Question 1 – Assignment Summary:

HELP International – a NGO, helps the people of backward countries. As a Data Analyst, we need to identify the top 5 countries that are in need of immediate help.

For the countries with **low GDPP, low income and high child mortality rate**, NGO could fulfill their basic needs. Using K-Means and Hierarchical clustering, we need to identify such countries to proceed with their aiding task.

In **EDA**, using bivariate analysis and univariate analysis, find the relationship and the distribution of the data. Remove the outliers if any.

Before clustering, perform scaling and check the Hopkins statistics. Hopkins statistics is used to measure the cluster tendency. If the Hopkins statistics>0.8, gives good cluster tendency.

<u>K-Means Clustering</u>

Elbow-curve/SSD and Silhouette score, choose the optimal value for K (number of clusters). After running the K-Means clustering algorithm, each country will be assigned to a cluster. Using profiling identify the cluster that has low GDPP, low income and high child_mortality and find the countries present in that cluster, which needs help.

<u>Hierarchical Clustering:</u>

<u>Two methods:</u>

<u>Single Linkage Method:</u>

Distance between 2 clusters is defined as the shortest distance between points in the two clusters.

We will not able to visualise the **dendrogram** clearly.

<u>Complete linkage method:</u>

Distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters. The **dendrogram** can be visualised clearly. Choose the complete linkage method.

Using the K-value which we used for K-Means, cut the dendrogram and identify the cluster labels. Each country has been assigned to cluster. Using profiling identify the cluster that has low GDPP, low income and high child_mortality and find the countries present in that cluster, which need help.

Using both K-Means and Hierarchical clustering, we will be able to provide the list of top-5 countries to the NGO.

Question 2: Clustering

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

Sl.No.	K-means Clustering	Hierarchical Clustering
1	K- Means is a method of cluster analysis, where the value of K has to be predefined. (i.e.) we need to know the number of clusters in advance to divide the entire dataset.	Hierarchical cluster analysis (HCA) will build a cluster of analysis which is referred to as dendrogram (a tree like structure). We don't need to have the knowledge of number of clusters in advance.
2	This works in two steps. i. Assignment step ii. Optimisation step Assignment Step - Randomly assign K- centre points and find the Euclidean distance. Assign the datapoints to the cluster that are closest to the cluster. Optimisation step - Now recalculate the cluster centre (for each cluster) and find the new centroids. Continue the Assignment and Optimisation step till algorithm converges (two consecutive iterations have same cluster centres).	This works either in agglomerative methods or divisive methods Agglomerative approach (bottom-up) – begins with n number of clusters and sequentially combine the clusters that are similar to each other until only one cluster is obtained. Divisive method (top-down) - all datapoints starts in one cluster (parent cluster), which divides into smaller cluster and keeps on dividing till each cluster has a single object to represent.
3	K- Means algorithm does not need much space. It works in a manner as the plate has	Hierarchical clustering works by building up a tree like structure. This occupies more memory space.

	datapoints which circles around and created centroids and clusters the data. So, it occupies less space.	
4	It is a non-linear process	It is a linear process
5	K-means algorithm can be used if the size of the data is big	Hierarchical clustering can be used if the data size is small.
6	K-means algorithm works well if the shape of the cluster is hyper spherical (circle or sphere)	This does not work well when the shape of the cluster is hyper spherical

b) Briefly explain the steps of K-Mean Clustering Algorithm

Steps involved:

To explain the steps let us take the value of K= 2.

