Classification des dépenses des ménages

Zemzmi Chaima

14/01/2021

Contents

Présentation du projet

L’objectif du projet est d’identifier les groupes des ménages tunisiens similaires en matière de dépenses, à l’aide de la base des données “DEPENSES_MENAGES”. Pour le faire on va appliquer et analyser les différentes méthodes de classification non supervisée afin de pouvoir choisir la meilleure classification.

Statistique descriptive

Description de la base des données ”“DEPENSES_MENAGES”

```
str(DEPENSES_MENAGES)
```

```
## tibble [11,281 x 28] (S3: tbl_df/tbl/data.frame)
##  $ ID            : num [1:11281] 788 1747 1751 705 738 ...
##  $ Taille        : num [1:11281] 6 4 4 7 4 4 6 5 7 3 ...
##  $ Extrp_Ind     : num [1:11281] 1443 908 908 1683 962 ...
##  $ Extrp_Mng     : num [1:11281] 240 227 227 240 240 228 220 313 313 227 ...
##  $ Seuil_PvrtBas : num [1:11281] 756962 570873 570873 756962 756962 ...
```

```
##  $ Seuil_PvrtHaut: num [1:11281] 1276806 820413 820413 1276806 1276806 ...
##  $ Region        : chr [1:11281] "Grand Tunis" "Grand Tunis" "Grand Tunis" "Grand Tunis" ...
##  $ Dep_Ind       : num [1:11281] 445989 526384 533623 534806 608661 ...
##  $ Csp_Prcpl     : num [1:11281] 6 8 8 6 6 5 6 4 9 6 ...
##  $ Taille_Cat    : chr [1:11281] "5 à 6 Personnes" "3 à 4 Personnes" "3 à 4 Personnes" "7 à 8 Personn
##  $ Milieu        : num [1:11281] 1 2 2 1 1 1 1 1 1 1 ...
##  $ Strate        : num [1:11281] 1 3 3 1 1 1 1 2 1 1 ...
##  $ Decile_Dep    : chr [1:11281] "Décile 1" "Décile 1" "Décile 1" "Décile 1" ...
##  $ Tranche_Dps   : chr [1:11281] "Mois de 500 DT" "500 à 750 DT" "500 à 750 DT" "500 à 750 DT" ...
##  $ Extr_Poor     : num [1:11281] 1 1 1 1 1 1 1 1 1 1 ...
##  $ Poor          : num [1:11281] 1 1 1 1 1 1 1 1 1 1 ...
##  $ Dep_Almtr     : num [1:11281] 109342 276887 253070 236535 156988 ...
##  $ Dep_TBA       : num [1:11281] 10400 0 0 49771 0 ...
##  $ Dep_Hab       : num [1:11281] 0 6500 6500 0 10000 ...
##  $ Dep_LogEner   : num [1:11281] 278303 201428 250353 225079 330600 ...
##  $ Dep_MblArtc   : num [1:11281] 10443 19012 12025 4950 10250 ...
##  $ Dep_HgSn      : num [1:11281] 0 21206 10325 3064 5525 ...
##  $ Dep_Trsp      : num [1:11281] 0 0 0 0 24050 ...
##  $ Dep_Telc      : num [1:11281] 0 0 0 714 0 ...
##  $ Dep_LC        : num [1:11281] 0 0 0 1320 0 0 0 0 857 0 ...
##  $ Dep_Ensgm     : num [1:11281] 37500 0 0 0 69897 ...
##  $ Dep_HR        : num [1:11281] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Dep_Atr       : num [1:11281] 0 1350 1350 0 1350 ...
```

D'après le résultat de la commande "str" on remarque que notre base est formée de 11281 individus et 28 variables dont 19 variables quantitatives (12 entre eux expriment les différentes dépenses), 2 variables qualitatives nominales ("Poor et Extr_Poor), 4 variables qualitatives ordinales ("ID, Milieu, strate et Csp Prcpl) et 3 variables qualitatives.

Nettoyage et prépation de la base

On va garder les 12 variables qui expriment les dépenses, puis on va regrouper les individus selon les variables qui on une influences sur les dépenses afin de réduire la dimension de la base d'étude.

Choix des variables de regroupement

```
#Création de la variable Dep_Moy
DEPENSES_MENAGES$Dep_Moy=rowMeans(DEPENSES_MENAGES[,17:28], na.rm=TRUE)
```

