E-commerce Sales & Customer Segmentation Analysis

Optimizing product offerings, marketing, and revenue retention

Created by - Deeksha Gosain

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- The project was conducted to understand **customer value distribution and engagement patterns** within an e-commerce business, with the objective of enabling **data-driven retention, upsell, and win-back strategies**.
- This project analyzes e-commerce transactional data to segment customers using **RFM analysis** and **K-Means clustering**.
- Customer segmentation was performed using RFM analysis, categorizing users into four business-relevant segments:
  - **Active High Value Customers** (77.23% revenue share with VVIP micro-segment found using K-Means Clustering)
  - **Inactive High Value Customers**
  - **Active Low Value Customers**
  - **Inactive Low Value Customers**
- The findings highlight a **strong dependence on a small group of high-value customers**, alongside clear growth opportunities in **Active Low-Value customers** through order-value expansion and in **Inactive High-Value customers** through targeted win-back initiatives.

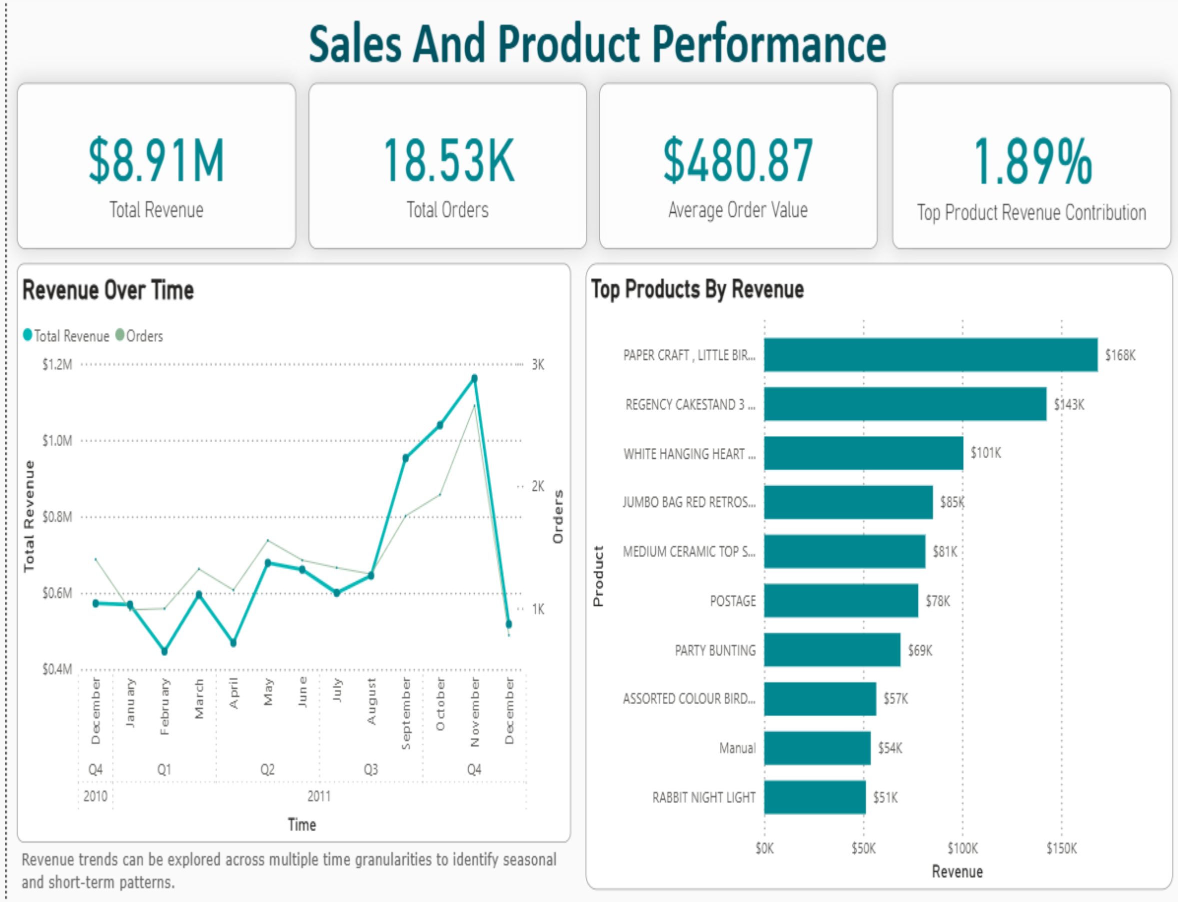
INTRODUCTION

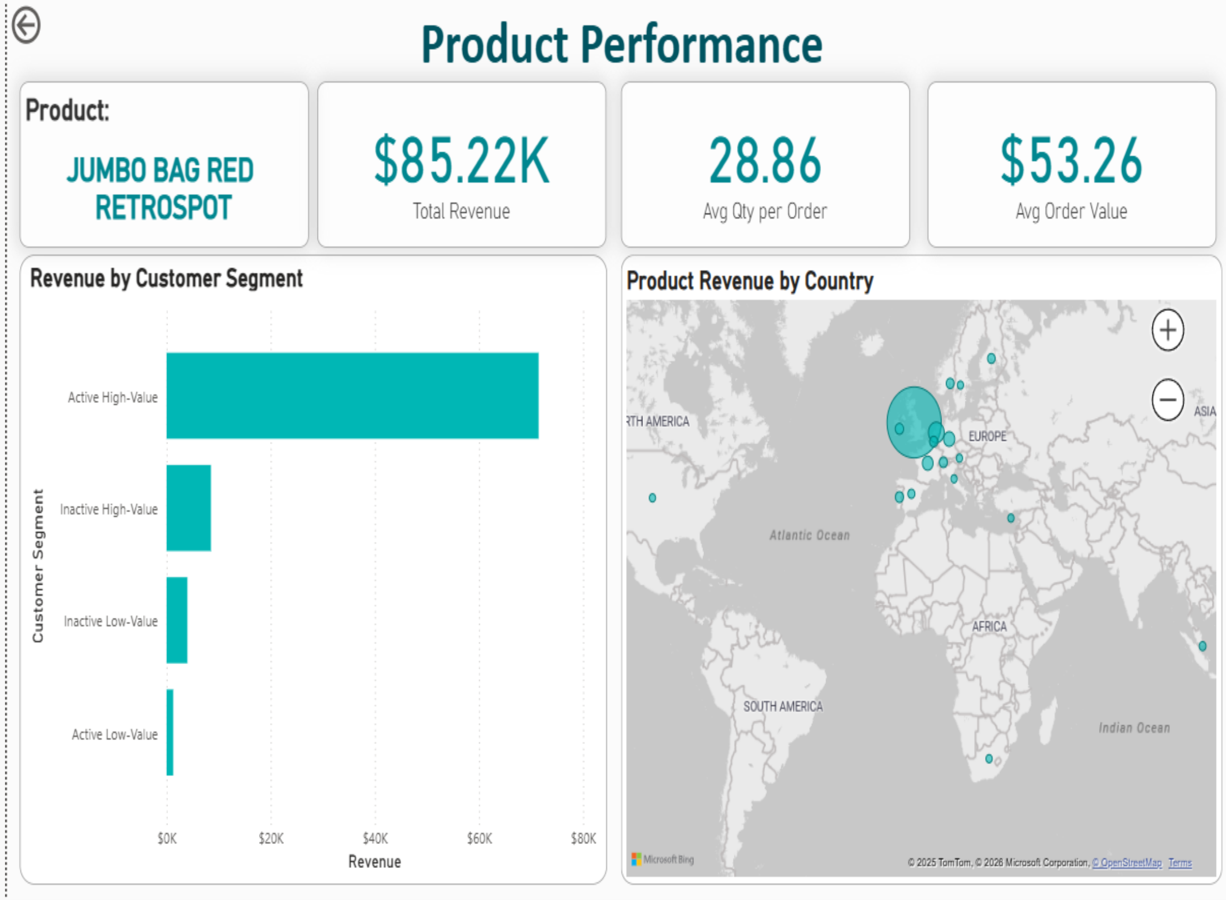
- This project uses UCI Online Retail data for a non-store online retail store based in UK between Dec 2010 and Dec 2011, which contains transactions from 4339 customers worldwide.
- Three key questions guide the analysis:
  - How are customers distributed based on recency, purchase frequency, and monetary value?
  - Which customer segments contribute the most to overall revenue?
  - Are there distinct customer groups within high-value customers that require differentiated engagement strategies?
- The goal of the analysis is to translate raw transaction data into actionable customer insights that support retention, upsell, and win-back decision-making.
- The findings provide a structured view of customer value concentration, highlight growth opportunities across segments, and demonstrate how customer-level analytics can inform strategic decisions in an e-commerce context.

