ENSIAS

جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

Projet d'évaluation

ACP : ANALYSE DONNES

- **Professeur :**

M.IBRAHIM AMRANI JOUTEI IDRISSI

- **Réalisé Par** :

Maataoui Mohamed

Bouibauan Mohamed

Sommaire

Table des matières

Chapitre 3

Introduction :

Après l'étape de la régression et la classification de notre jeu de données , On va par la suite étudier l'analyse en composante principale . Dans cette partie on essayera d'identifier la similarités entre les variables et les individues .

Partie 1 : **Application ACP normé** :

Question 1 : ACP normée sur le tableau des variables quantitatives .

```
library(FactoMineR)
library(ggplot2)
library(factoextra)
library(psych)
library("readxl")
library("corrplot")


# Set working directory
setwd("~/Projet")
# read Data
data <- read.csv2("data_projet6.csv",header=TRUE,sep=",",quote = "\"")
# convert char var to num var
data$turda <- as.numeric(as.character(data$turda))
data$X2_house_age <- as.numeric(as.character(data$X2_house_age))
data$X3_distance_to_the_nearest_MRT_station <- as.numeric(as.character(data$X3_distance_to_the_nearest_MRT_station))
data$X5_latitude <- as.numeric(as.character(data$X5_latitude))
data$X6_longitude <- as.numeric(as.character(data$X6_longitude))
data$dephm <- as.numeric(as.character(data$dephm))
data$SOLID <- as.numeric(as.character(data$SOLID))
data$Condi <- as.numeric(as.character(data$Condi))
data$Y_house_price_of_unit_area <- as.numeric(as.character(data$Y_house_price_of_unit_area))
# select var active
data.actifs <- data[,1:9]
# appliquer ACP

res <- PCA(data.actifs,scale.unit = TRUE,ncp = 5)
```

Figure 1 : Le script R total de l'application de l'ACP

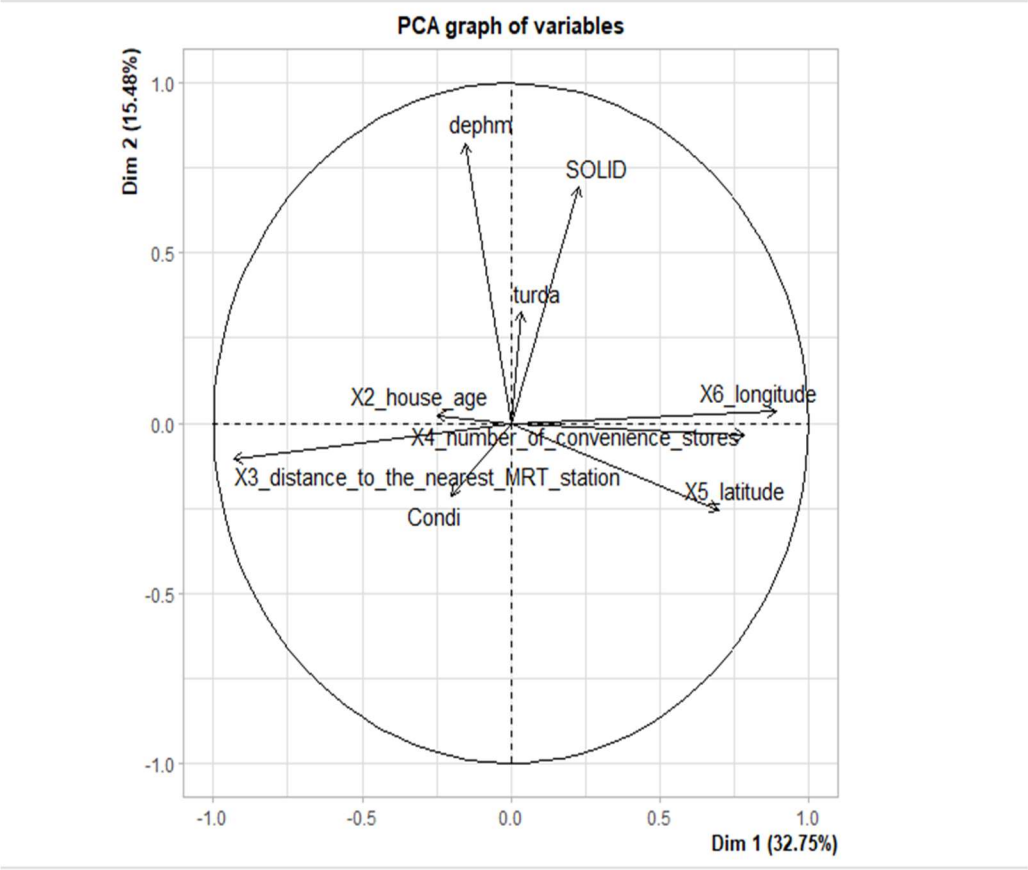
Pour appliquer l'ACP on a besoin de subdiviser le tableau en variable actif et variable supplémentaire actif avec la commande suivante :

```
# select var active
data.actifs <- data[,1:9]
```

Par la suite , on applique l'ACP normé :

```
res <- PCA(data.actifs,scale.unit = TRUE,ncp = 5)
```

La dernière commande donne le graphe ci-dessous comme résultat de traitement :



On voit que les variable X6 et X4 et X5 sont corrélées positivement par rapport à l'axe 1 alors que les variables X2 et X3 sont corrélées négativement par rapport à l'axe 1 . d'un autre coté les variables turda et Solid et dephm sont corrélées par rapport à l'axe 2 .

