**Question 1: Assignment Summary**

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly( what EDA you performed, which type of Clustering produced a better result and so on)

**Note**: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Ans. The problem statement is about, 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. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. We need to choosing the countries that are in the direst need of aid. So, to get countries we first gets idea about values dataset after that prepare data for modelling than with the help of graphs visualise the data set using heatmap and boxplots. By using boxplot we get to know about outliers in the data but as data set small so we go ahead with outliers. After that procced with scaling, It is good to scale the data set as some values are large and some are small so better to scaled them all. Then performed K-Mean clustering and Hierarchical Clustering. We have observed that it is good to go with Hierarchical Clustering as gives precise information. In Hierarchical more countries are come out that needs help as compare to K-Mean.

**Question 2: Clustering**

      a) Compare and contrast K-means Clustering and Hierarchical Clustering.

Ans.

|  |  |
| --- | --- |
| **K-Means Clustering** | **Hierarchical Clustering** |
| We need to have desired number of clusters ahead of time. | We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different  heights |
| It is a collection of data points in one cluster which are similar between them and not similar data points belongs to  another cluster. | Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single  branch. |
| Works very good in large dataset | Works well in small dataset and not good with large dataset |
| The main drawback of k-Means is it doesn’t evaluate  properly outliers. | Outliers are properly explained in hierarchical clustering |
| K-means only used for numerical. | Hierarchical clustering is used when we have variety of data  as it doesn’t require to calculate any distance. |

b) Briefly explain the steps of the K-means clustering algorithm.

Ans.

#### Step 1: Initialisation: Firstly, initialize any random points called as the centroids of the cluster.

#### Step 2: Cluster Assignment: After initialization, all data points are traversed and the distance between all the centroids and the data points are calculated. Now the clusters would be formed depending upon the minimum distance from the centroids.

#### Step 3: Moving Centroid: As the clusters formed in the above step are not optimized so we need to form optimized clusters. For this, we need to move the centroids iteratively to a new location. Take data points of one cluster, compute their average and then move the centroid of that cluster to this new location. Repeat the same step for all other clusters.

#### Step 4: Optimization: The above two steps are done iteratively until the centroids stop moving i.e. they do not change their positions anymore and have become static. Once this is done the k- means algorithm is termed to be converged.

#### Step 5: Convergence: Now this algorithm has converged and distinct clusters are formed and clearly visible. This algorithm can give different results depending on how the clusters were initialized in the first step.

 c) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Ans. ‘K’ value is chosen randomly in K-Means clustering based on statistical aspect. From business aspect, we need to first understand the dataset and based on that we decide number of ‘k’. for example, we have a dataset of variables like ‘pen’, ‘pencil’, ‘books’, ‘notebooks’, ‘mobiles’, ‘charger’, ‘laptop’. Now if we want to have k values based on statistical aspect, we can use silhouette score to determine that but based on business aspect, after viewing the dataset we can easily make cluster = 2, one in electronics category and another non-electronics.

 d) Explain the necessity for scaling/standardisation before performing Clustering.

Ans. Standardization comes into picture when features of input data set have large differences between their ranges, or simply when they are measured in different measurement units (e.g., Pounds, Meters, Miles … etc). These differences in the ranges of initial features causes trouble to many machine learning models. By applying standardisation/scaling will increase the performance of our model

 e) Explain the different linkages used in Hierarchical Clustering.

Ans. Linkage is a technique used in Agglomerative Clustering.

Linkage helps us to merge two data points into one using below linkage technique.

**Single linkage:** The distance between two clusters is calculated by the minimum distance between two points from each cluster.

**Complete linkage:** The distance between two clusters is calculated by the maximum distance between two points from each cluster.

**Average linkage:** The distance between two clusters is the average distance between every point of one cluster to the another every point of other cluster.

**Ward linkage:** The distance between clusters is calculated by the sum of squared differences with all clusters.