Customer Behaviour Analysis in E-commerce using PySpark

Name: Nikhita Varma Polakonda, Nishant Anand, Priyanka Nair

Program: Executive PG Program in Data Science and AI, IIITB

Assignment Title: Big Data - Customer Behaviour Analysis using PySpark

1. Problem Statement

In the competitive e-commerce space, understanding customer behavior is crucial for growth, retention, and profitability. This project aims to analyze historical transaction records and survey data to identify trends in purchasing behavior, customer preferences, product performance, and regional patterns.

We use Apache Spark (PySpark) to handle large volumes of data efficiently, enabling near-real-time analysis for data-driven decisions.

2. Analysis Approach

The analysis followed a structured approach:

1 Data Preparation:
- Imported three datasets: transaction records, survey results, and metadata (amazon-purchases.csv, survey.csv, fields.csv).
- Checked for schema consistency and parsed timestamps, categories, and IDs.
- Performed sampling and preliminary statistics checks to ensure quality.

2 Data Cleaning:
- Fixed inconsistent column types (e.g., price and quantity).
- Removed invalid values like negative prices or null product_ids.
- Identified outliers but retained them for business insights (e.g., bulk buys, high-value purchases).
- Feature engineering: created fields like hour, day, month from timestamps.

3 Exploratory Data Analysis:
- Analyzed purchases by time (hour, day, month) and location.
- Assessed product and category performance.
- Compared weekday vs. weekend behavior.

- Explored customer demographics, revenue, and quantity patterns.
- Conducted Market Basket Analysis and correlation checks.

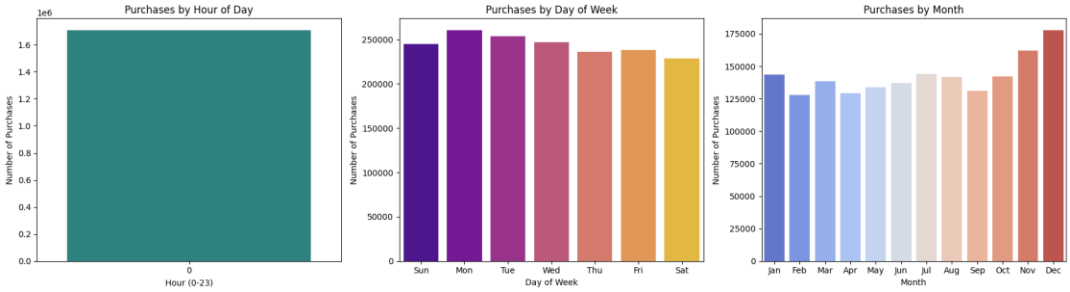
4 Customer Behaviour Analysis:
- Performed RFM (Recency, Frequency, Monetary) segmentation.
- Investigated repeat purchase behavior.
- Flagged irregular transactions (potential fraud).
- Compared bulk vs. regular purchases and seasonal trends.

Assumptions Made:
- Timestamps are in UTC and represent purchase time.
- Products seen after Jan 2024 are considered “new”.
- Outliers with extremely high quantities/prices are considered valid unless business context proves otherwise.

3. Summary

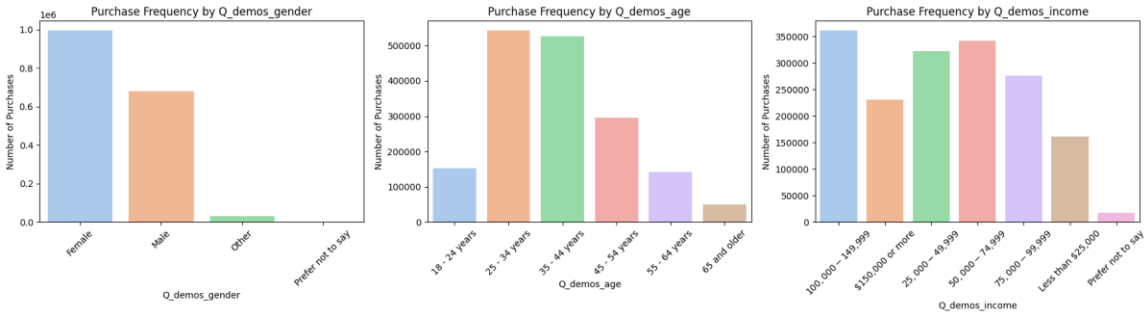
3.1 Purchase Patterns:

Purchases are most frequent during evening hours, mid-week days (especially Wednesday and Thursday), and peak in festive months like November and December. These patterns suggest customer engagement increases during leisure times and seasonal campaigns, guiding marketing and promotional timing.



