CLUSTERING OF COUNTRIES

Made By:

POOJA AGRAWAL

BUSINESS UNDERSTANDING

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 programmes, 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.

And this is where you come in as a data analyst. Your job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

DATA UNDERSTANDING

The Country-Data consists of the following columns:

Column Name	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services. Given as %age of the Total GDP
health	Total health spending as %age of Total GDP
imports	Imports of goods and services. Given as %age of the Total GDP
income	Net income per person
inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

Data Cleaning

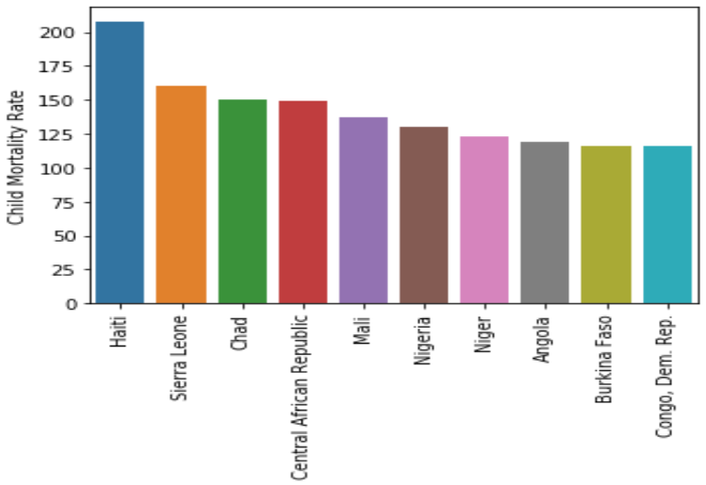
- Data set does not have any missing values nor any inconsistent data type.
- There is no duplicate values provided in dataset.
- We have converted imports, exports and health spending from percentage values to actual values of their GDP per capita. Because the percentage values don't give a clear picture of that country.

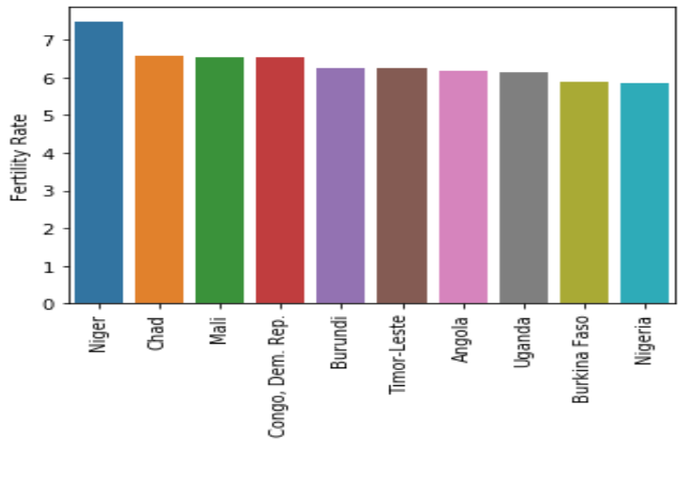
K Means Clustering

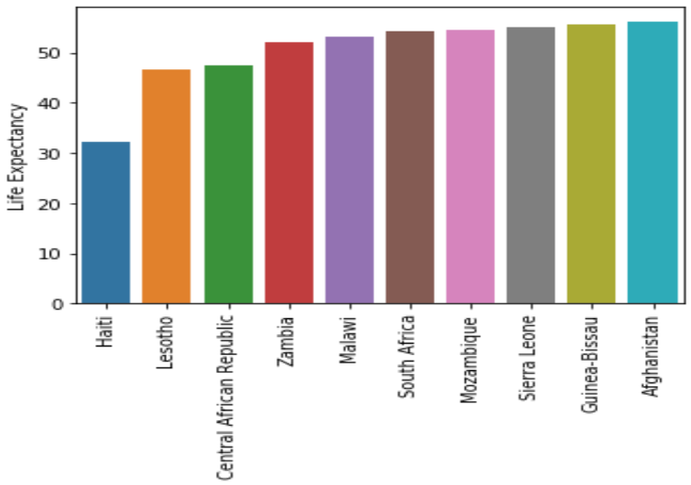
Heat Map

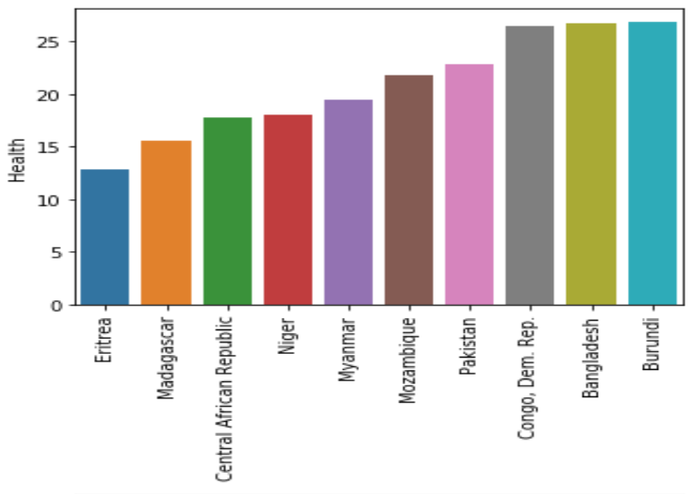
Cluster analysis or simply K Means clustering is done to achieve is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. It is a division of objects into clusters such that each object is in exactly one cluster, not several.. Lets have a look at how our correlation data looks before K Means clustering is performed.

