CUSTOMER BEHAVIOR ANALYSIS

A Data Analysis Project Report

1. EXECUTIVE SUMMARY.

This report presents an end-to-end analysis of 3,900 retail purchase transactions, covering customer demographics, shopping behavior, and product preferences. The analysis was conducted using Python for exploratory data analysis (EDA) and data cleaning, followed by structured querying in PostgreSQL to extract actionable business insights.

- **Key findings summary:**

- Male customers account for the majority of total revenue ($157,890 vs. $75,191 for female customers), though female subscribers represent an untapped growth opportunity.
- Young Adults (ages 18–31) are the highest-revenue age segment, contributing $62,143 — making them the most valuable demographic to target.
- Subscribed customers generate a comparable average spend ($59.49) to non-subscribers ($59.87), suggesting that the subscription program is not yet driving higher individual spend.
- Loyal customers (21+ previous purchases) dominate the customer base at 2,339 — but only 958 of repeat buyers subscribe, indicating significant room to grow the subscription program.
- Gloves, Sandals, and Boots are the highest-rated product categories, while Hats and Sneakers see the most discount-driven purchases.

2. DATASET OVERVIEW

The dataset contains 3,900 rows and 18 columns, covering three broad categories of information:

- Customer demographics (Customer ID,Location, Gender, Subscription status, Age)
- Shopping information (Discount applied, previous purchases, frequency of purchases, review rating ,promo code used, shipping type, payment method)
- Purchase information (item purchased, size, color, category, season, purchase amount).

Summary statistics for numeric columns:

```
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

3. DATA CLEANING & PREPARATION

3.1 Missing Values

The dataset was largely complete. The only column with missing values was Review Rating, which had 37 null entries (less than 1% of the dataset). These were handled by imputing the median rating of the corresponding product category — a deliberate choice over using the overall mean, since different product categories attract different levels of customer satisfaction.

```
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
```

3.2 Duplicate Records

No duplicate records were found in the dataset.

```
#check for duplicates
df.duplicated().sum()
```

```
np.int64(0)
```

3.3 Column Standardization

All column names were converted to snake_case (lowercase with underscores) for consistency and compatibility with SQL. The Purchase Amount (USD) column was additionally renamed to purchase_amount for cleaner referencing.

```
# we remove space within column names and replace it with
# underscore and change all characters to lower case
df.columns=df.columns.str.lower()
df.columns=df.columns.str.replace(' ','_')
df=df.rename(columns={'purchase_amount_(usd)':'purchase_amount'})
df.columns
```

```
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')
```

3.4 Redundant Columns

Upon inspection, the discount_applied and promo_code_used columns were found to contain identical values across all 3,900 rows. Since a promo code being used always corresponded to a discount being applied (and vice versa), the promo_code_used column was dropped to eliminate redundancy. The remaining discount_applied column captures the key information: whether or not a discount was granted.

```
df[['discount_applied','promo_code_used']].head(10)
```

	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

```
#checking if their values are all same. if it returns true, we delete one
(df['discount_applied']==df['promo_code_used']).all()
```

```
np.True_
```

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

3.5 Feature Engineering

Two new columns were created to support richer analysis:

- age_group: Customers were segmented into four equal quartile-based groups — Young Adult, Adult, Middle Aged, and Senior using pd.qcut().
- purchase_frequency_days: The text-based frequency_of_purchases column (e.g., 'Weekly', 'Monthly') was mapped to its numeric equivalent in days to enable quantitative comparisons.

```
# feature engineering. we create a column to categorize age into four groups
labels=['Young Adult','Adult','Middle Aged','Senior']
df['age_group']=pd.qcut(df['age'],q=4,labels=labels)
df[['age','age_group']].head(7)
```

	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

```
# create 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[['frequency_of_purchases','purchase_frequency_days']].head()
```

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

4. SQL ANALYSIS & FINDINGS

```
#pip install psycopg2-binary sqlalchemy
```

```
from sqlalchemy import create_engine
username="postgres"
password="I put my password here"
host="localhost"
port="5432"
database="customer_behavior_db"

engine= create_engine(f"postgresql+psycopg2://{username}:{password}@{host}:{port}/{database}")
table_name="customers"
df.to_sql(table_name,engine, if_exists="replace",index=False)

print(f"data successfully loaded into {table_name} table in {database}")
```

```
data successfully loaded into customers table in customer_behavior_db
```

After loading the cleaned dataset into a PostgreSQL database via SQLAlchemy, 12 analytical questions were explored.

Each question, its SQL query, result, and business insight are presented below.

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

