Clustering Assignment

By : AKSHAY MAGOTRA

PROBLEM STATEMENT

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
After the recent funding program's, they have been able to raise 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.

OBJECTIVE

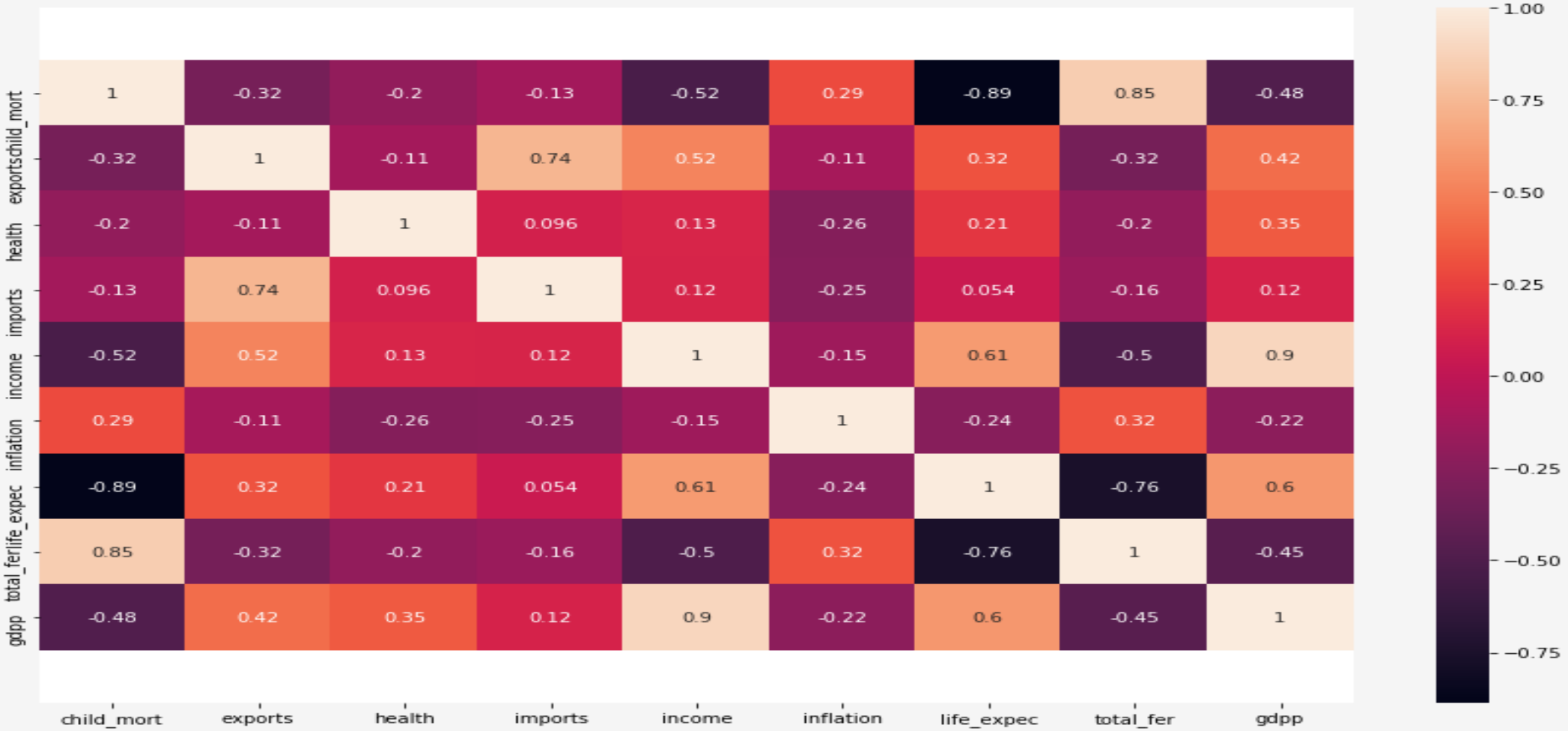
- To categories the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.

DATA PROCESSING

- ✓ It was found that there were no null values:
- ✓ There were also no duplicate values for country
- ✓ There were a few outliers and they were treated later on during PCA
- ✓ The data was standardized for Principal Component Analysis

Univariate Analysis

Heatmap to understand the attributes dependency

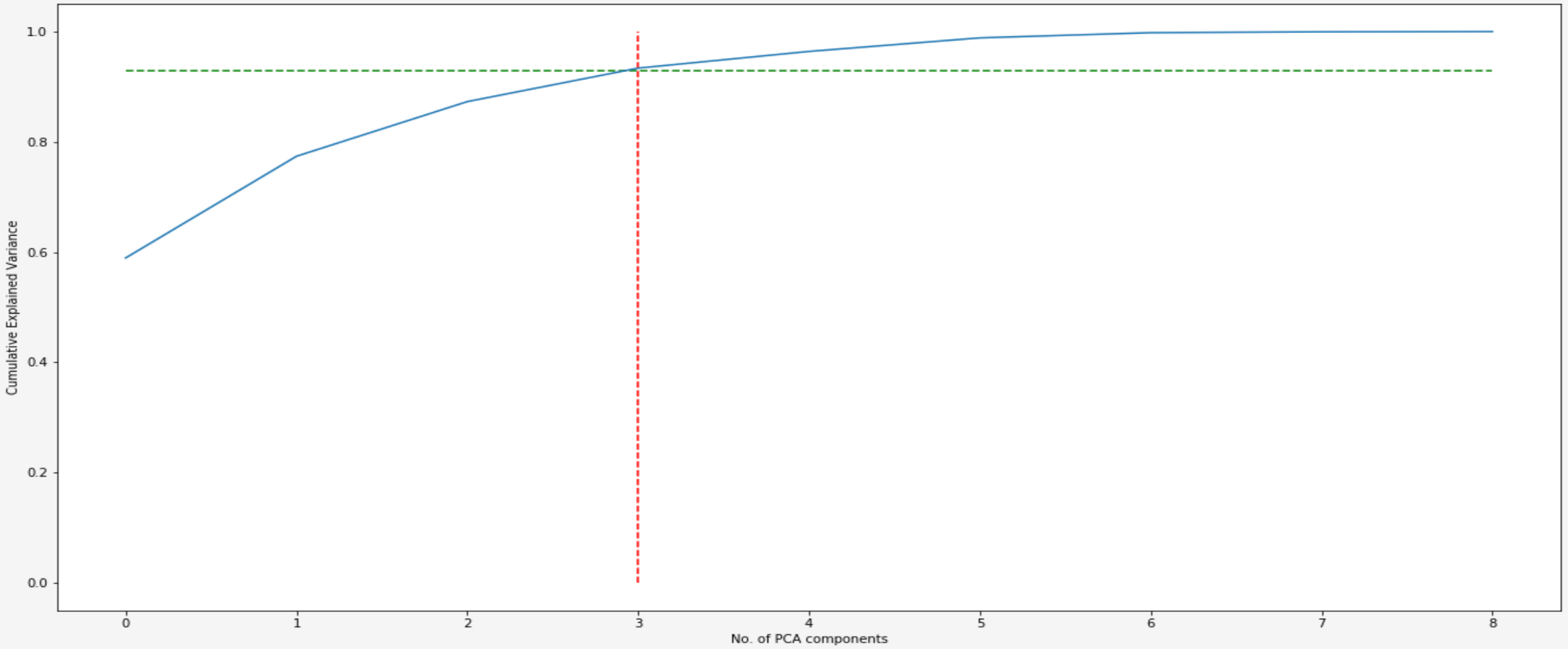


OUTCOMES

1.Child mortality and life_expentency are highly correlated with correlation of -0.89

