Question 1: Assignment Summary

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries (based on their GDPP, child mortality and income values) with basic amenities and relief during the time of disasters and natural calamities. The CEO of the company needs our help to identify the countries that are in the direst need of aid to use the recently acquired funds of $10 million strategically and effectively.

In order to accomplish this task, we did the univariate and bi-variate exploratory data analysis of each of the parameters provided in the given dataset to check the general tendency of data. We did the few outlier treatment based on business requirements/ intention of study. Further we took the clustering model approach to create cluster of countries based on the comparative parameters of child mortality, GDPP of the country and the income. The countries with high child mortality and low GDPP and income were separated by the clustering algorithm. To check whether the clustering was a feasible for the given data, we conducted Hopkin's test. The score of Hopkin's test is between 0 and 1 and the values above 0.85 is considered very good score. Our score of 50 such Hopkins test resulted in a 0.95 which was an indication of well-defined cluster.

Next, I scaled the data with a standard scaler to prepare the data and performed analysis to determine the most effective value of K (number of cluster) for the given data. Based on the outcome of both the sum of squared distance method and Silhouette score and the total number of countries, the K value of 3 was chosen. We then did the cluster profiling based on GDPP, child mortality and income. We observed by doing bi-variate analysis of clusters that the cluster 0 showed the high child mortality and low GDPP and income values. Out of the cluster 0, we further sorted the rows to get the top 10 countries that are in the direst need of aid.

Question 2: Clustering

Compare and contrast K-means Clustering and Hierarchical Clustering.

The following are the points of compare and contrast of K-means clustering and Hierarchical clustering methods:

- In K-means method, one must pre-specify the number of cluster whereas in Hierarchical, it need not be specified. Once the clustering is done, we can cut the dendrogram at any level for desired number of clusters. Hierarchical clustering creates a result dendrogram which can be interpreted, and we can select the number of clusters.
- The main principle of K-means clustering is using a pre-specified number of clusters, the method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance whereas Hierarchical clustering can be either divisive or agglomerative.
- The methods used in K-means clustering are less computationally intensive whereas the Hierarchical clustering is computationally more intensive
- K-means is suited for very large datasets whereas Hierarchical method is advisable for small datasets.
- The results of K-means clustering may differ since the starting point of choice of clusters are random whereas Hierarchical clustering produces consistent result that can be reproduced.

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

K-means is one of the simplest unsupervised learning algorithms for clustering problems. The procedure follows a simple and easy way to classify a given data set through a certain k number of clusters. One center is assigned randomly for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed. At this point we re-calculate k new centroids as center of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop, we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Here,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Steps of the algorithm

Let X = { x_1, x_2, x_3,……., x_n} be the set of data points and V = { v_1, v_2, v_3,……., v_c} be the set of centers.

1) Randomly select 'c' cluster centers.
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 of all 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' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

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

Statistical aspect: There are two statistical methods of choosing the k value as below:

Silhouette clustering

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

$$\text{silhouette score} = \frac{p - q}{max(p, q)}$$

p is the mean distance to the points in the nearest cluster that the data point is not a part of

q is the mean intra-cluster distance to all the points in its own cluster.

The value of the silhouette score range lies between -1 to 1.

A score closer to 1 indicates that the data point is very similar to other data points in the cluster,

A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

SSD - Elbow method

In the elbow method we run k-means clustering on the dataset for a range of values of k (say, k from 2 to 10), and for each value of k calculate the sum of squared distance (SSD) of the samples to their closest cluster center or the centroid.

Then, plot a line chart of the SSD for each value of k. If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best. The idea is that we want a small SSD, but that the SSD tends to decrease toward 0 as we increase k (the SSD is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). Here the intent is to choose a small value of k that still has a low SSD, and the elbow usually represents where we start to have diminishing returns by increasing k.

Business aspect: We came to an optimal value of K=3. We analyzed the dataset against both the Silhouette score method and the SSD or the elbow method and both these methods values and curves indicated the optimum values as 3 or 5 after which there was no significant improvements in the respective scores. Now out of 3 and 5, we chose 3 because our dataset was only 167 rows so for such a small dataset, it made more sense to go with the K value of 3.

Explain the necessity for scaling/standardization before performing Clustering.

