UNIVERSITÉ D'ANGERS

FACULTÉ DE DROIT, D'ÉCONOMIE ET DE GESTION

MASTER 1 IEE

Projet de Data Mining

Sujet : Les caractéristiques des pays du monde

Préparé : ADAYISSO Kokou

Remis le 13 Décembre 2019

TABLES DES MATIERES :

1 INTRODUCTION

Pendant longtemps secouée par de différentes crises socio-économiques, l'Afrique peine toujours à sortir de l'ornière ; on a souvent reproché aux décideurs publics africains de ne pas baser leur décision sur des réalités africaines. Dans l'optique d'aider les dirigeants à mieux adapter leur décision dans le contexte de leur pays en occurrence le cas d'Afrique, nous proposons une analyse pertinente axée sur une base de donnée tirée sur site **kaggle(www.kaggle.com)**.

Pour arriver à bout de l'analyse, nous nous sommes posé des questions :

Quelles sont les caractéristiques socio-économiques des pays ? Quelles caractéristiques socio-démo-économiques que l'Afrique partage en commun avec les autres continents ? Qu'est-ce qui différencie le continent africain des autres continents ?

L'approche de réponse à ces interrogations sera fournie à partir d'une analyse de donnée trouvée sur le site **Kaggle.** Cette base de donnée date depuis 2004**.** Ces données sont collectées par le gouvernement américain. La base de données brutes est manquante et ne contienne pas des variables telles que **Level-dev** (niveau de développement), **Eleccons** (niveau de consommation de l'électricité), **Elecprod** (niveau de production d'électricité) que nous avons créées et ajoutées nous-mêmes. Les valeurs ou les modalités de ces trois variables sont extraites dans une autre base de donnée sur le même site et collectée la même année par le gouvernement américain. Notre base de donnée est composée de 227 individus statistiques (les pays) et 22 variables dont 19 variables quantitatives et 3 variables qualitatives. Pour plus de détail sur la description des variables et des individus statistiques de la base de données, veuillez-vous référer à la partie annexes du dossier.

Nous allons nous contenter de faire les analyses avec cette base de donnée mais nous sommes toutefois conscients qu'il peut y avoir d'autres bases de données ou d'autres variables susceptibles de mieux expliquer les caractéristiques socio-économiques des pays.

Pour le traitement et l'analyse de ces données le logiciel R sera utilisé.

1.1 Difficultés :

Le traitement de données par le logiciel R n'est pas sans difficultés. Ces difficultés sont principalement liées à l'importation de donnée du fait des données manquantes dans la base et à la gestion des données manquantes du fait des modalités des variables qualitatives. Il va falloir donc utiliser une méthode non classique. Pour pallier à ce problème, nous avons dû passer par ces étapes ci-dessous :

-après l'ouverture R, on a installé le Package MissMDA

-importation de donnée via library(Rcmdr)

NB : Le script utilisé par la suite pour la gestion des données manquantes est consultable sur l'annexe

2 STATISTIQUES DESCRIPTIVES :

2.1 Cas de quelques variables

Pop	Areas	Literacy	Birthrate
Min. : 7026	Min. : 2	Min. : 17.60	Min. : 7.29
1st Qu. : 437624	1st Qu.: 4648	1st Qu.: 76.40	1st Qu. :12.72
Median: 4786994	Median: 86600	Median: 90.30	Median :18.90
Mean : 28740284	Mean : 598227	Mean : 82.84	Mean :22.11
3rd Qu.: 17497772	3rd Qu.: 441811	3rd Qu.: 97.80	3rd Qu. :29.77
Max. : 1313973713	Max. : 17075200	Max. :100.00	Max. :50.73

Le pays le moins peuplé de la base a une taille de population de 7026 et le plus peuplé se retrouve avec une population de taille de 1313973713(Chine). Au moins la moitié des pays de la base de donnée a une population inférieure ou égale à 4786994. En moyenne les pays de la base de donnée ont une taille de population estimée à 28740284.

De même, le pays le plus petit au monde a une superficie de 2 km carré (Monaco) et le plus vaste a une superficie de 17075200 km carré (Russie). Au moins la moitié des pays ont une superficie de 86600 km carré. Et les pays ont en moyenne une surface de 598227.

Le plus petit taux d'alphabétisation est de 17.60% (Niger)alors que le plus grand taux est de 100% (Danemark, Norway). Au moins 50% des pays ont un taux d'alphabétisation de 90.30% et le taux moyen est égal à 82.84%.

Le pays qui donne le moins de naissance a un taux de natalité de 7.29% et le pays où on donne plus d'enfants a un taux de natalité de 50,73%. En moyenne les pays de la base de donnée ont un taux de natalité qui est égale 22.11%.

2.2 Les relations entre les variables

Pearson corrélations : coefficients de corrélations							
	Agriculture	Arable	Areas	Birthrate	coastline	Crops	Deathrate
Agriculture	1.0000	-0.0341	-0.0504	0.6649	-0.0285	0.0514	0.3763
Birthrate	0.6649	-0.1834	-0.0664	1.0000	-0.0757	0.1179	0.3953
DGPpercap	-0.5707	0.0196	0.0722	-0.6488	0.0491	-0.2183	-0.2011
L.phone100	-0.5717	0.0596	0.0534	-0.7223	0.1520	-0.1495	-0.2597
Literacy	-0.5855	0.1019	0.0358	-0.7552	0.1137	0.0388	-0.3864
M.infant1000	0.6973	-0.1131	-0.0072	0.8450	-0.1365	-0.0612	0.6557
Other	0.0010	-0.8568	0.1389	0.0871	-0.0794	-0.5943	0.0616
Service	-0.6156	0.0886	-0.0541	-0.5436	0.1819	0.0520	-0.3516

