PROJET ANALYSE MULTIDIMENSIONNELLE ET CLUSTERING

PERENON Clément

REGAZZETTI Léa

TRIKI Arthur

Master 1 Informatique

Promotion 2020-2021

Table des matières

Introduction

Ce dossier s'intègre dans le cadre du contrôle continu des connaissances de nos études en première année de master informatique, et plus particulièrement lors de notre cours d'analyse multidimensionnelle et clustering.

Dans ce dossier, nous appliquerons nos connaissances acquises lors de notre formation dans le but de réaliser des traitements statistiques grâce au logiciel R sur un jeu de données de notre choix. Ainsi, dans un premier temps nous présenterons le jeu de données choisi, nous l'importerons dans R puis nous effectuerons un nettoyage des données.

Nous nous sommes alors posé la question suivante, comment peut-on étudier les caractéristiques des joueurs ? Pour cela, nous avons réalisé 3 traitements statistiques, tout d'abord une Analyse en Composantes Principales (ACP), ensuite une Analyse Factorielle des Correspondances (AFC), et pour terminer nous avons réalisé plusieurs méthodes de clustering.

Présentation, import et nettoyage du fichier de données

Le jeu de données provient du site Kaggle, une plateforme organisant des compétitions en data science. Le lien du dataset est le suivant : https://www.kaggle.com/karangadiya/fifa19. Dans ce jeu de données, on retrouve des informations sur différents joueurs de football.

```
joueurs =  read.csv("data.csv",sep=",",header = TRUE, row.names = 1, encoding
= "UTF-8")
str(joueurs)
## 'data.frame':    18207 obs. of  88 variables:
##  $ ID                      : int  158023 20801 190871 193080 192985 183277
177003 176580 155862 200389 ...
##  $ Name                    : chr  "L. Messi" "Cristiano Ronaldo" "Neymar J
r" "De Gea" ...
##  $ Age                     : int  31 33 26 27 27 27 32 31 32 25 ...
##  $ Photo                   : chr  "https://cdn.sofifa.org/players/4/19/158
023.png" "https://cdn.sofifa.org/players/4/19/20801.png" "https://cdn.sofifa.
org/players/4/19/190871.png" "https://cdn.sofifa.org/players/4/19/193080.png"
...
##  $ Nationality             : chr  "Argentina" "Portugal" "Brazil" "Spain"
...
```

Le fichier se compose de 18207 observations, qui sont les joueurs de football.

Toutefois, certains noms sont présents plusieurs fois dans le fichier de données, nous allons donc éliminer les joueurs en doublons en ne gardant que la première observation rencontrée dans le jeu de données pour les joueurs apparaissant plusieurs fois.

```
doublons = which(duplicated(joueurs$Name))
data = joueurs[-doublons,]
```

Par ailleurs, dans le cadre de ce projet, il n'est pas nécessaire d'avoir autant d'observations, nous allons donc réduire notre jeu de données à 1500 observations choisies aléatoirement. Mais pour permettre de reproduire notre travail, nous générons les mêmes données aléatoires.

```
set.seed(10)
x = sample(1:dim(data)[1],1500)
donnees = data[x,]
```

Puis ne garder que les joueurs n'ayant pas de valeurs manquantes.

```
donnees = donnees[which(complete.cases(donnees)),]
```

Nous pouvons alors indiquer que la deuxième colonne du fichier de données, qui contient les noms des joueurs, servira pour les étiquettes des lignes. De plus, de nombreuses variables ne sont pas nécessaires pour nos analyses, nous allons donc les enlever de notre jeu de données.

```
rownames(donnees) = donnees[,2]
d = donnees[,-c(1,2,4,6,10,28:53)]
str(d)
## 'data.frame':    1491 obs. of  57 variables:
##  $ Age                     : int  25 37 21 34 17 30 24 24 36 19 ...
##  $ Nationality             : chr  "Switzerland" "France" "England" "Brazil
" ...
##  $ Overall                 : int  80 69 59 67 65 71 67 69 72 67 ...
##  $ Potential               : int  83 69 69 67 79 71 69 75 72 80 ...
##  $ Club                    : chr  "Milan" "Valenciennes FC" "Morecambe" "P
araná" ...
```

Ainsi, notre fichier de données contient 1491 individus statistiques (les joueurs de football).

