upGrad

International Institute of Information Technology Bangalore

Author: Anish Mahapatra
Email: anishmahapatra01@gmail.com

Presentation for the CEO of HELP - International humanitarian NGO

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.

The factors on which the analysis was carried out to cluster the countries include the following features:

1. Child Mortality Rate
2. Exports
3. Health
4. Income
5. Inflation
6. Life Expectancy
7. Total Fertility
8. GDP

The data used was clean and there were no missing values. The data has considerable number of outliers that were found during outlier analysis, but, they were not removed as it was expected that clustering would put them in another cluster.

The data was normalized to get all the features on the same scale, post which a method called PCA was used to capture 95% of the variance. For this, the **scree plot** as shown below was used.

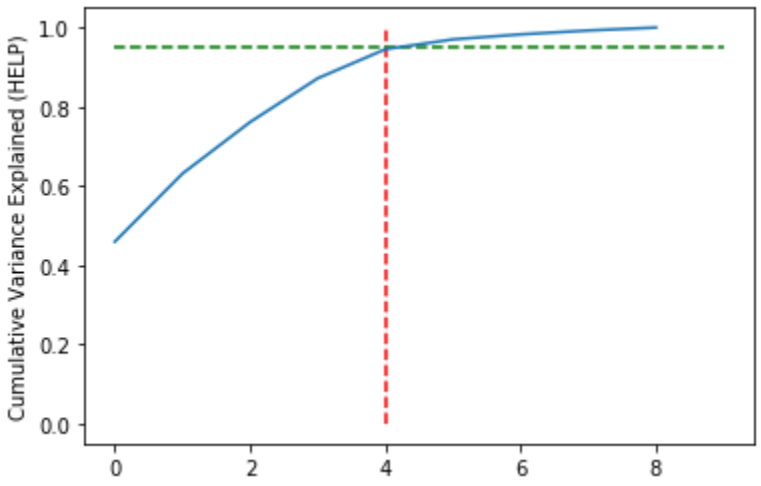


Fig 1: Scree Plot to indicate that 5 PCs capture 95% of variance

This plot indicated that 95% of the variance can be captured with the help of **5** Principal components (0 to 4).

To ensure that the Principal Components selected were uncorrelated (independent), the **heatmap** of the correlation matrix was plotted:

Fig 2: Heatmap of the Principal Components

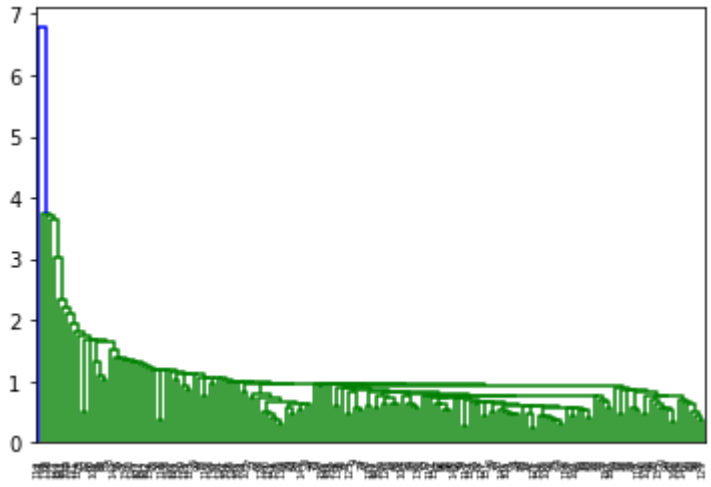
To assess if the new dataset with 5 principal components tends to cluster, **Hopkin's Statistics** was leveraged and a score between 0.75 to 0.9 was achieved. The closer the score to 1, the higher is the tendency to cluster.

For clustering, there are two main methods that were leveraged to find the number of clusters.

