Clustering Assignment

(Using K means & Hierarchical clustering method)

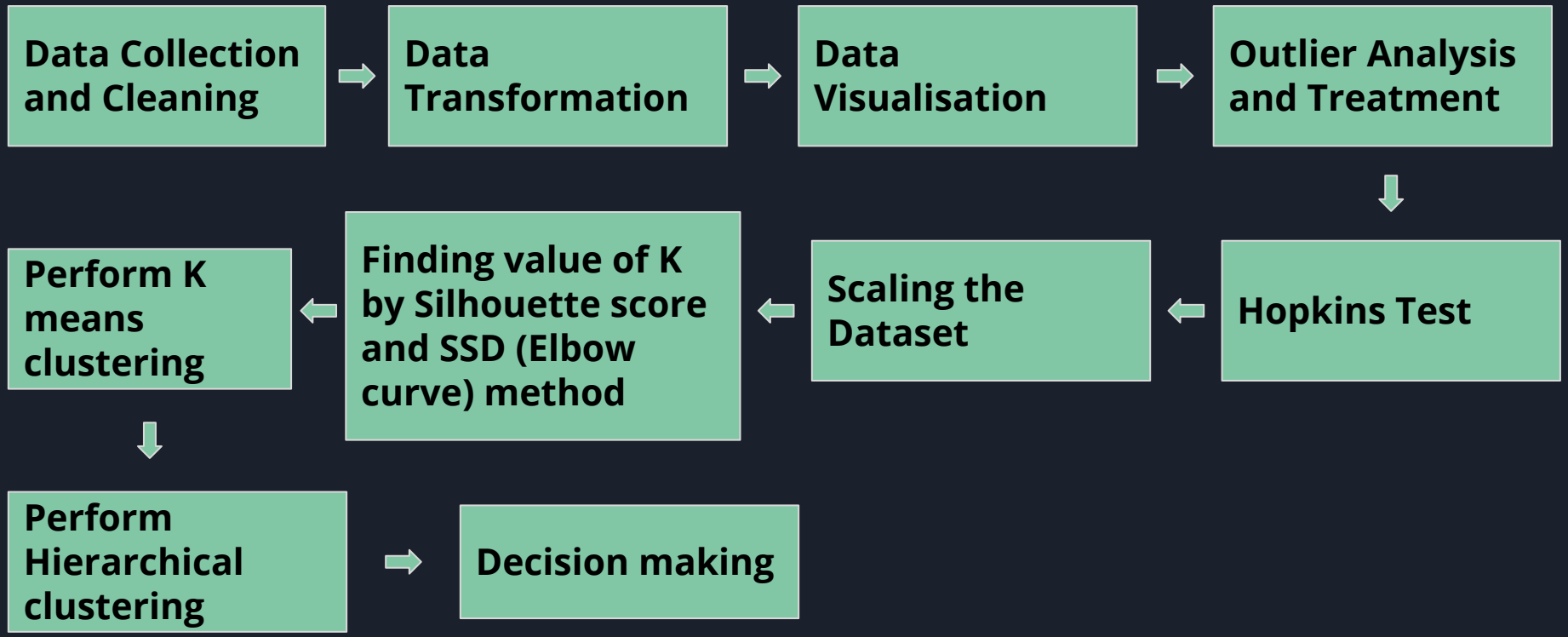
By: Shweta Patil

Problem statement:

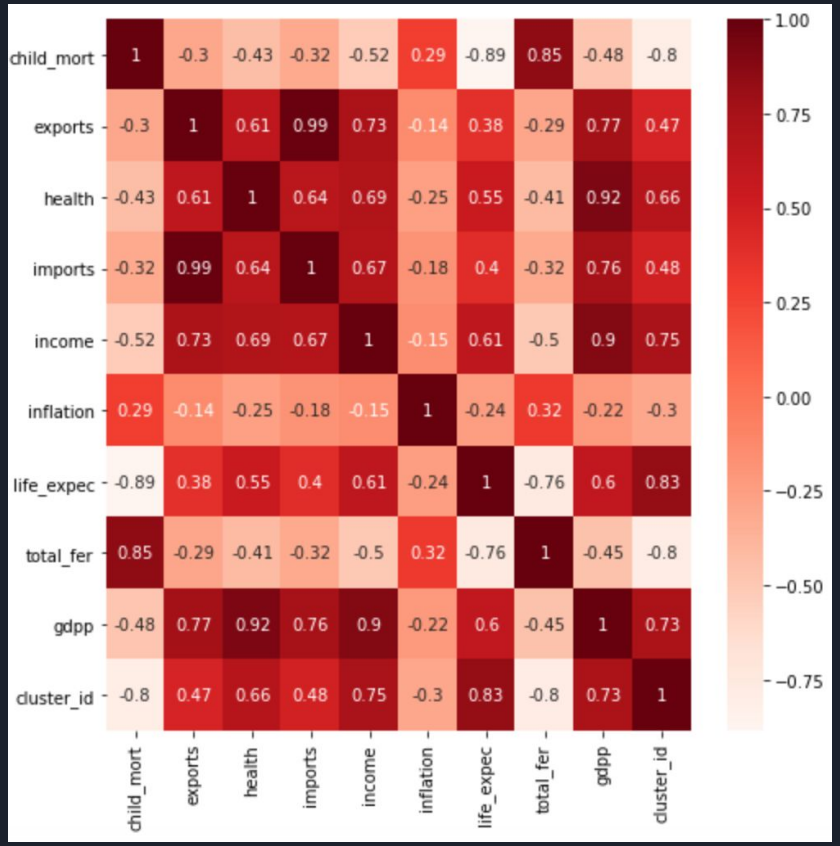
HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

During the recent fund raising program, $10 million funding was raised.
As an analyst, we have to come up with the list of countries that are in dire need of aid.

Analysis Approach



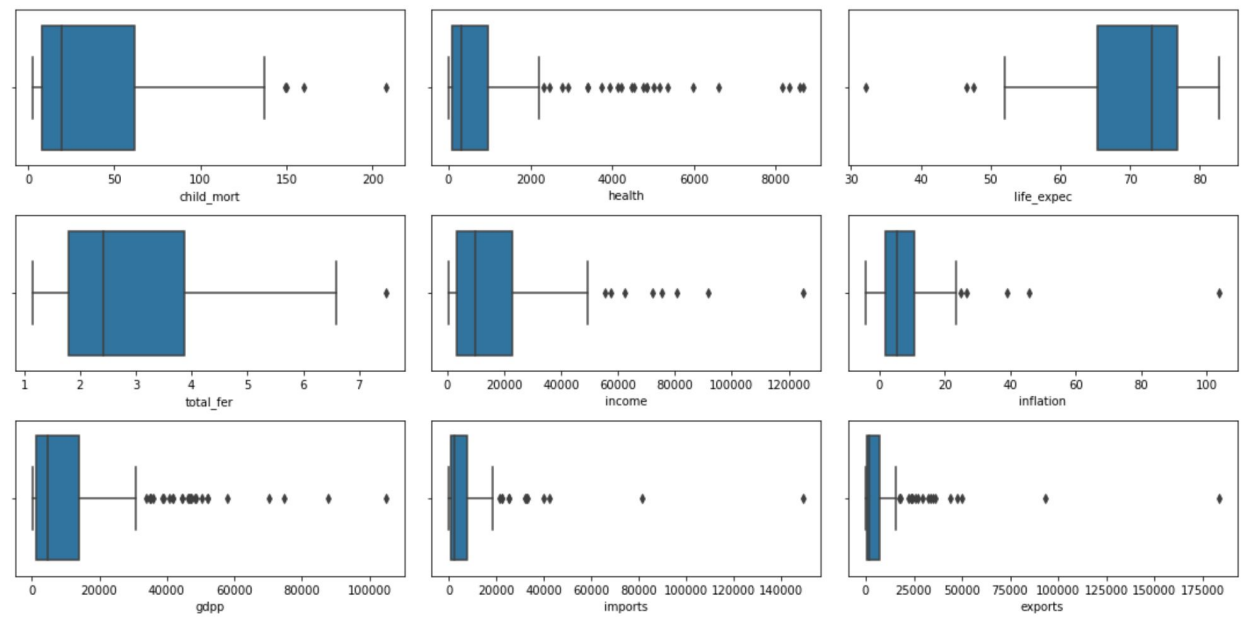
Data Visualisation - Correlation matrix



We can observe high correlation between the following pairs:

1. child_mort and total_fer
2. imports and exports
3. gdpp and health
4. income and gdpp

