Unsupervised Learning Project

Unsupervised statistical learning methods take as input unlabelled observations which are records about different variables without a response. Therefore the modelling is focused on the detection of correlations among the features rather than prediction or inference on a specific target or outcome. These techniques are useful to better visualize the data, to reduce the dimensionality of the attributes and to discover similarities among groups of observations. In this specific analysis they find application in the study of a dataset regarding different socioeconomic characteristics for several countries. First a principal component analysis is implemented to summarize the variables information into a lower number of components. Then a K-means clustering algorithm is applied to highlight similarities among countries, grouping those most similar together. Concretely such analysis could be supportive for some international organization, e.g. the International Monetary Fund or the World Bank, in deciding to which countries addressing most of their short and long term economic aid.

I. Data Set

The chosen data report information concerning 9 different attributes, 5 of economic type and the remaining about health conditions, describing 167 different worldwide countries. Here below a quick description of the features:

- **country:** Name of the country.
- **child_mort:** Number of dead children under 5 years of age per 1000 live births.
- **exports:** Percentage of GDP per capita related with exports of goods and services[1].
- **health:** Percentage of GDP per capita dedicated to health expenditure.
- **imports:** Percentage of GDP per capita related with imports of goods and services.
- **income:** Net annual income per person in US dollars.
- **inflation:** Annual GDP growth rate.
- **life_expec:** The average number of years a new born child would live with the current living conditions.
- **total_fer:** The average number of children per woman over her lifetime if the current age-fertility rates (computed for different age ranges) remain constant.[2]
- **gdpp:** Gross domestic product per capita.

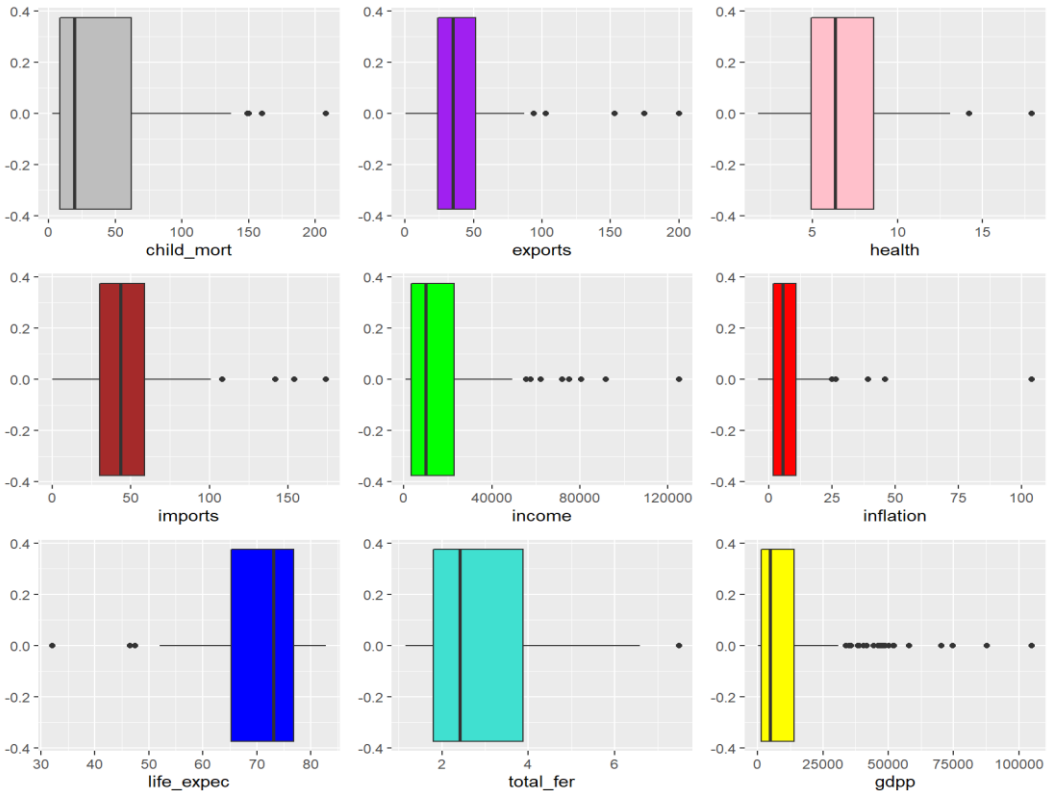
The dataset is already in tidy format that is each column represents a variable. In the next page a representative portion of the data set -table (1)-, together with the boxplots of the numerical values -figure (1), are reported. All the variables, except the countries' names, are continuous with feasible ranges and do not have missing values. The variable *inflation* is renamed as *gdp growth rate* for a matter of clarity. The analysis can begin, first by running a principal component analysis and then by clustering via K-means.

