Clustering of countries

By : Siddharth Singh

Objective:

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.

Problem Statement:

During the recent funding programs, NGO 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 Approach:

- **Data Collection And Data Cleaning:**

Importing the data then cleaning it , checking if there are any null values. We found out that some columns were in %of gdpp form so corrected them using the correct formulas.

- **Visualising data:**

We detected outliers by visualising the data , outliers were treated according to our problem statement.We also found out that some Variables are highly corelated to each other.

- **Outliers Detection and treatment:**

There were outliers in almost every columns. Some outliers like in gdpp column for example, have outliers on high end of spectrum which we can remove safely because high gdpp countries wont need urgent aid. For this analysis we capped the gdpp column at .95 quantile at upper end and .1 quantile at lower end. We did not capped other variable because that would effect our analysis for finding poorest countries with high child mortality.

- **Scaling data:**

Standardizing all the continuous variables.

- **Hopkins test:**

To check if data has tendency to form clusters.

- **Kmeans Clustering:**

Identifying the “k” through silhouette analysis and elbow curve. Then forming the cluster on scaled data the adding the cluster id on original data for better interpretation of data. And visualizing the clusters.

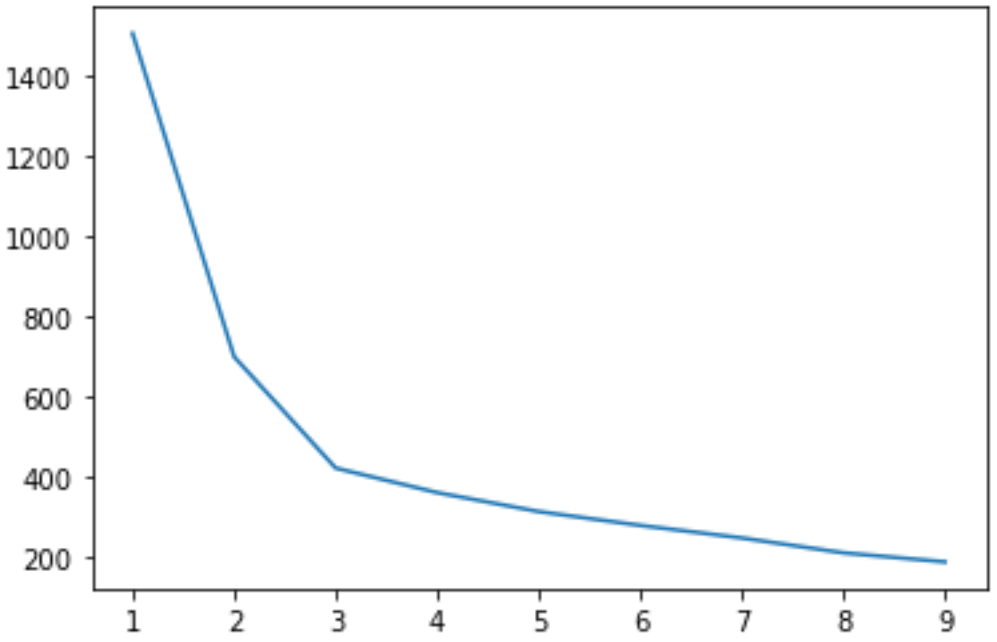
- **Hierarchical Clustering**

Identifying optimal number for k by analysing dendrogram. Then forming the cluster on scaled data and adding the cluster label to original data for better interpretation. Visualisation of clusters was also done.

