Customer Shopping Behavior Analysis

Project Description:

This project focuses on analyzing customer shopping behavior for a retail business to understand purchasing patterns, customer loyalty, and key factors influencing buying decisions.Using a real-world consumer behavior dataset, the analysis aims to identify trends across demographics, product categories, discounts, subscriptions, and shipping preferences.

The project follows an end-to-end data analytics workflow, starting from raw data preparation to business-ready insights. Python is used for data cleaning and exploratory analysis, SQL for structured business queries, and Power BI for interactive dashboard development. The final output supports data-driven decisions to improve sales performance, customer engagement, and marketing strategies.

Dataset Summary -

- Rows: 3,900
- Columns: 18
- Key Features: 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, Frequency of Purchases, Review Rating, Shipping Type)
- Missing Data: 37 values in Review Rating column

Exploratory Data Analysis using Python

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.

```
[4]: df.head()
```

[4]:

	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	Previous Purchases	Payn Met
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Ve
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Cr
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	Pa
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	Pa

- 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.isnull().sum()
```

```
Customer ID                 0
Age                         0
Gender                      0
Item Purchased              0
Category                    0
Purchase Amount (USD)       0
Location                    0
Size                        0
Color                       0
Season                      0
Review Rating              37
Subscription Status         0
Shipping Type               0
Discount Applied            0
Promo Code Used             0
Previous Purchases          0
Payment Method              0
Frequency of Purchases      0
dtype: int64
```

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

```
Customer ID                 0
Age                         0
Gender                      0
Item Purchased              0
Category                    0
Purchase Amount (USD)       0
Location                    0
Size                        0
Color                       0
Season                      0
Review Rating               0
Subscription Status         0
Shipping Type               0
Discount Applied            0
Promo Code Used             0
Previous Purchases          0
Payment Method              0
Frequency of Purchases      0
dtype: int64
```

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

```
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')
```

- Feature Engineering:
  Created age_group column by binning customer ages.
  Created purchase_frequency_days column from purchase data.

