CLUSTERING & PCA ASSIGNMENT

SUBMISSION

Submitted By : Barkha Garg

Objective

Business Context:

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 programs, 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.

The purposes of this assignment activity are the following:

- Analyse the available datasets for the available KPIs and choose the appropriate ones using concepts of PCA
- Cluster the countries into appropriate clusters using the KMeans & Hierarchical Clustering. Using PCA,Your
- Categorise the countries using some socio-economic and health factors that determine the overall development of the country. Suggest the countries need to be focused the most and are in direst need of money.

Approach

Approach:

The analysis was carried out using the country data provided. The language used to analyse and visualize is Python and the scripting tool used is Jupyter Notebook.

The entire process was divided into sequence of steps (individual tasks) with output of tasks helping us infer some useful insights about the problem and also look at them visually for a better understanding. These steps included

a) Reading & Cleaning the Data

b) Applying the technique of PCA to reduce the dimensionality and pick only the features that explain the maximum variance in the data and use them for creating the clusters (using explained variance ratio)

c) Used Sihoutte Analysis & Elbow Curve methods to find an optimum k (number of clusters)

d) Perform Clustering using the Kmeans algorithm

e) Perform Clustering using the Hierarchical Clustering algorithm

At the end of these tasks/steps, some useful inferences were drawn and finally arrive at identifying the countries where the fund should be disbursed.

Data Cleaning

Data Loading & Data Cleaning

Step 1: Loading the countries data (country-data.csv)

a) The request data was loaded in Data Frames **request_data.**

b) The total count of records in country_data data frame was 167

Step 2: Cleaning the Data

a) There were no nulls found in the data for any of the fields. So all records were dropped/imputed due to nulls.

b) There were no duplicates in the data and so nothing was dropped.

c) All the attributes were in appropriate data types and so no explicit conversion was done

d) There were a few outliers in the dataset, but were not too significant. So these were not dropped either.

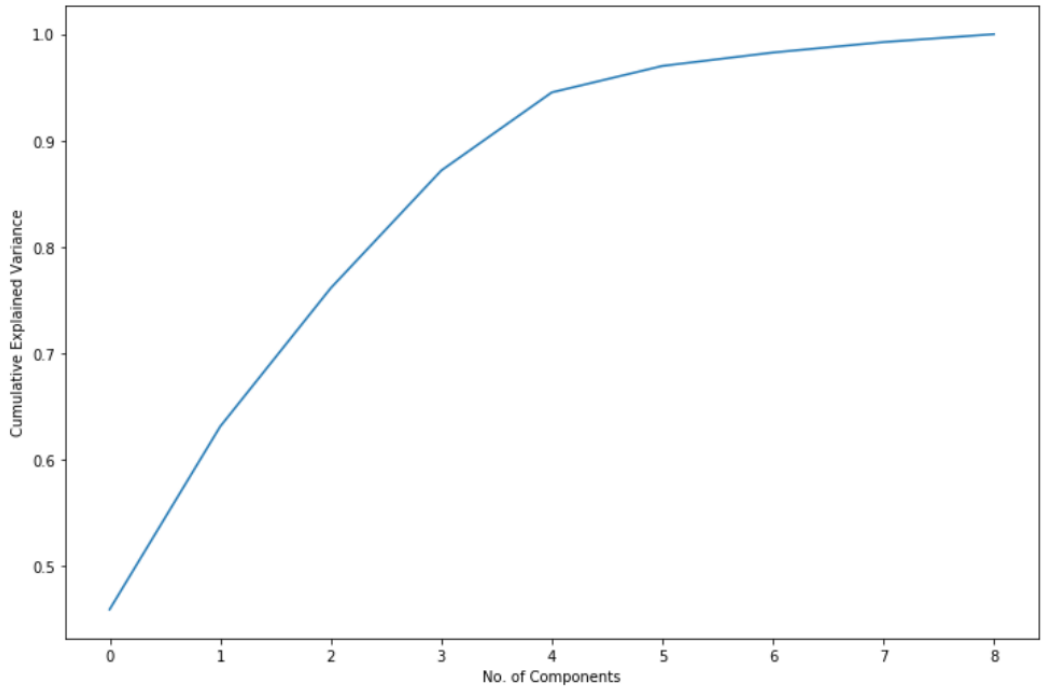
d) To proceed on with PCA, and to make sure the results are not biased due to the attributes values being on different scales, attribute values were scaled appropriately using Standard Scaler

PCA (Principal Component Analysis)

