Market Segmentation Project

For online retail store

Xiwen Mark

Table of Contents

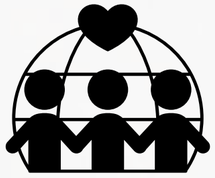
Xiwen Mark

Project Objectives

1. Understanding current customer base
2. Segment customers into groups based on purchasing behavior to target with better business strategies
3. Uncover deeper insights and the key pain points of current customers relationship management

Understanding Data

This dataset contains transactional records from a UK-based, non-store online retailer, covering all transactions between December 1, 2010, and December 9, 2011.



There are total 8 variables: {

InvoiceNo: a 6-digit integral number uniquely assigned to each transaction;

StockCode: a 5-digit integral number uniquely assigned to each distinct product;

Description: product name;

Quantity: the quantities of each product (item) per transaction

InvoiceDate: the day and time when each transaction was generated

UnitPrice: product price per unit;

CustomerID: a 5-digit integral number uniquely assigned to each customer

Country: the name of the country where each customer resides}

Data source: *UCI Machine Learning Repository*. (n.d.). https://archive.ics.uci.edu/dataset/352/online+retail

Analytical Approach

Segmentation Framework

- Used RFM (Recency, Frequency, Monetary) to measure customer behavior and value

Feature Engineering

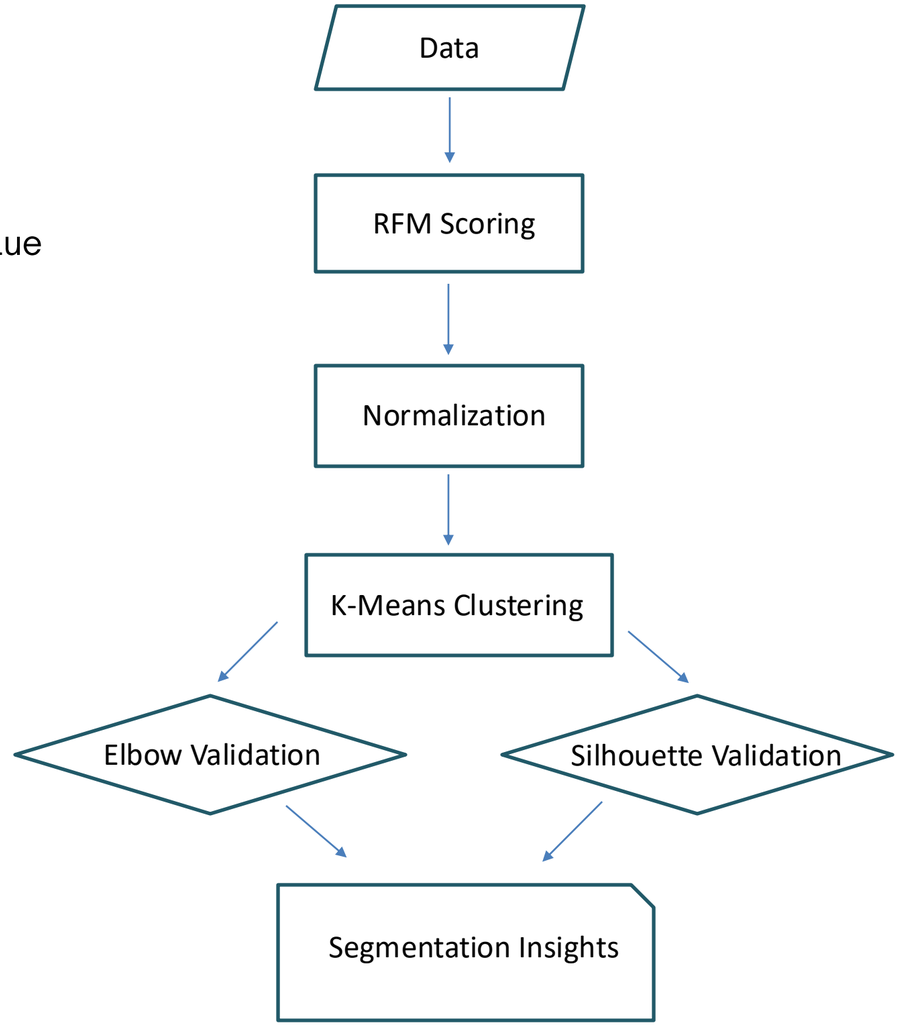
- Calculated RFM scores for each customer
- Normalized data for consistent scale across variables

Clustering Methodology

- Applied K-Means clustering for customer segmentation
- Determined optimal cluster number using
  - **Elbow Method** – minimizes within-cluster variance
  - **Silhouette Score** – measures cohesion and separation

Validation & Insights

- Interpreted clusters for business relevance
- Linked segments to actionable marketing strategies



EDA - Top Selling and Most Popular Products

Recognizing the top-selling and most popular products is crucial for understanding customer expectations and defining the company's market positioning. It helps identify what drives overall sales and what products customers purchase most frequently.

Top 10 Selling Products and Their Sales Share:

	Description	Sales Share
1	PAPER CRAFT , LITTLE BIRDIE	1.890
2	REGENCY CAKESTAND 3 TIER	1.600
3	WHITE HANGING HEART T-LIGHT HOLDER	1.127
4	JUMBO BAG RED RETROSPOT	0.956
5	MEDIUM CERAMIC TOP STORAGE JAR	0.914
6	POSTAGE	0.873
7	PARTY BUNTING	0.773
8	ASSORTED COLOUR BIRD ORNAMENT	0.635
9	Manual	0.603
10	RABBIT NIGHT LIGHT	0.576

Top 10 Products Generated the Most Quantity:

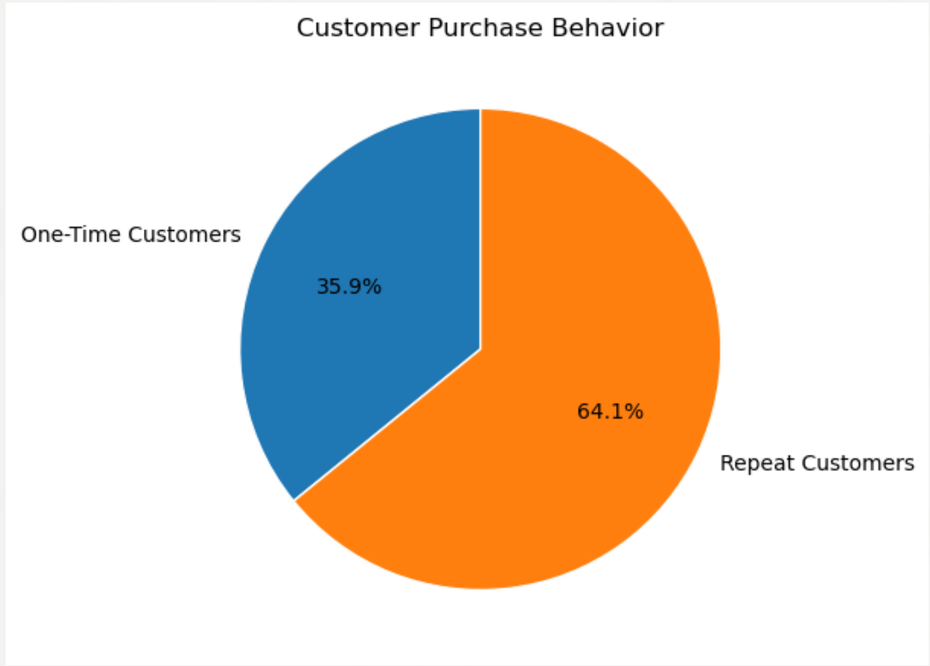
	Description
1	PAPER CRAFT , LITTLE BIRDIE
2	MEDIUM CERAMIC TOP STORAGE JAR
3	WORLD WAR 2 GLIDERS ASSTD DESIGNS
4	JUMBO BAG RED RETROSPOT
5	WHITE HANGING HEART T-LIGHT HOLDER
6	ASSORTED COLOUR BIRD ORNAMENT
7	PACK OF 72 RETROSPOT CAKE CASES
8	POPCORN HOLDER
9	RABBIT NIGHT LIGHT
10	MINI PAINT SET VINTAGE

EDA - Understanding One-Time vs. Repeat Customers

Percentage of customer that only purchased once:

A high percentage of one-time customers indicates potential issues with retention or loyalty. Understanding these patterns helps target re-engagement and loyalty-building strategies.

35.86%

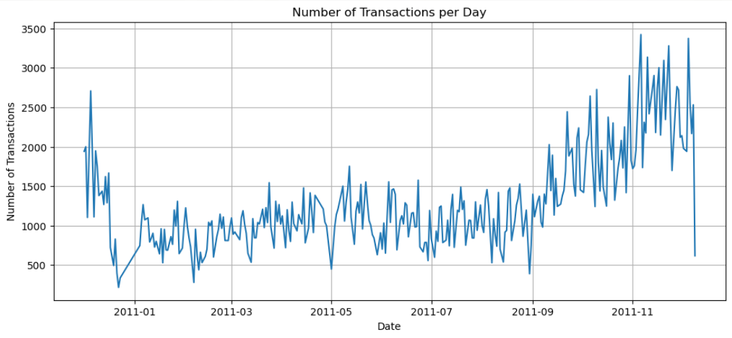


EDA - Explore Demand Spikes

Identify demand spikes and understand when customer activity is at its highest.

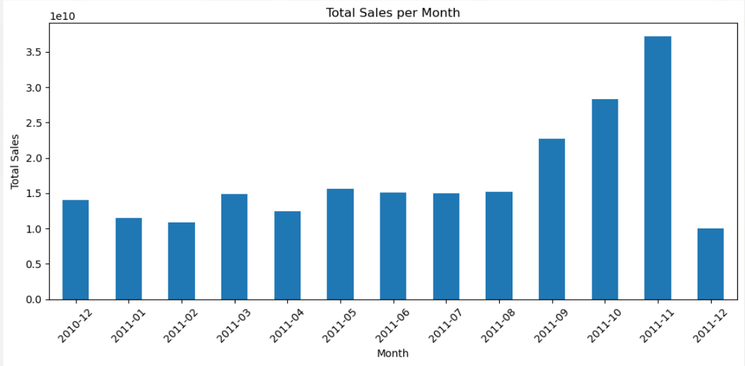
Detect Seasonality

- The dataset contains limited information, covering only about one year. Therefore, any inference about seasonality or demand spikes remains tentative. However, based on the available data, sales show a continuous increase from September to December, suggesting that the peak season for this online retail store occurs during **the fall months.**



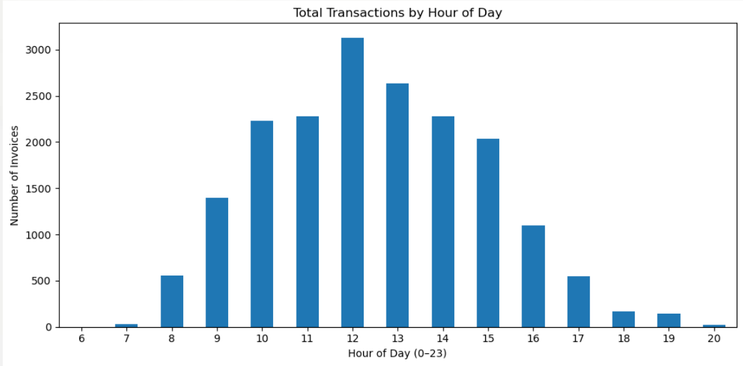
Total Sales per Month

- Total sales begin to rise in September and reach their peak in November.
- Data for December 2011 is incomplete, as it only covers the early part of the month.



Peak Shopping Hours

- **Peak shopping hours:** Between 11 AM and 2 PM.
- **Low activity:** Early morning (before 8 AM) and evening (after 6 PM).
- Most online retail transactions occur during standard working hours, suggesting that customers are most active around lunchtime.

