Clustering of Countries

Presented by : Mohit Patil
(mohitz4418@gmail.com)

Problem Statement

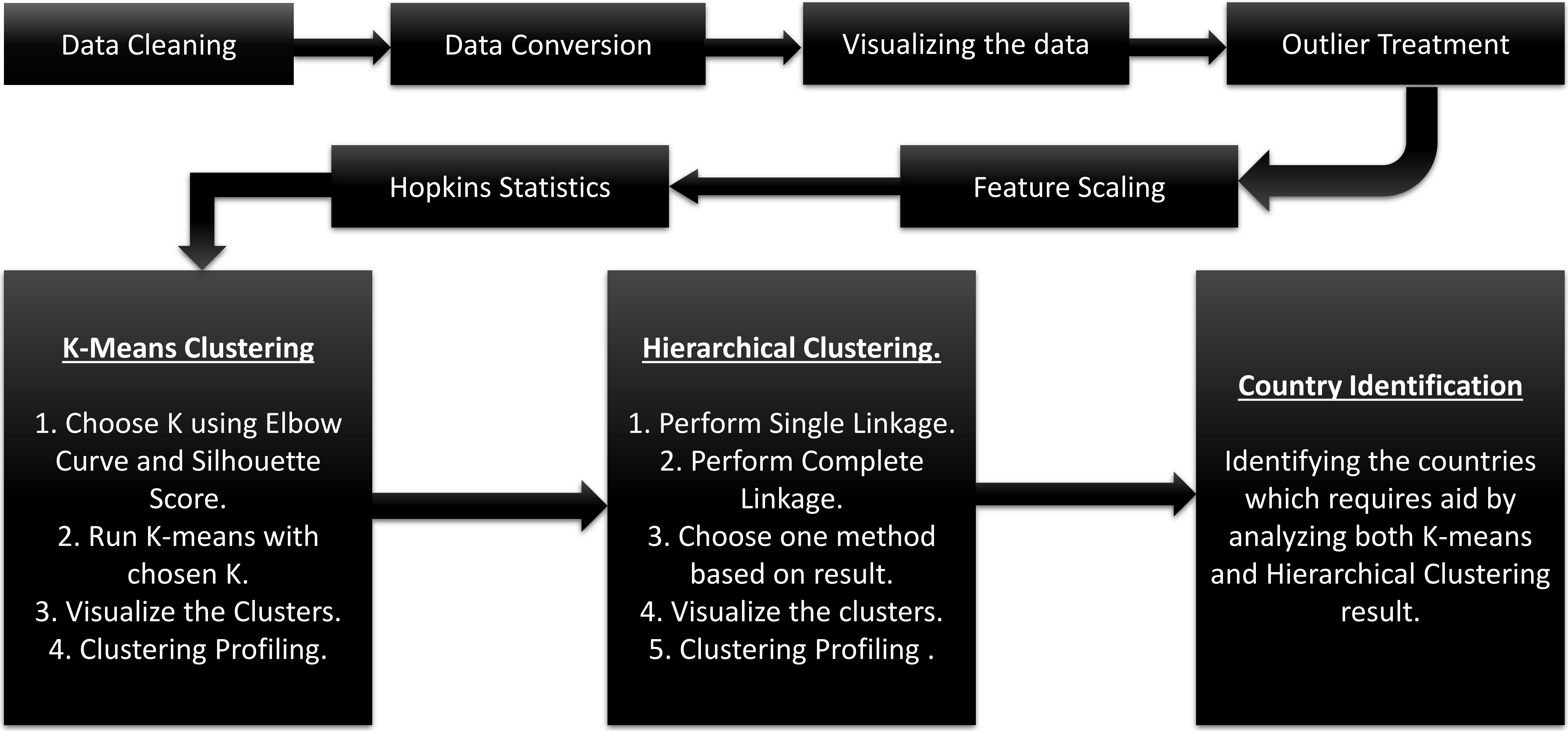
HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy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

After the recent funding programs, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money 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.

Objective:

To categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Analysis Methodology



Data Cleaning

Data Conversion

Visualizing the data

Outlier Treatment

Feature Scaling

Hopkins Statistics

K-Means Clustering

1. Choose K using Elbow Curve and Silhouette Score.
2. Run K-means with chosen K.
3. Visualize the Clusters.
4. Clustering Profiling.

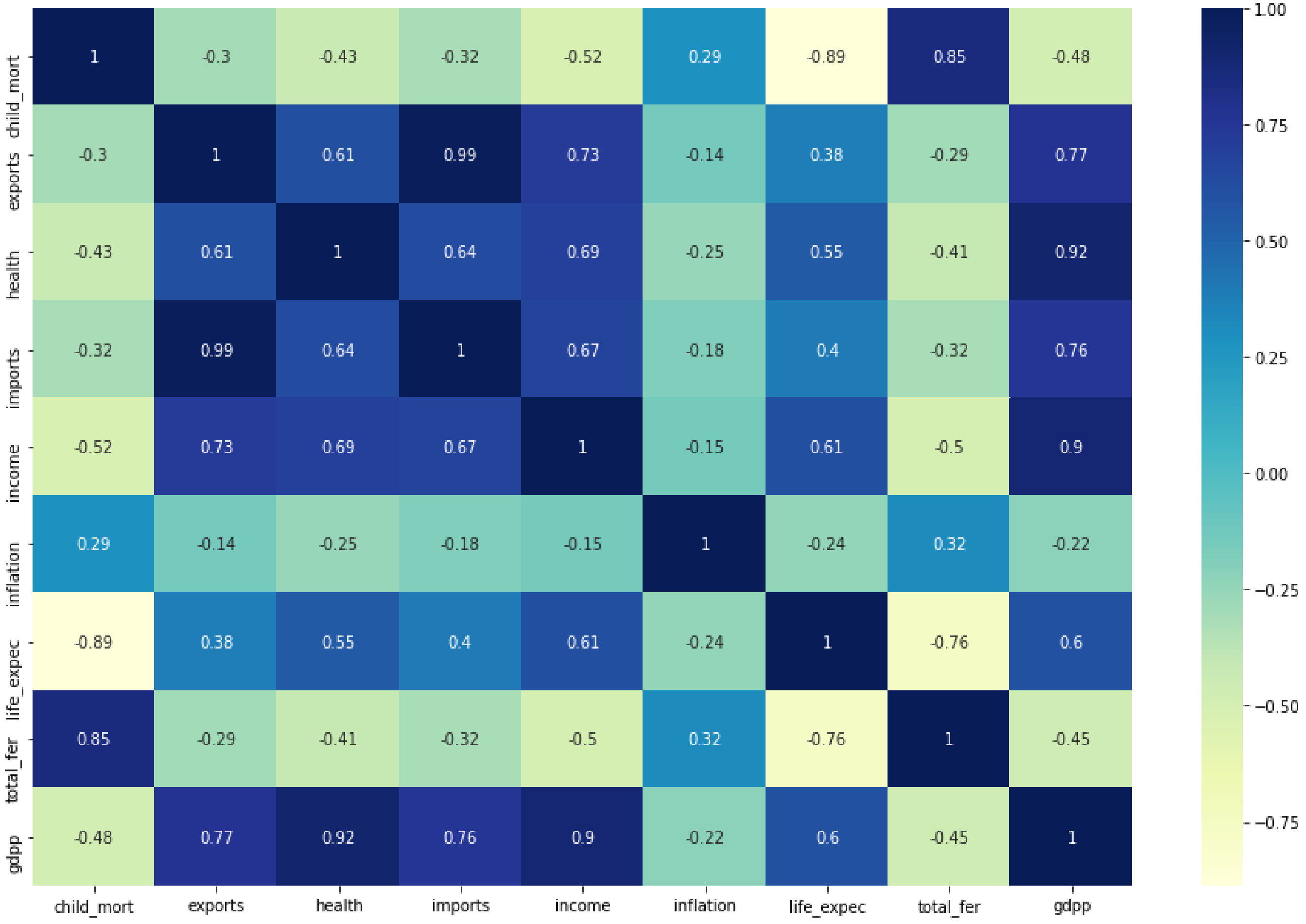
Hierarchical Clustering.

1. Perform Single Linkage.
2. Perform Complete Linkage.
3. Choose one method based on result.
4. Visualize the clusters.
5. Clustering Profiling .