METHODOLOGY

1. **Data Collection & Storage**
   - Downloaded retail transaction data from UCI Repository
   - Modeled data using a **star schema** (fact & dimension tables) for structured analysis
2. **Data Cleaning (SQL)**
   - Removed blanks and duplicates
   - Filtered logically incorrect transactions (negative quantity and unit price)
3. **Exploratory Data Analysis**
   - Analyzed distributions of **recency, frequency, and revenue**
   - Identified right-skewed behavior to validate **RFM segmentation**
4. **Customer Segmentation**
   - Applied **RFM analysis** to classify customers into four value-based segments
   - Performed **K-Means clustering** to uncover deeper behavioral patterns
5. **Dashboards**

DASHBOARD TAB 1





Interactive view of revenue trends, product performance, and drill-through product insights across customer segments and geographies.

DASHBOARD TAB 2

Customer Segmentation and Revenue Impact

1316	419	$986K	3.18%
Active High-Value Customers	Inactive High-Value Customers	Revenue at Risk	VVIP Revenue Share

Segment Behavioural Profile

Segments	Median Recency(days)	Median Frequency	Median Monetary
Active High-Value	15.00	133.00	2,358.89
Active Low-Value	18.00	38.00	462.43
Inactive High-Value	100.00	57.00	1,516.00
Inactive Low-Value	152.00	19.00	320.62

Customer Segment Preference for Top Products

Darker shading indicates products that are more heavily purchased by a given customer segment; values show absolute revenue.

Product	Active High-Value	Active Low-Value	Inactive High-Val
ASSORTED COLOUR BIRD ORNAMENT	$44.36K	$2.85K	$4.8
JUMBO BAG RED RETROSPOT	$71.44K	$1.27K	$8.5
Manual	$41.09K	$0.10K	$10.5
MEDIUM CERAMIC TOP STORAGE JAR	$3.29K	$0.23K	$77.6
PAPER CRAFT , LITTLE BIRDIE	$168.47K		
PARTY BUNTING	$48.96K	$1.28K	$12.1
POSTAGE	$60.30K	$2.55K	$8.4
RABBIT NIGHT LIGHT	$48.06K	$2.02K	$0.3
REGENCY CAKESTAND 3 TIER	$115.49K	$3.78K	$13.9
WHITE HANGING	$73.08K	$2.80K	$16.1

Customers By Segments

Revenue Share By Segment

Customer segmentation view highlighting revenue concentration, behavioral differences, and segment-level product preferences.

