Clustering Assignment

Identifying countries that are in direst need of aid.

By: **Rishik Patel**

Problem Statement

- International humanitarian NGO 'HELP International' gathered funds of $10 million through funding and need to identify the countries that are in direst need of financial aid.
- We need to list these countries based on socio-economic factors available for 167 countries.

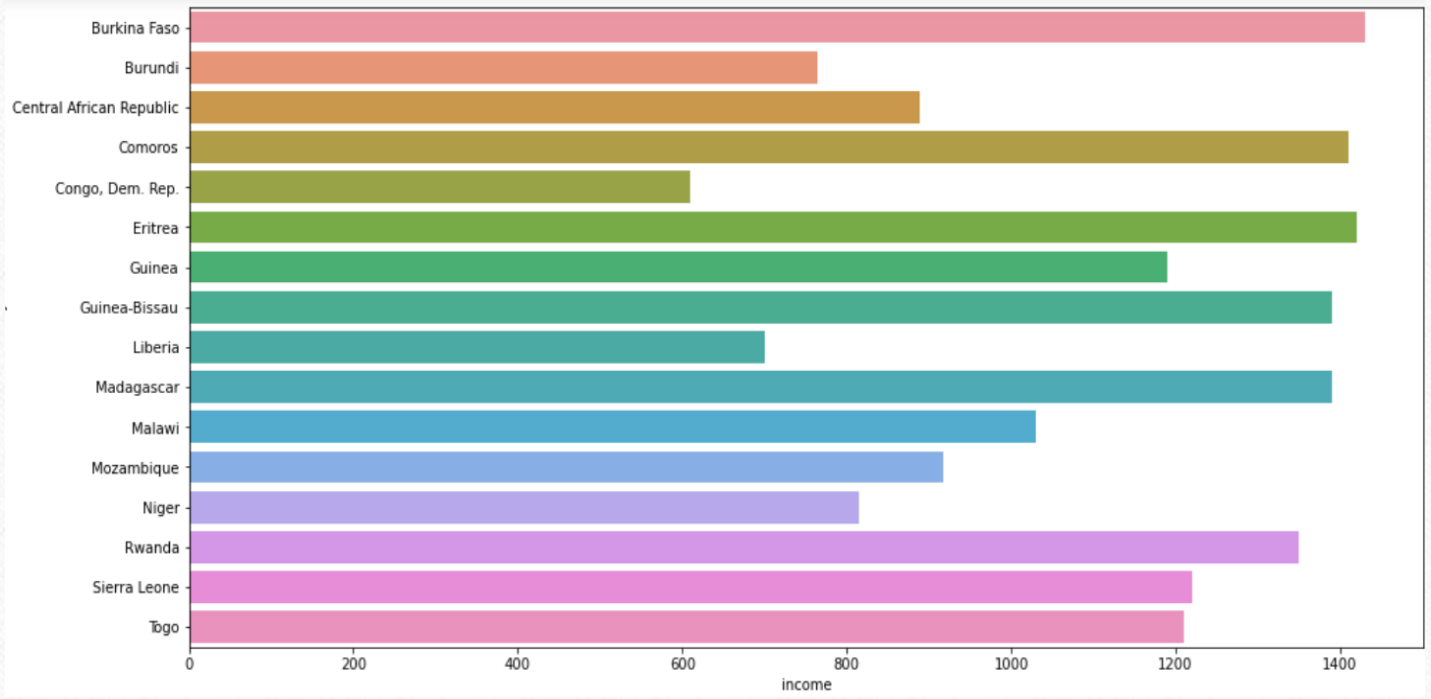
Analysis Approach

- We performed thorough EDA to verify nature and relationship of the available data.
- Further, we used clustering algorithms to determine different clusters present in the data set into which a specific characteristic country could be placed into.
- We used two different types of clustering algorithms (K-means & Hierarchical) to compliment and verify each others results.

