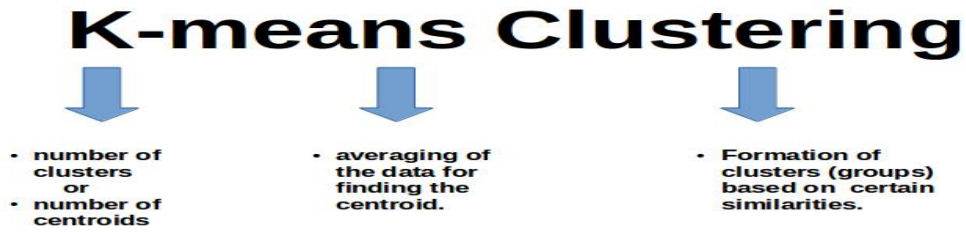
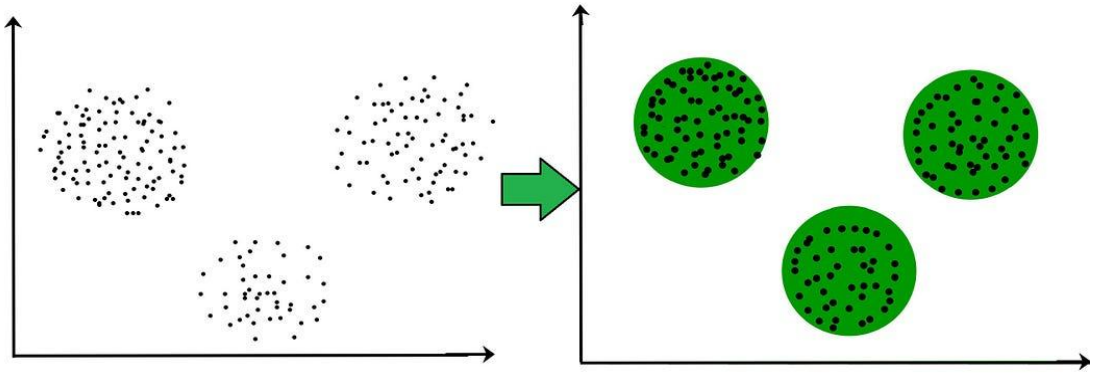
K-MEANS CLUSTERING ALGORITHM

- K-Means Clustering is an unsupervised learning algorithm that solves clustering problems in machine learning or data science.



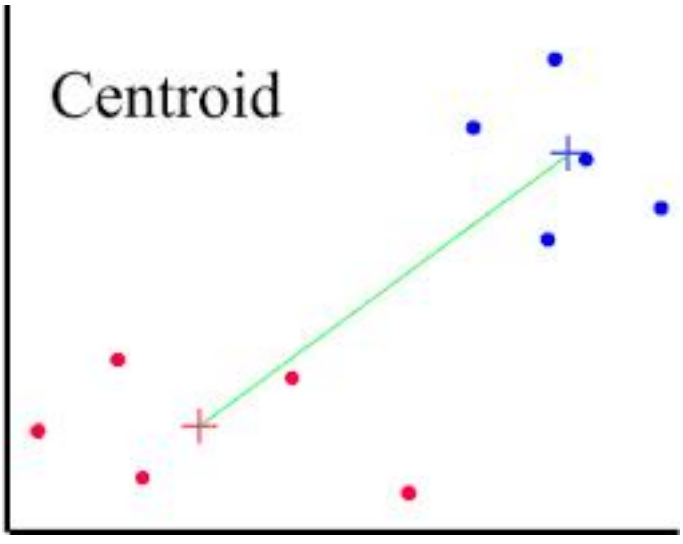
Clustering



- Clustering is a technique used in the K-means algorithm.
- In this algorithm, clustering refers to grouping similar data points based on their characteristics or features.
- The goal of clustering is to partition a set of data points into distinct clusters, where each cluster consists of data points that are more similar than those in other clusters.
- Clustering: grouping data based on similarity patterns based on distance
- The goal is to group similar instances into clusters.
- Clustering is an excellent tool for data analysis, customer segmentation, recommender systems, search engines, image segmentation, semi-supervised learning, dimensionality reduction, and more.

