Clustering Assignment

Shubham Rajendra Kinhikar

Table of Content

1. Problem Statement and Aim

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 programmes, NGO 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.

Aim:

- Cluster the countries based on economic condition and significant factors provided in dataset.
- List down the top 5 countries that are in the direst need of aid.

2. Data Preparation

Data Inspection:

- Inspected the dataset and found that there **167** records in it.

Null Count Validation:

- Validated the Null count for all the columns and observed that there no missing value as such.

Data Standardization:

- Converted following columns from percentage to absolute values:
  - **exports**
  - **Imports**
  - **health**

3. Handling Outliers

- child_mort, inflation and total_fer columns are having outlier at upper range. As our model is majorly influence by the higher values of these columns due to that doesn't fixed these outliers.
- Handled the all the remaining numeric columns outlier by capping technique.

4. Univariate and Bivariate Analysis

Univariate Analysis Observation:

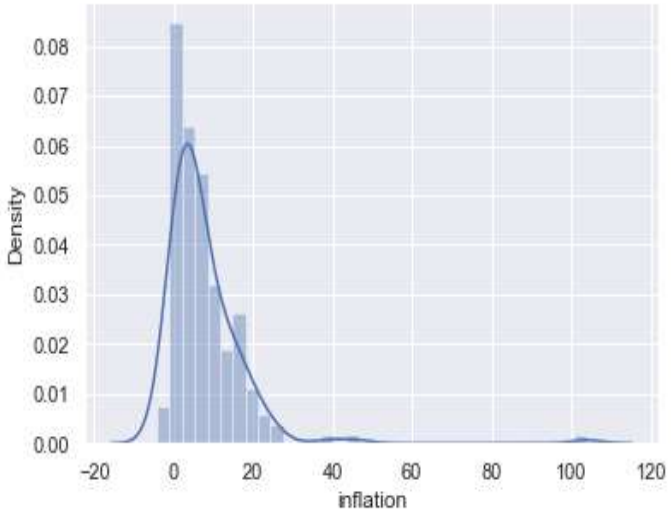
Most of the columns are normally distributed.

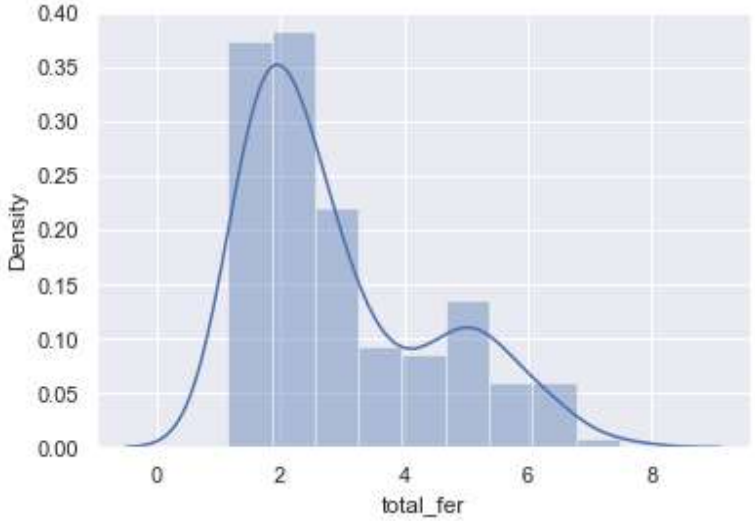
Bivariate Analysis Observation:

