E-Commerce Customer Behavior and Product Popularity Analysis

Presented By
Divya Prakash - 002055546
Madhumitha Nandhikatti - 002304994

Introduction

E-commerce growth has created massive customer interaction data. Every browse, order, and reorder provides behavioural patterns useful for improving inventory, pricing, recommendations, and marketing. Online grocery shopping is especially repetitive and dynamic, making data-driven insights essential.

Motivation

Customer Behaviour & Segmentation - Customers differ in frequency, spending, and reorder tendencies. Segmenting them enables targeted marketing, better retention, and optimized promotions.

Product Popularity Prediction - Some products consistently dominate sales. Understanding why helps with inventory management, assortment planning, and advertising.

Goals

Customer Segmentation using RFM behavioural features and **K-Means clustering**.

Product Popularity Prediction using machine learning models:

- Logistic Regression
- Decision Tree
- Random Forest

Methodology

Data Integration & Architecture

- Merged six Instacart tables (orders, products, aisles, departments, prior/train data) using product_id, order_id, user_id.
- Optimized memory for 32M+ rows using efficient Pandas operations and appropriate data types.
- Validated merged data by checking row counts, missing values, duplicates, and categorical consistency.

RFM Feature Computation

- Recency: days since last purchase
- Frequency: total number of orders
- Monetary: average basket size / spend
- All RFM features were scaled using StandardScaler before clustering to ensure equal contribution in K-Means.

Machine Learning Pipeline

- Selected key product features and engineered popularity labels using the 75th percentile threshold.
- Applied stratified train–test split to balance classes.
- Trained Logistic Regression, Decision Tree, and Random Forest models.
- Evaluated using accuracy, precision, recall, F1-score, and ROC-AUC.

Project Workflow Overview

1. Data Acquisition

• Loaded Instacart 2017 Online Grocery Dataset into Jupyter Notebook

• Imported all CSV files (orders, products, aisles, departments, prior/train data)

2. Data Cleaning & Merging

• Combined multiple relational tables using shared keys

• Built a unified dataset with product, user, order, and reorder information

3. Exploratory Data Analysis (EDA)

- Visualized:
  - Top departments & aisles
  - Daily and hourly shopping trends
  - Reorder ratio distributions

4. Feature Engineering

• Generated customer RFM features (Recency, Frequency, Monetary)

• Computed total orders and reorder ratios for products

• Encoded categorical fields (aisle, department)

5. Customer Segmentation (Unsupervised)

• Applied K-Means (k = 4) on standardized RFM features

• Interpreted clusters using scatter plots and box plots

6. Product Popularity Modelling (Supervised)

• Labelled products as High Popularity (top 25%) or Low Popularity

• Split into train–test sets

• Trained Logistic Regression, Decision Tree, Random Forest

• Evaluated via accuracy, precision, recall, F1-score, ROC-AUC

Dataset Description

Source: Kaggle – Instacart Market Basket Analysis

Files Used:

File Name	Description
orders.csv	Contains order metadata (user ID, order number, day, hour, day, recency)
order_products_prior.csv	Product-level data for all past orders
order_products__train.csv	Product-level data for labelled orders
products.csv	Product names and IDs
departments.csv	Department categories
aisles.csv	Aisle categories

Important Features: order_id, user_id, product_id, aisle_id, department_id, days_since_prior_order, reordered, order_hour_of_day

Scale: 3M+ rows , 200K+ users , 49K+ products

Exploratory Data Analysis Overview

Department-Level Trends

- Produce and Dairy & Eggs dominate orders → core revenue drivers
- These are frequently purchased staples → require strong in-stock availability
- High volume indicates Instacart is used primarily for routine grocery replenishment

