CLUSTERING COUNTRIES FOR HELP INTERNATIONAL NGO

By: Srijani Das

PROBLEM STATEMENT

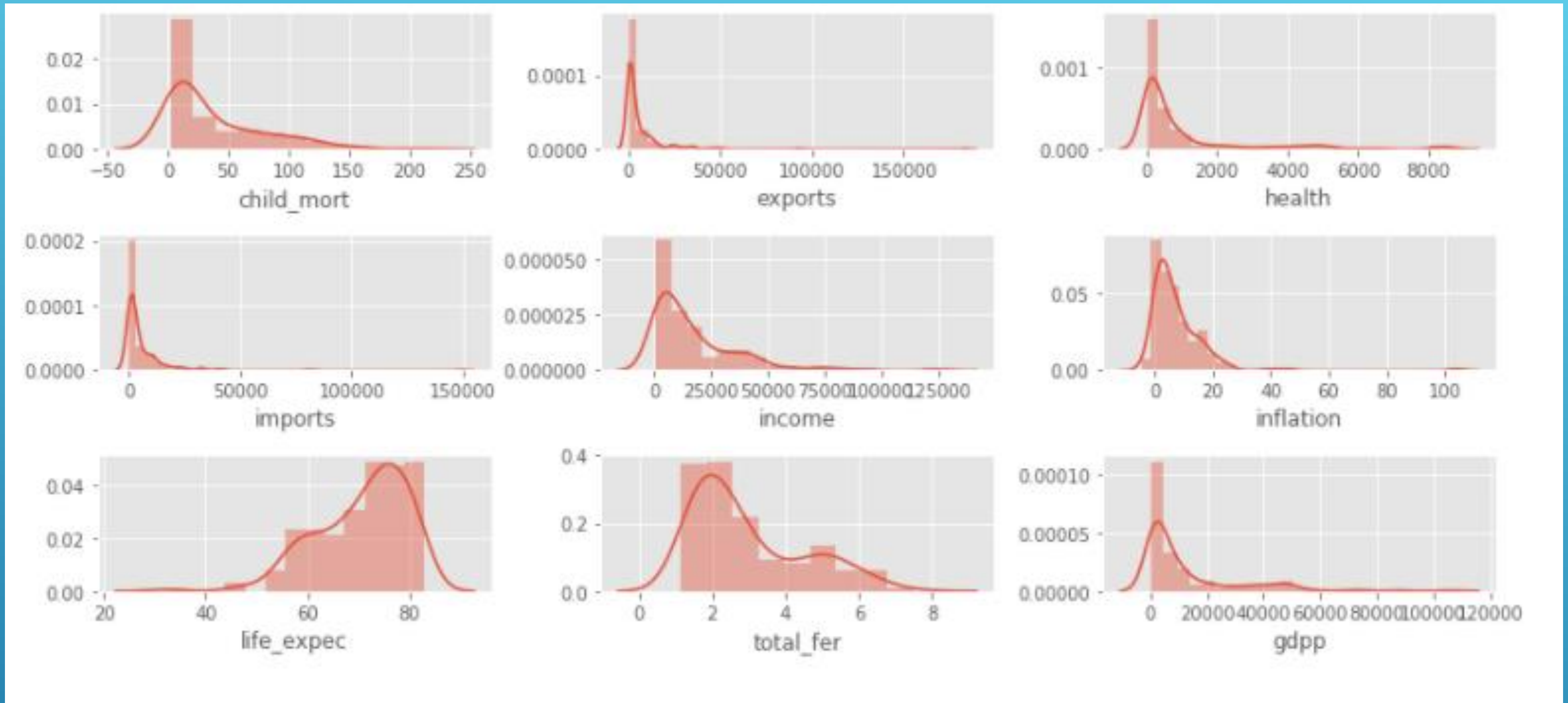
- HELP International NGO intends to provide under-developed countries with basic amenities and relief.
- The main objective of this case study is to identify the group of countries that needs relief the most by clustering the countries based on socio-economic and health factors.

