Customer Shopping Behavior Analysis

1. **Project Overview** This project analyzes customer shopping behavior using transactional data from 3,900 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: 3,900

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 We began with data preparation and cleaning in Python:

- Data Loading: Imported the dataset using pandas.
- Initial Exploration: Used df.info() to check structure and .describe() for summary statistics.

Statistics Summary of Numerical columns

```
[4]: df.describe(include= "all")
```

[4]:

	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
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	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.
- **Column Standardization**: Renamed columns 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 MySQL and loaded the cleaned DataFrame into the database for SQL analysis.

4. Data Analysis using SQL (Business Transactions)

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

Revenue by Gender

Q1. What is the total revenue generated by male vs. female customers?

```sql
/*Q1. What is the total revenue generated by male vs. female customers?*/

select gender, SUM(purchase_amount) as revenue
from mytable
group by gender;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

gender	revenue
Male	157890
Female	75191

High-Spending Discount Users

Q2. Which customers used a discount but still spent more than the average purchase amount?

```sql
/*Q2. Which customers used a discount but still spent more than the average purchase amount?*/
select customer_id,
 purchase_amount
from mytable
where discount_applied = 'Yes' and
purchase_amount >= (select AVG(purchase_amount) from mytable);
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

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

mytable 2

Output

Action Output

#	Time	Action	Message
1	15:51:23	select gender, SUM(purchase_amount) as revenue from mytable group by gender LIMIT 0, 1000	2 row(s) returned
2	15:52:08	select customer_id, purchase_amount from mytable where discount_applied = 'Yes' and purchase_amount >= (s...	839 row(s) returned

Top 5 Products by Rating

Q3. Which are the top 5 products with the highest average review rating?

```
16    /*Q3. Which are the top 5 products with the highest average review rating?*/
17 •  SELECT
18    item_purchased , ROUND(avg(review_rating),2) as "Average Product Rating"
19    from mytable
20    group by item_purchased
21    order by avg(review_rating) desc
22    limit 5;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

item_purchased	Average Product Rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Skirt	3.78

Shipping Type Comparison

Q4. Compare the average Purchase Amounts between Standard and Express Shipping.

```
25    /*Q4. Compare the average Purchase Amounts between Standard and Express Shipping.*/
26 •  SELECT shipping_type, ROUND(avg(purchase_amount),2) as "avg purchase amount"
27    FROM mytable
28    where shipping_type in ('Standard','Express')
29    group by shipping_type;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

shipping_type	avg purchase amount
Express	60.48
Standard	58.46

Subscribers vs. Non-Subscribers

Q5. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers.

```
31  /*Q5. Do subscribed customers spend more? Compare average spend and total revenue
32   between subscribers and non-subscribers.*/
33  SELECT subscription_status,
34         COUNT(customer_id) AS total_customers,
35         ROUND(AVG(purchase_amount),2) AS avg_spend,
36         ROUND(SUM(purchase_amount),2) AS total_revenue
37  FROM mytable
38  GROUP BY subscription_status
39  ORDER BY total_revenue,avg_spend DESC;
40
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

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

Discount-Dependent Products

Q6. Which 5 products have the highest percentage of purchases with discounts applied?

```
40
41  /*Q6. Which 5 products have the highest percentage of purchases with discounts applied?*/
42  SELECT item_purchased,
43         ROUND(100.0 * SUM(CASE WHEN discount_applied = 'Yes' THEN 1 ELSE 0 END)/COUNT(*), 2) AS discount_rate
44  FROM mytable
45  GROUP BY item_purchased
46  ORDER BY discount_rate DESC
47  LIMIT 5;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

item_purchased	discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

Customer Segmentation

Q7. Segment customers into New, Returning, and Loyal based on their total
-- number of previous purchases, and show the count of each segment.

```sql
/*Q7. Segment customers into New, Returning, and Loyal based on their total
-- number of previous purchases, and show the count of each segment.*/
 with customer_type as (
 SELECT customer_id, previous_purchases,
 CASE
     WHEN previous_purchases = 1 THEN 'New'
     WHEN previous_purchases BETWEEN 2 AND 10 THEN 'Returning'
     ELSE 'Loyal'
     END AS customer_segment
FROM mytable)
SELECT customer_segment, count(*) AS "Number of Customers"
FROM customer_type
GROUP BY customer_segment;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

customer_segment	Number of Customers
Loyal	3116
Returning	701
New	83

Top 3 Products per Category

Q8. What are the top 3 most purchased products within each category?

