

Analyse des Comportements et des Préférences sur LES ACTIVITES SPORTIVES

Yahia Chammami

2023-05-02

Introduction:

Ce projet a pour but d'étudier les différents avis des tunisiens sur les Activités Sportives .

Le questionnaire:

1.La période d'enquête : Mars - Avril 2022

2.L'enquête : L'enquête a été crée par google Forms et distribuée aux repondeurs à travers Facebook , les données ont été après collectées dans un fichier excel et traitées par la suite par le langage R.

3.Base de données : La base de données obtenue est une base de 45 lignes et 53 colonnes.

Importation des données:

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
##      filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      intersect, setdiff, setequal, union
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.2.2
```

```
Sport<-read.csv("C:\\Users\\Lenovo\\Desktop\\LES_ACTIVITES_SPORTIVES.csv")
```

```
sporttt<-read_xlsx("C:/Users/Lenovo/Desktop/LES_ACTIVITES_SPORTIVES.xlsx")
```

La base de données DE SPORTs est construite à partir d'un questionnaire destiné à des individus ayant des différentes catégories socio-professionnelles et plusieurs classes d'âge.

Les données contiennent principalement les domaines de preference sur les activités sportives dans un premier lieu et les critères sur lesquels s'appuient les utilisateurs en faisant leurs choix.

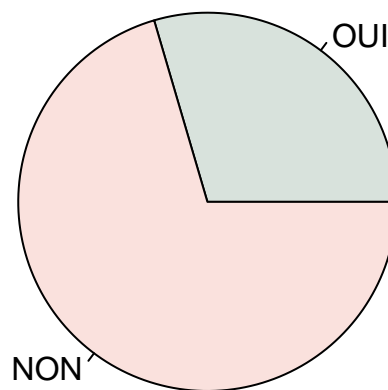
Statistiques descriptive:

1. Répartition selon les Reponses

```
library(dplyr)
rp <- Sport$Pratiquez.vous.une.ou.des.activités.sportives...
OUI = sum(Sport$rp == "Oui")
NON = sum(Sport$rp == "Non")

lbls2 <- c("OUI", "NON")
df1 <- data.frame(rp) %>% group_by(rp) %>% count(rp)
pie(table(rp), labels = lbls2, col=c("#d8e2dc", "#fae1dd"), main="Répartition selon les Reponses")
```

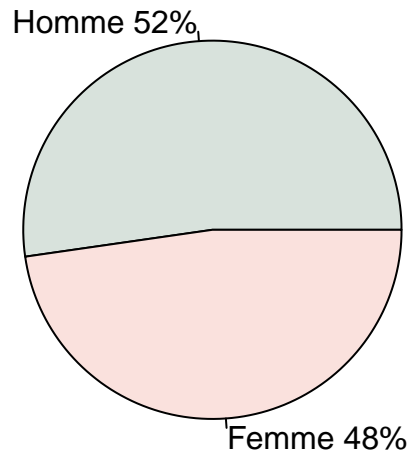
Répartition selon les Reponses



2. Echantillon selon le sexe:

```
sexe <- Sport$Vous.êtes..
freq1=table(sexe)
pct1 <- round(freq1/sum(freq1)*100)
lbls1 <- c("Homme", "Femme")
lbls1 <- paste(lbls1, pct1)
lbls1 <- paste(lbls1, "%", sep="")
pie(freq1, labels = lbls1, col=c("#d8e2dc", "#fae1dd"), main="Répartition selon le Sexe")
```

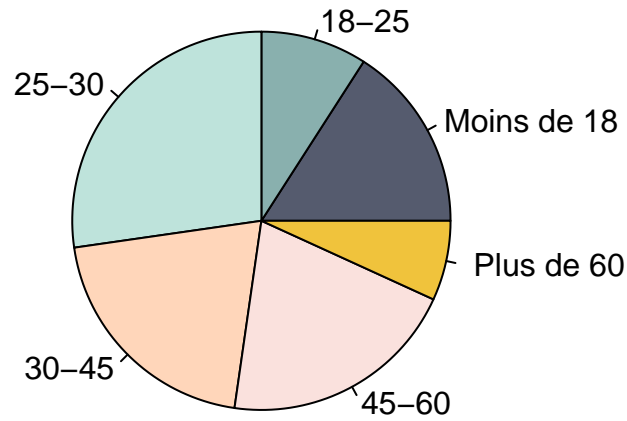
Répartition selon le Sexe



3.Echantillon selon l'age :

```
library(dplyr)
age <- Sport$Votre.âge...
lbls2 <- c("Moins de 18 ", "18-25", "25-30", "30-45", "45-60", " Plus de 60")
df1 <- data.frame(age) %>% group_by(age) %>% count(age)
colors=c("#555b6e", "#89b0ae", "#bee3db", "#ffd6ba", "#fae1dd", "#f0c33c")
pie(table(age), labels = lbls2, col=colors, main="Répartition selon Age")
```

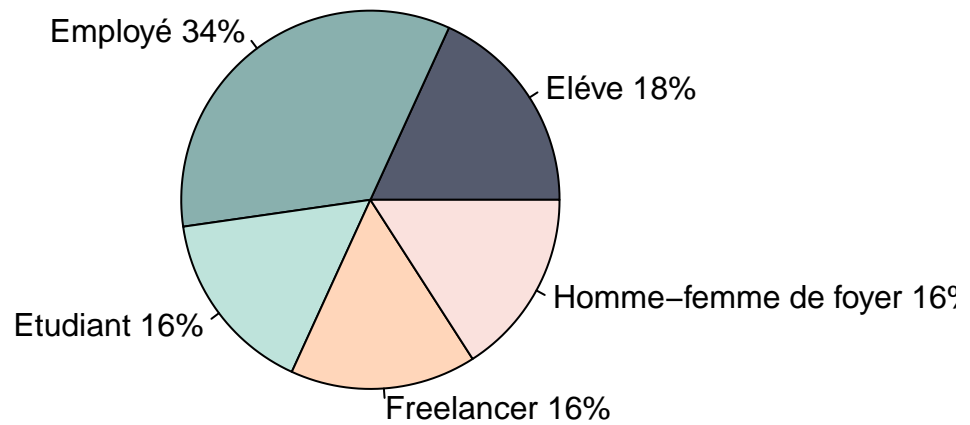
Répartition selon Age



4.La catégories socio-professionnelles de l'échantillon:

```
library(dplyr)
Categories <- Sport$Votre.Catégorie.socio.professionnelle..
freq4<-table(Categories)
pct4 <- round(freq4/sum(freq1)*100)
lbls4 <- names(freq4)
lbls4 <- paste(lbls4, pct4)
lbls4 <- paste(lbls4,"%",sep="")
colors=c("#555b6e","#89b0ae","#bee3db","#ffd6ba","#fae1dd")
pie(freq4,labels = lbls4, col=colors
    ,main="Répartition selon la Catégorie Socio_Professionnelle")
```

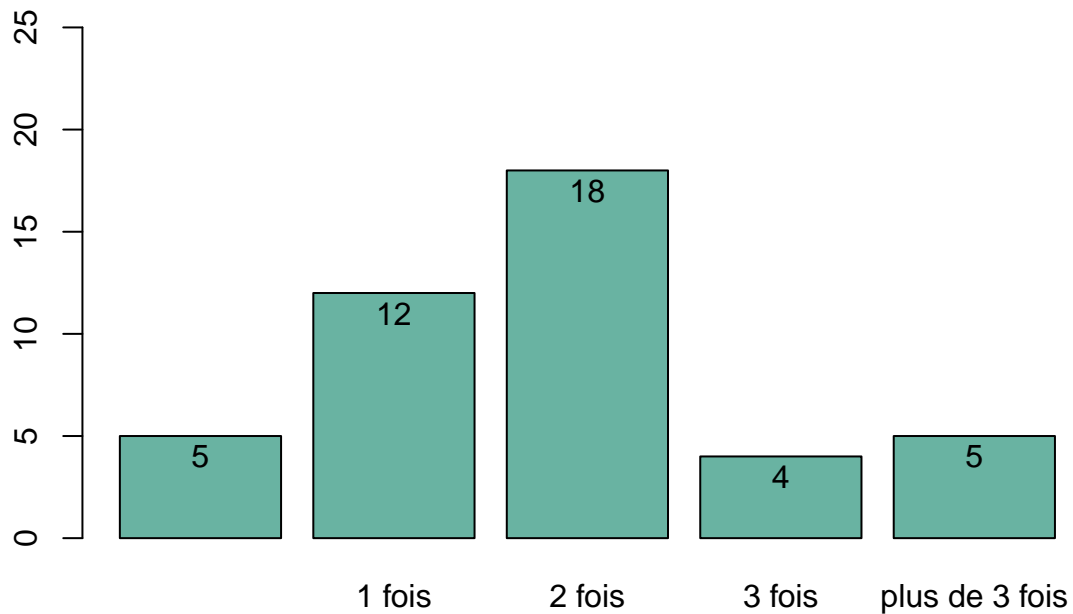
Répartition selon la Catégorie Socio_Professionnelle



5.le nombres de Pratiquer le Sport par semaine

```
library(dplyr)
Occupancy_Time <- Sport$Si.Oui..Combien.de.fois.par.semaine..
b <- barplot(table(Occupancy_Time),col="#69b3a2",ylim=c(0,25),main="Répartition selon le nombres de Pra
text(b[,1],table(Occupancy_Time)-1,table(Occupancy_Time))
```

Répartition selon le nombres de Pratiquer par semaine



Analyse en composantes principales:

1. But de l'ACP :

L'ACP a pour but d'étudier les différents critères selon lesquels les peuples choisissent leurs sports puis les regrouper selon un rapport d'importance .

```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.2.2
```

```
X=as.matrix(sportt[,32:51])  
res.pca1 <- PCA (X,graph=FALSE)
```

2. Les axes à conserver:

On commence par le critère du taux d'inertie cumulé et de Kaiser:

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.2
```

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

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

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

```
eig.val1 <- get_eigenvalue(res.pca1)
eig.val1
```

```
##      eigenvalue variance.percent cumulative.variance.percent
## Dim.1  7.96840526      39.8420263          39.84203
## Dim.2  3.28793568      16.4396784          56.28170
## Dim.3  2.04373256      10.2186628          66.50037
## Dim.4  1.36412595       6.8206297          73.32100
## Dim.5  0.95193931       4.7596966          78.08069
## Dim.6  0.82994625       4.1497312          82.23043
## Dim.7  0.51845642       2.5922821          84.82271
## Dim.8  0.45907789       2.2953894          87.11810
## Dim.9  0.38508187       1.9254094          89.04351
## Dim.10 0.34044812       1.7022406          90.74575
## Dim.11 0.31894577       1.5947288          92.34048
## Dim.12 0.29806798       1.4903399          93.83082
## Dim.13 0.24611232       1.2305616          95.06138
## Dim.14 0.22299623       1.1149812          96.17636
## Dim.15 0.20167446       1.0083723          97.18473
## Dim.16 0.18968191       0.9484096          98.13314
## Dim.17 0.14187073       0.7093537          98.84249
## Dim.18 0.11063002       0.5531501          99.39564
## Dim.19 0.07347926       0.3673963          99.76304
## Dim.20 0.04739199       0.2369599         100.00000
```

```
library(factoextra)
var1 <- get_pca_var(res.pca1)
var1
```

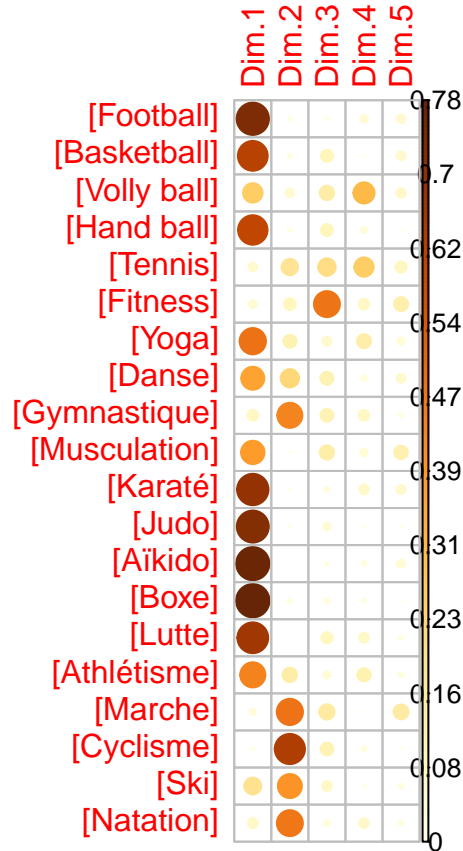
```
## Principal Component Analysis Results for variables
## =====
##      Name      Description
## 1 "$coord"    "Coordinates for the variables"
## 2 "$cor"      "Correlations between variables and dimensions"
## 3 "$cos2"     "Cos2 for the variables"
## 4 "$contrib"  "contributions of the variables"
```

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

```
## Warning: package 'corrplot' was built under R version 4.2.2
```

```
## corrplot 0.92 loaded
```

```
corrplot(var1$cos2, is.corr=FALSE)
```



Plus le cos2 est élevé plus la variable est mieux représentée .

