Customer Shopping Behavior Analysis

Created By:
Nikhil Mohan Tatfale

February 7, 2026

1. Project Overview

This project analyzes customer shopping behavior using e-commerce transactional data from 3,900 customer purchases across multiple product categories. The primary goal is to discover meaningful patterns in customer spending habits, identify different customer segments, understand product preferences, and analyze subscription trends to support data-driven business strategies.

The analysis workflow includes data collection from Kaggle, exploratory data anal- ysis (EDA) using Python, data cleaning and transformation, database integration with MySQL, business analytics through SQL queries, and visualization using Power BI dash- b oard.

2. Dataset Summary

Rows: 3,900

Columns: 18

Key Features:

- Customer Demographics: Age, Gender, Location, Subscription Status
- Purchase Information: Item Purchased, Category, Purchase Amount (USD), Season, Size, Color
- Shopping Patterns: Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type, Payment Method

Missing Data: 37 missing values identified in the Review Rating column

	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
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually

Figure 1: Sample view of the dataset

3. Exploratory Data Analysis using Python

The data preparation and cleaning process was conducted in Google Colab using Python and the Pandas library. Below are the key steps performed:

3.1 Data Loading and Initial Exploration

We started by importing the dataset using pandas and performing initial checks to understand the data structure.

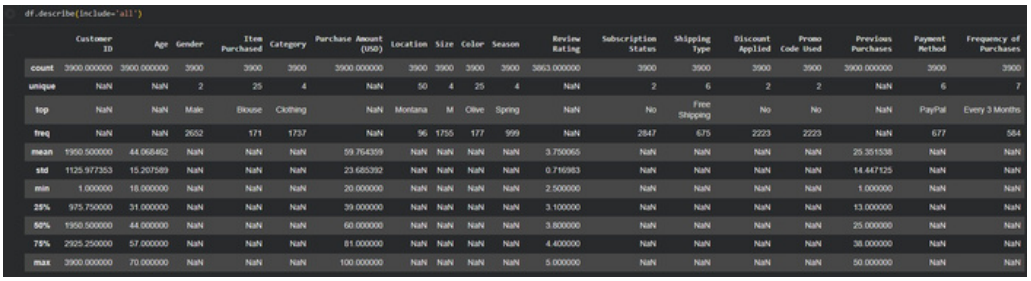
```python
import pandas as pd

# Load the dataset
df = pd.read_csv('shopping_behavior.csv')

# Check data structure
df.info()

# Generate summary statistics
df.describe()
```

Listing 1: Loading and exploring the dataset



df.describe(include='all')

	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

Figure 2: Statistical summary showing distribution of numerical features

Key observations from the initial exploration:

Customer age ranges from 18 to 70 years with a mean of approximately 44 years

Purchase amounts vary between $20 and $100, averaging around $59.76

Review ratings span from 2.5 to 5.0 with an average rating of 3.75

Customers have between 1 and 50 previous purchases

3.2 Missing Data Handling

We identified and addressed missing values in the dataset.

```
df.isnull().sum()
```

	0
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

Figure 3: Missing value detection - 37 null values in Review Rating

Solution: Missing values in the Review Rating column were filled using the median rating for each product category. This approach preserves category-specific rating patterns rather than using a global median.

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

# Verify no missing values remain
df.isnull().sum()
```

Listing 2: Imputing missing values with category-wise median

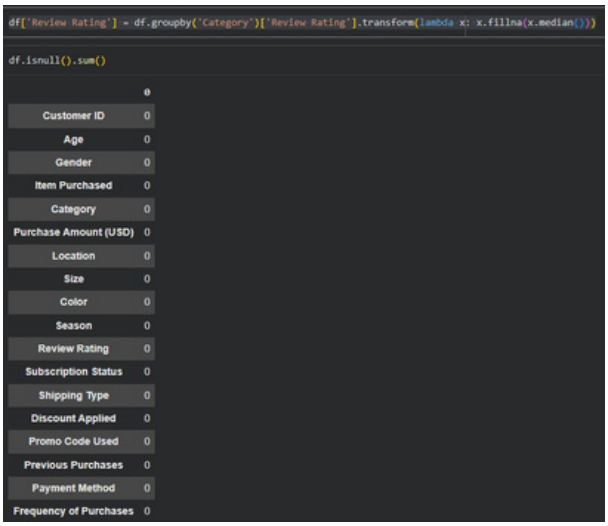


Figure 4: Verification - all missing values successfully handled

3.3 Column Standardization

All column names were converted to snake_case format for consistency and easier database integration.

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

Listing 3: Standardizing column names

