République de la Cote d'Ivoire

Union-Discipline-Travail

Ministère de l'enseignement supérieur et de la recherche scientifique

Année academique

2023-2024

Projet ACP (Analyse en composantes principales).

L'organisation non gouvernementale (ONG) HELP International a pu récolter environ 10 millions de dollars. Comment utiliser cet argent de manière stratégique et efficace pour assister les pays qui ont le plus besoin d'aide sur le plan du développement socio-économique et sanitaire ?

INSSEDS

Institut Supérieur de Statistique, d'Econométrie et de Datas Sciences

AVANT PROPOS

Dans le cadre de ce rapport, nous présentons les résultats d'une Analyse en Composantes Principales (**ACP**) menée dans le contexte d'un projet visant à aider HELP International à identifier les pays les plus nécessiteux pour une allocation efficace de ressources. HELP International, une organisation non gouvernementale (**ONG**), a réuni environ 10 millions de dollars et cherche à les utiliser de manière stratégique pour soutenir le développement socio-économique et sanitaire dans différents pays.

L'objectif principal de cette étude est d'identifier les facteurs socio-économiques et sanitaires qui déterminent le niveau de développement global des pays, afin de conseiller le PDG de HELP International dans le choix des pays prioritaires pour l'octroi d'aide. Pour ce faire, nous avons effectué une Analyse en Composantes Principales sur un ensemble de données comprenant plusieurs indicateurs socio-économiques et sanitaires pour un échantillon de pays.

Ce rapport présente en détail les différentes étapes de l'ACP, y compris la préparation des données, l'interprétation des composantes principales, l'analyse des clusters formés et la recommandation des pays prioritaires pour l'aide. Nous discutons également des implications de nos résultats et des recommandations pour les futures actions de HELP International.

Nous tenons à exprimer notre gratitude envers l'équipe de HELP International pour avoir confié cette tâche importante. Nous espérons que les résultats présentés dans ce rapport seront utiles pour guider les décisions stratégiques de l'organisation et contribuer à son objectif louable d'améliorer les conditions de vie dans les pays les plus vulnérables à travers le monde.

Notre remerciement à l'expert Mr. AKPOSSO DIDIER MARTIAL pour la qualité de formation qui nous a permit à réaliser ce travail.

Fait à Abidjan le 06/03/2024 par NGAGABA Bertrand GUERI

TABLE DES MATIERES

SOMMAIRE

Introduction

L'organisation non gouvernementale (**ONG**) HELP International a réussi à collecter environ 10 millions de dollars, une somme considérable destinée à soutenir des initiatives humanitaires à travers le monde. Face à ce financement, le PDG de l'organisation est confronté à la délicate tâche de déterminer la meilleure stratégie pour l'utiliser de manière efficace et ciblée. Pour ce faire, il doit identifier les pays qui présentent les besoins les plus pressants en termes d'aide humanitaire. En tant que data scientiste, notre mission consiste à analyser et à catégoriser les pays en fonction de divers facteurs socio-économiques et sanitaires qui influent sur leur développement global. Cette classification permettra ensuite de recommander au Président Directeur General (**PDG**) les pays sur lesquels concentrer les efforts de l'organisation non gouvernementale (**ONG**) afin d'avoir un impact significatif et durable sur les populations les plus vulnérables.

Objectif

Catégoriser les pays en fonction des facteurs socio-économiques et sanitaires qui déterminent le développement global du pays.

PARTIE I : TRAITEMENT DES DONNEES BRUTES

Cette partie présente le dictionnaire des données utilisées dans la présente analyse ainsi que les étapes de l'apurement de notre base de données. A cela, il faut ajouter que la méthodologie est basée sur une approche essentiellement descriptive en utilisant le logiciel **R** pour le traitement et l'analyse des données.

Dictionnaire des données de l'analyse

Le jeu de données s'appellera "**Help_I**". Il comporte des variables tant quantitatives que qualitatives. Ces variables sont listées dans le tableau de dictionnaire de données ci-dessous :

***Tableau 1*: Dictionnaire des données.**

VARIABLE	NATURE	DESCRIPTION	MODALITES
Pays	Qualitative	Nom du pays	Chaine de caractère
Enfant_mort	Quantitative	Décès d'enfants de moins de 5 ans pour 1000 naissances vivantes	Numerique decimal
exportations	Quantitative	Exportations de biens et services par habitant. Donné en pourcentage du PIB par habitant	Numerique decimal
dep_sante	Quantitative	Dépenses totales de santé par habitant. Données en pourcentage du PIB par habitant	Numerique decimal
importations	Quantitative	Importations de biens et services par habitant. Donné en pourcentage du PIB par habitant	Numerique decimal
revenu	Quantitative	Revenu net par personne	Numerique entier
taux_croissance	Quantitative	La mesure du taux d'inflation annuel du PIB total	Numerique decimal

VARIABLE	NATURE	DESCRIPTION	MODALITES
esperance_vie	Quantitative	Le nombre moyen d'années qu'un nouveau-né vivrait si les tendances de mortalité actuelles devaient rester les mêmes	Numerique decimal
total_fertilite	Quantitative	Le nombre d'enfants qui naîtraient à chaque femme si les taux de fécondité par âge actuels devaient rester les mêmes.	Numerique decimal
pib_par_hab	Quantitative	Le PIB par habitant. Calculé comme le PIB total divisé par la population totale.	Numerique entier

Importation du jeu de donées

Nous avons commencé par charger le jeu de données depuis son espace de stockage sur notre ordinateur avant de le lire. Cette base de données sera appelée « **Help_I** ». Par la suite, nous avons choisi d'en afficher un aperçu. Le jeu de données global contient 167 observations reparties sur 10 variables dont 9 de type quantitatifs et 1 de type qualitatif représentant les individus avec 3 modalités.

***Tableau 2*: Les 6 premières lignes du jeu de données.**

	enfant_mort	exportations	dep_sante	importations	revenu	taux_croissance	life_expec	total_fertilite	pib_par_hab
Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
Argentina	14.5	18.9	8.10	16.0	18700	20.90	75.8	2.37	10300

***Tableau 3*: Résumé des indicateurs et des valeurs manquantes (notée NAs).**

enfant_mort	exportations	dep_sante	importations	revenu	taux_croissance	life_expec	total_fertilite	pib_par_hab
Min. : 2.60	Min. : 0.109	Min. : 1.810	Min. : 0.0659	Min. : 49.6	Min. : -4.21	Min. : 2.04	Min. : 1.15	Min. : 1.47
1st Qu.: 7.85	1st Qu.: 23.800	1st Qu.: 4.950	1st Qu.: 28.8500	1st Qu.: 3170.0	1st Qu.: 1.81	1st Qu.:64.30	1st Qu.: 1.82	1st Qu.: 1310.00
Median : 19.20	Median : 35.000	Median : 6.330	Median : 42.9000	Median : 9940.0	Median : 5.45	Median :73.10	Median : 2.41	Median : 4660.00
Mean : 37.80	Mean : 41.353	Mean : 7.761	Mean : 45.6223	Mean : 17023.2	Mean : 130.49	Mean :69.30	Mean : 4.39	Mean : 12901.54

enfant_mort	exportations	dep_sante	importations	revenu	taux_croissance	life_expec	total_fertilite	pib_par_hab
3rd Qu.: 60.40	3rd Qu.: 51.400	3rd Qu.: 8.865	3rd Qu.: 58.0500	3rd Qu.: 22800.0	3rd Qu.: 11.15	3rd Qu.:76.80	3rd Qu.: 4.16	3rd Qu.: 14050.00
Max. :208.00	Max. :200.000	Max. :85.100	Max. :174.0000	Max. :125000.0	Max. :11400.00	Max. :82.80	Max. :74.00	Max. :105000.00
NA's :4	NA	NA	NA	NA	NA	NA	NA	NA

Le résumé sommaire de ce jeu de données permet de détecter qu'il existe des données manquantes matérialisées par des valeurs NAs. Ces valeurs doivent passer par une étape de traitement.

Traitement des valeurs manquantes dans les données

Cette section va servir à afficher les observations qui contiennent des valeurs manquantes. Elle servira également à réaliser tout le traitement qui y convient.

Proportion des valeurs manquantes dans les données

Elle permet d'obtenir des informations sur la structure de données, c'est-à-dire leurs dimensions, les class (ou types de variables), la présence de données manquantes. Il existe 4 valeurs manquantes sur l'ensemble des données. Et le taux global de valeurs manquantes est de 2.4%.

***Tableau 4*:Proportion de valeurs manquantes de chaque variable.**

Variables	Nombre	Proportion
enfant_mort	4	0.0239521
exportations	0	0.0000000
dep_sante	0	0.0000000
importations	0	0.0000000
revenu	0	0.0000000
taux_croissance	0	0.0000000
life_expec	0	0.0000000
total_fertilite	0	0.0000000
pib_par_hab	0	0.0000000

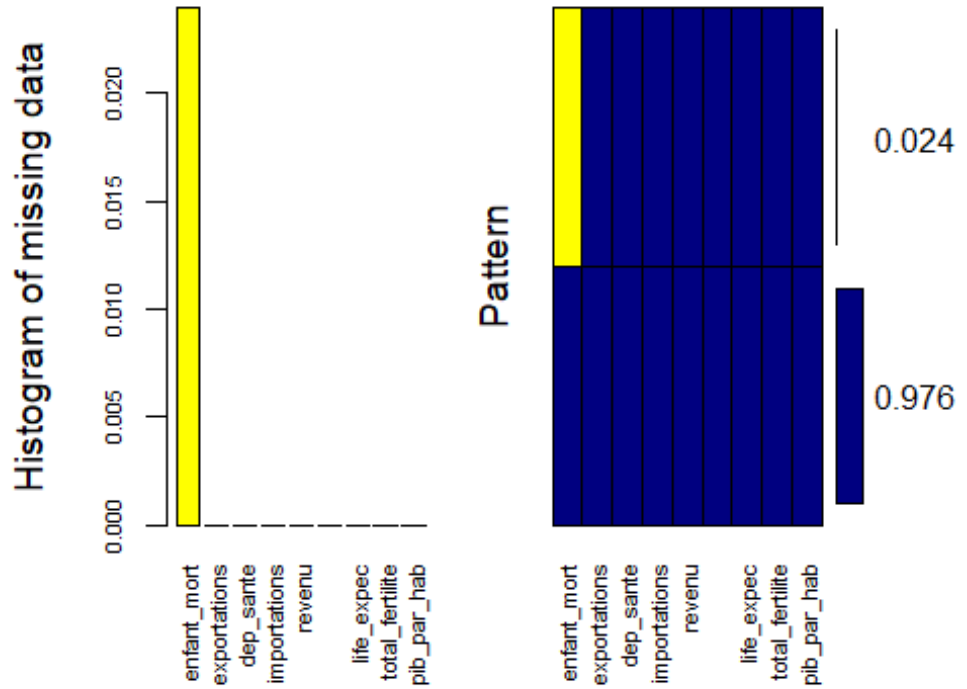
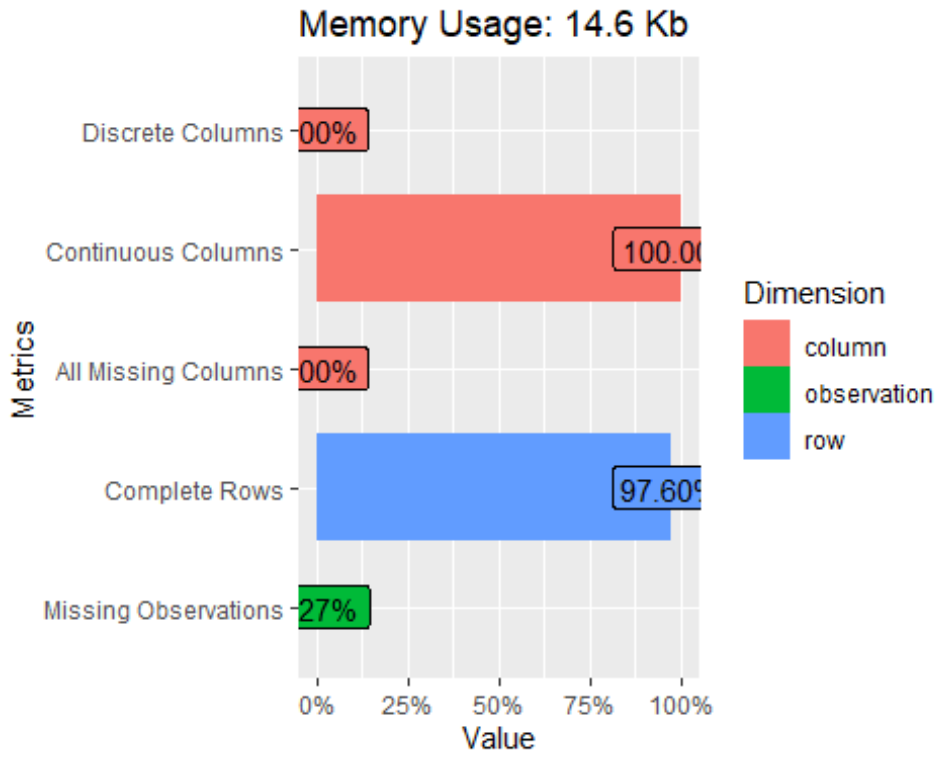
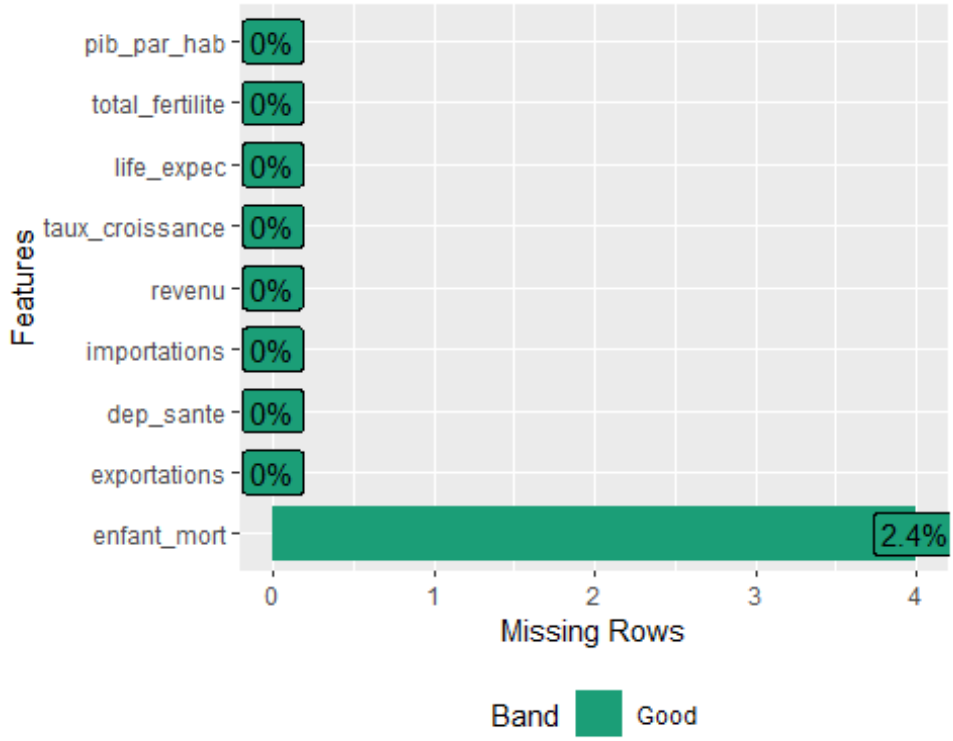


***Figure 1**: Visualisation des valeurs manquantes.*

Notre jeu de données à une capacité de 36,3Kb. En le visualisant, l'on se rend compte qu'en moyenne c'est 2.4% de données manquantes sur chacune des variables. La variable « enfant_mort » contient à elle seule ses 2.4% des données manquantes. Ce constat montre que, la perte des données sur chaque individu est très importante et assez significative dans l'ensemble pour ne pas nous donner le luxe de les supprimer.

