ENSIAS

جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

Projet d'évaluation

ACM : ANALYSE DONNES

- **Professeur :**

M.IBRAHIM AMRANI JOUTEI IDRISSI

- **Réalisé Par** :

Maataoui Mohamed

Bouibauan Mohamed

Table des matières

Partie 1 :

1.

2. **Question** : Tableau Disjonctif .

La méthode « DiscretizeDF » située dans le package « arules » permet de transformer les variables quantitatives en varibales qualitative de n modalités avec la méthode kmeans ou bien frequency .

```
data_tran <- discretizeDF(data, default = list(method = "cluster", breaks = 2, labels = c("0", "1")))
View(data_tran)
```

Le résultat :

	turda	X2_house_age	X3_distance_to_the_nearest_MRT_station	X4_number_of_convenience_stores	X5_latitude	X6_longitude	dephm	SOLID
0	1	1	0	1	1	1	0	0
1	0	0	0	1	1	1	0	0
2	1	0	0	1	1	1	0	0
3	1	0	0	1	1	1	0	0
4	0	0	0	1	1	1	0	0
5	0	0	0	0	0	0	0	0
6	1	1	0	1	1	1	0	0
7	0	0	0	1	1	1	0	0
8	0	1	1	0	0	0	0	0
9	1	0	0	0	0	0	0	0
10	0	1	0	0	1	1	0	0
11	0	0	0	1	1	1	0	1
12	0	0	0	1	0	1	0	0
13	0	0	0	1	0	0	0	0
14	0	0	0	1	1	1	0	1
15	0	1	0	0	1	1	0	1

Puis on construit le tableau disjonctif complet par la commande :

```
k <- tab.disjonctif(data_tran)
k
```

Le résultat :

```
> k
   0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1
0  0 1 0 1 1 0 0 1 0 1 0 1 1 0 1 0 0 1 0 1
1  1 0 1 0 1 0 0 1 0 1 0 1 1 0 1 0 1 0 0 1
2  0 1 1 0 1 0 0 1 0 1 0 1 1 0 1 0 1 0 0 1
3  0 1 1 0 1 0 0 1 0 1 0 1 1 0 1 0 1 0 0 1
4  1 0 1 0 1 0 0 1 0 1 0 1 1 0 1 0 1 0 0 1
5  1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0
6  0 1 0 1 1 0 0 1 0 1 0 1 1 0 1 0 0 1 0 1
7  1 0 1 0 1 0 0 1 0 1 0 1 1 0 1 0 0 1 0 1
8  1 0 0 1 0 1 1 0 1 0 1 0 1 0 1 0 0 1 1 0
9  0 1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0
10 1 0 0 1 1 0 1 0 0 1 0 1 1 0 1 0 1 0 0 1
11 1 0 1 0 1 0 0 1 0 1 0 1 1 0 0 1 1 0 0 1
12 1 0 1 0 1 0 0 1 1 0 0 1 1 0 1 0 1 0 0 1
13 1 0 1 0 1 0 0 1 1 0 1 0 1 0 1 0 1 0 1 0
14 1 0 1 0 1 0 0 1 0 1 0 1 1 0 0 1 1 0 1 0
15 1 0 0 1 1 0 1 0 0 1 0 1 1 0 0 1 1 0 0 1
```

Dans la figure ci-dessus, on constate que dans chaque ligne correspond à un individu, ce dernier est relevé par des 1 et des 0. Le 0 indique que l'individu n'est pas caractérise par cette modalité, par contre le 1 signifie que l'individu est caractérise par cette modalité.

3. **Question** : La fréquence des modalités .

Le calcul du fréquence de chaque modalité permet de détecter les modalités les plus fréquentes et celles les moins fréquentes. Les moins fréquentes possèdent une fréquence inférieur à 0.01 .

Avant le calcul de la fréquence de chaque modalité, on calcule le nombre d'effectif par modalité via la commande :

```
> apply(k,2,sum)
 0  1  0  1  0  1  0  1  0  1  0  1  0  1  0  1  0  1  0  1
31 39 47 23 62  8 26 44 29 41 17 53 53 17 48 22 39 31 29
 1
41
```

La somme des individus dans chaque modalité est égale aux nombres d'obéservations.

Après le calcul des effectifs de chaque modalité, on les divise par le nombre d'observation et on retrouve la fréquence de chaque modalité. Par la commande suivante

```
> propmod=apply(k,2,sum)/(nrow(k))
> propmod
        0         1         0         1         0         1         0         1         0         1         0         1
0.4428571 0.5571429 0.6714286 0.3285714 0.8857143 0.1142857 0.3714286 0.6285714 0.4142857 0.5857143 0.2428571 0.7571429
        0         1         0         1         0         1         0         1
0.7571429 0.2428571 0.6857143 0.3142857 0.5571429 0.4428571 0.4142857 0.5857143
```

Tout les fréquences sont supérieures à 0.01 .

