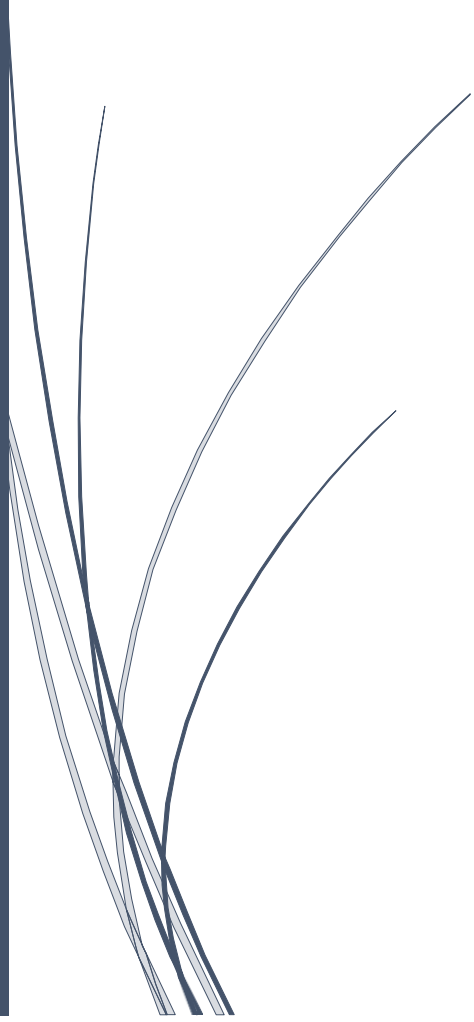
2023

Etude de 119 pays analyse du ‘bon vivre’

Méthodes descriptives



BENDRISS Oussama

Table des matières

Introduction

L'étude porte sur 119 pays et tente d'établir une catégorie de pays où il fait « bon vivre ». Étant subjective, la définition de cette notion est importante et passe par une évaluation basée sur nos propres critères : l'espérance de vie à la naissance, la superficie forestière, l'accès à l'eau potable et à l'électricité et à internet, la consommation de l'alcool, l'exposition à la pollution et la santé en étudiant les dépenses de chaque pays en la matière et différentes causes de décès.

Nous précisons que toutes les conclusions qui seront faites dans cette étude visent uniquement la liste des 119 pays[1]. Nous partons du postulat que la notion de « bon vivre » est également liée à la présence d'importantes surfaces forestières.

Plus concrètement, nous essayerons de répondre à plusieurs questions. Quels sont les pays où il fait « bon vivre » ? qu'en est-il de leurs caractéristiques ? Quels sont les spécificités des pays où il fait le moins bon vivre ? En quoi sont ils différents des autres pays ? Les causes de décès entre les pays sont-elles similaires ou différentes ? Existe-t-il des causes de décès attribuables à certains groupes plutôt que d'autres ?

Nous allons tenter de répondre à toutes ces questions à l'aide des méthodes d'analyse de données descriptives abordées lors des scéances STA101. Nous précisons que l'établissement des sorties et des résultats se fera sous R[2].

[1] Annexe 1 : liste des pays étudiés

[2] Code en Annexe 2

I. Construction de la base de données

1. Import de la première base de données

https://www.kaggle.com/datasets/vrec99/life-expectancy-2000-2015?sort=votes

Elle contient :

- 1904 observations : une liste de plusieurs pays sur 15 ans.
- 16 variables :

Nom de la variable	Type de variable	Contenu
Year	Quantitative	De 2000 a 2015
Continent	Qualitative	Nom des continents
Least Developed	Qualitative	Si la valeur est VRAI, le pays est classé comme « sous développé », si elle est FAUX, le pays n'est pas classé comme « sous développé ».
Life Expectancy	Quantitative	En nombre d'années par habitant
CO2 emissions	Quantitative	Nombre de tonnes d'émissions CO2 par habitant
Health expenditure	Quantitative	Dépenses en santé en pourcentage du PIB
Forest area	Quantitative	En % de la surface du pays couverte par la foret
GDP per capita	Quantitative	PIB divisé par le nombre d'habitants
Individuals using the Internet	Quantitative	En % de la population utilisant internet
Military expenditure	Quantitative	En % du PIB alloué aux dépenses militaires
People practicing open defecation	Quantitative	En % de la population
People using at least basic drinking water services	Quantitative	En % de la population
Obesity among adults	Quantitative	En % de la population
Health expenditure per capita	Quantitative	Dépenses en santé par personnes
Total alcohol consumption per capita	Quantitative	Litres d'alcool pur pour les personnes de 15 ans et plus

En ce qui concerne la variable 'Least Dev', elle sera traduite littéralement et par simplification dans la suite de notre étude par le terme « sous développé ». Cette notion est définie par les nations unis[3] pour désigner les pays qui présentent le score de développement le plus bas dans le monde qui est donc bien différente de la notion de pays « en voie de développement ». Il a été également fait le choix de ne pas enrichir cette variable par d'autres qualifications comme pays « en voie de développement », « pays émergeants » ou encore « développés » car nous souhaitant pas influencer notre choix de pays où il ferait « bon vivre » par ces modalités mais uniquement par la modalité « sous développé » définie par les nations unis.

2. Suppression des variables

Pour 3 principales raisons :

Neutraliser l'aspect temporel :

- 'Year' : Afin d'éliminer l'aspect temporel dans notre étude, nous décidons de nous restreindre à l'année 2014 cela semble être un bon compris entre l'exhaustivité des données (le faible taux de données manquantes) ainsi que leur récence.

Variables non exhaustives et/ou inadaptées pour notre étude :

[3] https://unctad.org/topic/least-developed-countries/list

- ‘Electric power consomption’ : sera remplacée par un indicateur sur l’accès a l’électricité. La consommation absolue de l’électricité ne pourrait constituer par un facteur déterminant pour notre étude.
- ‘Population’, ‘People practicing open defecation’ et ‘Military expendure’ ne présentent aucun lien avec les critères du « bon vivre » tel qu’on l’a défini au début de notre étude.

Variables non adaptées à notre étude :
- ‘Beer consumption per capita’ : variable qui nous parait non exhaustive, car prend pas en considération la consommation d’autres types d’alcool qui peuvent être plus ou moins significatifs en fonction des habitudes de consommation de chaque population.

3. Ajout de variables

Le premier data set ne permettra pas de répondre à l’ensemble des questions posées. À ce titre, il sera enrichi par des données (variables) provenant des bases de données de la banque mondiale : https://data.worldbank.org. Source de données ayant de nombreux critères sur la qualité de collecte et établissement des données statistiques. Cette source de données permet ainsi d’inclure plusieurs données de santé publique dans la base d’étude.

Nom de la variable	Type de variable	Contenu
Alcohol.Cons.Capita	Quantitative	Litres d'alcool pur pour les personnes de 15 ans et plus
Popul.polut.PM2.5	Quantitative	% de la population exposé a la pollution supérieur au seuil défini par le critere PM2,5 (*toutes les particules dans l'air ayant un diamètre aérodynamique inférieur ou égal à 2,5 µm*).
Mort.Water.Sanita.Hygiene	Quantitative	Taux de mortalité attribué à l'eau insalubre, à l'assainissement insalubre et au manque d'hygiène (pour 100 000 habitants).
Death.com.diseases	Quantitative	Cause de décès, par maladies transmissibles et conditions maternelles, prénatales et nutritionnelles (En % de la pop).
death.Non-com.diseases	Quantitative	Decès du a des maladies 'non transmissibles' : maladies chroniques et blessures. (En % de la pop).
Elec.Acess	Quantitative	En % de population

4. Base de données finale

Après consolidation, notre base de données se compose de 119 observations (pays) et des 17 variables suivantes :

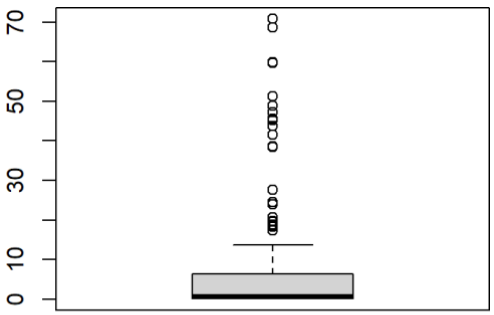
Country	Qualitative
Continent	Qualitative
Least.Dev	Qualitative
Life.Expect	Quantitative
CO2.Emiss	Quantitative
health expend_%_GDP	Quantitative
Forest.Area	Quantitative
GDP.Capita	Quantitative
Indiv.Inter	Quantitative
Obes.adul	Quantitative
Water.Acess	Quantitative
Alcohol.Cons.Capita	Quantitative
Popul.polut.PM2.5	Quantitative
Mort.Water.Sanita.Hygiene	Quantitative
Death.com.diseases	Quantitative
death.Non-com.diseases	Quantitative
Elec.Acess	Quantitative

II. Analyse univariée

1. Variables quantitatives

	Variables quantitatives	Min	Max	Moyenne	Ecart type	Coefficient de variat (Moy /
1	Life.Expect	53	83	73	7	
2	CO2.Emiss	0	33	5	5	
3	health expend_%_GDP	2	17	7		
4	Forest.Area	-	92	30		
5	GDP.Capita	850	107 860			
6	Indiv.Inter	1				
7	Obes.adul	3				
8	Water.Acess					
9	Alcohol.Cons.Capita					
10	Popul.polut.PM2.5					
11	Mort.Water.Sanita.H					
12	Death.com.di					
13	death.					
14						

Le faible coefficient de variation de 'Life.Expect' indique l'existence de valeurs bien proches et une faible dispersion. L'intuition nous amène /nous conduit à penser qu'il a une bonne homogénéité dans notre échantillon pour cette variable mais aussi un déséquilibre car on aurait des pays avec une espérance de vie plutôt proche.

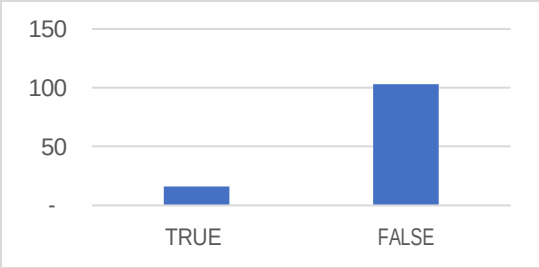


Mort.Water.Sanita.Hygiene

A contrario, la variable **'Mort.Water.Sanita.Hygiene'** présente un fort ratio en CV peut-être du à un déséquilibre dans la population étudiée (pays).
La visualisation des valeurs de cette variable sous forme de boite à moustache s'avère utile et indique l'existence d'un bon nombre de valeurs extrêmes qui dépassent les 1,5 fois la distance interquartiles. Les trois individus dont les valeurs sont les plus extrêmes sont le Niger, le Nigéria et la république démocratique du Congo. Les pays prenant les valeurs les plus extrêmes semblent présenter les espérances de vie les plus faibles de notre population et sont souvent des pays considérés « sous-développés ». Nous approfondirons ces liens dans la partie dédiée à l'analyse bivariée. Nous considérons donc que ces valeurs ne sont pas dues à des erreurs de mesure en les gardant dans notre étude.

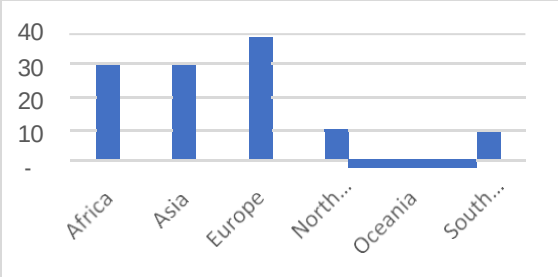
2. Variables qualitatives

‘Least Developed’



Notre population inclura 13 % de pays « sous-développés ».

‘Continent’



On observe que 81 % des pays étudiés se répartissent sur 3 continents. Les pays du continent européen sont les plus représentés (38 individus) et sont suivis par les pays du continent asiatique (30) puis africains (28). Deux pays sont du continent océanique.