Et puisque la technique de l'analyse en composantes principales (**ACP**) ne tolère pas les données manquantes, il convient donc de procéder au traitement de ces valeurs pour leur imputation, plutôt que de les supprimer de notre base de données.


Memory Usage: 14.6 Kb
Discrete Columns
Continuous Columns
All Missing Columns
Complete Rows
Missing Observations
Metrics
0%
25%
50%
75%
100%
Value
Dimension
column
observation
row


pib_par_hab
total_fertilite
life_expec
taux_croissance
revenu
importations
dep_sante
exportations
enfant_mort
0%
2.4%
Features
0
1
2
3
4
Missing Rows
Band
Good

Imputation de données manquantes par les K plus proches voisins

Nous avons d'abord dupliqué le jeu de données en le renommant "**Help_I**".

Pour l'imputation des valeurs manquantes parmi les variables quantitatives, nous avons utilisé la technique des K plus proches voisins consistant à remplir les valeurs manquantes en explorant les similitudes entre les cas. Ceci du fait que nous essayons de trouver les valeurs les plus probables pour chacune de ces inconnues. Pour mesurer donc cette proximité entre les observations, nous avons opté pour l'application d'une fonction de similarité reposant sur un calcul de distance. Cette fonction qui calcule la distance entre deux observations estime l'affinité entre les observations comme ceci : « Plus deux points sont proches l'un de l'autre, plus ils sont similaires. » Nous avons donc conservé les 10 observations du jeu de données qui sont les plus « proches » des observations à prédire. Par la suite nous nous sommes évertués à retrouver à quelle famille appartient les nouvelles données, en cherchant la famille sinon la classe majoritaire parmi les k données. Et nous avons retourné la valeur calculée comme étant la valeur qui a été prédite pour l'observation en entrée qui était inconnue.

Ci-dessous un bref aperçu des nouvelles données quantitatives après imputation des valeurs manquantes.

***Tableau 5*: Quantité de données après avoir imputer**

Individus/ Variables	enfant_mort	exportations	dep_sante	importations	revenu	taux_croissance	life_expec	total_fertilite	pib_par_hab
Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
Argentina	14.5	18.9	8.10	16.0	18700	20.90	75.8	2.37	10300

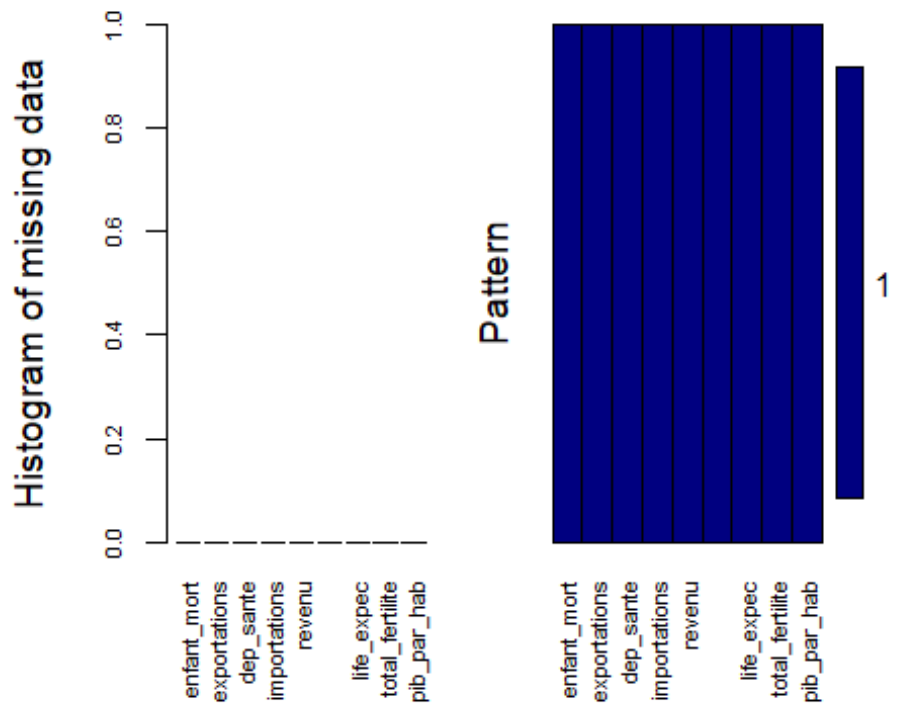


***Figure 2**: Visualisation des données après l'imputation.*

***Tableau 6**: Résumé des indicateurs après le traitement.*

enfant_mort	exportations	dep_sante	importations	revenu	taux_croissance	life_expec	total_fertilite	pib_par_hab
Min. : 2.60	Min. : 0.109	Min. : 1.810	Min. : 0.0659	Min. : 49.6	Min. : -4.21	Min. : 2.04	Min. : 1.15	Min. : 1.47
1st Qu.: 8.25	1st Qu.: 23.800	1st Qu.: 4.950	1st Qu.: 28.8500	1st Qu.: 3170.0	1st Qu.: 1.81	1st Qu.:64.30	1st Qu.: 1.82	1st Qu.: 1310.00
Median : 19.70	Median : 35.000	Median : 6.330	Median : 42.9000	Median : 9940.0	Median : 5.45	Median :73.10	Median : 2.41	Median : 4660.00
Mean : 39.25	Mean : 41.353	Mean : 7.761	Mean : 45.6223	Mean : 17023.2	Mean : 130.49	Mean :69.30	Mean : 4.39	Mean : 12901.54
3rd Qu.: 62.40	3rd Qu.: 51.400	3rd Qu.: 8.865	3rd Qu.: 58.0500	3rd Qu.: 22800.0	3rd Qu.: 11.15	3rd Qu.:76.80	3rd Qu.: 4.16	3rd Qu.: 14050.00
Max. :208.00	Max. :200.000	Max. :85.100	Max. :174.0000	Max. :125000.0	Max. :11400.00	Max. :82.80	Max. :74.00	Max. :105000.00

Traitement des valeurs aberrantes et extrêmes

Une seconde étape de l'exploration des données disponibles nous a conduit à traiter les valeurs aberrantes et extrêmes sur les variables quantitatives. Nous appelons valeur aberrante une valeur ou une observation qui est « distante » des autres observations effectuées sur le même phénomène, c'est-à-dire qu'elle contraste grandement avec les valeurs « normalement » mesurées. Leur présence dans les données peut conduire à des estimateurs de paramètres biaisés et, suite à la réalisation de tests statistiques, à une interprétation des résultats erronée. Pouvant être dû à plusieurs facteurs, nous avons pensé utile dans un premier temps de les détecter, puis de les imputer si elles existaient dans notre base de données. La détection desdites données est perceptible comme l'on peut le voir sur notre graphe. Nous avons utilisé les boites à moustache afin de visualiser les débordements observés.

Détection visuelle des valeurs aberrantes et extrêmes

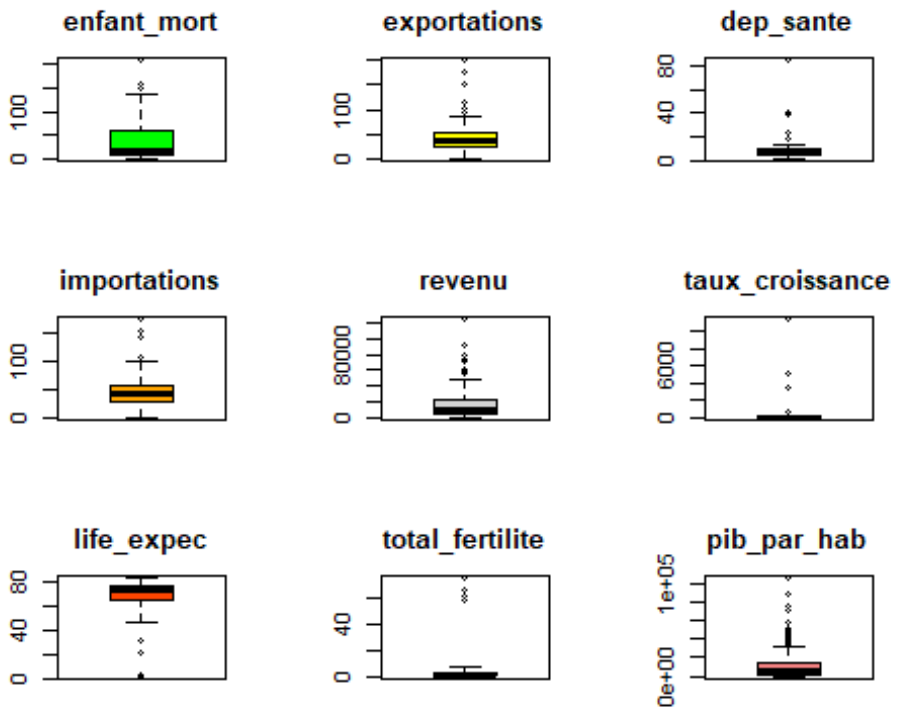


***Figure 3**: Visualisation des valeurs aberrantes et extrêmes.*

Phiques présentant des points au-dessus ou en dessous de la boite sont des valeurs aberrantes. Les autres variables qui présentent des données aberrantes seront retraitées.

Technique d'imputation de données aberrantes par winzorisation

Pour le traitement des données extrêmes nous avons utilisé la technique de Winzorisation en les ramenant dans les limites des bornes (inférieure et supérieure) des moustaches.

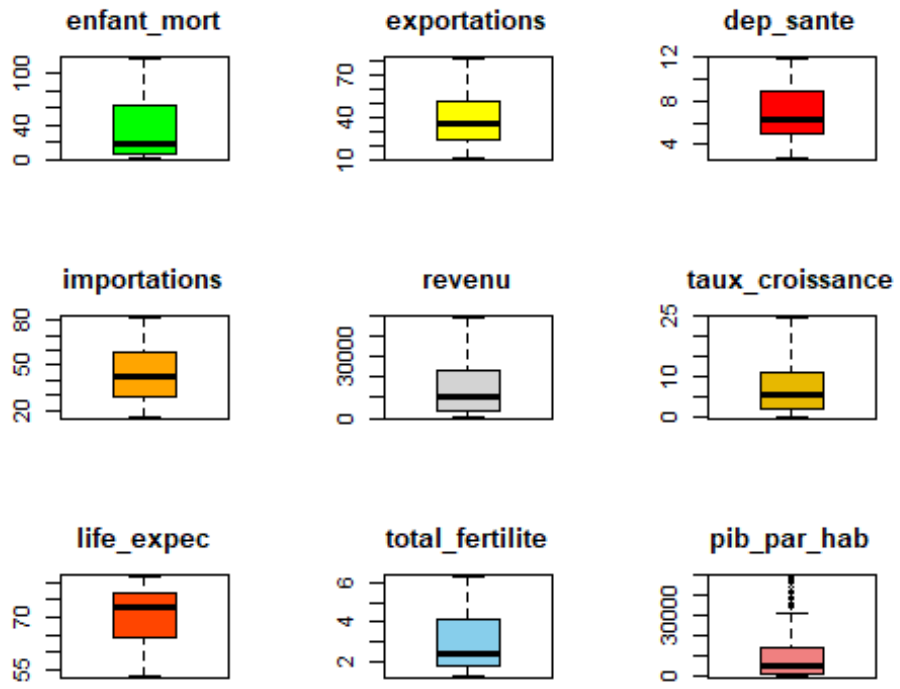


***Figure 4**: Visualisation des valeurs aberrantes et extrêmes après le traitement.*

Toutes les valeurs aberrantes et extrêmes ont été traitées comme le montre les graphiques ci-dessus.

PARTIE II : ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Dans cette partie nous allons réaliser l'analyse en composantes principales (**ACP**). Mais avant, nous avons souhaité baliser notre étude avec quelques visualisations préalables afin de justifier notre démarche.

Préalable à l'analyse en composantes principales (ACP)

Nous avons décidé de visualiser le sens de la distribution de nos données. Cette étape nous donnera la possibilité de normaliser les données au besoin.

