CUSTOMER SHOPPING BEHAVIOR ANALYSIS

---

1 Business Understanding

🎯 **Goal:** This project analyzes customer shopping behavior using transactional data from 3,900 purchases across various product categories. The objective is to uncover insights into spending patterns, customer segmentation, product preferences, and subscription behaviors to support data-driven strategic business decisions.

Business Questions:

1. What is total revenue generated by Female Vs Male customer?
2. Which customer used a discount but still spent more than the average purchase amount?
3. Which are the top 5 products with the highest average rating review?
4. Compare average purchase amounts between standard and express shipping.
5. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers?
6. Which 5 products have the highest percentage of purchases with discount applied?
7. Segment customers into new, returning and loyal based on their total number of previous purchases and show the count of each segment.
8. What are the top 3 most purchase product within each category?
9. Are customers who are repeated buyers(more than 5 previous purchases) also likely to subscribe?
10. What is the revenue contribution of each age group?

Key question:

"How can the company leverage consumer shopping data to identify trends, improve customer engagement, and optimize marketing and product strategies?"

---

2 Data Understanding (Exploration & Profiling)

Dataset summary:

- Data sources: POS(Point of Sales) System.
- Loaded Customer dataset Jupiter Notebook using python
- Data structure:
  - Rows: 3,900
  - Column: 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, Pervious purchases, Frequency of purchases, Review rating, Shipping type

- Data quality Checks:
  - Data types:
    - Float64 : 1
    - Int64: 4
    - Object: 13
  - Missing values:
    - 37 Values in Review rating
- Summarize descriptive statistics
  - Data loading: imported dataset using python
    - df = pd.read_csv("customer_shopping_behavior.csv")
  - Data structure/dimension: 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
```

  - Descriptive statistic summary: df.describe(include="all")

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Dis A
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	

Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3900	3900	3900	3900	3900.000000	3900	3900
2	6	2	2	NaN	6	7
No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
2847	675	2223	2223	NaN	677	584
NaN	NaN	NaN	NaN	25.351538	NaN	NaN
NaN	NaN	NaN	NaN	14.447125	NaN	NaN
NaN	NaN	NaN	NaN	1.000000	NaN	NaN
NaN	NaN	NaN	NaN	13.000000	NaN	NaN
NaN	NaN	NaN	NaN	25.000000	NaN	NaN
NaN	NaN	NaN	NaN	38.000000	NaN	NaN
NaN	NaN	NaN	NaN	50.000000	NaN	NaN

Tools: pandas, Excel, Python, Jupiter Notebook.

---

3 Data Cleaning (Data Preparation / Preprocessing)

🎯 **Goal:** Prepare **high-quality, consistent data** for analysis.

Data Cleaning:

- Handle missing data:
  - Checked for null values
    - df.isnull().sum()

```
Out[5]: 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
```

  - Imputed missing values in review rating column using the median rating of each product category.
    - df['Review Rating'] = df.groupby('Category')['Review Rating'].transform(lambda x: x.fillna(x.median()))

- Data consistent checks:
  - Verified that *discount_applied* and *promo_code_used* were redundant
    - (df['discount_applied'] == df['promo_code_used']).all()
  - Dropped *promo_code_used* column because of data redundancy.
    - df = df.drop('promo_code_used', axis=1)
- Standardized column name:
  - Renamed column name into snake case for better readability and documentation.

```
df.columns = df.columns.str.lower()
df.columns = df.columns.str.replace(' ','_')
df = df.rename(columns =
{'purchase_amount_(usd)':'purchase_amount'}) Treat outliers (IQR,
Winsorization).
```

- Feature Engineering:
  - Create age_group column by binning customer ages

```
# create a column age_group
labels = ['Young Adult','Adult','Middle-aged','Senior']
df['age_group'] = pd.qcut(df['age'], q=4, labels=labels)
```

  - Create purchase_frequency_days column from frequency_of_purchases column

```
# 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)
```

- Database integration:
  - Connected python script to SQLite and loaded the cleaned DataFrame into the database for SQL analysis

Deliverable: *Clean dataset, EDA Notebook.*

Tools: pandas, Python scripts, SQL.

---

4 Data analysis using SQL (Business Transaction)

🎯 **Goal:** Execute queries to find out the answers to the business questions.

Key insights:

- Revenue by gender: Compare total revenue generated by Male Vs Female
  - Q1. What is total revenue generated by Female Vs Male customer?

```
SELECT gender,
sum(purchase_amount) as revenue
FROM customer
GROUP by gender;
```

gender	revenue
Female	75191
Male	157890

- High spending discount users/ high-value customers: Identify the customers who used discounts but still spent above the average purchase.
  - Q2. Which customer used a discount but still spent more than the average purchase amount?

```
SELECT customer_id,
purchase_amount
FROM customer
WHERE discount_applied = 'Yes'
AND purchase_amount >= (SELECT
avg(purchase_amount) FROM
customer)
```

	customer_id	purchase_amount
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68
7	13	72

- Top 5 product by rating: Identify the products with highest rating.
  - Q3. Which are the top 5 products with the highest average rating review?

```
SELECT item_purchased,
round(avg(review_rating) ,2) as
"Average product rating"
FROM customer
GROUP by item_purchased
ORDER by avg(review_rating)
DESC
LIMIT 5;
```

	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

- Shipping type comparison: Compare average purchase amounts between Standard and Express shipping
  - Q4. Compare average purchase amounts between standard and express shipping.

```
SELECT shipping_type,
round(avg(purchase_amount),2)
as 'average purchas'
FROM customer
WHERE shipping_type in (
'Express', 'Standard')
GROUP by shipping_type;
```

	shipping_type	average purchas
1	Express	60.48
2	Standard	58.46

- Subscribe Vs Non subscriber: Compared average spend and total revenue across subscription status.
    - Q5. Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers?

```
SELECT subscription_status,
count(customer_id) as
total_customer,
round(avg(purchase_amount),2)
as average_spend,
round(sum(purchase_amount),2
) as total_revenue
FROM customer
GROUP by subscription_status
ORDER by total_revenue,
average_spend DESC;
```

	subscription_status	total_customer	average_spend	total_revenue
1	Yes	1053	59.49	62645.0
2	No	2847	59.87	170436.0

- Discount dependent products: Identify 5 products with the highest percentage of discount purchase
    - Q6.Which 5 products have the highest percentage of purchases with discount applied?

```
SELECT item_purchased, round(100*
sum(CASE WHEN discount_applied = 'Yes'
THEN 1 ELSE 0 END)/count(*),2) as
discount_rate
FROM customer
GROUP by item_purchased
ORDER by discount_rate desc
LIMIT 5;
```

	item_purchased	discount_rate
1	Hat	50.0
2	Sneakers	49.0
3	Coat	49.0
4	Sweater	48.0
5	Pants	47.0

- Customer Segmentation: Classified customer into New, returning, and loyal segments based on purchase history
    - 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 customer
)
 SELECT customer_segment, count(*) as
'Number of Customer'
 FROM customer_type
 GROUP by customer_segment;
```

	customer _segment	Number of Customer
1	Loyal	3116
2	NEW	83
3	returning	701

- Top 3 products by category: Listed the most purchased products within each category.
    - Q8. What are the top 3 most purchase product 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 customer
GROUP by category, item_purchased
)
SELECT item_rank, category,
item_purchased, total_orders
FROM item_counts
WHERE item_rank <= 3;
```

item_rank	category	item_purchased	total_orders
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
1	Clothing	Pants	171
2	Clothing	Blouse	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 and subscriptions: Checked customer with more than 5 purchases are more likely to subscribe.
    - Q9. Are customers who are repeated buyers(more than 5 previous purchases) also likely to subscribe?

```
SELECT subscription_status,
count(customer_id) as repeat_buyers
FROM customer
WHERE previous_purchases > 5
GROUP by subscription_status;
```

	subscription_status	repeat_buyers
1	No	2518
2	Yes	958

- Revenue by age: Calculated total revenue contribution of each age group.
    - Q10. What is the revenue contribution of each age group?

