UNIVERSITÉ RENNES 2

DIGISPORT GRADUATE SCHOOL

FOUILLE DE DONNÉES

Étienne Brière
Raphael Peus
Mattis Laperrousaz

2023/2024

Contexte :
Le jeu de données contient des informations sur les performances de différents joueurs de football des 5 ligues d'Europe au cours de la saison 2019-2020.

Variables du jeu de données :

- Player : Nom d'usage du joueur.
- Name : Nom complet du joueur.
- Club : Club de football pour lequel le joueur évolue durant la saison concernée.
- Âge : Âge du joueur.
- Main Position : Position principale sur le terrain du joueur.
- Apps : Nombre d'apparitions du joueur durant la saison.
- Mins : Minutes jouées par le joueur au cours de la saison.
- Goals : Nombre de buts marqués par le joueur.
- Assists : Nombre de passes décisives réalisées par le joueur.
- Shot per Game : Moyenne de tirs par match effectués par le joueur.
- Key Pass : Nombre moyen de passes clés (passes menant à une occasion de but) par match effectuées par le joueur.
- Drb : Moyenne de dribbles réussis par match par le joueur.
- Fouled : Moyenne de fautes subies par match par le joueur.
- Rating : Évaluation globale des performances du joueur sur la saison.

Objectif : Répartir les joueurs en différents groupes et décrire les principales caractéristiques de chaque groupe

Étapes à suivre :

1. Comprendre et pré-traiter les données
2. Classifier les joueurs en utilisant les différents algorithmes abordés en cours
3. Comparer les résultats obtenus avec les différentes méthodes
4. Conclure vis à vis des choix effectués
5. Suggérez une description des profils de joueurs. Il convient en particulier d'expliciter la relation entre les profils des joueurs, le club pour lequel ils évoluent, et leur position sur le terrain.

I.Pré-traitement des données

L'ensemble des données a été traité à l'aide des méthodes vu en cours. Dans un premier temps, la variable « Mins » (minutes jouées par le joueur au cours de la saison) a été convertie en heure afin de réduire l'amplitude des différences entre les observations. Dans un second temps, uniquement les variables quantitatives ont été retenues afin de réaliser une analyse générale des variables influençant la performance, pour le moment les clubs et les postes des joueurs étaient mis de côté.

En réalisant les diagrammes en boîte de chaque variable de ces nouvelles données (figure 1), nous remarquons que les données ne sont pas standardisées. De ce fait, l'ensemble des données

a subi une standardisation afin de pouvoir les traiter plus facilement par la suite. La figure 2 représente la nouvelle distribution des données après standardisation.

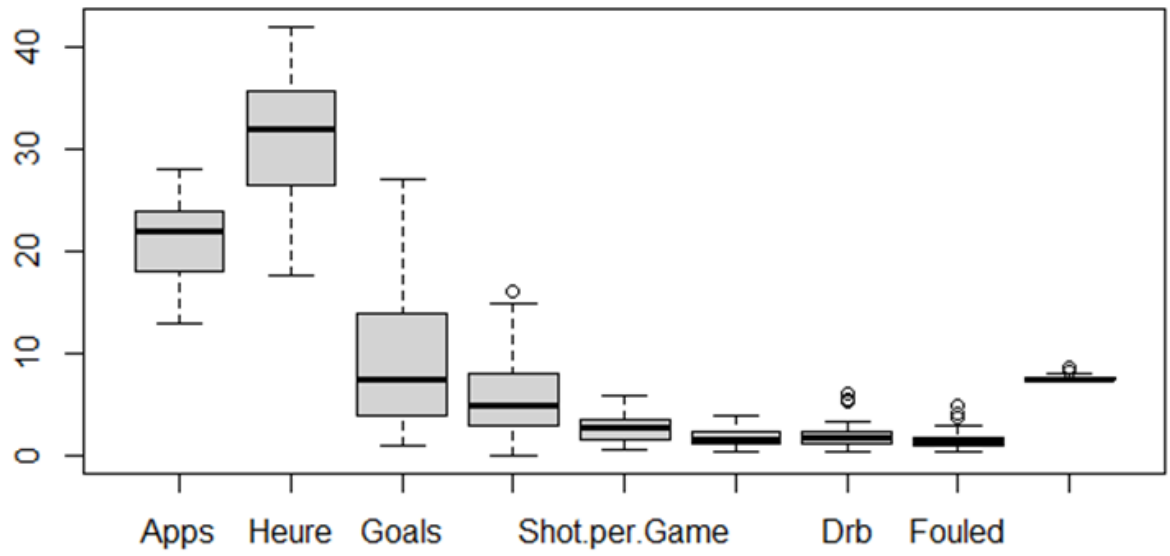

Figure 1 : Diagrammes en boîte de l'ensemble des données

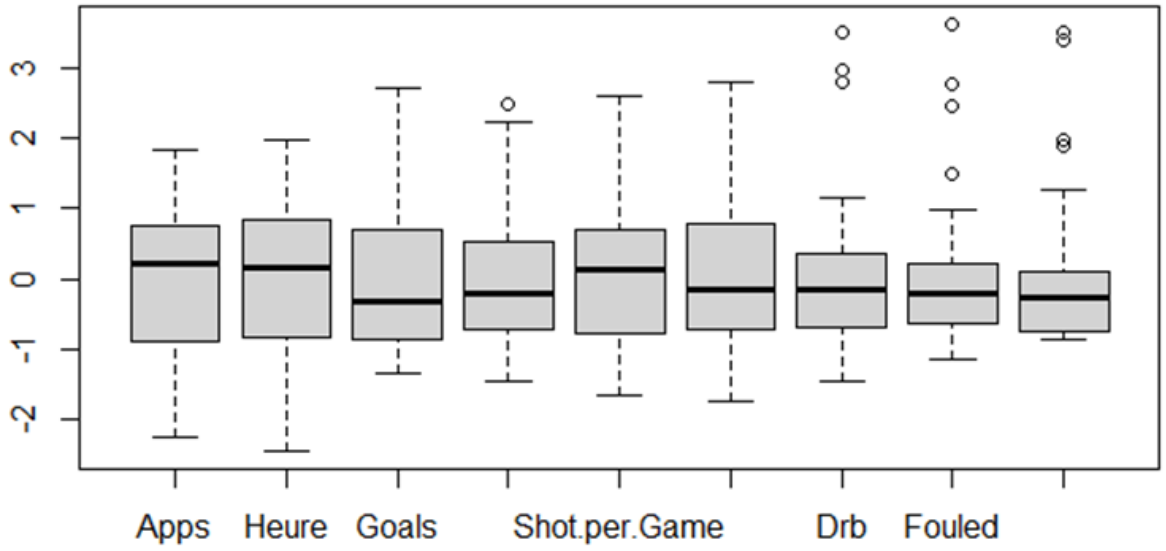

Figure 2 : Diagrammes en boîte de l'ensemble des données standardisées

II. Analyse générale

Une fois les données standardisées, une analyse en composantes principales (ACP) a été appliquée. La représentation des données dans le plan factoriel, qui est défini par les deux premiers axes, est peu satisfaisante. En effet, les deux premiers axes principaux expliquent seulement 50,13% de la variabilité totale des données (la part d'inertie expliquée sur la figure 3). Le dimension 1 (29,28%) regroupe une part légèrement plus importante que la dimension 2 (20,85%).

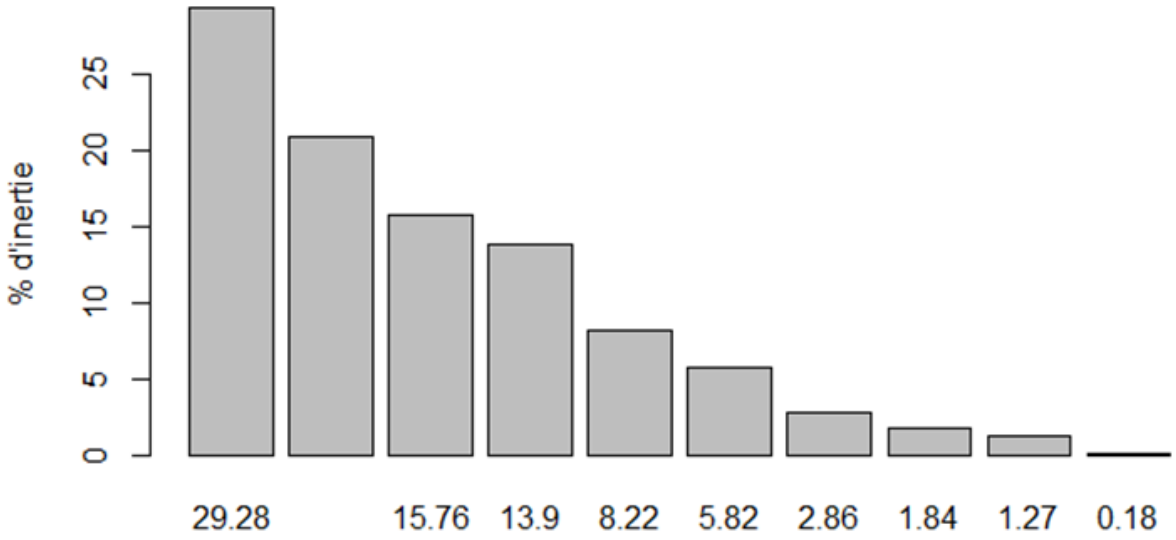

Figure 3 : Valeurs propres en % d'inertie

A partir du graphique représentant le cercle de corrélation (figure 4), il apparaît que le nombre d'apparition du joueur durant la saison (« Apps ») et que les heures jouées par le joueur durant la saison (« Heure ») n'est pas corrélé à l'ensemble des autres variables qui semblent déterminant pour la performance en football. Concernant les autres variables, il semblerait que l'âge du joueur (« Age ») soit positivement corrélé avec des variables déterminantes de la performance telles que les buts marqués (« Goals »), les tirs par match (« Shot per Game ») ou encore les passe décisives (« Key Pass »). Les passes (« Assists »), bien que moins influentes sur la composante principale, sont positivement corrélées avec les performances globales (« Rating »). De plus, les dribbles (« Drb ») et les fautes subies (« Fouled ») affichent également une corrélation positive, ce qui semble cohérent avec l'idée qu'un joueur est plus sujet aux fautes lorsqu'il dribble.

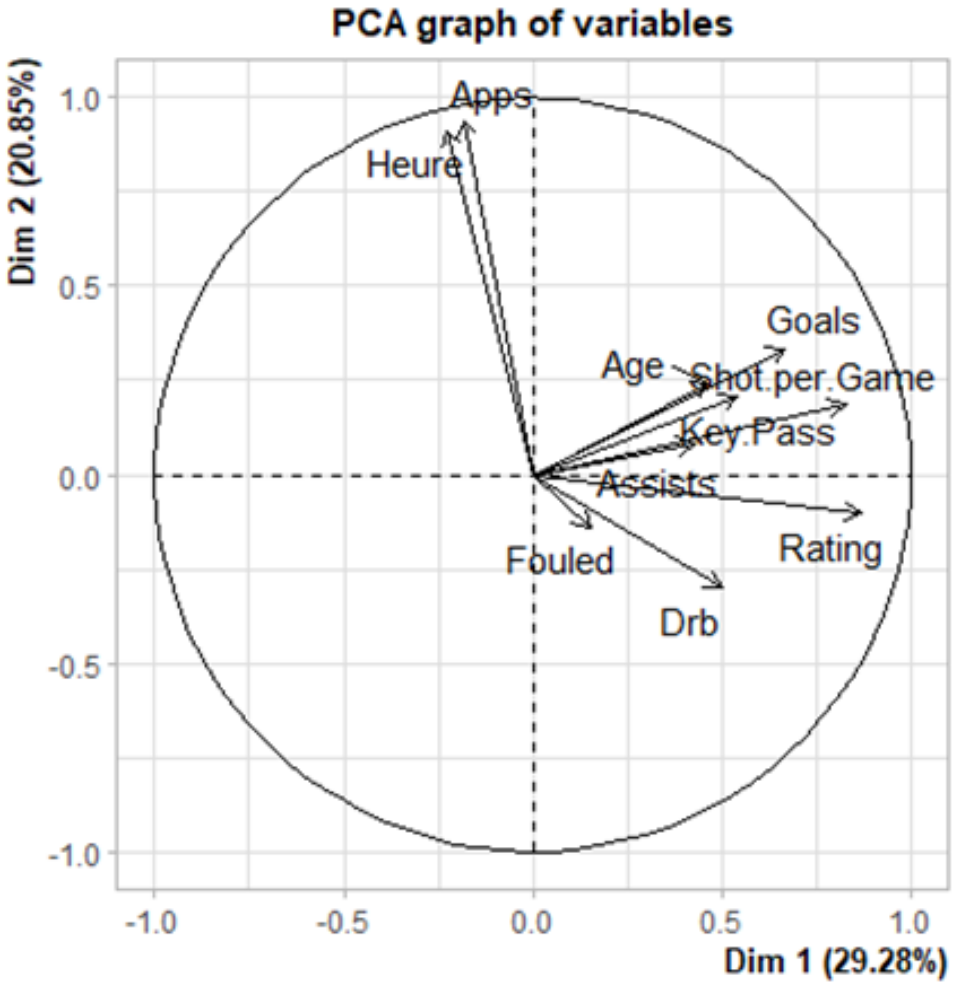

Figure 4 : Cercle de corrélation des variables

Sur la première dimension (Dim 1), des variables telles que "Goals", "Assists", "Key Passes", et "Shot per Game" se projettent de manière significative. Cet axe reflète principalement la contribution offensive des joueurs, indiquant que ceux qui se distinguent sur cet axe tendent à avoir des performances offensives supérieures. Sur la deuxième dimension (Dim 2), l'analyse met en évidence l'influence du nombre d'apparitions et des heures jouées, capturant ainsi la régularité et la durée de participation des joueurs aux matchs. Les variations le long de cet axe montrent des différences sur l'endurance des joueurs au cours de la saison, malgré que ces variables ne soient pas directement corrélées avec l'évaluation de la performance globale des joueurs, elles ne peuvent être classifiées de totalement non-déterminantes à la performance en football.

Dans l'analyse factorielle des individus (figure 5), La distribution des joueurs révèle que la majorité tend à se regrouper autour de la moyenne, bien qu'il y ait des exceptions notables. Par exemple, Ndidi se démarque par des statistiques offensives relativement basses. A l'opposé, deux cas particulièrement distincts sont ceux de Messi, qui, avec un temps de jeu moyen, affiche des statistiques offensives bien supérieures à celles des autres, et Neymar, qui, malgré un temps de jeu réduit en raison de blessures, présente des performances offensives nettement au-dessus de la moyenne. Ces observations soulignent les contributions uniques de ces joueurs à leurs équipes, malgré des contextes de jeu différents.

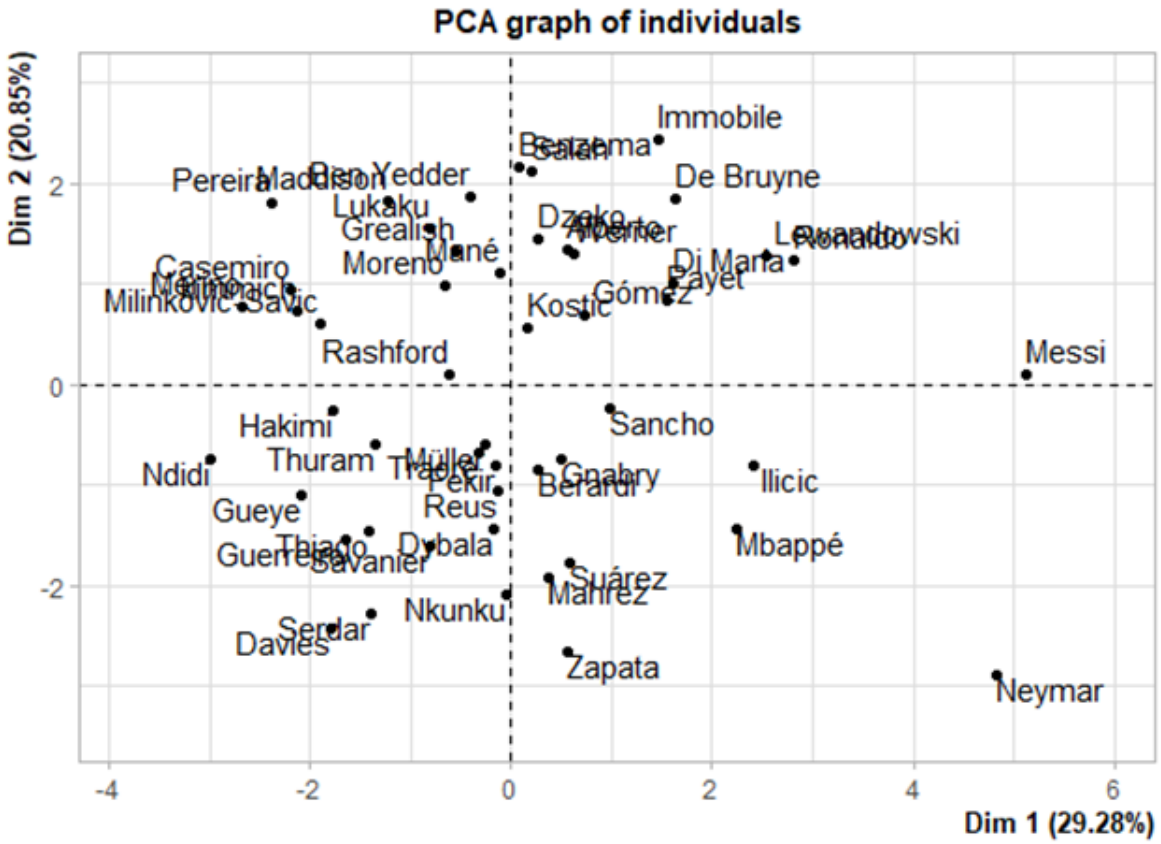

Figure 5 : Plan factoriel des individus

III. Classification des joueurs

Une première représentation (figure 6) des données a été d'exprimer les données d'une variable en fonction d'une autre (par paire). Malheureusement, sur cette représentation il est difficile d'observer une distinction de groupes.

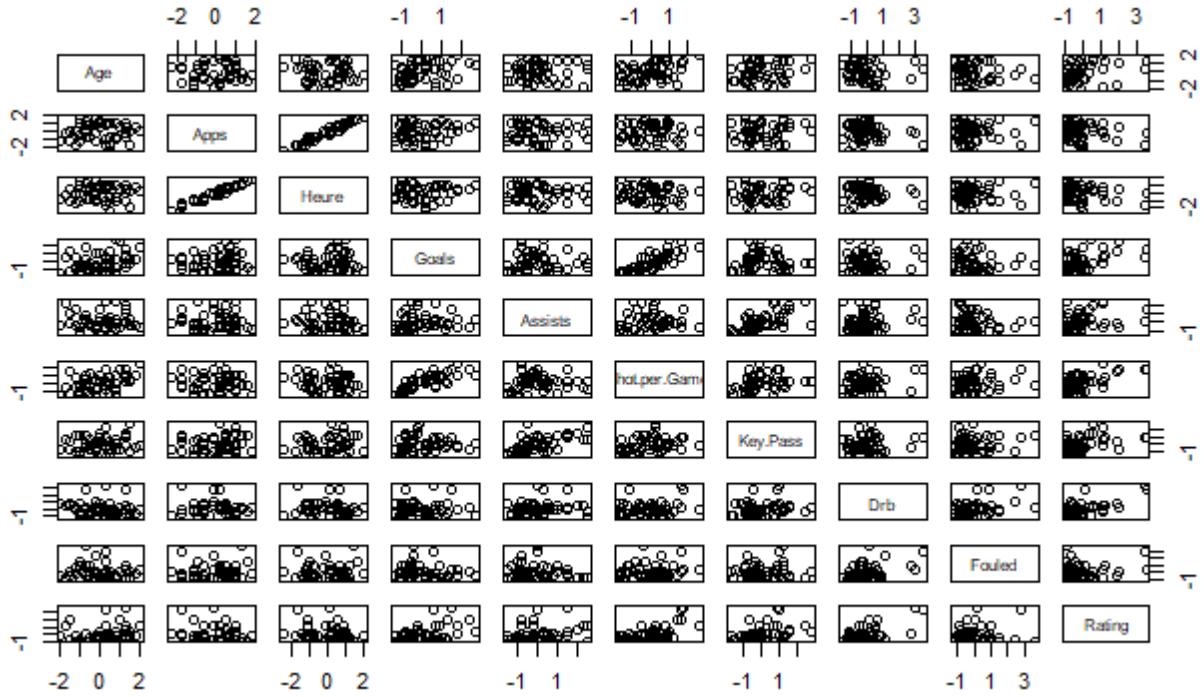

Figure 6 : Représentation par paire des données

Au vu des ces premiers résultats, il semblerait qu'on ne puisse trouver une structure de groupe avec l'ensemble des variables que nous disposons. De ce fait, il est préférable de traiter les données en différents points. Lors de l'ACP, nous avons remarqué que l'âge des joueurs était corrélé positivement avec de nombreuses variables déterminantes de la performance. D'autre part, avec ses données nous pouvons comparer les performances entre les clubs dans l'objectif de savoir quels clubs ont eu les joueurs les plus performants en termes de buts marqués, passes décisives et évaluation globale par exemple. Et enfin, nous allons nous intéresser à l'impact de la position des joueurs sur le terrain.

