

Rapport du Projet Data Analysis

Année Universitaire : 2022-2023

Intitulé du projet : Analyser une Dataset par la méthode ACP sous SPSS.

Le créneau du projet : Data Analysis

Réalisé par :

- ELGHAZI Soufiane N°13
- LABYADY Ibtissam N°23
- MAASRI Amine N°26

Encadré par :

- Dr. KHALIFI Hamid

Remerciements :

Nos remerciements s'adressent principalement à notre cher professeur Dr. KHALIFI Hamid de ses efforts remarquables afin de nous bien approcher des concepts de l'analyse des données et pour nous assigner ce projet dans l'intérêt de bien mobiliser nos prérequis.

Résumé :

Nous avons analysé un jeu de données du « Rapport *mondial sur le bonheur* » par la méthode ACP afin de tirer le maximum d'interprétations.

Ce projet nous a fait découvrir plus profondément la méthode ACP et l'appliquer comme concept de l'analyse des données.

Introduction Générale :

Dans le cadre de l'application de la méthode ACP sur une data set sous SPSS afin d'extraire le maximum des interprétations, on a choisi une data set de « *World Happiness Report* ».

A travers le présent rapport, nous allons commencer par décrire le Domain et la problématique de notre data set et les besoins d'analyse, Puis nous allons détailler les différentes étapes de l'analyse par ACP en interprétant les résultats obtenus, avant d'arriver vers la fin à parler des apports de celui-ci.

Table des matières

<u>REMERCIEMENTS :</u>	2
<u>RÉSUMÉ :</u>	2
<u>INTRODUCTION GÉNÉRALE :</u>	2
<u>TABLE DES MATIÈRES :</u>	3
<u>LISTE DES ABRÉVIATIONS :</u>	4
<u>LISTE DES FIGURES</u>	4
<u>CHAPITRE I : INTRODUCTION</u>	5
A. <u>DOMAINE :</u>	5
B. <u>PROBLÉMATIQUE:</u>	6
C. <u>BESOINS D'ANALYSE:</u>	6
D. <u>CONCLUSION:</u>	6
<u>CHAPITRE II: DICTIONNAIRE DES VARIABLES ET MODALITÉS</u>	7
A. <u>INTRODUCTION :</u>	7
B. <u>DICTIONNAIRE DES VARIABLES :</u>	7
C. <u>DICTIONNAIRE DES MODALITÉS :</u>	7
D. <u>CONCLUSION :</u>	7
<u>CHAPITRE III : ANALYSE EN COMPOSANTES PRINCIPALES(ACP)</u>	8
A. <u>INTRODUCTION :</u>	8
B. <u>DÉFINITION :</u>	8
C. <u>ILLUSTRATION :</u>	8
E. <u>CONCLUSION :</u>	13
<u>CONCLUSION :</u>	14
<u>BIBLIOGRAPHIE :</u>	14

Liste des abréviations :

ACP : Analyse en Composantes Principales.

SPSS: Statistical Package for the Social Sciences

Liste des figures :

<u>Figure1 : Affichage des variables</u>	7
<u>Figure2 : Matrice de corrélation</u>	8
<u>Figure3 : Indice KMO et test de Bartlett</u>	9
<u>Figure4 : Qualité de représentation</u>	9
<u>Figure5 : Variance totale expliqué</u>	10
<u>Figure6 : Graphique de valeurs propres</u>	10
<u>Figure7 : Matrice des composantes</u>	11
<u>Figure8 : Matrice des composantes après rotation</u>	11
<u>Figure9 : Matrice des composantes</u>	11
<u>Figure10 :Diagramme des composantes (variable éliminée)</u>	12
<u>Figure11 : Variance totale expliqué(variable éliminée)</u>	12
<u>Figure11 : Diagramme des pays</u>	13

Chapitre I : Introduction

A -Domaine :

Le Rapport sur le bonheur dans le monde est une enquête historique sur l'état du bonheur dans le monde. Le premier rapport a été publié en 2012, le deuxième en 2013, le troisième en 2015 et le quatrième dans la mise à jour 2016. Le World Happiness 2017, qui classe 155 pays selon leur niveau de bonheur, a été publié aux Nations Unies lors d'un événement célébrant la Journée internationale du bonheur le 20 mars. Le rapport continue de gagner en reconnaissance mondiale alors que les gouvernements, les organisations et la société civile utilisent de plus en plus les indicateurs de bonheur pour éclairer leurs décisions politiques. Des experts de renom dans tous les domaines - économie, psychologie, analyse d'enquêtes, statistiques nationales, santé, politiques publiques et autres - décrivent comment les mesures du bien-être peuvent être utilisées efficacement pour évaluer les progrès des nations.

