Customer Shopping Behavior Analysis

Comprehensive Data Analysis Report

Dataset: 3,900 Customer Transactions
Analysis Period: Multiple Seasons
Report Date: November 30, 2025

Analysis Tools:
Python (Pandas, NumPy) • PostgreSQL (pgAdmin) • Power BI

1. Project Overview

This project analyzes customer shopping behavior using transactional data from 3,900 purchases across multiple product categories. **The goal is to understand:**
• Who the customers are (age, gender, subscription status, location)
• How they shop (spend amount, discounts, shipping type, repeat purchases)
• Which products and segments drive the most revenue **The final deliverables are:**
• Cleaned dataset in Python
• Business queries answered in PostgreSQL (pgAdmin)
• An interactive Customer Behavior Dashboard in Power BI

2. Dataset Summary

Rows: 3,900
Columns: 18 **Key features include:**
• **Customer demographics:** Age, Gender, Location, Subscription Status
• **Purchase details:** Item Purchased, Category, Purchase Amount (USD), Season, Size, Color
• **Shopping behavior:** Discount Applied, Promo Code Used, Previous Purchases, Payment Method, Frequency of Purchases, Review Rating, Shipping Type **Missing data:** 37 missing values in the Review Rating column.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discou Appli
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	39
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	22
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	N
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	N
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	N
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	N
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	N
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	N
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	N

Figure 1: Summary Statistics - Main Dataset Features

Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3900	3900	3900.000000	3900	3900
2	2	NaN	6	7
No	No	NaN	PayPal	Every 3 Months
2223	2223	NaN	677	584
NaN	NaN	25.351538	NaN	NaN
NaN	NaN	14.447125	NaN	NaN
NaN	NaN	1.000000	NaN	NaN
NaN	NaN	13.000000	NaN	NaN
NaN	NaN	25.000000	NaN	NaN
NaN	NaN	38.000000	NaN	NaN
NaN	NaN	50.000000	NaN	NaN

Figure 2: Summary Statistics - Additional Features

3. Exploratory Data Analysis Using Python

Using pandas in Jupyter Notebook, comprehensive data preprocessing was performed: **Data loading:** Imported the CSV into a pandas DataFrame for analysis. **Initial exploration:** Used df.info() and df.describe() to understand data types, ranges, and summary statistics. **Handling missing values:** Checked nulls with df.isnull().sum(). Filled 37 missing values in Review Rating using the median rating within each product category to avoid the impact of outliers. **Column standardization:** Renamed columns to snake_case (e.g., Customer ID → customer_id, Purchase Amount (USD) → purchase_amount). **Feature engineering:** Created an age_group column by binning ages into: Young Adult, Adult, Middle-aged, Senior. Created other helper fields as needed for analysis. **Database integration:** Loaded the cleaned DataFrame into PostgreSQL using SQLAlchemy for further SQL-based analysis.

4. Data Analysis Using SQL (Business Questions)

All queries were run in PostgreSQL (pgAdmin) on the cleaned table. The following business questions were answered through detailed SQL analysis:

Q1. What is the total revenue generated by male vs. female customers?

Result: Female: $75,191 | Male: $157,890

	gender text	revenue numeric
1	Female	75191
2	Male	157890

Figure 3: Revenue by Gender

■ **Insight:** Male customers contribute roughly 2× more revenue than female customers, suggesting male-focused campaigns could drive additional sales, but the female segment still represents a meaningful share.

Q2. Which customers used a discount but still spent more than the average purchase amount?

Result: Identified 839 such orders where customers used a discount yet spent above the global average purchase amount ($59.76).

	customer_id bigint	purchase_amount bigint
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68
7	13	72
8	16	81
9	20	90
10	22	62
11	24	88

Total rows: 839 Query complete 00:00

Figure 4: High-Value Discount Users (Sample Data - 839 total rows)

■ **Insight:** High-value customers respond to discounts but still spend more than average—targeting them with personalized offers can increase revenue without hurting margins too much.

Q3. Which are the top 5 products with the highest average review rating?

Result: Top 5 by average rating: Gloves (3.86), Sandals (3.84), Boots (3.82), Hat (3.80), Skirt (3.78)

	item_purchased text	Average Product Rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

Figure 5: Top Products by Average Review Rating

■ **Insight:** These products are good candidates for being highlighted in marketing campaigns, bundles, and recommendation sections due to strong customer satisfaction.

Q4. Compare the average purchase amounts between Standard and Express shipping.

Result: Standard shipping: $58.46 | Express shipping: $60.48

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

Figure 6: Average Purchase Amount by Shipping Type

■ **Insight:** Customers choosing Express shipping spend slightly more on average, indicating higher purchase urgency or higher-value baskets.

Q5. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers.

	subscription_status text	total_customers bigint	avg_spend numeric	total_revenue numeric
1	Yes	1053	59.49	62645.00
2	No	2847	59.87	170436.00

Figure 7: Subscription Status Comparison

■ **Insight:** Average spend is very similar between subscribed and non-subscribed customers. Total revenue from non-subscribers is much higher due to their larger count. There is still room to convert heavy non-subscriber buyers into subscribers with targeted benefits.

Q6. Which 5 products have the highest percentage of purchases with discounts applied?

Result: Top 5 discount-dependent products: Hat (50.00%), Sneakers (49.66%), Coat (49.07%), Sweater (48.17%), Pants (47.37%)

	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

Figure 8: Products with Highest Discount Rates

■ **Insight:** These products rely heavily on discounts to sell. Pricing and promotion strategy should be reviewed to improve profitability.