ACP

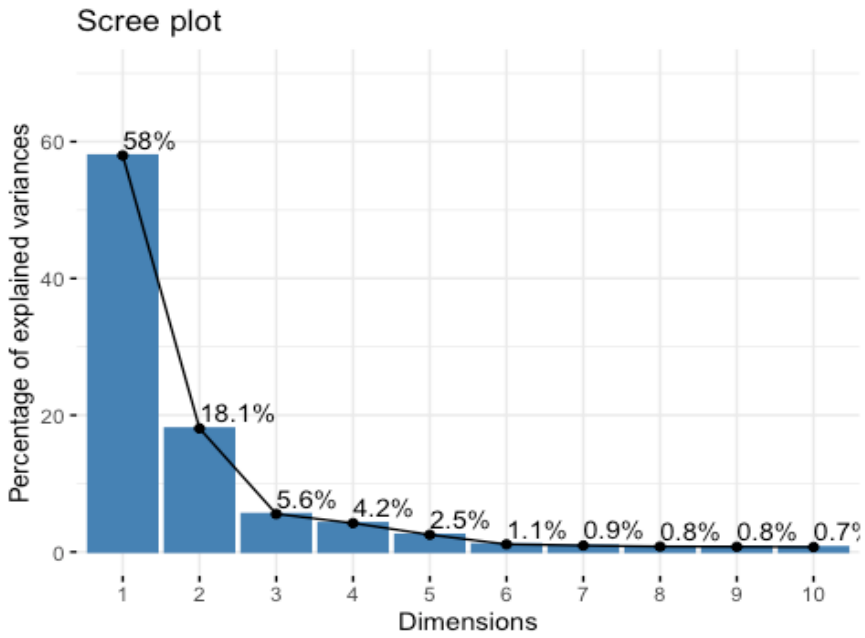
Dans un premier temps, on souhaite savoir si dans notre fichier de données il y a des joueurs qui se ressemblent. Pour cela, nous allons étudier différents attributs sur les joueurs, notés sur une échelle de 100. On pourra alors se demander si certains critères sont corrélés.

Réalisation de l'ACP

```
d.active <- d[, 23:56]
res.pca <- PCA(d.active,scale.unit=FALSE,graph = FALSE)
```

Calcul des valeurs propres

```
sum(eig.val[,1]) #Inertie totale
## [1] 9922.103
fviz_eig(res.pca,addlabels = TRUE, ylim = c(0, 70))
```



Nous sommes dans le cas d'une ACP non normée donc l'inertie moyenne est I/p. Dans ce cas c'est égal à 9922.103/34 = 291.82. Donc d'après la règle de Kaiser, on pourrait garder les 4 premiers axes car leur valeurs propres sont supérieures à 291.82.

Nous allons retenir et analyser les deux premiers axes qui restituent une majorité de l'information du jeu de données.

Analyse du premier axe

Analyse des individus

La qualité de représentation est d'autant plus grande que le cos2 est proche de 1, ce qui correspond à la situation où la distance projetée est fidèle à la distance initiale.

Ainsi, G. Viscarra, R. Ferguson, L. Grill caractérisent le côté négatif de l'axe 1 (ce sont des gardiens de but). Ces joueurs sont notés de la même façon, en comparaison à ceux qui caractérisent le côté positif de l'axe 1.

Analyse des variables

Meilleure est la contribution, plus sa position sera très à droite ou très à gauche et plus leur contribution dans la construction du graphique est importante. Crossing et finishing, BallControl, Stamina, ShortPassing sont représentés à droite de l'axe 1. GKGiving, GKKiking, GKPositioning et GKReflexes ont des coordonnées négatives, donc elles seront positionnées à gauche de l'axe 1. Elles sont caractérisées par les individus qui sont moins bien notés sur ces variables.

Plus le cos2 est élevé, plus la qualité de représentation est haute. Ainsi, BallControl, Dribbling et ShortPassing ont une excellente qualité de représentation sur l'axe 1.

Analyse du second axe

Analyse des individus

Le côté négatif de l'axe 2 est caractérisé par des joueurs tels que Rúben Dias, W. Orban, J. Pearce (des défenseurs). Mais nous constatons de la même manière qu'il est peu intéressant d'analyser les individus. On va donc analyser les variables.

