Université Gustave Eiffel

Master 2 Data Analyst

# Projet Analyse de données

Réalisé par

**Kadidiatou Bah, Jivaldo Julien, Ilyes Hadi**

Octobre 2021

# Table des matières

# Introduction et problématique

Les questions sur le bien-être ou encore le bonheur sont de plus en plus importantes à l'échelle du monde. En effet, nous sommes heureux lorsque plusieurs aspects de notre vie sont en accord avec la perception que l'on a du bonheur. Ces aspects font souvent référence à la vie économique, sociale ou encore psychologique à laquelle nous nous identifions.
Ainsi, pour évaluer le bonheur des différents pays du monde, le "World happiness report '' édité par le réseau de solutions pour le développement durable des Nations unis, essaye tant bien que mal de mesurer le bonheur des personnes au sein des différents pays.

Pour ce faire, elle utilise entre autres des indicateurs tels que : le PIB par habitant, l'espérance de vie, l'indice de générosité, la perception de la corruption gouvernementale ...
Le premier rapport a été publié en Avril 2012 dans le cadre de la réunion des Nations unis intitulée « Bien être et Bonheur ». Depuis cette date, le World happiness report est publié chaque année et classe les pays des plus heureux au moins heureux.

Nous avons décidé d'utiliser dans le cadre de ce projet, l'un de ces rapports, celui de l'année 2017 pour effectuer notre étude.
En effet, au moyen de la méthode dite d'ACP (Analyse en composante principale), nous essayerons de caractériser l'hétérogénéité entre les différents groupes de pays en fonction des variables, ce qui nous permettra ainsi de dégager les principales tendances issues de ces jeux de données.
Notre étude sera structurée autour de quatre grandes parties :

- Une première partie qui portera sur la description de la base de données ainsi que des statistiques descriptives pertinentes.
- Une partie sur l'analyse des données avec notamment une présentation de la méthode retenue et les éventuels traitements réalisés.
- Ensuite une présentation des résultats et les différentes interprétations qui en résultent.
- Enfin une conclusion sur le résultat global trouvé.

# Statistiques Descriptives

La base sur laquelle nous travaillons est une base composée de 155 observations représentant chacun un pays du monde. Pour ces 155 observations, la base de données possède 8 variables distinctes. Une variable **Région** qui va désigner la région où se situe le pays. Cette variable peut prendre 7 valeurs qui sont les suivantes : Africa, Arab States, Asia and Pacific, Europe, Middle East, North America, South/Latin America. Nous avons ensuite une variable **Happy.Score** , cette variable va représenter le score de bonheur. Cette variable est comprise entre 0 et 10 et correspond aux moyennes des réponses des individus questionnés dans les différents pays, la question est "Comment évalueriez- vous votre bonheur sur une échelle de 0 à 10". Nous précisons que les individus questionnés proviennent d'un échantillon représentatif à l'échelle nationale pour chaque pays. La troisième variable est la variable **PIB** qui va représenter la production intérieure brut par habitant pour chaque pays, la variable suivante est la variable **Family** (le soutien sociale)**,** ensuite nous avons la variable **Expect.Life** pour l'Esperance de vie. Une variable **Freedom** est également présente. S'en suit la variable **Generosity** (la générosité) ainsi que la variable **Corruption** (la confiance aux institutions)**.** L'agrégation des six variables précédentes, nous donne la première variable **Happy.Score,** ces 6 variables représentent donc leurs contributions aux scores de bonheur pour chaque observation.

**L'annexe 1** est une carte qui représente tous les pays du monde en fonction de la variable Happy.Score c'est-à-dire le score de bonheur de chaque pays. Chaque pays est coloré en fonction de son score. Les pays qui ont une couleur qui tendent vers rouge auront un happy score plus élever que ceux qui tendent vers le jaune.
On peut voir que les pays dont le score est plutôt faible se situent plutôt vers les pays du continent Africain et l'Asie de l'Ouest. A contrario les pays ayant un score de bonheur élevé se situent au niveau de l'Europe, de la zone Pacifique et de l'Amérique du Nord et du Sud. Le Moyen-Orient ainsi que les pays Arabe ont des couleurs qui varient, les scores de bonheur varient fortement dans ces régions du monde.

**L'annexe 2** est une figure représentant les "boxplots" des différentes variables de notre base de données, nous allons ici pouvoir analyser les 1er quartiles, les médianes, et les troisièmes quartiles. La variable Happy.Score à une médiane de 5,279 c'est-à-dire que 50% de nos observations seront soit au-dessus soit en dessous de cette valeur, le 1er quartile est de 4,505 donc 25% des plus faibles observations se trouve en dessous de cette valeur et le 3ème quartile à une valeur de 6,101 donc 25% des observations les plus fortes se trouvent au-dessus de cette valeur. Le minimum pour cette variable est de 2,693 et le maximum est de 7,537. A la suite de l'analyse de cette variable, nous pouvons dire qu'elle fluctue énormément en fonctions des pays, les scores de bonheurs diffèrent beaucoup dans cette base de données et ces scores ont l'air d'être équitablement dispersé dans l'ensemble de la base de données.

Une matrice de corrélation a été réalisée (**annexe 3** ) , grâce à cette matrice de corrélation nous allons pouvoir analyser la façon dont sont corrélées ou non les différentes variables entre elles. Les corrélations qui nous intéressent particulièrement sont les corrélations de la variable Happy.Score avec les autres variables qui la composent. Nous rappelons que la variable Happy.Score est l'agrégation des autres variables de la base de données mis à part la variable Region. Nous pouvons voir que la plus forte corrélation est la corrélation entre la variable Happy.Score et le PIB par habitant cette corrélation s'élève à 0,81, cela veut dire que le PIB par habitant explique une grande part de la variable Happy.Score, plus le PIB par habitant sera élevé plus le score de bonheur est élevé. La variable Familly et Expect.Life sont également positivement et fortement corrélés à la variable Happy.Score. La corrélation entre la variable Expect.Life et Happy.Score est de 0,78 cela veut dire que l'espérance de vie explique une grande part du score de bonheur, la corrélation entre la variable Family et la variable Happy.Score est de 0,75, cela veut dire que la Famille explique aussi une grande part du score de bonheur. Il y a ensuite la variable Freedom dont la corrélation s'élève à 0,57 avec la variable Happy.Score , cette variable explique donc moyennement le score de bonheur. La variable Corruption est quant à elle corrélé à notre variable Happy.Score à hauteur de 0,43. La variable qui explique le moins la variable Happy.Score est la variable Generosity, leurs corrélations ne s'élève qu'à une valeur de 0,16. On observe d'autre corrélation intéressante entre les variables notamment une corrélation entre la variable expect.Life et PIB , cette corrélation s'élève à 0,84. On a donc une bonne corrélation entre l'espérance de vie et le PIB par habitant