K-Mean Algorithm:

- K means comes under Unsupervised learning and is also called clustering algorithm.
- K mean is a clustering algorithm that is used to classify unlabeled data into groups/clusters based on similarity.
- Here K defines the number of pre-defined clusters that need to be created in the process if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- " It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties. "
- It allows us to cluster the data into different groups and is a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
- K-means then tries to determine different k-points called centroids, which are at the centre (least cumulative distance) from other points of the same class, but further away from points of another class.
- The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters.
- The value of k should be predetermined in this algorithm.



<u>Aim of the K-Mean Algorithm:</u>

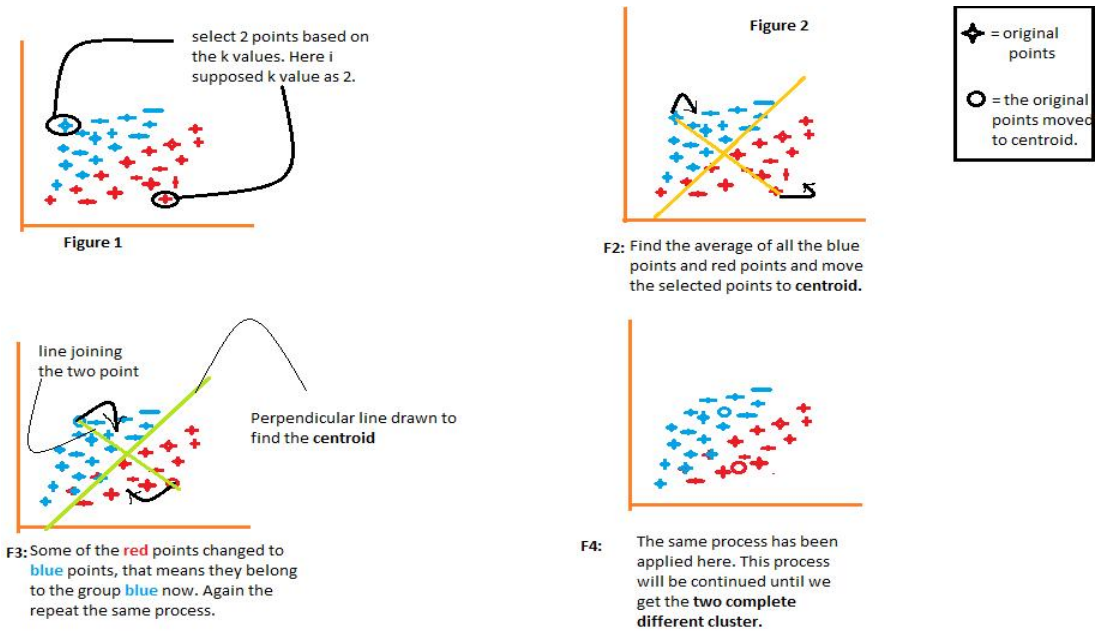
- The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K centre points or centroids by an iterative process.
- Assigns each data point to its closest K-center. Those data points which are near to the particular K-center, create a cluster.

How K-Mean Algorithm Works:

Plot Data

- Select the number K to decide the number of clusters.
- Select random K points or centroids. (It can be other from the input dataset).
- Assign each data point to their closest centroid, which will form the predefined K clusters.
- Repeat the fourth step, which means reassigning each data point to the new closest centroid of each cluster.
- Until you get a clearer cluster means no overlapping.