Visualisation de la distribution des données

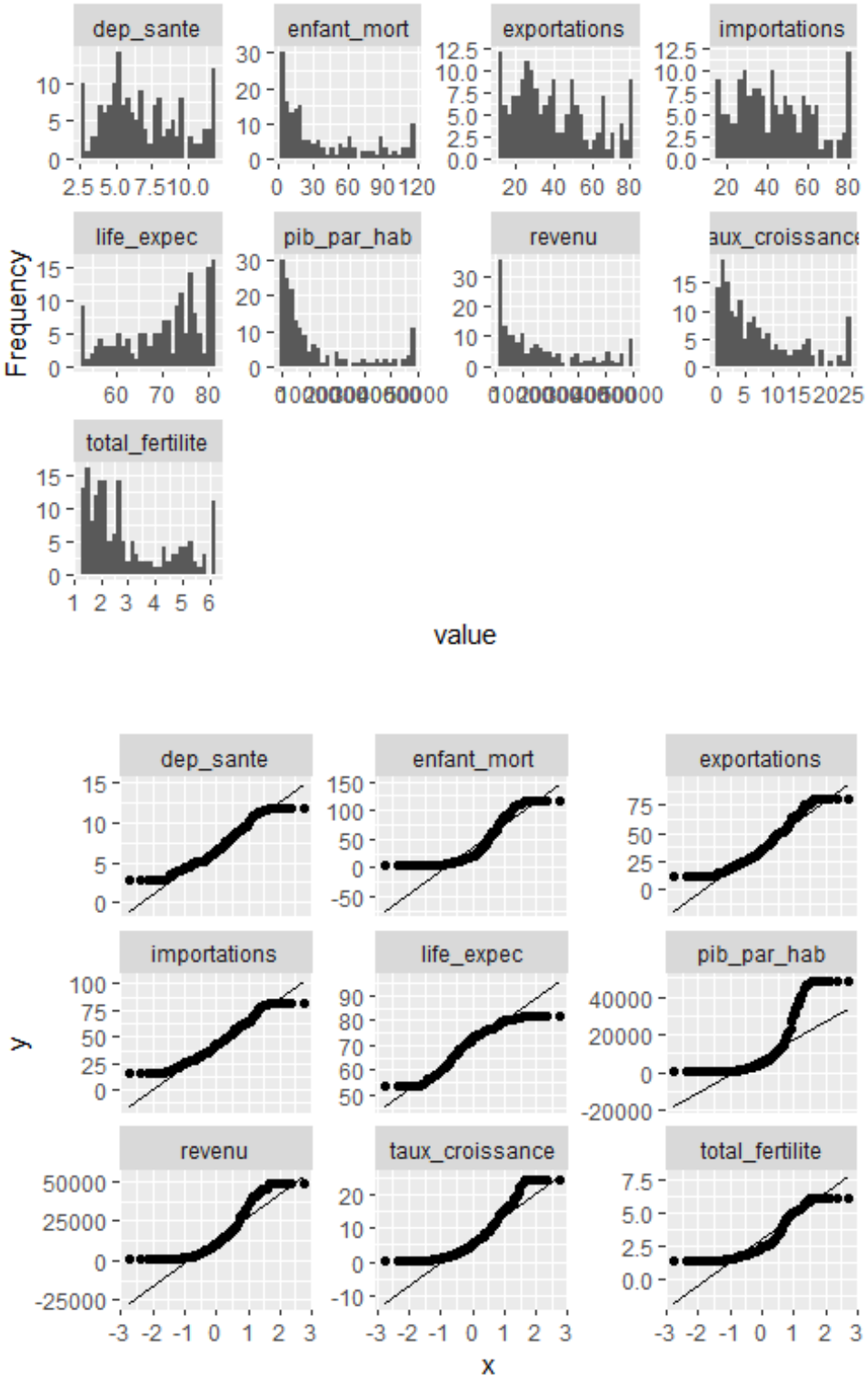


***Figure 5**: Visualisation des droites de régression.*

Les variables **dep_sante**, **importations** et **exportations** nous font soupçonner des distributions normales. Cette visualisation nous fait remarquer que ces données présentent des échelles différentes et ne peuvent être comparables dans leur état brut. Nous allons donc utiliser la technique statistique de normalisation pour les rendre comparables sur une échelle commune. L'idée c'est d'éviter par exemple que les données de la variable "**enfant_mort**" soient plus pesantes que celles de la variable "**importations**".

Corrélations des variables

La matrice des corrélations avec les coefficients nous indique de fortes corrélations positive entre **pib_par_pays** et le nombre de revenu. Il en est de même pour la liaison entre **total_fertilite** et **enfant_mort** ainsi qu'entre **life_expec** et la variable revenu, **pib_par_pays** et **life_expec** et en fin **exportation** et **importation** positive modérée. Ces variables augmentent et diminuent ensemble.

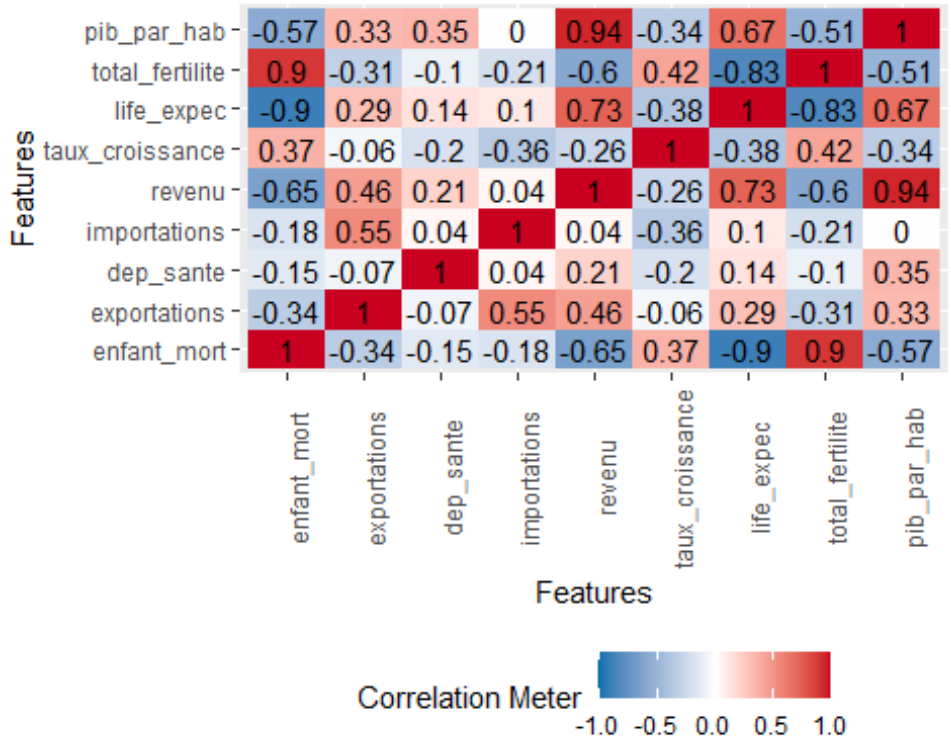


***Figure 6**: Matrice de corrélation.*

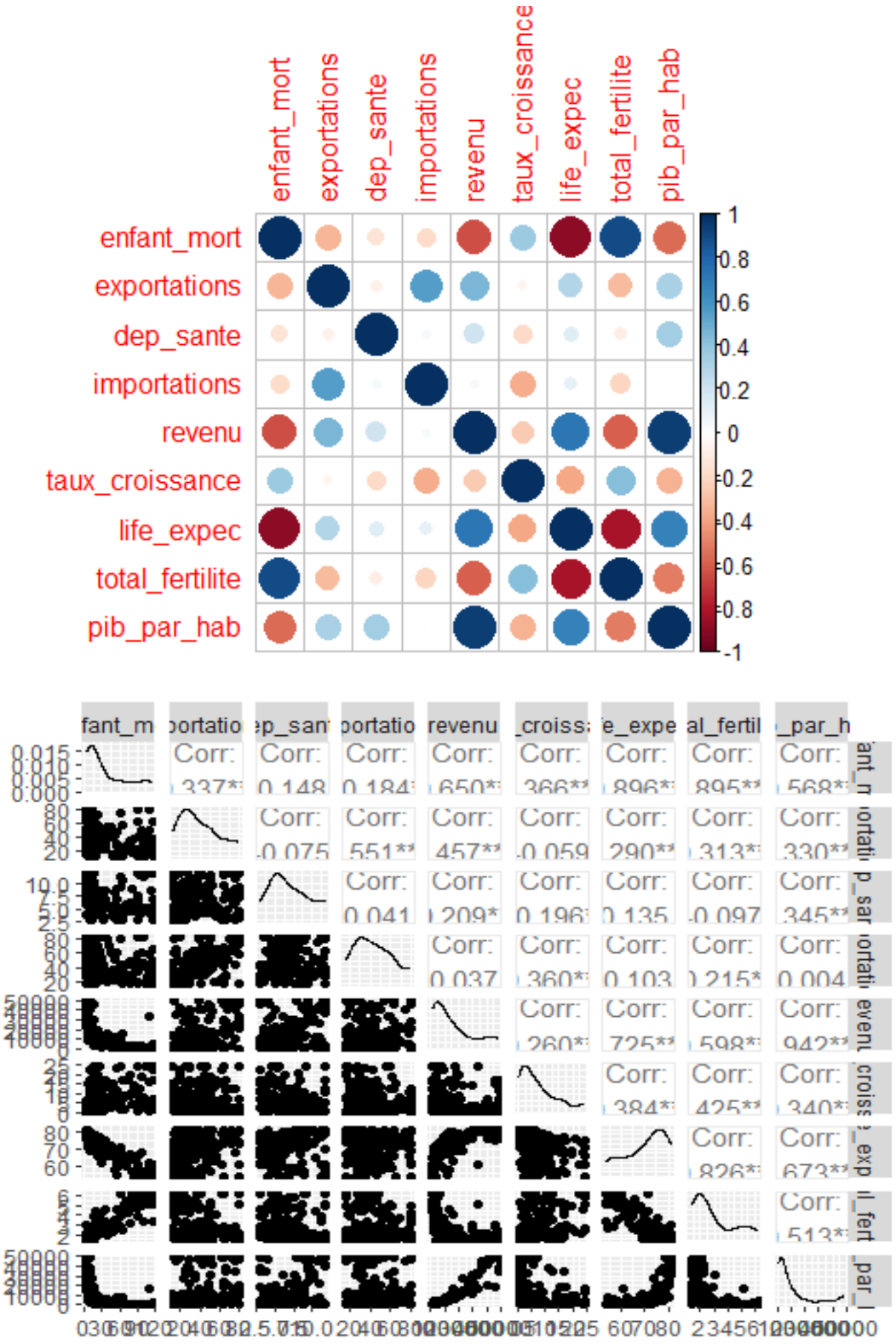


***Figure 7**: Graphes représentant les corrélations des variables.*

Le troisième graphique nous fait suspecter également une corrélation positive et significative entre revenu et dep_sante. Également pour les variables **dep_sante** et **exportations**.

Le quatrième graphique nous apporte plus de précision sur les coefficients de corrélation et leur significativité entre les variables.

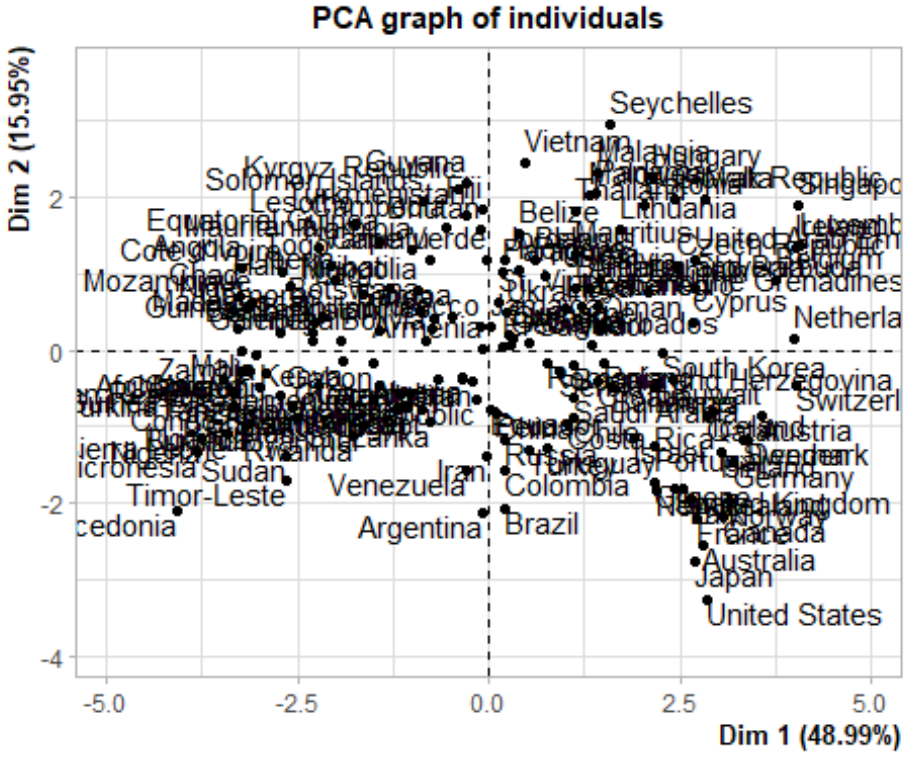
Cette visualisation des variables fortement corrélées nous indique qu'il existe une redondance dans les données. En raison donc de cette redondance, l'analyse en composantes principales (**ACP**) pourra être utilisée pour réduire d'origine en un nombre plus petit de nouvelles variables qui expliqueront la grande partie de la variance contenue dans les variables d'origine.