A detailed PCA was performed on the data

- There were 10 attributes in the initial dataset.
- Upon fitting them onto PCA, could see the following metrics of the explained variance ratio and the Scree Plot:

	Feature	Variance_%
0	child_mort	45.951740
1	exports	17.181626
2	health	13.004259
3	imports	11.053162
4	income	7.340211
5	inflation	2.484235
6	life_expec	1.260430
7	total_fer	0.981282
8	gdpp	0.743056



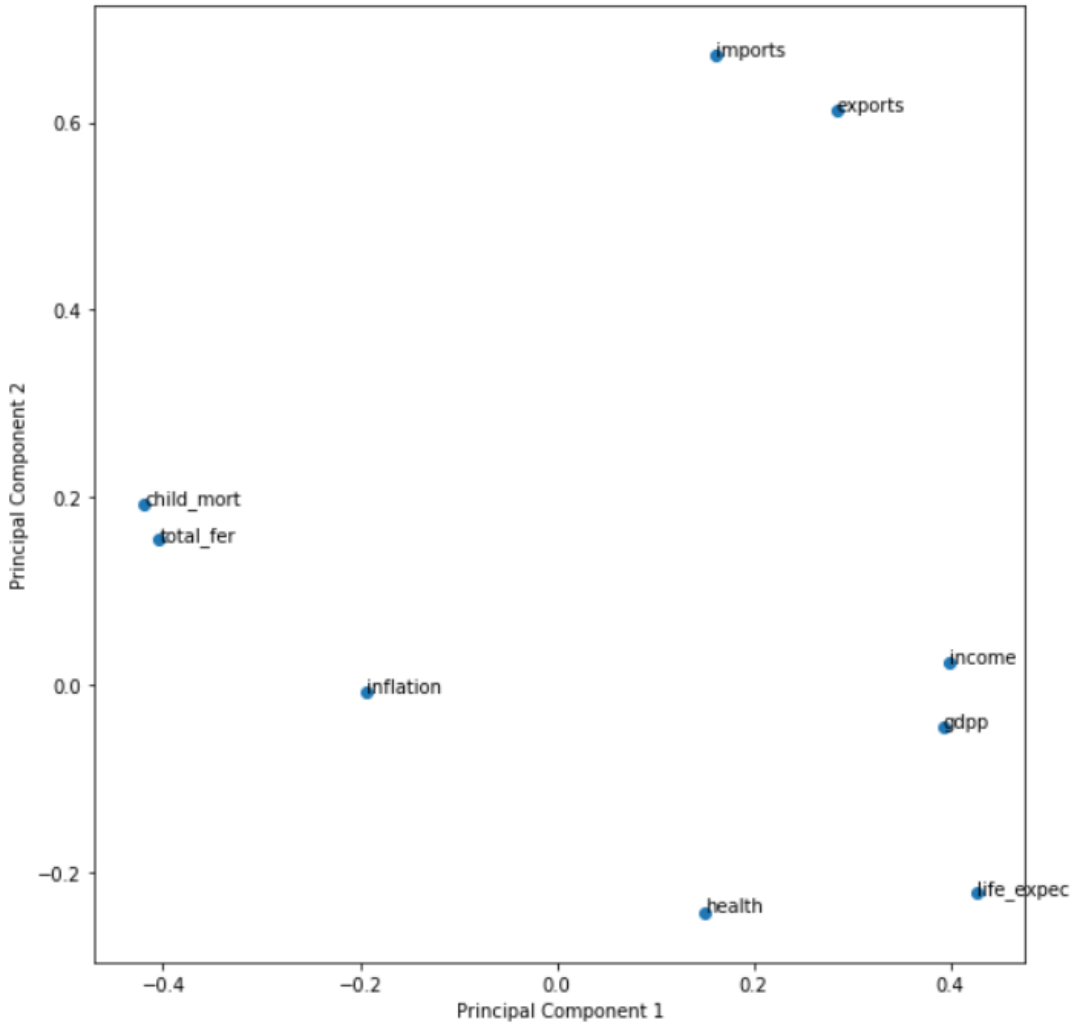
PCA (Cntd..)

- As seen from the Scree Plot, 5 variables appeared sufficient to go ahead with further analysis.
- These 5 could explain about 94.53 (~95%) of the total variance in data

	PC1	PC2	PC3	PC4	PC5
0	-2.913000	0.091969	-0.721242	1.001838	-0.146765
1	0.429870	-0.589373	-0.328611	-1.165014	0.153205
2	-0.285289	-0.452139	1.232051	-0.857767	0.191227
3	-2.932714	1.698771	1.525076	0.855595	-0.214778
4	1.033371	0.133853	-0.216699	-0.846638	-0.193186

PCA (Cntd..)

