Question 1: Assignment Summary

Problem Statement:

The international humanitarian NGO called HELP 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.

The CEO of the NGO needs a strategical and effective way to make use of 10 million dollars raised during funding programs and wants to choose which of the countries are in dire need of help.

Keeping in mind the socio-economic and health factors, countries must be classified to determine overall development of the country.

Solution methodology:

1. Firstly, required **libraries** are imported namely:
   A. NumPy and Pandas
   B. Matplotlib and Seaborn
   C. SKLearn libraries which includes:
      - StandardScaler
      - KMeans
      - Silhouette_score
   D. Scipy libraries for hierarchy which includes:
      - Linkage
      - Dendrogram
      - Cut_tree

2. Data is imported and all the necessary **data inspection** like null value and unique values checks are performed.

3. **EDA** is performed which includes univariate analysis using distplot which shows distribution of data points and bivariate analysis using pairplot which plots each and every variable against each other giving us the idea of how data points are stacked and heatmap which shows how correlated the variables with each other.

4. **Outlier analysis** is performed next where the outlier values are capped at different upper and lower boundaries depending on the column values, so that no important values is missed.

5. After outlier analysis, **scaling** of variables is done using StandardScaler library. Scaling is done to normalize the data within a range and if its not performed, output might get affected because of abnormal value range.

6. Next step in the process is **Clustering** the data points using K-Means and Hierarchical clustering process. First the number of cluster points are determined by plotting elbow curve method and silhouette score method from which we can get a brief idea of how many cluster points will be nominal. In Hierarchical clustering we get dendrograms from linkage library and the number of clusters needed can be finalized by cut_tree library. Number of cluster points has been chosen to be 3 by comparing all the methods mentioned above.

7. The 3 cluster points are renamed as Under-Developed, Developing and Developed countries and the clusters are **analysed** by plotting the variables GDP, child_mort and income to differentiate the clusters of developed and underdeveloped countries.

8. The **Visualization** using bar plots shows how GDP is affected from income, health, life expectancy, child mortality, imports, exports, and other factors.

9. Top 5 countries which are in dire need of help from NGO are listed from both methods by grouping and sorting cluster wise and came to final conclusion that the below mentioned countries need help based on health, child_mort, income, GDP and other factors.

    a. Burundi
    b. Liberia
    c. Congo, Dem. Rep.
    d. Niger
    e. Sierra Leone

Question 2: Clustering

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

K-means Clustering:

K-Means clustering is a clustering algorithm which groups data points in the same cluster whose values are like each other and the method is called distance measure. Here the number of clusters are decided beforehand.

K-means gives a line graph in either **elbow** curve method or **silhouette** method to show number of clusters.

The distance measure that is used in k-means is called Euclidean sum of squares. **Euclidean** distance is calculated from each data point to its nearest centroid.

K-means is not suitable for all shapes, sizes, and densities of clusters, it works well for spherical shape.

Hierarchical Clustering:

Hierarchical clustering visually describes the similarity or dissimilarity between the different data points and then decide the appropriate number of clusters instead of pre-defining.

Hierarchical clustering shows an inverted tree shaped structure called **dendrogram** and number of clusters can be decided using a library called cut tree.

Hierarchical clustering follows bottom up approach in **Agglomerative** clustering and top down approach in **Divisive** clustering.

Hierarchical clustering has different linkages like single linkage (the distance between 2 clusters is shortest), complete linkage (distance between 2 clusters is maximum) and average linkage (distance between 2 clusters is average distance between every point of one cluster to every other point of other cluster).

Hierarchical cluster cannot handle big data as well as k means cluster since k means is linear and hierarchical is quadratic.

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

The method in which any clustering algorithm goes about finding data points whose values are like each other to group them in same cluster is through the method of finding called 'distance measure'.

The distance measure that is used in k-means clustering is called Euclidean Sum of squared errors (SSE) is used as the objective function for K-means clustering with Euclidean distance.

$$D = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots (X_n - Y_n)^2}$$

The Euclidean distance is calculated from each data point to its nearest centroid (Centroids are essentially the cluster centres of a group of observations). These distances are squared and summed to obtain the SSE. The aim of the algorithm is to minimize the SSE. Note that SSE considers all the clusters formed using the K-means algorithm. distance measure. The 2 steps of choosing k random clusters are Assignment and Optimization.

The algorithm for K-means algorithm is as follows:

- Select initial centroids. The input regarding the number of centroids should be given by the user.
- Assign the data points to the closest centroid
- Recalculate the centroid for each cluster and assign the data objects again
- Follow the same procedure until convergence. Convergence is achieved when there is no more assignment of data objects from one cluster to another, or when there is no change in the centroid of clusters.
- The K-Means algorithm does not work with categorical data.
- The K-Means algorithm does not work with categorical data.