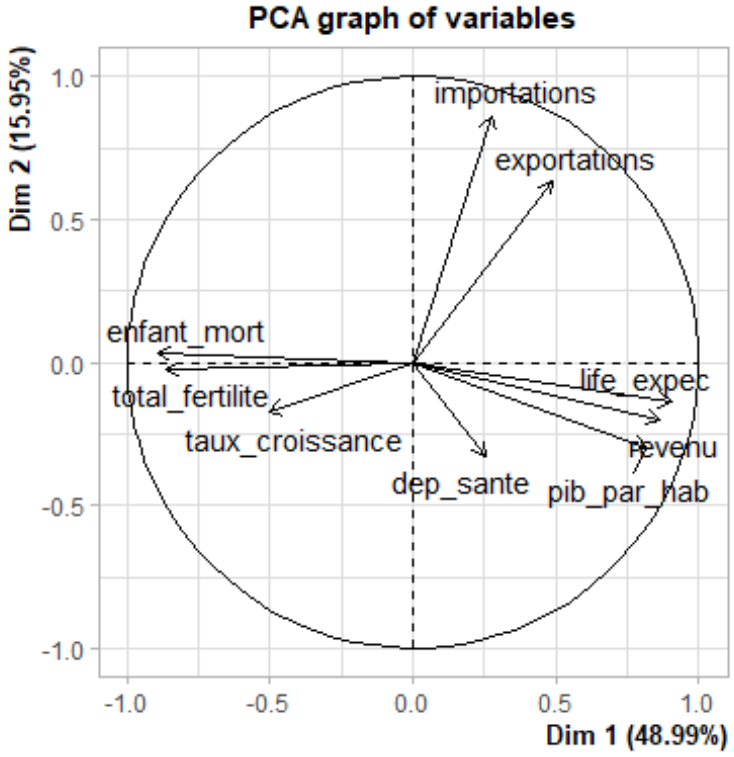
Analyse en composantes principales (ACP)

Cette étape va se consacrer à la projection et à la compression des données par l'analyse en composantes principales (**ACP**) pour n'en garder que celles qui portent suffisamment d'informations pour expliquer au mieux notre problématique. Comme expliqué précédemment, par défaut, la fonction '**PCA'** de '**FactoMineR'** centre et réduit les variables avant de réaliser l'analyse en composantes principales (**ACP**). Mais il était important pour nous de nous assurer de la normalisation de nos données. Ceci afin que toutes les variables aient le même poids dans la construction des plans de l'analyse en composantes principales (**ACP**).

Nous réaliserons l'analyse en composantes principales (**ACP**) uniquement sur les 9 variables quantitatives du jeu de données. Les plots des variables et des observations seront générés automatiquement.

Réalisation de l'ACP






```
PCA(X                                   =                                   Help_I)
Eigenvalues
```

```
##                                                    Dim.1      Dim.2      Dim.3      Dim.4      Dim.5      Dim.6      Dim.7
##  Variance                                   4.409      1.435      1.106      0.984      0.601      0.215      0.145
##  %  of  var.                                 48.987     15.948     12.294     10.932      6.674      2.392      1.612
##   Cumulative   %   of   var.       48.987      64.935     77.229     88.161     94.834     97.226     98.837
##                                                                                                      Dim.8           Dim.9
##     Variance                                                                                        0.072           0.032
##     %     of     var.                                                                               0.802           0.361
##          Cumulative          %           of           var.                                          99.639         100.000
##
##                    Individuals                    (the                       10                         first)
##                                              Dist       Dim.1       ctr      cos2       Dim.2       ctr      cos2
##  Afghanistan                       |     3.442  |   -3.129     1.330     0.827  |  -0.278     0.032     0.007  |
##  Albania                             |    1.605  |     0.528     0.038     0.108  |    0.084     0.003     0.003  |
##  Algeria                               |    1.994  |   -0.343     0.016     0.030  |  -0.348     0.051     0.030  |
##  Angola                                |    4.407  |   -3.221     1.409     0.534  |    1.074     0.482     0.059  |
##   Antigua   and   Barbuda   |     1.752   |      1.317       0.236      0.566   |      0.810      0.274      0.214   |
##  Argentina                          |    2.875  |  -0.091     0.001     0.001  |  -2.118     1.872     0.543  |
##  Armenia                              |    1.899  |  -0.085     0.001     0.002  |    0.017     0.000     0.000  |
##  Australia                         |    3.961  |     2.796     1.062     0.498  |  -2.536     2.683     0.410  |
##  Austria                              |    3.982  |     3.566     1.727     0.802  |  -0.869     0.315     0.048  |
##  Azerbaijan                        |    1.956  |  -0.204     0.006     0.011  |  -0.411     0.070     0.044  |
##                                                                                  Dim.3              ctr              cos2
##      Afghanistan                                                          1.050         0.596         0.093     |
##     Albania                                                                   0.236         0.030         0.022     |
##     Algeria                                                                 -1.692         1.549         0.720     |
##     Angola                                                                   -1.981         2.125         0.202     |
##         Antigua         and         Barbuda                0.167               0.015               0.009     |
##      Argentina                                                            -1.147         0.713         0.159     |
##     Armenia                                                                 -0.460         0.115         0.059     |
##     Australia                                                              0.326         0.058         0.007     |
##     Austria                                                                   0.889         0.428         0.050     |
##      Azerbaijan                                                          -1.381         1.033         0.499     |
##
##                                                                                                              Variables
##                                            Dim.1       ctr      cos2       Dim.2       ctr      cos2       Dim.3
##   enfant_mort                       |  -0.893   18.107     0.798  |     0.032     0.071     0.001  |     0.110
##   exportations                    |     0.487     5.378     0.237  |     0.635   28.091     0.403  |  -0.287
##  dep_sante                             |    0.257     1.496     0.066  |  -0.330     7.589     0.109  |     0.710
##   importations                  |     0.275     1.719     0.076  |     0.862   51.783     0.743  |     0.300
##  revenu                                    |     0.863   16.888     0.745  |  -0.201     2.823     0.041  |  -0.179
##    taux_croissance           |  -0.505     5.779     0.255  |  -0.174     2.107     0.030  |  -0.603
##   life_expec                           |     0.906   18.598     0.820  |  -0.137     1.316     0.019  |  -0.127
##    total_fertilite               |  -0.861   16.811     0.741  |  -0.025     0.042     0.001  |     0.077
##   pib_par_hab                       |     0.819   15.225     0.671  |  -0.298     6.178     0.089  |     0.006
##                                                                                                   ctr              cos2
##        enfant_mort                                                                     1.091           0.012     |
##        exportations                                                                   7.432           0.082     |
##       dep_sante                                                                         45.616           0.505     |
##        importations                                                                   8.117           0.090     |
##      revenu                                                                                     2.908           0.032     |
##           taux_croissance                                                     32.833                 0.363     |
##       life_expec                                                                           1.462           0.016     |
##           total_fertilite                                                           0.538              0.006     |
## pib_par_hab         0.003  0.000 |
```

Les analyse en composantes principales (PCA) effectuée sur un ensemble de données appelé Help_I :

1. **Valeurs propres :** Les valeurs propres représentent la quantité de variance expliquée par chaque dimension de l'analyse en composantes principales. Par exemple, la première dimension explique près de 49% de la variance totale des données, tandis que la deuxième dimension en explique environ 16%.

2. **Individus :** Cette section fournit des informations sur la contribution de chaque individu (pays dans notre cas) à chaque dimension de l'analyse. Par exemple, pour l'Afghanistan, les valeurs indiquent dans quelle mesure ce pays contribue à chaque dimension.

3. **Variables :** Il s'agit des variables originales de votre ensemble de données et de leur contribution à chaque dimension de l'analyse. Par exemple, la variable "**enfant_mort**"

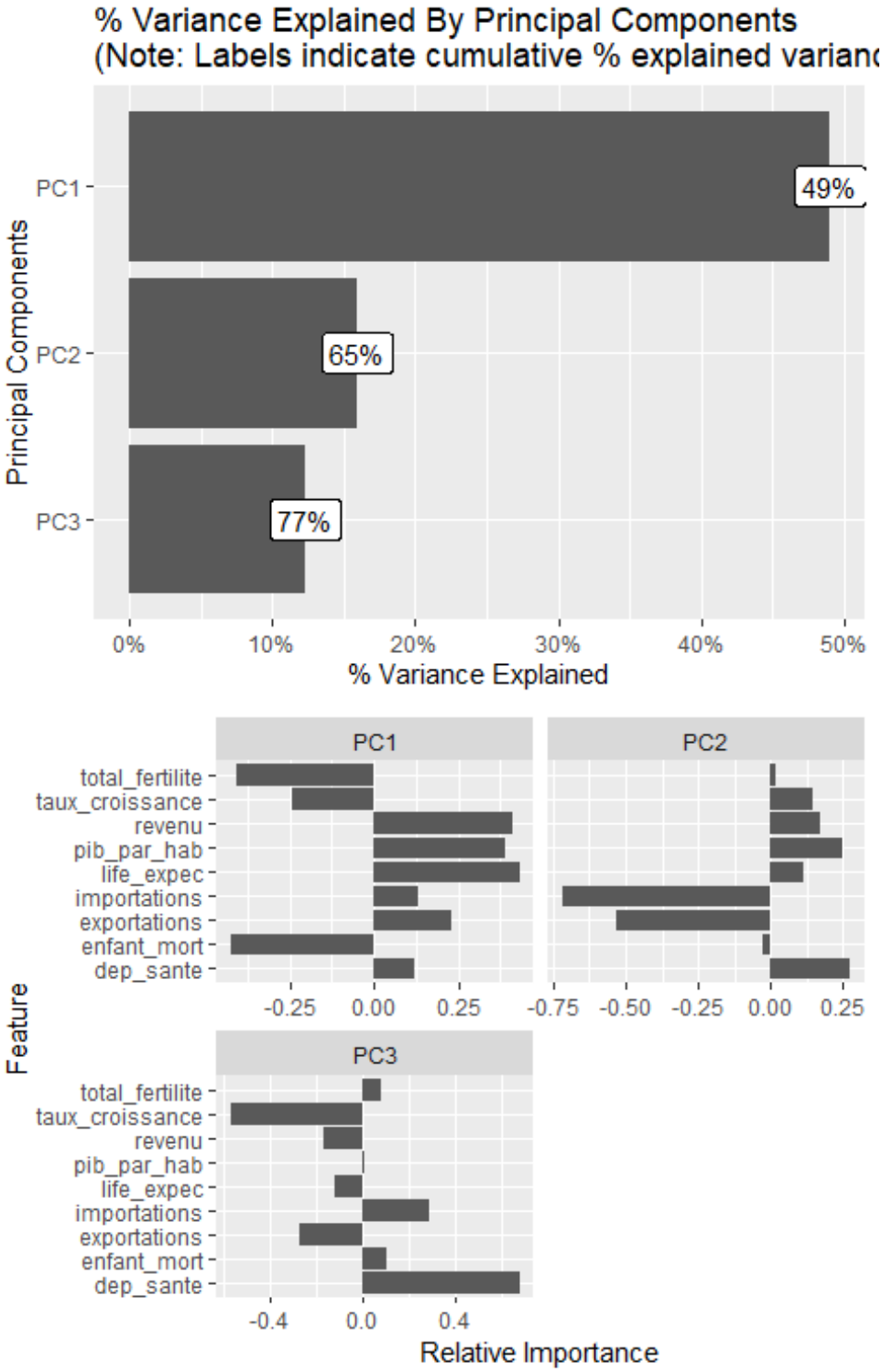
(taux de mortalité infantile) contribue négativement à la première dimension, ce qui signifie qu'elle est fortement associée à cette dimension.

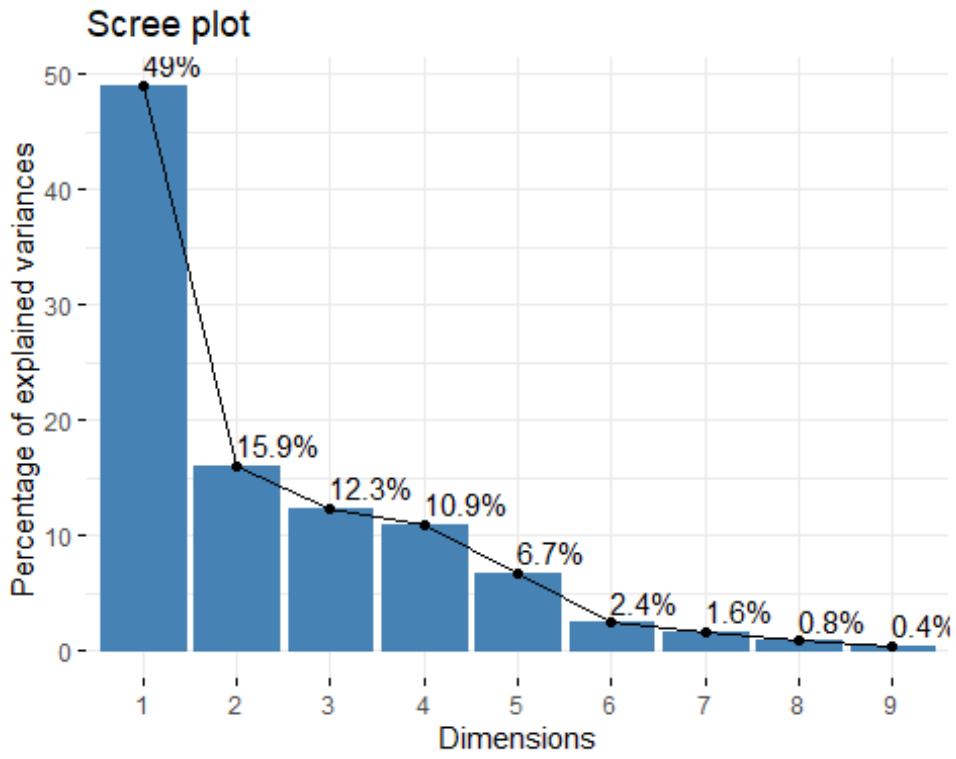
4. **Statistiques Hopkins :** Le test de Hopkins est utilisé pour évaluer la "clusterabilité" des données, c'est-à-dire leur aptitude à être regroupées en clusters distincts. Un score proche de 1 indique une forte propension à la clusterisation.

5. **Lien entre la variable de cluster et les variables quantitatives :** Cette section montre comment chaque dimension de l'analyse en composantes principales est liée aux variables quantitatives originales. Par exemple, une forte valeur d'Eta2 et une faible valeur de p indiquent une forte corrélation entre la dimension et la variable respective.

6. **Description de chaque cluster par les variables quantitatives :** Cette partie donne des informations sur les différences de moyenne entre chaque cluster pour chaque variable quantitative. Cela permet de comprendre comment les clusters se distinguent les uns des autres en termes de caractéristiques quantitatives.