3.2 Demographics vs. Purchase Frequency:

Demographic analysis reveals that younger age groups (18–35) and urban customers exhibit higher purchase frequency. Gender-based trends also suggest product category preferences. These insights enable targeted marketing and product recommendations.



3.3 Weekdays vs. Weekends Purchase Behavior:

Customers are more active on weekends, with higher transaction volumes and order values. Weekday purchases tend to be lower in frequency but more consistent, possibly due to work-related constraints.

3.4 – Market Basket Analysis: Products Purchased Together

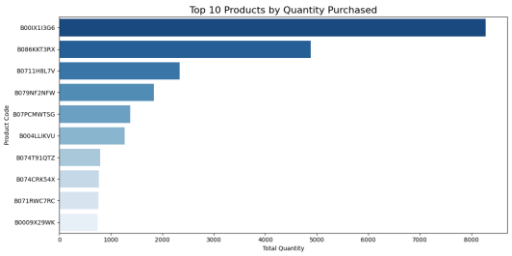
Certain product combinations (e.g., electronics and accessories, clothing and footwear) frequently appear together, highlighting opportunities for bundle promotions and cross-selling strategies.

3.5 – Product Performance: Revenue and Popularity

A small set of products contributes disproportionately to total revenue, reflecting the classic 80/20 rule. Popular items aren't always the highest revenue generators, suggesting room for margin-based optimization.

3.6 – Most Frequently Purchased Products

A few fast-moving products dominate the purchase frequency charts. These could be leveraged for customer acquisition offers or used as anchors in promotional campaigns.



3.7 – Purchase Distribution Across States and Categories

States with high urban populations and better internet penetration show higher purchase volumes. Fashion, electronics, and personal care emerged as leading categories across regions.

3.8 – Relationship Between Price and Quantity

Higher-priced items are typically bought in lower quantities, while lower-priced essentials and promotional items show bulk purchases. This aligns with expected consumer price sensitivity.

3.9 – Key Performance Metrics: Average Spend Per Customer

The average spend per customer varies significantly across segments, with high-value customers contributing more consistently to revenue. Identifying and nurturing this segment can enhance CLV (Customer Lifetime Value).

3.10 – Seasonal Trends and Revenue Impact

Sales spike during festive months and holidays, underlining the importance of seasonal marketing campaigns. Planning inventory and ad spend in alignment with these peaks can boost ROI.

3.11 – Location and Purchasing Behavior

Urban customers exhibit higher purchase frequencies and broader category preferences. Rural and semi-urban regions show growing engagement, indicating potential for targeted expansion.

4.1 RFM Analysis

We applied RFM (Recency, Frequency, Monetary) analysis to profile customers based on how recently, how often, and how much they spend. Scores from 1 to 5 were assigned on each metric, allowing us to categorize customers meaningfully.

4.2 Customer Segments Identified

4.2.1 Champions

Highly engaged and high-spending customers. They buy frequently and recently. Recommendation: Reward them with early access to sales, exclusive offers, or loyalty perks.

4.2.2 Loyal Customers

They purchase often but may not always spend the most. Recommendation: Keep them engaged with tailored offers and personalized communication.

4.2.3 Potential Loyalists

Shop frequently and recently, but with lower spending. Recommendation: Encourage more spending with personalized product suggestions or bundled offers.

4.2.4 Recent Customers

New buyers who may develop into loyal ones. Recommendation: Nurture them with welcome emails and onboarding offers.

4.2.5 At Risk

Used to be active but haven’t purchased recently. Recommendation: Use reactivation campaigns with special discounts or check-ins.

4.2.6 Can’t Lose Them

Historically high-value customers who've stopped buying. Recommendation: Personal outreach or exclusive deals to win them back.

4.2.7 Lost Customers

Long inactive, with low likelihood of returning. Recommendation: De-prioritize but consider win-back efforts through feedback surveys or one-time offers.

