

Customer Shopping Behavior Analysis

1. Project Overview

This project explores customer shopping habits by explaining 3900 transactions records across multiple categories. The main aim is to identify spending trends, customer groups, product choices and subscription patterns to support data driven business decisions.

2. About Dataset

- Rows: 3900
- Columns: 18
- Key Features:
 - a. Customer demographics (Age, Gender, Location, Subscription Status)
 - b. Purchase (Items purchased , Purchase Category, Purchase Amount, Season, Size, Color)
 - c. Shopping behavior (Discount applied, Promo code used, Previous purchases, Frequency of Purchases, Review Rating, Shipping Type)
 - d. Missing data: 37 values in Review Rating Column

3. Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

- a. **Data Loading:** Imported the csv file in python using pandas
- b. **Initial exploration:** Used df.info() to check 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
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually

- c. **Missing data:** Identified any missing entries and filled gaps in the "Review Rating" columns by applying median rating specific to each product category.
- d. **Column Standardization:** Renamed columns to snake case for better readability and documentation
- e. **Feature Engineering:**
 - i. Created age_group column by binning customer ages.
 - ii. Created purchase_frequency_days column from purchased data.

- f. **Data consistency check:** Verified if discount_applied and promo_code_used were redundant; dropped promo_code_used.
- g. **Database integration:** Connected python script to Postgre SQL and loaded the cleaned DataFrame into database for SQL analysis.

4. Data Analysis using SQL (Business Transactions)

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

- i. **Revenue by Gender:** Compared total revenue generated by male vs female

	gender text	revenue numeric
1	Female	75191
2	Male	157890

- ii. **High spending discount users:** Identified customers who used discounts but still spent 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

- iii. **Top 5 Products by Rating:** Found products with highest average review ratings.

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

iv. Shipping type comparison: Average purchased amounts between Standard shipping and Express shipping

	shipping_type	round
	text	numeric
1	Standard	58.46
2	Express	60.48

v. Subscribers vs Non Subscribers: Compared average spend and total revenue across subscription status.

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

vi. Discount dependent products: Identified 5 products with the highest percentage of discounted purchases.

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

vii. Customer Segmentation: Classified customers into New, Returning, and Loyal segments based on purchase history.

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

viii. Top 3 Products per Category: Listed the most purchased products within each category.

	item_rank	category	item_purchased	total_orders
	bigint	text	text	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

ix. Repeat Buyers & Subscriptions: Checked whether customers with >5 purchases are more likely to subscribe

	subscription_status	repeat_buyers
	text	bigint
1	No	2518
2	Yes	958

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

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

5. Power BI Dashboard



6. Business Recommendations

- Targeted Marketing:** Direct marketing efforts towards high spending age groups and customers who frequently choose express delivery
- Strengthen loyalty:** Introduce or improve loyalty rewards to increase repeat purchases and grow loyal customer groups.
- Increase subscription:** Highlight the unique perks and added value customers get when they subscribe.
- Revisit Discount strategy:** Adjust discounts to support sales growth without reducing overall profit margins.
- Improve product visibility:** Feature high rated and best selling items more prominently in marketing and promotional campaigns.