



Introduction to Clustering

Clustering is an *unsupervised machine learning* technique used to group similar data points together based on their features. Unlike classification, clustering doesn't rely on predefined labels—instead, it discovers inherent patterns or groupings in the data.

Each group formed is called a **cluster**, and the aim is to ensure that:

- Data points **within** the same cluster are more similar to each other.
- Data points **between** clusters are as different as possible.

Common Clustering Algorithms

- **K-Means**
- **Hierarchical Clustering**
- **DBSCAN**
- **Gaussian Mixture Models (GMM)**

Uses of Clustering in Business

Clustering is widely used in various business domains to segment data for strategic decision-making. Here are some key use cases:

1. Customer Segmentation

Goal: Group customers based on behavior to target marketing more effectively. **Example Dataset:** E-commerce data with features like purchase frequency, total spend, product categories, etc.

A retail company might use K-Means clustering to segment customers into:

- High-spenders
- Occasional buyers
- Deal seekers

These segments help in tailoring promotions and offers.

2. Anomaly Detection

Goal: Detect unusual behavior, such as fraud or outlier transactions. **Example Dataset:** Banking transaction data including amount, time, location, and device used.

If most customer behaviors fall into well-defined clusters, outliers can be flagged for investigation.

Here are **additional important business use cases of clustering**—beyond the ones you already listed. These cover multiple industries and practical applications where clustering directly drives revenue, efficiency, or risk reduction.

3. Product or Inventory Segmentation

Goal: Identify groups of products with similar sales patterns, profit margins, or lifecycles. **Example Dataset:** SKU-level sales, seasonality, return rates.

Use case example: A retailer clusters products into:

- Fast-moving items
- Seasonal products
- High-margin items
- Slow-moving or dead-stock

This guides stocking decisions, pricing, and liquidation strategies.

4. Churn Prediction Support (Behavioral Profiles)

Goal: Group users based on engagement or usage patterns to identify segments likely to churn. **Example Dataset:** Login frequency, feature usage, subscription renewal behavior.

Use case example: A SaaS company clusters customers into:

- Highly active
- Moderately active
- At-risk (low activity)

Marketing can then target at-risk groups with retention campaigns.

5. Supply Chain Optimization

Goal: Cluster suppliers, shipments, or routes for efficiency. **Example Dataset:** Delivery times, cost, reliability, shipment destinations.

Use case example: Logistics firms cluster delivery destinations to:

- Optimize route planning
- Reduce fuel costs
- Consolidate shipments

6. Image or Sensor Data Segmentation (Manufacturing)

Goal: Group patterns in quality-control images or sensor readings. **Example Dataset:** Machine vibration data, product inspection images.

Use case example: A factory clusters sensor signals into:

- Normal operation
- Early-stage wear
- Malfunction

This supports predictive maintenance and reduces downtime.

7. Healthcare Patient Segmentation

Goal: Group patients for targeted care paths, personalized medicine, or resource planning. **Example Dataset:** Symptoms, medical history, lab results, comorbidities.

Use case example: Hospitals cluster patients into:

- High-risk chronic disease group
- Low-risk routine care
- Emergency-prone group

Improving care delivery and resource allocation.

8. Financial Portfolio Segmentation

Goal: Group financial assets based on volatility, returns, risk profiles. **Example Dataset:** Historical prices, risk measures, correlations.

Use case example: A wealth management firm clusters stocks to build diversified portfolios or offer customized investment products.

9. Energy Usage Profiling

Goal: Group customers or equipment based on electricity consumption patterns. **Example Dataset:** Smart meter hourly usage, peak load times.

Use case example: Utility companies cluster consumers into:

- High daytime usage
- Evening peak users
- Low, steady users

Useful for demand forecasting and dynamic pricing plans.

10. Workforce or HR Segmentation

Goal: Cluster employees to understand performance, career trajectories, or training needs. **Example Dataset:** Years of experience, performance scores, skills, education.

Use case example: Companies cluster employees into:

- High performers ready for leadership
- Undertrained but high potential
- At-risk employees with morale issues

Useful for tailored development programs.

12. Pricing Segmentation

Goal: Group customers or products to identify differentiated pricing tiers. **Example Dataset:** Purchase history, market demand, customer willingness-to-pay.

Use case example: A company clusters customers based on price sensitivity for:

- Premium pricing
- Standard pricing
- Budget segment

Optimizing revenue with price discrimination.

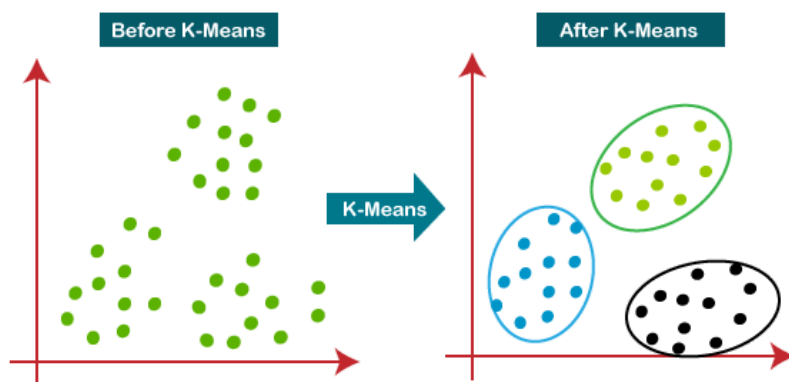
[More Examples](#)

1. Introduction to K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm used to group similar data points into a predefined number of clusters, K . The objective is to partition the dataset into K clusters, where each data point belongs to the cluster with the nearest mean, minimizing the variance within each cluster.

