Customer Shopping Behavior Analysis

Cover Page

Project Title: Customer Shopping Behavior Analysis
Subtitle: Understanding Consumer Patterns to Improve Engagement, Sales, and Retention

Prepared by: Abdou Salam Sisawo
Domain: Data Science / Business Analytics
Institution: Albukhary International University
Tool Stack: Python (Pandas, NumPy), MySQL, Power BI

Abstract / Executive Summary

This report presents a comprehensive analysis of customer shopping behavior using transactional retail data, with the objective of supporting data-driven decision-making in a modern retail environment. Understanding how customers interact with products, discounts, and subscription services is essential for improving sales performance, enhancing customer satisfaction, and building long-term customer loyalty. As retail organizations increasingly rely on data to guide strategic decisions, the ability to extract meaningful insights from consumer data has become a critical competitive advantage.

The primary aim of this analysis is to identify purchasing patterns, customer segments, and the key factors that influence consumer decision-making and repeat purchases. The study focuses on important variables such as subscription status, discount application, product categories, purchase frequency, and customer review ratings. By examining these factors, the analysis seeks to answer the overarching business question: *How can consumer shopping data be leveraged to identify trends, improve customer engagement, and optimize marketing and product strategies?*

To achieve these objectives, a structured analytical workflow was adopted using a combination of industry-relevant tools. Python was utilized for data preparation and preprocessing tasks, including data cleaning, handling missing values, standardizing variables, and creating derived features suitable for analysis. This step ensured that the dataset was accurate, consistent, and ready for further exploration. The processed data was then stored in a MySQL database, where structured queries were executed to aggregate information, segment customers, and compute key performance indicators such as total revenue, average purchase amounts, and customer counts across different categories.

Power BI was subsequently used to visualize the analytical results through interactive dashboards. These dashboards present insights in an intuitive and accessible manner, enabling stakeholders to quickly assess customer subscription distribution, purchasing trends, and product performance. The use of visual analytics enhances understanding by transforming complex datasets into clear and actionable representations.

Overall, this report demonstrates how the integration of Python, SQL, and Power BI can transform raw transactional data into valuable business insights. The findings provide practical recommendations for improving customer engagement, refining discount and subscription strategies, and optimizing product offerings. By leveraging these insights, retail organizations can enhance marketing effectiveness, strengthen customer relationships, and support sustainable business growth.

Customer Shopping Behavior Analysis.....i
Cover Page.....ii
Abstract / Executive Summary.....iii
2. Business Problem Statement.....1
3.1 Data Source.....2
3.2 Key Variables.....2
4. Methodology.....2
4.1 Data Preparation (Python).....2
4.2 Database Design & SQL Analysis.....3
5. Exploratory Data Analysis (EDA).....7
6. Key Insights & Findings.....8
7. Dashboard Interpretation.....8
8. Business Recommendations.....10
9. Limitations & Assumptions.....10
10. Conclusion.....10
12. Appendices.....11
Appendix A: SQL Queries.....11
Appendix B: Power BI Dashboard (Full View).....12

1. Introduction

Understanding customer behavior is critical for retail organizations seeking to improve sales performance, enhance customer satisfaction, and build long-term customer loyalty in an increasingly competitive market. As consumer preferences continue to evolve, retailers must gain deeper insights into how customers interact with products, pricing strategies, and promotional offers in order to remain relevant and responsive to market demands. Customer behavior analysis enables organizations to move beyond intuition-based decisions and adopt evidence-based strategies that are informed by real purchasing data.

This project analyzes consumer shopping data to uncover meaningful trends and behavioral patterns that influence purchasing decisions and repeat buying behavior. By examining transactional data across multiple dimensions such as product categories, subscription status, discounts, and purchasing frequency, the study aims to identify actionable insights that can support data-driven marketing initiatives and product optimization strategies. These insights help businesses better understand their customers, tailor offerings to specific customer segments, and design targeted interventions that enhance engagement, retention, and overall business performance.

2. Business Problem Statement

A leading retail company has observed notable changes in purchasing patterns across customer demographics, product categories, and sales channels, including both online and offline transactions. These shifts suggest that customer preferences and buying behaviors are evolving, potentially influenced by factors such as pricing strategies, promotional activities, subscription offerings, and overall customer experience. As a result, management recognizes the need to gain a clearer understanding of these emerging patterns in order to respond effectively to changing market conditions.

The management team seeks to identify the key factors that influence customer purchasing decisions and repeat buying behavior. In particular, there is a need to understand how elements such as discounts, subscription status, product preferences, and customer engagement impact overall sales performance and customer loyalty. Without a data-driven understanding of these factors, it becomes challenging for the company to design effective marketing strategies, optimize product offerings, and improve long-term customer retention. This project aims to address these challenges by analyzing customer shopping data to provide actionable insights that support informed strategic decision-making.

Key Business Question:
How can consumer shopping data be leveraged to identify trends, improve customer engagement, and optimize marketing and product strategies?

3. Data Overview

3.1 Data Source

The dataset consists of customer-level transactional data, including purchase history, product details, discounts, subscription status, and review ratings.

