CLUSTERING ASSIGNMENT

(K-MEANS AND HIERARCHICAL)

Contents

1. Background
2. Technical Approach
3. Outlier Treatment
4. Exploratory Data Analysis
5. Hierarchical Clustering
6. Finding the Optimal Number of Clusters
7. Cluster Summary
8. Final list of countries from the cluster – “Under-developed Countries”
9. Statistics
   (of our recommended countries)

Background

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programme, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

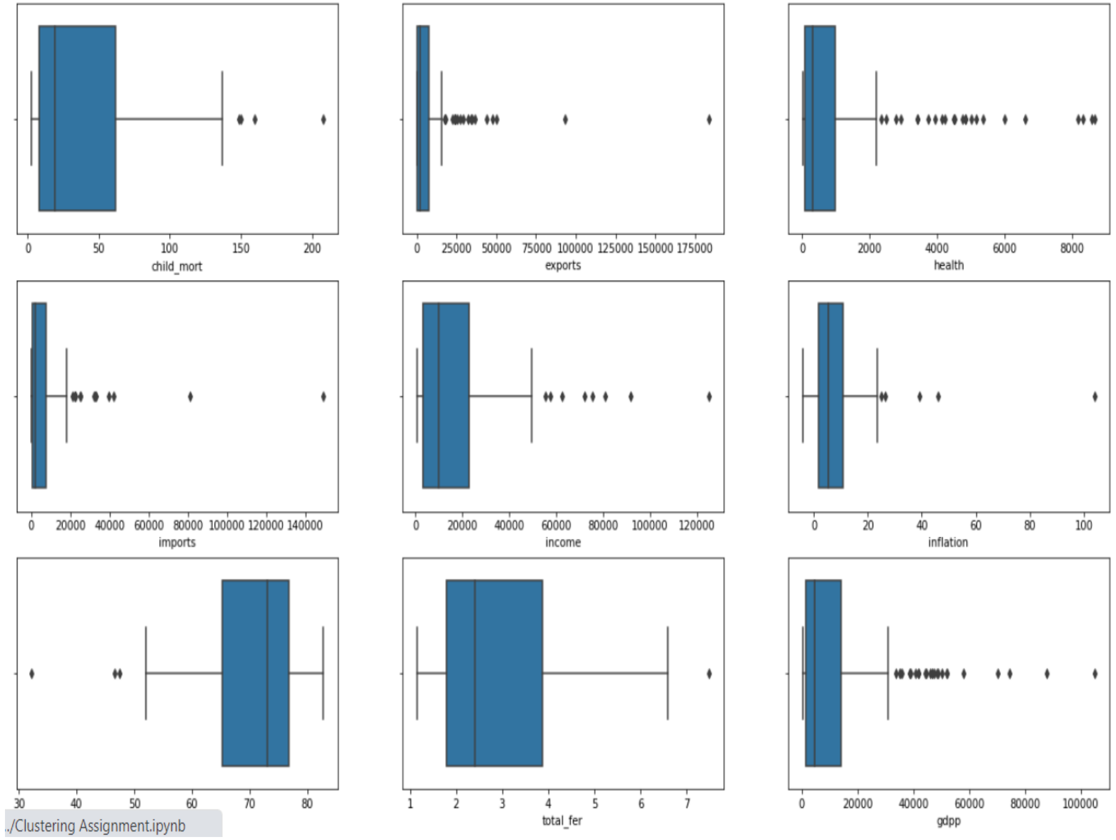
Technical Approach

- Using Hierarchical clustering to identify optimal cluster value.
- Using Silhouette and elbow method to validate the optimal cluster values.
- Using K-Means Cluster method to build the final cluster model.
- Identify appropriate cluster for financial aid using cluster mean method.
- Analyze the final cluster against all other clusters.
- Decision making on the final based on the descriptive statistics of our final cluster.
- Choose the top 10 countries from the final cluster based on higher child mortality and lower income and gdpp.
- Present the final report.