APPROACH

DATA COLLECTION AND PREPARATION

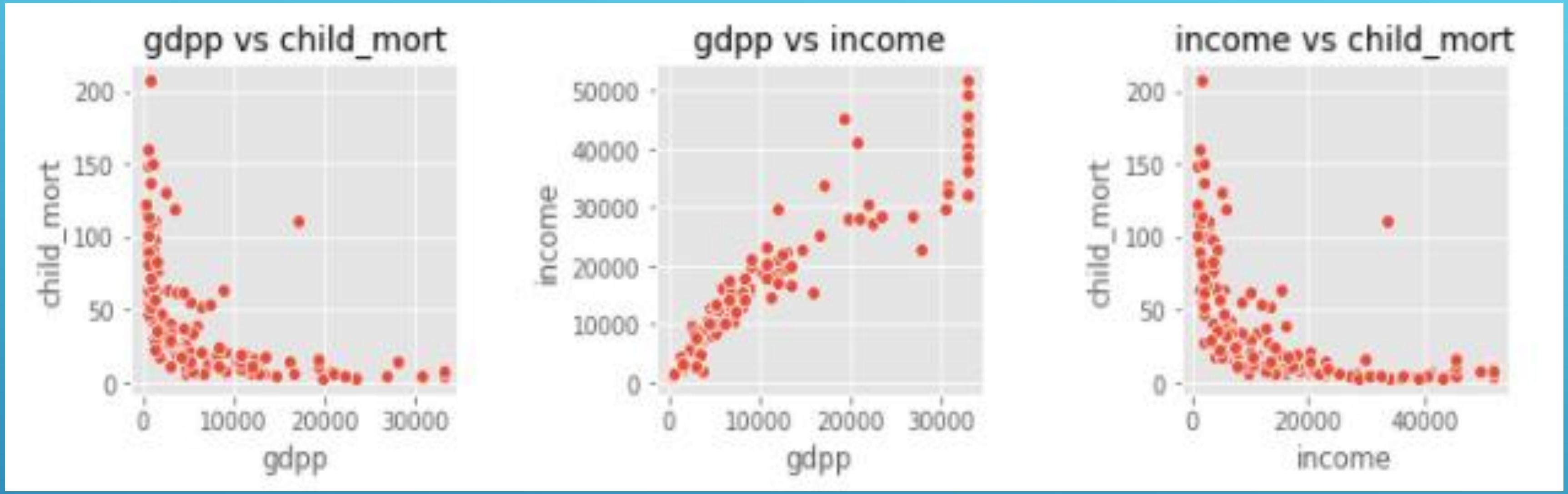
- Dataset used : Country-data.csv
- There is no missing values, repeated values and all the features have correct data type.
- 'exports', 'health', 'imports‘ are converted into actual values as they are given in percentage of gdpp
- Handling Outliers : All numerical features in the dataset have outliers.
- Flooring is skipped for ‘gdpp’ and ‘income’ columns and capping is skipped for ‘child_mort’- as these features are used for ordering the under-developed countries.
- Outliers are treated for all other features.