```
SELECT age_group,
sum(purchase_amount) as
total_revenue
From customer
Group by age_group
ORDER by total_revenue DESC
```

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

Deliverable: *EDA Notebook or report* summarizing insights and trends.

Tools: pandas, SQL Analytics.

---

5 Insight Generation & Interpretation

🎯 **Goal:** Translate data findings into **actionable business insights** such as spending patterns, customer segmentation, product preferences, and subscription behaviors

Insights from SQL data analytics:

- Spending patterns:
  - Female contributes nearly 68% of total revenue than male customers.
  - There are total of 839 customers who spends more than average spending amount even with after discount.
  - Slightly more customers opted for express shipping
- customer segmentation:
  - We have 701 returning customer, 83 new customer and 3116 loyal customers.
  - We have fairly equal number of customer across all age group.
- product preferences:
  - The most rated top five products are Gloves (3.86), Sandals (3.84), Boots(3.82), Hats(3.8) and skirts(3.78)
  - The most purchase top five product with discounts applied are Hats, sneakers, coat, suiter and pants
  - The top three product purchase in each categories are
    - accessories category: jewellery, sunglasses, belt, in
    - clothing category: pants, blouse and shirt.
    - Footwear category: sandals, shoes, sneakers
    - Outwear: Jacket and coat
- subscription behaviors:
  - The average spending by subscribers are slightly lower than non-subscriber and total revenue generated from subscribers are much lower than the non subscriber.
  - There is higher number of non subscribers who repeatedly purchase the products than the subscribers. It is less likely that the subscribers can repeat their purchase.

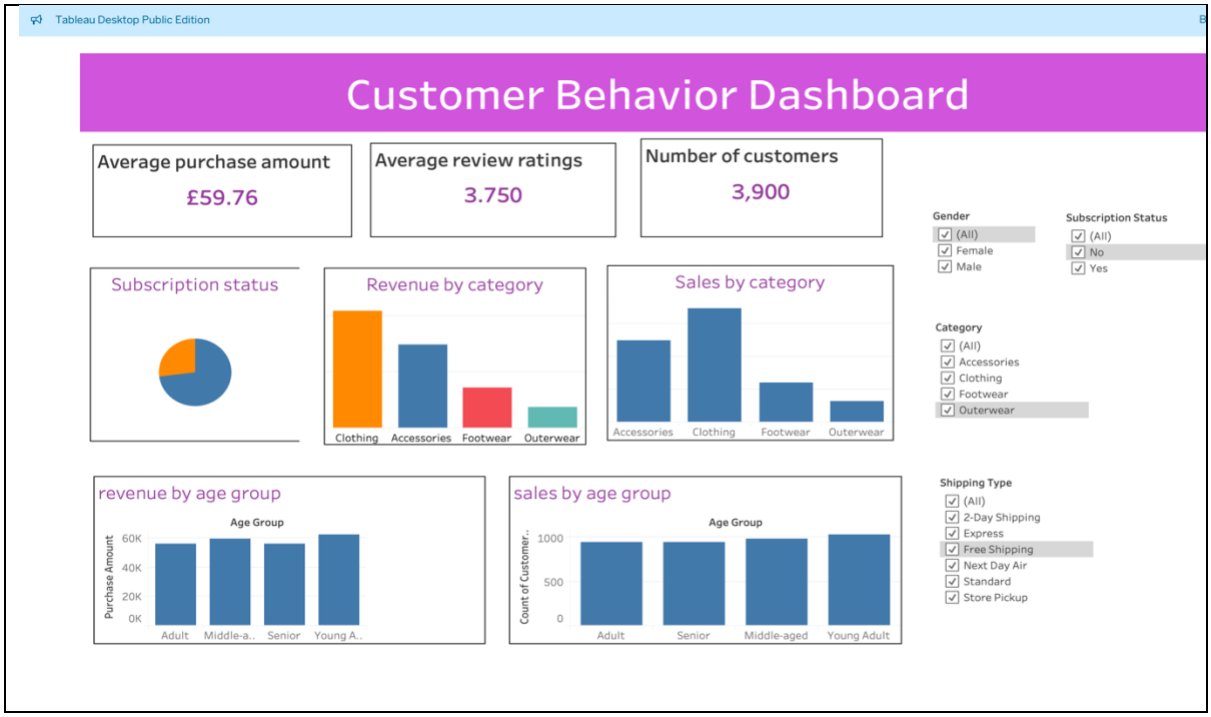
Deliverable: *Insight summary.*

---

6 Data Visualization & Storytelling

🎯 **Goal:** Communicate insights clearly to non-technical stakeholders.

Dashboard in Tableau:



Deliverable: *Dashboard*

---

7 Recommendation

🎯 **Goal:** Turn insights into business decisions.

Business recommendation:

- Boost Subscriptions – Promote exclusive benefits for subscribers..
- Customer Loyalty Programs – Reward repeat buyers to move them into the "Loyal"
- 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.

Deliverable: *Professional report or presentation deck.*