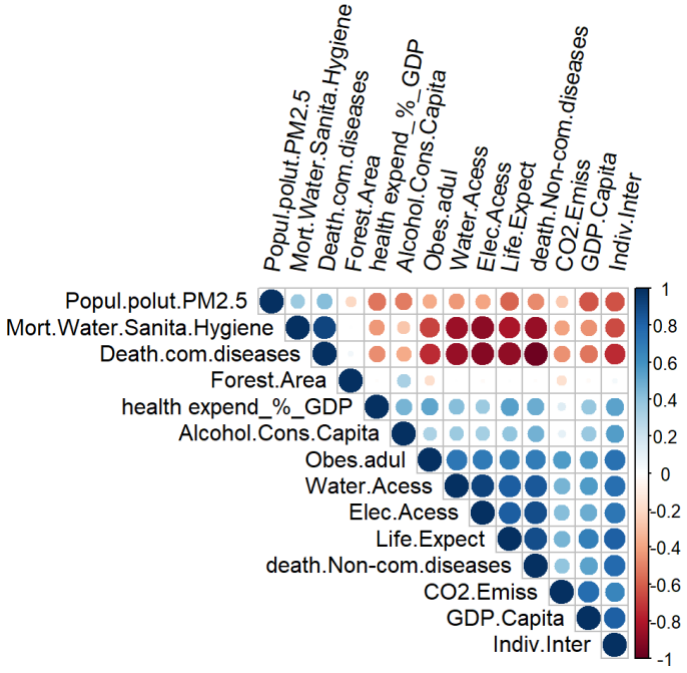
III. Analyse bivariée

1. Analyse du lien entre les variables quantitatives

Dans un premier temps, nous étudions la corrélation entre les variables quantitatives avec la **méthode Pearson** :

D’un côté nous constatons une corrélation positive forte entre ‘Mort.Water.Sanita.Hygiene’ et ‘Death.com.diseases’. De l’autre côté, une corrélation négative entre ces deux dernières et toutes les autres variables quantitatives.

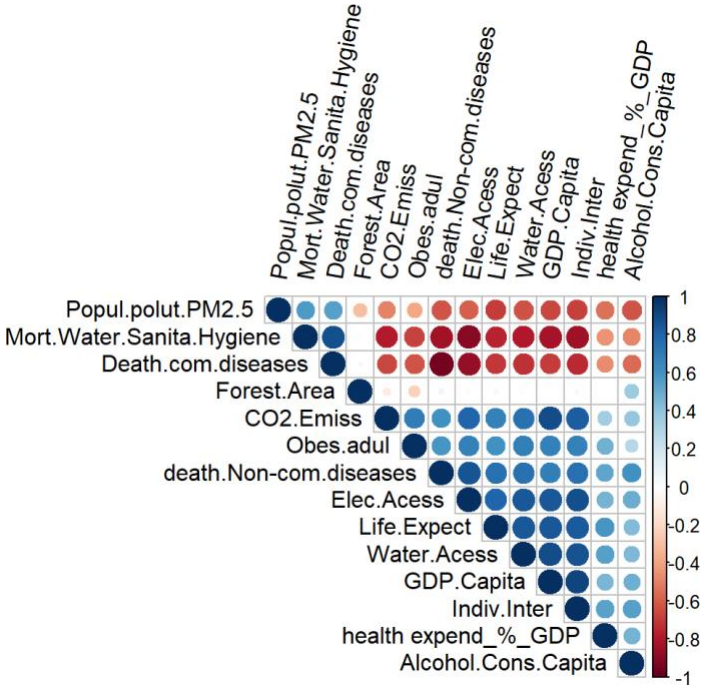
On observe également une faible corrélation entre la variable ‘Forest.Area’ et toutes les autres variables quantitatives.
‘Life.Expect’, ‘death.Non-com.diseases’ ‘Water.Acess’ ‘Indiv.Inter’ ‘death.Non-com.diseases’ et ‘Obes.adul’



forment un groupe de variables fortement et positivement corrélées.

On peut donc en déduire que les pays dont les habitants ont une mortalité élevée à cause de l'accès à l'eau et aux conditions d'hygiènes ont tendance à avoir aussi une mortalité élevée liée aux maladies infectieuses, aux conditions maternelles, périnatales et nutritionnelles. La surface forestière ne semble pas avoir un lien significatif sur l'espérance de vie ou les causes de mortalités. L'espérance de vie semble avoir un lien fort avec l'accès à l'eau, à l'électricité, à internet et aux morts dus à des maladies non transmissibles (maladies chroniques) et dans une moindre mesure et de façon contre-intuitive avec l'obésité chez les adultes. L'intuition nous suggère que la corrélation positive entre la mortalité due à des maladies non transmissibles (chroniques[4]) et est dû au fait que les maladies chroniques ont tendance a plus survenir a des âges plus élevés.

Dans un second temps, nous étudions la corrélation entre les variables quantitatives avec la **méthode Spearman** :



Les conclusions établies précédemment selon la méthode de Pearson se confirment.

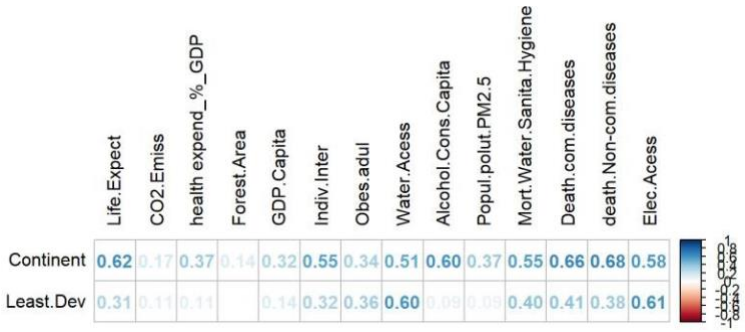
Les coefficients de Spearman sont globalement plus élevés que ceux de Pearson ce qui pourrait s'expliquer par des liaisons non-linéaires entre les variables.
À titre d'exemple : 'Popul.pomut.PM2.5' a une corrélation (Spearman) plus forte avec les variables : 'Water.Acess', 'Mort.Water.Sanita.Hygiene', 'Death.com.diseases', 'CO2.Emiss' , 'death.Non-com.diseases', 'Elec.Acess', 'Life.Expect' , 'GDP.Capita', 'Indiv.Inter' et 'Alcohol.Cons.Capita' ,

Il en va de même quant à la corrélation entre d'un côté 'Mort.Water.Sanita.Hygiene' et 'CO2.Emiss', 'GDP.Capita', 'Alcohol.Cons.Capita'.

[4] On entend par maladies non transmissibles les maladies chroniques : https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

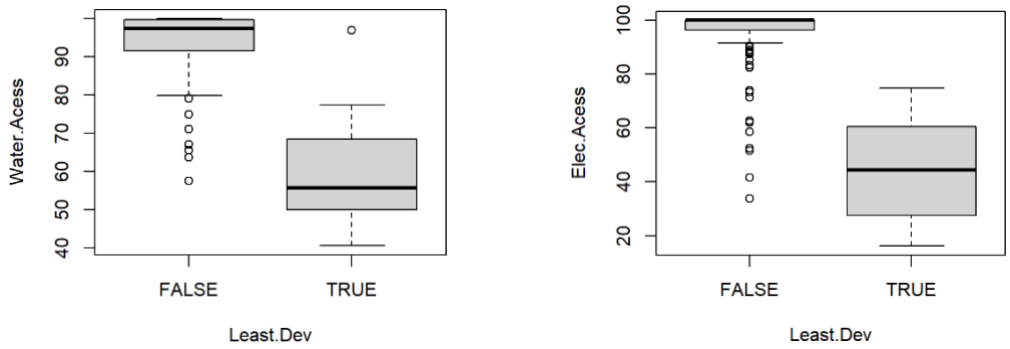
2. Analyse du lien entre variables qualitatives et variables quantitatives5

On utilise le rapport de corrélation eta2[6] pour mesurer l'intensité de la liaison entre une variable quantitative et une variable qualitative



En ce qui concerne la variable 'Least.Dev', les deux variables 'Elec.Acess' et 'Water.Acess' contribuent de façon relativement importante à sa variance. On peut donc dire qu'il y a un lien statistiquement significatif entre les modalités de la variable donc le fait que le pays soit « sous développé » pas et l'accès de la population à l'eau et à l'électricité.

La vue des boites à moustache confirme bien notre analyse :

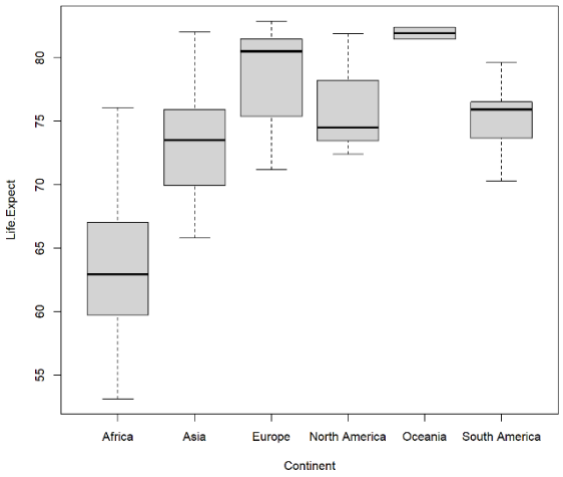


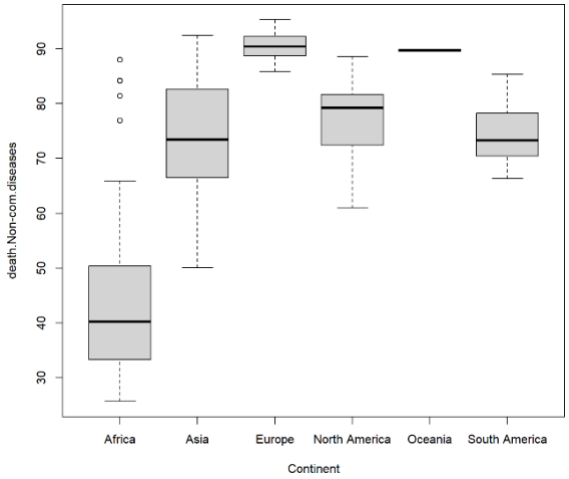
Pour ce qui est de la variable 'Continent', les trois variables qui contribuent le plus à la variance sont celles relatives à l'espérance de vie, le nombre de morts par les maladies transmissibles et le nombre de morts à cause des maladies non transmissibles. Ces dernières peuvent bien caractériser les continents auxquels se rattachent nos individus.

Nous établissons les boites à moustaches qui nous permettent de visualiser les valeurs des trois variables quantitatives en fonction des 6 modalités de la variable qualitative 'Continent'.

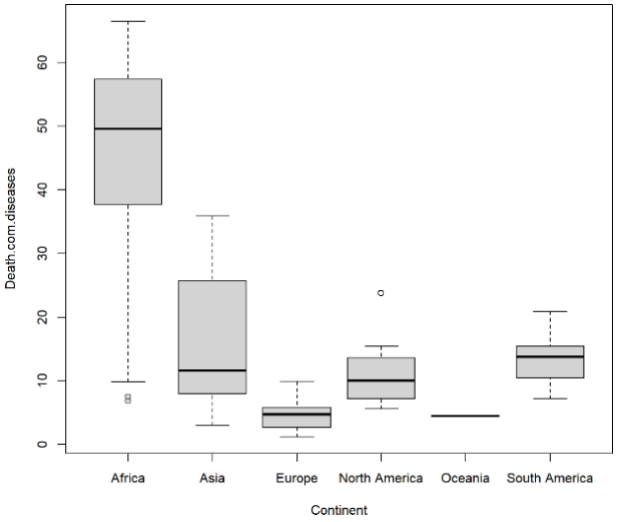
[5] Comment étudier la liaison entre une variable quantitative et une variable qualitative ? | ALLIAGE (alliage-ad.com)

[6] Le calcul est effectué avec la fonction eta2() sous R qui établit rapport de corrélation qui est une mesure d'association importante entre une variable quantitative et une variable qualitative.





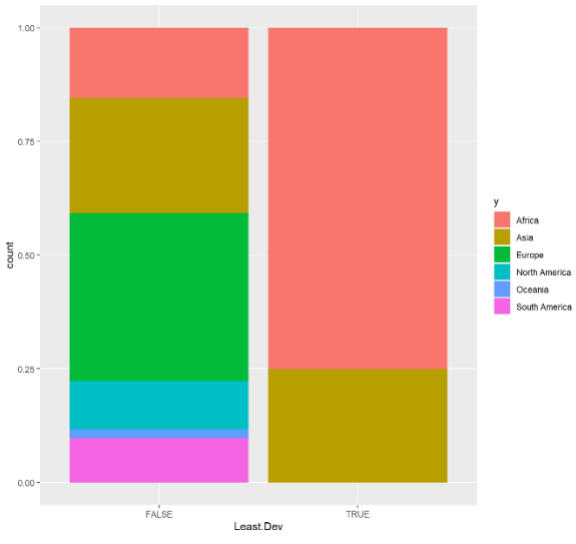
La modalité 'Africa' de la variable qualitative 'Continent' semble se distinguer significativement des autres modalités. Elle présente des valeurs bien différentes par rapport aux autres continents. Les pays africains présentent donc des caractéristiques différentes aussi bien en terme d'espérance de vie que des causes de mortalités qui les distinguent des pays des autres continents.



L'Amérique du nord et l'Amérique du sud semble présenter un profil relativement similaire comparé aux autres continents en termes d'espérance de vie et de causes de moralité. Les maladies transmissibles semblent avoir un important impact sur l'espérance de vie en Afrique et Asie.

