Customer Shopping Behavior Analysis

1. Introduction

This project presents a detailed analysis of customer shopping behavior using transactional retail data. The study aims to understand how customers interact with products across categories, seasons, and demographic segments. By leveraging Python, SQL, and Power BI, the project converts raw transactional data into meaningful business insights that can support strategic planning, marketing optimization, and customer retention initiatives.

2. Business Objective

The primary objective of this project is to analyze customer purchase behavior and identify patterns related to:

- Spending behavior and revenue drivers
- Product category performance
- Seasonal demand fluctuations
- Customer demographics and subscription behavior

The insights generated are intended to help businesses improve decision-making and enhance customer experience.

3. Dataset Description

The dataset consists of 3,900 customer purchase records with 18 attributes. It includes demographic details, transaction-level purchase data, and behavioral indicators such as subscription status, discount usage, shipping preference, and review ratings.

Key Columns:

- Customer ID, Age, Gender, Location
- Item Purchased, Category, Purchase Amount (USD)
- Season, Discount Applied, Promo Code Used
- Subscription Status, Shipping Type
- Review Rating, Purchase Frequency

4. Skills & Tools Demonstrated

- Python (EDA, Data Cleaning, Analysis)
- SQL (Data Extraction and Business Queries)
- Power BI (Interactive Dashboards)
- Data Visualization & Business Insight Generation

5. Data Quality & Preprocessing

Initial data inspection was conducted using Python. Data types were validated, and missing values were identified in the Review Rating column (37 records). These values were handled to ensure analytical accuracy. No duplicate records were found. Categorical variables were standardized for consistency.

6. Exploratory Data Analysis (EDA)

We began with data preparation and cleaning in Python.

Data Loading: Imported the dataset using pandas

Initial Exploration: Used df.info() to check structure and .describe() for summary statistics.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             3900 non-null   int64
 1   Age                     3900 non-null   int64
 2   Gender                  3900 non-null   object
 3   Item Purchased          3900 non-null   object
 4   Category                3900 non-null   object
 5   Purchase Amount (USD)   3900 non-null   int64
 6   Location                3900 non-null   object
 7   Size                    3900 non-null   object
 8   Color                   3900 non-null   object
 9   Season                  3900 non-null   object
 10  Review Rating           3900 non-null   float64
 11  Subscription Status     3900 non-null   object
 12  Shipping Type           3900 non-null   object
 13  Discount Applied        3900 non-null   object
 14  Promo Code Used         3900 non-null   object
 15  Previous Purchases      3900 non-null   int64
 16  Payment Method          3900 non-null   object
 17  Frequency of Purchases  3900 non-null   object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

```
df.describe(include='all')
```

	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
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3900.000000	3900	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.749949	NaN	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716223	NaN	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.700000	NaN	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN

Missing Data Handling: Checked for the null values and imputed missing values in the Review Rating column using the median rating of each product category.

Column Standardization: Rename columns to snake case for better readability and documentation.

Feature Engineering:

- Create age_group column by binning customer ages
- Create purchase_frequency_days column for purchase data.

Data Consistency Check: Verified if discount_applied and promo_code_used were redundant; dropped promo_code_used.

Database Integration: Connected Python Script to MySQL and loaded the cleaned Data Frame into the database for SQL analysis

6. SQL Query – Customer Shopping Behavior Analysis

Q1. Total Revenue Generated by Male vs Female Customers

Description:

This query calculates the total revenue generated by male and female customers by aggregating purchase amounts by gender. It helps identify which gender contributes more to overall revenue and supports demographic-based marketing and targeting strategies.

Q2. Customers Who Used Discounts but Spent Above Average

Description:

This query identifies customers who availed discounts yet spent more than the average purchase amount. It highlights high-value customers who remain strong spenders even when discounts are applied, helping optimize promotional strategies without impacting revenue.

Q3. Top 10 Products with the Highest Average Review Ratings

Description:

This query retrieves the top 10 products based on average customer review ratings. It helps identify best-rated products that can be promoted in marketing campaigns and positioned as premium or high-quality offerings.

Q4. Average Purchase Amount by Shipping Type

Description:

This query compares the average purchase amount between customers choosing Standard and Express shipping. It helps understand whether faster shipping options are associated with higher spending behavior.

Q5. Spending Behavior of Subscribed vs Non-Subscribed Customers

Description:

This query compares total customers, average spending, and total revenue between subscribed and non-subscribed customers. It evaluates whether subscription status influences customer spending and revenue contribution.

Q6. Products with the Highest Discount Usage Rate

Description:

This query identifies the top 5 products with the highest percentage of purchases where discounts were applied. It helps determine which products are most price-sensitive and frequently require discounts to drive sales.

Q7. Customer Segmentation Based on Previous Purchases

Description:

This query segments customers into **New**, **Returning**, and **Loyal** categories based on the number of previous purchases. It provides insights into customer retention levels and helps design targeted engagement and loyalty programs.

Q8. Top 3 Most Purchased Products Within Each Category

Description:

This query ranks products within each category based on total purchase count and identifies the top 3

best-selling products per category. It supports inventory optimization and category-level marketing decisions.