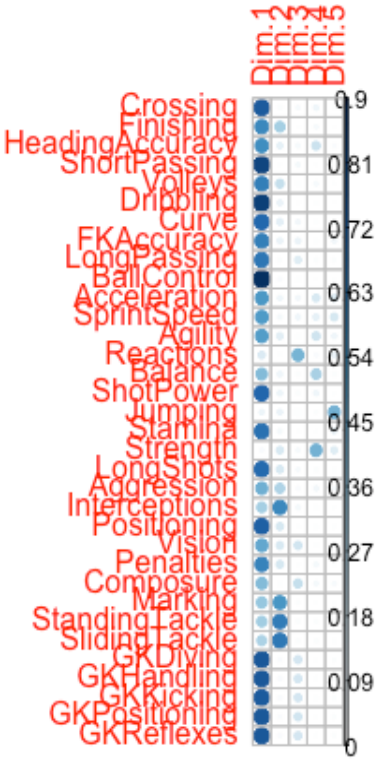
Analyse des variables

Les variables StandingTackle, SlidingTackle, Interceptions et marking contribuent fortement à la construction de l'axe 2. Ils sont également aux alentours de 0.6 de cos2, donc ils sont relativement bien représentés sur l'axe 2. Ainsi, le côté négatif de l'axe 2 est caractérisé par ces variables.

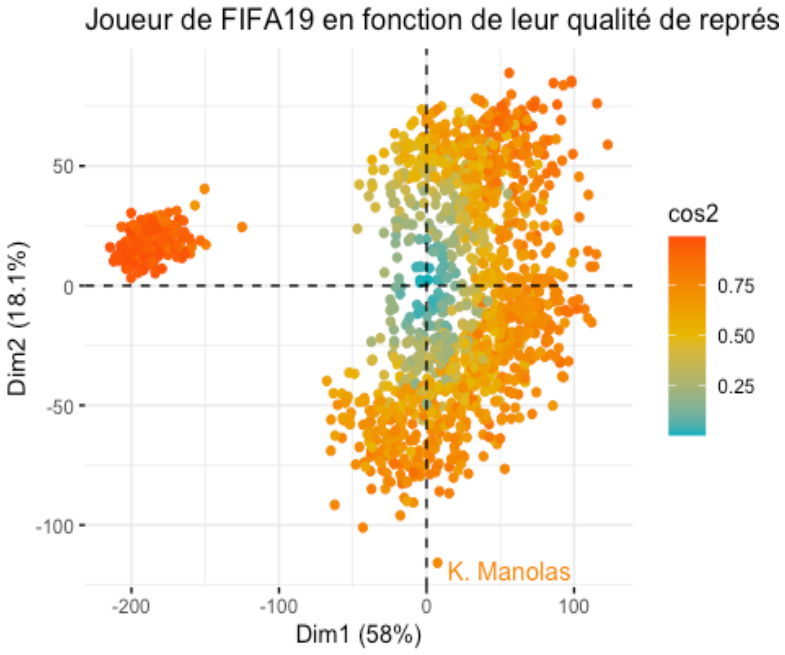
Analyse générale

On peut visualiser les cos2 des variables en fonctions des 5 dimensions.

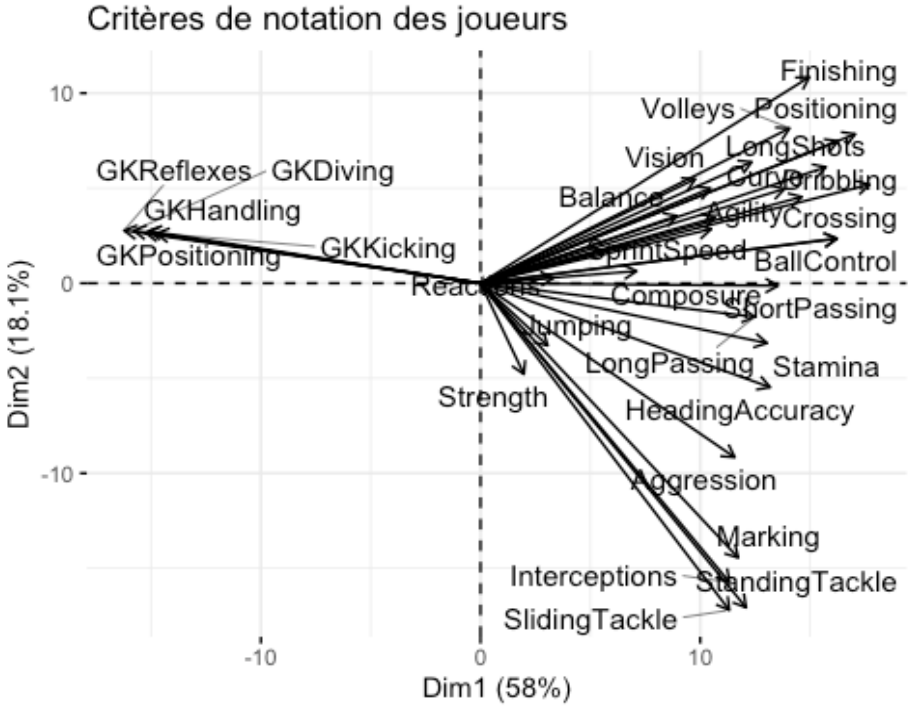
```
corrplot(var$cos2, is.corr=FALSE)
```