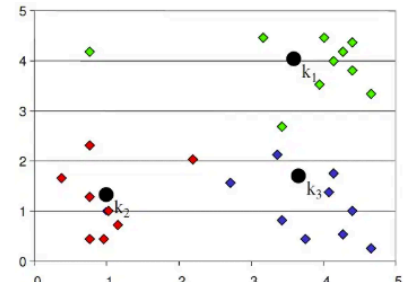
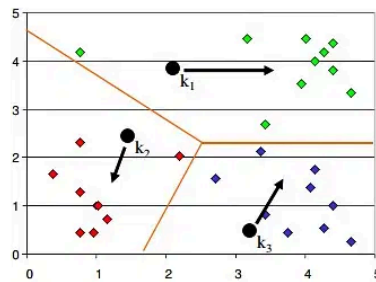
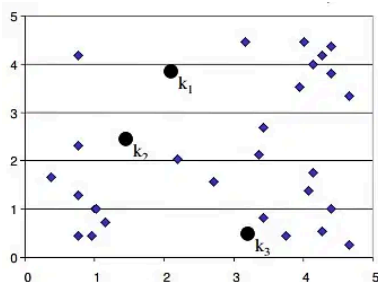
2. How K-Means Works

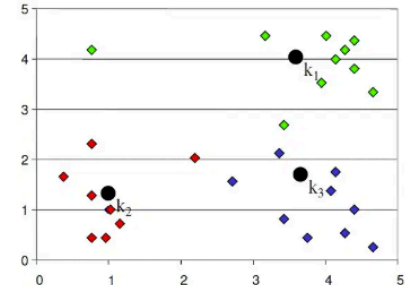
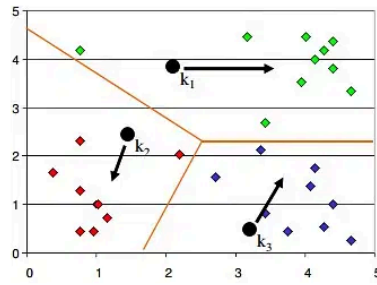
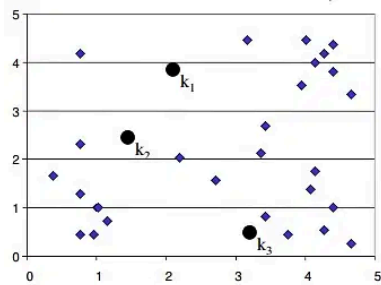
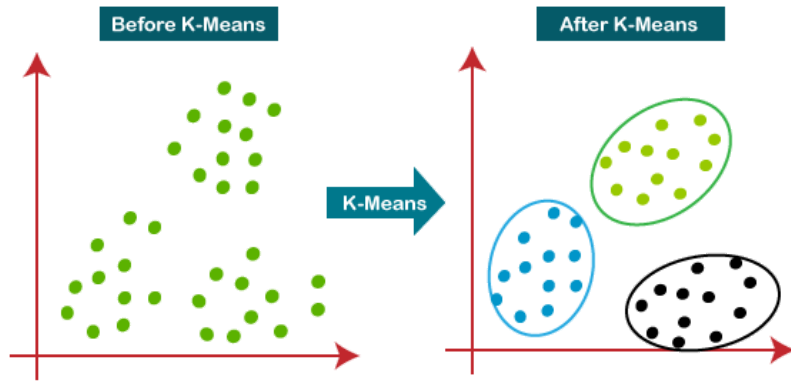
The K-Means algorithm works iteratively to partition a dataset into K clusters.



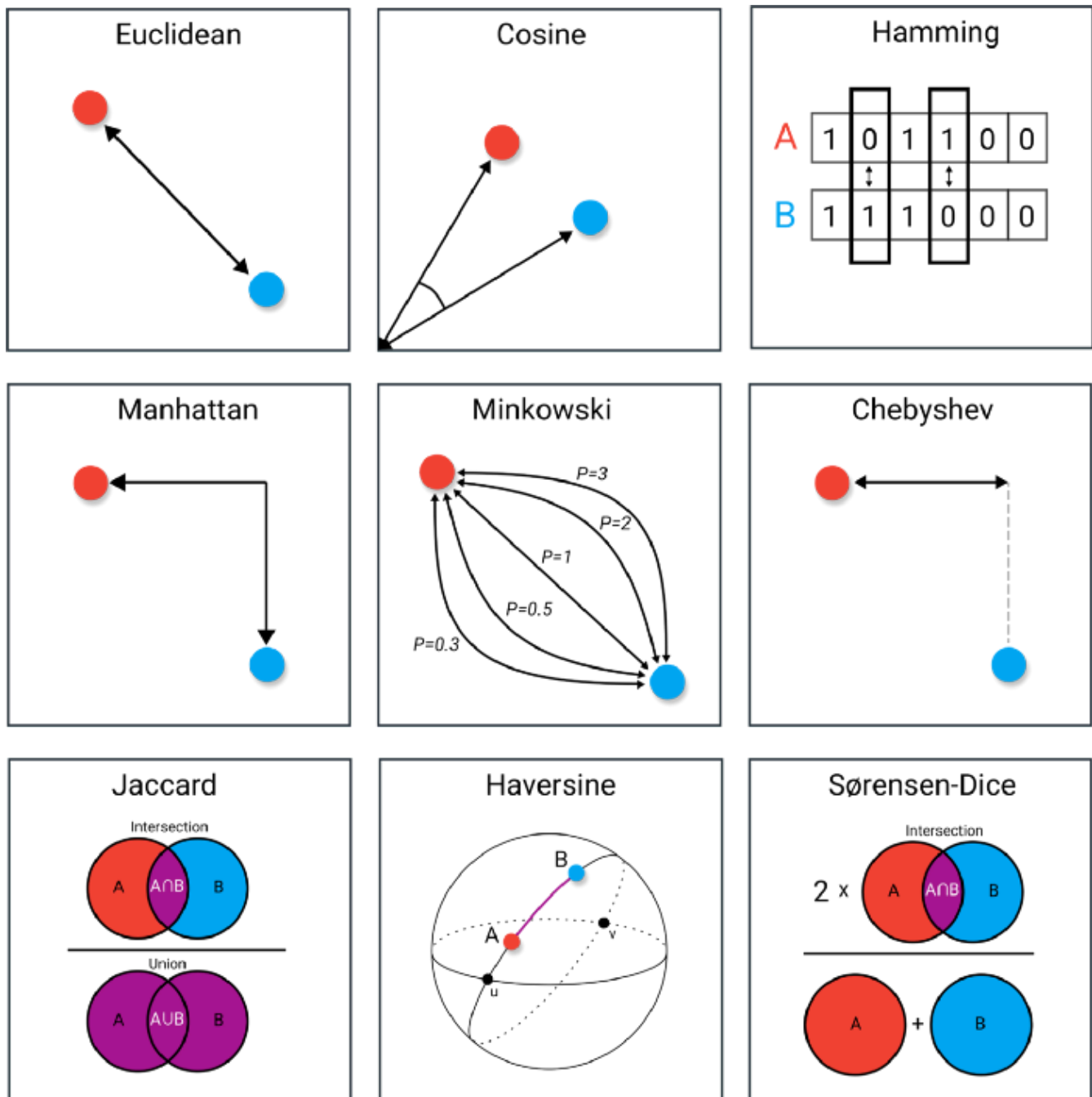
Here's how the algorithm proceeds:

