STAT-S401 : Analyse des performances des joueurs de football dans les cinq grands championnats européens

Travail réalisé par :

Aerts Robin
Bayala Darren
Ben Hamou Assia
Laloy Anouar
Tfeili Rayane

Encadré par :

Catherine Dehon

Université libre de Bruxelles

Année académique 2024–2025

Table des matières

1 Introduction **2**

2 Description de la base de données **2**

3 Description de la question de recherche **4**

4 Brève revue de littérature **4**

5 Analyses descriptives univariées et bivariées **5**

5.1 Analyses descriptives univariées 5

5.2 Analyses descriptives bivariées 6

6 Détection des valeurs aberrantes **7**

7 Analyse en Composantes Principales "Robuste" (ACP) **9**

7.0.1 Pré-traitement 9

8 Analyse Factorielle des Correspondances Binaires (ACOBI) **14**

8.1 Description des variables 14

8.2 Calculs Préliminaires 15

8.2.1 Table de contingence 15

8.2.2 Test d'indépendance 15

8.2.3 Matrice d'attraction/répulsion 16

8.3 Interprétation 16

9 Régression Robuste **17**

10 Limites de l'étude **18**

11 Conclusion **19**

12 Annexes **20**

12.1 Résumé statistique pour la détection des valeurs aberantes 20

12.2 Code R 21

1 Introduction

Nous vivons dans un monde où la data-driven est de plus en plus présent, et en particulier, le football n'y échappe pas . Aujourd'hui, de nombreux clubs intègrent désormais l'analyse de données dans leur stratégie, à l'image d'Arsenal, l'un des meilleurs clubs de Premier League, qui recrute des data scientists pour renforcer son département analytique. Face à cette montée en puissance de l'approche analytique, une nouvelle opportunité s'ouvre : mieux comprendre les profils des joueurs qui composent les effectifs des grands clubs européens. Ces profils peuvent être décrits à travers plusieurs caractéristiques : leur poste, leur nationalité, leurs performances, ou encore leur âge.

En Europe, les cinq grands championnats - la Premier League en Angleterre, la Liga en Espagne, la Serie A en Italie, la Bundesliga en Allemagne et la Ligue 1 en France — sont considérés comme les plus compétitifs et les plus médiatisés. Chacun de ces championnats possède ses propres caractéristiques : intensité du jeu, exigences physiques, style tactique ou encore rythme des matchs. Ces spécificités pourraient avoir une influence sur le type de joueur que chaque championnat attire ou valorise, notamment en termes d'âge.

C'est dans ce contexte que nous avons choisi d'analyser le profil d'âge dans les cinq ligues européennes . Dans un premier temps, nous présenterons plus en détail cette base de données afin d'en préciser le contenu et les limites. Nous décrirons ensuite la question de recherche de notre analyse statistique et ferons une brève revue littéraire pour voir où l'on peut se positionner par rapport à d'anciens travaux de recherches déjà réalisé sur le sujet d'intérêt. Nous proposerons par la suite une série d'analyses descriptives, d'abord univariées pour observer la distribution des principales variables et ensuite bivariées afin d'identifier d'éventuelles corrélations ou tendances entre les variables d'intérêt et par après appliquerons des méthodes d'analyses statistiques multivariées afin de répondre à notre question de recherche. Enfin, nous conclurons et répondrons à la question de recherche suivant les résultats obtenus.

2 Description de la base de données

Notre travail se base sur une base de données [1] comprenant 2 852 athlètes évoluant durant la saison 2023/2024 dans les cinq grands championnats européens : la Premier League d'Angleterre, la Bundesliga d'Allemagne, la Ligue 1 de France, la Serie A d'Italie et La Liga d'Espagne. Il est important de noter que l'on peut retrouver des doublons étant donnée qu'un joueur peut jouer dans deux championnats différents durant la période d'étude.
Notre ensemble reprend 37 variables tant qualitatives que quantitatives telles que l'âge, la nationalité, le poste, l'équipe, et le championnat dans lequel ils évoluent. De plus, on trouve des statistiques de performance, tant classiques (buts, passes décisives, minutes jouées, cartons) que plus avancées (buts attendus xG, passes décisives attendues xAG, et

1. Orkun Aktas, *All Football Players Stats in Top 5 Leagues 23/24*, Kaggle.

buts attendus hors pénalty npxG. Les statistiques "xG" et "xAG" permettent d'évaluer la qualité des occasions de buts et des passes décisives, indépendamment du résultat final). Vous pouvez retrouver ci-dessous un tableau regroupant l'ensemble des variables présentes dans la base de données.

Variable	Description (en français)
Player	Le nom du joueur.
Nation	La nationalité du joueur.
Pos	La position du joueur (par exemple, attaquant, milieu de terrain, défenseur).
Age	L'âge du joueur.
MP (Matches Played)	Le nombre total de matchs joués.
Starts	Le nombre de matchs dans lesquels le joueur a commencé.
Min (Minutes)	Le nombre total de minutes jouées (peut être le même que MP).
90s (90 Minutes Played)	L'équivalent en matchs de 90 minutes joués (par exemple, 1.5 = 135 minutes).
Gls (Goals)	Le nombre total de buts marqués par le joueur.
Ast (Assists)	Le nombre total de passes décisives effectuées par le joueur.
G+A (Goals + Assists)	Le nombre total de buts et passes décisives combinés.
G-PK (Goals - Penalty Kicks)	Le nombre total de buts marqués, excluant les pénaltys.
PK (Penalty Kicks)	Le nombre de buts marqués sur pénalty.
PKatt (Penalty Kicks Attempted)	Le nombre de pénaltys tentés par le joueur.
CrdY (Yellow Cards)	Le nombre de cartons jaunes reçus par le joueur.
CrdR (Red Cards)	Le nombre de cartons rouges reçus par le joueur.
xG (Expected Goals)	Le nombre de buts attendus à partir des tirs du joueur.
npxG (Non-Penalty Expected Goals)	Le nombre de buts attendus, excluant les pénaltys.
xAG (Expected Assists)	Le nombre de passes décisives attendues à partir des passes du joueur.
npxG+xAG (Non-Penalty xG + xAG)	La somme des buts attendus sans pénaltys et des passes décisives attendues.
PrgC (Progressive Carries)	Le nombre de fois où le joueur a porté le ballon vers l'avant.
PrgP (Progressive Passes)	Le nombre de passes effectuées par le joueur qui ont déplacé le ballon vers l'avant.
PrgR (Progressive Runs)	Le nombre de fois où le joueur a effectué des courses vers l'avant avec le ballon.
Gls (Goals)	(Répété) Le nombre total de buts marqués par le joueur.
Ast (Assists)	(Répété) Le nombre total de passes décisives effectuées par le joueur.
G+A (Goals + Assists)	(Répété) Le nombre total de buts et passes décisives combinés.
G-PK (Goals - Penalty Kicks)	(Répété) Le nombre total de buts marqués, excluant les pénaltys.
G+A-PK (Goals + Assists - Penalty Kicks)	Le total des buts et passes décisives, moins les buts sur pénalty.
xG (Expected Goals)	(Répété) Le nombre de buts attendus à partir des tirs du joueur.
xAG (Expected Assists)	(Répété) Le nombre de passes décisives attendues à partir des passes du joueur.
xG+xAG (Expected Goals + Expected Assists)	Le total des buts attendus et des passes décisives attendues.
npxG (Non-Penalty Expected Goals)	(Répété) Le nombre de buts attendus, excluant les pénaltys.
npxG+xAG (Non-Penalty xG + Expected Assists)	La somme des buts attendus sans pénaltys et des passes décisives attendues.