DATA VISUALIZATION



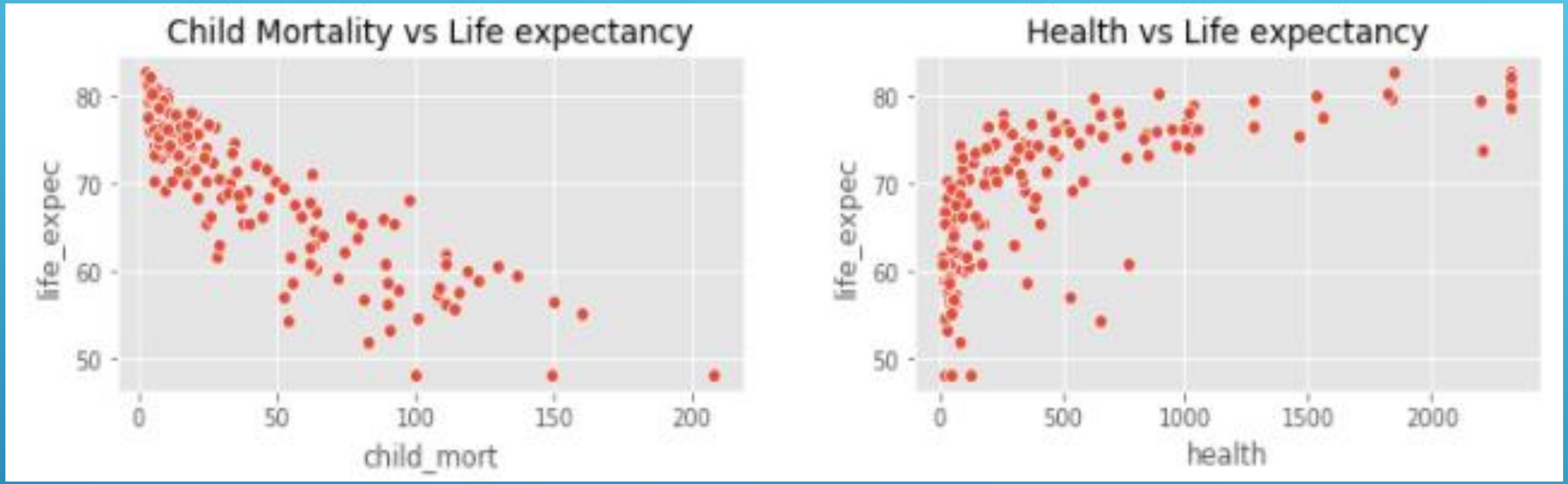
- Distribution of Child mortality (child_mort), GDP per capita (gdpp), exports, imports, income, inflation, total_fer feature are right-skewed, that means majority of the countries have low child mortality, low GDP, low income and so on.
- Most countries have good life expectancy (life_expec), so the distribution is left-skewed.

DATA VISUALIZATION



- Countries with high gdpp have low child mortality rate
- Countries with high income have low child mortality rate
- Gdpp and income has linear relationship

DATA VISUALIZATION

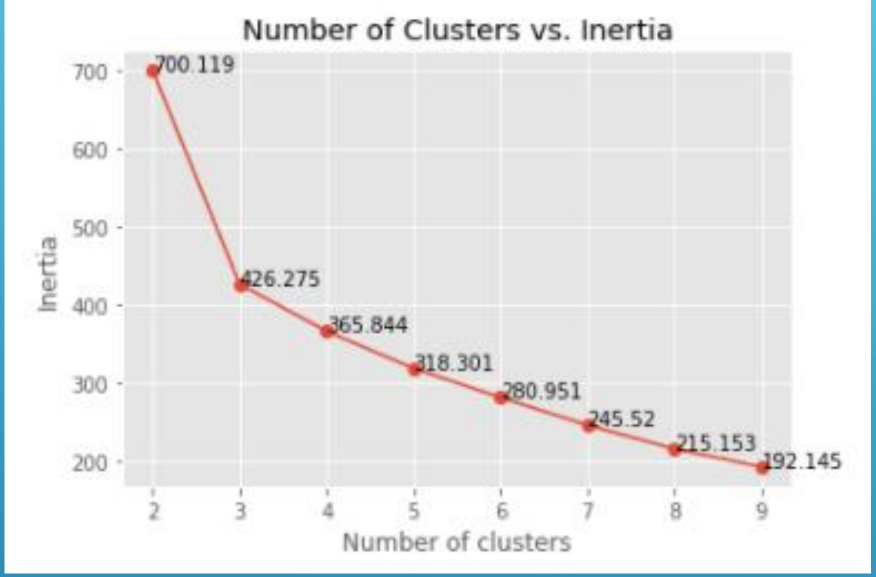


- High child mortality rate implies low life expectancy
- In countries where health expenditure per capita is high, life expectancy is high as well.