1. Hierarchical Clustering
2. K-Means Clustering

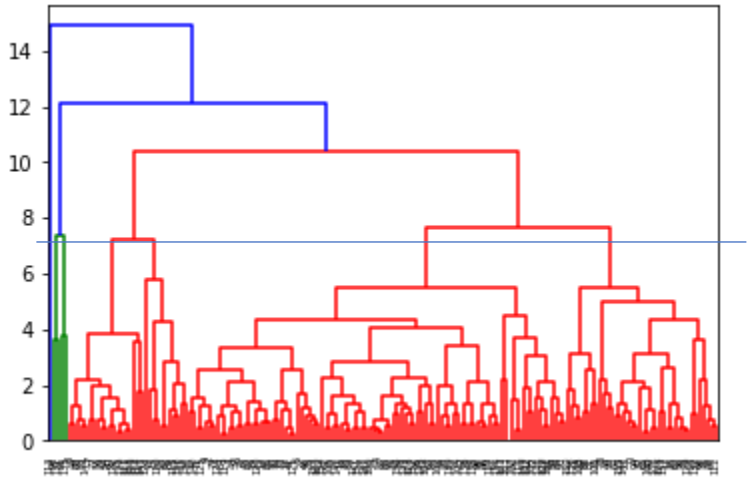
1**. Hierarchical Clustering**

Hierarchical Clustering using *Single Linkage*:



Nothing much can be interpreted from this.

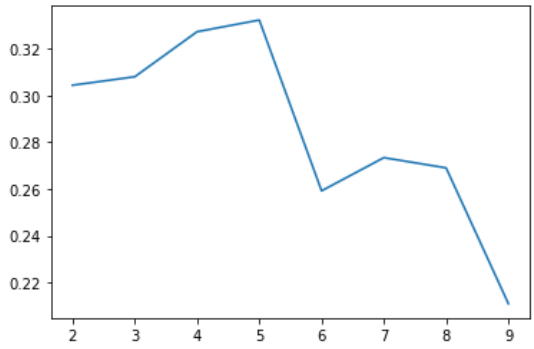
Hierarchical Clustering using *Complete Linkage*



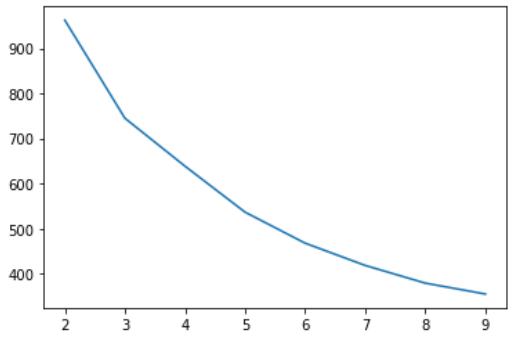
From this, we can interpret that the number of cluster can be between 4 and 6. As the green-colored ones are very close, they can be clustered as a single cluster.

2. K-Means Clustering:

Silhouette Score Plot:

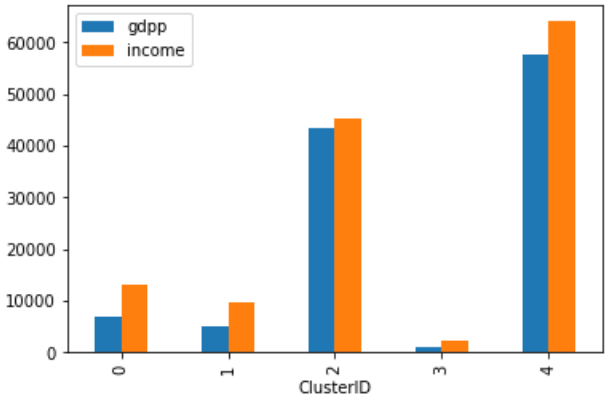


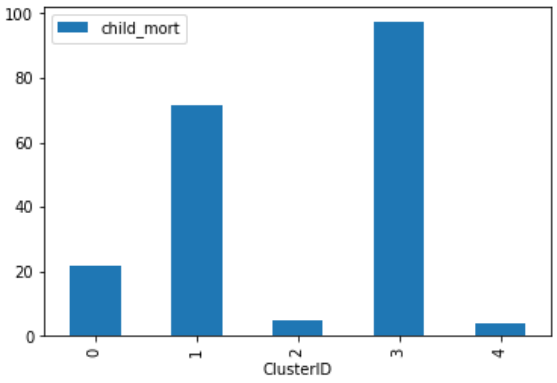
Elbow Curve:



Based on the analysis, we have chosen 5 clusters to be made.

