Clustering Assignment

Identify the five countries which are in the direst need of aid

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.

Objective:

After the recent funding programmes, they have been able to raise around $ 10 million. As a data analyst, our job is to come up with the list of five countries that are in the direst need of aid.

Analysis Approach

Read and Understand The Data

- Import the Data
- Find the data quality issues

Data Cleaning and Preparation for Modelling

- Converting the columns exports, health and imports from percentage to actual numbers
- check the distribution of each columns
- Outlier Treatment
- Scaling/ Standardization
- Hopkins Statistic to check cluster tendency

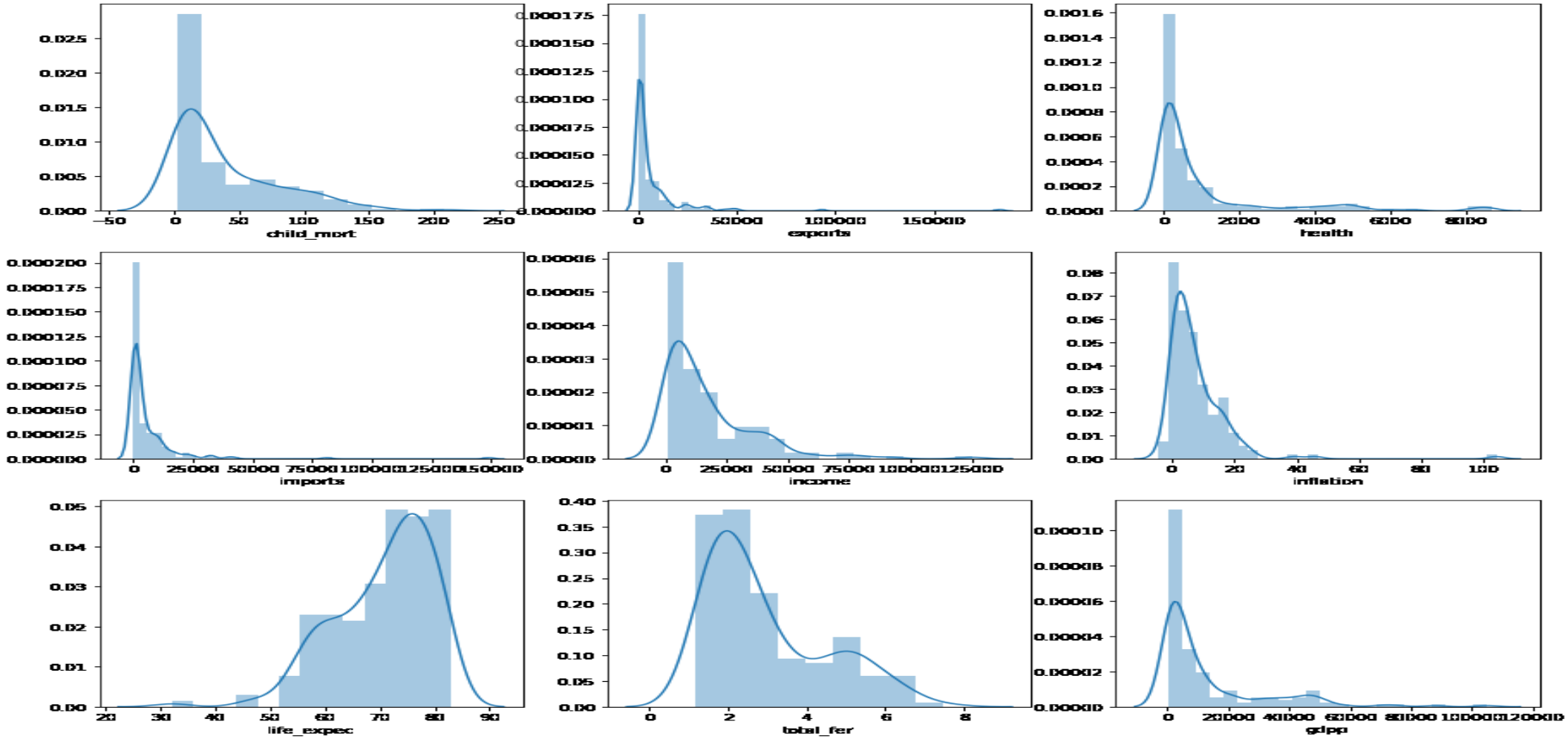
K Means Clustering

- Find the optimal number of clusters by using elbow curve method and silhouette analysis
- Perform K-Means Modelling
- Visualization on the clusters
- Cluster Profiling
- Identify the countries which require aid

