World Happiness and Freedom

Sofia Gervasoni

July 18, 2022

Abstract

This paper's aim is to cluster Countries with respect to their Happiness score and Freedom index. To reduce the dimesionality of dataset and starting analysing the similarities/dissimilarities between the different countries, PCA was introduced. Thanks to PCA it was possible to reduce the dimensionality of my dataset into two dimensions (two principal components), that already explained the 68% of the total variability. Moreover, it was possible to find out that the variable which mostly contribute in the explanation of the first dimension is the Happiness Score, whereas the second dimension is mostly explained by Dystopia residual. After that, using three different clustering techniques (hierarchical clustering, k-mean and k-medoid) it was possible to cluster the Countries in three groups: the happiest & most free Countries, the more-or-less happy & more-or-less free Countries and the saddest & least free Countries.

Contents

1 Goal of the analysis and a brief introduction **3**

2 Top 4 findings **3**

3 Data **4**
- 3.1 Why did I take only three variables from the data set *Freedom*? . . . 4
- 3.2 Descriptive analysis of the variables . . . 6
  - 3.2.1 Happiness Score . . . 6
  - 3.2.2 Lower Confidence Interval . . . 6
  - 3.2.3 Upper Confidence Interval . . . 6
  - 3.2.4 Economy (GDP per Capita) . . . 7
  - 3.2.5 Family . . . 7
  - 3.2.6 Health (Life expectancy) . . . 7
  - 3.2.7 Freedom . . . 7
  - 3.2.8 Trust (Government Corruption) . . . 7
  - 3.2.9 Generosity . . . 7
  - 3.2.10 Dystopia Residual . . . 7
  - 3.2.11 Personal Freedom (pf score) . . . 8
  - 3.2.12 Economic Freedom (ef score) . . . 8
  - 3.2.13 Human Freedom (hf score) . . . 8
- 3.3 Correlation . . . 9

4 Principal Component Analysis (PCA) **9**

5 Cluster analysis **14**
- 5.1 Hierarchical clustering . . . 14
- 5.2 K-Means . . . 17
- 5.3 K-Medoids . . . 21

6 Conclusions **24**

7 R Code **26**

8 Image appendix **33**

1 Goal of the analysis and a brief introduction

The aim of this paper is to find out how the factors of happiness and freedom of particular countries are affecting the well-being of inhabitants, and given these characteristics cluster the Countries. Different unsupervised learning technique will be used with this aim.

I started my analysis with the PCA that helped me reducing the dimensionality of the data. PCA was useful not only to reduce the dimesionality of my dataset, but also to start analysing the similarities and dissimilarities between the different countries (thanks to the score plot). With the loading plot, I was able to find out the correlation between the different variables and understand which contribute most to the first two principal components.

Then I started clustering the Countries and to do so I used three different techniques: hierarchical clustering, k-means and k-medoids. All these three methos pointed out more or less the same results, dividing the Countries in 3 clusters.

2 Top 4 findings

- In PCA the first two variables are enough to explain 68% of the total variance (78% is explained if we consider also the third component);
- The variable that contribute most to the first dimension is the *Happiness Score*, wheres the variable that contribute most to the second dimension is *Dystopia Residual*;
- Using the different clustering techniques, I was able to divide the countries in 3 different clusters based on their levels of happiness (and all the other variables that affect it) and freedom. It came out that the happiest Countries are also the one with the higher GDP per-capita, the higher Life expectancy (better health conditions) and the higher levels of freedom. But in terms of generosity the Countries in the least happy and least free countries cluster resulted more generous than the ones in the "middle cluster";
- The happiest countries are also the most free ones.

3 Data

To analyze the relationship between happiness and freedom and conduce my analysis on the different Countries in the world, I started from two different data sets that available on Kaggle: World happiness Report and Human freedom Index.

World happiness Report (2016), the first data set, gives us information about how economic factors (GDP), health, family, trust towards governments, generosity and life expectancy influence the happiness score. In this data set we have information about 157 countries within 13 variables. The different variables are summarized in three main scores which are personal freedom score, economic freedom score and human freedom score with values from 0 for the lowest degree of freedom to 10 being the maximum degree of freedom. The happiness scores and rankings use data from the Gallup World Poll. The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The scores are from nationally representative samples for the years 2013-2016 and use the Gallup weights to make the estimates representative. The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

The *Human freedom index* data set, the second data set used, contains information from 2016 to 2018 about freedom of 167 countries within 127 variables. The human freedom can be distinguished in personal (rule of law, freedom of expression, religion, security and safety, freedom of movement, freedom of association, identity and relationship) and economic (legal, trade, money, government and regulation) freedom. The Human Freedom Index gives each country a score from 0 to 10, wherein a score of 10 represents the most freedom and 0 represents no freedom at all, in each of the 82 indicators. These scores are carefully weighted and combined to determine the values for personal freedom and economic freedom, then those two values are averaged to determine each country's ultimate human freedom index score (see this link for more information about this index).

I merged the two data sets in order to draw relations between happiness and freedom. So, I obtained a data set with 137 countries (the intersection of both data sets) and 13 variables (I keep 10 variables from the first dataset and 3 from the second).

3.1 Why did I take only three variables from the data set *Freedom*?

As mentioned before, checking all the variables from the data set *Freedom*, I found out that all of them could be summarized in three independent variables: pf score (personal freedom), ef score (economic freedom) and hf score (human freedom).

This data set had information about different years, but I was analyzing only 2016, so that I kept only the first 162 rows by creating a new data set. In order to make it possible to analyze, firstly, I had to solve the missing values issue. Actually, this data set had 2081 missing values (see Figure 1) and most of them were summarized in 7 columns, so I dropped them off (they had only missing values).

After dropping these 7 columns, there still be 785 missing values located in different columns, distributed as shown in Figure 2. To avoid losing too many information, I decided to use the command *knnImputation()*, which helped me to easily attribute multiple missing values taking into consideration the correlation structure of the data.

Once the problem of missing data was solved, I started studying the correlation between variables. Particularly, I studied the correlation between the summary variables (the scores) and all the other variables, that are themselves a summary of different subcategories.

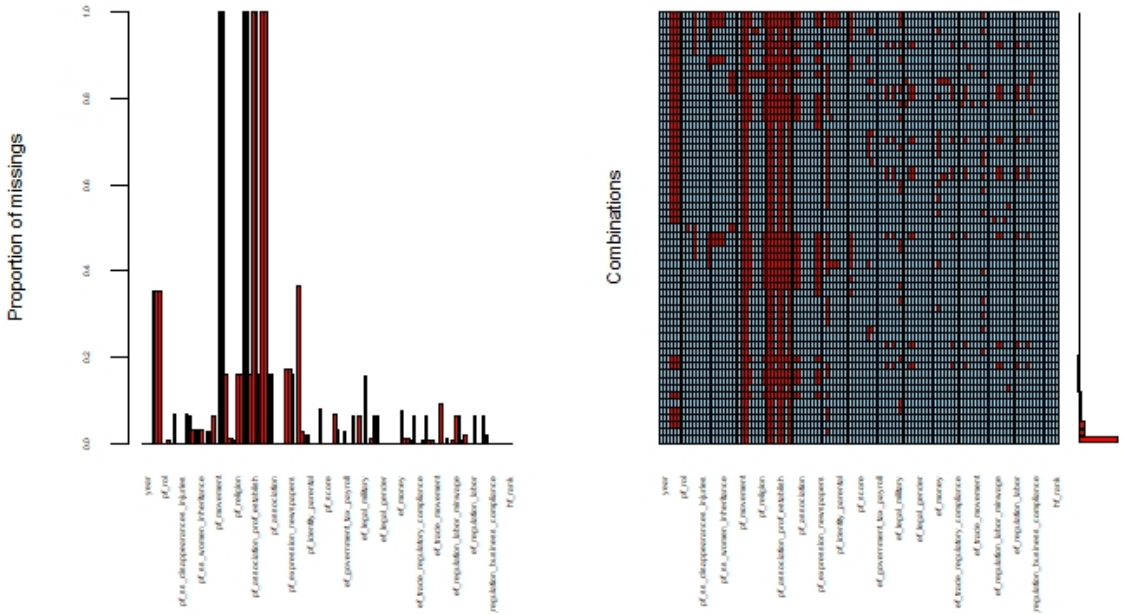

Figure 1: NAs distribution

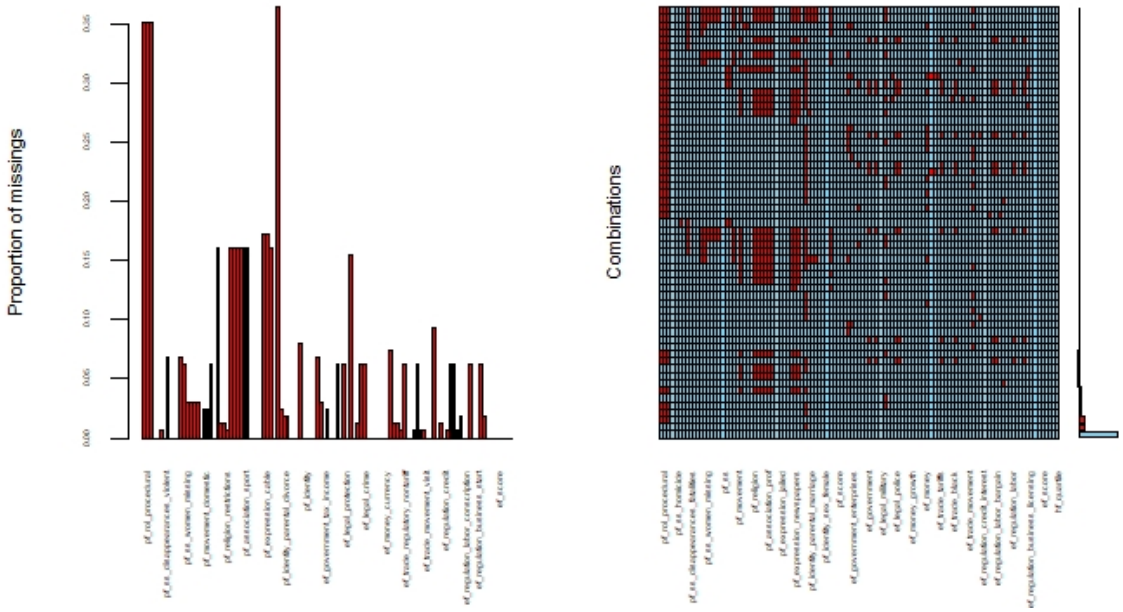

Figure 2: NAs distribution

First of all, I checked the correlation between "Personal Freedom" variables. They are grouped in 7 main categories, that are:

- "rol" - refers to the rule of law in the country (subcategories: procedural, civil and criminal)
- "ss" - refers to social security specifically for women and society overall (subcategories: disappearances, homicide, inheritance)
- "movement" - refers to the freedom to move (subcategories: domestic, foreign, women)
- "religion" - refers to freedom of religion (subcategories: estop, restrictions,harassment)

- "association" - refers to liberty of creating associations (subcategories: association, assembly, political, professional, sport)
- "expression" - refers to the possibility of free expression of opinion (subcategories: killed, jailed, influence, control, cable, newspapers, internet)
- "identity" - refers to to power to choose the social identity (subcategories: legal, marriage, divorce, male, female, identity divorce)

For each category I calculated the correlation between the variable that gives a summary for each category and the values for each subcategory. I obtained a highly positive correlation for all the variables, which means that each subcategory can be summarized in the variable related to the category. After that, I calculated the correlation between the "pf score" (value that summarizes personal freedom) and the different categories. Also in this case I obtained a highly positive correlation, so the variable "pf score" is enough to explain all the variables that refer to personal freedom.

I did the same with "Economic Freedom" (ef score). In this case there are 5 variables:

- "government" - subcategories: consumption, transfers, enterprises, tax income and tax payroll
- "legal" - subcategories: judicial, court, protection, military, integrity, enforcement, restrictions, police, crime, gender
- "money" - subcategories: growth, sd, inflation, currency
- "trade" - subcategories: tariffs, non tariff,regulatory, black trade, foreign movements, capital movement, visit movement
- "regulation" - subcategories: credit, labor, business

After all this analysis, it was clear (due to the strong correlation) that all these categories/subcategories can be summarized in "ef score". Moreover, there is a strong correlation between "ef score" & "hf score" and "pf score" & "hf score", so that both groups can be summarized in "hf score" (although I preferred keep all the three variables, to avoid losing too many information).

3.2 Descriptive analysis of the variables

All the variables in the data set are numerical, and there are no more missing values.

3.2.1 Happiness Score

The first variable "Happiness score" (a metric measured in 2016 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest") is symmetrically distributed as it is shown in the histogram (and box plot) and which is also confirmed in the fact that the mean (5.43) and the median (5.39) are close to each other and are located between the first and third quantile. All the values are ranging between 2.91 and 7.53 (see Figure 29).