4. **Question** : Application de L'ACM .

On supprime tt d'abord la variable Y_house_Price via la commande suivante :

```
> H <- subset(data_tran,select=-Y_house_price_of_unit_area)
```

Puis on applique ACM avec un nombre de composante principale égale à 5 .

```
> res<-MCA(H,ncp=8,graph=FALSE,axes=c(2.3))
> res
**Results of the Multiple Correspondence Analysis (MCA)**
The analysis was performed on 70 individuals, described by 9 variables
*The results are available in the following objects:

   name              description
1  "$eig"            "eigenvalues"
2  "$var"            "results for the variables"
3  "$var$coord"      "coord. of the categories"
4  "$var$cos2"       "cos2 for the categories"
5  "$var$contrib"    "contributions of the categories"
6  "$var$v.test"     "v-test for the categories"
7  "$ind"            "results for the individuals"
8  "$ind$coord"      "coord. for the individuals"
9  "$ind$cos2"       "cos2 for the individuals"
10 "$ind$contrib"    "contributions of the individuals"
11 "$call"           "intermediate results"
12 "$call$marge.col" "weights of columns"
13 "$call$marge.li"  "weights of rows"
```

Le nombre de variables étant égale à 10 trouvé via la commande S<-ncol(var) .

Le nombre d'individus est 70 trouvé acev la commande N<-nrow(var) .

Le nombre de modalité est égale à 10*2=20 .chaque variable posséde de modalité .

5. **Question** : Les valeurs propres .

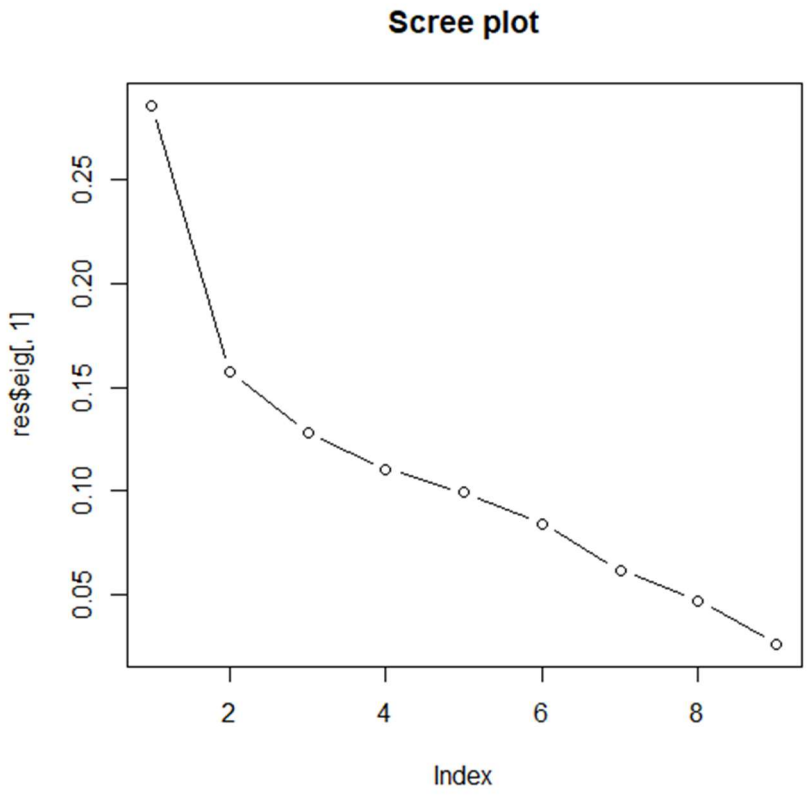
Le calcul des valeurs propres permet d'extraire les composantes ayant une portion significative de la variance totale,tout composante ayant une valeur propre supérieure à 1 est considéré comme sinificative .pour calculer les valeurs propres ,le pourcentage d'inertie de chaque valeur propre ainsi que le comul des pourcentages on utilise :

```
> res$eig
       eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.28578978              28.578978                           28.57898
dim 2 0.15717340              15.717340                           44.29632
dim 3 0.12800433              12.800433                           57.09675
dim 4 0.11057272              11.057272                           68.15402
dim 5 0.09909884               9.909884                           78.06391
dim 6 0.08435717               8.435717                           86.49962
dim 7 0.06186478               6.186478                           92.68610
dim 8 0.04718425               4.718425                           97.40453
dim 9 0.02595473               2.595473                          100.00000
```

6. **Question** : Graphique des valeurs propres .

