HELP International NGO Funding

Clustering Assignment

(K-means & Hierarchical Clustering)

Segmenting Countries Based On Socio-Economic Factors For Funding

Pawan Dixit

Index

- *Overview*
- *Technical Approach*
- *Steps for Analysis*
- Heat Matrix
- KMeans Clustering
  - K value
  - Scatter Plot Visualization – GDPP, INCOME & CHILD_MORT
  - Cluster Profiling
- Hierarchal Clustering
  - Single & Complete Linkage
  - N-Cluster = 3
  - N-cluster =4
- Summary

Overview

Background:

HELP International is an international humanitarian NGO that is committed to fight poverty and raise awareness in the people of backward countries

To provide people, children and their families from the chosen neediest countries during the time of disasters and natural calamities :

- Funding for basic amenities and relief equipment and tools
- Raising awareness and educating people

Problem Statement

To categorise the countries using some socio-economic and health factors that determine the overall development of the country.

To identify some countries which need to focus on the most.

Technical Approach

Data Collection &Cleaning

- Importing Data
- Identify data quality issues & clean data i.e. Converting data into absolute value

Outlier Treatment

- Removing outlier wherever required using Box plot

Data Visualization

- Visualisation of data variable using graph i.e. Displot/box plot/bar plot & Heat map

Scaling Data

- Standardising all continuous data using Standard scalar or Min/Max technique

Hopkin Test

- To Check if data has tendency to form clusters

KMean Clustering

- Visualizes data using Scatter plot

Hierarchical Clustering

- Identify single and complete linkage

Decision Making

- Identify the countries which requires aid by analysing both KMeans & Hierarchical Clustering results

Steps for Analysis

- Preprocessing:
  - Collect and clean data from any garbage values and duplicate records
  - Perform outlier analysis and cap the values in 5-95% quartile during processing, while considering original values for the final analysis
  - Visualize data to identify patterns or correlations, and select only relevant and important features for analysis
  - Scale up/down the continuous features to the same range, for correct working of ML algorithms
  - Perform Hopkins test to check if data has tendency to form clusters

- Cluster Analysis:
  - Perform K-means analysis and Hierarchical (complete-linkage and single-linkage) analysis
  - Identify optimum number of clusters

- Cluster Profiling:
  - Cluster data visualization
  - Scores
  - Story building around each of the clusters

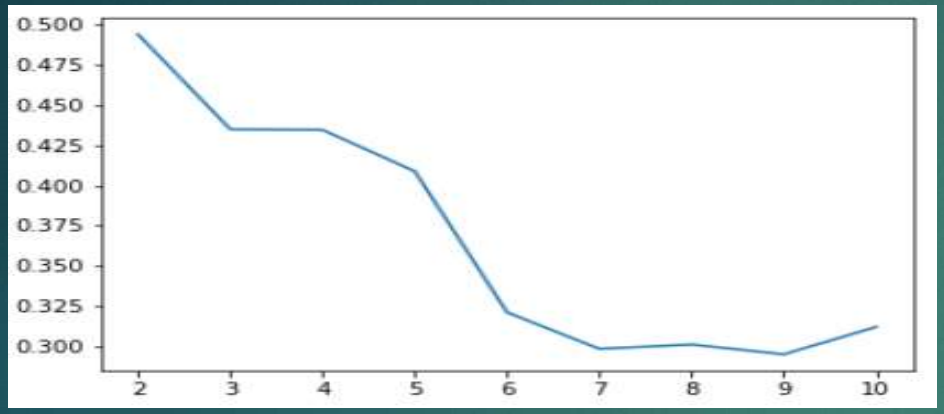
Visualization & Summary

Heat Matrix

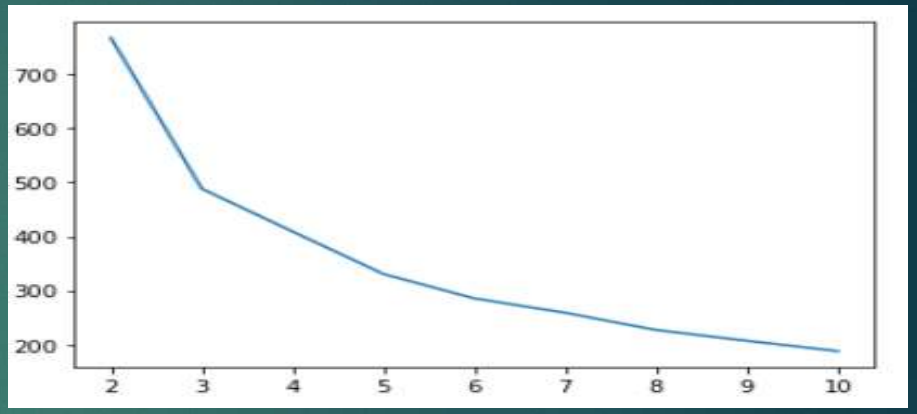
- After data cleaning , we capped lower range outlier of CHILD_MORT, INFLATION, TOTAL_FERTI. Moreover, capped upper range outliers for other columns.
- We did standardized scaling to standardize all parameters on cleaned, outlier removed data
- We see high correlation between `total_fer` and `child_mort`, between `gdpp` and `income`,and between `imports` and `exports`.

