Customer Purchasing Behavior Analysis

1. Project Summary

This project explores customer purchasing patterns using transaction records from **3,900 orders** spanning multiple product categories. The main objective is to extract meaningful insights about spending habits, customer groups, product trends, and subscription engagement to support smarter business strategies.

2. Dataset Overview

- **Total Records:** 3,900
- **Total Fields:** 18

Major Data Groups:

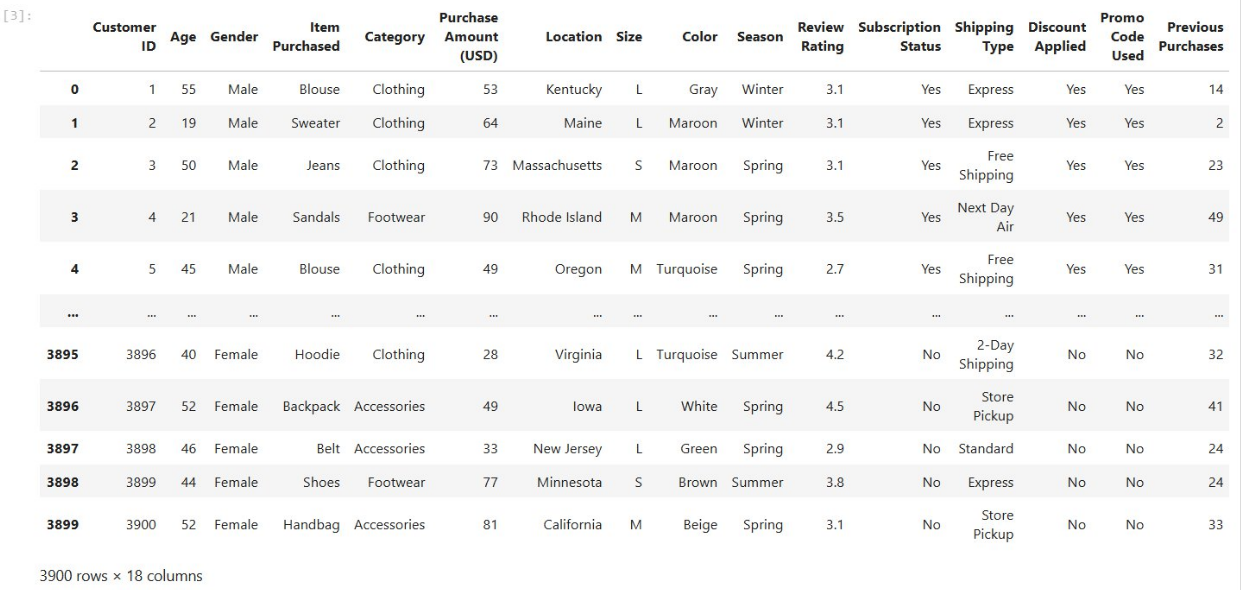
- **Customer Information:** Age, gender, location, and subscription status
- **Purchase Information:** Product name, category, purchase value, season, size, and color
- **Behavioral Metrics:** Discount usage, promo code usage, prior purchases, buying frequency, review scores, and shipping method

Data Quality Note:
There were **37 missing entries** in the *review rating* field.

3. Exploratory Data Analysis in Python

Data preparation and exploration were conducted using Python before moving into database analysis.

- **Data Import:** The dataset was loaded using the pandas library.

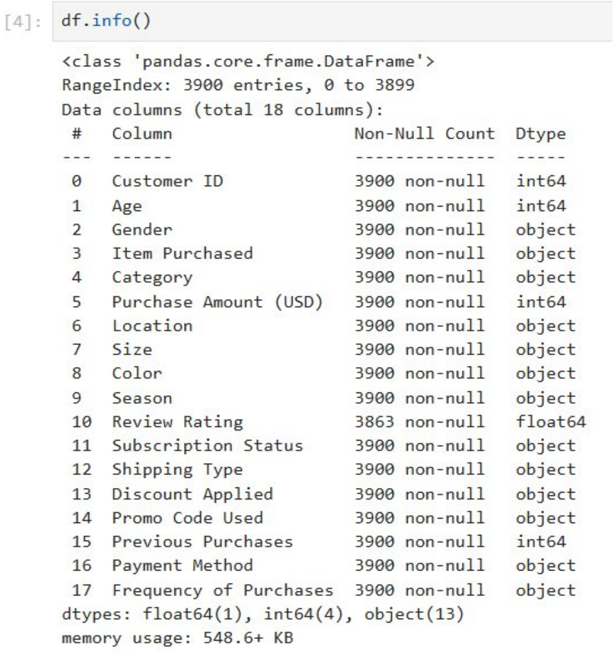


[3]:

	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
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	2-Day Shipping	No	No	32
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	Store Pickup	No	No	41
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	Standard	No	No	24
3898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No	Express	No	No	24
3899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No	Store Pickup	No	No	33

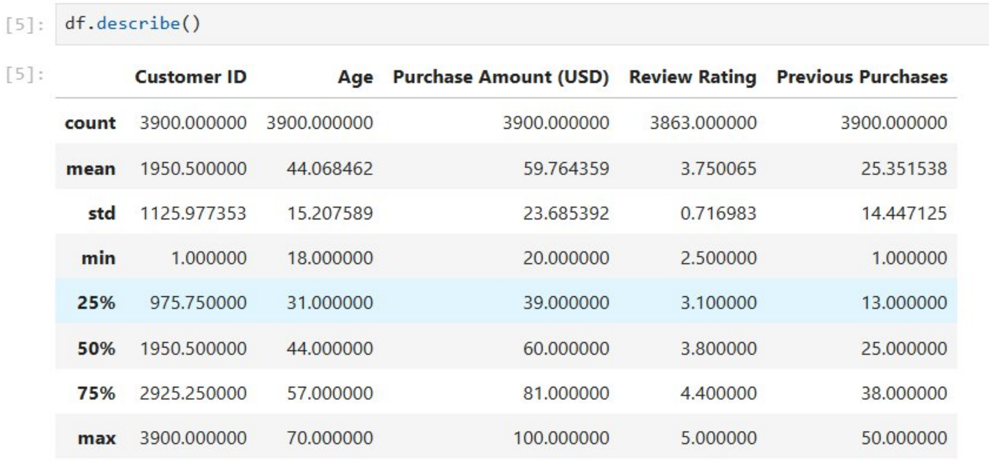
3900 rows × 18 columns

- **Preliminary Inspection:** Data structure and statistical summaries were reviewed using functions like .info() and .describe().



```
[4]: df.info()

<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
```



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

[5]:

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3863.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.750065	25.351538
std	1125.977353	15.207589	23.685392	0.716983	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.800000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

- **Handling Missing Values:** Missing review ratings were filled using the **median rating within each product category** to maintain consistency.
- **Column Formatting:** Column names were standardized into **snake_case** for clarity and uniformity.

Feature Creation:

- Generated an **age_group** variable by grouping customer ages into ranges.
- Derived a **purchase_frequency_days** feature to better understand buying intervals.

Data Validation:

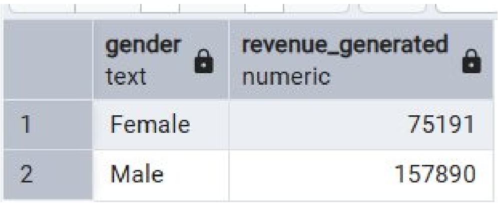
- Checked for overlap between *discount_applied* and *promo_code_used* fields; since both conveyed similar information, the *promo_code_used* column was removed.

Database Upload:
The cleaned dataset was transferred to a **PostgreSQL database** for advanced querying and structured analysis.

4. Business Analysis Using PostgreSQL

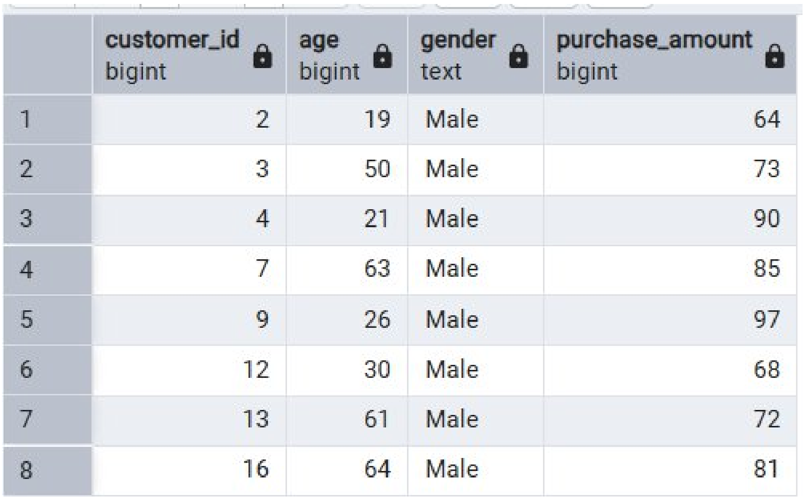
SQL queries were used to investigate key business performance questions:

1. **Revenue by Gender** – Measured total sales generated by different genders.