Cross-Category Insights

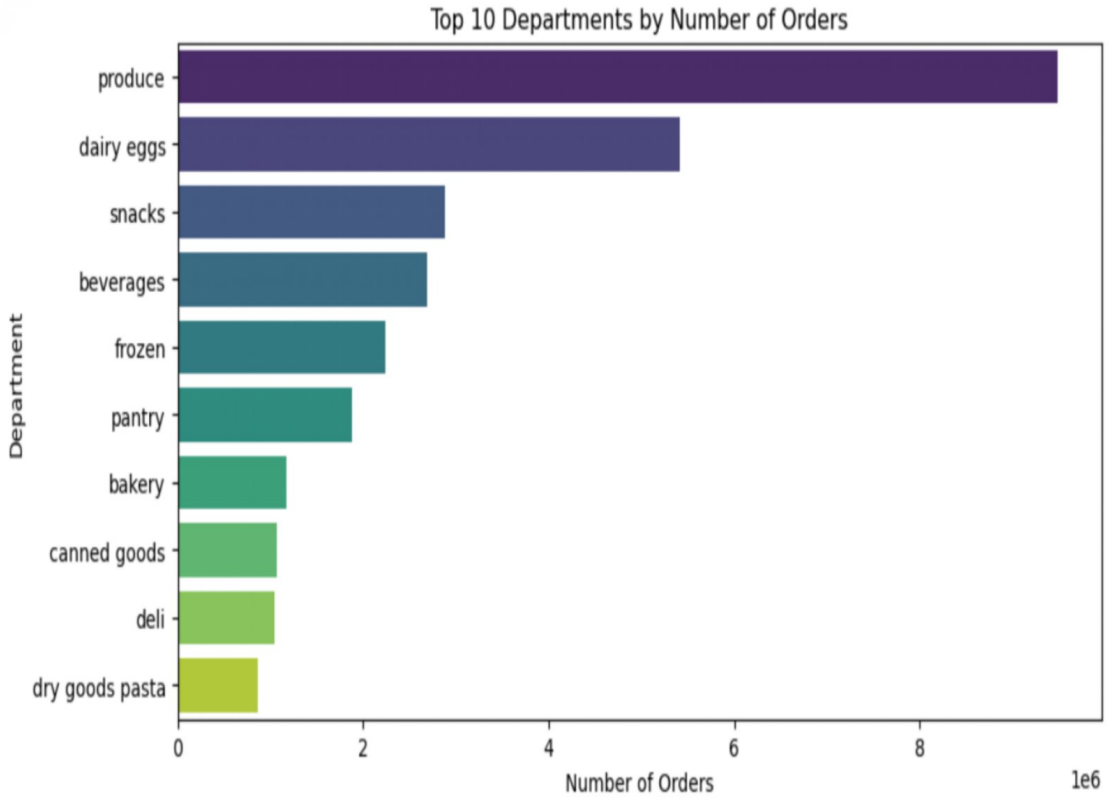
- Fresh fruits/vegetables strongly correlate with dairy, bakery, and packaged foods
- Opportunity for cross-selling (e.g., recommend yogurt with fresh fruit
- Basket patterns suggest customers buy complementary items together