Pour tracer le graphe d'évolution des valeurs propres on utilise la commande :

```
> plot(res$eig[,1],type="b",main="Scree plot")
```



7. **Question** : La dimension de sous espace .

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

```
> var(res$eig[4:9,1])*5/(var(res$eig[,1])*9)
[1] 0.09744785
```

Ce qui est un pourcentage très supérieur à 5 % .

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

```
 var(res$eig[5:9,1])*4/(var(res$eig[,1])*9)
1] 0.06320546
```

Ce qui est un pourcentage supérieur à 5% .
On appliquons la règles du rapport : la variance des 4 dernières valeurs propre par rapport à la variance de tout les valeurs propre on trouver le résultat suivant :

```
> var(res$eig[6:9,1])*3/(var(res$eig[,1])*9)
[1] 0.03390725
```

Ce qui est un pourcentage inférieur à 5% .
Donc la dimension du sous éspace est 5 .

Partie 2 :

8. **Question** : Cos^2 des modalité .

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 .
Les valeurs du cos^2 sur la dimension 5 du sous espace sont présentées dans la colonne dim5 .

```
> res$var$cos2
```

	Dim 4	Dim 5	Dim 6
turda_0	0.005477834	2.435232e-01	0.2803628398
turda_1	0.005477834	2.435232e-01	0.2803628398
X2_house_age_0	0.105493707	7.680910e-02	0.1008093868
X2_house_age_1	0.105493707	7.680910e-02	0.1008093868
X3_distance_to_the_nearest_MRT_station_0	0.001853868	1.951498e-02	0.0002151653
X3_distance_to_the_nearest_MRT_station_1	0.001853868	1.951498e-02	0.0002151653
X4_number_of_convenience_stores_0	0.010881055	1.076019e-06	0.0002333078
X4_number_of_convenience_stores_1	0.010881055	1.076019e-06	0.0002333078
X5_latitude_0	0.190492235	6.750058e-02	0.0050140905
X5_latitude_1	0.190492235	6.750058e-02	0.0050140905
X6_longitude_0	0.005109895	3.387647e-04	0.0245370237
X6_longitude_1	0.005109895	3.387647e-04	0.0245370237
dephm_0	0.002421224	8.200540e-02	0.2386228227
dephm_1	0.002421224	8.200540e-02	0.2386228227
SOLID_0	0.038992657	3.962376e-01	0.1030802227
SOLID_1	0.038992657	3.962376e-01	0.1030802227
Condi_0	0.634432047	5.958855e-03	0.0063396548
Condi_1	0.634432047	5.958855e-03	0.0063396548

9. **Question** : Distinction des modalités .

Les modalités turda_0 ,turda_1 ,Solid_0,Solid_1 sont bien représenté alors que les autres variables sont faiblement représentées .

Les modalités X2_house_age_0 ,X2_house_age1,X3_0,X3_1 dephm_0,dephm_1 sont moyennemnt représentées .

Les autre modalités X4_0,X4_1,X6_0,X6_1 sont faiblement représentées .

10. **Question** : La contribution des modalités .

