CLUSTERING OF COUNTRIES

BY-DYUTIMAYA DAS

Abstract:-

Objective: We, HELP International humanitarian NGO, committed to fight poverty and provide the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. We run a lot of operational projects from time to time, along with advocacy, drives to raise awareness as well as for funding purposes.

Problem statement: During the recent funding programs, we have been able to raise around $ 10 million. As an analyst, we have to come up with the countries list that are in the direst need of aid.

Analysis methodology:

Data collection and cleaning

- Import the data
- Identifying the data quality issues and clean the data

Outlier analysis and removal

- Removing the outlier where ever required as per understanding the problem statement.

Data Visualization:-

- Visualization of Data by using matplotlib and seaborn
- Do univariate and bivariate Heatmap,Pairplot analysis of the data

Scaling the data

- Standardizing all the continuous variables to one scale.

Hopkins Statistics

- To check if data has tendency to form clusters

K means clustering

- Identify the 'k' by silhouette analysis and by taking the consideration of Elbow curve.
- Forming 'k' no. of – clusters on scaled data
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

Analysis methodology:

Hierarchical Clustering

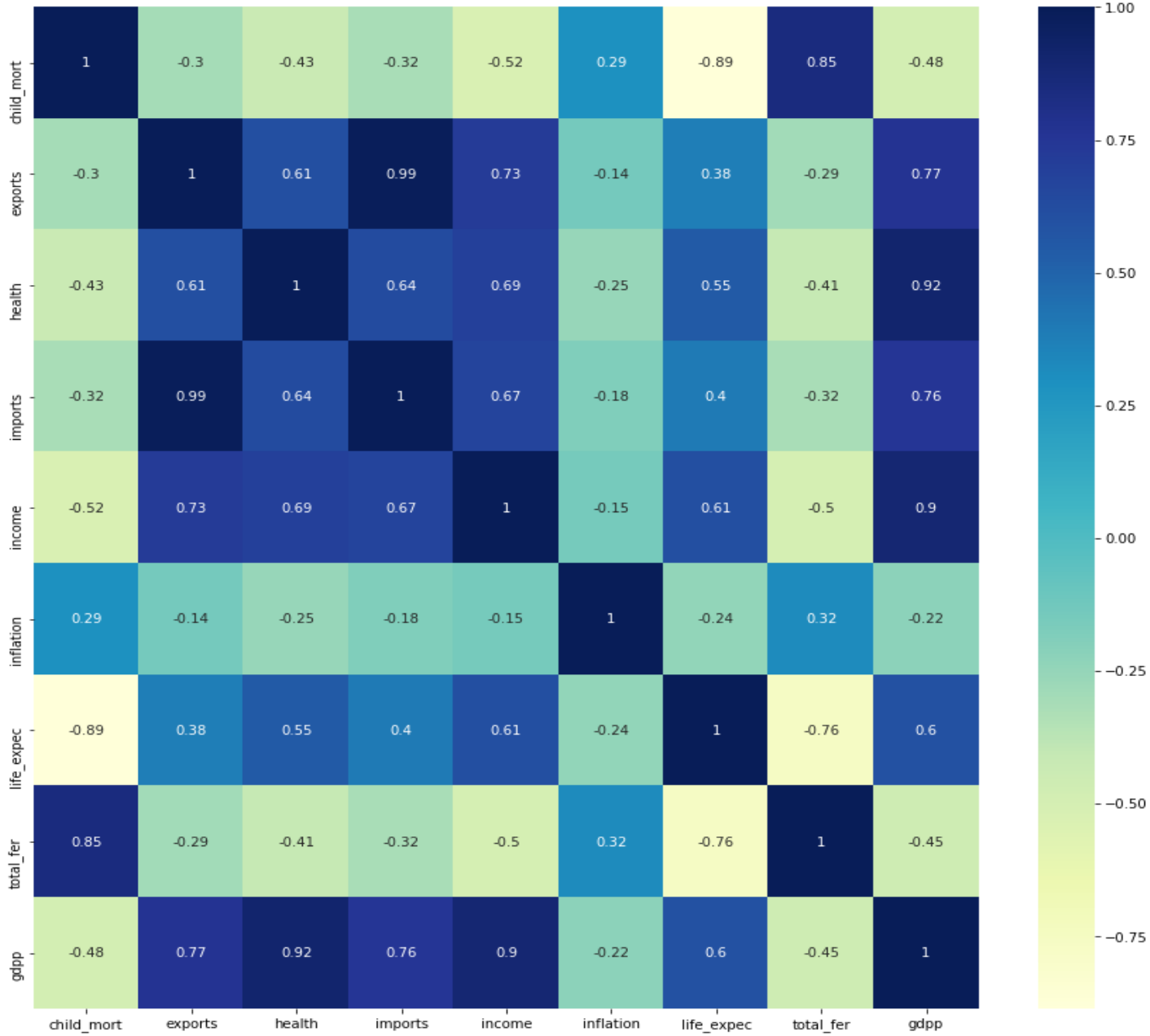
- Identify the 'k' via dendrogram analyse by both the method of single linkage and complete linkage.
- Forming k – clusters on scaled data.
- Visualizing the clusters with various variables.
- Analysing the clusters.
- Identifying the countries which requires aid.