Dans la première dimensions les variables qui ont un bon cos2 sont donc bien Dribbling, BallControl ressortent bien. On peut avoir une vue un peu plus globale avec ce graphique.



Enfin, le graphique des variables sur les deux premiers axes factoriels.



Ainsi, nous avons pu voir que certains critères étaient corrélés, notamment ceux concernant le gardien de but, et qu'ils s'opposaient aux autres critères, ce qui est logique. Concernant les joueurs, les gardiens caractérisent le côté négatif de l'axe 1. Et les défenseurs caractérisent le côté négatif de l'axe 2.

AFC

Maintenant, nous allons nous intéresser à un autre type de question. Est-ce que la position sur le terrain a un lien avec la morphologie ?

La position sur le terrain correspond à la variable “Position” et la morphologie à la variable “Body.Type”. Nous avons tout d’abord transformer les variables correspondantes.

Construction du tableau de contingence

```
contingence = table(d$Position,d$Body.Type)
```

Test d’indépendance du Chi²

```
chisq <- chisq.test(contingence)
chisq
##
##  Pearson's Chi-squared test
##
## data:  contingence
## X-squared = 102.72, df = 52, p-value = 3.513e-05
```

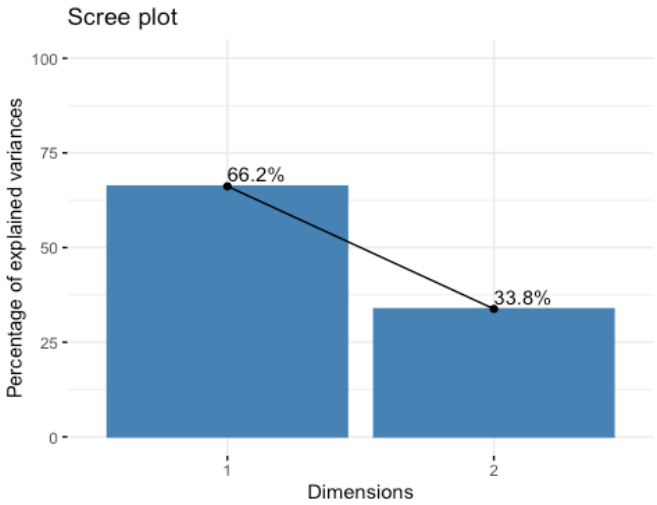
La pvalue est de 5.532e-05, ainsi, au seuil de 5% on rejette l’hypothèse d’indépendance. Les deux variables sont liées. Les joueurs n’ont pas la même position sur le terrain selon leur morphologie.

Réalisation de l’AFC

```
res.ca <- CA(contingence, graph = FALSE)  #calcul de l'AFC
```

Choix du nombre d’axes

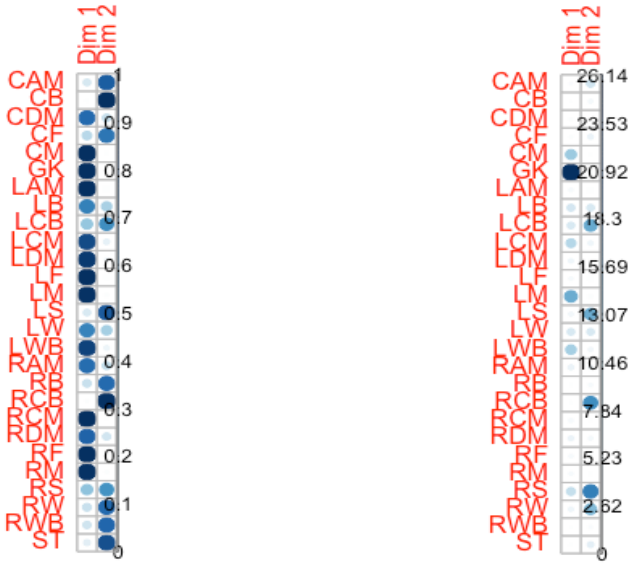
```
fviz_eig(res.ca, addlabels = TRUE, ylim = c(0, 100))
```



