SAISON 2020/2021 · BIG FIVE

# CINQ PROFILS DE TIREURS

Classification non supervisée des statistiques de tir des joueurs des cinq grands championnats européens.



**2 822**
JOUEURS DANS LA BASE

**811**
JOUEURS RETENUS

**5**
PROFILS

**75 %**
VARIANCE · 3 AXES

AUTEUR

**Nylsevens Lalanne**
Data Analyst

SOURCE

**fbref.com · StatsBomb**

# SOMMAIRE

Deux temps : comprendre d'abord ce que contiennent les données, puis laisser les algorithmes regrouper les joueurs qui tirent de la même façon.

PARTIE 1

## Découvrir les données


01 Introduction 03

02 Les données 04

03 Premier portrait 05

04 Le poste change tout 06


PARTIE 2

## Classer sans étiquette


05 Combien de groupes ? 08

06 L'ACP 09

07 Lire le plan factoriel 10

08 Les cinq profils 11

09 Valider les profils 13

10 Conclusion 14

# 01 INTRODUCTION

Dans le football, la donnée sert d'abord à **faire progresser les joueurs**, à mettre en lumière des comportements et à construire une tactique. On y suit la vitesse moyenne, les kilomètres parcourus, le nombre de tirs, de fautes, de corners ou de duels perdus.

Grâce à elle, un club peut dresser rapidement une liste courte de joueurs dont les statistiques collent au profil recherché, ou analyser les matchs et les séquences vidéo d'un groupe de joueurs. Il y gagne du temps et de l'argent.

C'est sur le **recrutement et la fidélisation** des joueurs que cette abondance soudaine de données a eu le plus d'impact.

L'objectif de ce projet est d'analyser une base de joueurs centrée sur **les tirs et les probabilités de marquer** pendant la saison 2020/2021.

Avec deux méthodes de classification, les k-means et la classification ascendante hiérarchique, on cherche **sur quels critères les joueurs se regroupent**. Des analyses univariées et bivariées complètent la lecture de cette saison dans les cinq grands championnats.

QUESTION DE DÉPART

## Peut-on reconnaître des familles de tireurs à partir de leurs seules statistiques ?

Q1
**Quelles statistiques distinguent vraiment les tireurs ?**
Pages 05 à 08

Q2
**Les groupes obtenus recoupent-ils les postes ?**
Page 09

Q3
**Dépendent-ils du championnat ?**
Page 09

SOURCE
### fbref.com
Données StatsBomb et Data Sports Group

PÉRIMÈTRE
### Big Five
Allemagne, Angleterre, Espagne, France, Italie

MÉTHODES
### CAH · k-means
Avec une ACP pour visualiser les groupes

# 02 LES DONNÉES

La base recense **2 822 joueurs** et leurs statistiques de tir dans les cinq grands championnats. Ses **18 variables** se rangent en trois familles.

| | | |
|---|---|---|
| **Description** GROUPE 1 | Player · Nation · Pos · Squad · Comp · Age · Born · 90s | Les variables qualitatives (championnat, pays, poste), plus l'âge et « 90s », le nombre de matchs complets de 90 minutes joués. |
| **Standard** GROUPE 2 | Gls · Sh · SoT · SoT% · Sh/90 · SoT/90 · G/Sh · G/SoT · Dist | Les données de tir : buts, tirs, tirs cadrés, distance moyenne de tir et ratios de buts par tir. |
| **Expected** GROUPE 3 | xG · npxG · npxG/Sh · G-xG · np:G-xG | Les probabilités qu'un tir finisse au fond. Elles tiennent compte du lieu du tir, de la partie du corps utilisée et de l'action qui l'a amené : contre-attaque, phase de possession… |

## Du fichier brut à l'échantillon

FILTRAGE

**2 822** JOUEURS · 18 VARIABLES

↓ AU MOINS 20 MATCHS COMPLETS JOUÉS (90s ≥ 20)
↓ GARDIENS RETIRÉS, ILS FAUSSERAIENT LES STATISTIQUES DE TIR

**811** JOUEURS RETENUS · 29 % DE LA BASE

## Nettoyage

VALEURS MANQUANTES

**24 NAN**

Seule la variable **G/SoT** (buts par tir cadré) a des valeurs manquantes. Les 24 joueurs concernés n'ont ni but ni tir cadré à leur actif : le ratio ne peut pas se calculer. Leur G/SoT vaut donc logiquement **0**, et c'est la valeur qui remplace les NaN.

