EE708, Assignment 3 Submission

Dhruv Gupta, Roll no: 240354

Problem 1

Question

a If s is a metric similarity measure on a set X with $s(x,y) \geq 0, \forall x,y \in X$ then $s(x,y)+a$ is also a metric similarity measure on set X, $\forall a \geq 0$.

b If d is a metric dissimilarity measure on a set X then $d+a$ is also a metric similarity measure on set X, $\forall a \geq 0$.

Answer

(a) A metric similarity measure $s(x,y)$ satisfies the following properties:

- **Non-negativity**: $s(x,y) \geq 0$ for all x, y.
- **Same point**: $s(x,y) = 0$ if and only if $x = y$.
- **Symmetry**: $s(x,y) = s(y,x)$.
- **Triangle Inequality**: $s(x,z) \leq s(x,y) + s(y,z)$.

Now, define $s'(x,y) = s(x,y) + a$.

- **Non-negativity**: Since $s(x,y) \geq 0$ and $a \geq 0$, we have $s'(x,y) \geq 0$.
- **Symmetry**: Since $s(x,y) = s(y,x)$, it follows that $s'(x,y) = s'(y,x)$.
- **Same point**: If $x = y$, then $s(x,y) = 0$, so $s'(x,y) = a$. This means that $s'(x,y) = 0$ only if $a = 0$
- **Triangle Inequality**: Since $s(x,y) \leq s(x,z) + s(z,y)$, adding a to both sides gives: $s(x,y) + a \leq (s(x,z) + a) + (s(z,y) + a) - a$.
  Since a appears on both sides and cancels out, the inequality still holds.

Thus, $s'(x,y)$ satisfies all properties except the same point condition unless $a = 0$. Therefore, for $a = 0$, it is a **metric similarity measure**, but even for $a > 0$, it can be called a relaxed metric similarity measure.

(b) A metric dissimilarity measure $d(x,y)$ satisfies:

1. **Non-negativity**: $d(x,y) \geq 0$.
2. **Symmetry**: $d(x,y) = d(y,x)$.
3. **Same point**: $d(x,y) = 0$ if and only if $x = y$.
4. **Triangle Inequality**: $d(x,y) \leq d(x,z) + d(z,y)$.

Now, define $d'(x,y) = d(x,y) + a$.

- **Non-negativity**: Since $d(x,y) \geq 0$ and $a \geq 0$, we have $d'(x,y) \geq 0$.
- **Symmetry**: Since $d(x,y) = d(y,x)$, it follows that $d'(x,y) = d'(y,x)$.
- **Same point**: If $x = y$, then $d(x,y) = 0$, so $d'(x,y) = a$. This means that $d'(x,y) = 0$ only if $a = 0$, otherwise it does not strictly satisfy the identity condition.
- **Triangle Inequality**: Since $d(x,y) \leq d(x,z) + d(z,y)$, adding a to both sides gives: $d(x,y) + a \leq (d(x,z) + a) + (d(z,y) + a) - a$.
  The inequality still holds, so the triangle inequality is preserved.

Thus, $d'(x,y)$ remains a valid metric dissimilarity measure for any $a \geq 0$, though it does not strictly satisfy the identity condition unless $a = 0$.

Problem 2

Question

Prove that the Euclidean distance satisfies the triangular inequality.
Hint: Use the Minkowski inequality, which states that for a positive integer p and two vectors $x = [x_1, ..x_l]^T$ and $y = [y_1, ..y_l]^T$ it holds that:

$$\left(\sum_{i=1}^{l} |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{l} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{l} |y_i|^p\right)^{1/p}$$

Answer

We need to prove that the **Euclidean distance** satisfies the **triangle inequality**, i.e.,

$$d(x, z) \leq d(x, y) + d(y, z)$$

for any points x, y, z in Euclidean space.

The Euclidean distance between two vectors $x = [x_1, x_2, \dots, x_n]^T$ and $y = [y_1, y_2, \dots, y_n]^T$ in an n-dimensional space is given by:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Similarly,

$$d(y, z) = \sqrt{\sum_{i=1}^{n} (y_i - z_i)^2}, \qquad d(x, z) = \sqrt{\sum_{i=1}^{n} (x_i - z_i)^2}.$$

The Minkowski inequality states that for any vectors x and y:

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p}.$$

For $p = 2$, this reduces to the Euclidean norm and implies that:

$$\sqrt{\sum_{i=1}^{n} (x_i - z_i)^2} \leq \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} + \sqrt{\sum_{i=1}^{n} (y_i - z_i)^2}.$$

Thus, we obtain:

$$d(x, z) \leq d(x, y) + d(y, z).$$

This proves that the Euclidean distance satisfies the **triangle inequality**.

