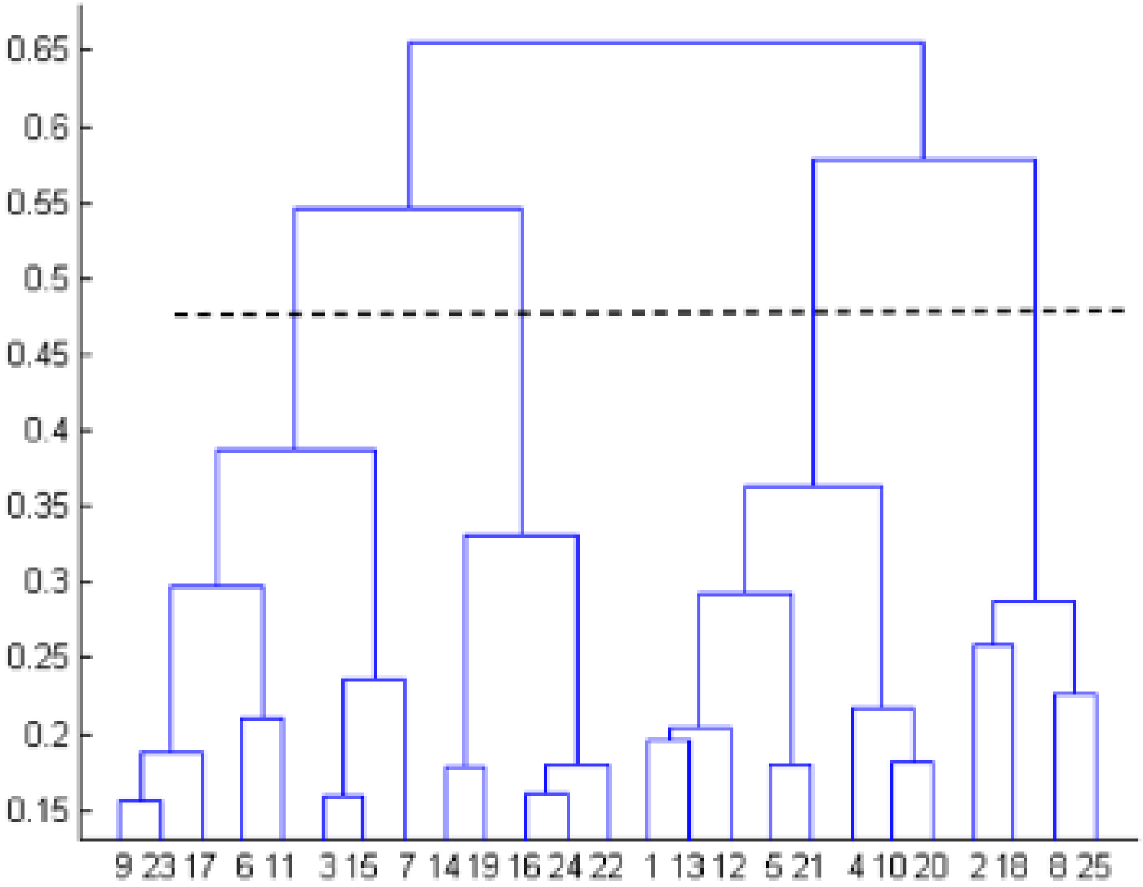
Data Analytics with Python

Week 10 Assignment Solution

Q1. **What is the most appropriate no. of clusters for the data points represented by the following dendrogram:**

Answer: 4

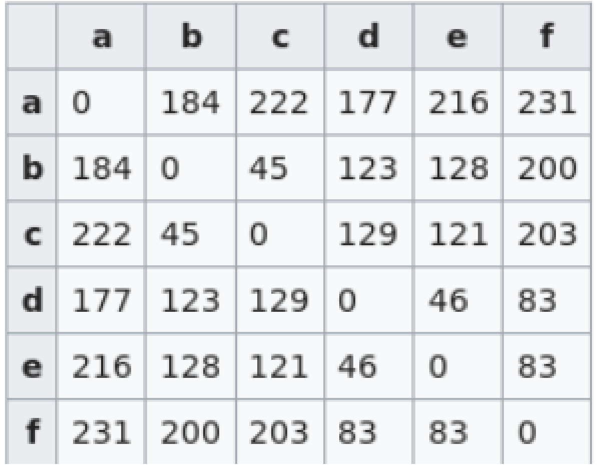
The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.



Q2.