There are two steps involved in K-Means Algorithm

- Assignment Step
- Optimisation Step

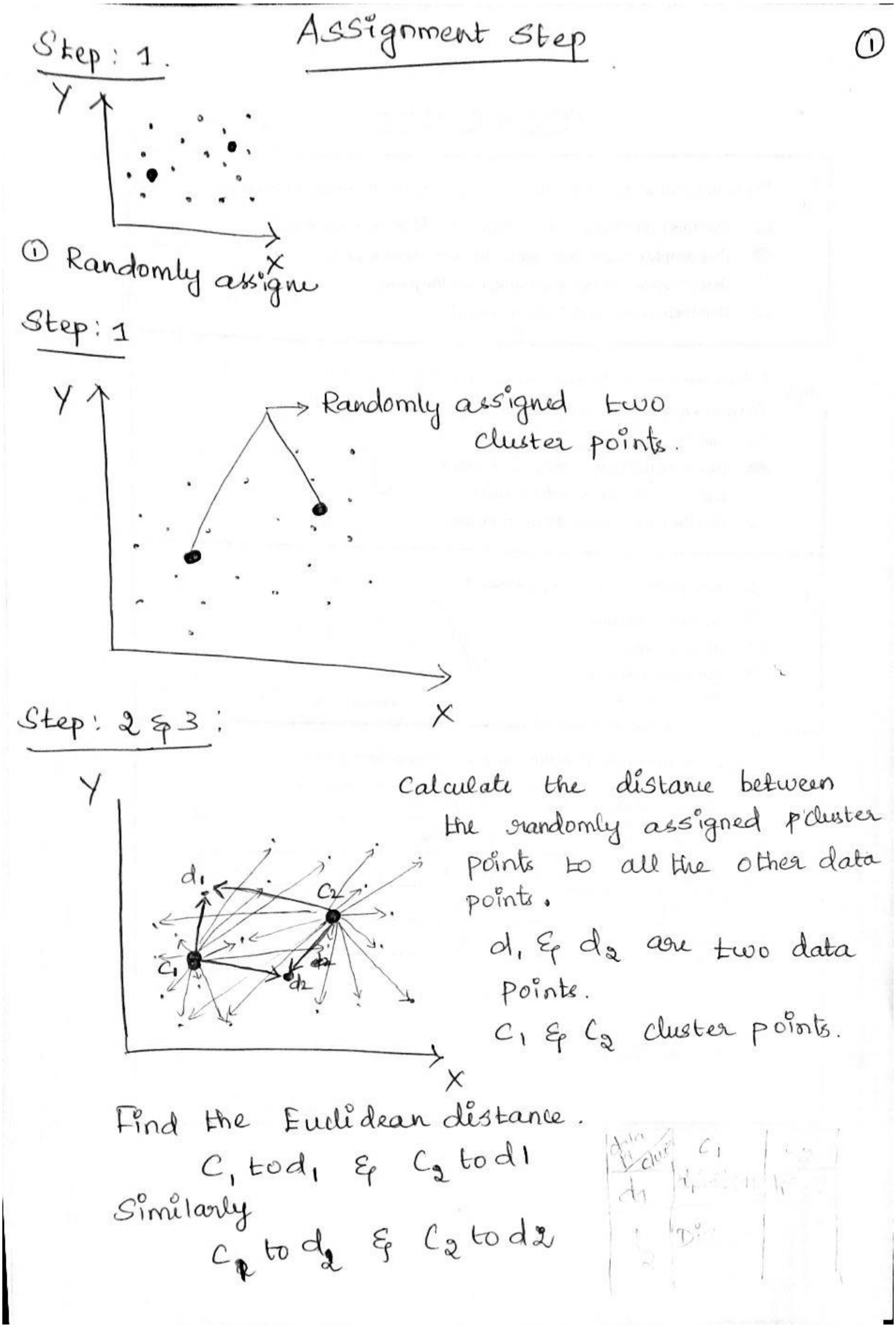
Assignment Step:

1. Randomly assign two centre points.
2. Now calculate the distance (Euclidean distance) between the between the two centre points and all the datapoints
3. After finding the distance assign the points to the cluster that has minimum distance.

Optimisation Step:

4. Now find the centroid for all the data points that belongs to cluster 1 and centroid for all the datapoints in the cluster 2. Now the we got new cluster centres.
5. Repeat the steps 2 to 4
6. Continue the process (step 5) until the algorithm converges (i.e.) two consecutive iterations have the same cluster.

Detailed Explanation:


Y
X
Y
X
Y
X
d_1
C_2
C_1
d_2

Assignment Step

①

Step: 1.

① Randomly assigne

Step: 1

Randomly assigned two cluster points.