3.2 Key Variables

- Customer ID
- Item Purchased
- Category
- Purchase Amount
- Discount Applied (Yes/No)
- Subscription Status
- Review Rating
- Payment Method
- Shipping Type

4. Methodology

4.1 Data Preparation (Python)

Python was used for data cleaning, transformation, and feature engineering. This included handling missing values, standardizing column names, and creating derived variables for analysis.

Data Cleaning & Preparation Code Output

```
# Final check of the cleaned dataset
df.head()
```

[21] ✓ 0.0s

	customer_id	age	gender	item_purchased	category	purchase_amount	location	size	color	season	review_rating
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7

```
# Final check of the cleaned dataset
df.head()
```

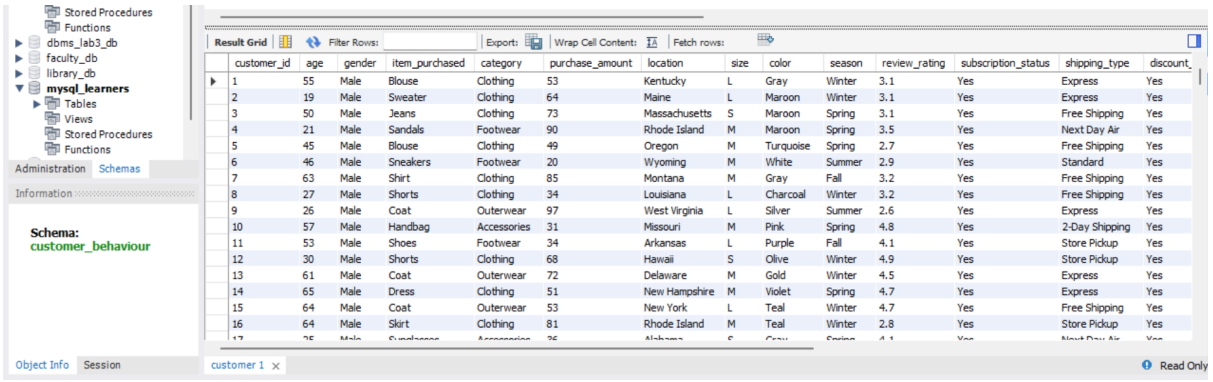
[21] ✓ 0.0s Python

...iew_rating	subscription_status	shipping_type	discount_applied	previous_purchases	payment_method	frequency_of_purchases	age_group	purchase_frequency_days
3.1	Yes	Express	Yes	14	Venmo	Fortnightly	Middle-aged	14
3.1	Yes	Express	Yes	2	Cash	Fortnightly	Young Adult	14
3.1	Yes	Free Shipping	Yes	23	Credit Card	Weekly	Middle-aged	7
3.5	Yes	Next Day Air	Yes	49	PayPal	Weekly	Young Adult	7
2.7	Yes	Free Shipping	Yes	31	PayPal	Annually	Middle-aged	365

4.2 Database Design & SQL Analysis

The cleaned dataset was loaded into a MySQL database. SQL queries were written to aggregate data, segment customers, and compute key metrics such as total revenue, average purchase amount, and discount utilization.

Table Schema

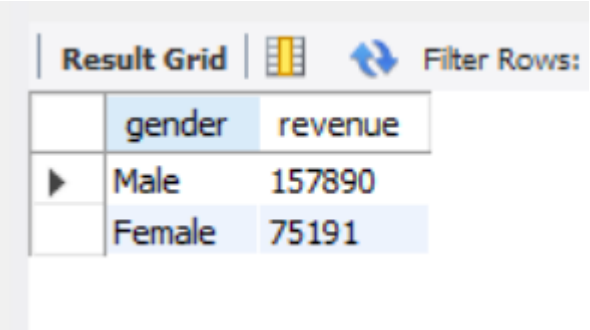


Export: | Wrap Cell Content: | Fetch rows:

shipping_type	discount_applied	previous_purchases	payment_method	frequency_of_purchases	age_group	purchase_frequency_days
Express	Yes	14	Venmo	Fortnightly	Middle-aged	14
Express	Yes	2	Cash	Fortnightly	Young Adult	14
Free Shipping	Yes	23	Credit Card	Weekly	Middle-aged	7
Next Day Air	Yes	49	PayPal	Weekly	Young Adult	7
Free Shipping	Yes	31	PayPal	Annually	Middle-aged	365
Standard	Yes	14	Venmo	Weekly	Middle-aged	7
Free Shipping	Yes	49	Cash	Quarterly	Senior	90
Free Shipping	Yes	19	Credit Card	Weekly	Young Adult	7
Express	Yes	8	Venmo	Annually	Young Adult	365
2-Day Shipping	Yes	4	Cash	Quarterly	Middle-aged	90
Store Pickup	Yes	26	Bank Transfer	Bi-Weekly	Middle-aged	14
Store Pickup	Yes	10	Bank Transfer	Fortnightly	Young Adult	14
Express	Yes	37	Venmo	Fortnightly	Senior	14
Express	Yes	31	PayPal	Weekly	Senior	7
Free Shipping	Yes	34	Debit Card	Weekly	Senior	7
Store Pickup	Yes	8	PayPal	Monthly	Senior	30