**L'annexe 4** est un graphique en barres représentant la moyenne de la variable Happy.Score en fonction des régions du monde, nous pouvons voir que la moyenne de la variable Happy.Score la plus élevée est celle l'Amérique du nord avec une valeur de 7,15, vient ensuite l'Europe avec une valeur de 6,03, puis l'Amérique du Sud avec une valeur de 5,96, vient ensuite le Moyen-Orient avec une valeur de 5,37 puis l'Asie et le Pacifique avec une valeur de 5,36. En avant dernière position on retrouve les pays Arabes avec une valeur pour la variable Happy.Score de 5,04 et enfin en dernière position l'Afrique avec une valeur de 4,08. Ces valeurs confirment l'analyse faite sur la carte du monde, les pays ayant le plus haut score de bonheur sont l'Amérique ainsi que l'Europe, et la zone ayant le score de bonheur le moins élevé est l'Afrique.

# Méthode de l'analyse en composante principale (ACP)

Nous utiliserons la méthode d'ACP pour l'analyse du jeu de données dont nous disposons. Cette méthode s'y prête bien pour différentes raisons que nous mentionnerons ci-après :

En effet, l'analyse en composante principale permet l'analyse des tableaux croisant des individus (qui dans notre cas correspond au différents pays) en lignes et des variables spécifiquement quantitatives en colonnes. Elle permet de traiter des données multidimensionnelles et de ce fait, elle caractérise l'hétérogénéité des différents groupes constitués. Ce qui permet de synthétiser l'information contenue dans une grande masse de données.

Le but ici, est de pouvoir rendre compte de la variabilité des indicateurs de bonheur entre les pays. Les individus actifs seront donc les 155 pays du classement. Par ailleurs, il n'y aura pas d'individus supplémentaires. Pour effectuer l'ACP nous avons gardé les 6 variables permettant de construire l'indicateur de bonheur à savoir : Le PIB par habitant, l'espérance de vie, la liberté, la confiance aux institutions (corruption), la générosité et le soutien social (Family). Enfin, en partant du fait que l'indicateur de bonheur soit une agrégation de toutes les variables nous avons décidé de l'intégrer en tant que variable supplémentaire quantitative.

# Méthode de classification

Nous effectuerons également une classification sur nos données. Elle nous permettra de classer les pays en groupes homogènes. En d'autres termes, cette méthode nous permettra de regrouper les pays qui se ressemblent et de séparer ceux qui sont éloignés sur l'ensemble des variables.

Nous utiliserons la méthode de classification non supervisée et plus spécifiquement la méthode des K-means et celle des CAH.

La méthode K-means nous permettra de définir le nombre de classe optimal et ainsi définir la moyenne de toutes les variables pour chaque classe. L'avantage de cette méthode est que le centre de gravité de chaque classe est recalculé à chaque fois qu'un nouvel individu est introduit dans la classe et non pas une fois seulement que tous les individus ont été affectés.

La méthode CAH (classification ascendante hiérarchique) quant à elle permet de fournir un ensemble de partition de nos individus(pays) en des classes de moins en moins fines obtenues par regroupements successifs de parties. Nous présenterons notre classification hiérarchique par un dendrogramme.

# Les résultats de l'analyse en composante principale

**Inertie et nombre de dimensions :**

L'inertie totale va nous permettre d'indiquer le nombre de dimensions dont nous avons besoin pour notre analyse. L'histogramme des valeurs propres en pourcentage d'inertie nous indique que les deux premières dimensions représentent respectivement 49% et 22,3% de l'inertie totale. Donc, près de ¾ de l'information est apportée par ces variables. Nous avons donc choisi de concentrer l'analyse uniquement sur les deux premières dimensions. (Voir Annexe 5, Annexe 6).

**Études des Individus et des variables :**

Lorsque l'on regarde notre nuage de point (annexe 7), globalement nous observons qu'il y a une concentration à gauche des pays en bas du classement et à droite des pays haut du classement. On peut voir que par exemple le premier et le dernier s'opposent sur l'axe 1. Ensuite, il y a une seconde opposition par rapport à l'axe 2 par exemple entre le 114ème et le 87ème du classement qui sont respectivement en bas et en haut de l'axe 2.

Concernant le cercle des corrélations (annexe 8), l'ensemble des variables sont concentrés à droite. La totalité des variables sont plutôt corrélées au premier axe. Enfin, concernant le second axe, seulement deux variables semblent y êtres corrélées.

Avec, le premier axe, nous pouvons voir qu'il oppose clairement les pays en termes de PIB par habitant, l'espérance de vie et le soutien familial. En effet, ces variables sont positivement corrélées au premier axe. Les pays situés à droite du nuage seront alors des pays dans lesquels ces indicateurs sont plutôt favorables. Pour ceux situé à gauche du nuage de point, ils auront plutôt des valeurs faibles pour ces indicateurs. On peut donc supposer que les 3 indicateurs joueront davantage que les autres sur le Happy score. On a alors ici un effet taille opposant les pays dans lesquels il fait bon vivre et ceux dans lesquels la population n'est pas forcément heureuse.

Le second axe va opposer les pays en fonction du niveau de générosité de la population mais aussi légèrement en fonction de la corruption et du niveau de liberté. Ces variables sont positivement corrélées au second axe. Les individus situés en haut de l'axe auront des indicateurs plutôt favorables et inversement.

