Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes 3,900 customer transactions across multiple product categories to understand purchase behavior, loyalty, and revenue drivers. The goal is to provide actionable business insights and help the company optimize marketing, product strategies, and customer engagement.

The analysis combines Python (Jupyter Notebook), PostgreSQL, and Power BI, with feature engineering, RFM scoring, segmentation, and statistical testing to uncover patterns beyond basic descriptive analytics.

2. Dataset Summary

- Rows / Columns: 3,900 / 18
- Key Features:
    - 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 values in Review Rating; imputed with category-wise median.

3. Exploratory Data Analysis & Feature Engineering

- **Data Loading:** Imported the dataset using pandas and viewed the first few rows with df.head().

	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	Pay M
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	V
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	

Payment Method	Frequency of Purchases
Venmo	Fortnightly
Cash	Fortnightly
Credit Card	Weekly
PayPal	Weekly
PayPal	Annually

- **Initial Exploration:** Used df.info() and df.describe() to check structure, data types, and summary statistics.
- **Missing Data Handling:** Found 37 missing values in Review Rating and imputed them with category-wise median.

```
df['Review Rating'] = df.groupby('Category')['Review Rating'].transform(lambda x: x.fillna(x.median()))
#Median review rating within each category instead of overall median
```

```
df.isnull().sum()
```

```
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             0
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
```

- **Column Standardization:** Converted column names to snake_case for readability and renamed purchase_amount_(USD) to purchase_amount.

Feature Engineering:

- Created age_group by binning ages: Young Adult, Adult, Middle-aged, Senior.

	age	age_group
0	55	Middle-aged
1	19	Young Adult
2	50	Middle-aged
3	21	Young Adult
4	45	Middle-aged
5	46	Middle-aged
6	63	Senior
7	27	Young Adult
8	26	Young Adult
9	57	Middle-aged

- Converted textual frequency of purchases into numeric days (purchase_frequency_days).

	purchase_frequency_days	frequency_of_purchases
0	14	Fortnightly
1	14	Fortnightly
2	7	Weekly
3	7	Weekly
4	365	Annually
5	7	Weekly
6	90	Quarterly
7	7	Weekly
8	365	Annually
9	90	Quarterly

- Verified discount_applied and promo_code_used were redundant; dropped promo_code_used.

Advanced Analysis:

- Calculated RFM scores (Recency, Frequency, Monetary) per customer.
- **Generated customer segments:** Champions, Loyal Customers, Big Spenders, Potential Loyalists, At Risk.

	customer_segment	total_customers	total_revenue	avg_spend	revenue_share_%
1	Big Spenders	921	77952	84.638436	33.444167
4	Potential Loyalist	1211	52733	43.545004	22.624324
2	Champions	866	51412	59.367206	22.057568
3	Loyal Customers	683	41572	60.866764	17.835860
0	At Risk	219	9412	42.977169	4.038081

- Conducted T-test to compare purchase amounts of subscribers vs non-subscribers.

```
T-stat: -0.4368012420060051
P-value: 0.6622796924526246
```

- **Database Integration:** Loaded the cleaned and feature-engineered DataFrame into PostgreSQL for SQL-based analysis.

4. SQL Analysis (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

Total rows: 839 | Query complete 00:00:00.466

3. **Top Products by Review:** 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 Comparison:** Compared average purchase amounts between Standard and Express shipping.

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. **Subscriber Behavior:** Compared average spend and total revenue across subscription status.

	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-Sensitive Products:** Identified 5 products with the highest percentage 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

7. **Customer Segmentation (SQL):** Classified 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

8. **Top Products per Category:** Listed the most purchased products within each 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

9. **Repeat Buyers & Subscription:** Checked whether customers with >5 purchases are more likely to subscribe.

