CUSTOMER SHOPPING TREND ANALYSIS

1. Project Overview

This analytical initiative is focused on leveraging 3,500 transactional records to understand core drivers of sales performance and customer retention. The primary objective is to dissect purchase behaviors, with a special emphasis on identifying high-value customer segments, evaluating the effectiveness of the current subscription model, and providing actionable insights to optimize marketing spend and product placement for sustained revenue growth.

2. Dataset Summary

- Rows: 3,500
- Columns: 18
- Missing Data: 37 values in Review Rating column
- 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)

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.

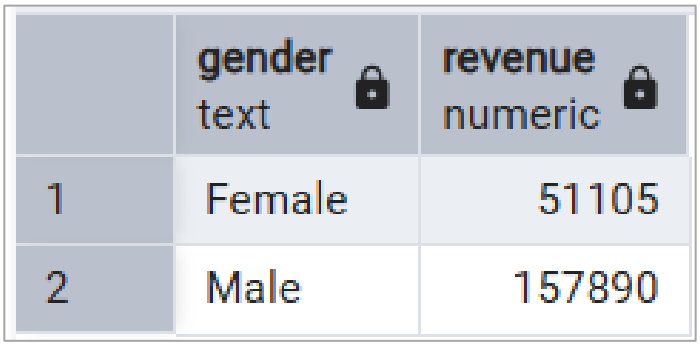
	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
count	3500.000000	3500.000000	3500	3500	3500	3500.000000	3500	3500	3500	3500	3463.000000	3500	3500	3500	3500	3500.000000	3500	3500
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN	6	7
top	NaN	NaN	Male	Pants	Clothing	NaN	Montana	M	Teal	Spring	NaN	No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
freq	NaN	NaN	2652	156	1560	NaN	88	1556	161	891	NaN	2447	601	1823	1823	NaN	610	523
mean	1750.500000	44.032571	NaN	NaN	NaN	59.712857	NaN	NaN	NaN	NaN	3.748859	NaN	NaN	NaN	NaN	25.373143	NaN	NaN
std	1010.507298	15.233519	NaN	NaN	NaN	23.713364	NaN	NaN	NaN	NaN	0.716415	NaN	NaN	NaN	NaN	14.441421	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	1.000000	NaN	NaN
25%	875.750000	31.000000	NaN	NaN	NaN	38.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000	NaN	NaN
50%	1750.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000	NaN	NaN
75%	2625.250000	57.000000	NaN	NaN	NaN	80.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000	NaN	NaN
max	3500.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	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.

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	51105
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	88

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	3.89
2	Sandals	3.85
3	Hat	3.84
4	Boots	3.82
5	Skirt	3.78

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

	shipping_type text	Average Purchase Amount numeric
1	Standard	58.44
2	Express	60.11

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

	subscription_status text	Total Customer bigint	AVG Spend numeric	Total Revenue numeric
1	Yes	1053	59.49	62645
2	No	2447	59.81	146350

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

	item_purchased text	Discount Rate numeric
1	Hat	59.00
2	Sneakers	55.00
3	Coat	52.00
4	Sweater	51.00
5	Pants	51.00

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

	customer_segment text	Numbers of Customer bigint
1	Loyal	2802
2	New	70
3	Returning	628

8. **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	Accessories	Belt	152
2	2	Accessories	Jewelry	152
3	3	Accessories	Sunglasses	144
4	1	Clothing	Pants	156
5	2	Clothing	Sweater	152
6	3	Clothing	Dress	151
7	1	Footwear	Boots	135
8	2	Footwear	Sandals	134
9	3	Footwear	Shoes	132
10	1	Outerwear	Coat	150
11	2	Outerwear	Jacket	147