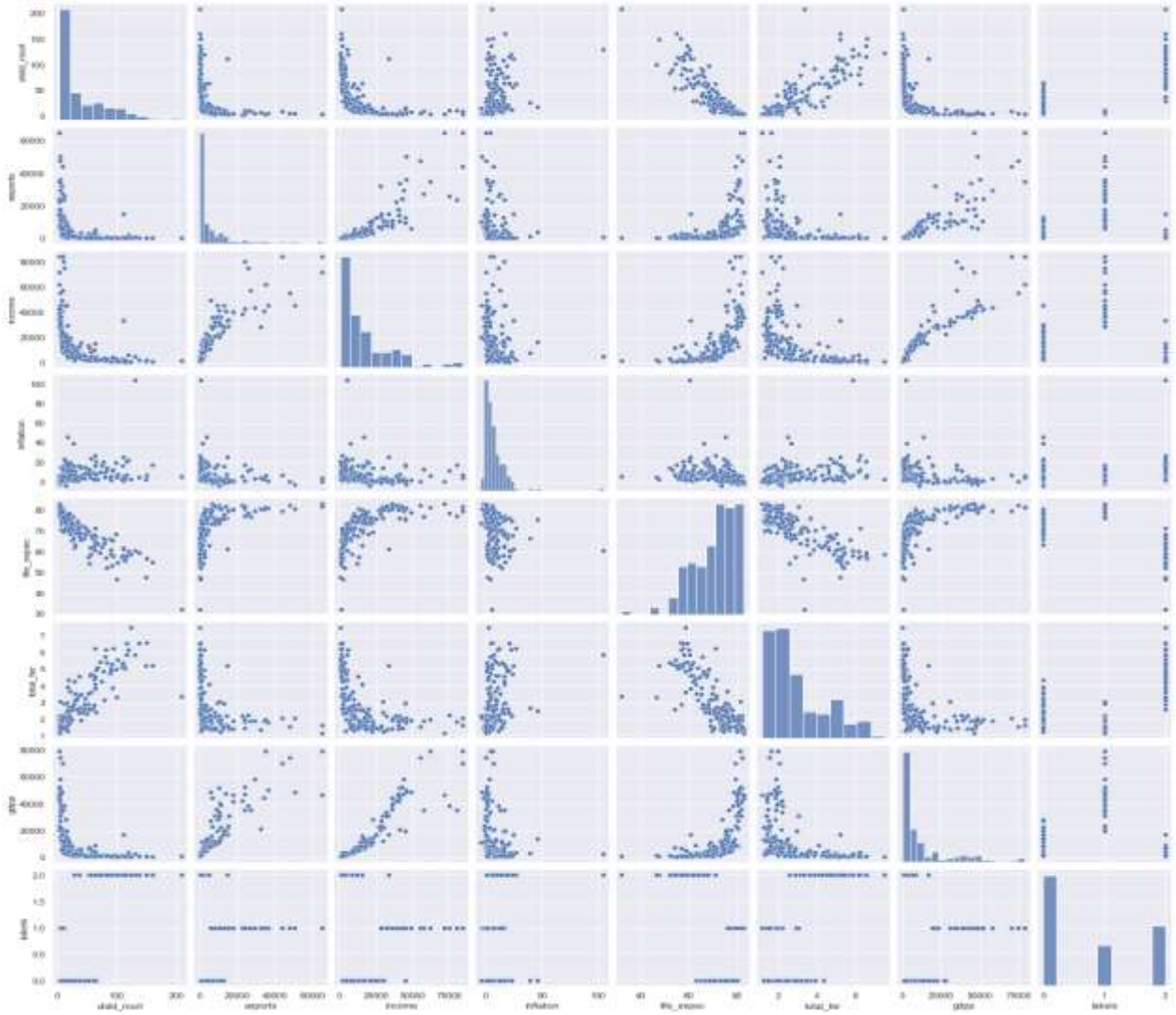
Pairplot has been made to examines whether data points differ significantly from uniformly distributed data in the two dimensional space.

Univariate Analysis Result:

Most of the columns are normally distributed.







Bivariate Analysis Result:

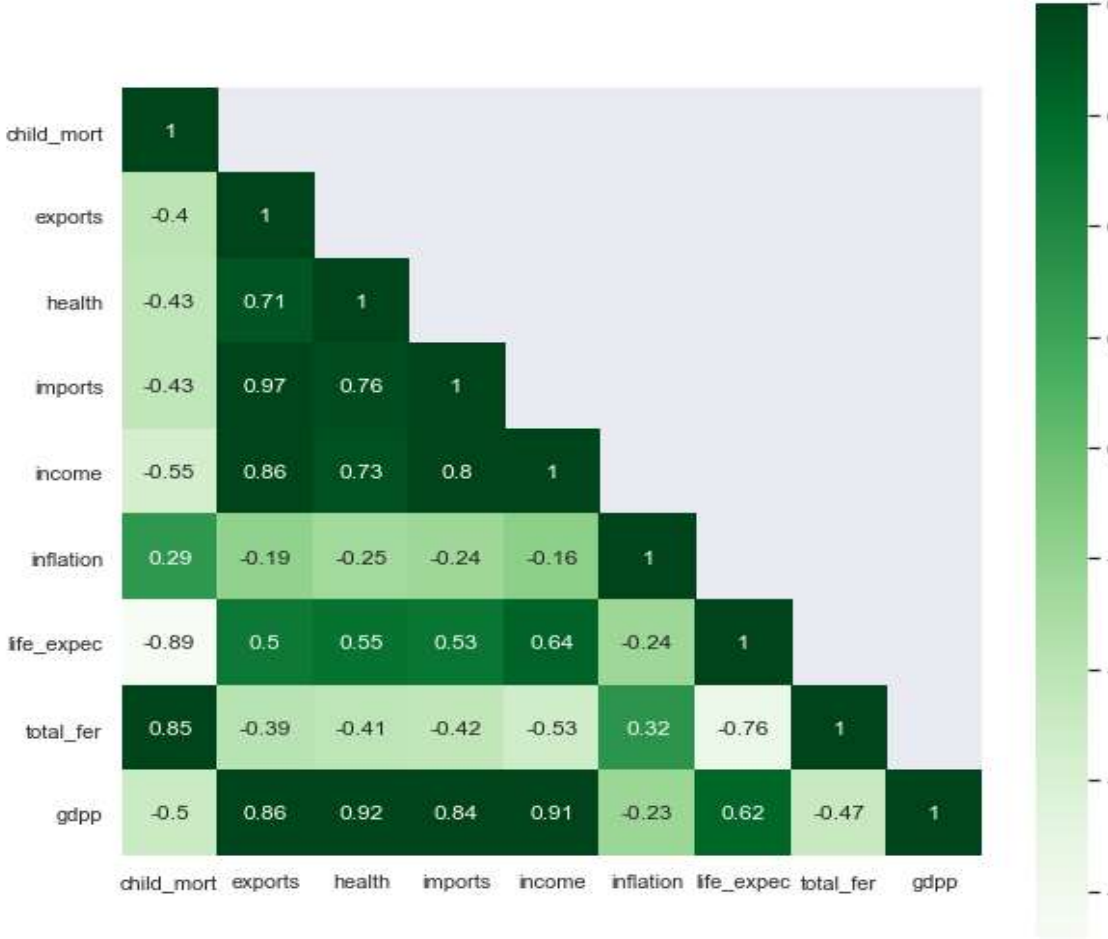
Data points are differ significantly from uniformly distributed data in the two dimensional space.

5. Correlation Between Attributes

Heatmap has been made to understand whether the variables correlate with each other to understand the genuinity of the data.

Dropped the imports and health columns

We can see that gdpp and income columns are highly colinear. As both columns are important for the clustering so didn't dropped any one of it.



6. Standardization

Scaled all the numeric columns using StandardScaler.

