Le travail à faire réalisé par *BAHRI Khalid*

```
library(learnr)
library(tidyverse)
library(stats)
#library(FactoMineR)
library(factoextra)
#library(corrplot)
tutorial_options(exercise.timelimit = 60)
knitr::opts_chunk$set(error = TRUE)
```

Introduction

Lecture des données

```
library(stats)
library(factoextra)
```

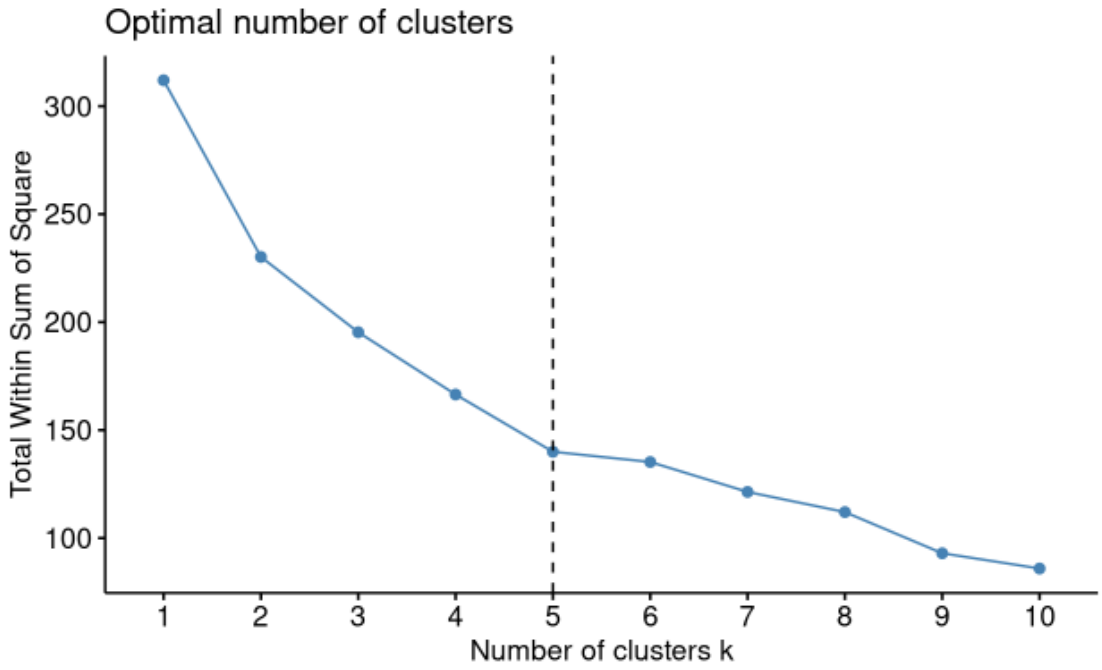
```
data(decathlon2[-13])
str(decathlon2[-13])
head(decathlon2[-13])
```

```
tail(decathlon2[-13])

colnames(decathlon2[-13])
```

K-means

Estimation du nombre optimal de clusters

```
library(factoextra)

data("decathlon2") # Loading the data set

df <-  scale(decathlon2[-13]) # Scaling the data


fviz_nbclust(df, kmeans, method  =  "wss") +

geom_vline(xintercept  =  5, linetype  =  2)
```



