**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**

Basically in this assignment our task is to find poorest countries which need financial aid the most with the help of socio-economic and health factors and present them to CEO who will provide funding.

Starting with Principal Component Analysis I have done using scree plot I could see that

Around 95% of the information is being explained by 3-4 components and thus took 3 components.

So I performed the clustering using the PCs and have allocated the cluster IDs back to each of the data points. thus I have put countries in 3 different clusters based on given variables like gdpp, income, child death rate and analysed which countries are in direst need of aid and present bottom five to CEO of the company.

Using both clustering techniques (k-means and Hierarchical) we got almost similar results. But in this case I chose to prefer hierarchical clustering since data is not big and assumptions not required.

Question 2: Clustering

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

Answer:

- Hierarchical clustering can’t handle big data well but K Means clustering can.
- In K Means clustering, we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are easily reproducible in Hierarchical clustering technique.
- Thus hierarchical clustering is usually preferable, as it is both more flexible and has fewer hidden assumptions about the distribution of the data.

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

- Initialize cluster centres (Randomly pick cluster-centres)
- Assign observations to the closest cluster center.
- Revise cluster centres as mean of assigned observations.
- Repeat step 2 and step 3 until convergence(no change in center further)

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

We use two techniques to choose no. of k

- First is using silhouette score
- Another is elbow curve method

Other than above methods we use

**Purpose Based**: we can run k-means clustering algorithm to get different clusters based on a variety of purposes.

we can partition the data on different metrics and see how well it performs for that particular case as per our business requirement.

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

In **Statistics**,

**Standardization** (data normalization or feature scaling) refers to the process of rescaling the values of the variables in our data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if we 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 vs 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 we are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable.

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 (because the field with the wider range of values likely has greater distances between values), 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**

Method of **single linkage** or nearest neighbour.

Proximity between two clusters is the proximity between their two closest objects.

Single linkage method controls only nearest neighbours similarity.

Method of **complete linkage** or farthest neighbour.

Proximity between two clusters is the proximity between their two most distant objects.

Method of between-group **average linkage**.

Proximity between two clusters is the arithmetic mean of all the proximities between the objects of one, on one side, and the objects of the other, on the other side.

**Question 3: Principal Component Analysis**

a) Give at least three applications of using PCA

**Ans.**

Main applications of PCA are

- data compression
- image processing
- visualization
- exploratory data analysis
- pattern recognition
- time series prediction

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

**Answer—**

**Basis Transformation--**

PCA simply takes points expressed in the standard basis and transforms them into points expressed in an eigenvector basis where vector tells direction of new feature and value tells magnitude.

In this process of transformation, some dimensions with low variance which are redundant are discarded and hence the resulting dimensional reduction.

**Variance as Information—**

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate.

Basically in PCA we tend to get more variance in order to get more information so we get our $1^{st}$ component having maximum variance i.e. most information is explained by that.

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

**Answer:**

- **Data standardization is must before PCA**
- **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:
- **Independent variables become less interpretable:** After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features**.**