DASHBOARD TAB 3

Revenue Growth and Retention

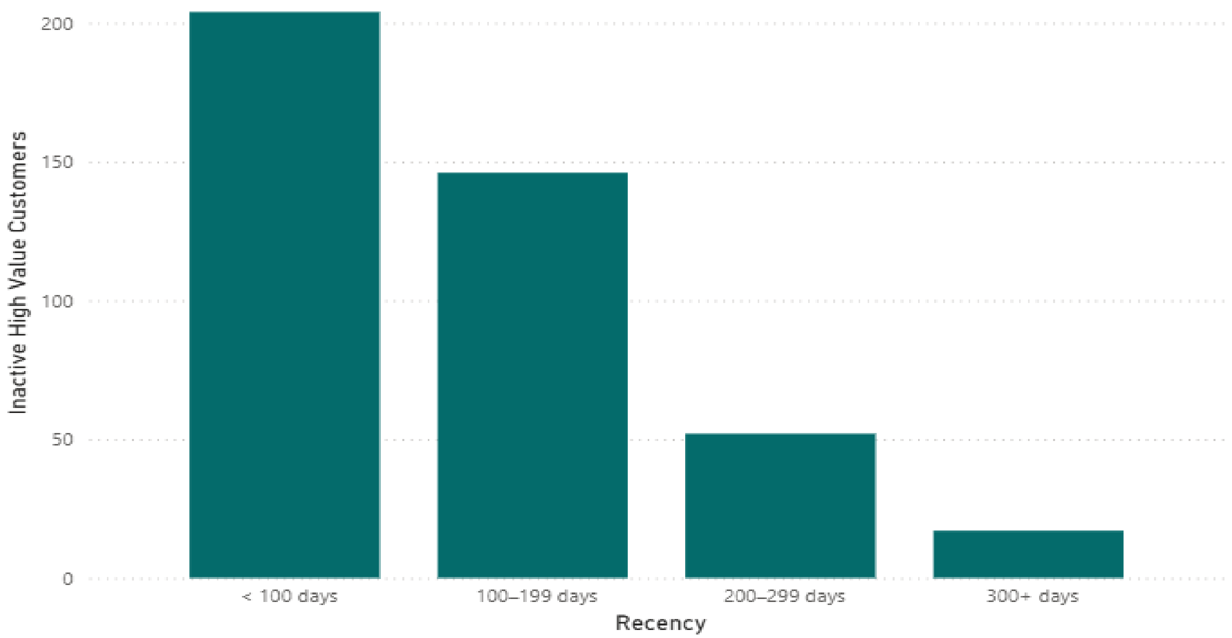
419
Win Back Opportunity Customers

768
Upsell Opportunity Customers

27.36%
% actionable customers

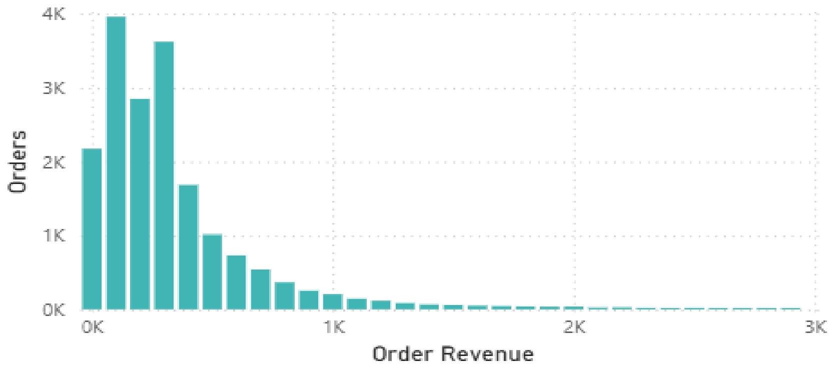
Win-Back Targeting Window

Identifies which inactive high-value customers should be prioritized for reactivation based on recency



Upsell Opportunity Window

Shows whether active customers usually place small orders, helping identify where to increase order size



VVIP Customers

Extreme high-value customers with frequent purchases, recent activity, and outsized revenue impact.

CustomerID	Revenue	Recency(Days)	Frequency
14,911	143.83K	1	5,675
14,096	65.16K	4	5,111
17,841	40.99K	1	7,847
12,748	33.72K	0	4,595

Action-oriented view identifying upsell, win-back, and high-impact customer retention opportunities.

OVERALL FINDINGS & IMPLICATIONS

Findings

- Revenue is highly concentrated among customer segments, with Active High-Value customers contributing 77% of total revenue, while Inactive High-Value customers represent $986K in revenue at risk.
- K-Means clustering identified a VVIP micro-segment (4 customers) generating 3.18% of total revenue, characterized by extremely high frequency and recent activity.
- Product revenue is highly diversified, with the top product contributing only 1.89% of total revenue, and some products generating high revenue from very low purchase frequency.
- The UK dominates order volume, while Active Low-Value customers place frequent but small orders, creating a clear upsell opportunity window.

Implications

- Retention of high-value customers is critical, and win-back strategies for inactive high-value customers offer high ROI compared to acquiring new customers.
- VVIP customers require **white-glove engagement** to protect from revenue impact.
- Top product's revenue share shows that business is not dependent on a single product
- **AOV expansion strategies** (bundles, minimum order thresholds, cross-sells) in the core UK market can unlock growth without increasing acquisition spend.

