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:

CLUSTERING OF COUNTRIES

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. 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 :

- To mention the countries that are in direst need of aid on the basis of socio-economic and health factors determining the overall development of the country.

- Suggesting the countries to CEO at least 5 countries which are in direst need of aid from the analysis work that you perform.

SOLUTION METHODOLOGY :

- First of all, data understanding, inspection and cleaning is done. In that, columns like export, health and import that are in % are changed to absolute values.
- Data Visualization using Scatter plot (for checking outliers), Pairplot and Heatmap (for the variable's correlations).
- In Data Preparation, a copy of original dataset is created and capping of outliers is been done. Columns like child_mort and total_fer is treated in lower range.
- Hopkins statistics is used for checking the cluster tendency of the dataset. Moreover, Hopkins score is above 0.85 which seems to be a good one for further clustering.
- After that, data scaling and K-Means clustering using Elbow curve method and silhoutte score analysis.
- Though optimal number for K means seems to be 3, further clustering process is performed by setting max_iter=50 and random_state=50 and cluster_id is created.
- Further visualization and profiling of clusters is shown in Scatter plot and boxplot. It is found that countries belonging to cluster_id =0 has low income & gdpp with high child mortality rate.
- Complete Linkage in Hierarchical clustering gave promising result and further visualization and profiling is been done that landed up with a result that cluster_label=0 has high child mortality rate with low income and gdpp. For that, 2 clusters been taken after cutting the dendogram.

- Finally, fetching the dataset that performed K-means clustering showed the list of the countries which are in direst need of aid.

Question 2: Clustering

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

Ans:

- In K-Means Clustering, it needs desired number of clusters. While in Hierarchical Clustering, number of clusters can be decided after the completion of plotting dendrogram , that to by cutting it at different heights.
- K-means requires a prior knowledge on number of centroid whereas Hierarchical cluster doesn't require any parameters. Instead, cut_tree() is used.
- K-Means clustering works on a very large dataset while Hierarchical clustering works on a small dataset.
- K-Means is only used for numerical while Hierarchical is used when we have variety of data.
- In comparison to Hierarchical clustering, K- Means is faster.
- K-Means doesn't evaluate outliers properly in comparison to Hierarchical clustering.

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

Ans:

- K points are selected randomly as an initial centroid.
- All the data points that are closer to the centroid will create cluster centre according to Euclidean distant function.
- Once all the points are assigned to each of k clusters, the cluster centres or centroid has to be updated.
- Above second and third steps should be repeated until cluster centres reach convergence.

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: As per the statistical aspect, K value is chosen randomly in K-Clustering. And it can be performed using Elbow Curve and Silhoutte Score. If it's about business aspect, comprehension of dataset is important. In that sense, we can decide how many clusters to be taken.

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

Ans: Though clustering techniques use Euclidean Distance, it will be wise to scale the variables. Ex:- heights in meters and weights in KGs before calculating the distance. It may create a big difference while calculating for K-Means and Hierarchical. This is because the cluster will tend to move variable having greater values or variances. In fact, most clustering algorithms are even highly sensitive to scaling. Rescaling the data can completely ruin the results.

e) Explain the different linkages used in Hierarchical Clustering.

Ans: In Agglomerative clustering, linkage plays a vital role in merging two clusters into one.

Different linkages that are used in Hierarchical clustering are as follows:

Single-Linkage : Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

Complete-Linkage: Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.

Average-Linkage : Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

Centroid-Linkage : Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more

similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn’t arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.