3. Analyse du lien entre variables qualitatives

Le constat est sans équivoque, sur la base du graphique et les résultats du test du chi deux (valeur de 30,28 avec une p-value à 1.294e-05) largement supérieur au chi deux critique (15 avec alpha à 0,01), on peut dire qu'il y a un lien de dépendance statistiquement significatif entre ces deux caractères.
Les pays « sous-développés » appartiennent majoritairement (75%) au continent africain et dans une moindre mesure au continent asiatique (25%).



IV. Choix des individus et des variables actifs et illustratifs

1. Critères de choix des variables illustratives

Pour cela nous considérons les variables ci-dessous comme des valeurs illustratives :

2. Sélection des variables illustratives

Pour mesurer le niveau de pollution, il nous avait semblé pertinent de choisir, lors de la collecte des données nécessaires, deux variables. La première 'CO2.Emiss' nous permet de mesurer les émissions de CO2 au niveau de chaque pays et 'Popul.pomut.PM2.5' pour mesurer l'exposition des habitants au niveau le plus élevé de la pollution. Il est donc intéressant de constater que ces deux variables sont faiblement corrélées avec la méthode Pearson (-0.26) et un peu mieux corrélées en Spearman (-0.49). En définitive, quelle que soit la méthode retenue, les deux variables ne pas sont pas assez corrélées pour en conclure si redondance il y a. En tout état de cause, l'intérêt est de mesurer le « bon vivre » dans les différents pays, nous préférerons donc nous baser sur la variable qui mesure le taux de la population exposé à la pollution aux particules fines et considérer la variable 'CO2 émission' en tant que variable illustrative.

'Continent' variable qualitative comme indicateur et ne nous apportera pas de caractéristique recherchée ou déterminante pour notre analyse. Elle sera donc considérée comme une variable illustrative. L'intérêt de son maintien dans le jeu de données est de pouvoir étudier les éventuelles ressemblances et différences avec les groupes d'individus que l'on constituera dans notre étude et les modalités de la variable 'Least.Dev'.

Nous garderons la variable quantitative 'Forest area' et la surveillerons car elle nous semblait, lors de l'analyse bivariée, se démarquer en raison de sa très faible corrélation avec les autres variables. Peut-être qu'elle pourrait donc contribuer à séparer les pays étudiés d'une façon différente.

In Fine, les variables actives étant à la fois quantitatives et qualitatives, le choix d'une AFDM pour une analyse factorielle s'impose.

3. Sélection des individus

Nous faisons le choix d'étudier tous les individus.

V. Analyse multidimensionnelle

1. Classification non-supervisée

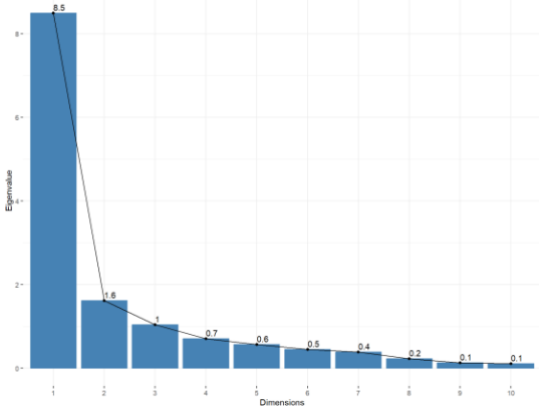
i. Choix du nombre de dimensions à retenir pour notre approche tandem

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	8.495	60.6785	60.6785
comp 2	1.6182	11.5586	72.2371
comp 3	1.0461	7.4722	79.7094
comp 4	0.7103	5.0734	84.7828
comp 5	0.5744	4.1026	88.8854
comp 6	0.4541	3.2437	92.1291
comp 7	0.392	2.8001	94.9292
comp 8	0.2307	1.6476	96.5768
comp 9	0.1327	0.9481	97.5249
comp 10	0.1122	0.8011	98.326
comp 11	0.1011	0.7223	99.0484
comp 12	0.073	0.5213	99.5697
comp 13	0.0469	0.3353	99.905
comp 14	0.0133	0.095	100

Tableau des valeurs propres

À première vue, le tableau des valeurs propres nous permet de constater que l'inertie est concentrée de façon significative sur les deux premiers axes (ils totalisent 72 % de l'inertie totale du nuage). Cette situation n'est pas typique de l'AFDM, néanmoins elle s'explique par la présence d'une seule variable quantitative active ('Least.Dev') et son faible nombre de modalités (2) à savoir 'True' ou 'False'.

L'établissement d'un éboulis nous permettra la visualisation des valeurs propres ce qui facilitera la sélection du nombre adéquat des dimensions à retenir.



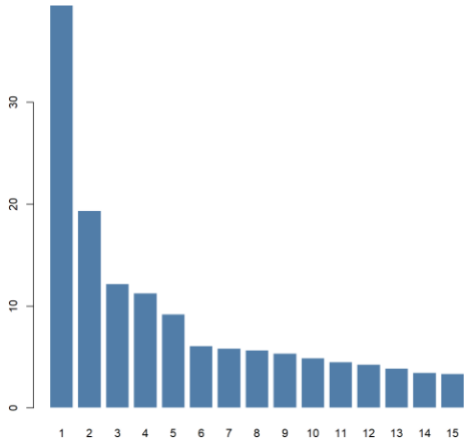
Éboulis des valeurs propres

Ce graphique nous permet de visualiser la concentration de l'inertie sur un nombre très limité d'axes. Il nous parait donc plus pertinent de privilégier la « règle du coude » sur les dernières dimensions ce qui maximisera notre inertie tout en minimisant le nombre d'axes plutôt que de fixer un seuil d'inertie à avoir et déterminer le nombre d'axes correspondant.

L'éboulis des valeurs propres nous montre un premier coude le 8ieme et 7ieme dimension. Néanmoins, ce nombre nous semble élevé compte tenu du niveau de concentration de l'inertie sur les premiers axes. Nous remarquons un second coude se démarquer entre la quatrième et troisième dimension. L'application de la « règle du coude » nous suggère cette fois de conserver les 3 premières dimensions totalisant ainsi une inertie proche de 80%. Cela impliquerait de perdre 20 % de l'information du nuage que l'on aurait obtenu en retenant 14 dimensions. Nous faisons le choix donc de retenir trois dimensions.

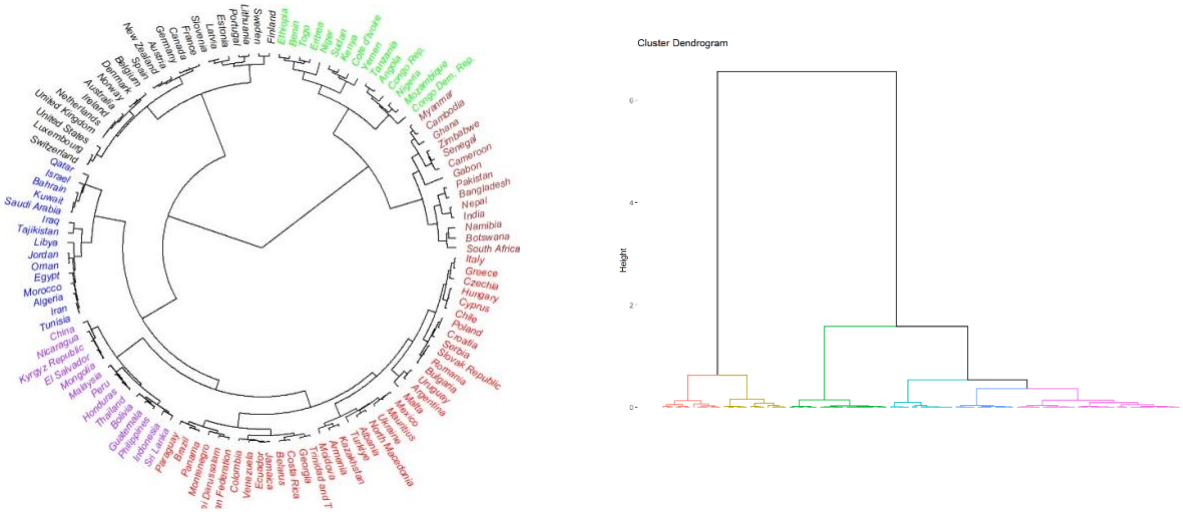
ii. Choisir le nombre de classes approprié

Sur la base des trois dimensions, nous procéderons à la réalisation d'une classification ascendante hiérarchique (CAH). Cette étape nous permettra de trouver le nombre de classe optimal pour notre étude.



Hauteurs de fusion en fonction du nombre agrégations

Le diagramme de gain d'inertie montre une première importante chute d'inertie interclasse entre la 5ieme et la 6ieme classe, cela veut dire qu'il est compliqué de passer de 6 à 5 classes. Nous décidons donc de retenir 6 classes et de visualiser la classification de nos individus en fonction de ce nombre de classes.



Dendrogrammes représentés de façon différentes le premier sous forme d'un ventilateur[7] après le découpage en 6 classes

Ayant défini le nombre de classes optimal, on réalise par la suite une consolidation par k-means. Nous obtenons les effectifs suivants :

[7] Ventilateur par référence au paramètre sous R utilisé dans la fonction plot l'argument type = "fan".

La classe 1 concentre le plus grand nombre de pays et est suivie par la classe 4. Les classes ayant le moins d'individus sont la classe 5 et 3.

2. Description des classes

On décrit les 6 classes retenues à partir des variables[8]. Cette étape est incontournable pour une bonne description des classes.

Pour l'interprétation de nos classes, nous fixons comme seuil de significativité aussi bien pour les variables quantitatives que qualitatives un v.test de :

- 2 pour les variables actives
- 3 pour les variables illustratives car ne participent pas par définition à la construction des axes.

iii. Variables qualitatives

Nous proposons de commencer par décrire nos classes en fonction des variables qualitatives actives et illustratives :

Classe 1 : Regroupe une bonne partie des pays européens (53% des pays européens sont dans la classe) et représentent 60% des individus de cette classe. A contrario, les pays asiatiques (6% de notre classe) et africain (inexistants) sont sous représentés dans ce groupe. Il est à noter qu'aucun individu de cette classe n'est « sous-développé ».

Classe 2 : Elle se caractérise par la surreprésentation des pays asiatiques (70% des individus de la classe[9]) et par l'absence des pays européens.

Classe 3 : Les pays africains sont surreprésentés car tous les pays de cette classe sont africains et représentent 53% de l'ensemble des pays de ce continent. Les pays « sous-développés » sont surreprésentés et constituent 73%[10] de la classe 3.

Classe 4 : La représentativité très élevé des pays du continent européen caractérise cette classe (75% de classe soit 47% de l'ensemble des pays de ce continent). Les pays du continent africain et asiatiques sont inexistants.

Classe 5 : Se caractérise par une surreprésentation des pays « sous-développés ».

Classe 6 : Est marquée par l'absence des pays européens.

[8] Voir résultats dans l'annexe 3
[9] Equivaut a 40% de l'ensemble des pays asiatiques
[10] Equivaut a 68% des pays « sous-développés » de notre population

iv. Variables quantitatives :

Nous continuons à décrire nos classes cette fois à l'aide des variables quantitatives actives et illustratives :

Classe 1 : Se caractérise principalement par une mortalité élevée pour cause de maladies chroniques (non transmissibles), en consommation d'alcool, accès électricité et eau mais aussi en surface forestière et une faiblesse marquée en mortalité à cause des conditions sanitaires et celle liée aux maladies transmissibles.

Classe 2 : Se démarque essentiellement par des valeurs bien élevées en obésité chez les adultes et en émission en Co2 et des valeurs très faibles en surface forestières et en consommation d'alcool.
Dans une bien moindre mesure la part des habitants exposés à la pollution aux particules fines et avec accès à l'eau et l'électricité y est relativement important. A contrario, relativement limité en mortalité due aux maladies transmissibles et celle due aux conditions sanitaires.

Classe 3 : Est singulière principalement par le niveau extrêmement élevé en mortalité pour causes sanitaires et celle due aux maladies transmissibles. Inversement, des valeurs extrêmement faibles en ce qui concerne l'accès à l'eau, l'électricité et internet mais aussi en maladies chroniques en obésité et en espérance de vie.
De façon moins importante, cette classe a des valeurs limitées en PIB par habitant et en émission CO2.

Classe 4 : Se caractérise principalement par le niveau élevé en PIB par habitant , de la part des dépenses en santé par apport au PIB et d'accès a internet. L'espérence de vie est importante parralellement a la consommation de l'acool et en mortalité due aux maladies chroniques. L'obésité pour les adultes y est relativement considérable. Cette classe a des valeurs extrêmement faibles en exposition en particules fines et dans une moindre mesure faible en mortalité pour causes de maladies transmissibles.

