Customer Shopping Behavior Analysis

• 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

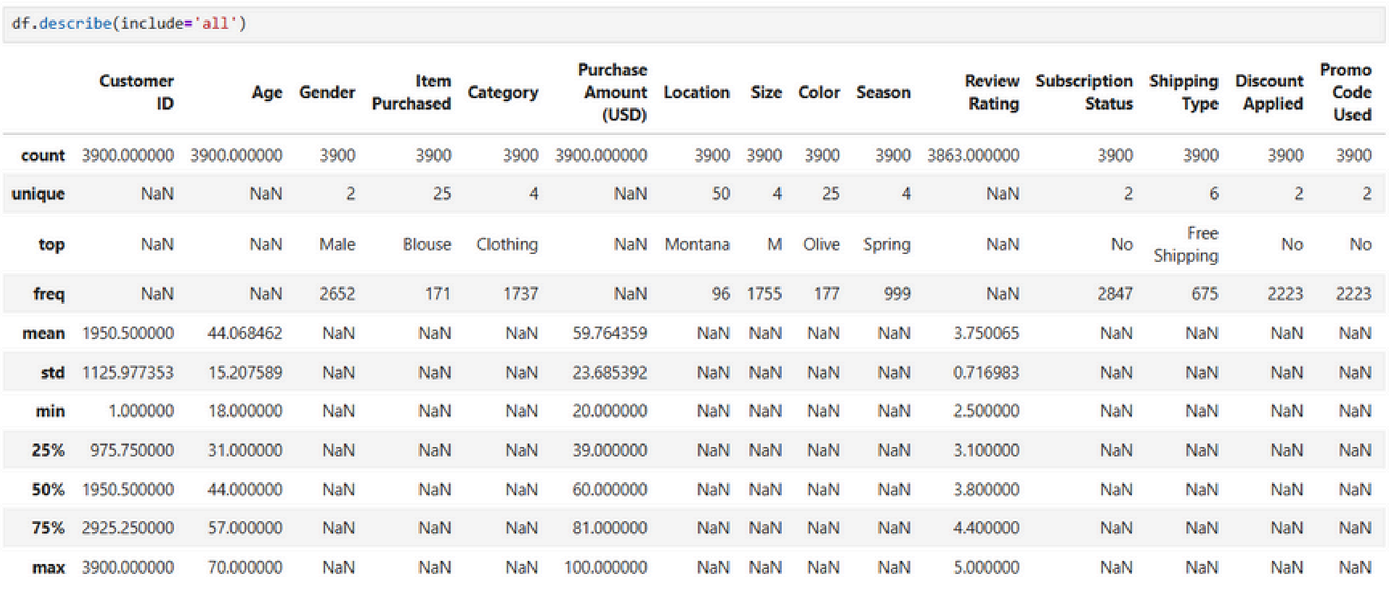
• Dataset Summary :

- **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

• Exploratory Data analysis using Python :

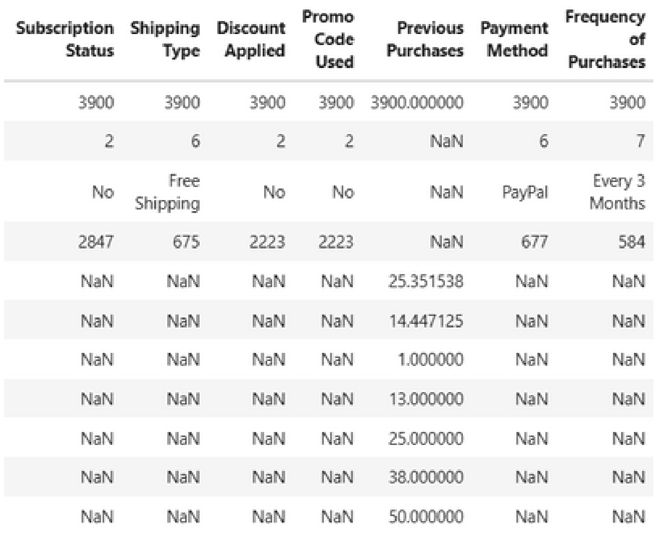
We began with data preparation and cleaning in Python:

- **Data Loading**: Imported the dataset using pandas.
- **Initial Exploration**: Used df.info() to check structure and .describe() for summary statistics.



```
df.describe(include='all')
```

	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
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN



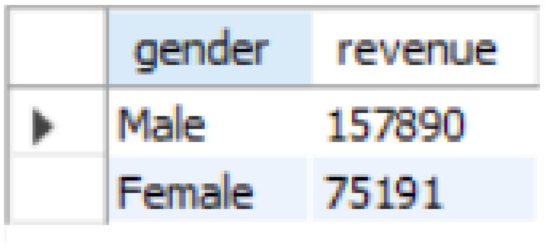
Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3900	3900	3900	3900	3900.000000	3900	3900
2	6	2	2	NaN	6	7
No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
2847	675	2223	2223	NaN	677	584
NaN	NaN	NaN	NaN	25.351538	NaN	NaN
NaN	NaN	NaN	NaN	14.447125	NaN	NaN
NaN	NaN	NaN	NaN	1.000000	NaN	NaN
NaN	NaN	NaN	NaN	13.000000	NaN	NaN
NaN	NaN	NaN	NaN	25.000000	NaN	NaN
NaN	NaN	NaN	NaN	38.000000	NaN	NaN
NaN	NaN	NaN	NaN	50.000000	NaN	NaN

- **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 PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

• SQL analysis using SQL (business transactions)

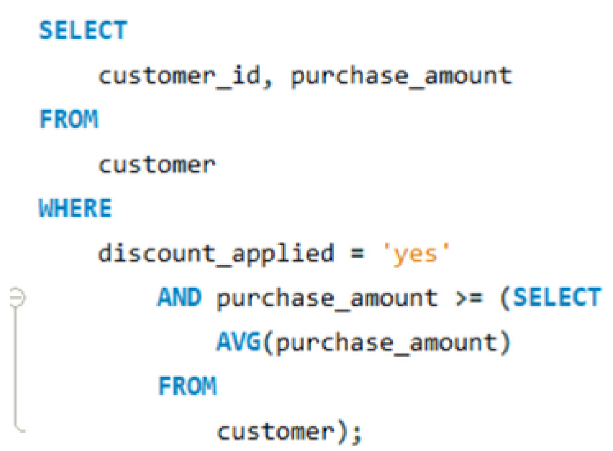
We performed structured analysis in MySQL to answer key business questions:

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

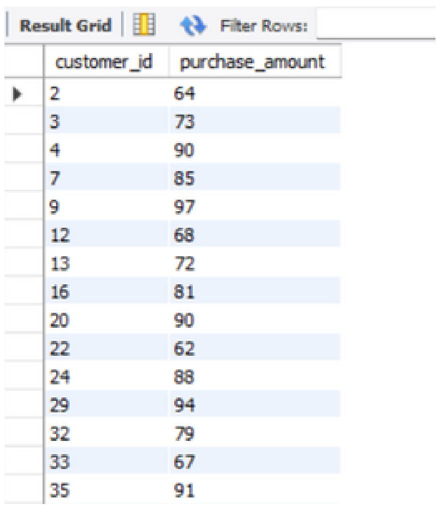


gender	revenue
Male	157890
Female	75191

2.**High Spending Discount users** : Identified Customers who usud a Discount on their order but still paid more than the average amount spent

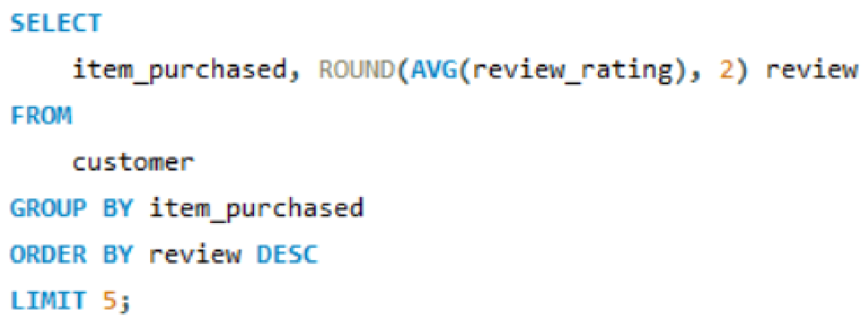


```
SELECT
    customer_id, purchase_amount
FROM
    customer
WHERE
    discount_applied = 'yes'
        AND purchase_amount >= (SELECT
            AVG(purchase_amount)
        FROM
            customer);
```

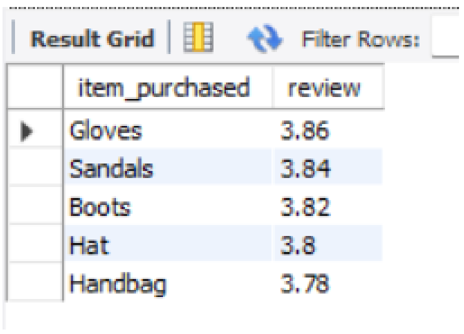


customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90
22	62
24	88
29	94
32	79
33	67
35	91

3. **Top 5 products by rating :** Analyzed top 5 products with highest average review rating

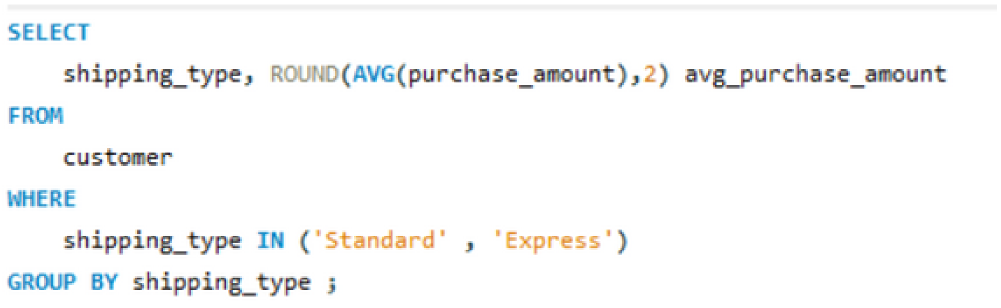


```
SELECT
    item_purchased, ROUND(AVG(review_rating), 2) review
FROM
    customer
GROUP BY item_purchased
ORDER BY review DESC
LIMIT 5;
```

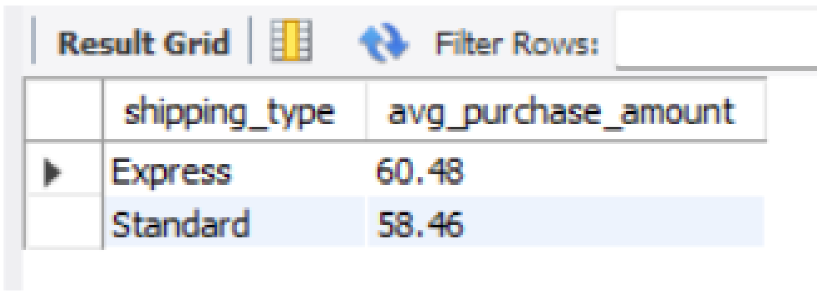


item_purchased	review
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Handbag	3.78

4. **Shipping Type Comparison:** compared average purchase amounts between Standard and express shipping

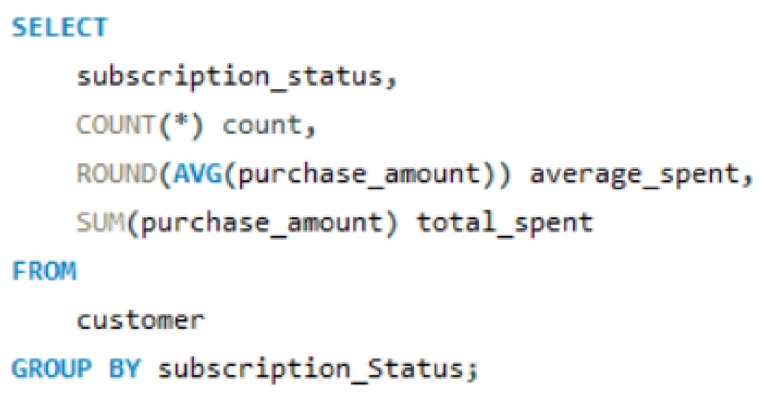


```
SELECT
    shipping_type, ROUND(AVG(purchase_amount),2) avg_purchase_amount
FROM
    customer
WHERE
    shipping_type IN ('Standard' , 'Express')
GROUP BY shipping_type ;
```

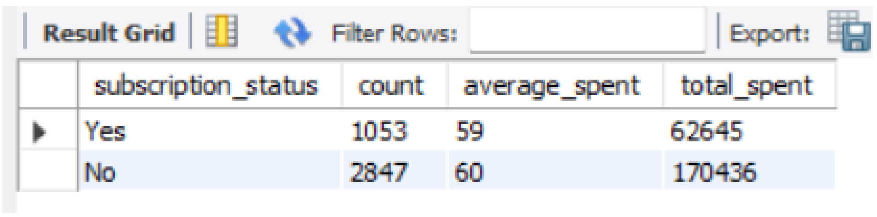


shipping_type	avg_purchase_amount
Express	60.48
Standard	58.46

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

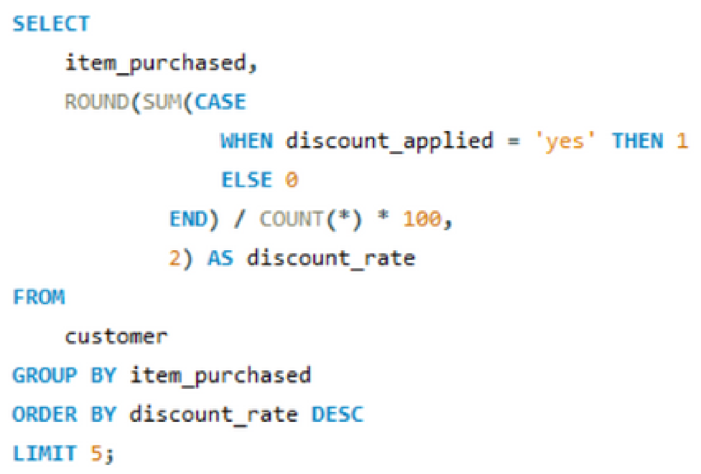


```
SELECT
    subscription_status,
    COUNT(*) count,
    ROUND(AVG(purchase_amount)) average_spent,
    SUM(purchase_amount) total_spent
FROM
    customer
GROUP BY subscription_Status;
```

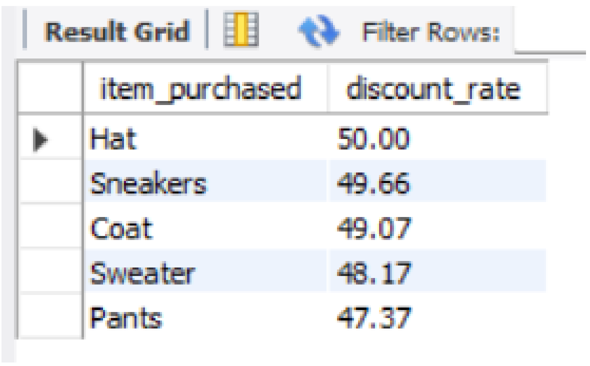


	subscription_status	count	average_spent	total_spent
▶	Yes	1053	59	62645
	No	2847	60	170436

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

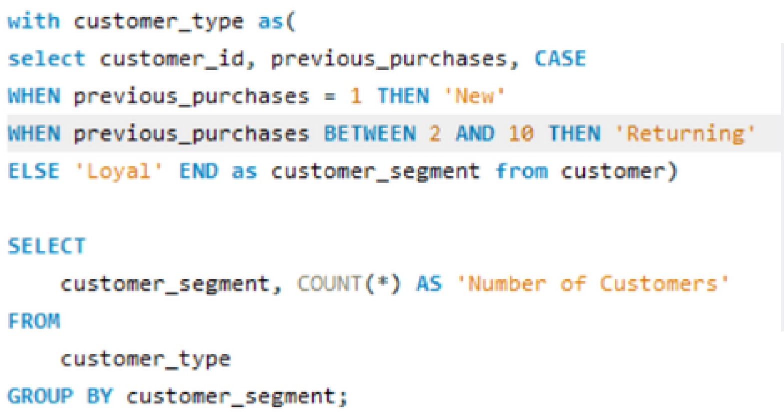


```
SELECT
    item_purchased,
    ROUND(SUM(CASE
                WHEN discount_applied = 'yes' THEN 1
                ELSE 0
            END) / COUNT(*) * 100,
            2) AS discount_rate
FROM
    customer
GROUP BY item_purchased
ORDER BY discount_rate DESC
LIMIT 5;
```

