Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior using transactional retail data to uncover patterns in spending, customer segments, product performance, and subscription behavior.
The objective is to derive **actionable business insights** that can support decisions related to marketing strategy, customer retention, and product positioning.

The analysis combines **Python (EDA & data cleaning)**, **SQL (business queries)**, and **Power BI (dashboard visualization)** to simulate an end-to-end analytics workflow.

2. Dataset Summary

- **Total Records:** 3,900 transactions
- **Total Columns:** 18

Key Attributes:

- **Customer Demographics:** Age, Gender, Location, Subscription Status
- **Purchase Details:** Item Purchased, Category, Purchase Amount, Season, Size, Color
- **Behavioral Data:** Discount Applied, Promo Code Used, Previous Purchases, Purchase Frequency, Review Rating, Shipping Type

Data Quality Notes:

- Missing values were observed in the **Review Rating** column.
- Certain columns showed overlapping information and required validation.

3. Exploratory Data Analysis & Data Cleaning (Python)

Data preparation was performed using **Python (pandas)** to ensure consistency and analytical readiness.

Steps Performed:

- **Data Loading:** Imported dataset using pandas.

	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

- **Initial Exploration:** Used df.info() and df.describe() to understand structure and distributions.
- **Missing Value Handling:**
    - Missing values in review_rating were imputed using the **median rating per product category**.

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

- **Column Standardization:**
    - Renamed columns to **snake_case** for readability and SQL compatibility.
- **Feature Engineering:**
    - Created age_group using quantile-based binning.
    - Derived purchase-related features for segmentation.
- **Redundancy Check:**
    - Identified overlap between discount_applied and promo_code_used.
    - Dropped promo_code_used to avoid duplication.
- **Database Integration:**
    - Loaded the cleaned dataset into **PostgreSQL** for further SQL-based analysis.

4. Business Analysis using SQL

Using PostgreSQL, multiple business-driven queries were executed to answer key analytical questions:

1. **Revenue by Gender** – Compared total revenue contribution across genders.

2. **High-Spending Discount Users** – Identified customers who used discounts but still spent above the average purchase value.
3. **Top 5 Products by Average Rating** – Evaluated product quality perception using customer ratings.
4. **Shipping Type Comparison** – Compared average purchase amounts between Standard and Express shipping.
5. **Subscription Analysis** – Compared average spend and total revenue between subscribed and non-subscribed customers.
6. **Discount-Dependent Products** – Identified products most frequently purchased with discounts.
7. **Customer Segmentation** – Classified customers into New, Returning, and Loyal groups based on purchase history.
8. **Top Products per Category** – Ranked top 3 most purchased products within each category using window functions.
9. **Repeat Buyers vs Subscription** – Analyzed whether frequent buyers are more likely to subscribe.
10. **Revenue by Age Group** – Evaluated revenue contribution across age segments.

These queries formed the foundation for dashboard visuals and insights.

4.1: SQL Queries & Outputs

The following SQL outputs were selected to highlight the most business-relevant findings that were later visualized in the dashboard.

- **Key Insight : Subscription Impact**
  Subscription status shows minimal impact on average purchase value, indicating subscriptions may support retention rather than higher transaction size.

	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

- **Key Insight : Top Products per Category**

Within each category, a small number of products account for the majority of purchases,indicating clear customer preferences and opportunities for focused product promotion.

	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

- **Key Insight : Age-based Revenue Concentration**
  Customers aged 26–35 contribute the highest revenue, making them a priority segment for focused marketing strategies.

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

5. Power BI Dashboard

An interactive Power BI dashboard was created to visually communicate insights.

Dashboard Highlights:

- Revenue distribution across customer segments
- Subscription and purchase behavior comparison
- Product and category performance
- Age group contribution analysis
- Interactive slicers for better exploration

The dashboard focuses on **clarity, minimal visuals, and business relevance** rather than excessive charts.

6. Business Recommendations

- **Optimize Subscription Strategy:** Focus on long-term engagement benefits rather than short-term revenue lift.
- **Strengthen Loyalty Programs:** Incentivize repeat buyers to transition into loyal customers.
- **Product Promotion:** Highlight top-rated and frequently purchased products in campaigns.
- **Targeted Marketing:** Allocate marketing efforts based on age-group revenue contribution.

7. Tools & Technologies Used

- **Python:** pandas, numpy
- **SQL:** PostgreSQL
- **Visualization:** Power BI
- **Version Control:** Git & GitHub