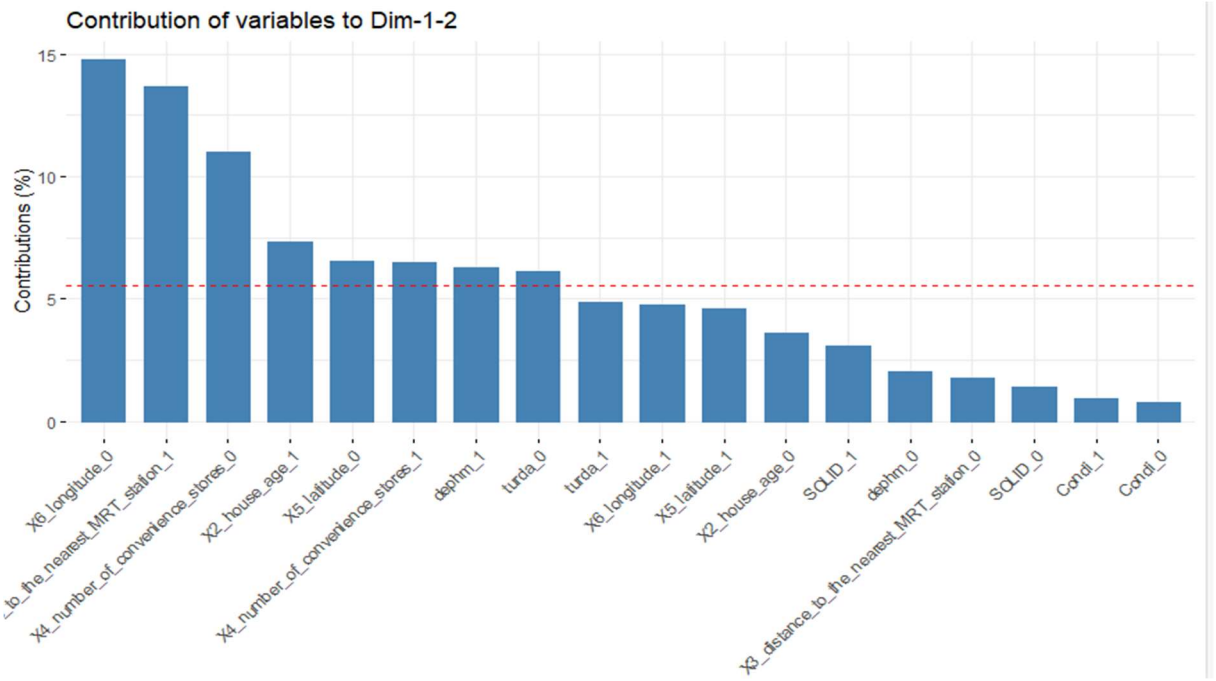
Afin de calculer la contribution des modalités dans chaque sous espace on utilise cette commande :

```
> res$var$contrib
```

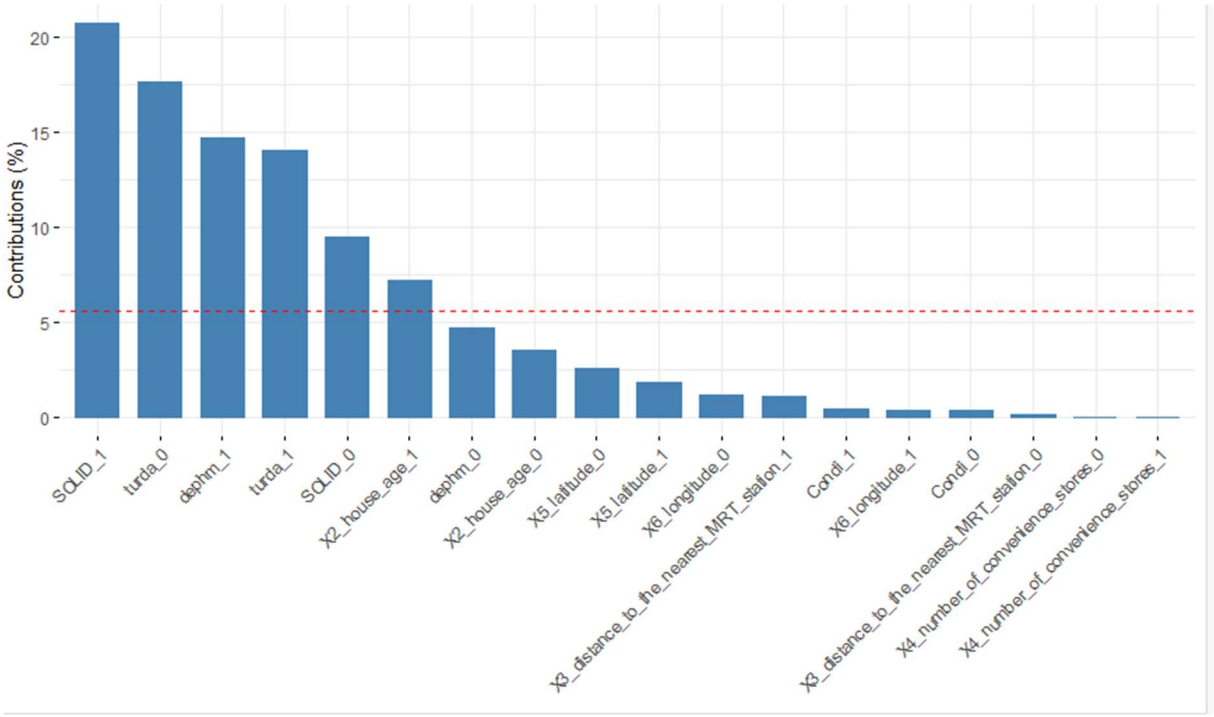
	Dim 4	Dim 5	Dim 6
turda_0	0.30667960	1.521233e+01	20.574179063
turda_1	0.24377096	1.209185e+01	16.353834640
X2_house_age_0	3.48309909	2.829641e+00	4.362809670
X2_house_age_1	7.11763727	5.782311e+00	8.915306717
X3_distance_to_the_nearest_MRT_station_0	0.02129023	2.500628e-01	0.003238915
X3_distance_to_the_nearest_MRT_station_1	0.16499925	1.937986e+00	0.025101594
X4_number_of_convenience_stores_0	0.68728225	7.583391e-05	0.019316098
X4_number_of_convenience_stores_1	0.40612133	4.481095e-05	0.011414058
X5_latitude_0	11.21172855	4.432842e+00	0.386824063
X5_latitude_1	7.93024702	3.135425e+00	0.273607264
X6_longitude_0	0.38877582	2.875841e-02	2.447006995
X6_longitude_1	0.12470168	9.224397e-03	0.784889036
dephm_0	0.05908745	2.232967e+00	7.633054409
dephm_1	0.18421382	6.961603e+00	23.797169628
SOLID_0	1.23145047	1.396270e+01	4.267126195
SOLID_1	2.68680103	3.046406e+01	9.310093517
Condi_0	28.23307913	2.958798e-01	0.369798176
Condi_1	35.51903504	3.722359e-01	0.465229963

