ASSIGNMENT PART 2

Question 1:

The problem statement was to find the list of countries which are in dire need of aid so that Help International can help these countries with its funding programs. Below are the steps to achieve the solution:

a. Removing outliers from the dataset gdpp and income above 95 percentiles. Doesn't have any significance as they are well developed countries.
b. Standardization of data to perform PCA.
c. I have selected 4 Principal components as I could explain above 80% of the data by these 4 PCs.
d. Hopkins statistics was 0.83 showing high tendency of the dataset to cluster.
e. Based on the Silhouette score and Sum of squares we concluded that number of clusters can be 3 for a better model.
f. Performed KMeans on the PCA dataset and derived the Clusters 0, 1 and 2. Cluster 0 seems to be the Bad cluster having all the feature values very low.
g. **Gdpp , total_fertility, inflation and income** are the important features which classify these clusters.
h. Based on these important features we got the below set of coutries from KMeans algorithm which are in dire need of aid: **Albania, Bosnia and Herzegovina, China, Macedonia, FYR, Tunisia.**
i. Performed Hierarchical clustering on the PCA dataset and found another set of countries: **China, Dominican Republic, Morocco, Peru, Philippines, and Romania.**
j. The features which are important in order to determine the above list of countries are **inflation, imports, health and child mortality.**

Question 2:

The three shortcomings of Principal Component Analysis:

1. PCA is a linear combination of original values
2. PCA requires Principal components to be independent and perpendicular to eachother. This othonormal tendency sometimes not feasible when the data needs to be collinear. PCA becomes a bad choice in this case.
3. Low variance features are not important in PCA, which is indeed not useful at times.

Question 3:

K Means clustering is non linear and doesn't iterate for maximum iterations so has better performance. In the other hand Hierarchical clustering is linear process. In this we need to process the similar size of data every iteration. It slows down the performance.

KMeans for its non linearity is more preferred for larger datasets whereas hierarchical clustering is preferred when the dataset is small.

KMeans clustering we predefine the number of clusters (k), but in Hierarchical clustering instead of predefining the number of clusters, we first visualize the similarities and dissimilarities between the different data points and then decide the appropriate number of clusters on the basis of these similarities and dissimilarities ie Linkages.