Clustering Countries Assignment

-BY SHRUTI PAWAR

Case Study Objective

Problem Statement

- 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.They have raised 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.

- We need to analyze the data and cluster the countires based on -some socio-economic and health factors that determine the overall development of the country.
- -provide the list of top 10 countries to which the financial aid should be given

Approach for Analysis

- Data Reading and Understanding
- Exploratory Data Analysis – perform outlier analysis , understand the correlation between columns , univariate , bivariate analysis done to understand the data and identify the patterns
- Data Preparation for Building The model – Checking the Hopkins stats, scaling of data.
- Building the model using K means – use elbow curve and silhouette score to choose the value of k and perform clustering
- Building the model using Hierarchical clustering –used single and complete linkage method to choose the number of clusters
- Compare and contrast results from the clusters obtained and conclude

K means clustering

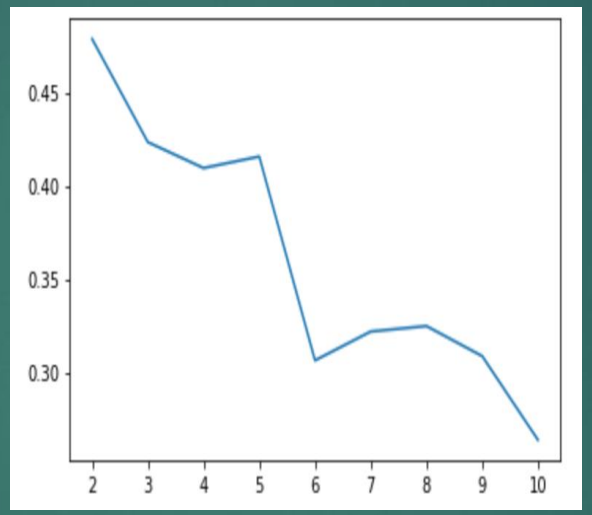
Choosing the value ok K:

Used the silhouette score and elbow curve to determine the value of k.

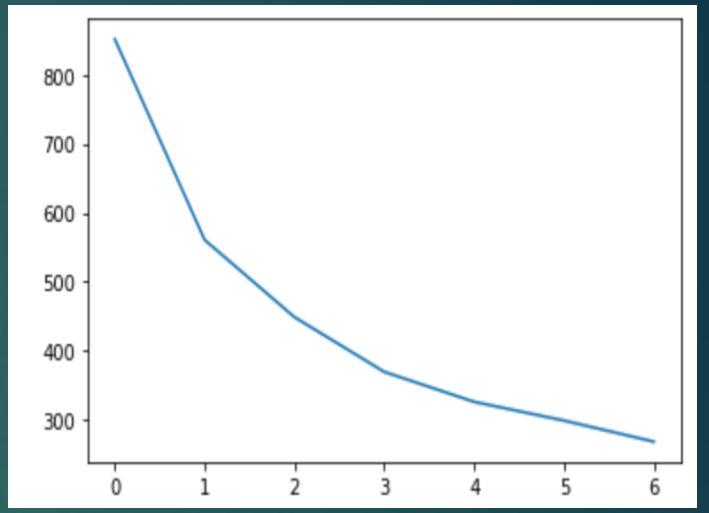
The silhouette score at 3 is the second maximum and also seems to be a decent number for cluster selection.

The second elbow curve also is formed at 3 , so based on above two methods we chose the value of k as 3.

silhouette score



Elbow Curve



K means clustering

Cluster formations:

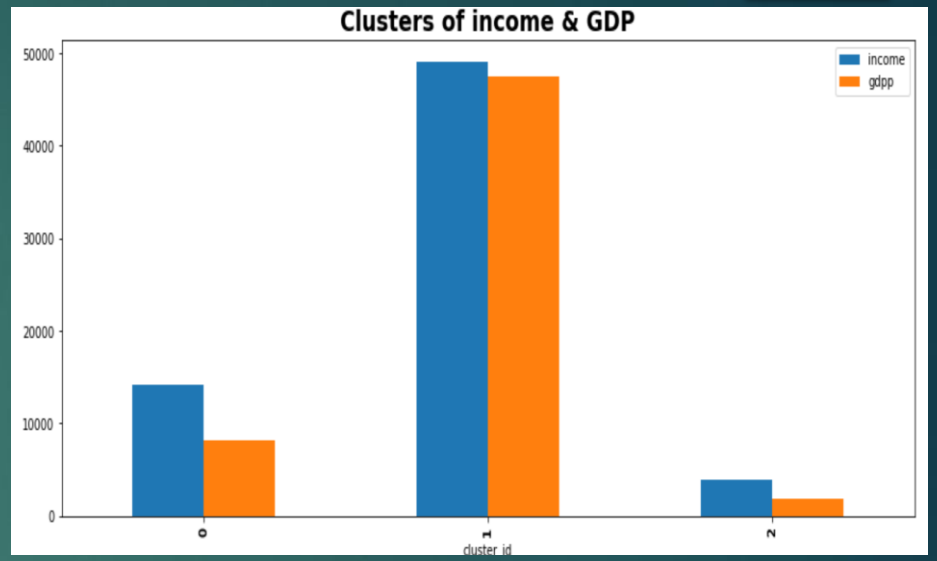
We chose columns gdpp,income and total mortality rate for cluster profiling and we see that

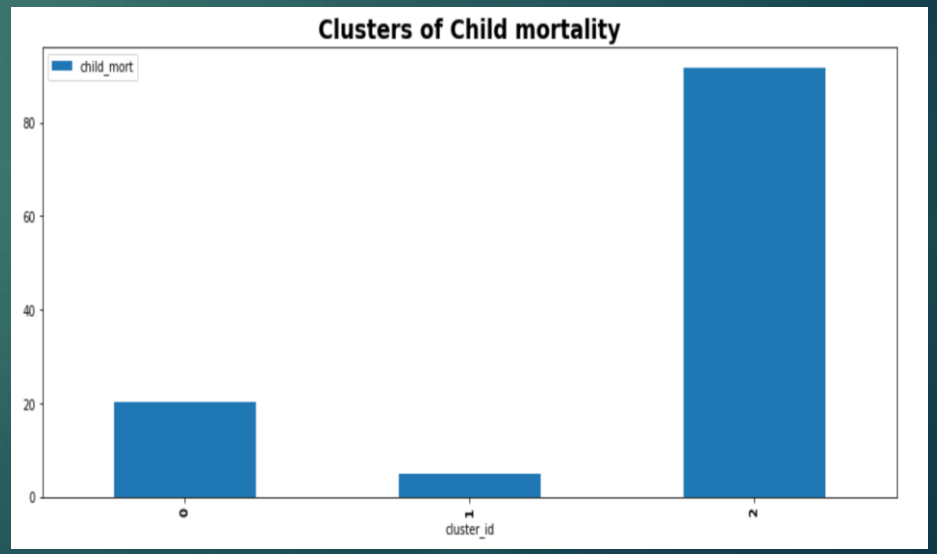
clusters formed can be used categorize countries into:

- Underdeveloped (Cluster with low gdpp and low income having high mortality rate)

- Developing (Cluster with moderate gdpp and moderate income and moderate mortality rate)

- Developed (Cluster with high gdpp and high income and low mortality rate)
- out of 167 countries - 48 countries were clustered as Underdeveloped which has a very low income and gdpp and high rate of child mortality
- - Top 10 countries found from this list are -'Haiti' 'Sierra Leone' 'Chad' 'Central African Republic' 'Mali' 'Nigeria''Niger' 'Angola' 'Congo, Dem. Rep.' 'Burkina Faso'

