Answer 1:

Clustering of Countries:

Problem Statement and Approach:

HELP International is an international humanitarian NGO that want to know names of countries which are in direst need of aid. Solving this problem can be achieved by finding most backward and poor countries list. Countries which have High Rate in child mortality rate, Fertility rate, inflation rate, health spending, life expectancy average and Low Rate in GDP per capita, Import, Export, Income per capita , depict that those countries are poor and backward.

Data Preparation: Data cleaning process done for data preparation where I checked for missing values, null values, and data info etc.

EDA:

1-Find out top 10 countries name which have high rate in child mortality rate, Fertility rate, inflation rate, health spending, life expectancy average.

2- Find out bottom 10 countries name which have low rate in GDP per capita, Import, Export, Income per capita.

Outliers Checking and Removing for range 0.5-0.95:

Scaling the Data: Scaling done using standard scalar.

PCA:

1- Feature variables:- all except of country, Response Variable :- country.
2- Two principal component created- PC1,PC2.
3- the screeplot created for checking cumulative variance against the number of components
4- I got result of 4 components for modelling.

K-Means Clustering:

Hopkins Statistics: provided result 0.7311293910509467 that clearly state that It has High Tendency to Cluster.

K-Means: Five clusters created for kmeans.

Silhouette Analysis: Found result of clustering with cluster id. Plot created for visualization.

Hierarchical Analysis:

1- Used “complete” method for merging as “single” method not giving clear result.
2- Dedrogram is observed that cutting it at n = 5 is most optimum.
3- Concatenation performed with clustered data and countries name.
4- Find mean value and cluster id from clustered data.
5- Result found as a list of Backward countries by printing cluster id zero values.

Answer 2:

Shortcomings of using Principal component Analysis-

PCA doesn't have any disadvantages. If used correctly, it should filter out the noise, and you should be left with a stronger signal.

However there are some limitations of PCA-

1- **Linearity**: PCA assumes that the PCA are a linear combination of the original features. If this is not true, PCA will not give you sensible results.

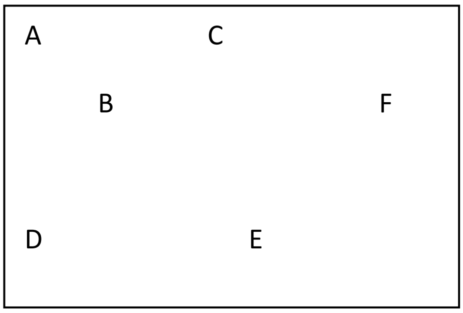
2- **Large Variance implies more structure**: PCA uses variance as the measure of how important a particular dimension is. So, high variance axes are treated as principle components, while low variance axes are treated as noise.

3- **Orthogonality**: PCA assumes that the principle components are orthogonal.
In real dataset where PCA fails because the above assumptions do not hold.

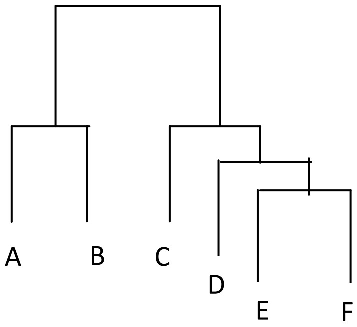
Answer 3:

Compare and Contrast K-Means Clustering and Hierarchical clustering:

Hierarchical Clustering :
In hierarchical clustering, clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster. Hierarchical Clustering can be potentially very useful for various data mining tasks. A hierarchical clustering scheme produces a sequence of clustering in which each clustering is needed into the next clustering sequence. Since hierarchical clustering is a greedy search algorithm based on a local search, the merging decision made early in the agglomerative process are not necessarily the right ones. One possible solution to this problem is to refine a clustering produced by the agglomerative hierarchical algorithm to potentially correct the mistakes made early in the agglomerative process. Hierarchical methods are commonly used for clustering in Data Mining. A hierarchical clustering scheme produces a sequence of clustering in which each clustering is nested into the next clustering in the sequence.



Before Clustering

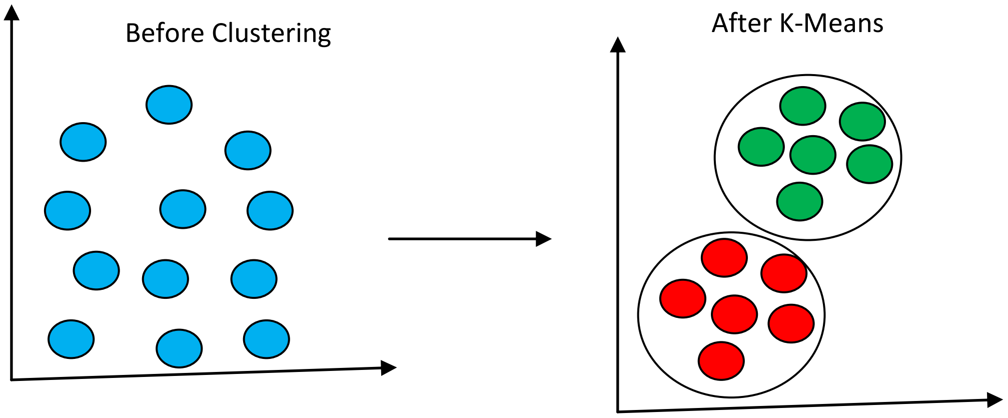


Dendrogram

[Hierarchical Clustering]

K-Means Clustering: is a well known partitioning method. In this objects are classified as belonging to one of K-Groups. The results of partitioning method are a set of K clusters, each object of dataset belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where consider real-valued data, the arithmetic mean of the vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases.

Ex- A cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset. K-Means is a data mining algorithm which performs clustering of the data samples. The division of a dataset into a number of groups such that similar items falls or belong to same groups. In order to cluster the database, K-Means algorithm uses an iterative approach.



[K-Means Clustering]