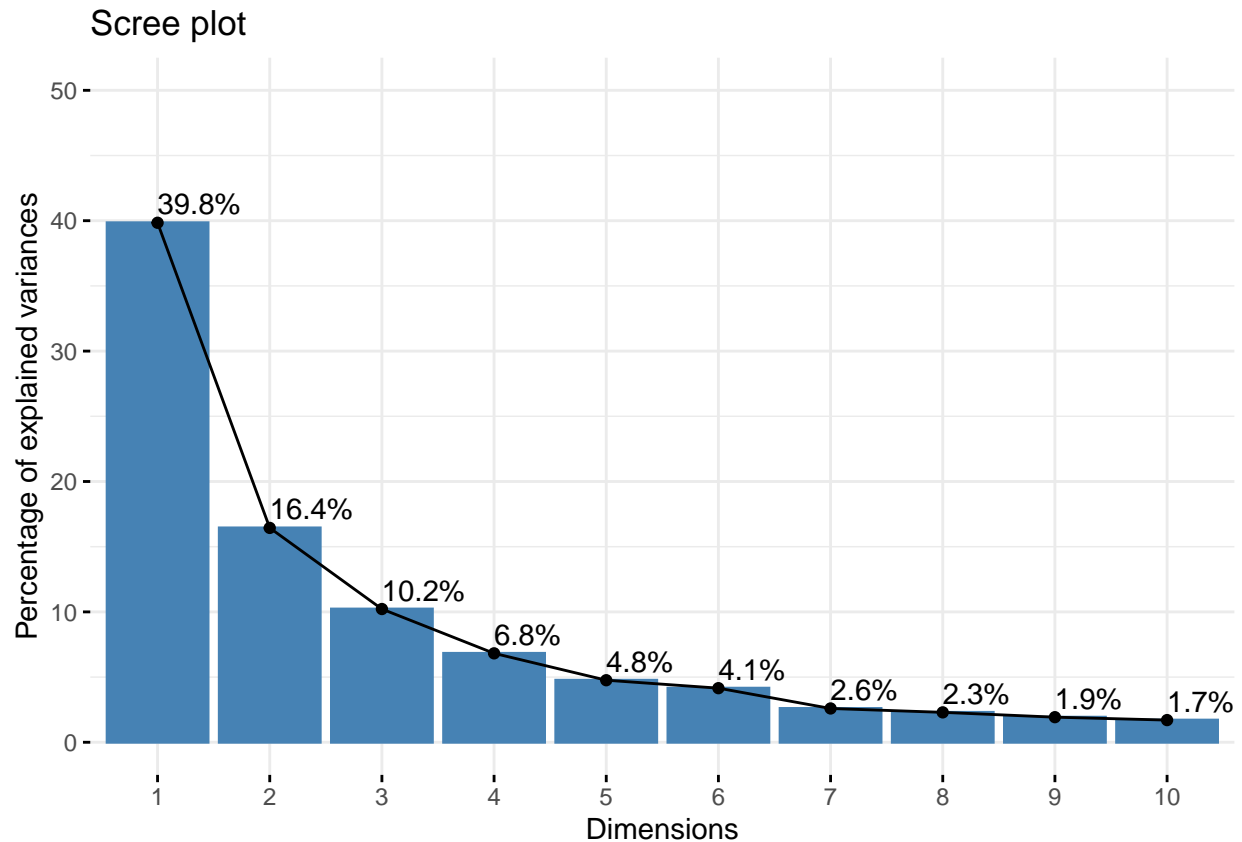
Par exemple, l'accès est bien présenté sur le 3ème axe alors qu'il est mal présenté sur le 1er axe. Sa coordonnée sur le 3ème axe est plus importante pour interpréter les composantes principales en considération alors que sa coordonnée sur le 1er est moins importante.

```
res.pca1$var$coord
```

##	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
## [Football]	0.8529024	-0.096233107	0.08419502	-0.20363651	-0.22907634
## [Basketball]	0.7872228	-0.111395025	0.31843446	-0.02342953	-0.24035728
## [Volly ball]	0.5136734	0.226058381	0.37958532	-0.55756149	0.25144090
## [Hand ball]	0.7773592	-0.104850558	0.31605740	0.16778640	0.01948238
## [Tennis]	0.2295445	0.433908886	0.45999755	-0.50976800	0.29462353
## [Fitness]	-0.2060933	0.311461077	0.68527630	0.28416345	-0.36580001
## [Yoga]	-0.6924791	0.343794193	0.23776106	0.37170436	0.18931766
## [Danse]	-0.6071530	0.482256524	0.34287573	0.14390221	0.23399548
## [Gymnastique]	-0.2868251	0.660510051	0.33799442	0.28160605	0.11430042
## [Musculature]	0.6167864	-0.020223507	-0.37796128	0.19297160	0.35524658
## [Karaté]	0.8279146	-0.032750416	-0.12741608	0.25640464	0.24036182
## [Judo]	0.8450469	-0.048441034	0.20395813	0.03369381	-0.05634320
## [Aikido]	0.8756363	0.003361166	-0.09177476	0.12940011	0.20575372
## [Boxe]	0.8809965	0.001009564	0.14687136	0.09630414	0.02240723
## [Lutte]	0.8128494	-0.001257960	0.29365480	0.25924726	-0.12137025
## [Athlétisme]	0.6633991	0.380388916	-0.18915108	0.34948427	0.13318161
## [Marche]	0.1679206	0.690721051	-0.40116055	-0.03113806	-0.39258056
## [Cyclisme]	0.0807039	0.797017161	-0.33763718	-0.12460994	0.04512309


```
## [Ski]          0.4441098  0.633815441 -0.26124813  0.07081194 -0.13324658
## [Natation]     0.2680805  0.685021210 -0.16080260 -0.25159569 -0.13211292
```

```
library(factoextra)
fviz_eig(res.pca1, addlabels = TRUE, ylim = c(0, 50))
```



D'après le critère du Coude, on observe un décrochement (coude) suivi d'une décroissance . On sélectionne les axes avant le coude. Selon le tracé de l'éboulis présenté ci-dessus, le nombre d'axe à conserver est de 4.

Critère de kaiser : On a 4 valeurs propres supérieures à 1

Critère de coude : Le coude est au niveau de deuxième axe.

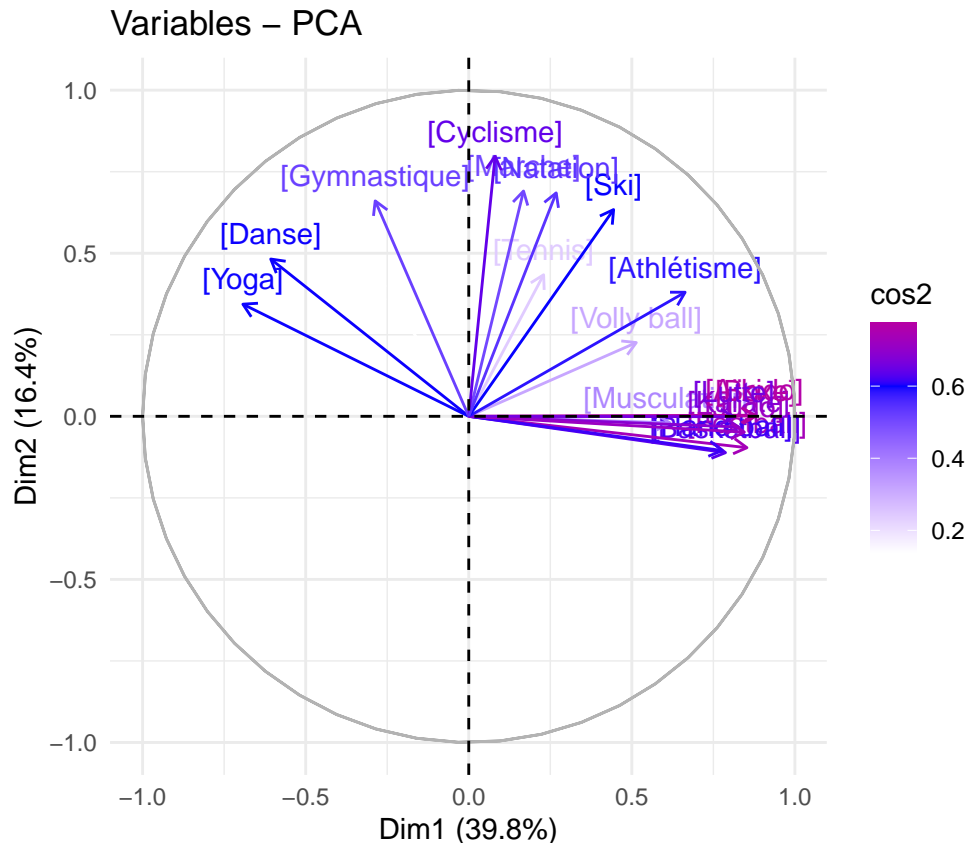
Critère d'inertie cumulée : Les deux premiers axes ont un taux d'inertie cumulée de 56%.

Décision: On retient les Deux premiers axes.

D'après ces trois critères le nombre d'axe à conserver est de 2 axes.

3.interprétation des axes :

```
fviz_pca_var(res.pca1, axes=c(1,2), col.var="cos2") +
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=0.6) +
  theme_minimal()
```



Interprétation de cette première carte des variables (i.e. axes 1 et 2) :

Les deux premières dimensions contiennent 70% de l'inertie totale

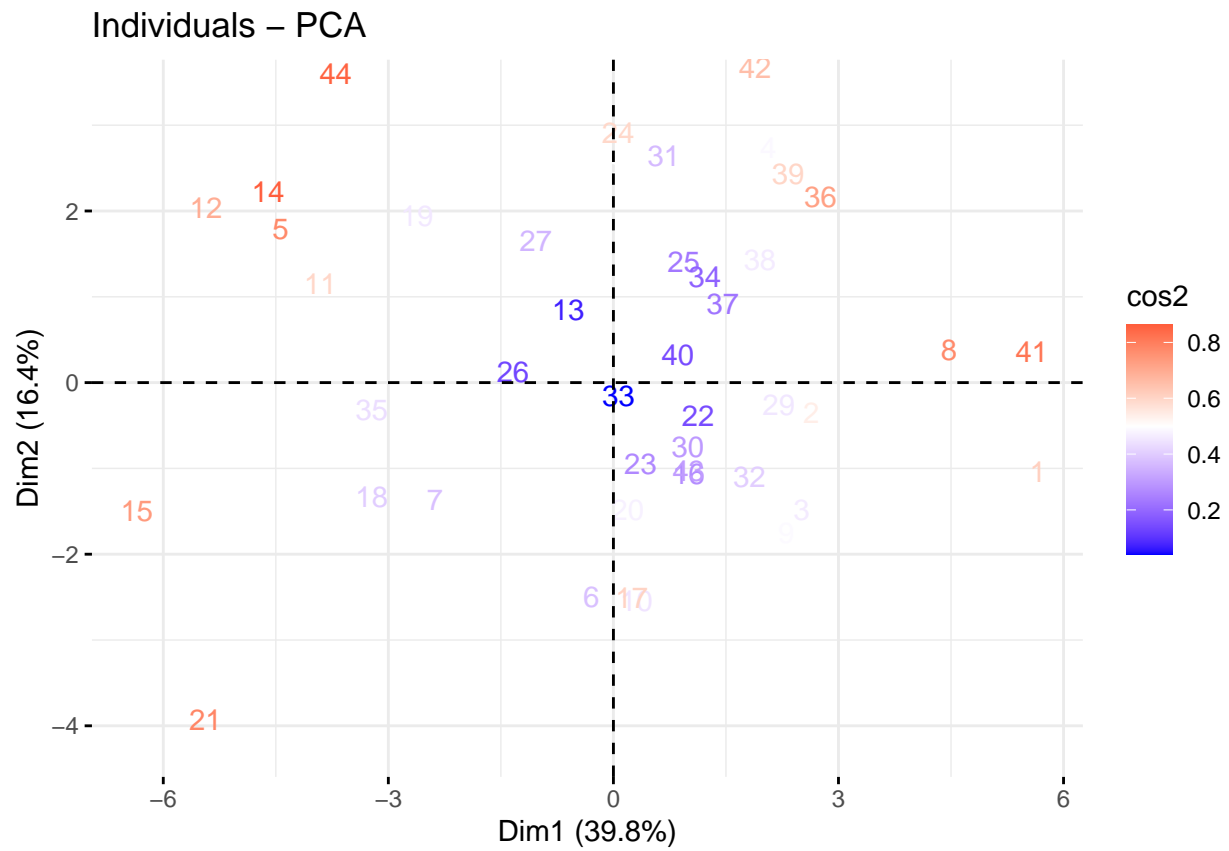
Les variables qui correspondent aux Football, Basketball, Musculation, Karaté, Boxe, Aikido, lutte sont corrélées négativement avec les variables Danse, Yoga, Gymnastique, Fitness.

On constate qu'après l'analyse du résultat que les variables Football, Basketball, Musculation, Karaté, Boxe, Aikido et lutte interviennent dans la contribution et la construction du premier axe. On conclura que la première axe oppose Les personnes qui préfèrent les activités Sportives que sont considérés comme plus appropriées aux Hommes.