Pour mieux comprendre et interpréter le second axe, nous allons analyser les indicateurs de L'Ouzbékistan et l'Italie. Le second axe oppose dans notre nuage de point, le 47ème et 48ème pays du classement (annexe 11) à savoir respectivement L'Ouzbékistan et l'Italie. Les pays ont tous les deux un score similaire, ils sont donc à droite du plan en d'autres termes, ils ont un score plutôt correct. Néanmoins, lorsque l'on regarde nos données, on voit que les pays s'opposent nettement sur certains indicateurs. Ce ne sont pas les mêmes indicateurs qui vont expliquer les scores des deux pays. L'Italie va avoir de meilleurs indicateurs en termes de PIB par habitant et d'espérance de vie et l'Ouzbékistan va avoir de meilleurs indicateurs en termes de liberté, générosité et de confiance aux institutions. Donc, on peut supposer que le second axe oppose les pays en fonction de la générosité, de la liberté et de la confiance aux institutions. Effectivement, le second axe va nous permettre de comprendre quels facteurs jouent le plus sur les scores des pays. Dans notre exemple, L'Ouzbékistan et l'Italie se situent respectivement au-dessus et au-dessous du second axe. Donc ce qui oppose les deux pays c'est la combinaison d'indicateurs qui favorise le score de bonheur à savoir d'un côté les indicateurs plus orientés sur le confort sociale du pays (générosité, confiance aux institutions, liberté) et de l'autre des indicateurs sur le confort de vie (PIB par habitant, l'espérance de vie et le soutien sociale).

Pour discuter ces résultats, nous pouvons parler de la qualité de représentation des variables (annexe 9). Dans la première dimension, les variables sont plutôt bien représentées. Seulement la générosité, le niveau de corruption et la liberté sont mal représentés dans cet axe. Dans le second axe, la générosité est la seule variable bien représentée donc, on peut remettre en cause notre précédente conclusion et dire que le second axe oppose en réalité les pays en fonction de l'indicateur de générosité. Ce résultat prend d'autant plus de sens lorsque l'on étudie les relations entre variables (annexe 3). La générosité est la variable qui est la moins corrélée avec les autres. De plus, lorsque nous regardons la contribution aux axes, on peut voir que la générosité ne participe pas vraiment à la construction du premier axe, toutefois elle contribue majoritairement du second axe. Nous pouvons globalement confirmer les résultats de l'ACP, notamment ceux du premier axe. En effet, les variables les plus corrélées entre elles sont effectivement le PIB par habitant, le soutien social et l'espérance de vie. Par ailleurs, ce sont aussi les variables les plus corrélées au score de bonheur.

# Les résultats de la classification

Lorsque nous appliquons une K-means à nos données (annexe 12), nous observons selon le critère du coude, qu'il y' a potentiellement quatre classes. Les quatre classes ainsi constituées sont assez différentes lorsque nous mettons un accent sur la moyenne de chaque classe pour chaque variable. En effet, sur l'ensemble des variables, le groupe 3 à relativement les moyennes les plus élevées en comparaison aux trois autres groupes. Par opposition, nous observons que le groupe 4 à les moyennes les plus faibles relativement aux autres groupes.

Ces résultats semblent tout à fait cohérents car lorsque nous regardons la répartition des groupes dans notre table qui attribue à chaque pays sa classe d'appartenance (annexe 13), nous observons que le groupe 3 se situe en tête de liste et le groupe 4 en bas de liste, Sachant que les pays sont classés par niveau de score décroissant.

Pour faire le lien avec l'ACP, nous pouvons voir que les groupes 1 et 2 s'opposent en fonction de deux groupes de variable. Le premier groupe représente les variables corrélées à l'axe 1 (PIB par habitant, situation familiale et espérance de vie) et le second groupe représente celles corrélées à l'axe 2 (Générosité, niveau de corruption et liberté).

Lorsque nous appliquons maintenant une CAH, nous observons avec le dendrogramme (annexe 14), un niveau de découpage plutôt net jusqu'à quatre segments ce qui est en phase avec les résultats trouvés lors de la mise en place de notre K-means. Au-delà, le découpage n'est pas suffisamment marqué. En d'autres termes, les partitions de nos individus(pays) sont en des classes de moins en moins fines.

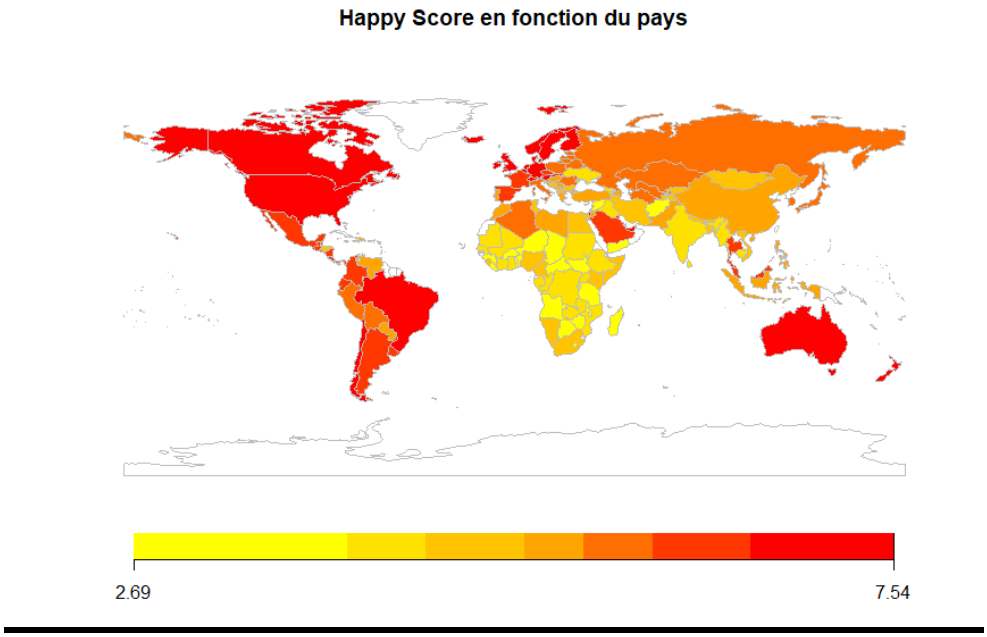
# Conclusion

La base de données que nous avons utilisée dans le cadre de ce projet provient du World happiness report. Cette base nous renseigne sur le score de bonheur de tous les pays du monde en prenant en compte différents aspects économiques, sociaux et psychologiques.