Country Identification

Identifying the countries which requires aid by analyzing both K-means and Hierarchical Clustering result.

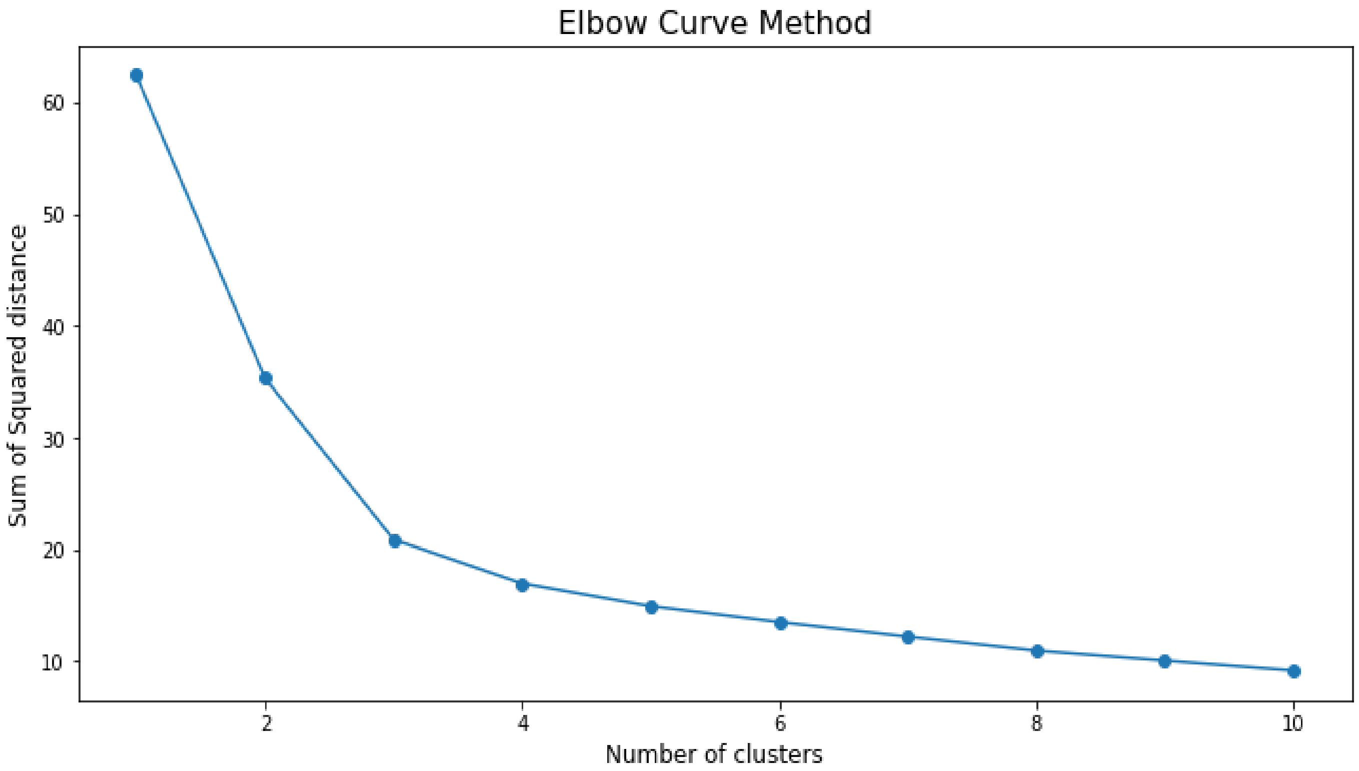
Correlation of data.



- After data cleaning, we cannot remove data since we had less data so, we cap the data.
- After cleaning we scale the data by using min max scaling.
- Looking at heatmap we see that few variables like income, gdpp, child_mort, total_fer and imports, exports have high correlation.

K-means Clustering.

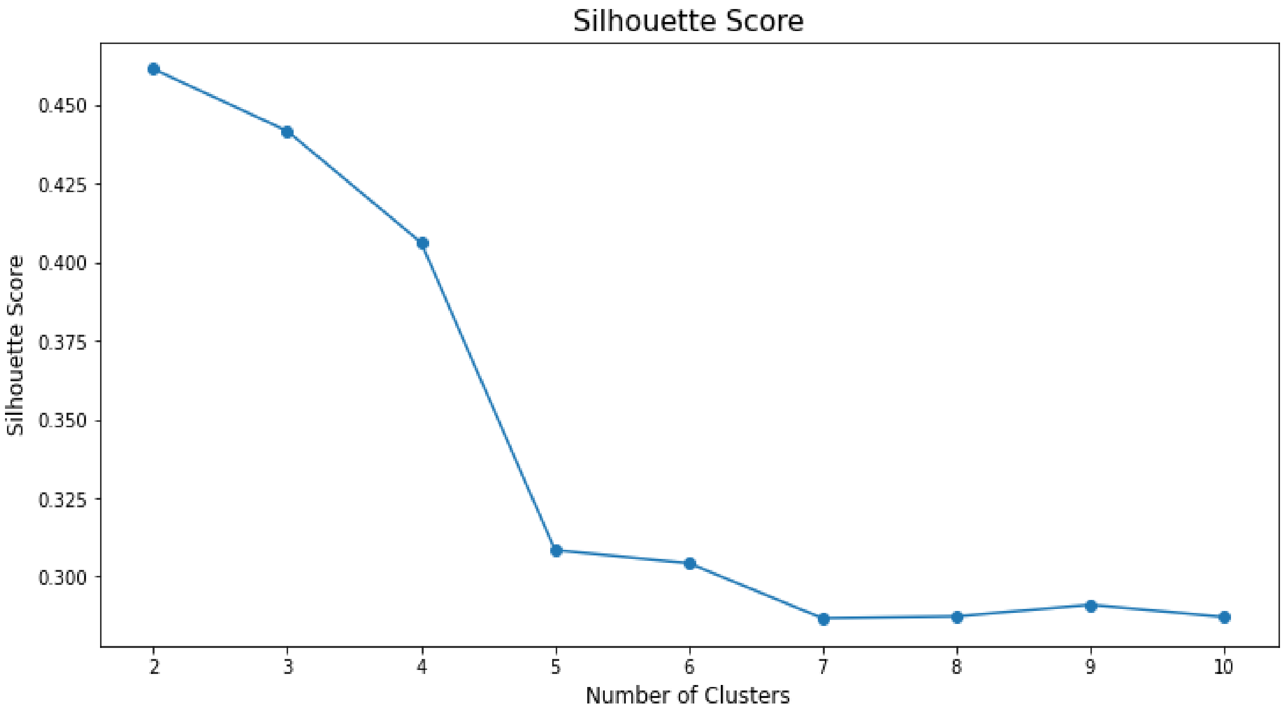
1. Elbow Curve.



- In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.
- From the elbow curve method we have found that the optimal number of clusters are 3.

K-means Clustering.