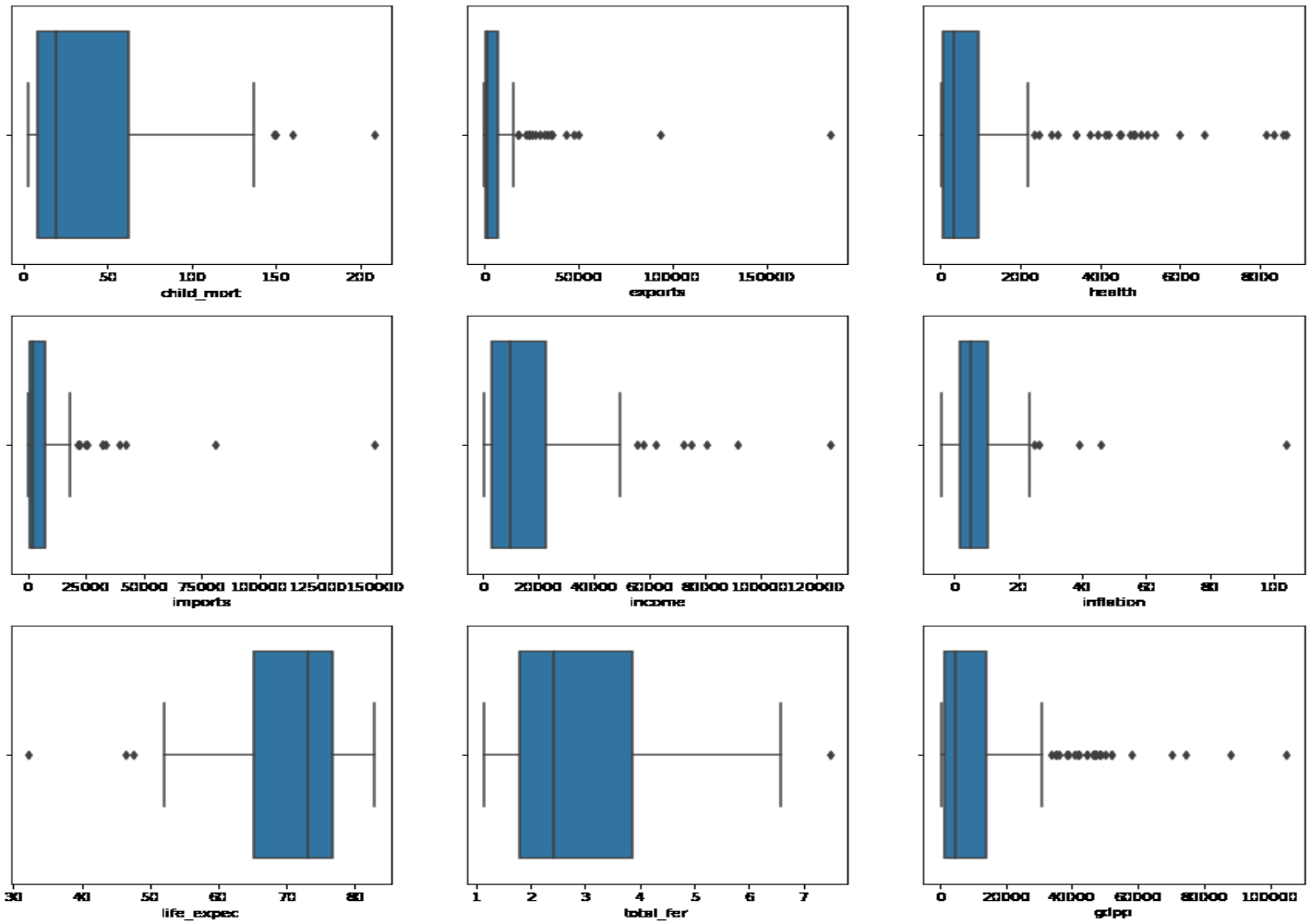
Hierarchical Clustering

- Create Dendrograms Using Single and Complete Linkage
- Visualization on the clusters
- Cluster Profiling
- Identify the countries which require aid

