International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

From Transactions to Trends: An Integrated Analysis of Customer Shopping Behavior

ASTHA JHA, Prof. Kiran Sharma ,

Assistant Professor, Department of Computer Science and Engineering,

Parul Institute of Technology, Parul University, Gujarat, India

Students of Computer Science and Engineering, Parul Institute of

Engineering and Technology, Parul University, Gujarat, INDIA

Abstract

This paper presents a comprehensive analysis of customer shopping behavior through the examination of 3,900 transactional records derived from a retail dataset. The study employs a multi-tool analytical framework integrating Python for data preprocessing and feature engineering, SQL for structured business-oriented querying, and Power BI for the development of interactive dashboards and visualizations. The investigation encompasses five primary dimensions: demographic influences on purchasing decisions, spending patterns across various segments, product preferences and performance metrics, subscription behavior, and discount utilization patterns. Key findings include statistically notable revenue differentials by gender and age group, identification of top-rated and best-selling product categories, and the delineation of distinct customer loyalty segments. The integrated analytical approach yielded actionable business insights, culminating in strategic recommendations including subscription growth initiatives, evidence-based discount policy reforms, and targeted marketing strategies. The research demonstrates the practical value of combining multiple data analytics tools to derive meaningful intelligence from transactional retail data.

Keywords: *Customer Behavior Analytics, Retail Data Mining, Python, SQL, Power BI, Transactional Analysis, Business Intelligence, Segmentation, Subscription Behavior, Discount Optimization*

1. Introduction

The rapid proliferation of e-commerce platforms and digital point-of-sale systems has generated an unprecedented volume of transactional data. Retailers and businesses are increasingly recognizing the strategic value of this data as a resource for understanding customer preferences, optimizing inventory, and refining marketing strategies. However, the transformation of raw transactional records into actionable business intelligence requires a robust analytical framework capable of handling structured queries, statistical computations, and intuitive visualizations.

Customer behavior analysis has emerged as a critical domain within business intelligence, enabling organizations to move beyond reactive decision-making toward proactive, data-driven strategy formulation. Traditional methods of market research, reliant on surveys and focus groups, have been supplemented and in many cases supplanted by computational analysis of large-scale transactional datasets. The convergence of database management systems, programming languages optimized for data manipulation, and business intelligence visualization tools has opened new avenues for deriving insights that were previously inaccessible.

This research addresses the challenge of holistically analyzing customer shopping behavior by leveraging three complementary analytical tools: Python, SQL, and Power BI. The dataset comprises 3,900 transactional records capturing a wide spectrum of customer interactions, including demographic attributes, product categories, payment methods, subscription status, and review ratings. By integrating these tools within a unified analytical pipeline, the study aims to produce a multi-dimensional understanding of customer behavior that transcends what any single tool could achieve independently.

The primary objectives of this research are: (1) to analyze demographic influences on purchasing patterns and revenue generation; (2) to identify high-performing product categories and understand product preference dynamics; (3) to investigate subscription behavior and its correlation with customer engagement; (4) to evaluate discount utilization patterns and their impact on sales volume; and (5) to segment customers into loyalty tiers based on behavioral indicators.

2. Literature Review

The field of customer behavior analytics has a rich body of literature spanning disciplines including marketing science, information systems, and data engineering. Early work by Kotler and Keller (2016) established foundational frameworks for understanding consumer decision-making processes, emphasizing the role of demographic, psychological, and social factors. These conceptual models have since been operationalized through quantitative methods as transactional data has become more abundant.

Recency, Frequency, and Monetary (RFM) analysis, introduced by Hughes (1994), remains one of the most widely applied methodologies for customer segmentation in retail analytics. RFM models categorize customers based on how recently they made a purchase, how frequently they transact, and how much monetary value they contribute. Subsequent studies have extended RFM with machine learning approaches, including K-means clustering and decision trees, to achieve more nuanced segmentation (Hosseini et al., 2010; Sohrabi & Khanlari, 2007).

The integration of SQL-based data warehousing with Python analytics has been examined extensively in the context of business intelligence pipelines. Kimball and Ross (2013) documented best practices for dimensional modeling in data warehouses, while the emergence of pandas and NumPy libraries in Python has democratized exploratory data analysis. Studies by McKinney (2012) demonstrated Python's efficacy in handling large-scale tabular datasets, feature engineering, and statistical profiling.