```
df[['purchase_frequency_days','frequency_of_purchases']].head(10)
```

	purchase_frequency_days	frequency_of_purchases
0	14.0	Fortnightly
1	14.0	Fortnightly
2	7.0	Weekly
3	7.0	Weekly
4	365.0	Annually
5	7.0	Weekly

	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

- Data Consistency Check: Verified if discount_applied and promo_code_used were redundant; dropped promo_code_used.
- Database Integration: Connected Python script to SQLserver and loaded the cleaned DataFrame into the database for SQL analysis.

Data Analysis using Sql :

1. **Revenue by gender :** Total revenue generated by male vs. female customers :

	gender	revenue
1	Male	157890
2	Female	75191

2. **High spending discount user** : customers used a discount but still spent more than the average purchase amount :

	customer_id	purchase_amount
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
11	24	88
12	29	94
13	32	79
14	33	67
15	35	91
16	37	69
17	40	60
18	41	76
19	43	100

3. **Higher Review Rating Product :** product with more than high average review rating

	item_purchased	avg_review_rating
1	Gloves	3.86142857142857
2	Sandals	3.844375
3	Boots	3.81875
4	Hat	3.8012987012987
5	Skirt	3.78481012658228

4. **Compare Standard and Express Shipping. :** compare with there average purchased amount

Results | Messages

	shipping_type	average_purchase_amount
1	Standard	58
2	Express	60

5. Subscriber VS Non - Subscriber : who spent More

	subscription_status	total_customer	total_spent	Average_spent
1	No	2847	170436	59
2	Yes	1053	62645	59

6. Top 5 Product :

	item_purchased	total_purchases	discounted_purchases	discount_percentage
1	Hat	154	77	50.000000000000
2	Sneakers	145	72	49.655172413793
3	Coat	161	79	49.068322981366
4	Sweater	164	79	48.170731707317
5	Pants	171	81	47.368421052631

7. Customer Segmentation : Segment customers into New, Returning, and Loyal based on their total number of previous purchases .

	customer_segment	Number of Customers
1	Returning	701
2	Loyal	3116
3	New	83

8. Top 3 Product in Each Category :

	item_rank	item_purchased	category	purchase_count
1	1	Jewelry	Accessories	171
2	2	Belt	Accessories	161
3	3	Sunglasses	Accessories	161
4	1	Blouse	Clothing	171
5	2	Pants	Clothing	171
6	3	Shirt	Clothing	169
7	1	Sandals	Footwear	160
8	2	Shoes	Footwear	150
9	3	Sneakers	Footwear	145
10	1	Jacket	Outerwear	163
11	2	Coat	Outerwear	161

9. Repeat Buyers &Subscription: Checked whether customers with >5 purchases are more likely to subscribe.

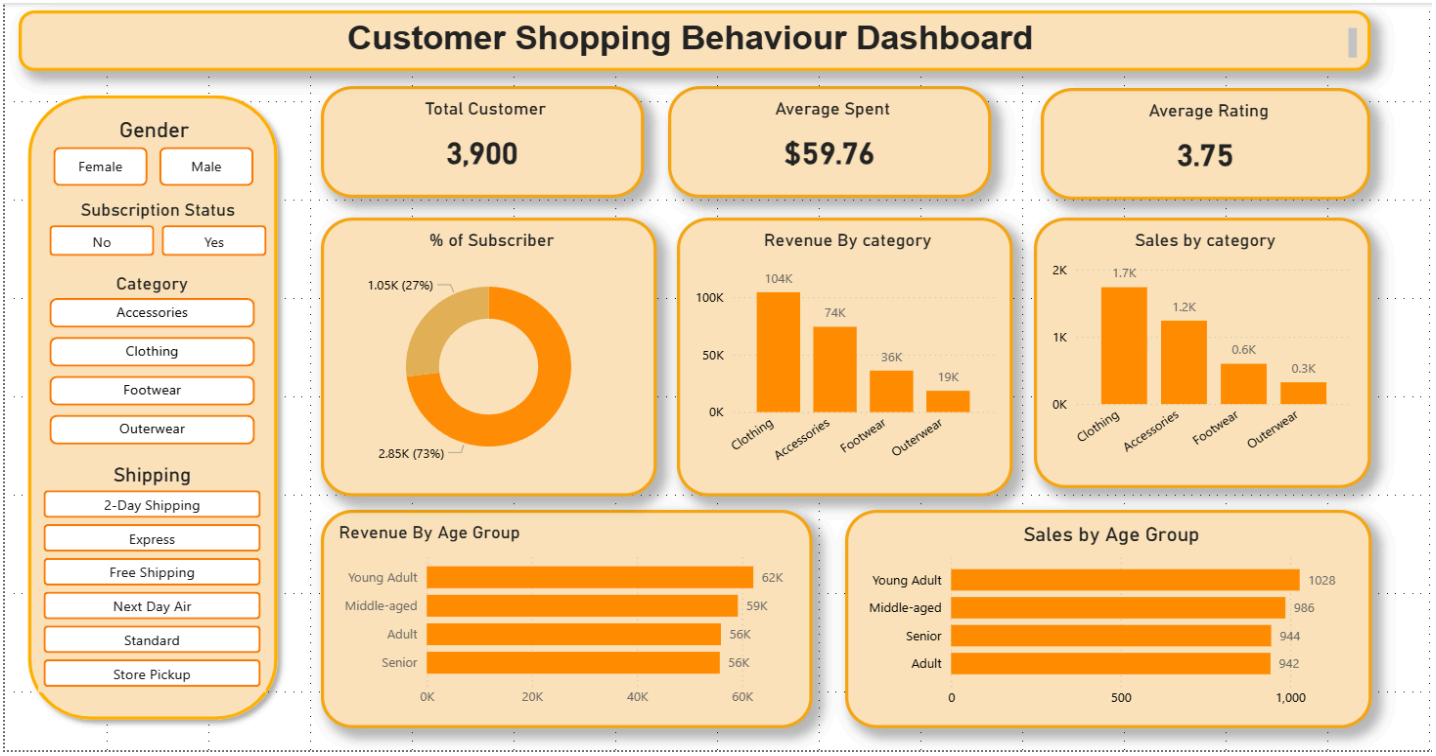
	subscription_status	repeat_buyers
1	No	2518
2	Yes	958

10 . Revenue Contribution By Age Group

Results | Messages

	age_group	total_spent
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

Power Bi : Customer Behaviour Dashboard



Business Recommendations

Based on the customer shopping behavior analysis, the following strategic recommendations are proposed to improve revenue growth, customer retention, and operational efficiency.

1. Strengthen Customer Retention

Customer segmentation indicates a strong base of loyal customers but relatively fewer new customers. The business should introduce structured loyalty programs and targeted onboarding offers to retain high-value customers while improving new customer acquisition.

Impact: Increases customer lifetime value and reduces acquisition costs.

2. Optimize Discount Strategy

Several high-selling products show heavy reliance on discounts, which may negatively impact profit margins. Discount usage should be strategically limited for top-performing products and redirected toward low-performing or seasonal items.

Impact: Maintains sales volume while improving profitability.

3. Focus on High-Performing Categories and Products

Revenue and sales analysis shows that a few product categories and top-ranked items contribute disproportionately to overall revenue. Marketing and promotional efforts should prioritize these categories while using cross-selling strategies to improve basket size.

Impact: Maximizes return on marketing investment.

4. Improve Subscription Value Proposition

Subscribed and non-subscribed customers show similar average spending, indicating the need for stronger subscription incentives. Enhancing subscription benefits such as exclusive offers or faster shipping can improve adoption and recurring revenue.

Impact: Drives predictable revenue and customer engagement.

5. Target High-Revenue Age Groups

Young adult and middle-aged customers contribute the highest share of revenue. Personalized marketing campaigns and targeted product offerings should focus on these segments.

Impact: Improves conversion rates and campaign effectiveness.

6. Leverage Product Ratings and Shipping Insights

High-rated products perform better in sales, and express shipping is associated with higher average spend. These insights should be used for product positioning and premium service offerings.

Impact: Enhances customer experience and increases average order value.