Question 2 : les raisons pour centrer et réduire les variables .

Le centrage et la réduction des variables forme des variables sans unité et par la suite les variables devient comparable .

Question 3 : l'indice de KMO et les indices MSAI .

L'indice KMO mesure l'importance des coefficients de corrélation linéaire par rapport à la somme des coefficients de corrélation linéaire partiel et les coefficients de corrélation linéaire plus le KMO est grand plus les coefficients de corrélation partiel sont faible , plus le KMO est faible plus les coefficients de corrélation partiel sont gras dans ce cas il difficile de déduire la dimension .

Le KMO de notre tableau nous donne le résultats suivant après la saisie de cette commande en R :

```
> KMO(cor(data.actifs))
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(data.actifs))
Overall MSA =  0.64
```

Figure 22 : Résultat de KMO commande en R

```
MSA for each item =
                                turda                                X2_house_age
                                 0.42                                        0.63
X3_distance_to_the_nearest_MRT_station        X4_number_of_convenience_stores
                                 0.61                                        0.75
                           X5_latitude                           X6_longitude
                                 0.82                                        0.63
                                dephm                                      SOLID
                                 0.43                                        0.65
                                Condi
                                 0.33
```

Figure 3 : Le MSA pour chaque variable

Le résultat obtenue pour le KMO total : 0.64 ce qui est une valeur entre médiocre et moyen .

Le MSA pour chaque variable on voit indice très faible pour les deux variables « dephm » et « Condi » .

Question 4 : Calcul des valeurs propres.

Afin de calculer les valeurs propres , le pourcentage d'inertie de chaque valeur propre et le cumul des pourcentages d'inertie on utilise la commande suivante :

```
> res$eig
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 2.94710240              32.7455822                           32.74558
comp 2 1.39318768              15.4798631                           48.22545
comp 3 1.23761620              13.7512911                           61.97674
comp 4 1.02098988              11.3443319                           73.32107
comp 5 0.72784783               8.0871981                           81.40827
comp 6 0.68360877               7.5956530                           89.00392
comp 7 0.46751128               5.1945698                           94.19849
comp 8 0.44150107               4.9055674                           99.10406
comp 9 0.08063489               0.8959432                          100.00000
```

Figure 4 : Calcul des valeurs propre de notre tableau

On a 9 composante c'est le nombre des variable actifs .

La première colonne du graphe en dessus désigne la valeurs propres pour chacune des composantes .

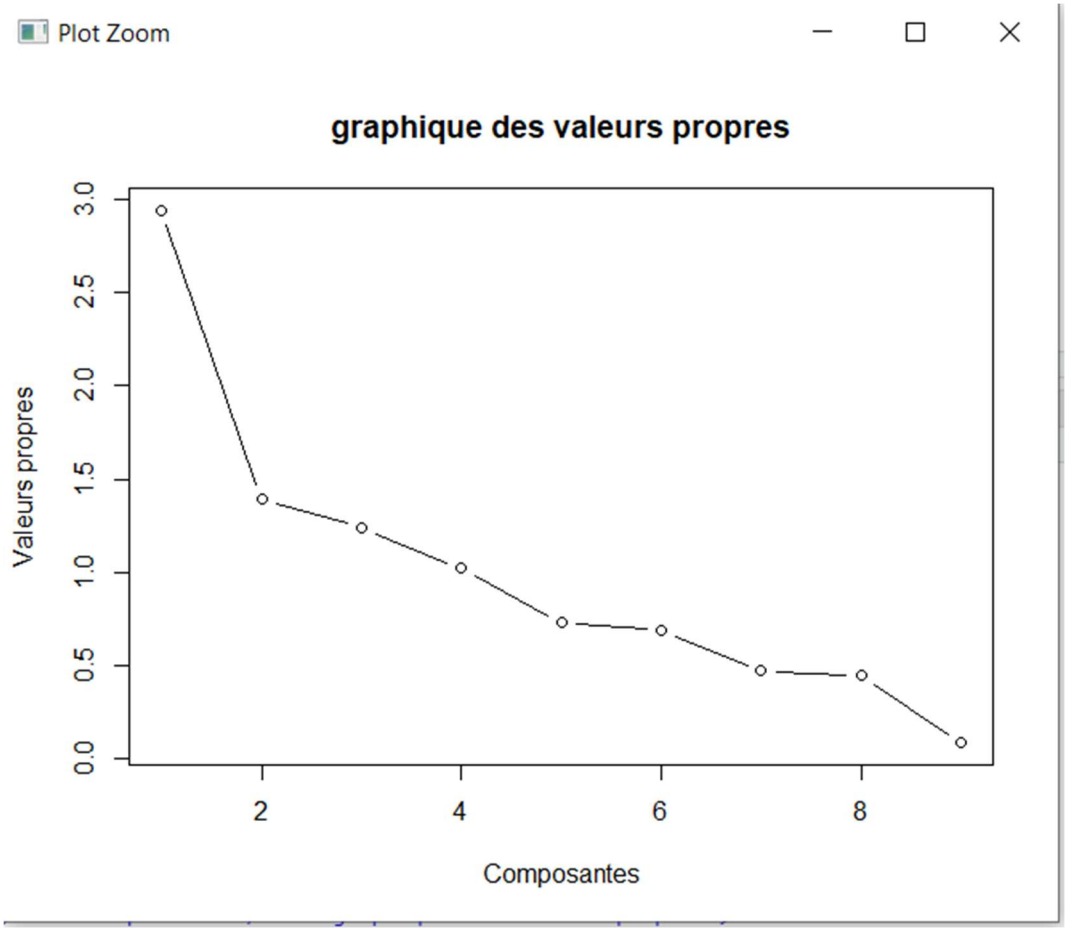
La deuxième colonne montre la variance et la troisième montre la variance cumulées .