Pairwise two-sided p-values : Test de significativité des coefficients de corrélation							
	Agriculture	Arable	Areas	Birthrate	coastline	Crops	Deathrate
Agriculture		0.6095	0.4500	<.0001	0.6695	0.4412	<.0001
Birthrate	<.0001	0.0056	0.3193		0.2559	0.0764	<.0001
DGPpercap	<.0001	0.7685	0.2789	<.0001	0.4616	0.0009	0.0023
L.phone1000	<.0001	0.3714	0.4237	<.0001	0.0219	0.0243	<.0001
Literacy	<.0001	0.1257	0.5917	<.0001	0.0876	0.5608	<.0001
M.infant1000	<.0001	0.0891	0.9146	<.0001	0.0399	0.3590	<.0001

Other	0.9886	<.0001	0.0365	0.1909	0.2334	<.0001	0.3554
Service	<.0001	0.1832	0.4174	<.0001	0.0060	0.4358	<.0001

2.2.1 Interprétations de la matrice de corrélations et de son test de validité :

L'interprétation va porter seulement sur quelques variables **:**

Les variables Agriculture et Birthrate sont positivement corrélés car leur coefficient de corrélation (0.6649), en plus d'être positif, il est également significativement différent de zéro du fait que p-value (<.0001) est inferieur 5%. Ce qui signifie que ces deux variables vont suivre la même direction dans le cercle de corrélations.

La variable service est faiblement et négativement corrélée à la variable Birthrate car le coefficient de corrélations (-0.5436) est négatif et n'est pas trop proche de 1 et de plus il est significativement différent de zéro car le P-value (<.0001) est inférieur à 5%. Les deux variables vont suivre des directions inverses.

Les variables Other et Arable sont fortement et négativement corrélés par ce que leur coefficient de corrélations (-0.8568) porte le signe négatif et est proche de 1 et en plus de cela le coefficient de corrélations est statistiquement significatif car la probabilité critique(P-value=<.0001) est inférieure à 5%. Ce qui implique les deux variables auront des directions opposées dans le cercle de corrélations.

Par contre on constate une absence de liaison linéaire entre les variables M. infant1000 et Areas par ce que leur coefficient de corrélation (-0.0072) est proche de zéro et n'est pas statistiquement significatif car P-value est supérieur à 5%.

NB : Dans la matrice de corrélations, il est globalement clair que les variables ne sont pas toutes corrélées positivement deux à deux. Il n'y aura donc pas le problème d'effet de taille.

3 ANALYSE DE DONNEES :ACP

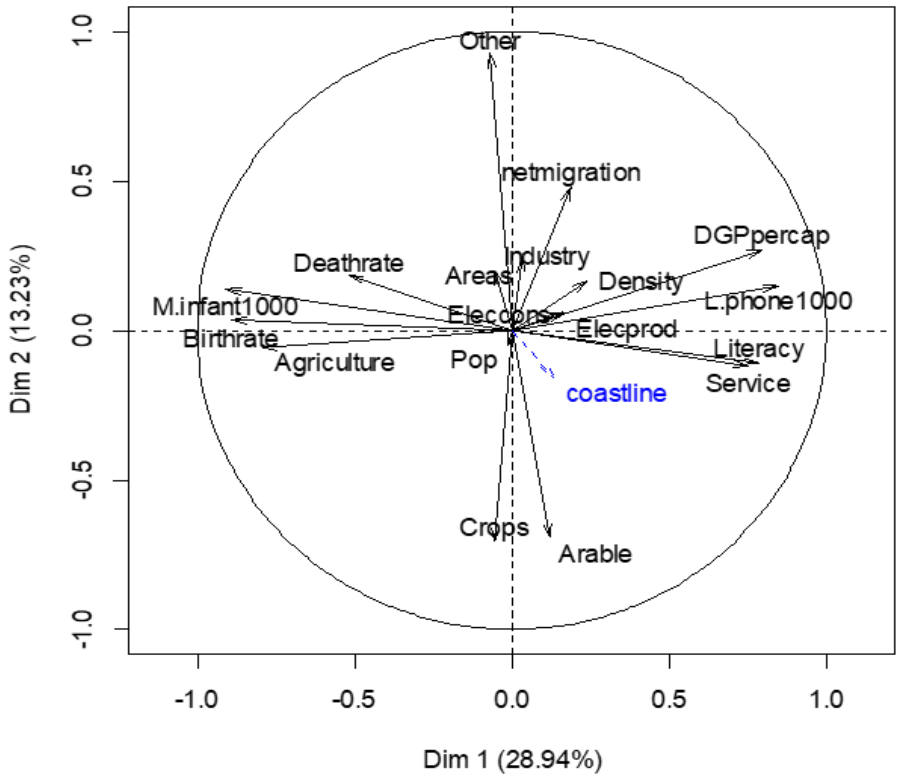
La méthode appliquée pour le traitement de cette base est Analyse des Correspondances Principale (ACP) du fait que nos variables d'intérêt sont principalement quantitatives en colonnes d'une part et d'autre part on a des individus (Pays) en ligne.

NB : Notre analyse sera faite sur une base de données centrée et réduite dans le but de donner le même poids à chaque variable et individu. Le logiciel R le fait automatiquement.

Cependant on a décidé de mettre la variable Coastline en supplémentaire car cela n'a pas grand intérêt sur le traitement de notre problématique et de plus on n'a pas parvenu à trouver sa vraie description dans la base de données.

3.1 Interprétation des graphiques

3.1.1 Cercle de corrélation :



Sur ce graphique nous avons deux axes qui explique globalement 42,16% l'inertie de la base dont l'axe 1 explique à lui seul 13,23% l'inertie et l'axe 2 explique 28.94%.

Sur l'axe1 nous avons une opposition entre les variables mortalité infantile, taux de natalité, la population agricole, taux de mortalité à gauche et des variables PIB par tête, nombre de ligne téléphoniques, taux d'alphabétisation, la population dans le secteur service à droite. On peut donc dire plus le taux de mortalité, le taux de mortalité, la mortalité infantile (caractéristiques des pays situés plus à gauche) sont élevés plus faible seront le taux d'alphabétisation, le nombre de ligne téléphoniques, PIB par tête (caractéristiques des pays situés plus à droite).

