Customer Segmentation with Unsupervised Clustering

Victor Jansen | AI2 | 25 – 03 - 2025

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

Understanding the Data

The dataset used for this analysis consists of transactional records from an online retail store, spanning from **December 1, 2010** to **December 9, 2011**. It includes **18,532 unique orders** placed by **4,338 customers** across **37 countries**. A total of **3,665 different products** were sold during this period, generating a combined revenue of approximately **€8.9 million**.

Data Cleaning & Preperation

- Removed rows with missing CustomerID
  - Required for customer-level clustering
- Removed rows with negative or zero values
  - To exclude returns, cancellations, and invalid transactions
- Dropped rows with missing Description
  - For dataset consistency
- Converted InvoiceDate to datetime format
  - Required for calculating **Recency**
- Reset DataFrame index
  - For clean structure after filtering

Feature Engineering

Create a new **total price** column

- Quantity * UnitPrice

Created **RFM features** for each customer:

- **Recency**: Days since last purchase
  - Days between the most recent purchase and a fixed reference date (the day after the latest invoice)
- **Frequency**: Number of unique orders
  - Count of unique InvoiceNo per customer
- **Monetary**: Total spending across all orders
  - Sum of TotalPrice per customer

Feature Scaling for Clustering

Clustering is distance-based

larger values dominate the result

We normalized the data to make all features contribute equally

Clustering method

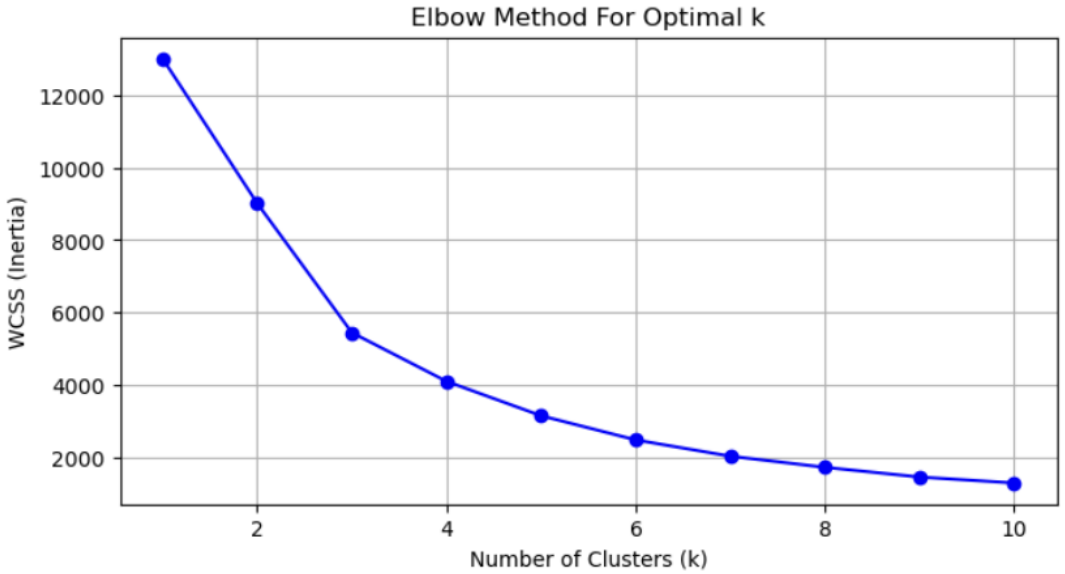
K-Means Clustering (Primary Method)

- **Why?**
  - Scalable to large datasets (~4,000+ customers)
  - Easy to interpret and visualize
  - Works well with RFM data (numerical, normalized)
- **Distance-based** using Euclidean distance
- Fast and efficient

Hierarchical Clustering (Comparison)

- **Why?**
  - Doesn't require choosing k in advance
  - Dendrogram provides intuitive visual for cluster selection
  - Good for verifying cluster structure found by K-Means
- **Why not primary method?**
  - Less scalable — $O(n^2)$ complexity not ideal for large datasets

