Suggestion Of Countries To HELP International

RECOMMENDATION LIST OF BACKWARD COUNTRIES THOSE WHO ARE IN DIREST NEED OF AID

BY USING DIFFERENT CLUSTERING TECHNIQUES

Submitted by: SOURAV KARMAKAR

PROBLEM STATEMENT:

HELP International, an International organization needs suggestion of countries that are in direst need of aid by assessing some socio-economic factors of all the countries. The countries need to be categorised using some socio-economic and health factors determining the overall development of the countries. We need to suggest the countries which the CEO needs to focus on the most for providing the aid.

ANALYSIS APPROACH:

The data has been analyzed following the below approach:

1. The data is read and understood and cleaned if required
2. Data is prepared for creating the required model for categorizing
3. Creating the model
4. Final analysis
5. Recommendation

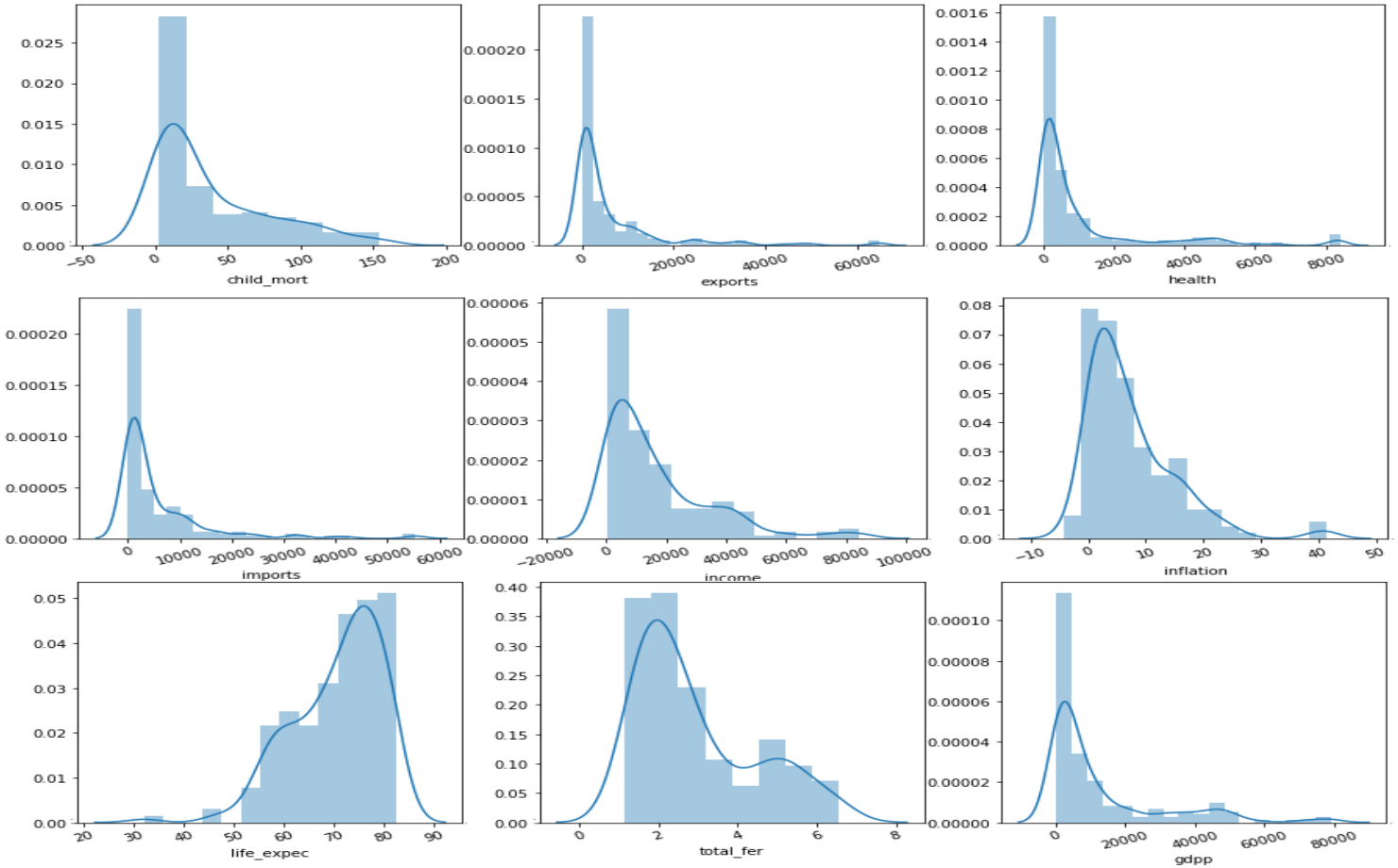
CLEANING THE DATA:

By looking into the dataset we found that there is no missing data in any of the fields. However, we found some higher values in some of the fields. Though these higher values are actual data and not any wrong entry, we cap them to a certain soft limit at the higher end so that these higher values do not effect our analysis.

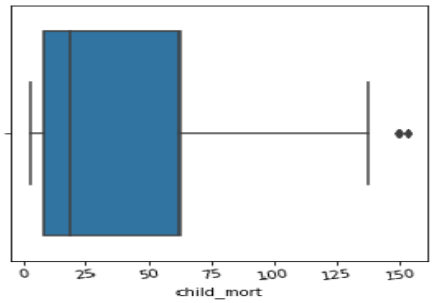
UNDERSTANDING THE DATA(1):

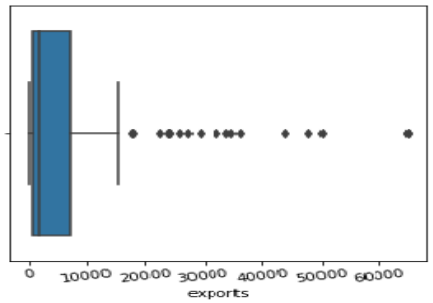
The plots shows the distribution of the data of socio-economic and health related fields.

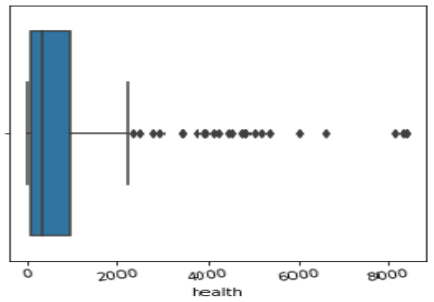
It gives us a clear idea that the data can be segmented into three different segments.

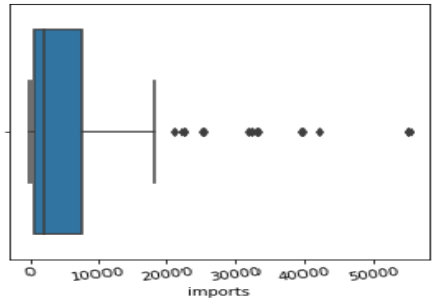


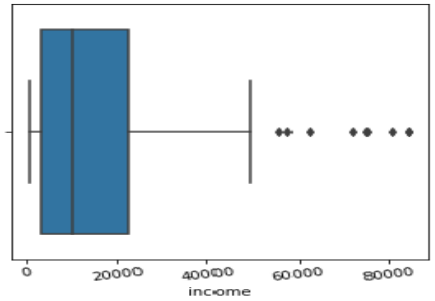
UNDERSTANDING THE DATA(2):