EDA – Univariate Analysis

Net Income per person

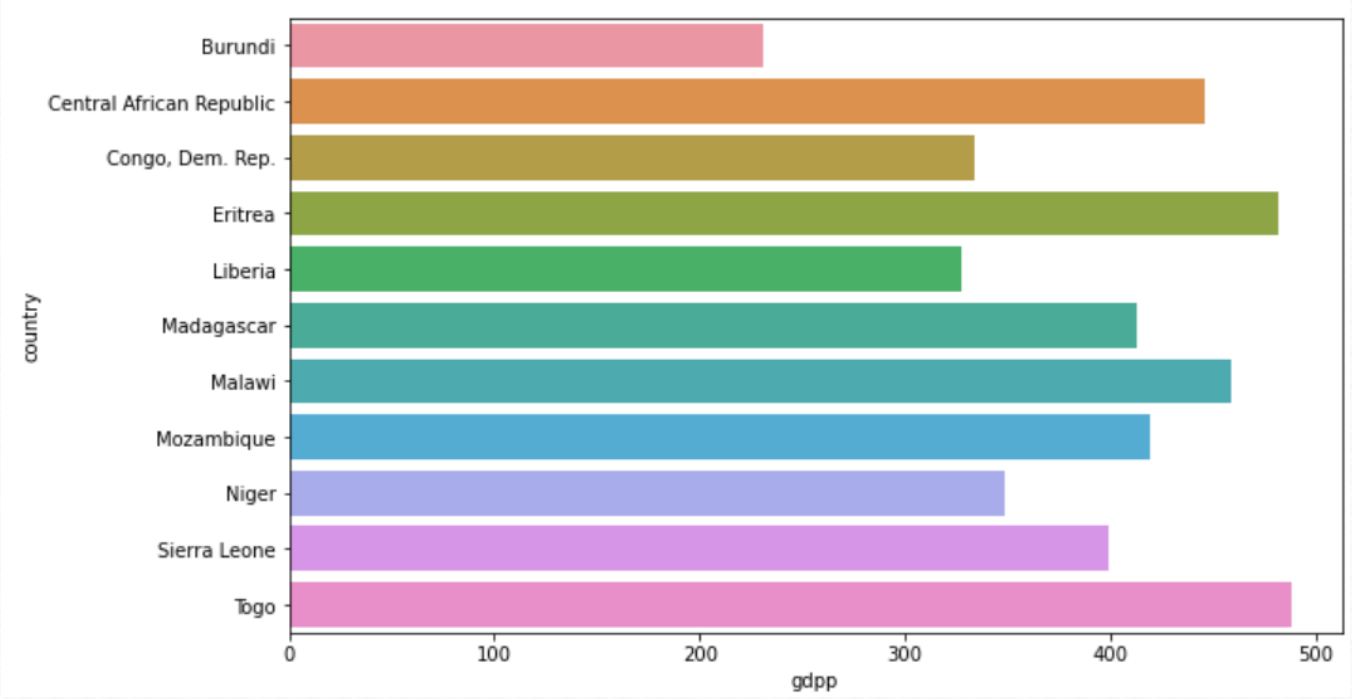
- We found several countries with net Income per person less than 1500 (Ref. plot)
- Ranged as *low* as 609 : **Congo, Dem Rep**
- Ranged as *high* as 125000: **Qatar**