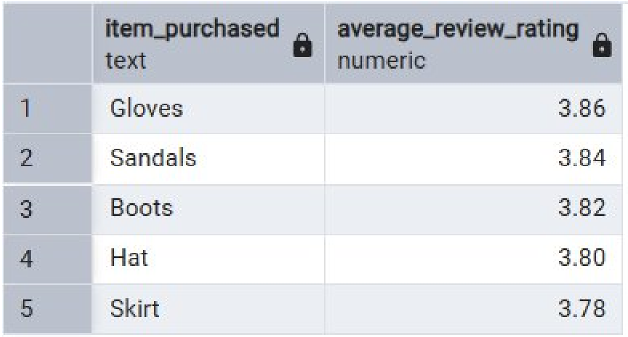
	gender text	revenue_generated numeric
1	Female	75191
2	Male	157890

2. **Big Spenders Using Discounts** – Found customers who received discounts but still spent more than the overall average.



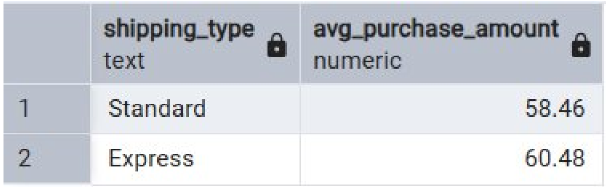
	customer_id bigint	age bigint	gender text	purchase_amount bigint
1	2	19	Male	64
2	3	50	Male	73
3	4	21	Male	90
4	7	63	Male	85
5	9	26	Male	97
6	12	30	Male	68
7	13	61	Male	72
8	16	64	Male	81

3. **Highest-Rated Products** – Identified the top five products based on average customer ratings.



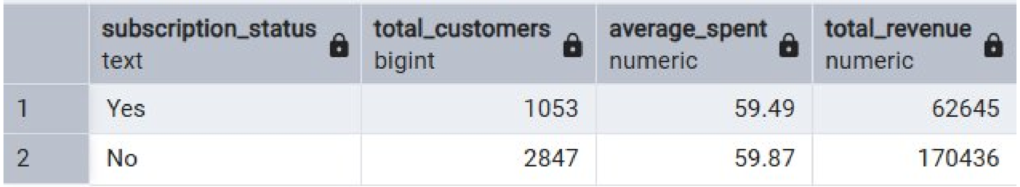
	item_purchased text	average_review_rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

4. **Impact of Shipping Type** – Compared spending patterns between standard and express delivery users.



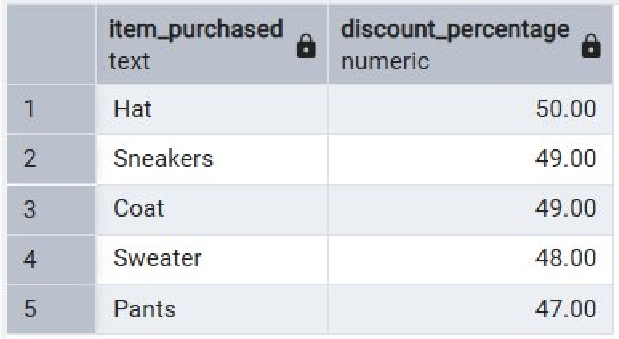
	shipping_type text	avg_purchase_amount numeric
1	Standard	58.46
2	Express	60.48

5. **Subscribers vs Non-Subscribers** – Evaluated differences in average purchase value and total contribution to revenue.



