Q1. Assignment Summary

Problem Statement:

HELP International, a humanitarian NGO is committed to fight poverty and provide people of backward countries basic amenities and relief.

It has been able to raise funds of $10 Million. We need to decide which countries are in direst need of aid

A dataset of 167 countries has been provided along with their socio economic and health factors.

Steps:

1. Data Quality Check: The dataset did not contain any null values. Attributes like exports, health, imports were given in % of gdpp, so converted them to absolute values

2. Univariate/ Bivariate Analysis: Sorted top 10 attribute values and plotted a graph with respect to countries. Plotted a correlation matrix for all attribute values.

3. Outlier treatment: Performed lower capping for child_mort, inflation and total_ferti and upper capping for all other variables like gdpp, health etc as these countries don't lie in the consideration of dire need

4. Scaling: Performed standard scaling on cleaned, outlier removed data

5. Checked the cluster Tendency(Hopkins Test): The Hopkins score was coming to be >85% which is a good measure

6. Find best value of k using SSD, Silhouette Score: Performed SSd, Silhoutte analysis and decided to take k =4

7. Using final value of k perform k means analysis

8. Visualise cluster using scatter plots: Plotted a scatter plot for gdpp, income, child_mort

9. Perform cluster profiling: Compared gdpp, income, child_mort values for all the clusters formed

10. Hierarchical Clustering: Constructed a dendrogram for both single and complete linkage. Decided to go with complete linkage as it was much clear. Chose value of k = 3

11. Listing out the target countries: The top 5 target countries came out to be same in both kmeans and Hierarchical clustering. The top 5 countries are:

- Congo, Dem. Rep.

- Liberia

- Burundi

- Niger

- Central African Republic

Q2

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

Ans:

KMeans Clustering	Hierarchical Clustering

1. We need to know the number of clusters before we proceed with the algorithm.	1.We can stop at any number of clusters and can find the appropriate one by interpreting the dendrogram.
2.Use median or mode as cluster centre	2.Start with n clusters and sequentially combine similar clusters until one cluster is obtained (agglomerative)
3.Less computationally intensive	3.High computationally intensive.
4. Clusters formed are circular, spherical in nature(hyper spherical)	4.Clusters formed can be non hyper spherical

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

K – Means Algorithm is as follows:

Ø Start by choosing k initial points (k = number of clusters we want)

Ø Initial choice of centre clusters is random

Ø Assignment: We allocate each data point to the nearest cluster use Euclidean Distance

Ø Optimization step: Re Compute cluster centre

Keep iterating the process of assignment and optimization till convergence.

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

To find optimal number of clusters (statistical way) we use two techniques:
Silhouette Score
Elbow curve method

Elbow Curve:

- Compute the clustering algorithm for different values of k.
- In this method we calculate the sum of squared distances of a point to its nearest cluster centre
- After that a curve is plotted of ssd to the number of clusters.
- We see that when the number of clusters are increased the 'ssd' value go down
- We check that at what value of k the rate of drop is significant.
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Silhouette Score

silhouette score = (p-q)/max(p,q)

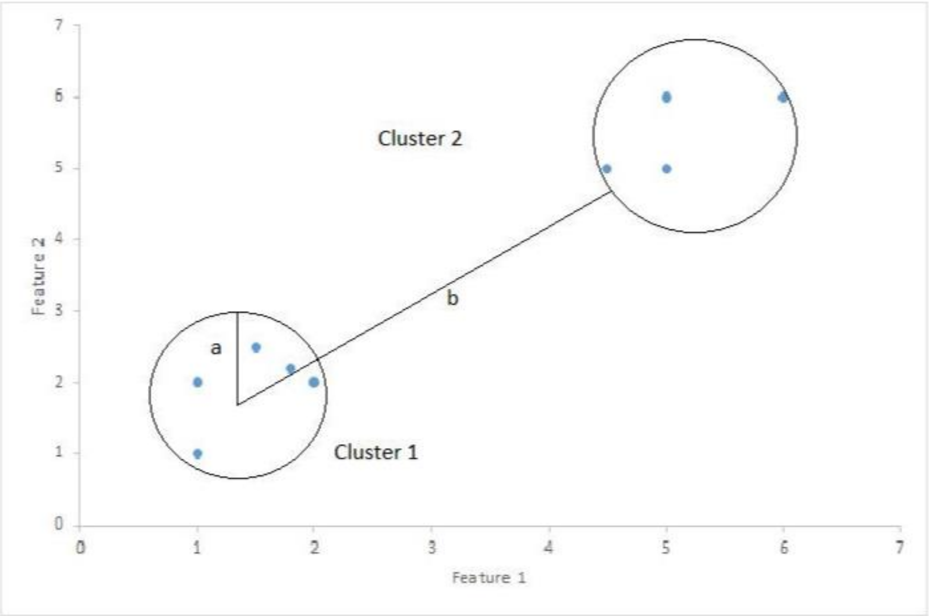
p is the mean distance to the points in the nearest cluster that the data point is not a part of

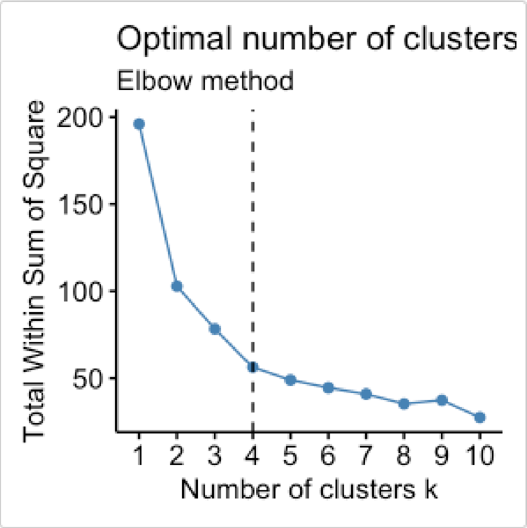
q is the mean intra-cluster distance to all the points in its own cluster.

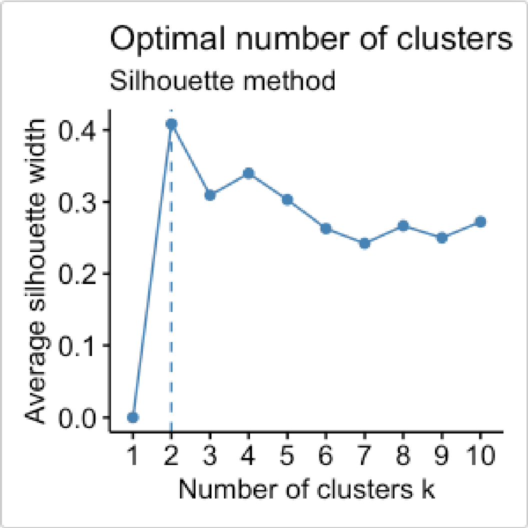
* The value of the silhouette score range lies between -1 to 1.

* 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.







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

Standardization prevents variables with larger scale from dominating on how clusters are defined. It allows all variables in an algorithm to be considered with equal importance.
Since k means is a distance based algorithm, if we use exact values before scaling, the distances would appear much larger leading to incorrect clusters in most cases.

Income values (15000 vs 20000) would show up distance as 5000 than age (15 vs 20) which is 5 even though the percentage difference is same. Thus derivatives tend to align along directions with higher variance leading to poorer convergence.

d) Explain the different linkages used in Hierarchical Clustering.

Three types of linkages are there:

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.