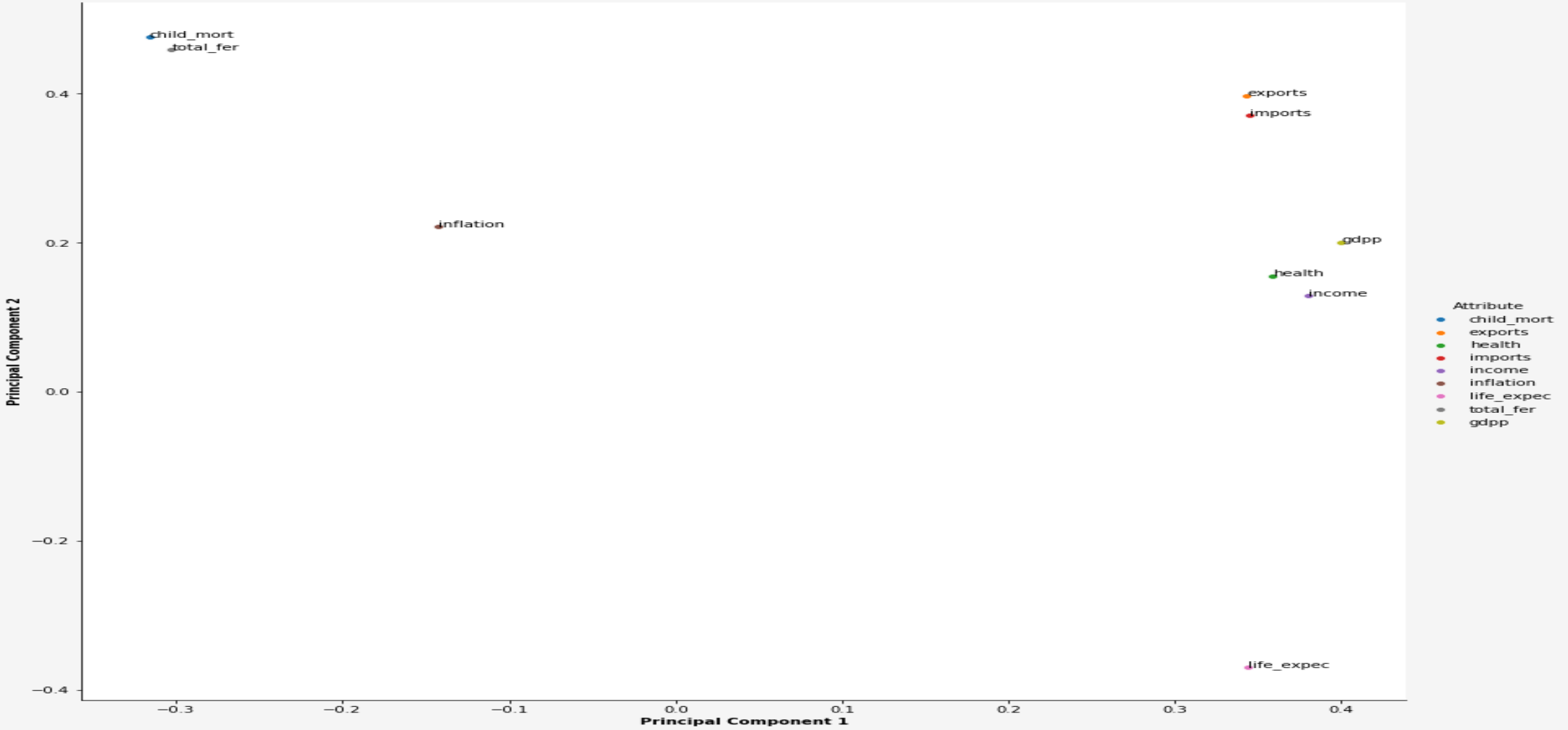
2.Child mortality and total_fertility are highly correlated with correlation of 0.85

Scree plot to visualize the Cumulative variance against the Number of components

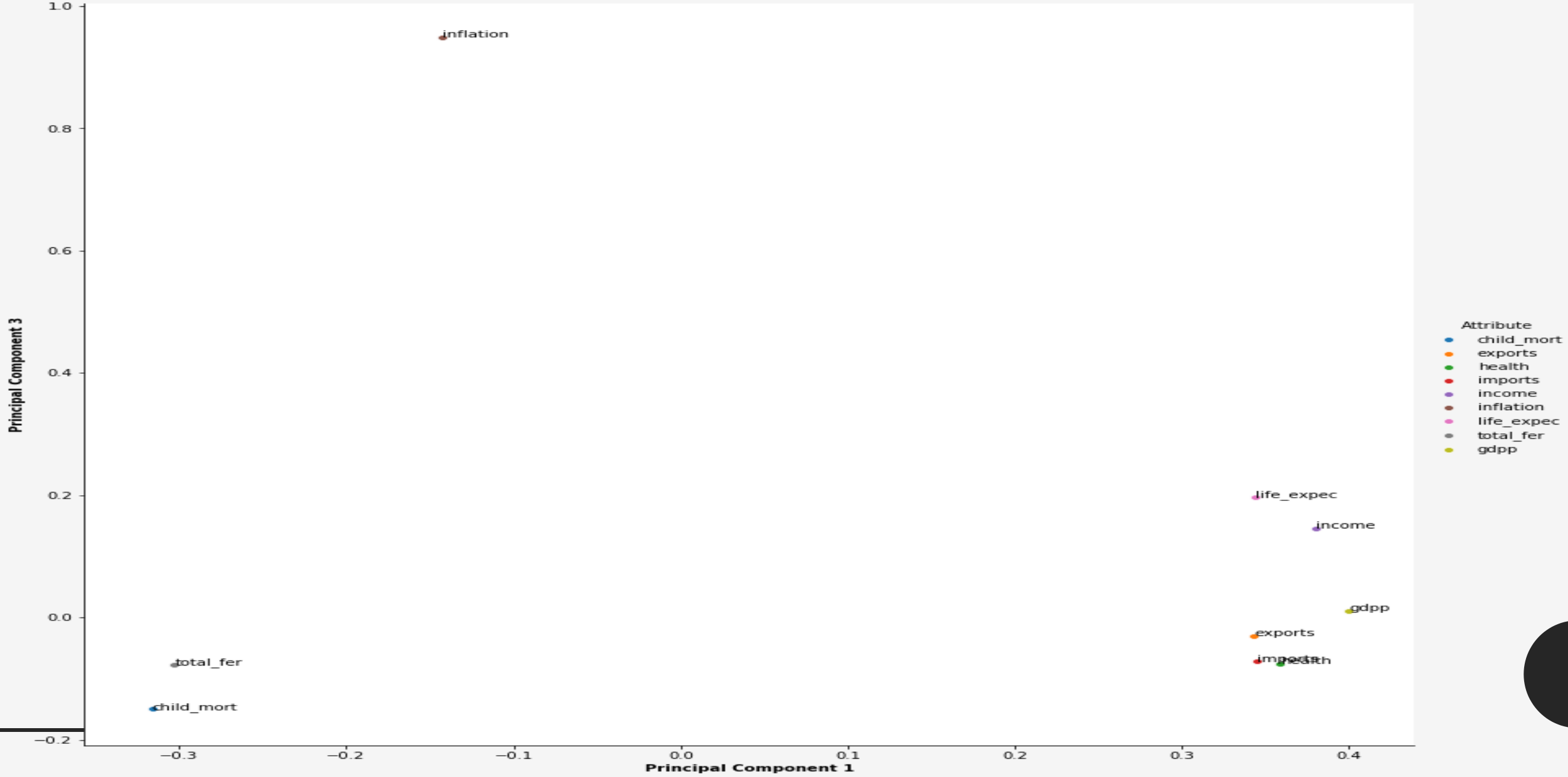


We can clearly see that from the above Scree plot that more than 90% variance is explained by the first 3 principal components. Thus, we will use these components only going forward for Clustering process.

