CLUSTERING & PCA ASSIGNMENT

ASSIGNMENT PART II

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:
CEO of HELP International humantarian NGO needs to decied how they can use the money strategically and more effectively by choosing the countries that are in the direst need of aid.

Methodology:
With the given “country dataset” ,we did
-data cleaning and inspection if data
#Where we observe that there are no null” values in the dataset.
#there is sudden spike in the data values after 99th percentile. Hence ,removed the highest value which is usually high for developed countries,leaving behind lower values in the dataset whenever applicable.

-carried univariate analysis
# we didnt observe any abnormility in the box plot .

-bivariate analysis
#After which we got to know,few socio-economic factors are high for countries with high income such as GDPP,Health.
Also, we observe the inflation is high for countries with high fertility rate and mortality rate..etc.

Later we performed,PCA ,where we observe there is zero correlation amomgst the principle components as expected.
We also performed visualisation on clusters that have been formed by plotting scatter plot of all countries and differentiating the clusters.

Further ,we calculated Hopkins score,performed Silhouette analysis and elbow curve.

With help of Hierarchical clustering, we completed our findings for final list of countries .

Question 2

State at least three shortcomings of using Principal Component Analysis.

Answer:

i. <u>Relies on linear assumptions</u>

PCA is focused on finding orthogonal projections of the dataset that contains the highest variance possible in order to 'find hidden LINEAR correlations' between variables of the dataset. This means that if you have some of the variables in your dataset that are linearly correlated, PCA can find directions that represents your data.

ii. <u>Mean and covariance doesn't describe some distributions</u>

There are many statistics distributions in which mean and covariance doesn't give relevant information of them

iii. <u>Relies on orthogonal transformations</u>

Sometimes consider that principal components are orthogonal to the others it's a restriction to find projections with the highest variance:

Question 3

State at least three shortcomings of using Principal Component Analysis.

Answer:

- Hierarchical clustering can’t handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n^2).
- 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