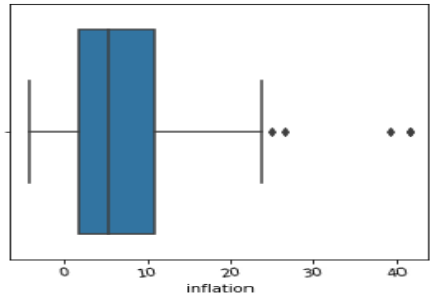


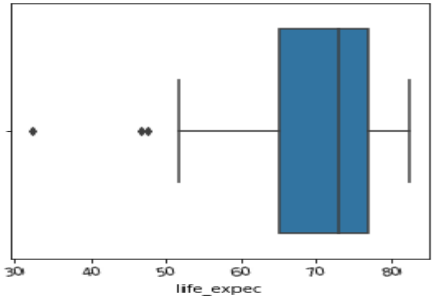


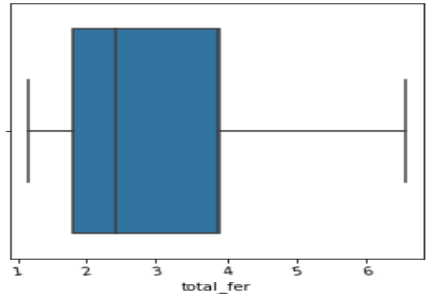


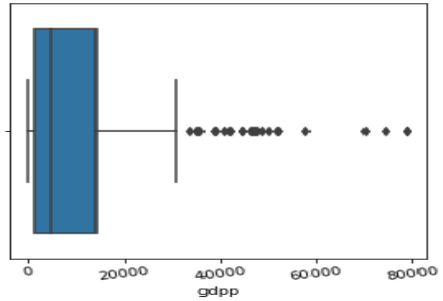












These box plots for each numeric fields show that some of them contains outliers.

However, we do not remove them since these are actual data and not any mistake. Also these fields with high values indicate the developed countries.

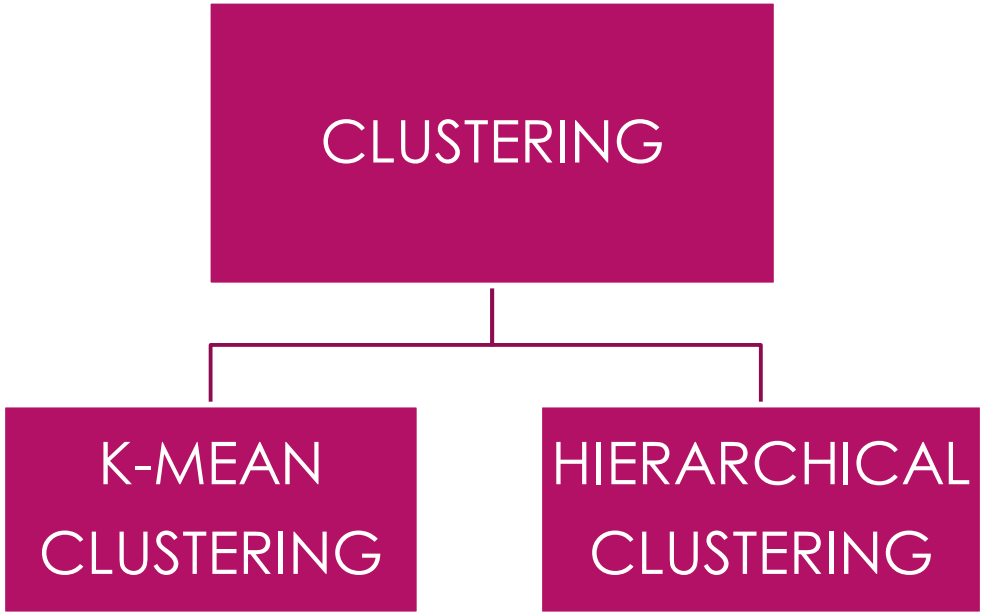
PREPERATION OF THE DATA

The data is scaled or standardized.

After standardizing the, the data set is measured for the clustering tendency using Hopkin's Test which returned a good score indicating clustering is possible for the given data set.

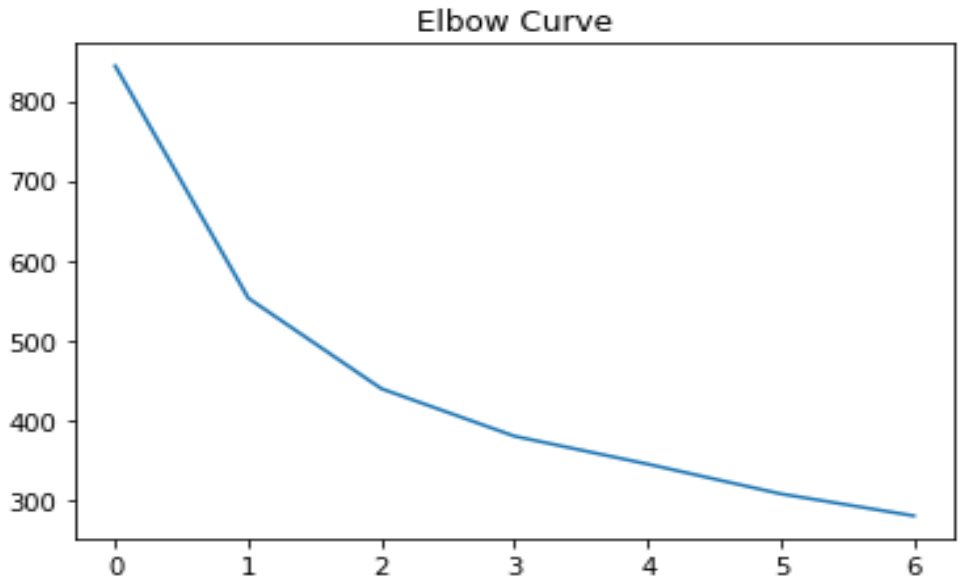
CREATING THE MODEL

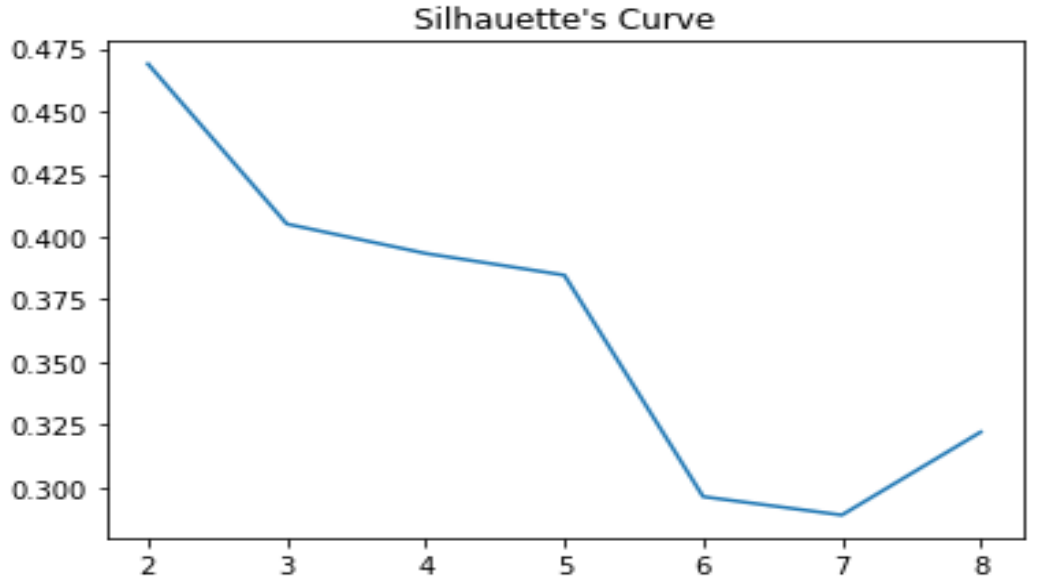
Countries can be categorized or clustered in two different ways of clustering techniques:



K-MEAN CLUSTERING

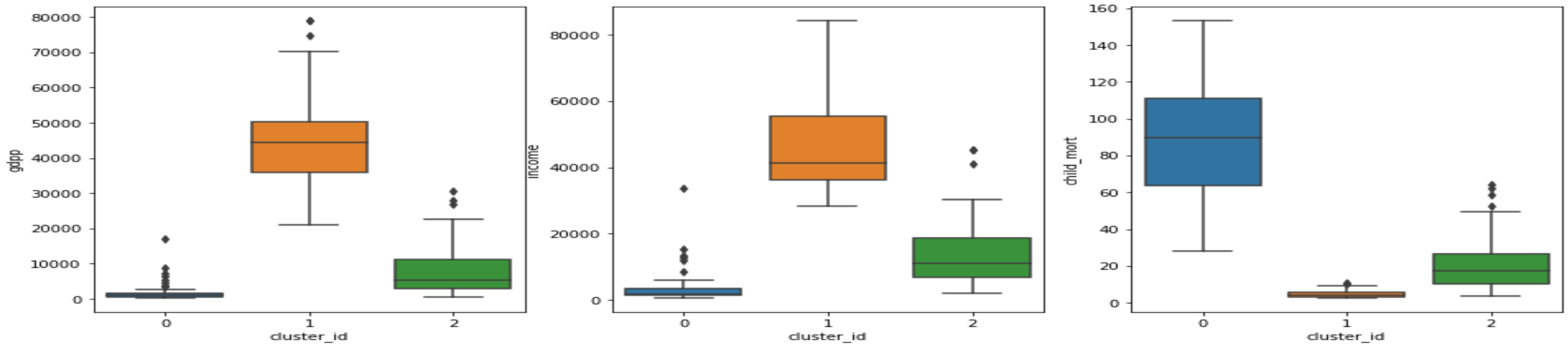
Deciding the number of Clusters using Elbow Curve and Silhouette's Curve:





The optimal number of cluster has been decided as 3 after analyzing the above curves.

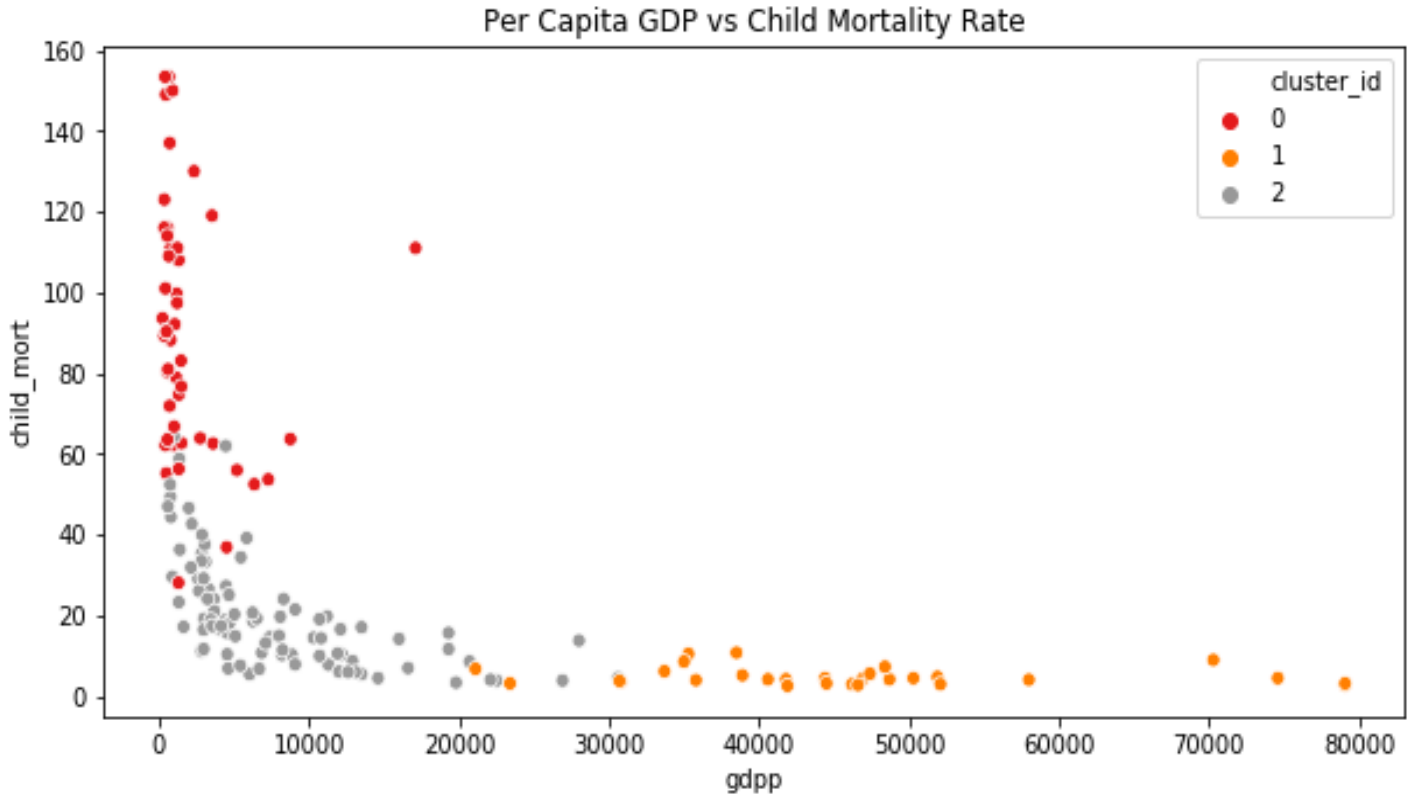
K-MEAN CLUSTERING: Analyzing three clusters



From the above three plots, we can easily distinguish the three clusters.

- Cluster 0 contains those countries whose GDP/capita and income are the lowest but the child mortality rate is highest.
- Cluster 1 contains those countries whose GDP/capita and income are very high and the child mortality rate is very low.
- Cluster 2 contains those countries whose GDP/capita and income are moderate and also the child mortality rate is moderate.