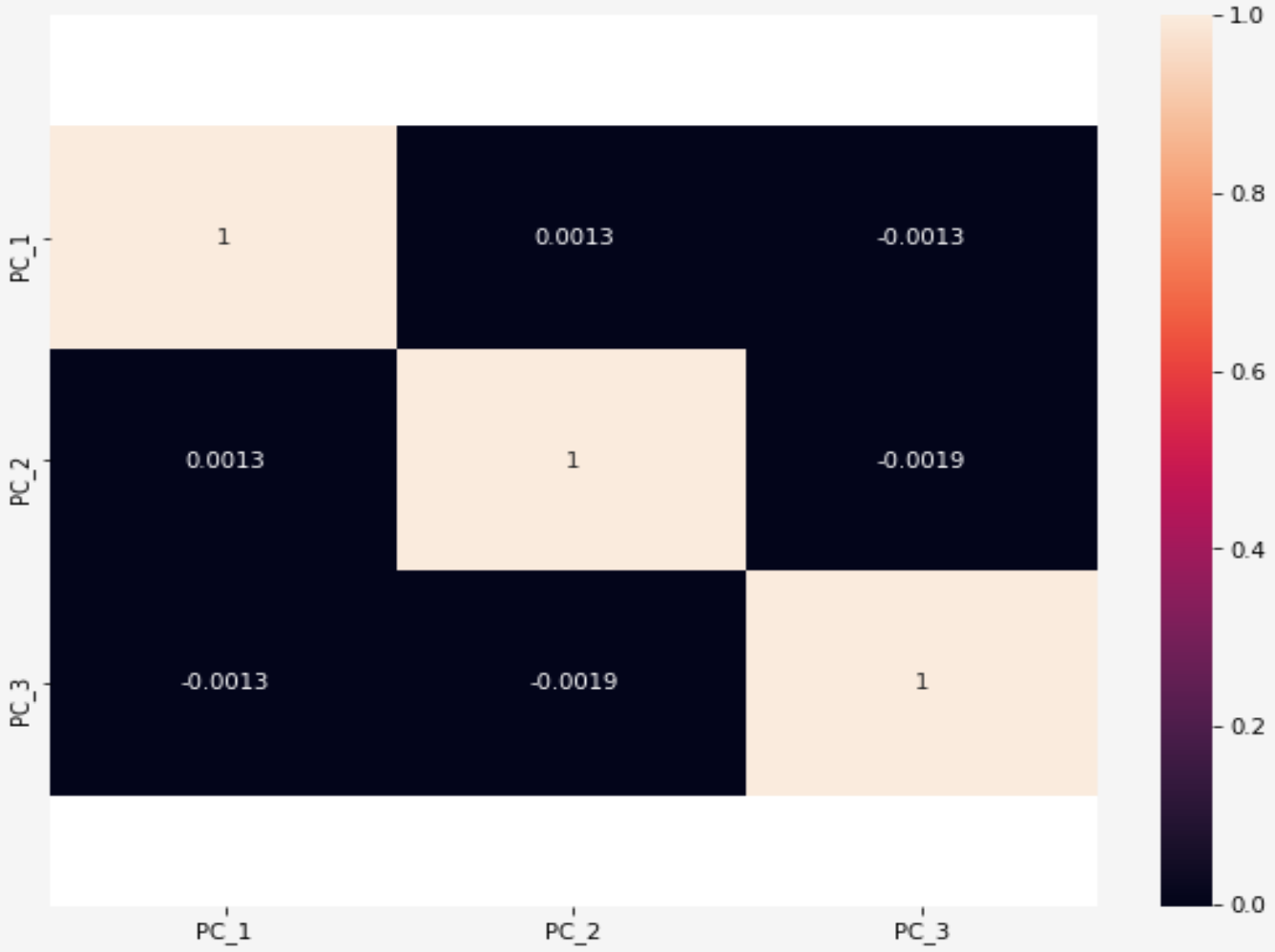
Plotting t visualization graph with PC1 and PC2



With PC1 and PC3



Plotting Heatmap to check the dependency in the dataset.



By the Heat-Map we can clearly see that the correlation among the attributes is almost 0, Thus we can proceed with this data-frame.

Hopkins Statistics:

- The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.
- If the value is between {0.01, ...,0.3}, the data is regularly spaced.
- If the value is around 0.5, it is random.
- If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.

We have 0.78 as a Hopkins Score . This is a good Hopkins score for Clustering.

K-means clustering :

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

The algorithm works as follows:

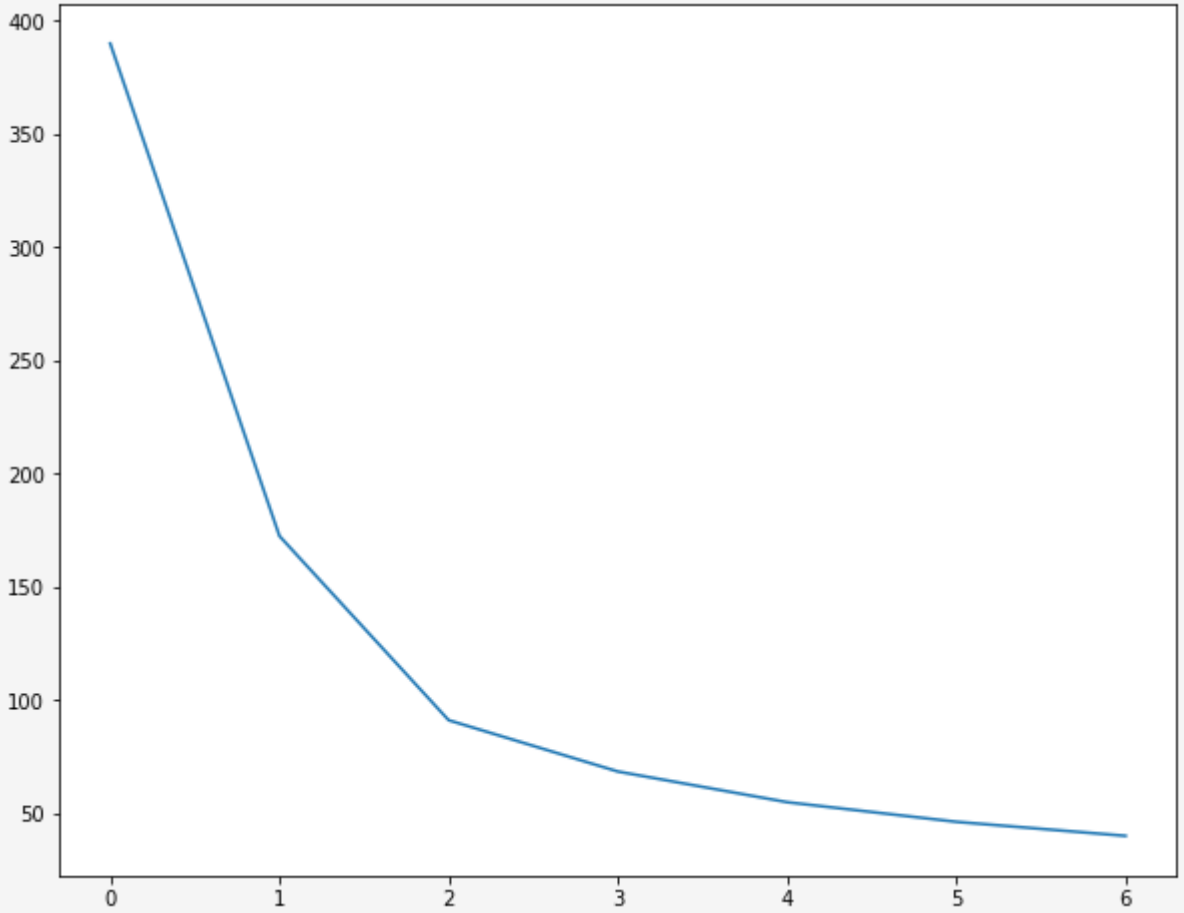
First we initialize k points, called means, randomly.