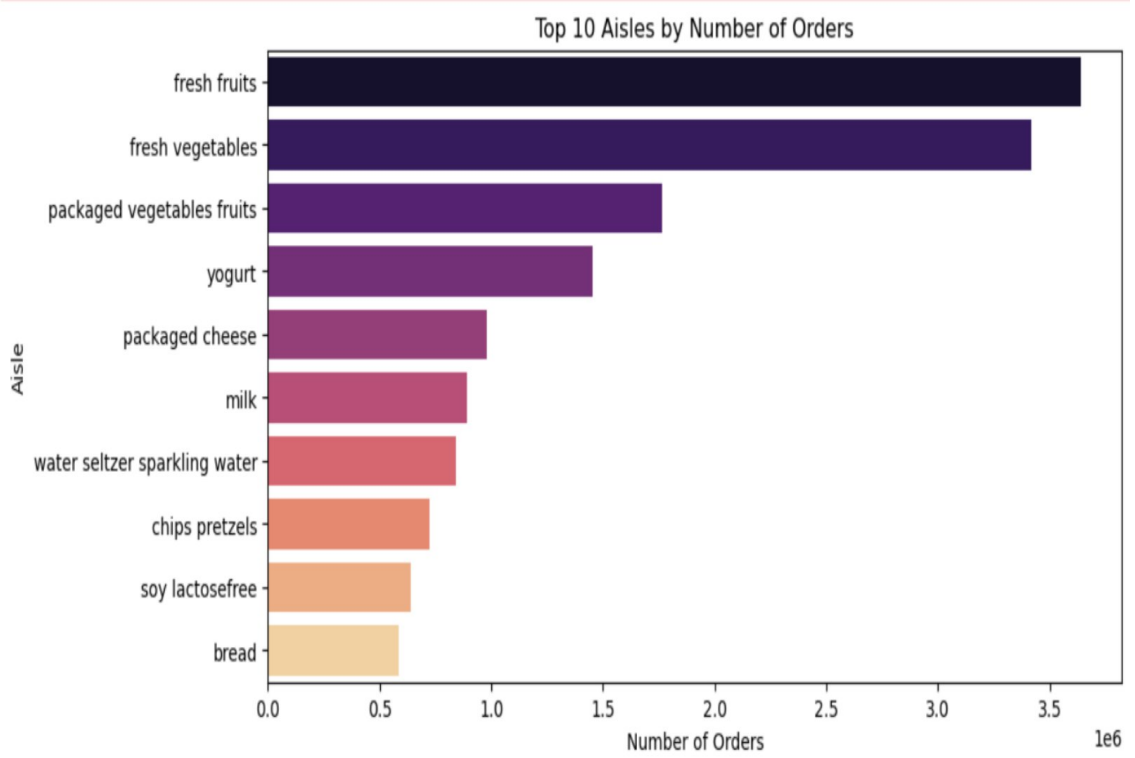
Temporal Patterns

- Peak ordering: 10 AM – 4 PM
- Operational implication: allocate more staff, delivery slots, and inventory restocking during midday
- Early morning and late nights show low demand → no need for extended operational hours

Weekly Order Behaviour

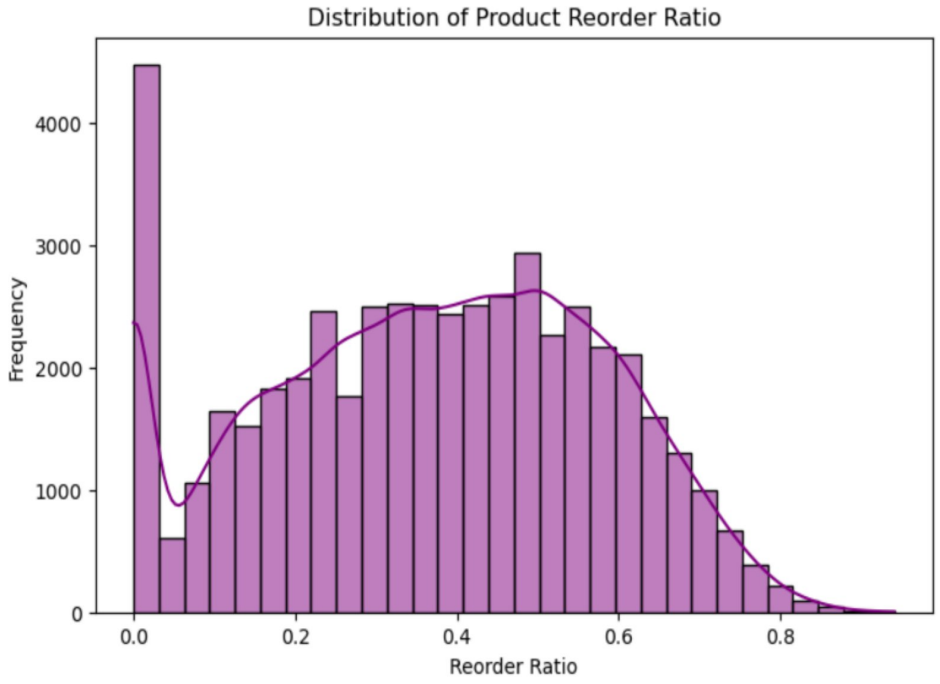
- Sundays & Mondays → highest order volume
- Reflects typical weekly grocery routines
- Opportunity: weekend promotions, family bundles, recipe kits

