Customer Purchase Behavior Analysis

Project Summary

This project focuses on understanding customer purchasing behavior by analyzing a dataset containing 3,900 retail transactions across multiple product categories. The aim is to uncover patterns in spending habits, product preferences, customer groups, and subscription tendencies. These insights will support the company in making informed and data-driven business decisions related to marketing, product strategy, and customer retention.

2. Dataset Overview

Total Rows: 3,900

Total Columns: 18

Key Data Attributes

Customer Demographics: age, gender, location, subscription status

Transaction Details: product name, category, purchase amount, season, size, color

Behavioral Indicators: discount applied, promo code usage, previous purchases, purchase frequency, review rating, shipping preference

Missing Data:

37 missing values were found in the *review_rating* column.

3. Exploratory Data Analysis (Python)

To begin the analysis, the dataset was cleaned and prepared using Python.

Data Loading

Imported the dataset using `pandas` for initial exploration.

Data Structure Review

Used `.info()` and `.describe()` to understand variable types, distributions, and summary statistics.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	...
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

Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequ... 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

4.Handling Missing Ratings

Imputed the missing review ratings using the **median rating of each product category**, ensuring category-based consistency.

5.Column Standardization

Converted all column names into **snake_case** for better readability and smoother integration with SQL and Python scripts.

6.Feature Engineering

age_group: Created age brackets to study spending patterns across age segments.

purchase_frequency_days: Calculated time gaps between customer purchases to estimate buying frequency.

7.Consistency Check

Evaluated the overlap between *discount_applied* and *promo_code_used*. Since *promo_code_used* was redundant, it was removed.

8.Database Integration

Loaded the cleaned dataset into a **PostgreSQL database** for structured SQL-based analysis.

4. SQL Analysis (Business Transactions)

A series of SQL queries were executed to answer key business questions and simulate real-world retail decision-making.

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

	gender text	revenue numeric
1	Female	75191
2	Male	157890

Revenue by Gender – Compared total revenue generated by male vs. female customers.

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

	customer_id bigint	purchase_a 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

Total rows: 839 Query complete 00:00

High-Spending Discount Users – Identified customers who used discounts but still spent above the average purchase amount.

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

	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

Top 5 Products by Rating – Found products with the highest average review ratings.

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

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

Shipping Type Comparison – Compared average purchase amounts between Standard and Express shipping.

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

Subscribers vs. Non-Subscribers – Compared average spend and total revenue across subscription status.

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	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

Discount-Dependent Products – Identified 5 products with the highest percentage of discounted purchases.

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

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

Customer Segmentation – Classified customers into New, Returning, and Loyal segments based on purchase history.

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

	item_rank bigint	category text	item_purchased text	total_or 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

Listed the most purchased products within each category.

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

	subscription_status text	repeat_buyer bigint
1	No	2518
2	Yes	958

Repeat Buyers & Subscriptions – Checked whether customers with >5 purchases are more likely to subscribe.

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	Senior	55763

5. Data Visualization in Power BI

A fully interactive dashboard was developed to bring insights to life visually.

This dashboard allows stakeholders to explore:

Revenue trends

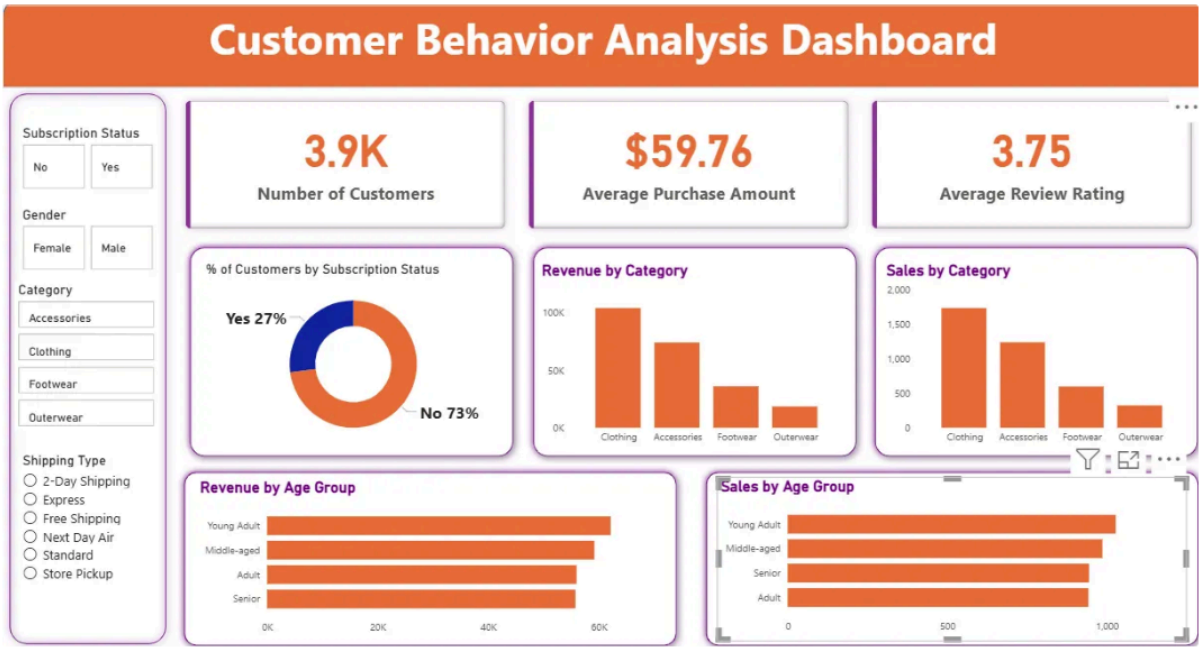
Demographic breakdowns

Product performance

Subscription behavior

Shipping patterns

Discount usage



It serves as a decision-making companion for product, marketing, and finance teams.

6. Strategic Business Insights

Based on the analysis, several actionable recommendations emerged:

1.Strengthen Subscription Offers Introduce better perks and exclusive benefits to increase subscription adoption.

2.Enhance Loyalty Initiatives Implement reward programs to encourage repeat purchases and elevate customers into the **Loyal** segment.

3.Optimize Discount Strategy Review discount-related policies to ensure higher sales do not harm profit margins.

4.Promote High-Performing Products Feature top-rated and best-selling items more prominently in marketing campaigns.

5.Target High-Value Segments Focus promotional efforts on strong revenue-generating age groups and customers who prefer express delivery.