Customer Shopping Behavior Analysis

1. 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.

2. 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

3. 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 `.describe()` for summary statistics.

[27]:

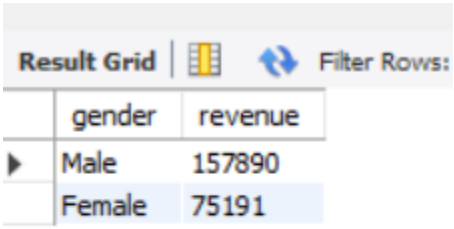
	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN

- **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 SQL analysis.

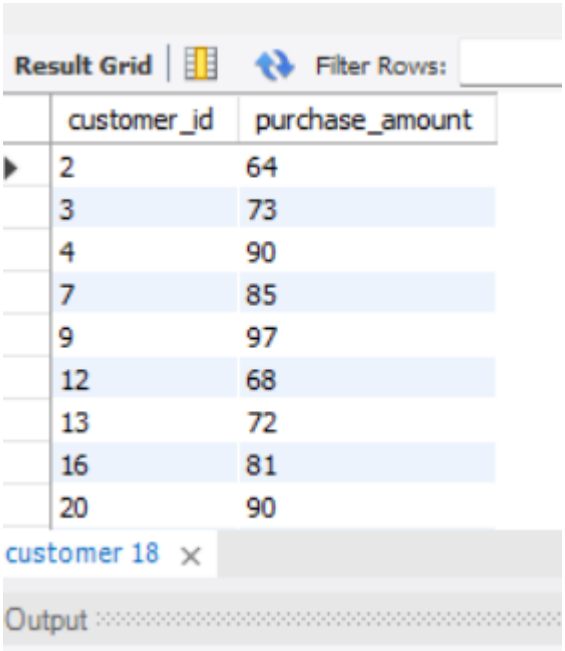
4. Data Analysis using SQL (Business Transactions)

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

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



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

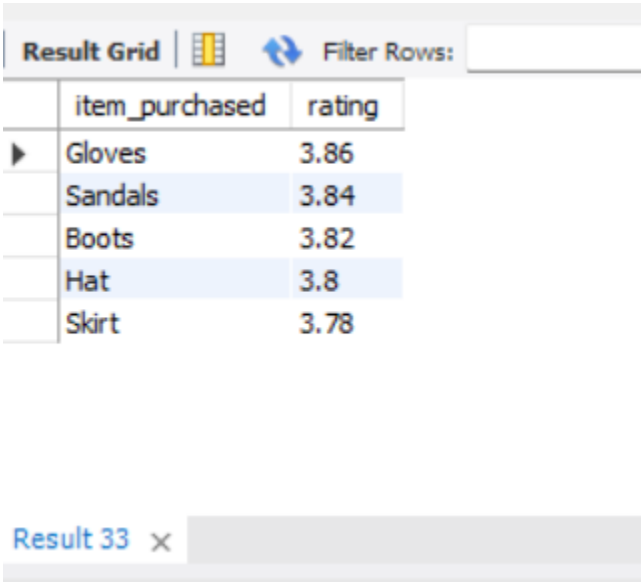

Result Grid | Filter Rows:

customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90

customer 18 ×

Output

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


Result Grid | Filter Rows:

item_purchased	rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Skirt	3.78

Result 33 ×

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

Result Grid | Filter Rows:

shipping_type	ROUND(AVG(purchase_amount),2)
Express	60.48
Standard	58.46

Result 20 ×

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

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

subscription_status	total_customers	avg_spend	total_revenue
Yes	1053	59.49	62645
No	2847	59.87	170436

Result 34 ×

Output

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

Result Grid | Filter Rows:

item_purchased	discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

Result 22 ×

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

Result Grid | Filter Rows:

customer_segment	segment_count
Loyal	3116
Returning	784

Result 23 ×

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

Result Grid | Filter Rows: | Export: | Wrap C

item_rank	category	item_purchased	total_orders
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
1	Clothing	Blouse	171
2	Clothing	Pants	171
3	Clothing	Shirt	169
1	Footwear	Sandals	160
2	Footwear	Shoes	150
3	Footwear	Sneakers	145

Result 24

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

Result Grid | Filter Rows:

subscription_status	repeat_buyers
Yes	958
No	2518

Result 25

Output

10. **Revenue by Age Group** – Calculated total revenue contribution of each age group.