- **Step 1:** Define the number K to decide the number of clusters.
- **Step 2:** Initialize K centroids **randomly**.
- **Step 3:** Assign each data point to the nearest centroid, forming K clusters.
- **Step 4:** Recalculate the centroids as the mean of all data points assigned to each cluster.
- **Step 5:** Repeat steps 2 and 3 until the centroids no longer change, or the algorithm converges.





4. Common Distance measures



4.1. Euclidean Distance

- Definition: The straight-line distance between two points in Euclidean space.
- Formula:

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Here, (x_i) and (y_i) are the (i) -th components (features) of the two points (x) and (y) ; the term $((x_i - y_i)^2)$ is the **squared difference** for that component, which (a) makes all contributions non-negative and (b) penalizes larger differences more strongly than smaller ones.
- For example, if $(x = (x_1, x_2, x_3, x_4))$ and $(y = (y_1, y_2, y_3, y_4))$, then $[\sum_{i=1}^4 (x_i - y_i)^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2]$
- Use Case: Commonly used in clustering (e.g., K-means) and nearest neighbor algorithms.

4.2. Cosine Similarity/Distance

- Definition: Measures the cosine of the angle between two vectors. It evaluates orientation rather than magnitude.

- Formula:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}| |\mathbf{B}|}$$

- Use Case: Used in text similarity and high-dimensional spaces (e.g., document similarity).

4.3. Hamming Distance

- Definition: Counts the number of positions at which corresponding symbols differ between two strings of equal length.
- Formula: Count mismatches between (A) and (B).
- Use Case: Applied in binary and categorical data, like error detection in coding.

4.4. Manhattan Distance

- Definition: The sum of absolute differences between the coordinates of two points.
- Formula:

$$\text{Manhattan Distance} = \sum_{i=1}^n |x_i - y_i|$$

- In this formula, (x_i) and (y_i) are again the (i) -th components of the points, and $(|x_i - y_i|)$ is the **absolute difference** for that component, which treats all deviations linearly and ignores direction (only the magnitude of the difference matters).
- For example, if $(x = (x_1, x_2, x_3, x_4))$ and $(y = (y_1, y_2, y_3, y_4))$, then $\sum_{i=1}^4 |x_i - y_i| = |x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3| + |x_4 - y_4|$.
- Use Case: Used when movement is restricted to horizontal and vertical paths (e.g., grid-based games or city blocks).

4.5. Minkowski Distance

- Definition: Generalized distance metric, including Euclidean ($(p=2)$) and Manhattan ($(p=1)$) distances as special cases.
- Formula:

$$\text{Minkowski Distance} = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

- Use Case: Offers flexibility to adjust distance metrics based on parameter (p) .

4.6. Chebyshev Distance

- Definition: The maximum absolute difference between two coordinates.
- Formula:

$$\text{Chebyshev Distance} = \max_i (|x_i - y_i|)$$

- Use Case: Used in chess for king's moves and grid-based pathfinding.

4.7. Jaccard Index/Distance

- Definition: Measures similarity or dissimilarity between two sets.
- Formula:

$$\text{Jaccard Index} = \frac{|A \cap B|}{|A \cup B|}$$

$$\text{Jaccard Distance} = 1 - \text{Jaccard Index}$$

- Use Case: Applied in text analysis, image segmentation, and set comparison.

4.8. Haversine Distance

- Definition: Calculates the shortest distance between two points on a sphere, considering curvature (e.g., Earth).
- Formula:

$$d = 2r \arcsin \sqrt{\sin^2 \left(\frac{\Delta \phi}{2} \right) + \cos \phi_1 \cos \phi_2 \sin^2 \left(\frac{\Delta \lambda}{2} \right)}$$

- Use Case: Used in geographic applications like GPS.

4.9. Sørensen-Dice Coefficient

- Definition: Measures the similarity between two sets.
- Formula:

$$\text{Sørensen-Dice Coefficient} = \frac{2|A \cap B|}{|A| + |B|}$$

- Use Case: Common in biology and text analysis.

