Clustering & PCA Assignment

Nithya V

Abstract

Business Understanding :

To suggest HELP International an international humanitarian NGO, to decide how to use recent funding money strategically and effectively to fight poverty.

Business Objectives :

To categorise the countries using some socio-economic and health factors that determine the overall development of the country.

Then suggest the CEO of HELP International identify which countries needs most focus.

Approach

Cluster the countries using some socio-economic and health factors that determine the overall development of the country.

Perform PCA on the data.

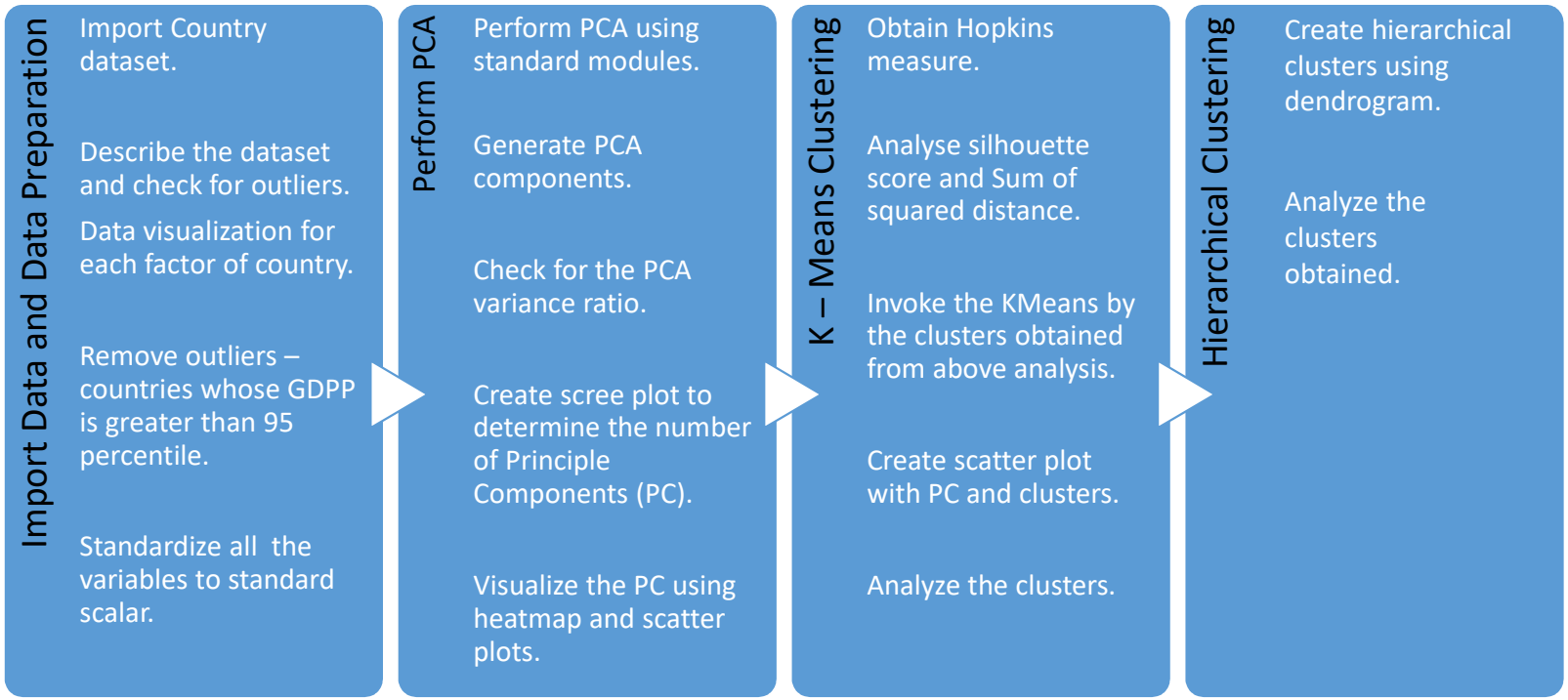
After PCA, continue both K-means and Hierarchical clustering and create clusters.

