Customer Purchase Behavior Analysis

Project Overview

This project analyzes **3,900 retail transactions** to understand customer shopping behavior and identify key factors that influence spending, product preferences, subscription adoption, and customer loyalty.

Using **Python**, **SQL (PostgreSQL)**, and **Power BI**, the project provides a complete workflow—from data cleaning and modeling to business analysis and dashboard visualization—designed to support data-driven decision-making in retail.

Dataset Summary

The dataset contains **3,900 rows** and **18 columns**, covering:

Customer Demographics

- Age
- Gender
- Location
- Subscription Status

Purchase Details

- Item Purchased
- Category
- Purchase Amount
- Season
- Size
- Color

Shopping Behavior

- Discount Applied
- Promo Code Used
- Previous Purchases
- Purchase Frequency
- Review Rating
- Shipping Type

Missing Data

- **37 missing values** in *Review Rating* column.

Data Preparation & Exploratory Analysis (Python)

All data cleaning and preprocessing were performed in Python before moving the dataset to SQL.

1. Data Loading & Initial Checks

- Loaded the dataset using pandas.
- Used **.info()** to inspect structure and **.describe()** to review summary statistics.

2. Handling Missing Values

- Identified 37 missing ratings.
- Imputed missing *Review Rating* values using the **median review rating by product category**.

3. Column Standardization

- Converted column names to **snake_case** for better readability and consistency.

4. Feature Engineering

- **age_group:** Created age segments using binned age ranges.
- **purchase_frequency_days:** Derived from purchase behavior to understand frequency patterns.

5. Data Consistency Checks

- Reviewed **discount_applied** vs **promo_code_used** and **confirmed redundancy.**
- Removed promo_code_used from the dataset.

6. Database Integration

- Connected Python to **PostgreSQL** and inserted the cleaned DataFrame into the database for SQL-based business analysis.

Business Analysis Using SQL

SQL queries were executed in PostgreSQL to answer key business questions related to revenue, customer behavior, product performance, and subscription trends.

Below is a summary of the analyses conducted:

1. Revenue by Gender

Compared total revenue generated by male and female customers.

	gender text	revenue numeric
1	Female	75191
2	Male	157890

Total rows: 2 | Query complete 00:00:00.075

2. High-Spending Discount Users

Identified customers who used discounts *and* spent above the **average purchase amount**.

	customer_id bigint	purchase_amount bigint
1	2	64
2	3	73
3	4	90
4	7	85

Total rows: 839 | Query complete 00:00:00.099

3. Top 5 Products by Rating

Found the five products with the highest **average review ratings**.

	item_purchased text	Average Product Rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.81
4	Hat	3.81
5	T-shirt	3.78

Total rows: 5 | Query complete 00:00:00.088

4. Shipping Type Comparison

Compared average purchase amounts between:

- Standard Shipping
- Express Shipping

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

Total rows: 2 | Query complete 00:00:00.085

5. Subscribers vs Non-Subscribers

Compared:

- Number of customers
- Average spend
- 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

Total rows: 2 Query complete 00:00:00.071

6. Discount-Dependent Products

Identified five 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

Total rows: 5 Query complete 00:00:00.096

7. Customer Segmentation

Classified customers as:

- New
- Returning
- Loyal

based on previous purchase counts.

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

Total rows: 3 Query complete 00:00:00.057

8. Top 3 Products per Category

Used ranking logic to list the **top 3 most purchased items** in 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

Total rows: 11 Query complete 00:00:00.079

9. Repeat Buyers & Subscriptions

Evaluated whether customers with more than 5 previous purchases were more likely to subscribe.

