Projet d'apprentissage non supervisé

Marie Guibert - Clémence Chesnais

2023-03-28

Environnement de travail

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stargazer)
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(gridExtra)
```

```
##
## Attachement du package : 'gridExtra'
##
## L'objet suivant est masqué depuis 'package:dplyr':
##
##     combine
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(cluster)
library(NbClust)
```

Question 1

Importation des données

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

```
## '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(data)
```

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.

STANDARDISATION ?

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

Statistiques descriptives

```
sum(is.na(data))
```

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

Le jeu de données ne présente pas de valeur manquante, nous n'avons pas besoin de faire de modification de ce point de vue.

Résumé des données :

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

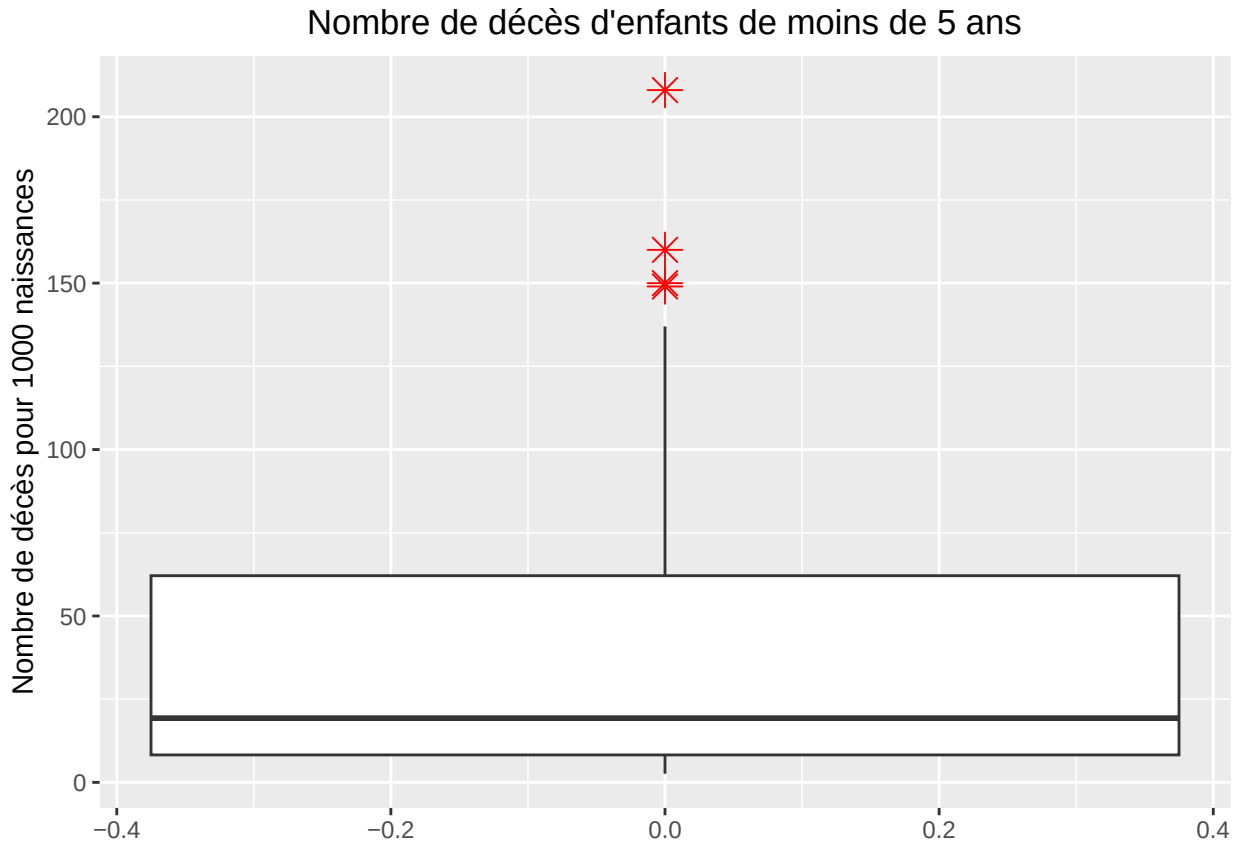
```
##
## Résumé des données
## ===================================================
## Statistic    N     Mean     St. Dev.   Min     Max
## ---------------------------------------------------
## enfant_mort 167   38.270     40.329   2.600  208.000
```

```
## 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.

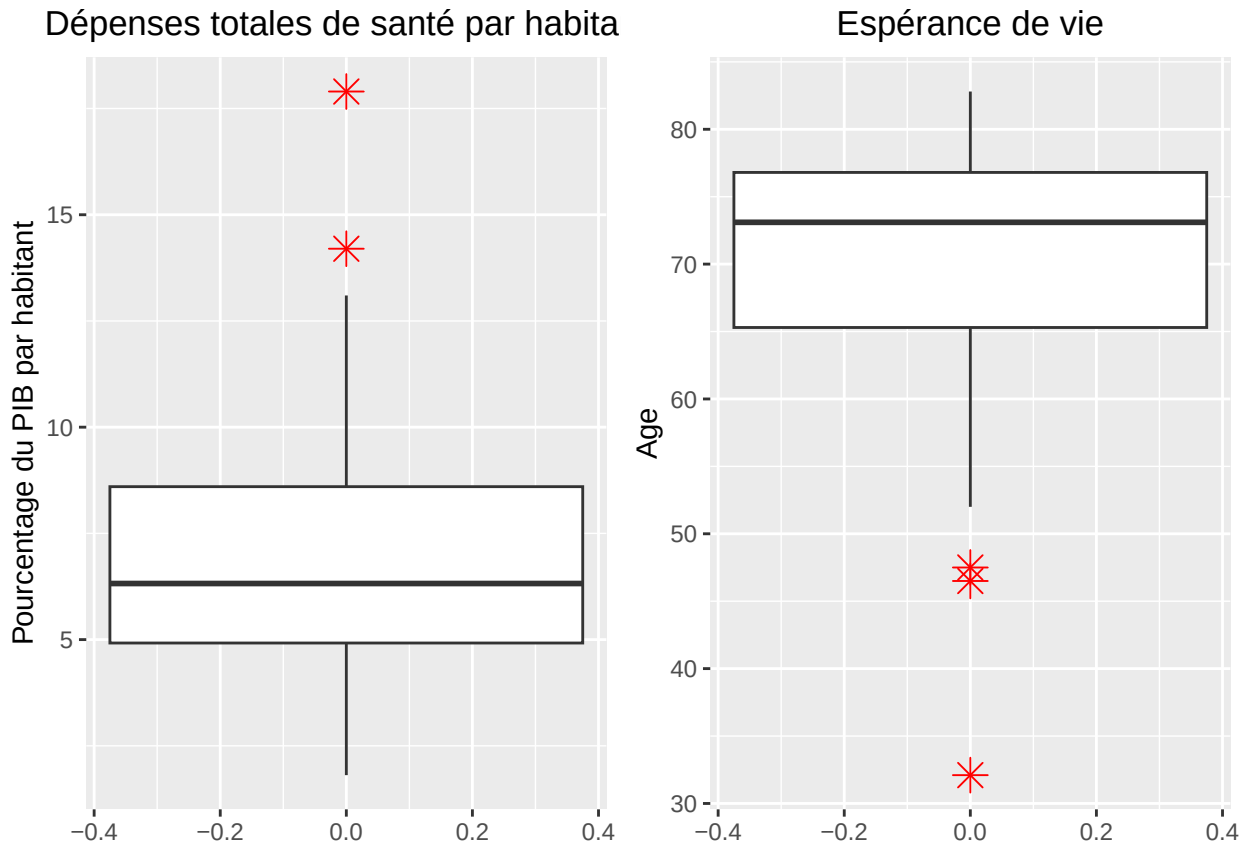
Graphiques :

```
ggplot(data=data, 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))
```