Sur l'axe2 oppose principalement les variables pourcentage des terres des autres cultures, l'immigration nette (partie supérieure) aux variables pourcentage des terres des cultures vivrières et pourcentage des terres de culture industrielles (partie inferieure). On peut dire par exemple que plus un pays a un pourcentage de terre des autres cultures plus élevé moins élevé sera son pourcentage des terres de cultures vivrières.

3.1.2 Graphique des individus :

Il n'est pas sorti. Passons donc à l'interprétation des sorties des résultats.

3.2 Interprétation des résultats chiffrés:

3.2.1 Valeurs propres : Eigenvalues

Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6

Eigenvalues	5.210	2.381	2.012	1.668	1.457	1.170
Variance % of var.	28.944	13.228	11.178	9.267	8.093	6.501
Cumul % of var.	28.944	42.172	53.350	62.617	70.709	77.210

D'après la règle de Kaiser, on retient les axes qui ont les valeurs propres supérieures à 1.

Dans notre cas ici 6 axes (Dim1 jusqu'à Dim6) sont ceux dont les valeurs propres sont supérieures à 1. Ces six axes expliquent 77.21% l'information.

3.2.2 Variables quantitatives

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Pop	-0.016	0.005	0.000	-0.051	0.109	0.003	0.212	2.239	0.045
Areas	0.035	0.023	0.001	0.243	2.480	0.059	0.145	1.048	0.021
Density	0.241	1.115	0.058	0.166	1 .163	0.028	-0.179	1.590	0.032
netmigration	0.188	0.677	0.035	0.479	9.628	0.229	0.003	0.001	0.000
M.infant1000	-0.908	15.834	0.825	0.138	0.801	0.019	0.054	0.146	0.003
DGPpercap	0.794	12.104	0.631	0.270	3.055	0.073	0.008	0.003	0.000
Literacy	0.782	11.737	0.611	-0.110	0.511	0.012	0.000	0.000	0.000
L.phone1000	0.846	13.726	0.715	0.149	0.932	0.022	-0.076	0.285	0.006
Arable	0.121	0.280	0.015	-0.692	20.120	0.479	0.166	1.376	0.028
Crops	-0.057	0.063	0.003	-0.708	21.031	0.501	-0.043	0.091	0.002

3.2.2.1 Contributions

Les variables qui contribuent plus à la définition de dl'axe1 sont la mortalité infantile, nombre de ligne téléphoniques, PIB par tête et le taux d'alphabétisation. Pour l'axe2 c'est la proportion des terres de cultures industrielles, vivrières et l'immigration nette. Quant à l'axe3, on peut citer population, densité, superficie et la proportion des terres de cultures vivrières. Raison pour laquelle la longueur des vecteurs qui représentent ces variables est la plus longue possible dans le cercle de corrélation.

3.2.2.2 Coordonnées :

Le long de l'axe1

On remarque une opposition entre la variable mortalité infantile de coordonnée négative et les variables nombre de ligne téléphoniques, PIB par tête et le taux d'alphabétisation de coordonnées positives.

Le long de l'axe2.

Par contre sur l'axe2, on a une opposition entre les variables proportion des terres de cultures vivrières et industrielles qui sont de coordonnées négatives et l'immigration nette qui est de coordonnée positive.

Le long de l'axe3

Ensuite l'axe3 oppose les variables population, superficie, la proportion des terres de cultures vivrières de coordonnées positives à la variable densité de coordonnée négative.

3.2.3 Variables catégorielles

$Dim.1$category		
	Estimate	p.value
Region=WESTERN EUROPE	2.3622675	6.741310e-14
Climate=H subtrop&cont	1.0554167	3.505797e-09
Region=LATIN AMER. & CARIB	0.3266312	6.521313e-03
Region=NORTHERN AMERICA	2.1769460	7.862190e-03
Climate=dry tropical	-1.6714686	2.013846e-03
Climate=wet tropical	-1.2073255	7.400335e-05
Region=SUB-SAHARAN AFRICA	-3.2914083	1.406940e-29

Les régions et le type de climat qui contribuent plus et positivement à la définition de l'axe1 sont l'Europe de l'Ouest, Amérique latine et Caribes, Amérique du Nord.

3.2.4 Les individus

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Afghanistan	7.644	-4.971	2.090	0.423	2.513	1.168	0.108	0.159	0.006	0.000
Albania	2.179	-0.041	0.000	0.000	-1.349	0.337	0.384	-0.412	0.037	0.036
Algeria	3.702	-0.764	0.049	0.043	1.486	0.409	0.161	0.258	0.015	0.005
AmSamoa	4.805	0.367	0.011	0.006	-2.376	1.044	0.244	-0.586	0.075	0.015
Andorra	3.132	2.061	0.359	0.433	1.648	0.502	0.277	-0.358	0.028	0.013
Angola	7.343	-5.358	2.427	0.532	2.499	1.155	0.116	0.696	0.106	0.009
Anguilla	3.692	1.976	0.330	0.286	1.743	0.562	0.223	-0.811	0.144	0.048
Anti&Bar	2.707	1.738	0.256	0.412	-0.916	0.155	0.115	-0.677	0.100	0.063
Argentina	1.924	0.874	0.065	0.206	0.647	0.078	0.113	0.232	0.012	0.015
Armenia	2.407	-0.127	0.001	0.003	-0.710	0.093	0.087	-0.150	0.005	0.004

Ces pays sont ceux qui contribuent plus et de manière significative à la définition des axes.

3.2.4.1 Contributions et coordonnées :

Les individus qui contribuent le plus à définir l'axe1 sont essentiellement Afghanistan et Angola mais cependant il n'y a pas d'opposition entre ces deux pays le long de cet axe.

Par contre sur l'axe2, les pays qui contribuent plus à sa définition sont Afghanistan, American Samoa et Angola et on remarque une opposition entre Afghanistan et American Samoa le long de cet axe.