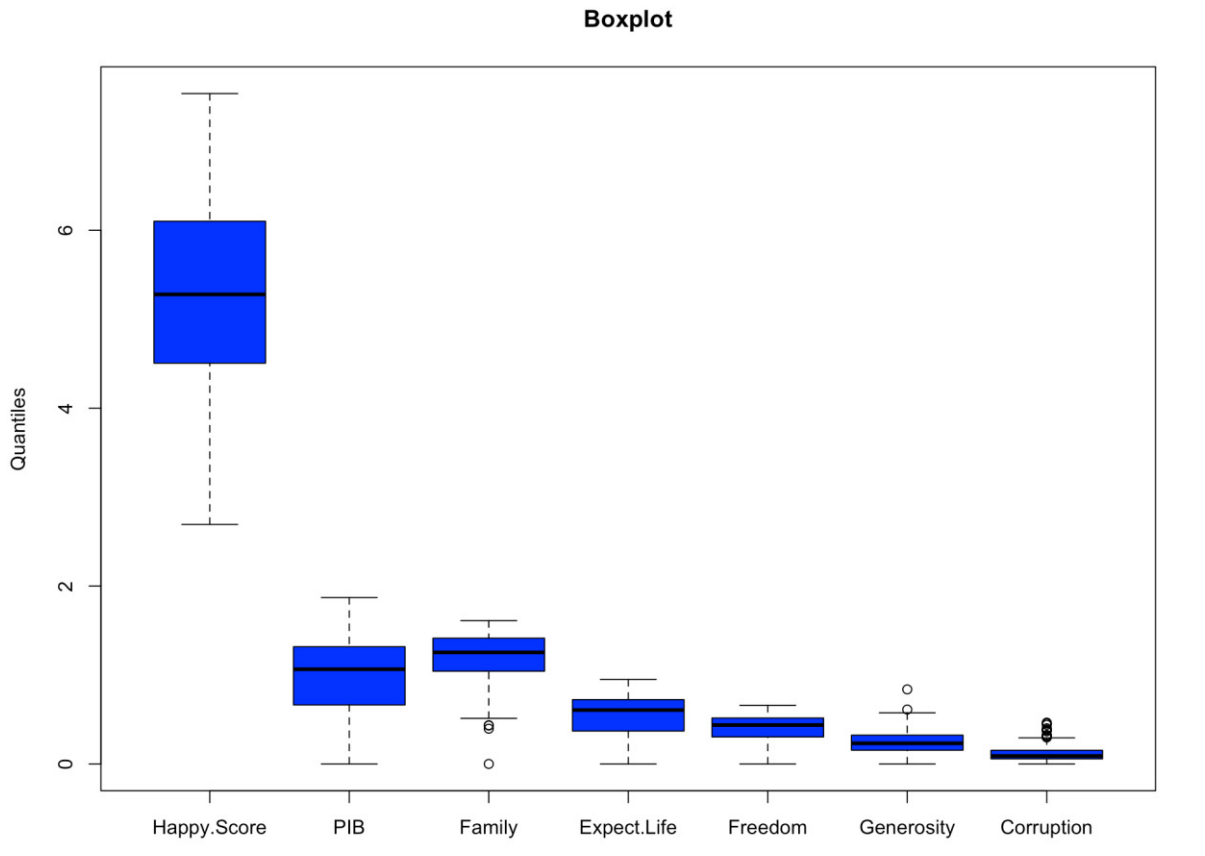
A l'issue de nos analyses et par la méthode d'analyse en composante principale et la méthode de classification, nous trouvons des résultats plutôt intéressants.

En effet, grâce à l'analyse en composante principale nous avons pu mettre en évidence que sur l'axe 1 les pays vont être opposés en fonction du PIB par habitant, de l'espérance de vie et la situation familiale. Sur l'axe 2, ils vont plutôt être opposés en fonction de l'indicateur de générosité. En effet, les pays situés à droite du plan vont avoir des indices de PIB par habitant, de l'espérance de vie et la situation familiale élevés et ceux à gauche auront des indices plutôt faibles pour ces indicateurs (Effet Taille). Quant à l'axe 2, les pays en haut du plan seront plus 'généreux' et ceux en bas le seront moins (Effet Forme). Concernant les variables, on observe une forte corrélation entre les variables qui constituent l'axe 1 (les indices de PIB par habitant, de l'espérance de vie et la situation familiale). Mais des corrélations plutôt faibles entre variables qui constituent l'axe 2 (Générosité, liberté et niveau de corruption).