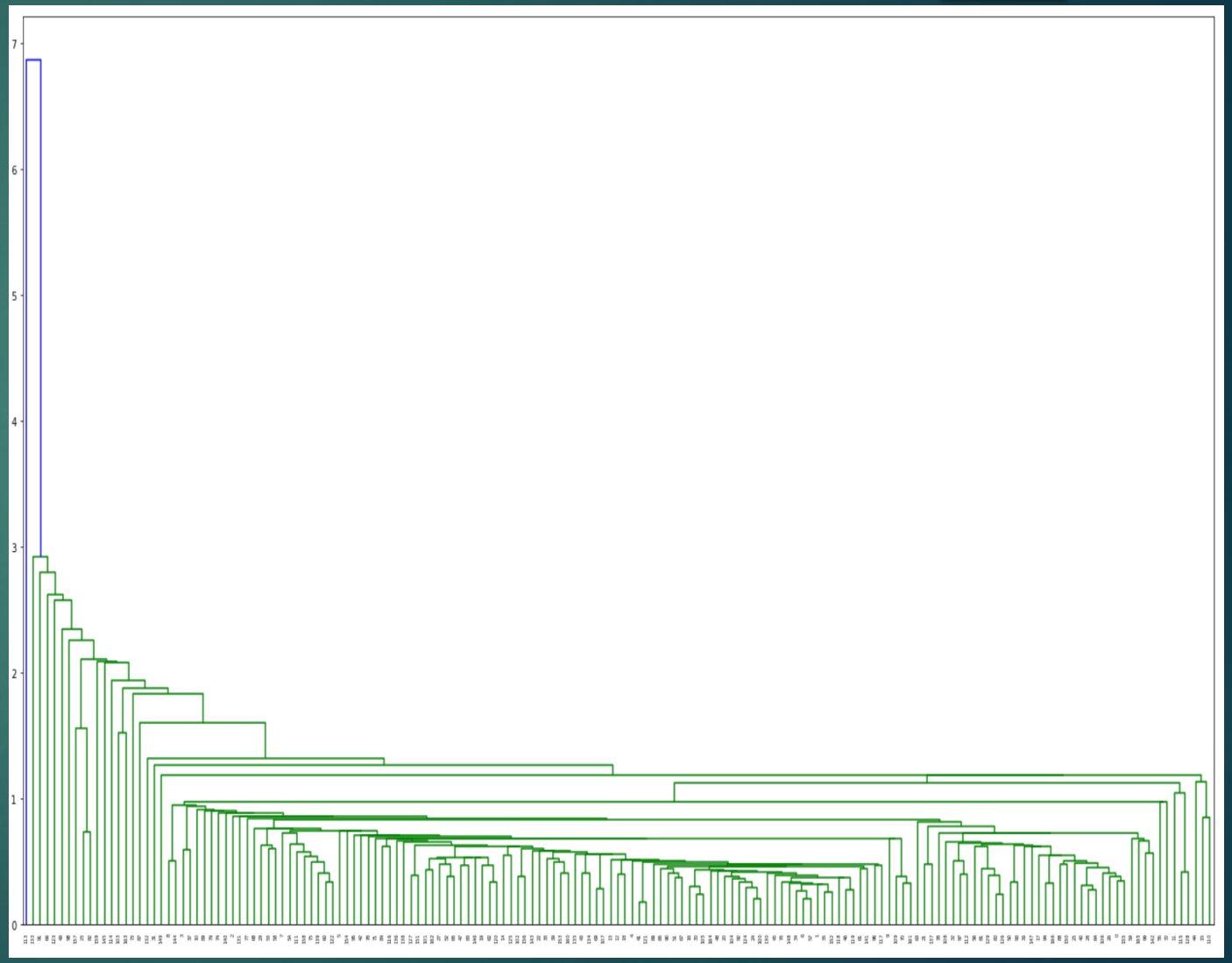




Hierarchical clustering

Single Linkage Method:

Alongside is the dendrogram obtained after using the single linkage method. The dendrogram does not really help in determining the number of clusters and where to cut.