TABLE 1 – Description des variables de la base de données

Sur base de vérification, nous avons constaté que certaines informations étaient manquantes. C'est dans ce cadre que nous avons décidé de compléter les informations suivantes, afin de garantir la fiabilité de nos analyses :

- Les joueurs Marco Pellegrino, Max Moerstedt et Max Svensson n'avaient pas de date de naissance dans la base. On leur a attribué respectivement 2002, 2006 et 2001 comme année de naissance, ce qui correspond à un âge de 21, 17 et 21 ans.
- La nationalité de deux joueurs a également été complétée manuellement (ex : "ar ARG" pour Pellegrino, "tr TUR" pour Kücüksahin).

Ce nettoyage permet de limiter la perte d'information lorsqu'on supprimera les valeurs manquantes par la suite.

Pour mieux capturer chaque profil d'âge par ligue, nous avons décidé de regrouper les joueurs par tranches d'âge, plutôt que de les comparer individuellement. Cela rend l'analyse plus fluide et permet d'observer des tendances générales plutôt que des comportements spécifiques. Ainsi, une nouvelle variable qualitative, appelée *cat_age*, a été ajoutée. Elle classe les joueurs en quatre catégories d'âge :

- **Jeune** ($\leq$ 21ans) : avant 21 ans, les joueurs sont considérés comme des talents en développement, en phase post-formation ou d'intégration dans le monde pro-

fessionnel ;
— **Moyen** (22 à 26 ans) : les joueurs commencent à atteindre un certain niveau physique et tactique ;
— **Expérimenté** (27 à 31 ans) : cette tranche correspond souvent au pic de la carrière pour un joueur offensif ;
— **Âgé** ($\geq$ 32 ans) : au-delà de 30 ans, le déclin physique commence à se faire sentir, et les profils évoluent souvent vers des rôles plus tactiques ou physiquement moins intenses.

3 Description de la question de recherche

Dans ce travail, nous cherchons à comprendre comment des variables tels que l'âge et le championnat dans lequel évolue un joueur peuvent influencer ses performances offensives.

On le sait tous : l'âge des sportifs est un facteur crucial dans les sports de haut niveau, notamment le football. Parmi les sportifs, très peu d'athlètes arrivent à conserver un haut niveau de performance jusqu'à un âge avancé, à cause du poids physique qu'ils supportent.

De plus , les exigences physiques et tactiques ne sont pas les mêmes en Premier League, en Liga ou en Serie A, et cela peut influencer la réussite ou non de certains profils de joueurs. En effet, La Premier League est souvent connue pour son intensité physique et son rythme soutenu, tandis que la Liga accorde davantage d'importance à la possession de balle et à la créativité offensive. De son côté, la Serie A se distingue par sa rigueur tactique et sa densité défensive.

Ainsi, en combinant ces deux dimensions , il devient alors intéressant de se demander comment l'âge influence les performances individuelles des athlètes et plus particulièrement les performances offensives selon le championnat dans lequel il évolue

C'est donc dans cette optique que nous avons orienté notre étude : observer comment l'âge influence les performances offensives des joueurs selon le championnat dans lequel ceux-ci évoluent. Tout en gardant un cadre structuré, nous restons ouverts à l'exploration d'autres relations pertinentes entre les variables, afin d'enrichir la compréhension globale des facteurs qui déterminent la réussite des joueurs au plus haut niveau

4 Brève revue de littérature

Une étude [2] nous montre déjà les différentes tendances en termes d'âge des joueurs dans les différentes ligues européennes. Parmi les cinq championnats, la Ligue 1 se distingue comme la plus jeune. La Bundesliga suit une tendance similaire. En Liga, l'âge moyen

2. CIES Football Observatory, *Répartition des minutes par âge – Rapport hebdomadaire n°349*, 2021

reste assez équilibré même si certains clubs misent aussi sur les jeunes joueurs. La Serie A, quant à elle, est composée de joueurs plus âgés occupant des postes clés et donc reste la ligue la plus expérimentée des cinq. Enfin, la Premier League se situe entre deux, avec une forte intensité physique nécessitant un bon équilibre entre jeunesse et expérience.

Une autre étude a analysé les performances de 154 joueurs de LaLiga entre les saisons 2012-2013 et 2019-2020. Les résultats ont montré que les joueurs âgés de 30 ans et plus réduisaient leur distance totale parcourue et leur nombre d'efforts à haute intensité avec l'âge. Cependant, leur précision de passe augmentait de manière significative, compensant partiellement la baisse des performances physiques [3]

5 Analyses descriptives univariées et bivariées

5.1 Analyses descriptives univariées

Une première analyse peut être faite sur deux de nos variables d'intérêt dans notre base de données : la catégorie d'âge et le championnat. Dans le premier graphique ci-dessous, on peut voir la répartition de nos joueurs présents dans notre base de données dans tous les championnats.

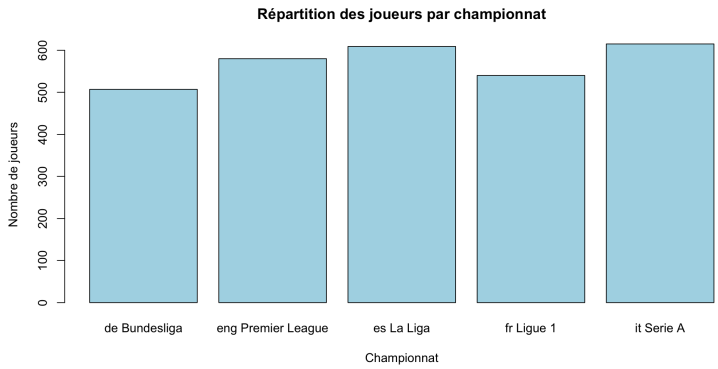

FIGURE 1 – Répartition des joueurs par championnat

On observe une répartition plutôt équilibrée où chaque championnat accueille une part équivalente de l'effectif global. Les championnats qui accueillent le plus de joueurs en Europe sont la Liga et la Serie A avec respectivement 21.4% et 21.6%. La Bundesliga arrive en dernier avec 17.8% de l'effectif globale.

La répartition par catégorie d'âge est un peu plus nuancé : on voit que la catégorie d'âge moyen (22-26 ans) est la plus representée en Europe avec 39.1% de l'effectif globale. Par contre, on n'observe pas énormément de joueurs agés dans les 5 championnats : il

3. Bollier AC et al., *Association Between Professional Football Participation and Long-term Health Outcomes*, PubMed, 2022

ne représente que 9.5% de la totalité des joueurs présents.

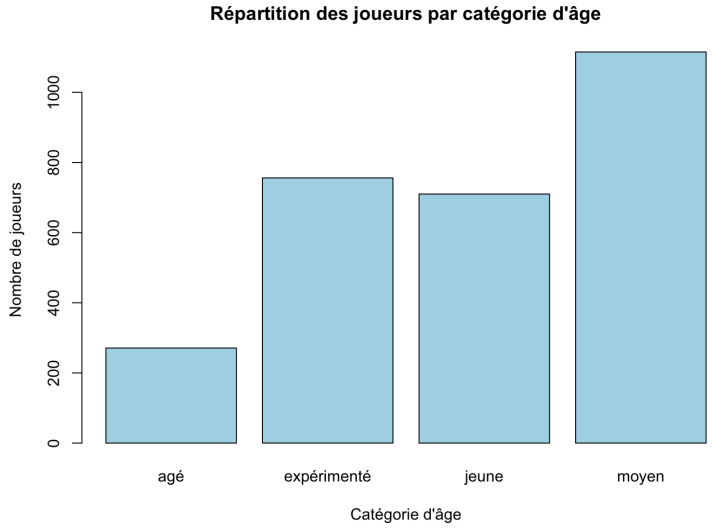

FIGURE 2 – Répartition des catégories d'âges parmi nos joueurs

5.2 Analyses descriptives bivariées

Etant donné que l'on s'intéresse aux liens entre les performances offensives et l'âge, il est intéressant d'étudier la répartition des buts par catégorie d'âge