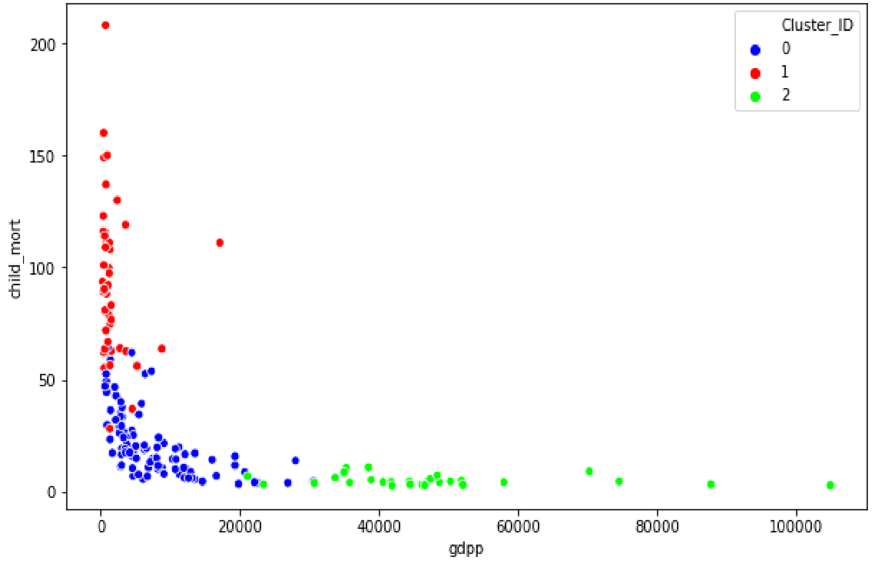
Silhouette Score



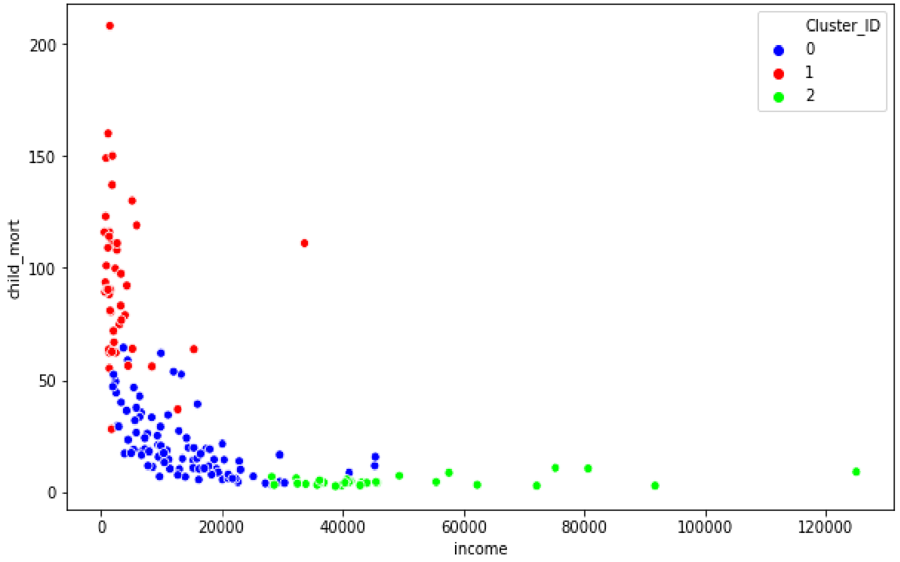
- The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b).

- From the Silhouette Score method we have found that the optimal number of clusters are 3.

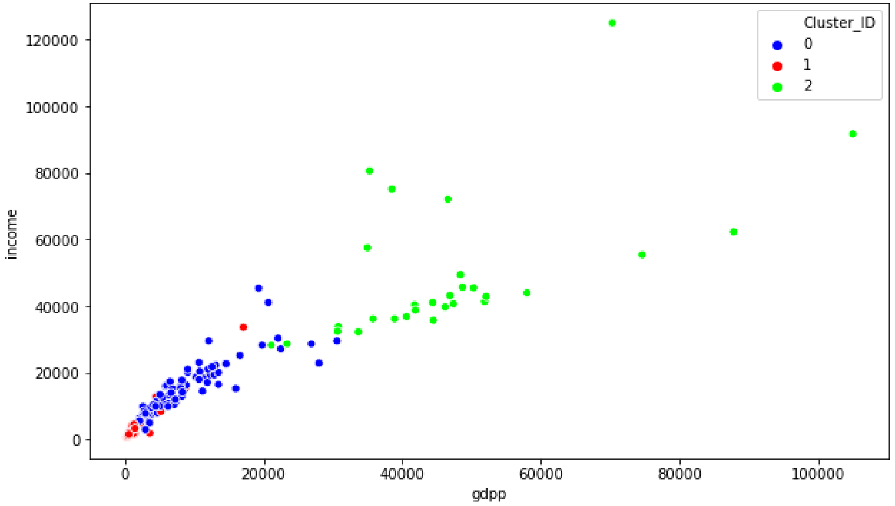
K-means Clustering



1. child_mort vs gdpp



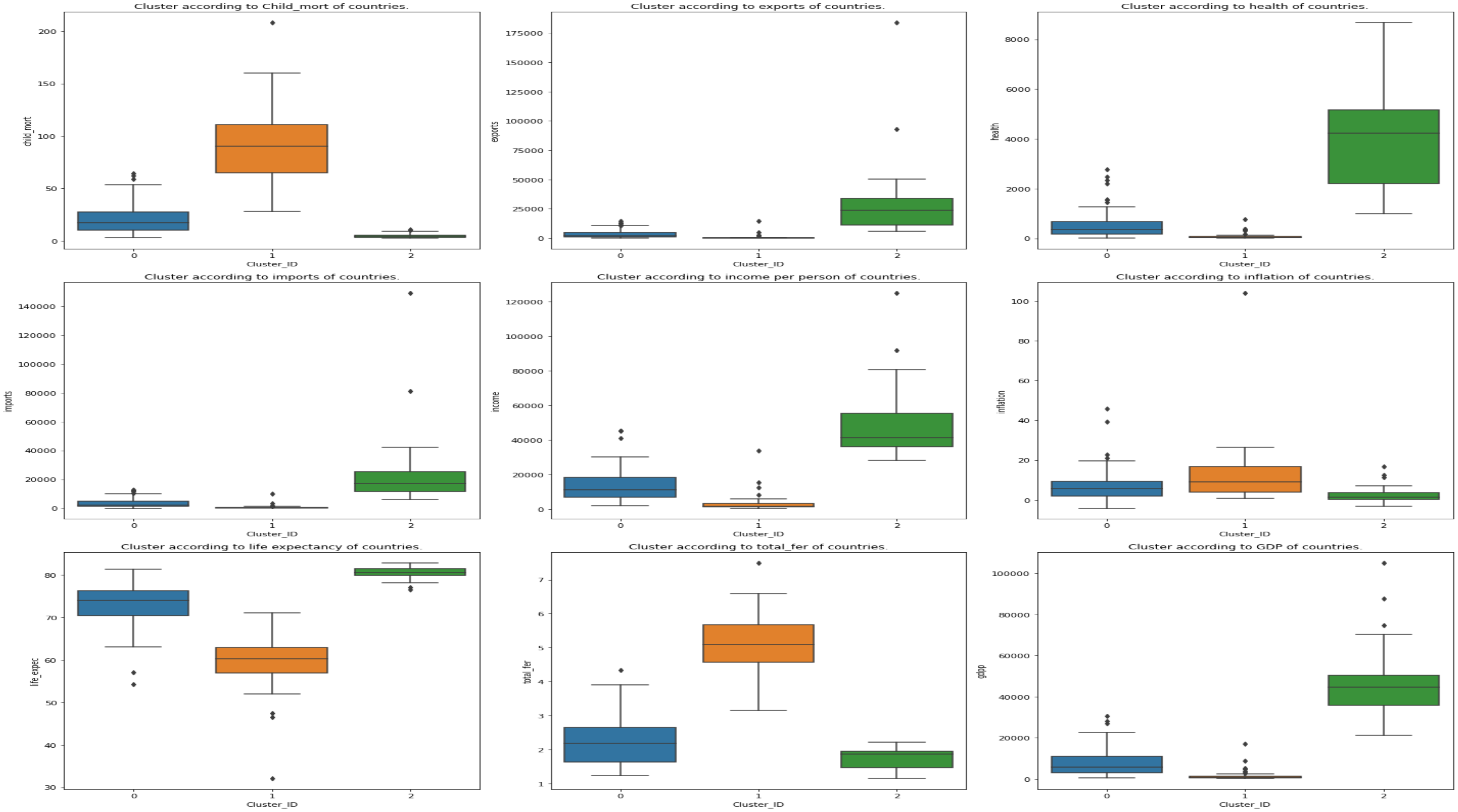
3. child_mort vs income



2. income vs gdpp

As we can see the we formed 3 cluster showing scatter plot between mainly three features income, GDP and child mortality rate.

All columns after K-means clustering with its respective clusters.



Cluster according to Child_mort of countries.

Cluster according to exports of countries.

Cluster according to health of countries.

Cluster according to imports of countries.

Cluster according to income per person of countries.

