Unsupervised learning analysis on world countries data

Alfredo Funicello

July 20, 2022

Abstract

HELP International, source of the data, is an ONG playing an active role in efforts to improve the quality of life of some of the most vulnerable populations in the world. The following analysis will be employing socio-economic markers from 167 world countries and exploiting unsupervised learning techniques as PCA, k-means and k-medoids, to extract and understand underlying similarities between living conditions in countries.

1 Data Presentation

The dataset is composed by set of economic indicators and a set of socio-medical indicators from 167 countries.

The dataset has been uploaded under the name "Unsupervised Learning on Country Data" on the Kaggle platform. The scope of the analysis is to understand if it's possible to uncover any meaningful logical structure of countries in order to explore the defining characteristics of the said groups.

The dataset is composed by 167 rows and features the following variables:

- Country, name of the country
- gdpp, GDP per capita. Calculated as the Total GDP divided by the total population.
- income, net income per person.
- inflation, measurement of the annual growth rate of the Total GDP
- exports, exports of goods and services per capita. Given as %age of the GDP per capita
- imports, imports of goods and services per capita. Given as %age of the GDP per capita
- health, total health spending per capita. Given as %age of GDP per capita
- life_exp, average number of years a new born child would live if the current mortality patterns are to remain the same
- child_mort, amount of deaths of children under 5 years of age per 1000 live births
- total_fer, amount of children that would be born from each woman considered current age-fertility rate

2 Data Manipulation

The data has already been cleaned and doesn't present any missing or incoherent values.

The `continent` variable has been added from the `Country-data.csv` file.

3 Data Exploration

From what can be see in Figure 1 the data presents heterogeneity of scale and many outliers in all of the features.

The different measurements scales between the features will be handled by using the `scale()` function to standardize values in highly scale-dependent techniques such as PCA or clustering.

The outliers are decided to be kept because they hold relevant information on the characteristics of the countries and excluding them would be detrimental to the scope of the analysis.

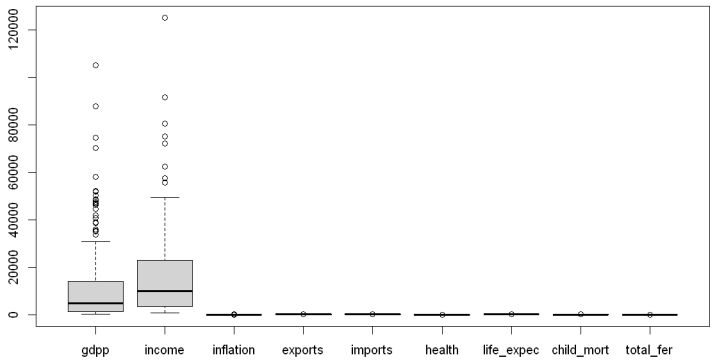

Figure 1: Boxplot of the features in the dataset

The correlation heatmap in Figure 2 highlights some important relations between the features. The most interesting correlations are those between socio-medical variables and economic ones; as it can be seen `life_expec` is positively correlated with `gdpp`, which shows that people living in richer countries live longer, but negatively correlated with `total_fer` and `child_mort` which are common characteristics of underdeveloped countries. `total_fer` is in fact negatively correlated with `gdpp`.

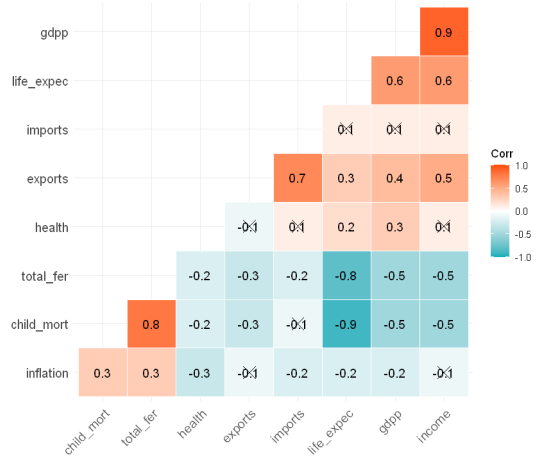

Figure 2: Correlation heatmap between the features

4 PCA Exploration

In this section PCA will be used for exploration and visualization purposes.

PCA has been applied to the scaled dataset.

The Scree plot in Figure 3 plots the percentage of variance explained by the PCs; PC1 and PC2 explain a cumulative 63% of variance, the cumulative value reaches 94% by including up to the 5th dimension.

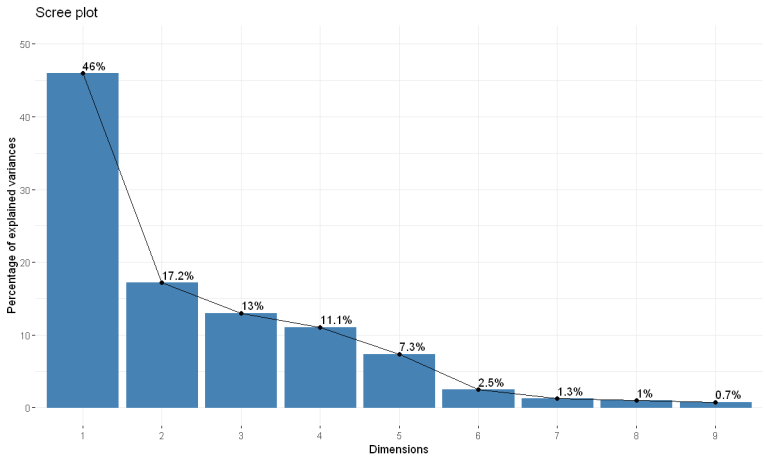

Figure 3: Scree plot

Through the variable correlation plot in Figure 4 it is possible to understand the relation between the PCs in output from the PCA and the variables. The previously explored positive correlations between variables are shown in variables grouped together, while negatively correlated ones lie in opposite directions. Most of the variables are well represented by PC1 and PC2, except for `inflation` and `health`. Figure 5 shows in fact that they're mostly contributing to PC3, PC4 and PC5.

