End-to-End Customer Shopping Behavior Analysis

Python → SQL → Power BI

Name: Harsh patwa
Tools: Python, PostgreSQL (SQL), Power BI
Dataset: Customer Shopping Behavior Dataset

---

1. Project Overview

This project demonstrates a complete **end-to-end data analytics workflow**, starting from raw data cleaning to business analysis and dashboard creation.

The objective of this project is to:

- Clean and prepare raw customer data using **Python**
- Store and analyze data using **SQL**
- Solve real-world **business questions**
- Visualize insights using an interactive **Power BI dashboard**

This project reflects a real industry-style data analyst workflow.

---

2. Dataset Description

The dataset contains customer-level shopping information including:

- Customer demographics (age, gender, age group)
- Purchase behavior (purchase amount, previous purchases)
- Product information (item purchased, category)

- Marketing factors (discount applied, subscription status)
- Logistics details (shipping type)
- Customer feedback (review rating)

Total records: ~3,900 customers.

---

3. Data Cleaning & Preparation (Python)

Purpose

Python was used to clean, standardize, and prepare the raw dataset before loading it into SQL for analysis.

Steps Performed

- Loaded the dataset using Pandas
- Checked for missing values and duplicates
- Standardized categorical fields (Yes/No values)
- Ensured correct data types for numeric columns
- Created derived fields such as age groups
- Exported the cleaned dataset for SQL analysis

Sample Code

```
import pandas as pd

df = pd.read_csv("customer_shopping_behavior.csv")

df['discount_applied'] = df['discount_applied'].str.title()
df['subscription_status'] = df['subscription_status'].str.title()

df.to_csv("customer_cleaned.csv", index=False)
```

---

4. Data Storage (SQL)

The cleaned dataset was loaded into a **PostgreSQL** database as a table named `customer`.

SQL was used to:

- Perform aggregations
- Segment customers
- Analyze revenue and behavior patterns
- Answer business-driven analytical questions

---

5. Business Questions & SQL Analysis

Q1. Total Revenue by Gender

```
select sum(purchase_amount) as revenue, gender
from customer
group by gender;
```

	revenue numeric	gender text
1	75191	Female
2	157890	Male

Insight: Male contributes more to total revenue.

Q2. Customers Who Used Discounts but Spent Above Average

```
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
7	13	72
8	16	81

Total rows: 839 Query complete 00:00:(

Insight: Customers who use discounts but still spend above the average demonstrate strong purchasing power and are less price-sensitive.

Q3. Top 5 Products by Average Review Rating

```
select item_purchased, avg(review_rating) as avg_review_rating
from customer
group by item_purchased
order by avg_review_rating desc
limit 5;
```

	item_purchased text	avg_review_rating double precision
1	Gloves	3.8614285714285725
2	Sandals	3.8443750000000003
3	Boots	3.8187500000000005
4	Hat	3.8012987012987005
5	Skirt	3.784810126582278

Insight: These products have the highest average customer ratings, indicating strong customer satisfaction and consistent product quality.

Q4. Average Purchase Amount by Shipping Type

```
select shipping_type, avg(purchase_amount) as avg_purchase_amount
from customer
where shipping_type in ('Express','Standard')
group by shipping_type;
```

	avg_purchase_amount numeric	shipping_type text
1	58.4602446483180428	Standard
2	60.4752321981424149	Express

Insight: Express shipping users generally show higher spending behavior.

Q5. Spending Comparison: Subscribers vs Non-Subscribers

```
select
subscription_status,
count(customer_id) as total_customers,
avg(purchase_amount) as avg_purchase_amount,
sum(purchase_amount) as total_revenue
from customer
group by subscription_status;
```

	total_customer bigint	avg_purchase_amount numeric	total_money_spend numeric	subscription_status text
1	2847	59.8651211801896733	170436	No
2	1053	59.4919278252611586	62645	Yes

Insight: Non Subscribed customers generate higher average and total revenue.

Q6. Products with Highest Discount Usage