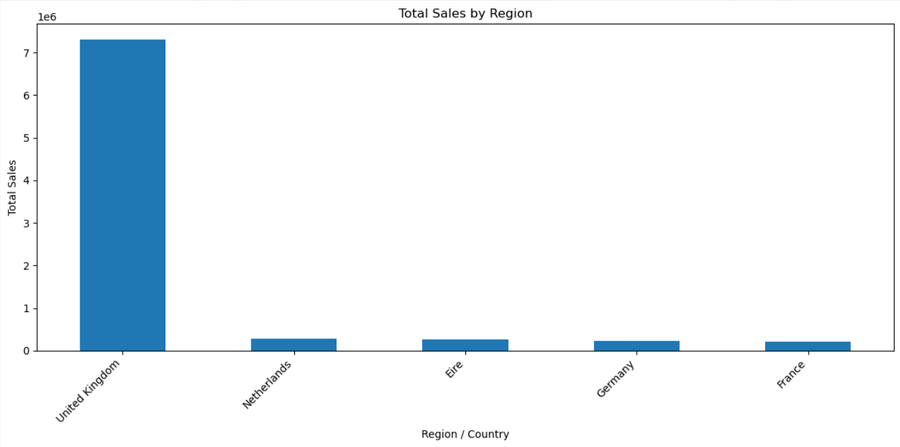


EDA - Regional Performance Insights

Understanding both sales and customer distribution by region helps align marketing, inventory, and operational strategies to regional demand patterns.

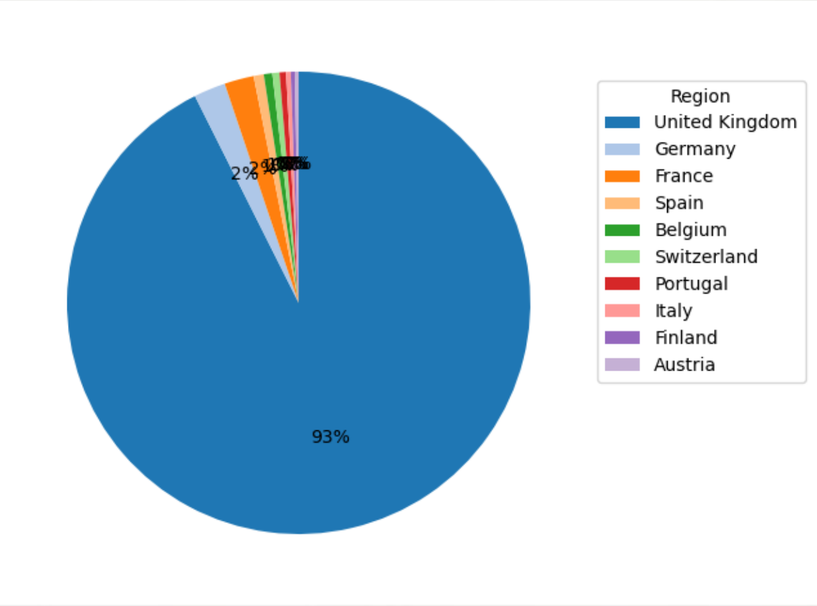
Total Sales by Region

- The United Kingdom generates the highest revenue, reflecting its position as the company's primary market.
- It is followed by the Netherlands, Ireland, Germany, and France, which show similar sales levels and are geographically close to the UK.



Customers by Region

- The United Kingdom accounts for approximately 93% of the customer base, followed by Germany and France, which each represent around 2%.



RFM Analysis

RFM stands for Recency, Frequency, and Monetary Value — a marketing analysis framework used for customer segmentation and behavioral targeting. It helps businesses identify their most valuable customers based on purchasing behavior, enabling data-driven decisions to improve strategy and performance.

Recency

- How recently a customer made their last purchase.

Frequency

- How often a customer makes purchases.

Monetary Value

- How much a customer spends.

Objectives:

- Identify high-value customers and understand purchasing patterns. By leveraging these insights, businesses can optimize marketing strategies, improve product positioning, and increase revenue through better customer engagement.

K-Means Clustering

Objectives:

- Group similar data point, in our scenarios, which is grouping customer into different segments.
- It aim to minimize the inertia which is the distance between each data points to the centroids, which is what within cluster distance (using the Euclidean distance). Tight knit clusters showing a higher cohesiveness of the group.
- Then K-Means try to maximizing between clusters, which means each cluster can be more distinct.

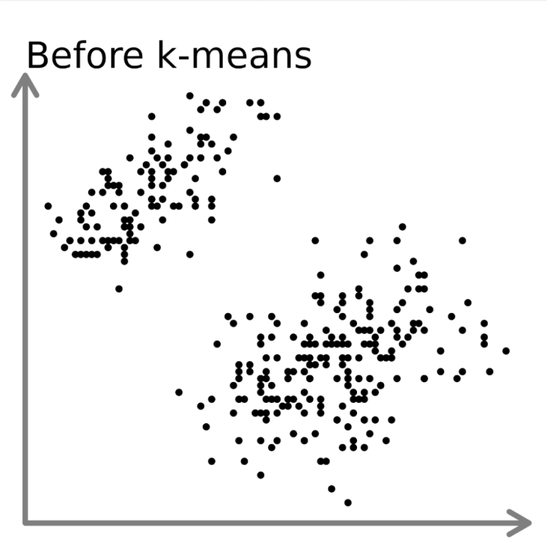
Key concepts:

- **Inertia:** how far the points within a cluster are.

--> lower the inertia is, better the K-Mean clustering.

- **Dunn Index:** how far between clusters.

--> higher the Dunn Index, better the K-Mean clustering.



K-Means. (n.d.). Datacamp. https://media.datacamp.com/legacy/v1725630538/image_9c867e067e.png

Data Pre-Processing

Data Preprocessing is crucial for K-Mean Clustering as the raw data usually highly skewed and containing outliers. K-Means clustering works by minimizing the sum of squared distances between data points and cluster centroids. As K-Means uses Euclidean distance, it is sensitive to large values, which the outliers.

Method	Purpose
Log Transformation	Reduce skewness and compress large values while keeping the order of magnitude.
Z-Score Normalization	Scale features to have mean = 0 and standard deviation = 1, so all features contribute equally to clustering.