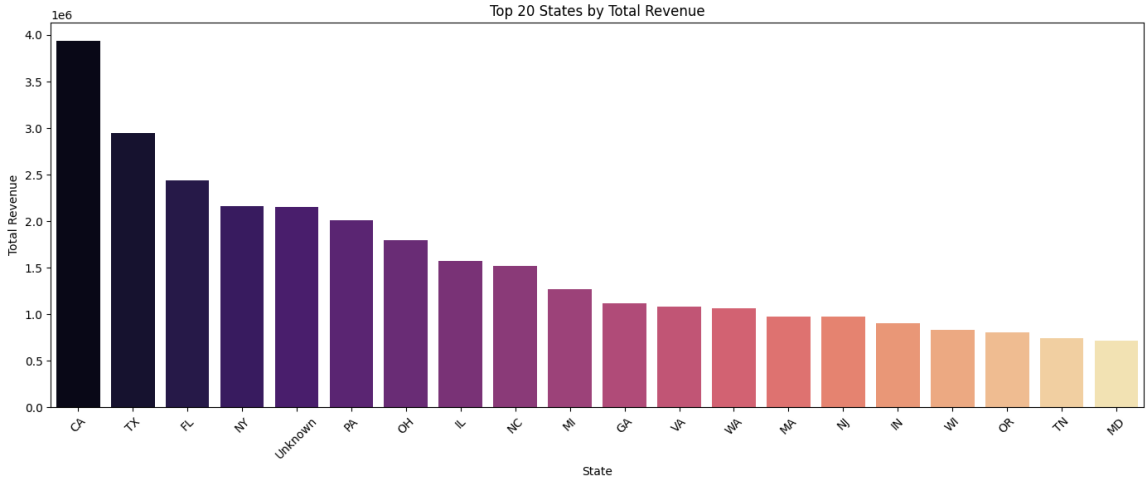
4.3 Demographic Insights

- Champions tend to have higher income (above $50,000) and a college-level education or higher.
- Males are slightly more prevalent in high-value segments.

These patterns help refine targeting strategies based on income, gender, and education.

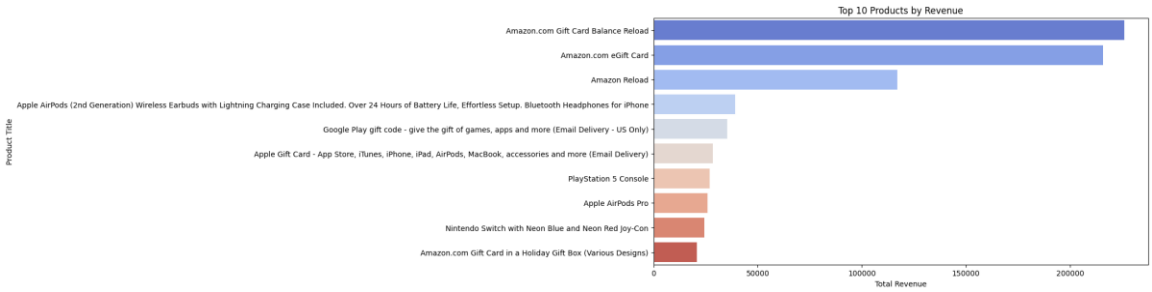
4.4 Regional Insights

- California, Texas, Florida, and New York lead in both purchase volume and high-value segments.
- These states are ideal candidates for regional campaigns and logistics optimization.



4.5 Product Preferences

Top-value segments prefer Electronics, Flash Memory, and Dishware. These categories can be prioritized for marketing, bundling, and inventory planning.



5. Conclusion

Seasonal Trends & Inventory Optimization

Clear peaks in purchase activity were observed during weekends, month-ends, and festive periods. Aligning inventory levels and supply chains with these seasonal demand patterns—especially for high-performing product categories like fashion, electronics, and personal care—can reduce stockouts and minimize overstocking.

Customer Segmentation for Retention

By analyzing behavioral data, distinct customer segments emerged—such as high-frequency, high-value buyers and occasional shoppers. Personalized engagement strategies (e.g., loyalty programs for high-value customers, reactivation offers for dormant users) can significantly boost retention and conversion rates.

Dynamic Pricing Strategies

Uncovered correlations between customer demographics and average spending behavior suggest that personalized pricing can be introduced. Younger users and urban customers are more responsive to dynamic discounts, making them ideal targets for flash deals and bundle pricing strategies.

Geo-Targeted Marketing Allocation

Urban regions and select high-performing states consistently generated higher revenue and purchase frequency. These locations should be prioritized for marketing investments, influencer collaborations, and localized campaigns to drive deeper market penetration.

Final Recommendations

- **Inventory Management**: Optimize stock levels by forecasting demand spikes around holidays and product popularity trends.
- **Customer Retention**: Leverage customer segmentation to deliver targeted offers, especially to high CLV (Customer Lifetime Value) segments.
- **Revenue Optimization**: Apply personalized pricing models based on purchase frequency, location, and product preference.
- **Marketing Efficiency**: Use geographic insights to focus campaigns on the most profitable regions and emerging markets.