Question 5 : Graphe des valeurs propres .

Le traçage de graphe des valeurs propres se fait par la commande R suivante :

```
plot(1:9,res$eig[,1],type="b",ylab="Valeurs propres",xlab="Composantes",main="graphique des valeurs propres")
```

Ce qui donne le résultat suivant :



On remarque que le graphe prend une forme plus au moins coudé et qu'on peut distinguer entre la partie horizontale dans laquelle on a une décroissance lente qui tend vers le 0 et la partie verticale caractérisé par une diminution forte de valeur propre .

Question 6 : Dimension du sous espace

On appliquons la règles du rapport :

la variance des 8 dernières valeurs propre par rapport à la variance de tout les valeurs propre on trouver le résultat suivant :

```
> var(res$eig[2:9,1])*700/(var(res$eig[,1])*8)
[1] 24.07561
```

Figure 5: Règle du rapport appliqué à les 8 derniers valeurs propres

Ce qui est un pourcentage très supérieur à 5 % .

la variance des 6 dernières valeurs propre par rapport à la variance de tout les valeurs propre on trouver le résultat suivant :

```
> var(res$eig[4:9,1])*500/(var(res$eig[,1])*8)
[1] 9.037854
```

Figure 6 : Règle du rapport appliqué à les 6 derniers valeurs propres

Ce qui est un pourcentage très supérieur à 5 % .

la variance des 5 dernières valeurs propre par rapport à la variance de tout les valeurs propre on trouver le résultat suivant :

```
> var(res$eig[5:9,1])*400/(var(res$eig[,1])*8)
[1] 4.699813
```

Figure 7 :Règle du rapport appliqué à les 5 derniers valeurs propres

On a trouver un pourcentage inférieur à 5 % .

Donc la dimension du sous éspace est 4

Partie 2 : **Nuage des variables** .

Question 7 : Cos^2 des variables .

Un cos2 élevé indique une bonne représentation de la variable sur les axes principaux en considération .

Le résultat obtenue par R des valeurs de cos2 sont les suivantes :

```
> res$var$cos2
                                        Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
turda                                0.001138326 0.1065822951 0.577192479 0.0021191128 0.277308250
X2_house_age                         0.060890906 0.0005574914 0.442591382 0.1854055582 0.254681519
X3_distance_to_the_nearest_MRT_station 0.864615635 0.0108574765 0.002051841 0.0021654634 0.017694245
X4_number_of_convenience_stores      0.616843591 0.0013230751 0.002128178 0.0231658840 0.061839221
X5_latitude                          0.490091090 0.0649761006 0.029868826 0.0757662330 0.031964834
X6_longitude                         0.796010900 0.0011394897 0.006394967 0.0167083367 0.001343555
dephm                                0.024201663 0.6782106426 0.001310131 0.0186210414 0.057993684
SOLID                                0.053031047 0.4843267572 0.071891197 0.0003819772 0.007985820
Condi                                0.040279242 0.0452143497 0.104187202 0.6966562686 0.017036703
```

Figure 7 : Cos2 des variables

Question 8 : Cos^2 des variables .

Les variables X3 , X4 et X6 ont une valeur de cos2 élevé dont il sont bien représentés sur les axes principales . alors que les variables Condi , dephm ,turda ,X2 et Solid ont une très faible valeur de cos2 dont il sont faiblement présenté sur ces axes principales . la X5 est moyennement présentée

Question 9 : La contribution des variables .

Pour avoir la contribution des variables dans chaque sous espace on utilise la commande suivante :

```
> res$var$contrib
                                          Dim.1       Dim.2      Dim.3       Dim.4      Dim.5
turda                                 0.03862525  7.65024675 46.6374371  0.20755473 38.0997563
X2_house_age                          2.06612795  0.04001553 35.7616021 18.15939244 34.9910391
X3_distance_to_the_nearest_MRT_station 29.33782128  0.77932619  0.1657897  0.21209450  2.4310363
X4_number_of_convenience_stores       20.93051096  0.09496747  0.1719579  2.26896315  8.4961743
X5_latitude                           16.62959149  4.66384405  2.4134159  7.42086037  4.3916919
X6_longitude                          27.00995053  0.08179011  0.5167165  1.63648407  0.1845928
dephm                                 0.82120196 48.68049390  0.1058592  1.82382234  7.9678309
SOLID                                 1.79943007 34.76392771  5.8088442  0.03741244  1.0971826
Condi                                 1.36674050  3.24538829  8.4183774 68.23341596  2.3406957
```

Figure 8 : contribution des variables

Plus la valeur de la contribution est importante, plus la variable contribue à la composante principale en question .Afin de visualiser cette contribution pour simplifier la lecture on utilise la commande R suivante :

```
> corrplot(res$var$contrib, is.corr=FALSE)
```

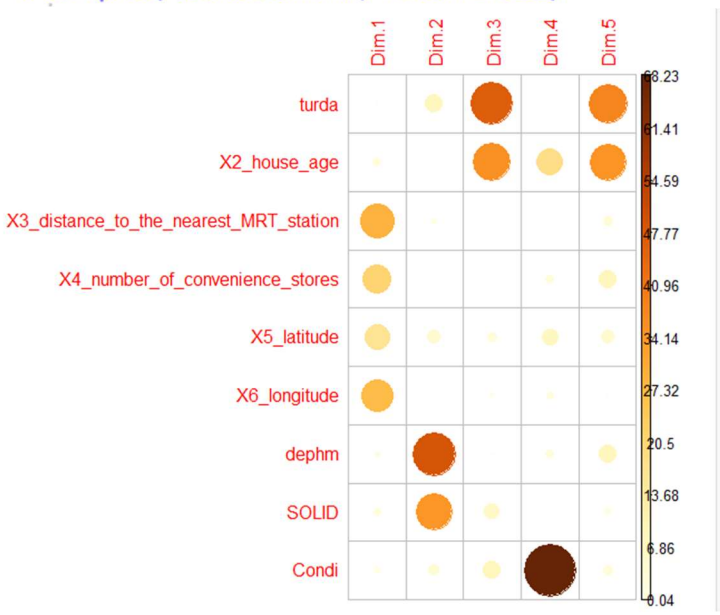
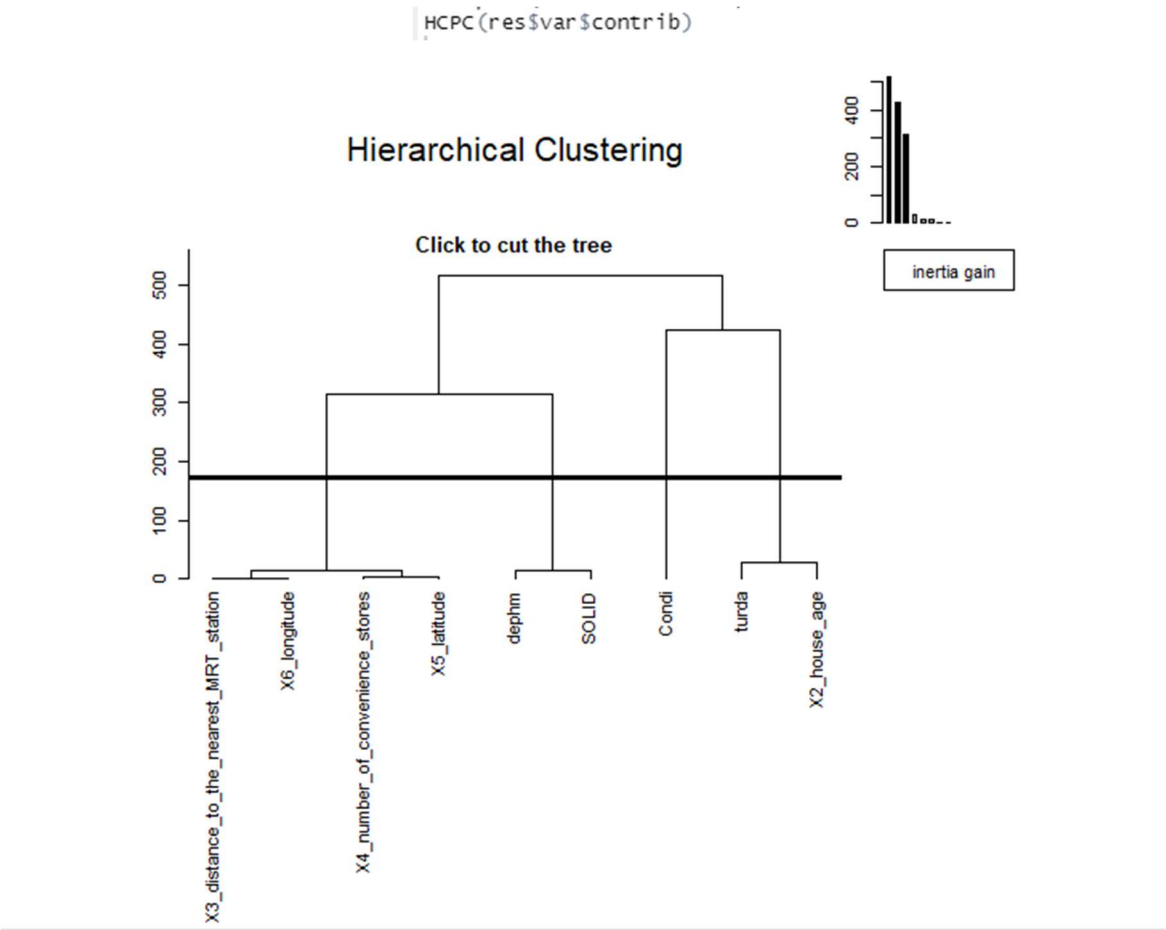


Figure 9 : Visualisation de la contribution des variable

Question 10 : La CAH

On applique la CAH à notre tableau des contributions avec la commande suivante :



D'après la figure on voit que la CAH à regrouper 4 classes :

Classe 1 : contient les variables X3 ,X6,X4 et X5 .

Classe 2 : contient les variables dephm ,SOLID .

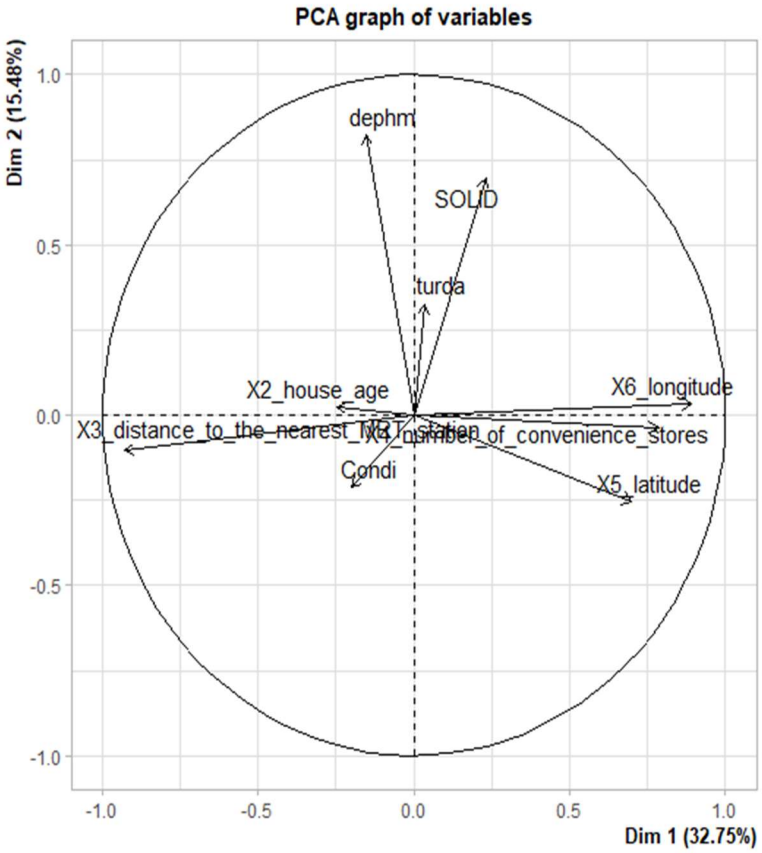
Classe 3 : la variable Condi .

Classe 4 : turda , X2 .

Question 11 : Le nuage des variables projeté sur les 2 premières axes .

Pour Afficher le nuage des variables projeté sur les 2 premières axes on utilise la commande suivante :

```
> res<- PCA(data.actifs,ncp=5,axes=c(1,2))
```



Question 12 : Les variables bien corrélées :

On conclut d'après la figure que les variables X5 , X4 et X6 sont positivement bien corrélées avec l'axe_1 alors que les variables X2 , X3 sont négativement bien corrélées avec l'axe_1 .

Partie3 : **Nuage des individus** .