Problem 3

Question

Prove whether $d(x, y) = |x - y|^2$ satisfies the properties of a valid distance metric.

Answer

We need to check whether the function $d(x, y) = |x - y|^2$ satisfies the properties of a valid **distance metric**.
As stated in Question 1, we will again check the four properties of a valid distance metric:
Checking Non-Negativity Since $d(x, y) = |x - y|^2$ is a square of the absolute difference, and a square is always non-negative, we have:

$$|x - y|^2 \geq 0 \quad \text{for all } x, y.$$

Thus, non-negativity is satisfied.

Checking case of same points
For this property to hold, $d(x, y) = 0$ should imply $x = y$:

$$|x - y|^2 = 0.$$

Since x and y are real numbers or vectors, this is only possible when $x = y$, hence this property is satisfied.

Checking Symmetry A valid metric should satisfy:

$$d(x, y) = d(y, x)$$

Since

$$d(x, y) = |x - y|^2 = |y - x|^2 = d(y, x)$$

Thus, symmetry holds.

Checking Triangle Inequality
A valid metric should satisfy: $\quad d(x, z) \leq d(x, y) + d(y, z).$

Checking with our function: $\quad |x - z|^2 \overset{?}{\leq} |x - y|^2 + |y - z|^2.$
This inequality **does not always hold**.

For example, if $x = 0$, $y = 1$, and $z = 2$, we get: $|0 - 2|^2 = 4, \quad |0 - 1|^2 + |1 - 2|^2 = 1 + 1 = 2.$
Since $4 \not\leq 2$, the inequality fails.
Thus, **the triangle inequality is not satisfied**.

Conclusion
Since $d(x, y) = |x - y|^2$ **fails the triangle inequality**, it is **not a valid distance metric**.

Problem 4

Question

In many clustering schemes, a vector **x** is assigned to a cluster C, considering the proximity between x and C, D(x, C), which can be defined as:

a $D_{\min}(x, C) = \min_{v \in C}(\delta(x, v))$ (Single-Linkage Clustering)
b $D_{\text{avg}}(x, C) = \langle \delta(x, v) \rangle_{v \in C}$ (Average-Linkage Clustering)
c $D_{\max}(x, C) = \max_{v \in C}(\delta(x, v))$ (Complete-Linkage Clustering)

Let $C = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$, where
$x_1 = [1.5, 1.5]^T$ $x_4 = [1.5, 2]^T$ $x_7 = [2, 3]^T$
$x_2 = [2, 1]^T$ $x_5 = [3, 2]^T$ $x_8 = [3.5, 3]^T$
$x_3 = [2.5, 1.75]^T$ $x_6 = [1, 3.5]^T$
and let $x = [6, 4]^T$. Assume that the Euclidean distance measures the dissimilarity between two points. Then find $D_{\min}(X, C), D_{\max}(X, C), D_{avg}(X, C)$.

Answer

We need to compute the proximity measures between $x = [6, 4]^T$ and the cluster C using **Euclidean distance**.

Compute Distances
Using the **Euclidean formula** $\delta(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$, distances are:

$$\delta(x, x_i) = \sqrt{(6 - x_{i1})^2 + (4 - x_{i2})^2}$$

$$\begin{aligned}&\delta(x, x_1) = 5.15, \quad \delta(x, x_2) = 5, \quad \delta(x, x_3) = 4.16, \quad \delta(x, x_4) = 4.92,\\ &\delta(x, x_5) = 3.61, \quad \delta(x, x_6) = 5.03, \quad \delta(x, x_7) = 4.12, \quad \delta(x, x_8) = 2.69.\end{aligned}$$

Proximity Measures:

- Single-Linkage (Min Distance):

$$D_{\min}(x, C) = \min\{5.15, 5, 4.16, 4.92, 3.61, 5.03, 4.12, 2.69\} = 2.69.$$

- Complete-Linkage (Max Distance):

$$D_{\max}(x, C) = \max\{5.15, 5, 4.16, 4.92, 3.61, 5.03, 4.12, 2.69\} = 5.15.$$

- Average-Linkage (Mean Distance):

$$D_{\text{avg}}(x, C) = \frac{5.15 + 5 + 4.16 + 4.92 + 3.61 + 5.03 + 4.12 + 2.69}{8} \approx 4.34.$$

Thus, the final results are:

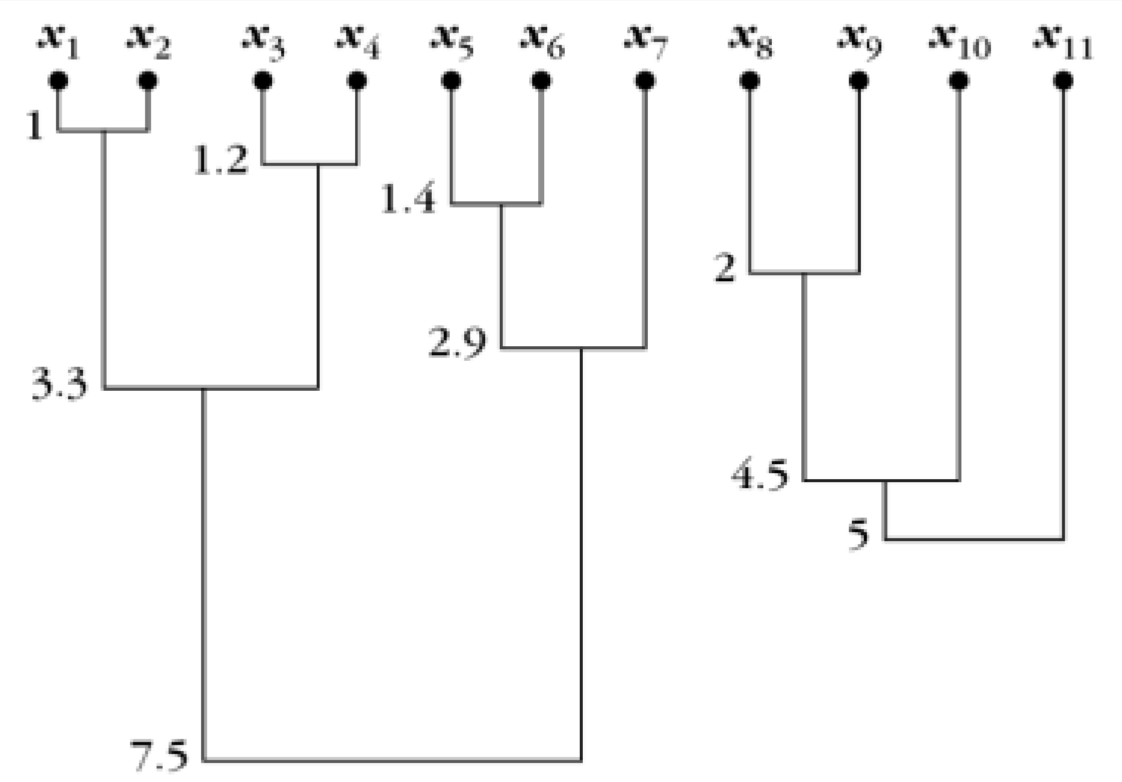
$$D_{\min}(x, C) = 2.69, \quad D_{\max}(x, C) = 5.15, \quad D_{\text{avg}}(x, C) \approx 4.34.$$

