Customer Shopping Behavior Analysis

1. Project Overview

This project focuses on analysing **customer shopping behaviour** using transactional and demographic data to uncover actionable business insights. The analysis integrates **Python (EDA & data preparation)**, **SQL (business-driven querying)**, and **Power BI (interactive visualization)** to deliver a complete analytics workflow.

The dataset contains approximately **3,900 customer transactions**, covering multiple product categories, age groups, subscription statuses, and shipping preferences. The final output is an interactive **Customer Behaviour Dashboard** that supports strategic decision-making in marketing, product planning, and customer retention.

2. Dataset Summary

Source: Customer Shopping Behaviour dataset (CSV)

Size & Structure

- Rows: ~3,900
- Columns: 18

Key Attributes

- **Customer Demographics:** Age, Gender, Location, Subscription Status
- **Purchase Details:** Category, Item Purchased, Purchase Amount, Season, Size, Colour
- **Behavioural Metrics:** Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases
- **Experience Metrics:** Review Rating
- **Logistics:** Shipping Type

Data Quality Notes

- Missing values identified in the **Review Rating** column
- Categorical values standardized for consistency across tools

3. Exploratory Data Analysis using Python

Python (Jupyter Notebook) was used as the first analytical layer to clean, explore, and validate the dataset before loading it into the database and Power BI.

Key Steps Performed

- **Data Loading:** Imported CSV data using Pandas
- **Initial Exploration:** Used `info ()` and `describe ()` to understand structure and distributions

	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

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 Data Handling:** Checked for null values and imputed missing values in the Review Ratingcolumn 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_appliedand promo_code_used were redundant; dropped promo_code_used.

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

Outcome
This step ensured the dataset was clean, consistent, and analytically reliable before moving to SQL-based business analysis.

4. Data Analysis using SQL (Business Transactions)

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

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

	gender text	revenue numeric
1	Female	75191
2	Male	157890

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

	customer_id bigint	purchase_amount 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	[illegible]

Total rows: 839 Query complete 00:00:(

3. **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

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

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. **Subscribers vs. Non-Subscribers** – Compared average spend and 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

6. **Discount-Dependent Products** – Identified 5 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

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

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

8. **Top 3 Products per Category** – Listed the most purchased products within each

	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

category.

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

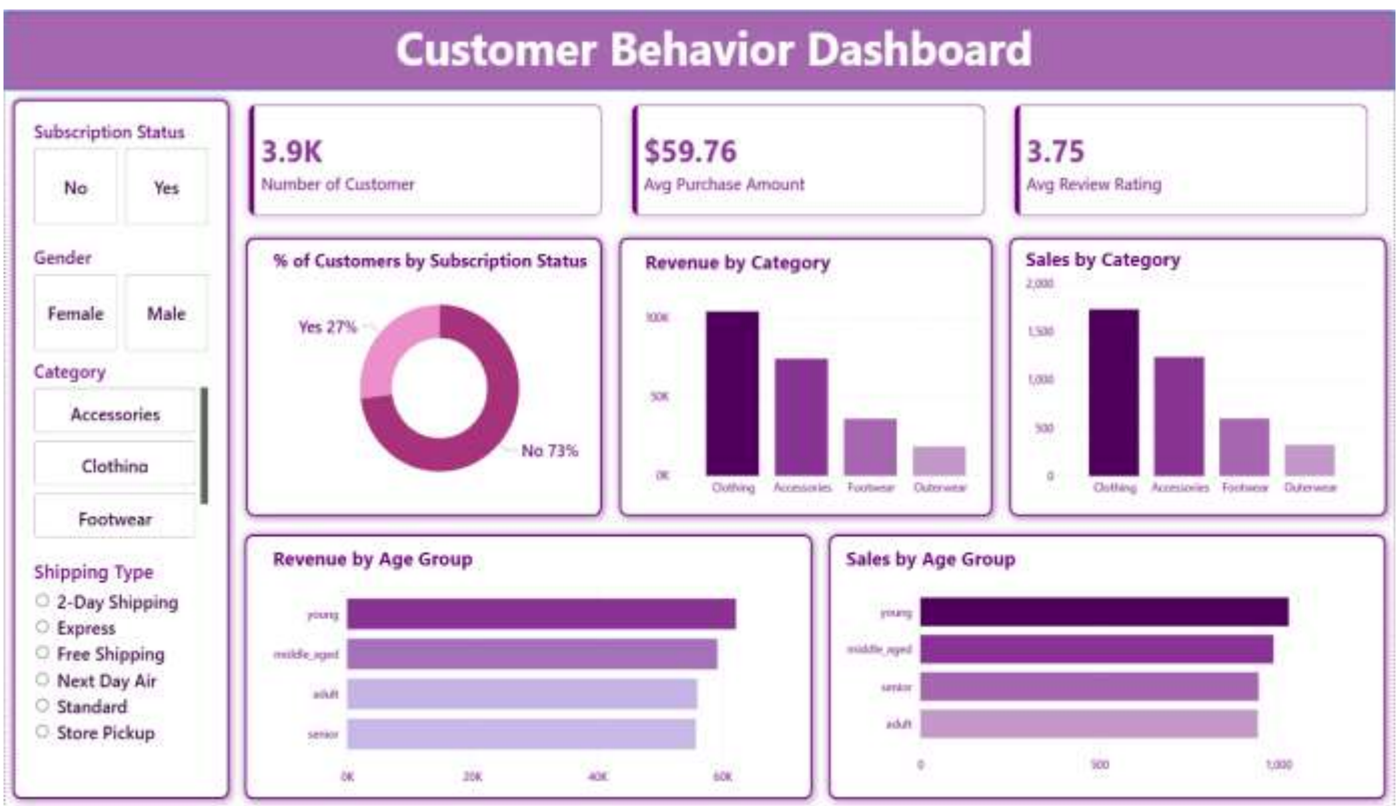
	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

10. **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

5. Dashboard in Power BI

Finally, we built an interactive dashboard in **Power BI** to present insights visually.



An interactive **Customer Behaviour Dashboard** was developed in Power BI using the processed dataset.

Key KPIs

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

Visual Components

- Subscription Status Distribution (Donut Chart)
- Revenue by Category (Bar Chart)
- Sales by Category (Bar Chart)
- Revenue by Age Group (Horizontal Bar Chart)
- Sales by Age Group (Horizontal Bar Chart)

Interactive Filters

- Gender
- Subscription Status
- Product Category
- Shipping Type

Tools & Features Used

- Power Query for data transformation
- DAX measures for KPIs and aggregations
- Slicers for dynamic filtering
- KPI cards and comparative visuals

6. Key Insights

- **Clothing** is the highest revenue-generating category
- **Young and middle-aged customers** contribute the most to revenue and sales
- Only **27% of customers are subscribers**, indicating strong growth potential
- Subscribers show higher average purchase value compared to non-subscribers
- Express and faster shipping options are associated with higher spending

7. Business Recommendations

- **Increase Subscription Adoption:** Offer exclusive discounts, early access, or free shipping for subscribers
- **Strengthen Loyalty Programs:** Target returning customers to convert them into loyal segments
- **Optimize Discount Strategy:** Balance promotional offers to protect profit margins
- **Product Strategy:** Promote top-rated and best-selling products in campaigns
- **Targeted Marketing:** Focus on high-value age groups and high-spend shipping preferences