Customer Behavior Analysis - Business Report

1. Executive Summary

This project provides an end-to-end analysis of customer shopping behavior for a retail organization with the objective of identifying revenue drivers, customer segments, loyalty dynamics, discount dependency, and behavioral risks such as churn. The dataset consists of 3,900 transactions with 18 demographic, transactional, and behavioral variables.

Key Business Outcomes:

- Total Revenue generated: ~233,081 USD.
- Loyal customers (customers with more than 5 previous purchases) contribute approximately 89% of total revenue.
- Male customers generate 67.7% of revenue while female customers generate 32.3%; however, Average Order Value (AOV) across genders is almost identical at approximately 60 USD.
- Adults (30–44) and Middle-Aged customers (45–59) are the strongest revenue-generating age groups.
- Subscription status currently does not increase revenue, indicating ineffective monetization of the subscription model.
- Products such as Hats, Sneakers, Coats, Sweaters, and Pants show extremely high dependency on discounts.
- The highest-rated products based on average customer reviews are Gloves, Sandals, Boots, Hats, and Skirts.

Strategic business priorities emerging from this analysis include strengthening customer loyalty, redesigning subscription value propositions, optimizing discount strategy, promoting high-rated products, and expanding high-potential customer segments.

2. Business Context & Objectives

The retail company operates across multiple product categories including Clothing, Footwear, Outerwear, and Accessories. Management has observed variations in purchasing patterns across age groups, genders, and promotional conditions. The business seeks to utilize customer data to improve sales performance, enhance customer satisfaction, and strengthen long-term brand loyalty.

The primary business question guiding this project is:

"**How can the company leverage consumer shopping data to identify trends, improve customer engagement, and optimize marketing and product strategies?**"

Specific Business Objectives:
- Understand customer segmentation based on demographics, loyalty, and purchasing behavior.
- Analyze revenue distribution across gender, age groups, and product categories.
- Evaluate the influence of discounts, subscriptions, and shipping choices on customer spend.
- Identify top-performing products using ratings, volumes, and revenue metrics.
- Provide actionable recommendations for marketing, pricing, loyalty programs, and product placement.

3.Tech Stack Used

Tools & Technologies:
Python (Pandas, NumPy), PostgreSQL, SQL, SQLAlchemy, Power BI, DAX

4. Data Overview & Quality Assessment

Source File: customer_shopping_behavior.csv
Number of Records: 3,900
Number of Attributes: 18

Data Categories:
- Demographics: customer_id, age, gender, location
- Product Information: item_purchased, category, size, color, season
- Transactional Attributes: purchase_amount_(usd), shipping_type, discount_applied, promo_code_used
- Behavioral Indicators: review_rating, subscription_status, previous_purchases, payment_method, frequency_of_purchases

Data Quality Summary:
- The dataset contained minimal missing values, predominantly in the review_rating field.
- Missing review ratings were imputed using median values at the product category level.
- All variables were validated for data type consistency.
- The dataset is cross-sectional in nature and does not contain transaction timestamps, limiting time-based analysis.

5. Data Engineering & Methodology

The analytical pipeline was executed using a combination of Python for data preparation, PostgreSQL for business querying, and Power BI for insight visualization.

Key Data Engineering Steps:
- Standardized all column names into snake_case format.
- Removed special characters from column headers for SQL compatibility.
- Imputed missing values for review_rating using category-level grouped statistics.
- Created derived features including:
  - age_group (Young Adult, Adult, Middle-Aged, Senior)
  - customer_loyalty_segment (New, Returning, Loyal)
- The cleaned dataset was exported into PostgreSQL using SQLAlchemy.
- All business logic queries were executed in SQL and validated against Python outputs.
- Final dashboards were built in Power BI for stakeholder consumption.

6. Core Business Analysis (SQL Findings Q1–Q10)

Revenue by Gender:
Male customers generated approximately 157,890 USD across 2,652 transactions, contributing 67.7% of total revenue. Female customers generated approximately 75,191 USD across 1,248 transactions, contributing 32.3%. Average Order Value remains consistent across genders at around 60 USD, indicating equal spending power and highlighting female customer acquisition as a major growth opportunity.

Revenue by Age Group:
Adults and Middle-Aged customers account for the majority of revenue. AOV remains stable across all age groups, demonstrating that revenue variation is driven primarily by transaction volume rather than basket size.

Customer Loyalty Segmentation:
Loyal customers dominate the business, accounting for nearly 89% of total revenue. Returning and New customers contribute less than 11% combined, indicating weak early-stage customer monetization and poor conversion into loyal segments.

Subscription Impact:
Subscribed customers do not demonstrate higher spend or order value compared to non-subscribers. This confirms that the current subscription program lacks financial incentive value and does not materially influence purchasing behavior.

Product Review Performance:

Products such as Gloves, Sandals, and Boots achieved the highest average customer ratings, making them prime candidates for promotion, homepage placement, and bundling.

