Clustering Assignment

Arushi Shree

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. After the recent funding programmes, they have been able to raise around $ 10 million.

- We as a data analyst have to categorise the countries which are in need of fund using some socio-economic and health factors that determine the overall development of the country. We need to report these countries to CEO of NGO by performing clustering algorithm.

Clustering Approach

Following steps are performed to get required group of countries:

1. **Data Understanding:**

   Importing data, understanding quality and basic summary of data.

2. **Data Cleaning and Visualisation:**

   Data cleaning and visualizing data using heat maps to get correlation,
   Boxplot to understand outliers in data.

3. **Data Preparation:**

   Removing outliers in data, scaling variables for standardization.
   Calculating Hopkins score to identify cluster tendency.

Clustering Approach

4. **Model Building :**

K-Means Clustering:

Understanding optimum number of clusters using two methods

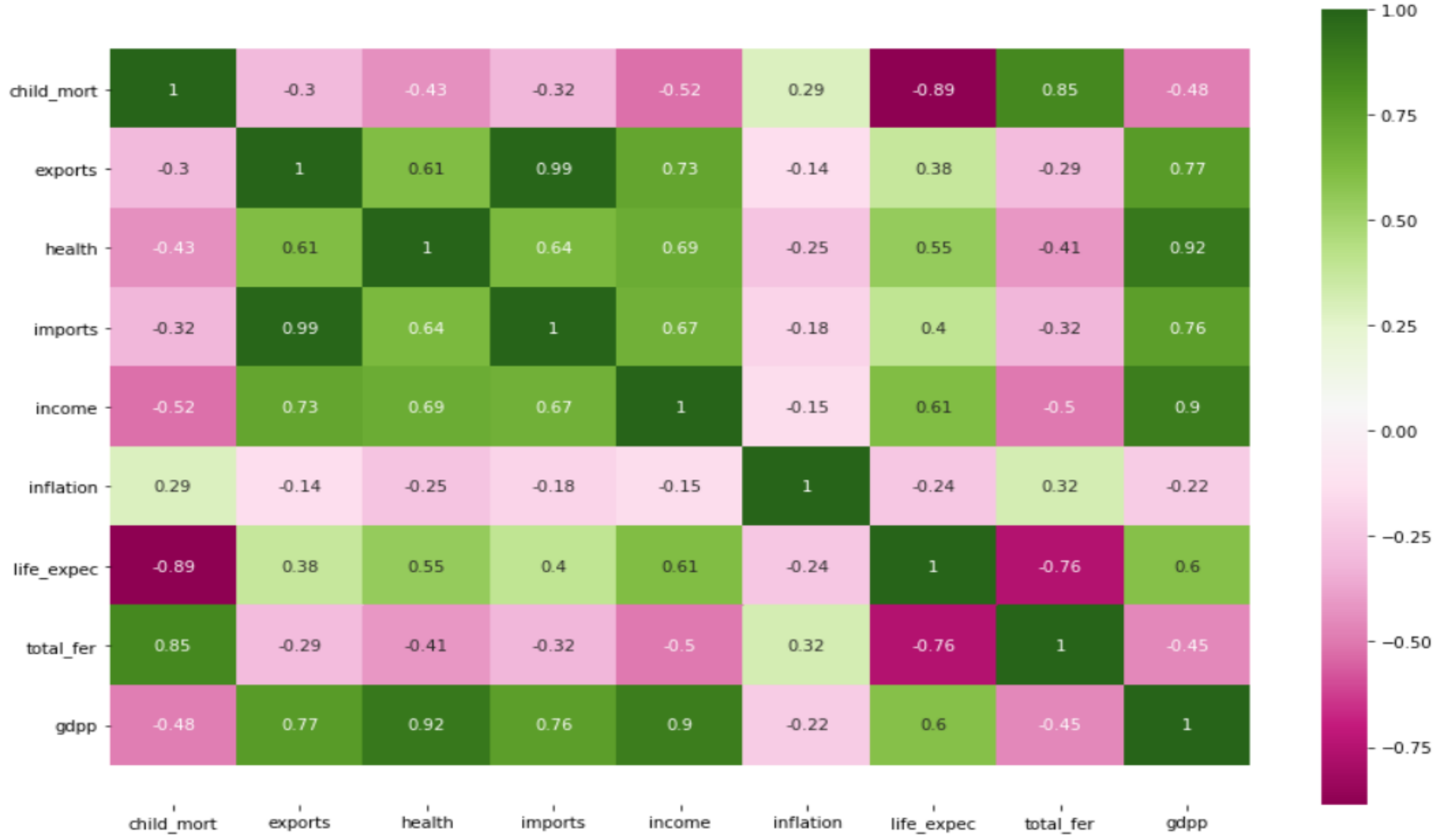
1. Elbow curve method 2. Silhouette score method

Model building using optimum clusters, labeling countries with cluster formed. Analyzing cluster groups by plotting scatter plot and box plot with respect to valuable features. Identifying group of countries through above analysis.