Problem 5

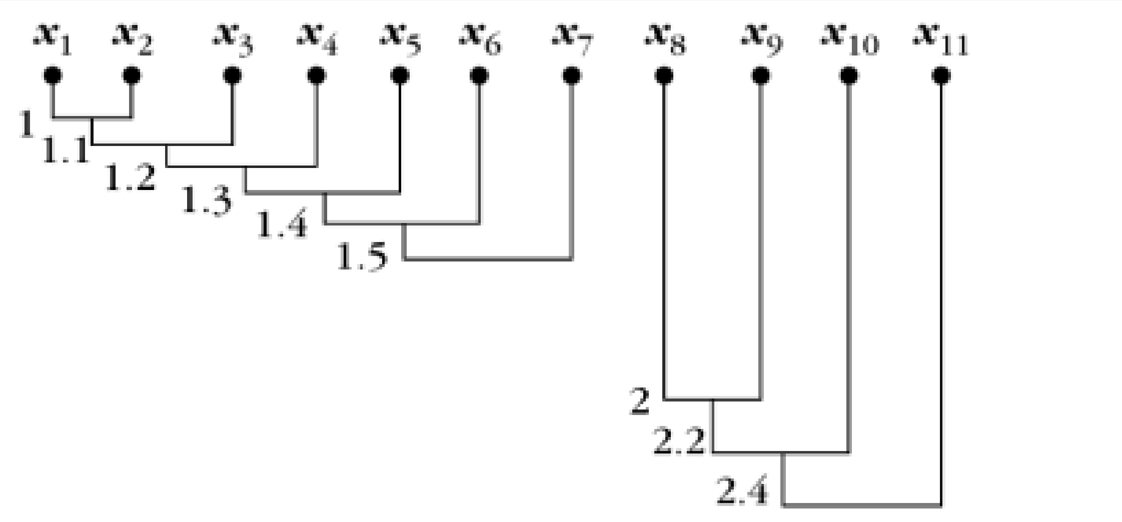
Question

Consider the data set shown in the figure. The first seven points form an elongated cluster, while the remaining four form a rather compact cluster. Draw the corresponding dendrograms based on dissimilarity.

Complete-Link Method



Single-Link Method



Problem 6

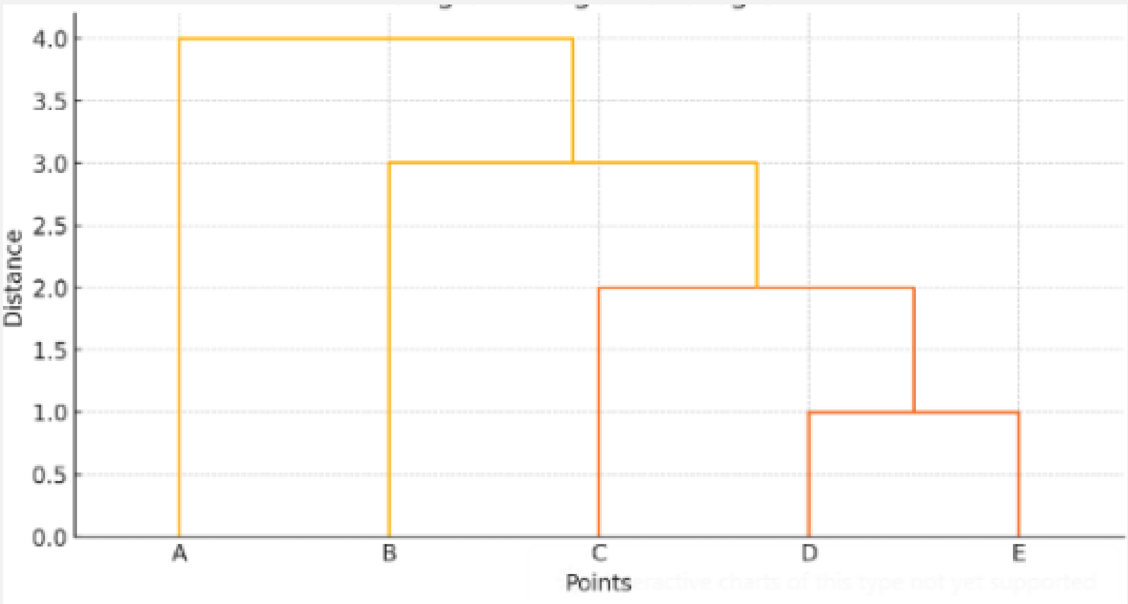
Question

For a dataset $\mathcal{C}$ with 5 samples, consider the dissimilarity matrix:

$$P = \begin{bmatrix} 0 & 4 & 9 & 6 & 5 \\ 4 & 0 & 3 & 8 & 7 \\ 9 & 3 & 0 & 3 & 2 \\ 6 & 8 & 3 & 0 & 1 \\ 5 & 7 & 2 & 1 & 0 \end{bmatrix}$$