[1] GDP=Consumption+Investments+Government Spending+Exports-Imports.
[2] According to the OECD a total fertility rate of 2.1 is required to maintain constant the level of the population.

Table (1). Head of the dataset.

country	child_mo	export	health	imports	income	gdp_gr.r	life_expec	total_fer	gdpp
Afghanistan	90.2	10	7.58	44.9	1610	9.44	56.2	5.82	553
Albania	16.6	28	6.55	48.6	9930	4.49	76.3	1.65	4090
Algeria	27.3	38.4	4.17	31.4	12900	16.1	76.5	2.89	4460
Angola	119	62.3	2.85	42.9	5900	22.4	60.1	6.16	3530
Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200
Argentina	14.5	18.9	8.1	16	18700	20.9	75.8	2.37	10300
Armenia	18.1	20.8	4.4	45.3	6700	7.77	73.3	1.69	3220
Australia	4.8	19.8	8.73	20.9	41400	1.16	82	1.93	51900
Austria	4.3	51.3	11	47.8	43200	0.873	80.5	1.44	46900
Azerbaijan	39.2	54.3	5.88	20.7	16000	13.8	69.1	1.92	5840

Figure (1). Box plots of the features.



I. Principal Component Analysis

The principal component analysis, from now on referred as PCA, is an unsupervised statistical method used for representing data with a lower dimensionality. That occurs by building *m* components out of the original *p* features, where *m* is lower than *p*. For instance the first component assumes the following form:

$$z_{1i} = \phi_{11}x_{1i} + \phi_{21}x_{2i} + \cdots + \phi_{p_1}x_{ip}{}^{3}, \text{ where } \sum_{J=1}^{p} \phi_{j1}^{2} = 1$$

[3] Where $i \in [0,n]$, where n is the sample size.

Since the new component is a linear combination of the original variables is very important to scale[4] the values before applying such a technique. If that is ignored the variables having a larger variance would be overloaded. The loads *phi* are selected with the objective of maximizing the variance[5] of the projections in order to incorporate within the first component the greatest amount of information out of all possible linear combinations. The same reasoning is applied to the subsequent components with the further constrain that they must be uncorrelated among each other's. The geometrical intuition behind this concept is that the vectors of weights should be orthogonal[6] between all the different components.

The PCA grounds its theoretical foundations in the Singular Value Decomposition. The SVD decomposes the original matrix of the data X into three different matrices as shown here below:

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nr} \end{bmatrix} \begin{bmatrix} s_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{rr} \end{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \ddots & \vdots \\ v_{r1} & \cdots & v_{rp} \end{bmatrix}$$ [7]

Where U is a unitary matrix[8] collecting the left singular vectors[9], S is a diagonal matrix of singular values[10] and V^T is the transposed of a unitary matrix of right singular vectors[11]. The loads *phi* of the components are obtained by multiplying matrices U and S. As a rule of thumb in deciding how many components should be considered a common approach is to retain at least the 70%-80% of the energy, or total variability. That implies considering only the first *k* singular values:

$$\sum_{i=1}^{k} s_i^2 \geq 0.7 \sum_{i=1}^{r} s_i^2$$

In figure (2) the loads for the first five principal components are visualized graphically. For example the first component is negatively impacted by the features *total fertility*, *gdp growth rate* and *child mortality*. The largest difference among the loadings is between *child mortality* and *life expectancy* therefore meaning that countries having an high score for one of the two variables tend to have a low score for the other. On the other hand the largest loads in absolute terms in the second principal component are *imports* and *exports* which have a negative impact on it. Their most relevant counterparts are instead *health* and *life expectancy*. Similarly in PC3 the highest negative values are scored by *gdp growth rate* and *income* as opposed to *health* and *imports*.