3.2.2 Lower Confidence Interval

The second variable "Lower Confidence Interval" (Lower Confidence Interval of the Happiness Score) is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (5.33) and the median (5.29) are close to each other and are located between the first and third quantile. All the values are ranging from 2.73 to 7.46 (see Figure 30).

3.2.3 Upper Confidence Interval

The third variable "Upper Confidence Interval" (Upper Confidence Interval of the Happiness Score) is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (5.52) and the median (5.48) are close to each other and are located between the first and third quantile. All the values are ranging from 3.08 to 7.67 (see Figure 31).

3.2.4 Economy (GDP per Capita)

The fourth variable "GDP per capita" (the extent to which GDP contributes to the calculation of the Happiness Score) is more or less symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (0.98) and the median (1.05) are close to each other and mean is closer to median than the first and third quantile. All the values are ranging from 0.07 to 1.82 (see Figure 32).

3.2.5 Family

The fifth variable "Family" (the extent to which Family contributes to the calculation of the Happiness Score, it is a measure of how much families contributes in the life of the country, it is a social aspect), judging from the histogram seems to be not symmetrically distributed, but when it comes to mean (0.80) and median (0.86), which are very close to each other, we can see that the variable is actually symmetrically distributed. What is more, the mean is closer to median than the first and third quantile. All the values are ranging from 0 to 1.18 (see Figure 33).

3.2.6 Health (Life expectancy)

The sixth variable "Health (Life Expectancy)" (the extent to which Life expectancy contributed to the calculation of the Happiness Score), judging from the histogram seems to be not symmetrically distributed, but when it comes to mean (0.57) and median (0.62), they seems very close to each other. What is more, the mean is closer to median than the first and third quantile. All the values are ranging from 0 to 0.95 (see Figure 34).

3.2.7 Freedom

The seventh variable "Freedom" (the extent to which Freedom contributed to the calculation of the Happiness Score), judging from the histogram seems to be not symmetrically distributed, but when it comes to mean (0.38) and median (0.4), they seems to be very close to each others. What is more, the mean is closer to median than the first and third quantile. All the values are ranging from 0 to 0.6 (see Figure 35).

3.2.8 Trust (Government Corruption)

The eight variable "Trust(Government Corruption)" (the extent to which Perception of Corruption contributes to Happiness Score) seems to be not symmetrically distributed as it is observed in the histogram, but in terms of mean (0.14) and median (0.11) it is nearly symmetrically distributed, because they are close to each other and are located between the first and third quantile. All the values are ranging from 0 to 0.51 (see Figure 36).

3.2.9 Generosity

The ninth variable "Generosity" (the extent to which Generosity contributed to the calculation of the Happiness Score) is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (0.24) and the median (0.22) are close to each other and are located between the first and third quantile. All the values are ranging from 0 to 0.82 (see Figure 37).

3.2.10 Dystopia Residual

The tenth variable "Dystopia Residual" (the extent to which Dystopia Residual contributed to the calculation of the Happiness Score) is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (2.31) and the median (2.28) are close to each other and are located between the first and third quantile. All the values are ranging from 0.82 to 3.56 (see Figure 38).

3.2.11 Personal Freedom (pf score)

The eleventh variable "pf score (Personal Freedom index)" is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (7.04) and the median (6.96) are close to each other and are located between the first and third quantile. All the values are ranging from 2.51 to 9.4 (see Figure 39).

3.2.12 Economic Freedom (ef score)

The twelfth variable "ef score (Economic freedom index)" is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (6.85) and the median (6.95) are close to each other and are located between the first and third quantile. All the values are ranging from 2.88 to 8.97 (see Figure 40).

3.2.13 Human Freedom (hf score)

The thirteenth variable "hf score (Human freedom index)" is symmetrically distributed as it is shown in the histogram and which is also confirmed in the fact that the mean (6.95) and the median (6.85) are close to each other and are located between the first and third quantile. All the values are ranging from 3.77 to 8.89(see Figure 41).

I treated all the variables as numeric (since we are dealing with scores). To have a better idea of how the variables are distributed, I summarized them in Figure 3.

```
 HappinessScore     LowerCI         UpperCI       Economy_GDP          Family       Health_LifeExpectancy    Freedom
 Min.   :2.905   Min.   :2.732   Min.   :3.078   Min.   :0.06831   Min.   :0.0000   Min.   :0.0000        Min.   :0.0000
 1st Qu.:4.459   1st Qu.:4.371   1st Qu.:4.547   1st Qu.:0.69429   1st Qu.:0.6450   1st Qu.:0.4249        1st Qu.:0.2767
 Median :5.389   Median :5.295   Median :5.483   Median :1.05266   Median :0.8597   Median :0.6201        Median :0.4027
 Mean   :5.429   Mean   :5.331   Mean   :5.527   Mean   :0.98463   Mean   :0.8031   Mean   :0.5729        Mean   :0.3764
 3rd Qu.:6.361   3rd Qu.:6.227   3rd Qu.:6.465   3rd Qu.:1.27964   3rd Qu.:1.0217   3rd Qu.:0.7301        3rd Qu.:0.4877
 Max.   :7.526   Max.   :7.460   Max.   :7.669   Max.   :1.82427   Max.   :1.1833   Max.   :0.9528        Max.   :0.5961
 Trust_GovernmentCorruption   Generosity     Dystopya_residual    pf_score        ef_score        hf_score
 Min.   :0.00000            Min.   :0.0000   Min.   :0.8179    Min.   :2.512   Min.   :2.880   Min.   :3.766
 1st Qu.:0.06126            1st Qu.:0.1546   1st Qu.:1.9982    1st Qu.:6.041   1st Qu.:6.290   1st Qu.:6.275
 Median :0.10398            Median :0.2225   Median :2.2809    Median :6.975   Median :6.950   Median :6.848
 Mean   :0.13739            Mean   :0.2438   Mean   :2.3105    Mean   :7.044   Mean   :6.846   Mean   :6.945
 3rd Qu.:0.17554            3rd Qu.:0.3147   3rd Qu.:2.6646    3rd Qu.:8.259   3rd Qu.:7.500   3rd Qu.:7.858
 Max.   :0.50521            Max.   :0.8197   Max.   :3.5591    Max.   :9.399   Max.   :8.970   Max.   :8.887
```

Figure 3: Variables summary

3.3 Correlation

In general, it is possible to say that the correlation between the variables is mostly positive, even if the correlation between most of the variables is moderate (see Figure 4). Although we can say that Happiness Score is mostly correlated with Economy (GDP per Capita) and Social aspects (like how families are involved in the life of the country and the life expectancy of the inhabitants). Actually, the correlation between Happiness score and:

- GDP is 0.80
- Family is 0.74
- Life expectancy is 0.76

It is also interesting to notice that there is a strong positive correlation (o.83) between Life Expectancy (health) and the GDP (Economy): as the GDP increase, the level of health (so the life expectancy) of a Country increase. The variables related to freedom do not seem to be strongly correlated to the happiness score, with the exception of the economic freedom (ef score) that seems to be positively correlated to happiness. So, in general, we could say that the happiness rate of a country is affected the most by the economic situation of the country (even if social factors still have their importance). As suspected there is a strong correlation between pf score, ef score and hf score. It is also possible to notice that there is a strong positive correlation between Happiness Rate and its lower and upper bounds. The variable Generosity is the least correlated with the others. Happiness score and its lower and upper bound (as expected) are strongly correlated as they refer exactly to the same object. As I have a strong correlation between some of the variables implemented the PCA, which is a way to deal with highly correlated variables, so there is no need to remove them.

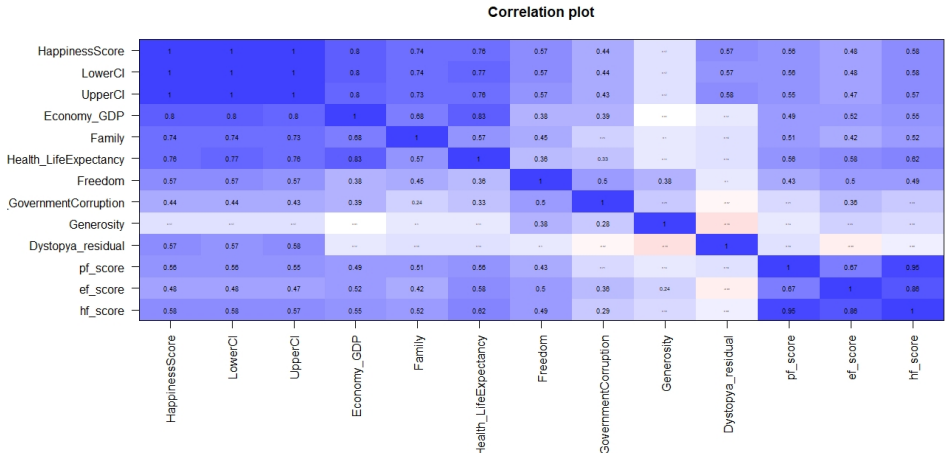

Figure 4: Correlation

4 Principal Component Analysis (PCA)

Principal component analysis, or PCA, is a statistical procedure that allows to summarize the information content in large data tables by means of a smaller set of "summary indices" that can be more easily visualized and analyzed. While reducing the number of variables we have to make sure to have principal components that maximize the explanation of the variance (in order to lose the lowest number of information as possible). The goal is to extract the important information from the data and to express this information as a set of summary indices called principal components.

It is important to scale the data before starting the analysis, so that it is possible to obtain a symmetric distribution (normalized).

In R we can run the PCA using the command *prcomp()*, and using *scale=T* to normalize the variables (variance=1). Running this command on our dataset we obtain the standard deviations of the principal components and the rotation matrix (Figure 5).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Happiness.Score	0.353	0.234	0.105	-0.016	-0.042	-0.037	0.009	-0.027	0.040	0.004	-0.371	0.000	0.816
Lower.Confidence.Interval	0.354	0.229	0.103	-0.016	-0.040	-0.039	0.003	-0.024	0.023	0.707	-0.373	0.000	-0.411
Upper.Confidence.Interval	0.352	0.238	0.107	-0.017	-0.044	-0.035	0.015	-0.030	0.058	-0.707	-0.368	0.000	-0.406
Economy..GDP.per.Capita.	0.313	0.075	-0.108	0.464	-0.068	-0.092	0.100	0.065	0.702	0.019	0.386	0.000	0.000
Family	0.287	0.068	-0.065	0.194	-0.372	0.590	-0.407	-0.270	-0.288	-0.006	0.251	0.000	0.000
Health..Life.Expectancy.	0.309	0.009	-0.148	0.316	-0.118	-0.396	0.294	0.308	-0.621	-0.016	0.214	0.000	0.000
Freedom	0.244	-0.198	0.365	-0.188	0.156	0.553	0.555	0.273	-0.001	0.005	0.140	0.000	0.000
Trust..Government.Corruption.	0.186	-0.178	0.484	0.261	0.624	-0.124	-0.450	-0.010	-0.112	-0.008	0.109	0.000	0.000
Generosity	0.085	-0.352	0.538	-0.253	-0.600	-0.338	-0.123	-0.013	0.093	0.002	0.133	0.000	0.000
Dystopia.Residual	0.128	0.576	0.059	-0.532	0.166	-0.166	-0.003	-0.166	-0.040	0.005	0.531	0.000	0.000
pf_score	0.277	-0.217	-0.357	-0.347	0.056	0.019	-0.344	0.465	0.088	-0.005	0.000	0.532	0.000
ef_score	0.263	-0.395	-0.194	-0.075	0.152	-0.130	0.283	-0.712	-0.032	0.000	0.000	0.321	0.000
hf_score	0.296	-0.309	-0.322	-0.266	0.101	-0.040	-0.118	0.023	0.047	-0.004	0.000	-0.784	0.000

Figure 5: prcomp() output

Returning the summary, we get the values to evaluate the relevance of each principal component. In our case, it is possible to observe that the first 2 principal components explain the 68% (54.8% expained by the first component) of the total variability. With the third component the total variability explained is 78% and with the fourth it grows to 85%. To understand the number of principal component to keep, I used the scree plot (Figure 6).