Analyze the clusters and identify the countries which are in dire need of aid.

Create visualizations on the clusters that have been formed by plotting Principal Components and the clusters.

Summary of the PCA and clustering – to provide top 5 countries which are in direst need of aid.

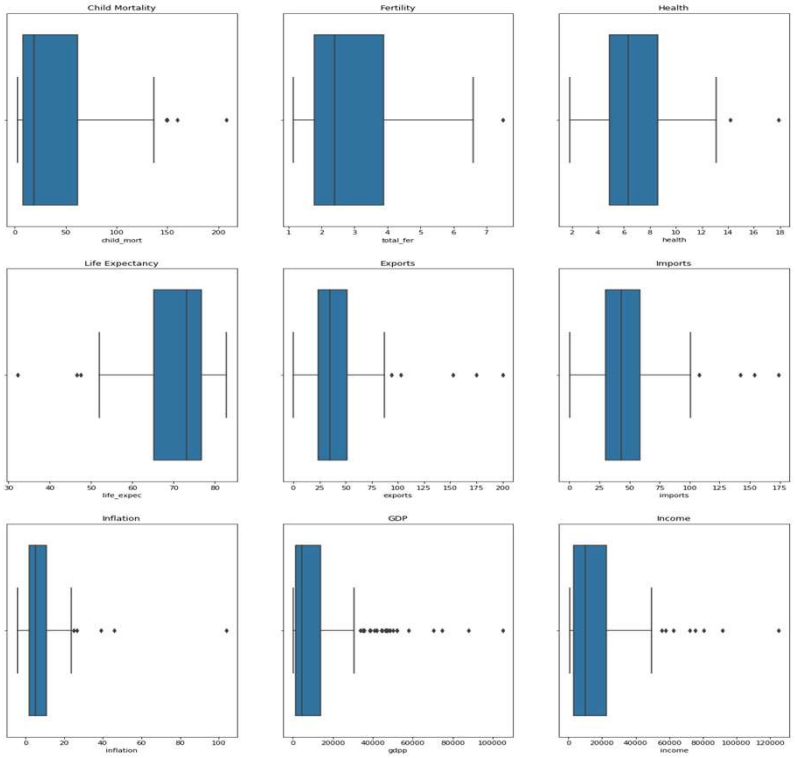
Detailed Approach



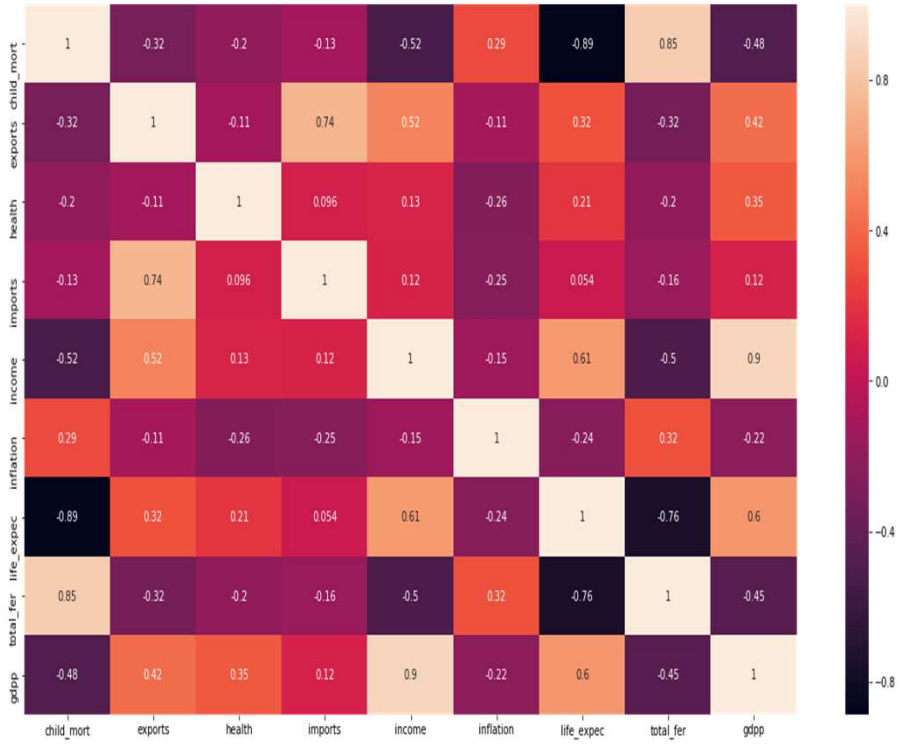
Country Data

Data Visualization of the country factors

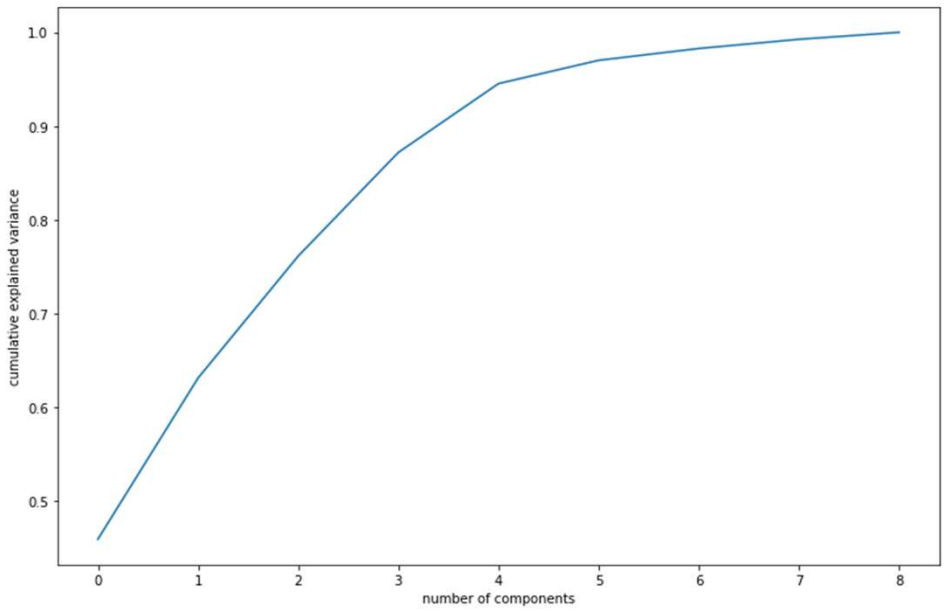
Boxplots

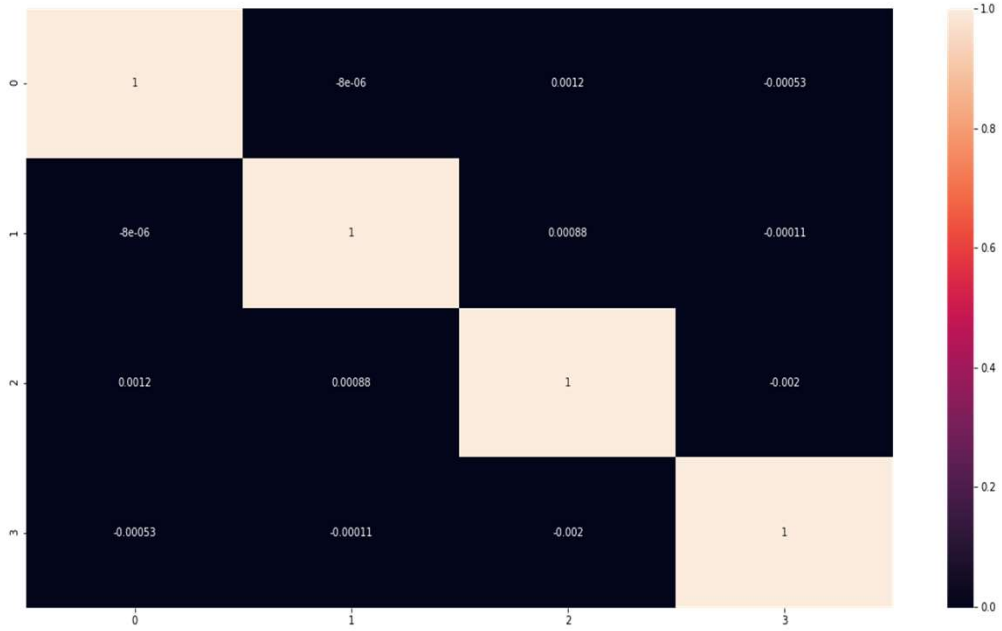


Heatmap



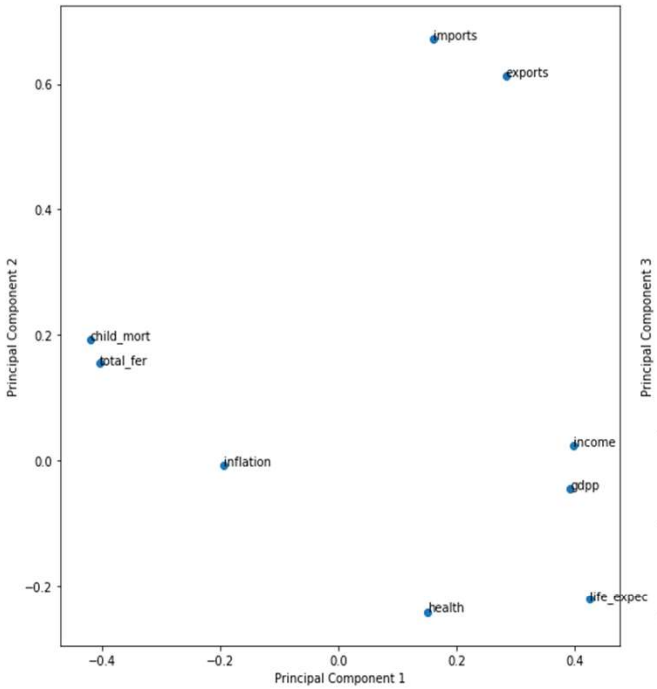
