# Retail Stock Market Behaviour Phase-2: Customer Segmentation and Predictive Analytics


1st Tejmul Movin
*Department of CS and AI*
*Rishihood University*
Sonipat, India
sansar.t23csai@nst.rishihood.edu.in

2nd A Jithendranath
*Department of CS and AI*
*Rishihood University*
Sonipat, India
akula.j23csai@nst.rishihood.edu.in

3rd M Sree Sai Nath
*Department of CS and AI*
*Rishihood University*
Sonipat, India
meesala.s23csai@nst.rishihood.edu.in



***Abstract*—This study presents a comprehensive multi-phase analytical framework applied to a UK-based online retail dataset containing 541,909 transaction records spanning December 2010 to December 2011. The research integrates data preprocessing, Exploratory Data Analysis (EDA), Recency-Frequency-Monetary (RFM) based customer segmentation using K-Means clustering, Market Basket Analysis (MBA) comparing Apriori and FP-Growth algorithms, and supervised learning using XGBoost. The analysis identifies four distinct customer segments and achieves 96.9% accuracy in predicting high-value transactions exceeding £500, demonstrating effective machine learning applications in retail analytics.**



***Index Terms*—Data Mining, K-Means Clustering, RFM Analysis, XGBoost, Market Basket Analysis, Customer Segmentation, E-commerce Analytics.**


## I. Introduction

### A. Background and Motivation

The exponential growth of online retail generates massive transactional datasets that unlock valuable insights into customer behavior and market trends. In competitive e-commerce environments, businesses must distinguish between high-value loyal customers and those at risk of churning. Traditional retail analytics using descriptive statistics and manual segmentation fail to capture the complex, multidimensional nature of customer behavior, while machine learning techniques offer powerful tools for pattern discovery and predictive modeling.

### B. Research Objectives

This research addresses key retail analytics challenges through five objectives: (1) implement robust data preprocessing handling missing values, duplicates, and outliers; (2) conduct extensive EDA identifying temporal, geographical, and behavioral patterns; (3) apply unsupervised learning for customer segmentation using RFM metrics; (4) utilize Market Basket Analysis for product association mining; and (5) develop supervised learning models for high-value transaction prediction.

## II. Methodology

### A. Data Preprocessing Pipeline

A systematic seven-stage preprocessing pipeline ensured data integrity:

1) **Remove Duplicates:** We removed 5,268 duplicate records to prevent inflated counts and incorrect metrics.
2) **Handle Missing Customer IDs:** We identified and removed 135,080 records (24.9%) that lacked a `CustomerID`, as behavioral segmentation is impossible without unique user identifiers.
3) **Remove Cancelled Orders:** We filtered out 9,288 invoices starting with 'C' to prevent revenue distortion. Fig. 1 illustrates the products with the highest cancellation frequency.

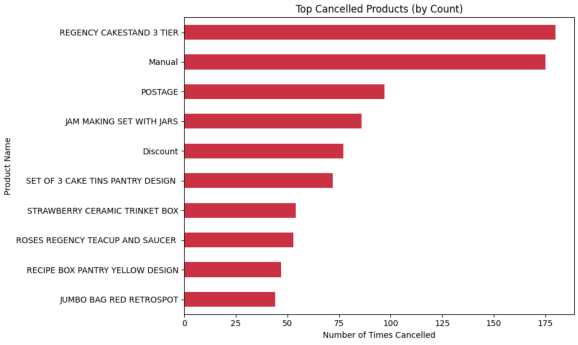


Fig. 1. Top Cancelled Products showing items with highest return rates.