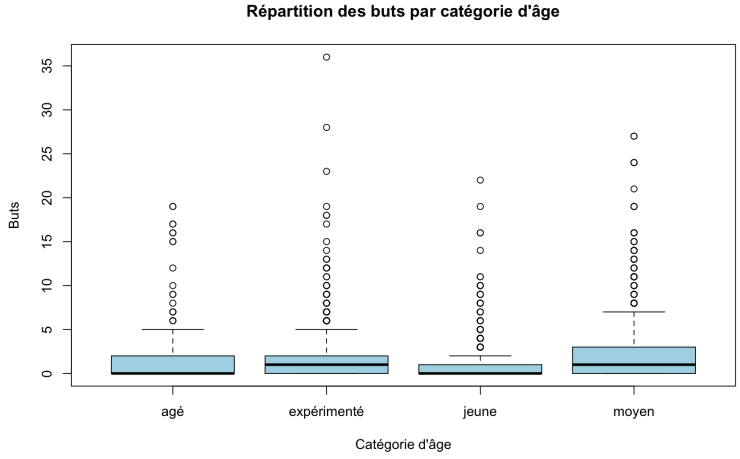

FIGURE 3 – Répartition des buts par catégorie d'âge

Le boxplot ci-dessus représente la distribution du nombre de buts inscrits en fonction de quatre tranches d'âge : jeune, moyen, expérimenté et âgé. Les catégories « expérimenté » et « moyen » montrent une plus grande variabilité, indiquant que ces catégories marquent le plus régulièrement En revanche, les « jeunes » et les « âgés » montrent des répartitions plus concentrées autour de faibles valeurs. Au final, on observe des valeurs

aberrantes (outliers) dans chaque catégorie, ce qui illustre l'influence de certains joueurs aux performances exceptionnelles pour chaque tranche d'âge.

La table suivante illustre la distribution des joueurs en fonction de leur tranche d'âge dans les cinq principaux championnats européens : la Bundesliga, la Premier League, La Liga, la Ligue 1 et la Serie A. Il est observable que la plus grande partie des joueurs dans chaque ligue se classe dans la catégorie expérimentée, suivie de près par celle moyenne. Les groupes jeunes et agés constituent un pourcentage plus petit de la population totale.

Quant à la distribution spécifique par ligue, la Serie A et la Premier League semblent recevoir un nombre assez important de jeunes joueurs, alors que la Bundesliga et la Ligue 1 présentent une proportion plus importante de joueurs considérés comme intermédiaires. En ce qui concerne La Liga, elle se caractérise par une distribution assez équilibrée entre les diverses catégories.

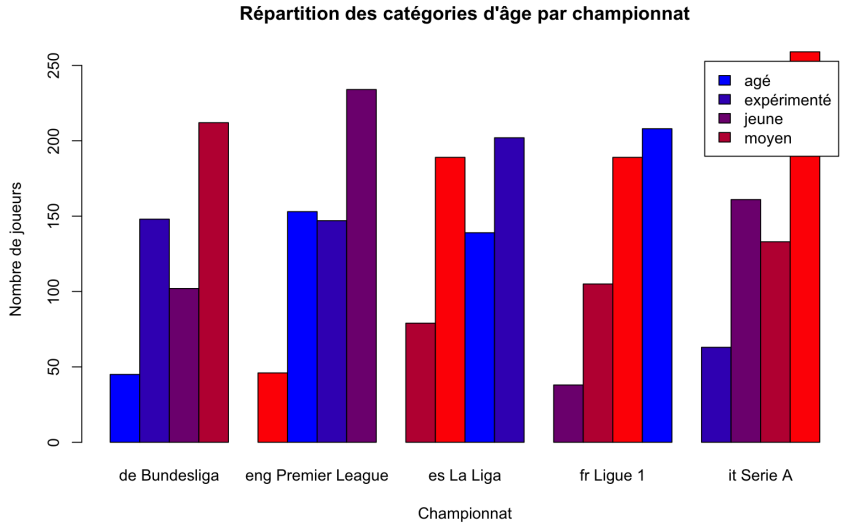

FIGURE 4 – Répartition des catégories d'âges parmi les championnats

6 Détection des valeurs aberrantes

Avant de commencer notre analyse multivariée, il est important de détecter les valeurs aberrantes présentes dans notre base de données. Dans un premier temps, nous avons filtré nos données afin d'avoir des données pertinentes à cette analyse. Nous nous sommes concentrés uniquement sur les joueurs ayant joué au moins 90 minutes sur la saison et ayant le poste d'attaquant ou de milieu de terrain ou les deux.

Après avoir examiné le résumé statistique de notre nouvel ensemble de données (cfr. annexe 1), on observe des écarts significatifs entre la médiane et la moyenne des variables *Gls* (buts) et *Ast* (passes décisives). Ces différences reflètent potentiellement la présence de valeurs aberrantes comme le montre les deux boxplots ci-dessous.

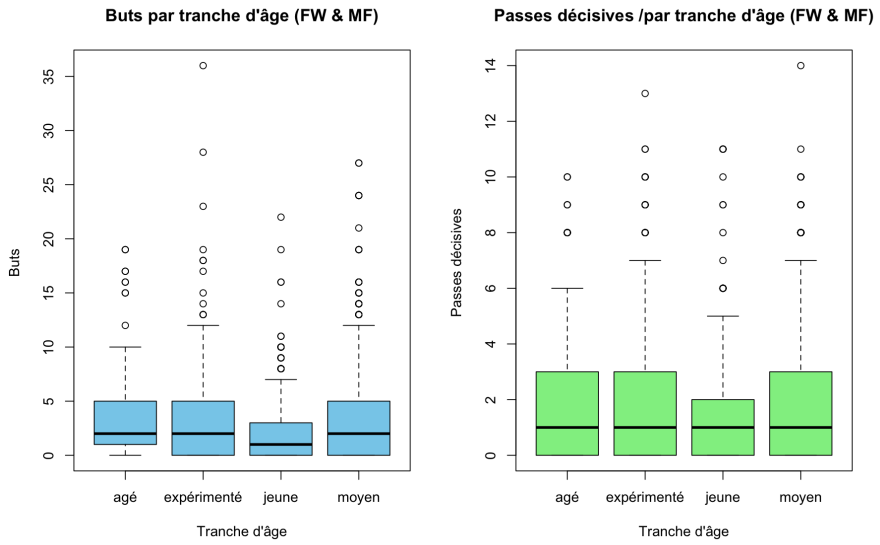

FIGURE 5 – Répartition des buts et passes décisives par catégorie d'âge

Néanmoins, ces analyses bivariées ne permettent pas de repérer toutes les valeurs aberrantes, puisque les variables interagissent entre elles. Par conséquent, afin de considérer ces interactions et de repérer toutes les valeurs aberrantes, nous avons utlisé la Distance de Mahalanobis robuste (MCD). Ici, le seuil fixé est la valeur du quantile à 97.5% de la distribution du χ^2 avec n degrés de liberté, où n représente le nombre de variables.

Pour mieux visualiser et analyser les résultats, nous avons tracé un graphique où chaque observation est représentée par un point :

— Les points rouges correspondent aux valeurs aberrantes (la distance est supérieure au seuil défini)
— Les points bleus correspondent aux observations considérées comme normales.

