Projet d'apprentissage non supervisé

Clémence CHESNAIS - Marie GUIBERT

2023-04-20

Environnement de travail

```
library(tidyverse)
library(stargazer)
library(gridExtra)
library(corrplot)
library(cluster)
library(NbClust)
library(ade4)
library(FactoMineR)
library(sf)
```

Nous initialisons une graine aléatoire afin d'obtenir les mêmes résultats à chaque exécution.

```
set.seed(1234)
```

Question 1

Importation des données

```
donnees <- read.csv("Pays_donnees.csv",sep=",",dec=".",stringsAsFactors = T,row.names="pays")
str(donnees)
```

```
## 'data.frame':    167 obs. of  9 variables:
##  $ enfant_mort: num  90.2 16.6 27.3 119 10.3 14.5 18.1 4.8 4.3 39.2 ...
##  $ exports    : num  10 28 38.4 62.3 45.5 18.9 20.8 19.8 51.3 54.3 ...
##  $ sante      : num  7.58 6.55 4.17 2.85 6.03 8.1 4.4 8.73 11 5.88 ...
##  $ imports    : num  44.9 48.6 31.4 42.9 58.9 16 45.3 20.9 47.8 20.7 ...
##  $ revenu     : int  1610 9930 12900 5900 19100 18700 6700 41400 43200 16000 ...
##  $ inflation  : num  9.44 4.49 16.1 22.4 1.44 20.9 7.77 1.16 0.873 13.8 ...
##  $ esper_vie  : num  56.2 76.3 76.5 60.1 76.8 75.8 73.3 82 80.5 69.1 ...
##  $ fert       : num  5.82 1.65 2.89 6.16 2.13 2.37 1.69 1.93 1.44 1.92 ...
##  $ pib_h      : int  553 4090 4460 3530 12200 10300 3220 51900 46900 5840 ...
```

```
# summary(donnees)
```

Dans ce jeu de données, nous pouvons observer 10 variables dont 9 numériques et 1 facteur comprenant les différents pays (individus). Nous avons choisi de transformer la variable pays en facteur pour simplifier nos traitement des données.

Prétraitement des données

Données manquantes

```
sum(is.na(donnees))
```

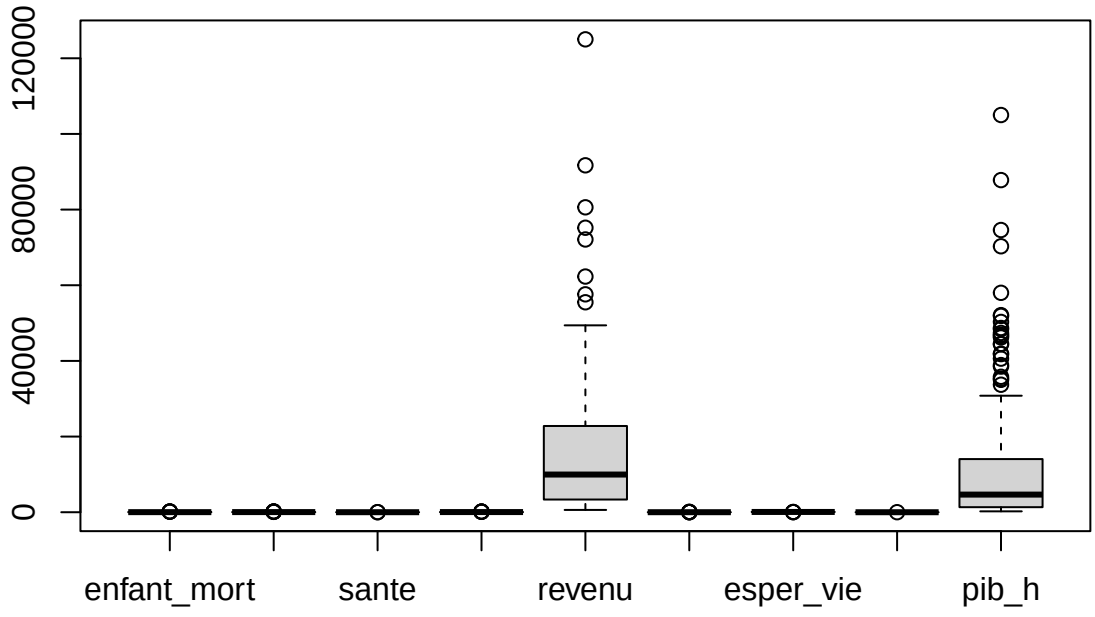
```
## [1] 0
```

Le jeu de données ne présentent pas de valeur manquante. Nous n'avons pas besoin de faire de modification de ce point de vue.

Standardisation des données

Nous pouvons remarquer que les données sont dans des unités différentes et les ordres de grandeur sont très variables. Ces boxplots nous montrent une différence notable entre les variables et nous incitent à standardiser les données.

```
boxplot(donnees)
```



Nous passons donc à la standardisation :

```
donnees.sc <- scale(donnees)
```

Afin de pouvoir analyser ces données, nous allons réaliser des statistiques descriptives de base.

Statistiques descriptives

On effectue les statistiques descriptives sur les valeurs avant standardisation.

Résumé des données :

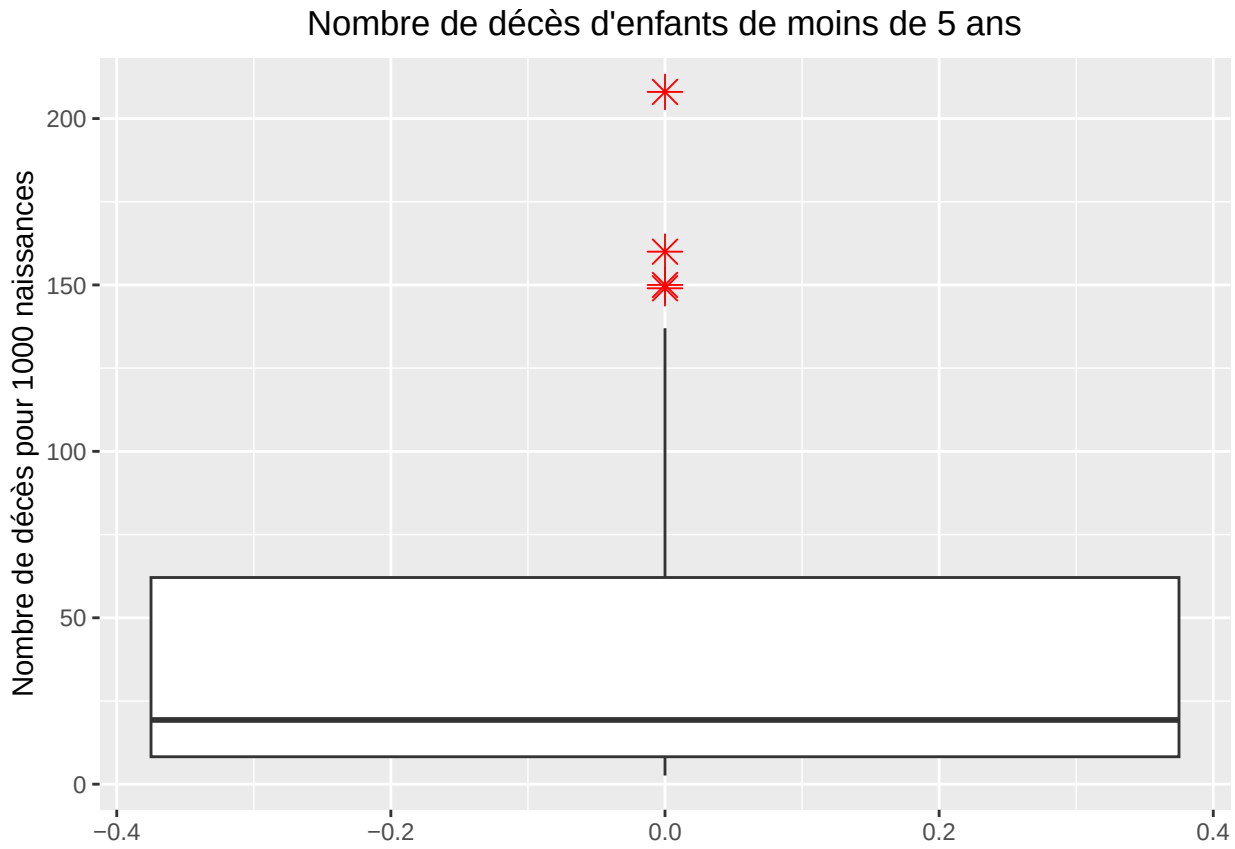
```
stargazer(donnees,type="text",title="Résumé des données",out="resume_donnnees.txt")
```

```
##
## Résumé des données
## ===================================================
## Statistic    N     Mean     St. Dev.   Min     Max
## ---------------------------------------------------
## enfant_mort 167   38.270     40.329   2.600  208.000
## exports     167   41.109     27.412   0.109  200.000
## sante       167   6.816      2.747    1.810  17.900
## imports     167   46.890     24.210   0.066  174.000
## revenu      167 17,144.690 19,278.070  609   125,000
## inflation   167   7.782      10.571   -4.210 104.000
## esper_vie   167   70.556     8.893    32.100 82.800
## fert        167   2.948      1.514    1.150   7.490
## pib_h       167 12,964.160 18,328.710  231   105,000
## ---------------------------------------------------
```

Ce résumé statistique nous permet d'avoir une vue d'ensemble sur les données.
Notre jeu de données est composé de 167 pays très hétérogènes. En effet, nous pouvons observer une assez grande différence entre le minimum et le maximum de chaque variable, ce qui prouve la diversité de notre échantillon.

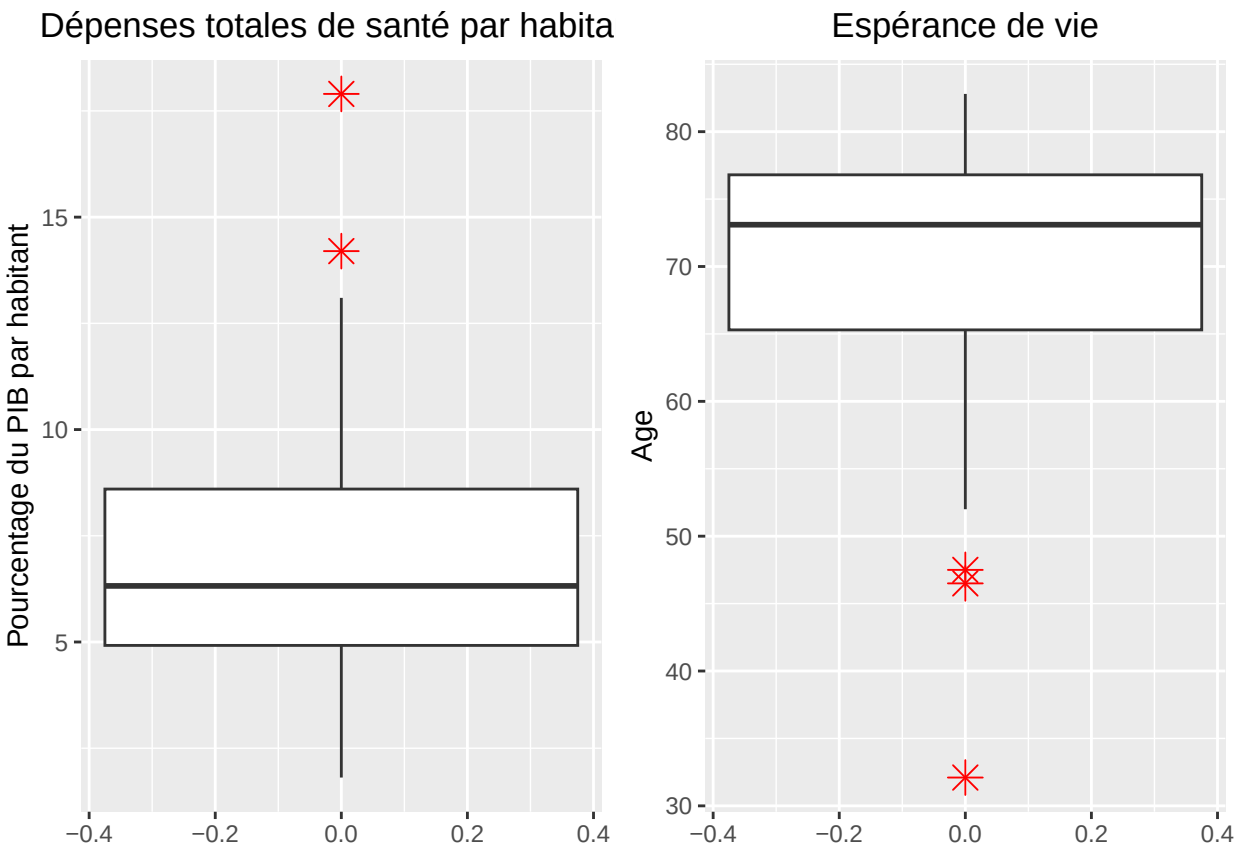
Graphiques :

```
ggplot(data=donnees, aes(y=enfant_mort)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Nombre de décès d'enfants de moins de 5 ans",y="Nombre de décès pour 1000 naissances")+
  theme(plot.title = element_text(hjust=0.5))
```