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

	subscription_status text	Repeat Buyers bigint
1	No	2174
2	Yes	958

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

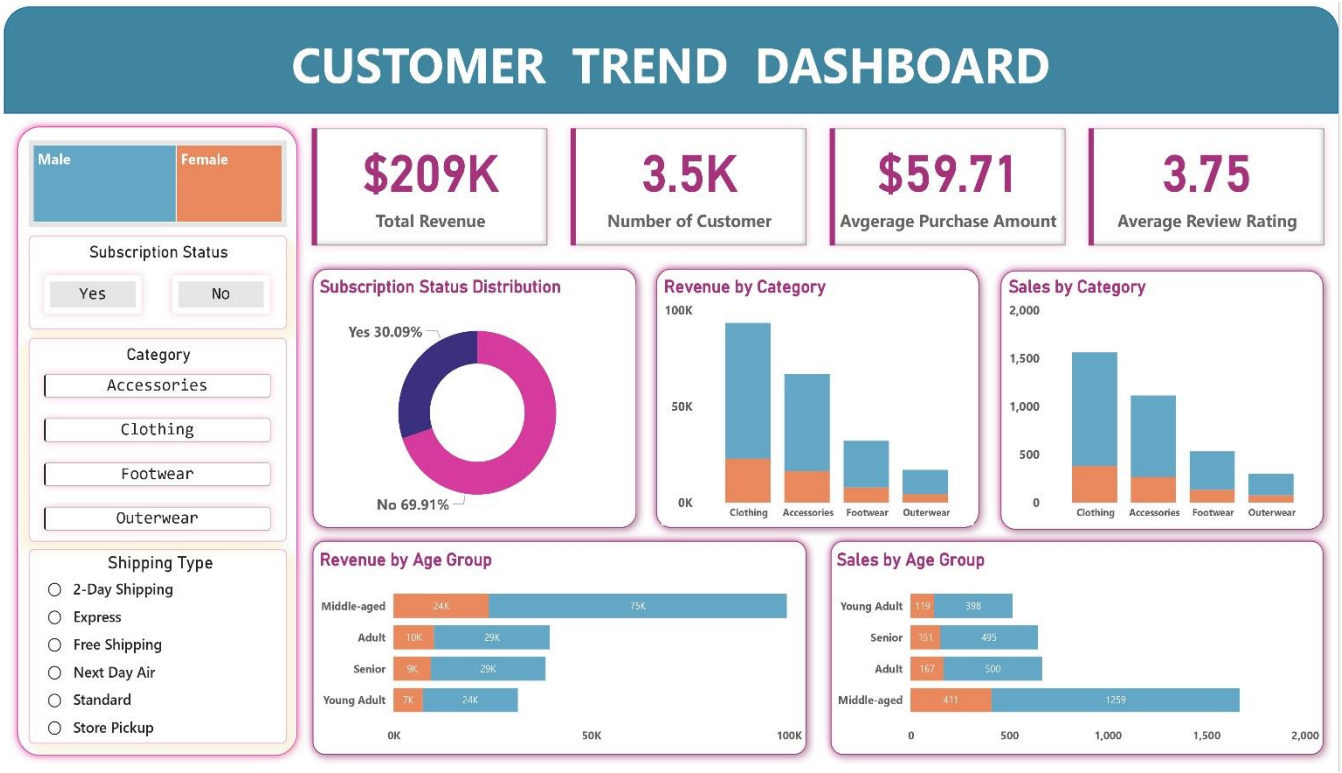
	age_group text	Total Revenue numeric
1	Middle-aged	99530
2	Adult	39555
3	Senior	38458
4	Young Adult	31452

5. Dashboard in Power BI

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

The primary Key Performance Indicators (KPIs) are:

- Total Revenue: $209K
- Number of Customers: 3.5K
- Average Purchase Amount: $59.71
- Average Review Rating: 3.75
- Subscription Status Distribution: Yes - 30.09% | No - 69.91%



6. Business Recommendation

I. Strategic Business Recommendations

Based on the SQL query results, the following actions are the most impactful for revenue growth and margin control:

1. **Maximize Gender/Age Focus (Targeting):**
   - Action: Invest in highly targeted marketing campaigns aimed at the Male segment and the Middle-aged group (highest revenue contributors).
   - Goal: Capitalize on the most valuable existing customer segments for maximum return on ad spend (ROAS).

2. **Address Discount-Dependency (Margin Control):**
    - Action: Re-evaluate the discount strategy for high-volume, discount-dependent items like Hat (59.00% discount rate) and Sneakers (55.00% discount rate).
    - Goal: Protect profit margins by substituting deep discounts with value-added bundles or loyalty points.

3. **Optimize Loyalty Conversion (Retention):**
    - Action: Design targeted campaigns (e.g., email sequences, special offers) to convert the 628 Returning customers into the highly stable Loyal segment (2,802 customers).
    - Goal: Increase the Customer Lifetime Value (CLTV) by moving repeat buyers into the most retained group.

❗ **NOTE ON SUBSCRIPTION DATA:** The raw SQL data showed Non-Subscribers had a negligibly higher average spend ($59.81) than Subscribers ($59.49). The largest opportunity lies in converting the high volume of Non-Subscribers (69.91% of base).

II. Secondary Insights & Product Health

- **Top Revenue Categories:** The Clothing and Accessories categories are the highest revenue drivers.
- **Top-Rated Products:** Prominently feature top-rated products like Gloves (3.89 avg. rating) and Sandals (3.85 avg. rating) in marketing to build trust.