CONCLUSION

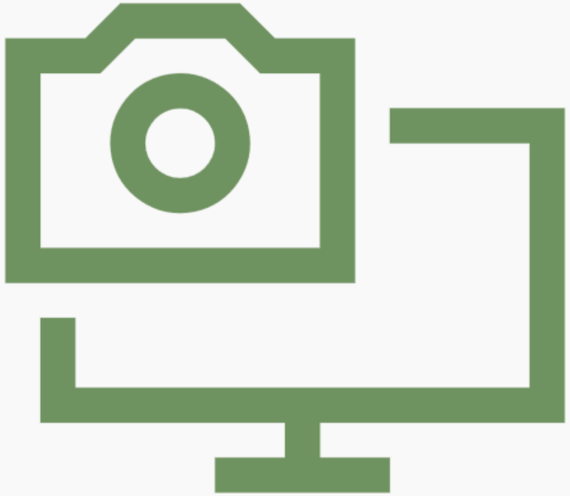
- The analysis demonstrates that e-commerce revenue is primarily driven by customer behavior rather than individual products, with a small group of high-value customers accounting for a disproportionate share of total revenue.

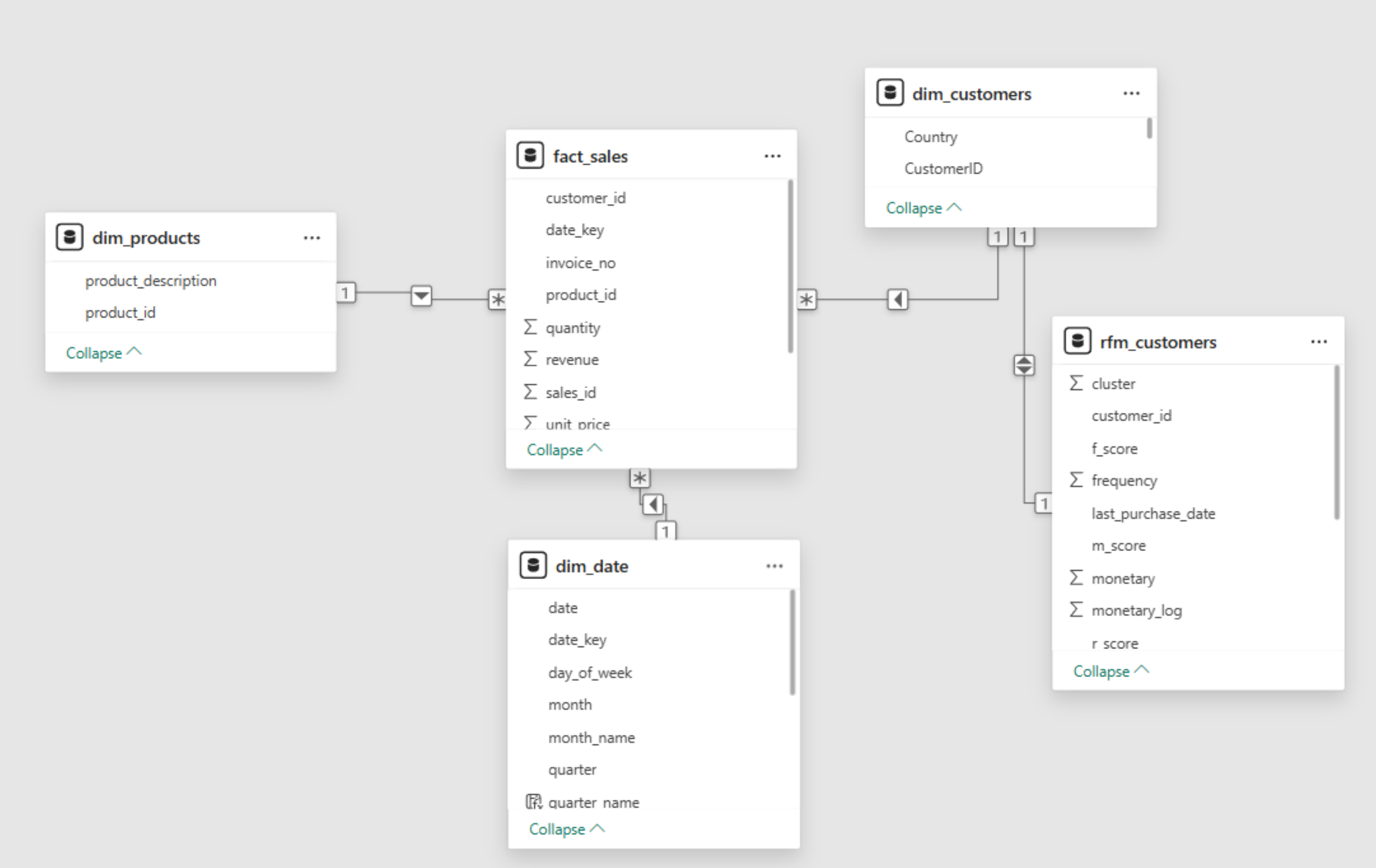
- Customer segmentation reveals a highly concentrated revenue structure, where retaining and protecting active high-value customers is critical, while inactive high-value customers present a strong win-back opportunity with higher expected returns than new customer acquisition.