Hierarchical Clustering:

Building model using correct number of clusters through dendrograms. Labeling countries with cluster formed. Analyzing cluster groups by plotting scatter plot and boxplot with respect to valuable features. Identifying group of countries through above analysis.

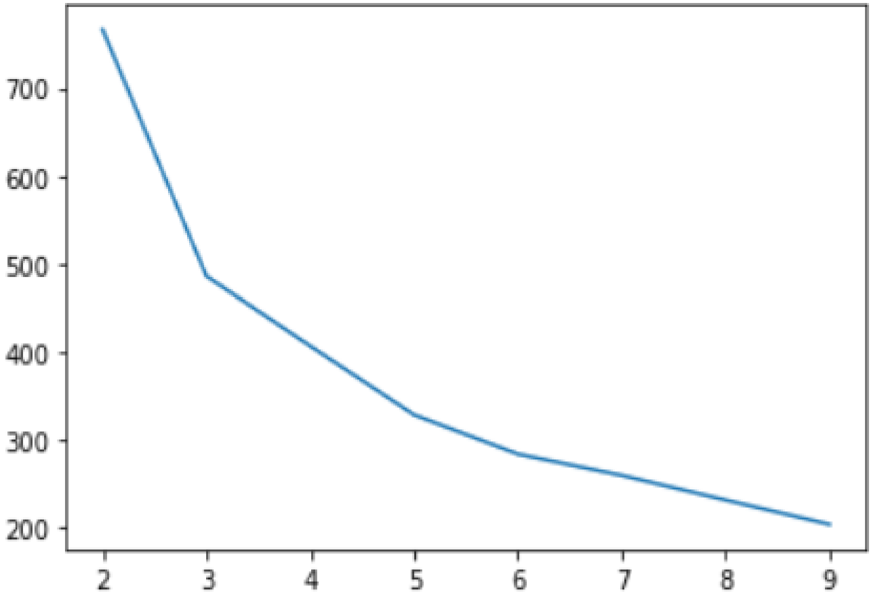
Correlation in the data



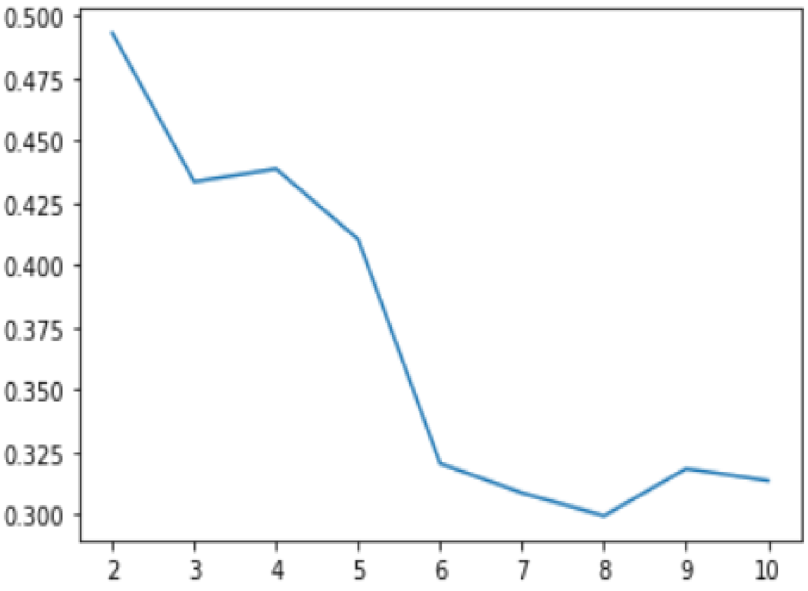
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52	0.73	0.69	0.67	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.14	-0.25	-0.18	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.38	0.55	0.4	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.29	-0.41	-0.32	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.77	0.92	0.76	0.9	-0.22	0.6	-0.45	1

- “Gdpp” is positively correlated with “Income” which is obvious.
- “Child mortality” shows high positive correlation with total fertility.
- “Import” and “Exports” are also positively correlated.
- “Health”, “Income”, “Gdpp” shows high negative correlation with child mortality.

