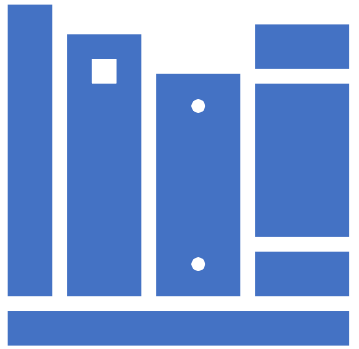


UpGrad

Countries Clustering

Yatin Bajaj

Business Objectives and Strategy

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 programs, 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.

Business Objective :

- Aim is to categorize the countries using some socio-economic and health factors that determine the overall development of the country.

- Then need to suggest the countries which the CEO needs to focus on the most.

Strategy :

- Perform PCA on the dataset and obtain the new dataset with the Principal Components.
- Use K-means and Hierarchical Clustering to form clusters.
- Analyse the clusters and identify the ones which are in dire need of aid.

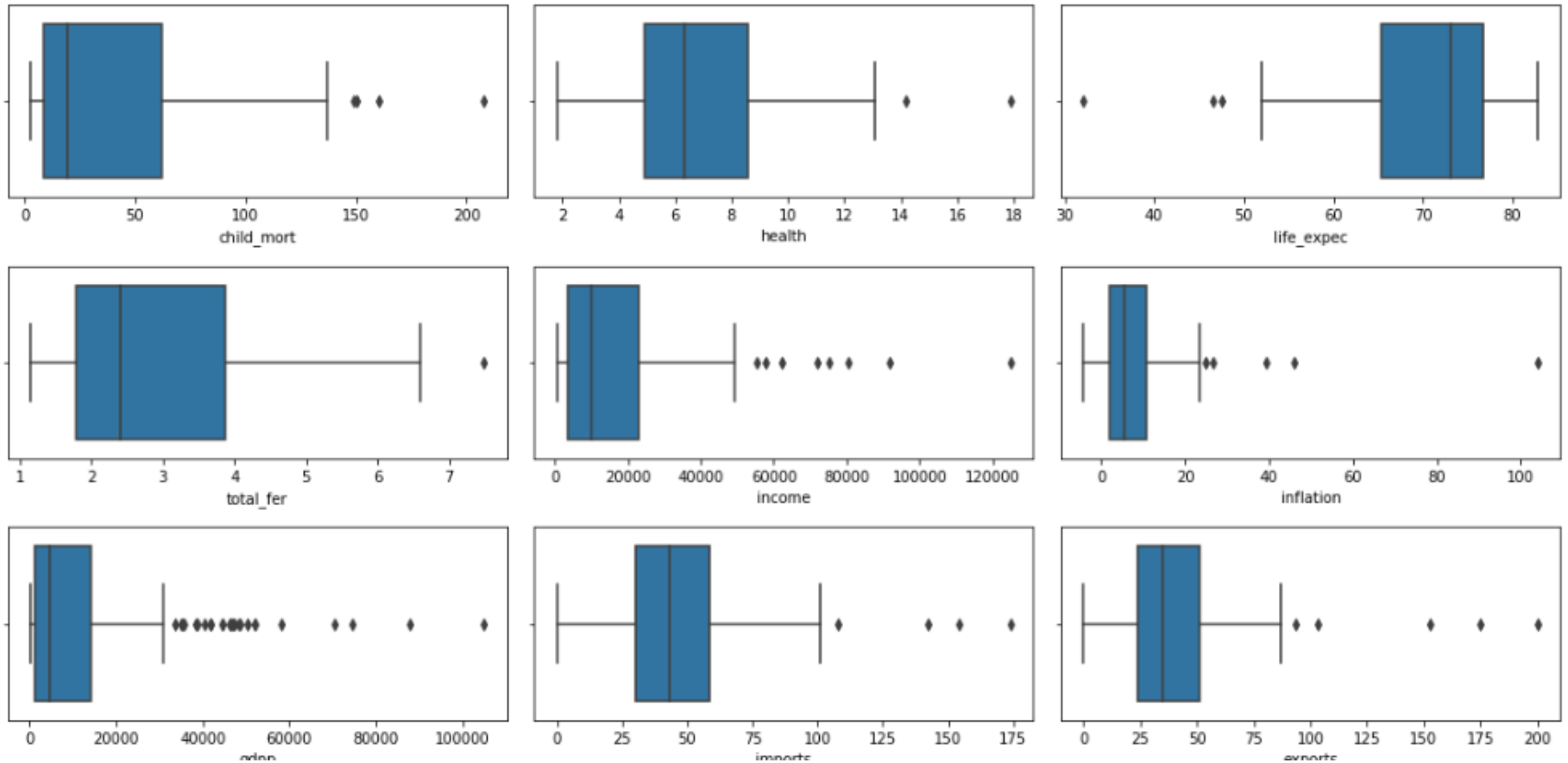
Assumption :