PCA on Country Dataset - 1

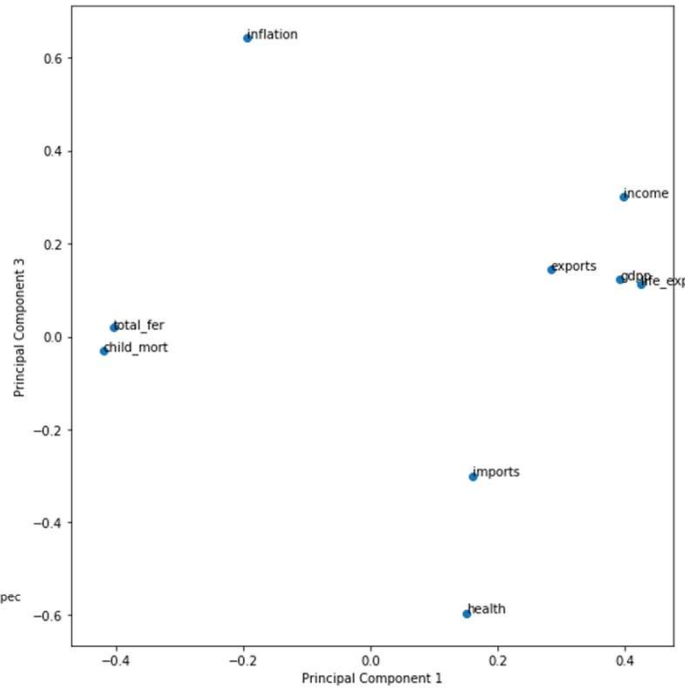


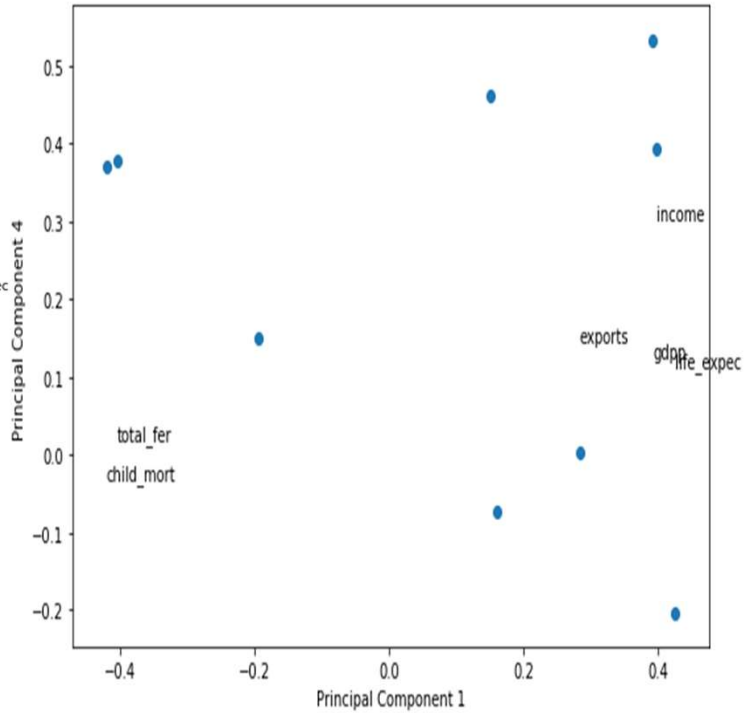


- The scree plot - plotting the cumulative variance against the number of principal components.
- Decide on the count of principal components which would explain around 95% of the total variance in the data.
