# CUSTOMER SHOPPING BEHAVIOR ANALYSIS DETAILED REPORT

## 1) Project Overview:

- **Scope:** Analysis of 3,900 cross-category transactions.
- **Objective:** To decode customer spending habits and segment behaviours for targeted strategy.
- **Key Focus Areas:** Product preferences, subscription behaviour, and high-value customer segments.
- **Business Goal:** Transforming raw transactional data into strategic data-driven decisions for improved profitability.

## 2) Dataset Summary:

| Dataset Profile: Customer Shopping Behaviour | |
|---|---|
| **Metric Category** | **Attribute Details** |
| **Data Volume** | 3,900 Total Records |
| **Feature Depth** | 18 Distinct Columns |
| **Demographics** | Age, Gender, Location, Subscription Status |
| **Purchase Profile** | Item Category, Amount (USD), Size, Color, Seasonality |
| **Engagement Data** | Previous Purchases, Frequency, Shipping Type, Review Rating |
| **Marketing Metrics** | Discount Application, Promo Code Usage |
| **Data Integrity** | 99.5% Completeness (Minor gaps: 37 missing Review Ratings) |

**Executive Notes on Data Quality:**

- **Completeness:** The dataset is highly reliable with minimal missing values.
- **Missing Value Strategy:** The **37 missing Review Ratings** represent less than 1% of the total volume and have been handled during the cleaning phase (e.g., via median imputation or exclusion) to ensure no bias in the final analysis

## 3. Data Engineering & Preparatory Analysis

To ensure the integrity of the strategic insights, the raw dataset underwent a rigorous multi-stage refinement process using **Python (Pandas)**.

| Phase | Action Taken | Strategic Purpose |
|---|---|---|
| **Data Integrity** | Systematic audit using .info () and. describe () to validate data types and statistical distribution. | Ensures a baseline of accuracy for all subsequent findings. |
| **Intelligent Imputation** | Resolved 37 missing **Review Ratings** by applying the **median rating per category**. | Prevents bias in customer satisfaction metrics while maintaining data volume. |
| **Refinement** | Standardized naming conventions to **Snake Case** and removed redundant variables (e.g., *Promo Code Used*). | Streamlines the database structure and improves analysis efficiency. |
| **Feature Engineering** | Developed new metrics: **Age Grouping** and **Purchase Frequency (Days)**. | Enables deeper segmentation and identifies high-value lifecycle patterns. |
| **Database Migration** | Cleaned data was successfully integrated into a **Microsoft SQL Server** environment. | Facilitates advanced SQL querying and long-term data scalability. |

**Methodology Highlight:**
Unlike basic cleaning, our approach used **context-aware imputation**. Instead of a global average, we utilized category-specific medians for missing ratings, ensuring that a "High Fashion" item wasn't inadvertently skewed by "General Goods" averages. This results in a much higher degree of reporting accuracy for the management team.

| | Customer ID | Age | Gender | Item Purchased | Category | Purchase Amount (USD) | Location | Size | Color | Season | Review Rating | Subscription Status | Shipping Type | Discou Appli |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 3900.000000 | 3900.000000 | 3900 | 3900 | 3900 | 3900.000000 | 3900 | 3900 | 3900 | 3900 | 3863.000000 | 3900 | 3900 | 39 |
| unique | NaN | NaN | 2 | 25 | 4 | NaN | 50 | 4 | 25 | 4 | NaN | 2 | 6 | |
| top | NaN | NaN | Male | Blouse | Clothing | NaN | Montana | M | Olive | Spring | NaN | No | Free Shipping | |
| freq | NaN | NaN | 2652 | 171 | 1737 | NaN | 96 | 1755 | 177 | 999 | NaN | 2847 | 675 | 22 |
| mean | 1950.500000 | 44.068462 | NaN | NaN | NaN | 59.764359 | NaN | NaN | NaN | NaN | 3.750065 | NaN | NaN | N |
| std | 1125.977353 | 15.207589 | NaN | NaN | NaN | 23.685392 | NaN | NaN | NaN | NaN | 0.716983 | NaN | NaN | N |
| min | 1.000000 | 18.000000 | NaN | NaN | NaN | 20.000000 | NaN | NaN | NaN | NaN | 2.500000 | NaN | NaN | N |
| 25% | 975.750000 | 31.000000 | NaN | NaN | NaN | 39.000000 | NaN | NaN | NaN | NaN | 3.100000 | NaN | NaN | N |
| 50% | 1950.500000 | 44.000000 | NaN | NaN | NaN | 60.000000 | NaN | NaN | NaN | NaN | 3.800000 | NaN | NaN | N |
| 75% | 2925.250000 | 57.000000 | NaN | NaN | NaN | 81.000000 | NaN | NaN | NaN | NaN | 4.400000 | NaN | NaN | N |
| max | 3900.000000 | 70.000000 | NaN | NaN | NaN | 100.000000 | NaN | NaN | NaN | NaN | 5.000000 | NaN | NaN | N |