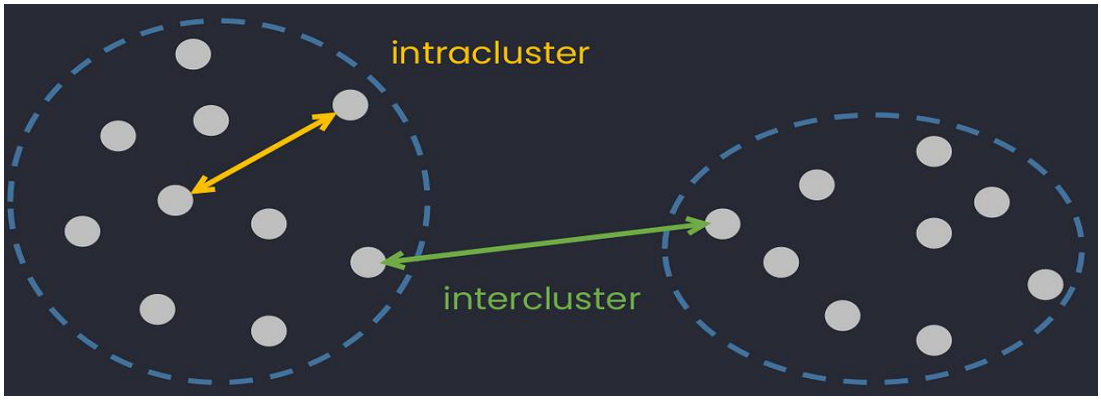


- We will understand each figure one by one.

- Figure 1 shows the representation of data from two different items. the first item has shown in blue colour and the second item has shown in red colour. Here I am choosing the value of K randomly as 2. There are different methods by which we can choose the right k values.
- In Figure 2, Join the two selected points. Now to find out the centroid, we will draw a perpendicular line to that line. The points will move to their centroid. If you will notice there, then you will see that some of the red points are now moved to the blue points. Now, these points belong to the group of blue colour items.
- The same process will continue in Figure 3. We will join the two points and draw a perpendicular line to that and find out the centroid. Now the two points will move to its centroid and again some of the red points get converted to blue points.
- The same process is happening in Figure 4. This process will be continued until and unless we get two completely different clusters of these groups.

Main points:

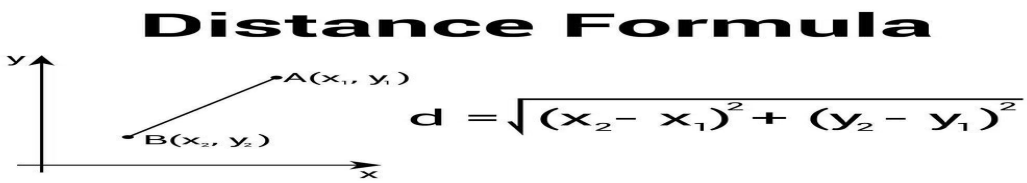
- **Intercluster distance should be high**: The distance between observations in two clusters should be High.
- **Intracluster Distance should be Very Less**: The distance of observation within the cluster should be very less.



Measuring Distance:

- **<u>Euclidean Distance Measure:</u>**

- The most common case is determining the distance between two points. If we have point P and point Q, the Euclidean distance is an ordinary straight line. It is the distance between the two points in Euclidean space.
- The formula for the distance between two points is shown below:



- **<u>Manhattan Distance Measure:</u>**

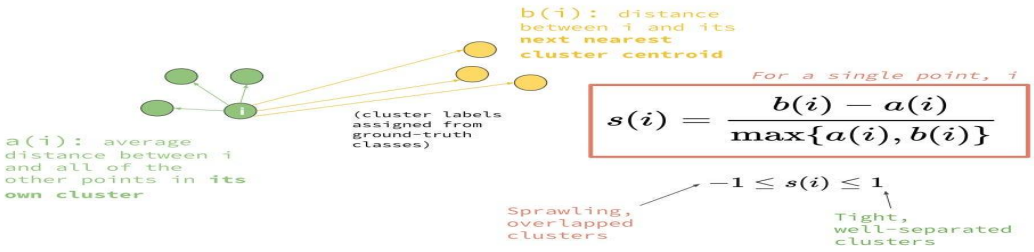
- The Manhattan distance is the simple sum of the horizontal and vertical components or the distance between two points measured along axes at right angles.
- The formula is shown below:

Manhattan $\sum_{i=1}^{k} |x_i - y_i|$

How to evaluate K-Mean Model?

Silhouette Coefficient:

- Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.
- 1: Means clusters are well apart from each other and clearly distinguished.
- 0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
- 1: Means clusters are assigned in the wrong way.



Important Points:

- It will use a Distance measure.
- Scaling is very important
- Handling outliers is also Important.

How to select the optimal value for k?

Elbow Method:

- The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster.