We categorize each item to its closest mean and we update the mean’s coordinates, which are the averages of the items categorized in that mean so far.

We repeat the process for a given number of iterations and at the end, we have our cluster

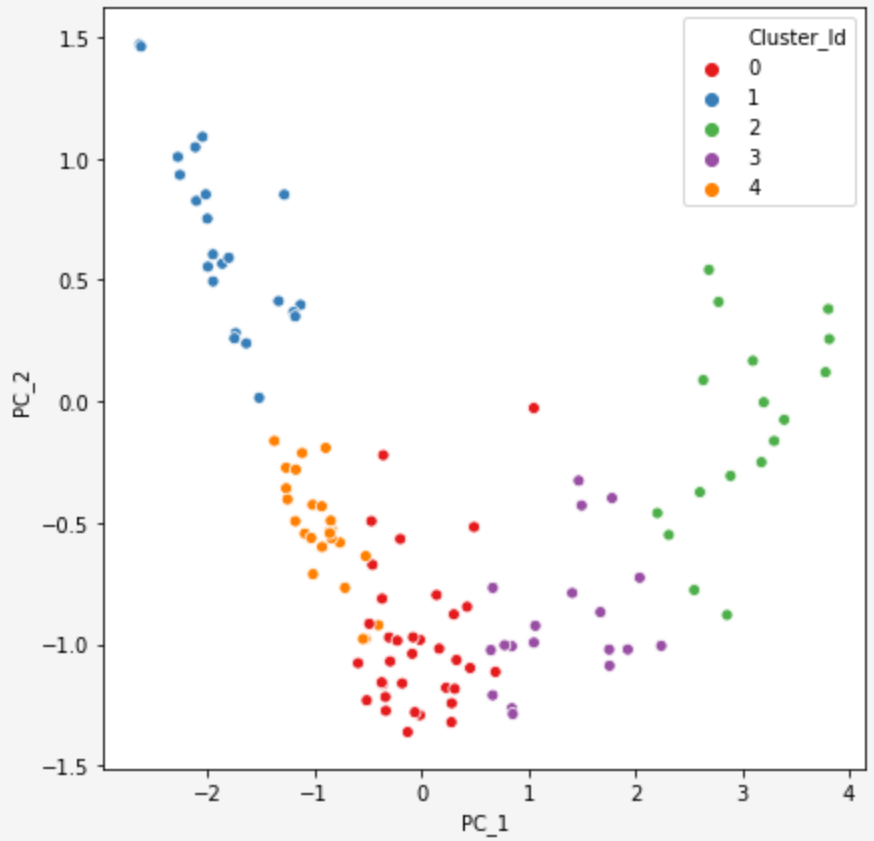
Use of Elbow Method

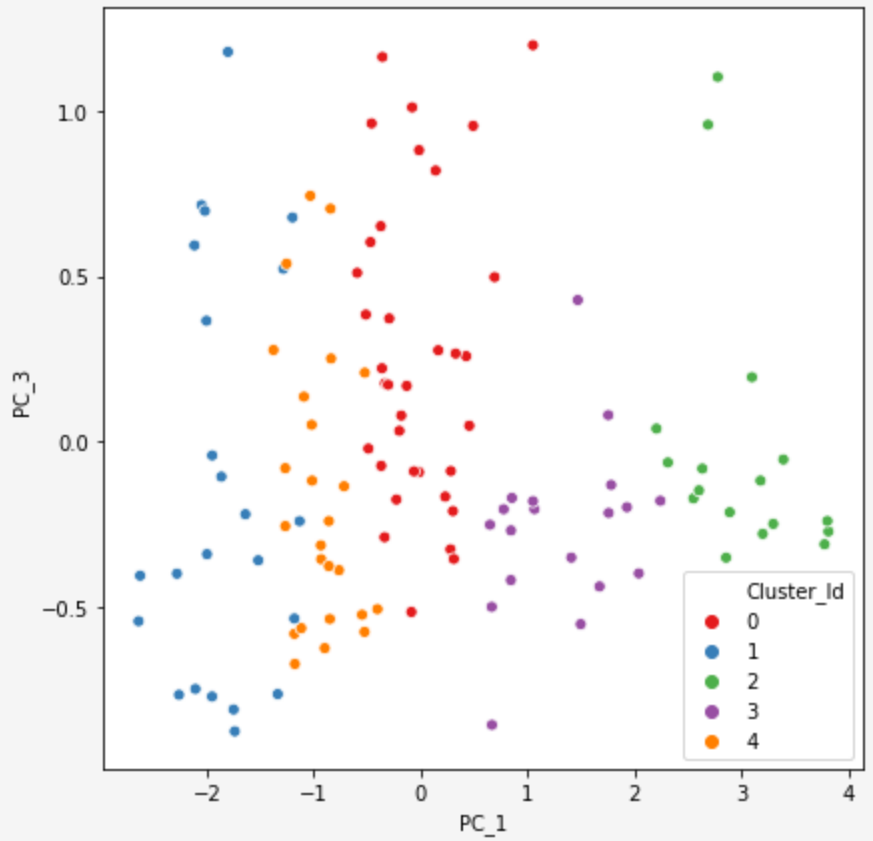
The Elbow Method is one of the most popular methods to determine this optimal value of k. We now demonstrate the given method using the K-Means clustering technique using the Sklearn library of python.



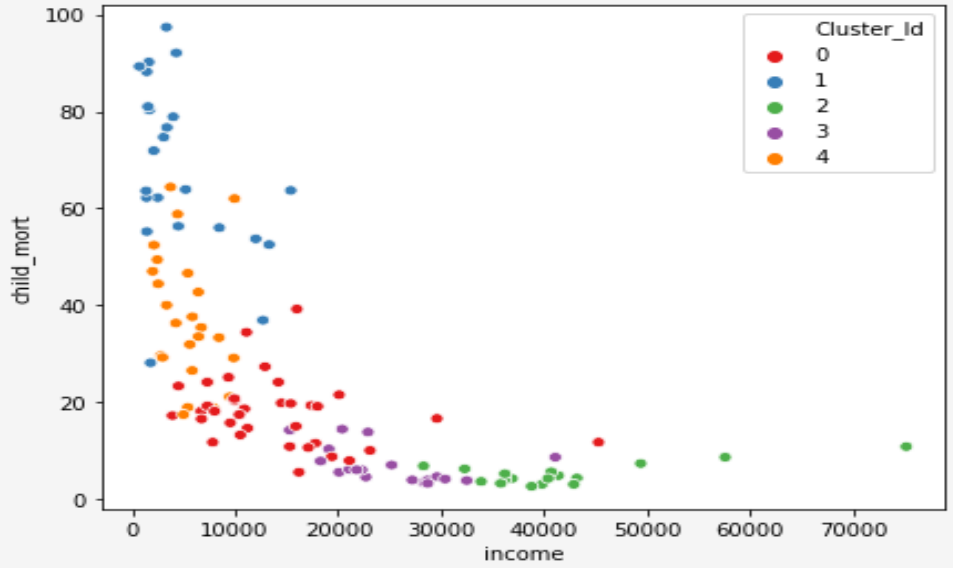
We can see that ,it looks good to proceed with either 4 or 5 clusters.