> **CE QU'ON EN TIRE**
> Le filtre ne garde que 29 % de la base. Les profils décrivent donc des **joueurs réguliers** : pour un remplaçant qui n'a tiré que quelques fois, des ratios comme les buts par tir cadré seraient trop instables pour être interprétés.

# 03 PREMIER PORTRAIT

Avant de classer qui que ce soit, trois questions simples : **d'où viennent les joueurs**, à quels postes jouent-ils, et l'âge rend-il plus efficace devant le but ?

## Joueurs retenus par championnat

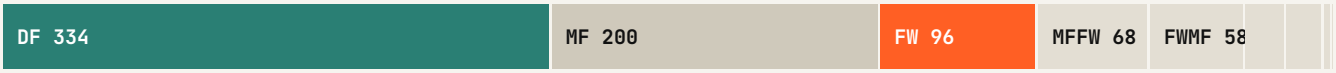
811 JOUEURS · ≥ 20 MATCHS




La Bundesliga pèse logiquement moins : c'est le seul championnat à 18 clubs, contre 20 ailleurs. La Premier League compte la plus grande part de joueurs ayant disputé au moins 20 matchs. Deux hypothèses : c'est de loin le championnat le plus riche, capable d'entretenir des effectifs plus fournis, et le seul à disputer trois compétitions nationales (le championnat et deux coupes, contre une coupe ailleurs). D'où le besoin d'effectifs « riches », qui laissent les titulaires habituels se reposer.

## Répartition par poste

CODES FBREF

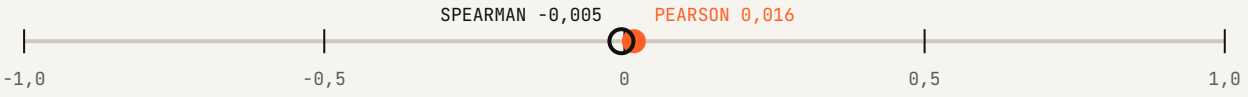



DF défenseur · MF milieu · FW attaquant · les doubles codes (MFFW, FWMF…) désignent les joueurs polyvalents, poste principal en premier.

## L'âge rend-il plus efficace ?

BUTS PAR TIR CADRÉ × ÂGE

L'idée : vérifier si l'expérience accumulée au fil d'une carrière aide à mieux convertir ses tirs.




Le coefficient de Pearson, un peu supérieur à celui de Spearman, suggère au mieux une relation linéaire positive **très faible**. Mais la p-valeur (**0,66**) est largement supérieure au seuil de 5 % : impossible d'affirmer qu'il existe une relation significative entre l'âge et l'efficacité.

> CE QU'ON EN TIRE
>
> **L'âge n'est pas un critère de finition.** Un recruteur ne peut pas compter sur l'expérience pour améliorer la conversion d'un joueur. Autre point de vigilance : la Premier League fournit plus de joueurs réguliers que les autres ligues, ce qu'il faut garder en tête en lisant les profils championnat par championnat.

# 04 LE POSTE CHANGE TOUT

On attend d'un attaquant qu'il soit plus précis qu'un défenseur, et d'un milieu ou d'un défenseur qu'il tente sa chance de plus loin. **Le poste influence-t-il vraiment la façon de tirer ?**

## Ce qui va ensemble

CORRÉLATIONS DE PEARSON

CORRÉLATION FORTE +

Assez logiquement, les **buts (Gls)** évoluent avec les **buts attendus (xG)**, les **tirs cadrés (SoT)** et le **nombre de tirs par 90 minutes (Sh/90)**.