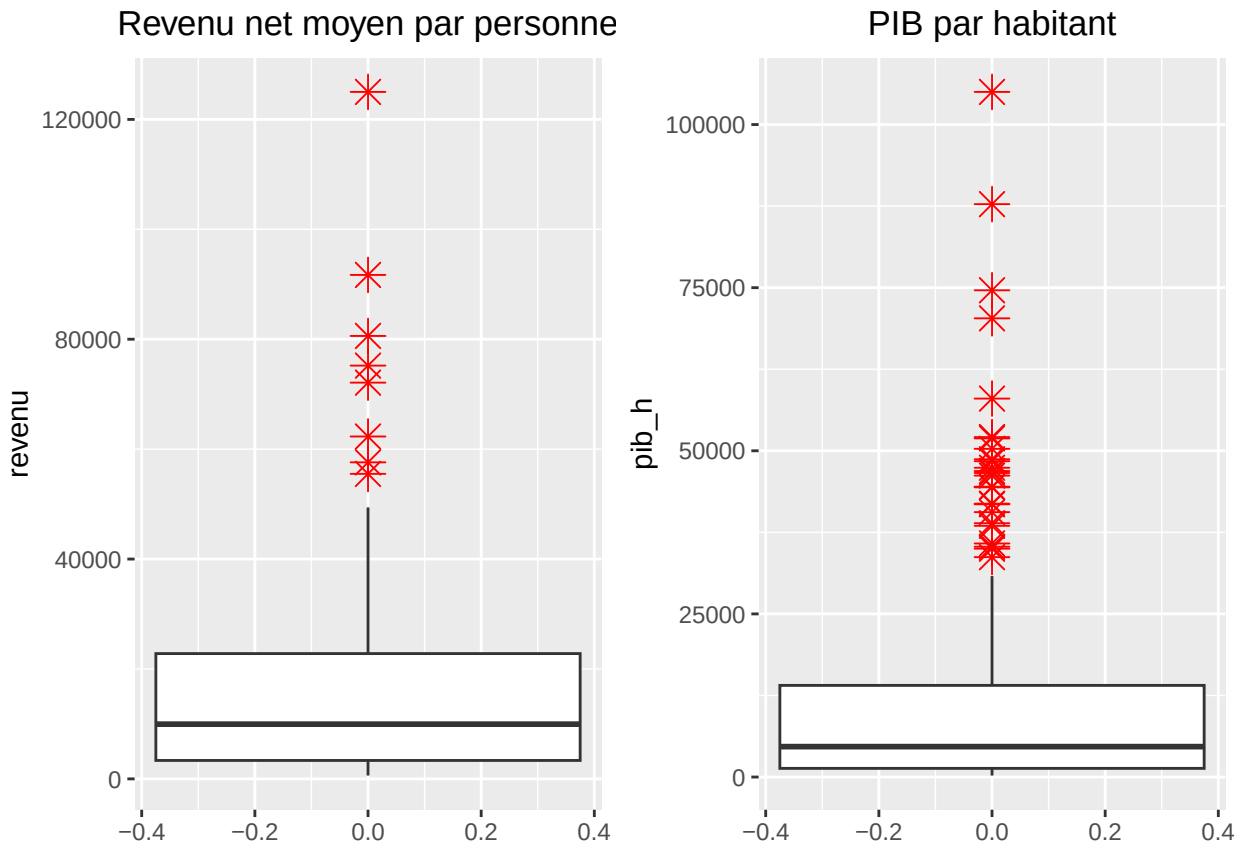
Ce graphique nous permet de voir que notre échantillon est assez hétérogène car le taux de mortalité infantile est très variable selon les pays. De plus, nous pouvons observer quelques pays avec des valeurs très élevées qui correspondent à des pays plutôt défavorisés.

```
sante <- ggplot(data=donnees, aes(y=sante)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Dépenses totales de santé par habitant",y="Pourcentage du PIB par habitant")+
  theme(plot.title = element_text(hjust=0.5))

esperance <- ggplot(data=donnees, aes(y=esper_vie)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Espérance de vie",y="Age")+
  theme(plot.title = element_text(hjust=0.5))

grid.arrange(sante,esperance,ncol=2)
```



Ici, nous avons voulu représenter une autre caractéristique : la santé. Globalement, les dépenses de santé sont assez homogènes parmi tous nos pays. Cependant, l'espérance de vie est plus faible dans certains pays de notre échantillon. Nous pouvons aussi nous interroger sur la pauvreté du pays en faisant ce constat.

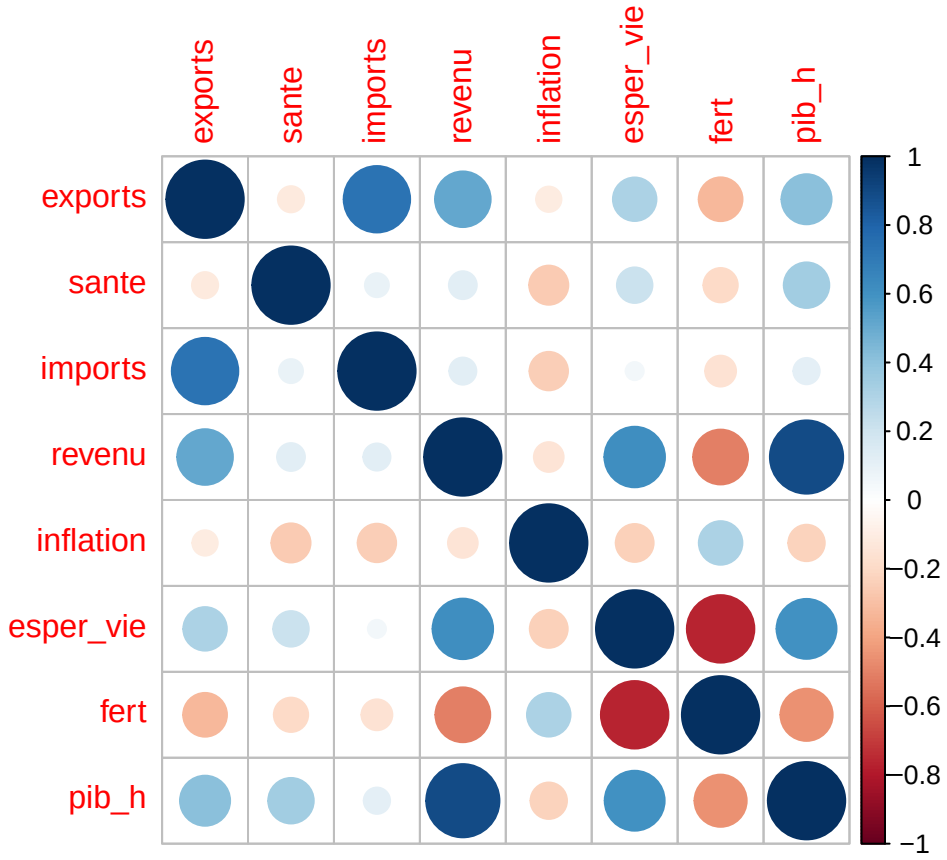
```
revenu_net <- ggplot(data=donnees, aes(y=revenu)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Revenu net moyen par personne")+
  theme(plot.title = element_text(hjust=0.5))

pib_hab <- ggplot(data=donnees, aes(y=pib_h)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="PIB par habitant")+
  theme(plot.title = element_text(hjust=0.5))

grid.arrange(revenu_net,pib_hab,ncol=2)
```



La dimension économique est aussi très importante à étudier pour cibler les pays à aider. Notre première approche avec ces deux graphiques nous permettent de voir que certains pays seront à écarter de suite puisqu'ils sont financièrement autonomes. Nous allons donc devoir nous attarder sur la moyenne basse pour leur venir en soutien.

Matrice de corrélation :

```
corrplot(cor(donnees[-1]),method="circle")
```



Grâce à cette matrice de corrélation, nous pouvons observer une corrélation négative, entre l'espérance de vie et le nombre d'enfants par femme. Par ailleurs, une corrélation positive,proche de 1, apparaît entre le PIB par habitant et le revenu net moyen par personne.

Question 2

Matrice de dissimilarité :

Afin de chercher les individus similaires, on peut calculer une matrice de distance / dissimilarité.

```
MD <- as.matrix(dist(donnees.sc,method = "euclidean"))
# MD <- as.matrix(dist(donnees.sc,method = "minkowski"))
# MD <- as.matrix(dist(donnees.sc,method = "manhattan"))
which(MD == min(MD[row(MD)!=col(MD)]),arr.ind=TRUE)
```

```
##         row col
## Poland  122  42
## Croatia  42 122
```

Avec la méthode de la distance euclidienne, de manhattan et minkowski, les deux pays les plus proches / similaires sont la Pologne et la Croatie.

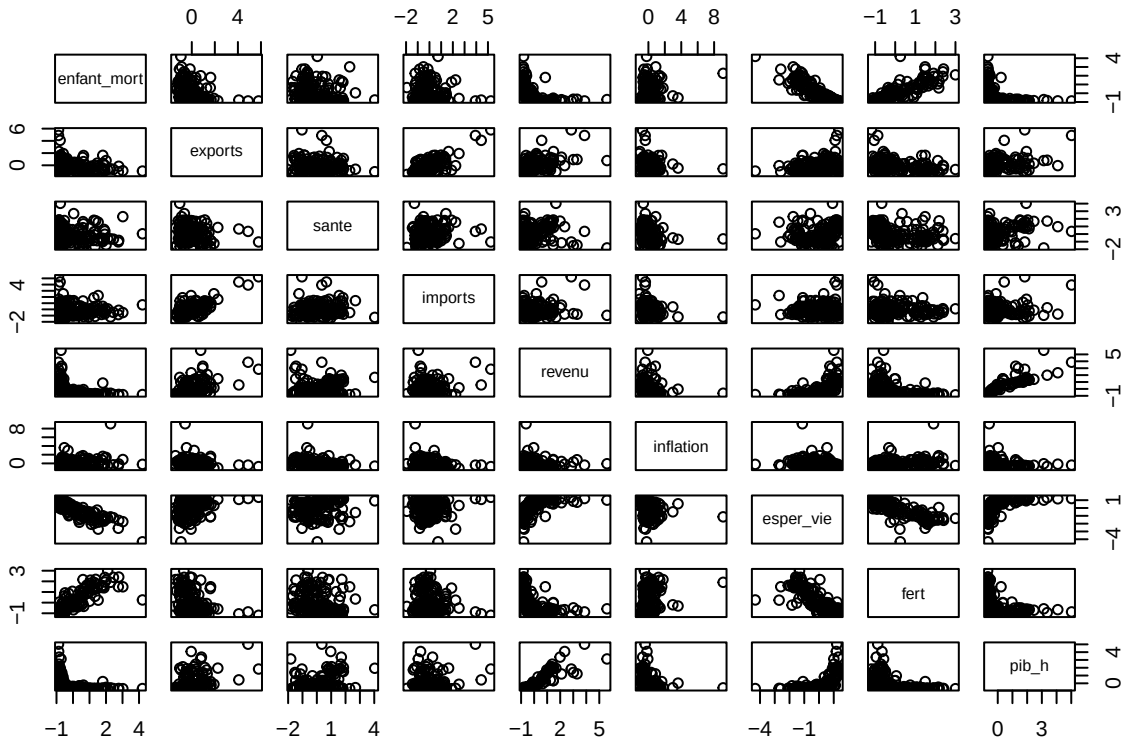
Dans notre situation, les variables sont quantitatives, nous pouvons donc utiliser une approche en termes de distances. On cherche à partitionner les pays en groupes distincts et homogènes afin de déterminer leur besoin d'aide. L'objectif est de former des groupes compacts avec une faible variabilité au sein des groupes.

Première approche : CAH

Classification Ascendante Hiérarchique :

Cette première représentation nous permet d'observer de potentiels groupes de pays. Ayant beaucoup de variables, cette analyse est un peu plus compliquée et aucune partition ne semble se démarquer.

```
pairs(donnees.sc)
```

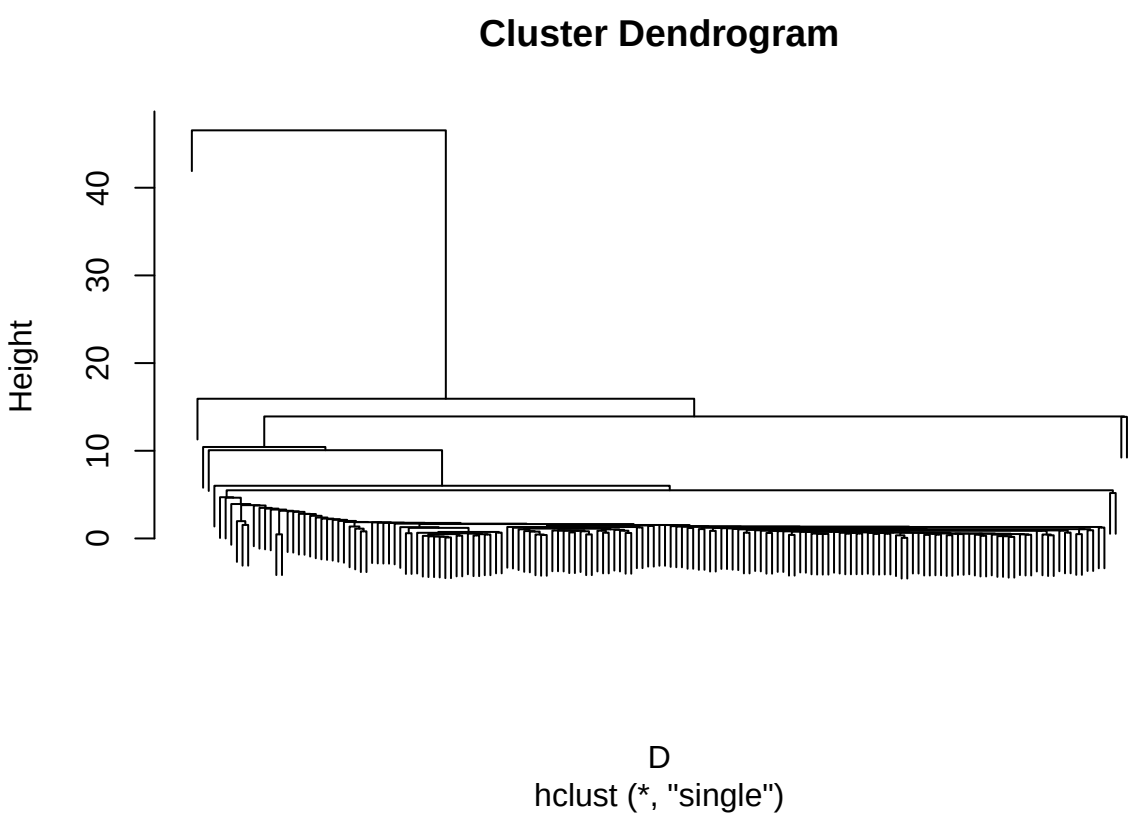


On calcule d'abord la distance euclidienne au carré. Le calcul de la distance euclidienne nous permettra par la suite d'être optimal lors de l'utilisation de la stratégie de Ward.

```
D <- dist(donnees.sc,method="euclidean")^2
```