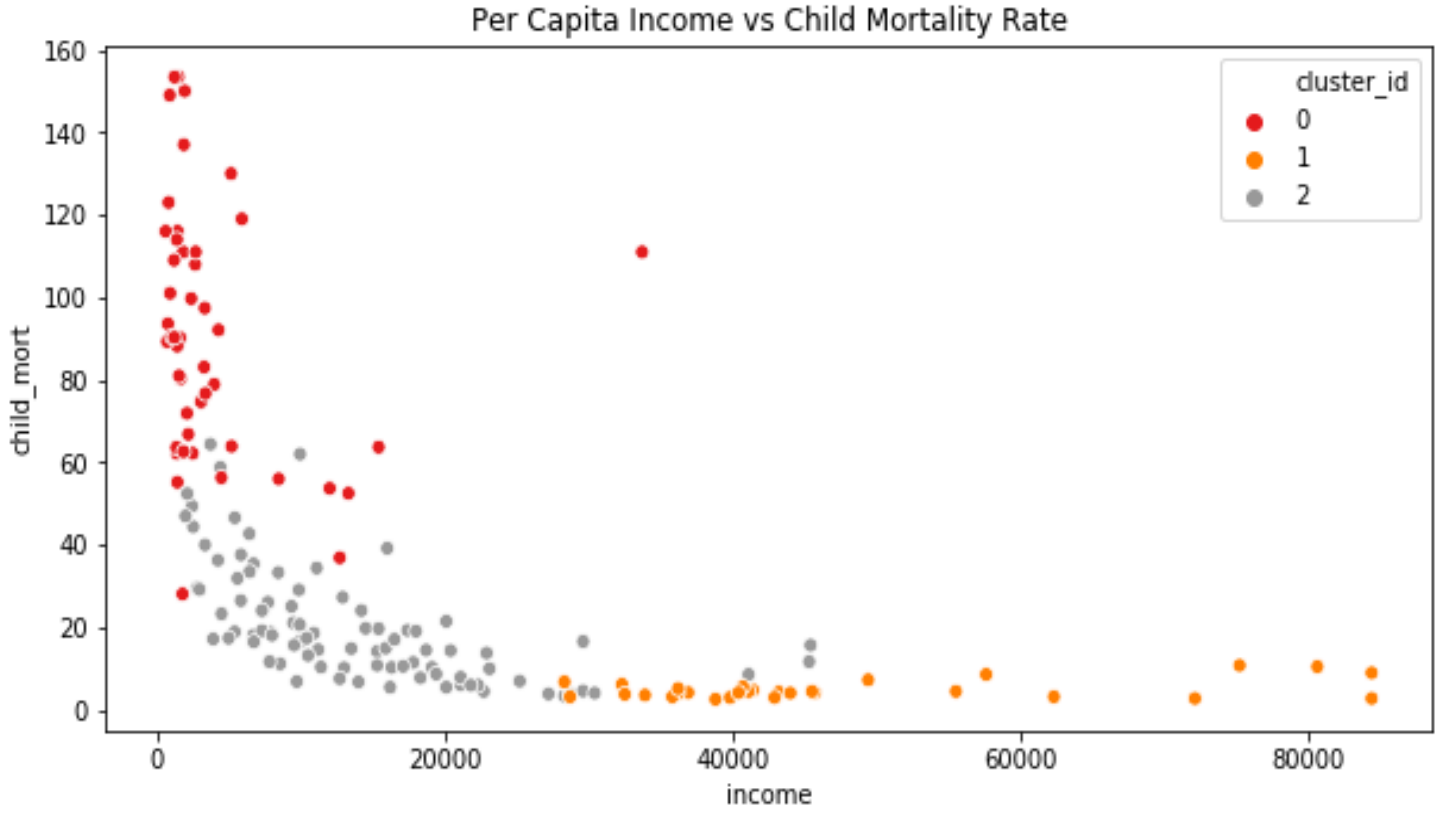
K-MEAN CLUSTERING: Analyzing three clusters



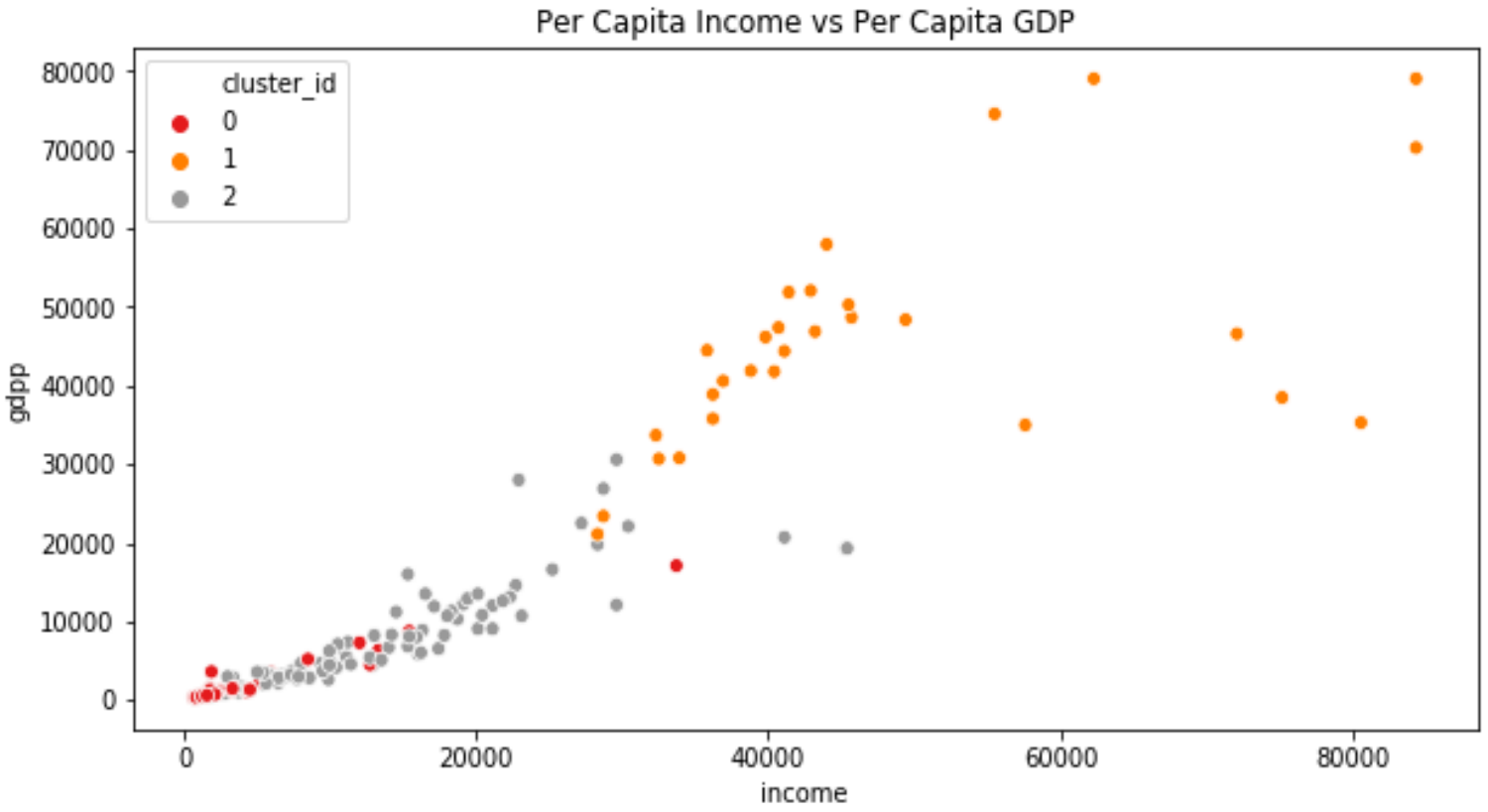
From the plot between per capita GDP and child mortality rate, we can conclude that the cluster 0 contains countries with lower GDP/capita but higher child mortality rate. Cluster 1 contains countries with higher GDP/capita but lower child mortality rate. Cluster 2 contains countries with moderate GDP/capita and moderate child mortality rate.

K-MEAN CLUSTERING: Analyzing three clusters

From the plot between per capita income and child mortality rate, we can conclude that the cluster 0 contains countries with lower income/capita but higher child mortality rate. Cluster 1 contains countries with higher income/capita but lower child mortality rate. Cluster 2 contains countries with moderate income/capita and moderate child mortality rate.