Power BI has been recognized as an industry-leading tool for business intelligence visualization. Research by Turban et al. (2017) highlights the importance of interactive dashboards in enabling non-technical stakeholders to engage with analytical findings. Recent literature has emphasized the value of multi-tool analytical frameworks; Singh and Dhall (2020) demonstrated that combining Python for preprocessing, SQL for aggregation, and BI tools for visualization produces more comprehensive insights than siloed approaches.

3. Methodology

3.1 Dataset Description

The dataset utilized in this research comprises 3,900 transactional records collected from a retail environment. Each record captures a discrete customer transaction and includes the following attributes: Customer ID, Age, Gender, Item Purchased, Category, Purchase Amount (USD), Location, Size, Color, Season, Review Rating, Subscription Status, Payment Method, Shipping Type, Discount Applied, Promo Code Used, Previous Purchases, Preferred Payment Method, and Frequency of Purchases. The dataset encompasses both categorical and numerical variables, providing a comprehensive profile of each transaction and the associated customer.

The dataset exhibits demographic diversity with customers ranging in age from 18 to 70 years, with a mix of male and female respondents. Product categories include Clothing, Footwear, Outerwear, and Accessories. Geographically, the data spans multiple US states, offering regional diversity relevant to localized marketing analysis.

3.2 Data Preprocessing with Python

Python version 3.10, in conjunction with the pandas and NumPy libraries, was employed for data preprocessing. The preprocessing pipeline encompassed the following stages: missing value detection and imputation, duplicate record removal, data type normalization, and outlier identification using the interquartile range (IQR) method. Categorical variables were encoded using label encoding and one-hot encoding as required by subsequent analytical procedures.

Feature engineering was performed to derive new analytical variables from existing attributes. Notable engineered features include Age Group (binned into categories: 18-25, 26-35, 36-50, 51+), Season-Category Interaction, and a Customer Value Score computed as a normalized composite of Purchase Amount, Frequency of Purchases, and Previous Purchases.

3.3 SQL-Based Business Queries

Structured Query Language (SQL) was utilized to execute targeted business queries on the cleaned dataset, which was loaded into a SQLite database for this purpose. Queries were designed to address specific analytical questions, including revenue aggregation by gender and age group, identification of top-selling and highest-rated products, subscription rate analysis, and cross-tabulation of discount usage against purchase volume. SQL's declarative syntax facilitated efficient aggregation and filtering operations on the dataset.

3.4 Power BI Visualization

Power BI Desktop was employed to create an interactive multi-page dashboard for visual exploration of the analytical findings. The dashboard comprised five primary report pages: Demographic Analysis, Product Performance, Spending Patterns, Subscription and Loyalty Analysis, and Discount and Promotional Analysis. DAX (Data Analysis Expressions) measures were authored to compute KPIs including Total Revenue, Average Order Value, Subscription Rate, and Customer Retention Proxy.

3.5 Analytical Framework

The analytical framework adopted in this study follows a structured pipeline: raw data ingestion, Python-based preprocessing and feature engineering, SQL-based querying for business-specific aggregations, and Power BI for visualization and insight communication. This end-to-end pipeline reflects industry-standard practices in business intelligence and ensures reproducibility and traceability across analytical stages.

4. Results and Findings

4.1 Demographic Analysis

The demographic analysis revealed significant patterns in purchasing behavior across gender and age group dimensions. Male customers constituted approximately 68% of the transaction volume, while female customers accounted for the remaining 32%. However, average purchase value per transaction was marginally higher for female customers, suggesting higher per-unit spending propensity. The following table summarizes revenue distribution by gender:

Gender	Transaction Count	Total Revenue (USD)	Avg. Purchase Value (USD)
Male	2,652	$230,840	$87.04
Female	1,248	$109,760	$87.95
Total	3,900	$340,600	$87.33

Table 1: Revenue Distribution by Gender

Age group analysis indicated that the 26-35 cohort generated the highest total revenue, followed by the 36-50 group. The 18-25 segment exhibited the highest frequency of purchases but lower average transaction values, consistent with budget-conscious spending behavior typical of younger demographics. The 51+ segment demonstrated lower transaction frequency but consistently higher average purchase values.