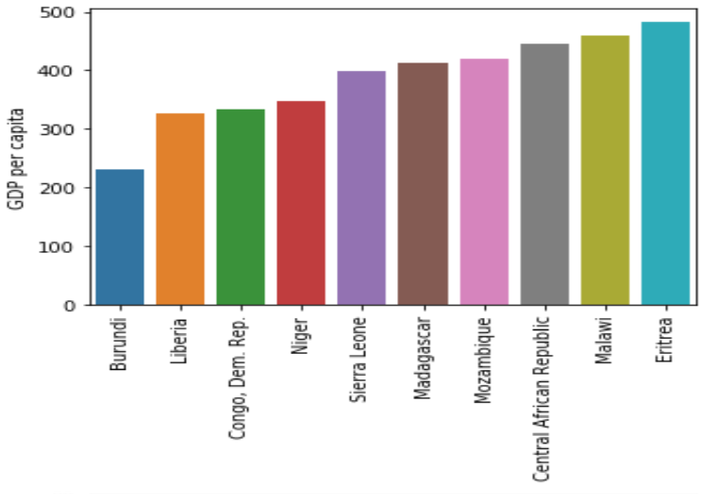
Exploratory Data Analysis

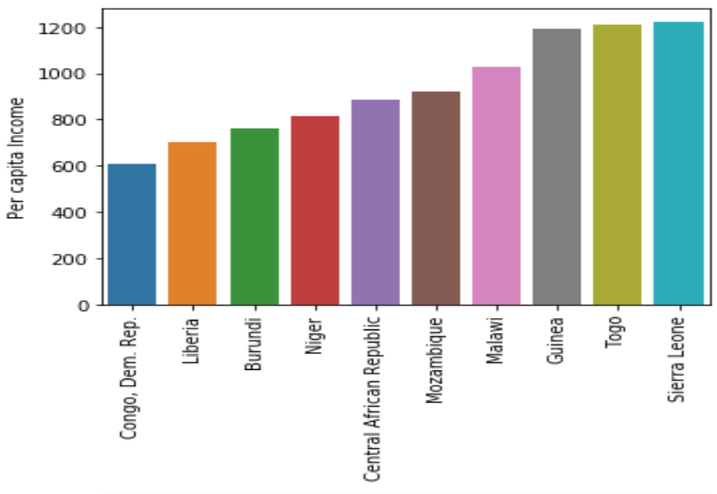


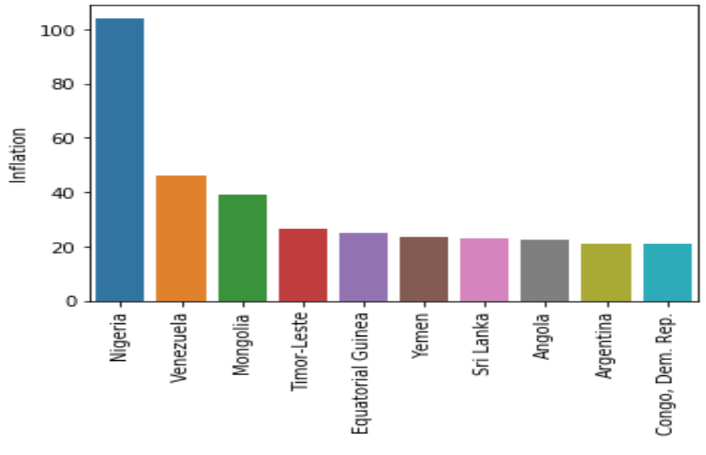


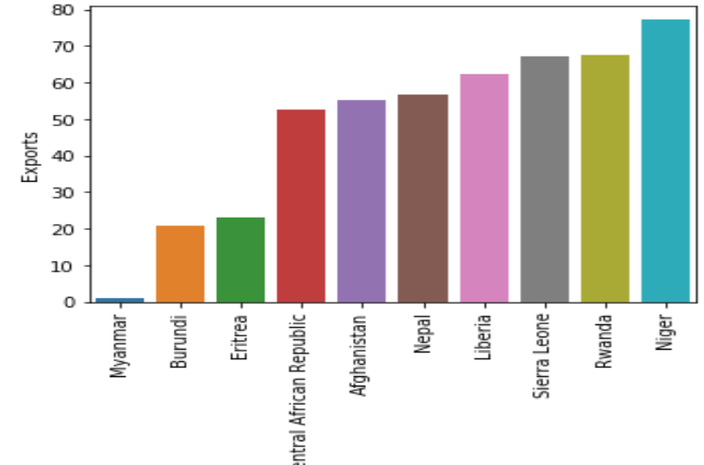


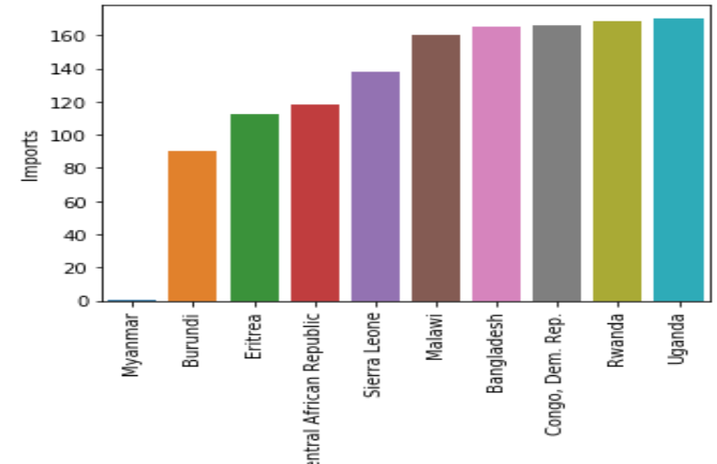




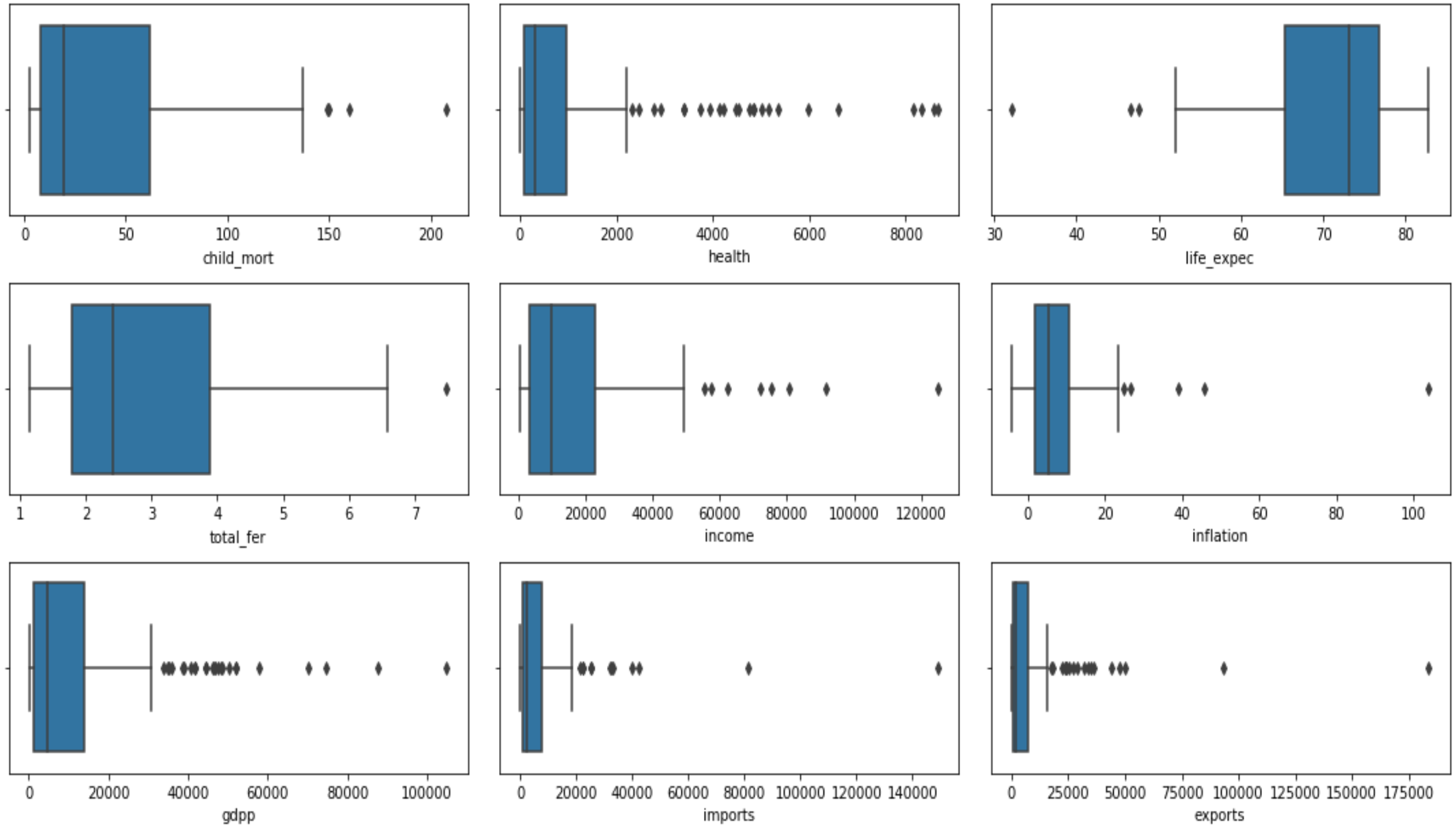




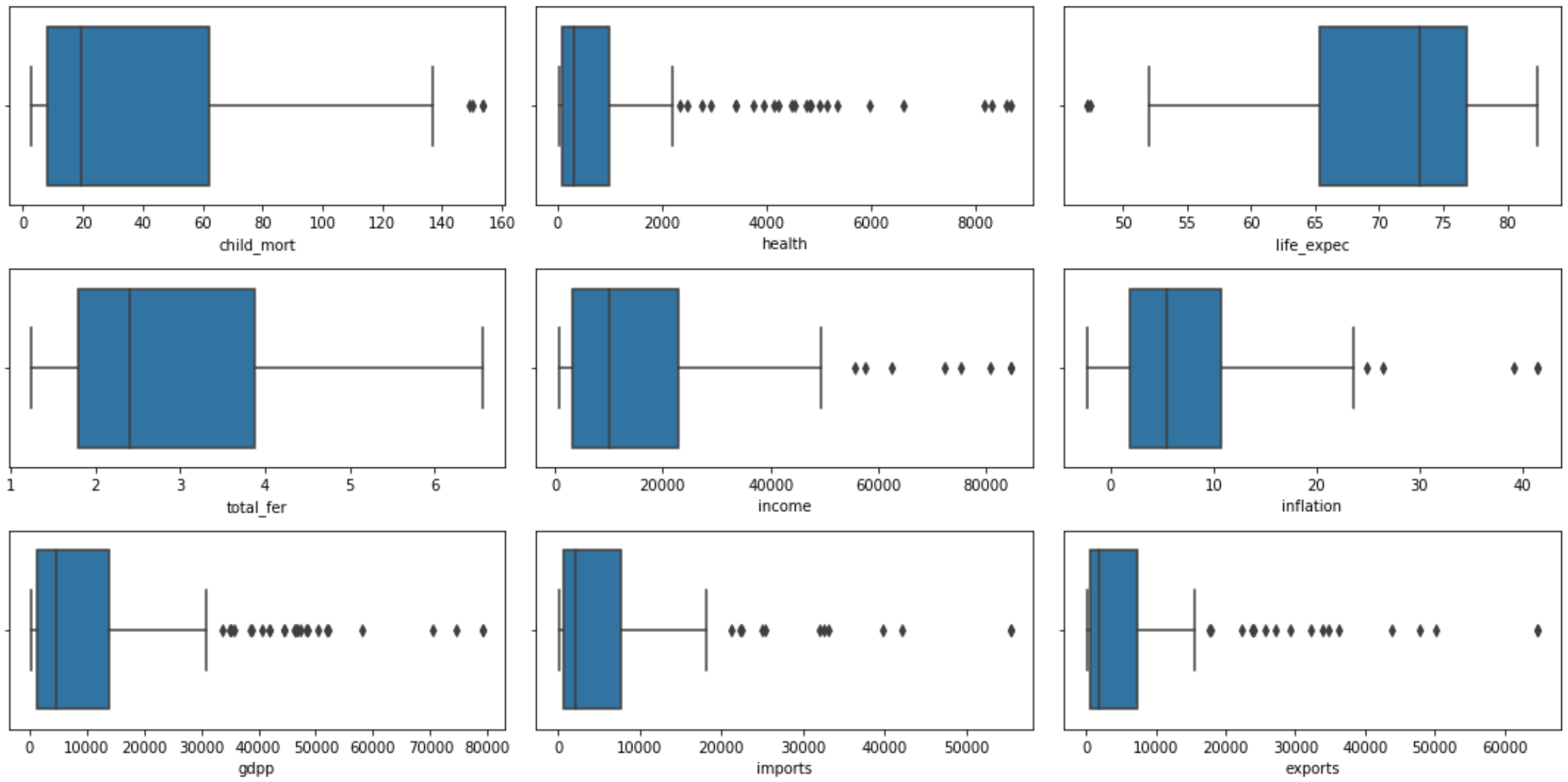




Outlier Analysis



Outlier Treatment



Keeping in mind we need to identify backward countries based on socio economic and health factors. We will cap the outliers to values accordingly for analysis.

Hopkins Check

The Hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) 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.

Inference: 0.92 is a good Hopkins score for Clustering.

Scaling

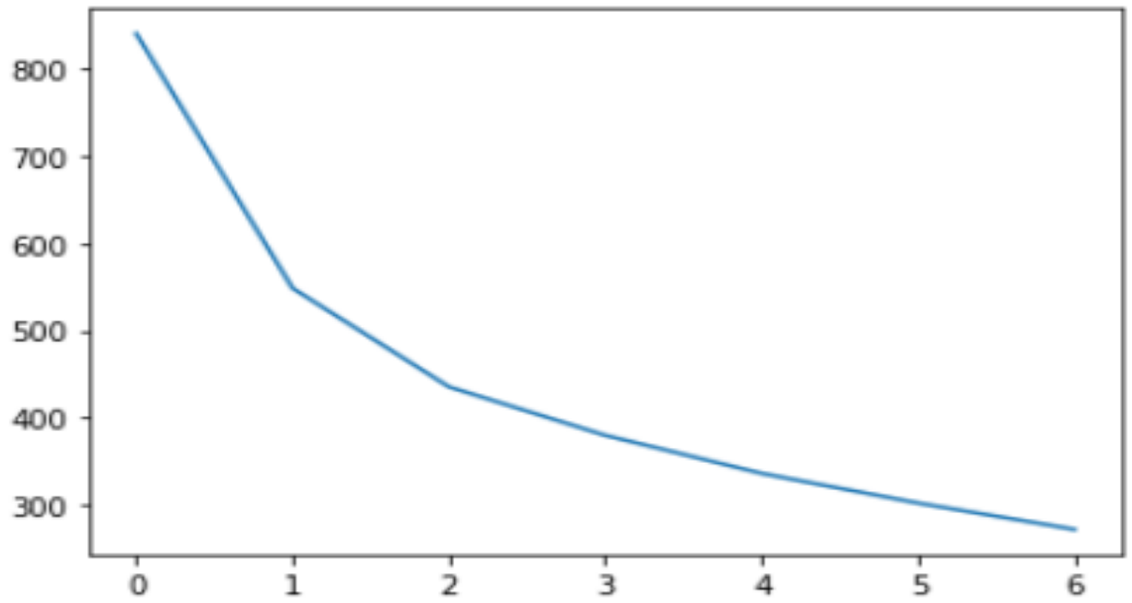
The Euclidean distance is calculated by taking the square root of the sum of the squared differences between observations. This distance can be greatly affected by differences in scale among the variables. So I performed scaling of the variables.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.344012	-0.569638	-0.565164	-0.598844	-0.851772	0.263649	-1.693799	1.926928	-0.702314
1	-0.547543	-0.473873	-0.439335	-0.413679	-0.387025	-0.375251	0.663053	-0.865911	-0.498775
2	-0.272548	-0.424015	-0.484946	-0.476198	-0.221124	1.123260	0.686504	-0.035427	-0.477483
3	2.084186	-0.381264	-0.532486	-0.464070	-0.612136	1.936405	-1.236499	2.154642	-0.531000
4	-0.709457	-0.086754	-0.178874	0.139659	0.125202	-0.768917	0.721681	-0.544433	-0.032079

