PROJECT REPORT HEADINGS: CUSTOMER SHOPPING BEHAVIOR

1. Introduction and Project Setup

Project Overview: This project uses a dataset of 3,900 customer transactions across diverse categories to perform a thorough analysis of shopping habits. The primary objective is to reveal key behavioral insights regarding purchasing frequency, customer demographics, popular items, and the influence of subscriptions. The ultimate purpose is to equip the business with actionable data for informing strategic decision-making and optimizing customer engagement efforts.

Business Problem & Objective: A major retailer seeks deeper clarity on its consumer base to boost sales, enhance satisfaction, and cultivate lasting loyalty. The leadership team has observed shifting trends in purchases across diverse demographics, product lines, and sales platforms (physical and digital). Specifically, there is high interest in determining which key variables—including discounts, customer reviews, seasonality, or payment methods—are the main drivers of both immediate buying decisions and sustained repurchase behavior.
This project is therefore focused on analyzing the comprehensive consumer behavior data to address the central business inquiry:
"How can the organization utilize consumer shopping insights to pinpoint emerging trends, strengthen customer relationships, and refine its marketing and product development strategies?"

2. Data Preparation and Quality

Data Summary: This analysis is based on a structured dataset containing 3,900 transactional records and 18 distinct features.
The dataset encompasses several key feature groups for comprehensive analysis:
Customer Demographics: Details such as Age, Gender, Location, and Subscription Status.
Purchase Specifics: Information covering the Item Purchased, Product Category, Purchase Amount, Season, Size, and Color.
Shopping Behavior Indicators: Metrics including whether a Discount Applied or Promo Code Used, the count of Previous Purchases, the Frequency of Purchases, the Review Rating, and the Shipping Type.
Data Cleaning and Preprocessing: Based on the dataset summary, the data cleaning and preprocessing steps focused on addressing the identified missing values and ensuring consistent data types and formats across all 18 features before proceeding to detailed analysis.
Data Quality: There were 37 missing entries specifically within the Review Rating column that required preprocessing.
The approach to handling these missing entries was carried out with python by using the median rating of the product category of the review rating column since median is immune to outliers. I filled the 37 missing values with this calculated median.
I then verified and converted the data types of key columns to ensure they were suitable for numerical analysis, categorical grouping and other analysis.
I also renamed and converted all columns using both lower and snake casings for better readability and documentation.

```
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')
```

3. Exploratory Data Analysis (EDA)

Data Loading: Loaded the csv file into jupyter notebook using pandas as pd

```
df
```

	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
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14
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
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	2-Day Shipping	No	No	32
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	Store Pickup	No	No	41
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	Standard	No	No	24

Descriptive Statistics: I used df.info() and df.describe() to give an insight of the data summary

```
df.describe()
```

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3863.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.750065	25.351538
std	1125.977353	15.207589	23.685392	0.716983	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.800000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

```
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 (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
```

Feature Engineering: I created a new variable from age where the several numerical ages were put into manageable groups titled age_group

```
# create a new 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

I also created a new column titled purchase_freguency_days from purchase_data.

```
# create new column purchase_frequency_days

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)
```

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

	purchase_frequency_days	frequency_of_purchases
0	14	Fortnightly
1	14	Fortnightly
2	7	Weekly
3	7	Weekly
4	365	Annually

For data consistency check, I ensured that both the discount_applied and promo_code_used columns were the same and dropped promo_code_used.

Tools Used for EDA (Python/SQL): I used Python's pandas for statistical tests and MSSQL for aggregation queries.

4. Key Findings and Insights

1. Revenue by gender - compared the revenue generated by both female and male customers.

	gender	revenue
1	Male	157890
2	Female	75191

2. High spending discount customers – identified customers who used discount codes but still spent more than the average purchase amount.

customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90
22	62
24	88
29	94
32	79
33	67
35	91
37	69
40	60
41	76
43	100

3. Average purchase amount between shipping types – compared average purchase amounts between standard and express shipping

shipping_type	(No column name)
Standard	58
Express	60

4. Top 5 products by review rating – identified products with the highest average review ratings.

item_purchased	Average Product Rating
Gloves	3.860000
Sandals	3.840000
Boots	3.820000
Hat	3.800000
Skirt	3.780000

5. Subscribers vs Non-Subscribers – compared the average spend and total revenue between types of subscribers.

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

6. Top 5 discounted products – identified top 5 products with the highest percentage of discounts.

item_purchased	discount_rate
Hat	50.000000000000
Sneakers	49.660000000000
Coat	49.070000000000
Sweater	48.170000000000
Pants	47.370000000000

7. Customer segmentation – identified the number of customers across returning, loyal and new segments.

customer_segment	Number of Customers
Returning	701
Loyal	3116
New	83

8. Top 3 products per category – identified the most purchased products within different categories.

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

9. Repeat buyers and subscribers – identified if customers with more than 5 purchases were more likely to subscribe

subscription_status	repeat_buyers
Yes	958
No	2518

10. Revenue by age group – calculated revenue generated by different age groups.

age_group	total_revenue
Young Adult	62143
Middle-aged	59197
Adult	55978
Senior	55763

11. Peak seasons and location analysis – identified the total amount of purchases for clothing items in Kentucky across different seasons.

season	TotalClothingPurchases
Winter	11
Spring	9
Fall	8
Summer	2

12. Subscription value analysis – calculated and compared the average purchase amount for both subscribed and non-subscribed customers

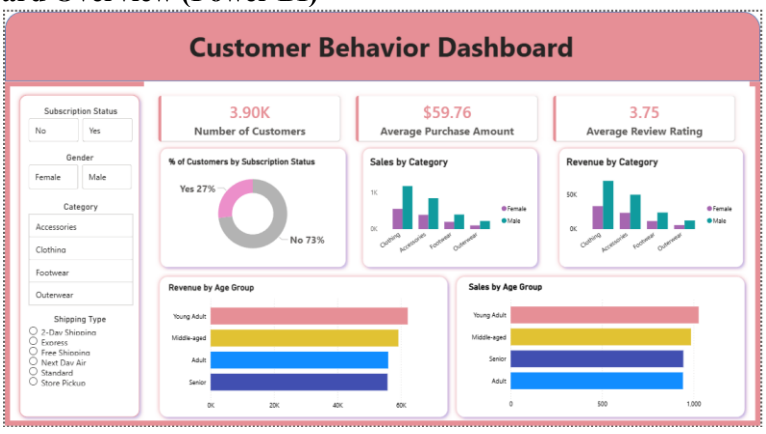
subscription_status	AvgPurchaseValue
Yes	59
No	59

13. Popular payment methods across demographics - identified various methods preferred by gender and age groups

AgeGroup	Gender	payment_method	TotalTransactions
18-25 Young Adult	Female	Debit Card	19
18-25 Young Adult	Female	Venmo	26
60+ Senior	Female	PayPal	26
18-25 Young Adult	Female	Credit Card	27
18-25 Young Adult	Female	Bank Transfer	30
60+ Senior	Female	Debit Card	31
18-25 Young Adult	Female	Cash	35
18-25 Young Adult	Female	PayPal	36
60+ Senior	Female	Bank Transfer	36
60+ Senior	Female	Venmo	36
60+ Senior	Female	Cash	48
60+ Senior	Female	Credit Card	51
26-40 Adult	Female	Cash	53
26-40 Adult	Female	Debit Card	54
26-40 Adult	Female	Bank Transfer	54
18-25 Young Adult	Male	PayPal	59
18-25 Young Adult	Male	Venmo	63
26-40 Adult	Female	Credit Card	65
18-25 Young Adult	Male	Bank Transfer	66
26-40 Adult	Female	Venmo	66
26-40 Adult	Female	PayPal	67
18-25 Young Adult	Male	Credit Card	68
18-25 Young Adult	Male	Debit Card	70
18-25 Young Adult	Male	Cash	72
41-60 Middle Age	Female	Cash	76
41-60 Middle Age	Female	Debit Card	77
60+ Senior	Male	Bank Transfer	79
60+ Senior	Male	Venmo	80
41-60 Middle Age	Female	Venmo	80
41-60 Middle Age	Female	Credit Card	80
60+ Senior	Male	Debit Card	82
60+ Senior	Male	PayPal	82
41-60 Middle Age	Female	Bank Transfer	83
60+ Senior	Male	Cash	84
60+ Senior	Male	Credit Card	88
41-60 Middle Age	Female	PayPal	92
26-40 Adult	Male	Bank Transfer	115
26-40 Adult	Male	Credit Card	118
26-40 Adult	Male	Cash	123
26-40 Adult	Male	PayPal	127
26-40 Adult	Male	Debit Card	130
26-40 Adult	Male	Venmo	131
41-60 Middle Age	Male	Bank Transfer	149
41-60 Middle Age	Male	Venmo	152
41-60 Middle Age	Male	Debit Card	173
41-60 Middle Age	Male	Credit Card	174
41-60 Middle Age	Male	Cash	179

5. Data Visualization and Presentation

Dashboard Overview (Power BI)



6. Conclusion and Recommendations

What was set to be achieved from data ingestion through EDA in python to analysis in SQL and visualization via PowerBI were achieved and relevant business insights were identified.
Be what it may, one can recommend that further handling of dataset should include

- Exploration of more metrices that could enhance business value.
- More discount rates should be offered especially to returning customers.

- Focus should be given to express shipping strategy
- Marketing can be done to promote express shipping among customers.
- Loyalty points should be awarded to repeat customers to encourage more shopping.
- Marketing should be done to attract high revenue generating age groups.

Ruby Ineba Briggs
ML Engineer/Data Scientist