```
SELECT gender, SUM(purchase_amount) AS total_purchase_amount
```

```
FROM customers
GROUP BY gender;
```

	gender text	total_purchase_amount numeric
1	Female	75191
2	Male	157890

Insight: Male customers contribute over twice the revenue of female customers. This likely reflects a higher proportion of male customers in the dataset (as confirmed in Q12), rather than a difference in individual spend. This gap represents a strategic opportunity: female customers are underrepresented and under-converted.

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

```
SELECT customer_id, purchase_amount
FROM customers
WHERE purchase_amount > (SELECT ROUND(AVG(purchase_amount),2) FROM
customers)
AND discount_applied='Yes'
GROUP BY customer_id, purchase_amount
LIMIT 10;
```

	customer_id bigint	purchase_amount bigint
1	475	76
2	878	65
3	1517	68
4	1360	68
5	654	68
6	627	79
7	128	89
8	983	62
9	1198	92
10	1365	75

Insight: These high-value discount users spend above average even with a discount applied. They are ideal candidates for loyalty rewards or premium membership targeting, as they demonstrate both price sensitivity and high purchasing power.

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

```
SELECT item_purchased,ROUND(AVG(review_rating)::numeric,2)AS average_review
FROM customers
GROUP BY item_purchased
```

```
ORDER BY average_review DESC
LIMIT 5;
```

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

Insight: Accessories and footwear dominate the top-rated products. These items represent reliable quality signals and could be highlighted in marketing materials or used as entry points for upselling.

Q4. How do average purchase amounts compare across shipping types?

```
SELECT shipping_type,ROUND(AVG(purchase_amount),2) AS average_purchase_amount
FROM customers
WHERE shipping_type IN('Standard','Express','Store Pickup')
GROUP BY shipping_type
ORDER BY average_purchase_amount DESC;
```

	shipping_type text	average_purchase_amount numeric
1	Express	60.48
2	Store Pickup	59.89
3	Standard	58.46

Insight: The differences across shipping types are marginal (within $2), suggesting that shipping preference does not strongly predict spending level. Express shipping customers spend slightly more on average, possibly indicating a preference for convenience among higher-spend customers.

Q5. Do subscribed customers spend more than non-subscribers?

```
SELECT subscription_status, COUNT(*) AS total_customers,
        ROUND(AVG(purchase_amount),2)AS avarage_revenue,
        SUM(purchase_amount)AS total_revenue
FROM customers
GROUP BY subscription_status;
```

	subscription_status text	total_customers bigint	avarage_revenue numeric	total_revenue numeric
1	No	2847	59.87	170436
2	Yes	1053	59.49	62645

Insight: Interestingly, subscribed customers do not spend significantly more per transaction than non-subscribers ($59.49 vs. $59.87). The subscription program appears to drive loyalty and retention rather than increased per-visit spending. Focus should be on converting non-subscribers — who represent 73% of customers — into subscribers.

Q6. Which 5 products have the highest rate of discount-driven purchases?

```
SELECT item_purchased,ROUND(100*SUM(CASE WHEN discount_applied='Yes' THEN 1
ELSE 0 END)/COUNT(*),2)AS percent_of_purchases_on_discount
FROM customers
GROUP BY item_purchased
ORDER BY percent_of_purchases_on_discount DESC
LIMIT 5;
```

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

Insight: Nearly half of Hat, Sneaker, and Coat purchases involve a discount. These products may be struggling to convert at full price, which warrants investigation into pricing strategy or perceived value.

Q7. Which 5 products sell most often at full price (no discount)?

```
SELECT item_purchased,ROUND(100*SUM(CASE WHEN discount_applied='No' THEN 1
ELSE 0 END)/COUNT(*),2)AS percent_of_purchases_without_discount
FROM customers
GROUP BY item_purchased
ORDER BY percent_of_purchases_without_discount DESC
LIMIT 5;
```

	item_purchased text	percent_of_purchases_without_discount numeric
1	Socks	67.00
2	Blouse	66.00
3	Sandals	63.00
4	Skirt	61.00
5	Jacket	60.00

Insight: Socks, Blouses, and Sandals sell predominantly at full price, demonstrating strong inherent demand. These items should be protected from unnecessary discounting to preserve margin.

Q8. How are customers segmented by loyalty level?

