PROJET D'ANALYSE DES DONNEES

LA QUALITE DE VIE DANS LE MONDE

M1 ISIFAR

RAMIRES ALIAGA CHRISTIAN LUIS

BEHZOD ISRALOV

SOMMAIRE

PRESENTATION

Dans notre projet, on analysera **la qualité de vie** sur 140 pays en 2016 par rapport au PIB par habitant, au soutien social, à la liberté de faire de choix de vie, à la perception de la corruption, à l'espérance de vie, aux dons d'argent et aux continents. Ces données on les a pris sur le site Kaggle qui les avaient pris sur **le rapport du bonheur dans le monde publié par l'ONU**, accessible sur Wikipédia, à l'exception de la variable Région. On a aussi transformé la variable quantitative Famille en une variable qualitative. C'est ainsi qu'on a obtenu 5 variables quantitatives et 2 variables qualitatives.

A l'aide de plusieurs méthodes d'analyse des données, comme par exemple l'analyse en composantes principales, on dévoilera les relations pouvant exister entre les variables et les pays. Avant de commencer, on va décrire ci-dessous ce qui représentent chacune des variables.

ECONOMIE : **le PIB par habitant.**

FAMILLE : **le soutien social** est la moyenne nationale des réponses à la question : si vous avez des problèmes, avez-vous des parents ou des amis sur qui vous pouvez compter pour vous aider à chaque fois que vous avez besoin d'eux ou pas ?

LIBERTE : **la liberté de faire des choix de vie** est la moyenne nationale des réponses binaires à la question : êtes-vous satisfait insatisfait de votre liberté de choisir ce que vous faites de votre vie ?

CORRUPTION : **la perception de la corruption** est la moyenne de réponses binaires à deux questions : la corruption est-elle répandue dans tout le pays ? Au gouvernement ou pas ? Et la corruption est-elle répandue dans les entreprises ou pas ?

SANTE : **l'espérance de vie** en bonne santé à la naissance sont construits à partir de données de l'Organisation Mondiale de la Santé.

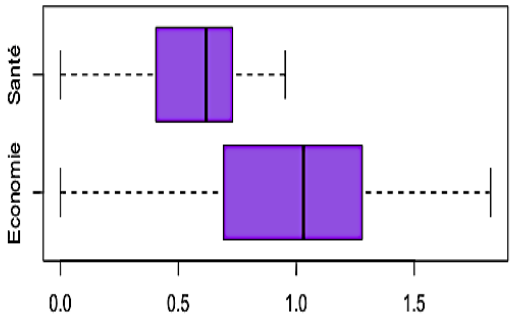
GENEROSITE : la générosité est le résidu de la régression la moyenne nationale des réponses à la question : avez-vous fait un **don argent à un organisme de bienfaisance** le mois dernier ?

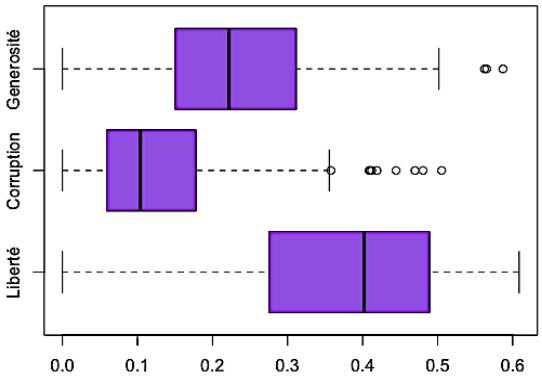
REGION : on expliquera ceci plus tard.

I. ANALYSE UNIVARIEE

1. VARIABLES QUANTITATIVES

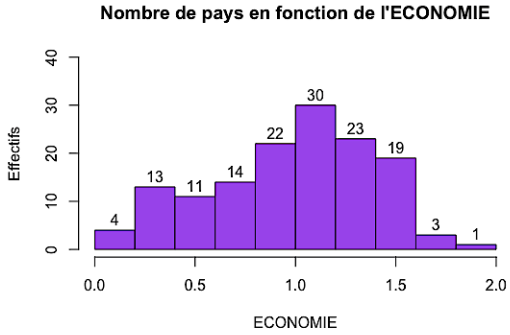
Caractéristiques de position de chaque variable quantitative

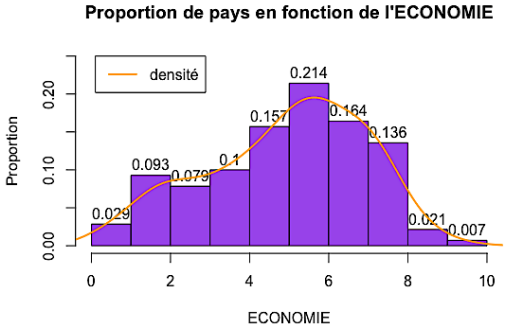




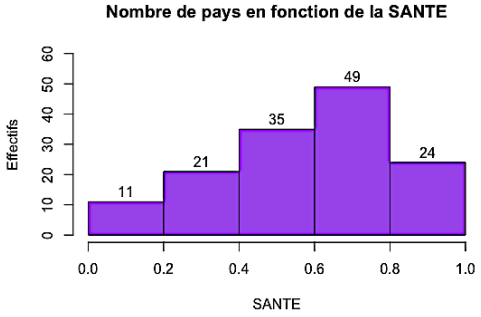
INDICATEURS STATISTIQUES					
	MIN	MAX	MEDIANNE	MOYENNE	VARIANCE
ECONOMIE	0.06831	1.8243	1.04351	0.98258	0.1592083
SANTE	0.03824	0.9528	0.61854	0.57304	0.04720166
LIBERTE	0.00589	0.6085	0.40277	0.38298	0.01912964
CORRUPTION	0.00322	0.50521	0.10580	0.13931	0.01263129
GENEROSITE	0.02025	0.5870	0.22121	0.23997	0.01618896

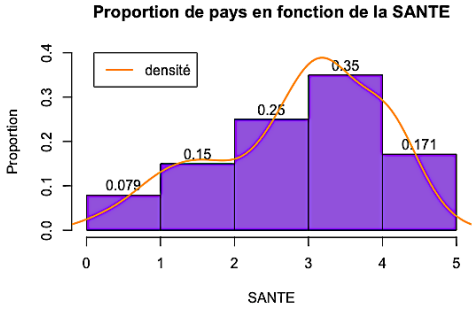
Analysons par rapport aux classes :



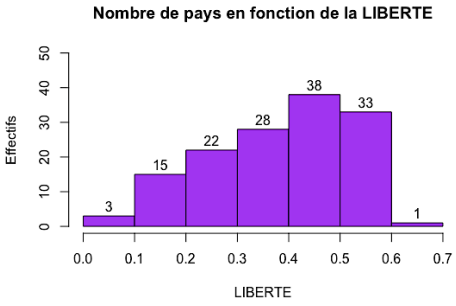


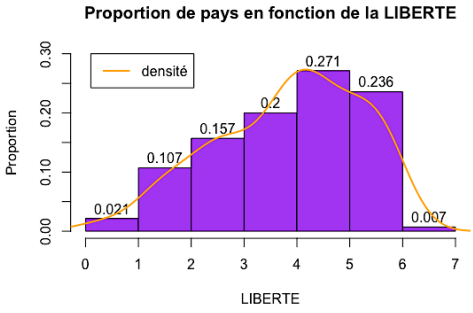
On constate qu'il y a 89 pays représentant 64 % des individus qui ont une économie moyenne, 23 pays représentant 16 % des individus qui ont une économie importante et 28 pays représentant 20 % des individus qui ont une économie mauvaise.