3.2.5 Fonction Dimdesc :

3.2.5.1 Cas des variables quantitatives

$Dim.1		
$Dim.1$quanti		
	Corrélation	p. value
L.phone1000	0.8456513	2.791503e-63
DGPpercap	0.7941227	1.457745e-50
Literacy	0.7819669	4.374705e-48
Service	0.7502084	2.653442e-42
Density	0.2409670	2.476880e-04
netmigration	0.1877449	4.535033e-03
Elecprod	0.1657489	1.239152e-02
Eleccons	0.1634741	1.366314e-02
coastline	0.1327233	4.577255e-02
Deathrate	-0.5196642	4.217305e-17
Agriculture	-0.7936112	1.867537e-50
Birthrate	-0.8890647	2.654921e-78
M.infant1000	-0.9082533	4.263397e-87

Les variables qui définissent plus et positivement l'axe 1 sont : nombre de ligne téléphoniques, PIB par tête, taux d'alphabétisation, population dans le secteur de service, densité, immigration nette, la production électrique, la consommation électrique et celles qui le définissent plus et négativement sont taux de mortalité, la proportion de la population agricole, taux de natalité et mortalité infantile. Ces deux groupes de variables s'opposent le long de l'axe1.Par exemple les pays qui ont un nombre de ligne téléphoniques, PIB, taux d'alphabétisation élevés s'oppose au pays qui ont un taux de mortalité, taux de natalité une population agricole élevés.

3.2.5.2 Cas des variables qualitatives

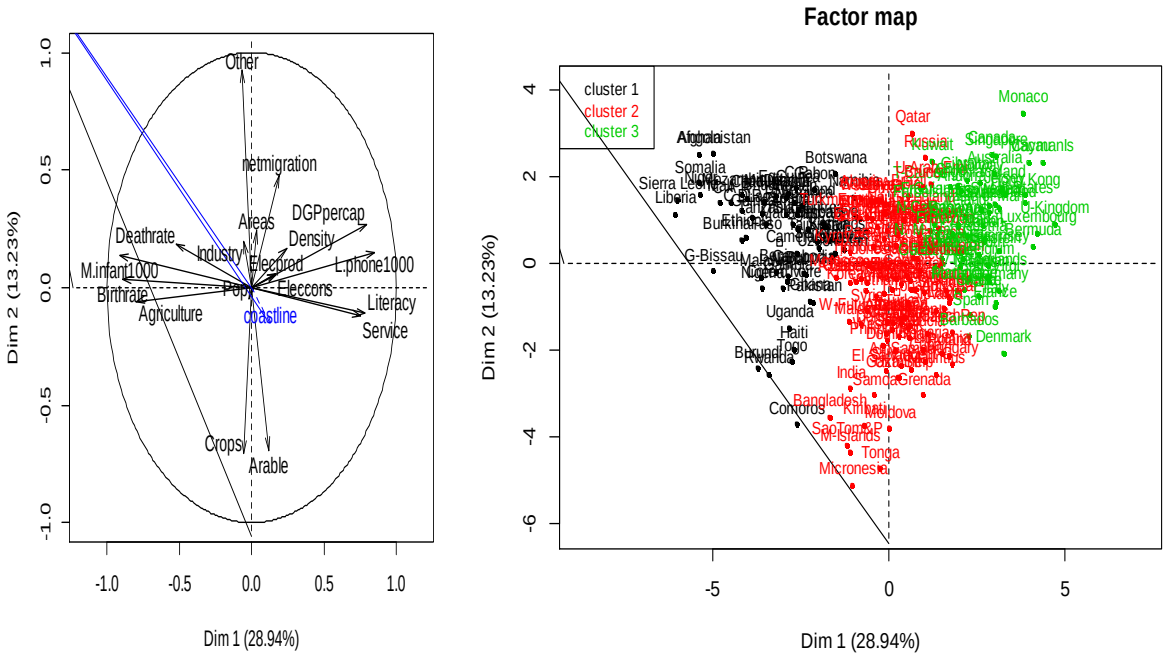
$Dim.1$quali

	R2	p. value
Région	0.6124152	3.126703e-39
Climate	0.2035573	1.027994e-09

Pour les variables qualitatives, l'axe1 et les variables régions et climat sont dépendants au seuil de 5%.

4 CLASSIFICATION HIERARCHIQUE

4.1 Graphique de cercle de corrélation superposé au graphe des clusters



Graphiquement on a trois type de clusters (groupe de pays) :

-Cluster1 en noir des pays d'Afrique caractérisés par un taux de mortalité, taux de natalité, mortalité infantile et une population agricole très élevés.

-Cluster2 en rouge des pays composé principalement des pays d'Asie caractérisé la variable migration nette, la proportion de terres de cultures vivrières, industrielles et autres cultures.

-Cluster3 en vert des pays occidentaux caractérisés par contre par un PIB par tête un taux d'alphabétisation, une population dans le secteur de service et nombre de ligne téléphonique pour 1000 mille habitant élevés.

NB : Malgré que Afghanistan soit un pays d'Asie se trouve dans le cluster1 par ce qu'il a la variable taux de mortalité très élevé et qui le rapproche plus au pays d'Afrique qui sont caractérisés par cette variable.

4.2 Interprétation des sorties de clusters :

Ici c'est seulement les variables qui sortent de manière significative qui apparaissent c'est à dire des variables qui sont significatives au seuil de 5%.

4.2.1 Lien entre les variables qualitatives et les clusters

	p.value	df
Region	3.649036e-36	20
Climate	3.299465e-06	10

C'est seulement les variables climat et régions qui sont dépendantes des clusters au seuil de 5% (P-value <5%). Ceux qui veut dire que si j'ai deux pays qui sont de régions différentes, la probabilité qu'ils soient dans de clusters différents est élevée. Par exemple, si je prends le Nigeria et la France, la probabilité que je retrouve ces deux pays ensemble dans le cluster1 est faible.

Par contre la dernière variable qualitative « niveau de développement » n'est pas significativement dépendante des clusters c'est pourquoi elle n'apparait pas ici.

