OBJECTIVE

CEO, of the HELP INTERNATIONAL NGO needs to decide how to use $10 million fund raised by it, 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 job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO needs to focus on the most.

PROBLEM STATEMENT

Data inspection and EDA tasks suitable for this dataset – data cleaning, univariate analysis, bivariate analysis, outlier analysis etc.

Try both k-means and hierarchical clustering(both single and complete linkage) on this dataset to create the clusters.

Analyze the clusters and identify the ones which are in dire need of aid, by comparing how these three variables - [**gdpp, child mortality** and **income**] vary for each cluster of countries.

VIEWING AND UNDERSTANDING THE DATASET

First 5 Rows of Dataset

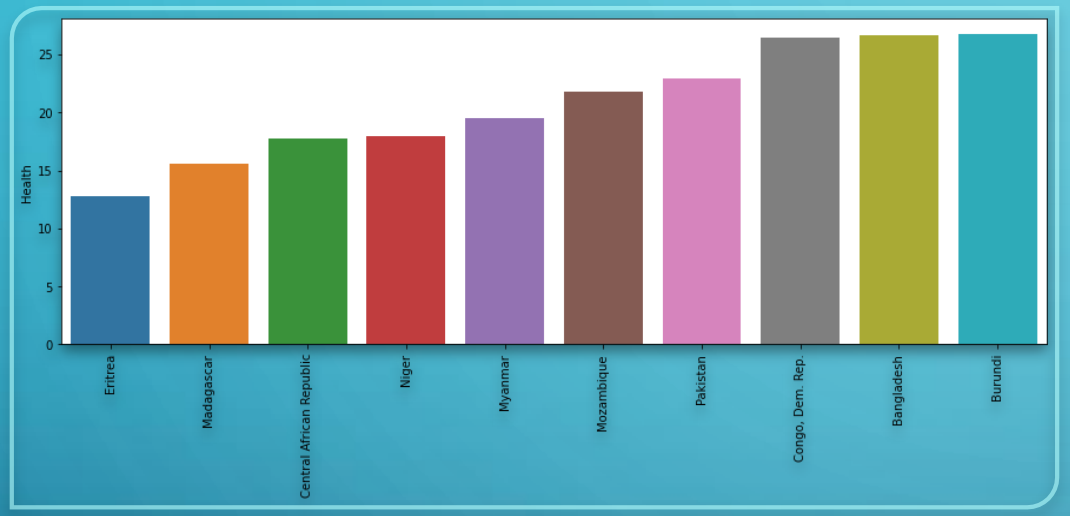
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