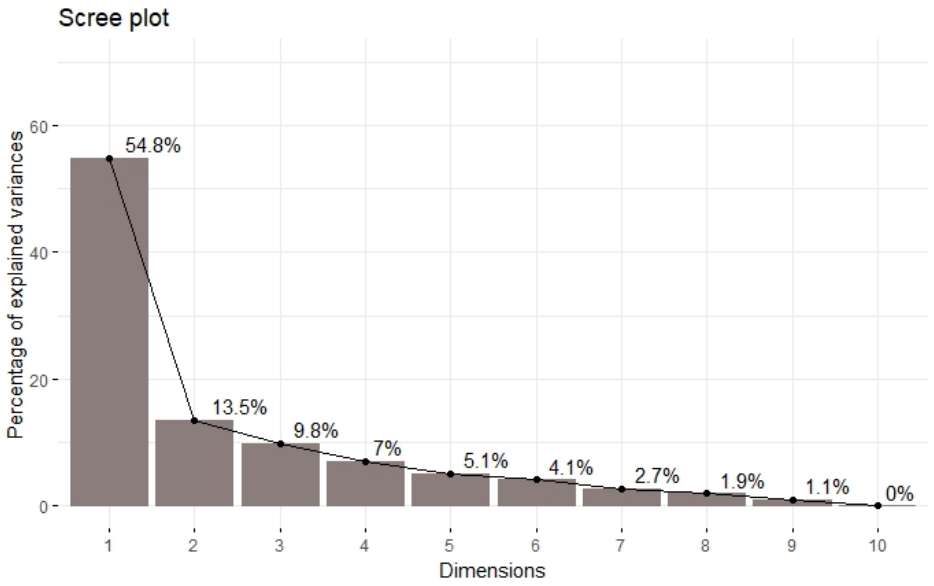

Figure 6: ScreePlot

The scree plot criterion looks for the "elbow" in the curve and selects all components just before the line flattens out. When the eigenvalues drop dramatically in size, an additional factor would add relatively little to the information already extracted. According to the screeplot in Figure 6, we could say that the "elbow" is located in correspondence of the second/third principal component, so we could say that these three explain most of the variability of the model (the first three PC explain 78% of the total variance). Actually, also two components could be enough, as they already explain 68% of the total variability.

To visualize the magnitude and sign of each variable's contribution to the first two principal components, and to represent each observation in terms of those components, we use the biplot. The plot shows the observations as points in the plane formed by the first 2 principal components (synthetic variables). Like for any scatterplot we may look for patterns, clusters, and outliers. In addition to

the observations the plot shows the original variables as vectors (arrows). They begin at the origin [0,0] and extend to coordinates given by the loading vector (plot of the direction vectors that define the model, it shows how the original variables contribute to creating the principal component). These vectors can be interpreted in three ways:

- The orientation (direction) of the vector, with respect to the principal component space, in particular, its angle with the principal component axes: the more parallel to a principal component axis is a vector, the more it contributes only to that PC.
- The length in the space; the longer the vector, the more variability of this variable is represented by the two displayed principal components; short vectors are thus better represented in other dimension.
- The angles between vectors of different variables show their correlation in this space: small angles represent high positive correlation, right angles represent lack of correlation, opposite angles represent high negative correlation.

Looking at the *biplot* in Figure 7, *Health (life expectancy)* is perfectly parallel with the PC1 axes, this means that it only contribute to that PC, but it is perfectly perpendicular to the PC2, so it do not contribute to it. Referring to the length of the vectors, *Dystopia* and *Happiness score* (with its lower and upper bounds) , seem to be the longest vectors, this means that most of the variability of these two variables is displayed by the first two PCs. Whereas the shortest vector seems to be *Generosity*, whose variability would be better explained by other principal components. Lastly, referring to the angles between vectors, we could say that there is almost perfect correlation between the variables Happiness score and its lower and upper bounds. Whereas we could say that there is no correlation between Dystopia and corruption (that seems highly correlated with freedom, trust in government and economic freedom). There is almost negative correlation between dystopia and generosity.

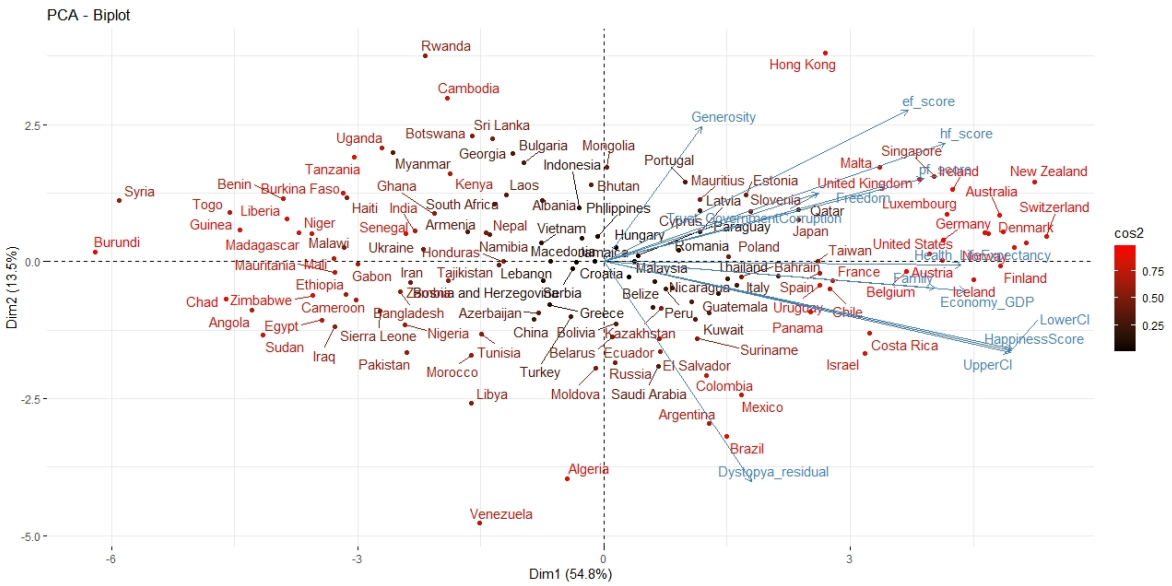

Figure 7: Biplot

Already from the biplot it is also possible to understand something about the level of happiness in the different countries. For example we could say that Syria is the country with the lowest Happiness Score, because it is exactly opposite to the direction of the vector of Happiness score. We can also see that countries with positive values for GDP and health life expectancy (e.g. Iceland, Finland, Austria) are also the ones that have the higher values for Happiness score.
This analysis could be deepened splitting the biplot in its two components: the score plot and the loading plot.

With the **score plot**, we can form a first idea about how to cluster the Countries. Countries close to

each other have similar characteristics, whereas those far from each other are dissimilar. In particular, from Figure 8, we can observe that most of the Countries located in the right side of this plot, are European Countries, wherease in the left side there are mostly Asian or African Countries. We could observe for example that Countries like Iceland or Finland are completely in the opposite side with respect to Burundi and Syria.

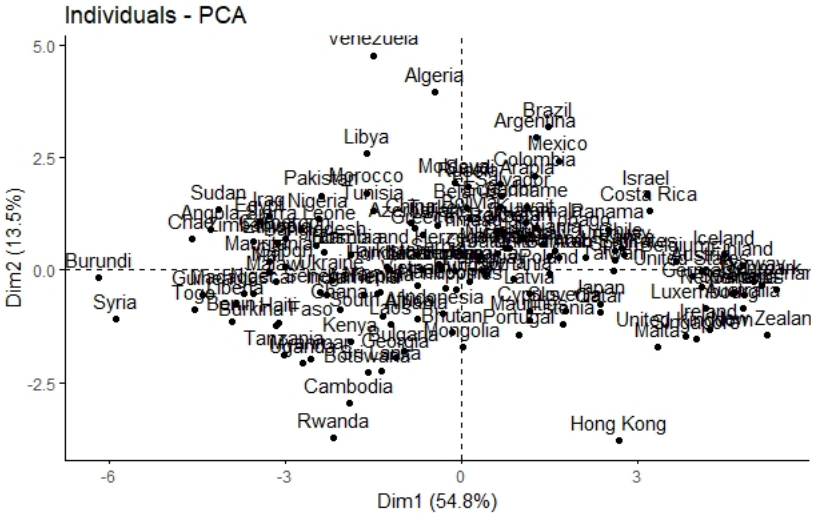

Figure 8: Score plot

With the **loading plot** (in Figure 9) we can better visualize the correlation between variables. But what is made clear from this graph is how much of the variability of the variable is represented by the two displayed principal components (length of the vectors). The lighter the vector, the higher the variability of the variable is represented by the two displayed principal components. So, for example the variability Happiness Score is the most represented by the first two principal components, whereas Generosity is the least represented by these two components.

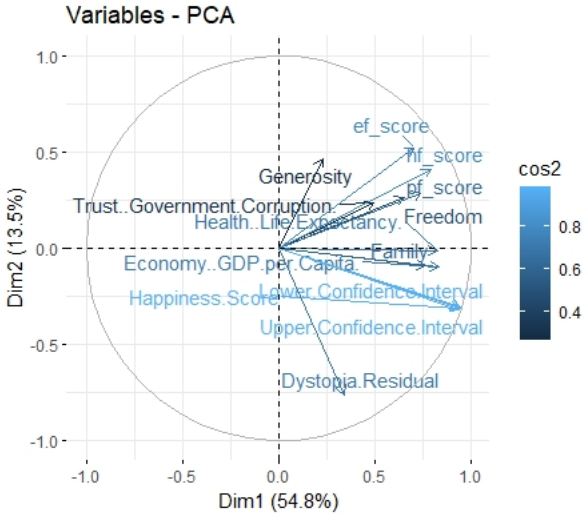

Figure 9: Loading plot

The following graphs represent the contribution of the different variables to the first dimension (Figure 10), the second dimension (Figure 11) and both the dimensions (Figure 12).

