Customer Shopping Behavior Analysis

1. Executive Summary

This analysis examines the behavior of customer shopping to understand revenue drivers, customer loyalty, discount dependency, and product performance. Using python for data preparation, SQL for analytical querying, and Power BI for visualization, the project delivers actionable insights for pricing, retention, and marketing strategy.

Key highlights:

- Identified high-value customer segments based on purchase history
- Evaluated the effectiveness of discounts on spending behavior
- Analyzed product and category performance
- Assessed subscription behavior among repeat buyers

2. Business Context & Analytical Objectives

Retail businesses rely on customer behaviour data to optimize revenue, retention, and promotions. This analysis focuses on answering the following business questions:

- Which customer segments generate the most revenue?
- Do discounts attract high-value customers or primarily increase volume?
- Which products dominate sales within each category?
- Are repeat buyers more likely to subscribe?
- Which age groups contribute most to overall revenue?

3. Data Overview

Source: Retail transactional dataset
Observations: ~3,900 transactions
Features: Customer demographics, purchase details, discounts, subscriptions, and behavioral metrics

Key Data Attributes

- Demographics: Age, Gender, Location
- Purchase Data: Item Purchased, Category, Purchase Amount
- Behavioural Indicators: Previous Purchases, Purchase Frequency, Discount Applied
- Feedback: Review Rating
- Logistics: Shipping Type, Season

Data Quality Notes

- Missing values in review ratings
- Redundant discount indicators identified
- Inconsistent categorical labels in purchase frequency

```
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         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
```

Figure 1:Dataset schema and data types after initial loading.

4. Data Preparation & Feature Engineering (Python)

4.1 Data Cleaning

- Standardized column names using snake_case
- Identified and handled missing values in review ratings using category-wise median imputation
- Verified consistency between discount and promo code indicators

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

```
df.columns
```

```
Index(['customer_id', 'age', 'gender', 'item_purchased', 'category',
       'purchase_amount', 'location', 'size', 'color', 'season',
       'review_rating', 'subscription_status', 'shipping_type',
       'discount_applied', 'promo_code_used', 'previous_purchases',
       'payment_method', 'frequency_of_purchases'],
      dtype='object')
```

Figure 2:Data cleaning and standardization performed in Python.

4.2 Feature Engineering

- Created age_group feature using quantile-based binning
- Transformed categorical purchase frequency into purchase_frequency_days
- Classified customers based on purchase history

```
# create a column age_group

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

```
df[['age', 'age_group']].head(10)
```

	age	age_group
0	55	Middle-aged
1	19	Young Adult
2	50	Middle-aged
3	21	Young Adult
4	45	Middle-aged
5	46	Middle-aged
6	63	Senior
7	27	Young Adult
8	26	Young Adult
9	57	Middle-aged

Figure 3:Feature engineering for age segmentation and purchase frequency normalization.

5. Analytical Methodology (SQL)

5.1 Revenue & Spend Analysis

- Revenue distribution by gender
- Revenue contribution across age groups
- Comparison of average spend between subscribers and non-subscribers

age_group	total_revenue
45+	114960.0
25–34	45400.0
35–44	43463.0
Under 25	29258.0

Figure 4: Revenue distribution by key customer attributes.

5.2 Product & Category Performance

- Top products within each category using window functions
- Products with highest discount dependency
- Impact of shipping type on average purchase value

Category	Item Purchased	purchase_count
Accessories	Jewelry	171
Accessories	Belt	161
Accessories	Sunglasses	161
Clothing	Pants	171
Clothing	Blouse	171
Clothing	Shirt	169
Footwear	Sandals	160
Footwear	Shoes	150
Footwear	Sneakers	145
Outerwear	Jacket	163
Outerwear	Coat	161

Figure 5:Top products per category based on purchase frequency.

5.3 Customer Behavior & Loyalty

- Customer segmentation: New, Returning, Loyal
- Relationship between repeat purchases and subscription status

Advanced SQL concepts applied include:

- Conditional logic using CASE

- Aggregations and subqueries
- Window functions (ROW_NUMBER)

customer_segment	customer_count
Loyal	3116
Returning	701
New	83

Figure 6: Customer segmentation and subscription relationship.

6. Key Insights

- Loyal customers represent the largest segment, highlighting strong retention potential
- Certain products show high reliance on discounts, posing potential margin risks
- Repeat buyers demonstrate higher subscription likelihood
- Revenue contribution varies significantly across age groups
- Express shipping users tend to have slightly higher average spending

7. Data Visualization

An interactive dashboard was developed to communicate insights effectively, featuring:

- KPI cards for customer count, average spend, and ratings
- Revenue breakdowns by category and age group
- Customer segmentation visuals
- Filters for gender, category, and shipping type

Figure 7:Power BI dashboard summarizing customer behavior insights

8. Business Recommendations

- Strengthen loyalty programs to convert repeat buyers into subscribers
- Review discount strategies for products with high discount dependency
- Focus marketing efforts on high-revenue age groups
- Optimize shipping incentives linked to higher spending behaviour

9. Limitations & Future Work

- No time-based data available for cohort or churn analysis
- Lack of cost data prevents margin analysis
- Future work could include predictive modelling for subscription conversion

10. Tools & Skills Demonstrated

- **Python**: pandas, data cleaning, feature engineering
- **SQL**: BigQuery, aggregations, window functions, CASE logic
- **Power BI**: dashboard design, KPI reporting
- **Analytics Skills**: business reasoning, segmentation, insight communication