UFR de Mathématiques
Master 1 Ingénierie Statistique et Informatique
de la Finance de l'Assurance et du Risque

Rapport Projet

Analyse de Données 2020
Sujet : Fifa 2020

Rédigé et présenté par :

Diamé SENGHOR
Sabrine DAMAK

Année académique : 2020/2021

Table des matières

1 Introduction

Dans le cadre de notre projet d'analyse de données, nous sommes amenés à effectuer une étude sur un jeu de données. Ainsi nous avons décidé de mener notre étude sur les 199 meilleurs joueurs de Fifa 2020. Etant tous les deux fans de football nous avons trouvé intéressant de joindre l'utile à l'agréable en menant cette étude. Notre jeu de données est extrait de Kaggle. Nous analyserons les relations existantes entre les variables et regarderons plus particulièrement, quelles variables explicatives influent le plus sur les salaires des joueurs. Afin d'observer si il existe une éventuelle explication aux différences de salaires des joueurs. Pour répondre à ces questions nous allons faire une étude détaillée de certaines variables mais aussi faire une analyse bivariée sur d'autres. La population étudiée est l'ensemble des 199 meilleurs jours du monde de l'année 2020. L'unité statistique est un joueur parmi les joueurs étudiés. La nature des variables est donnée par le tableau suibvant :

Variables quantitatives	**Variables qualitatives**
Age	Nom
Taille	Nationalité
Poids	Club
Valeur	Position
Release_clause	preferred_foot
Valeur	championnat
Salaire	
Drible	
Mouvement_sprint	
Mouvement_agility	

Notre jeu de données n'a pas de valeurs manquantes et nous pouvons le verifier par la commande suivante :

```
> sum(is.na(data))
0
```

2 Analyse Univariée

```
> mode=which.max(prop.table(table(data$club)))
Fc Bayern Munich
15
```

Les joueurs du Fc bayern de Munich sont les plus représentés parmis les 199 meilleurs joueurs de l'année 2020.

Nous pouvons voir un résumé des indicateurs de position des variables ci dessous :

```
> summary(data)
     nom                 age          taille.cm.      poids.kg.
 Length:199         Min.   :19.00   Min.   :163.0   Min.   :59.00
 Class :character   1st Qu.:25.00   1st Qu.:178.0   1st Qu.:72.00
 Mode  :character   Median :27.00   Median :183.0   Median :77.00
                    Mean   :27.32   Mean   :182.8   Mean   :77.45
                    3rd Qu.:30.00   3rd Qu.:188.0   3rd Qu.:83.00
                    Max.   :37.00   Max.   :199.0   Max.   :97.00
 nationalitÃ.          club             potentiel         valeur
 Length:199         Length:199         Min.   :84.00   Min.   :  6500000
 Class :character   Class :character   1st Qu.:85.00   1st Qu.: 30500000
 Mode  :character   Mode  :character   Median :87.00   Median : 36500000
                                       Mean   :87.26   Mean   : 40570352
                                       3rd Qu.:89.00   3rd Qu.: 47000000
                                       Max.   :95.00   Max.   :105500000
    salaire          position         preferred_foot     release_clause_eur
 Min.   : 15000   Length:199         Length:199         Min.   : 13000000
 1st Qu.: 78000   Class :character   Class :character   1st Qu.: 55200000
 Median :125000   Mode  :character   Mode  :character   Median : 68100000
 Mean   :141472                                         Mean   : 76580402
 3rd Qu.:190000                                         3rd Qu.: 92300000
 Max.   :565000                                         Max.   :195800000
   dribbling     movement_sprint_speed movement_agility
 Min.   : 1.00   Min.   :33.00         Min.   :34.0
 1st Qu.:68.50   1st Qu.:65.50         1st Qu.:62.0
 Median :81.00   Median :74.00         Median :75.0
 Mean   :69.67   Mean   :72.58         Mean   :72.6
 3rd Qu.:85.00   3rd Qu.:83.00         3rd Qu.:84.0
 Max.   :96.00   Max.   :96.00         Max.   :96.0
>
```

La moyenne et la médiane des variables "age", "poids", "potentiel" sont toutes égales ce qui montre une distribution symétrique des données.
Nous pouvons voir qu'elles ont un coéfficient d'asymetrie très faible (proche de zéro)

```
> skewness(data$age)
0.08474896
> skewness(data$poids.kg)
0.189138
> skewness(data$potentiel)
0.4889009
```

2.1 Variable qualitative POSITION

Intéressons nous un peu à la variable position

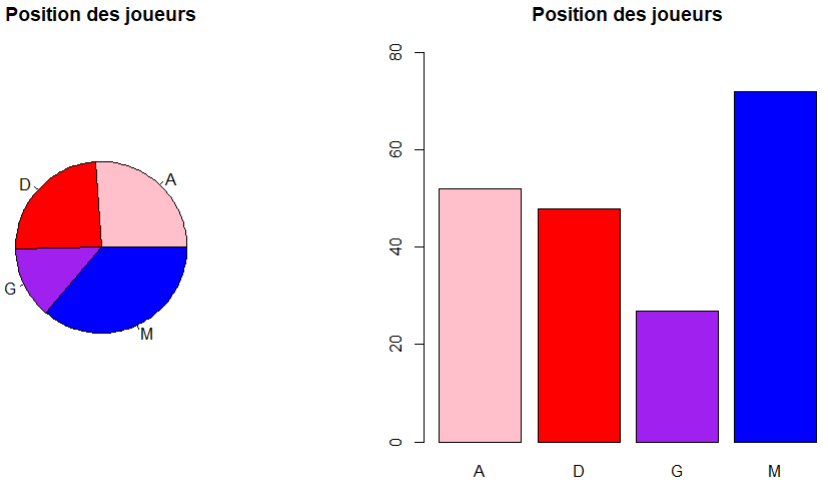
```
eff.position=table(data$position)
A  D  G  M
52 48 27 72
```

Parmi les 199 joueurs étudiés, on compte 52 attaquants, 48 défenseurs, 27 gardiens et 72 milieux de teraain

```
prop.position=prop.table(table(data$position))
A         D         G         M
0.2613065 0.2412060 0.1356784 0.3618090
```

On voit directement au vue des résultats que les Milieux et les Attaquants ont une plus grande proportion que les autres postes.

```
par(mfrow=c(1,2))
pie(prop.position, main='Position des joueurs', col= c("pink","red","purple","blue",14,10))
barplot(eff.position, main="Position des joueurs", ylim = c(0, 80), col=c("pink","red","purple","blue",14,10))
```



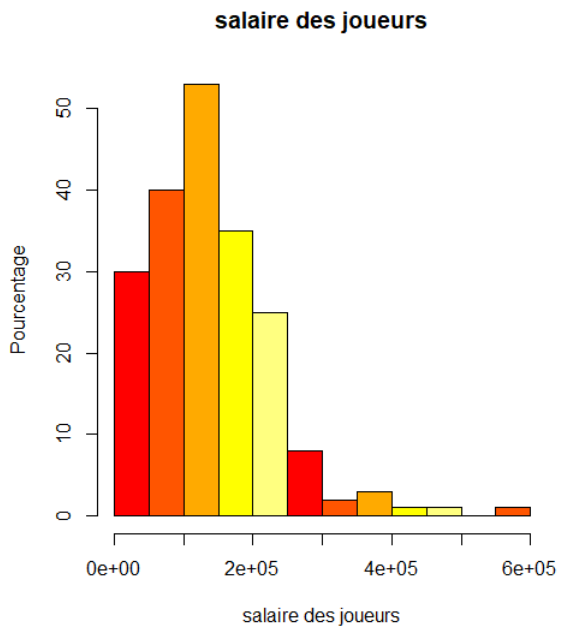
Avec les graphiques, le résultat est beaucoup plus représentatif. On observe bien sur le cercle et le diagramme que les milieux et les attaquants représentent la plus grosse part des joueurs dans notre échantillon.
Intéressons nous maintenant à la variable SALAIRE

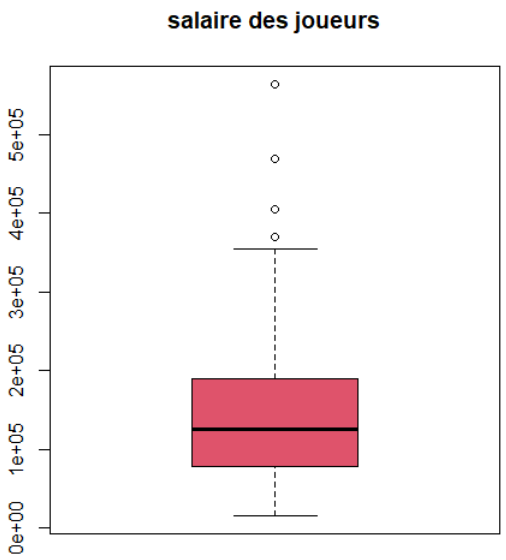
2.2 Variable quantitative SALAIRE

```
summary(data$salaire)
```

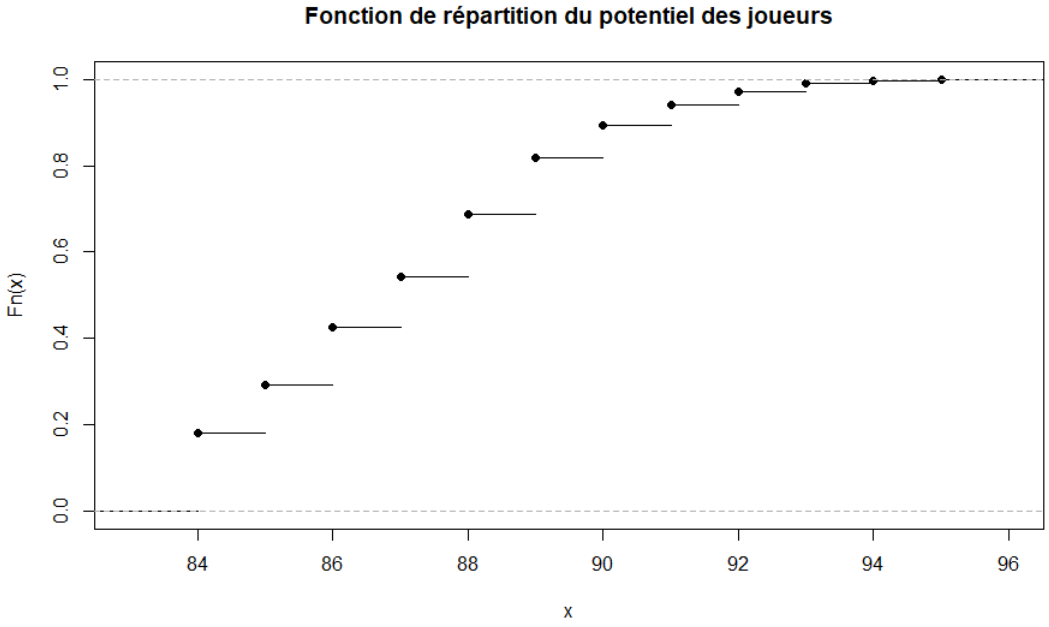
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15000	78000	125000	141472	190000	565000

```
par(mfrow=c(1,2))
hist(data$salaire,main="salaire des joueurs",xlab='salaire des joueurs',ylab='Pourcentage',col=heat.colors(5))
boxplot(data$salaire,main="salaire des joueurs",col=10)
```