Classe 5 : Se démarque essentiellement par l'importance en mortalité pour maladies transmissibles et relativement supérieure en exposition aux particules fines. Inversement, Bien faible en obésité chez les adultes, mortalité en maladies non transmissibles et accès internet et en espérance de vie.

Classe 6 : Est relativement faible en accès internet, en obésité et PIB par habitant.

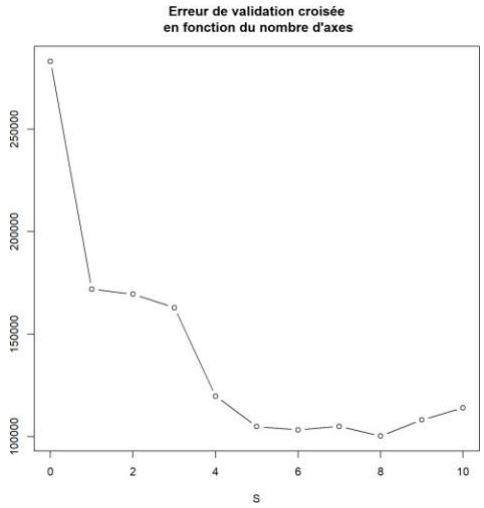
3. Analyse factorielle (AFDM)

Nous proposons d'intégrer les classes préalablement constituées en tant que variable illustrative dans notre jeu de donnée. Dès lors, nous pourrions en tirer avantage pour une

analyse factorielle des données mixtes. Ainsi nous aurons exploité la complémentarité entre d'un côté la classification non supervisée et notre AFDM.

i. Nombre de dimensions pour AFDM

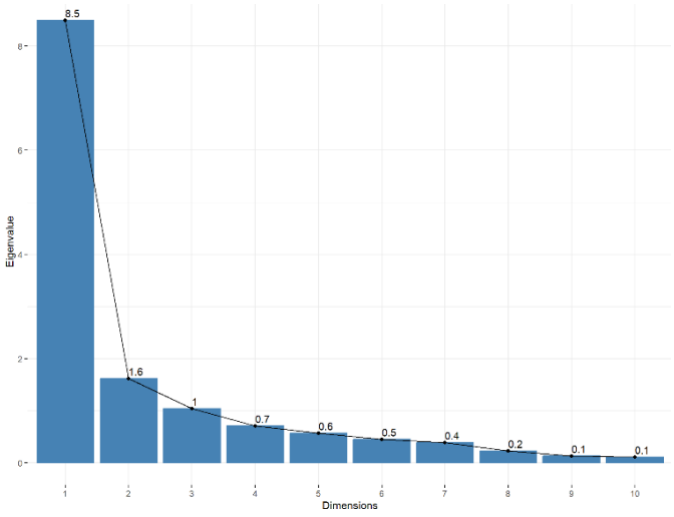
En utilisant la technique de validation croisée pour choisir le nombre d'axes on obtient les résultats suivants :



La technique de validation croisée obtenue avec la fonction 'fviz_mfa_ind' sous R nous suggère de considérer entre 5 et 8 dimensions pour notre AFDM car ces derniers minimiseraient l'erreur. Nous comptions donc chercher le nombre de dimension optimal en visualisant le tableau ainsi que l'éboulis des valeurs propres.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	8.495	60.6785	60.6785
comp 2	1.6182	11.5586	72.2371
comp 3	1.0461	7.4722	79.7094
comp 4	0.7103	5.0734	84.7828
comp 5	0.5744	4.1026	88.8854
comp 6	0.4541	3.2437	92.1291
comp 7	0.392	2.8001	94.9292
comp 8	0.2307	1.6476	96.5768
comp 9	0.1327	0.9481	97.5249
comp 10	0.1122	0.8011	98.326
comp 11	0.1011	0.7223	99.0484
comp 12	0.073	0.5213	99.5697
comp 13	0.0469	0.3353	99.905
comp 14	0.0133	0.095	100

Tableau des valeurs propres



Éboulis des valeurs propres

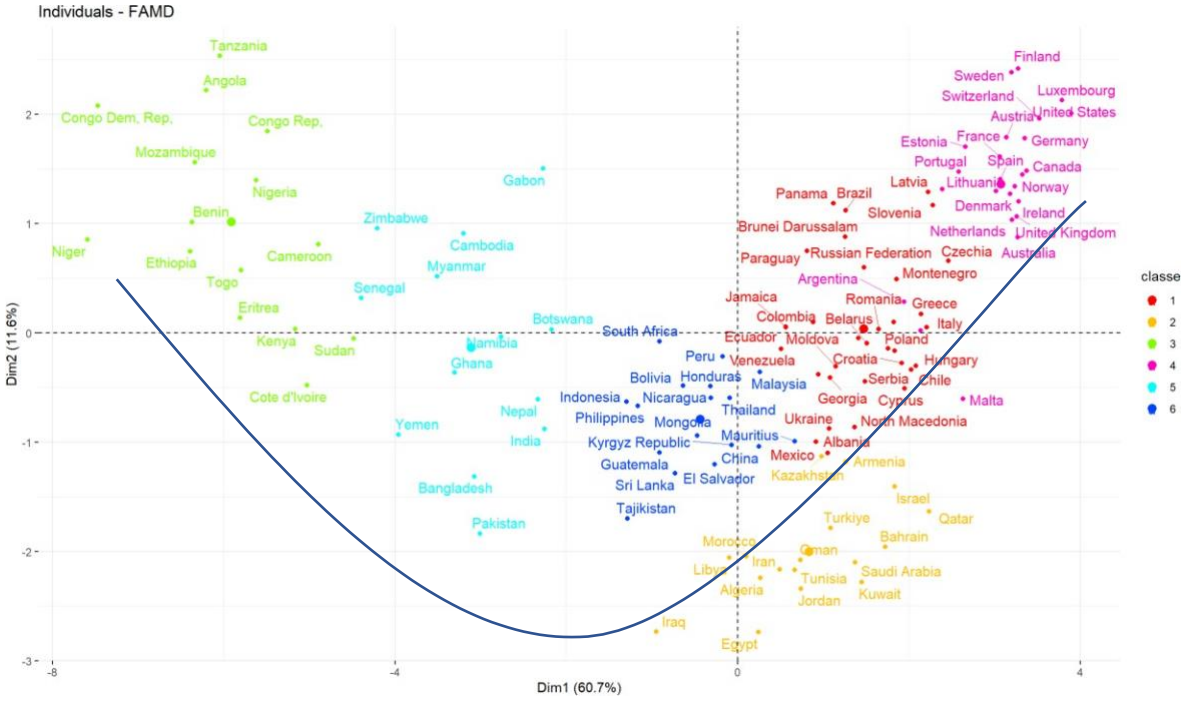
La visualisation de l'éboulis des valeurs propres nous montre que considérer 6 axes peut être excessif au regard du coude bien marqué entre le 1er et deuxième axe quand bien même il minimiserait notre erreur de classification. Le premier plan constitué des deux premières dimensions porte 72% d'inertie.

La troisième méthode « à la Kaiser » nous suggère de retenir les 3 premières dimensions car supérieures à la moyenne des valeurs propres dont la valeur est 1.

Après l'étude des résultats des 3 méthodes, nous faisons le choix de maintenir les deux premières dimensions et de baser notre étude sur ce plan car étudier un second plan ne serait pas bénéfique au regard des pourcentage de variances.

ii. Interprétation à l'aide des classes

Nous établissons notre AFDM en nous basant sur les deux premières dimensions tout en coloriant les individus en fonction des classes auxquelles ils appartiennent. Commençons par observer le graphe des individus en fonction des classes.



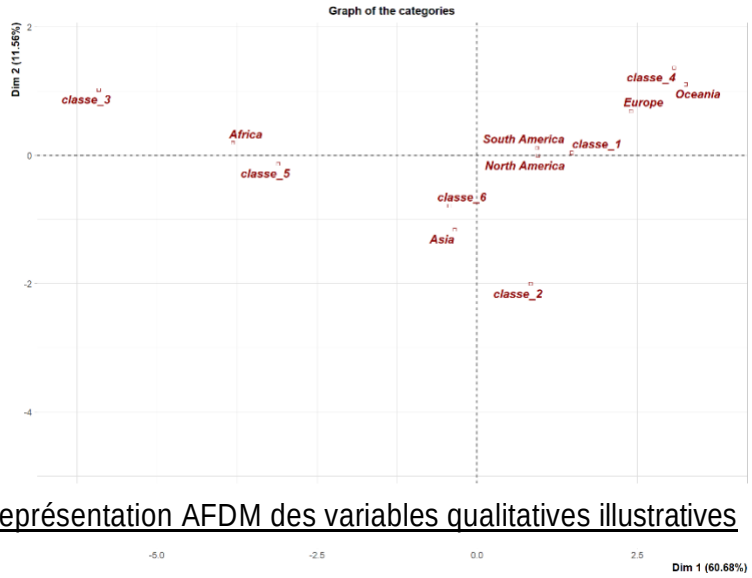
Représentation AFDM des individus en fonction des classes

À première vue, nous constatons que nos individus forment une parabole, ce phénomène est nommé l'effet Gutmann. En effet, le premier axe factoriel (60.7%) est bien plus fort que le deuxième axe (11.6%). Ce dernier semble moins enclin à séparer nos individus. Cela pourrait être dû à la faiblesse en corrélation entre nos variables et cet axe ou à leur faible qualité de projection. Nous tâcherons de surveiller cet aspect et y apporter plus d'explications durant notre étude.

La lecture de l'AFDM nous indique que 3 classes semblent s'opposer : la **classe 3** avec des valeurs négatives extrêmes sur la première dimension et positives sur la deuxième quand bien même éparpillé sur la deuxième dimension, la **classe 4** se caractérise par des individus avec de fortes valeurs aussi bien sur la 1ere que sur la 2de dimension. La **classe 2** extrêmement négative sur le deuxième axe et en majeure partie positive sur le premier.

Nous décidons de tracer les barycentres des 6 classes sur le plan factoriel, cela nous sera fort profitable pour l'interprétation de nos axes et en complément de représenter la variable qualitative illustrative continent. Cela nous permettra d'enrichir un peu plus notre analyse.

La visualisation de ce graphe appelle à plusieurs constatations :
La **classe 5** semble proche de la modalité 'Africa' qui est dans la partie négative de la première dimension.
La **classe 6** parait proche de la modalité 'Asia', le retour à l'AFDM des individus (graphe précédent) confirme que les individus de cette classe demeurent dans la partie négative de l'axe 2.

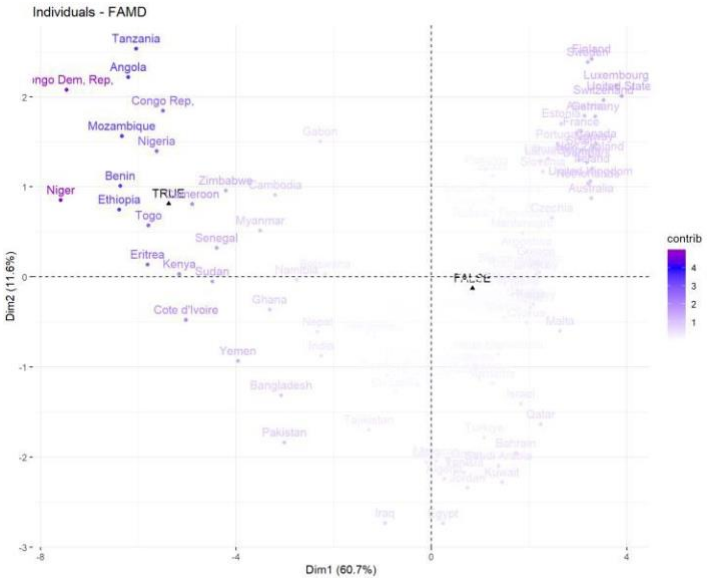


Représentation AFDM des variables qualitatives illustratives

Les modalités 'South America' et 'North America' sont extrêmement proches entre elles tout en étant aux abords de la **classe 1.**
La **classe 4** semble très proche de la modalité 'Oceania' et dans une moindre mesure avec 'Europe'.
Les modalités 'Oceania' et 'Africa' s'opposent entre elles sur le premier axe, alors que 'Asia' s'oppose moindrement sur le deuxième axe avec ce même groupe de modalités.

iii. Interprétation à l'aide des individus

Afin d'interpréter nos axes à l'aide des individus nous souhaitons nous appuyer exclusivement sur les individus les plus contributifs. On commence donc par étudier leur contribution en visualisant les individus en fonction de leur contribution a la construction des deux axes :



Représentation des individus en fonction de leur contribution

L'examen de cette représentation fait apparaitre une forte hétérogénéité dans la répartition de la contribution entre nos individus. En effet, la contribution des pays situés à l'extrémité négative de l'axe 1 qui se répartissent sur la partie positive de l'axe 2 semble très forte. Ces pays correspondent à la **classe 3**, sont africains et semblent proches de la modalité « sous développé ». Nous avions préalablement constaté lors de l'analyse bivarié la situation particulière du continent africain pour ce qui de l'espérance de vie et de la mortalité singulière pour les causes de maladies transmissibles et chroniques, il en va de même pour la **classe 3.** Ces facteurs peuvent avoir contribué à cet effet. Les pays s'opposant à ce groupe, sur le premier axe, semblent contribuer mais dans une bien moindre mesure que le groupe précédemment identifié.

