Question 1: Assignment Summary

Ans: <u>Problem Statement:</u>

IDEAL: CEO of HELP International NGO would be able to decide how to use raised money strategically and effectively by getting the countries who are in dire need of aid.

REALITY: In reality we are having data for 167 companies with factors affecting it's development. From the raw dataset it will be difficult to decide that which courtiers are in dire need of aid.

CONSEQUENCES: May be raised fund will go to the countries not in dire need of aid and countries in dire need of aid will get skip.

PROPOSAL: Clustering will be done to categorize the companies in different cluster based on the country development factors (Variables)

<u>Solution methodology:</u> Below are the steps /methodology used to drive the solution

Step1: Data understanding: To understand the data have read the csv and inquired about the shape of data and type of variables. For 167 countries data is available with factors '**child_mort', 'exports', 'health', 'imports', 'income', 'inflation', 'life_expec', 'total_fer', 'gdpp'**

Step2: Data Cleaning: Data cleaning was no needed as the data type was appropriate for all available variables and also no missing values are there

Step 3: Data Visualisation: Data visualisation has been done by using boxplot (mainly to get outliers), bar graph (to check distribution for variable), scatter plot.

Box plot and histogram are used for univariate analysis. From analysis its visible like variable are having outliers which need to be treated and are not distributed equally.

Scatter plot and correlation (Heat map) has been used for Bivariate analysis, which conclude that child_mort is directly corelated to inflation and total_fer and indirectly related to health, gdpp, import, export, income, life_expec

Step 4: Outliers treatment: As in univariate analysis it's clear that variables are having outliers. For outlier treatment below caping has been performed

- Child_mort lower outliers has been capped with 01 percentile and have left higher as usual because as per the problem statement may be those need aid.
- Other variables higher outliers have been capped with 99 percentiles.

Step 5: Hopkins score: Have checked Hopkins score for the data to check whether clustering is possible or not. AS Hopkins score is >80 its good, data is good for clustering

Step 6: Clustering: Both K-means and hierarchal clustering has been performed to do the clustering.

For K-means to get value of k, elbow curve and silhouette score has been calculated and based on that 3 cluster has been created.

For Hierarchal also from dendrogram, value of k 3 has been decided and for same clustering has been performed.

From both of the method clustering has been done based on below criteria:

- Cluster 1: high export, high health, high imports, high income, low inflation, high life_expec, low total_fer, high gdpp, low child_mort
- Cluster 2: low export, low health, low imports, low income, high inflation, low life_expec, high total_fer, low gdpp, high child_mort
- Cluster 3: avg export, avg health, avg imports, avg income, avg inflation, avg life_expec, avg total_fer, avg gdpp, avg child_mort

Step 6: Conclusion: In terms of no countries available in all clusters created via K-Means and Hierarchical Clustering, K-means is having relative balanced no of records. Hence for final approach will consider K-means algorithm

Step 7: Summary/Solution: Countries related to cluster having high child_mort will in dire need for aid. Below are top 5 companies which are in dire need of aid as per K-mean clustering

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
Haiti	208.0	101.286	45.7442	428.314	1500.0	5.45	32.1	3.3300	662.0	0
Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.20	55.0	5.2000	399.0	0
Chad	150.0	330.096	40.6341	390.195	1930.0	6.39	56.5	6.5636	897.0	0
Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.01	47.5	5.2100	446.0	0
Mali	137.0	161.424	35.2584	248.508	1870.0	4.37	59.5	6.5500	708.0	0

Question 2: Clustering

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

Ans: Comparison b/w K-means and Hierarchal Clustering:

Properties	K-means	Hierarchical Clustering
Definition	K-means clustering generates specific number of disjoint, flat clusters	Hierarchical clustering method construct a hierarchy of clustering, not just a single partition of objects
Performance	K-means clustering is usually more efficient run-time wise	Hierarchical clustering can be slow (has to make several merge/split decisions)

Sensitive to noise	K-means is very sensitive to noise in the dataset	It is less sensitive to noise in dataset
Cluster	K-means clustering requires prior knowledge of k	Hierarchical clustering does not require prior knowledge of k. It can be determined by Dendrogram
Quality	K-Means algorithm shows less quality	Hierarchical algorithm shows more quality
Data set	K-means algorithm is good for large data set because time complexity of K-means is linear O(n)	Hierarchical is good for small data set because time complexity of K-means is quadratic $O(n^2)$
Cluster result	In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ	Results are reproducible in Hierarchical clustering.

