Customer Shopping Behavior Analysis

Problem Statement

A leading retail company wants to better understand its customers' shopping behavior in order to improve sales, customer satisfaction and long-term loyalty. The management team has noticed changes in purchasing patterns across demographics, product categories and sales channels (online vs offline). They are particularly interested in uncovering which factors, such as discounts, reviews, seasons or payment preferences, drive consumer decisions and repeat purchases.

You are tasked with analyzing the company's consumer behavior dataset to answer the following overarching business question:

"How can the company leverage consumer shopping data to identify trends, improve customer engagement, and optimize marketing and product strategies?"

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:
  1. Customer demographics (Age, Gender, Location, Subscription Status)
  2. Purchase details (Item Purchased, Category, Purchase Amount, Season, Size, Color)
  3. Shopping behavior (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)
  4. Missing Data: 37 values in Review Rating column

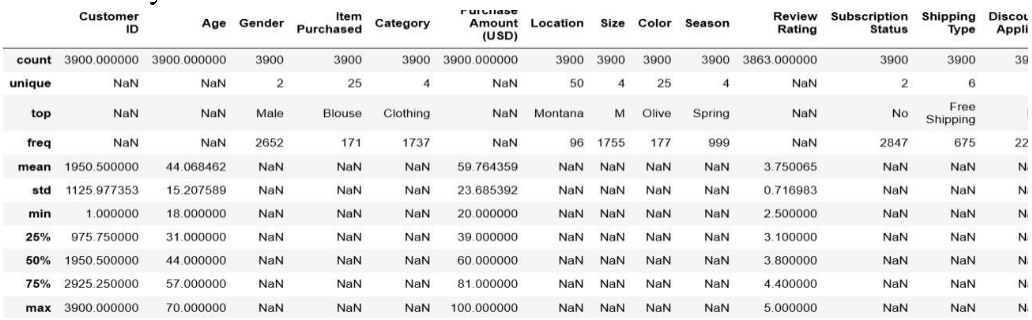
Deliverables

1. **Data Preparation & Modeling (Python):** Clean and transform the raw dataset for analysis.
2. **Data Analysis (SQL):** Organize the data into a structured format, simulate business transactions, and run queries to extract insights on customer segments, loyalty, and purchase drivers.
3. **Visualization & Insights (Power BI):** Build an interactive dashboard that highlights key patterns and trends, enabling stakeholders to make data-driven decisions.
4. **Report and Presentation:** Write a clear project report summarizing your key findings and business recommendations. Prepare a presentation that visually communicates insights and actionable recommendations to stakeholders.
5. **GitHub Repository:** Include all Python scripts, SQL queries, and dashboard files in a well-structured repository.

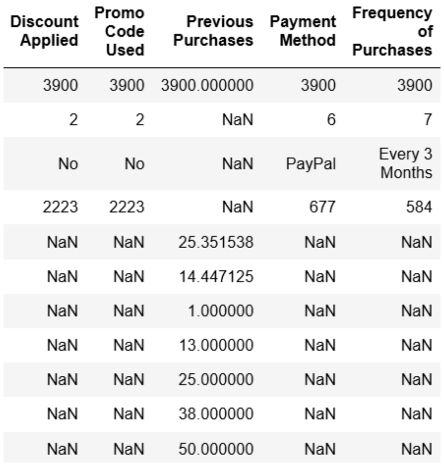
Exploratory Data Analysis using Python

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

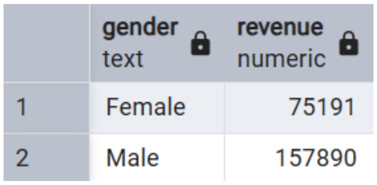
- **Missing Data Handling:** Checked for null values and imputed missing values in the Review Ratingcolumn using the median rating of each product category.
- **Column Standardization:** Renamed columns to **snake case** for better readability and documentation.
- **Feature Engineering:**
  - Created **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 PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

