HELP International Clustering Assignment

NAVAEEP JAGATHKARI

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.

Problem Approach

- As a Data Analyst, Our job is to categorize the countries with the help of data available, to help the CEO to reach out the right countries which are dire need.
- The data available regarding the countries in the dataset is., Income, GDPP, Child_mortality_rate etc.,
- Steps to come out with solution:
  - 1. Read, clean the data if required and visualize.
  - 2. outliers treatment if required
  - 3. Data processing
  - 4. Data Modelling
    - KMeans
    - Hierarchical Clustering
    - Visualization of clusters
  - 5. Sorting Top Countries

EDA

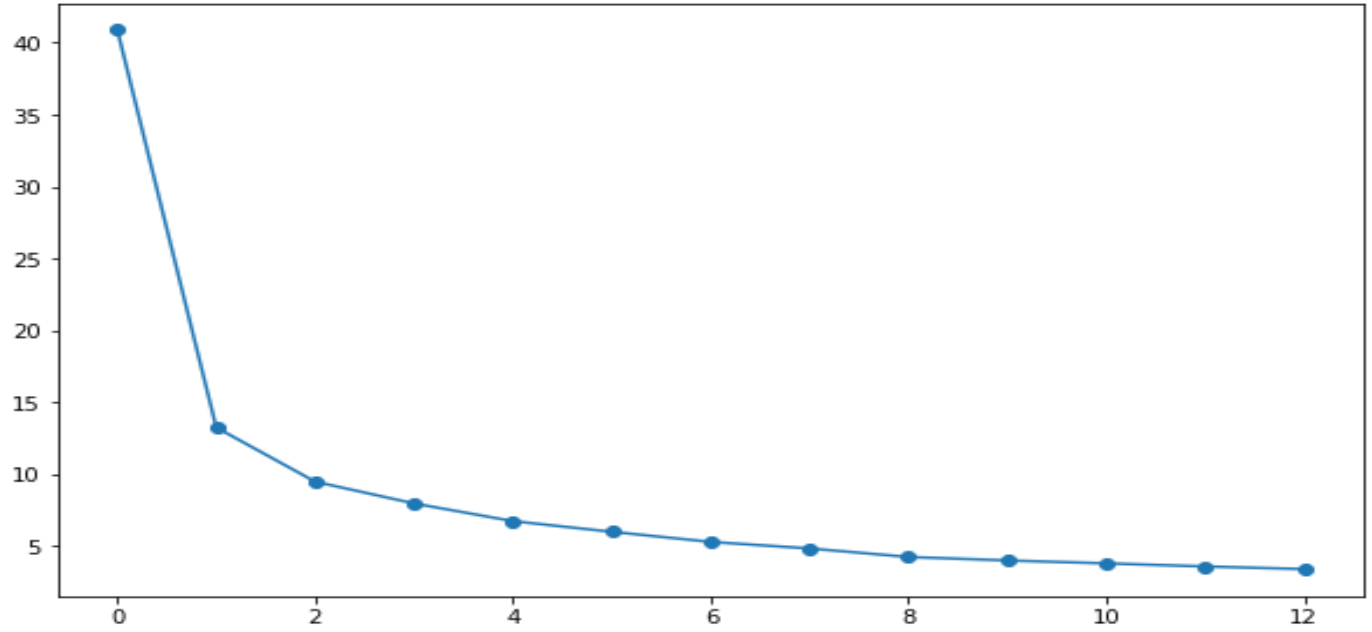
- The dataset had no null-values. So skipped do data cleaning
- In the dataset all variables are in correct datatype. No changes required
- Outliers are present in all variables, Outliers treatment to this dataset can lead to loss of important data. As each row in the dataset represent the data of a country. So outliers treatment is negated.
- In Data processing step, as the data is skewed the StandardScaler will not scale the data as required for further process. I tried doing using StandardScaler, RobustScaler etc.,except with MinMaxScaler I find skewed results in all scaling techniques.