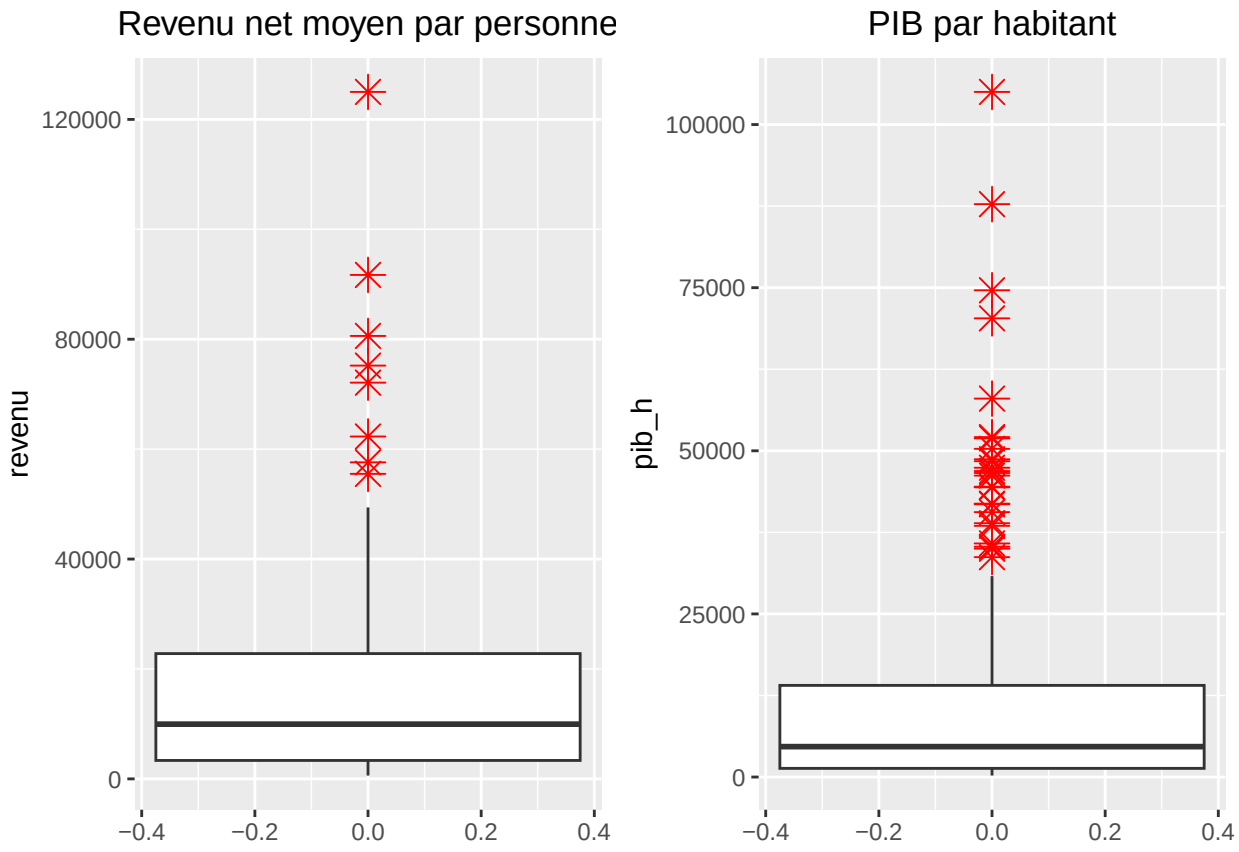



```
sante <- ggplot(data=data, aes(y=sante)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Dépenses totales de santé par habitant",y="Pourcentage du PIB par habitant")+
  theme(plot.title = element_text(hjust=0.5))
```

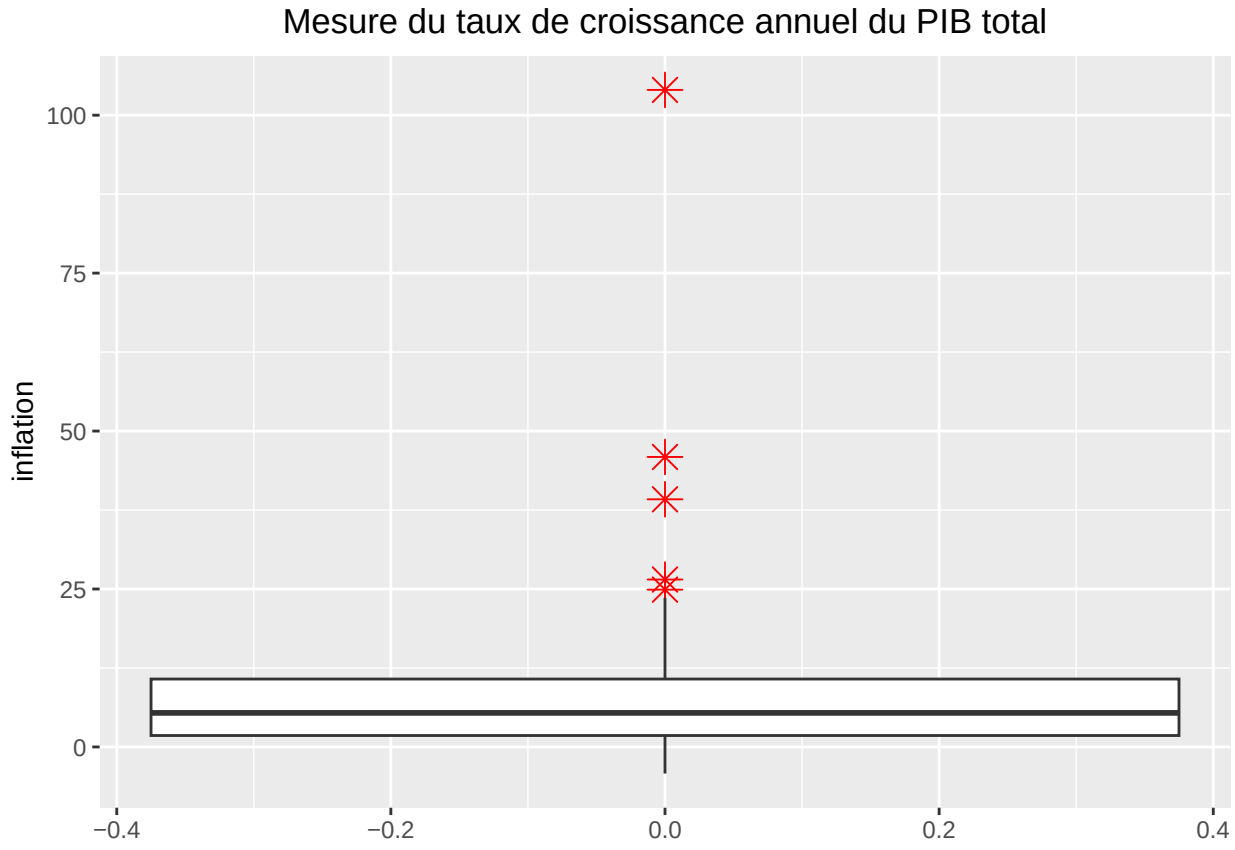
```
esperance <- ggplot(data=data, aes(y=esper_vie)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Espérance de vie",y="Age")+
  theme(plot.title = element_text(hjust=0.5))

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



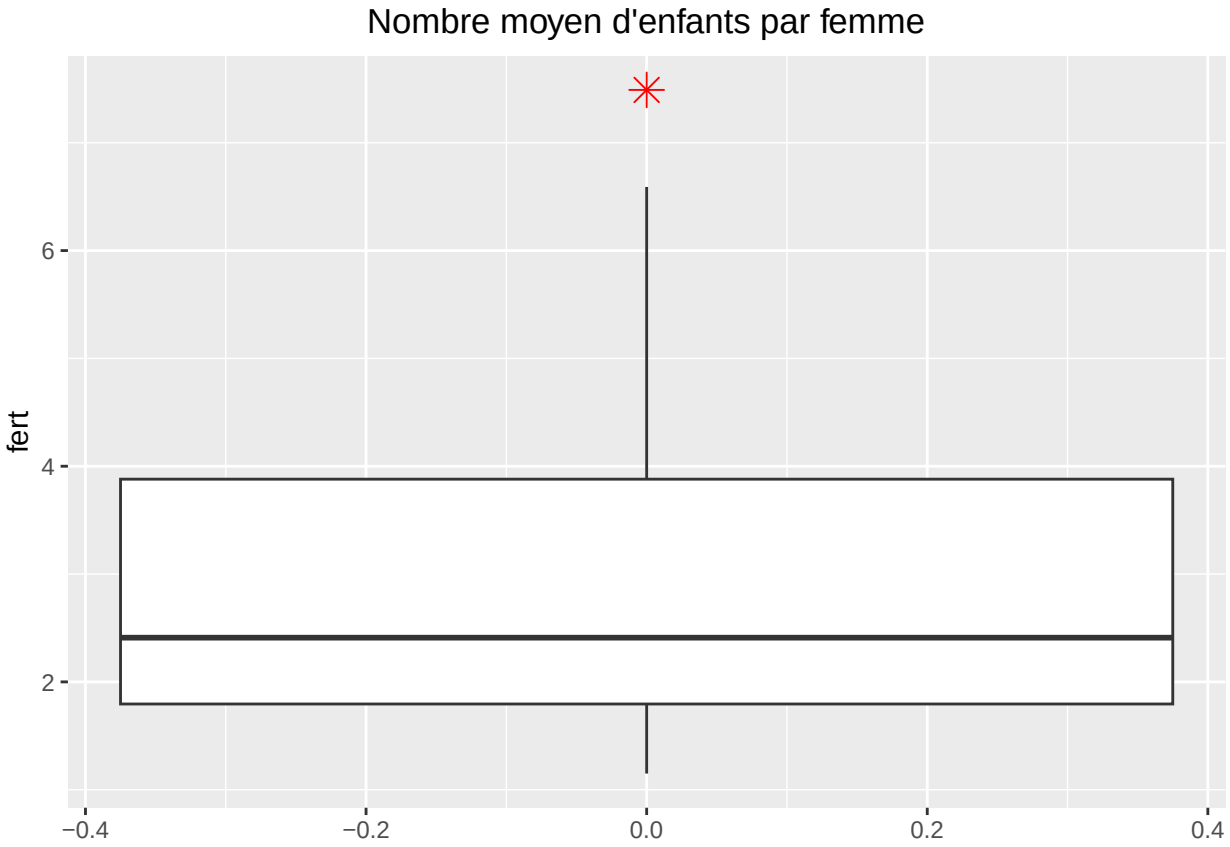

