# Customer Shopping Behavior & Sales Analysis

## Project Overview

This project examines customer shopping behavior using transactional data from over 3,900 purchases across multiple product categories. It uncovers key insights into spending patterns, customer segmentation, product preferences, and subscription behavior to support data-driven business decisions.

## Dataset Summary

**Dataset Size:** 3,900 records with 18 variables
**Key Attributes Included:**

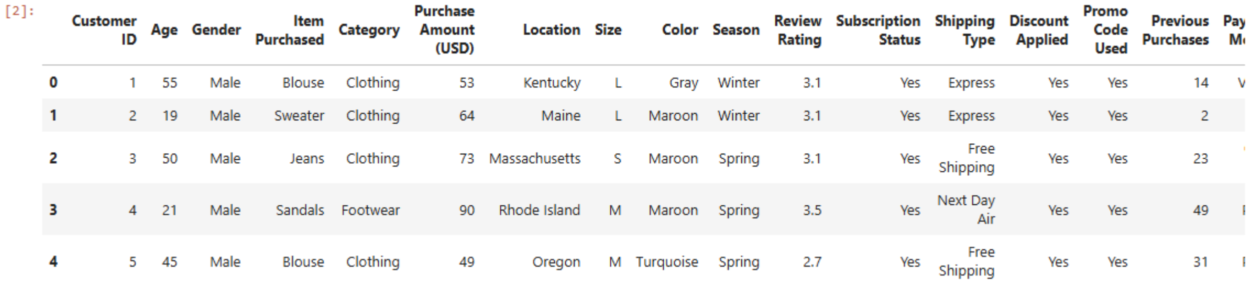
- **Customer Demographics:** Age, Gender, Location, Subscription Status
- **Purchase Information:** Item Purchased, Product Category, Purchase Amount, Season, Size, Color
- **Shopping Behavior Metrics:** Discount Applied, Promo Code Usage, Purchase Frequency, Previous Purchases, Review Ratings, Shipping Type

**Data Quality Note:** 37 missing values identified in the *Review Rating* column

## Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

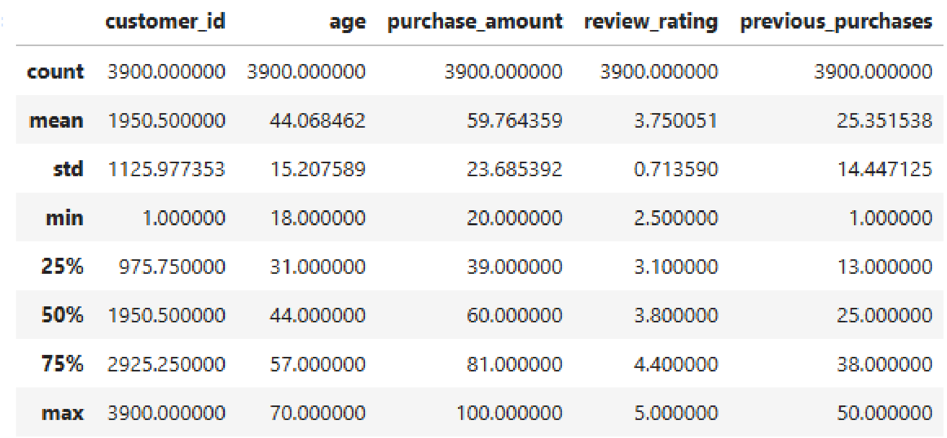
- **Data Loading:** Imported the dataset using pandas.
- **Initial Exploration:**
  - df.head() : to quickly inspecting the data, verifying column names, and checking that the data has been loaded correctly



[2]:

| | 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 | Pay M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 55 | Male | Blouse | Clothing | 53 | Kentucky | L | Gray | Winter | 3.1 | Yes | Express | Yes | Yes | 14 | V |
| **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 | I |
| **4** | 5 | 45 | Male | Blouse | Clothing | 49 | Oregon | M | Turquoise | Spring | 2.7 | Yes | Free Shipping | Yes | Yes | 31 | I |

