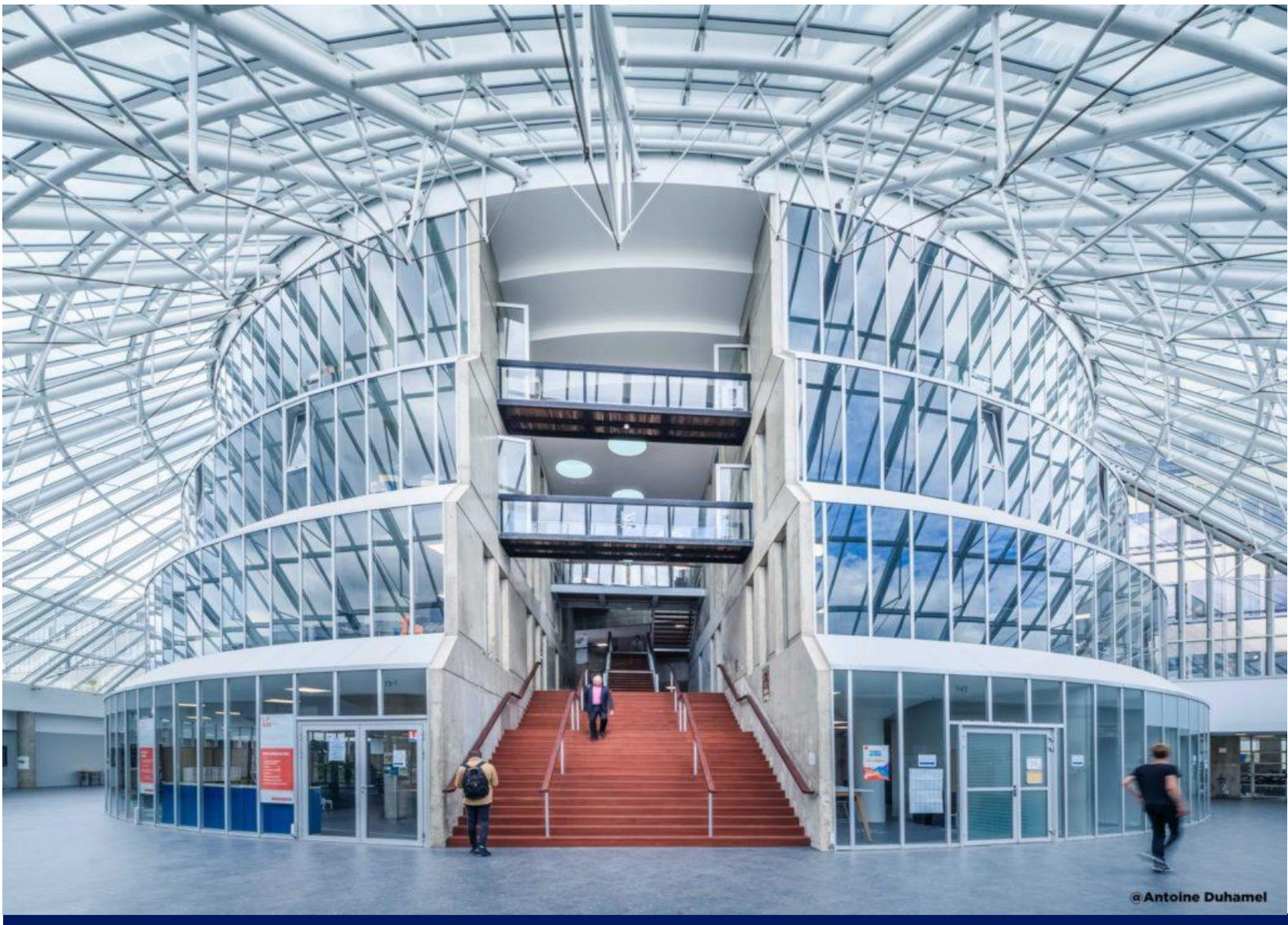


Analyses Multidimensionnelles

Ulysse Blondy & Agossouvo Bernice

LP Dataming

Enseignant : M. Hervé CLEMENT

TABLE DES MATIERES

I. INTRODUCTION

La visualisation des données est un élément clé du traitement de l'information. En effet, il est tout à fait possible de visualiser un jeu de données contenant deux (2) ou trois (3) variables quantitatives voire étudier les relations entre elles. Maintenant, supposons que notre jeu de données comporte plusieurs variables avec N dimensions, il devient beaucoup plus difficile d'analyser et visualiser le nuage de point.

La méthode factorielle de représentation permet de résoudre ce problème de visualisation des données. Cette méthode est basée sur la notion d'Inertie de nuage de point (somme distance au carré de chaque point à leur centre de gravité G). En d'autres termes, on peut dire cette méthode repose sur la notion de variance, c'est-à-dire plus l'inertie du nuage point sera grande plus les points seront dispersés (on parle aussi de variabilité).

Dans ce mini projet, trois (3) méthodes factorielles de représentation seront abordées. Dans un premier, l'analyse en composantes principales (ACP) (dont les variables sont quantitatives), ensuite l'analyse factorielle correspondances (AFC) (dont les variables sont qualitatives) et enfin l'analyse des correspondances multiples (ACM) une extension d'AFC permettant de représenter et visualiser une table contenant plus de deux (2) variables qualitatives.

L'objectif consiste à fournir des éléments de réponse aux questions suivante :

Quels sont les axes (facteurs) à retenir ?

Quels sont les variables corrélées ? Y-a-t-il un lien entre elles ?

Quelles sont les variables importantes (qualité de représentation) dans mon jeu de données ?

Quels sont les individus contribuent et mieux représentés dans mon jeu de données ?

Etc…

II. STATISTIQUES DESCRIPTIVES

Le jeu de données comporte 07 indicateurs clé de 55 pays. Parmi ces indicateurs nous avons le coefficienit de gini, le revenu par personne, l'espérance de vie, le niveau de perception de la corruption, l'utlisation d'internet et le taux d'emploi de la population agé de plus de 15 ans en 2012.

Variables	Descriptions
GINI	Indice de gini en % en 2012
Income	Revenu par personne en 2012
HDI	Indice de développement humain des pays en 2012
Life_exp	Esperance de vie de chaque pays en % en 2012
CPI	Score de la perception de corruption en % en 2012
TUI	Taux d'utilisateurs d'internet par pays en % en 2012
TE	Taux d'emploi en % en 2012

Tableau 1 : Description des variables

II.1 Statistiques univariées

Variables	Sommaire statistique					
	min	max	mean	sd	median	CoefVar
GINI	27.30	56.20	40.11	7.11	41.30	0.17
Income	706.00	89500.00	14054.44	16795.35	7140.00	**1.19**
HDI	0.37	0.93	0.64	0.18	0.65	0.27
Life_exp	49.60	83.40	69.89	8.91	69.80	0.12
CPI	8.00	90.00	40.22	19.52	34.00	0.48
TUI	0.80	92.30	30.46	29.33	18.20	0.96
TE	0.37	0.88	0.59	0.12	0.58	0.19

Tableau 2: Sommaire statistiques

Ce tableau ci-dessus donne les statistiques des variables. Par exemple, en terme de revenu par personne, on constate qu'en moyenne les pays ont 14054,44$ annuel par personne et que cette somme varie de 760 à 89500$.

On constate qu'il ya une part de variabilité plus importante autour de de la moyenne pour la variable revenu par resonne(Income) soit 1,19 et la variable taux d'utilisateurs d'internet (TUI) de 0,48. Par contre, on peut noter que cette variabilité est moins importante pour les variables indice de gini(GINI) 0,17, esperance de vie (Life_exp) 0,12 et le taux d'emploi (TE) soit 0,19.

II.2 Statistiques bi-variées

II.2.1 Matrice de corrélation

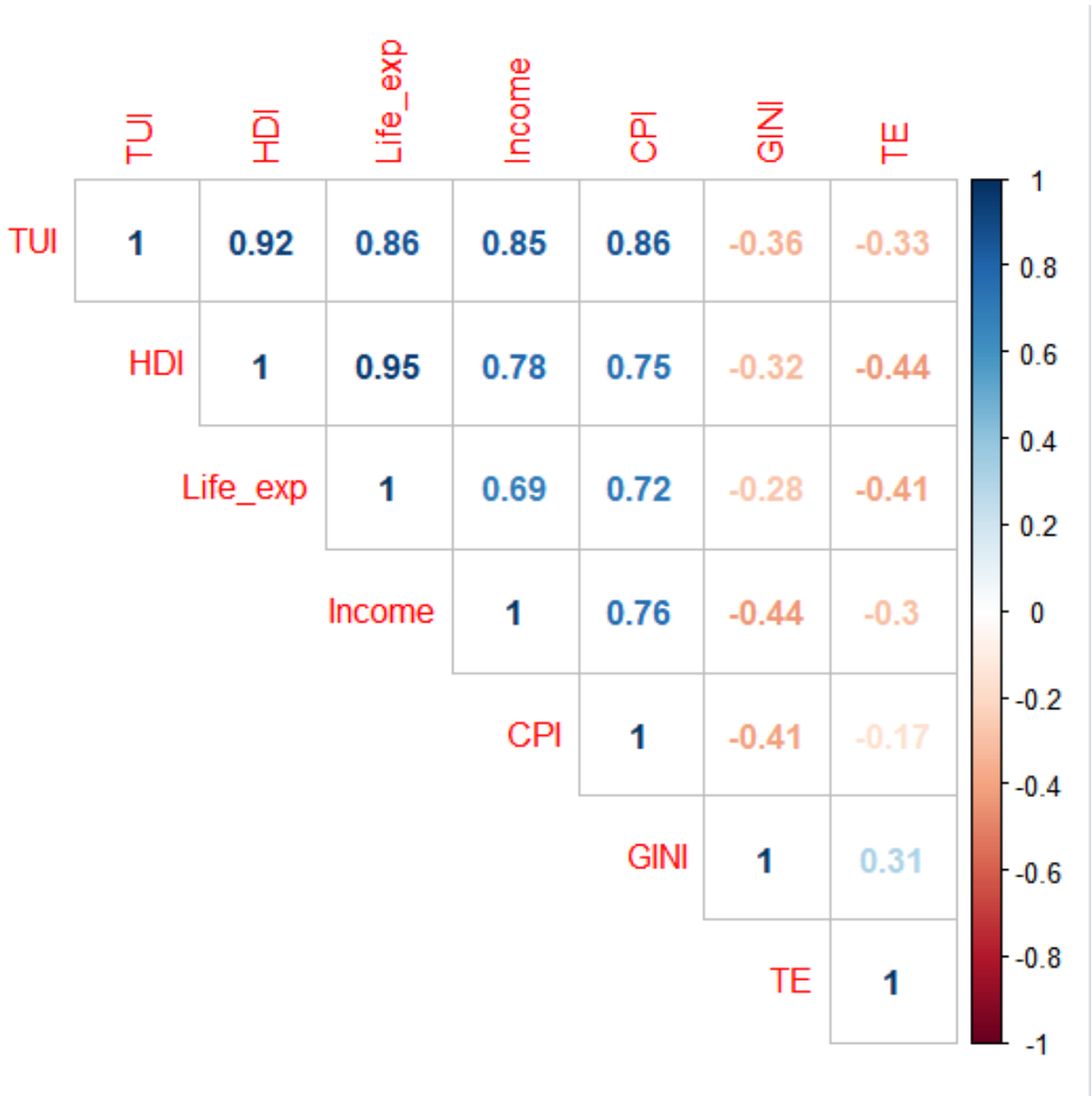


Tableau 3: Matrice de corrélation

Ce graphique ci-dessus donne la matrice de corrélation linéaire des variables. La palette de couleurs nous indique le sens de corrélation à savoir positive ou négative. La couleur bleue indique que les deux variables sont corrélées positivement et en rouge négativement.