Visualization & Summary

KMean Clustering: Find the best value of k:



Silhouette Score



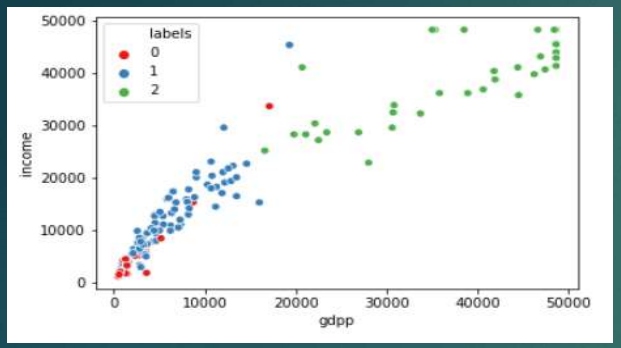
SSD: Sum of Squared Distance

By looking Silhouette analysis, we see the highest peak is at k = 2 and in sum of squared distances graph, we see that the comfortable elbow value is at 3. Hence Final KMeans with K = 3.

Visualization & Summary

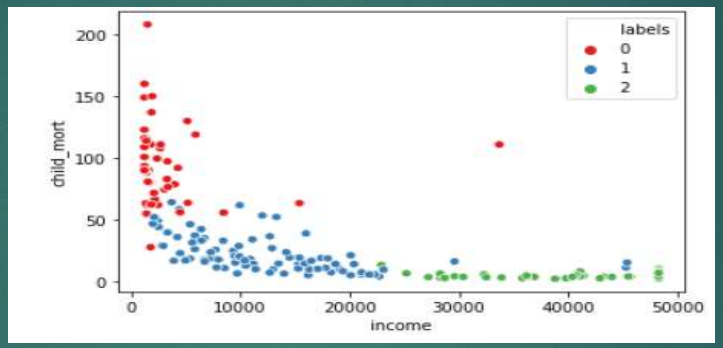
KMean Clustering: Scatter Plot visualization

GDPP Vs INCOME



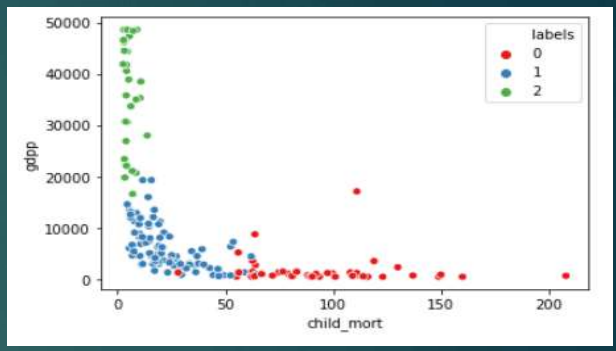
Scatter plot of income, gdpp for 3 cluster. We can see that cluster 0, both gdpp and income per person is very low

INCOME Vs CHILD_MORT



Scatter plot of income, child mort for 3 cluster. We can see that cluster 0 income is very low and child mort is high.