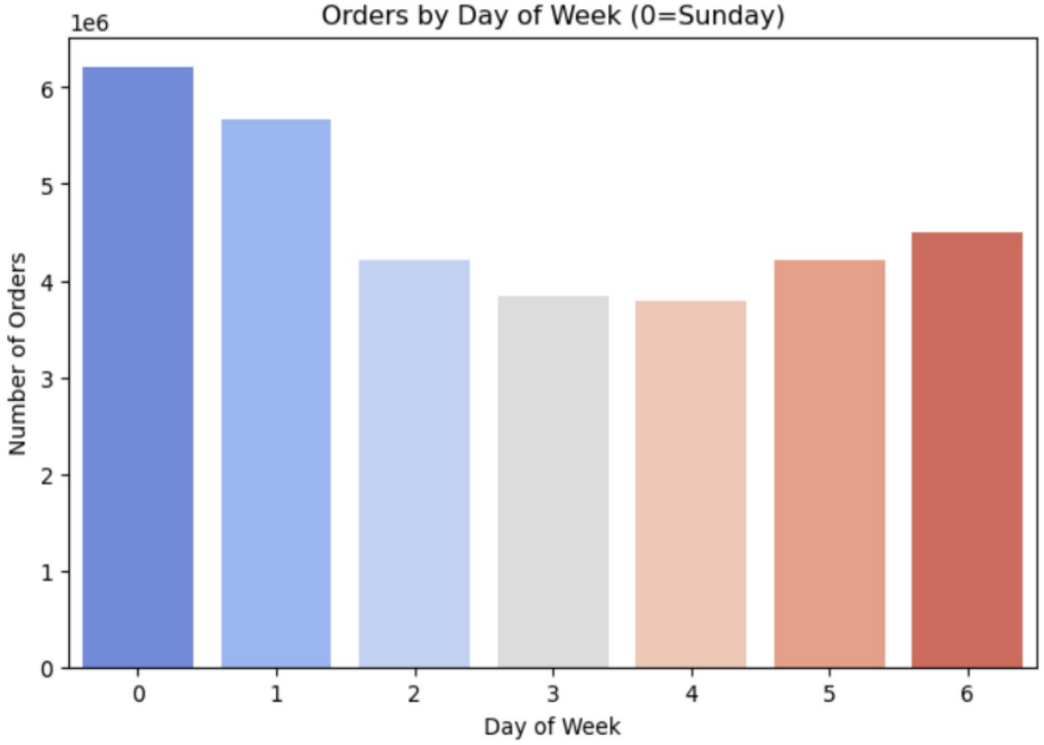




Insights:
Produce is the most frequently ordered department.
Dairy & eggs, snacks, beverages follow closely.
Shows customers prioritize fresh groceries.

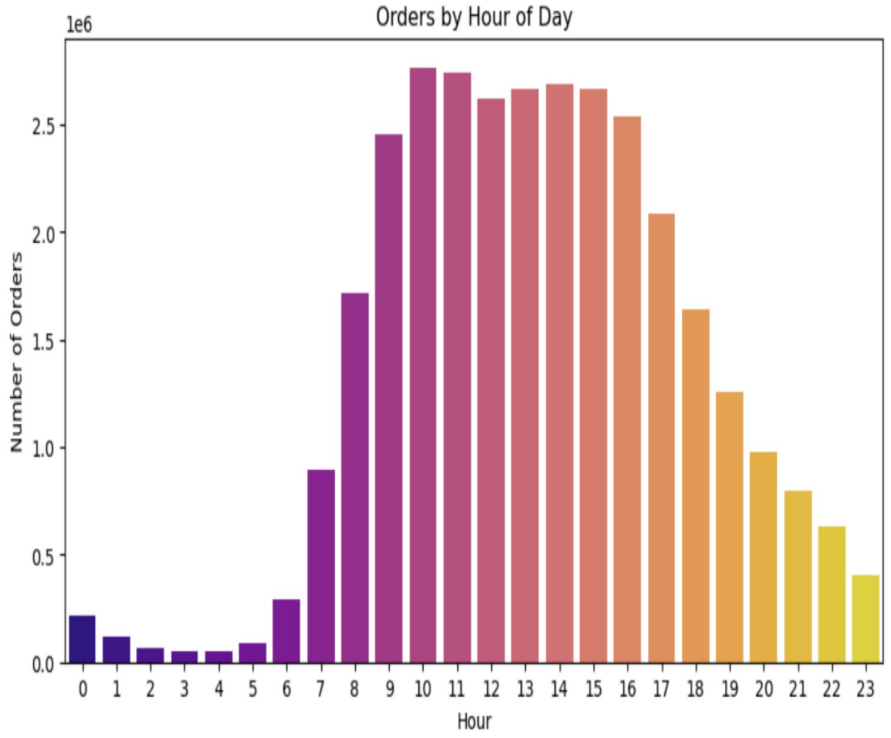
Insights:
Fresh fruits and vegetables dominate
Packaged foods and dairy products also appear frequently