Outlier Treatment

There seems to be outliers in every single variable. This is a very delicate situation in terms of Business problem statement & Clustering analysis.

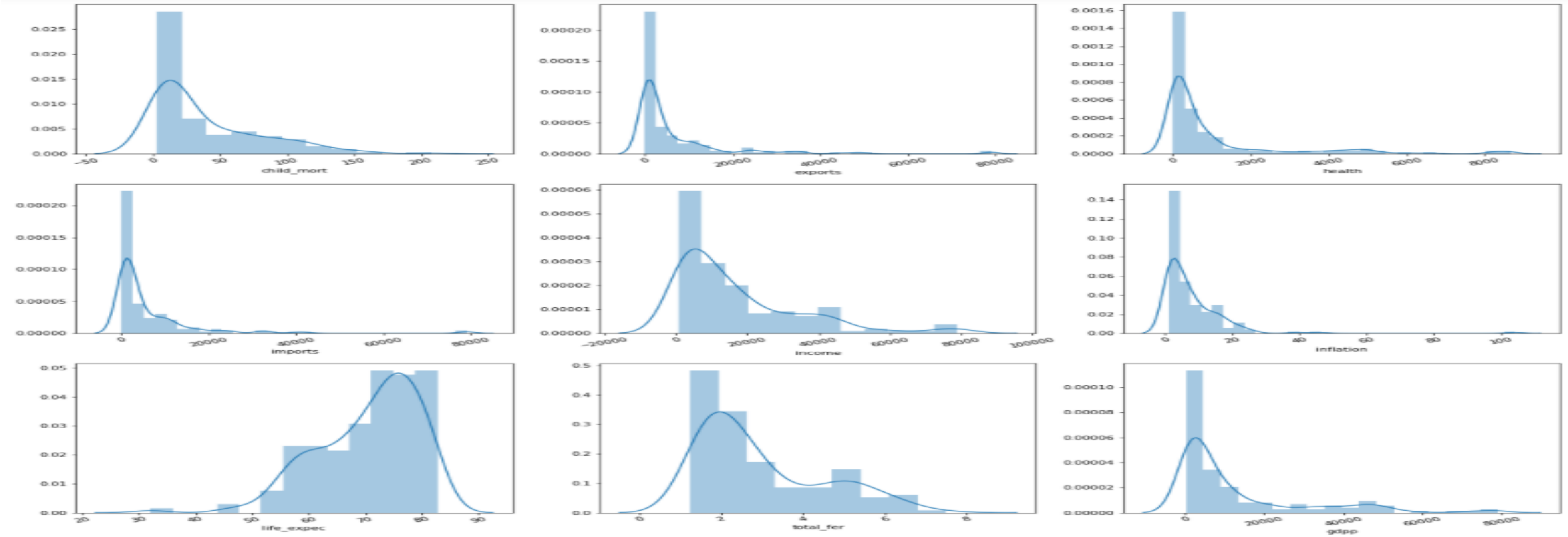
- If we apply outlier treatment by CAPPING this will change the ranking of few countries with respect to requirement of Financial Aid. Also we will still have some outliers present after Capping, so we need to make a wise decision considering this.
- If we apply outlier treatment by Deletion based on IQR values, this will remove few countries from the list that would have really deserved the Financial Aid.
- If we do not apply Outlier treatment, it can impact the clustering model, as the presence of Outlier can change the CENTROID (K-Means) of the cluster.
- After considering all these scenarios, I've decided to treat the outliers by soft capping to 99th percentile for upper limit and 1 percentile for lower limit.



Exploratory Data Analysis

Most of the data point are 'NOT Normally' distributed.

