Customer Behaviour Analysis Report

---

1. Summary

This report analyzes customer shopping behavior using a dataset of **3,900 customer transactions**. The objective is to understand purchasing patterns, customer demographics, product performance, subscription impact, and revenue distribution. The analysis was conducted using **Python for exploratory data analysis (EDA)**, **PostgreSQL for advanced querying**, and **Power BI for interactive visualization**.

Key findings indicate that **Clothing is the highest revenue-generating category**, **subscribed customers contribute higher average spend and total revenue**, and **younger age groups are the most valuable customer segment**.

---

2. Exploratory Data Analysis (EDA) Using Python

EDA was performed using **Pandas** to assess data quality, structure, and key statistical properties.

Dataset Overview

- Total records: **3,900**
- Total features: **18**
- Data types:
    - Numerical: Age, Purchase Amount, Review Rating, Previous Purchases
    - Categorical: Gender, Category, Shipping Type, Subscription Status, etc.

Data Loading

- **I imported the dataset using Pandas library.**
- **Initial inspection was performed to understand the dataset structure.**

Initial Exploration

- **I used df.info() to examine:**
    - **Number of records and columns**
    - **Data types**
    - **Presence of missing values**

```
<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
```

- **I used df.describe() to generate summary statistics for both numerical and categorical features, including:**
    - **Mean age (~44 years)**
    - **Average purchase amount (~$59.76)**
    - **Average review rating (~3.75)**
    - **Distribution of categories, genders, and purchase frequency**

	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

- **Identified 37 missing values in the review_rating column.**
- **Missing ratings were imputed using the median review rating of each product category, ensuring minimal distortion of customer satisfaction metrics.**

```
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
```

Column Standardization

- **I converted column names to snake_case for better readability and documentation.**
- **Renamed purchase_amount_(usd) to purchase_amount for consistency.**

```
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

- **I created a new age_group column by binning customer ages into:**
  - **Young Adult**
  - **Adult**
  - **Middle-aged**
  - **Senior**

	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
7	27	Young Adult
8	26	Young Adult
9	57	Middle-aged

- **I created purchase_frequency_days by mapping textual purchase frequency (e.g., Weekly, Monthly) to numeric day values.**

	purchase_frequency_days	frequency_of_purchases
0	14	Fortnightly
1	14	Fortnightly
2	7	Weekly
3	7	Weekly
4	365	Annually
5	7	Weekly
6	90	Quarterly
7	7	Weekly
8	365	Annually
9	90	Quarterly

Data Consistency Check

- **Verified that discount_applied and promo_code_used contained identical values across all records.**
- **Dropped the redundant promo_code_used column to simplify the dataset.**

	discount_applied	promo_code_used
0	Yes	Yes
1	Yes	Yes
2	Yes	Yes
3	Yes	Yes
4	Yes	Yes
5	Yes	Yes
6	Yes	Yes
7	Yes	Yes
8	Yes	Yes
9	Yes	Yes

Database Integration

- **The cleaned dataset was connected to PostgreSQL using SQLAlchemy.**
- **Data was successfully loaded into a relational table for advanced SQL analysis.**

```python
from sqlalchemy import create_engine

# Step 1: Connect to PostGreSQL
username = "postgres"  # Default user
password = "---------"  # Password set during installation
host = "localhost"  # if running locally
port = "5432"  # Default PostGreSQL port
database = "customer_behavior"  # Database created in PgAdmin4

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

# Step 2: Load DataFrame into PostGreSQL
table_name = "customer"
df.to_sql(table_name, engine, if_exists="replace", index=False)

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

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

Key EDA Insights

- Average purchase amount: **$59.76**
- Average review rating: **3.75**
- Majority of customers are **male**.
- Clothing is the most frequently purchased category.
- Most customers are **repeat buyers**, indicating strong retention.

---

3. Data Analysis Using SQL

I loaded the cleaned dataset into PostgreSQL for business-driven analytical queries.

Key Analytical Findings

- **Revenue by Gender**:
  Male customers generate higher total revenue due to higher transaction volume.

	gender text	revenue numeric
1	Female	75191
2	Male	157890

- **Discount Effectiveness**:
  Several customers using discounts still spent **above the average purchase amount**, showing discounts do not necessarily reduce revenue.

	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

Total rows: 839 | Query complete 00:0

- **Product Performance**:
  The top 5 products with the highest average review ratings demonstrate strong customer satisfaction.