```
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')
```

Figure 5: Standardized column names in snake case

3.4 Feature Engineering

Two new features were created to enhance analytical capabilities:

1. Age Group Column: Customers were segmented into four age groups using quartiles.

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

Listing 4: Creating age group categories

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

df[['age',"age_group"]].head()
```

	age	age_group
0	55	Middle-aged
1	19	Young Adult
2	50	Middle-aged
3	21	Young Adult
4	45	Middle-aged

Figure 6: Age group segmentation

2. Purchase Frequency Days: Converted categorical frequency values to numerical days for quantitative analysis.

```
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)
```

Listing 5: Mapping purchase frequency to days

```
# create the column of 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[['purchase_frequency_days',"frequency_of_purchases"]].head(10)
```

	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

Figure 7: Purchase frequency converted to numerical days

3.5 Data Consistency Check

We identified redundant columns by checking if discount applied and promo code used contained identical information.

```
(df['discount_applied'] == df['promo_code_used']).all()
# Returns: True

# Drop redundant column
df = df.drop('promo_code_used', axis=1)
```

Listing 6: Checking for redundant columns

```
# check wheather discount_applied and promo_code_use is same data or not
(df['discount_applied'] == df['promo_code_used']).all()
```

```
np.True_
```

```
# so let drop the promo_code_used
df = df.drop('promo_code_used', axis=1)
```

Figure 8: Removing redundant promo code used column

3.6 Database Integration

After cleaning and preparing the data, the final dataset was exported and loaded into a MySQL database for structured querying and business analysis.

4. Data Analysis using SQL (Business Questions)

The cleaned dataset was uploaded to MySQL to perform structured analysis and answer critical business questions. Below are the key queries and findings:

1. Revenue by Gender

Business Question: What is the total revenue generated by male versus female customers?

```sql
SELECT gender, SUM(purchase_amount) AS revenue
FROM updated_customer_shopping_behavior
GROUP BY gender;
```

Listing 7: Analyzing revenue contribution by gender

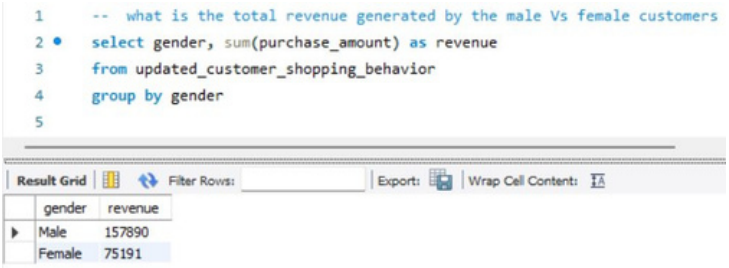


Figure 9: Total revenue comparison between genders

Key Finding: Male customers contribute $157,690 in revenue compared to $75,191 from female customers. This reveals a significant gender gap presenting an opportunity for targeted female customer acquisition.

2. High-Spending Discount Users

Business Question: Which customers used discounts but still spent above the average purchase amount?

```sql
SELECT customer_id, discount_applied, purchase_amount
FROM updated_customer_shopping_behavior
WHERE purchase_amount >= (SELECT AVG(purchase_amount)
                          FROM updated_customer_shopping_behavior
    )
      AND discount_applied = 'yes'
ORDER BY purchase_amount ASC;
```