# Annexes




*Annexe 1 : Carte du monde légendé en fonction de la variable Happy.Score*



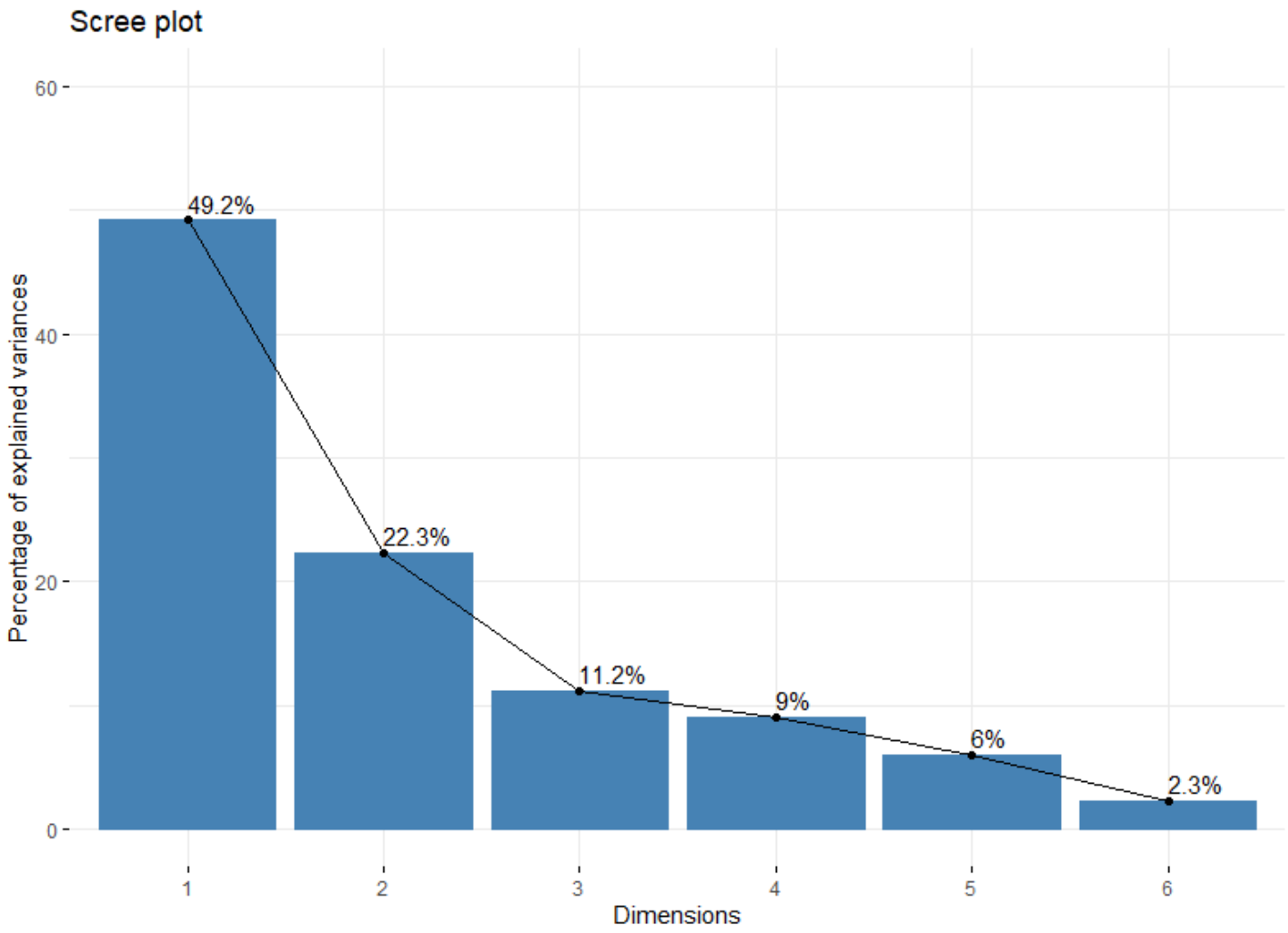

*Annexe 2 : Boxplot des variables de la base de données*


*Annexe 3 : matrice de corrélation*

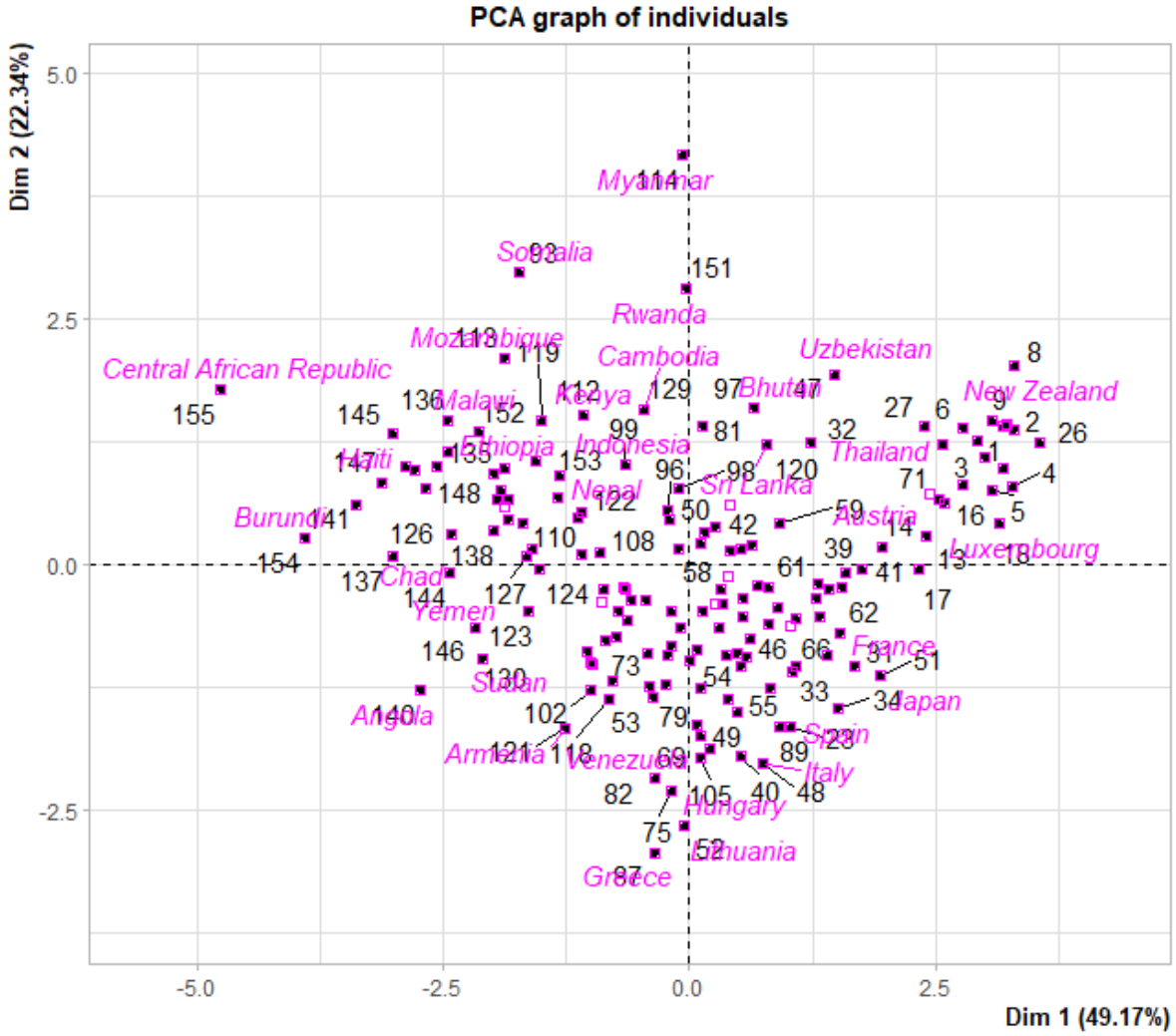

*Annexe 4 : Graphique en barres des moyennes de la variable Happy.Score en fonction de la region*

| | eigenvalue | percentage of variance |
|---|---|---|
| comp 1 | 2.95 | 49.17 |
| comp 2 | 1.34 | 22.34 |
| comp 3 | 0.67 | 11.17 |
| comp 4 | 0.54 | 9.04 |
| comp 5 | 0.36 | 6.01 |
| comp 6 | 0.14 | 2.26 |
| | cumulative | percentage of variance |
| comp 1 | | 49.17 |
| comp 2 | | 71.51 |
| comp 3 | | 82.69 |
| comp 4 | | 91.73 |
| comp 5 | | 97.74 |
| comp 6 | | 100.00 |

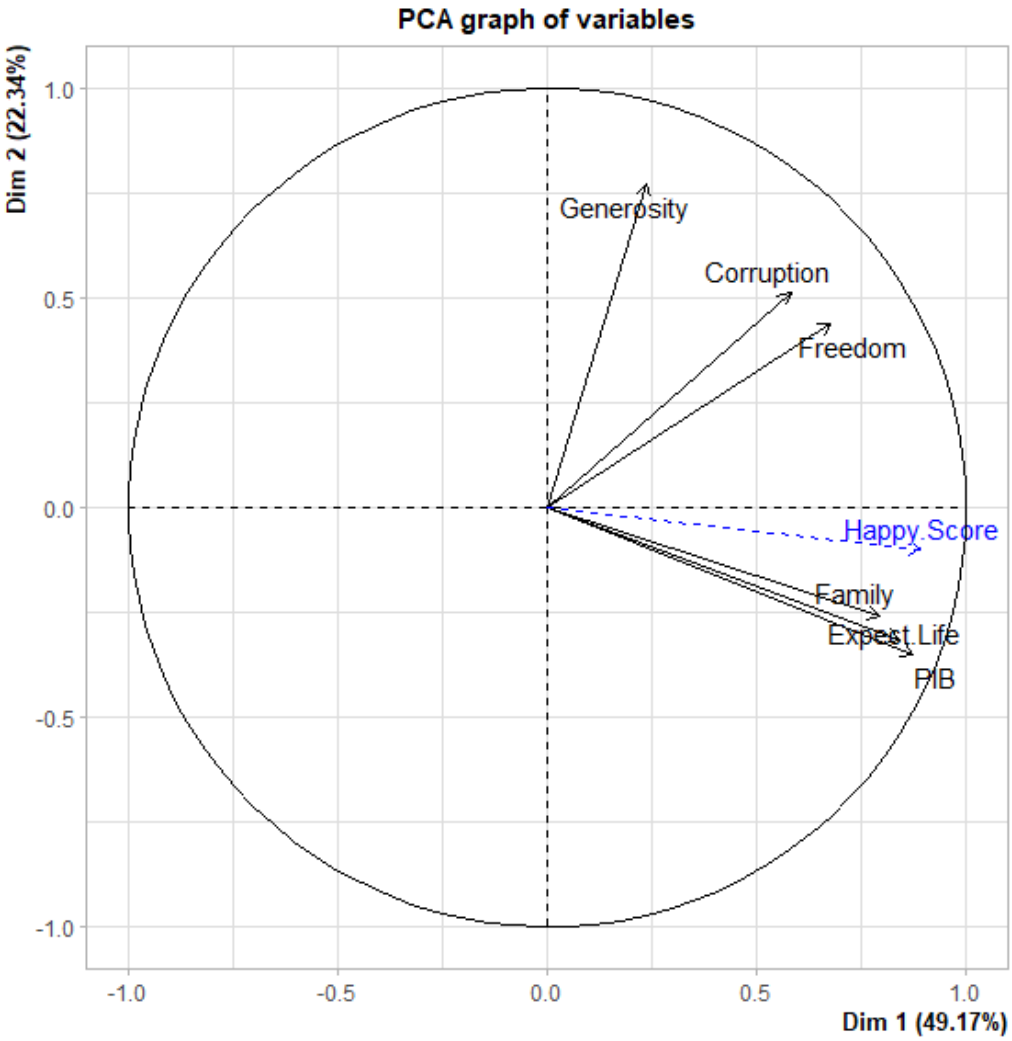
*Annexe 5 : Inertie totale*




*Annexe 6 : Histogramme des valeurs propres en pourcentage d'inertie*



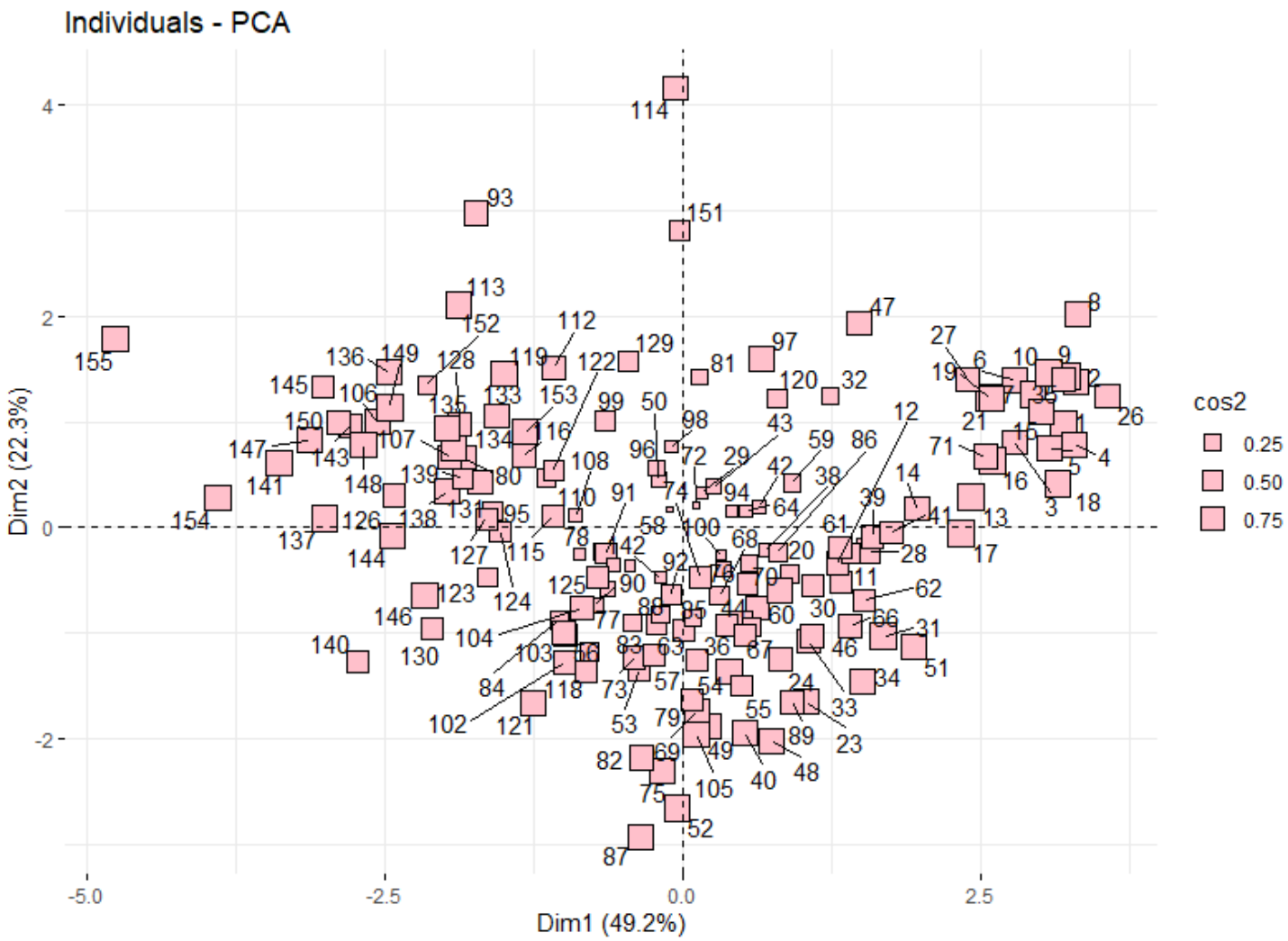

*Annexe 7 : Graphique des individus*




*Annexe 8 : Cercle des corrélations*

| | Dim.1 | Dim.2 | Dim.1 | Dim.2 | Dim.1 | Dim.2 |
|---|---|---|---|---|---|---|
| PIB | 0.87 | -0.35 | 0.76 | 0.12 | 25.87 | 9.07 |
| Family | 0.79 | -0.26 | 0.63 | 0.07 | 21.39 | 5.06 |
| Expect.Life | 0.84 | -0.32 | 0.70 | 0.10 | 23.88 | 7.48 |
| Freedom | 0.68 | 0.44 | 0.46 | 0.19 | 15.51 | 14.36 |
| Generosity | 0.24 | 0.77 | 0.06 | 0.60 | 1.90 | 44.41 |
| Corruption | 0.58 | 0.51 | 0.34 | 0.26 | 11.44 | 19.62 |

*Annexe 9 : Concaténation coordonnées, qualité de représentation et contribution des variables*