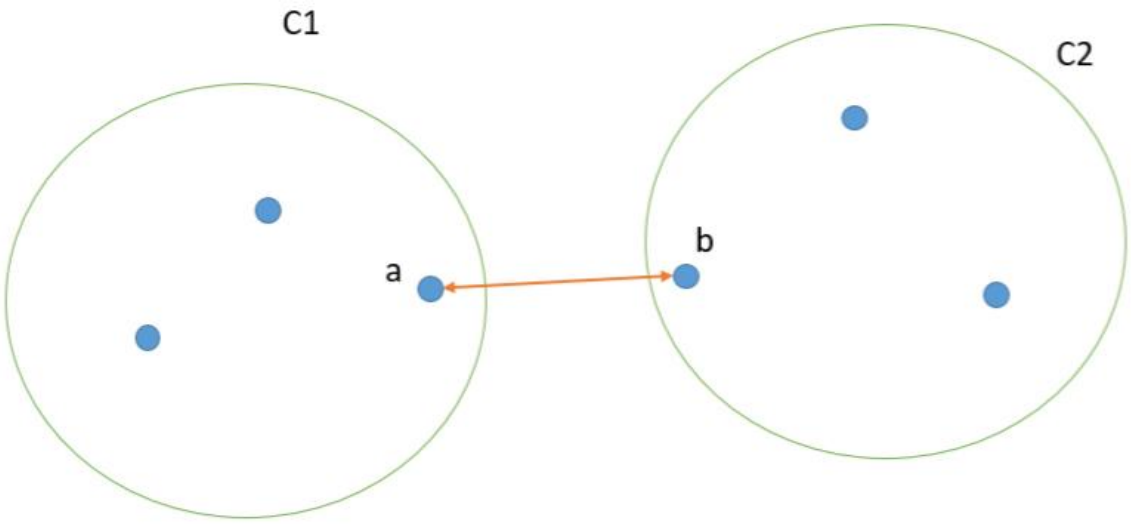
In real life we often get data in different types of units. For example, weight of a player could be in kgs with possible values between 45-70 kgs, whereas in the same data, the number of runs scored by the player could be in terms of 100-1000 or even more depending on the sport they play. The

process of re-scaling the values of variables in data is done so that they share a common scale. It is done as pre-processing step during cluster analysis because the clustering algorithms are based on distance between points in mathematical space. Thus, standardization helps to make the relative value of each variable equal by converting each variable to a unitless measure or a relative distance.

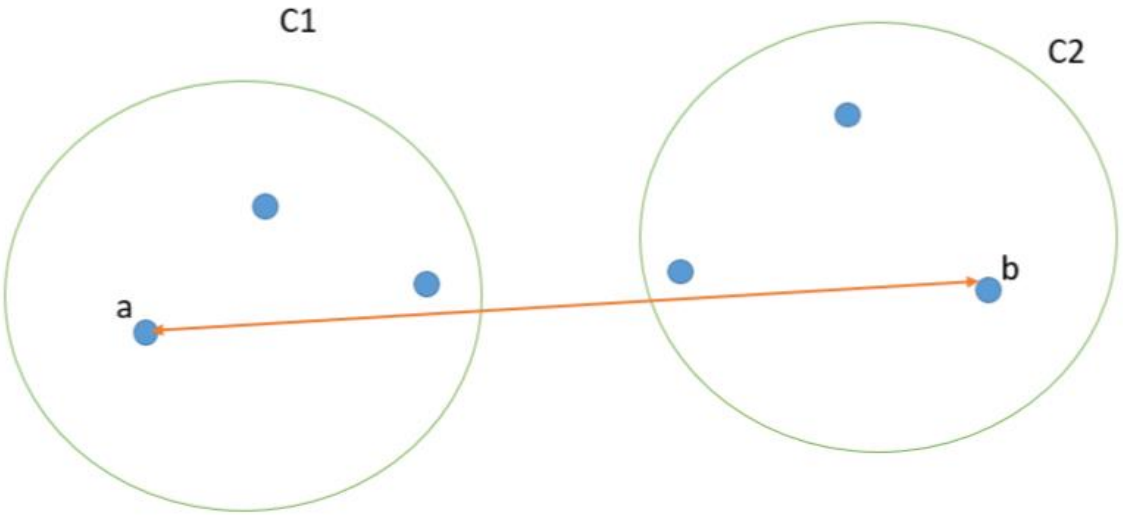
Explain the different linkages used in Hierarchical Clustering.

The process of Hierarchical Clustering involves either clustering sub-clusters (data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster into smaller sub-clusters in a top-down manner. During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed. The different types of linkages describe the different approaches to measure the distance between two sub-clusters of data points. The different types of linkages are:

1. Single Linkage: For two clusters C1 and C2, the single linkage returns the minimum distance between two points a and b such that a belongs to C1 and b belongs to C2.



2. Complete Linkage: For two clusters C1 and C2, the single linkage returns the maximum distance between two points a and b such that a belongs to C1 and b belongs to C2.



3. Average Linkage: For two clusters C1 and C2, first for the distance between any data-point a in C1 and any data-point b in C2 and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.