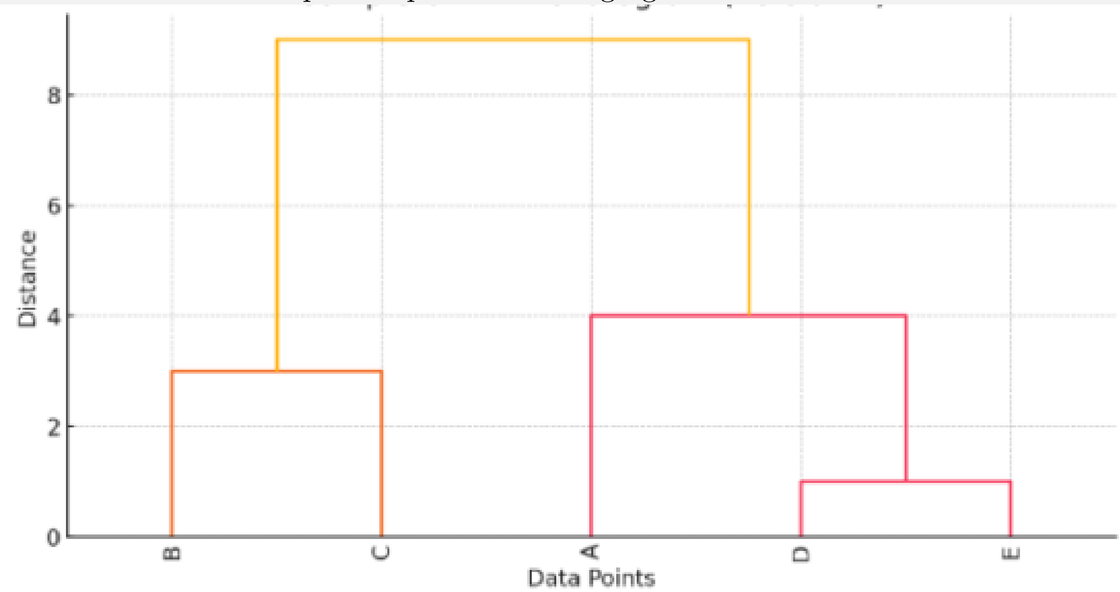
where $P_{i,j} = \delta(x_i, x_j)$. Determine all possible dendrograms resulting from applying the single-link and complete-link algorithms to P and comment on the results.

Single-Link Clustering

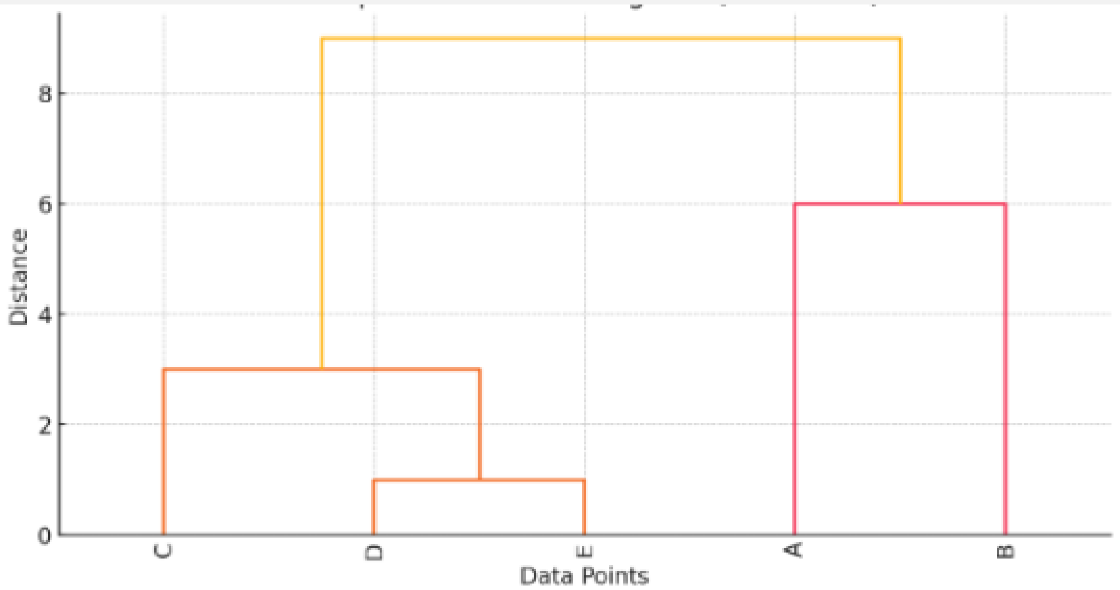


Complete-Link Clustering

We observe that during complete-link hierarchal clustering, there are two possible dendrograms obtained because during the second merge, there are two equally spaced options for merging the clusters.