K means versus hierarchical clustering

- Hierarchical clustering does not assume a particular value of ‘k’, as needed by k-means clustering
- The generated tree may correspond to a meaningful taxonomy
- Only a distance or “proximity” matrix is needed to compute the hierarchical clustering



	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

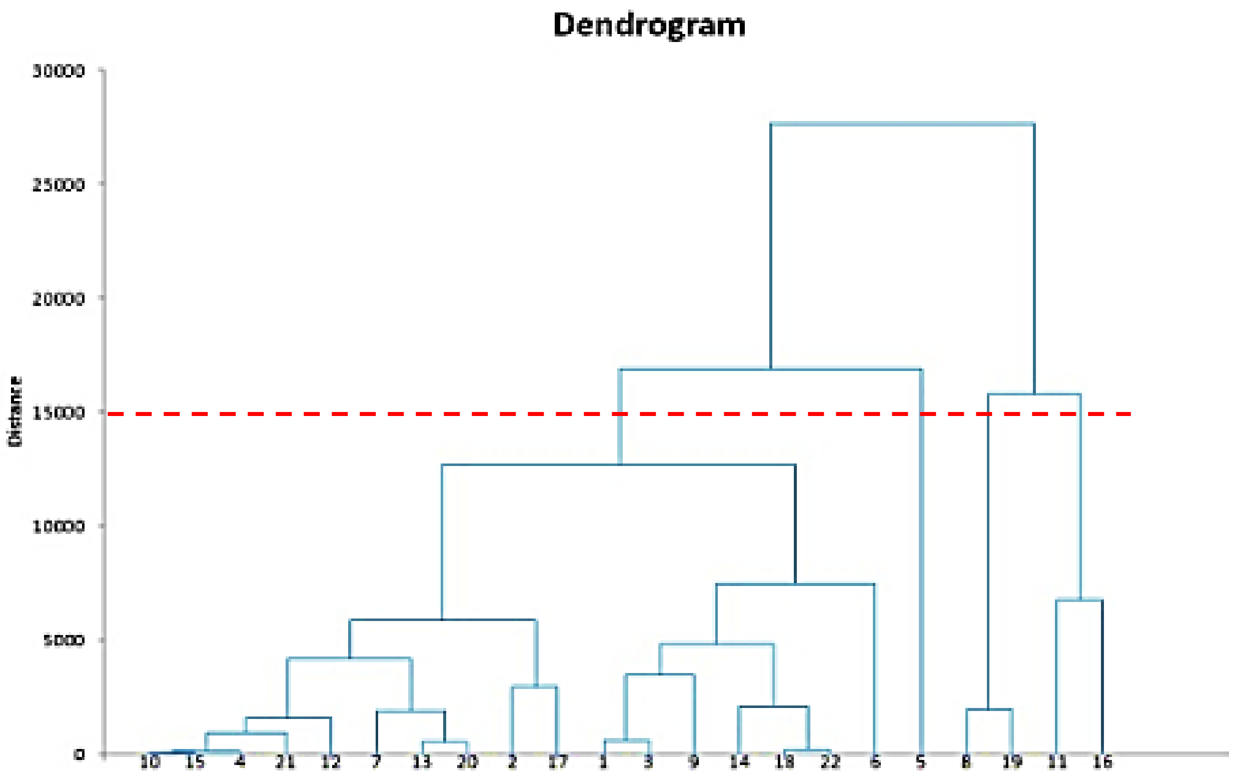
Proximity matrix

Q3. For the given dendrogram, what would be the threshold value for a total number of clusters equal to 4?

Answer: 15000

Explanation:

Look at the height at which a horizontal line will cut 4 vertical lines.



Q4. Which of the following clustering requires merging approach?

a. Partitional
b. Hierarchical
c. Naive Bayes
d. None of the mentioned

Answer: Hierarchical

In hierarchical clustering, the data points are grouped together based on the similarity between them. The clustering process starts with each data point forming its own cluster, and then iteratively merging the most similar clusters until a single cluster containing all data points is obtained.