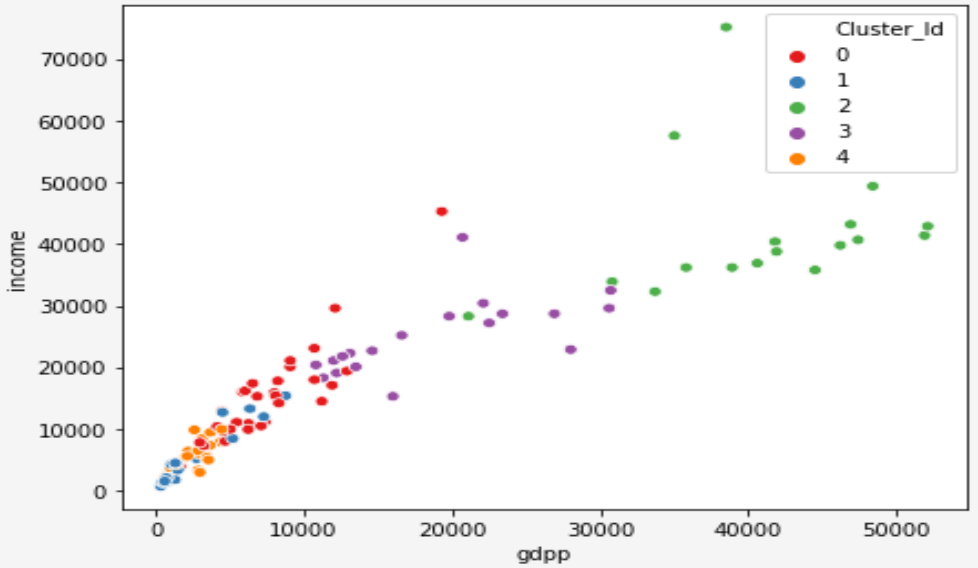
Scatter plot on Principal Components

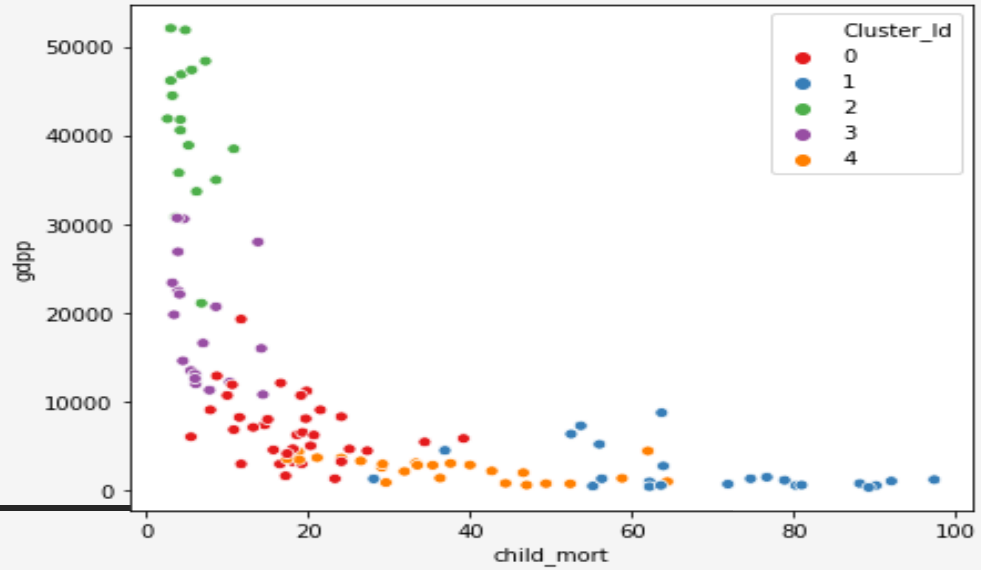


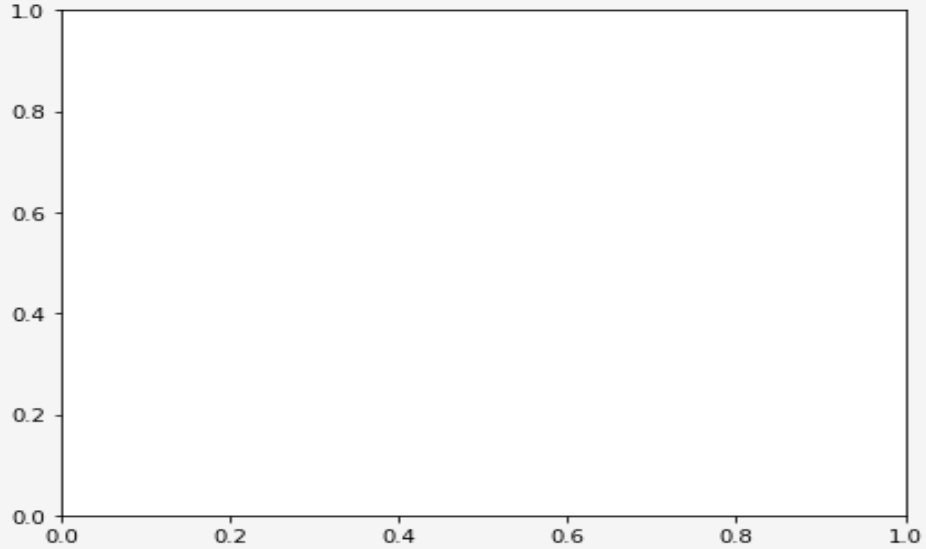


Scatter plot on Original attributes

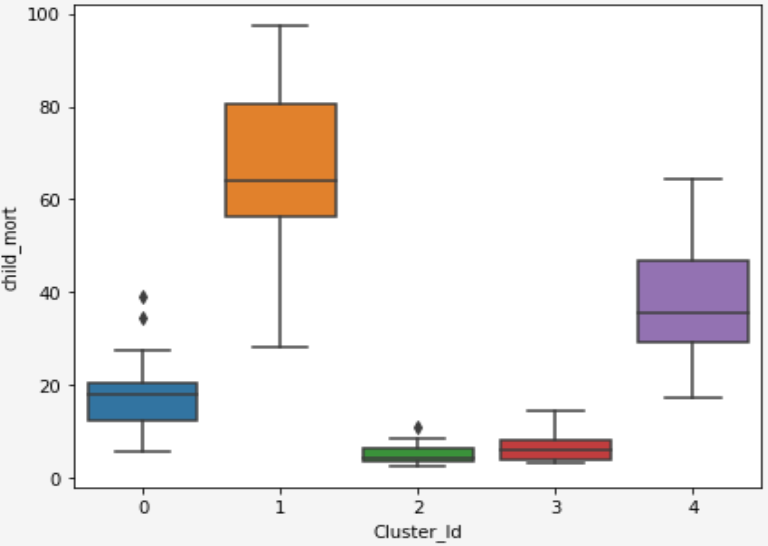


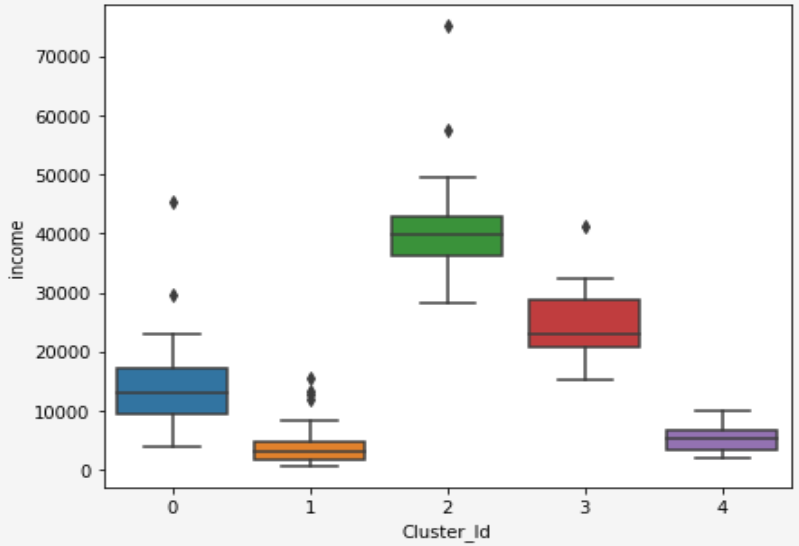


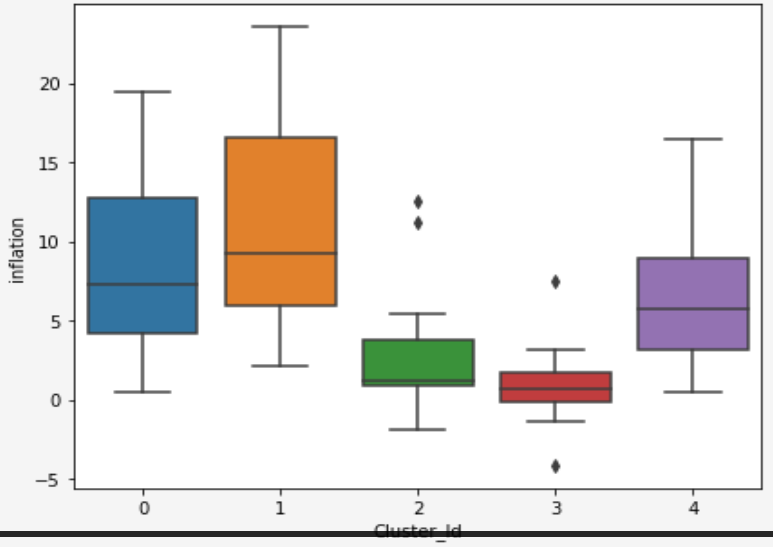