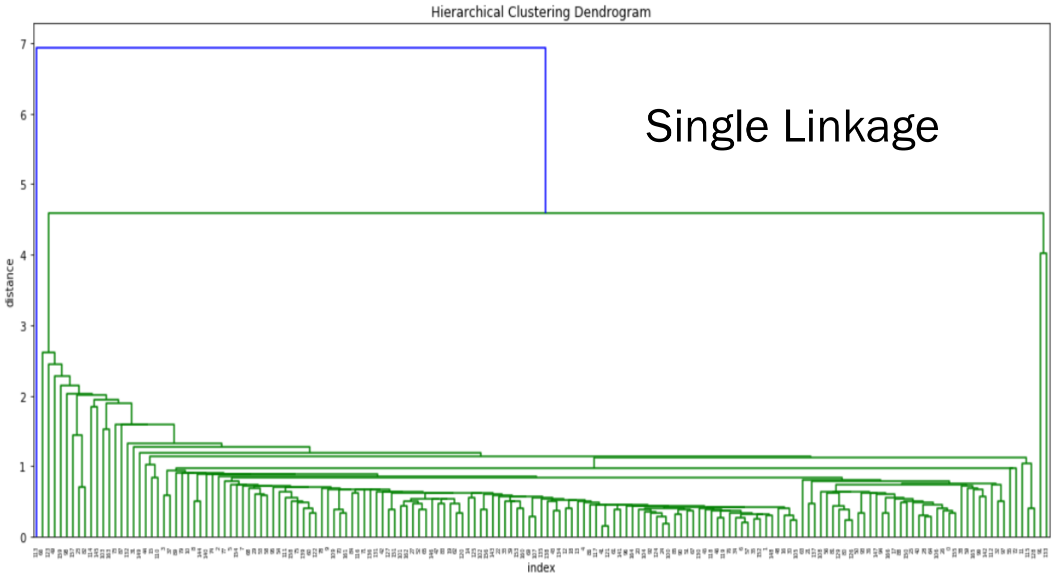
- Their variance is also different.
- Their range are also different.



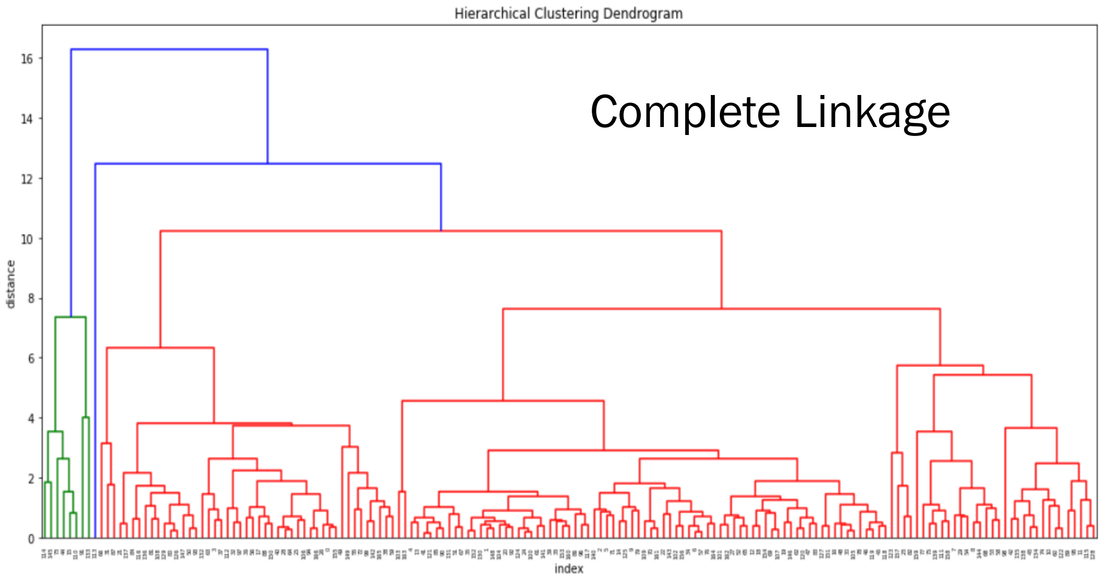
Hierarchical Clustering

We made use of Hierarchical Clustering to identify appropriate cluster size with a good split of data (Max Intra-Cluster distance & Min Inter-Cluster Distance).