Q5. **The Euclidean and Manhattan distance between X (3,4) and the origin is _______ respectively**

$$\begin{aligned}&\mathit{Given\ points}\\&X=(3,4)\\&Y=(0,0)\end{aligned}$$

The Euclidean distance between two points in a two-dimensional space is given as

$$d \quad = \quad \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

$$d \quad = \quad \sqrt{(3-0)^2+(4-0)^2} \quad = \quad \sqrt{9+16} \quad = \quad \mathbf{5}$$

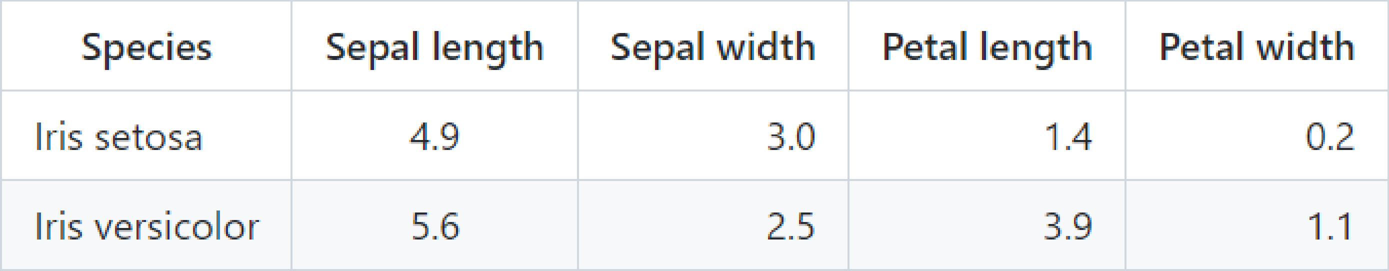
The Manhattan distance, also known as taxicab distance or L1 distance, is given as

$$d \quad = \quad |x_2-x_1| + \ |y_2-y_1|$$

$$d \quad = \quad |3-0| + \ |4-0| \quad = \quad 3+4 \quad = \quad \mathbf{7}$$

$\boldsymbol{Answer}$: $\boldsymbol{5\ and\ 7}$

Q6.



Species	Sepal length	Sepal width	Petal length	Petal width
Iris setosa	4.9	3.0	1.4	0.2
Iris versicolor	5.6	2.5	3.9	1.1

To calculate the Euclidean distance between the two objects, we first need to compute the differences between their corresponding features.

The difference in sepal length
d1 = (4.9 - 5.6) = -0.7

The difference in sepal width is:
d2 = (3.0 - 2.5) = 0.5

The difference in petal length is:
d3 = (1.4 - 3.9) = -2.5

And the difference in petal width is:
d4 = (0.2 - 1.1) = -0.9

Then use these differences to calculate the Euclidean distance using the following formula:

$$d = \sqrt{(d1)^2 + (d2)^2 + (d3)^2 + (d4)^2}$$

$$d = \sqrt{(-0.7)^2 + (0.5)^2 + (-2.5)^2 + (-0.9)^2}$$

$$d = \sqrt{0.49 + 0.25 + 6.25 + 0.81}$$

$$d = \sqrt{7.8}$$

$$\boldsymbol{d = 2.793 \approx 2.8}$$

Answer : 2.8

Q7. Which of the following is required by K-means clustering?

a. defined distance metric
b. number of clusters
c. initial guess as to cluster centroids
d. All of these

Answer: All of these

K-means clustering is a partitioning method that divides the data points into a fixed number of non-overlapping clusters. To perform K-means clustering, we need to specify the following three things:

1.Distance metric: a method for measuring the similarity between data points..

2.Number of clusters: the desired number of clusters to be formed.

3.Initial guess as to cluster centroids: a set of initial cluster centroids that will be iteratively updated during the algorithm.

Q8. Considering the K-means algorithm, after current iteration, we have 3 clusters A, B and C with centroids (0, 1) (2, 1), (-1, 2) respectively. Which cluster will the points (2, 3) and (2, 0.5) be assigned to in the next iteration?

a. Cluster A

b. Cluster B

c. Cluster C

d. None of these

Answer: Cluster

The points (2, 3) and (2, 0.5) will be assigned to Cluster B in the next iteration because it is closer to the centroid of Cluster B, which is (2, 1), than to the centroids of Clusters A and C.

Q9. State True or False: The K-means algorithm is sensitive to outliers

a. True

b. False

Answer: True

The K-means algorithm is sensitive to outliers, as they can heavily influence the position of the centroids and the overall clustering.

Q10. State True or False: Irrespective of the initial choice of parameters, K-means algorithm will always converge to the same final clusters

a. True

b. False

Answer: False

The final clusters obtained by the K-means algorithm can vary depending on the initial choice of parameters. The algorithm tries to minimize the sum of squared distances between the data points and their assigned cluster centroids, and the final solution can depend on the initial location of these centroids. Therefore, different initializations of the algorithm can result in different final clusters