Toutes ces idées dégagent 3 problématiques majeures :

- Comment les performances des joueurs diffèrent-elles en fonction de leur âge ? Existe-t-il des tendances claires de performance pour différents groupes d'âge, tels que les jeunes joueurs en développement par rapport aux vétérans expérimentés ?
- Comment les performances des joueurs varient-elles d'un club à l'autre ? Existe-t-il des caractéristiques spécifiques aux clubs qui influencent les performances individuelles des joueurs ?
- Comment les performances des joueurs varient-elles en fonction de leur position principale sur le terrain ?

1. Comment les performances des joueurs varient-elles en fonction de leur position principale sur le terrain ?

Pour étudier la variation des performances en fonction de la position des joueurs sur le terrain, une Classification Ascendante Hiérarchique (CAH) a été utilisée dans un premier temps suivie

d'un K-means afin de comparer les résultats de ces deux méthodes. Les joueurs présents dans l'ensemble de nos données sont soit des milieux de terrain (M) soit des attaquants (FW). La CAH a été réalisée seulement sur les variables quantitatives déterminantes de la performance, telles que les heures jouées, les buts, les tirs…Après avoir standardisé les données, la matrice des distances a été calculée et à servi pour entreprendre la CAH. La figure 7 représente le dendrogramme obtenu. Nous identifions sur ce dendrogramme que les milieux et les attaquants sont quasiment bien séparés, cependant au vue de la silhouette il est difficile d'affirmer une structure à 2 groupes.

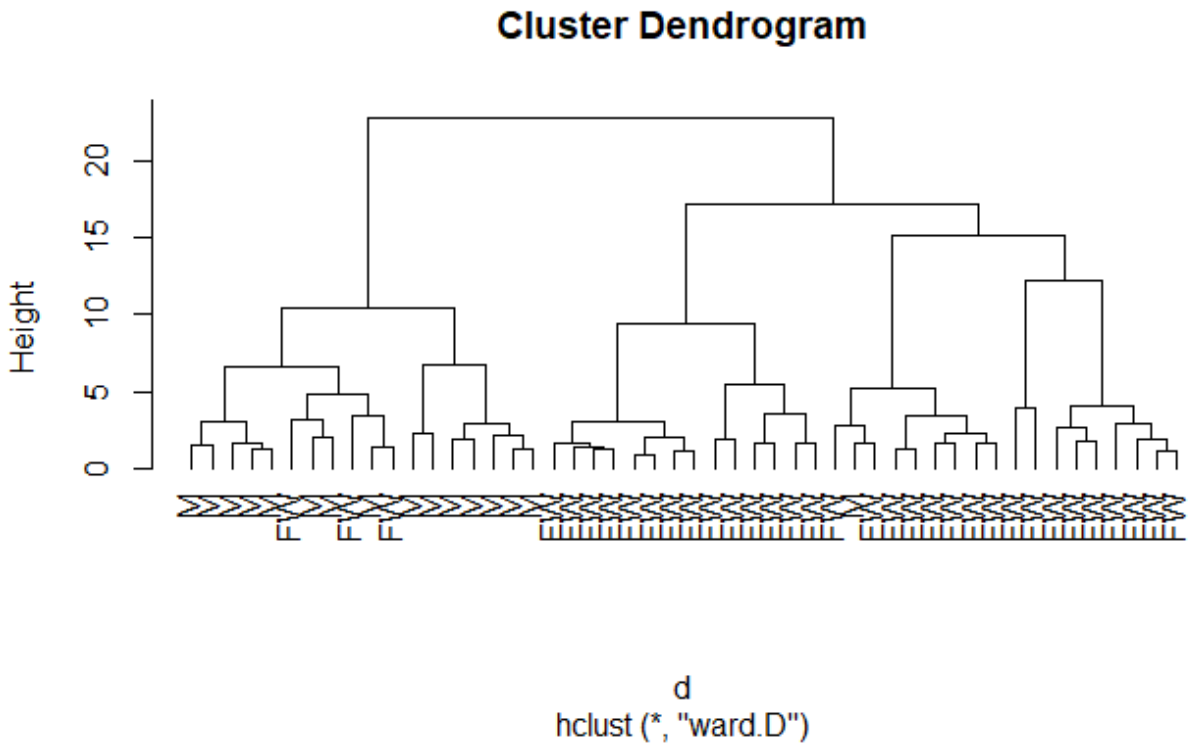

Figure 7 : Dendrogramme

Pour vérifier la structure des groupes, nous avons réalisé la méthode du coude (figure 8). Sur la graphique résultant, une structure à 2 groupes semble apparaître malgré un saut non réellement significatif. Nous nous sommes aidés de critères automatiques calculés dans le package "NbClust" afin de vérifier une énième fois cette structure, il apparaît qu'une répartition en 2 groupes est raisonnable même si cette méthode en recommande 5.

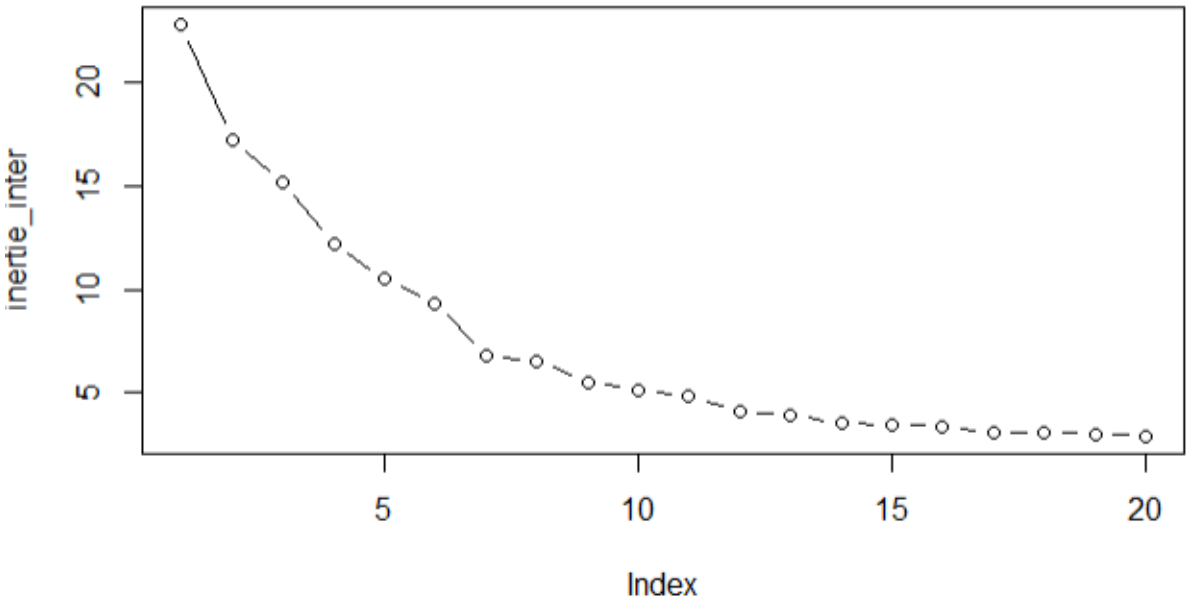

Figure 8 : Méthode du coude

Ensuite, la répartition par paire de variables en fonction des 2 groupes (figure 9) nous a permis d'identifier les paires de variables qui semblaient le plus se répartir en 2 groupes.

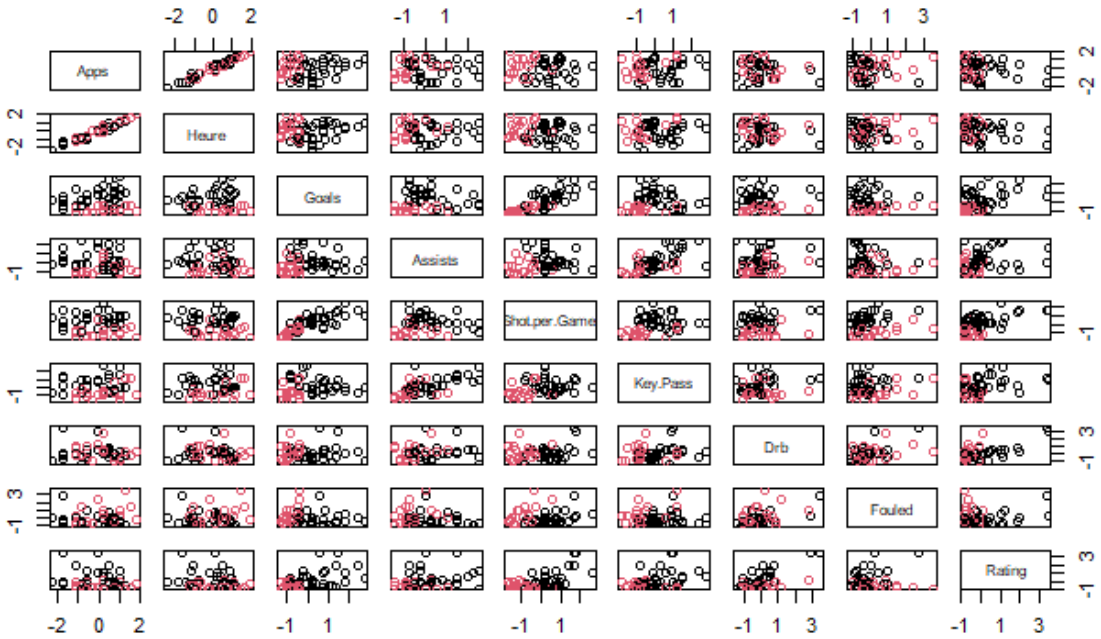

Figure 9 : Dispersion par paires en fonction des groupes

Suite à ce résultat, nous avons décidé de traiter les paires suivantes : Apps & Heure ; Goals & Shot.per.Game, Assists & Key.Pass ; Drb & Fouled. Les figures 10 et 11 représentent les graphiques obtenus avec la fonction *cutree()*.