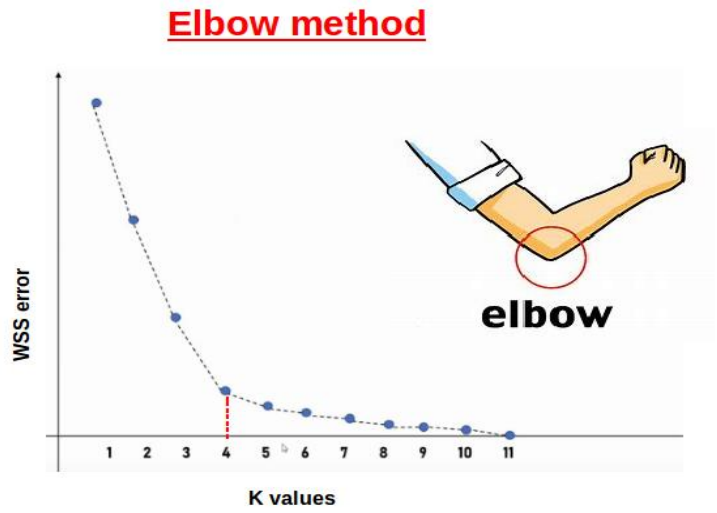
$$WCSS(k) = \sum_{j=1}^{k} \sum_{x_i \in \text{cluster } j} ||x_i - \bar{x}_j||^2,$$

where $\bar{x}_j$ is the sample mean in cluster j

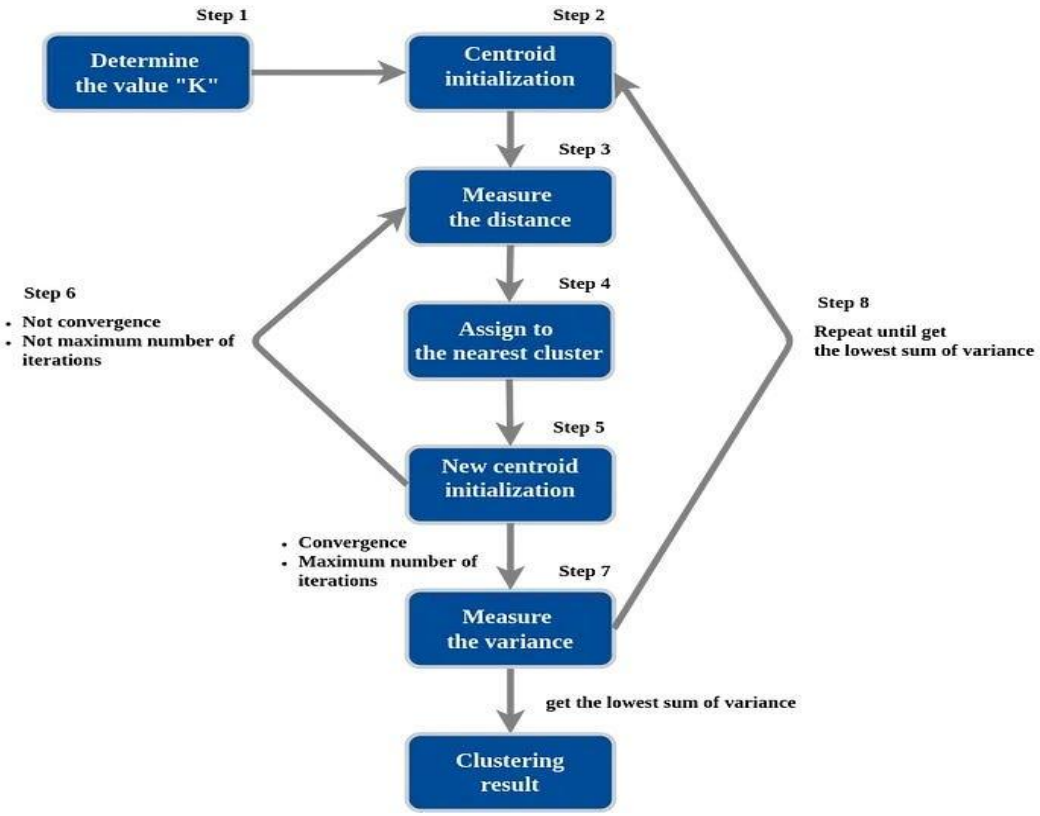
How does it work?

