Q1

**Business Objective**: 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. A nine feature socio-ecomic factor are provided for each of the 167 countries. Using these features in the data we need to identify countries which are in the direst need of aid.

**Methodology**:

Clean the data and perform EDA. Check for duplicates, null and outliers etc

We will use derived metrics where suitable eg % Health converted to Heath per person

Check for collinearity – indeed features are collinear as seen by correlation matrix

Data is standardized as features are different units and scale to ensure mean is zero and standard deviation is equal to one.

Using Principal Component Analysis (PCA) we will attempt to reduce the dimension while retaining the information/ variance. This is done by creating Scree plot. Here at 5 Principal component variance explained is greater than 95% and good practice to have at least half the features as PC, hence 5 is optimum.

From PC converted data we will attempt to cluster using unsupervised learning techniques like Kmeans and Hierarchical clustering cluster countries based on their socio-economic factors

We will treat the outliers i.e. countries with very high or very low development characteristics to enable clustering algorithm to work. This is done by removing values of PC above 95% quartile and below 5% quartile.

KMeans clustering is done by using Silhoutte measure and elbow curve to identify the optimum K. From the analysis we find that K=3 as optimum for silhouette score analysis AND k=2 OR 3 seems given by elbow. Hence choosing K=3, i.e. 3 cluster KMeans algorithm,

Once under-developed country cluster is identified we will use is centroid/ mean/ characteristics to find the most under developing countries from the original database (includes outliers) which require aid the most

A comparison of Kmeans and hierarchical clustering will be done and if variations seen will try to explain them.

Q2

Three shortcomings of PCA are

1. The PCs are linear combination of the features. Hence this can be restrictive
2. PCs must be orthogonal and uncorrelated which is also restrictive
3. PCA considers low variance variable are not useful and removes them which can lead to loss in information. This is particularly harmful in imbalanced variable for eg bank fraud detection as fraud events are rare and hence data is imbalance. PCA would ignore these events.

Q3

Hierarchical clustering is more intuitive and can help visualize the different segmentation results. We can look at a big cluster and see when it breaks what happens. Generally for business we cannot afford to have very large cluster to take action i.e. K>15, so dendogram can be cut below 15 cluster and visualizaed

Once the we have a good cluster grouping we figure out their centroid (mean/ median). Then we could use the K-means and initialize these centroids as initialized points and algorithm will converge fast.

Advantage hierarchical clustering: There is no need to decide K. Elbow and silhouette analysis.

Disadvantage hierarchical clustering: Need higher computation power to run on larger dataset. Hence K-means is tool of choice in big dataset.