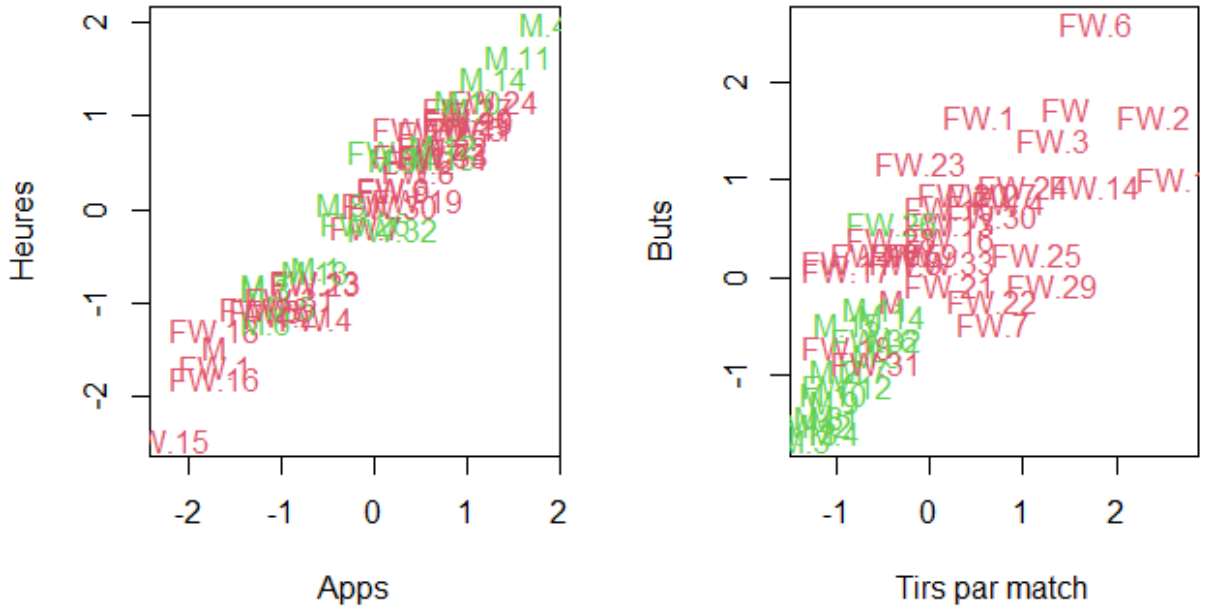

Figure 10

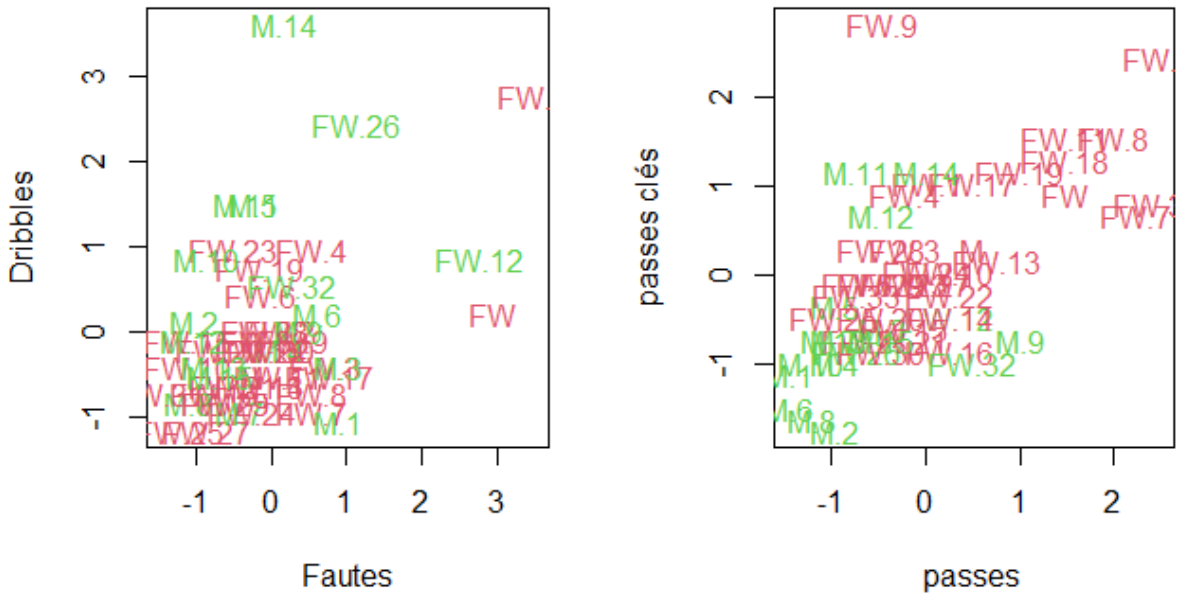

Figure 11

Avant d'interpréter ces résultats, nous avons vérifié la qualité de la classification précédente. Nous avons constaté que les attaquants (FW) et milieux (M) étaient à 4 exceptions près classés dans le bon groupe d'appartenance (figure 12). Les graphiques précédents montrent que d'heures et d'apparition au cours de la saison, les milieux de terrain ont tendance à être légèrement plus performants en termes d'heures et d'apparition sur le terrain que les attaquants. A l'opposé, les attaquant par leur position plus proche des buts ont comme le démontrent le graphique (figure 10 à droite) ont en très grande majorité plus d'occasion pour tirer et pour marquer des buts que les joueurs de milieux de terrain. Il en va de même pour les passes clé lors des matchs, même si certains milieux arrivent à se démarquer malgré tout. En ce qui concerne les dribbles et les fautes, le graphique ne montre aucune différenciation entre les deux postes.

	gpe.ward	
Foot.postes	1	2
FW	31	3
M	1	15

Figure 12 : Vérification de la qualité de la classification

La méthode des K-means a ensuite été réalisée pour chaque paire de variables. Comparé à la CAH, les K-means n'ont pas montré de grand intérêt pour étudier la variation des performances en fonction de la position des joueurs sur les terrains. Les figures 13 et 14 représentent les résultats des K-means.

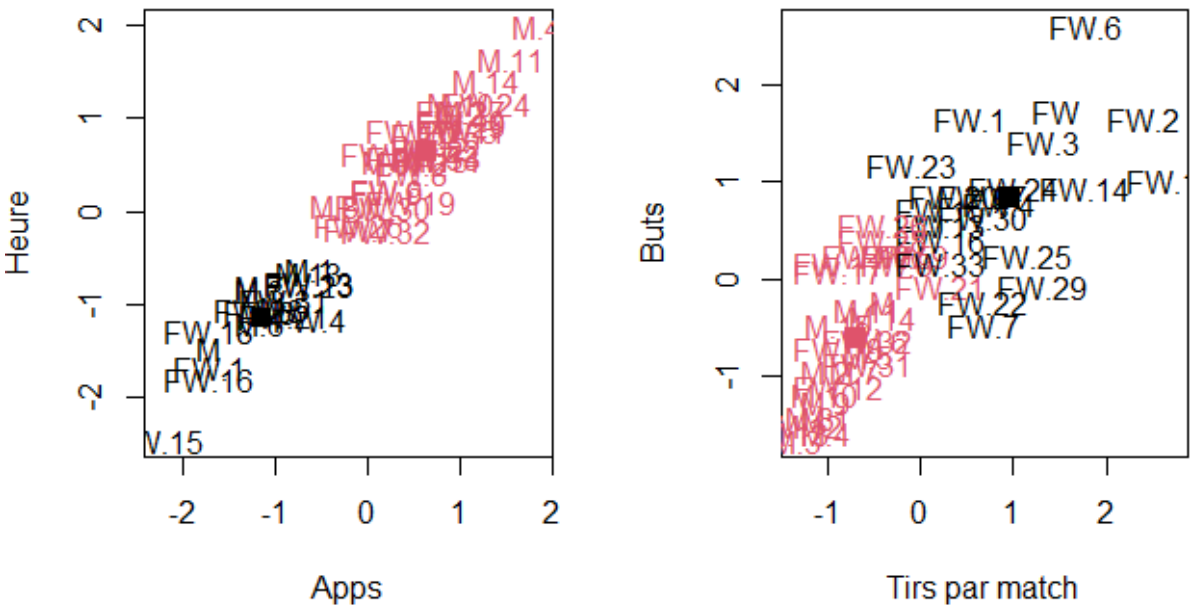

Figure 13

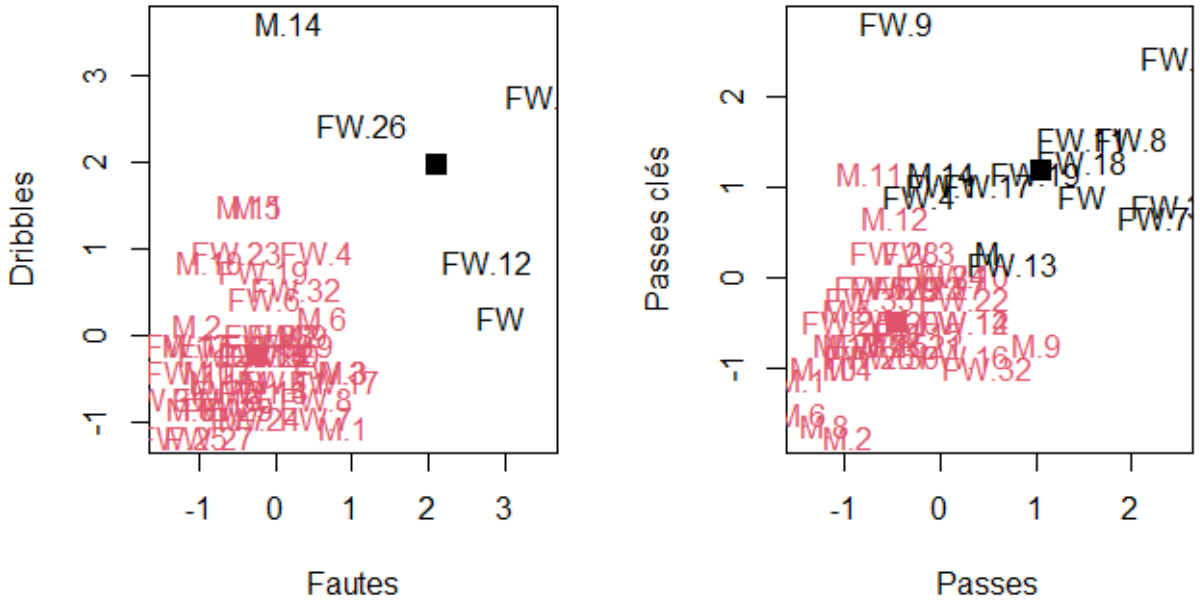

Figure 14

2. Comment les performances des joueurs varient-elles d'un club à l'autre, d'un championnat à l'autre ? Existe-t'il des caractéristiques spécifiques aux clubs qui influencent les performances individuelles des joueurs.

Pour analyser les performances des joueurs en fonction de leur club, il est nécessaire de préparer un jeu de données spécifique. Ce nouveau jeu de données regroupe les joueurs par club. Pour chaque club, nous calculerons la moyenne des différentes statistiques pertinentes, comme les buts marqués, les passes décisives, etc. Le résultat final sera un jeu de données contenant 27 lignes, correspondant aux 27 clubs différents présents dans notre jeu de données initial. Ce regroupement et ce calcul des moyennes permettent une évaluation claire des performances moyennes par club. Ensuite,une Classification Ascendante Hiérarchique (CAH) et un K-means seront utilisés afin de comparer les résultats de ces méthodes. A noter qu'un ACP à été tenté mais sans résultat significatif . Les joueurs appartiennent donc à 27 clubs différents. Voici la liste des clubs correspondant à leurs ID :

1 : Aston Villa
2 : Atalanta
3 : Barcelona
4 : Bayern munich
5 : Borussia Dortmund
6 : Borussia M.Glasdbach
7 : Eintrach Francfort
8 : Inter
9 : Juventus
10 : Lazio
11 : Leiceister
12 : Liverpool

13 : Manchester City
14 : Manchester united
15 : Marseille
16 : Monaco
17 : Montpellier
18 : Paris saint germain
19 : Leipzig
20 : Real Betis
21 : Real Madrid
22 : Real sociedad
23 : Roma
24 : Sassuolo
25 : Schalke 04
26 : Villarreal
27 : Wolverhampton

Comme précédemment, l'analyse en Clusters Hiérarchique (CAH) a été effectuée uniquement avec les variables quantitatives essentielles à l'évaluation de la performance. Le dendrogramme présenté à la figure 15 illustre les résultats de cette analyse. Sur ce dendrogramme, il est clair que certains clubs se distinguent nettement des autres. Nous pouvons distinguer trois groupes distincts, ce qui nous permet de supposer avec une certaine assurance l'existence de ces groupements.