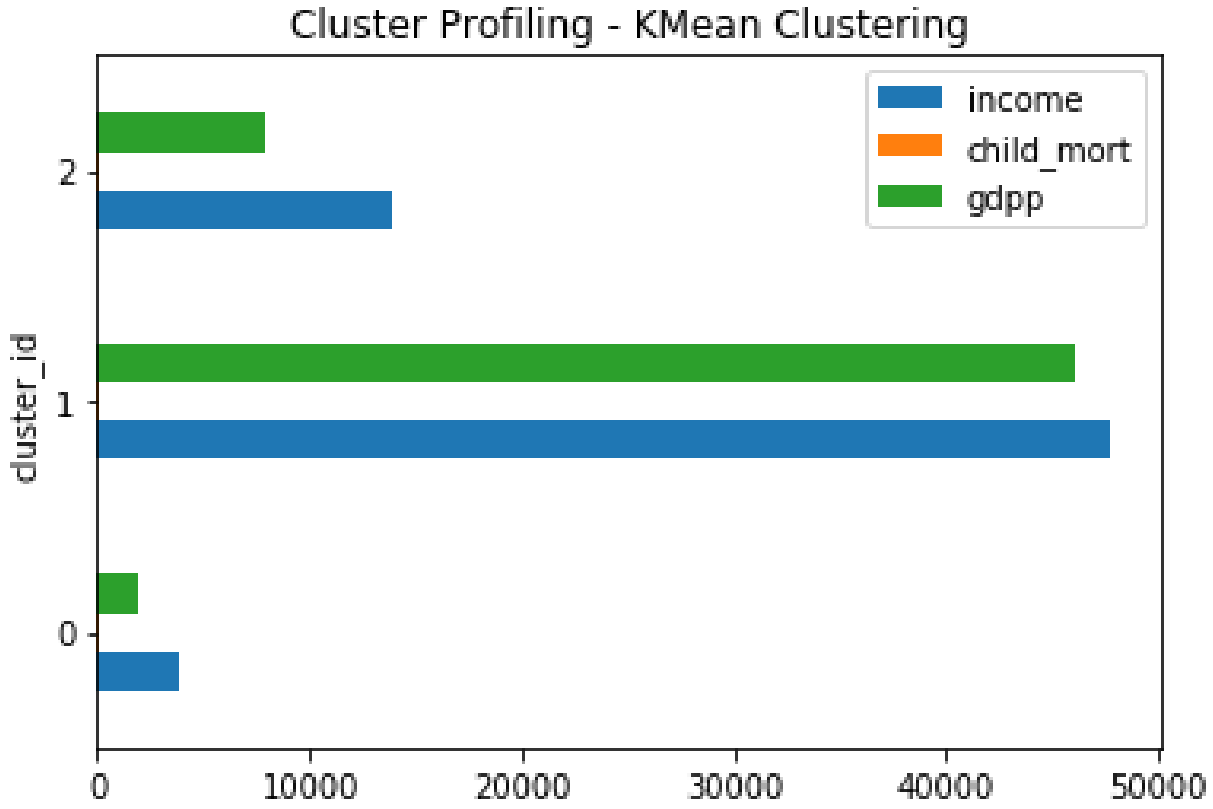


K-MEAN CLUSTERING: Analyzing three clusters



From the plot between per capita income and per capita GDP, we can conclude that the cluster 0 contains countries with lower per capita income and hence lower per capita GDP. Cluster 2 contains countries with moderate per capita income and moderate per capita GDP. Cluster 1 contains countries with higher per capita income and higher per capita GDP.

K-MEAN CLUSTERING: Cluster Profiling

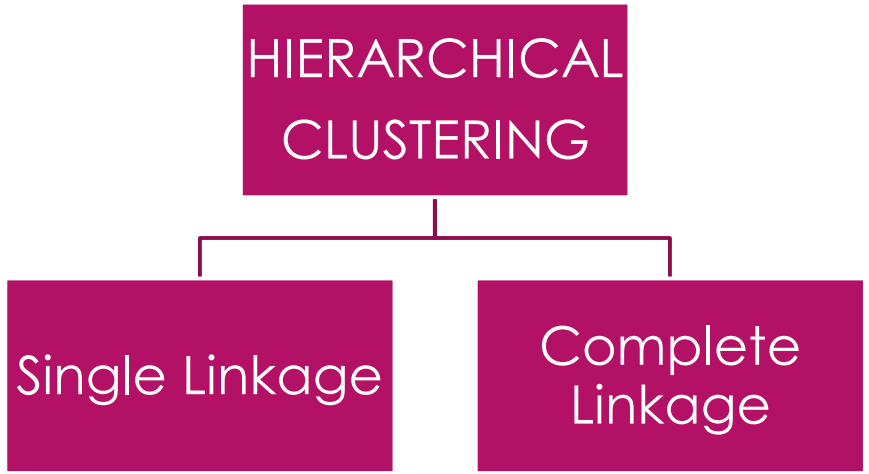


From the plot, we can conclude that the cluster 1 contains the developed countries, cluster 2 contains the developing countries and cluster 0 contains the underdeveloped countries that need to be provided with Aid.

HIERARCHICAL CLUSTERING

In Hierarchical Clustering, we used both the methods to analyze the data and to decide the number of clusters:

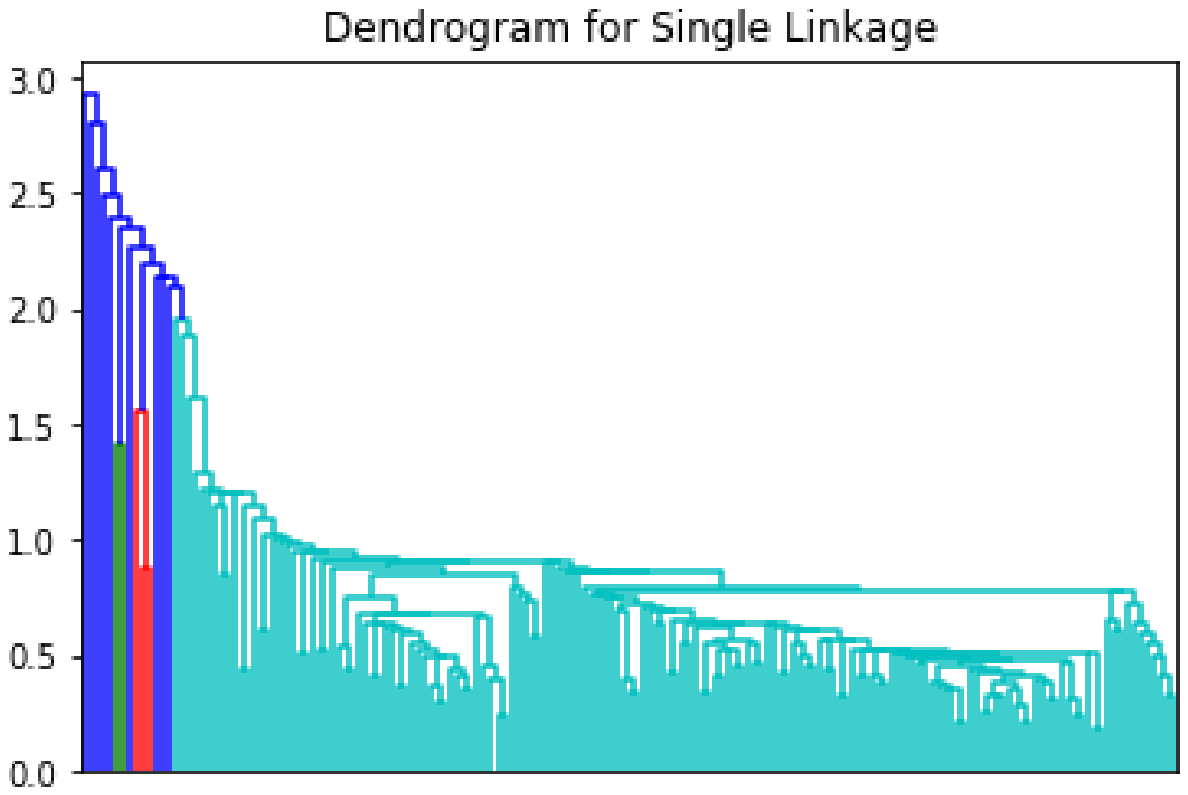
- Single Linkage Method
- Complete Linkage Method



HIERARCHICAL CLUSTERING: Single Linkage

The dendrogram for Hierarchical Clustering Single linkage method is not interpretable and hence cannot be decided the number of clusters to be chosen.

