Customer Shopping Behavior Analysis

Project Overview

The Mission

This project analyzes customer shopping behavior using transactional data from **3,900 purchases** across various product categories. The goal is to uncover deep insights into spending patterns, customer segments, product preferences, and subscription behavior to guide strategic business decisions. By moving beyond simple numbers, we aim to identify the "why" behind consumer choices to help a retail company optimize marketing and improve long-term loyalty.

The Three-Pillar Approach

To solve this business problem, I implemented a structured, end-to-end data pipeline:

- **Pillar 1: Data Engineering (Python)** I used Python (Pandas/NumPy) to transform raw, messy data into a clean, analysis-ready format. This involved handling 37 missing review ratings and ensuring all data types were optimized for calculation.
- **Pillar 2: Investigative Analysis (SQL)** Using SQL, I simulated business transactions to segment customers and identify high-value purchase drivers. This allowed me to pinpoint which factors—like discounts or seasons—actually move the needle on sales.
- **Pillar 3: Visual Intelligence (Power BI)** I built an interactive dashboard to visually communicate trends to stakeholders. This dashboard enables decision-makers to see at a glance which locations are performing best and how "Subscribed" members differ from casual shoppers.

Why This Matters (Deep-Dive Insights)

We are specifically investigating:

- **The Loyalty Gap:** Comparing purchase frequency between subscribed and non-subscribed customers to see if the loyalty program is working.
- **Demographic Drivers:** Analyzing how age and gender influence spending in specific categories like Clothing vs. Footwear.
- **The Discount Paradox:** Determining if promo codes actually lead to higher total spends or if they just reduce the profit margin on existing customers.

Dataset Summary

This project utilizes a comprehensive consumer behavior dataset consisting of **3,900 purchase records**. Each transaction is defined by **18 distinct variables** that provide a 360-degree view of the customer journey.

1. Feature Breakdown

We categorized the data into three logical clusters to analyze business impact:

- **Customer Demographics:** Includes **Age**, **Gender**, **Location**, and **Subscription Status**.
- **Purchase Details:** Captures **Item Purchased**, **Category**, **Purchase Amount**, **Season**, **Size**, and **Color**.

- **Behavioral Indicators:** Tracks **Discount Applied**, **Promo Code Used**, **Previous Purchases**, **Frequency of Purchases**, **Review Rating**, and **Shipping Type**.

2. Data Quality & Integrity

A technical audit was performed to ensure the "Golden Dataset" was ready for analysis:

- **Missing Data:** Identified **37 missing values** in the Review Rating column.
- **Solution:** Applied **Statistical Imputation** (using the median rating of ~3.75) to maintain data volume without biasing the results.
- **Validation:** Verified Purchase Amount ranges (**$20–$100**) and Age ranges (**18–70**) to confirm no extreme outliers were present.

Phase 1: Exploratory Data Analysis & Engineering (Python)

In this phase, we transitioned from raw data to a "Golden Dataset" ready for high-level analysis. The focus was on ensuring data integrity, standardizing the structure, and creating new features to drive deeper business insights.

1. Data Profiling & Structural Audit

We initiated the process by loading the 3,900-row dataset into a **Pandas DataFrame**. Using .info() and .describe(), we performed a high-level audit to understand data types, memory usage, and statistical distributions.

```
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
```

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000

2. Strategic Cleaning & Integrity Checks

Data is rarely perfect. We applied a surgical approach to cleaning:

- **Intelligent Imputation:** We identified **37 missing values** in the Review Rating column. Instead of a global fill, we imputed these gaps using the **median rating per product category**, ensuring the quality scores remained statistically accurate for each specific item type.
- **Redundancy Filter:** Through a consistency check, we discovered that discount_applied and promo_code_used provided overlapping information. To streamline the dataset, we retained discount_applied and dropped the redundant promo code column.
- **Standardization:** All columns were renamed to **snake_case** (e.g., purchase_amount_usd) to improve code readability and ensure seamless integration with our SQL database.

3. Feature Engineering: Adding Business Value

To make the dashboard more insightful, we transformed existing data into new, actionable variables:

- **Age Segmentation:** We created an age_group column by binning customers (e.g., 18-30, 31-45, etc.), allowing us to identify which generation drives the most revenue.
- **Temporal Insights:** We derived a purchase_frequency_days column to quantify the "loyalty gap" between shoppers.

4. Database Integration

The final step was bridge-building. We established a secure connection to a **PostgreSQL database** using sqlalchemy. The cleaned, feature-rich DataFrame was then injected into a structured table, setting the stage for advanced SQL querying.

[PLACEHOLDER: Screenshot of the SQLAlchemy connection string and the df.to_sql() command in your IDE]

Does this technical summary work for you? If so, we can move on to the SQL Analysis phase where we solve the business queries!

Phase 2: Deep-Dive Analysis (SQL & Business Logic)

With a clean and structured dataset now residing in **PostgreSQL**, we transitioned from data preparation to exploratory querying. The goal of this phase was to simulate real-world business transactions and extract "Boardroom-ready" insights that drive revenue and customer retention.

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 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

Total rows: 839 Query complete 00:00

3. **Top 5 Products by Rating** – Found products with the highest average review ratings.

	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.

