Customer Shopping Behavior Analysis Report

1. Project overview

The objective of this project is to analyze consumer shopping data to understand shifting purchasing patterns across various demographics, product categories, and sales channels. By uncovering the key drivers behind consumer decisions—such as discounts, reviews, and previous purchase —the company aims to improve sales, boost customer satisfaction, and foster long-term loyalty. The final goal is to provide actionable recommendations to optimize marketing and product strategies.

2. Dataset Summary

Following the initial ingestion and validation of the raw data, we have summarized the structural components of the dataset. This profile serves as the baseline for all subsequent analytical modeling and dashboard reporting.

Data Architecture & Scope

- **Total Volume:** The dataset comprises **3,900 unique transaction records**.
- **Feature Set:** There are **18 specialized variables** capturing a 360-degree view of the customer journey, including demographic, transactional, and behavioral dimensions.

Key Variables Included:

- **Demographics:** Age, Gender, and Location.
- **Transactional Data:** Item Purchased, Category, Purchase Amount (USD), and Payment Method.
- **Behavioral Indicators:** Review Rating, Frequency of Purchases, and Previous Purchases.
- **Logistics & Incentives:** Shipping Type, Season, Subscription Status, Discount Applied, and Promo Code Used.

Data Quality: The dataset is highly complete, with only 37 missing values identified in the "Review Rating" column.

3. Exploratory Data Analysis Using Python (EDA)

We begin with data preparation and cleaning with python using pandas library.

Data loading: Imported the dataset using python (Pandas)

Initial Exploration: Used **data.info()** for data structure and **data.describe()** for summary ststistics.

```
data.describe()
```
✓ 0.1s

	# Customer ID	# Age	# Purchase Amount (USD)	# Review Rating
count	3900.0	3900.0	3900.0	3863.0
mean	1950.5	44.06846153846154	59.76435897435898	3.750064716541548
std	1125.9773532358456	15.207589127162382	23.685392250875307	0.7169829842073633
min	1.0	18.0	20.0	2.5
25%	975.75	31.0	39.0	3.1
50%	1950.5	44.0	60.0	3.8
75%	2925.25	57.0	81.0	4.4
max	3900.0	70.0	100.0	5.0

Check for Missing Values: Used **data.isnull().sum()** to find the missing values and replace it with grouped median values.

```
# Fill the na values using grouped median value
data['Review Rating'] = data.groupby('Category')['Review Rating'] \
    .transform(lambda x: x.fillna(x.median()))
```
✓ 0.2s

Column Standardization: Rename column to snake case to better readability and documentation.

```
# Cleanup the column and make it readbale and consistent
data.columns = data.columns.str.lower()
data.columns = data.columns.str.replace(' ', '_')
```

Feature Engineering:

- ❖ Create age_group column from age by binning customer age.
- ❖ Create purchase_frequency_days from purchase data.

```
# Create a column of age groups for the customers
labels = ['Young Adult', 'Adult', 'Middle Aged', 'Senior']
data['age_group'] = pd.qcut(data['age'], q=4, labels=labels)
```

```python
# Create a column for purchase_frequency_days
frequency_mapping = {
    'Fortnightly' : 14,
    'Weekly' : 7,
    'Monthly' : 30,
    'Quarterly' : 90,
    'Annually' : 365,
    'Bi-Weekly' : 14,
    'Every 3 Months' : 90
}

data['purchase_frequency_days'] = data['frequency_of_purchases'].map(frequency_mapping)
```

Data Consistency Check: Verified the column discount_applied and promo_code_used were redundant then remove promo_code_used

Database Integration: Write the appropriate python scrip to connect to Postgres SQL and load the data into database.

4. Business Question Analysis Using SQL

1.1 Gender-Based Revenue Contribution Analysis

gender	totla_revene
Female	75,191
Male	157,890

1.2 High-Value Consumer Discount Utilization Trends

customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68

1.3 Product Quality Benchmarking: Top 5 Highest-Rated Items

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

1.4 Logistical Performance: Shipping Method Spend Comparison

average_purchased_amount	shipping_type
60.48	Express
58.46	Standard

1.5 Subscription ROI: Member vs. Non-Member Spend Metrics

total_customer	average_purchased_amount	total_purchased_amount	subscription_status
2,847	59.87	170,436	No
1,053	59.49	62,645	Yes

1.6 Price Elasticity: Highest Discount-Sensitive Products

item_purchased	percentage_purchased
Hat	50
Sneakers	49
Coat	49
Sweater	48
Pants	47

1.7 Customer Lifecycle Segmentation (New vs. Loyal Profiles)

segment	total_customer
Loyal	3,116
Returning	701
New	83

1.8 Hero Product Performance Across Categories

item_purchased	category	total_order	customer_ranking
Jewelry	Accessories	171	1
Sunglasses	Accessories	161	2
Belt	Accessories	161	3
Blouse	Clothing	171	1
Pants	Clothing	171	2
Shirt	Clothing	169	3
Sandals	Footwear	160	1

