1. **Assignment Summary (Clustering of Countries):**

   Ans: CLUSTERING ASSIGNMENT of countries Identify top - 5 countries that are direst need of AID

   Methodology for solving problem:

   Importing .csv Files: country-data.csv

   - Inspect the data frame / Check the structure of the data
   - converting the %age values "exports, health, imports " into actual values
   - df.shape , df.info ( ), df.describe()

   Data Quality Check and Missing values/Cleaning the Data: Inspect Null values (both in columns and rows of data frame)

   Data Visualisation/Perform EDA to understand various variables

   From the plots we can observe following: cluster profiling is possible on child_mort, inflation, GDPP, exports,imports, Income , life_expec, Total_fer

Outlier Treatment: From plots we can observe the following:

Upper outliers exist for child_mort, exports, imports, inflation, health, income, total_fer, and GDPP Lower Outliers exist in Life_expec As we need to find the direst need of AID, so we should not treat the upper outliers of Child_mort and Inflation. For analysis purpose we are treating the upper outliers using Capping We can observe that all the columns having upper outliers are capped to 99% except Child Mort and Inflation

Hopkins stats run: 100 times The Mean Value of Hopkins is: 0.91, Consider Data is very good for Cluster

Scaling is performed on the columns

Find the K Value used for analysis:

- Silhouette Score
- Elbow Curve

From the plots of we can observe the following:

- Silhouette Score for n_clusters 3 (excluding n_clusters = 2) is high as compared to others
- from Elbow curve we can observe the bend at point 3 (n_clusters = 3)

As 3 is satified in both the plots , we are considering the no of culters as 3 for analysis

KMean Clustering

Cluster Profiling

Visualization: GDPP - Income, Income - Child_mort, GDPP - Child_mort

From the above plots we can observe the following:

- Inflation effect on child_mort and Life_expec :
  - except on data point there in no much impact on inflation
- GDPP effect on Child_mort and Life_expec:
  - higher the GDPP lower the child_mort and higher life_expec
- Higher spending on health there is a lower child_mort and Higher Life_expec
- Higher the total_fer, lower in life_expec and higher child_mort

Hierarchical Clustering

- Single Linkage
- Complete Linkage

The following are observed from the above clustring Profile and Value counts of

Single Linkage *:*

- Only one cluster is dominating other clusters i.e. cluster label 0
- The total count of cluster label 0 is 165
- other cluster label is having one count each
- Hence Sinlge Linkage Cluster is not used in the further analysis
-

Complete Linkage

Similar points can be observed as seen in KMeans Clusters:

- Inflation effect on child mort and Life_expec :
  - except on data point there in no much impact on inflation
- GDPP effect on Child mort and Life_expec:
  - higher the GDPP lower the child mort and higher life_expec

- Higher spending on health there is a lower child mort and Higher Life_expec
- Higher the total_fer, lower in life_expec and higher child mort

Comparing the KMeans and Hierarchical Clustering

From the above comparison, Both clusters show the same top - 5 countries. They are:(from highest child mort to least)

- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali

Conslusion:

- Inflation effect on child mort and Life_expec :
    - Except on data point there in no much impact on inflation
- GDPP effect on Child mort and Life_expec:
    - Higher the GDPP lower the child mort and higher life_expec
- Higher spending on health there is a lower child mort and Higher Life_expec
- Higher the total_fer, lower in life_expec and higher child mort
- The top 5 Countries which are in need of direst AID are:
    - Haiti
    - Sierra Leone
    - Chad
    - Central African Republic
    - Mali
- The most of African countries are in top - 5 which are in need of direst AID¶

Question 2: Clustering

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

Ans: The following are the comparison between K-Means Cluster and Hierarchical Cluster:

1. Hierarchical clustering can’t handle big data well but K Means clustering can. This is

because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).

2. In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
3. K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
4. K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.

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

Ans:The algorithm for K-means algorithm is as follows:

1. Clusters the data into k groups where k is predefined.
2. Select k points at random as cluster centers.
3. Assign objects to their closest cluster center according to the Euclidean distance function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

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 following are mostly used to optimize the value of k:

1. Silhouette method: Silhouette refers to a method of interpretation and validation of consistency within clusters of dataSilhouette analysis can be used to study the separation distance between the resulting clusters. 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.
2. Elbow Curve Method:It involves running the algorithm multiple times over a loop,

with an increasing number of cluster choice and then plotting a clustering score as a function of the number of clusters. The curve plotted in the shape of elbow and hence its called as elbow curve (use inertia to determine the value to plot against each k value).

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

Ans:All such distance based algorithms are affected by the scale of the variables.