La visualtisation de la ontribution dans le premier axe :

```
> contrib <- res$var$contrib
> fviz_contrib(res,choice="var",axes=1:2,top=36)
```



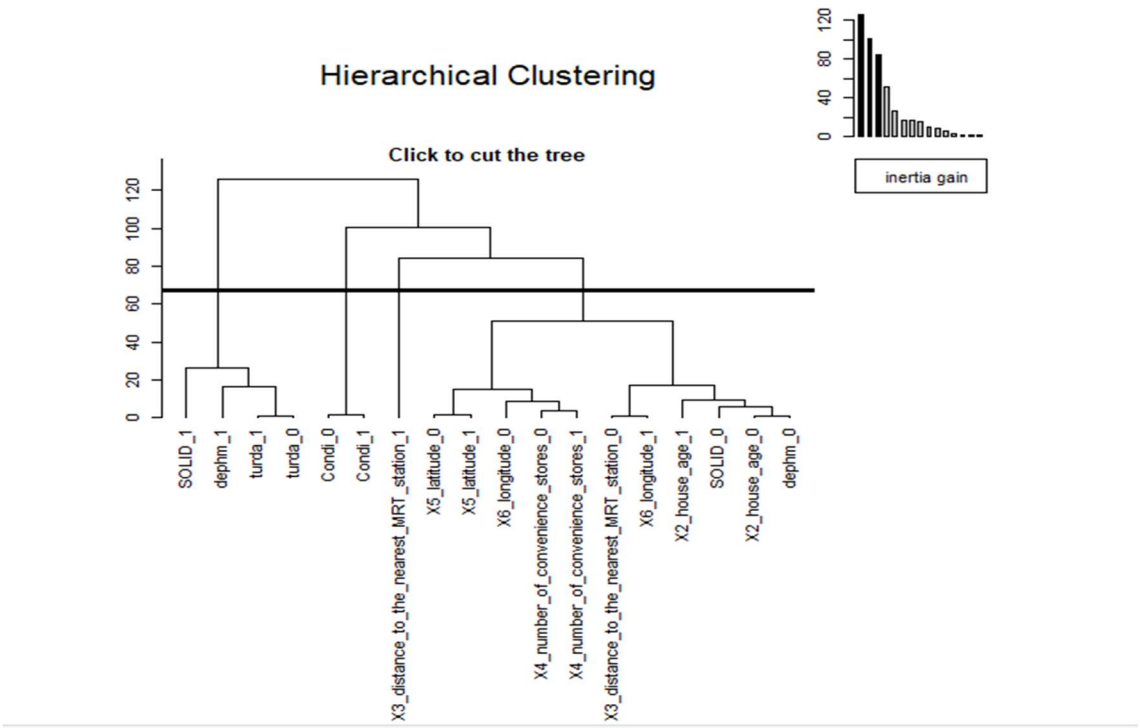
La visualisation de la contribution dans sous espace Dim 5 :



11. **Question** : Application de la CAH .