5. K-Means Clustering Step by Step Example

Step 1: Euclidean Distance Formula

The Euclidean distance measures the straight-line distance between two points in a 2D space. The formula is:

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

We will use this formula to compute the distance between each point and the centroids.

Problem Setup

- **Data Points:** (x, y)
 - A1: (2, 10)
 - A2: (2, 5)
 - A3: (8, 4)
 - A4: (5, 8)
 - A5: (7, 5)
 - A6: (6, 4)
 - A7: (1, 2)
 - A8: (4, 9)
- **Initial Centroids:**
 - Mean 1: (2, 10)
 - Mean 2: (5, 8)
 - Mean 3: (1, 2)

We will calculate the distances between each point and the centroids and assign each point to the closest cluster.

Iteration 1

Steps:

1. **Compute Distances:** Use the Euclidean distance formula to calculate the distance from each point to all centroids.
2. **Assign Clusters:** Each point is assigned to the cluster of the nearest centroid.

Results:

- Distances and cluster assignments are shown in the table below.

Point	Coordinates	Dist Mean 1 ($\sqrt{(2-x)^2 + (10-y)^2}$)	Dist Mean 2 ($\sqrt{(5-x)^2 + (8-y)^2}$)	Dist Mean 3 ($\sqrt{(1-x)^2 + (2-y)^2}$)	Cluster
A1	(2, 10)	$\sqrt{(2-2)^2 + (10-10)^2} = 0.00$	$\sqrt{(5-2)^2 + (8-10)^2} = 3.61$	$\sqrt{(1-2)^2 + (2-10)^2} = 8.25$	1
A2	(2, 5)	$\sqrt{(2-2)^2 + (10-5)^2} = 5.00$	$\sqrt{(5-2)^2 + (8-5)^2} = 3.61$	$\sqrt{(1-2)^2 + (2-5)^2} = 3.16$	3
A3	(8, 4)	$\sqrt{(2-8)^2 + (10-4)^2} = 9.43$	$\sqrt{(5-8)^2 + (8-4)^2} = 5.00$	$\sqrt{(1-8)^2 + (2-4)^2} = 7.28$	2
A4	(5, 8)	$\sqrt{(2-5)^2 + (10-8)^2} = 3.61$	$\sqrt{(5-5)^2 + (8-8)^2} = 0.00$	$\sqrt{(1-5)^2 + (2-8)^2} = 6.71$	2
A5	(7, 5)	$\sqrt{(2-7)^2 + (10-5)^2} = 8.06$	$\sqrt{(5-7)^2 + (8-5)^2} = 3.61$	$\sqrt{(1-7)^2 + (2-5)^2} = 6.08$	2
A6	(6, 4)	$\sqrt{(2-6)^2 + (10-4)^2} = 8.94$	$\sqrt{(5-6)^2 + (8-4)^2} = 4.47$	$\sqrt{(1-6)^2 + (2-4)^2} = 5.39$	2
A7	(1, 2)	$\sqrt{(2-1)^2 + (10-2)^2} = 8.00$	$\sqrt{(5-1)^2 + (8-2)^2} = 7.62$	$\sqrt{(1-1)^2 + (2-2)^2} = 1.00$	3
A8	(4, 9)	$\sqrt{(2-4)^2 + (10-9)^2} = 2.24$	$\sqrt{(5-4)^2 + (8-9)^2} = 1.41$	$\sqrt{(1-4)^2 + (2-9)^2} = 7.07$	2

Iteration 2

Steps:

1. **Recalculate Centroids:** For each cluster, compute the new centroid as the mean of all points in that cluster.
For each cluster, the new centroid is calculated as the mean of the (x)- and (y)-coordinates of all points in the cluster:

$$\text{Centroid}_i = \left(\frac{\sum x_{\text{cluster}}}{n}, \frac{\sum y_{\text{cluster}}}{n} \right)$$

- **New Centroids:** (2.0, 10.0), (6.0, 6.0), (1.5, 3.5)
- 2. **Reassign Clusters:** Recompute distances to the updated centroids and assign points to the nearest cluster.

Point	Coordinates	Dist Mean 1 ($\sqrt{(x1-x)^2 + (y1-y)^2}$)	Dist Mean 2 ($\sqrt{(x2-x)^2 + (y2-y)^2}$)	Dist Mean 3 ($\sqrt{(x3-x)^2 + (y3-y)^2}$)	Cluster
A1	(2, 10)	$\sqrt{(2-2)^2 + (10-10)^2} = 0.00$	$\sqrt{(6-2)^2 + (6-10)^2} = 5.39$	$\sqrt{(1.5-2)^2 + (3.5-10)^2} = 8.83$	1
A2	(2, 5)	$\sqrt{(2-2)^2 + (10-5)^2} = 5.00$	$\sqrt{(6-2)^2 + (6-5)^2} = 4.47$	$\sqrt{(1.5-2)^2 + (3.5-5)^2} = 2.50$	3