Start with some k
k=[2,3,4,5,6,…10]
If k=2 apply k-mean

It will find WCSS
Then step repeat for different k values
Plot a graph of k versus WCSS.
Choose the k value after which the WCSS value is constant.



K-mean Algorithm Overview:



Application OF K-Mean Algorithm:

- The k-means algorithm is a popular clustering algorithm used in various fields to group data points into distinct clusters. Here are some common applications of the k-means algorithm:
- Image Segmentation: In image processing, k-means can be used to segment an image into different regions based on colour similarity. Each cluster represents a distinct region in the image, allowing for further analysis or processing.
- Customer Segmentation: In marketing and customer analytics, k-means can be used to segment customers into groups based on their buying behaviour, demographics, or other relevant factors. This information can help businesses target specific customer segments with tailored marketing strategies.
- Anomaly Detection: K-means can be used to identify outliers or anomalies in a dataset. By clustering the data, the algorithm can identify data points that deviate significantly from the rest of the data, which can be useful in detecting fraudulent transactions, network intrusions, or other irregularities.
- Document Clustering: In natural language processing (NLP), k-means can be applied to group similar documents together based on their content. This can be useful for tasks such as document organization, topic modeling, and information retrieval.
- Recommendation Systems: K-means can be used in collaborative filtering-based recommendation systems to cluster users with similar preferences or behaviours. By identifying similar user clusters, personalized recommendations can be generated based on the preferences of users in the same cluster.

EXAMPLE

	X	Y
0	8	7
1	3	7
2	4	9
3	9	6
4	8	5
5	5	8
6	7	3
7	8	4
8	7	5
9	4	5

Step 1: Initialization

- Choose k = 2 clusters.
- Let's pick initial centroids randomly:

- C1 = (4,5)
- C2 = (8,5)

Step 2: Assignment

Point	Dist to (4,5)	Dist to (8,5)	Cluster
0 (8,7)	$\sqrt{((8-4)^2+(7-5)^2)}=\sqrt{20}=4.47$	$\sqrt{((8-8)^2+(7-5)^2)}=2.00$	Cluster 2
1 (3,7)	$\sqrt{((3-4)^2+(7-5)^2)}=\sqrt{5}=2.24$	$\sqrt{((3-8)^2+(7-5)^2)}=\sqrt{29}=5.39$	Cluster 1
2 (4,9)	$\sqrt{((4-4)^2+(9-5)^2)}=4.00$	$\sqrt{((4-8)^2+(9-5)^2)}=\sqrt{32}=5.66$	Cluster 1
3 (9,6)	$\sqrt{((9-4)^2+(6-5)^2)}=\sqrt{26}=5.10$	$\sqrt{((9-8)^2+(6-5)^2)}=\sqrt{2}=1.41$	Cluster 2
4 (8,5)	$\sqrt{((8-4)^2+(5-5)^2)}=4.00$	0.00	Cluster 2
5 (5,8)	$\sqrt{((5-4)^2+(8-5)^2)}=\sqrt{10}=3.16$	$\sqrt{((5-8)^2+(8-5)^2)}=\sqrt{18}=4.24$	Cluster 1
6 (7,3)	$\sqrt{((7-4)^2+(3-5)^2)}=\sqrt{13}=3.61$	$\sqrt{((7-8)^2+(3-5)^2)}=\sqrt{5}=2.24$	Cluster 2
7 (8,4)	$\sqrt{((8-4)^2+(4-5)^2)}=\sqrt{17}=4.12$	$\sqrt{((8-8)^2+(4-5)^2)}=1.00$	Cluster 2
8 (7,5)	$\sqrt{((7-4)^2+(5-5)^2)}=3.00$	$\sqrt{((7-8)^2+(5-5)^2)}=1.00$	Cluster 2
9 (4,5)	0.00	4.00	Cluster 1

- Compute distances of each point to both centroids and assign to the nearest one.
- Cluster 1 (C1): {1,2,5,9}
- Cluster 2 (C2): {0,3,4,6,7,8}

Step 3: Update Centroids

Recalculate centroids as the **mean of points in each cluster**.