__Calcul du clustering k-means __

```
set.seed(123)

df <-  scale(decathlon2[-13])

km.res <-  kmeans(df, 5, nstart  =  25)

print(km.res)
```

```
K-means clustering with 5 clusters of sizes 3, 4, 5, 8, 7

Cluster means:
        X100m  Long.jump   Shot.put  High.jump       X400m
1 -1.40429576  1.7161275  1.5502646  0.9608257 -1.21617001
2 -0.17652313  0.1110324  0.2351301  0.0978260  0.28291397
3  1.46898803 -0.9607294 -0.8384979 -0.8174767  0.92859451
4 -0.01676477 -0.4027678  0.2216799  0.8039166  0.01558585
5 -0.32740603  0.3476111 -0.4531805 -0.8025330 -0.32154359
  X110m.hurdle     Discus   Pole.vault   Javeline      X1500m
1  -0.94999079  1.5464099 -0.008528911  1.3264499  0.14897127
2  -0.50335333  0.6279123  0.496809090 -0.2949200  1.64003581
3   1.25342068 -0.6486684 -0.059702380 -0.5562381 -0.09990692
4   0.04962638  0.2116043 -1.067180040  0.1942150 -0.68523610
5  -0.25724697 -0.8000531  0.982043228 -0.2245998 -0.14651909
         Rank      Points
1 -1.23822680  2.02600474
2 -0.72651063  0.03369463
3  0.80863791 -1.39475228
4 -0.01579371  0.11437516
5  0.38626900 -0.02200466

Clustering vector:
     SEBRLE        CLAY     BERNARD      YURKOV   ZSIVOCZKY
          2           2           5           4           4
   McMULLEN   MARTINEAU       HERNU      BARRAS        NOOL
          4           3           3           3           3
```

```
set.seed(123)
```

```
df <-  scale(decathlon2[-13])
```

```
km.res <-  kmeans(df, 5, nstart  =  25)
```

```
aggregate(decathlon2[-13], by=list(cluster=km.res$cluster), mean)
```

```
aggregate(decathlon2[-13], by=list(cluster=km.res$cluster), sd)
```

```
decathlon2_C=cbind(decathlon2[-13], cluster=as.factor(km.res$cluster))
```

```
ggplot(decathlon2_C, aes(y=X400m, fill=cluster)) + geom_boxplot()
```

```
ggplot(decathlon2_C, aes(y=Shot.put, fill=cluster)) + geom_boxplot()
```

```
for(i in  c(1:5)){
```

```
var=colnames(decathlon2_C)[i]
```

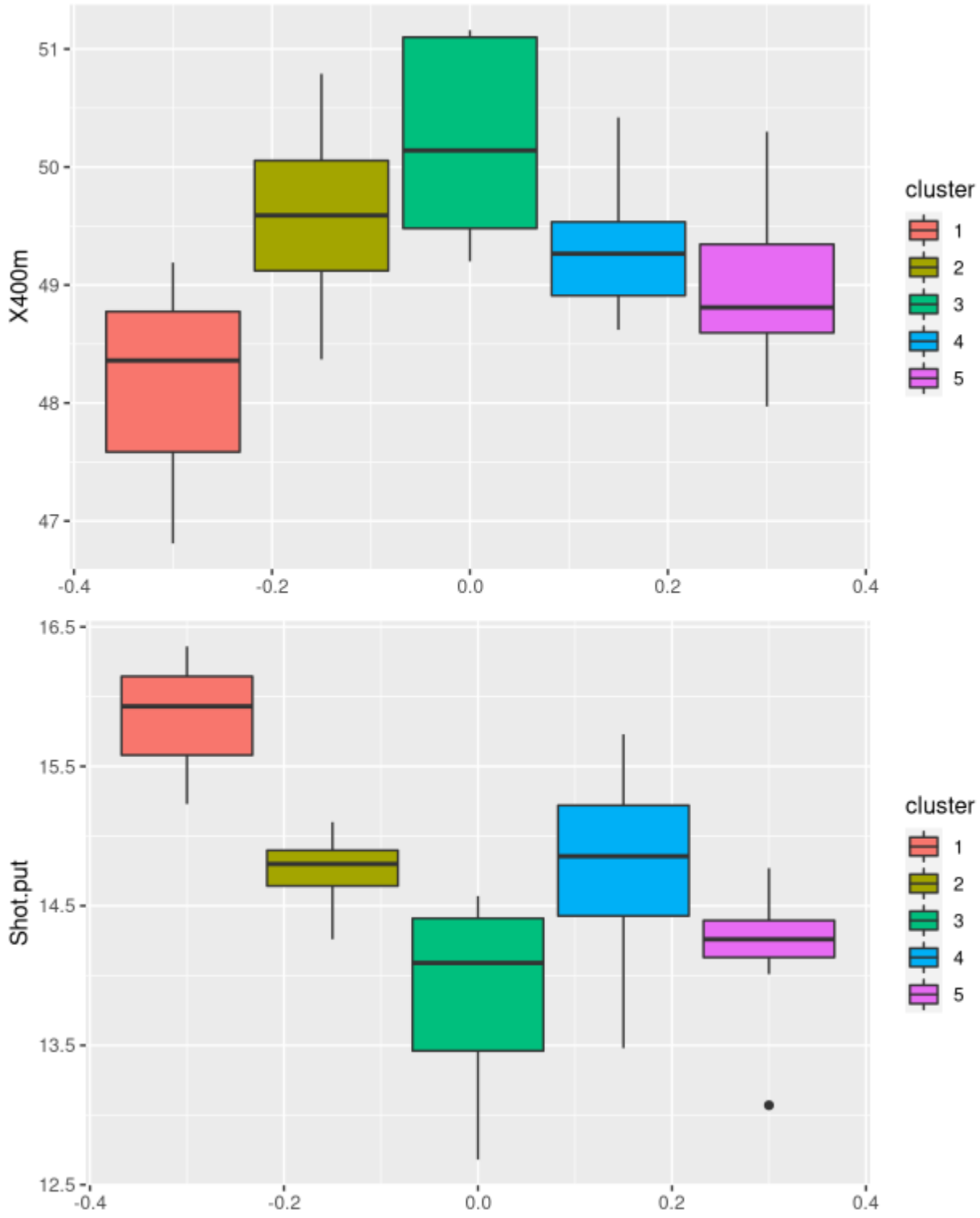
```
print(ggplot(decathlon2_C, aes(y=decathlon2_C[[i]], fill=cluster)) +

geom_boxplot()+ ylab(var))

}
```