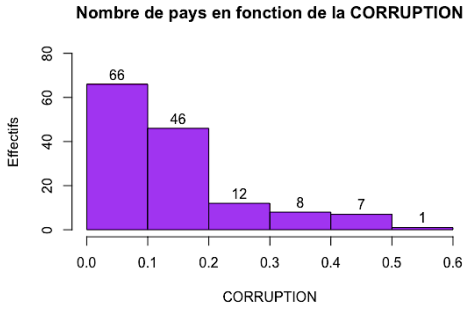


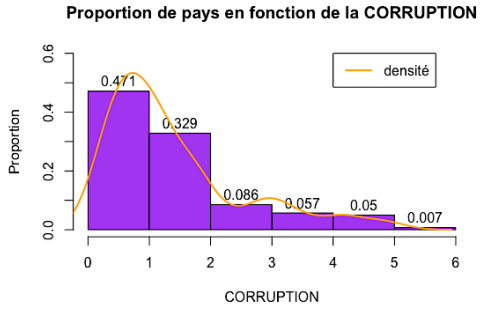
On constate qu'il y a 84 pays représentant 60 % des individus qui ont une espérance de vie moyenne, 24 pays représentant 17 % des individus qui ont une espérance de vie élevée et 32 pays représentant 23 % des individus qui ont une espérance de vie basse.



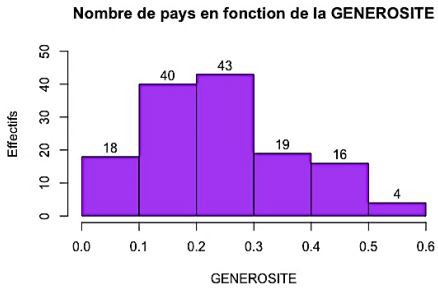


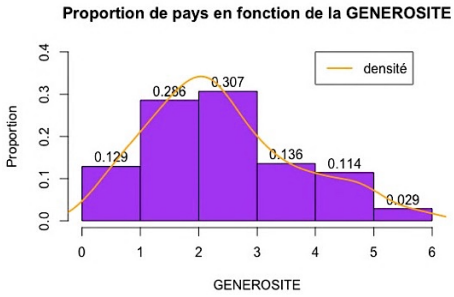
On constate qu'il y a 88 pays représentant 63 % des individus qui ont une liberté de choix de vie moyenne, 34 pays représentant 24 % des individus qui ont une liberté importante et 18 pays représentant 13 % des individus qui ont une liberté médiocre.





On constate qu'il y a 112 individus, représentant 80 % de pays, ont une perception de la corruption dans la politique et les entreprises faible ; 20 individus, représentant 14 % de pays, ont une perception moyenne ; 8 pays, représentant 6 % des individus, ont une perception élevée.





On constate que chez 58 individus, représentant 41 % de pays, ne font pas souvent un don d'argent à un organisme de bienfaisance ; 62 individus, représentant 45 % de pays, font plus ou moins un don d'argent ; 20 individus, représentant 14 % de pays, font fréquemment un don d'argent.

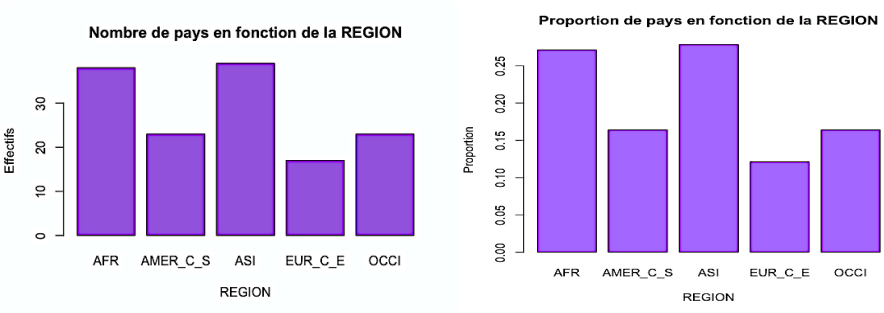
Les variables suivent une loi normale ?

On a fait un test de Shapiro et on a obtenu p-value < 5 %. Ceci veut dire que les valeurs de chacune de nos variables quantitatives ne suivent pas une loi normale.

2. VARIABLES QUALITATIVES

On peut distinguer ci-dessous 5 modalités pour la variable REGION :

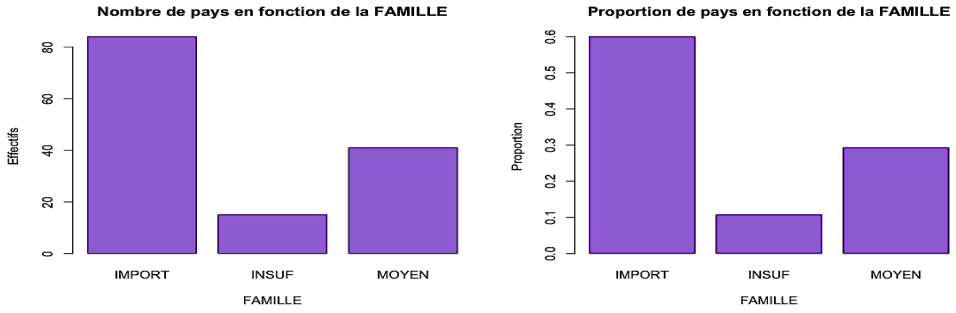
AFRI : Afrique
AMER_C_S : Amérique Centrale et du Sud
ASI : Asie
EUR_C_E : Europe Centrale et de l'Est
OCCI : Europe de l'Ouest, États-Unis, Canada, Australie et Nouvelle Zélande.



Les régions qui contiennent le plus d'effectifs sont l'Afrique et l'Asie qui représentent les 27 % et 28 % des individus respectivement. Les 3 régions restant contiennent à peu près 20 pays chacune et représentent à peu près 15 % chacune.

On peut distinguer ci-dessous 3 modalités pour la variable FAMILLE :

IMPORT : important
INSUF : insuffisant
MOYEN : moyen



Concernant la variable FAMILLE, on distingue 3 modalités. Dans 84 pays représentant 60 % des pays, le soutien social est important. Par contre, on constate que dans 41 pays représentant 29 %, le soutien social est moyen et dans les 15 pays restant le soutien social est insuffisant.

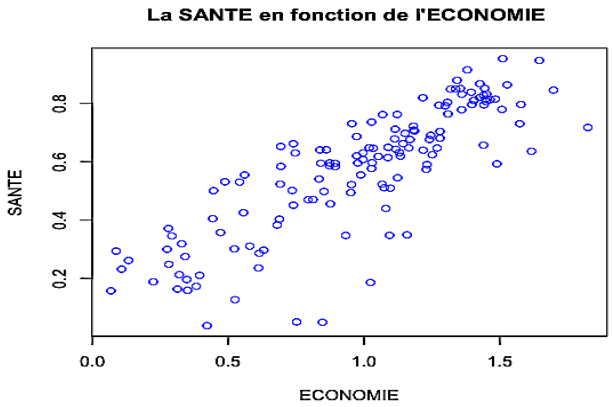
II. ANALYSE MULTIVARIEE

1. VARIABLES QUANTITATIVE ET QUANTITATIVE

a. ANALYSE BIVARIEE

Tableau de corrélation de Pearson					
	ECONOMIE	SANTE	LIBERTE	CORRUPTION	GENEROSITE
ECONOMIE	1	0.83177096	0.41142980	0.38537096	0.0369734
SANTE	0.83177096	1	0.4152955	0.3233234	0.1414462
LIBERTE	0.41142980	0.4152955	1	0.5039882	0.3792825
CORRUPTION	0.38537096	0.3233234	0.5039882	1	0.2929421
GENEROSITE	0.04995262	0.1414462	0.3792825	0.2929421	1

