HELP International NGO Funding

Clustering Assignment

(K-means & Hierarchical Clustering)

Segmenting Countries Based On Socio-Economic Factors For Funding

Created by Sheetal Atre

1. Overview

- *Project statement and Rationale*
- 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

- *Project Objectives*
- 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.

2. Technical Approach

Segmenting the countries into aided and non-aided using "clustering“ algorithm

- Use K-Means Cluster method to build the cluster model.
- Use Hierarchical clustering to build the cluster model and compare results with K-means Cluster method.
- Use Silhouette and Elbow method to validate the optimal cluster values .
- Identify the cluster representing the countries which are in the direst need of aid using cluster mean method.
- Analyze the final cluster statistics against other clusters
- Decision making on the final list based on the descriptive statistics of the selected cluster
- Choose the top 10 countries from the final cluster based on high child mortality, low income and low gdpp.

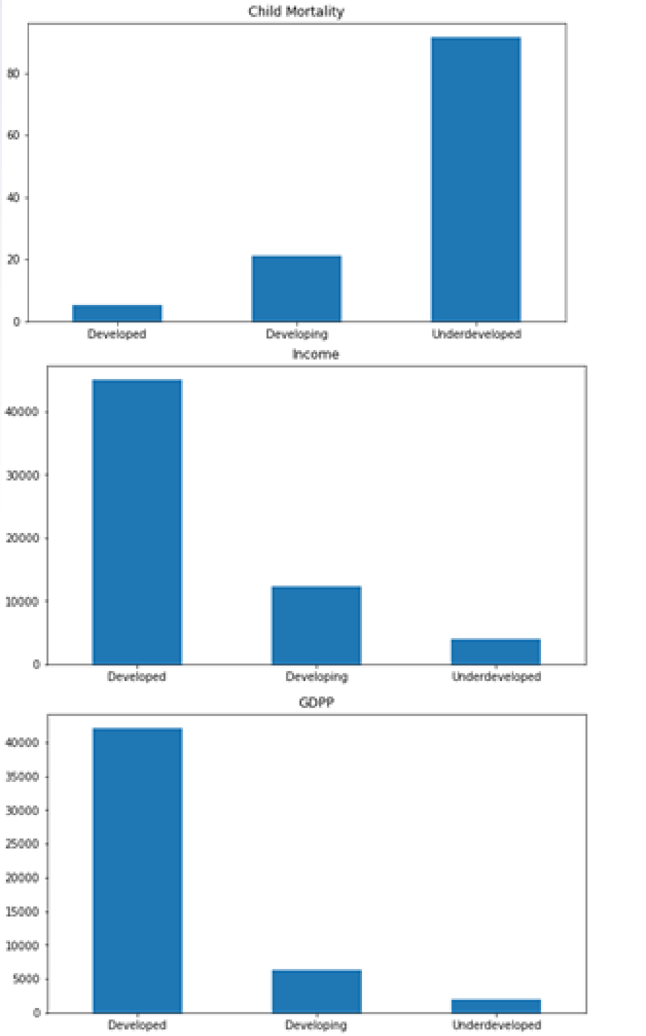
3. Steps for analysis

- 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

4. Cluster Summary

- The final model generated k=3 clusters.
- Based on descriptive analysis, we have identified clusters as:
  - Underdeveloped countries
  - Developing countries
  - Developed countries

- Countries in Underdeveloped cluster have:
  - Lowest average GDPP (approx. 1879.14)
  - Lowest average income(approx. 3900.47)
  - Highest average child mortality(approx. 91.55)