Question 13 : Le cos2 des individus sur le sous espace :

Pour afficher le cos^2 des individus sur le sous espace on utilise la commande R suivante :

```
> res$ind$cos2
          Dim.1        Dim.2        Dim.3        Dim.4        Dim.5
0  0.165265491 0.3375172193 2.442119e-01 2.195572e-02 1.324163e-01
1  0.294584254 0.4179303751 2.002572e-03 1.639079e-03 1.319549e-01
2  0.543456732 0.1465255631 1.758697e-02 1.361773e-01 1.230296e-01
3  0.399919822 0.0462222641 3.829053e-02 3.688103e-01 1.265026e-01
4  0.241573427 0.1014101172 5.755436e-01 7.071474e-02 1.210316e-04
5  0.366144002 0.2588960248 2.589759e-01 1.131291e-02 3.602520e-02
6  0.134044597 0.2066208872 3.983426e-01 1.082032e-02 1.569322e-01
7  0.243109560 0.4271230287 8.485764e-02 1.562154e-01 5.161807e-02
8  0.731098041 0.1735098134 9.093857e-03 4.864353e-04 5.854509e-02
9  0.332137809 0.3514277459 9.679353e-03 3.221839e-02 3.585944e-02
10 0.016998008 0.1306143414 7.230666e-02 4.506326e-01 2.570152e-02
11 0.524374149 0.0045986263 2.465292e-01 1.710748e-02 9.999992e-04
12 0.046502562 0.1675258351 4.517850e-01 5.604191e-02 8.855879e-02
13 0.408187667 0.1005907203 2.405894e-01 9.252792e-02 6.683164e-02
14 0.130596591 0.0284340435 2.990110e-01 3.480018e-01 1.094362e-02
```

Question 14 : Les individus bien représentées , moyennement représentées et faiblement représentées sur chaque sous espace :

Notre tableaux de donnés contient 70 lignes donc il est difficile d'identifier chaque individus est-ce qu'il est bien représenté ou moyennement ou faiblement mais en générale les individus avec une faible valeurs de cos^2 sont faiblement représentés et ceux avec une grande valeur de cos^2 sont bien représentées .

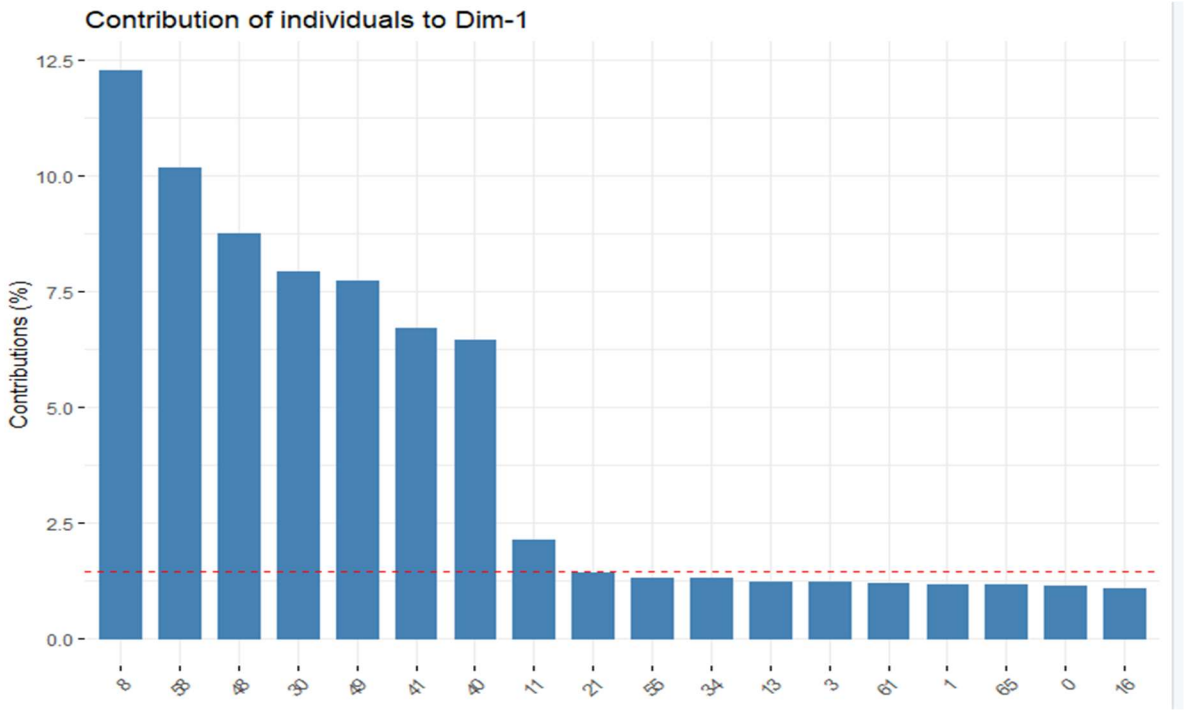
Question 15 : La contribution des individus dans chaque axe du sous espace :