Pour conforter cette interprétation, nous établissons la représentation des 20 individus les plus contributeurs :

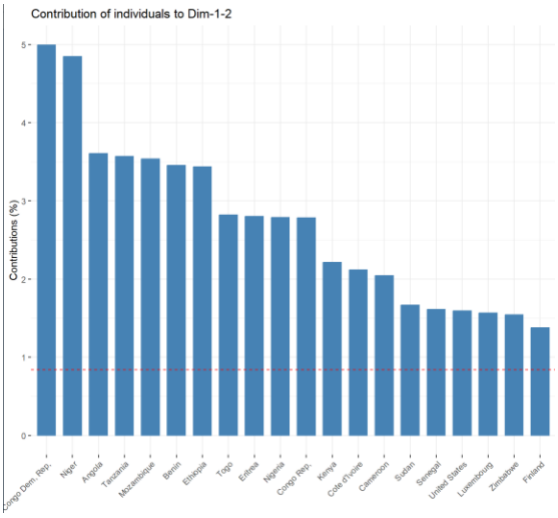


Diagramme des 20 individus les plus contributifs

Ce diagramme nous conforte dans notre analyse, nous constatons que :

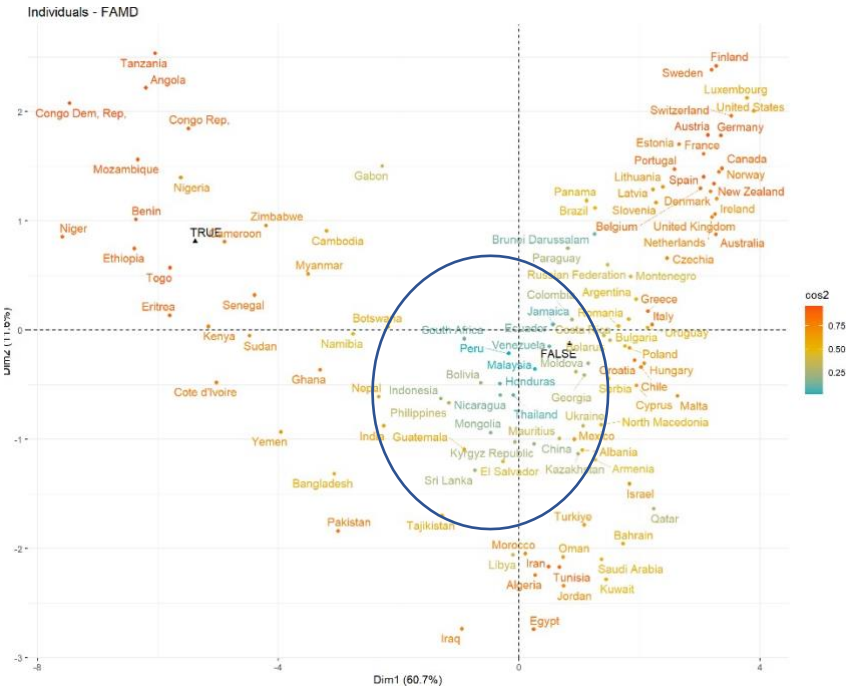
- Deux pays[11] qui représentent 1,68 % de notre population totalisent environ 10%. Ces individus sont donc extrêmes.
- Les 11 individus les plus contributifs (10% de la population) sont tous des pays africains de la **classe 3** qui se situent dans la partie négative du premier axe. Ils totalisent 38% de la contribution totale.
- Ce n'est qu'après avoir visualisé les 20 individus les plus contributifs que 3 pays européens apparaissent.

Nous pouvons donc dire que la première dimension est très marquée par la contribution des pays situés à son extrémité négative par les valeurs extrêmes des pays africains de la **classe 3** et dans une moindre mesure par les valeurs extrêmes des pays de la **classe 4** à son extrémité positive. Le deuxième axe est plus faible car les deux groupes identifiés ne contribuent que sur sa partie positive.

[11] République démocratique du Congo et le Niger

Pour juger la proximité des individus, on choisit de se baser sur la qualité de leur représentation et donc sur leur Cos2.

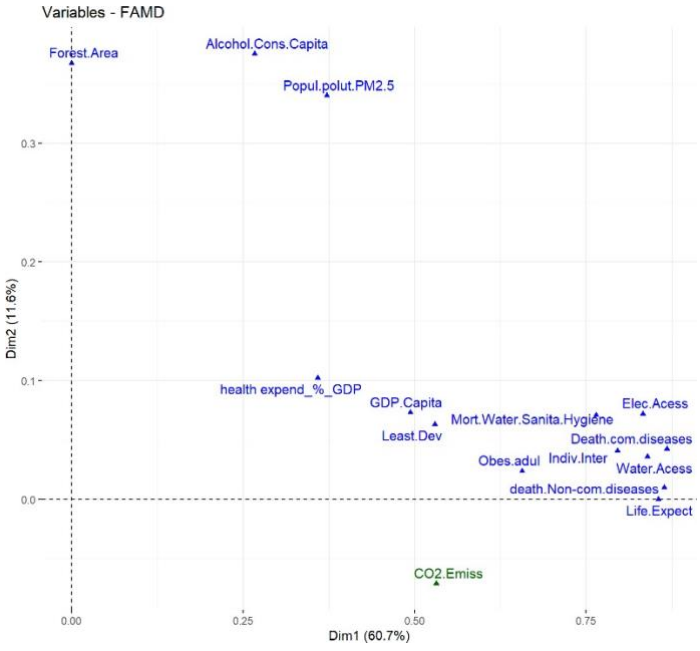
Nous remarquons que les individus les moins bien représentés sont situés au centre de notre AFDM et sous forme d'un cercle. Nous tacherons donc de ne pas interpréter la proximité entre les individus de cette partie du graphique.



iv. Interprétation à l'aide des variables

Nous réalisons qu'il y a beaucoup plus de variables liées à la première dimension comparé à la deuxième. Ceci constitue un premier élément mais n'est toujours pas suffisant pour nous expliquer la faiblesse de notre deuxième dimension.

En ce qui concerne le premier axe plusieurs variables semblent contribuer à sa construction. Les cinq variables les plus contributives[12] partagent le même niveau de contribution soit à environ 10%, elles sont suivies de près par trois variables dont le niveau de contribution reste supérieur à la moyenne de contributions si toutes les variables avaient la même contribution.



Graph des variables sur deux axes en fonction de leur contribution

La situation est bien différente pour notre deuxième axe. La contribution des variables à la construction de cet axe est concentrée sur trois variables[13] qui concentrent plus que 60% de la contribution.

[12] Death.com.diseases, death.non-com.diseases, Life.Expect, Water.Acess et Electric.Acess

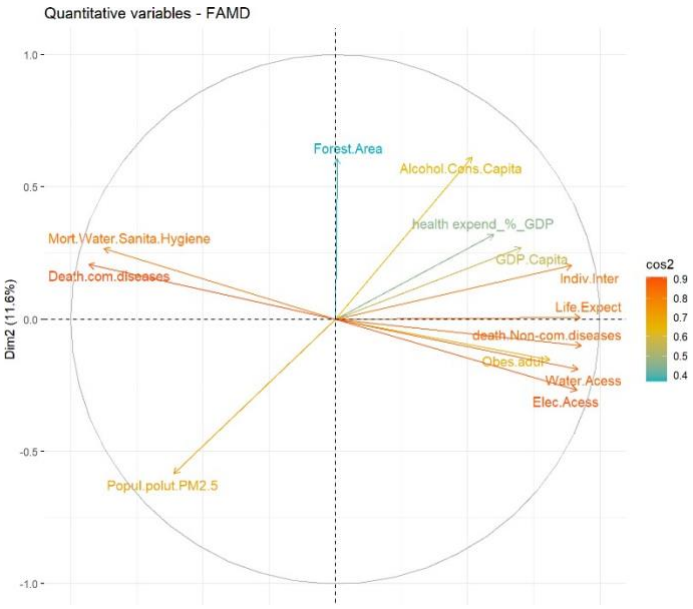
[13] Alchool.Cons.capital, Forest.Area, Popul.polut.PM.2.5

Notons également que la variable sur la surface forestière contribue exclusivement à l'axe 2 de la même façon que l'espérance de vie contribue exclusivement à l'axe 1. Ceci est cohérence avec les résultats de l'analyse bivarié entre ces deux variables.

Les variables 'Continent' et 'Co2.Emiss' sont nos variables illustratives et ne participent donc pas par définition a la construction de nos axes.

Une fois les contributions établies, il nous parait essentiel de chercher à savoir si elles sont bien projetées.

Les couleurs établies dans le cercle des corrélations nous permettent d'évaluer la qualité de projection de façon plus précise grâce à la valeur du Co2 de l'angle de nos variables. Ainsi, nous pouvons expliquer l'effet Gutmann constaté dans la représentation de l'AFDM du a la faiblesse de l'axe 2 par la faible qualité de projection des trois variables les plus contributives (plus de 60%) a ce même axe dont fait partie la variable 'Forest.Area' quand bien même elle s'avère exclusive à cet axe.



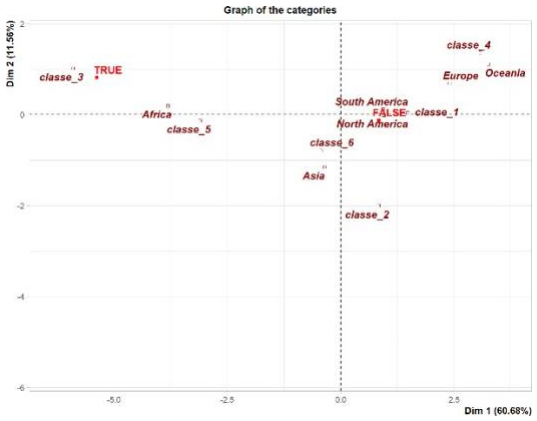
Cercle des corrélations des variables en fonction de l'angle du Cos2

Les variables concernant la part des dépenses santé en PIB, le PIB par habitant et la surface forestière ne pas assez bien projetées pour pouvoir les interprétées et donc ne feront pas intégrées dans la suite de notre analyse.

Les deux variables sur l'exposition aux particules fines et la consommation de l'alcool s'opposent diamétralement et participent à la construction des deux axes, elles sont assez bien projetées.

Cas des variables qualitatives

La modalité 'TRUE' qui correspond à la notion « sous-développé » est proche de la **classe 3**. Elle caractérise les individus se trouvant sur le côté extrême négatif de l'axe 1.
Quant à la modalité 'FALSE' de la même variable elle se situe sur la partie positive de l'axe 1.



v. Description automatique des axes[14]

La **1**ere **dimension** se caractérise par une corrélation positive forte avec la mortalité due aux maladies non transmissibles (chroniques), l'espérance de vie, l'accès à l'eau, à l'électricité et à internet et à contrario une corrélation négative forte avec la mortalité liée aux conditions sanitaires et les maladies transmissibles. Cette dimension est très bien expliquée par la variation des classes.

Pour la **2**iem **dimension** nous constatons des corrélations bien moins importantes qu'avec la première dimension. La deuxième dimension est positivement corrélée avec la consommation de l'alcool et négativement corrélée avec la pollution aux particules fines. Quand bien même l'axe 2 est moins enclin à séparer nos classes que l'axe 1, sa corrélation avec les variables identifiées nous parait relativement significatif.

VI. Synthèse

Sur la base des critères établis pour notre étude, on peut attester que les pays où il ferait « bon vivre » sont les pays de la classe 4[15]. Ces pays jouissent d'un bon accès à l'électricité, à l'eau et à internet. L'espérence de vie y est importante parrallèlement à la consommation de l'acool, chercher à établir le lien entre ces deux variables serez sujet à l'effet cigogne. Les maladies à l'origine des décès de ces habitants sont les maladies chroniques, ce qui est intuitivement cohérent avec le niveau élevé de l'espérence de vie. L'obésité pour les adultes y est relativement considérable. Contre intuitivement, malgré un niveau d'industrialisation (donnée exogène a l'étude) qui serait supérieur l'exposition en particules fines y est limité. Ce résultat peut s'expliquer par la présence de lois encadrants le traitement des émissions Co2 et la proximité des régions industrielle des zones d'habitation. La mortalité due aux maladies transmissibles et pour les causes sanitaires y est bien limitée.