4) **Remove Invalid Quantity or Unit Price:** Records with negative or zero `Quantity` or `UnitPrice` were removed to ensure analytical consistency.
5) **Convert Data Types:** We converted `InvoiceDate` to datetime objects and `CustomerID` to integers for efficient processing.
6) **Create New Analytical Features:** We engineered `TotalPrice` (Quantity $\times$ UnitPrice) and extracted temporal features including `Year`, `Month`, `Hour`, `DayOfWeek`, and `TimeOfDay`.
7) **Normalization:** We applied `StandardScaler` to `Quantity`, `UnitPrice`, and `TotalPrice` to standardize features achieving $\mu = 0$ and $\sigma = 1$, ensuring equal weight in machine learning algorithms.
8) **Outlier Removal (IQR Method):** Using the Interquartile Range method with bounds $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$, we removed 2,485 extreme outliers to

stabilize the statistical distribution, as shown in Fig. 2.

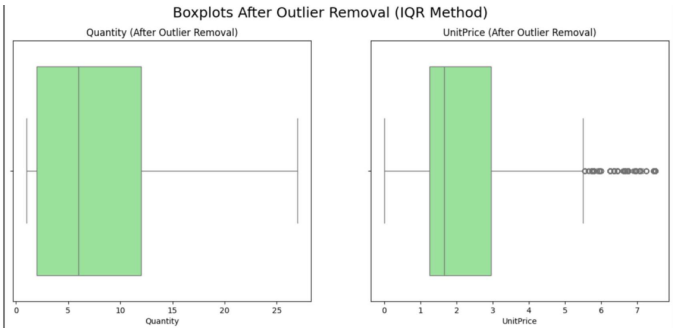


Fig. 2. Boxplots of Quantity and UnitPrice after applying IQR outlier removal, showing stabilized data distribution.

The final dataset summary is presented in Table I.

TABLE I
FINAL DATASET SUMMARY

| Metric | Count / Value |
|---|---|
| Total Rows | 333,234 |
| Unique Customers | 4,191 |
| Unique Products | 3,575 |
| Countries Covered | 37 |
| Total Revenue | £4,299,476.18 |

### B. Exploratory Data Analysis (EDA)

*1) Temporal Analysis:* Fig. 3 depicts the monthly sales trend. We observed steady growth throughout 2011, culminating in a significant peak in November (Q4), validating seasonal pre-holiday shopping behavior (Black Friday/Christmas).

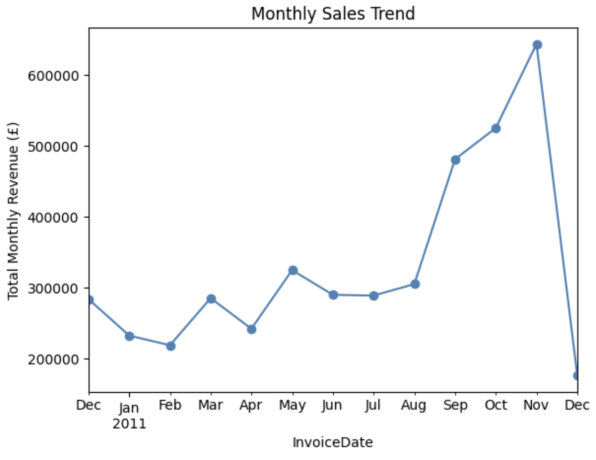


Fig. 3. Monthly sales trend showing significant revenue peaks in Q4 (November), indicating strong seasonality.

Transaction volume analysis by Day and Hour revealed "hot zones" on Wednesday and Thursday afternoons (12:00 PM – 3:00 PM), as shown in Fig. 4. Weekends and early mornings showed minimal activity, enabling optimized server scaling and customer support staffing.

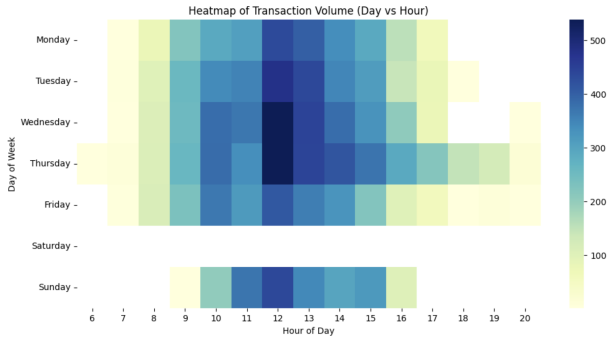


Fig. 4. Transaction volume heatmap identifying Wednesday and Thursday afternoons as peak operational windows.