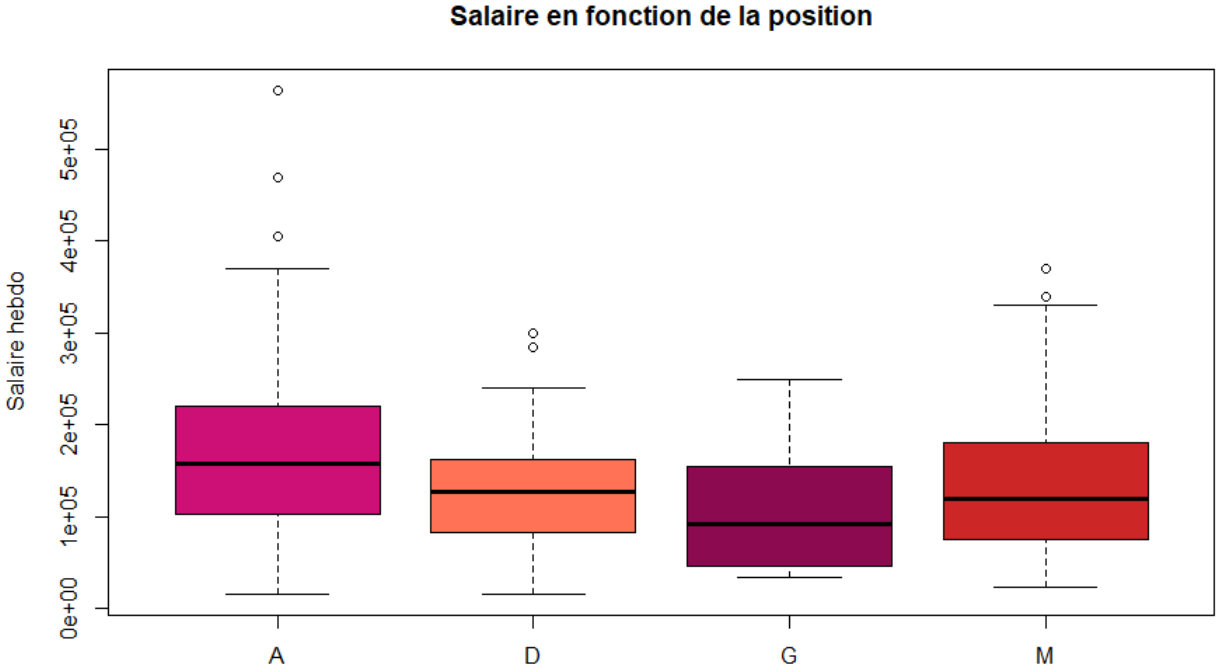
Le salaire des joueurs est plutôt hétérogène : le joueur le mieux payé gagne jusqu'à plus 500 mille euros tandis que le moins payé a un salaire qui tourne autour de 15 mille euros. On rappelle que les salaire sont hebdomadaires.L'étendue est donc très importante. Le salaire moyen est de 120 mille euros environ. On utilise une boîte à moustache car elle permet une bonne représentation, en effet, elle est délimitée par le premier et le troisième quartile. On remarque aussi que plus de la moitié" des joueurs (environ 60%) a un salaire superieur à 100 mille euros,
Voici la fonction de répartition des potentiels des joueurs



3 Analyse Bivariée

3.1 Lien Variables Quantitative/Qualitative

Nous étudions dans cette étape le rapport pouvant exister entre le salaire du joueur et sa postition sur le terrain.
>boxplot(data$salaire data$position,main='Salaire en fonction de la position',ylab='Salaire hebdo',col=c("deeppink3","coral1","deeppink4","firebrick3"))



Grâce aux boites à moustaches ci-dessus, nous observons des différences de salaires des joueurs en fonction du poste occupé, ces boîtes à moustaches illustrent des données asymétriques. On observe également certaines valeurs "atypiques" sauf pour les gardiens. Pour plus de détails on exécute la commande tapply pour déterminer la moyenne conditionnelle des salaire selon la position.
tapply(data$salaire,data$position,mean)

A	D	G	M
177500.0	130562.5	109518.5	134708.3

Le salaire moyen des joueurs est très different selon la position. En effet on remarque clairement que les attaquants gagnent beaucoup plus que les autres en moyenne suivi des milieux de terrain et enfin les défenseurs et les gardiens qui sont en derniere position. Néamoins parmi les attaquants, les défenseurs et les milieux il y a un grand écart à cause de certains joueurs qui sont largement au dessus de la moyenne comme Messi et Christiano pour les attaquants, Ramos et Van Djik pour les défenseurs et efin De Bruyne et Modric pour les milieux. Ces joueurs sortent du lot parce qu'ils ont des salaires très supérieurs à la moyenne ce qui s'explique parce qu'ils font partis des meilleurs joueurs au monde et cumulent presque toutes les récompenses de l'année. Pour plus de détails faisons un résumé des indicateurs de position du salaire selon la position.

tapply(data$salaire,data$position,summary)

```
$A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15000  103750  157500  177500  220000  565000

$D
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16000   85500  127500  130563  161250  300000

$G
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34000   46000   92000  109519  155000  250000

$M
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23000   75250  120000  134708  180000  370000

>
```

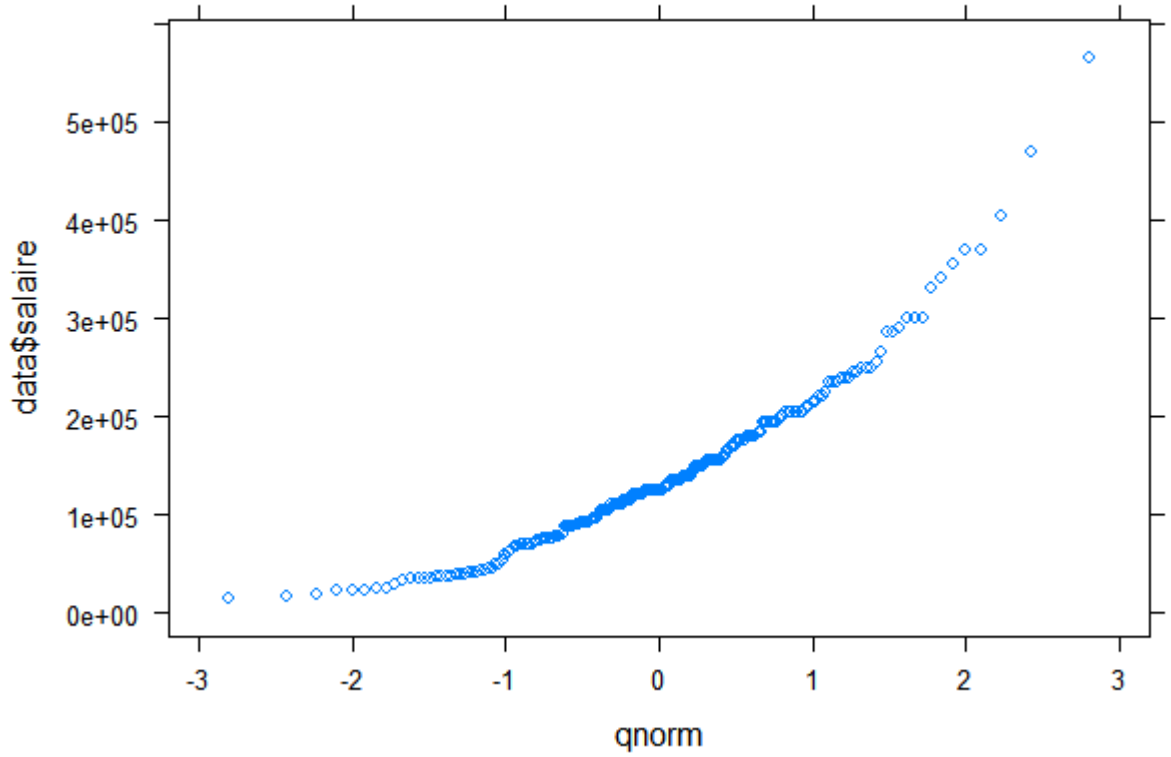
On voit directement que le joueur ayant le plus gros salaire est un attaquant s'élevant à 565 mille euros et que le plus petit est aussi pour un attaquant s'élevant à 15 mille euros. Concernant les salaires minimum,le plus grand des salaires minimum est 34 mille euros pour les gardiens. Généralement, les attaquants ont tendances à avoir de plus gros salaires et une étendue plus importante que les autres postes (1st Qu. :103750 - 3rd Qu. :22000) tandis que les défenseurs ont tendance à gagner beaucoup moins et dont les données sont les moins dispersés (1st Qu. :85500 - 3rd Qu. :161250). On peut donc supposer qu'il existe un lien entre le salaire et la position occupée. Pour déterminer si cette hypothèse est vérifiée faisons un test de Fisher avec comme hypothèse H0 (Les deux variables sont indépendantes).

```
summary(lm(data$salaire data$position))
```

```
Call:
lm(formula = data$salaire ~ data$position)

Residuals:
    Min      1Q  Median      3Q     Max
-162500  -60541  -12500   45292  387500

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       177500      11731  15.131  < 2e-16 ***
data$positionD    -46938      16932  -2.772 0.006110 **
data$positionG    -67982      20067  -3.388 0.000852 ***
data$positionM    -42792      15395  -2.780 0.005976 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84590 on 195 degrees of freedom
Multiple R-squared:  0.0694,    Adjusted R-squared:  0.05509
F-statistic: 4.848 on 3 and 195 DF,  p-value: 0.002822
```

La p-value est très faible et inférieure au seuil(5%) donc on peut rejeter l'hypothèse d'indépendance car la probabilité de se tromper est trop faible. On peut donc dire que le salaire et la position d'un joueur sont dépendantes si le test de Fisher est valide c'est à dire la distribution conditionnelle de la variable position sur la modalité salaire est gaussienne.Pour cela vérifions si les résidus sont gaussiens. Traçons d'abord le diagramme Quantile-Quantile plot

qqmath(data$salaire)



Ce graphique peut aussi être retrouvé par la commande suivante plot(lm(data*salaire* *data*position))

D'après le graphe des QQ-plot il semblerait que la distribution est gaussienne. Vérifions cela avec un test statistique pour pouvoir conclure.
shapiro.test(residuals(lm(data*salaire* *data*position)))

```
        Shapiro-Wilk normality test

data:  residuals(lm(data$salaire ~ data$position))
W = 0.94091, p-value = 2.917e-07
```

Avec une p-value très inférieure au seuil(0.05) on ne peut qu'infirmer notre hypoyhèse de gaussienneté des résidus. Par conséquent notre test de Fisher n'est pas valide et donc on ne peut pas confirmer un probable lien entre les variables salaires et position même si toutes les observations semblent montrer le contraire.

3.2 Lien Variables Qualitative/Qualitative

A priori on pourrait penser que la position d'un joueur et le championnat n'a pas de rapport malgrè que chaque entraineur adopte sa tactique autrement dit comment les joueurs de son

équipe se positionnent sur le terrain. Intéressons nous à étudier les deux variables qualitatives Position et Championnat pour voir si elles ont un lien.
Pour cela commençons par afficher le tableau de fréquences des deux variables

3.2.1 Tableau de contingence

```
table(data$position,data$championnat)
```

```
  Bundesliga CSL Eredivisie Liga Liga_NOS Ligue1 MLS PremierLigue Serie_A
A          4   0          1   13        0      6   1           13      14
D          8   0          0   12        3      2   0           13      10
G          7   0          0    6        0      2   0            7       5
M         13   5          1   17        2      3   0           20      10

  Superligue
A          0
D          0
G          0
M          1
```

