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 objective is to uncover insights into spending patterns, customer segments, product preferences, and subscription behavior to support data-driven business decisions.

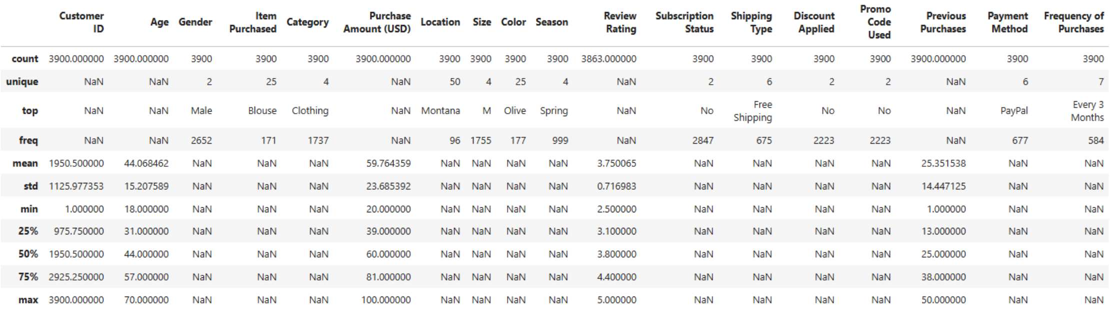
2 Dataset Summary

- 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 the *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 inspect structure and `df.describe()` for summary statistics.

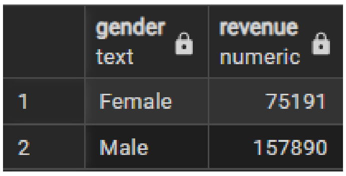


	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3900.000000	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN	6	7
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	NaN	677	584
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	25.351538	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	14.447125	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%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	50.000000	NaN	NaN

- **Missing Data Handling:** Imputed missing values in the *Review Rating* column using the median rating of each product category.
- **Column Standardization:** Renamed columns to snake_case for improved readability.
- **Feature Engineering:**
  - Created `age_group` by binning customer ages.
  - Created `purchase_frequency_days` from purchase history.
- **Data Consistency Check:** Verified redundancy between `discount_applied` and `promo_code_used`; removed `promo_code_used`.
- **Database Integration:** Loaded the cleaned dataset into PostgreSQL for SQL analysis.

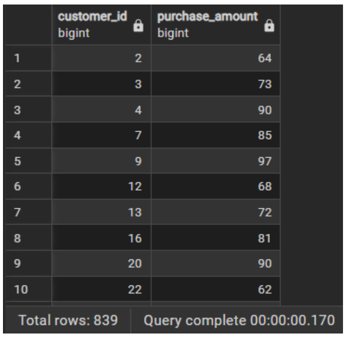
4 SQL-Based Analysis

- Revenue by gender



	gender text	revenue numeric
1	Female	75191
2	Male	157890

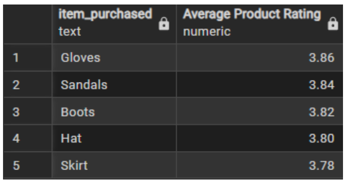
- High-spending discount users



	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

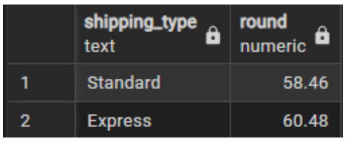
Total rows: 839 Query complete 00:00:00.170

- Top five products by average rating



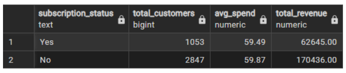
	item_purchased text	Average Product Rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

- Shipping type comparison



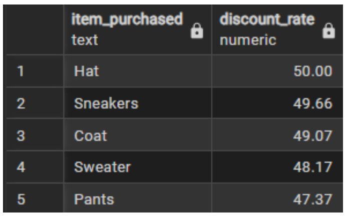
	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

- Subscribers vs. non-subscribers analysis



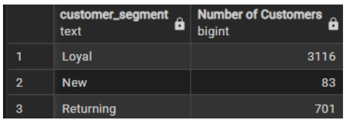
	subscription_status text	total_customers bigint	avg_spend numeric	total_revenue numeric
1	Yes	1053	59.49	62645.00
2	No	2847	59.87	170436.00