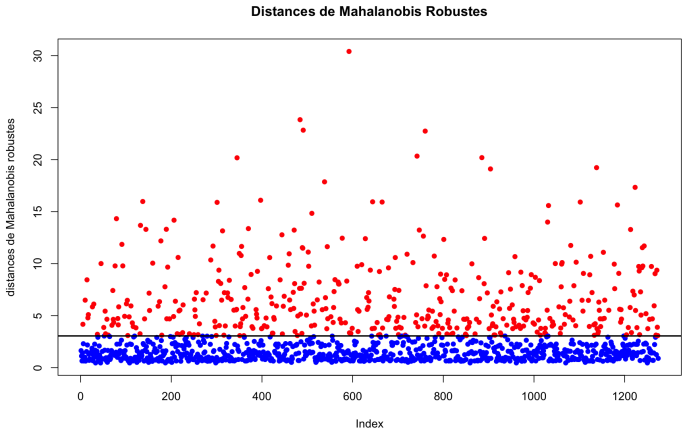

FIGURE 6 – Distance de Mahalanobis Robuste

On en conclut que notre ensemble de données filtrées comporte énormément de valeurs aberrantes. C'est pourquoi, dans la suite de ce travail, nous sommes obligés d'aborder

des méthodes robustes afin d'avoir des résultats concluants concernant notre question de recherche.

7 Analyse en Composantes Principales "Robuste" (ACP)

7.0.1 Pré-traitement

Avant de procéder à l'ACP, les variables sont : **Centrées** (moyenne nulle) et **Réduites** (variance unitaire), afin de neutraliser l'effet des unités de mesure.
De plus, nous ne sélectionnons qu'un nombre limité de variables de notre base de données afin de plus avoir de linéarité entre les variables. En effet, sans cette étape préliminaire , notre matrice ne sera pas de rang maximal et donc dans ce cas procéder à un ACP n'aurait pas été possible . De ce fait nous créons une nouvelles base de données nommée *data_pca* contenant les variables suivantes : ("Born","MP", "Starts", "Min","Gls", "Ast", "PK", "PKatt" , "CrdY" , "CrdR", "xG" , "npxG","xAG" , "PrgC", "PrgP" , "PrgR").

Méthodologie de l'ACP

L'ACP repose sur l'étude de la matrice de corrélation des variables standardisées.

Étape 1 : Valeurs propres et vecteurs propres

La table suivante présente les cinq valeurs propres $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ associées aux cinq composantes principales sélectionnées.

λ_1	λ_2	λ_3	λ_4	λ_5
4.0883	1.0148	0.5396	0.2069	0.1621

TABLE 2 – Valeurs propres associées aux composantes principales

Ces valeurs propres représentent la variance expliquée par chacune des composantes principales. Puisque nous avons initialement cinq variables quantitatives, il est logique d'obtenir cinq valeurs propres. Chaque valeur propre représente l'importance de la composante dans la représentation de l'information contenue dans les données.

Étape 2 : Détermination du nombre de composantes principales à conserver

Plusieurs règles sont utilisées pour décider du nombre de composantes principales à retenir :

Règle 1 de Kaiser : Conserver les composantes principales dont la valeur propre est supérieure ou égale à 1. Nous observons que :

$$\lambda_1 = 4.088 \quad \text{et} \quad \lambda_2 = 1.015$$

Ces deux valeurs propres sont supérieures à 1. Nous décidons donc de conserver les **deux premières composantes principales** pour l'interprétation.

Règle 2 du Pourcentage d'inertie expliquée :

	λ_1	λ_2	λ_3	λ_4	λ_5
Standard deviation	2.0220	1.0074	0.7346	0.4548	0.4026
Proportion of Variance	0.5901	0.1390	0.1011	0.04508	0.04191
Cumulative Proportion	0.5901	0.7291	0.8302	0.87524	0.91715

TABLE 3 – Tableau des proportions d'inertie expliquée et cumulée par composante principale

D'après ce tableau, les trois premières composantes λ_1, λ_2 et λ_3 expliquent environ 83.0% de l'inertie totale. Ainsi, selon la règle 2, il est pertinent de conserver les **trois premières composantes principales**, car elles expliquent ensemble une part majoritaire de la variance du jeu de données.

Règle 3 du coude : Tracer le graphe des valeurs propres et repérer le **coude** marquant un changement de structure.

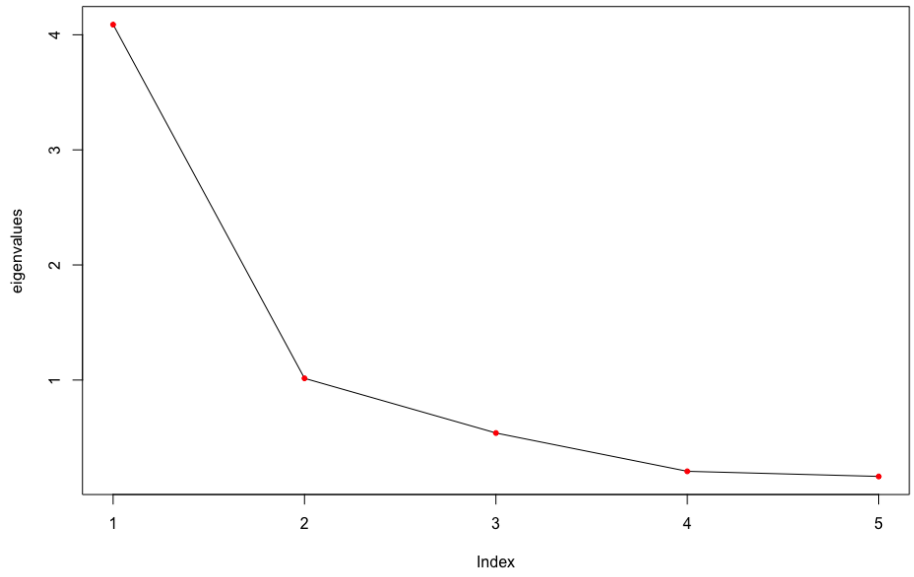

FIGURE 7 – Graphique des valeurs propres (méthode du coude)

Nous observons :

— Une forte décroissance entre la première et la deuxième composante,
— Puis une décroissance beaucoup plus progressive et quasi constante à partir de la troisième composante.

Le coude se forme donc après la deuxième composante. Ainsi, selon la règle du coude, il est judicieux de conserver uniquement les **deux premières composantes principales**.

Synthèse de la sélection des composantes principales

La décision finale se repose sur l'application combinée des trois règles précédentes. Pour des raisons d'interprétation et de visualisation, nous décidons de **conserver uniquement les deux premières composantes principales** dans la suite de notre analyse.

Étape 3 : Interprétation des composantes principales

L'interprétation des composantes principales se fait grâce aux **corrélations** entre les composantes principales retenues et les variables initiales.

Variable	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5
Born	-0.162	-0.969	0.175	0.018	0.015
MP	0.426	-0.051	0.041	0.237	0.201
Starts	0.430	-0.038	0.138	0.373	0.094
Min	0.435	-0.039	0.142	0.388	0.108
Gls	0.082	-0.039	-0.138	-0.217	0.416
Ast	0.158	-0.069	-0.336	-0.082	-0.244
PK	0.003	0.009	-0.011	-0.026	0.058
PKatt	0.004	0.012	-0.021	-0.029	0.070
CrdY	0.376	0.011	0.505	-0.672	-0.095
CrdR	$\sim$0	$\sim$0	$\sim$0	$\sim$0	$\sim$0
xG	0.094	-0.037	-0.150	-0.213	0.440
npxG	0.107	-0.046	-0.167	-0.237	0.488
xAG	0.181	-0.066	-0.352	-0.105	-0.198
PrgC	0.214	-0.155	-0.397	-0.051	-0.227
PrgP	0.350	-0.057	0.046	-0.156	-0.407
PrgR	0.157	-0.117	-0.456	-0.087	0.016

TABLE 4 – Vecteurs propres de la matrice de corrélation

Par ailleurs, la corrélation entre la h-ième composante principale et l'ensemble des variables initiales peut s'exprimer de manière vectorielle :

$$\text{Cor}[\phi_p, X] = \sqrt{\lambda_p} \times u_p$$

L'analyse des corrélations permet d'identifier quelles variables influencent le plus chaque axe factoriel et guide l'interprétation géométrique du plan principal.

