Question 1: Assignment Summary

I first cleaned the data by checking null values in the data, if any, then made sure that the attributes are in correct data format and also checked for duplicates in the data and prepared it for analysis. Then performed univariate and bivariate analysis on the data by plotting pair plots and heat map of correlation matrix for all numeric variables.

From the heatmap, I deduced that there are some attributes that have high correlation among them. There is an option to drop those attributes. But this way, I may lose some information that tends to be important in our analysis.

So, I performed PCA (Principal Component Analysis) to handle this issue of multicollinearity in the data without losing or compromising any important information.

I then scaled data using StandardScaler() method before building the model and then performed outlier analysis on the scaled data. There, certainly, are few outliers in the data, that were observed by plotting boxplot. Since the aim is to find the countries that are in direst need of aid, I cannot lose any kind of data and would like to analyse as it is. *So, considering the business need, outliers are not removed here.*

I then built a model using K-means clustering algorithm by considering 3, 4, 5 clusters consecutively and found that the result are best when 3 clusters are formed. This conclusion is derived by following Elbow method and finding Silhouette score.

I then built a model using Hierarchical clustering by plotting a dendrogram and cutting it to obtain 3 clusters first and then 4 clusters. Observed that results are best when 3 clusters are considered.

One major difference I observed between results of hierarchical clustering and K-means clustering is that the "income" and "gdpp" are well clustered/ explained using k-Means clustering algorithm than hierarchical clustering. So I considered K-Means clustering algorithm in this analysis and final analysis to identify the countries that are in the direst need of aid.

Conclusion:

From the business understanding, I identified few factors like child mortality, GDP Per capita, income, etc as vital attributes to determine the socio-economic standard of each country. I succeeded in identifying 12 countries

that are in direst need of aid. This list is subject to changes as it is based on few factors like number of clusters chosen, number of attributes considered based on business understanding, the clustering algorithm used to build the model.

Question 2: Clustering

A) Compare and contrast K-means clustering and Hierarchical Clustering.

K-means Clustering	Hierarchical Clustering
Assigns records to each cluster to find mutually exclusive clusters based on Euclidian distance	These methods can be divisive or agglomerative.
Needed an advance knowledge of number of clusters one wants to divide the data into. i.e., the value of K.	Can stop at any number of clusters and can find appropriate number of clusters by looking at the dendrogram.
Cluster centre can be mean or median value of the data	Agglomerative method begins with "n" no. of clusters and sequentially combine similar clusters till only one cluster is obtained.
These methods do not involve intensive calculations and hence can be best in case of larger datasets.	Divisive method is exactly opposite to agglomerative approach. Start with a single cluster and the target is to arrange the clusters in natural hierarchy.
Running the algorithm multiple times results in varying values as the initial value of K is randomly chosen.	Results are reproducible in this method of clustering.
K-means algorithm results in non-overlapping clusters such that each data point is in exactly one cluster.	Result is a set of nested clusters that are arranged in the form of a tree.

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

K-means clustering algorithm in the process of dividing "n" data points into "K" clusters.

Steps of K-means clustering algorithm:

1. Choose a random number of K (i.e., random number of clusters) as the initial cluster centres.

2. Calculate the “Euclidian” distance between the datapoints and assign each data point to its nearest cluster centre.
3. Now compute the new cluster centre for each cluster. This will be the mean of all the cluster members.
4. Considering the new cluster centres, reassign the data points to the clusters.
5. Repeat steps 3, 4 till no further changes in clusters and cluster centres are observed.
6. Result is the optimal clusters.

C) How is the value of “k” chosen in K-means clustering? Explain both statistical as well as the business aspects of it.

The value of “K” in K-means clustering i.e., number of clusters depends on the method we use to find similarities between datapoints and the parameters used for clustering. Elbow method and silhouette method are two widely used methods to determine the initial value of “k”.

Elbow method:

Compute the K-means clustering algorithm for different values of K. Ex, varying K from 1 to 10.

i. Calculate the total within-cluster sum of square (wss) for each K value.
ii. For each varying value of K, plot the curve of wss.
iii. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Silhouette method:

Compute the K-means clustering algorithm for different values of K. Ex, varying K from 1 to 10.

i. Calculate average silhouette of observations (avg sil) for each K value.
ii. For each varying value of K, plot the curve of avg sill.
iii. The location of maximum is considered as the appropriate number of clusters.

➔ K value is chosen randomly based on statistical aspect. From business aspect, we first need to understand the dataset clearly and based on our understanding, we decide on number of clusters i.e., the initial K value.
➔ If we want to know the k value based on statistical aspect, we use any of the above mentioned methods (elbow method or silhouette method) to determine the initial value of K i.e., initial number of clusters.

D) Explain the necessity for scaling/standardization before performing clustering.

K-means algorithm works by calculating the Euclidian distance between the datapoints. To ensure that the attributes with larger range of values do not outweigh those with smaller range, scaling the attribute to the same normal scale is necessary. Also, the attributes need not necessarily be in the same units. So, standardizing the data helps in making the attributes uniform and unit-free.

E) Explain the different linkages used in Hierarchical clustering.

Common types of linkages in Hierarchical clustering:

1. **Single linkage –** The distance between two clusters is defined as the shortest distance between datapoints in the two clusters.
2. **Complete linkage –** The distance between two clusters is defined as the maximum distance between any tow datapoints in the clusters.
3. **Average linkage –** The distance between two clusters is defined as the average distance between every datapoint of the cluster to every other datapoint of the other cluster.