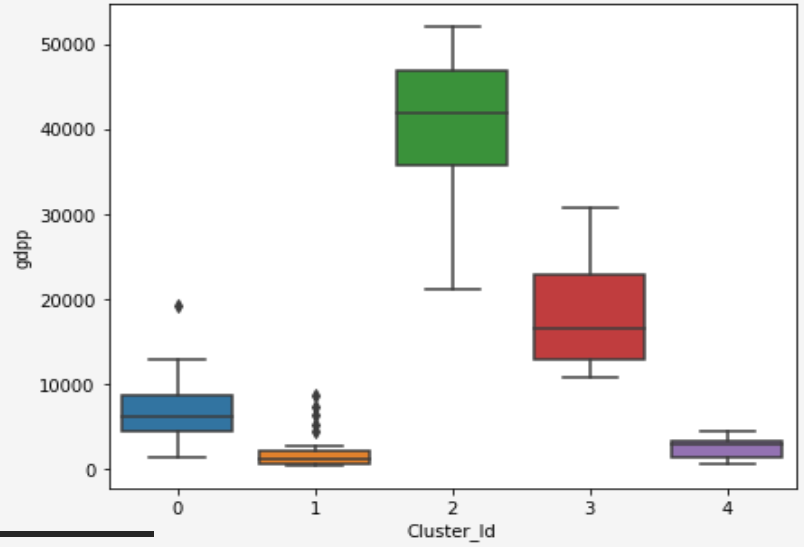


Box plot on Original attributes









Outcomes: We can see that

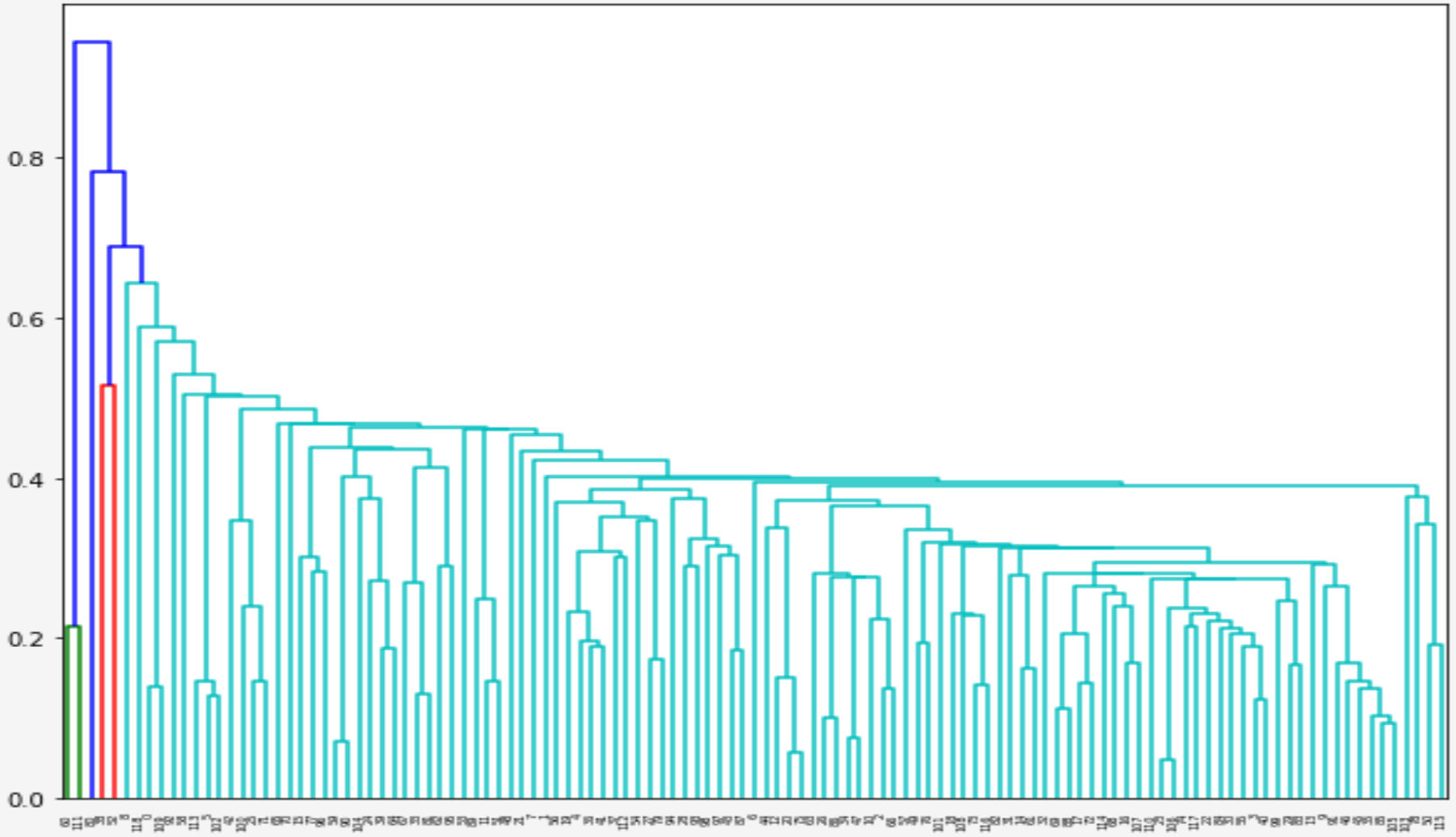
1.Child Mortality is highest for Cluster 0 and Cluster 3.These clusters need some aid.

2.Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development.