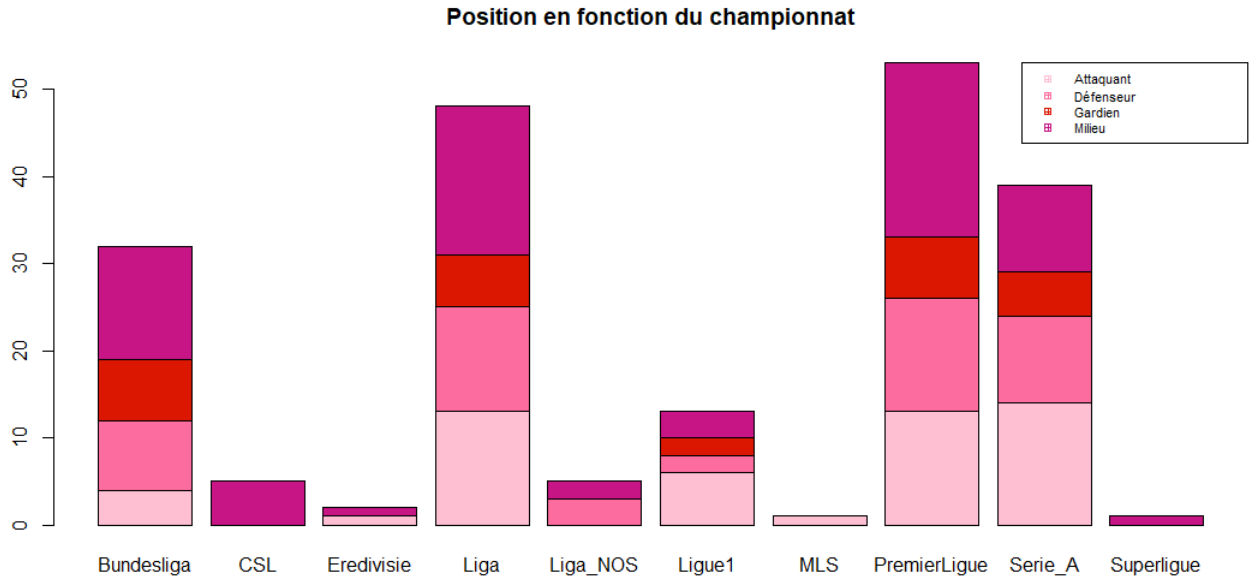
3.2.2 Distibution conditionnelle

```
prop.table(table(data$position,data$championnat))
```

```
   Bundesliga         CSL  Eredivisie        Liga    Liga_NOS      Ligue1
A 0.020100503 0.000000000 0.005025126 0.065326633 0.000000000 0.030150754
D 0.040201005 0.000000000 0.000000000 0.060301508 0.015075377 0.010050251
G 0.035175879 0.000000000 0.000000000 0.030150754 0.000000000 0.010050251
M 0.065326633 0.025125628 0.005025126 0.085427136 0.010050251 0.015075377

           MLS PremierLigue     Serie_A  Superligue
A 0.005025126  0.065326633 0.070351759 0.000000000
D 0.000000000  0.065326633 0.050251256 0.000000000
G 0.000000000  0.035175879 0.025125628 0.000000000
M 0.000000000  0.100502513 0.050251256 0.005025126
```

Représentons les deux variables sur un graphique.

```
barplot(table(data$ position,data$ championnat),main = 'Position en fonction du championnat',col = c("#FEBFD2","#FD6C9E","#DB1702","#C71585"))
par(xpd=TRUE)
legend("topright",expression('Attaquant','Défenseur','Gardien','Milieu'),pch=c(12),
col =c("#FEBFD2","#FD6C9E","#DB1702", "#C71585"), cex=0.7)
```



Au vu du schéma on a indédépendance entre la postition et le championat.Poussons notre étude pour confirmer cette hypothèse d'indépendance avec un test khi-deux.

```
chisq.test(data$ position,data$ championnat)
```

```
        Pearson's Chi-squared test

data:  data$position and data$championnat
X-squared = 29.373, df = 27, p-value = 0.3431
```

Test d'indépendance du Khi-deux
La P-value obtenue nous permet de valider notre hypothèse nulle (H0) car elle est supérieure au seuil (5%).La probabilité de se tromper en réjetant l'hypothèse d'indépendance est très forte. On peut donc dire que la postition d'un joueur n'a pas de lien avec le championnat dans lequel il joue.
Validation du test de Khi-deux

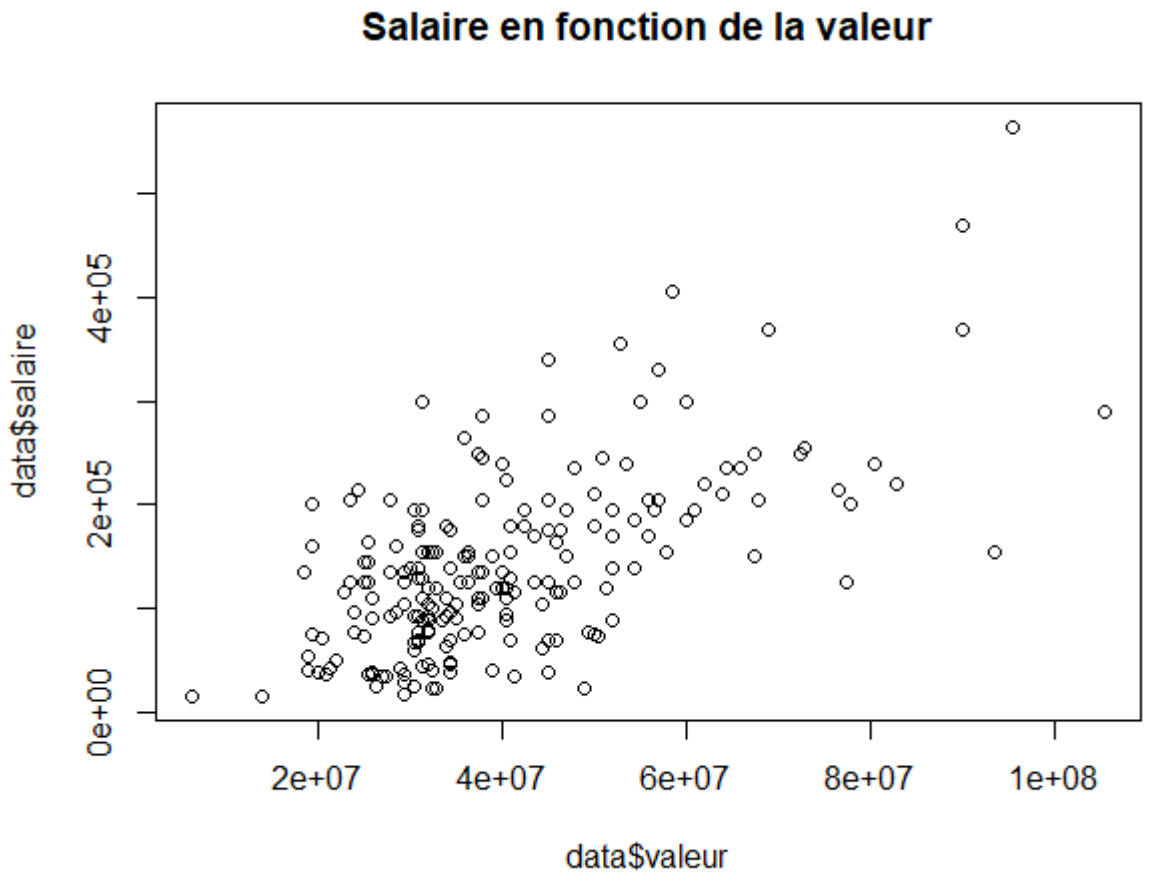
```
> chisq.test(data$position,data$championnat)$expected
             data$championnat
data$position Bundesliga      CSL Eredivisie      Liga Liga_NOS   Ligue1
            A   8.361809 1.306533  0.5226131 12.542714 1.306533 3.396985
            D   7.718593 1.206030  0.4824121 11.577889 1.206030 3.135678
            G   4.341709 0.678392  0.2713568  6.512563 0.678392 1.763819
            M  11.577889 1.809045  0.7236181 17.366834 1.809045 4.703518
             data$championnat
data$position       MLS PremierLigue   Serie_A Superligue
            A 0.2613065    13.849246 10.190955  0.2613065
            D 0.2412060    12.783920  9.407035  0.2412060
            G 0.1356784     7.190955  5.291457  0.1356784
            M 0.3618090    19.175879 14.110553  0.3618090
Warning message:
```

Notre test n'est pas valide car on constate qu'il y a des effectifs théoriques inférieurs à 5.

3.3 Lien Variables Quantitative/Quantitative

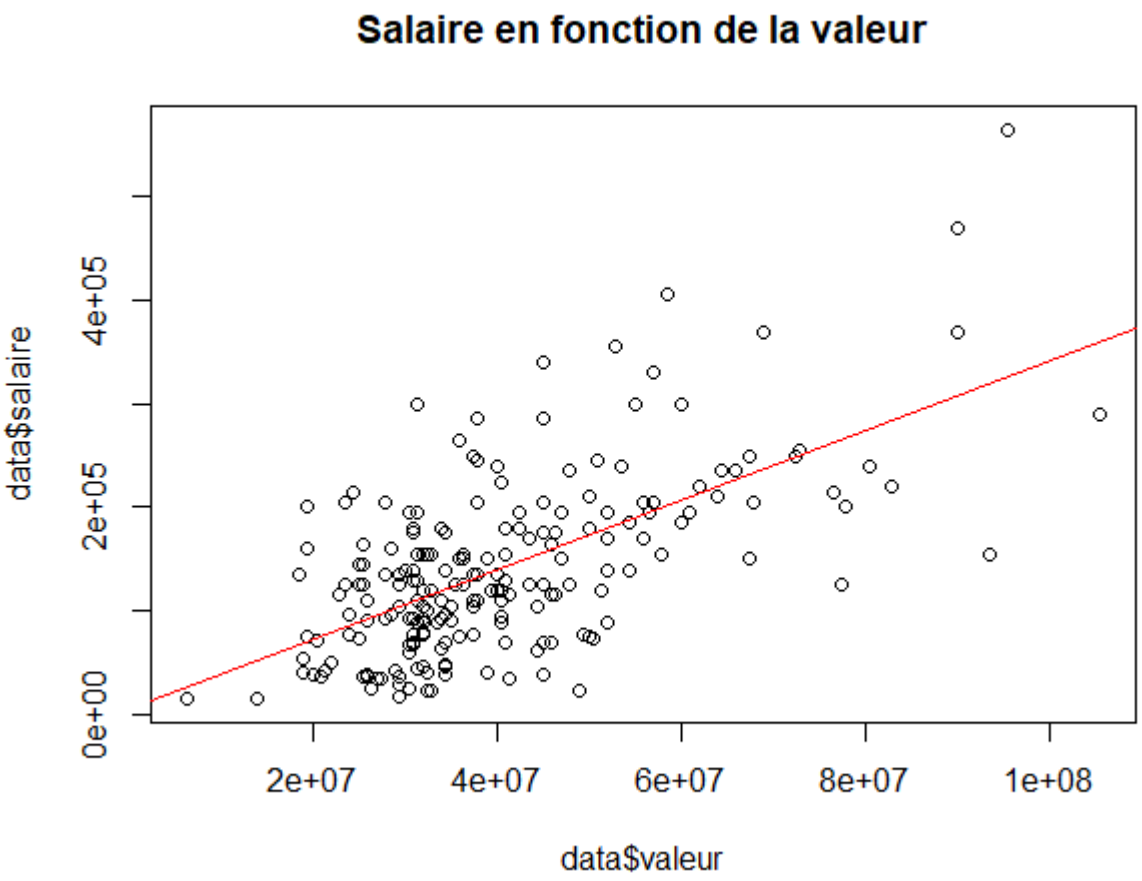
Dans cette partie nous essayerons de faire une étude entre les deux variables salaire et valeur des joueurs pour déterminer si l'une a une influence sur l'autre. Tout d'abord, essayons de représenter le graphe des salaires et la valeur d'un joueur pour voir si les deux variables ont un lien.

```
plot(data$valeur,data$salaire, main='Salaire en fonction de la valeur')
```



D'après le graphe, il semble qu'il y ait un lien entre salaire et valeur. Le nuage de points semble suivre une droite. Ajoutons au graphe la droite de régression linéaire en utilisant la méthode des moindres carrées.

```
a <- cov(data$valeur,data$salaire)/var(data$valeur)
b <- mean(data$salaire)-a*mean(data$valeur)
abline(b,a,col="red")
```



Cette droite de regression semble confirmer la relation linéaire existante entre ces deux variables. Afin de valider cette interprétation , on fait un test de significativité du coefficient de corrélation de Pearson :
cor(data$salaire,data$valeur)

0.6246238

On peut dire qu'il ya un lien entre salaire et valaur au vu du désultat du test.En effet le coefficient de corrélation est largement supérieur à 0. Par ailleurs certains points du nuage sont proches de la ligne, mais d'autres en sont éloignés, ce qui indique seulement une relation linéaire modérée entre les variables. Notre corrélation n'est pas statistiquement significative. Vérifions cette validité par un test de fisher