La force de corrélation dépend de sa valeur. Selon Jacob Cohen, on peut interpréter le coefficient de corrélation comme le suivant.

Autour de **0,10**	effet de petite taille	corrélation faible
Autour de **0,30**	effet de taille moyenne	corrélation moyenne
Plus de **0,50**	effet de grande taille	corrélation forte

Par exemple, on peut dire qu'il existe une forte corrélation positive entre l'indice de développement humain et l'Esperance de vie soit 0,95. Cela signifie plus l'espérance de vie d'un pays est élevée plus l'IDH est importante. En revanche, entre le score de la perception de corruption et l'indice de Gini, il y a une corrélation moyenne négative soit -0,41. Cela signifie, lorsque la perception de corruption augmente, l'indice de Gini diminue en même temps. A noter que, plus la perception augmente moins il y a de corruption.

II.2.2 Lien linéaire significatif

TABLE N° 15 — **LIMITES D'ACCEPTATION DU COEFFICIENT DE CORRÉLATION LINÉAIRE r LORSQUE $\rho = 0$**

ν \ α	0,10	0,05	0,02	0,01
1	0,9877	0,9969	0,9995	0,9999
2	0,9000	0,9500	0,9800	0,9900
3	0,8054	0,8783	0,9343	0,9587
4	0,7293	0,8114	0,8822	0,9172
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3384	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0,2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540

• La table ci-contre donne pour $|r|$ les limites d'acceptation de l'hypothèse H_0 où le coefficient ρ de corrélation est nul dans la population à deux dimensions, en fonction du risque α de 1ère espèce et du paramètre ν qui dépend de la taille de l'échantillon :

— pour une population normale à 2 variables $\nu = n - 2$

— pour une population normale à p variables où s des $(p - 2)$ autres variables sont fixées $\nu = n - 2 - s$

Tableau 4: Table donnant les limites d'acceptions de r afin de conclure à un lien linéaire significatif (source cours intro)

Notre jeu de données comporte 55 observations reparti sur 7 variables. Pour rappel, un test statistique est construit sur deux hypothèses : l'hypothèse nulle (Ho) qui indique dans notre cas qu'il existe aucun lien linéaire en deux variables et l'hypothèse alternative (H1) qui consiste à rejeter Ho c'est-à-dire qu'il y a bien un lien linéaire entre deux variables dans ma table de données.

On va donc chercher à savoir pour quelle valeur absolue le coefficient de corrélation doit atteindre pour qu'on puisse conclure à un lien linéaire significatif, sachant que le risque de se tromper est de 5%.

D'après le tableau ci-dessus, on a :

V=55-2-(7-2) =55-2-5=48, par consequent pour un risque de 5% la valeur absolue du coefficient de corrélation doit etre supérieur à 0,2732.

Autrement dit, $|r|>0,2732$.

En s'appuyant sur cette méthode, on peut donc en déduire seules deux variables n'ont pas de lien linéaire entre elles. Le score de perception de corruption et le taux d'emploi n'auraient aucun lien linéaire. Cela signifie que plus le taux de perception de corruption augmente plus le taux d'emploi augmente également.

III. ANALYSE EN COMPOSANTES PRINCIPALES

Dans cette partie, nous allons travailler sur des variables quantitatives. On va donc chercher à représenter notre jeu de données pour pouvoir visualiser les individus (c'est-à-dire détection d'individus ou groupes d'individus atypiques) et les variables (liaisons et sélection). En ce sens, on pourra procéder à la réduction du nombre de variables et réduire la variabilité.

III.1 Choix des axes

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.63	66.08	66.08
comp 2	0.96	13.78	79.86
comp 3	0.80	11.48	91.34
comp 4	0.28	4.02	95.36
comp 5	0.22	3.17	98.53
comp 6	0.07	1.00	99.54
comp 7	0.03	0.46	100.00

Tableau 5 : Tableau des Valeurs Propres

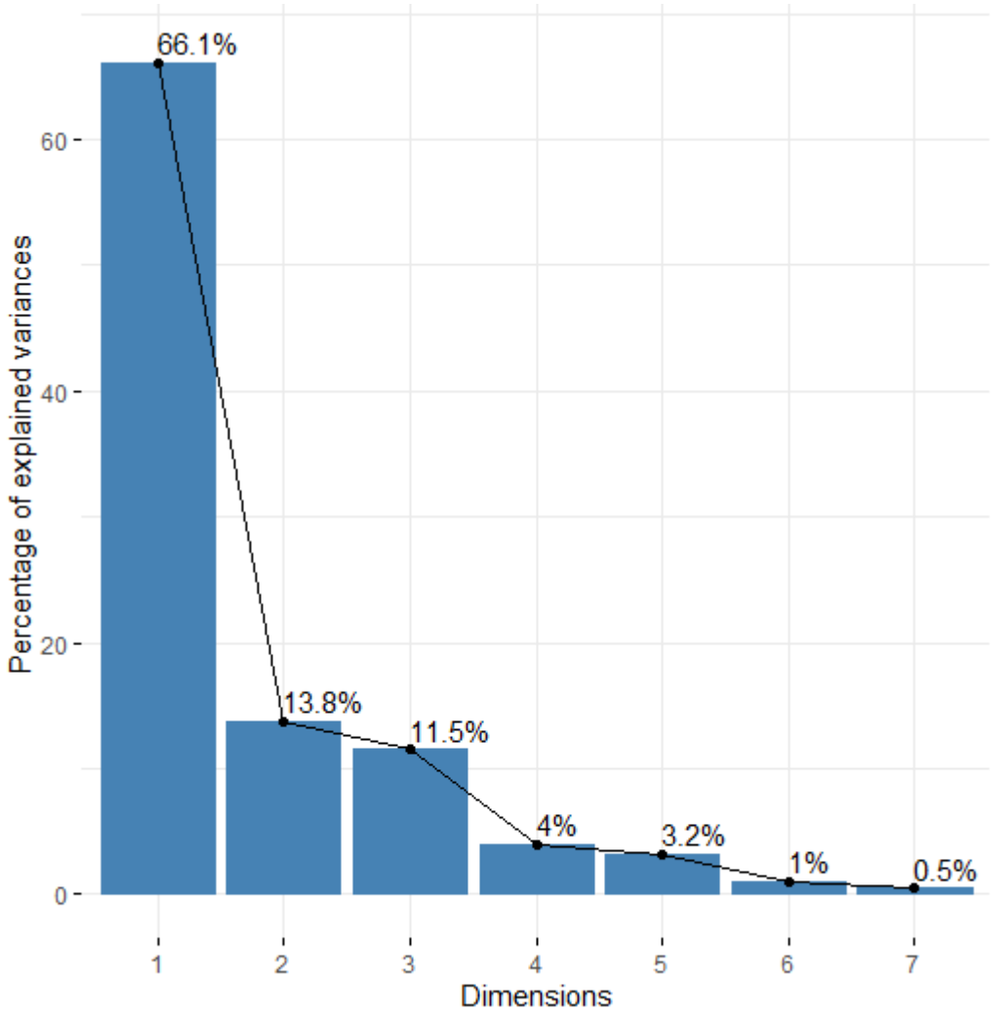


Figure 1 : Visualisation des valeurs propres

Le critère de Kaiser et éboulis des valeurs propres(coude) sont deux méthodes permettant de sélectionner le nombre d'axes à retenir pour la suite de l'analyses en composantes principales. Dans notre cas, l'éboulis des valeurs propres n'est pas recommandé.

Selon le critère de Kaiser qui consiste à prendre toutes les valeurs propres supérieurs à 1, on devrait garder qu'un seul axe. Par ailleurs, comme on peut le voir dans le tableau ci-dessus *tableau 5* l'axe 2 est proche de 1, par consequent on peut choisir les deux. De ce fait, les deux premiers axes concentrent 79,86% de l'inertie des nuages de points ce qui est très significatif.