*2) Correlation Analysis:* A correlation heatmap revealed that `Quantity` and `UnitPrice` have very low correlation (near zero), indicating bulk buyers don't necessarily purchase cheaper items. However, `Quantity` shows strong positive correlation with `TotalPrice`, identifying it as the primary revenue driver. This insight is relevant for pricing strategies: increasing unit prices doesn't strongly affect purchase quantity, while bulk discounts could significantly boost revenue.

*3) Customer Retention Analysis:* Cohort Analysis (Fig. 5) illustrates retention rates by first purchase month. Retention drops significantly after the first month (often below 25%), highlighting critical need for better onboarding and re-engagement strategies immediately following initial purchase.

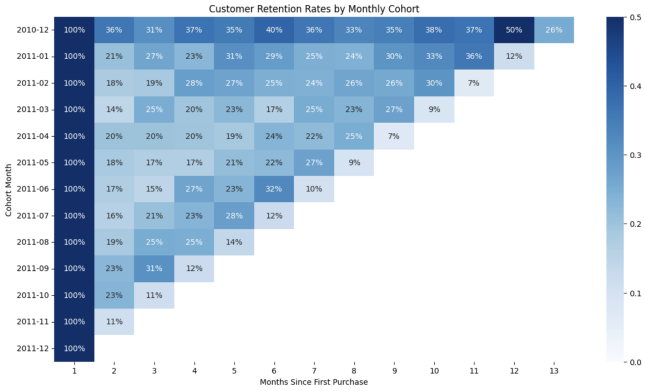


Fig. 5. Cohort Analysis Heatmap showing customer retention rates over time. The steep drop-off in Month 1 indicates need for better onboarding strategies.

## III. Customer Segmentation

### A. RFM Analysis

For each `CustomerID`, we calculated three metrics:

- **Recency (R):** Days since last purchase from reference date (December 9, 2011)
- **Frequency (F):** Number of unique orders (`InvoiceNo`)
- **Monetary (M):** Total revenue contribution (`TotalPrice`)

**Data Transformation:** Raw RFM data exhibited severe right-skewness (Fig. 6). We applied Log transformation $\log(1+x)$ to normalize distributions for clustering analysis.

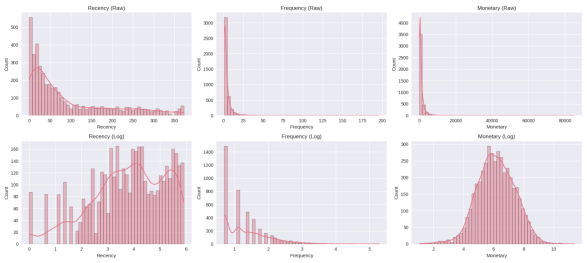


Fig. 6. Comparison of Raw vs. Log-Transformed RFM distributions. Raw data shows severe right-skew; log transformation normalizes for analysis.

### B. RFM Scoring Methodology

We utilized Quantile Discretization (`pd.qcut`) to assign scores on a 1-5 scale:

- **R Score:** Calculated using 5 quintiles where 1 = highest recency (worst), 5 = lowest recency (best)
- **F Score:** Calculated using 5 quintiles where 1 = lowest frequency, 5 = highest frequency (best)
- **M Score:** Calculated using 5 quintiles where 1 = lowest monetary, 5 = highest monetary (best)

The final metrics were calculated as:

$$\text{RFM\_Score} = \text{R\_Score} \parallel \text{F\_Score} \parallel \text{M\_Score} \quad (1)$$

$$\text{RFM\_Value} = \text{R\_Score} + \text{F\_Score} + \text{M\_Score} \quad (2)$$

where $\parallel$ denotes string concatenation for `RFM_Score` (e.g., "555" for best customers).

### C. Optimal Cluster Determination

Using Elbow Method and Silhouette Analysis on log-transformed and standardized data, $k = 4$ was identified as optimal (Fig. 7). The "elbow" in Sum of Squared Errors (SSE) at $k = 4$ coincides with highest Silhouette Score, confirming optimal mathematical separation.

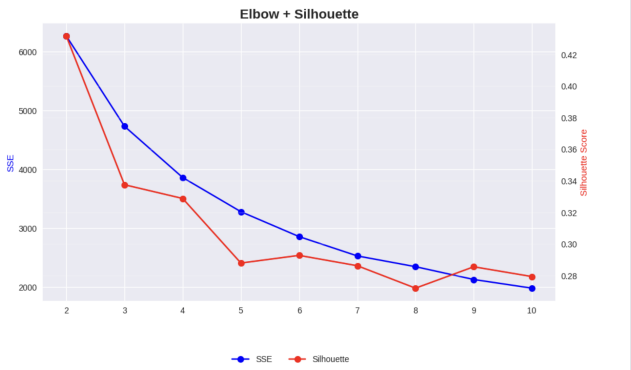


Fig. 7. Elbow Method and Silhouette Score analysis converging on $k = 4$ as optimal number of clusters.

### D. Cluster Profiles and Results

K-Means clustering with $k = 4$ produced four distinct segments (Table II).

TABLE II
CLUSTER PROFILES AND RFM AVERAGES

| Cluster | R | F | M (£) | % | Segment |
|---|---|---|---|---|---|
| 0 | 20.8 | 1.9 | 336.3 | 19.5 | Champions |
| 1 | 10.0 | 12.7 | 3461.4 | 15.9 | Loyal |
| 2 | 63.3 | 4.1 | 1137.7 | 28.4 | At Risk |
| 3 | 188.0 | 1.3 | 235.3 | 36.1 | Lost |

Fig. 8 (Snake Plot) visualizes standardized RFM values per cluster, clearly separating Champions (low Recency, high Frequency/Monetary) from Lost customers.

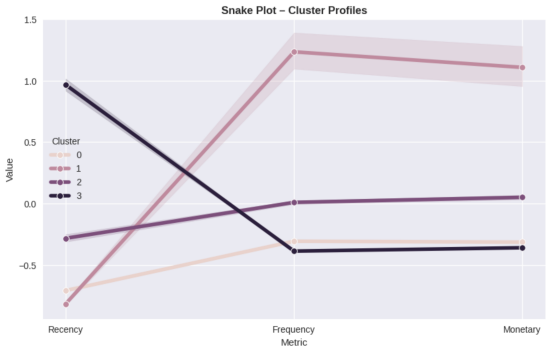


Fig. 8. Snake Plot showing standardized RFM values per cluster. Champions clearly separate with optimal metrics.

Fig. 9 provides 3D visualization confirming spatial cluster separation.

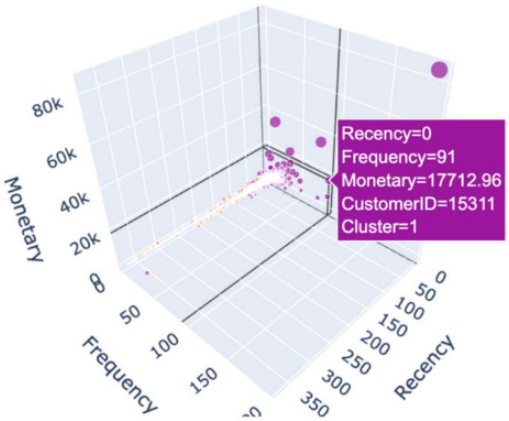


Fig. 9. 3D Visualization of Customer Segments showing clear spatial separation of four clusters.