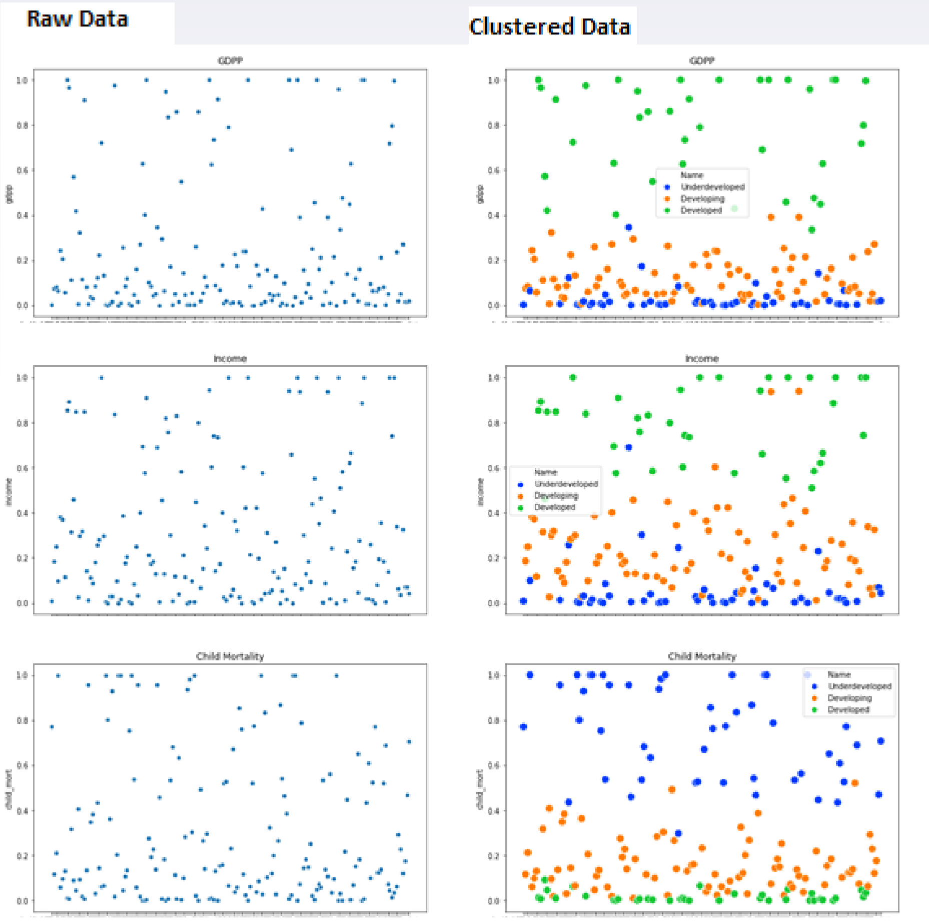


5. Raw Data Vs Clustered Data

Cluster_Id	Child Mortality	Income	GDPP
Developed	5.24	45056.76	42102.70
Developing	21.13	12406.67	6359.70
Underdeveloped	91.55	3900.47	1879.14

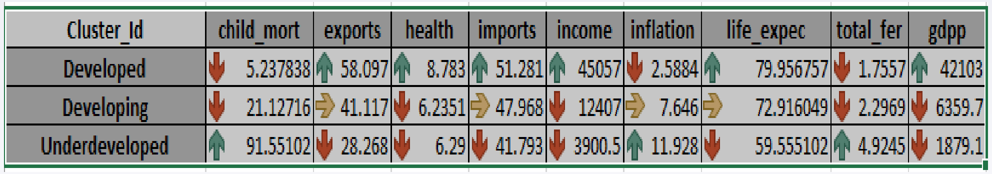
- We have clustered the raw data according to GDPP, Income and Child Mortality values. This has produced three distinct clusters.

- The selected clustering algorithm(K-means) aims to maximise the similarity between the data points in the same cluster and minimise the similarity between data points in different clusters.



