RAPPORT D'ANALYSE DE DONNÉES

Stanislas ROLLAND - Pierre LECHAT

Septembre 2025

Table des matières

Introduction

Dans le cadre de notre formation en BUT Informatique à l'IUT de Lannion, nous avons été amenés à réaliser un projet d'analyse de données.

Nous avons choisi d'analyser un jeu de données concernant les statistiques des joueurs de football évoluant dans les cinq meilleures ligues européennes durant la saison 2024 - 2025. Ce choix a été motivé par notre intérêt commun pour le football et par la richesse des données disponibles dans ce domaine.

Le présent rapport détaille les différentes étapes de notre analyse, depuis la présentation du jeu de données jusqu'aux techniques d'analyse utilisées, en passant par les résultats obtenus et leur interprétation.

Présentation du jeu de données

Description du jeu de données

Le jeu de données que nous avons choisi d'analyser est un ensemble de données concernant les statistiques des joueurs de football sur la saison 2024 - 2025 les cinq meilleures ligues européennes (Premier League, La Liga, Serie A, Ligue 1, Bundesliga).

Notre jeu de données comporte 197 entrées (une par joueur) et les 15 variables suivantes :

Variables qualitatives :

- **Player** : Nom et prénom du joueur.
- **Nation** : Nationalité du joueur.
- **Pos** : Position principale du joueur sur le terrain (ex : Attaquant, Défenseur, Gardien).
- **Squad** : Équipe actuelle du joueur.
- **Comp** : Compétition dans laquelle le joueur évolue (ex : Premier League, La Liga etc).

Variables quantitatives :

- **Id** : Identifiant unique du joueur.
- **Age** : Âge du joueur.
- **Born** : Année de naissance du joueur.
- **MP** : Nombre de matchs joués par le joueur.
- **Starts** : Nombre de matchs commencés comme titulaire.
- **Min** : Nombre total de minutes jouées.
- **90s** : Nombre de périodes de 90 minutes jouées (équivalent à des matchs complets).
- **Gls** : Nombre total de buts marqués.
- **Ast** : Nombre total de passes décisives.
- **G+A** : Somme des buts et des passes décisives.

Exemple d'une entrée du jeu de données

```
Id,Player,Nation,Pos,Squad,Comp,Age,Born,MP,Starts,Min,90s,Gls,Ast,G+A
1,Bukayo Saka,eng ENG,"FW,MF",Arsenal,Premier League,22.0,2001,25,20,1729,19.2,6,10,16
```

Figure 1 –

Présentation des techniques d'analyse de données

Analyse quantitative des données : l'ACP

Définition

L'Analyse en Composantes Principales (ACP) est une technique statistique utilisée pour réduire la dimensionnalité d'un jeu de données tout en conservant le maximum d'information possible. Elle permet de transformer un ensemble de variables corrélées en un ensemble de variables non corrélées appelées composantes principales.

L'ACP est particulièrement utile lorsque l'on travaille avec des données multivariées, c'est-à-dire des données comportant plusieurs variables quantitatives. En réduisant le nombre de dimensions, l'ACP facilite la visualisation et l'interprétation des données, tout en aidant à identifier les structures sous-jacentes et les relations entre les variables.

Cercle de corrélation

Définition

Le cercle de corrélation est un outil graphique utilisé dans le cadre de l'ACP pour visualiser les relations entre les variables originales et les composantes principales. Chaque variable est représentée par un vecteur dans un plan défini par les deux premières composantes principales.

La position et la longueur des vecteurs permettent d'interpréter la contribution de chaque variable aux composantes principales. Par exemple, des vecteurs proches les uns des autres indiquent des variables fortement corrélées, tandis que des vecteurs perpendiculaires suggèrent une absence de corrélation. La distance d'un vecteur à l'origine reflète l'importance de la variable dans la formation des composantes principales.