- Data considered is for 167 countries.
- Analysis is performed based on 9 socio-economic factors.

Basic Strategy and Methodology

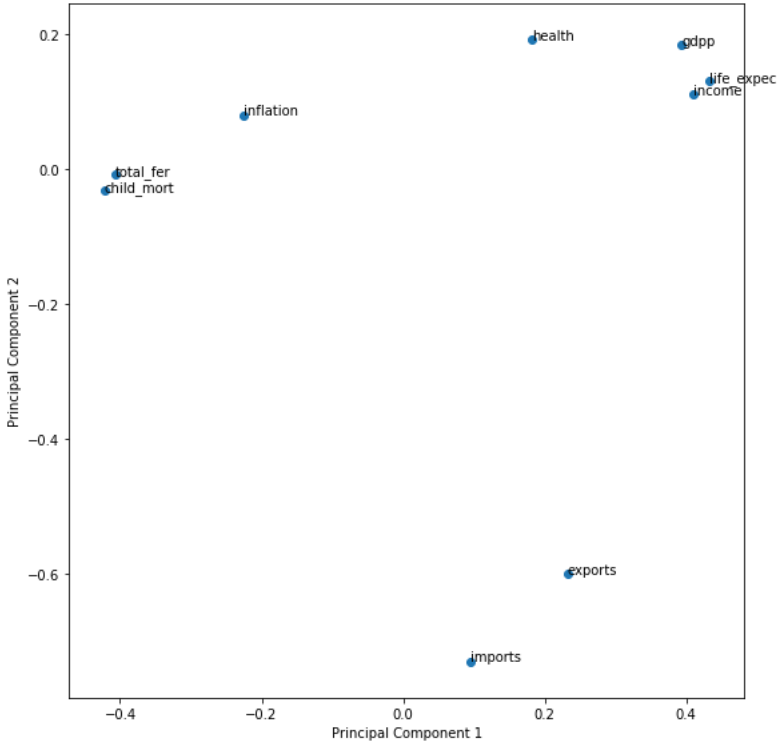
Data Preparation

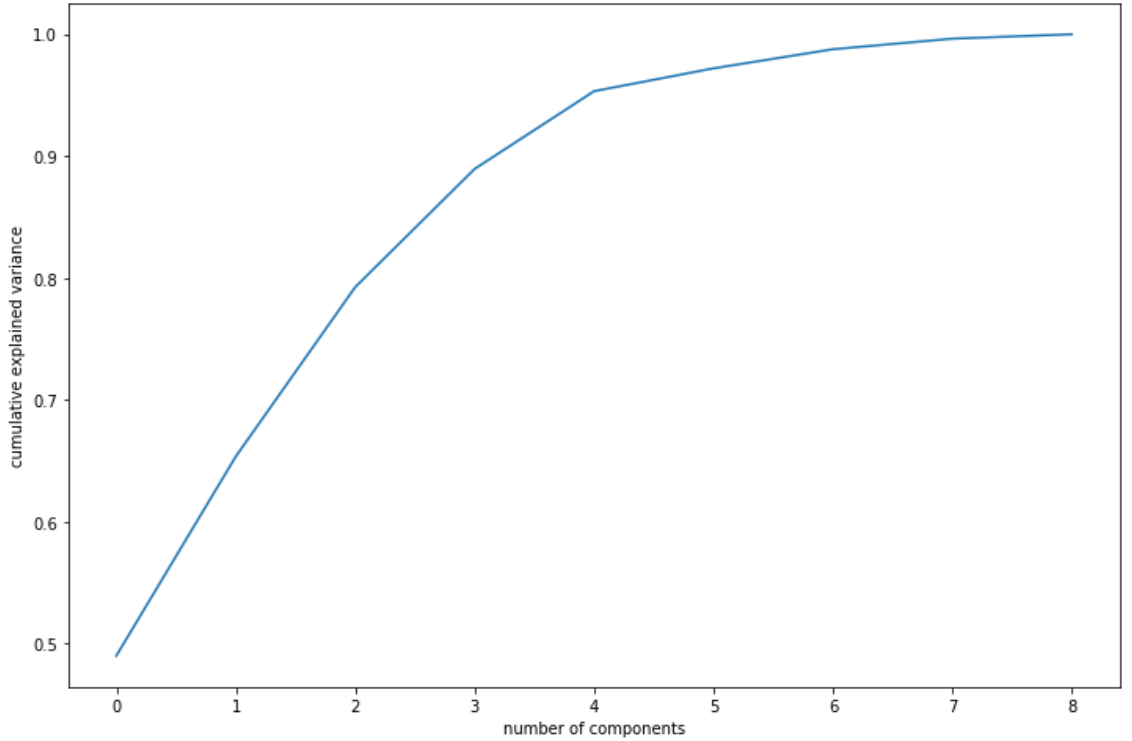
- We performed scaling of variables to ensure that they are on the same scale. This is one of the pre-requisites for PCA.
- Due to the presence of large number of outlier values, we performed Outlier Treatment.