7. Check Cluster Tendency

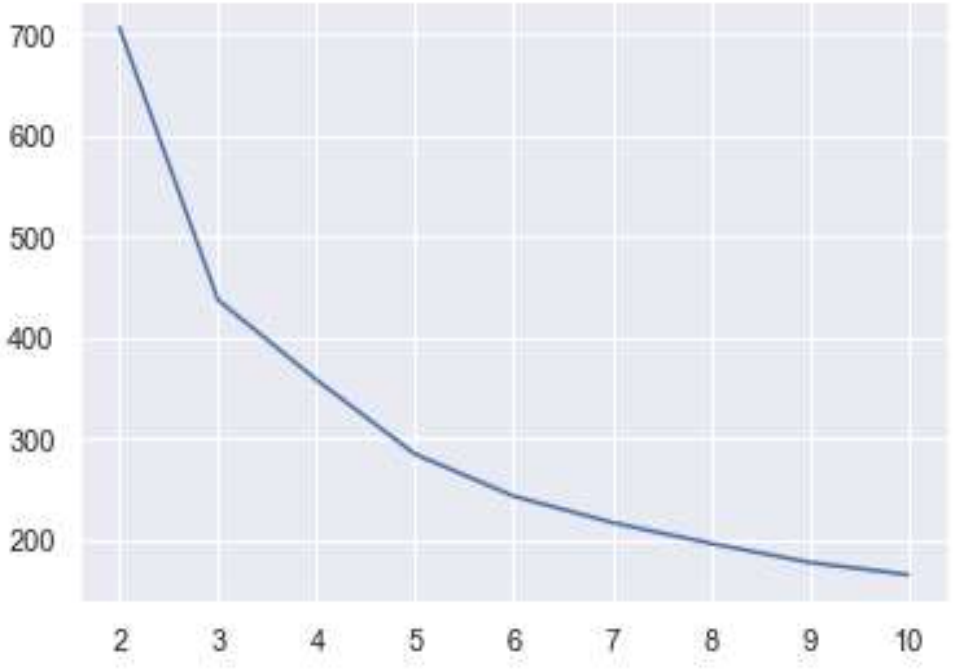
Hopkins test has been performed to examines whether data points differ significantly from uniformly distributed data in the multidimensional space.

Hopkins Score: **0.92**

8. Find the Optimal value of K

Applied Silhouette Analysis and Elbow-curve / SSD technique for detecting the optimal value of K.

Elbow Curve / SSD



From Elbow curve it has been observed that after 3 clusters the SSD value is not decreasing drastically so finalize the value of K=3

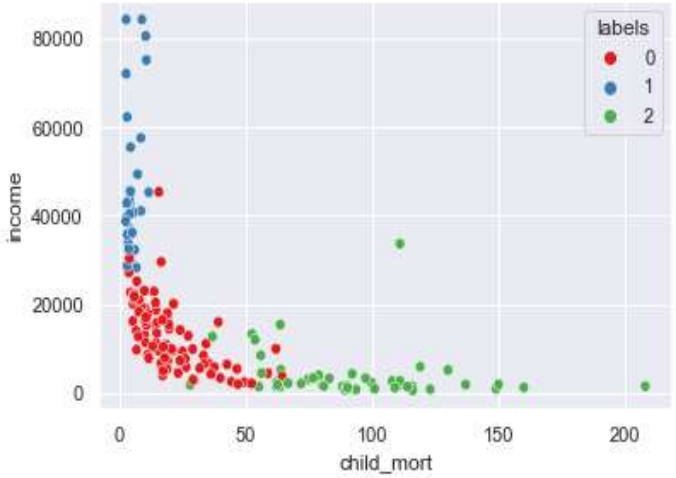
Silhouette Score

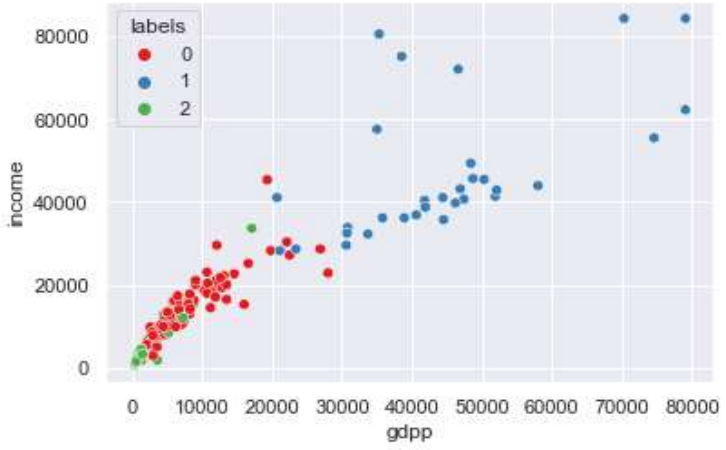
9. Analyzing the Clusters Created by K-means Algorithm