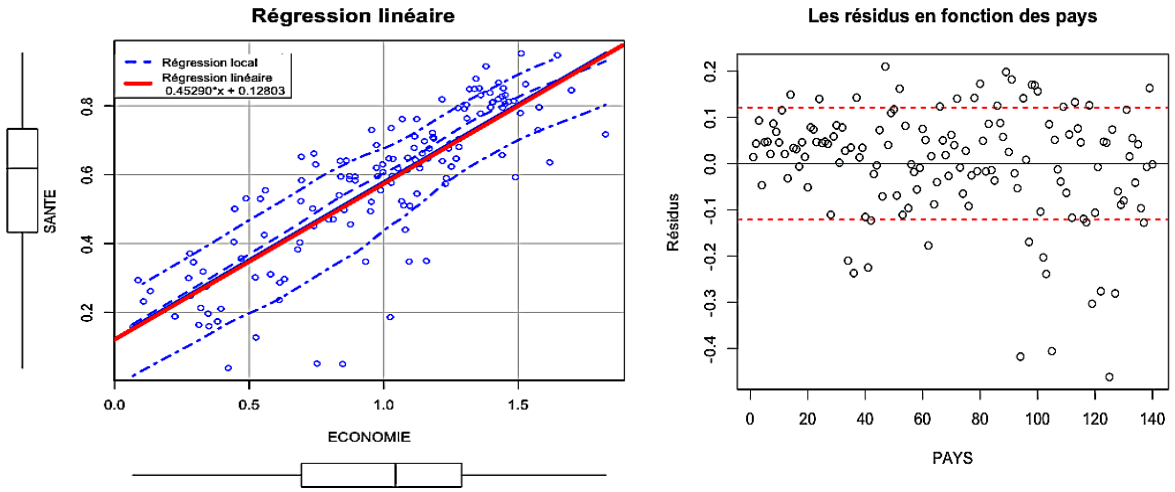
On constate qu'il y a une forte corrélation entre les variables ECONOMIE et SANTE. On va donc analyser la relation qu'il y a entre ces 2 variables.



A-T-on 2 variables suivant une loi normale ?

Les distributions de 2 variables ne suivaient pas une loi normale. On ne peut donc comparer les 2 variances des variables à l'aide du test de Fisher.

Faisons une régression linéaire :



La commande lm nous donne : Multiple R-squared : 0.6918, p-value : < 2.2e-16.

Ici, la variable ECONOMIE représente 69% de l'information de la variation de la variable SANTE et la valeur de p-value veut dire qu'on rejette l'hypothèse d'indépendance. Ceux-ci nous dit que ces 2 variables sont fortement liées.

Que peut-on dire sur les résidus ?

Pour tester les hypothèses de normalité, d'indépendance et d'homogénéité, on pourrait analyser des graphiques mais on peut aussi se servir des tests de R.

L'hypothèse de normalité des résidus est rejetée car notre test de Shapiro-Wilk nous donne p-value = 4.2e-07 < 5 %. L'hypothèse d'indépendance des résidus est acceptée car p-value = 0.53 > 5% grâce au test de Durbin-Watson. L'hypothèse d'homogénéité est acceptée car p-value = 0.29609 > 5% grâce au test de Breush-Pagan.

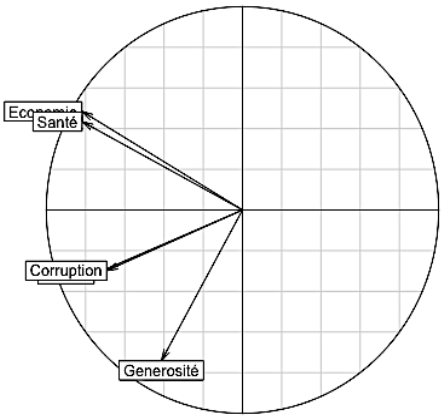
b. ANALYSE EN COMPOSANTES PRINCIPALES

On va commencer tout d'abord pour analyser les valeurs propres et les composantes principales :

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.6008	1.0788	0.8055	0.68433	0.39581
Proportion of Variance	0.5125	0.2328	0.1298	0.09366	0.03133
Cumulative Proportion	0.5125	0.7452	0.8750	0.96867	1.00000

On voit que les 1er, 2ème, 3èmes composantes principales contiennent 51 %, 23% et 13 % de l'information respectivement. On va garder 3 axes qui représentent 88 % de l'information. On pourrait aussi garder 2 axes qui représenteront 74 % de l'information.

Cercle de corrélation entre le 1er et 2ème axe

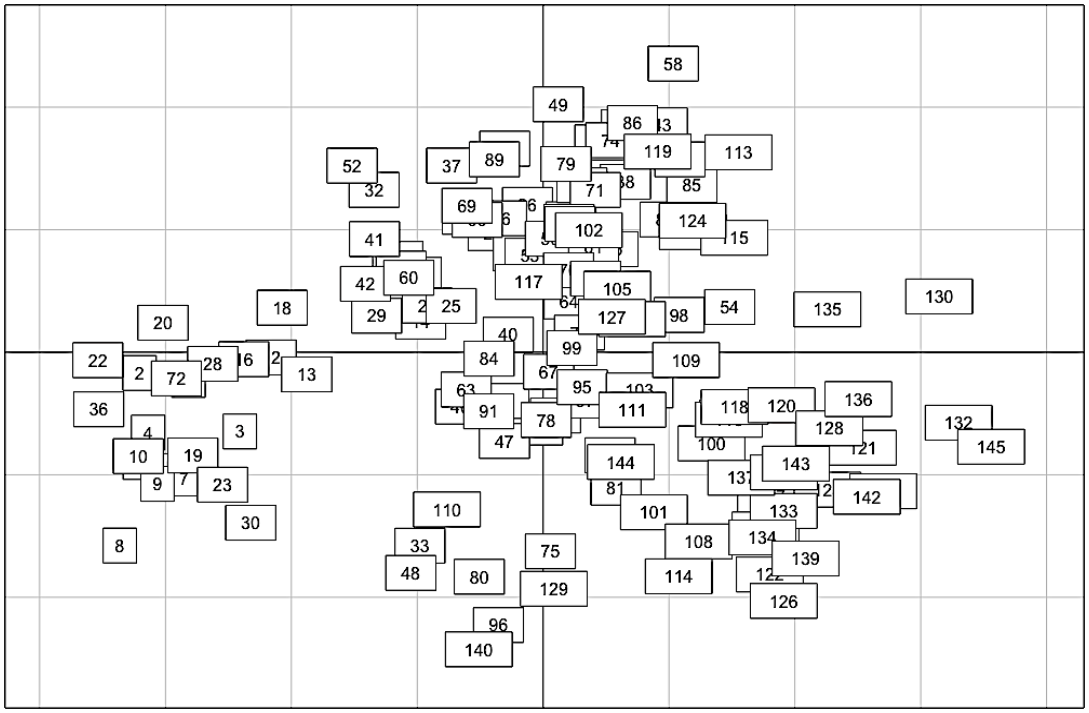


Contribution			
	Axe 1	Axe 2	Axe 3
ECONOMIE	0.262	0.202	0.013
SANTE	0.261	0.162	0.101
LIBERTE	0.223	0.086	0.023
CORRUPTION	0.187	0.077	0.496
GENEROSITE	0.067	0.473	0.368

On constate que les variables ECONOMIE et SANTE sont corrélées. Par contre, elles sont non corrélées avec la variable GENEROSITE. De plus, les variables CORRUPTION et LIBERTE sont aussi corrélées.

On voit aussi que les variables ECONOMIE, SANTE, LIBERTE et CORRUPTION contribuent le plus à la création de la 1ère composante et à peu près avec la même importance. Les variables CORRUPTION et LIBERTE contribuent très peu à la 2ème composante. Et, la variable GENEROSITE a un rôle important à la création de la 2ème composante.

Plan factoriel entre le 1er et 2ème axe



A l'aide de notre plan factoriel et de notre tableau des contributions des individus, qu'on n'affichera pas sur ce document car les données sont nombreuses, on analysera l'importance des certains individus à la création des composantes principales.

Les individus 1, 2, 4, 6, 22, 20, 36, 72, 130, 132, 145 contribuent fortement à la construction de la 1ère composantes et quasiment pas à la 2ème composante. Les individus 43, 48, 49, 74, 80, 86, 96, 114, 129 et 140 ont un rôle important sur le 2ème axe et quasiment pas sur le 1er axe. L'individu 8 contribue de manière importante aux 2 composantes.