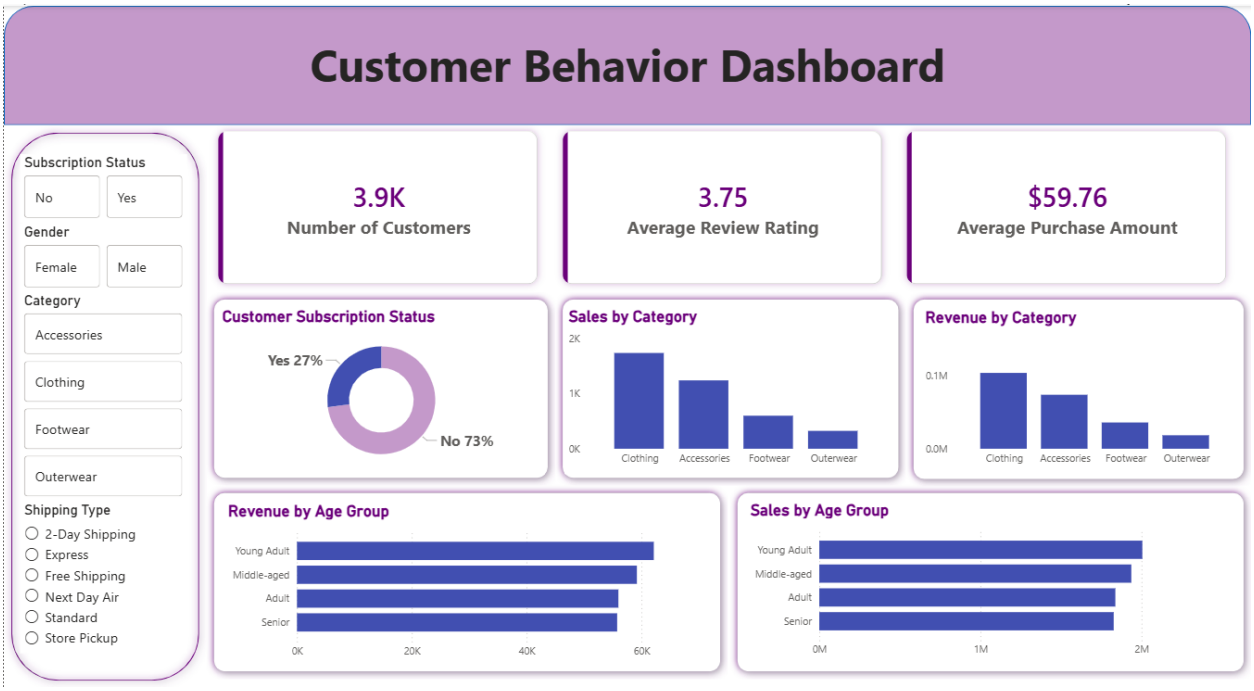
	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

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

5. Power BI Dashboard

- Finally, we built an interactive dashboard in Power BI to present insights visually.
- Enabled real-time insights for business decision-making.



6. Key Insights

- **High-Value Segments:** Big Spenders & Champions drive most revenue; Potential Loyalists = growth opportunity.
- **Age & Gender:** Young Adults & Middle-aged are top contributors; male/female spending similar.
- **Discount Behavior:** Top discounted products: Hat, Sneakers, Coat; some customers spend above average even with discounts.
- **Subscription Impact:** No significant difference in spend between subscribers and non-subscribers (T-test $p>0.05$).
- **Product Preferences:** Top-rated: Gloves, Sandals, Boots; Most purchased: Blouse, Pants, Sandals, Jacket, Jewelry.
- **Actionable Recommendations:** Boost subscriptions & loyalty programs, optimize discounts, highlight high-demand products, target high-revenue age groups.

7. Business Recommendations

1. **Boost Subscriptions:** Convert repeat buyers to subscribers with targeted offers.
2. **Loyalty Programs:** Reward Champions and Big Spenders to retain high-value customers.
3. **Optimize Discounts:** Focus promotions on items where discounts drive actual incremental revenue.
4. **Product Marketing:** Highlight high-performing and top-rated products.
5. **Targeted Marketing:** Personalize campaigns using age, gender, and purchase patterns.
6. **Retention Strategies:** Engage At-Risk customers identified via RFM to reduce churn.