Complete-Link Clustering



Problem 7

Question

Having generated a dendrogram, can we “prune” it? If yes, how?

Answer

Yes, a dendrogram can be ”pruned” to obtain better results for clustering.
We can prune the dendrogram based on a certain height or threshold we set, below which we consider the clustering to have stopped.

- If this threshold is set at a higher height, we will have a larger amount of clusters as the merging would have stopped earlier
- Similarly, if this threshold is set at a lower height, we will have a smaller amount of clusters as the merging would have stopped later.

(Here, height refers to the dissimilarity value during hierarchical clustering.)

Problem 8

Question

How can we make k-means robust to outliers?

Answer

K-Means Clustering is sensitive to outliers because of two main reasons:

a It uses a centroid-update method, so outliers can shift the clusters quite disproportionately

b It assumes the clusters to be perfect and spherically shaped, outliers can distort this shape and make ellipses or other shapes.

To make K-Means better and robust to outliers we can consider doing the following:

a Using K-Medians or taking the median of each cluster as the centroid instead of the mean, this is better when there are significant outliers in the dataset as it gives a better measure for central tendency.

b Identifying the outliers from the dataset and then clipping them while performing centroid calculations.

c Using other distance measures than Euclidean, for example, probabilistic distance measures like Mahalonobis distance.

Problem 9

Question

K-means clustering: Using the dataset in A3-P1.csv, implement K-means clustering and determine the number of clusters using the Elbow method.

a Plot the inertia (Within-Cluster Sum of Squares - WCSS) for number of clusters ranging from 1 to 15.

b Find the optimal number of clusters using the elbow method.

c Perform clustering using the optimal number of clusters, plot the clustering results, with each cluster data in a different colour, and highlight the cluster centres.

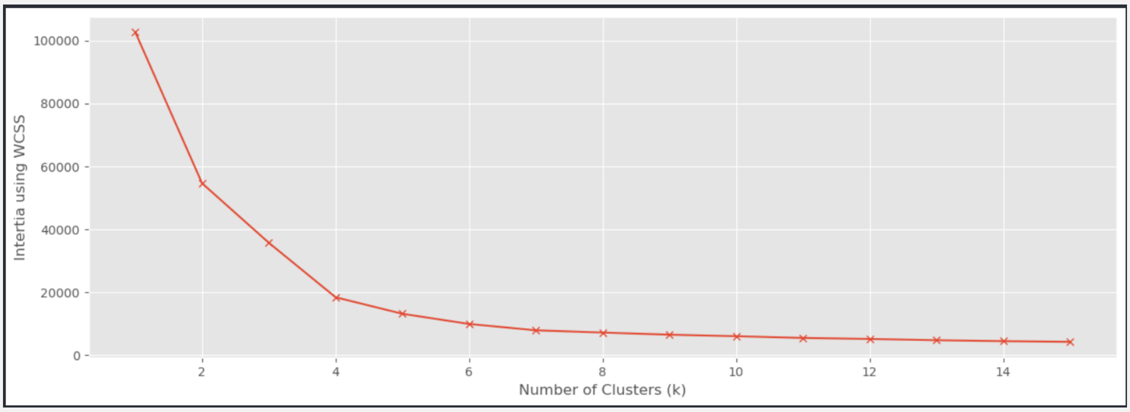
Answer: Part A

Computing inertia using within cluster sum of squared distances (WCSS)
I computed the value of inertia for each k between 1 and 15 manually using this method:

$$\text{WCSS} = \sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - \mu_j\|^2$$