Les scores de bonheur et les classements utilisent les données du Gallup World Poll. Les scores sont basés sur les réponses à la principale question d'évaluation de la vie posée dans le sondage. Cette question, connue sous le nom d'échelle de Cantril, demande aux répondants de penser à une échelle avec la meilleure vie possible pour eux étant un 10 et la pire vie possible étant un 0 et d'évaluer leur propre vie actuelle sur cette échelle. Les scores proviennent d'échantillons représentatifs au niveau national pour l'année 2018 et utilisent les pondérations Gallup pour rendre les estimations représentatives. Les colonnes suivant le score de bonheur estiment dans quelle mesure chacun des **six facteurs** - *production économique, soutien social, espérance de vie, liberté, absence de corruption et générosité* - contribue à rendre les évaluations de la vie plus élevées dans chaque pays qu'elles ne le sont dans *Dystopie*, un pays hypothétique qui a des valeurs égales aux moyennes nationales les plus basses du monde pour chacun des six facteurs. Ils n'ont aucun impact sur le score total rapporté pour chaque pays, mais ils expliquent pourquoi certains pays se classent plus haut que d'autres.

B - Problématique :

Quels pays ou régions se classent au premier rang pour le bonheur global et chacun des six facteurs contribuant au bonheur ?

Quels sont les facteurs qui influent sur le score de bonheur des pays ?

C - Besoins d'analyse :

Qu'est-ce que la Dystopie ?

La dystopie est un pays imaginaire qui compte les personnes les moins heureuses du monde. Le but de l'établissement de Dystopie est d'avoir une référence par rapport à laquelle tous les pays peuvent être comparés favorablement (aucun pays n'obtient de moins bons résultats que Dystopie) en termes de chacune des six variables clés, permettant ainsi à chaque sous barre d'avoir une largeur positive. Les scores les plus bas observés pour les six variables clés caractérisent donc la dystopie. Étant donné que la vie serait très désagréable dans un pays avec les revenus les plus bas du monde, l'espérance de vie la plus basse, la générosité la plus faible, le plus de corruption, le moins de liberté et le moins de soutien social, on parle de « dystopie », contrairement à l'utopie.

Que décrivent les colonnes qui succèdent au score de bonheur (comme la famille, la générosité, etc.) ?

Les colonnes suivantes : **PIB par habitant, Famille, Espérance de vie, Liberté, Générosité, Confiance Gouvernement Corruption** décrivent dans quelle mesure ces facteurs contribuent à évaluer le bonheur dans chaque pays.

Si vous additionnez tous ces facteurs, vous obtenez le score de bonheur, il peut donc être peu fiable de les modéliser pour prédire les scores de bonheur.

D - Conclusion :

Dans ce qui précède, nous avons décrit le Domain, la problématique et les besoins d'analyse de notre data set, dans le chapitre suivant nous présenterons notre dictionnaire des variables et des modalités.

Chapitre II : Dictionnaire des variables et modalités







A. Introduction :

Après avoir introduire notre problématique dans le chapitre précédent, nous présentons les dictionnaires des données, cette phase a pour but de bien positionner afin de comprendre les interprétations obtenues.

B. Dictionnaire des Variables :

Le Rapport sur le bonheur dans le monde est une enquête sur l'état du bonheur mondial. Les scores de bonheur (0-10) sont basés sur les réponses à la principale question d'évaluation de la vie posée dans le sondage.

Colonnes de score de bonheur :

-  GDP per capita : PIB par habitant
-  Social support : Famille - Soutien social
-  Healthy life expectancy : Espérance de vie
-  Freedom to make life choices: Liberté
-  Generosity : Générosité
-  Perceptions of corruption : Confiance Gouvernement Corruption.