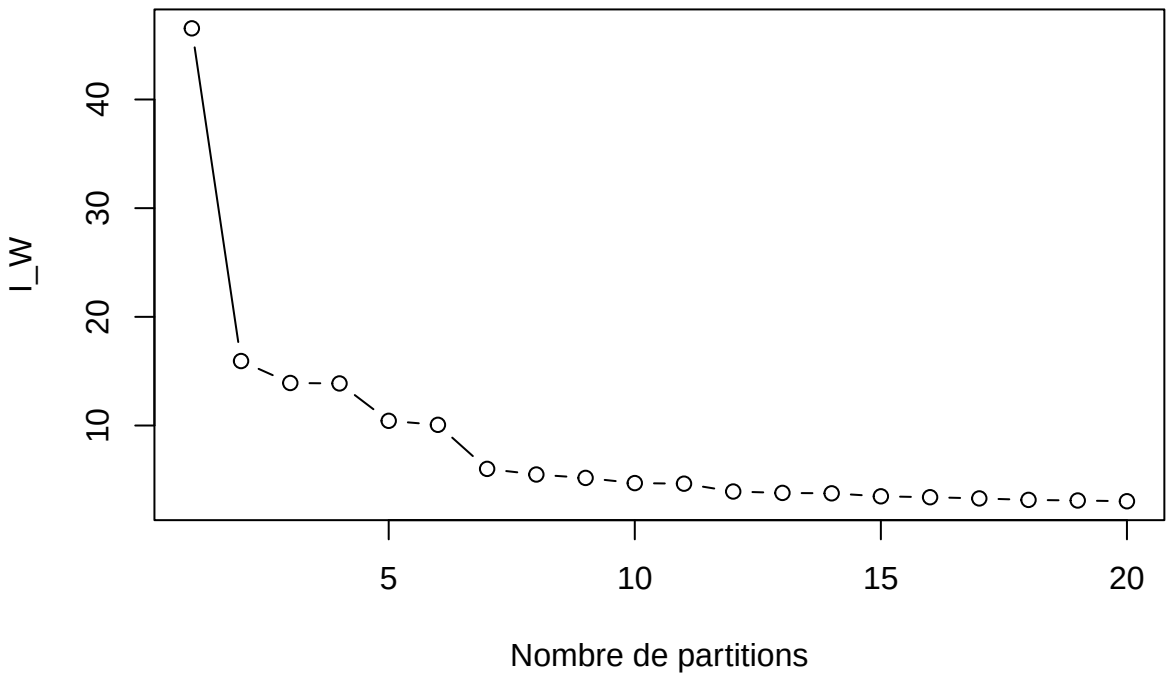
Méthode CAH avec le saut minimal (single linkage) :

```
CAH_min <- hclust(d= D,method="single")
plot(CAH_min, labels = FALSE)
```



Ce premier dendrogramme n'est pas très explicite et ne nous permet pas de faire un choix de partition clair.

```
plot(rev(CAH_min$height)[1:20],type="b",xlab="Nombre de partitions",ylab="I_W")
```

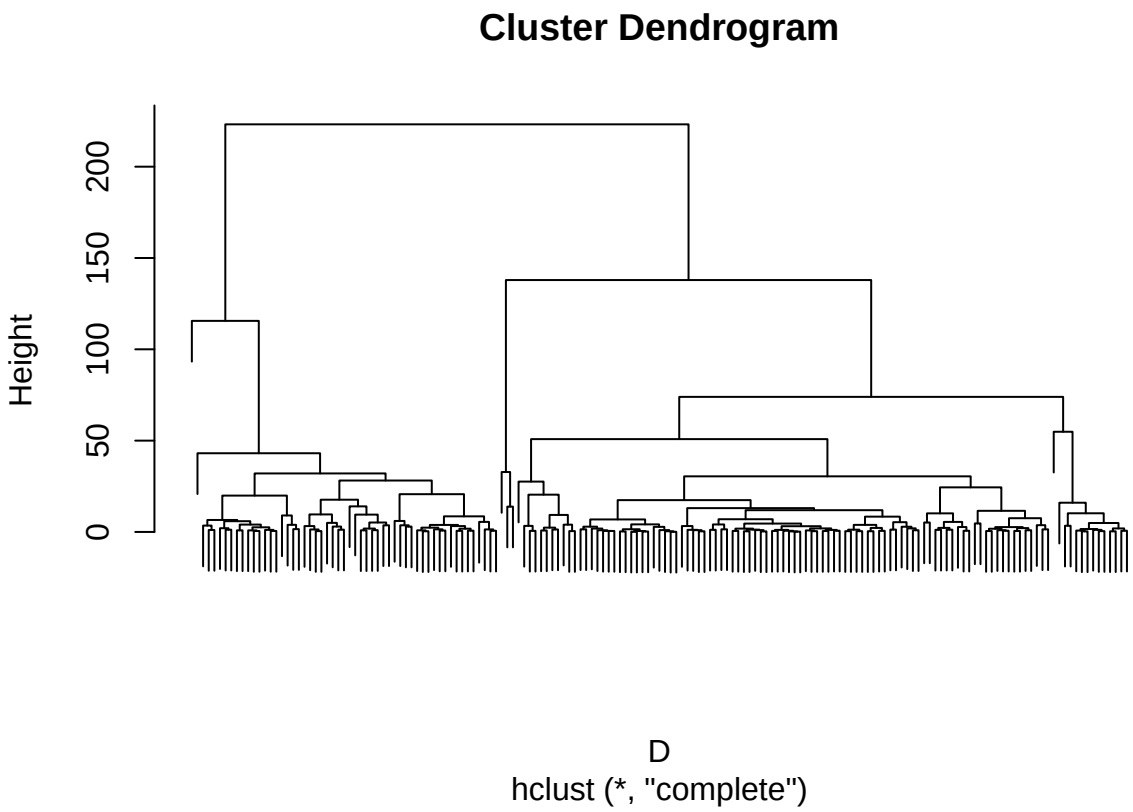


Cependant, le tracé de la pertie d'inertie nous suggère de choisir une partition en 2 groupes. Nous avons choisi de représenter seulement les 20 premières valeurs pour ne pas "noyer" l'information importante. Chaque coupure correspond à un saut important d'inertie intra-classes.

Faisons maintant les mêmes graphiques avec la méthode de distance de saut maximal (complet linkage).

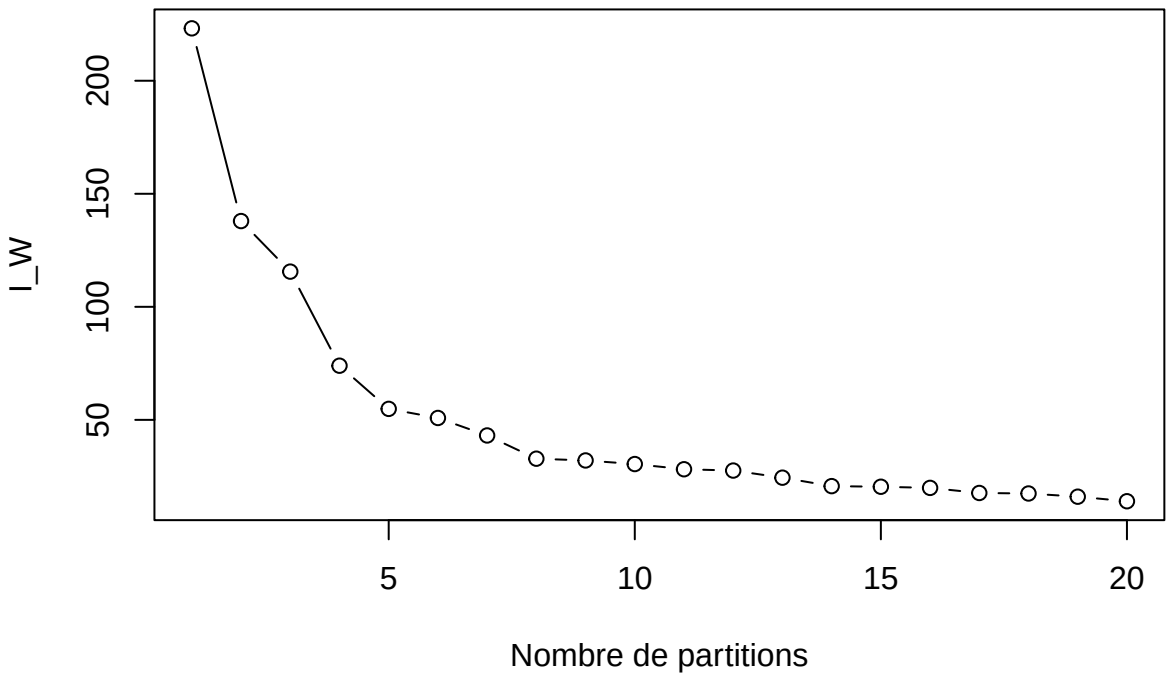
Méthode CAH avec le saut maximal :

```
CAH_max <- hclust(d= D,method="complete")
plot(CAH_max, labels = FALSE)
```



En analysant ce dendrogramme, nous pouvons distinguer 3 groupes de pays : un à droite, un au milieu et un à gauche.

```
plot(rev(CAH_max$height)[1:20],type="b",xlab="Nombre de partitions",ylab="I_W")
```

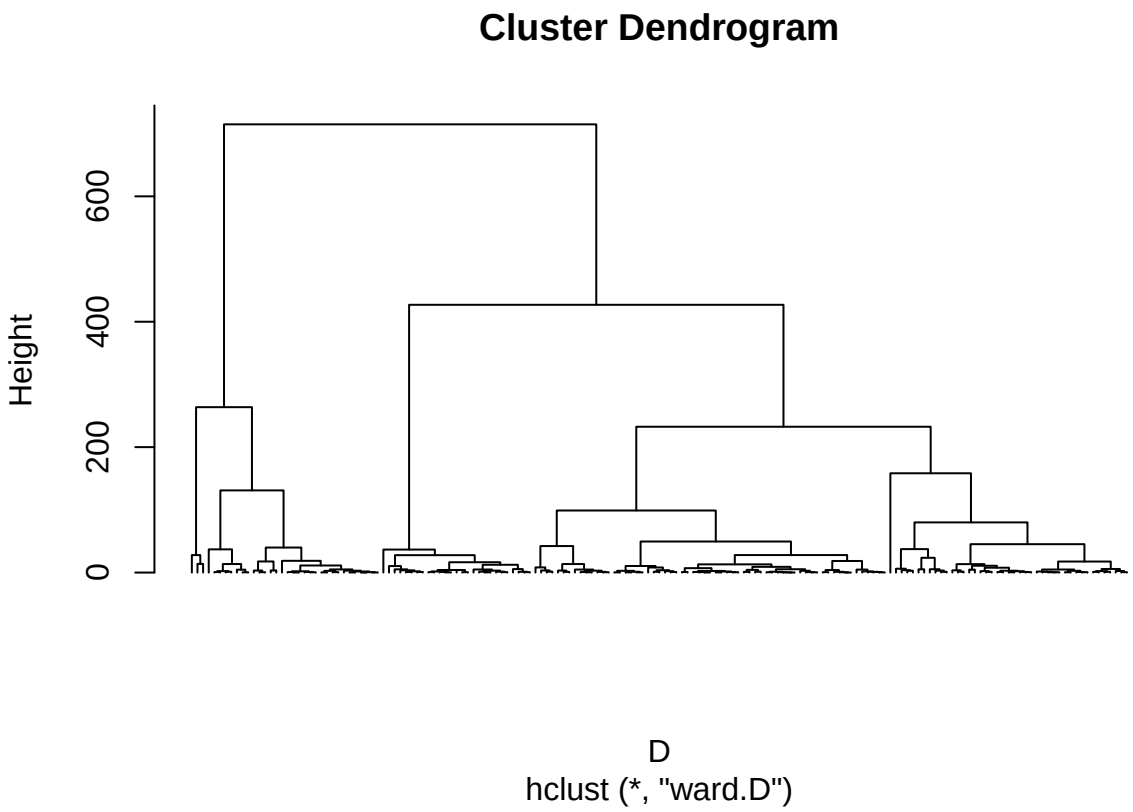


Le graphique ci-dessus n'est pas très concluant quant à l'hypothèse de partition en deux groupes. Nous nous posons la question d'une éventuelle partition en cinq groupes de pays en observant les sauts d'inertie intra-classes. En effet, ayant un budget limité, cette approche nous permettrait de le diviser au mieux. Nous continuons donc nos analyses.