EDA – Univariate Analysis

GDPP

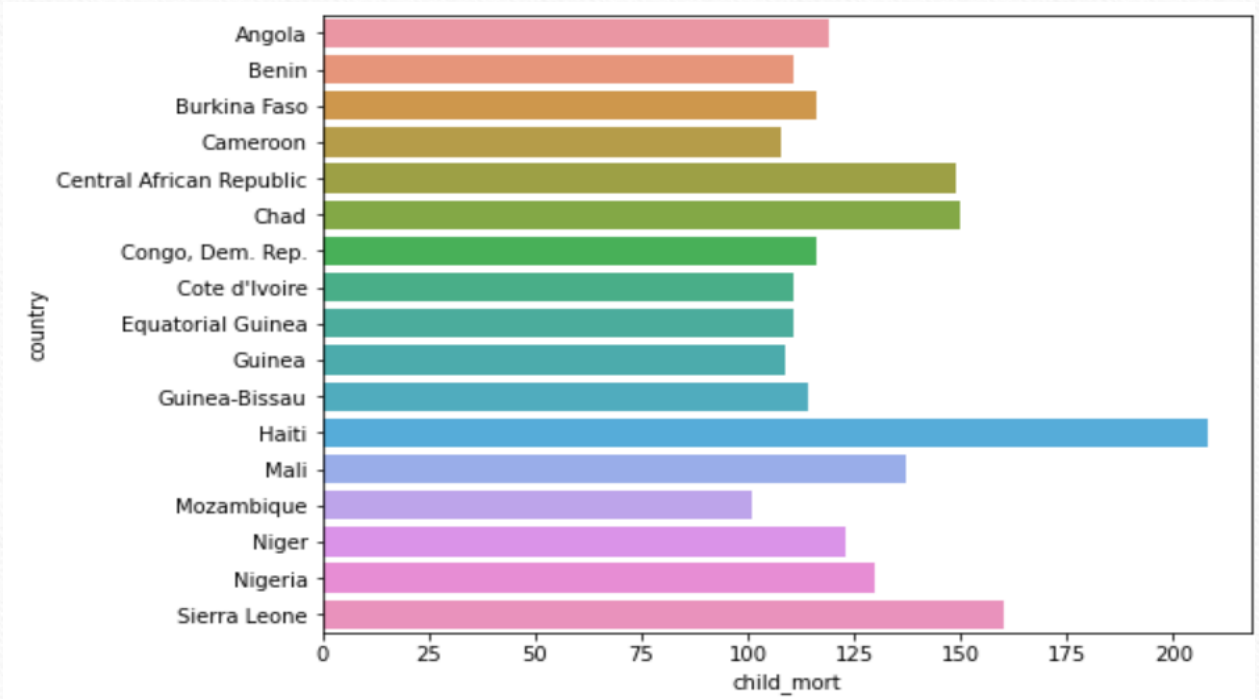
- We found several countries with GDP per capita less than 500 (Ref. plot)
- Ranged as *low* as 231 : **Burundi**
- Ranged as *high* as 105000: **Luxembourg**



EDA – Univariate Analysis

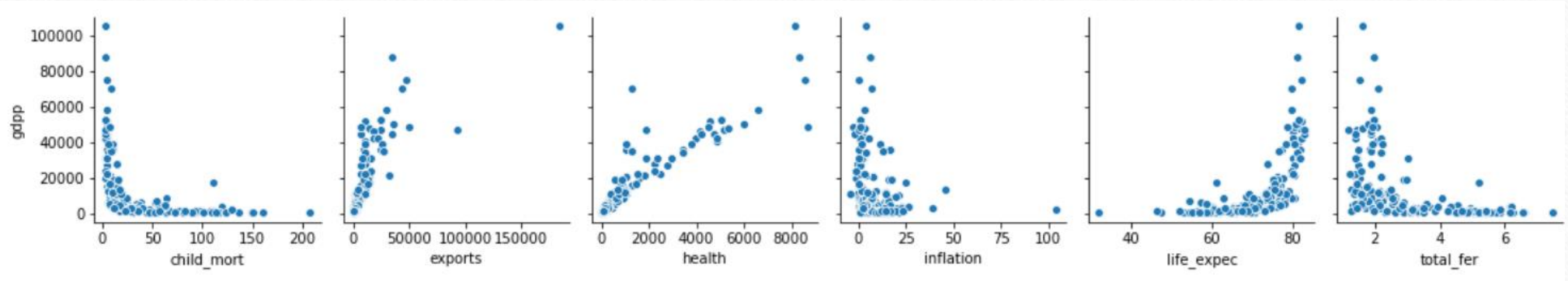
Child Mortality

- We found several countries with Child Mortality more than 100 (Ref. plot)
- Ranged as *high* as 208: **Haiti**
- Ranged as *low* as 2.6 : **Iceland**