Principal Component Analysis (PCA)

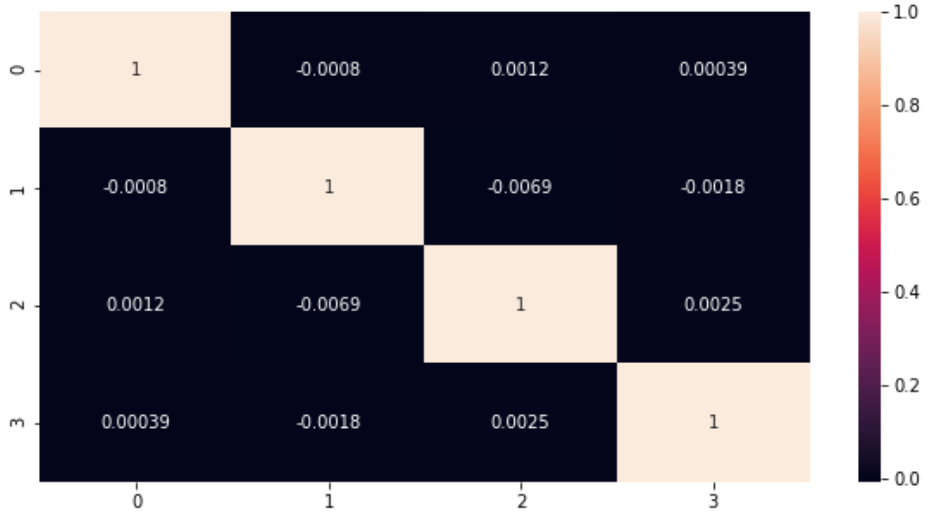
- First, we visualize the original factors with 2 principal components.
- We plot a Scree Plot by plotting the cumulative variance of data against the number of Principal components.
- We observe that 4 Principal components amounts to 90% variance in data and also significant reduction in dimensionality.