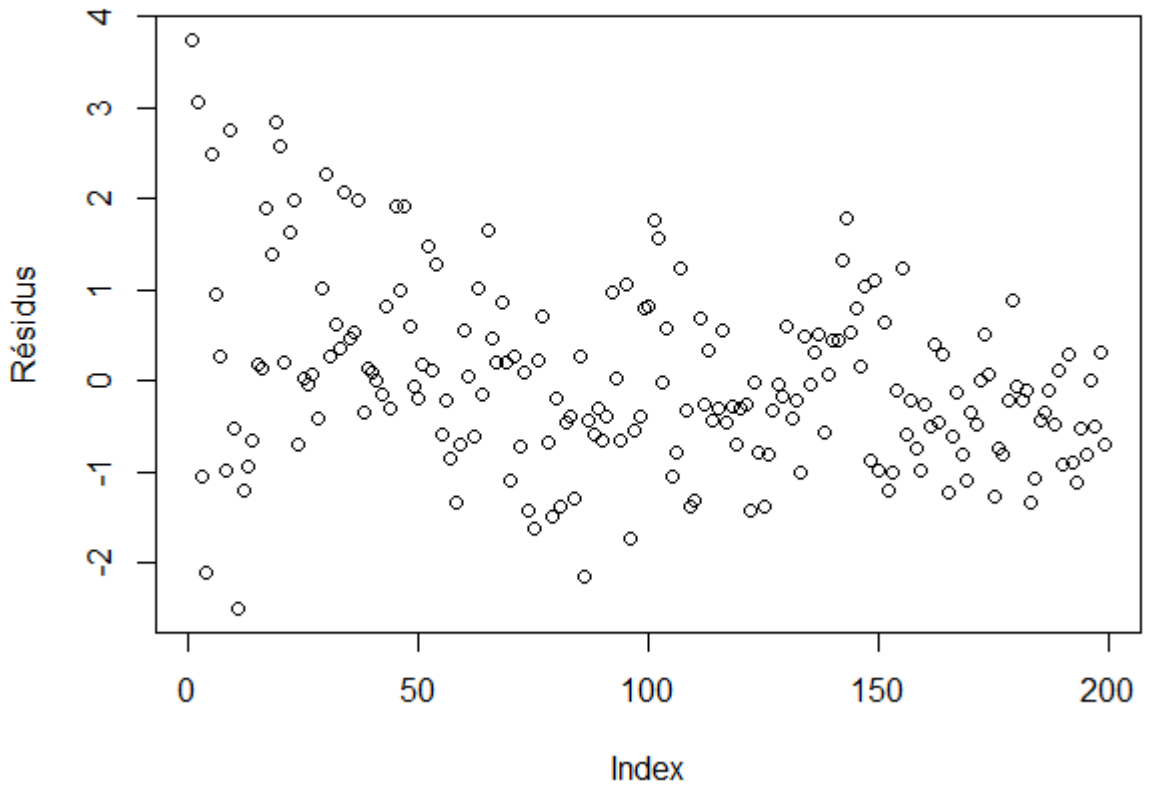
summary(lm(data*salaire* *data*valeur))

```
Call:
lm(formula = data$salaire ~ data$valeur)

Residuals:
    Min      1Q  Median      3Q     Max
-163788  -44762   -9461   33976  239512

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.561e+03  1.303e+04   0.427     0.67
data$valeur 3.350e-03  2.984e-04  11.226   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68130 on 197 degrees of freedom
Multiple R-squared:  0.3902,    Adjusted R-squared:  0.3871
F-statistic:   126 on 1 and 197 DF,  p-value: < 2.2e-16
```

Le test de Fisher nous donne un coefficient de détermination égale à 0.3902 ce qui n'est pas négligeable. En effet près de 39% de la variance de salaire est expliquée par la régression linéaire.

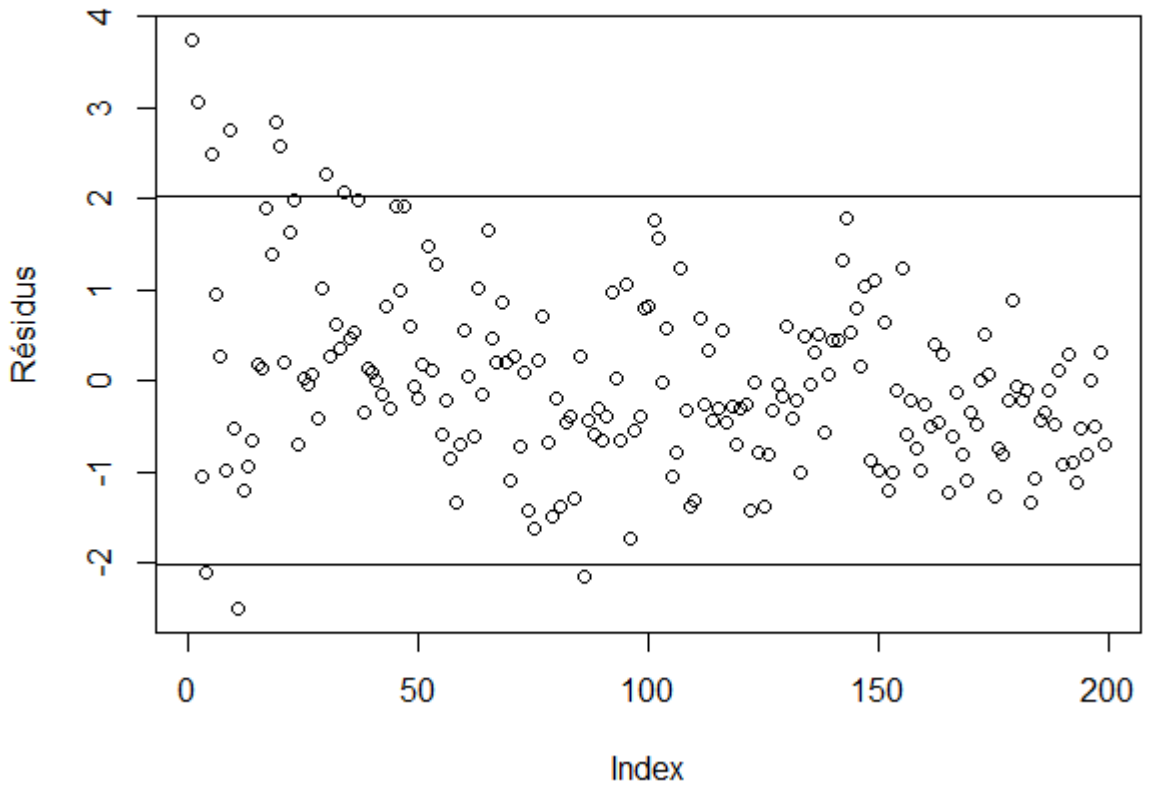
Analysons maintenant les résidus

```
reg=lm(data$salaire~data$valeur)
reg
residus=rstudent(reg)
residus
```



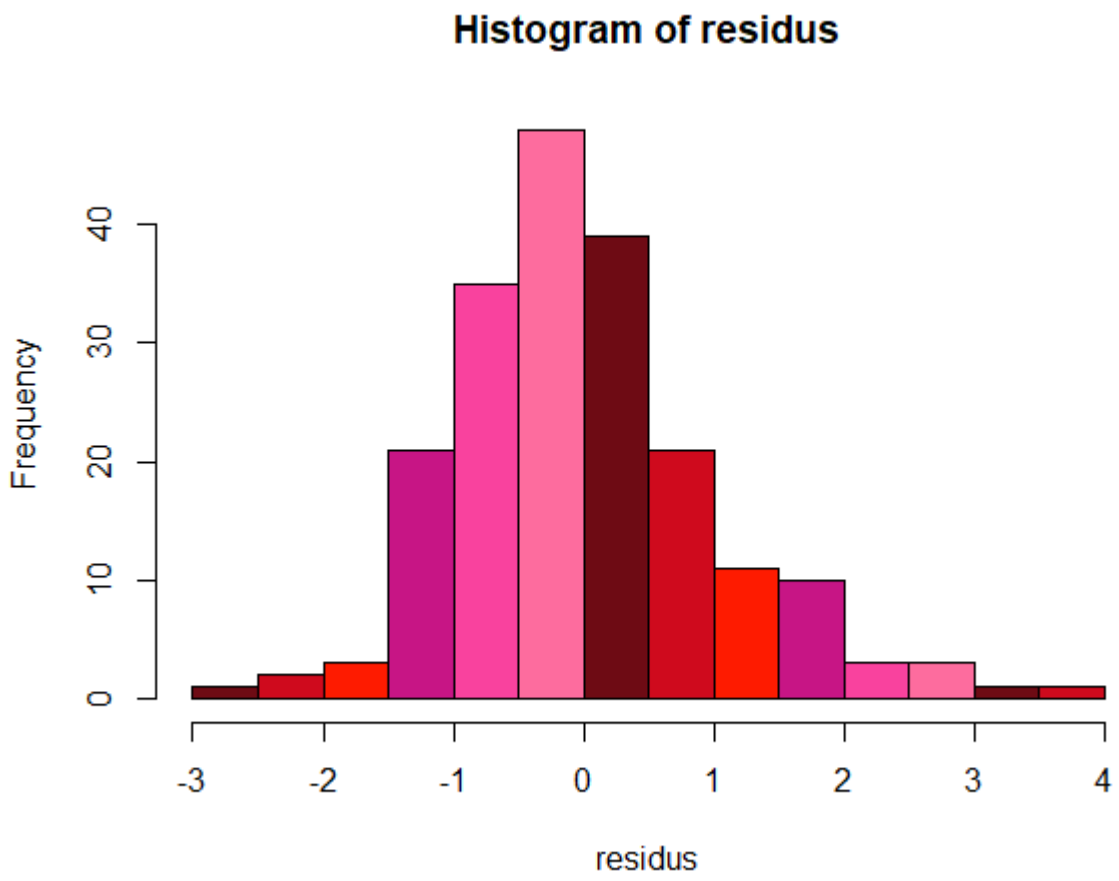
On peut observer une structure particulière dans le nuage des résidus. Cela signifie qu'il reste une information dans les résidus que la relation linéaire ne prend pas en compte. Par conséquent le modèle est mal adapté.
Afin de localiser les points du nuage qui sont mal expliqués par la relation linéaire, on trace les deux droites parallèles à l'axe des abscisses

```
plot(residus,ylab = 'Résidus')
sigma=sd(residus)
sigma
abline(-2.026,0,2.026)
```



On constate qu'il existe des points mal expliqués par la relation linéaire et qu'il existe encore une information dans les résidus qui n'est pas prise en compte par l'ajustement linéaire de la variable valeur par rapport à la variable salaire.
On peut constater que 10 individus sont en dehors de l'intervalle ce qui fait que plus de 95% des individus sont bien représentés dans le nuage des résidus.Les individus mal représentés sont des individus extrêmes.
Maintenant représentons les résidus sous forme d'histogramme pour essayer de déterminer la nature de sa distribution.

```
hist(residus,col = c("#6E0B14", "#CF0A1D", "#FE1B00","#C71585", "#F9429E", "#FD6C9E"))
```



Cette distribution semble être une distribution gaussienne
Vérifions cela par un test Shapiro.
res=residuals(reg) res shapiro.test(res)
Shapiro-Wilk normality test

data : res
W = 0.96504, p-value = 7.561e-05
On ne peut pas confirmer que la distribution des individus est gaussiènne car la p-value est très faible(inférieure au seuil 0,05) ce qui contredit notre hypothèse de départ(H0). La probabilité de se tromper en rejetant l'hypoyhèse nulle(hypothèse de normalité) est très faible.Donc l'hypoyhèse de normalité est rejetée.