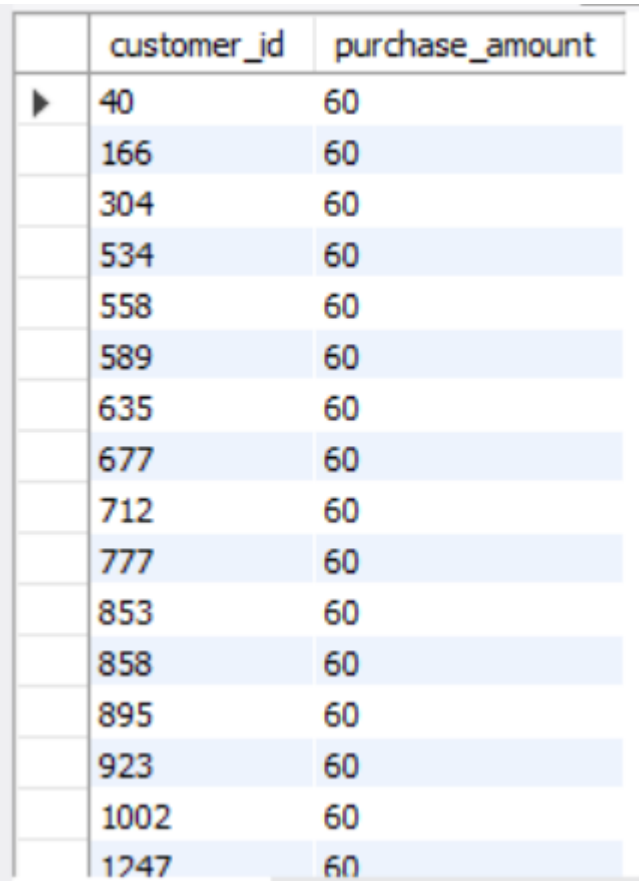
SQL Query Results – Customer Segmentation / Discount Analysis

1. **Revenue by Gender** – Compared total revenue generated by male vs. female customers.


Result Grid | Filter Rows:

gender	revenue
Male	157890
Female	75191

2. **High-Spending Discount Users** – Identified customers who used discounts but still spent above the average purchase amount. **839 rows**



customer_id	purchase_amount
40	60
166	60
304	60
534	60
558	60
589	60
635	60
677	60
712	60
777	60
853	60
858	60
895	60
923	60
1002	60
1247	60

3. **Top 5 Products by Rating** – Found products with the highest average review ratings.

item_purchased	Average Product Rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Skirt	3.78

4. **Shipping Type Comparison** – Compared average purchase amounts between Standard and Express shipping.

Result Grid | Filter Rows:

shipping_type	Average purchase Amount
Express	60.48
Standard	58.46

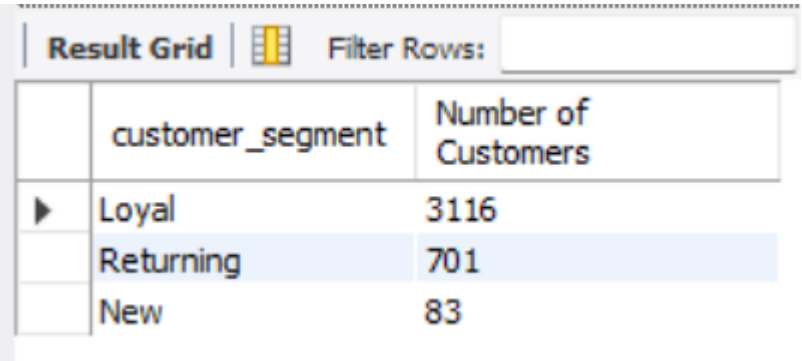
5. **Subscribers vs. Non-Subscribers** – Compared average spend and total revenue across subscription status.

Subscription_status	Total_Customers	revenue	Average Purchase Amount
Yes	1053	62645	59.4919
No	2847	170436	59.8651

6. **Discount-Dependent Products** – Identified 5 products with the highest percentage of discounted purchases.