Le tableau ci-dessous présente les corrélations entre les variables initiales et les deux premières composantes principales λ_1 et λ_2 :

Variable	**Corrélation avec ϕ_1**	**Corrélation avec ϕ_2**
Born	-0.327	-0.976
MP	0.860	-0.051
Starts	0.870	-0.038
Min	0.879	-0.039
Gls	0.166	-0.039
Ast	0.320	-0.069
PK	0.006	0.009
PKatt	0.007	0.012
CrdY	0.761	0.011
CrdR	$\sim$0	$\sim$0
xG	0.190	-0.037
npxG	0.216	-0.046
xAG	0.365	-0.067
PrgC	0.432	-0.156
PrgP	0.707	-0.058
PrgR	0.318	-0.118

TABLE 5 – Corrélations entre variables initiales et composantes principales

Les variables *PK*, *PKatt*, *CrdR* ne sont pratiquement pas corrélées avec ces axes. Elles ne seront donc **pas prises en compte dans l'interprétation** des 2 premières composantes principales.

Interprétation de la première composante principale ϕ_1

La première composante principale ϕ_1 est fortement corrélée positivement avec plusieurs variables offensives et de temps de jeu. Cela indique que ϕ_1 représente un **indice global de performance et de contribution offensive** : les joueurs obtenant des valeurs élevées sur cet axe sont ceux qui cumulent un grand nombre de minutes, de titularisations, d'actions offensives (passes, buts attendus, etc.). Inversement, les joueurs ayant des scores faibles sur cet axe sont ceux avec peu de temps de jeu et/ou peu de contribution offensive.

Interprétation de la deuxième composante principale ϕ_2

Toutes les variables sont corrélées négativement avec la deuxième composante principale ϕ_2. La variable *Born* domine négativement, donc la composante représente un classement des individus par leur âge : plus on est en bas plus on est jeune et plus on est vers le haut, plus on est âgé.

Étape 4 : Projection des individus

Projection selon le championnat

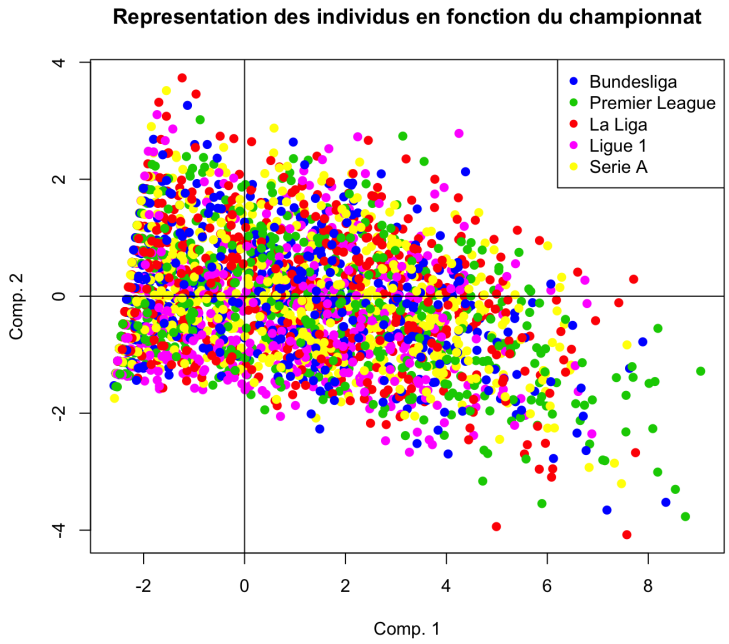

FIGURE 8 – Projection selon le championnat

Sur ce premier graphique, les joueurs sont colorés selon leur championnat d'appartenance. Nous observons que les joueurs issus des différents championnats sont relativement dispersés, sans formation de groupes nettement distincts.

Cela indique que les stratégies de jeu et les profils offensifs ne varient pas drastiquement d'un championnat à l'autre dans notre base de données. Chaque championnat semble contenir une grande hétérogénéité de profils de joueurs en termes d'intensité offensive et de temps de jeu. On ne peut rien dégager en fonction des championnats.

Projection selon la catégorie d'âge

Bien que la représentation graphique par catégories d'âge montre quelques tendances générales, il est important de noter que les groupes d'âge se répartissent principalement selon des segments horizontaux relativement proches les uns des autres sur le plan factoriel. Ces segments se chevauchent fortement, notamment entre les catégories *jeune*, *moyen* et *expérimenté*.

Cela suggère que, globalement, l'âge n'a pas un impact déterminant sur les performances offensives globales dans notre base de données. Des jeunes joueurs peuvent atteindre des performances proches de celles des joueurs plus expérimentés, et inversement, certains joueurs plus âgés peuvent avoir des statistiques moindres.

Ainsi, le facteur âge seul n'explique pas entièrement la variabilité des performances offensives, même s'il existe des tendances secondaires.

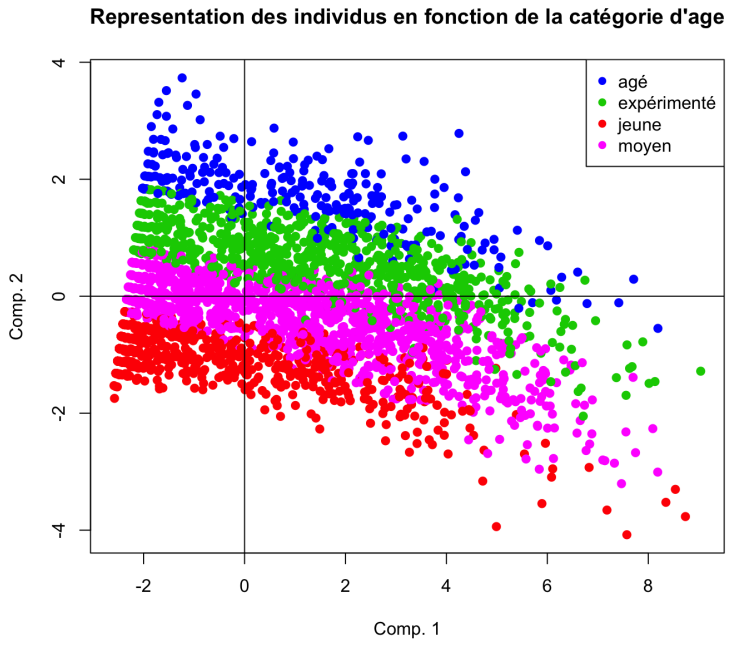

FIGURE 9 – Projection selon la catégorie d'âge

8 Analyse Factorielle des Correspondances Binaires (ACOBI)

Dans cette partie, nous étudions le lien entre la catégorie d'âge des joueurs et le championnat dans lequel ils évoluent avec une Analyse en Correspondances Binaires.

8.1 Description des variables

Soient les deux variables qualitatives :

- X, le *championnat*, à cinq modalités A_j $(j = 1, \ldots, 5)$,
- Y, la *catégorie d'âge*, à quatre modalités B_k $(k = 1, \ldots, 4)$, correspondant aux tranches d'âge définies dans le jeu de données.

8.2 Calculs Préliminaires

8.2.1 Table de contingence

Comp	agé	expérimenté	jeune	moyen	Total
de Bundesliga	45	148	102	212	507
Premier League (eng)	46	153	147	234	580
La Liga (es)	79	189	139	202	609
Ligue 1 (fr)	38	105	189	208	540
Serie A (it)	63	161	133	259	616
Total	271	756	710	1115	2852

TABLE 6 – Table de contingence des effectifs par championnat et catégorie d'âge.

8.2.2 Test d'indépendance

Avant de poursuivre l'ACOBI, il convient de vérifier l'indépendance statistique des variables X (championnat) et Y (catégorie d'âge). Si X et Y étaient indépendantes, l'ACOBI n'apporterait aucune information pertinente sur leur association.

Hypothèses

— $\mathcal{H}_0$: X et Y sont indépendantes.
— $\mathcal{H}_1$: X et Y sont dépendantes.