Distribution of each columns using Density Plot

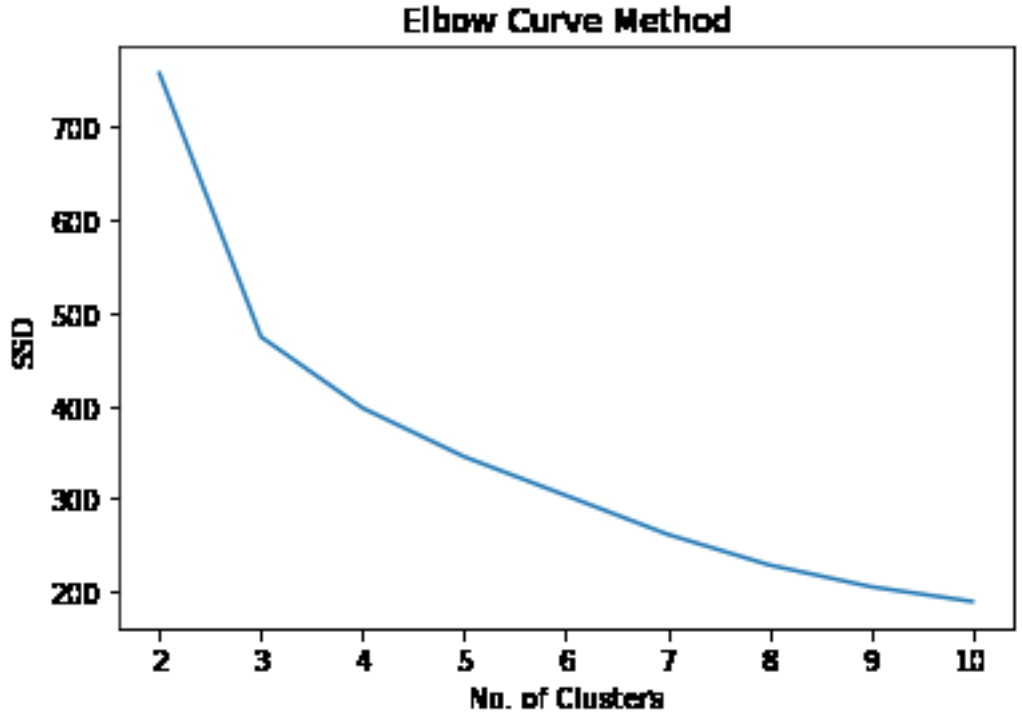


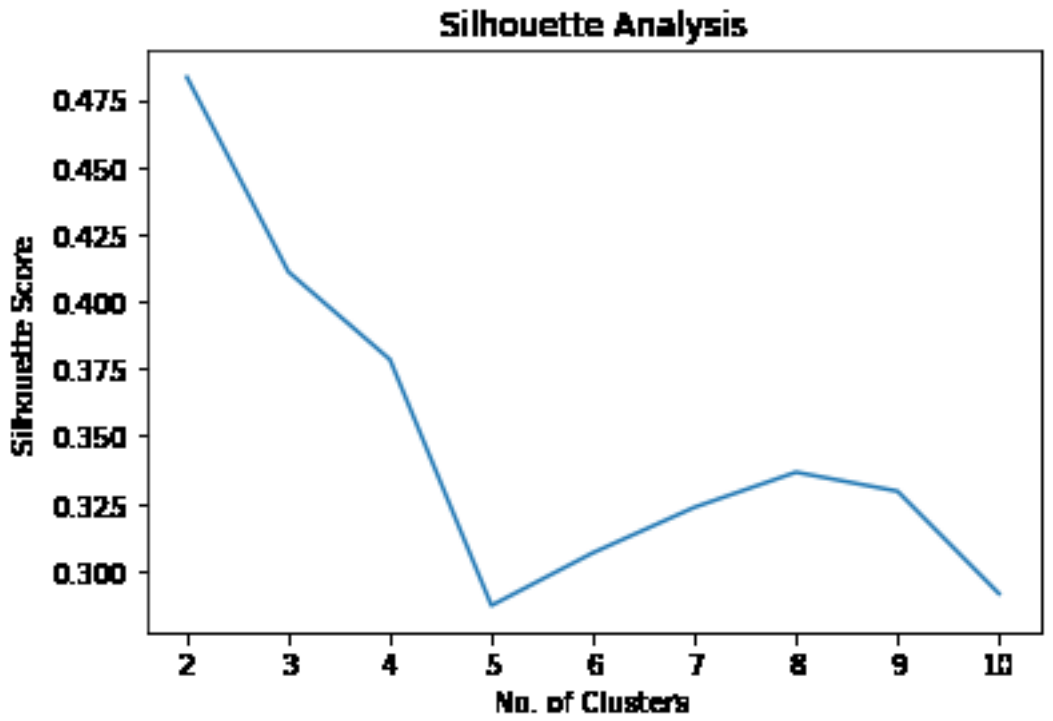
Treatment of Outliers



- All the columns contain outliers.
- As we have to find out the countries which are in the direst need of aid, we cannot just remove the outliers. Outlier must be capped depending upon the column.
- We consider, if child_mort is high, then the country needs immediate aid. So this column will not be capped for extremely high value outliers.
- My approach here is to cap the high value outliers for each column except child_mort.