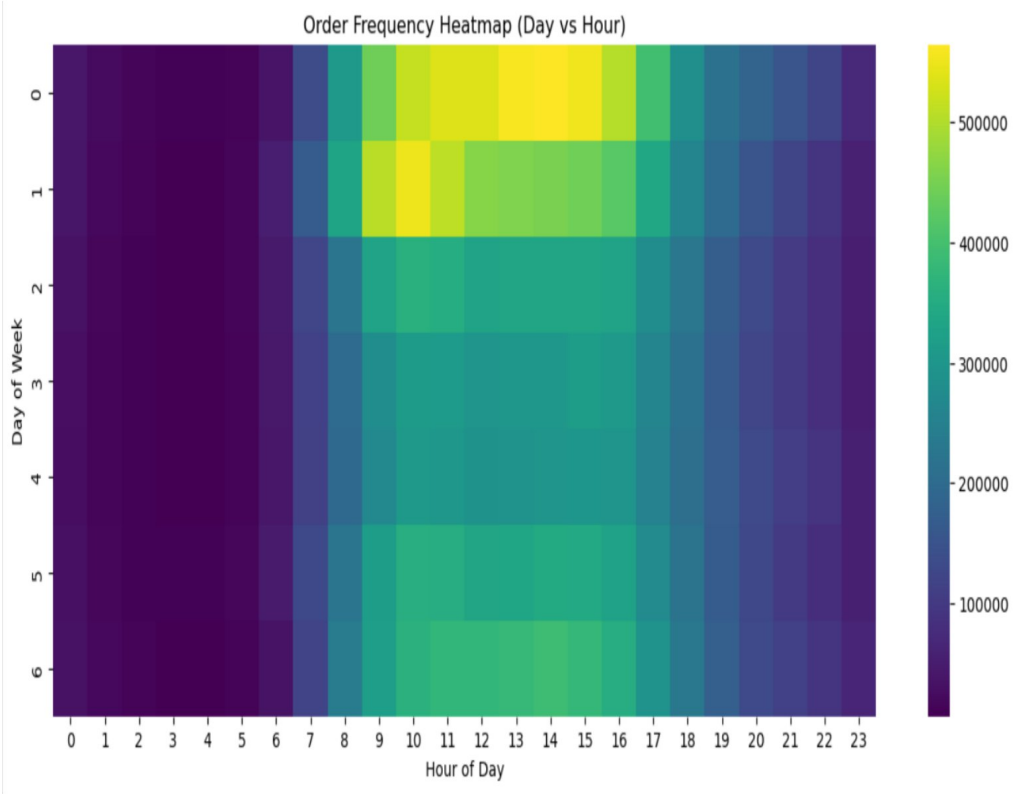




Insights:
Many products have low reorder ratio
A smaller number have very high loyalty
Indicates selective repeating patterns

Insights:
Highest order volume on Sundays and Mondays
Mid-week orders drop significantly





Insights:
Most orders occur between 10 AM – 4 PM
Early morning and late night have low activity

Insights:
Peak ordering hours: 10 AM – 2 PM
Sunday is the busiest day across most hours

Customer Segmentation

Cluster Behaviour Insights

- Four clusters show clear behavioural differences.
- Cluster 3: High-frequency, highly loyal → Instacart's "VIP customers."
- Cluster 1: Low-value users with minimal activity → require low-cost engagement strategies.
- Helps prioritize marketing resources toward high-value customers.