Listing 8: Identifying discount users with high spending

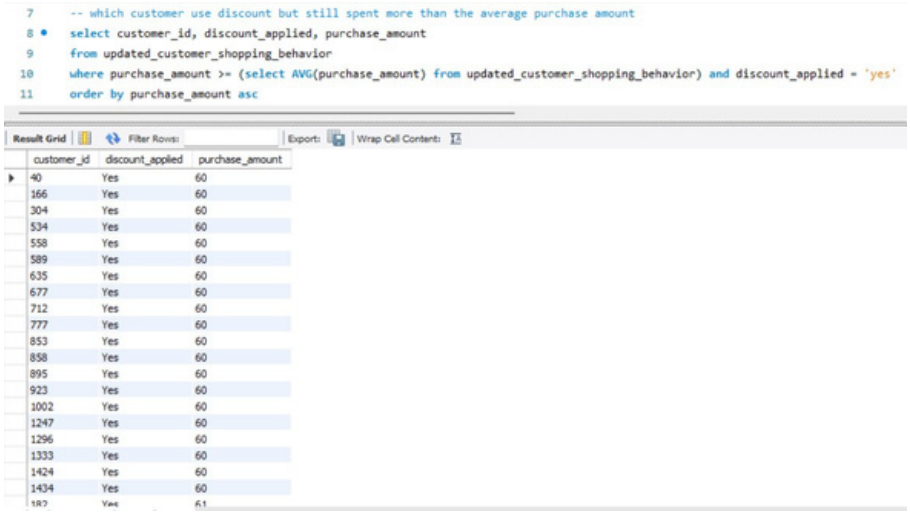


Figure 10: Customers using discounts while maintaining above-average spending

Key Finding: Multiple customers spend $60 or more even with discounts applied, indicating that strategic discounting doesn't necessarily erode revenue.

3. Top 5 Products by Rating

Business Question: Which products have the highest average customer review ratings?

```
SELECTitem_purchased, ROUND(AVG(review_rating)2)  AS '
   review_rating_average'
FRO  updated_customer_shopping_behavior
M     B  item_purchased
GRO   Y  AVG(review_rating)  DESC
UP    B
ORD   Y
ER    5;
LIMI
T
```

Listing 9: Finding top-rated products

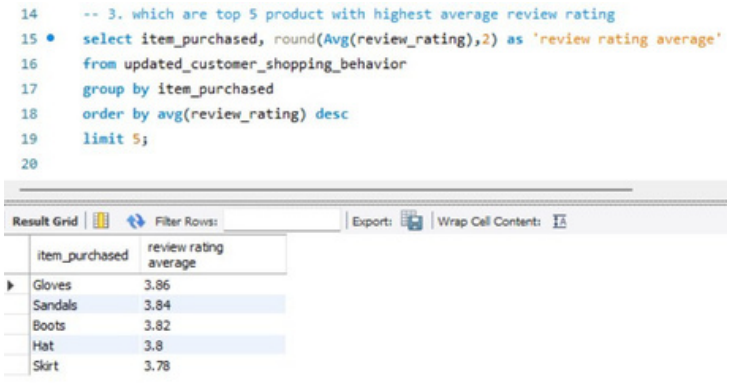


Figure 11: Top 5 products by average review rating

Key Finding: Gloves (3.86), Sandals (3.84), and Boots (3.82) receive the highest customer satisfaction ratings and should be prominently featured in marketing efforts.

4. Shipping Type Comparison

Business Question: Is there a difference in average purchase amount between standard and express shipping customers?

```
SELECTshipping_type, ROUND(AVG(purchase_amount), 2) AS '
    average_purchase_amount'
FRO  updated_customer_shopping_behavior
M      shipping_type IN ('Express',   'Standard')
WHE  BY shipping_type;
RE
GRO
UP
```

Listing 10: Comparing purchase amounts by shipping type

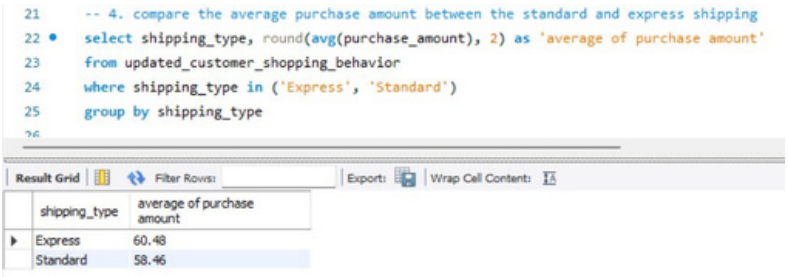


Figure 12: Average purchase amount by shipping method

Key Finding: Express shipping customers spend slightly more on average ($60.48 vs $58.46), suggesting a correlation between urgency and spending.

