## Question1: 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.

**Methodology:**

As a Data Analytics, we need to categorize the countries with some socio-economical and health factors. That determines the overall development of the country. So, CEO suggested that which top 10 countries he need to focus the most to provide for immediate help.

So, Data set is provided, process is

1.The first step is to collect the data and read the data

2.To clean the data.

3.after cleaning the data, we need to handle the outliers present or not and then we need to do scaling.

4.By using Hopkins test we can check the cluster tendency check.

5.finding the value of K-means by using Silhouette and sum of squared distance.

6.Then we done with cluster profiling.

7.Then we implement hierarchical clustering with both complete single and complete linkage.

8.As we done with two methods, these two methods show the same 10 countries which are direst need of help.

## Question2: Clustering

### 1.Compare and contrast K-means clustering and Hierarchical clustering.

- In K-means Clustering we need to identify the location of elbow on the X-axis.
- In the above plot, the elbow seems to be on point 3 of X-axis.so the optimal number of clusters will be 3 for k-means algorithm.
- After finding the optimal clusters we need to fit the K-means clustering Model to the dataset.
- In Hierarchical clustering the dendrograms are used for this purpose.

**2.Breifly Explain the steps of the K-means clustering Algorithm.**

- K-means is one of the simplest and popular unsupervised machine learning Algorithm.
- Objective of K-means is group similar datapoints together and discover underlying patterns.

**Steps:**

- select initial centroids. The input regarding the number of centroids should be given by the user.
- Assign the data points to the closet centroid
- Recalculate the centroid for each cluster and the data objects again
- Follow the same procedure until convergence is achieved.

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

The ways we can choose the K-value is by two methods:
-Elbow curve method
- silhouette method

**1.Elbow Curve:**

- Elbow method runs the k-means clustering in the range of 1-10.
- For each k-value we need to calculate the average distances to centroids across all the data points.
- It uses the sum of squared distances method

**2.Silhouette Method:**

- It is measure of how similar a data point is with in the cluster (cohesion) compared to other clusters.
- It is also having a range of values of k.

Equation for silhouette method is:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

S(i)=coefficient of data point

a(i)= avg. distance between i and all other data point.

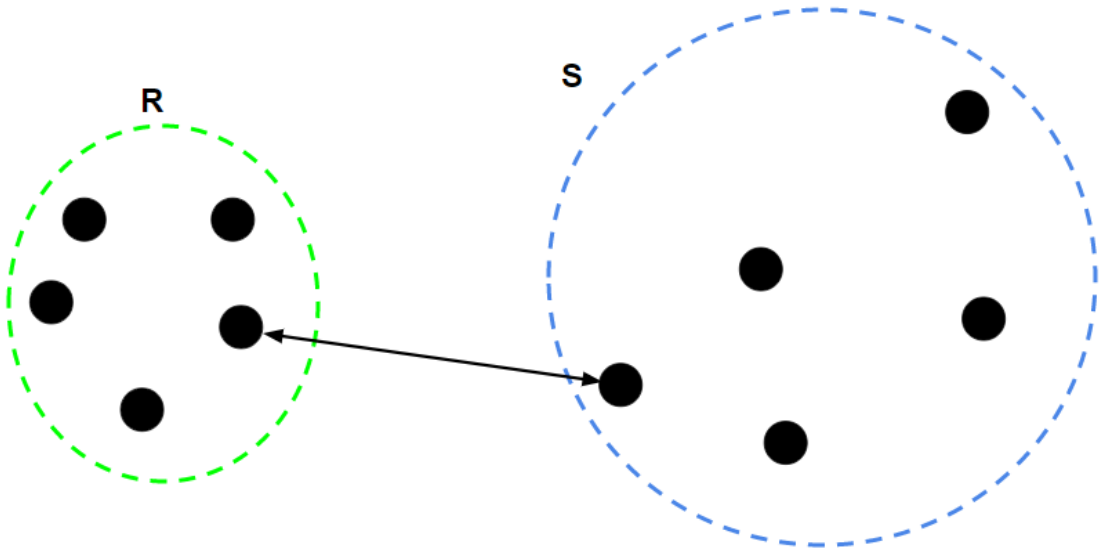
## 4.Explain the necessity for scaling/standardisation before performing Clustering?

- Data scaling ensures the feature/attribute is being weighted equally by the clustering algorithm.
- The necessity of performing the Standardisation before clustering to rescale the values of the variables in dataset so they share a common scale.
- Standardisation is important, where each variable has a different unit or where the scales of each Variables are very different from one another.
- Normalisation is done on each column separately.

## 5.Explain the different linkages used in Hierarchical clustering?

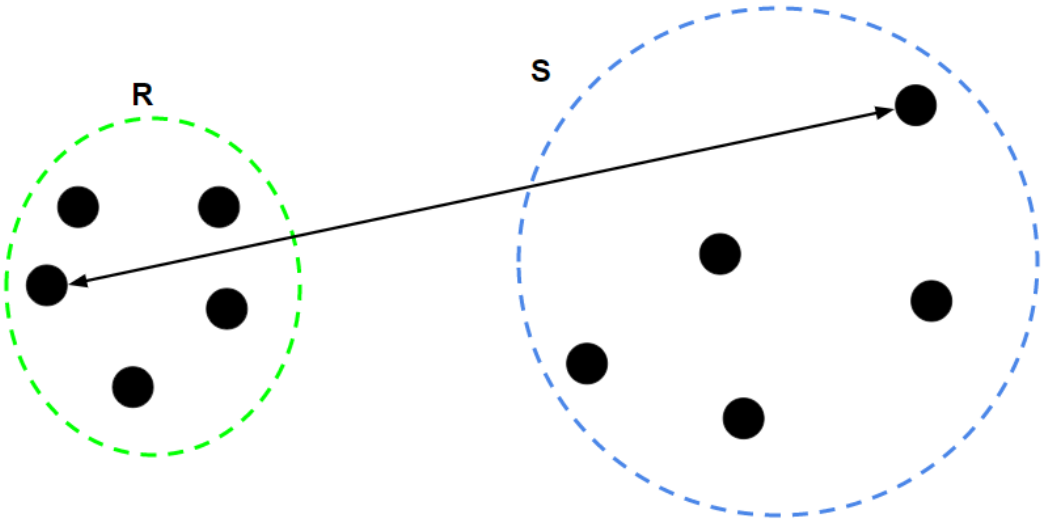
**Single linkage:**

For two clusters, the single linkage defines the minimum distance of the two points.




**Complete Linkage:**

For two clusters, the complete linkage defines Maximum distance of the two points.




**Average Linkage:**

For two clusters, the distance between two data points is calculated by arithmetic mean.