4 Analyse en Composantes Principales

L'analyse des composantes principales (ACP) est une méthode pour bien simplifier notre jeu de données et nous permet de traiter plusieurs variables en même temps. De plus, on l'utilise pour trouver les situations d'indépendance entre les variables. Nous allons réaliser une ACP réduite sur toutes nos variables quantitatives.
res.acp=prcomp(data1,scale=T,center=T)
res.acp

4.0.1 Matrice de corrélation

D'abord, on retire quelques variables quantitatives (taille(cm) , poids(kg) ,potentiel ,valeur ,salaire ,release_clause_eur, dribbling, movement_sprint_speed, movement_agility) de notre jeu de données original pour créer un nouveau jeu de données de 10 variables et 199 observations pour l'ACP. On utilise ce jeu pour créer la matrice de corrélations ci - dessous.

```
cor(data1)
```

```
                            age  taille.cm.   poids.kg.    potentiel      valeur
age                  1.00000000  0.03045009  0.09765381 -0.418807178 -0.25195029
taille.cm.           0.03045009  1.00000000  0.80738629  0.066394533 -0.19473595
poids.kg.            0.09765381  0.80738629  1.00000000  0.110022919 -0.08470958
potentiel           -0.41880718  0.06639453  0.11002292  1.000000000  0.75777014
valeur              -0.25195029 -0.19473595 -0.08470958  0.757770142  1.00000000
salaire              0.24515928 -0.19438073 -0.05326219  0.453113384  0.62462375
release_clause_eur  -0.27585924 -0.19265343 -0.08483205  0.764503215  0.97215718
dribbling           -0.10852751 -0.55316524 -0.53362832 -0.007674849  0.22884587
movement_sprint_speed -0.30100646 -0.43163673 -0.39272177  0.134716967  0.26335768
movement_agility    -0.11187728 -0.74483139 -0.70093112  0.040234289  0.29432335
                        salaire release_clause_eur    dribbling
age                  0.24515928        -0.27585924 -0.108527505
taille.cm.          -0.19438073        -0.19265343 -0.553165237
poids.kg.           -0.05326219        -0.08483205 -0.533628316
potentiel            0.45311338         0.76450321 -0.007674849
valeur               0.62462375         0.97215718  0.228845872
salaire              1.00000000         0.63253537  0.222186458
release_clause_eur   0.63253537         1.00000000  0.196241839
dribbling            0.22218646         0.19624184  1.000000000
movement_sprint_speed  0.13759428         0.23151655  0.653011495
movement_agility     0.26798926         0.26315840  0.752555973
                    movement_sprint_speed movement_agility
age                           -0.3010065      -0.11187728
taille.cm.                    -0.4316367      -0.74483139
poids.kg.                     -0.3927218      -0.70093112
potentiel                      0.1347170       0.04023429
valeur                         0.2633577       0.29432335
salaire                        0.1375943       0.26798926
release_clause_eur             0.2315165       0.26315840
dribbling                      0.6530115       0.75255597
movement_sprint_speed          1.0000000       0.67947261
movement_agility               0.6794726       1.00000000
```

Vérifions si la somme des valeurs propres de la matrice de corrélation fait 10 qui est le nombre de variable et la trace de la matrice de corrélation.

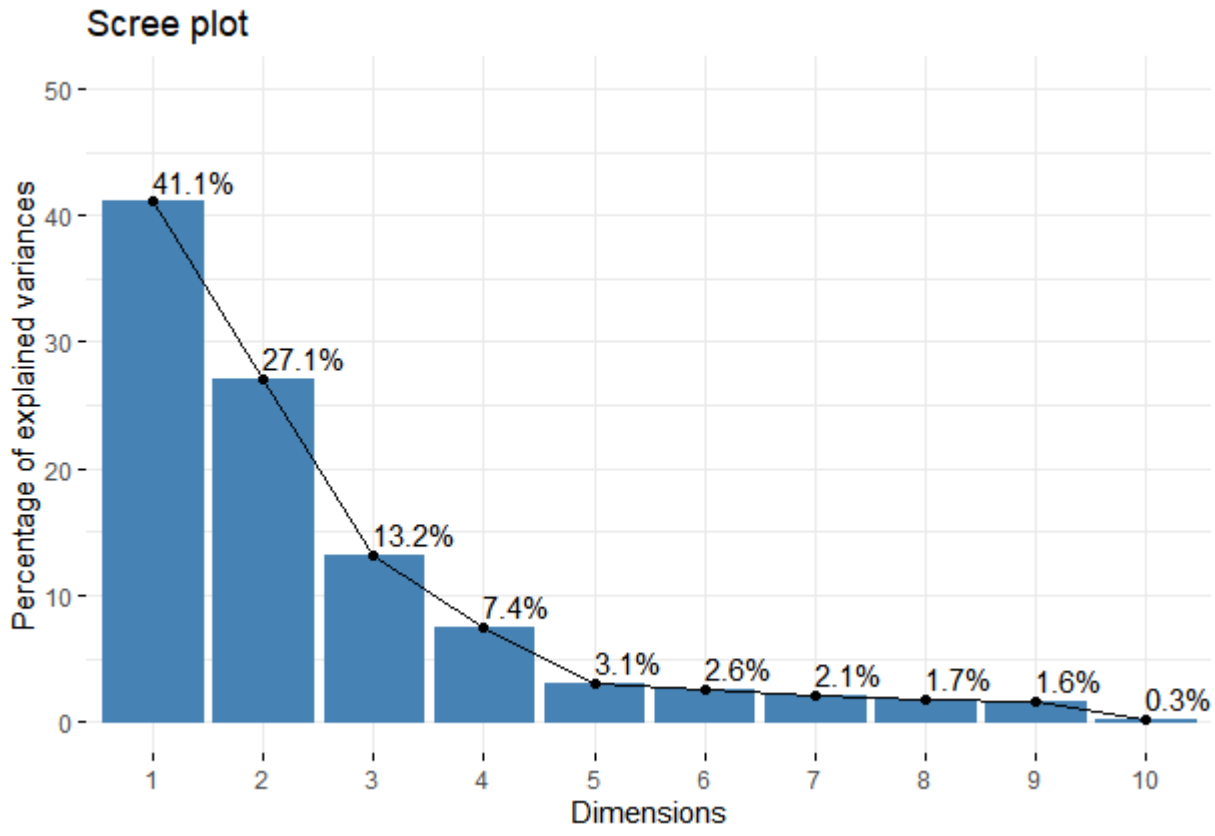
```
> sum(eigen(cor(data1))$values)
10
```

```
> summary(res.acp)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.0263 1.6447 1.1473 0.86092 0.55593 0.50821 0.45325 0.41546
Proportion of Variance 0.4106 0.2705 0.1316 0.07412 0.03091 0.02583 0.02054 0.01726
Cumulative Proportion  0.4106 0.6811 0.8127 0.88687 0.91777 0.94360 0.96414 0.98140
                          PC9   PC10
Standard deviation     0.4012 0.1582
Proportion of Variance 0.0161 0.0025
Cumulative Proportion  0.9975 1.0000
> |
```

Si nous nous intéressons à la part d'inertie totale expliquée par les composantes principales, on voit bien que notre premier composante principale explique 41,06% de l'inertie totale, notre deuxième composante 27,05%, la troisème quant à elle explique 13,16% de l'inertie. On peut aussi obtenir ce résultat de manière plus claire par cette commande :

```
> get_eigenvalue(res.acp)
       eigenvalue variance.percent cumulative.variance.percent
Dim.1  4.10588446       41.0588446                    41.05884
Dim.2  2.70519068       27.0519068                    68.11075
Dim.3  1.31639716       13.1639716                    81.27472
Dim.4  0.74118476        7.4118476                    88.68657
Dim.5  0.30905448        3.0905448                    91.77712
Dim.6  0.25827876        2.5827876                    94.35990
Dim.7  0.20543308        2.0543308                    96.41423
Dim.8  0.17260410        1.7260410                    98.14027
Dim.9  0.16095586        1.6095586                    99.74983
Dim.10 0.02501667        0.2501667                   100.00000
> |
```

Les trois premières composantes expliquent 81,27% de l'inertie totale on va donc gerder ces trois composantes. Le graphique des valeurs propres peut être généré à l'aide de la fonction fviz_eig()



4.0.2 Matrice des contributions des individus à la construction des axes principaux

```
> (z^2%*%diag(1/res.acp$sdev)^2*(1/199))[1:10,1:4]
              [,1]          [,2]          [,3]          [,4]
 [1,] 0.0469956519 0.0239933526 4.804458e-02 9.040435e-05
 [2,] 0.0067653410 0.0081356157 2.120733e-02 3.164106e-02
 [3,] 0.0396826804 0.0135260384 2.301622e-03 1.885272e-03
 [4,] 0.0001147907 0.0384662006 7.882801e-04 1.174464e-02
 [5,] 0.0351909155 0.0149751532 1.694176e-02 5.785705e-04
 [6,] 0.0181466814 0.0160765808 9.201136e-03 1.156382e-03
 [7,] 0.0006105664 0.0449262244 1.399730e-03 1.841079e-02
 [8,] 0.0009760921 0.0284217597 6.254789e-05 1.221422e-02
 [9,] 0.0084375633 0.0001456736 2.568775e-02 2.936502e-03
[10,] 0.0203036431 0.0029326232 6.134598e-04 3.941420e-05
>
```

On peut voir sur la matrice que certains individus (comme le premier) contribuent plus que d'autres à la construction des axes principaux.Ce qui est logique et que nous allons tenter de le montrer plus clairement sur la représentation du plan factoriel.Ici nous avons pris un extrait de 10 individus.

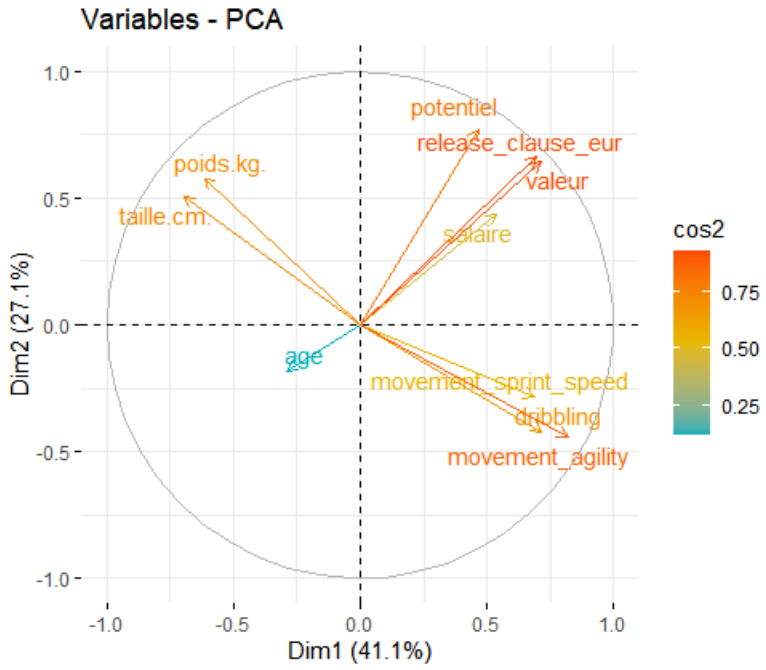
4.0.3 Matrice des contributions des axes principaux à la représentation des individus (matrice des cos2(ϕ)).

```
> (diag(1/rowSums(z^2))%*%z^2)[1:10,1:4]
              PC1          PC2          PC3          PC4
 [1,] 0.594446869 0.199957390 0.1948408358 0.0002064260
 [2,] 0.232826284 0.184469213 0.2339960672 0.1965680967
 [3,] 0.752960944 0.169095958 0.0140018561 0.0064575135
 [4,] 0.003553704 0.784594089 0.0078241176 0.0656347721
 [5,] 0.678379619 0.190197437 0.1047082705 0.0020133452
 [6,] 0.541658336 0.316164471 0.0880541586 0.0062308789
 [7,] 0.017327059 0.840007187 0.0127355074 0.0943157304
 [8,] 0.040961182 0.785821282 0.0008415396 0.0925267139
 [9,] 0.426199142 0.004848056 0.4160081416 0.0267759950
[10,] 0.853593245 0.081231491 0.0082688162 0.0002991228
>
```

L'individus 1 sera le mieux representé sur le plan factoriel (1.2) car sa proportion sur PC1 et PC2 est assez importante (0.6+0.2=0.8)