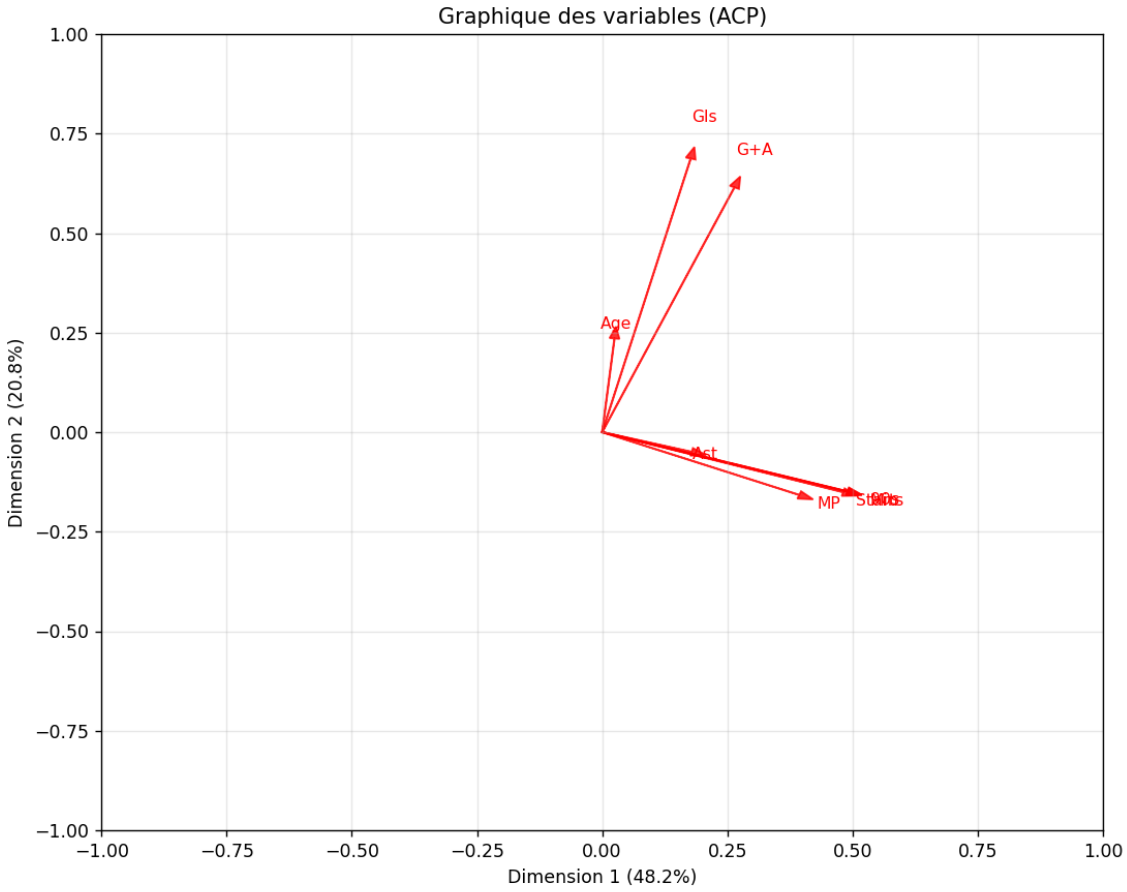

FIGURE 2 –

Interprétation

Dans la figure ci-dessus, on peut observer que les variables "Gls" et "G+A" sont plus ou moins corréléeses car "G+A" est la somme de "Gls" et "Ast". En revanche la variable "Ast" n'est pas corrélées avec ces dernières, ce qui peut être expliqué par le fait que les joueurs qui inscrivent beaucoup de buts inscrivent généralement beaucoup moins de passes décisives.

Quant à elle, les variables "MP", "Starts", "Min" et "90s" sont également fortement corrélées entre elles, ce qui reflète le fait que ces variables mesurent différentes facettes du temps de jeu des joueurs.

La variable "Age", elle, semble être très faiblement corrélée avec les autres variables, ce qui peut indiquer que l'âge des joueurs n'a pas une influence directe sur leurs performances statistiques dans ce jeu de données.

