SUBJECTIVE QUESTIONS

Question 1: Assignment Summary

Problem Statement: HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes. After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

1. Reading and Understanding the Data
2. Data Cleansing
   - Missing Value check
   - Data type check
   - Duplicate check
3. Data Visualization
   - Heatmap(Visualization of the correlation of data)
   - Pairplot
4. Data Preparation
   - Rescaling
5. PCA Application
   - Principal Components Selection (After comparing variance ratio and cumulative variance against number of PCA components, it is inferred that: more than 90% variance is explained by the first 3 principal components)
   - Outlier Analysis and Treatment
6. Hopkins Statistics Test

   Hopkins Score Calculation (Hopkins Score calculation gave a score of 0.8215813724135352)

7. Model Building
   - K-means Clustering
   - Elbow Curve(It is inferred that 4 or 5 clusters would be good, Also we used the boxplot visualisation of clusters against child mortality, income and GDP, we inferred that clusters 0 and clusters 3 are in need of aid.)
   - Silhouette Analysis

- Hierarchical Clustering (After performing the same process of scatter plot visualisation in the case of hierarchical clustering, it seems that K-means provided better information than Hierarchical Clustering. So, we proceed for the final analysis with K-Means. We merged the countries of cluster 0 and cluster 3 for final analysis.)

8. Final List Preparation
    - Final Country list Preparation

Question 2: Clustering

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

- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n^2).
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.

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

K-Means starts by randomly defining *k* centroids. From there, it works in iterative (repetitive) steps to perform two tasks:

1. Assign each data point to the closest corresponding centroid, using the standard Euclidean distance.
2. For each centroid, calculate the mean of the values of all the points belonging to it. The mean value becomes the new value of the centroid.

Once step 2 is complete, all of the centroids have new values that correspond to the means of all of their corresponding points. These new points are put through steps one and two producing yet another set of centroid values. This process is repeated over and over until there is no change in the centroid values, meaning that they have been accurately grouped. Or, the process can be stopped when a previously determined maximum number of steps has been met.

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 value of K to perform the K-Means Clustering Algorithm. K-means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc.

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

Scaling/Standardisation is necessary to give equal weight to the data. It controls the variability of the dataset; it converts data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms.

e) Explain the different linkages used in Hierarchical Clustering.

Single Linkage:
In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster.
Complete Linkage
In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

PCA is used as a dimensionality reduction technique in domains like facial recognition, computer vision and image compression.

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

Change of **basis** via **PCA**. We can **transform** the original data set so that the eigenvectors are the **basis** vectors and find the new coordinates of the data points with respect to this new **basis.**

In case of **PCA**, "**variance**" means summative **variance** or multivariate variability or overall variability or total variability.

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

Disadvantages of Principal Component Analysis

1. 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.

2. Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.