Determining K Clusters

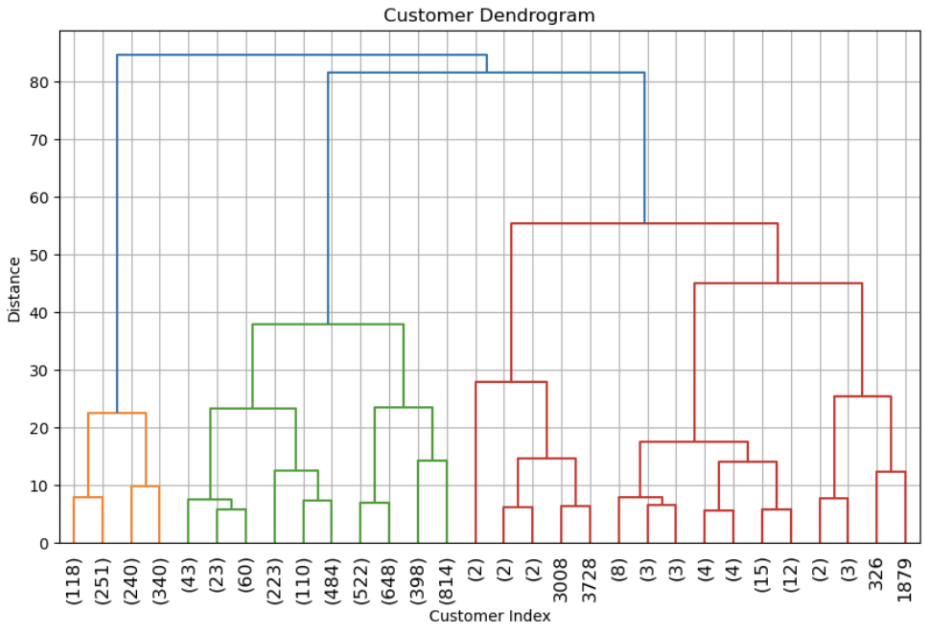


Elbow Method for optimal K

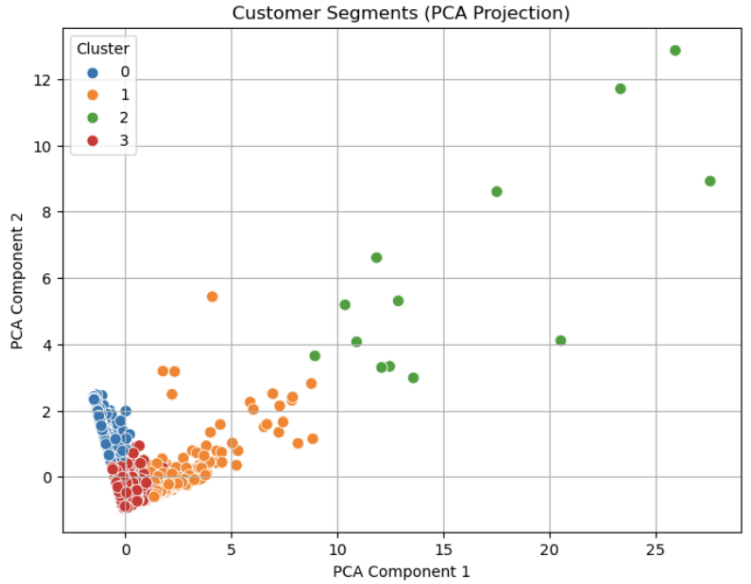
- Plotted Within-Cluster Sum of Squares (WCSS) for k = 1 to 10
- Sharp decrease in WCSS levels off around **k = 4**
- Indicates that 4 clusters balance model complexity and compactness
- Supports selection of **k = 4** for K-Means

Dendrogram for Hierarchical Clustering

- Shows how customer clusters merge at increasing distance levels
- A horizontal cut at distance ≈ 50 yields **4 distinct clusters**
- Confirms the structure observed with the Elbow Method

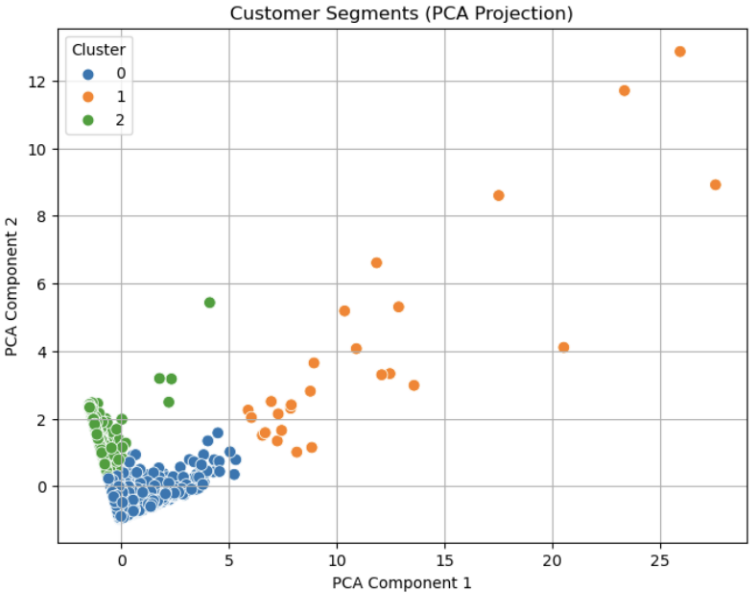


Comparing k Clusters



3 vs 4 k Cluster

When comparing k = 3 vs. k = 4, we observed that the fourth cluster in the upper plot (red) captures a distinct group that is otherwise merged into a broader segment in the 3-cluster version. This supports our decision to use 4 clusters, enabling more precise and actionable segmentation



Interpreting clusters

Cluster	Recency	Frequency	Monetary	Count
1	248.6	1.6	$478.1	1062
2	15.7	22.0	$12453.2	211
3	7.4	82.5	$127338.3	13
4	43.9	3.7	$1350.1	3052

1 - Dormant Customers

Low frequency and haven't purchased in a long time

2 - Loyal high-value customers

Highly engaged and profitable group

segments

3 - VIP/ whales

Tiny segment wit massive value

4 - Casual Shoppers

Massive segment, with moderate value

Marketing Strategies – Dormant Customer

GOAL: Reengage customers

Strategy: Reactivation campaigns

Time-limited discounts	“We miss you” emails	Product recommendations based on last purchase

Marketing Strategies – Loyal high-value customer

GOAL: Retain and Reward loyalty

Strategy: Loyalty & retention

Loyalty perks	Exclusive offers	birthday gifts

Marketing Strategies – VIP/ Whales

GOAL: Retain and Protect relationships

Strategy: Personalisation & care

Dedicated account manager	Elite rewards	Exclusive private sales

Marketing Strategies – Casual shoppers

GOAL: Grow spending

Strategy: Nurturing & cross-sell

Bundle offers	Product suggestions	Free shipping thresholds