Assignment: Part II

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, where we as an analyst needs to find the top 5 countries who needs help.

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)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Answer

Solution Methodology:

Step 1(Scaling and Normalizing):

Primarily I have imported the 'Country' Dataset after importing I have performed basic checks like checking the shape, amount of nulls present in the dataset, While describing the data set and plotting the boxplot for each variable I have noticed the presence of outliers In the dataset .Used 1.5xIQR rule for removing the outliers from all the columns, After the removal of outliers I have scaled the data using standard scaled(X-mean/SD), Then finally checked the distribution of data.

Step 2(Selection of component and performing PCA):

In this step I have used PCA library to finding the principal components which can explains our original data in reduced dimension, here I have performed PCA 2 time one is for identifying the number of component and second time to select only the required component, For identifying the n_component I have used scree plot where I have noticed around 95% of the variance can be explained by 4 PC, s then I have applied incremental PCA and found the 4 vectors. Also plotted the heatmap between PC, s to check if the vectors are orthogonal and there should be no relation.

Step 3(Performing Clustering):

Before performing the clustering I have performed Hopkins statistics to understand if this dataset is eligible for clustering , then I have calculated the silhouette score and plotted elbow curve using inertia to find the number of centroids required for the cluster .In my case 3 is doing fine for implementing KMeans and hierarchical clustering .

In case of hierarchical clustering the dendogram will not be good if we plot it using single linkage (min distance) whereas complete linkage (max distance) is doing just fine

Step 4(Analysis on actual data frame):

Once I got the cluster ID, s I have merged it with the actual dataset for analysis.

My countries are (1. Sierra Leone 2. Niger 3. Congo, Dem. Rep 4. Chad 5. Central African Republic)

Question 2:

Answer:

State at least three shortcomings of using Principal Component Analysis.

1) The PCs have to be linear combinations of the original column: This means it will only work with the linear algorithms as it expects there should be some linear relationship between the variables.
2) PCA requires the PCs to be uncorrelated/orthogonal/perpendicular: It assumes that there should not be any relationship between the dependent variable.
3) PCA assumes low variance components are not very useful: It means if the column is having a low variance suppose a date field is having only one date for a dataset then it is not useful.

Question 3

Compare and contrast K-means Clustering and Hierarchical Clustering.

1. KMeans needs a prior knowledge of number of centroid (K) whereas hierarchical cluster do not need this kinds of parameters where as you use cut_tree () function to create the number of clusters of your choice
2. KMeans clustering is a lazy learner as every time the algorithm will calculate the centroid.
3. KMeans is fast compare to hierarchical and also hierarchical clusters need more ram to run.