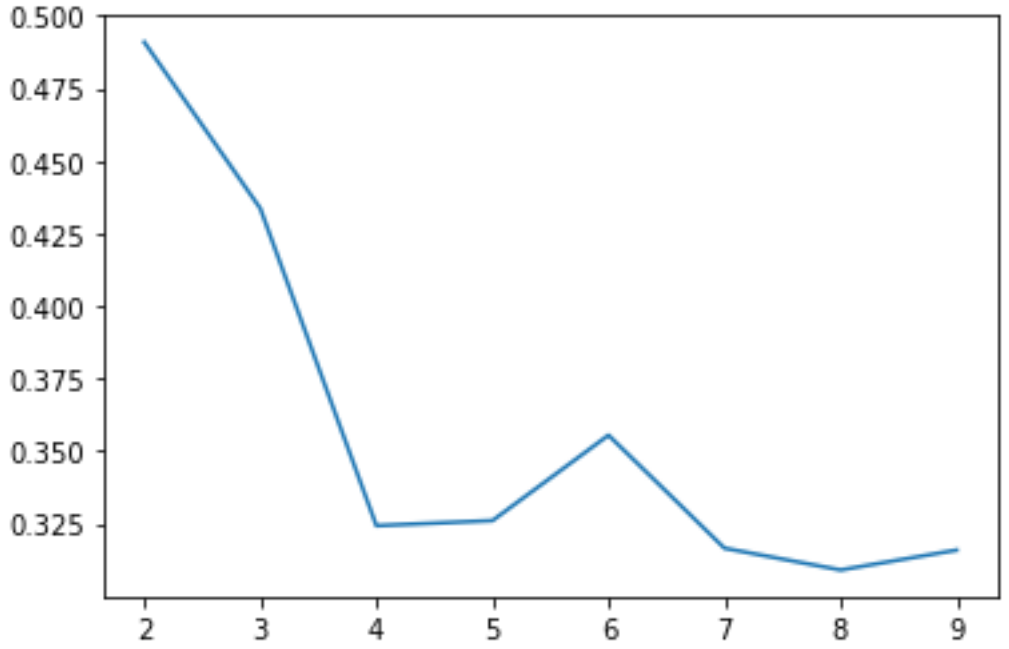
- **Decision Making:**

Successfully identified the top 10 countries by analysing both model which are in dire need of Aid.