De plus, les individus 31, 35, 39, 75 et 129 sont mal représentés par rapport au 1er axe, et les individus 12, 99 et 109 sont mal représentés par rapport au 2ème axe car leurs valeurs de qualité de représentation sur le plan factoriel sont proches de 0.

Grace à notre cercle de corrélation et au plan factoriel on peut extraire des informations importantes :

Les pays comme la Suisse (2), la Belgique (18), le Luxemburg (20), le Singapour (22), la France (32) et le Hong-Kong (72) ont une espérance de vie et un soutien social élevés. Par contre, les pays comme le Malawi (122), le Haïti (126), le Chad (132), la Guinée (139), l'Afghanistan (142) et le Burundi (145) ont une espérance de vie et un soutien social très fiable.

Les pays ayant une perception de la corruption et une liberté de choix de vie assez importantes sont la Norvège (4), le Nederland (7), la Nouvelle Zélande (8), l'Australie (9), la Suède (10), le Singapore (22) et la Malte (30). Du côté opposé on trouve les pays comme l'Arménie (113), l'Angola (130), le Chad (132) et le Madagascar (136) qui ont une perception de la corruption et une liberté de vie élevés.

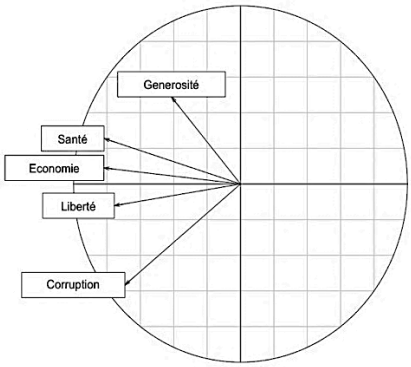
En ce qui concerne le don d'argent, La Nouvelle Zélande (8), l'Australie (9), la Suède (10) font de dons assez couramment. Contrairement, les pays comme la Lituanie (58), l'Arménie (113) et l'Angola les dons d'argent ne sont pas fréquents.

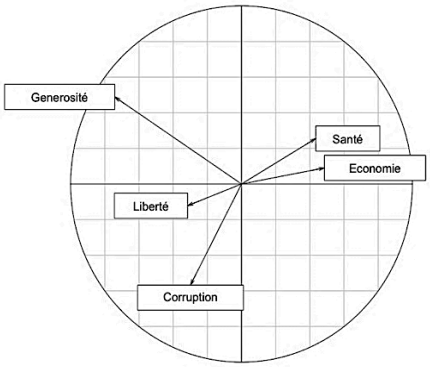
Finalement, on peut constater que les pays qui ont une qualité de vie élevés se trouvent à gauche dans notre plan factoriel et les pays ayant une qualité de vie insuffisante se trouvent à droite.

Peut-on analyser les autres composantes principales ?

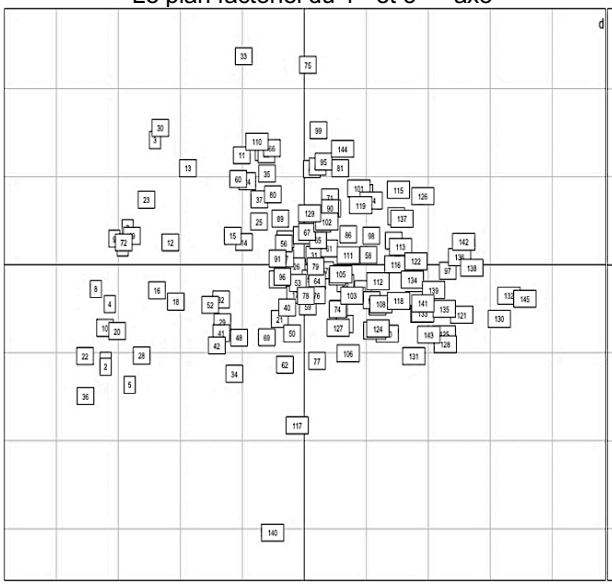
Le cercle de corrélation du 1er et 3ème axe

Le cercle de corrélation du 2ème et 3ème axe

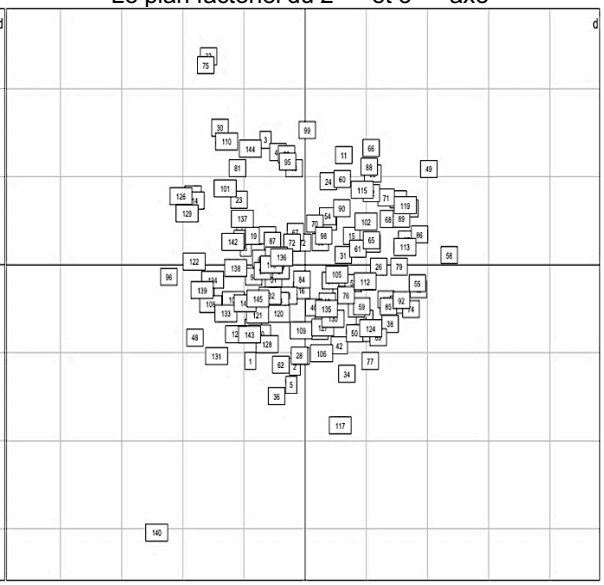




Le plan factoriel du 1er et 3ème axe



Le plan factoriel du 2ème et 3ème axe



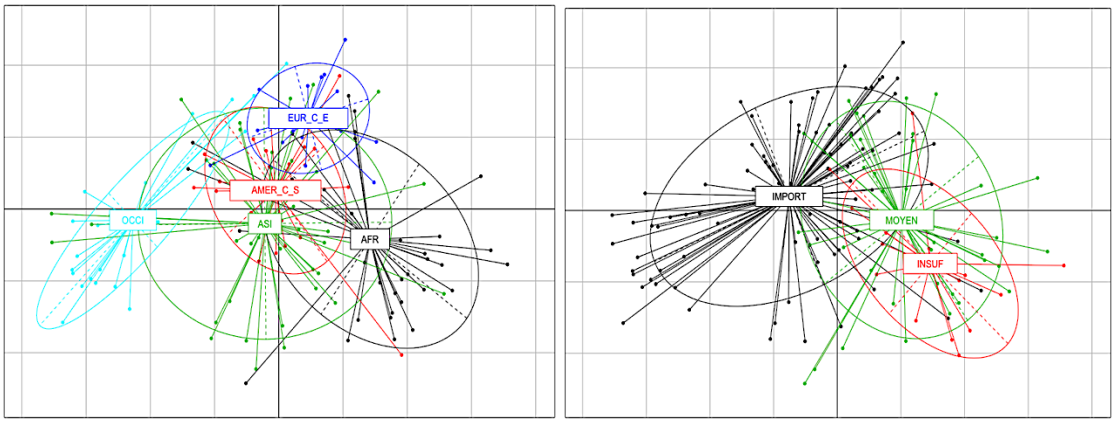
On pourrait également analyser les cercles de corrélation et les plans factoriels ci-dessus qui représentent le 64% de l'information entre le 1er et 3ème axe, et 36 % entre le 2ème et 3ème. Cependant on ne détaillera pas cette partie de l'analyse pour ne pas dépasser le nombre maximal de pages.

c. VARIABLES QUALITATIVES

Peut-on ajouter des variables qualitatives sur le plan factoriel ?

Oui, on peut le faire. Sur les plans factoriels ci-dessous, on peut distinguer la répartition des individus par rapport aux régions et au soutien social. On fera une analyse plus détaillée dans la suite de notre projet.

La REGION dans le plan factoriel du 1er et 2ème axe La FAMILLE dans le plan factoriel du 1er et 2ème axe



2. QUALITATIVE ET QUALITATIVE

a. ANALYSE BIVARIEE

Ici, on va analyser la relation entre les variables REGION et FAMILLE.

