Customer Behaviour ETL & Analytics Project

1. Executive Summary

This project presents an end-to-end **ETL and business intelligence solution** designed to help a leading retail company better understand customer shopping behaviour. The analysis leverages **Python for data extraction, cleaning, and feature engineering**, **SQL for business-focused analysis**, and **Power BI for interactive visualization**.

The goal of this project was to transform raw customer behaviour data into actionable insights that can guide **marketing strategies, customer engagement initiatives, and product optimization decisions**. The final deliverable is a clean analytical dataset, a set of insightful SQL-based answers to key business questions, and a polished Power BI dashboard for decision-makers.

---

2. Problem Statement

A leading retail company wants to better understand its customers' shopping behaviour 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, shipping preferences, and payment methods**, drive consumer decisions and repeat purchases.

Overarching Business Question:

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

---

3. Data Overview

The dataset contains detailed customer-level transaction data, including:

- Customer demographics (age, gender)
- Purchase behaviour (product, category, purchase amount)
- Discounts and promotions
- Reviews and ratings
- Subscription status
- Shipping preferences
- Historical purchase count (repeat behaviour)

This dataset enables both **descriptive and behavioural analysis** of customer shopping patterns.

---

4. ETL Process Overview

4.1 Extract (Python)

- Imported the raw customer behaviour dataset into Python.
- Inspected data structure, data types, and overall quality.
- Identified missing values, duplicates, and inconsistencies.

	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	Previ Purcha
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	

4.2 Transform (Data Cleaning & Feature Engineering)

Using Python (Pandas & NumPy), the following transformations were applied:

Data Cleaning

- Handled missing values using appropriate strategies (imputation).
- Standardized column names to snake case.

Feature Engineering

- Created customer segments: **New, Returning, Loyal** based on previous purchase count.
- Derived age groups for demographic analysis.
- Prepared aggregated fields to support SQL-based business queries.

```
|: 1 #age-group
   2 labels=['Young','Adult', 'Middle Aged', 'Senior']
   3 bins=[0,18,30,50,80]
   4
   5 df['age_group']=pd.cut(df['age'],bins=bins, labels=labels, right=True)
```

```
|: 1 df[['age', 'age_group']].head(20)
```

|:

	age	age_group
0	55	Senior
1	19	Adult
2	50	Middle Aged
3	21	Adult
4	45	Middle Aged
5	46	Middle Aged
6	63	Senior
7	27	Adult
8	26	Adult
9	57	Senior
10	53	Senior
11	30	Adult
12	61	Senior
13	65	Senior
14	64	Senior
15	64	Senior
16	25	Adult
17	53	Senior
18	52	Senior
19	66	Senior

4.3 Load (SQL)

- Exported the cleaned and enriched dataset from Python.
- Loaded the data into a SQL database.

- Designed queries to answer business-driven analytical questions efficiently.

```
select top 20 * from customer
```

110 %

Results | Messages

	customer_id	age	gender	item_purchased	category	purchase_amount	location	size	color	season	review_rating	subscription_status	shipping_typ
1	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express
2	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express
3	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shippi
4	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day A
5	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shippi
6	6	46	Male	Sneakers	Footwear	20	Wyoming	M	White	Summer	2.9	Yes	Standard
7	7	63	Male	Shirt	Clothing	85	Montana	M	Gray	Fall	3.2	Yes	Free Shippi
8	8	27	Male	Shorts	Clothing	34	Louisiana	L	Charcoal	Winter	3.2	Yes	Free Shippi
9	9	26	Male	Coat	Outerwear	97	West Virginia	L	Silver	Summer	2.6	Yes	Express
10	10	57	Male	Handbag	Accesso...	31	Missouri	M	Pink	Spring	4.8	Yes	2-Day Ship
11	11	53	Male	Shoes	Footwear	34	Arkansas	L	Purple	Fall	4.1	Yes	Store Picku
12	12	30	Male	Shorts	Clothing	68	Hawaii	S	Olive	Winter	4.9	Yes	Store Picku
13	13	61	Male	Coat	Outerwear	72	Delaware	M	Gold	Winter	4.5	Yes	Express
14	14	65	Male	Dress	Clothing	51	New Hamps...	M	Violet	Spring	4.7	Yes	Express
15	15	64	Male	Coat	Outerwear	53	New York	L	Teal	Winter	4.7	Yes	Free Shippi
16	16	64	Male	Skirt	Clothing	81	Rhode Island	M	Teal	Winter	2.8	Yes	Store Picku
17	17	25	Male	Sunglasses	Accesso...	36	Alabama	S	Gray	Spring	4.1	Yes	Next Day A
18	18	53	Male	Dress	Clothing	38	Mississippi	XL	Lavender	Winter	4.7	Yes	2-Day Ship

---

5. Business Analysis Using SQL

The transformed dataset was analyzed using SQL to answer key business questions.