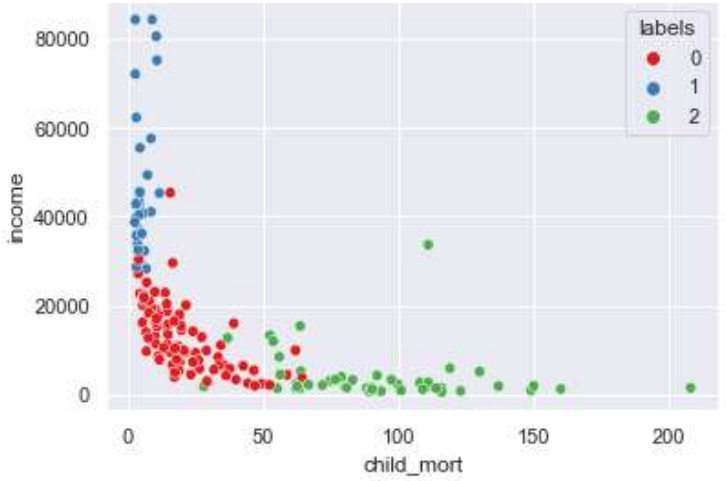
- Scatter plot has been made to analyze the clusters created by K-means Algorithm at **K=3**

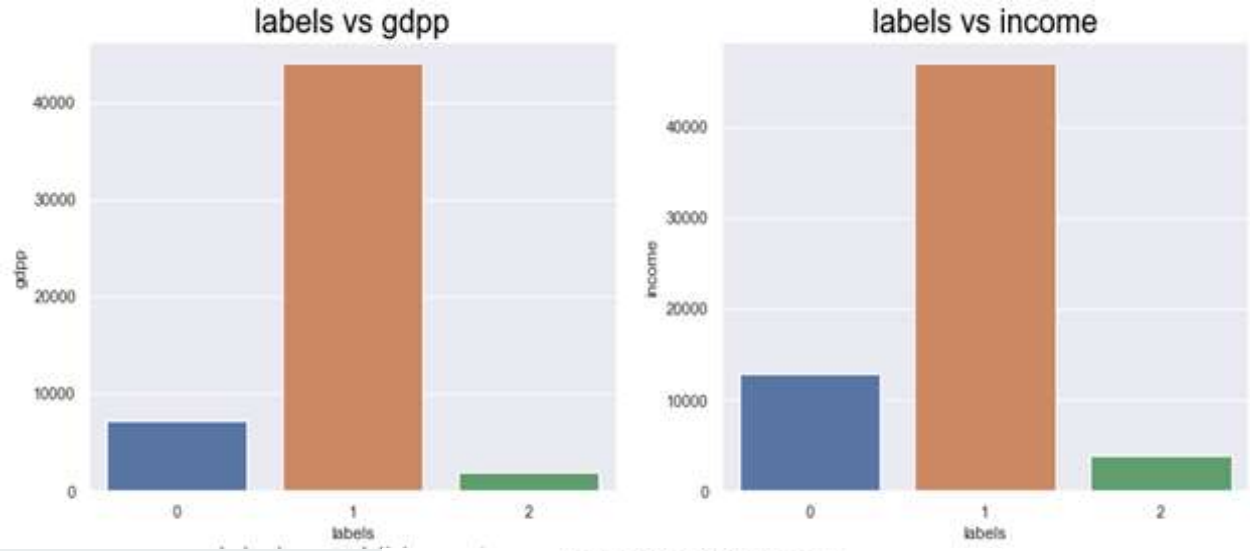
- Compared all the 3 clusters based on gdpp, income, child_mort variables and observed that **Cluster-2** countries are in dire need of aid
- Sorted the Cluster-2 countries base **gdpp(Ascending), income(Ascending), child_mort(descending)** variables to identify the top five countries which are in dire need of aid.

