Question 1)

We imported and prepared the data for EDA. We performed univariate analysis to find out which countries are affected by what factors. We then looked for variables having high correlations. We performed outlier analysis using Percentile Capping i.e. Winsorization. We then performed a Hopkins Statistics Test to find out if the data is suitable for clustering or not. After that, we rescaled the features using Standardization method to prevent any outliers from making out analysis biased. In the next step, we used Kmeans clustering algorithm for model building. Using elbow curve and Silhouette method, we found out that 3 clusters are ideal for out dataset. We assigned the labels and did cluster profiling to find out cluster number 2 has the countries which are the least developed and requires most focus. We found out 5 countries in cluster Id 2 . Now, we moved onto heirarical clustering and performed single linkage and complete linkage. We cut the dendogram at no. of clusters = 3, assigned the labels and did cluster profiling. Cluster 0 had the least developed nations. We got 5 countries who needed the most focus from heirarichal clustering. K means does not need prior knowledge about the number of clusters. Hence, we will proceed with the clusters formed by Kmeans and based on that information provided by the final clusters , we will deduce the final list of countries which are in need of aid and needs the most found out that the 5 countries from both K means and Heirarichal clustering were the same got out final list of 5 countries in the final analysis.

Question 2)

a) Ans:

- 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. $O(n^2)$.
- 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.
- In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm many times may differ. In Hierarchical Clustering, results are reproducible in Hierarchical clustering.

b) Ans:

Step 1: Choose the number of clusters *k*

Step 2: Select k random points from the data as centroids

Step 3: Assign all the points to the closest cluster centroid

Step 4: Recompute the centroids of newly formed clusters

Step 5: Repeat steps 3 and 4 until centroids of newly formed clusters do not change.

c) Ans:

The value of 'k' in k means clustering is decided by two methods:

1) The elbow curve method

The elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

The intuition behind the Elbow curve is that the explained variation changes rapidly until the number of groups you have in the data and then it slows down leading to an elbow formation in the graph as shown below. The Elbow point is the number of clusters you should use for your K-Means algorithm.

2) Silhouette curve method

This is a better measure to decide the number of clusters to be formulated from the data. It is calculated for each instance and the formula goes like this:

Silhouette Coefficient = (x-y)/ max(x,y)

where, **y** is the mean intra cluster distance: mean distance to the other instances in the same cluster. **x** depicts mean nearest cluster distance i.e. mean distance to the instances of the next closest cluster.

The coefficient varies between -1 and 1. A value close to 1 implies that the instance is close to its cluster is a part of the right cluster. Whereas, a value close to -1 means that the value is assigned to the wrong cluster.

Business Consideration: We also should get an idea of number of value of k by understanding the business and its factors**.** Your goal shouldn't be to just create

clusters from your data. It should be to create meaningful, accurate clusters that you can use to generate insights about your business.

d) Ans:
In statistics, standardization (sometimes called data normalization or feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

When you are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

e) Ans:

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.