CUSTOMER SHOPPING BEHAVIOR ANALYSIS

Project Overview

This project analyzes customer shopping behavior using transactional data from 3,900 purchases across various product categories. The goal is to uncover insights into spending patterns, customer segments, product preferences and subscription behavior to guide strategic business decisions.

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 Purchase, Review Rating, Shipping Type)

-**Missing Data**: 37 values in Review Rating column

Exploratory Data Analyst 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.

[1]:

	OrderID	OrderDate	CustomerID	CustomerName	ProductID	ProductName	Category	Brand	Quantity	UnitPrice	Discount	Tax	ShippingCost	TotalAmc
0	ORD0000001	2023-01-31	CUST001504	Vihaan Sharma	P00014	Drone Mini	Books	BrightLux	3	106.59	0.00	0.00	0.09	31
1	ORD0000002	2023-12-30	CUST000178	Pooja Kumar	P00040	Microphone	Home & Kitchen	UrbanStyle	1	251.37	0.05	19.10	1.74	25
2	ORD0000003	2022-05-10	CUST047516	Sneha Singh	P00044	Power Bank 20000mAh	Clothing	UrbanStyle	3	35.03	0.10	7.57	5.91	10
3	ORD0000004	2023-07-18	CUST030059	Vihaan Reddy	P00041	Webcam Full HD	Home & Kitchen	Zenith	5	33.58	0.15	11.42	5.53	15
4	ORD0000005	2023-02-04	CUST048677	Aditya Kapoor	P00029	T-Shirt	Clothing	KiddoFun	2	515.64	0.25	38.67	9.23	82

[15]:
```
df.describe()
```

[15]:

	Quantity	UnitPrice	Discount	Tax	ShippingCost	TotalAmount
count	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	3.001400	302.905748	0.074226	68.468902	7.406660	918.256479
std	1.413548	171.840797	0.082583	74.131180	4.324057	724.508332
min	1.000000	5.000000	0.000000	0.000000	0.000000	4.270000
25%	2.000000	154.190000	0.000000	15.920000	3.680000	340.890000
50%	3.000000	303.070000	0.050000	45.250000	7.300000	714.315000
75%	4.000000	451.500000	0.100000	96.060000	11.150000	1349.765000
max	5.000000	599.990000	0.300000	538.460000	15.000000	3534.980000

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.

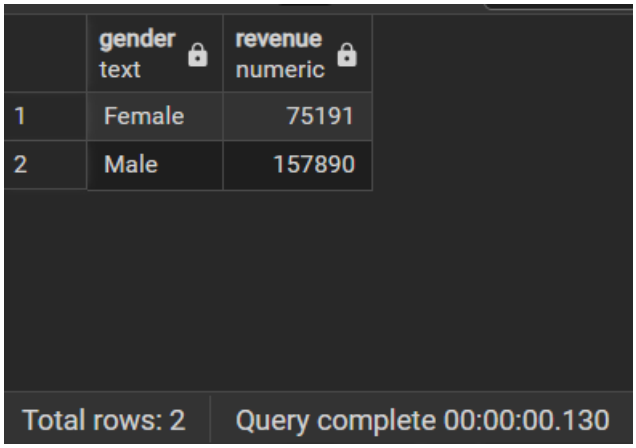
Data Consistency Check: Verified if discount_applied and promo_code_used were redundant ; dropped promo_code_used