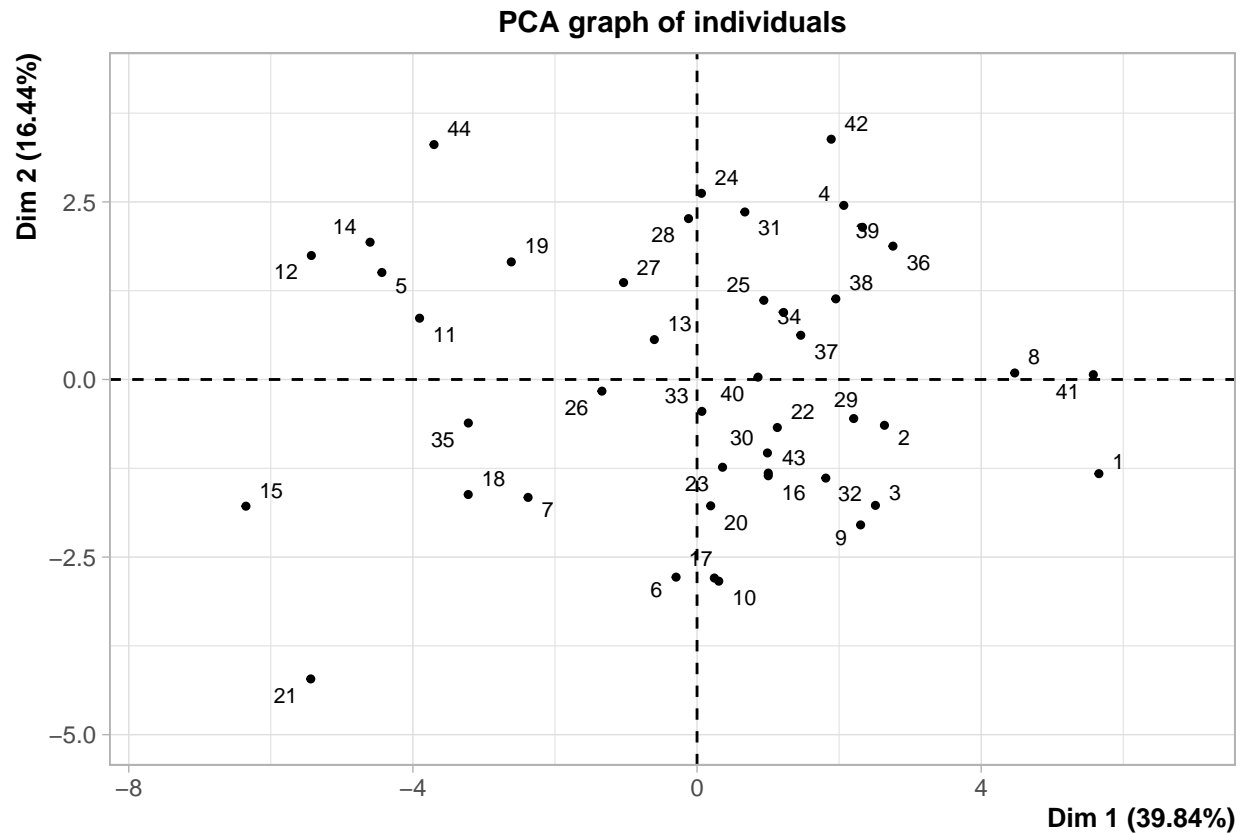
le deuxième axe opposé Les personnes qui preferent la Danse, Yoga, Gymnastique, Fitness. Ces choix alimentaires reflètent une préférence pour les activités Sportives que sont considérés comme plus appropriées aux Femmes .

4.interprétation de carte des individus :

```
fviz_pca_ind(res.pca1, geom = "text", col.ind = "cos2") +
scale_color_gradient2(low = "blue", mid = "white",
                      high = "red", midpoint = 0.5)
```



```
plot.PCA(res.pca1, axes=c(1, 2), choix="ind", cex=0.7)
```



Analyse de correspondance Multiple:

But de l'ACM:

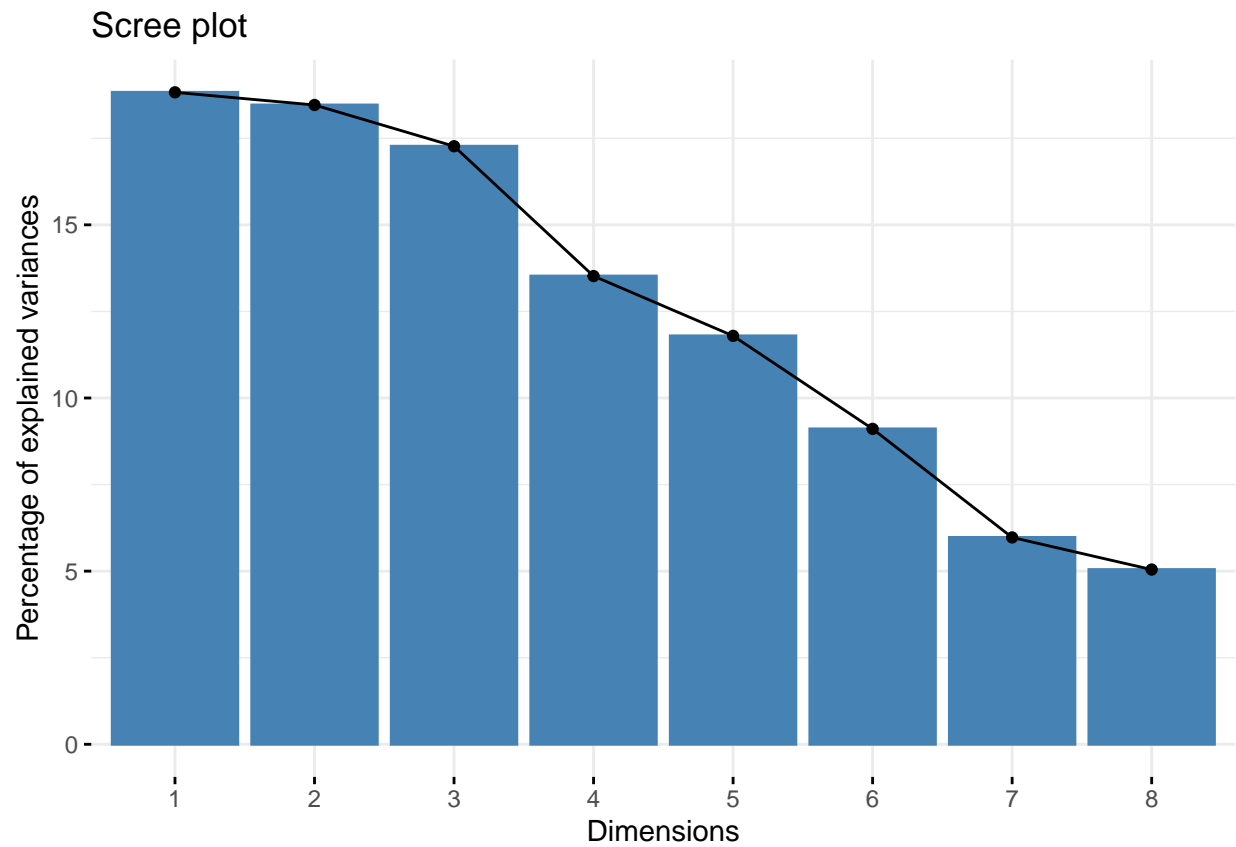
Résumer et visualiser le tableau de données contenant 6 niveaux de type sport lu par an et 6 types de contraintes rencontrées lors la lecture et les classes d'age. - L'objectif est d'identifier: * Un groupe de personnes ayant un profil similaire dans leurs réponses aux questions. * Les associations entre les catégories des variables.