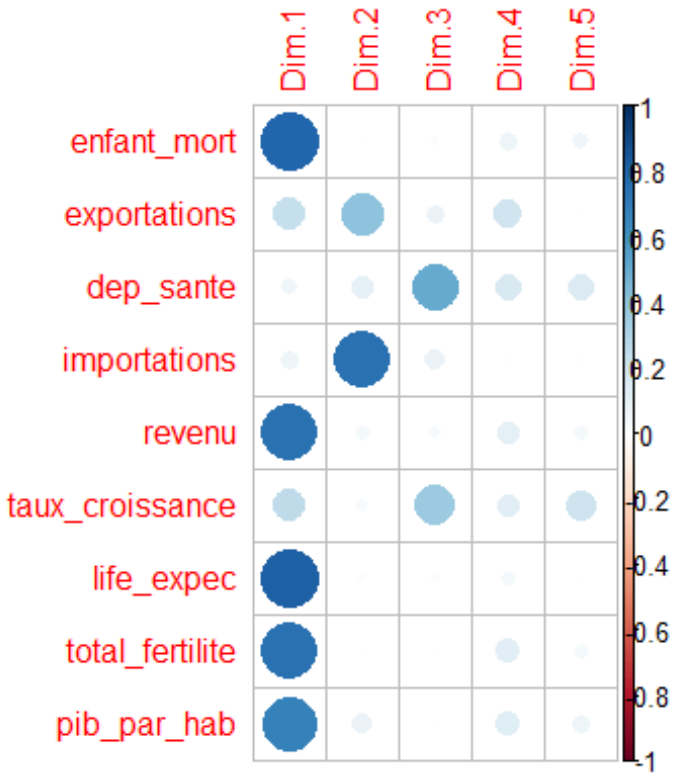
Ces données fournissent des informations précieuses sur la structure des données, les relations entre les variables et les individus, ainsi que sur la clusterisation potentielle des données.

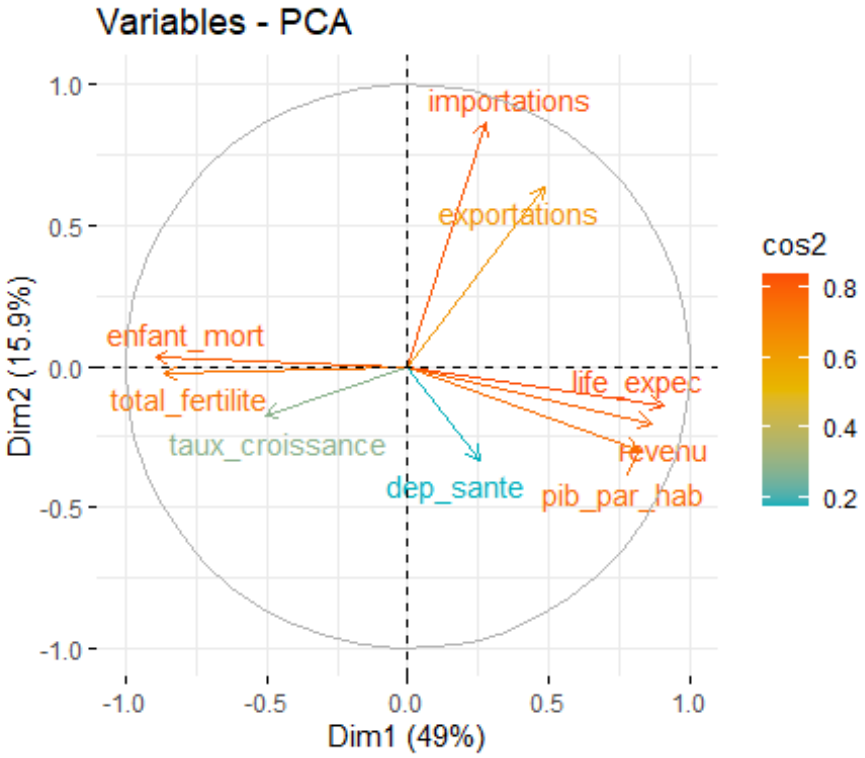

% Variance Explained By Principal Components
(Note: Labels indicate cumulative % explained varianc
Principal Components
PC1
PC2
PC3
49%
65%
77%
0%
10%
20%
30%
40%
50%
% Variance Explained
PC1
PC2
PC3
total_fertilite
taux_croissance
revenu
pib_par_hab
life_expec
importations
exportations
enfant_mort
dep_sante
-0.25
0.00
0.25
-0.75
-0.50
-0.25
0.00
0.25
-0.4
0.0
0.4
Feature
Relative Importance

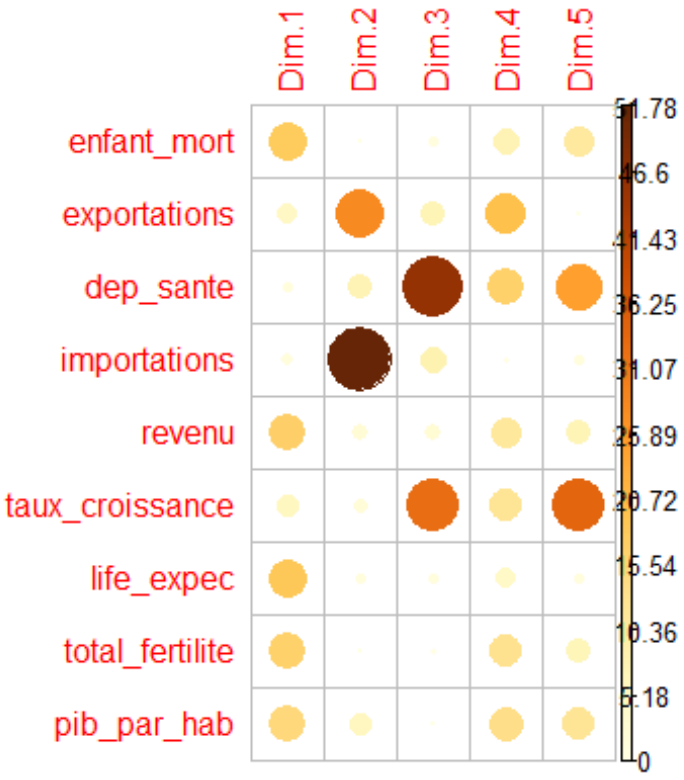


Visualisation de la qualité de la représentation des variables

Les variables libellées sont celles les mieux représentées sur le plan.

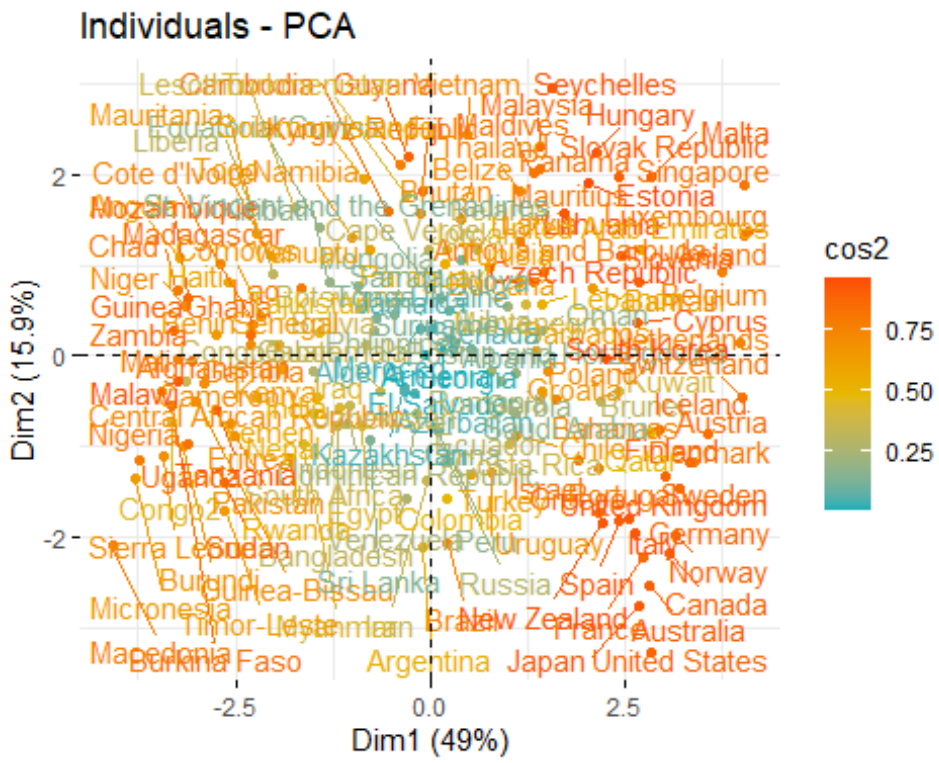

Dim.1
Dim.2
Dim.3
Dim.4
Dim.5
enfant_mort
exportations
dep_sante
importations
revenu
taux_croissance
life_expec
total_fertilite
pib_par_hab
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Variables - PCA
importations
exportations
enfant_mort
total_fertilite
taux_croissance
dep_sante
life_expec
revenu
pib_par_hab
cos2
0.8
0.6
0.4
0.2
1.0
0.5
0.0
-0.5
-1.0
Dim2 (15.9%)
Dim1 (49%)



Visualisation de la qualité de la représentation des observations

Les individus libellés sont ceux ayant la plus grande contribution à la construction du plan.



PARTIE III : LA CLASSIFICATION DES DONNEES

Détermination de l'échantillon

Présentons un échantillon de 50 enregistrements à partir du jeu de données afin que le graphe ne soit pas Surchargé.

Classification des données

Le résumé statistique nous montre que l'échelle sur toutes les colonnes est la même donc nous n'avons pas besoin de normaliser les données. Il est important de toujours vérifier l'échelle des différentes variables avant de procéder à la segmentation car la plupart des algorithmes de Clustering se base sur le calcul de la distance entre des points. De ce fait, si les valeurs ne sont pas à la même échelle le calcul sera biaisé et on n'aura pas un résultat exploitable.

Faisabilité de l'analyse de clustering

De plus il faut vérifier la faisabilité de l'analyse de clustering en calculant la statistique Hopkins.

```
$hopkins_stat
[1]                                                    0.8824324
$plot
NULL
```

Le résultat du test de Hopkins est de 0.8824324, ce qui suggère une tendance élevée des données à former des clusters. Cela indique que les données sont propices à la clusterisation.

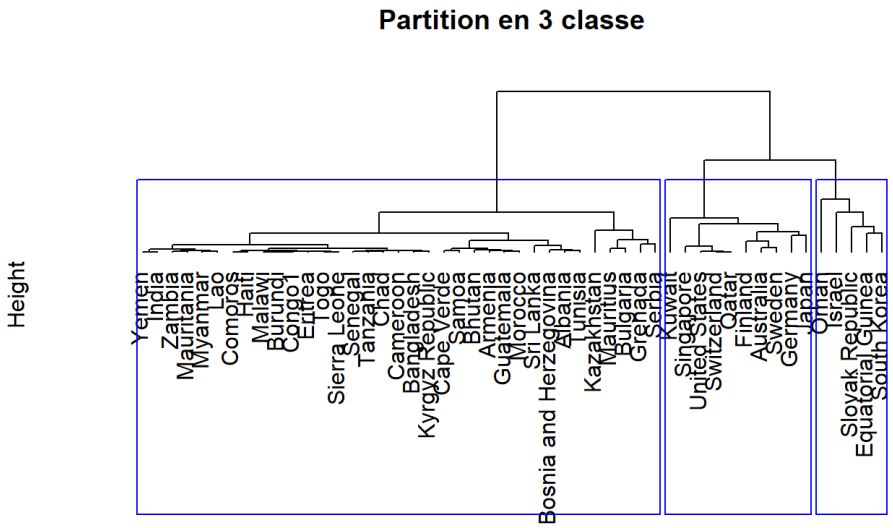
Détermination du nombre optimal de clusters

FactoMiner nous suggère d'utiliser la partition ayant la plus grande perte relative d'inertie. Nous allons calculer cet indicateur avec best.cutree. L'extension JLutils propose une fonction best.cutree qui permet de calculer cette indicatrice à partir de n'importe quel dendrogramme calculé avec hclust ou agnes. Par défaut, best.cutree regarde quelle serait la meilleure partition entre 3 et 20 classes.

Nombre optimal de cluster

En l'occurence il s'agirait d'une partition en 3 classes.

Visualisation de la classification



Affectation des individus à chaque classe

Individus	x
Cape Verde	1
Bangladesh	1
Myanmar	1
Germany	2

Individus	x
Finland	2
Bosnia and Herzegovina	1
Burundi	1
Oman	3
Togo	1
Sierra Leone	1
Yemen	1
Mauritius	1
Cameroon	1
Switzerland	2
Qatar	2
Albania	1
Tanzania	1
Malawi	1
Tunisia	1
Congo1	1
Zambia	1
Kazakhstan	1
Grenada	1
Lao	1
Bhutan	1
Kuwait	2
Mauritania	1
Israel	3
India	1
Guatemala	1
Senegal	1
Samoa	1
Singapore	2
Equatorial Guinea	3
South Korea	3
Slovak Republic	3
Chad	1

Individus	x
Australia	2
Sri Lanka	1
Japan	2
Comoros	1
Kyrgyz Republic	1
Sweden	2
Haiti	1
Armenia	1
Serbia	1
Eritrea	1
Morocco	1
United States	2
Bulgaria	1

Implémentation de l'algorithme K-means avec k = best

Nous pouvons à présent implémenter l'algorithme k-means à l'ensemble de nos données en considérant Best comme nombre de clusters optimal.

```
K-means clustering with 3 clusters of sizes 96, 31, 40

Cluster means :

enfant_mort exportations dep_sante importations    revenu taux_croissance

1   57.418750    32.35260  6.400469    43.73542  5099.004       8.940896

2    5.765161    49.43903  8.261677    41.92000 42042.258       4.139697

3   15.353000    48.39025  6.973250    47.15550 20487.500       6.503700

life_expec total_fertilite pib_par_hab

1   65.61188       3.771365    2345.575

2   79.90645       1.911290   40477.097

3   74.85000       1.954250   12884.500

Clustering vector:

Afghanistan                 Albania

1                       1

Algeria                 Angola

1                       1

Antigua and Barbuda                 Argentina

3                       3

Armenia                 Australia

1                       2

Austria                 Azerbaijan
```

