Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior 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 behavior to guide strategic business decisions.

2. Dataset Summary

-Rows: 3900

-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.

3. Exploratory Data Analysis using Python

Began with data preparation and cleaning using python.

- **Data Loading**: Imported dataset using pandas.

- **Initial Exploration**: Used df.info() to check 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

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

```
# check if missing data or null values are present

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
```

```
# imputting missing and null values of the rating review column with median considering the category

df['Review Rating'] = df.groupby('Category')['Review Rating'].transform(lambda x: x.fillna(x.median()))
```

```
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**: Renames columns to **snake case** for better readability and documentation.

```
df.columns = df.columns.str.lower()
df.columns = df.columns.str.replace(' ', '_')
```

```
df.columns
```

```
Index(['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'],
      dtype='object')
```

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

- **Data Consistency** : Verified if discount_applied and promo_code_used were redundant; dropped promo_code_used.

```
(df['discount_applied'] == df['promo_code_used']).all()
```

```
np.True_
```

```
df = df.drop('promo_code_used', axis=1)
```

```
df.columns
```

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

- **Database Integration**: Connected Python script to PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

```
# Connect to PostgreSQL
#Replace placeholders with actual details
from sqlalchemy import create_engine
username = "postgres"
password = "1234"
host = "localhost"
port = "5432"
database = "customer_behavior"

engine = create_engine(f"postgresql+psycopg2://{username}:{password}@{host}:{port}/{database}")

# Load DataFrame into PostgreSQL
table_name = "customer"  #choose a table name
df.to_sql(table_name, engine, if_exists="replace", index=False)

print(f"Data Successfullt loaded into table '{table_name}' in database '{database}'.")
```

```
Data Successfullt loaded into table 'customer' in database 'customer_behavior'.
```

4. Data Analysis using SQL (Business Transactions)

We performed structured analysis in PostgreSQL to answer key business questions:

1. **Revenue by Gender**

	gender text	revenue numeric
1	Female	75191
2	Male	157890

2. **High Spending Discount Users**

	customer_id bigint	purchase_amount bigint
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

Total rows: 839 | Query complete 00:00:00.115

3. **Top 5 Products by Rating**

	item_purchased text	Average Product Rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

4. **Shipping Type Comparison**

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

5. **Subscribers vs. Non-Subscribers**

	subscription_status text	avg_spend numeric	total_revenue numeric	total_customers bigint
1	Yes	59.49	62645.00	1053
2	No	59.87	170436.00	2847

6. **Discount-dependent Products**

	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.00
3	Coat	49.00
4	Sweater	48.00
5	Pants	47.00

7. **Customer Segmentation**

	customer_segment text	Number_of_Customers bigint
1	Loyal	3116
2	New	83
3	Returning	701

8. Top 3 Products per Category

	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	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 & Subscriptions

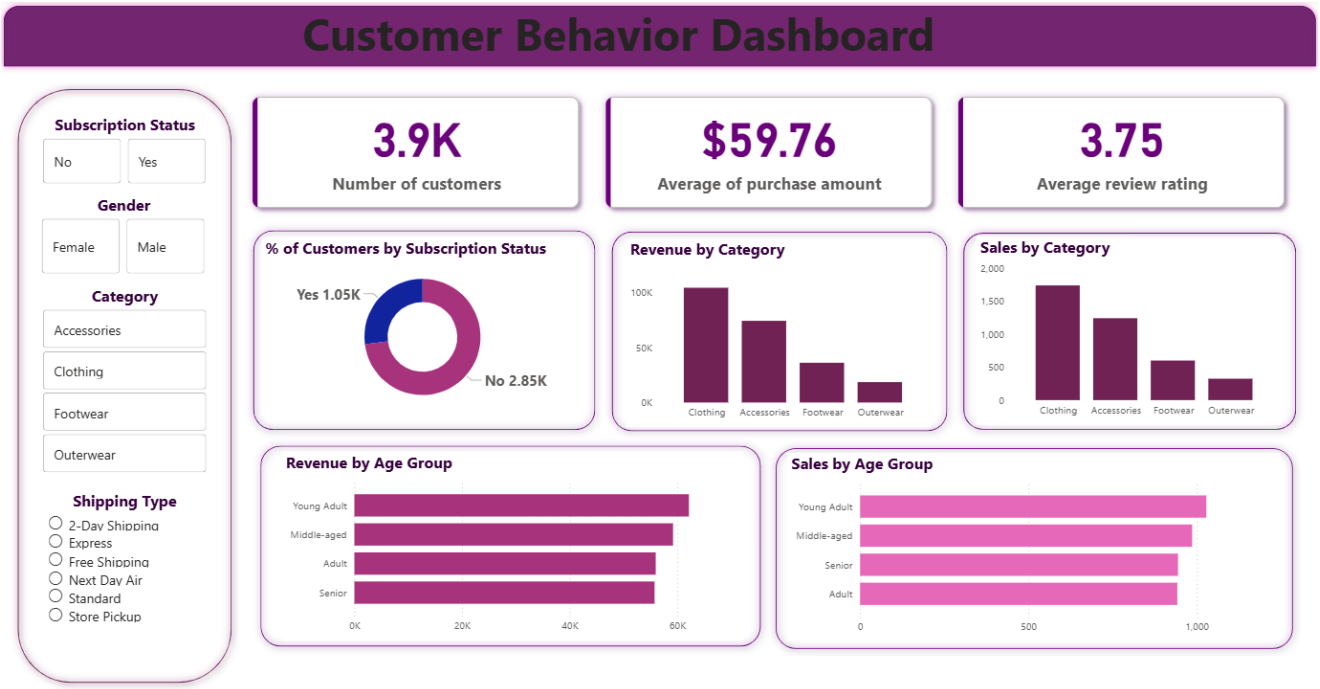
	subscription_status text	repeat buyers bigint
1	No	2518
2	Yes	958

10. Revenue by Age Group

	age_group text	total_revenue numeric
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

5. Dashboard in Power BI

Here is a dynamic dashboard created in Power BI to present insights visually.



6. Business Recommendations

- **Launch a Subscription Conversion Campaign**: With the customers not subscribed, creating a targeted campaign offering an incentive to convert this large segment into recurring revenue.
- **Grow the Female Customer Segment:** Customer base is 68% male, which represents a significant opportunity. Analyzing the purchasing habits of current female customers which is 32% and launch targeted marketing campaigns to attract and retain more female shoppers.
- **Implement a Cross-Sell Strategy:** Since clothing is the top category, using it as a base to cross-sell items for Accessories and Footwear. Recommend matching accessories or shoes at checkout to increase the average order value.

- **Segment Marketing by Purchase Frequency**: Stopping sending the same marketing emails to everyone. Creating different email cadences: send “New Arrivals” to weekly shoppers and “Seasonal Stock-up” offers to quarterly or annual shoppers.