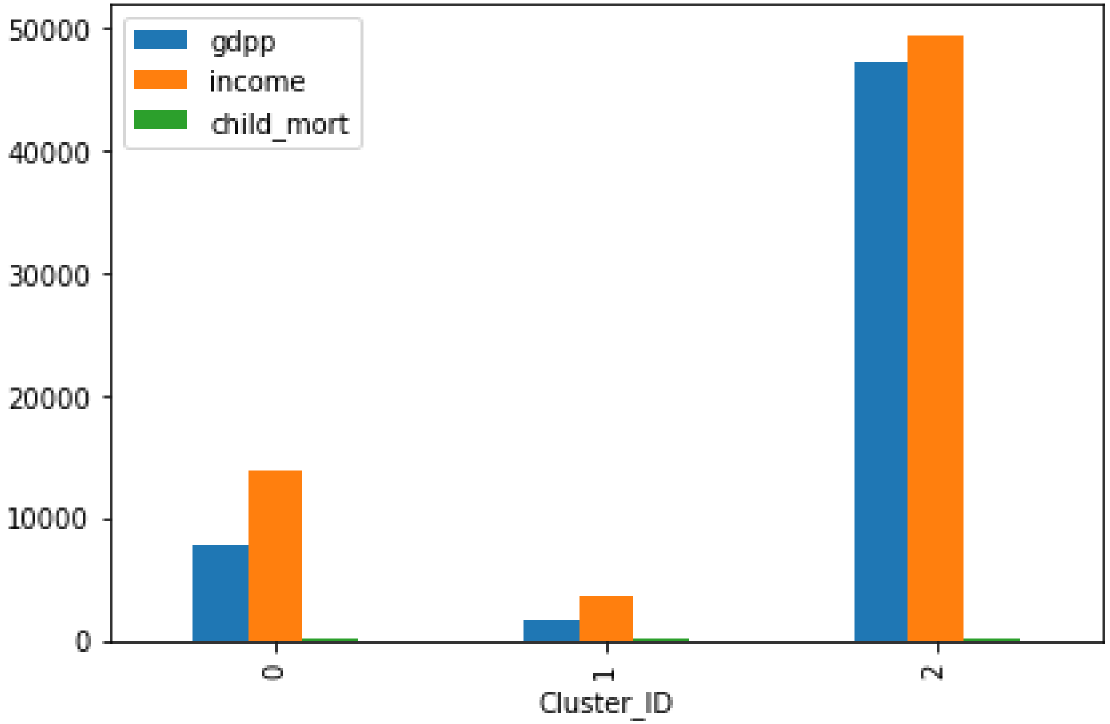
Cluster according to inflation of countries.

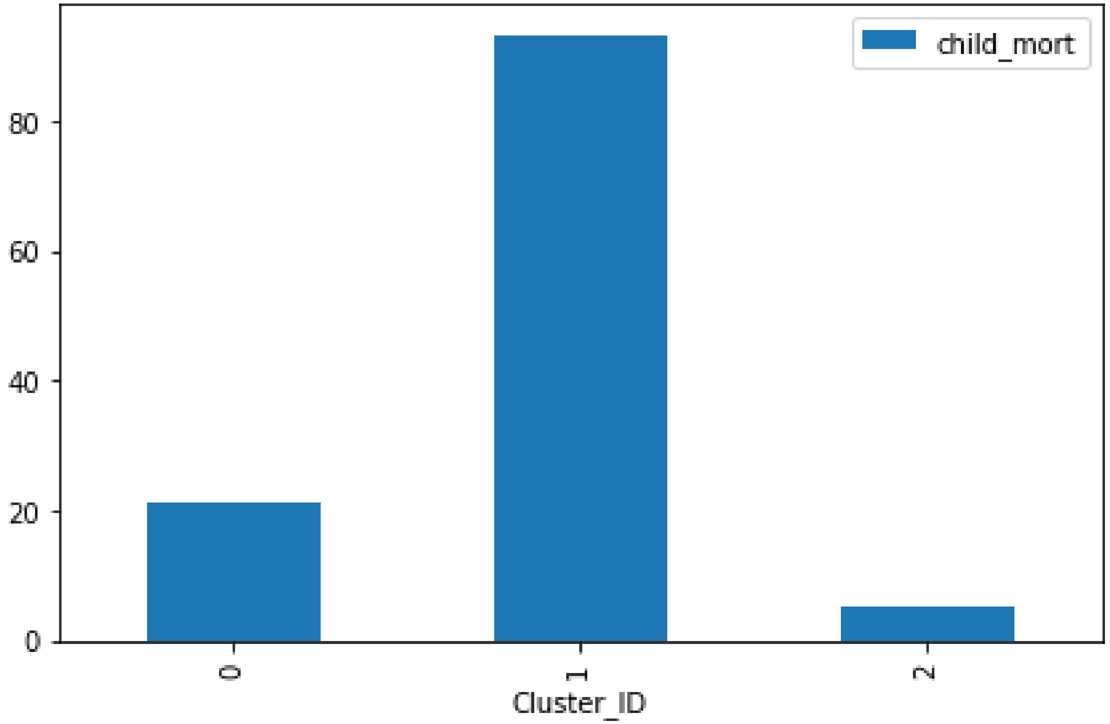
Cluster according to life expectancy of countries.

Cluster according to total_fer of countries.

Cluster according to GDP of countries.

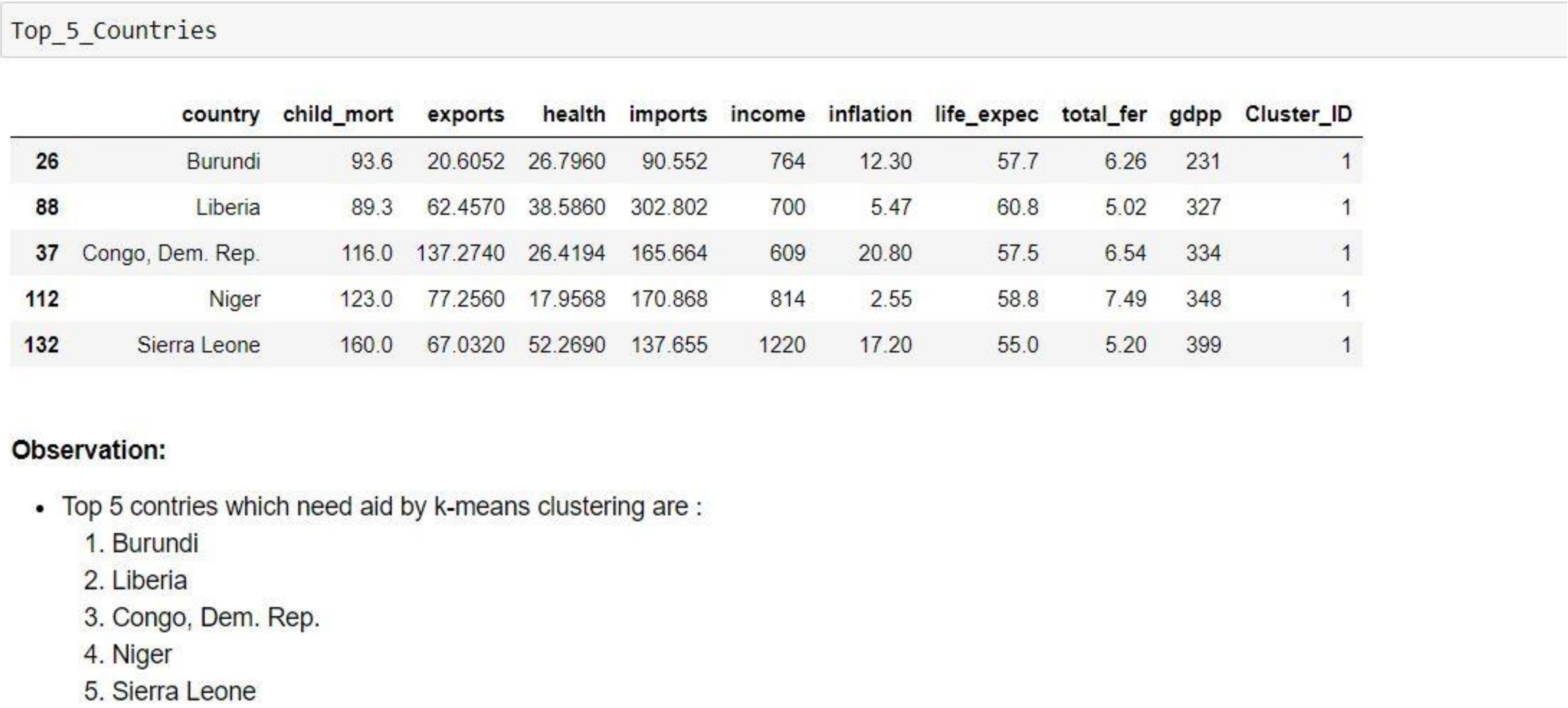
Clustering Profiling after K-means clustering.