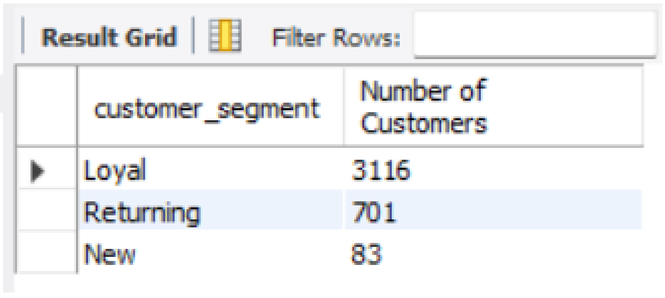


	item_purchased	discount_rate
▶	Hat	50.00
	Sneakers	49.66
	Coat	49.07
	Sweater	48.17
	Pants	47.37

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



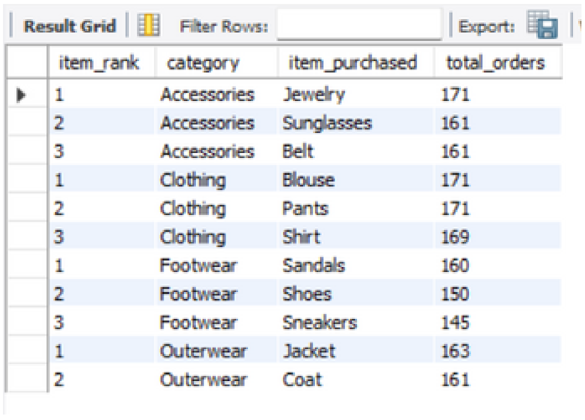
```
with customer_type as(
select customer_id, previous_purchases, CASE
WHEN previous_purchases = 1 THEN 'New'
WHEN previous_purchases BETWEEN 2 AND 10 THEN 'Returning'
ELSE 'Loyal' END as customer_segment from customer)

SELECT
    customer_segment, COUNT(*) AS 'Number of Customers'
FROM
    customer_type
GROUP BY customer_segment;
```



	customer_segment	Number of Customers
▶	Loyal	3116
	Returning	701
	New	83

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

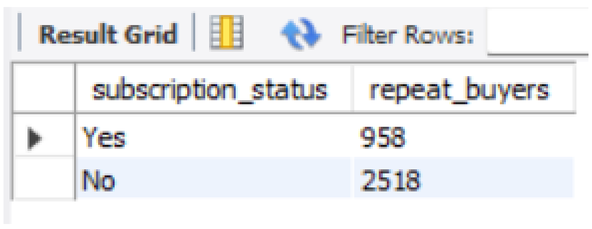
```
WITH item_counts AS (
    SELECT category,
           item_purchased,
           COUNT(customer_id) AS total_orders,
           ROW_NUMBER() OVER (PARTITION BY category ORDER BY COUNT(customer_id) DESC) AS item_rank
    FROM customer
    GROUP BY category, item_purchased
)
SELECT item_rank,category, item_purchased, total_orders
FROM item_counts
WHERE item_rank <=3;
```



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

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

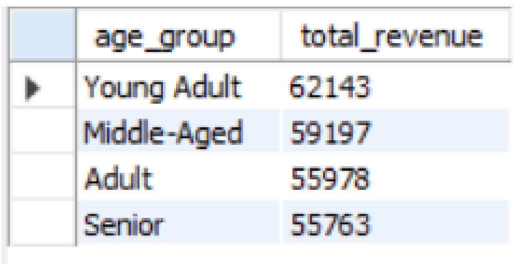
```
SELECT subscription_status,
       COUNT(customer_id) AS repeat_buyers
FROM customer
WHERE previous_purchases > 5
GROUP BY subscription_status;
```



subscription_status	repeat_buyers
Yes	958
No	2518

10. **Revenue by Age Group :** calculated total revenue contribution by age groups

