CUSTOMER BEHAVIOUR ANALYSIS USING SQL, PYTHON & POWER BI

Author: Ganesh Shiva Kuppaswamy

---

1. PROJECT OVERVIEW

This project analyzes customer shopping behaviour using transactional retail data.
The goal is to identify spending patterns, customer segments, product preferences, loyalty trends, and discount behaviours using:

- **Python (Jupyter Notebook)** → Data cleaning, preparation, and structured EDA
- **MySQL Workbench** → Business-driven analytical queries
- **Power BI** → Interactive visual dashboards and insights

This end-to-end approach reflects a real industry workflow used by data analysts in e-commerce and retail.

---

2. BUSINESS PROBLEM STATEMENT

A retail company has observed changes in customer shopping habits across product categories, age groups, subscription preferences, and discount usage.
Management wants to understand:

- Which customer groups drive revenue?
- How discounts influence spending?
- Which products are most preferred?
- How to convert returning customers into loyal ones?
- How subscription behavior relates to repeat purchases?

Business Question:

"How can customer shopping data be used to identify behaviour patterns and improve marketing, loyalty, and revenue growth strategies?"

---

3. DATASET SUMMARY

The dataset contains transactional records for individual customer purchases.

Dataset Structure

- **Rows:** 3,900
- **Columns:** 18

Key Columns

- **Demographics:** Gender, Age_Group, Subscription_Status
- **Purchases:** Purchase_Amount, Discount_Applied, Previous_Purchases
- **Products:** Category, Item_Purchased
- **Shipping:** Shipping_Type
- **Review:** Review_Rating

	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	Previous Purchases
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31

---

4. PYTHON DATA PREPARATION & EDA

Performed in **Jupyter Notebook** using pandas.

4.1 Data Loading

Loaded the raw CSV file and inspected the structure:

```
# Quick information about the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             3900 non-null   int64
 1   Age                     3900 non-null   int64
 2   Gender                  3900 non-null   object
 3   Item Purchased          3900 non-null   object
 4   Category                3900 non-null   object
 5   Purchase Amount (USD)   3900 non-null   int64
 6   Location                3900 non-null   object
 7   Size                    3900 non-null   object
 8   Color                   3900 non-null   object
 9   Season                  3900 non-null   object
 10  Review Rating           3863 non-null   float64
 11  Subscription Status     3900 non-null   object
 12  Shipping Type           3900 non-null   object
 13  Discount Applied        3900 non-null   object
 14  Promo Code Used         3900 non-null   object
 15  Previous Purchases      3900 non-null   int64
 16  Payment Method          3900 non-null   object
 17  Frequency of Purchases  3900 non-null   object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

```
# Quick Summary statistics
df.describe()
```

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3863.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.750065	25.351538
std	1125.977353	15.207589	23.685392	0.716983	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.800000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

4.2 Missing Value Handling

- Identified missing values in **Review_Rating**
- Filled missing values using **median rating per item/category**

```
# Replacing the null values with the median of the column grouped by its category
df['Review Rating'] = df.groupby('Category')['Review Rating'].transform(lambda x: x.fillna(x.median()))
```

```
df.isnull().sum()
```

```
Customer ID               0
Age                       0
Gender                    0
Item Purchased            0
Category                  0
Purchase Amount (USD)     0
Location                  0
Size                      0
Color                     0
Season                    0
Review Rating             0
Subscription Status       0
Shipping Type             0
Discount Applied          0
Promo Code Used           0
Previous Purchases        0
Payment Method            0
Frequency of Purchases    0
dtype: int64
```

4.3 Data Cleaning

- Standardized column names to snake_case
- Converted data types (numeric, categorical)
- Cleaned inconsistent values (Yes/No → yes/no)

4.4 Feature Engineering

Created additional features for richer analysis:

- **age_group** (Creating a new categorical column as "age_group" from the numerical column "age")
- **purchase_frequency_days** (Creating new numerical column as "purchase_frequency_days" from the categorical column "frequency_of_purchases")
- **customer_segment** (used later in SQL)

---

5. DATA ANALYSIS USING SQL (MySQL Workbench)

All core business questions were analyzed using SQL queries to extract patterns in customer spending, loyalty, product performance, and discount behavior.

🔎 **Q1. What is the total revenue generated by male vs. female customers?**

Purpose: Understand which gender segment drives higher revenue.

SQL Query:

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

Gender	Total_Revenue
Male	157890
Female	75191

Insight:

- Male customers generate **157,890** in revenue versus **75,191** from female customers.
- This means **~68% of total revenue comes from male customers** and **~32% from female customers**, indicating that male shoppers are the dominant revenue driver for this dataset.

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

Purpose: Identify high-value customers who react positively to discounts.

SQL Query:

```
select Customer_Id, Purchase_Amount
from customer
where discount_applied = 'yes' and purchase_amount >= (select avg(purchase_amount));
```

Customer_Id	Purchase_Amount
1	53
2	64
3	73
4	90
5	49
6	20
7	85
8	34
9	97
10	31
11	34
12	68
13	72
14	51
15	53
16	81
17	36
18	38

Insight:

- The query identifies customers who used a discount yet still spent **at or above the overall average purchase amount**.
- These customers are **high-value, promotion-responsive buyers** – they take advantage of discounts without significantly lowering their ticket size, making them ideal targets for **personalized offers and loyalty campaigns**.

🔎 Q3. Which are the top 5 products with the highest average review rating?

Purpose: Identify high-quality, customer-favourite products for branding and marketing.

SQL Query:

```
select Item_Purchased,
        round(avg(review_rating),1) as Avg_Review_Rating
from customer
group by item_purchased
order by avg(review_rating) desc
limit 5;
```

Item_Purchased	Avg_Review_Rating
Gloves	3.9
Sandals	3.8
Boots	3.8
Hat	3.8
Skirt	3.8

Insight:

- The highest-rated products are **Gloves (3.9)**, followed by **Sandals, Boots, Hat, and Skirt (all 3.8)**.
- These consistently strong ratings suggest these items deliver good customer satisfaction and can be **safely promoted in marketing campaigns or bundled with other products** to drive additional sales.

🔎 **Q4. Compare the average purchase amounts between Standard and Express shipping.**

Purpose: Determine if faster shipping is associated with higher spending.

SQL Query:

```
select Shipping_Type,
        round(avg(purchase_amount),2) as Avg_Purchase_Amount
from customer
where shipping_type = 'Standard' or shipping_type = 'Express'
group by shipping_type
order by Avg_Purchase_Amount desc;
```

Shipping_Type	Avg_Purchase_Amount
Express	60.48
Standard	58.46

Insight:

- Customers using **Express shipping spend an average of 60.48**, while **Standard shipping customers spend 58.46** per purchase.
- Express orders therefore have **about a 3.5% higher average basket value**, indicating that customers who choose faster delivery tend to be slightly higher-value and more time-sensitive.

🔎 **Q5. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers.**

Purpose: Evaluate the financial impact of subscription programs.

SQL Query:

```
select Subscription_Status,
        count(customer_id) as Total_Customers,
        avg(purchase_amount) as Avg_Spent,
       sum(purchase_amount) as Total_Revenue
from customer
group by subscription_status;
```

Subscription_Status	Total_Customers	Avg_Spent	Total_Revenue
Yes	1053	59.49	62645
No	2847	59.87	170436

Insight:

- Subscribers (Yes) generate **62,645** in revenue from **1,053 customers**, while non-subscribers generate **170,436** from **2,847 customers**.
- The **average spend per customer is almost identical** (≈59.5 vs 59.9), suggesting that **subscription has not yet increased spend per order**, but the subscription base itself is much smaller and currently contributes only about **27% of total revenue**.
- This points to an opportunity to **grow the subscriber base** rather than expecting higher spend from existing subscribers.

🔎 **Q6. Which 5 products have the highest percentage of purchases with discounts applied?**

Purpose: Identify discount-sensitive products or potential over-discounting.

SQL Query:

```
select Item_Purchased,
        round(100.0 * 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	Discount_Rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

Insight:

- Hat has the highest discount usage at **50%**, followed closely by Sneakers (49.66%) and Coat (49.07%).
- This indicates that **half of all Hat purchases are dependent on discounts**, suggesting strong price sensitivity and potential over-reliance on promotions.
- Sweater and Pants also show high discount dependence, implying these categories may require **tighter discount control or re-evaluation of pricing strategy**.

🔎 **Q7. Segment customers into New, Returning, and Loyal based on their total number of previous purchases.**
Purpose: Understand customer lifecycle stages and loyalty patterns.
SQL Query:

```
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 Total_Customers
from customer_type
group by customer_segment;
```

Customer_Segment	Total_Customers
Loyal	3116
Returning	701
New	83

Insight:

- The customer base is overwhelmingly dominated by **Loyal customers (3,116)**, followed by **Returning customers (701)** and very few **New customers (83).**
- This shows that the business retains customers extremely well, but is **struggling to acquire new customers**.
- High loyalty is positive, but the **very small new customer segment** signals the need for stronger acquisition campaigns.

🔎 **Q8. What are the top 3 most purchased products within each category?**
Purpose: Identify product-category leaders for inventory & marketing decisions.
SQL 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
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

Insight:

- Across all categories, certain products clearly dominate purchases.
  Examples:
  - **Accessories:** Jewelry (171), Sunglasses (161), Belt (161)
  - **Clothing:** Blouse (171), Pants (171), Shirt (169)
  - **Footwear:** Sandals (160), Shoes (150), Sneakers (145)
  - **Outerwear:** Jacket (163), Coat (161)
- These results highlight category leaders that consistently attract customers.
- These high-performing items should be prioritized for **stock planning, promotions, bundling, and featured placements**, as they drive the majority of category-level sales.