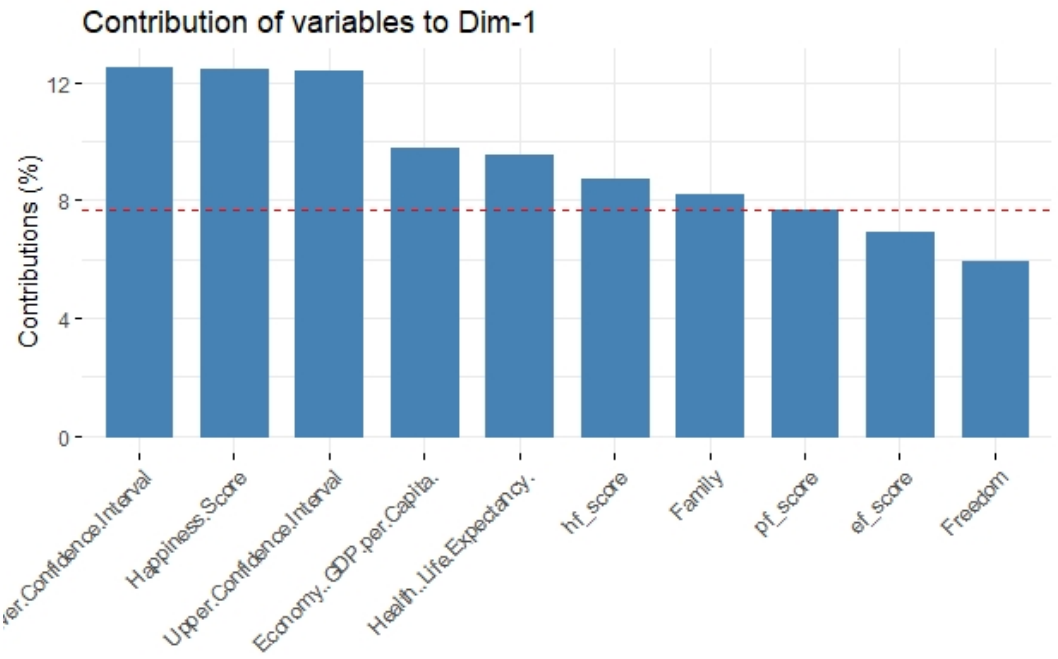

Figure 10: Variable contribution to dimension 1

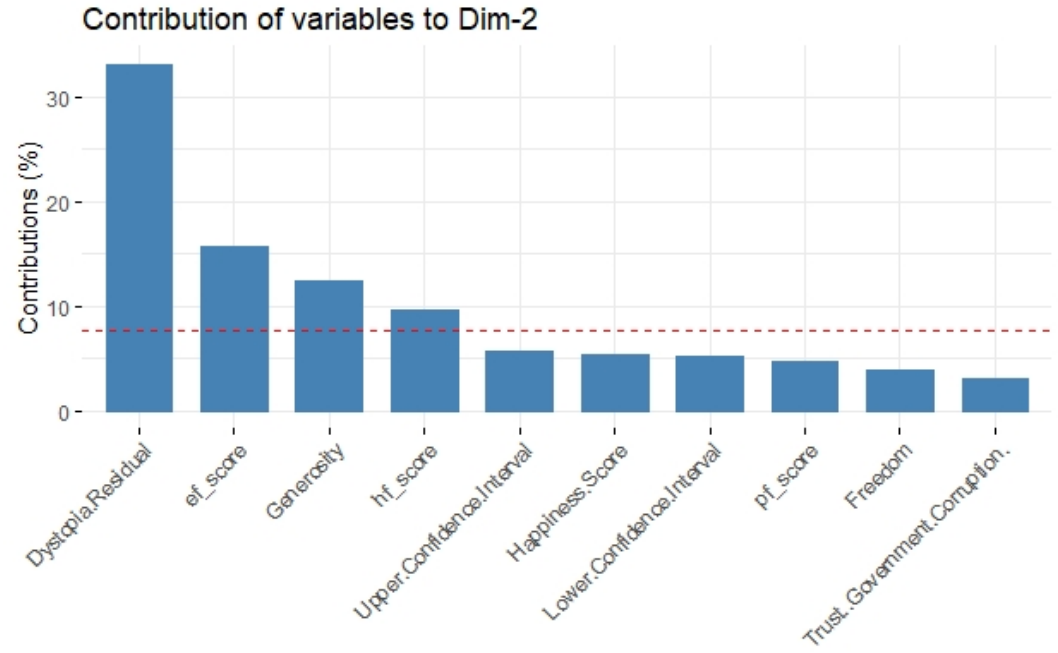

Figure 11: Variable contribution to dimension 2

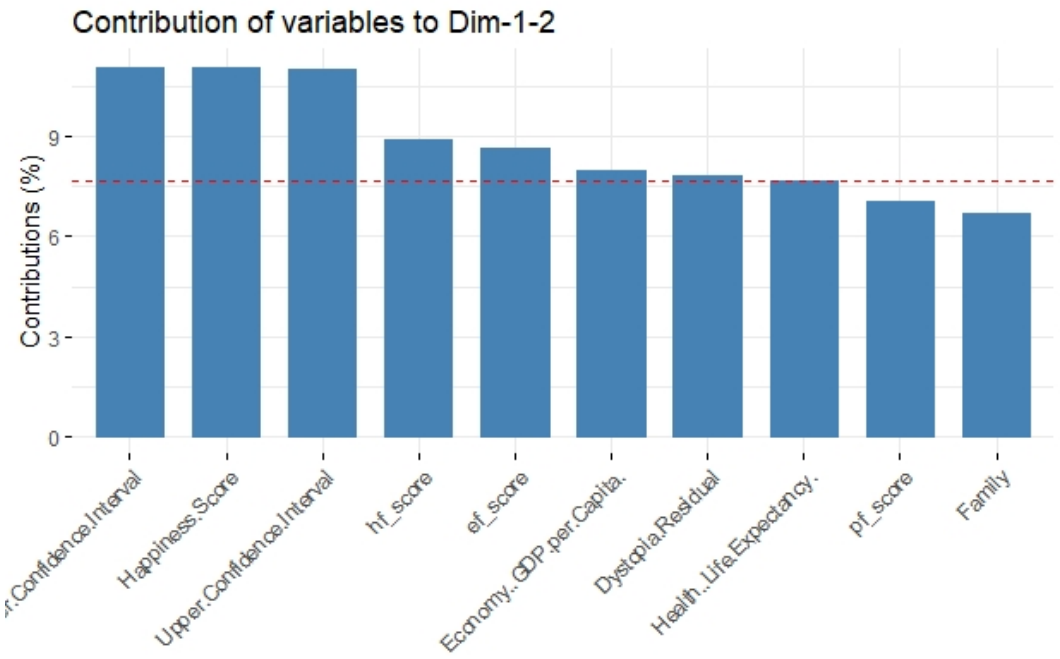

Figure 12: Variable contribution to both dimensions

5 Cluster analysis

A cluster is collection of data objects, similar to one another within the same cluster, but dissimilar to the objects in other cluster. Cluster analysis is a statistical method for processing data. It works by organizing items into groups, or clusters, on the basis of how closely associated they are. It is an unsupervised learning algorithm, meaning that you don't know how many clusters exist in the data before running the model. A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. Similarity is expressed in terms of a distance function, which is typically metric: *d(i, j)*.

5.1 Hierarchical clustering

Before starting, I have conducted an analysis to understand which is the best method (between ward, single, complete and average) to apply to *hclust()* (the command in R to run the hierarchical clustering). In order to do so, I applied the cophenetic distance and I found out that the higher correlation between the euclidean distance and cophenetic distance is obtained with the "average method" (0.6572907). Although, applying the average method I obtained a chaining problem (as you can see in Figure 13).

The second method suggested by the cophenetic distance is the "complete" method (0.6206036), in this case we do not have problems of chaining and the dedrogram is easier to read. According to the dendrogram obtained (Figure 14), in my opinion, the ideal number of clusters is 3.

The dendrogram with 3 clusters shows a division between countries that are underdeveloped (i.e. many countries in Africa and Asia), developed (i.e. countries in North Europe, North America and Oceania) and in development (i.e. South America, Eastern Europe and certain countries in Asia).

- In the first cluster we have **the more-or-less happy and free Countries**, some Countries with strong economies but low rights and some Countries with weak economies and more rights;
- In the second cluster we have **the saddest and least free Countries** , with low values both for economy and rights;
- In the third cluster we have **the happiest and most free Countries**, developed both for economy and rights.

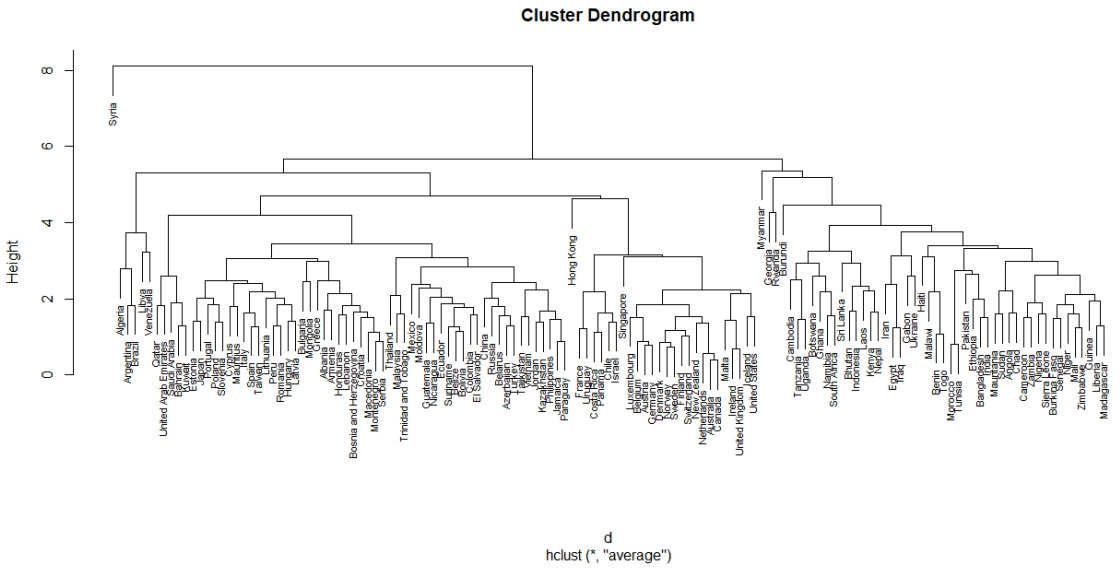

Figure 13: Dendrogram (average method)

But this is just broadly speaking, we will see later the characteristics for the different clusters.

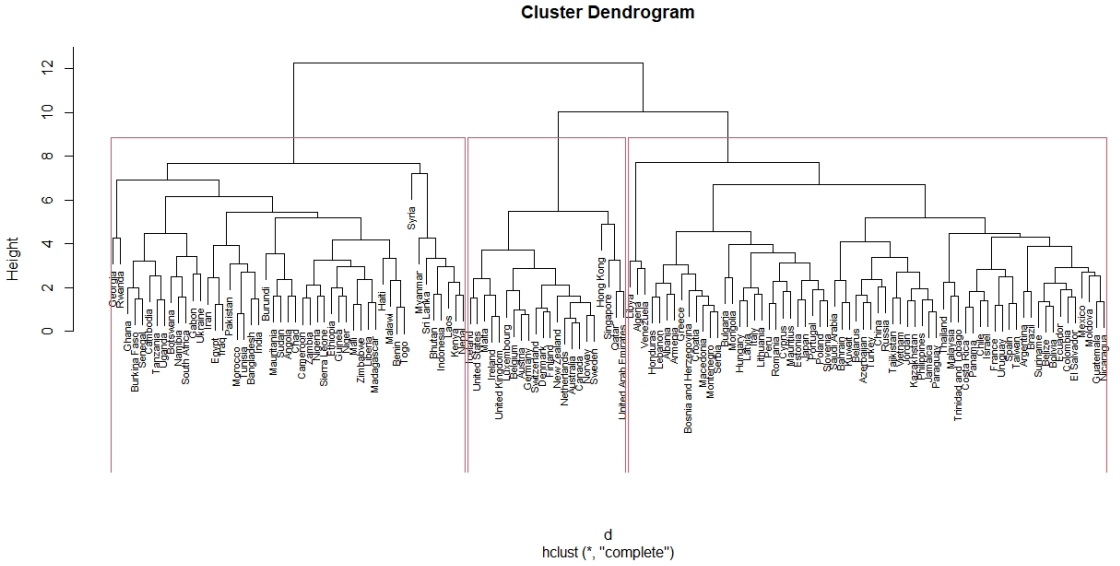

Figure 14: Dendrogram (complete method)

To have a visual representation of what obtained from hierarchical clustering, I plotted the cluster on a geographical map, in Figure 15. From this map is more clear the division of the Countries in clusters. We can notice that most of the Countries in Africa are group in the second cluster like others State in the south-west of Asia (i.e. India, Pakistan, Iraq, Syria and Iran), Ukraine and Malaysia. In the first cluster we have Countries located in South America, Russia, Asia, North Africa and South of Europe. Lastly, in the third cluster, we have Countries located in North America, North Europe and Oceania.

To better understand the characteristics of the different clusters, in Figure 16, there is a table with the mean for the different variables in each cluster.

Analysing the values in the table (Figure 16), it is possible to notice that the happiest Countries are

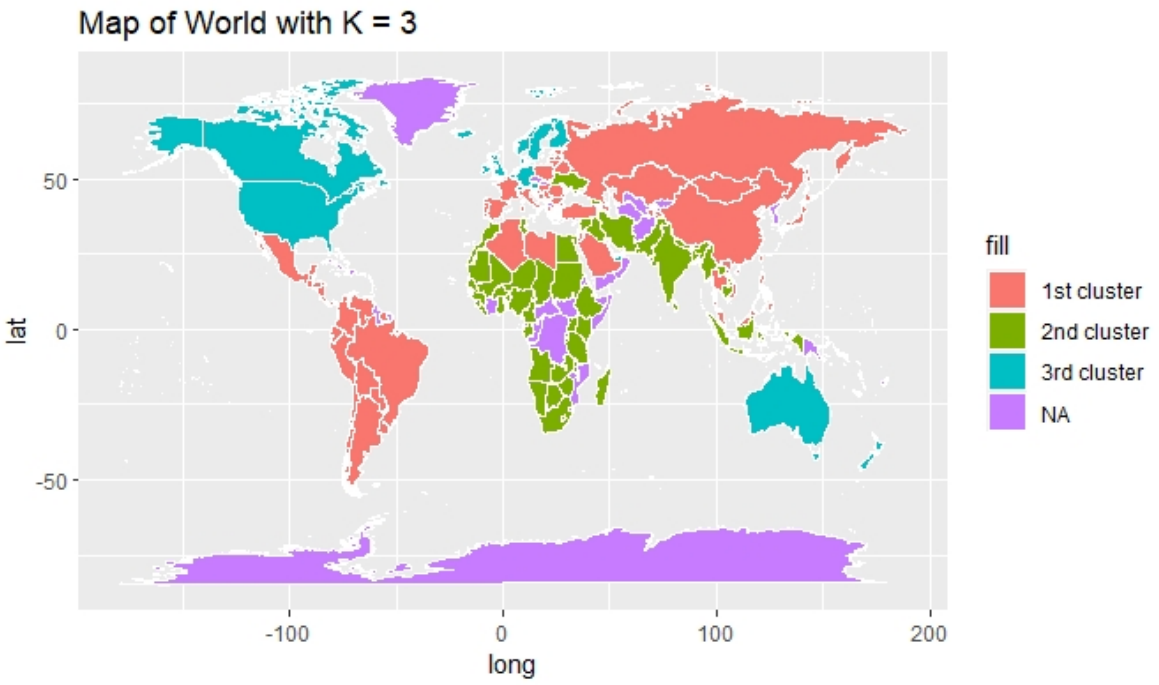

Figure 15: Geographical map clustering (hierarchical clustering)

	[,1]	[,2]	[,3]
HappinessScore	5.78515152	4.2351224	7.0186364
LowerCI	5.68512121	4.1321429	6.9365455
UpperCI	5.88518182	4.3381020	7.1007273
Economy_GDP	1.10559803	0.5940245	1.4917336
Family	0.88939606	0.5698627	1.0637136
Health_LifeExpectancy	0.66108030	0.3415080	0.8238473
Freedom	0.36692121	0.3119529	0.5484577
Trust_GovernmentCorruption	0.09315212	0.1123349	0.3259382
Generosity	0.18530242	0.2603453	0.3827155
Dystopya_residual	2.48369970	2.0450918	2.3822618
pf_score	7.23988945	6.0328401	8.7106859
ef_score	6.93106061	6.2746939	7.8659091
hf_score	7.08547503	6.1537670	8.2882975

Figure 16: Means by variables (for each cluster)

also the most free and they are the ones that belong to the third group, while the least happy Countries are the ones located in the second group. This is reflected in almost all the other variables such as, the Economy, Health and Freedom (personal, economic and health freedom). An interesting thing to notice is the Trust in Government, actually, for this variable we have lower values in correspondence to the first group, that have even lower values than the second group. Another thing to notice are the values of Generosity (how much a Country help the other Countries); in this case the first group is the one with the lowest values. Even if the Country in the second group are poorer than the ones in the first group, they seems to help more the other countries. Regarding to the Dystopia residual, the first group is the one with the higher results that is the farthest from Dystopia (an imaginary country that has the world's least-happy people).