4.2 Product Performance

Analysis of product categories revealed that Clothing constituted the dominant category by transaction volume, accounting for approximately 44.7% of all purchases. Footwear ranked second at 26.3%, followed by Outerwear at 16.1% and Accessories at 12.9%. Examination of review ratings indicated that Footwear received the highest

average customer rating (3.82/5.0), suggesting a positive relationship between product satisfaction and category specialization.

Category	Transactions	% Share	Avg Rating	Total Revenue (USD)
Clothing	1,743	44.7%	3.74	$152,185
Footwear	1,026	26.3%	3.82	$89,705
Outerwear	628	16.1%	3.69	$54,880
Accessories	503	12.9%	3.75	$43,830

Table 2: Product Category Performance Metrics

At the item level, Blouse, Jewelry, and Pants emerged as the top three best-selling items by transaction count. Seasonal analysis indicated peak purchasing activity during the Fall and Winter seasons, with Outerwear demonstrating the highest seasonal demand concentration in Winter months.

4.3 Spending Patterns

Spending pattern analysis demonstrated that the majority of purchase amounts were distributed within the $20-$100 range, with a mean purchase value of $87.33 and a standard deviation of $27.91. No extreme outliers were identified following IQR-based filtering. Geographic analysis highlighted Montana, California, and Idaho as the top three states by total revenue contribution. Credit Card and PayPal were the most frequently utilized payment methods, collectively accounting for 58% of transactions.

4.4 Subscription Behavior

Subscription analysis revealed that 27.4% of customers (n=1,069) held active subscription status. Subscribed customers exhibited significantly higher purchase frequency and total lifetime value compared to non-subscribers. The average number of previous purchases for subscribed customers was 26.4, compared to 22.1 for non-subscribed customers. Subscribed customers also demonstrated higher utilization of promo codes (62% vs. 38%), indicating a correlation between subscription engagement and promotional responsiveness.

4.5 Discount and Promotional Analysis

Discount application analysis indicated that 43.6% of transactions involved a discount, while 62.4% utilized a promo code. Cross-analysis of discount usage against purchase amounts revealed that discounted transactions had a marginally lower average purchase value ($84.12) compared to non-discounted transactions ($89.91), suggesting potential price elasticity effects. However, discounted transactions occurred at higher frequency within the subscription segment, partially offsetting revenue reduction through volume increase.

4.6 Customer Loyalty Segmentation

Customers were segmented into three loyalty tiers based on a composite score derived from purchase frequency, previous purchase count, and subscription status. The resulting segments were designated as High-Value Loyalists (18.2%), Engaged Regulars (41.6%), and Occasional Shoppers (40.2%). High-Value Loyalists generated disproportionately high revenue relative to their population share, contributing approximately 34% of total revenue. This finding underscores the importance of retention-focused strategies for this segment.

5. Discussion

The findings of this study carry significant implications for retail business strategy. The demographic analysis confirming higher transaction volumes among male customers, while female customers exhibit higher average spend, suggests that segmented marketing campaigns tailored to gender-specific purchasing behaviors could improve campaign ROI. Age-group analysis provides a basis for lifecycle-oriented product recommendations, with younger cohorts potentially benefiting from value-oriented promotions and older cohorts from premium product positioning.

The dominance of Clothing in transaction volume, combined with Footwear's superior review ratings, presents a strategic opportunity: cross-promotional campaigns linking high-satisfaction Footwear with high-volume Clothing purchases could drive incremental revenue. The seasonal demand concentration in Fall and Winter warrants proactive inventory planning and targeted seasonal campaigns, particularly for Outerwear.

The relatively low subscription penetration rate of 27.4%, despite the demonstrably higher lifetime value of subscribed customers, represents a significant missed opportunity. The data suggests that converting even a modest proportion of non-subscribed high-frequency shoppers to subscription status could yield substantial revenue uplift. This finding aligns with existing literature on subscription economy dynamics (Tzuo and Weisert, 2018), which documents the compounding revenue effects of subscription model adoption.

The analysis of discount utilization highlights a nuanced relationship between promotional activity and revenue. While discounts are associated with slightly lower average transaction values, their higher prevalence among subscribed, high-frequency customers suggests that targeted, subscription-linked discounts may be more effective than broad promotional campaigns. This supports a recommendation to migrate from blanket discount policies toward personalized, behavior-triggered promotional strategies.