CAH avec la distance de Ward :

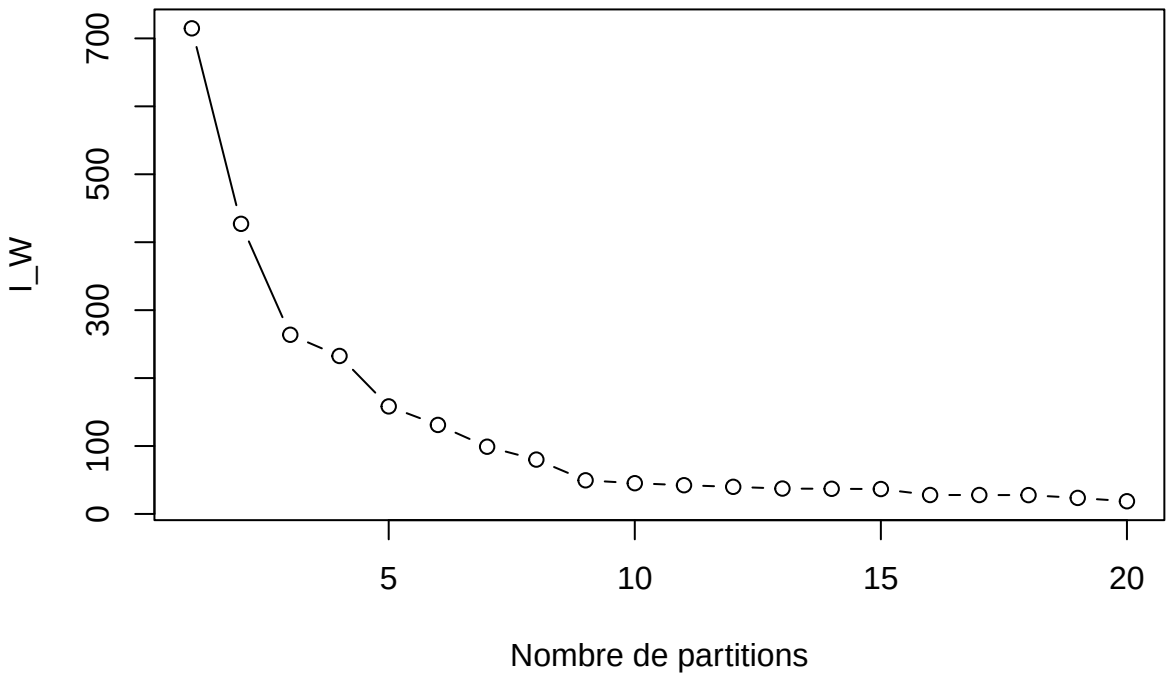
Enfin, avec la distance de ward, on obtient les résultats suivants :

```
CAH_ward <- hclust( d = D,method="ward.D")
plot(CAH_ward,hang=-1,labels=FALSE)
```



Le dendrogramme nous permet aussi de supposer l'existence de cinq groupes de pays.

```
plot(rev(CAH_ward$height)[1:20],type="b",xlab="Nombre de partitions",ylab="I_W")
```



Ce tracé de la perte d'inertie intra-classes nous indique quatre voire cinq groupes.

En conclusion, nous avons décidé de chosisir une partition en cinq groupes afin de diviser au mieux notre budget et de se focaliser sur les pays les plus nécessiteux.

Critère automatique à partir du package 'NbClust' :

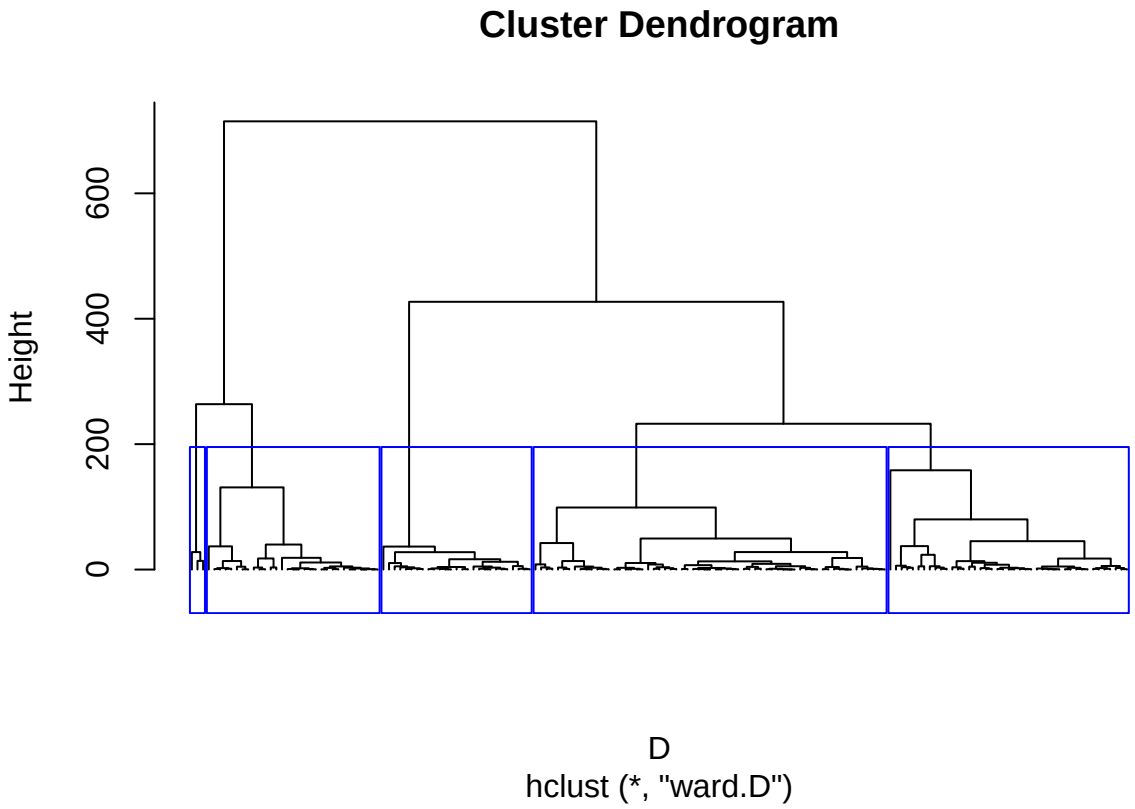
```
NbClust(donnees.sc,min.nc = 2,max.nc = 15,method="ward.D",index="all")
```

Cette étape nous permet de conclure que le choix le plus pertinent serait une partition en 2 groupes. Cependant, notre budget étant limité, nous allons devoir diviser notre échantillon en 5 groupes.

Interprétation des groupes : Représentation graphique des clusters avec un dendrogramme :

Dans notre cas, on fixe K = 5 car nous avons choisi de réaliser une partition en 5 groupes.

```
K=5
plot(CAH_ward,hang=-1,labels=FALSE)
rect.hclust(CAH_ward,K,border="blue")
```



A l'aide de ce dendrogramme, nous supposons que le groupe ayant le plus faible effectif (à gauche) est composé des pays les plus défavorisés. Nous allons analyser cet aspect par la suite.

Représentation graphique des clusters avec la fonction cutree :

La fonction cutree permet de faire apparaitre visuellement les groupes.

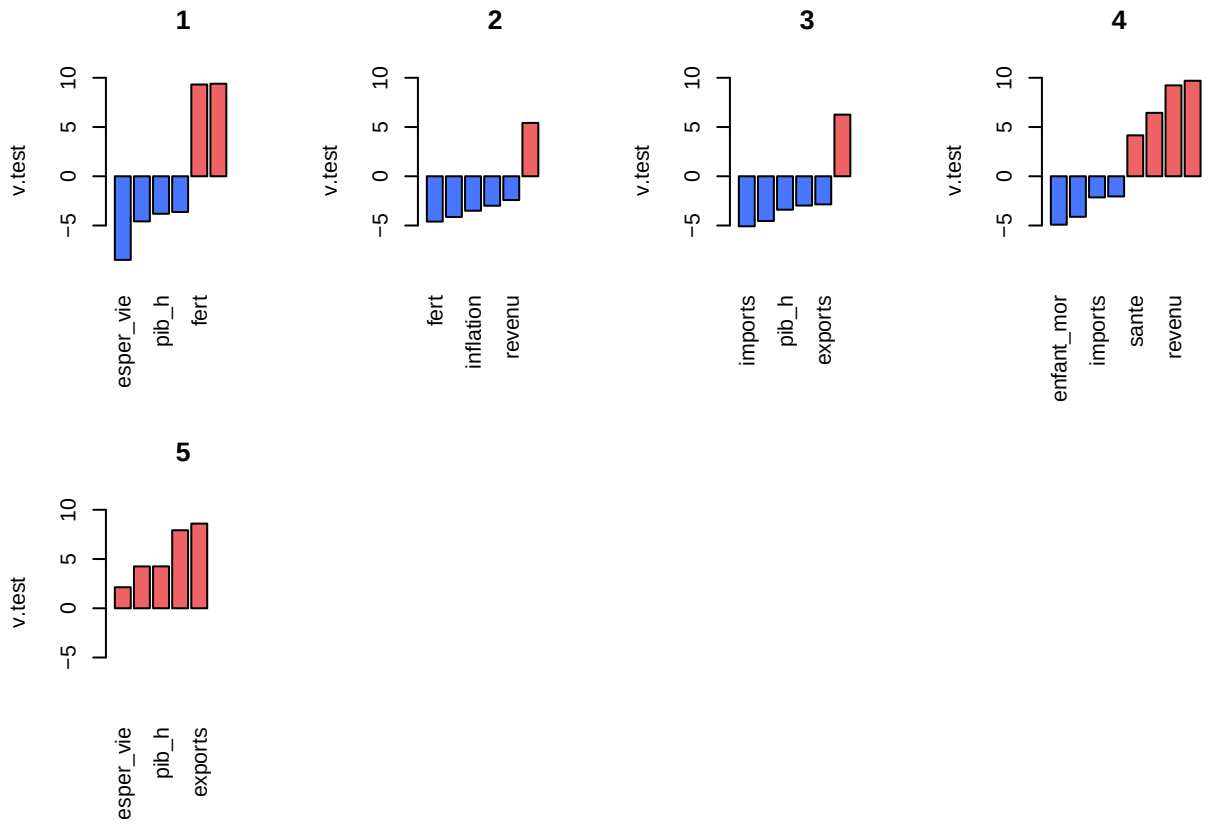
```
K = 5
gpe.ward <-  cutree(CAH_ward,k=K)
donnees  <- cbind(donnees,gpe.ward)
donnees$gpe.ward <- as.factor(donnees$gpe.ward)
# plot(gpe.ward)
interp_cah <- catdes(donnees,num.var = 10)
# interp_cah
```

```
nrow(donnees[donnees$gpe.ward==1,])
```

```
## [1] 27
```

Le groupe nécessiteux est composé de 27 pays.

```
plot.catdes(interp_cah,barplot = T,output = "figure")
```



A l'aide des deux sorties précédentes, nous remarquons que les groupes présentent des caractéristiques opposées. Le premier groupe est caractérisé par un faible taux de PIB par habitant, de revenu et d'espérance de vie. De plus, le taux d'enfants morts est plus important dans ce groupe-ci. Nous pouvons donc supposer que ce groupe comporte des pays plus défavorisés et en difficultés.

Par ailleurs, le second groupe présente des caractéristiques communes comme le PIB par habitant et le revenu. Le taux d'importation étant assez élevé nous pouvons nous demander si ces pays dépendants des économies extérieurs.

Le troisième et quatrième groupe correspondent globalement à des pays avec un niveau de vie moyen et ne sont pas notre cible première.

Le dernier groupe comporte des pays plus riches avec une production et des exportations importantes.

Représentation sur les deux premiers axes d'une ACP:

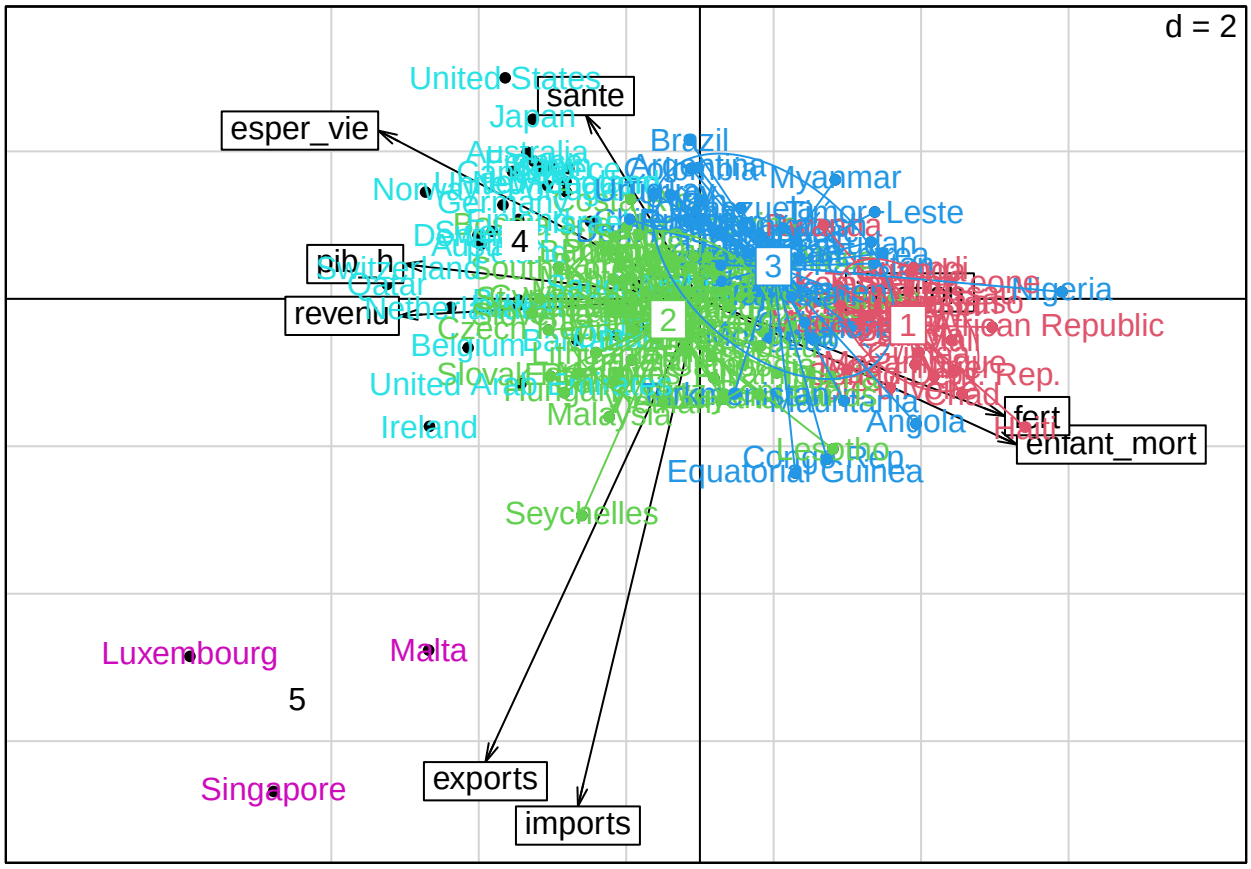
```
CCpca = dudi.pca(donnees.sc,scannf=FALSE,nf=2)
cumsum(CCpca$eig)/sum(CCpca$eig) # ?% de variabilité expliquée sur les deux premiers axes
```

```
## [1] 0.4595174 0.6313337 0.7613762 0.8719079 0.9453100 0.9701523 0.9827566
## [8] 0.9925694 1.0000000
```

```
scatter(CCpca,posieig = "none",clab.row=0,pch=NA)
```

```
## NULL
```