```
revenu_net <- ggplot(data=data, aes(y=revenu)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Revenu net moyen par personne")+
  theme(plot.title = element_text(hjust=0.5))

pib_hab <- ggplot(data=data, 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)
```




```
ggplot(data=data, aes(y=inflation)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Mesure du taux de croissance annuel du PIB total")+
  theme(plot.title = element_text(hjust=0.5))
```

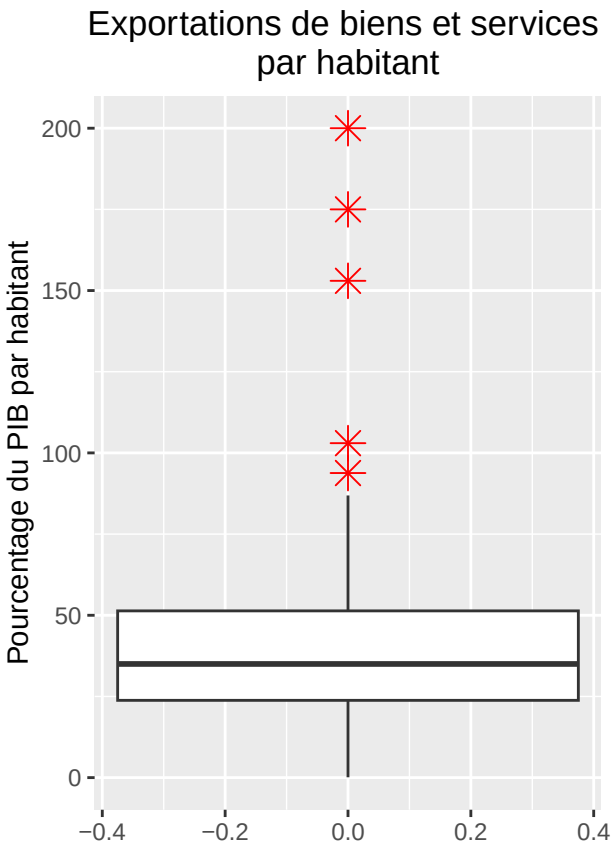


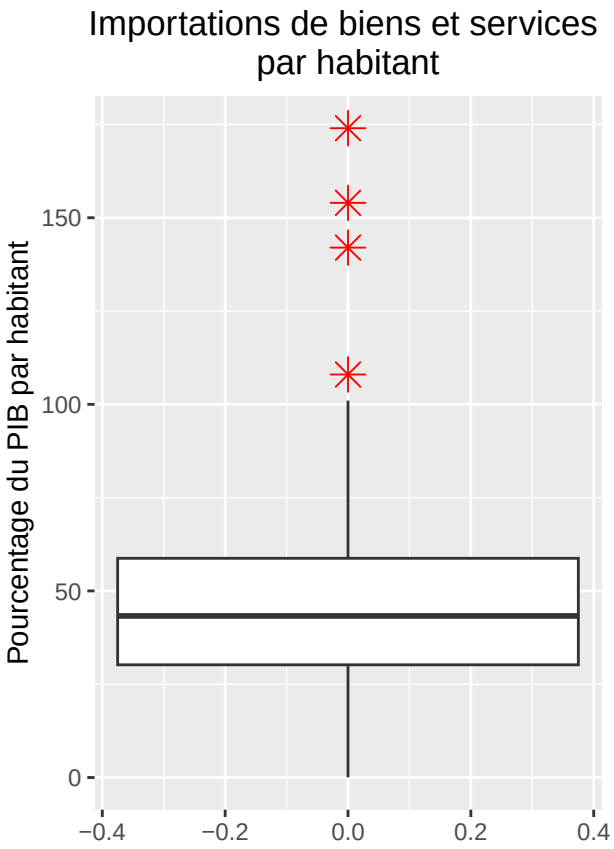