  - df.info() : to understanding its structure.
  - df.isnull().sum() : to count the number of missing (null/NaN) values in each column of a DataFrame
  - df.describe() : to generates a statistical summary of the data



| | customer_id | age | purchase_amount | review_rating | previous_purchases |
|---|---|---|---|---|---|
| **count** | 3900.000000 | 3900.000000 | 3900.000000 | 3900.000000 | 3900.000000 |
| **mean** | 1950.500000 | 44.068462 | 59.764359 | 3.750051 | 25.351538 |
| **std** | 1125.977353 | 15.207589 | 23.685392 | 0.713590 | 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 |

- **Missing Data Handling:** Identified null values and imputed missing entries in the Review Rating column using the median rating for each product category.

- **Column Standardization:** Renamed columns using ***snake_case*** naming conventions to improve readability and maintain consistent documentation standards.

- **Feature Engineering:**
  - Created an age_group column by categorizing customer ages into meaningful bins.
  - Engineered a purchase_frequency_days column derived from purchase data to measure buying frequency.

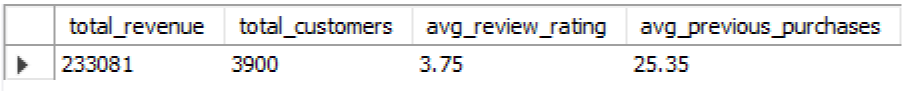
- **Data Consistency Check:** Found that promo_code_used repeated the same information as *discount_applied*, so it was removed.

- **Database Integration:** Connected a Python script to MySQL and loaded the cleaned DataFrame into the database to enable SQL-based analysis.

## Data Analysis using SQL

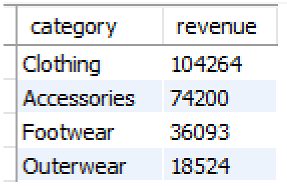
Performed in-depth MySQL analysis using structured queries to uncover answers to critical business questions:

IMP KPI'S (Total revenue, Total Customers, Average Review Rating, Repeat Purchase Rate)



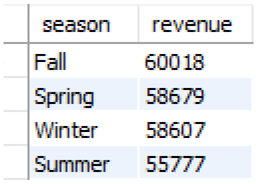
| | total_revenue | total_customers | avg_review_rating | avg_previous_purchases |
|---|---|---|---|---|
| ▶ | 233081 | 3900 | 3.75 | 25.35 |

Q1. Category-wise total sales



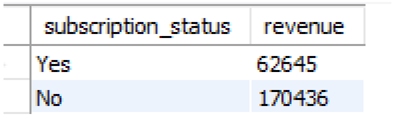
| category | revenue |
|---|---|
| Clothing | 104264 |
| Accessories | 74200 |
| Footwear | 36093 |
| Outerwear | 18524 |

## Q2. Revenue by season



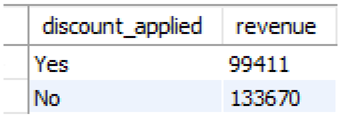
| season | revenue |
|---|---|
| Fall | 60018 |
| Spring | 58679 |
| Winter | 58607 |
| Summer | 55777 |

## Q3. Subscription vs non-subscription revenue



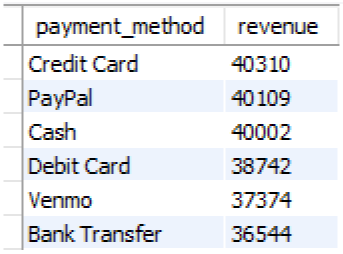
| subscription_status | revenue |
|---|---|
| Yes | 62645 |
| No | 170436 |

## Q4. Impact of discounts on sales



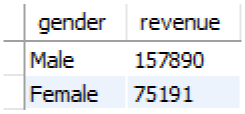
| discount_applied | revenue |
|---|---|
| Yes | 99411 |
| No | 133670 |