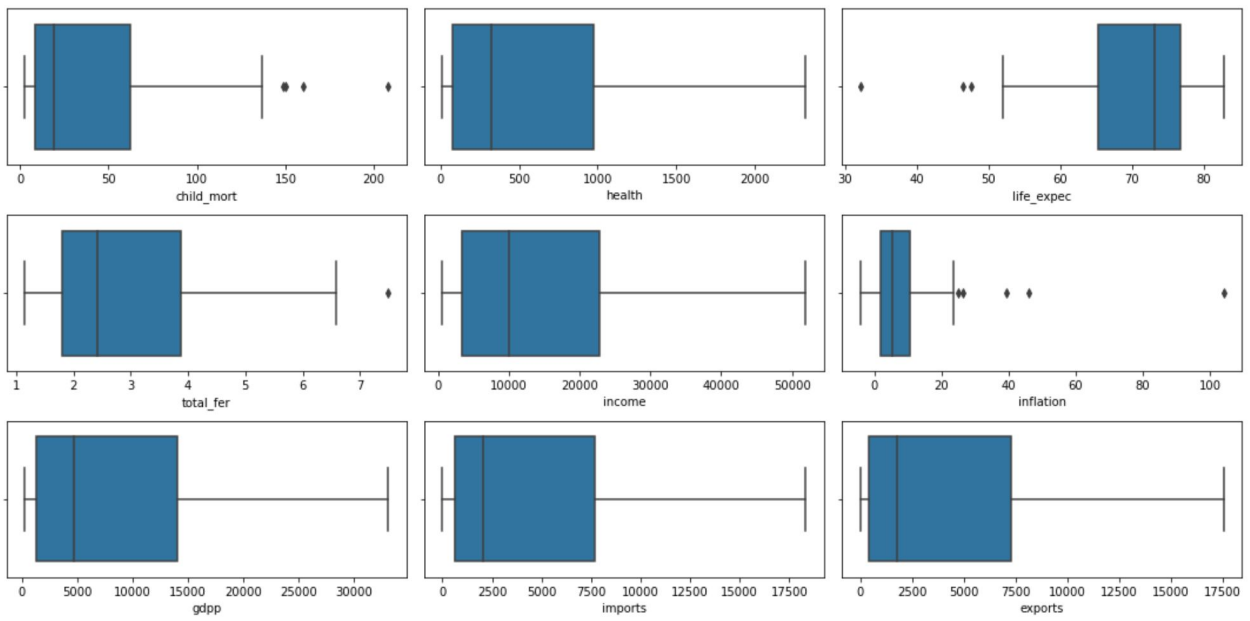
Data Visualisation - Outlier Analysis



We can see that all the columns are having Outliers.

We cannot exclude these outliers as we have only 167 samples with us and excluding outliers can lead to loss of Data.

Data Visualisation - Outlier Treatment Result

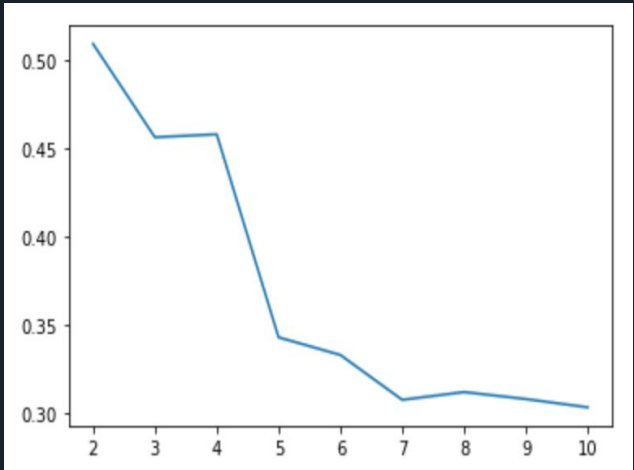


1. For columns such as child_mort, inflation, total_fer we should not do anything to the upper range outliers but deal with the lower range outlier through capping.

2. For the rest of the columns, we should not do anything for the lower range outliers but deal with the upper range outliers through capping.