Tableau observé			
REGION	**FAMILLE**		
	IMPORT	INSUF	MOYEN
AFRI	9	7	22
AMER_C_S	18	1	4
ASI	22	7	10
EUR_C_E	13	0	4
OCCI	22	0	1

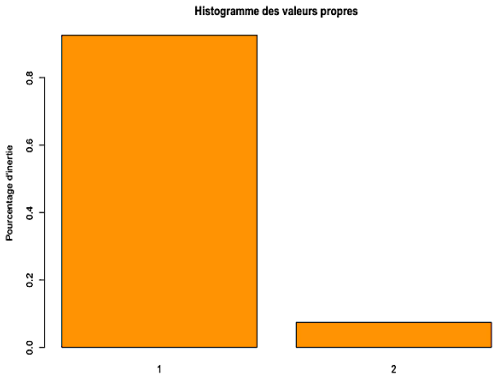
Tableau théorique			
REGION	**FAMILLE**		
	IMPORT	INSUF	MOYEN
AFRI	22.8	4.1	11.13
AMER_C_S	13.8	2.5	6.7
ASI	23.4	4.2	11.4
EUR_C_E	10.2	1.8	5
OCCI	13.8	2.5	6.7

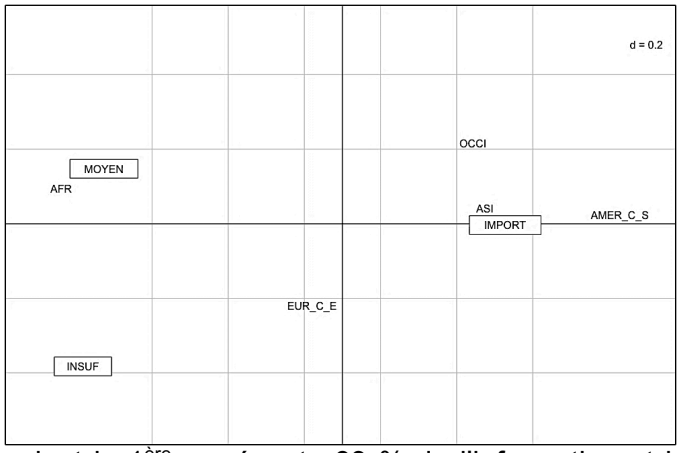
Existe-t-il dépendance entre ces 2 variables quantitatives ?

On ne peut pas utiliser le test du Chi2 car notre tableau théorique contient des cases ayant des effectifs inférieurs à 5. Par contre, on va utiliser le test exact de Fisher qui n'a pas des conditions pour son usage. On constate ainsi que le p-value < 5% et on peut donc dire que les 2 variables qualitatives sont liées.

b. ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC)

Plan factoriel





Ici, on a que 2 composantes principales dont la 1ère représente 93 % de l'information et la 2ème 7%. Les modalités AFRI et AMER_C_E sont très importantes pour la création du 1er axe, et EUR_C_E et OCCI jouent un rôle important à la création du 2ème axe. On constate aussi que EUR_C_E et AMER_C_S sont mal représentés par rapport au 1er et 2ème axe respectivement.

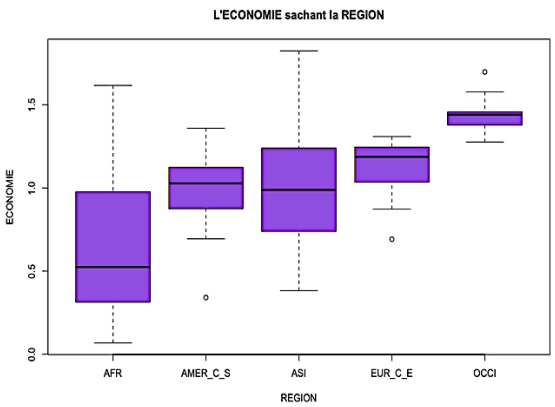
Faisons un choix pour l'analyser du plan factoriel :

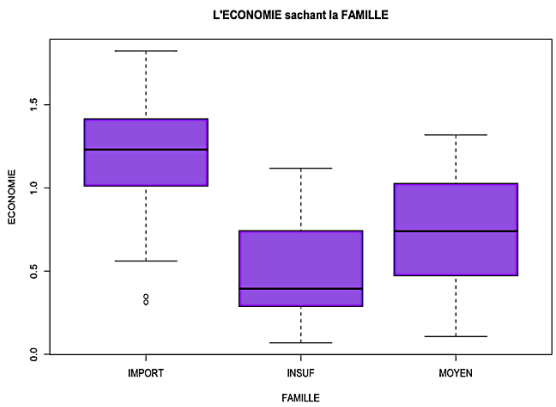
Par rapport à la variable REGION, l'Asie et l'Amérique Centrale et du Sud sont 2 régions qui se caractérisent pour la présence d'un soutien social assez important. Contrairement, en Afrique il y a plus des pays où le soutien social est moyen. De plus, on voit que dans les pays occidentaux il y a surtout des individus présentant un soutien social importante et aucun pays est caractérisé par un soutien insuffisant.

3. VARIABLES QUANTITATIVES ET QUALITATIVES

a. ANALYSE BIVARIEE

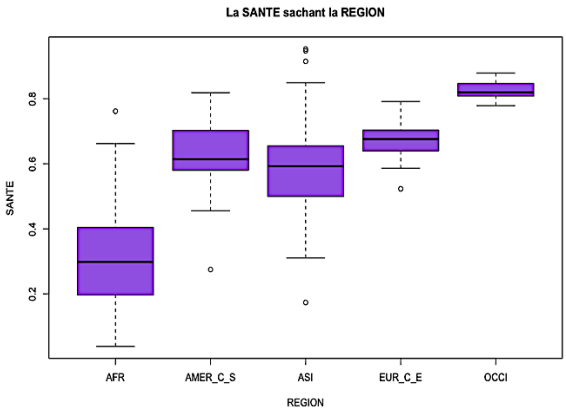
Faisons une analyse de chaque variable quantitative par rapport aux variables qualitatives :

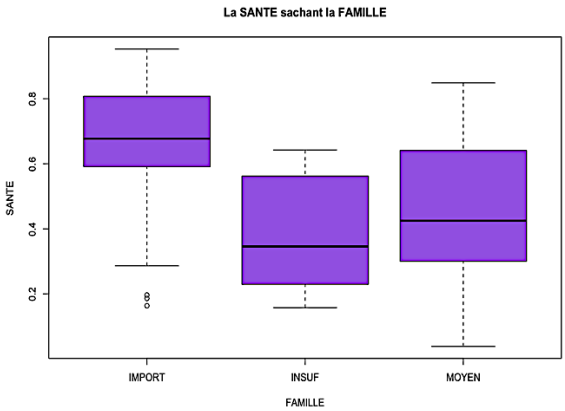




Dans le premier graphique ci-dessus, on constate que chez les pays occidentaux, le PIB par habitant est plus élevé que dans les autres régions. Par contre, les pays qui ont un PIB faible se trouvent en Afrique. On voit aussi qu'en Asie la plupart de pays ont une économie moyenne. En revanche, on constatera que le Singapour et le Hong-Kong se trouvent en Asie, et qu'ils ont une richesse par habitant aussi important que les pays occidentaux.

Dans le deuxième graphique ci-dessus, on voit que chez les pays où le soutien social est important l'économie est plus élevée que chez les pays où le soutien est insuffisant ou moyen. Par contre, le Mali et l'Uganda sont des pays où les personnes ont un soutien social important mais qui ont un PIB par habitant très faible.