4.2.1.1 Description des clusters et des modalités des variables qualitatives

Cluster1					
$`1`					
	Cla/Mod	**Mod/Cla**	**Global**	**p.value**	**v.test**
Region=SUB-SAHARAN AFRICA	86.274510	74.576271	22.466960	1.288079e-26	10.6781
Climate=wet tropical	34.210526	66.101695	50.220264	4.789622e-03	2.820853
Climate=dry tropical	40.540541	25.423729	16.299559	3.531156e-02	2.104767
Region=EASTERN EUROPE	0.000000	0.000000	5.286344	2.427796e-02	-2.252702
Region=WESTERN EUROPE	0.000000	0.000000	12.334802	1.145277e-04	-3.857553
Climate=H subtrop&cont	4.166667	3.389831	21.145374	2.024134e-05	-4.262212
Region=LATIN AMER. & CARIB	2.222222	1.694915	19.823789	4.987288e-06	-4.565322

86,27% des pays de SUB-SAHARA Africains sont dans le cluster1.Il y a également 2.22% des pays de l'Amérique latine et Caraïbes dans le culster1.Le cluster1 est constitué à 74,58% des pays de SUB-SAHARA Africain et à 1 ,7% des pays de l'Amérique latine et Caraïbes. Il n'a pas de pays de l'Europe de l'Est, ni de l'Ouest dans le cluster1. La base de données est composée de 22,47% des pays de SUB-SAHARA Africain, de 19,82% des pays de l'Amérique latine et Caraïbes, de 12,33% de pays de l'Europe de l'Ouest, de 5,29% pays de l'Europe de l'Est.

34,21% des pays de climat tropical humide sont dans le cluster1, et 40,54% des pays de climat tropical sec dans le cluster1 et enfin 4,17% des pays de climat subtropical humide tempéré et continental dans le cluster1. Le cluster1 est par contre constitué à 66,10% des pays de climat tropical humide, à 25,42% des pays de climat tropical sec et à 3,4% des pays de climat subtropical humide tempéré et continental. La base de données est composée de 50,22% de pays à climat tropical humide, de 16,3% des pays à climat tropical sec, à 19,82% climat subtropical humide tempéré et continental.

Cluster2

$2`	Cla/Mod	Mod/Cla	Global	p.value	v.test
Region=LATIN AMER. & CARIB	68.88889	28.971963	19.823789	0.001195615113136	3.239924
Region=EASTERN EUROPE	91.66667	10.280374	5.286344	0.001449347388448	3.184639
Region=NORTHERN AFRICA	100.00000	5.607477	2.643172	0.010165459337687	2.570150
Region=C.W. OF IND. STATES	83.33333	9.345794	5.286344	0.011357029884675	2.531517
Region=WESTERN EUROPE	0.00000	0.000000	12.334802	0.000000003282487	-5.916808
Region=SUB-SAHARAN AFRICA	11.76471	5.607477	22.466960	0.000000002336907	-5.972468

Il n'y a que des modalités des variables qualitatives régions qui sortent de manière significatives pour le cluster2.

On remarque que 100% des pays de l'Afrique du Nord se trouve dans le cluster2 mais ne représentent que 5,6% des pays de cluster2 et 2.64% des pays de la base.

Par contre 11,76% des pays de SUB-SAHARA Africain se retrouve dans le cluster2 mais qui ne représentent que 5.6% de l'ensemble des pays du cluster2.

La majorité (91.67%) des pays de l'Europe de l'est sont dans le cluster2 et représentent 28,97% des pays du cluster2.

Dans le cluster2 il n'y a pas de pays de l'Europe de l'Ouest et ils se retrouvent tous dans le cluster3

Cluster3					
$`3`					
	Cla/Mod	Mod/Cla	Global	p.value	v.test
Region=WESTERN EUROPE	100.000000	45.901639	12.334802	3.567740e-19	8.949600
Climate=H subtrop&cont	54.166667	42.622951	21.145374	5.557960e-06	4.542543
Region=C.W. OF IND. STATES	0.000000	0.000000	5.286344	2.092391e-02	-2.309355
Climate=wet tropical	19.298246	36.065574	50.220264	1.024071e-02	-2.567594
Climate=dry tropical	5.405405	3.278689	16.299559	4.877553e-04	-3.487392
Region=SUB-SAHARAN AFRICA	1.960784	1.639344	22.466960	2.717236e-07	-5.142073

On peut dire que 100% des pays de l'Europe de l'Ouest sont dans le cluster3 et représentent 45,49% des pays de ce cluster. Par contre les pays de SUB-SAHARA Africain (1,96%)sont minoritaires dans ce cluster et représentent environ 1,64% des pays du cluster.

Par ailleurs, 54,17% des pays à climat méditerranéen sort de manière significative dans ce cluster et représente approximativement 42,62% de l'ensemble des pays qui le compose.

4.2.2 Lien entre les clusters et les variables quantitatives : Indice de corrélation entre clusters et les variables quantitatives

	Eta2	P-value
M.infant1000	0.73276511	6.486859e-65
Birthrate	0.67764968	8.565412e-56
L.phone1000	0.67137470	7.421848e-55
DGPpercap	0.66738355	2.868778e-54

Literacy	0.51642778	4.567418e-36
Agriculture	0.48584228	4.393318e-33
Service	0.46680911	2.575521e-31
Deathrate	0.39862748	1.836882e-25
netmigration	0.22198820	6.172375e-13
Crops	0.08851308	3.104914e-05
Other	0.07257638	2.163481e-04
Industry	0.06901186	3.324639e-04
Density	0.05560455	1.649056e-03
coastline	0.03419551	2.030452e-02

Cette sortie nous donne des variables quantitatives qui sortent de façon significative et participent plus à la construction des clusters. C'est des variables qui permettent beaucoup plus en fait de séparer des pays. Ces variables sont classées selon le pouvoir contributif décroissant sont : mortalité infantile, taux de natalité, le nombre de lignes téléphoniques pour 1000 habitants, PIB par tête, taux de d'alphabétisation population agricole, population dans le secteur service, taux de mortalité migration nette, proportion de terre de cultures industrielles et proportion de terres d'autres culture, la population dans le secteur industriel, densité de la population.