- Discount-dependent products



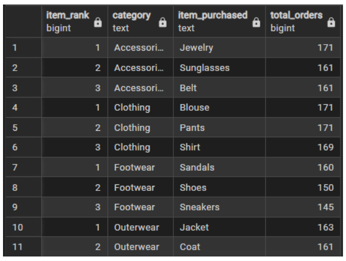
	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

- Customer segmentation (New, Returning, Loyal)



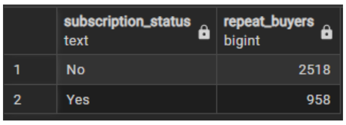
	customer_segment text	Number of Customers bigint
1	Loyal	3116
2	New	83
3	Returning	701

- Top products per category



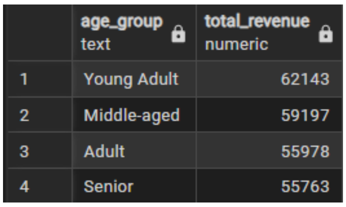
	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	Belt	161
4	1	Clothing	Blouse	171
5	2	Clothing	Pants	171
6	3	Clothing	Shirt	169
7	1	Footwear	Sandals	160
8	2	Footwear	Shoes	150
9	3	Footwear	Sneakers	145
10	1	Outerwear	Jacket	163
11	2	Outerwear	Coat	161

- Repeat buyers and subscription likelihood



	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

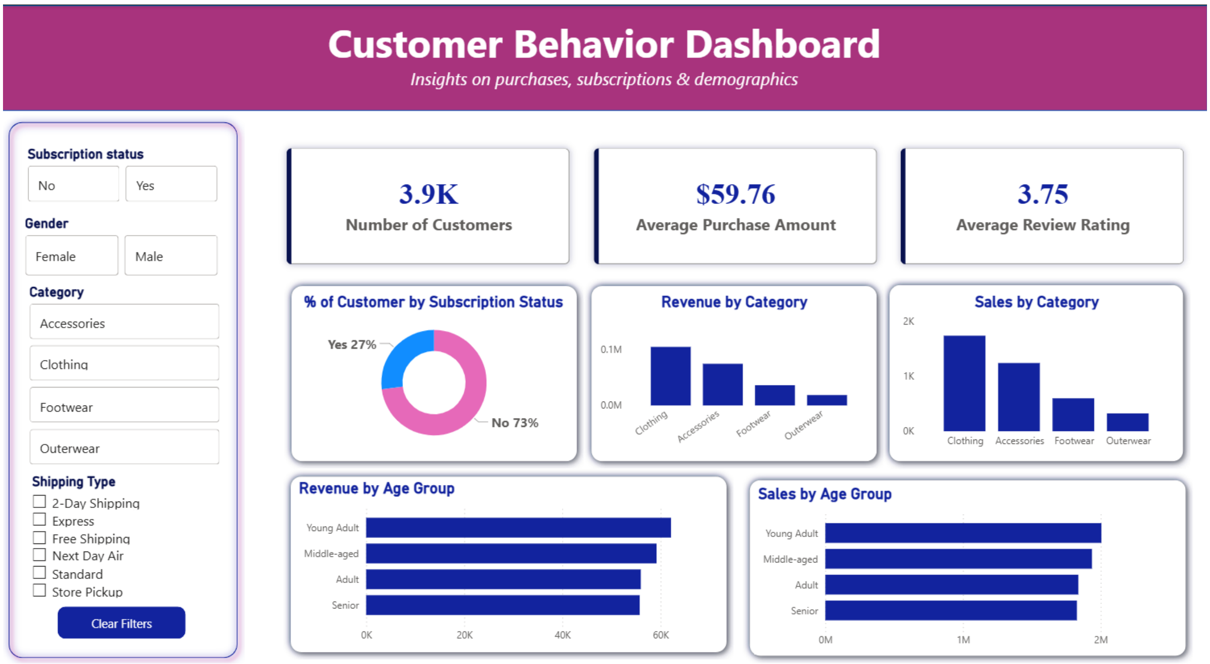
- Revenue by age group



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

5 Dashboard

An interactive Power BI dashboard was developed to visualize customer behavior, revenue distribution, subscription trends, and category performance.



6 Business Recommendations

- Promote exclusive subscription benefits
- Incentivize repeat buyers through loyalty programs
- Balance discount strategies with profitability
- Highlight top-rated and best-selling products
- Focus marketing on high-revenue customer segments