K –Means Clustering



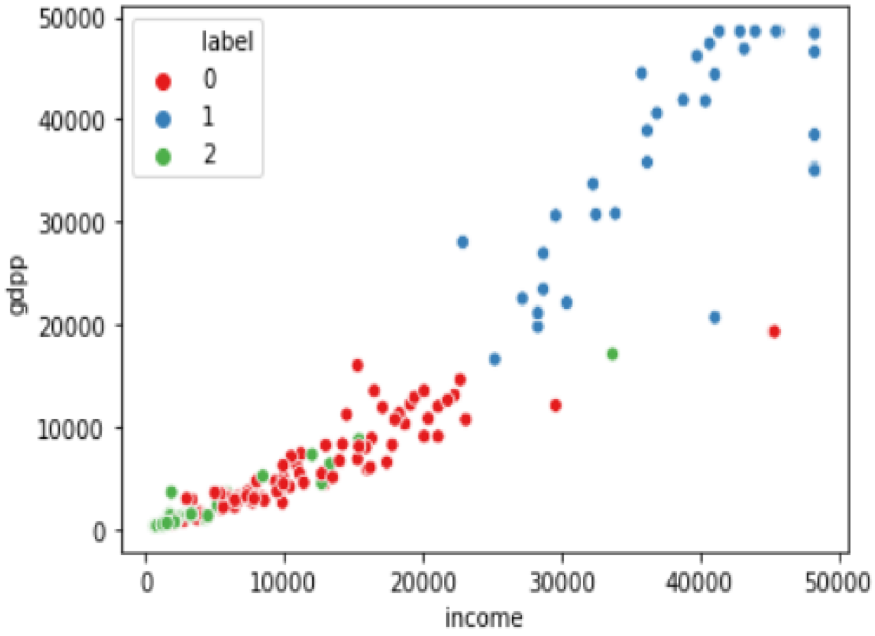
Elbow curve method

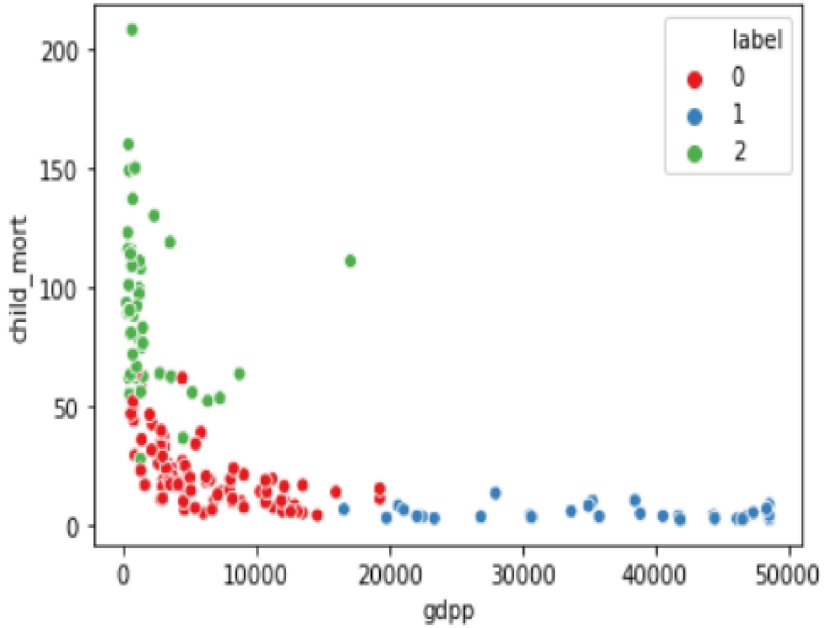


Silhouette score method

- Both “Elbow curve method” and “Silhouette score method “gives 3 as optimum number of clusters.
- Although 2 clusters are more likely indicated by both the methods but as per business aspect we can’t just create two clusters. So we chose 3 as optimum number of clusters.