2 3

Bahamas Bahrain

3 2

Bangladesh Barbados

1 3

Belarus Belgium

3 2

Belize Benin

1 1

Bhutan Bolivia

1 1

Bosnia and Herzegovina Botswana

1 1

Brazil Brunei

3 2

Bulgaria Burkina Faso

3 1

Burundi Cambodia

1 1

Cameroon Canada

1 2

Cape Verde Central African Republic

1 1

Chad Chile

1 3

China Colombia

1 1

Comoros Congo1

1 1

Congo2 Costa Rica

1 3

Cote d'Ivoire Croatia

1 3

Cyprus Czech Republic

2 3

Denmark Dominican Republic

2 1

Ecuador Egypt

1 1

El Salvador Equatorial Guinea

1 3

Eritrea Estonia

1 3

Fiji Finland

1 2

France Gabon

2 3

Gambia Georgia

1 1

Germany Ghana

2 1

Greece Grenada

3 1

Guatemala Guinea

1 1

Guinea-Bissau Guyana

1 1

Haiti Hungary

1 3

Iceland India

2 1

Indonesia Iran

1 3

Iraq Ireland

1 2

Israel Italy

2 2

Jamaica Japan

1 2

Jordan Kazakhstan

1 3

Kenya Kiribati

1 1

Kuwait Kyrgyz Republic

2 1

Lao Latvia

1 3

Lebanon Lesotho

3 1

Liberia Libya

1 3

Lithuania Luxembourg

3 2

Macedonia Madagascar

1 1

Malawi Malaysia

1 3

Maldives Mali

1 1

Malta Mauritania

3 1

Mauritius Micronesia

3 1

Moldova Mongolia

1 1

Montenegro Morocco

3 1

Mozambique Myanmar

1 1

Namibia Nepal

1 1

Netherlands New Zealand

2 2

Niger Nigeria

1 1

Norway Oman

2 2

Pakistan Panama

1 3

Paraguay Peru

1 1

Philippines Poland

1 3

Portugal Qatar

3 2

Romania Russia

3 3

Rwanda Samoa

1 1

Saudi Arabia Senegal

2 1

Serbia Seychelles

1 3

Sierra Leone Singapore

1 2

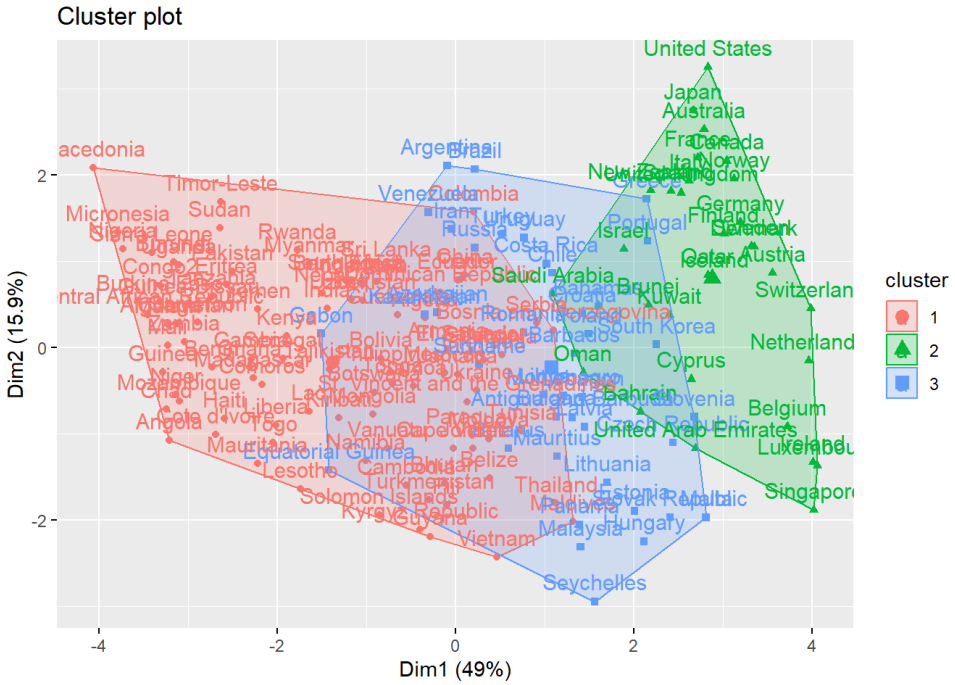
Slovak Republic Slovenia

```
3                    3
Solomon Islands              South Africa
1                    1
South Korea                  Spain
3                    2
Sri Lanka St. Vincent and the Grenadines
1                    1
Sudan                Suriname
1                    3
Sweden               Switzerland
2                    2
Tajikistan               Tanzania
1                    1
Thailand             Timor-Leste
1                    1
Togo                 Tonga
1                    1
Tunisia              Turkey
1                    3
Turkmenistan                 Uganda
1                    1
Ukraine        United Arab Emirates
1                    2
United Kingdom           United States
2                    2
Uruguay              Uzbekistan
3                    1
Vanuatu              Venezuela
1                    3
Vietnam                  Yemen
1                    1
Zambia
1
Within cluster sum of squares by cluster:
 [1] 1677396973 3526160050 2409405779
 (between_SS / total_SS =  89.8 %)
Available components:
 [1] "cluster"    "centers"    "totss"      "withinss"    "tot.withinss"
 [6] "betweenss"  "size"       "iter"       "ifault"
```

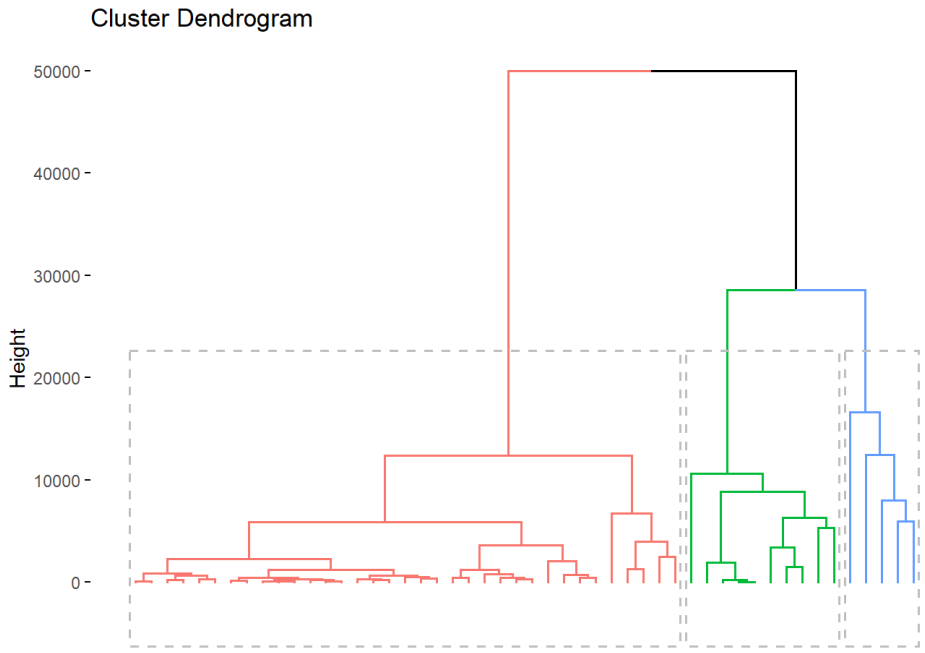
Évidemment, trois clusters sont formés et cela est conforme à notre paramètre k qui a été transmis à l'algorithme. Nous voyons que le premier groupe contient 96 observations, le

deuxième groupe contient 31 observations et le troisième groupe contient 40 observations. Chacune des observations de notre ensemble des données est affectée à l'un de ces trois groupes (1, 2 ou 3).

Visualisation des résultats avec K-means



Visualisation du clustering par le dendrogramme



Qualité de la segmentation

En Apprentissage supervisé, on dispose de données étiquetées ce qui permet de comparer les prédictions du modèle aux réelles observations. En apprentissage non-supervisé, comme c'est le cas ici, cette comparaison ne peut se faire car les données ne sont pas étiquetées. Comment mesurer alors la qualité de notre partitionnement ?

L'objectif du clustering est d'obtenir des clusters de bonne qualité. Le clustering est de haute qualité si la distance dans les observations (intra-cluster) d'un cluster donné est minimale et la distance séparant les clusters eux-mêmes (inter-cluster) est maximale. Malheureusement, il n'y aucune valeur seuil (ou valeur de référence) pour dire la distance intra-cluster est minimale ou si la distance inter-cluster est maximale.

Pour mesurer la qualité du clustering, on peut calculer le coefficient de Silhouette pour chacun des trois groupes. L'indice Silhouette est calculé en utilisant la distance intra-cluster moyenne, a, et la distance moyenne du cluster le plus proche, b, pour chacune des observations participant à l'exercice de clustering. Pour une observation, l'indice Silhouette est donné par la formule : (b - a) /max (a, b).

```
Silhouette of 167 units in 3 clusters from silhouette.default(x = kmeans_out$cluster, dist = dist(Help_I, "euclidean")) :
Cluster sizes and average silhouette widths:
96     31     40
```

```
0.7144976 0.5912875 0.4481125

Individual silhouette widths:

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

0.02487 0.53888 0.72813 0.62782 0.80236 0.82581
```

La silhouette de 167 unités réparties en 3 clusters à partir de la fonction silhouette.default (x = kmeans_out$cluster, dist = dist (**Help_I**, “euclidean”)) indique les tailles des clusters et les largeurs moyennes des silhouettes :

- Le premier cluster contient 96 unités avec une largeur moyenne de silhouette de 0,7144976.
- Le deuxième cluster contient 31 unités avec une largeur moyenne de silhouette de 0,5912875.
- Le troisième cluster contient 40 unités avec une largeur moyenne de silhouette de 0,4481125.

Les largeurs de silhouette individuelles varient de manière significative : - La plus faible largeur de silhouette est de 0,02487. - Le premier quartile est de 0,53888. - La médiane est de 0,72813. - La moyenne est de 0,62782. - Le troisième quartile est de 0,80236. - La plus grande largeur de silhouette est de 0,82581.

Ces informations permettent d’évaluer la cohésion et la séparation des clusters. Une valeur de silhouette proche de 1 indique une bonne séparation entre les clusters, tandis qu’une valeur proche de 0 indique un chevauchement entre les clusters.

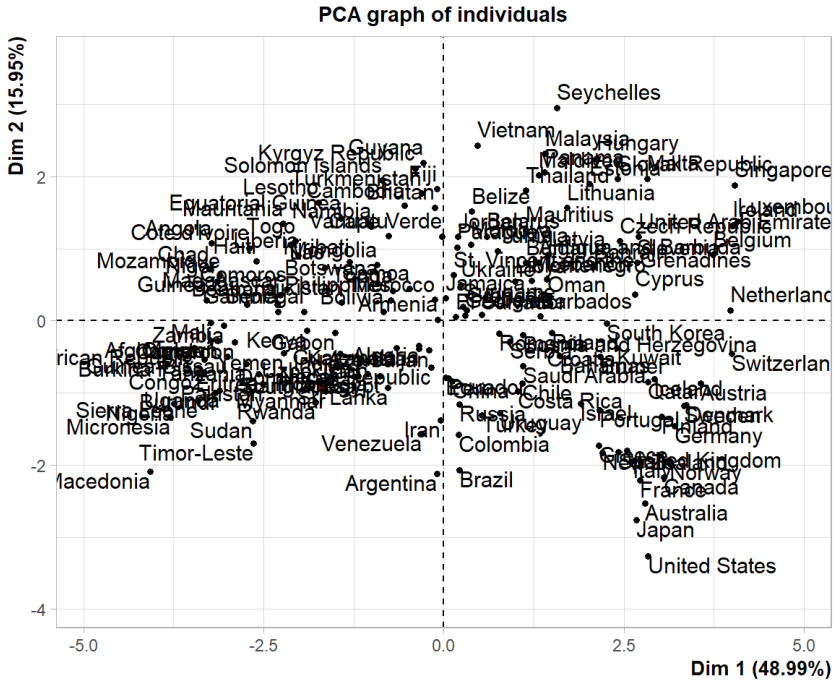
Centre de gravité de chaque cluster

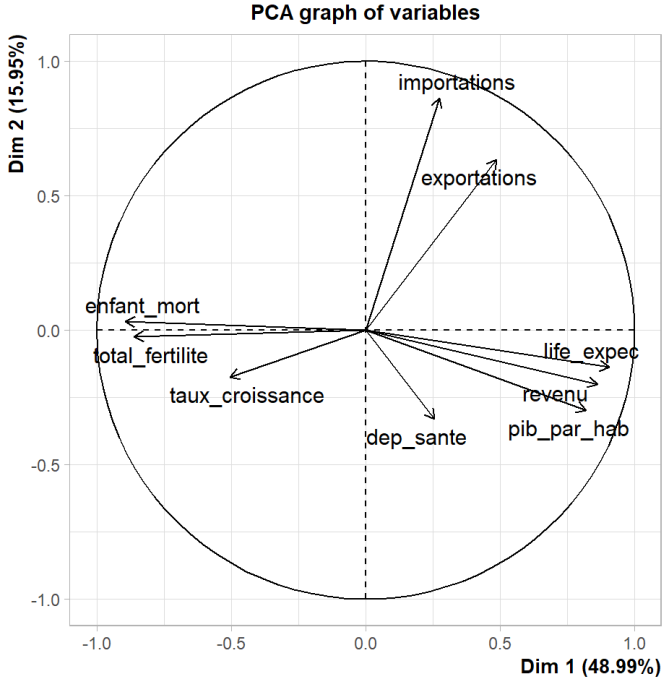
```
enfant_mort exportations dep_sante importations    revenu taux_croissance

1   57.418750    32.35260  6.400469    43.73542  5099.004      8.940896

2    5.765161    49.43903  8.261677    41.92000 42042.258      4.139697

3   15.353000    48.39025  6.973250    47.15550 20487.500      6.503700

life_expec total_fertilite pib_par_hab

1   65.61188      3.771365   2345.575

2   79.90645      1.911290   40477.097

3   74.85000      1.954250   12884.500
```