- **Champions (Cluster 0):** Lowest Recency, highest Frequency/Monetary—most valuable customers
- **Loyal Customers (Cluster 1):** Consistent buyers with moderate-to-high spending

- **At Risk (Cluster 2):** High historical spending but increasing recency—churn risk
- **Lost/Hibernating (Cluster 3):** High recency, low frequency/monetary—minimal engagement

## IV. Market Basket Analysis

We compared Apriori and FP-Growth algorithms on UK transactions with minimum support = 0.02 (Table III).

TABLE III
ALGORITHM PERFORMANCE COMPARISON

| Metric | Apriori | FP-Growth |
|---|---|---|
| Min Support | 0.02 | 0.02 |
| Itemsets Found | 235 | 235 |
| Rules (Lift $>$ 1.5) | Identical | Identical |
| Time (sec) | $\approx$ 0.3 | $\approx$ 0.6 |

Both algorithms identified 235 frequent itemsets and identical association rules, validating results. FP-Growth demonstrated 50% computational efficiency improvement through tree-based structure eliminating candidate generation. Analysis revealed strong associations (Lift $>$ 20) between complementary items like "Green Regency Teacup" and "Pink Regency Teacup," enabling product bundling and cross-selling strategies.

## V. Predictive Analysis

### A. Problem Definition

We developed a binary classification model predicting high-value orders: BigOrder $= 1$ if TotalPrice $> £500$, else $0$.

### B. Feature Engineering and Model Training

Orders were aggregated by `InvoiceNo`, creating features: `Quantity` (total items), `UnitPrice` (mean), `StockCode` (unique products), and `Hour` (temporal effects). XGBoost Classifier was selected for effectiveness on imbalanced tasks and feature importance capabilities. Dataset split: 80/20 (training/testing) with stratified sampling.

### C. Results and Feature Importance

The model achieved 96.9% accuracy with 0.844 F1-Score, demonstrating strong precision-recall balance. Fig. 10 reveals feature importance hierarchy.

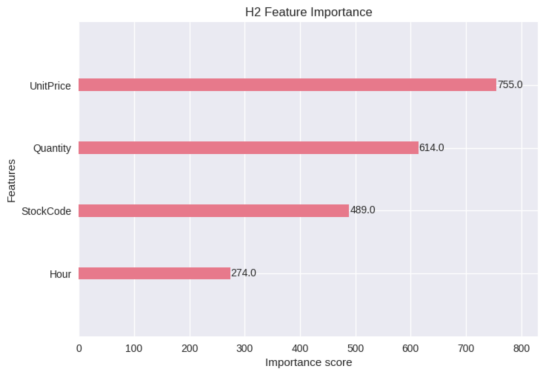


Fig. 10. XGBoost Feature Importance. Quantity dominates as predictor of high-value orders, followed by UnitPrice.

Analysis reveals:

- `Quantity` (F-score $\approx$ 500): Primary driver—bulk purchases indicate high-value transactions
- `UnitPrice`: Secondary importance—premium items contribute but less than volume
- `StockCode`: Moderate—order diversity trends toward higher values
- `Hour`: Minimal—temporal factors insignificant for high-value prediction

This enables real-time high-value transaction identification through quantity and unit price monitoring.

## VI. Conclusion

This research demonstrated comprehensive data mining through: (1) rigorous preprocessing improving data quality 26%; (2) EDA uncovering actionable temporal patterns and retention challenges; (3) RFM-based K-Means segregating 4,191 customers into four segments; (4) MBA identifying 235 frequent itemsets with FP-Growth showing 50% efficiency improvement; and (5) XGBoost achieving 96.9% accuracy for high-value forecasting.

Findings enable data-driven strategies: personalized reactivation for "At Risk" customers, premium programs for "Champions," strategic bundling based on associations, real-time high-value order detection, and bulk-discount incentives leveraging the Quantity-driven revenue model.

## Acknowledgment

The authors acknowledge the UCI Machine Learning Repository for providing the Online Retail Dataset.

## References

[1] D. Chen, S.L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation," *Journal of Database Marketing & Customer Strategy Management*, vol. 19, no. 3, pp. 197–208, 2012.

[2] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1–12, 2000.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD*, pp. 785–794, 2016.