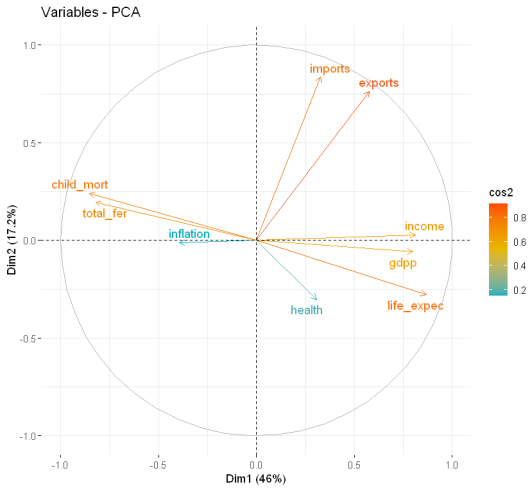

Figure 4: Correlations between the PCs and the variables

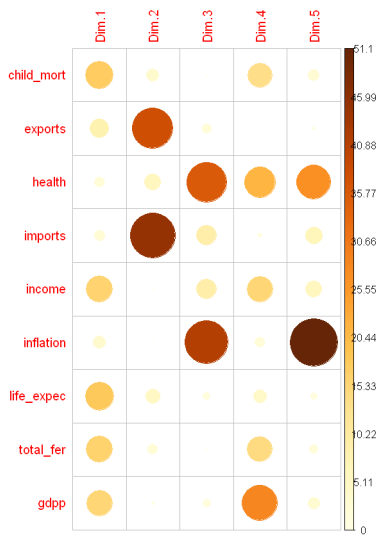

Figure 5: Correlations between the PCs and the variables

The Biplot in Figure 6 is already providing substantial information about continent characteristics. Bigger points in the plot are the centroids for each of the continents; African countries are distributed on the opposite direction to `gdpp` and `life_expec` but rank high on `child_mort` and `total_fer`. South America is interestingly distributed on the opposite direction from `imports` and `exports` suggesting that it's a continent characterized by isolationist economies. European countries are all placed away from `child_mort` and `total_fer` while being in the direction of `gdpp` which describes well developed economies with slow growing populations. Asia shows to have high internal variance on the markers.

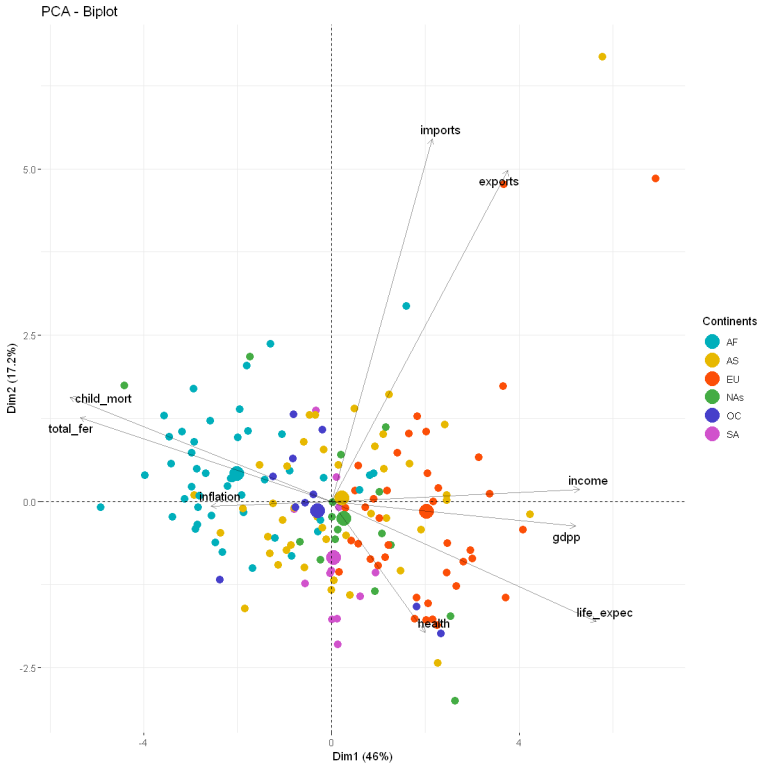

Figure 6: Biplot showcasing variables and individuals

Figure 7 is the same Biplot featuring country names of the points that are distributed away from the center.