Point	Coordinates	Dist Mean 1 ($\sqrt{((x1-x)^2 + (y1-y)^2)}$)	Dist Mean 2 ($\sqrt{((x2-x)^2 + (y2-y)^2)}$)	Dist Mean 3 ($\sqrt{((x3-x)^2 + (y3-y)^2)}$)	Cluster
A3	(8, 4)	$\sqrt{((2-8)^2 + (10-4)^2)} = 9.22$	$\sqrt{((6-8)^2 + (6-4)^2)} = 3.61$	$\sqrt{((1.5-8)^2 + (3.5-4)^2)} = 6.86$	2
A4	(5, 8)	$\sqrt{((2-5)^2 + (10-8)^2)} = 3.61$	$\sqrt{((6-5)^2 + (6-8)^2)} = 2.24$	$\sqrt{((1.5-5)^2 + (3.5-8)^2)} = 6.92$	2
A5	(7, 5)	$\sqrt{((2-7)^2 + (10-5)^2)} = 8.06$	$\sqrt{((6-7)^2 + (6-5)^2)} = 2.24$	$\sqrt{((1.5-7)^2 + (3.5-5)^2)} = 5.92$	2
A6	(6, 4)	$\sqrt{((2-6)^2 + (10-4)^2)} = 8.94$	$\sqrt{((6-6)^2 + (6-4)^2)} = 2.83$	$\sqrt{((1.5-6)^2 + (3.5-4)^2)} = 5.22$	2
A7	(1, 2)	$\sqrt{((2-1)^2 + (10-2)^2)} = 8.06$	$\sqrt{((6-1)^2 + (6-2)^2)} = 7.21$	$\sqrt{((1.5-1)^2 + (3.5-2)^2)} = 1.80$	3
A8	(4, 9)	$\sqrt{((2-4)^2 + (10-9)^2)} = 2.24$	$\sqrt{((6-4)^2 + (6-9)^2)} = 3.61$	$\sqrt{((1.5-4)^2 + (3.5-9)^2)} = 6.80$	1

Comments:

- Notice how centroids shift toward the center of their clusters.
- Some points might change clusters as centroids update.

Iteration 3

Steps:

1. **Recompute Centroids:** Repeat the process of recalculating centroids.
 - **New Centroids:** (3.0, 9.5), (6.5, 5.25), (1.5, 3.5)
2. **Reassign Clusters:** Update cluster assignments based on new centroids.

Point	Coordinates	Dist Mean 1	Dist Mean 2	Dist Mean 3	Cluster
A1	(2, 10)	1.12	5.94	8.83	1
A2	(2, 5)	4.53	4.18	2.50	3
A3	(8, 4)	8.86	3.00	6.86	2
A4	(5, 8)	2.55	2.28	6.92	2
A5	(7, 5)	7.90	2.05	5.92	2
A6	(6, 4)	8.55	2.65	5.22	2
A7	(1, 2)	7.55	7.07	1.80	3
A8	(4, 9)	1.80	3.74	6.80	1

Comments:

- Observe how centroids continue to adjust.
- Points begin to stabilize in their respective clusters.

Iteration 4

Steps:

1. **Recalculate Centroids:** Adjust centroids as the mean of their cluster points.
2. **Check for Convergence:** If centroids do not change, the algorithm stops.

Results:

- Distances and final cluster assignments are shown in the table below.
- **Final Centroids:** (3.67, 9.0), (7.0, 4.33), (1.5, 3.5)

Point	Coordinates	Dist Mean 1	Dist Mean 2	Dist Mean 3	Cluster
A1	(2, 10)	1.60	6.12	8.83	1
A2	(2, 5)	4.12	4.01	2.50	3
A3	(8, 4)	8.64	2.62	6.86	2
A4	(5, 8)	2.36	2.03	6.92	2
A5	(7, 5)	7.68	1.75	5.92	2
A6	(6, 4)	8.28	2.35	5.22	2
A7	(1, 2)	7.23	7.07	1.80	3
A8	(4, 9)	1.41	3.77	6.80	1

Comments:

- The algorithm converges as centroids stabilize and no further changes occur.

Final Results

- Each point has been assigned to its nearest cluster.
- The process has converged after 4 iterations.

Key Observations

1. **Centroid Movement:** Centroids progressively adjust to the cluster's center during each iteration.
2. **Convergence:** K-Means stops when the centroids no longer move.

6. Python Code Example

Here's how to implement K-Means clustering in Python using `scikit-learn` and Mall Customers Information. We will use a using [Mall Customers Dataset](#) to demonstrate kmeans clustering using scikit Learn Library.

Index	CustomerID	Genre	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	Female	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72
10	11	Male	67	19	14
11	12	Female	35	20	99
12	13	Male	58	20	15
13	14	Female	24	20	77
14	15	Male	22	20	79

Import Libraries and Dataset

```

1 # Import necessary libraries
2 import pandas as pd
3 from sklearn.cluster import KMeans
4 import matplotlib.pyplot as plt
5
6 # Step 1: Get the data
7 df = pd.read_csv('https://raw.githubusercontent.com/msfasha/307304-Data-Mining/refs/heads/main/datasets/mall_customers.csv')
8 df

```

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

Apply K-Means with K = 3

```

1 # Step 2: Select relevant features (Annual Income and Spending Score) for clustering
2 features = df[["Annual Income (k$)", "Spending Score (1-100)"]]
3
4 # Step 3: Apply K-Means with K=3
5 kmeans = KMeans(n_clusters=3, random_state=42)
6 kmeans.fit(features)
7
8 # Step 4: Add the cluster labels to the DataFrame
9 df["Cluster"] = kmeans.labels_

```

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)	Cluster
0	1	Male	19	15	39	2
1	2	Male	21	15	81	2
2	3	Female	20	16	6	2
3	4	Female	23	16	77	2
4	5	Female	31	17	40	2
...
195	196	Female	35	120	79	1
196	197	Female	45	126	28	0
197	198	Male	32	126	74	1
198	199	Male	32	137	18	0
199	200	Male	30	137	83	1

200 rows × 6 columns

▼ Print the updated dataframe showing the assigned cluster for each record

1 df

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)	Cluster
0	1	Male	19	15	39	2
1	2	Male	21	15	81	2
2	3	Female	20	16	6	2
3	4	Female	23	16	77	2
4	5	Female	31	17	40	2
...
195	196	Female	35	120	79	1
196	197	Female	45	126	28	0
197	198	Male	32	126	74	1
198	199	Male	32	137	18	0
199	200	Male	30	137	83	1