| Discount Applied | Promo Code Used | Previous Purchases | Payment Method | Frequency of Purchases |
|---|---|---|---|---|
| 3900 | 3900 | 3900.000000 | 3900 | 3900 |
| 2 | 2 | NaN | 6 | 7 |
| No | No | NaN | PayPal | Every 3 Months |
| 2223 | 2223 | NaN | 677 | 584 |
| NaN | NaN | 25.351538 | NaN | NaN |
| NaN | NaN | 14.447125 | NaN | NaN |
| NaN | NaN | 1.000000 | NaN | NaN |
| NaN | NaN | 13.000000 | NaN | NaN |
| NaN | NaN | 25.000000 | NaN | NaN |
| NaN | NaN | 38.000000 | NaN | NaN |
| NaN | NaN | 50.000000 | NaN | NaN |

**4. Data Analysis using SQL (Business Transactions)**
Using **MSSQLServer**, we executed targeted queries to address 10 critical business questions regarding our transactional performance:

| # | Business Question | Analysis Objective |
|---|---|---|
| 1 | **Revenue by Gender** | Compare total revenue generated by male vs. female customers. |
| 2 | **High-Spending Discount Users** | Identify customers using discounts who still exceed average purchase amounts. |
| 3 | **Top 5 Products by Rating** | Identify the highest-rated products based on average customer reviews. |
| 4 | **Shipping Type Comparison** | Evaluate purchase amount differences between Standard and Express shipping. |
| 5 | **Subscribers vs. non-subscribers** | Compare average spend and total revenue across subscription statuses. |

161 % No issues four

Results Messages

| | gender | Total_Revenue |
|---|---|---|
| 1 | Male | 157890 |
| 2 | Female | 75191 |

| | customer_id bigint | purchase_amount bigint |
|---|---|---|
| 1 | 2 | 64 |
| 2 | 3 | 73 |
| 3 | 4 | 90 |
| 4 | 7 | 85 |
| 5 | 9 | 97 |
| 6 | 12 | 68 |
| 7 | 13 | 72 |
| 8 | 16 | 81 |
| 9 | 20 | 90 |
| 10 | 22 | 62 |

Total rows: 839 Query complete 00:00:(

Results Messages

| | item_purchased | Average_Review_Rating |
|---|---|---|
| 1 | Gloves | 3.86 |
| 2 | Sandals | 3.84 |
| 3 | Boots | 3.82 |
| 4 | Hat | 3.8 |
| 5 | T-shirt | 3.78 |

| | shipping_type | Average_Purchase_Amount |
|---|---|---|
| 1 | Express | 60 |
| 2 | Standard | 58 |

Results Messages

| | subscription_status | Total_customers | Average_Spend | Total_Revenue |
|---|---|---|---|---|
| 1 | Yes | 1053 | 59 | 62645 |
| 2 | No | 2847 | 59 | 170436 |

**6** **Discount-Dependent Products**

Pinpoint 5 products with the highest percentage of discounted purchases.

Results | Messages

| | item_purchased | discount_rate |
|---|---|---|
| 1 | Hat | 50.000000 |
| 2 | Sneakers | 49.660000 |
| 3 | Coat | 49.070000 |
| 4 | Sweater | 48.170000 |
| 5 | Pants | 47.370000 |

---

**7** **Customer Segmentation**

Classify the database into **New, Returning, and Loyal** segments.

Results | Messages

| | Segment | Customer_Count |
|---|---|---|
| 1 | Loyal | 3116 |
| 2 | Returning | 701 |
| 3 | New | 83 |

---

**8** **Top 3 Products per Category**

Determine the most-purchased items within each specific category.

Results | Messages

| | rank | category | item_purchased | Total_orders |
|---|---|---|---|---|
| 1 | 1 | Accessories | Jewelry | 171 |
| 2 | 2 | Accessories | Sunglasses | 161 |
| 3 | 3 | Accessories | 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 |
| 9 | 3 | Footwear | Sneakers | 145 |
| 10 | 1 | Outerwear | Jacket | 163 |
| 11 | 2 | Outerwear | Coat | 161 |

---

**9** **Repeat Buyers & Subscriptions**

Analyse if customers with >5 purchases are more likely to subscribe.