From the below Dendrograms, we made use of 'Complete Linkage’ as number of clusters formed is easy to interpret and the hierarchy is clearly differentiable.



Single Linkage



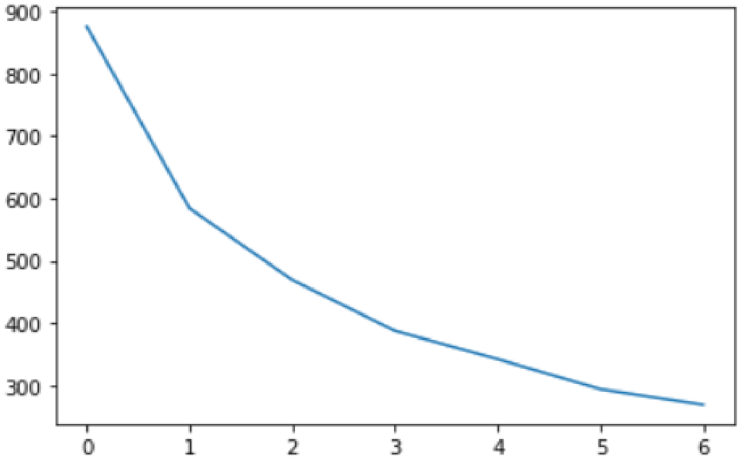
Complete Linkage

Finding the Optimal Number of Clusters

SSD (Elbow curve):

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set**.**

We could observe the elbow pattern for number of clusters = 3



Silhouette score:

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters

The silhouette score for 3 clusters is **0.43**

$$\text{silhouette score} = \frac{p - q}{max(p, q)}$$

Cluster Summary

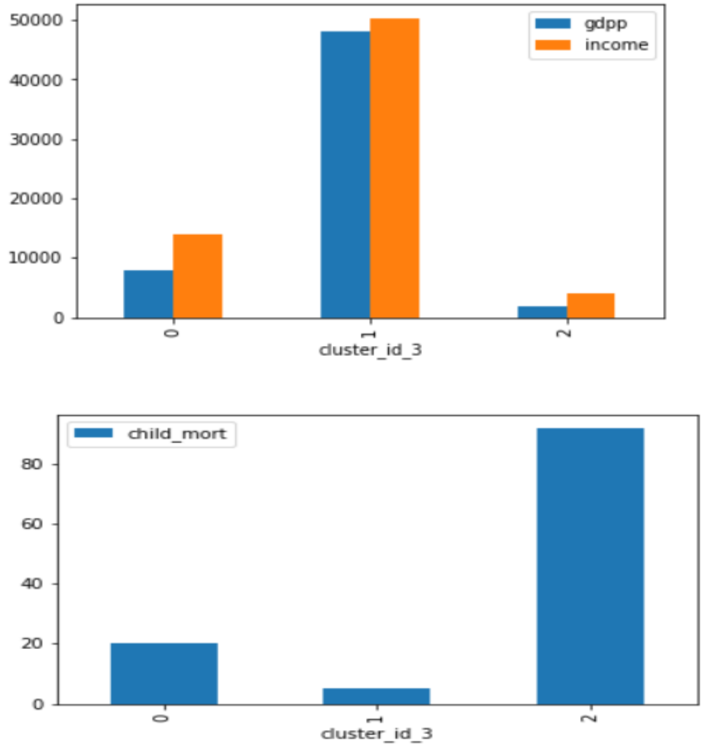
The final model generated 3 clusters. Based on their descriptive statistics, we could identify them as

1. Under developed countries
2. Developing countries
3. Developed countries

Cluster “Under developed countries” has the **highest average child Mortality rateof -92** when compared to 10 other clusters and **lowest average GDPP and Income of -1909 and 3897** respectively.

All these figures clearly makes this cluster the best candidate for the financial aid from NGO.