item_purchased	discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

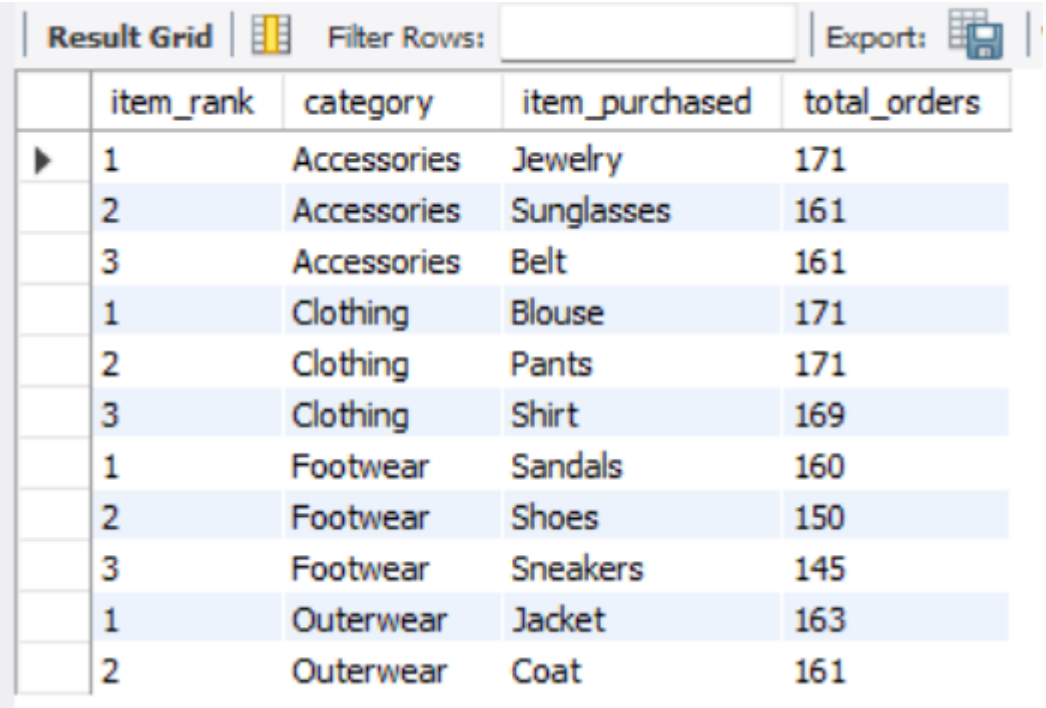
7. **Customer Segmentation** – Classified customers into New, Returning, and Loyal segments based on purchase history.



Result Grid | Filter Rows:

customer_segment	Number of Customers
Loyal	3116
Returning	701
New	83

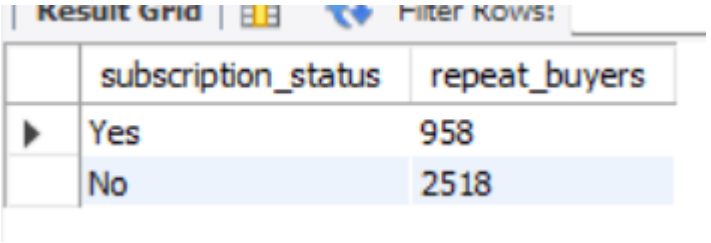
8. **Top 3 Products per Category** – Listed the most purchased products within each category.



Result Grid | Filter Rows: | Export:

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

9. **Repeat Buyers & Subscriptions** – Checked whether customers with >5 purchases are more likely to subscribe.



Result Grid | Filter Rows:

subscription_status	repeat_buyers
Yes	958
No	2518

10. **Revenue by Age Group** – Calculated total revenue contribution of each age group.

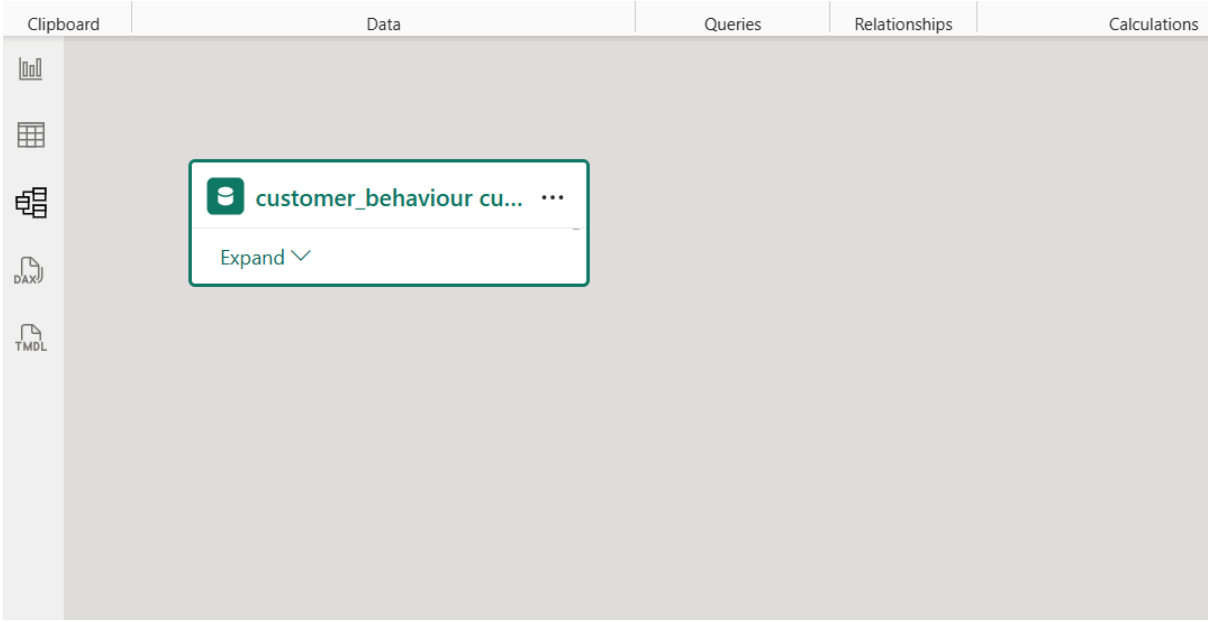
Result Grid | Filter Rows:

age_group	Revenue_Total
Young Adult	62143
Middle-aged	59197
Adult	55978
Senior	55763