```sql
/*Q8. What are the top 3 most purchased products within each category?*/
WITH item_counts AS (
    SELECT category,
           item_purchased,
           COUNT(customer_id) AS total_orders,
           ROW_NUMBER() OVER (PARTITION BY category ORDER BY COUNT(customer_id) DESC) AS item_rank
    FROM mytable
    GROUP BY category, item_purchased
)
SELECT item_rank,category, item_purchased, total_orders
FROM item_counts
WHERE item_rank <=3;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

item_rank	category	item_purchased	total_orders
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	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

Repeat Buyers & Subscriptions

Q9. Are customers who are repeat buyers (more than 5 previous purchases) also likely to subscribe?

```
76    /*Q9. Are customers who are repeat buyers (more than 5 previous purchases) also likely to subscribe?*/
77 •  SELECT subscription_status,
78           COUNT(customer_id) AS repeat_buyers
79    FROM mytable
80    WHERE previous_purchases > 5
81    GROUP BY subscription_status;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

subscription_status	repeat_buyers
Yes	958
No	2518

Revenue by Age Group

Q10. What is the revenue contribution of each age group?

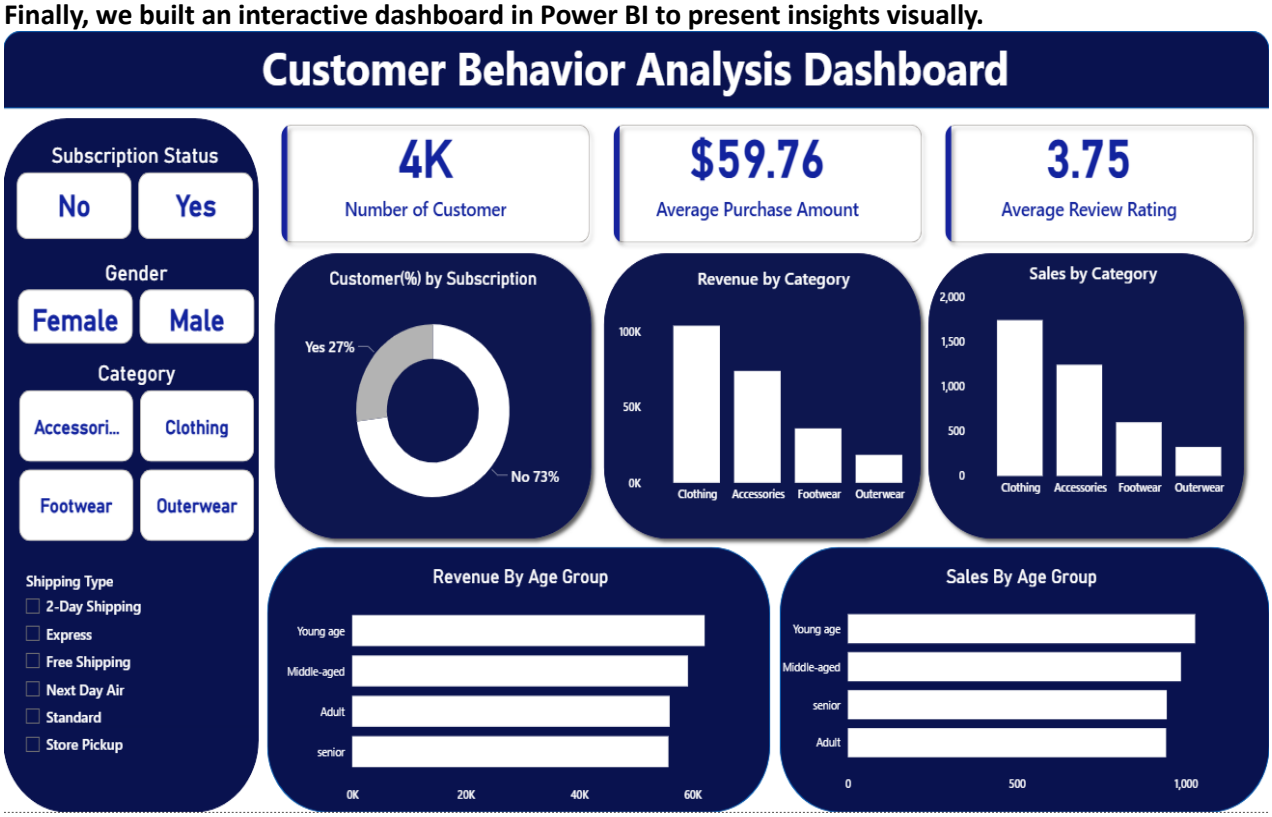
```
82
83    /*Q10. What is the revenue contribution of each age group?*/
84 •  SELECT age_group, SUM(purchase_amount) AS total_revenue
85    FROM mytable
86    GROUP BY age_group
87    ORDER BY total_revenue DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

age_group	total_revenue
Young age	62143
Middle-aged	59197
Adult	55978
senior	55763

5. Dashboard in Power BI

Finally, we built an interactive dashboard in Power BI to present insights visually.



6. 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.**