$coord

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
GINI	-0.4938416	0.55109581	0.65190546	0.15186967	-0.064473445
Income	0.8829379	0.08274818	-0.14958551	0.38247780	0.193061847
HDI	0.9492206	0.07744488	0.20347663	-0.12636137	0.116173571
Life_exp	0.9067603	0.09583461	0.24196126	-0.28135238	0.090314643
CPI	0.8620857	0.24445934	-0.22387899	0.02043750	-0.377072418
TUI	0.9593436	0.17831677	0.02507171	0.03951384	-0.009715805
TE	-0.4638508	0.73974364	-0.45363325	-0.12252470	0.129524030

$cor

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
GINI	-0.4938416	0.55109581	0.65190546	0.15186967	-0.064473445
Income	0.8829379	0.08274818	-0.14958551	0.38247780	0.193061847
HDI	0.9492206	0.07744488	0.20347663	-0.12636137	0.116173571
Life_exp	0.9067603	0.09583461	0.24196126	-0.28135238	0.090314643
CPI	0.8620857	0.24445934	-0.22387899	0.02043750	-0.377072418
TUI	0.9593436	0.17831677	0.02507171	0.03951384	-0.009715805
TE	-0.4638508	0.73974364	-0.45363325	-0.12252470	0.129524030

$cos2

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
GINI	0.2438795	0.303706587	0.4249807245	0.0230643981	4.156825e-03
Income	0.7795793	0.006847262	0.0223758253	0.1462892701	3.727288e-02
HDI	0.9010197	0.005997709	0.0414027392	0.0159671957	1.349630e-02
Life_exp	0.8222143	0.009184272	0.0585452520	0.0791591612	8.156735e-03
CPI	0.7431918	0.059760370	0.0501218036	0.0004176914	1.421836e-01
TUI	0.9203401	0.031796870	0.0006285907	0.0015613434	9.439686e-05
TE	0.2151576	0.547220653	0.2057831288	0.0150123017	1.677647e-02

$contrib

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
GINI	5.272635	31.4880525	52.86894756	8.1942255	1.87128714
Income	16.854376	0.7099185	2.78362350	51.9730566	16.77921314
HDI	19.479898	0.6218376	5.15063183	5.6727603	6.07565855
Life_exp	17.776137	0.9522178	7.28321469	28.1233447	3.67193527
CPI	16.067683	6.1959066	6.23531105	0.1483957	64.00710853
TUI	19.897600	3.2966737	0.07819867	0.5547077	0.04249484
TE	4.651672	56.7353932	25.60007270	5.3335095	7.55230252

Figure 2 : Coordonnées et qualités de représentation des variables

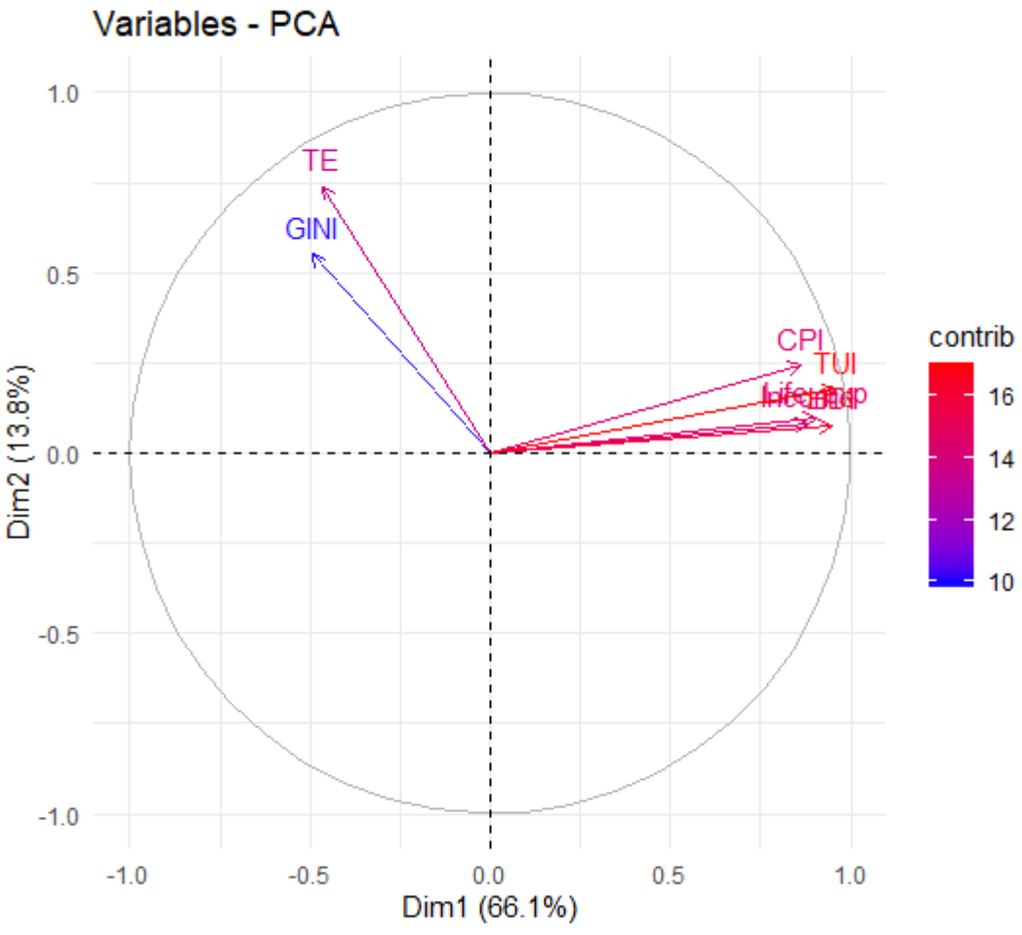


Figure 3 : Projection des variables sur le premier plan factoriel

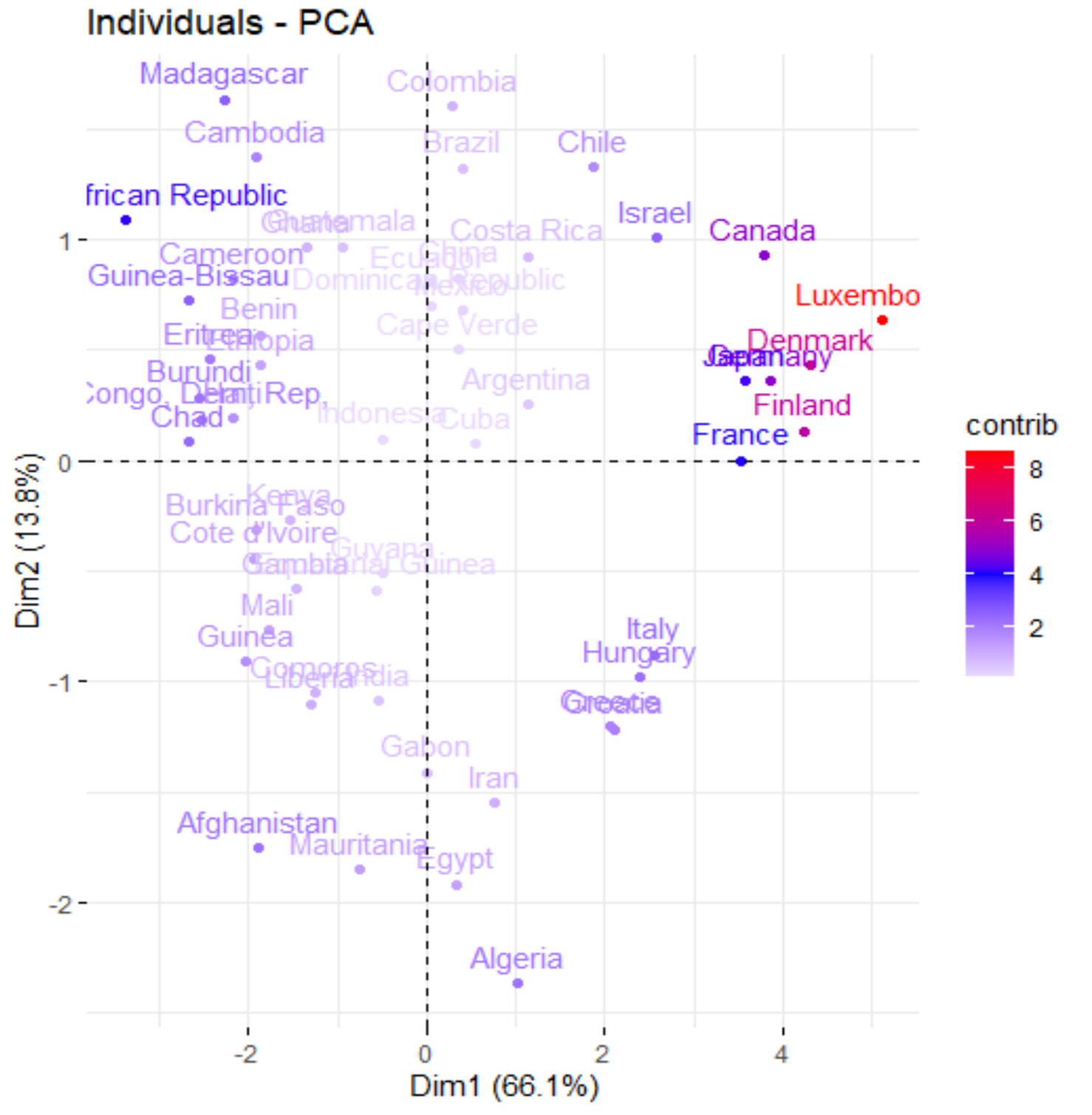


Figure 4 : Visualisation des individus (pays) sur le premier plan factoriel

	Dim.1	Dim.2			
Afghanistan	FALSE	TRUE			
Algeria	FALSE	TRUE	Ethiopia	FALSE	FALSE
Argentina	FALSE	FALSE	Finland	TRUE	FALSE
Benin	FALSE	FALSE	France	TRUE	FALSE
Brazil	FALSE	TRUE	Gabon	FALSE	TRUE
Burkina Faso	FALSE	FALSE	Gambia	FALSE	FALSE
Burundi	TRUE	FALSE	Germany	TRUE	FALSE
Cambodia	FALSE	TRUE	Ghana	FALSE	FALSE
Cameroon	TRUE	FALSE	Greece	FALSE	TRUE
Canada	TRUE	FALSE	Guatemala	FALSE	FALSE
Cape Verde	FALSE	FALSE	Guinea	FALSE	FALSE
Central African Republic	TRUE	TRUE	Guinea-Bissau	TRUE	FALSE
Chad	TRUE	FALSE	Guyana	FALSE	FALSE
Chile	FALSE	TRUE	Haiti	TRUE	FALSE
China	FALSE	FALSE	Hungary	TRUE	FALSE
Colombia	FALSE	TRUE	India	FALSE	TRUE
Comoros	FALSE	TRUE	Indonesia	FALSE	FALSE
Congo, Dem, Rep,	TRUE	FALSE	Iran	FALSE	TRUE
Costa Rica	FALSE	FALSE	Israel	TRUE	TRUE
Cote d'Ivoire	FALSE	FALSE	Italy	TRUE	FALSE
Croatia	FALSE	TRUE	Japan	TRUE	FALSE
Cuba	FALSE	FALSE	Kenya	FALSE	FALSE
Denmark	TRUE	FALSE	Liberia	FALSE	TRUE
Dominican Republic	FALSE	FALSE	Luxembourg	TRUE	FALSE
Ecuador	FALSE	FALSE	Madagascar	TRUE	TRUE
Egypt	FALSE	TRUE	Mali	FALSE	FALSE
Equatorial Guinea	FALSE	FALSE	Mauritania	FALSE	TRUE
Eritrea	TRUE	FALSE	Mexico	FALSE	FALSE

Figure 5 ; Visualisation des individus ayant contribué sur l'axe 1 et axe 2, sachant True signifie contribuent et False ne contribuent pas.