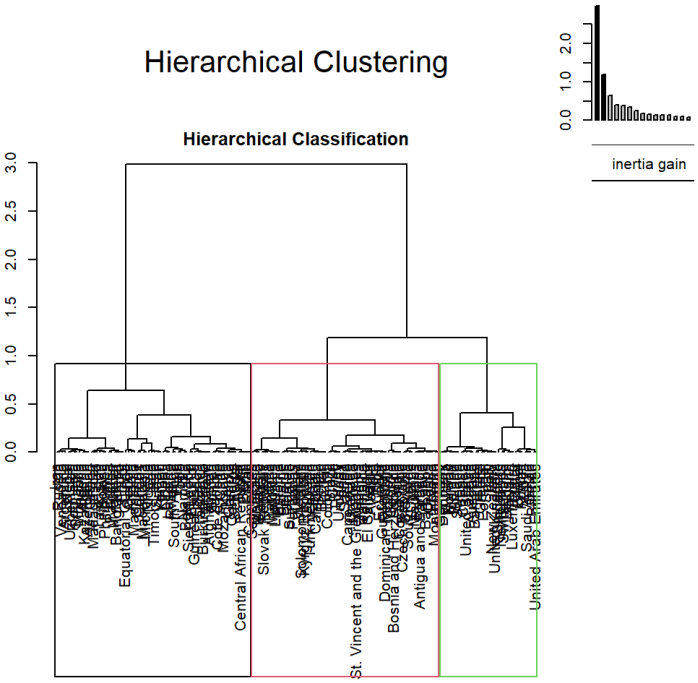
Indice moyen observé chez les patients pour chaque groupe

```
[1] 7662.811

[1] 82710.7

[1] 33573.18
```

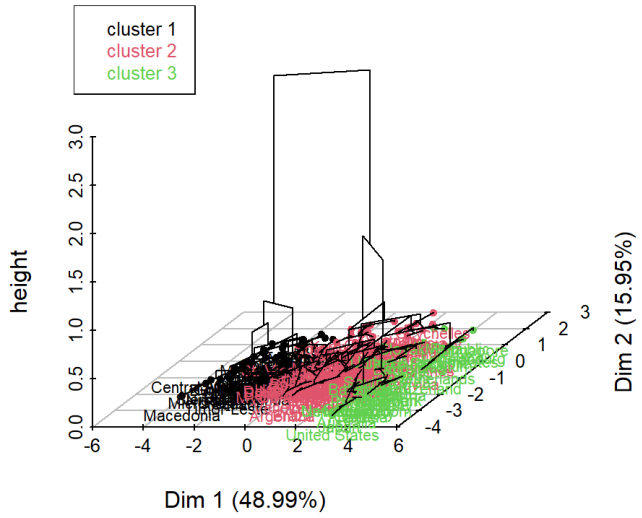
Catégorisation par classification ascendante hiérarchique

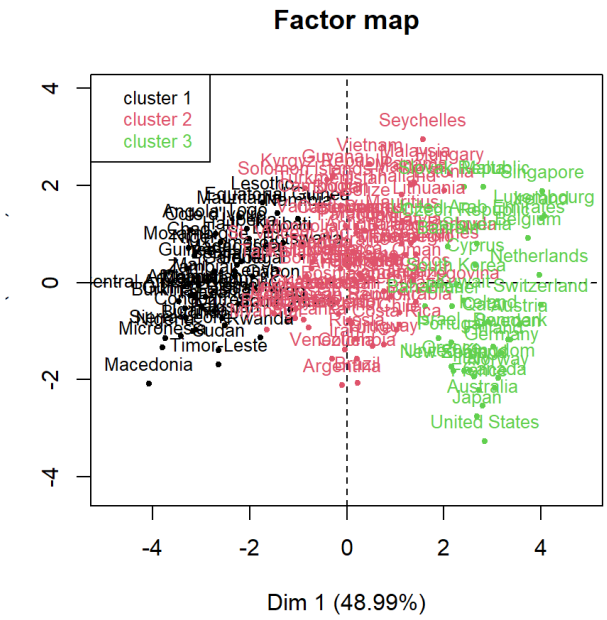






Hierarchical clustering on the factor map





CARACTERISATION DES CLUSTERS PAR LES VARIABLES

Link between the cluster variable and the quantitative variables

Eta2 P-value

pib_par_hab 0.8416152 2.379638e-66

enfant_mort 0.8235035 1.707462e-62

life_expec 0.7969290 1.687023e-57

total_fertilite 0.7709906 3.220838e-53

revenu 0.7511760 2.905162e-50

taux_croissance 0.1748644 1.429152e-07

dep_sante 0.1470039 2.175984e-06

exportations 0.1292176 1.181924e-05

Description of each cluster by quantitative variables

$`1`

v.test Mean in category Overall mean sd in category

enfant_mort 11.468078 88.359184 37.754731 22.2005071

total_fertilite 11.167942 5.033694 2.990844 0.9729316

taux_croissance 4.237241 11.012143 7.465896 7.9790748

importations -2.261705 39.153878 44.217605 16.5790304

exportations -3.999366 29.726531 39.365689 17.3103089

pib_par_hab -5.550731 1822.310204 11948.174850 2879.4230142

revenu -6.663677 3754.579592 15642.601198 5498.4217687

life_expec -10.749312 59.466122 70.478084 4.6881704

Overall sd p.value

enfant_mort 36.636095 1.908508e-30

total_fertilite 1.518708 5.852115e-29

taux_croissance 6.948594 2.262836e-05

importations 18.588543 2.371561e-02

exportations 20.010552 6.351251e-05

pib_par_hab 15145.827443 2.844773e-08

revenu 14811.762768 2.670606e-11

life_expec 8.505407 5.970509e-27

$`2`

v.test Mean in category Overall mean sd in category

life_expec 3.529649 72.879012 70.478084 3.9195211

revenu -2.824607 12296.666667 15642.601198 8134.9568811

dep_sante -3.820241 6.089160 6.883156 1.9564435

pib_par_hab -4.671509 6289.654321 11948.174850 4458.1991912

enfant_mort -5.395231 21.946914 37.754731 13.6917934

total_fertilite -5.575090 2.313704 2.990844 0.6791779

Overall sd p.value

life_expec 8.505407 4.161119e-04

revenu 14811.762768 4.733865e-03

dep_sante 2.598810 1.333214e-04

pib_par_hab 15145.827443 2.989945e-06

enfant_mort 36.636095 6.843540e-08

total_fertilite 1.518708 2.474023e-08

$`3`

v.test Mean in category Overall mean sd in category

pib_par_hab 11.707600 37745.675676 11948.174850 1.046410e+04

revenu 10.705217 38711.081081 15642.601198 7.840640e+03

life_expec 7.537809 79.805405 70.478084 1.888650e+00

dep_sante 4.657323 8.644027 6.883156 2.660329e+00

exportations 3.568197 49.753514 39.365689 2.124743e+01

taux_croissance -4.581525 2.834373 7.465896 3.766201e+00

total_fertilite -5.535255 1.767838 2.990844 3.473073e-01

enfant_mort -6.080769 5.344324 37.754731 2.473523e+00

Overall sd p.value

pib_par_hab 15145.827443 1.165322e-31

revenu 14811.762768 9.620539e-27

life_expec 8.505407 4.779350e-14

dep_sante 2.598810 3.203470e-06

exportations 20.010552 3.594468e-04

taux_croissance 6.948594 4.615964e-06

total_fertilite 1.518708 3.107752e-08

enfant_mort 36.636095 1.196072e-09

Lien entre la variable de cluster et les variables quantitatives

- **Eta2** : Il s'agit du carré du coefficient de corrélation entre chaque variable quantitative et la variable de cluster. Plus la valeur d'Eta2 est proche de 1, plus la variable est fortement liée au cluster.
- **P-value** : Il s'agit de la significativité statistique de la relation entre chaque variable quantitative et la variable de cluster. Une valeur de P-value faible (inférieure au seuil de signification généralement fixé à 0.05) indique une relation significative.

Description de chaque cluster par variables quantitatives

Chaque cluster est décrit en termes de moyenne et de déviation standard pour chaque variable quantitative, ainsi que des tests statistiques associés (v-test) pour déterminer si les moyennes dans chaque cluster diffèrent significativement de la moyenne globale.

Cluster 1:

- Ce cluster se caractérise par des valeurs élevées pour les variables telles que le taux de mortalité infantile, le taux de fertilité total, le revenu, l'espérance de vie, etc.
- Les variables telles que les exportations et les importations semblent plus faibles dans ce cluster par rapport à la moyenne globale.
- Les valeurs-p des tests statistiques suggèrent que les différences observées dans ces variables entre ce cluster et la moyenne globale sont statistiquement significatives.

Cluster 2:

- Ce cluster se caractérise par des valeurs intermédiaires pour la plupart des variables, mais surtout pour l'espérance de vie, le revenu, et la dépense de santé.
- Les différences observées dans les variables entre ce cluster et la moyenne globale sont également statistiquement significatives.

Cluster 3:

- Ce cluster présente des valeurs élevées pour des variables telles que le revenu, l'espérance de vie, le PIB par habitant, etc.
- Les valeurs-p des tests statistiques indiquent des différences significatives dans ces variables entre ce cluster et la moyenne globale.

L'analyse montre que les trois clusters se distinguent par leurs caractéristiques socio-économiques, avec des différences significatives dans plusieurs variables clés.

CARACTERISATION DES CLUSTERS PAR LES DIMENSIONS (axe)

Link between the cluster variable and the quantitative variables

Eta2 P-value

Dim.1 0.85172538 1.065319e-68

Dim.4 0.40472634 3.363306e-19

Dim.5 0.19785898 1.407860e-08

Dim.2 0.08943539 4.607687e-04

Dim.3 0.05620328 8.710317e-03

Description of each cluster by quantitative variables

$`1`

v.test Mean in category Overall mean sd in category Overall sd

Dim.4 4.845353 0.5788746 -2.145240e-16 1.0131782 0.9919052

Dim.3 2.283473 0.2892994 -2.446168e-16 1.1507254 1.0518717

Dim.5 -2.978308 -0.2780086 -7.739204e-16 0.9113270 0.7749963

Dim.1 -10.119543 -2.5592345 -1.718270e-15 0.7799347 2.0997146

p.value

Dim.4 1.263870e-06

Dim.3 2.240253e-02

Dim.5 2.898444e-03

Dim.1 4.525248e-24

$`2`

v.test Mean in category Overall mean sd in category Overall sd

Dim.5 5.683290 0.3522509 -7.739204e-16 0.6172850 0.7749963

Dim.2 3.266904 0.3130171 3.947275e-18 1.1504803 1.1980611

Dim.3 -3.011056 -0.2532994 -2.446168e-16 0.9109187 1.0518717

Dim.4 -8.188987 -0.6496105 -2.145240e-16 0.6001998 0.9919052

p.value

Dim.5 1.321283e-08

Dim.2 1.087306e-03

Dim.3 2.603411e-03

Dim.4 2.634336e-16

$`3`

v.test Mean in category Overall mean sd in category Overall sd

Dim.1 9.341889 2.8537227 -1.718270e-15 0.6402982 2.0997146

Dim.4 4.542422 0.6555026 -2.145240e-16 0.7038195 0.9919052

Dim.2 -3.439248 -0.5994586 3.947275e-18 1.4124789 1.1980611

Dim.5 -3.574017 -0.4029703 -7.739204e-16 0.4903900 0.7749963

p.value

Dim.1 9.463012e-21

Dim.4 5.561161e-06

Dim.2 5.833333e-04

```
Dim.5 3.515458e-04
```

IDENTIFICATION DES PARAGONS (individu le plus proche du centre des classes)

```
Cluster: 1
Malawi    Zambia   Tanzania Afghanistan    Guinea
0.7115560  0.8591448  0.8666673  0.9925416  1.1170247
Cluster: 2
Suriname                 Jamaica
0.6085087                0.6847724
Paraguay                 Tunisia
0.7196014                0.8865682
St. Vincent and the Grenadines
0.9643442
Cluster: 3
Iceland   Finland   Sweden   Austria   Germany
0.5819603 0.9242841 0.9412343 1.1801986 1.3226342
$dist
Cluster: 1
Macedonia      Micronesia        Congo1        Congo2
5.744174       5.574780       5.477122       5.371868
Equatorial Guinea
4.897425
Cluster: 2
Vietnam   Moldova Venezuela Sri Lanka Argentina
4.348344  4.310139  4.108363  4.043310  4.013278
Cluster: 3
United States     Ireland    Singapore  Netherlands   Luxembourg
4.925092     4.845664     4.800796     4.790307     4.751361
```

CONCLUSION

En conclusion de l'analyse précédente de **l'ACP** (Analyse en Composantes Principales) et des analyses subséquentes :

1. **Analyse des Composantes Principales (PCA) :**
   - L'ACP a permis de réduire la dimensionnalité des données en identifiant les combinaisons linéaires les plus significatives des variables initiales.
   - Les neuf premières dimensions (ou axes) ont expliqué plus de 90% de la variance totale des données.
   - Les individus ont été représentés dans un espace multidimensionnel, mettant en évidence leurs positions relatives par rapport aux axes principaux.
2. **Silhouette:**

- L'analyse de silhouette a révélé la qualité de la segmentation des données en clusters.
- Les résultats indiquent une bonne séparation des clusters, avec des valeurs de silhouette proches de 1 pour certains clusters, ce qui suggère une bonne cohésion intra-cluster et une bonne séparation inter-cluster.

3. **Hopkins Statistic:**
   - La statistique de Hopkins a été utilisée pour évaluer la tendance des données à former des clusters.
   - Une valeur élevée de la statistique de Hopkins (0.88 dans ce cas) suggère que les données ont une structure significative pouvant être regroupée en clusters.
4. **Analyse de Clustering:**
   - Les données ont été regroupées en trois clusters distincts.
   - Chaque cluster a été caractérisé par des profils socio-économiques spécifiques, avec des différences significatives dans des variables telles que le revenu, l'espérance de vie, la mortalité infantile, etc.