Si nous additionnons tous ces facteurs, nous obtiendrons le score de bonheur, il pourrait donc être peu fiable de les modéliser pour prédire les scores de bonheur.

Nom	Type	Largeur	Décimales	Etiquette	Valeurs	Manquant	Colonnes	Align	Mesure	Rôle
Overallrank	Numérique	3	0	Classement général	Aucun	Aucun	8	Centre	Echelle	Entrée
Countryorregion	Chaîne	24	0	Pays ou région	Aucun	Aucun	25	Centre	Nominales	Entrée
Score	Numérique	5	3	Score	Aucun	Aucun	10	Centre	Echelle	Entrée
GDPpercapita	Numérique	5	3	PIB par habitant	Aucun	Aucun	12	Centre	Echelle	Entrée
Socialsupport	Numérique	5	3	Aide sociale	Aucun	Aucun	12	Centre	Echelle	Entrée
Healthylifeexpectancy	Numérique	5	3	Espérance de vie en bonne santé	Aucun	Aucun	15	Centre	Echelle	Entrée
Freedomtomakelifecoices	Numérique	5	3	Liberté de choix de vie	Aucun	Aucun	18	Centre	Echelle	Entrée
Generosity	Numérique	5	3	Générosité	Aucun	Aucun	10	Centre	Echelle	Entrée
Perceptionsofcorruption	Numérique	5	3	Perceptions de corruption	Aucun	Aucun	15	Centre	Echelle	Entrée

Figure 1 : Affichage des variables

C. Dictionnaires de modalités :

Les données contiennent 9 colonnes et 156 lignes. Les variables sont de type quantitatif, ce qui rend la méthode ACP possible. Notre objectif principal est de faire une analyse exploratoire des facteurs qui rendent les gens heureux.

D. Conclusion :

Nous avons présenté notre dictionnaire des données, dans le chapitre suivant on va faire une analyse en utilisant ACP afin d'interpréter les résultats obtenus sous SPSS.

Chapitre III : Analyse en Composantes Principale (ACP)

A. Introduction :

Cette partie contient le dernier volet de ce rapport. Elle a pour objectif d'exposer le travail achevé. Nous nous focalisons sur l'aspect analytique des données, cette phase a pour but de faire apparaître la méthode ACP afin d'interpréter les résultats.

B. Définition :

ACP : algorithme de réduction dimensionnelle non supervisé capable d'identifier les corrélations dans un jeu de données et de le transformer en un ensemble de données avec un nombre réduit de variables en minimisant la perte d'information.

SPSS : est un logiciel utilisé pour l'analyse statistique.

C. Illustration :

Analyser les résultats d'un ACP, c'est répondre à trois questions :

1. Les données sont-elles factorisables ?

- Pour répondre à cette question, dans un premier temps, il convient d'observer la matrice des corrélations. Si plusieurs variables sont corrélées (> 0.5), la factorisation est possible. Si non, la factorisation n'a pas de sens et n'est donc pas conseillée.
- Dans un deuxième temps, il faut observer l'indice de KMO (Kaiser-Meyer-Olkin) qui doit tendre vers 1. Si ce n'est pas le cas, la factorisation n'est pas conseillée.
- Enfin, on utilise le test de sphéricité de Bartlett. : si la signification (Sig) tend vers 0.000, c'est très significatif, inférieur à 0.05 significatif, entre 0.05 et 0.10 acceptable et au-dessus de 0.10, on rejette.

Si l'ACP satisfait à au moins deux de ces trois conditions, on pourrait continuer.

Matrice de corrélation							
		PIB par habitant	Aide sociale	Espérance de vie en bonne santé	Liberté de choix de vie	Générosité	Perceptions de corruption
Corrélation	PIB par habitant	1,000	,728	,866	,366	-,013	,320
	Aide sociale	,728	1,000	,675	,405	,019	,218
	Espérance de vie en bonne santé	,866	,675	1,000	,359	,021	,316
	Liberté de choix de vie	,366	,405	,359	1,000	,299	,462
	Générosité	-,013	,019	,021	,299	1,000	,362
	Perceptions de corruption	,320	,218	,316	,462	,362	1,000

Figure 2 : Matrice de corrélation

On a d'après la matrice de corrélation plusieurs variables qui ne se sont pas supérieures à 0.5. Donc, on peut dire que la factorisation n'est pas conseillée selon la matrice de corrélation.