Les pays où il ferait moins « bon vivre » sont les pays de la classe 3. Ils ont un niveau extrêmement élevé en mortalité pour causes sanitaires et celle due aux maladies transmissibles. Inversement, des valeurs extrêmement faibles pour l'accès à l'eau, l'électricité et internet mais aussi en maladies chroniques, en obésité et en espérance de vie. Le PIB par habitant et les émission CO2 sont bien moins importantes que dans les pays où il fait « bon vivre ». Ainsi les causes de décès entre ces deux groupes sont bien différentes et caractérisent même leur différence.

Nous avons pu donc répondre aux questions posées au début de notre étude. Nous pourrions également approfondir cette étude en enrichissant encore plus notre base de données pour prendre en considération d'autres paramètres et essayer d'établir une catégorisation plus précise.

[14] Voir résultats dans l'annexe 10

[15] Finlande, Suède, Luxembourg, USA, Autriche, Suisse, Allemagne, France, Estonie, Espagne, Canada, Portugal, Lituanie, Norvège, Danemark, Irlande, UK, Australie, Pays-Bas, Argentine, Nouvelle Zélande, Uruguay, Belgique et Malte.

VII. Annexes

Annexe 1 : Liste des pays étudiés

Albania Algeria Angola Argentina Armenia Australia Austria Bahrain Bangladesh Belarus Belgium Benin Bolivia Botswana Brazil Brunei Darussalam Bulgaria Cambodia Cameroon Canada Chile China Colombia Congo Dem, Rep, Congo Rep, Costa Rica Cote d'Ivoire Croatia Cyprus Czechia Denmark Ecuador Egypt El Salvador Eritrea Estonia Ethiopia Finland France Gabon Georgia Germany Ghana Greece Guatemala Honduras Hungary India Indonesia Iran Iraq Ireland Israel Italy Jamaica Jordan Kazakhstan Kenya Kuwait Kyrgyz Republic Latvia Libya Lithuania Luxembourg Malaysia Malta Mauritius Mexico Moldova Mongolia Montenegro Morocco Mozambique Myanmar Namibia Nepal Netherlands New Zealand Nicaragua Niger Nigeria North Macedonia Norway Oman Pakistan Panama Paraguay Peru Philippines Poland Portugal Qatar Romania Russian Federation Saudi Arabia Senegal Serbia Slovak Republic Slovenia South Africa Spain Sri Lanka Sudan Sweden Switzerland Tajikistan Tanzania Thailand Togo Trinidad and Tobago Tunisia Turkiye Ukraine United Kingdom United States Uruguay Venezuela Yemen Zimbabwe

Annexe 2 : Code R

```
library(Factoshiny)
library(FactoMineR)
library(dplyr)
library(tidyverse)
library(ellipse)
library(corrplot)
library(stats)
library(Hmisc)
library(BioStatR)
library(missMDA)
library(factoextra)
library(ggplot2)
library(ggcorrplot)
library(ape)
library(ggplot2)
library(ggdendro)

# chargement dataframe
df <- as.data.frame(pure_df19)
is.na.data.frame(df)

#fixer les indivs
rownames(df) <- df$Country
df$Country <- NULL
head(df)

#Pearson corr pr variables quanti
mat_pear <- cor(df[3:16])
corrplot(mat_pear, type = "upper", order = "hclust", tl.col = "black", tl.srt = 80)

## Sperman corr pr variables quanti
mat_sper <- cor(df[3:16],method = "spearman")
corrplot(mat_sper, type = "upper", order = "hclust", tl.col = "black", tl.srt = 80)

## cor de deux variables quanti
cor.test(df$CO2.Emiss, df$Popul.polut.PM2.5, method="spearman")


#chi-deux
chisq.test(df$Least.Dev,  df$Continent)
```

```
tab <- table(df$Least.Dev, df$Continent)
barplot(tab, beside=TRUE, legend=TRUE)
mosaicplot(tab, shade=TRUE)

ggplot(df, aes(x = Least.Dev, fill = Continent)) +
 geom_bar(position = "fill") +
 scale_fill_discrete(name = "y")

####lien entre variables quanti et quali
#transfo des variables charact en factor
df$Continent <- as.factor(df$Continent)
df$Least.Dev <- as.factor(df$Least.Dev)

quali <- which(sapply(df, is.factor))
quanti <- which(sapply(df, is.numeric))

# creation d'une matrice vide avec en ligne les variables quantitatives et en colonne les variables qualitatives
mateta2 <- matrix(NA,length(quali),length(quanti))
rownames(mateta2) <- names(quali)
colnames(mateta2) <- names(quanti)

# calcul des différents eta carré
for(ii in seq(nrow(mateta2))){
for(jj in seq(ncol(mateta2))){
   mateta2[ii, jj]<-eta2(df[, colnames(mateta2)[jj]],
                df[, rownames(mateta2)[ii]])
 }
}
# pot des résultats quali vs quanti
corrplot(mateta2, tl.col = "black", tl.srt = 40)
corrplot(mateta2, tl.col = "black", method = 'number')

## FAMD facto
#Factoshiny(df)

# code FAMD
res.famd<-FAMD(df,sup.var=c(1,4),graph=FALSE) #co2 emiss illustrative + continent var illustrative
plot.FAMD(res.famd,invisible=c('ind.sup'),title="Graphe des individus et des modalités")
plot.FAMD(res.famd,axes=c(1,2),choix='var',title="Graphe des variables")
plot.FAMD(res.famd, choix='quanti',title="Cercle des corrélations")


###  Classification - Construction de la partition :
# construc FAMD
c <- FAMD(df, ncp = Inf,
          graph = FALSE,
          sup.var = c(1,4))  #variable illus!!!
## chop valeurs propres
round(res.famd$eig, 3)

## éboulis des valeurs propres
barplot(res.famd$eig[,1], las = 2, cex.names = .5)

## éboulis des valeurs propres V2
fviz_eig(res.famd, choice= "variance", addlabels = TRUE)
fviz_eig(res.famd, choice= "eigenvalue", addlabels = TRUE)

## nbr pour cah / classif
ncp <- 3
D <- dist(res.famd$ind$coord[,1:ncp])#distance euclidienne entre observations
res.hclust <- hclust(D,method = "ward.D2")#CAH par méthode de Ward

## hauteurs de fusion en fonction des différentes étapes d'agrégation des classes
barplot(sort(res.hclust$height,decreasing = TRUE)[1:15],
     names.arg = 1:15,
     xlab = "index",
     ylab = "hauteur de fusion")

barplot(sort(res.hclust$height,decreasing = TRUE)[1:15],
     names.arg = 1:15,
```

```
      xlab = "index",
      ylab = "hauteur de fusion",
      col =rgb(0.2,0.4,0.6,0.85),
      border= "white",
      width=0.5)

##CAH
res.famd2<-FAMD(df,sup.var=c(1,4),graph=FALSE, ncp=3) #nouveau famd pour incl 3 axes
res.hcpc <- HCPC(res.famd2, nb.clust = 6)
plot(res.hcpc, choice = "3D.map")

#Factor map
fviz_cluster(res.hcpc,           #nbr de cluts intégré
          repel = TRUE,       # Avoid label overlapping
          show.clust.cent = TRUE, # Show cluster centers
          palette = "jco",       # Color palette see ?ggpubr::ggpar
          ggtheme = theme_minimal(),
          main = "Factor map"
)

# Dendrogram v1
fviz_dend(res.hcpc, show_labels = TRUE, repel= TRUE, ggtheme = theme_minimal())

# Dendrogram v2 +  kmeans
colors = c("red", "blue", "green", "black", "brown", "purple")
clus4 = cutree(res.hclust, k = nbclasse)
plot(as.phylo(res.hclust), type = "fan", tip.color = colors[clus4],
    label.offset = 1, cex = 0.7)


nbclasse <- 6
partition <-  cutree(res.hclust, k = nbclasse) #élagage de l'arbre

#Conso
centres.gravite <- by(res.famd$ind$coord[,1:ncp],
                 INDICES = partition,
                 FUN = colMeans)

#donne un objet de type "matrix", nécessaire pour pouvoir utiliser ces centres comme des valeurs initiales pour la fonction kmeans
centres.gravite <- do.call(rbind, centres.gravite)

#kmeans
res.kmeans <- kmeans(res.famd$ind$coord[,1:ncp],
                centers = centres.gravite)

part.finale <- as.factor(res.kmeans$cluster)

table(part.finale)
plot(part.finale)

## description des classes a partir des variables
df_part <- cbind.data.frame(df, classe = part.finale)#on concatène le jeu de données avec la nouvelle variable classe
catdes(df_part, num = ncol(df_part))


 #intégrer class en illustratif
 res.famd <- FAMD(df_part,
            ncp = Inf,
              graph = FALSE,
              sup.var =  c(ncol(df_part),1,4)) #variable illus!!
##plot le tt
p <- fviz_famd_ind(res.famd, habillage=df_part$classe,
               addEllipses=TRUE, ellipse.level=0.95)


#SAVE
write.infile(res.famd,  file="resultats_afdm.csv")

#eboulis
barplot(res.famd$eig[,1], las = 2, cex.names = .5)
```

```
round(res.famd$eig, 3)
# Nombre d'axes par validation croisée
res.ncp <- estim_ncpFAMD(df_part[,-c(1,4,ncol(df_part))],
        # on retire les variables illustratves qui ne sont pas gérées par la fonction
                ncp.max = 10,
                method.cv = "Kfold",
                nbsim = 100 #augmenter ce nombre améliore la précision des résultats, mais aussi le temps de calcul

)
## plot tt ca
plot(x = as.numeric(names(res.ncp$crit)),
    y = res.ncp$crit,
    xlab = "S",
    ylab = "Erreur",
    main = "Erreur de validation croisée\n en fonction du nombre d'axes",
    type = "b")

## Interprétation à l'aide des classes
fviz_mfa_ind(res.famd,
          habillage = "classe", # couleurs selon les modalités de la variable classe
          palette = c("#FF0000", "#FFBF00", "#80FF00", "#FF00BF", "#00FFFF", "#0040FF", "#8000FF"),# définition des couleurs
          repel = TRUE
)

# tracer les barycentres des XX classes sur le plan factoriel, très utile pour l'interprétation des axes.
plot(res.famd,
    choix = "quali",
    invisible = c("quali","ind")
)

#Interprétation à l'aide des individus top 11-20-17
###### méthode 1
plot(res.famd, choix = "ind", invisible = "quali", select = "contrib 2")
res.famd$ind$contrib

fviz_contrib(res.famd, choice = "ind", axes = c(1,2), top= 10)
fviz_contrib(res.famd, choice = "ind", axes = c(1,2),  sort.val= "asc")

###### méthode 2
fviz_famd_ind(res.famd, col.ind="contrib", ) +
 scale_color_gradient2(low="white", mid="blue",
                high="red", midpoint=4, space ="Lab")



###### méthode 3
fviz_famd_ind(res.famd, select.ind = list(contrib = 17))

fviz_famd_ind(res.famd, alpha.ind="contrib") +
 theme_minimal()

#Interprétation à l'aide des variables
##### métho 1
plot(res.famd,choix = "var", select = "contrib 30")

##### métho 2
# Couleur par valeurs cos2: qualité sur le plan des facteurs
fviz_famd_var(res.famd, "quanti.var", col.var = "cos2",
          gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
          repel = TRUE)

# Graphique des variables
fviz_famd_var(res.famd, repel = TRUE, col.var="blue")

# Contribution à la première dimension
fviz_contrib(res.famd, "var", axes = 1)
# Contribution à la deuxième dimension
fviz_contrib(res.famd, "var", axes = 2)

#co2 des vars
fviz_cos2(res.famd, "quanti.var", axes= c(1,2))
```

```
res.famd$qua
#Pour les variables quantitatives uniquement
plot(res.famd, choix = "quanti")

#Quant aux modalités des variables qualitatives
plot(res.famd, choix = "quali")
fviz_cos2(res.famd, "quali.var", axes= c(1,2))

##########représenter qu'une partie des modalités privilégier FActoshiny

#Description automatique des axes
res.dimdesc <- dimdesc(res.famd)
lapply(res.dimdesc$Dim.1, round, 3)# pour arrondir à 3 décimales les résultas portant sur la première dimension

lapply(res.dimdesc$Dim.2, round, 3)# pour la seconde dimension

lapply(res.dimdesc$Dim.3, round, 3)# pour la trois dimension

plot(res.famd,choix = "ind", select = "cos2 0.5", autoLab = "yes", habillage = "classe")
#Factoshiny(df_part)
```

