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.

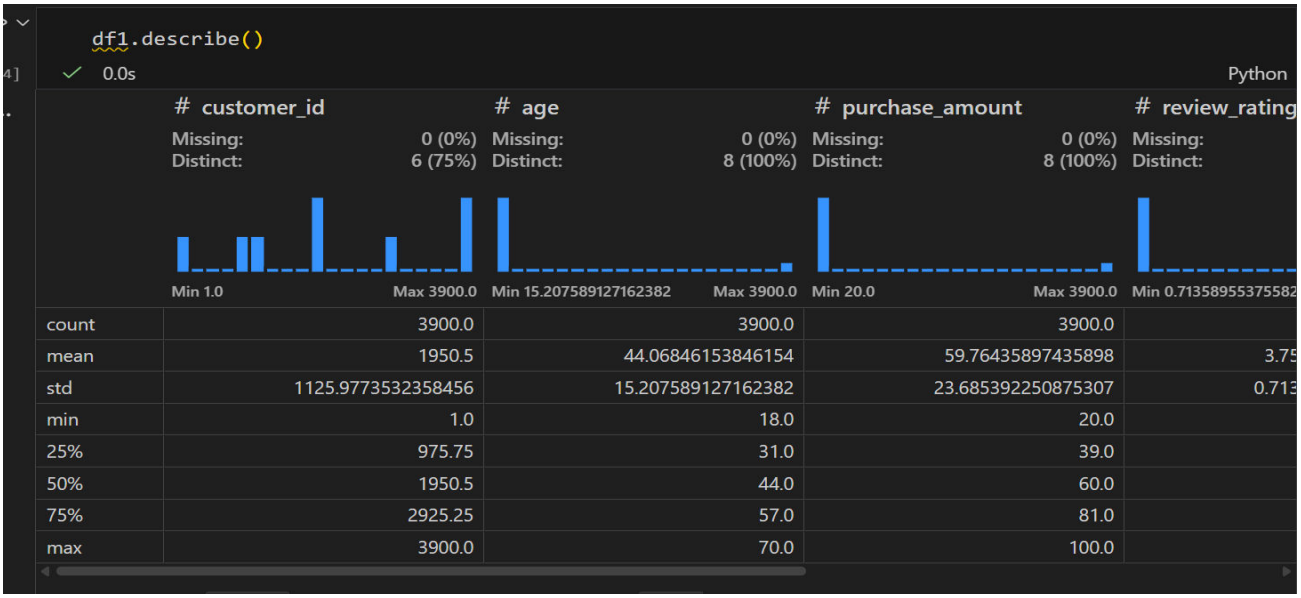
2. Dataset Summary

- 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	purchase_amount	review_rating
count	3900.0	3900.0	3900.0	
mean	1950.5	44.06846153846154	59.76435897435898	3.75
std	1125.9773532358456	15.207589127162382	23.685392250875307	0.713
min	1.0	18.0	20.0	
25%	975.75	31.0	39.0	
50%	1950.5	44.0	60.0	
75%	2925.25	57.0	81.0	
max	3900.0	70.0	100.0	

# review_rating	# previous_purchases	# frequency_of_purchases_...
Missing: 0 (0%)	Missing: 0 (0%)	Missing: 0 (0%)
Distinct: 8 (100%)	Distinct: 8 (100%)	Distinct: 8 (100%)
Min 0.7135895537558273	Min 1.0	Min 7.0
Max 3900.0	Max 3900.0	Max 3900.0
3900.0	3900.0	3900.0
3.750051282051282	25.35153846153846	89.13307692307693
0.7135895537558273	14.447125170462309	119.03756558017905
2.5	1.0	7.0
3.1	13.0	14.0
3.8	25.0	30.0
4.4	38.0	90.0
5.0	50.0	365.0

- **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:** 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.

```
pd.read_sql("SELECT * FROM cust01 LIMIT 5;", engine)
```

Python

	# customer_id	# age	gender	item_purchas
0	1	55	Male	Blouse
1	2	19	Male	Sweater
2	3	50	Male	Jeans
3	4	21	Male	Sandals
4	5	45	Male	Blouse

4. Data Analysis using SQL (Business Transactions)

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	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	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
	32	79
	33	67
	35	91

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

	item_purchased	avg_rating
▶	Gloves	3.86
	Sandals	
	Boots	3.82
	Hat	3.8
	Handbag	3.78

Sandals

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

	shipping_type	avg_p_ampount
▶	Express	60.48
	Standard	58.46

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

	subscription_status	total_customers	total_revenue	avg_revenue
▶	Yes	1053	62645	59.49
	No	2847	170436	59.87

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

item_purchased	disapplied
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37
Boots	46.53
Jeans	45.97
Dress	45.18
Hoodie	45.03
Backpack	44.76
Belt	44.72
Jewelry	44.44
Shorts	43.31
Gloves	42.86
Scarf	42.68

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

csegment	Number of customer
Loyal	3279
Returning	538
New	83

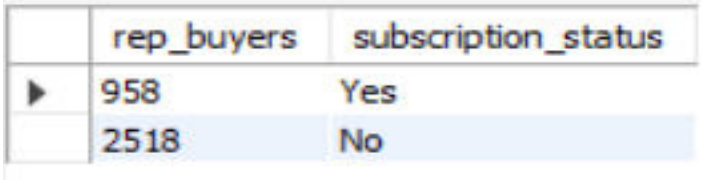
8. **Top 10 Cities with revenue :** Listed most revenue generated cities.

total_revenue	location
5784	Montana
5617	Illinois
5605	California
5587	Idaho
5514	Nevada
5261	Alabama
5257	New York
5220	North Dakota
5174	West Virginia
5172	Nebraska

9. **Top 3 Products per Category** – Listed the most purchased products within each category.

item_purchased	category	total_orders
Belt	Accessories	161
Shirt	Clothing	169
Sneakers	Footwear	145

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



rep_buyers	subscription_status
958	Yes
2518	No

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

	total_revenue	age_group
▶	62143	Young adult
	59197	Middle aged
	55978	Adult
	55763	Senior

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.