Graphique des individus

Définition

Le graphique des individus est un outil visuel utilisé dans le cadre de l'ACP pour représenter les observations (individus) dans l'espace défini par les composantes principales. Chaque point sur le graphique correspond à une observation du jeu de données, et la position de chaque point est déterminée par ses coordonnées sur les axes des composantes principales.

Ce graphique permet d'identifier des groupes d'individus similaires, des tendances générales, ainsi que des observations atypiques ou des outliers. En analysant la distribution des points, on peut tirer des conclusions sur la structure sous-jacente des données et sur les relations entre les différentes observations.

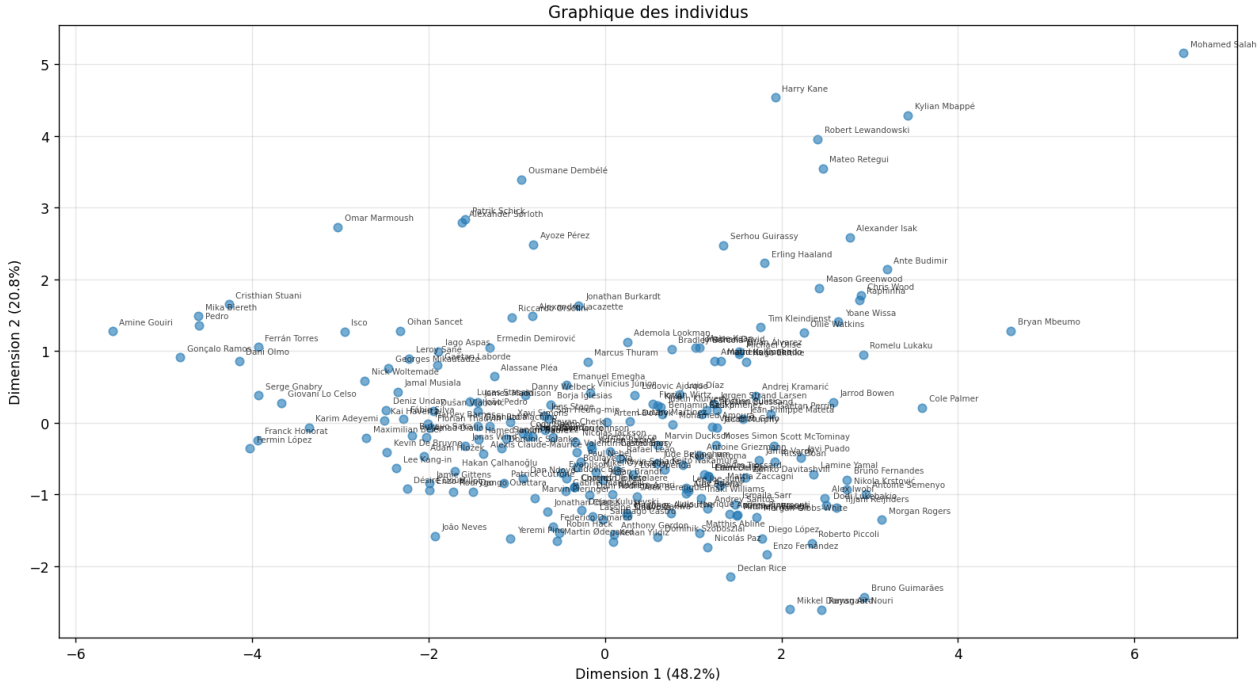

Figure 3 –

Interprétation

Dans la figure ci-dessus, on peut observer que les joueurs sont répartis en plusieurs groupes distincts. Par exemple, on peut identifier un groupe de joueurs situés dans la partie supérieure droite du graphique, qui sont probablement des attaquants ou des milieux offensifs, car ils ont des valeurs élevées pour les variables "Gls" et "G+A".

En revanche, les joueurs situés dans la partie inférieure gauche du graphique ont des valeurs plus faibles pour ces variables, ce qui suggère qu'ils occupent des postes plus reculés et / ou orientés sur la défense.