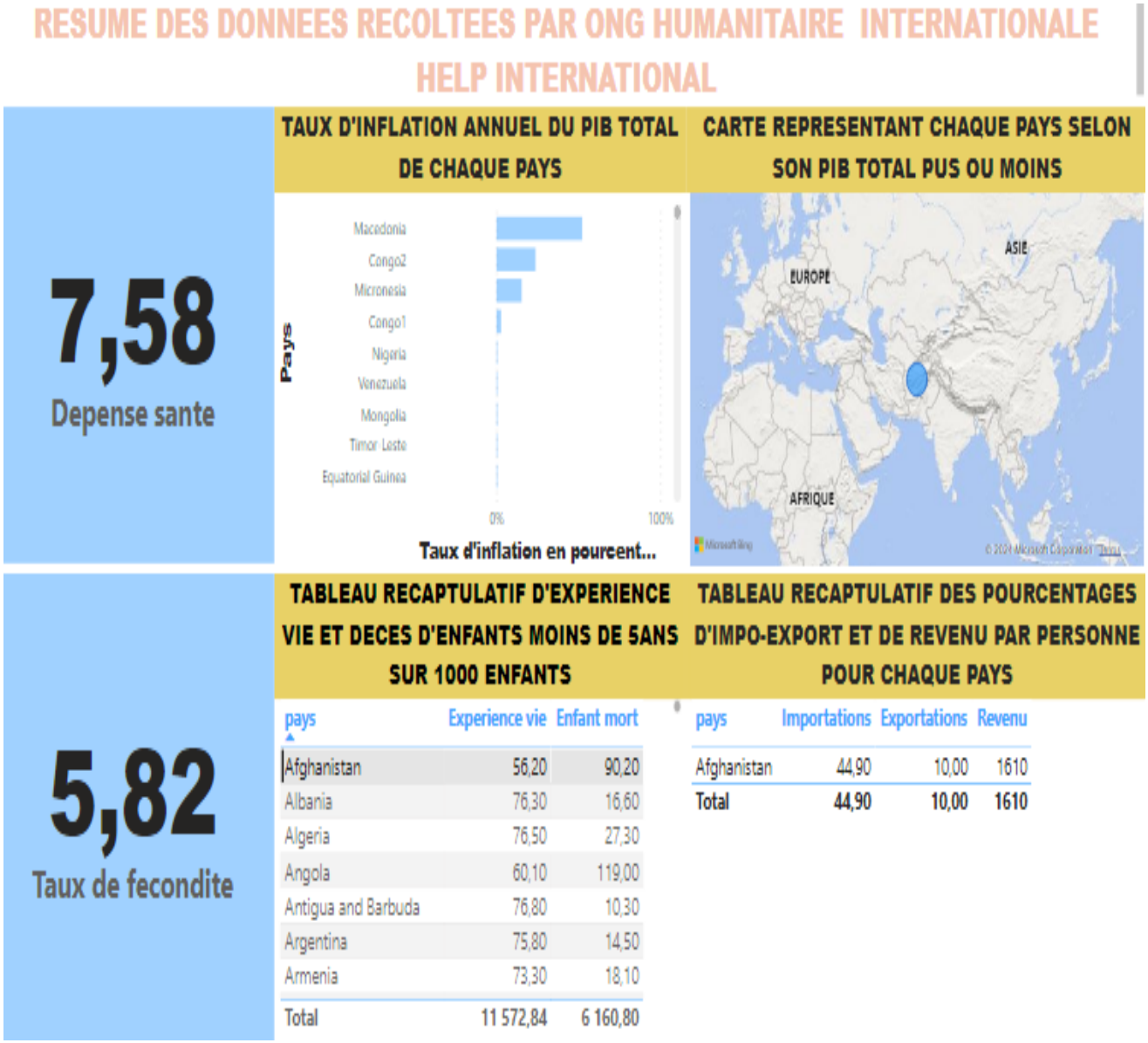
L'ACP et les analyses subséquentes ont permis de mieux comprendre la structure des données socio-économiques étudiées, en identifiant des tendances, des relations et des regroupements significatifs parmi les pays inclus dans l'étude. Ces résultats fournissent des indications précieuses pour la segmentation des pays en fonction de leurs caractéristiques socio-économiques et pour l'élaboration de politiques ciblées en conséquence. En portant notre attention sur les pays qui ont le plus besoin d'aide, les conclusions de l'analyse précédente de l'ACP et des analyses subséquentes fournissent des indications cruciales. Voici quelques points clés pour orienter les choix des pays nécessitant une assistance :

1. **Cluster 1:**
   - Ce cluster est caractérisé par des indicateurs socio-économiques relativement bas, tels qu'une forte mortalité infantile, un faible revenu, et une espérance de vie réduite.
   - Les pays inclus dans ce cluster nécessitent une assistance humanitaire et des investissements dans des domaines tels que la santé, l'éducation et le développement économique.
2. **Cluster 2:**
   - Les pays de ce cluster présentent des indicateurs socio-économiques moyens à modérés, mais avec des défis spécifiques tels qu'une espérance de vie inférieure à la moyenne ou des taux de fécondité élevés.
   - Bien que ces pays puissent ne pas nécessiter une aide d'urgence, ils pourraient bénéficier d'une assistance ciblée pour améliorer leur infrastructure de santé, leur éducation et leur développement économique à long terme.
3. **Cluster 3:**

- Ce cluster regroupe des pays avec des indicateurs socio-économiques relativement élevés, mais qui pourraient encore avoir besoin d'une aide ciblée dans certains domaines.
- Ces pays pourraient bénéficier d'un soutien pour renforcer leurs systèmes de santé, réduire les inégalités sociales, et promouvoir un développement économique durable.

En conclusion, les résultats de l'analyse fournissent une base solide pour identifier les pays nécessitant une aide prioritaire. Il est recommandé de concentrer les efforts sur les pays du Cluster 1, tout en fournissant un soutien supplémentaire et une coopération stratégique aux pays des Clusters 2 et 3 pour accélérer leur progression vers un développement socio-économique durable.

N.Bertrand Gueri. (2023-2024). *Help International.* Abidjan: INSSEDS.



Code source

```
library(knitr)

Dictionnaire=tibble::tribble(

 ~VARIABLE, ~NATURE, ~DESCRIPTION, ~MODALITES,

 "Pays", "Qualitative", "Nom du pays", "Chaine de caractère",

 "Enfant_mort", "Quantitative", "Décès d'enfants de moins de 5 ans pour 1000 naissances vivantes", "Numerique decimal",

 "exportations", "Quantitative", "Exportations de biens et services par habitant.Donné en pourcentage du PIB par habitant", "Numerique decimal",

 "dep_sante", "Quantitative", "Dépenses totales de santé par habitant.Données en pourcentage du PIB par habitant", "Numerique decimal",

 "importations", "Quantitative", "Importations de biens et services par habitant. Donné en pourcentage du PIB par habitant", "Numerique decimal",

 "revenu", "Quantitative", "Revenu net par personne", "Numerique entier",

 "taux_croissance", "Quantitative", "La mesure du taux d'inflation annuel du PIB total", "Numerique decimal",

 "esperance_vie", "Quantitative", "Le nombre moyen d'années qu'un nouveau-né vivrait si les tendances de mortalité actuelles devaient rester les mêmes", "Numerique decimal",

 "total_fertilite", "Quantitative", "Le nombre d'enfants qui naîtraient à chaque femme si les taux de fécondité par âge actuels devaient rester les mêmes.", "Numerique decimal",

 "pib_par_hab", "Quantitative", "Le PIB par habitant. Calculé comme le PIB total divisé par la population totale.", "Numerique entier"



)

knitr::kable(Dictionnaire)

#importation du Dataset

Help_I=read.csv("C:/Users/HP/Desktop/INSSEDS/Cours R/MiniProjetAnalyseMultidim/help_international.csv",sep=";",dec=".",row.names=1)

# Afficher les premières lignes pour vérifier

knitr::kable(head(Help_I))

##Structure du jeu de données

knitr::kable(str(Help_I))

knitr::kable(summary(Help_I))

##Visualisation et traitement des valeurs manquantes

# Fonction pour calculer la proportion de valeurs manquantes par variable

proportion_valeurs_manquantes <- function(data) {

# Calcul du nombre de valeurs manquantes par colonne

 nb_valeurs_manquantes <- sapply(data, function(x) sum(is.na(x)))

 # Calcul de la proportion de valeurs manquantes
```

```
  proportion_manquantes <- nb_valeurs_manquantes / nrow(data)
  # Création d'un dataframe pour le résultat
  resultat <- data.frame(Nombre = nb_valeurs_manquantes, Proportion = proportion_manquantes)
  return(resultat)
}
# Utilisation de la fonction avec votre base de données
resultat <- proportion_valeurs_manquantes(Help_I)
# Affichage du résultat
knitr::kable(resultat)
##visualisation
# Charger le package VIM
if (!require(dplyr)) install.packages("VIM")
library(VIM)
# Utilisation de la fonction aggr() pour visualiser les valeurs manquantes
aggr(Help_I, col=c('navyblue','yellow'), numbers=TRUE, sortVars=TRUE,
     labels=names(Help_I), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))
library("DataExplorer")
introduce(Help_I)
plot_intro(Help_I)
plot_str(Help_I)
plot_missing(Help_I)
###Traitement
library(DMwR2)
Help_I= knnImputation(Help_I, k = 10, scale = TRUE, meth = "median")
knitr::kable(head(Help_I))
##Visualisation après traitement
aggr(Help_I, col=c('navyblue','yellow'), numbers=TRUE, sortVars=TRUE,
labels=names(Help_I), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))
##Resume sans les valeurs manquantes
knitr::kable(summary(Help_I))
```

```
##Visualisation et traitement des valeurs extrême

##Visualisation

library(rpart)

par(mfrow=c(3,3), mar=c(3,3,3,3))

boxplot(Help_I$enfant_mort, col = "green",main="enfant_mort")

boxplot(Help_I$exportations, col = "yellow",main="exportations")

boxplot(Help_I$dep_sante, col = "red",main="dep_sante")

boxplot(Help_I$importations, col =  "orange",main="importations")

boxplot(Help_I$revenu,main="revenu")

boxplot(Help_I$taux_croissance, col = "#E7B800",main="taux_croissance")

boxplot(Help_I$life_expec, col = "orangered",main="life_expec")

boxplot(Help_I$total_fertilite, col = "skyblue",main="total_fertilite")

boxplot(Help_I$pib_par_hab, col = "lightcoral",main="pib_par_hab")

par(mfrow=c(1,1), mar=c(3,3,3,3))

#library(DescTools)

##Traitement des valeurs extrêmes et aberrantes

Help_I$enfant_mort=Winsorize(Help_I$enfant_mort)

Help_I$exportations=Winsorize(Help_I$exportations)

Help_I$dep_sante=Winsorize(Help_I$dep_sante)

Help_I$importations=Winsorize(Help_I$importations)

Help_I$revenu=Winsorize(Help_I$revenu)

Help_I$taux_croissance=Winsorize(Help_I$taux_croissance)

Help_I$life_expec=Winsorize(Help_I$life_expec)

Help_I$total_fertilite=Winsorize(Help_I$total_fertilite)

Help_I$pib_par_hab=Winsorize(Help_I$pib_par_hab)

###Visualisation après traitement

par(mfrow=c(3,3), mar=c(3,3,3,3))

boxplot(Help_I$enfant_mort, col = "green",main="enfant_mort")

boxplot(Help_I$exportations, col = "yellow",main="exportations")

boxplot(Help_I$dep_sante, col = "red",main="dep_sante")
```

```
boxplot(Help_I$importations, col = "orange",main="importations")

boxplot(Help_I$revenu,main="revenu")

boxplot(Help_I$taux_croissance,col = "#E7B800",main="taux_croissance")

boxplot(Help_I$life_expec, col = "orangered",main="life_expec")

boxplot(Help_I$total_fertilite, col = "skyblue",main="total_fertilite")

boxplot(Help_I$pib_par_hab, col = "lightcoral",main="pib_par_hab")

par(mfrow=c(1,1), mar=c(3,3,3,3))

plot_histogram(Help_I)

plot_qq(Help_I)

library("DataExplorer")

plot_correlation(Help_I, cor_args = list("use" = "pairwise.complete.obs"))

library(corrplot)

cor=cor(Help_I)

corrplot(cor)

library(GGally)

ggpairs(Help_I)

##ACP

library(FactoMineR)

res.pca=PCA(Help_I)

knitr::kable(summary(res.pca))

plot_prcomp((Help_I), nrow = 2L, ncol = 2L)

library(factoextra)

library(corrplot)

fviz_eig(res.pca, addlabels = TRUE)

res=get_pca_var(res.pca)

corrplot(res$cos2)

fviz_pca_var(res.pca, col.var = "cos2",

             gradient.cols=c("#00AFBB","#E7B800", "#FC4E07"),

             repel=TRUE

             )
```

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

fviz_pca_ind(res.pca,col.ind = "cos2",

             gradient.cols=c("#00AFBB","#E7B800", "#FC4E07"),

             repel=TRUE)

#Classification des donnees

Classe=sample(1:dim(Help_I)[1],50)

Help_Isample=Help_I[Classe,]

#Calcul de la statistique Hopkins

get_clust_tendency(Help_I,graph = FALSE, n=50, seed = 123)
```

```
src=source(url("https://raw.githubusercontent.com/larmarange/JLutils/master/R/clustering.R"))

best=best.cutree(hc)

best

library(dendextend)

plot(hc, main= "Partition en 3 classe", xlab="", sub="", axes=FALSE, hang= - 1)

rect.hclust(hc, k=best, border = "blue")

groups= cutree(hc,k=best)

knitr::kable(groups)

set.seed(123)

#Executation de l'algorithme k-means avec k=3

kmeans_out=kmeans(Help_I, centers = best, nstart = 50)

kmeans_out

# Visualisation des résultats

fviz_cluster(kmeans_out, data = Help_I)

library(factoextra)

fviz_dend(hc, k=best, show_labels = FALSE, rect = TRUE)

#Calcul de l'indice Silhouette

library("cluster")

si<-silhouette(kmeans_out$cluster, dist(Help_I, "euclidean"))

summary(si)

#Centre de gravité de chaque cluster

centroide <- kmeans_out$centers

centroide

#Indice moyen observé chez les patients pour chaque groupe

for (i in 1:3) {

 print(sum(centroide[i, ]))

}

library(FactoMineR)

res<-PCA(Help_I)

res.hcpc<-HCPC(res,nb.clust=best,graph=TRUE)
```

```
names(res.hcpc)

# AFFECTATION DE CHAQUE INDIVIDU A UNE CLASSE (CLUSTER)

cluster = res.hcpc$data.clust

knitr::kable(head(cluster))

#extraction des individus par groupe

cluster = res.hcpc$data.clust

groupe_1 = subset(cluster, clust ==1)

groupe_2 = subset(cluster, clust ==2)

groupe_3 = subset(cluster, clust ==3)

# CARACTERISATION DES CLUSTERS PAR LES VARIABLES

res.hcpc$desc.var

# CARACTERISATION DES CLUSTERS PAR LES DIMENSIONS (axe)

res.hcpc$desc.axes

# IDENTIFICATION DES PARAGONS (individu le plus proche du centre des classes

res.hcpc$desc.ind
```