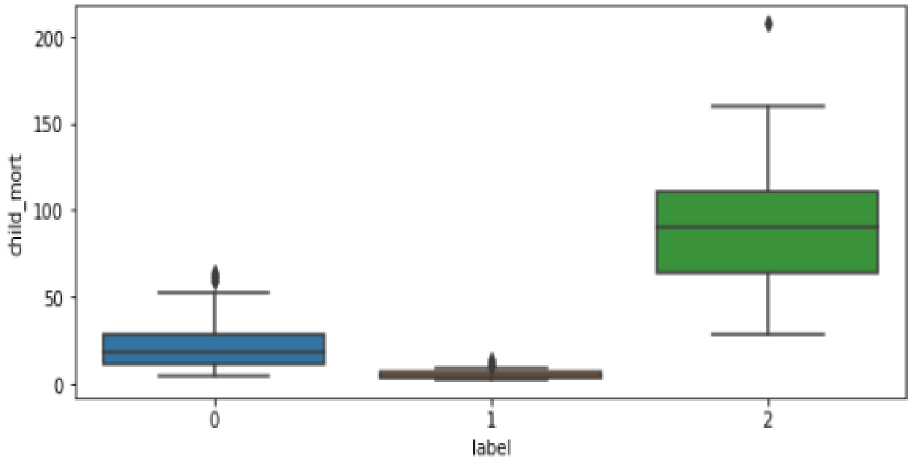
K-means Analysis

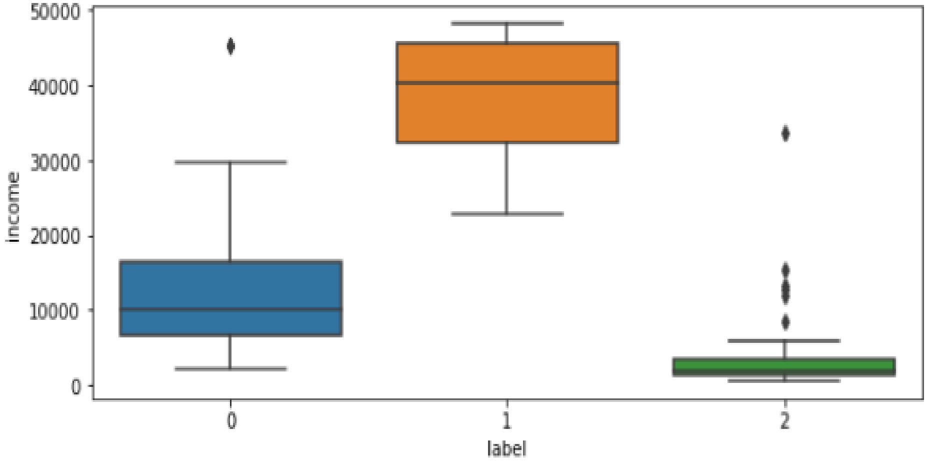


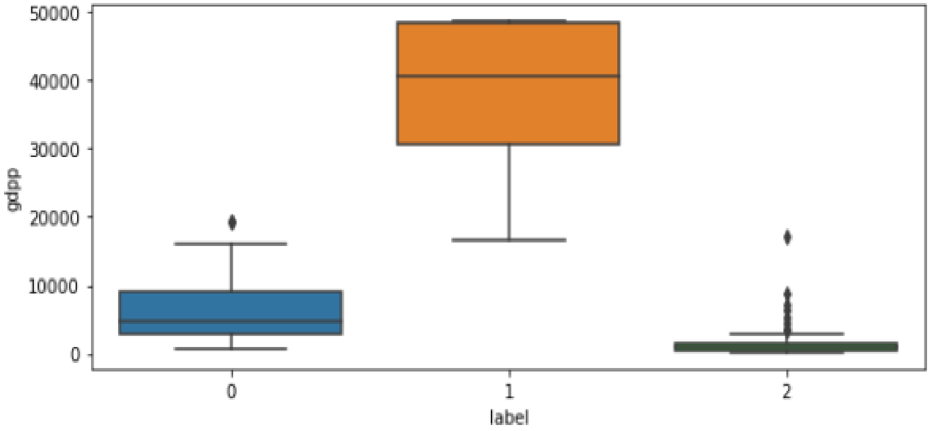


- In scatter plot of “child mortality” vs “gdpp”, we can see that cluster-2 countries has high child mortality rate and low gdpp.
- In scatter plot of “gdpp” vs “Income” cluster – 2 countries has low income and low gdpp.
- So cluster-2 countries are our target countries.

