

Devoir_6

EL Hadrami N'DOYE

30/12/2020

```
library("FactoMineR")
library("factoextra")
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library("corrplot")
```

```
## corrplot 0.84 loaded
```

```
library("fpc")
```

Application de CAH et K-means sur les données Decathlon

```
# chargement des données decathlon2
data("decathlon2")
# Les variables actives
decathlon2.active <- decathlon2[1:23,1:10]
head(decathlon2.active,4)
```

```
##      X100m Long.jump Shot.put High.jump X400m X110m.hurdle Discus Pole.vault
## SEBRLE 11.04      7.58   14.83     2.07 49.81      14.69 43.75      5.02
## CLAY   10.76      7.40   14.26     1.86 49.37      14.05 50.72      4.92
## BERNARD 11.02      7.23   14.25     1.92 48.93      14.99 40.87      5.32
## YURKOV 11.34      7.09   15.19     2.10 50.42      15.31 46.26      4.72
##      Javeline X1500m
## SEBRLE 63.19 291.7
## CLAY   60.15 301.5
## BERNARD 62.77 280.1
## YURKOV 63.44 276.4
```

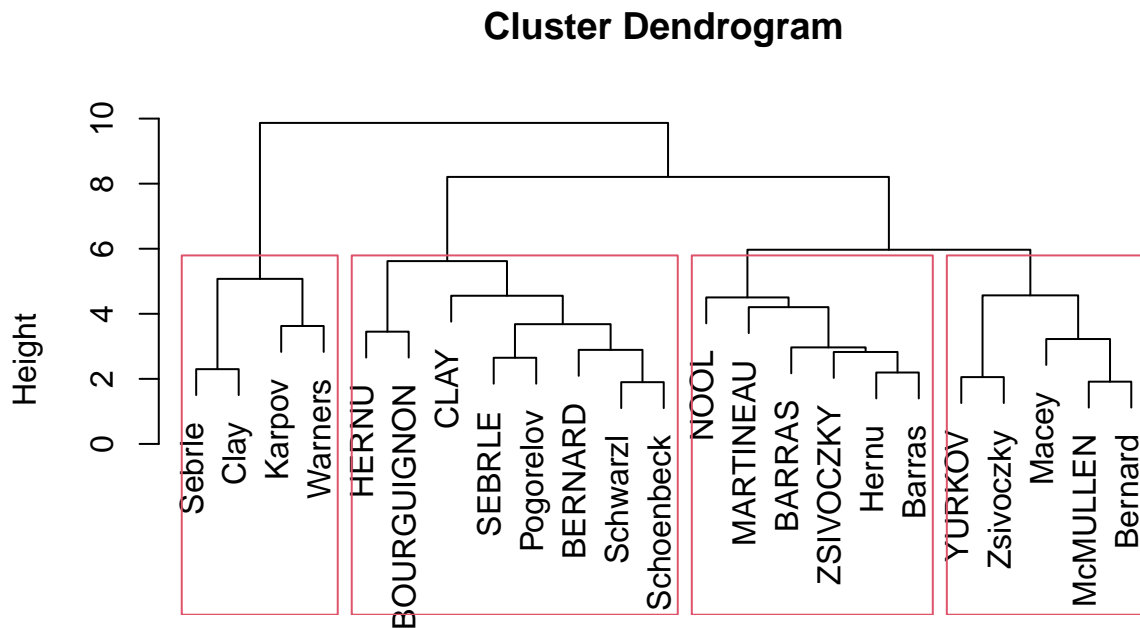
```
summary(decathlon2.active)
```

```
##      X100m      Long.jump      Shot.put      High.jump
## Min.   :10.44 Min.   :6.800 Min.   :12.68 Min.   :1.860
## 1st Qu.:10.84 1st Qu.:7.165 1st Qu.:14.17 1st Qu.:1.940
## Median :10.97 Median :7.310 Median :14.65 Median :2.010
## Mean   :11.00 Mean   :7.350 Mean   :14.62 Mean   :2.007
## 3rd Qu.:11.23 3rd Qu.:7.525 3rd Qu.:15.14 3rd Qu.:2.095
## Max.   :11.64 Max.   :7.960 Max.   :16.36 Max.   :2.150
##      X400m      X110m.hurdle      Discus      Pole.vault
## Min.   :46.81 Min.   :13.97 Min.   :37.92 Min.   :4.400
## 1st Qu.:48.95 1st Qu.:14.17 1st Qu.:43.74 1st Qu.:4.610
## Median :49.40 Median :14.37 Median :44.75 Median :4.820
## Mean   :49.43 Mean   :14.53 Mean   :45.16 Mean   :4.797
```

```
## 3rd Qu.:50.02 3rd Qu.:14.94 3rd Qu.:46.93 3rd Qu.:5.000
## Max. :51.16 Max. :15.67 Max. :51.65 Max. :5.320
## Javeline X1500m
## Min. :52.33 Min. :262.1
## 1st Qu.:55.40 1st Qu.:268.8
## Median :57.44 Median :278.1
## Mean :59.11 Mean :277.9
## 3rd Qu.:62.98 3rd Qu.:283.6
## Max. :70.52 Max. :301.5
```