Conditions d'application

1. Taille de l'échantillon : $n \geq 30$ (ici $n = 2852$).
2. Tous les $n^*_{jk} \geq 1$.
3. Au moins 80 % des n^*_{jk} sont ≥ 5 (ici 100 %).

Résultats Comme $\chi^2 = 68{,}123 > \chi^2_{12,\,0.95} = 21{,}03$ et `p-value` $\ll 0{,}05$, on rejette $\mathcal{H}_0$ au niveau $\alpha = 5\%$. Il existe donc une association statistiquement significative entre le championnat et la catégorie d'âge.

8.2.3 Matrice d'attraction/répulsion

Comp	agé	expérimenté	jeune	moyen
de Bundesliga	0.9340815	1.1012387	0.8081340	1.0695536
eng Premier League	0.8346609	0.9951560	1.0180767	1.0319592
es La Liga	1.3651804	1.1707718	0.9168297	0.8484158
fr Ligue 1	0.7405767	0.7335391	1.4059155	0.9852450
it Serie A	1.0763167	0.9859909	0.8672855	1.0754586

TABLE 7 – Matrice d'attraction/répulsion entre championnats et catégories d'âge.

Exemple d'interprétation :

- La Liga et la Serie A attirent plus que prévu les joueurs expérimentés ($d > 1$),
- La Ligue 1 attire plutôt les jeunes joueurs et repousse davantage les joueurs âgés et expérimentés ($d < 1$),
- La Bundesligua attire plus les joueurs expérimentés et ceux d'âge moyen.

8.3 Interprétation

Nos championnats et catégories d'ages sont extrêmement bien représenté sur notre plan factoriel définit par nos deux composantes principale.

On utilise ici la représentation Pseudo-Barycentrique : elle permet d'interpréter les modalités lignes et colonnes qui se trouvent dans les mêmes quadrants factoriels ou dans des cadrants opposés (par exemple 1 et 3).

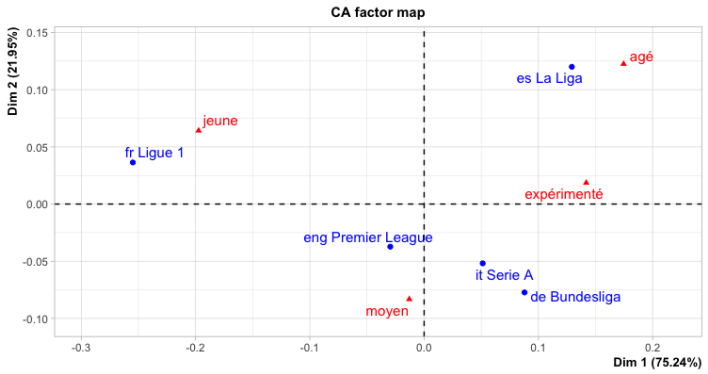

FIGURE 10 – Factor MAP CA

L'analyse factorielle des correspondances binaires (ACOBI) a permis d'étudier la relation entre les championnats européens et les catégories d'âge des joueurs. Le test du χ^2 a révélé une association statistiquement significative, justifiant la poursuite de l'analyse.

Le plan factoriel fondé sur les deux premiers axes (97,2 % de l'inertie) met en évidence une organisation contrastée :

- *Ligue 1* est dans le même cadrant que la modalité jeune, donc clairement elle attire cette modalité, confirmant son rôle de recrue de jeunes talents.
- *Liga* : attirent les modalités « expérimenté » et « âgé », traduisant une stratégie tournée vers l'expérience.
- *Premier League* : située dans le cadrant 3 attire plus les joueurs d'age moyens, tout en restant proche du centre, suggérant peut etre une répartition équilibrée des âges.
- *Serie A* et *Bundesliga* : rejette les profils « jeune », cependant on peut rien dire sur les profils qu'ils attirent étant donné qu'ils sont seuls dans leur cadrant.

Ces observations sont cohérentes avec la matrice d'attraction/répulsion, qui confirme :

$$D_{\text{Ligue 1, jeune}} \approx 1,41, \quad D_{\text{Liga, âgé}} \approx 1,37.$$

Enfin, rappelons qu'une représentation pseudo-barycentrique ne permet une lecture fiable que pour :

- les modalités situées dans un même quadrant (proches statistiquement) ;
- ou celles situées dans des quadrants opposés (opposition nette).

Les modalités se trouvant dans des quadrants adjacents ne peuvent pas être directement comparées de manière rigoureuse.

9 Régression Robuste

Grâce à l'analyse préliminaire de détection des valeurs aberrantes,faites précédemment, nous sommes certain de la présence d'outliers dans notre base de données. Nous souhaitons alors expliquer la variable Gls (le nombre de buts marqué) en utilisant une régression. Or, une régression simple ne fera pas l'affaire dans notre cas, dû justement à la présence d'outliers. Ainsi, nous allons légèrement sortir du cadre du cours, en utilisant une régression robuste, via la fonction *rlm* dans R, qui repose sur les M-estimateurs. Ces derniers permettent de réduire l'influence des outliers en leur attribuant un poids plus faible.

Dans un premier temps, nous effectuons une régression robuste avec une seule variable explicative : l'âge du joueur. Le pseudo R^2 obtenu est -0.015, étant très proche de zero, on peut raisonablement conclure que l'âge n'explique pas le nombre de but marqués.

Paramètre	**Valeur**	**Erreur Std.**	**t-value**
Intercept	0.5171	0.4845	1.0672
Age	0.0793	0.0191	4.1485

Résidus : Min = -3.37, 1Q = -2.10, Médiane = -0.82, 3Q = 1.87, Max = 33.10

Erreur standard des résidus : 3.118, **ddl :** 1272

Nous rajoutons des variables explicatives à notre modèle : le nombre de passes décisives, le championnant dans lequel le joueur évolue, ainsi que le nombre de minutes jouées. Les résultats (en utilisant la fonction summary()), montrent que les variables ayant un impact significatif sont nombre de passes décisives et le temps de jeu. Pour le temps de jeu cela semble plutôt intuitif ; plus un attaquant joue, plus il a d'opportunités de marquer des buts.

Paramètre	**Valeur**	**Erreur Std.**	**t-value**
(Intercept)	-0.1111	0.4147	-0.2680
Age	0.0048	0.0156	0.3058
Ast	0.5460	0.0339	16.1054
Compeng Premier League	0.1501	0.2074	0.7236
Compes La Liga	-0.1870	0.2054	-0.9103
Compfr Ligue 1	-0.0146	0.2104	-0.0695
Compit Serie A	0.1467	0.2066	0.7102
Min	0.0012	0.0001	13.2806

Résidus : Min = -8.63, 1Q = -1.19, Médiane = -0.25, 3Q = 1.44, Max = 28.10
Erreur standard des résidus : 1.88, **ddl :** 1266

10 Limites de l'étude

Dans cette section, nous allons mettre en avant certaines limites de notre étude.

Tout d'abord, cette étude se base uniquement sur des statistiques liées à la saison 2023/2024, ce qui ne permet pas une généralisation temporelle de nos résultats.

Ensuite, notre base de données ne recense que les matchs de championnat des cinq grands championnats européens. Les matchs de coupe nationale, de coupe d'Europe, ainsi que les matchs internationaux ne sont pas pris en compte, ce qui nous prive d'informations potentiellement utiles.

Par ailleurs, notre base de données comporte principalement des données offensives des joueurs. Il manque de nombreuses données concernant les milieux de terrain, les défenseurs et les gardiens de but. Cela ne permet donc pas de réaliser des analyses complètes sur ces types de joueurs.

De plus, nous constatons la présence de nombreuses valeurs aberrantes dans notre jeu de données. Ces anomalies peuvent biaiser certaines de nos analyses statistiques et nous ont poussé à adopter une ACP robuste.