Clustering

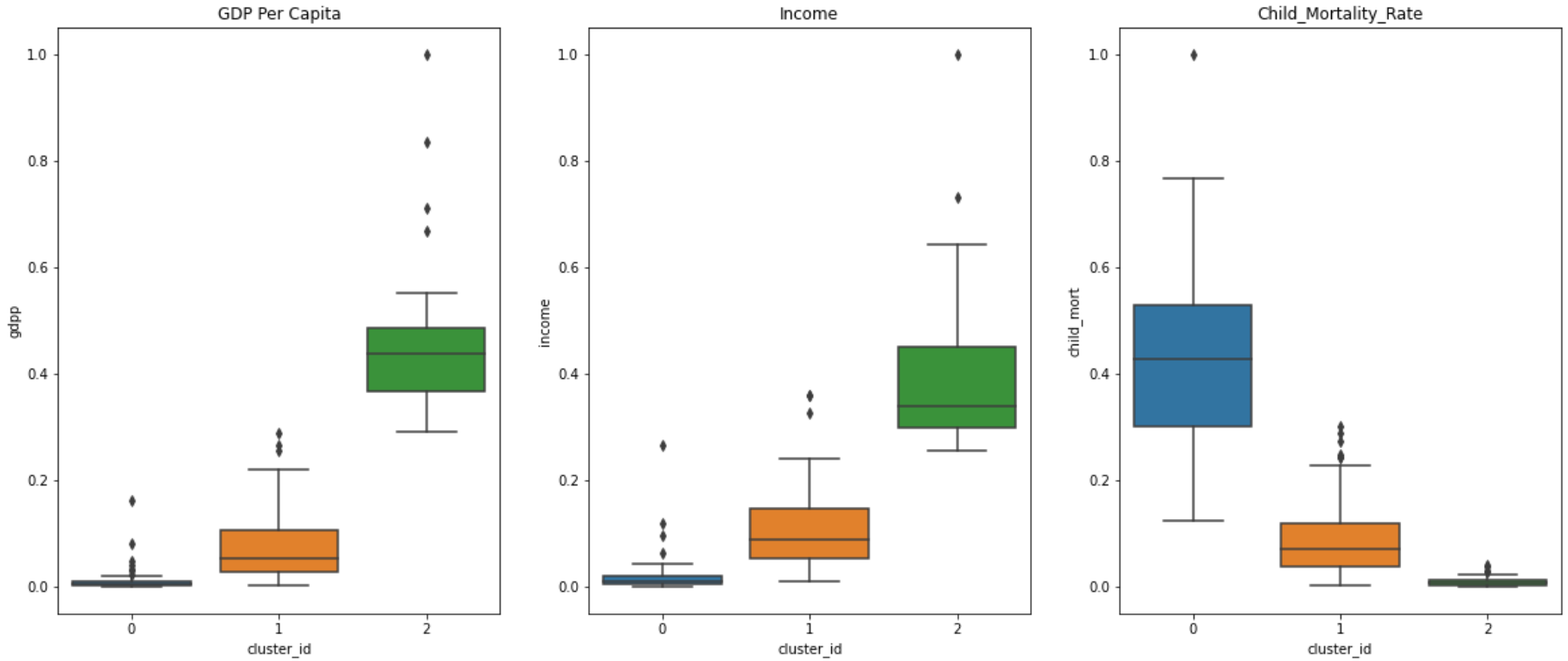
- KMeans
  - First took maximum clusters = 4, after k-means clustering and after plotting elbow-curve came to final number of clusters = 3
  - Used fit_transform approach
  - init = k-means++
  - random_state=0
- Hierarchical
  - Metric is Euclidean(used to find distance)
  - Plotted dendrogram using linkage, of method “single”
  - Again, plotted dendrogram using linkage, of method “complete”
  - We can clearly observe the number of clusters = 3
  - Used cut_tree with n_clusters = 3 to get the labels of the clusters formed.

