Question 1: Assignment Summary

Ans:

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.
Categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

EDA:

As part of EDA we have converted values of export , health and import columns from %gdpp to actual values . Checked distribution of values for each column to check feasibility of clustering data set.

After EDA we have performed outlier treatment using capping technique.

We have build the models using K-Means algorithm and Hierarchical clustering . Out of these two algorithms K-Means with 3 clusters provides clear insights as it divided data points based gdpp ,income and child-mortality clearly

Question 2: Clustering

Using K-Means algorithm we have grouped datapoint into 3 cluster but 2 clusters into 2 hierarchical clustering. Both models provides exact same results for the top five countries can be considered for funding.

Steps of the K-Means Clustering:

1. **Choose initial number of clusters using silhouette score and elbow curve**
2. **Build model on scaled data using K-Means Algorithm with initial number of clusters**
3. **Visualize data for clusters**
4. **Do the cluster profiling and identify final results**

Before building model we need to scale/standardize the data as all the values in the data set to be on the same scale to get accurate results. If values are in different scales then model may not give correct results.

Types of Linkage in Hierarchical Clustering

Single Linkage: Returns the minimum distance between two data points belongs to different clusters

Complete Linkage: Returns the maximum distance between two data points belongs to different clusters

Average Linkage : Returns value of Arithmetic mean between clusters