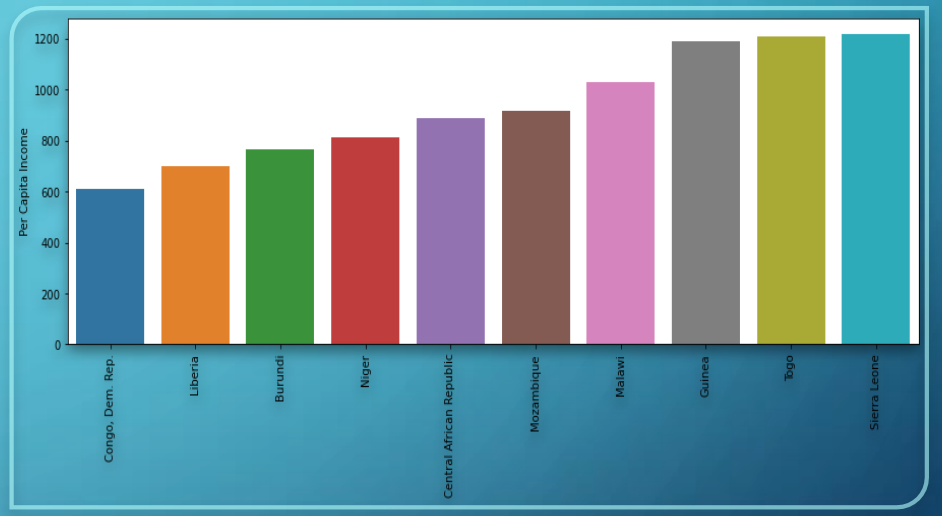
Description of the Dataset

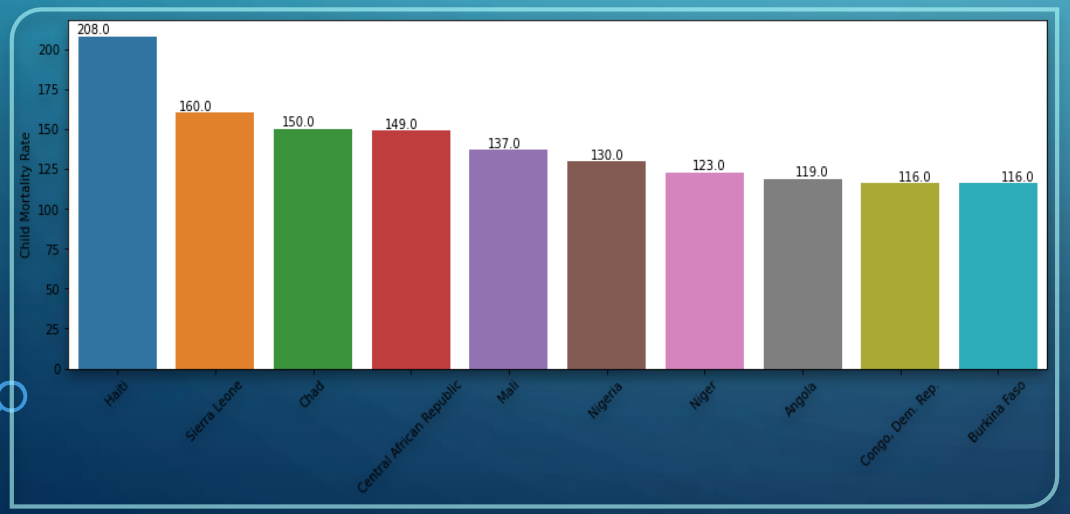
	Column Name	Description
0	country	Name of the country
1	child_mort	Death of children under 5 years of age per 1000 live births
2	exports	Exports of goods and services per capita. Given as %age of the GDP per capita
3	health	Total health spending per capita. Given as %age of GDP per capita
4	imports	Imports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

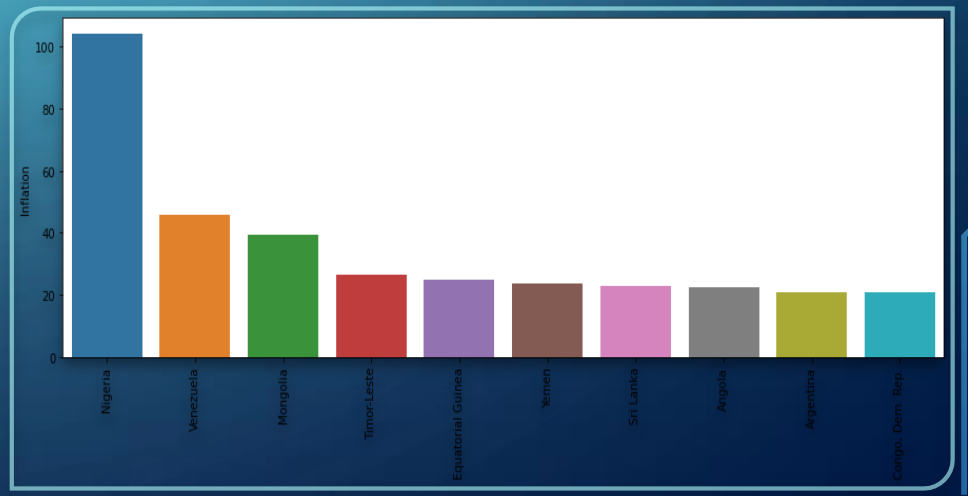
DATA VISUALIZATION FOR TOP 10 UNDER DEVELOPED COUNTRIES.