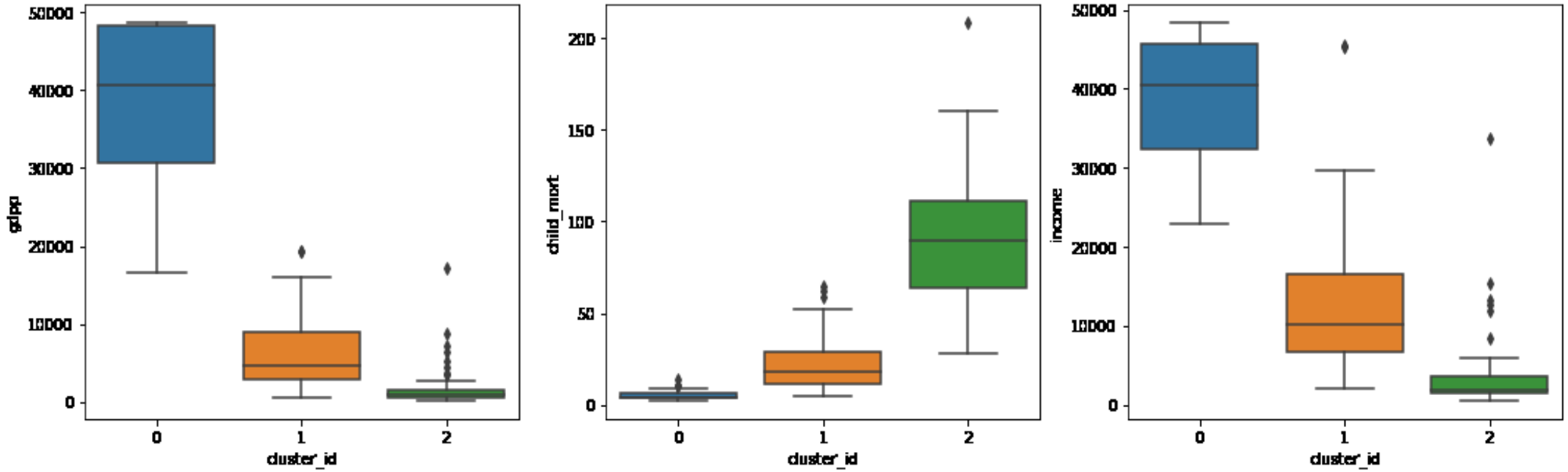
K Means Clustering – Find The Optimal Number of Clusters





- The optimal number of clusters is 3.

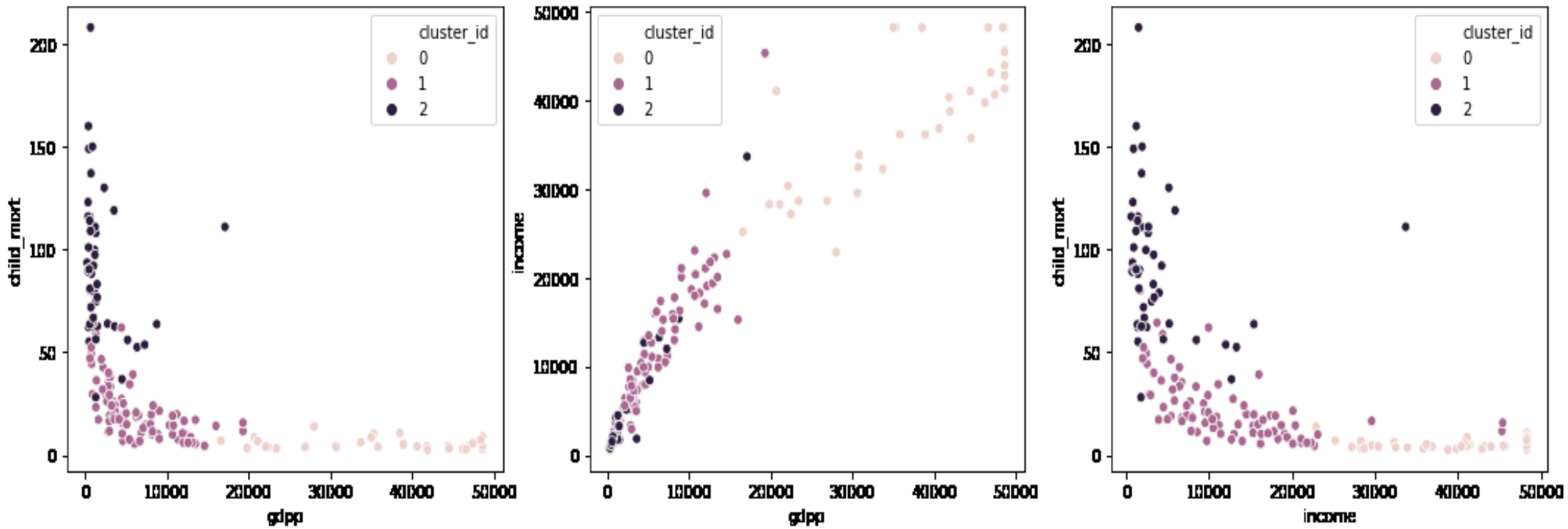
Visualization of Clusters(1/2)



- Cluster 0 contains countries with high gdpp, low child_mort and high income.
- Cluster 1 contains countries with low gdpp, medium child_mort and medium income.
- Cluster 2 contains countries with the lowest gdpp, highest child_mort and the lowest income.