## Q5. Most used payment method



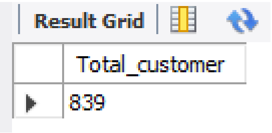
| payment_method | revenue |
|---|---|
| Credit Card | 40310 |
| PayPal | 40109 |
| Cash | 40002 |
| Debit Card | 38742 |
| Venmo | 37374 |
| Bank Transfer | 36544 |

## Q6. What is the total revenue generated by male vs. female customers?

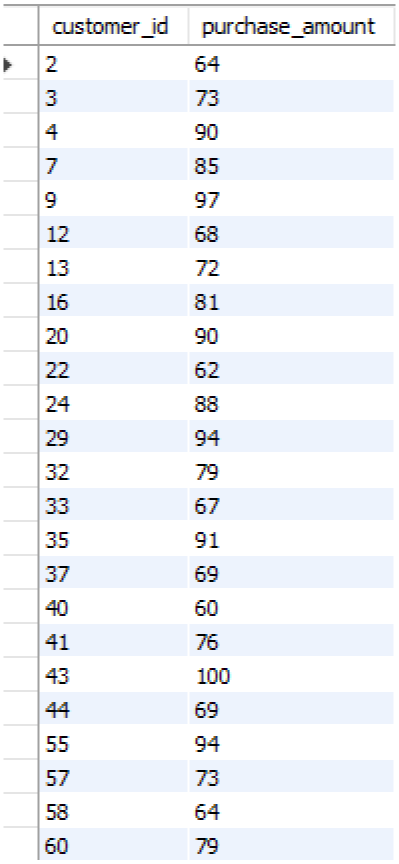


| gender | revenue |
|---|---|
| Male | 157890 |
| Female | 75191 |

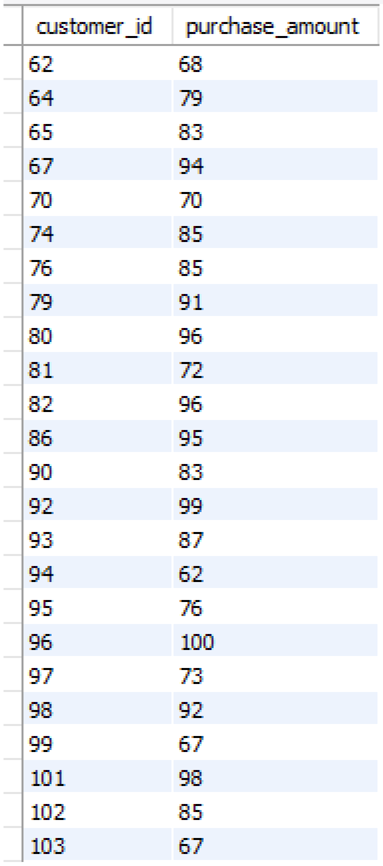
## Q7. Which customers used a discount but still spent more than the average purchase amount?



