Assignment: Part 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:
Problem Statement: From the given data we need to choose the countries that are in the direst need of aid.
We have been provided with the dataset which contains details about Child Mortality, Income of the country, Export done from country, import done in country, Gross Domestic Product Per Capita etc.
Looking at this data we need to identify which countries are in need to aid. And report top 4 to 5 countries which are in need and can be helped using fund money.

Solution:
Step 1: First we need to learn and understand the data. (Data Checking: Whether data contains null values or not, check for Data types of data and also perform the basic data exportation steps. Etc.)

Step 2: Check whether all the data is in same scale or not (if not perform Scaling to scale the data and keep all the columns in one scale)

Step 3: Dimensionality Reduction: Using PCA we need to keep only those importance components which we have obtained by performing PCA.

Step 4: Now perform K-Means Cluster to identify the number of clusters needed to be formed for this data.

Step 5: Once we got the cluster then we need to merge it with the original dataset so to identify which columns re import and which re not.

Step 6: Based on Cluster we decide which columns are important and which are not

In the above assignment we have done dimensionality reduction using PCA and based on clusters we have identified which countries need aid

Question 2: Clustering

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

Answer:

K – Means Clustering	Hierarchical Clustering
K-Means clustering is an easy way to identify cluster.	Hierarchical clustering outputs a hierarchy, i.e. a structure that is more informative than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrogram.
K-Means may be than hierarchical clustering when the data is large.	Hierarchical clustering is easy to implement.
An instance can change cluster (move to another cluster) when the centroids are recomputed.	It is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around.
Difficult to predict the number of clusters (K-Value)	Time complexity: not suitable for large datasets

Initial seeds have a strong impact on the final results	Initial seeds have a strong impact on the final results
The order of the data has an impact on the final results	The order of the data has an impact on the final results
	Very sensitive to outliers

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

Different steps involved in implementing K-Means Clustering Algorithm are:

1. Missing values treatment: Imputing missing values its mean or medium or with some other values.
2. Data transformation: Converting % to the real numbers
3. Outlier treatment: If outliers re detected in data treating the outlier is important
4. Data standardization: All features in data frame should be in same scale.
5. Finding the optimal value of K: Finding optimal k value of clustering
6. Implementing K Means algorithm
7. Analyzing the clusters of customers to obtain business insights: Once we have identified the K clusters then based on this k cluster we need to give business recommendation.

Once we are through with the data preparation, the K-means algorithm is quite easy to implement. All it takes is running the KMeans () function. Here we need to decide the number of required clusters beforehand and run the algorithm multiple times to get the most optimal number of clusters.

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:
There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.

d) Explain the necessity for scaling/standardization before performing Clustering.
Answer:
In statistics, standardization (sometimes called data normalization or feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale.
Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 v/s 0-1000).
The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

When working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable.
For example, One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another, it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unit less measure or relative distance.

e) Explain the different linkages used in Hierarchical Clustering.
Answer:
In hierarchical clustering there are 3 types of linkages used. They are as follows:

1. **Single-Link**
2. **Complete Linkage**

3. **Average Linkage**

Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
Average Linkage: Here, the distance between 2 clusters is defined as the 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.
Answer:
1. PCA is used as a dimensionality reduction technique in domains like facial recognition, computer vision and image compression. It is also used for finding patterns in data of high dimension in the field of finance, data mining, bioinformatics, psychology, etc.
2. PCA can be applied to disease control, especially when you research focus transcends beyond just the investigation to the causes and analysis of one type of disease.
3. PCA also has varied applications in Quantitative Finance with particular bias in evaluating stock portfolio, energy pricing and stock selection for technical trading.

b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.
- Basis Transformation: In PCA we change the basis of data (i.e. we find new basis) in which the representation becomes much more useful to us. It is the fundamental concept behind PCA.
- Variance as information: Variance is a numerical measure for the variation which occurs in that particular field or column. If a column has very little variance then the information contained in that column is very low and vice-versa.

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

1. **Independent variables become less interpretable:**
- After implementing PCA on the dataset, original features will turn into Principal Components. Principal Components are the linear combination of original features. Principal Components are not as readable and interpretable as original features.
2. **Data standardization is must before PCA:**
- We need to standardize the data before implementing PCA; otherwise PCA will not be able to find the optimal Principal Components.
3. **Information Loss:**
- Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features.