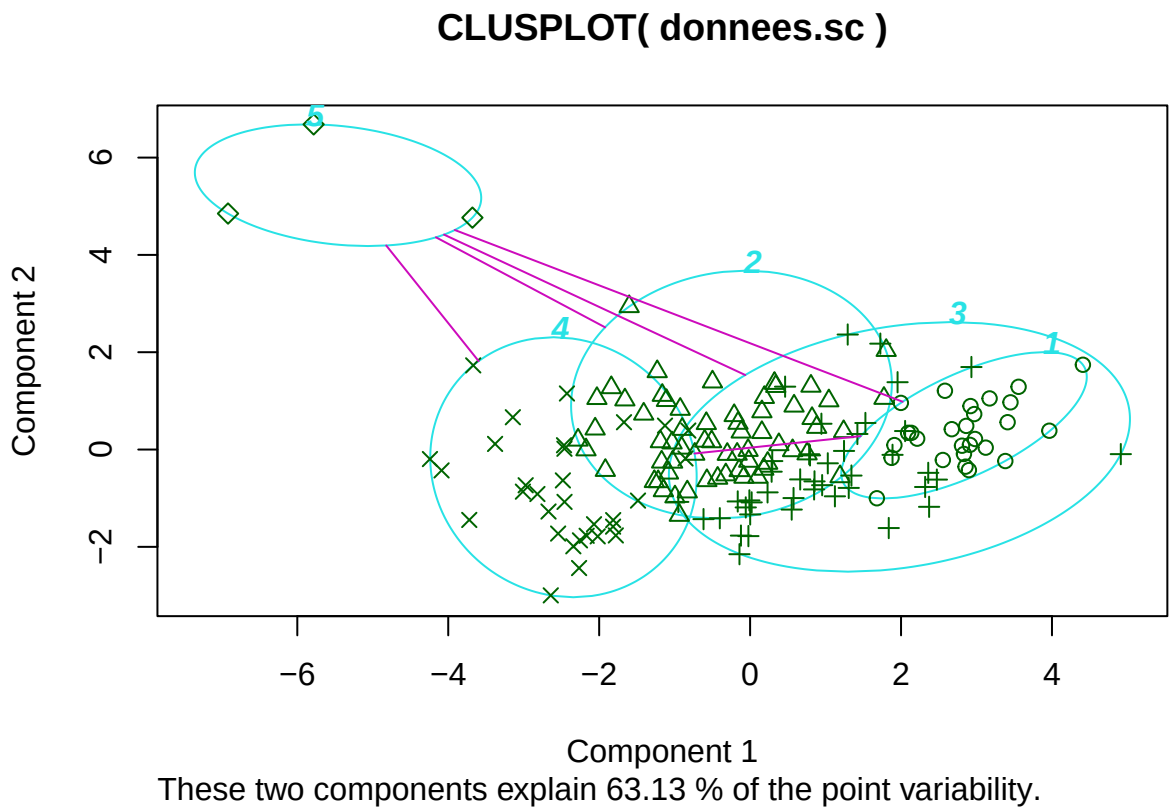
```
text(CCpca$li[,1], CCpca$li[,2],labels =row.names(donnees),col=gpe.ward+1,xpd=TRUE)
s.class(CCpca$li, factor(gpe.ward), col = 2:4, add.plot = TRUE,clabel = 1)
```



L'ACP nous présente les deux axes déterminants pour nos pays.
L'axe 1 est caractérisé par un nombre moyen d'enfants par femme important et de décès d'enfants de moins de 5 ans. Ces deux composantes sont synonymes de pauvreté et d'un faible développement du pays. Nous devons donc aider en priorité les pays corrélés positivement avec l'axe 1.
De plus, lorsque les pays sont corrélés négativement avec l'axe 2, les dépenses de santé sont très faibles.

Représentation des groupes avec la fonction clusplot :

```
K = 5
clusplot(donnees.sc,cutree(CAH_ward,K),labels=4)
```



Ce graphe correspond à la représentation des groupes sur les deux premiers axes principaux d'une ACP. De plus, des ellipses de contour autour des groupes sont tracées. Ici, nous pouvons voir 5 groupes.

Deuxième approche : Agrégation autour de centres mobiles

Pour stabiliser nos résultats, nous réalisons la méthode des kmeans. Cette fonction donne le résultat de l'algorithme d'agrégation autour des centres mobiles.

```
K = 5 # 5 groupes
c <- kmeans(donnees.sc,K,nstart=50)
```

Ici, nous avons initialisé nstart à 50 pour répéter la procédure plusieurs fois et garder la partition avec la plus faible inertie intra classes.

On récupère les groupes :

```
gpe.kmeans <- as.factor(c$cluster)
donnees.sc <- cbind(donnees.sc, gpe.kmeans)
```

Représentation sur les deux premiers axes d'une ACP:

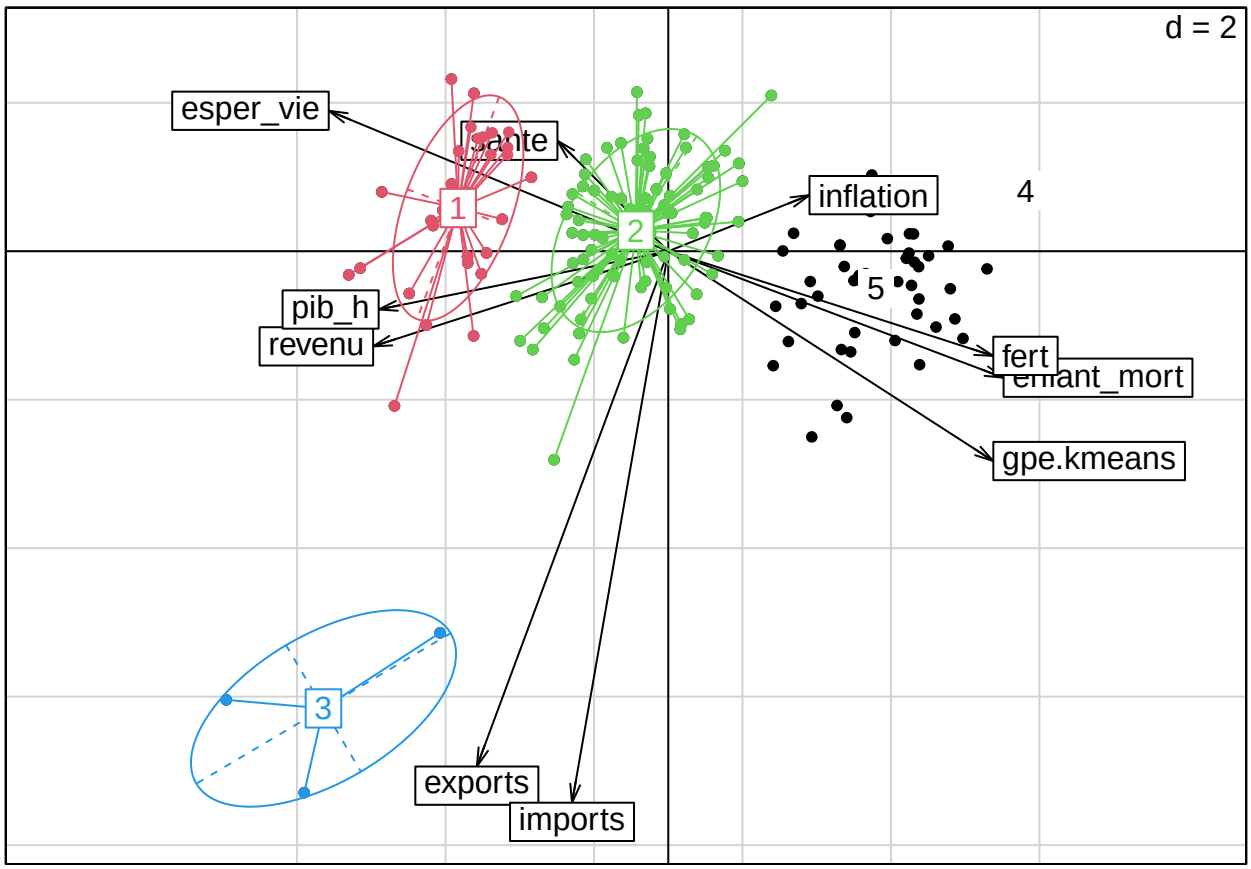
```
CCpca = dudi.pca(donnees.sc,scannf=FALSE,nf=2)
cumsum(CCpca$eig)/sum(CCpca$eig) # ?% de variabilité expliquée sur les deux premiers axes
```

```
##  [1] 0.4817222 0.6474280 0.7653582 0.8704973 0.9369205 0.9599541 0.9747029
##  [8] 0.9855549 0.9933596 1.0000000
```

```
scatter(CCpca,posieig = "none",clab.row=0,pch=NA)
```

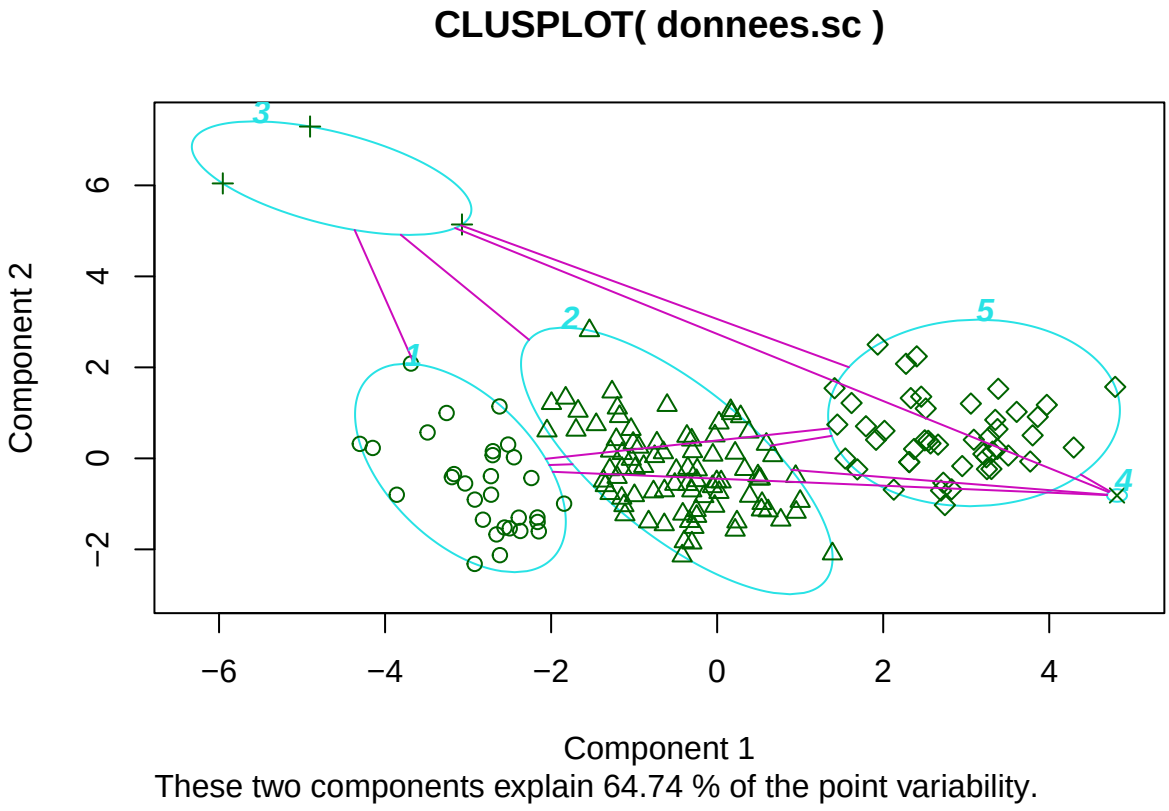
```
## NULL
```

```
text(CCpca$li[,1], CCpca$li[,2],labels = "",col=gpe.kmeans+1,xpd=TRUE) # labels = "" pour une meilleure
s.class(CCpca$li, factor(gpe.kmeans), col = 2:4, add.plot = TRUE,clabel = 1)
```



L'ACP présente ici est similaire à celle obtenu avec la méthode de la CAH.

```
clusplot(donnees.sc,c$cluster,labels=4)
```



Le cluster situé le plus à droite est caractérisé par un nombre important d'enfants morts de moins de 5 ans, une fertilité significative et une espérance de vie assez faible. Il correspond donc au groupe à aider en priorité.
Nous observons un unique point (à droite) correspondant à un pays très en difficulté. Notre budget lui sera donc en partie destiné.
Dans un second temps, nous aiderons le deuxième cluster regroupant les pays en besoin d'aide.