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,738
Test de sphéricité de Bartlett	Khi-deux approximé	431,179
	ddl	15
	Signification de Bartlett	,000

Figure 3 : Indice KMO et test de Bartlett

On a l'indice de KMO= 0.738 est proche de 1 ainsi la Signification de Bartlett =0.000, c'est très significatif. Donc la factorisation est conseillée.

L'ACP satisfait à au moins deux de ces trois conditions (KMO et Bartlett), Donc on peut continuer.

Qualité de représentation

	Initial	Extraction
PIB par habitant	1,000	,885
Aide sociale	1,000	,745
Espérance de vie en bonne santé	1,000	,841
Liberté de choix de vie	1,000	,597
Générosité	1,000	,687
Perceptions de corruption	1,000	,638

Figure 4 : Qualité de représentation

En outre, dans le tableau Qualité de représentation toutes les variables sont supérieures à 0,5 donc toutes les variables sont prises en compte dans l'ACP.

2. Combien de facteurs retenir ? :

Trois règles sont applicables :

- 1ere règle : la règle de Kaiser qui veut qu'on ne retienne que les facteurs aux valeurs propres supérieures à 1.
- 2eme règle : on choisit le nombre d'axe en fonction de la restitution minimale d'information que l'on souhaite.

Pour ces deux premières règles, on examine le tableau « Variance totale expliquée ».

Variance totale expliquée

Composante	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	2,988	49,802	49,802	2,988	49,802	49,802
2	1,406	23,428	73,230	1,406	23,428	73,230
3	,596	9,928	83,158			
4	,563	9,378	92,536			
5	,320	5,326	97,862			
6	,128	2,138	100,000			

Méthode d'extraction : Analyse en composantes principales.

Figure 5 : Variance totale expliquée

D'après la 1ere règle de Kaiser et d'après le tableau Variance totale expliquée et dans la colonne 'cumulés', on voit qu'on restitue 73,230% si on retient 2 facteurs.

- 3eme règle : test du coude. On observe le graphique des valeurs propres et on ne retient que les valeurs qui se trouvent à gauche du point d'inflexion. Graphiquement, on part des composants qui apportent le moins d'information (qui se trouvent à droite), on relie par une droite les points presque alignés et on ne retient que les axes qui sont au-dessus de cette ligne.

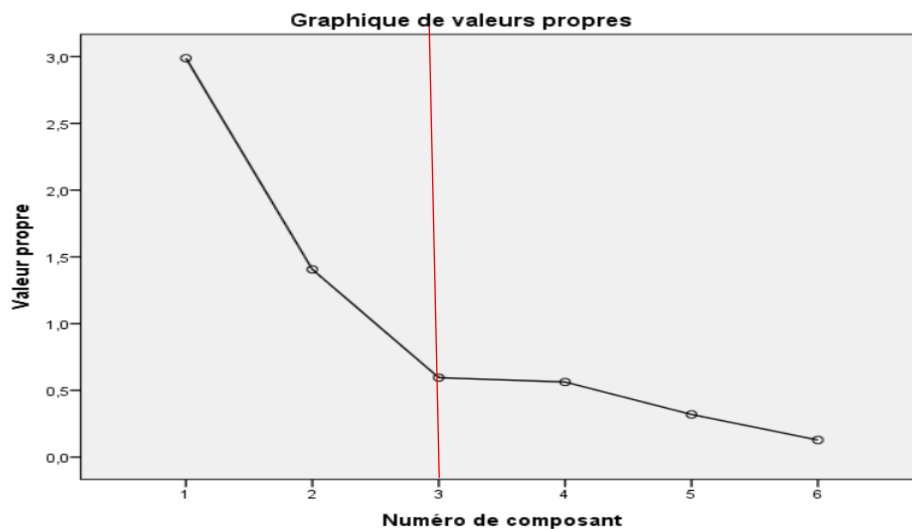


Figure 6 : Graphique de valeurs propres

Le Tracé d'effondrement montre que les deux premières composantes : composante 1 et composante 2 sont les composantes qu'il faut retenir parce que le point de changement du tracé est le composante 3.