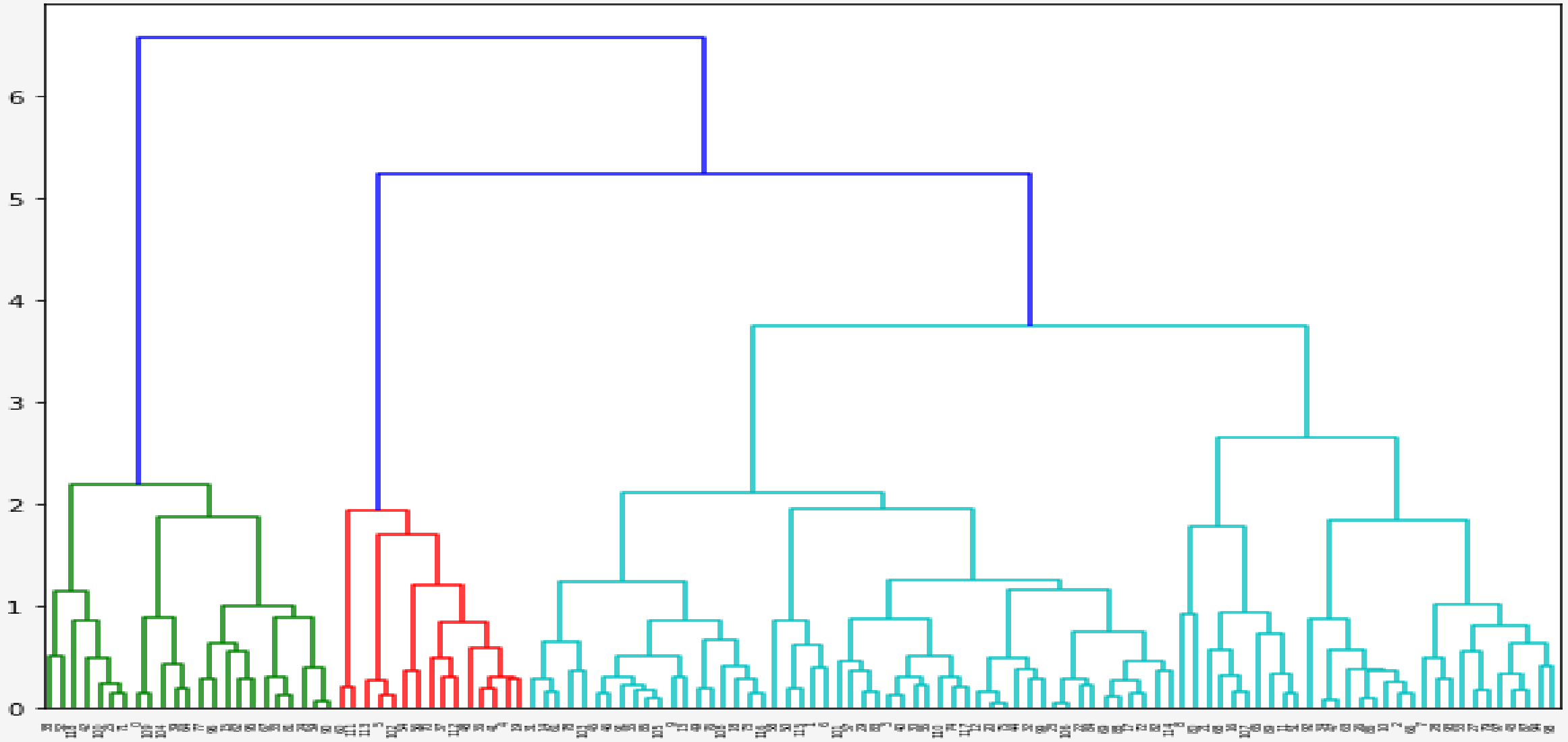
3.Income per capita and gdpp seems lowest for countries in clusters 0 and 3. Hence, these countries need some help.