Finding the Optimal Number of Clusters

Method 1: Elbow-Curve/SSD to get the right no. of clusters

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.



Conclusion:

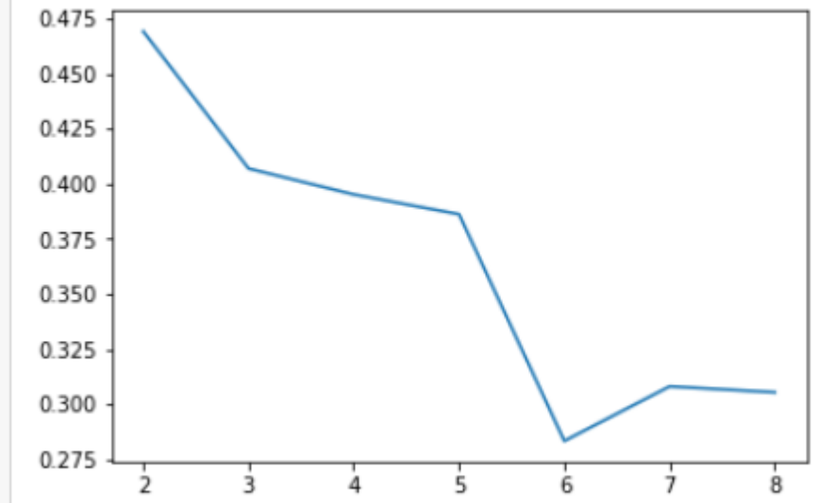
After observing this curve I selected my optimum value of cluster as 3.

Finding the Optimal Number of Clusters
Method 2: Silhouette Analysis

Before proceeding into clustering the data, we find out the optimal number of clusters by the following two methods

The value of the silhouette score range lies between -1 to 1.
A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

```
For n_clusters=2, the silhouette score is 0.4691904548751326
For n_clusters=3, the silhouette score is 0.40696069407925417
For n_clusters=4, the silhouette score is 0.39516613113615756
For n_clusters=5, the silhouette score is 0.3696001191189621
For n_clusters=6, the silhouette score is 0.2814243590064648
For n_clusters=7, the silhouette score is 0.28881689583186215
For n_clusters=8, the silhouette score is 0.29184801498765806
```



Conclusion:
The silhouette score reaches a peak at around 3 clusters indicating that it might be the ideal number of clusters. (k=3)

Also I observed this curve and silhouette score and then I selected my optimum value of cluster as 3.

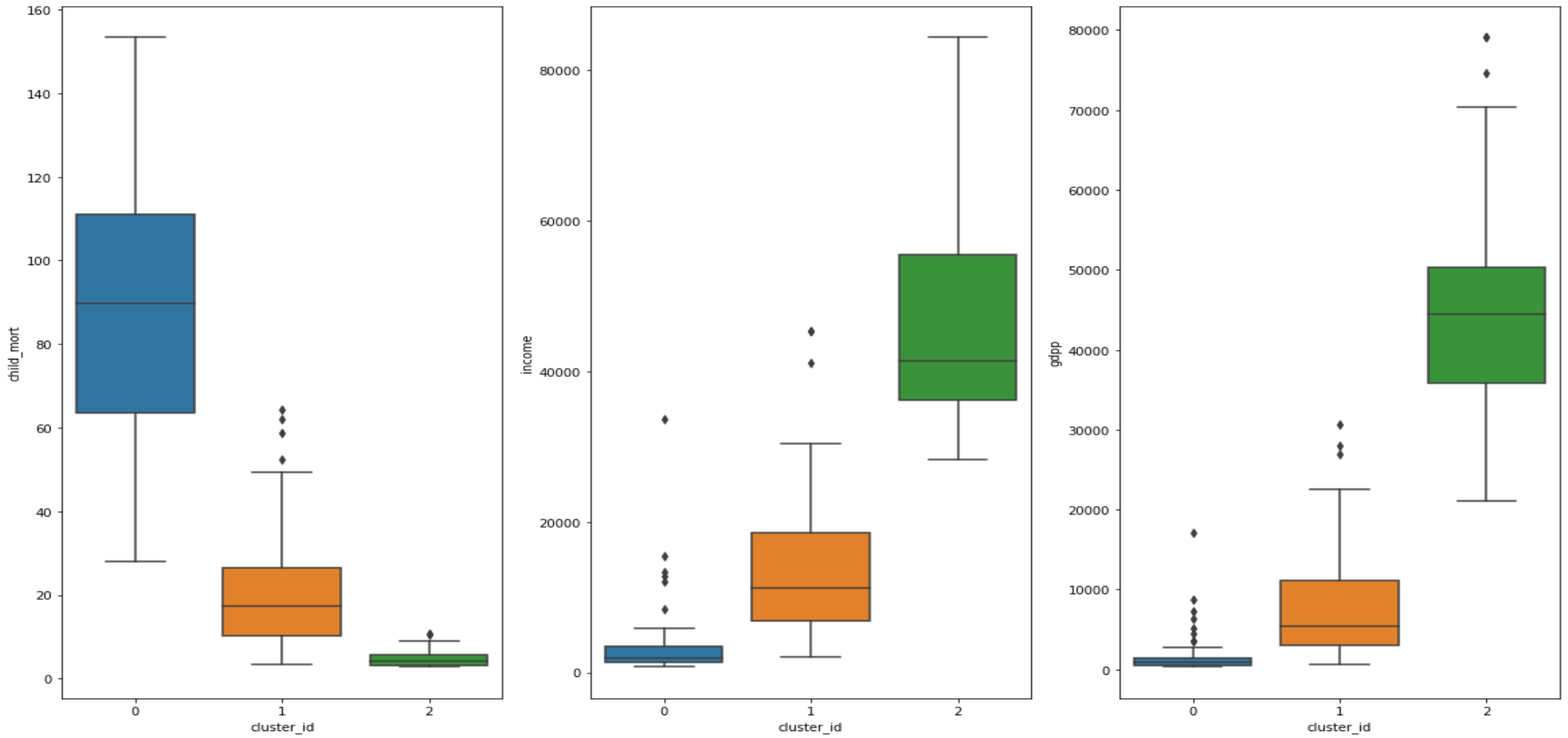
CLUSTERING

Proceeding ahead with 3 Clusters we get the following number of records in each clusters and our data head looks as follows:

Cluster ID	Value Counts
1	90
0	48
2	29

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	Afghanistan	90.2	55.30	41.9174	248.297	1610.0	9.44	56.2	5.82	553.0	0
1	Albania	16.6	1145.20	267.8950	1987.740	9930.0	4.49	76.3	1.65	4090.0	1
2	Algeria	27.3	1712.64	185.9820	1400.440	12900.0	16.10	76.5	2.89	4460.0	1
3	Angola	119.0	2199.19	100.6050	1514.370	5900.0	22.40	60.1	6.16	3530.0	0
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100.0	1.44	76.8	2.13	12200.0	1

From the business understanding we have learnt that Child_Mortality, Income, Gdpp are some important factors which decides the development of any country. Hence, we will proceed with cluster profiling by using these 3 variables



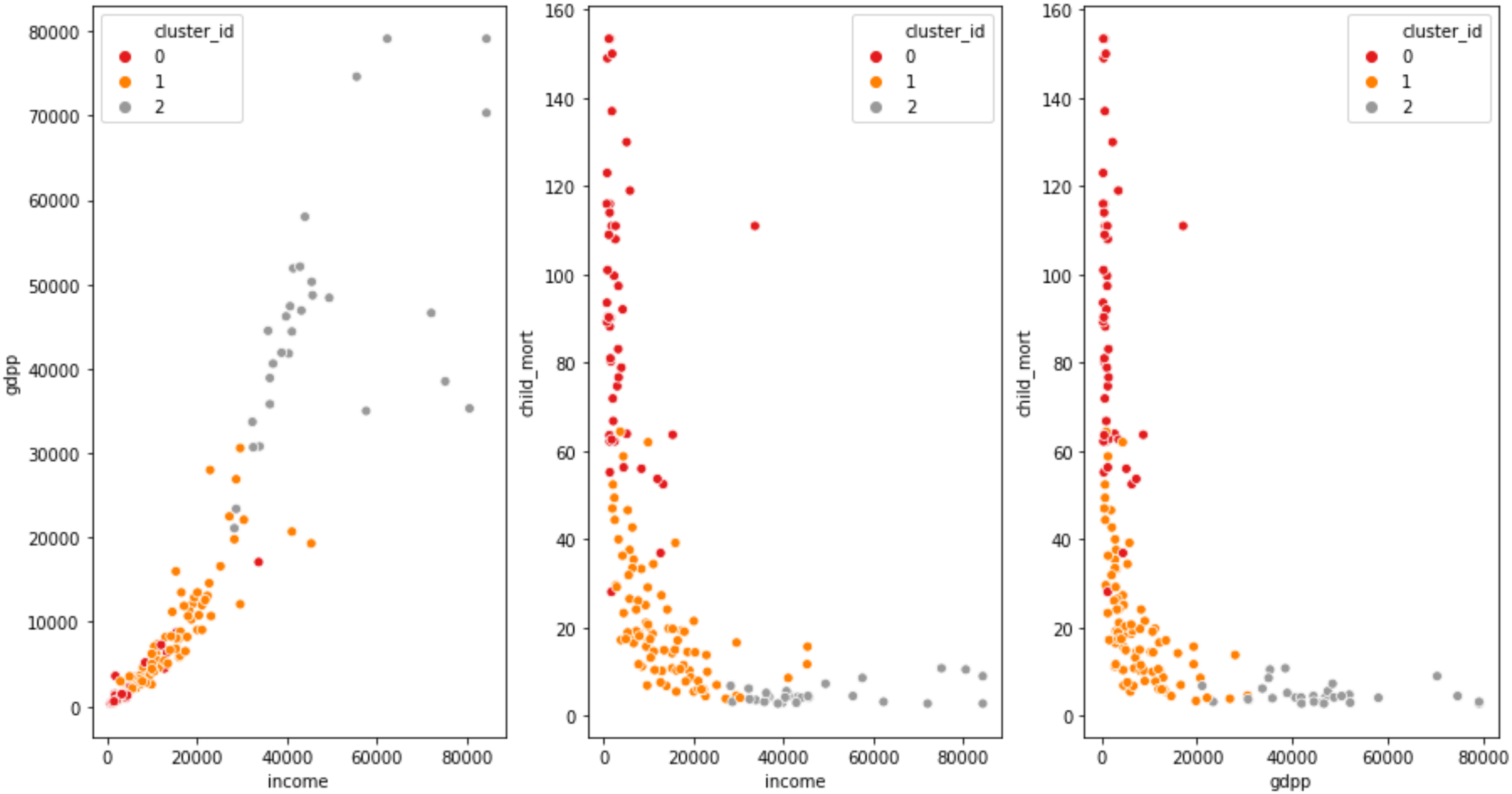
Inference:

Cluster 0 has lowest GDPP so we can say that the countries in cluster 0 must be in high Aid.

Cluster 0 has lowest income so we can say that the countries in cluster 0 must be in high Aid.

Cluster 0 has highest mortality rate so we can say that the countries in cluster 0 must be in high Aid.

Cluster wise comparison of Income''gdpp''child_mort'