Let consider the 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:

Emp-ID	Age	Income (Rs)
E001	50	110,000
E002	25	80,000
E003	40	90,000
E004	45	50,000

Here the Age of the person ranges from 25 to 50 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:

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

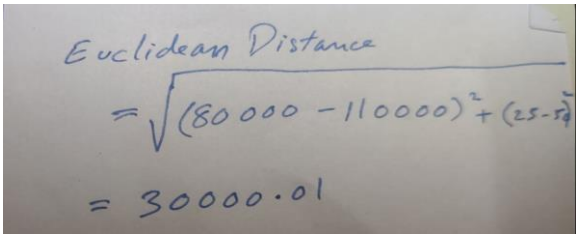
Here,

n = number of variables

p1,p2,p3,… = features of first point

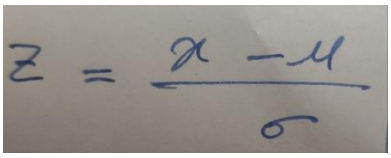
q1,q2,q3,… = features of second point

The Euclidean distance between observation 1 and 2 will be given as:

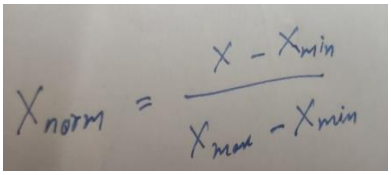


It can be noted here that the high magnitude of income affected the distance between the two points. This will impact the performance of all distance based model as it will give higher weightage to variables which have higher magnitude (income in this case).

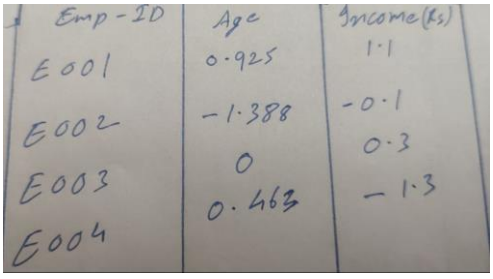
We do not want our algorithm to be affected by the magnitude of these variables. The algorithm should not be biased towards variables with higher magnitude. To overcome this problem, we can bring down all the variables to the same scale. One of the most common technique to do so is normalization where 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:



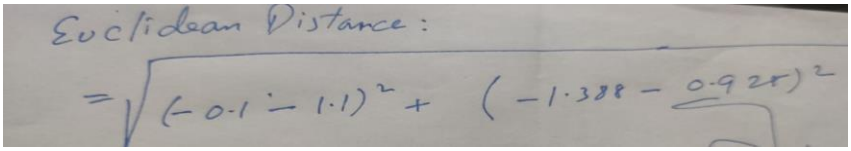
Apart from normalization, there are other methods too to bring down all the variables to the same scale. For example: Min-Max Scaling. Here the scaling is done using the following formula:



For now, we will be focusing on normalization. Let's see how normalization can bring down these variables to same scale and hence improve the performance of these distance based algorithms.



Let's again calculate the Euclidean distance between observation 1 and 2:



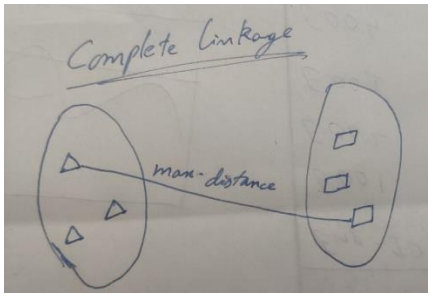
This time the distance is around 2.565. We can clearly see that the distance is not biased towards the income variable. It is now giving similar weightage to both the variables. Hence, it is always advisable to bring all the features to the same scale for applying distance based algorithms like K-Means.

e) Explain the different linkages used in Hierarchical Clustering.

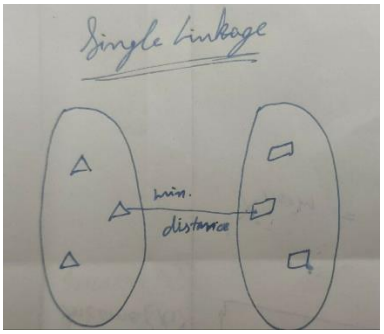
Ans:

There are several ways to measure the distance between clusters in order to decide the rules for clustering, and they are often called Linkage Methods. Some of the common linkage methods are:

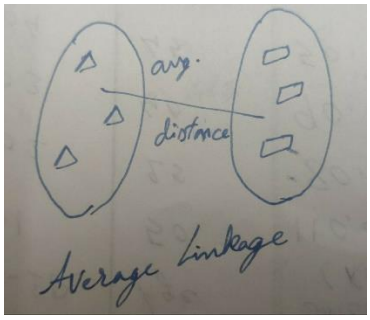
Complete-linkage: calculates the maximum distance between clusters before merging.



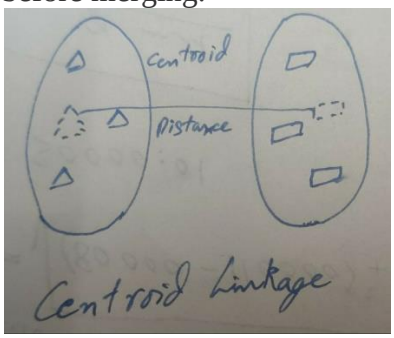
Single-linkage: calculates the minimum distance between the clusters before merging. This linkage may be used to detect high values in your dataset which may be outliers as they will be merged at the end.



Average-linkage: calculates the average distance between clusters before merging.



Centroid-linkage: finds centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.



The choice of linkage method entirely depends on businessproblem and there is no hard and fast method that will always give you good results. Different linkage methods lead to different clusters.