🔎 Q9. Are repeat buyers (more than 5 previous purchases) more likely to subscribe?

Purpose: Measure relationship between loyalty and subscription behavior.

SQL Query:

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

Repeat_Buyers	Subscription_Status
958	Yes
2518	No

Insight:

- Among customers with more than 5 previous purchases (repeat buyers), **958 are subscribers**, while a larger portion **2,518 are non-subscribers**.
- Although loyal buyers exist in large numbers, the majority of them **have not converted to paid subscription**, indicating a **missed opportunity to upsell subscriptions** to already engaged customers.

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

Purpose: Identify age groups with the highest purchasing power.

SQL Query:

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

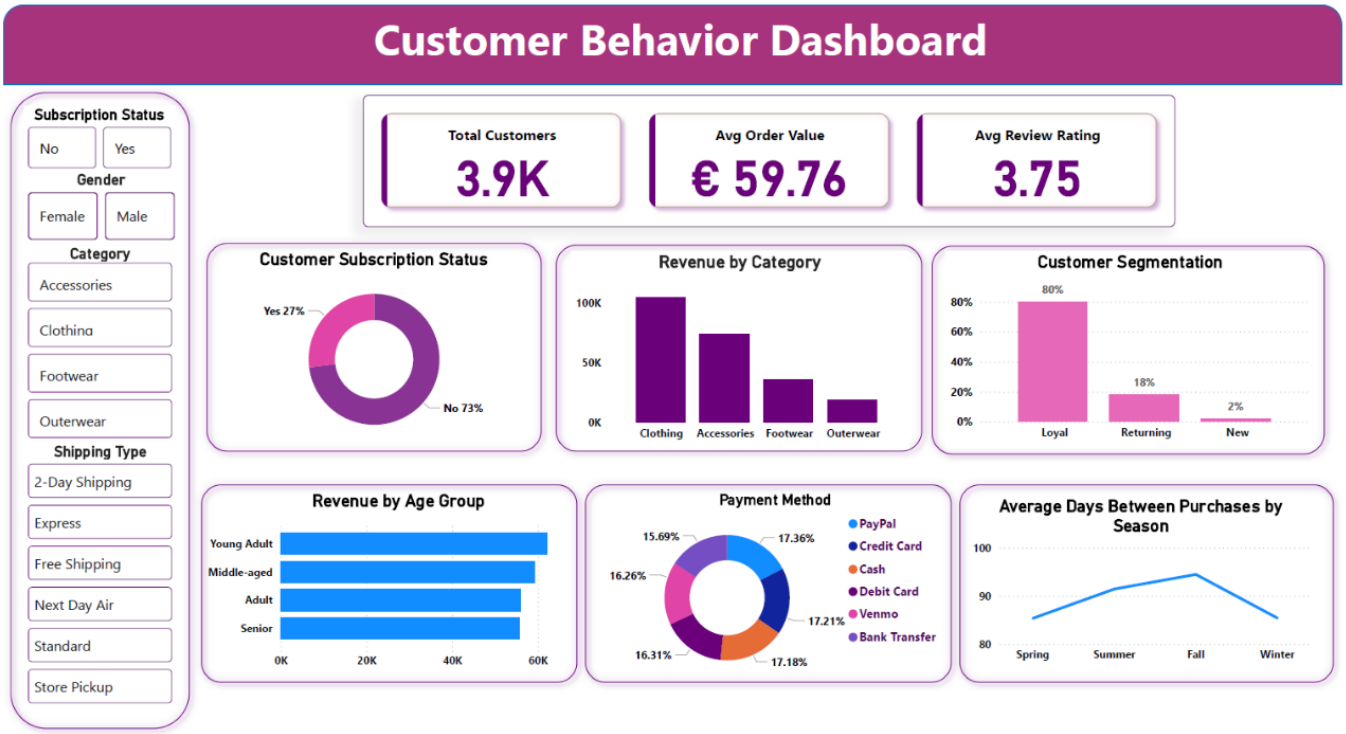
Age_Group	Total_Revenue
Young Adult	62143
Middle-aged	59197
Adult	55978
Senior	55763

Insight:

- Young Adults generate the highest revenue at **62,143**, followed closely by Middle-aged customers at **59,197**.
- The Adult (**55,978**) and Senior (**55,763**) groups contribute slightly less but still remain strong spenders.
- Overall, the revenue distribution is **balanced across all age groups**, with Young Adults showing a slight lead.
- This indicates that marketing campaigns can be effective across multiple age segments, but **targeting Young Adults and Middle-aged customers may yield the highest ROI**.