```
ggplot(data=data, aes(y=fert)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Nombre moyen d'enfants par femme")+
  theme(plot.title = element_text(hjust=0.5))
```



Imports et Exports :

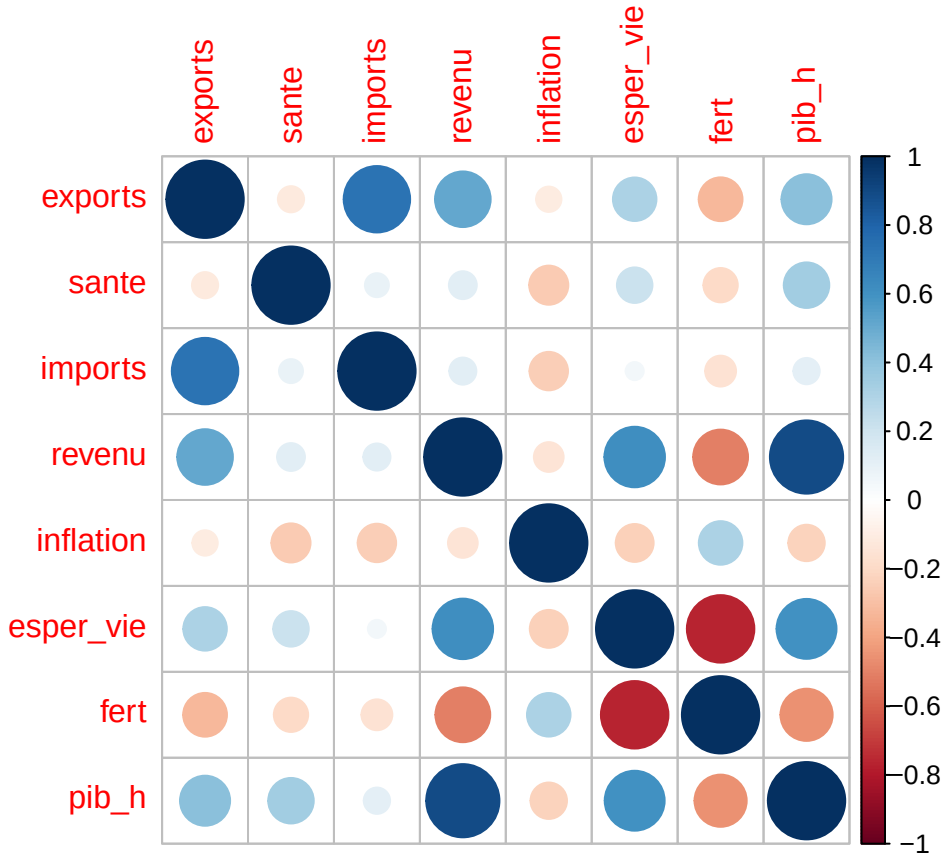
```
imports <- ggplot(data=data, aes(y=imports)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Importations de biens et services \npar habitant",y="Pourcentage du PIB par habitant")+
  theme(plot.title = element_text(hjust=0.5))

exports <- ggplot(data=data, aes(y=exports)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,outlier.size=4)+
  labs(title="Exportations de biens et services \npar habitant",y="Pourcentage du PIB par habitant")+
  theme(plot.title = element_text(hjust=0.5))

grid.arrange(exports, imports, ncol=2)
```





Matrice de corrélation :

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

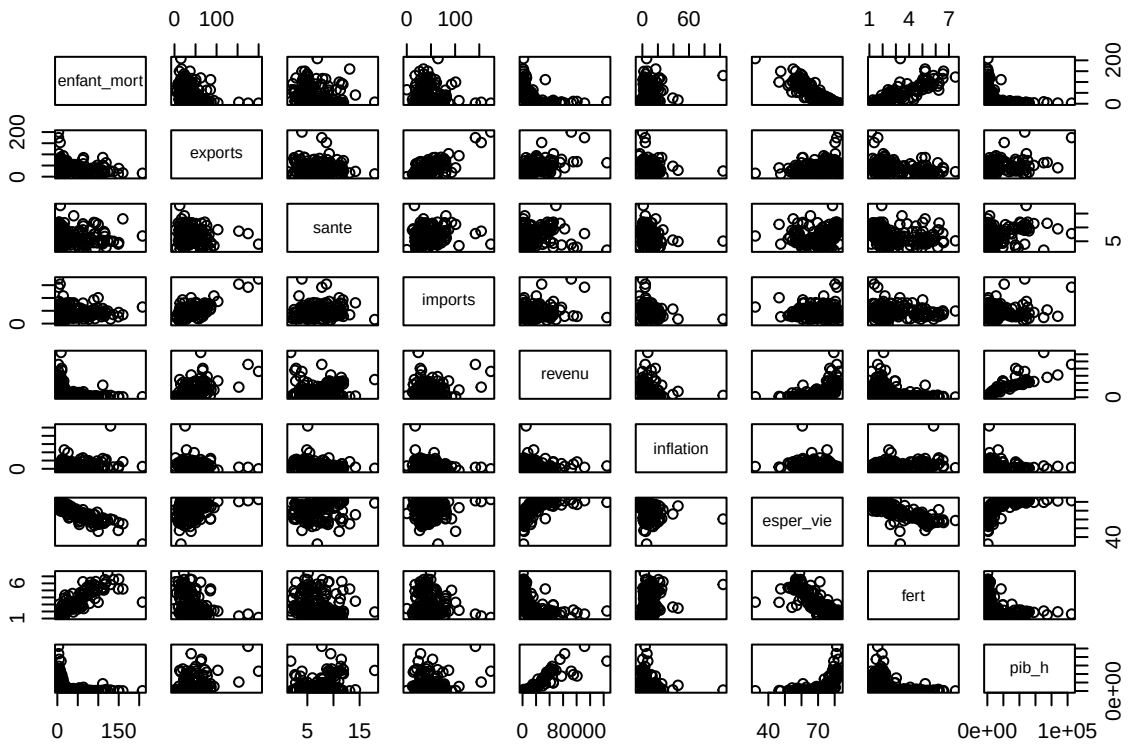
```
##                row col
## Guinea-Bissau   65  26
## Burkina Faso    26  65
```

Avec la distance euclidienne, de manhattan et mikowski, les deux pays les plus proches / similaires sont le Burkina Faso et la Guinée-Bissau.

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(data)
```

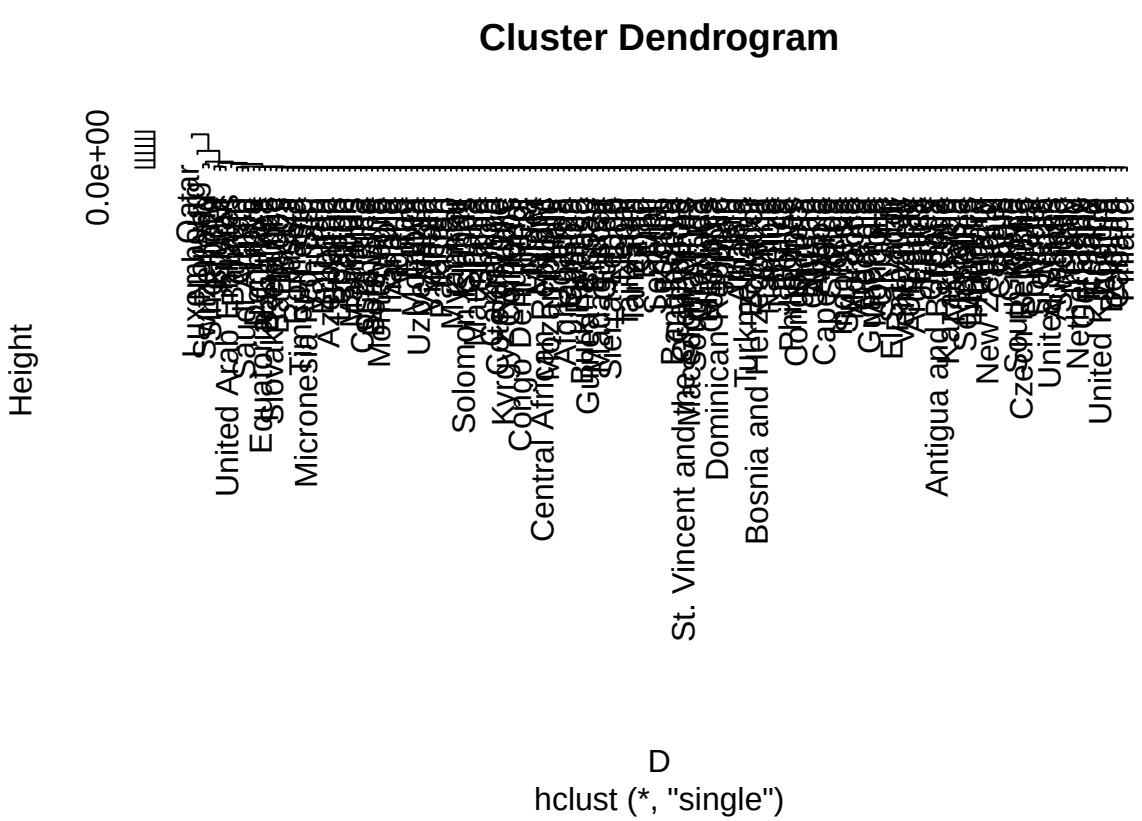


On calcule d'abord la distance euclidienne au carré :

```
D <- dist(data,method="euclidean")^2
```

