Customer Shopping Behaviour Analysis

Sangotade Idowu

December 24, 2025

1 Project Overview

This project analyzes customer shopping behaviour using transactional retail data consisting of 3,900 purchase records across multiple product categories. The primary objective is to uncover insights related to spending patterns, customer segmentation, product preferences, discount effectiveness, and subscription behaviour. These insights are intended to support data-driven business decisions related to marketing strategy, customer retention, and revenue optimization.

The analysis follows a complete analytics workflow, beginning with exploratory data analysis and data cleaning in Python, followed by structured querying in PostgreSQL, and concluding with interactive data visualization using Power BI.

2 Dataset Summary

The dataset used in this analysis contains transactional-level customer shopping data with the following characteristics:

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

2.1 Key Feature Groups

- **Customer Demographics:** Age group, gender, location, subscription status
- **Purchase Details:** Item purchased, product category, purchase amount, season, size, color
- **Shopping Behaviour:** Discount applied, promo code usage, previous purchases, purchase frequency, review rating, shipping type

2.2 Data Quality

Exploratory data analysis identified 37 missing values in the *review_rating* column. These were addressed during the Python data cleaning phase. No duplicate rows were found, and all categorical variables were standardized to ensure consistency before loading into the database.

3 Exploratory Data Analysis and Data Preparation

Initial exploratory data analysis (EDA) was conducted in Python using Pandas and Matplotlib. This stage involved:

- Inspecting data structure and data types
- Identifying missing values and handling them appropriately
- Validating numerical ranges for purchase amounts and ratings
- Verifying category distributions across products and demographics

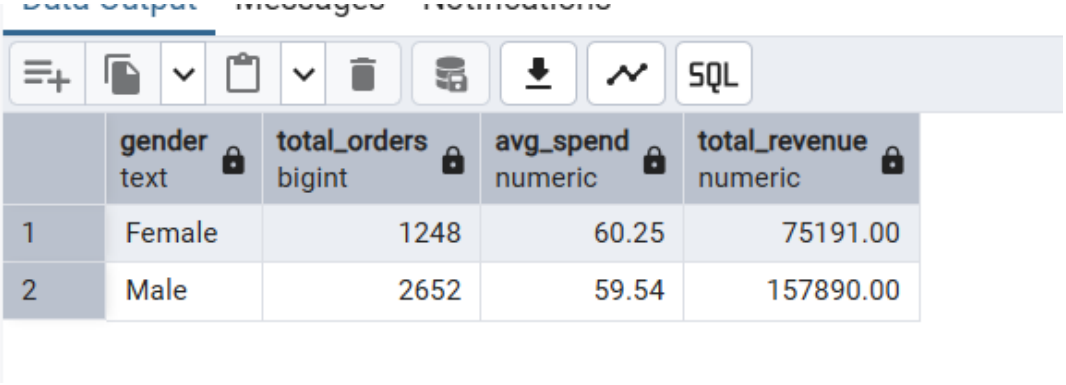
Following EDA, the cleaned dataset was exported and loaded into a PostgreSQL database. All subsequent analytical queries were executed using SQL to ensure reproducibility and scalability.

4 Business Questions and Analysis

4.1 Q1: Revenue Contribution by Gender

Business Question: Is there a meaningful difference in revenue contribution between male and female customers, and is this difference driven by higher spending or higher transaction volume?

SQL Query Result:



	gender text	total_orders bigint	avg_spend numeric	total_revenue numeric
1	Female	1248	60.25	75191.00
2	Male	2652	59.54	157890.00

Figure 1: Revenue and order distribution by gender

Interpretation: Male customers generate significantly higher total revenue compared to female customers. This difference is primarily driven by a higher number of transactions rather than substantially higher average spending per order, indicating volume-based revenue dominance rather than value-based purchasing.

4.2 Q2: Impact of Discounts on High-Value Purchases

Business Question: Do discounts attract high-value purchases, or are they primarily used for low-value transactions?

SQL Query Result:

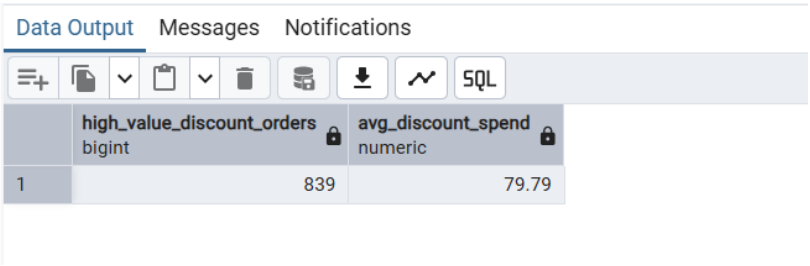


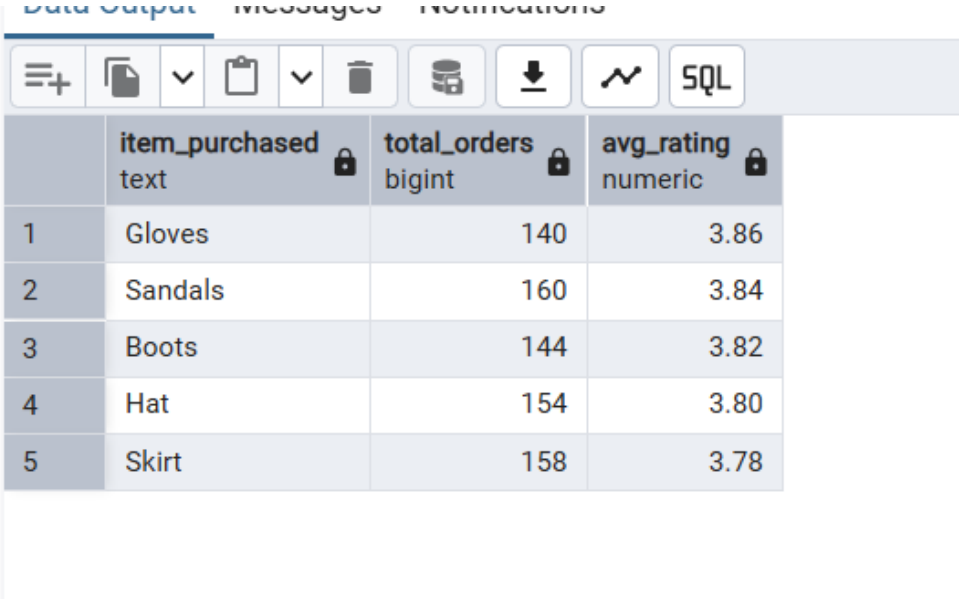
Figure 2: High-value orders with discounts applied

Interpretation: A notable number of transactions with discounts exceed the average purchase amount. This suggests that discounts are not solely used to drive low-value sales, but can also successfully incentivize higher-value purchases.

4.3 Q3: Products with High Satisfaction and Volume

Business Question: Which products receive consistently high customer satisfaction, and do these products also have significant purchase volume?

SQL Query Result:



	item_purchased text	total_orders bigint	avg_rating numeric
1	Gloves	140	3.86
2	Sandals	160	3.84
3	Boots	144	3.82
4	Hat	154	3.80
5	Skirt	158	3.78

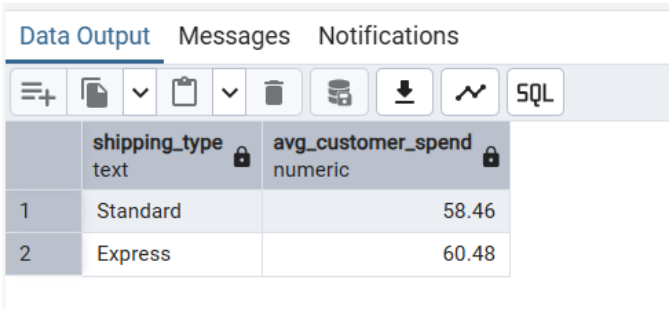
Figure 3: Top-rated products by average review score

Interpretation: Products such as gloves, sandals, and boots achieve both high average review ratings and strong purchase volumes. This alignment indicates successful product-market fit and suggests opportunities for focused promotion or inventory expansion.

4.4 Q4: Shipping Speed and Customer Spending

Business Question: Does faster shipping correlate with higher customer spending?

SQL Query Result:



	shipping_type text	avg_customer_spend numeric
1	Standard	58.46
2	Express	60.48

Figure 4: Average spending by shipping type

Interpretation: Customers using express shipping exhibit higher average spending than those selecting standard shipping. This suggests that convenience-oriented customers may also be more willing to spend, supporting premium shipping offerings.

4.5 Q5: Subscription Status and Customer Value

Business Question: How does subscription status influence customer lifetime value?

SQL Query Result:

Data Output | Messages | Notifications

	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

Figure 5: Spending and revenue by subscription status

Interpretation: Although non-subscribers contribute more total revenue due to larger population size, subscribers demonstrate comparable average spending. This highlights potential growth opportunities through improved subscription conversion strategies.

4.6 Q6: Discount Dependency by Product

Business Question: Which products rely most heavily on discounts to drive sales?

SQL Query Result:

	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

Figure 6: Products with highest discount usage rates

Interpretation: Items such as hats and sneakers exhibit high discount dependency, suggesting greater price sensitivity. These products may benefit from optimized pricing strategies or targeted promotional campaigns.

4.7 Q7: Customer Lifecycle Segmentation

Business Question: How is the customer base distributed across lifecycle stages?