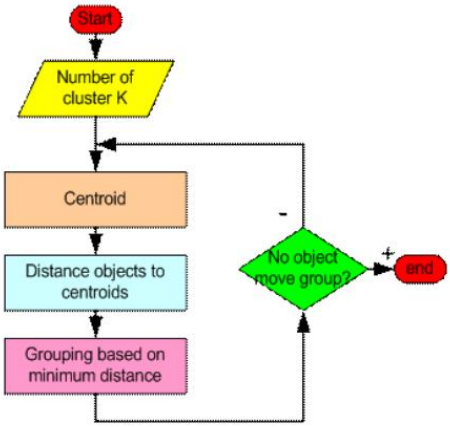
b) Briefly explain the steps of the K-means clustering algorithm.

Ans: <u>Steps of K-mean clustering algorithm:</u> Below are the steps which are getting followed in K-mean clustering

1) Randomly select 'c' cluster centre
2) Calculate the distance between each data point and cluster.
3) Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centre
4) Recalculate the new cluster centre by using below formula:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

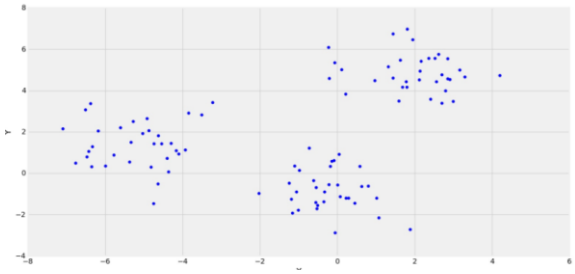
   Where Ci represents no of data points in ith cluster
5) Recalculate the distance between each data point and new obtained data centres
6) If no data point was reassigned then stop otherwise repeat from step 3.



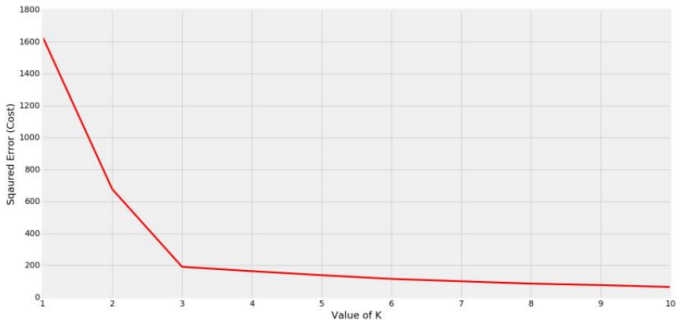
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Ans: Statistical way to choose k:

- **elbow method: elbow method** 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.



Clearly the elbow is forming at K=3. So the optimal value will be 3 for performing K-Means.

- **Silhouette analysis:** Silhouette coefficient is a measure of how similar a data point is to its own cluster (cohesion) compared to other clusters (separation).

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

  Where p is the mean distance to the point in the nearest cluster that the datapoint is not the part of

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

  The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

  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. 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.

Business Aspect to select k for K-mean clustering: We can get the value of k from above 2 approaches but with this business aspect is more important as how many cluster business needed or how it will cater problem statement. Based on statistical and business aspect value of k can be decided.

