Question 1: Assignment Summary

Problem Statement

The NGO **HELP International - an international humanitarian NGO**, needs to decide how to use the funds of **$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**.

Task - To suggest the countries which are in dire need of aid, which the CEO needs to focus, by **categorizing** the countries using some socio-economic and health factors that determine the overall development of the country.

Solution Methodology

I divided my assignment in the following steps: -

1) Data Inspection and Pre-Processing
2) Exploratory Data Analysis
3) Outlier treatment
4) Checking cluster's tendency - Hopkin's test
5) Scaling the data
6) K means clustering and Hierarchical Clustering
    a. Cluster visualization
    b. Cluster Profiling
7) Results based on both clustering techniques

Data Inspection and Pre-Processing

- Inspected data in terms of shape of dataset, null values, statistical values of numeric variables
- For pre-processing, some columns like exports, health and imports were expressed as percentage of gdpp. Hence, I converted them to absolute values.

Exploratory Data Analysis

- Univariate Analysis
    - Plotted histogram for all the variables and noted down observations
- Bivariate Analysis
    - Plotted pair plot between all variables and identified few relationships amongst them
- Multivariate Analysis
    - Plotted a correlation heatmap and observed some significant correlations between some variables

Outlier Treatment

- According to our business use case, treated the outliers accordingly
    - For example, for columns like child_mort, inflation and total_fer, we should not treat the higher range outliers as it is critical to our business use case. These columns, if the values are high, suggest that the countries are in dire need of aid.

    - For other columns, we should not treat the lower range outliers for the same reason of them being critical to our business use case. These columns, if the values are low, suggest that the countries are in dire need of aid.

Checking cluster's tendency using Hopkins test

- After the data was cleaned, outliers treated and EDA performed, it was imperative to check if the dataset really had clusters. I used Hopkins test to determine how dissimilar is our dataset from randomly sampled data (non-clusterable data). A value in the range of 90-95% showed that our data was clusterable and very dissimilar than random sampled data

Scaling Data

Since the performance of K means model is impacted as it will give higher weightage to variables which have higher magnitude it was imperative to scale the data first.

K means clustering

- The first step is to decide on the best k value or the number of clusters. This was done using Silhouette scores and SSD. Both gave a best value of k=3
- K means clustering was done for both k=3 and k=4(to see if better clusters are obtained from business point of view)
- Clusters were visualised based on 3 important variables(specified in the task) gdpp, child_mort and income.
- It was decided that k=3 gave the best clusters.
- For cluster profiling, the cluster with lowest gdpp, highest child_mort and lowest income was the ideal cluster as the countries in this cluster were in dire need of aid
- Top 5 countries from this cluster was selected.

Hierarchical clustering

- Single linkage and Complete linkage were used. Complete linkage gave a better interpretable dendrogram.
- K=2 was chosen as the best k value by interpreting the dendrogram and also after analysing cluster labels for k=3 and k=4 which had only 1 point in last cluster(1 point doesn't constitute a cluster)
- For cluster profiling, the cluster with lowest gdpp, highest child_mort and lowest income was the ideal cluster as the countries in this cluster were in dire need of aid
- Top 5 countries from this cluster was selected.

Results

- Both clustering techniques gave same list of countries in same order which are in dire need of aid. Hence they were in the final list to be shown to CEO of the NGO.

Question 2: - Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.
b) Briefly explain the steps of the K-means clustering algorithm.
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
d) Explain the necessity for scaling/standardisation before performing Clustering.
e) Explain the different linkages used in Hierarchical Clustering.

a)

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 needs advanced 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 can find an appropriate threshold to cut the dendrogram by interpreting it.
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 starts with a random choice of clusters, the results produced by running the algorithm many times may differ.	In Hierarchical Clustering, results are reproducible
K- means clustering is 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.

b) **Algorithmic steps for k-means clustering:**
Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} 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 the 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, ‘ci’ 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.

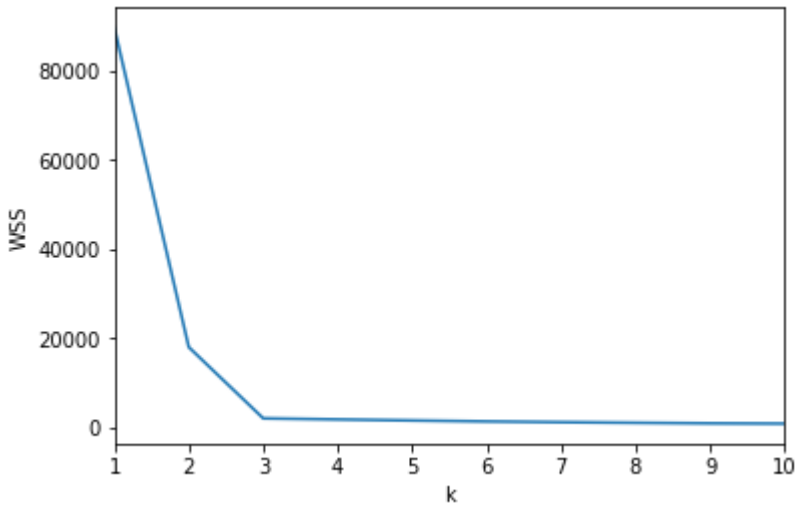
c) Statistical Sense

Two methods that can be useful to find this mysterious k in K-Means.
These methods are:

a. The Elbow Method
b. The Silhouette Method

The Elbow method
This is probably the most well-known method for determining the optimal number of clusters. It is also a bit naive in its approach.
Within-Cluster-Sum of Squared Errors (WSS) is calculated for **different values of k** and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an **elbow.**
Within-Cluster-Sum of Squared Errors calculation

1) The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.
2) The WSS score is the sum of these Squared Errors for all the points.
3) Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.



