Question 1:

* **Problem statement**

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries .

After the recent funding programs, the NGO have raised around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively.

The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

* Business Objectives:

Our Objective is to categorize the countries using some socio-economic and health factors that determine the overall development of the country.

Then suggest the countries which the CEO needs to focus on the most providing with basic amenities and relief during the time of disasters and natural calamities.

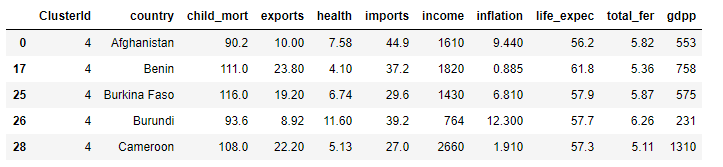
* **Analysis Approach**
* Step 1: Data Understanding (Understanding Problem statement, Columns provided and data dictionary )
* Step 2: Data Cleaning/Standardizing (Eliminating multi co-linearity on highly correlated data)
* Step 3: Using PCA Module(imported) to perform PCA on the data set. Then merging data with original dataset, following with outlier handling.
* Step 3: Setting the clusters for the dataset. Followed by performing Hopkins and analyzing the best cluster to use with good tendency
* Step 4: Accordingly applying Silhouette Analysis (for nearest cluster) and sum of squares of distance between them
* Step5 : At the end applying Hierarchical clustering on dataset representing through Dendogram.

Hence we conclude by this analysis that :

1. After clustering cluster 4 has the lowest frequency, meaning this cluster contains the countries in need.

* (Please note this clustering may change dynamically everytime we run the code.)

1. We have seen total of 23 country under this cluster who's socio economic growth is little, which needs help from NGO
2. Below is the countries list which needs the help:



Question 2:

* *Assume that, we have two-dimensional data (i.e., two features) and the joint distribution of the data follows multivariate normal distribution. One of the important properties of multivariate normal distribution is that, if the correlation between the features is zero, it means that features are orthogonal.* The main job of PCA is to represent the data in lower dimensional by removing the redundant features. It achieves that through finding **orthogonal**principal components. The above property is not applicable if the joint distribution of data (not individual distribution of feature) follows other distribution instead of multivariate normal distribution. We also use a covariance matrix (covariance matrix is a function of correlation matrix) to find the principal components. The only one distribution (zero-mean probability distribution) which allows us to represent the whole data in a compact form is Gaussian distribution. It shows that PCA make an implicit assumption that data should follows Gaussian distribution. If data didn’t follow Gaussian distribution, it would be difficult to extract independent statistical components by PCA.

Question 3: The most important difference is the hierarchy. Actually, there are two different approaches that fall under this name: top-down and bottom-up.

In top-down hierarchical clustering, we divide the data into 2 clusters (using k-means with k=2, for example). Then, for each cluster, we can repeat this process, until all the clusters are too small or too similar for further clustering to make sense, or until we reach a preset number of clusters.

In bottom-up hierarchical clustering, we start with each data item having its own cluster. We then look for the two items that are most similar, and combine them in a larger cluster. We keep repeating until all the clusters we have left are too dissimilar to be gathered together, or until we reach a preset number of clusters.

In k-means clustering, we try to identify the best way to divide the data into k sets simultaneously. A good approach is to take k items from the data set as initial cluster representatives, assign all items to the cluster whose representative is closest, and then calculate the cluster mean as the new representative, until it converges (all clusters stay the same).