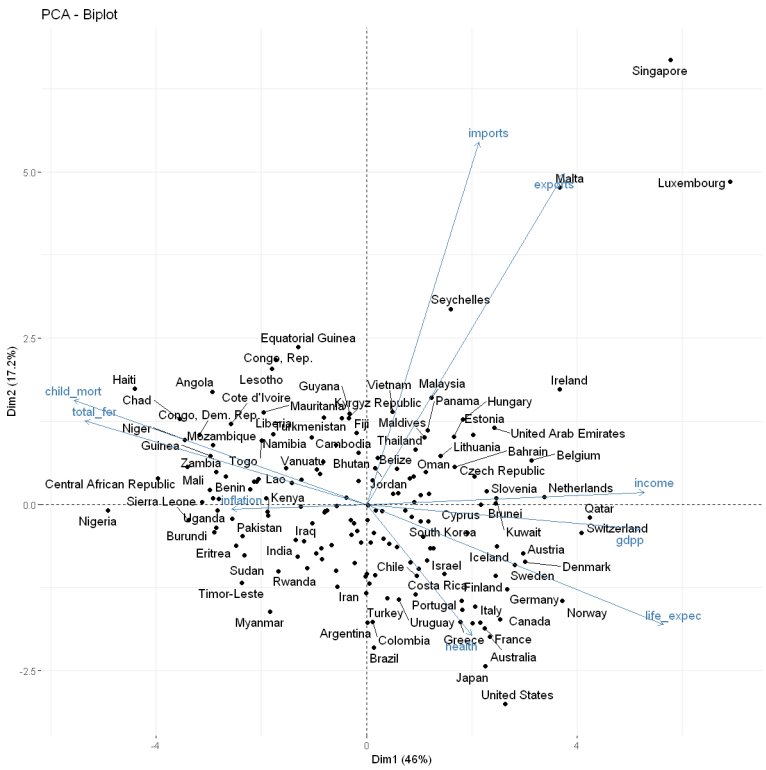

Figure 7: Biplot showcasing variables and individuals

5 K-Means Clustering Exploration

First step of the clustering process is assessing the clustering tendency. The Hopkins statistic has been chosen to do so, the dataset has a Hopkins value close to 1 (approx. 0.98) with 0 p-value under the null hypothesis of spatial randomness.

Figure 8 and 9 are the output plots of the elbow and silhouette methods ran on K-Means. The elbow method seem to suggest 6 clusters, while the silhouette has highest values between 2 and 5.

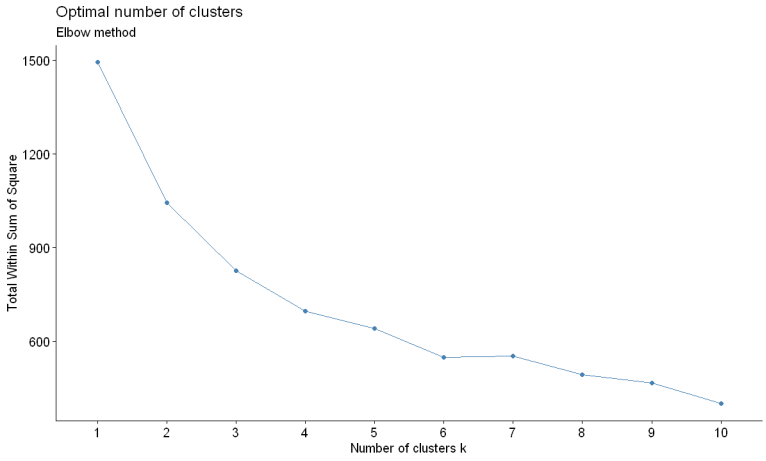

Figure 8: Elbow method on K-Means

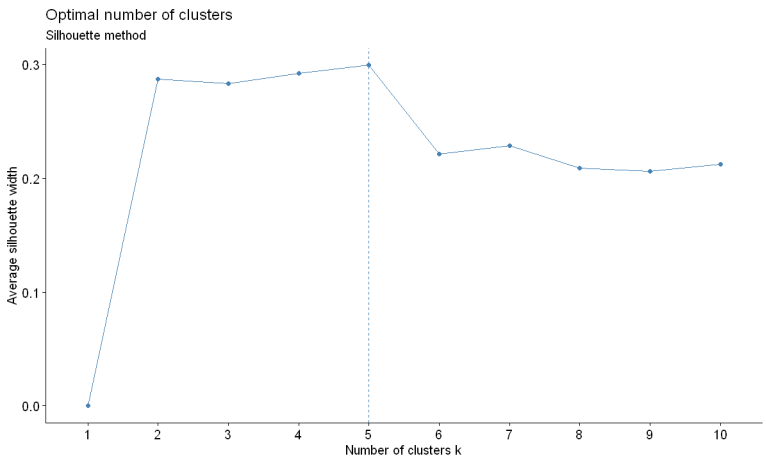

Figure 9: Silhouette method on K-means

Figure 11 shows the Kmeans clusters evolution from k=2 to k=7. Most of the variance that leads to splitting the clusters lays almost parallel to PC1, reason is that PC1, as we found out in the PCA step, is by far the most explaining principal component, explaining 46% of the variance, compared to only 17% for PC2. From what is possible to see in Figure 10, where the representation of k=5 clusters are overlaid on representations of the variable loadings the separation between the clusters is mostly because of differences on the `income-gdpp-life-expec` to `child_mort-total_fer` axis. Given that high child mortality is associated with poverty[1] and that high fertility is tied to low economic development[2], we understand the splits on the said axis to be creating separation between the countries based on their development level.

As it's possible to see in Figure 11 the clusters when k=2 are clearly separated on PC1, in k=3 we see a central cluster emerging, which might be interpreted as having characteristics that are a middle ground between the left and right clusters. k=4 leads to having a separate cluster for the 3 exports-imports outliers in the upper right corner. k=5 leads to a singleton split on Nigeria. k=6 causes a split in the central cluster. k=7 further splits the middle-ground countries. (More exploration on the nature of these close to the center clusters is done in the next chapter).