- Although product revenue is diversified, clear seasonal demand patterns and frequent low-value purchasing behavior indicate that future growth will be driven by average order value expansion, targeted retention, and personalized engagement strategies.

- Overall, the findings highlight the importance of data-driven customer segmentation and behavioral analytics in guiding revenue growth, risk mitigation, and strategic decision-making in an e-commerce environment.

APPENDIX





Data Model used for this project

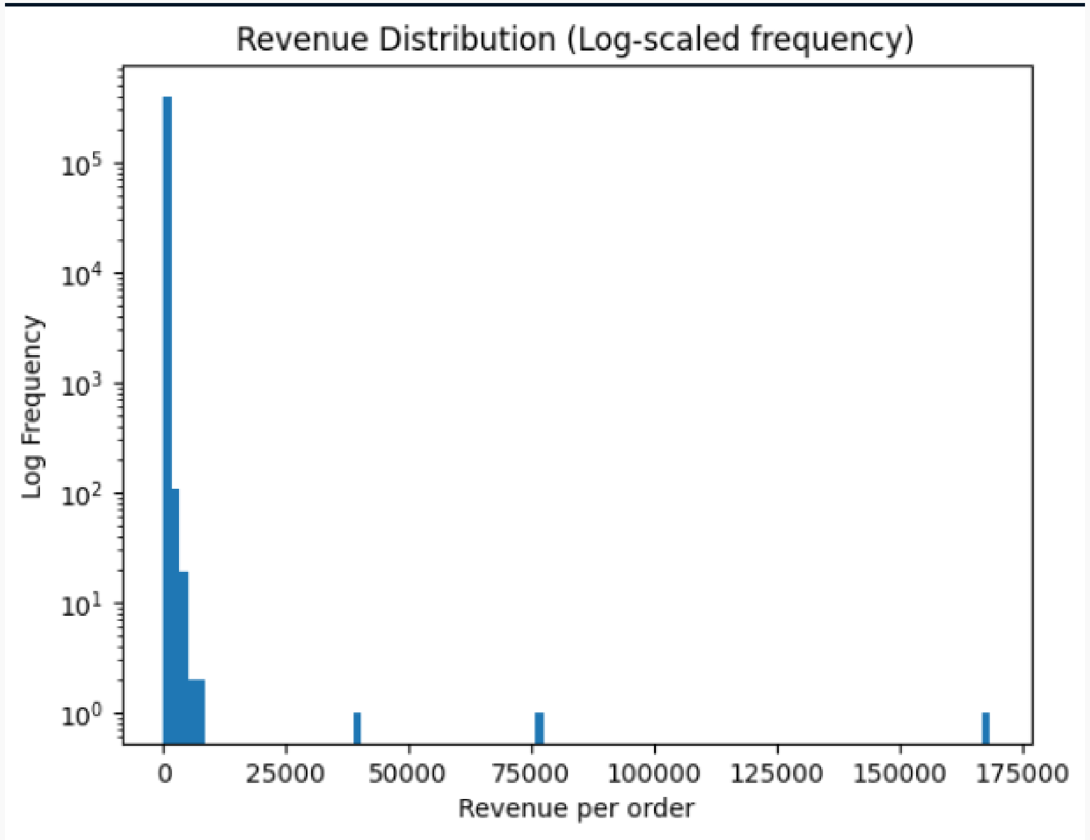
RFM Segmentation

```
# Activeness: recent OR (moderately recent AND frequent)
    active = (r >= 4) or (r >= 3 and f >= 4)

    # Value: high frequency OR high monetary
    high_value = (m >= 4)
```