4.3 Visualization Approach (Power BI)

Power BI was used to create an interactive dashboard that visually communicates insights. Key visuals include donut charts, bar charts, and summary cards to support managerial decision-making.

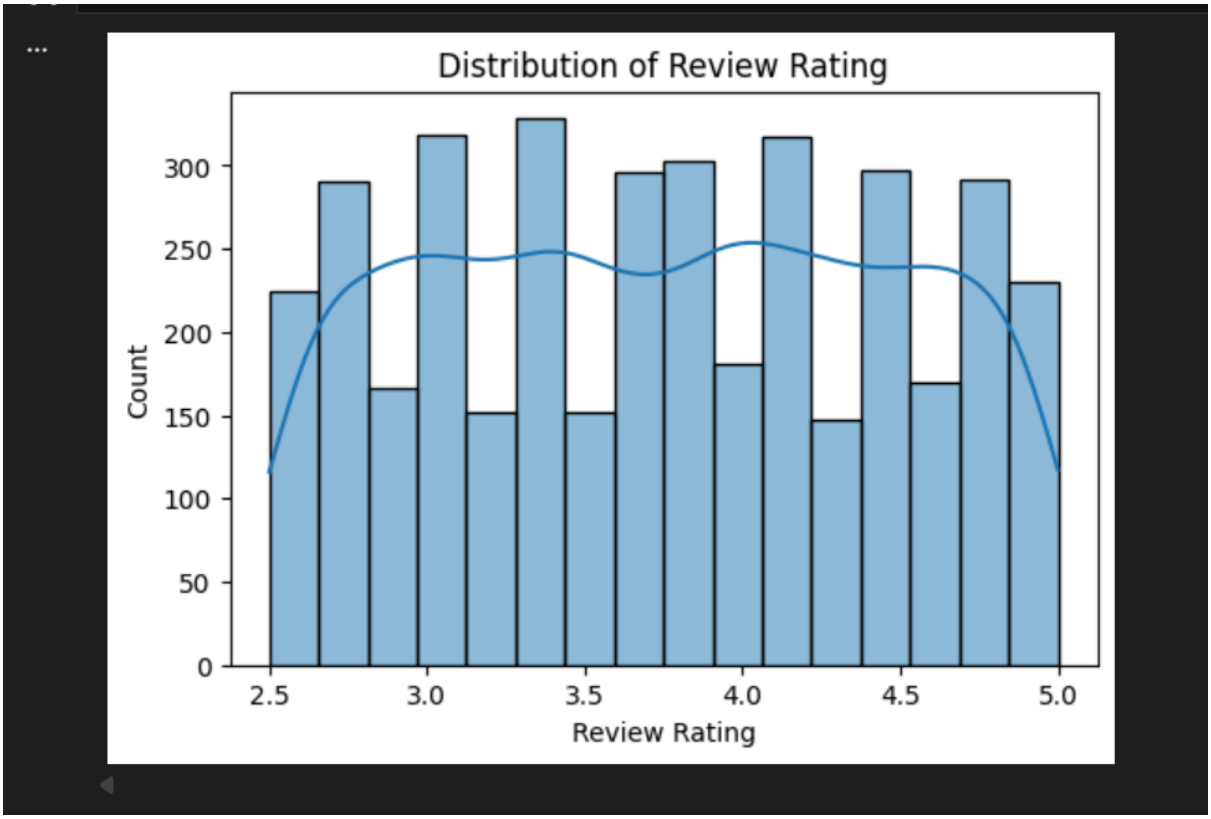
Power BI Data Model View



5. Exploratory Data Analysis (EDA)

Exploratory analysis was conducted to understand the distribution of key variables such as purchase amount, review ratings, and subscription status. Visual inspections helped guide analytical decisions, including aggregation methods and feature encoding.

Figure 1: Distribution of Review Ratings



6. Key Insights & Findings

- Subscribed customers demonstrate higher average spending compared to non-subscribers.
- Certain product categories show stronger discount dependency.
- Review ratings are generally skewed towards higher values, indicating positive customer sentiment.
- Subscription status plays a significant role in repeat purchasing behavior.