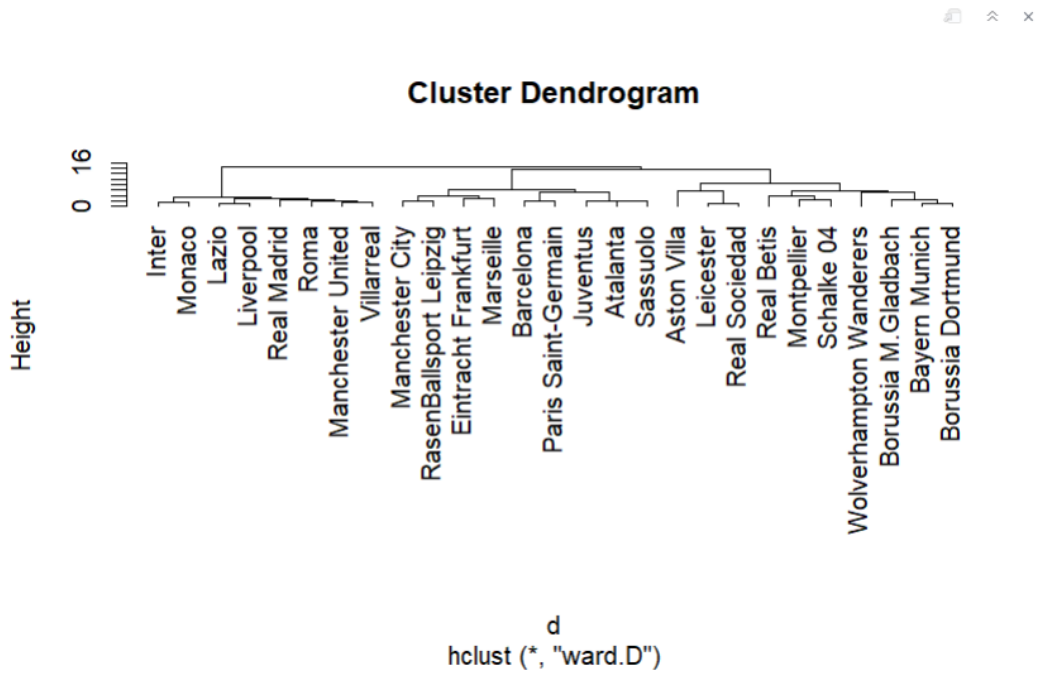

Figure 15

Nous avons donc fait une méthode du coude (figure 16) pour affirmer le nombre de groupe mais le résultat n'est pas probant . Nous nous sommes donc aidés du package "NbClust" afin de vérifier cette structure, il apparaît qu'une répartition en 3 groupes est recommandée. (figure 17).

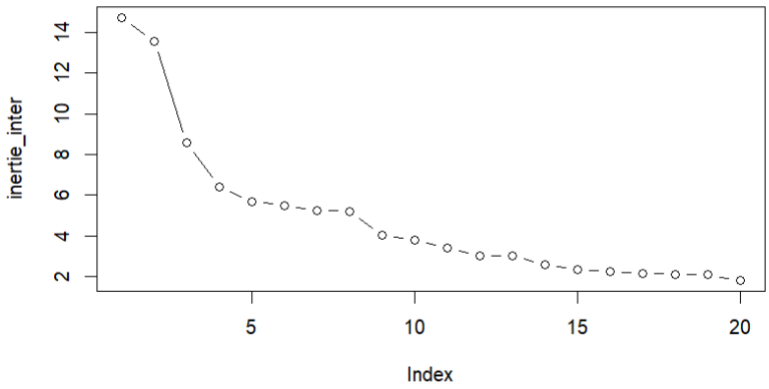

Figure 16

```
* Among all indices:
* 3 proposed 2 as the best number of clusters
* 6 proposed 3 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 3 proposed 9 as the best number of clusters
* 3 proposed 10 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 1 proposed 13 as the best number of clusters
* 4 proposed 15 as the best number of clusters
```

Figure 17

Ensuite, la répartition par paire de variables en fonction des 2 groupes (figure 18) nous a permis d'identifier les paires de variables qui semblaient le plus se répartir en 2 groupes.

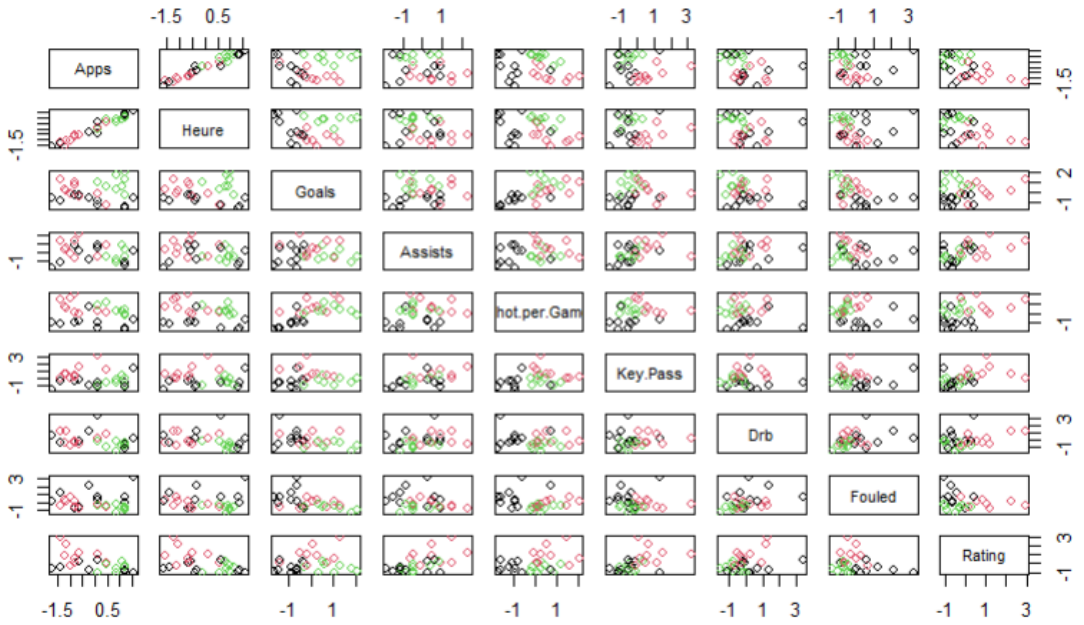

Figure 18 Dispersion par paires en fonction des groupes

Pour analyser les résultats, nous avons utilisé une méthode de clustering appelée gpe.ward, illustrée par la figure 19. Cette approche nous aide à identifier les clubs qui présentent des statistiques similaires et qui sont ainsi regroupés dans l'un des trois groupes identifiés. Par exemple, Aston Villa et Atalanta, ainsi que la Lazio et Liverpool, sont classés dans le même groupe, indiquant une similitude dans leurs performances et leurs apparitions

Foot.Club	gpe.ward		
	1	2	3
Aston Villa	1	0	0
Atalanta	0	1	0
Barcelona	0	1	0
Bayern Munich	1	0	0
Borussia Dortmund	1	0	0
Borussia M.Gladbach	1	0	0
Eintracht Frankfurt	0	1	0
Inter	0	0	1
Juventus	0	1	0
Lazio	0	0	1
Leicester	1	0	0
Liverpool	0	0	1
Manchester City	0	1	0
Manchester United	0	0	1
Marseille	0	1	0
Monaco	0	0	1
Montpellier	1	0	0
Paris Saint-Germain	0	1	0
RasenBallsport Leipzig	0	1	0
Real Betis	1	0	0
Real Madrid	0	0	1
Real Sociedad	1	0	0
Roma	0	0	1
Sassuolo	0	1	0
Schalke 04	1	0	0
Villarreal	0	0	1
Wolverhampton Wanderers	1	0	0

Figure 19 gpe.ward

Suite à ce résultat, nous avons décidé de traiter les paires suivantes : Apps & Heure ; Goals & Shot.per.Game, Assists & Key.Pass ; Drb & Fouled. Les figures 20 et 21 représentent les graphiques obtenus.

