Question 1: Assignment Summary

Problem

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 programs, 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.

Approach

- Data Understanding
    - Checking for Null Values
    - Using description method to understand the data
- Data Cleaning
    - None of the rows have missing values
    - All variables are int64/float64 excluding country, no categorical variables
    - No Data Preparation or Dummy variable creation is required for the columns in this dataset
- Performing PCA
    - Data Scaling and Standardization
    - Performed PCA and choose the PCs that have *more than 85% variance*.
        - *We identified 5 PCS based on score*
    - Run the PCA with the 5 chosen PCs.
- Perform Clustering
    - Data Preparation for Clustering
        - Outlier Treatment
        - Hopkins Check
    - Clustering
        - K-Means Clustering
            - Run K-Means and choose K using both Elbow and Silhouette method.
            - Run K-Means with chosen K
            - Visualize the Clusters
            - Clustering Profiling
        - Hierarchical Clustering
            - Single Linkage
            - Complete Linkage
            - Visualize the Clusters
            - Clustering Profiling
- Country Identification
    - K-Means Clustering
    - Hierarchical Clustering
    - Combing the Results
    - 5 Countries which need immediate attention in 3 area.
        - *Child Mortality* : Mozambique, Malawi, Togo, Guinea, Comoros
        - *Income*: Congo, Dem. Rep., Niger, Central African Republic, Sierra Leone, Burkina Faso, Haiti
        - *GDPP*: Congo, Dem. Rep., Niger, Central African Republic, Sierra Leone, Burkina Faso, Haiti
        - 20 Countries based on socio-economic and health factors.

Question 2: Clustering

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

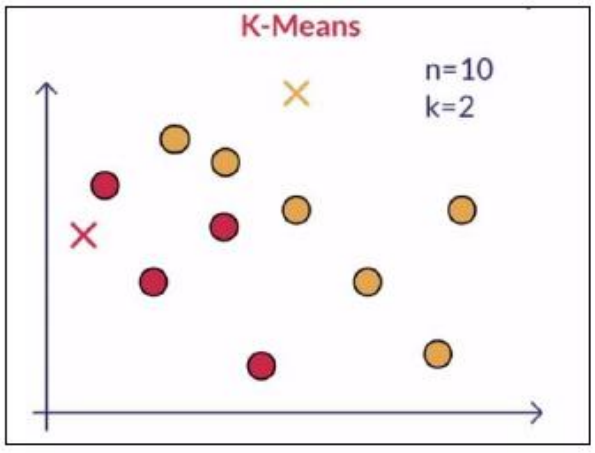
K-means Clustering	Hierarchical Clustering
Non-Linear process	Linear method
Randomly selects n points to start with nearest neighborhood.	Either start with all data points (agglomerative) as initial n clusters or decisive(top-down)
Recalculates its centroid- allows reorganization, happens in iteration.	There is no reorganization leafs.
Iteration need to be mentioned, might take more (suboptimal solution) or less (optimal solution).	One leaf once connected to a branch there is no way it can go back to other branch segment.
Does not consume more RAM within the system. Can be done in low infrastructure.	Lots of RAM, keeps on growing. Stops running on bigger size of Data in normal infrastructure. If on cloud then can be done on bigger server space.
Preferred for mall Large of Data, computationally less intensive.	Preferred for mall Set of Data, Hierarchical clustering generally produces better clusters, but is more computationally intensive.

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

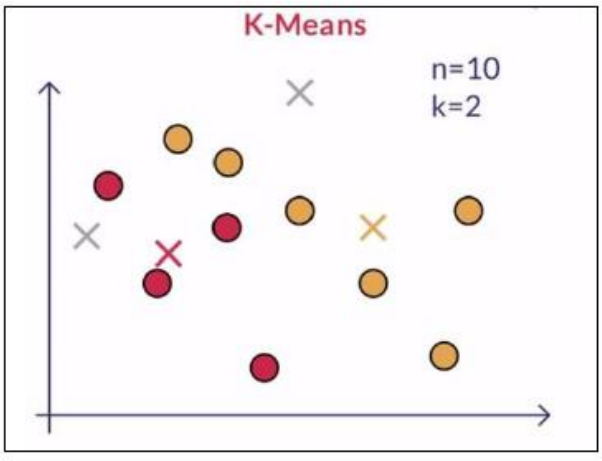
K-Means algorithm is the process of dividing the N data points into K groups or clusters.

Here the steps of the algorithm are:

- Start by choosing K random points the initial cluster centres.
- Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
- For each cluster, compute the new cluster centre which will be the mean of all cluster members.
- Now re-assign all the data points to the different clusters by taking into account the new cluster centres.
- Keep iterating through the step 3 & 4 until there are no further changes possible.



Assigning each data point to their nearest cluster



Updating the Cluster Centers

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

There are a number of pointers that can help us decide the K for our K-means algorithm:

Compute clustering algorithm different values of k. For instance, by varying k from 1 to 10 clusters

1. Elbow method:

- For each k, calculate the total within-cluster sum of square (wss).
- Plot the curve of wss according to the number of clusters k.
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

2. Average silhouette Method

- For each k, calculate the average silhouette of observations (*avg.sil*).
- Plot the curve of *avg.sil* according to the number of clusters k.
- The location of the maximum is considered as the appropriate number of clusters.

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

It is majorly used for in K-Means algorithm. Standardization of data, that is, converting them into z-scores with mean 0 and standard deviation 1, is important for 2 reasons in K-Means algorithm:

1. Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale helps in this process.

2. The different attributes will have the measures in different units. Thus, standardization helps in making the attributes unit-free and uniform.

e) Explain the different linkages used in Hierarchical Clustering.

- 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.

- PCA is a linear transformation method and works well in tandem with linear models such as linear regression, logistic regression etc., though it can be used for computational efficiency with non-linear models as well.
  It should not be used forcefully to reduce dimensionality (when the features are not correlated).
- PCA is predominantly 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.

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

- **Basis transformation**
  Using the analogy of basis as a unit of representation, different basis vectors can be used to represent the same observations, just like you can represent the weight of a patient in kilograms or pounds.

$\begin{bmatrix} 1ft \\ 0lbs \end{bmatrix}$ in ft/lbs space = $\begin{bmatrix} 30.48cm \\ 0kg \end{bmatrix}$ in cm/kg space and
$\begin{bmatrix} 0ft \\ 1lbs \end{bmatrix}$ in ft/lbs space = $\begin{bmatrix} 0cm \\ 0.45kg \end{bmatrix}$

Example $\begin{bmatrix} 165 \\ 55 \end{bmatrix}$ is same as $\begin{bmatrix} 5.4 \\ 121.3 \end{bmatrix}$

- **Variance as Information**
  In statistics, explained variation measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given data set. Often, variation is quantified as variance; then, the more specific term explained variance can be used.

c) State at least three shortcomings of using 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.
For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results.

Also, for standardization, all the categorical features are required to be converted into numerical features before PCA can be applied. PCA is affected by scale, so you need to scale the features in your data before applying PCA. Use **StandardScaler** to standardize the dataset features onto unit scale (mean = 0 and standard deviation = 1) which is a requirement for the optimal performance of many Machine Learning algorithms.

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.