1.Choix de nombre d'axes :

```
res.mca <- MCA (sportt[,c(24:31)], graph = FALSE)
res.mca$eig
```

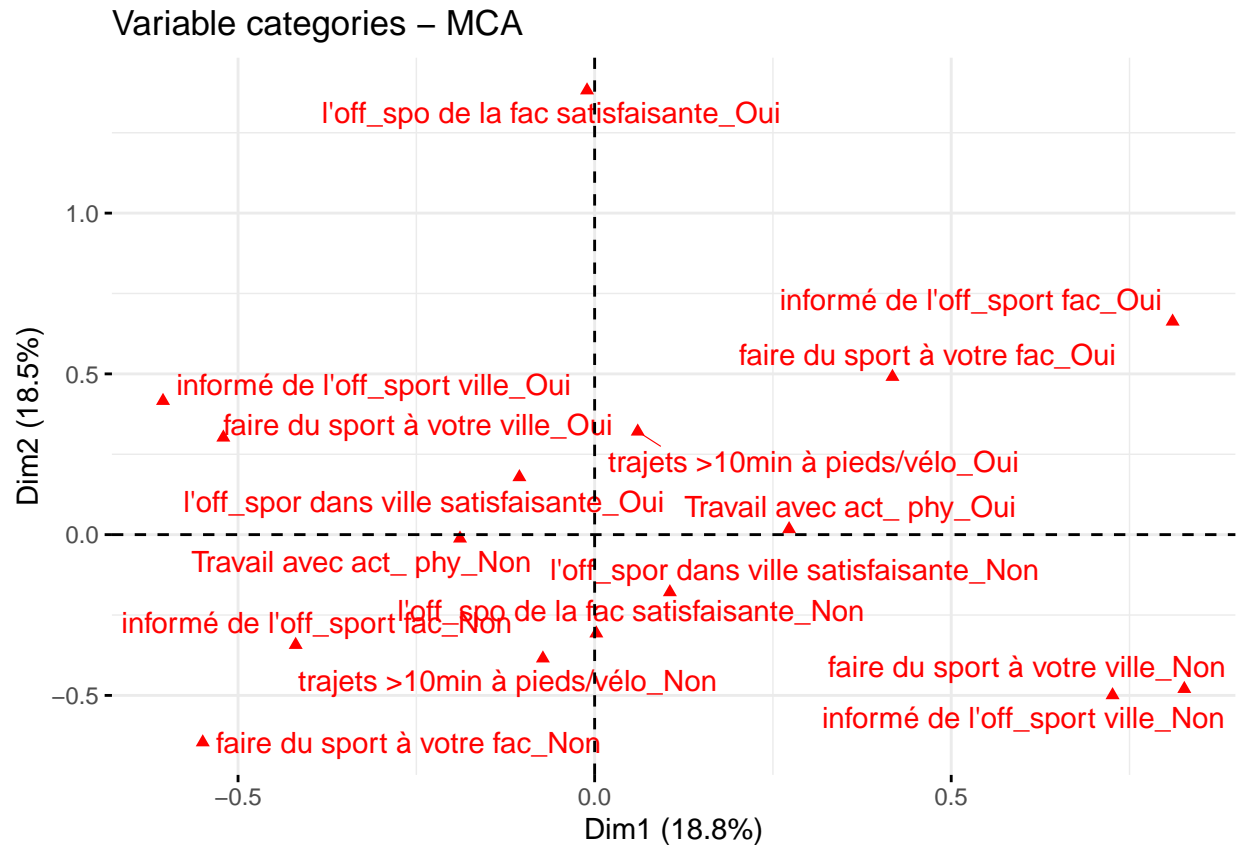
```
##      eigenvalue percentage of variance cumulative percentage of variance
## dim 1 0.18826456          18.826456          18.82646
## dim 2 0.18461073          18.461073          37.28753
## dim 3 0.17269821          17.269821          54.55735
## dim 4 0.13521063          13.521063          68.07841
## dim 5 0.11794209          11.794209          79.87262
## dim 6 0.09107891           9.107891          88.98051
## dim 7 0.05973623           5.973623          94.95414
## dim 8 0.05045865           5.045865         100.00000
```

```
fviz_screplot(res.mca)
```



2. Interprétation de la première carte des modalités.

```
fviz_mca_var (res.mca,  
              repel = TRUE,  
              ggtheme = theme_minimal ())
```