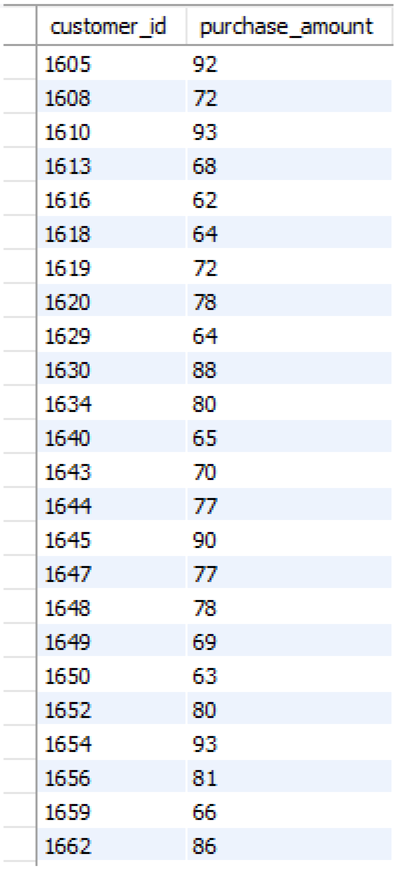
| Total_customer |
|---|
| 839 |



| customer_id | purchase_amount |
|---|---|
| 2 | 64 |
| 3 | 73 |
| 4 | 90 |
| 7 | 85 |
| 9 | 97 |
| 12 | 68 |
| 13 | 72 |
| 16 | 81 |
| 20 | 90 |
| 22 | 62 |
| 24 | 88 |
| 29 | 94 |
| 32 | 79 |
| 33 | 67 |
| 35 | 91 |
| 37 | 69 |
| 40 | 60 |
| 41 | 76 |
| 43 | 100 |
| 44 | 69 |
| 55 | 94 |
| 57 | 73 |
| 58 | 64 |
| 60 | 79 |

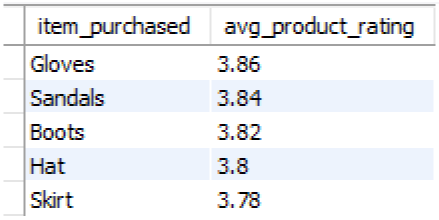


| customer_id | purchase_amount |
|---|---|
| 62 | 68 |
| 64 | 79 |
| 65 | 83 |
| 67 | 94 |
| 70 | 70 |
| 74 | 85 |
| 76 | 85 |
| 79 | 91 |
| 80 | 96 |
| 81 | 72 |
| 82 | 96 |
| 86 | 95 |
| 90 | 83 |
| 92 | 99 |
| 93 | 87 |
| 94 | 62 |
| 95 | 76 |
| 96 | 100 |
| 97 | 73 |
| 98 | 92 |
| 99 | 67 |
| 101 | 98 |
| 102 | 85 |
| 103 | 67 |



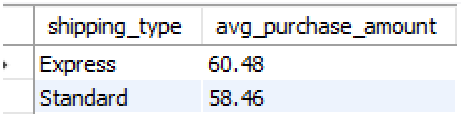
| customer_id | purchase_amount |
|---|---|
| 1605 | 92 |
| 1608 | 72 |
| 1610 | 93 |
| 1613 | 68 |
| 1616 | 62 |
| 1618 | 64 |
| 1619 | 72 |
| 1620 | 78 |
| 1629 | 64 |
| 1630 | 88 |
| 1634 | 80 |
| 1640 | 65 |
| 1643 | 70 |
| 1644 | 77 |
| 1645 | 90 |
| 1647 | 77 |
| 1648 | 78 |
| 1649 | 69 |
| 1650 | 63 |
| 1652 | 80 |
| 1654 | 93 |
| 1656 | 81 |
| 1659 | 66 |
| 1662 | 86 |

## Q8. Which are the top 5 products with the highest average review rating?



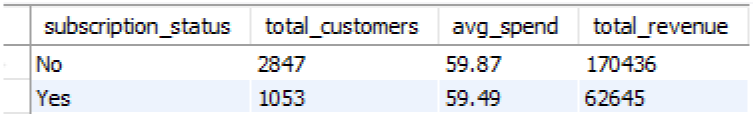
| item_purchased | avg_product_rating |
|---|---|
| Gloves | 3.86 |
| Sandals | 3.84 |
| Boots | 3.82 |
| Hat | 3.8 |
| Skirt | 3.78 |

## Q9. Compare the average Purchase Amounts between Standard and Express Shipping.



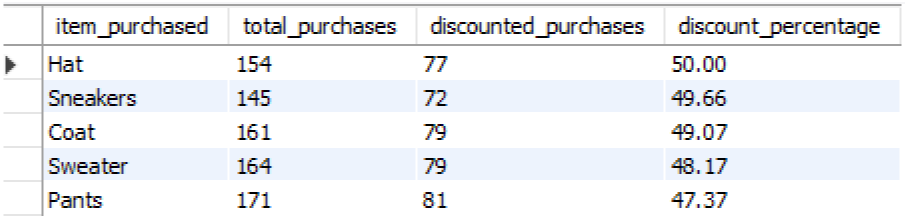
| shipping_type | avg_purchase_amount |
|---|---|
| Express | 60.48 |
| Standard | 58.46 |