- **Cluster1 (points 1,2,5,9):**

$$X_{mean} = \frac{3+4+5+4}{4} = 4, \quad Y_{mean} = \frac{7+9+8+5}{4} = 7.25$$

→ New centroid = **(4, 7.25)**

- **Cluster2 (points 0,3,4,6,7,8):**

$$X_{mean} = \frac{8+9+8+7+8+7}{6} = 7.83, \quad Y_{mean} = \frac{7+6+5+3+4+5}{6} = 5$$

→ New centroid = **(7.83, 5)**

Step 4: Re-Assignment

- Recompute distances with new centroids (4,7.25) and (7.83,5).

Point	Coordinates	Dist to C1 (4, 7.25)	Dist to C2 (7.83, 5)	Cluster
0	(8,7)	$\sqrt{((8-4)^2+(7-7.25)^2)} = \sqrt{(16+0.06)} \approx 4.01$	$\sqrt{((8-7.83)^2+(7-5)^2)} = \sqrt{(0.03+4)} \approx 2.01$	C2
1	(3,7)	$\sqrt{((3-4)^2+(7-7.25)^2)} = \sqrt{(1+0.06)} \approx 1.03$	$\sqrt{((3-7.83)^2+(7-5)^2)} = \sqrt{(23.3+4)} \approx 5.12$	C1
2	(4,9)	$\sqrt{((4-4)^2+(9-7.25)^2)} = \sqrt{(0+3.06)} \approx 1.75$	$\sqrt{((4-7.83)^2+(9-5)^2)} = \sqrt{(14.7+16)} \approx 5.47$	C1
3	(9,6)	$\sqrt{((9-4)^2+(6-7.25)^2)} = \sqrt{(25+1.56)} \approx 5.20$	$\sqrt{((9-7.83)^2+(6-5)^2)} = \sqrt{(1.37+1)} \approx 1.56$	C2
4	(8,5)	$\sqrt{((8-4)^2+(5-7.25)^2)} = \sqrt{(16+5.06)} \approx 4.52$	$\sqrt{((8-7.83)^2+(5-5)^2)} = \sqrt{(0.03+0)} \approx 0.18$	C2
5	(5,8)	$\sqrt{((5-4)^2+(8-7.25)^2)} = \sqrt{(1+0.56)} \approx 1.22$	$\sqrt{((5-7.83)^2+(8-5)^2)} = \sqrt{(8+9)} \approx 4.12$	C1
6	(7,3)	$\sqrt{((7-4)^2+(3-7.25)^2)} = \sqrt{(9+18.06)} \approx 5.35$	$\sqrt{((7-7.83)^2+(3-5)^2)} = \sqrt{(0.69+4)} \approx 2.18$	C2
7	(8,4)	$\sqrt{((8-4)^2+(4-7.25)^2)} = \sqrt{(16+10.56)} \approx 5.10$	$\sqrt{((8-7.83)^2+(4-5)^2)} = \sqrt{(0.03+1)} \approx 1.02$	C2
8	(7,5)	$\sqrt{((7-4)^2+(5-7.25)^2)} = \sqrt{(9+5.06)} \approx 3.60$	$\sqrt{((7-7.83)^2+(5-5)^2)} = \sqrt{(0.69+0)} \approx 0.83$	C2
9	(4,5)	$\sqrt{((4-4)^2+(5-7.25)^2)} = \sqrt{(0+5.06)} \approx 2.25$	$\sqrt{((4-7.83)^2+(5-5)^2)} = \sqrt{(14.7+0)} \approx 3.83$	C1

Cluster Assignment

- Cluster 1 (centroid 4, 7.25): {1,2,5,9}
- Cluster 2 (centroid 7.83, 5): {0,3,4,6,7,8}

Result

- The assignments remain stable after recomputing distances with the updated centroids.
- Final centroids: (4, 7.25) and (7.83, 5)

Final clusters:

- Cluster 1 → {1,2,5,9}
- Cluster 2 → {0,3,4,6,7,8}

