

Customer Shopping Behavior Analysis

Project Overview

This project focuses on understanding customer shopping behavior using transactional data from 3,900 purchases across multiple product categories. The objective is to identify spending patterns, key customer groups, subscription activity, and product interests. These insights help shape decisions around marketing, customer retention, and product strategy.

Dataset Summary

The dataset contains information on 3,900 transactions with 18 features. It includes:

- **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 missing entries found in the Review Rating column

This data gives a clear view of how customers shop, what drives their decisions, and which factors influence repeat purchases.

Exploratory Data Analysis using Python

We started the analysis by preparing and cleaning the dataset in Python.

- **Data Loading:** Imported the dataset using pandas.
- **Initial Exploration:** Used df.info() to check structure and .describe() for summary statistics.

```
<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
```

[12]:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900
unique	Nan	Nan	2	25	4	Nan	50	4	25	4
top	Nan	Nan	Male	Blouse	Clothing	Nan	Montana	M	Olive	Spring
freq	Nan	Nan	2652	171	1737	Nan	96	1755	177	999
mean	1950.500000	44.068462	Nan	Nan	Nan	59.764359	Nan	Nan	Nan	Nan
std	1125.977353	15.207589	Nan	Nan	Nan	23.685392	Nan	Nan	Nan	Nan
min	1.000000	18.000000	Nan	Nan	Nan	20.000000	Nan	Nan	Nan	Nan
25%	975.750000	31.000000	Nan	Nan	Nan	39.000000	Nan	Nan	Nan	Nan
50%	1950.500000	44.000000	Nan	Nan	Nan	60.000000	Nan	Nan	Nan	Nan
75%	2925.250000	57.000000	Nan	Nan	Nan	81.000000	Nan	Nan	Nan	Nan
max	3900.000000	70.000000	Nan	Nan	Nan	100.000000	Nan	Nan	Nan	Nan

Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3863.000000	3900	3900	3900	3900	3900.000000	3900	3900
Nan	2	6	2	2	Nan	6	7
Nan	No	Free Shipping	No	No	Nan	PayPal	Every 3 Months
Nan	2847	675	2223	2223	Nan	677	584
3.750065	Nan	Nan	Nan	Nan	25.351538	Nan	Nan
0.716983	Nan	Nan	Nan	Nan	14.447125	Nan	Nan
2.500000	Nan	Nan	Nan	Nan	1.000000	Nan	Nan
3.100000	Nan	Nan	Nan	Nan	13.000000	Nan	Nan
3.800000	Nan	Nan	Nan	Nan	25.000000	Nan	Nan
4.400000	Nan	Nan	Nan	Nan	38.000000	Nan	Nan
5.000000	Nan	Nan	Nan	Nan	50.000000	Nan	Nan

- **Missing Data Handling:** Checked for null values and imputed missing entries in the Review Rating column by using the median rating of each product category. This helped maintain realistic review patterns without introducing bias.
- **Column Standardization:** Renamed all column names into **snake case** to make them consistent, easier to reference in code, and more readable for documentation.
- **Feature Engineering:**
 - Created an **age_group** field by grouping customer ages into categories (example: **Young Adult**, **Middle-aged**, **Senior**).
 - Generated a **purchase_frequency_days** field to measure how often a customer buys, based on timestamps and previous purchase records.
- **Data Consistency Check:** Reviewed **discount_applied** and **promo_code_used** to see if both were delivering the same signal. Since the **promo_code column** provided similar information and caused duplication, it was removed.
- **Database Integration:** The cleaned DataFrame was connected to PostgreSQL using SQLAlchemy and loaded as a dedicated database table. This allowed structured SQL queries to simulate business transactions and extract segment-based insights.

Data Analysis using SQL (Business Transactions)

As mentioned above, after preparing the data in Python, the dataset was loaded into PostgreSQL to simulate real business transactions and run structured queries. The goal of this step was to extract insights that directly support decisions around marketing, pricing, retention, and product strategy.