La méthode de la CAH permet de classifier les modalités selon leur ressemblance, et grâce au critère de coupe elle nous donne le nombre de classes à garder. On applique le CAH au tableau des contributions des modalités, on trouve le dendrogramme suivant avec la ligne de découpage

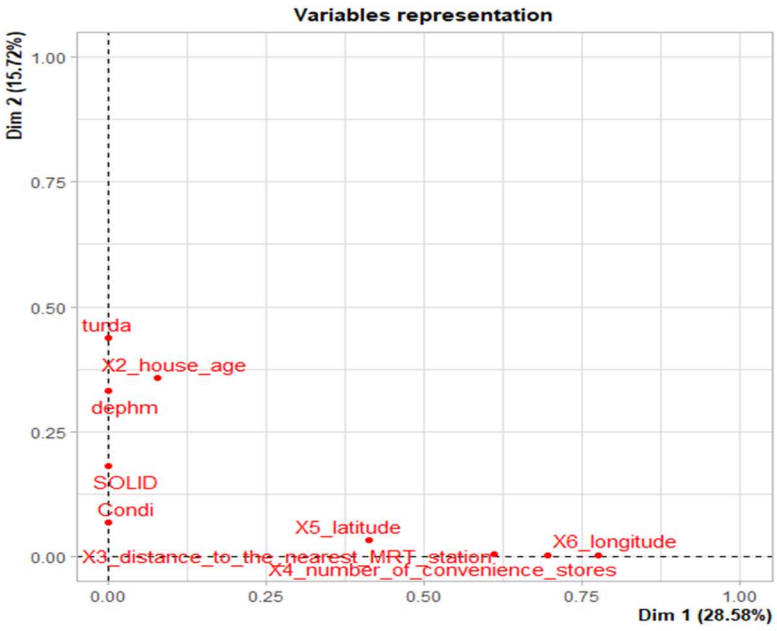
```
> HCPC(res$var$contrib)
[1] "click on the graph to cut the tree"
```



On voit d'après la figure que le CAH découpe l'ensemble des modalités en 4 classe , l'une est très grande par rapport au autres, c'est à dire qu'une grande partie des individus sont caractérisés par ces modalités et partagent les mêmes ca ractéristiques.Les modalités qui constituent des petites classes sont les moins contri butifs et peu d'individus qui possèdent leurs caractéristiques.

12. **Question** : Le nuage Des modalités .

```
> plot(res,choix="var")
```



Partie 3 :

13. **Question** : Cos^2 des individus :