## Q10. Do subscribed customers spend more? Compare average spend and total revenue



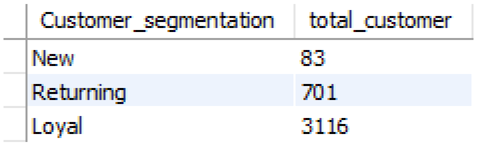
| subscription_status | total_customers | avg_spend | total_revenue |
|---|---|---|---|
| No | 2847 | 59.87 | 170436 |
| Yes | 1053 | 59.49 | 62645 |

## Q11. Which 5 products have the highest percentage of purchases with discounts applied?



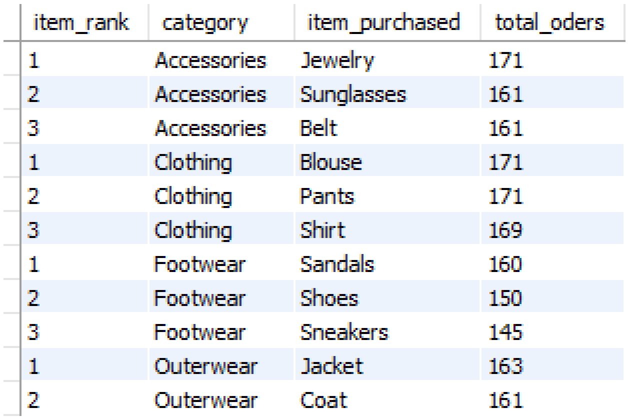
| item_purchased | total_purchases | discounted_purchases | discount_percentage |
|---|---|---|---|
| Hat | 154 | 77 | 50.00 |
| Sneakers | 145 | 72 | 49.66 |
| Coat | 161 | 79 | 49.07 |
| Sweater | 164 | 79 | 48.17 |
| Pants | 171 | 81 | 47.37 |

Q12. Segment customers into New, Returning, and Loyal based on their total number of previous purchases, and show the count of each segment.



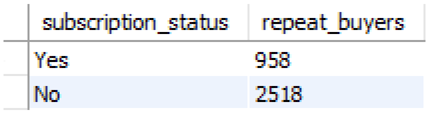
| Customer_segmentation | total_customer |
|---|---|
| New | 83 |
| Returning | 701 |
| Loyal | 3116 |

Q13. What are the top 3 most purchased products within each category?



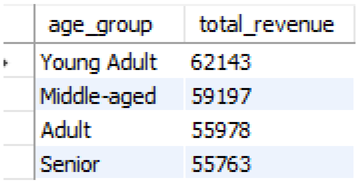
| item_rank | category | item_purchased | total_oders |
|---|---|---|---|
| 1 | Accessories | Jewelry | 171 |
| 2 | Accessories | Sunglasses | 161 |
| 3 | Accessories | Belt | 161 |
| 1 | Clothing | Blouse | 171 |
| 2 | Clothing | Pants | 171 |
| 3 | Clothing | Shirt | 169 |
| 1 | Footwear | Sandals | 160 |
| 2 | Footwear | Shoes | 150 |
| 3 | Footwear | Sneakers | 145 |
| 1 | Outerwear | Jacket | 163 |
| 2 | Outerwear | Coat | 161 |

Q14. Are customers who are repeat buyers (more than 5 previous purchases) also likely to subscribe?



| subscription_status | repeat_buyers |
|---|---|
| Yes | 958 |
| No | 2518 |

Q15. What is the revenue contribution of each age group?