5. Subscribers vs. Non-Subscribers

Business Question: Do subscribed customers spend more? Compare total revenue and average spend.

```
SELECTsubscription_status,
        SUM(purchase_amount) AS revenue,
        ROUND(AVG(purchase_amou 2)  AS 'average_spend',
        nt),
        COUNT(customer_id) AS 'total_customer'
FROMupdated_customer_shopping_behavior
GROUPBY subscription_status;
```

Listing 11: Analyzing subscription impact on spending

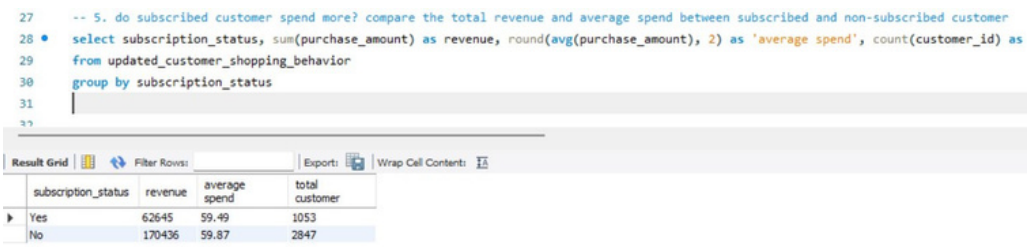


Figure 13: Comparison of subscribed vs non-subscribed customers

Key Finding: Only 27% of customers are subscribed. While average spending per transaction is similar, the low subscription rate represents a significant growth opportunity.

6. Discount-Dependent Products

Business Question: Which products have the highest percentage of purchases with discounts applied?

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

Listing 12: Calculating discount rate by product

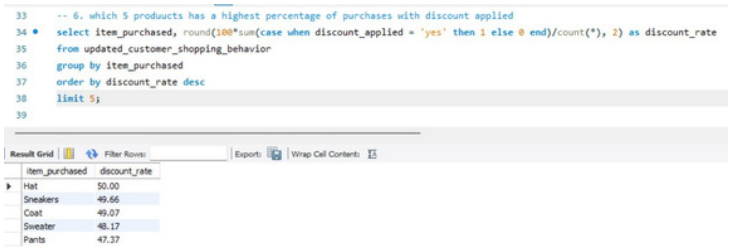


Figure 14: Products with highest discount dependency

Key Finding: Hat (50%), Sneakers (49.66%), and Coat (49.07%) show very high discount rates, suggesting potential over-reliance on promotional pricing.

7. Customer Segmentation

Business Question: How are customers distributed across New, Returning, and Loyal segments?

```sql
WITH customer_type AS
(
    SELECT customer_id, previous_purchases, purchase_amount,
        CASE
            WHEN previous_purchases = 1 THEN 'New'
            WHEN previous_purchases BETWEEN 2 AND 10 THEN 'Returning'
            ELSE 'Loyal'
        END AS customer_segmented
    FROM updated_customer_shopping_behavior
)
SELECT customer_segmented,
       COUNT(*) AS 'number_of_customers',
       ROUND(SUM(purchase_amount),2) AS 'revenue'
FROM customer_type
GROUP BY customer_segmented;
```

Listing 13: Segmenting customers by purchase history

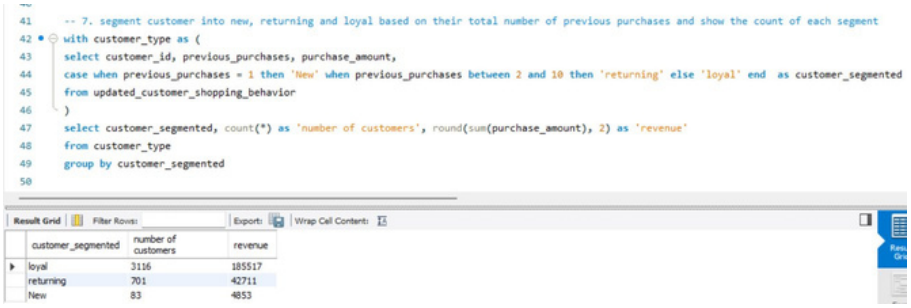


Figure 15: Customer distribution across segments

Key Finding: Loyal customers (80%) drive the majority of revenue, but new customer acquisition is weak at only 2% of the customer base.

8. Top 3 Products per Category

Business Question: What are the most purchased products within each category?

```sql
WITH item_count 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 updated_customer_shopping_behavior
    GROUP BY category, item_purchased
)
SELECT item_rank, category, item_purchased, total_orders
FROM item_count
WHERE item_rank <= 3;
```

Listing 14: Identifying best-sellers by category

```sql
-- 8. what are the top 3 most purchase products within each the category
with item_count 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 updated_customer_shopping_behavior
group by category, item_purchased
)
select item_rank, category, item_purchased, total_orders
from item_count
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

Figure 16: Top 3 products in each category

Key Finding: Jewelry leads Accessories, Blouse leads Clothing, Sandals lead Footwear, and Jacket leads Outerwear. These items should be prioritized in inventory management.

9. Repeat Buyers & Subscriptions

Business Question: Are customers with more than 5 previous purchases more likely to subscrib e?