200 rows × 6 columns

▼ Print out the centroids of the clusters

1 kmeans.cluster_centers_

```

array([[87.          , 18.63157895],
       [86.53846154, 82.12820513],
       [44.15447154, 49.82926829]])

```

Format centroids in a dataframe for better visualization

```

1 # Assuming features and KMeans object already defined
2 centroids = pd.DataFrame(kmeans.cluster_centers_, columns=features.columns)
3

```

```

4 # Display centroids for interpretation
5 print("Centroids for each cluster:")
6 print(centroids)

```

```

Centroids for each cluster:
  Annual Income (k$)  Spending Score (1-100)
0          87.000000          18.631579
1          86.538462          82.128205
2          44.154472          49.829268

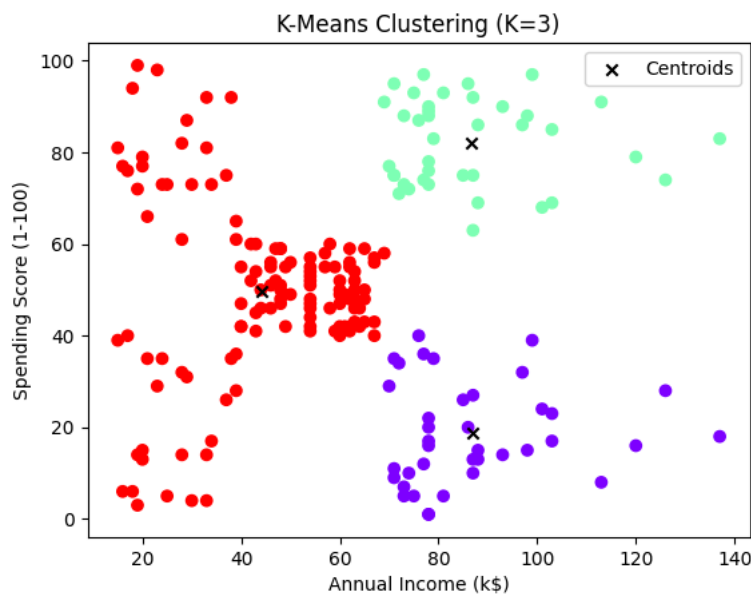
```

▼ Display The Scatter Plot and the DataFrame

```

1 # Step 5: Plot the clusters and centroids
2 plt.scatter(features["Annual Income (k$)"], features["Spending Score (1-100)"], c=df["Cluster"], cmap="rainbow")
3 plt.scatter(kmeans.cluster_centers_[0, 0], kmeans.cluster_centers_[0, 1], color="black", marker="x", label="Centroids")
4 plt.title("K-Means Clustering (K=3)")
5 plt.xlabel("Annual Income (k$)")
6 plt.ylabel("Spending Score (1-100)")
7 plt.legend()
8 plt.show()

```



Explanation:

- **Step 1:** We create a simple 2D dataset.
- **Step 2:** We apply K-Means clustering with $K = 3$.
- **Step 3:** We assign each data point to a cluster and store the result in the DataFrame.
- **Step 4:** We plot the clusters with different colors and mark the centroids with black 'x's.

Observations for (k=3):

1. Cluster Formation:

- The data points are divided into **three distinct clusters**, differentiated by their spending scores and annual incomes.
- The clusters reflect distinct patterns in customer behavior.

2. Cluster Characteristics:

- **Cluster 1 (e.g., Red):**
 - Represents customers with **low to moderate income** and a **wide range of spending scores**.
 - Includes customers who might be more price-conscious or have varied spending behaviors.
- **Cluster 2 (e.g., Purple):**
 - Comprises customers with **high income** but **low spending scores**.
 - Indicates potentially frugal or savings-oriented individuals, despite higher incomes.
- **Cluster 3 (e.g., Green):**

- Includes customers with **high income** and **high spending scores**.
- Likely represents high-value customers who are ideal for premium marketing campaigns.

3. Centroid Placement:

- The centroids (black "X" markers) represent the mean position of each cluster.
- The distance between centroids suggests clear separation between the groups.

4. Insights:

- This segmentation aligns well with typical customer segmentation strategies in marketing:
 - **Cluster 1:** Target with discounts or budget-friendly options.
 - **Cluster 2:** Upsell higher-end products or encourage spending.
 - **Cluster 3:** Retain as loyal high-value customers with exclusive offers.

5. Potential Next Steps:

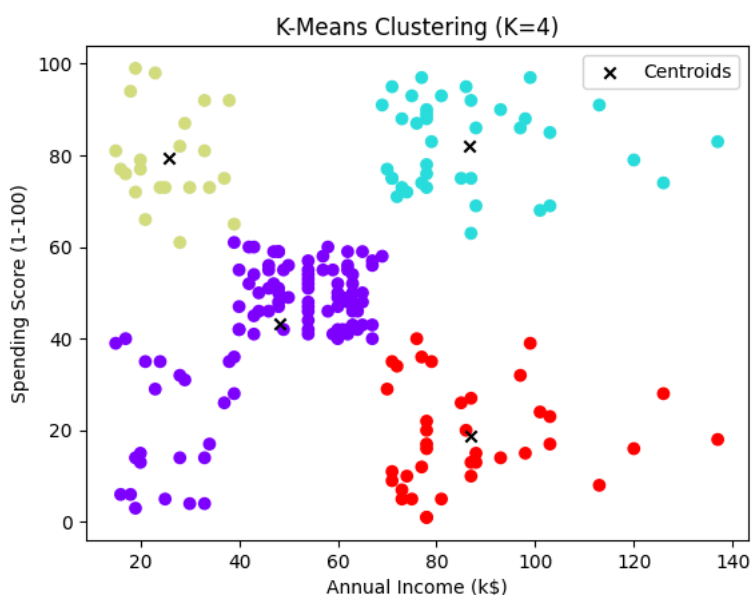
- Validate the clusters with real-world customer behavior (e.g., demographic or shopping preferences).
- Experiment with **k=4** to check for further sub-clusters.
- Perform **3D clustering** by adding another feature (e.g., Age or Gender).