---

[4] $x_i \rightarrow \frac{x_i - u_x}{\sigma_x},\ i = 1,2,\dots,n.$

[5] $\max_{\emptyset} \left[\frac{1}{n}\sum_{i=1}^{n} z_{1i}^2\right]$, variance for centred values.

[6] $\phi_i^T \phi_j = 0, \forall i,j$. Where *i* and *j* represents two different components.

[7] $X=USV^T$.

[8] $UU^{-1}=I$.

[9] $u_i = \frac{Xv_i}{s_i}$.

[10] $s_i = |Xv_i|$. Each singular value is the norm of a *n×1* vector. Each singular value represents the amount of variance from the original matrix summarized by the corresponding component.

[11] $argmax_v \sum_{i=1}^{n} |x_i v|^2 = argmax_v |Xv|^2$. In short the right singular vectors are those maximizing the projections of the original entries. Equivalently they identify the best fit line minimizing the overall distance of the entries from it.

Figure (2). First five partial components loadings.

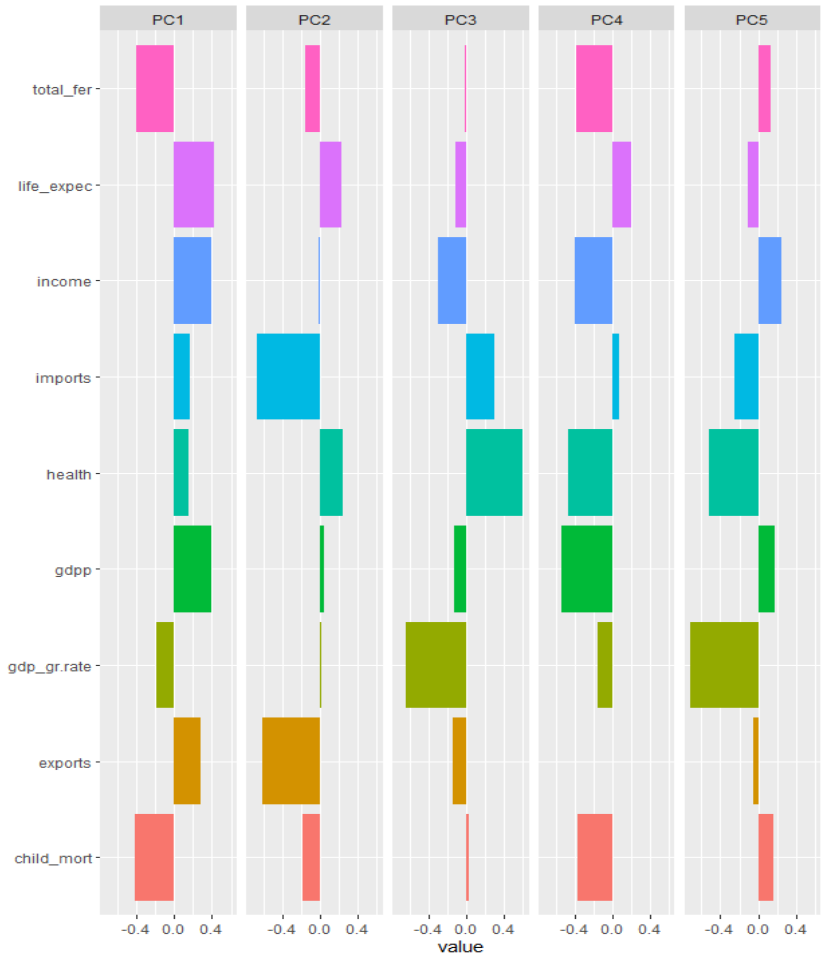
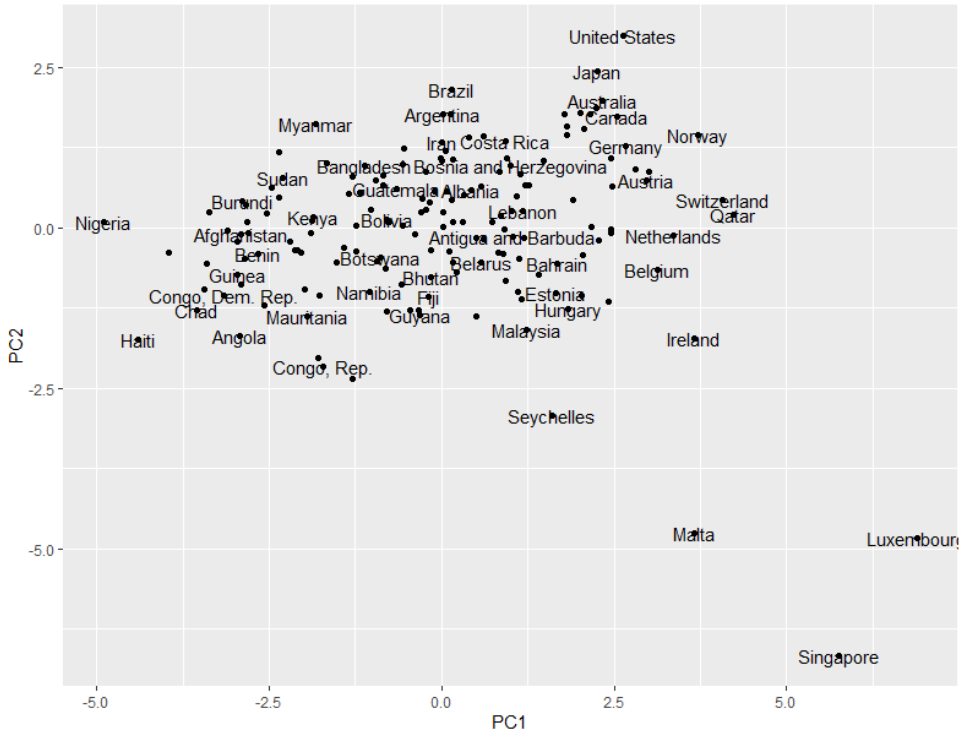