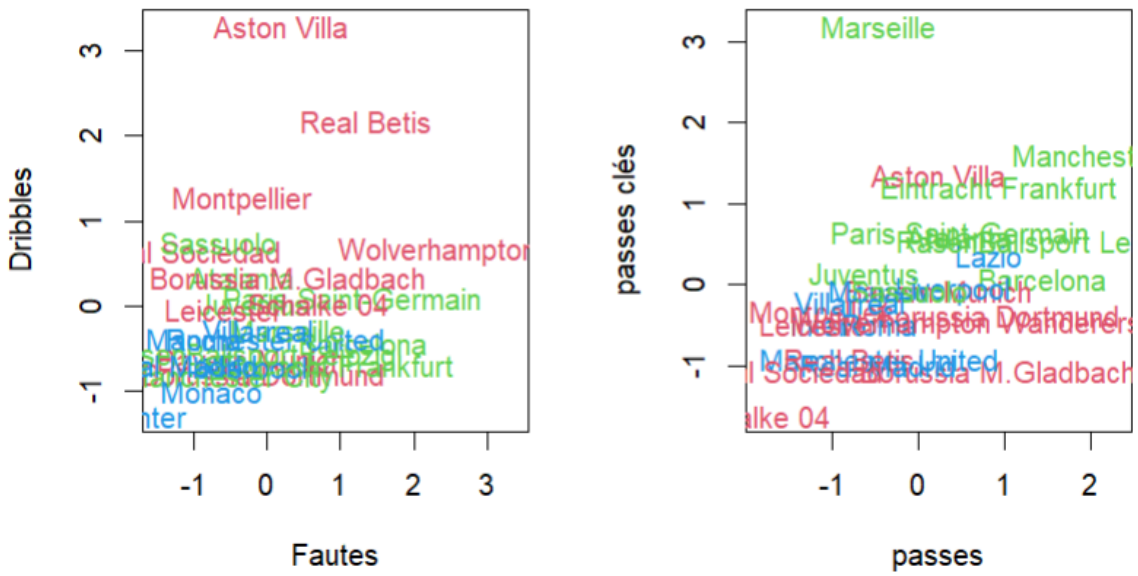

Figure 20

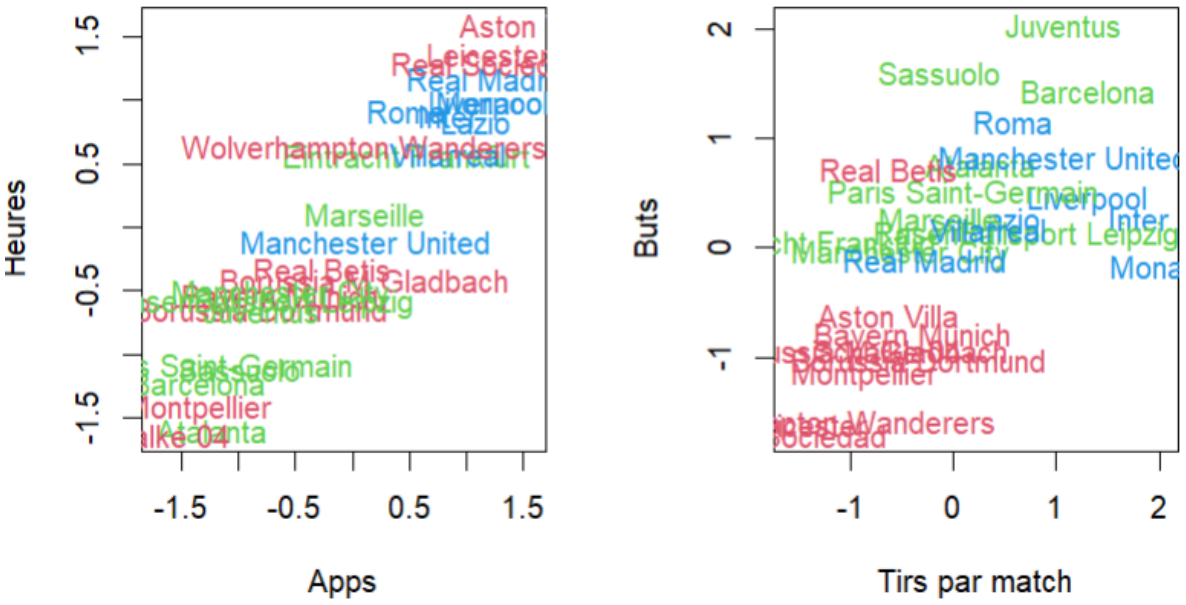

Figure 21

Les graphiques analysés permettent de tirer plusieurs conclusions intéressantes sur le comportement et les performances des équipes de football.

Premièrement, il apparaît logiquement que les équipes qui effectuent un grand nombre de dribbles, comme Wolverhampton et le Real Betis, tendent également à subir davantage de fautes. Cette tendance se confirme pour la majorité des équipes mais présente des exceptions notables. Par exemple, Aston Villa se distingue par un nombre élevé de dribbles sans pour autant subir beaucoup de fautes, tandis qu'Eintracht Francfort subit de nombreuses fautes sans dribbler fréquemment.

Concernant les statistiques de passes et de passes clés, nous observons généralement que les équipes qui réalisent beaucoup de passes tendent aussi à effectuer un grand nombre de passes clés. Schalke 04 illustre le cas inverse, avec peu de passes clés en raison d'un faible nombre de passes, tandis que Manchester City montre une forte corrélation entre un nombre élevé de passes et de passes clés. Marseille se démarque en générant un nombre élevé de passes clés malgré un nombre relativement bas de passes, ce qui suggère une stratégie efficace de contre-attaque.

Les graphiques sur les heures de jeu et les apparitions montrent une corrélation directe, sans surprise notable. En ce qui concerne les buts et les tirs par match, la tendance générale indique que les équipes qui tirent fréquemment au but, comme la Juventus, marquent également plus de buts. Cependant, de rares exceptions existent, telles que Sassuolo, qui malgré un nombre moins élevé de tirs, affiche une efficacité remarquable devant le but.

En résumé, ces analyses graphiques révèlent des tendances claires dans le comportement des équipes sur le terrain, tout en mettant en évidence certaines stratégies uniques et efficacités particulières.

La méthode des Kmeans, illustrée dans les figures 22 et 23, a été appliquée en choisissant de segmenter les données en deux clusters pour tous les graphiques, à l'exception de celui analysant les heures de jeu et les apparitions. Cette approche a produit des résultats particulièrement intéressants, les deux groupes étant nettement plus faciles à distinguer les uns des autres.

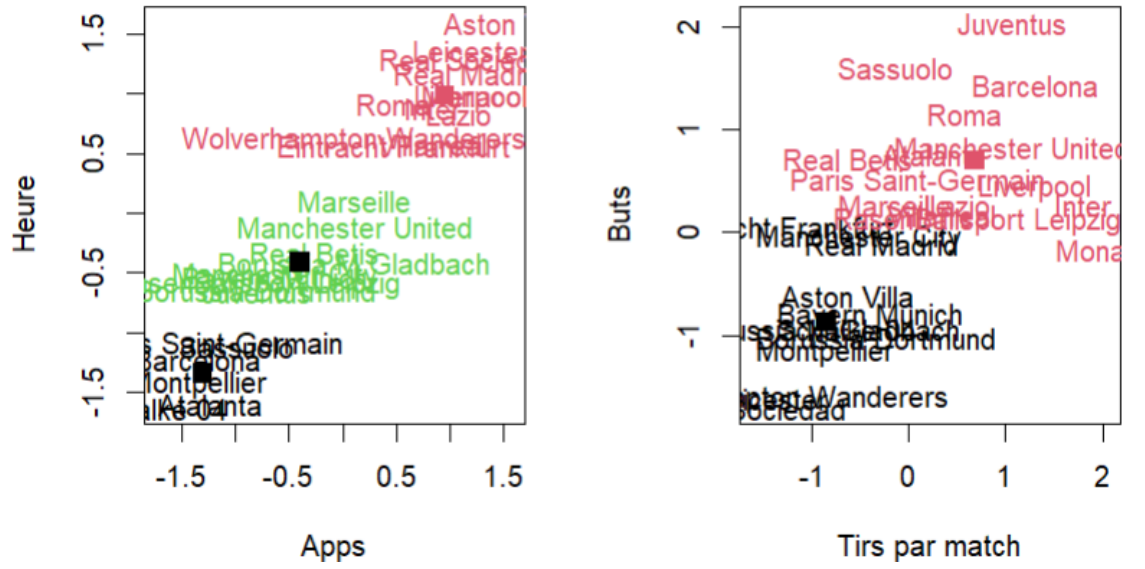

Figure 22

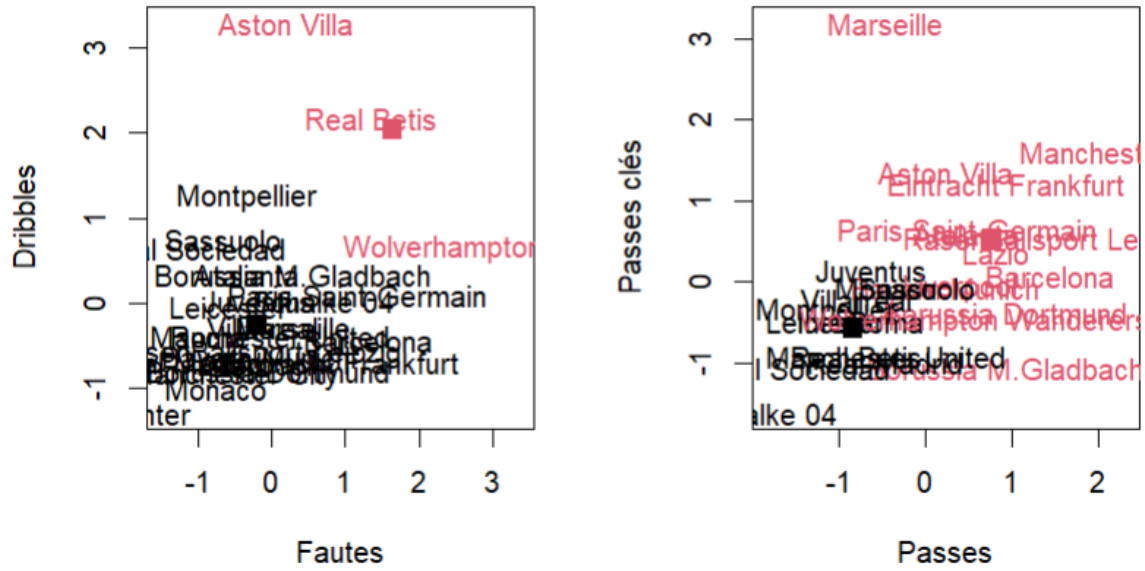

Figure 23

On distingue (figure 24) grâce à la fonction ‘clusplot’3 groupes sont complètement distincts malgré 3 observations qui sont à la frontières entre 2 groupes.

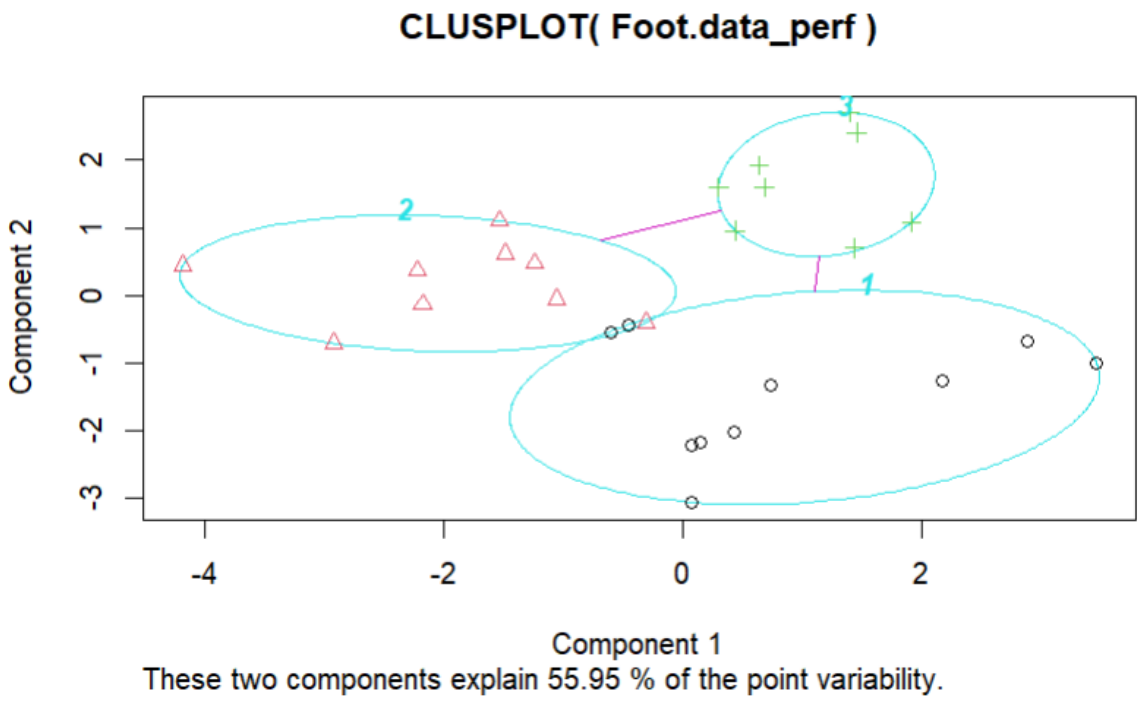

Figure 24

3. Comment les performances des joueurs diffèrent-elles en fonction de leur âge ?

Pour étudier la performance des joueurs en fonction de leur âge, d'abord, une Classification Ascendante Hiérarchique (CAH) a été utilisée, suivie de la méthode des K-means. Les données comprennent des joueurs âgés de 19 à 25 ans. Les joueurs ont été séparés en 2 tranches d'âge, de 19 à 26 les jeunes et de 27 à 35 ans les vétérans. Le même processus que pour la première problématique a été employée, en premier les données ont été standardisées puis une matrice de distance a été calculée et utilisée pour la CAH. Le dendrogramme est représenté par la figure 25, les jeunes et les vétérans semblent globalement séparés même s'ils ne sont pas clairement distincts.