De plus, on peut remarquer que certains joueurs sont situés loin du centre du graphique, ce qui indique qu'ils ont des performances statistiques atypiques par rapport à la majorité des joueurs. Ces observations peuvent être considérées comme des outliers et méritent une attention particulière pour comprendre les raisons de leurs performances exceptionnelles ou médiocres.

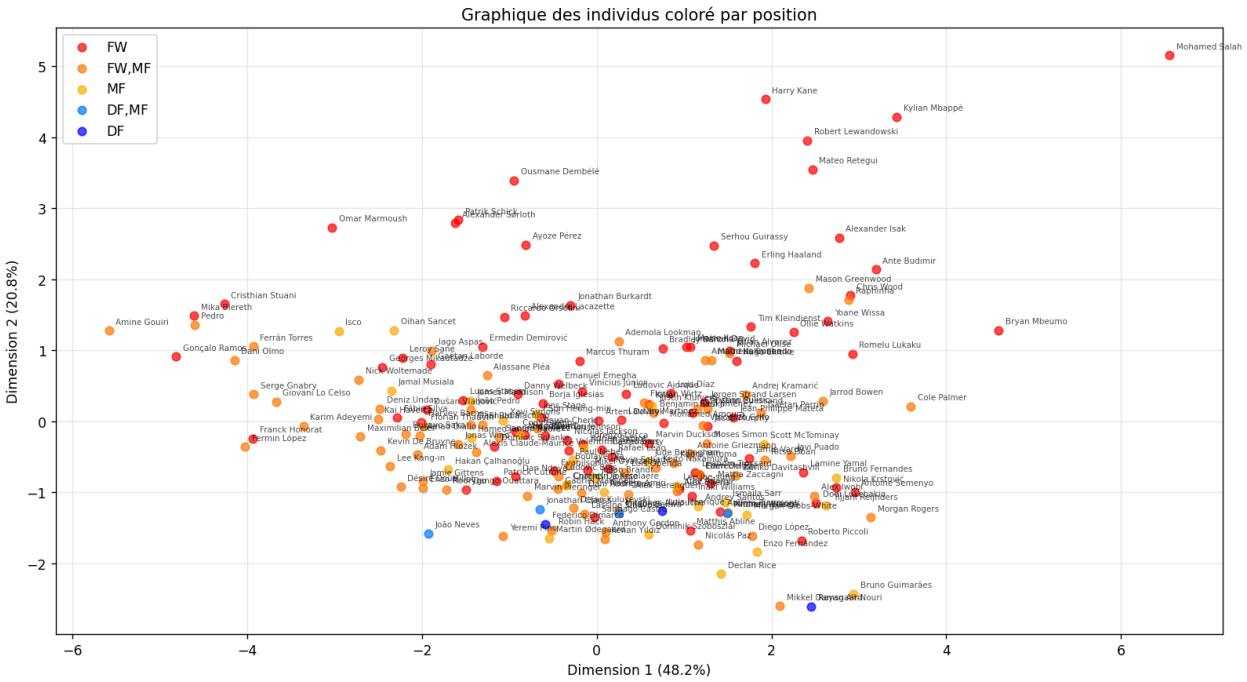

FIGURE 4 –

Interprétation

La figure ci-dessus confirme notre interprétation précédente, en effet, on peut observer que les joueurs représentés par des couleurs chaudes comme le rouge et l'orange se démarquent par rapport aux autres car ils occupent des postes avancés sur le terrain et donc qu'ils inscrivent plus de buts que les joueurs occupant des postes reculés.

Les joueurs représentés avec des couleurs plus froides, notamement le bleu foncé, sont situés dans la partie inférieure , ce qui confirme donc que même les meilleurs joueurs occupant des postes reculés inscrivent très peu de buts comparé aux joueurs avec des postes avancés.