The below plots shows the countries which have low per capita income and high inflation whereas low expenditure on health and a child mortality rate. These can be considered as under developed countries. By far now, from these plots, we can consider these countries as good candidate for NGO's aids





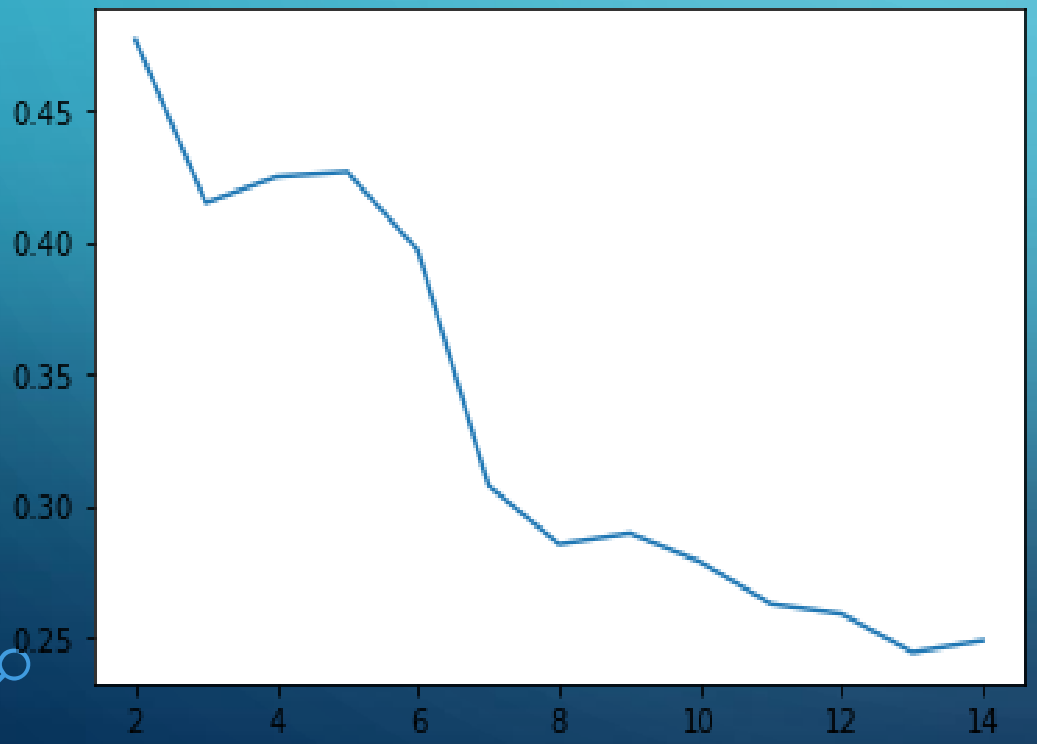




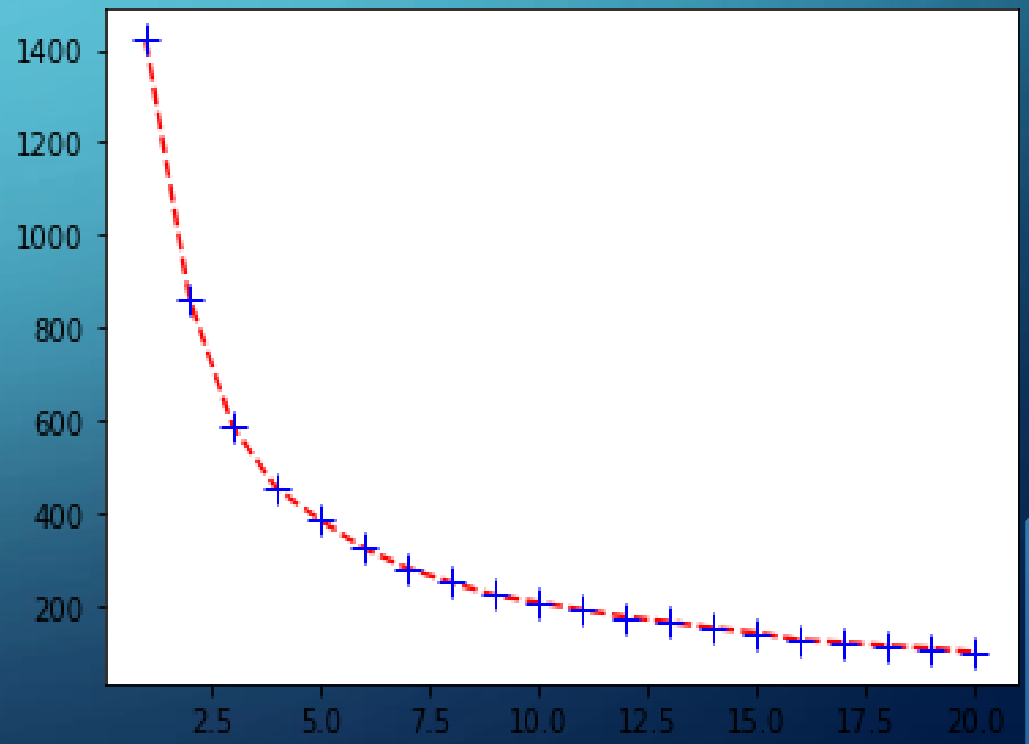
K-MEANS CLUSTERING

By looking at the silhouette analysis, we see the highest peak is at k=4 and in sum of squared distances graph, we see that the elbow is in the range of 3 to 5, so we are going ahead with 4 clusters