[1]Poverty, urban-rural classification and term infant mortality: a population-based multilevel analysis https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6343321/

[2]The Link between Fertility and Income, https://www.stlouisfed.org/on-the-economy/2016/december/link-fertility-income

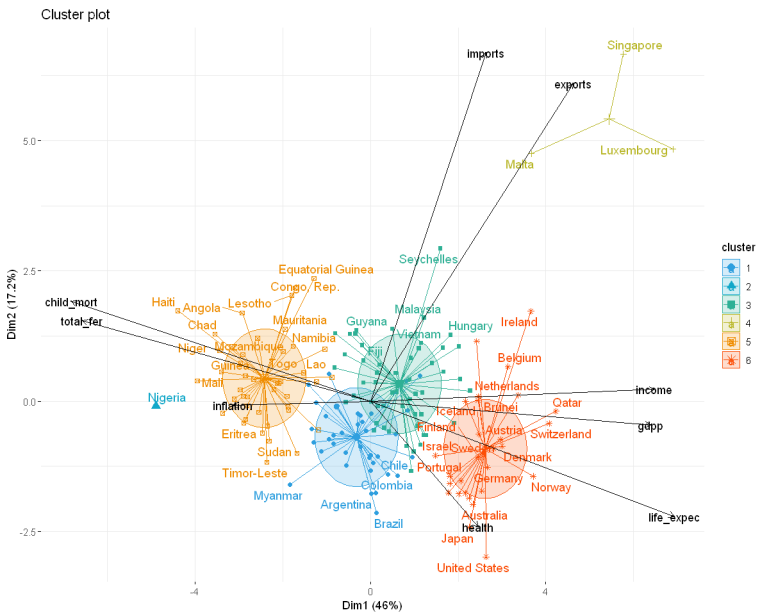

Figure 10: K-Means clusters with K=5

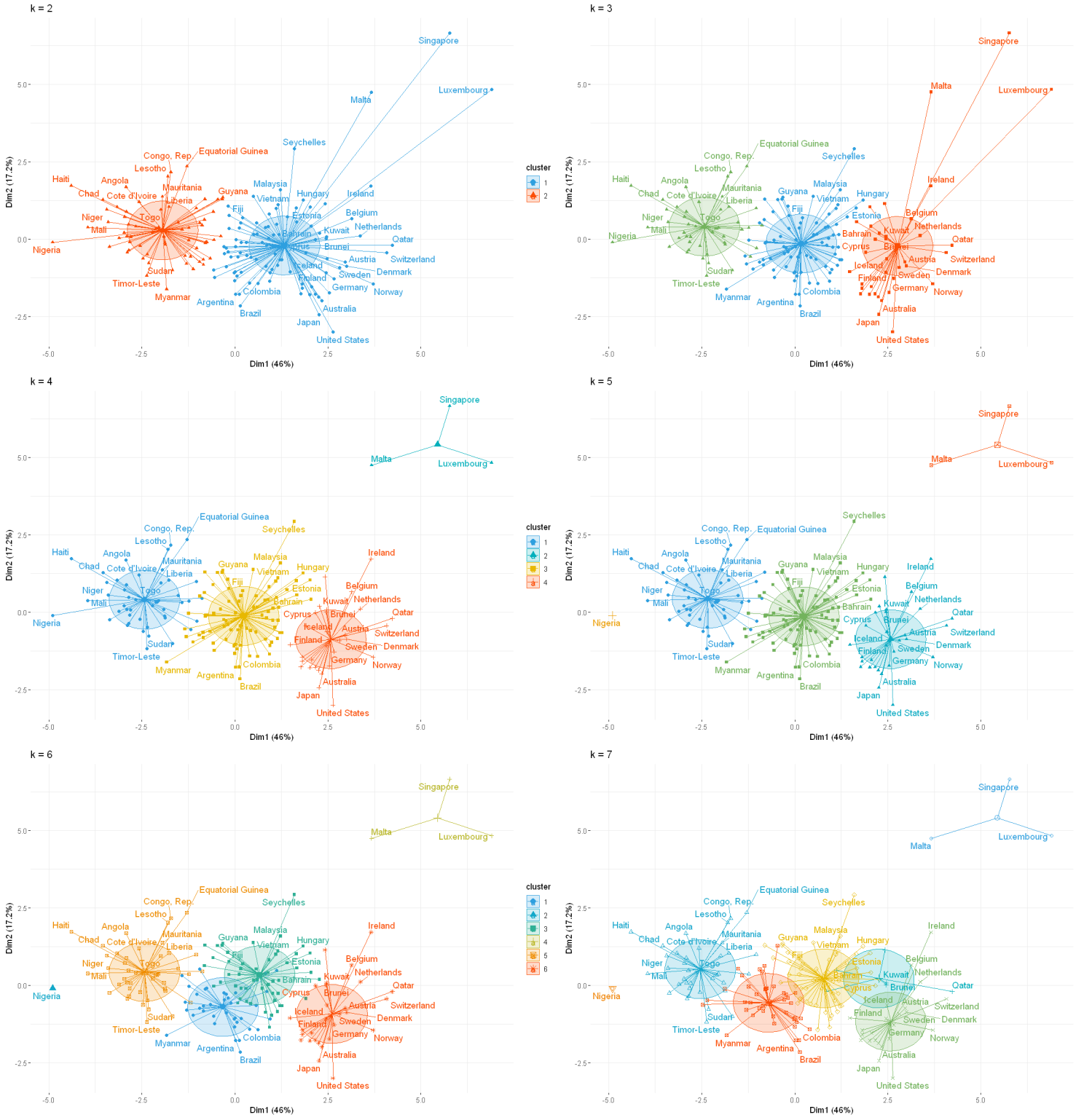

Figure 11: Evolution of K-Means

6 K-Medoids Clustering Exploration

The nature of the analysis doesn't allow for any fix for the outlier values and some of the clusters created by the K-means algorithm seem to be largely dictated by outlier values; examples are the singleton cluster on Nigeria and the import-export outliers cluster. K-Medoids is a possible outlier robust alternative to K-means