1. **Revenue by Gender:** Compared total revenue generated by male and female customers to understand spending differences across demographics.

	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, highlighting opportunities for premium upsell campaigns.

Data Output Messages Notifications

The screenshot shows a PostgreSQL client interface with a toolbar at the top containing icons for Data Output, Messages, Notifications, and various file operations. Below the toolbar is a table with two columns: 'customer_id' and 'purchase_amount'. The table has 13 rows, each containing a unique customer ID and their corresponding purchase amount. At the bottom of the interface, there is a status bar displaying 'Total rows: 839' and 'Query complete 00:00:00.127'.

	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
12	29	94
13	32	79

Total rows: 839 Query complete 00:00:00.127

3. **Top 5 Products by Rating:** Ranked products based on average review scores to understand which items drive strong customer satisfaction.

Data Output Messages Notifications

A screenshot of a data output interface. At the top, there are tabs for "Data Output", "Messages", and "Notifications". Below the tabs is a toolbar with various icons. The main area displays a table with two columns: "item_purchased" (text) and "Average Product Rating" (numeric). The table has 5 rows, each representing a product with its name and its average rating. The products listed are Gloves (3.86), Sandals (3.84), Boots (3.82), Hat (3.80), and Skirt (3.78).

	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. **Shipping Type Comparison:** Compared average purchase amounts between Standard and Express shipping users to see whether faster delivery correlates with higher spending.

Data Output Messages Notifications

A screenshot of a data output interface. At the top, there are tabs for "Data Output", "Messages", and "Notifications". Below the tabs is a toolbar with various icons. The main area displays a table with two columns: "shipping_type" (text) and "round" (numeric). The table has 2 rows, each representing a shipping type and its average purchase amount. The shipping types listed are Standard (58.46) and Express (60.48).

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. **Subscribers vs. Non-Subscribers:** Analyzed how subscription status affects revenue by comparing total spend and average purchase amounts between both groups.

Data Output Messages Notifications

A screenshot of a data output interface. At the top, there are tabs for "Data Output", "Messages", and "Notifications". Below the tabs is a toolbar with various icons. The main area displays a table with four columns: "subscription_status" (text), "total_customers" (bigint), "avg_spend" (numeric), and "total_revenue" (numeric). The table has 2 rows, each representing a subscription status and its corresponding metrics. The subscription statuses listed are Yes (1053 customers, avg spend 59.49, total revenue 62645.00) and No (2847 customers, avg spend 59.87, total revenue 170436.00).

	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. **Discount-Dependent Products:** Identified five products that are most frequently purchased with discounts, helping reveal which items may rely heavily on price promotions.

Data Output Messages Notifications

The screenshot shows a software interface for managing data. At the top, there are tabs for "Data Output", "Messages", and "Notifications". Below the tabs is a toolbar with various icons for file operations like creating, opening, saving, and deleting files, as well as a SQL editor icon. The main area displays a table with two columns: "item_purchased" (text) and "discount_rate" (numeric). The table contains five rows of data, each representing a product and its corresponding discount rate.

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

7. **Customer Segmentation:** Grouped customers into New, Returning, and Loyal segments based on their purchase counts to support targeted retention strategies.

Data Output Messages Notifications

The screenshot shows a software interface for managing data. At the top, there are tabs for "Data Output", "Messages", and "Notifications". Below the tabs is a toolbar with various icons for file operations like creating, opening, saving, and deleting files, as well as a SQL editor icon. The main area displays a table with two columns: "customer_segment" (text) and "Number of Customers" (bigint). The table contains three rows of data, each representing a segment and its count of customers.

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

8. **Top 3 Products per Category:** Listed the most purchased products within each category to understand category-level favorites and demand concentration.