- Moyenne des dépenses selon la variable milieu

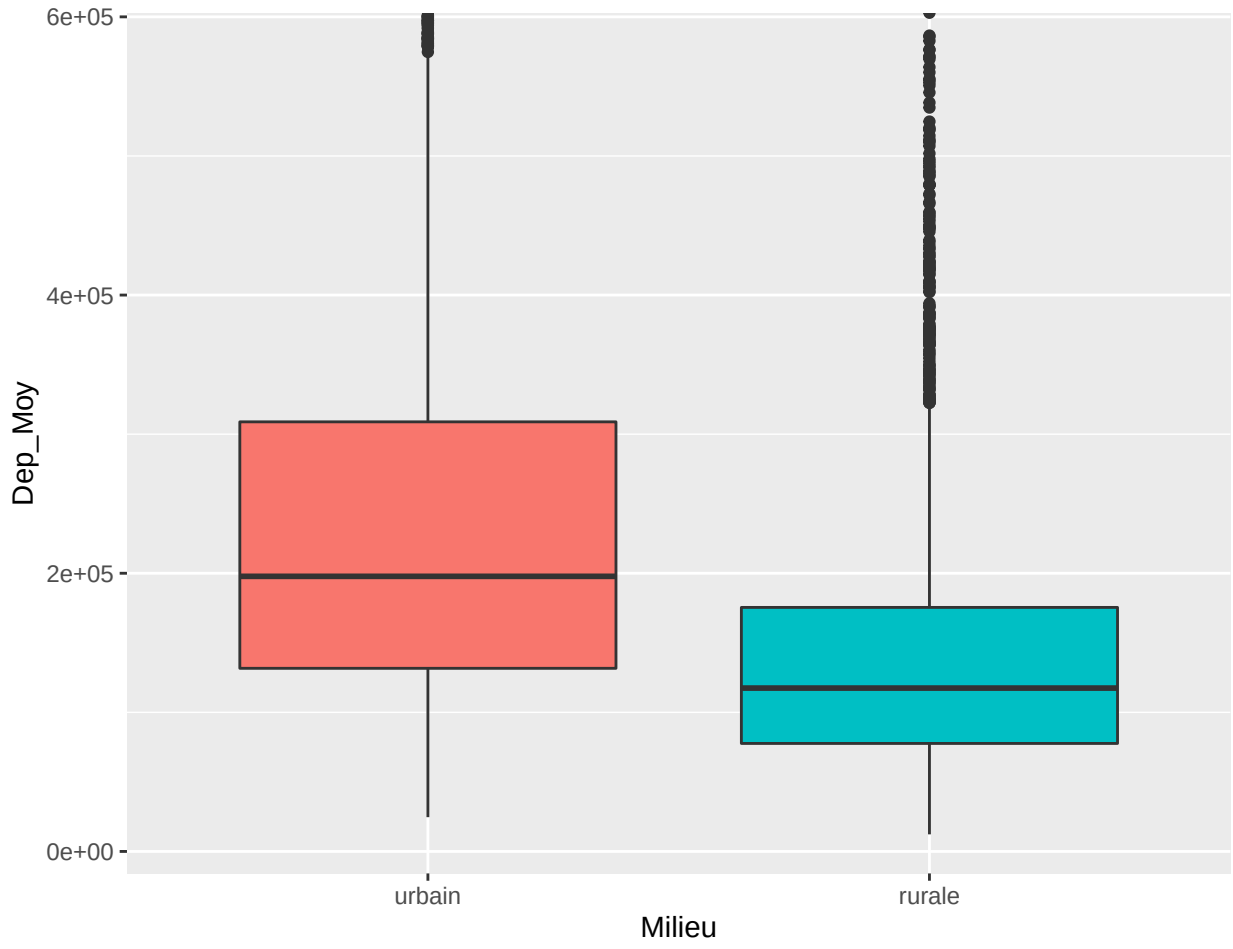
```
#recodage de la varaible milieu
library(labelled)
```

```
## Warning: package 'labelled' was built under R version 4.0.3
```

```
DEPENSES_MENAGES$Milieu=to_factor(labelled(DEPENSES_MENAGES$Milieu,
                                  c("urbain" = 1, "rurale" = 2)))

#Boxplot de Dep_Moy selon le Milieu
library(ggplot2)
ggplot(DEPENSES_MENAGES, aes(x=Milieu, y=Dep_Moy, fill=Milieu))+
geom_boxplot()+
coord_cartesian(ylim =range(boxplot(DEPENSES_MENAGES$Dep_Moy~DEPENSES_MENAGES$Milieu,
                                    plot=FALSE)$stats))+
xlab("Milieu") +
```

```
theme(legend.position="none") +
ylab("Dep_Moy")
```

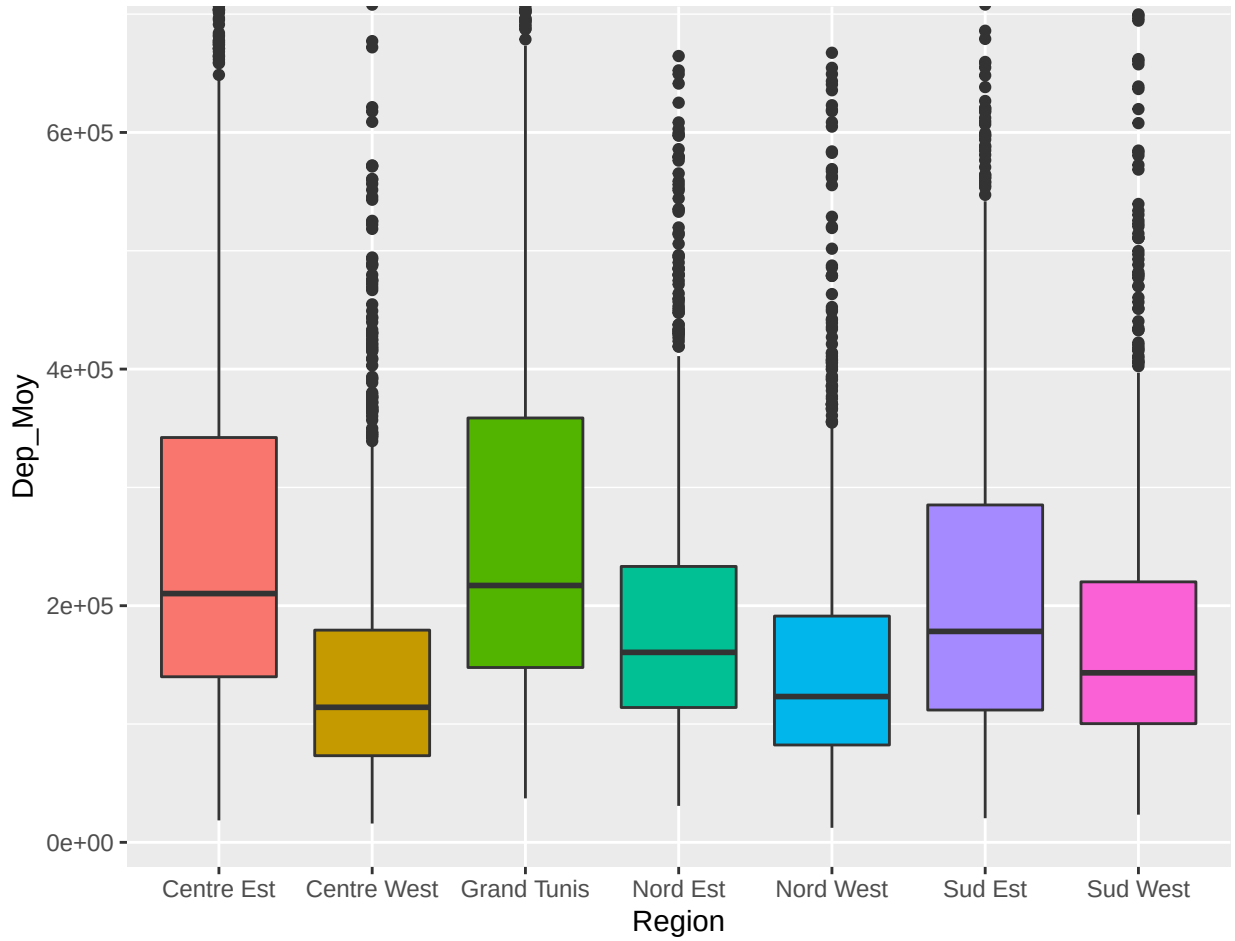


- Moyenne des dépenses selon la variable Région

```
DEPENSES_MENAGES$Region=as.factor(DEPENSES_MENAGES$Region)
levels(DEPENSES_MENAGES$Region)
```

```
## [1] "Centre Est"  "Centre West" "Grand Tunis" "Nord Est"    "Nord West"
## [6] "Sud Est"     "Sud West"
```

```
#Boxplot de Dep_Moy selon la Région
library(ggplot2)
ggplot(DEPENSES_MENAGES, aes(x=Region, y=Dep_Moy, fill=Region)) +
geom_boxplot() +
coord_cartesian(ylim =range(boxplot(DEPENSES_MENAGES$Dep_Moy~DEPENSES_MENAGES$Region,
                                    plot=FALSE)$stats))+
xlab("Region") +
theme(legend.position="none") +
ylab("Dep_Moy")
```