```
WITH customer_type AS(SELECT customer_id,previous_purchases,
                        CASE
                            WHEN previous_purchases<=5 THEN 'New'
                            WHEN previous_purchases BETWEEN 6 AND 20 THEN
'Returning'
                            ELSE 'Loyal'
                            END AS customer_segment
                            FROM customers)
SELECT customer_segment,COUNT(*) AS number_of_customers
FROM customer_type
GROUP BY customer_segment;
```

	customer_segment text	number_of_customers bigint
1	Loyal	2339
2	New	424
3	Returning	1137

Insight: The majority of customers (60%) are classified as Loyal, indicating strong retention. However, this also means that attracting and nurturing New customers (only 11% of the total) is an area of relative weakness that the business should address.

Q9. What are the top 3 most purchased products in each category?

```
WITH product_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 product_rank
        FROM customers
        GROUP BY category,item_purchased)

SELECT product_rank, category,item_purchased,total_orders
FROM product_counts
WHERE product_rank<=3;
```

	product_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessories	Jewelry	171
2	2	Accessories	Sunglasses	161
3	3	Accessories	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

Insight: Purchase volume is remarkably consistent across top products within each category, suggesting broad and even demand rather than a few hero products dominating. This is a healthy indicator of product portfolio balance.

Q10. Are repeat buyers (5+ purchases) more likely to subscribe?

```
SELECT subscription_status,COUNT(*)as repeat_customers
FROM customers
WHERE previous_purchases>5
GROUP BY subscription_status
```

	subscription_status text	repeat_customers bigint
1	No	2518
2	Yes	958

Insight: Even among highly engaged repeat buyers, only about 28% hold a subscription. This is a major conversion gap — loyal customers who already shop frequently are the most natural candidates for a subscription program but are not being converted at scale.

Q11. What is the revenue contribution of each age group?

```
SELECT age_group,SUM(purchase_amount) AS total_revenue_contri
FROM customers
GROUP BY age_group
ORDER BY total_revenue_contri DESC;
```

	age_group text	total_revenue_contri numeric
1	Young Adult	62143
2	Middle Aged	59197
3	Adult	55978
4	Senior	55763

Insight: Revenue contribution is distributed relatively evenly across age groups, but Young Adults lead. This is noteworthy given that younger customers may have lower average incomes. They may be shopping more frequently or in higher volumes. Tailoring product offerings and marketing to this segment could yield good returns.

Q12. How do subscription rates break down by gender?