Annexe 3 : Résultats pour la description des classes sous R en fonction des variables qualitatives et quantitatives

```
Link between the cluster variable and the categorical variables (chi-square test)
================================================================================
               p.value df
Continent 1.287321e-21 25
Least.Dev 3.636241e-13  5

Description of each cluster by the categories
=============================================
$`1`
                          Cla/Mod    Mod/Cla    Global      p.value    v.test
Continent=Europe        52.631579  60.606061 31.932773 6.520486e-05  3.993137
Least.Dev=FALSE         32.038835 100.000000 86.554622 3.592908e-03  2.911854
Continent=South America 60.000000  18.181818  8.403361 3.134236e-02  2.152699
Least.Dev=TRUE           0.000000   0.000000 13.445378 3.592908e-03 -2.911854
Continent=Asia           6.666667   6.060606 25.210084 1.772744e-03 -3.125879
Continent=Africa         0.000000   0.000000 23.529412 2.499737e-05 -4.214823

$`2`
                Cla/Mod  Mod/Cla   Global      p.value    v.test
Continent=Asia       40 70.58824 25.21008 2.582156e-05  4.207496
Continent=Europe      0  0.00000 31.93277 7.879678e-04 -3.356987

$`3`
                   Cla/Mod   Mod/Cla   Global      p.value    v.test
Continent=Africa 53.571429 100.00000 23.52941 9.045731e-12  6.820923
Least.Dev=TRUE   68.750000  73.33333 13.44538 4.822608e-09  5.853180
Continent=Asia    0.000000   0.00000 25.21008 9.219895e-03 -2.603791
Continent=Europe  0.000000   0.00000 31.93277 1.967532e-03 -3.095090
Least.Dev=FALSE   3.883495  26.66667 86.55462 4.822608e-09 -5.853180

$`4`
                    Cla/Mod    Mod/Cla    Global      p.value    v.test
Continent=Europe   47.36842  75.000000 31.932773 1.474564e-06  4.814652
Least.Dev=FALSE    23.30097 100.000000 86.554622 2.051217e-02  2.316845
Continent=Oceania 100.00000   8.333333  1.680672 3.931064e-02  2.060920
Least.Dev=TRUE      0.00000   0.000000 13.445378 2.051217e-02 -2.316845
Continent=Africa    0.00000   0.000000 23.529412 6.869340e-04 -3.394742
Continent=Asia      0.00000   0.000000 25.210084 3.708940e-04 -3.559972
$`5`
                   Cla/Mod  Mod/Cla   Global     p.value    v.test
Least.Dev=TRUE   31.25000 38.46154 13.44538 0.017810252  2.369540
Continent=Asia   23.33333 53.84615 25.21008 0.022356872  2.284249
Least.Dev=FALSE   7.76699 61.53846 86.55462 0.017810252 -2.369540
Continent=Europe  0.00000  0.00000 31.93277 0.004806547 -2.819721

$`6`
                Cla/Mod  Mod/Cla   Global      p.value    v.test
Continent=Asia       30 52.94118 25.21008 0.0091758451  2.605432
Continent=Europe      0  0.00000 31.93277 0.0007879678 -3.356987
```

```
Link between the cluster variable and the quantitative variables
================================================================
                                 Eta2      P-value
Death.com.diseases           0.8991989 1.401241e-54
Mort.Water.Sanita.Hygiene    0.8928560 4.359009e-53
Elec.Acess                   0.8837221 4.367579e-51
death.Non-com.diseases       0.8520408 3.383433e-45
Water.Acess                  0.8381993 5.167460e-43
Life.Expect                  0.8202279 1.922508e-40
Indiv.Inter                  0.7810109 1.243682e-35
Obes.adul                    0.7806207 1.374298e-35
Popul.polut.PM2.5            0.7599303 2.150206e-33
Alcohol.Cons.Capita          0.5895015 2.159624e-20
GDP.Capita                   0.5582201 1.264613e-18
health.expend_%_GDP          0.5004548 1.118053e-15
CO2.Emiss                    0.3457213 2.745498e-09
Forest.Area                  0.3043414 7.323691e-08
```

```
Description of each cluster by quantitative variables
=====================================================
$`1`
                             v.test Mean in category Overall mean sd in category Overall sd
death.Non-com.diseases     4.299321         86.1278877    73.521091      7.8350500  19.731206
Alcohol.Cons.Capita        3.736947          8.9563636     6.592168      3.1241055   4.257113
Elec.Acess                 3.612839         99.2955951    87.384624      1.7046530  22.184380
Forest.Area                3.541206         41.0104810    30.232829     14.6294566  20.479611
Water.Acess                3.387911         96.9013697    89.276094      3.0887073  15.145104
Obes.adul                  2.991446         23.4969697    19.575630      3.1765700   8.820678
Indiv.Inter                2.859001         62.9494632    51.335877      9.7142886  27.333844
Life.Expect                2.605310         76.0031013    73.306871      2.8816985   6.963801
Mort.Water.Sanita.Hygiene -3.328304          0.4242424     8.497479      0.4748399  16.322004
Death.com.diseases        -4.177932          6.5134807    17.793523      4.3499143  18.167624
                               p.value
death.Non-com.diseases    1.713222e-05
Alcohol.Cons.Capita       1.862682e-04
Elec.Acess                3.028626e-04
Forest.Area               3.983025e-04
Water.Acess               7.042701e-04
Obes.adul                 2.776594e-03
Indiv.Inter               4.249769e-03
Life.Expect               9.179128e-03
Mort.Water.Sanita.Hygiene 8.737655e-04
Death.com.diseases        2.941717e-05
```

```
$`2`
                             v.test Mean in category Overall mean sd in category Overall sd
Obes.adul                  4.170399         27.870588    19.575630      4.6163093   8.820678
CO2.Emiss                  4.088862         10.057462     5.052212      9.0020422   5.428610
Popul.polut.PM2.5          2.792911         95.692436    71.400318     16.6759978  38.572136
Elec.Acess                 2.149307         98.136402    87.384624      6.2114302  22.184380
Water.Acess                1.991829         96.078444    89.276094      4.5196968  15.145104
Mort.Water.Sanita.Hygiene -2.091416          0.800000     8.497479      0.8554325  16.322004
Death.com.diseases        -2.129036          9.071528    17.793523      3.0118137  18.167624
Alcohol.Cons.Capita       -5.243816          1.558353     6.592168      1.6447343   4.257113
Forest.Area               -5.554596          4.581564    30.232829      7.1060241  20.479611
                               p.value
Obes.adul                 3.040670e-05
CO2.Emiss                 4.334954e-05
Popul.polut.PM2.5         5.223608e-03
Elec.Acess                3.161010e-02
Water.Acess               4.638981e-02
Mort.Water.Sanita.Hygiene 3.649079e-02
Death.com.diseases        3.325125e-02
Alcohol.Cons.Capita       1.572892e-07
Forest.Area               2.782547e-08
```

$`3`

	v.test	Mean in category	Overall mean	sd in category	Overall sd
Mort.Water.Sanita.Hygiene	9.718151	46.9466667	8.497479	13.724813	16.322004
Death.com.diseases	8.554546	55.4660858	17.793523	7.194040	18.167624
Popul.polut.PM2.5	3.058845	100.0000000	71.400318	0.000000	38.572136
Alcohol.Cons.Capita	-2.283936	4.2353333	6.592168	3.454805	4.257113
CO2.Emiss	-3.527769	0.4100663	5.052212	0.314997	5.428610
health.expend_%_GDP	-3.853695	4.2299576	6.576373	1.412629	2.511873
GDP.Capita	-3.921827	3204.2486863	21751.438809	1928.018813	19510.104968
Indiv.Inter	-5.946067	11.9391114	51.335877	7.115271	27.333844
Obes.adul	-6.127953	6.4733333	19.575630	1.661913	8.820678
Life.Expect	-7.567587	60.5326667	73.306871	3.474165	6.963801
death.Non-com.diseases	-8.053636	35.0020354	73.521091	6.399399	19.731206
Water.Acess	-8.980986	56.3055345	89.276094	9.565249	15.145104
Elec.Acess	-9.049502	38.7212706	87.384624	13.933930	22.184380

	p.value
Mort.Water.Sanita.Hygiene	2.523208e-22
Death.com.diseases	1.183331e-17
Popul.polut.PM2.5	2.221920e-03
Alcohol.Cons.Capita	2.237531e-02
CO2.Emiss	4.190784e-04
health.expend_%_GDP	1.163486e-04
GDP.Capita	8.787992e-05
Indiv.Inter	2.746611e-09
Obes.adul	8.901690e-10
Life.Expect	3.802193e-14
death.Non-com.diseases	8.037027e-16
Water.Acess	2.683504e-19
Elec.Acess	1.436216e-19

$`4`

	v.test	Mean in category	Overall mean	sd in category	Overall sd
GDP.Capita	6.822867	46131.886070	21751.438809	1.783414e+04	19510.104968
health.expend_%_GDP	6.774846	9.693197	6.576373	2.122547e+00	2.511873
Indiv.Inter	6.440024	83.576497	51.335877	9.053965e+00	27.333844
Life.Expect	5.789936	80.691613	73.306871	2.189602e+00	6.963801
Alcohol.Cons.Capita	5.482733	10.867083	6.592168	1.958006e+00	4.257113
death.Non-com.diseases	4.234089	88.822408	73.521091	2.970382e+00	19.731206
Obes.adul	3.961139	25.975000	19.575630	3.918891e+00	8.820678
Water.Acess	3.704235	99.551193	89.276094	9.226344e-01	15.145104
Elec.Acess	3.098009	99.972286	87.384624	9.271411e-02	22.184380
CO2.Emiss	2.837562	7.873512	5.052212	4.126732e+00	5.428610
Mort.Water.Sanita.Hygiene	-2.771423	0.212500	8.497479	1.301041e-01	16.322004
Death.com.diseases	-3.564895	5.931464	17.793523	2.692053e+00	18.167624
Popul.polut.PM2.5	-8.975128	7.994438	71.400318	1.374198e+01	38.572136

	p.value
GDP.Capita	8.924148e-12
health.expend_%_GDP	1.245387e-11
Indiv.Inter	1.194546e-10
Life.Expect	7.041330e-09
Alcohol.Cons.Capita	4.188040e-08
death.Non-com.diseases	2.294797e-05
Obes.adul	7.459319e-05
Water.Acess	2.120293e-04
Elec.Acess	1.948258e-03
CO2.Emiss	4.545958e-03
Mort.Water.Sanita.Hygiene	5.581193e-03
Death.com.diseases	3.640015e-04
Popul.polut.PM2.5	2.830219e-19

$`5`

	v.test	Mean in category	Overall mean	sd in category	Overall sd
Death.com.diseases	4.163966	37.679439	17.793523	9.356880e+00	18.167624
Popul.polut.PM2.5	2.820235	99.995916	71.400318	1.258245e-02	38.572136
health.expend_%_GDP	-2.303090	5.055655	6.576373	2.002187e+00	2.511873
Alcohol.Cons.Capita	-2.511174	3.782000	6.592168	2.902677e+00	4.257113
Water.Acess	-2.536380	79.178293	89.276094	1.136932e+01	15.145104
CO2.Emiss	-2.704748	1.192493	5.052212	9.681057e-01	5.428610
GDP.Capita	-3.056893	6073.817976	21751.438809	4.157522e+03	19510.104968
Elec.Acess	-3.391125	67.608946	87.384624	1.436581e+01	22.184380
Life.Expect	-3.901276	66.165308	73.306871	3.129867e+00	6.963801
Indiv.Inter	-4.103581	21.850683	51.335877	9.720596e+00	27.333844
death.Non-com.diseases	-4.127238	52.114213	73.521091	9.864729e+00	19.731206
Obes.adul	-4.743505	8.576923	19.575630	4.519727e+00	8.820678