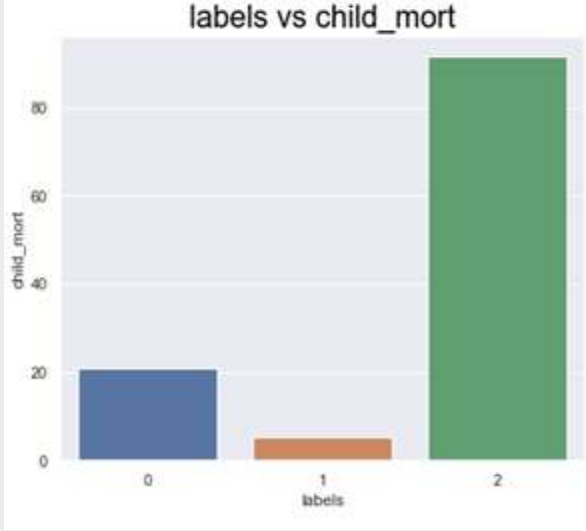
Visualizing the clusters after applying K-means Algorithm:











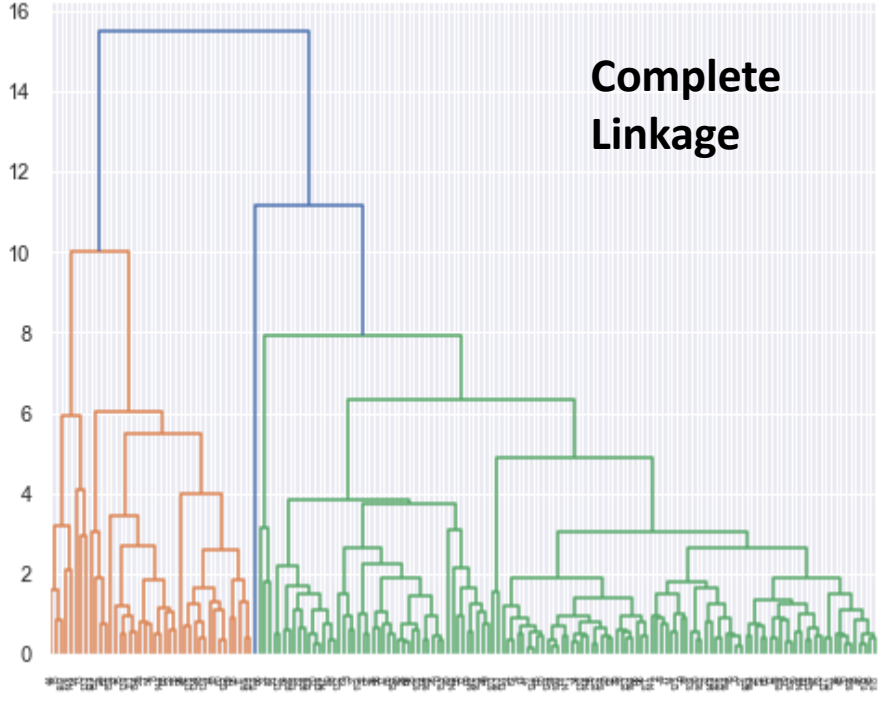
Compared All the 3 Clusters:

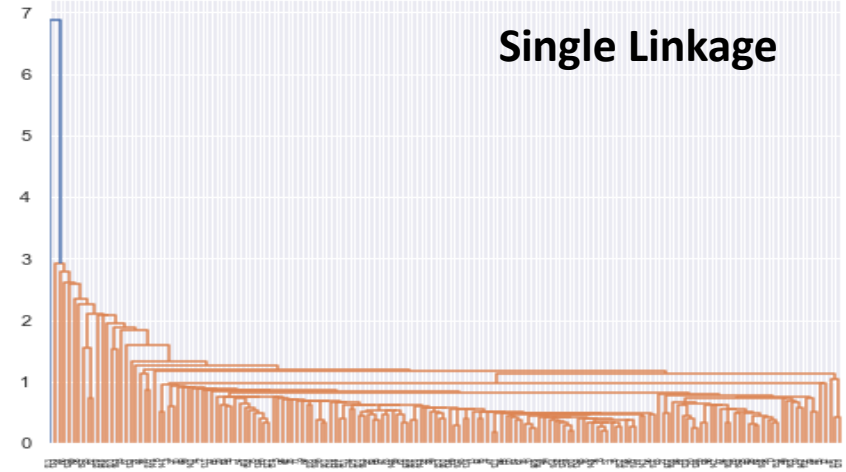
Cluster-2 countries are in dire need of aid