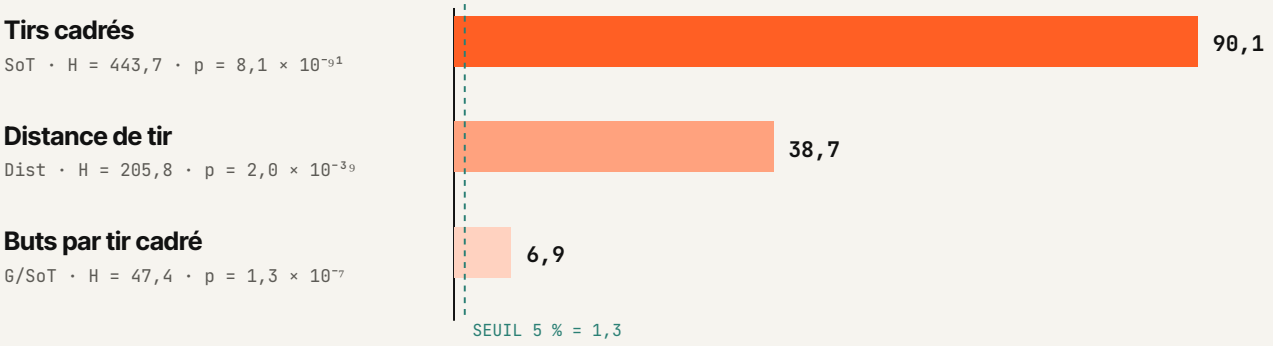
CORRÉLATION FORTE -

La **distance de tir** s'oppose au **xG par tir hors penalty (npxG/Sh)**, qui mesure la capacité d'un joueur à frapper en bonne position. Plus on tire de loin, moins on se trouve en position de marquer.

## Le poste, testé variable par variable

KRUSKAL-WALLIS · $-\log_{10}(p)$

Trois variables sont testées : **SoT** (tirs cadrés), **G/SoT** (buts par tir cadré) et **Dist** (distance au but au moment du tir). Pour chacune, l'hypothèse nulle dit que les médianes sont égales quel que soit le poste. Plus la barre est longue, plus la p-valeur est petite et plus l'effet du poste est net.




À RETENIR

**Pour les trois variables, on rejette l'hypothèse de médianes égales au seuil alpha de 5 %. La position d'un joueur compte pour sa distance de tir, sa capacité à cadrer et sa capacité à convertir en but.**

CE QU'ON EN TIRE

L'effet du poste est très inégal. Il est massif sur le **volume** (H = 444 pour les tirs cadrés) et bien plus modeste sur l'**efficacité** (H = 47 pour les buts par tir cadré). Le poste dit surtout combien un joueur tire, beaucoup moins comment il convertit : la finition reste en grande partie individuelle, et c'est là qu'un recruteur peut faire la différence.

PARTIE 2 · CLASSIFICATION NON SUPERVISÉE

# CLASSER SANS ÉTIQUETTE

En apprentissage non supervisé, les données **n'ont pas d'étiquette**. L'algorithme doit découvrir seul les structures cachées qu'elles contiennent.

Sans étiquette, l'algorithme ne peut pas calculer de façon certaine un score de réussite. C'est cette absence d'annotation qui distingue l'apprentissage non supervisé de l'apprentissage supervisé. Il en existe deux grandes familles, et ce rapport utilise une méthode de chacune.

PARTITIONNEMENT

## K-MEANS

On fixe le nombre de groupes, puis l'algorithme répartit les joueurs autour de centres. La même famille compte les k-médoïdes.

REGROUPEMENT HIÉRARCHIQUE

## CAH

On fusionne pas à pas les joueurs les plus proches, jusqu'à obtenir un arbre complet que l'on coupe au bon niveau.

# 05 COMBIEN DE GROUPES ?

Les deux méthodes ont besoin d'un nombre de groupes. On le cherche de deux façons indépendantes, et **les deux pointent vers cinq**.

## Dendrogramme

CAH · WARD

La CAH cherche des groupes dont les membres se ressemblent le plus possible (homogénéité intra-classe), tout en étant aussi différents que possible les uns des autres (hétérogénéité inter-classe). Avec la méthode de Ward, fondée sur le carré des distances, l'arbre suggère un découpage en **5 groupes**.