```
SELECT gender , subscription_status,COUNT(*) AS total_by_gender
FROM customers
GROUP BY gender, subscription_status;
```

	gender text	subscription_status text	total_by_gender bigint
1	Female	No	1248
2	Male	No	1599
3	Male	Yes	1053

Insight: All 1,053 subscribers in this dataset are male. No single female customer holds a subscription. Female customers shop and spend at comparable levels to male customers, yet the subscription program has completely failed to convert them. Whether due to targeting, messaging, or the nature of the benefits offered, the subscription program is a male-only product in practice.

5. POWER BI DASHBOARD

The cleaned dataset was visualized in Power BI to provide an interactive summary of the analysis for non-technical stakeholders. The dashboard allows users to slice and filter all visuals dynamically by Shipping Type, Product Category, Subscription Status, and Gender — enabling quick exploration of how these variables affect key metrics.

5.1 KPI Cards

Five headline metrics are displayed at the top of the dashboard to give an immediate snapshot of overall business performance:

KPI	Value	What It Tells Us
Average Purchase Amount	$59.76	Customers spend consistently around the $60 mark — a useful baseline for discount and pricing strategy
Average Review Rating	3.75	Satisfaction is moderately positive across all products, with room for improvement
Number of Customers	3,900	Total transactions analyzed in this dataset
Total Unique Items	25	A manageable product catalogue with broad category coverage
Total Revenue	$233,081	Combined purchase amount across all 3,900 purchases

5.2 Visuals & Insights

The dashboard contains six core visuals, each designed to answer a specific business question at a glance:

Total Items & Revenue per Category

Two bar charts show purchase volume and revenue by product category. Clothing leads with 1,737 purchases and $104,264 in revenue, followed by Accessories (1,240 purchases, $74,200). Footwear and Outerwear trail significantly. Notably, Accessories generates strong revenue relative to its purchase volume, suggesting higher average transaction values in that category.

Total Revenue by Age Group

A horizontal bar chart compares revenue across the four age segments. Young Adults lead at $62,143, with all four groups falling within a $6,000 range of each other — confirming broad appeal across age ranges, with Young Adults as the highest-value demographic to target.

Items Purchased per Age Group

Purchase volume is nearly even across all age groups: Young Adults (1,028), Middle Aged (986), Senior (944), and Adult (942). This near-equal split confirms that the product offering resonates across the full customer age spectrum — a sign of portfolio breadth rather than niche appeal.

Customers by Subscription Status

A donut chart shows the subscriber split: 2,847 customers (73%) are non-subscribers versus 1,053 (27%) who subscribe. The visual immediately communicates the scale of the conversion opportunity — nearly three quarters of the customer base is unconverted. When filtered by Gender, the chart reveals that all 1,053 subscribers are male, with zero female subscribers.

5.3 Interactive Filters

Four slicers on the right panel allow any visual on the dashboard to be filtered in real time, enabling business users to explore the data without needing SQL or Python:

Filter	Options	Example Use Case
Shipping Type	2-Day, Express, Free Shipping, Next Day Air, Standard, Store Pickup	Compare spend patterns between Express and Standard shipping customers
Category	Accessories, Clothing, Footwear, Outerwear	Isolate revenue and purchase trends for a single product category
Subscription Status	No / Yes	See how revenue and volume differ between subscribers and non-subscribers
Gender	Female / Male	Surface the zero female subscriber finding instantly

Figure 1 customer behavior analysis - power bi dashboard

6. KEY FINDINGS

6.1 Revenue & Demographics

- Male customers generate 68% of total revenue ($157,890 out of $233,081 combined), though this reflects a larger male customer base rather than higher per-transaction spend.
- Young Adults are the top revenue-contributing age group ($62,143), followed closely by Middle Aged customers ($59,197). Revenue is fairly balanced across age groups.
- The average purchase amount is $59.76, with purchases ranging from $20 to $100.

6.2 Subscription & Loyalty

- Only 27% of customers (1,053 out of 3,900) are subscribers. Subscribed and non-subscribed customers spend nearly the same amount per transaction (~$59.50 vs. ~$59.87).
- 72% of the customer base are Loyal (2,339 customers with 21+ previous purchases), yet only 958 repeat buyers subscribe so there's a significant missed conversion opportunity.
- All 1,053 subscribers are male so there are zero female subscribers in the dataset. Female customers spend comparably but have not been converted by the subscription program at all.

6.3 Product Performance

- Gloves, Sandals, and Boots are the highest-rated products, while Socks, Blouses, and Sandals sell most consistently at full price thus indicating strong perceived value.
- Hats, Sneakers, and Coats see the highest rates of discount usage (approximately 49–50%), suggesting potential pricing or value perception issues.
- Purchase volumes are well-distributed across the top products within each category, with no single item dominating implying a balanced product portfolio.

6.4 Shipping & Discounts

- Shipping type has minimal impact on average spend (less than $2 difference between Express and Standard shipping).
- Discounts do not appear to suppress spend as high-value customers who use discounts still purchase above the average transaction amount ($59.76).

7. RECOMMENDATIONS

7.1 Grow the Subscription Program

- Launch a targeted subscription conversion campaign specifically for Loyal and Returning customers (3,476 combined) who have not yet subscribed. These customers already demonstrate commitment and are the most natural fit.

- The subscription program has zero female subscribers despite female customers being active shoppers with comparable spend meaning this is a conversion failure, not a data anomaly. A dedicated female-targeted subscription campaign with different benefit framing, product curation, or communication channels should be treated as a top priority.
- Rethink the subscription value proposition. Since subscribers do not currently spend more per transaction, the program should be reframed around exclusive perks, early product access, or members-only discounts to make subscription status feel premium.

7.2 Optimize Pricing & Discounting

- Protect high-demand, full-price products (Socks, Blouses, Sandals) from promotional discounting to preserve margins as these items sell well without incentivization.
- Investigate why Hats, Sneakers, and Coats rely heavily on discounts (~50% of purchases). Consider repositioning, bundling, or adjusting price points rather than defaulting to discounts.
- Target high-value discount users (those spending above average even with a discount) for premium loyalty programs or early-access events, as they show both price sensitivity and high spending capacity.

7.3 Demographic Targeting

- Double down on Young Adult marketing, as this group leads in total revenue contribution. Digital-first campaigns (social media, influencer partnerships) are likely to resonate most with this age segment.
- Develop strategies to attract and convert new customers, who currently represent only 11% of the base. First-purchase incentives or onboarding programs could help grow this group.
- Explore why female customers — while spending comparably when they do shop — are not subscribing. A gender-segmented analysis of product preference and purchase frequency could reveal opportunities for more targeted product curation.

8. TOOLS & TECHNOLOGIES

- Python(pandas): Data loading, exploration, cleaning, and feature engineering
- Jupyter Notebooks: Interactive analysis environment
- SQLAlchemy: Database connection and data ingestion
- PostgreSQL: Structured querying and business analysis
- SQL (CTEs, Window Functions, Subqueries): Advanced analytical queries
- POWER BI: Dashboard

— End of Report — **BY: AINAMANI DICKSON**