Question 3

Ayant un jeu de données assez restreint, les **kmeans** sont utiles mais la CAH est encore faisable.
Dans notre cas, la méthode des kmeans permet une convergence plus rapide. Nous savons d'ores et déjà que nous allons apporter notre aide au Nigéria en priorité. Nous souhaitons aider d'autres pays, c'est pourquoi nous avons sélectionné le deuxième cluster (dont on a parlé ci-dessus). Or, ce deuxième groupe contient 47 pays. Puisque notre budget est limité, nous allons réaliser une deuxième CAH pour le répartir le plus efficacement possible.

On souhaite donc récupérer les pays correspondant à ce deuxième cluster.
Puisque la méthode des kmeans nous donne des numéros de groupe variable à chaque éxecution du programme, nous devons récupérer le bon groupe correspondant aux pays nécessiteux.
On se base sur un pays étant dans le groupe des défavorisés et on récupère tous les autres du même groupe.
On utilise le Benin comme référence pour accéder aux autres pays du même groupe (défavorisés)

```r
df_donnees_sc <- as.data.frame(donnees.sc)
# On récupère le numéro du groupe du Bénin :
num_grp <- df_donnees_sc$gpe.kmeans[rownames(df_donnees_sc)=="Benin"]
cat("Le numéro du groupe du bénin est",num_grp)
```

```
## Le numéro du groupe du bénin est 5
```

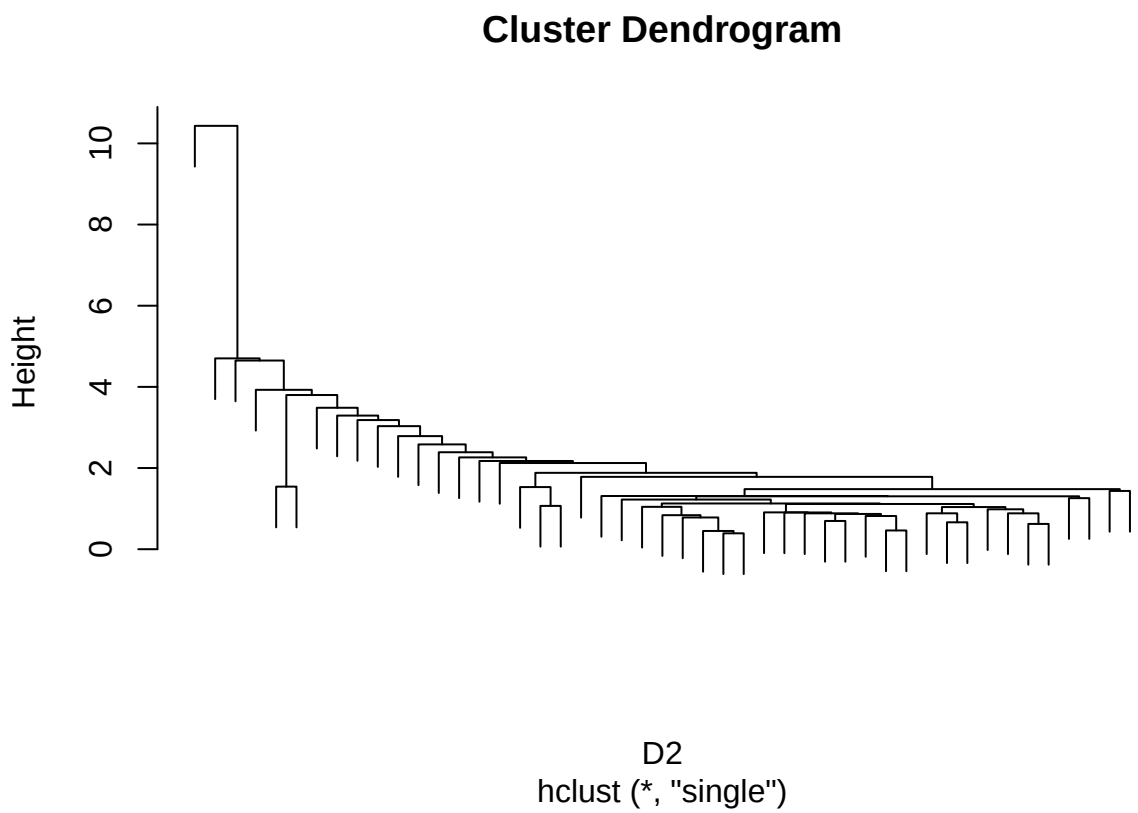
```r
# On récupère les données associées à ce groupe :
donnees.sc_defav <- df_donnees_sc[df_donnees_sc$gpe.kmeans==num_grp,]
donnees.sc_defav <- donnees.sc_defav[,-10] # on retire la colonne correspondant au groupe
```

On réalise maintenant une deuxième CAH :

```r
D2 <- dist(donnees.sc_defav,method="euclidean")^2
```

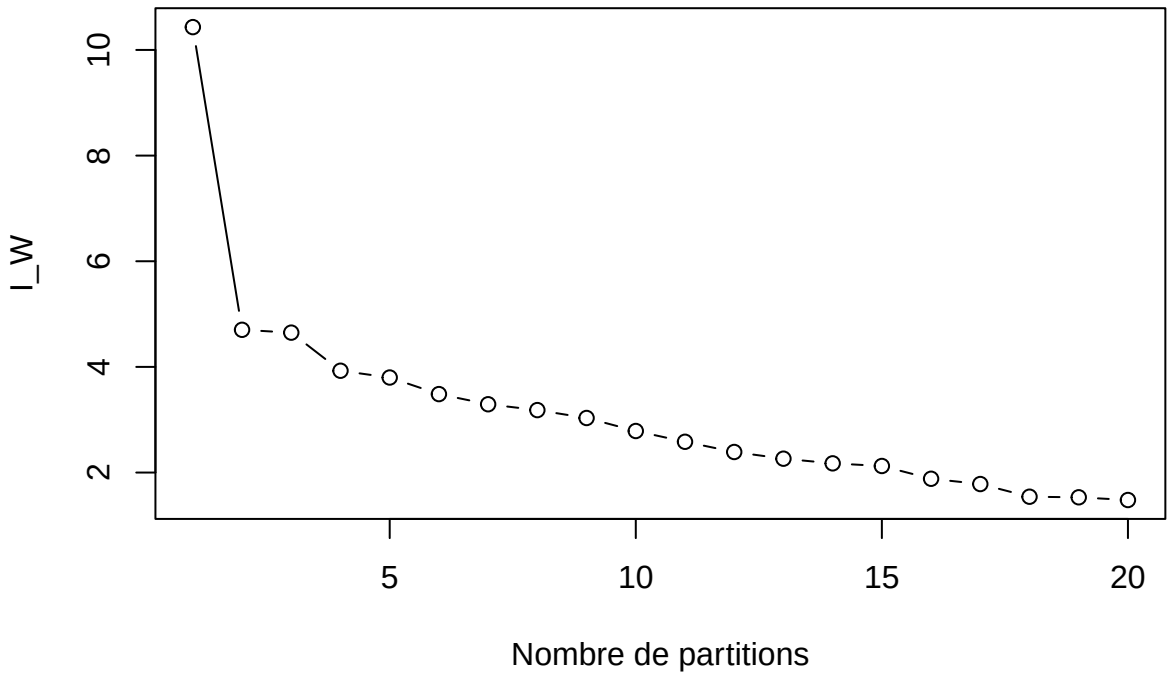
Méthode CAH avec le saut minimal (single linkage) :

```r
CAH_min2 <- hclust(d= D2,method="single")
plot(CAH_min2, labels = FALSE)
```



Ce premier dendrogramme n'est pas très explicite et ne nous permet pas de faire un choix de partition clair.

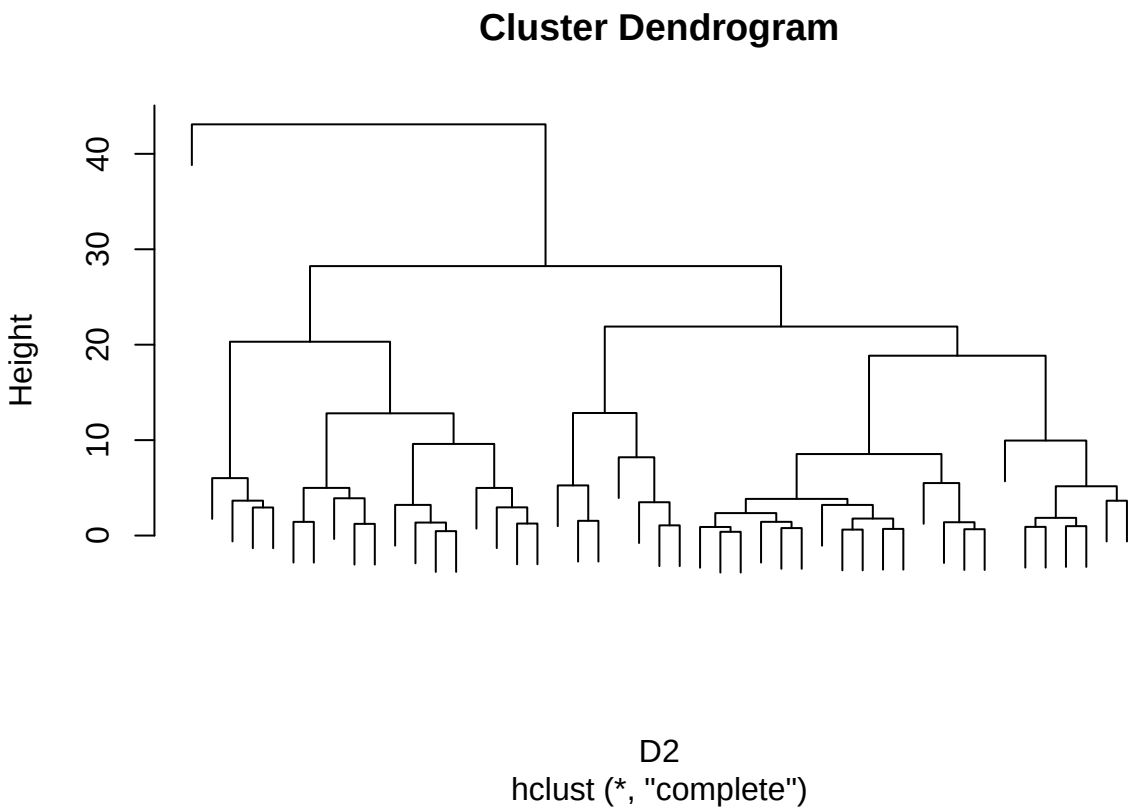
```
plot(rev(CAH_min2$height)[1:20],type="b",xlab="Nombre de partitions",ylab="I_W")
```



Cependant, le tracé de la pertie d'inertie nous suggère de choisir une partition en 2 ou 3 groupes.

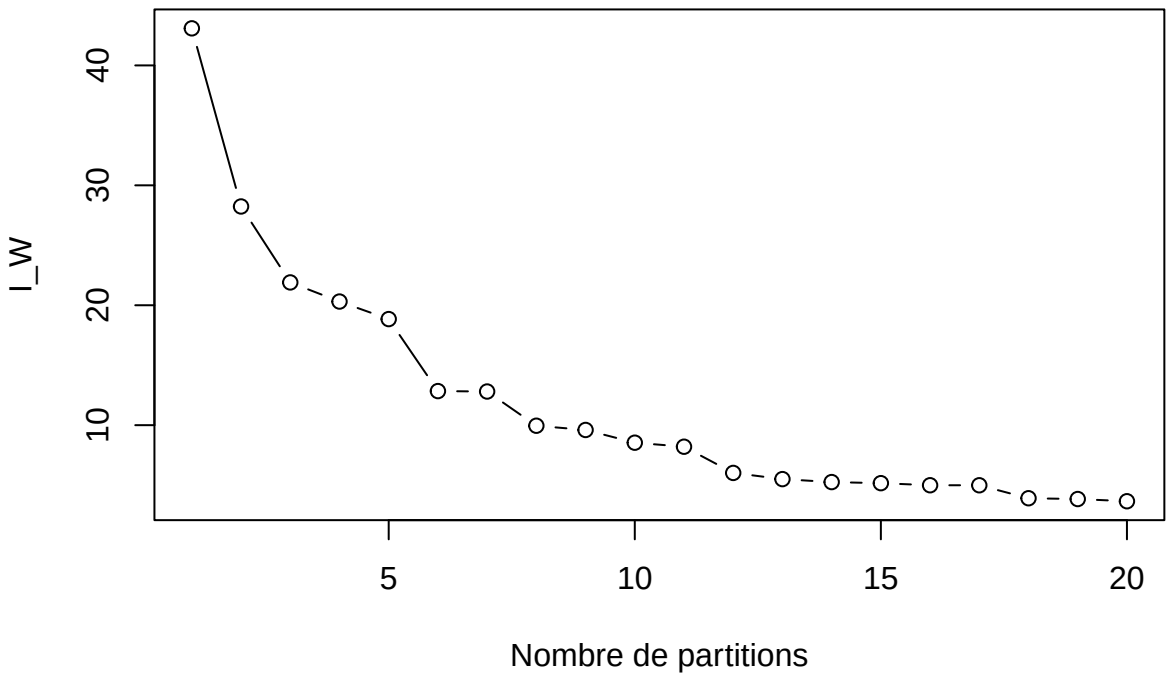
Méthode CAH avec le saut maximal :

```
CAH_max2 <- hclust(d= D2,method="complete")
plot(CAH_max2, labels = FALSE)
```



En analysant ce dendrogramme, nous pouvons distinguer 3 groupes de pays (un à droite, un au milieu et un à gauche).

```
plot(rev(CAH_max2$height)[1:20],type="b",xlab="Nombre de partitions",ylab="I_W")
```