Data Analysis using SQL (Business Transactions)

Q1. What is the total revenue generated by male vs. female customers? (Comparing revenue across demographics)

select gender, SUM (purchase_amount) as revenue from customer

group by gender



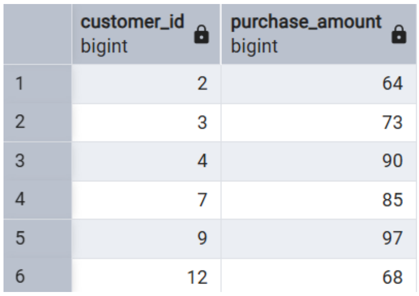
	gender text	revenue numeric
1	Female	75191
2	Male	157890

Q2. Which customers used a discount but still spent more than the average purchase amount?

select customer_id, purchase_amount

from customer

where discount_applied='Yes' and purchase_amount >= (select AVG (purchase_amount) from customer)

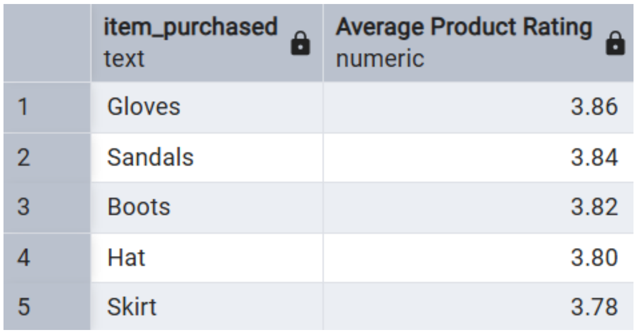


	customer_id bigint	purchase_amount bigint
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68

...

Q3. Which are the top 5 products with the highest average review rating? (These products can be highlighted in marketing campaigns and also be sold at premium price)

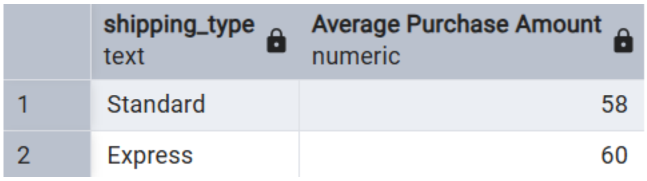
```
select item_purchased,

ROUND (AVG (review_rating::numeric), 2) as "Average Product Rating"

from customer

group by item_purchased

order by avg (review_rating) desc

limit 5;
```



	item_purchased text	Average Product Rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

Q4. Compare the average Purchase Amounts between Standard and Express Shipping. (Helps to decide better shipping type)

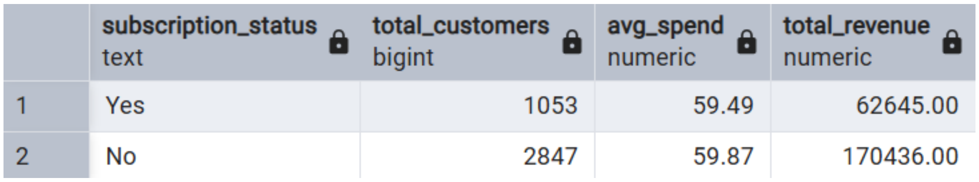
```
select shipping_type,

ROUND (avg(purchase_amount)) as "Average Purchase Amount"

from customer

where shipping_type in ('Standard', 'Express')

group by shipping_type;
```



	shipping_type text	Average Purchase Amount numeric
1	Standard	58
2	Express	60

Q5. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers. (Tells us whether subscriptions are generating good returns)

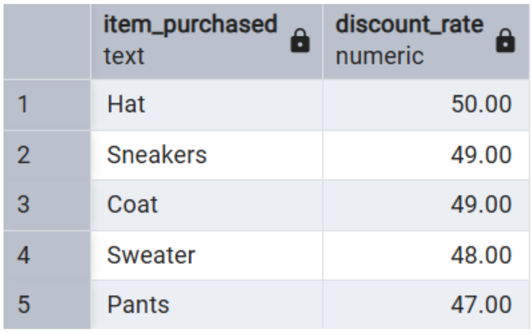
```
select subscription_status,
COUNT (customer_id) as total_customers,
ROUND (avg(purchase_amount), 2) as "avg_spend",
ROUND (SUM (purchase_amount), 2) as "total_revenue"
from customer
group by subscription_status
order by total_revenue, avg_spend desc;
```



	subscription_status text	total_customers bigint	avg_spend numeric	total_revenue numeric
1	Yes	1053	59.49	62645.00
2	No	2847	59.87	170436.00

Q6. Which 5 products have the highest percentage of purchases with discounts applied? (Tells which products rely heavily on discount to sell)

```
select item_purchased,
ROUND (100*SUM (CASE WHEN discount_applied='Yes' THEN 1 ELSE 0 END)/COUNT (*), 2) as discount_rate
from customer
group by item_purchased
order by discount_rate desc
limit 5;
```



	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.00
3	Coat	49.00
4	Sweater	48.00
5	Pants	47.00

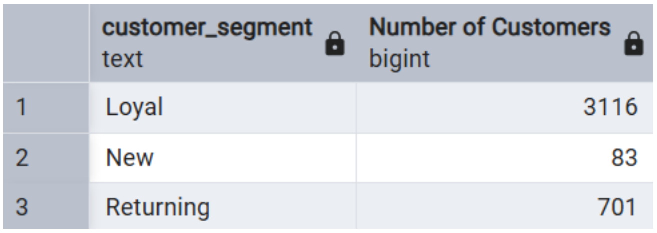
Q7. Segment customers into New, Returning, and Loyal based on their total number of previous purchases, and show the count of each segment. (Helps understand customer loyalty)

Soln. We will segment customers according to the following:

- Customer bought only once → New
- Previous purchases = 2-10 times → Returning
- Previous purchases > 10 times → Loyal

Creating a **Common Table Expression (CTE)** for this purpose.

```
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;
```

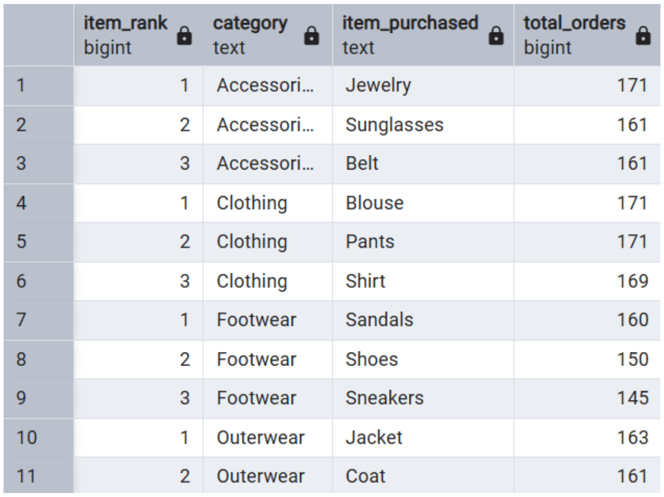


	customer_segment text	Number of Customers bigint
1	Loyal	3116
2	New	83
3	Returning	701

Q8. What are the top 3 most purchased products within each category?

Soln. A window function (Here, row_number() as for many items, the count of total_orders are same but we need three different ranks) to rank products by total orders within each category. Then picking the top 3. CTE is also used to write the query.