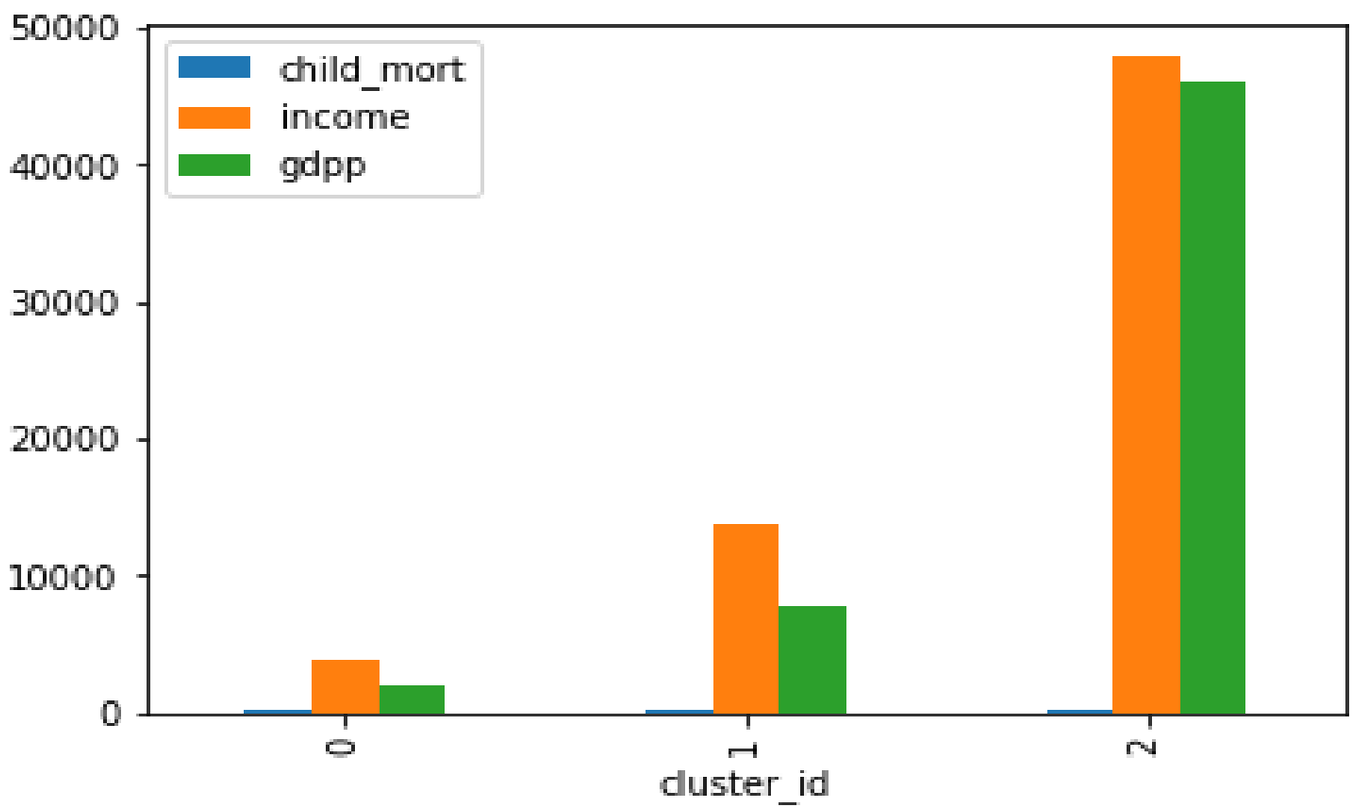


Final Inference:

Child Mortality is highest for Cluster 0, These clusters need some aid. Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in clusters 0. Hence, these countries need some help.

Cluster Profiling

cluster_id	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	90.335417	879.097657	115.348635	827.327888	3901.010000	10.608604	59.567083	4.972233	1911.400833
1	20.547778	3477.250726	528.925228	3589.291996	13804.333333	7.131624	73.393333	2.242591	7808.577778
2	4.989655	25405.359310	4253.879655	21316.695862	47784.413793	2.906731	80.453103	1.757352	46068.137931



Cluster Analysis

1. We observe that the best country cluster is cluster 0 based on our three important columns.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	Afghanistan	90.2	55.300	41.9174	248.297	1610.0	9.440	56.2	5.82	553.0	0
3	Angola	119.0	2199.190	100.6050	1514.370	5900.0	22.400	60.1	6.16	3530.0	0
17	Benin	111.0	180.404	31.0780	281.976	1820.0	0.885	61.8	5.36	758.0	0
21	Botswana	52.5	2768.600	527.0500	3257.550	13300.0	8.920	57.1	2.88	6350.0	0
25	Burkina Faso	116.0	110.400	38.7550	170.200	1430.0	6.810	57.9	5.87	575.0	0

Final list of top 10 countries that needs Aid from CEO by K Means Clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
88	Liberia	89.3	62.457000	38.586000	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.419400	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.796000	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	22.243716	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	22.243716	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	22.243716	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.248100	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.946400	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.332000	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.269000	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

Hierarchical Clustering

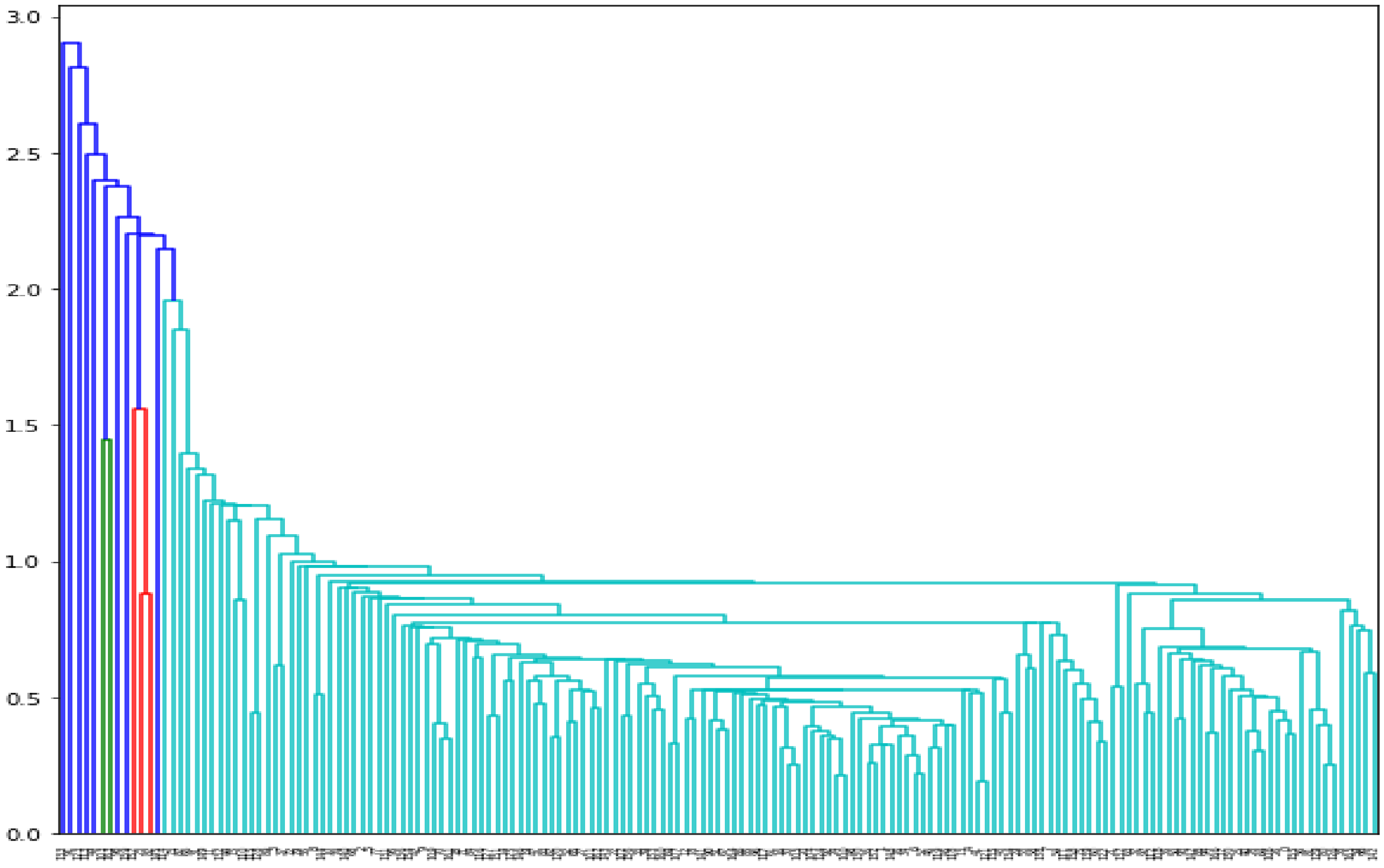
Hierarchical Clustering has one advantage over K-means Clustering which is that we don't have to select the initial number of clusters before performing clustering.