Step: 2 & 3:

Calculate the distance between the randomly assigned p'cluster points to all the other data points.

d_1 & d_2 are two data points.

C_1 & C_2 cluster points.

Find the Euclidean distance.

C_1 to d_1 & C_2 to d1

Similarly

C_1 to d_2 & C_2 to d2

Data Points / clusters	Cluster C1	Cluster C2
d_1	d_1-dist_C1	d_1-dist_C_2
d_2	d_2-dist_C1	d_2-dist_C_2
⋮	⋮	⋮

Now, d_1-dist.C1 → dist. b/w. C_1 & d_1

d_1 - dist C2 → dist. b/w C_2 & d_1

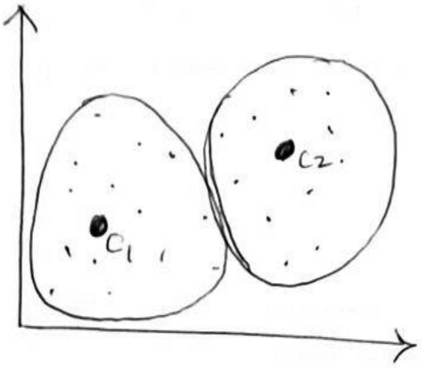
Same follows for $d_2, \ldots, d_n$.

Now, check ~~the~~ whether to which cluster we need to assign the datapoint. If the distance d_1-dist-C1 < d_1-dist-C_2 assign the data point d1 to cluster C1 else assign it to cluster C2.

Follow this for all the other datapoints and assign to cluster C1 or to C2 based on the euclidean distance (minimum).

Optimisation Step.

Step : 4 :



Now, find the centroid (new cluster point).

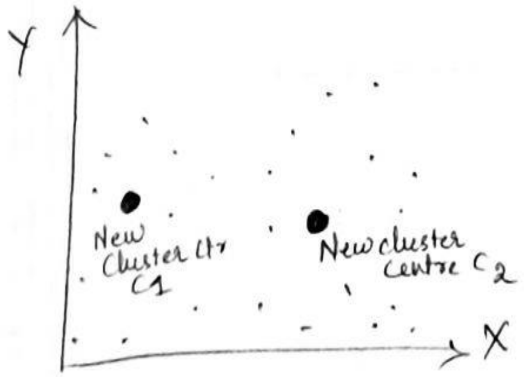
Say. Cluster C_1 has $d_1, d_2, d_3, d_4, d_6, d_7$.

Cluster C_2 has $d_5, d_8, d_9, d_{10}, d_{11}, d_{12}$.

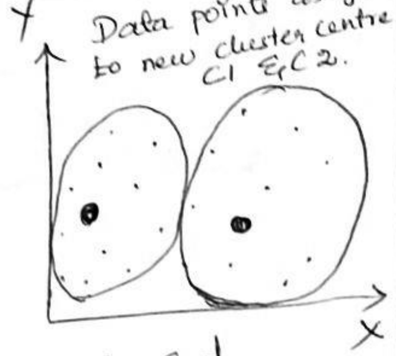
New cluster $C_1 = d_1+d_2+d_3+d_4+d_6+d_7 / 6$

New Cluster $C_2 = d_5+d_8+d_9+d_{10}+d_{11}+d_{12} / 6$

Use the ~~m~~ new cluster points C_1 & C_2 Calulate the euclidean distance b/w the datapoints. Based on the distance calculated assign the datapoints to the clusters.



Data points / New cluster centre	C1	C2
d_1	d_1-dist-c1	d_1-dist-c_2
d_2	d_2-dist-c1	d_2-dist-c_2
⋮ dn	⋮	⋮



Data points assigned to new cluster centre C1 & C2.

d_1-dist_ C1 → dist b/w new cluster C_1 & d_1.

d_2-dist_ C2 → dist b/w new cluster C_2 & d_1.

If d_1-dist_C1 < d_2-dist_ C_2 → assign d_1 to new cluster centre ~~d~~ C_1 else assign it to cluster centre C_2.

Continue the same process for all the data points.