RFM Insights for Personalization

- High-frequency users respond well to personalized offers, free delivery, and loyalty programs.
- High-monetary customers (Cluster 0) may prefer bulk discounts, recipe bundles, or meal kits.
- Cluster 2 (recent but light shoppers) are strong candidates for first-time coupons and onboarding promotions.

Business Impact

- Enables differentiated customer journeys and targeted communication.
- Cluster 3 users are ideal for subscription services (e.g., Instacart Express).
- At-risk users (e.g., Cluster 1 with high recency) can be targeted with win-back campaigns.
- Improves retention, customer lifetime value, and marketing ROI.

Visualization Interpretation

- Scatter plots clearly show natural group separation (Recency vs Frequency, Frequency vs Monetary).
- Cluster 3 appears as a distinct high-frequency group.
- Boxplots confirm Cluster 0's high monetary behaviour.
- Visual patterns validate that clustering results are meaningful and actionable.

Customer Segmentation

Step 1: Feature Calculation

- Recency = days since last purchase
- Frequency = total number of orders
- Monetary = average basket size

Step 2: Clustering

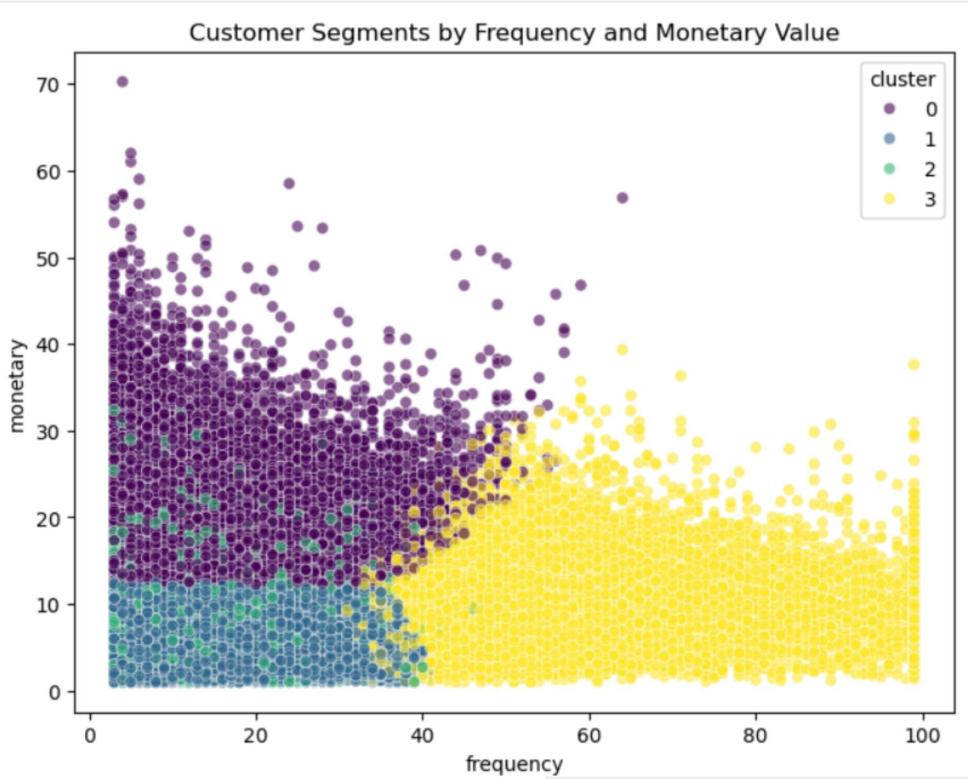
We used *K=4* clusters based on silhouette stability and interpretability.