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        Dim 6        Dim 7        Dim 8
0  0.138316599 1.767050e-01 6.326401e-01 0.0001308937 0.0003162733 1.141784e-04 0.0005938568 0.0149422331
1  0.359295009 4.730005e-01 2.895022e-04 0.1244613703 0.0092754865 1.671730e-02 0.0145051645 0.0021972566
2  0.380581043 3.994597e-02 1.467572e-02 0.1933344497 0.1514088709 1.886867e-01 0.0241223576 0.0029004282
3  0.380581043 3.994597e-02 1.467572e-02 0.1933344497 0.1514088709 1.886867e-01 0.0241223576 0.0029004282
4  0.359295009 4.730005e-01 2.895022e-04 0.1244613703 0.0092754865 1.671730e-02 0.0145051645 0.0021972566
5  0.342097161 3.592743e-01 6.458997e-02 0.0086727472 0.0153065956 5.527713e-04 0.1051295146 0.0396199257
6  0.138316599 1.767050e-01 6.326401e-01 0.0001308937 0.0003162733 1.141784e-04 0.0005938568 0.0149422331
7  0.331663639 2.218360e-01 1.470417e-01 0.1195006730 0.0262923404 4.079444e-02 0.0332171948 0.0727667254
8  0.715427275 1.446370e-02 1.190161e-01 0.0123777250 0.0391442438 6.104806e-02 0.0309689087 0.0065697438
9  0.369473139 6.022082e-02 2.629801e-02 0.0209249875 0.2241141237 1.436429e-01 0.0972228011 0.0171954358
10 0.001257156 1.977748e-02 9.479019e-02 0.3758417236 0.0921930414 1.575485e-01 0.2026308965 0.0271891820
11 0.268042223 8.942672e-02 1.761729e-01 0.0199280246 0.4020634100 3.718917e-02 0.0043035126 0.0002914987
12 0.049941359 5.938405e-01 5.202298e-02 0.0015252226 0.0202070536 3.601481e-02 0.0347354533 0.2110381241
13 0.064385503 4.443649e-01 8.186488e-02 0.0006258631 0.0174778124 1.583762e-04 0.0001844942 0.0248568146
```

14. **Question** : Distinction des individus .

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 .

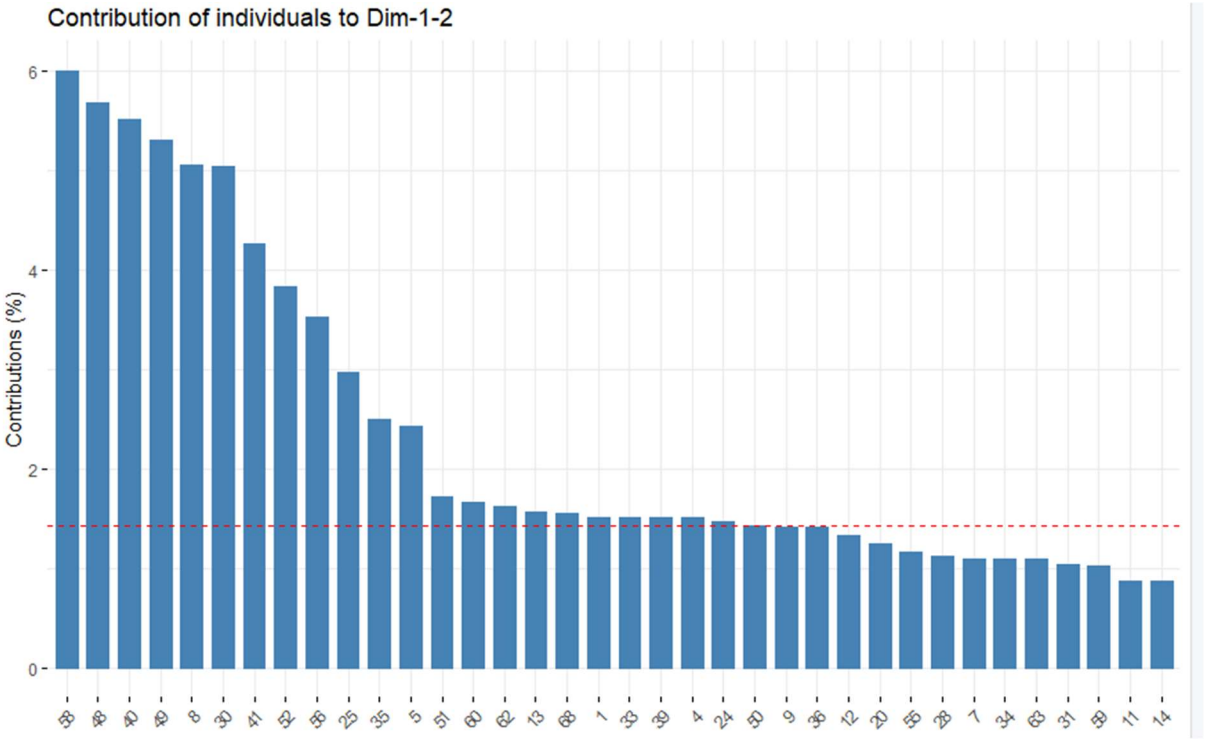
15. **Question** : Calcul des contributions .

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        Dim 5        Dim 6        Dim 7        Dim 8
0  0.508833428 1.1820009895 5.196130448 0.001244568 0.003355382 0.0014230161 0.010092211  0.332940404
1  1.011627055 2.4215779587 0.001819882 0.905737835 0.075315451 0.1594630290 0.188666647  0.037471380
2  0.973650997 0.1858218462 0.083825786 1.278392130 1.117083450 1.6353931984 0.285088130  0.044943609
3  0.973650997 0.1858218462 0.083825786 1.278392130 1.117083450 1.6353931984 0.285088130  0.044943609
4  1.011627055 2.4215779587 0.001819882 0.905737835 0.075315451 0.1594630290 0.188666647  0.037471380
5  1.838128091 3.5101051746 0.774842423 0.120443029 0.237182721 0.0100622815 2.609481273  1.289405804
6  0.508833428 1.1820009895 5.196130448 0.001244568 0.003355382 0.0014230161 0.010092211  0.332940404
7  1.019152700 1.2394841670 1.008797401 0.949096625 0.232996169 0.4246851601 0.471528094  1.354328104
8  7.673888266 0.2820960914 2.850215147 0.343154334 1.210865823 2.2184341253 1.534543142  0.426823255
9  1.890171280 0.5601864419 0.300374630 0.276682810 3.306478112 2.4895860932 2.297680520  0.532820761
10 0.005393808 0.1542925730 0.908012517 4.167831035 1.140729600 2.2900524461 4.016199904  0.706565187
11 1.011277098 0.6134820746 1.483980297 0.194325544 4.374613089 0.4753440936 0.075005526  0.006661211
12 0.160210479 3.4639211877 0.372604270 0.012646271 0.186944069 0.3914140112 0.514762328  4.100540841
13 0.306564527 3.8471686293 0.870268919 0.007702147 0.239993245 0.0025547500 0.004058072  0.716851080
```

On visualise La contribution des individus par exemple sur Dim1 :

