles qui pourraient l'expliquer (âge, poste, taille, etc.). L'analyse de la performance en fonction de l'âge a mis en évidence une courbe non linéaire, avec un pic de performance autour de 24 ans, confirmé par un bon ajustement du modèle IMAP1 ($R^2 = 0,822$).

L'étude des caractéristiques biométriques a montré que la taille et le poids peuvent jouer un rôle, notamment pour les rebonds offensifs, mais qu'ils ne suffisent pas à expliquer toutes les différences de performance entre les joueurs. Le poste occupé reste un facteur clé : chaque poste présente un profil de performance bien spécifique, comme l'ont montré les analyses croisées et les visualisations.

Enfin, l'analyse factorielle a permis d'avoir une vue d'ensemble du jeu de données et de mieux comprendre comment les différentes dimensions de la performance se structurent entre elles. On distingue notamment un axe offensif (liée au scoring) et un axe plutôt défensif ou physique (rebonds, interceptions, fautes, etc.).

En résumé, la performance d'un joueur de basket-ball dépend de plusieurs facteurs qui interagissent entre eux : l'âge, les capacités physiques, le rôle sur le terrain et probablement aussi des aspects moins visibles comme l'expérience ou la stratégie de jeu. Cette étude ouvre donc la voie à des analyses plus fines, voire à une modélisation plus complète incluant des données longitudinales ou comportementales.