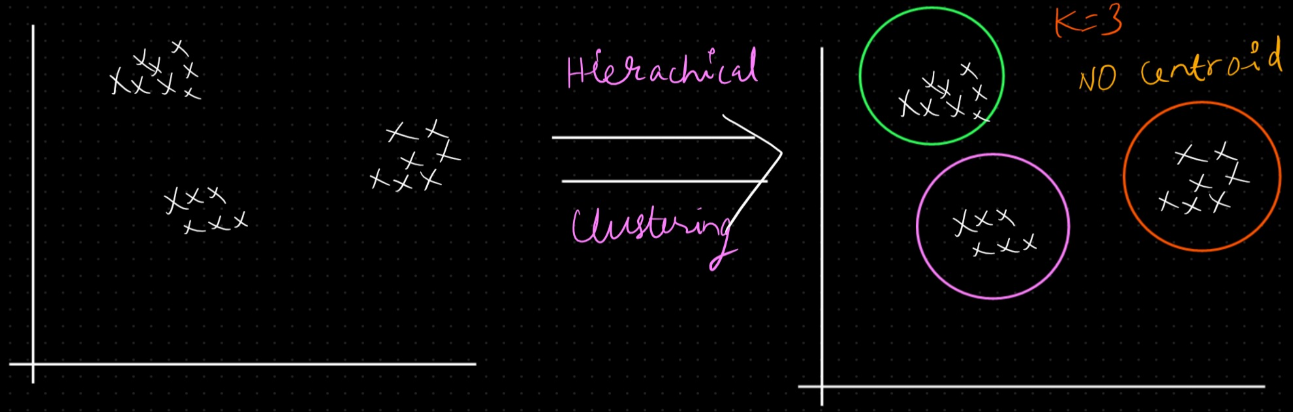
Hiearchical Clustering

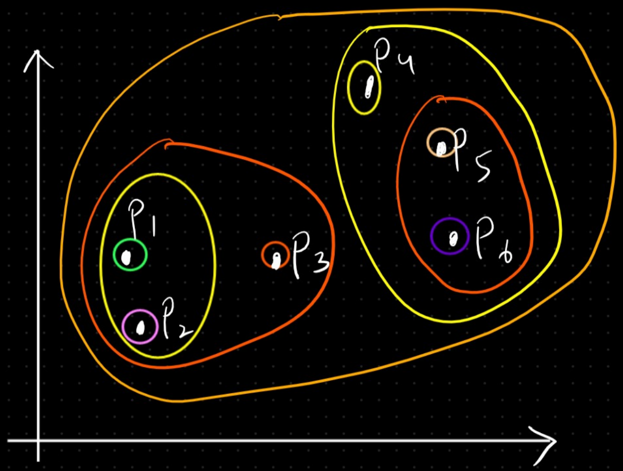
- → Hierarchical clustering is a method of cluster analysis that builds a hierarchy of clusters
- → It's based on the idea of continuously merging or splitting clusters until a certain criterion is met.
- → The two main approaches to hierarchical clustering are agglomerative (bottom-up) and divisive (top-down).



K-Means → Centroid

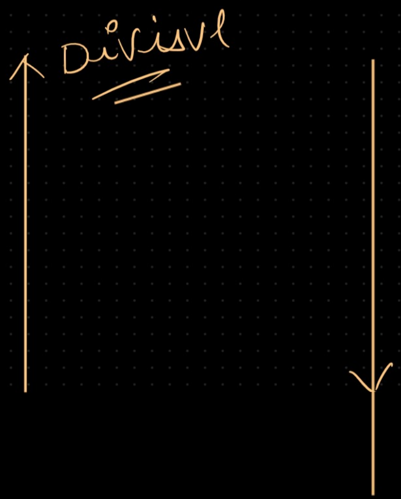
① Agglomerative

② Divisive



(i) Initialization :

Start by treating each data point as a single cluster. So if you have
n
n data points, you'll start with
n
n clusters, each containing one point.



(2) Compute Distance

Calculate the distance between.each pair of clusters. This could be done using various distance metrics, like Euclidean distance or Manhattan distance, depending on the nature of your data.

↓ ↓

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

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

(3) Merge closest clusters

Find the two clusters that are closest to each other based on the distance metric chosen in step 2. Merge these two clusters into a single cluster.