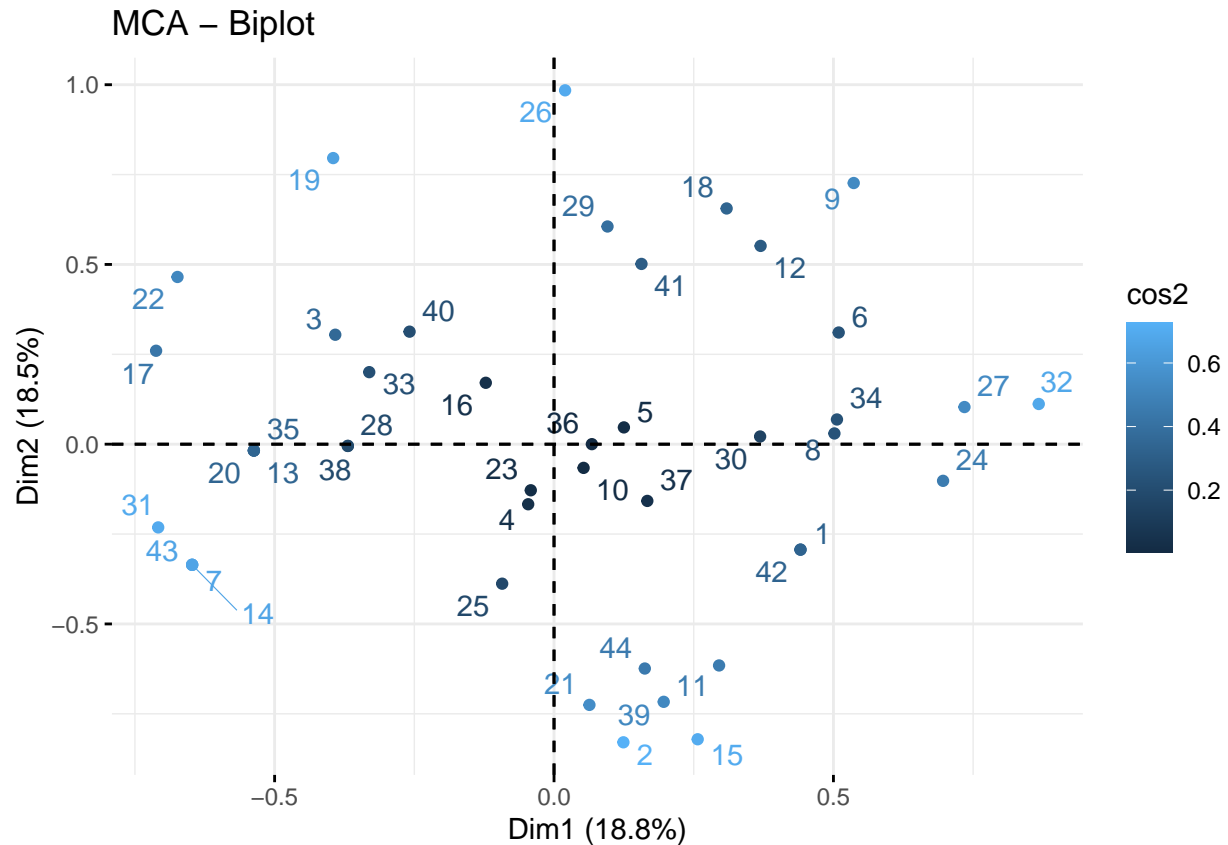


Il suffit d'identifier les groupes de modalités qui sont proches dans la carte. On a un premier groupe de modalités à gauche constitué des modalités: Non_informé par l'offre sportif , Ne le font pas dans la fac ,et Ne font pas un trajets au moins de 10min ,dont l'interprétation est la suivante : les individus qui ne sont pas intéressés par les activités sportives dans les facultés.

On a aussi un deuxième groupe à droite constitué de Non_informé par l'offre sportif dans la ville, le font dans la ville et font pas un trajets au moins de 10min dont l'interprétation est la suivante : les individus qui sont intéressés par les activités sportives organisées dans la ville.

3.Interprétation de la première carte des individus.

```
fviz_mca_biplot(res.mca,
  col.ind="cos2",
  invisible ="var",repel = TRUE,
  ggtheme = theme_minimal ())
```



Classification:

1. But de Classification

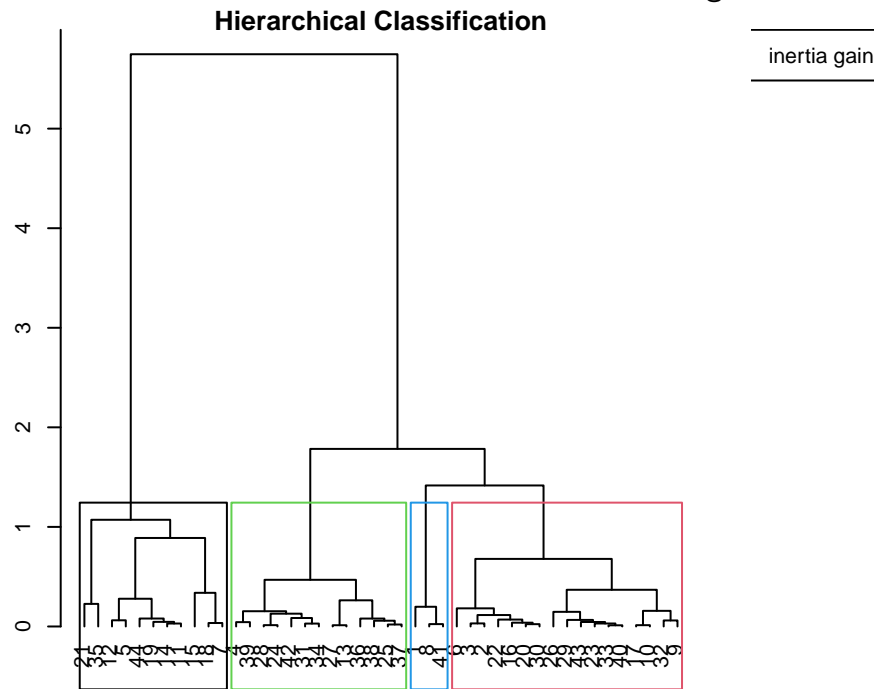
L'idée de classification de différents éléments d'un jeu de données en plusieurs catégories est de regrouper les datas en fonction de leur similarité. Comme les datas présentent des caractéristiques communes, il est plus facile de prédire leur comportement.

2. Classification

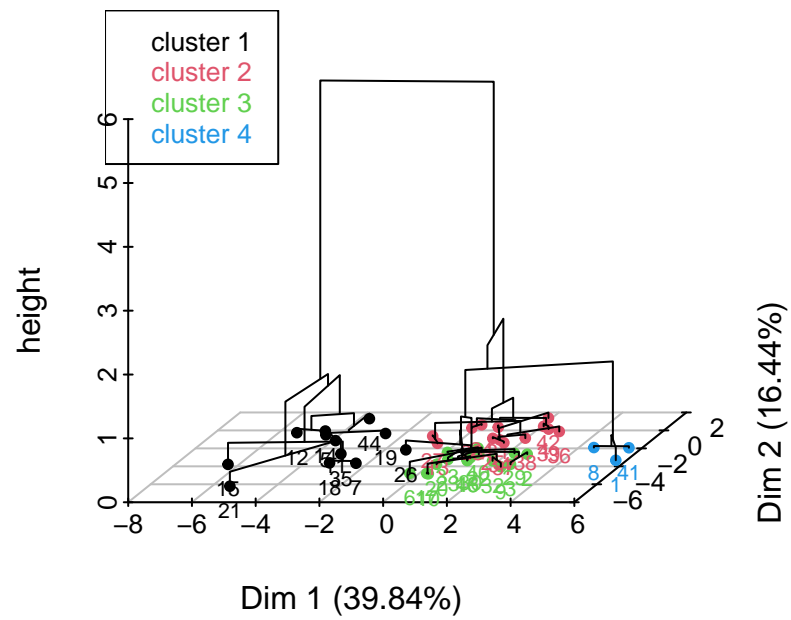
```
res.pca <- PCA(X, ncp = 4, graph = FALSE)
res.hcpc <- HCPC(res.pca, nb.clust = 4, graph = FALSE)
```

```
HCPC (res.pca1, nb.clust = 4, graph = TRUE)
```