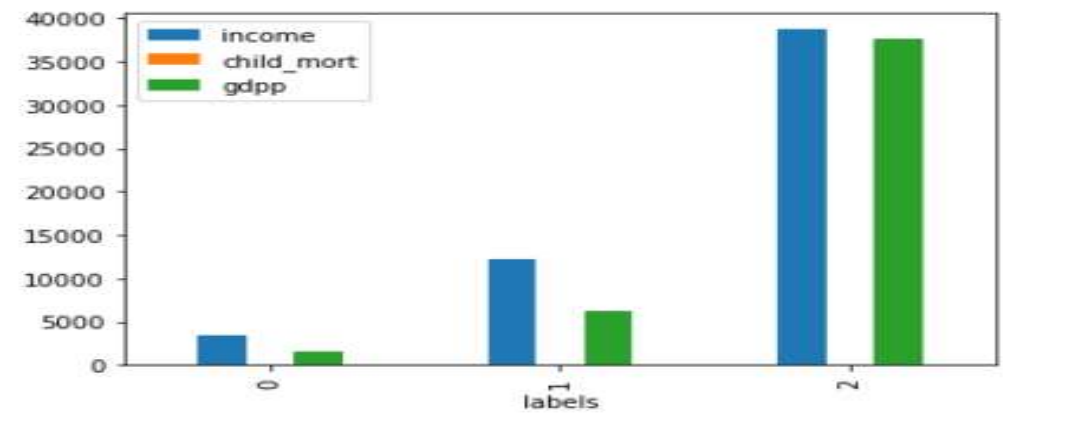
CHILD MORT Vs GDPP



Scatter plot of child mort, gdpp for 3 cluster. We can see that cluster 0 gdpp is low and child mort is high.

Visualization & Summary

KMean Clustering: Cluster Profiling

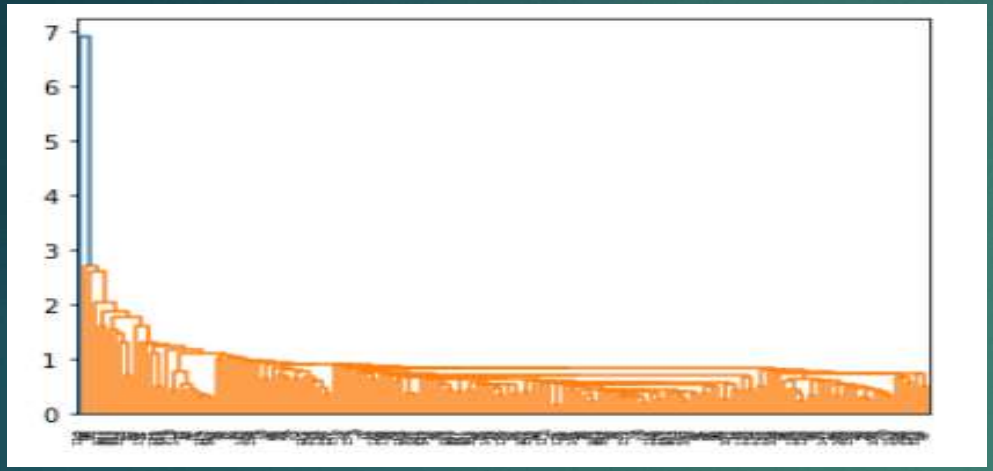


labels	income	child_mort	gdpp
0	3374.822222	94.537778	1651.757778
1	12317.529412	22.860000	6278.847059
2	38711.081081	5.237838	37745.675676

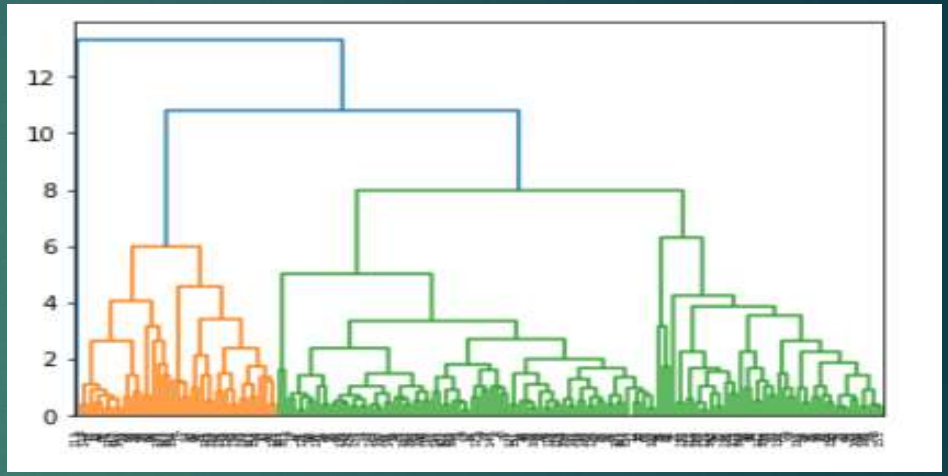
As per K Mean Clusters, Cluster 0 is area of concern as Low income (3374.82, High Child mort (94.53) and low gdpp (1651.75)