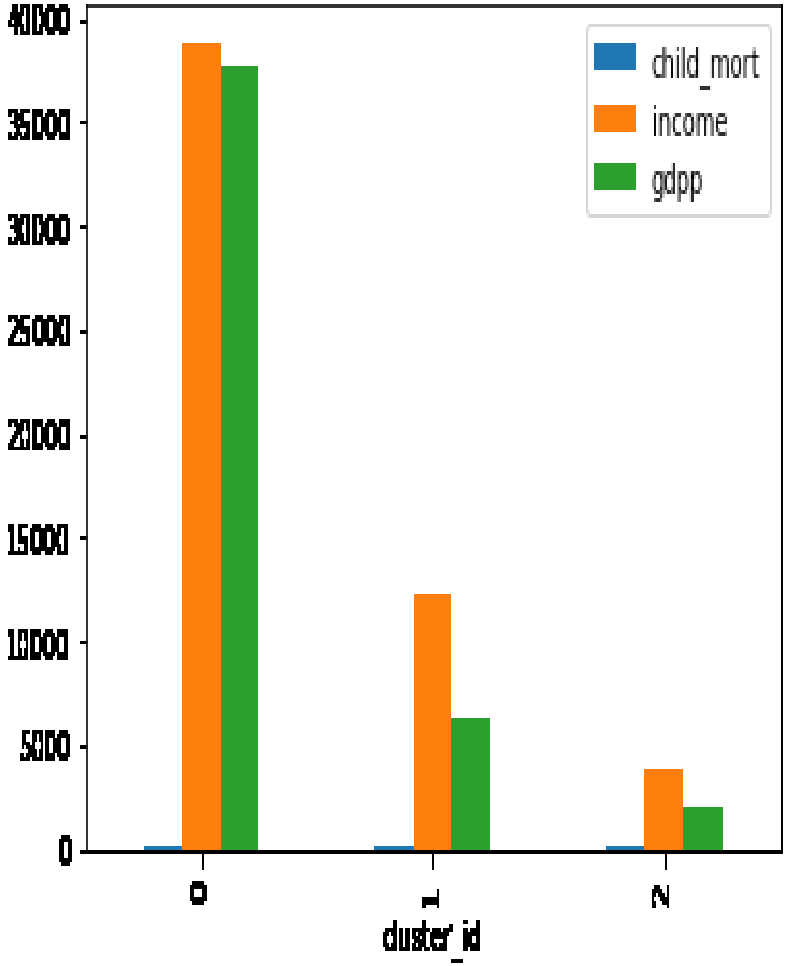
So, the countries in cluster 2 are in the direst need of aid

Visualization of Clusters (2/2)



Scatter plot of the clusters shows that the countries in cluster 2 are in the direst need of aid.

Cluster Profiling and Conclusion from K Means Clustering



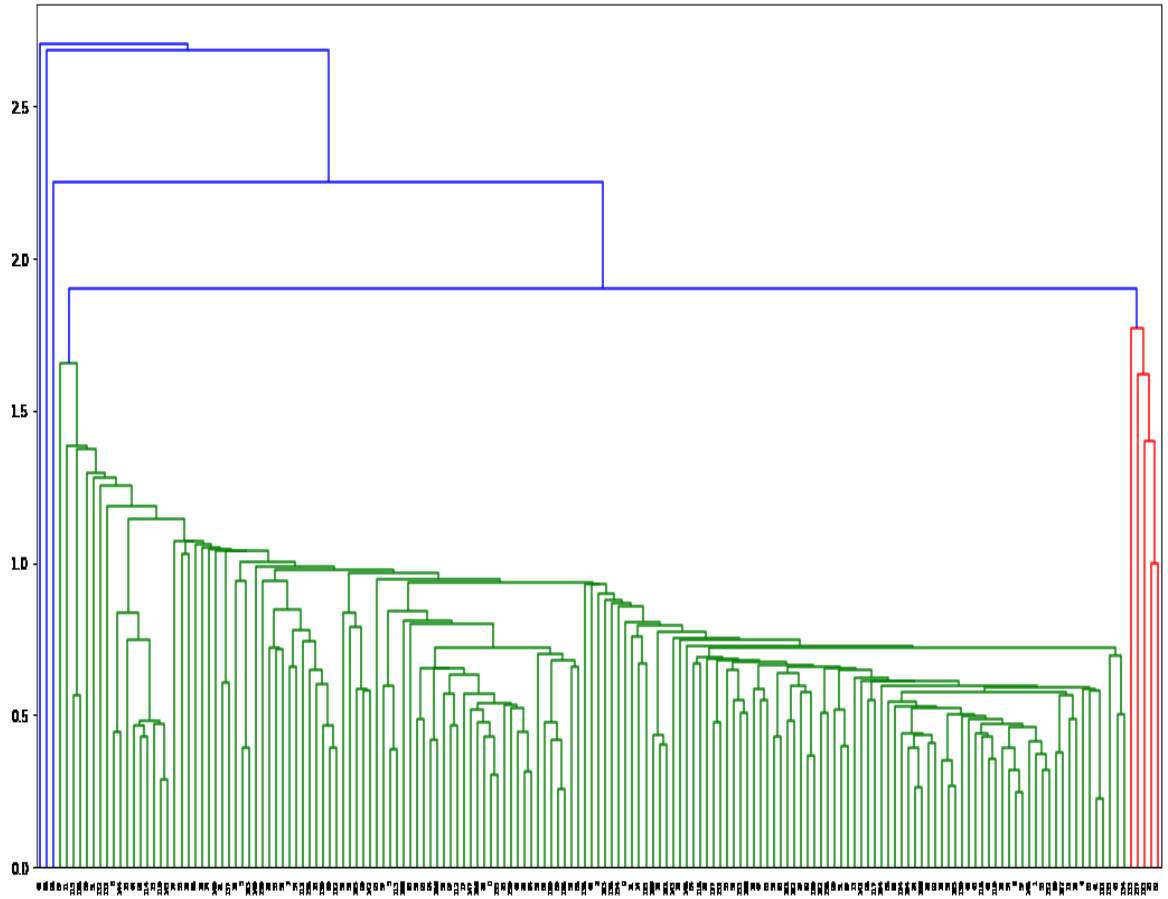
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	5.861	231	2
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.020	327	2
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	5.861	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