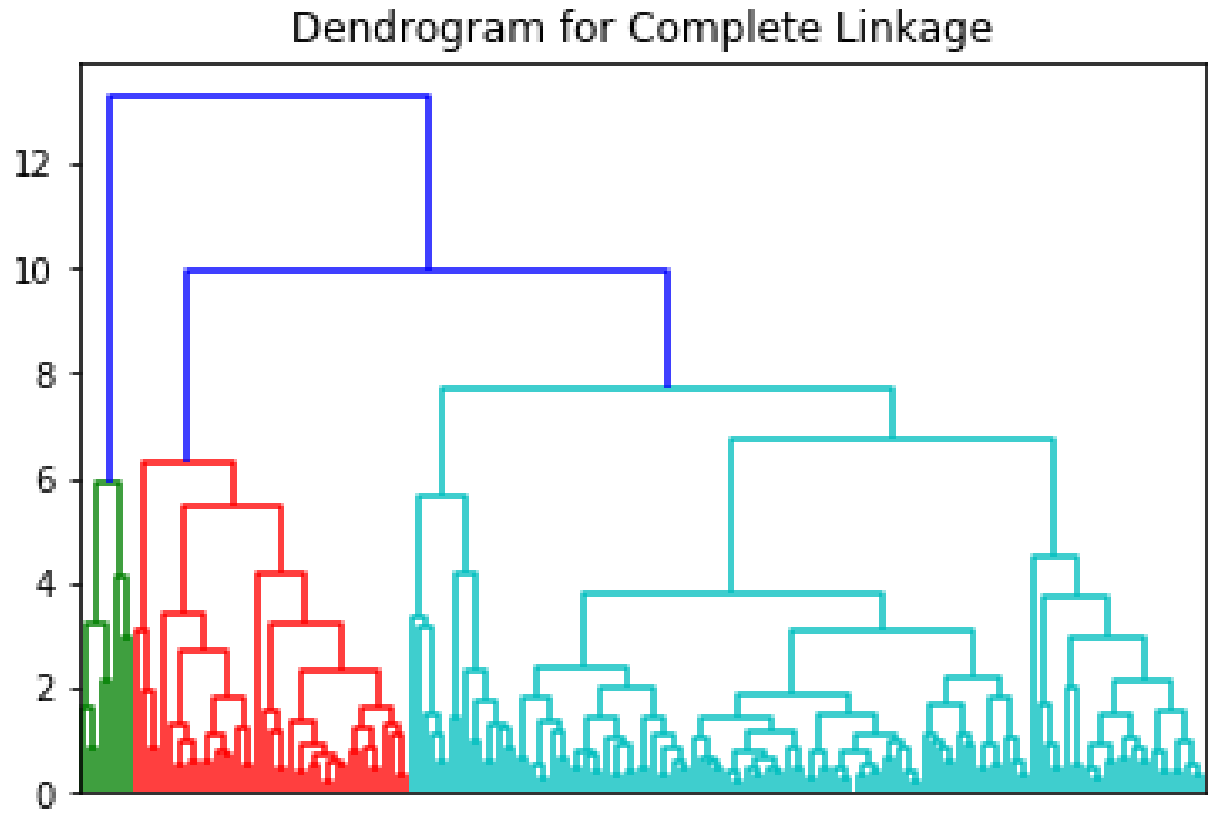
Hence, we try Hierarchical Clustering with complete linkage method.



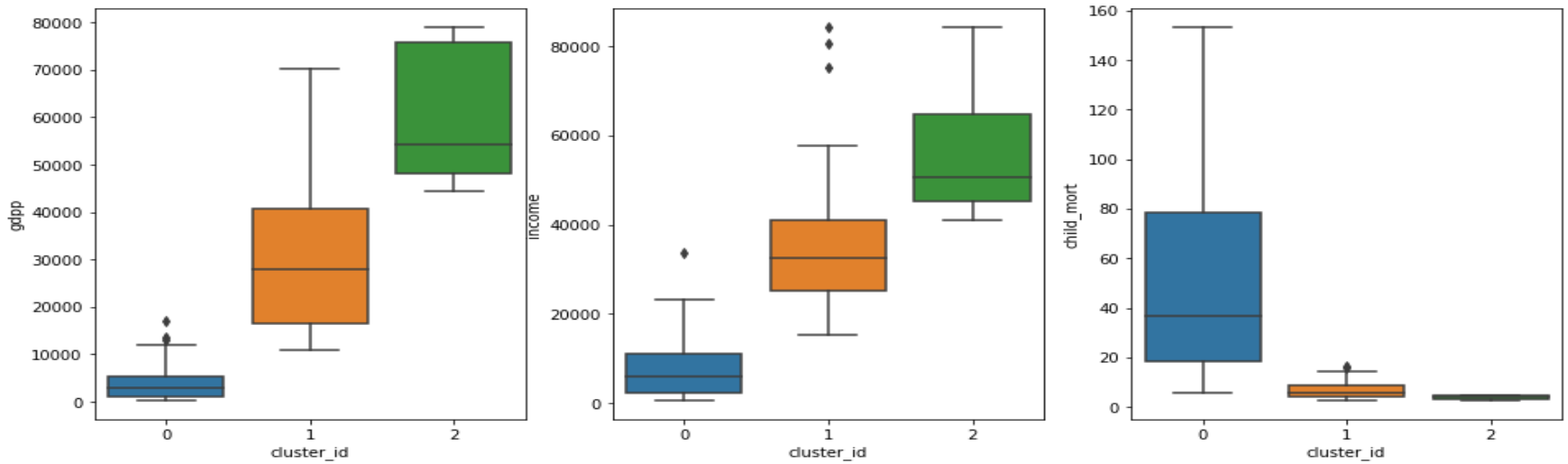
HIERARCHICAL CLUSTERING: Complete Linkage

The dendrogram for Hierarchical Clustering with Complete Linkage method clearly distinguishes the clusters.

Hence, we choose 3 number of clusters as depicted with color gradient in the diagram.



HIERARCHICAL CLUSTERING: Analyzing three clusters

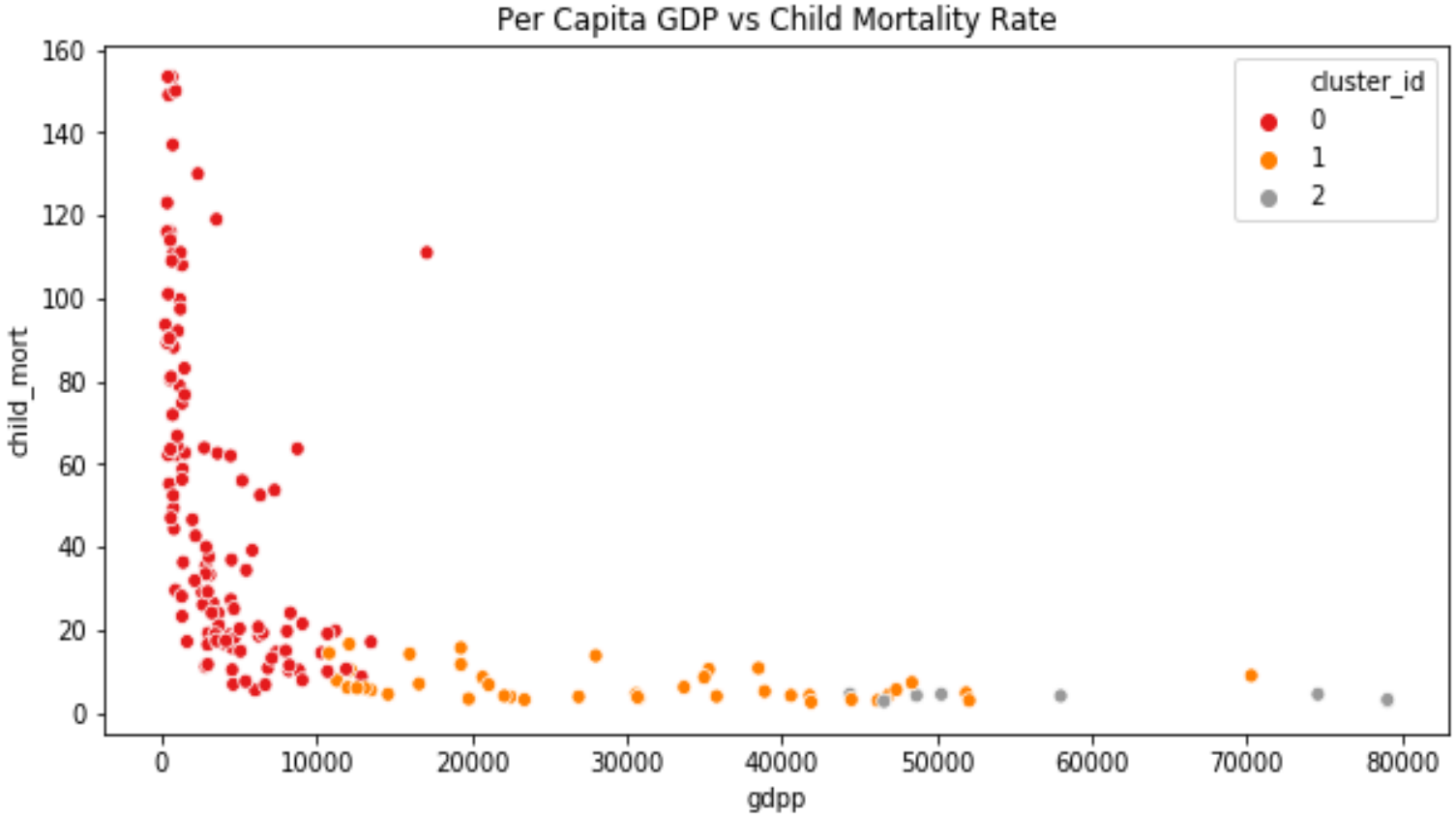


From the above three plots, we can easily distinguish the three clusters.

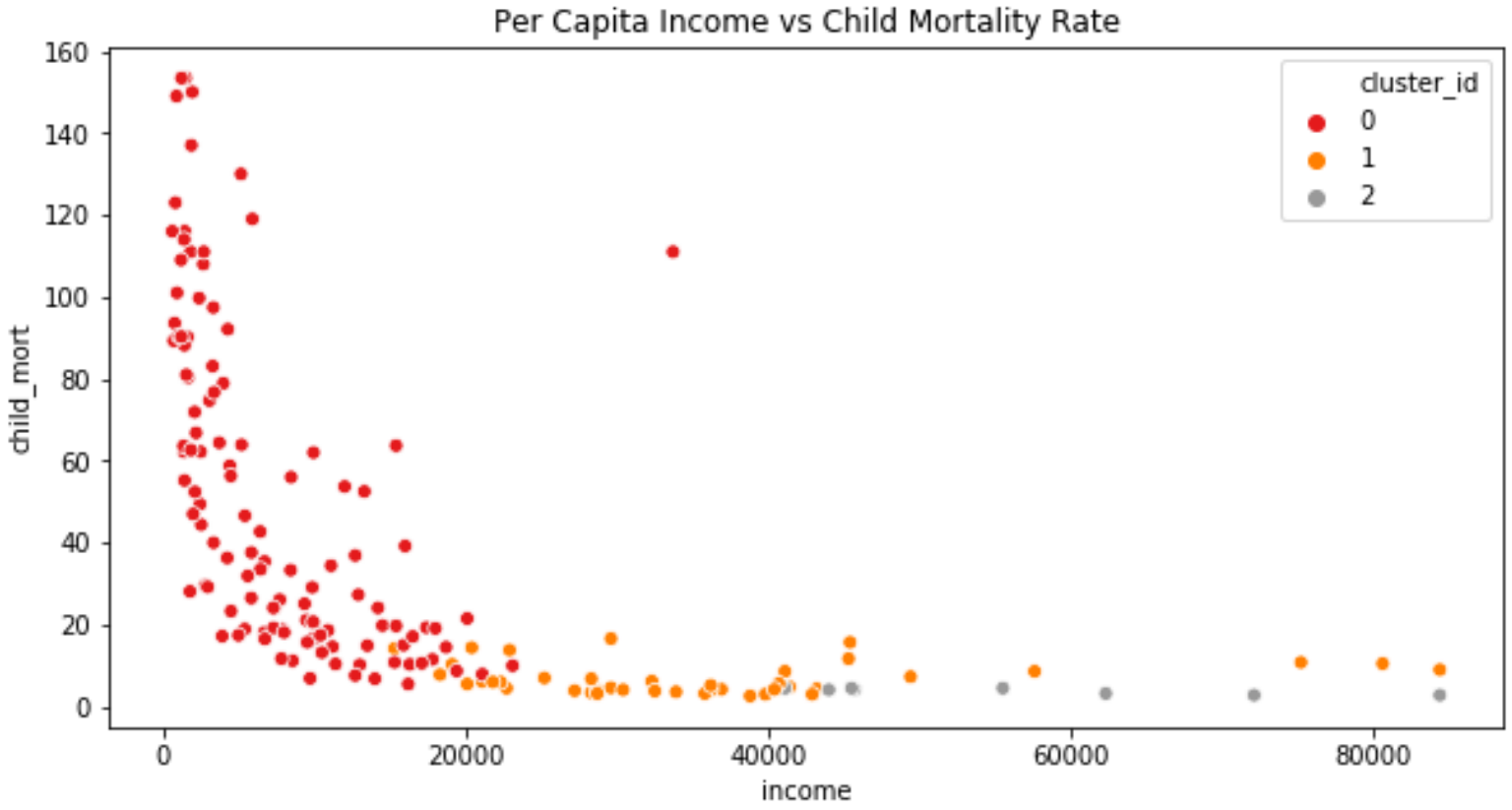
- Cluster 0 contains those countries whose GDP/capita and income are the lowest but the child mortality rate is highest.
- Cluster 1 contains those countries whose GDP/capita and income are moderate and the child mortality rate is very lower than cluster 0.
- Cluster 2 contains those countries whose GDP/capita and income are very high but the child mortality rate is lowest among all clusters.

HIERARCHICAL CLUSTERING: Analyzing three clusters

From the plot between per capita GDP and child mortality rate, we can conclude that the cluster 0 contains countries with lower GDP/capita but higher child mortality rate. Cluster 2 contains countries with higher GDP/capita but lower child mortality rate. Cluster 1 contains countries with moderate GDP/capita and moderate child mortality rate.



