CLUSTERING ASSIGNMENT

Piyush Rajarshi

Business Overview

- Help International Humanitarian NGO is committed to fighting poverty & providing people of backward countries with basic amenities & relief during time of disasters & natural calamities.
- Recently this organization is able to raise $10M fund.
- CEO of organization wants to use this fund strategically & effectively to provide aid to the countries which are in urgent need of aid.

Analysis Approach

- Imported required libraries to perform the analysis.
- Data in csv file format is read in the system and dimension of the dataframe is checked i.e. rows & columns are understood.
- Data is cleaned by correcting the name, check for the missing values present in the data if any.
- Actual values were obtained for the column exports, imports, health as they were present in the data frame as % of the GDP per capita.
- Statistical understanding of the data frame is build by describing the dataframe in terms of metric like mean, median, minimum, maximum & %tile(from 0 to 100).
- Correlation of different variables present in the dataframe is visualized.
- Various plots are plotted to understand the percentage of countries(%) distributed in the range present.
  - Some of the plots are :
  - GDP Range vs No. of Countries (in %)
  - Bar plot of the child mortality rate

Analysis Approach (continued…)

- Income Range vs No. of Countries (in %)
- Inflation Range vs No. of Countries (in %)
- Life Expectancy Range vs No. of Countries (in %)
- Total Fertility Range vs No. of Countries (in %)

- Outliers are visualized using boxplot and treated for the variable present in the data frame.
- For child mortality column outliers are not treated as it is possible that countries which are in require of aid might get hampered, therefore it is left as it is.
- Hopkins score is checked to find out how fit the provided data is to build and form clusters.
- Continuous variables are scaled using the StandardScaler so as to bring these variable in the same scale as to remove the possibility of the values being in different range.
- Initially clustering modelling is done using the KMeans algorithm so initially value of "K" is determined using the silhouette score & elbow curve.

Analysis Approach (Continued…)

- Silhouette score determines k value by separation distance between the clusters.
- Elbow curve determines the k value by calculating the sum of squared errors (SSE).

- Model is formed using KMeans & Clusters are determined (clusters thus obtained are 4).
- Clusters & countries are visualized.
- Model is formed using Hierarchical clustering and clusters & dendrogram obtained using complete linkage are cut to form 4 clusters.
- Five Countries which are in urgent need of aid are filtered and displayed.

