ASSIGNMENT: Clustering & PCA Assignment

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)

ANSWER:

Problem Statement: To categorize the countries using some socio-economic and health factors that determine the overall development of the country, and to suggest the countries that are in the direst need of aid so that the NGO can decide how to use this money strategically and effectively.

Solution Methodology: The following approach has been used to arrive at the final list of countries.

1. EDA of the Country Data, incl. Outliers treatment
2. Principal Component Analysis, to reduce dimensionality before applying clustering techniques
3. K-means Method, to arrive at no. of clusters first and then categorizing the countries under each cluster
4. Hierarchical Clustering (Complete Linkage is preferred) to determine an optimal no. of clusters from the dendrogram.
5. Analyzing the distribution of Child Mortality, GDPP and Income of each Cluster to identify the list of countries which are in dire need of aid.

QUESTION 2: CLUSTERING

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer:

K-means Clustering	Hierarchical Clustering
- No. of clusters is decided using elbow curve and Silhouette Score - Clustering is carried out after deciding on the no. of clusters - Clusters are collection of similar objects, and dissimilar objects are not clustered together. - It is a low-cost method, provided the no. of maximum iterations is set to a finite/small value. - The members in the final clusters may vary (for the same no. of clusters) based on the initial centroid.	- No. of clusters is decided using dendrograms - Data is clustered hierarchically either using top-down (divisive) or bottoms-up (agglomerative) approach - The clusters have a tree-like structure, dissimilar objects get clustered together at higher levels of the tree. - It is a high-cost method as it hierarchically clusters all the available data. - The resulting cluster will not vary for the same approach selected. - Types of linkages: Single, Average, Complete

b) Briefly explain the steps of the K-means clustering algorithm.

Answer: Below are the steps:

1. No. of clusters identified (denoted by k)

2. In the dataset, k cluster-centers are placed randomly and far away from each other. These cluster-centers are called Centroids
3. Each point in the dataset is allocated to the nearest Centroid (using Euclidean distance). This is the Assignment Step.
4. The Centroids are re-calculated based on the mean of the data-points belonging to each cluster. This is the Optimisation Step.
5. The Step #3 (Assignment Step) and Step #4 (Optimisation Step) are repeated until the Centroids no longer change.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Answer: The 'k' value is normally chosen by either of the methods.
Statistical Method: A range of k-values are selected and the Silhouette Score for each k-value is determined. A higher silhouette score means data-points in one cluster poorly match with other clusters but well matched within the same clusters.
The elbow-curve of the k-values is plotted to find out the relation between Sum of Separated Errors (SSE) and no. of clusters. An optimal no. of cluster is the one where the curve straightens to the lower value of SSE.
The no. of clusters is decided by finding an optimal k-value which has a high Silhouette Score and low SSE (by referring to the elbow-curve).

d) Explain the necessity for scaling/standardization before performing Clustering.

Answer: In any dataset, the dimensions are highly likely to have different units/base and therefore the range of values would vary from dimension-to-dimension. A high numerical value in one dimension may give a wrong impression of higher weightage than a dimension with lower range of values, thus incorrectly clustering the data-points and vice-verse. It is therefore necessary to scale/standardize all the numerical dimensions and bring them to the same numerical range so that the impact/weightage of each dimension can be determined correctly and used for accurate clustering.

e) Explain the different linkages used in Hierarchical Clustering.

Answer: Below are the 3 types of linkages used in Hierarchical Clustering.

- Single: Distance between two clusters is defined as the shortest distance between two points of each cluster.
- Complete: Distance between two clusters is defined as the longest distance between two points of each cluster.
- Average: Distance between two clusters is defined as the average distance between each point of the cluster to every point of the other cluster.

QUESTION 3: PRINCIPAL COMPONENT ANALYSIS

a) Give at least three applications of using PCA.

Answer: PCA is method of dimensionality reduction in datasets with high no. of dimensions. PCA is used in the following:

- Facial Recognition and other image processing techniques: The no. of components in an image data is huge and for faster comparison of 2 or more images, PCA is used.
- Financial Data Analysis: For risk management based on different portfolios and customer data, where the no. of attributes/dimension is high.

- Data visualization: Visualizing data with multiple dimensions and analyzing the same is not practical. PCA helps in reducing the no. of dimensions to 2 or 3 which can be plotted for visual reference and analysis.

b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

Answer: Basis is the fundamental units in which the data is defined. In case of multiple dimensions, the basis is transformed in order to achieve a more linear relation between the dimensions. Variance in dimension indicates that the dimension contains more information or not; higher variance implying more information. In PCA, instead of selecting one or more dimension and ignoring the others, the direction of maximum variance is determined in order to reduce the no. of dimensions while considering the variance from each of the dimensions. In that way, information from each dimension is retained based on their variance instead of completely eliminating the dimensions of lower variance.

c) State at least three shortcomings of using Principal Component Analysis.

Answer:

1. PCA cannot be applied if data is not linear or dimensions are not linearly correlated
2. In order to achieve maximum variance, PCA loads large no. of dimensions in the initial PCs, sometime making it difficult to identify the dimensions which are important.
3. PCA can be applied on numerical variables and not on categorical variables.