Q7. Segment customers into New, Returning, and Loyal based on total number of previous purchases.

Result: Loyal: 3,116 customers | New: 83 customers | Returning: 701 customers

	customer_segment text	Number of Customers bigint
1	Loyal	3116
2	New	83
3	Returning	701

Figure 9: Customer Segmentation

■ **Insight:** The majority of customers are Loyal, indicating strong retention. However, the New segment is small (only 83), suggesting that acquiring completely new customers may be an area to focus on.

Q8. What are the top 3 most purchased products within each category?

Result: Accessories: Jewelry (171), Sunglasses (161), Belt (161) | Clothing: Blouse (171), Pants (171), Shirt (169) | Footwear: Sandals (160), Shoes (150), Sneakers (145) | Outerwear: Jacket (163), Coat (161)

	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessories	Jewelry	171
2	2	Accessories	Sunglasses	161
3	3	Accessories	Belt	161
4	1	Clothing	Blouse	171
5	2	Clothing	Pants	171
6	3	Clothing	Shirt	169
7	1	Footwear	Sandals	160
8	2	Footwear	Shoes	150
9	3	Footwear	Sneakers	145
10	1	Outerwear	Jacket	163
11	2	Outerwear	Coat	161

Figure 10: Top 3 Products by Category

■ **Insight:** These products are the core catalog items and should be kept in stock, prominently featured, and considered for bundles.

Q9. Are customers who are repeat buyers (more than 5 previous purchases) also likely to subscribe?

Result: Subscription Status - No: 2,518 repeat buyers | Yes: 958 repeat buyers

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

Figure 11: Repeat Buyers by Subscription Status

■ **Insight:** Many repeat buyers still do not subscribe, which is a missed opportunity. There is a strong candidate group (2,518 repeat buyers without subscription) for subscription-focused campaigns.

Q10. What is the revenue contribution of each age group?

Result: Young Adult: $62,143 | Middle-aged: $59,197 | Adult: $55,978 | Senior: $55,763

	age_group text	total_revenue numeric
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

Figure 12: Revenue by Age Group

■ **Insight:** All age groups generate similar revenue, but Young Adults contribute slightly more. Marketing can be balanced but with a slight focus on younger users, especially for digital channels.

5. Interactive Dashboard in Power BI

An interactive Customer Behavior Dashboard was built in Power BI to present the insights visually.
The dashboard includes: KPIs:
• Number of customers (~3.9K)
• Average purchase amount ($59.76)
• Average review rating (3.75) **Visualizations:**
• % of customers by subscription status
• Revenue by category
• Sales by category
• Revenue by age group
• Sales by age group **Interactive Slicers for:**
• Subscription Status
• Gender
• Category
• Shipping Type

Figure 13: Customer Behavior Dashboard - Power BI

6. Business Recommendations

Based on the comprehensive analysis, the following key recommendations are proposed:

1. Increase Subscription Conversions

Many heavy buyers are not subscribed. Offer targeted incentives (exclusive discounts, early access, loyalty rewards) to convert them. Focus on the 2,518 repeat buyers without subscriptions as high-priority targets.

2. Optimize Discount Strategy

Products like Hat, Sneakers, Coat, Sweater, and Pants are highly discount-dependent. Test reduced discount levels or value bundles to protect margins while keeping volume. Consider dynamic pricing strategies.

3. Leverage Top-Rated and Best-Selling Products

Highlight Gloves, Sandals, Boots, Hat, and Skirt in marketing campaigns, homepages, and recommendation engines. These products have the highest customer satisfaction and should be featured prominently.

4. Focus on High-Value Segments

Young Adult and Middle-aged groups drive the most revenue. Design age-specific promotions for these segments. Develop targeted digital campaigns for younger demographics.

5. Strengthen New Customer Acquisition

The New segment is very small (83 customers). Invest in awareness/acquisition campaigns and then use the loyalty engine to move them into Returning and Loyal segments. Develop referral programs leveraging loyal customers.

6. Promote Express Shipping for Upsell

Express shipping customers spend slightly more. Encourage upgrades to Express with small fees or subscription perks. Consider free express shipping thresholds to increase basket sizes.

7. Gender-Targeted Marketing

Male customers generate 2× the revenue of female customers. While maintaining female-focused initiatives, consider expanding male-oriented product lines and marketing channels.

8. Product Bundle Strategy

Create strategic bundles using top-performing products from different categories (e.g., Jewelry + Blouse + Sandals) to increase average order value and cross-category purchases.

7. Conclusion

This comprehensive analysis of 3,900 customer transactions has revealed critical insights into shopping behavior, customer segmentation, and revenue drivers. **Key Findings:**
• A strong loyal customer base (3,116 customers) provides a solid foundation for growth
• Significant opportunity exists in converting 2,518 repeat buyers to subscribers
• Product performance varies significantly, with clear winners that should be leveraged
• All age groups contribute meaningfully to revenue, enabling diversified marketing strategies
• Discount dependency for certain products presents both a risk and an optimization opportunity **Next Steps:**
• Implement A/B testing for subscription conversion campaigns
• Review and optimize pricing strategy for discount-heavy products
• Develop targeted acquisition campaigns to grow the new customer segment
• Create personalized marketing journeys based on customer segments
• Monitor KPIs through the Power BI dashboard for continuous improvement The combination of Python data preprocessing, SQL-based business intelligence, and Power BI visualization provides a robust framework for data-driven decision making. Regular updates to this analysis will ensure strategies remain aligned with evolving customer behaviors and market conditions.