III.2 Interprétation de l'axe 1 selon les variables et les individus

Variables

Une variable contribue à l'élaboration d'un axe si sa valeur de contribution est supérieure à la moyenne, à savoir 1/n c'est-à-dire 1/7 ou encore 0,143 soit 14,3%. Plus sa valeur est élevée, et plus la variable contribue à la construction de l'axe.

L'axe 1 oppose les pays dont le revenu par personne, le taux d'utilisation d'internet, l'Esperance de vie, le niveau de perception de la corruption, et sont les plus importants à ceux qui ne l'ont pas. En revanche, les contribution du taux d'emploi et le taux d'inéquality sont assez faibles, par consequent elles n'interviennent pas dans l'interpretation du l'axe1 mais plutôt l'axe 2.

A noter que, plus le niveau de perception de corruption est élevé, moins il y a de corruption.

En terme de qualité de representation (c'est-à-dire proche du cercle de correlation) , toutes les variables ne sont pas bien representées. En effet, seules les variables indidice de developpement humain (HDI) et le taux d'utilisation d'internet (TUI) sont bien representées.

Le second axe est construit par le taux d'inégalité et le taux d'emploi et sont mal representées.

Individus

De la même façon, les individus qui contribuent à l'axe 1 sont ceux dont leurs contributions sont supérieures à la moyenne, à savoir 1/n c'est-à-dire 1/55 ou encore 1,818%.

D'après les sortie R des *figures 4 et 5*, les individus qui contribuent à la construction de l'axe 1 sont :

Gauche	Droite
Madagascar, Burundi, Cameroon, Central African Republic, Chad, Congo, Eritrea, Guinea-Bisseau, Haiti	Canada, Israel, Italy, Luxembourg, Denmark, Finland, France Germany, Hungary, Japan

On peut donc penser que les pays faisant partie de la colonne droite ont des revenus par personne, taux d'utilisation d'internet, Esperance de vie, niveau de perception de la corruption plus importants, alors que les pays à gauche ne le sont pas. A noter que, plus le niveau de perception de corruption est élevé, moins il y a de corruption.

III.3 Interprétation de l'axe 2 selon les variables et les individus.

Variables

Sur l'axe 2, seules les variables taux d'inégalité et taux d'emploi contribuent à la création du second l'axe. L'axe 2 oppose les pays dont GINI et TE sont élevés et ceux qui ne le sont pas.

Sur cet axe, aucune des deux variables sont bien représentées.

Individus

Gauche	Droite
Afghanistan, Mauritania Egypte, Gabon, Comoros, Croatia, Greece, India, Iran, Liberia,	Israel, Madagascar, Central African Republic, Brazil Colombia, Cambodia, chili

On peut penser que les pays de la colonne gauche ont un taux d'inégalité de revenu et taux d'emploi assez élevés.

III.4 Interprétation des individus sur le premier plan factoriel

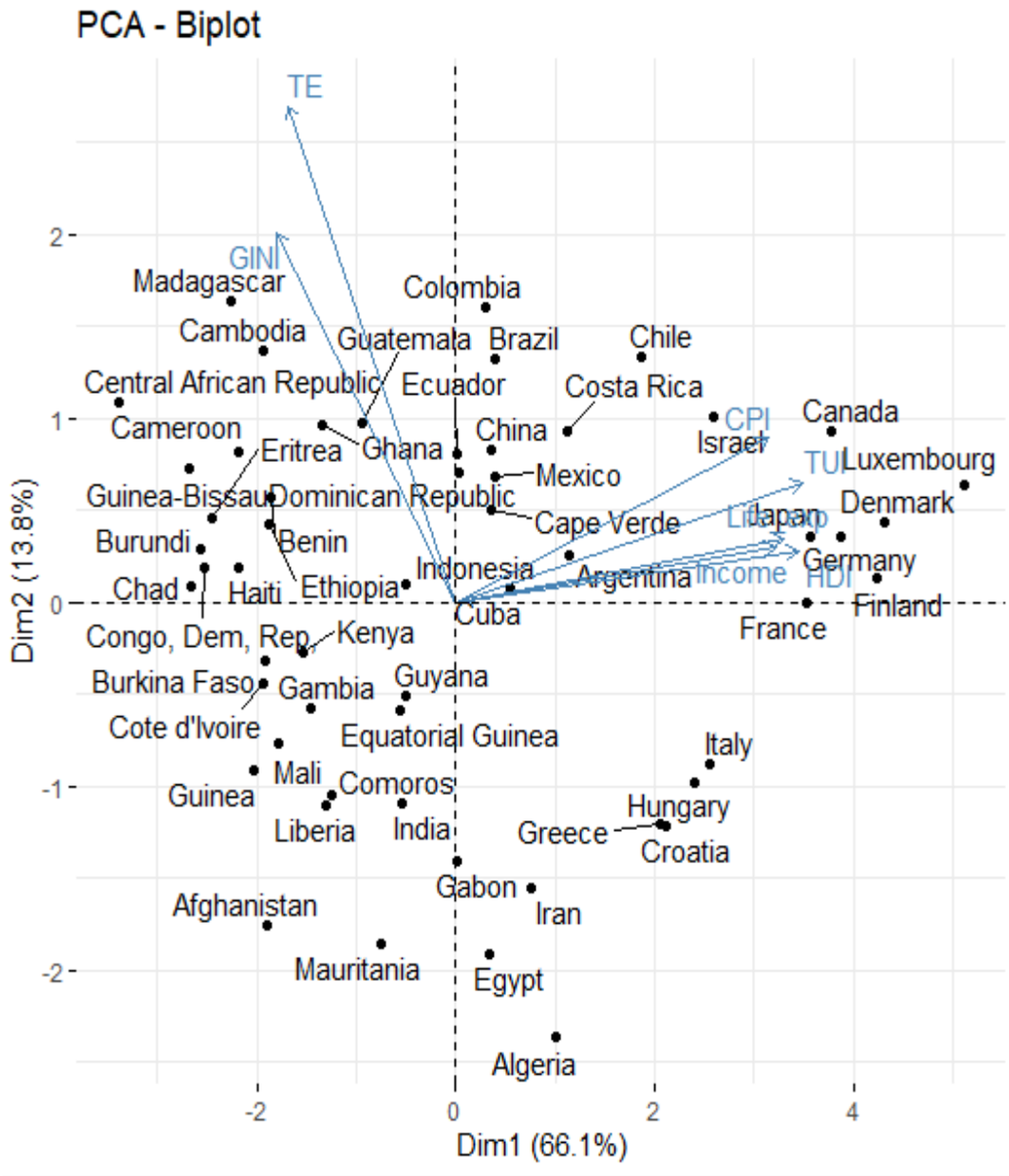


Figure 6: Représentation des individus et les variables sur le premier plan factoriel

On remarque les pays ayant un taux de revenu(income) et niveau de perception de corruption (CPI) élevés ont un taux d'espérance de vie, indice de développement et un taux d'utilisation d'internet également. Ces pays sont majoritairement des pays d'Europe et développés. En revanche, ceux ayant un indice de Gini c'est-à-dire un niveau d'inégalité de revenu et taux d'emploi élevé sont des pays d'Afrique et sous-développés.

IV. ANALYSE FACTORIELLE DES CORRESPONDANCES

Dans cette partie, nous allons aborder l'analyse factorielle des correspondances. En effet, on va visualiser les correspondances entre les modalités d'une même variable et représenter simultanée des modalités des deux (2) variables afin d'analyser les liens entre les deux variables.

On peut se demander, pourquoi ne pas effectuer directement un ACP sur nos deux variables catégorielles ?

On ne le fait pas, parce que la distance euclidienne entre deux modalités n'a pas de sens. On cherchera plutôt à représenter les conditionnelles de chaque modalité. En fait, effectuer une AFC revient à faire un ACP sur le tableau des profils lignes (avec n>=p) ou profils colonnes (avec p>=n).

IV.1 Tableaux de contingence, fréquences et du Khi2

GINI	CPI				Total
	[8,28)	[28, 34)	[34, 45.5)	[45.5, 90)	
[27.3-34.3)	0	2	3	9	14
[34.3- 41.3)	5	3	5	0	13
[41.3-44.75)	4	3	5	2	14
[44.75-56.2)	4	4	3	3	14
Total	13	12	16	14	55

Tableau 6 : Tableau des effectifs observés

GINI	CPI				Total
	[8,28)	[28, 34)	[34, 45.5)	[45.5, 90)	
[27.3-34.3)	3,31	3,05	4,07	3,56	14
[34.3- 41.3)	3,07	2,84	3,78	3,31	13
[41.3-44.75)	3,31	3,05	4,07	3,56	14
[44.75-56.2)	3,31	3,05	4,07	3,56	14
Total	13	12	16	14	55

Tableau 7 : Tableau des effectifs théoriques

GINI	CPI				Total
	[8,28)	[28, 34)	[34, 45.5)	[45.5, 90)	
[27.3-34.3)	0%	4%	5%	16%	25%
[34.3- 41.3)	9%	5%	9%	0%	24%
[41.3-44.75)	7%	5%	9%	4%	25%
[44.75-56.2)	7%	7%	5%	5%	25%
Total	24%	22%	29%	25%	100%

Tableau 8 : Tableau des fréquences observées

GINI	CPI				Total
	[8,28)	[28, 34)	[34, 45.5)	[45.5, 90)	
[27.3-34.3)	6%	6%	7%	6%	25%
[34.3- 41.3)	6%	5%	7%	6%	24%
[41.3-44.75)	6%	6%	7%	6%	25%
[44.75-56.2)	6%	6%	7%	6%	25%
Total	24%	22%	29%	25%	100%

Tableau 9 : Tableau de fréquences théoriques

GINI	CPI				Total
	[8,28)	[28, 34)	[34, 45.5)	[45.5, 90)	
[27.3-34.3)	3,31	0,36	0,28	8,29	12,25
[34.3- 41.3)	1,21	0,01	0,39	3,31	4,92
[41.3-44.75)	0,14	0,00	0,21	0,69	1,04
[44.75-56.2)	0,14	0,29	0,28	0,09	0,81
Total	5	1	1	12	19

Tableau 10 : Tableau des valeurs Khi2

IV.1.1 Comparaison du Khi2 sortie R et méthode fréquences observées & théoriques

```
**Results of the Correspondence Analysis (CA)**
The row variable has  4  categories; the column variable has 4 categories
The chi square of independence between the two variables is equal to 19.01971 (p-value =  0.0250258 ).
*The results are available in the following objects:
```

D'après cette sortie de R, on a un Khi2 très faible soit 19,01. En comparant cette valeur à celle de la méthode des fréquences observées & théoriques, on remarque qu'elles sont égales soit 19,01~=19.

