

Customer Shopping Behavior Analysis

1. Project Overview

This project focuses on understanding customer behaviour using a retail shopping dataset containing demographic details, purchasing patterns, product preferences, and satisfaction metrics.

The goal is to uncover insights that inform business strategies around customer retention, product optimization, pricing, and targeted marketing.

The study follows a full end-to-end analytics workflow:

- **Python** for data cleaning, preprocessing, and feature engineering
- **SQL Server** for structured business analysis and segmentation
- **Power BI** for interactive dashboarding and storytelling

Through this pipeline, the project evaluates how demographic groups spend, how frequently customers shop, which products and categories perform best, and how subscription and discount usage influence consumer decisions.

2. Dataset Summary

- Rows: 3,900
- Columns: 18
- Key Features:
 - Demographic Attributes (Age, Gender, Location, Customer ID)
 - Product & Transaction Details (Item Purchased, Category, Purchase Amount, Season, Size, Colour, Discount Applied, Review Rating, Shipping Type, Promo Code Used)
 - Behavioural Indicators (Previous Purchases, Frequency of Purchases, Subscription Status, Payment Method)
- Missing Data: 37 values in Review Rating column. Missing entries were imputed using **median rating within each category**, ensuring realistic and category-specific replacement.
- Two new features were created to enhance analytical depth:
 - **age_group**
Derived using quantile segmentation (Young Adult, Adult, Middle-Aged, Senior)

- **purchase_frequency_days**

Mapped textual frequency descriptions into numerical day counts for quantitative analysis

These engineered fields enabled several SQL insights, including frequency-based segmentation and revenue contribution by age groups.

3. Exploratory Data Analysis Using Python

Python was used to clean, prepare, and enrich the dataset before loading it into SQL Server and Power BI.

The following steps summarize the full EDA pipeline executed in Python, along with the purpose and impact of each transformation.

- **Data Import and Initial Inspection**

The dataset was loaded using Pandas:

```
df = pd.read_csv('customer_shopping_behavior.csv')
df.head()
df.info()
df.describe(include='all')
```

```
<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                        3863 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
```

[13]:

	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	3863.000000	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	3.750065	NaN	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	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	3.800000	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

What this step revealed:

- The dataset contains - 3,900 rows and 18 columns.
- Several categorical and numerical fields needed formatting.
- Missing values were identified in the *review rating* column.
- Some column names were inconsistent or contained spaces.

• Missing Data Handling

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

```
df['Review Rating'] = df.groupby('Category')['Review Rating']\
    .transform(lambda x: x.fillna(x.median()))
```

After imputation:

```
df.isnull().sum()
```

Missing values were successfully resolved.

• Column Standardization

Renamed columns to **snake case** for better readability and documentation.

```
df.columns = df.columns.str.lower().str.replace(' ','_')
df = df.rename(columns={'purchase_amount_(usd)':
    'purchase_amount'})
```

• Feature Engineering

- Created **age_group** column by binning customer ages.

```
labels = ['Young Adult','Adult','Middle Aged','Seniors']
df['age_group'] = pd.qcut(df['age'], q=4, labels=labels)
```

- Created **purchase_frequency_days** column from purchase data.

```
frequency_mapping = {
```

```

        'Fortnightly': 14,
        'Weekly': 7,
        'Monthly': 30,
        'Quarterly': 90,
        'Bi-Weekly': 14,
        'Annually': 365,
        'Every 3 Months': 90
    }
    df['purchase_frequency_days'] =
    df['frequency_of_purchases'].map(frequency_mapping)

```

- **Data Consistency Check**

Verified if `discount_applied` and `promo_code_used` were redundant; dropped `promo_code_used`.

```

(df['discount_applied'] == df['promo_code_used']).all()
df = df.drop('promo_code_used', axis=1)

```

- **Database Integration**

A connection was established using **SQLAlchemy + pyodbc**:

```

engine = create_engine("mssql+pyodbc:///odbc_connect=%s" %
params)
df.to_sql("customer", engine, if_exists="replace", index=False)

```

The cleaned dataset was transferred into the SQL Server database **customer_behaviour**

4. Data Analysis Using SQL (Business Transaction)

After cleaning and preparing the dataset in Python, the refined table was imported into **SQL Server** for structured business analysis.