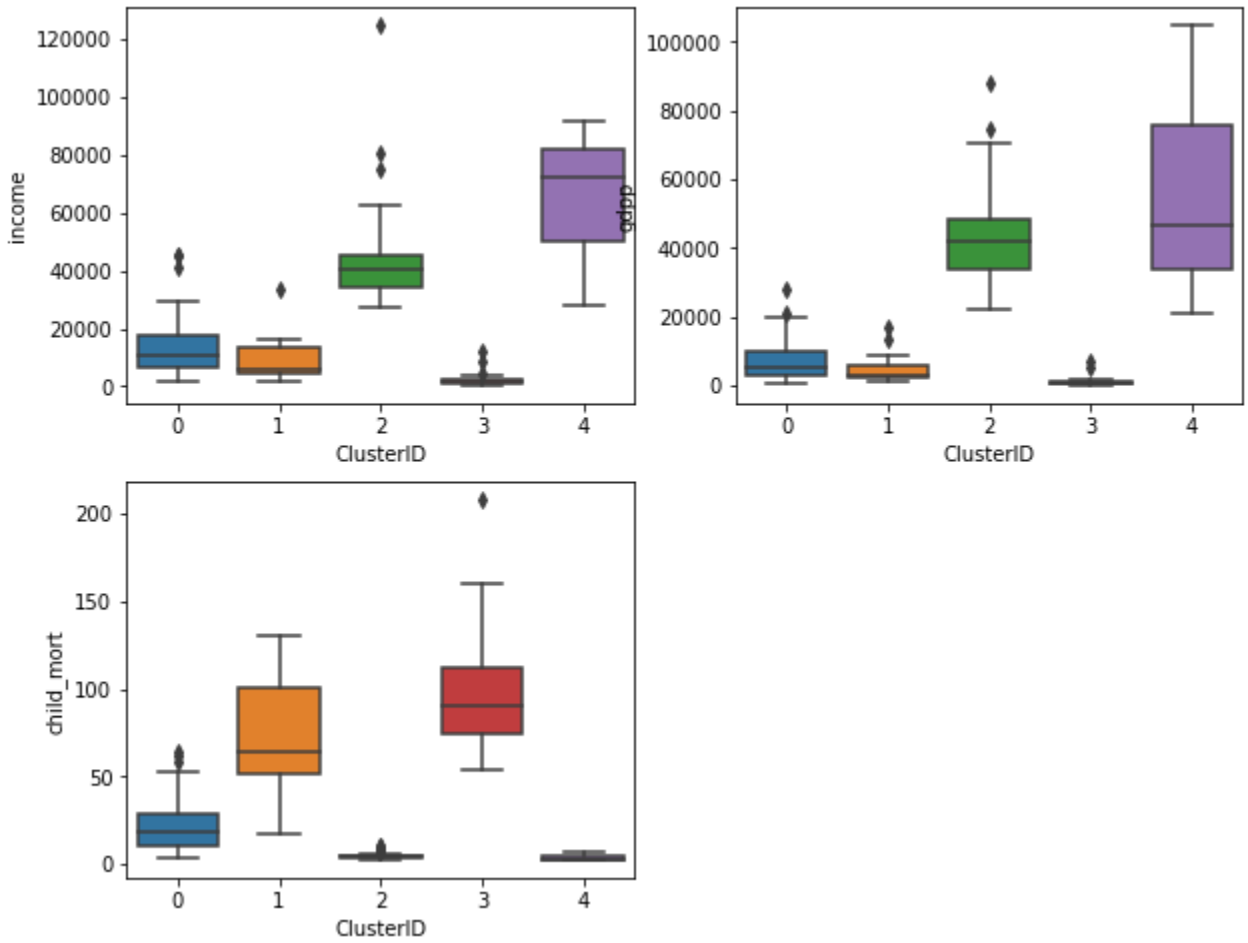
Let us now analyze the GDP, income and mortality rates of the chosen clusters from 0 to 4 (5 clusters)





The GDP of Cluster 3 is the lowest and the mortality rate (average) of Cluster 3 is the highest, hence, indicating that it houses the countries that are in the most dire need for monetary relief.

Let us now observe the distribution of the GDP, income and mortality rates across the clusters.



Cluster 3 has the highest mortality rate, lowest income and GDP.