It has a different concept of linkage through which it performs the clustering operations. There are two types of Linkage:

1. Single Linkage
2. Complete Linkage

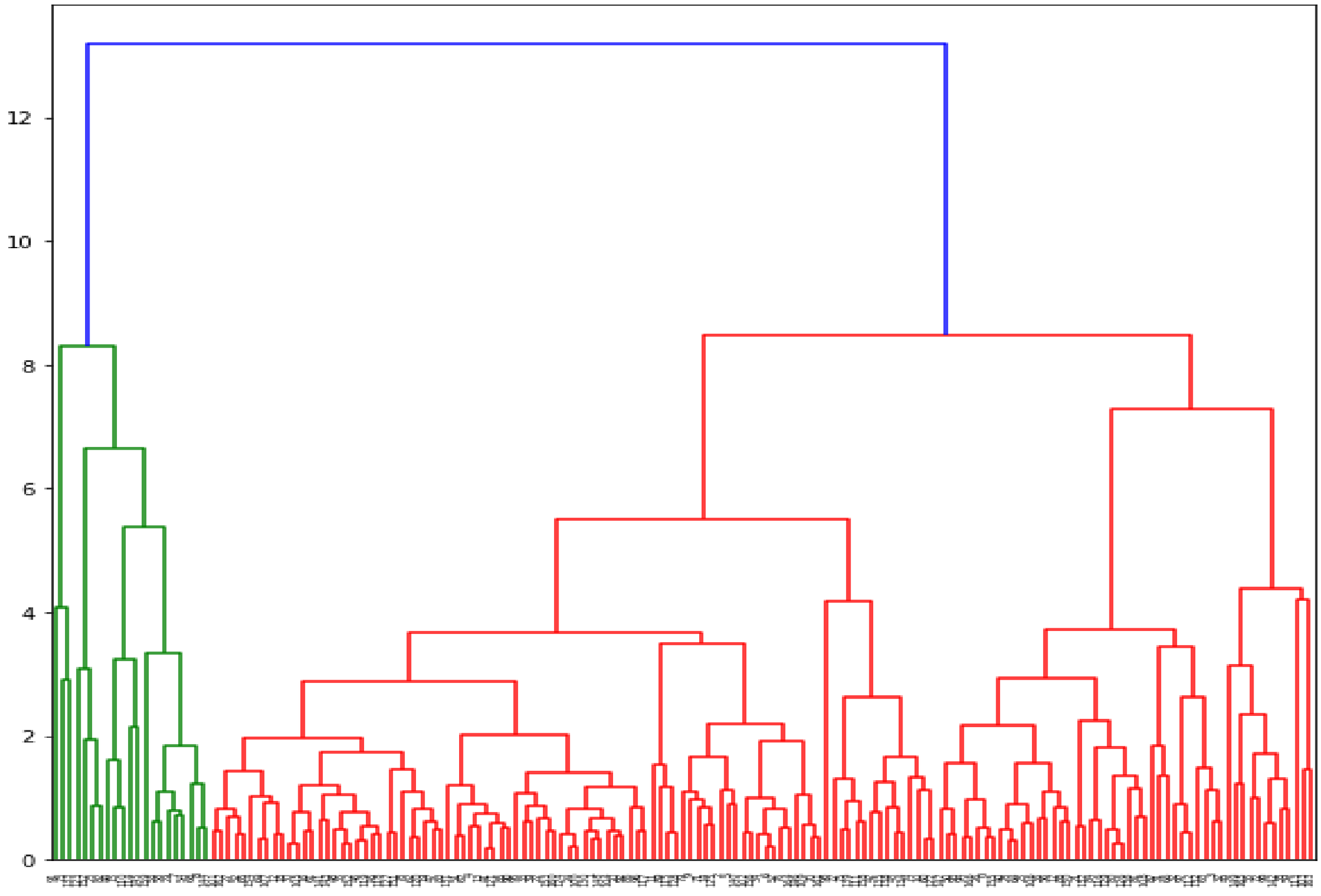
Lets try both the methods on our country data and see if the results are good enough.

Single Linkage



Single Linkage do not give clear cluster formation so we have to try complete linkage in the next step.

Complete Linkage



Now we see some good amount of clusters getting formed.

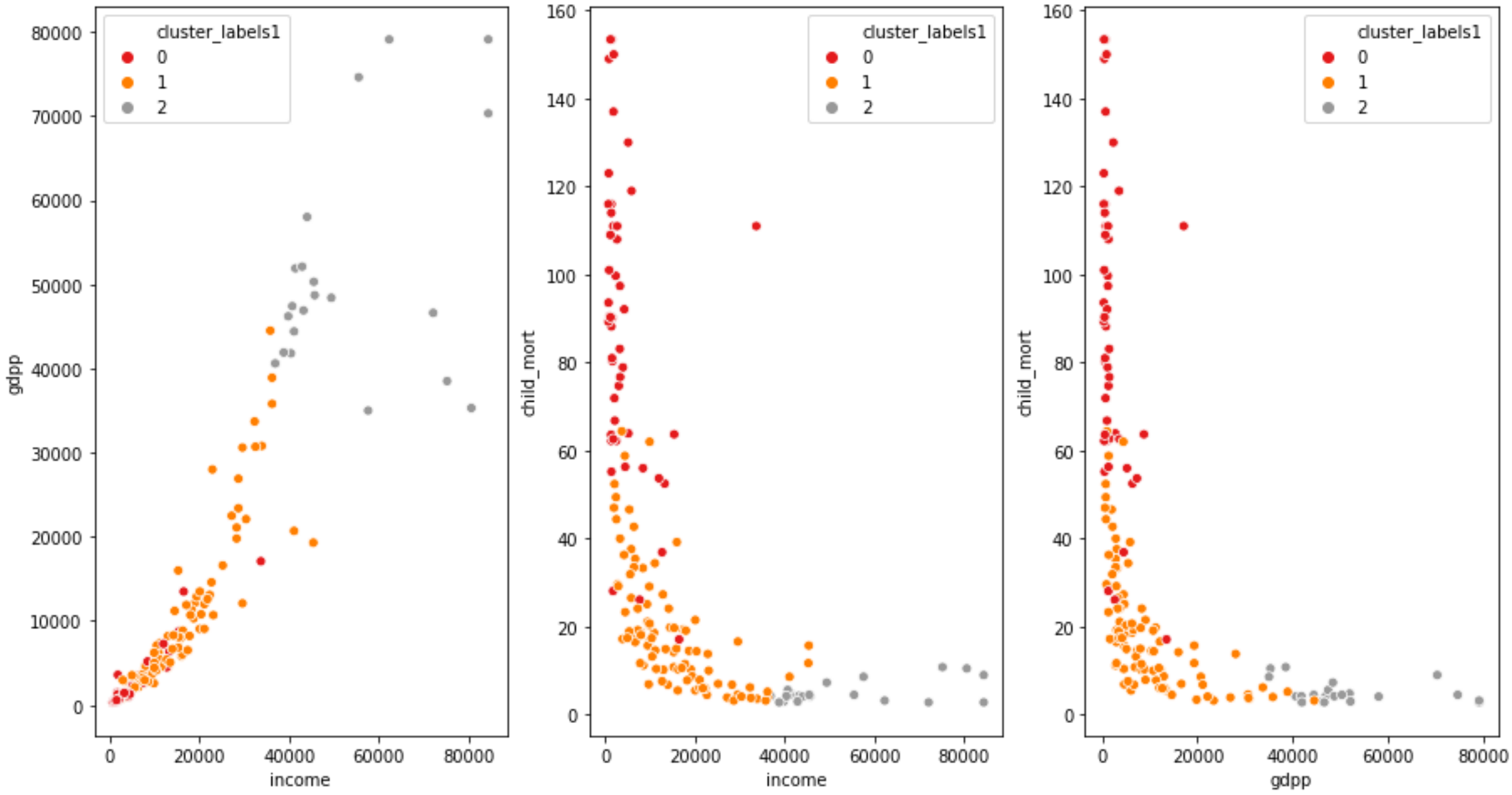
Now if we cut the tree at 3 clusters and look at our data head after assigning the cluster ids. The results are as such:

Cluster ID	Value Counts
1	96
0	50
2	21

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels1
0	Afghanistan	90.2	55.30	41.9174	248.297	1610.0	9.44	56.2	5.82	553.0	0
1	Albania	16.6	1145.20	267.8950	1987.740	9930.0	4.49	76.3	1.65	4090.0	1
2	Algeria	27.3	1712.64	185.9820	1400.440	12900.0	16.10	76.5	2.89	4460.0	1
3	Angola	119.0	2199.19	100.6050	1514.370	5900.0	22.40	60.1	6.16	3530.0	0
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100.0	1.44	76.8	2.13	12200.0	1

Original Dataset Data Head

We observe that the clustering on the other clusters are similar as that in K-means.

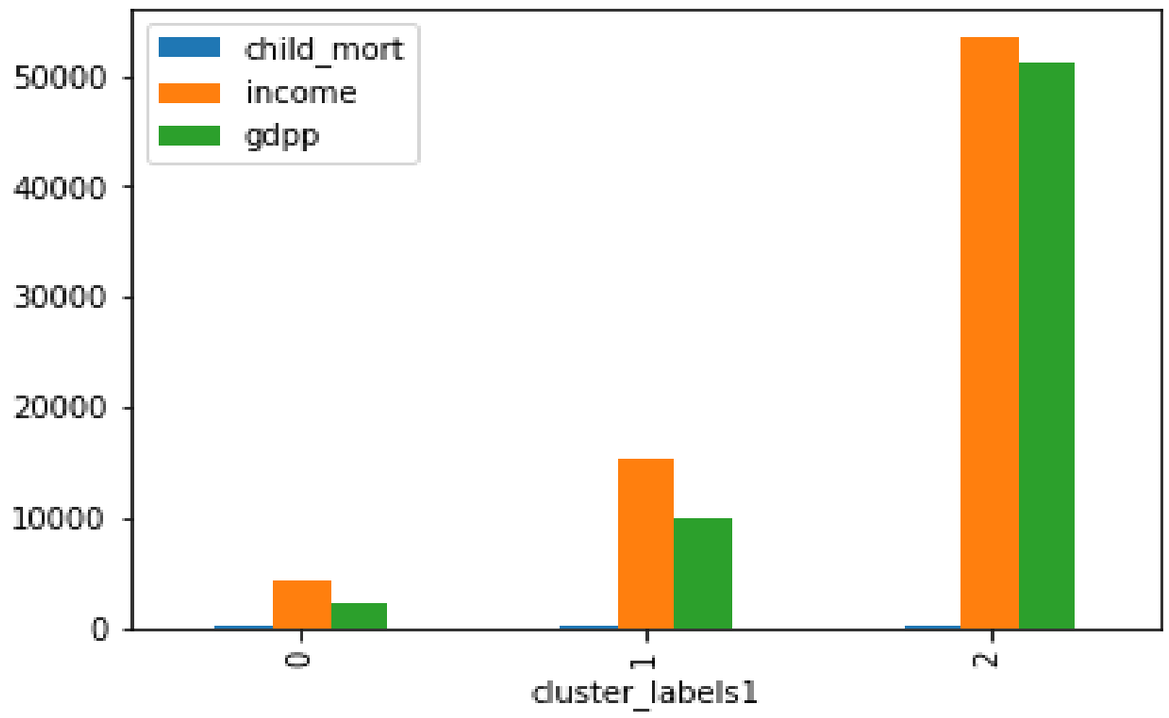


Inference:

Child Mortality is highest for Cluster 0. These cluster need some aid. Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in clusters 0. Hence, these countries need some help.

Cluster Profiling

cluster_labels1	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	87.586000	945.634750	126.874890	871.805773	4229.169600	11.797820	60.016400	4.875544	2157.944800
1	19.188542	4326.711618	733.118130	4474.111767	15438.333333	5.936460	74.069479	2.181075	9849.187500
2	5.176190	29964.696190	4751.401429	24182.246667	53421.333333	3.598248	80.298571	1.823962	51289.333333



Inference:

Child Mortality is highest for Cluster 0, These clusters need some aid. Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. However Income per capita and gdpp seems lowest for countries in clusters 0. Hence, countries in cluster 0 need some help.

Cluster Analysis

1. We observe that the best country cluster is cluster 0 based on our 3 important columns.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels1
0	Afghanistan	90.2	55.300	41.9174	248.297	1610.0	9.440	56.2	5.82	553.0	0
3	Angola	119.0	2199.190	100.6050	1514.370	5900.0	22.400	60.1	6.16	3530.0	0
17	Benin	111.0	180.404	31.0780	281.976	1820.0	0.885	61.8	5.36	758.0	0
21	Botswana	52.5	2768.600	527.0500	3257.550	13300.0	8.920	57.1	2.88	6350.0	0
25	Burkina Faso	116.0	110.400	38.7550	170.200	1430.0	6.810	57.9	5.87	575.0	0

Final list of top 10 countries that needs Aid from CEO

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels1
88	Liberia	89.3	62.457000	38.586000	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.419400	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.796000	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	22.243716	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	22.243716	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	22.243716	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.248100	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.946400	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.332000	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.269000	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

Conclusion

Based on my analysis I followed below observations to choose the countries that are in need of aid:

So firstly I have analyzed both K-means and Hierarchical clustering and found clusters formed are also identical by both the methods The clusters formed in both the cases are great and I can choose anyone of the method. So, I will proceed with the clusters formed by hierarchical clustering as we know whenever we have smaller data set we should go with hierarchical clustering and based on the information provided by the final clusters I will deduce the final list of countries which are in need of aid.

FINAL LIST OF TOP 10 COUNTRIES TO FOCUS ON

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels1
88	Liberia	89.3	62.457000	38.586000	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.419400	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.796000	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	22.243716	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	22.243716	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	22.243716	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.248100	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.946400	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.332000	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.269000	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

THANK YOU