	1 vs 2 (%)	3 vs 1(%)	3 vs 2(%)
HappinessScore	26.8	17.6	39.7
LowerCI	27.3	18.0	40.4
UpperCI	26.3	17.1	38.9
Economy_GDP	46.3	25.9	60.2
Family	35.9	16.4	46.4
Health_LifeExpectancy	48.3	19.8	58.5
Freedom	15.0	33.1	43.1
Trust_GovernmentCorruption	-20.6	71.4	65.5
Generosity	-40.5	51.6	32.0
Dystopya_residual	17.7	-4.3	14.2
pf_score	16.7	16.9	30.7
ef_score	9.5	11.9	20.2
hf_score	13.1	14.5	25.8

Figure 17: Relative differences in variables between clusters

To better understand the difference in mean between the three clusters about the different variables, in Figure 17, I summarized relative differences between the different clusters. From this table, and starting our analysis by looking at the relative difference between the third cluster (the happiest) and the second cluster (the saddest), so looking at the last column of the table, it is possible to notice that the greatest change is on the variables *Trust in government* (65.5%) and *Economy - GDP* (60.2%). So we could say that this are the factors that mostly contribute in the differentiation of these two clusters. When it comes to the relation between the third cluster (happiest countries) and the first one (more or less happy countries), it is possible to notice that the greatest difference is about *Trust in government* (71.4%) and *Generosity* (51.6%). In this case we have a negative coefficient for *Dystopya residual* (-4.3%), meaning that the countries in the middle cluster (cluster 1) are farther from the Dystopya characteristics than the happiest countries, this was not expected as the countries of the third group are the ones which should be farther from Dystopya (the saddest country in the world). Lastly, comparing the first cluster (the one in the middle in terms of happiness) with the second cluster (the one with the saddest countries), it is possible to notice that the greatest relative difference is for the variables *Health and life expectancy* (48.3%) and *Economy - GDP* (46.3%). In this case, it is interesting to notice that there is a negative relative difference both for *Trust - Government corruption* (-20.6%) and *Generosity* and a positive coefficient *Dystopya residual* (17.7%). As stated before, even if the countries of the first group are richer (in terms of GDP) than the countries of the second group, they are less generous with the other countries. Moreover in the second group trust more in the government than the countries of the first group.As correctly expected, in this case Dystopya residual is higher for the happier countries (higher for group 1).

5.2 K-Means

Now we move onto the K-Mean method. By applying this method, we have to choose the number of clusters in advance, and in order to do so we use the "elbow" chart. To build this chart, I implemented several time the analysis, using different number of initial clusters (from 1 to 15), and each time I calculated the sum of squares within groups. According to Figure 18 the ideal number of clusters is between 3 and 4, actually in these point there is the elbow.

After choosing 3 as the ideal number of clusters (from elbow chart), I calculate the centroids (with the function *aggregate()* and at the end I plotted the scatter-plot in Figure 19. Also in this case I scaled the data before starting the analysis. Already from this plot, we can extract useful information to understand the characteristics of the countries inside the different groups. In this case we have many variables and the pair plot could result hard to read, but this will be useful to understand some characteristics that will be deeply explained later. In general, we could say that most of the variables

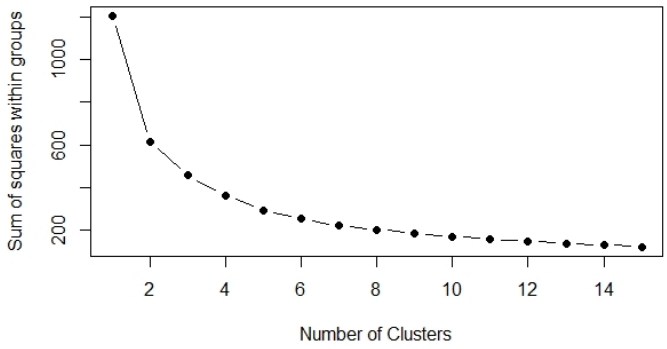

Figure 18: Elbow chart

have low values for the "pink group", middle values for the "black group" and high values for the "green group". But there are some variables (i.e. Generosity) in which the "pink group" seems to have higher values than the "black group". We will see it more in detail later on.

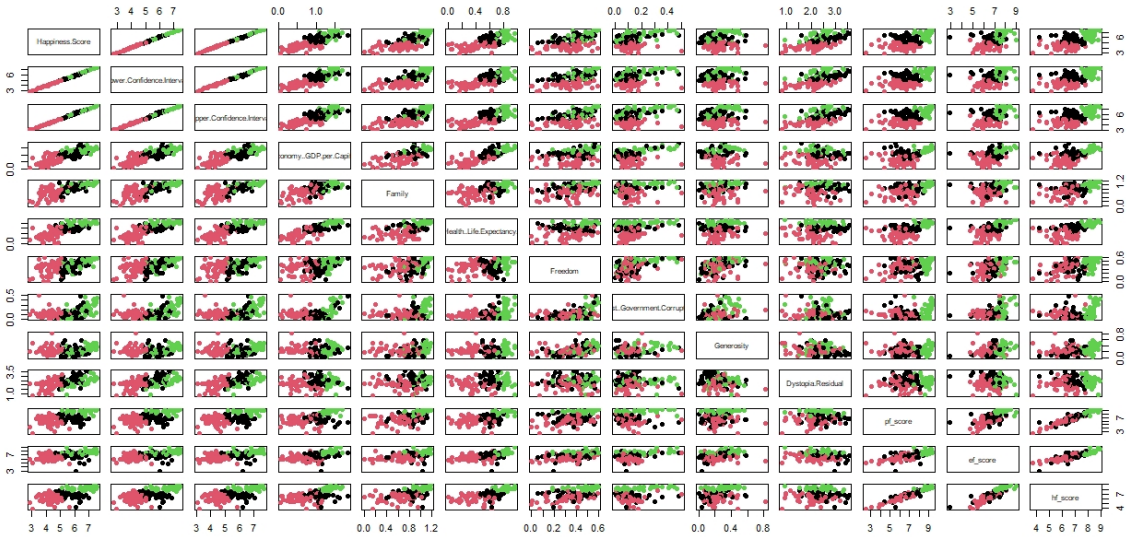

Figure 19: Scatter-plot (3 clusters)

Lastly, the findings from the Clusplot (Figure 20) seem to confirm what I found out from the previous analysis with hierarchical method, except for a few cases. Actually, in cluster 3 (the one on the right) we have countries that are developed in terms of economic as well as human rights issues, for these countries we have generally very high values for the Happiness Score and freedom. In cluster 1 we have countries with middle values for the Happiness Score and freedom. In group 2 (the one on the left) we have mostly underdeveloped countries both for economy and human rights and here the values for the Happiness Score and freedom are the lowest. When analysing the cusplot, we have to bear in mind that it takes into consideration the fist two principal components (from PCA), and not all the original variables (see Chapter 4 for the PCA).
Also in this case, I plotted the results in geographical map, to obtain a better visualization of the output. In particular, in Figure 21, we can observe that most of the European countries (mostly the Western ones) are now grouped in the third cluster (in Figure 15, for hierarchical clustering, we saw that only Countries in the north of Europe belonged to the first cluster). Also in this case almost all

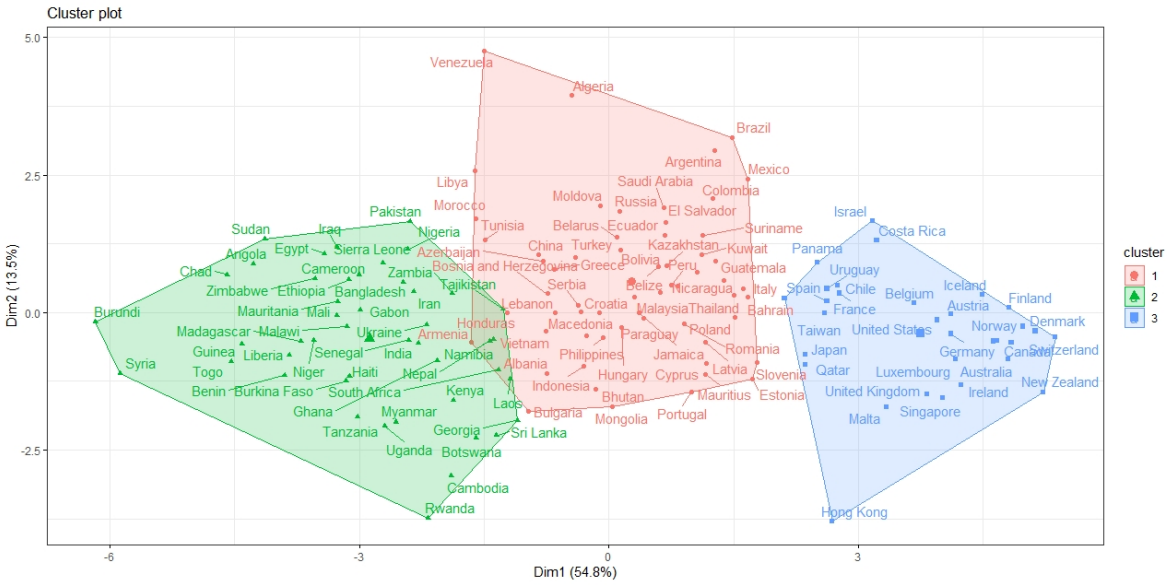

Figure 20: Cusplot (K-Means)

the countries of Africa belong to the same group with most of the countries in South Asia. Russia, most of the Countries in the Eastern Asia and most of the countries in South America still belong to the same group.

As in Hierarchical clustering, also in this case, to better understand the characteristics of the different

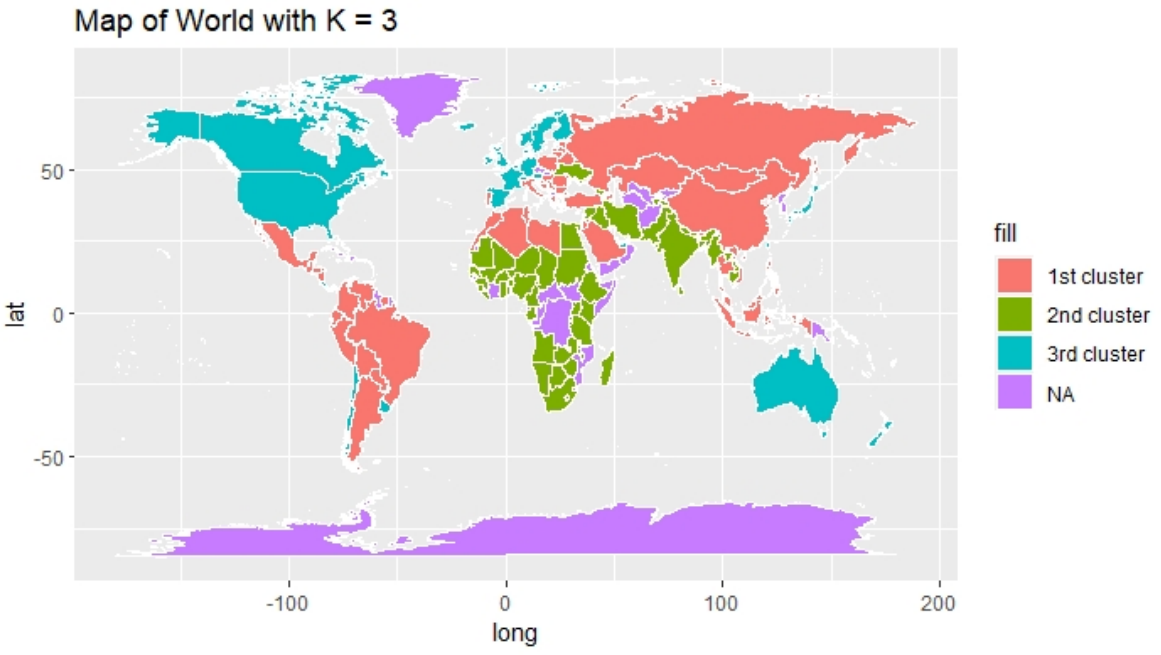

Figure 21: Geographical map clustering (K-Mean clustering)

clusters, in Figure 22, there is a table with the mean for the different variables in each cluster.