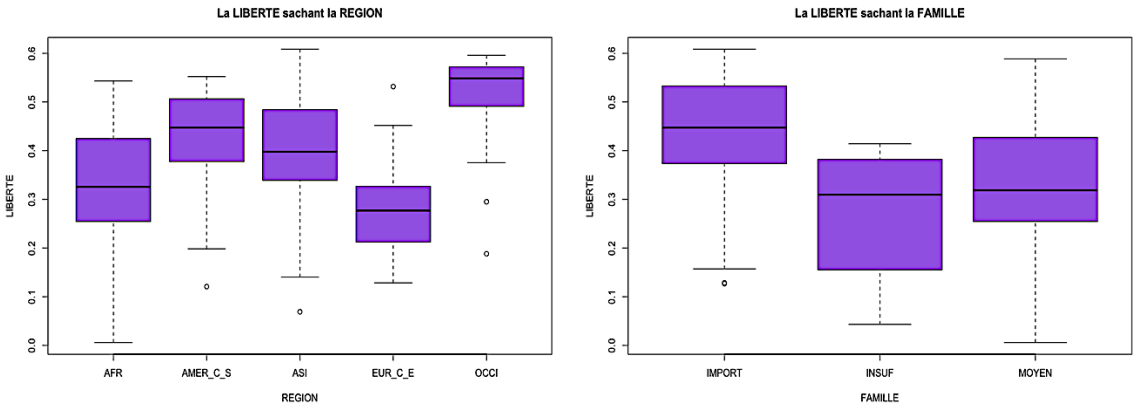




Dans le premier graphique ci-dessus, on constate que chez les pays occidentaux l'espérance de vie est plus importante que dans le reste de régions. Un autre élément à signaler est que chez les pays africains l'espérance de vie est très faible si on le compare à la moyenne globale des pays. En Asie, les pays se caractérisent par une espérance de vie moyenne si on ne prend pas en compte, par exemple, le Hong-Kong et le Singapour qui ont une espérance de vie très élevé. Dans ce sens, on peut aussi signaler qu'en Amérique du Sud et en Europe Centrale et de l'Est l'espérance de vie est moyenne.

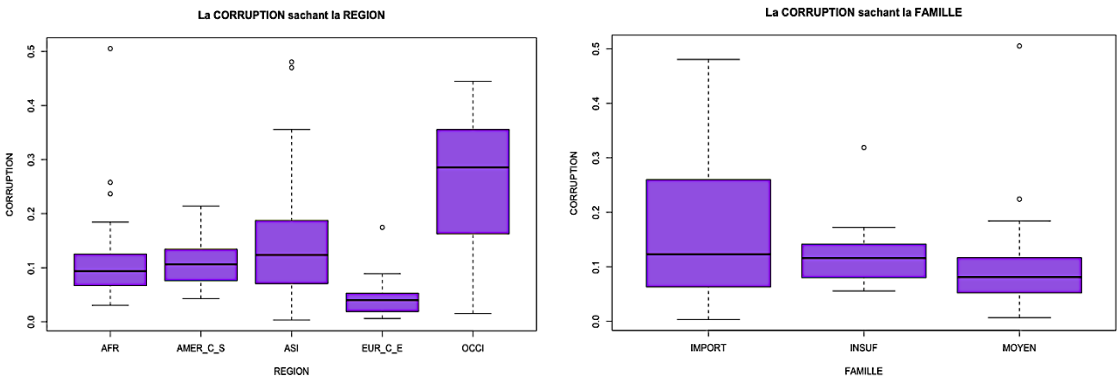
Dans le deuxième graphique ci-dessus, on constate que chez les pays où le soutien social est important l'espérance de vie est très élevée si on le compare avec les pays présentant un soutien insuffisant ou moyen. Par contre, on peut aussi voir que les pays comme le Mali,

le Sud-Afrique et l'Uganda, présentant un soutien social important, ont une espérance de vie médiocre.



Dans le premier graphique ci-dessus, on constate que chez les pays occidentaux la liberté de choix de vie est bien plus importante que dans le reste du monde. Cependant on trouve qu'il y a deux pays, le Portugal et le Chypre où le choix de vie est faible. Dans ce sens, on voit aussi que chez les pays de l'Europe de l'Est et Centrale le choix de vie est le plus faible que dans les autres régions si on ne prend pas en compte la Slovénie où les personnes se sentent plus libre à faire un choix de vie personnel.

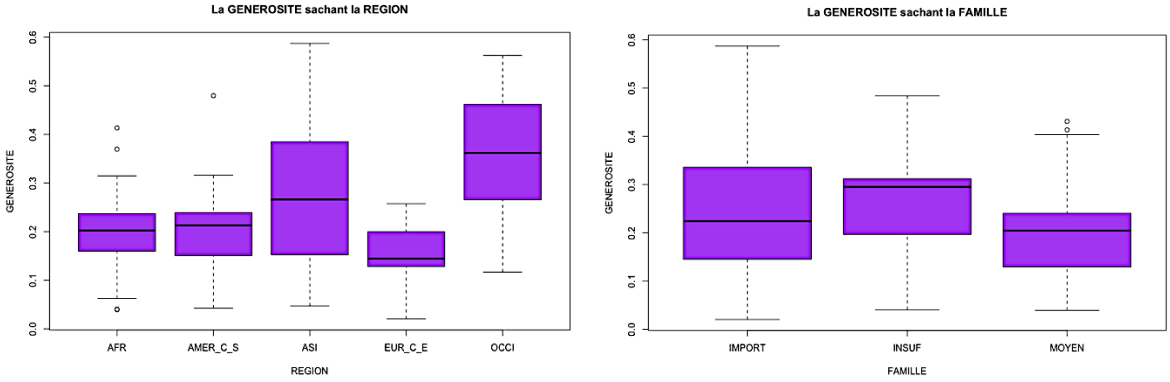
Dans le deuxième graphique ci-dessus, les pays ayant un soutien social important ont une liberté de vie très importante, à l'exception de l'Ukraine et la Mauritanie par exemple. Contrairement, pour la plupart de pays se caractérisant par un soutien insuffisant la liberté de choix de vie est moyenne. Dans ce sens, on voit aussi que chez les pays où le soutien social est moyen les personnes ont une liberté de choix de vie moyenne.



Dans le premier graphique ci-dessus, on constate que chez les pays occidentaux la perception de la corruption est assez élevée par rapport aux autres régions, qui en moyenne, ont une perception faible. Par contre, on trouve aussi des pays comme, par exemple, le

Rwanda, le Qatar, le Singapour et les Émirats Arabes Unis qui ne se trouvent pas dans les pays occidentaux mais qui présentent une perception de la corruption élevée.

Dans le deuxième graphique ci-dessus, la plupart des pays où le soutien social est important, la perception de la corruption est à peu près la même que dans les pays ayant un soutien insuffisant et moyen. Par contre, on constate que chez le pays avec un soutien social important, certains ont une perception de la corruption élevée. De plus, dans les pays comme le Rwanda, la Géorgie, le Laos, le Japon et l'Estonie ont une perception de la corruption élevée même s'ils n'appartiennent pas aux pays où soutien social est important.

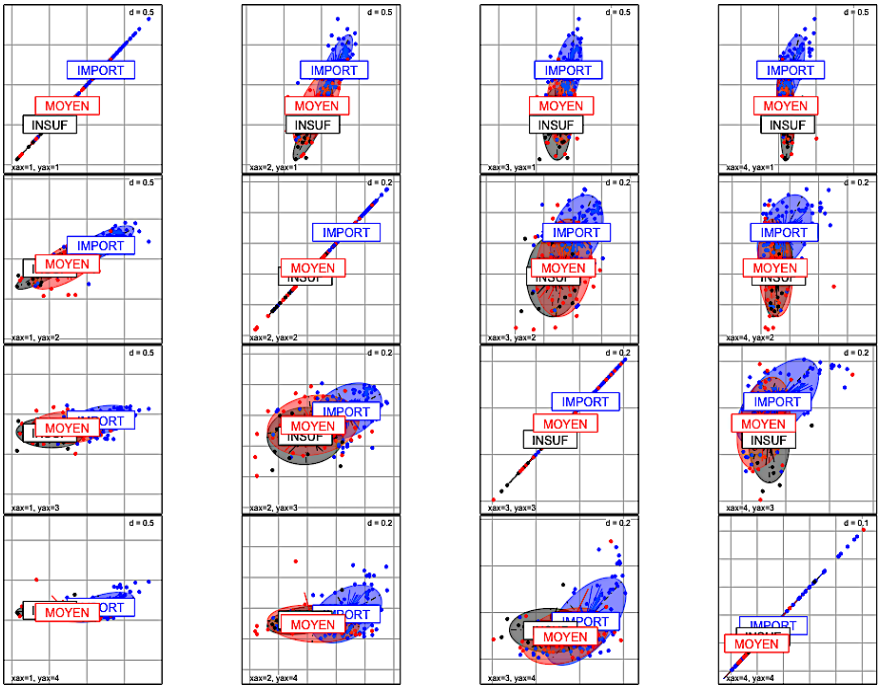


Dans le premier graphique ci-dessus, on constate qu'en Occident les personnes font des dons d'argent à un organisme de bienfaisance bien plus fréquemment que dans le reste du monde. Dans ce sens, on voit aussi que dans les pays comme l'Haïti, le Chili, le Kenya et le Maurice il y a des dons d'argent assez fréquent même s'ils ne se trouvent pas en Occident. Par opposition, en Europe de l'Est et Centrale les dons d'argent se font moins souvent que dans les autres régions.