Kmeans Clustering

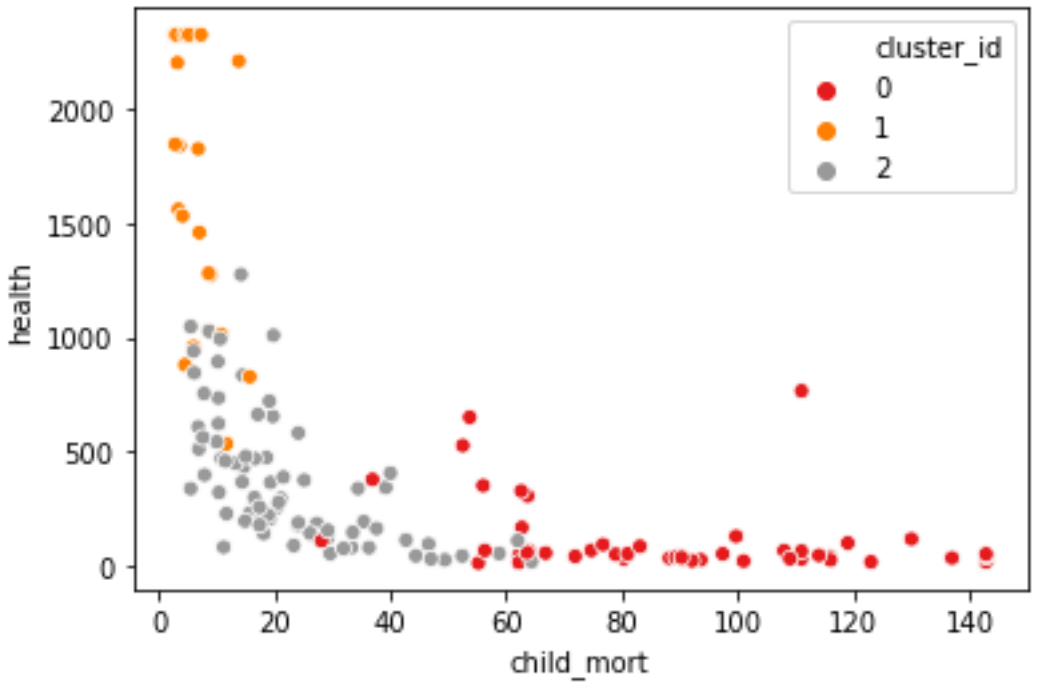


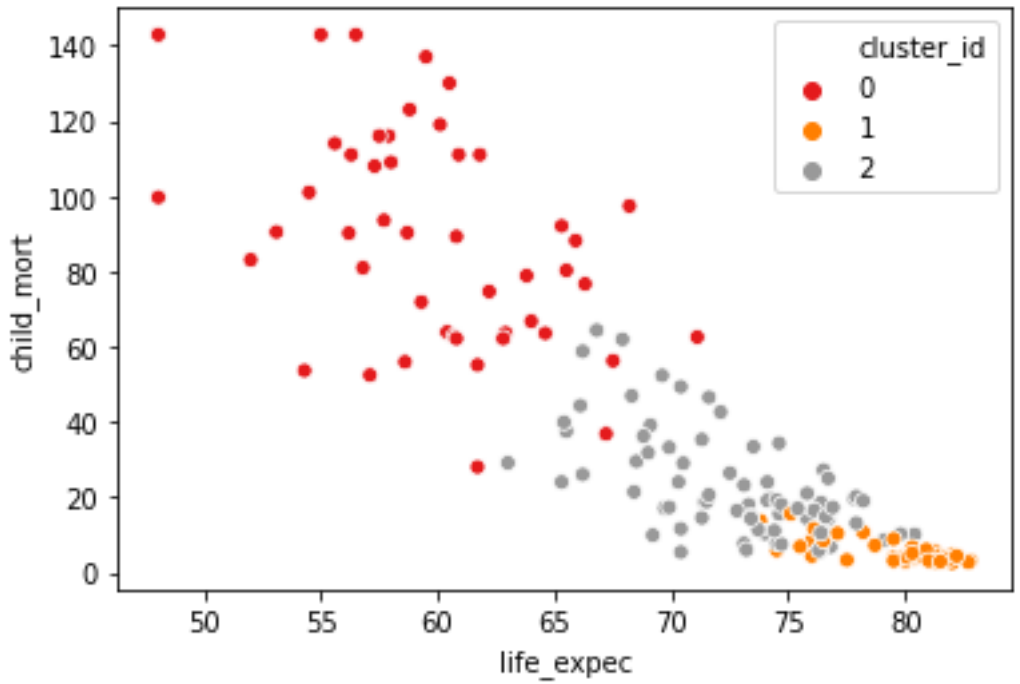
Elbow Curve



Silhouette Analysis

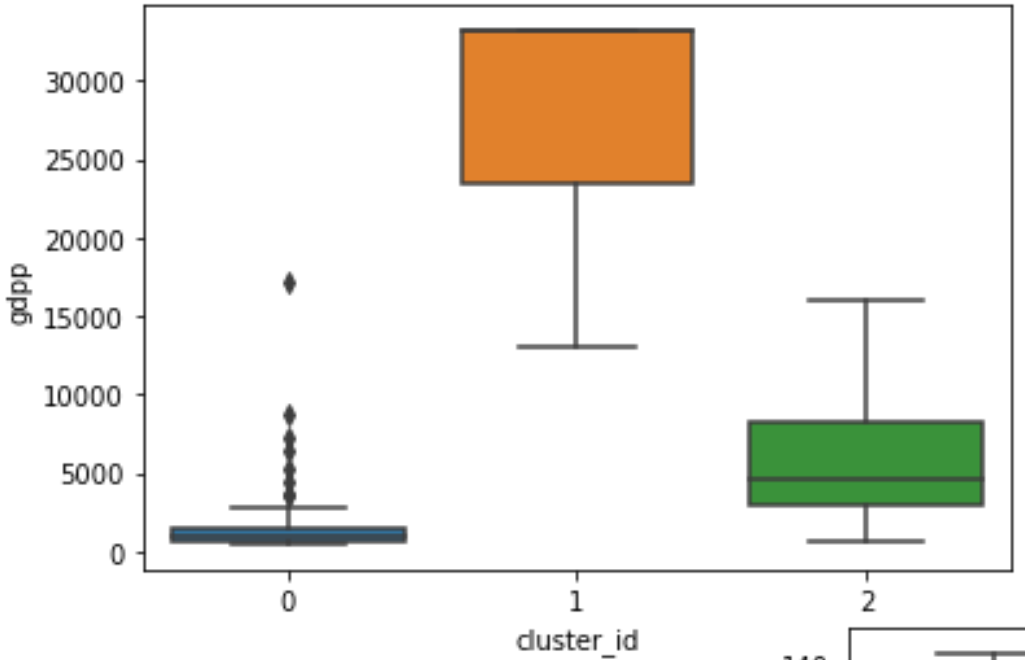
We can see in silhouette analysis that highest peak is at 2, but 2 is never a good number for clustering, on the other hand elbow curve has elbow at 3. So we will go with k=3.