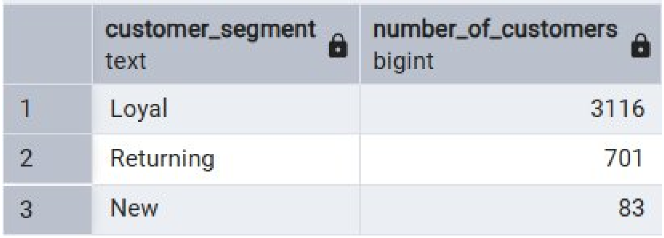
	subscription_status text	total_customers bigint	average_spent numeric	total_revenue numeric
1	Yes	1053	59.49	62645
2	No	2847	59.87	170436

6. **Products Most Reliant on Discounts** – Determined the five products most frequently purchased with discounts.



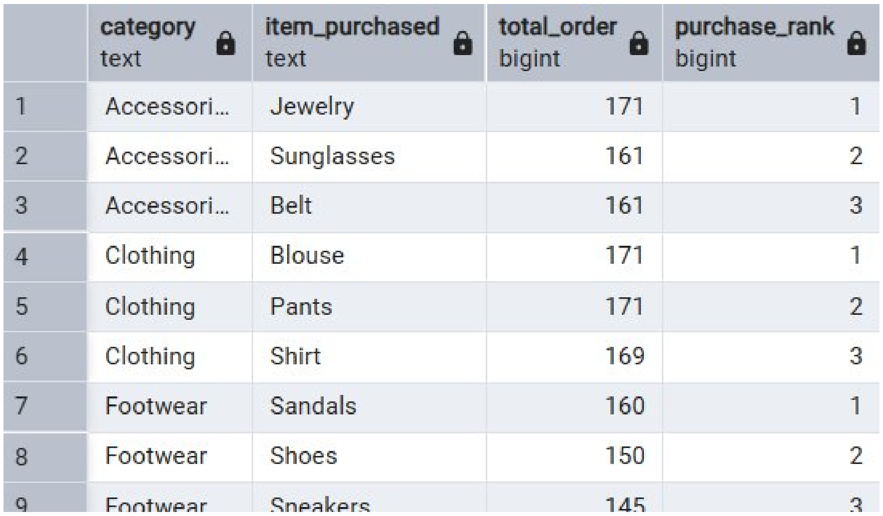
	item_purchased text	discount_percentage numeric
1	Hat	50.00
2	Sneakers	49.00
3	Coat	49.00
4	Sweater	48.00
5	Pants	47.00

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



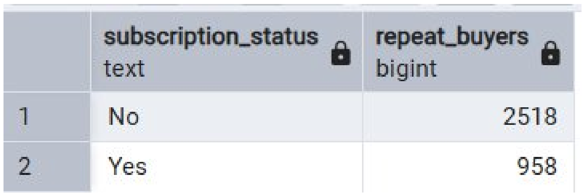
	customer_segment text	number_of_customers bigint
1	Loyal	3116
2	Returning	701
3	New	83

8. **Top Products by Category** – Selected the three most purchased items in each product category.



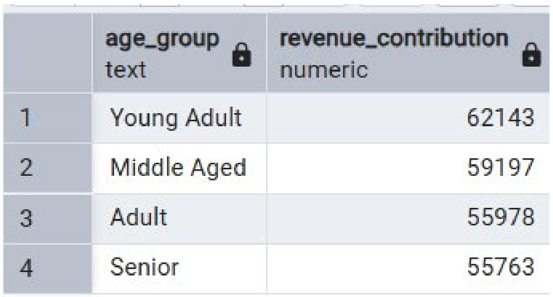
	category text	item_purchased text	total_order bigint	purchase_rank bigint
1	Accessori...	Jewelry	171	1
2	Accessori...	Sunglasses	161	2
3	Accessori...	Belt	161	3
4	Clothing	Blouse	171	1
5	Clothing	Pants	171	2
6	Clothing	Shirt	169	3
7	Footwear	Sandals	160	1
8	Footwear	Shoes	150	2
9	Footwear	Sneakers	145	3

9. **Repeat Purchase Behavior & Subscriptions** – Assessed whether frequent buyers (more than five purchases) were more likely to hold subscriptions.



