Customer Shopping Behavior Analysis

1. Project Overview:

The project analyzes customer shopping behavior using transitional data from 3,900 purchase across various product categories. The goal 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 Purchase, 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 and cleaning in Pandas.
- **Initial Exploration:** Used df.info () to check structure and .describe() for summary statistics.

	customer_id	age	purchase_amount	review_rating	previous_purchases	purchase_frequency_days
count	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000	3316.000000
mean	1950.500000	44.068462	59.764359	3.750051	25.351538	88.980398
std	1125.977353	15.207589	23.685392	0.713590	14.447125	129.097211
min	1.000000	18.000000	20.000000	2.500000	1.000000	7.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000	14.000000
50%	1950.500000	44.000000	60.000000	3.800000	25.000000	30.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000	90.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000	365.000000

- **Missing Data Handling:** Checked for null values and imputed missing values in the Review Rating column using the median rating of each product category.
- **Column Standardization:** Rename 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.
- **Data Integration:** Connected Python script to MySQL and loaded the cleaned DataFrame into the databases for SQL analysis.

4. Data Analysis using SQL:

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

- **Revenue by Gender** : Compared total revenue generated by male vs. female customers.

gender	revenue
Male	157890
Female	75191

- **High-Spending-Discount-user:** Identified customers who used discounts but still spent above the average purchase amount.

customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90
22	62
24	88
29	94

- **Top 5 Products by Rating:** Found Products with highest average review rating.

item_purchased	average_product_rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Skirt	3.78

➢ **Shopping Type Comparison:** Compared average purchase amounts between Standard and Express shipping.

shipping_type	avg_purchased_amount
Express	60.48
Standard	58.46

➢ **Subscribe vs. Non-Subscribe:** Compared average spend and total_ravenue across subscription status.

subscription_status	total_customer	avg_spend	total_revenue
Yes	1053	59.49	62645
No	2847	59.87	170436

➢ **Discount-Dependent Products:** Identifies 5 products with the highest percentage of discount purchase.

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

➢ **Customer Segmentation:** Classified customer into New, Returning, and Loyal segments based on purchase history.

customer_segment	number_of_customers
Loyal	3116
Returning	701
New	83

➢ **Top 3 Products per Category:** Listed the most purchased products within each category.

item_rank	category	item_purchased	total_orders
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
1	Clothing	Blouse	171
2	Clothing	Pants	171
3	Clothing	Shirt	169
1	Footwear	Sandals	160
2	Footwear	Shoes	150
3	Footwear	Sneakers	145
1	Outerwear	Jacket	163
2	Outerwear	Coat	161

➤ **Repeat Buyers & Subscriptions:** Checked whether customers with >5 purchases are more likely to subscribe.
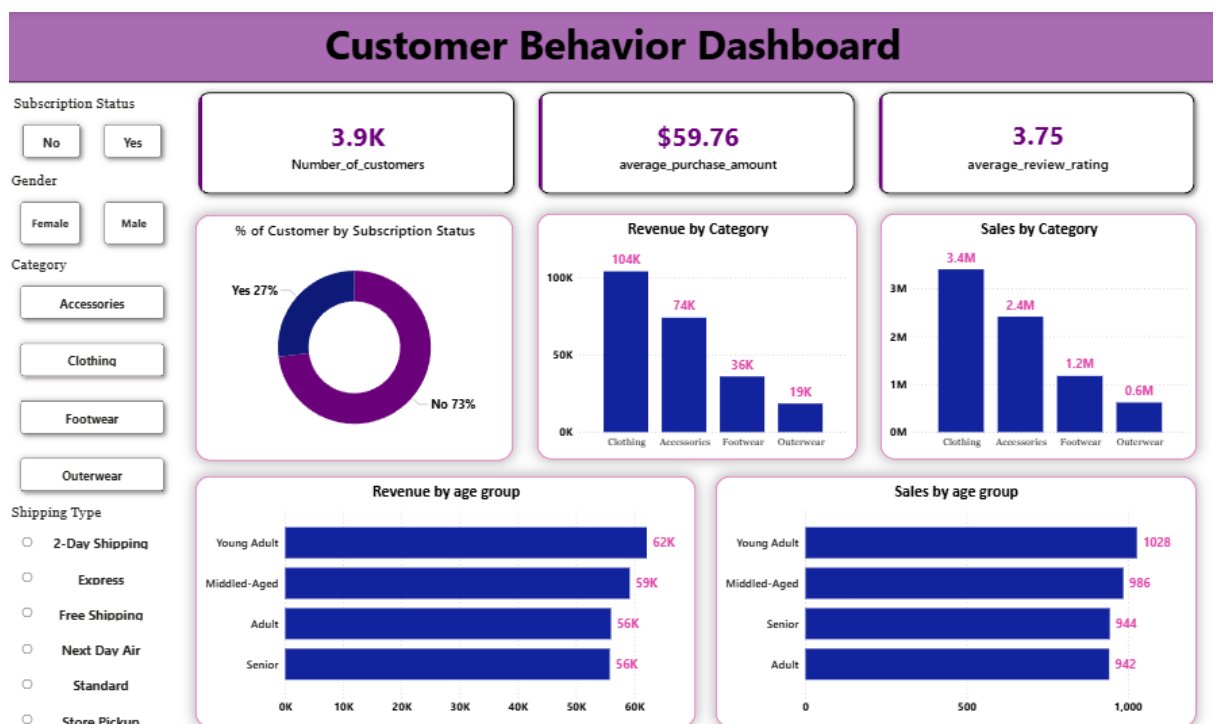
subscription_status	repeat_buyers
Yes	958
No	2518

➤ **Revenue by Age Group:** Calculated total revenue contribution of each age group.

age_group	total_revenue
Young Adult	62143
Middled-Aged	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.