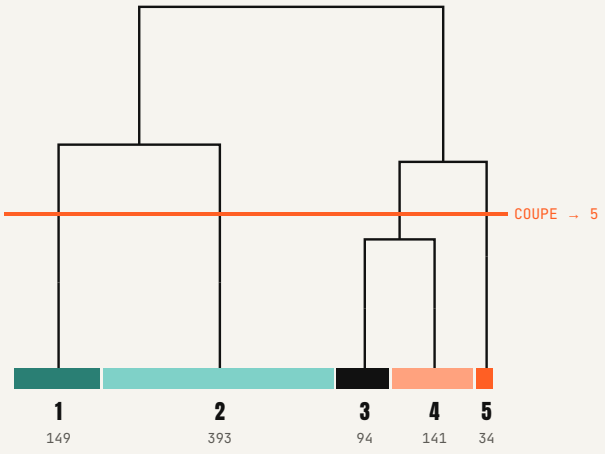
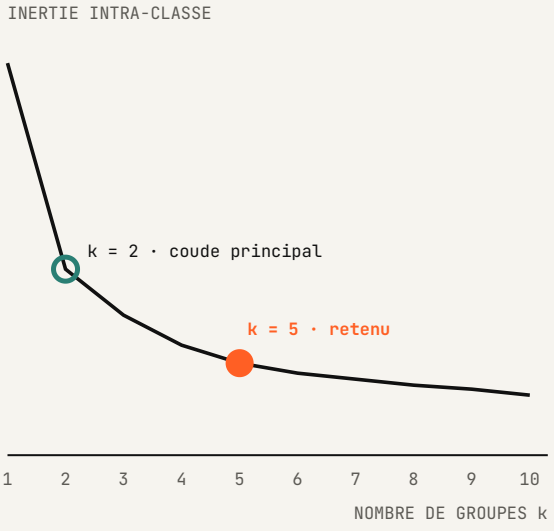



Schéma simplifié · largeur des feuilles proportionnelle à l'effectif de chaque groupe.

## Méthode du coude

K-MEANS

Les k-means ne trouvent pas seuls le nombre optimal de groupes : il faut le fixer en amont. L'algorithme renvoie l'inertie intra-classe, qui mesure la dispersion à l'intérieur des groupes. Plus elle est proche de 0, meilleure est la partition, mais elle vaudrait 0 avec un groupe par joueur. On cherche donc **un point d'inflexion** sur la courbe.




Courbe retracée d'après le notebook.

**À RETENIR**

**Le coude le plus marqué est à 2 groupes, un chiffre sans intérêt pour décrire des profils de joueurs. Le point d'inflexion suivant, à 5, concorde avec le dendrogramme : on retient 5 groupes.**

**CE QU'ON EN TIRE**

Deux groupes reviendraient, pour l'essentiel, à séparer les joueurs défensifs des joueurs offensifs, une distinction que le poste donne déjà. Passer à cinq groupes, c'est ce qui fait apparaître **des manières de tirer différentes à l'intérieur de chaque camp**.

# 06 L'ACP

L'analyse en composantes principales transforme des variables très corrélées en quelques **composantes principales**, non corrélées entre elles. Elle résume l'information d'une large base de données en un petit nombre de variables synthétiques.

## Combien d'axes garder ?

ÉBOULIS · KAISER-GUTTMAN

L'éboulis des valeurs propres et la règle de Kaiser-Guttman recommandent de conserver deux axes. Ils n'expliquent pourtant que **62 % de la variance**. Pour en retenir davantage, on garde trois axes, qui en expliquent **75 %**.




PART DE LA VARIANCE EXPLIQUÉE, CUMULÉE

## Ce que mesure chaque axe

COS² · PART DE LA VARIABLE PORTÉE PAR L'AXE

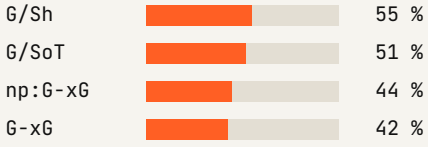
AXE 1

### Le volume offensif



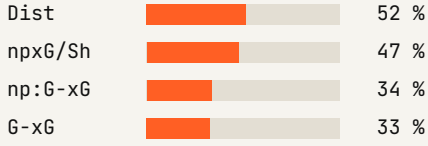

AXE 2

### L'efficacité




AXE 3

### La distance et la qualité des occasions




## CAH ou k-means ?

PREMIER PLAN FACTORIEL

On projette les joueurs sur les trois premiers plans factoriels, colorés d'abord selon les groupes du dendrogramme, puis selon ceux des k-means, pour voir quelle méthode distingue le mieux les groupes.