EDA – Bivariate Analysis

GDPP vs rest



Countries with increasing (or higher) GDPP have **low**:

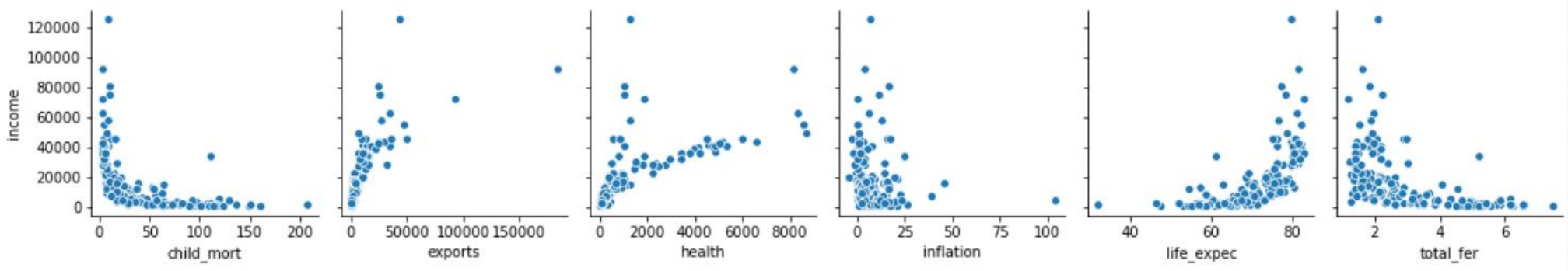
- Child Mortality rate
- Total Fertility rate
- Inflation

Countries with increasing (or higher) GDPP have **high**:

- Exports & Imports
- Health spending
- Life Expectancy

EDA – Bivariate Analysis

INCOME vs rest



Countries with increasing (or higher) Income have **low**:

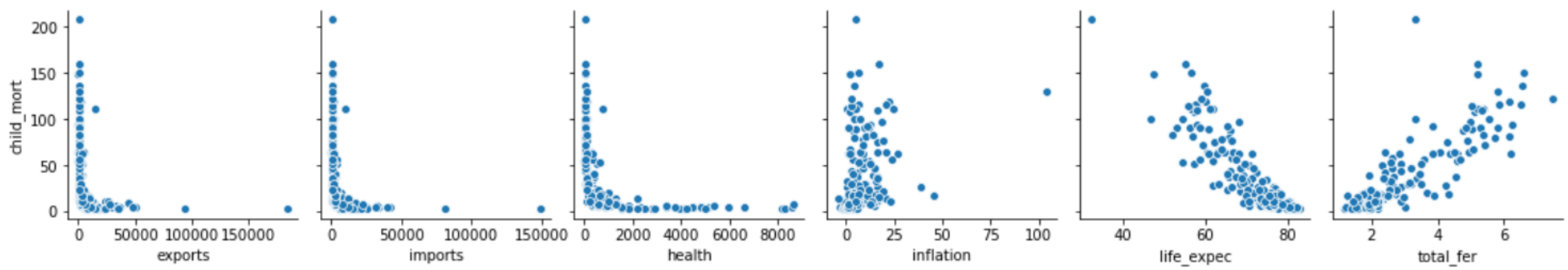
- Child Mortality rate
- Total Fertility rate
- Inflation

Countries with increasing (or higher) Income have **high**:

- Exports & Imports
- Health spending
- Life Expectancy

EDA – Bivariate Analysis

Child Mortality vs rest

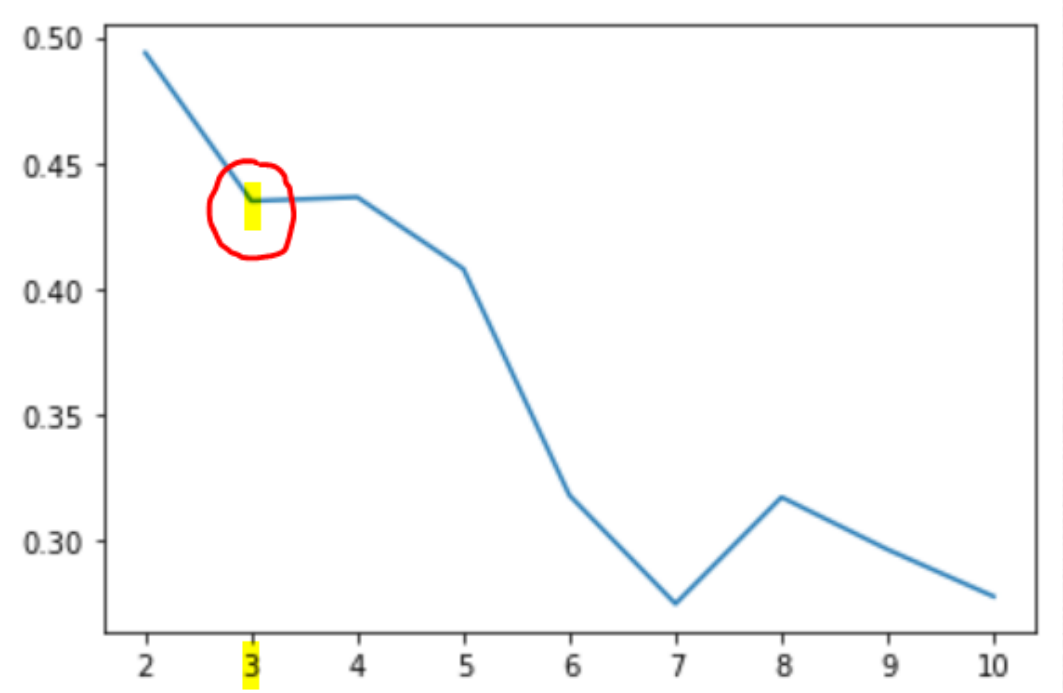


- Countries least involved in exports and imports are seen with highest child mortality rate. The health spending is also the least for countries with higher child mortality. Clearly, they are financially worst.
- Child mortality rate correctly decreases with increase in life expectancy.
- Child mortality rate correctly increases with increase in total fertility, showcasing the financial burden it puts on a family

Selection of Number of Clusters

Silhouette Score curve

- Silhouette Score gives the measure of how similar a data point is to its own cluster than the rest.
- The higher the Silhouette score for a cluster, more the cohesion between it's member data points and more the separation between the data points of different cluster.
- Chose **'three'** clusters as it's grossing highest score of all available clusters.