▼ Apply K-Means with K = 4

```

1 # Select relevant features (Annual Income and Spending Score) for clustering
2 features = df[["Annual Income (k$)", "Spending Score (1-100)"]]
3
4 # Apply K-Means with K=4
5 kmeans = KMeans(n_clusters=4, random_state=42)
6 kmeans.fit(features)
7
8 # Add assigned clusters to each row
9 df["Cluster"] = kmeans.labels_
10
11 # Plot the clusters and centroids
12 plt.scatter(features["Annual Income (k$)"], features["Spending Score (1-100)"], c=df["Cluster"], cmap="rainbow")
13 plt.scatter(kmeans.cluster_centers_[0], kmeans.cluster_centers_[1], color="black", marker="x", label="Centroids")
14 plt.title("K-Means Clustering (K=4)")
15 plt.xlabel("Annual Income (k$)")
16 plt.ylabel("Spending Score (1-100)")
17 plt.legend()
18 plt.show()

```



With (k=4) clusters, the customer segmentation becomes more refined. Here are the observations and insights:

Observations

1. Cluster Formation:

- The data is divided into four distinct clusters, with clear separation along **Annual Income (k\$)** and **Spending Score (1-100)**.

2. Cluster Characteristics:

- **Cluster 1 (Light Green):**
 - **Low Annual Income, High Spending Score.**
 - Represents customers who spend a large proportion of their income, possibly impulsive spenders or younger customers.
- **Cluster 2 (Purple):**
 - **Moderate Annual Income, Moderate Spending Score.**
 - Likely represents the "average" customer with balanced income and spending behavior.
- **Cluster 3 (Light Blue):**
 - **High Annual Income, High Spending Score.**
 - Represents premium, high-value customers with significant purchasing power.
- **Cluster 4 (Purple):**
 - **High Annual Income, Low Spending Score.**
 - Likely represents frugal or savings-oriented individuals who may prioritize saving over spending despite their high income.

3. Centroids:

- The centroids (black "X" markers) represent the average characteristics of each cluster.
- Clear separation of centroids confirms meaningful clusters.

▼ Insights

1. Marketing Strategies:

- **Cluster 1 (Light Green):**
 - Target with discounts, promotions, or budget-friendly offers to maximize spending potential.
- **Cluster 2 (Purple):**
 - Maintain regular engagement, as they represent a balanced customer segment.
- **Cluster 3 (Light Blue):**
 - Focus on premium products and services. These customers are ideal for exclusive campaigns.
- **Cluster 4 (Purple):**
 - Consider introducing tailored promotions or incentivizing spending for this segment.

2. Segmentation Insights:

- Including (k=4) provides finer segmentation, highlighting distinct behavioral groups like "impulsive spenders" and "savings-oriented high-income customers."
- This segmentation can inform personalized marketing and product offerings.

3. Potential for Further Analysis:

- **3D Clustering:** Add a third feature like **Age** to see if it further refines the clusters.
- **Exploring Gender:** Overlay gender information to analyze behavioral patterns by gender.

▼ 8. Cluster Data Using 3 Features

```
1 # Step 2: Select relevant features (Annual Income and Spending Score) for clustering
2 features = df[["Annual Income (k$)", "Spending Score (1-100)", "Age"]]
3
4 # Step 3: Apply K-Means with K=3
5 kmeans = KMeans(n_clusters=3, random_state=42)
6 kmeans.fit(features)
7
8 # Step 4: Add the cluster labels to the DataFrame
9 df["Cluster"] = kmeans.labels_
```

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)	Cluster
0	1	Male	19	15	39	2
1	2	Male	21	15	81	2
2	3	Female	20	16	6	0
3	4	Female	23	16	77	2
4	5	Female	31	17	40	2
...
195	196	Female	35	120	79	1
196	197	Female	45	126	28	1
197	198	Male	32	126	74	1
198	199	Male	32	137	18	1
199	200	Male	30	137	83	1

200 rows × 6 columns

Display Centroids

```

1 centroids = pd.DataFrame(kmeans.cluster_centers_, columns=features.columns)
2
3 # Display centroids for interpretation
4 print("Centroids for each cluster:")
5 print(centroids)

```

```

Centroids for each cluster:
   Annual Income (k$)  Spending Score (1-100)    Age
0          59.879032          35.427419  44.483871
1          88.731707          79.243902  32.975610
2          29.971429          68.514286  25.771429

```

3D Scatter Plot

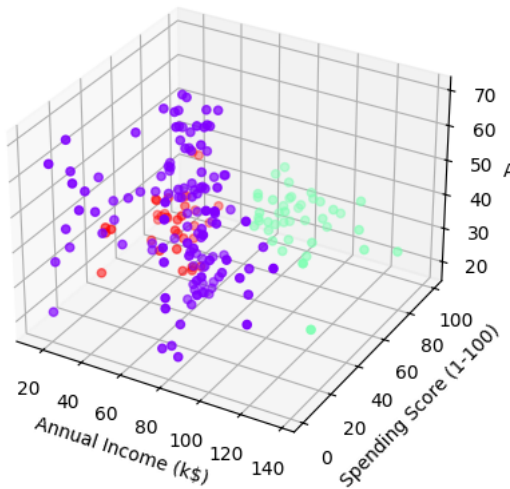
We can use a 3D scatter plot to visualize three features together.

```

1 from mpl_toolkits.mplot3d import Axes3D
2
3 fig = plt.figure()
4 ax = fig.add_subplot(111, projection='3d')
5
6 # Plotting points
7 ax.scatter(
8     df["Annual Income (k$)"],
9     df["Spending Score (1-100)"],
10    df["Age"],
11    c=df["Cluster"], cmap="rainbow"
12 )
13
14 # Label axes
15 ax.set_title("K-Means Clustering (3 Features)")
16 ax.set_xlabel("Annual Income (k$)")
17 ax.set_ylabel("Spending Score (1-100)")
18 ax.set_zlabel("Age")
19 plt.show()