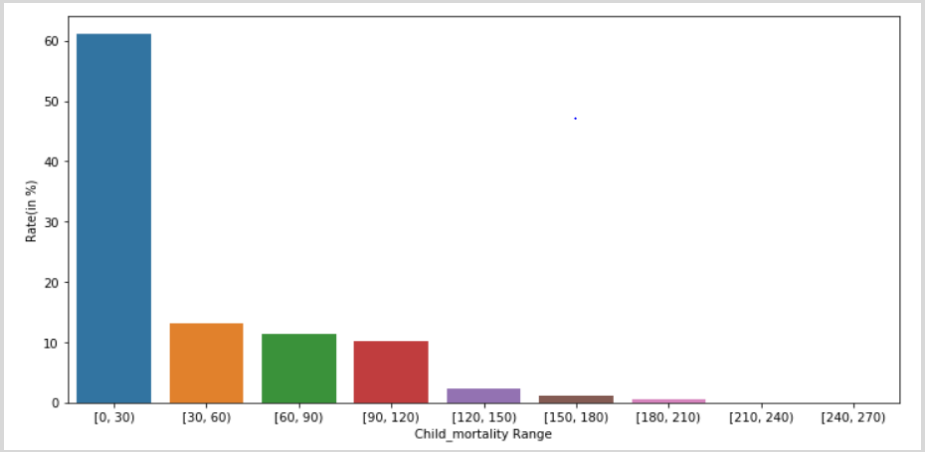
Outcomes of Model Formed

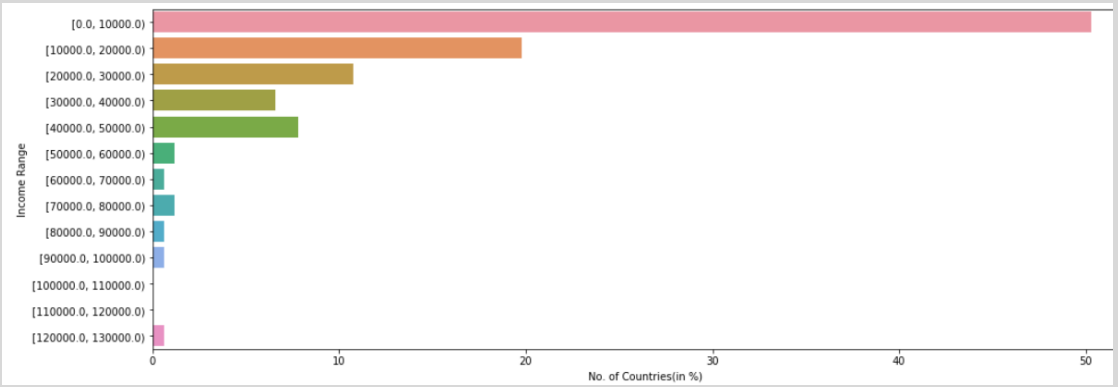
- KMeans Cluster
  - The clusters obtained after formation of model is 4.
  - Cluster formed are divided as below :-
    - Cluster – 0 : Moderately High Child Mortality, Moderate Low (Income & GDP)
    - Cluster – 1 : Low Child Mortality, High (Income & GDP)
    - Cluster – 2 : Moderately Low Child Mortality, Moderate High (Income & GDP)
    - Cluster – 3 : High Child Mortality, Low (Income & GDP)
- Hierarchical Cluster
  - The clusters obtained after formation of model is 4.
  - Cluster formed are divided as below :-
    - Cluster – 0 : High Child Mortality, Low (Income & GDP)
    - Cluster – 1 : Moderately High Child Mortality, Moderately Low (Income & GDP)
    - Cluster – 2 : Moderately Low Child Mortality, Moderately High (Income & GDP)
    - Cluster – 3 : Low Child Mortality, High (Income & GDP)

Visualizations

Child Mortality Range vs Rate(in %) here rate is no of countries.

- Around 61% of countries have child mortality range in 0 – 30.
- Around 31% of countries have child mortality range greater than 120.



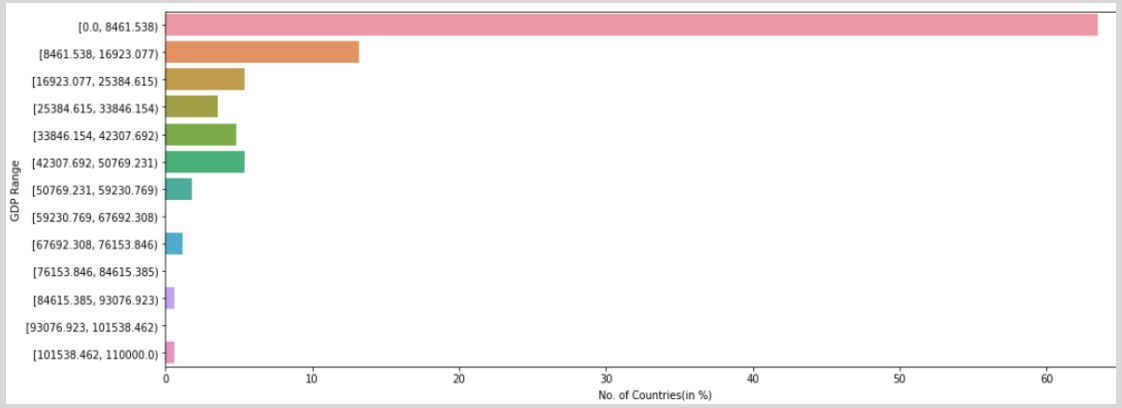


Income Range vs No. of Countries(in %)

- Around 50% countries have income within 10k.
- Approx. 3.7% countries have income greater than 60k.
- Approx. 47% countries have income in range 10k to 60k.

