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 NGO provides the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. We would do some analysis on the Socio-economic and health factors of these countries and find the ones that are in the direst need of aid.

Analysis

We have 167 Countries having 9 Factors each

During Principal Component Analysis (reduce factors to the principal components that explains most of the data without losing information), using a scree plot, we are able to conclude that 5 principal components would explain 95% of the (variance in) data.

The Heat map shows that there is negligible or no correlation between the principal components.

Determine Clusters of countries that are doing similar in terms of these Factors.

KMeans Clustering

We would determine if the data set has tendencies to from good clusters using Hopkins analysis and find the no. of clusters using Silhouette Analysis after which it was determined that we could form 3 – 5 clusters . When we form 5 clusters, 2 of these were removed because they had very low data-points, we had 3 good clusters.

Hierarchical Clustering

Using an agglomerative clustering, a Dendrogram – complete method, Euclidean distance, we cut 5 clusters, two of which have very low data-points therefore tried to fit them into 3 clusters, however it ended up grouping the bigger clusters into 1 keeping the small clusters. Therefore we concluded with 5 Clusters.

Bar graphs using the mean value of each variable in each cluster helped understand the countries in need of aid.

The below analysis was done on features within the choosen cluster.

1. Inflation v/s Income v/s GDP
2. Life Expectancy v/s Child Mortality v/s Health Spending v/s Fertility
3. Imports v/s Exports

Question 2

State at least three shortcomings of using Principal Component Analysis.

Answer

The three shortcomings of Principal Component analysis are -

1. PCA is limited linear techniques only. For non-linear techniques, PCA could be used for computational efficiency only.

2. PCA needs all the components to be perpendicular to each other. In the case when the components are not perpendicular, this might not be a good solution.

3. PCA assumes that the columns with low variance are not useful which might not be true for some business case or a classification problem.

Question 3

Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer

K-means

1. One has to determine the no. of Clusters before one can fit data into the model without knowing the data.

2. It is not computation heavy, the distance between each data point and the cluster centriods needs to be computed.

3. The k - centroids must be randomly initialized, the distance between every data-point and the centroid should be made Iteratively until the k-centroids do not move, where we now obtain our k clusters.

Hierarchical

1. One can see from the dendrogram, how many clusters would one is likely to get depending on where the analyst wants to cut this tree.

2. It is computation heavy, the distance between each data point and every other data point needs to be computed.

3. Each cluster is initially considered as a single cluster, we then group these clusters that have a minimum distance iteratively until we get only 1 cluster.