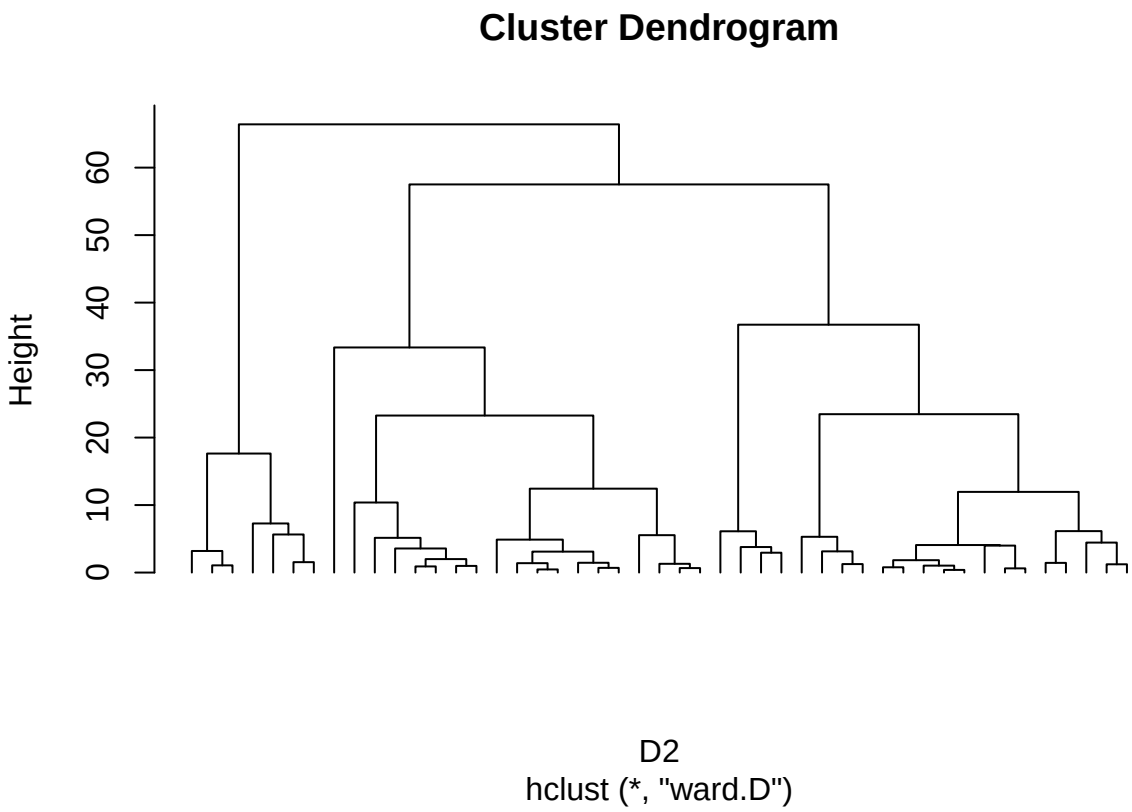


Le graphique ci-dessus n'est pas très concluant quant à l'hypothèse de partition en deux groupes. Nous nous posons la question d'une éventuelle partition en 3 groupes de pays en observant les sauts d'inertie intra-classes.

CAH avec la distance de Ward :

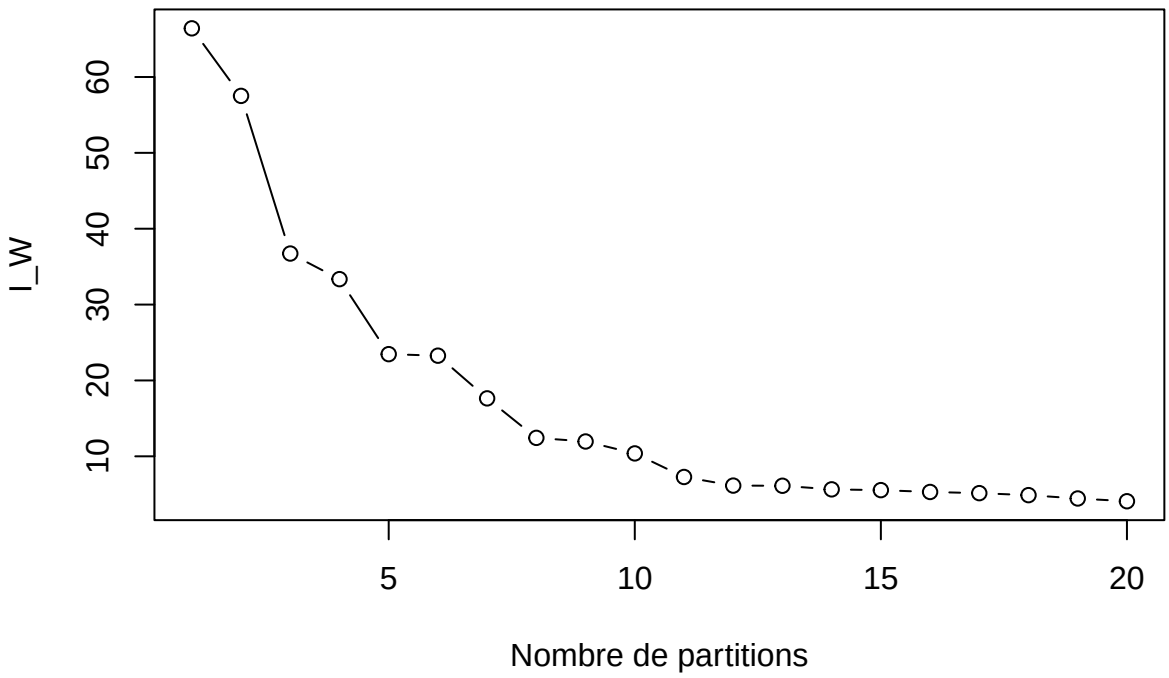
Enfin, avec la distance de ward, on obtient les résultats suivants :

```
CAH_ward2 <- hclust( d = D2,method="ward.D")
plot(CAH_ward2,hang=-1,labels=FALSE)
```



Le dendrogramme nous permet aussi de supposer l'existence de 3 groupes de pays.

```
plot(rev(CAH_ward2$height)[1:20],type="b",xlab="Nombre de partitions",ylab="I_W")
```



Ce tracé de la perte d'inertie intra-classes nous indique 2 groupes.

En conclusion, nous avons décidé de chosisir une partition en 3 groupes afin de diviser au mieux notre budget et de se focaliser sur les pays les plus nécessiteux.

Critère automatique à partir du package 'NbClust' :

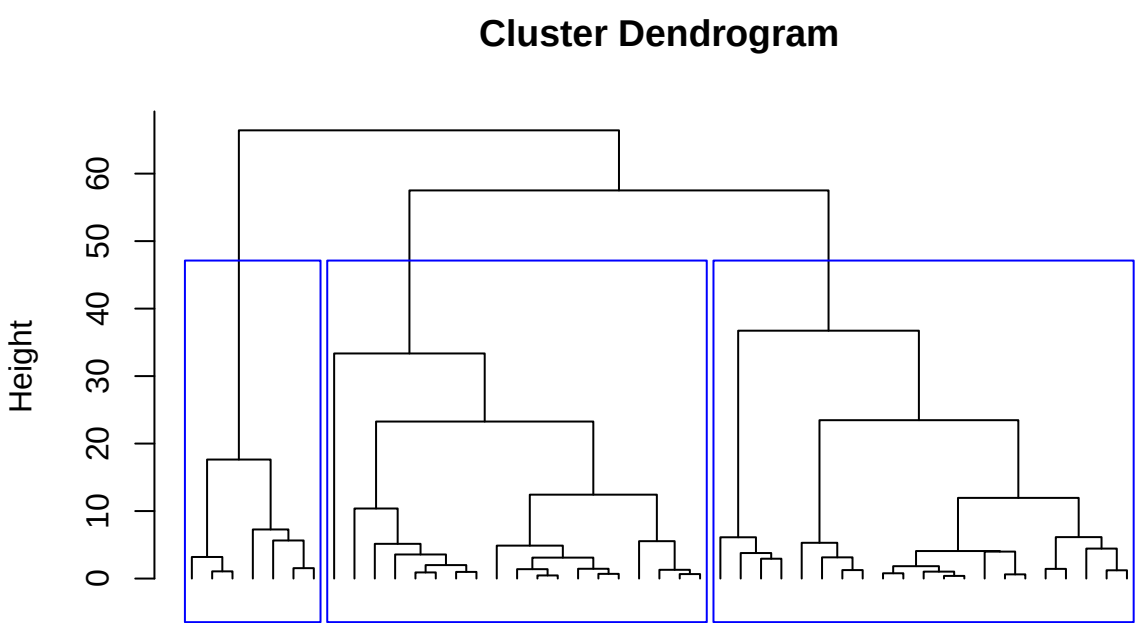
```
NbClust(donnees.sc_defav,min.nc = 2,max.nc = 15,method="ward.D",index="all")
```

Même si cette méthode propose à la majorité une partition en 2 groupes, nous choisissons de diviser en 3 pour répartir au mieux notre budget.

Interprétation des groupes : Représentation graphique des clusters avec un dendrogramme :

Dans notre cas, on fixe K = 3 car nous avons choisi de réaliser une partition en 3 groupes.

```
K=3
plot(CAH_ward2,hang=-1,labels=FALSE)
rect.hclust(CAH_ward2,K,border="blue")
```



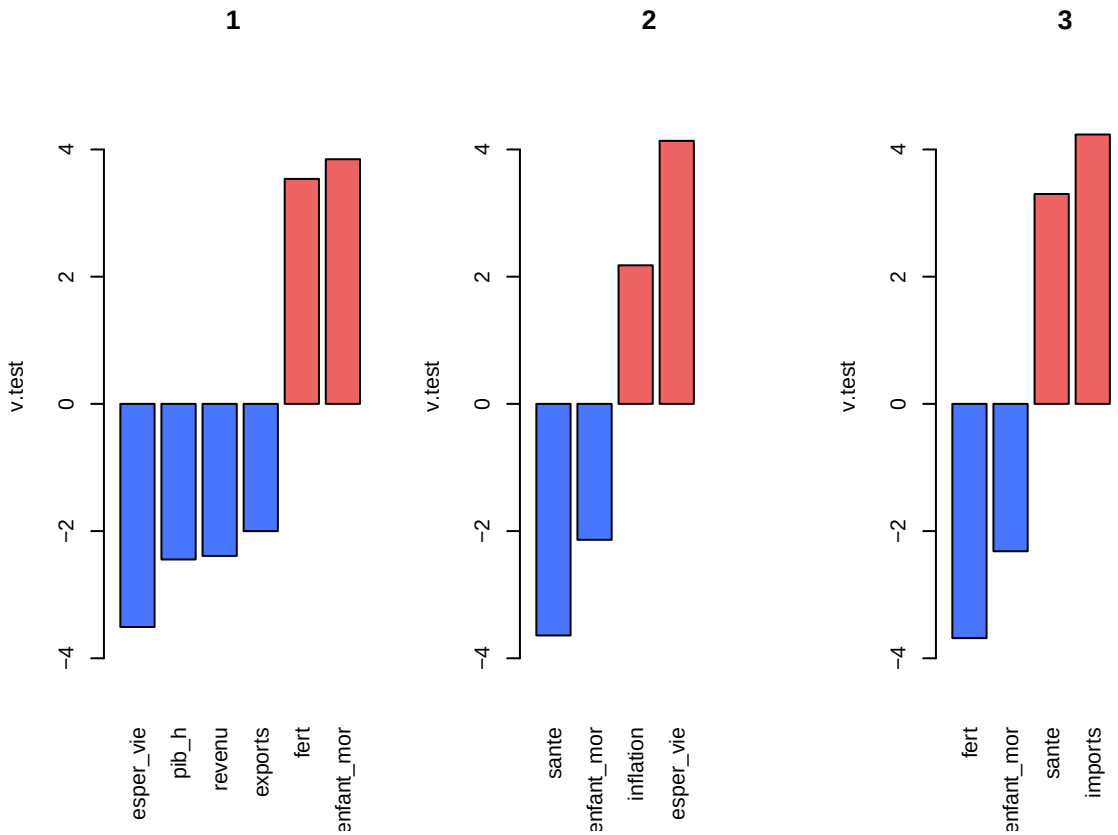
A l'aide de ce dendrogramme, nous supposons que le groupe ayant le plus faible effectif (à gauche) est composé du pays le plus défavorisé. Nous allons analyser cet aspect par la suite.

Représentation graphique des clusters avec la fonction cutree :

La fonction cutree permet de faire apparaitre visuellement les groupes.

```
K = 3
gpe.ward2 <-  cutree(CAH_ward2,k=K)
donnees.sc_defav2  <- cbind(donnees.sc_defav,gpe.ward2)
donnees.sc_defav2$gpe.ward2 <- as.factor(donnees.sc_defav2$gpe.ward2)
# plot(gpe.ward2)
interp_cah2 <- catdes(donnees.sc_defav2,num.var = 10)
# interp_cah2
```

```
plot.catdes(interp_cah2,barplot = T)
```



A l'aide des deux sorties précédentes, nous remarquons que les groupes présentent des caractéristiques. Le premier groupe est représenté par un faible taux d'importations et des dépenses de santé faibles. Cependant, il semble que ce groupe ait une fertilité importante.
Concernant le second groupe, nous observons une opposition dans ces trois domaines. De plus, le nombre d'enfants morts de moins de 5 ans est faible.
Enfin, le dernier groupe présente une seule caractéristique principale : un taux d'inflation très élevé.

Il est difficile d'identifier quel groupe correspond au groupe le plus défavorisé, nous allons donc continuer notre étude.

