

Customer Purchase Behavior Analytics

1. Project Summary

→ This project explores customer purchasing behavior using real-world transactional data. The analysis focuses on understanding how demographics, product choices, discounts, and subscription status influence spending patterns. Insights from this study can support data-driven decisions in marketing, pricing, and customer retention strategies.

2. Dataset Overview

- **Total records:** 3900 transaction
- **Total attributes:** 18 columns

Key Data Dimensions

- **Customer information:** age, gender, location, subscription status
- **Transaction details:** product name, category, purchase value, season, size, color
- **Behavioral signals:** discounts, promo usage, purchase frequency, review ratings, shipping method

Data quality note: A small number of **missing values** were found in the review ratings field and handled during preprocessing.

3. Data Cleaning & Preparation (Python)

→ The dataset was cleaned and prepared using **Python (Pandas, NumPy)** with the following steps:

- Verified data types and dataset structure

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Customer ID      3900 non-null   int64  
 1   Age               3900 non-null   int64  
 2   Gender            3900 non-null   object  
 3   Item Purchased    3900 non-null   object  
 4   Category          3900 non-null   object  
 5   Purchase Amount (USD) 3900 non-null   int64  
 6   Location          3900 non-null   object  
 7   Size               3900 non-null   object  
 8   Color              3900 non-null   object  
 9   Season             3900 non-null   object  
 10  Review Rating     3863 non-null   float64 
 11  Subscription Status 3900 non-null   object  
 12  Shipping Type     3900 non-null   object  
 13  Discount Applied   3900 non-null   object  
 14  Promo Code Used    3900 non-null   object  
 15  Previous Purchases 3900 non-null   int64  
 16  Payment Method     3900 non-null   object  
 17  Frequency of Purchases 3900 non-null   object  
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

- Handled missing review ratings using category-level statistics

```

Customer ID      0
Age              0
Gender           0
Item Purchased   0
Category          0
Purchase Amount (USD) 0
Location          0
Size              0
Color              0
Season             0
Review Rating     37
Subscription Status 0
Shipping Type     0
Discount Applied   0
Promo Code Used   0
Previous Purchases 0
Payment Method     0
Frequency of Purchases 0
dtype: int64

```

- Standardized column names for consistency

```

Index(['customer_id', 'age', 'gender', 'item_purchased', 'category',
       'purchase_amount', 'location', 'size', 'color', 'season',
       'review_rating', 'subscription_status', 'shipping_type',
       'discount_applied', 'promo_code_used', 'previous_purchases',
       'payment_method', 'frequency_of_purchases'],
      dtype='object')

```

- Created new analytical features:

- Age groups** to enable demographic comparisons

	age	age_group
0	55	Middle-aged
1	19	Young Adult
2	50	Middle-aged
3	21	Young Adult
4	45	Middle-aged

- Purchase frequency** metrics derived from transaction history

frequency_of_purchases	
Every 3 Months	584
Annually	572
Quarterly	563
Monthly	553
Bi-Weekly	547
Fortnightly	542
Weekly	539
Name: count, dtype: int64	

- Removed redundant fields after validating overlapping information

	discount_applied	promo_code_used
0	Yes	Yes
1	Yes	Yes
2	Yes	Yes
3	Yes	Yes
4	Yes	Yes

- Exported the cleaned dataset to **PostgreSQL** for structured querying

4. Analytical Questions & SQL Insights

→ Using PostgreSQL, several business-focused questions were explored:

A. Spending patterns by gender

→ Compared total revenue contribution across genders.

	gender	total_revenue
1	Female	75191
2	Male	157890

B. High-value customers using discounts

→ Identified customers who still spent above-average amounts despite discounts.

	customer_id	purchase_amount
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
12	29	94
13	22	70

Total rows: 839 | Query complete 00:00:01.023

C. Top-rated products

→ Ranked items based on average customer review scores.

	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

D. Shipping method vs. spending behavior

→ Analyzed differences in average order value between standard and express shipping.

	shipping_type	Average Purchase AMount
text		numeric
1	Standard	58.46
2	Express	60.48

E. Subscription impact on revenue

→ Compared customer count, average spend, and total revenue for subscribers vs non-subscribers.

	subscription_status	Total Customer	Average Spend	Total Revenue
text		bigint	numeric	numeric
1	No	2847	59.8651211801896733	170436
2	Yes	1053	59.4919278252611586	62645

F. Products most dependent on discounts

→ Highlighted items with the highest share of discounted purchases.

	item_purchased text	Discount Rate numeric
1	Hat	50.00
2	Sneakers	49.00
3	Coat	49.00
4	Sweater	48.00
5	Pants	47.00
6	Boots	46.00
7	Jeans	45.00
8	Dress	45.00
9	Hoodie	45.00
10	Belt	44.00

G. Customer segmentation

→ Grouped customers into **New**, **Returning**, and **Loyal** segments based on purchase history.

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

H. Best-selling products by category

→ Identified top-performing products within each product 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

I. Repeat buyers and subscription likelihood

→ Checked whether frequent buyers were more likely to subscribe.

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

J. Revenue contribution by age group

→ Measured which age segments generated the most revenue.

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

5. Visualization & Dashboard (Power BI)

→ An interactive Power BI dashboard was built to present findings clearly and intuitively.



Dashboard Highlights

- Displays **3.9K total customers, \$59.76 average purchase amount, and 3.75 average review rating**
- Visualizes **revenue and sales volume by product category**, highlighting top-performing categories
- Shows **subscription status distribution**, comparing subscribed vs non-subscribed customers
- Analyzes **seasonal purchasing patterns**, comparing total purchase amounts and order volume across seasons
- Includes slicers for **subscription status, gender, category, and shipping type** to enable drill-down analysis

The dashboard supports interactive exploration, allowing users to identify trends across multiple customer and transaction dimensions.

6. Key Business Insights

- Non-subscribed customers make up the majority of the customer base, indicating an opportunity for subscription growth
- Clothing and Accessories generate the highest revenue and sales volume among all categories
- Purchase activity is relatively consistent across seasons, with slightly higher spending observed during Fall and Spring
- Customers using faster shipping options tend to exhibit higher average purchase amounts
- Strong repeat purchasing behavior suggests a stable and loyal customer segment

7. Business Recommendations

- Expand subscription offerings with clear value propositions to increase adoption
- Introduce targeted loyalty programs to convert repeat customers into long-term subscribers
- Review discount dependency for certain products to ensure margin sustainability
- Focus marketing efforts on high-performing categories such as Clothing and Accessories
- Leverage seasonal and demographic insights to design more personalized marketing campaigns

8. Tools & Technology

- **Python:** Pandas, NumPy
- **SQL:** PostgreSQL
- **BI Tool:** Power BI
- **Environment:** Jupyter Notebook