Customer Shopping Behavior Analysis Report

1. Business Problem

A leading retail company wants to better understand customer shopping behavior to improve sales, satisfaction, and loyalty.
The management team has observed variations in purchasing patterns across demographics, product categories, and sales channels.
They seek to identify how factors such as age, discounts, seasons, product reviews, and subscription status influence purchasing decisions and repeat buying behavior.

The overarching question is:

"How can the company leverage customer shopping data to identify trends, improve engagement, and optimize marketing and product strategies?"

---

2. Dataset Overview

Source: Customer Shopping Behavior Dataset (3,900 records)

Structure:

- **Rows:** 3,900
- **Columns:** 18
- **Key Attributes:**
    - Customer demographics (Age, Gender, Location, Subscription Status)
    - Purchase details (Item Purchased, Category, Purchase Amount, Season, Size, Color)
    - Shopping behavior (Discount Applied, Promo Code Used, Previous Purchases, Review Rating, Shipping Type, Frequency of Purchases)

Data Issues Identified:

- 37 missing values in the `review_rating` column.
- Inconsistent column naming format.
- Redundant column (`promo_code_used`) found to overlap with `discount_applied`.

---

3. Data Preparation (Python)

Steps Performed:

1. **Data Import & Exploration**
    - Loaded dataset using Pandas.
    - Used `.info()` and `.describe()` for initial structure and summary.
2. **Handling Missing Values**
    - Filled null values in `review_rating` using the median rating within each product category.
3. **Column Standardization**
    - Renamed columns to lower snake_case for consistency and readability.
4. **Feature Engineering**
    - Created `age_group` by binning the `age` column into four quantiles: Young Adult, Adult, Middle Age, Senior.
    - Created `purchase_frequency_days` from frequency descriptions such as Weekly, Monthly, Annually.
5. **Data Consistency**
    - Verified redundancy between `discount_applied` and `promo_code_used`, and dropped the latter.

6. **Database Integration**
    - Connected to SQL Server using SQLAlchemy and uploaded the cleaned dataset for SQL analysis.

---

4. SQL Analysis and Insights

a. Revenue by Gender

Analyzed total and average purchase amounts by gender.
Result: Female customers generated slightly higher average purchase amounts compared to male customers.

b. Top Rated Products

Query identified products with the highest average review ratings.
Result: Items such as *Sneakers*, *Blouses*, and *Watches* had the highest satisfaction levels.

c. Discount-Dependent Products

Determined which products most frequently relied on discounts for sales.
Result: Apparel items had the highest proportion of discounted purchases.

d. Customer Segmentation

Segmented customers based on `previous_purchases`:

- New: ≤ 1
- Returning: 2–10
- Loyal: > 10

Result: Returning customers formed the largest segment, but Loyal customers contributed the most to total revenue.

e. Revenue by Age Group

Calculated total revenue across the newly created `age_group` feature.
Result: Adults and Middle-Aged customers contributed the highest revenue share.

f. Shipping Type Comparison

Compared average purchase amounts by shipping type.
Result: Customers using Express Shipping spent significantly more on average.

---

5. Power BI Dashboard

The Power BI dashboard consolidates key metrics and interactive visuals from the cleaned dataset.
Key pages and visuals include:

- **Revenue by Age Group** (bar chart)
- **Top Products by Category** (stacked bar)
- **Customer Segment Distribution** (pie or donut chart)
- **Total Revenue by Gender and Category**
- **Average Purchase Amount by Shipping Type**
- **Discount Impact on Sales Volume**
- **Seasonal Purchase Trends**

These visuals enable management to filter and explore data by category, gender, age group, and subscription status.

---

6. Key Findings

1. Female customers tend to have higher average purchase amounts.
2. Express shipping users spend more than standard shipping users.
3. Adults and Middle-Aged customers are the most profitable segments.
4. Apparel products dominate both total orders and discount usage.

5. Loyal customers provide the highest revenue despite being fewer in number.

6. High review-rated products correlate with repeat purchases.

---

7. Business Recommendations

1. **Promote Subscriptions**
   Introduce exclusive subscriber discounts or early access offers to increase customer retention.

2. **Encourage Loyalty**
   Implement loyalty programs that reward frequent buyers to move them into the “Loyal” segment.

3. **Optimize Discount Strategies**
   Limit discount offers to high-margin or low-conversion products to balance revenue and profit.

4. **Highlight High-Rated Products**
   Feature top-rated products prominently in advertisements and campaigns.

5. **Focus Marketing Efforts on Key Demographics**
   Target Adult and Middle-Aged customers, who demonstrate the highest purchasing power.

6. **Enhance Express Shipping Options**
   As express users are high spenders, offering better deals or convenience for them can boost sales further.

---

8. Deliverables Summary

Deliverable	Description
Data Preparation (Python)	Cleaned and transformed the raw CSV data
Database Integration (SQL Server)	Structured data for querying and insights

Data Analysis (SQL)	Generated business insights and customer segmentation
Visualization (Power BI)	Built an interactive dashboard using the `.pbix` file
Reports	Final written documents summarizing findings and recommendations

---

9. Conclusion

The integration of Python-based data preparation, SQL-driven analysis, and Power BI visualization provides a comprehensive view of customer shopping behavior.
The insights support actionable business strategies, including targeted marketing, refined discount policies, and loyalty-building initiatives.
This project demonstrates a complete data-driven approach to understanding and optimizing customer engagement and revenue growth.