Segmentation Logic

- Recency= days since last purchase
- Frequency = total orders per customer
- Monetary = total revenue per customer



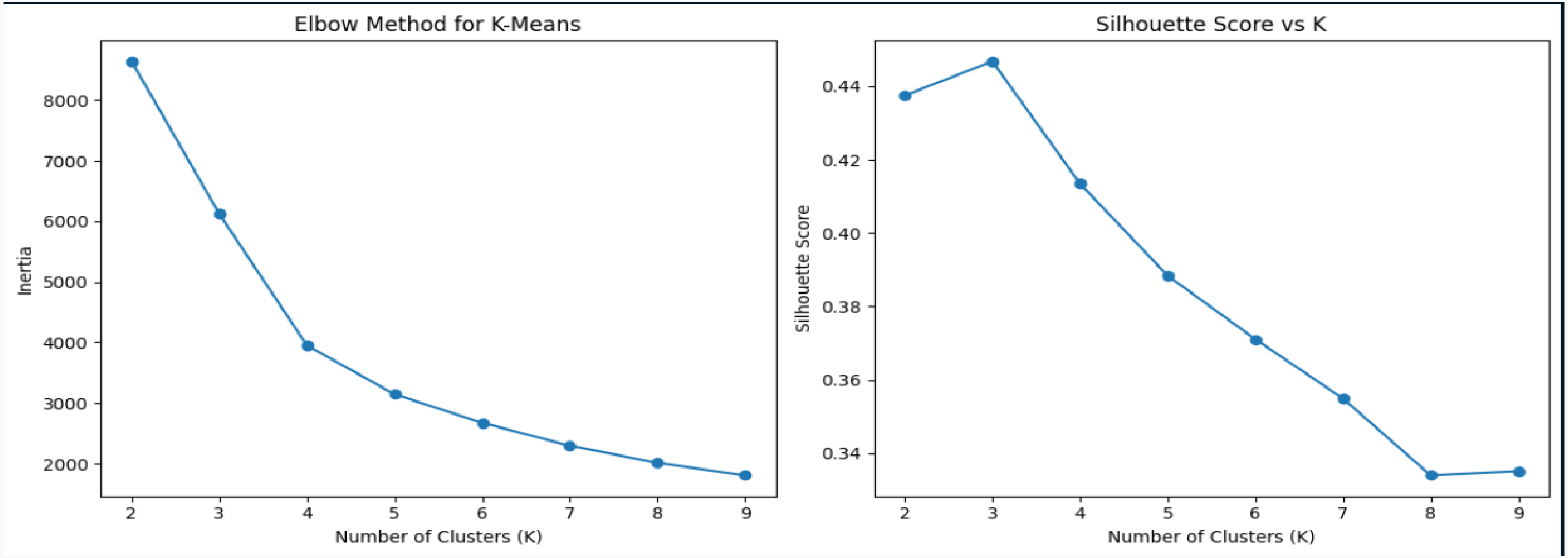
- The adjacent distribution shows that revenue is **highly right-skewed**.
- To prevent this skewness from disproportionately influencing the K-Means clustering results, a **log transformation** was applied to the monetary variable.
- This transformation was not required for RFM segmentation, as RFM relies on **relative scoring (1–5)**, which is robust to skewed distributions.

K-Means Clustering

- Cluster 3 → Ultra High-Value Outliers(4 customers making up more than 3% of the revenue)
- Cluster 2 → Active High-Value Customers
- Cluster 1 → Active Low-Value customers
- Cluster 0 → Inactive Low-Value customers

Interpretation of clusters based on centroid analysis

- Recency= days since last purchase
- Frequency = total orders per customer
- Monetary_log = log transformed total revenue per customer



Primarily Elbow method was used to determine K and silhouette score method was used as secondary method to determine k

From above methods value of **K= 4** was used