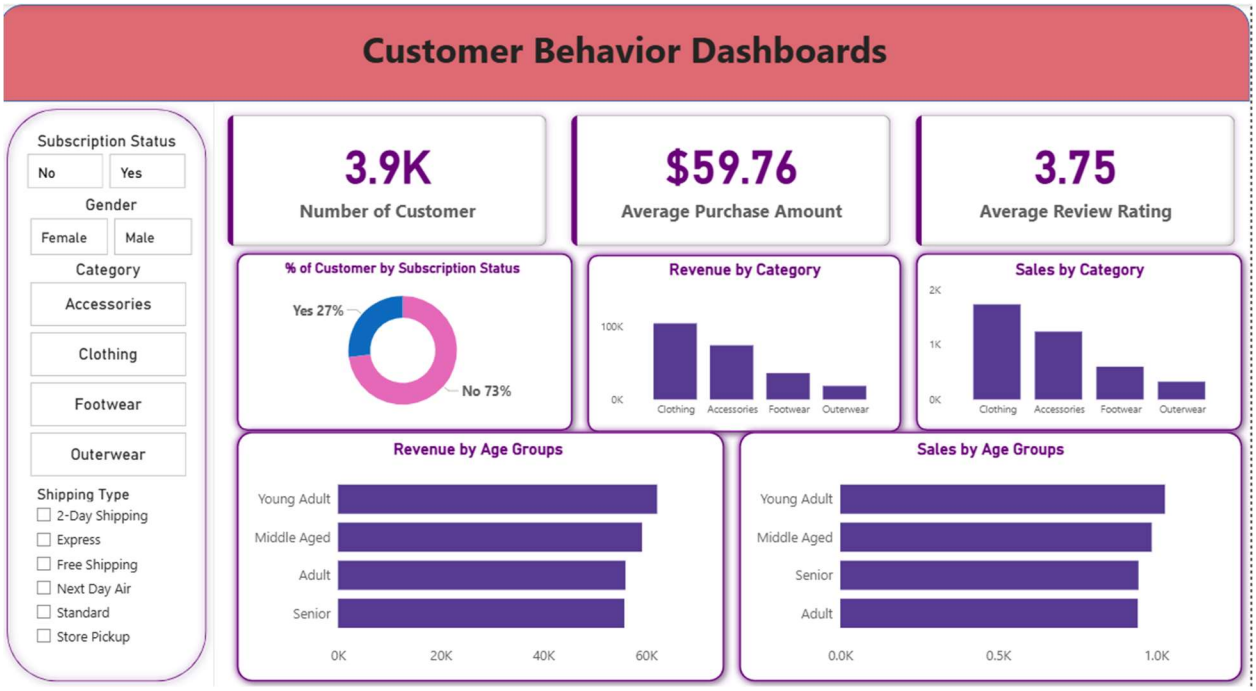
1.9 Retention Opportunity: Repeat Buyer Conversion Rate

subscription_status	repeat_buyer
No	2,518
Yes	958

2.0 Demographic Revenue Distribution by Age Cohort

age_group	revenue
Young Adult	62,143
Middle Aged	59,197
Adult	55,978
Senior	55,763

5. Customer Behavior Dashboards Using Power BI



6. Business Recommendations

Based on the quantitative analysis of the consumer behavior dataset, we have identified several strategic opportunities to optimize revenue and customer lifetime value:

- **Targeted Loyalty Conversion:** Our analysis shows that **72.4% of repeat buyers** (those with >5 purchases) are not currently subscribed to our membership program. We recommend a targeted email campaign offering a one-time "loyalty bonus" to convert these high-frequency shoppers into subscribers, securing long-term retention.
- **Male-Centric Growth Strategy:** Since **Male customers contribute 68% of total revenue** ($157,890), marketing budgets should be prioritized toward male-focused social media channels and product categories (Blouses, Pants, Shirts) to maximize immediate ROI.
- **Strategic Discounting for High-Elasticity Items:** Products like **Hats and Sneakers** show a near **50% discount-dependency rate**. We should utilize these specific items as "loss leaders" in promotional banners to drive traffic, while cross-selling high-margin, low-discount items like Jewelry and Belts.
- **Premium Shipping Upsells:** Data indicates that customers using **Express Shipping** have a higher average order value ($60.48) than those using Standard Shipping ($58.46). Implementing a "Free

Express Shipping on orders over $75" threshold could further incentivize larger basket sizes among our most active shoppers.

- **Quality-Led Marketing:** With **Gloves, Sandals, and Boots** consistently receiving the highest satisfaction ratings (3.82+ stars), these products should be featured in "Customer Choice" or "Best Rated" marketing modules to reduce purchase hesitation for new customers.
- **Age-Specific Campaign Timing:** Revenue is remarkably consistent across the **26-65 age cohorts**. We recommend a broad-reach branding strategy for this demographic, while testing more aggressive "first-purchase" incentives for the **18-25 and 66-75 groups**, which currently represent our lowest revenue contributors.