5.1 Revenue by Gender

Question: What is the total revenue generated by male vs female customers?

Insight:

- Revenue contribution was compared across genders, helping identify which demographic drives higher sales.
- This insight can support targeted marketing and promotional campaigns.

```
select gender, SUM(purchase_amount) as Revenue
from customer
group by gender
```

110 %

Results | Messages

	gender	Revenue
1	Male	157890
2	Female	75191

5.2 High-Value Discount Users

Question: Which customers used a discount but still spent more than the average purchase amount?

Insight:

- Identified price-insensitive customers who respond well to discounts.
- These customers represent an opportunity for **premium promotions and loyalty programs**.

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

10 %

Results | Messages

	customer_id	purchase_amount
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68
7	13	72
8	16	81
9	20	90
10	22	62
11	24	88
12	29	94

5.3 Top Rated Products

Question: Which are the top 5 products with the highest average review rating?

Insight:

- High-rated products can be promoted more aggressively.
- These products may also justify premium pricing or bundling strategies.

```
select top 5 item_purchased, Round(AVG(review_rating),2) as "Product Average Rating"
from customer
group by item_purchased
order by [Product Average Rating] desc
```

110 %

Results | Messages

	item_purchased	Product Average Rating
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.8
5	Handbag	3.78

5.4 Shipping Type vs Spend

Question: Compare the average purchase amounts between standard and express shipping.

Insight:

- Customers opting for express shipping tend to show different spending behaviour.
- Shipping preferences can be used as a proxy for urgency or willingness to pay.

```
select shipping_type,
AVG(purchase_amount) as "Average Purchase Amount"
from customer
where shipping_type ='Standard' or shipping_type ='Express'
group by shipping_type
```

Messages

shipping_type	Average Purchase Amount
Standard	58
Express	60

5.5 Subscription Impact on Spending

Question: Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers.

Insight:

- Average spend per purchase is identical for subscribed and non-subscribed customers, indicating similar purchasing behaviour at the transaction level.
- Non-subscribed customers generate higher total revenue due to a larger customer base, suggesting subscription strategies should focus on retention and engagement rather than immediate spend increase.

```
select subscription_status,
Count(customer_id) as [Total Customers],
AVG(purchase_amount) as [Average Spend],
SUM(purchase_amount) as [Total Revenue]
from customer
group by subscription_status
order by [Total Revenue], [Average Spend] desc
```

Results Messages

subscription_status	Total Customers	Average Spend	Total Revenue
Yes	1053	59	62645
No	2847	59	170436

5.6 Customer Segmentation

Question: Segment customers into New, Returning, or Loyal based on total previous purchases.

Insight:

- Clear behavioral segmentation allows personalized marketing strategies.
- Loyal customers contribute a disproportionate share of revenue.

```sql
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
order by [Number of Customers] desc
```

10 %

Results | Messages

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

5.7 Top Products by Category

Question: What are the top 3 most purchased products within each category?

Insight:

- Identifies category leaders and customer preferences.
- Supports inventory optimization and cross-selling strategies.

```sql
with item_counts as (
select category,
item_purchased,
count(customer_id) as [Total Purchases],
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 Purchases]
from item_counts
where item_rank <=3
```

Results | Messages

item_rank	category	item_purchased	Total Purchases
1	Accessories	Jewelry	171
2	Accessories	Belt	161
3	Accessories	Sunglasses	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

5.8 Repeat Buyers & Subscription Behaviour

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

Insight:

- Repeat buyers are more likely to be non-subscribers, as a larger portion of repeat customers do not hold a subscription.
- This highlights a missed conversion opportunity, where frequent buyers can be strategically targeted with subscription incentives to improve long-term retention.

```
select subscription_status, count(customer_id) as [Repeat Buyers]
from customer
where previous_purchases > 5
group by subscription_status
```

10 %

Results | Messages

	subscription_status	Repeat Buyers
1	Yes	958
2	No	2518

5.9 Revenue by Age Group

Question: What is the revenue contribution of each age group?