since it uses observations as medoids for the clustering.

In this section results from the PAM algorithm, ran on Manhattan distance, an outlier robust alternative to Euclidean distance, will be analyzed with the objective of comprehending if the outliers are heavily influencing the clusters.

As with K-means the TWSS and the silhouette average width values are plotted (Figure 12 and 13): the elbow method is not so clear about the optimal number of clusters, there seems to be a slight elbow at k=5, the silhouettes suggest k=2 and k=3 with another increase at k=6.

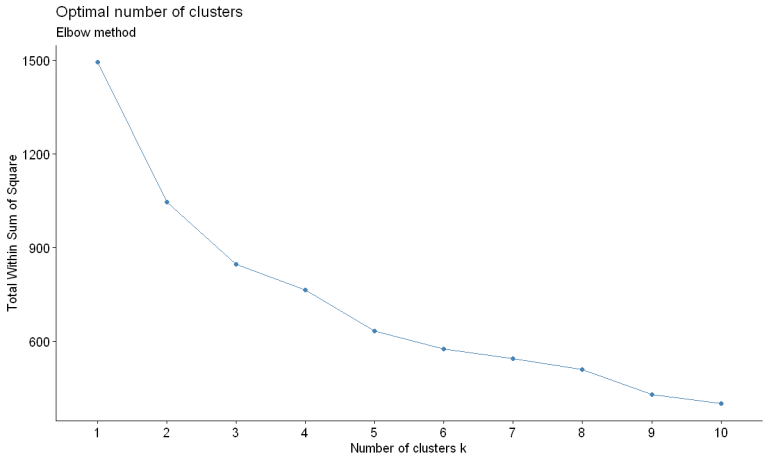

Figure 12: Elbow method on PAM

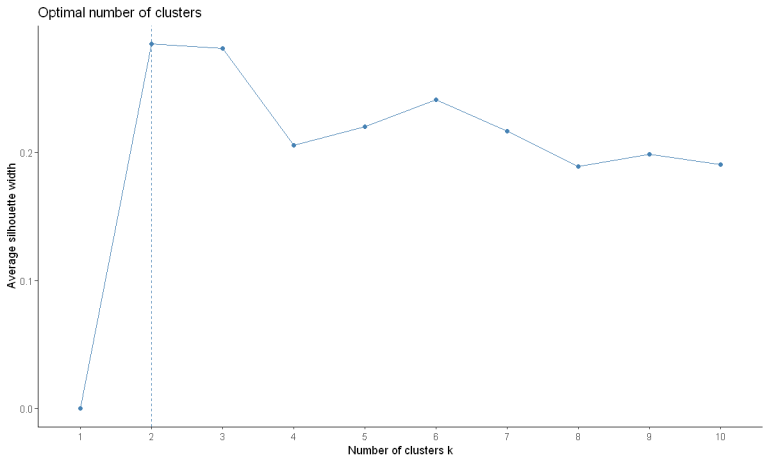

Figure 13: Silhouette method on PAM

Figure 14 showcases the evolution of the clusters as k rises. The tendency to split horizontally on PC1 seems to be maintained on lower amount of clusters, but on k=4 and k=5 the clusters are more well formed and less influenced by the outliers.