cluster <int>	X100m <dbl>	Long.jump <dbl>	Shot.put <dbl>	High.jump <dbl>	X400m <dbl>	X110m.hurdle <dbl>	Discus <dbl>	Pole.vault <dbl>
1	10.59667	7.870000	15.84000	2.090000	48.12000	14.05000	50.160	4.833333
2	10.94250	7.397500	14.74000	2.007500	49.58500	14.26000	47.005	4.965000
3	11.40600	7.082000	13.84200	1.920000	50.21600	15.08600	42.620	4.820000
4	10.98750	7.246250	14.72875	2.075000	49.32375	14.52000	45.575	4.557500
5	10.90000	7.467143	14.16429	1.921429	48.99429	14.37571	42.100	5.091429

5 rows | 1-9 of 13 columns

cluster <int>	X100m <dbl>	Long.jump <dbl>	Shot.put <dbl>	High.jump <dbl>	X400m <dbl>	X110m.hurdle <dbl>	Discus <dbl>
1	0.2214347	0.07937254	0.5703508	0.03000000	1.2080149	0.0800000	1.465640
2	0.1276388	0.12971122	0.3507136	0.09912114	1.0041746	0.2946184	3.364684
3	0.1320227	0.32782617	0.7762538	0.05612486	0.9017095	0.4401477	3.785320
4	0.2059646	0.17573824	0.7648611	0.07348469	0.6058274	0.4089359	1.278917
5	0.1603122	0.17717358	0.5362791	0.04336995	0.7865294	0.3792913	1.558707