Discount Dependency:
Products including Hats, Sneakers, Coats, Sweaters, and Pants show nearly 50% discount dependency. This suggests customer price conditioning and long-term margin risks.

Shipping Type Behavior:
Premium shipping options do not yield significantly higher AOV, indicating that faster shipping reflects urgency rather than purchasing power.

7. Advanced Behavioral Analysis (SQL Findings Q11–Q15)

Purchase Frequency Segments:
Customers purchasing at intervals below 30 days generate significantly higher revenue than infrequent buyers. Revenue concentration increases as purchase frequency increases.

Previous Purchase History:
Customers with high historical purchases contribute the highest total revenue and exhibit strongest repeat behavior. Low-history customers show adequate AOV but poor repeat improvement.

Frequency vs Purchase Amount:
Higher purchase frequency strongly correlates with higher total customer value, confirming behavioral momentum effects.

Discount Behavior vs Frequency:
Discount-heavy customers drive immediate revenue volume but exhibit reduced long-term consistency and higher churn risk.

Churn Risk Identification:
Customers with high previous purchases but average purchase gaps exceeding 45 days represent the highest churn risk segment. These users must be prioritized for proactive retention strategies.

8. Power BI Dashboard Summary

The Power BI dashboard presents:
- Revenue distribution across age groups and genders.
- Product performance and discount dependency.
- Subscription vs non-subscription analysis.
- Customer loyalty segmentation.
- Purchase frequency behavior.

-A dedicated High Churn-Risk Customers KPI built using DISTINCT customer-level churn flags to identify retention risk in real time.

The dashboard supports dynamic slicing by age group, gender, category, and subscription status.

SQL → DAX → Dashboard → Decision.

9. Strategic Business Recommendations

Loyalty Expansion:
Implement tiered loyalty rewards, personalized offers, and early-access benefits for loyal customers.

Subscription Redesign:
Introduce tangible benefits such as shipping privileges, exclusive collections, and priority discounts.

Discount Control:
Gradually reduce deep discounting on high-risk SKUs and transition to bundle-based or threshold promotions.

Product Promotion Strategy:
Feature top-rated products with “Best Seller” and “Top Rated” badges.

Target Segment Growth:
Increase female customer acquisition and aggressively market to adults and middle-aged customers.

Churn Prevention:
Execute win-back campaigns for high-history inactive customers using personalized incentives.

10. Business Limitations

- Absence of timestamp limits trend and seasonal behavior analysis.
- No cost or profit data restricts margin optimization.
- Customer Lifetime Value modeling is constrained without longitudinal data.

11. Future Analytics Roadmap

- Introduce time-series transaction tracking.
- Implement RFM and CLV modeling.
- Develop churn prediction models.
- Build discount elasticity forecasting.
- Design subscription propensity scoring.

12. Predictive Modeling Extension (Churn Prediction)

To extend the behavioral insights derived from descriptive analytics, a supervised machine learning model was developed to predict customer churn risk using the same customer-level dataset. The objective of this extension was to transition from retrospective analysis to proactive identification of at-risk customers, enabling targeted retention strategies.

Churn Definition & Target Engineering
Given the cross-sectional nature of the dataset and absence of transaction timestamps, churn was defined using a multi-factor behavioral disengagement framework. Customers were classified as churn-risk if they exhibited at least two of the following conditions:

- Low purchase frequency (quarterly or annual purchasing)
- Low historical purchases (≤ first quartile threshold)
- Lack of active subscription
- Low spending level (≤ median purchase value)

This multi-signal definition provided a more realistic representation of disengagement compared to single-metric churn proxies.

Feature Engineering & Preprocessing
Predictive features included demographic attributes, product category behavior, purchase history tiers, subscription status, promotion usage, and spending level. Categorical variables were encoded using one-hot encoding, and missing review ratings were imputed using median values to preserve distributional integrity. Data leakage controls were applied by excluding direct frequency-based fields used in churn labeling.

Model Development & Evaluation
Two classification models were trained using scikit-learn:

- Logistic Regression
- Random Forest

Models were evaluated using accuracy and ROC-AUC metrics on an 80/20 train-test split.

Model Performance

- Logistic Regression :- Accuracy: 80.5%, ROC-AUC: 0.91
- Random Forest:- Accuracy: 79.0%, ROC-AUC: 0.91

The engineered behavioral churn definition improved predictive separability from approximately ROC-AUC 0.53 (frequency-only baseline) to 0.91, demonstrating the importance of multi-factor engagement signals in churn modeling.

Key Predictive Drivers
Model interpretability indicated that churn risk is primarily associated with:

- Low purchase history
- Absence of subscription
- Reduced behavioral engagement

These findings align with earlier SQL-based behavioral insights and reinforce the importance of loyalty depth and subscription participation in customer retention.

Business Implications
The predictive model enables the organization to move from descriptive churn identification to proactive churn prevention by:

- Flagging high-risk customers for targeted retention campaigns
- Prioritizing disengaged but historically valuable customers
- Supporting subscription redesign and loyalty interventions
- Enabling data-driven marketing allocation