	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

9. **Repeat Buyers and Subscription Relationship:** Checked whether customers with more than five purchases show higher likelihood of subscribing, helping guide loyalty program strategy.

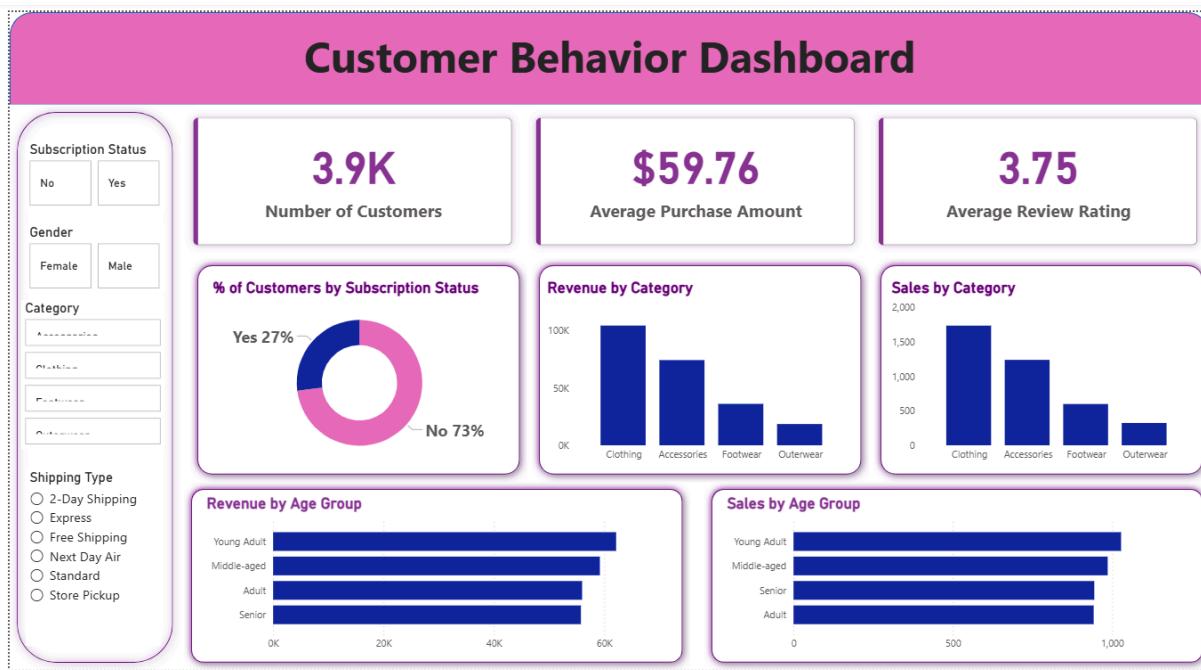
	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

10. **Revenue by Age Group:** Calculated total revenue contributed by each age group to understand which demographic segments drive the most value.

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

Dashboard in Power BI

Finally, an interactive dashboard was built in Power BI to present the insights visually, allowing stakeholders to explore trends, compare customer groups, and quickly understand which factors influence spending and repeat purchases.



Business Recommendations

Based on the insights generated from Python, SQL analysis, and the Power BI dashboard, these actions are recommended to support growth in revenue, retention, and customer engagement:

- **Boost Subscriptions:** Strengthen subscription adoption by promoting clear benefits, such as free shipping, early access, or personalized offers.
- **Customer Loyalty Programs:** Introduce rewards or tiered perks to encourage repeat buyers and help transition shoppers into the Loyal segment.
- **Review Discount Policy:** Evaluate products that rely heavily on discounts and consider controlled promotions to maintain profitability without losing demand.
- **Product Positioning:** Feature top-rated and most-purchased products in marketing campaigns, product pages, and homepage placements to maximize exposure.
- **Targeted Marketing Efforts:** Focus campaigns on age groups that generate the highest revenue and customers who prefer Express shipping, as they tend to spend more.