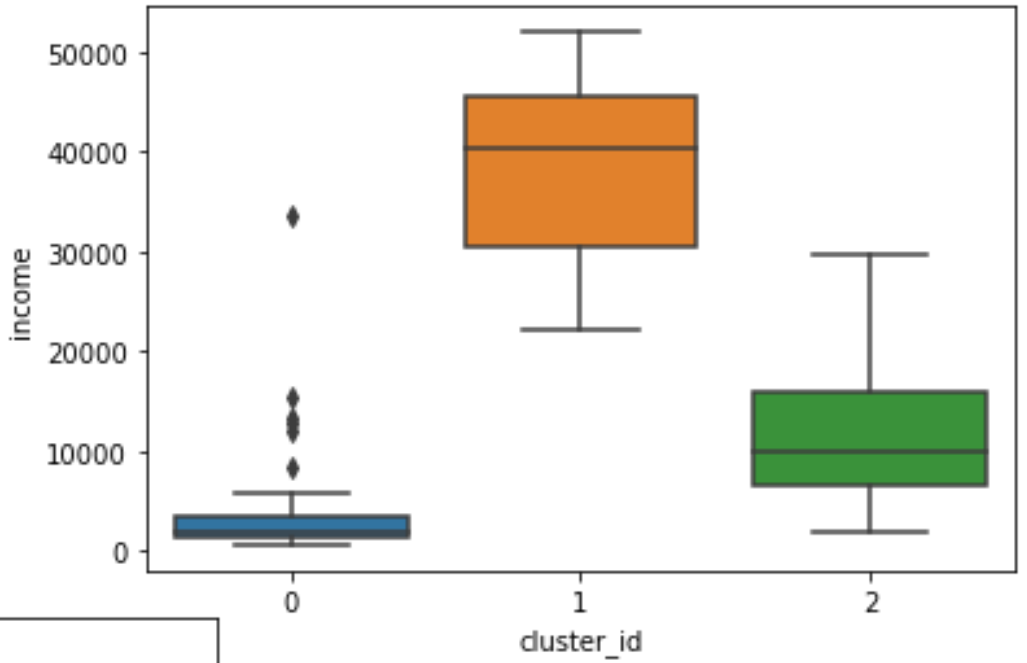


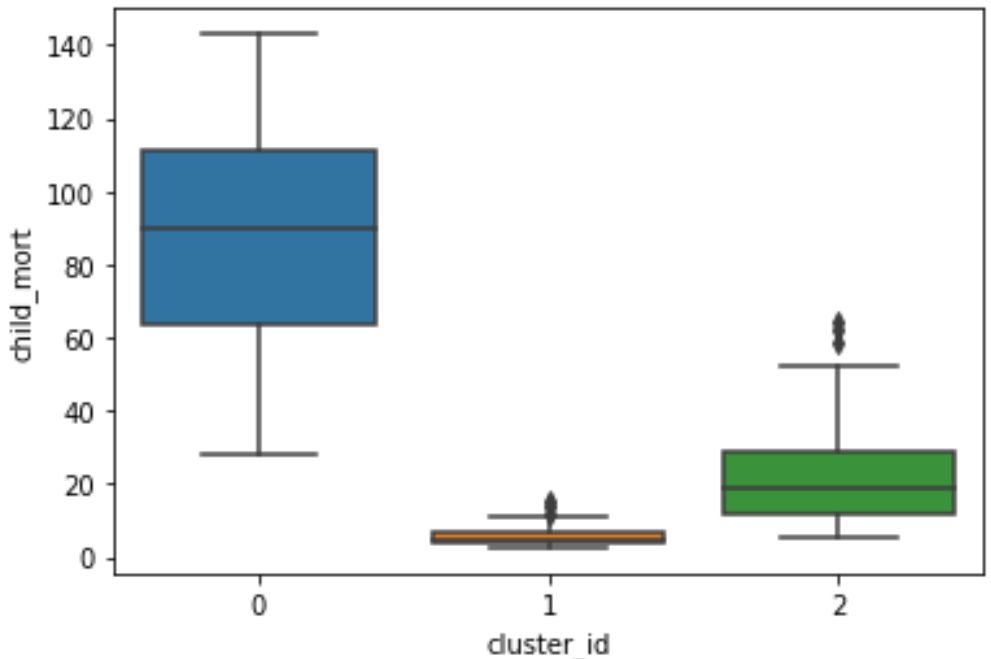


For Cluster 0 ,health expenditure is very low and child mortality is very high.

No Surprise that for cluster 0 child mortality is high and life expectancy is low.





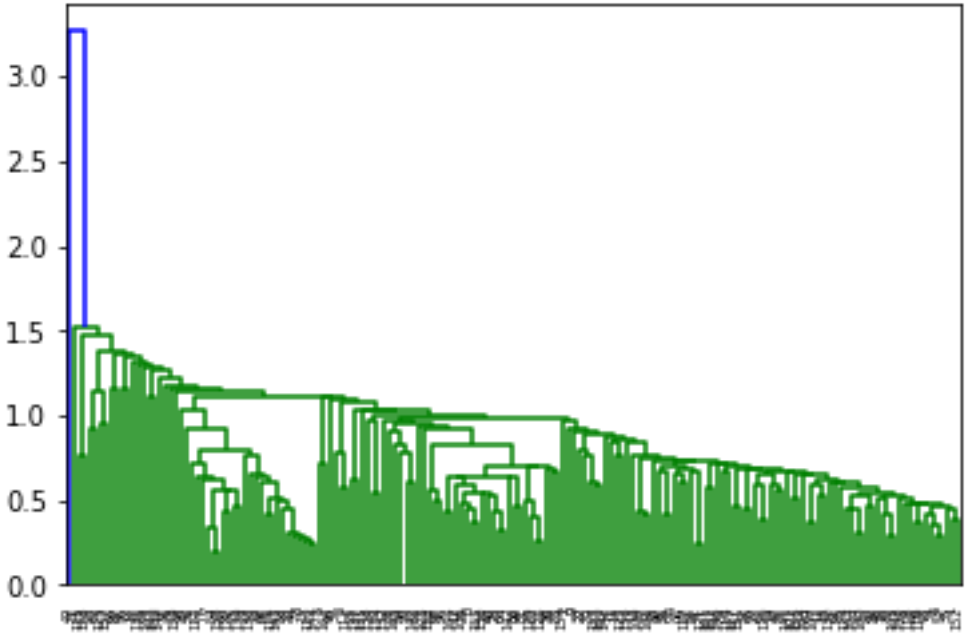


For cluster id 0 both gdpp and income are very low and child mortality is very high.
So this cluster is our main focus for providing aid to countries.