We could also see that cluster “Under developed countries” comprises of **-29% of overall data** and has -48 observations in comparison to 167 total observations.



Graphical representation of GDPP, Income and child mortality of 3 clusters.

Final list of countries from the cluster – “Under-developed Countries”

We concluded on the top 10 list of countries from the final cluster (“Under Developed Countries”) based on cluster median values of gdpp, income and child_mortality.

We filtered the countries with (in below order) -

- Lowest gdpp
- Lowest income
- Highest child_mortality

Countries
Burundi
Congo, Dem. Rep
Niger
Sierra Leone
Mozambique
Central African Republic
Malawi
Togo
Guinea-Bissau
Afghanistan

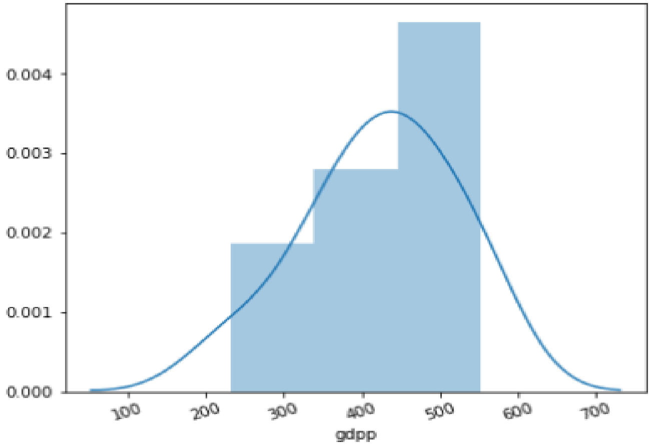
Statistics

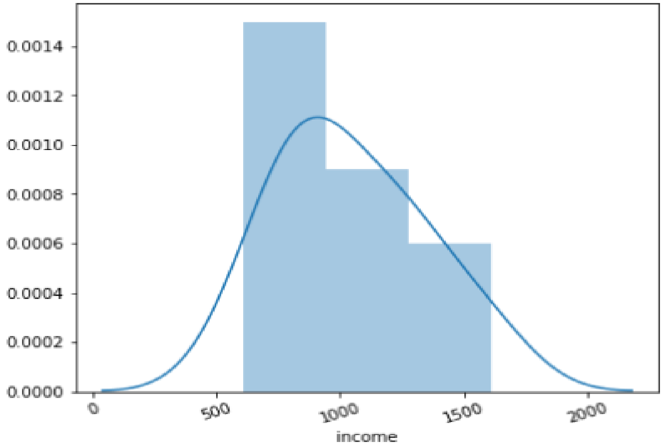
(of our recommended countries)

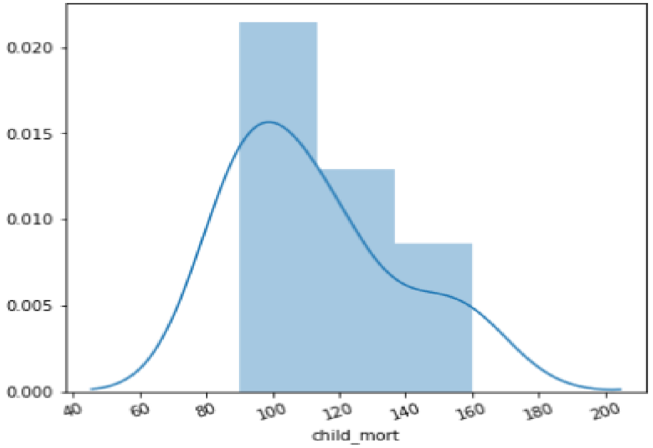
Min_GDPP= 231 Max_GDPP= 553 Median_GDPP= 432.5

Min_INCOME= 609 Max_INCOME= 1610 Median_INCOME= 974

Min_CHILD_MORT= 90 Max_CHILD_MORT = 160 Median_CHILD_MORT = 107


gdpp


income


child_mort