- In the above plot the bar of child_mort is not get visualized properly
- So plot a new graph between child_mort and Cluster labels.
- From above plot we can observe that the cluster 1 has highest child mortality rate and lowest income and GDP rate countries.
- So the countries which are needed the aid are belongs from cluster 1.

After K-means clustering we get following countries which needed aid.



```
Top_5_Countries
```

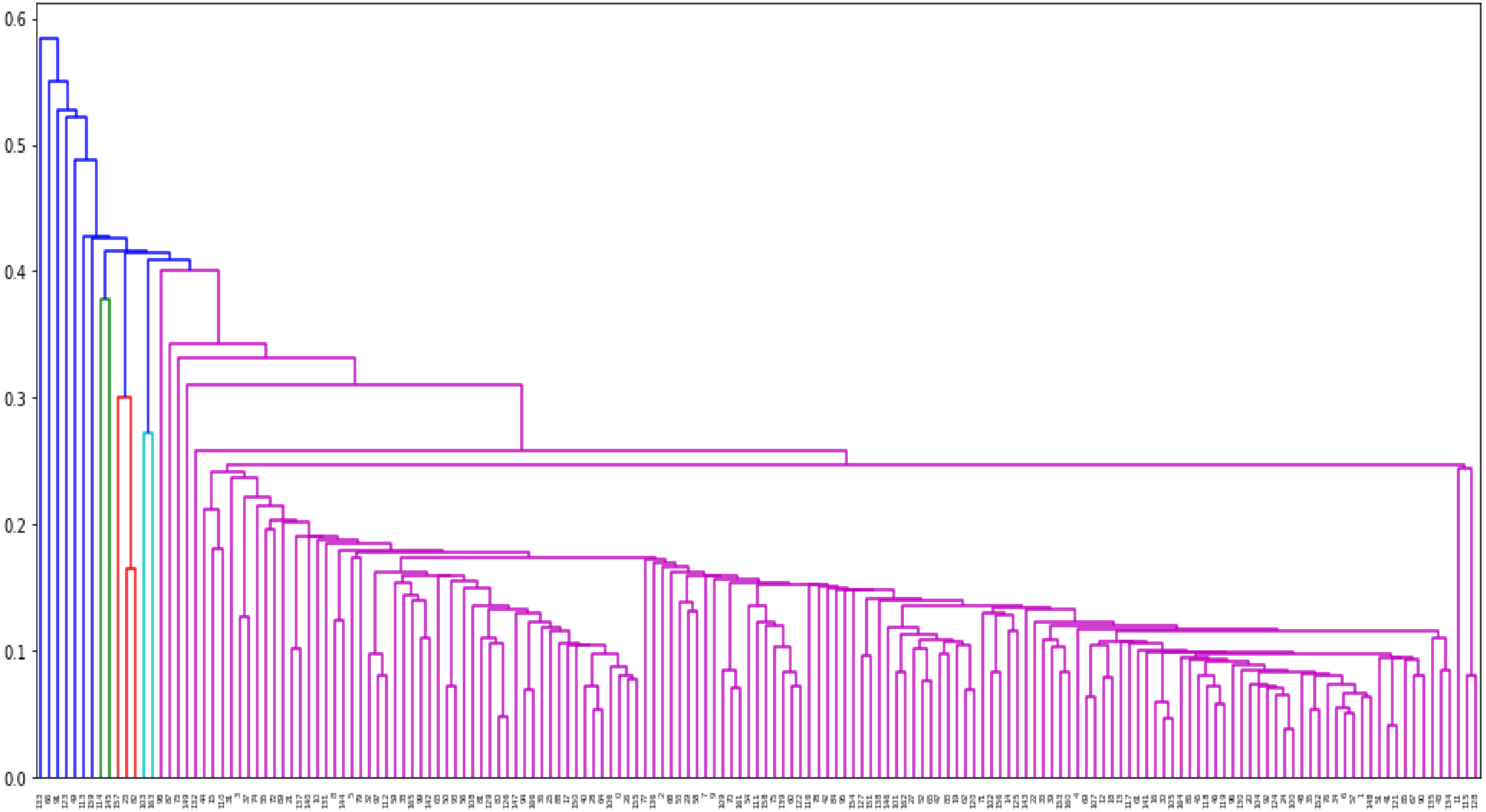
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Cluster_ID
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	1
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	1
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	1
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	1
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399	1

Observation:

- Top 5 contries which need aid by k-means clustering are :
    1. Burundi
    2. Liberia
    3. Congo, Dem. Rep.
    4. Niger
    5. Sierra Leone

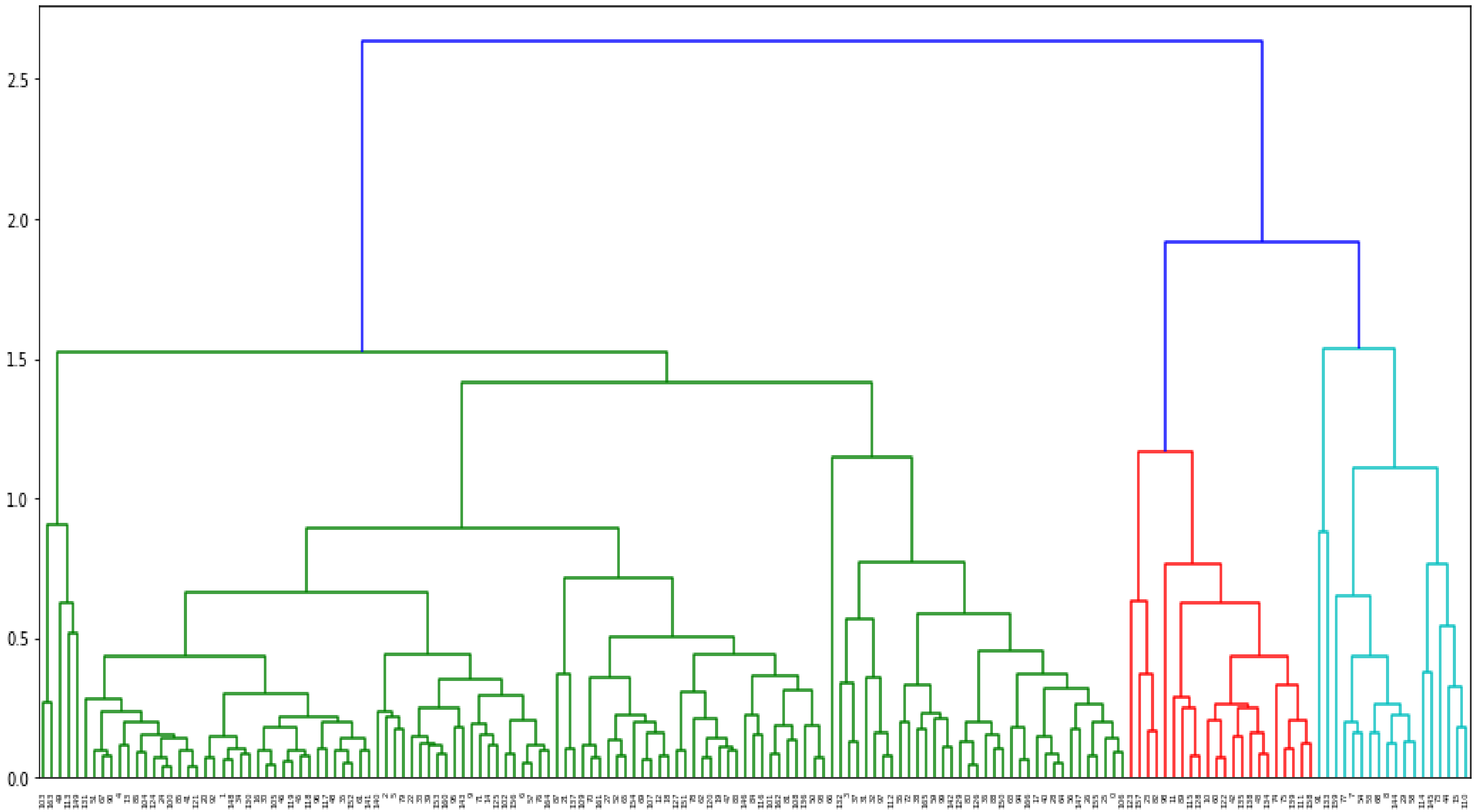
Hierarchical Clustering.