5 rows | 1-8 of 13 columns



Accès aux résultats de la fonction `kmeans()`

```
km.res$cluster
```

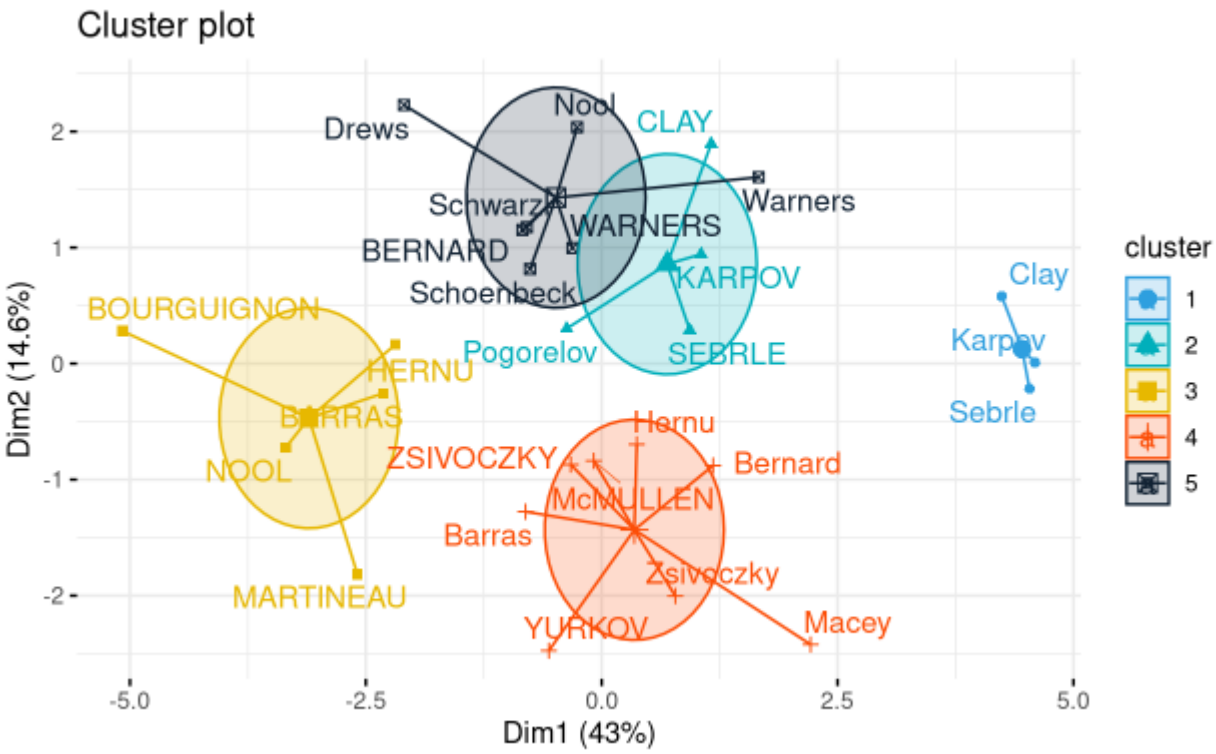
```
km.res$centers
```

```
     SEBRLE        CLAY     BERNARD      YURKOV   ZSIVOCZKY 
          2           2           5           4           4 
   McMULLEN   MARTINEAU       HERNU      BARRAS        NOOL 
          4           3           3           3           3 
BOURGUIGNON      Sebrle        Clay      Karpov       Macey 
          3           1           1           1           4 
    Warners   Zsivoczky       Hernu     Bernard    Schwarzl 
          5           4           4           4           5 
  Pogorelov  Schoenbeck      Barras      KARPOV     WARNERS 
          2           5           4           2           5 
       Nool       Drews 
          5           5 
```

```
        X100m  Long.jump   Shot.put  High.jump       X400m
1 -1.40429576  1.7161275  1.5502646  0.9608257 -1.21617001
2 -0.17652313  0.1110324  0.2351301  0.0978260  0.28291397
3  1.46898803 -0.9607294 -0.8384979 -0.8174767  0.92859451
4 -0.01676477 -0.4027678  0.2216799  0.8039166  0.01558585
5 -0.32740603  0.3476111 -0.4531805 -0.8025330 -0.32154359
  X110m.hurdle     Discus   Pole.vault   Javeline      X1500m
1  -0.94999079  1.5464099 -0.008528911  1.3264499  0.14897127
2  -0.50335333  0.6279123  0.496809090 -0.2949200  1.64003581
3   1.25342068 -0.6486684 -0.059702380 -0.5562381 -0.09990692
4   0.04962638  0.2116043 -1.067180040  0.1942150 -0.68523610
5  -0.25724697 -0.8000531  0.982043228 -0.2245998 -0.14651909
         Rank      Points
1 -1.23822680  2.02600474
2 -0.72651063  0.03369463
3  0.80863791 -1.39475228
4 -0.01579371  0.11437516
5  0.38626900 -0.02200466
```

Visualisation des clusters produits par `kmeans()`

```
fviz_cluster(km.res, data  = df,

palette  =  c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07","#112233"),

ellipse.type  =  "euclid", # Concentration ellipse

star.plot  =  TRUE, # Add segments from centroids to items

repel  =  TRUE, # Avoid label overplotting (slow)

ggtheme  = theme_minimal()

)
```



Clustering hiérarchique

```
set.seed(123)
```

```
df <-  scale(decathlon2[-13])
```

```
km.res <-  kmeans(df, 5, nstart  =  25)
```