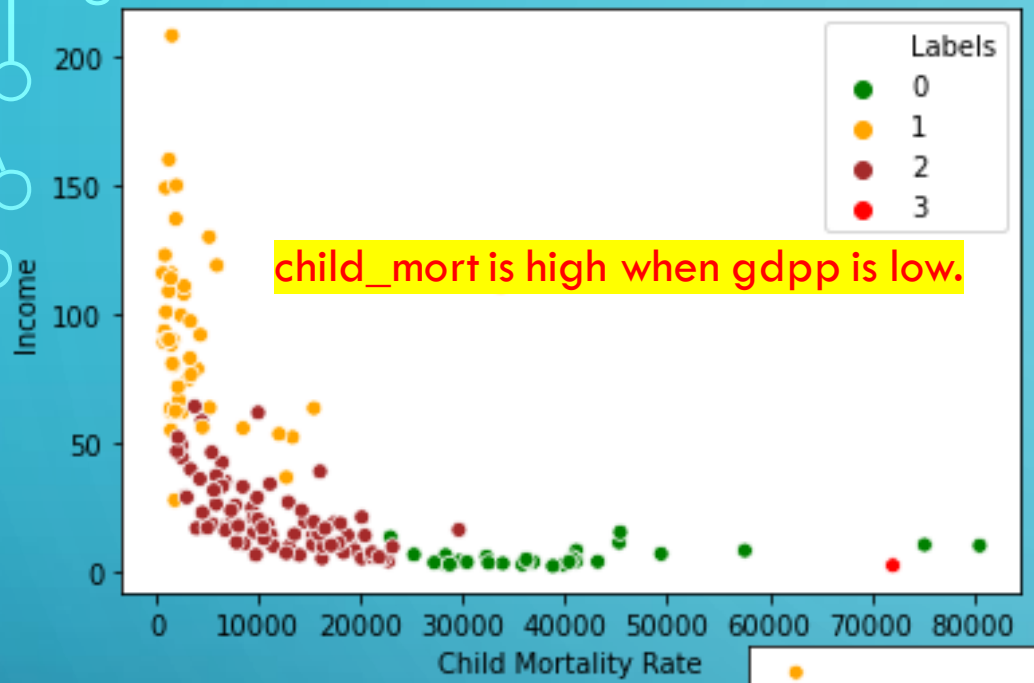
SILHOUETTE SCORE

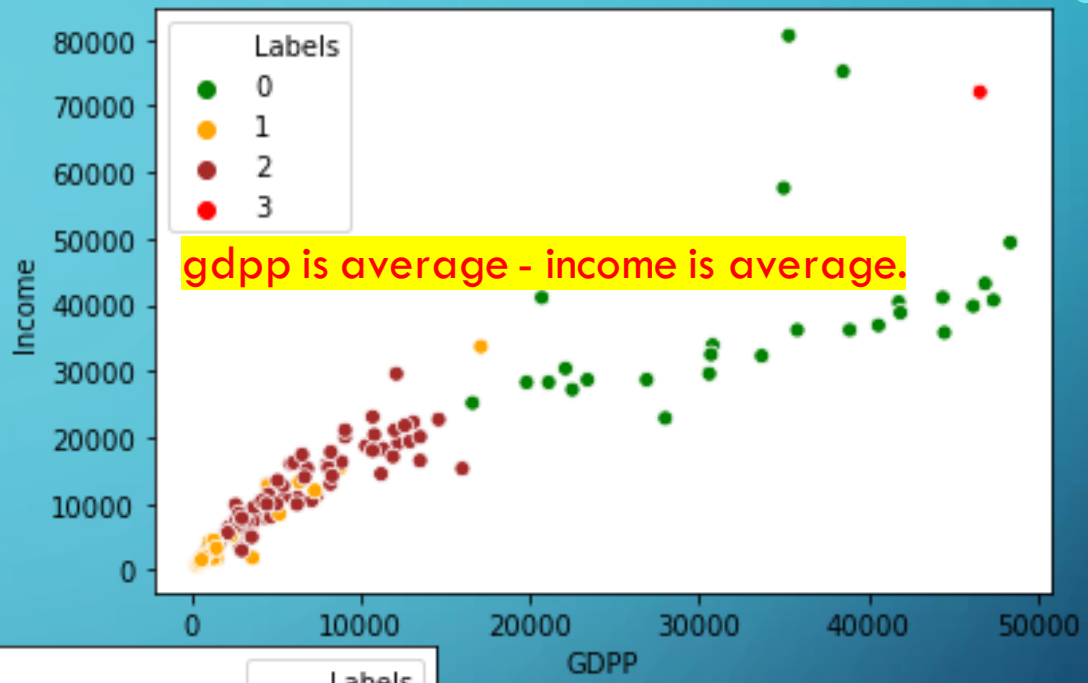


