Business Problem Statement

A retail e-commerce company wants to improve revenue growth and customer retention by understanding how customer demographics, subscription status, discounts, and purchase behavior influence spending patterns.

- The objective of this project is to analyze customer shopping behavior to:
- Identify key revenue drivers
- Evaluate the impact of subscriptions and discounts on customer spending
- Segment customers based on loyalty and purchasing frequency
- Use customer and ROI data to improve targeting, personalization, retention, and long-term value.

Key Business Questions

1. Which product categories and age groups generate the highest revenue?
2. Do subscribers spend more and purchase more frequently than non-subscribers?
3. How does discount usage affect average order value and repeat purchases?
4. Which customer segments contribute most to long-term revenue?
5. Are repeat buyers more likely to convert into subscribers?

Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior using transactional data from 3,900 purchases to understand revenue drivers, customer segments, and loyalty patterns.

Using Python for data cleaning, PostgreSQL for business analysis, and Power BI for visualization, the project delivers an end-to-end analytical workflow from raw data to actionable business insights.

The goal is to support strategic decisions related to marketing optimization, subscription growth, discount policy, and customer retention.

2. Dataset Summary

- Records: 3,900 transactions
- Features: 18 attributes

The dataset includes:

- Customer demographics: Age, Gender, Location, Subscription Status
- Purchase details: Category, Item, Purchase Amount, Season, Size, Color
- Behavioral variables: Discount Applied, Frequency of Purchases, Previous Purchases, Review Rating, Shipping Type
- Missing values were limited to the Review Rating column (37 records) and were handled using category-wise median imputation.

3. Exploratory Data Analysis using Python

The following preprocessing steps were performed using Python:

- **Data Loading:** Imported the dataset using pandas.
- **Initial Exploration:** Used df.info()to check structure and .describe() for summary statistics.

	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	Na
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	Na
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	Na
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	Na
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	Na
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	Na
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	Na

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

- **Missing Value Treatment:** Imputed missing review ratings using the median rating of each product category to preserve rating distribution.
- **Column Standardization:** Renamed columns to **snake case** for better readability and documentation.

- Feature Engineering:
  - ❖ Created age_group by binning customer ages into meaningful segments.
  - ❖ Converted purchase frequency into numerical values (days) for quantitative analysis as purchase_frequency_days.
- Data Consistency Check:
  - ❖ Identified redundancy between discount_applied and promo_code_used and removed promo_code_used to avoid duplication.
- Database Integration:
  - ❖ Loaded the cleaned dataset into PostgreSQL for structured SQL-based business analysis.

4. Data Analysis using SQL (Business Transactions)

We performed structured analysis in PostgreSQL to answer key business questions:

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

	gender text	revenue numeric
1	Female	75191
2	Male	157890

2. **High-Spending Discount Users** – Identified customers who used discounts but still spent above 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

Total rows: 839 Query complete 00:00:

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

	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

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

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

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

	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

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

	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

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

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

8. **Top 3 Products per Category** – Listed the most purchased products within each category.

	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

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

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

10. **Revenue by Age Group** – Calculated total 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

Key Insights from SQL Analysis

• **Subscriber vs Non-Subscriber Behavior**
Subscribers show higher average purchase value and repeat purchase frequency, indicating stronger long-term customer value despite lower total population.

• **Discount Impact**
Discount usage does not significantly increase average order value, suggesting discounts primarily drive volume rather than premium purchases. Certain products show high dependency on discounts, indicating margin sensitivity.

• **Customer Loyalty**
Customers with more than five previous purchases are significantly more likely to subscribe, confirming a strong relationship between loyalty and subscription adoption.

• **Revenue Concentration**
Young and middle-aged customers contribute the highest revenue share, making them priority segments for retention and personalized marketing.

• **Product Performance**
Clothing and Accessories generate the highest revenue and contain the majority of top-rated products, making them key focus categories for growth.