Summary of Key Metrics

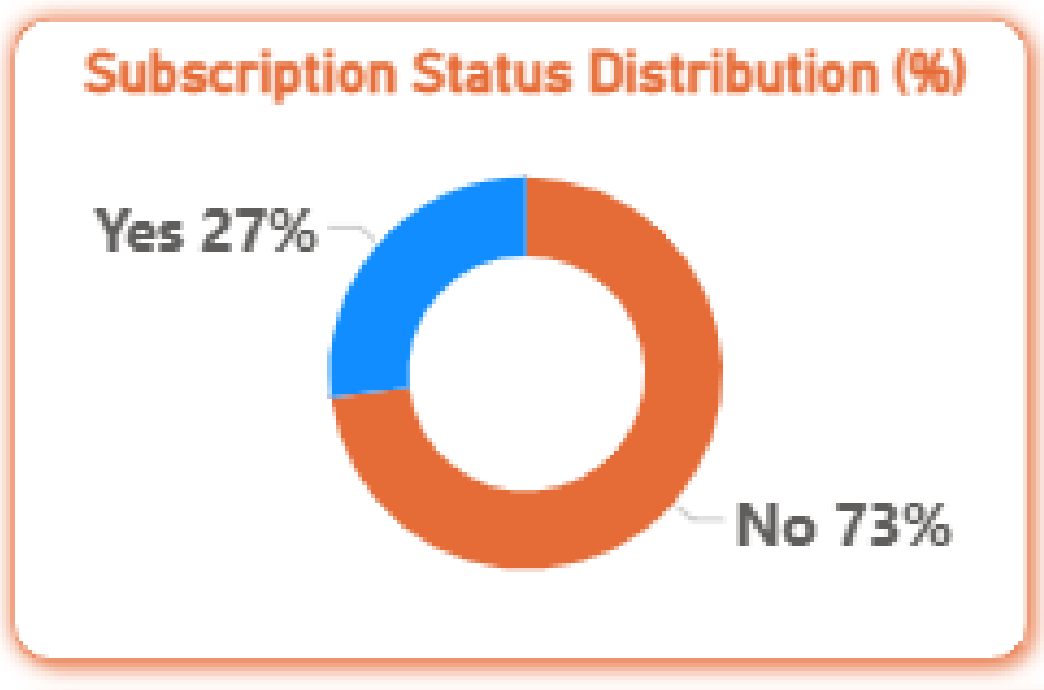
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_customers	total_orders	total_revenue	avg_purchase_amount	subscribed_customers	subscription_rate_percent
3900	3900	233081	59.76	1053	27.00

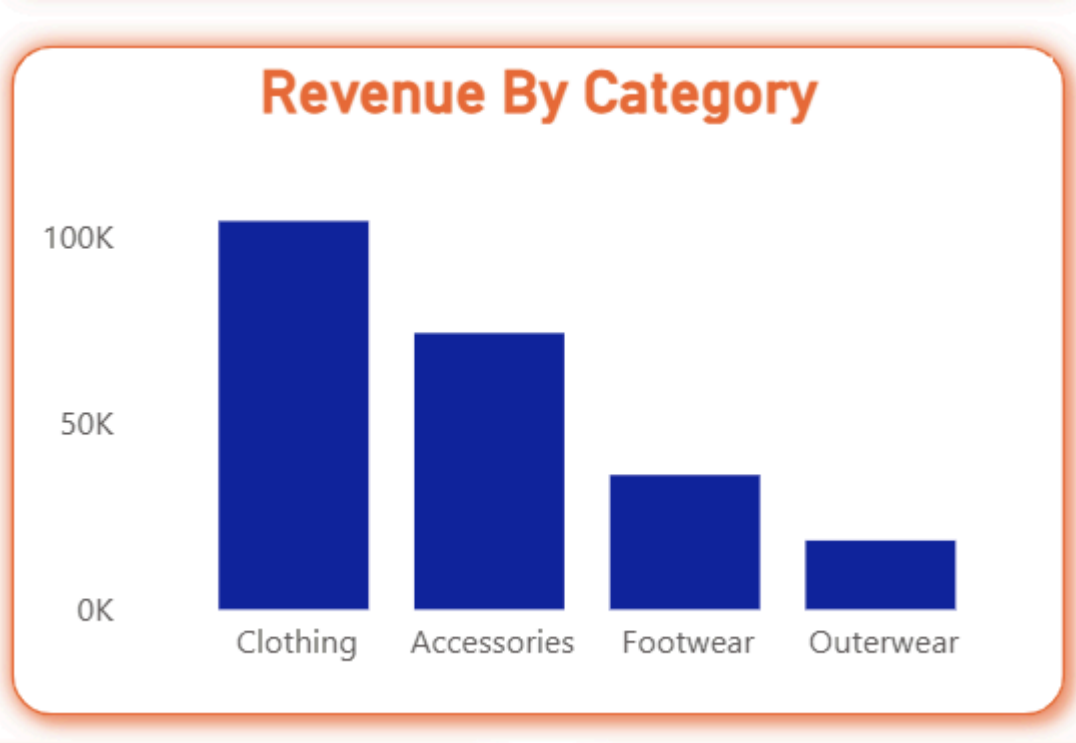
7. Dashboard Interpretation

The Power BI dashboard provides an at-a-glance view of customer behavior. Donut charts illustrate subscription distribution, while bar charts highlight top-performing products and categories.