GDP vs No. of Countries(in %)

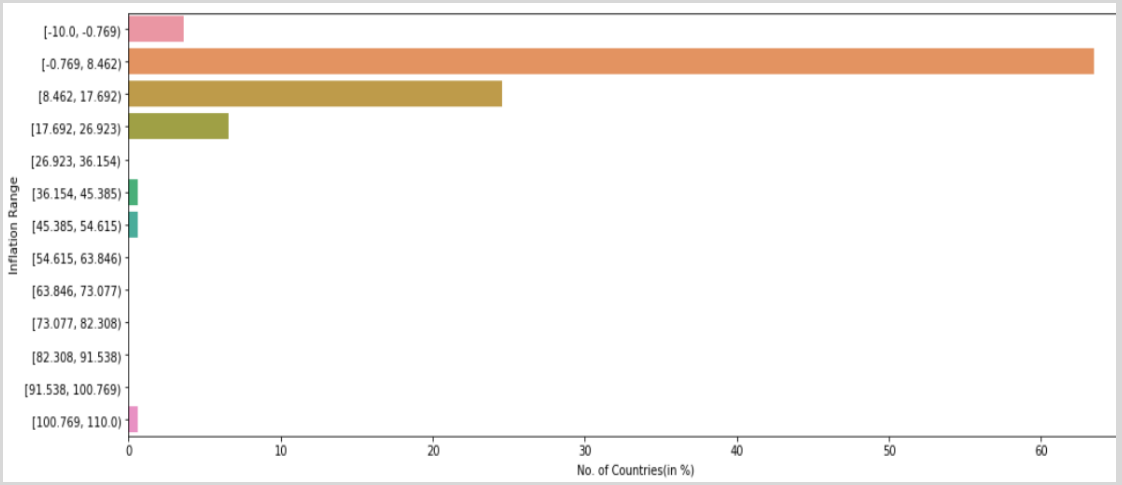
- Around 63% countries have GDP within 9k.
- Approx. 4.2% countries have GDP greater than 50k.
- Approx. 33% countries have GDP in range 9k to 50k.



Visualizations

Inflation vs No. of Countries

- 3.6% countries have negative inflation rate in range -10 to -0.77
- 63% countries have inflation rate in between -0.77 to 8.4.
- 24% countries have inflation rate in between 8.4 – 17.7.
- 8% countries have inflation rate greater than 17.7.