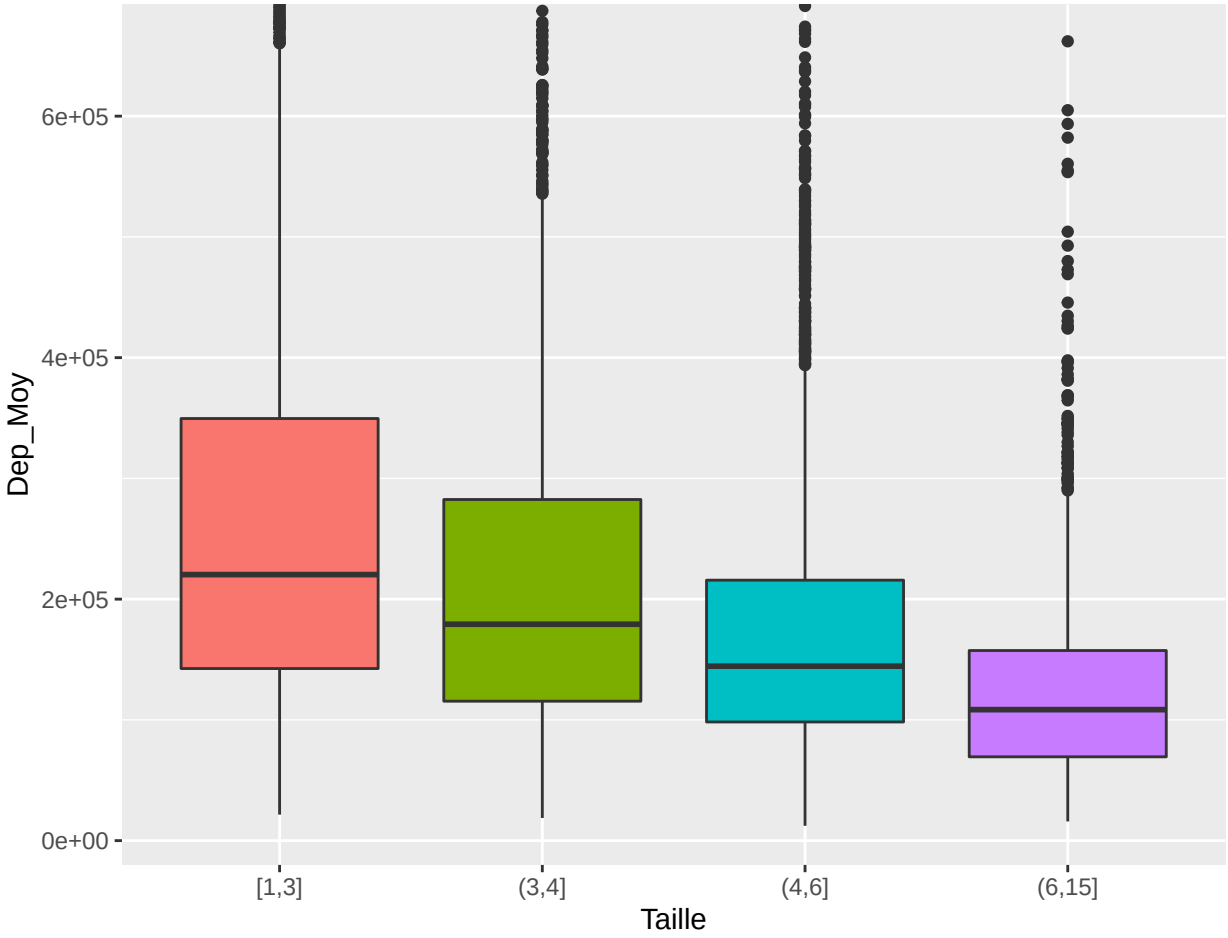
- Moyenne des dépenses selon la variable Taille

```
#Regroupement de la varaible taille selon les 4 quartiles
DEPENSES_MENAGES$Taille=as.factor(cut(DEPENSES_MENAGES$Taille, breaks=quantile(DEPENSES_MENAGES$Taille,
c(0, 0.25, 0.5, 0.75, 1)), include.lowest = TRUE))

levels(DEPENSES_MENAGES$Taille)
```

```
## [1] "[1,3]"  "(3,4]"  "(4,6]"  "(6,15]"
```

```
#Boxplot de Dep_Moy selon la Taille
ggplot(DEPENSES_MENAGES, aes(x=Taille, y=Dep_Moy, fill=Taille)) +
geom_boxplot() +
coord_cartesian(ylim =range(boxplot(DEPENSES_MENAGES$Dep_Moy~DEPENSES_MENAGES$Taille,
                                    plot=FALSE)$stats))+
xlab("Taille") +
theme(legend.position="none") +
ylab("Dep_Moy")
```



Conclusion D'aprés les boxplots on peut conclure que les variables "Milieu, Region et Taille" influent la variables moyenne des dépenses donc on va les retenir pour le regrepement des individus.

Regrepement des individus

```
library(doBy)
```

```
## Warning: package 'doBy' was built under R version 4.0.3
```

```
DEPENSES_MENAGES=DEPENSES_MENAGES[,-c(1,3,4,5,6,8,9,10,12,13,14,15,16,29)]
DEPENSES_MENAGES=summaryBy( .~Region+Milieu+Taille,data=DEPENSES_MENAGES,FUN=c(mean),keep.names = TRUE )
names=paste(DEPENSES_MENAGES$Region,DEPENSES_MENAGES$Milieu,DEPENSES_MENAGES$Taille)
DEPENSES_MENAGES=DEPENSES_MENAGES[,-c(1,2,3)]
rownames(DEPENSES_MENAGES)=names
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
# dim de la nouvelle base
dim(DEPENSES_MENAGES)
```

```
## [1] 56 12
```

La base finale est constituée de 12 variables quantitatives exprimant les différentes dépenses et 56 lignes présentant les groupes d'individus selon la variable "Milieu" qui prend deux valeurs (urbaine, rurale), la

variable “taille” qui contient 4 classes (“[1,3]” “(3,4]” “(4,6]” “(6,15]”) et la variable Région qui contient 7 modalités (“Centre Est” “Centre West” “Grand Tunis” “Nord Est” “Nord West” “Sud Est” “Sud West”)