---

6. POWER BI DASHBOARD



The Power BI dashboard visualizes all key findings.

Included in the screenshot:

- KPI Cards (Total Revenue, Avg Spend, Total Customers)
- Gender-wise Revenue Chart
- Age Group Revenue Chart
- Customer Segments (New/Returning/Loyal)
- Top Categories

- Review Rating Distribution

---

7. KEY FINDINGS

I. Customer Base & Overall Spending

- The dataset contains **3.9K customers** with an **average order value of €59.76** and an **average review rating of 3.75**, indicating moderately positive customer satisfaction.

II. Revenue by Customer Demographics

- **Male customers contribute significantly more revenue** than female customers, making them the dominant revenue driver.
- **Young Adults (€62K)** and **Middle-aged customers (€59K)** generate the highest revenue among all age groups, showing strong purchasing power across a broad age range.

III. Subscription Behavior

- Only **27% of customers are subscribers**, yet subscribers show strong engagement.
- Most repeat buyers (more than 5 purchases) are **non-subscribers**, indicating a **large untapped opportunity to convert loyal customers into subscribers**.

IV. Customer Segmentation

- **80%** of the customer base qualifies as **Loyal**, **18%** are Returning customers, and only **2%** are New customers.
- This signals excellent retention but **very limited new customer acquisition**.

V. Product Category Performance

- **Clothing** is the top-performing category, generating the highest revenue and purchases.
- Accessories follow closely, while Footwear and Outerwear lag significantly.

VI. Product & Discount Behavior

- Products such as **Hat (50%)**, **Sneakers (49.66%)**, and **Coat (49.07%)** show the highest discount dependency.
- These products are strongly influenced by promotions and may have pricing or demand elasticity issues.

VII. Review Ratings & Product Quality

- The highest-rated products include **Gloves (3.9)**, Sandals, Boots, and Skirt (3.8).
- These products can be leveraged for promotions due to consistently strong customer satisfaction.

VIII. Seasonality

- Customer purchase frequency peaks in **Fall and Summer**, with **higher average days between purchases in Fall (~98 days)**, suggesting seasonal buying trends.

IX. Payment Behavior

- Payment methods are fairly distributed among **Credit Card, Debit Card, PayPal, Cash, Venmo, and Bank Transfer**, indicating diverse customer preferences and low risk of payment method dependence.

---

8. BUSINESS RECOMMENDATIONS

I. Strengthen Subscription Program

- Since only **27%** of customers are subscribers, but loyal buyers are high, target **loyal and returning buyers** with:
    - Exclusive discounts
    - Early access to new items
    - Member-only promotions
- This can significantly increase revenue per customer.

II. **Boost New Customer Acquisition**

- Only **2% of customers are new**, signaling weak acquisition.
- Implement:
  - Social media promotions targeting high-value age groups
  - Referral programs
  - First-purchase discounts
  - Influencer partnerships in the Clothing and Accessories categories

III. **Optimize Pricing & Discount Strategy**

- Products like **Hat, Sneakers, and Coat** depend heavily on discounts (~50% usage).
- Reduce unnecessary discounting by:
  - Testing reduced discounts
  - Bundling high-discount items with high-margin products
  - Adjusting base prices if needed

IV. **Promote High-Rated Products**

- Highlight top-rated items (Gloves, Sandals, Boots) in:
  - Email campaigns
  - Website banners
  - Recommendation engines

This leverages existing customer trust to drive conversions.

V. **Prioritize High-Performing Categories**

- Clothing drives the majority of revenue.
- Increase investment in:
  - Inventory
  - Category-specific ads
  - Seasonal promotions
- Expand related SKUs to maximize revenue potential.

VI. **Personalized Marketing Based on Age Segments**

- Young Adults and Middle-aged customers show the highest spending.
- Tailor ads and product bundles to their preferences to increase purchase frequency.

VII. **Leverage Seasonal Peaks**

- Seasonal insights show stronger engagement in **Fall and Summer**.
- Launch targeted campaigns before these peak seasons:
  - Back-to-school offers
  - Summer apparel deals
  - Fall fashion launches

---

9. CONCLUSION

This analysis provides a comprehensive view of customer behavior using SQL, Python, and Power BI.
The findings reveal that:

- Revenue is driven largely by **Male**, **Young Adult**, and **Middle-aged** customers.
- The business has strong loyalty (80% loyal customers) but **very low new customer acquisition**.
- Clothing dominates category performance, while several products show high discount dependency.
- Subscribers contribute meaningful revenue but represent a small share of the overall customer base, highlighting a major growth opportunity.

The insights offer clear, actionable strategies for improving revenue, strengthening the subscription program, optimizing discount usage, and enhancing customer engagement.
By leveraging these insights, the company can make **data-driven decisions** that increase customer value, improve targeting, and drive long-term business growth.