CE QU'ON EN TIRE

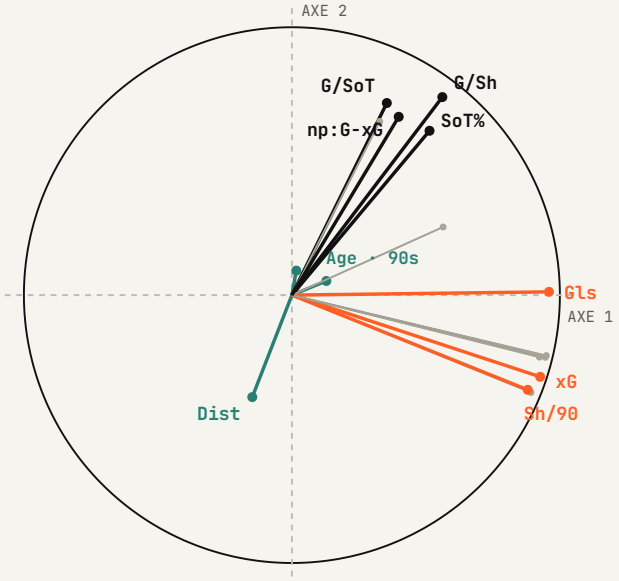
Volume et efficacité sont portés par **deux axes distincts**, donc largement indépendants : tirer beaucoup ne garantit pas de bien convertir. Les buts se lisent à 92 % sur l'axe du volume : sur une saison, un joueur marque beaucoup d'abord parce qu'il tire beaucoup. L'efficacité est un second critère, à évaluer à part.

CHOIX RETENU

**Sur le premier plan, les groupes de la CAH se séparent un peu plus nettement que ceux des k-means. C'est donc la classification issue du dendrogramme qui est retenue pour la suite.**

# 07 LIRE LE PLAN FACTORIEL

Le cercle des corrélations projette chaque variable sur le premier plan. Une variable proche du bord est **bien représentée** : elle est fortement corrélée aux deux axes qui forment ce plan.




Flèches reconstruites à partir des cos² du notebook.

BIEN ET MAL REPRÉSENTÉES

**Age** et **90s** sont très mal représentées sur ce plan. À l'inverse, **xG**, **G/Sh**, **Sh/90** ou **Gls** atteignent presque le bord du cercle. Gls se confond avec l'axe 1, signe de son poids dans la construction de cet axe.

LE CAS DE LA DISTANCE

**Dist** semble fortement corrélée, en négatif, aux ratios de buts par tir (cadré ou non) et aux écarts entre buts et buts attendus. Elle forme en revanche un angle presque droit avec le nombre de tirs par 90 minutes : **les deux ne sont pas corrélées**.

### Qualité de représentation

COS²

Le cos² d'une variable sur un axe est le carré de sa corrélation avec cet axe. Additionné sur toutes les composantes, il vaut 1, soit 100 %. Gls est ainsi représentée à **92 %** par l'axe 1, ce qui rejoint la lecture du cercle.

### Groupes et variables

INDIVIDUS × CERCLE

Mettre le nuage des joueurs en regard du cercle montre comment les groupes se sont formés. Le groupe 2 est tiré par la **distance**, cohérent avec sa médiane de distance, la plus élevée de tous.