IV.2 Choix valeurs propres

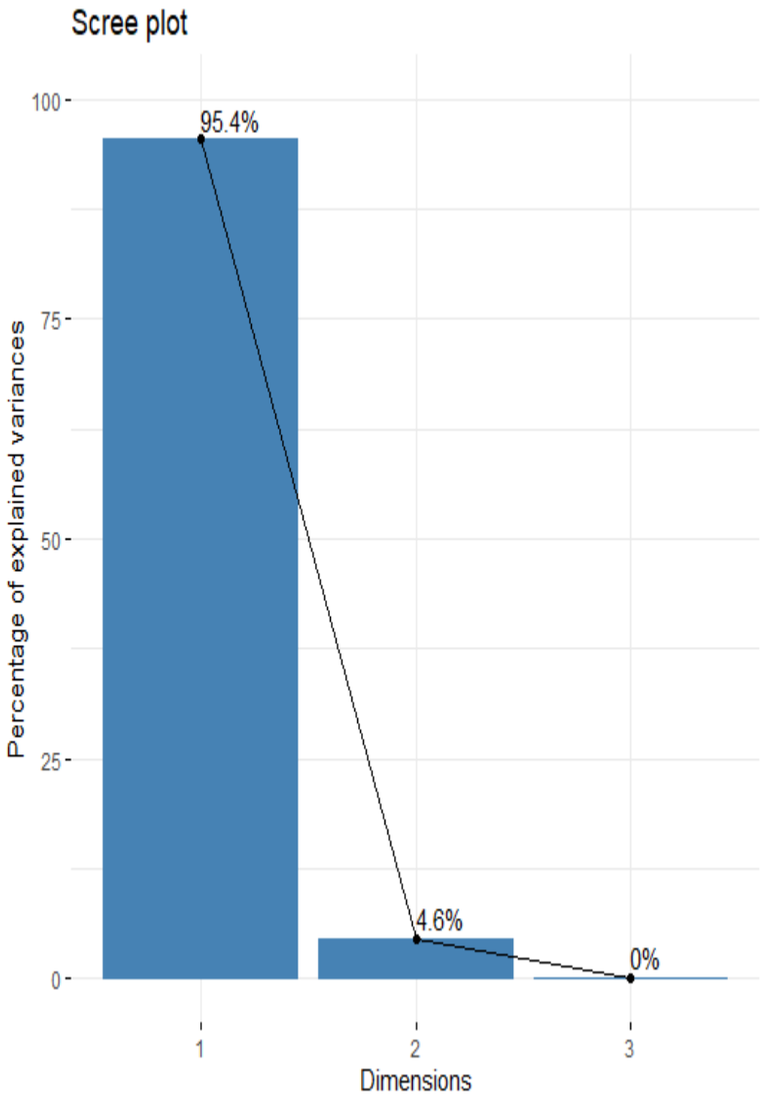


Figure 7 : Visualisation des Valeurs Propres

Valeur propre	Valeur	% variance	% cumulé
λ_1	0,330	95,406	95,406
λ_2	0,016	4,559	99,965
λ_3	0,000	0,035	100,000
Total	0,346		

Tableau 11 : Tableau des valeurs Propres

Le choix des axes dépend de nombre de valeurs propres retenus. En effet, selon la méthode Min (n, p) -1=Min (4,4) -1=4-1=3, cela prouve bien que l'analyse factorielles des correspondances fournit bien 3 valeurs propres.
Comme on peut le voir dans le tableau ci-dessus *tableau 10* ainsi que *la figure 7,* les deux premiers axes concentrent 89,63% de l'inertie des nuages de points. Par consequent, on ne gardera que les deux premiers axes pour l'analyse.
La dernière ligne du *tableau 10* est la somme des valeurs propres et permet d'en déduire le Khi2.

Lignes	Coordonnées		Contributions		Cosinus carrés	
%GINI	Dim1	Dim2	Dim1	Dim2	Dim1	Dim2
[27.3-34.3)	-0.934	-0.048	67.321	3.768	O.997	0.003
[34.3- 41.3)	0.610	-0.081	26.632	9.732	0.982	0.017
[41.3-44.75)	0.257	-0.090	5.098	13.104	0.887	0.109
[44.75-56.2)	0.111	0.213	0.949	73.396	0.213	0.787

Colonnes	Coordonnées		Contributions		Cosinus carrés	
CPI	Dim1	Dim2	Dim1	Dim2	Dim1	Dim2
[8,28)	0.605	0.055	26.255	4.523	0.991	0.008
[28, 34)	0.171	0.162	1.924	36.371	0.523	0.473
[34, 45.5)	0.203	-0.178	3.628	58.789	0.563	0.436
[45.5, 90)	-0.940	0.014	68.193	0.318	1.000	0.000

Figure 8 : Coordonnées, contributions et cosinus carrés des modalités lignes et colonnes.

Ce tableau ci-dessus montre dans un premier temps les indicateurs des points-lignes sur les deux premiers axes (premier plan factoriel). Sur le premier axe, le point [27.3-34.3) de Gini s'oppose à tous les autres points, en particulier le point « [34.3-41.3) ». Elle a une contribution de 67,32% et un cosinus près de 0.997. En effet, il est quasiment sur l'axe et aura surement une représentation très faible sur l'axe 2. Par ailleurs, on peut noter que le point [44.75-56.2) est peu représenté sur l'axe 1 avec une contribution de moins de 1% soit 0,959%.

Le second axe, concentre 4,6% d'inertie sur le premier plan factoriel. Il est construit essentiellement par la modalité [44.75-56.2) de Gini avec une contribution de 73,37% et s'oppose simultanément aux points « [41.3-44.75) » et « [34.3-41.3) » (contributions respectives de 13,10% et 9,73%). La modalité « point [44.75-56.2) » c'est le seul point bien représenté sur l'axe 2 avec une cos2 de 0,787.

Pour les points-colonnes, le premier axe est essentiellement construit par la modalité du taux d'emploi « [45.5-90) » soit 68,19% et s'oppose simultanément à la modalité « [8-28) » et sont très bien représentées (cos2 respectives 1 et 0.991) ». En revanche l'axe 2 est lié aux classes « [34-45.5) » et « [28-34) ».

IV.3 Interprétation des points lignes et colonnes sur le premier plan factoriel

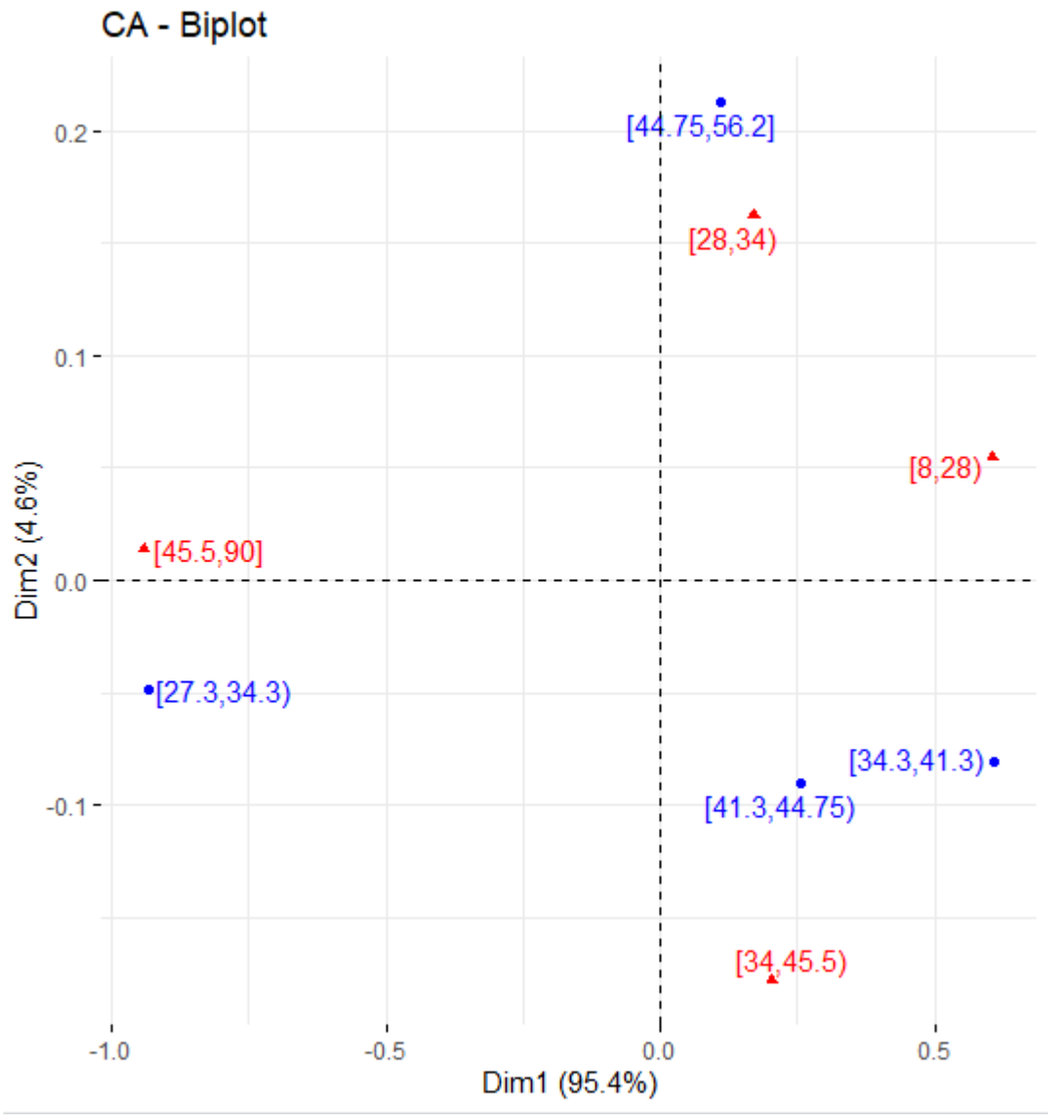


Figure 9 : Visualisation des modalités lignes et colonnes sur le premier plan factoriel

Ce graphique ci-dessus représente simultanément les points lignes et colonnes. En effet, on constate que certaines modalités des deux variables se rapprochent et d'autres s'éloignent. Par exemple sur l'axe 2, on remarque le point « [44.75-56.2) » est plus excentré que loin « [8-28) ». Cela signifie que les pays ayant un pourcentage d'inégalité de revenu dans cet intervalle ont un taux de perception de corruption plus élevé proportionnellement.