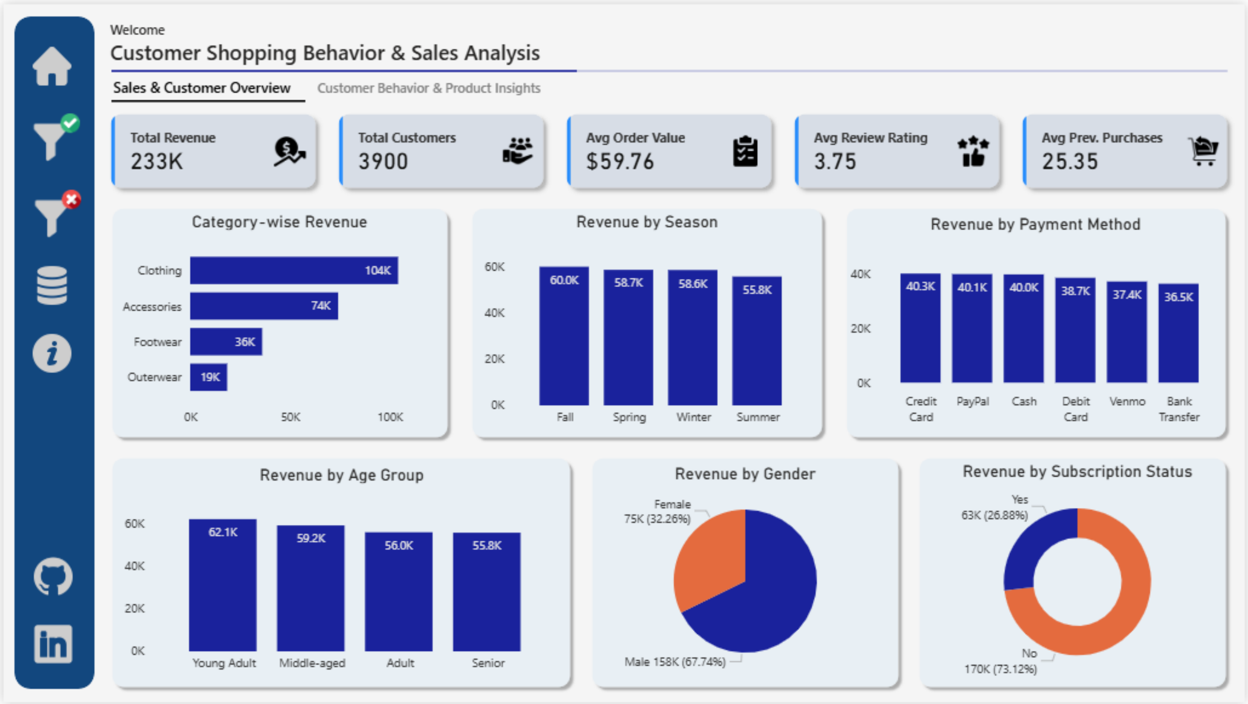


| age_group | total_revenue |
|---|---|
| Young Adult | 62143 |
| Middle-aged | 59197 |
| Adult | 55978 |
| Senior | 55763 |

## Dashboard in Power BI

Finally, an interactive Power BI dashboard was developed to visually present insights, featuring two pages: *Sales & Customer Overview* and *Customer Behavior & Product Insights*.

