Question 1: Assignment Summary:

Ans:

The problem statement deals with providing recommendation to HELP International, an organisation that is committed to fighting poverty and providing basic amenities and relief during the time of disasters and natural calamities to under-developed countries, so that they can make strategic decisions on choosing the 'countries' which are in dire need of aid.

The Solution comprises of multiple activities being performed, which are listed below:

- **Deciding the model:** As our end goal is to provide list of countries which are in dire need of aid, so we have performed analysis and group the countries based on:
    - their social profile
    - their economic profile
    - their infrastructure & facilities

  So ***Clustering*** is the way forward here.
- **EDA:** As part of data check we performed following analysis -
    - Performing basic checks –
        - Number of Rows/columns (shape)
        - Datatype check for columns (info)
        - Checking range, mean, median etc. of numerical columns (describe)
        - Null check (isna)
    - ***Duplicate*** check on 'Country' field to ensure we have only entry per country
    - 3 features (exports, health and imports) were given as % of gdpp, there were converted to actual values.
    - Checked & treated ***Outliers*** –
        - As we are looking for under-developed countries who doesn't have infrastructure and resources, so we can remove the countries which are fully developed and have strong economy.
        - We capped our data up to 95% percentile of 'gdpp' feature value.
    - Checked ***Correlation*** among features using –
        - Heatmaps
        - Pairplots
- **Reset of Index:**
    - Country names were made as index as they were unique
    - We only need numerical features for clustering.
- **Model Building:**
    - ***Cluster Tendency check (Hopkins Test):*** Before applying clustering technique, we need to assess our data and find if it has clustering tendency present OR it is just a random distribution.
    - **Scaling of data:** After performing necessary EDA, we are now ready to build models but do that we need to bring our numerical features in similar value range. ***StandardScaler*** method was used to do so.
    - ***Model 1 – KMeans:***
        - Value of K was determined using – *Silhouette analysis* and *Elbow curve. (K=4)*
        - Model was built and labels/clusters were assigned.

        - Results analysis were performed, using barplots, Cluster which has those countries which are under-developed and have various socio-economic problem, were determined.
        - Top 5 Countries were determined from that cluster.
    - ***Model 2 - Hierarchical:***
        - Model was built and ***dendogram*** was plotted.
            - Using ***Single*** Linkage
            - Using ***Complete*** Linkage
        - Based on the ***dendogram,*** where to perform cut was decided and labels were assigned. (n_clusters = 4)
        - Results analysis were performed, using barplots, Cluster which has those countries which are under-developed and have various socio-economic problem, were determined.
        - Top 5 Countries were determined from that cluster.
- **Final analysis:** Both Clustering models produced same results i.e., same countries were recommended which are in dire need to Aid.

Question 2: Clustering

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

Ans:

K- means clustering is division of the set of data objects into non- overlapping subsets (**clusters**) such that each data object is in exactly one subset.

A **hierarchical clustering** is a set of nested **clusters** that are arranged as a tree.

K Means clustering needed advance knowledge of K i.e., no. of clusters one wants to divide the data. As one starts with random choice of clusters, the results produced by running the algorithm many times may differ. The method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance (Euclidean). Methods used are normally less computationally intensive and are suited with very large datasets.

On the contrast, in ***Hierarchical clustering*** one can stop at any number of clusters, one finds appropriate by interpreting the dendrogram and results are reproducible in Hierarchical clustering. Hierarchical methods can be either divisive or agglomerative.

Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.

Divisive methods work in the opposite direction, beginning with one cluster that includes all the records

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

Ans:

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid.

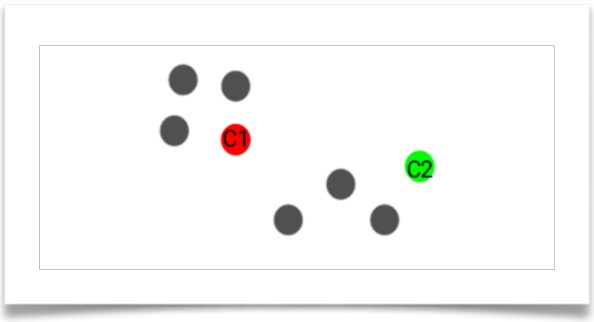
Steps involved in K-means are:

Step 1: Choose the number of clusters *k:*

The first step in k-means is to pick the number of clusters = K. This can be done in 2 ways - *Random selection* or using *Statistical method*s (SSD/Elbow curve or Silhouette Analysis)

Step 2: Select K random points from the data as centroids:

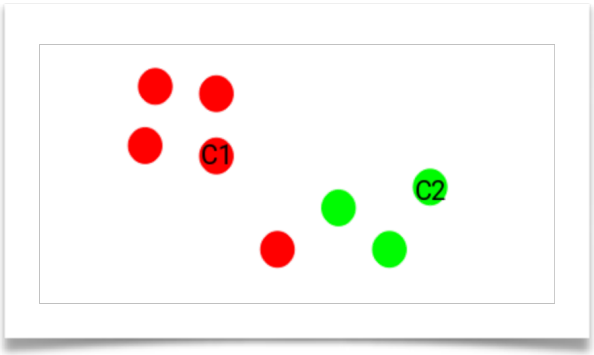
Next, we randomly select the centroid for each cluster. Let's say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid:



Example: Above diagram shows red and green circles which represent the centroid for these clusters.

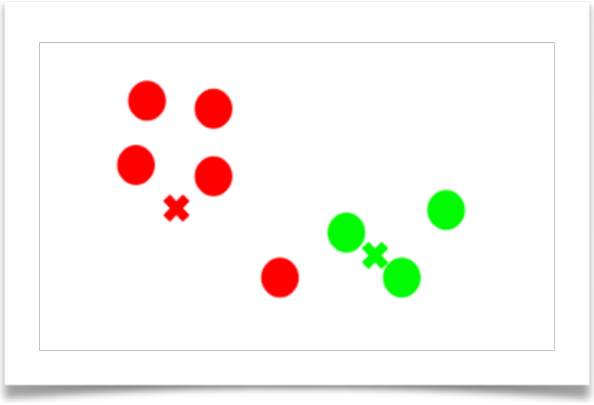
Step 3: Assign all the points to the closest cluster centroid:

Once we have initialised the centroids, we assign each point to the closest cluster centroid:



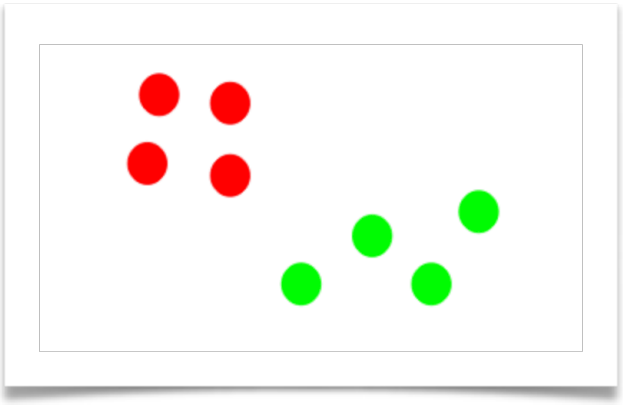
Here you can see that the points which are closer to the red point are assigned to the red cluster whereas the points which are closer to the green point are assigned to the green cluster.

Step 4: Recompute the centroids of newly formed clusters
Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters:



Here, the red and green crosses are the new centroids.

Step 5: Repeat steps 3 and 4
We then repeat steps 3 and 4:



Stopping Criteria for K-Means Clustering
There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:
1. Centroids of newly formed clusters do not change
2. Points remain in the same cluster
3. Maximum number of iterations are reached

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

Ans:
When we use a K-means clustering algorithm, we need to select the number of clusters we would like to work with. We can select the number of clusters using industry-related/ Business knowledge or using different statistical methods.

Industry related:

Indeed, choosing the value of K is often a **business decision**. This decision is made considering various aspects like What type of data we have? On what factors we want our data to be segmented? How many clusters we want?

Business always wants to understand and profile data in such way that they can see the patterns and work on it. This drives the decision for number of segmentation.

Of course we have statistical methods as well to determine this. Some of them are explained below.

Statistical Methods:

1. **The Elbow Method:** To determine the optimal number of clusters, we will need to run the k-means algorithm for different values of k (number of clusters). For each value of k, we will then calculate the total within-cluster sum of squares distance (ssd). We will then plot the values of ssd on the y-axis and the number of clusters (k) on the x-axis. We choose the value of K at the position when the decrease in the ssd for values of K begins increasing.
2. **The Silhouette coefficient:** To determine the optimal number of clusters, we will need to measure the quality of the clusters that were created. This value determines how closely each data point is to the centroid of its cluster. This measure has a range of [-1, 1]. The Silhouette Coefficient is calculated using the mean within-cluster distance/ variation (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). A high average silhouette coefficient indicates successful clusters. This method checks the silhouette coefficient for different values of k. The optimal number of clusters is, therefore, the maximised silhouette value for the data set.
3. **The Gap Statistic:** To determine the optimal number of clusters, we will need to know the variation between clusters for different values of k with their expected values of distribution with no clusters.