As expected, the **plot looks like an arm with a clear elbow at k = 3.**

The Silhouette method
The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

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.

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

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

Note: s(i) is defined to be equal to zero if i is the only point in the cluster. This is to prevent the number of clusters from increasing significantly with many single-point clusters.

Here, a(i) is the measure of similarity of the point i to its own cluster. It is measured as the average distance of i from other points in the cluster.

For each data point $i \in C_i$ (data point i in the cluster C_i), let

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

Similarly, b(i) is the measure of dissimilarity of i from points in other clusters.

For each data point $i \in C_i$, we now define

$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j)$$

d(i, j) is the distance between points i and j. Generally, Euclidean Distance is used as the distance metric.

Business sense
Sometimes, if we employ statistical means to derive the best value of k, it may be possible that in business sense there may not be so many clusters as to what the statistical methods like silhouette score or elbow method gives. For example, if we arrive at the k value of 12 from the dataset using silhouette and elbow methods, maybe later the product or marketing team can suggest that only 11 clusters make sense. Hence the business/industry knowledge is critical in deciding the number of clusters too.

d) All the distance-based clustering algorithms like K-means clustering are affected by the scale of the variables.
For example, 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 like each other.

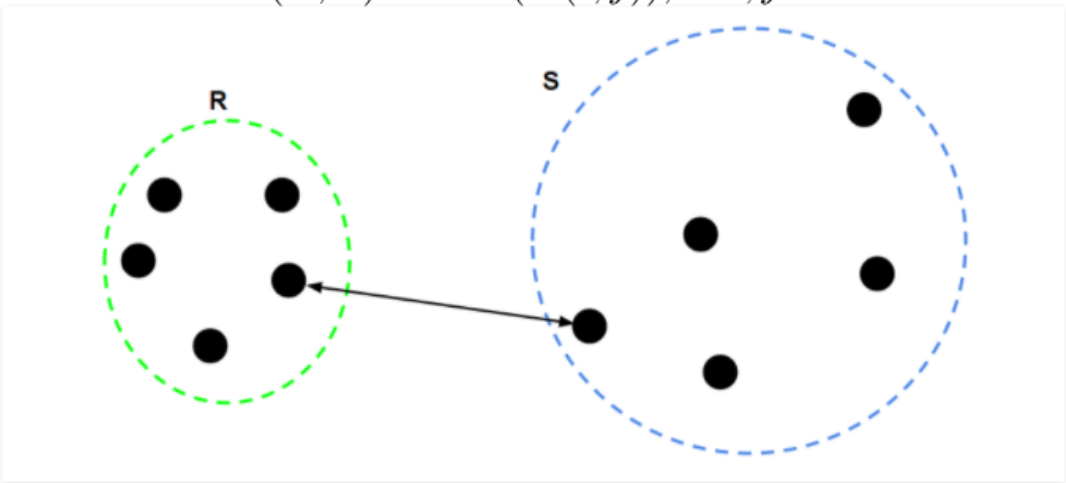
Euclidean Distance = [(100000–80000)^2 + (30–25)^2]^(1/2)

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 clustering 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 by scaling/standardization.

e) The types of linkages in Hierarchical Clustering are as follows:

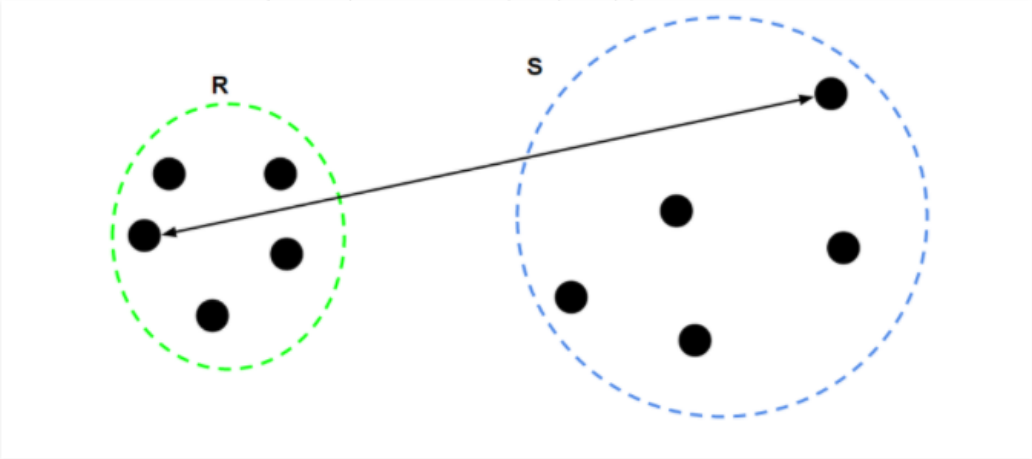
Single Linkage: For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.

$$L(R,S) = min(D(i,j)), i \epsilon R, j \epsilon S$$



Complete Linkage: For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.

$$L(R,S) = max(D(i,j)), i \epsilon R, j \epsilon S$$



Average Linkage: For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.

$$L(R,S) = \frac{1}{n_R + n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(i,j), i \epsilon R, j \epsilon S$$

where

n_R - Number of data-points in R

n_S - Number of data-points in S