The integration of Python, SQL, and Power BI within a unified analytical framework proved synergistically effective. Python's flexibility in data preprocessing and feature engineering ensured data quality and enrichment; SQL provided computationally efficient and business-aligned querying; and Power BI translated complex findings into accessible visual narratives suitable for non-technical stakeholders. This multi-tool approach is consistent with contemporary best practices in enterprise business intelligence.

6. Recommendations

Based on the analytical findings, the following strategic recommendations are proposed:

- Subscription Growth Initiatives: Implement targeted in-app and email campaigns offering trial subscriptions to non-subscribed customers who exhibit high purchase frequency. Offer first-month free or discounted subscription tiers to reduce adoption friction.
- Refined Discount Policy: Transition from broad discount campaigns to personalized, behavior-triggered promotions. Prioritize discount offers for customers in the Occasional Shopper loyalty segment to stimulate re-engagement, while preserving full-price transactions with High-Value Loyalists.
- Targeted Demographic Marketing: Develop gender and age-group specific marketing campaigns. For younger demographics (18-25), emphasize value-for-money propositions and trend-driven content. For the 51+ segment, focus on quality, comfort, and premium positioning.
- Seasonal Inventory Optimization: Increase Outerwear inventory procurement ahead of Fall and Winter seasons based on demonstrated seasonal demand patterns. Develop Winter-specific bundling promotions linking Outerwear with Accessories.
- Cross-Category Promotions: Leverage Footwear's high satisfaction ratings to drive cross-sell campaigns with Clothing, the highest-volume category. Bundle recommendations in e-commerce interfaces can capitalize on this relationship.
- Loyalty Tier Recognition Programs: Formalize the three-tier loyalty segmentation model within a customer relationship management (CRM) system. Provide exclusive benefits to High-Value Loyalists to reinforce retention, and implement re-engagement workflows for Occasional Shoppers.

7. Conclusion

This research has demonstrated the efficacy of an integrated multi-tool analytical framework for deriving actionable insights from retail transactional data. Through the systematic application of Python for data preprocessing, SQL for business-oriented querying, and Power BI for interactive visualization, the study produced a comprehensive characterization of customer shopping behavior across 3,900 records. Key findings include gender and age-based revenue differentials, product category performance metrics, subscription behavior patterns, discount utilization dynamics, and a three-tier customer loyalty segmentation.

The study confirms that transactional data, when subjected to rigorous and multi-dimensional analysis, constitutes a valuable strategic asset for retail organizations. The recommendations derived from this analysis are empirically grounded and addressable through operational changes in marketing strategy, inventory management, and customer relationship management. Future research directions may include the application of machine learning models for predictive customer lifetime value estimation, real-time analytics integration, and longitudinal analysis to track behavioral shifts over time.

The methodology and framework presented in this paper are designed to be transferable across retail contexts, offering a replicable blueprint for organizations seeking to enhance their data analytics capabilities. The findings

contribute to the growing body of literature on applied retail analytics and underscore the strategic imperative of data-driven decision-making in contemporary commerce.

References

[1] Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Systems with Applications, 37(7), 5259-5264.

[2] Hughes, A. M. (1994). Strategic Database Marketing. Probus Publishing Company.

[3] Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley.

[4] Kotler, P., & Keller, K. L. (2016). Marketing Management (15th ed.). Pearson Education.

[5] McKinney, W. (2012). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.

[6] Singh, A., & Dhall, R. (2020). A comprehensive survey on big data analytics frameworks. Journal of Big Data, 7(1), 1-41.

[7] Sohrabi, B., & Khanlari, A. (2007). Customer lifetime value (CLV) measurement based on RFM model. Iranian Accounting & Auditing Review, 14(47), 7-20.

[8] Turban, E., Sharda, R., & Delen, D. (2017). Decision Support and Business Intelligence Systems (10th ed.). Pearson Education.

[9] Tzuo, T., & Weisert, G. (2018). Subscribed: Why the Subscription Model Will Be Your Company's Future. Portfolio/Penguin.

[10] Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques (3rd ed.). Elsevier.