K-means Analysis





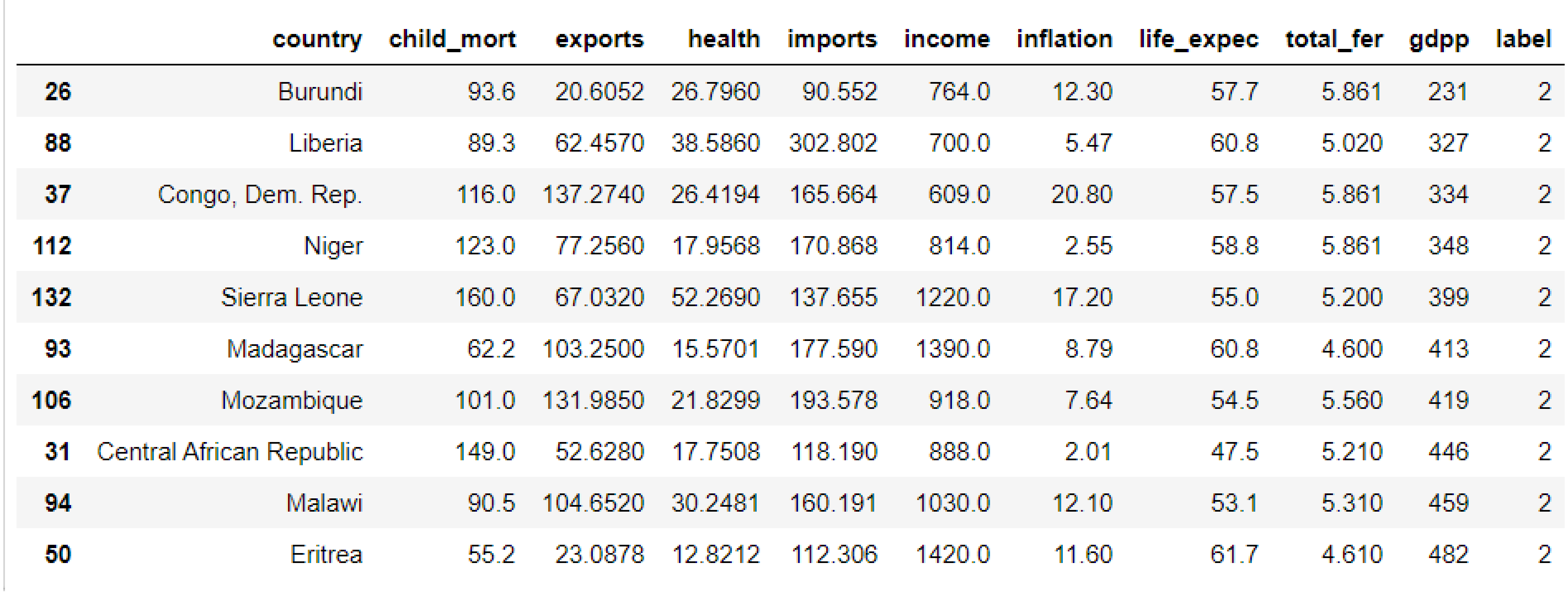


K-means Analysis

- From above boxplot of (Child mortality, Income, gdpp) vs Cluster groups, it is clear that cluster-2 group countries have high avg. child mortality rate, low income and low gdpp.
- So these are the groups of countries which are in need of fund.

Top 10 Countries – K Means

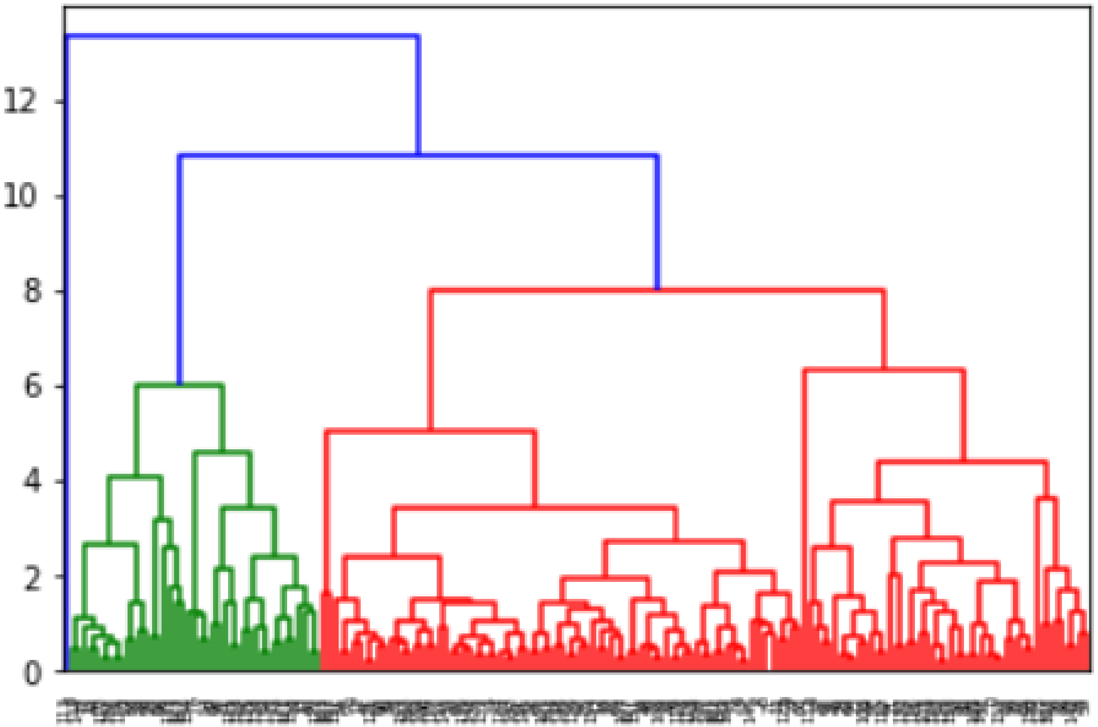
Top 10 countries which are in direst need of aid obtained from K-Means clustering along with other important factors are as follow:



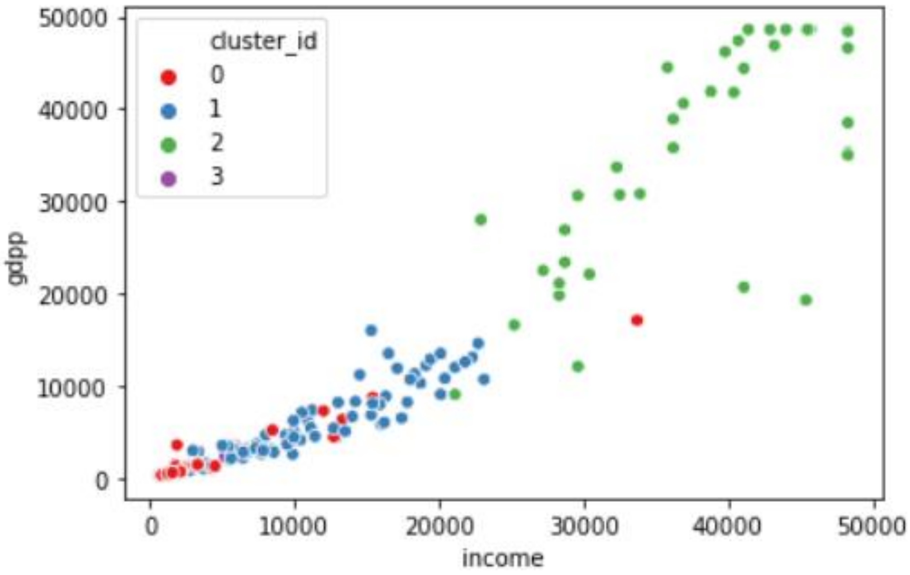
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	5.861	231	2
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.020	327	2
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	5.861	334	2
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	5.861	348	2
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.200	399	2
93	Madagascar	62.2	103.2500	15.5701	177.590	1390.0	8.79	60.8	4.600	413	2
106	Mozambique	101.0	131.9850	21.8299	193.578	918.0	7.64	54.5	5.560	419	2
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.210	446	2
94	Malawi	90.5	104.6520	30.2481	160.191	1030.0	12.10	53.1	5.310	459	2
50	Eritrea	55.2	23.0878	12.8212	112.306	1420.0	11.60	61.7	4.610	482	2

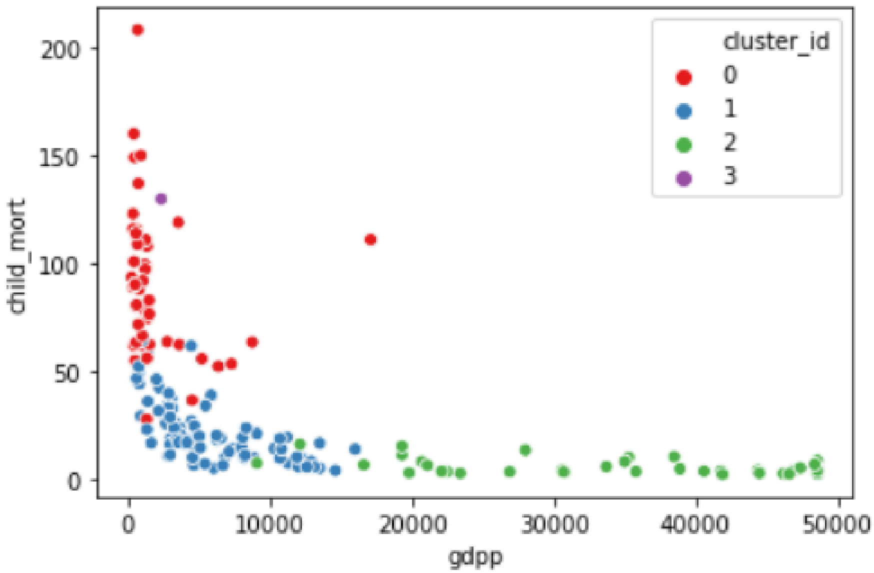
Hierarchical Clustering

Out of all linkage method (single, complete, average) complete linkage gives better dendrogram. Here we cut tree at height 6 to get 4 clusters.