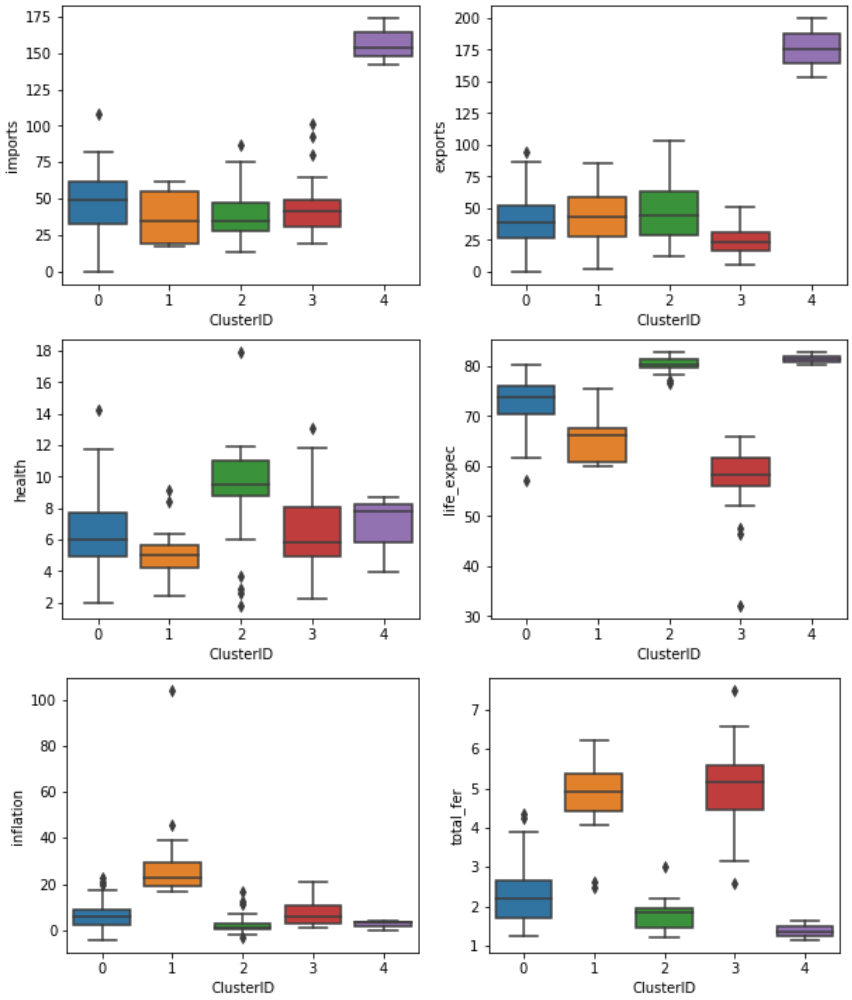
Find below the observations cluster-wise:

1. Cluster 0: The child mortality rate is low, the income and GDP are barely above Cluster 1
2. Cluster 1: The child mortality rate is alarmingly high; The GDP and income are the second-lowest
3. Cluster 2: The child mortality rate is quite low, the income and GDP are the second highest
4. Cluster 3: The child mortality rate is the highest (very bad), the income and GDP are the lowest of all the clusters.
5. Cluster 4: The child mortality rate is the lowest, the income and the GDP are the highest of all the clusters

Now, when we are going to proceed to suggest to the NGO, we need to focus on the countries that have the highest child mortality rate, lowest income and lowest GDP on average. Based on this, the focus should be as mentioned below:

Cluster-3 > Cluster-1 > Cluster-0 > Cluster-2 > Cluster-4

We shall now analyze the rest of the features to understand if there is a deviation in our understanding.



From this, we can make the following observations about Cluster-3:

- Cluster 3 has the lowest income
- It also has the lowest GDP of all the clusters
- The highest mortality rate is present in cluster 3
- It has a low amount of exports as compared to the other cluster
- By far, it has the lowest life expectancies, around 60, with multiple countries having even lower expectancies

<u>Countries to Focus on:</u>

Afghanistan, Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo, Dem. Rep., Cote d'Ivoire, Eritrea, Gambia, Ghana, Guinea, Guinea-Bissau, Haiti, Kenya, Kiribati, Lao, Lesotho, Liberia, Madagascar, Malawi, Mali, Mozambique, Namibia, Niger, Pakistan, Rwanda, Senegal, Sierra Leone, South Africa, Tanzania, Togo, Uganda, Zambia

A lot of the countries that we have picked up belong to South Africa. Hence, we should focus there first.
The top 5 countries that we can focus on to begin with can include the following:

- Uganda
- Zambia
- Mozambique
- Congo
- Afghanistan