Figure (3). First two partial components scores.



Those countries scoring high in PC1 like Luxembourg, Singapore or Qatar are those having the higher life expectancy and overall better economic conditions such as high gdp per capita and high personal income. At the other end of the spectrum countries like Nigeria, Haiti or Congo have a high fertility rate, a high child mortality and worst economic conditions. Higher values in terms of second component mean better living conditions such as higher health expenditure and greater life expectancy. Among the others there are countries like the United States and Japan. On the contrary a low score means that the country is a top exporter/importer and that its economy is highly dependent from foreign ones. Examples are Malta or Singapore.

The first component represents by itself about 46% of the total energy of the data. This means that half of the information included in this dataset can be summarized by one single component whose greater loadings in absolute terms are those regarding fertility, child mortality, income and life expectancy. With the second component it is grasped another 17% of the total variability and with the third, somewhat much more tricky in the interpretation, the 13%.

Resuming the PCA permitted to reduce significantly the dimensionality of the data, from 9 features to 3 components, with the adverse effect of losing about 24% of the variance.

III. K-Means Clustering

K-means is a clustering algorithm useful for detecting similarities among the observations by defining *k* non-overlapping subgroups in which these observations can be partitioned, where *k* is decided a priori.

The algorithm applies the following steps:

1. Randomly assign each observation to one of the *k* clusters.
2. For each cluster compute the centroid's coordinates[12].
3. Once the centroids are defined assign each observation to the cluster whose centroid is the closest[13].
4. Iterate from step 2 until there are no more changes in the partitions.

The algorithm will drive to the minimum the within-cluster variation for each of the *k* clusters while maximizing the between variation among them. The within-variance is defined as the sum between the squared differences of all the pairwise observations belonging to a specific cluster:

$$w(C_k) = \frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{J=1}^{p}\left(x_{iJ} - x_{i'_J}\right)^2 \quad (1)$$

[12] A vector of means, computed for each predictor *p.*

[13] Based on the Euclidean distance. Let C_k (c_{1k} , $c_{2k, \ldots,}$ c_{pk}) be the coordinates of centroid *k* and X_i (x_{1i}, x_{2i},…,x_{pi}) the coordinates in $\mathbb{R}^p$ of the i^{th} observation. Their Euclidean distance is: $\sqrt{\sum_{J=1}^{p}(c_{kJ} - x_{iJ})^2}$

[14] Number of observations in cluster *k.*

[15] Squared Euclidean distance.

The formula just mentioned considering pairwise distances is exactly the same as twice the summation of the differences between each observation and the cluster means, that are ultimately the values of the current centroids[16].

