Customer Shopping Behavior Analysis

1. Project Overview

This project focuses on understanding how customers shop by examining real purchase data collected from **3,900 transactions** across multiple product categories. The main objective is to identify meaningful patterns in customer spending, purchasing habits, product choices, and subscription usage. These insights can help businesses make better decisions related to marketing, customer engagement, and overall business strategy.

2. Dataset Summary

The dataset consists of **3,900 records** and **18 different variables**, offering a detailed view of customer behavior.

Key information included in the dataset:

- **Customer details:** age, gender, location, and whether the customer has an active subscription
- **Purchase information:** items bought, product category, amount spent, season of purchase, size, and color preferences
- **Shopping behavior insights:** use of discounts or promo codes, number of previous purchases, purchase frequency, customer review ratings, and preferred shipping methods

During data review, a small amount of missing information was identified. Specifically, **37 entries** in the Review Rating column were incomplete and required attention during the data preparation stage.

3. Exploratory Data Analysis Using Python

The analysis began in Python, where the dataset was prepared and cleaned to ensure accuracy and consistency before deeper analysis.

	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

First, the data was loaded using the **pandas** library. An initial review was carried out to understand the structure of the dataset, including data types and basic statistical summaries. This helped identify potential issues such as missing values and inconsistencies.

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 values were mainly found in the *Review Rating* column. To address this, the missing ratings were filled using the **median rating of the corresponding product category**, ensuring that the data remained balanced and unbiased.

To improve readability and maintain good documentation practices, all column names were standardized into **snake_case** format. Additional features were also created to enhance analysis. Customer ages were grouped into meaningful **age categories**, and purchase frequency was converted into a **days-based metric** for better behavioral insights.

A consistency check was performed on discount-related fields. Since *discount_applied* and *promo_code_used* conveyed similar information, the *promo_code_used* column was removed to avoid redundancy.

Once the data was cleaned and enriched, the final dataset was connected to a **SQL Server database** directly from Python, enabling structured and efficient SQL-based analysis.

4. Data Analysis Using SQL (Business Transactions)

1. **What is the total revenue generated by male and female customers?**
   → Used to compare spending behavior across genders.

	gender	revenue
1	Male	157890
2	Female	75191

2. **Which customers applied discounts but still spent more than the overall average purchase**
   Helps identify high-value customers who are not price-sensitive.

	customer_id	purchase_amount
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

3. **Which are the top 5 products with the highest average customer review ratings?**
Identifies best-performing products based on customer satisfaction.

	item_purchased	Average_Product_Rating
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.8
5	Handbag	3.78
6	Skirt	3.78
7	T-shirt	3.78
8	Sweater	3.76
9	Sneakers	3.76
10	Belt	3.76

4. **How does the average purchase amount differ between Standard and Express shipping methods?**
Analyzes the impact of shipping preference on spending behavior.

	shipping_type	(No column name)
1	Standard	58
2	Express	60

5. **Do subscribed customers spend more than non-subscribed customers?**
→ Compares average spend and total revenue by subscription status.

	subscription_status	total_customers	avg_spend	total_revenue
1	Yes	1053	59	62645
2	No	2847	59	170436

6. **Which five products have the highest percentage of purchases made with discounts?**

Identifies products that are most dependent on discount strategies.

	item_purchased	discount_rate
1	Hat	50.000000000000
2	Sneakers	49.660000000000
3	Coat	49.070000000000
4	Sweater	48.170000000000
5	Pants	47.370000000000

7. **How can customers be segmented into New, Returning, and Loyal groups based on purchase history?**

Supports customer lifecycle and retention analysis.

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

8. **What are the top three most purchased products within each product category?**

→ Helps understand category-wise product demand.

	item_rank	category	item_purchased	total_orders
1	1	Accessories	Jewelry	171
2	2	Accessories	Belt	161
3	3	Accessories	Sunglasses	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. **Are repeat buyers (customers with more than five previous purchases) more likely to subscribe?**

Evaluates the relationship between customer loyalty and subscriptions.

	subscription_status	repeat_buyers
1	Yes	958
2	No	2518

10. **Which age groups contribute the highest total revenue?**

Identifies high-value demographic segments.

	age	total_revenue
1	49	5552
2	69	5484
3	25	5372
4	41	5282
5	54	5282
6	57	5200
7	28	5104
8	19	4941
9	50	4930
10	31	4864

11. **What is the total revenue and average purchase amount for each product category?**

Assesses category-level performance.

	category	total_revenue	avg_purchase
1	Clothing	104264	60
2	Accessories	74200	59
3	Footwear	36093	60
4	Outerwear	18524	57

12. **Which customers fall into the top 10% of spenders based on total purchase value?**

Identifies premium and high-value customers.

	customer_id	total_spent
1	1	53
2	2	64
3	3	73
4	4	90
5	5	49
6	6	20
7	7	85
8	8	34
9	9	97
10	10	31

13. **What is the average purchase amount segmented by gender and subscription status?**

Analyzes combined demographic and behavioral trends.

	gender	subscription_status	avg_spend
1	Female	No	60
2	Male	No	59
3	Male	Yes	59

14. **Which products generate the highest revenue per order?**

Highlights products with high per-transaction value.

	item_purchased	revenue_per_order
1	Boots	62
2	Dress	62
3	T-shirt	62
4	Shirt	61
5	Shoes	61
6	Shorts	60
7	Jeans	60
8	Gloves	60
9	Hat	60
10	Backpack	60