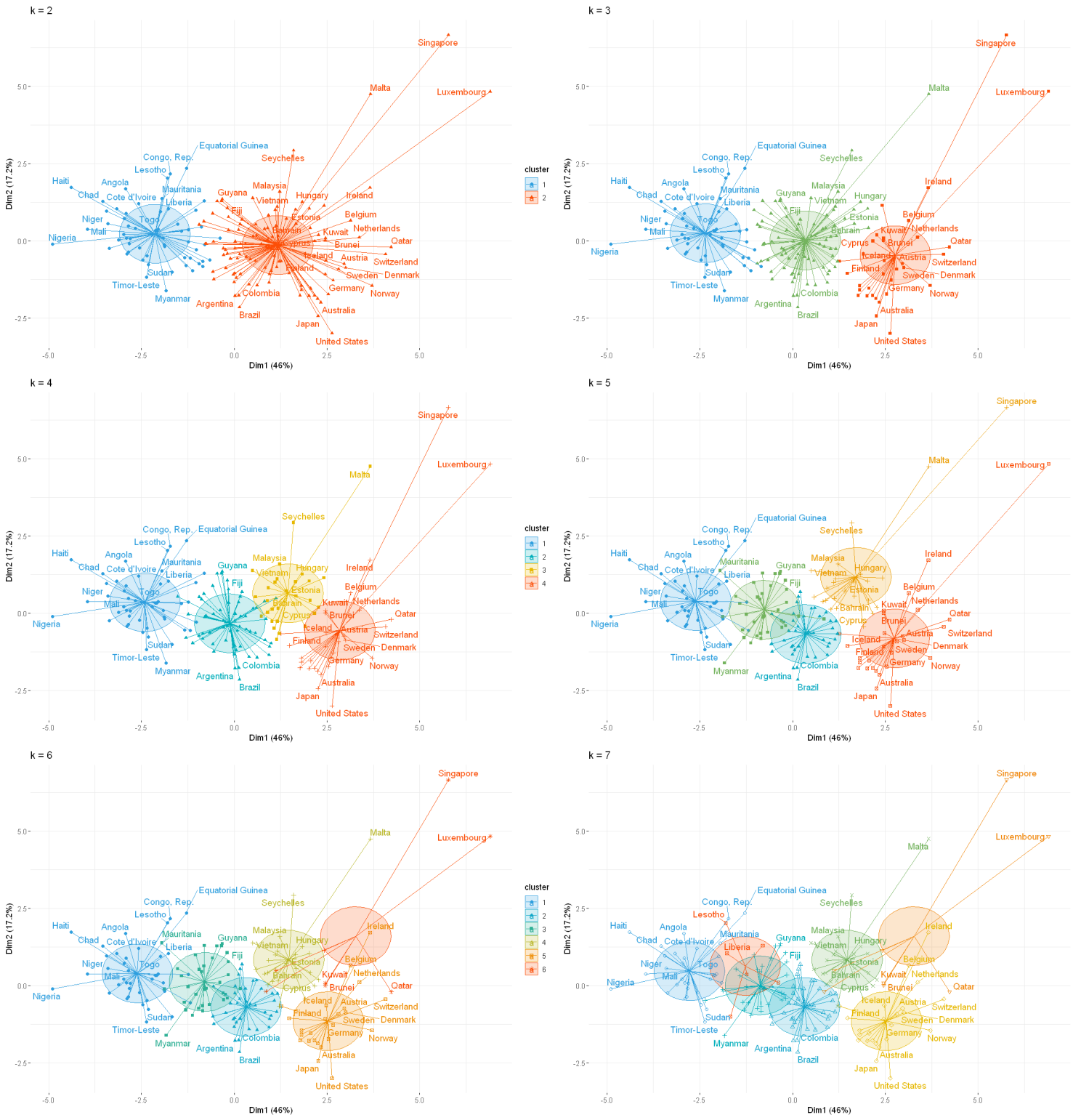

Figure 14: Evolution of the clusters from PAM

Figure 15, 16 and 17 show the outline of values for the medoids from each cluster. While actual values can bee seen in figure 18, the values shown in the figures are standardized, the zero threshold is set as the mean point for the column and the values have to be interpreted as relative to the average country.

As it can be seen cluster 1 has low economic indicators, low health spending, low life expectancy but high

values for child mortality and fertility. Cluster 2 has still negative values for economic indicators, but the magnitude is less important, life expectancy is positive and child mortality and fertility are under control. Cluster 3 has very positive economic indicators, possibly negative inflation, very positive health spending and very negative child mortality and fertility.

Full distributions for each of the variables for the 3 clusters is depicted in Figure 20.