Visualization & Summary

Hierarchal Clustering : Single & Complete Linkage



Single Linkage method Hierarchal Clustering



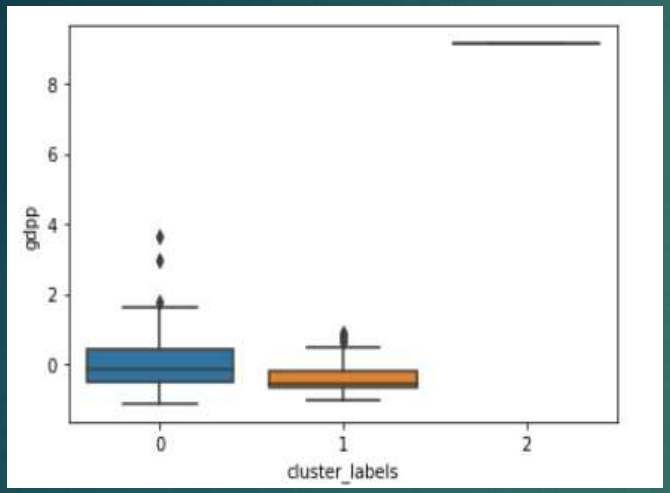
Complete Linkage method Hierarchal Clustering

We are going to use this method as Single linkage is not clear. By looking at Dendrogram taking n-cluster as 3

Visualization & Summary

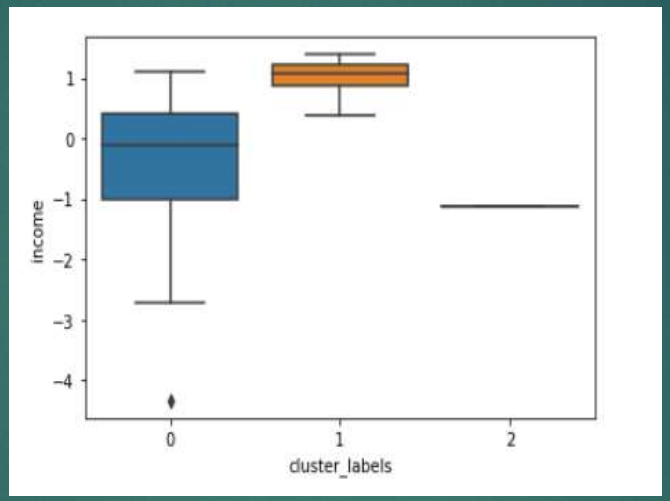
Hierarchal Clustering: n-cluster = 3

Custer Labels Vs GDPP



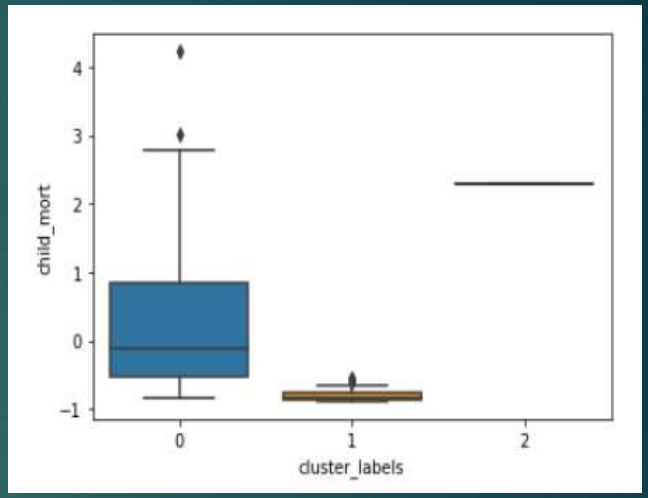
Box plot of gdpp and cluster labels for n-cluster as 3, we can see that cluster 0 Q3 is higher than other clusters.

Custer Labels Vs INCOME



Box plot of income and labels for n-cluster as 3, we can see that cluster 0 low income compare to cluster 1

Custer Labels Vs CHILD_MORT

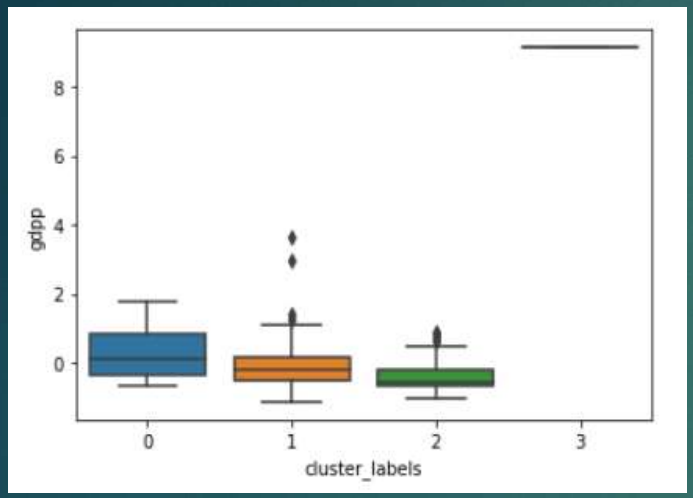


Box plot of child mort and labels for n-cluster as 3, we can see that cluster 0 is high in child mort as compare to other clusters.