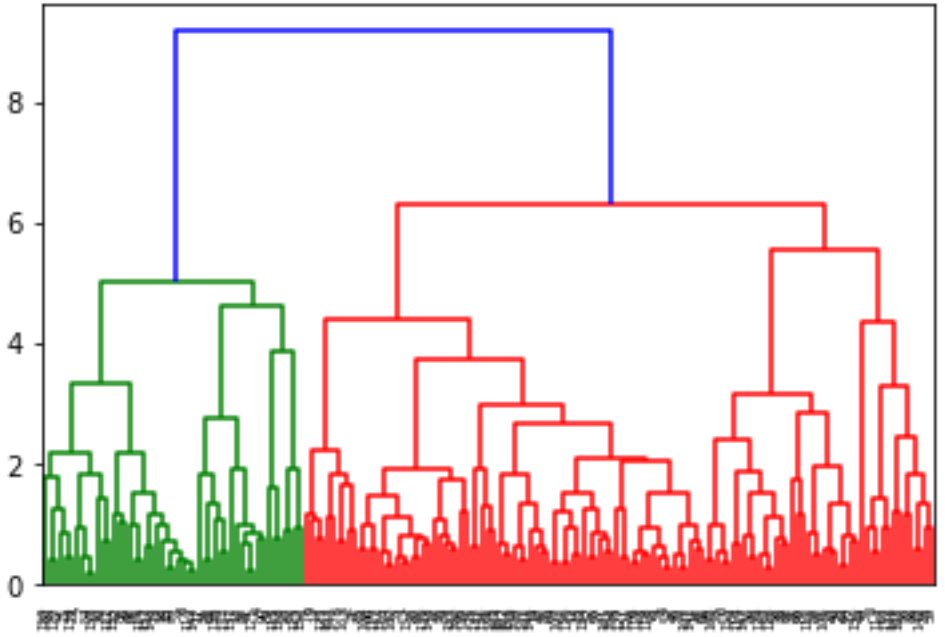
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	Sierra Leone	142.875	67.032	52.2690	137.655	1220.0	17.20	55.00	5.2000	399.0	0
1	Central African Republic	142.875	52.628	17.7508	118.190	888.0	2.01	48.05	5.2100	446.0	0
2	Haiti	142.875	101.286	45.7442	428.314	1500.0	5.45	48.05	3.3300	662.0	0
3	Chad	142.875	330.096	40.6341	390.195	1930.0	6.39	56.50	6.5900	897.0	0
4	Mali	137.000	161.424	35.2584	248.508	1870.0	4.37	59.50	6.5500	708.0	0
5	Nigeria	130.000	589.490	118.1310	405.420	5150.0	24.16	60.50	5.8400	2330.0	0
6	Niger	123.000	77.256	17.9568	170.868	814.0	2.55	58.80	7.0075	348.0	0
7	Angola	119.000	2199.190	100.6050	1514.370	5900.0	22.40	60.10	6.1600	3530.0	0
8	Congo, Dem. Rep.	116.000	137.274	26.4194	165.664	609.0	20.80	57.50	6.5400	334.0	0
9	Burkina Faso	116.000	110.400	38.7550	170.200	1430.0	6.81	57.90	5.8700	575.0	0

Top 10 Countries obtained from K-Means Clustering in dire need of Aid, sorted by child mortality as decreasing order and income and Gdpp as increasing order.