Decision Making

- Identifying the countries which requires aid by analyzing both K-means and Hierarchical Clustering results.

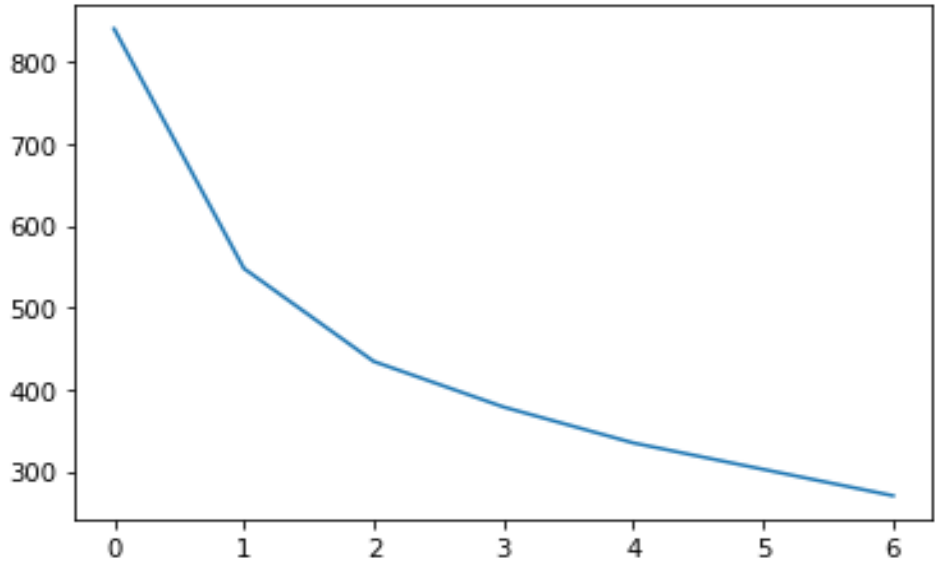
CORRELATION IN DATA BY USING HEATMAP

- After data cleaning , we chose to do nothing to the outliers.
- We did standardized scaling to standardize all parameters on cleaned data.
- Looking at the heat map, we see that few variables like (total fertility, child mortality) , (income , gdpp) and (imports and exports) have high correlation.