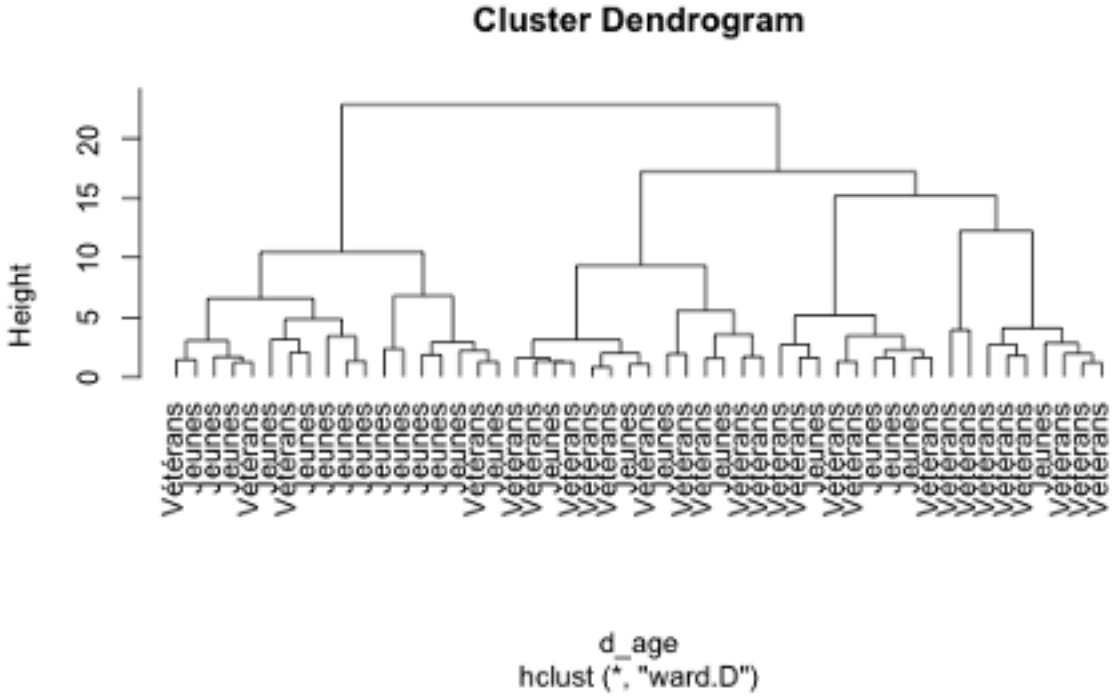

Figure 25 Dendrogramme

Afin de vérifier la composition des groupes, la méthode du coude a été réalisée (figure 26). On observe qu'une structure à 2 groupes semble être cohérente même si le saut n'est pas très marqué. La fonction NbClust de même package a été utilisée afin de donner une répartition des groupes et de vérifier notre choix, cette méthode conseille d'utiliser 5 groupes.

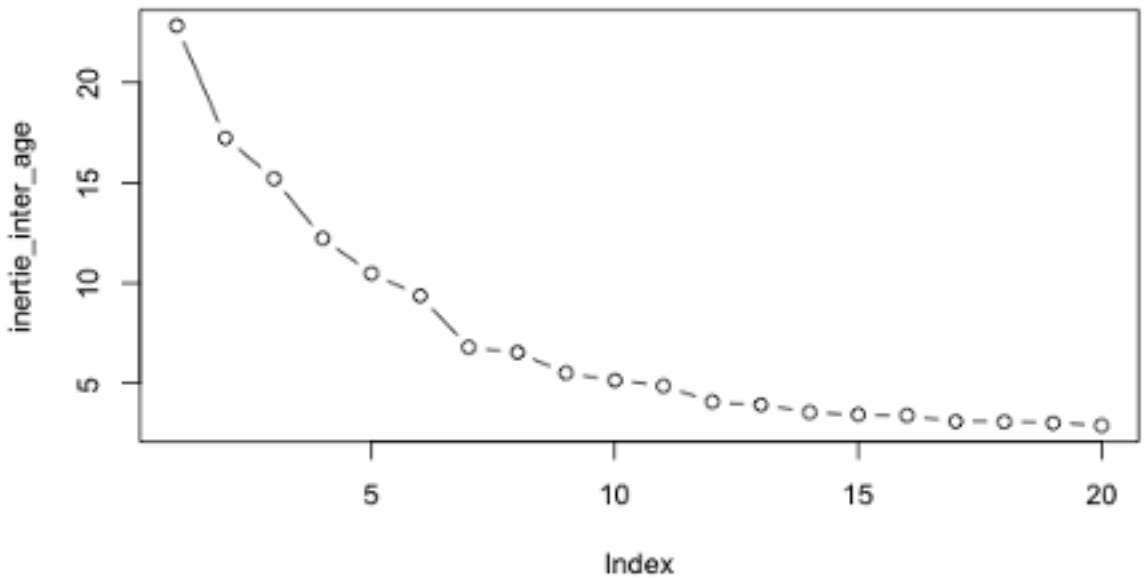

Figure 26 Méthode du coude

Comme pour la répartition par poste, par la suite, la fonction par paire a été utilisée par rapport aux 2 groupes d'âges afin d'identifier les paires de variables. (figure 27)

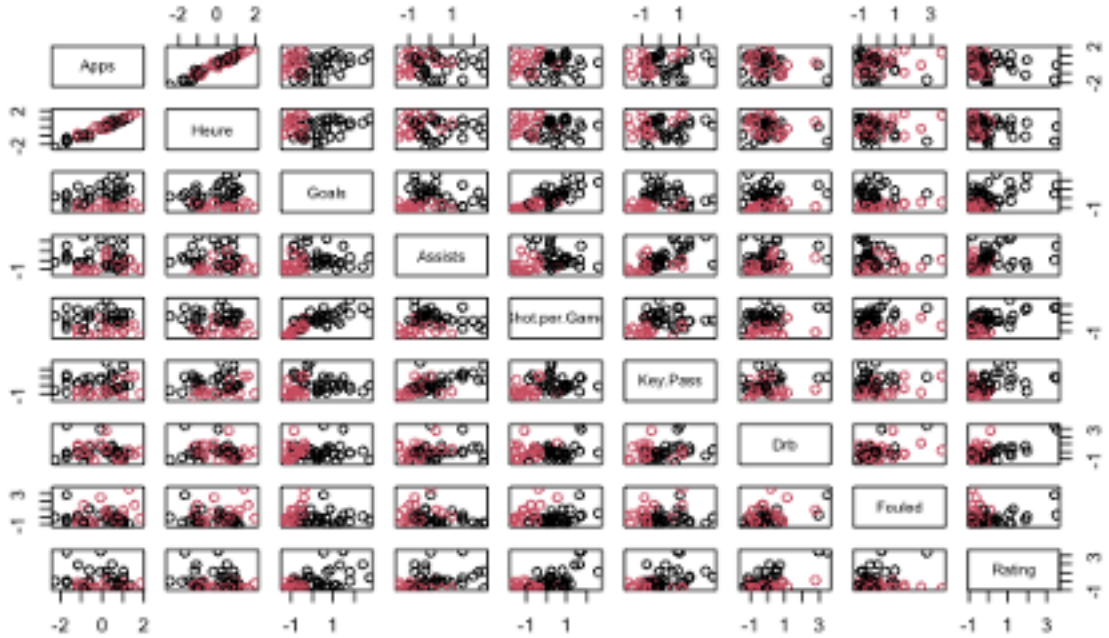

Figure 27 Dispersion par paires en fonction des groupes

Similairement à la première problématique, les paires suivantes seront associées : Apps & Heure ; Goals & Shot.per.Game, Assists & Key.Pass ; Drb & Fouled. Les figures 28 et 29 représentent les graphiques obtenus avec la fonction *cutree()*.

