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:

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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes. After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Objective:

The requisite is:

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

Method followed:

-Data Processing:

- It was found that there were no null values
- There were also no duplicate values for country
- There were a few outliers and they were treated later on during PCA
- The data was standardized for Principal Component Analysis

-Screeplot:

4 components are good enough to get a 95% of variance in the data. So PC is selected to be 4.

-Clustering:

- Both the methods K means and Hierarchical Clustering was used on the 4 PCA components
- For K means , K= 3 was taken using the elbow dip and silhouette analysis .
- While doing the Hopkins Statistics a value of 0.77 was attained.
- If the Hopkins Statistics values are:

- 0.3 : Low chase of clustering
- around 0.5 : Random
- 0.7 - 0.99 : High chance of clustering

-Finally using all these values clusters of 3 were formed and the countries are split into clusters.

Question 2

State at least three shortcomings of using Principal Component Analysis.

Answer:

The three shortcomings of using Principal Component Analysis:

- The PCs have to be linear combinations of the original column: This means it will only work with the linear algorithms since this method expects to have some linear relationship between the variables.
- PCA requires the PCs to be uncorrelated/orthogonal/perpendicular: It assumes that there will not be any relationship between the dependent variable.
- PCA assumes low variance components are not very useful: It means if the column is having a low variance.
  Eg: If there is a variable with only one element it is not considered to be useful.

Question 3

Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer:

- K Means needs a prior knowledge of number of centroid (K) whereas hierarchical cluster do not need this kinds of parameters. cut_tree () function is used to create the number of clusters of any choice.
- In K Means clustering the algorithm will calculate the centroid each time.
- K Means is fast compare to hierarchical clustering
- Hierarchical clusters need more ram to run.

x------x------x------x------x