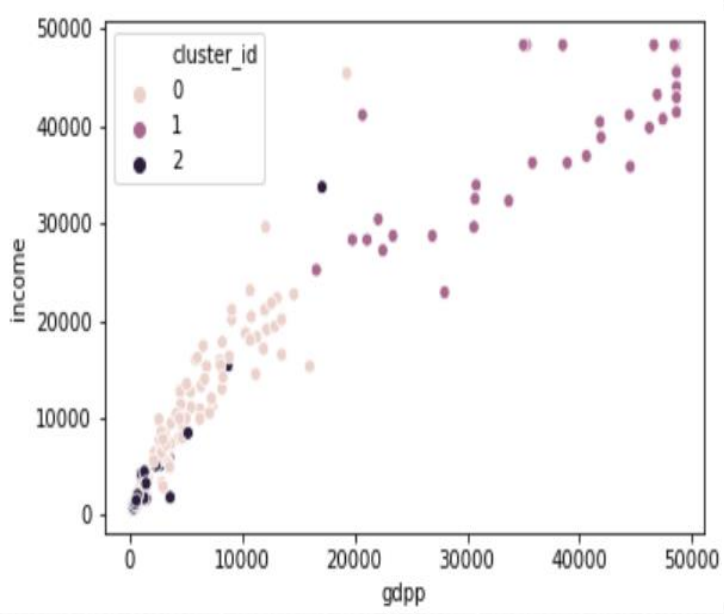


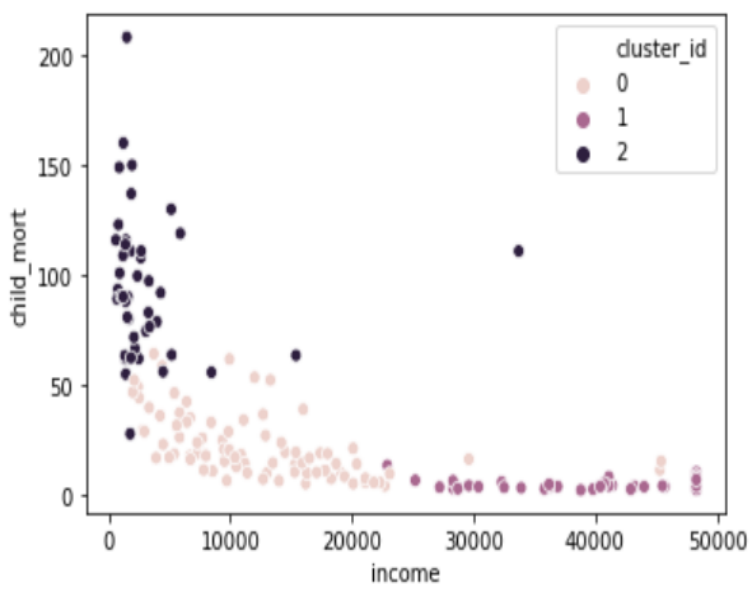
Selection of Number of Clusters

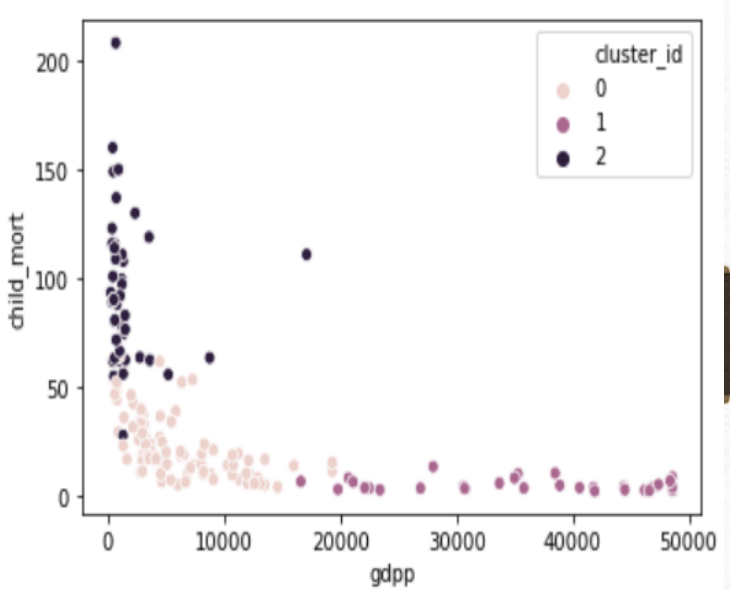
SSD or Elbow curve

- It’s the Sum of squared distances between data points and their closest cluster center & Distortion as the average of the squared distances of each points from the respective cluster centers.
- It’s plotted for each cluster and the point or cluster where the distortion value changes most significantly is taken as ‘k’. In our case **‘three’**
- It’s termed as elbow curve as the plot makes steep change in slope at that cluster and is seen as an elbow.