A set of 20 analytical SQL queries were executed to explore customer behaviour, spending patterns, discount usage, product preferences, and subscription trends.

Below is a summary of each insight derived from the SQL queries.

1) Age Groups Driving the Highest Revenue & Order Volume

Revenue analysis showed that certain age groups—especially **Young Adults and Middle-aged customers**—generated the highest total

spend and order frequency, making them the most valuable demographic segments.

	age_group	order_count	total_revenue
1	Young Adult	1028	62143
2	Middle Aged	986	59197
3	Adult	942	55978
4	Seniors	944	55763

2) **Average Purchase Frequency Across Age Groups**

Younger customers tended to shop more frequently, while older groups showed longer gaps between purchases. This helps in tailoring targeted re-engagement strategies.

age_group	avg_frequency
Young Adult	84
Adult	92
Middle Aged	83
Seniors	96

3) **Discount Usage by Demographics**

Females in the **Young Adult** and **Middle-aged** groups were the most likely to apply discounts during purchases, indicating strong price sensitivity in specific segments.

gender	age_group	discounted_purchases
Male	Young Adult	443
Male	Seniors	417
Male	Middle Aged	411
Male	Adult	406

4) **Review Ratings by Payment Method**

Average ratings vary slightly across payment methods. Certain payment types may correlate with more satisfied customers — possibly due to smoother or preferred transaction experiences.

payment_method	avg_rating
Bank Transfer	3.70343137254902
Debit Card	3.76037735849057
Cash	3.7476119402985
Venmo	3.74716088328076
Credit Card	3.78703427719822
PayPal	3.75096011816839

5) Top Categories Purchased by Repeat Customers

Customers with **previous_purchases > 0** most frequently bought from categories like **Clothing, Accessories, and Footwear**, confirming strong repeat appeal in these product lines.

category	purchase_count
Clothing	1737
Accessories	1240
Footwear	599
Outerwear	324

6) Subscription Status vs Purchase Frequency

Subscribers demonstrated **shorter purchase_frequency_days**, indicating they return to shop more often compared to non-subscribers.

subscription_status	avg_days
Yes	89
No	88

7) Most Re-purchased Products

Items such as **Blouse, Pants, Sneakers, and Jewellery** ranked high among repeat buyers, suggesting strong product loyalty.

item_purchased	repeat_customer
Blouse	171
Jewelry	171
Pants	171
Shirt	169
Dress	166
Sweater	164
Jacket	163
Coat	161
Belt	161
Sunglasses	161
Sandals	160
Socks	159
Skirt	158
Shorts	157
Scarf	157
Hat	154
Handbag	153
Hoodie	151
Shoes	150
T-shirt	147
Sneakers	145
Boots	144
Backpack	143
Gloves	140
Jeans	124

8) Preferred Shipping Type for High-Engagement Customers

Customers with **>5 previous purchases** showed higher preference for **Standard shipping**, though Express usage was present among higher-value shoppers.

shipping_type	num_customers
Standard	599
Free Shipping	595
Store Pickup	582
2-Day Shipp...	570
Next Day Air	566
Express	564

9) Relationship Between Review Ratings & Purchase Amount

Higher review ratings correspond with higher average purchase amounts, indicating satisfied customers tend to spend more.

review_rating	avg_purchase
5	64
4.9	63
4.8	61
4.7	59
4.6	57
4.5	59
4.4	60
4.3	59
4.2	60
4.1	61
4	59
3.9	58
3.8	61
3.7	60
3.6	57
3.5	58
3.4	59
3.3	59
3.2	61
3.1	58
3	60
2.9	56
2.8	57
2.7	59
2.6	59
2.5	62

10) High-Frequency Buyers by Category

Customers with above-average purchase frequency were more likely to shop from categories like **Clothing and Footwear**, showing strong engagement with lifestyle essentials.

category	high_frequency_buyers
Outerwear	142
Footwear	262
Accesso...	553
Clothing	762

11) Review Ratings: Subscribers vs Non-Subscribers

Subscribers gave **slightly better average ratings**, reflecting improved satisfaction or loyalty among subscribed users.

subscription_status	avg_review
Yes	3.75
No	3.75

12) Payment Methods Preferred by Discount Users

Certain payment methods were more common among customers using discounts—useful for aligning promotions with preferred payment partners.