5. Dashboard and Visualization (Using Power BI)

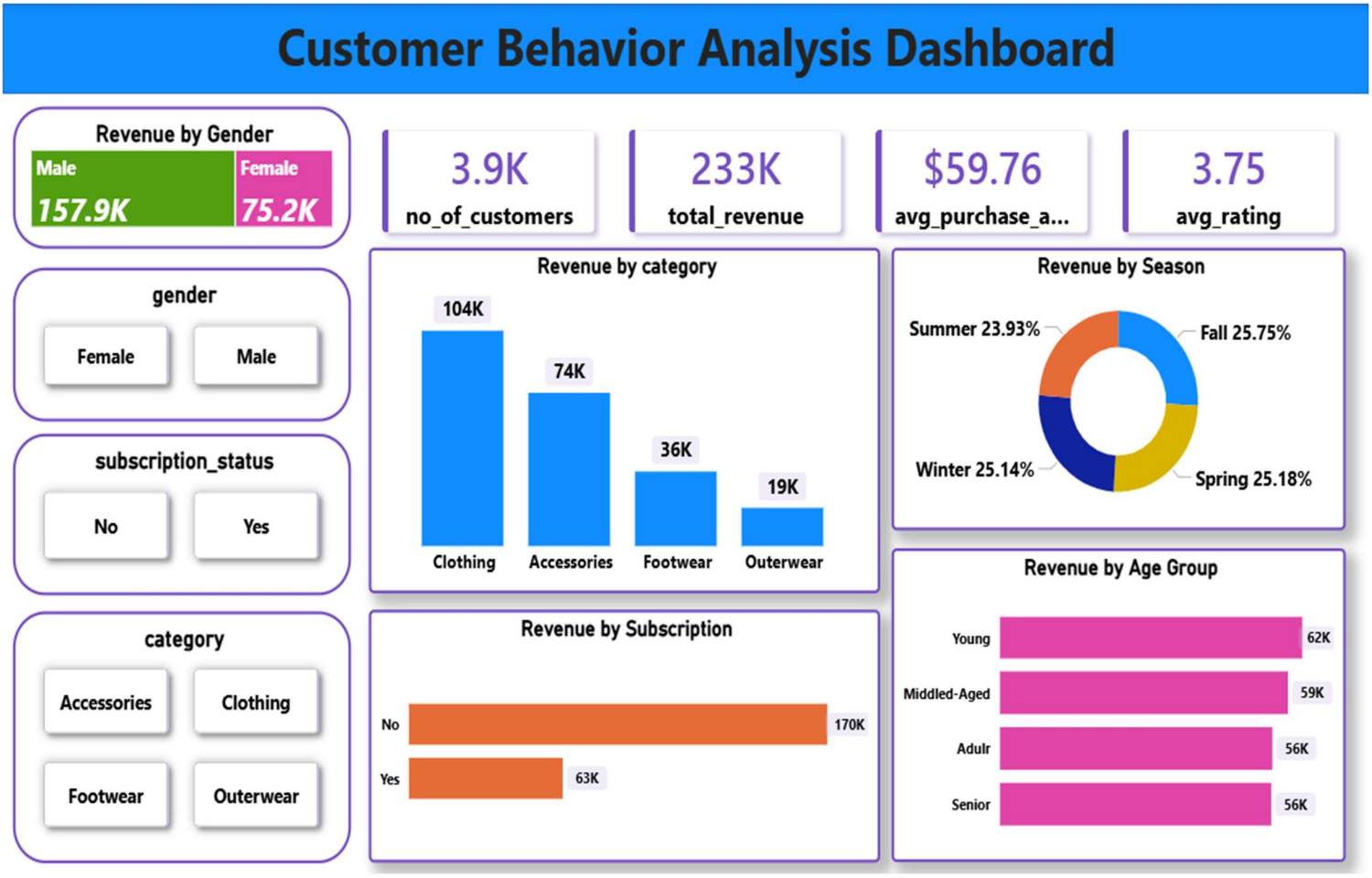
An interactive Power BI dashboard was developed to provide a consolidated view of customer behavior and revenue performance.

Key features include:

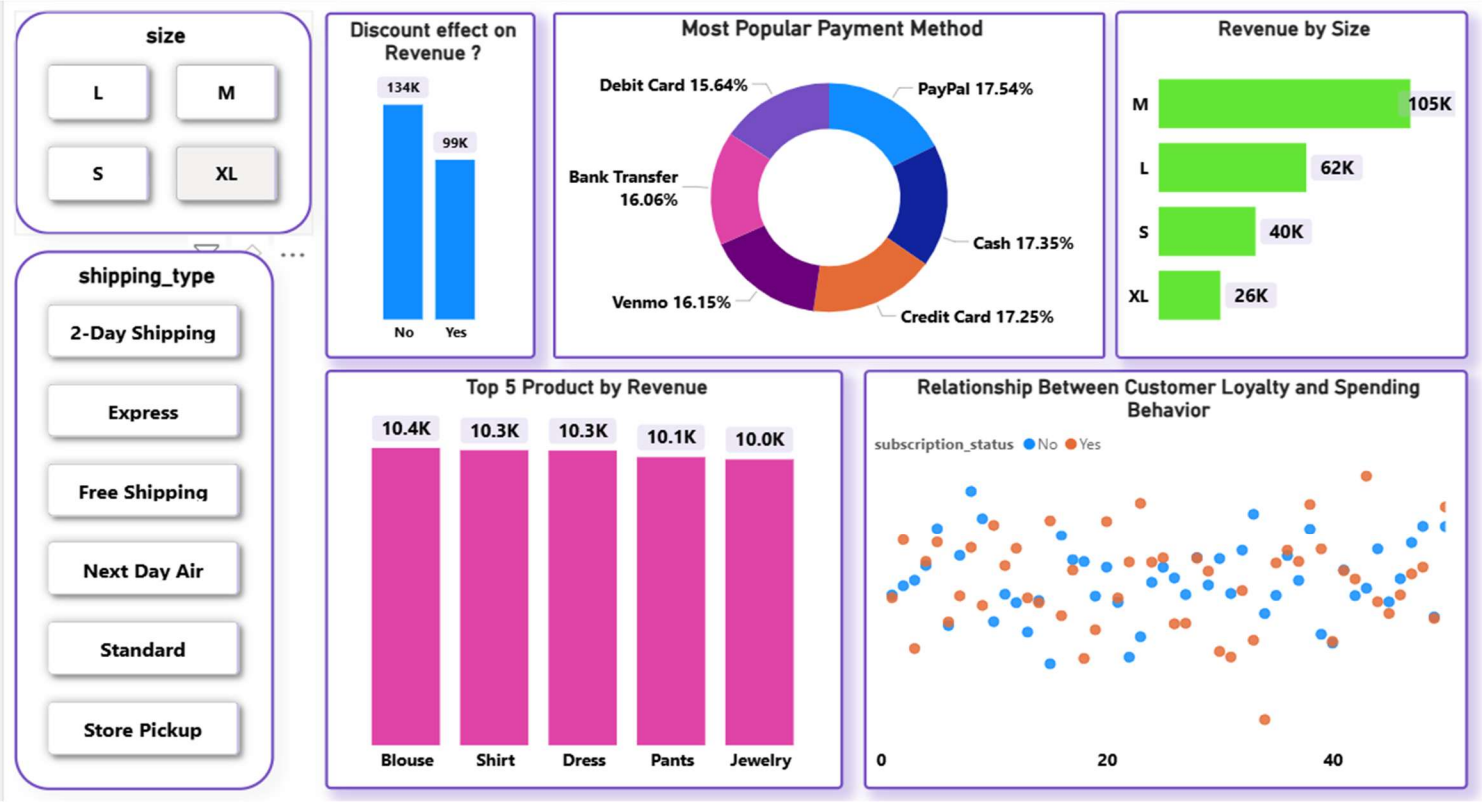
- ✓ Executive KPIs for revenue, customer base, average order value, and subscription rate.
- ✓ Revenue analysis by category, age group, season, and subscription status.
- ✓ Customer behavior analysis by purchase frequency, discount usage, and loyalty segments.
- ✓ Dynamic filters for gender, category, shipping type, and subscription status.

The dashboard enables business users to quickly identify high-value segments, evaluate discount effectiveness, and monitor subscription performance.

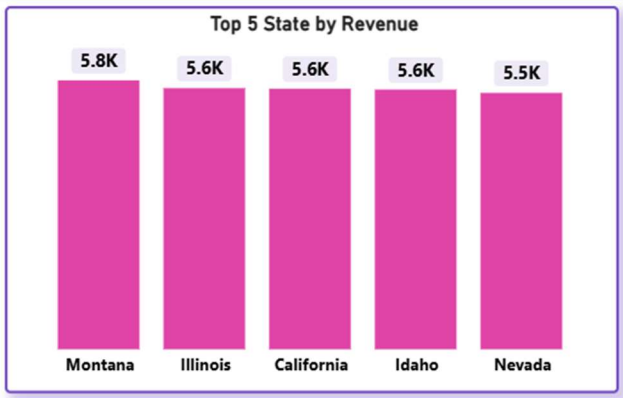
Customer Revenue & Segmentation Overview

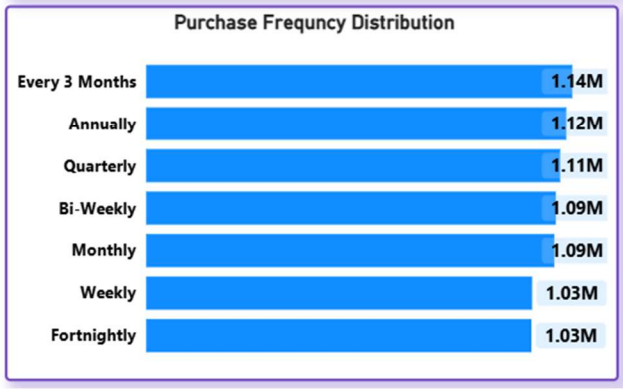


Customer Behavior & Loyalty Analysis



Geographic & Transaction Insights





Business Recommendations

1. Strengthen Subscription Acquisition
Target repeat buyers with personalized subscription offers, as loyal customers show significantly higher conversion probability and long-term lifetime value.

2. Optimize Discount Strategy
Reduce blanket discounting and focus promotions on discount-sensitive products to protect margins while sustaining sales volume.

3. Prioritize High-Value Segments
Focus retention campaigns on young and middle-aged customers who contribute the highest revenue share.

4. Improve Customer Retention
Introduce loyalty rewards, early-access benefits, and personalized recommendations to convert returning customers into long-term subscribers.

5. Product Marketing Strategy
Prioritize promotion of high-revenue and top-rated categories such as Clothing and Accessories through cross-selling and personalized campaigns.