Représentation sur les deux premiers axes d'une ACP:

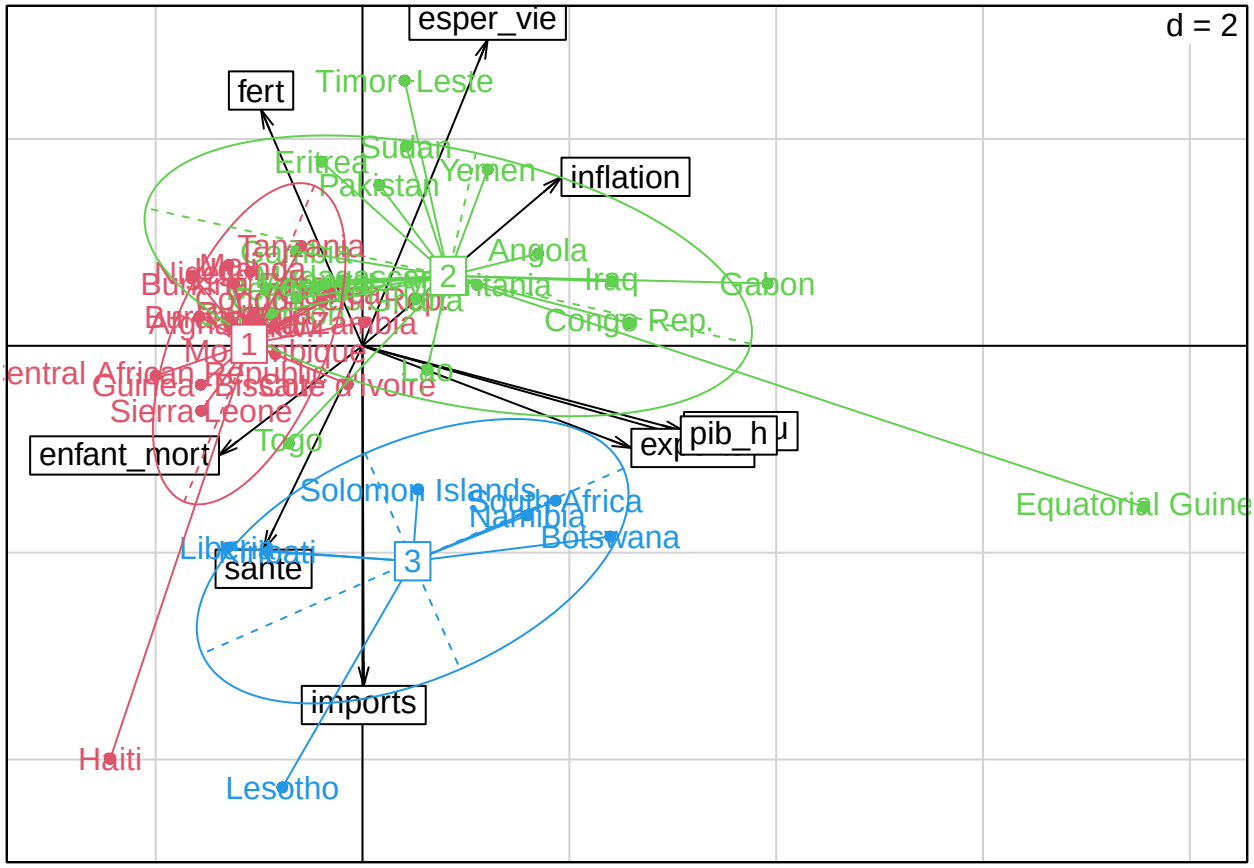
```
CCpca = dudi.pca(donnees.sc_defav,scannf=FALSE,nf=2)
cumsum(CCpca$eig)/sum(CCpca$eig) # ?% de variabilité expliquée sur les deux premiers axes
```

```
## [1] 0.3316051 0.5289347 0.6937215 0.8009149 0.8949914 0.9524356 0.9809640
## [8] 0.9975829 1.0000000
```

```
scatter(CCpca,posieig = "none",clab.row=0,pch=NA)
```

```
## NULL
```

```
text(CCpca$li[,1], CCpca$li[,2],labels =row.names(donnees.sc_defav),col=gpe.ward2+1,xpd=TRUE)
s.class(CCpca$li, factor(gpe.ward2), col = 2:4, add.plot = TRUE,clabel = 1)
```

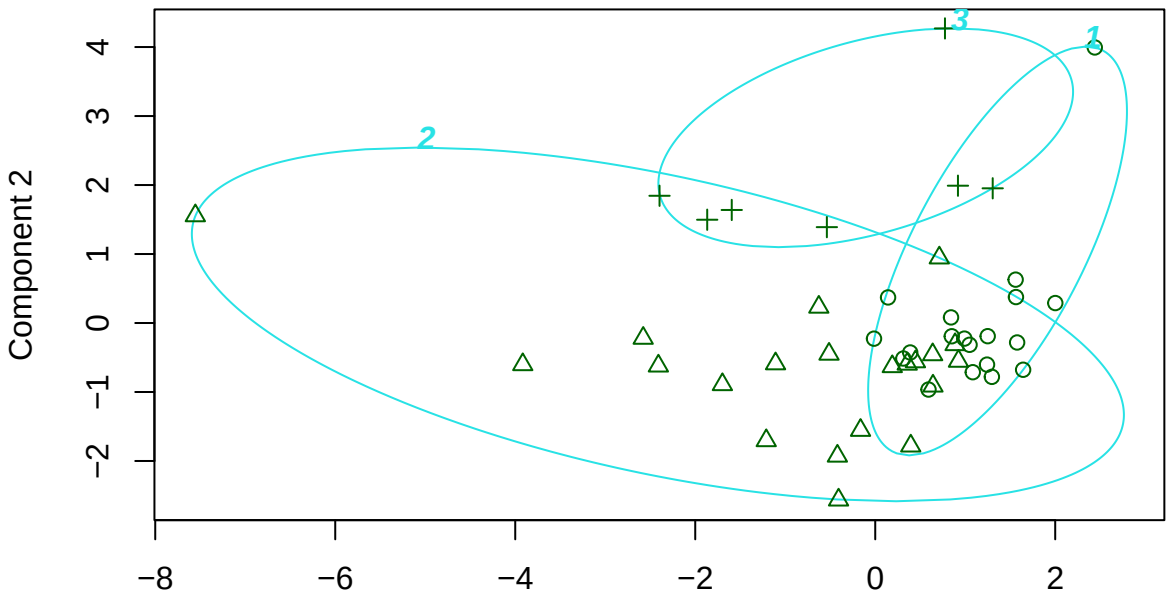


L'ACP nous présente les deux axes déterminants pour nos pays.
L'axe 1 est corrélé positivement avec des exportations et un PIB par habitant importants, négativement au nombre d'enfants morts de moins de 5 ans. L'axe 2, quant à lui, est corrélé positivement à la fertilité et l'espérance de vie et négativement aux importations et aux dépenses de santé.

Nous devons donc aider en priorité les pays corrélés négativement à l'axe 1 et 2.

Représentation des groupes avec la fonction clusplot :

```
K = 3
clusplot(donnees.sc_defav,cutree(CAH_ward2,K),labels=4)
```



Ce graphe correspond à la représentation des groupes sur les deux premiers axes principaux d'une ACP. De plus, des ellipses de contour autour des groupes sont tracées. Ici, nous pouvons voir 3 groupes.
La méthode du clusplot permet d'identifier différents groupes mais nous décidons d'utiliser l'ACP pour déterminer le groupe de pays que nous allons aider. Ainsi, nous allons choisir celui avec le plus important taux de mortalité infantile et un faible PIB par habitant.

Nous répertorions les pays de ce groupe en prenant comme point de référence un pays que nous connaissons, très pauvre, le Burkina Faso:

```
# On récupère le numéro du groupe du Burkina Faso :
num_grp_help <- donnees.sc_defav2$gpe.ward2[rownames(donnees.sc_defav2)=="Burkina Faso"]
cat("Le numéro du groupe du Burkina Faso est",num_grp_help)
```

```
## Le numéro du groupe du Burkina Faso est 1
```

```
# On récupère les données associées à ce groupe :
pays_help <- donnees.sc_defav2[donnees.sc_defav2$gpe.ward2==num_grp_help,]
# Affichage des pays
liste_noms <- split(rownames(pays_help)," ")
cat("Les pays à aider sont :", paste(liste_noms$` `,","))
```

```
## Les pays à aider sont : Afghanistan , Burkina Faso , Burundi , Central African Republic , Chad , Cong
```

Question 4

Finalement, après avoir réalisé différentes méthodes, nous sommes arrivées à plusieurs conclusions.
Tout d'abord, notre première étape, la CAH et les K-means nous ont permis de dégager un seul pays très en difficulté, le Nigéria.
Ce résultat ne nous a pas paru optimal puisque nous souhaitions aider plusieurs pays. Grâce à la réalisation d'une ACP, nous avons pu identifier des caractéristiques socio-économiques similaires entre certains pays. Notre objectif a été de les regrouper en clusters homogènes. Ensuite, nous avons pu évaluer les niveaux de développement économique, social et humain dans chacun de ces clusters et identifier les pays les plus pauvres qui ont besoin d'une aide. Suite à cela, un autre groupe de pays défavorisé s'est dégagé. Nous avons donc effectué une seconde CAH sur celui-ci. Grâce à notre choix de partionnement en 3 groupes, nous avons obtenu une liste de pays plus restreinte permettant d'allouer au mieux notre budget.
En somme, les différentes méthodes employées, la Classification Ascendante Hiérarchique, les algorithmes de clustering comme les K-means sont utiles pour comprendre les similitudes et différences entre les pays défavorisés. Aussi, cela nous a permis d'identifier les pays les plus nécessiteux et qui ont besoin d'une aide et d'un soutien supplémentaires.

Question 5

En conclusion, deux choix s'offrent à nous : nous pouvons proposer de venir en aide seulement au Nigéria, ou bien, nous pouvons répartir notre budget entre plusieurs pays. En prenant considération du niveau de pauvreté des pays sélectionnés, nous allons répartir au mieux nos 10 millions d'euros. Dans un premier temps, nous allons donner 1/3 de notre budget au Nigéria, soit environ 3 400 000 euros. Le reste sera divisé entre les autres pays. Puisque nous avons sélectionné 19 pays dans notre seconde étape, nous allons donner environ 315 000 euros à chacun.

1. Nigéria
2. Afghanistan
3. Burkina Faso
4. Burundi
5. Central African Republic
6. Chad
7. Congo Dem. Rep.
8. Cote d'Ivoire
9. Guinea
10. Guinea-Bissau
11. Haiti
12. Malawi
13. Mali
14. Mozambique
15. Niger
16. Rwanda

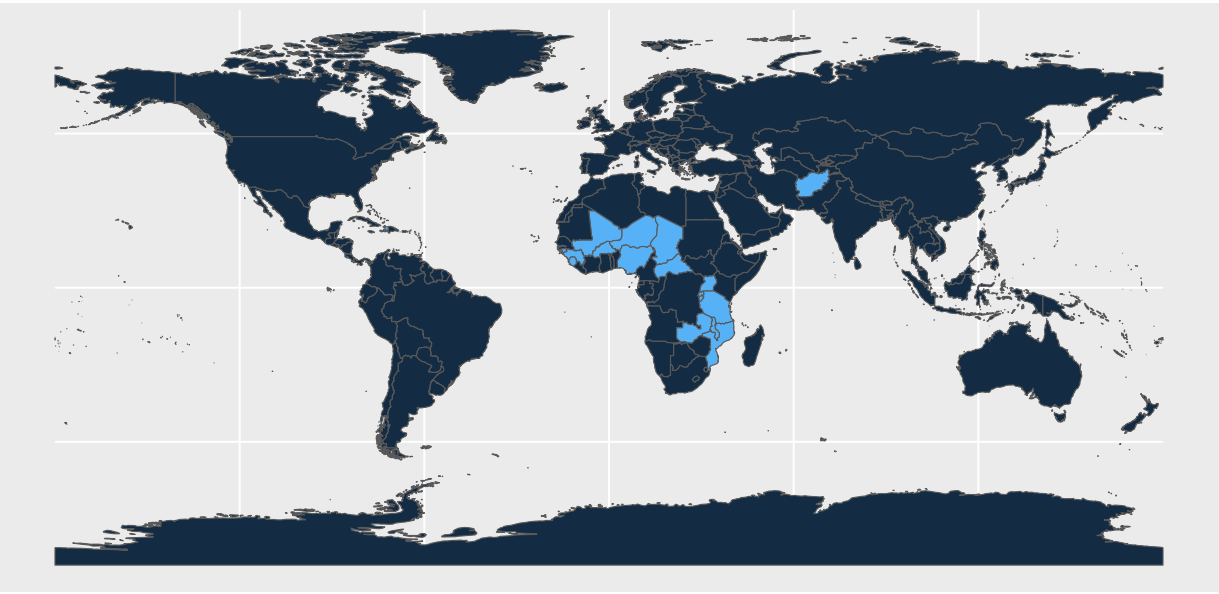
17. Sierra Leone
18. Tanzania
19. Uganda
20. Zambia

Carte des pays à aider en priorité

```
library(sf)
library(rnaturalearth)
# Espace de travail et data.frame
world <- rnaturalearth::ne_countries(scale = "medium", returnclass = "sf")
class(world)
```

```
## [1] "sf"         "data.frame"
```

```
liste_nom_pays <- c(liste_noms$` `,'Nigeria',"Congo, Rep"," Côte d'Ivoire") # pour merge
# data_pays_carte <- data.frame(
#   nom = liste_nom_pays,
#   y = 1
# )
world$pays_defav <- ifelse(world$name_sort %in% liste_nom_pays, 1, 0)

# Création de la carte
ggplot(data = world) + geom_sf(aes(fill=pays_defav)) + theme(legend.position="none")+
  labs(title = "Les pays défavorisés à aider en priorité")+
  theme(plot.title = element_text(hjust = 0.45))
```



Cette carte nous montre bien que les pays nécessiteux se trouve majoritairement en Afrique.

Améliorations possibles

Notre étude n'étant pas exhaustive, des améliorations peuvent être envisagées. Nous aurions pu ajouter d'autres variables explicatives afin d'étoffer les "profils" des pays. Par exemple, l'Indice de Développement Humain aurait pu être petinent comme indicateur ou encore une variable binaire comme certains droits (droit à l'avortemenent) ou l'accès à l'eau potable. Aussi, le partitionnement des pays lors de la réalisation d'une CAH est arbitraire et nous aurions pu prendre K = 2 pour aborder un point de vue différent.