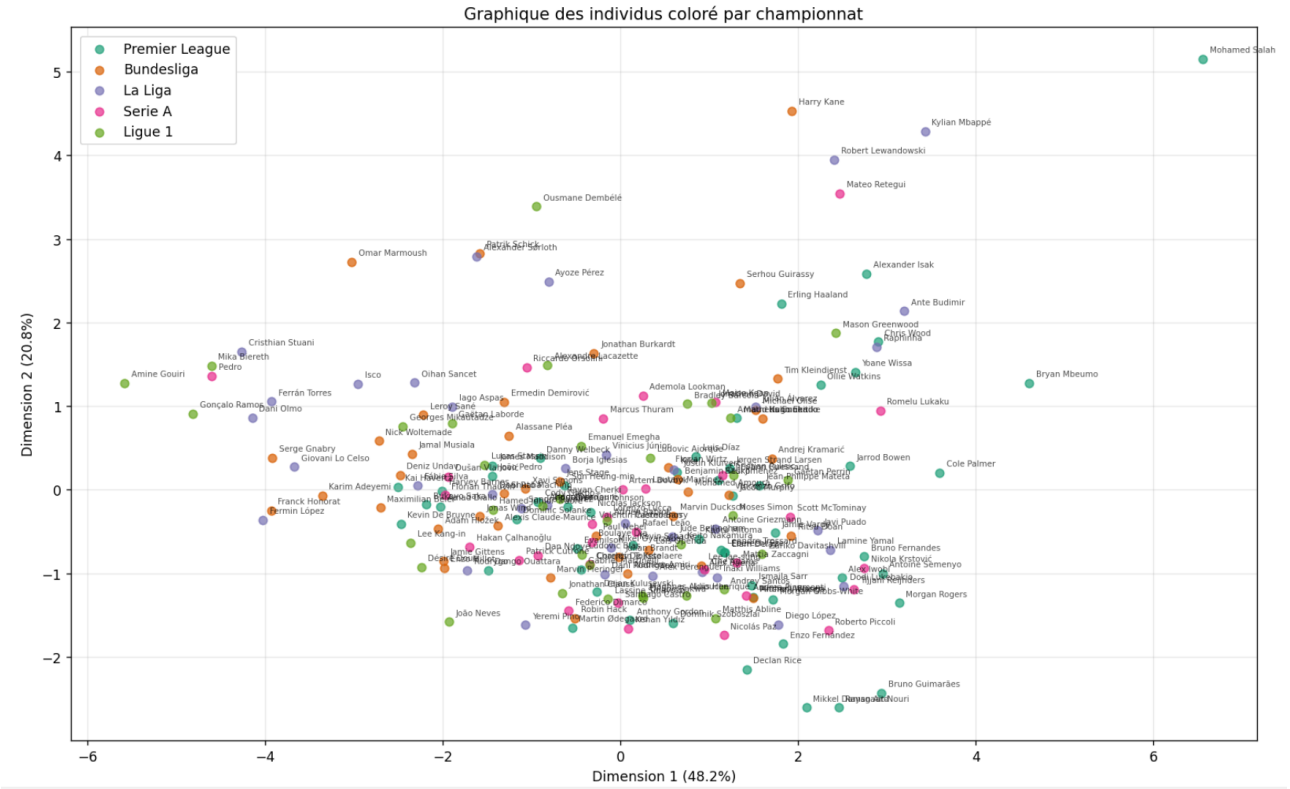

Figure 5 –

Interprétation

Dans la figure ci-dessus, on constate qu'il est difficile de distinguer les joueurs en fonction des championnats dans lesquels ils évoluent. En effet, les joueurs de différents championnats sont dispersés de manière plutôt homogène sur le graphique, ce qui suggère que les performances statistiques des joueurs ne sont pas fortement influencées par le championnat dans lequel ils jouent.

Valeurs propres

Définition

Les valeurs propres mesurent l'importance de chaque composante principale dans une ACP. Plus la valeur est grande, plus la composante explique une part importante de la variance des données. À l'inverse, une petite valeur signifie que la composante apporte peu d'information.

En général, on retient les composantes principales dont les valeurs propres sont supérieures à 1, car elles expliquent plus de variance qu'une variable originale.