Result Grid | Filter Rows:

age_group	total_revenue
Young Adult	62143
Middle-aged	59197
Adult	55978
Senior	55763

Result 26 ×

11.**Seasonal Revenue Breakdown** - A table ranking **total revenue** by category and season, showing **Clothing** as the top earner and **Spring** as the peak season for that category.

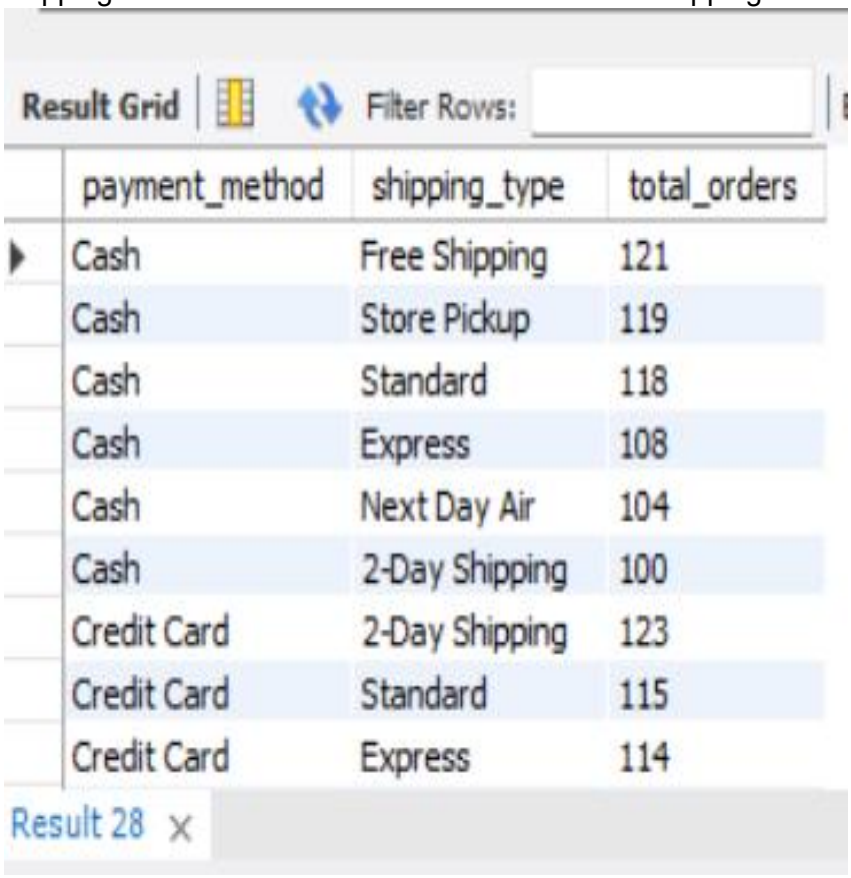
Result Grid | Filter Rows:

season	category	total_revenue
Spring	Clothing	27692
Winter	Clothing	27274
Fall	Clothing	26220
Summer	Clothing	23078
Fall	Accessories	19874
Summer	Accessories	19028
Winter	Accessories	18291
Spring	Accessories	17007
Spring	Footwear	9555

Result 27 ×

Output

11. **Shipping Preferences Analysis** - This data shows credit card users prefer 2-day shipping while cash users most often choose free shipping.