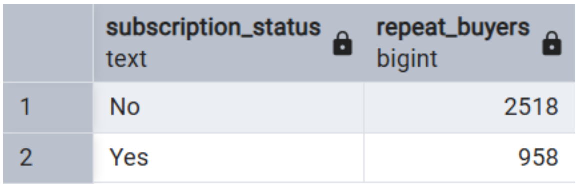
```
with item_counts as (
select category,
item_purchased,
COUNT (customer_id) as total_orders,
ROW_NUMBER () over (partition by category order by count(customer_id) DESC) as item_rank
from customer
group by category, item_purchased
)
select item_rank, category, item_purchased, total_orders
from item_counts
where item_rank<=3;
```



	item_rank	category	item_purchased	total_orders
	bigint	text	text	bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	Belt	161
4	1	Clothing	Blouse	171
5	2	Clothing	Pants	171
6	3	Clothing	Shirt	169
7	1	Footwear	Sandals	160
8	2	Footwear	Shoes	150
9	3	Footwear	Sneakers	145
10	1	Outerwear	Jacket	163
11	2	Outerwear	Coat	161

Q9. Are customers who are repeat buyers (more than 5 previous purchases) also likely to subscribe?

```
select subscription_status,
COUNT (customer_id) AS repeat_buyers
from customer
WHERE previous_purchases>5
group by subscription_status;
```



	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

Q10. What is the revenue contribution of each age group?

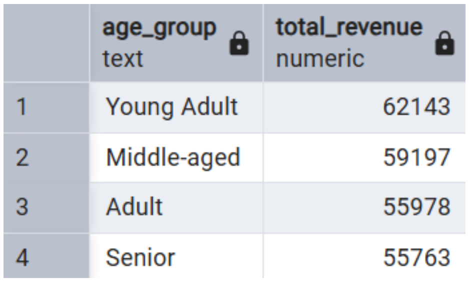
select age_group,

SUM (purchase_amount) as total_revenue

from customer

group by age_group

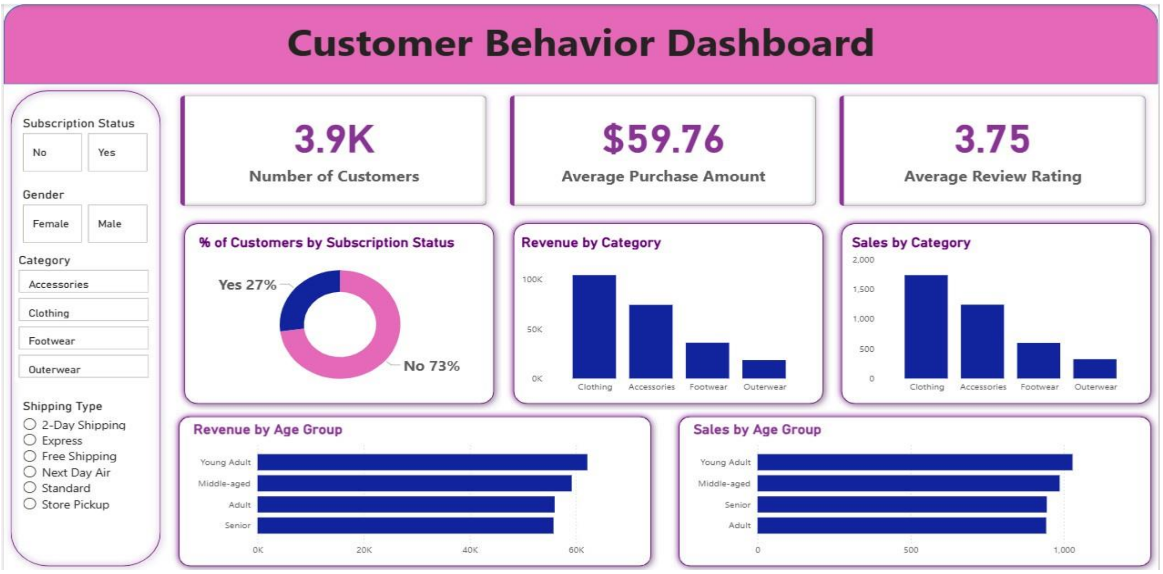
order by total_revenue DESC;



	age_group text	total_revenue numeric
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

Dashboard in Power BI

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



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.