15. **How do customers rank based on total spending across all purchases?**

Used for customer prioritization and targeting.

	customer_id	total_spent	spending_rank
1	43	100	1
2	96	100	1
3	194	100	1
4	205	100	1
5	244	100	1
6	249	100	1
7	456	100	1
8	519	100	1
9	582	100	1
10	616	100	1

16. **Do customers with higher numbers of previous purchases spend more on average?**

Examines the link between purchase history and spending behavior.

	previous_purchases	avg_spend
1	1	58
2	2	60
3	3	58
4	4	61
5	5	64
6	6	55
7	7	61
8	8	66
9	9	63
10	10	58

17. **What percentage of total revenue comes from discounted versus non-discounted purchases?**

Measures the financial impact of discounts.

	discount_applied	revenue_percentage
1	Yes	42.650000000000000
2	No	57.350000000000000

18. **Which shipping type generates the highest total revenue overall?**

Supports logistics and shipping strategy decisions.

	shipping_type	total_revenue
1	Free Shipping	40777
2	Express	39067
3	Store Pickup	38931
4	Standard	38233
5	2-Day Shipping	38080
6	Next Day Air	37993

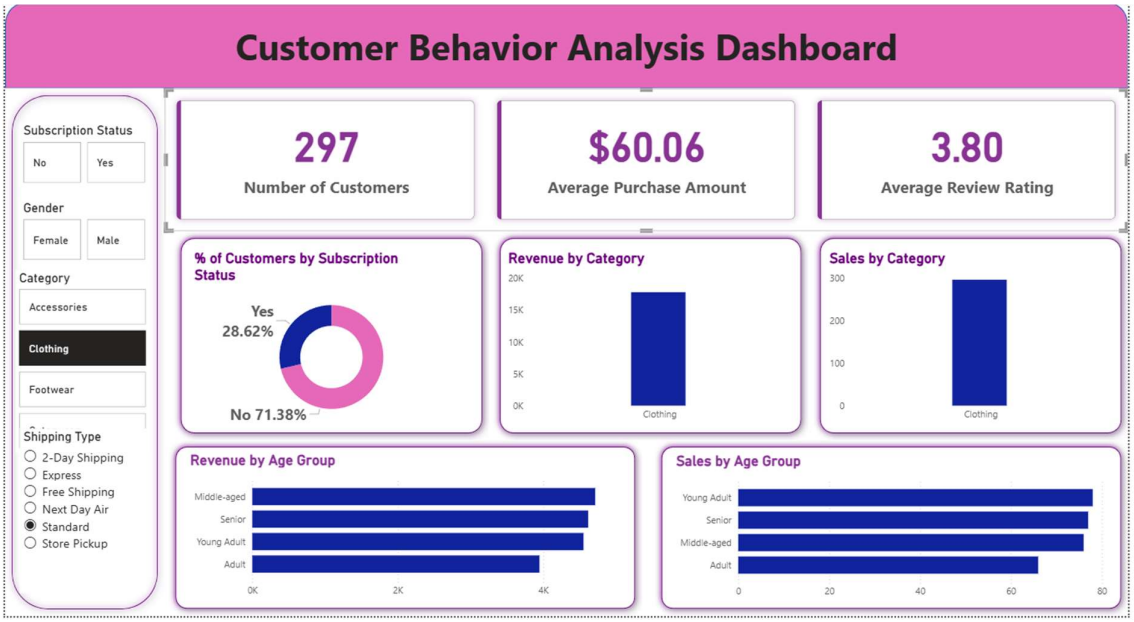
19. **Which age group has the highest average purchase value per transaction?**

Identifies age groups with stronger per-order spending power.

	age	avg_purchase
1	53	67
2	49	66
3	21	64
4	44	64
5	61	64
6	28	64
7	54	63
8	51	63
9	65	63
10	25	63

5. Dashboard Creation in Power BI

To present the findings in a clear and visual manner, an **interactive dashboard** was built using Power BI. The dashboard allows users to explore key metrics such as revenue, customer segments, product performance, and shopping behavior through dynamic filters and visuals, making insights easy to understand for both technical and non-technical stakeholders.



6. Business Recommendations

Based on the insights generated from the analysis, several practical recommendations were made:

- **Strengthen Subscription Programs:**
  Encourage more customers to subscribe by offering meaningful benefits such as exclusive discounts, early access to new products, and special rewards for members. This helps increase repeat purchases and long-term customer value.
- **Build Strong Customer Loyalty:**
  Introduce loyalty and reward programs that recognize repeat buyers and motivate them to continue shopping. By rewarding frequent customers, the business can gradually move them into a long-term loyal customer segment.
- **Use Discounts More Strategically:**
  Apply discounts carefully to boost sales without hurting profit margins. Instead of offering broad discounts, focus on targeted promotions that attract customers while maintaining overall profitability.

- **Promote High-Performing Products:**
  Highlight best-selling and highly rated products in marketing campaigns to build customer trust and drive higher conversions. These products can act as key drivers for revenue growth.
- **Focus on High-Value Customer Segments:**
  Direct marketing efforts toward age groups and customers who contribute the most revenue, particularly those who prefer express shipping. This ensures better returns on marketing spend and improves overall efficiency.