Clusters formed using K-Means Algorithm

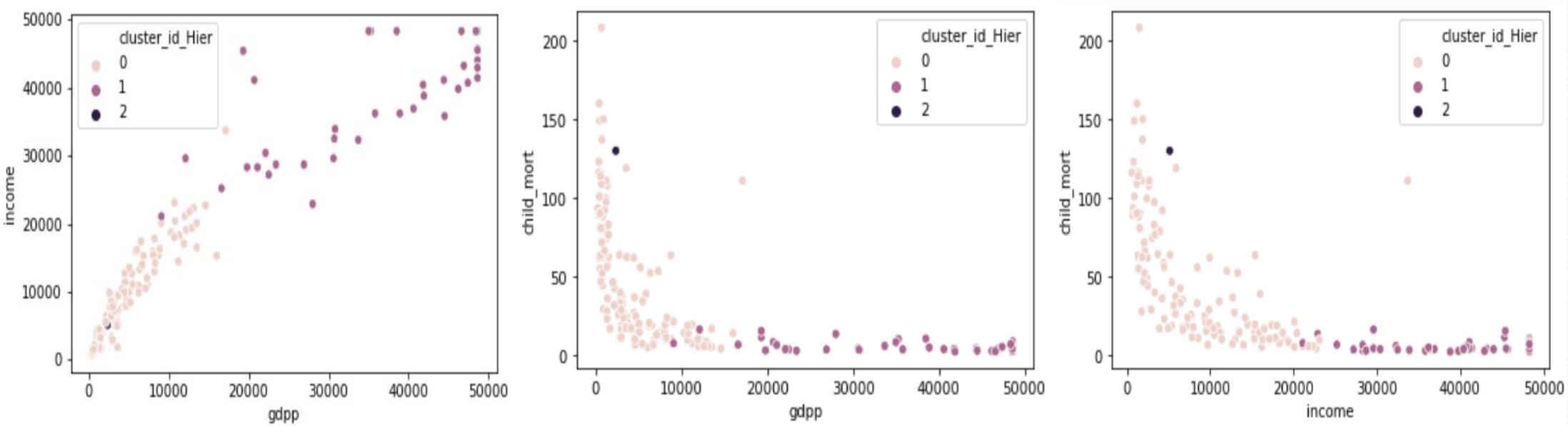






- Cluster 1 represents high income, high GDPP and least child mortality – Developed nations
- Cluster 2 represents least income, least GDPP and highest child mortality – **Under Developed**
- Cluster 0 represent moderate income, GDPP and child mortality – Developing nations

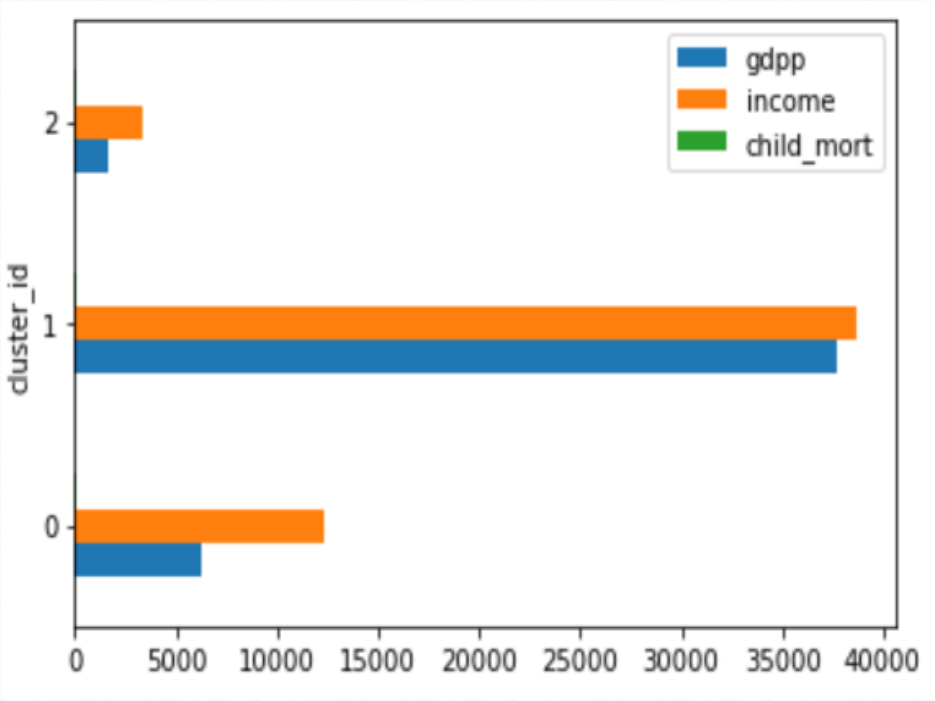
Clusters formed using Hierarchical Clustering Algorithm