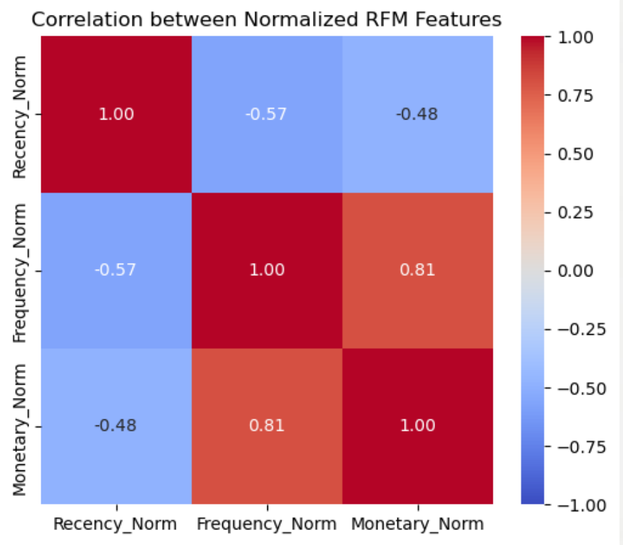
RFM Score Correlation After Standardization

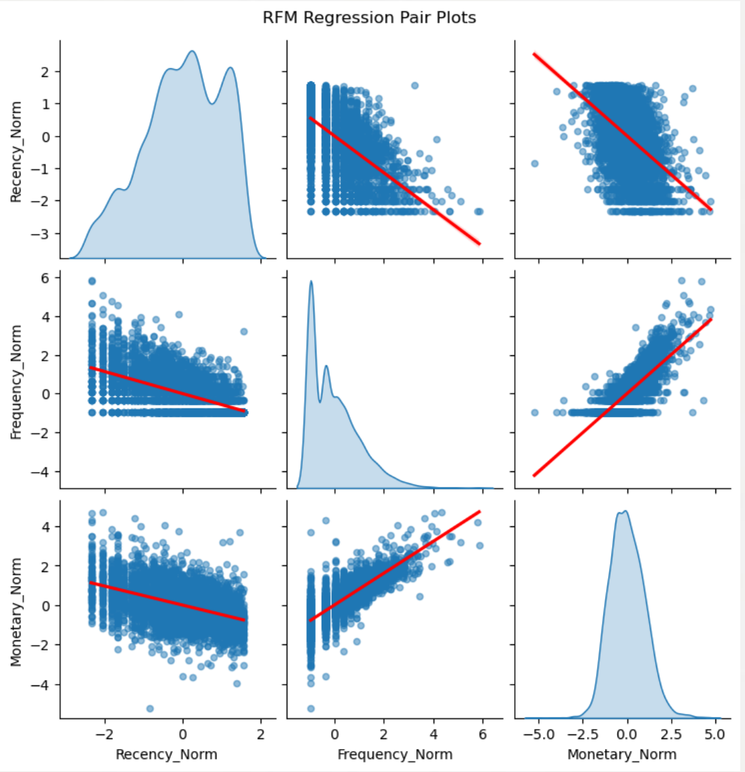
- Strong positive relationship between Monetary Value and Frequency (correlation = 0.81

  → Customers who purchase more frequently tend to contribute higher sales.

- Moderate negative relationship between Recency and Frequency (correlation = -0.57), as well as between Recency and Monetary Value (correlation = -0.48).

  → Longer the time since a customer's last purchase, the less frequently they make purchases, and the lower the total amount they spend.





K-Means Clustering – Elbow Method

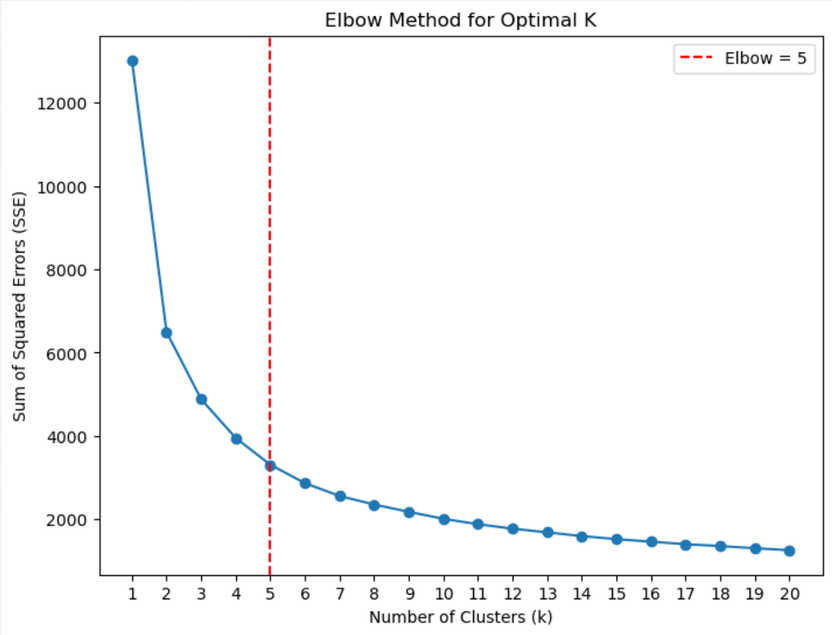
Elbow method is widely used technique in clustering (especially K-Means) to help determine the optimal number of clusters (k).

Concepts:

- It run K-Means for a range of K values, and for each K, it calculate the Within_Cluster Sum of Squares (WCSS), aka inertia.

- Optimal number of cluster is the one that balancing cluster compactness and simplicity. Where the WCSS curve starts to flatten.

The result of the optimal number of clusters is 5.