### Contribution des individus

AXE 1

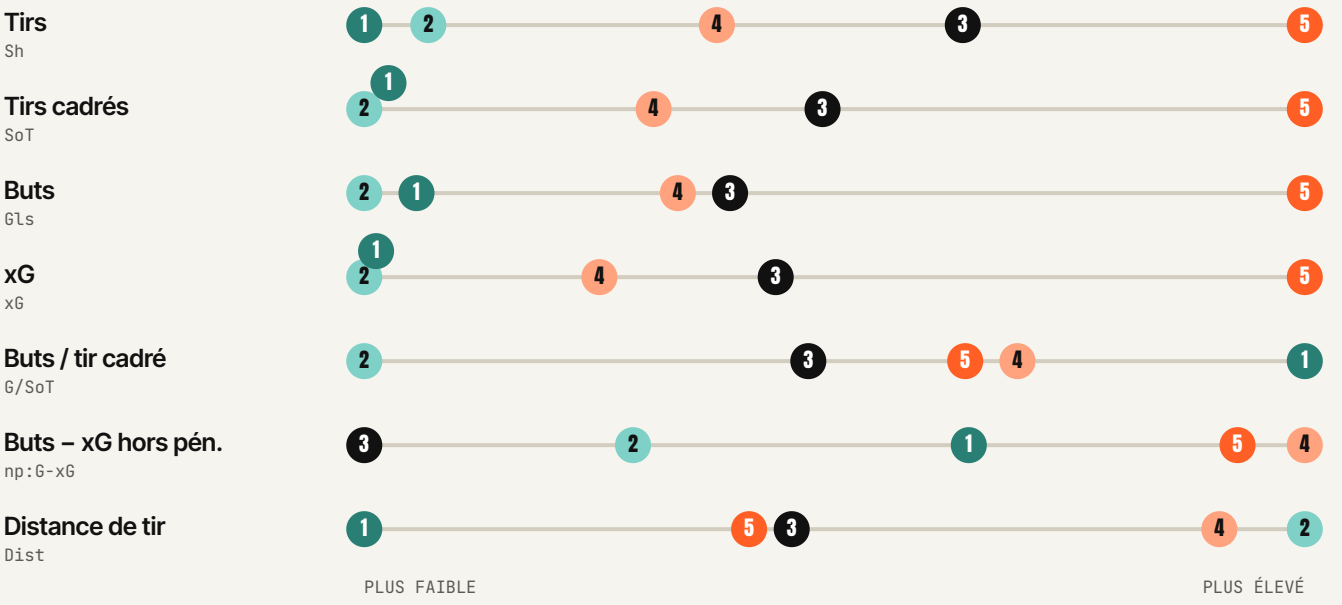
Les contributions désignent les joueurs qui pèsent le plus dans chaque axe. **Les dix premiers contributeurs de l'axe 1 appartiennent tous au groupe 5**, celui des meilleurs attaquants. C'est logique : leurs statistiques dépassent la moyenne sur de nombreuses variables, dans des données entièrement centrées sur le tir.

# 08 LES CINQ PROFILS

Pour définir chaque groupe, on affiche les boxplots de chaque variable, cluster par cluster, et on retient les plus parlantes. **Cinq profils de tireurs se dessinent**, des défenseurs opportunistes aux attaquants d'élite.

