HELP INTERNATIONAL – Clustering Countries

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 programmes, 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.

Problem Statement :

- The task is to cluster the countries based on some socio-economic factors . To check final set of countries, I considered columns such as child_mort, gdpp, income

Steps Followed:

- Read analyze and understand the data
- Exploratory data analysis
    - Perform uni-variate analysis using dist-plot to understand the distribution of data
    - Perform Pair-Plot to understand the pair of columns
- Outlier Analysis :
    - Total no of countries is 167, removing every outliers would cause more data loss even that would possibly direct our analysis in an uncertain way. So `Removing Outliers is not a good option'
- Hopkins Statistics – score is above 90%. So we can proceed with Clustering
- Perform Scaling on data using Standard Scaler – as it keeps the value in the same scale – required for clustering
- **K-Means Algorithm :**
    - Check for Silhouette score and elbow curve
    - Finalize the optimal number of cluster as 3
    - Visualizing the clusters – using scatter plot
    - Cluster Profiling using columns such as child_mort, income, gdpp
- Cluster 1 is the chosen cluster as it has low inome, low gdpp and high child_mort
- Countries chosen are :
    - Haliti
    - Sierra Leone
    - Chad
    - Central African Republic
    - Mali
- **Hierarchical Clustering :**
    - Single -linkage has no significant clustering
    - Perform complete linkage – check for dendrogram

    - Upon looking at dendrogram, choosing n_clusters = 3
    - Visualizing the clusters – using scatter plot
    - Cluster Profiling using columns such as child_mort, income, gdpp
- Cluster 0 is the chosen cluster as it has low inome, low gdpp and high child_mort
- Countries chosen are :
    - Haliti
    - Sierra Leone
    - Chad
    - Central African Republic
    - Mali