Realisation d'une CAH

```
decathlon2.active.cr <- scale(decathlon2.active,center=T,scale=T)
dist.dec2 <- dist(decathlon2.active.cr)
clust.dc2 <- hclust(dist.dec2,method = "ward.D2")
plot(clust.dc2)
rect.hclust(clust.dc2,k=4)
```



dist.dec2
hclust (*, "ward.D2")

```
print(sort(cutree(clust.dc2,k=4)))
```

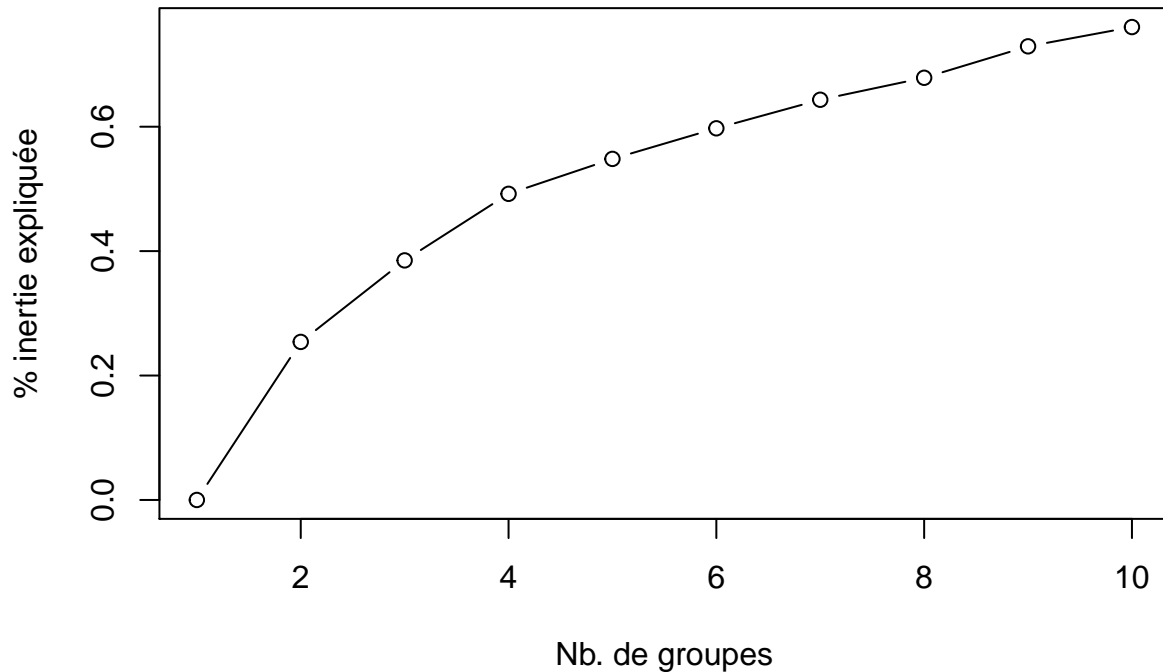
```
## SEBRLE CLAY BERNARD HERNU BOURGUIGNON Schwarzl
## 1 1 1 1 1 1
## Pogorelov Schoenbeck YURKOV McMULLEN Macey Zsivoczky
## 1 1 2 2 2 2
## Bernard ZSIVOCZKY MARTINEAU BARRAS NOOL Hernu
## 2 3 3 3 3 3
## Barras Sebrle Clay Karpov Warners
## 3 4 4 4 4
```

Le dendrogramme suggere un decoupage en 4 groupes, nous verrons ensuite la methode de K-means pour

trouver le nombre de découpages (k) optimal afin de confirmer ou rejeter le nombre de découpages trouvés sur la méthode de classification.