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

Data Analysis using SQL

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

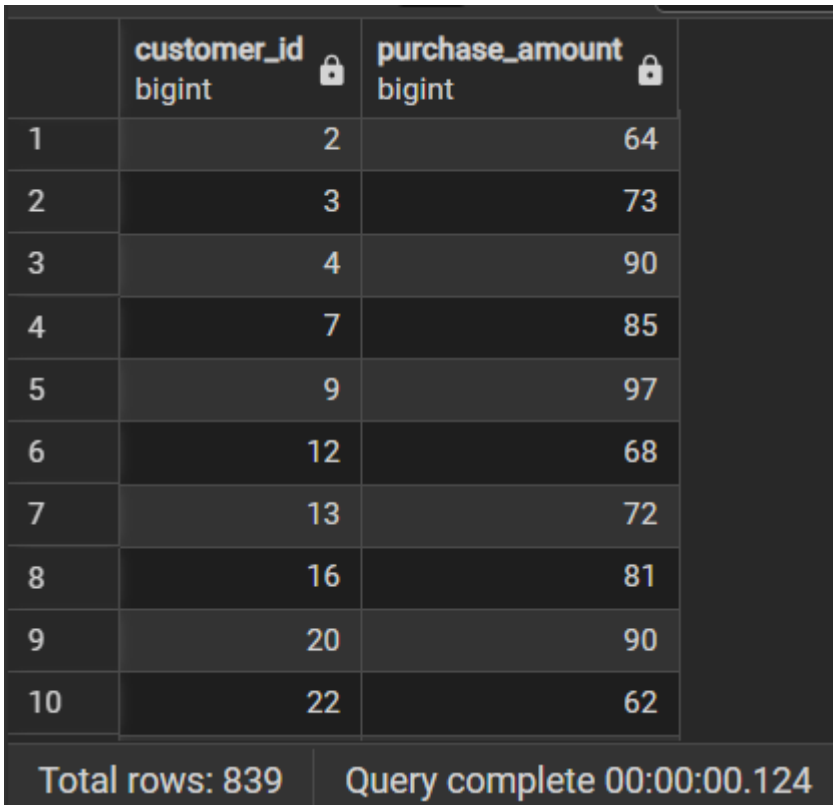
Revenue by Gender – Compared total revenue generated by male vs female customers.



	gender text	revenue numeric
1	Female	75191
2	Male	157890

Total rows: 2 | Query complete 00:00:00.130

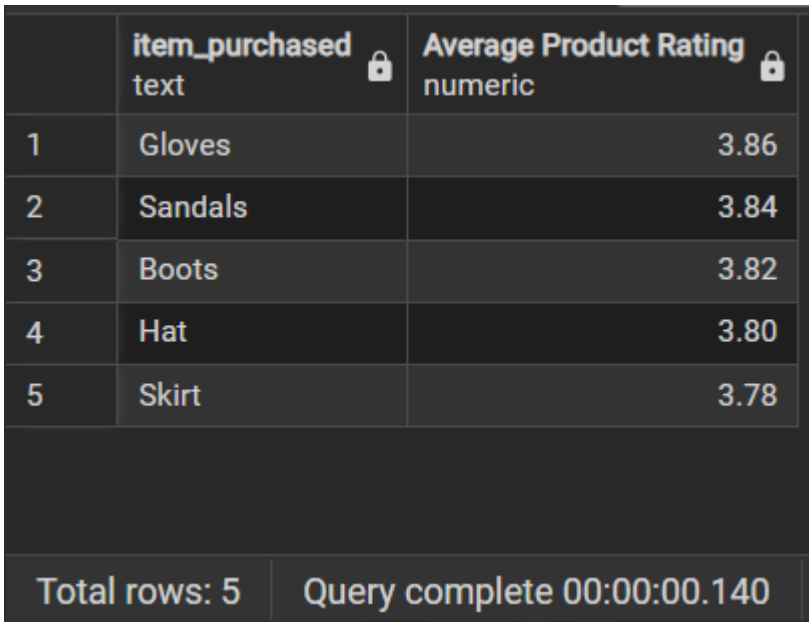
High-Spending Discount Users – Identified customers who used discounts but still spend above the average purchase amount.



	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

Total rows: 839 | Query complete 00:00:00.124

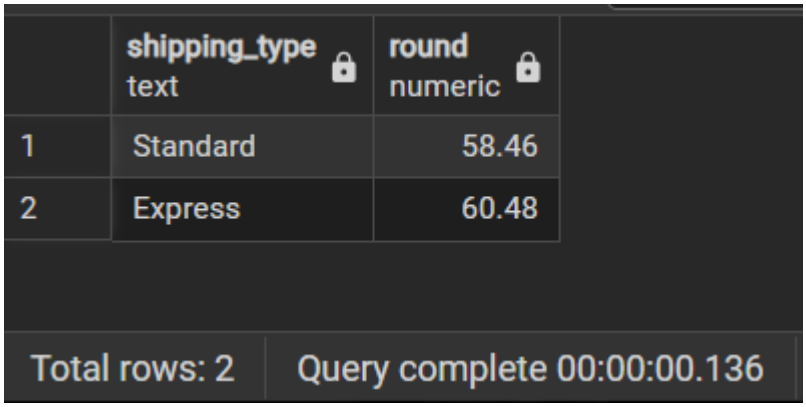
Top 5 Products by Rating – Found products with the highest average review ratings.