4.2.2.1 Description des clusters par les variables quantitatives :

Cluster1						
$`1`						
	v.test	**Mean in category**	**Overall mean**	**sd in category**	**Overall sd**	**p.value**
M.infant1000	12.479189	84.6406780	35.5069643	28.9882	35.076	9.696964e-36
Birthrate	11.740724	36.7137288	22.1147321	7.2440788	11.0778	7.880868e-32
Agriculture	9.987317	0.3095058	0.1508443	0.1612512	0.141529	1.732065e-23
Deathrate	9.475611	14.4900000	9.2413453	5.8584156	4.9347	2.652051e-21
Other	2.455693	86.0579375	81.6383111	14.0715613	16.033	1.406132e-02
Crops	-2.067634	2.6365122	4.5642222	4.5749880	8.3060	3.867449e-02
coastline	-2.354820	2.1005085	21.1653304	7.4417283	72.1274	1.853169e-02
DGPpercap	-7.089026	1728.8135	9689.8230	1517.47458	10004.77	1.350593e-12
Service	-8.229844	0.4175811	0.5652830	0.1257295	0.1598896	1.874572e-16
L.phone1000	-8.679805	16.3932203	236.06143	16.47495	225.4669243	3.964482e-18
Literacy	-10.542472	60.4979077	82.8382775	19.0578322	18.8787616	5.503320e-26

Dans le cluster1 on a deux types de pays :

- Ceux qui sont ,d'une part, surreprésentés (V-test positif) c'est à dire des pays dont les caractéristiques moyennes dépassent les caractéristiques moyennes de l'ensembles des pays de la base. Nous pouvons dire les pays de cluster1 ont en moyenne une mortalité infantile pour 1000 naissance égale à 84,64

alors que l'ensemble des pays de la base de données a en moyenne une mortalité infantile pour 1000 naissance qui est égale à 35,51. Les pays du cluster1 ont une mortalité infantile pour 1000 naissance plus élevée que la moyenne des pays de la base. On peut également dire Les pays du cluster1 ont un taux de natalité moyenne estimé à 36,71% pendant que l'ensemble des pays de la base de données a en moyenne un taux de natalité évalué à 22,11%. Par conséquent, les pays du cluster1 possède un taux de natalité supérieure à celui de l'ensembles des pays du monde. On peut aussi interpréter que les pays du cluster1 ont une population agricole moyenne égale à 0,31% pendant que la population agricole moyenne de l'ensemble des pays du monde est de 0,15%. Les pays du cluster1 ont alors une population plus agricole par rapport à la moyenne des pays du monde.

-Et ceux qui sont d'autre part sous-représentés (V-test négatif) c'est-à-dire des pays qui ont de caractéristiques moyennes surpassées par les caractéristiques moyennes de l'ensembles des pays du monde. On constate que les pays du cluster1 ont en moyenne un pourcentage de terre de culture industrielles égal à 2,64% contre 4,56% pour l'ensemble des pays du monde en moyenne. En moyennes les terres des pays du cluster1 sont donc moins favorables aux cultures industrielles que la moyenne des pays du monde. Les pays du cluster1 ont un PIB par tête moyen égal à 1728,81$ américain alors qu'en moyenne l'ensemble des pays du monde ont un PIB par tête égal à 9689,82$ américain. Les pays de cluster1 sont en moyenne moins riches que la moyenne des pays du monde. Bref ils sont les plus pauvres de la planète. Ensuite Les pays de ce cluster ont en moyenne un taux de population dans le secteur service égale à 0,42% alors que le taux moyen de la population dans le secteur service pour l'ensemble des pays est estimé à 0,57%. Les pays du cluster1 sont donc des pays à faible concentration de la population dans le secteur service que la moyenne de la population du monde. Les pays de cluster1 possèdent un taux d'alphabétisation moyen estimé à 60.50% pendant que l'ensemble des pays du monde ont en moyenne des taux d'alphabétisation égale à 82,84%. Donc les pays du cluster1 sont moins éduqué que la moyenne des pays du monde.

NB : les pays du cluster1 sont majoritairement des pays de SUB-SAHARA Africain. Ils auront approximativement ces caractéristiques ci-dessus.

Cluster2						
$`2`						
	v.test	**Mean in category**	**Overall mean**	**sd in category**	**Overall sd**	**p.value**
Crops	4.443893	7.1643925	4.5642222	10.836004	8.306	0.0000088
Literacy	3.943718	88.0830166	82.8382775	12.19614	18.8787	0.00008
Industry	3.590558	0.3163495	0.2827109	0.11446133	0.13299	0.0003
Arable	2.154525	15.7631776	13.7971111	14.422889	12.95	0.0311
L.phone1000	-2.386457	198.15779	236.0614	128.22189	225.46	0.017
Agriculture	-2.960997	0.1213234	0.1508443	0.0763297	0.1415	0.00306
Birthrate	-3.327844	19.5177962	22.1147321	6.919555	11.0778	0.00087
Other	-4.043262	77.0714953	81.6383111	17.60473	16.0338	0.00005
DGPpercap	-4.344468	6627.9422	9689.823	4853.126	10004.77	0.00001
M.infant1000	-4.3789	24.6868308	35.5069643	14.95830	35.076	0.0000119
Deathrate	-5.758397	7.2395830	9.2413453	3.04587451	4.9347639	0.0000000084
Netmigration	-6.084709	-2.0390245	0.0381250	4.65940152	4.8460005	0.00000000

1

En se servant la même logique utilisée au niveau de cluster1, nous remarquons ici que les pays de cluster2 ont en moyenne des taux de mortalité (7.24%), de natalité (19.52%), de population agricole (0.12%) en dessous de la moyenne de l'ensembles des pays du monde contrairement aux pays de cluster1.

Les pays de ce cluster sont moins riches que la moyenne de l'ensemble des pays du monde mais par contre ils sont plus riches que les pays du clusters1 car leur PIB par tête moyen est de 6627.94 $ américain contre 1728.81$ américain pour les pays de cluster1.