En ce sens, on peut donc penser que les pays ayant d'inégalité de revenu dans cette tranche sont plus susceptible d'etre corrompus que l'inverse.

V. ANALYSE DES CORRESPONDANCES MULTIPLES

V.1 Choix des axes

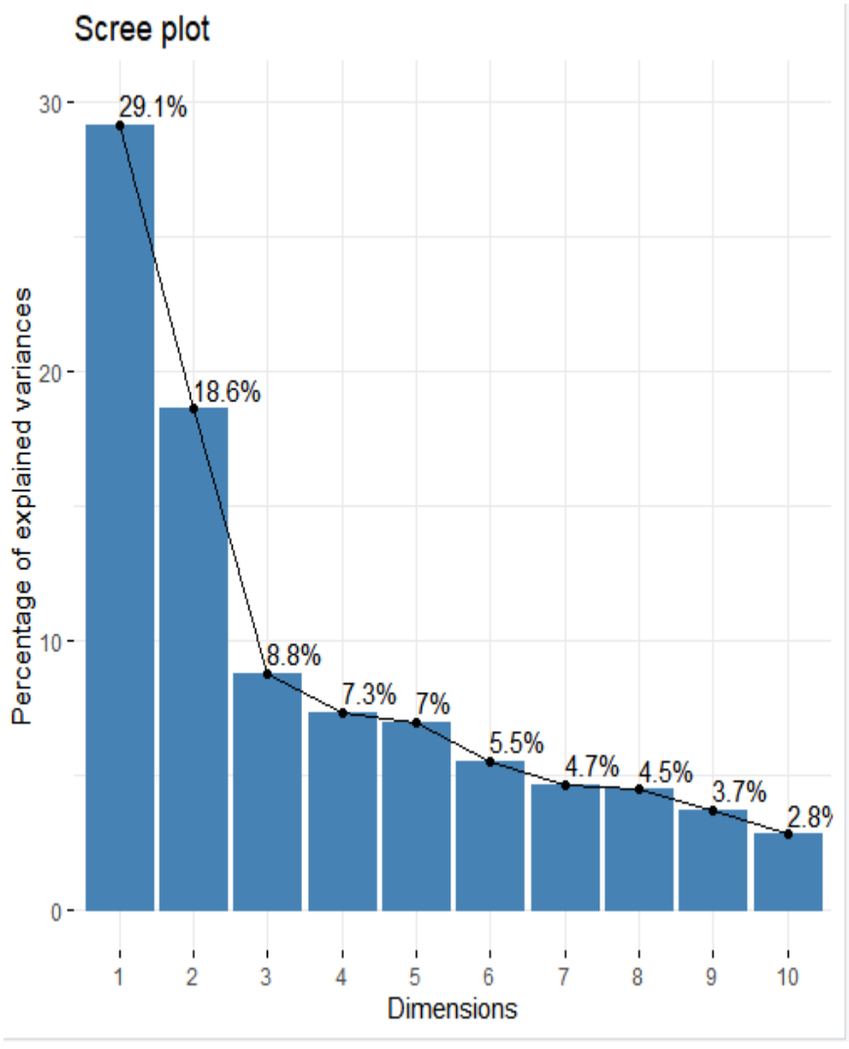


Figure 10: Visualisation des valeurs propres

Tout comme ACP et AFC, le choix des axes peut se faire le critère d'éboulis des valeurs propres (coude) et la règle de kaiser. Ce graphique ci-dessus permet de visualiser les pourcentages des variances expliquées afin de pouvoir procéder au choix des axes. Selon le critère de kaiser, on garde tous les facteurs dont sa valeur propre est supérieure à la moyenne des valeurs propres à savoir 100/7 ou encore 14,28%. Par consequent, cela revient à garder les deux premiers facteurs (respectivement 29,1% et 18,6%) ce qui implique que l'analyse se fera que sur le premier plan factoriel.

V.2 Interprétation des résultats ACM sur le premier plan factoriel

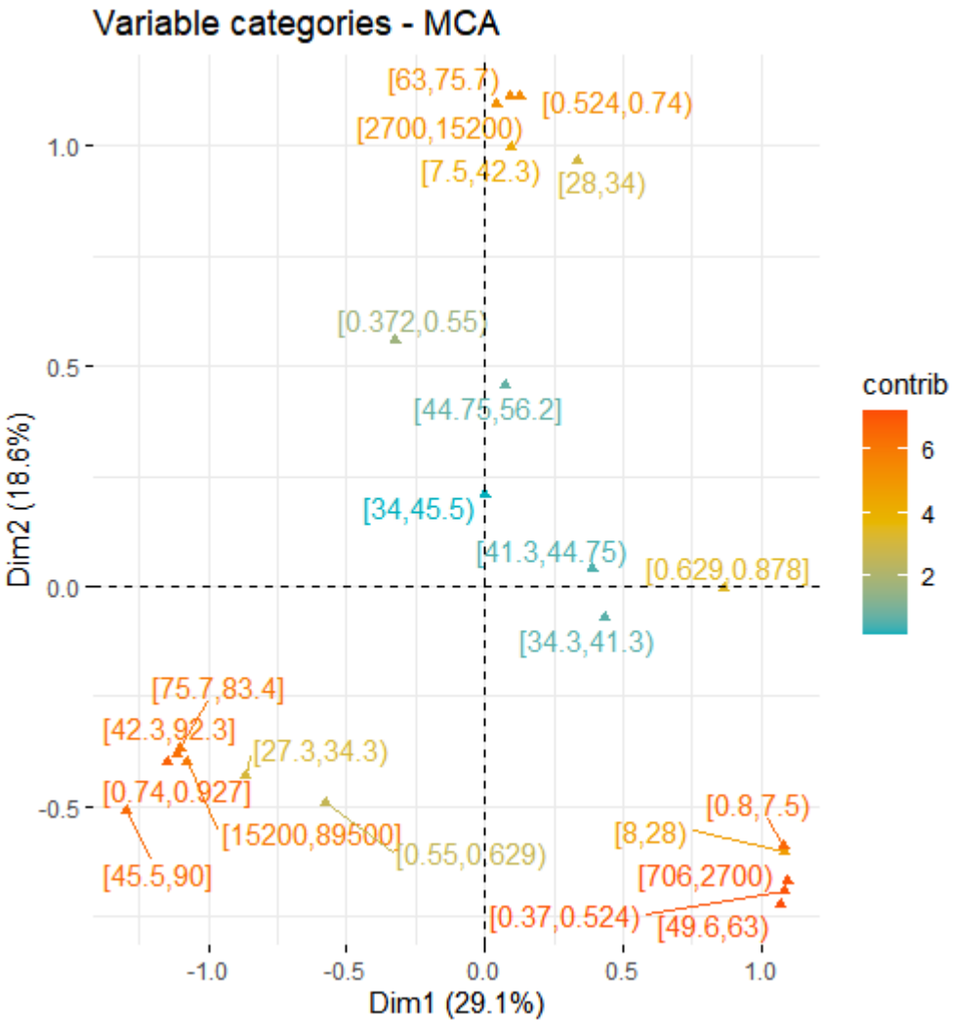


Figure 11 : Contribution des modalités sur le plan factoriel

Ce graphique ci-dessus met en évidence la contribution des modalités des variables sur le premier plan factoriel. En effet, on peut identifier 4 groupes d'individus homogènes.

En revanche, il est évident que les catégories en bas à gauche (sauf [273,34.3]) ont une contribution négative importante à la création du premier axe tandis que les modalités comme [0.37,0.524], [49.6, 63] [63,75.5] [0.8,07.5], [706,2700] sont fortement contribuées à la création premier axe positivement.

Sur le second axe, on retrouve les modalités [0.37,0.524], [49.6, 63] [63,75.5] qui contribuent négativement et [0.524,0.74], [2700, 15200] positivement.