Insight:

- Certain age groups contribute significantly more to total revenue.
- Enables age-targeted marketing and product positioning.

```
select age_group, sum(purchase_amount) as [Total Revenue]
from customer
group by age_group
order by [Total Revenue] desc
```

0 %

Results | Messages

	age_group	Total Revenue
	Senior	88480
2	Middle Aged	87322
3	Adult	53140
4	Young	4139

---

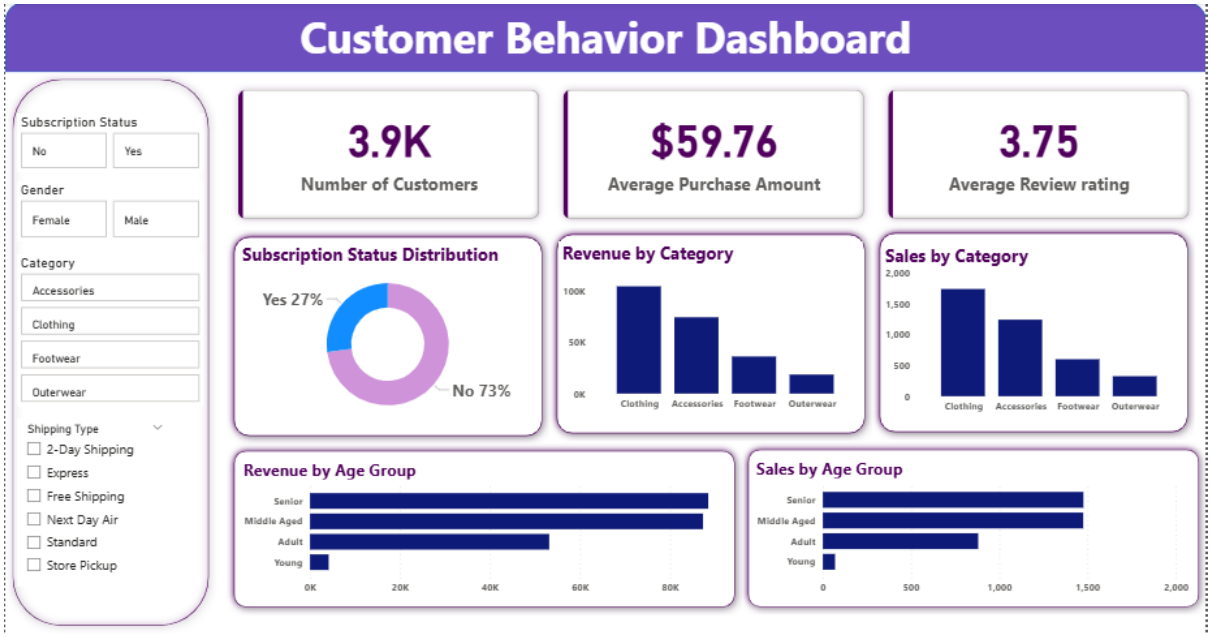
6. Power BI Dashboard

The final step of the project involved building an **interactive Power BI dashboard** to visualize insights derived from SQL analysis.

Dashboard Highlights

- Revenue breakdown by gender and age group
- Subscriber vs non-subscriber performance
- Top products and categories
- Customer segmentation overview
- Shipping preferences and spending patterns

The dashboard allows stakeholders to filter and explore data dynamically, enabling **data-driven decision-making**.



---

7. Key Insights & Recommendations

- Revenue is primarily driven by customer volume rather than subscription status, as non-subscribed customers contribute higher total revenue despite similar average spending behaviour.
- Discount strategies are effective across customer segments, including higher-spending customers, indicating that well-targeted promotions can boost revenue without eroding value.
- Products with higher average review ratings consistently perform better, making them strong candidates for featured promotions and strategic positioning.
- A large portion of repeat buyers are not subscribed, presenting a clear opportunity to convert frequent purchasers into subscribers through tailored incentives.
- Demographic factors such as age and gender show distinct revenue contributions, enabling more refined customer segmentation and personalized marketing efforts.

---

8. Conclusion

This project demonstrates how an end-to-end ETL and analytics pipeline can transform raw customer behavior data into meaningful business insights. By integrating **Python for data preparation**, **SQL for structured analysis**, and **Power BI for visualization**, the analysis uncovers key trends in customer spending, product performance, and engagement patterns.

The findings highlight that while subscriptions do not directly increase average spending, they represent a strategic opportunity for improving retention and long-term customer value. Additionally, insights into discount effectiveness, repeat purchasing behavior, and demographic contributions can help the company optimize its marketing, loyalty programs, and product strategies. Overall, this analysis equips decision-makers with data-driven insights to enhance customer engagement and drive sustainable growth.

---

9. Tools & Technologies Used

- **Python** (pandas, NumPy) – Data cleaning & feature engineering
- **SQL** – Business analysis & querying
- **Power BI** – Data visualization & dashboarding