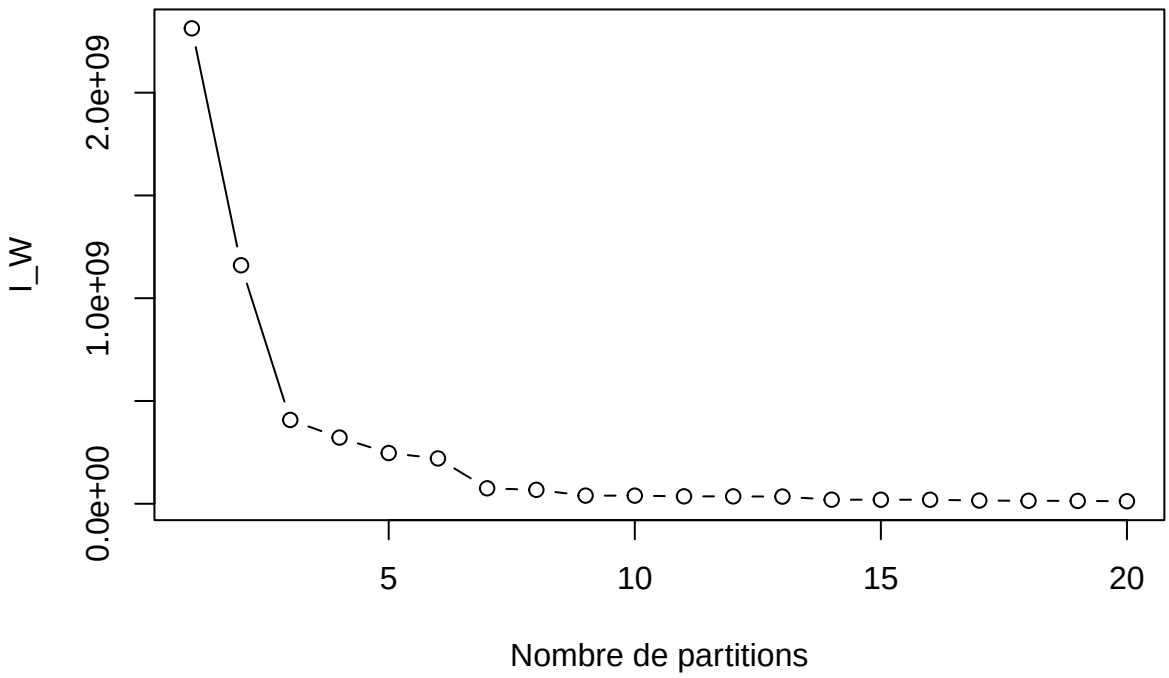
Méthode CAH avec le saut minimal (single linkage) :

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



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

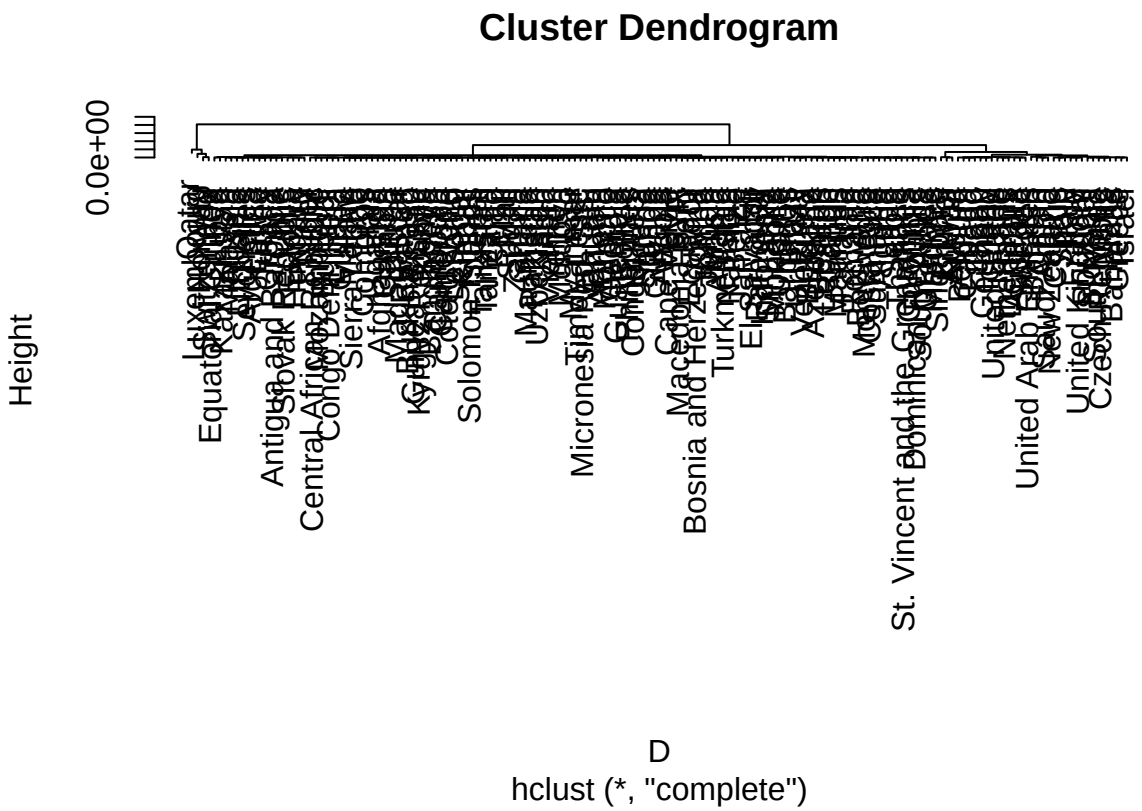
```
plot(rev(CAH_min$height)[1:20],type="b",xlab="Nombre de partitions",ylab="I_W")
```



Cependant, le tracé de la pertie d'inertie nous suggère de choisir une partition en 3 ou 4 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.

