<u>Question 1:</u> 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

<u>Solution 1:</u>

<u>Assignment Summary:</u>

Help the HELP an international humanitarian NGO's CEO to make the right decision on how to use 10 Mn fund strategically and effectively on which countries. Need to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Need to suggest the countries which the CEO needs to concentrate on the most.

<u>Analysis Approach:</u>

1. First imported the data in python notebook for analysis.
2. Perform Normal Check,
    - Checked if there are some missing data or null values.
    - Checked for Duplicate values.
3. We did the outlier analysis and found that there are outliers in some columns; removing the outliers we lose significant amount of data related to different countries; So we Treated them after PCA and removed Nearly 7% Of data which are not necessary for our analysis.
4. Then we did data scaling before PCA on the dataset and found the principal components as 3 principal components, as these 3 are explaining 93% of the variance in the data we have taken n-components as 3.
    - Checked on a heat map for collinearity between columns and found zero correlation.
5. We now can do the K-means clustering on the data.
    - Computed Silhouette score
    - Then we performed SSD/elbow curve and got optimal no of clusters as.
    - In K means we took K=4 clusters and performed the K means.
    - Elbow curve for optimal number of K = 4
6. Analysis of K-Means and Hierarchical clustering
    - Computed K means
    - Computed two types of Hierarchical clustering
        - Single linkage
        - complete linkage
        - Formed dendrogram
    - Visualized clusters
7. Provided final list of countries

Which type of clustering produced the better result

● We checked both the methods for clustering i. e Hierarchical Clustering and k-means clustering and both produced the similar result. And we can see from our analysis that countries pointed out by hierarchical method and k-means method were commonly present in both, so we choose top-20 common countries.

● I would say hierarchical clustering is usually preferable, as it is both more flexible and has fewer hidden assumptions about the distribution of the underlying data.

● But for us we can say that both hierarchical & k-means method produced similar and better results, But for recommendation I choose results from Hierarchical method.

<u>Question 2:</u> Clustering

<u>Solution 2:</u>

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

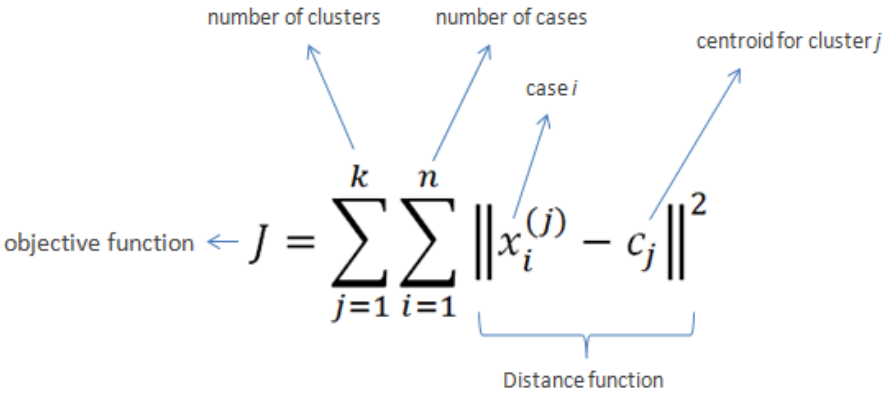
K-means clustering is the process of dividing the data objects into subsets where in each cluster has similarity among them and dissimilarity between two clusters. While using k-means, you need to have a sense of desired number of clusters needed. K-means gives different results such as data is not well separated into sphere like clusters and may be the K value we picked may not be suited for shape of the data.

In contrast, Hierarchical Clustering calculates the distance of each pair of data points. Hierarchical Clustering joins the nearby data points into the cluster and successfully adds the nearby points into groups, finally forms Dendogram. We can decide the clusters numbers based on the dendogram.

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

Steps to be followed:

1. Specifying the number of clusters (for example, k=1,2,3,..etc).
2. Initialising the random centroids at first then calculating the Euclidean distance between the random centroid and data points.
3. Calculate the mean and reshuffle the centroid using mean value and iterating the same until the centroids position is not changing.



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

The K-value is chosen using Elbow method and silhouette analysis.

Elbow method-we will choose K value based on the sum of squared distance between the data points and their centroids. We will plot the curve and choose the best K from the curve plotted.

Silhouette analysis is used to determine the degree of separation between the clusters. We will compute Average between the data points with in the same clusters (a) and Average distance between the clusters (b) and calculating the coefficient of the cluster points. The one with coefficient '0' means the dataset has close neighbouring clusters.

Coefficient is given by = (b-a)/max (a, b)

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

The Scaling is necessary because the dataset or sample given may have different variables with different range. For example, a variable may have data ranging from 100 to 1000, the other may have in lacs, that is why the dataset or sample is scaled using standardisation before clustering. Standardisation scales the data in such a way that the mean becomes zero and standard deviation becomes 1. The following is the formulae for standardisation.

$$z = \frac{x_i - \mu}{\sigma}$$

e) Explain the different linkages used in Hierarchical clustering.

The hierarchical Clustering has three types of linkages. They are:

1. Single linkage.
2. Complete linkage.
3. Average linkage

In the single linkage, we will merge the two clusters with the smallest minimum pairwise distance.

In the complete linkage, we will merge the two cluster having smallest maximum pairwise distance.

Average linkage is the compromise between the sensitivity of complete linkage to clusters and single linkage tendency of forming long chains which does not have impact on outliers.

Question 3: Principle Component Analysis

a) Give three applications of using PCA.

1. PCA is one of the dimensionality reduction technique.

2. PCA aims to transform the correlated variables into a smaller set of linearly uncorrelated variables. Hence this will reduce the multicollinearity.

3. Data visualization of PCA is used to show the different cluster variations.

b) Briefly discuss the 2 important building blocks of PCA – Basis Transformation and variance as information.

Basis Transformation: we can transform the original dataset or sample so that the Eigen vectors becomes the new basis vectors and find the new coordinates of the data by using the old coordinates with respect to new basis. The basic formula is given by:

New Coordinates= M * old coordinates.

$$PX = \begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} = \begin{bmatrix} p_1 \cdot x_1 & \cdots & p_1 \cdot x_n \\ \vdots & \ddots & \vdots \\ p_m \cdot x_1 & \cdots & p_m \cdot x_n \end{bmatrix}$$

For example, for the matrix A we will decompose these matrix into Eigen vectors and Eigen values and this new Eigen vector matrix is known as A' and A' is multiplied by matrix A in order to get the New coordinates. This process iterates till the Eigen vector does change. This is known as Basis of Transformation in the PCA.

Variance as information: In PCA, Variance is the cumulative variance or multivariate variance or overall variability. Variance as information means as the in the process of iteration in each and every step the new coordinates forms and thus plotted on the space as Eigen vectors changes the variance also changes. The one where the new

coordinate's does change has a lot variance and this will reduce the dimensionality. In covariance matrix, the variances are in diagonal. Sum of diagonal values is overall variability. This covariance matrix is the symmetric matrix.

c) State at least three shortcomings of using Principle Component Analysis.

1. Independent variables are not interpreted (not readable).

After implementing PCA on dataset, the independent variables are converted into the principle components of linear combination where the variables are uncorrelated to each other. But the data cannot be in readable or understandable way.

2. Data Standardisation must be used before PCA.

Scaling is important in PCA because the dataset may have a lot of variables with different mathematical formats such as percentages, kilogram. If this the case the PCA will not be able to find the optimal principle components. That is why standardisation is important in order to find the optimum principle components.

3. Information loss.

If Principle components are not chosen with care the information available may be lost.