Customer Shopping Behaviour Analysis

Python (Pandas) | SQL (MySQL) | Power BI

1. Project Overview

This project analyzes customer shopping behavior to understand spending patterns, the impact of subscriptions and discounts, and customer loyalty. The goal is to identify high-value customer segments and suggest data-driven business improvements.

2. Analysis Workflow

- Data cleaning and exploration using Python

- Business analysis using SQL

- Visualization and storytelling using Power BI

3. Exploratory Data Analysis using 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	Discou Appli
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	39
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	22
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	Na
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	Na
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	Na
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	Na
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	Na
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	Na
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	Na

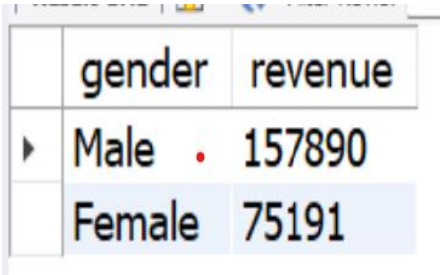
Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3900	3900	3900.000000	3900	3900
2	2	NaN	6	7
No	No	NaN	PayPal	Every 3 Months
2223	2223	NaN	677	584
NaN	NaN	25.351538	NaN	NaN
NaN	NaN	14.447125	NaN	NaN
NaN	NaN	1.000000	NaN	NaN
NaN	NaN	13.000000	NaN	NaN
NaN	NaN	25.000000	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 Ratingcolumn 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_appliedand promo_code_used were redundant; dropped promo_code_used.
- **Database Integration:** Connected Python script to MySQL and loaded the cleaned DataFrame into the database for SQL analysis.

Data Analysis using SQL

We performed structured analysis in SQL to answer key business questions:

1. **Revenue by Gender** – Compared total revenue generated by male vs. female customers.



gender	revenue
Male	157890
Female	75191

Insight:

- Male customers generate more than twice the total revenue compared to female customers.

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

customer_id	purchase_amount
1411	93
1413	100
1414	79
1416	90
1419	75
1420	75
1422	100
1424	60
1428	90
1429	97
1432	88
1434	60
1435	64
1436	95

Insight:

- Some customers use discounts but still spend above the average purchase amount.

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

item_purchased	Average Product Rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Skirt	3.78

Insight:

- Product ratings are clustered between ~3.7 and ~3.9.
- No product shows extremely high ratings, indicating moderate customer satisfaction overall.

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

Shipping_type	Average Purchase Amount
Express	60.48
Standard	58.46

Insight:
- Express shipping users spend slightly more on average than standard shipping users.

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

subscription_status	total_customers	avg_spend	total_revenue
Yes	1053	59.5	62645
No	2847	59.9	170436

Insight:
- Average spend is almost the same for subscribers and non-subscribers.
- Non-subscribers generate higher total revenue due to larger customer volume.

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

item_purchased	discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

7. **Customer Segmentation** – Classified customers into New, Returning, and Loyal segments based on purchase history.

customer_segmentation	no_of_customers
Loyal	3116
Returning	701
New	83

Insight:
- Majority of customers are classified as Loyal.

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

subscription_status	repeated_customers
Yes	958
No	2518

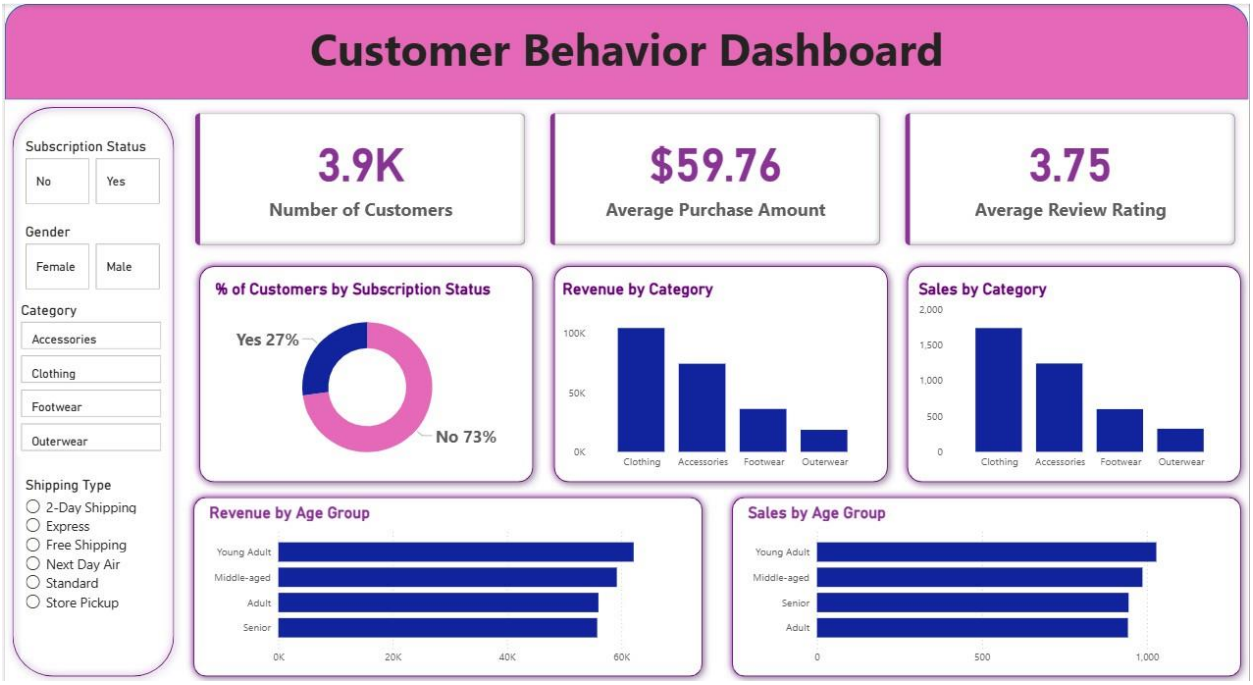
Insight:
- Most repeat buyers are not subscribed.

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

age_group	revenue
Young Adult	62143
Middle-aged	59197
Adult	55978
Senior	55763

Insight:
- Revenue is fairly evenly distributed across age groups.
- Young Adults contribute the highest revenue

Dashboard in Power BI



Key Business Insights

- Revenue is gender-skewed, with **male customers contributing significantly more**.
- Several products are highly dependent on discounts, posing margin risks.
- Customer base shows strong loyalty but weak new customer acquisition.
- Revenue contribution is balanced across age groups.

Business Recommendations

- Target repeat buyers for subscription conversion to increase long-term value.
- Use discounts selectively for price-sensitive products to protect margins.
- Improve acquisition strategies to grow the new customer segment.
- Focus marketing efforts on high-performing customer segments without over-relying on discounts.