```
select
item_purchased,
sum(case when discount_applied='Yes' then 1 else 0 end) * 100.0 /
count(*)
as discount_percentage
from customer
group by item_purchased
order by discount_percentage desc
limit 5;
```

	item_purchased text	discount_percentage numeric
1	Hat	50.0000000000000000
2	Sneakers	49.6551724137931034
3	Coat	49.0683229813664596
4	Sweater	48.1707317073170732
5	Pants	47.3684210526315789

Insight: These products are frequently purchased with discounts, highlighting strong discount dependency.

Q7. Customer Segmentation Based on Previous Purchases

```
select
case
when previous_purchases < 2 then 'New'
```

```
when previous_purchases between 2 and 10 then 'Returning'
else 'Loyal'
end as customer_segment,
count(*) as customer_count
from customer
group by customer_segment;
```

	customer_segment text	customer_count bigint
1	returning	701
2	Loyal	3116
3	New	83

Insight: Loyal customers constitute the largest customer segment, indicating strong customer retention and repeat purchasing behavior, while new customers form a very small proportion of the customer base.

Q8. Top 3 Most Purchased Products per 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 category, item_purchased, total_orders
from item_counts
where item_rank <= 3;
```

	category text	item_purchased text	total_orders bigint
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
4	Clothing	Blouse	171
5	Clothing	Pants	171
6	Clothing	Shirt	169
7	Footwear	Sandals	160
8	Footwear	Shoes	150
9	Footwear	Sneakers	145

Total rows: 11 Query complete 00:00:00.230

Insight: These are the category-wise top-performing products.

Q9. Repeat Buyers and Subscription Status

```
select
subscription_status,
count(customer_id) as repeat_buyers
from customer
where previous_purchases > 5
group by subscription_status;
```

	repeat_buyers bigint	subscription_status text
1	2518	No
2	958	Yes

Insight: Repeat buyers are more likely to be non-subscribed customers, suggesting that repeat purchasing does not necessarily translate into subscription adoption.

Q10. Revenue Contribution by Age Group

```
select
age_group,
sum(purchase_amount) as total_revenue
from customer
group by age_group
order by total_revenue desc;
```

	total_revenue numeric	age_group text
1	62143	young_adult
2	59197	middle_age
3	55978	adult
4	55763	senior

Insight: Young adults generate the highest revenue, followed by middle-aged customers, indicating that younger segments are the primary revenue drivers.

---

6. Dashboard Development (Power BI)

An interactive Power BI dashboard was created to visualize key insights.

Key Performance Indicators (KPIs)

- **Total Customers:** Displays the total number of unique customers in the dataset.
- **Average Purchase Amount:** Shows the average amount spent per transaction.
- **Average Review Rating:** Represents the overall customer satisfaction score based on product reviews.