Enfin, notre base de données ne contient qu'une trentaine de variables. Cela reste insuffisant pour mener une étude vraiment fiable, car le football est un sport très complexe. D'autres variables auraient permis d'approfondir notre étude et d'enrichir nos résultats (par exemple : le nombre de jours d'indisponibilité pour blessure, le nombre de titres remportés à la fin de la saison, le nombre de prix d'homme du match reçus tout au long de la saison, etc.).

11 Conclusion

À travers cette analyse, nous avons analysé l'impact de l'âge sur les performances offensives des joueurs de football évoluant dans les cinq grands championnats européens lors de la saison 2023/2024. Grâce à des méthodes descriptives, robustes (ACP robuste) et exploratoires (ACOBI), nous avons mis en évidence plusieurs résultats.

Nos résultats montrent que si l'âge influe sur certains paramètres de jeu, il n'explique pas à lui seul la performance offensive globale. Des jeunes joueurs peuvent atteindre des niveaux de performance similaires à ceux des joueurs plus expérimentés. En revanche, l'analyse des correspondances binaires révèle des spécificités par championnat : la Ligue 1 attire majoritairement des jeunes talents, tandis que la Liga privilégie davantage les joueurs expérimentés et âgés. La Premier League présente une structure composée de joueur d'âge moyen, tandis que pour la Serie A et la Bundesliga, on ne peut rien conclure sur le profil d'âge de leurs joueurs.

Cependant, notre étude présente des limites, notamment l'analyse sur une seule saison, la focalisation sur des données majoritairement offensives et la présence de nombreuses valeurs aberrantes.

En conclusion, cette étude met en évidence la complexité à identifier des facteurs influençant la performance des joueurs professionnels et souligne l'importance d'une approche multivariée et robuste dans l'analyse de données sportives.

12 Annexes

12.1 Résumé statistique pour la détection des valeurs aberantes

Variable	Min	1st Qu.	Median	Mean	Max
Rk	2.0	696.2	1404.5	1414.6	2852.0
Player	-	-	-	-	-
Nation	-	-	-	-	-
Pos	-	-	-	-	-
Squad	-	-	-	-	-
Comp	-	-	-	-	-
Age	15	22	25	25	37
Born	1985	1995	1998	1998	2007
MP	1.00	14.00	25.00	22.65	38.00
Starts	0.00	5.00	14.00	14.65	38.00
Min	90	501	1230	1312	3325
90s	1.00	5.60	13.70	14.58	36.90
Gls	0.000	0.000	2.000	3.173	36.000
Ast	0.000	0.000	1.000	1.903	14.000
G+A	0.000	1.000	3.000	5.076	44.000
G-PK	0.000	0.000	2.000	2.835	31.000
PK	0.0000	0.0000	0.0000	0.3375	10.0000
PKatt	0.0000	0.0000	0.0000	0.4199	10.0000
CrdY	0.000	1.000	2.000	3.016	17.000
CrdR	0.0000	0.0000	0.0000	0.1209	2.0000
xG	0.000	0.700	1.850	3.191	30.600
npxG	0.000	0.600	1.700	2.858	26.700
xAG	0.000	0.500	1.250	1.913	11.800
npxG+xAG	0.000	0.160	0.290	0.3229	1.4400
PrgC	0.00	9.00	20.00	30.38	218.00
PrgP	0	15	39	57	392
PrgR	0.0	22.0	49.0	72.8	508.0
Gls_90	0.0000	0.0000	0.1400	0.1979	1.4300
Ast_90	0.0000	0.0000	0.0800	0.1183	1.0100
$G+A_90$	0.0000	0.0900	0.2700	0.3161	1.4300
$G\text{-}PK_90$	0.0000	0.0000	0.1200	0.1814	1.4300
$G+A\text{-}PK_90$	0.0000	0.0900	0.2500	0.2997	1.4300
xG_90	0.0000	0.0800	0.1700	0.2138	1.2300
xAG_90	0.0000	0.0500	0.1000	0.1263	1.0200
$xG+xAG_90$	0.0000	0.1600	0.3100	0.3403	1.4400