Even if some Countries move to a cluster to another (with respect to the Hierarchical clustering), the characteristics of the three cluster still the same. Higher values of happiness are reflected in almost all the other variables such as, the Economy, Health and Freedom (personal, economic and health freedom). Also in this case, for Trust in Government, actually, we have lower values in correspondence

	[,1]	[,2]	[,3]
HappinessScore	5.63476667	4.1698043	6.8985161
LowerCI	5.53305000	4.0652174	6.8165806
UpperCI	5.73648333	4.2743913	6.9804516
Economy_GDP	1.07666883	0.5646520	1.4296997
Family	0.85802850	0.5668930	1.0473006
Health_LifeExpectancy	0.63324683	0.3278972	0.8197339
Freedom	0.35279600	0.3124928	0.5169719
Trust_GovernmentCorruption	0.08737467	0.1154117	0.2668281
Generosity	0.18597300	0.2583987	0.3342548
Dystopya_residual	2.44068217	2.0240576	2.4837374
pf_score	7.04148249	5.9938038	8.6088075
ef_score	6.83033333	6.2465217	7.7677419
hf_score	6.93590791	6.1201628	8.1882747

Figure 22: Means by variables (for each cluster)

to the second group, that have even lower values than the first and third group. Another thing to notice are the values of Generosity (how much a Country help the other Countries); also in this case the first group is the one with the lowest values. Even if the Country in the second group are poorer than the ones in the third group, they seems to help more the other countries. Regarding to the Dystopia residual, the third group still be the one with the higher results that is the farthest from Dystopia (an imaginary country that has the world's least-happy people).

	1 vs 2 (%)	3 vs 1(%)	3 vs 2(%)
HappinessScore	26.0	18.3	39.6
LowerCI	26.5	18.8	40.4
UpperCI	25.5	17.8	38.8
Economy_GDP	47.6	24.7	60.5
Family	33.9	18.1	45.9
Health_LifeExpectancy	48.2	22.7	60.0
Freedom	11.4	31.8	39.6
Trust_GovernmentCorruption	-32.1	67.3	56.7
Generosity	-38.9	44.4	22.7
Dystopya_residual	17.1	1.7	18.5
pf_score	14.9	18.2	30.4
ef_score	8.5	12.1	19.6
hf_score	11.8	15.3	25.3

Figure 23: Relative differences in variables between clusters

Even in this case, to better understand the difference in mean between the three clusters about the different variables, in Figure 23, I summarized relative differences between the different clusters. In this case the major change with respect to the hierarchical clustering method is that the greatest relative difference between the third and the second cluster is for the variables *Economy - GDP* (60.5%) and *Health and Life Expectancy* (60%). With respect to the relation between the third and the first group, it is interesting to notice that in this case the third group the third group has higher values for this variable, as expected. There are no significant changes in the relation between the first and second

cluster.

5.3 K-Medoids

The last clustering technique used is the K-Medoids method. This was useful, because it is more robust to noise and outliers. The method is also called "PAM". Also in this case we have to choose the number of clusters in advance. In this case I choose the number of clusters that maximize the silhouette width. As it is possible to observe observed from Figure 24, the optimal amount of clusters is 3.

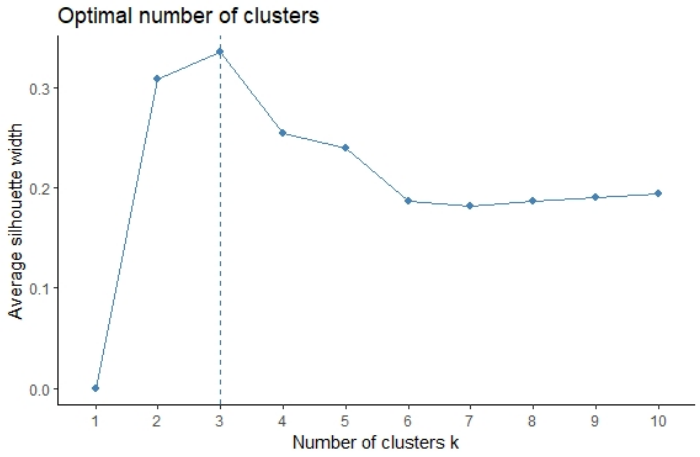

Figure 24: Optimal n. of clusters (K-Medoids)

Once again I divide the observations in three clusters because then the clusters can be better differentiated from each other and less overlap. Once again I am ready to confirm that it proves the earlier findings from the previous analysis.

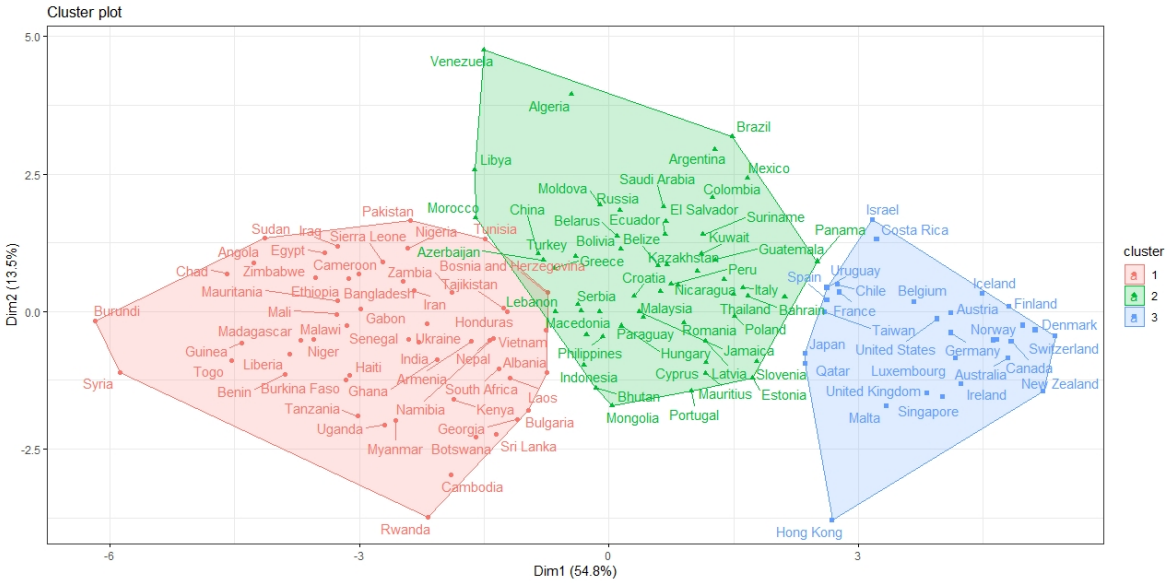

Figure 25: 3 Clusters (K-Medoids)

Also in this case, to have a better idea of how the Countries are located in the different cluster, I plot the geographical map with different color for the Countries in the different clusters. As we can see

from Figure 26 (here the colours for the first and second group are inverted with respect to the other two methods), the results are very similar to the once obtained for the K-Mean method. For all the results it is possible to confirm what already said for the other two clustering methods.

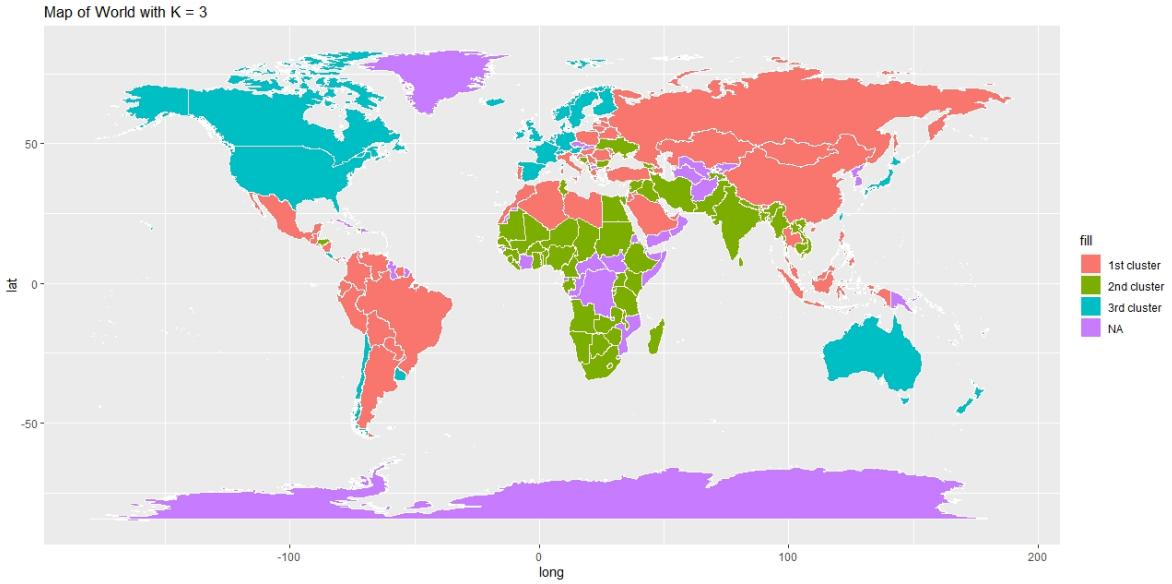

Figure 26: Geographical map clustering (K-Medoids clustering)

Moreover I present the table with the means for each variable (Figure 27) in order to attribute characteristics to the different clusters.

	[,1]	[,2]	[,3]
HappinessScore	5.78160000	4.2487358	6.9165517
LowerCI	5.67987273	4.1450000	6.8351379
UpperCI	5.88332727	4.3524717	6.9979655
Economy_GDP	1.11057745	0.6084706	1.4332452
Family	0.88486727	0.5802142	1.0553814
Health_LifeExpectancy	0.63336909	0.3713408	0.8266714
Freedom	0.37073145	0.3057287	0.5163693
Trust_GovernmentCorruption	0.09673964	0.1069921	0.2700631
Generosity	0.19088055	0.2463019	0.3397997
Dystopya_residual	2.49446400	2.0296740	2.4750069
pf_score	7.01490012	6.1353398	8.7615325
ef_score	6.83654545	6.3439623	7.7834483
hf_score	6.92572279	6.2396510	8.2724904

Figure 27: Means by variables (for each cluster)

Also in this case, to better understand the difference in mean between the three clusters about the different variables, in Figure 28, I summarized relative differences between the different clusters. In this case, compering the third and second cluster, the highest relative difference is for the variable *Trust in Government* (60.4%) and *Economy - GDP* (57.5%). Looking at the relative differences between the third and first cluster, it is possible to notice that even in this case the greatest change in for the variable *Trust in Government* (64.2%). As in the case of the hierarchical clustering, also in this case the change for the variable *Dystopya residual* (-0.8%) is negative. For the relative differences

between the first and second group, also in this case we record results similar to the ones obtained for the previous two methods.

	1 vs 2 (%)	3 vs 1(%)	3 vs 2(%)
HappinessScore	24.0	16.4	38.6
LowerCI	24.4	16.9	39.4
UpperCI	23.5	15.9	37.8
Economy_GDP	42.2	22.5	57.5
Family	31.4	16.2	45.0
Health_LifeExpectancy	41.4	23.4	55.1
Freedom	12.7	28.2	40.8
Trust_GovernmentCorruption	-20.3	64.2	60.4
Generosity	-31.6	43.8	27.5
Dystopya_residual	16.5	-0.8	18.0
pf_score	12.9	19.9	30.0
ef_score	7.1	12.2	18.5
hf_score	10.1	16.3	24.6

Figure 28: Relative differences in variables between clusters

6 Conclusions

In conclusion, we could say that there is a positive correlation between happiness and freedom as the Countries with the higher values of Happiness score are also the ones with the higher values of Freedom. Thanks to the clustering methods used (hierarchical clustering, k-mean and k-medoid), it was possible to cluster the Countries in three different cluster.

The three clusters differs for average values of the different variables and in particular, we could divide the countries in:

- In the first cluster we have **the more-or-less happy and free Countries**, some Countries with strong economies but low rights and some Countries with weak economies and more rights;
- In the second cluster we have **the saddest and least free Countries** , with low values both for economy and rights;
- In the third cluster we have **the happiest and most free Countries**, developed both for economy and rights.