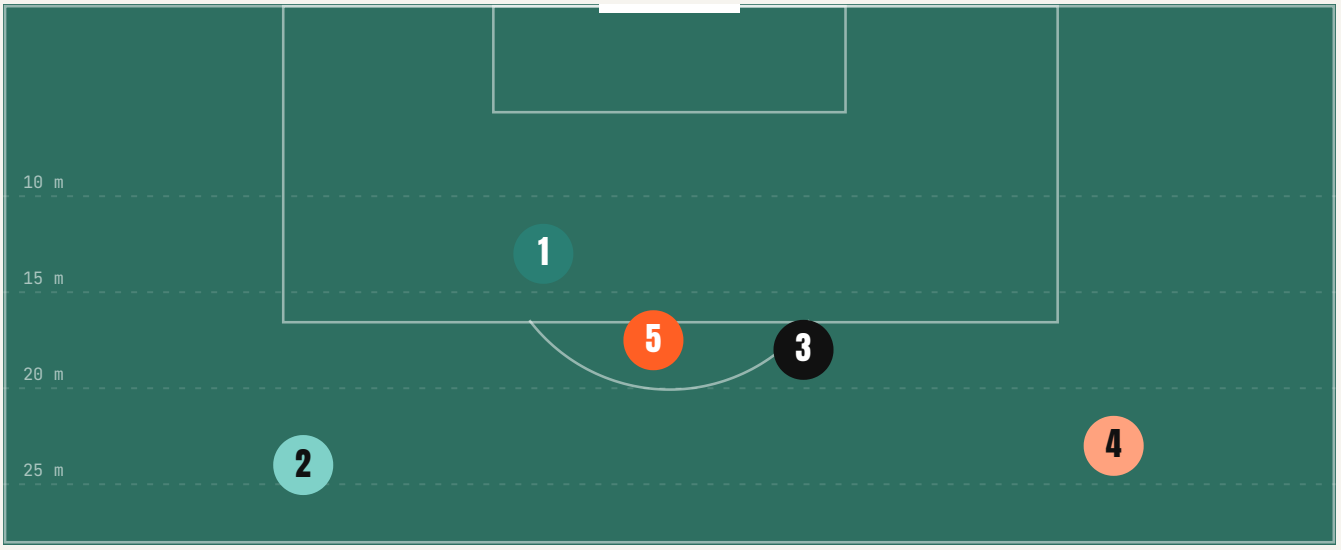
## Où se place chaque profil

MÉDIANE DE CHAQUE CLUSTER, DU PLUS FAIBLE AU PLUS ÉLEVÉ




## D'où tirent-ils ?

DISTANCE MÉDIANE AU BUT




Médianes lues sur les boxplots du notebook (ordres de grandeur) · placement latéral illustratif.

### 1 Défenseurs opportunistes

DÉFENSEURS · 149 JOUEURS

Surtout des défenseurs, et des milieux assez défensifs. Rarement en bonne position, ils tirent peu. Ce qui les distingue du groupe 2, c'est leur **opportunisme** : près du but, sur corner ou coup franc direct, ils savent être précis, avec **la meilleure médiane de buts par tir cadré**. Sans avoir le profil du buteur type, leur médiane de buts – xG reste positive.

### 2 Milieux & latéraux défensifs

MILIEUX · LATÉRAUX · 393 JOUEURS

Des milieux de terrain et des latéraux à profil défensif. Ils tirent peu et cadrent peu, ce qui s'explique par la distance depuis laquelle ils tentent leur chance : **la médiane la plus haute**. Or plus on tire de loin, plus il est difficile de marquer. Sans surprise, ce groupe a **la plus faible médiane de buts par tir cadré**.

### 3 Attaquants de métier

ATTAQUANTS · 94 JOUEURS

Les purs attaquants. Ils tirent et cadrent plus que tous les autres, à l'exception du groupe 5. Souvent en position préférentielle, ils ont **un xG élevé**. Ils tirent **de près**, car ils passent l'essentiel de leur temps dans la surface adverse, à l'affût des centres et des passes en retrait.

### 4 Offensifs excentrés

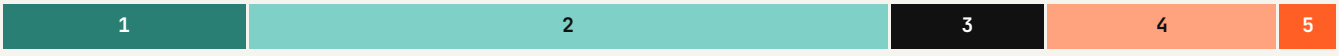
AILIERS · MILIEUX OFF. · 141 JOUEURS

Des profils offensifs, voire très offensifs, qui gravitent autour de l'attaquant : milieux offensifs, ailiers, latéraux offensifs. Ils tirent et cadrent régulièrement, mais moins que les groupes 3 et 5, et **de plus loin**, car ils jouent plus excentrés. Leur **excellent écart buts – xG** montre pourtant une vraie efficacité.

### 5 Attaquants d'élite

CLASSE MONDIALE · 34 JOUEURS

Proches du groupe 3, mais ultra-performants. Deux variables les distinguent : le **xG** et l'**écart buts – xG**. Ils se retrouvent plus souvent que les autres en position idéale, et dépassent aussi les attentes statistiques. Leur nombre de **buts** marque nettement l'écart avec les autres groupes.




TAILLE DES GROUPES · 811 JOUEURS 149 · 393 · 94 · 141 · 34

**CE QU'ON EN TIRE**

Deux lectures à nuancer. Le ratio record du groupe 1 repose sur **très peu de tirs cadrés** (environ 4 par joueur) : il est fragile. Le groupe 3 se crée les meilleures occasions après le groupe 5 mais **marque un peu moins que son xG**, quand le groupe 4 fait l'inverse en tirant de plus loin. Sur une seule saison, ces écarts tiennent en partie à la variance : ils sont à confirmer sur plusieurs saisons avant d'en faire un critère de recrutement.

# 09 VALIDER LES PROFILS

Les profils ont été construits **sans connaître le poste ni le championnat**. S'ils ont un sens, ils devraient s'aligner sur le premier, et pas forcément sur le second.