- The heat map of all the principal components shows that correlations are almost zero.

PCA on Country Dataset - 2

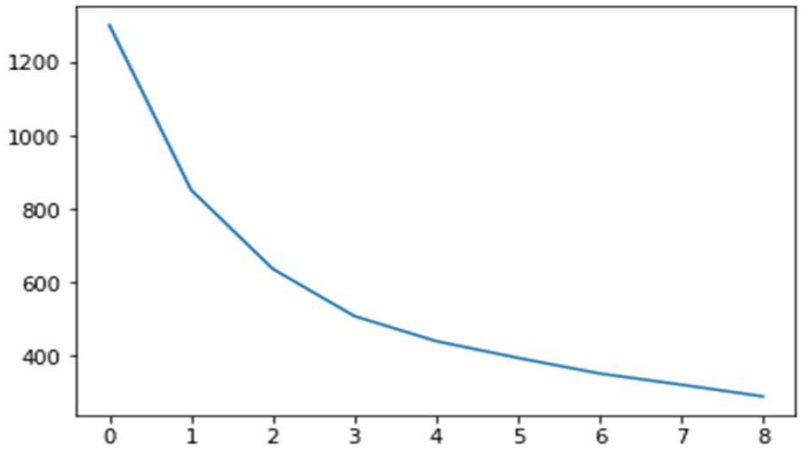




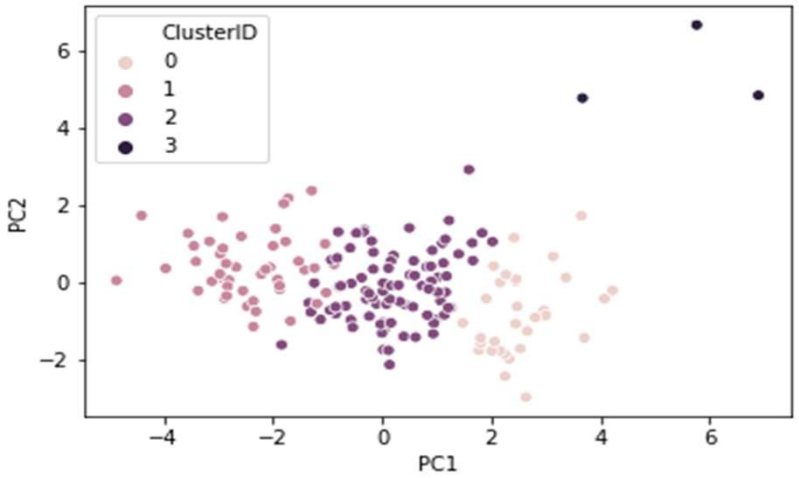


Scatter plots of the principle component one against all the other principal components.