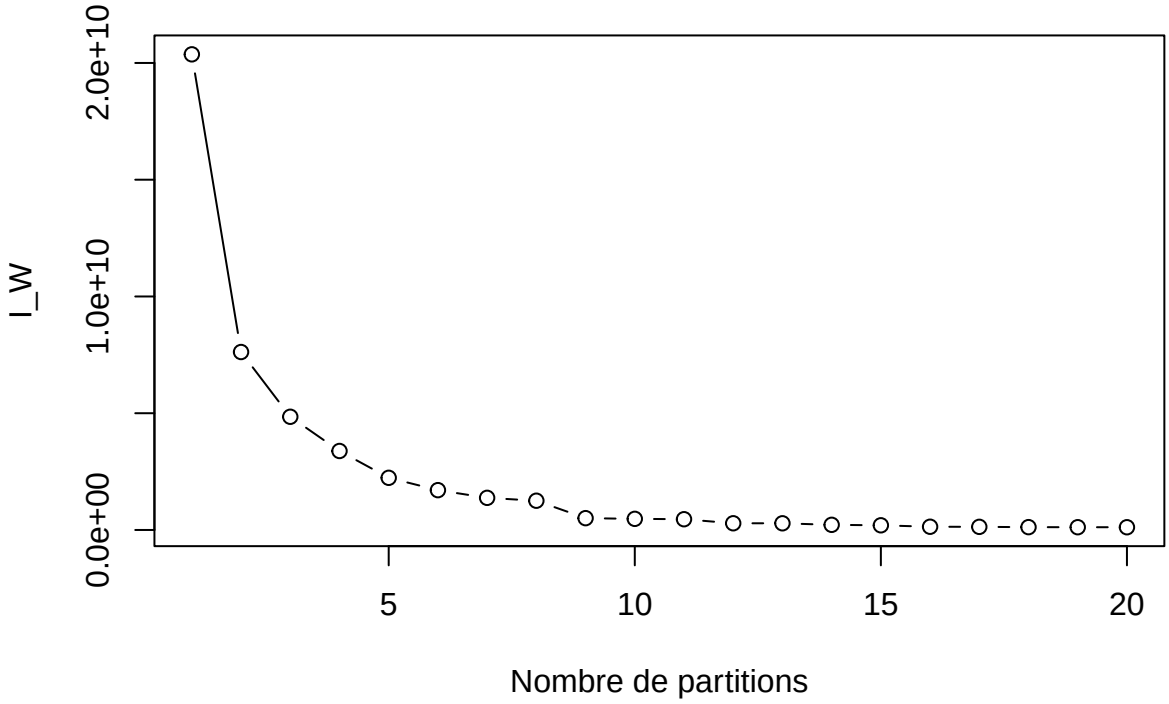
Faisons maintant les mêmes graphiques avec la méthode de distance de saut maximal (complet linkage):

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



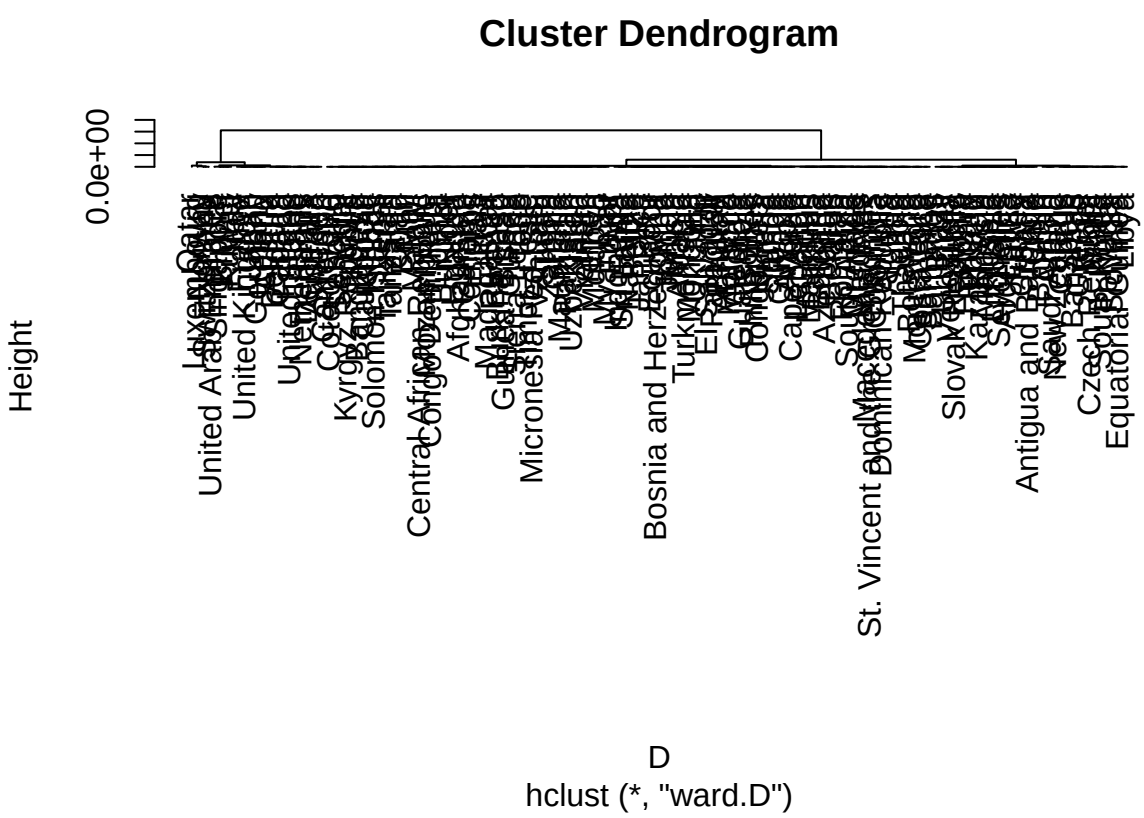
En analysant ce dendrogramme, nous pouvons distinguer 3 groupes de pays différents. De plus, le graphique suivant nous permet de confirmer cette hypothèse.

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



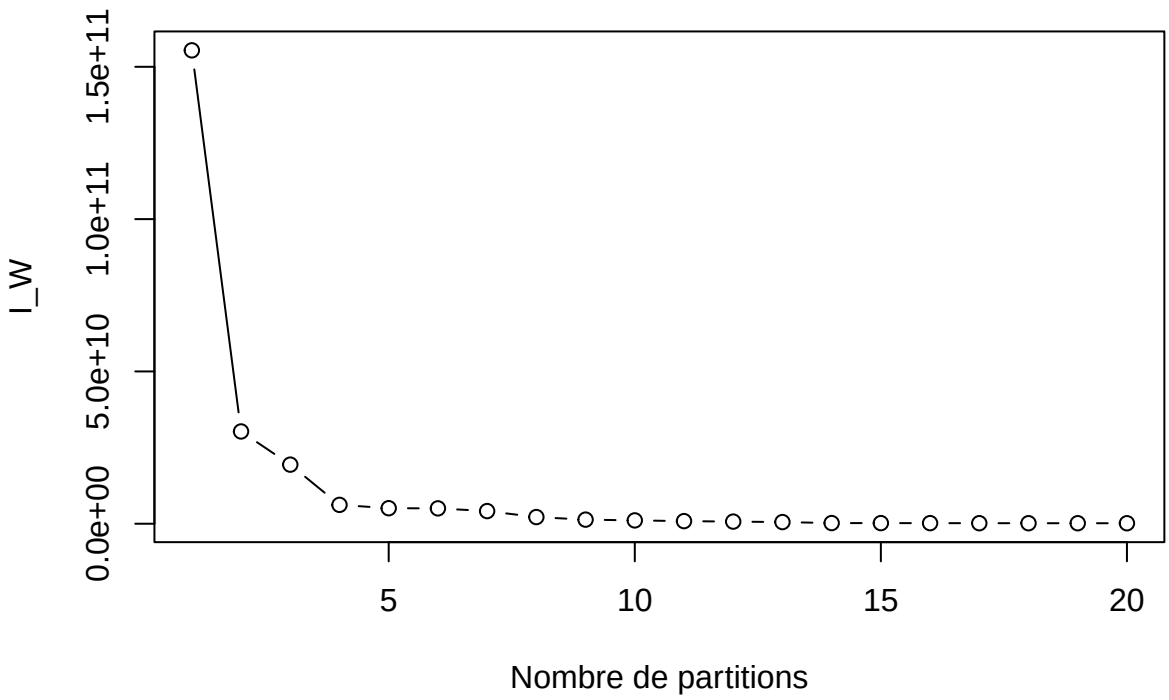
Enfin, avec la distance de ward, on obtient les résultats suivants:

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



Le tracé de la parte d'inertie nous permet de distinguer 3 groupes.