The five countries which are in the direst need of aid are the countries having the lowest gdpp, highest child_mort and the lowest income. Those 5 countries are as follows.

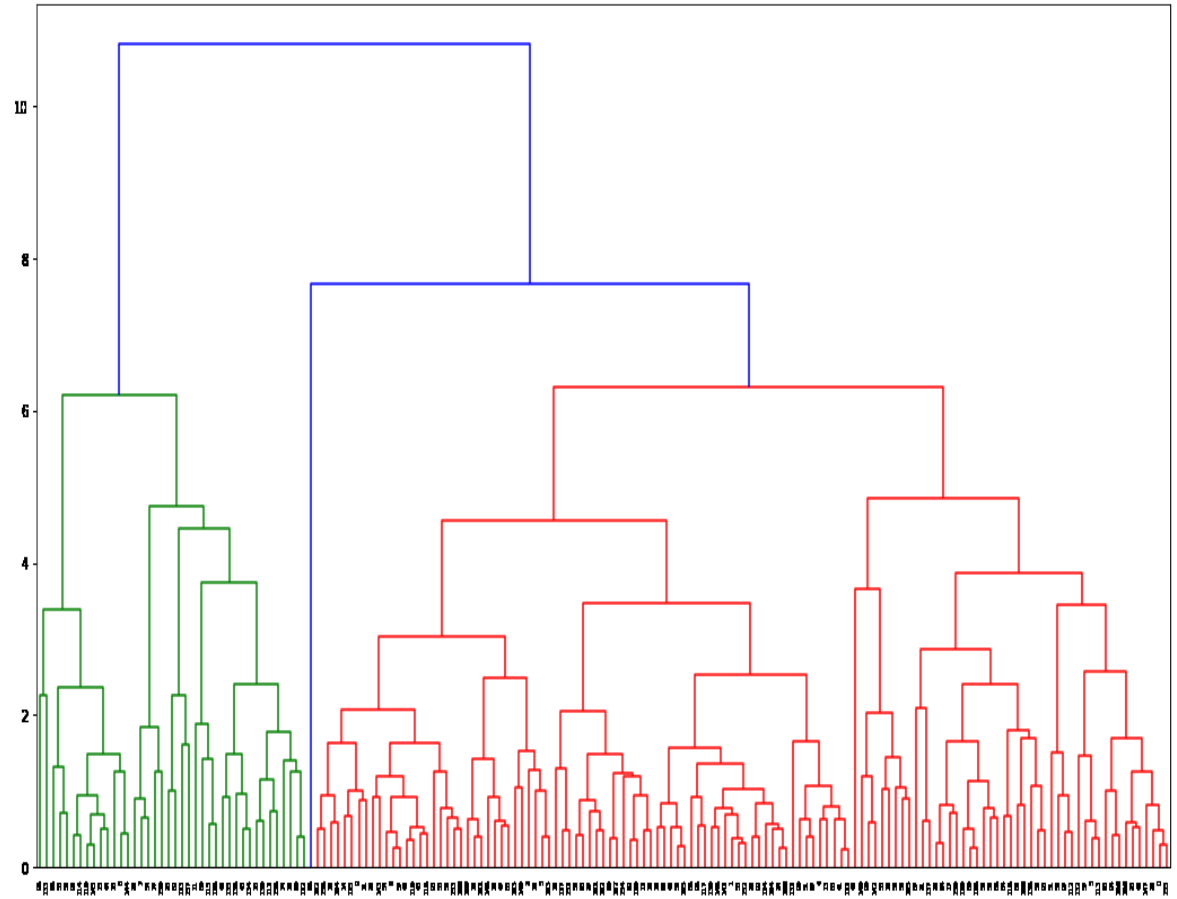
- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone

Hierarchical Clustering - Dendrograms

Single Linkage

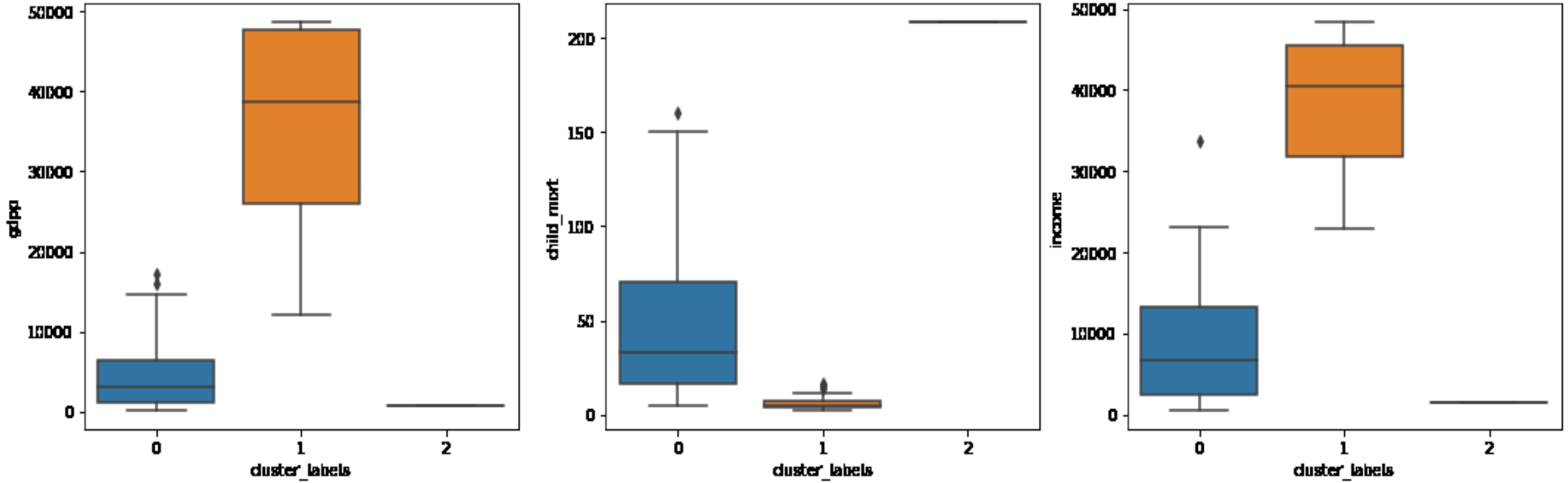


Complete Linkage



Complete linkage dendrograms seems to easily interpretable and clear. Looking at the height of the dendrograms, the optimal number of clusters will be 3.

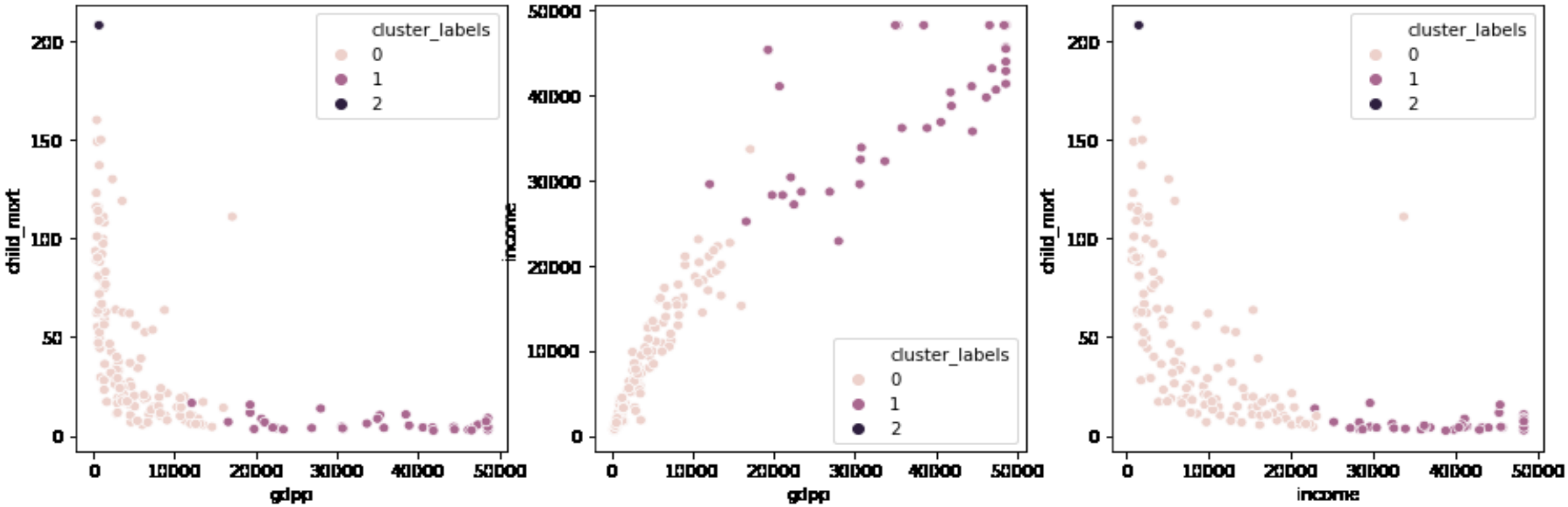
Visualization of Clusters(1/2)