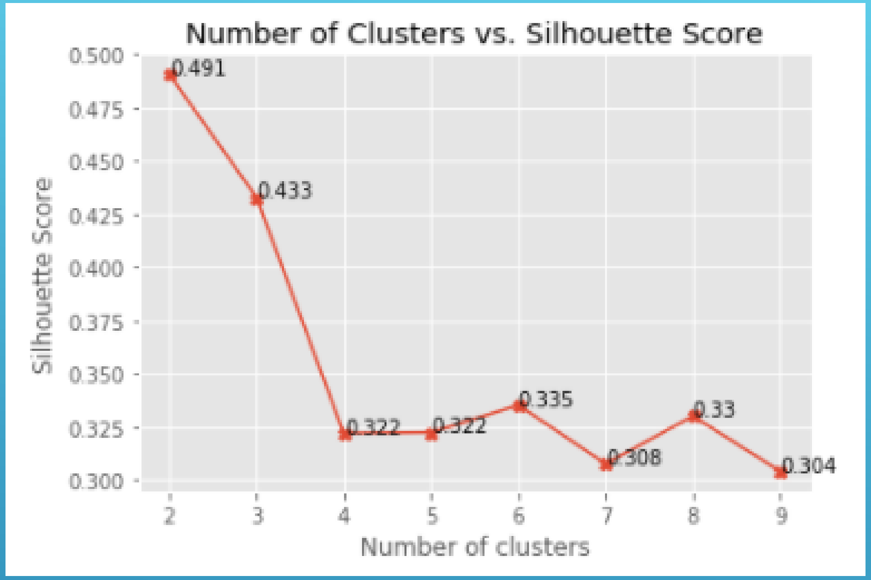
FEATURE SCALING AND HOPKINS TEST

- Standardized the data for better clustering.
- Performed Hopkins test to verify suitability of data for clustering.
- High Hopkins score (0.885) confirms high clustering tendency in the given data.

CLUSTERING : KMEANS



Elbow method

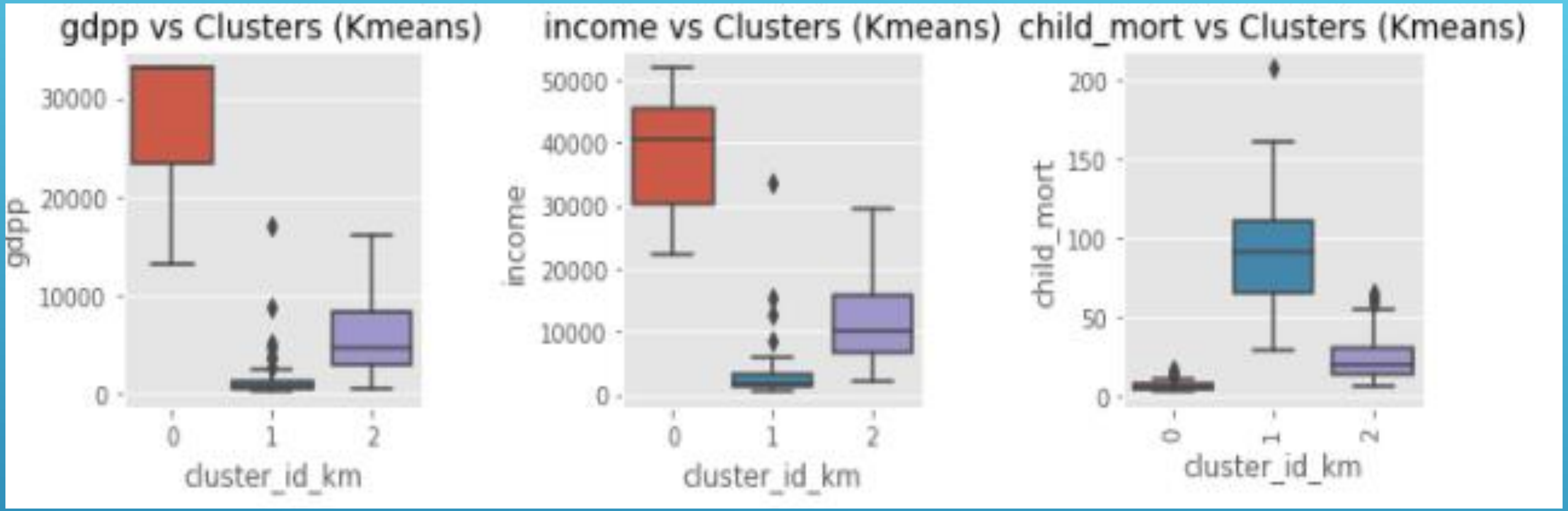


Silhouette Score

Choosing optimum number of clusters, K

- Elbow method suggest K=3 is optimum
- K=2 has highest Silhouette score. It will be useful to take 3 clusters as –
- under-developed, developing and developed countries