3. Comment interpréter les résultats ?

C'est la phase la plus délicate de l'analyse. On donne un sens à un axe grâce à une recherche lexicale à partir des coordonnées des variables et des individus. Ce sont les éléments extrêmes qui concourent à l'élaboration des axes.

Figure 7 : Matrice des composantes

D’après la matrice des composantes on voit que 88% de la variable PIB par habitant est représenté par l’axe 1, et le reste par l’axe 2, ainsi de suite pour les autres variables.

Matrice des composantes ^a		
	Composante	
	1	2
PIB par habitant	,886	-,316
Espérance de vie en bonne santé	,871	-,287
Aide sociale	,816	-,282
Liberté de choix de vie	,649	,420
Générosité	,211	,801
Perceptions de corruption	,560	,570

Méthode d'extraction : Analyse en composantes principales.
a. 2 composantes extraites.

Matrice des composantes après rotation ^a		
	Composante	
	1	2
PIB par habitant	,934	,111
Espérance de vie en bonne santé	,908	,130
Aide sociale	,856	,111
Générosité	-,168	,811
Perceptions de corruption	,247	,760
Liberté de choix de vie	,394	,665

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.
a. La rotation a convergé en 3 itérations.

Figure 8 : Matrice des composantes après rotation

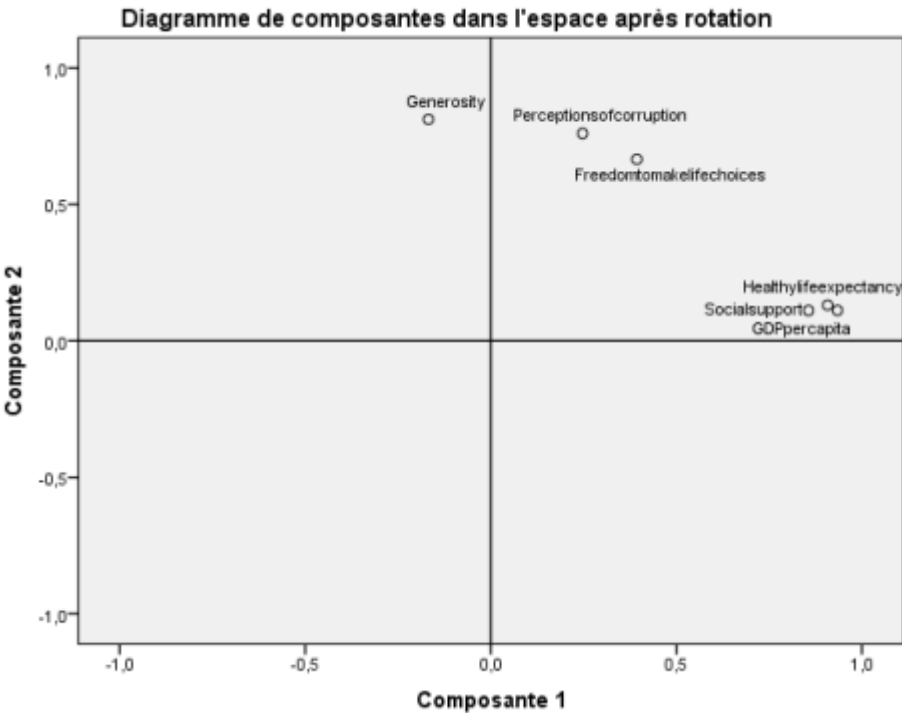


Figure 9 : Diagramme de composantes

Le Tracé des composantes dans l'espace après rotation montre que la composante 1 se compose principalement des variables : PIB par habitant, soutien social et Espérance de vie .

