Question1:

Problem:
The problem at hand was to identify the countries which require assistance because of their poor socio economic factors.
Methodology:
The basic idea for solving the problem is to group by the similar countries into clusters and find the cluster which is performing poorly across the given factors. This cluster contains the countries which require assistance.

I order to do the above, PCA is performed in the data set to reduce dimensionality and multi collinearity. To decide on number of components, scree curve was plotted and from that it was inferred that 93% of the variance can be explained by 2 PCs. Hence 2 principal components are chosen for the analysis. Hopkins coefficient was calculated to see if the data can be clustered and the value obtained was greater than 0.7 indicating that clustering is possible.

For clustering both KMeans and Hierarchical Clustering are performed. In k-Means the number of cluster was decided using silhouette score and from the plot it was inferred that 3 cluster has the highest silhouette score. The clusters formed also made sense as there was distinct separation between very rich, moderate and backward countries. For Hierarchical clustering both 3 clusters and 5 clusters were compared and it was observed that the cluster on interest remained unchanged in both the cases, hence 3 clusters was chosen for the remaining analysis.

After clustering is done, different parameters were measured among the clusters and the cluster with low gdp, high mortality,low life expectancy was chosen as the cluster of interest. 10 countries are selected from the chosen cluster on basis of gdpp of the country.

Question 2 :

Limitations of Principal Component Analysis:

1. PCA is based on linear relations, hence if the data is not linearly correlated then PCA doesn't do a very good job
2. All the principal components found through PCA are required to be normal to each other, Hence these components found may not always be the directions of highest variance
3. Principal components are less interpretable as they are linear combination of the original features. Hence interpretability of the results in not easy
4. While used for dimensionality reduction some information is lost because only few PCs are selected

Question 3:

KMeans vs Hierarchical clustering

1. For Kmeans clustering we need to have an idea of number of clusters before hand where are Hierarchical clustering requires not presumptions
2. The clusters formed though KMeans are highly dependent on the initial choice of centroids. Therefore, the results can change drastically even with small change of initial centroids. Hierarchical clustering is similar to a tree approach starting either from one cluster to many clusters or vice-versa. Hence it is independent of initial conditions.
3. Hierarchical clustering is very computationally expensive and requires large memory, where as KMeans is less expensive if we know the number of clusters before hand.
4. KMeans can form bad/unintuitive clusters if the data is not well separated as circles or spheres, whereas Hierarchical clustering has no such issues.