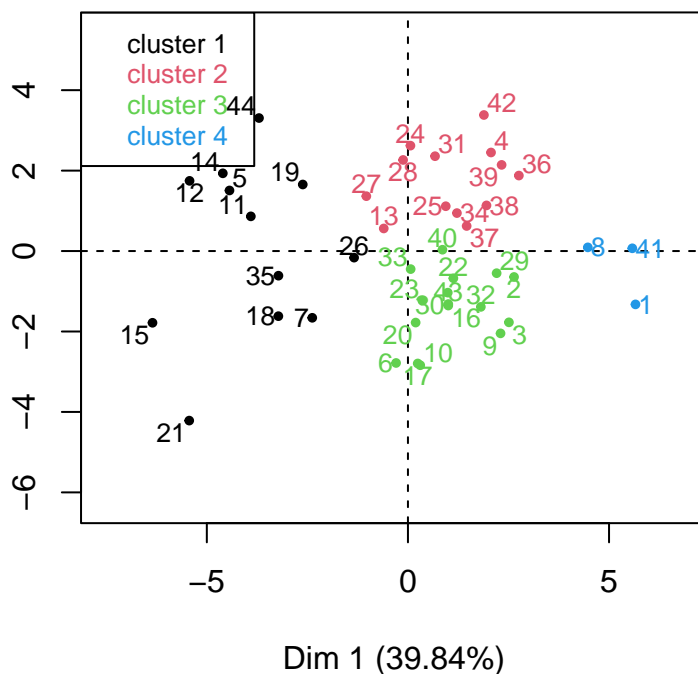
Hierarchical Clustering



Hierarchical clustering on the factor map



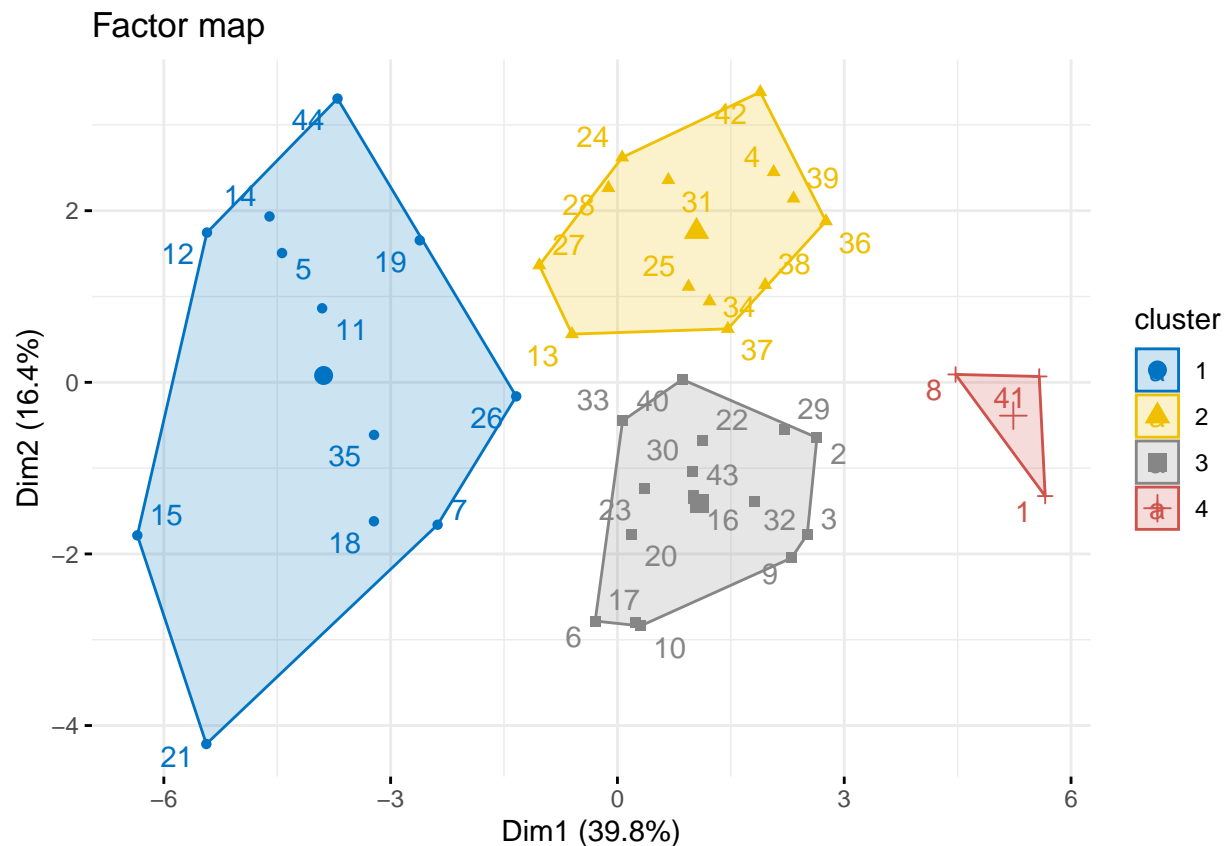
Factor map



```
## **Results for the Hierarchical Clustering on Principal Components**
##   name
## 1  "$data.clust"
## 2  "$desc.var"
## 3  "$desc.var$quanti.var"
## 4  "$desc.var$quanti"
## 5  "$desc.axes"
## 6  "$desc.axes$quanti.var"
## 7  "$desc.axes$quanti"
## 8  "$desc.ind"
## 9  "$desc.ind$para"
## 10 "$desc.ind$dist"
## 11 "$call"
## 12 "$call$t"
##   description
## 1  "dataset with the cluster of the individuals"
## 2  "description of the clusters by the variables"
## 3  "description of the cluster var. by the continuous var."
## 4  "description of the clusters by the continuous var."
## 5  "description of the clusters by the dimensions"
## 6  "description of the cluster var. by the axes"
## 7  "description of the clusters by the axes"
## 8  "description of the clusters by the individuals"
## 9  "parangons of each clusters"
## 10 "specific individuals"
## 11 "summary statistics"
```

```
## 12 "description of the tree"
```

```
fviz_cluster(res.hcpc,
  repel = TRUE,
  show.clust.cent = TRUE,
  palette = "jco",
  ggtheme = theme_minimal(),
  main = "Factor map"
)
```



3.Les Classes Observées :

- Cluster_1 [12-5-4-19-11-35-18-7-15-26-21]:la categorie des personnes qui préfèrent les ACT_SPORs de fitness et esthétiques (Yoga-Fitness-Danse),constituée majoritairement par les Femmes .
- Cluster_2 [24-13-14-25-34-37-38-39-42-31-28-27]:la categorie des personnes qui préfèrent les ACT_SPORs de relaxation et assez esthétiques(Marche-Tennis-VollyBall) ,constituée majoritairement par les femmes agés de 45ans.
- Cluster_3 [40-33-30-20-22-23-29-2-32-43-9-3-17-6-10]:la categorie des personnes qui préfèrent les ACT_SPORs de relaxation et assez Puissantes(Natation-Cyclisme-Athlétisme) ,constituée majoritairement par les Hommes ages de plus de 30ans.
- Cluster_4 [8-41-1]: la categorie des personnes qui préfèrent les ACT_SPORs de Combat(Boxe-Aikido-Karaté-Judo-Musculation) ,constituée majoritairement par les Hommes agés moins de 30ans.

Merci pour Votre attention - Yahia Chammami -