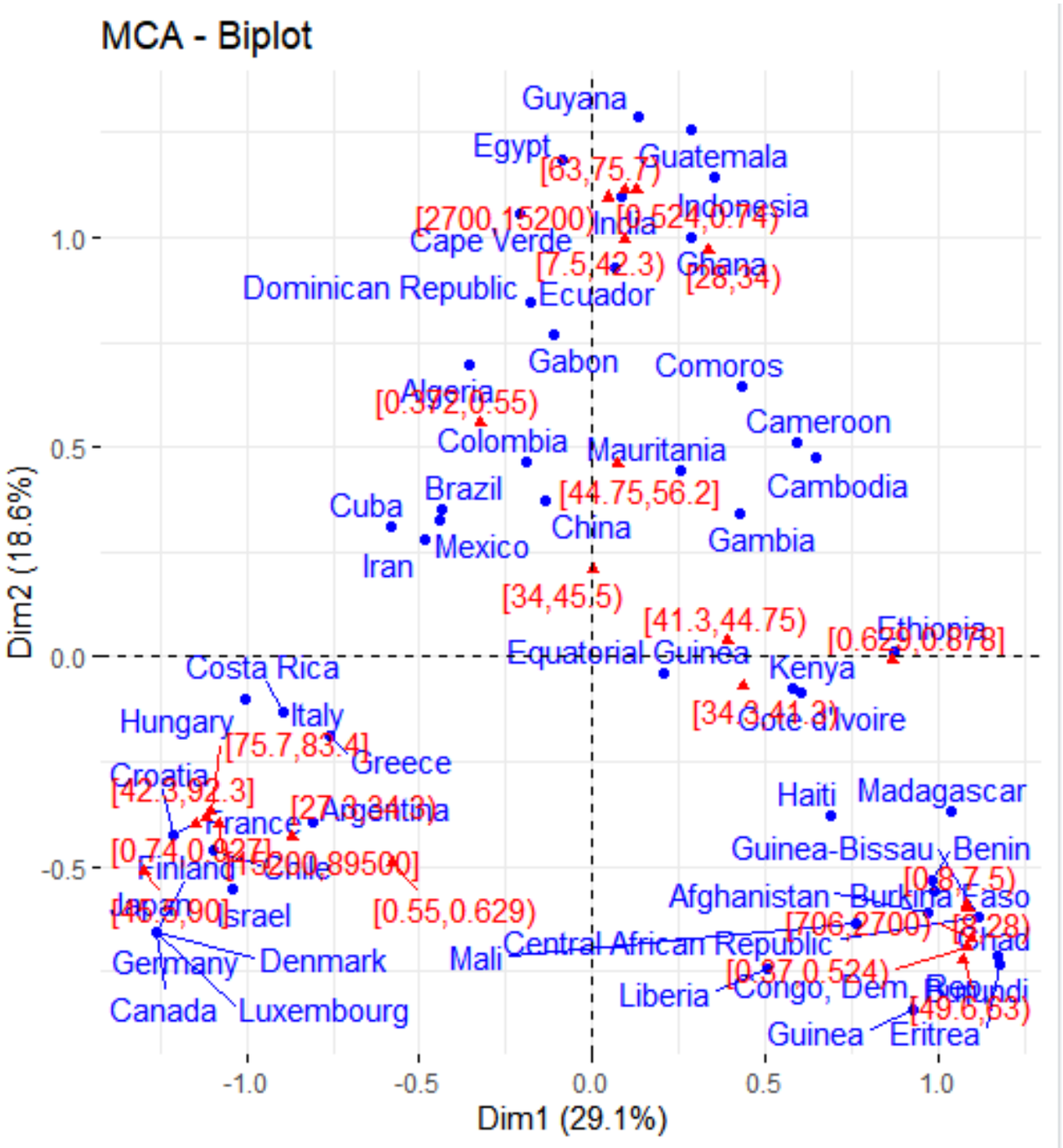


Figure 12 : Visualisation des individus et modalités sur le premier plan factoriel

Ce graphique ci-dessus met en évidence les individus et les différentes modalités. Graphiquement, on peut identifier des quatre (4) groupes d'individus homogènes. Par exemple, les modalités de Revenu on peut toute suite différencier les pays de chaque groupe.

Pour un revenu [706, 2700], on remarque que ce sont presque tous des pays d'Afrique (Benin, Congo, Guinée…).

Pour un revenu [2700, 15000], on retrouve des pays comme Egypte, Guyana, Cape Verde, Inde …

Enfin ceux qui ont des revenus plutôt élevés [15000, 89500] sont majoritairement des pays Européens comme la France, Allemagne, Canada, Finlande …

VI. BILAN & CONCLUSION

Ce projet a été très intéressant et enrichissant, car la méthode factorielle de représentation est très demandée et importante lorsqu'on traite une table contenant plusieurs variables. Basée sur la notion d'inertie de nuage de point, elle permet de résoudre le problème de visualisation des données au sein d'un jeu de données.

Sur les trois (3) méthodes étudiées, on a pu constater des groupes d'individus homogènes. Par exemple, les pays Européens et développés ont un revenu par personne plus important que les autres [2700, 15000]. Inversement, les pays d'africains ont plus tendance à etre dans un intervalle de [706, 2700].

On a aussi traité la notion de corrélation et la significativité linéaire entre deux variables.

Ce projet a été plus que productif, car il nous a permis de familiariser un peu plus sur l'utilisation des logiciels et la programmation.

VII. TABLE DES ILLUSTRATIONS

VIII. BIBLIOGRAPHIE

http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/75-acm-analyse-des-correspondances-multiples-avec-r-l-essentiel/