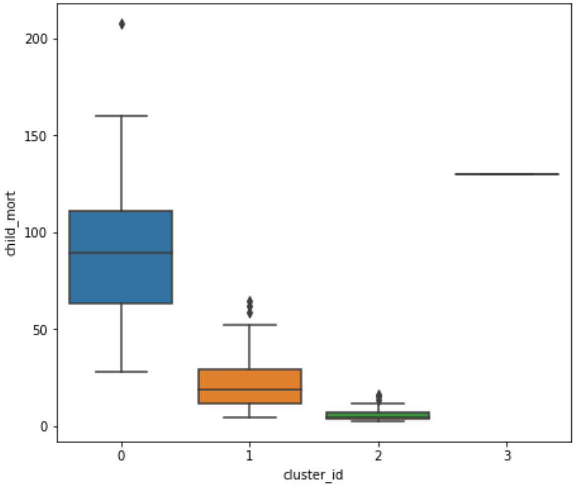
Hierarchical Analysis

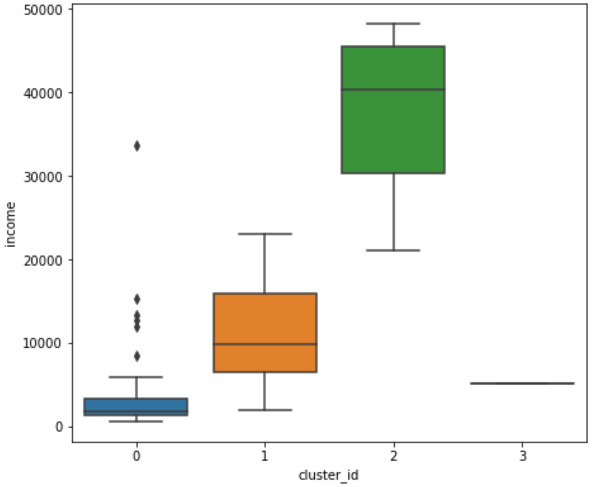


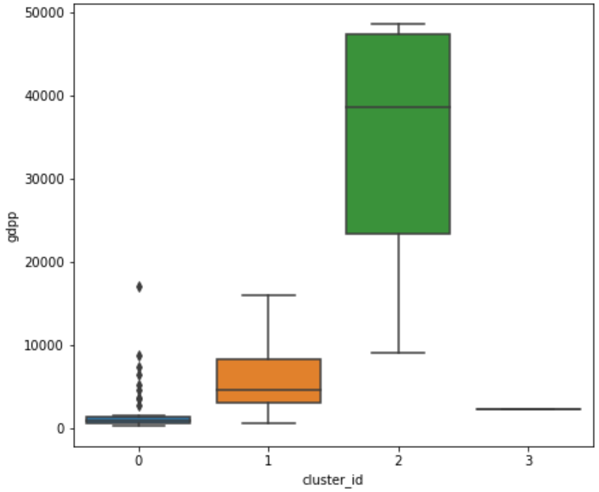


- In scatter plot of “gdpp” vs “Income” cluster – 0 countries has low income and low gdpp.
- In scatter plot of “child mortality” vs “gdpp”, we can see that cluster-0 countries has high child mortality rate and low gdpp.
- So cluster-0 countries are our target countries.

Hierarchical Analysis

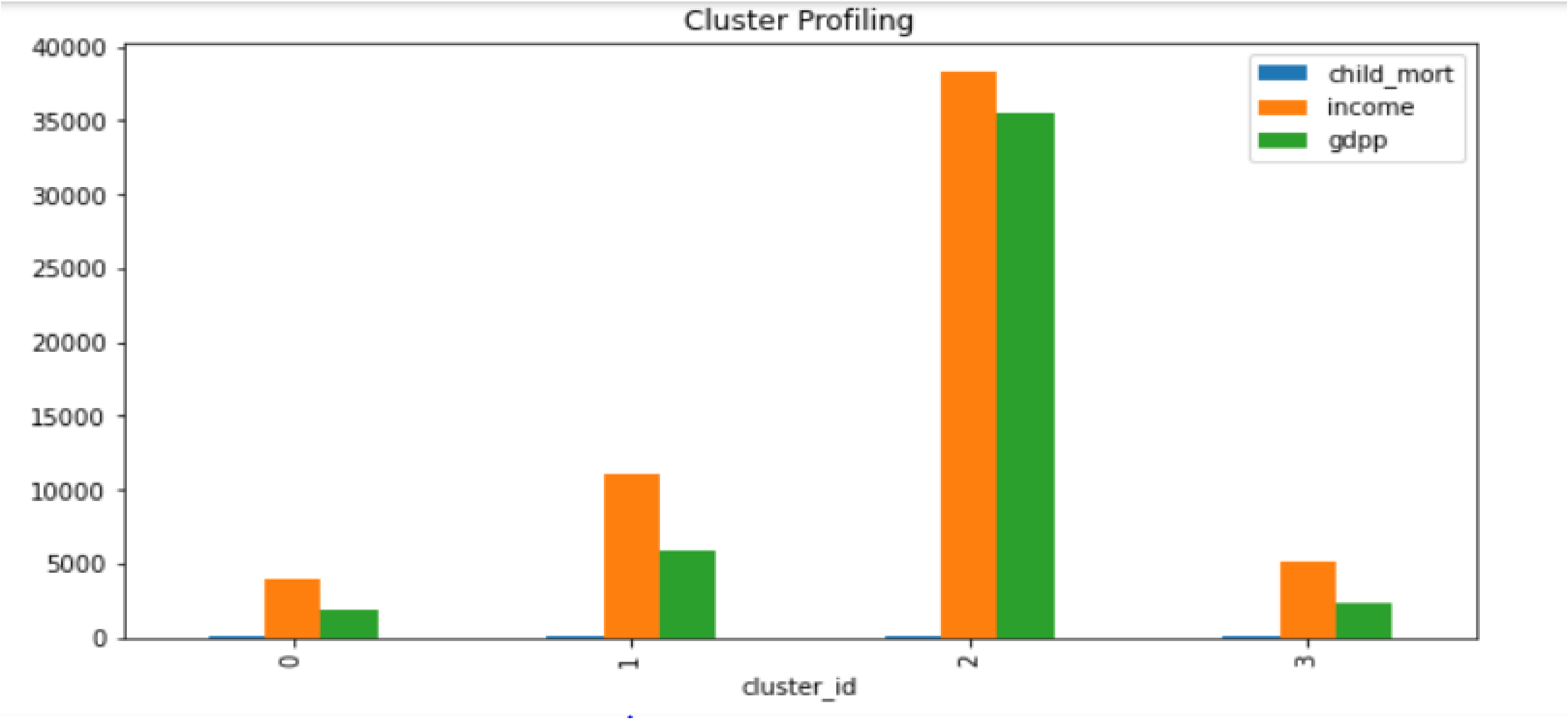






- From above boxplot of (Child mortality, Income, gdpp) vs Cluster groups, it is clear that cluster-0 group countries have high avg. child mortality rate, low income and low gdpp.
- So these are the groups of countries which are in need of fund.

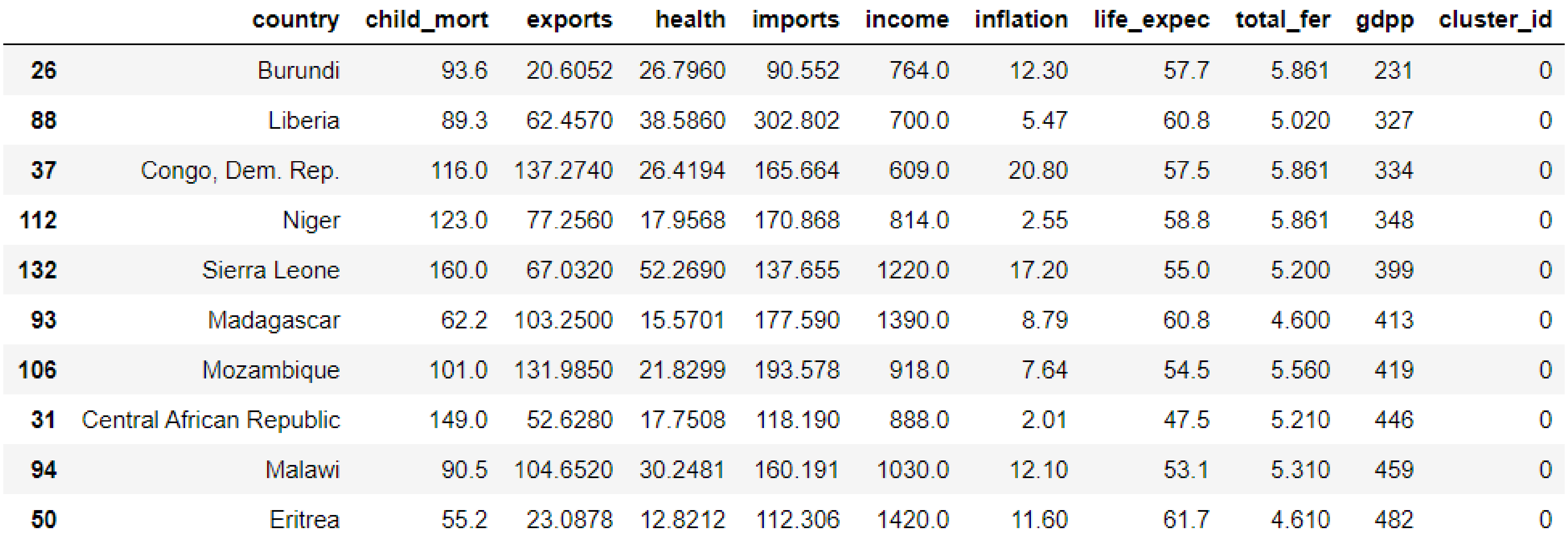
Hierarchical Analysis



- From the above ‘Cluster Profiling graph of (Child mortality, Income, gdpp) vs Cluster ID, we can conclude that cluster-0 group countries have high avg. child mortality rate, low income and low gdpp.
- So these are the groups of countries which are in need of fund.

Top 10 Countries (Hierarchical Clustering)

Top 10 countries which are in direst need of aid obtained from hierarchical clustering along with other important factors are as follow:



	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	5.861	231	0
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.020	327	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	5.861	334	0
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	5.861	348	0
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.200	399	0
93	Madagascar	62.2	103.2500	15.5701	177.590	1390.0	8.79	60.8	4.600	413	0
106	Mozambique	101.0	131.9850	21.8299	193.578	918.0	7.64	54.5	5.560	419	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.210	446	0
94	Malawi	90.5	104.6520	30.2481	160.191	1030.0	12.10	53.1	5.310	459	0
50	Eritrea	55.2	23.0878	12.8212	112.306	1420.0	11.60	61.7	4.610	482	0