

# Analysis of Customer Purchasing Behavior

## 1. Project Overview

This project focuses on analyzing retail customer purchase data to understand how consumers shop, what influences their buying decisions, and how businesses can use these insights to improve sales and customer loyalty. The main objective is to transform raw transactional data into meaningful insights that help management make smarter marketing, product, and customer-engagement decisions.

The study combines **Python for data preparation**, **SQL for structured analysis**, and **Power BI for visualization**, creating a complete end-to-end analytics workflow.

## 2. Business Objective

A retail company aims to better understand customer behavior across demographics, product types, and purchasing channels. Management wants to discover:

- What drives repeat purchases
- How discounts and promotions affect spending
- Which customer groups generate the most revenue
- How seasonal and product preferences influence buying decisions

The core question addressed in this project is:

**How can customer shopping data be used to identify trends, strengthen engagement, and optimize marketing and product strategies?**

## 3. Data Summary

The dataset represents several thousand customer purchase transactions and includes both demographic and behavioural information.

### Key Characteristics

- Customer Details: Age, Gender, Location, Subscription Status
- Purchase Details: Product Name, Category, Amount Spent, Season, Size, Color
- Behavior Indicators: Discount Usage, Purchase Frequency, Previous Purchases, Review Ratings, Shipping Type
- Minor missing values observed in review ratings

This combination of attributes allows both descriptive and predictive style analysis of consumer behavior.

## 4. Data Preparation and Exploration (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
<b>count</b>	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	39
<b>unique</b>	Nan	Nan	2	25	4	Nan	50	4	25	4	Nan	2	6	
<b>top</b>	Nan	Nan	Male	Blouse	Clothing	Nan	Montana	M	Olive	Spring	Nan	No	Free Shipping	
<b>freq</b>	Nan	Nan	2652	171	1737	Nan	96	1755	177	999	Nan	2847	675	22
<b>mean</b>	1950.500000	44.068462	Nan	Nan	Nan	59.764359	Nan	Nan	Nan	Nan	3.750065	Nan	Nan	Nan
<b>std</b>	1125.977353	15.207589	Nan	Nan	Nan	23.685392	Nan	Nan	Nan	Nan	0.716983	Nan	Nan	Nan
<b>min</b>	1.000000	18.000000	Nan	Nan	Nan	20.000000	Nan	Nan	Nan	Nan	2.500000	Nan	Nan	Nan
<b>25%</b>	975.750000	31.000000	Nan	Nan	Nan	39.000000	Nan	Nan	Nan	Nan	3.100000	Nan	Nan	Nan
<b>50%</b>	1950.500000	44.000000	Nan	Nan	Nan	60.000000	Nan	Nan	Nan	Nan	3.800000	Nan	Nan	Nan
<b>75%</b>	2925.250000	57.000000	Nan	Nan	Nan	81.000000	Nan	Nan	Nan	Nan	4.400000	Nan	Nan	Nan
<b>max</b>	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

- **Missing Data Handling:** Checked for null values and imputed missing values in the `Review Rating` column 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_applied` and `promo_code_used` were redundant; dropped `promo_code_used`.
- **Database Integration:** Connected Python script to PostgreSQL and loaded the cleaned Data Frame into the database for SQL analysis.

## 5. 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	75191
2	Male	157890

**2. High-Spending Discount Users** – Identified customers who used discounts but still spent above the average purchase amount.

	customer_id	purchase_amount
	bigint	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	00

Total rows: 839    Query complete 00:00:00

3. **Top 5 Products by Rating** – Found products with the 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

4. **Shipping Type Comparison** – Compared average purchase amounts between Standard and Express shipping.
5. **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

6. **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

7. **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

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

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

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

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

## 6. Dashboard in Power BI



## 7. Business Recommendations

- Boost Subscriptions** – Promote exclusive benefits for subscribers.
- Customer Loyalty Programs** – Reward repeat buyers to move them into the “Loyal” segment.
- Review Discount Policy** – Balance sales boosts with margin control.
- Product Positioning** – Highlight top-rated and best-selling products in campaigns.
- Targeted Marketing** – Focus efforts on high-revenue age groups and express-shipping users.