SQL Query Result:

Data Output | Messages | Notifications

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

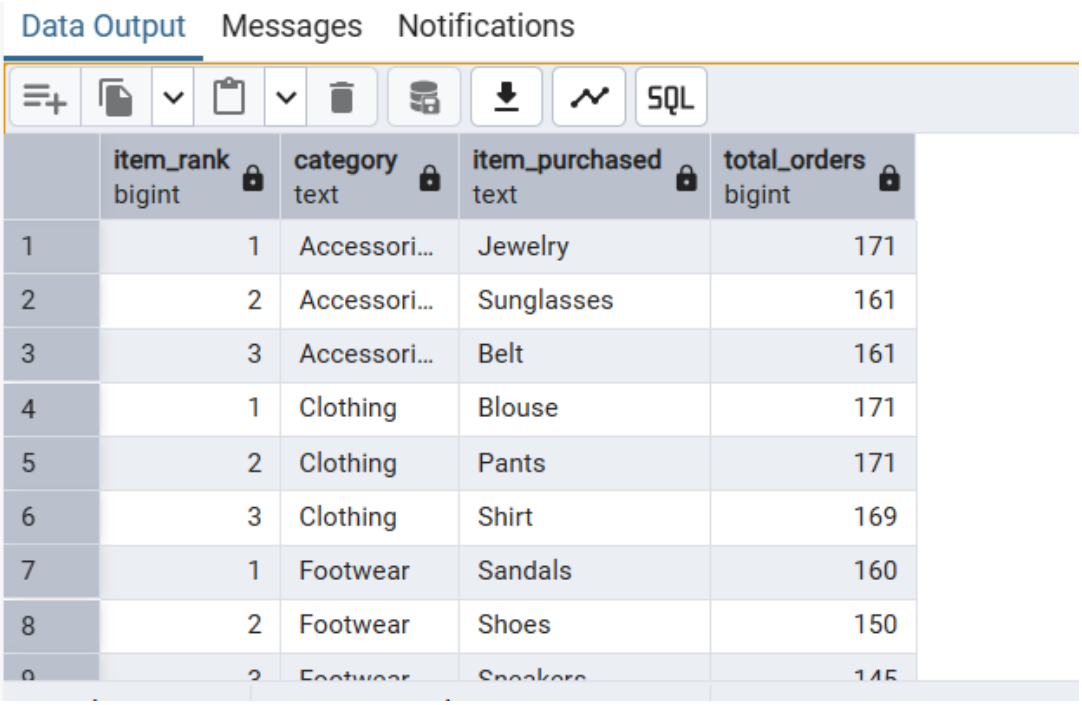
Figure 7: Customer distribution by lifecycle segment

Interpretation: The majority of customers fall into the loyal segment, indicating strong retention. However, the relatively small number of new customers highlights the need for enhanced acquisition strategies.

4.8 Q8: Top Products by Category

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

SQL Query Result:



	item_rank bigint	category text	item_purchased text	total_orders 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

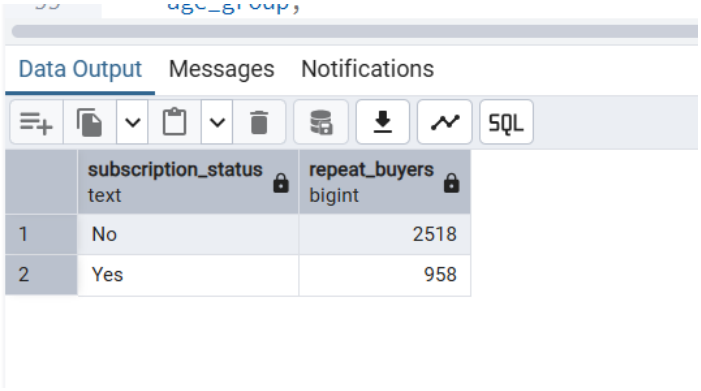
Figure 8: Top products by category

Interpretation: Each category shows clear top-performing products, which can inform inventory prioritization and targeted marketing efforts.

4.9 Q9: Repeat Buyers and Subscription Likelihood

Business Question: Are repeat buyers more likely to subscribe?

SQL Query Result:



	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

Figure 9: Subscription status among repeat buyers

Interpretation: Repeat buyers are more frequently non-subscribers, suggesting untapped potential to convert high-engagement customers into subscribers.

4.10 Q10: Revenue Contribution by Age Group

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

SQL Query Result:

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

Figure 10: Revenue distribution by age group

Interpretation: Young adults and middle-aged customers contribute the highest revenue, indicating that these demographics should be prioritized in marketing and product development strategies.

5 Power BI Dashboard Insights

To complement SQL-based analysis, an interactive Power BI dashboard was developed to provide a holistic overview of customer behaviour.

Figure 11: Customer Behaviour Power BI Dashboard

The dashboard highlights key metrics including total customers (3.9K), average purchase amount ($59.76), and average review rating (3.75). Visualizations reveal that clothing generates the highest revenue and sales volume, while young adults represent the most valuable age group. Subscription adoption remains relatively low at 27%, reinforcing findings from the SQL analysis.

6 Conclusion

This project demonstrates how combining Python-based data preparation, SQL-driven analysis, and business intelligence dashboards can deliver actionable insights. Key opportunities include improving subscription conversion among loyal customers, optimizing discount strategies for price-sensitive products, and targeting high-value demographics through tailored marketing initiatives.

Future work may include predictive modeling for churn or customer lifetime value estimation.