CLUSTER PROFILING : KMEANS



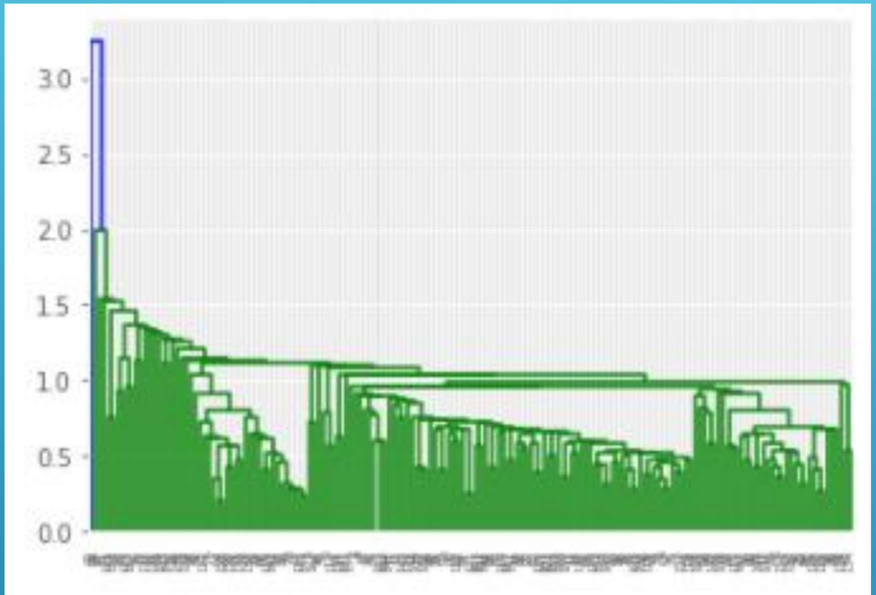
Cluster 0 : 41 Developed Countries (high gdpp and income, low child mortality)

Cluster 1 : 46 Under-developed Countries (low gdpp and income, high child mortality)

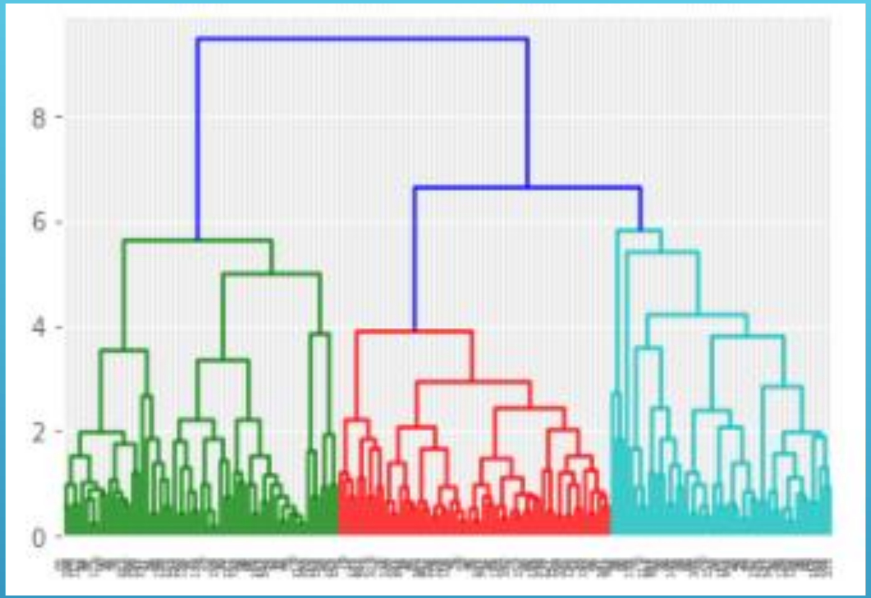
Cluster 2 : 80 Developing Countries (medium gdpp and income, medium child mortality)

Countries in Cluster 1 (KMeans) are suitable for getting relief fund.

