Customer Shopping Behavior Analysis

1. Project Overview

This project analyzes customer shopping behavior using transactional data from 3,900 purchases across diverse product categories. The objective is to get actionable insights from spending habits, customer segments, product preferences, and subscription trends to make data-driven business strategies and enhance decision-making efficiency.

2. Dataset summary

2.1 Dataset features

- Customer demographics (Age, Gender, Location)
- Product details (Item Purchased, Category, Color, Season, Size)
- Transactional attributes (Purchase Amount, Shipping Type, Payment Method, Discount Applied, Promo Code Used, Previous Purchases, Subscription Status, Review Rating, Frequency of Purchases).

```
 #   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)
```

2.2 Key descriptive statistics

- Age range: 18 to 70 years, mean age is around 44.
- Purchase amounts ranged from 20 to 100 USD, with a mean of about 60 USD.
- Review ratings had a mean near 3.75 out of 5, post-imputation.
- The most frequent item category was "Clothing", with "Male" being the most common

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000

3. Exploratory Data analysis

The analysis focused on cleaning missing values, feature engineering, and preparing it for further modeling or business intelligence tasks.

3.1. Data cleaning

The dataset was loaded using Python's pandas library.

```
data = pd.read_csv('customer_shopping_behavior.csv')
```

The first data quality check revealed 37 missing values in the "Review Rating" column.

```
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
```

These null values were filled with the median review rating within each product category, which preserved distributional integrity and prevented the bias.

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

Unnecessary or redundant columns (like “Promo Code Used” since it was always 'Yes') were dropped to streamline analysis.

```
(df['discount_applied'] == df['promo_code_used']).all()
```

```
np.True_
```

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

Column names were converted to a consistent format (snake case) for better accessing and readability.

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

3.2. Feature Engineering

Age groups were created by segmenting customers into quartiles:

- Young Adult
- Adult
- Middle Aged
- Senior

```
labels =  ["Young Adult", "Adult", "Middle Aged", "Senior"]

df['age_group'] = pd.qcut(df['age'], q=4, labels=labels)
```

	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

A numeric mapping for purchase frequency was added (e.g., "Fortnightly" to 14, "Weekly" to 7), enabling quantitative analysis of repeat customer value and retention.

```
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)
```

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

3.3. Publishing to MySQL database

The final step in your workflow involves publishing the cleaned and engineered dataset to a MySQL database so we can gather better business insights with SQL queries.

```
import pandas as pd
from sqlalchemy import create_engine

# Create engine
engine = create_engine('mysql+pymysql://root:admin@localhost/customer_behavior')

# Send DataFrame to MySQL
df.to_sql('customers', con=engine, if_exists='replace', index=False, method='multi')

print("\n...DataFrame successfully sent to 'customers' table in your MySQL database...")
```

4. Business analysis using MySQL

This report presents a comprehensive analysis of customer purchase behavior using transactional data from the store's customer dataset. The objective is to analyze meaningful insights about revenue patterns, product performance, customer segmentation, discount effectiveness, and subscription engagement. By examining multiple dimensions such as gender, age, product category, and customer loyalty, the analysis provides actionable findings to guide marketing, pricing, and retention strategies.

4.1. Revenue by Gender

Male customers generated a total revenue of 157,890, while female customers generated 75,191. This shows that male buyers contribute more than double the revenue compared to female buyers.

gender	revenue
Male	157890
Female	75191

4.2. Discount Users Spending Above Average

There are 839 customers who used a discount and still spent more than the average purchase amount. This indicates that discounted items do not necessarily lead to low-value transactions, suggesting strong product appeal.

discount_applied_more_than_avg_purchase
839

4.3. Top Products by Average Review Rating

The top-rated products based on customer reviews are gloves with an average rating of 3.86, followed by sandals with 3.84, and boots with 3.82, hat with 3.8 and skirt with 3.78. These items demonstrate customer satisfaction, particularly in the footwear and accessories.

item_purchased	average_review
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Skirt	3.78

4.4. Purchase Amount Comparison by Shipping Type

The average purchase amount for express shipping is 60.48, slightly higher than standard shipping at 58.46. Customers opting for express delivery tend to spend marginally more overall.

	shipping_type	avg_puchase_amount
▶	Express	60.48
	Standard	58.46

4.5. Subscription Impact on Spending

Subscribed customers have an average spend of 59.49, contributing a total revenue of 62,645. Unsubscribed customers spend slightly more on average at 59.87, contributing total revenue of 170,436. This shows that unsubscribed customers not only spend more but also generate higher total sales, implying the subscription program may not be influencing spending behavior effectively.

	subscription_status	avg_spend	total_revenue
▶	Yes	59.49	62645
	No	59.87	170436

4.6. Products with the Highest Discount Application

Hats have the highest percentage of discounted purchases at 50 percent, followed closely by sneakers at 49.66 percent, coats at 49.07 percent, sweaters at 48.17 percent, and pants at 47.37 percent. This suggests discounts are most frequently applied to apparel and outerwear items.