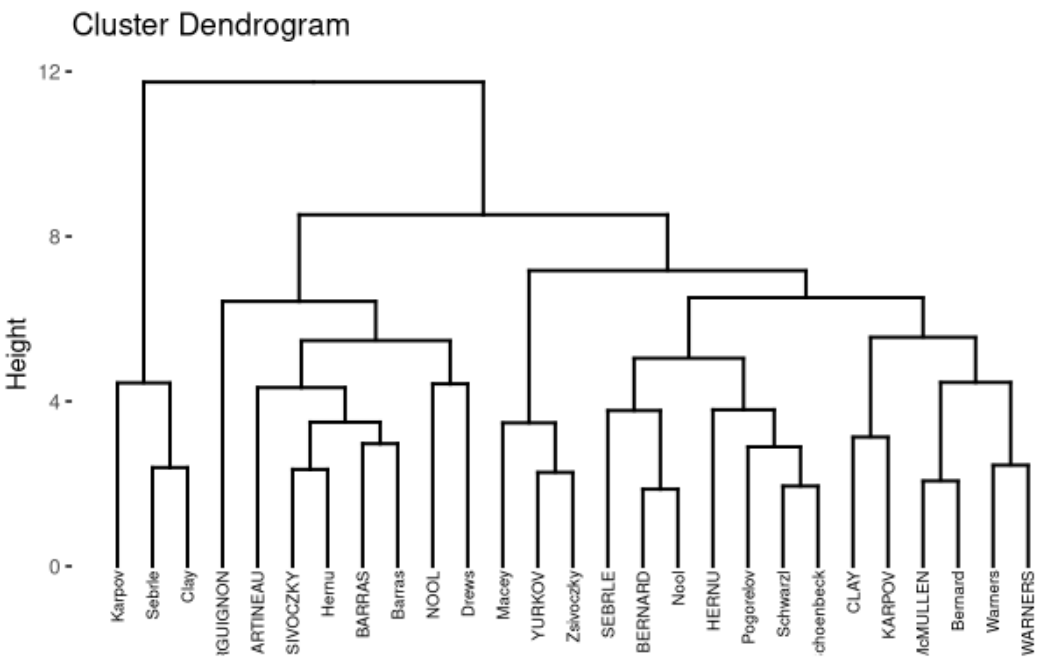
```
res.dist <-  dist(df, method  =  "euclidean")
```

```
res.hc <-  hclust(d  = res.dist, method  =  "ward.D2")
```

```
grp <-  cutree(res.hc, k  =  5)
```

Dendrogramme

```
fviz_dend(res.hc, cex  =  0.5)
```



Vérification de l'arborescence du clustering

```
res.coph <-  cophenetic(res.hc)

cor(res.dist, res.coph)
```

```
[1] 0.6123268
```

```
res.hc2 <-  hclust(res.dist, method  =  "average")

cor(res.dist, cophenetic(res.hc2))
```

```
[1] 0.7250826
```

découpage du dendrogramme en clusters

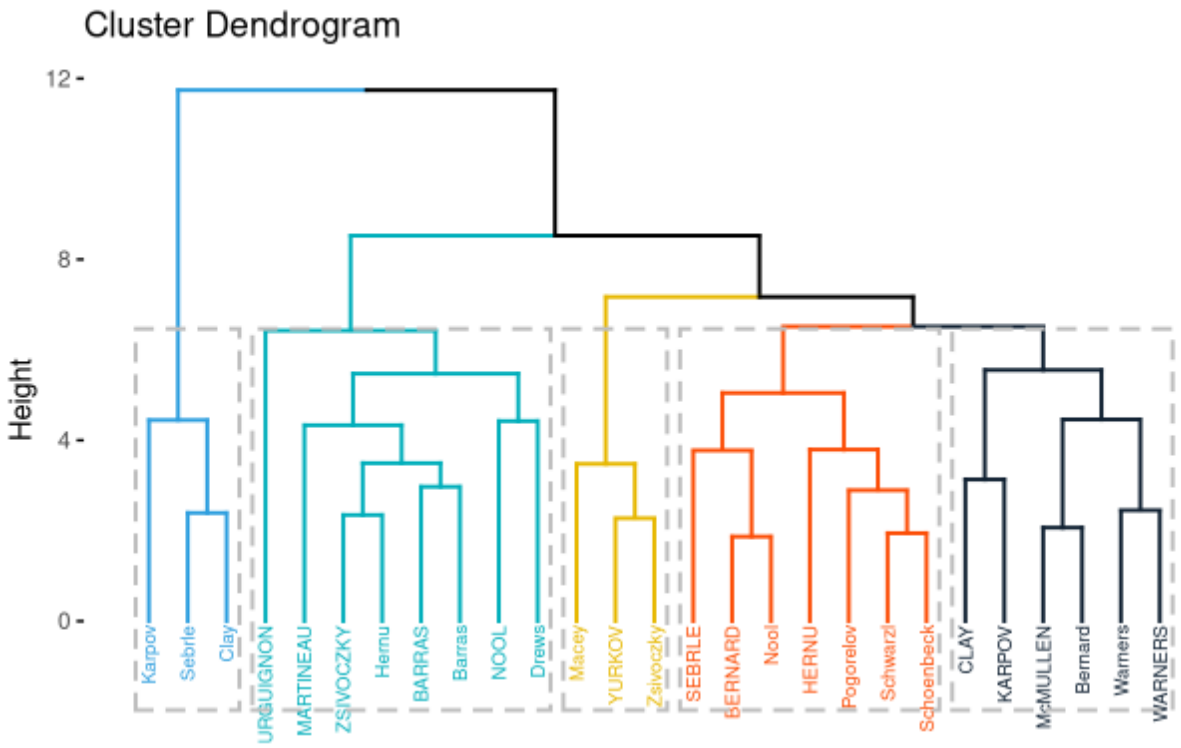
```
grp <-  cutree(res.hc, k  =  5)

head(grp, n  =  10)

table(grp)

rownames(df)[grp ==  1]
```

```
   SEBRLE      CLAY   BERNARD    YURKOV ZSIVOCZKY  McMULLEN MARTINEAU
       1         2         1         3         4         2         4
    HERNU    BARRAS      NOOL
       1         4         4
```