Elbow Curve

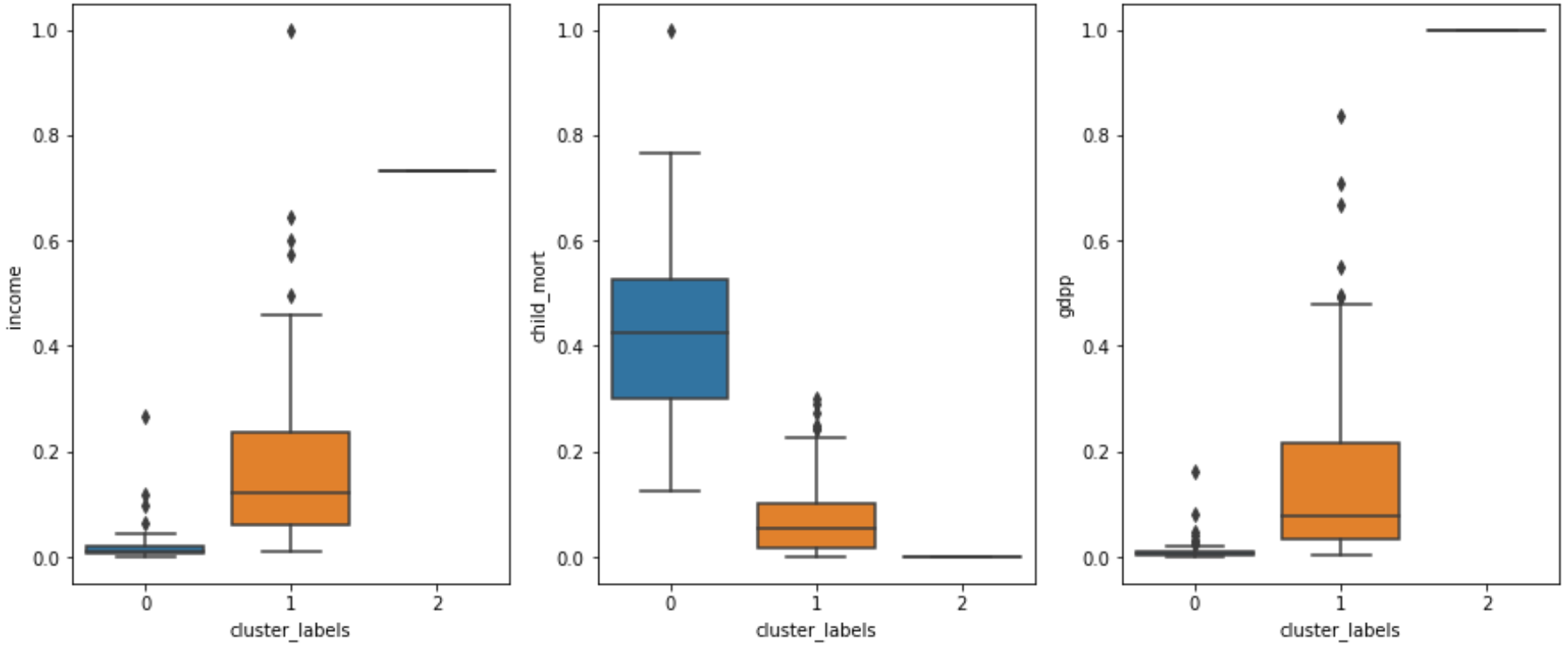
- Elbow curve is used to find the optimal number of clusters are required for the data give.
- Optimal number of clusters can be Identified, If the curve has sudden elbow at point which represents the cluster number.
- From the curve decided to take optimal number of clusters=3