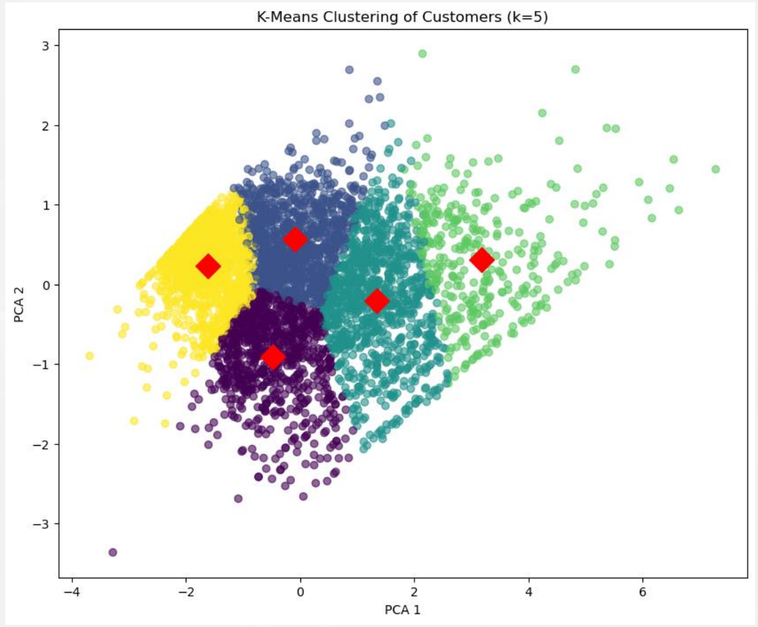


$$WCSS = \sum_{i=1}^{k} \sum_{x \epsilon C_i} ||x - \mu_i||^2$$

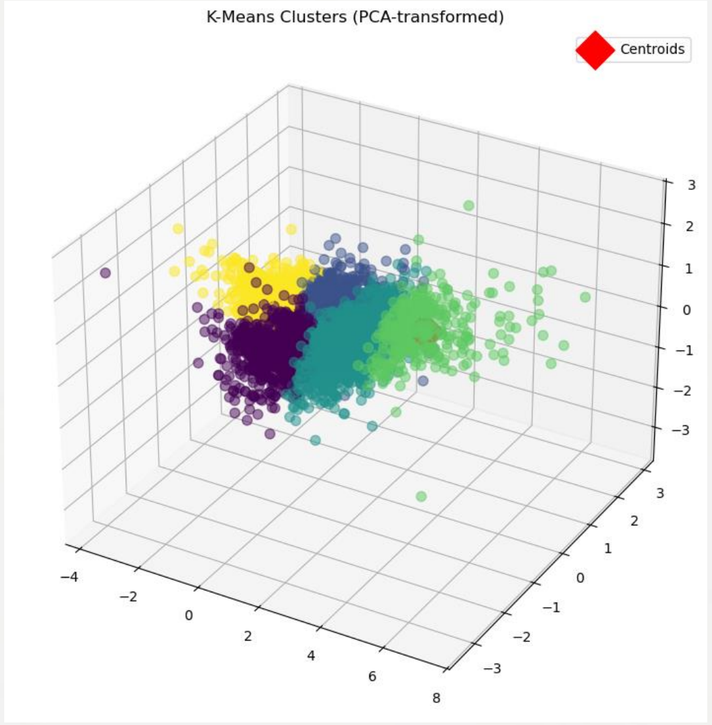
Within-cluster sum of squares (WCSS) equation. | Image: Built In

K-Means Clustering – Elbow Method

Data Visualization



2 Dimensions



3 Dimensions

K-Means Clustering – Silhouette Method

Silhouette method is another method to evaliate clustering quality and help to find the optimal number of clusters(k) by measuring how similar a data point is to its own cluster compared to other clusters.

Silhouette Score Evaluation and range:

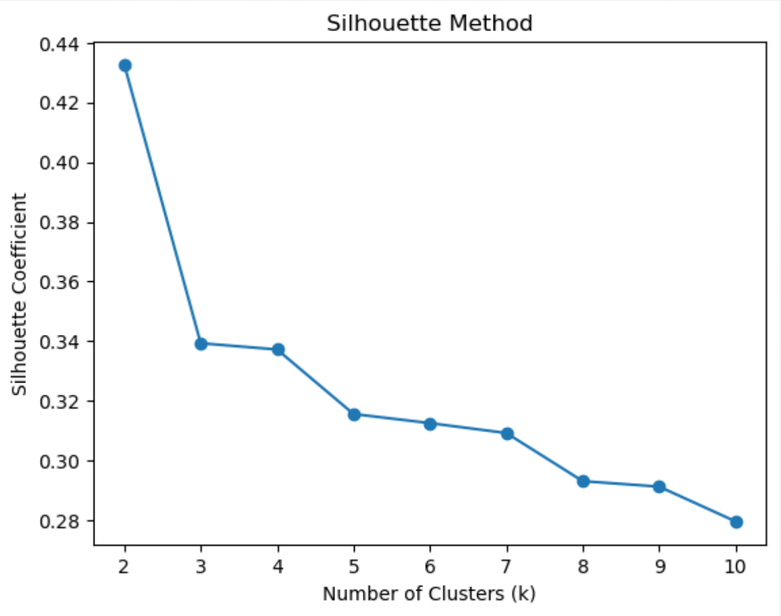
0.71-1: Strong structure and clusters well-seperated;

0.51-0.7: Reasonable structure;

0.26-0.5: Weak structure, overlapping clusters

< 0.25: No substantial structure

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$



Miller, C. (n.d.). *Training systems using Python Statistical modeling*. O'Reilly Online Learning. https://www.oreilly.com/library/view/training-systems-using/9781838823733/f6058e32-7d77-4256-abb5-ebae0e679d56.xhtml

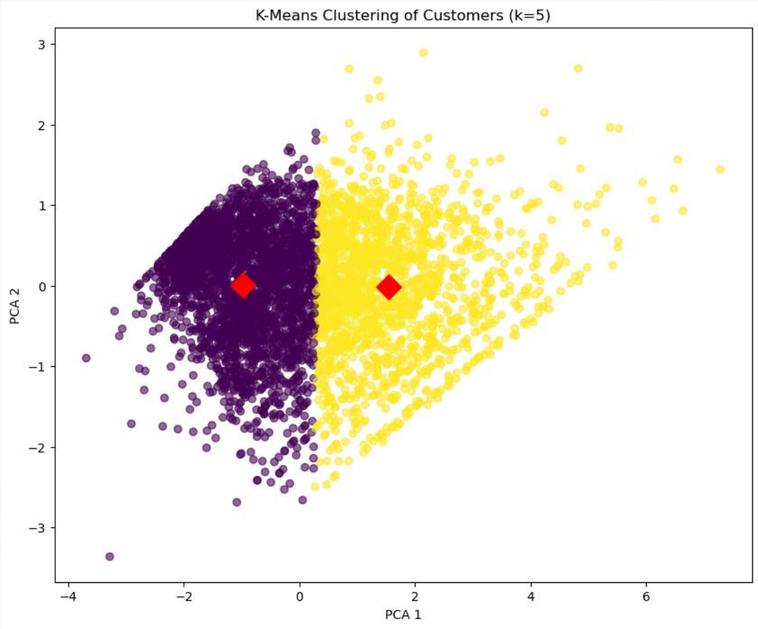
→ Usually higher silhouette Score means the data is well separated.