Étant donné que le nombre de modalités de la variable morphologie est de 3, nous n’obtenons que 2 valeurs propres, ainsi ce jeu de données ne permet pas de réellement appliquer la démarche visant à déterminer le nombre d’axes à interpréter. Nous allons donc interpréter les deux axes.

Analyse des profils lignes

```
row<-get_ca_row(res.ca)
corrplot(row$cos2, is.corr = FALSE)
corrplot(row$contrib, is.corr=FALSE)
```

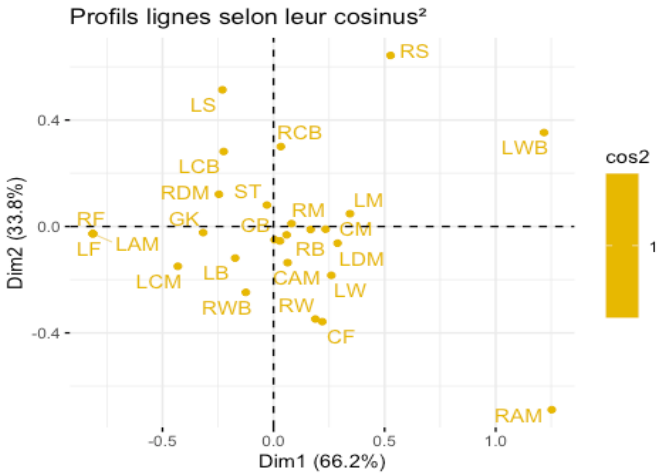


Le profil des "attaquants droits" (RS : right striker) et "attaquants gauches" (LS : left striker) caractérisent le côté positif de l'axe 1. Le profil des "ailiers droits" (RW : Right winger) caractérise le côté négatif de l'axe 1. De ce fait, la distribution de la morphologie des "ailiers droits" n'est pas la même que celle des attaquants.

Le côté négatif de l'axe 2 est caractérisé par le profil des "centres arrières gauches" (LCB : left center back) tandis que les "ailiers arrières gauches" (LWB : left winger back) caractérisent le côté positif de l'axe 2.

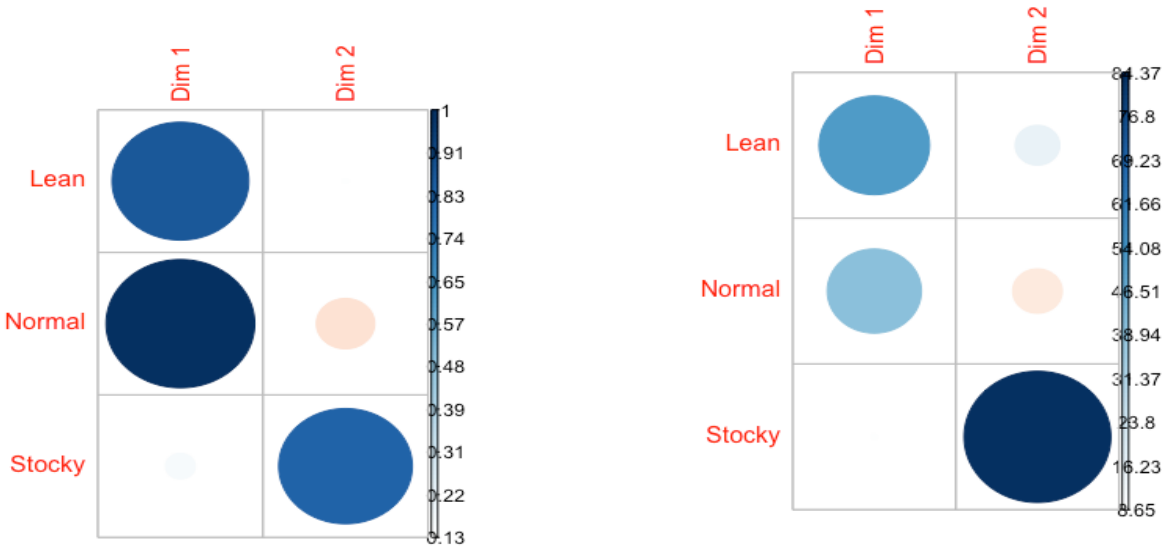
La représentation graphique des profils lignes sur le 1e plan factoriel est la suivante :

```
fviz_ca_row (res.ca, col.row = "cos2",title ="Profils lignes selon leur cosinus²",gradient.cols = c ("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)
```



