Question 1. **Problem Statement**: The problem at hand is to categorize the given countries using socio-economic and health factors that determine the overall development of the countries. This categorization is then to be used to suggest the countries which are in direst need of aid. These suggestions will be used for strategic and efficient spending of the funds that have been raised by an 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.

Solution Methodology: The main approach taken to solve this problem was to first identify the most significant principal components. Outlier treatment wasn't done since the dataset is very small and there's a risk of losing out information. The data was then scaled to bring the data on a uniform scale. There are variables that are related to each other, like income is closely related to GDP per capita etc. This means that the variation in the data can be explained by fewer variables than given in the dataset. On performing PCA, it was found that 4 principal components were explaining ~94% of variation. The first 4 PCs were selected for further analysis. Hopkins statistic was found to be close to 0.9 telling that the data has a high clustering tendency. Silhouette Analysis and Sum of Squared Distances were plotted and it was found that the appropriate number of clusters is 2 to 5. On selecting K=4 for K-Means it was found that one cluster had only 2 countries. Hence K=3 was selected for K-Means. After performing K-Means with K=3, a clear difference among the clusters was observed. The difference observed was in terms of the mean values of different features like life expectancy, child mortality etc. One of the clusters was identified as grouping the countries which were under-developed, hence in need of aid. As a next step, hierarchical clustering was done. Numbers of clusters was chosen 5 for hierarchical clustering. The results obtained were quite like K-Means clustering.

These clusters were then analyzed based on values of variables like Child Mortality rate, Life Expectancy, total fertility and health spending etc.

Question 2: State at least three shortcomings of using Principal Component Analysis.

Answer: Following are the shortcomings of PCA:

- PCA assumes that columns with low variance are not useful. This might not be true in prediction steps. Columns with low variance might also be useful for predicting the dependent variable. This problem is seen in classification problems with class imbalance.
- Since PCA relies on orthogonal transformations, it needs the components to be perpendicular, though in some cases, that may not be the best solution.
- Since PCA relies on linear assumptions, it limited to linearity. If the data is not linearly correlated (e.g. spiral), PCA is not enough.

Question 3: Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer: Following are some comparisons between K-Means and Hierarchical clustering:

- K-Means begins by creating K centroids and then iterates between assign and update steps till no data points can be reassigned any further. Assign step assigns the samples to a centroid, and the update step recalculates the centroid as the mean of all the samples that are assigned to it. Hierarchical clustering on the other hand build clusters incrementally. It produces a dendogram. The algorithm begins by assigning each sample to its own cluster and then merging the most

similar clusters. The merging continues till all the samples are in the same cluster. The layers then can be navigated to see which number of cluster makes the most sense.

- The data points in K-Means is clustered such that each object is in exactly one cluster containing other objects like it. All clusters have objects which are dissimilar to those in other clusters. Whereas, in Hierarchical clustering, clusters have a tree-like structure or a parent-child relationship.
- K-Means is usually preferred if the number of clusters to be created is known beforehand, while Hierarchical clustering can be done if the number of clusters to be created is not known.