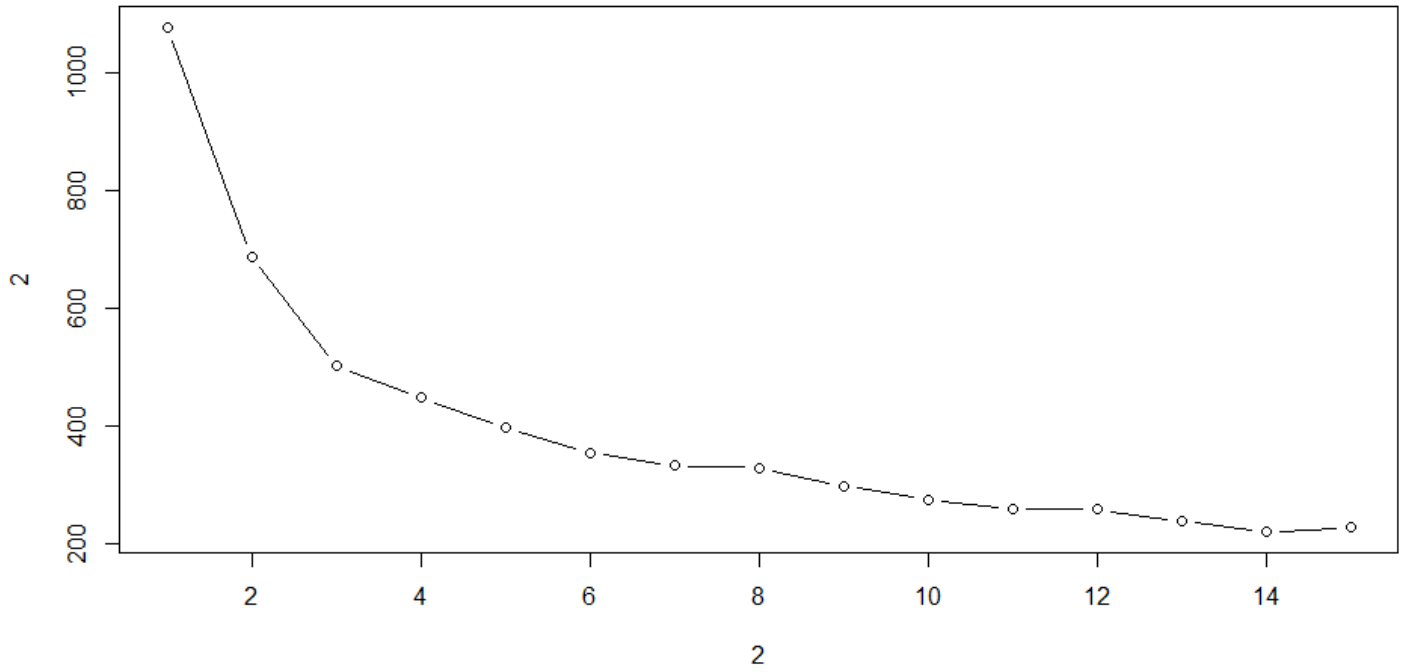

*Annexe 10 : Qualité de représentation des individus*

| Country | Region | Happiness.Rank |
|---|---|---|
| Uzbekistan | Asia & Pacific | 47 |
| Italy | Europe | 48 |

| Happiness.Score | Economy..GDP. | Family | Health..Life.Expectancy. |
|---|---|---|---|
| 5.97100019454956 | 0.786441087722778 | 1.54896914958954 | 0.498272627592087 |
| 5.96400022506714 | 1.39506661891937 | 1.44492328166962 | 0.853144347667694 |

| Freedom | Generosity | Trust..Government.Corruption. |
|---|---|---|
| 0.658248662948608 | 0.415983647108078 | 0.246528223156929 |
| 0.256450712680817 | 0.17278964817524 | 0.0280280914157629 |

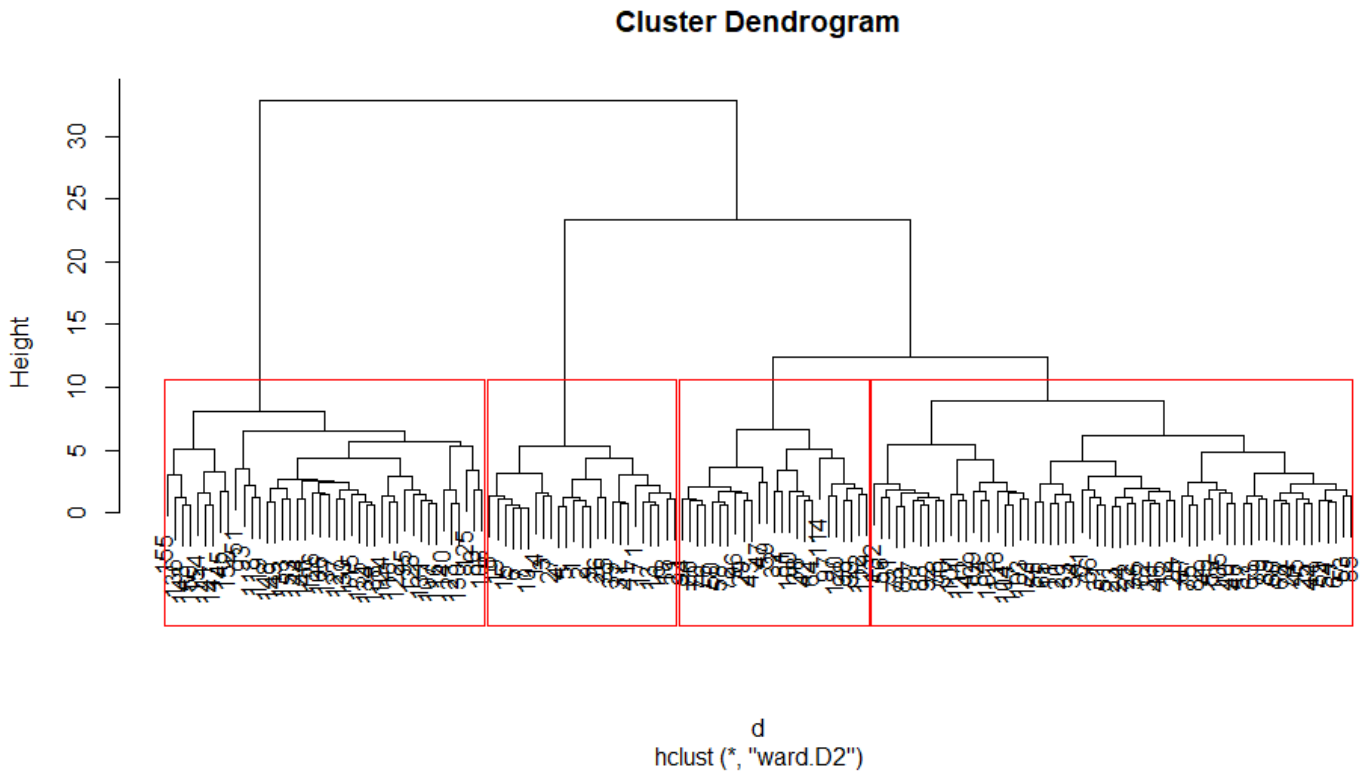
*Annexe 11 : Opposition individu 47 et 48 : Italie et Ouzbékistan*




*Annexe 12 : Répartition des classes avec la la méthode des K-means*

| Group.1 | Happy.Score | PIB | Family | Expect.Life | Freedom | Generosity | Corruption | fit.cluster |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.2167086 | 0.310322 | 0.3044157 | 0.3805422 | -0.1156444 | -0.514448585 | -0.4320568 | 1 |
| 2 | -0.2828950 | -0.579067 | 0.2545677 | -0.2360700 | 0.7683772 | 1.919737973 | -0.2525379 | 2 |
| 3 | 1.3710921 | 1.309757 | 0.8798638 | 1.0354591 | 1.0961925 | 0.843357583 | 1.7188414 | 3 |
| 4 | -1.1512292 | -1.218057 | -1.1496832 | -1.2668588 | -0.6206743 | -0.003674516 | -0.1606006 | 4 |

*Annexe 13 : Moyenne de chaque groupe pour chaque variable*




*Annexe 14: Classification Ascendante Hiérarchique: le Dendrogramme*