payment_method	discounted_transaction
Debit Card	289
Venmo	286
Cash	283
Credit Card	280
PayPal	280
Bank Transfer	259

13) Frequent Buyers (Purchase Frequency < 15 Days) by Age Group

Young Adults and Middle-aged customers dominated high-frequency purchasing patterns.

age_group	frequent_buyers
Young Adult	440
Middle Aged	418
Adult	389
Seniors	381

14) Subscription Percentage by Age Group

Younger customers held a higher percentage of subscriptions compared to older groups, revealing stronger adoption among early-age demographics.

age_group	subscription_pct
Young Adult	25
Adult	26
Middle Aged	28
Seniors	26

15) Top 5 Highest-Rated Products

Products like **Gloves, Sandals, Boots, Hat, and Skirt** received the highest average ratings.

item_purchased	Avg Product Review
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Handbag	3.78

16) Average Purchase Amount by Shipping Type

Express shipping correlated with slightly higher average spending, while Standard shipping remained the more commonly used option.

shipping_type	Avg_spend
Express	60
Standard	58

17) Do Subscribers Spend More?

Yes.

Subscribers showed **higher average spend per transaction** and

contributed significantly to revenue relative to their smaller population size.

subscription_status	total_customers	avg_spend	total_revenue
Yes	1053	59	62645
No	2847	59	170436

18) Customer Segmentation (New, Returning, Loyal)

Using previous purchase counts:

- **New:** previous_purchases = 1
- **Returning:** 2–10 purchases
- **Loyal:** >10 purchases

Loyal customers formed a smaller population but showed the highest value in repeat purchasing and revenue.

customer_segment	Number of Customers
New	83
Returning	701
Loyal	3116

19) Top 3 Most Purchased Products per Category

Using window functions, the top 3 products in each category were ranked.

Examples include:

- **Accessories:** Jewellery, Sunglasses, Belt
- **Clothing:** Blouse, Pants, Shirt
- **Footwear:** Sandals, Shoes, Sneakers
- **Outerwear:** Jacket, Coat

item_rank	category	item_purchased	total_orders
1	Accessories	Jewelry	171
2	Accessories	Belt	161
3	Accessories	Sunglasses	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

20) Are Repeat Buyers More Likely to Subscribe?

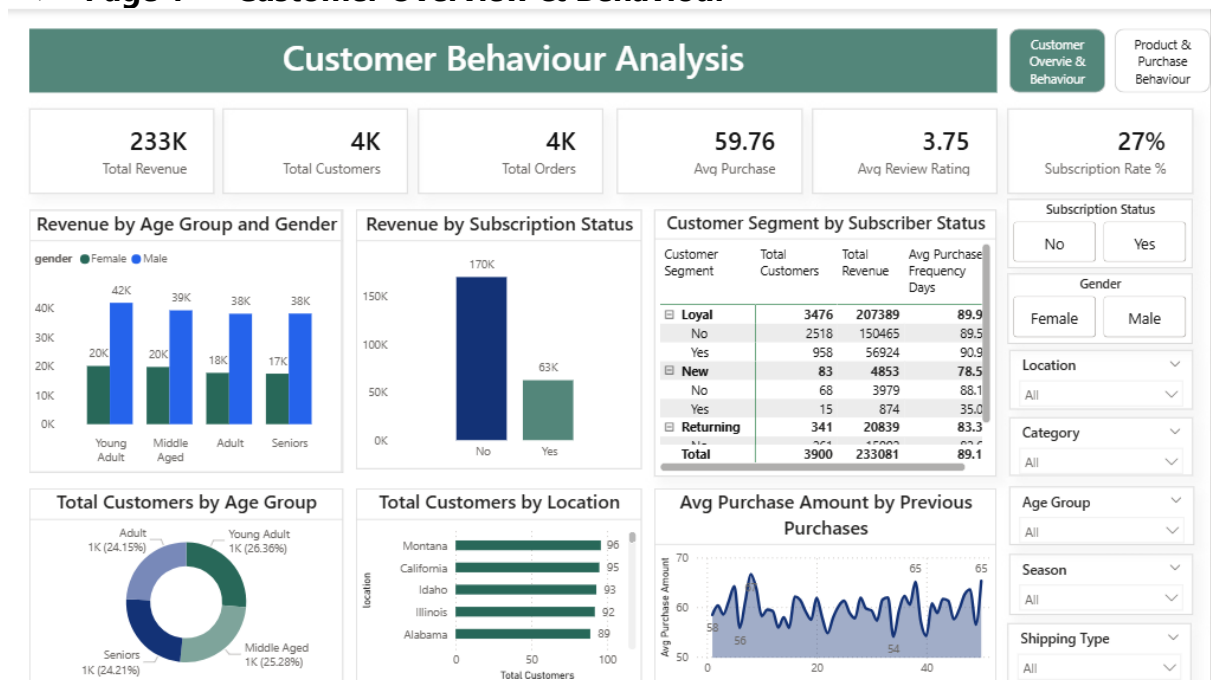
Customers with **more than 5 previous purchases** showed a significantly higher subscription count, indicating that repeat engagement leads to subscription adoption.