```

K-Means Clustering (3 Features)



▼ Use Plotly Package to Create Dynamic Plots

```

1  import plotly.express as px
2
3  # Create an interactive 3D scatter plot
4  fig = px.scatter_3d(
5      df,
6      x="Annual Income (k$)",
7      y="Spending Score (1-100)",
8      z="Age",
9      color="Cluster", # Color points by cluster
10     title="K-Means Clustering (3 Features)",
11     labels={
12         "Annual Income (k$)": "Annual Income (k$)",
13         "Spending Score (1-100)": "Spending Score (1-100)",
14         "Age": "Age"
15     },
16     width=1000, # Set the width of the plot
17     height=800  # Set the height of the plot
18 )
19
20 # Show the interactive plot
21 fig.show()
22

```

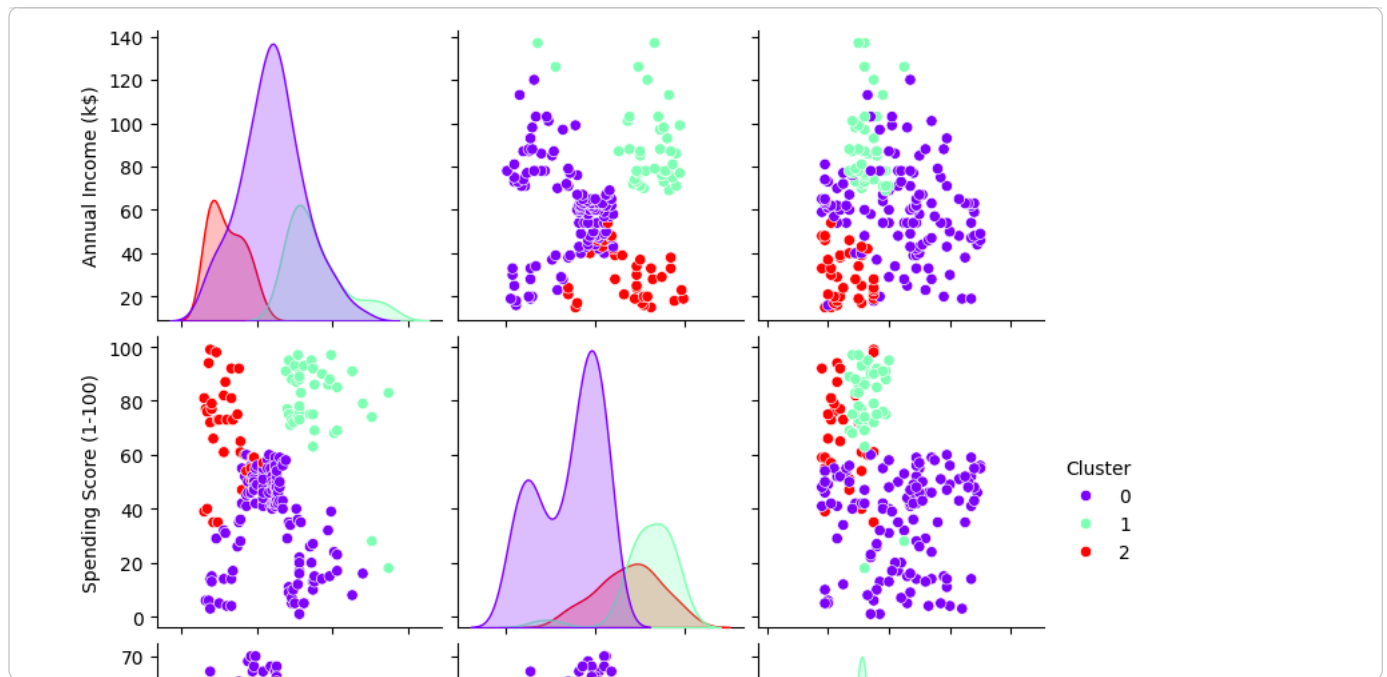
▼ Pairwise Scatter Plots

We can create scatter plots for each pair of the three features, highlighting clusters using different colors.

```

1  import seaborn as sns
2
3  # Add Cluster information to the dataframe
4  sns.pairplot(df, vars=["Annual Income (k$)", "Spending Score (1-100)", "Age"], hue="Cluster", palette="rainbow")
5  plt.show()
6

```



This pair plot visualizes the relationships between **Annual Income (k\$)**, **Spending Score (1-100)**, and **Age** across the identified customer clusters.

Observations

1. Cluster Separation:

Cluster 0 (Purple):

- Spreads across a wide range of **Annual Income**, but tends to have **low Spending Scores**.
- Includes a mix of ages, but younger individuals dominate.

Cluster 1 (Green):

- Consists of customers with **high Spending Scores** and moderate to high **Annual Income**.
- Likely includes customers who spend consistently regardless of age.

Cluster 2 (Red):

- Concentrated among customers with **low Annual Income** and **high Spending Scores**.
- Mostly young customers, indicating impulsive or budget-focused spenders.

2. Feature Interactions:

Annual Income vs. Spending Score:

- Cluster 0 includes customers with higher income but lower spending.
- Cluster 1 indicates high-spending customers across various income levels.
- Cluster 2 reflects high spending at lower income levels, highlighting a younger demographic.