1. 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. Objective: The requisite is: To categorise the countries using some socio-economic and health factors that determine the overall development of the country. To suggest the countries which the CEO needs to focus on the most. Method followed: -Data Processing: It was found that there were no null values There were also no duplicate values for country There were a few outliers and they were treated later on during PCA The data was standardized for Principal Component Analysis -Screeplot: 4 components are good enough to get a 95% of variance in the data. So PC is selected to be 4. -Clustering: Both the methods K means and Hierarchical Clustering was used on the 4 PCA components For K means , K= 3 was taken using the elbow dip and silhouette analysis . While doing the Hopkins Statistics a value of 0.77 was attained. If the Hopkins Statistics values are: - 0.3 : Low chase of clustering - around 0.5 : Random - 0.7 - 0.99 : High chance of clustering
2. a) Hierarchical clustering can’t handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).

   In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.

   K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).

   K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

   b) Specify number of clusters *K*.

   Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.

1. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.

- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

c) Elbow method is used to determine e k-means clustering.

d) these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, to prevent this we use standardization.

e) **Single Linkage**: For two clusters R and S, the single linkage returns the minimum distance between two points

Complete Linkage: For two clusters R and S, the single linkage returns the maximum distance between two points

Average Linkage: For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated