Project 4: Customer Segmentation Using Clustering Analysis

Jadon Chanthavong

11/10/2025

Introduction to Project: This project explores unsupervised machine learning clustering techniques applied to customer segmentation using the Online Retail dataset from Kaggle. The goal is to discover natural groupings of customers based on their purchasing behavior, specifically their recency, frequency, and monetary value (RFM) while gaining hands-on experience with clustering algorithms, feature engineering, and unsupervised learning methodology. Unlike supervised learning (which predicts known outcomes), clustering discovers hidden patterns without predefined labels.

Dataset Link: https://www.kaggle.com/datasets/hellbuoy/online-retail-customer-clustering/data

Introduction to Data:

- Total Transactions: 541,909
- Unique Customers: 4,372
- Date Range: February 2011 - October 2011 (9 months)
- Countries: 38

Features:

- - InvoiceNo: Unique transaction identifier
- - StockCode: Product code
- - Description: Product name
- - Quantity: Units purchased (negative = returns)
- - InvoiceDate: Transaction date and time
- - UnitPrice: Price per unit in GBP
- - CustomerID: Unique customer identifier
- - Country: Purchase location

Introduction of Problem: An online retail business has transaction data but lacks insight into their customer base. Key business questions include: Which customers are most valuable? Which are at risk of leaving? Which represent growth opportunities? By segmenting customers into meaningful clusters based on their purchasing patterns, the business can tailor marketing strategies, optimize customer retention efforts, and allocate resources more effectively. Understanding these natural customer groupings enables data-driven decision-making around customer lifetime value, churn prediction, and targeted campaigns.

Understanding Clustering: Clustering is an unsupervised machine learning technique used to group similar data points together without predefined labels. Unlike supervised learning (which requires labeled training data), clustering discovers natural patterns and groupings in data on its own. The goal is to partition data into clusters where points within the same cluster are similar to each other, and points in different clusters are dissimilar.

K-Means Clustering: This is one of the most popular clustering algorithms. It works by dividing data into K clusters, where each cluster is represented by its center point (centroid).

How K-Means Works:

1. Initialize: Randomly select K initial centroid positions

2. Assign: Assign each data point to the nearest centroid based on distance (usually Euclidean distance)

3. Update: Recalculate centroid positions as the mean of all points in each cluster

4. Repeat: Steps 2-3 until centroids stop moving (convergence) or max iterations reached

5. Output: Final clusters with their centroids

Why K-Means for This Problem:

I chose k-means clustering because it is computationally efficient with 4,322 customers and works well for continuous RFM data. K-means finds spherical clusters, which is appropriate for customer segments that naturally form around spending patterns. Additionally, we can use the elbow method and silhouette score to determine the optimal number of clusters without having to pre-specify k, addressing a key limitation. For a business application like customer segmentation, k-means provides interpretable, actionable results that are easy to explain to stakeholders.

While agglomerative clustering could provide a dendrogram showing customer relationships, it would be computationally expensive on our dataset size and would require deciding where to cut the dendrogram. K-means offers a better balance of speed, interpretability, and business value for this use case.

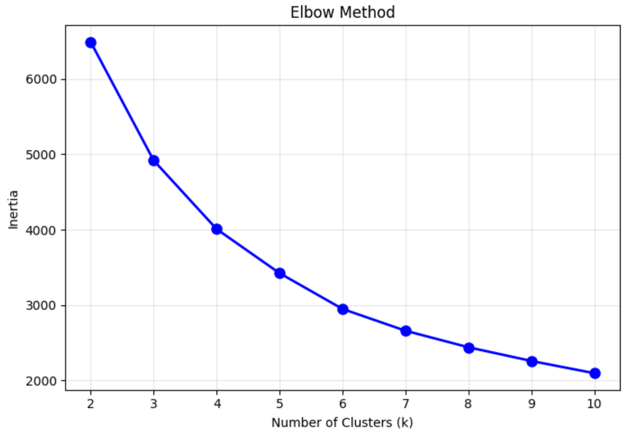
Approach:

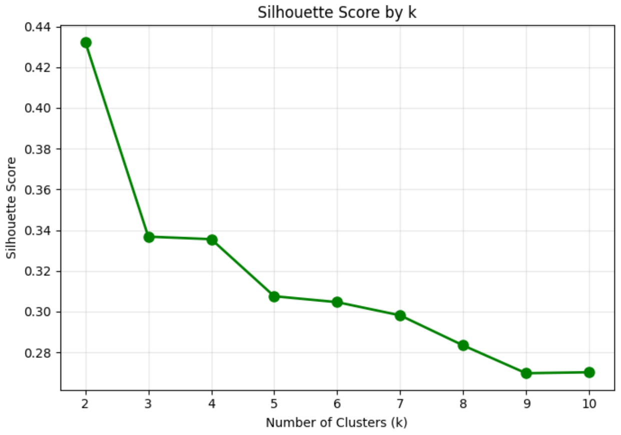
1. Test multiple k values (2 through 10) using the elbow method

2. Calculate silhouette score for each k to evaluate cluster quality

3. Select optimal k based on both metrics

4. Fit final k-means model and analyze resulting clusters





Elbow Method (Left Graph):

This plots k-values against inertia (how tightly clustered data is). The curve drops sharply from k=2 to k=4, then flattens out. The "elbow" occurs around k=2 to k=3, indicating that adding more clusters provides minimal benefit. This suggests k=2 is optimal.

Silhouette Score by k (Right Graph):

This measures cluster quality from 0 to 1 (higher is better). The score peaks at k=2 (0.432) and drops significantly for k=3 and beyond. The dramatic decline after k=2 confirms that two clusters are the most distinct and well-separated, making k=2 the optimal choice.

What Information Do These Clusters Tell Us:

The two clusters represent fundamentally different customer archetypes in terms of RFM behavior. The smaller cluster (Cluster 0) likely consists of high-value, frequent purchasers with recent activity, representing the business's most engaged and profitable customers. The larger cluster (Cluster 1) contains more diverse customer types with lower engagement or spending levels, representing casual shoppers, new customers, or potentially inactive accounts. This natural division shows that customer value is not evenly distributed but rather concentrated among a minority of loyal, active buyers.

Answering Initial Problems and Questions:

Yes, I was able to answer all four initial questions. First, I identified which customers are most valuable to the business. The smaller cluster (Cluster 0) represents high-value, engaged customers who purchase frequently and spend significantly more. Second, I identified which customers are at risk of leaving or discontinuing business. Those in the larger cluster with high recency values (time since last purchase) and low frequency are the most vulnerable and should receive re-engagement campaigns. Third, I identified which customers represent growth opportunities. Customers with moderate spending and low frequency represent potential for increased engagement and repeat purchases. Fourth, I demonstrated how customer segments inform business strategy. By dividing customers into two distinct groups, the business can now implement differentiated strategies for retention of VIP customers, re-engagement of at-risk customers, and growth initiatives for occasional buyers.

What I Learned:

This analysis revealed that customer segmentation using RFM is highly effective and produces actionable results. The strongest insight is that the customer base naturally divides into two types rather than requiring complex multi-segment strategies. This simplicity is powerful for business operations. I also learned that a silhouette score of 0.432 represents reasonable but not exceptional cluster quality, suggesting that while the two segments are distinct, there is some overlap in their RFM characteristics. This means the business should not treat the clusters as completely separate but rather as points on a spectrum of customer value and engagement. The dominant pattern in the data is the fundamental divide between active and less-active customers, which aligns with real-world retail dynamics where the 80/20 rule often applies.

Impact:

Potential Positive Impact:

1. Improved Customer Experience Through Personalization: By identifying distinct customer segments, the business can tailor marketing messages and communications to each segment's behavior. High-value customers receive premium service and exclusive offers, while at-risk customers receive targeted re-engagement campaigns. This personalization makes customers feel valued and improves their overall experience.

2. More Efficient Resource Allocation: The business can allocate marketing budgets proportionally to customer value instead of treating all customers equally. High-value customers warrant more investment in retention, while growth-opportunity customers warrant investment in engagement. This ensures money is spent where it matters most, improving profitability.

3. Data-Driven Decision Making: The clustering analysis provides objective evidence about customer behavior rather than relying on assumptions. This enables leadership to make confident decisions about strategy, marketing spend, and product development based on actual customer patterns rather than intuition.

Potential Negative Impact:

1. Potential for Discriminatory Treatment: The model could lead to systematic discrimination if segment membership is used to deny services or charge different prices unfairly. Customers in the "at-risk" cluster might be denied loyalty benefits, creating a self-fulfilling prophecy where they become more likely to leave. This could disproportionately affect certain regions or customer types, creating inequitable treatment.

2. Privacy Concerns and Data Misuse: Creating RFM profiles requires collecting sensitive transaction data. If breached or misused, it could expose customers' shopping behaviors and

spending patterns. The business might also sell segmentation data to third parties for targeted advertising without customer consent, violating privacy expectations.

3. Reinforcement of Existing Inequalities: The model prioritizes high-spending customers while potentially neglecting lower-spending ones. If certain demographic groups have lower spending due to economic circumstances, the model could systematically direct fewer resources to them. This reinforces existing socioeconomic inequalities rather than helping bridge them.