Subscription Status Distribution Donut Chart



Top Products by Revenue



8. Business Recommendations

Based on the analysis, the following recommendations are proposed:

- Strengthen subscription-based incentives to improve customer retention.
- Optimize discount strategies for high-impact product categories.
- Leverage customer reviews to guide product placement and marketing campaigns.
- Use data-driven segmentation to personalize customer engagement.

9. Limitations & Assumptions

- The dataset represents a limited time period and may not capture long-term seasonal effects.
- Some variables, such as discount application, showed limited variation.
- Behavioral insights are constrained by the available features in the dataset.

10. Conclusion

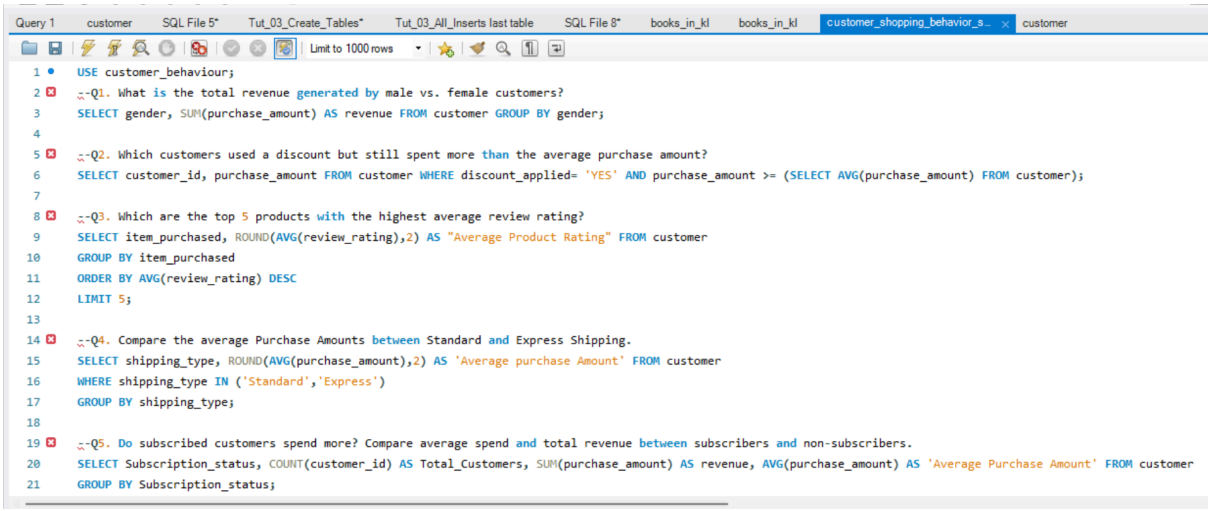
This project demonstrates how integrating Python, SQL, and Power BI can transform raw customer data into actionable business insights. The findings highlight the importance of subscription models, discount strategies, and customer sentiment in shaping purchasing behavior.

11. References

- IBM Data Science Professional Certificate Materials
- MySQL Documentation
- Microsoft Power BI Documentation

