Project Overview

Understanding customer behaviours is central to designing effective business strategies. This project explores how customers interact with products, services, and brands, focusing on the drivers of decision-making, loyalty, and satisfaction. The insights will guide marketing, product development, and customer experience improvements. This project analyses customer shopping behavior using transactional data from 3,900 purchases across various product categories.

Dataset Summary

- Rows: 3,900

- Columns: 18

- Key Features:

- 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, Frequency of Purchases, Review Rating, Shipping Type.

- Missing Data: 37 values in Review Rating column.

Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

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

* ***Initial Exploration:*** Used df.info() to check structure and df.describe() for summary statistics.

```
df.head()
```

Python

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Customer ID              3900 non-null   int64
 1   Age                      3900 non-null   int64
 2   Gender                   3900 non-null   object
 3   Item Purchased           3900 non-null   object
 4   Category                 3900 non-null   object
 5   Purchase Amount (USD)    3900 non-null   int64
 6   Location                 3900 non-null   object
 7   Size                     3900 non-null   object
 8   Color                    3900 non-null   object
 9   Season                   3900 non-null   object
 10  Review Rating            3863 non-null   float64
 11  Subscription Status      3900 non-null   object
 12  Shipping Type            3900 non-null   object
 13  Discount Applied         3900 non-null   object
 14  Promo Code Used          3900 non-null   object
 15  Previous Purchases       3900 non-null   int64
 16  Payment Method           3900 non-null   object
 17  Frequency of Purchases   3900 non-null   object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

*_Missing Data Handling:_ Checked for null values and imputed missing values in the Review Rating column using the median rating of each product category.

*_Column Standardization:_ Renamed columns to **snake case** for better readability and documentation.

*Feature Engineering: -Created **age_group** column by binning customer ages.

-Created **purchase_frequency_days** column from purchase data.

*_Data Consistency Check:_ Verified if discount_applied and promo_code_used were redundant. Dropped promo_code_used column as it contained the same info as discount_applied.

*_Database Integration:_ Connected Python script to PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

Data Analysis using SQL (Business Transactions)

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

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

	gender text	revenue numeric
1	Female	75191
2	Male	157890

b. **High-Spending Discount Users** – Identified customers who used discounts but still spent above the average purchase amount.

	customer_id bigint	purchase_amounts 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
11	24	88
12	29	94
13	32	79

c. **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

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

	shipping_type text	average_purchase numeric
1	Standard	58.46
2	Express	60.48

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

f.

	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

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

	products text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.00
3	Coat	49.00
4	Sweater	48.00
5	Pants	47.00

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

i.

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

j. **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	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

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

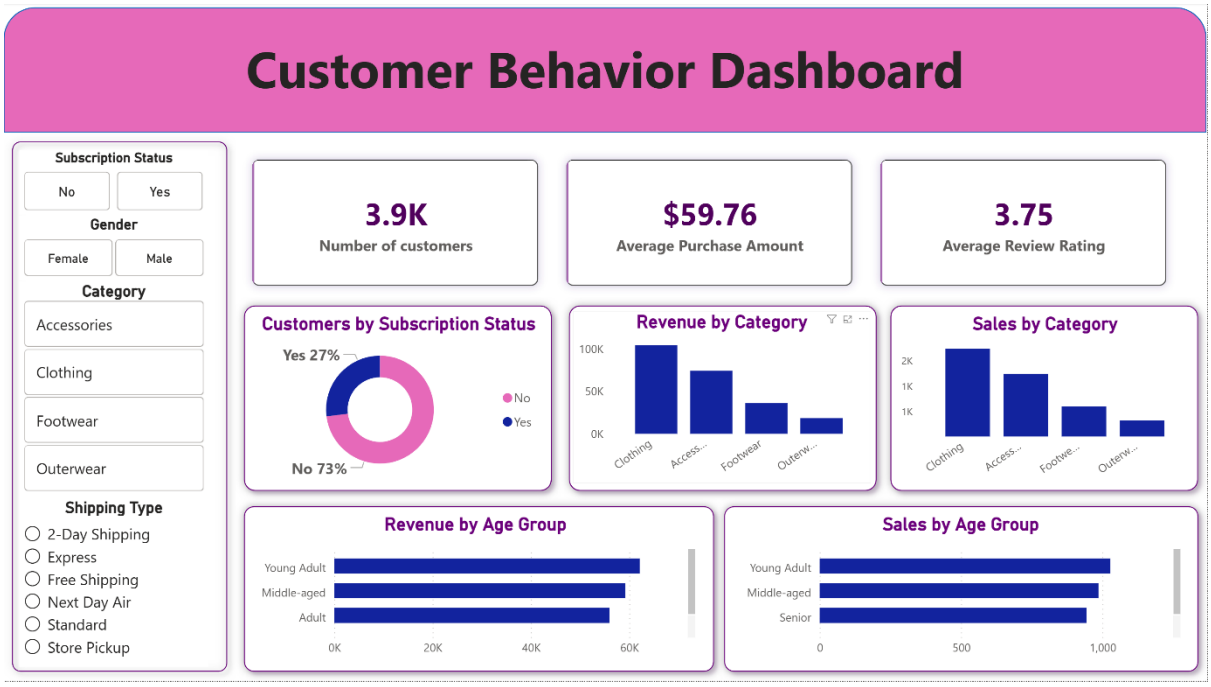
	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

l. **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

Dashboard in Power BI

An interactive dashboard in **Power BI** to present insights visually.



Business Recommendations

- ✓ Boost Subscriptions – Promote exclusive benefits for subscribers.
- ✓ Customer Loyalty Programs – Reward repeat buyers to move them into the "Loyal" segment.
- ✓ Review Discount Policy – Balance sales boosts with margin control.
- ✓ Product Positioning – Highlight top-rated and best-selling products in campaigns.
- ✓ Targeted Marketing – Focus efforts on high-revenue age groups and express-shipping users.