6. Country Segments

(Story-building around countries in the clusters)



Cluster_Id	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Developed	↓ 5.237838	↑ 58.097	↑ 8.783	↑ 51.281	↑ 45057	↓ 2.5884	↑ 79.956757	↓ 1.7557	↑ 42103
Developing	↓ 21.12716	⇨ 41.117	↓ 6.2351	⇨ 47.968	↓ 12407	⇨ 7.646	⇨ 72.916049	↓ 2.2969	↓ 6359.7
Underdeveloped	↑ 91.55102	↓ 28.268	↓ 6.29	↓ 41.793	↓ 3900.5	↑ 11.928	↓ 59.555102	↑ 4.9245	↓ 1879.1

Developed Countries:

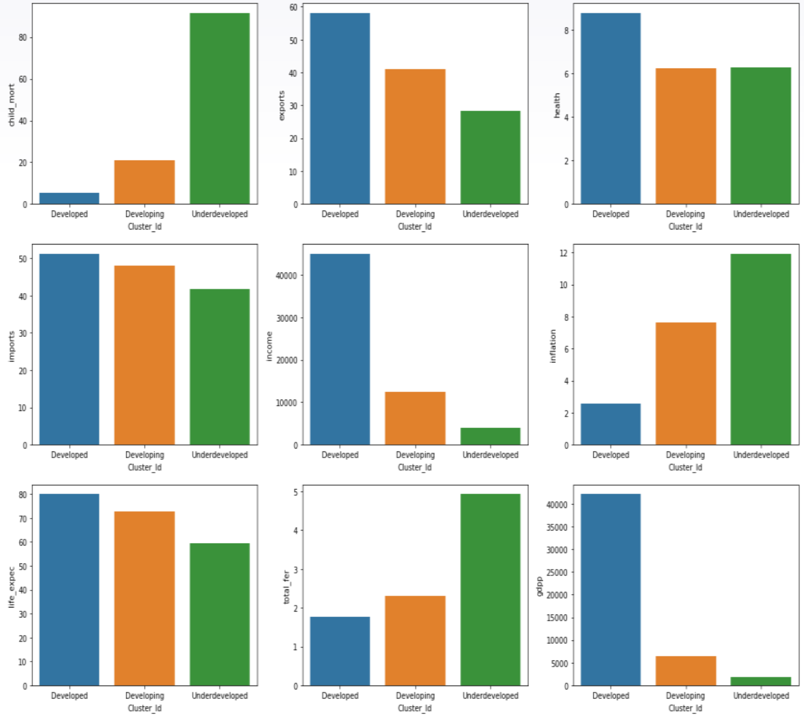
This cluster has higher percentage of healthy population and higher life expectancy. They have low child mortality rates and their fertility rate is also lowest. Consequently they are able to provide the best services in terms of imports/exports which is clearly reflected in their high mean income and very low inflation rates. From their GDPP these countries do not need aid at all.

Developing Countries:

This cluster has healthy population, good life expectancy and fairly good fertility rate. However they have high child mortality rates. These countries are able to provide good services in terms of exports, but their imports are higher than exports. This is clearly reflected in their lower mean income and high inflation rates. Though these countries have much lower GDPP, many of the are not in need of aid compared to the other underdeveloped countries.

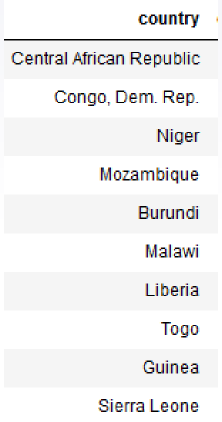
Underdeveloped Countries:

This cluster consists of healthy population and very high fertility. Yet they have very high child mortality, low life expectancy. They have very low exports, and comparatively they have very high imports of services. They very low income, very high inflation and extremely low GDPP. These countries need in dire need of external financial help and support. Hence we will focus on these countries.