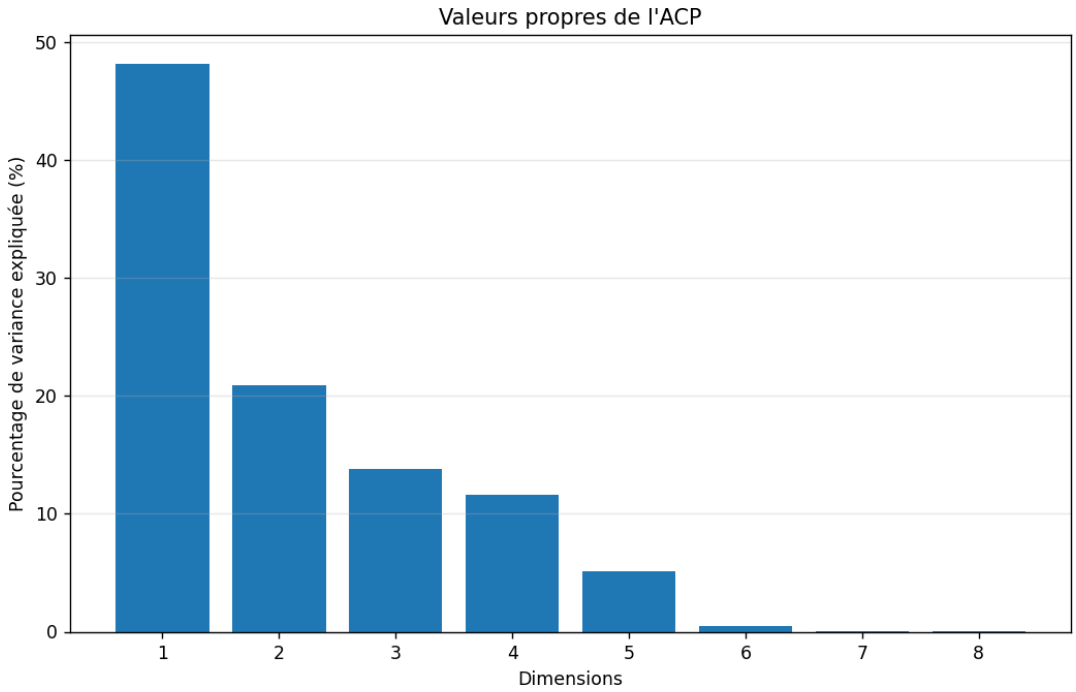

FIGURE 6 –

Interprétation

Dans la figure ci-dessus, on peut observer que les deux premières composantes principales ont des valeurs propres significativement plus élevées que les autres (près de 70% de la variance totale expliquée). Cela indique que ces deux composantes capturent la majorité de l'information contenue dans les données et donc prouvent que l'ACP réduit efficacement la dimension du jeu de données.

Cependant, bien que ces deux premières couvrent la majeure partie du jeu de données, il est préférable de conserver les quatres premières composantes principales pour obtenir une représentation plus fidèle de notre jeu de données (environ 90%).

Analyse catégorielle des données : AFC

Définition

L'Analyse Factorielle des Correspondances (AFC) est une technique statistique utilisée pour analyser des données qualitatives, souvent présentées sous forme de tableaux de contingence. Elle permet de visualiser les relations entre les catégories de deux variables qualitatives en les représentant dans un espace à deux dimensions.

L'AFC est particulièrement utile pour identifier des associations entre les catégories, détecter des groupes similaires et comprendre la structure sous-jacente des données qualitatives. En réduisant la complexité des données, l'AFC facilite l'interprétation et la communication des résultats.

Graphique des associations

Définition

Le graphique des associations est un outil visuel utilisé dans le cadre de l'AFC pour représenter les relations entre les catégories de deux variables qualitatives. Chaque catégorie est représentée par un point dans un plan défini par les deux premières dimensions de l'analyse.

Ce graphique permet d'identifier des associations entre les catégories, des tendances générales, ainsi que des groupes similaires. En analysant la distribution des points, on peut tirer des conclusions sur la structure sous-jacente des données et sur les relations entre les différentes catégories.