Dans le deuxième graphique ci-dessus, chez les pays où le soutien social est important il y a à peu près la même quantité des pays où les dons se font assez fréquemment que des pays où les dons se font moins fréquemment. De plus, chez la plupart des pays où le soutien social est insuffisant les dons d'argent se font plus fréquemment que dans les autres catégories. Par contre, on voit aussi que les pays se caractérisant pour un soutien social moyen comme le Laos, le Kenya et le Cambodge, les dons d'argent se font assez souvent.

b. ANALYSE FACTORIELLE DISCRIMINANTE

Faisons une analyse de la dispersion des individus par rapport à la variable FAMILLE :



On constate que les différences de dispersion et d'orientation sont faibles entre les 3 modalités de la variables FAMILLE. Aucune des mesures ne permet de les distinguer les 3 modalités sans risque de confusion.

Et si on fait une prédiction, a-t-on une bonne approximation ?

Ici, on va prendre 110 pays au hasard sur 140 pays et on fera une prédiction sur les 30 pays non pris et on verra à quelle modalité de la variable REGION et FAMILLE appartiennent chacun des pays.

On obtient le tableau ci-dessous :

REGION		Appartenance réelle		
		IMPORT	INSUF	MOYEN
Appartenance prédite	IMPORT	**19**	0	4
	INSUF	0	**1**	1
	MOYEN	1	1	**3**

Le résultat est décevant car on obtient 7 des pays mal prédits qui représentent 23% des valeurs prédits.

FAMILLE		Appartenance réelle				
		AFRI	AMER_C_S	ASI	EUR_C_E	OCCI
Appartenance prédite	AFRI	**5**	0	2	0	0
	AMER_C_S	0	**2**	3	1	0
	ASI	4	0	**2**	1	0
	EUR_C_E	0	0	0	**2**	1
	OCCI	0	0	0	0	**7**

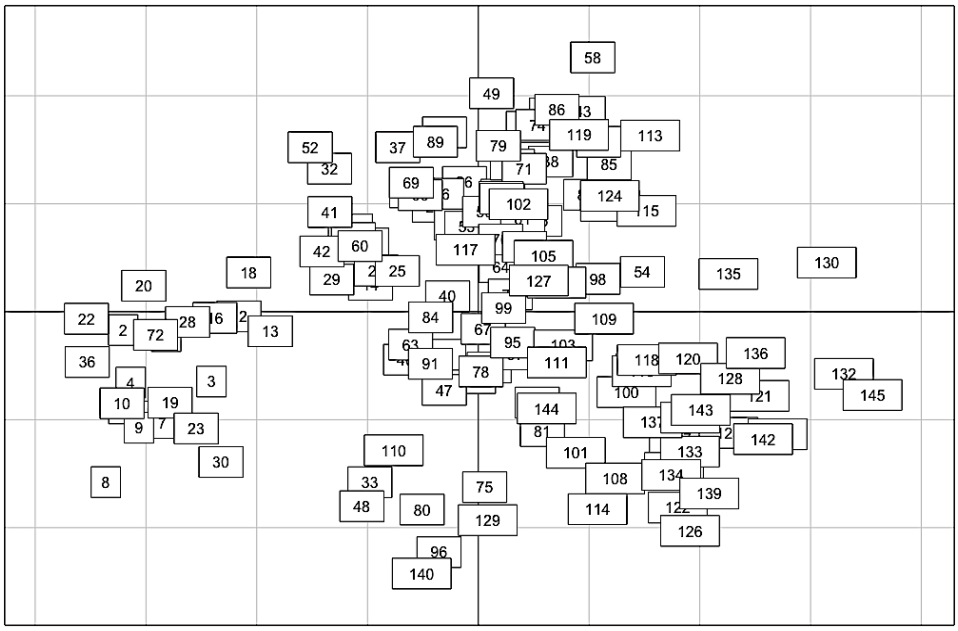
Le résultat est décevant car on obtient 12 des pays mal prédits qui représentent 40% des valeurs prédits.

On pourrait aussi répéter cette expérience et analyser si les valeurs obtenues précédemment sont dû au hasard ou si elles ont toujours le même pourcentage d'erreur de prédiction, c'est-à-dire, 23% et 40% pour les variables REGION et FAMILE respectivement.

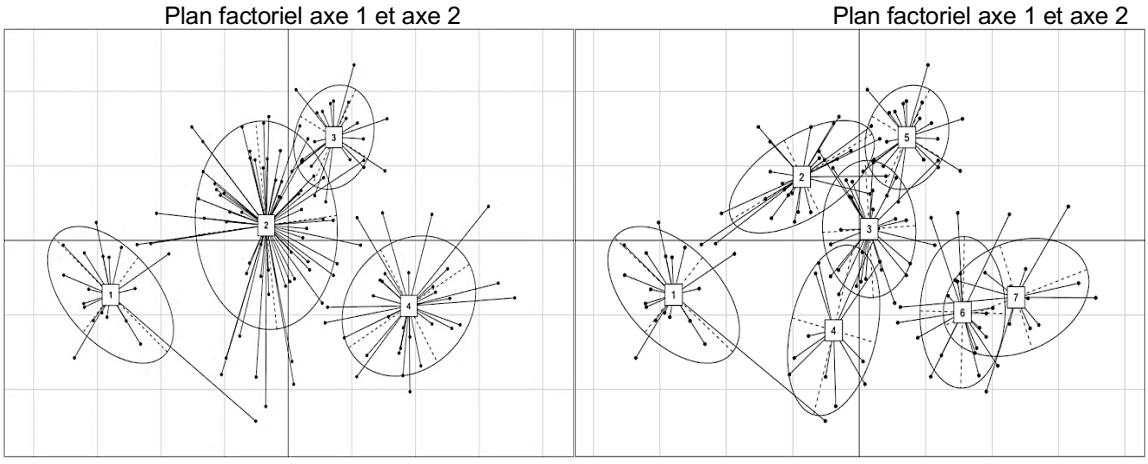
III. CLASSIFICATION HIERARCHIQUE

Rappelons la distribution des pays :