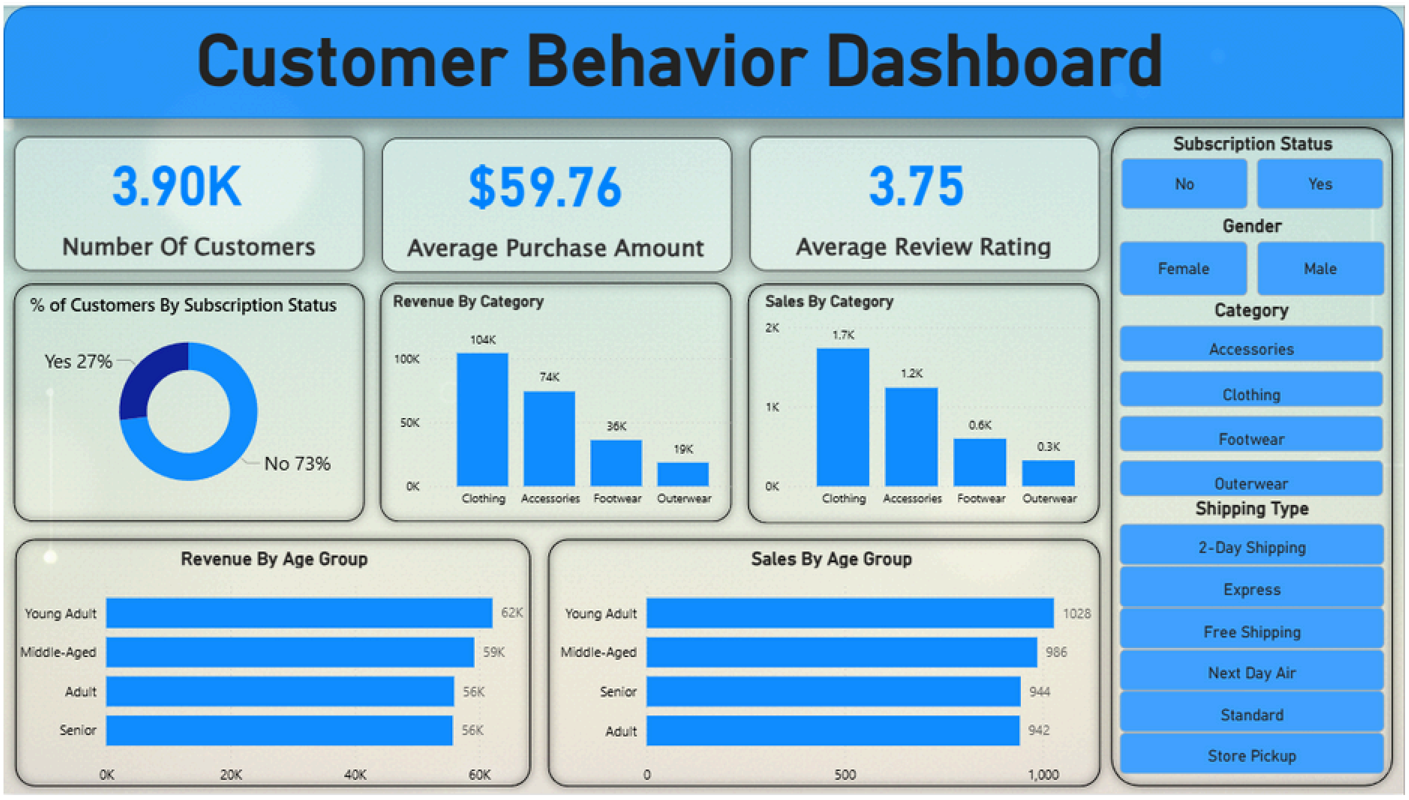
```
SELECT
    age_group, SUM(purchase_amount) total_revenue
FROM
    customer
GROUP BY age_group
ORDER BY total_revenue DESC;
```



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

- ## Dashboard (Power BI)

Finally We built a Completely Interactive Power BI Dashboard



- ## Business Recommendations

  - **Boost Subscriptions :** Promote exclusive benifits 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**.**