Clustering & PCA Assignment II

Question 1: Assignment Summary

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: -

Below was my job

To identify top countries that are direst need of aid with the provided dataset

Steps Followed:

1) Reviewed the details of the dataset using pandas .head, .info & .describe commands
2) Removed the outliers for high GDP and income.
3) Since the objective of the NGO is to allocate money to countries who are dire need, I didn't remove the outliers who are lagging behind economically and socially
4) Also I removed the outliers who are much ahead econimically and socially. Otherwise these countries will push away other countries from their cluster and we might get high number of countries in the cluster we are looking for.
5) Did PCA after applying standard sclaler
6) Reviewed explained_variance_ratio_ and plotted Scree plot to identify how many PCA components are required.
7) Decided to take 5 PC's
8) Reviewed both Silhouette & Elbow curve to check optimum number of cluster in K-Means cluster
9) Went with 3 cluster in K-Means
10) After that moved with Hierarchical Clustering by Complete linkage since simple linkage didn't derive any conclusion.

Post this I compared the final list generated by K-Means clustering and analysis and Hierarchical Clustering by Complete linkage and concluded both lists are almost identical.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.
b) Briefly explain the steps of the K-means clustering algorithm.
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
d) Explain the necessity for scaling/standardisation before performing Clustering.
e) Explain the different linkages used in Hierarchical Clustering.

Answer: -

a) In K-Means clustering the number of cluster needs to be decided beforehand whereas in Hierarchical clustering we can decide at the time execution

   K-means clustering can be used for big size of data whereas Hierarchical Clustering will have performance issues.

   K Means works well if shape of the clusters is hyper spherical

b) Algorithm of K-means clustering –

   We have N data points X = {x1,x2,x3,….,xn} with centres V = {v1,v2,……vc}.

   c cluster centres are selected randomly.

   Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.

   Repeat the same process again and continue in these steps.

c) If the value of K increases, there will be fewer elements in the cluster.

   So average distortion will decreases in this case. The lesser number of elements means closer to the centroid.

   In statistics terminology this is a task of grouping set of objects in such a way that similar or more similar objects comes closer and create a group and its centroid. This groups are called clusters.

d) There are 3 types of Hierarchical Clustering –

   Single Linkage – shortest distance between points in the two clusters

   Complete Linkage – maximum distance between any 2 points in clusters

Average Linkage – average distance between every point of one cluster to every other point of the other cluster

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.
b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.
c) State at least three shortcomings of using Principal Component Analysis.

Answer: -

a) 3 application of using PCA –
   Facial recognition or object detection
   Video file analysis
   Data mining
b) Building block of PCA –
   1) Lowers noise sensitivity and increased efficiency gives the processes taking place in a smaller dimension.
   2) Standardization is a crucial step in PCA
   3) Dimension reduction

c) 3 shortcomings of PCA –

   1) Difficult to represent original feature using linear combination of PC's
   2) High of features with Small number of variances can be problematic for PCA
   3) Standardization is a must for PCA