Results:

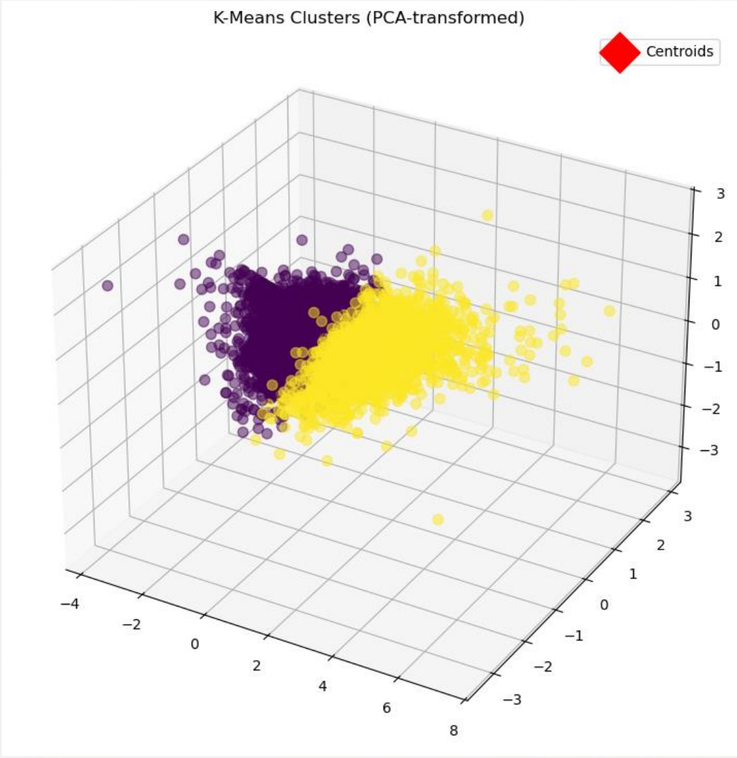
K=2, Silhouette Score=0.433
K=3, Silhouette Score=0.336
K=4, Silhouette Score=0.336
K=5, Silhouette Score=0.316
K=6, Silhouette Score=0.313
K=7, Silhouette Score=0.309

K-Means Clustering – Silhouette Method

Data Visualization



2 Dimensions



3 Dimensions

Data statistical interpretation (5 Clusters)

VIPs / Champions (≈8% of customers, 53% of revenue)

- Small group driving over half of total revenue — confirms Pareto principle (80/20 rule).
- Highly engaged (avg. recency: 12 days).
- Action: Prioritize retention and exclusive offers to sustain engagement.

Potential Loyalists (22% of customers, 16% of revenue)

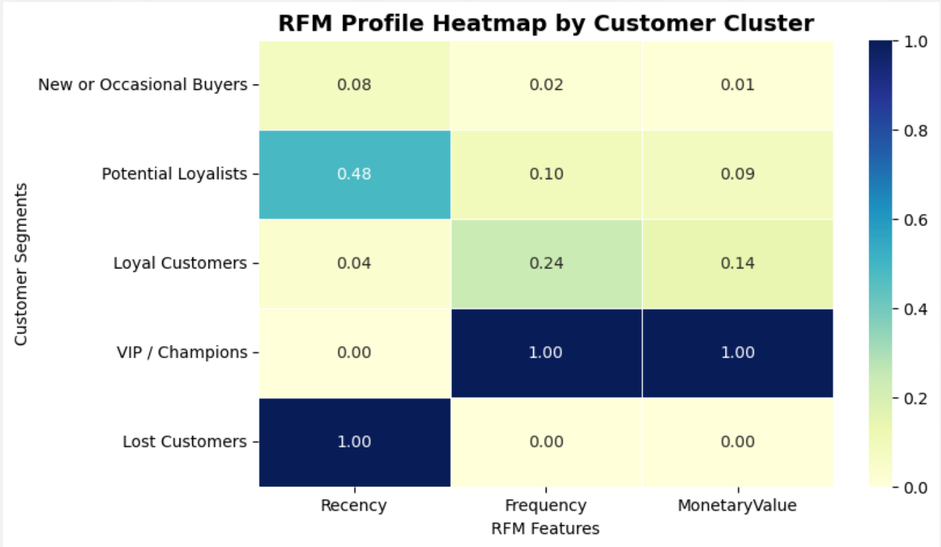
- Moderate activity; declining recency (R = 0.48).
- **Risk:** Becoming inactive soon.
- **Action:** Re-engage through reminders, special offers, or win-back campaigns.

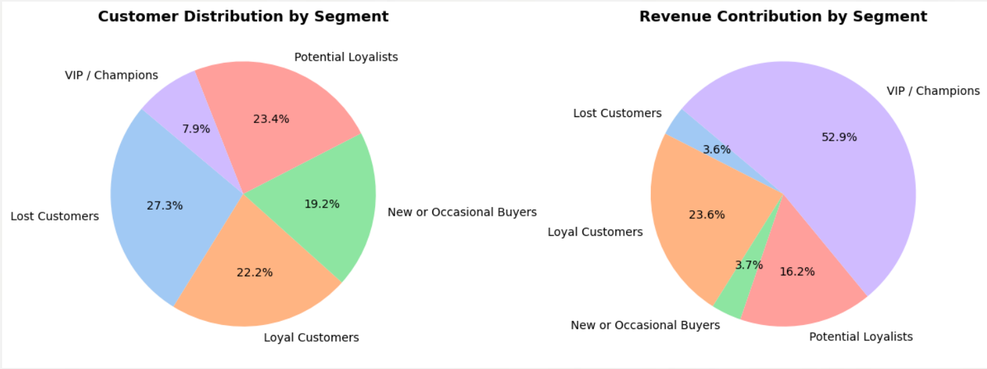
Loyal & Potential Loyalists (≈45% of customers,
40% of revenue)

- Strong mid-tier segment with scalable potential.
- **Action:** Move them toward VIP tier through personalization, loyalty rewards, and targeted campaigns.

New / Occasional Buyers & Lost Customers (46% of customers, <8% of revenue)

- Weak retention: low purchase frequency (1–1.6 avg).
- *Lost Customers* are the largest group (27%), indicating **high churn**.
- **Action:** Improve onboarding, satisfaction, and post-purchase experience.