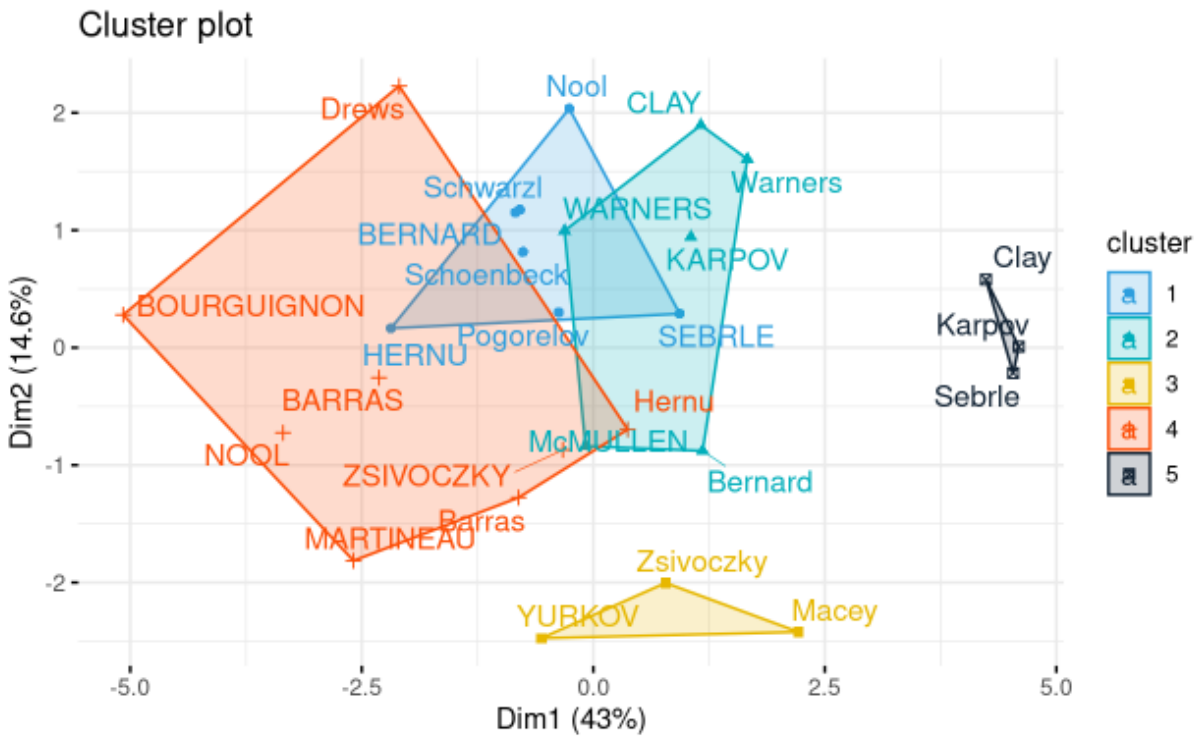
```
grp
1 2 3 4 5
7 6 3 8 3
```

```
[1] "SEBRLE"     "BERNARD"    "HERNU"      "Schwarzl"   "Pogorelov"
[6] "Schoenbeck" "Nool"
```

```
fviz_dend(res.hc, k  =  5,

cex  =  0.5,

k_colors  =  c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07","#112233"),

color_labels_by_k  =  TRUE,

rect  =  TRUE

)
```




```
fviz_cluster(list(data  = df, cluster  = grp),

palette  =  c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07","#112233"),

ellipse.type  =  "convex",

repel  =  TRUE,

show.clust.cent  =  FALSE, ggtheme  = theme_minimal())
```