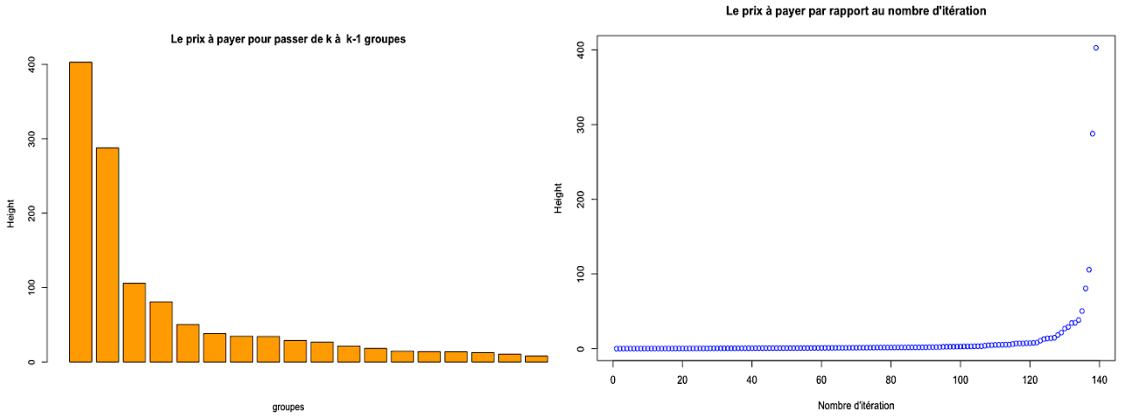
Plan factoriel axe 1 et axe 2



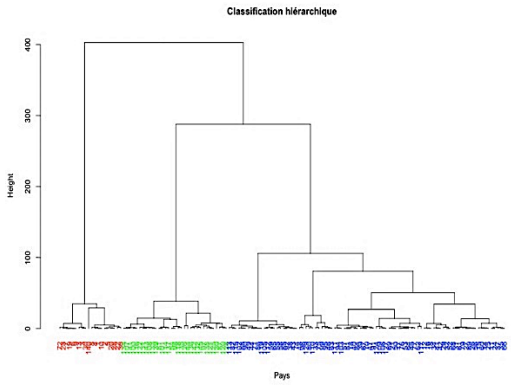
Voyons ci-dessous un regroupement des individus en 4 et 7 groupes :



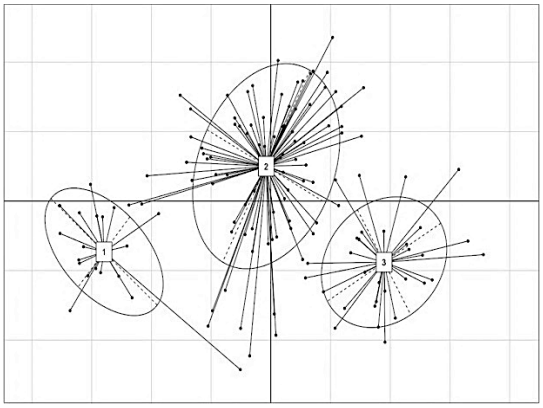
Quelle est le nombre de groupe optimal ?



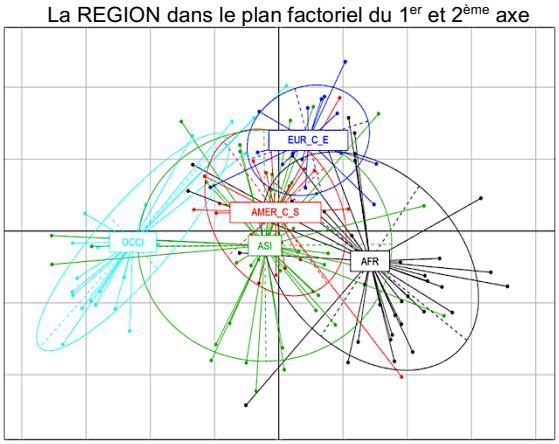
On voit que l'idéal serait de regrouper les pays en 3 groupes.



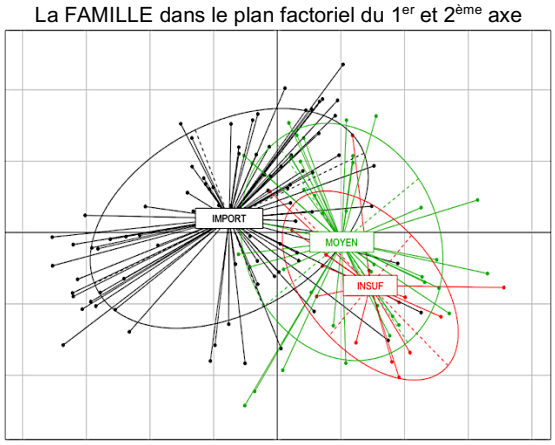
Plan factoriel axe 1 et axe 2



La REGION dans le plan factoriel du 1er et 2ème axe



La FAMILLE dans le plan factoriel du 1er et 2ème axe

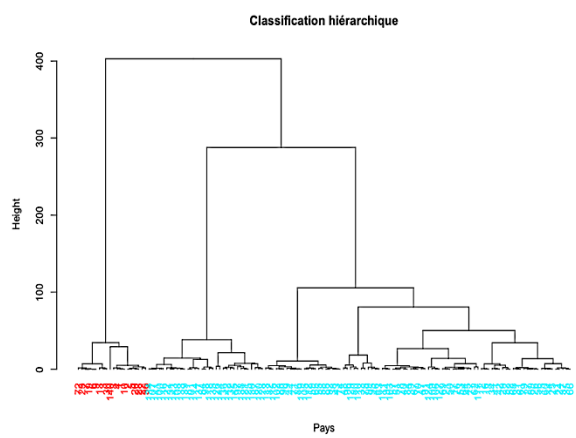


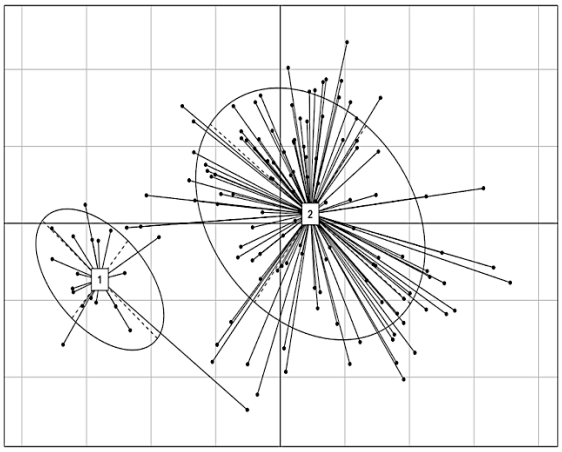
Que peut-on dire sur ces 3 groupes ?

Ici, *le groupe 1* contient surtout des pays occidentaux et les pays présentant un soutien social important. *Le groupe 3* contient surtout les pays de l'Afrique et ne peut rien dire sur le soutien social. De plus, on ne peut pas assurer que *le groupe 2* représente une région car il contient des pays appartenant à des différentes régions. Dans ce sens, on constate aussi qu'on ne peut pas dire que *le groupe 2* représente un type de soutien social.

Et si on regroupe les pays en 2 groupes ?

Plan factoriel du 1er et 2ème axe





Ici, le groupe 1 pourrait être interprété comme la zone idéale à vivre et le groupe 2 comme la zone moins idéale. De plus, ceci a déjà été dit lorsqu'on a fait une ACP.

CONCLUSION

Les différentes méthodes d'analyses, qui ont été fait dans le but d'expliquer la qualité de vie à l'aide des 7 variables, nous ont permis d'apprécier les meilleurs pays pour vivre. La plupart de ces pays se trouvent en Europe de l'Ouest et en Amérique du Nord. En Asie, on peut aussi trouver des pays comme le Singapour ou le Qatar se caractérisant pour une qualité de vie élevé. En revanche, la plupart de pays présentant une qualité de vie mauvaise se trouvent en Afrique.

Pour aller plus loin, les résultats obtenus pourraient être améliorer si on avait un tableau des données avec le taux de criminalité, le taux de chômage, la réussite professionnelle, la qualité de l'éducation, etc. Malheureusement on n'a pas pu trouver un tableau contenant autant d'information pour ainsi interpréter plus précisément et plus efficacement la qualité de vie dans le monde.

REFERENCES

1. https://fr.wikipedia.org/wiki/World_Happiness_Report

2. https://s3.amazonaws.com/happiness-report/2016/HR-V1_web.pdf (page 17)

3. https://www.kaggle.com/koki25ando/data-analysis-of-world-happiness-report/data (Les données de la variable région sont fausses, on a dû les corriger)