Hierarchical Clustering

Single linkage



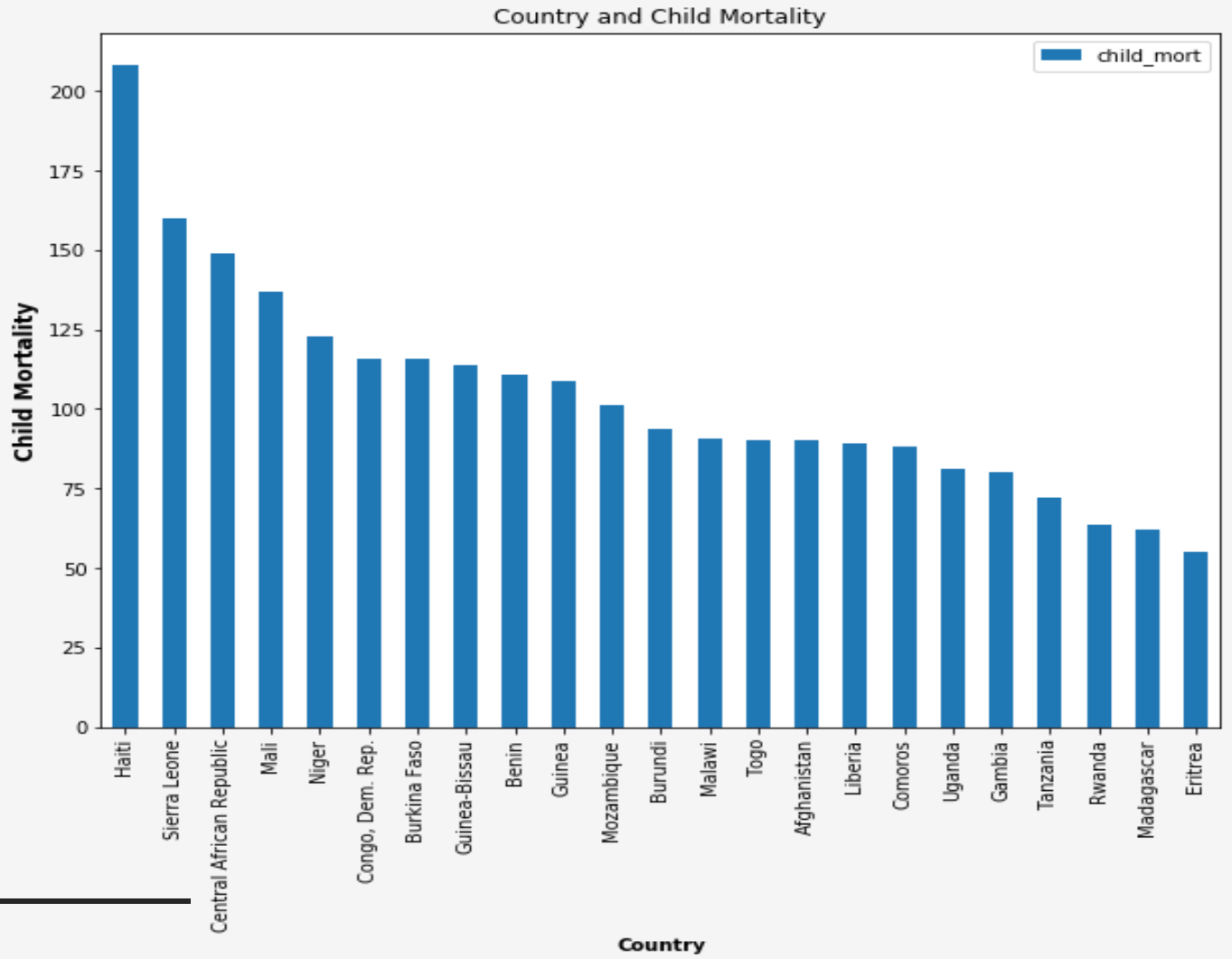
Complete Linkage



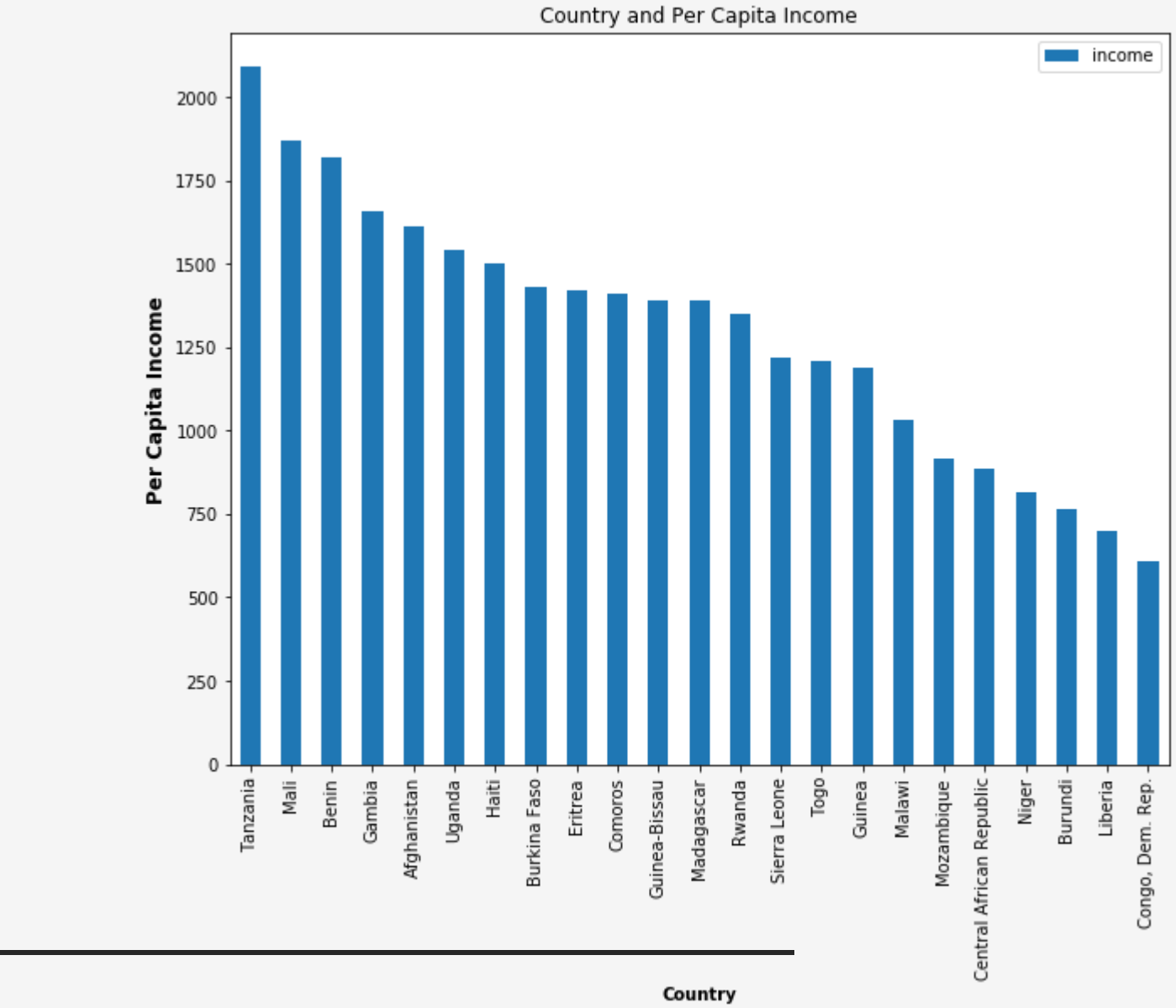
Bar Plot

For the countries which are in need of aid

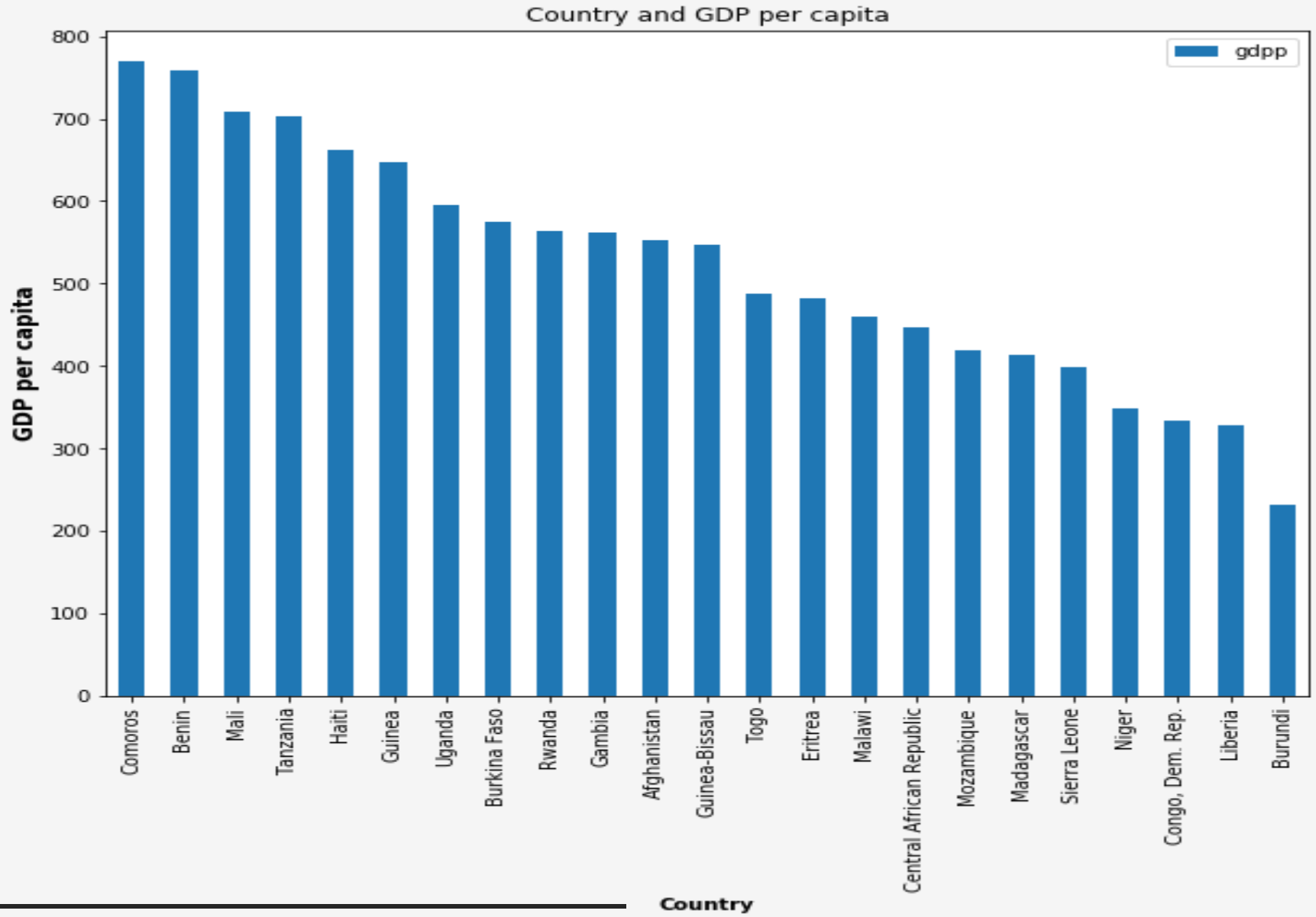
For Child Mortality



For Per Capita Income



For Per Capita GDP



Closing Statement

In this scenario :
1.We have used PCA above to reduce the variables involved and then done the clustering of countries based on those Principal components.
2.Then later we identified few factors like child mortality, income etc which plays a vital role in deciding the development status of the country and builded clusters of countries based on that.
3.Based on those clusters we have identified the below list of countries which are in dire need of aid. The list of countries are subject to change as it is based on the few factors like Number of components chosen, Number of Clusters chosen, Clustering method used etc.which we have used to build the model.

Final countries list's

Recommendations

Countries on which we require to focus more are :

{Afghanistan ,Benin ,Burkina Faso ,Burundi ,Central African Republic ,Comoros Congo-Dem. Rep ,Eritrea ,Gambia ,Guinea ,Guinea-Bissau ,Haiti,Liberia,Madagascar,Malawi,Mali,Mozambique,Niger,Rwanda,Sierra Leone ,Tanzania, Togo ,Uganda }