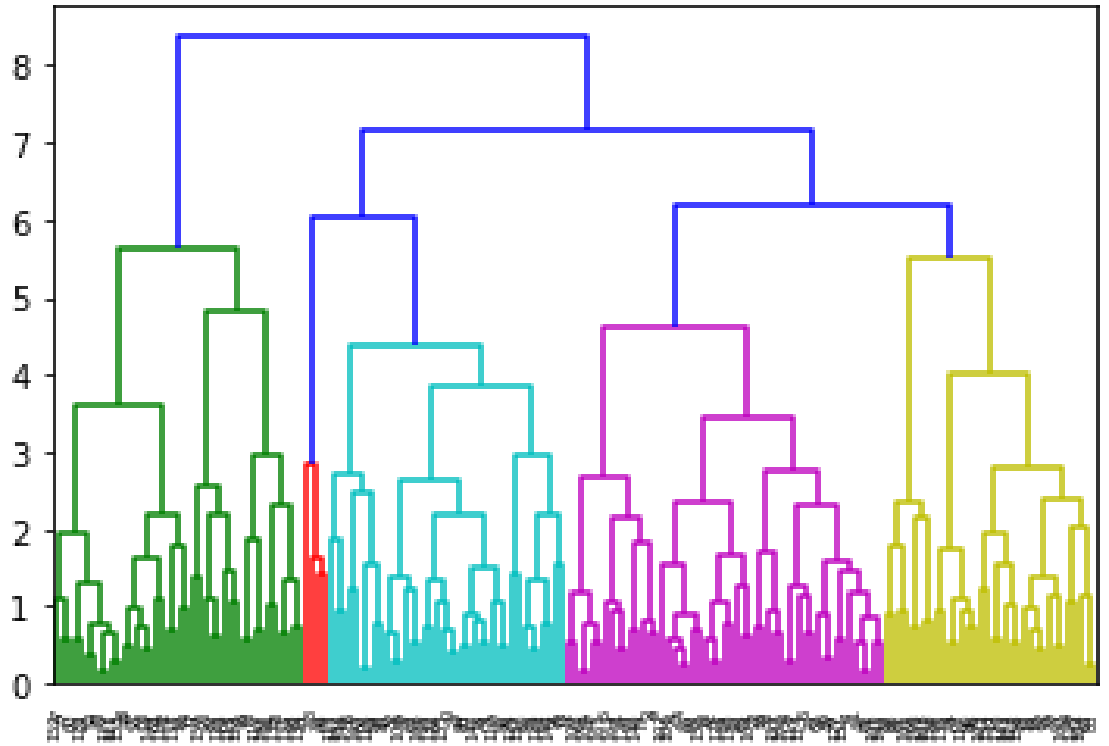
K-Means Clustering & Hopkins Test

- We performed Hopkins Statistics Test to ensure that the given data has some meaningful clusters and is not random. It examines whether data points differ significantly from uniformly distributed data in the multidimensional space and whether it makes sense to create clustering.
- Values of Hopkins test range from 0 to 1.
- If the value is between {0.01, ...,0.3}, the data is regularly spaced.
- If the value is around 0.5, it is random.
- If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.
- FOR OUR ANALYSIS, THE VALUE COMES OUT TO BE ~0.72
- Correlation Matrix-



