Customer Shopping Behavior Analysis Report

1. Project Overview

This project analyzes customer shopping behavior using transactional data from 3,900 purchases across multiple product categories. The objective is to extract actionable insights related to spending patterns, customer segmentation, product performance, and subscription behavior, enabling data-driven business decisions.

2. Dataset Summary

The dataset consists of 3,900 rows and 18 columns, capturing a mix of customer demographics, purchase details, and behavioral attributes.

Key data fields include:

- Customer demographics: age, gender, location, subscription status
- Purchase details: item purchased, category, purchase amount, season, size, color
- Shopping behavior: discount applied, promo code usage, previous purchases, purchase frequency, review rating, shipping type

Missing data was identified in the review_rating column (37 records).

3. Data Exploration and Preprocessing (Python)

Data exploration and cleaning were performed using Python in a Jupyter Notebook environment.

The preprocessing steps were as follows:

First, missing values were identified. The missing review ratings were handled by imputing median values, calculated separately for each product category to preserve category-level distribution.

Second, column names were standardized by removing spaces, converting all characters to lowercase, and replacing spaces with underscores to ensure consistency and compatibility across tools.

Third, a new feature called age_group was created by categorizing customers into Young Adult, Adult, Middle-Aged, and Senior groups based on age ranges. This enabled clearer demographic segmentation in later analysis.

Fourth, purchase frequency values were transformed into numerical equivalents measured in days. For example, weekly purchases were mapped to 7 days, fortnightly to 14 days, monthly to 30 days, and similar conversions for other frequencies. This allowed frequency to be treated as a quantitative variable.

4. Data Storage and SQL Analysis (PostgreSQL)

After preprocessing, the cleaned dataset was imported into PostgreSQL for structured querying and deeper analytical exploration.

Using SQL, ten critical business questions were answered to better understand customer behavior and revenue drivers:

Q1. Total revenue generated by male vs. female customers

	gender text	revenue numeric
1	Female	75191
2	Male	157890

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

	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

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

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

Q4. Compare the average Purchase Amounts between Standard and Express Shipping.

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

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

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

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

	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.00
3	Coat	49.00
4	Sweater	48.00
5	Pants	47.00

Q7. Segment customers into New, Returning, and Loyal based on their total number of previous purchases, and show the count of each segment.

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

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

	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	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

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

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

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

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

SQL query outputs and result tables were used to validate findings and support downstream visualization.

5. Data Visualization (Power BI)

The PostgreSQL database was connected directly to Power BI to create an interactive customer shopping behavior dashboard.

The dashboard highlights key performance indicators such as:

- Average review rating
- Number of customers
- Average purchase amount

It also visualizes revenue and sales distribution by category, age group, gender, and subscription status. Interactive slicers allow filtering by subscription status, gender, product category, and shipping type, enabling dynamic exploration of customer behavior.

6. Key Insights

Analysis reveals noticeable differences in spending behavior across customer segments. Subscribed customers contribute a higher share of total revenue, certain product categories consistently outperform others, and age groups exhibit distinct purchasing patterns. Discounts influence purchase behavior differently across products, and repeat buyers show stronger alignment with subscription adoption.

7. Business Recommendations

Based on the insights derived from Python analysis, SQL queries, and the Power BI dashboard, the following business recommendations are proposed:

First, focus retention efforts on subscribed and repeat customers. The analysis shows that subscribed customers contribute a higher share of total revenue and demonstrate higher average spending. Offering loyalty-based incentives, early access to products, or exclusive discounts can further strengthen retention and increase customer lifetime value.

Second, optimize discount strategies at the product level. Certain products show a high percentage of purchases with discounts applied, yet still generate strong revenue. This indicates price elasticity for those items. Discounts should be targeted toward these products rather than applied broadly, ensuring promotional spend drives incremental revenue instead of eroding margins.

Third, tailor marketing strategies by age group. Revenue contribution varies significantly across age segments, suggesting different purchasing priorities. Marketing campaigns should be personalized by age group, with younger customers targeted through promotions and digital engagement, while older segments may respond better to value-based or quality-focused messaging.

Fourth, leverage shipping preferences to improve conversion. Differences in average purchase amounts between Standard and Express shipping indicate that faster shipping correlates with higher spending. Encouraging Express Shipping through limited-time upgrades or loyalty perks could increase average order value.

Fifth, prioritize high-performing product categories. Categories that consistently generate higher revenue and sales volume should receive greater focus in inventory planning, advertising spend, and product expansion. Underperforming categories should be reviewed for pricing, placement, or potential discontinuation.

Finally, convert repeat buyers into subscribers. The relationship between frequent purchases and subscription likelihood suggests an opportunity to target repeat buyers with personalized subscription offers. Timely prompts after multiple purchases can accelerate subscription adoption and stabilize recurring revenue.

These recommendations transform analytical findings into actionable strategies that can support revenue growth, customer retention, and operational efficiency.

8. Conclusion

This project demonstrates a complete analytics workflow, from raw data exploration and cleaning in Python, to structured analysis using SQL, and finally to interactive visualization in Power BI. The insights derived can support targeted marketing strategies, product optimization, and improved customer retention initiatives. Future work could include predictive modeling to forecast customer lifetime value and churn.