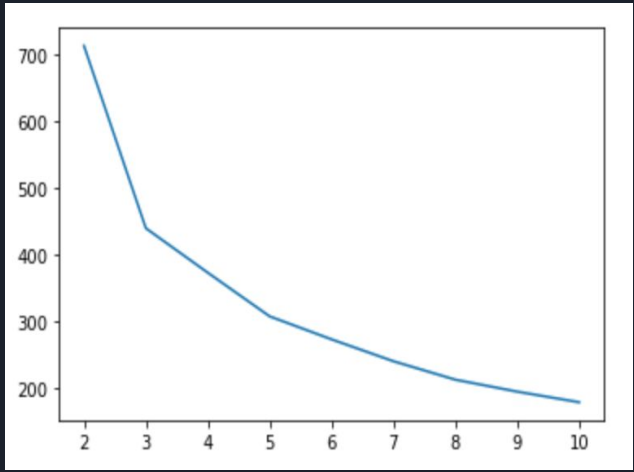
Finding optimal value of K

Method 1: Silhouette Score



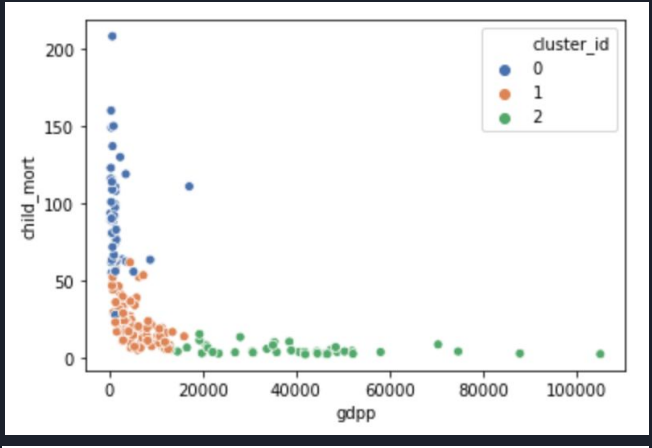
We look at the maximum value in case of Silhouette Score. From the graph we can infer that ideal number of cluster value can be 3 or 4.

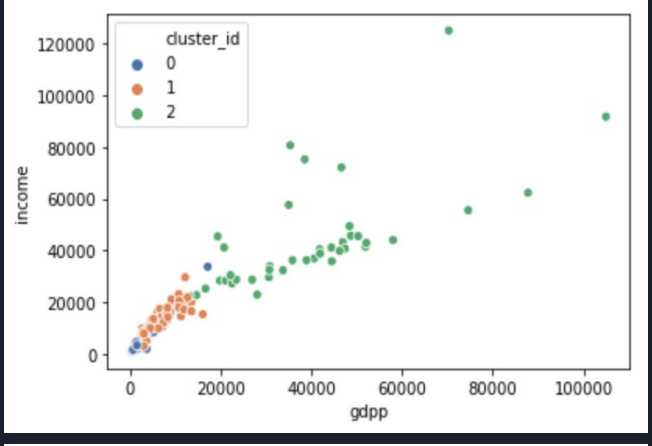
Method 2: SSD (Elbow curve)

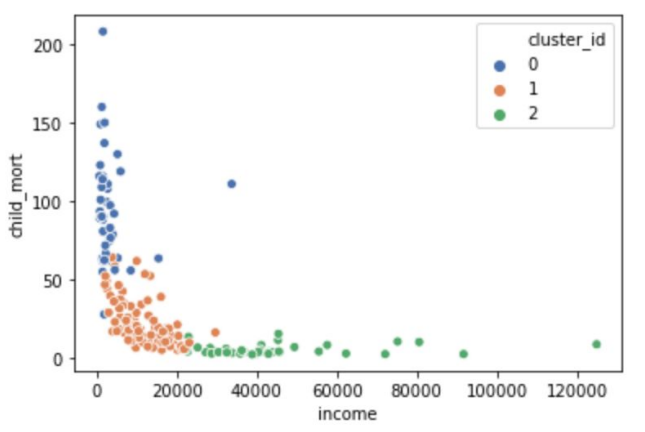


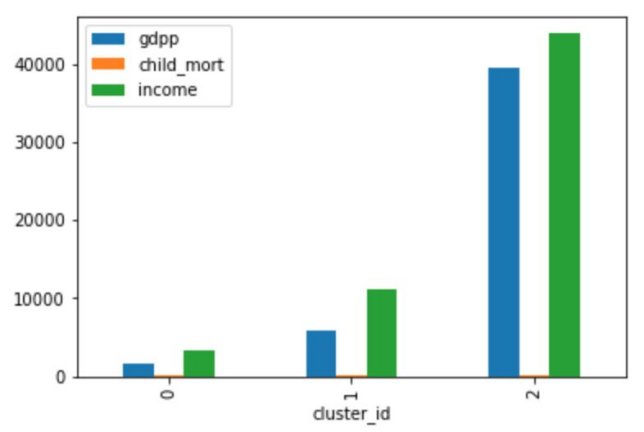
We can observe that the elbow is formed between 3 to 5 and hence value between 3-5 will be suitable for cluster formation.

