Question 1: Assignment Summary

Sol. the main objective of clustering of countries is to provide the support from the NGO. Following are the steps taken while doing clustering:

- Data preparation
- EDA and visualizing the data
- Principal component analysis
- Checking the Hopkins statistics to check the overall variance of the data
- K- means or Hierarchical clustering whichever gives the best results
- Final report by giving the suggestions or recommendations

In here k means clustering seemed to give better clustering.

Question 2: Clustering

A. Compare and contrast K-Means and hierarchical clustering

Sol.

- K-means is used for huge data where hierarchical clustering is used when there is small data because the run time is very high, so we prefer k-means clustering when the data is huge
- In k-means we can use elbow curve to find optimum number of clusters and in hierarchical clustering we first take every data point as a cluster itself and join the nearest data point and this keeps on going until the no. of clusters reaches to 2. After seeing the plot we can cut the tree of clusters as per the business requirement

B. Briefly explain the steps of the K-means clustering algorithm.

Sol.

- We need to check silhouette scores or elbow curve. Observing at the curve or values we can choose the number of clusters.
- Initialize the no. of cluster centers and assign observations to the closest cluster center
- Check the distances from the data points to cluster centers and assign to the closest one and keep on iterating this until all the points are no longer changing clusters.

C. How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Sol.

- we can choose k value in two ways either silhouette scores or by elbow curves
- In statistical aspect we can see where the max variance is and using that we can do the clustering.
- But in business aspect we need to show the data and ask them how many categories they want to do on their data

D. Explain the necessity for scaling/standardization before performing Clustering.

Sol. The features of the data that we were given will not be of same units or of same scale, without scaling them or bringing down to same scale, the clustering will be difficult and cannot say that we are making good clusters. For example if we are having weights in kgs and heights in centimeters, in this the heights will be of big numbers may be float kind but weights will be of different scale. So normalizing the variables will bring all the variables to the same scale.

E. Explain the different linkages used in Hierarchical Clustering.

Sol. There are three types of linkages in hierarchical clustering. They are :

1. Simple linkage: This links the two clusters with shortest distance between the observations of each cluster.
2. Average linkage: This links the two clusters with mean distance from a data point from one cluster to a data point of the other cluster.
3. Complete linkage: this links with the maximum distance of 2 data points from 2 different clusters.

Question 3: Principal Component Analysis

A. **Give at least three applications of using PCA**.

Sol. 1. Doing PCA gives better visualization than the actual data frame.

2. Improves the speed of running the algorithm.

3. Reduces highly correlated variables.

4. Over-fitting will be minimized.

B. Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

Sol. 1. The points taken in standard basis will be transformed into eigenvector basis. By this transformation of dimensions variables with low variance will be discarded, which leads to dimensionality reduction

2. In PCA we check the individual variance to the overall variance of components which leads to know the importance of variables.

C. State at least three shortcomings of using Principal Component Analysis.

Sol.1. PCA assumes that principal components are linear combination of original features. If this condition doesn't satisfy, we will not get good results.

2. High variance variables are treated as Principal components, if all variables have approximately of same variance we might lose data.

3. It assumes that the principal components are orthogonal (statistically independent).