Analyse des profils colonnes

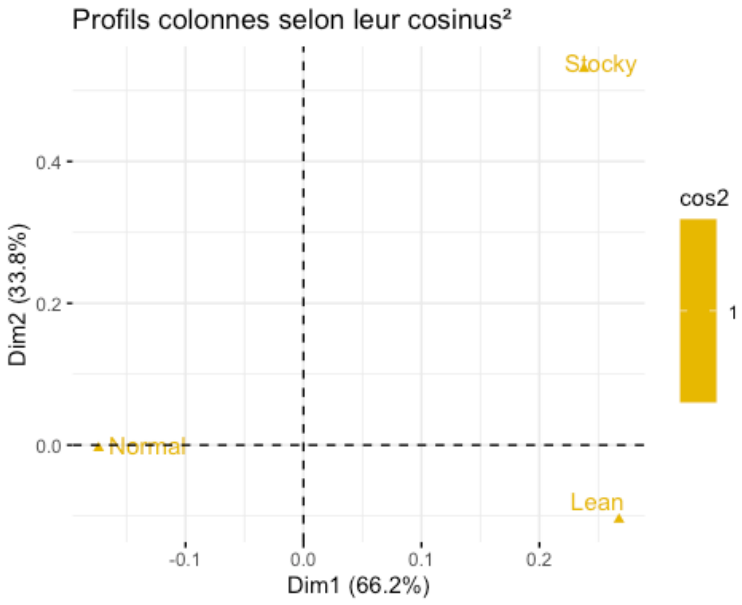
```
col<-get_ca_col(res.ca)
corrplot(col$cos2, is.corr = FALSE)
corrplot(col$contrib, is.corr=FALSE)
```



Le profil des "trapus" (stocky) caractérise le côté positif de l'axe 1 par opposition au profil "normal" qui caractérise le côté négatif de l'axe 1. Cela signifie que les trapus n'ont pas la même position sur le terrain que les joueurs ayant une morphologie normale. Le côté positif de l'axe 2 est caractérisé par le profil des "minces" (lean).

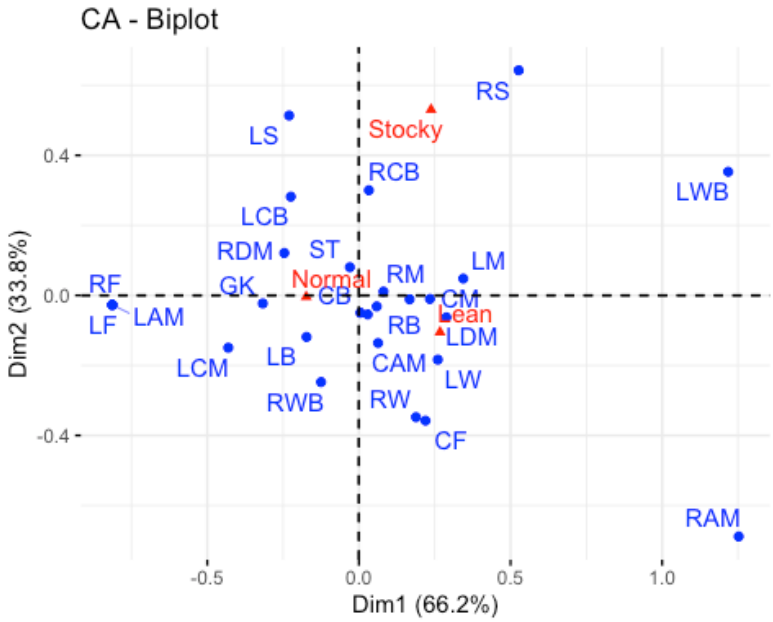
La représentation graphique des profils colonnes sur le 1e plan factoriel est la suivante :

```
fviz_ca_col(res.ca, col.col = "cos2",title ="Profils colonnes selon leur cosinus²",gradient.cols = c ("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)
```



Pour terminer, nous pouvons représenter simultanément les profils lignes et les profils colonnes.

```
fviz_ca_biplot (res.ca, repel = TRUE)
```



Ainsi, nous avons pu voir selon leur position sur le terrain, les joueurs n'avaient pas la même morphologie.