TABLE 8 – Résumé des statistiques des différentes variables

12.2 Code R

```
####Importation du fichier + petit clean :
library(dplyr) ## est utilisé afin de créer une nouvelle variable (catégorie d'age)
library(MASS)
library(robustbase)

##### Lecture des données et études préliminaires
data = read.csv("top5-players.csv")

str(data)
summary(data)

# Rajout des années de naissance, âge et nationalité
data$Born[data$Player == "Marco Pellegrino"] <- 2002
data$Born[data$Player == "Max Moerstedt"] <- 2006
data$Born[data$Player == "Max Svensson"] <- 2001

data$Age[data$Player == "Marco Pellegrino"] <- 21
data$Age[data$Player == "Max Moerstedt"] <- 17
data$Age[data$Player == "Max Svensson"] <- 21

data$Nation[data$Player == "Marco Pellegrino"] <- "ar ARG"
data$Nation[data$Player == "Mahmut Kücüksahin"] <- "tr TUR"

## Création d'une nouvelle variable contenant les catégories d'âge
data <- data %>%
  mutate(cat_age = case_when(
    Age <= 21 ~ "jeune",
    Age >= 22 & Age <= 26 ~ "moyen",
    Age >= 27 & Age <= 31 ~ "expérimenté",
    Age >= 32 ~ "agé"
  ))

data = na.omit(data) # Retire le joueur avec de nombreux NA


#### #ANALYSES PRELIMINAIRES

# Répartition par championnat avec pourcentages
comp_table <- table(data$Comp)
comp_pct <- round(100 * comp_table / sum(comp_table), 1)
pie(comp_table, labels = paste0(names(comp_table), " (", comp_pct, "%)"),
    main = "Répartition par championnat")

# Répartition par catégorie d'âge avec pourcentages
age_table <- table(data$cat_age)
```

```
age_pct <- round(100 * age_table / sum(age_table), 1)
pie(age_table, labels = paste0(names(age_table), " (", age_pct, "%)"),
    main = "Répartition par catégorie d'âge")

table(data$cat_age, data$Comp)

# Calcul des statistiques descriptives
age_buts_summary <- data %>%
  group_by(cat_age) %>%
  summarise(
    mean_buts = mean(Gls, na.rm = TRUE),
    median_buts = median(Gls, na.rm = TRUE),
    sd_buts = sd(Gls, na.rm = TRUE),
    n = n()
  )

# Affichage des statistiques descriptives
print(age_buts_summary)

# Boxplot des buts par catégorie d'âge
boxplot(Gls ~ cat_age, data = data,
        main = "Répartition des buts par catégorie d'âge",
        ylab = "Buts", xlab = "Catégorie d'âge", col = "lightblue")


boxplot(data_FW$Gls,
        main = "Répartition des buts - Attaquants",
        ylab = "Nombre de buts",
        col = "lightblue")

age_table <- table(data$cat_age)

# Barplot classique
barplot(age_table,
        col = "lightblue",
        border = "black",
        main = "Répartition des joueurs par catégorie d'âge",
        xlab = "Catégorie d'âge",
        ylab = "Nombre de joueurs")

barplot(comp_table,
        col = "lightblue",
        border = "black",
        main = "Répartition des joueurs par championnat",
        xlab = "Championnat",
        ylab = "Nombre de joueurs")
```

```
prop_contingence <- prop.table(table(data$cat_age, data$Comp), margin = 2) * 100
print(prop_contingence)

colors <- colorRampPalette(c("blue", "red"))(ncol(table(data$cat_age, data$Comp)))

# Créer un barplot avec le dégradé de couleurs
barplot(table(data$cat_age, data$Comp),
        beside = TRUE,  # Barres côte à côte
        col = colors,   # Appliquer le dégradé de couleurs
        legend = TRUE,
        main = "Répartition des catégories d'âge par championnat",
        xlab = "Championnat",
        ylab = "Nombre de joueurs")


#####Détection d'outliers :

data_clean <- data[data$Min >= 90 & data$Pos %in% c("FW", "MF","MF,FW","FW,MF"), ]

summary(data_clean)

# Boxplot des buts
boxplot(Gls ~ cat_age, data = data_clean,
        col = "skyblue", main = "Buts par tranche d'âge (FW & MF)",
        xlab = "Tranche d'âge", ylab = "Buts ")


# Boxplot des passes décisives
boxplot(Ast ~ cat_age, data = data_clean,
        col = "lightgreen", main = "Passes décisives /par tranche d'âge (FW & MF)",
        xlab = "Tranche d'âge", ylab = "Passes décisives ")


data_numerique <- data_clean[, c("Gls", "Ast","Age")]

MCD <- covRob(data_numerique, estim="mcd")

distrob=sqrt(mahalanobis(data_numerique, MCD$center, MCD$cov))
cutoff=sqrt(qchisq(0.975,df=ncol(data_numerique)))
plot(distances, pch = 16, col = ifelse(distances > cutoff,  "red", "blue")
,main = "Distances de Mahalanobis Robustes")
abline(h=cutoff,col = 'black')

####PCA robuste:

colonne_pca = c("Born","MP", "Starts", "Min","Gls", "Ast", "PK", "PKatt" , "CrdY"
, "CrdR", "xG" , "npxG","xAG" , "PrgC", "PrgP" , "PrgR")
```

```
data_pca=na.omit(data[,colonne_pca])

res_robust <- PcaHubert(data_pca, k = 5, scale = TRUE)
summary(res_robust)

#Deuxième étape: déterminer le nombre de composantes principales à conserver
#Règle 1: garder les composantes principales
dont les variances sont plus grandes ou égales à 1
eigenvalues=res_robust@eigenvalues
eigenvalues
#on garde comp. 1 et comp.2

#Règle 2: garder les composantes principales qui ont un pourcentage d'inertie élevé
summary(res_robust)
#on garde comp. 1 et comp.2 et comp.3

#Règle 3: règle du coude à l'aide d'un graphique des valeurs propres
plot(eigenvalues, type='line')
points(eigenvalues, pch=20, col='red')
#on garde comp. 1 et comp.2

#Conclusion des trois règles: on garde comp. 1 et comp.2

eigenvectors=res_robust@loadings
#valeurs manquantes car R décide de ne pas les afficher
eigenvectors

plot(res_robust@scores[,1:2])
#on projette sur les première et deuxième composantes principales
abline(h=0,v=0)


#Objectif: vérifier que l'interprétation de comp.1 et comp.2 de l'étape 3
soit cohérente
#Des graphiques plus explicites peuvent s'obtenir de la manière suivante:
# représentation des individus en fonction du championnat
# S'assurer que 'Comp' est un facteur
data$Comp <- as.factor(data$Comp)

# Palette de couleurs (5 couleurs pour les 5 ligues)
palette_couleurs <- c("blue", "green3", "red", "magenta", "yellow")

#### on trace 3 graphe comp.1/comp.2 , comp.1/comp.3 et comp.2/comp.3

plot(res_robust@scores[,1],
     res_robust@scores[,2],
     xlab = "Comp. 1",
```

```
     ylab = "Comp. 2",

     main = "Representation des individus en fonction du championnat",
     pch = 19,
     col = palette_couleurs[as.numeric(data$Comp)])
abline(h=0,v=0)
legend("topright",c("Bundesliga","Premier League","La Liga","Ligue 1" , "Serie A")
,cex=1,col=c("blue", "green3", "red", "magenta", "yellow"),pch=16)


# représentation des individus en fonction de la catégorie d'age
# S'assurer que 'Comp' est un facteur
data$cat_age <- as.factor(data$cat_age)

# Palette de couleurs (4 couleurs pour les 4 catégorie d'age)
palette_couleurs <- c("blue", "green3", "red", "magenta")

#### on trace 3 graphe comp.1/comp.2 , comp.1/comp.3 et comp.2/comp.3
# Tracer le scatter plot
plot(res_robust@scores[,1],
    res_robust@scores[,2],
     xlab = "Comp. 1",
     ylab = "Comp. 2",
     main = "Representation des individus en fonction
     de la catégorie d'age"
     pch = 19,
     col = palette_couleurs[as.numeric(data$cat_age)])
abline(h=0,v=0)
legend("topright",c("agé","expérimenté","jeune","moyen")
,cex=1,col=c("blue", "green3", "red", "magenta"),pch=16)

####ACOBI:

########### quel est le lien entre la catégorie d'âge et le championnat ?

## Y-a-t'il des liens entre l'age du joueur et le championnat
( en terme du nombre de joueurs présent)
##ici nous allons faire une ACOBI.

###création de la table de contingence
aggregation <- with(data, table(Comp, cat_age))
print(aggregation)

#Test d'indépendance entre le championnat et l'age (hypothèse vérifiée)
test=chisq.test(aggregation)
test
## petite p-valeur donc en rejette que les variables sont indépendente.
```

```
on peut donc continuer

library(FactoMineR)
bca=CA(aggregation) #ACOBI sur aggregation
bca$eig #valeurs propres, pourcentage d'inertie, pourcentage d'inertie cumulé

#Valeurs propres
bca$eig[,1]
barplot(bca$eig[,1])

#Pourcentage d'inertie expliquée par les axes principaux
bca$eig[,2:3]

#Coordonnées des modalités sur les axes principaux
bca$row$coord #pour les profils lignes
bca$col$coord #pour les profils colonnes

#Contribution des modalités à la construction des axes principaux
bca$row$contrib #pour les profils lignes
bca$col$contrib #pour les profils colonnes

#Qualité de représentation des modalités sur les axes principaux
bca$row$cos2 #pour les profils lignes
bca$col$cos2 #pour les profils colon


#### régression robuste:
#Après avoir fait l'analyse de présence d'outlier, il est donc justifié d'utiliser
une regression robuste
On essaye d'abord d'expliqué le nombre de goal avec la variable âge en faisant
une régression linéaire classique et robuste
(juste pour voir la différence entre les deux)

#regression robuste univariée
reg_robuste <- rlm(Gls ~ Age, data = data_clean)
abline(reg_robuste, col = "darkgreen", lwd = 2, lty = 2)  # Robuste
summary(reg_robuste)
rss <- sum(resid(reg_robuste)^2)
tss <- sum((data_plot$Gls - mean(data_plot$Gls))^2)
pseudo_r2 <- 1 - rss / tss
pseudo_r2

#regression robuste multivariée
reg_robuste_multi <- rlm(Gls ~ Age + Ast+ Comp +Min, data = data_clean)
summary(reg_robuste_multi)
rss_multi <- sum(resid(reg_robuste_multi)^2)
tss_multi <- sum((data_clean$Gls - mean(data_clean$Gls))^2)
```

```
pseudo_r2_multi <- 1 - rss_multi / tss_multi
pseudo_r2_multi
```

Références

[1] CIES Football Observatory. *Répartition des minutes par âge – Rapport hebdomadaire n°349*. 2021. Disponible à : https://football-observatory.com/IMG/sites/b5wp/2021/wp349/fr .

[2] Bollier AC, et al. *Association Between Professional Football Participation and Long-term Health Outcomes*. PubMed, 2022. Disponible à : https://pubmed.ncbi.nlm.nih.gov/35307260/ .

[3] Aktas, Orkun. *All Football Players Stats in Top 5 Leagues 23/24*. Kaggle. Disponible à : https://www.kaggle.com/datasets/orkunaktas/all-football-players-stats-in-top-5-leagues-2324.

[4] Todorov, V., & Filzmoser, P. (2022). *rrcov : Scalable Robust Estimators with High Breakdown Point*. R package version 1.7-11. Retrieved from https://CRAN.R-project.org/package=rrcov.