Question 1:

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. Hence as a analyst, We need help the organisation in choosing the countries that are in dire need.

Solution:

We have data of all countries which contains information related to child mortality, gdpp,income,health, total fertility, imports,exports,inflation.

Since few columns are highly correlated in the dataframe , To avoid multi-collinearity and reduce variables , We perform PCA on the variables.

If we do outlier analysis on the data frame and remove any country, There might be a case that we might miss the country that is in direst need monetary fund. Hence we do not perform outlier analysis on this dataframe.

We then do silhouette analysis and find the 5 clusters are needed to explain 95% of variance. Hence we choose 5 clusters and perform run Kmeans algorithm.

We then map the cluster ids ot the dataframe frame and then check gdpp,income and child_motality against the cluster ids. We find that cluster 1 and 4 contains countries that are in direst need,

We then filter countries from cluster 1 and 4 and sort it against child_mortality ,income,gdpp and then we take top 15 countries that in are in direst need of monetary support.

Clustering :

1. Difference between kmeans and hiearachial algorithm

   Hierachical Clustering cant handle big data well but k means clustering can.This is because the time complexity of K means is linear i.e of O(n) while that of hierarchial clustering is quadratic.

   In K means clustering, since we start with random choice of clusters, the results produced by runnin the algorithm multiple times might differ.While results are reproduxible in hierarchial clustering.

K means isfound 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 dendogram.

Kmeans algorithm:

Using the silhouette analysis , We find the number of clusters that explain maximum variance.

kmeans = KMeans(n_clusters=5, max_iter=50)

kmeans.fit(df)

We then get that cluster labels using

kmeans.labels_

Using this we perform cluster analysis.

Need of standardisation:

We need to perform standardisation to get all the variables in same units.

There are 3 types of hierarchical clustering.

Single Linkage

In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.

Complete Linkage

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.

Average Linkage

In average linkage hierarchical clustering, the distance between two clusters is defined as the average

distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters “r” and “s” to the left is equal to the average length each arrow between connecting the points of one cluster to the other.

PCA

3 applications of pca

- For data visualisation and EDA
- For creating uncorrelated features that can be input to a prediction model: With a smaller number of uncorrelated features, the modelling process is faster and more stable as well.
- Finding latent themes in the data: If you have a data set containing the ratings given to different movies by Netflix users, PCA would be able to find latent themes like genre and, consequently, the ratings that users give to a particular genre.
- Noise reduction

Shortcomings of pca

- PCA is limited to linearity, though we can use **non-linear techniques such as t-SNE** as well (you can read more about t-SNE in the optional reading material below).
- PCA needs the components to be perpendicular, though in some cases, that may not be the best solution. The alternative technique is to use **Independent Components Analysis.**
- PCA assumes that columns with low variance are not useful, which might not be true in prediction setups (especially classification problem with a high class imbalance).