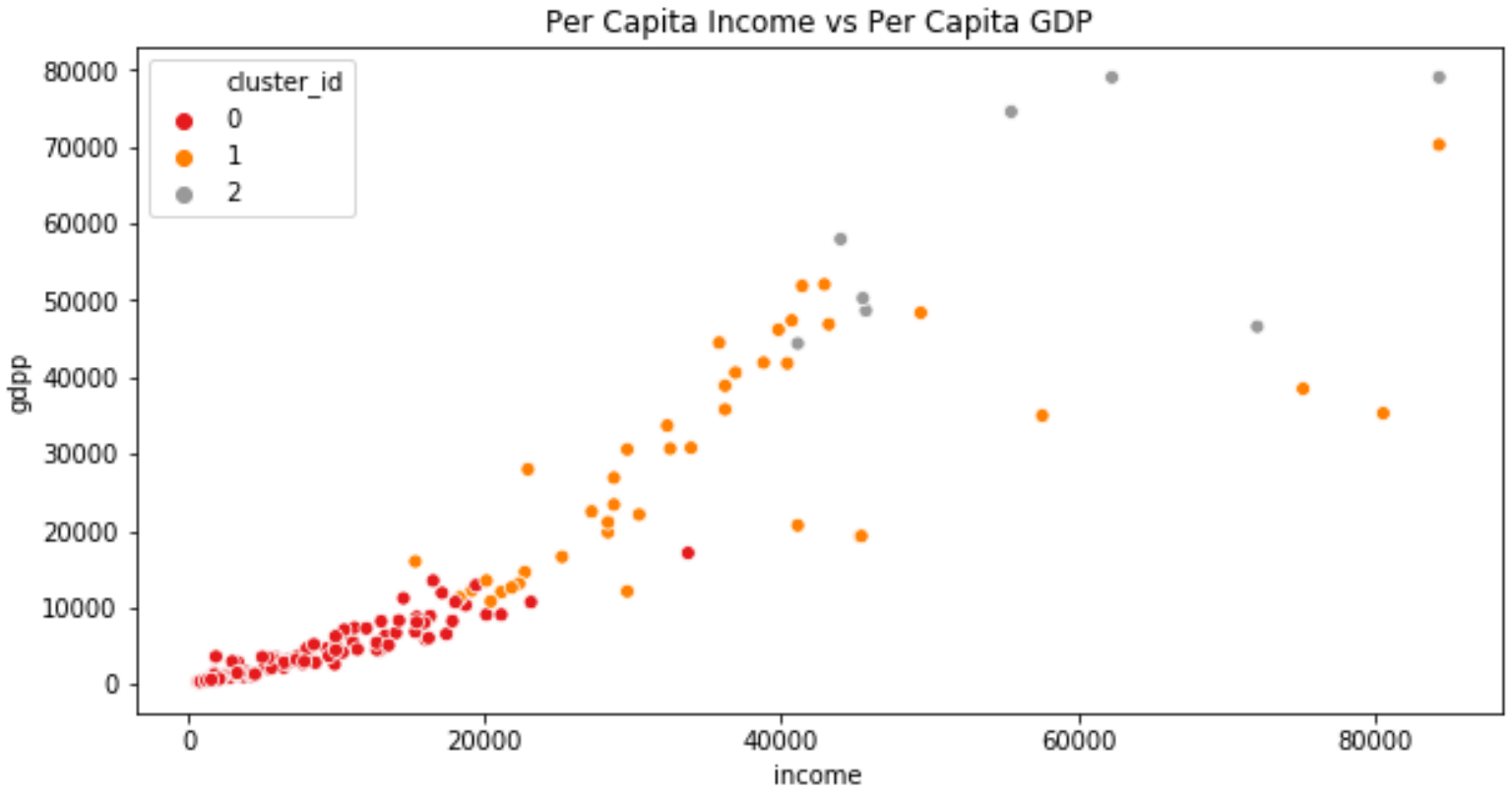
HIERARCHICAL CLUSTERING: Analyzing three clusters



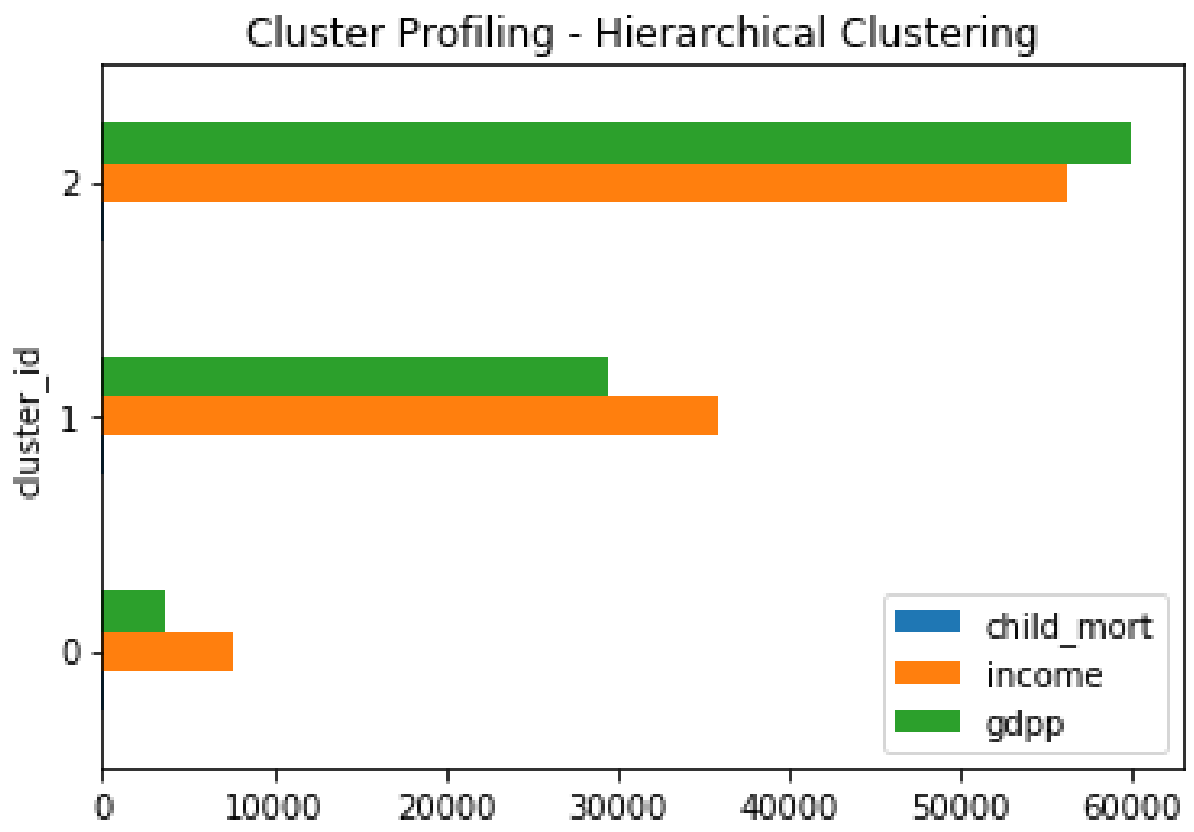
From the plot between per capita income and child mortality rate, we can conclude that the cluster 0 contains countries with lower income/capita but higher child mortality rate. Cluster 2 contains countries with higher income/capita but lower child mortality rate. Cluster 1 contains countries with moderate income/capita and moderate child mortality rate.

HIERARCHICAL CLUSTERING: Analyzing three clusters

From the plot between per capita income and per capita GDP, we can conclude that the cluster 0 contains countries with lower per capita income and hence lower per capita GDP. Cluster 1 contains countries with moderate per capita income and moderate per capita GDP. Cluster 2 contains countries with higher per capita income and higher per capita GDP.



HIERARCHICAL CLUSTERING: Cluster Profiling



From the plot, we can conclude that the cluster 2 contains the developed countries, cluster 1 contains the developing countries and cluster 0 contains the underdeveloped countries that need to be provided with Aid.

RECOMMENDATION

After performing both the clustering methods, we choose KMean Clustering method over Hierarchical Clustering method as in Hierarchical method, the inter-cluster similarities are not that prominent and also countries are not evenly distributed over the clusters.

Hence, the result that we finally achieved from KMean clustering is as below:

The top 10 countries that are in direst need for support are:

1. Burundi
2. Liberia
3. Congo, Dem. Rep.
4. Niger
5. Sierra Leone
6. Madagascar
7. Mozambique
8. Central African Republic
9. Malawi
10. Eritrea

Thank You….