Analyses en composantes principales “ACP”

Dans cette partie, on va appliquer l’ACP afin de regrouper les dépenses corrélées et de réduire le nombre de variables, ainsi que d’analyser les dépenses pour chaque groupe d’individus.

```
library("FactoMineR")
#pca sur des données standardisées
pca <- PCA(DEPENSES_MENAGES, scale.unit = TRUE, graph = FALSE)
```

Choix de nombre d’axes

Critère de Kaiser

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

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

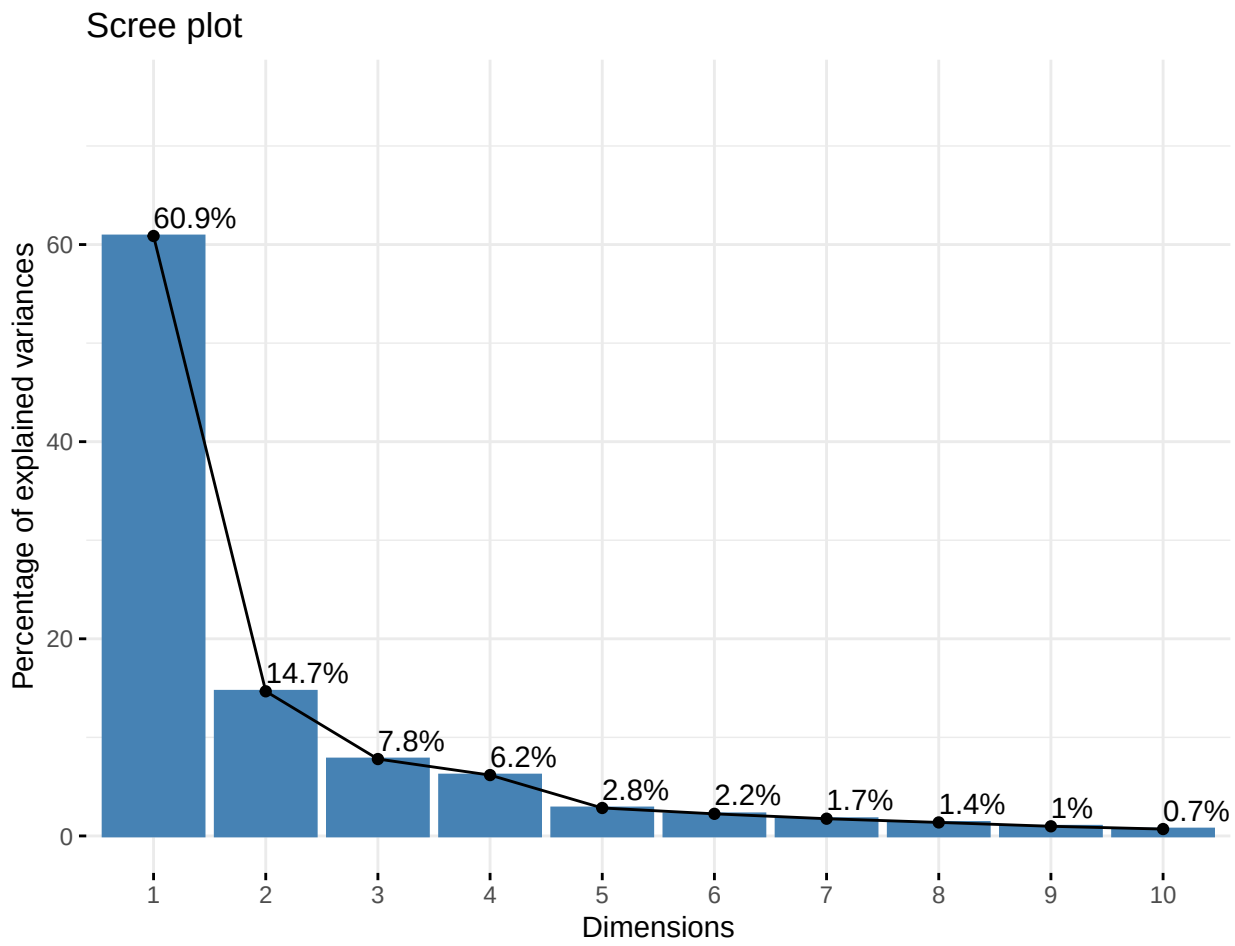
```
eig.val <- get_eigenvalue(pca)
eig.val
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1  7.30421282       60.8684402                    60.86844
## Dim.2  1.76081697       14.6734747                    75.54191
## Dim.3  0.93618450        7.8015375                    83.34345
## Dim.4  0.74098344        6.1748620                    89.51831
## Dim.5  0.34049475        2.8374563                    92.35577
## Dim.6  0.26912081        2.2426735                    94.59844
## Dim.7  0.20972772        1.7477310                    96.34618
## Dim.8  0.16443808        1.3703174                    97.71649
## Dim.9  0.11739422        0.9782852                    98.69478
## Dim.10 0.08375284        0.6979404                    99.39272
## Dim.11 0.04884693        0.4070578                    99.79978
## Dim.12 0.02402690        0.2002242                   100.00000
```

Selon le critère de Kaiser on va retenir les 2 premiers axes qui expliquent 75,54% de la variance totale