```sql
SELECTsubscription_status,  COUNT(customer_id)AS repeat_buyers
FRO updated_customer_shopping_behavior
M    previous_purchases > 5
WHE BY subscription_status;
RE
GRO
UP
```

Listing 15: Analyzing subscription behavior among repeat buyers

```sql
-- 9. are customer who are repeat buyer( more than 5 previous orders) are also likely to subscribed
select subscription_status, count(customer_id) as repeat_buyers
from updated_customer_shopping_behavior
where previous_purchases > 5
group by subscription_status

-- 10. what is the revenue contribution of each age group
```

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

subscription_status	repeat_buyers
Yes	958
No	2518

Figure 17: Subscription status among repeat buyers

Key Finding: Even among repeat buyers, 72.5% are not subscribed, indicating subscription benefits may not be compelling enough.

10. Revenue by Age Group

Business Question: What is the revenue contribution of each age group?

```sql
SELECT age_group, ROUND(SUM(purchase_amount), 2) AS revenue
FROM updated_customer_shopping_behavior
GROUP BY age_group
ORDER BY revenue DESC;
```

Listing 16: Calculating revenue by age group

```
71    -- 10. what is the revenue contribution of each age group
72    select age_group, round(sum(purchase_amount), 2) as revenue
73    from updated_customer_shopping_behavior
74    group by age_group
75    order by revenue desc
```

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

age_group	revenue
Young Adult	62143
Middle-aged	59197
Adult	55978
Senior	55763

Figure 18: Revenue distribution across age groups

Key Finding: Revenue is relatively balanced across all age groups, with Young Adults contributing slightly more ($62,143), suggesting broad market appeal.

5. Dashboard in Power BI

An interactive Power BI dashboard was created to visualize key metrics and enable dynamic exploration of the data.

Figure 19: Power BI Dashboard - Customer Shopping Behavior Overview

5.1 Dashboard Highlights

Key Performance Indicators:

Total Customers: 3,900

Average Purchase Amount: $59.8

Average Review Rating: 3.75/5.0

Subscription Rate: 27%

Visual Components:

Revenue by Category - Clothing dominates with approximately $100K

Revenue by Age Group - Senior customers generate the highest revenue ($91.4K)

Sales by Category - Clothing accounts for 44.54% of all sales

Sales by Age Group - Young Adults show the highest purchase frequency

Subscription Status Distribution - Clear visualization of the subscription gap

Interactive Filters: The dashboard includes slicers for Subscription Status, Gender, Category, and Shipping Type, enabling users to drill down into specific customer segments and analyze behavior patterns dynamically.

6. Business Recommendations

Based on the comprehensive data analysis, the following strategic recommendations are prop osed:

1. Expand Female Customer Base

Male customers generate 68% of total revenue. Implement targeted marketing campaigns for female demographics, expand product offerings that appeal to women, and create female-focused promotional events to bridge this revenue gap.

2. Enhance Subscription Value Proposition

Only 27% of customers subscribe. Implement exclusive subscriber-only discounts, free express shipping for members, loyalty points system, and clear communication of subscription benefits to increase penetration.

3. Optimize Discount Strategy

Products like Hats and Sneakers show 50% discount dependency. Test higher base prices with selective discounting, implement time-limited flash sales, create product bundles to maintain value, and reserve heavy discounts for customer acquisition.

4. Strengthen Customer Acquisition

Only 2% of customers are new. Accelerate growth through referral programs, first- purchase discount codes, enhanced digital marketing presence (SEO, social media, paid ads), and strategic brand partnerships.

5. Leverage High-Rated Products

Products with ratings above 3.80 (Gloves, Sandals, Boots) should be featured prominently on the homepage, highlighted in email campaigns, showcased in customer testimonials, and prioritized in inventory decisions.

6. Implement Age-Targeted Marketing

Different age groups show distinct behaviors. Target Young Adults with social media and influencer partnerships, Middle-aged customers with email campaigns and family bundles, and Seniors with easy-to-use interfaces and traditional marketing channels.