Question 1: Assignment Summary

The problem statement- HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries. Now the CEO of the NGO needs to decide how to use $ 10 million 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. So here I come as a Data Analyst and my job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country.

The solution methodology-

1. **Understand the Data** – Here I checked the basic information of data like null values, data type, columns, etc.
2. **Preparing the Dataset**- Here I converted health, import and export columns into a uniform form, so that analysis can be performed easily.
3. **Performing EDA**- I performed EDA of data and visualized the dataset in order to have better understanding of data set.
4. **Treating Outliers**- During EDA, found that various columns have outliers, so I treated the outliers.
5. **Rescaling Dataset-** To standardise the values, rescaling was performed on dataset.
6. **Hopkins Check-** Hopkins check is performed 10 times on dataset and the value was above 80% in all the cases.
7. **K-means Clustering-** Elbow curve analysis and Silhouette score was performed to determine the value of 'k'. The value of k was obtained as 2.
8. **Hierarchical Clustering-** Hierarchical clustering was performed, both Single Linkage and Complete Linkage was performed, and here also the value of k was obtained as 2.
9. **Final Conclusion-** Both the clustering suggests same top 5 countries that require aid. Thus, we draw conclusion that **Burundi, Liberia, Congo, Niger and Sierra Leone** are top 5 countries that require AID from HELP International.

Question 2: Clustering

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

Ans. k-means is method of cluster analysis using a pre-specified no. of clusters. It requires advance knowledge of 'K'.

Hierarchical clustering also known as hierarchical cluster analysis (HCA) is also a method of cluster analysis which seeks to build a hierarchy of clusters without having fixed number of clusters.

K-means Clustering	Hierarchical Clustering
k-means, 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.	Hierarchical methods can be either divisive or agglomerative.

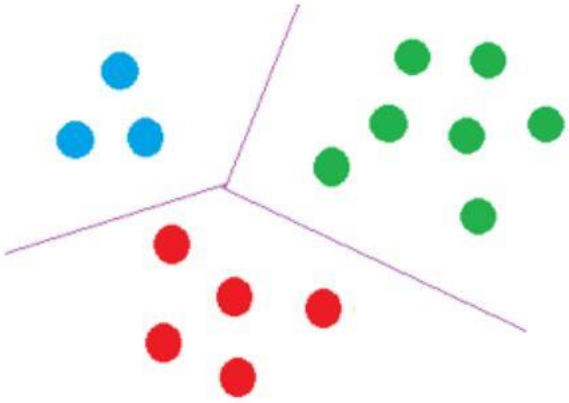
K Means clustering needed advance knowledge of K i.e., no. of clusters one wants to divide your data.	In hierarchical clustering one can stop at any number of clusters, one finds appropriate by interpreting the dendrogram.
One can use median or mean as a cluster centre to represent each cluster.	Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
Methods used are normally less computationally intensive and are suited with very large datasets.	Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy.
In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm many times may differ.	In Hierarchical Clustering, results are reproducible in Hierarchical clustering
K- means clustering a simply a 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 is found to work well when the structure of the clusters is hyper spherical (like circle in 2D, sphere in 3D).	Hierarchical clustering doesn't work as well as, k means when the shape of the clusters is hyper spherical.

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

Ans.

1) **Initialize cluster centers:** We randomly pick three points A1, A2 and A3 and assign them as cluster centers.
2) **Assign observations to the closest cluster center:** Once we have these cluster centers, we can assign each point to the clusters based on the minimum distance to the cluster center. For a point, compute its distance to A1, A2 and A3, respectively. By comparing shortest euclidean distance, we assign point to the nearest cluster. We then move to nest point and follow the same procedure.
3) **Revise cluster centers as mean of assigned observations:** Now we've assigned all the points based on which cluster center they were closest to. Next, we need to update the cluster centers based on the points assigned to them. For instance, we can find the center mass of a cluster by summing over all the points in that cluster and dividing by the total number of points. The result will be a new center. Similarly, we can find the new centers for other clusters.
4) **Repeat step 2 and step 3 until convergence:** The last step of k-means is just to repeat the above two steps. We keep on iterating between assigning points to cluster centers, and

updating the cluster centers until convergence. Finally, we may get a solution like shown below:

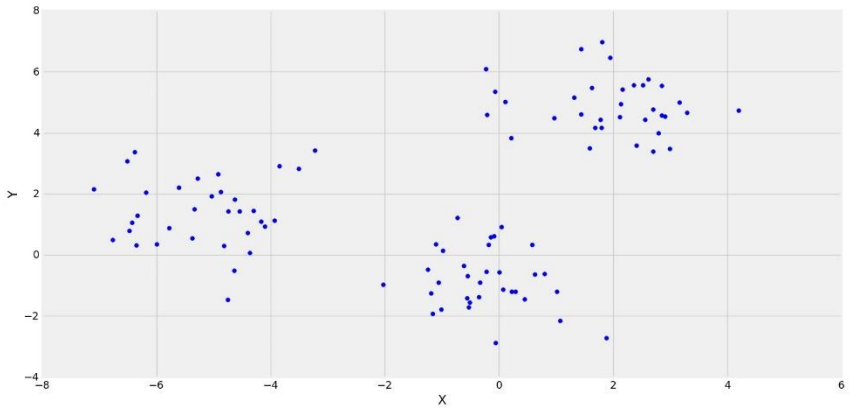


c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

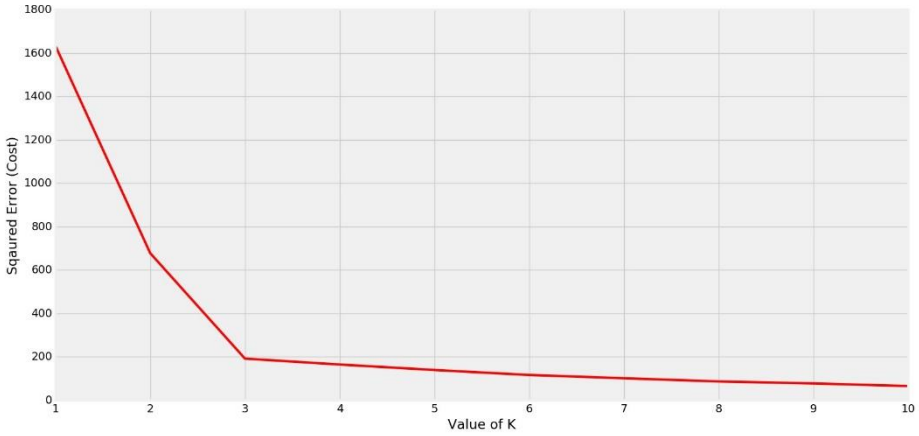
Ans. The value of 'k' is chosen on the basis of statistical and business aspect:

Statistical aspect: The elbow curve and Silhouette score are statistical tools which helps in determining the value of 'k'**:**

Elbow curve: - There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.



In the above figure, its clearly observed that the distribution of points are forming 3 clusters. Now, let's see the plot for the squared error(Cost) for different values of K.



Silhouette score- The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters. The Silhouette Value s(i) for each data point i is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

Business aspect: The selection of value of 'k' also depends upon business for which cluster is being performed. We need to understand the business model and need of client before determining the value of k.

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

Ans. All such distance-based algorithms are affected by the scale of the variables. Consider your data has an age variable which tells about the age of a person in years and an income variable which tells the monthly income of the person in rupees:

ID	Age	Income(rupees)
1	25	80,000
2	30	100,000
3	40	90,000
4	30	50,000
5	40	110,000

Here the Age of the person ranges from 25 to 40 whereas the income variable ranges from 50,000 to 110,000. Let's now try to find the similarity between observation 1 and 2. The most common way is to calculate the Euclidean distance and remember that smaller this distance closer will be the points and hence they will be more similar to each other. Just to recall, Euclidean distance is given by:

$$D = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

Here,
n = number of variables
p1, p2,p3,… = features of first point
q1, q2,q3,… = features of second point
The Euclidean distance between observation 1 and 2 will be given as:
Euclidean Distance = [(100000–80000) ^2 + (30–25) ^2] ^ (1/2)
which will come out to be around ***20000.000625***. It can be noted here that the high magnitude of income affected the distance between the two points. This will impact the performance of all distance-based model as it will give higher weightage to variables which have higher magnitude (income in this case).
We do not want our algorithm to be affected by the magnitude of these variables. The algorithm should not be biased towards variables with higher magnitude. To overcome this problem, we can bring down all the variables to the same scale. One of the most common technique to do so is normalization where we calculate the mean and standard deviation of the variable. Then for each observation, we subtract the mean and then divide by the standard deviation of that variable:

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

Apart from normalization, there are other methods too to bring down all the variables to the same scale. For example: Min-Max Scaling. Here the scaling is done using the following formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

For now, we will be focusing on normalization. You can try min-max scaling as well. Let's see how normalization can bring down these variables to same scale and hence improve the performance of these distance-based algorithms. If we normalize the above data, it will look like:

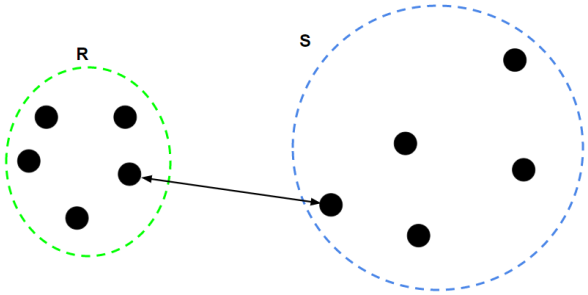
ID	Age	Income(rupees)
1	-1.192	-0.260
2	-0.447	0.608
3	1.043	0.173
4	-0.447	-1.563
5	1.043	1.042

Let's again calculate the Euclidean distance between observation 1 and 2:
Euclidean Distance = [(0.608+0.260) ^2 + (-0.447+1.192) ^2] ^ (1/2)
This time the distance is around ***1.1438***. We can clearly see that the distance is not biased towards the income variable. It is now giving similar weightage to both the variables. Hence, it is always advisable to bring all the features to the same scale for applying distance-based algorithms like KNN or K-Means.

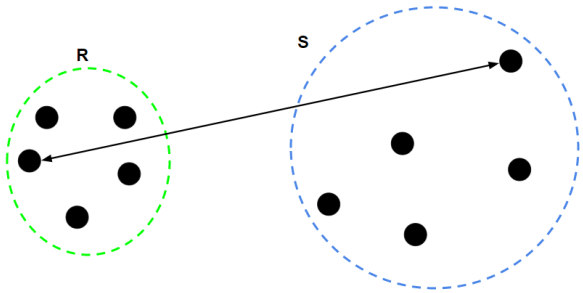
e) Explain the different linkages used in Hierarchical Clustering.
Ans. The process of Hierarchical Clustering involves either clustering sub-clusters (data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster

into smaller subclusters 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: -

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.