Where: k is the number of clusters, μ_j is the centre of the jth cluster, C_j

Output: Part A

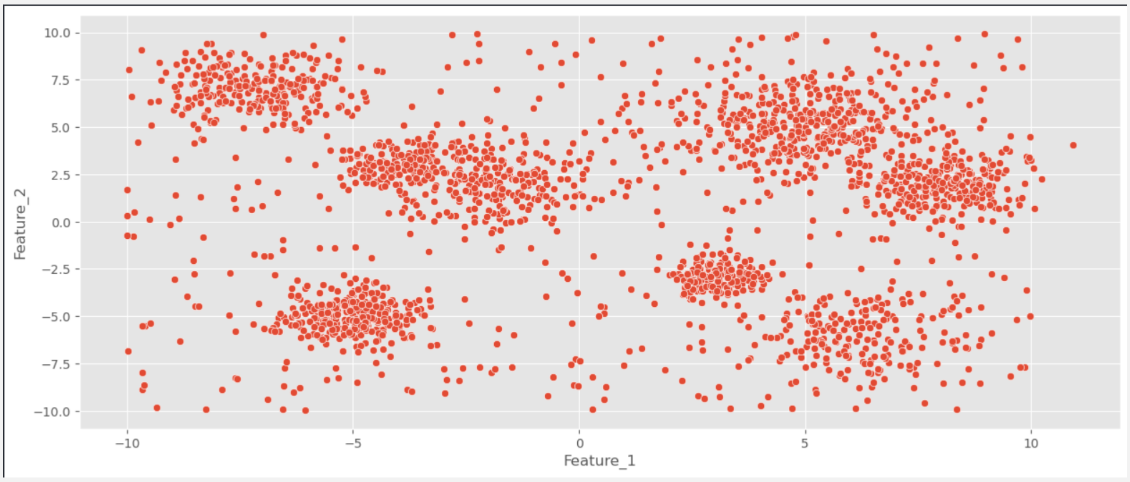


Plot of WCSS Inertia vs Number of clusters (k)

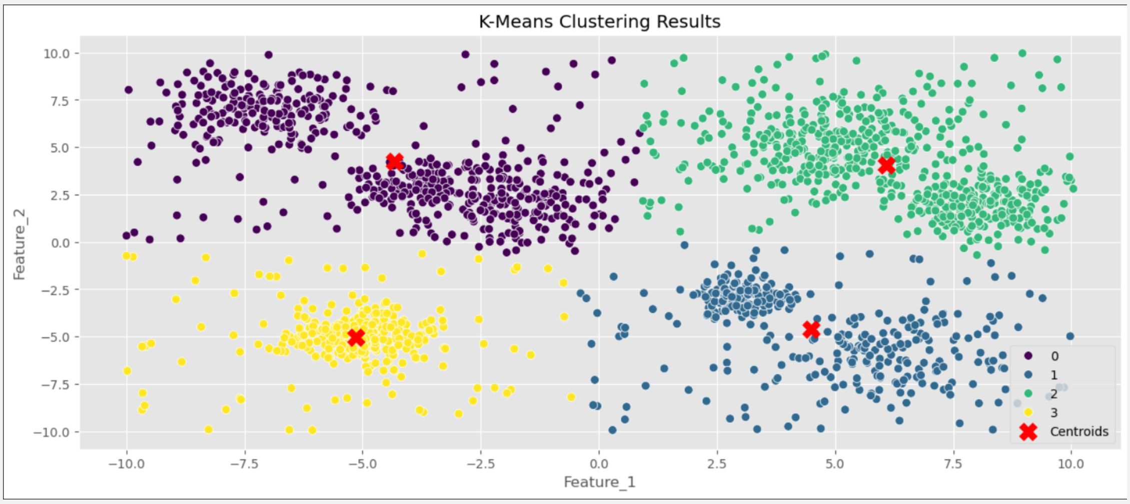
Answer: Part B

To find the optimal number of clusters, we observe above plot and then use the elbow method: We can see that the decrease in inertia is not that much after k=4, hence **k = 4 marks the most optimal number of clusters**.
k = 5 also can be used as the optimal number of clusters when we want a more fine-tuned model.