	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

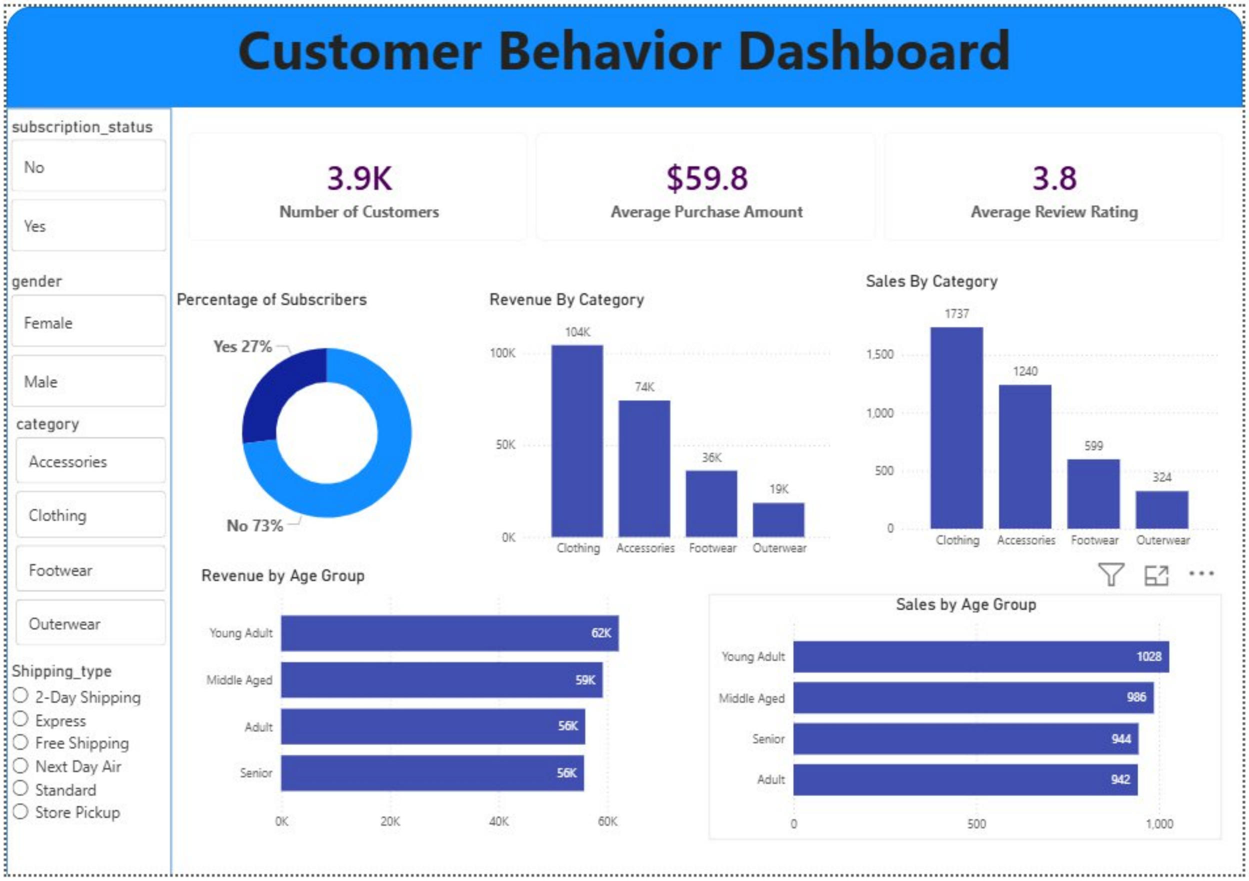
10. **Revenue by Age Segment** – Calculated how much each age group contributes to overall sales.



	age_group text	revenue_contribution numeric
1	Young Adult	62143
2	Middle Aged	59197
3	Adult	55978
4	Senior	55763

5. Power BI Dashboard

An interactive **Power BI dashboard** was developed to visually communicate findings, allowing stakeholders to explore trends, customer segments, and product performance dynamically.



Customer Behavior Dashboard

3.9K
Number of Customers

$59.8
Average Purchase Amount

3.8
Average Review Rating

6. Strategic Business Suggestions

- **Encourage Subscriptions:** Introduce special perks and incentives to convert more customers into subscribers.
- **Strengthen Loyalty Programs:** Offer rewards and retention benefits to frequent shoppers to build long-term loyalty.
- **Optimize Discount Strategies:** Use discounts strategically to drive sales without significantly impacting profit margins.
- **Promote Strong Products:** Feature top-rated and high-demand products in marketing campaigns.
- **Refine Target Marketing:** Direct campaigns toward age groups and customer segments that contribute the most revenue, including those preferring express shipping.