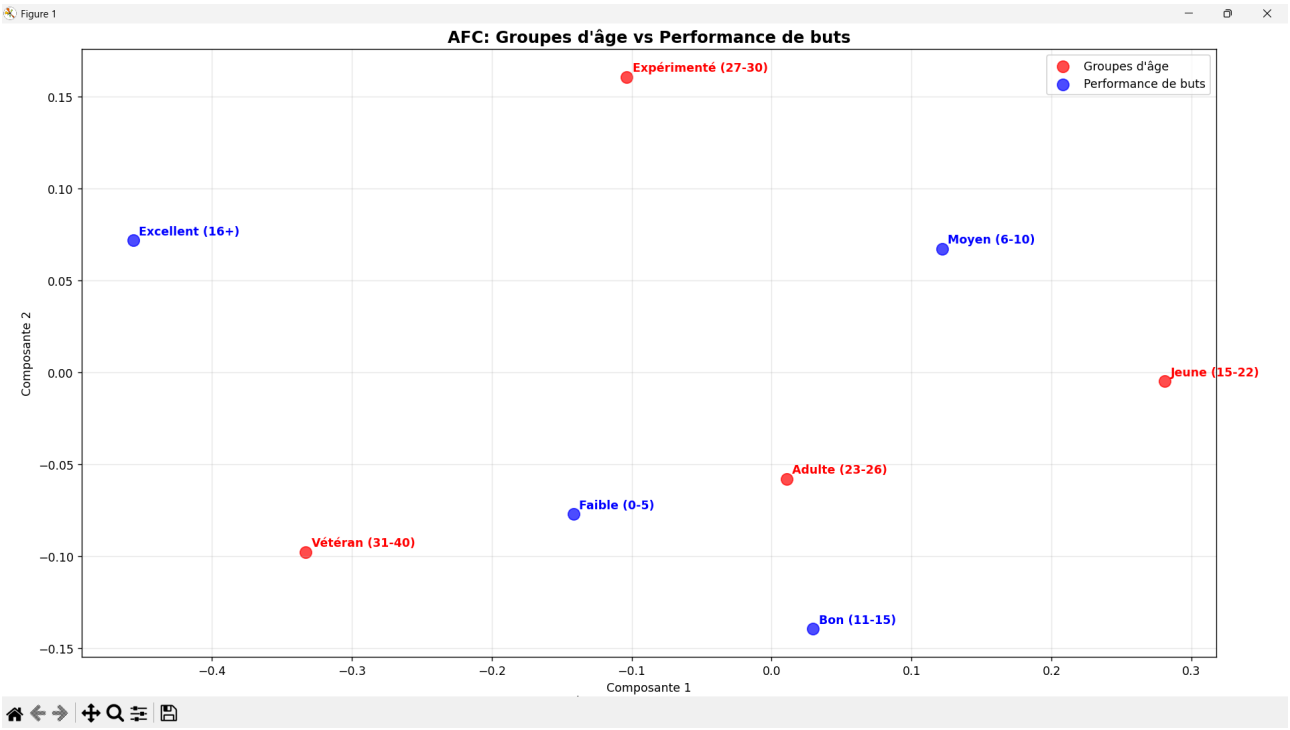

FIGURE 7 –

Interprétation

Dans la figure ci-dessus, on peut observer qu'il existe une progression logique des performances en fonction de l'âge. En effet, les joueurs jeunes (15–22 ans) se situent à proximité de la catégorie "Moyen (6–10 buts)", indiquant qu'ils obtiennent généralement des résultats corrects mais encore en phase de progression. Les adultes (23–26 ans) se rapprochent de la catégorie "Bon (11–15 buts)", reflétant une amélioration notable liée à une plus grande expérience de jeu.