K – Means Clustering - 1

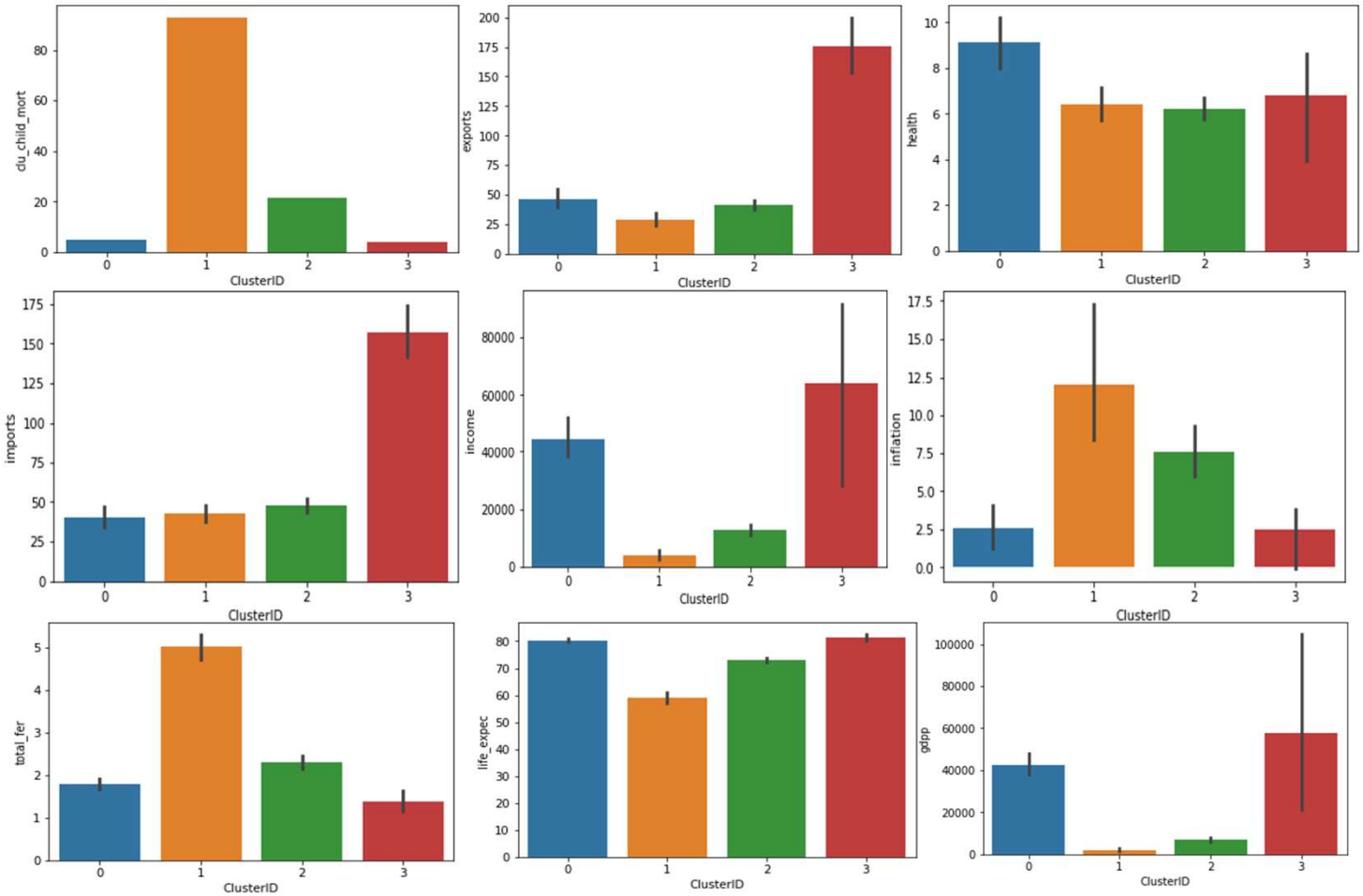


The scree plot with which we decide on the number of Principal components.