Pour afficher la contribution des individus dans tous les sous espace on utilise la commande suivante :

```
> res$ind$contrib
          Dim.1       Dim.2       Dim.3       Dim.4
0   1.137316073  4.91338204 4.0019816256 0.4361345728
1   1.147898245  3.44495379 0.0185819510 0.0184360388
2   1.033350942  0.58936164 0.0796311839 0.7474142790
3   1.199072215  0.29316305 0.2733839267 3.1918996781
4   0.919832369  0.81682010 5.2185156199 0.7772187279
5   0.995897663  1.48961410 1.6773795909 0.0888200134
6   0.312416926  1.01869709 2.2108091356 0.0727946159
7   0.525581644  1.95333504 0.4368560383 0.9748457123
8  12.278449213  6.16421165 0.3636848713 0.0235812491
9   0.789412647  1.76688145 0.0547824325 0.2210361863
10  0.056058391  0.91121125 0.5678454386 4.2898194617
11  2.127314431  0.03946428 2.3815952632 0.2003320057
12  0.106750358  0.81350429 2.4696393576 0.3713463586
13  1.211152242  0.63136773 1.6999047175 0.7924747870
14  0.440487456  0.20287378 2.4015857309 3.3881057242
15  0.207022200  0.03434056 0.3561325039 4.2856305929
16  1.079221458  0.01856350 0.1827481477 0.9291039875
17  0.009930898  0.09984701 0.1320913404 0.5888067336
```

Pour mieux visualiser la contribution des individus on utilise la commande suivante :

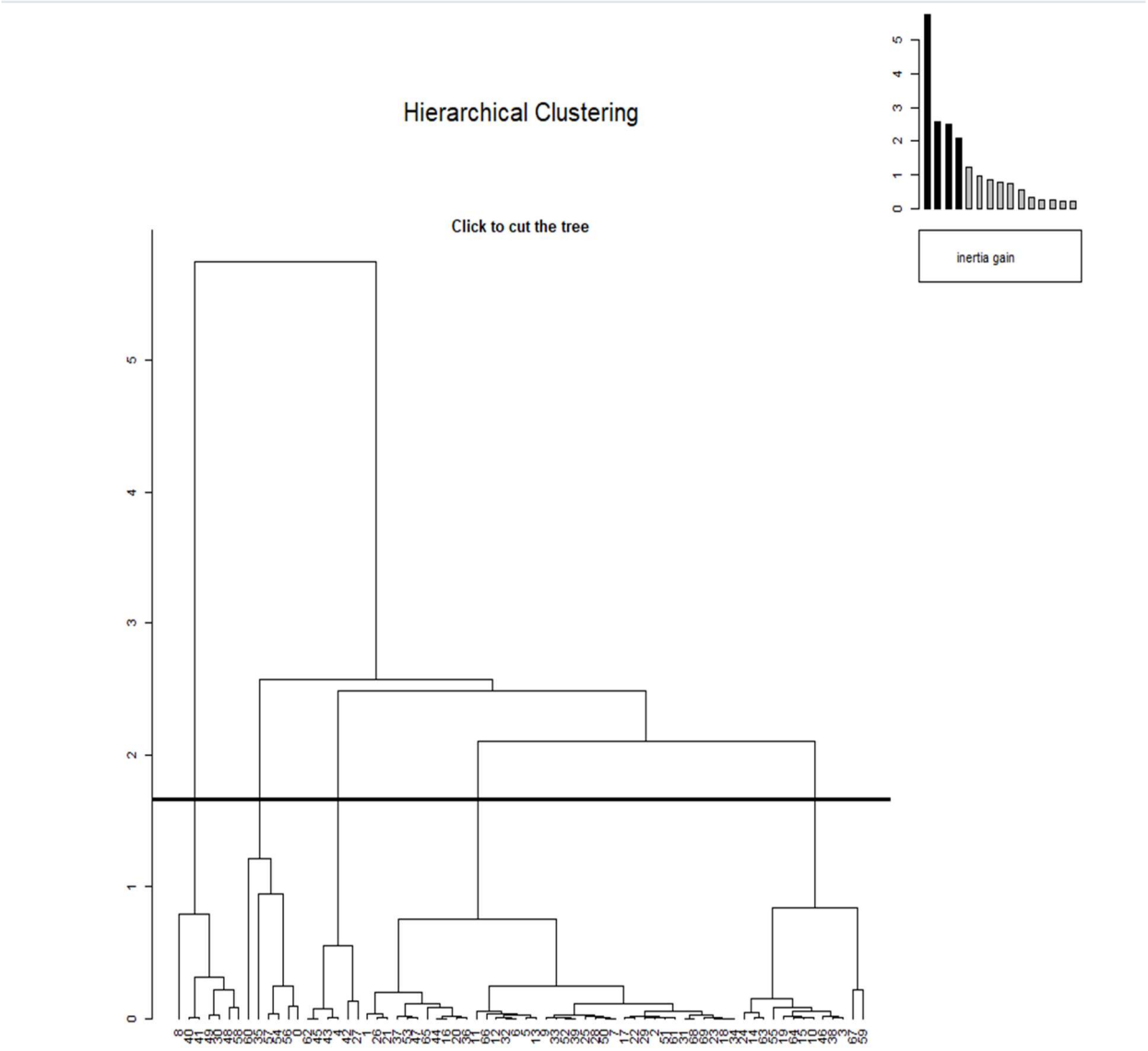
Fziz_contrib(res,choixe= 'ind')



On voit que l'individu 8 est le plus contributif dans le cas de la dimension 1 suivi par 58 .