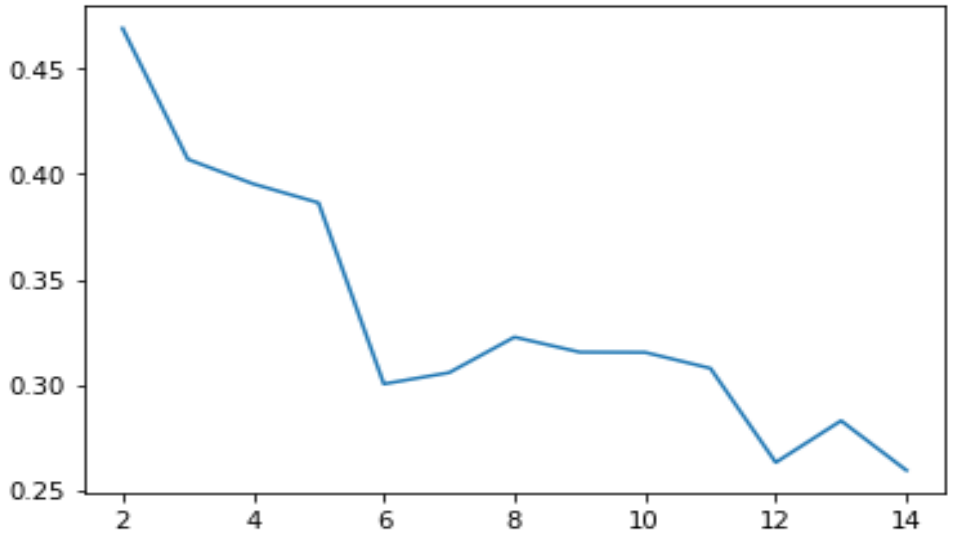


K-MEANS CLUSTERING ALGORITHM

SUM OF SQUARED DISTANCES

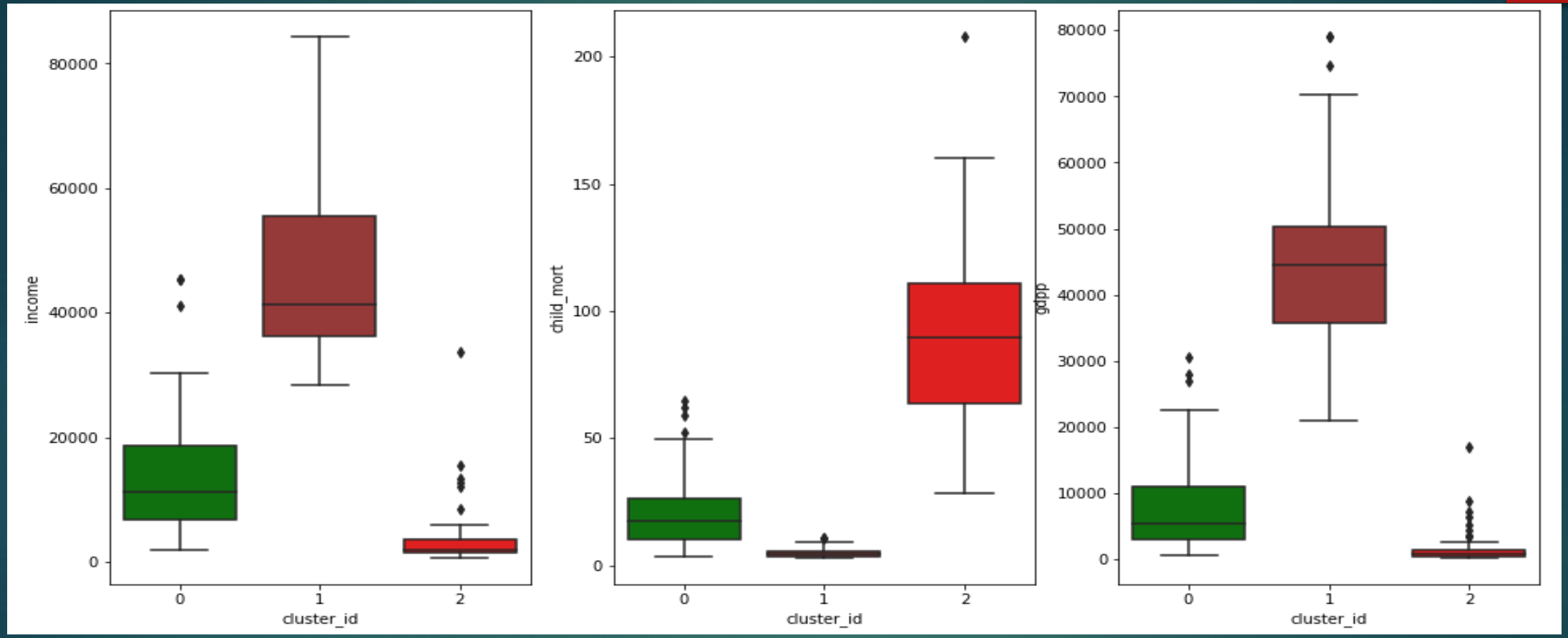


SILHOUETTE ANALYSIS



By Considering Elbow Curve and Silhouette Score we can take the value of k=3.