The scatter plot of PC1 and PC2 with all the clusters generated.

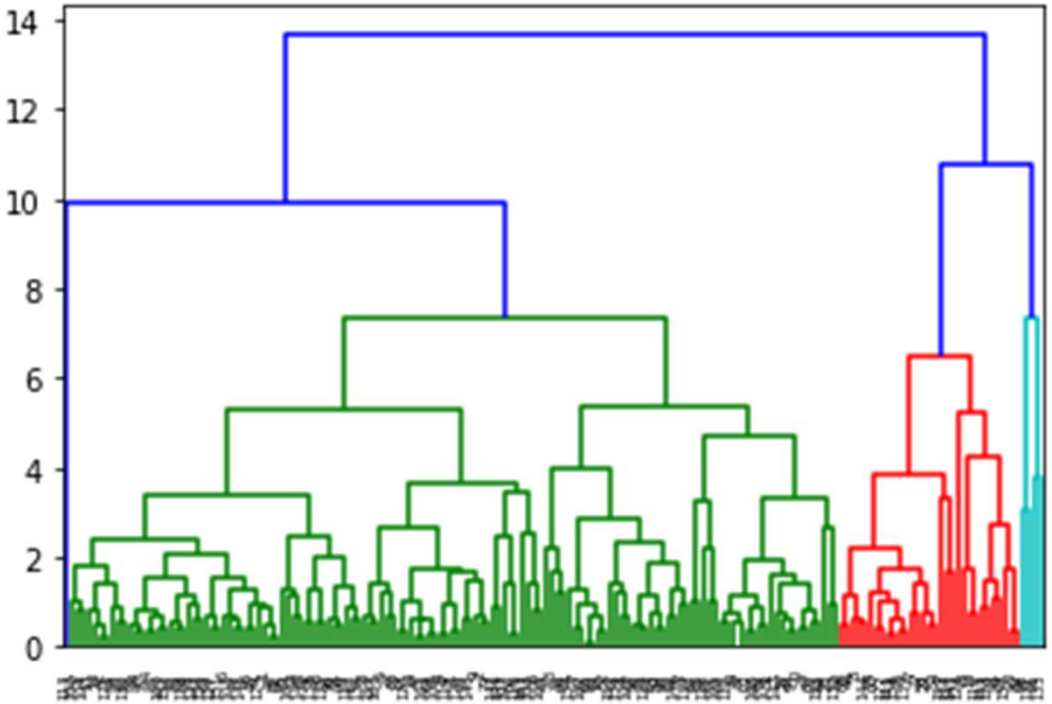
K – Means Clustering - 2



From the clusters obtained we observe that cluster 2 countries need more funding from HELP International.

Cluster 1 : Low Funding
Cluster 2 : High Funding
Cluster 3 : Moderate Funding
Cluster 4 : No Funding needed