- Cluster 0 contains countries with very low gdpp,high child_mort and low income.
- cluster 1 contains countries with high income, low child_mort and high income.
- cluster 2 contains countries with low gdpp, highest child_mort and very low income. It seems cluster 2 contains less number of data points also.

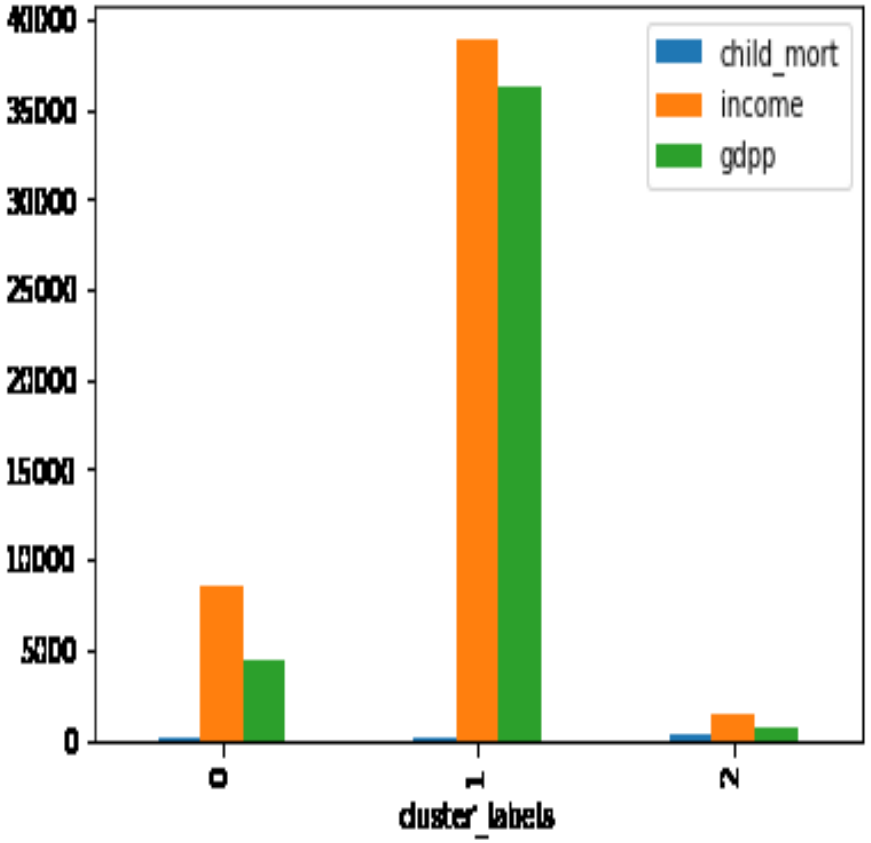
So cluster 0 and cluster 2 are in the direst need of aid

Visualization of Clusters (2/2)



Scatter plot of the clusters shows that the countries in cluster 0 and cluster 2 are in the direst need of aid.

Cluster Profiling and Conclusion from Hierarchical Clustering



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_labels
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	5.861	231	2	0
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.020	327	2	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	5.861	334	2	0
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	5.861	348	2	0
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.200	399	2	0

The five countries which are in the direst need of aid are the countries having the lowest gdpp, highest child_mort and the lowest income. Those 5 countries are as follows.

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

Conclusion

I found both the results from K Means and Hierarchical clustering to be same for the five countries which are in the direst need of aid. Those 5 countries are as follows.

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

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	5.861	231
Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.020	327
Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	5.861	334
Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	5.861	348
Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.200	399