CLUSTERING : HIERARCHICAL CLUSTERING



Single Linkage

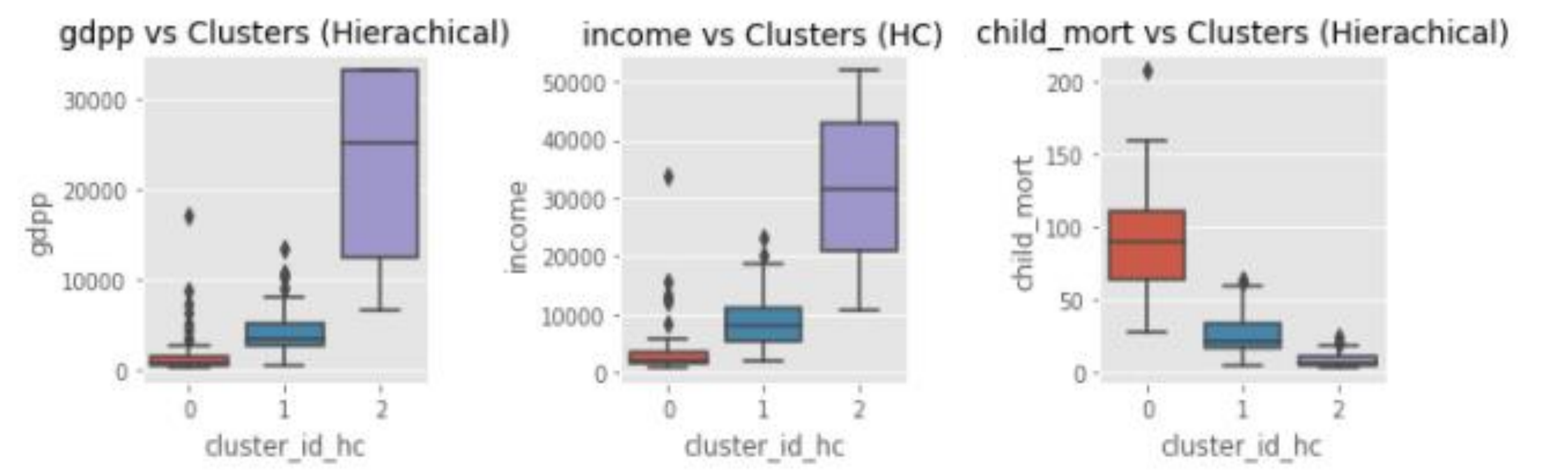


Complete Linkage

Hierarchical Clustering with Complete Linkage (right) yielded stable and more interpretable dendogram than that of Single Linkage (left)

Complete Linkage method is chosen here.

CLUSTER PROFILING : HIERARCHICAL CLUSTERING



Cluster 0 : 48 Under-developed Countries (low gdpp and income, high child mortality)

Cluster 1 : 59 Developing Countries (medium gdpp and income, medium child mortality)

Cluster 2 : 60 Developed Countries (high gdpp and income, low child mortality)

Countries in Cluster 0 (Hierarchical Clustering) are suitable for getting relief fund.

FINAL MODEL

Main Driving Factors : gdpp, income and child_mort are 3 main driving factors for clustering.

Low gdpp and income imply high rate of child mortality.

Life expectancy in the under-developed countries is low because of high child mortality rate.

Final Model: Hierarchical Clustering is chosen for final model as

- It does not need pre-specified number of clusters
- Always produce the same clusters unlike Kmeans
- Generates an inverted tree-like structure (dendogram) which helps in visualization

TOP 15 UNDER-DEVELOPED COUNTRIES

This is the list of top 15 under-developed countries.

1. Burundi
2. Liberia
3. Congo, Dom. Rep.
4. Niger
5. Sierra Leone
6. Madagascar
7. Mozambique
8. Central African Republic
9. Malawi
10. Eritrea
11. Togo
12. Guinea –Bissau
13. Afghanistan
14. Gambia
15. Rwanda

This list is sorted in

- Ascending order of gdpp (GDP per capita) and income
- Descending order of Child mortality

Top 15 countries are same for both K-Means and Hierarchical Clustering.

THANK YOU