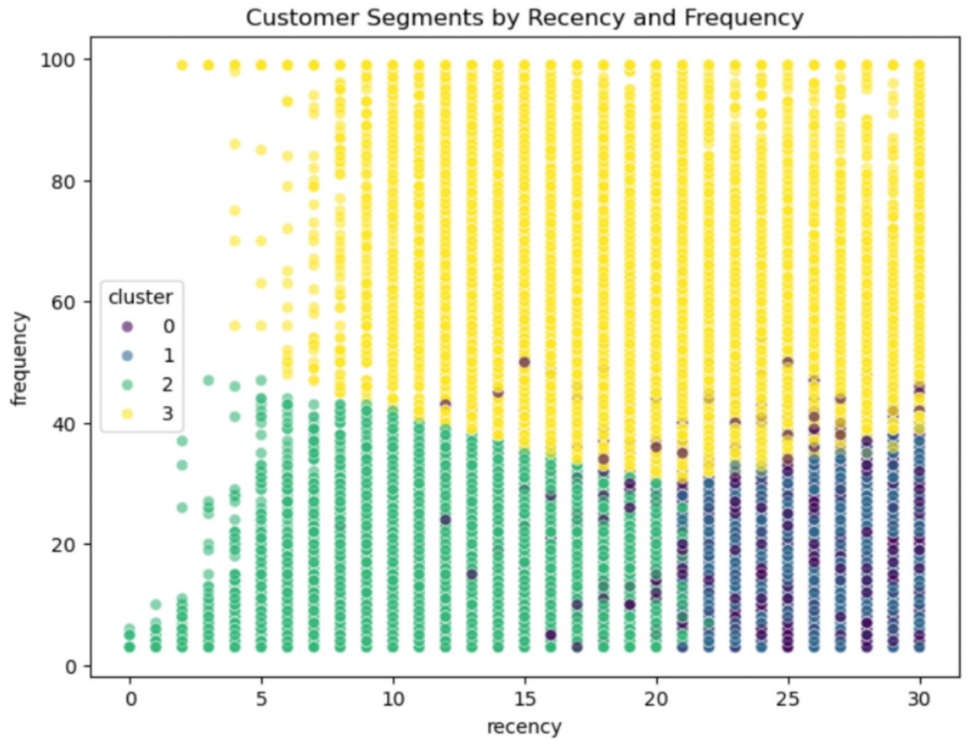
Step 3: Cluster Interpretation

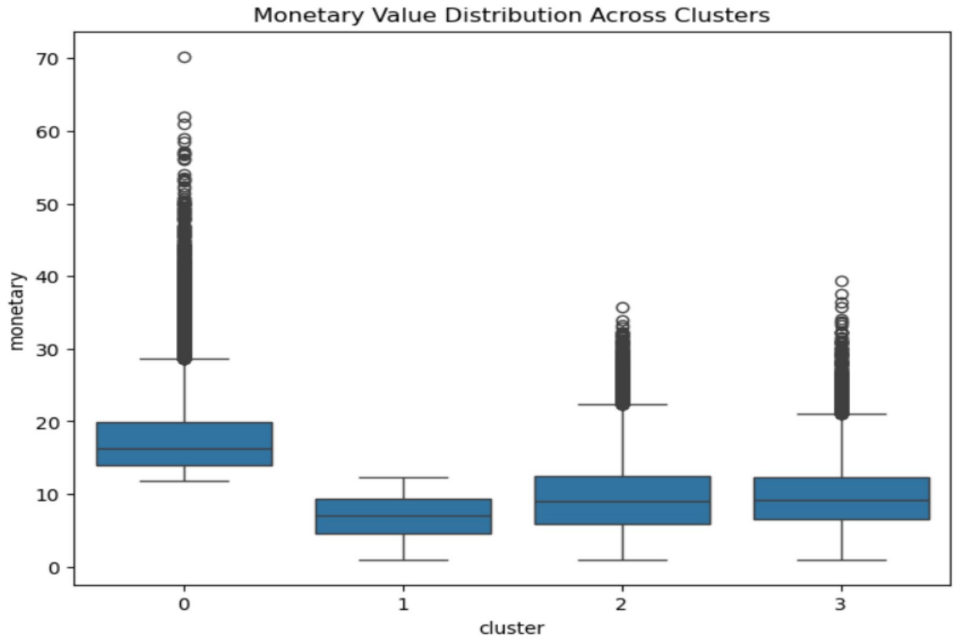
cluster	recency	frequency	monetary
0	28.866997	12.582353	17.700981
1	29.473048	11.386878	6.918750
2	13.158978	8.712878	9.584936
3	22.230458	57.112410	9.825112

Example interpretation:

Cluster	Description
Cluster 0	High monetary, medium frequency customers
Cluster 1	Low monetary, low frequency
Cluster 2	Medium monetary, low frequency
Cluster 3	High frequency, moderate monetary


Customer Segments by Frequency and Monetary Value
cluster
0
1
2
3
monetary
frequency
0
10
20
30
40
50
60
70
0
20
40
60
80
100


Customer Segments by Recency and Frequency
cluster
0
1
2
3
frequency
recency
0
20
40
60
80
100
0
5
10
15
20
25
30


Monetary Value Distribution Across Clusters
monetary
cluster
0
10
20
30
40
50
60
70
0
1
2
3

Product Popularity Prediction

Step 1: Popularity Label

- Popular = Top 25% of products by total historical orders
- Not Popular = Remaining 75%
- Reflects retail long-tail effect: few products drive most sales

Class Distribution

- Popular: ~12,424 products
- Not Popular: ~37,253 products
- Slight imbalance, but suitable for supervised ML tasks

Models Used (Supervised ML)

1. **Logistic Regression (Baseline)**

• Fast, interpretable linear model

• Provides benchmark

• Limited for nonlinear patterns

2. Decision Tree

• Learns hierarchical rules

• Captures nonlinear interactions

• Simple but prone to overfitting

3. Random Forest (Best Performing)

• Ensemble of many trees (majority vote)

• Reduces overfitting, improves stability

• Handles category interactions and nonlinear behaviour

Why Random Forest Performed Best

• Captures complex relationships

• Robust to noise & class imbalance

• Models feature interactions

• Achieved highest ROC-AUC and accuracy

Performance Insights

Key Features Used

• Reorder Ratio — strongest predictor

• Department Encoding — broad category influence

• Aisle Encoding — granular category signals

Feature combination helps models understand how much, how often, and in what context products are purchased.

	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
0	Logistic Regression	0.754630	0.521500	0.229376	0.318614	0.771790
1	Decision Tree	0.786333	0.591322	0.471630	0.524737	0.824053
2	Random Forest	0.789452	0.607436	0.447082	0.515067	0.839120

Final Outcome of Supervised Learning

- Product popularity is highly predictable using behavioral + category features
- Random Forest provides the most reliable ranking of popularity probability
- Enables:

– Smarter inventory planning

– More effective promotions

– Personalized product recommendations

– Better demand forecasting

Feature Importance

	Feature	Importance
0	reorder_ratio	0.727388
2	dept_encoded	0.146130
1	aisle_encoded	0.126481

Example insights:

– Reorder ratio is the strongest predictor of popularity

– Department and aisle categories have moderate influence

Conclusion

Customer Insights

- RFM + K-Means successfully segmented customers into 4 distinct groups based on recency, frequency, and basket size.
- Segments reveal differences in loyalty, engagement, and spending.
- Results support personalized marketing, retention strategies, and customer lifecycle management.

Product Insights

- ML models classified products into high vs low popularity tiers.
- Random Forest delivered the strongest performance across all metrics.
- Feature importance shows reorder ratio is the most powerful predictor of popularity.
- Helps optimize inventory, reduce stockouts, and enhance recommendations.

Conclusion

Operational Insights

- Peak ordering times and weekly rhythms highlight opportunities for staffing, replenishment, and delivery optimization.
- Produce and dairy dominance underscores the need for strong supply chain control for perishables.
- Cross-category patterns support better cross-selling and bundling strategies.

Overall Impact

This study demonstrates that combining EDA, customer segmentation, and predictive modelling creates a strong analytic framework that supports strategic decision-making in e-commerce.

Data-driven approaches improve customer satisfaction, operational efficiency, and long-term competitiveness.

Future Extensions

Future Work

- Incorporate pricing, discounts, and review sentiment
- Add time-series forecasting for product demand
- Build deep learning models for personalized recommendations
- Use sequence models to analyze shopping behaviour over time
- Compare model performance across regions or customer cohorts

These extensions would enhance practical value and deepen insights for large-scale e-commerce applications.

References

- Instacart Online Grocery Shopping Dataset, Kaggle.
- Scikit-learn Documentation: https://scikit-learn.org/
- Seaborn Visualization Library
- Pandas Documentation
- Matplotlib Documentation

Thank you