K-MEANS CLUSTERING



From the above plot we can observe that in cluster-2

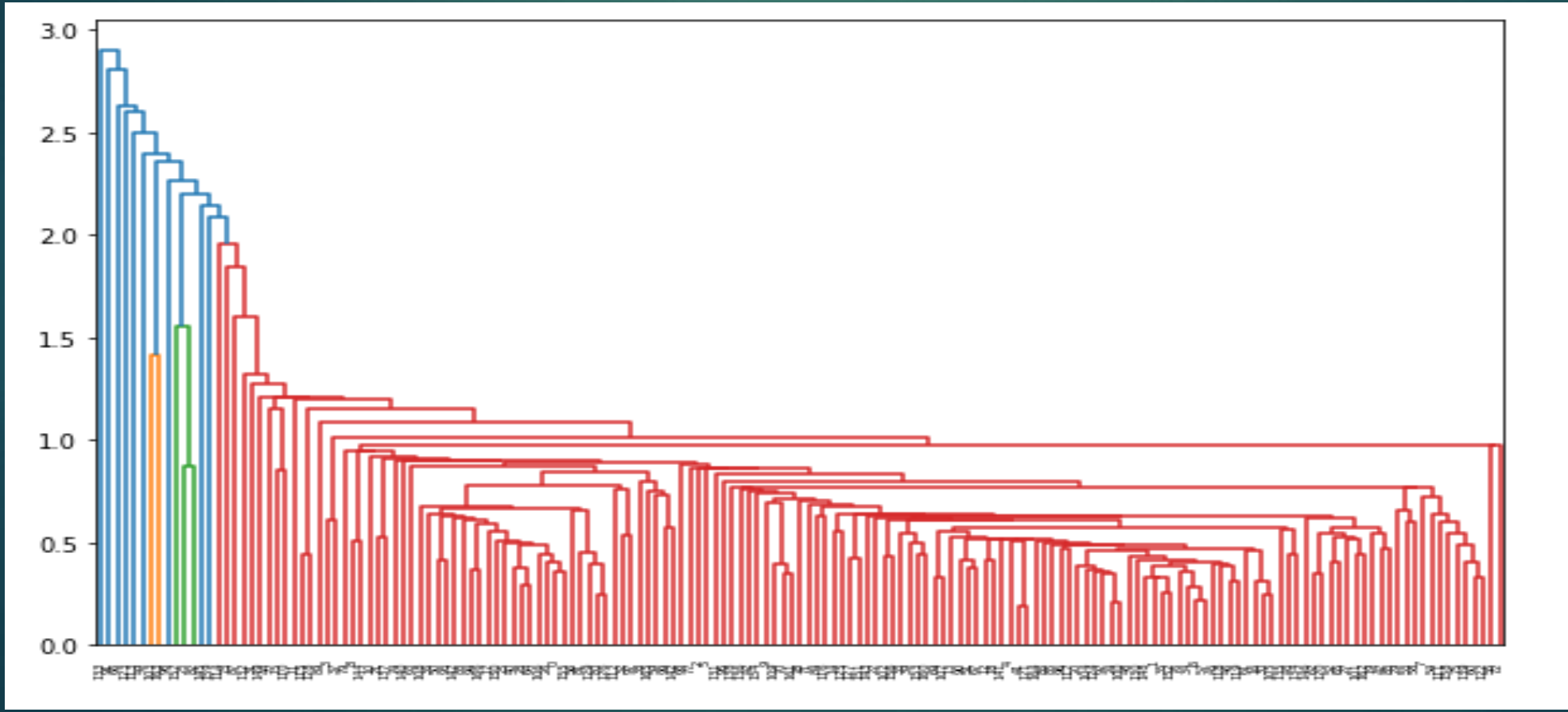
- High Child_mort rate
- Lowest gdpp
- Lowest income.

K-means clustering :
10 countries Under cluster-2 are:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.26	231.0	2
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327.0	2
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334.0	2
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348.0	2
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.20	399.0	2
93	Madagascar	62.2	103.2500	15.5701	177.590	1390.0	8.79	60.8	4.60	413.0	2
106	Mozambique	101.0	131.9850	21.8299	193.578	918.0	7.64	54.5	5.56	419.0	2
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.21	446.0	2
94	Malawi	90.5	104.6520	30.2481	160.191	1030.0	12.10	53.1	5.31	459.0	2
50	Eritrea	55.2	23.0878	12.8212	112.306	1420.0	11.60	61.7	4.61	482.0	2