Output: Part C



Scatterplot of initial dataset



Plot of Clustering using optimal value of k (k=4)

Problem 10

Question

Hierarchical Clustering: Using the dataset in A3-P2.csv, implement bottom-up hierarchical clustering from scratch using Euclidean distance as the distance metric. Compute the distance between two clusters using the following methods:

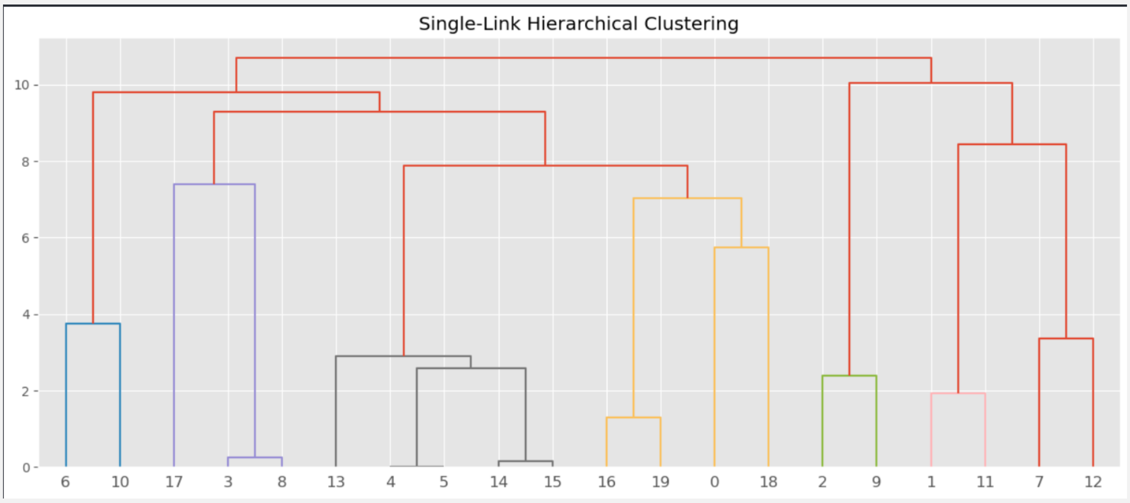
a $D_{\min}(A, B) = \min_{u \in A, v \in B}(\delta(u, v))$ (Single-Linkage Clustering)

b $D_{\text{avg}}(A, B) = \langle \delta(u, v) \rangle_{u \in A, v \in B}$ (Average-Linkage Clustering)

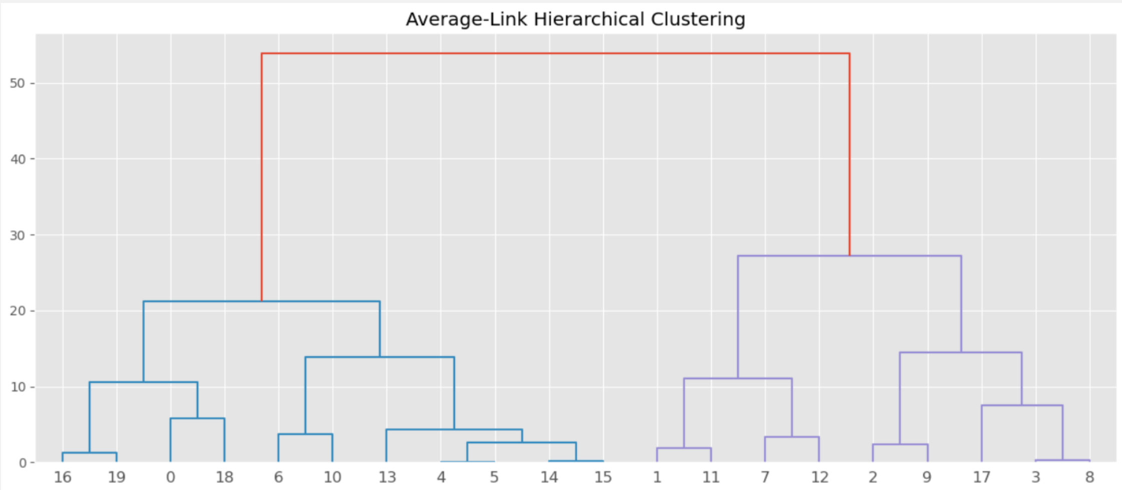
c $D_{\max}(A, B) = \max_{u \in A, v \in B}(\delta(u, v))$ (Complete-Linkage Clustering)

Plot dendrograms for each clustering method to visualize the hierarchical clustering process.

Output: Part A



Output: Part B



Output: Part C