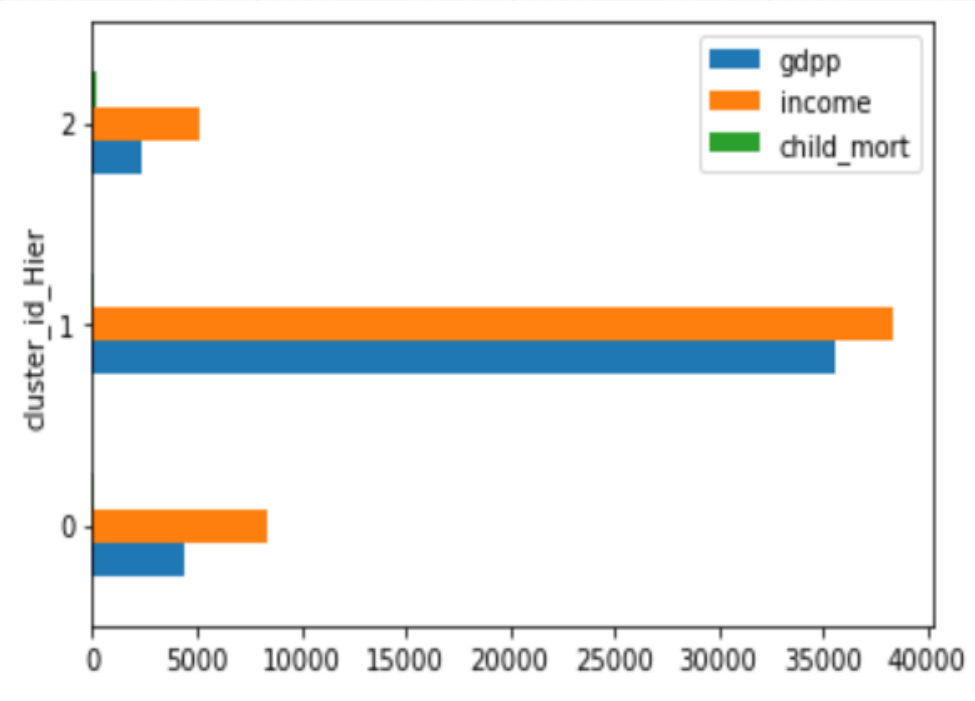
- Cluster 1 represents high income, high GDPP and least child mortality – Developed nations
- Cluster 2 represents least income, least GDPP and highest child mortality – **Under Developed**
- Cluster 0 represent moderate income, GDPP and child mortality – Developing nations

Cluster profiling achieved through

K-means Clustering



Hierarchical Clustering (Complete linkage)



Recommendation

- Based on the socio-economic factors of various countries we found following listed countries to be the TOP 10 worst affected of all (sorted in ascending order by GDPP, Income & descending order by Child mortality):
- From top → bottom: In Decreasing order of urgency
- These countries are in dire need of basic amenities and relief

country	child_mort	income	gdpp
Congo, Dem. Rep.	116	609	334
Liberia	89.3	700	327
Burundi	93.6	764	231
Niger	123	814	348
Central African Republic	149	888	446
Mozambique	101	918	419
Malawi	90.5	1030	459
Guinea	109	1190	648
Togo	90.3	1210	488
Sierra Leone	160	1220	399