Step: 5 & 6:

(1) Repeat the steps 2 to 4 again.

(2) Stop iterating at a point where the K-Means algorithm converges. (i.e) two iterations converge have same cluster.

This means, when you are calulating the new cluster centre, it will remain same as the previous cluster centre. & We can stop at this point.

Now, our data has been clustered intwo clusters C_1 & C_2.

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

Statistical Approach:

The optimal value of K can be chosen by using two methods.

1. Elbow-curve/SSD
2. Silhouette Analysis

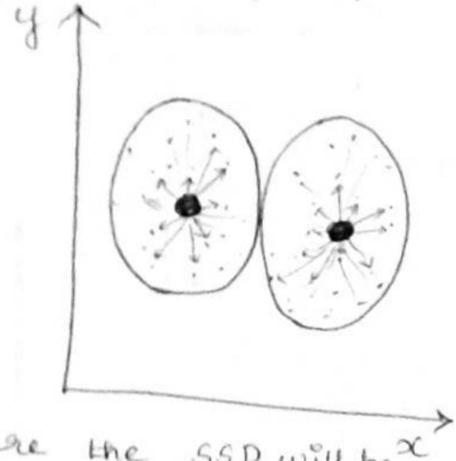
Elbow-curve/SSD:

This Elbow-curve method involves calculating sum of squared distance (SSD). SSD is the calculation of sum of squared distance of the point to their closest cluster centre.

If the value of K = 2, the sum of squared distance (SSD) of the datapoint to the cluster will be higher than to the sum of squared distance (SSD) of the datapoint to the cluster centre when the value of K = 3. So, if the number of clusters (value of K) increases the value of SSD decreases. At particular point there will no significant change even if the value of K increases. We can stop at the point and decide the value of K, which can be considered as the optimal value of K.

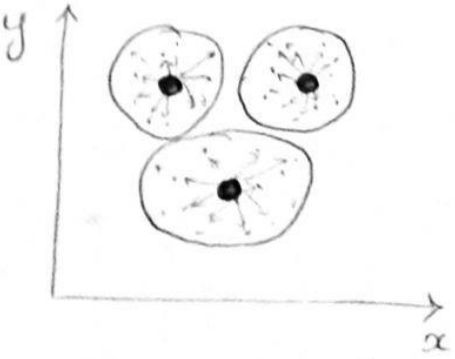
Let us discuss with a small representation for our better understanding.

K = 2 (no. of Clusters)



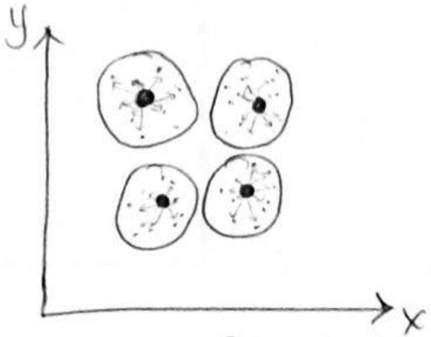
Here the SSD will be higher as we have 2 clusters.

K = 3 (no. of clusters)