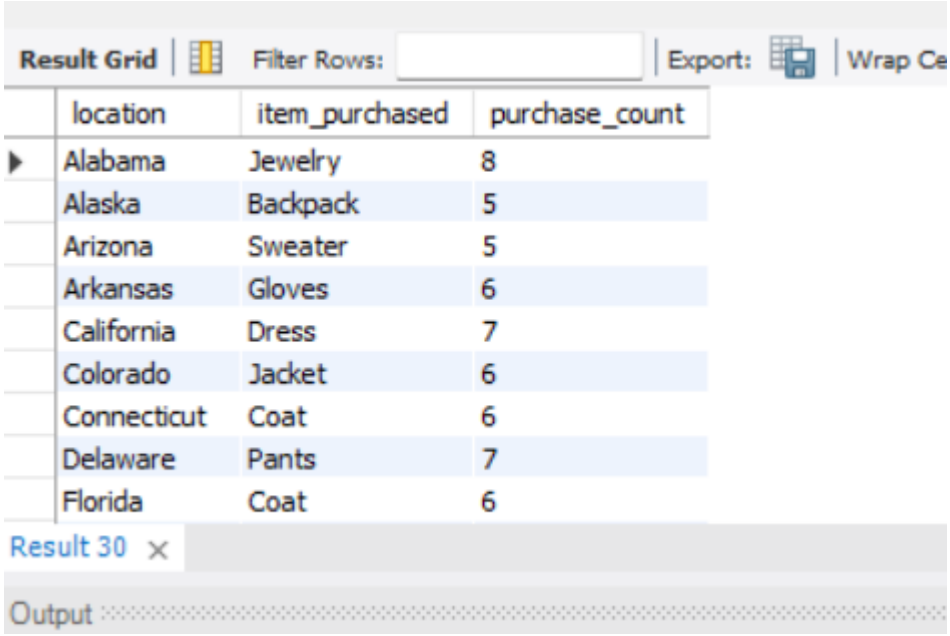


Result Grid | Filter Rows:

payment_method	shipping_type	total_orders
Cash	Free Shipping	121
Cash	Store Pickup	119
Cash	Standard	118
Cash	Express	108
Cash	Next Day Air	104
Cash	2-Day Shipping	100
Credit Card	2-Day Shipping	123
Credit Card	Standard	115
Credit Card	Express	114

Result 28

12. **Top Seller by Location** -This data identifies the most frequently purchased item in each state. It shows which specific products lead in sales across different geographical areas.



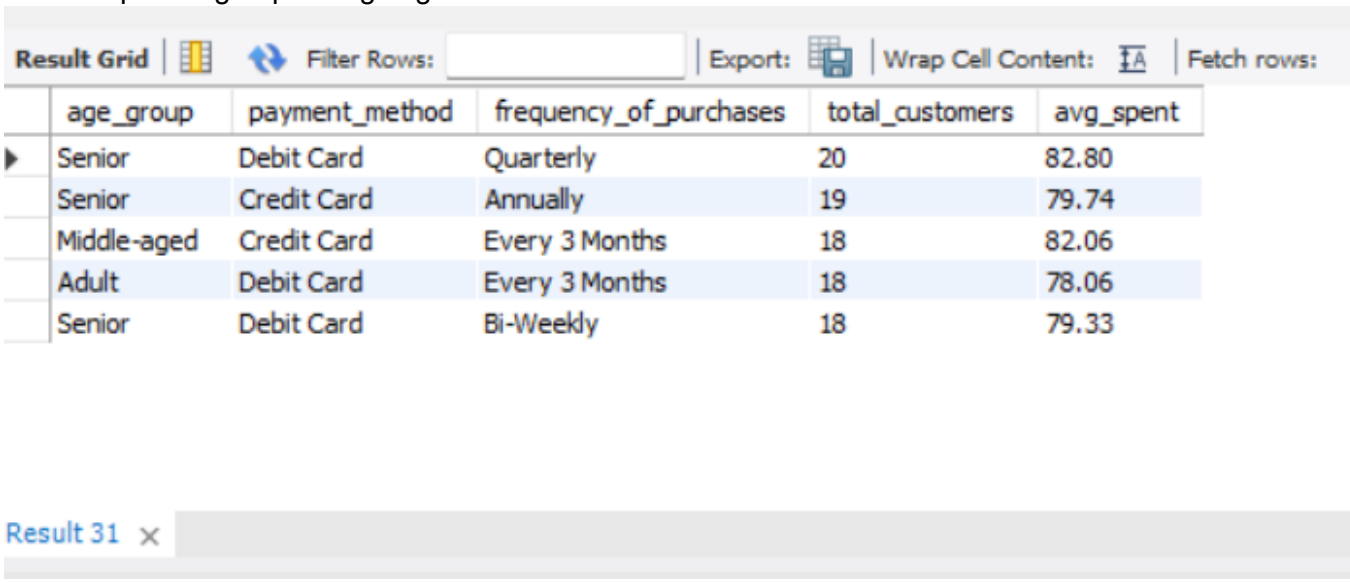
Result Grid | Filter Rows: | Export: | Wrap Ce

location	item_purchased	purchase_count
Alabama	Jewelry	8
Alaska	Backpack	5
Arizona	Sweater	5
Arkansas	Gloves	6
California	Dress	7
Colorado	Jacket	6
Connecticut	Coat	6
Delaware	Pants	7
Florida	Coat	6

Result 30

Output

14. **Premium Customer Profile** -Seniors using debit cards on a quarterly basis represent the most frequent high-spending segment.



