Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words.

Problem Statement – An NGO has got 10M funding and wants to use this money in various countries fighting from poverty. They want to decide which are the most backward countries and are in the direst need of aid from the list of countries.

Solution Methodology - First performed the PCA on data, out of 9 principal components selected only 5 components as they were explaining the 95% variance in the data. Then performed k means and hierarchical clustering, in both the models divided the data into three clusters so as to classify countries as under-developed, developing and developed. Hierarchical clustering clustered most of the countries in single cluster so results were not much reliable. K means cluster model produced better results. Visualized the clusters of k means based on gdpp, child mortality and income. Then selected that cluster in which income and gdpp were minimum and child mortality was maximum.
This cluster countries to be presented to NGO which are in direst need of aid.

Question 2: Clustering

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

In hierarchical clustering, we start with each data item having its own cluster. We then look for the two items that are most similar and combine them in a larger cluster. We keep repeating until all the clusters we have left are too dissimilar to be gathered together, or until we reach a preset number of clusters.

In k-means clustering, we divide the data into k sets at the same time and then assign all items to the cluster, which is closest, and then calculate the cluster mean as the new representative, until it converges.

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

1) Randomly select k cluster centers or as required according to business need.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum from the cluster centers.

4) Recalculate the new cluster center using:

$$v_i = (1 / c_i) \sum_{j=1}^{c_i} x_i$$

where, ‘c_i’ is the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) Repeat from step 3 until no data point is reassigned.

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

Elbow method 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 K value increases, there will be less number of data points in the cluster. So average distortion will be less. The lesser number of data points means closer to the centroid. So, the point where this distortion declines the most is the elbow point.

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

Standardisation controls the variation in the dataset, it converts data into specific range using linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms.

e) Explain the different linkages used in Hierarchical Clustering.

In single-link clustering the similarity of two clusters is the similarity of their most similar members. This single-link merge criterion is local. We pay attention solely to the area where the two clusters come closest to each other.

In complete-link clustering or complete-linkage clustering, the similarity of two clusters is the similarity of their most dissimilar members. This is equivalent to choosing the cluster pair whose merge has the smallest diameter. This complete-link merge criterion is non-local; the entire structure of the clustering can influence merge decisions.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

- It allows you to understand the data and reduce the dimension of data.
- In terms of disciplines, 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.
- It 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.

- In PCA, basis transformation is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.
- Variance means summative variance or multivariate variability or overall variability or total variability.

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

Model performance: PCA can lead to a reduction in model performance on datasets with no or low feature correlation or does not meet the assumptions of linearity.
Classification accuracy: Variance based PCA framework does not consider the differentiating characteristics of the classes. Also, the information that distinguishes one class from another might be in the low variance components and may be discarded.
Outliers: PCA is also affected by outliers, and normalization of the data needs to be an essential component of any workflow.