	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

Total rows: 5 Query complete 00:00:00.140

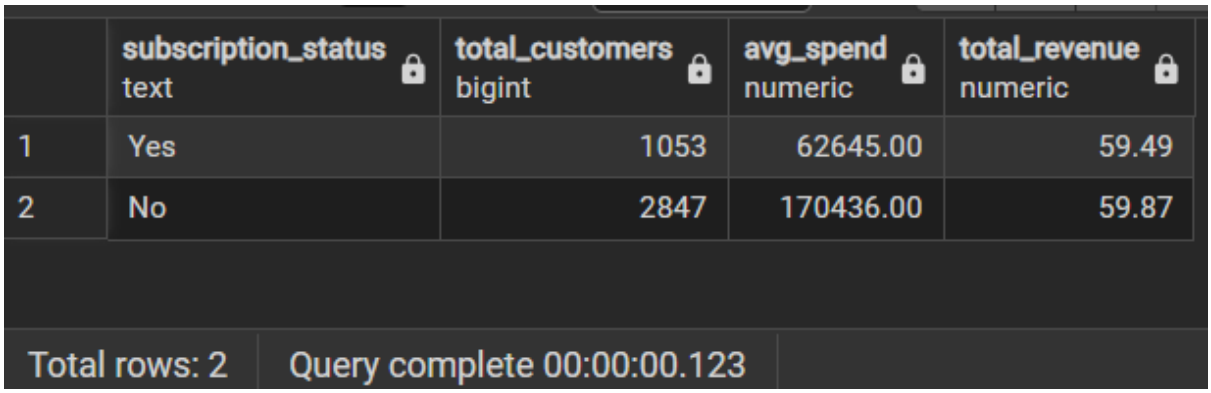
Shipping type Comparison – Compared average purchase amounts between standard and express shipping.



	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

Total rows: 2 Query complete 00:00:00.136

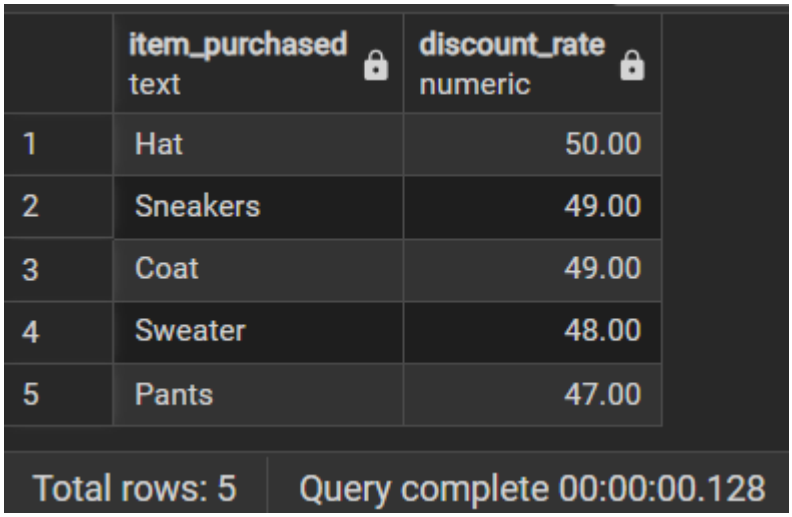
Subscribers vs Non Subscribers – Compared average spend and total revenue across subscription status.



	subscription_status text	total_customers bigint	avg_spend numeric	total_revenue numeric
1	Yes	1053	62645.00	59.49
2	No	2847	170436.00	59.87