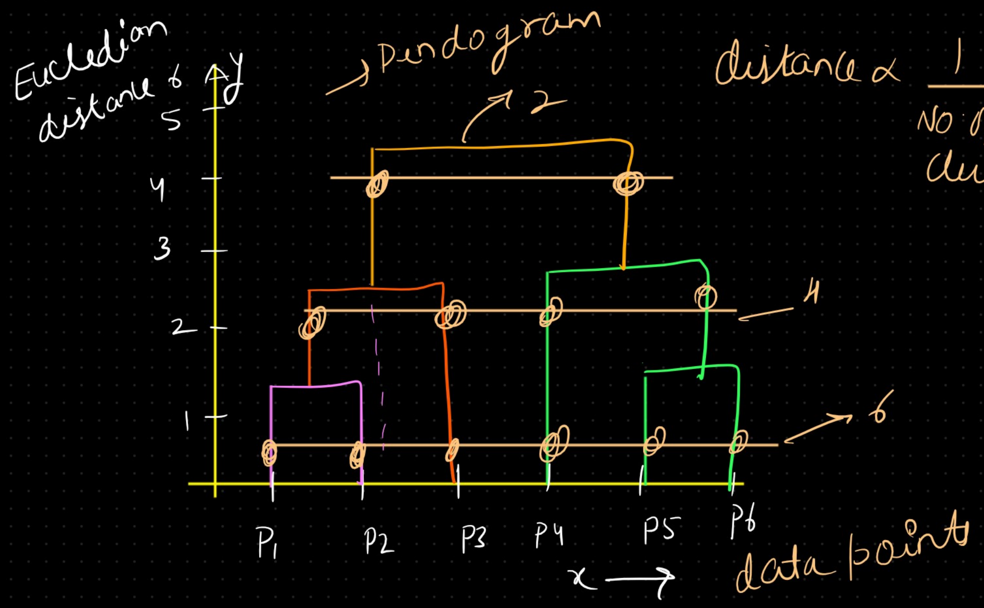
(4) update distance

Recalculate the distances between this new cluster and all the other clusters. This is typically done using linkage methods like single linkage (nearest neighbor), complete linkage (furthest neighbor), or average linkage.

(5) Repeat:

Repeat steps 3 and 4 until all data points belong to a single cluster, or until a stopping criterion is met (e.g., a certain number of clusters is reached, or a specific distance threshold is reached).

How many clusters? → Dendogram



$\text{distance} \propto \frac{1}{\text{No of cluster}}$

→ A dendrogram is a diagram that shows the arrangement of the clusters produced by hierarchical clustering. It looks like a tree structure, where the leaves represent individual data points (or clusters at the lowest level), and the branches represent the merging of clusters as the algorithm progresses.

In a dendrogram:

x – data points
y → Eucledian dist ance

The x-axis typically represents the individual data points or clusters.
The y-axis represents the distance or dissimilarity between clusters.
The height of each vertical line segment (or "branch") in the dendrogram indicates the distance or dissimilarity at which clusters are merged.

Dendrograms are useful for visualizing the hierarchical relationships between clusters and for determining the appropriate number of clusters by identifying the level at which to cut the dendrogram. Cutting the dendrogram at a particular height results in a certain number of clusters, each represented by the branches below the cut.

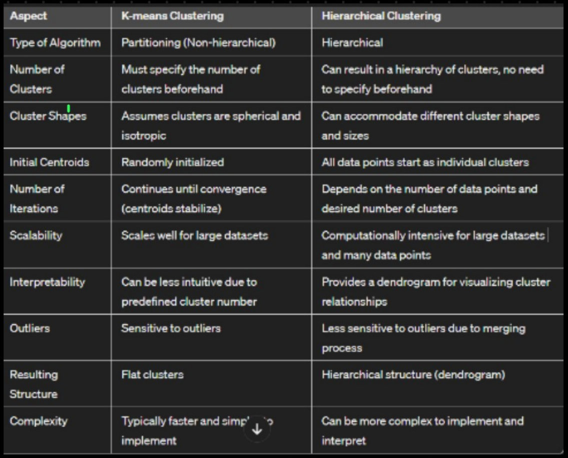
K Means Vs Hiearchical Clustering

① Dataset size

- → Large → K Means
- → Small → Hiearchical

② Types of Data

- → Numerical – Both
- → Variety of Data – Hiearchical



Aspect	K-means Clustering	Hierarchical Clustering
Type of Algorithm	Partitioning (Non-hierarchical)	Hierarchical
Number of Clusters	Must specify the number of clusters beforehand	Can result in a hierarchy of clusters, no need to specify beforehand
Cluster Shapes	Assumes clusters are spherical and isotropic	Can accommodate different cluster shapes and sizes
Initial Centroids	Randomly initialized	All data points start as individual clusters
Number of Iterations	Continues until convergence (centroids stabilize)	Depends on the number of data points and desired number of clusters
Scalability	Scales well for large datasets	Computationally intensive for large datasets and many data points
Interpretability	Can be less intuitive due to predefined cluster number	Provides a dendrogram for visualizing cluster relationships
Outliers	Sensitive to outliers	Less sensitive to outliers due to merging process
Resulting Structure	Flat clusters	Hierarchical structure (dendrogram)
Complexity	Typically faster and simp[illegible]o implement	Can be more complex to implement and interpret