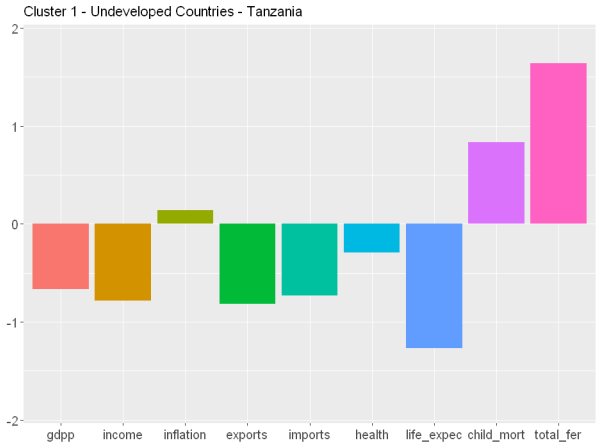

Figure 15: Distribution of values for Cluster 1

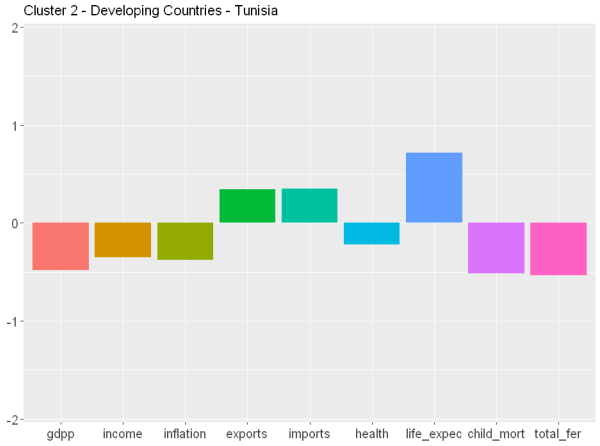

Figure 16: Distribution of values for Cluster 2

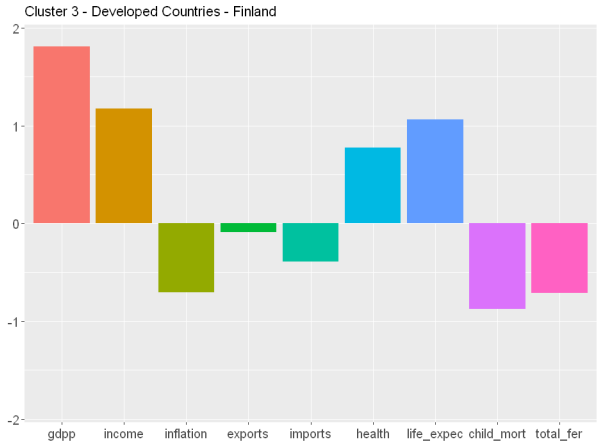

Figure 17: Distribution of values for Cluster 3

	cluster	gdpp	income	inflation	exports	imports	health	life_expec	child_mort	total_fer
Tanzania	1	702	2090	9.250	18.7	29.1	6.01	59.3	71.9	5.43
Tunisia	2	4140	10400	3.820	50.5	55.3	6.21	76.9	17.4	2.14
Finland	3	46200	39800	0.351	38.7	37.4	8.95	80.0	3.0	1.87

Figure 18: Values for the Medoids of the 3 clusters

The barplot in Figure 19 illustrates the distribution of continents throughout the clusters; cluster 1 is mostly dominated by African countries, while cluster 2 has many Asian and European countries, cluster 3 is mainly European countries.

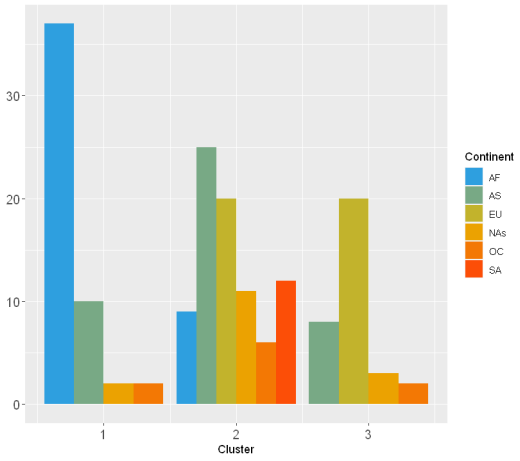

Figure 19: Distribution of continent in the 3 clusters

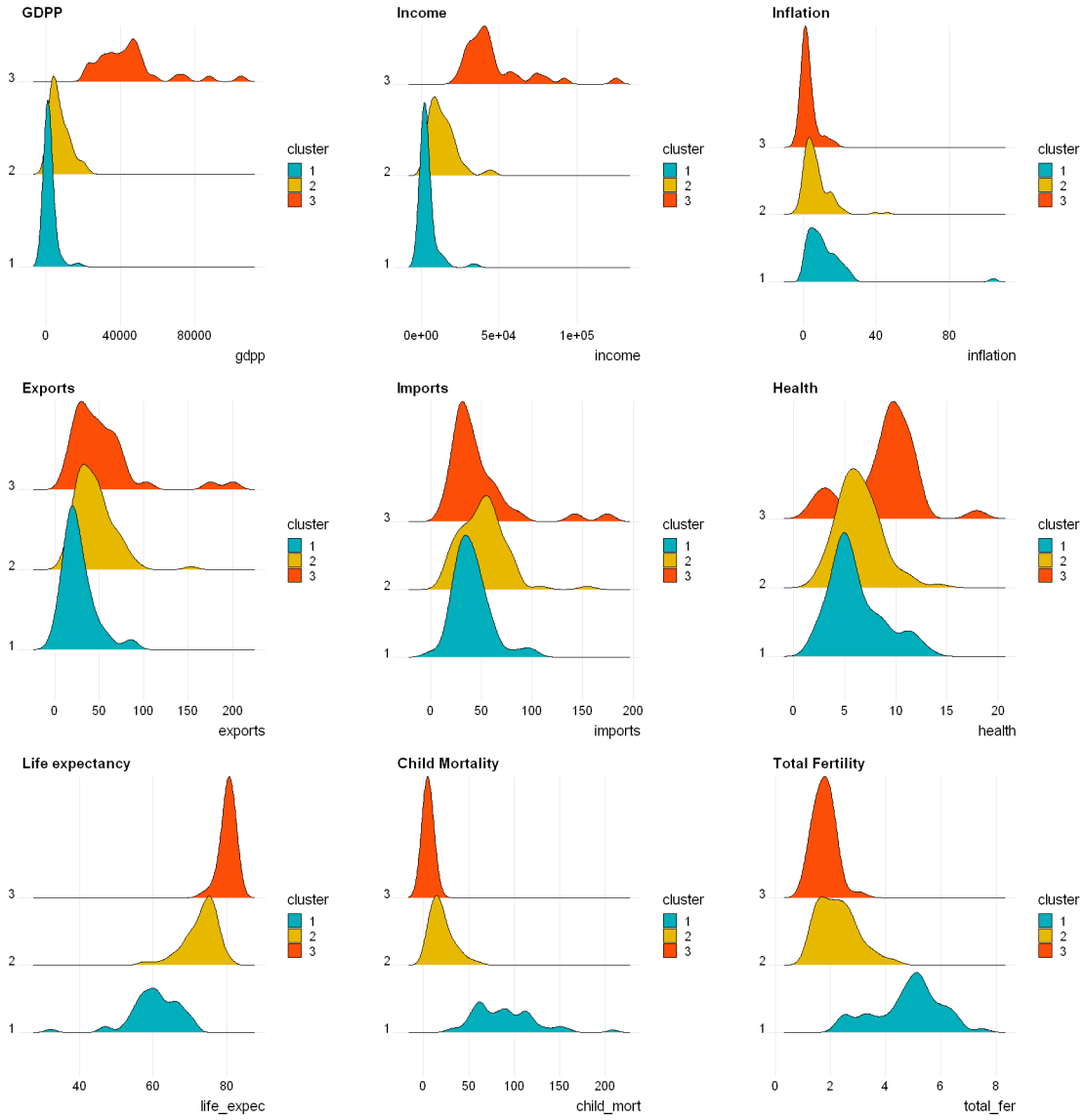

Figure 20: Distributions of the values for each of the features when k=3

In order to take a more detailed look at the characteristics of the countries the exploration will further continue on k=5.

