Customer Shopping Behavior Analysis

1. Project Overview

This project focuses on exploring customer shopping behavior using transactional data from 3,900 purchases across different product categories. The aim is to identify key insights into how customers spend, what products they prefer, and how their purchasing habits vary across segments. By analyzing these patterns — including spending behavior and subscription trends — the project provides data-driven recommendations to support smarter marketing, sales, and business strategy decisions.

2. Dataset Summary

The dataset consists of **3,900 records** and **18 columns**, capturing detailed information about customer demographics, purchase details, and shopping behavior.

Key Features include:

- **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

3. Exploratory Data Analysis using Python

The first step involved preparing and cleaning the dataset in Python.

- **Data Loading:** The dataset was imported using the **pandas** library for easy manipulation and analysis.
- **Initial Exploration:** The **df.info()** function was used to understand the dataset's structure, data types, and missing values, while **df.describe()** provided summary statistics to get an overview of key numerical features.

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.749949	25.351538
std	1125.977353	15.207589	23.685392	0.716223	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.700000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

Data Preparation and Feature Engineering

- **Column Standardization:** Renamed all column names to *snake_case* format to improve readability and maintain consistent documentation.
- **Feature Engineering:**
  - Created an **age_group** column by categorizing customer ages into defined ranges.
  - Generated a **purchase_frequency_days** column to measure the average time gap between customer purchases.
- **Data Consistency Check:** Checked for redundancy between **discount_applied** and **promo_code_used** columns and removed **promo_code_used** after confirming overlap.
- **Database Integration:** Connected the Python environment to an **SQLite** database within **Jupyter Notebook** and loaded the cleaned DataFrame for performing SQL-based analysis.

4. Data Analysis using SQL

- We conducted a structured analysis in **SQLite** to address key business questions and uncover insights from transactional data.

1. **Revenue by Gender:** Analyzed and compared the total Revenue generated by **Male** and **Female** customers to understand gender-based spending patterns.

	Gender	Revenue
0	Female	75191
1	Male	157890

2. **High-Spending Discount Users:** Identified customers who applied discounts yet had spending amounts higher than the overall average purchase value.

	customer_id	purchase_amount
0	2	64
1	3	73
2	4	90
3	7	85
4	9	97
...	...	...
834	1667	64
835	1671	73
836	1673	73
837	1674	62
838	1676	90

339 rows × 2 columns

3. **Top 5 Products by Rating:** Identified the five products with the highest average customer review ratings.

	item_purchased	AVG_PRODUCT_RATING
0	Gloves	3.86
1	Sandals	3.84
2	Hat	3.81
3	Boots	3.81
4	T-shirt	3.78

4. **Shipping Type Comparison:** Compared the average purchase amounts for orders using **Standard** versus **Express** shipping.

	shipping_type	purchase_amount
0	Express	60.48
1	Standard	58.46

5. **Subscribers vs. Non-Subscribers:** Compared average spending and total revenue between subscribed and non-subscribed customers.

	SUBSCRIPTION_STATUS	TOTAL_CUSTOMER	AVERAGE_SPEND	TOTAL_REVENUE
0	No	2847	59.865121	170436
1	Yes	1053	59.491928	62645

6. **Discount-Dependent Products:** Identified the five products with the highest proportion of purchases made using discounts.

	item_purchased	DISCOUNT_RATE
0	Hat	50.00
1	Sneakers	49.66
2	Coat	49.07
3	Sweater	48.17
4	Pants	47.37

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

	CUSTOMER_SEGMENTS	Number_of_Customer
0	Loyal	3116
1	New	83
2	Returning	701

8. **Top 3 Products per Category:** Listed the three most frequently purchased products within each product category.

	item_rank	category	item_purchased	total_orders
0	1	Accessories	Jewelry	171
1	2	Accessories	Sunglasses	161
2	3	Accessories	Belt	161
3	1	Clothing	Pants	171
4	2	Clothing	Blouse	171
5	3	Clothing	Shirt	169
6	1	Footwear	Sandals	160
7	2	Footwear	Shoes	150
8	3	Footwear	Sneakers	145
9	1	Outerwear	Jacket	163
10	2	Outerwear	Coat	161

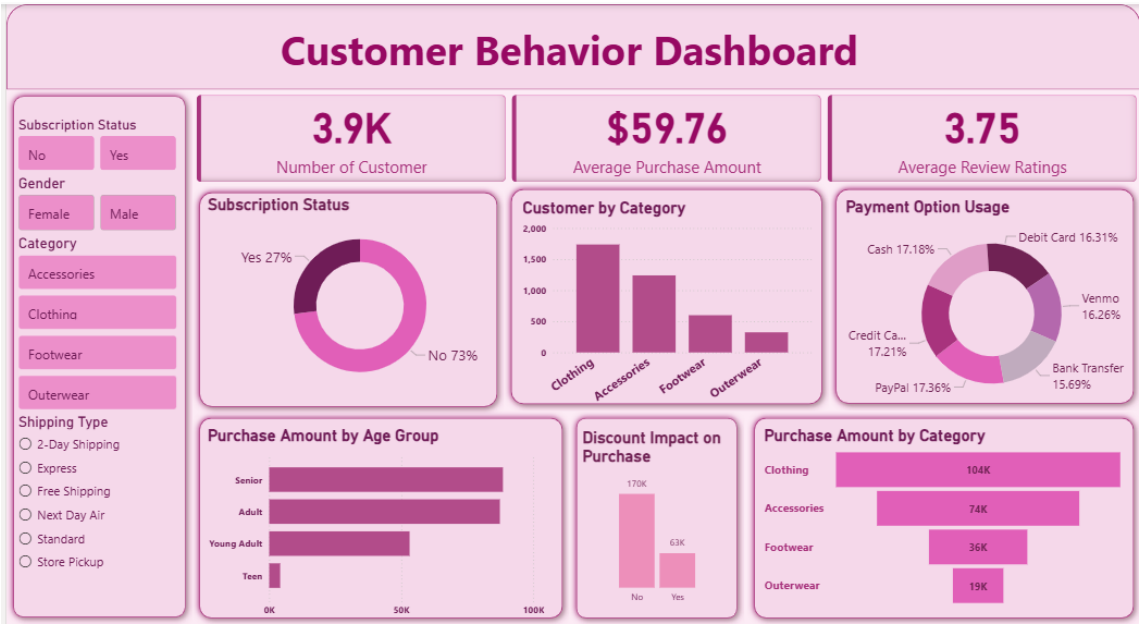
9. **Repeat Buyers & Subscriptions:** Analyzed whether customers with more than five purchases are more likely to subscribe.

	COUNT(customer_id)	subscription_status
0	2583	No
1	980	Yes

10. **Revenue by Age Group:** Calculated the total revenue contributed by each age group to understand spending patterns across demographics.

	age_group	REVENUE
0	Adult	55978
1	Middle-aged	59197
2	Senior	55763
3	Young Adult	62143

5. Dashboard in Power BI



6. Business Recommendations

- **Boost Subscriptions:** Promote exclusive benefits to encourage more customers to subscribe.
- **Customer Loyalty Programs:** Reward repeat buyers to help them transition into the **"Loyal"** customer segment.
- **Review Discount Policy:** Adjust discount strategies to drive sales while maintaining healthy profit margins.
- **Product Positioning:** Feature top-rated and best-selling products prominently in marketing campaigns.
- **Targeted Marketing:** Focus promotional efforts on high-revenue age groups and customers who use **Express** shipping.