La méthode de K-means

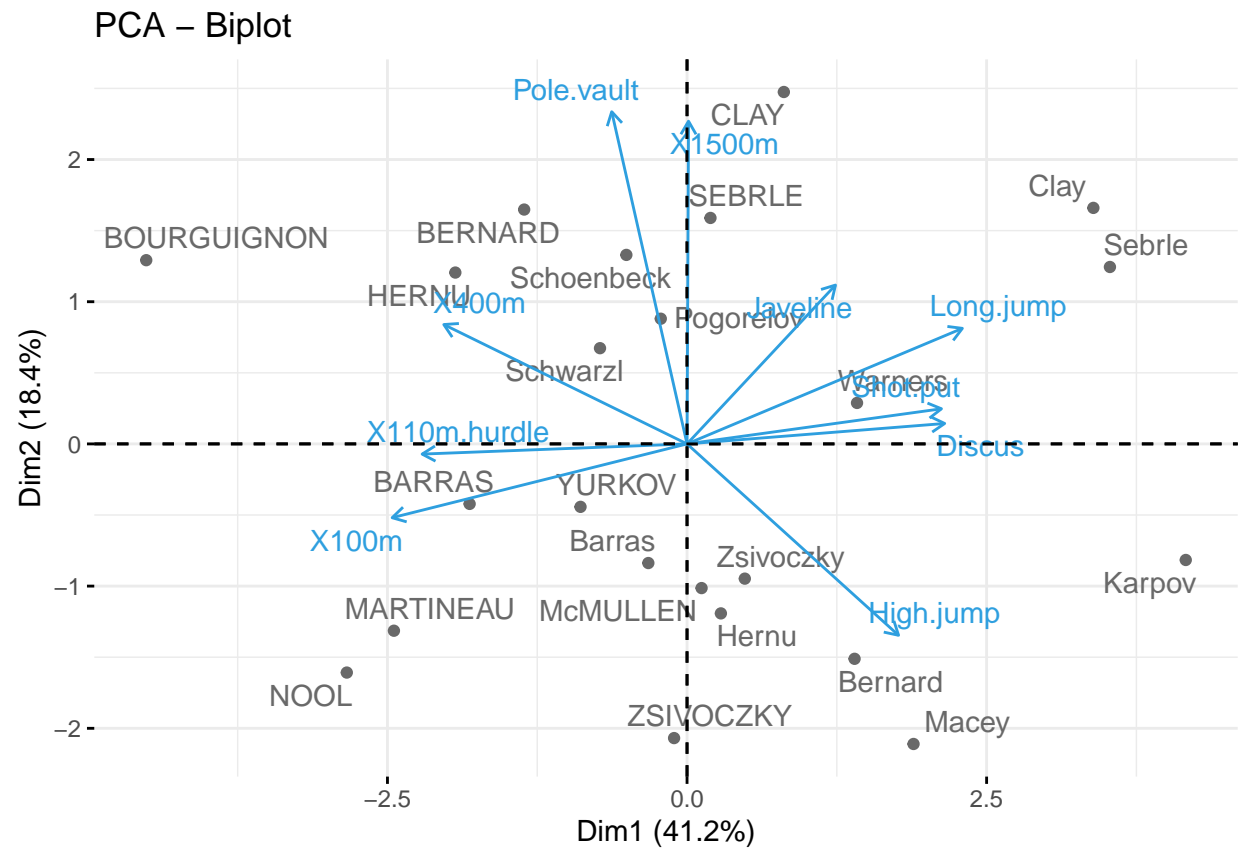
```
groupes.kmeans <- kmeans(decathlon2.active.cr,centers=4,nstart=5)
inertie <- rep(0,times=10)
for (k in 2:10){
  group <- kmeans(decathlon2.active.cr,centers = k,nstart=5)
  inertie[k] <- group$betweenss/group$totss
}
plot(1:10,inertie,type="b",xlab="Nb. de groupes",ylab="% inertie expliquée")
```



Le graphe ci - dessus montre l'évolution de la proportion d'inertie, on constate qu'à partir de $k = 4$ classes, l'adjonction d'un groupe supplémentaire n'augmente pas «significativement» la part d'inertie expliquée par la partition.

Construction d'un diagramme d'individus et des variables

```
res.da <- PCA(decathlon2.active,scale.unit = TRUE,graph = FALSE)
fviz_pca_biplot(res.da,repel = TRUE,col.var = "#2E9FDF", col.ind = "#696969")
```



Le biplot confirme aussi que les individus peuvent être regroupés pour former quatre classes.