Clustering

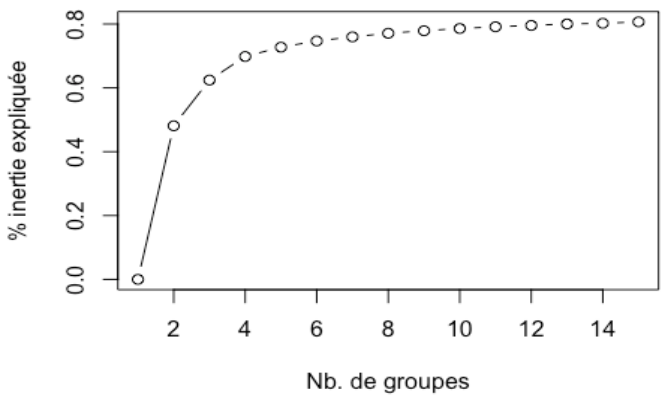
Pour terminer, nous allons réaliser une classification des joueurs en fonction des attributs que l'on avait utilisés lors de l'ACP. Comme précédemment, il n'est pas nécessaire de centrer et réduire les données puisqu'elles sont toutes dans la même unité.

Matrice des distances euclidiennes entre les individus

```
distances<-dist(d.active)
```

Détermination du nombre de classe à prendre par la méthode de l'inertie inter-classe

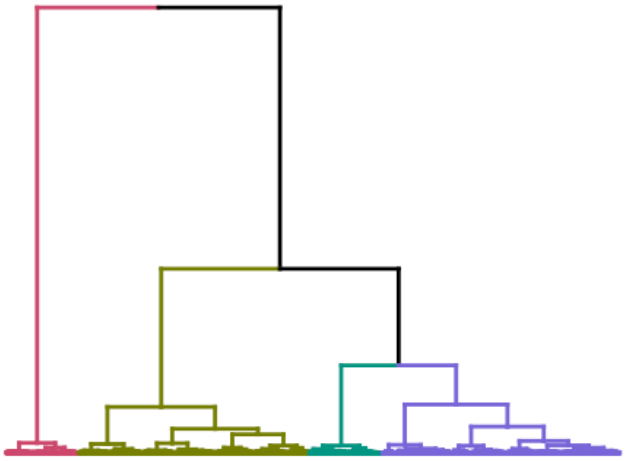
```
inertie.expl <- rep(0,times=15)
for (k in 2:15){
  clus <- kmeans(d.active,centers=k,nstart=5)
  inertie.expl[k] <- clus$betweenss/clus$totss
}
plot(1:15,inertie.expl,type="b",xlab="Nb. de groupes",ylab="% inertie expliquée")
```



A partir de k = 4 classes, l'adjonction d'un groupe supplémentaire n'augmente pas assez la part d'inertie expliquée par la partition. On choisit k=4.

CAH avec le critère de Ward

```
res.cahward<-hclust(distances,method="ward.D") #réalisation de la CAH
ggdendrogram(res.cahward)
ggplot(color_branches(res.cahward, k = 4), labels = FALSE, rect = TRUE)
```



```
groupes.cah<-cutree(res.cahward,k=4)
```

Au vu du nombre important de joueurs, un dendrogramme ne semble pas adapté ou du moins pas optimal.

Méthode des k-MEANS

```
res.kmeans<-kmeans(d.active,centers=4,nstart=5)
fviz_cluster(res.kmeans,d.active, palette = c("red", "blue", "orange","green",
"purple"), geom = "point",ellipse.type = "convex", ggtheme = theme_bw())
```

Nous pouvons remarquer qu'un cluster de joueurs semble se détacher des autres.

Conclusion

Pour conclure, nous pouvons dire que les caractéristiques des gardiens de but sont corrélées négativement aux caractéristiques des autres joueurs. De même, sur le nuage de point des clusters, ils sont regroupés entre eux. On peut donc conclure que les caractéristiques des joueurs de but sont bien spécifiques à leur poste et ne sont pas les mêmes que celles des joueurs conventionnels. Par ailleurs, nous avons mis en évidence un lien entre la position du joueur sur le terrain et sa morphologie.