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

Ans:

Standardisation is an important step of Data preprocessing. It controls the variability of the dataset, it convert data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms.

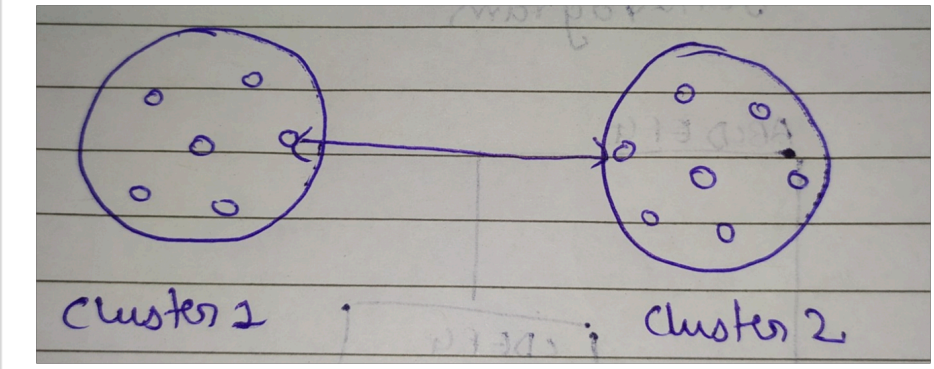
The idea is that if different components of data (features) have different scales of measurements, Standardisation prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance. Hence, it is always advisable to bring all the features to the same scale for applying distance based algorithms like KNN or K-Means.

e) Explain the different linkages used in Hierarchical Clustering.

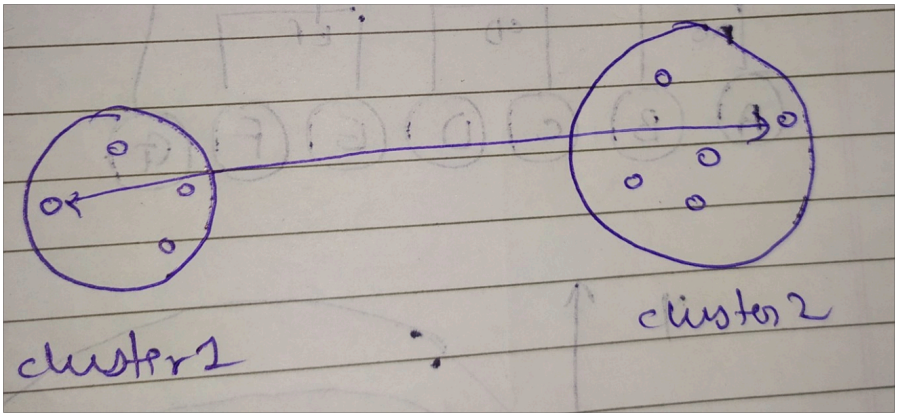
Ans:

Hierarchical clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster.

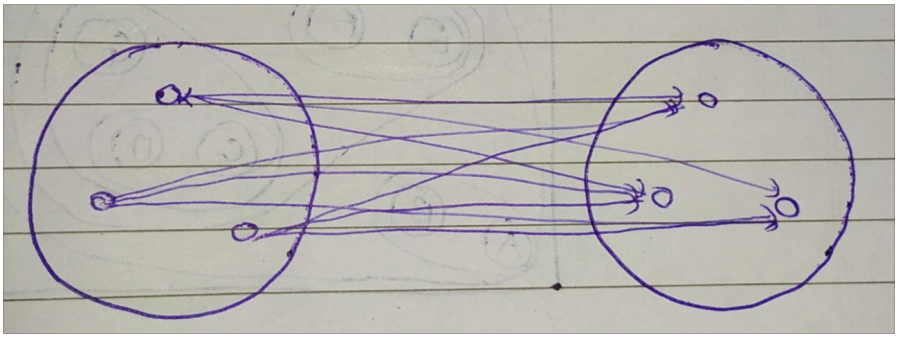
1. **Single linkage** we merge in each step the two clusters whose two closest members have the smallest distance



2. **Complete linkage** we merge in each step the two clusters whose merger has the smallest maximum pairwise distance



3. **Average linkage** returns the average of distances between all pairs of data point . It merges in each iteration the pair of clusters with the highest cohesion.



4. **Centroid linkage** returns the distances between between centroid of Clusters.