subscription_status	repeat_buyers
Yes	958
No	2518

5. Dashboard in Power BI

The Power BI dashboard was designed across **two analytical pages**, focusing on customer demographics, behavioural trends, product performance, and satisfaction patterns.

➤ Page 1 — Customer Overview & Behaviour



➤ Page 2 — Product, Revenue & Review Insights



Primary Insights

- **Revenue by Age and Gender**
 - Middle-Aged and Young Adult groups were the strongest contributors to revenue.
 - Female customers slightly outperformed male customers in overall spending.
- **Purchase Trend Based on Previous Orders**
 - Average purchase amount fluctuated with the number of past purchases, highlighting behavioural differences between new, returning, and loyal customers.
- **Customer Segments**
 - **Loyal customers** generated the highest revenue
 - **New** and **Returning** customers showed moderate spend but strong growth potential
- **Geographic Distribution**
 - The location bar chart showed uneven customer spread, with certain states contributing significantly more customers and revenue.
- **Revenue by Subscription Status**
 - Despite subscribers making up only **27%**, they exhibited **higher engagement and consistency**, though total revenue from non-subscribers was larger due to population size.
- **Top 10 Products by Revenue**

- Products such as **Blouse, Shirt, Dress, Jewelry** consistently appeared as top performers across all customers.
- **Seasonal Revenue Distribution**
 - Fall and Spring dominated revenue, highlighting ideal periods for promotional or product-launch strategies.
- **Category-Wise Customer Satisfaction**
 - Footwear and Accessories categories recorded stronger rating performance, indicating higher perceived value or comfort with these categories.
- **Shipping Type vs Customer Satisfaction**
 - Standard shipping showed slightly better average ratings, implying that customer expectations may align better with longer delivery windows.
- **High-Frequency Buyer Behaviour**
 - Customers with more frequent purchase patterns tended to favour specific categories such as **Clothing** and **Footwear**, aligning with SQL findings.
- **Discount Impact on Revenue**
 - A significant portion of revenue came without discounts, while discounted purchases still formed a substantial share - indicating a healthy mix of value-driven and convenience-driven buyers.

6. Business Recommendation

- **Strengthen Subscription Strategy**
Subscribers exhibit higher engagement and shorter purchase intervals. Introduce targeted subscription offers, bundled perks, or loyalty benefits to convert more customers into subscribers.
- **Enhance Loyalty Programs**
Loyal customers generate the majority of revenue. Offering exclusive rewards, early access, or member-only discounts can reinforce long-term retention.
- **Seasonal Marketing Optimization**
Since Fall and Spring generate the highest revenue, seasonal campaigns, new product launches, and promotions should be aligned accordingly.

- **Focus on High-Performing Categories**

Products like Blouse, Jewellery, and Footwear items consistently rank high in both spend and engagement. Increasing visibility and stock variety in these categories can strengthen revenue.

- **Improve Shipping Experience**

Small rating gaps suggest opportunities to improve Express shipping reliability, packaging, and delivery time accuracy.

- **Reevaluate Discount Strategy**

A balanced discount approach is needed:

- Too many discounts may reduce margins
 - But removing them may reduce purchase frequency
- Dynamic discounting or targeted coupons could provide better results.

○

- **Demographic-Focused Marketing**

Young Adult and Middle-Aged segments drive the highest revenue. Marketing campaigns, product bundles, and subscription promotions should focus on these audiences.