Visualization & Summary

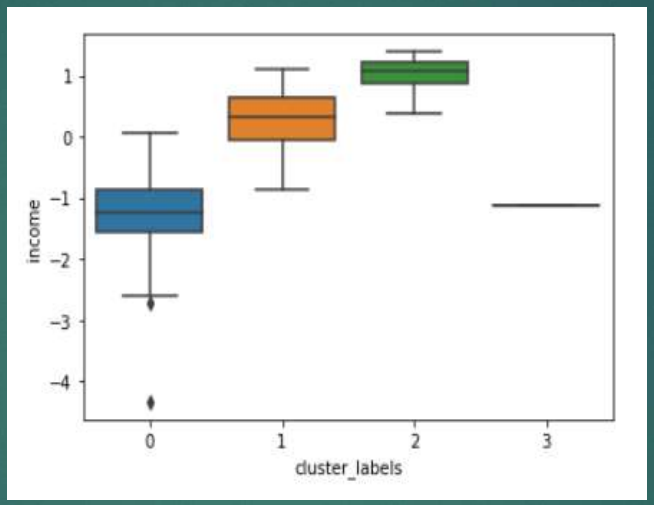
Hierarchal Clustering: n-cluster = 4

Cluster Labels Vs GDPP



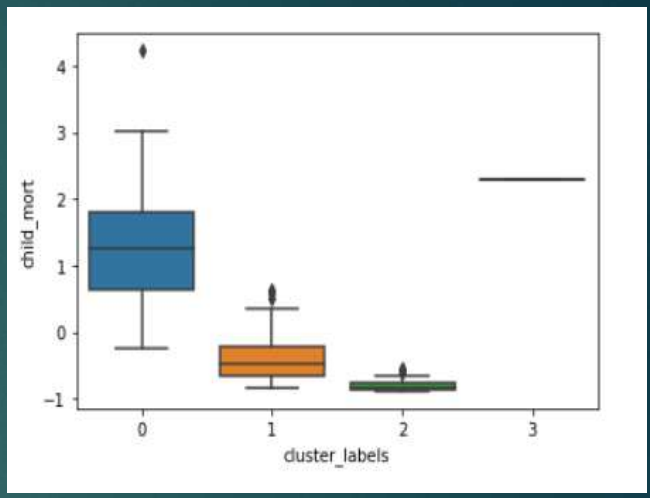
Box plot of gdpp and lables for n-cluster as 4, we can see that cluster 2 have low gdpp compare to others clusters. Its also observed there are some outliers in cluster 1.

Cluster Labels Vs INCOME



Box plot of income and label n-cluster as 4, we can cluster 0 have low income as compare to other clusters.

Cluster Labels Vs CHILD_MORT



Box plot of child mort and label n-cluster as 4, we can child mort is high in cluster 0 as compare to other clusters.

Summary:

The following are the countries which are in direst need of aid by considering socio – economics factor into consideration:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	labels
132	Sierra Leone	160.0	70.4688	52.26900	169.281	1220.0	17.20	55.0	5.20	465.9	0
31	Central African Republic	149.0	70.4688	26.71592	169.281	1213.0	2.01	47.5	5.21	465.9	0
112	Niger	123.0	77.2560	26.71592	170.868	1213.0	2.55	58.8	7.49	465.9	0
37	Congo, Dem. Rep.	116.0	137.2740	26.71592	169.281	1213.0	20.80	57.5	6.54	465.9	0
106	Mozambique	101.0	131.9850	26.71592	193.578	1213.0	7.64	54.5	5.56	465.9	0
26	Burundi	93.6	70.4688	26.79600	169.281	1213.0	12.30	57.7	6.26	465.9	0
94	Malawi	90.5	104.6520	30.24810	169.281	1213.0	12.10	53.1	5.31	465.9	0
88	Liberia	89.3	70.4688	38.58600	302.802	1213.0	5.47	60.8	5.02	465.9	0
93	Madagascar	62.2	103.2500	26.71592	177.590	1390.0	8.79	60.8	4.60	465.9	0
50	Eritrea	55.2	70.4688	26.71592	169.281	1420.0	11.60	61.7	4.61	482.0	0

1. Sierra Leone
2. Central African Republic
3. Niger
4. Congo, Dem. Rep
5. Mozambique
6. Burundi
7. Malawi
8. Liberia
9. Madagascar
10. Eritrea