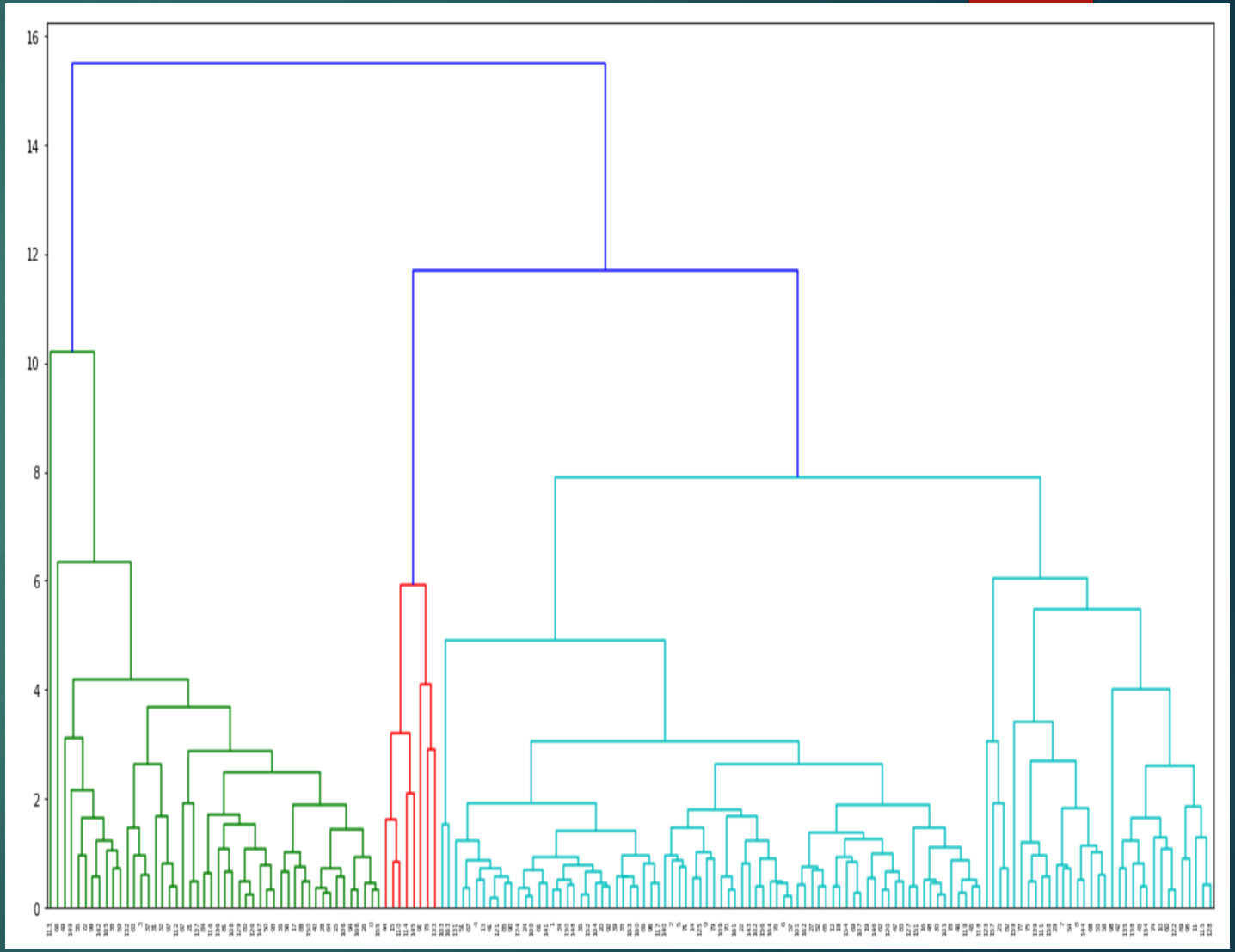
Hierarchical clustering

Complete Linkage Method:

Alongside is the dendrogram obtained after using the complete linkage method.

we see that if cut the tree at around 11 , then we would get 3 distinct clusters for our data, which is a good value for number of clusters as 4/5 would be high and 2 would not suit .

Hence we choose to cut the tree at 11 and make 3 clusters of data.



Hierarchical clustering

Complete Linkage Method:

- Highest Income and GDPP are for countries belonging to cluster 2

- Highest Mortality rate is for countries belonging to cluster 0

- Cluster 0 - high mortality rate , low gdpp,low income

- Cluster 1 - moderate child mortality rate , moderate gdpp ,moderate income

- Cluster 2 - low mortality rate, high gdpp,high income

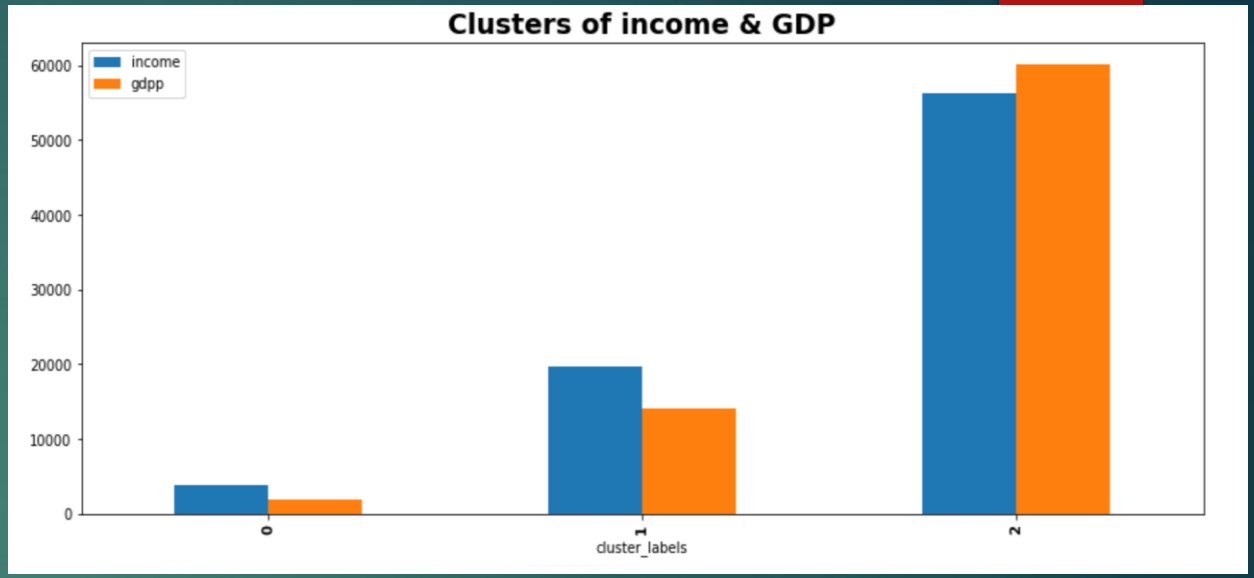
From the clusters formed above we can categorise countries into:

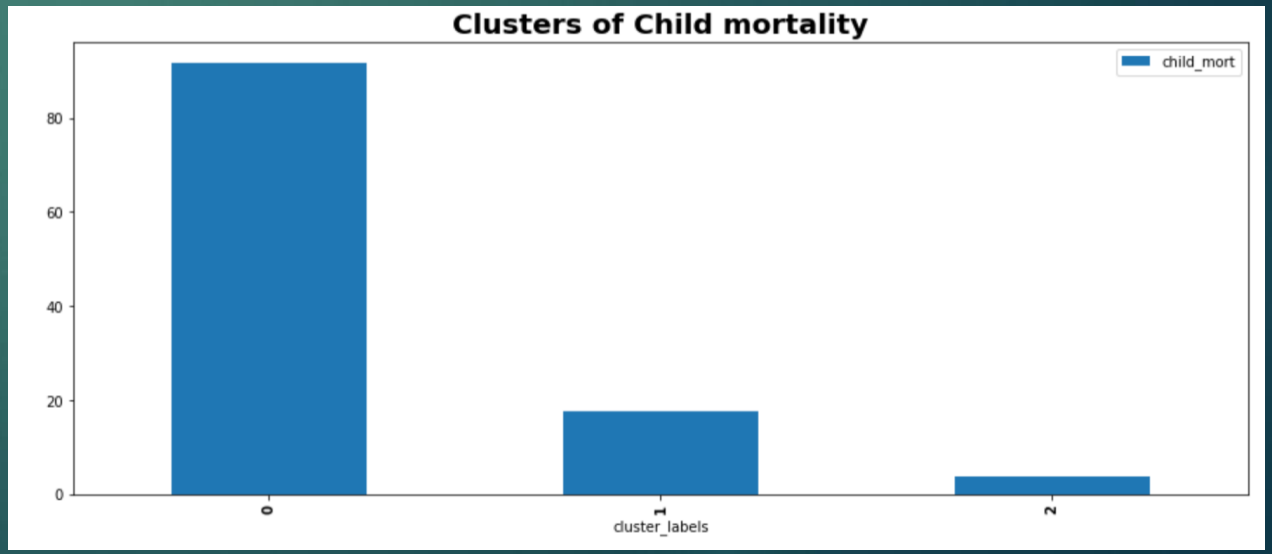
- Underdeveloped (Cluster with low gdpp and low income having high mortality rate)

- Developing (Cluster with moderate gdpp and moderate income and moderate mortality rate)

- Developed (Cluster with high gdpp and high income and low mortality rate)

- Top 10 countries found from this list are -'Haiti' 'Sierra Leone' 'Chad' 'Central African Republic' 'Mali' 'Nigeria' 'Niger' 'Angola' 'Congo, Dem. Rep.' 'Burkina Faso'





Conclusion

We observe that both the clustering techinques have given us the same list of top 10 countries which are in dire need of funds and these can be reported to the CEO for further action

Top 10 countries found from this list are -'Haiti' 'Sierra Leone' 'Chad' 'Central African Republic' 'Mali' 'Nigeria' 'Niger' 'Angola' 'Congo, Dem. Rep.' 'Burkina Faso'