From the different clustering methods we obtain more or less the same results, except for some Countries (i.e. Spain, France, Morocco etc.) that moves from a cluster to another changing the method. The following table resume the Countries that changed cluster changing the clustering method:

State	**Hierarchical**	**Kmean**	**Kmedoid**
Albania	1	1	2
Armenia	1	1	2
Bosnia and Herzegovina	1	1	2
Bulgaria	1	1	2
Chile	1	3	3
France	1	3	3
Honduras	1	1	2
Indonesia	2	1	1
Japan	1	3	3
Moroco	2	1	1
Panama	1	3	1
Spain	1	3	3
Taiwan	1	3	3
Tajikistan	1	2	2
Tunisia	2	1	2
United Arab Emirates	3	3	1
Uruguay	1	3	3
Vietnam	1	1	2

About the characteristics of the different groups we can say that the first cluster (**the more-or-less happy and free Countries**) is the one with values of the variable that is in the middle for most of the variables. But for this group we always have the lowest values for *Generosity* and *Trust in Government* and the highest values for Dystopia residual (except for the K-Means method). This means that even if they are richer than the countries in the second cluster (**the saddest and least free Countries**), they tend to help less the other countries and they trust less in their Government. With respect to Dystopia residual, having the highest values means that they are the most distant from Distopya (the imaginary saddest country in the world). Looking at the relative distances between the means for the third and second country, *Economy - GDP* is always between the variables that present the greatest relative difference. Nevertheless, the variable that presented the greatest difference is *Trust in Government*, for Hierarchical and K-medoids, whereas is *Economy - GDP* followed by *Health and Life Expectancy* for the K-mean method.

Before starting the clustering of the Countries, a PCA was presented. Thanks to PCA it was possible to reduce the dimensionality of the data and trying to understand which factors are the most influential for the first two dimensions. It was interesting to notice that the first two principal components

already explained the 68% of the total variability (55% explained by the first component) and adding the third component 78% of the total variability is explained. The variable that mostly contribute to explain the first dimension is the *Happiness Score*, whereas the second dimension is mostly explained by *Dystopia residual*.

7 R Code

```
#Libraries
library(corrplot)
library(VIM)
library(naniar)
library(FactoMineR)
library(factoextra)
library(cluster)
library(plmf)
library(anacor)
library(ggplot2)
library(grid)
library(ca)
library(FactoMineR)
library(dplyr)
library(ClustGeo)
library(dplyr)


#Data
rm(list=ls())
d1=read.csv(choose.files(), header=T) #2016
d2=read.csv(choose.files(), header=T) #hfi_cc_2018

#d1
rownames(d1)=d1[,1]
d1.new=d1[,-c(1,2,3)]
which(is.na(d1.new)) #no missing values
sum(is.na(d1.new))

#d2
which(d2[,1]==2016) #from 1 to 162, data of 2016
new.d2=d2[1:162,]
rownames(new.d2)=new.d2[,3]

rownames(new.d2)=d2[1:162,3]
new.d2=new.d2[,-3]


#corrplot.mixed(cor(new.d2), order="hclust", tl.col="black")
#install.packages("DMwR")
#new.d2=new.d2[,-c(1,2,3,26,27,34,35,37,38,40,41)]

aggr(new.d2)
any_na(new.d2) #TRUE there are missing values
miss_var_summary(new.d2)

nomissing=knnImputation(new.d2[, !names(new.d2) %in% "medv"])
#corrplot.mixed(cor(nomissing), order="hclust", tl.col="black")

#column to keep: 7, 14, 20 ,21, 25, 31, 43, 51, 60, 61, 69, 78, 83, 98, 117, 118
d2.new=new.d2[,c(7, 14, 20 ,21, 25, 31, 43, 51, 60, 61, 69, 80, 85, 98,
                 117, 118,120)]
#sum(is.na(d2.new)) #32 missing values
```

```
d2.off=d2.new[,c(10,16,17)]
is.na(d2.off)
sum(is.na(d2.off))


#aggr(d2.new) #Missing values in only one variable
#library(naniar)
#any_na(d2.new) #TRUE there are missing values
#miss_var_summary(d2.new)
#26 missing values in pf_association
#5 missing values in pf_ss_women
#1 missing value in pf_religion


#substitute the missing value in pf_religion with the mean of the column
#which(is.na(d2.new$pf_religion)) #160
#d2.new$pf_religion[160]=mean(d2.new$pf_religion[-160])
#miss_var_summary(d2.new)

#substitute the missing values in pf_ss_women with the mean of the column
#which(is.na(d2.new$pf_ss_women)) #9 12 29 101 129
#d2.new$pf_ss_women[c(9,12,29,101,129)]=mean(d2.new$pf_ss_women[-c(9,12,29,101,129)])
#miss_var_summary(d2.new)

#fill the missing values in pf_association
#library(DMwR)
#d2.nomissing=knnImputation(d2.new[, !names(d2.new) %in% "medv"])
#any_na(d2.nomissing) #no missing values left

a=merge(d1.new,d2.off, by = 0) #by=0 => by="row.names"
row.names(a)=a[,1]
a.new=a[,-1]

colnames(a.new)=c("HappinessScore", "LowerCI", "UpperCI", "Economy_GDP",
                  "Family", "Health_LifeExpectancy", "Freedom",
                  "Trust_GovernmentCorruption", "Generosity",
                  "Dystopya_residual", "pf_score", "ef_score", "hf_score")

summary(a.new)


#Descriptive analysis-------------------------------------------------------------

dim(a.new) #137 rows and 13 columns
str(a.new) #all numerical

sum(is.na(a.new)) #no missing data
pairs(a.new, col="mistyrose4", pch=19)

par(mfrow=c(1,2))
summary(a.new[,1])
colors()
hist(a.new[,1], main="", xlab="Happiness Score", col="mistyrose4")
quantile(a.new[,1])
boxplot(a.new[,1], col="mistyrose4")
```

```
summary(a.new[,2])
hist(a.new[,2], main="", xlab="Lower confidence interval", col="mistyrose4")
quantile(a.new[,2])
boxplot(a.new[,2], col="mistyrose4")

summary(a.new[,3])
hist(a.new[,3], main="", xlab="Upper confidence interval", col="mistyrose4")
quantile(a.new[,3])
boxplot(a.new[,3], col="mistyrose4")

summary(a.new[,4])
hist(a.new[,4], main="", xlab="Economy (GDP)", col="mistyrose4")
boxplot(a.new[,4], col="mistyrose4")

summary(a.new[,5])
hist(a.new[,5], main="", xlab="Family", col="mistyrose4")
boxplot(a.new[,5], col="mistyrose4")

summary(a.new[,6])
hist(a.new[,6], main="", xlab="Health (Life expectancy)", col="mistyrose4")
boxplot(a.new[,6], col="mistyrose4")

summary(a.new[,7])
hist(a.new[,7], main="", xlab="Freedom", col="mistyrose4")
boxplot(a.new[,7], col="mistyrose4")

summary(a.new[,8])
hist(a.new[,8], main="", xlab="Trust (government corruption)", col="mistyrose4")
boxplot(a.new[,8], col="mistyrose4")

summary(a.new[,9])
hist(a.new[,9], main="", xlab="Generosity", col="mistyrose4")
boxplot(a.new[,9], col="mistyrose4")

summary(a.new[,10])
hist(a.new[,10], main="", xlab="Dystopia residual", col="mistyrose4")
boxplot(a.new[,10], col="mistyrose4")

summary(a.new[,11])
hist(a.new[,11], main="", xlab="pf_score", col="mistyrose4")
boxplot(a.new[,11], col="mistyrose4")

summary(a.new[,12])
hist(a.new[,12], main="", xlab="ef_score", col="mistyrose4")
boxplot(a.new[,12], col="mistyrose4")

summary(a.new[,13])
hist(a.new[,13], main="", xlab="hf_score", col="mistyrose4")
boxplot(a.new[,13], col="mistyrose4")

hist(a.new)
x11()
par(mfrow=c(1,1))
hist(scale(a.new), main="", col="mistyrose4" )
```

```
round(cor(a.new),2)

library(corrplot)
x11()
corPlot(a.new, cex = 0.4, xlas=2,show.legend=F)

#PCA-------------------------------------------------------------------

pca=prcomp(a.new, scale=T)
round(pca$rotation,3)
screeplot(pca,type=c("lines"))
summary(pca)
#the first 2 principal component explain the 68% of the total variability
#3 --> 78%
#4--> 85%
x11()
biplot(pca, cex=.9, col= c("black","darkmagenta"))
plot(pca, col="mistyrose4")


res.pca = PCA(a.new,graph=F)

(bip1.2 = fviz_pca_biplot(res.pca, col.ind="cos2", repel = TRUE, axes = c(1,2))+
  scale_color_gradient2(low="green", mid="black", high="red"))
(bip2.3 = fviz_pca_biplot(res.pca, col.ind="cos2", repel = TRUE, axes = c(2,3))+
  scale_color_gradient2(low="green", mid="black", high="red"))

fviz_eig(res.pca, addlabels=TRUE, hjust = -0.3, barcolor = "mistyrose4",
         barfill="mistyrose4") +
  ylim(0, 70)
fviz_pca_ind(prcomp(scale(a.new)),ggtheme=theme_classic(),legend="bottom")
#score plot
fviz_pca_var(res.pca,col.var="cos2",repel=T) #Loading plot

fviz_contrib(res.pca,choice="var",axes=1,top=10)
fviz_contrib(res.pca,choice="var",axes=2,top=10)
fviz_contrib(res.pca,choice="var",axes=1:2,top=10)

#(var = get_pca_var(res.pca))
#res.km = kmeans(var$coord, centers = 4, nstart = 25)
#grp = as.factor(res.km$cluster)
#fviz_pca_var(res.pca, col.var=grp, legend.title="cluster")

#Cluster analysis-------------------------------------------------------

##Hierarchical method##

d=dist(scale(a.new), method = "euclidean") #important to scale the data

#choose the right method (using cophenetic distance)
fit=hclust(d, method="ward.D2")
coph=cophenetic(fit)
cor(d,coph) #0.5646204

fit.single=hclust(d, method="single")
coph.s=cophenetic(fit.single)
```

```
cor(d,coph.s) #0.4090292

fit.c=hclust(d, method="complete")
coph.c=cophenetic(fit.c)
cor(d,coph.c) # 0.6206036

fit.a=hclust(d, method="average")
coph.a=cophenetic(fit.a)
cor(d,coph.a) #0.6572907

#use the method "average" to do the analysis
plot(fit.a, cex=.7) #not good because we have chaining

#so try with complete
plot(fit.c, cex= .7)
rect.hclust(fit.c, k=3)
#rect.hclust(fit.c, k=4)

#plot(fit, cex=.7)
#rect.hclust(fit, k=3)
#rect.hclust(fit, k=4)

(groups.c=cutree(fit.c, k=3))
a.newmap=cbind(a.new,groups.c)
names(groups.c)

rownames(a.newmap)=ifelse(rownames(a.newmap)=='United States', 'USA',
                          rownames(a.newmap))
rownames(a.newmap)=ifelse(rownames(a.newmap)=='United Kingdom', 'UK',
                          rownames(a.newmap))
map = map_data('world')
unique(map$region)

thismap=mutate(map, fill = ifelse(region %in% row.names(subset(a.newmap, subset = groups
                                  ifelse(region %in% row.names(subset(a.newmap, subse
                                         ifelse(region %in% row.names(subset(a.newmap
                                                'NA'))))
unique(thismap$region)

ggplot(thismap, aes(long, lat, fill = fill, group=group)) +
  geom_polygon(colour="white") +
  ggtitle("Map of World with K = 3")

hier=cbind(apply(a.new[a.newmap$groups.c==1,],2, mean),
           apply(a.new[a.newmap$groups.c==2,], 2, mean),
           apply(a.new[a.newmap$groups.c==3,], 2,mean))

h32=((hier[,3]-hier[,2])/hier[,3])*100
h31=((hier[,3]-hier[,1])/hier[,3])*100
h12=((hier[,1]-hier[,2])/hier[,1])*100

round(cbind("1 vs 2 (%)"=h12,"3 vs 1(%)"= h31,
      "3 vs 2(%)"= h32),1)


##K-means##
```

```
SSV=vector(mode = "numeric", length = 15)
SSV[1]=(n - 1) * sum(apply(scale(a.new),2,var))
for (i in 1:15)SSV[i]=sum(kmeans(scale(a.new),centers=i,nstart=200)$withinss)
plot(1:15, SSV, type="b", xlab="Number of Clusters",
     ylab="Sum of squares within groups",pch=19, col="black")

#1.chose 3 the ideal number of clusters (from elbow chart)
kclust=kmeans(scale(a.new), 3, nstart= 200)
#2. calculate centroids
centr= aggregate(scale(a.new), by=list(kclust$cluster), FUN=mean)
#3. scatterplot of clusters
nk=3
pairs(scale(a.new), col=kclust$cluster, pch=19)
points(kclust$centers, col= 2:nk+1, pch=19, cex=2)

data.kclust=kclust$cluster

x11()
fviz_cluster(kclust, data = a.new,
             palette = "set2",
             geom = c("point", "text"),
             ellipse.type = "convex",
             ggtheme = theme_bw(), repel = TRUE
)

a.newmap[,15]=kclust$cluster

#map with clusters
map = map_data('world')
unique(map$region)

thismap <- mutate(map, fill = ifelse(region %in% row.names(subset(a.newmap, subset = V15
                                     ifelse(region %in% row.names(subset(a.newmap, subse
                                            ifelse(region %in% row.names(subset(a.newmap
                                                   'NA'))))
unique(thismap$region)

ggplot(thismap, aes(long, lat, fill = fill, group=group)) +
  geom_polygon(colour="white") +
  ggtitle("Map of World with K = 3")

k=cbind(apply(a.new[a.newmap$V15==1,],2, mean),
        apply(a.new[a.newmap$V15==2,], 2, mean),
        apply(a.new[a.newmap$V15==3,], 2,mean))

k32=((k[,3]-k[,2])/k[,3])*100
k31=((k[,3]-k[,1])/k[,3])*100
k12=((k[,1]-k[,2])/k[,1])*100

round(cbind("1 vs 2 (%)"=k12,"3 vs 1(%)"= k31,
            "3 vs 2(%)"= k32),1)

##K-medoids##

fviz_nbclust(scale(a.new), pam, method = "silhouette")+
```

```
  theme_classic()


x11()
pam.res = pam(scale(a.new),3)
fviz_cluster(pam.res, data = a.new,
             palette = "set2",
             geom = c("point", "text"),
             ellipse.type = "convex",
             ggtheme = theme_bw(), repel = TRUE
)


c=pam.res$clustering

c = case_when(
  c == 1 ~ 2,
  c == 2 ~ 1,
  TRUE ~ 3
)

a.newmap[,16]=c

thismap <- mutate(map, fill = ifelse(region %in% row.names(subset(a.newmap, subset = V16
                                     ifelse(region %in% row.names(subset(a.newmap, subse
                                            ifelse(region %in% row.names(subset(a.newmap
                                                   'NA'))))
unique(thismap$region)
x11()
ggplot(thismap, aes(long, lat, fill = fill, group=group)) +
  geom_polygon(colour="white") +
   ggtitle("Map of World with K = 3")

km=cbind(apply(a.new[a.newmap$V16==1,],2, mean),
        apply(a.new[a.newmap$V16==2,], 2, mean),
        apply(a.new[a.newmap$V16==3,], 2,mean))

km32=((km[,3]-km[,2])/km[,3])*100
km31=((km[,3]-km[,1])/km[,3])*100
km12=((k[,1]-km[,2])/km[,1])*100

round(cbind("1 vs 2 (%)"=km12,"3 vs 1(%)"= km31,
            "3 vs 2(%)"= km32),1)

a=cbind("Hierachical"=a.newmap[,14], "Kmean"=a.newmap[,15], "Kmedoid"=a.newmap[,16])
rownames(a)=row.names(a.new)
```