D'après le Tracé après rotation on voit que la variable générosité est éloignée donc il faut l'éliminer et refaire l'analyse factorielle et ainsi on aura le diagramme suivant

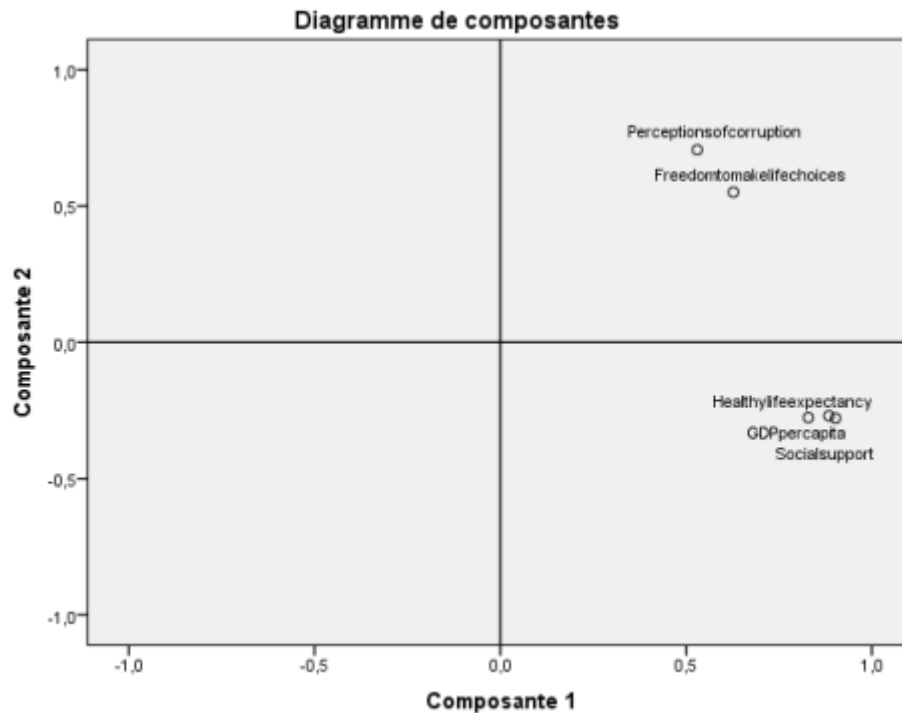


Figure 10 : Diagramme de composante (variable éliminée)

Ainsi l'élimination de la variable générosité a augmenté l'information retenue par les 2 composantes principal de 73,230 à 79,753

Variance totale expliquée						
Composante	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	2,959	59,187	59,187	2,959	59,187	59,187
2	1,028	20,565	79,753	1,028	20,565	79,753
3	,563	11,255	91,007			
4	,320	6,407	97,415			
5	,129	2,585	100,000			

Méthode d'extraction : Analyse en composantes principales.

Figure 11 : Variance totale expliqué (variable éliminée)

D'après le diagramme de composante ci-dessus on peut remarquer clairement que les trois variables PIB par habitant, espérance de vie, aide social sont fortement corrélée avec la première composante, cela indique que ces trois variables varient ensemble si l'une augmente, les trois autres ont tendance à faire la même chose. Et plus précisément on pourrait affirmer que sur la base de la nouvelle corrélation de 0,903 pour le Pib qu'on déduit que c'est lui le facteur principal qui influent sur le taux de bonheur dans ces pays, plus le PIB

par habitant augmente plus les autres variables augmente tout de même et plus le score de bonheur dans chaque pays est élevé.

De même pour aussi pour les deux variables perceptions de corruption et liberté de choix de vie qui forment un cluster aussi et ceci indique qu'ils sont corrélés aussi, mais moins corrélés que les trois autres variables sur lesquels on a parlé dans le paragraphe précédent.