4.0.4 Premier Cercle de corrélation

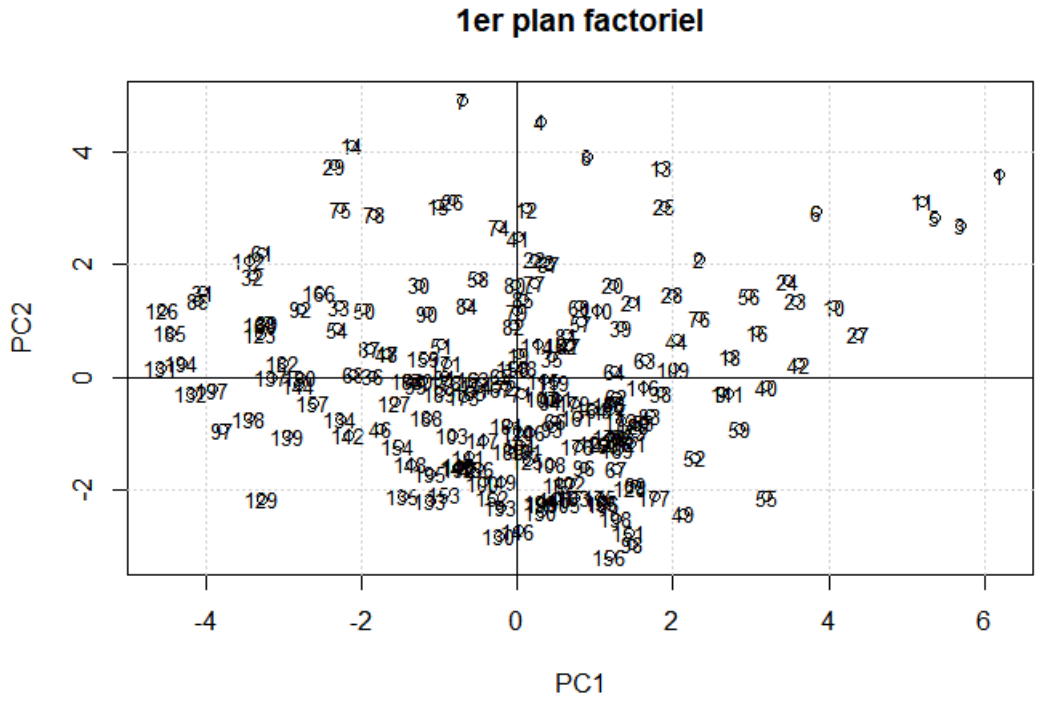
Le cercle de corrélation sert à représenter la matrice de corrélation. Les variables sont représentées dans la base construite à partir des composantes principales qui ont été normalisées, et qui constitue une base orthonormée de l'espace des variables. Les variables positivement corrélées sont regroupées. Une variable est d'autant mieux représentée sur un axe qu'elle est proche du bord du cercle des corrélations et de l'axe, d'autant plus mal représentée qu'elle est proche de l'origine



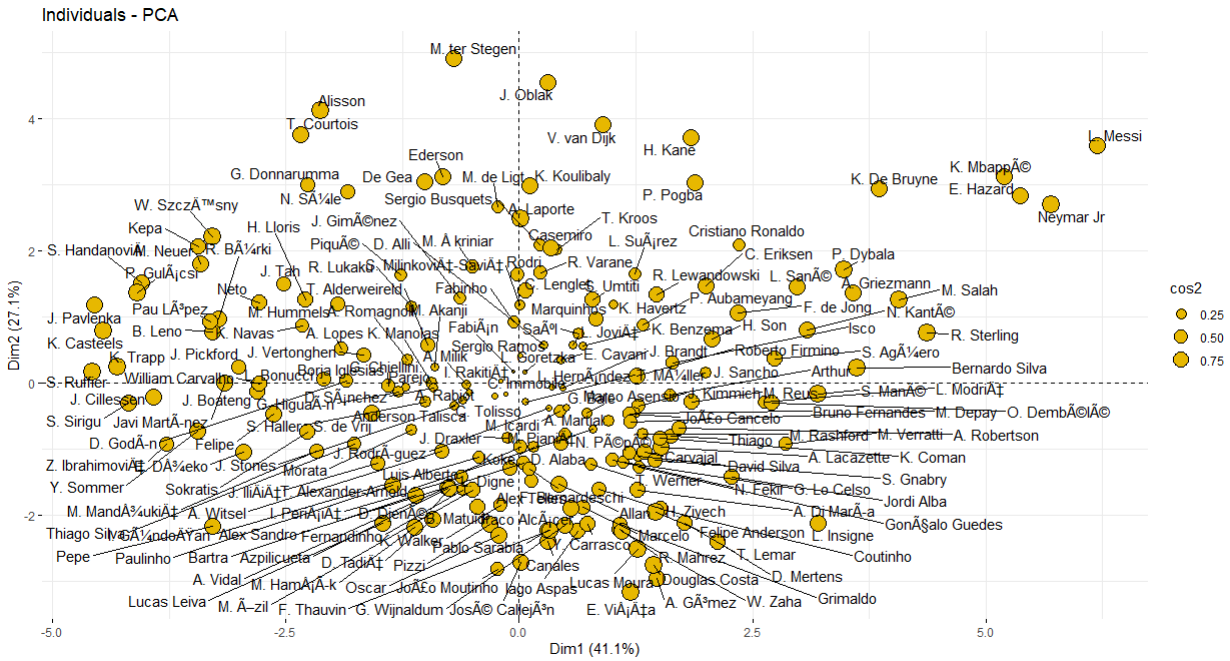
A l'exeption de la variable "age", toutes les variables sont plus ou moins bien représentées sur le cercle de corrélation.
Les variables les plus corrélées par la première composante principale(PC1) :
-mouvement_sprint_speed
-dribling
-mouvement_agility
Ces trois variables sont regroupées entre elles parce qu'elles sont toutes liées à la technique du joueur.
-valeur et release_clause_eu sont bien représentées par PC1, salaire et potentiel un peu moins et elles sont toutes corrélées postivement à PC1. En revanche ce dernier groupe de variables est décorrélée par rapport au groupe de variable (mouvement_agility,dribbling,mouvement_sprint_speed) qui sont regroupées car elles sont toutes liées à l'efficacité du joueur
Ce qui est intéressant ici, c'est d'interpréter la composante principale. Ici, il se trouve que toutes les variables les plus représentées à PC1 ont un "mode" commun, une notion qui les unit et cette notion pourrait être interprétée comme performance du joueur. En effet plus un joueur est performant plus sa valeur augmente par conséquent sa clause libératoire aussi augmente car tous ces paramètres sont très liés.
Les variables tailles et poids sont bien representées par PC1 aussi mais sont toutes les deux décorrélées par rapport aux autres variables et ça se comprend aussi car cs variables n'ont pas de lien avec la performance d'un joueur ou très peu. Si on fait de même pour la deuxième composante principale, les variables potentiel, release_clause et valeur sont les mieux représentées par PC2 et sont corrélées positivement aussi.
Ces variables ont sont toutes liées au prix du joueur Sachant que notre objectif est de regrouper les variables en variables synthétiques, on peut résumer notre jeu de données en deux variables. L'une réprésentant les performance des joueurs et l'autre leur valeur.
Bien entendu on ne peut pas rendre compte de la compléxité de la réalité de nos individus en deux phrases car quand on résume, on perd de l'information forcément.Cependant, ces 2 phrases sont les 2 phrases « optimales », c'est-à-dire que ce sont celles qui résument le mieux l'échantillon.Bien sûr, on peut aller plus loin et rajouter des phrases à notre résumé en analysant les

composantes PC3,PC4 etc. Ce qui nous permettrait de mieux représenter les autres variables qui sont mal représentées dans le premier cercle comme la variable age.

4.0.5 Premier plan factoriel

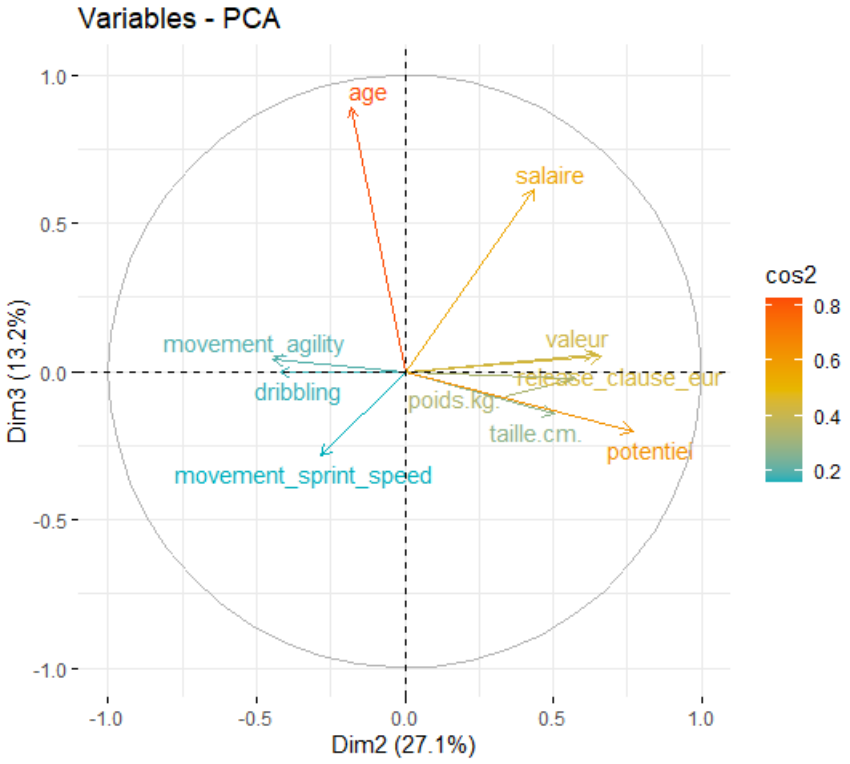


Un point est dit bien représenté sur un axe ou un plan factoriel si il est proche de sa projection sur l'axe ou le plan. S'il est éloigné, on dit qu'il est mal représenté. La proximité dans l'espace entre deux individus bien représentés traduit la ressemblance de ces deux individus du point de vue des valeurs prises par les variables.Bien que le nombre de données ne nous permet pas pas d'avoir une lisibilité clair du graphique, on peut bien distinguer l'individu 1 qui est bien représenté sur PC1 et aussi les individus 4 et 7 à PC2.Certains individus comme l'individus 131 (bien représenté sur PC1) sont dans la direction opposée à l'individu 1 sur l'axe, cela veut dire qu'ils sont très differents.Les observations qui ont les coordonnées sur les axes les plus extrêmes (élevés) sont celles qui ont la contribution la plus élevée. Nous allons essayer de rendre plus lisible notre représentation des individus sur le plan factoriel et montrer la validité de leur représentation dans un plan factoriel cos2.