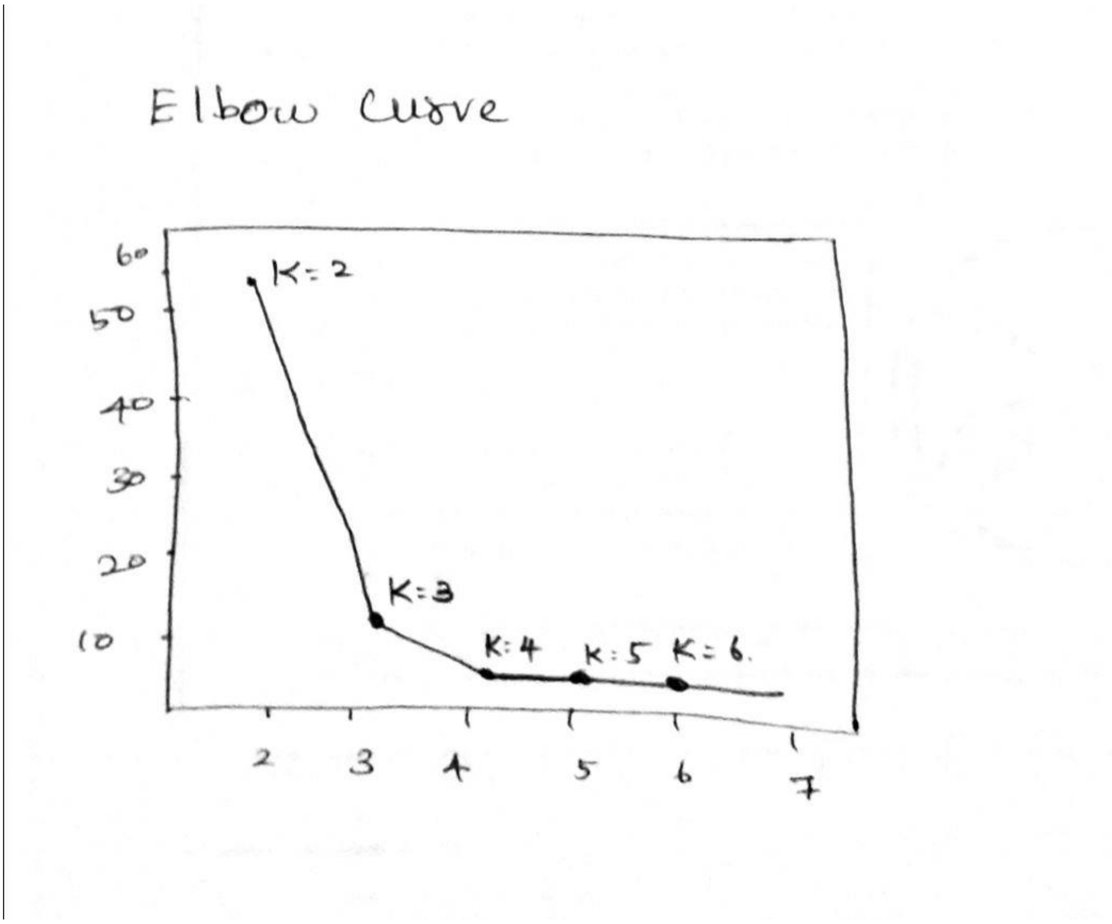
As, we have defined another cluster, few datapoints has been moved to the new cluster (closest to them). Hence, the SSD will get reduced.

K = 4 (no. of clusters)



Now, on introducing another new cluster, few datapoints has been moved to their nearest cluster. Again our SSD value got reduced

To visualise these, we can plot and see the values of K.



From the above visualisation, we can view that there is a decrease in the value SSD when the cluster value changes from K = 2 to K =3. And when the value changes from K = 3 to K = 4, there is no significant change and the same pattern follows for the subsequent K values.

So, we can choose the value of K as 3 or 4. In this way, we can use the Elbow-curve/ SSD to choose optimal value for K.

Now, there is another method to choose the value of K, which is the Silhouette Analysis.

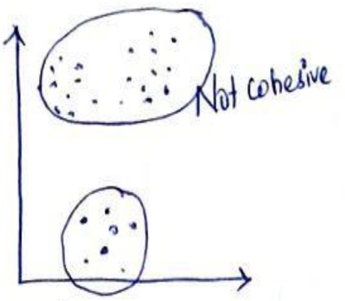
Silhouette Analysis:

The value of K which we have chosen should satisfy the below two conditions.

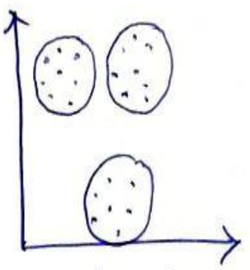
The K, which we chose should satisfy two Conditions,

(1). All the datapoints in the cluster (group) should have minimum distance (Close to each other).