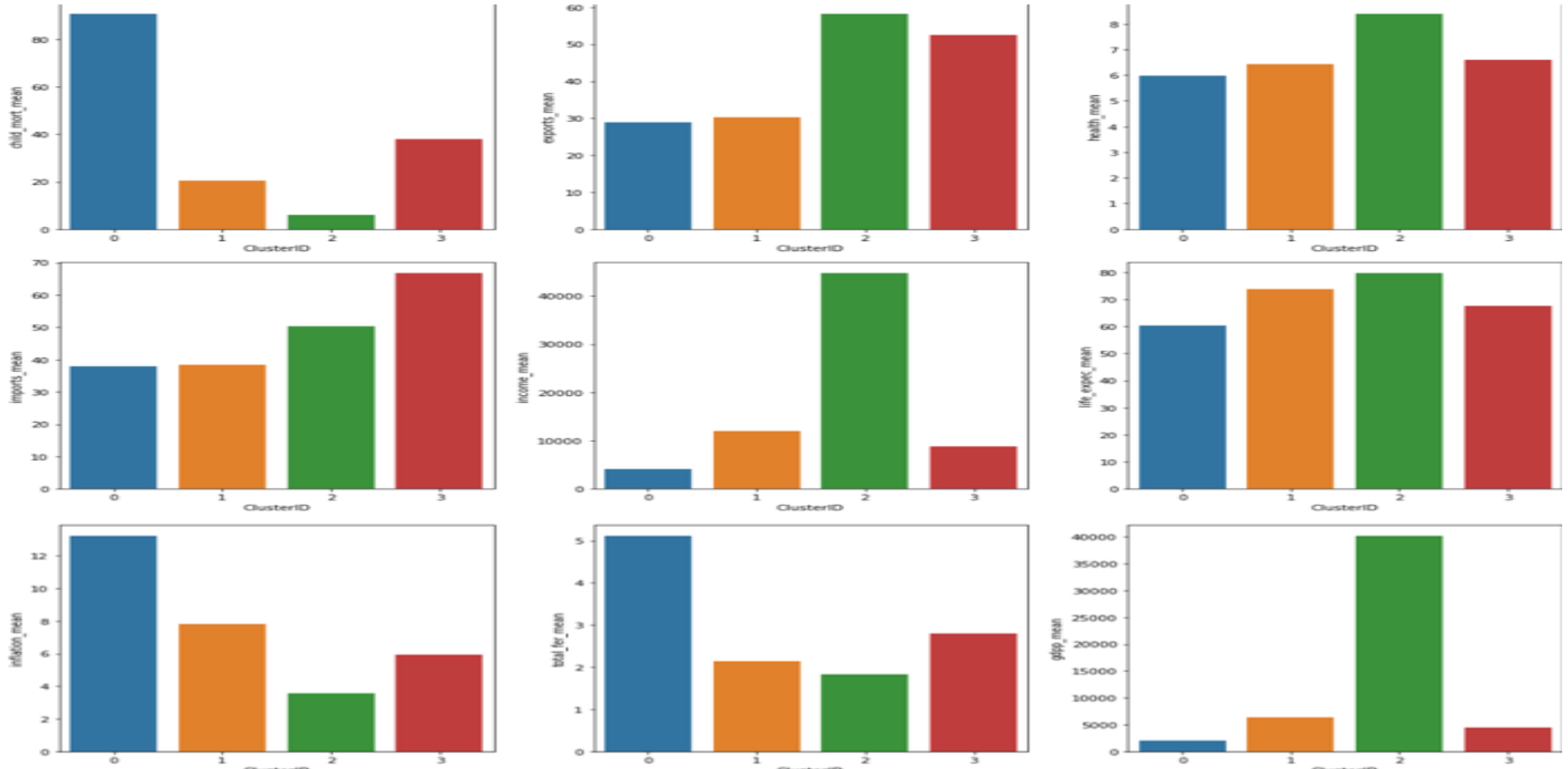
Hierarchical Clustering

- Hierarchical Clustering is performed as it does not have the restriction of deciding the value of K(no of clusters) beforehand.
- Based on the dendrogram plotted below, we cut the line at n=4



Clusters Comparison

- All 4 clusters we identified have been plotted against each of the variables in the original data-set.
- Countries belonging to Cluster ‘0’ have high child mortality rates for children under 5 years of age.
- Cluster ‘0’ has very low gdpp value followed by clusters ‘5’, ‘2’ and ‘1’.



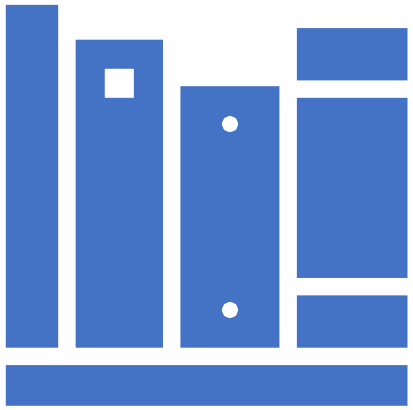
Conclusion & Recommendations

- Based on results from Clustering, we identified that countries in Cluster 0 are in dire need of aid.

- Cluster 0 contains 42 countries. These are listed below:

    - 'Afghanistan', 'Angola', 'Benin', 'Botswana', 'Burkina Faso',
    - 'Burundi', 'Cameroon', 'Central African Republic', 'Chad',
    - 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.', "Cote d'Ivoire",
    - 'Equatorial Guinea', 'Eritrea', 'Gabon', 'Gambia', 'Ghana',
    - 'Guinea', 'Guinea-Bissau', 'Iraq', 'Kenya', 'Lao', 'Madagascar',
    - 'Malawi', 'Mali', 'Mauritania', 'Mozambique', 'Niger', 'Nigeria',
    - 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'South Africa',
    - 'Sudan', 'Tajikistan', 'Tanzania', 'Timor-Leste', 'Uganda',
    - 'Yemen', 'Zambia'

- Based on the data presented data, the company of the NGO needs to decide how to use this money strategically and effectively for the above listed companies.




Data Science IIITB

Yatin Bajaj