12. Appendices

Appendix A: SQL Queries



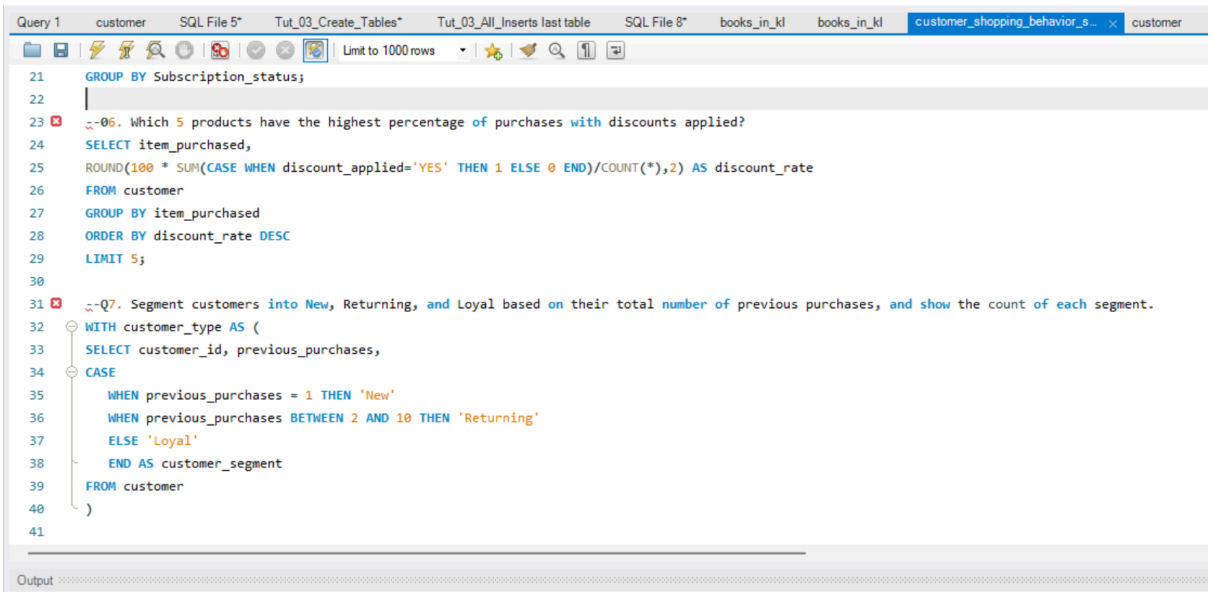
```sql
USE customer_behaviour;
--Q1. What is the total revenue generated by male vs. female customers?
SELECT gender, SUM(purchase_amount) AS revenue FROM customer GROUP BY gender;

--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);

--Q3. Which are the top 5 products with the highest average review rating?
SELECT item_purchased, ROUND(AVG(review_rating),2) AS "Average Product Rating" FROM customer
GROUP BY item_purchased
ORDER BY AVG(review_rating) DESC
LIMIT 5;

--Q4. Compare the average Purchase Amounts between Standard and Express Shipping.
SELECT shipping_type, ROUND(AVG(purchase_amount),2) AS 'Average purchase Amount' FROM customer
WHERE shipping_type IN ('Standard','Express')
GROUP BY shipping_type;

--Q5. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers.
SELECT Subscription_status, COUNT(customer_id) AS Total_Customers, SUM(purchase_amount) AS revenue, AVG(purchase_amount) AS 'Average Purchase Amount' FROM customer
GROUP BY Subscription_status;
```



```sql
GROUP BY Subscription_status;

--Q6. Which 5 products have the highest percentage of purchases with discounts applied?
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;

--Q7. Segment customers into New, Returning, and Loyal based on their total number of previous purchases, and show the count of each segment.
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
)
```

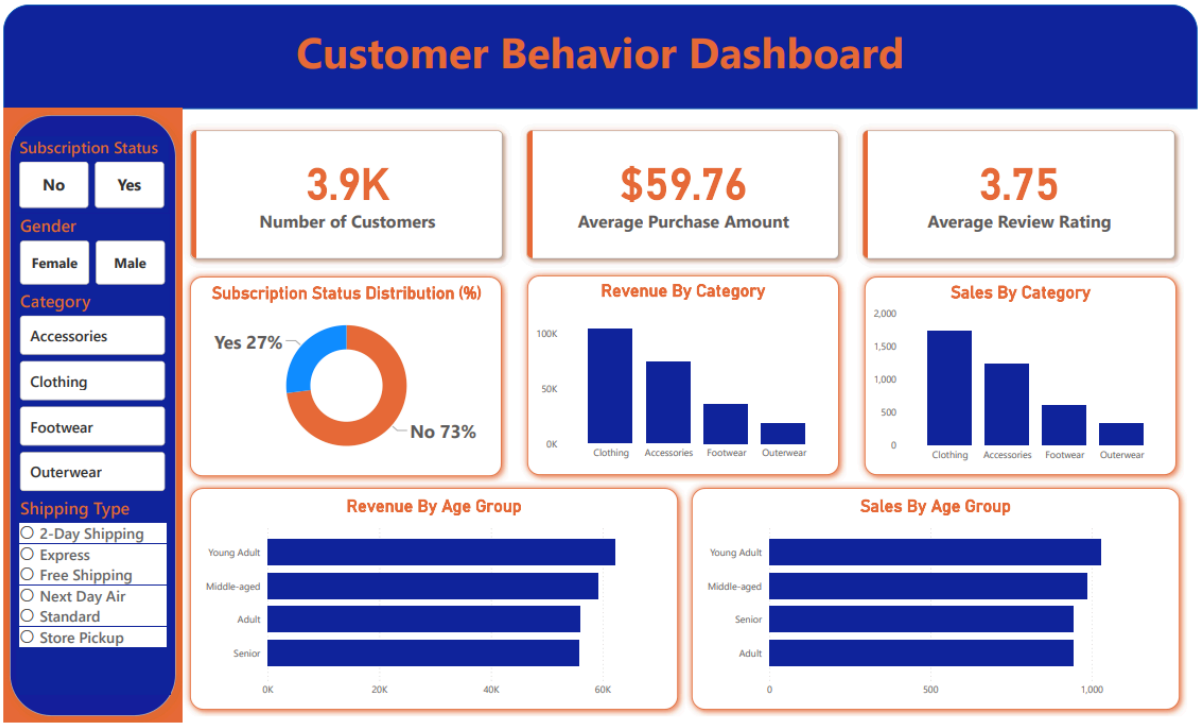
```sql
SELECT customer_segment, COUNT(*) AS "Number of Customers"
FROM customer_type
GROUP BY customer_segment

--Q8. What are the top 3 most purchased products within each 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 item_rank, category, item_purchased, total_orders
FROM item_counts
WHERE item_rank<=3;

--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
```

Appendix B: Power BI Dashboard (Full View)



End of Report