Question 16 : CAH au tableau de contribution des individus .

On applique le CAH au tableau des contributions des individus on trouve le résultat suivant :

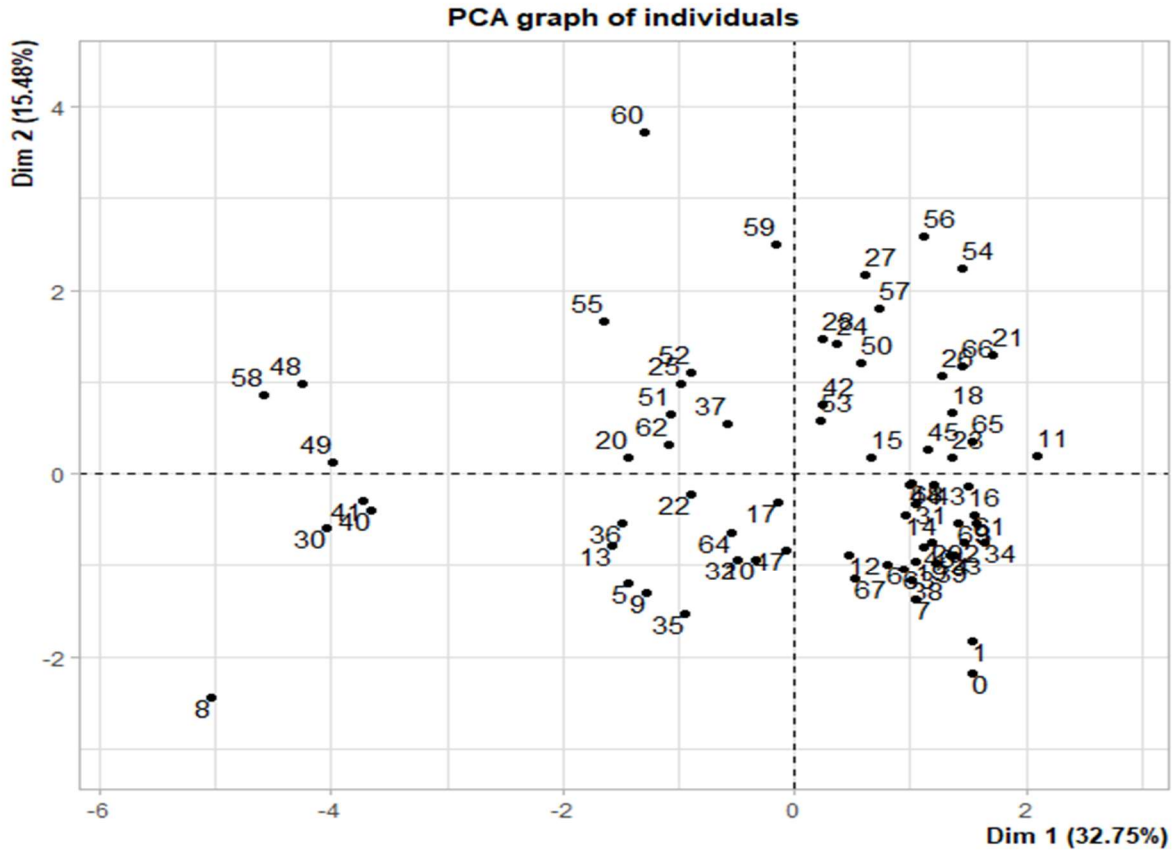


Le CAH appliqué aux tableaux de contribution des individus à permet de découper les individus en 5 classes ,une grande partie des individus se trouve dans la 4 classe cad que la la grande partie des individus de la classe 4 partage des caractéristiques semblables .

Question 17 : Nuage des individus .

On utilise la commande suivante pour afficher le nuage des individus

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



Conclusion

Dans Une première étape on a commencé par appliqué ACP normé puis on a calculé indice KMO globale et on a calculé les indice MSAI de chaque variable . la détermination de ces indices nous a donné une vision sur la corrélation entre les variables . Par la suite on a déterminé la dimension du sous espace en se basant sur la méthode du rapport .

Le calcul de cos2 à permit de connaitre les variables et les individus les mieux représentés et les mieux contributifs dans le sous espace d'étude .et finalement le CAH a permit de déterminer les classes des variables et des individus .