Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior using transactional data from 3,900 purchases across various product categories. The goal is to uncover insights into spending patterns, customer segments, product preferences, and subscription behavior to guide strategic business decisions.

2. Dataset Summary

- Rows: 3,900 - Columns: 18
- Key Features:

  - 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 values in Review Rating column

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	N
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	N
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	N
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	N
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	N
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	N
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	N

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:** Checkedfor null values and imputed missing values in the `Review Rating` column using themedian 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_applied and 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.

4. Data Analysis using SQL (Business Transactions)

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

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

gender	total_revenue
Male	157890
Female	75191

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

customer_id	discount_appli...	purchase_amount
2	Yes	64
3	Yes	73
4	Yes	90
7	Yes	85
9	Yes	97
12	Yes	68
13	Yes	72
16	Yes	81
20	Yes	90
22	Yes	62
24	Yes	88
29	Yes	94
32	Yes	79
33	Yes	67
35	Yes	91
37	Yes	69

stomer_behavior_data 3

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

item_purchas...	avg_review_rating
Gloves	3.86142857142857 25
Sandals	3.8443750000000003
Boots	3.8187500000000005
Hat	3.80129870129870 05
Skirt	3.784810126582278

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

shipping_ty...	avg_purchase_amo...
Express	60.4752
Standard	58.4602

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

total_custom...	subscription_sta...	avg_spend	total_revenue
1053	Yes	59.4919	62645
2847	No	59.8651	170436

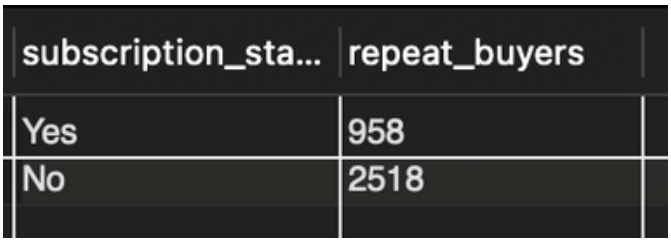
6. **Discount-Dependent Products** – Identified 5 products with the highest percentage of discounted purchases.

item_purchas...	discount_ra...
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.

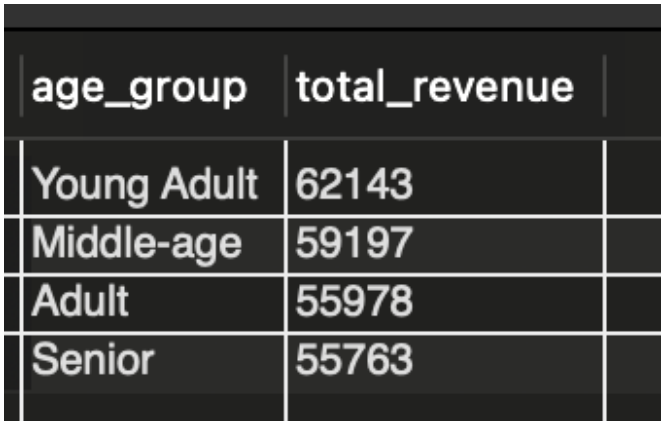
COUNT(customer_i...	customer_segm...
3116	Loyal
701	Returning
83	New

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



subscription_sta...	repeat_buyers
Yes	958
No	2518

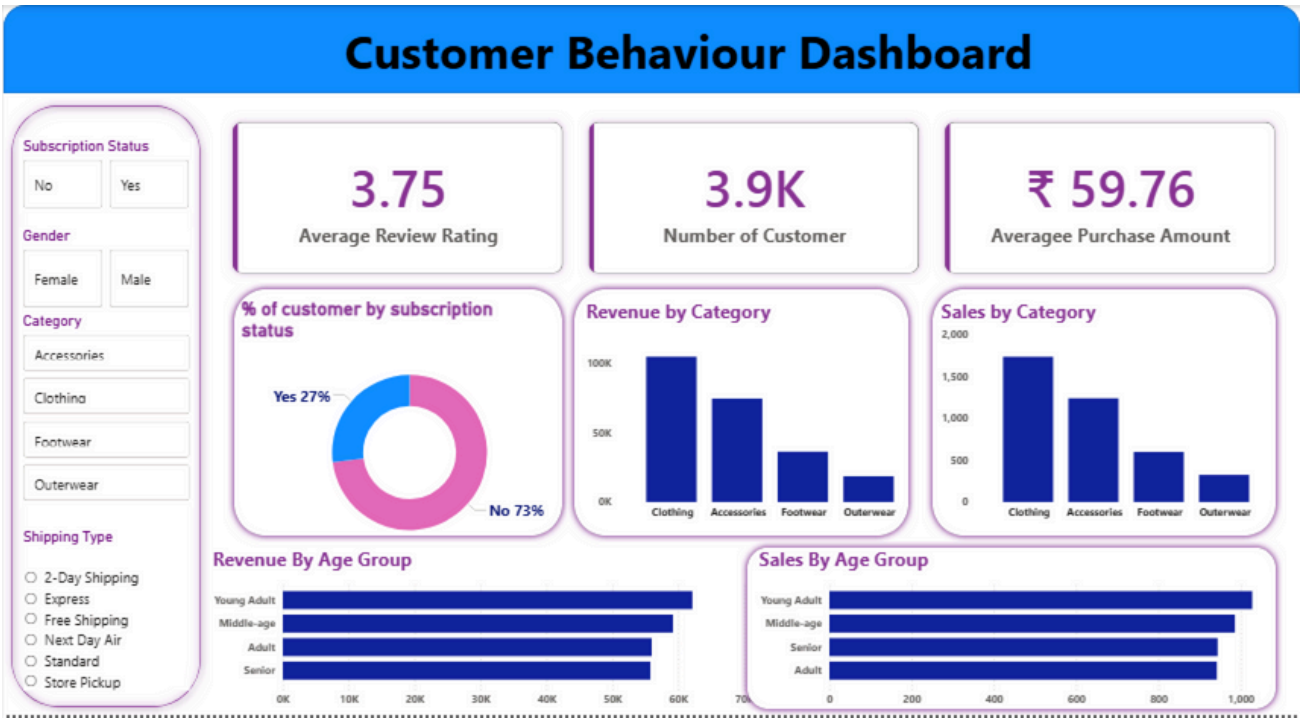
9. **Revenue by Age Group** – Calculated total revenue contribution ofeach age group.



age_group	total_revenue
Young Adult	62143
Middle-age	59197
Adult	55978
Senior	55763

5. Dashboard in Power BI

Finally, we built an interactive dashboard in **Power BI** to present insights visually.



6. Business Recommendations

- **Boost Subscriptions** – Promote exclusive benefits for subscribers.
- **Customer Loyalty Programs** – Reward repeat buyers to move them into the “Loyal” segment.
- **Review Discount Policy** – Balance sales boosts with margin control.
- **Product Positioning** – Highlight top-rated and best-selling products in campaigns.
- **Targeted Marketing** – Focus efforts on high-revenue age groups and express-shipping users.