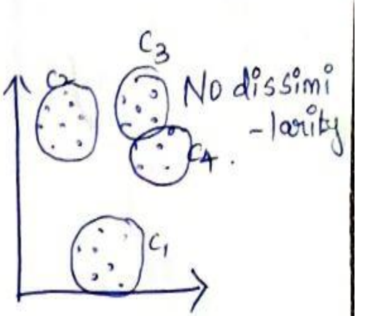
(2). The behaviour of datapoints in the diff clusters Should show dissimilarity (i.e). The behaviour of datapoints in cluster 1 should be different from the behaviour of datapoints in Cluster 2.



k = 2



K = 3



k = 4

When K = 2, there is ~~a~~ no cohesiveness between the datapoints. (1^{st} condition fails).

When k = 4, Cluster C_3 and C_4 have same behaviour.

So, it is optimal to have K = 3.

This can be found by using Silhouette Metric.

Silhouette Metric :

For every ith data point,

a(i) → Average distance from (b/w datapoints) Own cluster (as minimum as possible). [Cohesion]

b(i) → Average distance from the nearest neighbouring cluster (as large as possible).

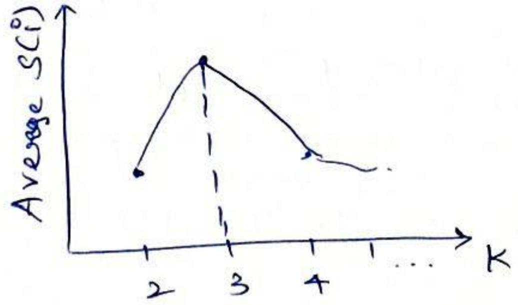
[dissimilarity].

$$S(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

$a(i) << b(i)$.

Find the average.

$\boxed{\text{Mean of all } S(i)}$ (i) For every K.



The highest point is the optimal Value of K.

The range of the Silhouette value is between +1 and -1. A **high value is acceptable** 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.

Business aspects:

Consider a company that has three departments and you want to cluster the employees based on the departments. On running Elbow-curve/SSD and Silhouette score, we get an optimal value of K = 5.

In this case, we need to think from the business point of view as well. There are only 3 departments in the company. Making 5 clusters does not make any sense. So, we have to decide based on the business aspect and go ahead with K = 3.

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

It is must to perform the scaling/standardisation before performing clustering which prevents the variables with larger scales from dominating in the formation of clusters.

Consider the below example. Below table shows the average number of people travelling in the train at Sydney between 7 am to 7 pm.

Time in hours	Total number of people
7 am	10000
8 am	11000
9 am	9000
10 am	6000
11 am	3000
12 noon	2000
1 pm	1000
2 pm	1000
3 pm	500
4 pm	700
5 pm	6000
6 pm	12000
7 pm	10000

Here the variable Time is in hrs (small in range) and Total number of people in thousands (high in range).

Say, if the randomly chosen cluster centre is (6000,10) and (9000,5). If we calculate the Euclidean distance

Euclidean Distance = [(10000–6000)^{2} + (7–10)^{2}]^(1/2) = 4000

Euclidean Distance = [(10000–9000)^{2} + (7–5) ^{2}]^(1/2) = 1000

So, our cluster will be formed in such a way that more significance will be given to the variable that is high in range. In our case it is total number of people. In other words, total number of people has become the primary driver in dividing the clusters.

After standardisation/scaling, both the variables will be in same unit and both of them will have same influence in forming the clusters and the performance of our model will be good.

After standardisation/scaling (rough calculations) both the features will be in same unit.

Time in hours	Total number of people
1.2	0.87
1.7	0.99
1.11	1.0
1.4	1.2
0.98	1.4
1.2	0.96
1.1	1.6
1.0	1.5
0.99	0.923
1.5	1.4
1.6	0.98
1.6	1.7
0.96	0.93

Say, if the randomly chosen cluster centre is (0.9,1) and (1.7,0.9). If we calculate the Euclidean distance

Euclidean Distance = $[(1.2-0.9)^{2} + (0.87-1)^{2}]$ ^ (1/2) = 0.326

Euclidean Distance = $[(1.2-1.7)^{2} + (0.87-0.9)^{2}]$ ^ (1/2) = 0.499

Now, both the variables have equal influence in forming the clusters. Both of them are in same range.