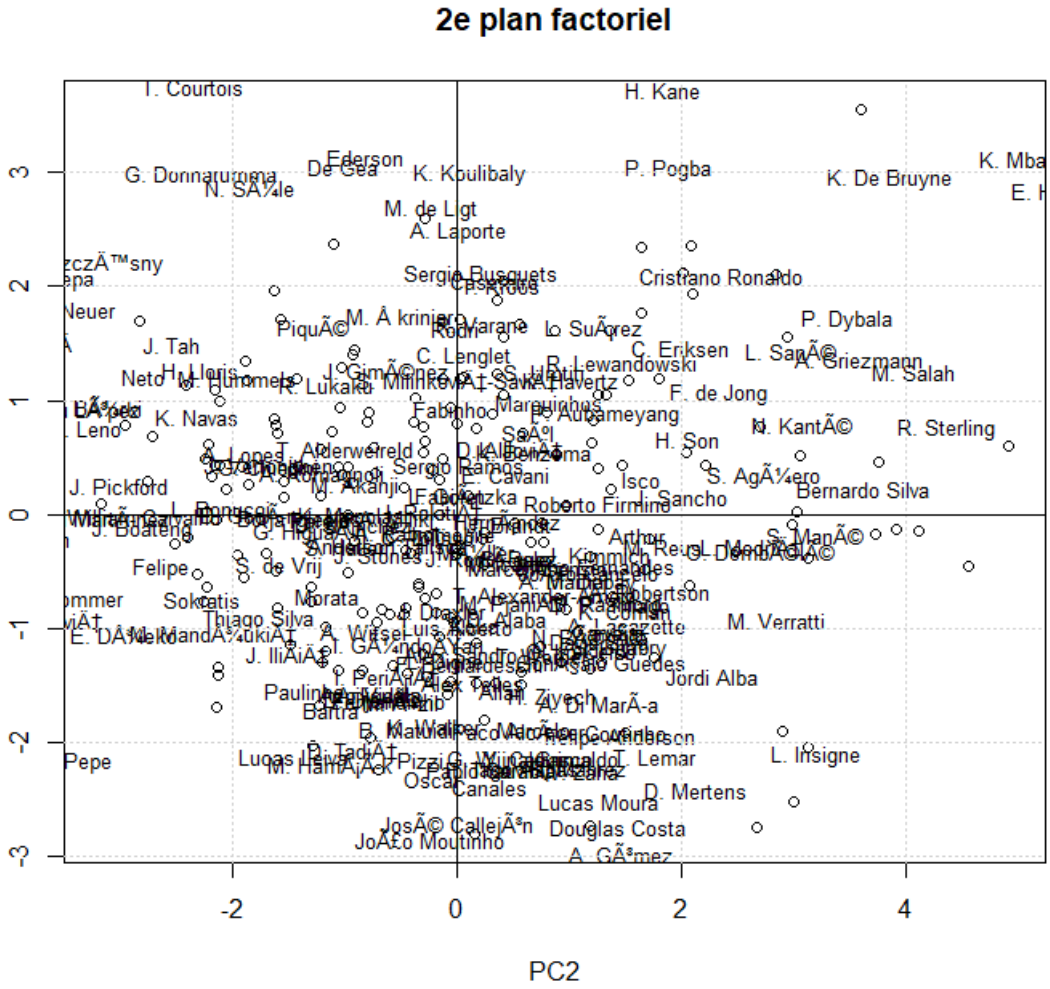
Les valeurs de cos2 sont utilisés pour estimer la qualité de la représentation. Les variables qui sont proches du centre du graphe sont moins importantes pour les premières composantes.Un cos2 élevé indique une bonne représentation de l'individu sur les axes principaux en considération, un faible cos2 indique que la variable n'est pas parfaitement représentée par les axes principaux. Dans ce cas, l'individu est proche du centre du cercle

4.0.6 Deuxième cercle de corrélation



Dans ce plan, seules les variables age(qui était mal représentée dans le premier cercle), salaire et potentiel sont interprétables car sont les mieux représentées par rapport aux deux axes principaux
La deuxième composante principale contribue bien à la représentation des variables potentiel,Valeur et salaire ce qui est vérifié dans le premier cercle de corrélation. On note que PC3 contribue bien à la représentation de la Variable Age .
Ici les variables potentiel et age sont décorrélées ainsi que potentiel et salaire.On peut aussi constater que toutes ces trois variables sont corrélées positivement par rapport aux variables PC2 et PC3 sauf potentiel qui est corrélée négativement par rapport à PC3.

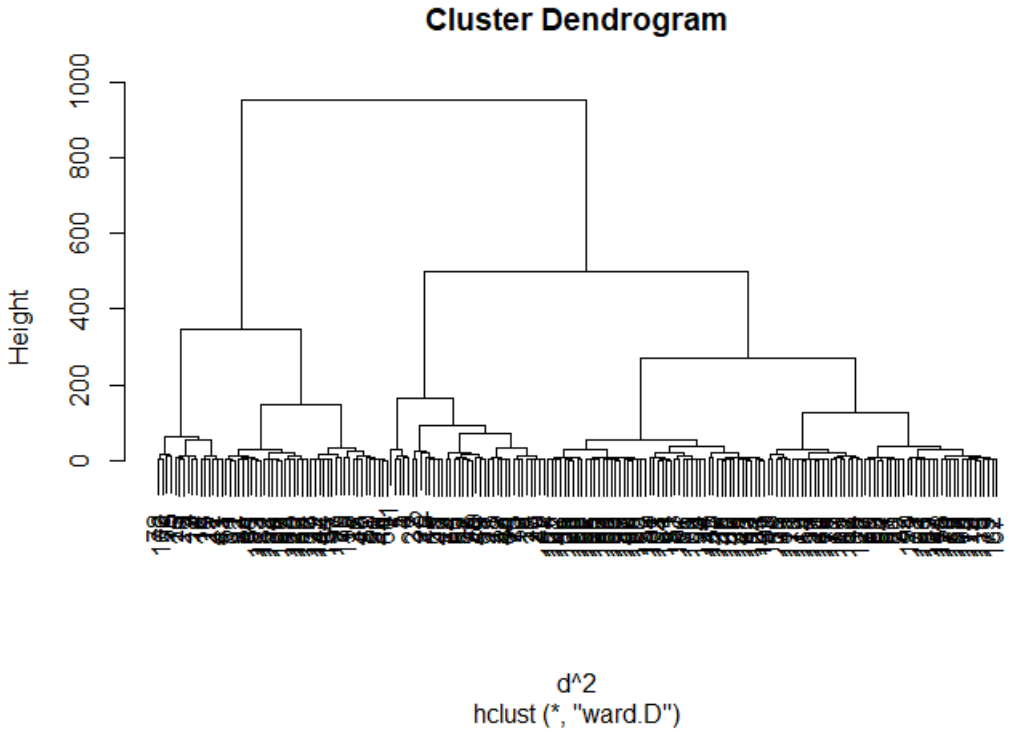
4.0.7 Deuxième plan factoriel



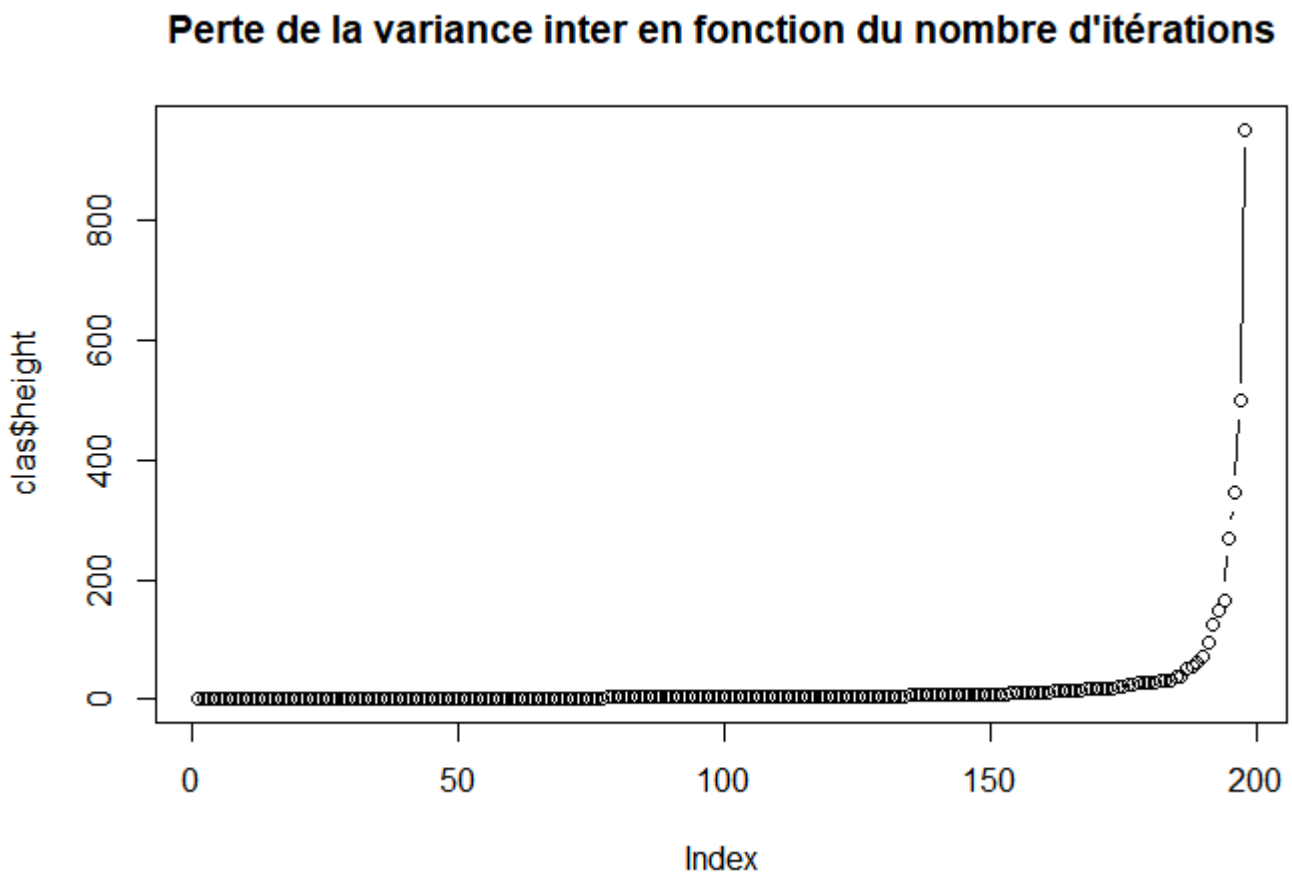
Le deuxième plan ne semble pas se distinguer pourtant du premier parce qu'une majorité est située au milieu et quelques individus comme sur le premier plan sont bien représentés par rapport aux deux axes. D'autres sont très bien représentés sur un seul axe.

5 Classification

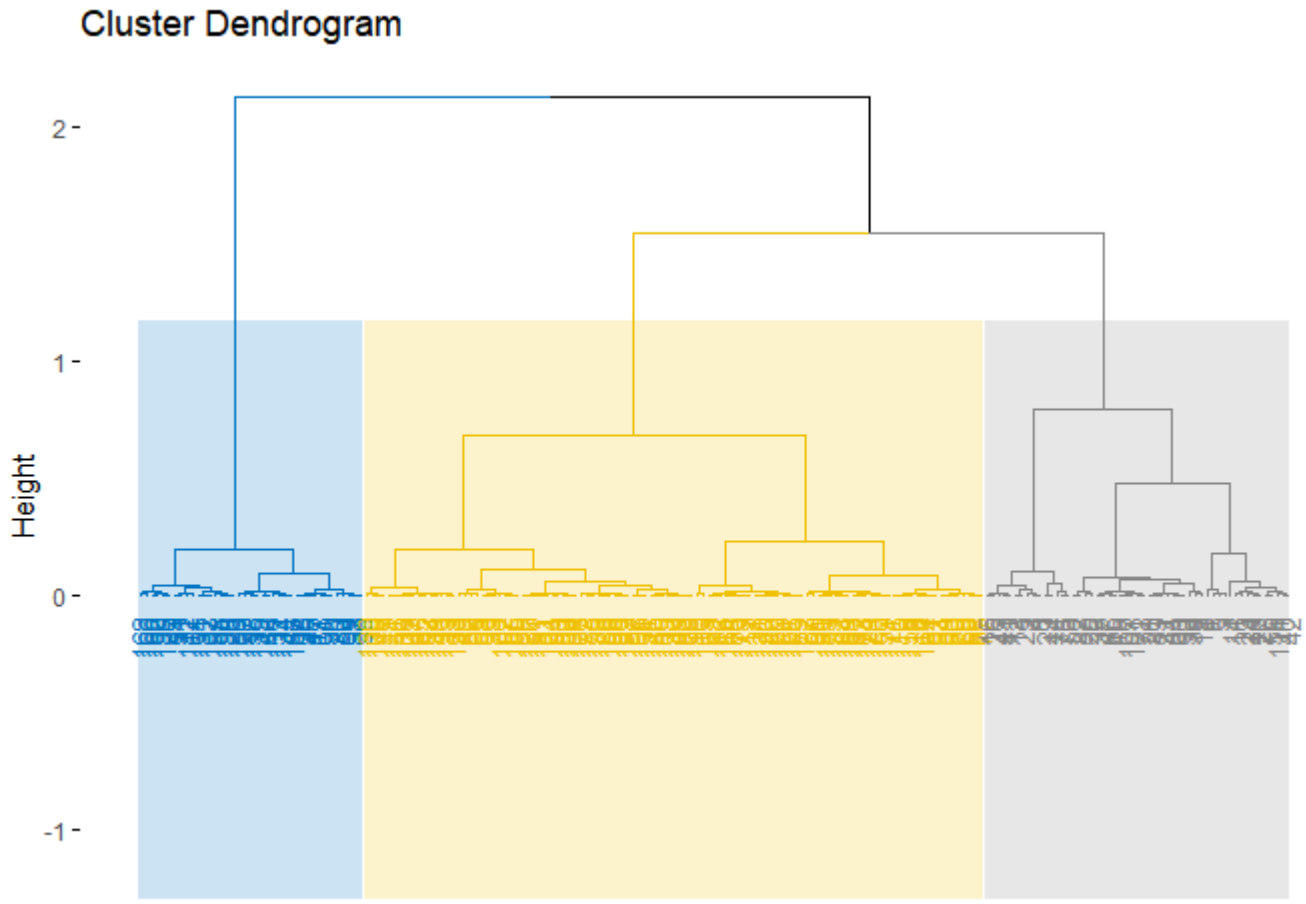
La classification est l'une des approches les plus importantes pour l'exploration des données multivariées. L'objectif est d'identifier des groupes (i.e., clusters) d'objets similaires dans notre jeu de données.On effectue une classification hiérarchique ascendante. On agrège les individus selon la méthode Ward où la distance entre 2 individus correspond à la perte de variance inter-classe qui résulterait de leur fusion. On regroupe donc les individus dont la fusion ferait perdre le minimum de variance interclasse.
On a 199 individus, il y aura maximum 198 étapes.
A l'étape 0 la variance interclasse vaudra 0, à l'étape 198 elle vaudra la variance totale
On utilise la fonction scale() pour centrer et réduire des données, pour éviter que les variables à forte variance pèsent indûment sur les résultats. Le principe de la Classification hierarchique est de rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existant entre chaque individu pris deux à deux.
Deux observations identiques auront une distance nulle.
On utilise la fonction dist() pour calculer la distance.