1. Single Linkage.



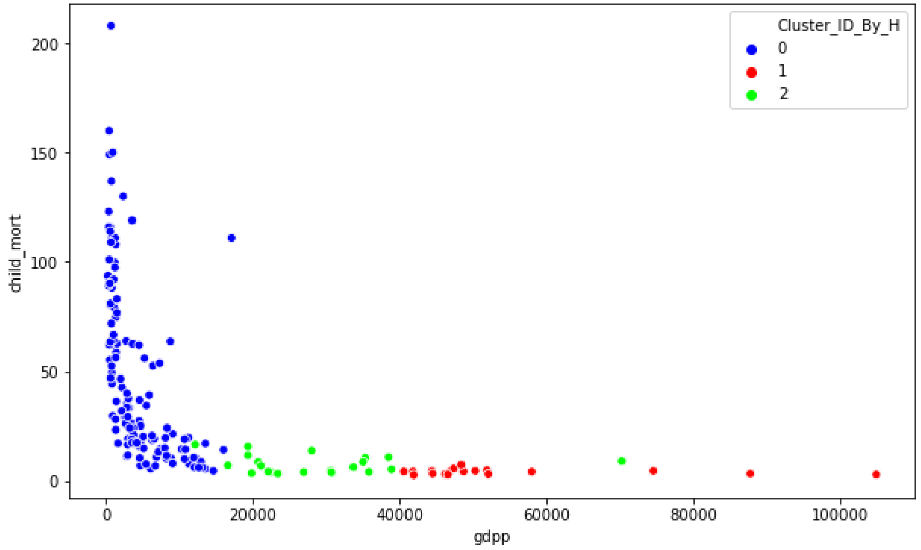
- From above dendrogram we cannot clearly get the clusters.
- In above plot some clusters has very few values.
- By using single linkage we cannot get a proper cut_tree so we can form meaningful clusters.
- Hence we need to ahead and try Complete Linkage method.

2. Complete Linkage.

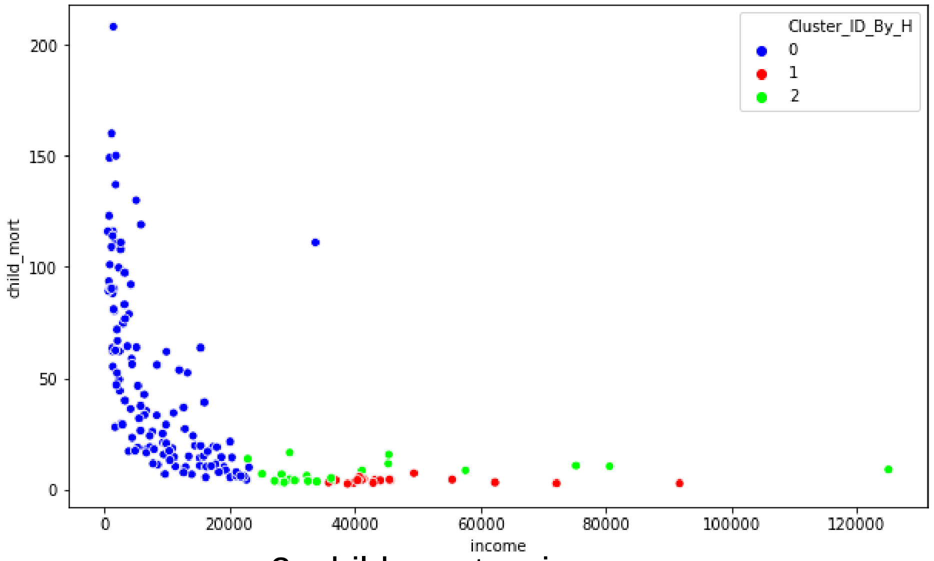


- From above dendrogram we can observe that the complete linkage providing better clusters than single linkage.
- By using above dendrogram we can cut_tree in 3 clusters.

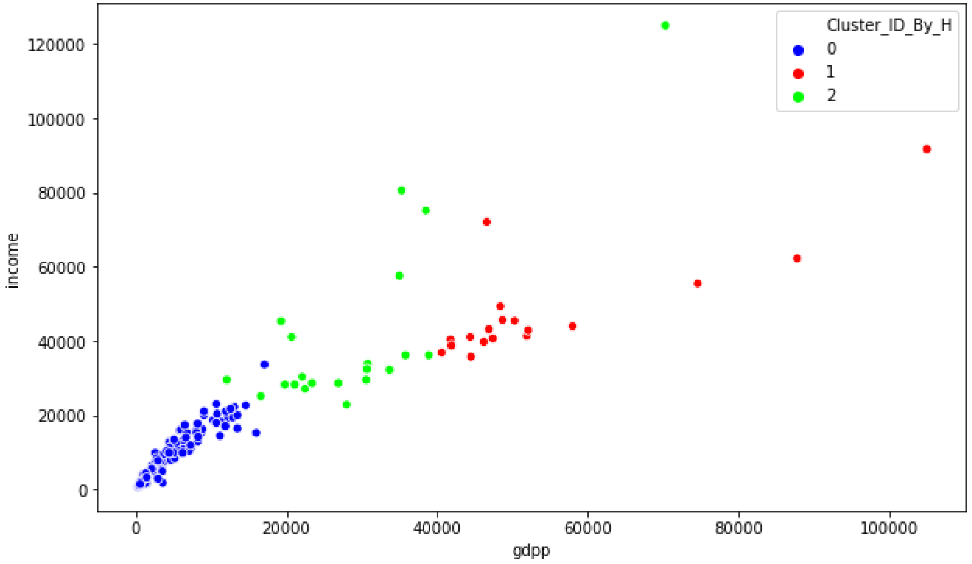
Hierarchical Clustering.



1. child_mort vs gdpp



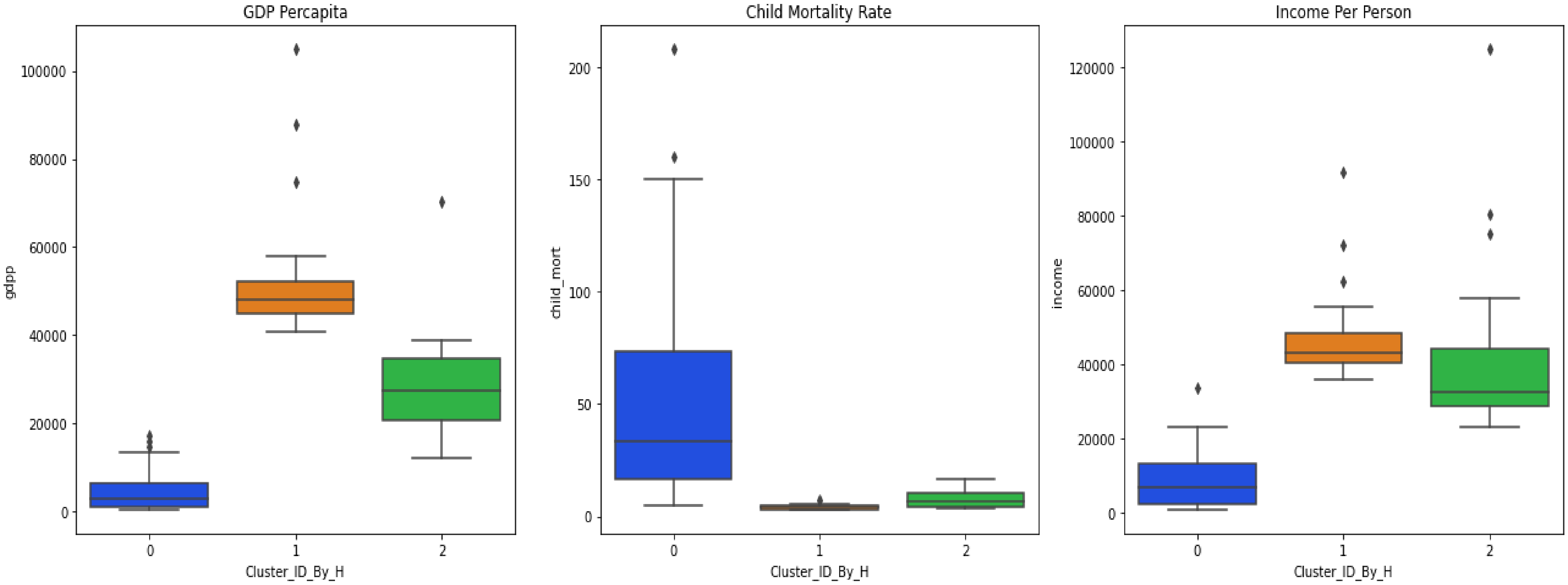
2. child_mort vs income



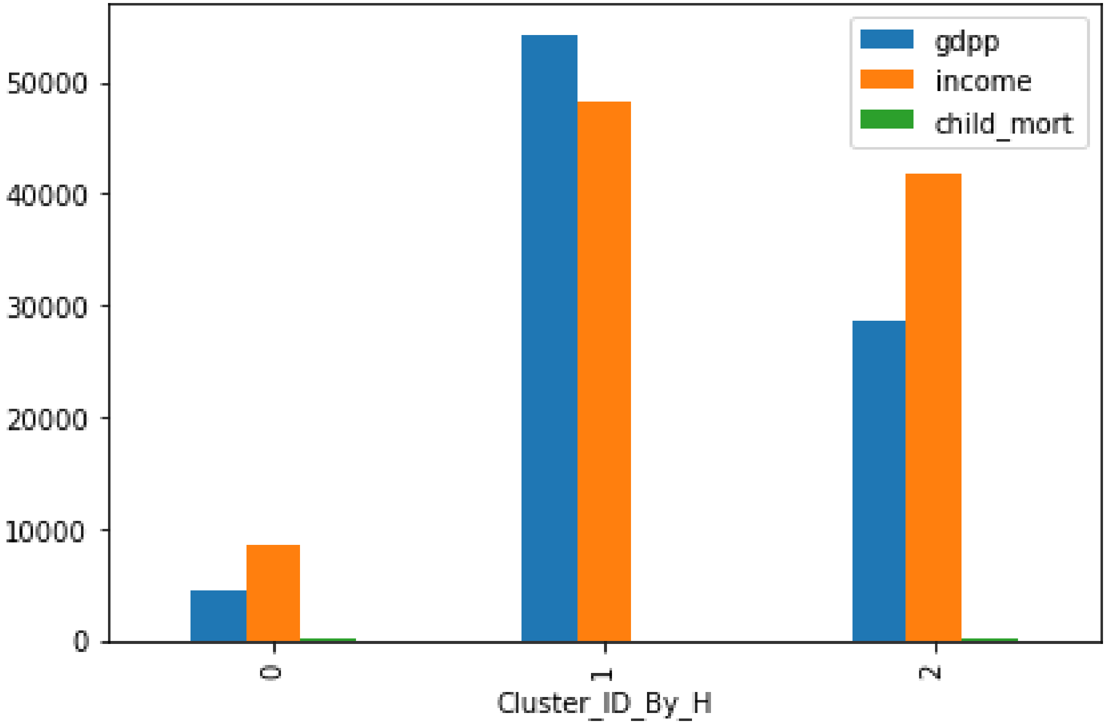
3. income vs gdpp

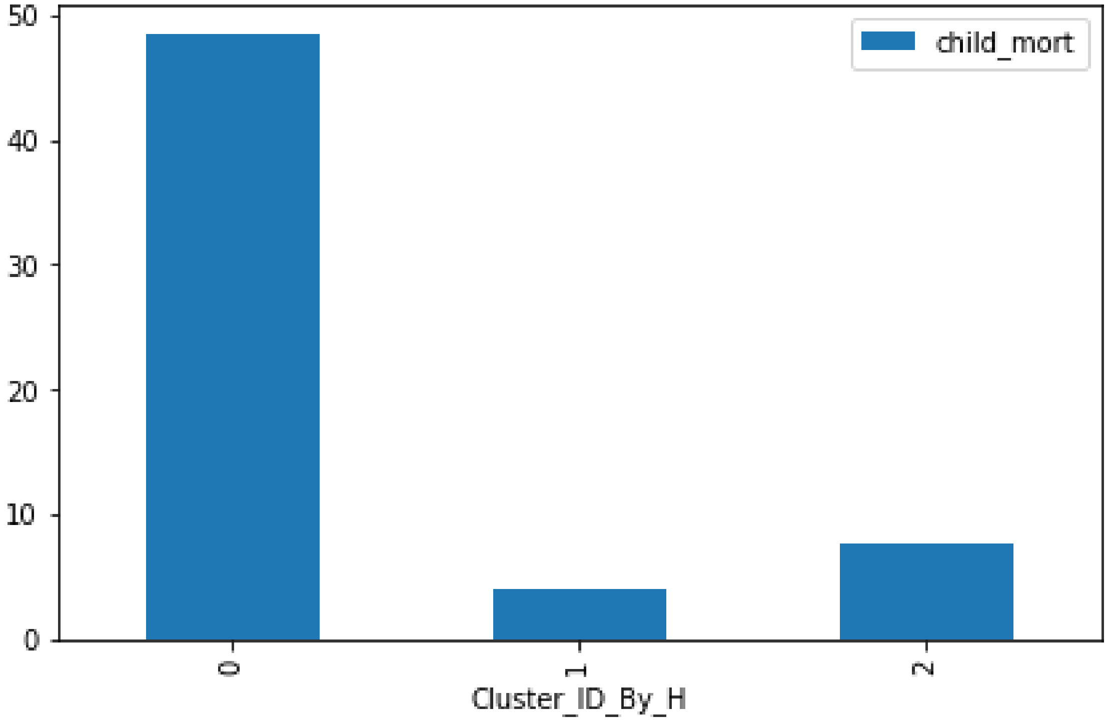
As we can see the we formed 3 cluster showing scatter plot between mainly three features income, GDP and child mortality rate.

gdpp, income, child_mort columns after Hierarchical clustering with its respective clusters.



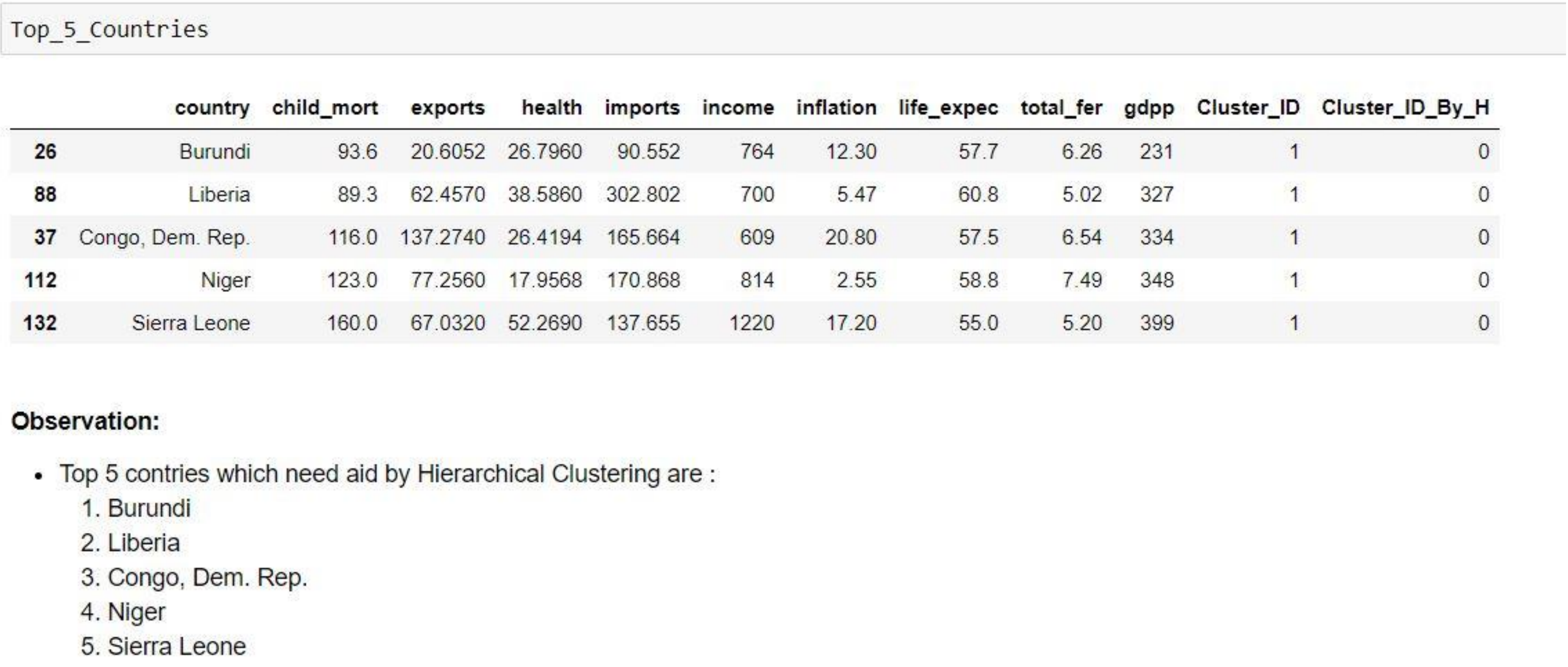
Clustering Profiling after K-means clustering.