- Next, the driving attributes influencing the clusters were identified.
- As shown in the chart beside, variables child_mort, total_fer, income, gdpp, Life_expec have the highest magnitude along the first Principal Component. Henc
- These would be the once whose values we would be considering in picking the final values from the clusters that we would generate



Hopkins Statistics

- Before going ahead with performing Clustering on the PCA, did a check on the Hopkins score that gives a measure of how well can a given dataset be clustered (a score >0.5 is desirable)
- Obtained Hopkins score was over 0.5 and hence the clustering on this dataset can be performed.

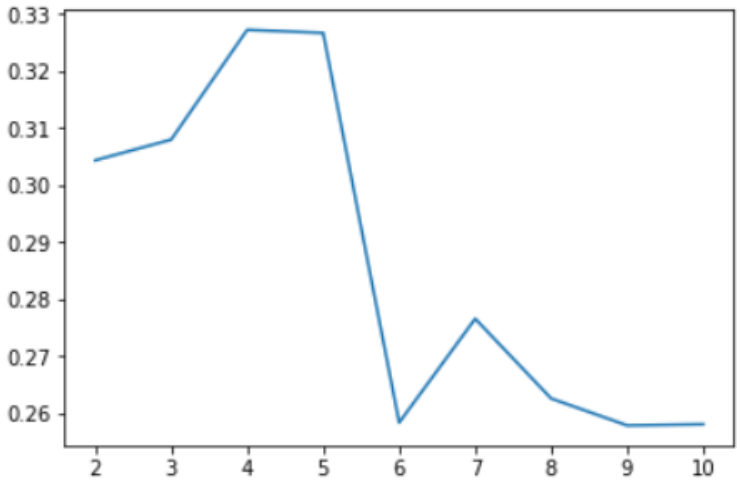
```
In [840]:  1  hopkins(df_train_pca)

Out[840]: 0.8639415662061457
```

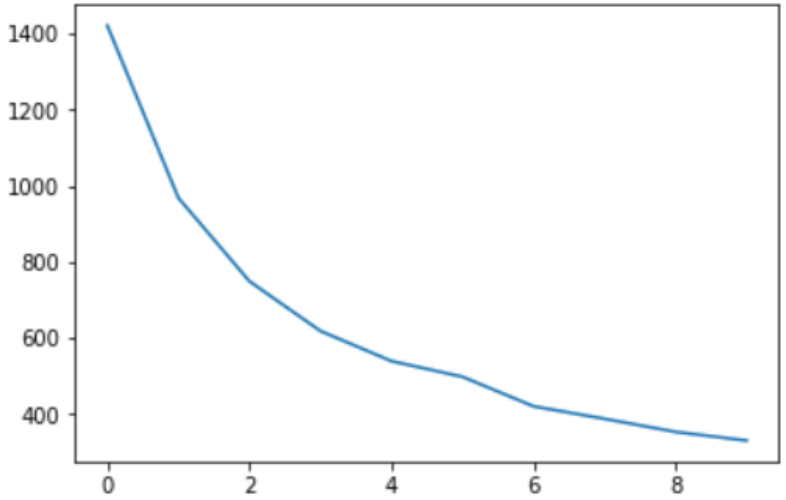
KMeans - Clustering

Finding the Optimal K (number of clusters)

Using the Silhoutte score and Elbow curve methods, observed the below trend and came to a conclusion that the data can be clustered in 4 distinct clusters:



Sihouette Score

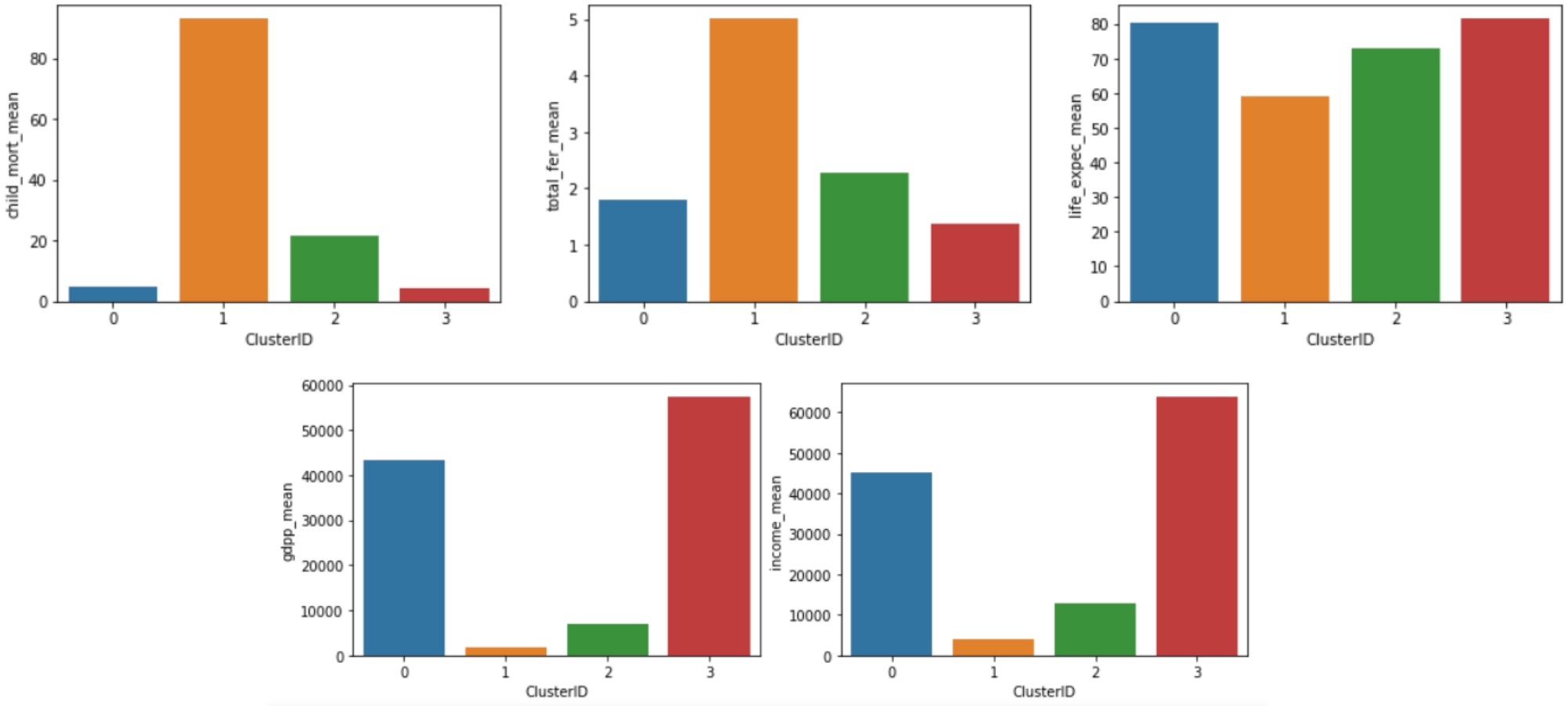


Elbow Curve

KMeans – Clustering (Cntd)

Finding the Optimal K (number of clusters)

Using the Silhoutte score and Elbow curve methods, observed the below trend and came to a conclusion that the data can be clustered in 4 distinct clusters:



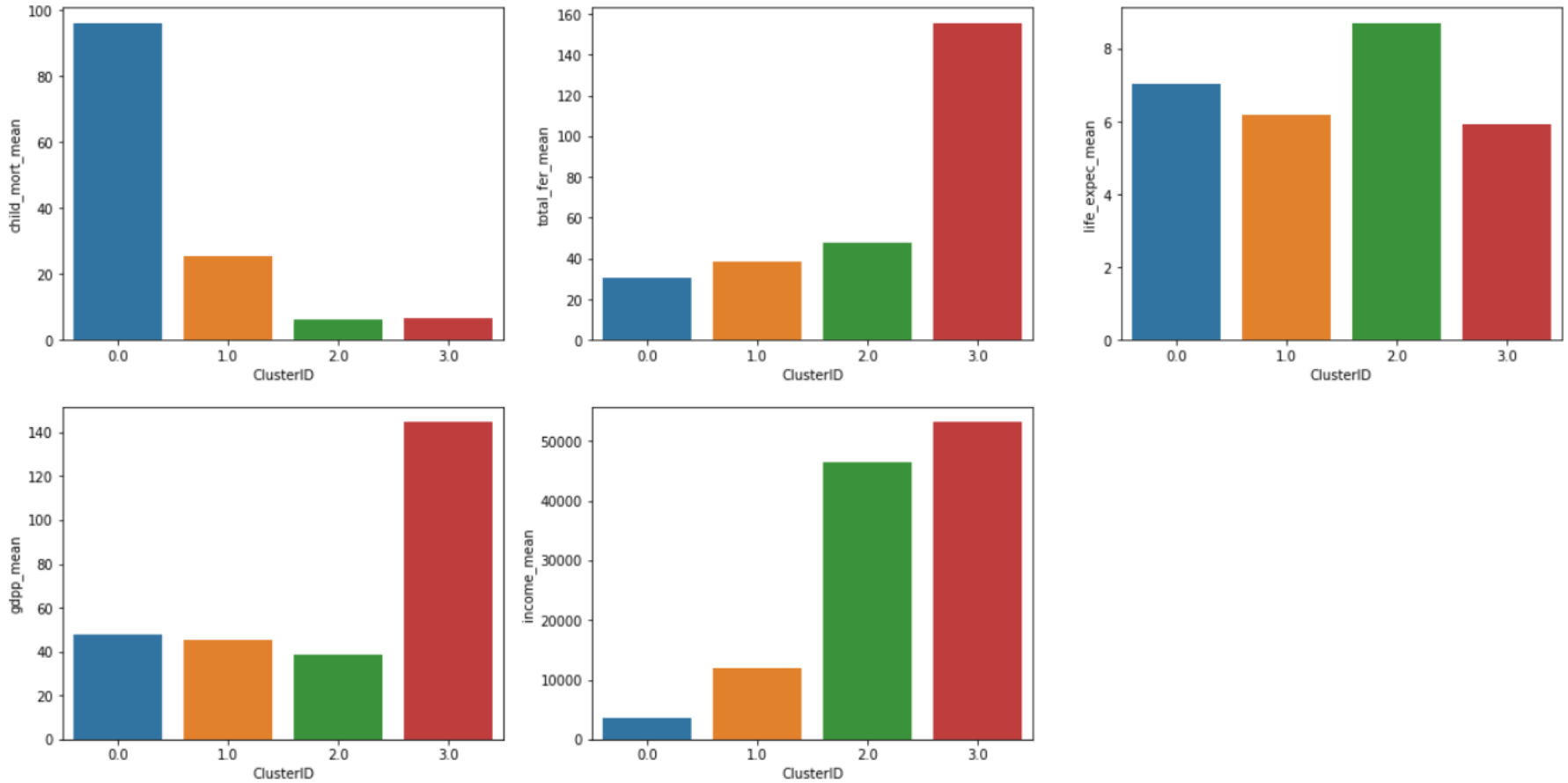
KMeans – Clustering (Cntd)

The Clusters thus formed are:

- Cluster 1: These are the countries with a very low child mortality, high life expectancy, high gdp & high income groups. So these countries seem to be doing good on their own.
- Cluster 2: These are the countries with a very high rate of child mortality with a very low gdpp & low income. These countries actually need immediate attention
- Cluster 3: This group seems to be doing OK having a relatively low value of child mortality, high income and gdp. So these are good on their own too
- Cluster 4: This appears to represent some really developed countries having good income, gdp and the least child mortality. These ones definitely are not striving for additional aid.

Hierarchical Clustering

Similar Clustering activity was performed using the Hierarchical Technique on the PCA dataset and creating 4 clusters.



Hierarchical Clustering (Cntd)

The Clusters formed in Hierarchical Clustering are:

- Cluster 1: These are the countries with a very high child mortality, high life expectancy, average gdp & a very low income groups. These countries do need some aid to develop
- Cluster 2: These are the countries with a relatively low child mortality, average gdp and lower than average income. These countries still are not in as much need of immediate aid as the previous group of countries
- Cluster 3: This group seems to be doing good having decent values in all the KPIS. There are almost developed nations with a good per capita income
- Cluster 4: This appears to represent some really developed countries having good income, gdp and the least child mortality. These ones definitely are not striving for additional aid.

Conclusions

- As per the clusters obtained from the K Means and Hierachical clustering methods, we see that countries in Cluster 2 as per K Means and countries in Cluster 4 as per Hierarchical Clustering are the ones that are struggling the most and hence need the funds.
- HELP International should consider disbursing funds to them. Few such countries are as follows:

As per K – Means:

Burkina Faso	Burundi	Cameroon	Central African Republic
Chad	Congo, Dem. Rep.	Cote d'Ivoire	Guinea
Guinea-Bissau	Haiti	Lesotho	Mali
Mauritania	Mozambique	Niger	Sierra Leone

As per Hierarchical:

Central African Republic	Chad	Haiti
Mali	Sierra Leone	