Les pays du cluster2 sont plus éduqués que la moyenne de l'ensemble des pays de base de données. En moyenne, ils ont une proportion de population dans le secteur industriel plus élevé que la moyenne des pays de la base. Et ils sont plus éduqués que les pays du cluster1.

Ils émigrent plus que la moyenne de l'ensemble des pays du monde. En moyenne, ils quittent plus leur pays plus qu'ils n'en reçoivent d'étrangers.

NB : Caractéristiques sont principalement pour les pays de l'Europe de l'Est et de l'Afrique du Nord, et de l'Amérique Latine et Caraïbes.

Cluster3					
$`3`					
	v.test	**Mean in category**	**Overall mean**	**sd in category**	**Overall sd**	**p.value**
DGPpercap	11.9058	2.27606e+04	9.689e+03	8.9640e+03	1.0004e+04	1.103e-32
L.phone1000	11.27	5.1501e+02	2.3606e+02	1.818e+02	2.254e+02	1.745e-29
Service	8.69	7.1778e-01	5.65e-01	1.184e-01	1.598e-01	3.566e-18
Netmigration	6.45	3.4711e+00	3.8125e-02	4.375634e+00	4.846e+00	1.075618e-10
Literacy	5.98	9.524e+01	8.28e+01	5.399e+00	1.887e+01	2.1036e-09
Density	3.5234	1.0195e+03	3.790e+02	3.053e+03	1.656e+03	4.260e-04
Eleccons	2.41	1.520432e+05	7.268e+04	5.4647e+05	2.9994e+05	1.59e-02
Elecprod	2.39	1.620275e+11	7.855e+10	5.751322e+11	3.169506e+11	1.63e-02
coastline	2.22	3.877e+01	2.116e+01	6.984e+01	7.212e+01	2.609e-02
Other	2.12	8.537e+01	8.163e+01	1.2332e+01	1.603382e+01	3.372273e-02
Deathrate	-2.89	7.67e+00	9.241e+00	2.242e+00	4.934e+00	3.844e-03
Crops	-2.95	1.86e+00	4.5642e+00	2.79e+00	8.306e+00	3.091e-03
Industry	-3.40	2.33e-01	2.827e-01	1.086105e-01	1.329939e-01	6.646e-04
Agriculture	-6.54	4.916e-02	1.5084e-01	5.4372e-02	1.415299e-01	5.872690e-11
M.infant1000	-7.41	6.96e+00	3.55e+01	4.279e+00	3.5076e+01	1.210358e-13
Birthrate	-7.86	1.254e+01	2.211e+01	3.5306e+00	1.107e+01	3.5854e-15

Les pays du cluster3 sont les plus riche de la planète car ils ont en moyenne 22761$ américain comme PIB par tête contre 6627.94$ américain pour les pays du cluster2. Ils ont en moyenne un PIB par tête supérieure par rapport à la moyenne de l'ensemble des pays du monde (9689,82$). Ils ont le nombre de lignes téléphoniques pour 1000 habitant le plus élevé au monde car ils sont au-dessus de la moyenne des pays de base pour cette variable. Ils possèdent également la migration nette, une proportion de population dans le secteur service, la production électrique, consommation électrique les plus élevés au monde. Ils sont les plus éduqués au monde car en moyenne les pays de ce cluster ont un taux d'alphabétisation de 95,24% contre 88.08% des pays du cluster2.

Contrairement aux pays de cluster1, les pays de cluster3 ont le taux de mortalité (7.67%), de natalité (12,55%), mortalité infantile pour 1000 naissances (6,96), proportion de population industrielle (0,23%) et agricole (0,049%) et un pourcentage de terre de cultures industrielles (1,87%) les plus faibles.

NB : Ces caractéristiques correspondent approximativement aux pays de l'Europe de l'Ouest car le cluster3 est constitué essentiellement des pays de l'Europe de l'Ouest.

4.2.3 Les individus parangons :

Les individus parangons sont des individus qui ont les caractéristiques les plus proches de caractéristiques moyennes d'un cluster c'est-à-dire les individus les plus représentatifs d'un cluster.

$para

Cluster : 1

Zambia	Tanzania	Sudan	Madagascar	Nepal
0.6603073	0.7041119	0.7543129	0.9459742	0.9712305

Pour le cluster1, les pays les plus représentatifs ou bien les pays les plus proches du barycentre du cluster1 sont Zambie, Tanzanie, Soudan, Madagascar, Nepal

Cluster : 2

Macedonian	Montserrat	Armenia	Bos&Herz	Tunisia
0.5775304	0.6238328	0.6915792	0.7750285	0.9191934

Macédoine, Montserrat, Arménie, Bosnie & Herzégovine et Tunisie sont les pays les plus représentatifs du cluster2.

Cluster: 3

Austria	Belgium	Sweden	Japan	N-Zealand
0.6509251	0.9202906	1.0013722	1.0334714	1.0700018

Les individus parangons du cluster3 sont Australie, Belgique, Suède, Japon et Nouvelle-Zélande.

4.2.4 Les individus spécifiques :

Les individus spécifiques sont des individus les plus éloigné du centre de gravité des autres clusters c'est-à-dire les individus les plus éloignés des individus les plus représentatifs des autres clusters

$dist

Cluster : 1

Afghanistan	Liberia	Somalia	Sierra Leone	Niger
7.445399	7.235655	6.892419	6.888514	6.551263

Pour le cluster1, les pays les plus éloignés des pays les plus représentatifs des autres clusters sont :

Afghanistan, Liberia, Somalie, Sierra Leone et Niger.

Cluster: 2

China	India	Micronesia	Russia	Tonga
10.586558	8.767595	6.884691	6.693177	6.137795

Chine, Inde, Micronésie, Russie et Tongo sont les pays les plus éloignés des pays qui ont des caractéristiques moyennes des autres clusters.

Cluster: 3

U-Kingdom	Chile	Monaco	Macau	CaymanIs
17.409484	10.063442	7.063124	6.854802	6.406226

Les individus spécifiques du cluster3 sont : Royaumes unis, Chili, Monaco, Macau et Cayman Islands.