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

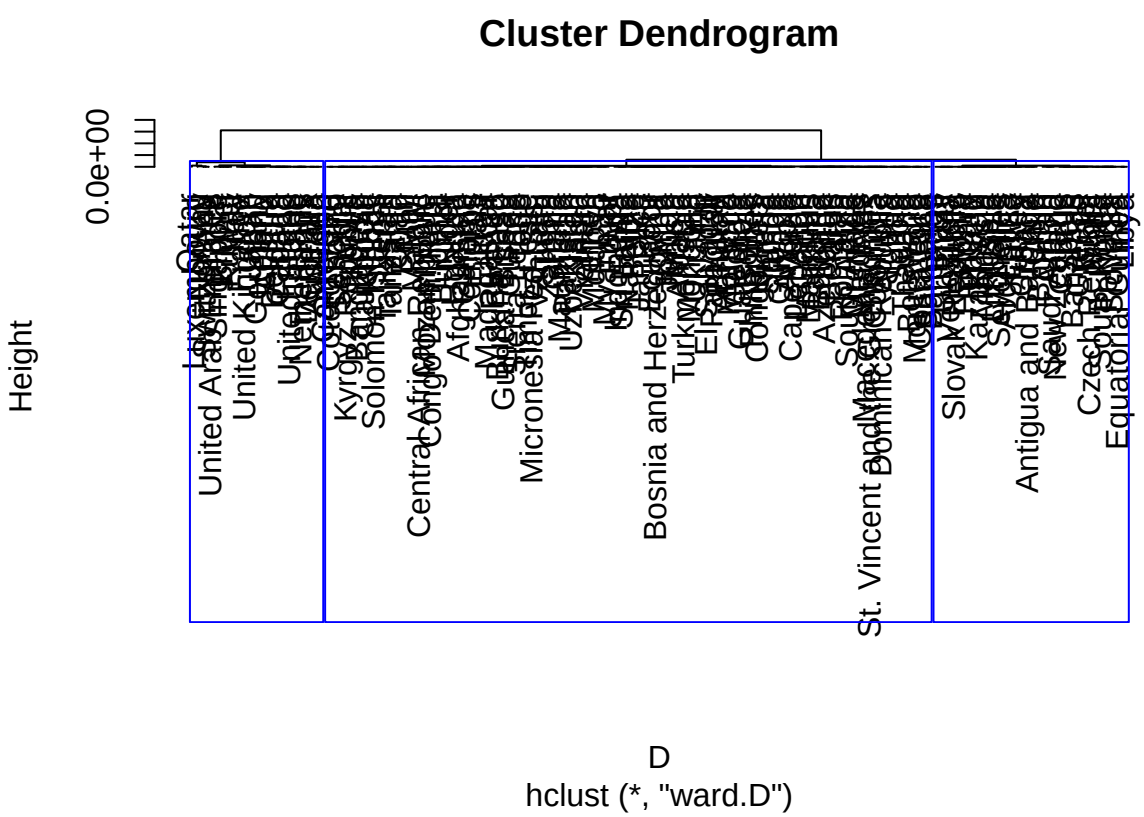


Ainsi, cette dernière classification ascendante hiérarchique nous permet de confirmer notre hypothèse de partition.

Représentation graphique des groupes avec un dendrogramme :

Ensuite, ce dendrogramme nous permet de distinguer clairement les groupes de pays. Notre hypothèse est donc valide.

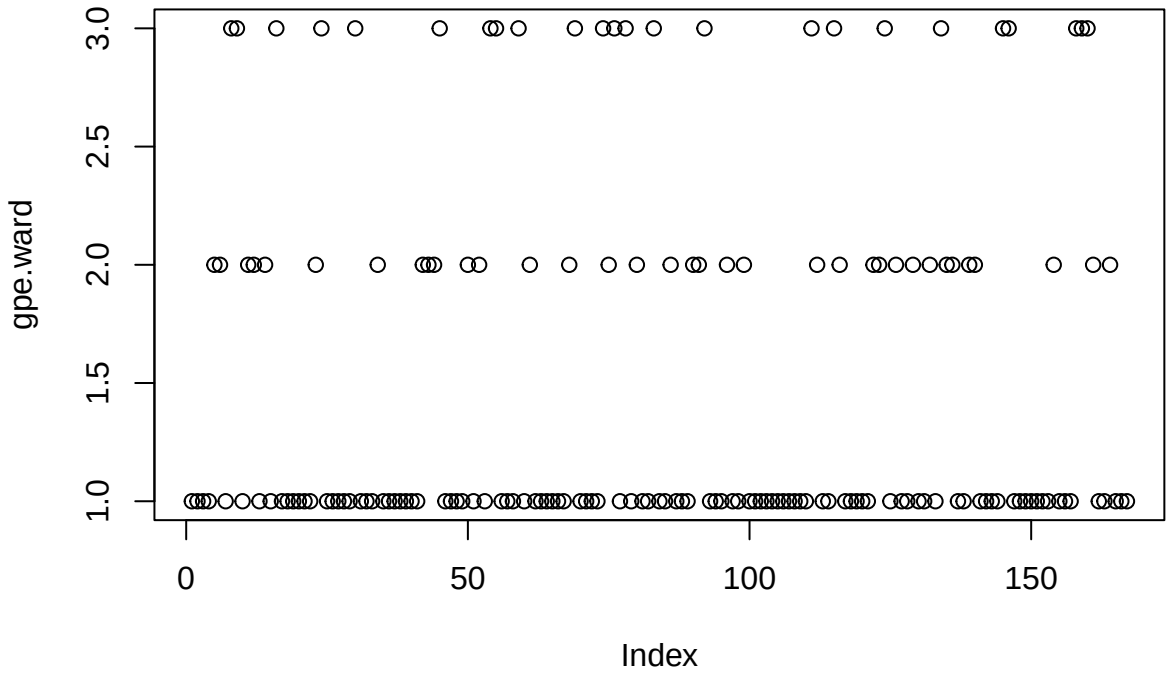
```
K=3
plot(CAH_ward,hang=-1)
rect.hclust(CAH_ward,K,border="blue")
```



Représentation graphique des clusters avec la fonction cutree :

La fonction cutree permet de faire apparaitre visuellement les groupes. Dans notre cas, on fixe K = 3 car nous avons choisi de réaliser une partition en 3 groupes.