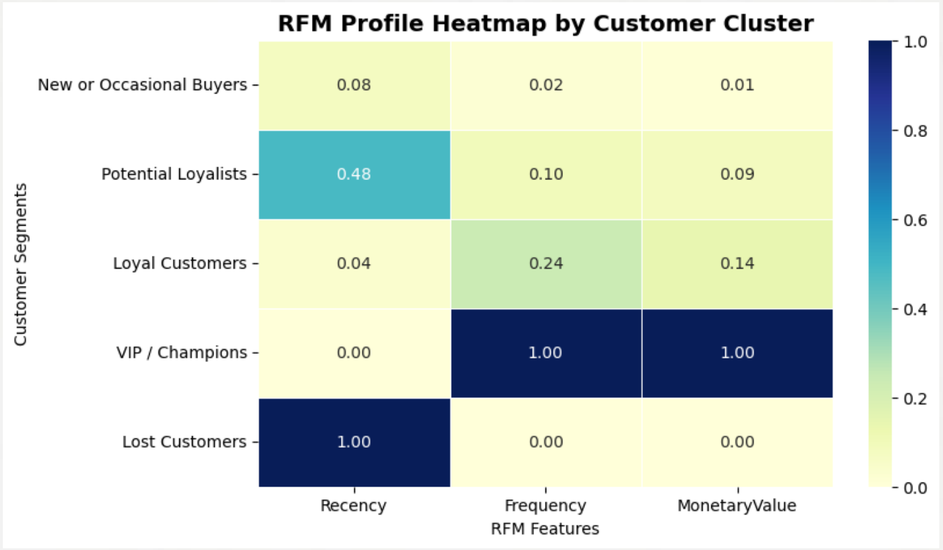


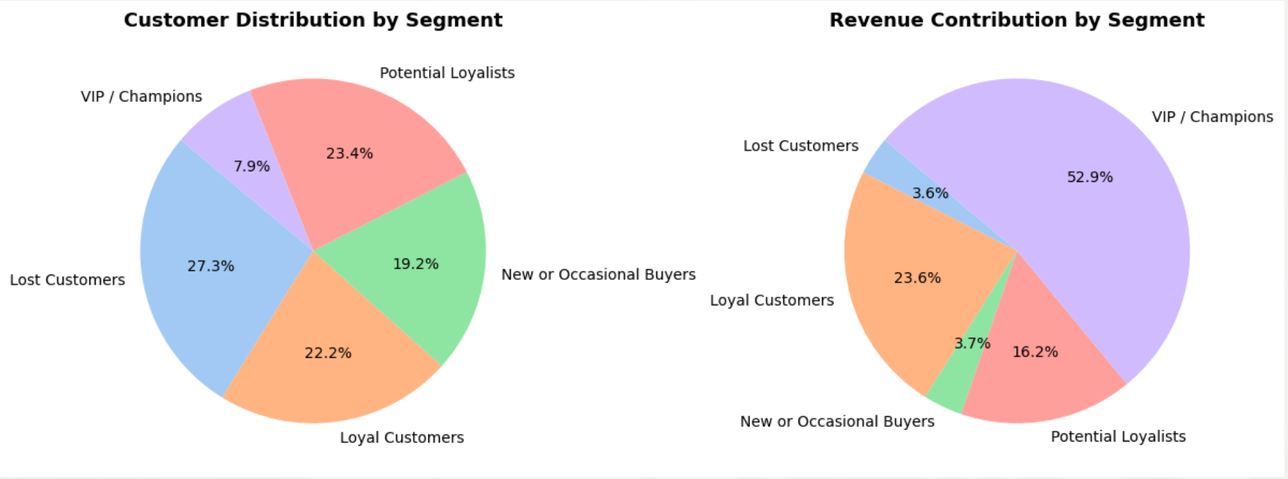
Data statistical interpretation (5 Clusters)

Summary Insight:

The company’s revenue is heavily concentrated among a small elite group (VIPs) while nearly half the customers are inactive or at risk.

Strategic focus should shift toward retention and reactivation to reduce churn and balance revenue distribution.

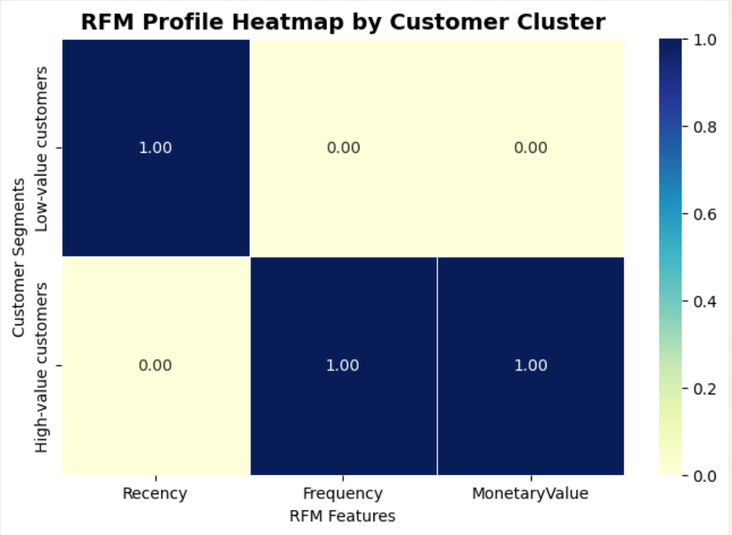




Data statistical interpretation (2Clusters)

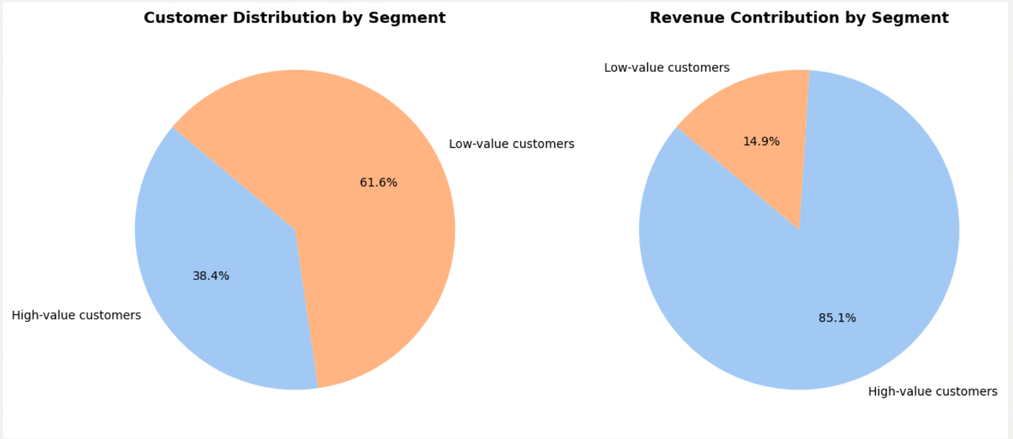
Low-Value Customers (≈62% of customer base, ≈15% of revenue)

- Average purchase frequency: **1.67**, indicating low retention (most leave after 1–2 purchases).
- Represent the **majority of customers** but contribute **minimal revenue**.
- **Risk:** High customer churn and weak long-term engagement.
- **Action:** Strengthen onboarding, reactivation, and cross-selling strategies.