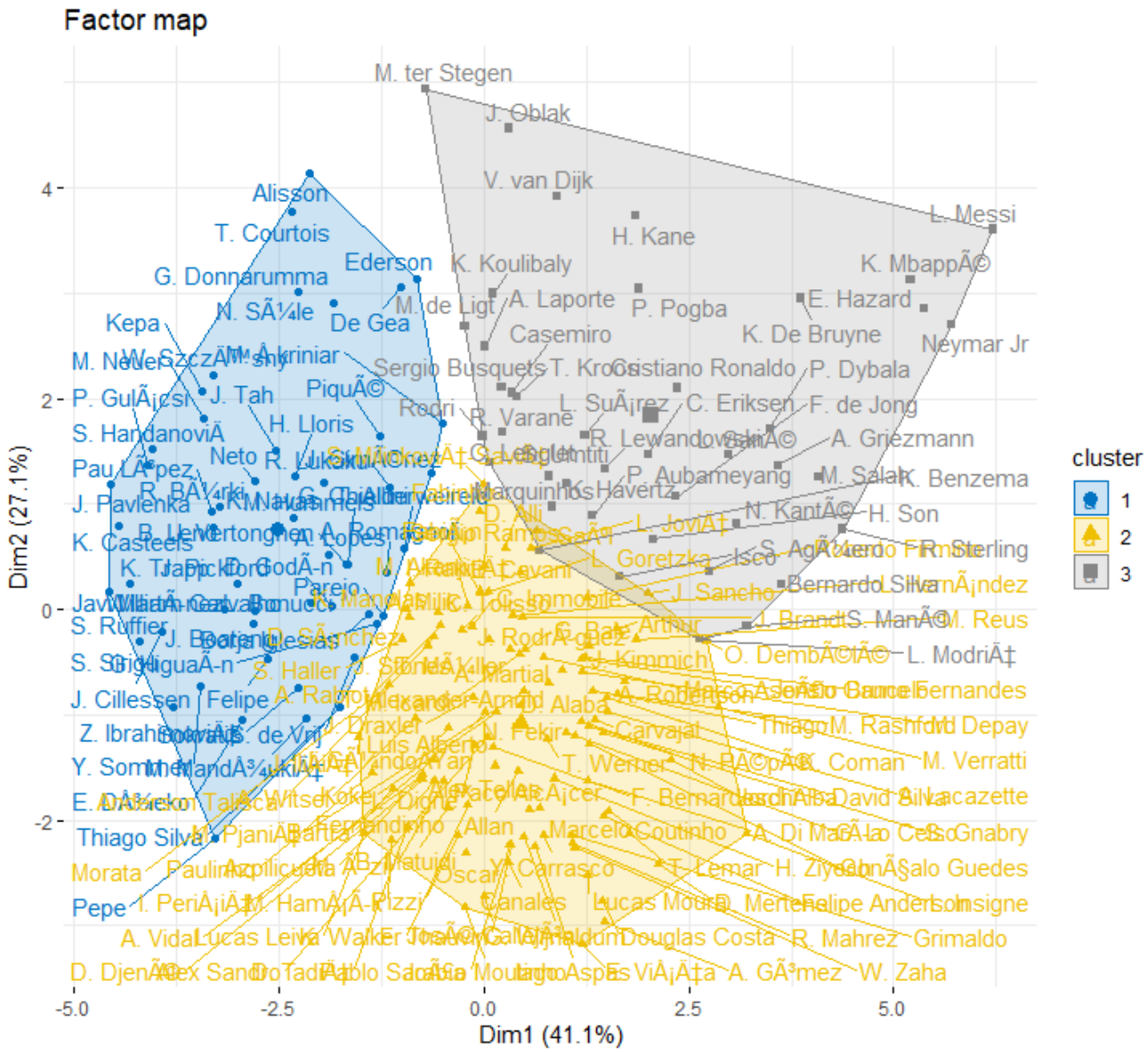
En premier lieu, une analyse de la forme du dendrogramme pourra nous donner une indication sur le nombre de classes à retenir. Dans notre exemple, deux branches bien distinctes apparaissent sur l'arbre
On va essayer de tracer ensuite la courbe correspondant à la perte de variance inter en fonction du nombre d'itérations afin de choisir un nombre de classes.



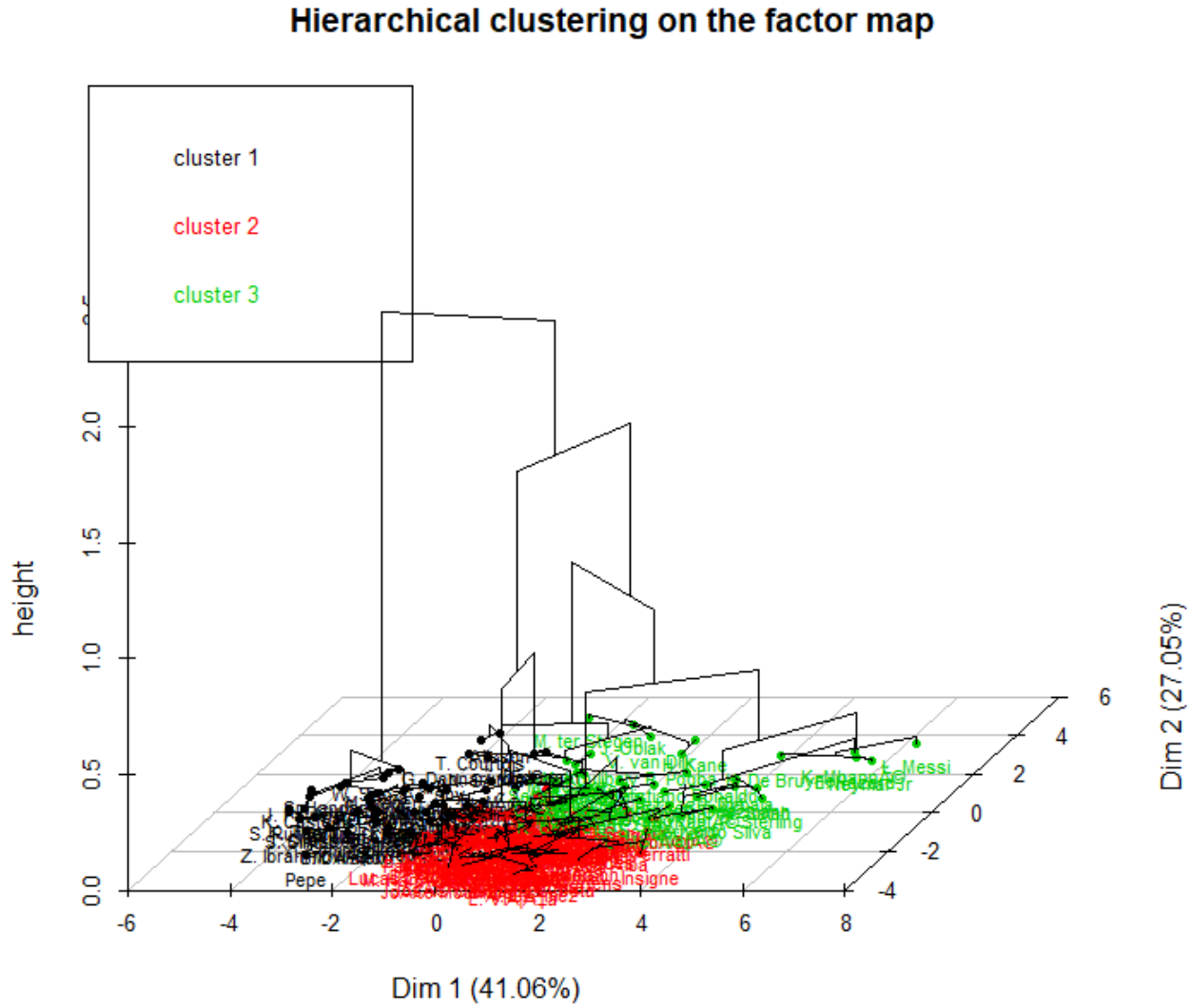
Ce graphique est une méthode qui permet de déterminer le nombre de groupes.Ainsi on peut voir que trois groupes se distinguent après visualisation du grapghe. On va afficher à nouveau le dendrogramme avec la fonction fviz_dend() pour voir si notre choix de trois groupes est logique.



Le dendrogramme suggère une solution à 3 groupes. Nous pouvons aussi visualiser les individus et les colorer par groupes pour une meilleure représentation.



D'après ce graphique les joueurs ne sont pas classés selon leur position sur le terrain donc il pourrait y avoir d'autres critère de ressemblace pour classifier les joueurs. Ce critèe pourrait être le potentiel ou encore le salaire ou d'autres critèe de ressemblance. On peut faire le lien de notre classification avec l'ACP en representant les classes obtenues dans un plan.



Même si le graphique n'est pas très lisible il confirme quand même que la classification ne s'est pas faite sur le critère position car on peut voir Lionel Messi (qui est un attaquant) se retrouver dans la même classe que Oblak (qui est un gardien).

6 Conclusion

Les différentes méthodes d'analyse que nous avons utilisées ont établit des liens entre certaines variables.La variété et la richesse de ces méthodes nous ont permis de faire une analyse efficace de nos variables et voir la dépendance ou non de celles-ci ainsi que leur lien. Ainsi cette étude nous a permis d'avoir des résultats justifiés dans le monde réel.C'est à dire que les attaquants et les milieux sont souvent les plus connus et les mieux payés.Ce qui est un fait dans le monde du foot.Ce qui encore une fois rend crédible nos tests.Cependant aucune véritable réponse n'a pu être amené à cause de l'invalidité de la plupart des test.Le salaire d'un joueur pourrait par exemple avoir un lien avec le club ou campionnat dans lequel il joue et pas forcément sa

position sur le terrain.