CUSTOMER SHOPPING BEHAVIOR

1.0 Project Overview

This project conducts a comprehensive analysis of customer shopping behavior using a dataset of 3,900 transactions. The goal was to extract actionable insights regarding revenue generation, customer segmentation, product performance, and the impact of marketing strategies like discounts and subscriptions.

The project followed a complete data analytics pipeline:

1. **Data Preprocessing & Cleaning** (Python, Pandas, Jupyter Notebook)
2. **Exploratory Data Analysis (EDA) & Business Intelligence** (MySQL)
3. **Data Visualization & Dashboarding** (Power BI)
4. **Reporting & Documentation** (This Report)

2.0 Data Preprocessing & Feature Engineering

The initial dataset was cleaned and prepared for analysis using a Python script (preprocessing_data.ipynb). The key steps included:

- **Data Loading & Inspection:** The dataset was loaded, and an initial check confirmed 3,900 entries with 18 columns. No duplicate records were found.
- **Handling Missing Values:** 37 missing values in the Review Rating column were imputed using the median rating for each respective product Category, ensuring no bias in the average ratings analysis.
- **Data Transformation:**
    - Column names were standardized to lowercase with underscores for compatibility (e.g., Purchase Amount (USD) became purchased_amount).
- **Feature Engineering:**
    - **Age Groups:** Customers were segmented into four quantile-based groups for more nuanced analysis:
        - Young Adult (18-31)
        - Adult (31-44)
        - Middle Aged (44-57)
        - Old (57-70)

    - **Purchase Frequency:** The categorical frequency_of_purchases column (e.g., "Weekly", "Annually") was mapped to a numerical purchase_frequency_days column, representing the approximate days between purchases. This allows for future analysis of purchase cadence.

- **Data Validation:** A check confirmed that the discount_applied and promo_code_used columns were identical. Consequently, the redundant promo_code_used column was dropped to simplify the dataset.

The cleaned and enhanced dataset was successfully exported to a MySQL database named customer_behavior into a table called customer.

3.0 Exploratory Data Analysis (MySQL Queries)

A series of SQL queries were executed to answer key business questions. The findings are summarized below:

Q1: What is the total revenue generated by male vs female customers?

- **Purpose:** Understand gender-based spending patterns.
- **Insight:** This query provides a direct comparison of total revenue contribution by gender, highlighting which demographic drives more sales.

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

- **Purpose:** Identify high-value customers who are less price-sensitive.
- **Insight:** This segment is highly valuable for targeted marketing, as they are motivated by discounts but have a high spending threshold.

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

- **Purpose:** Identify best-performing products by customer satisfaction.
- **Insight:** These products can be highlighted in marketing materials and used to understand what attributes lead to high customer satisfaction.

Q4: Compare the average purchase amounts between standard and express shipping.

- **Purpose:** Analyze the relationship between shipping choice and spending.
- **Insight:** Customers opting for faster shipping might have a higher average order value, indicating a potential for upselling shipping options.

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

- **Purpose:** Evaluate the financial impact of the subscription program.
- **Insight:** This is a critical metric for assessing customer loyalty program success. It shows whether subscribers have a higher Customer Lifetime Value (CLV).

Q6: Which 5 products have the highest percentage of purchases with discounts applied?

- **Purpose:** Identify products that are most discount-driven.
- **Insight:** This helps in planning promotional strategies and managing inventory for products that rely heavily on discounts to sell.

Q7: Segment customers into new, returning, and loyal based on their total number of previous purchases, and show the count of each segment.

- **Segmentation Logic:**
    - **New Customers:** 1 previous purchase
    - **Returning Customers:** 2-10 previous purchases
    - **Loyal Customers:** 11+ previous purchases
- **Insight:** Understanding the composition of the customer base allows for tailored communication and retention strategies for each segment.

Q8: What are the top 3 most purchased products within each category?

- **Purpose:** Identify best-sellers in each product category (Clothing, Footwear, Accessories).
- **Insight:** Essential for inventory management, sales strategy, and identifying market leaders within each category.

Q9: Are customers who are repeat buyers (More than 5 previous purchases) also likely to subscribe?

- **Purpose:** Investigate the correlation between repeat purchase behavior and subscription uptake.
- **Insight:** If a strong correlation exists, marketing can target repeat buyers for subscription sign-ups to further cement their loyalty.

Q10: What is the revenue contribution of each age group?

- **Purpose:** Understand which age demographic contributes the most to revenue.
- **Insight:** This informs targeted advertising, product development, and marketing messaging to the most valuable age segments.

4.0 Key Findings & Business Implications

1. **Customer Segmentation is Powerful:** The project successfully segments customers by loyalty (New, Returning, Loyal) and age (Young Adult, Adult, Middle Aged, Old), enabling highly targeted marketing campaigns.
2. **Data Quality is Crucial:** The cleaning process, especially handling missing review ratings, was essential to ensure the accuracy of all subsequent analysis, particularly for product ratings.
3. **Discounts Drive Specific Behaviors:** The analysis can pinpoint which products are most discount-sensitive and identify customers who spend above average even with discounts, revealing different purchasing psychographics.
4. **Subscriber Value:** The comparison between subscribers and non-subscribers directly measures the ROI of the subscription program, a key business metric.
5. **Product & Category Performance:** Insights into top-rated and best-selling products per category provide clear directives for marketing and inventory planning.

5.0 Conclusion

This project demonstrates a full-cycle data analysis, from raw data to actionable insights. The cleaned dataset and comprehensive SQL queries provide a strong foundation for understanding customer behavior, optimizing marketing strategies, and driving data-informed business decisions. The code is structured, documented, and ready for review on GitHub.

CUSTOMER BEHAVIOR DASHBOARD

Subscription Status

No | Yes

Gender

Female | Male

Category

Accessories

Clothing

Footwear

Outerwear

Shipping Type

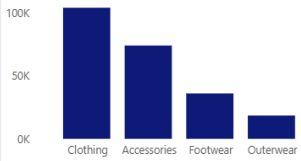
- 2-Day Shipping
- Express
- Free Shipping
- Next Day Air
- Standard
- Store Pickup

3.9K
Number of Customers

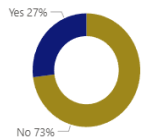
$59.76
Average Purchase Amount

3.75
Average Review Rating

Revenue by Category



Customers by Subscription Status



Sales by Category

Revenue by Age Group

Sales by Age Group