Les joueurs expérimentés (27–30 ans) se distinguent clairement par leur proximité avec la catégorie "Excellent (16+)", suggérant que cette tranche d'âge correspond au pic de performance dans la carrière des joueurs contrairement aux vétérans (31–40 ans) qui, eux, se trouvent du côté opposé du plan factoriel, proches de la catégorie "Faible (0–5 buts)" ce qui traduit une diminution naturelle et logique des performances avec l'âge.

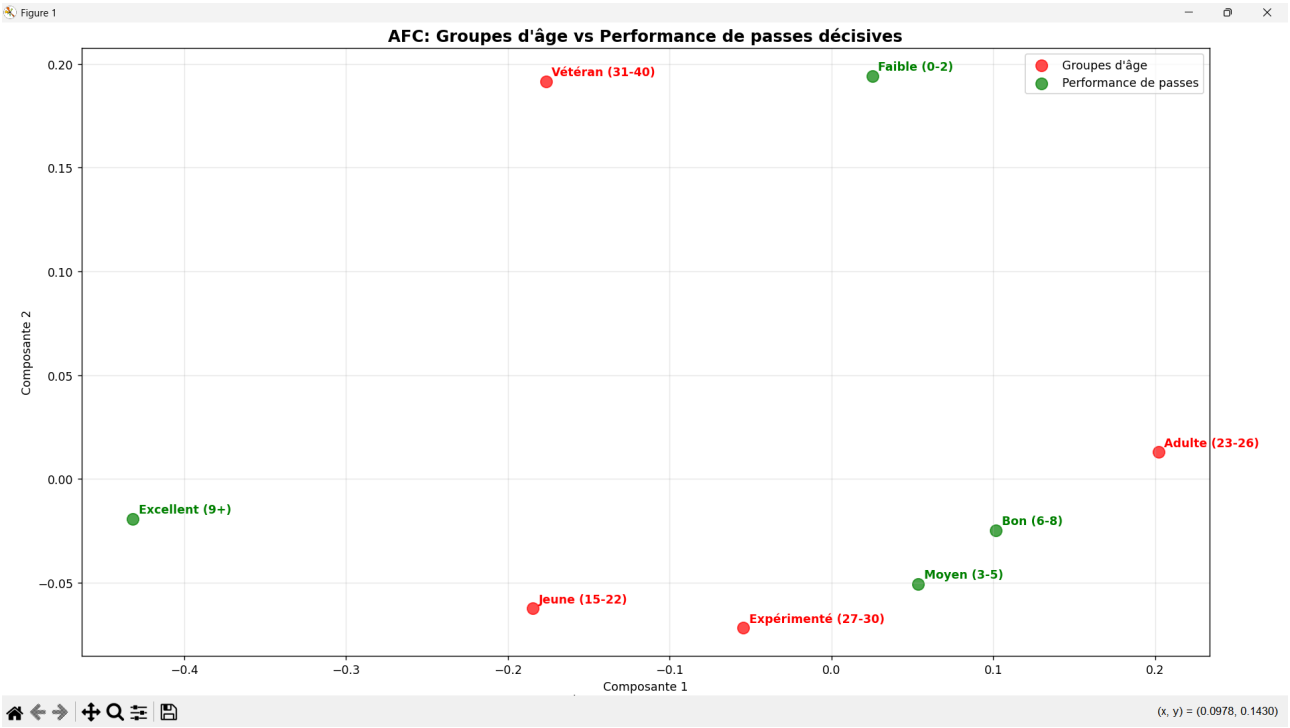

FIGURE 8 –

Interprétation

Dans la figure ci-desssus, on peut observer une tendance

Analyse mixte des données : ACM

Définition

L'Analyse des Correspondances Multiples (ACM) est une technique statistique utilisée pour analyser des données qualitatives comportant plusieurs variables. Elle permet de représenter graphiquement les relations entre les différentes catégories des variables, en les projetant dans un espace à deux dimensions.

L'ACM est particulièrement utile pour identifier des associations entre les catégories, détecter des groupes similaires et comprendre la structure sous-jacente des données qualitatives. En réduisant la complexité des données, l'ACM facilite l'interprétation et la communication des résultats.

Interprétation

Conclusion