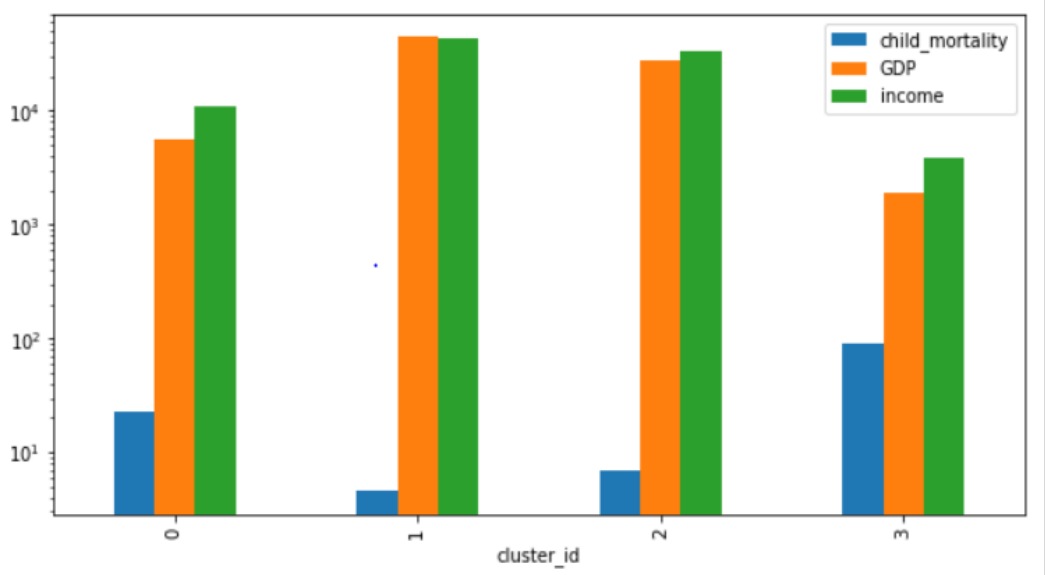


KMeans model Outcome

- Cluster – 0 : Moderately High Child Mortality, Moderate Low (Income & GDP)
- Cluster – 1 : Low Child Mortality, High (Income & GDP)
- Cluster – 2 : Moderately Low Child Mortality, Moderate High (Income & GDP)
- Cluster – 3 : High Child Mortality, Low (Income & GDP)

Country of various cluster as below :-

Cluster – 0	**Cluster – 1**	**Cluster – 2**	**Cluster – 3**
• Myanmar • Turkmenistan • India • Tajikistan • Bangladesh	• Qatar • UAE • Malta • Canada • Switzerland	• Saudi Arabia • Seychelles • Barbados • Bahamas • Oman	• Haiti • Sierra Leone • Chad • Central African Republic • Mali

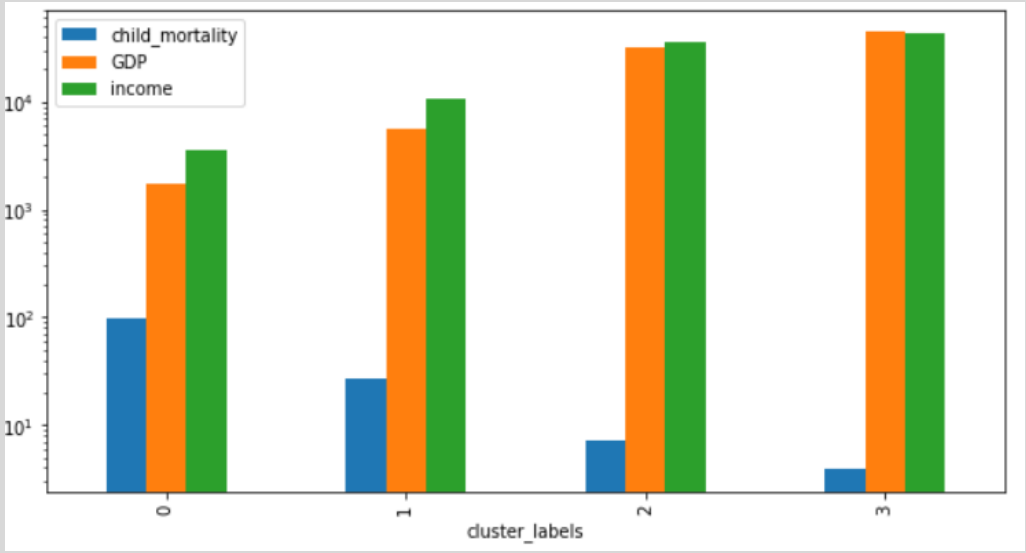


Hierarchical Cluster Model Outcome

- Cluster – 0 : High Child Mortality, Low (Income & GDP)
- Cluster – 1 : Moderately High Child Mortality, Moderately Low (Income & GDP)
- Cluster – 2 : Moderately Low Child Mortality, Moderately High (Income & GDP)
- Cluster – 3 : Low Child Mortality, High (Income & GDP)

Countries of the clusters as below :-

Cluster – 0	Cluster – 1	Cluster – 2	Cluster – 3
• Haiti • Sierra Leone • Chad • Central African Republic • Mali	• Lesotho • Pakistan • Lao • Myanmar • Kiribati	• Libya • Saudi Arabia • Bahamas • Oman • Kuwait	• Malta • Belgium • Netherlands • Switzerland • Austria



Thank You !!!!