- From above plot we can observe that the cluster 0 has highest child mortality rate and lowest income and GDP rate countries.
- So the countries which are needed the aid are belongs from cluster 0.
- From this we can say that three clusters defining the 3 types of countries.

After Hierarchical clustering we get following countries which needed aid.



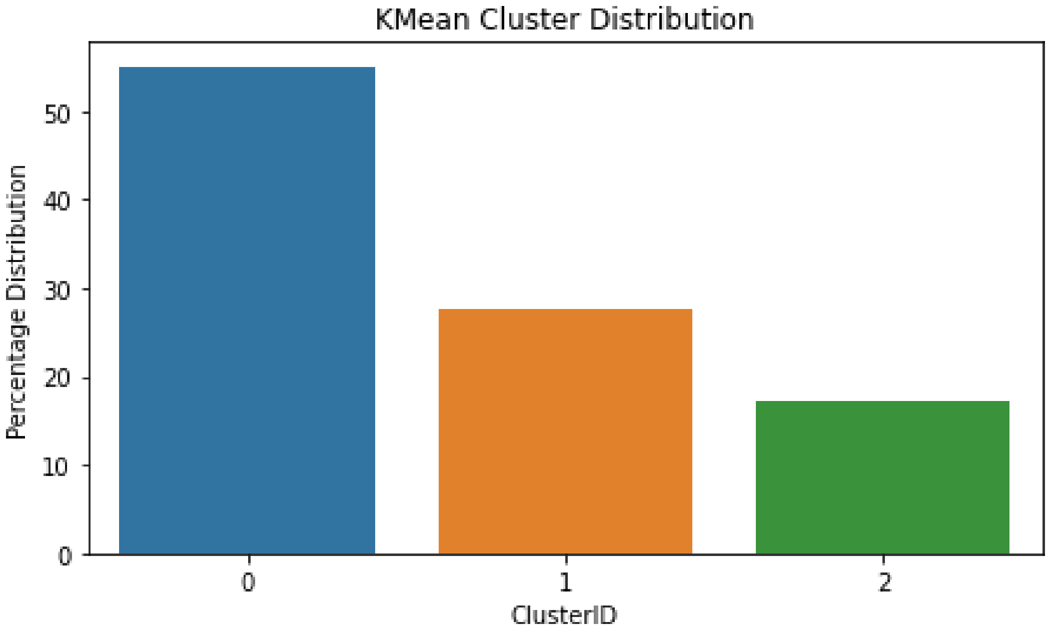
```
Top_5_Countries
```

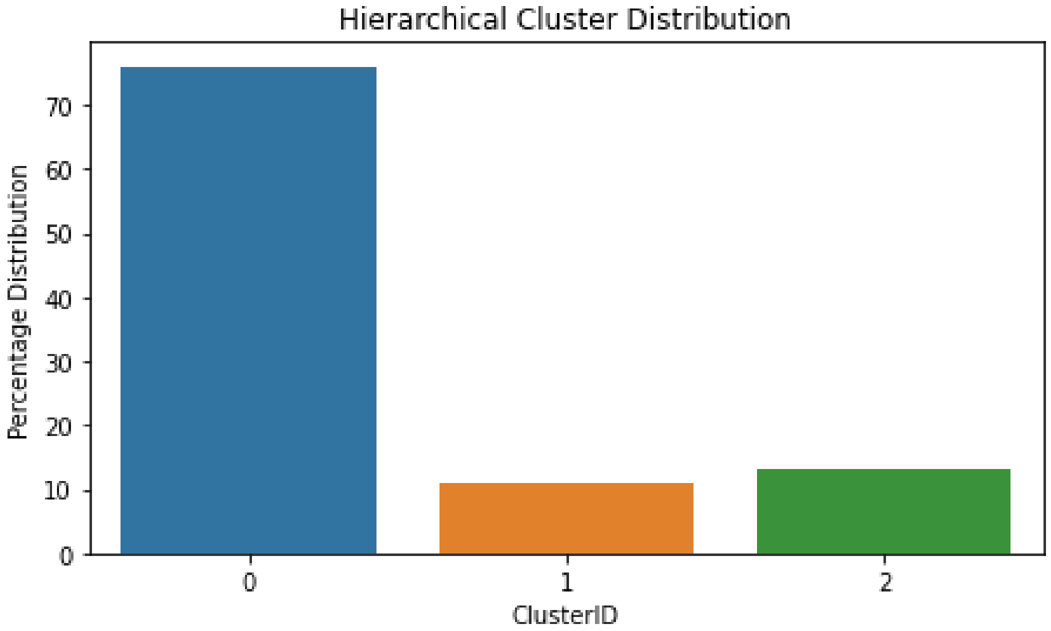
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Cluster_ID	Cluster_ID_By_H
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	1	0
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	1	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	1	0
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	1	0
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399	1	0

Observation:

- Top 5 contries which need aid by Hierarchical Clustering are :
  1. Burundi
  2. Liberia
  3. Congo, Dem. Rep.
  4. Niger
  5. Sierra Leone

Comparison of cluster formed by both K-means and Hierarchical Clustering.

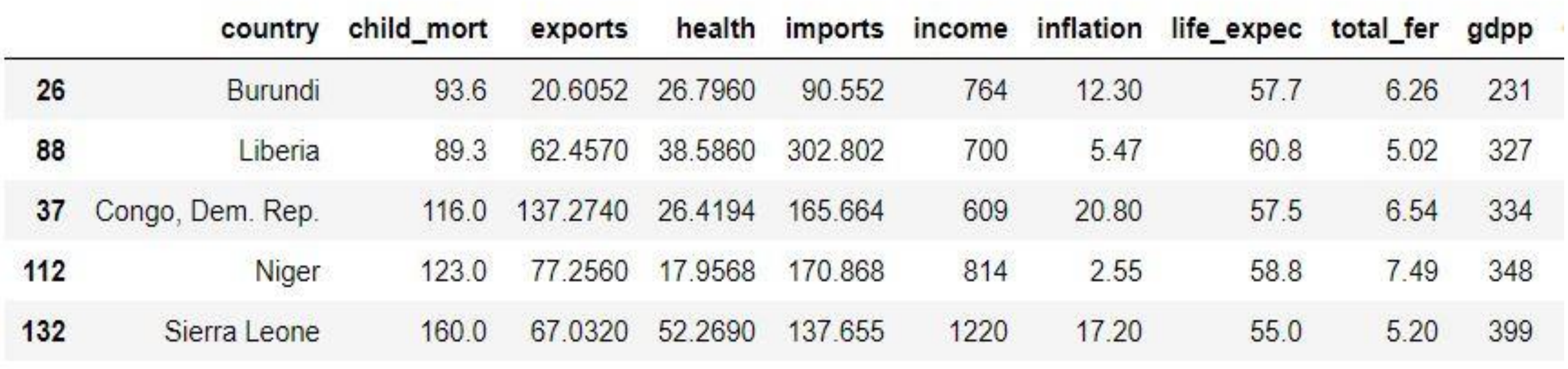




- Above plot shows the percentage of values in every cluster by both K-means and Hierarchical Clustering Method.
- From above plot we can observe that the K-means method providing better cluster classification for this case.
- Since both K-means and Hierarchical methods providing the same countries which needed aid both method providing good result.
- But for overall result the K-means result is better in this case.

Summary

- As by both K-means and Hierarchical Clustering method – we have got same countries which requires aid as a result.
- The following are the countries which are needed aid by considering economic factor into consideration:



	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