Critére de coude

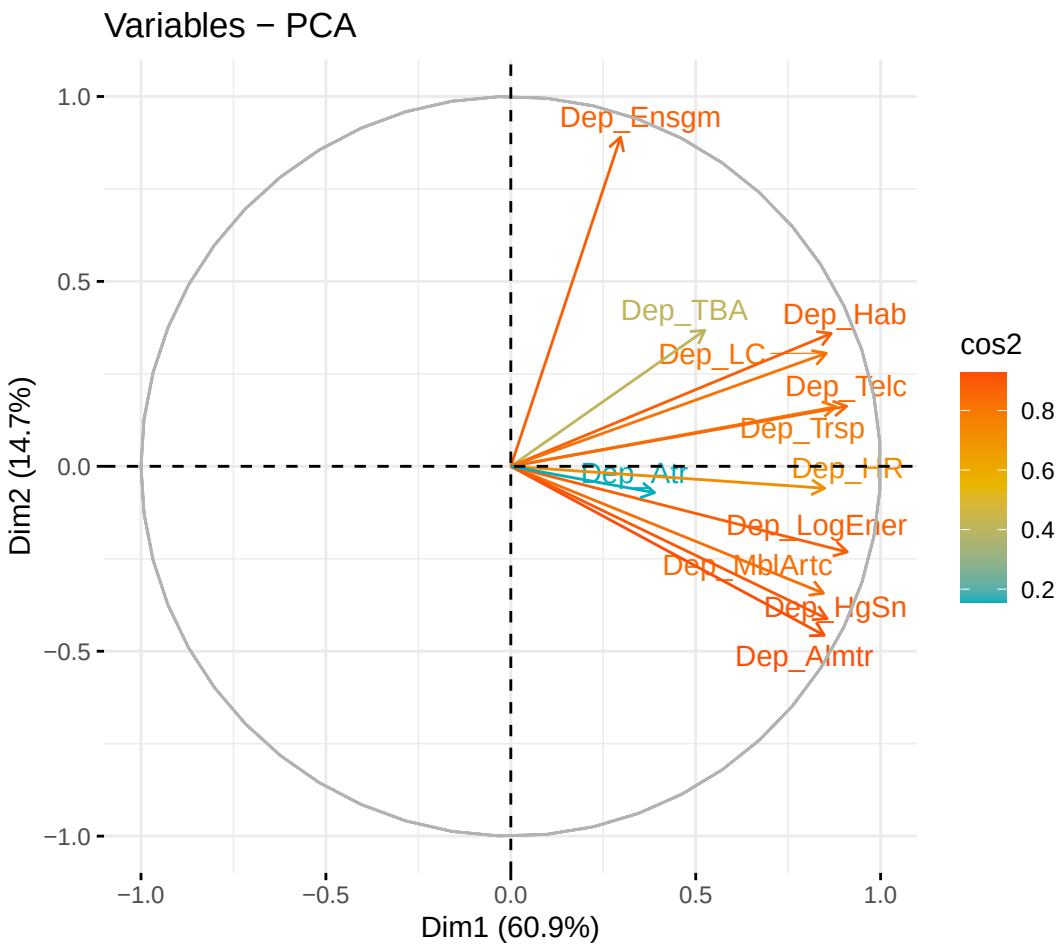
```
fviz_eig(pca, addlabels = TRUE, ylim = c(0,75))
```



Selon le critére de coude, nous conserverons les deux premiers axes.

Cercle de corrélation

```
#Colorer en fonction du cos2: qualité de représentation
fviz_pca_var(pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Évite le chevauchement de texte
             )
```



A partir du cercle de corrélation, on remarque que toutes les variables sont bien représentées par L'ACP à l'exception des variables Dep_Atr (autres dépenses) et Dep_TBA (dépenses en boissons alcoolisées et tabac).

toutes les autres dépenses sont bien représentées par le premier axe à l'exception de la variable Dep_Ensgm qui est representée par le deuxieme axe

Description des dimensions

```
res.desc <- dimdesc(pca, axes = c(1,2), proba = 0.05)
list(res.desc$Dim.1,res.desc$Dim.2)
```

```
## [[1]]
## $quanti
##             correlation      p.value
## Dep_LogEner   0.9098770 2.759659e-22
## Dep_Telc      0.9081482 4.506469e-22
## Dep_Trsp      0.8792170 4.999454e-19
## Dep_Hab       0.8665315 6.260252e-18
## Dep_HgSn      0.8552414 4.820055e-17
## Dep_LC        0.8532493 6.787088e-17
## Dep_HR        0.8486941 1.457085e-16
## Dep_Almtr     0.8476261 1.736631e-16
## Dep_MblArtc   0.8454987 2.453677e-16
## Dep_TBA       0.5247956 3.294119e-05
```

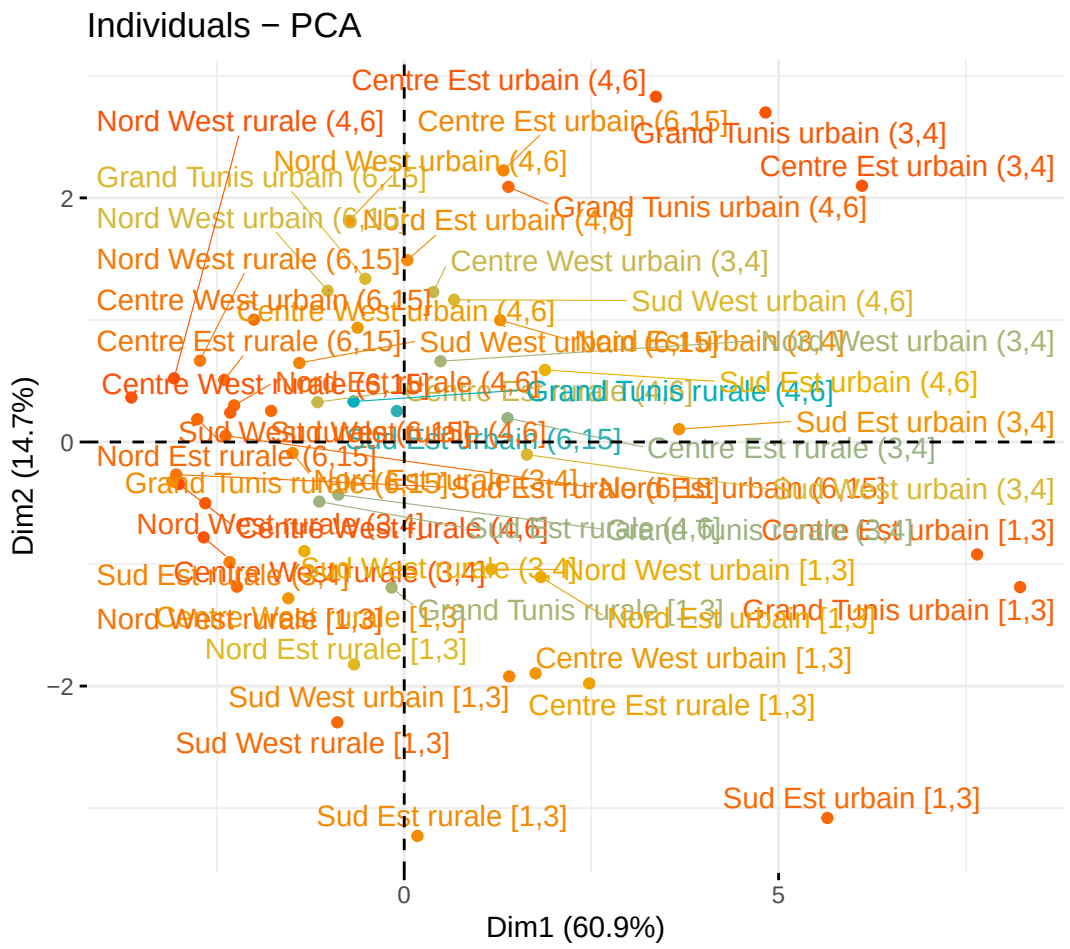
```
## Dep_Atr       0.3886618 3.074206e-03
## Dep_Ensgm     0.2968905 2.628484e-02
##
## attr(,"class")
## [1] "condes" "list "
##
## [[2]]
## $quanti
##             correlation      p.value
## Dep_Ensgm     0.8900213 4.594634e-20
## Dep_TBA       0.3678652 5.282218e-03
## Dep_Hab       0.3592333 6.547221e-03
## Dep_LC        0.3057579 2.192531e-02
## Dep_MblArtc  -0.3431081 9.631616e-03
## Dep_HgSn     -0.4119122 1.609065e-03
## Dep_Almtr    -0.4570043 3.991113e-04
##
## attr(,"class")
## [1] "condes" "list "
```