K Means clustering with K = 3









We can observe that cluster Id 0 has low GDPP, low Income and high Child mortality rate and hence cluster 0 contains data points of our interest (Countries that are in urgent need of aid)

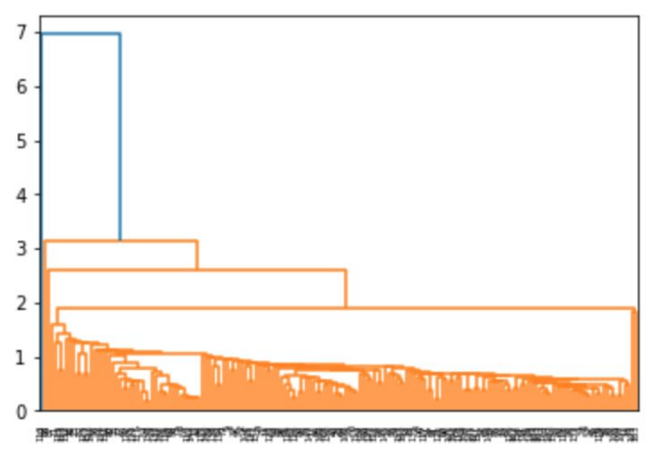
K Means clustering results

Top 10 countries that has Low GDPP, Low INCOME and High CHILD_MORT

	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	12.30	57.7	6.26	231	0
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	0
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	0
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399	0
93	Madagascar	62.2	103.2500	15.5701	177.590	1390	8.79	60.8	4.60	413	0
106	Mozambique	101.0	131.9850	21.8299	193.578	918	7.64	54.5	5.56	419	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.01	47.5	5.21	446	0
94	Malawi	90.5	104.6520	30.2481	160.191	1030	12.10	53.1	5.31	459	0
50	Eritrea	55.2	23.0878	12.8212	112.306	1420	11.60	61.7	4.61	482	0