Hierarchical Clustering

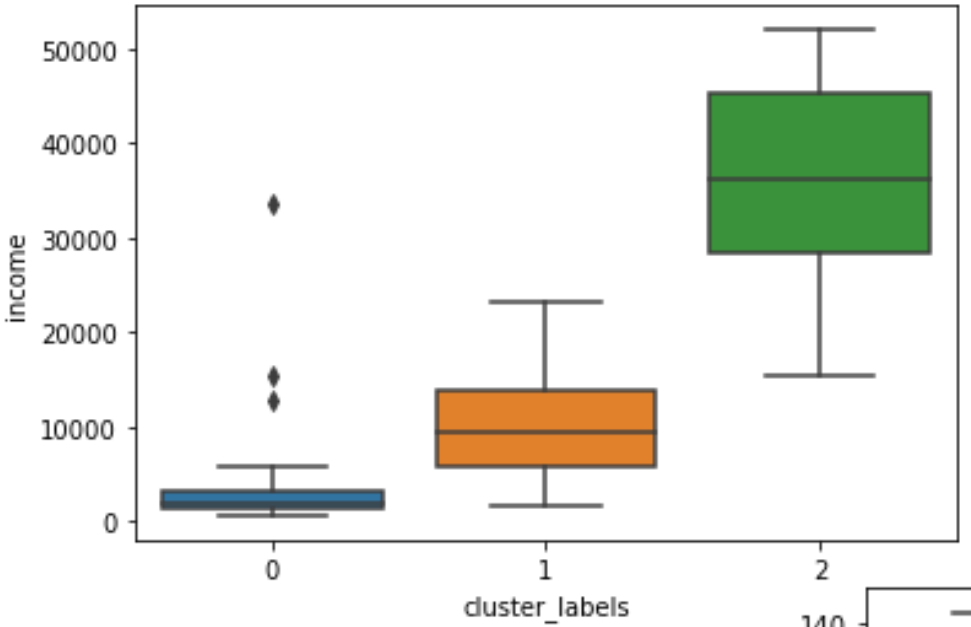


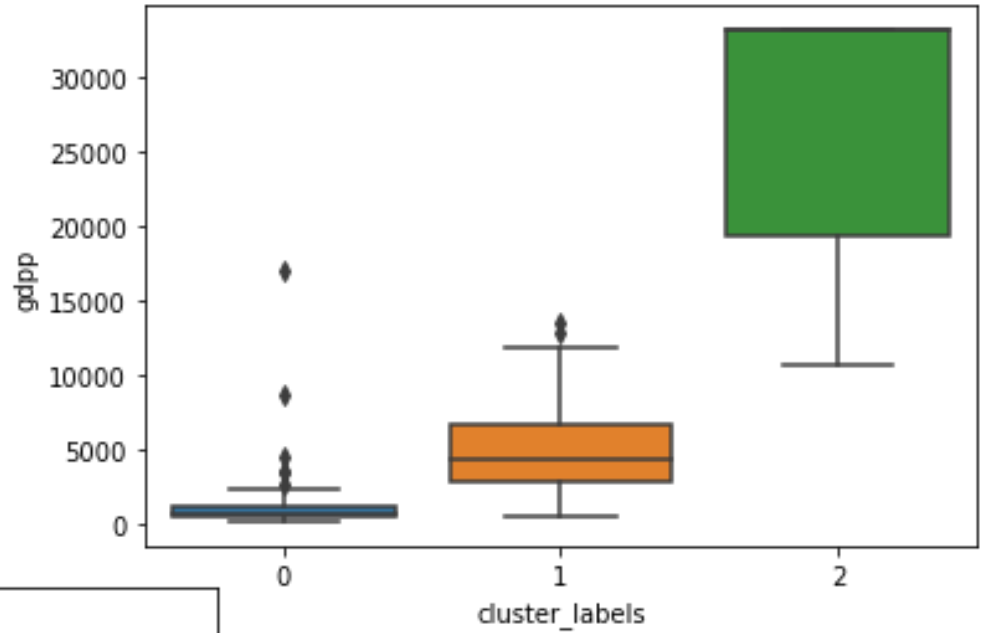
Single linkage Hierarchical clustering

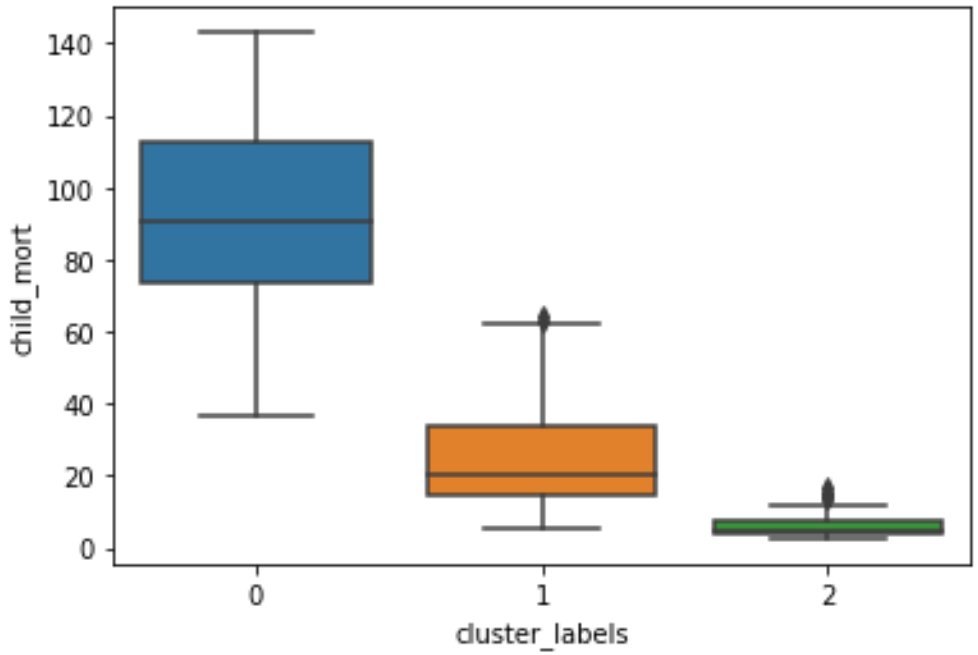


Complete linkage hierarchical clustering

We will go with the complete linkage hierarchical clustering because single linkage is more complex and cluster formation was not good in single linkage.







For Cluster 0 both income and gdpp is vey low and child mortality is very high.
So this cluster will be our focus for countries that are in dire need of aid.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
0	Sierra Leone	142.875	67.032	52.2690	137.655	1220.0	17.20	55.00	5.2000	399.0	0
1	Central African Republic	142.875	52.628	17.7508	118.190	888.0	2.01	48.05	5.2100	446.0	0
2	Haiti	142.875	101.286	45.7442	428.314	1500.0	5.45	48.05	3.3300	662.0	0
3	Chad	142.875	330.096	40.6341	390.195	1930.0	6.39	56.50	6.5900	897.0	0
4	Mali	137.000	161.424	35.2584	248.508	1870.0	4.37	59.50	6.5500	708.0	0
5	Nigeria	130.000	589.490	118.1310	405.420	5150.0	24.16	60.50	5.8400	2330.0	0
6	Niger	123.000	77.256	17.9568	170.868	814.0	2.55	58.80	7.0075	348.0	0
7	Angola	119.000	2199.190	100.6050	1514.370	5900.0	22.40	60.10	6.1600	3530.0	0
8	Congo, Dem. Rep.	116.000	137.274	26.4194	165.664	609.0	20.80	57.50	6.5400	334.0	0
9	Burkina Faso	116.000	110.400	38.7550	170.200	1430.0	6.81	57.90	5.8700	575.0	0

Top 10 countries obtained from hierarchical clustering in dire need, sorted by child mortality in decreasing order and income and gdpp in increasing order.

Summary

We got same top 10 countries from both hierarchical and k-means model that are in dire need of Aid.

Following are the countries name requiring aid.

1. Sierra Leone
2. Central African Republic
3. Haiti
4. Chad
5. Mali
6. Nigeria
7. Niger
8. Angola
9. Congo , Dem. Rep.
10. Burkina Faso