les variables “Dep_Almtr”,“Dep_Hab”,“Dep_LogEner”,“Dep_MblArtc”,“Dep_HgSn”,“Dep_Trsp”,“Dep_Telc”,“Dep_LC”,“I sont significativement associées avec la premiére composante principale ce qui montre que la premiére composante représente Dépenses pour les besoins quotidiens.

Seule la variable des dépenses d’enseignement est fortement corrélée avec la deuxième composante ce qui montre que la deuxieme composante represente les depenses d’enseignement.

Carte des individus

```
#Coloré en fonction de la qualité et la contribution
fviz_pca_ind (pca, col.ind = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE # Évite le chevauchement de texte
              )
```

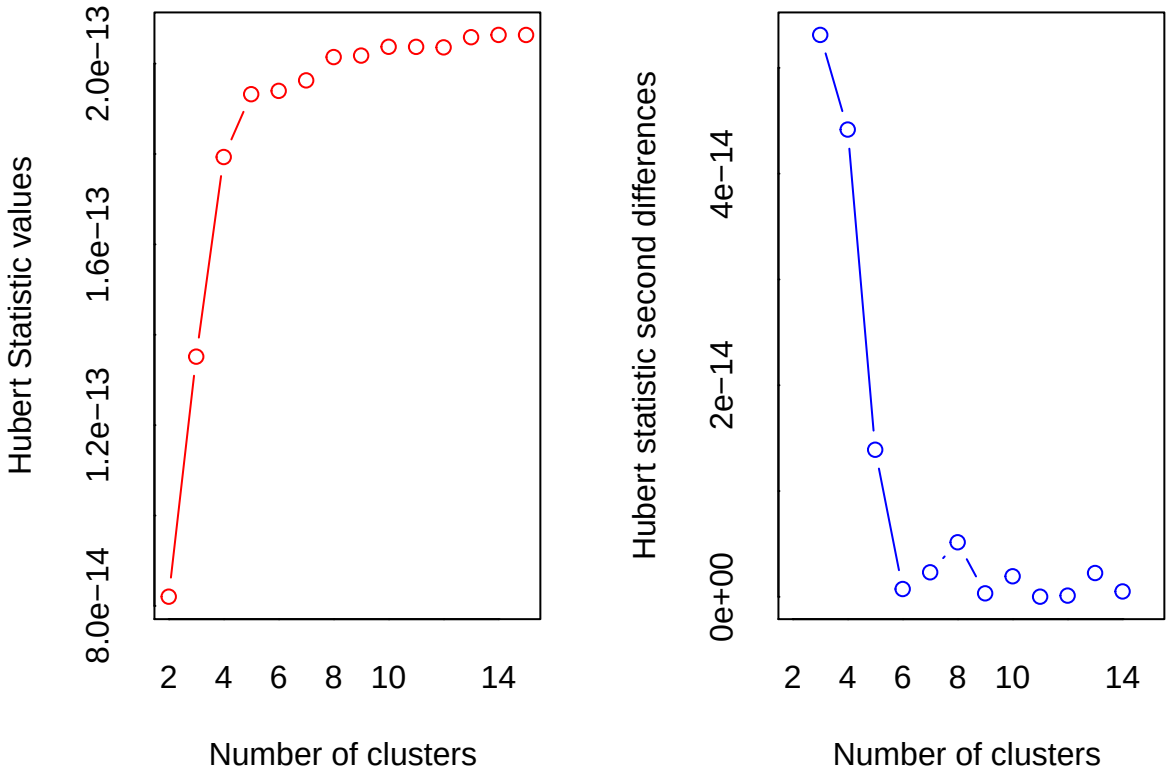


Classification

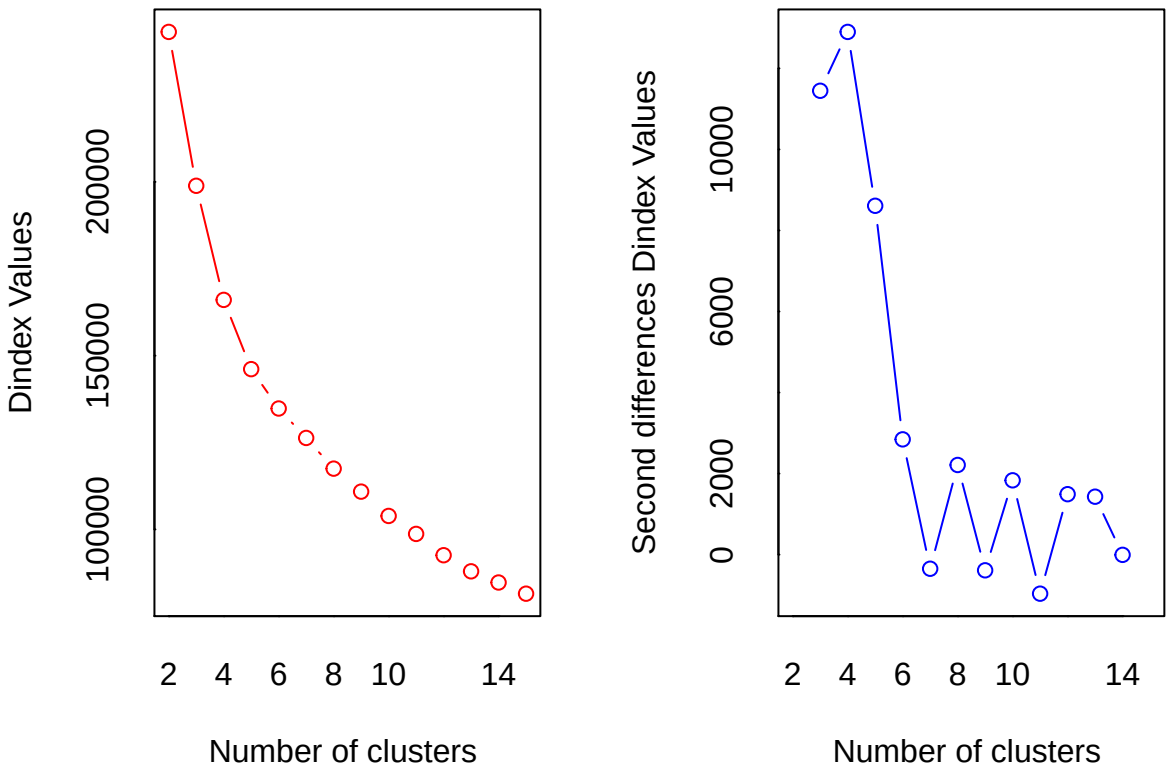
Méthode ward.D

Afin de déterminer le nombre optimal de classes, j'utiliserai le package Nbclust (qui est basé sur les critères d'adéquation)

```
library(NbClust)
Ward=NbClust(DEPENSES_MENAGES,method="ward.D")
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##
```




```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 4 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 9 proposed 4 as the best number of clusters
## * 2 proposed 5 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 1 proposed 10 as the best number of clusters
## * 1 proposed 12 as the best number of clusters
## * 2 proposed 15 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  4
##
##
## *******************************************************************
```

Le nombre des classes donnée par les indices d'adéquation est 4 ** Les classes

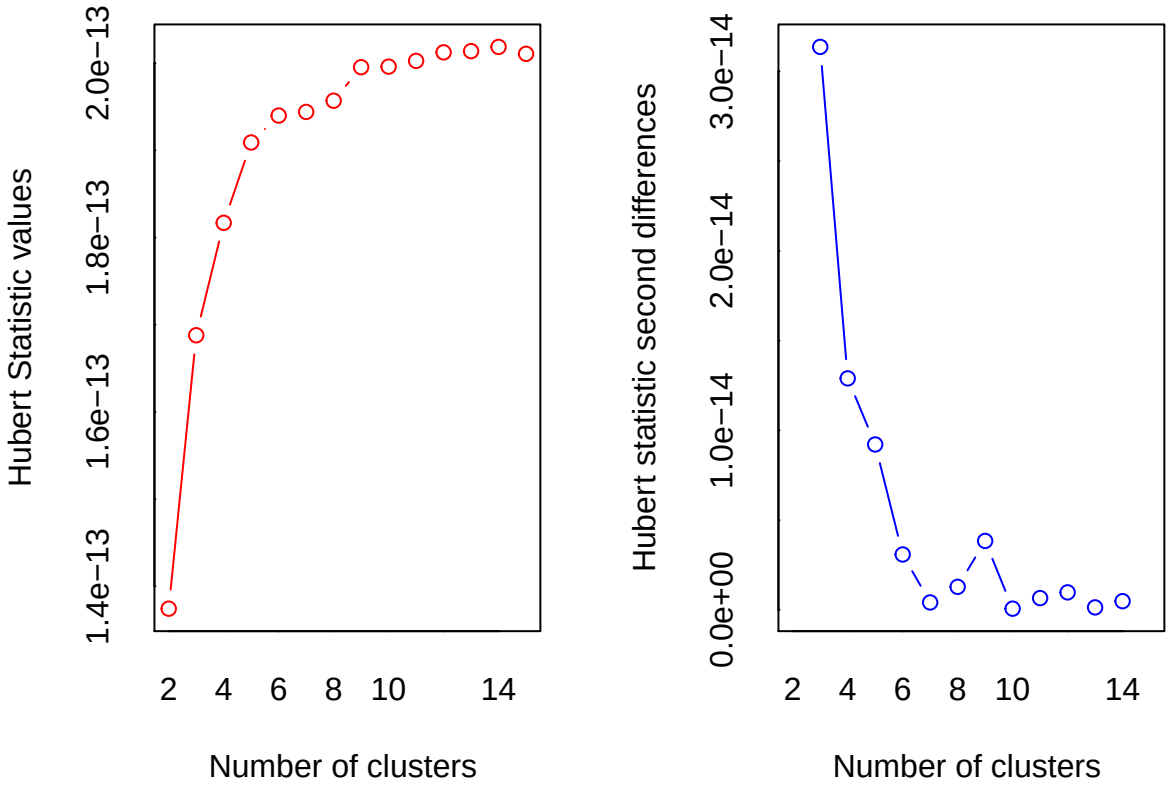
```r
print(sort(Ward$Best.partition))
```

```
##   Centre Est urbain [1,3]  Grand Tunis urbain [1,3]      Sud Est urbain [1,3]
##                         1                         1                         1
##   Centre Est urbain (3,4]   Centre Est rurale [1,3]  Centre West urbain [1,3]
##                         2                         2                         2
##  Grand Tunis urbain (3,4]  Grand Tunis rurale [1,3]     Nord Est urbain [1,3]
##                         2                         2                         2
##    Nord West urbain [1,3]      Sud Est rurale [1,3]     Sud West urbain [1,3]
##                         2                         2                         2
##     Sud West urbain (3,4]   Centre Est urbain (4,6]  Centre Est urbain (6,15]
##                         2                         3                         3
##   Centre Est rurale (3,4]  Centre West urbain (3,4]  Centre West rurale [1,3]
##                         3                         3                         3
##  Grand Tunis urbain (4,6]  Grand Tunis rurale (3,4]     Nord Est urbain (3,4]
##                         3                         3                         3
##     Nord Est urbain (4,6]     Nord Est rurale [1,3]    Nord West urbain (3,4]
##                         3                         3                         3
##    Nord West rurale [1,3]      Sud Est urbain (3,4]      Sud Est urbain (4,6]
##                         3                         3                         3
##     Sud West urbain (4,6]     Sud West rurale [1,3]     Sud West rurale (3,4]
##                         3                         3                         3
##   Centre Est rurale (4,6]  Centre Est rurale (6,15]  Centre West urbain (4,6]
##                         4                         4                         4
## Centre West urbain (6,15]  Centre West rurale (3,4]  Centre West rurale (4,6]
##                         4                         4                         4
## Centre West rurale (6,15] Grand Tunis urbain (6,15]  Grand Tunis rurale (4,6]
##                         4                         4                         4
## Grand Tunis rurale (6,15]    Nord Est urbain (6,15]     Nord Est rurale (3,4]
##                         4                         4                         4
##     Nord Est rurale (4,6]    Nord Est rurale (6,15]    Nord West urbain (4,6]
##                         4                         4                         4
##   Nord West urbain (6,15]    Nord West rurale (3,4]    Nord West rurale (4,6]
##                         4                         4                         4
##   Nord West rurale (6,15]     Sud Est urbain (6,15]      Sud Est rurale (3,4]
##                         4                         4                         4
##      Sud Est rurale (4,6]     Sud Est rurale (6,15]    Sud West urbain (6,15]
##                         4                         4                         4
##     Sud West rurale (4,6]    Sud West rurale (6,15]
##                         4                         4
```