	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

- **Shipping Type Comparison**:
  Average purchase amounts for **Standard** and **Express** shipping are comparable, indicating shipping speed does not significantly impact spend.

	shipping_type text	average_puchase_amount numeric
1	Standard	58.46
2	Express	60.48

- **Subscription Impact**:
  Subscribed customers:
  - Spend more on average
  - Generate higher total revenue
  - Represent a smaller but more valuable customer segment

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

- **Customer Segmentation**:
    - New customers (1 purchase)
    - Returning customers (2–10 purchases)
    - Loyal customers (>10 purchases)

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

- **Category-Level Insights**:
  Clothing dominates both **sales volume and revenue**, followed by Accessories and Footwear.

	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

- **Age Group Revenue Contribution**:
  Young Adults contribute the highest total revenue among all age groups.

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

---

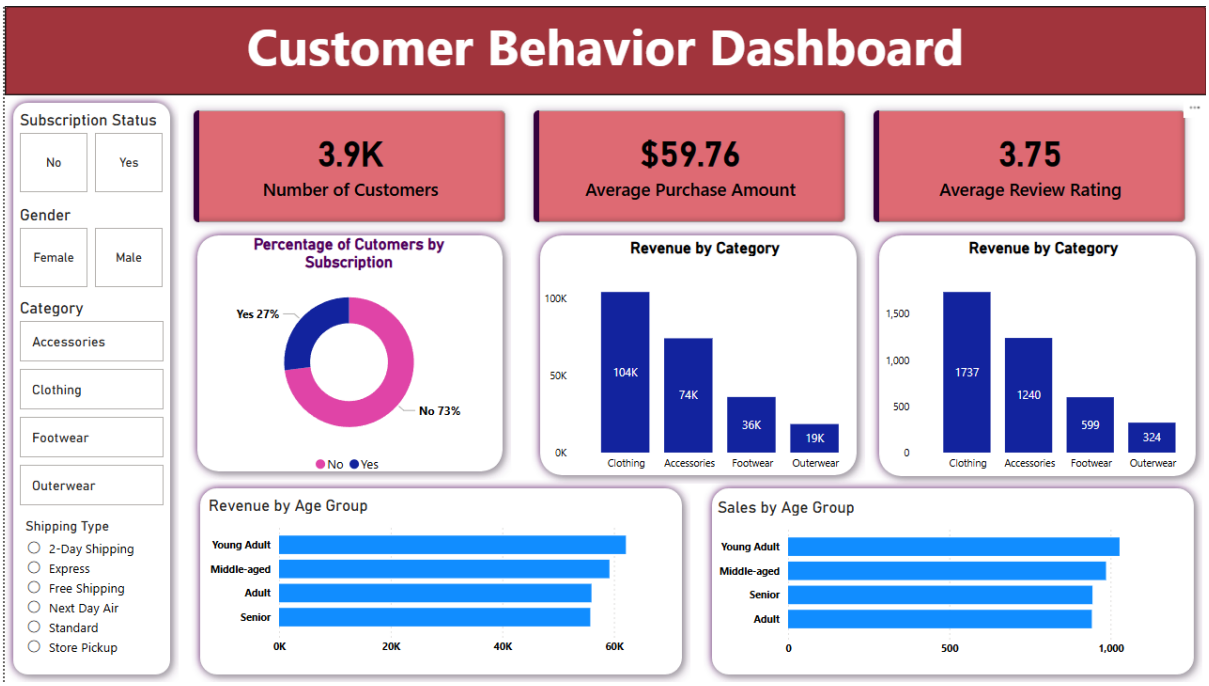
4. Dashboard

An interactive **Power BI dashboard** was developed to visualize customer behavior and business KPIs.

Dashboard Highlights

- Total customers: **~3.9K**
- Average purchase amount: **~$59.76**
- Average review rating: **~3.75**
- Subscription distribution:
    - 73% Non-subscribers
    - 27% Subscribers
- Revenue and sales breakdown by:
    - Product category
    - Age group
    - Subscription status

The dashboard allows dynamic filtering by **gender, category, subscription status, and shipping type**, enabling deeper exploratory analysis.



---

5. Business Recommendations

1. **Focus on Clothing Category**
   Allocate more inventory and marketing resources to Clothing, as it consistently generates the highest revenue and sales volume.

2. **Expand Subscription Programs**
   Encourage repeat buyers to subscribe through targeted offers, as subscribers demonstrate higher customer lifetime value.

3. **Optimize Discount Strategy**
   Use discounts strategically on high-performing products to increase basket size rather than applying broad discounts.
4. **Target Young Adult Segment**
   Design personalized campaigns for Young Adults, who contribute the highest revenue and purchase frequency.
5. **Leverage High-Rated Products**
   Promote top-rated products with premium positioning instead of discount-heavy promotions.
6. **Retention-Based Marketing**
   Develop loyalty programs aimed at converting Returning customers into Loyal customers.