8 Image appendix

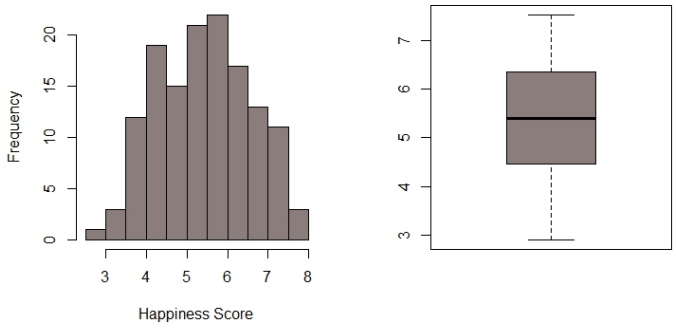

Figure 29: Happiness Score

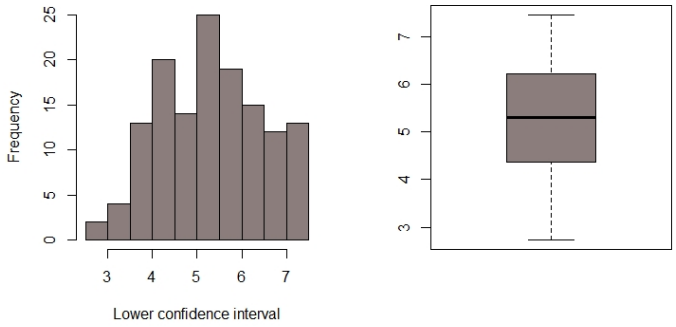

Figure 30: Lower confidence interval

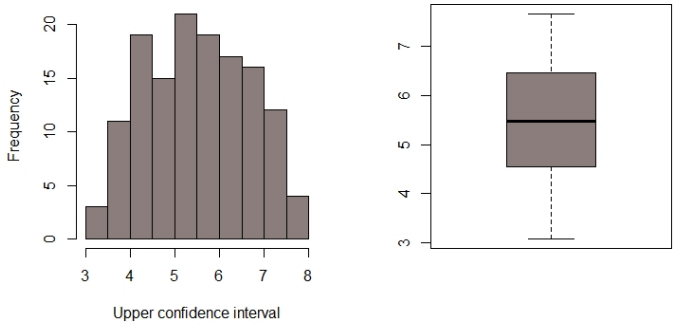

Figure 31: Upper confidence interval

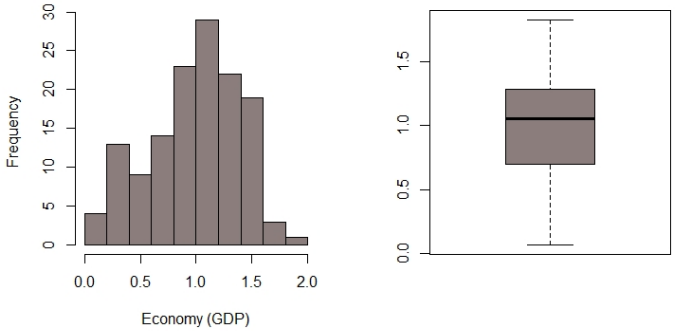

Figure 32: Economy (GDP)

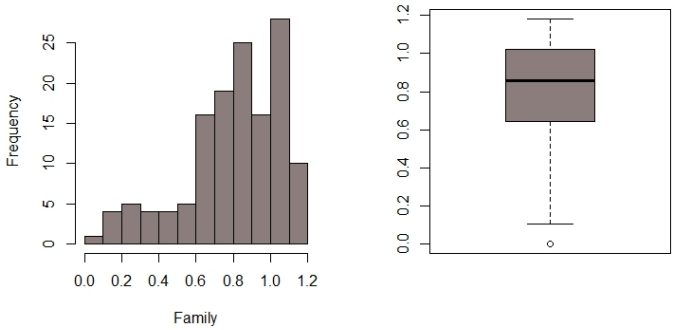

Figure 33: Family

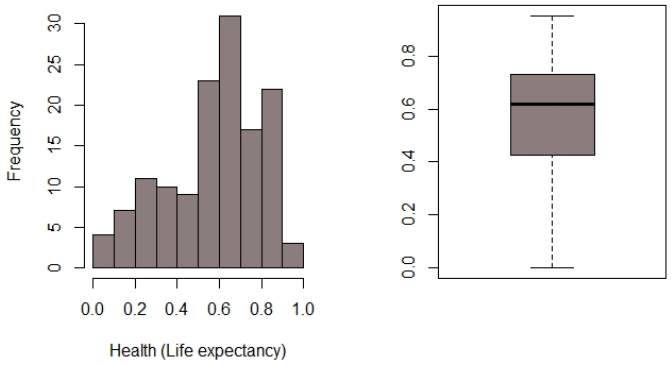

Figure 34: Health (Life Expectancy)

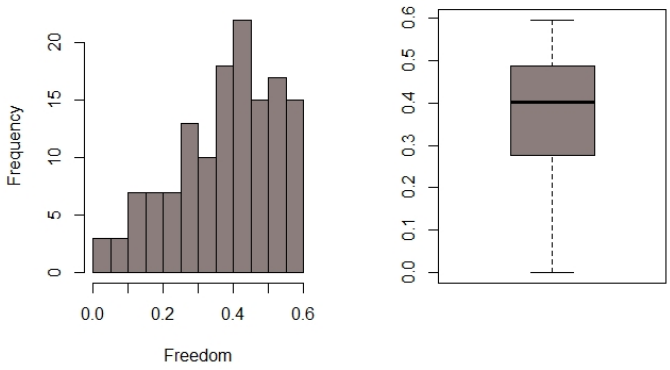

Figure 35: Freedom

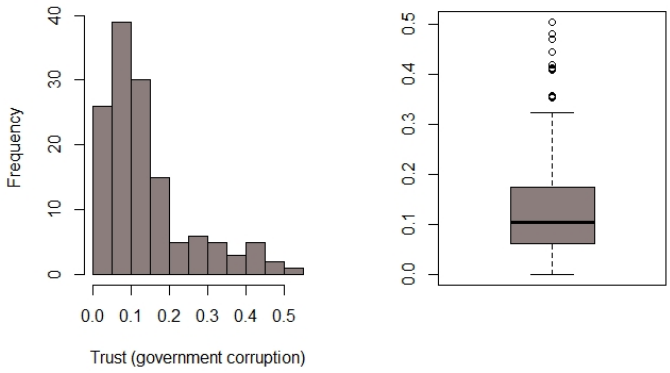

Figure 36: Trust (Government Corruption)

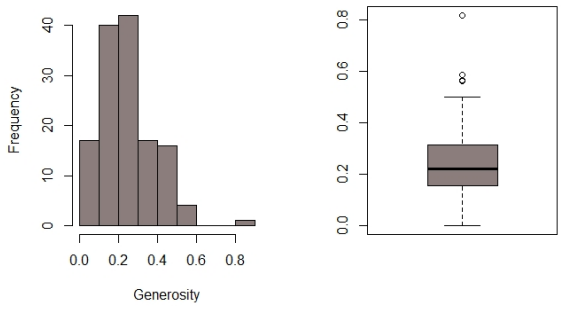

Figure 37: Generosity

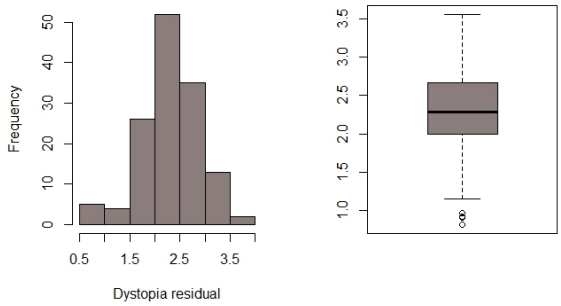

Figure 38: Dystopia residual

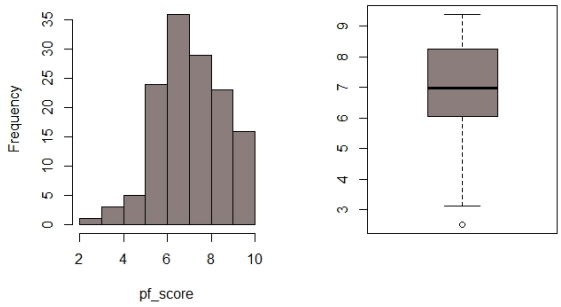

Figure 39: Personal Freedom score (pf score)

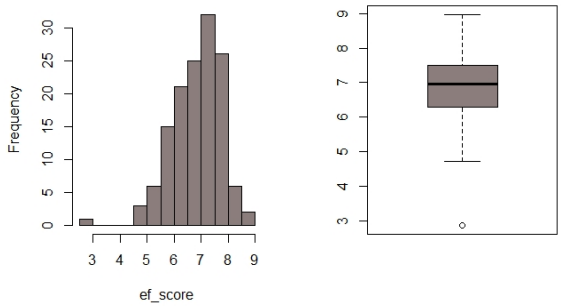

Figure 40: Economic Freedom (ef score)

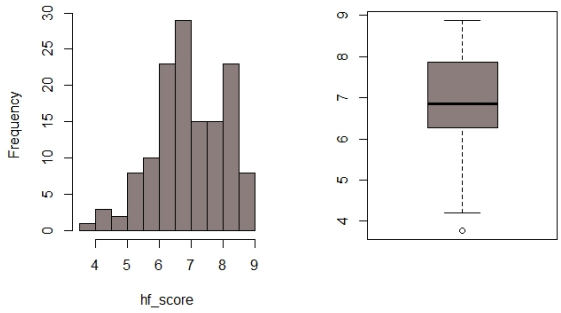

Figure 41: Human Freedom (hf score)