https://explorable.com/fr/la-correlation-statistique

http://larmarange.github.io/analyse-R/recodage.html#renommer-des-variables

https://mtes-mct.github.io/parcours-r/m4/lacp.html#principe-de-lacp

IX. ANNEXE

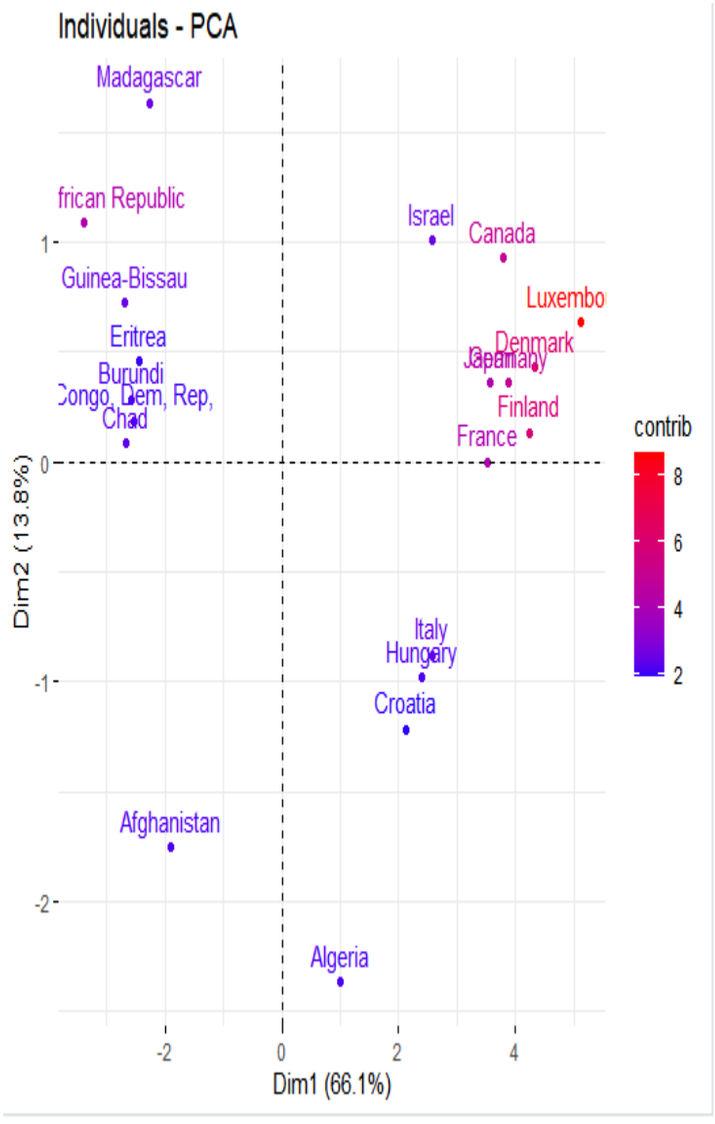


Figure 13 : Les 20 individus ayant le plus contribués

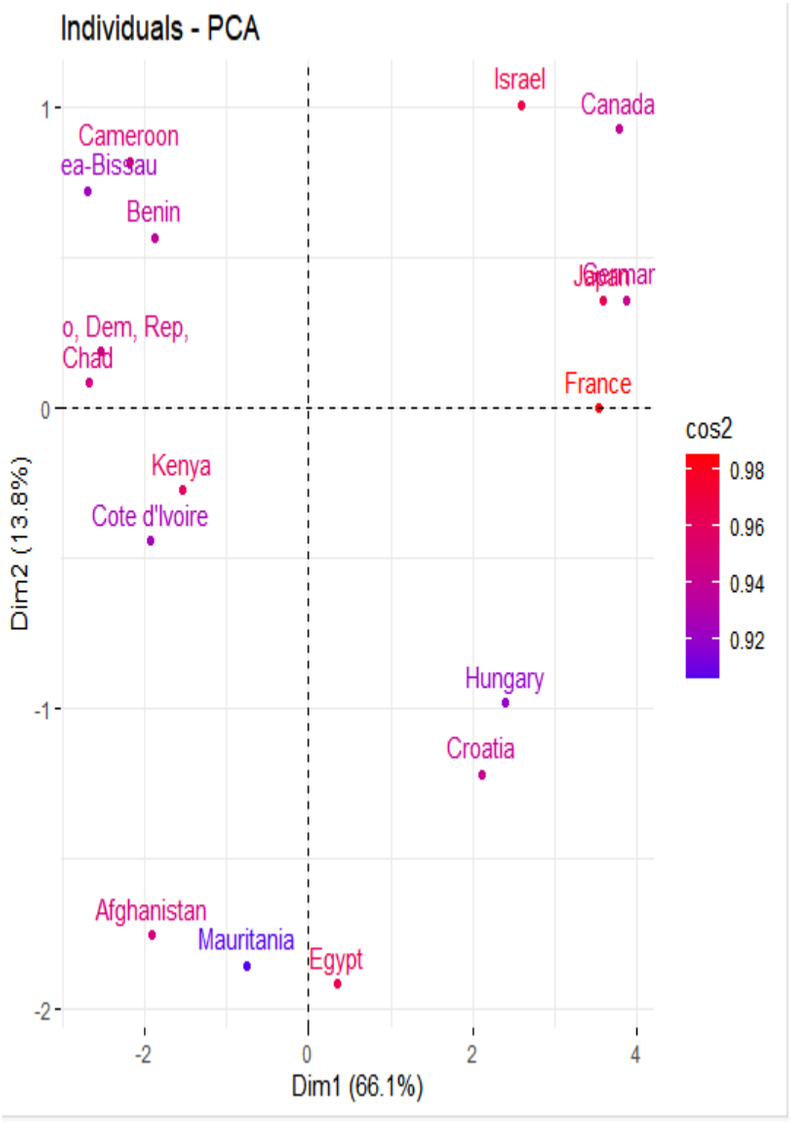


Figure 14 ; Pays ayant une bonne qualité de représentation

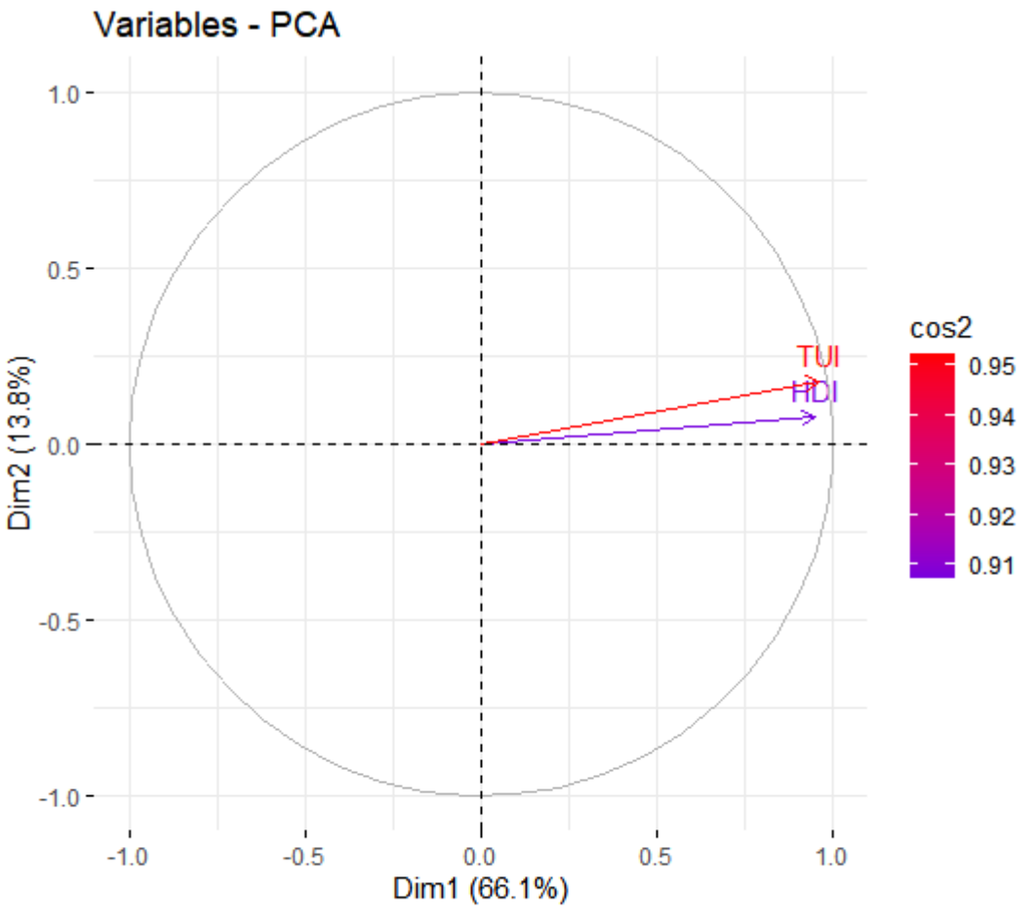
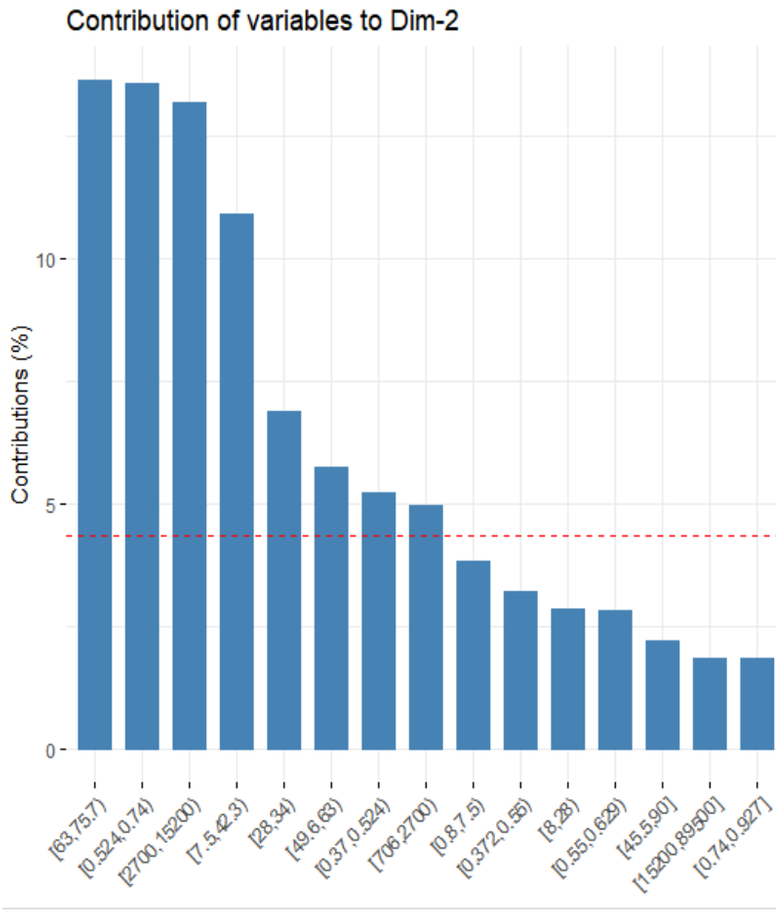
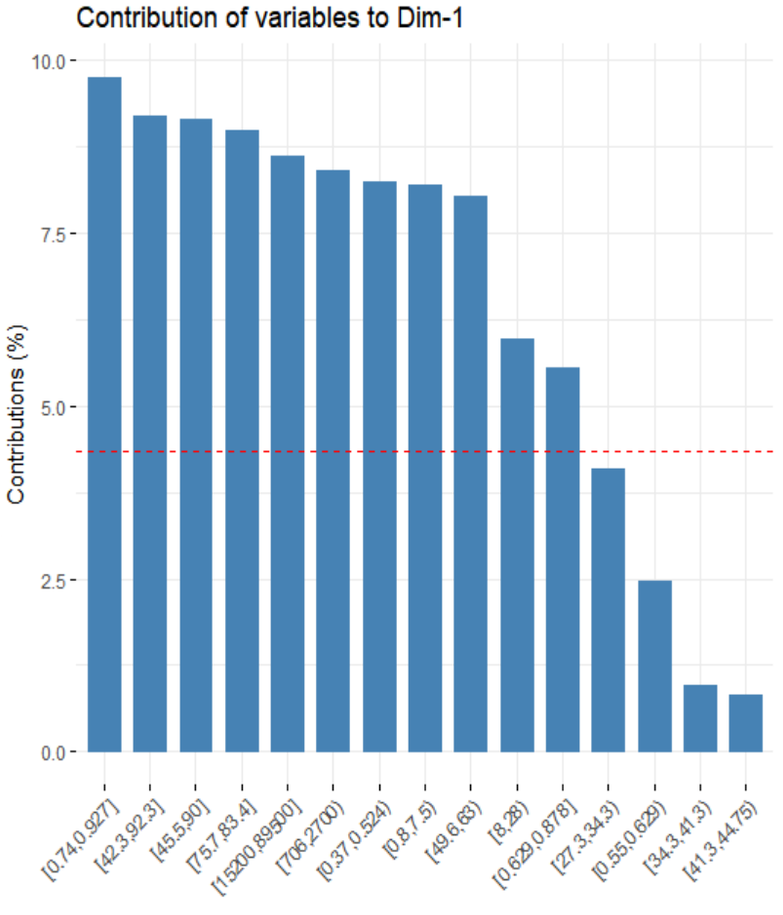
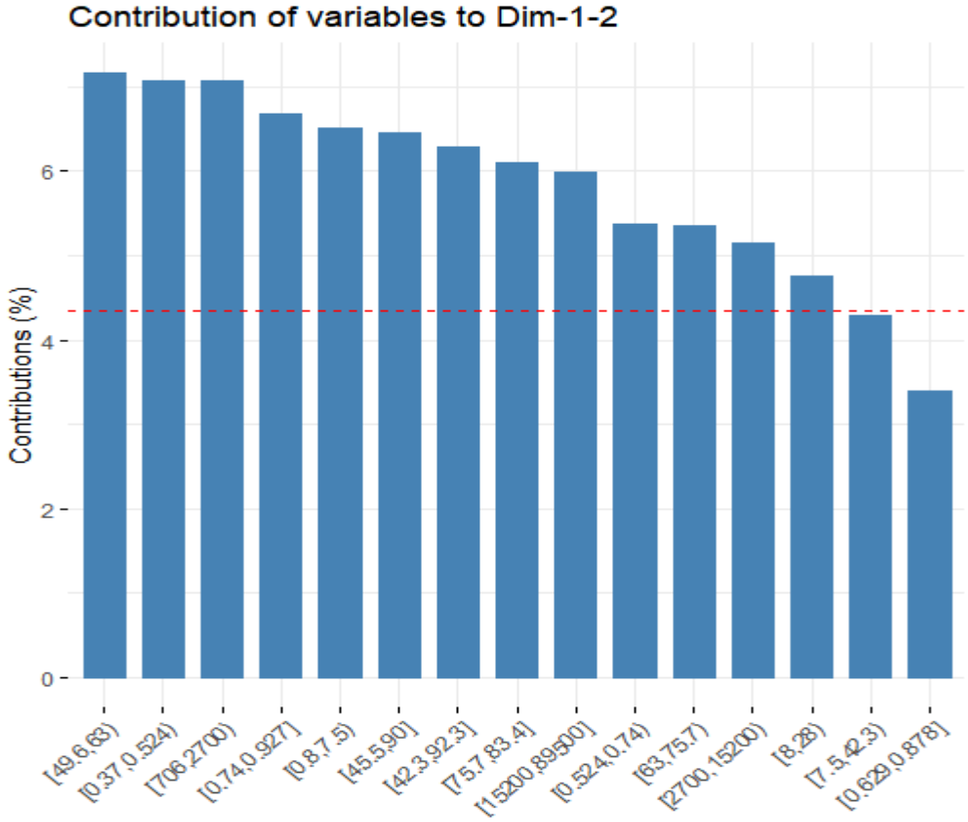


Figure 15 : Les variables les mieux représentées, cos2>0,85


Contribution of variables to Dim-2
Contributions (%)
10
5
0
[63,75.7]
[0.524,0.74)
[2700,15200)
[7.5,42.3)
[28,34)
[49.6,63)
[0.37,0.524)
[706,2700)
[0.8,7.5)
[0.372,0.55)
[8,28)
[0.55,0.629)
[45.5,90]
[15200,89500]
[0.74,0.927]


Contribution of variables to Dim-1
Contributions (%)
10.0
7.5
5.0
2.5
0.0
[0.74,0.927]
[42.3,92.3]
[45,590]
[75.7,83.4]
[15200,89500]
[706,2700)
[0.37,0.524)
[0.8,7.5)
[49.6,63)
[8,28)
[0.629,0.878]
[27.3,34.3)
[0.55,0.629)
[34.3,41.3)
[41.3,44.75)


Contribution of variables to Dim-1-2
Contributions (%)
6
4
2
0
[49.6,63)
[0.37,0.524)
[706,2700)
[0.74,0.927]
[0.8,7.5)
[45.5,90]
[42.3,92.3]
[75.7,83.4]
[15200,89500]
[0.524,0.74)
[63,75.7)
[2700,15200)
[8,28)
[7.5,42.3)
[0.629,0.878]