## *Page 1: Sales & Customer Overview*



Welcome

Customer Shopping Behavior & Sales Analysis

Sales & Customer Overview | Customer Behavior & Product Insights

Total Revenue 233K

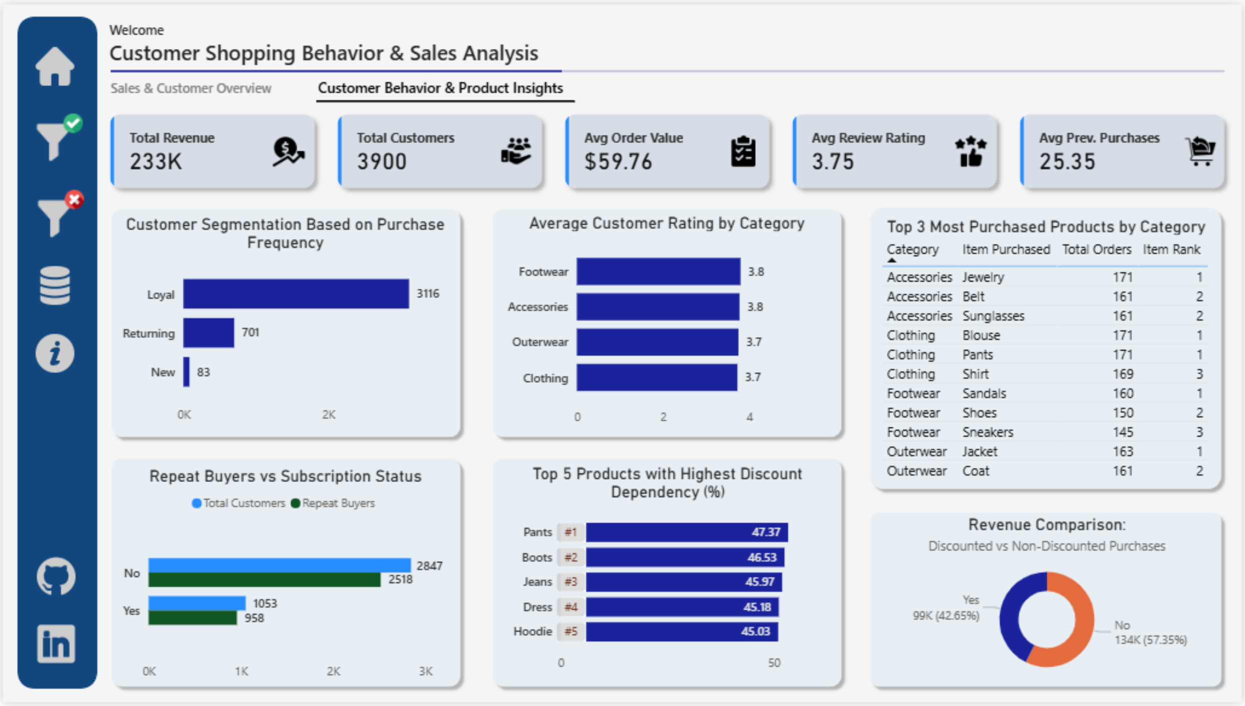
Total Customers 3900

Avg Order Value $59.76

Avg Review Rating 3.75

Avg Prev. Purchases 25.35

## *Page 2: Customer Behavior & Product Insights*



Welcome

Customer Shopping Behavior & Sales Analysis

Sales & Customer Overview | Customer Behavior & Product Insights

Total Revenue 233K

Total Customers 3900

Avg Order Value $59.76

Avg Review Rating 3.75

Avg Prev. Purchases 25.35

Top 3 Most Purchased Products by Category

| Category | Item Purchased | Total Orders | Item Rank |
|---|---|---|---|
| Accessories | Jewelry | 171 | 1 |
| Accessories | Belt | 161 | 2 |
| Accessories | Sunglasses | 161 | 2 |
| Clothing | Blouse | 171 | 1 |
| Clothing | Pants | 171 | 1 |
| Clothing | Shirt | 169 | 3 |
| Footwear | Sandals | 160 | 1 |
| Footwear | Shoes | 150 | 2 |
| Footwear | Sneakers | 145 | 3 |
| Outerwear | Jacket | 163 | 1 |
| Outerwear | Coat | 161 | 2 |

## Strategic Recommendations

- **Boost Subscriptions**: Promote exclusive benefits for subscribers to increase customer lifetime value.
- **Customer Loyalty Programs:** Reward repeat buyers to transition them into the Loyal customer segment.
- **Review Discount Policy:** Balance short-term sales uplift with long-term margin control.
- **Product Positioning:** Highlight top-rated and best-selling products in marketing campaigns.
- **Targeted Marketing:** Focus efforts on high-revenue age groups and express-shipping users for better conversion.

## Business Impact

✓ Improved customer engagement

✓ Higher retention & repeat purchase rate

✓ Optimized promotional spend

✓ Stronger revenue predictability