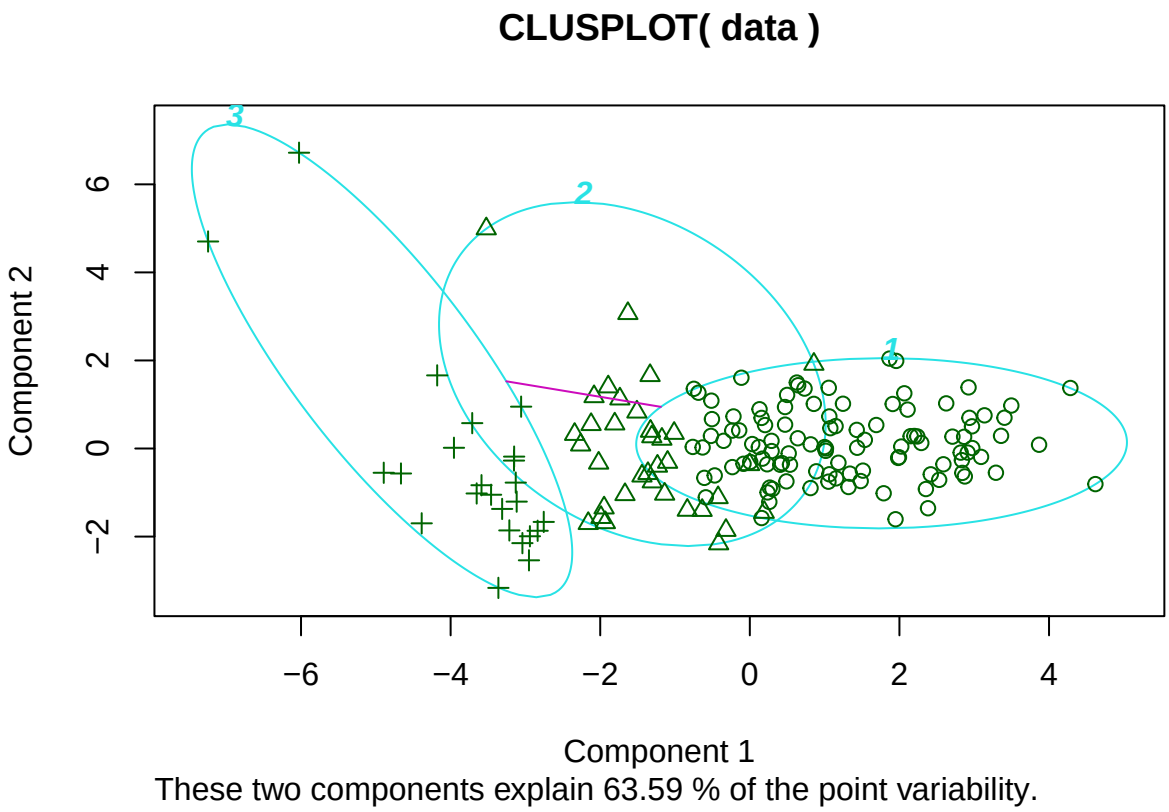
```
gpe.ward = cutree(CAH_ward,k=K)
plot(gpe.ward)
```



Sur ce graphique, on remarque correctement 3 groupes distincts.

Représentation des groupes avec la fonction clusplot :

```
data$gpe.ward <- gpe.ward
clusplot(data,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 3 groupes. En colorant les points avec leur vraie classe, nous pouvons observer que

A FINIR

Agrégation autour de centres mobiles :

```
c <- kmeans(x = data,3)
c
```

```
## K-means clustering with 3 clusters of sizes 35, 23, 109
##
## Cluster means:
##   enfant_mort  exports    sante  imports    revenu inflation esper_vie     fert
## 1   11.888571 52.08286 6.958000 49.00571 26125.714  6.503714  76.29143 1.973143
## 2    5.082609 61.64783 8.973043 50.36957 53991.304  3.233130  80.42174 1.805217
## 3   53.744037 33.25137 6.314771 45.47675  6485.899  9.152055  66.63211 3.502110
##       pib_h gpe.ward
## 1 18135.429 2.028571
## 2 51960.870 3.000000
## 3  3074.991 1.009174
##
## Clustering vector:
```

```
##                 Afghanistan                     Albania
##                           3                           3
##                     Algeria                     Angola
##                           3                           3
##          Antigua and Barbuda                   Argentina
##                           1                           1
##                     Armenia                   Australia
##                           3                           2
##                     Austria                  Azerbaijan
##                           2                           3
##                     Bahamas                     Bahrain
##                           1                           1
##                  Bangladesh                    Barbados
##                           3                           1
##                     Belarus                     Belgium
##                           3                           2
##                      Belize                       Benin
##                           3                           3
##                      Bhutan                     Bolivia
##                           3                           3
##       Bosnia and Herzegovina                    Botswana
##                           3                           3
##                      Brazil                      Brunei
##                           3                           2
##                    Bulgaria                Burkina Faso
##                           3                           3
##                     Burundi                    Cambodia
##                           3                           3
##                    Cameroon                      Canada
##                           3                           2
##                  Cape Verde      Central African Republic
##                           3                           3
##                        Chad                       Chile
##                           3                           1
##                       China                    Colombia
##                           3                           3
##                     Comoros              Congo Dem. Rep.
##                           3                           3
##                  Congo Rep.                  Costa Rica
##                           3                           3
##               Cote d'Ivoire                     Croatia
##                           3                           1
##                      Cyprus                Czech Republic
##                           1                           1
##                     Denmark            Dominican Republic
##                           2                           3
##                     Ecuador                       Egypt
##                           3                           3
##                 El Salvador             Equatorial Guinea
##                           3                           1
##                     Eritrea                     Estonia
##                           3                           1
##                        Fiji                     Finland
##                           3                           2
```

```
##                  France                   Gabon
##                       2                       3
##                  Gambia                 Georgia
##                       3                       3
##                 Germany                   Ghana
##                       2                       3
##                  Greece                 Grenada
##                       1                       3
##               Guatemala                  Guinea
##                       3                       3
##           Guinea-Bissau                  Guyana
##                       3                       3
##                   Haiti                 Hungary
##                       3                       1
##                 Iceland                   India
##                       2                       3
##               Indonesia                    Iran
##                       3                       3
##                    Iraq                 Ireland
##                       3                       2
##                  Israel                   Italy
##                       1                       1
##                 Jamaica                   Japan
##                       3                       2
##                  Jordan              Kazakhstan
##                       3                       1
##                   Kenya                Kiribati
##                       3                       3
##                  Kuwait          Kyrgyz Republic
##                       2                       3
##                     Lao                  Latvia
##                       3                       1
##                 Lebanon                 Lesotho
##                       3                       3
##                 Liberia                   Libya
##                       3                       1
##              Lithuania                Luxembourg
##                       1                       2
##           Macedonia FYR                Madagascar
##                       3                       3
##                  Malawi                Malaysia
##                       3                       1
##                Maldives                    Mali
##                       3                       3
##                   Malta                Mauritania
##                       1                       3
##               Mauritius       Micronesia Fed. Sts.
##                       3                       3
##                 Moldova                Mongolia
##                       3                       3
##              Montenegro                 Morocco
##                       3                       3
##              Mozambique                 Myanmar
##                       3                       3
```

```
##                   Namibia                     Nepal
##                         3                         3
##               Netherlands                New Zealand
##                         2                         1
##                     Niger                   Nigeria
##                         3                         3
##                    Norway                      Oman
##                         2                         1
##                  Pakistan                    Panama
##                         3                         3
##                  Paraguay                      Peru
##                         3                         3
##               Philippines                    Poland
##                         3                         1
##                  Portugal                     Qatar
##                         1                         2
##                   Romania                    Russia
##                         3                         1
##                    Rwanda                     Samoa
##                         3                         3
##              Saudi Arabia                   Senegal
##                         1                         3
##                    Serbia                  Seychelles
##                         3                         1
##              Sierra Leone                 Singapore
##                         3                         2
##            Slovak Republic                  Slovenia
##                         1                         1
##            Solomon Islands              South Africa
##                         3                         3
##                South Korea                     Spain
##                         1                         1
##                  Sri Lanka St. Vincent and the Grenadines
##                         3                         3
##                     Sudan                  Suriname
##                         3                         3
##                    Sweden                Switzerland
##                         2                         2
##                 Tajikistan                  Tanzania
##                         3                         3
##                   Thailand                Timor-Leste
##                         3                         3
##                      Togo                     Tonga
##                         3                         3
##                   Tunisia                    Turkey
##                         3                         1
##              Turkmenistan                    Uganda
##                         3                         3
##                   Ukraine          United Arab Emirates
##                         3                         2
##            United Kingdom             United States
##                         2                         2
##                   Uruguay                Uzbekistan
##                         1                         3
```

```
##                         Vanuatu                       Venezuela
##                               3                               1
##                         Vietnam                           Yemen
##                               3                               3
##                          Zambia
##                               3
##
## Within cluster sum of squares by cluster:
## [1]  4134801203 17074173287  3244806848
##  (between_SS / total_SS =  79.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

La fonction **kmeans** nous permet d'obtenir le partitionnement final. Dans notre cas, elle nous rend 3 clusters composés de 35, 109 et 23 individus.