Hierarchical clustering - 1

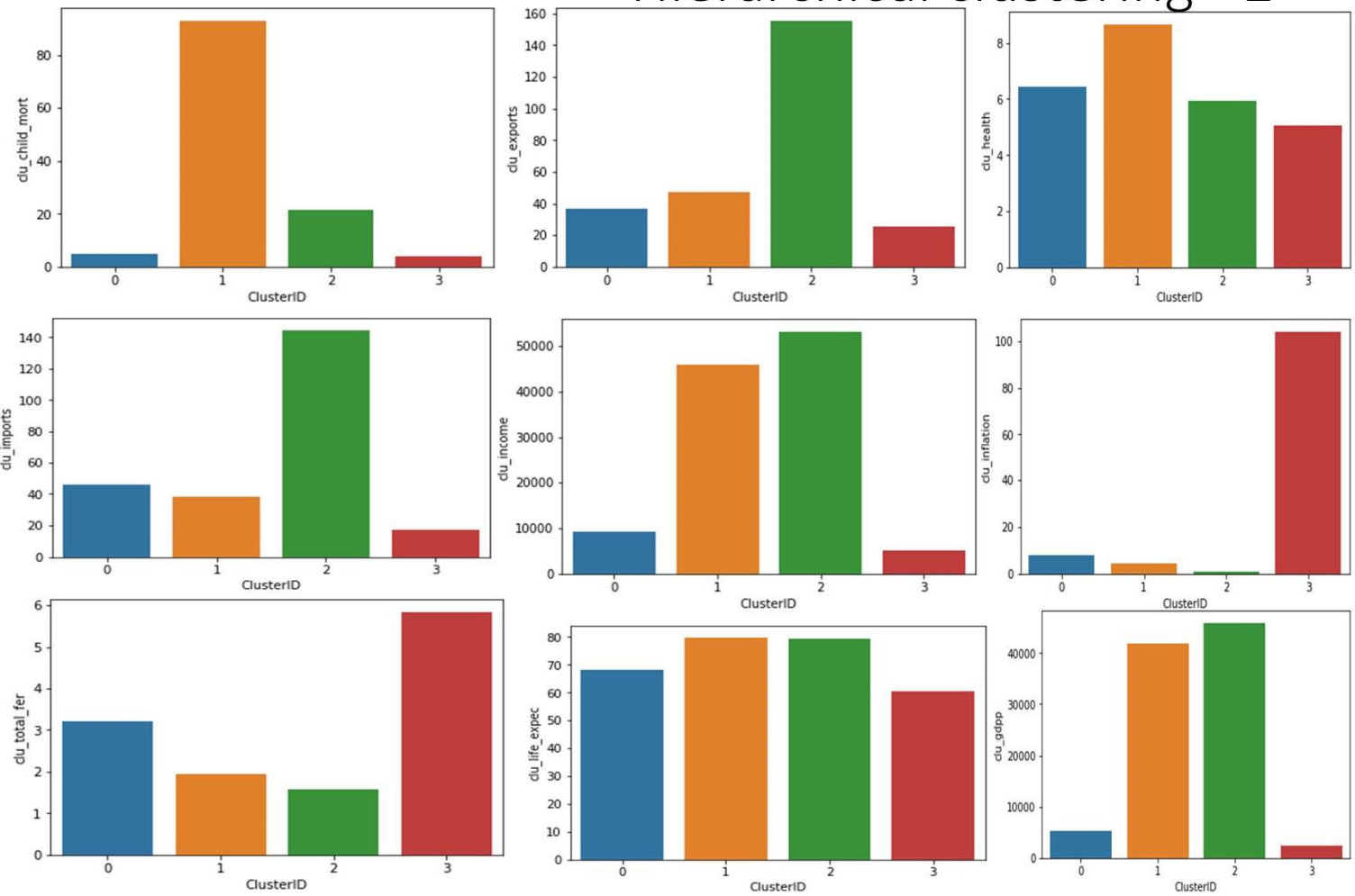


The demand and gap is high compared to the supply.

Huge traffic in the morning peak and evening peak results in more cancellations and no cars available.

Low airport pickup demand during the day results in the high wait time for the drives at the airport.

Hierarchical clustering - 2



From the clusters obtained we observe that cluster 1 countries need more funding from HELP International.

Cluster 1 : High Funding

Cluster 2 : Low Funding

Cluster 3 : No Funding

Cluster 4 : Moderate Funding

Summary

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
Burundi	93.6	8.92	11.60	39.2	764	12.300	57.7	6.26	231	0
Liberia	89.3	19.10	11.80	92.6	700	5.470	60.8	5.02	327	0
Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.800	57.5	6.54	334	0
Niger	123.0	22.20	5.16	49.1	814	2.550	58.8	7.49	348	0
Sierra Leone	160.0	16.80	13.10	34.5	1220	17.200	55.0	5.20	399	0

From K-Means and Hierarchical clustering below are the **top 5 countries which are in direst need** of aid:

1. **Burundi.**
2. **Liberia.**
3. **Congo, Dem. Rep.**
4. **Niger.**
5. **Sierra Leone.**