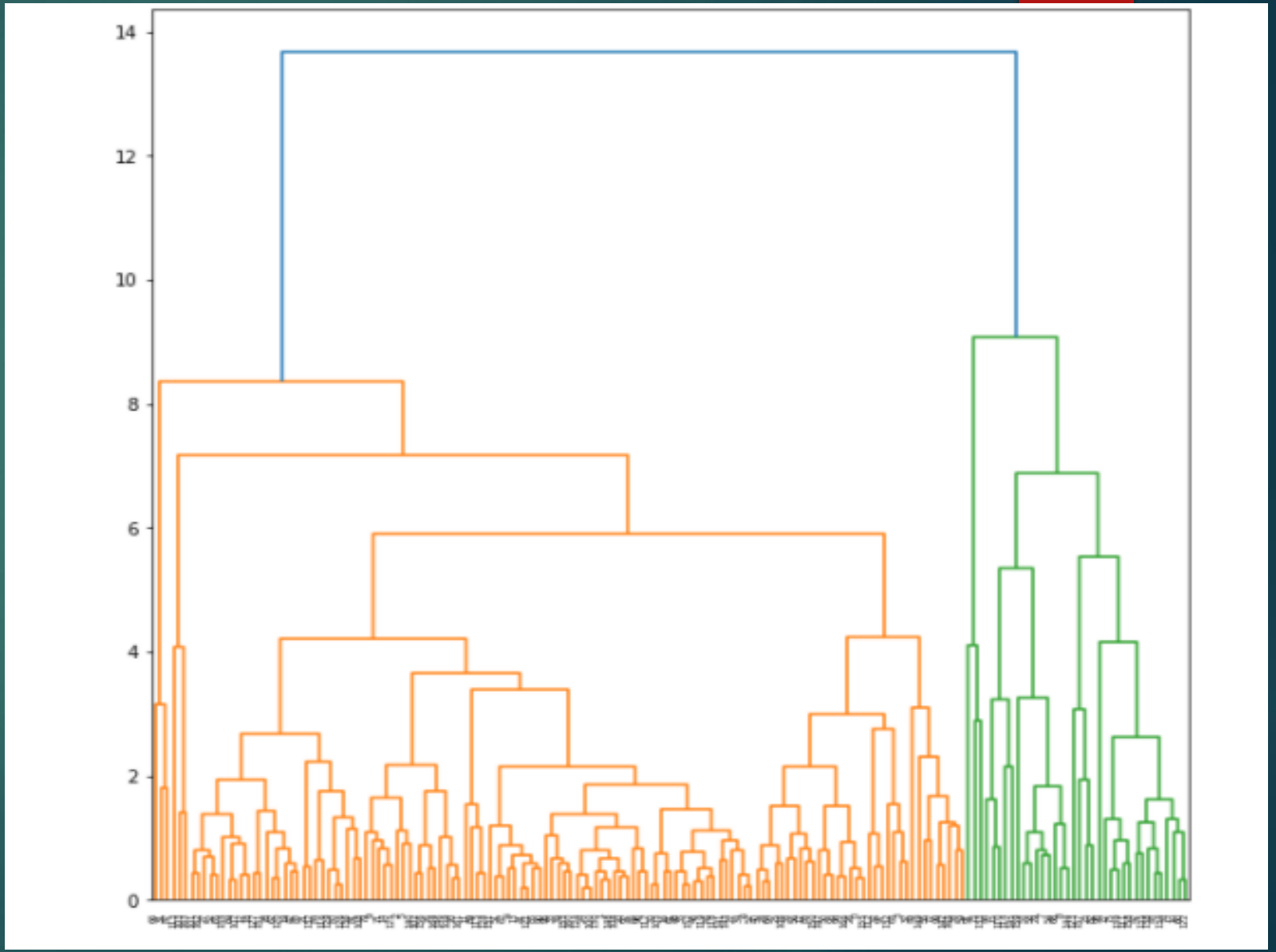
Hierarchical Clustering:

- **Single Linkage method of Hierarchical clustering**



Hierarchical Clustering:
(**Complete Linkage Method)**

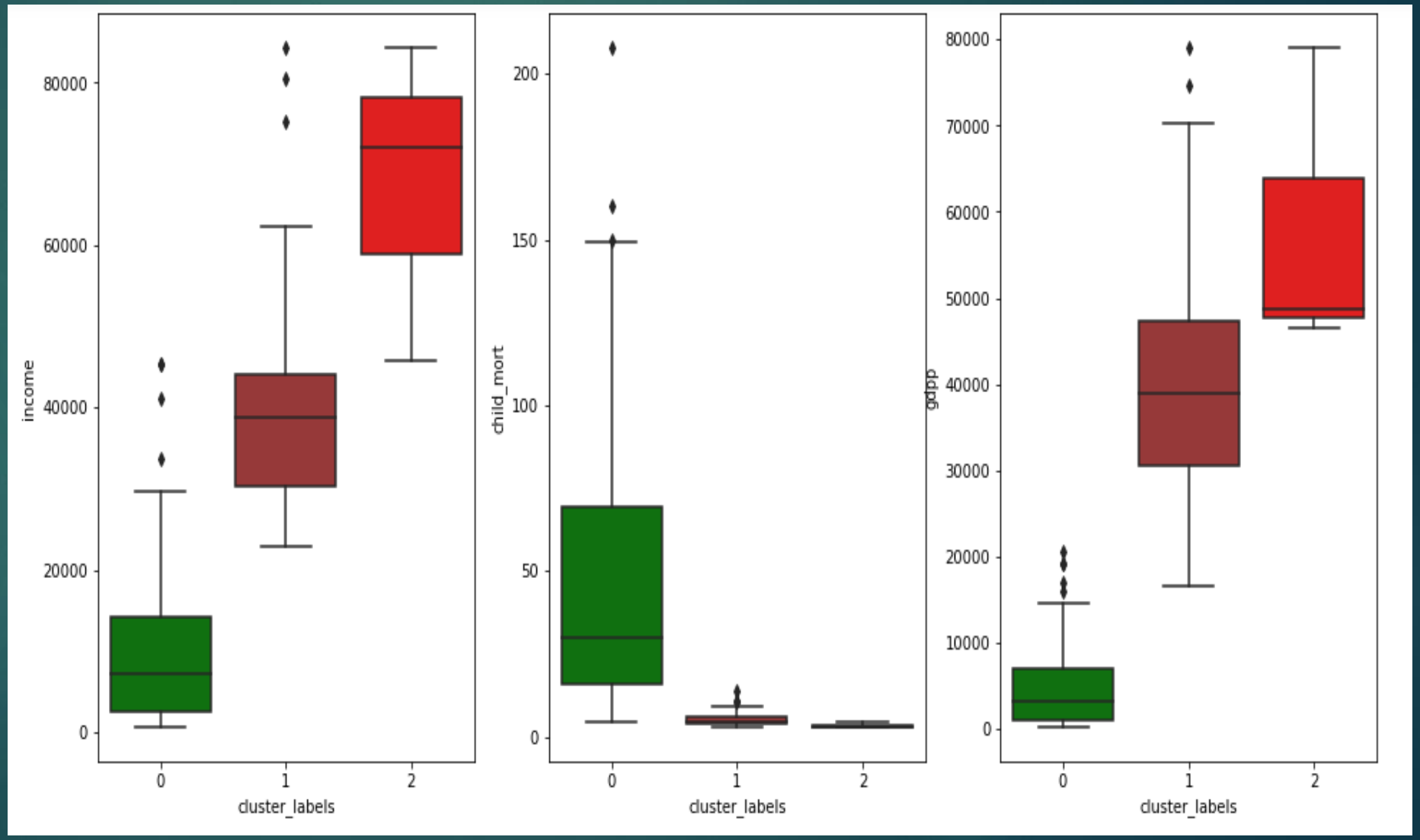
- As the Single method hierarchical Clustering was not so clear so we went for complete linkage method hierarchical clustering.
- By looking at this dendrogram taking k-clusters as 3.



Hierarchical Clustering:

From the above plot we can observe that in cluster-0 Having

- Lowest income
- Highest Child_mort
- Lowest gdpp.



Hierarchical Clustering:

Top 10 countries need to consider for:-

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_labels
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.26	231.0	2	0
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327.0	2	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334.0	2	0
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348.0	2	0
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.20	399.0	2	0
93	Madagascar	62.2	103.2500	15.5701	177.590	1390.0	8.79	60.8	4.60	413.0	2	0
106	Mozambique	101.0	131.9850	21.8299	193.578	918.0	7.64	54.5	5.56	419.0	2	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.21	446.0	2	0
94	Malawi	90.5	104.6520	30.2481	160.191	1030.0	12.10	53.1	5.31	459.0	2	0
50	Eritrea	55.2	23.0878	12.8212	112.306	1420.0	11.60	61.7	4.61	482.0	2	0

Summary:-

- Considering both K-means and Hierarchical clustering method - we have got same countries which requires aid. The following are the countries which are in direst need of aid by considering socio – economic factor into consideration:

  - Burundi
  - Liberia
  - Congo, Dem. Rep.
  - Niger
  - Sierra Leone
  - Madagascar
  - Mozambique
  - Central African Republic
  - Malawi
  - Eritrea