Q9. Subscription Behavior of Repeat Buyers

Description:

This query analyzes whether repeat buyers (customers with more than 5 previous purchases) are more likely to subscribe. It helps assess the relationship between customer loyalty and subscription adoption.

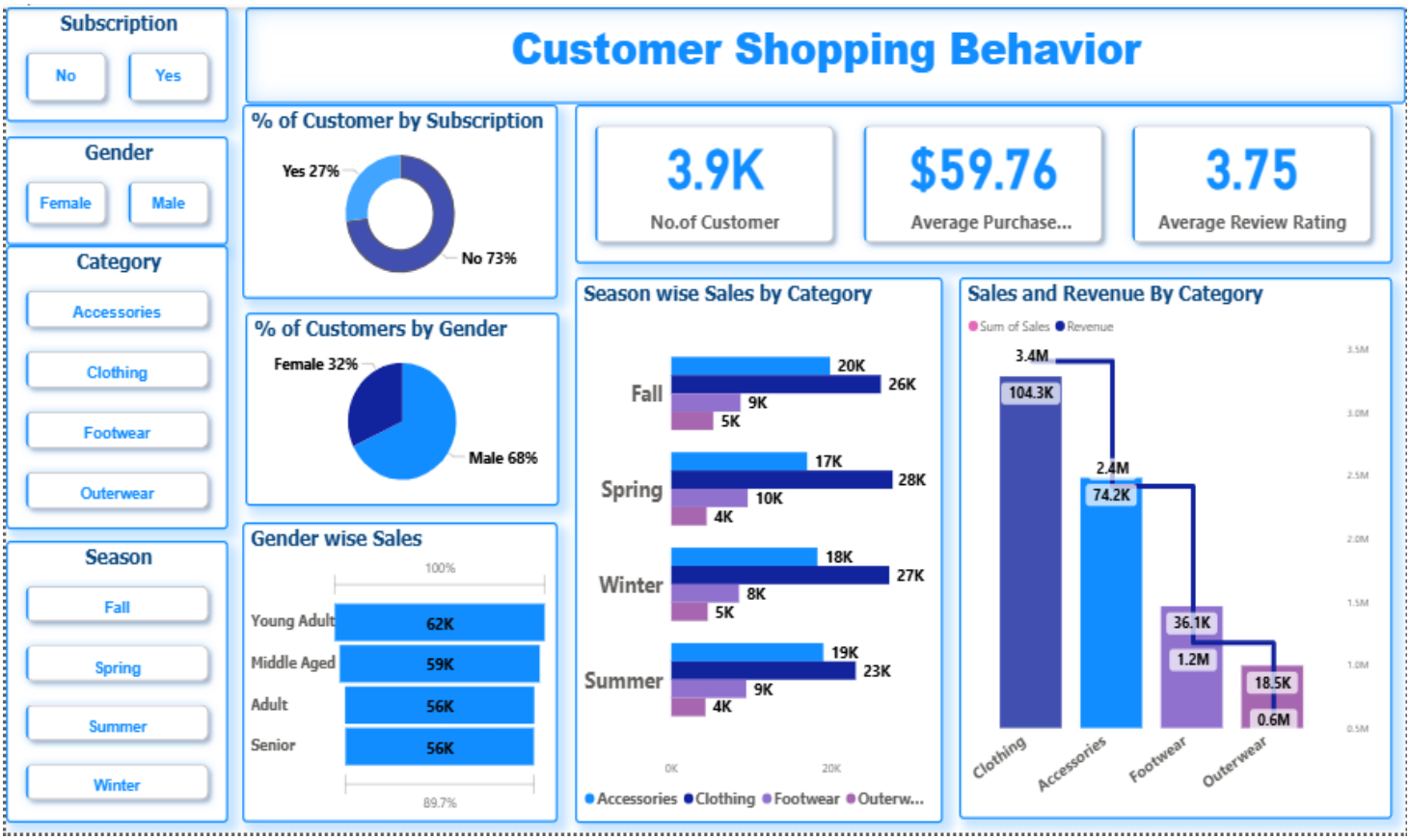
Q10. Revenue Contribution by Age Group

Description:

This query calculates total revenue generated by each age group. It helps identify high-value age segments and supports targeted marketing and personalized offers based on customer demographics.

10. Power BI Dashboard Development

An interactive Power BI dashboard was developed to visualize KPIs and trends. The dashboard includes customer count, average purchase value, average review rating, category-wise sales and revenue, and season-wise performance. Slicers allow users to dynamically filter data by gender, season, category, and subscription status.



11. Business Insights

- ✓ High revenue concentration in Clothing and Accessories
- ✓ Strong seasonal dependency of sales
- ✓ Underutilized subscription model
- ✓ High engagement from young and middle-aged demographics

12. Business Recommendations

- Introduce subscription-based benefits to increase adoption
- Launch loyalty programs for repeat customers
- Optimize discount strategies to balance revenue and margins
- Focus marketing efforts on high-performing categories
- Design seasonal promotions aligned with demand trends

13. Conclusion

This project demonstrates a complete end-to-end data analytics lifecycle, from raw data exploration to actionable business recommendations. The integration of Python, SQL, and Power BI showcases strong analytical, visualization, and business storytelling skills relevant for Data Analyst and Business Analyst roles.