## Postes × profils

EFFECTIFS · COULEUR = PART DE LA LIGNE

| | DF | DFFW | DFMF | FW | FWDF | FWMF | MF | MFDF | MFFW | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 120 | 1 | 3 | 0 | 0 | 0 | 19 | 4 | 2 | 149 |
| 2 | 201 | 2 | 18 | 3 | 0 | 3 | 133 | 14 | 19 | 393 |
| 3 | 1 | 0 | 1 | 43 | 1 | 27 | 8 | 0 | 13 | 94 |
| 4 | 12 | 2 | 1 | 24 | 1 | 21 | 39 | 7 | 34 | 141 |
| 5 | 0 | 0 | 0 | 26 | 0 | 7 | 1 | 0 | 0 | 34 |

| $X^2$ | V DE CRAMÉR | P-VALEUR |
|---|---|---|
| **664,3** | **0,45** | **≈ 0** |
| forte dépendance | liaison forte | $2{,}9 \times 10^{-119}$, très inférieure à 0,05 |

Le tableau confirme la caractérisation : **les profils défensifs se concentrent dans les groupes 1 et 2**, les profils offensifs dans les groupes 3 et 4, et les attaquants de classe mondiale forment le groupe 5. Le lien est fort, mais pas total : **39 milieux et 12 défenseurs** se retrouvent dans le groupe offensif 4, et 19 milieux dans le groupe 1. Le profil de tir capte des rôles que l'étiquette de poste ne dit pas, comme les latéraux offensifs.

## Championnats × profils

EFFECTIFS

| | BUNDESLIGA | PREMIER LEAGUE | LA LIGA | LIGUE 1 | SERIE A | |
|---|---|---|---|---|---|---|
| 1 | 22 | 27 | 40 | 30 | 30 | 149 |
| 2 | 54 | 93 | 81 | 88 | 77 | 393 |
| 3 | 15 | 28 | 12 | 26 | 13 | 94 |
| 4 | 24 | 27 | 33 | 22 | 35 | 141 |
| 5 | 5 | 7 | 7 | 4 | 11 | 34 |

| $X^2$ | V DE CRAMÉR | P-VALEUR |
|---|---|---|
| **22,0** | **0,08** | **0,14** |
| faible | liaison négligeable | > 0,05 : non significatif |

Les championnats ne séparent pas les joueurs (p = 0,14, au-dessus de 5 %) : seule la Bundesliga semble un peu sous-représentée dans le groupe 2. Conséquence pratique : **la même grille de profils permet de comparer des joueurs d'une ligue à l'autre**.

# 10 CONCLUSION

Comparer plusieurs méthodes de classification non supervisée sur ce jeu de données s'est révélé **particulièrement instructif**. Les deux approches ont permis de créer des groupes de joueurs cohérents, fondés sur leurs statistiques et leurs particularités de tir.

La classification non supervisée peut donc intéresser les cellules de recrutement des clubs : elle les aide à **affiner leurs recherches** et à se concentrer directement sur les joueurs qui répondent à leurs attentes, notamment en qualité de tir et de finition.

## Les cinq profils

CAH · 811 JOUEURS

| 01 | 02 | 03 | 04 | 05 |
|---|---|---|---|---|
| **Défenseurs opportunistes** | **Milieux & latéraux défensifs** | **Attaquants de métier** | **Offensifs excentrés** | **Attaquants d'élite** |
| 149 joueurs | 393 joueurs | 94 joueurs | 141 joueurs | 34 joueurs |

## Réponses aux trois questions

**Q1** Le **volume de tirs** et la **qualité des occasions** séparent le plus les joueurs ; l'efficacité et la distance affinent les profils.

**Q2** **Oui, largement** (V = 0,45), mais pas totalement : le profil révèle des rôles que le poste officiel ne dit pas.

**Q3** **Non** (p = 0,14) : un profil de tireur se retrouve dans toutes les ligues.

## Limites et pistes

**Une seule saison.** Les écarts entre buts et xG varient d'une année à l'autre : il faudrait vérifier que chaque joueur garde son profil sur plusieurs saisons.

**Petits volumes.** Les ratios des défenseurs reposent sur quelques tirs cadrés seulement.

**Joueurs réguliers uniquement.** Les moins de 20 matchs sont exclus : les jeunes qui percent en cours de saison échappent à l'analyse.