Standardisation is the process of converting them into Z-scores with mean 0 and standard deviation 1.

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

Min-Max scaling can also be done using the below formula.

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

e) Explain the different linkages used in Hierarchical Clustering.

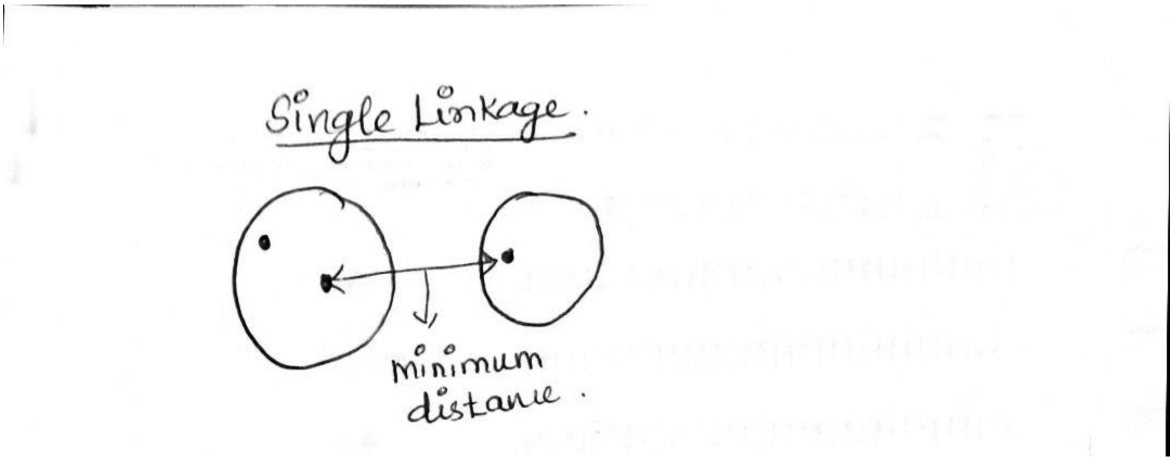
Linkage: Linkage is the measure of dissimilarity or similarity between the clusters having multiple observations.

There are three types of Linkages.

1. Single Linkage (Shortest distance)
2. Complete Linkage (Largest distance)
3. Average Linkage (Average distance)

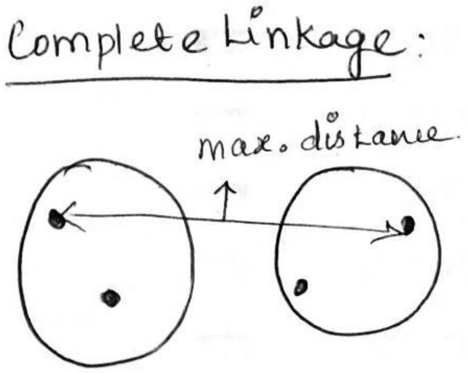
Single Linkage:

The distance between two clusters is defined as the shortest distance between points in the two clusters.



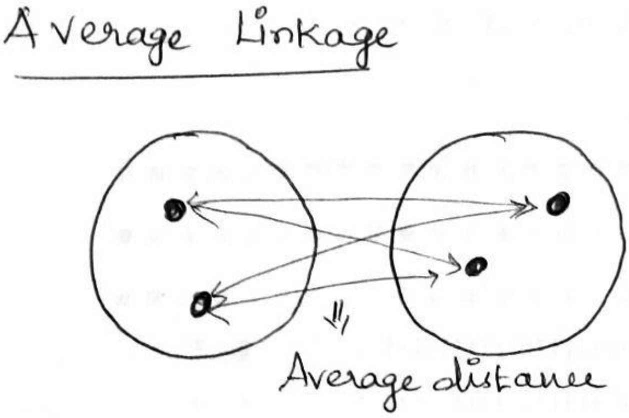
Complete Linkage:

The distance between two clusters is defined as the maximum distance between points in the two clusters.



Average Linkage:

The distance between two clusters is the average distance between every point of one cluster to every other point in the other cluster.


Average Linkage
Average distance