Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior using transactional data from 3900 purchases across various product categories. The aim is to uncover insights into spending patterns, customer segments, product preferences, and subscription behavior to guide strategic business decisions.

2. Dataset Summary

- Rows: 3900

- Columns: 20

- 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

2 Dataset Collection

The data was sourced from Kaggle and was loaded into python for EDA

```
[1]: import pandas as pd
     import numpy as np
```

```
[2]: df=pd.read_csv('shopping_trends .csv')
```

```
[3]: df
```

[3]:

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method	Shipping Type	Discount Applied	Promo Code Used
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Credit Card	Express	Yes	Yes
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Bank Transfer	Express	Yes	Yes
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Cash	Free Shipping	Yes	Yes
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	PayPal	Next Day Air	Yes	Yes
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Cash	Free Shipping	Yes	Yes

3. Data Cleaning & Preparation
• *Missing Data Handling:*

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

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 19 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  Payment Method           3900 non-null   object
 13  Shipping Type            3900 non-null   object
 14  Discount Applied         3900 non-null   object
 15  Promo Code Used          3900 non-null   object
 16  Previous Purchases       3900 non-null   int64
 17  Preferred Payment Method 3900 non-null   object
 18  Frequency of Purchases   3900 non-null   object
dtypes: float64(1), int64(4), object(14)
memory usage: 579.0+ KB
```

```
df.isnull()
```

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3895	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3896	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3897	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3898	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3899	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False

3900 rows × 19 columns

Column Standardization:

Renamed the column to SNAKE CASE for better coding and readability

```
### columns tillesn need to be professionl with uderscore and lowcase ####
df.columns =df.columns.str.lower()
df.columns=df.columns.str.replace(' ','_')
```

```
df
```

	customer_id	age	gender	item_purchased	category	purchase_amount_(usd)	location	size	color	season	review_rating	subscription_status	payme
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Ba
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	
...	...	...	...	...	...	...	...	...	...	...	...	...	
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	
3898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No	
3899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No	Ba

4. Feature creation

• Created age group column by binning customer ages.

• Created purchase_frequency_days column from purchase data.

5. Data Consistency

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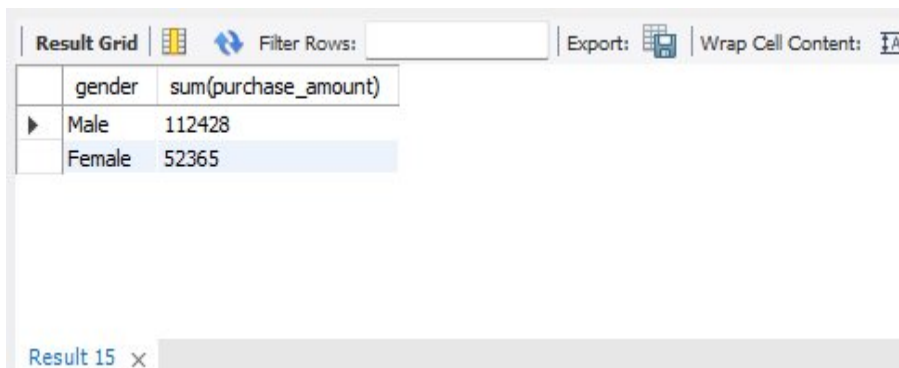
6.Data Export

Saving data into a file format (Excel / CSV) and exported to MySQL

7. Data Analysis using SQL

Structured analysis was performed in MySQL to answer key business questions such as:

• Revenue by Gender – Compared total revenue generated by male vs. female customers.

gender	sum(purchase_amount)
Male	112428
Female	52365

Result 15

• Top 5 Products with the highest revenue by comparing total revenue generated across all products.

item_purchased	average_rating
Gloves	3.89
Sandals	3.85
Boots	3.82
Hat	3.81
Shorts	3.80

• Average price of the mode of shipping used

shipping_type	round(avg(purchase_amount),2)
Express	61.34
Standard	58.21

• Comparing average spend and total revenue between Subscribers and non-Subscribers to find out if more customers subscribe

subscription_status	total_customers	avg_spend	total_revenue
Yes	759	59.12	44869
No	2010	59.66	119924

• Item which have the highest number of purchases with discounts applied

item_purchased	discount_rate
Sneakers	51.40
Jewelry	50.00
Hat	49.57
Sweater	49.56
Coat	47.97

- Top 3 most purchased products within each category

category	item_purchased	purchase_count
Accessories	Sunglasses	121
Accessories	Hat	117
Accessories	Scarf	116
Clothing	Pants	127
Clothing	Shirt	126
Clothing	Dress	118
Footwear	Sandals	112

- Buyers who make more than five purchases that are more likely to subscribe

subscription_status	total_customers
Yes	689
No	1787

- Revenue Contribution by Group

age_group	total_revenue
young adult	42342
middle-aged	42058
adults	40329
Senior	40064

8. Data visualization with Power BI

This Power BI report analyzes customer behavior to identify key purchasing patterns, revenue drivers, and customer segments.

The dashboard shows that Clothing and Accessories generate the highest sales and revenue, with Young Adults contributing the most to overall sales. Most customers are non-subscribers, highlighting an opportunity to improve subscription adoption and customer retention.

Overall, the report provides actionable insights to support data-driven marketing, sales optimization, and customer engagement strategies.

9. Business Recommendations

• Boost Subscriptions – Promote exclusive benefits for subscribers.

• Customer Loyalty Programs – Reward repeat buyers to move them into the 'Loyal' segment.

• Review Discount Policy – Balance sales boosts with margin control.

• Product Positioning – Highlight top-rated and best-selling products in campaigns.

• Targeted Marketing – Focus efforts on high-revenue age groups and express shipping users.