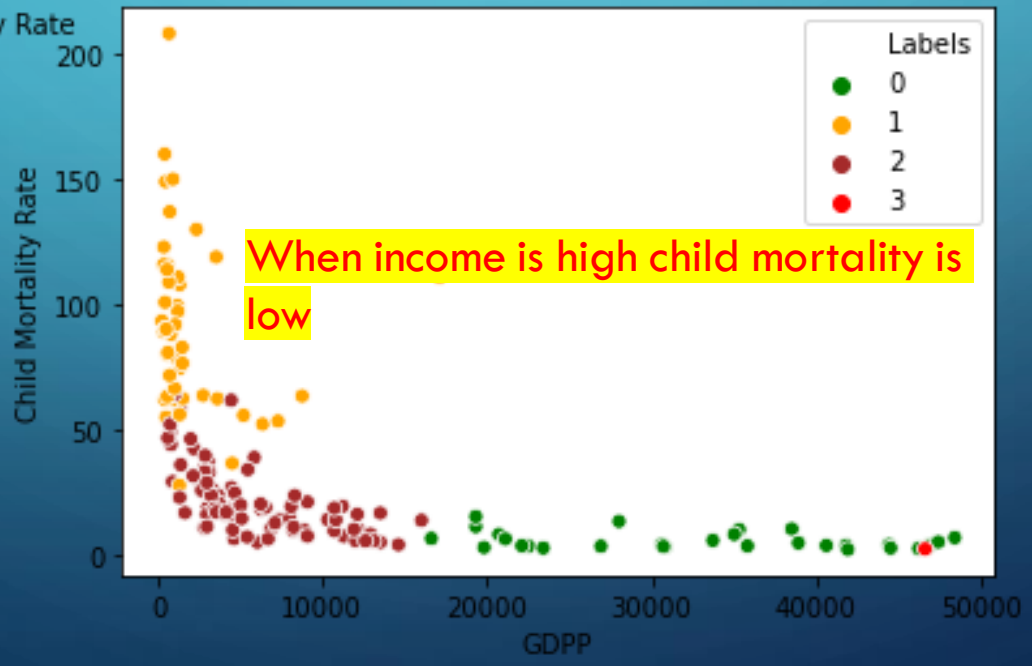
ELBOW SCORE

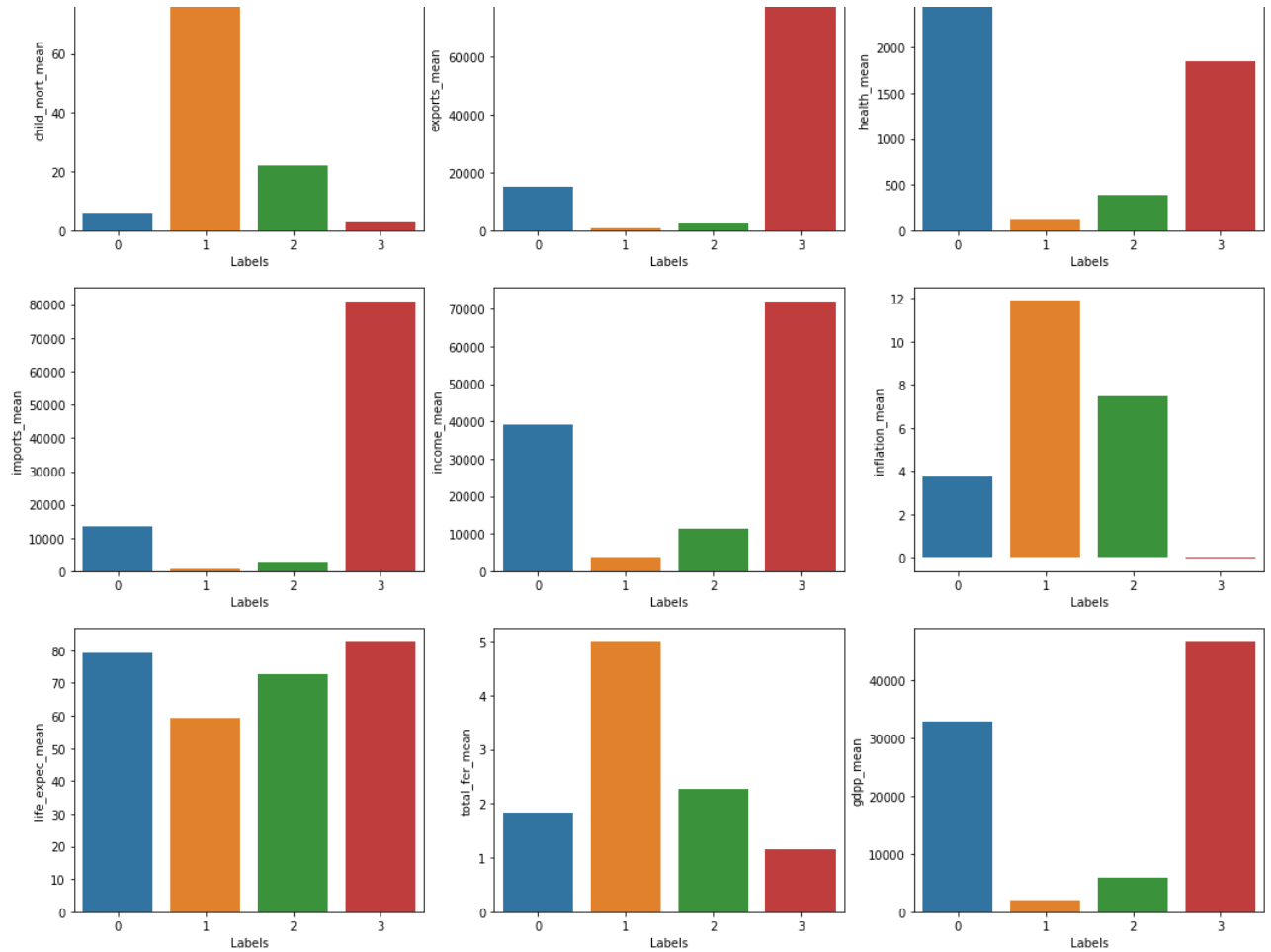


K MEANS CLUSTER PROFILING - CLUSTER LABEL 1









As per our k-means analysis cluster 1 is our area of concern due to below reasons:

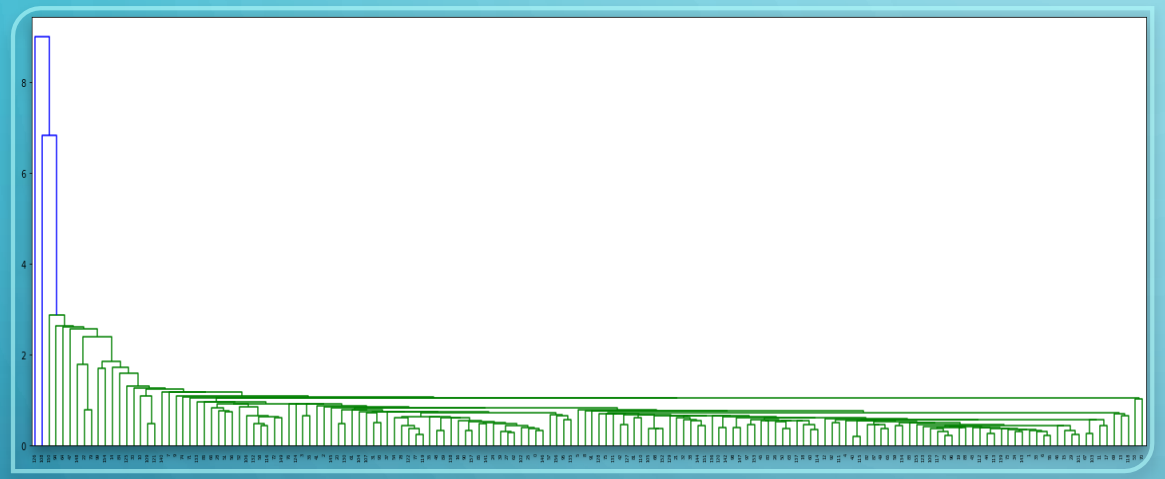
1. low gdpp
2. low income
3. high child mortality
4. high inflation
5. high total fertility

K MEANS RESULTS - TOP 5 COUNTRIES NEEDING NGO'S AID

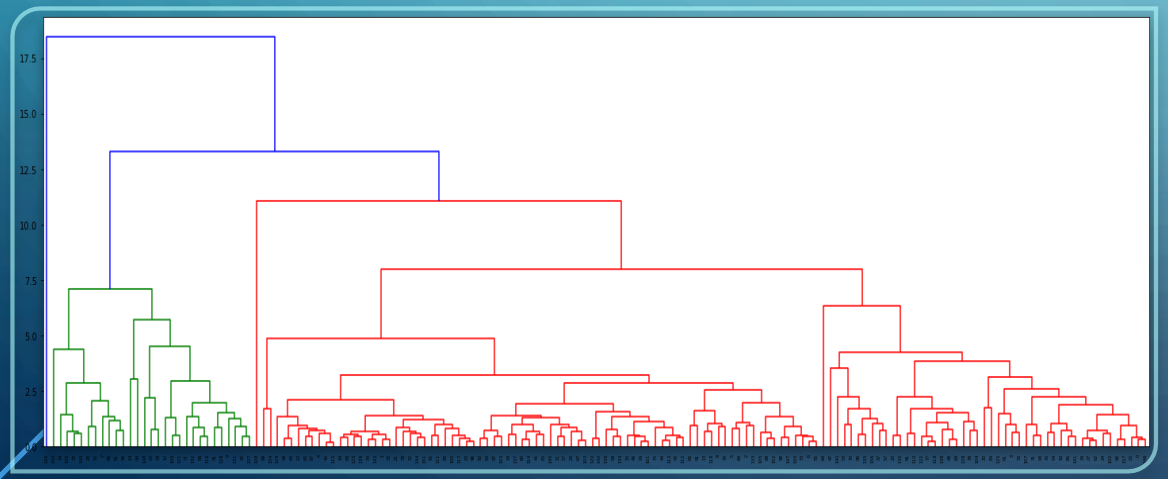
- Burundi
- Liberia
- Congo, Dem. Rep
- Niger
- Sierra Leone

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399

HIERARCHICAL CLUSTERING ANALYSIS



- Single method clustering did not create well defined clusters.