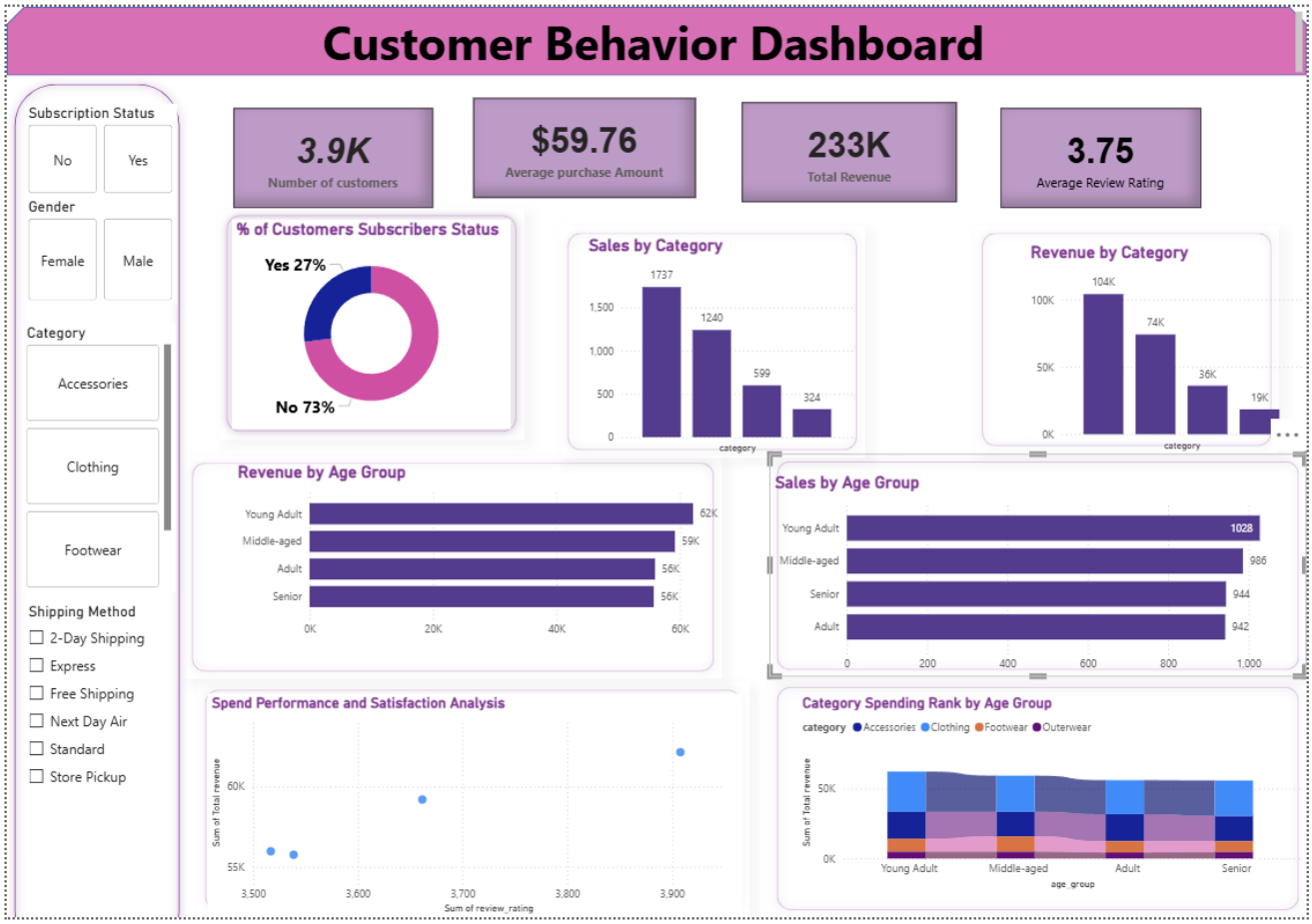
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

age_group	payment_method	frequency_of_purchases	total_customers	avg_spent
Senior	Debit Card	Quarterly	20	82.80
Senior	Credit Card	Annually	19	79.74
Middle-aged	Credit Card	Every 3 Months	18	82.06
Adult	Debit Card	Every 3 Months	18	78.06
Senior	Debit Card	Bi-Weekly	18	79.33

Result 31 ×

5. Dashboard in Power BI

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



6. 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.