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

Ans: Standardization refers to the process of rescaling the values of the variables in your data set so they share a common scale, hence prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

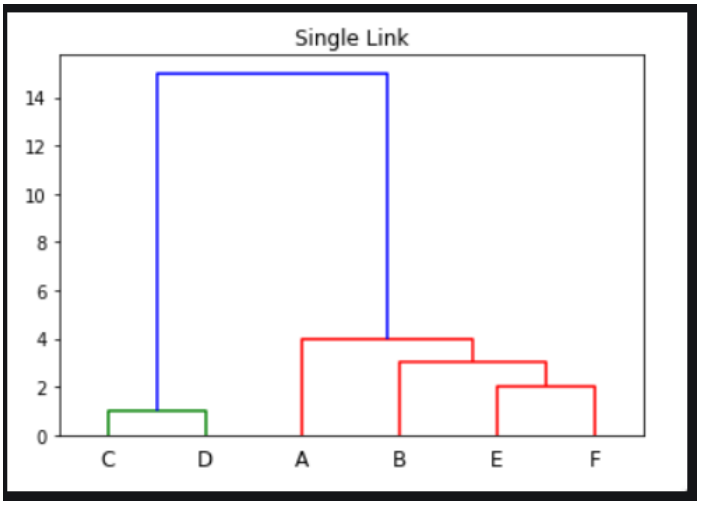
When we are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.
Since, clustering techniques use **Euclidean Distance** to form the cohorts, it will be wise e.g. to scale the variables having heights in meters and weights in KGs before calculating the distance.

e) Explain the different linkages used in Hierarchical Clustering.

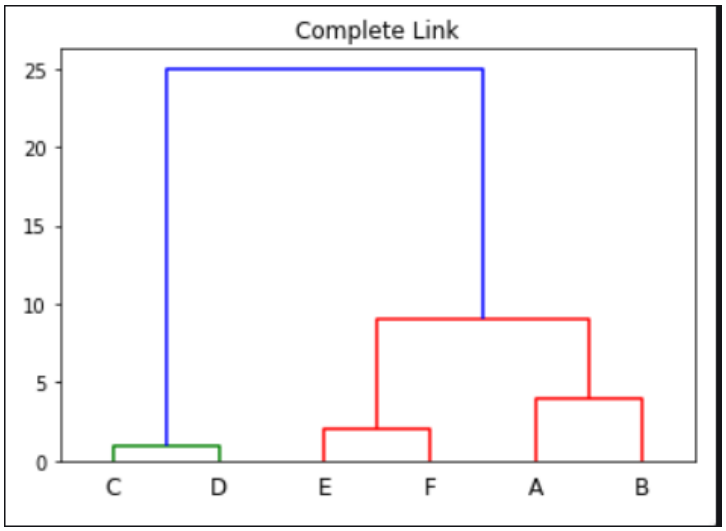
Ans: Hierarchical Clustering: Hierarchical clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster. A hierarchical clustering is represented as a dendrogram.

There are three type of linkage used in Hierarchal Clustering

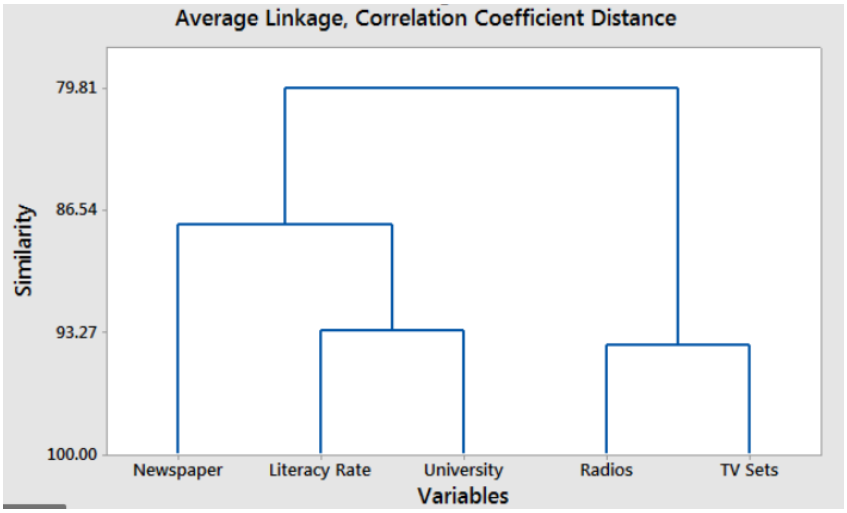
- **Single Linkage:** Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters



- **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.



Usually, single linkage type will produce dendrograms which are not structured properly, whereas complete or average linkage will produce clusters which have a proper tree-like structure.