

CUSTOMER SHOPPING BEHAVIOR ANALYSIS

1. Project Objective

The objective of this project is to analyze customer shopping behavior and identify patterns that can support data-driven business decisions related to revenue growth, customer retention, subscription strategy, and product performance.

This project simulates a real-world analytics workflow where raw data is cleaned, structured, analyzed using SQL, and visualized through an interactive dashboard for stakeholders.

2. Dataset Overview

- Total Records: ~3,900 customer transactions
- Total Features: 18
- Data Type: Transactional retail data

Key Data Categories:

- **Customer Attributes:** Age, Gender, Location, Subscription Status
- **Purchase Details:** Item Purchased, Category, Purchase Amount, Season
- **Behavioral Indicators:** Previous Purchases, Discount Applied, Shipping Type, Review Rating

Data Quality Notes:

- Minor missing values were identified in the Review Rating column.
- Some categorical columns had inconsistent formatting, which required standardization.

3. Data Preparation & Feature Engineering (Python)

Data cleaning and preparation were performed in Jupyter Notebook using Pandas.

Key Steps:

- Loaded and explored data structure using `df.info()` and summary statistics.
- Identified and handled missing values to maintain data integrity.
- Standardized column names for consistency and easier SQL querying.
- Engineered new features to improve analytical depth:
 - Age Group: Grouped customers into meaningful age segments.

- Purchase Frequency (Days): Converted frequency labels into numeric values.
- Performed redundancy checks and removed non-essential columns.
- Prepared the final dataset for database storage.

This step ensured that the data was analysis-ready before being pushed to the database.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	39
unique	Nan	Nan	2	25	4	Nan	50	4	25	4	Nan	2	6	
top	Nan	Nan	Male	Blouse	Clothing	Nan	Montana	M	Olive	Spring	Nan	No	Free Shipping	
freq	Nan	Nan	2652	171	1737	Nan	96	1755	177	999	Nan	2847	675	22
mean	1950.500000	44.068462	Nan	Nan	Nan	59.764359	Nan	Nan	Nan	Nan	3.750065	Nan	Nan	Nan
std	1125.977353	15.207589	Nan	Nan	Nan	23.685392	Nan	Nan	Nan	Nan	0.716983	Nan	Nan	Nan
min	1.000000	18.000000	Nan	Nan	Nan	20.000000	Nan	Nan	Nan	Nan	2.500000	Nan	Nan	Nan
25%	975.750000	31.000000	Nan	Nan	Nan	39.000000	Nan	Nan	Nan	Nan	3.100000	Nan	Nan	Nan
50%	1950.500000	44.000000	Nan	Nan	Nan	60.000000	Nan	Nan	Nan	Nan	3.800000	Nan	Nan	Nan
75%	2925.250000	57.000000	Nan	Nan	Nan	81.000000	Nan	Nan	Nan	Nan	4.400000	Nan	Nan	Nan
max	3900.000000	70.000000	Nan	Nan	Nan	100.000000	Nan	Nan	Nan	Nan	5.000000	Nan	Nan	Nan

Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3900	3900	3900.000000	3900	3900
2	2	Nan	6	7
No	No	Nan	PayPal	Every 3 Months
2223	2223	Nan	677	584
Nan	Nan	25.351538	Nan	Nan
Nan	Nan	14.447125	Nan	Nan
Nan	Nan	1.000000	Nan	Nan
Nan	Nan	13.000000	Nan	Nan
Nan	Nan	25.000000	Nan	Nan
Nan	Nan	38.000000	Nan	Nan
Nan	Nan	50.000000	Nan	Nan

4. SQL-Based Business Analysis (PostgreSQL)

The cleaned dataset was loaded into **PostgreSQL** to simulate enterprise-level data analysis.

Using SQL, the following business questions were answered:

1. Revenue comparison between male and female customers

	gender text	revenue numeric
1	Female	75191
2	Male	157890

2. Identification of high-spending customers who used discounts

	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 00:00:0

3. Top-rated products based on customer reviews

	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

4. Impact of shipping type on average purchase amount

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. Spending behavior of subscribed vs. non-subscribed customers

	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

6. Products most dependent on discounts

	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

7. Customer segmentation into New, Returning, and Loyal groups

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

8. Top products within each category

	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessories	Jewelry	171
2	2	Accessories	Sunglasses	161
3	3	Accessories	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

9. Subscription likelihood among repeat buyers

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

10. Revenue contribution by age group

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

This step demonstrated the use of **aggregation, filtering, grouping, subqueries, and CTEs** to extract actionable insights.

5. Data Visualization & Dashboard (Power BI)

An interactive dashboard was built using **Power BI** to communicate insights effectively.



7. Business Recommendations

Based on the analysis, the following recommendations are proposed:

- Strengthen subscription benefits to improve customer lifetime value.

- Implement loyalty programs for repeat customers.
- Optimize discount strategies for high-revenue categories.
- Use customer segmentation to enable targeted marketing campaigns.
- Focus product promotions on top-rated and high-performing items.

8. Tools & Technologies Used

- Python (Pandas, NumPy) – Data cleaning and feature engineering
- PostgreSQL – Business-focused SQL analysis
- Power BI – Interactive dashboard and data storytelling