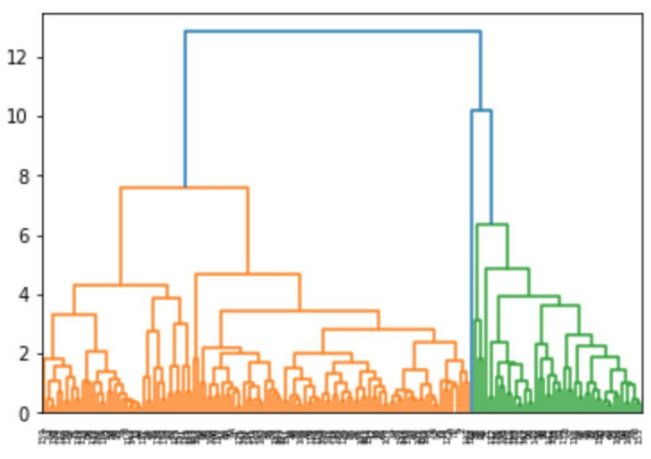
Hierarchical clustering

Method 1: Single Linkage



Hierarchy is not clear using Single Linkage Method

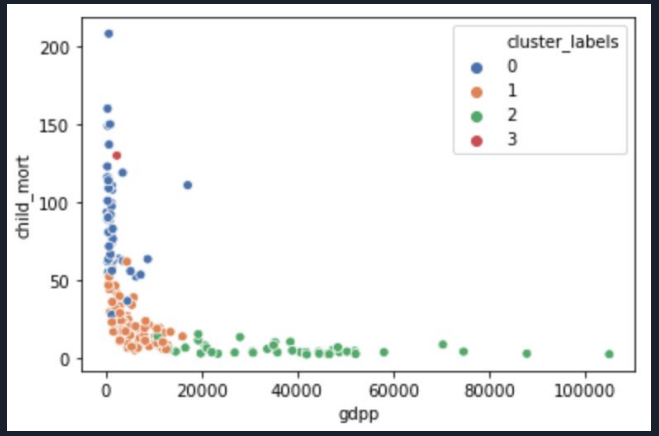
Method 2: Complete Linkage

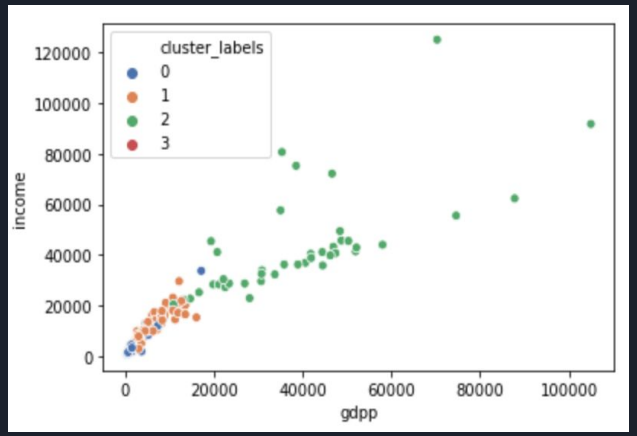


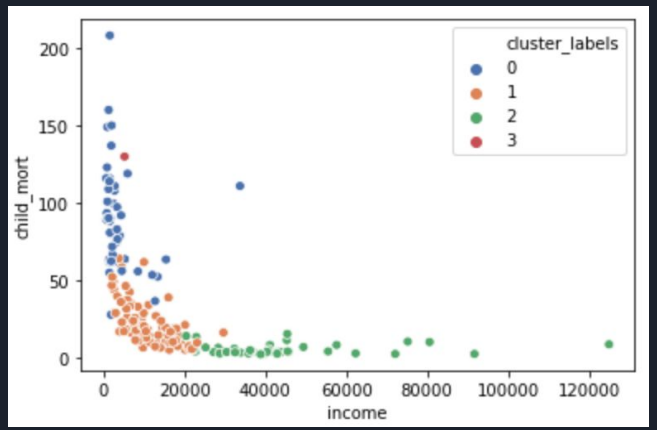
Hierarchy is clear using Complete Linkage Method.

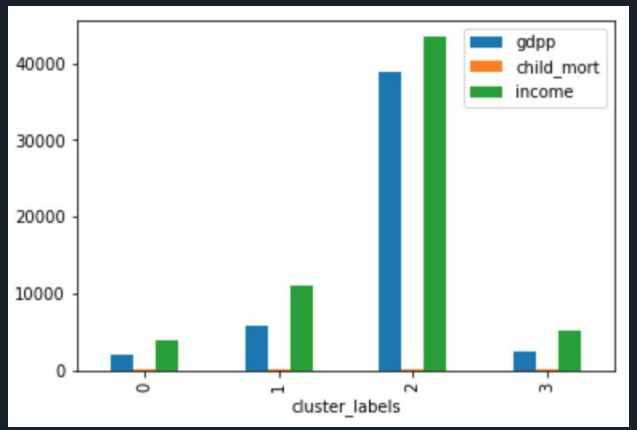
We can form 3 or 4 clusters for analysis.

Hierarchical clustering with K = 4









We can observe that cluster Id 0 has low GDPP, low Income and highest Child mortality rate and hence cluster 0 contains data points of our interest (countries in urgent need of aid)

Hierarchical clustering results

Top 10 countries that has Low GDPP, Low INCOME and High CHILD_MORT

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	0
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	0
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	0
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399	0
93	Madagascar	62.2	103.2500	15.5701	177.590	1390	8.79	60.8	4.60	413	0
106	Mozambique	101.0	131.9850	21.8299	193.578	918	7.64	54.5	5.56	419	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.01	47.5	5.21	446	0
94	Malawi	90.5	104.6520	30.2481	160.191	1030	12.10	53.1	5.31	459	0
50	Eritrea	55.2	23.0878	12.8212	112.306	1420	11.60	61.7	4.61	482	0

Conclusion

We performed analysis on the Dataset using K means clustering and Hierarchical clustering.

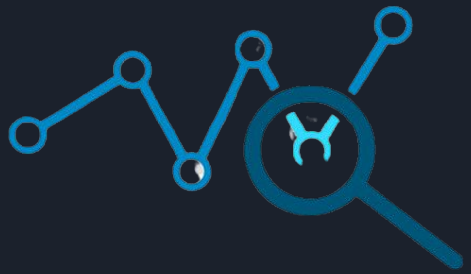
Both the approaches are giving same results and hence we can say that following countries are in urgent need of aid:

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

“Information is the oil of the 21st century, and analytics is the combustion engine.”

- Peter Sondergaard

Thanks!




PLAN
Business
PUZZLE
Marketing
Process
SUCCESS
Problems
GROW
Team
50,000.00