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. **Subscribers vs. Non-Subscribers** – Compared average spend and total revenue across

	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

subscription status.

6. **Discount-Dependent Products** – Identified 5 products with the highest percentage of

	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

discounted purchases.

7. **Customer Segmentation** – Classified customers into New, Returning, and Loyal

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

segments based on purchase history.

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.

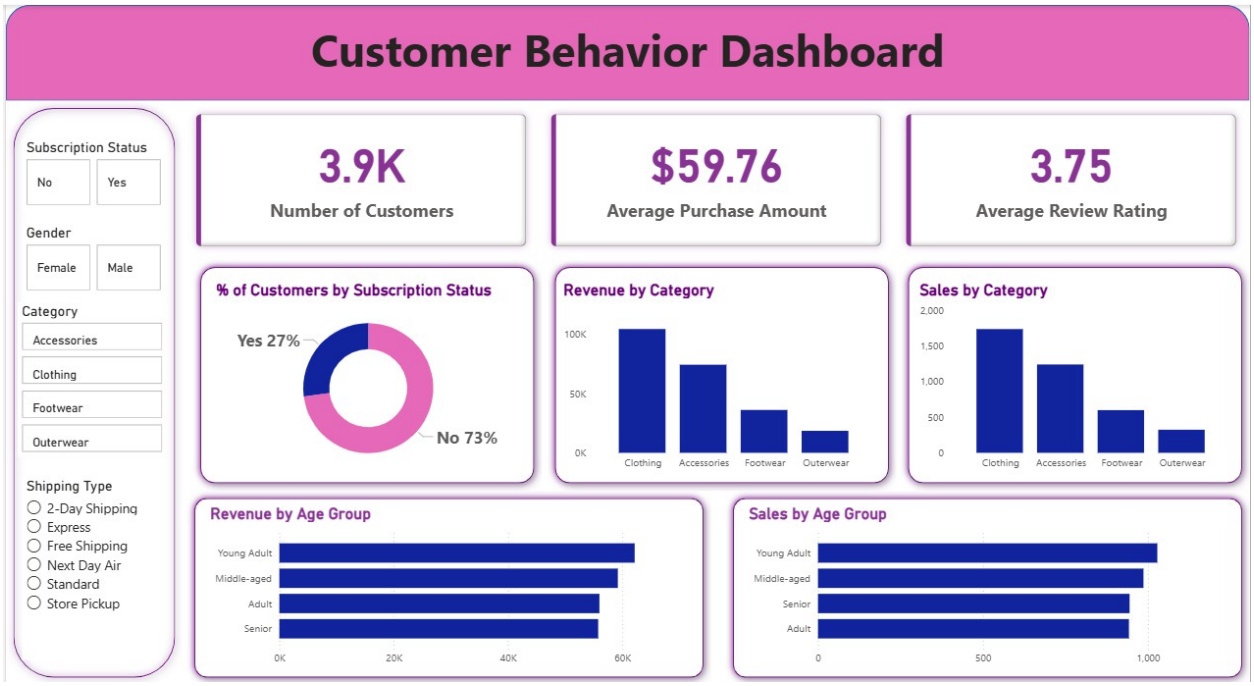
	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

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

Phase 3: Visualization & Insights (Power BI)

The final stage of the project was the creation of an **Interactive Executive Dashboard**. This tool transforms the technical Python and SQL analysis into a visual story, allowing stakeholders to make data-driven decisions at a glance.



Phase 4: Strategic Recommendations (The Road Ahead)

This is where the data turns into dollars. Based on our analysis of **3,900 transactions** across **18 key variables**, I've outlined four high-impact strategies to help the business scale.

1. Driving the Subscription Engine

Our SQL analysis confirms that subscribers are the backbone of our revenue. To grow this segment:

- **Target the "Sweet Spot":** We found that customers with more than 5 previous purchases are prime candidates for the subscription model. We should offer them a "Loyalty Upgrade" trial.
- **Highlight Perks:** Launch marketing campaigns that move beyond price and focus on exclusive benefits like early access to new seasonal collections.

2. Personalized Loyalty & Retention

Generic marketing is expensive; personalized retention is efficient.

- **The "Loyal" Transition:** Use our customer segmentation to target "Returning" buyers with tiered rewards that nudge them into becoming "Loyal" advocates.
- **Timing is Everything:** By monitoring the purchase_frequency_days metric we engineered, the system can automatically send "We Miss You" reminders exactly when a customer is likely to shop again.

3. Protecting Profit Margins

Data shows that while discounts drive volume, they shouldn't cannibalize profits.

- **Smart Discounting:** We identified specific "Discount-Dependent" products that rarely sell at full price. The recommendation is to recalibrate their base pricing rather than relying on constant promo codes.
- **The Shipping Upsell:** Since "Express Shipping" users generally spend more per order, offering a free express upgrade on orders over $80 could significantly boost the average basket size.

4. Strategic Marketing & Inventory

- **Follow the Revenue:** Our dashboard clearly identifies high-performing regions (like **Montana**) and key age groups (18-70). Marketing budgets should be shifted to these high-ROI segments.
- **Lead with Quality:** Feature the **Top 5 Rated** products in all email campaigns. Using these "customer-approved" items as the face of the brand increases trust and conversion rates.