Before applying the algorithm it is very important to standardize the values. That must be done with all the algorithms based on the concept of distance, as already argued for PCA. For computational reasons the K-means is run with a maximum number of iterations equal to 30. Moreover it is applied 9 times with different number of clusters with the aim to detect the optimal *k*. To do so the elbow rule, a heuristic approach, is adopted. It consists in retaining the clusters which drive down the within variance more quickly. That occurs when the curve in figure (3) shows the most pronounced "elbow". As can be seen after 4 centroids the within variation declines more slowly. It is not a global optimum but rather a local one. A still image of an interactive three-dimensional plot of the 4 clusters is reported in figure (4). For a matter of visualization only three dimensions at a time can be explored.

Figure (3). Within variance for each *k*.

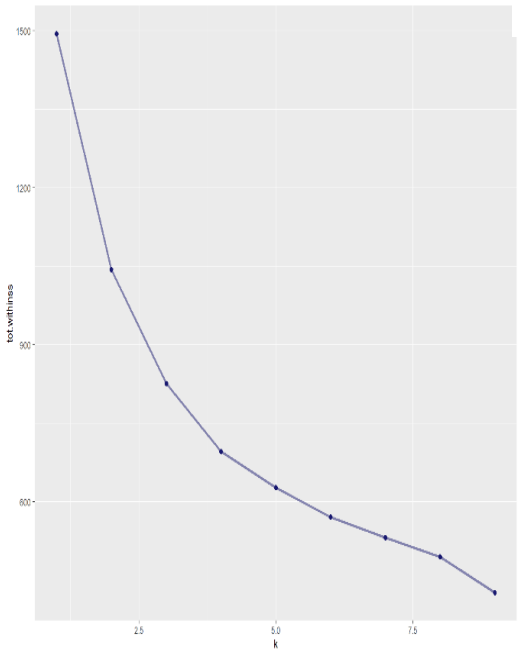
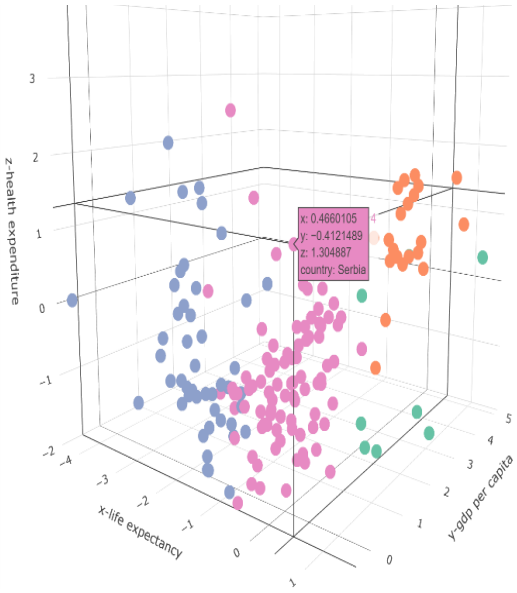


Figure (4). Three-dimensional plot of the 4 clusters.



The left most groups share similar values considering both the gross domestic product per capita and the health expenditure while they differentiate considerably on the life expectancy aspect. While the blue group, predominantly populated by African countries, has a low life expectancy, the pink one, grouping mainly Slavic, south American and Asian countries, has a greater one. Among 167 countries the vast majority belongs to these two groups, about the 80%. The rest is clustered in other two groups having an higher per capita GDP. The orange one, the biggest in size among the two, collects Commonwealth and European countries, having higher relative values for health expenditures with respect those green, e.g. Qatar, Kuwait or United Arab Emirates.

[16] $2\sum_{i\in C_k}\sum_{j=1}^{p}\left(x_{iJ}-\bar{x}_{Jk}\right)^2$. Equal to expression (1).

IV. Trimmed K-Means

The traditional k-means algorithm however assumes no outliers since their presence might be too much influential, leading to ineffectual results. Since in this analysis there are several outliers, as evident from picture (1), a robust method named trimmed k-means is introduced. The intuition behind the trimmed k-means is straightforward: exclude the furthest *n(α-1)* observations from the initial centroids. In this way $n \times \alpha$ observations will be left unassigned avoiding to affect the clustering procedure. The minimization problem becomes of the kind:

$$arg\min_{y}{}^{17} \min_{C_1 \dots C_k} \left\{ \sum_{k=1}^{k} \frac{1}{|C_k|} \sum_{ii' \in Ck} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 \right\}$$

Except for the first term the minimization problem remains the same. Given that the outliers to be dropped depend on the initial centroids[18] it is reasonable to run the algorithm multiple times before selecting the model with the lowest within variance.