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

Total rows: 2 Query complete 00:00:00.076

10. Revenue by Age Group

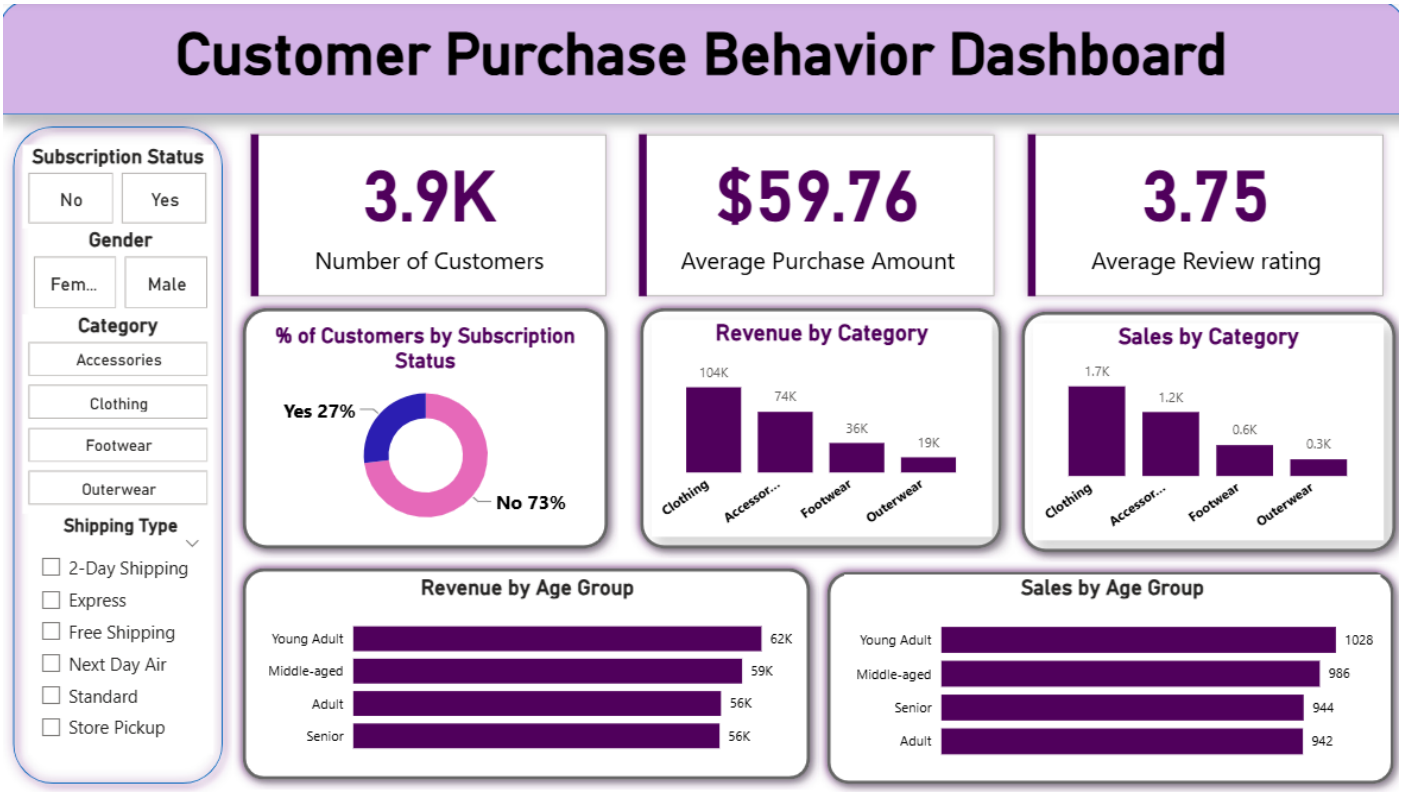
Calculated total revenue contributed by each age segment.

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

Total rows: 4 Query complete 00:00:00.110

Power BI Dashboard

An interactive Customer Behavior Dashboard was built in Power BI to visualize the key findings from SQL.



The dashboard includes:

Key Metrics

- Total Customers: **3.9K**
- Average Purchase Amount: **$59.76**
- Average Review Rating: **3.75**

Visual Insights

- % of Customers by Subscription Status
- Revenue by Category
- Sales by Category
- Revenue by Age Group
- Sales by Age Group

Interactive Filters

- Subscription Status
- Gender
- Category
- Shipping Type

This allows stakeholders to explore customer behavior dynamically and uncover hidden patterns.

Business Recommendations

Based on the Python, SQL, and dashboard findings, the following recommendations were derived:

1. Boost Subscriptions

Offer exclusive incentives to increase subscription adoption, as subscribers show stronger spending patterns.

2. Strengthen Customer Loyalty Programs

Reward returning and repeat buyers to encourage movement into the *Loyal* segment.

3. Review Discount Strategy

since discount usage is high for certain products, evaluate discount margins to balance revenue and profitability.

4. Improve Product Positioning

Highlight high-rated and best-selling products in marketing campaigns.

5. Target High-Value Segments

Focus marketing efforts on:

- Young Adults
- Middle-aged customers
- Express shipping users

who contribute significantly to revenue.