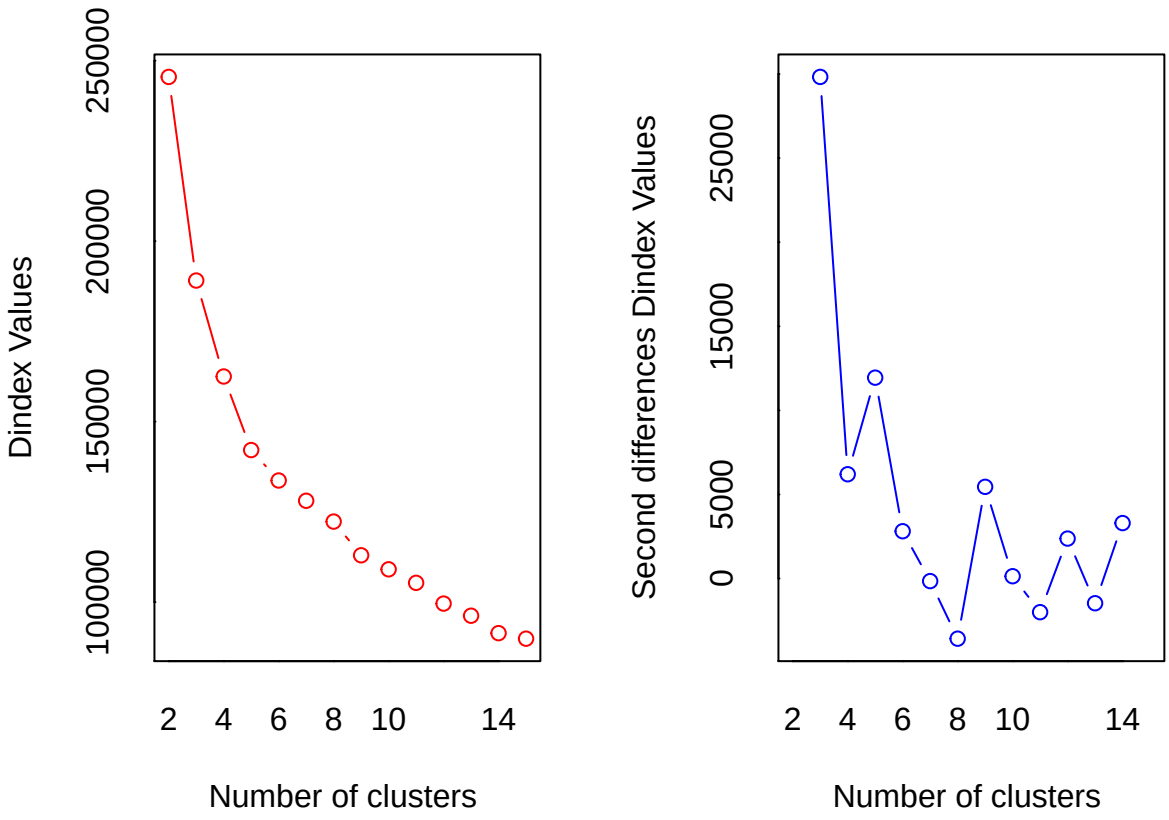
Kmeans

```r
Kmeans=NbClust(DEPENSES_MENAGES,method="kmeans")
```

```
## Warning in pf(beale, pp, df2): production de NaN
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##
```




```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 6 proposed 2 as the best number of clusters
## * 12 proposed 3 as the best number of clusters
## * 1 proposed 4 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 1 proposed 9 as the best number of clusters
## * 2 proposed 14 as the best number of clusters
## * 1 proposed 15 as the best number of clusters
##
##                   ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************
```

le package NbClust propose 3 classes selon la methode Kmeans

```r
print(sort(Kmeans$Best.partition))
```

```
##   Centre Est urbain (4,6]   Centre Est rurale [1,3]   Centre Est rurale (3,4]
##                         1                         1                         1
##  Centre West urbain [1,3]  Centre West urbain (3,4]  Grand Tunis urbain (3,4]
##                         1                         1                         1
##  Grand Tunis urbain (4,6]  Grand Tunis rurale [1,3]  Grand Tunis rurale (3,4]
##                         1                         1                         1
##     Nord Est urbain [1,3]     Nord Est urbain (3,4]     Nord Est rurale [1,3]
##                         1                         1                         1
##    Nord West urbain [1,3]    Nord West urbain (3,4]      Sud Est urbain (3,4]
##                         1                         1                         1
##      Sud Est urbain (4,6]      Sud Est rurale [1,3]     Sud West urbain [1,3]
##                         1                         1                         1
##     Sud West urbain (3,4]     Sud West rurale [1,3]  Centre Est urbain (6,15]
##                         1                         1                         2
##   Centre Est rurale (4,6]  Centre Est rurale (6,15]  Centre West urbain (4,6]
##                         2                         2                         2
## Centre West urbain (6,15]  Centre West rurale [1,3]  Centre West rurale (3,4]
##                         2                         2                         2
##  Centre West rurale (4,6] Centre West rurale (6,15] Grand Tunis urbain (6,15]
##                         2                         2                         2
##  Grand Tunis rurale (4,6] Grand Tunis rurale (6,15]     Nord Est urbain (4,6]
##                         2                         2                         2
##    Nord Est urbain (6,15]     Nord Est rurale (3,4]     Nord Est rurale (4,6]
##                         2                         2                         2
##    Nord Est rurale (6,15]    Nord West urbain (4,6]   Nord West urbain (6,15]
##                         2                         2                         2
##    Nord West rurale [1,3]    Nord West rurale (3,4]    Nord West rurale (4,6]
##                         2                         2                         2
##   Nord West rurale (6,15]     Sud Est urbain (6,15]      Sud Est rurale (3,4]
##                         2                         2                         2
##      Sud Est rurale (4,6]     Sud Est rurale (6,15]     Sud West urbain (4,6]
##                         2                         2                         2
##    Sud West urbain (6,15]     Sud West rurale (3,4]     Sud West rurale (4,6]
##                         2                         2                         2
##    Sud West rurale (6,15]   Centre Est urbain [1,3]   Centre Est urbain (3,4]
##                         2                         3                         3
##  Grand Tunis urbain [1,3]      Sud Est urbain [1,3]
##                         3                         3
```