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

The K-Means clustering is used to find the number of optimal clusters required for a dataset. In this initially centroids are selected as per user input.

Business point of view:

Since the data set will not be labelled initially, we must cluster the data using k-means algorithm by initially providing cluster centers. The distance between cluster centres are calculated based on Euclidian sum of squares method.

Many a times the number of clusters must be formulated based on the business decisions and profiteering of organisation. To avoid expense initialisation of cluster points are carried out on a subset of data.

Statistical point of view:

Initiation of the centroids in a cluster is one of the most important steps of the K-means algorithm. Many times, random selection of initial centroid does not lead to an optimal solution.

To overcome this problem, the algorithm is run multiple times with different random initialisations. The sum of squared errors (SSE) are calculated for different initial centroids.

This procedure will ensure that the selection is random, and the centroids are far apart. The disadvantage of this method is that calculating the farthest point will be expensive.

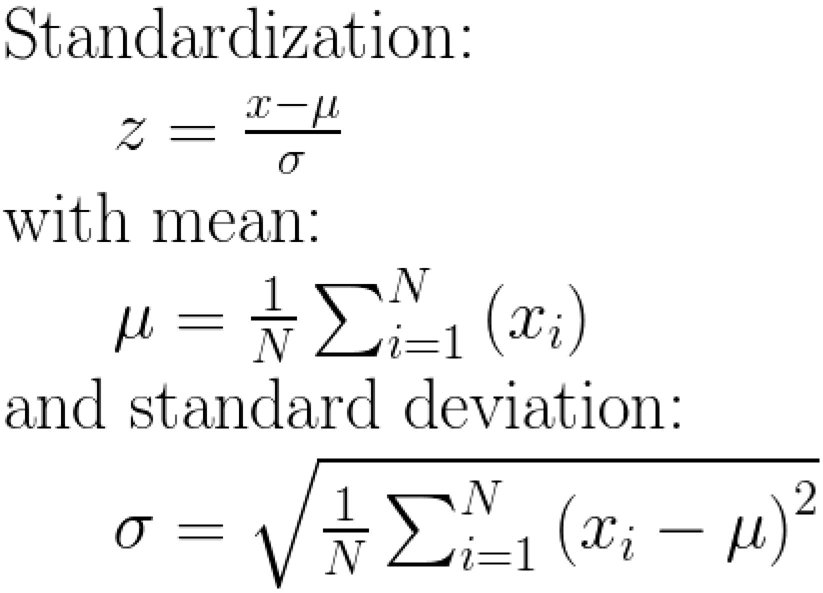
d. Explain the necessity for scaling/standardisation before performing Clustering.

There are few processes of scaling and MinMaxScaler and StandardScaler are popular of the lot.

Scaling helps to normalize the data within a range. It is also called as feature scaling. It helps to speed up the calculation.

Scaling is implemented in scikit learn that implements the Transformer API to compute the mean and standard deviation on a training set to be able to later reapply the same transformation on the testing set.

The calculation of standard scaler is shown below.



Standardization:

$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

e. Explain the different linkages used in Hierarchical Clustering.

The Linkage is measure of dissimilarity between clusters having multiple observations.

The following three methods differ in how the distance between each cluster is measured.

Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters

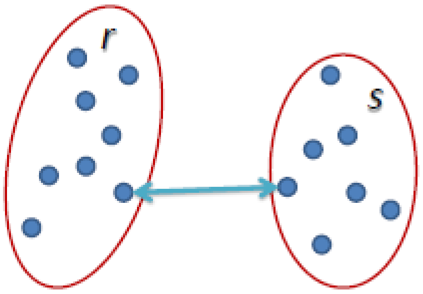
Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

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.

Average and Complete linkage methods give a well-structured dendrogram, whereas single linkage gives us dendrograms which are not very well structured.

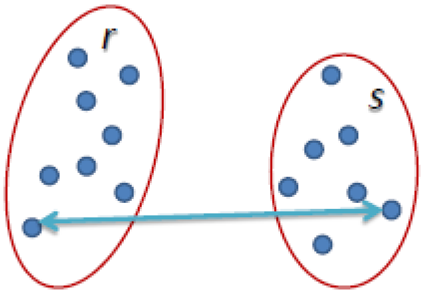
The visualization of all 3 types of linkages are as follows.

Single Linkage



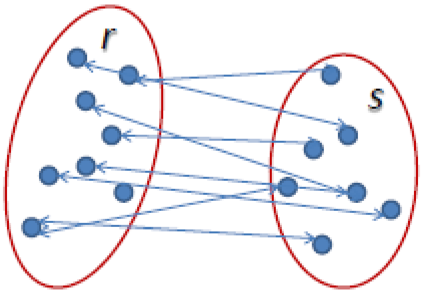
$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete Linkage



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average Linkage



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$