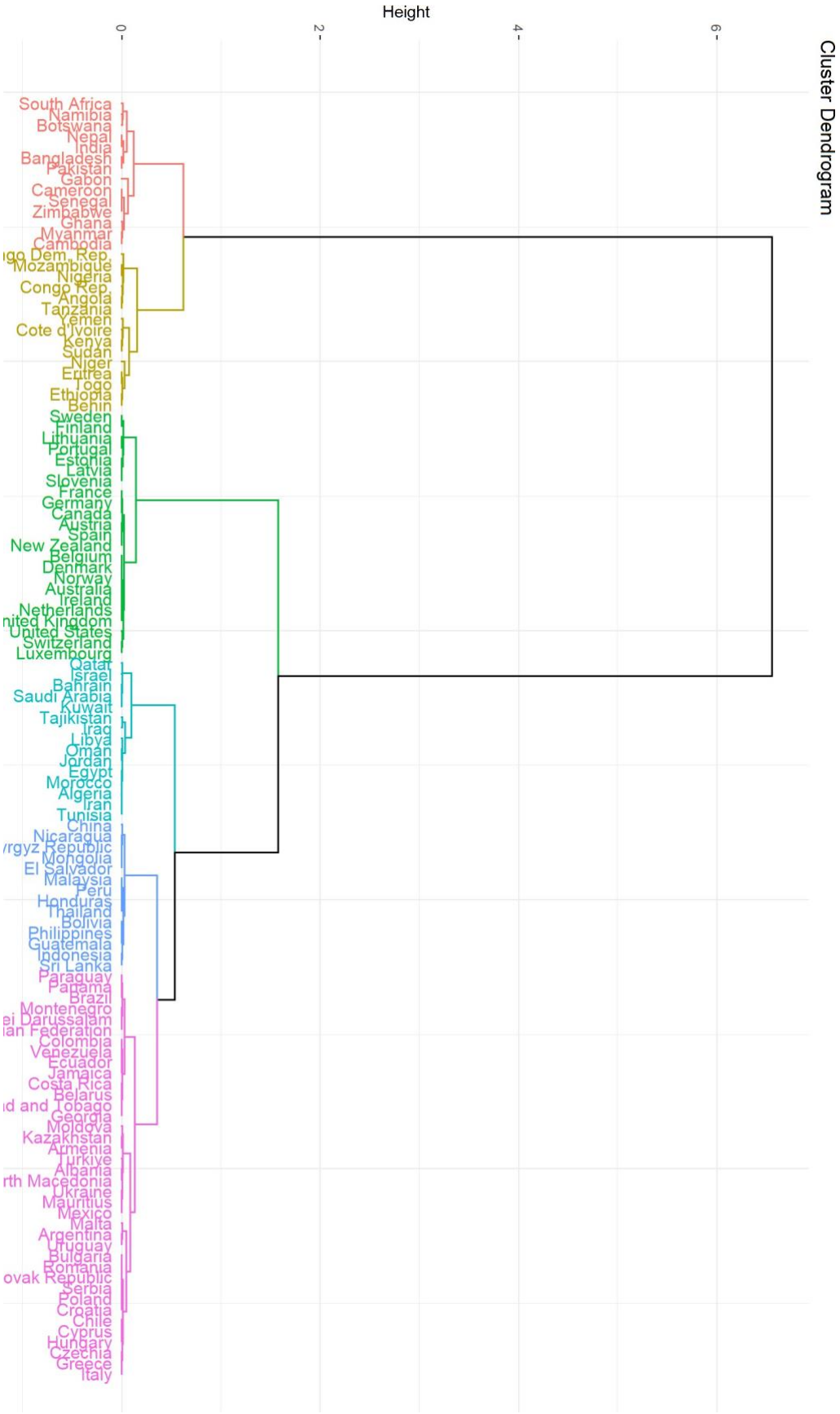
	p.value
Death.com.diseases	3.127664e-05
Popul.polut.PM2.5	4.798854e-03
health.expend_%_GDP	2.127378e-02
Alcohol.Cons.Capita	1.203303e-02
Water.Acess	1.120051e-02
CO2.Emiss	6.835621e-03
GDP.Capita	2.236438e-03
Elec.Acess	6.960631e-04
Life.Expect	9.568702e-05
Indiv.Inter	4.068035e-05
death.Non-com.diseases	3.671471e-05
Obes.adul	2.100512e-06

$`6`

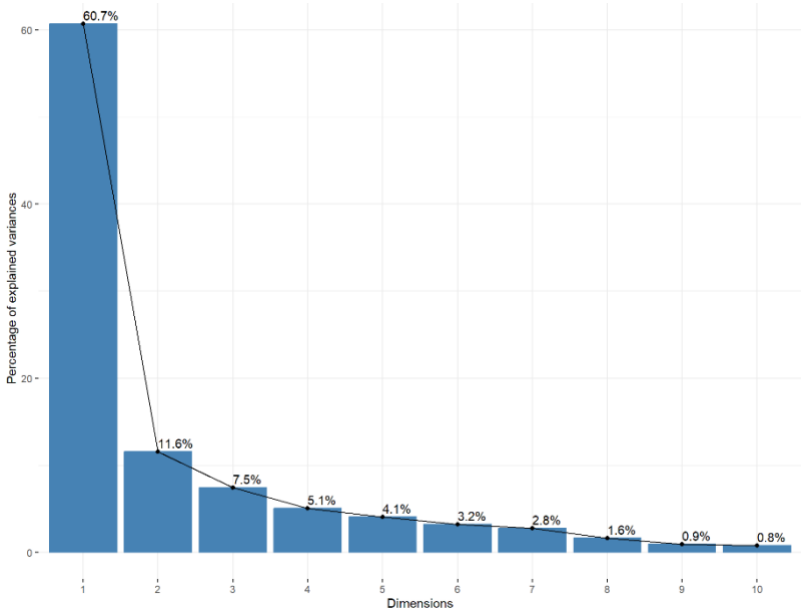
	v.test	Mean in category	Overall mean	sd in category	Overall sd
Popul.polut.PM2.5	1.995580	88.75743	71.40032	14.960587	38.572136
Obes.adul	-2.498606	14.60588	19.57563	6.232408	8.820678
GDP.Capita	-2.563184	10474.94856	21751.43881	5511.058672	19510.104968
Indiv.Inter	-2.733558	34.48729	51.33588	14.627897	27.333844

	p.value
Popul.polut.PM2.5	0.045979629
Obes.adul	0.012468289
GDP.Capita	0.010371718
Indiv.Inter	0.006265411

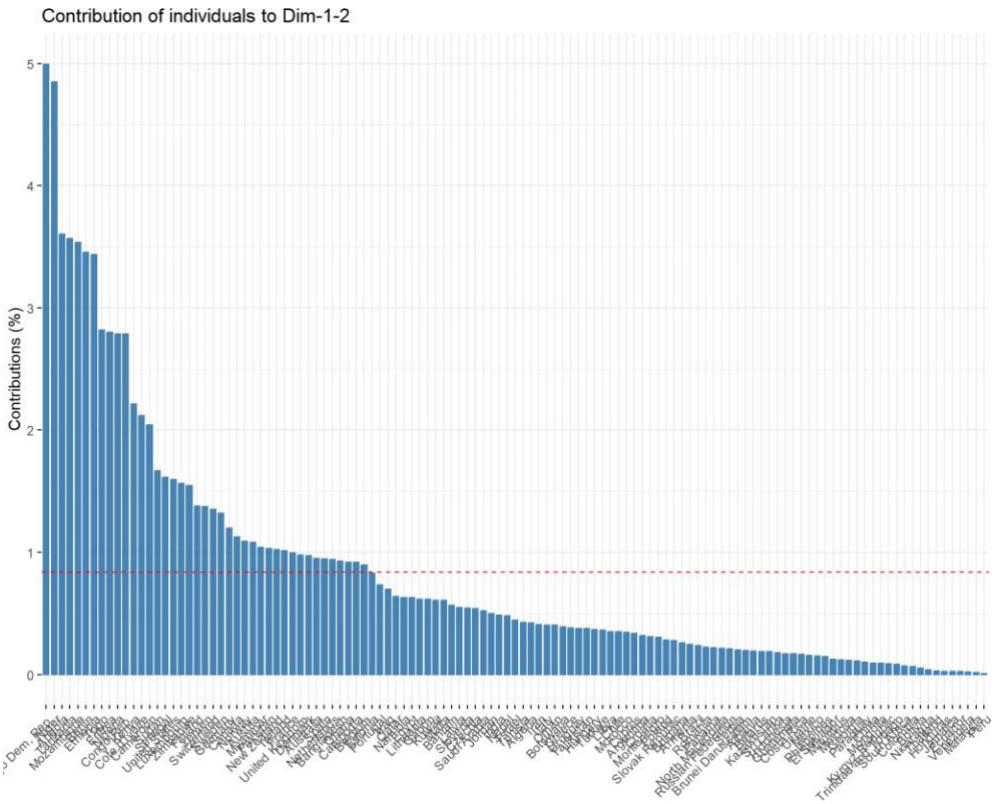
Annexe 4 : Dendrogramme détaillé en fonction des individus



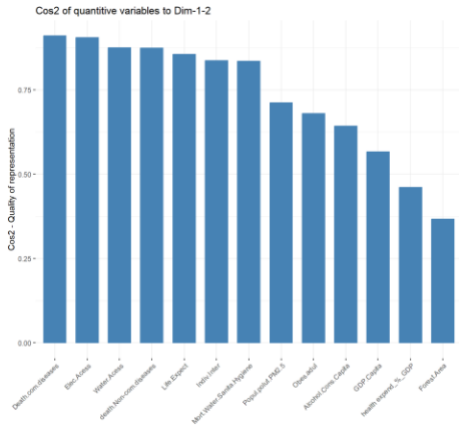
Annexe 5: Eboulis de la variance



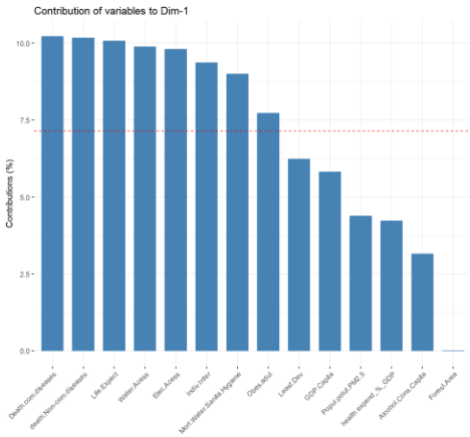
Annexe 6 : Diagramme de tous individus en fonction de leur contribution



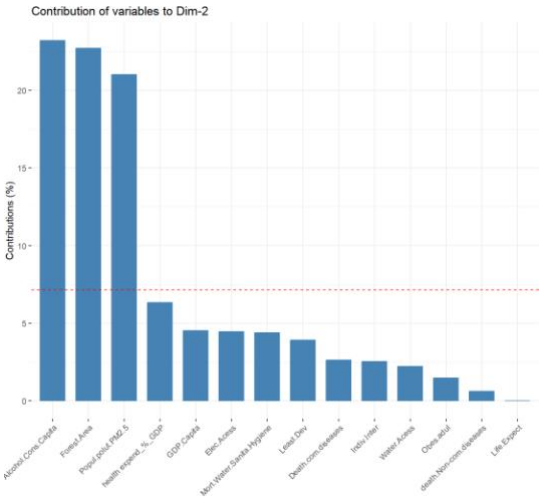
Annexe 7 : Histogramme contribution des variables pour les deux premières dimensions



Annexe 8 : Histogramme de la contribution des variables pour la première dimension



Annexe 9 : Histogramme de la contribution des variables à la deuxième dimension



Annexe 10 : Résultats de la description automatique des axes

Axe 1

```
$quanti
                          correlation p.value
death.Non-com.diseases          0.929       0
Life.Expect                     0.925       0
Water.Acess                     0.916       0
Elec.Acess                      0.913       0
Indiv.Inter                     0.892       0
Obes.adul                       0.810       0
GDP.Capita                      0.702       0
health expend_%_GDP             0.599       0
CO2.Emiss                       0.532       0
Alcohol.Cons.Capita             0.517       0
Popul.polut.PM2.5              -0.610       0
Mort.Water.Sanita.Hygiene      -0.874       0
Death.com.diseases             -0.932       0

$quali
             R2 p.value
classe    0.953       0
Continent 0.663       0
Least.Dev 0.529       0

$category
                  Estimate p.value
Least.Dev=FALSE      3.108   0.000
Continent=Europe     1.840   0.000
classe=4             3.755   0.000
classe=1             2.153   0.001
classe=5            -2.430   0.000
Continent=Africa    -4.376   0.000
Least.Dev=TRUE      -3.108   0.000
classe=3            -5.231   0.000
```

Axe 2

```
$quanti
                          correlation p.value
Alcohol.Cons.Capita             0.613   0.000
Forest.Area                     0.606   0.000
health expend_%_GDP             0.320   0.000
GDP.Capita                      0.270   0.003
Mort.Water.Sanita.Hygiene       0.266   0.003
Death.com.diseases              0.206   0.025
Indiv.Inter                     0.203   0.027
Water.Acess                    -0.189   0.039
Elec.Acess                     -0.269   0.003
Popul.polut.PM2.5              -0.583   0.000

$quali
              R2 p.value
classe    0.722   0.000
Continent 0.321   0.000
Least.Dev 0.063   0.006

$category
                 Estimate p.value
classe=4            1.446   0.000
Continent=Europe    0.532   0.000
classe=3            1.102   0.001
Least.Dev=TRUE      0.469   0.006
Least.Dev=FALSE    -0.469   0.006
classe=6           -0.704   0.005
Continent=Asia     -1.313   0.000
classe=2           -1.919   0.000
```