Figure 21: Visualization of the points from the 5 clusters on PCA

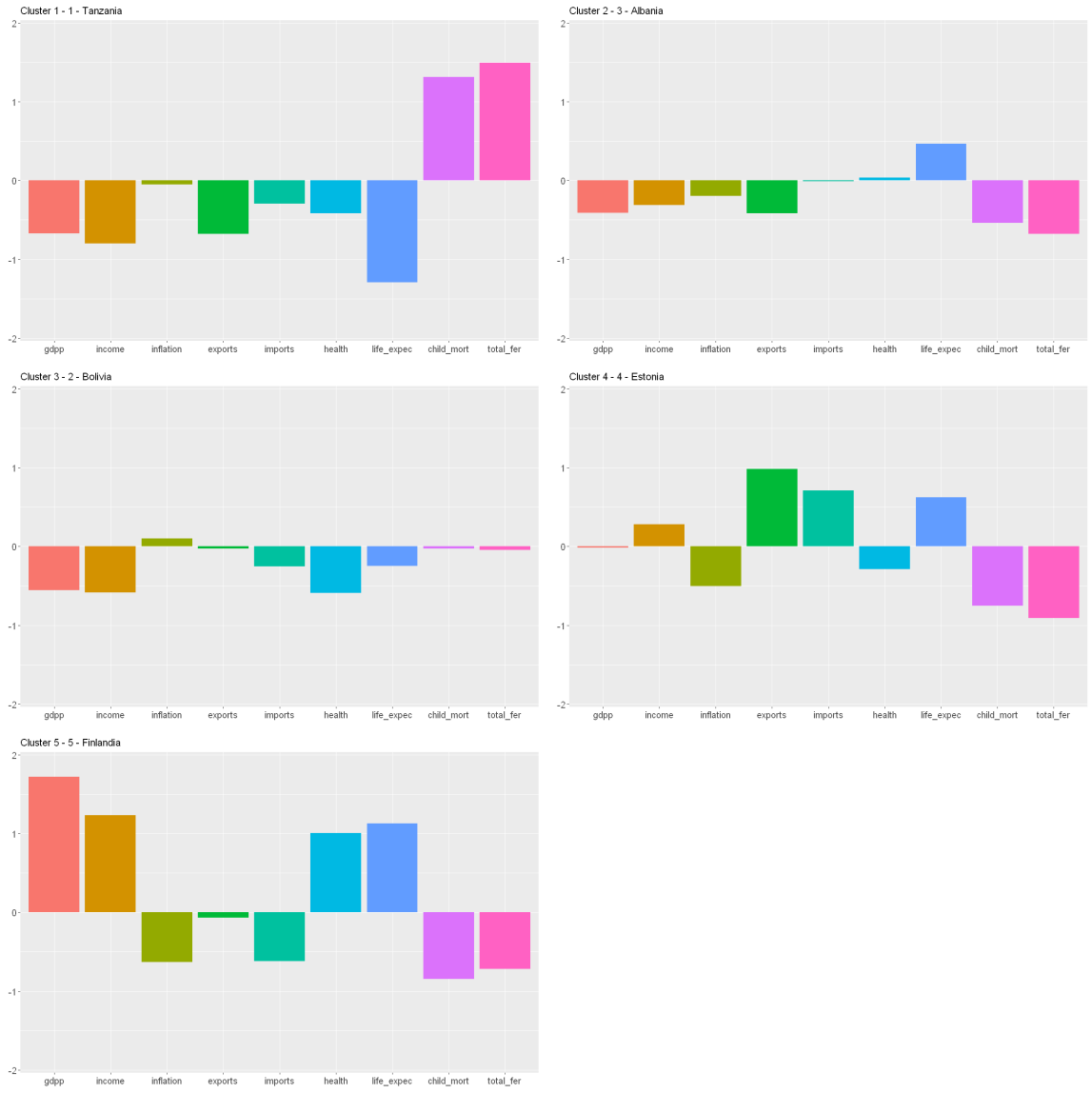

Figure 22: Distribution of variables for each cluster when k=5

	cluster	gdpp	income	inflation	exports	imports	health	life_expec	child_mort	total_fer
Tanzania	1	702	2090	9.250	18.7	29.1	6.01	59.3	71.9	5.43
Bolivia	3	1980	5410	8.780	41.2	34.3	4.84	71.6	46.6	3.20
Albania	2	4090	9930	4.490	28.0	48.6	6.55	76.3	16.6	1.65
Estonia	4	14600	22700	1.740	75.1	68.7	6.03	76.0	4.5	1.72
Finland	3	46200	39800	0.351	38.7	37.4	8.95	80.0	3.0	1.87

Figure 23: Values for the Medoids of the 3 clusters