Total rows: 2 | Query complete 00:00:00.123

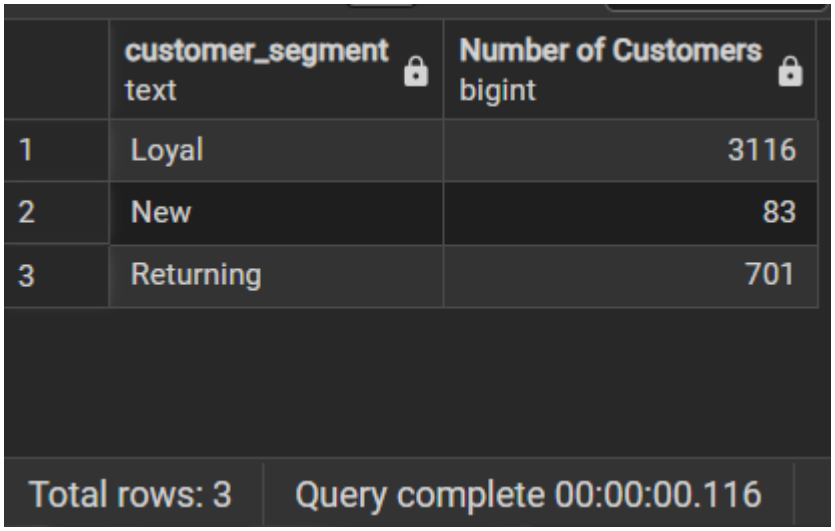
Discount-Dependent Products – Identified 5 products with the highest percentage of discounted purchases.



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

Total rows: 5 | Query complete 00:00:00.128

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



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

Total rows: 3 | Query complete 00:00:00.116

Top 3 Products per Category – Listed the main 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

Total rows: 11 Query complete 00:00:00.158

Repeat buyers and Subscriptions – Checked whether customers with >5 purchases are more likely to subscribe.

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

Total rows: 2 Query complete 00:00:00.159

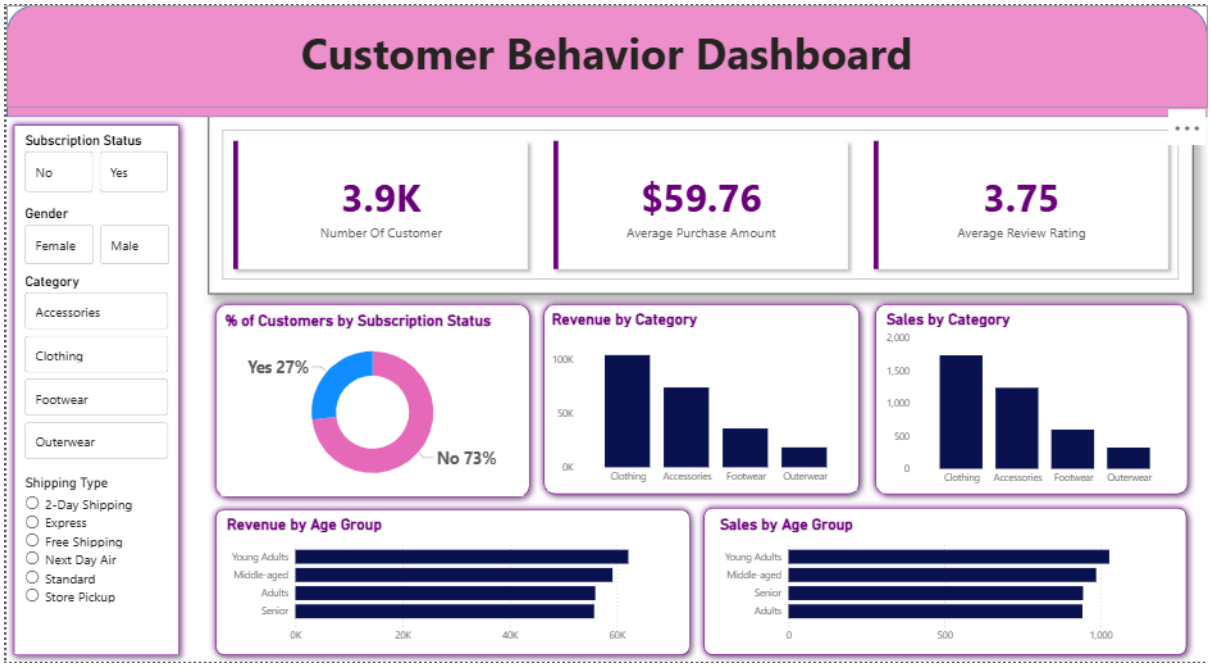
Revenue by Age Group – Calculated total revenue contributed of each age group.

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

Total rows: 4 | Query complete 00:00:00.116

Dashboard with Power BI

Finally we built an interactive dashboard in power bi to present insights visually.



Buisness Recommendations

- ✓ **Boost Subscriptions** – Promotes exclusive benefits for subscribers.
- ✓ **Customer Loyalty Programs** - Rewards 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.