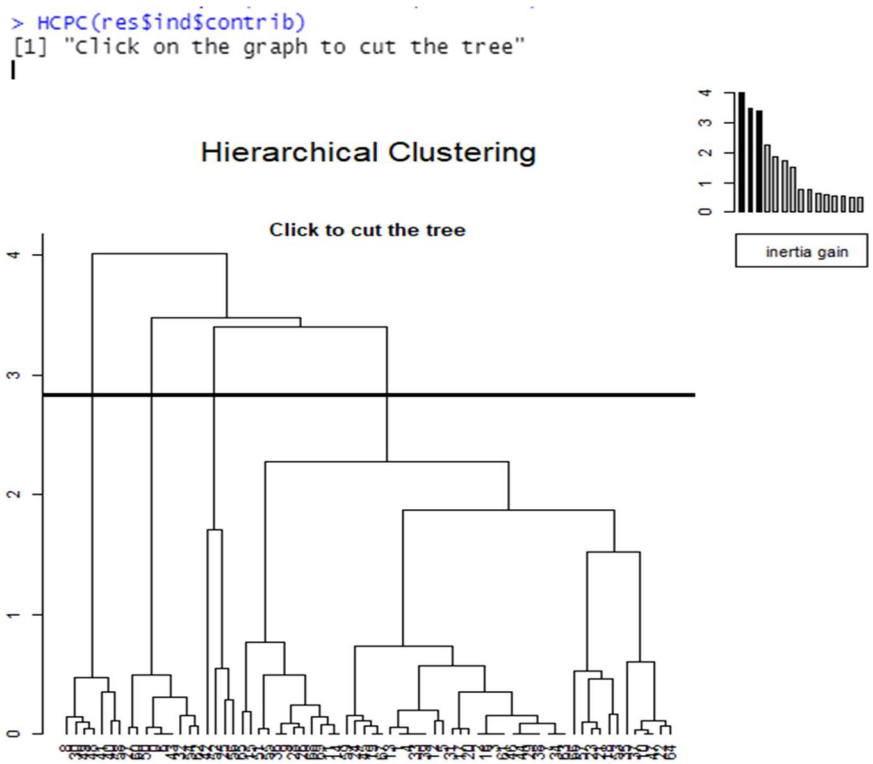
```
> contrib <- res$var$ind
> fviz_contrib(res,choice="ind",axes=1:2,top=36)
```



Dans cette figure on voit que l'individus 58 est plus contributif sur ce sous espace .

16. **Question** : Application de la CAH .

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



Partie 4 :

17. **Question** : Calcul des coefficient de corrélation des variables .

Le graphique ci-dessous permet d'identifier les variables les plus corrélées avec chaque axe. La commande qui permet de calculer les coefficients de corréelation des variables avec les projections sur les axes du sous espace est la sui vante :resvareta2 avec facto représente la sortie de la fonction MCA. La sortie de la commande nous donne le tableau suivant :

```
> res$var$eta2
                                          Dim 1       Dim 2       Dim 3       Dim 4        Dim 5        Dim 6
turda                                 4.150638e-04 0.437453195 0.025294589 0.005477834 2.435232e-01 0.2803628398
X2_house_age                          7.847363e-02 0.357396276 0.182228999 0.105493707 7.680910e-02 0.1008093868
X3_distance_to_the_nearest_MRT_station 6.102758e-01 0.004385361 0.031475723 0.001853868 1.951498e-02 0.0002151653
X4_number_of_convenience_stores       6.955942e-01 0.002526246 0.000600815 0.010881055 1.076019e-06 0.0002333078
X5_latitude                           4.120395e-01 0.031621423 0.072667003 0.190492235 6.750058e-02 0.0050140905
X6_longitude                          7.749121e-01 0.002588898 0.017763832 0.005109895 3.387647e-04 0.0245370237
dephm                                 3.815844e-04 0.330500124 0.294033305 0.002421224 8.200540e-02 0.2386228227
SOLID                                 3.021137e-08 0.180320694 0.276672270 0.038992657 3.962376e-01 0.1030802227
Condi                                 1.593319e-05 0.067768423 0.251302469 0.634432047 5.958855e-03 0.0063396548
```

Conclusion :

Dans ce chapitre, on a pu appliquer l'analyse en composante multiple appliqué à notre table de données. On a commencé tout d'abord par l'application des k_means pour déterminer les variables catégorielles qui correspondent à nos variables quantitatives, le nombre de modalités a été fixé sur deux. Après, on a passé à l'étude des modalités et leurs caractéristiques en construisant le tableau disjonctif, et en calculant les valeurs propres et d'autres valeurs descriptifs appliqué aux modalités.