Visualizations

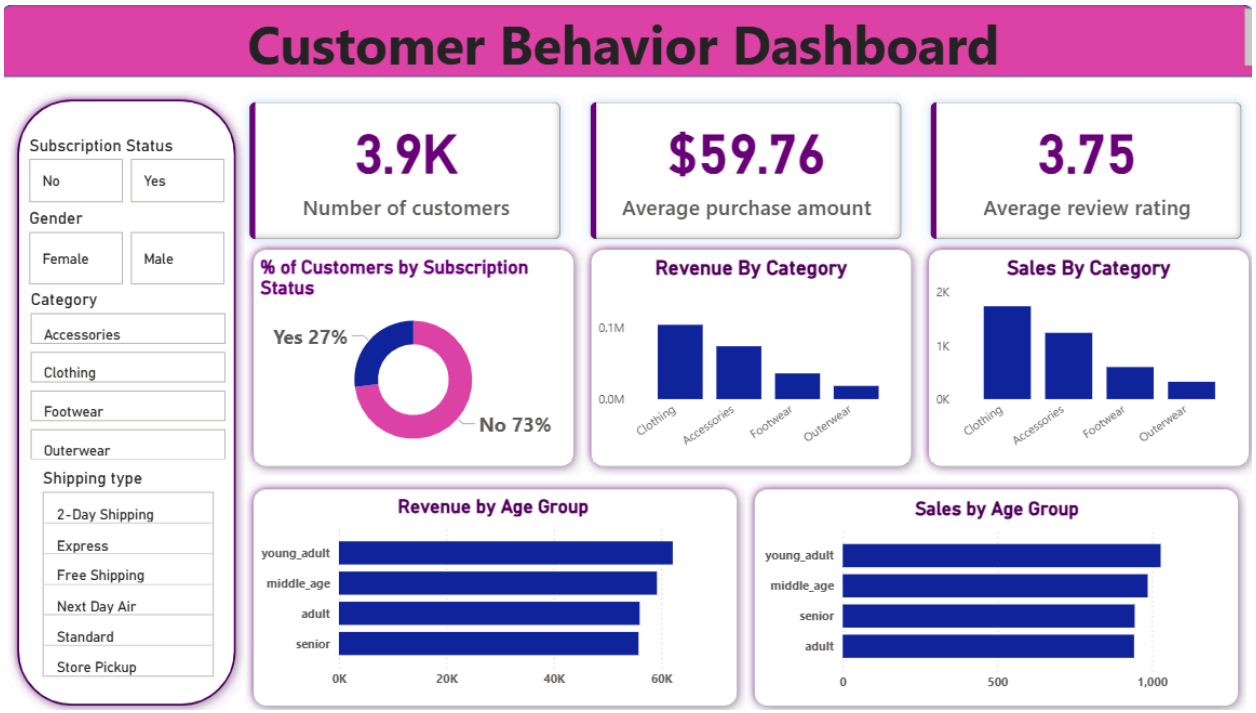
- **Revenue by Category:** Highlights total revenue generated across different product categories to identify high-performing segments.
- **Sales by Category:** Shows the number of purchases per category, enabling comparison of sales volume.

- **Subscription Status Distribution:** Visualizes the proportion of subscribed versus non-subscribed customers.
- **Revenue by Age Group:** Displays revenue contribution from different age groups.
- **Sales by Age Group:** Compares purchase frequency across age groups.

Interactive Filters (Slicers)

- **Gender:** Filters dashboard metrics by customer gender.
- **Category:** Allows analysis by product category.
- **Subscription Status:** Enables comparison between subscribed and non-subscribed customers.
- **Shipping Type:** Filters results based on the selected shipping method.
-

The dashboard allows dynamic filtering and clear storytelling.



7. Key Insights Summary

- **Non-subscribed customers generate higher average and total revenue**, indicating strong purchasing behavior even without subscription benefits.
- **Loyal customers form the largest customer segment**, reflecting strong retention and repeat purchasing behavior.
- **Several products rely heavily on discounts to drive sales**, suggesting price-sensitive demand for these items.
- **Customers using faster shipping options, such as Express and Next Day Air, tend to have higher purchase amounts**, indicating urgency-driven or high-intent purchases.
- **Young adult and middle-aged customers contribute the majority of revenue**, making them the most valuable age groups, while categories like Clothing and Accessories dominate overall sales and revenue.

---

8. Conclusion

This project successfully demonstrates a **complete data analytics lifecycle**:

- Data cleaning using Python
- Business analysis using SQL
- Insight visualization using Power BI

The project reflects practical, real-world data analyst responsibilities and decision-making processes.

---

9. Future Enhancements

- Incorporate **time-based analysis** to study customer behavior trends over different periods.

- Develop **customer lifetime value (CLV) models** to better understand long-term customer profitability.

- Automate the **data ingestion and refresh pipeline** to support near real-time analysis.

- Extend the analysis by integrating **additional customer attributes**, such as location or seasonality.

- Enhance the Power BI dashboard with **advanced interactivity**, including drill-through pages and tooltip insights.

- Deploy the dashboard to **Power BI Service** for scheduled refreshes and broader stakeholder access.