Hierarchical K-Means Clustering

```
res.hk <-hkmeans(df, 5)

names(res.hk)

res.hk
```

Visualisation des résultats de `hkmeans`

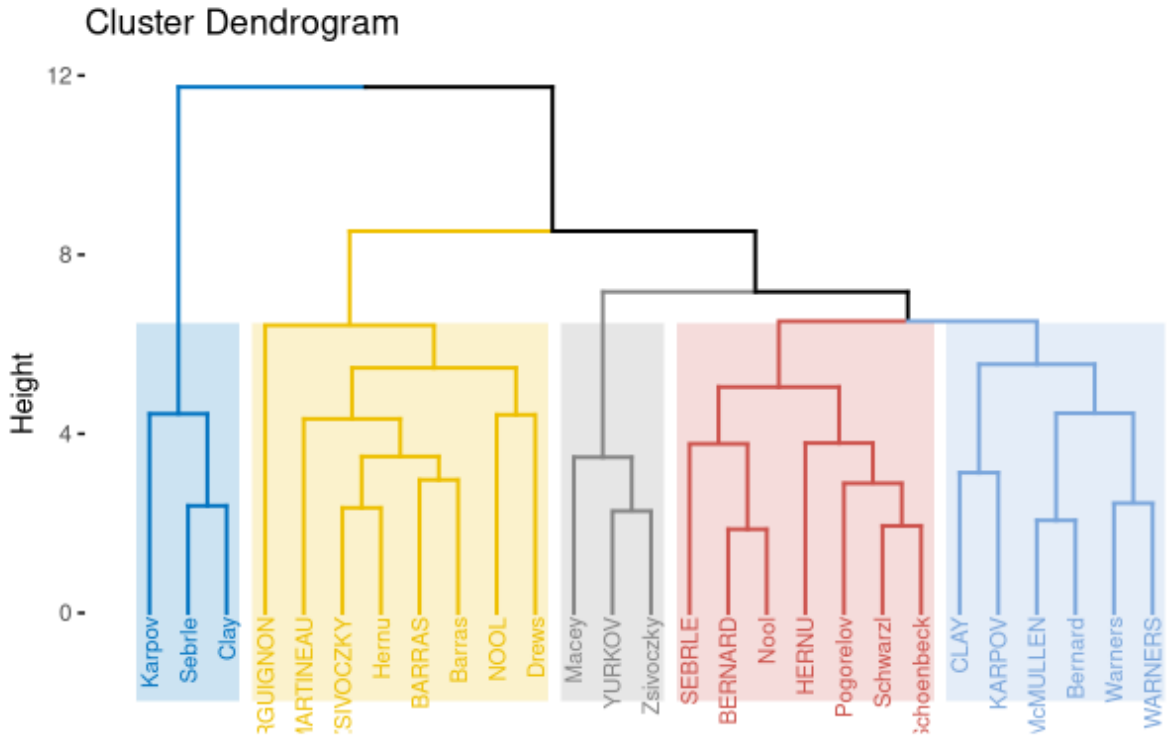
```
set.seed(123)

df <-  scale(decathlon2[-13])

km.res <-  kmeans(df, 5, nstart  =  25)

res.dist <-  dist(df, method  =  "euclidean")
```