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

4.7. Customer Segmentation by Purchase History

The customer segmentation analysis shows there are 83 new customers with only one prior purchase, 701 returning customers with between two and ten purchases, and 3,116 loyal customers who made more than ten purchases. The large number of loyal customers indicates strong retention and positive customer experience.

	customer_segment	Number of Customers
▶	Loyal	3116
	Returning	701
	New	83

4.8. Top 3 Products within Each Category

In accessories, the most popular items are jewelry, sunglasses, and belts. For clothing, the leading products are belts, blouses, and pants. In footwear, sandals, shoes, and sneakers dominate, while in outerwear, jackets and coats rank highest. These findings highlight specific products that drive sales within each category.

	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

4.9. Subscription Status Among Repeat Buyers

Among customers with more than five previous purchases, 958 are subscribed, but most repeat buyers remain unsubscribed. This indicates that customers who frequently buy are not significantly motivated to subscribe, suggesting the subscription offer may need enhancements.

	subscription_status	repeat_buyers
▶	Yes	958
	No	2518

4.10. Revenue Contribution by Age Group

Young adults are the largest contributors to total revenue, accounting for 62,143. This age group forms the core of the customer base and represents the primary driver of sales performance.

	age_group	total_revenue
▶	Young Adult	62143
	Middle Aged	59197
	Adult	55978
	Senior	55763

5. Power Bi Dashboard for customer analysis

This dashboard delivers a concise view of retail customer activity, segmenting data by subscription status, purchase category, and age group to spotlight participation and spending trends.

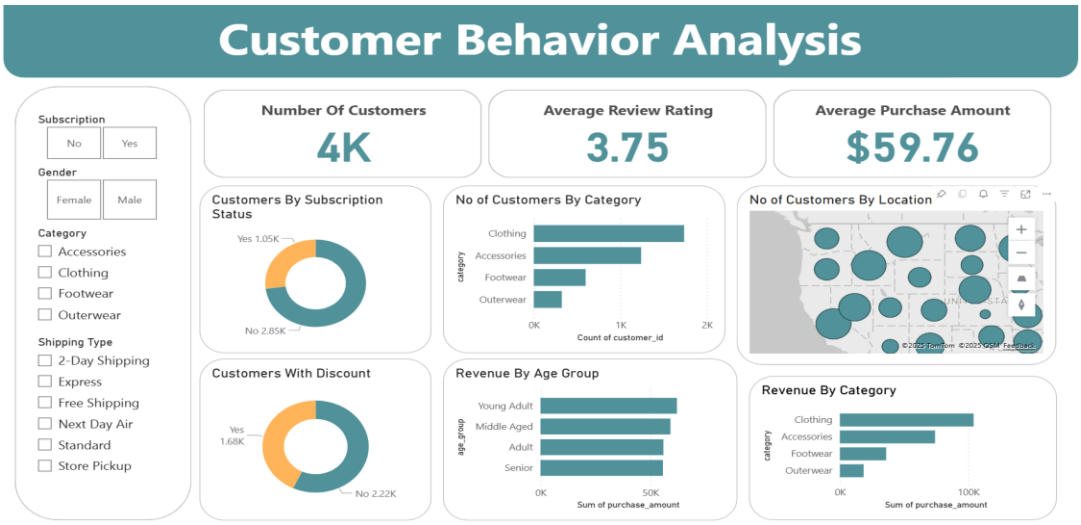
- 4,000 customers contribute to an average purchase value of $59.76 and a review rating of 3.75.

- Most customers are not subscribed or using discounts, suggesting opportunities for loyalty or promotional strategies.

- Clothing is the leading category by both customer engagement and revenue, followed by accessories and footwear.

- Young adults are the top spenders among age groups, making them an influential segment for marketing focus.

- The customer base is widely distributed but shows concentration in certain regions, helping drive targeted outreach efforts.



6. Conclusion

In conclusion, these insights offer a strategic framework for driving sustainable growth in the retail sector. By leveraging targeted marketing strategies, enhancing customer retention initiatives, and implementing effective product management practices, businesses can establish a competitive advantage and achieve long-term success in an retail market.

7. Business Recommendations

- **Target Young Adults:** Focus marketing on young adults using social media and influencer channels.
- **Revise Subscription Strategy:** Add new perks to boost subscription sign-ups and spending.
- **Expand Discount Offers:** Offer discounts and bundle deals on popular apparel items.
- **Enhance Top Products:** Prioritize inventory investment in highly rated items.
- **Nurture Loyal Customers:** Reward loyal customers with personalized programs.
- **Encourage Subscriptions:** Target repeat buyers with tailored subscription offers.
- **Optimize Shipping:** Incentivize express shipping to increase high-value purchases.
- **Regional Marketing:** Use localized campaigns in high-density customer areas.