5 CONCLUSION

L'analyse des données nous a montré qu'on peut catégoriser les pays du monde du monde en trois groupes selon des caractéristiques socio-démo-économiques définies par les variables : Cluster1, cluster2 et cluster3.

Cependant le cluster1 est constitué majoritairement des pays de SUB-SAHARA Africain caractérisés par un taux de mortalité infantile, un taux de mortalité, un taux de natalité, une proportion de population agricole et une part de terre d'autres cultures élevés contre un taux d'alphabétisation, un nombre de ligne téléphoniques pour 1000 habitants, un PIB par tête, une population dans le secteur service et une part de terres de cultures industrielles relativement faible par rapport à l'ensemble des pays du monde. On peut retrouver quelques pays de l'Amérique Latine et Caraïbes dans ce groupe mais qui reste minoritaires. Et les pays qui représentent mieux ce groupe sont Zambie, Tanzanie, Soudan, Madagascar, Népal.

Ensuite les pays de l'Amérique Latine et Caraïbes, de l'Europe de l'Est et de l'Afrique du Nord sont majoritairement représentés dans le cluster2 et sont définis par un taux d'alphabétisation moyennement élevé, une proportion de terre de cultures industrielles et vivrières, une population dans l'industrie plus élevés contre un taux de mortalité infantile, et de natalité, un PIB par tête, un nombre de lignes téléphoniques et une population dans le secteur agricole moyennement faible.

Et Macédoine, Montserrat, Arménie, Bosnie & Herzégovine et Tunisie sont les pays les plus représentatifs de ce cluster.

Enfin le culster3 est essentiellement constitués des pays de l'Europe de l'Ouest et sont caractérisés par un taux de d'alphabétisation, PIB par tête, une migration nette, une proportion de population dans le secteur service, la production électrique, consommation électrique et un nombre de lignes téléphoniques les plus élevés au monde contre un taux de mortalité , de natalité, mortalité infantile pour 1000 naissances , proportion de population industrielle et agricole et un pourcentage de terre de cultures industrielles les plus faibles. Les pays les plus représentatifs de ce groupe sont Royaumes unis, Chili, Monaco, Macau et Cayman Islands.

Par ailleurs on remarque une absence des pays de l'Asie dans l'analyse est dû au fait qu'ils ne contribuent de manière significative à la construction de ces trois clusters.

D'après les résultats de cette analyse il est clair que les pays de l'Afrique surtout ceux des pays de SUB-SAHARA Africain n'ont pas assez de points communs avec les pays occidentaux sur le plan socio-démo-économique par contre ils s'opposent. Il est donc judicieux que ces pays trouvent et définissent des modèles et des politiques qui marchent et qui s'adaptent à la réalité de leur pays.

6 ANNEXE

6.1 Script de gestion de données manquantes :

```
# Chargement des packages

library(missMDA)

missingdata <- country[,c(-1,-14, -20)]

# Gestion des données manquantes (à la fois quantitative et qualitative)
```

- *Visualiser les données manquantes*

```
nb <- estim_ncpPCA(missingdata, scale=TRUE)

nb
```

- *Imputation multiple*

```
comp <- imputePCA (missingdata , ncp=nb$ncp, scale=TRUE)

comp$completeObs

data <- merge(country[,c(1,14,20)], comp$completeObs, by="row. names", all.x=TRUE, all.y=TRUE)

str(data)
```

6.2 Description des variables et des individus

Variables	
Pop	population
Areas	Superficie en Km2
Density	Densité de la population
Coastline	
Netmigration	immigration nette
M-infant1000	mortalité infantile pour 1000 naissances
DGPpercap	PIB par tête
Literacy	taux d'alphabétisation
L-phones1000	le nombre de lignes téléphoniques pour 1000 habitants
Arable	pourcentage des terres de cultures vivrières(riz, maïs, blé)
Crops	pourcentage des terres de cultures industrielles(agrumes, café, caoutchouc)
Other	pourcentage des terres d'autres cultures
Birthrate	taux de natalité
Deathrate	taux de mortalité
Agricuture	proportion de population agricole
Industrie	proportion de population dans le secteur industriel
Service	proportion de population dans le secteur service
Level dev	niveau de développement
Elecconso	quantité d'électricité consommée en KWh
Elecprod	quantité d'électricité produite en KWh
Climate	type de climat
1, Dry tropical	=climat tropical et sec
2, Wet tropical	=climat tropical humide.
3, H subtro&cont	= climat subtropical humide tempéré et continental
4, DH Sum&wint wint	= climat méditerraneen

La liste des pays recodés

COUNTRY	
American Samoa	AmSamoa
Antigua & Barbuda	Anti&Bar
Bahamas, The	Bahamas
Bosnia & Herzegovina	Bos&Herz
British Virgin Is	BritisVirgi
Cayman Islands	CaymanIs
Central African Rep	C A R
Congo, Dem. Rep.	C D R
Congo, Repub. of the	C R
Czech Republic	CzechRep
Dominican Republic	DominikR
Equatorial Guinea	EqGuinea
French Guiana	FrGuiana
French Polynesia	FrPolynesia
Gambia, The	Gambia
Guinea-Bissau	G-Bissau
Marshall Islands	M-Islands
Micronesia, Fed. St.	Micronesia
Netherlands Antilles	N-Antilles
New Caledonia	N-Caledonia
New Zealand	N-Zealand
N. Mariana Islands	N. M-Islands
Papua New Guinea	PN-Guinea
Saint Kitts & Nevis	St-Kit&N
St Pierre & Miquelon	St-PMiquelon
St VincentGrenadines	St-Vincent-G
Sao Tome & Principe	SaoTom&P
Solomon Islands	S-Islands
Trinidad & Tobago	T-Tobago
Turks & Caicos Is	T-Caicos
United Arab Emirates	U- Arab-Emi
United Kingdom	U-Kingdom
United States	U-States
Virgin Islands	V-Islands
Wallis and Futuna	W-Futuna
Western Sahara	W-Sahara