High-Value Customers (≈38% of customer base, ≈85% of revenue)

- **Highly engaged:** average recency **26 days** vs. 134 for low-value customers.
- Purchase **~8.4× more often** and spend **~9× more** (avg. $4,548 vs. $498).
- **Insight:** This group drives the business — loyal, active, and profitable.
- **Action:** Prioritize retention, loyalty programs, and personalized experiences.



Data statistical interpretation (2Clusters)

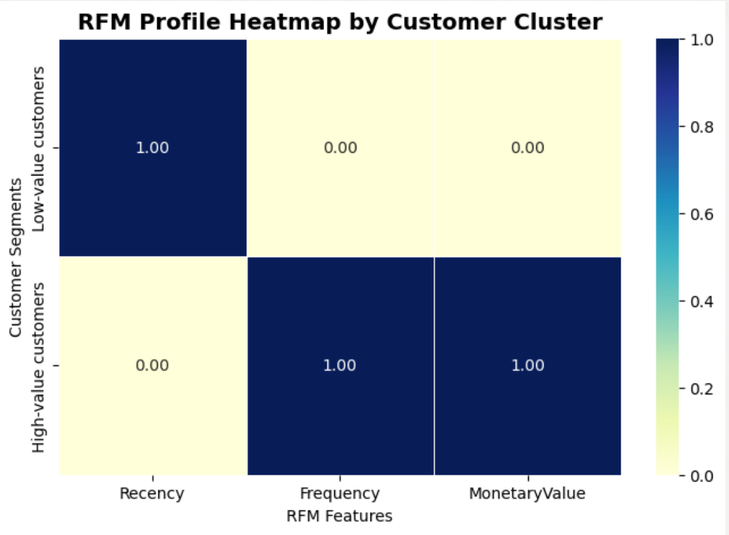
Summary Insight:

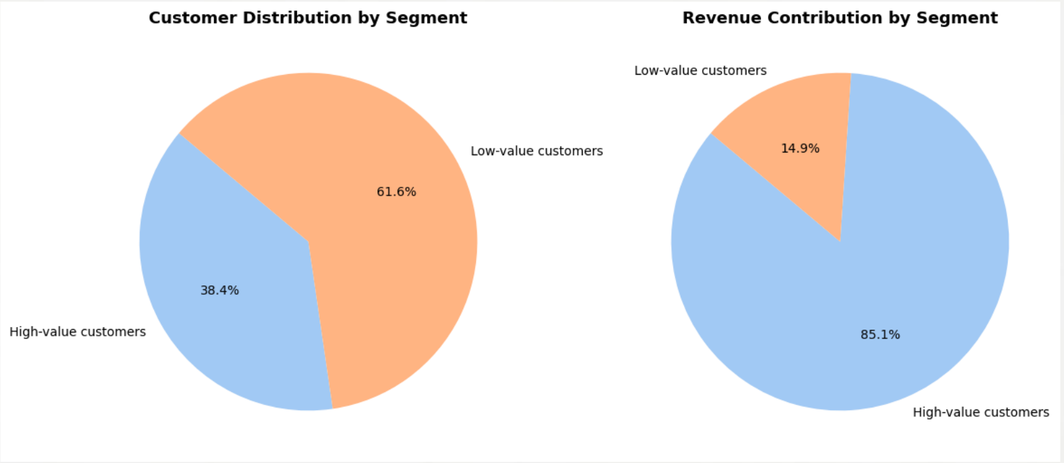
The business faces a high customer concentration risk —

revenue is heavily dependent on a small group of high-value customers.

Strengthening retention among low-value customers and

protecting high-value relationships is critical for sustainable growth.





Business Recommendations

Retain VIPs and High-Value Customers

- Focus on active, loyal, high-revenue customers through tailored loyalty programs.
- Offer exclusive rewards, early access, and personalized offers to strengthen relationships.
- Gather insights on needs and satisfaction to refine offerings and maintain engagement.
- Introduce advocacy or referral programs to leverage loyal customers as brand promoters.

Increase Retention Rate

- Launch reactivation campaigns using past purchase and preference data.
- Analyze the customer journey to identify friction points and service gaps.
- Implement personalized promotions and early loyalty incentives to encourage repeat purchases.
- Set up automated campaigns triggered by inactivity to re-engage dormant customers.

Sources

Dataset:

UCI Machine Learning Repository. (n.d. b). https://archive.ics.uci.edu/dataset/352/online+retail

References:

- Mgmarques. (2018, November 27). *Customer segmentation and market basket analysis*. Kaggle. https://www.kaggle.com/code/mgmarques/customer-segmentation-and-market-basket-analysis
- DipraCh. (n.d.). *GitHub - DipraCh/Customer-Segmentation-RFM-Analysis-and-K-Means-Clustering*. GitHub. https://github.com/DipraCh/Customer-Segmentation-RFM-Analysis-and-K-Means-Clustering/tree/main

Images: (Presentation Only)

- *K-Means*. (n.d.). Datacamp. https://media.datacamp.com/legacy/v1725630538/image_9c867e067e.png
- Miller, C. (n.d.). *Training systems using Python Statistical modeling*. O'Reilly Online Learning. https://www.oreilly.com/library/view/training-systems-using/9781838823733/f6058e32-7d77-4256-abb5-ebae0e679d56.xhtml
- *UCI Machine Learning Repository*. (n.d.). https://archive.ics.uci.edu/dataset/352/online+retail

The End

THANK YOU FOR YOUR TIME

Xiwen Mark