Question 1

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly(why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

Answer:

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. Categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Solution Methodology:

Seps involved:

1. Data Cleaning
2. EDA
3. Perform PCA on the dataset and obtain the new dataset with the Principal Components. Choose the appropriate number of components k.

Scaling of variable by standard scalar method Applying PCA on the data find out the pricipal component screen to plot to find the number of pca, from the graph the best values for components are 3 where 90% of variance is being explained. Remove outliers in the data if there in the data. use scatter plot to observe the distribution. calculate the hoopikins stat value how good the pca's are

4. Apply K-means clustering to form clusters on the data set.

Use silhouette score to see how the data is being distribution. Also use elbow curve to find the k values to identify the number of cluster that can be formed. Apply the k.means

clustering on the data set obtained after pca's being applied to find the best clusters. After the clusters being formed observe the clusters by plotting the distribution. create the cluster means w.r.t to the various variables mentioned in the question and plot and see how they are related

5. Apply Hierarchal clustering on the data set obtained to form clusters on the output df of step 1:

There are two methods ,use the single linkage procedure it does not perform well so use the complete linkage method. Define the cluster ids and find the clusters observe whether clusters are good or not. Create the cluster means w.r.t variables and see how they are related

6. Analyze the clusters and identify the ones which are in dire need of aid.

Identify the cluster which are in need.

Question 2

State at least three shortcomings of using Principal Component Analysis.

Answer:

PCA has 3 major assumptions/simplifications embedded –

1. The PCs have to be linear combinations of the original columns. Why limit ourselves to linearity when we can go non-linear. t-SNE is an alternative, although computationally very expensive

2. PCA requires the PCs to be uncorrelated/orthogonal/perpendicular . Sometimes the data demands that correlated components to represent the data . ICA (Independent Component Analysis) overcomes this drawback, but is several times slower than PCA

3. PCA assumes low variance components are not very useful - In supervised learning situations, this can lead to loss of valuable information. This is especially true for highly imbalanced classes/variables.

Question 3

Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer:

Difference between K Means and Hierarchical clustering

- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found 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 dendrogram.