Customer Shopping Behaviour Analysis

Problem Statement

A leading retail company wants to better understand its customer's shopping behaviour in order to improve sales, customer satisfaction, and long-term loyalty. The management team has noticed changes in purchasing patterns across demographics, product categories and sales channels (online vs offline). They are particularly interested in uncovering which factors, such as discounts, revies, seasons, or payment preferences, drive consumer decisions and repeat purchases.

The client has asked **"How can the company leverage consumer shopping data to identify trends, improve customer engagement, and optimise marketing and product strategies?"**

1. Project overview

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

This will show a full pipeline of analysis starting from Excel → Python → SQL → Power Bi

2. Dataset Summary

Rows	3900
Columns	18
Customer demographics	Age, Gender, Location, Subscription Status
Purchasing details	Item purchased, Category, Purchase Amount, Season, Size, Colour
Shopping behaviour	Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type
Missing data	37 values in Review Rating column

3. Exploratory Data Analysis using Python

Begin with data preparation and cleaning in Python

- Load data: imported the dataset with pandas
- Initial exploration: Used df.info () to check the structure and .describe() for summary statistics

	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	Payment Method	Frequency of Purchases
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3900.000000	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN	6	7
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	NaN	677	584
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	25.351538	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	14.447125	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	1.000000	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	50.000000	NaN	NaN

```
<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
```

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

```
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
```

```
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 standardisation:** the columns were renamed to snake case for better readability and documentation.
- **Feature engineering:**
    - Created age_group column by binning customer ages
    - Created purchase_frequency_days column from 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 PostgreSQL and loaded the cleaned Data Frame into the database for SQL analysis.

4. Data Analysis using SQL

Structured analysis in Microsoft SQL server was performed to answer key business questions:

1. **Revenue by Gender**: compared total revenue generated by male vs female customers

	gender	Revenue
1	Male	157890
2	Female	75191

2. **High-spending discount users**: identified customers who used discounts but still spent above the average purchase amount (top 25 shown)

	customer_id	purchase_amount_usd
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
20	44	69
21	55	94
22	57	73
23	58	64
24	60	79
25	62	68

3. **Top 5 products by rating**: found products with the highest revenue ratings.

	item_purchased	Average Product Rating
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.8
5	Skirt	3.78

4. **Shipping type comparison**: compared average purchase amounts between standard and express shipping.

	shipping_type	Average purchase amount
1	Express	60.48
2	Standard	58.46

5. **Subscribers vs non-subscribers**: compared average spend and total revenue across subscription status

	subscription_status	total_customers	avg_spend	total_revenue
1	Yes	1053	59.49	62645
2	No	2847	59.87	170436

6. **Discount-dependent purchases**: identified 5 products with the highest percentage of discounted purchases.

	item_purchased	discount_rate
1	Hat	50.000000000000
2	Sneakers	49.660000000000
3	Coat	49.070000000000
4	Sweater	48.170000000000
5	Pants	47.370000000000

7. **Customer segmentation**: classified customers into New, Returning and Loyal segments based on purchase history.

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

8. **Top 3 products per category**: listed the most purchased products within each category.

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

9. **Repeat buyers and subscriptions:** checked whether customers with >5 purchases are more likely to subscribe.

	subscription_status	repeat_buyers
1	Yes	958
2	No	2518

10. **Revenue by age group**: calculated total revenue contribution of each age group.

	age_group	total_revenue
1	Young Adult	62143
2	Middle Aged	59197
3	Adult	55978
4	Senior	55763

5. Power BI Dashboard

An interactive dashboard was built in Power BI to present some insights visually.

6. Business Recommendations

Boost Subscriptions – Promote exclusive benefits for subscribers

Customer loyalty programmes – reward repeat buyers to move them into the ‘Loyal’ category

Review discount policy – balance sales boosts with margin control

Product positioning – highlight top-rated and best-selling products with campaigns

Target marketing – focus efforts on high- revenue age groups and express shipping users