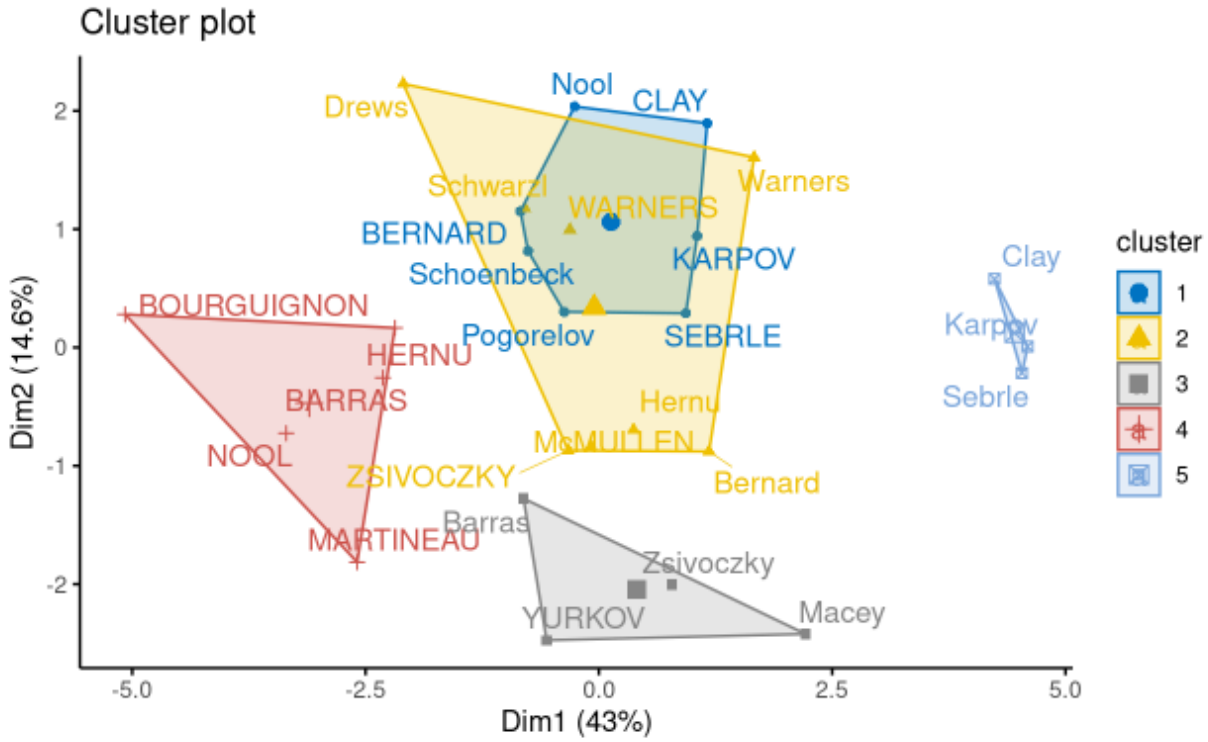
```
res.hk <-hkmeans(df, 5)
```

```
 [1] "cluster"      "centers"      "totss"        "withinss"
 [5] "tot.withinss" "betweenss"    "size"         "iter"
 [9] "ifault"       "data"         "hclust"
```

```
3 -0.1010797  0.23959426
4  0.8086379 -1.39475228
5 -1.2382268  2.02600474

Clustering vector:
     SEBRLE        CLAY     BERNARD      YURKOV   ZSIVOCZKY
          1           1           1           3           2
   McMULLEN   MARTINEAU       HERNU      BARRAS        NOOL
          2           4           4           4           4
BOURGUIGNON      Sebrle        Clay      Karpov       Macey
          4           5           5           5           3
    Warners   Zsivoczky       Hernu     Bernard    Schwarzl
          2           3           2           2           2
  Pogorelov  Schoenbeck      Barras      KARPOV     WARNERS
          1           1           3           1           2
       Nool       Drews
          1           2

Within cluster sum of squares by cluster:
[1] 45.05161 41.45161 16.59780 32.40713 12.74734
 (between_SS / total_SS =  52.5 %)

Available components:

 [1] "cluster"      "centers"      "totss"        "withinss"
 [5] "tot.withinss" "betweenss"    "size"         "iter"
 [9] "ifault"       "data"         "hclust"
```

```
fviz_dend(res.hk, cex  =  0.6, palette  =  "jco",rect  =  TRUE,

rect_border  =  "jco", rect_fill  =  TRUE)

fviz_cluster(res.hk, palette  =  "jco", repel  =  TRUE,

ggtheme  = theme_classic())
```


Cluster Dendrogram
Height
12
8
4
0
Karpov
Sebrle
Clay
BOURGUIGNON
MARTINEAU
ZSIVOCZKY
Hernu
BARRAS
Barras
NOOL
Drews
Macey
YURKOV
Zsivoczky
SEBRLE
BERNARD
Nool
HERNU
Pogorelov
Schwarzl
Schoenbeck
CLAY
KARPOV
McMULLEN
Bernard
Warners
WARNERS


Cluster plot
Dim2 (14.6%)
Dim1 (43%)
2
1
0
-1
-2
-5.0
-2.5
0.0
2.5
5.0
Drews
Nool
CLAY
Warners
Schwarzl
WARNERS
BERNARD
KARPOV
Schoenbeck
BOURGUIGNON
Pogorelov
SEBRLE
HERNU
BARRAS
Hernu
NOOL
McMULLEN
ZSIVOCZKY
Bernard
Barras
MARTINEAU
Zsivoczky
YURKOV
Macey
Clay
Karpov
Sebrle
cluster
1
2
3
4
5