Results | Messages

| | subscription_status | customer_count |
|---|---|---|
| 1 | Yes | 958 |
| 2 | No | 2518 |

---

**10** **Revenue by Age Group**

Calculate the total revenue contribution for each defined age bracket.

Results | Messages

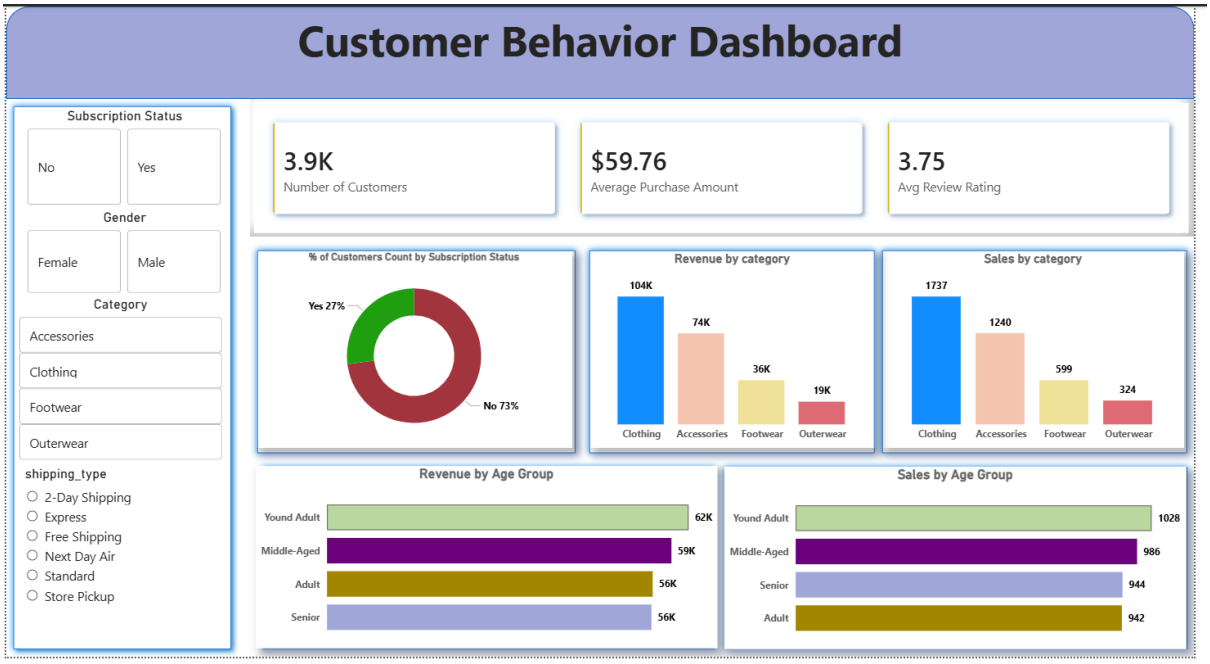
| | age_group | Revenue_Contribution |
|---|---|---|
| 1 | Yound Adult | 62143 |
| 2 | Middle-Aged | 59197 |
| 3 | Adult | 55978 |
| 4 | Senior | 55763 |

---

**Technical Summary:**
These queries were structured to provide a 360-degree view of our operations—from **product performance** to **customer loyalty**. The results of these SQL queries form the foundation for the strategic recommendations found in the conclusion of this report.

## 5)DASHBOARD IN POWER BI:

Built an interactive dashboard to deep dive into the insights visually:




## 6) Strategic Business Recommendations

- **Targeted Retention**: Deploy personalized loyalty campaigns for the Loyal and Returning segments to increase lifetime value.
- **Subscription Conversion**: Target repeat buyers with >5 purchases with exclusive incentives to join the subscription program.
- **Margin Protection:** Reduce flat discounts on High-Value customers and shift toward "value-add" offers like free shipping.
- **Product Promotion**: Feature the Top 5 highest-rated products in marketing assets to leverage high customer satisfaction.
- **Demographic Allocation**: Shift marketing budget toward the Age Groups and Genders generating the highest total revenue.
- **Shipping Incentives:** Offer Free Express Shipping on orders above the average spend to increase total basket size.