Here below a demonstration by applying both a normal and a trimmed k-means using only two dimensions is given. The selected features are *GDP per capita* and *health expenditure*. Both clustering procedures are run with 4 centroids and maximum 30 iterations. It can be seen from picture (6) that the black countries are classified by the trimmed k-means as outliers and therefore excluded from any cluster. Such an exclusion however improves the clustering results. For instance Ireland, that was classified within the red cluster in figure (5), the one including very few countries as Qatar or Kuwait, is then classified together with the other European countries which indeed score more similar results for these two features.

Figure 2. Normal k-means.

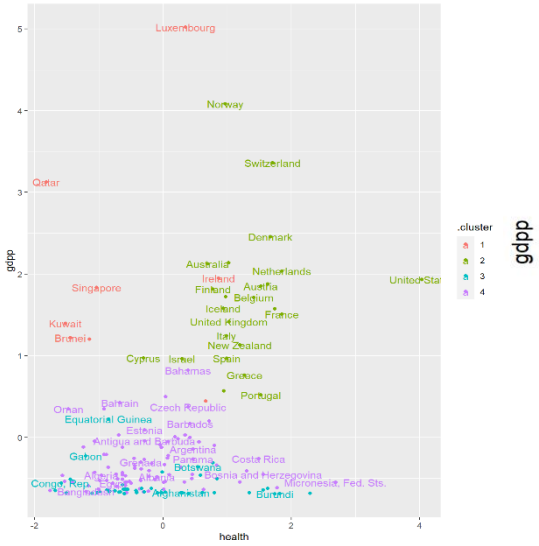
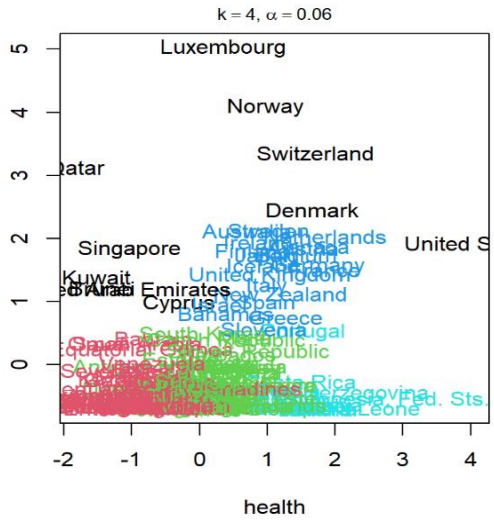


Figure 1. Trimmed k-means.



[17] Where y ranges from 1 to *n(1-α)*.

[18] The initial centroids coincide with some random observations. Given such an initialization it is expected that the true outliers will be detected. That could not be assumed if the initialization method considered a random value within the ranges of the observations instead of the observations themselves.

V. Conclusions

The analysis of aggregate socio-economic factors at national level can be useful in a bunch of different contexts. For instance an international organization committed in fostering growth in developed economies might be interested in finding which are the characteristics that mostly discriminate between rich and developing economies. In addition it might find useful grouping together those most similar countries in order to categorize them in underdeveloped, developing and rich economies. To answer these questions two unsupervised statistical learning methods have been applied to a data set providing 9 features for each of the 167 nations.

The principal component analysis revealed that 3 components summarize well the 76% of the total amount of information. An high score in the first principal component is determined by a high *life expectancy*, a high net *income*, a low *child mortality* as well as a low *fertility rate*. The second component instead contrasts *health expenditure* and *life expectancy* with the country's attitude in trading with foreign economies. Finally the last one is positively related with *health expenditure* and negatively with the *GDP annual growth rate*.

In the second part of the analysis the k-means algorithm grouped together the most alike countries. Four clusters have emerged, one for African countries, one for Asian and South American, one for European and one for the richest Arab countries, together with some other small and rich nations as Luxembourg and Singapore. However some inconsistencies arised in the results. For example Ireland was assigned to that very last group despite having similar results to the rest of European countries. To settle this issue, caused by the presence of influencing outliers, a trimmed k-means have been applied. Such a technique avoids to classify those considered extreme observations leading to more robust results.