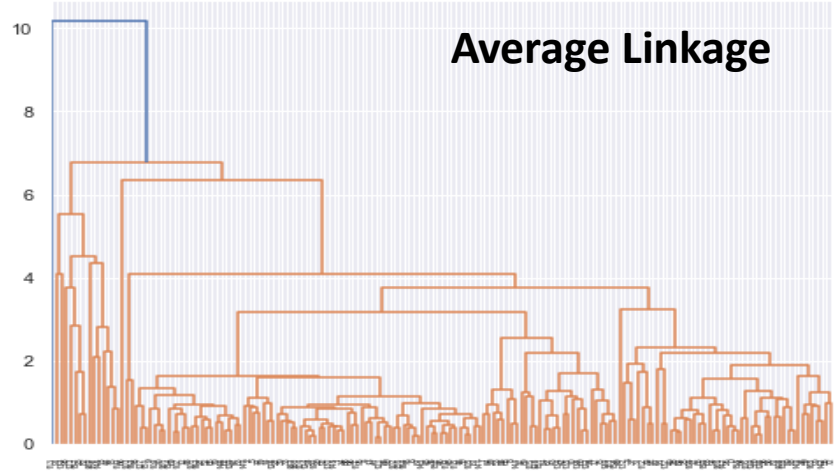
10. Hierarchical Clustering

- Applied Hierarchical Clustering algorithm with all the 3 linkage.
- It has been observe that Single and Average linkage are not making much sense So considered Complete linkage tree for further analysis
- Cut the Complete Linkage dendrogram to obtain 3 and 2 clusters

Considered Complete Linkage for Further Analysis:

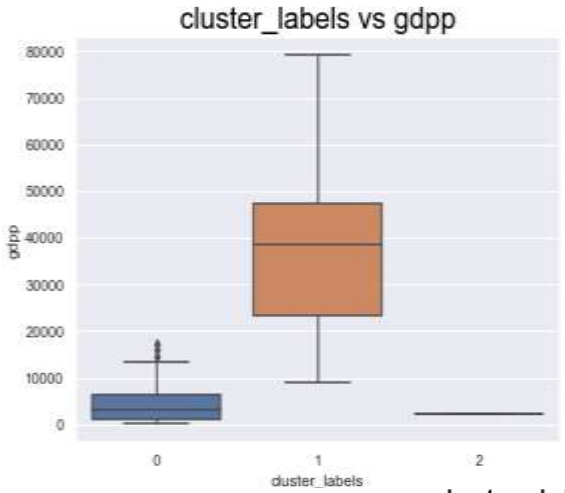


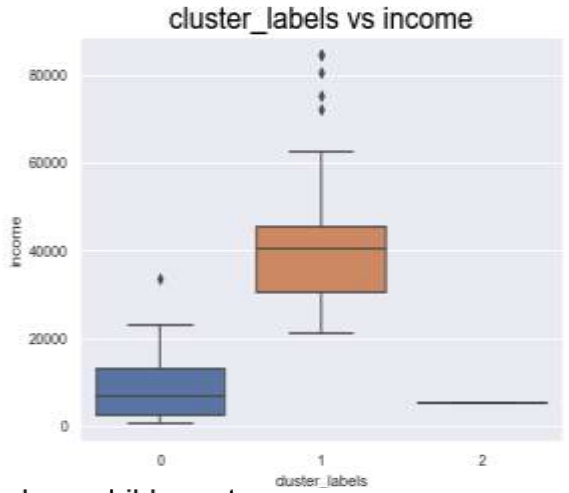


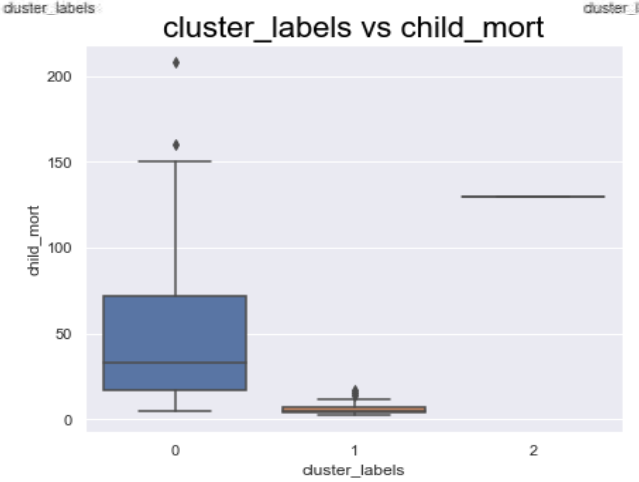


Cut the Complete Linkage dendrogram to obtain 3 Clusters:

Cluster-2 countries are in dire need of aid but it contain only one country

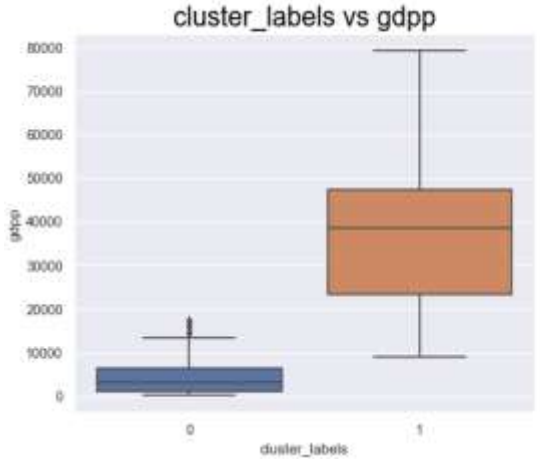


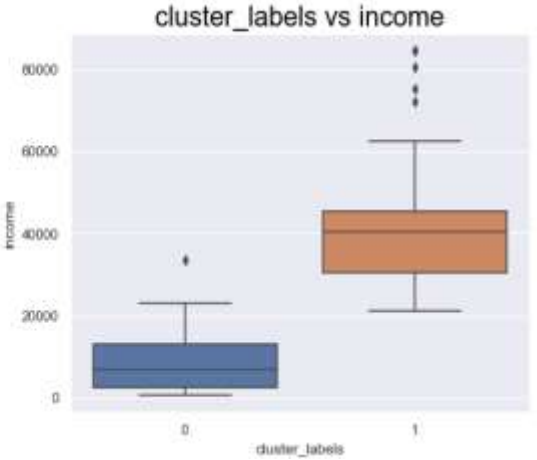


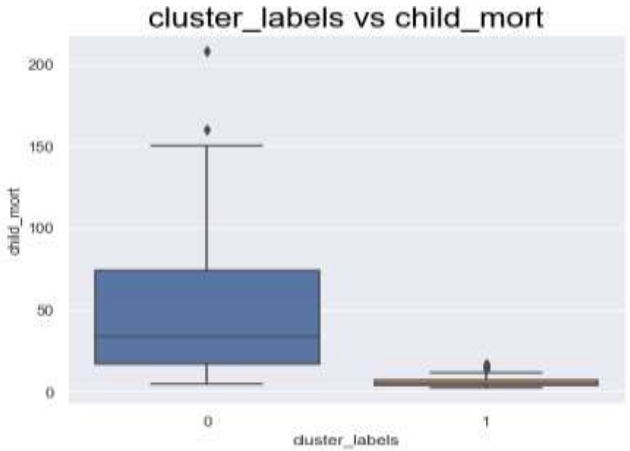


Cut the Complete Linkage dendrogram to obtain 2 Clusters:

Cluster-1 countries are in dire need of aid.







11. Identify the top 5 countries

- Compared the K-means and Hierarchical clustering results at **K=3** and both are giving the different results.
- But if we comapred the **top 5 countries** which are in direst need of aid at **K=3 in K-means** and with **2 clusters in Hierarchical clustering** then the countries are matching. So finalized the same list of countries which are as follows:
    - Burundi
    - Liberia
    - Congo, Dem. Rep.
    - Niger
    - Sierra Leone

Thank you

- Shubham Rajendra Kinhikar

Post Graduate Diploma in Data Science
International Institute of Information Technology - Bangalore
Upgrad