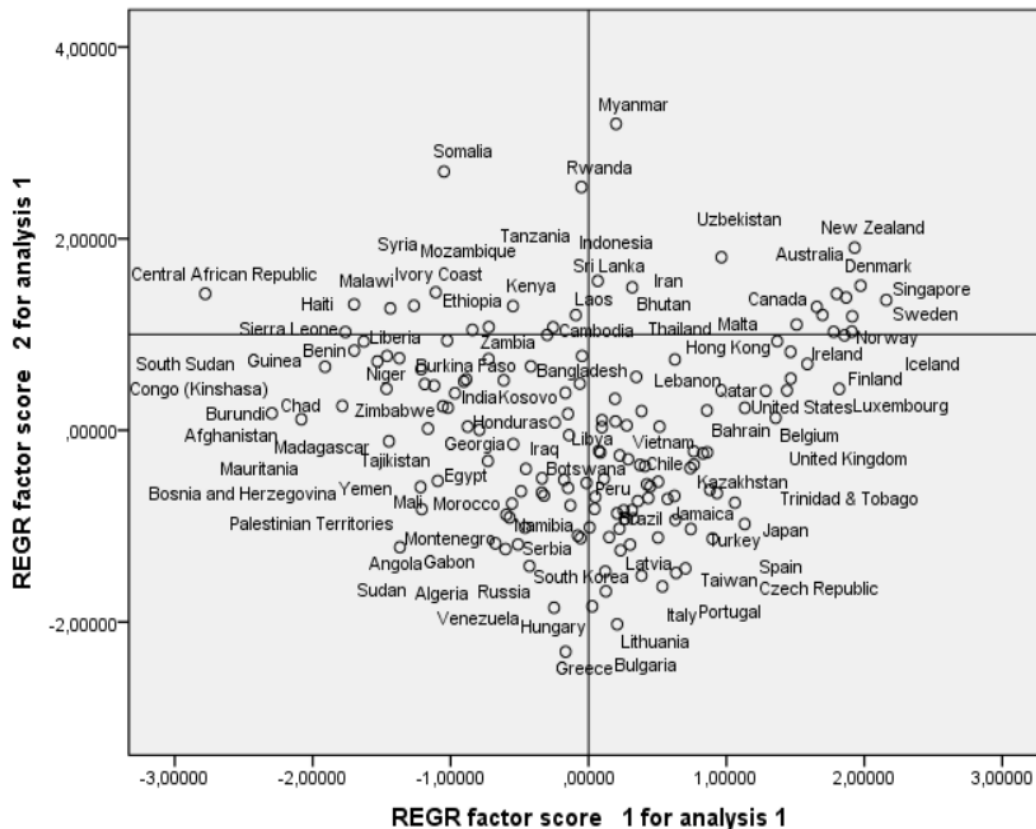


Figure 12 : Diagramme des pays

D'après ce diagramme on observe que plus les pays sont à droites plus leurs score de bonheur est élevée et cela met l'accent aussi sur leur développement au niveau de plusieurs domaines et ceci explique leurs valeurs importants des indices sur lesquelles nous nous sommes basées pour mener cette analyse et en déduire que les peuples des pays scandinaves sont les pays les plus heureux du monde, par contre ceux des pays africaines sont les plus malheureux.

E. Conclusion :

À ce stade, nous atteignons la fin de l'étude du projet. Dans ce dernier chapitre, nous avons à la fois illustrer, analyser et interpréter notre résultat obtenu par ACP sur SPSS. À présent, nous passerons, dans la partie suivante à la conclusion globale du projet.

Conclusion :

Dans notre projet nous avons mis en œuvre l'analyse d'une data set de « World Happiness Report » sous SPSS, l'objectif est d'appliquer la méthode ACP afin d'extraire le maximum possible d'interprétations.

Pour aboutir à ce résultat, on a passé d'abord d'introduire les données à analyser vers l'analyse et l'interprétation.

L'introduction (description), est une étape majeure et un appui nécessaire, nous aide à simplifier la réalisation du projet et rend clair le Domain de notre data set ainsi notre analyse.

Quant à l'étape de l'analyse elle consiste à mettre en œuvre les prérequis du cours. Elle permet d'avoir le résultat final souhaité. Elle prend généralement le volume horaire majeur dans le traitement du sujet vu son importance et sa complexité.

En conclusion alors, ce projet nous a permis également de développer notre esprit d'interprétation et de la réflexion, Il nous a permis d'enrichir et d'approfondir nos connaissances de l'analyse des données.

Bibliographie :

- [World happiness report 2018 data set](#)
- [World happiness report 2019 data set](#)