7. Top 10 Recommended Countries
(Statistically)

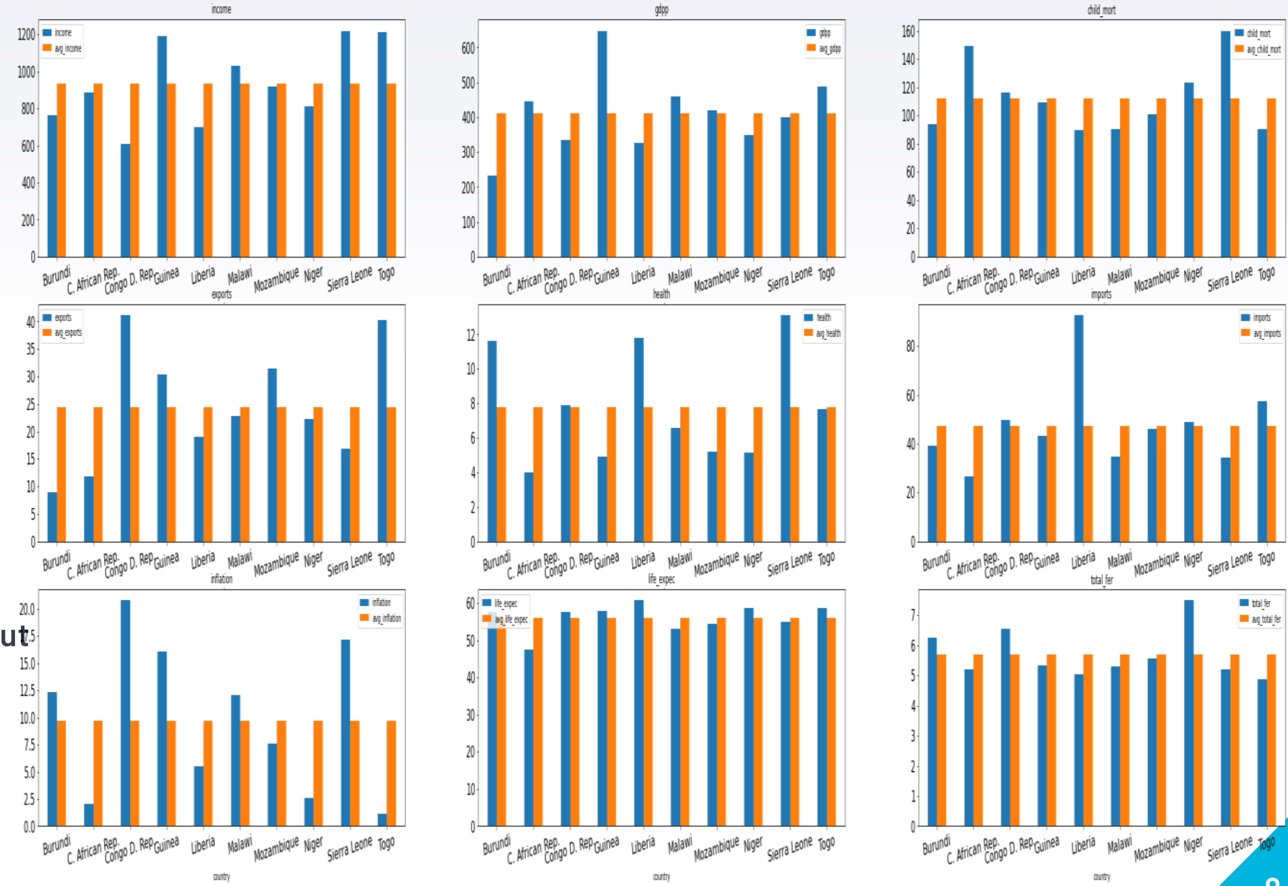
- We recommend following top 10 countries from the Underdeveloped segment for funding



country
Central African Republic
Congo, Dem. Rep.
Niger
Mozambique
Burundi
Malawi
Liberia
Togo
Guinea
Sierra Leone

- We have selected these countries out of all the other 49 countries of Underdeveloped segment based on lowest mean GDPP, lowest mean income and highest average child mortality rate

Recommended countries - Comparison of other factors with cluster means



THANKS!