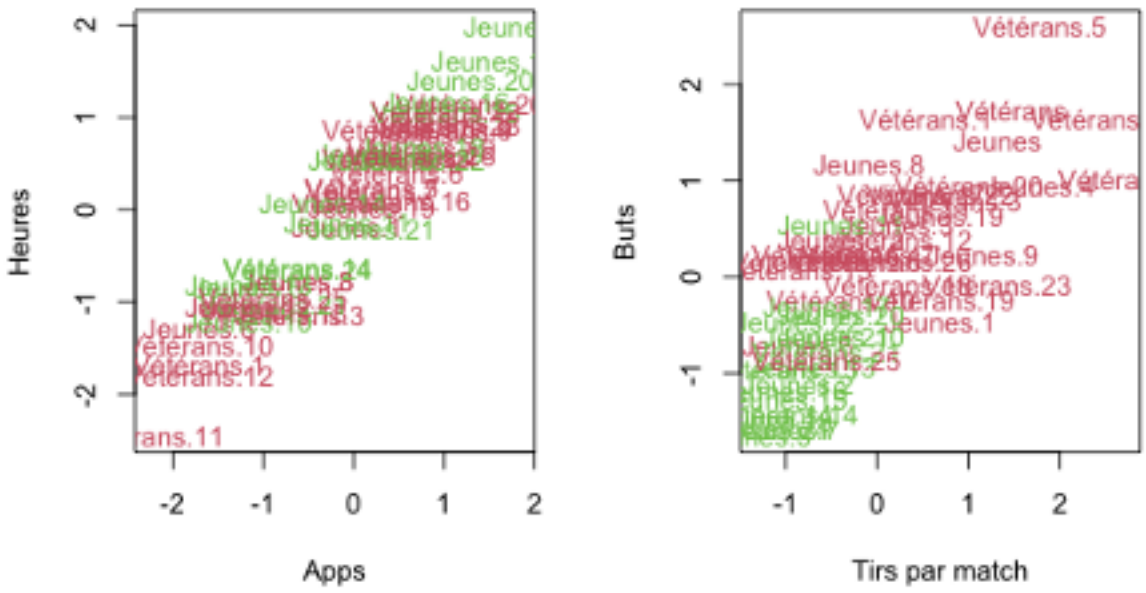

Figure 28

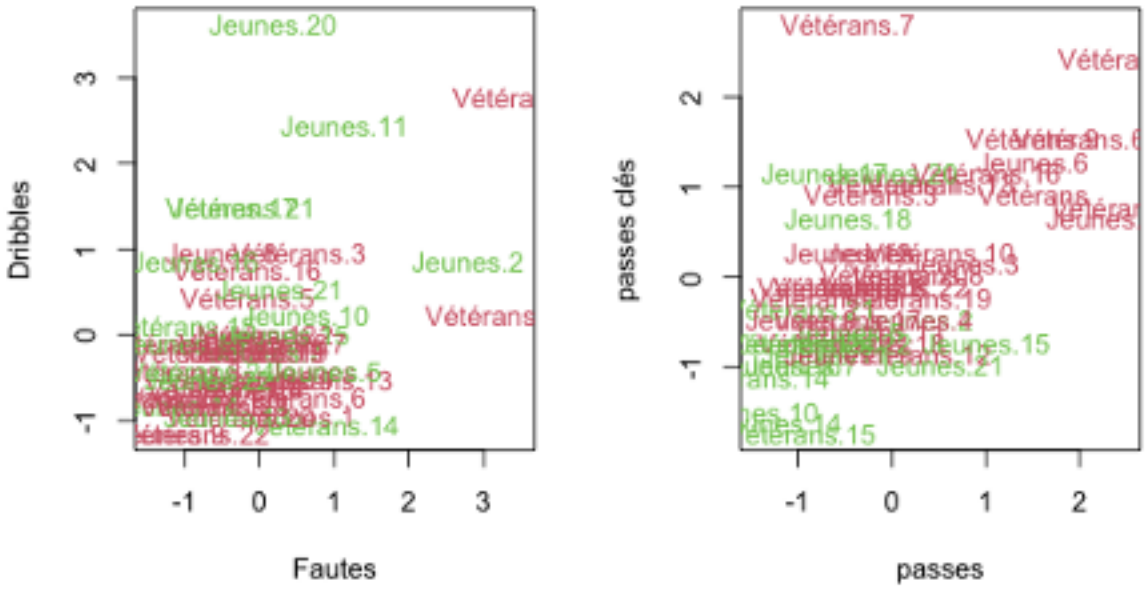

Figure 29

Ensuite, la répartition des groupes a été vérifiée et on voit que les jeunes et les vétérans sauf 13 sont dans le bon groupe (figure 30). Dans les prochains graphiques, une tendance de

regroupement en fonction de l'âge des joueurs est perceptible. Les joueurs classés dans la catégorie "Jeunes" ont tendance à avoir moins d'apparitions et moins d'heures de jeu par rapport aux joueurs classés dans la catégorie "Vétérans". Les joueurs ‘Jeunes’ semblent moins tirés de fois au but par match et moins marqués de buts contrairement aux joueurs ‘Vétérans’. Les joueurs sont dispersés uniformément sur l'ensemble du graphique, laissant entendre que les comportements de dribble et les infractions commises ne sont pas fortement influencés par l'âge des joueurs. Les joueurs les plus habiles dans la distribution de passes, et donc potentiellement plus créatifs, pourraient être plus jeunes en moyenne.

```
             gpe.ward_age
Foot.ages     1  2
  Jeunes      9 14
  Vétérans 23  4
```

Figure 30 Vérification de la qualité de la classification

De la même manière que pour la première problématique, la méthode des K-means ne semble pas montrer grand intérêt pour étudier la variation des performances par rapport à l’âge des joueurs (figures 31 et 32).

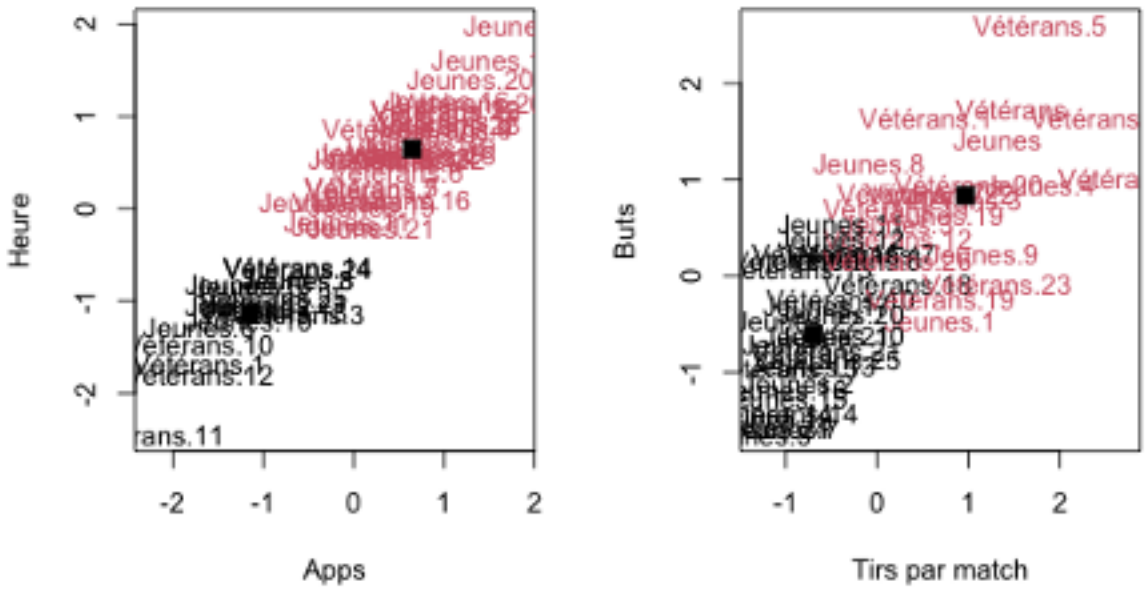

Figure 31

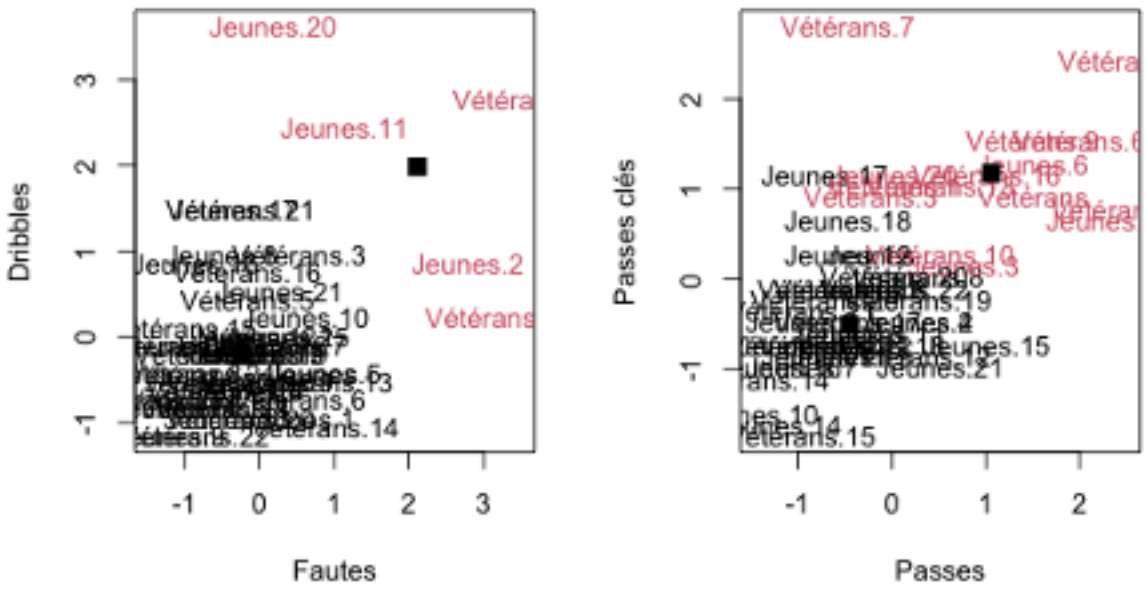

Figure 32

On distingue (figure 33) grâce à la fonction ‘clusplot’ de la library ‘cluster’, 2 groupes sont assez distincts même si quelques individus, environ 15, se retrouvent entre les 2. On retrouve le même nombre d'individus qui n’étaient pas dans le bon groupe figure 30.

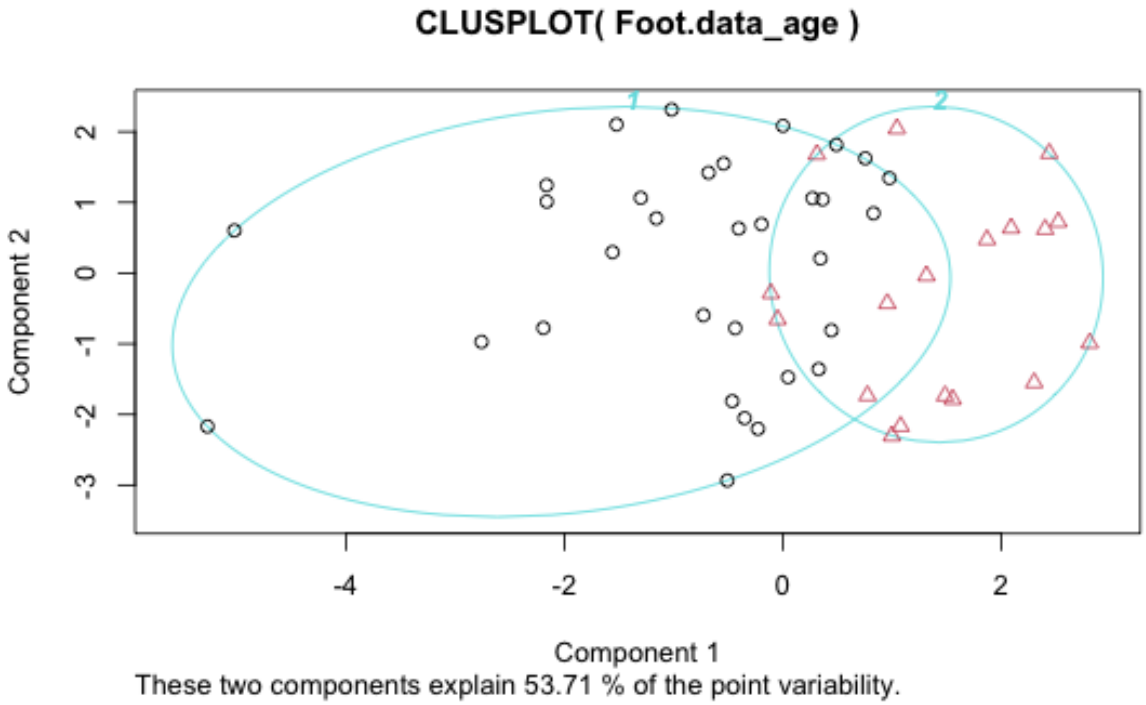

Figure 33