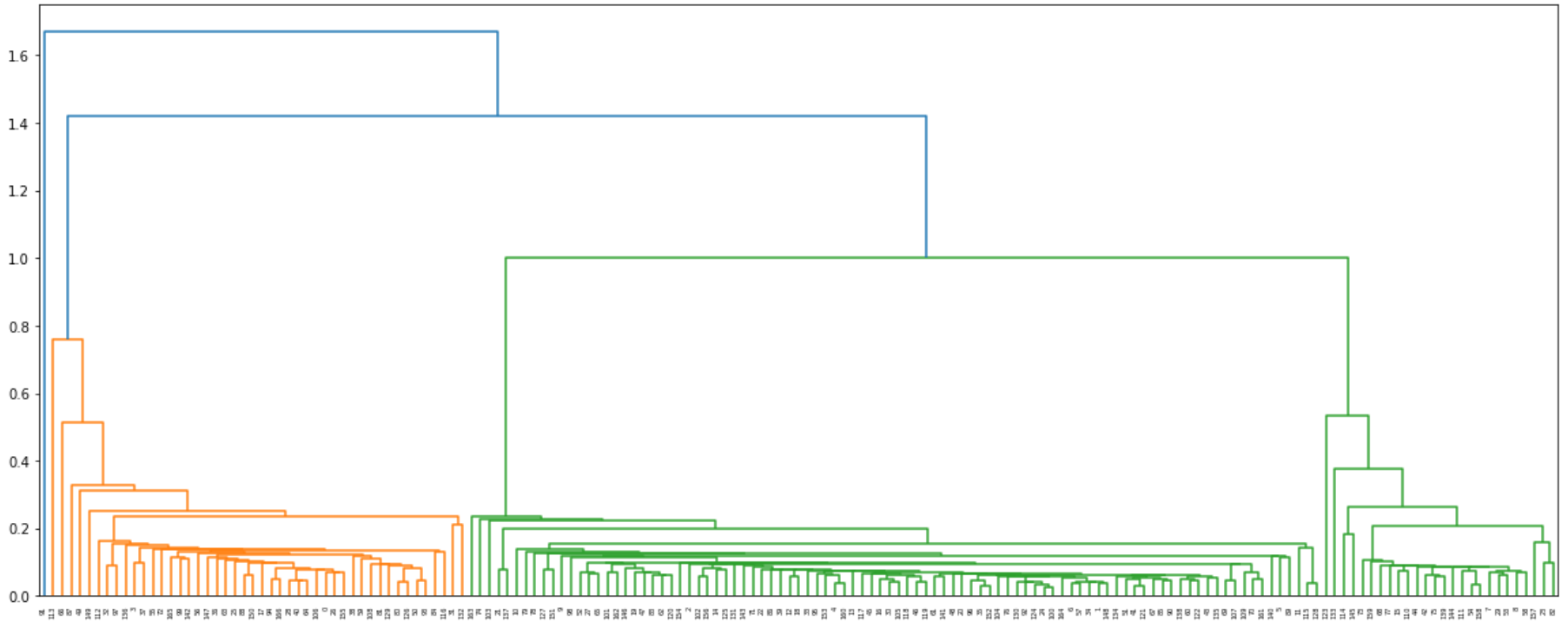
KMeans Clusters plot



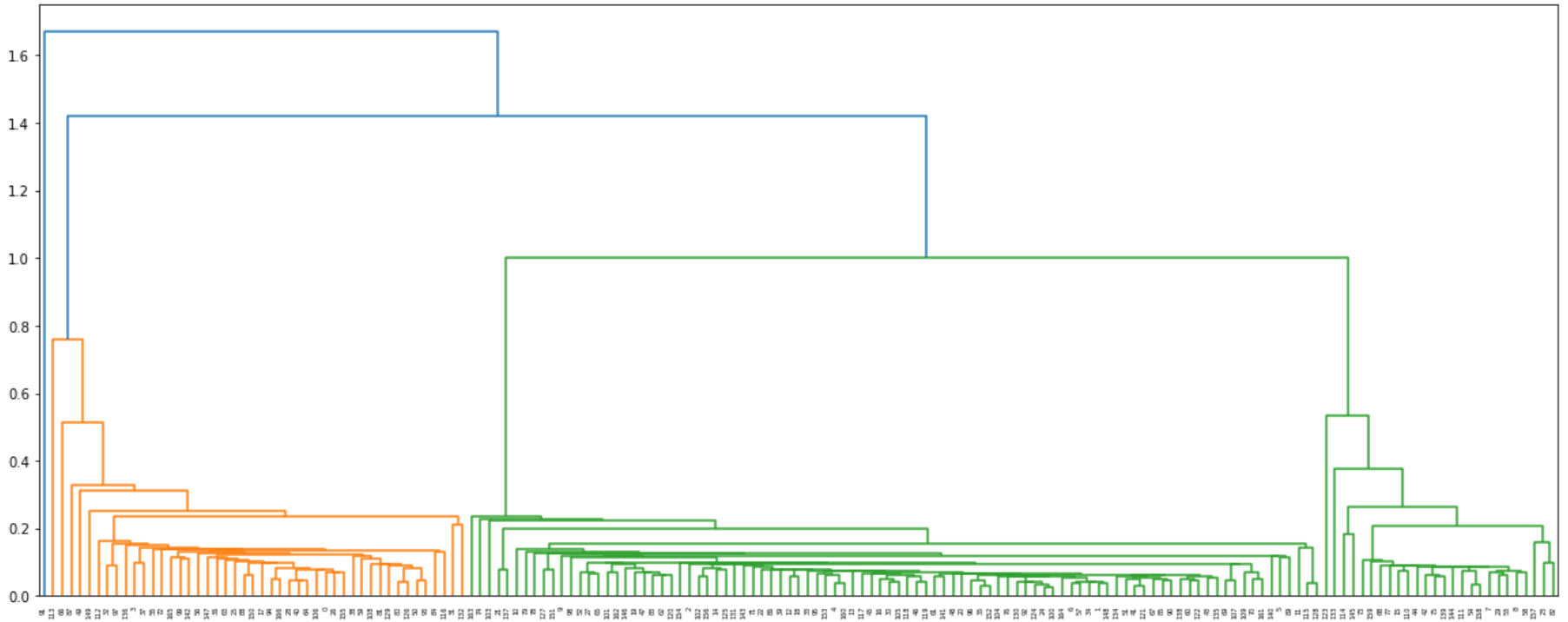
Hierarchical Clusters plot



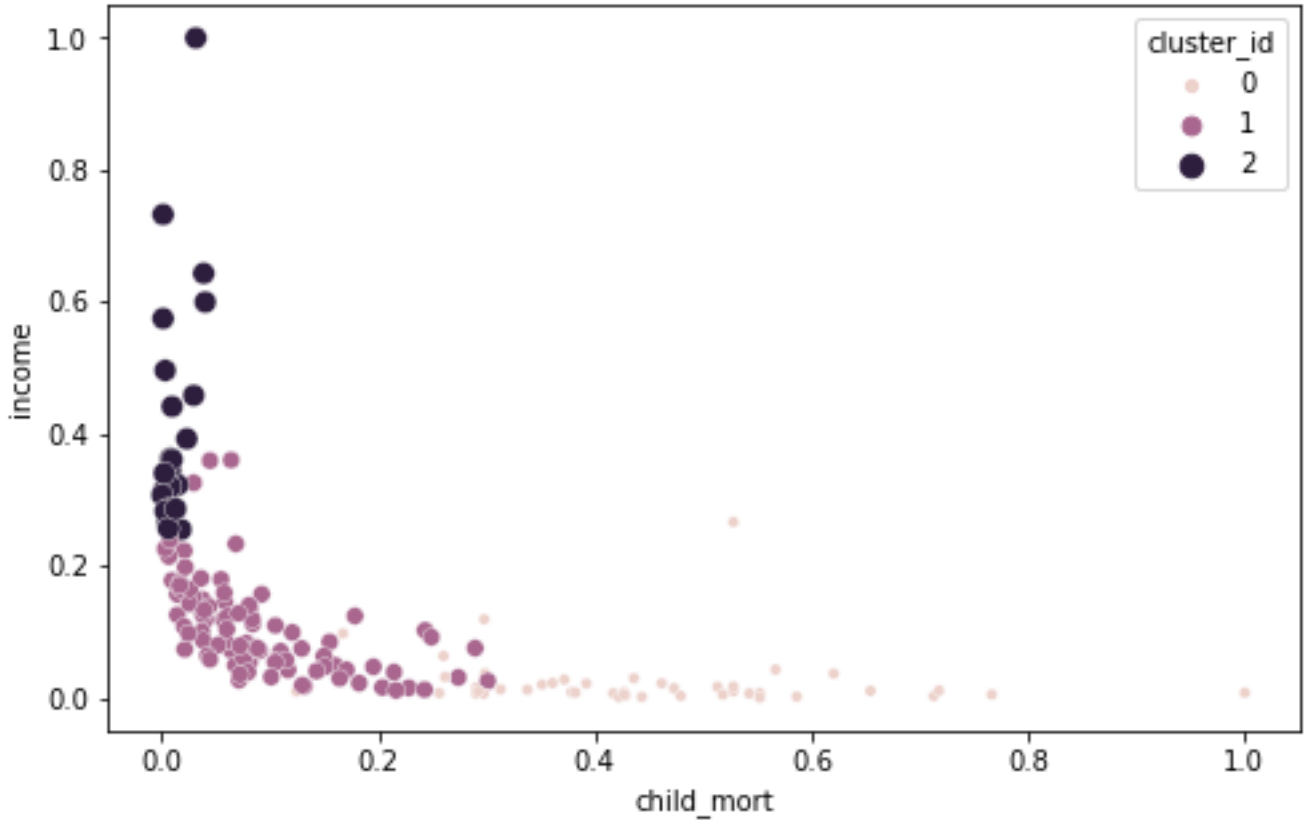
Dendrogram – Single Linkage



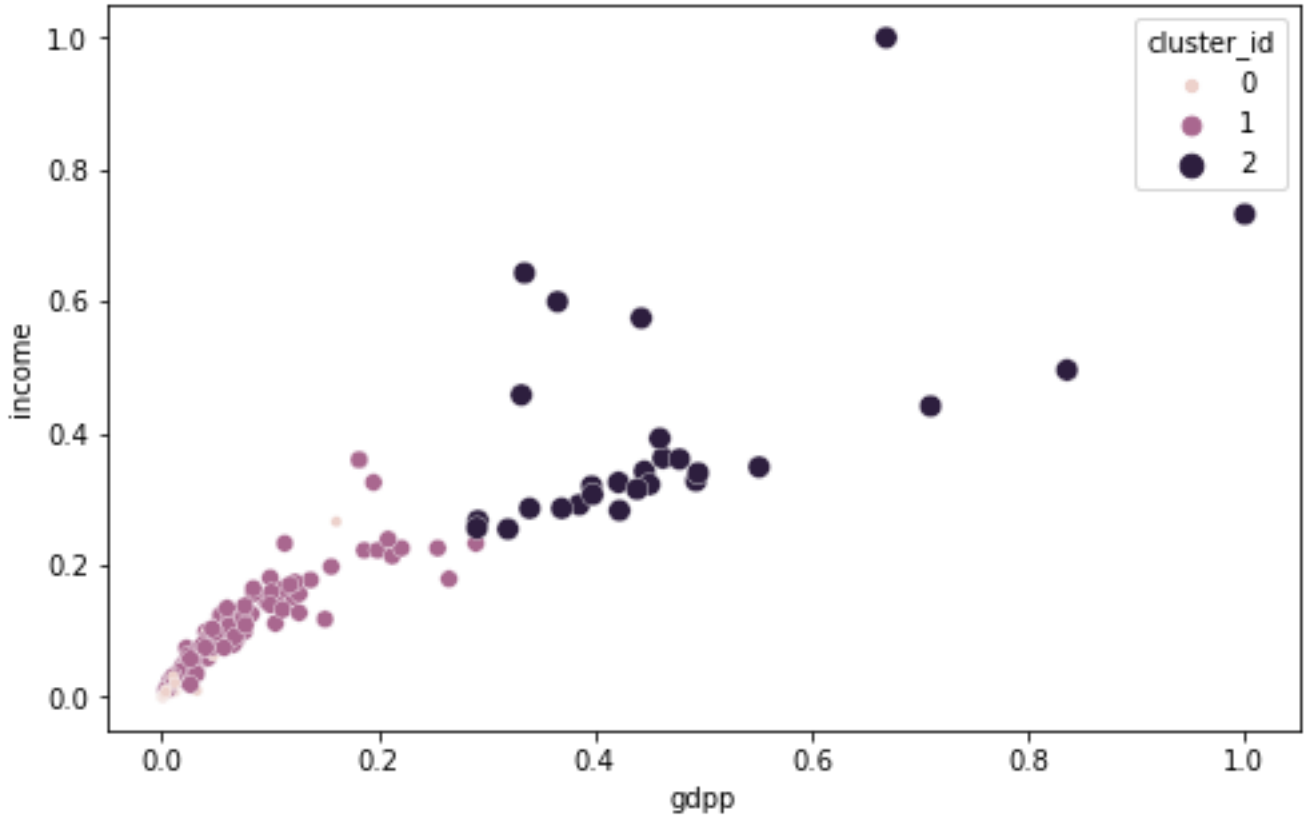
Dendrogram-Complete Linkage



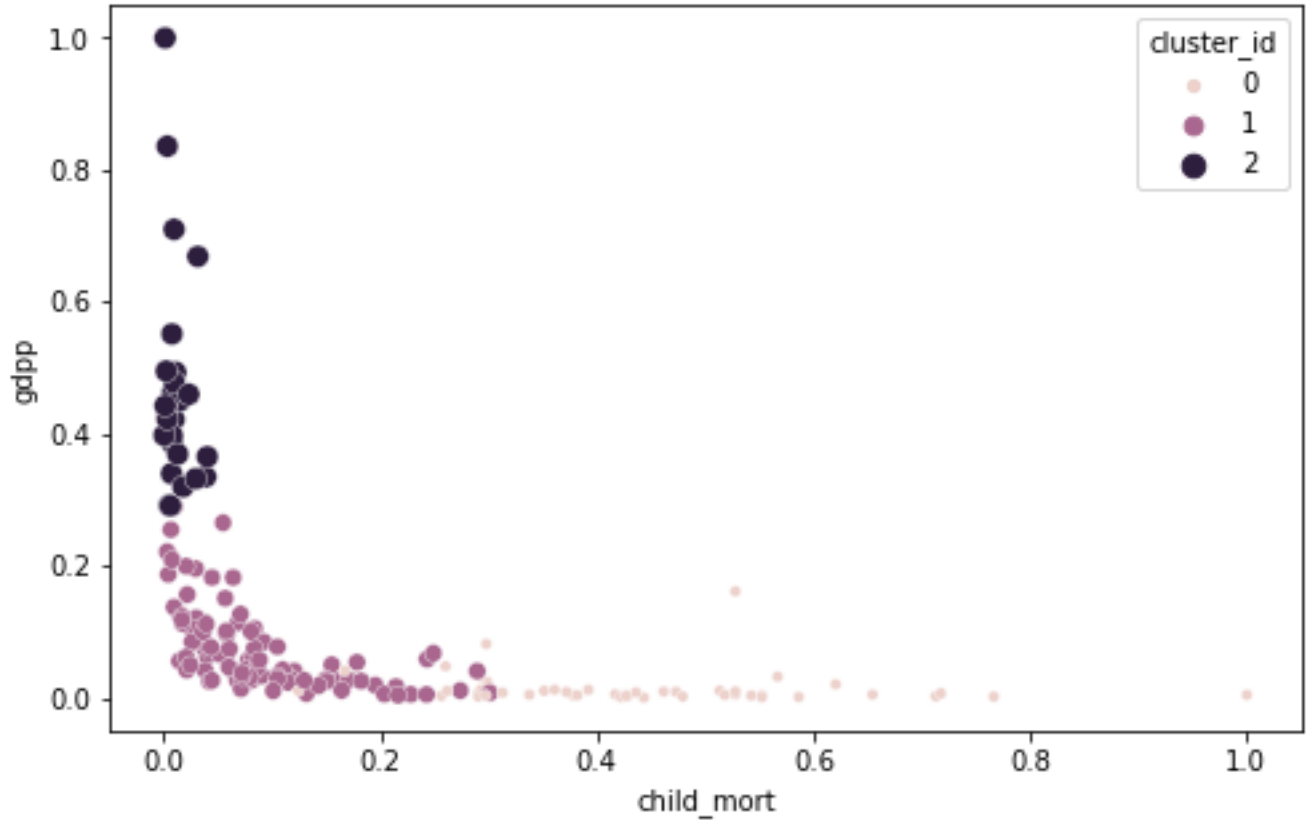
Scatter plot(to visualize clustering points for)- child_mort vs income



Scatter plot(to visualize clustering points for)- GDPP vs income



Scatter plot(to visualize clustering points for)- child_mort vs GDPP



Summary

- From the clustering techniques and the graphs obtained we can come to an understanding that
- From the labels of clusters formed:
  - cluster 0 = High Mortality rate, Low income and Low GDP
  - cluster 1 = Average Mortality rate, Average income, Average GDP
  - cluster 2 = Low Mortality rate, High income, High GDP
- So, It is concluded that when Mortality is high, Income and GDP of the country is poor .

Top countries to get help

- From the results obtained, we got the following poor countries which are to be concentrated as per the business requirement.
- As mentioned, sorted the countries based on gdpp, income and child_mort categories.
- 1. Congo,Dem Rep
- 2.Liberia
- 3. Burundi
- 4. Niger
- 5. Cental African Republic

Conclusion:

CEO of Help International is advised to concentrate on the above countries which are in dire need of help.