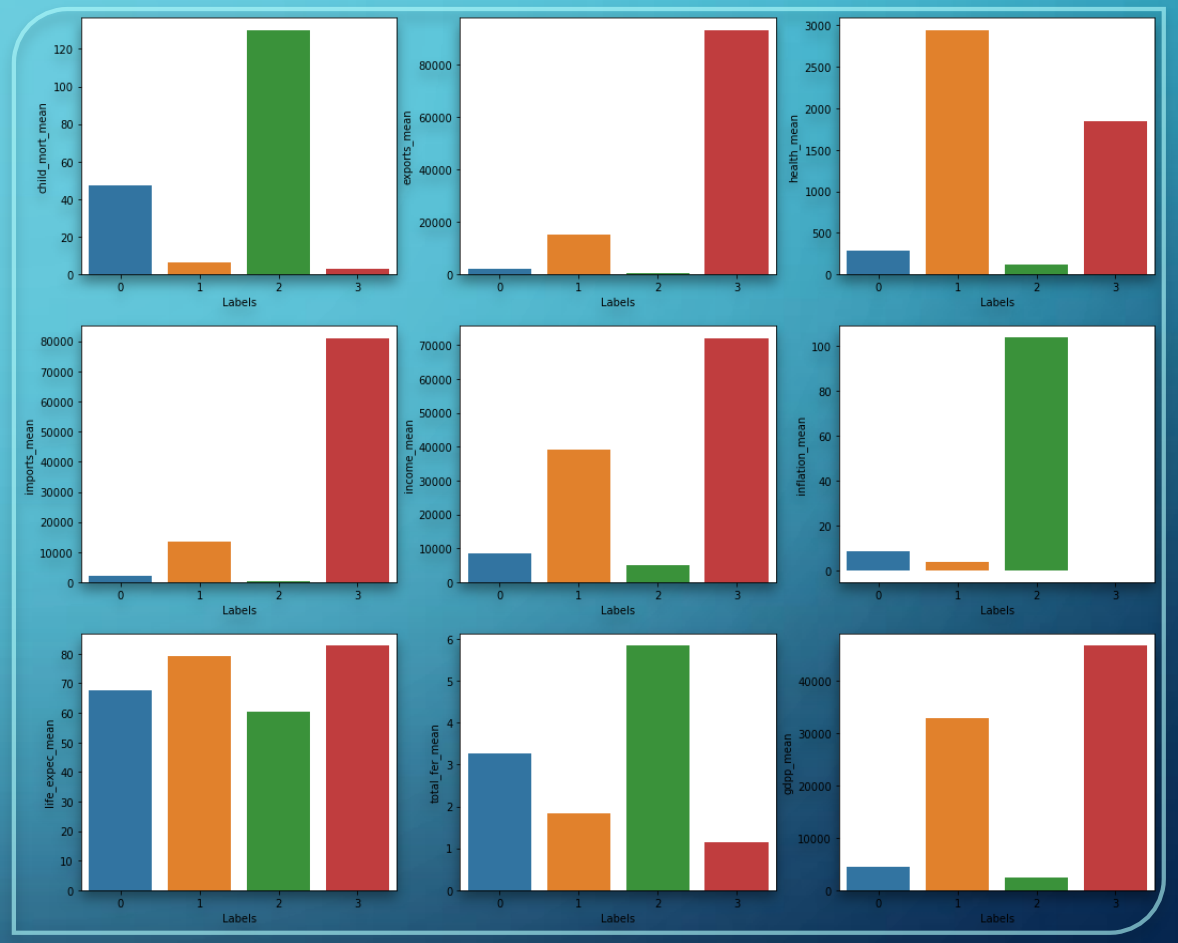


- So we went for a ‘complete’ method dendogram. By looking at it, we decided to cut the dendogram to form 4 clusters

HIERARCHICAL RESULTS - TOP 5 COUNTRIES NEEDING NGO'S AID

- As per our hierarchical clustering analysis cluster 0 is our area of concern due to below reasons:

1. low gdpp
2. low income
3. high child mortality
4. high inflation
5. high total fertility



HIERARCHICAL CLUSTERING RESULTS - TOP 5 COUNTRIES NEEDING NGO'S AID

- Burundi
- Liberia
- Congo, Dem. Rep
- Niger
- Sierra Leone

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399

SUMMARY

Since our objective was to assist the NGO in deciding which countries needed their aid the most, I took gdpp, income and child mortality (as mentioned in the question), in to account while doing the cluster profiling.

After comparing the results of KMeans and Hierarchical clustering methods, I found that the final result obtained by both the methods are same.

The countries were grouped into 4 clusters by taking the socio-economic and health factors, we can determine the over all development of the countries. In the developed countries the GDP per capita is high whereas child mortality rate is low. In the under developed countries, it was vice versa.

The outcomes of the clustering exercise and my analysis are the following:

- In countries like, Haiti, Sierra Leone, Chad etc, the death rate of children below 5 years of age per 1000(child_mort) id high.
- Countries like Burundi, Congo, Niger, etc., GDP per capita income is very low which in turn affects the income per person which is low as well for these poor countries.
- The NGO should focus on these countries by spending on their health infrastructure and economic conditions by giving them the proper aid.