Internship: ALFIDO Tech Data Analytics Internship

Submitted By: Palak Ramesh Kale

Table Of Content

Abstract

Customer retention and value optimization are critical challenges for e-commerce businesses in a highly competitive market. This project presents a comprehensive analysis of customer transaction and behavior data to identify purchasing patterns, customer segments, and churn risks. Using a large-scale e-commerce dataset, the study applies data preprocessing, exploratory data analysis, and RFM (Recency, Frequency, Monetary) analysis to segment customers based on their engagement and spending behavior.

The analysis explores category-wise sales performance, month-over-month revenue trends, active customer patterns, and churn distribution to evaluate overall business performance. Customers are classified into meaningful segments such as Champions, Loyal Customers, At-Risk Customers, and Lost Customers, enabling a clear understanding of value contribution and retention status. Visualisations are used to highlight differences in spending behavior across segments and identify periods of declining engagement.

The findings reveal that a small group of high-value customers contributes a substantial share of total revenue, while a significant proportion of customers exhibit churn behavior, indicating the need for proactive retention strategies. The project demonstrates how customer segmentation and behavioral analysis can support data-driven decision-making, helping businesses improve customer engagement, reduce churn, and maximise customer lifetime value.

➢ Introduction

In today's competitive e-commerce environment, understanding customer behavior is essential for improving revenue, retention, and long-term business sustainability. Customers interact with online platforms in different ways—some purchase frequently and generate high value, while others gradually disengage and eventually churn. Identifying these patterns early allows businesses to design targeted marketing, retention, and engagement strategies.

This project focuses on customer analytics using transactional e-commerce data to study purchasing behavior, customer value, and retention trends. The dataset includes customer demographics, purchase history, product categories, payment methods, monetary value, and churn indicators across multiple time periods. By applying RFM (Recency, Frequency, Monetary) analysis, customers are segmented into meaningful groups such as Champions, Loyal Customers, At-Risk Customers, and Lost Customers.

In addition to segmentation, the project analyzes month-over-month active customers, repeat purchasing behavior, revenue trends, and churn distribution. Visualizations and statistical summaries are used to understand sales performance, customer engagement, and category-wise contribution to revenue. These insights help identify high-value customers, detect early signs of churn, and evaluate overall business health.

The outcomes of this project demonstrate how data-driven customer segmentation and retention analysis can support strategic decision-making, enabling businesses to improve customer lifetime value, optimize marketing efforts, and strengthen long-term growth.

➢ Problem Statement

Task 1 : Customer Behavior Analysis

Dataset: https://www.kaggle.com/datasets/bhanupratapbiswas/customer-behavior-analysis

Goal: Analyze customer transactions & behavior to identify segments, purchase patterns, and churn risks.

Requirements:

- Perform data cleaning and feature engineering
- Segment customers (RFM or clustering) and profile each segment
- Visualize purchase patterns and retention trends

➢ Objectives

1. To analyze customer transaction data in order to understand purchasing behavior and spending patterns over time.
2. To perform data preprocessing and feature engineering to ensure data quality and derive meaningful customer-level metrics.
3. To segment customers using RFM (Recency, Frequency, Monetary) analysis and classify them into meaningful behavioral groups.
4. To profile each customer segment based on recency, frequency, and monetary value to identify high-value, loyal, and at-risk customers.
5. To analyze month-over-month customer activity and revenue trends in order to assess customer retention and engagement.
6. To identify customer churn by defining inactivity-based churn criteria and analyzing churn distribution.
7. To study product category-wise sales performance and identify key revenue-generating categories.
8. To derive actionable business insights and recommendations that support customer retention, engagement, and revenue growth.

➢ Dataset Description

Dataset Name : Customer Behavior Analysis Dataset

Source : Kaggle – *Customer Behavior Analysis*
This dataset contains transaction-level customer purchase data designed to analyze customer behavior, purchasing patterns, retention, and churn risk.
Each record represents a single customer transaction, capturing demographic, transactional, and behavioral attributes**.**

Column Name	Description
Customer ID	Unique identifier for each customer
Purchase Date	Date on which the transaction occurred
Product Category	Category of the purchased product
Product Price	Price of a single product
Quantity	Number of units purchased
Total Purchase Amount	Total transaction value (Price × Quantity)
Payment Method	Mode of payment used (e.g., Card, UPI, Cash)
Customer Age	Age of the customer
Returns	Indicates whether the product was returned

➢ Data Preprocessing

Data preprocessing was carried out to ensure data quality and analytical consistency. The dataset was first inspected to understand its structure, data types, and completeness. The Purchase Date variable was converted into datetime format to enable time-based analysis such as recency calculation and monthly trend analysis. Missing values in numerical and categorical variables were handled using appropriate techniques, and duplicate or invalid records were identified and removed.

Feature engineering was performed to derive meaningful customer-level metrics, including Recency (days since last purchase), Frequency (number of transactions), and Monetary value (total spending). Transaction-level data was aggregated at the customer level, and a churn flag was created to identify customers who had not made purchases within a defined inactivity period. These steps ensured that the dataset was well-prepared for customer segmentation, retention analysis, and churn trend evaluation.

➢ Methodology and Implementation

1. Importing Libraries and Customer Behavior Data

```
# Importing necessary libraries
import pandas as pd
import numpy as np
import datetime
```

```
# Importing Customer behaviour data
data=pd.read_csv('/content/ecommerce_customer_data_custom_ratios.csv')
```

```
data.head(10)
```

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
0	46251	2020-09-08 09:38:32	Electronics	12	3	740	Credit Card	37	0.0	Christine Hernandez	37	Male	0
1	46251	2022-03-05 12:56:35	Home	468	4	2739	PayPal	37	0.0	Christine Hernandez	37	Male	0
2	46251	2022-05-23 18:18:01	Home	288	2	3196	PayPal	37	0.0	Christine Hernandez	37	Male	0
3	46251	2020-11-12 13:13:29	Clothing	196	1	3509	PayPal	37	0.0	Christine Hernandez	37	Male	0
4	13593	2020-11-27 17:55:11	Home	449	1	3452	Credit Card	49	0.0	James Grant	49	Female	1
5	13593	2023-03-07 14:17:42	Home	250	4	575	PayPal	49	1.0	James Grant	49	Female	1
6	13593	2023-04-15 03:02:33	Electronics	73	1	1896	Credit Card	49	0.0	James Grant	49	Female	1
7	13593	2021-03-27 21:23:28	Books	337	2	2937	Cash	49	0.0	James Grant	49	Female	1
8	13593	2020-05-05 20:14:00	Clothing	182	2	3363	PayPal	49	1.0	James Grant	49	Female	1
9	28805	2023-09-13 04:24:00	Electronics	394	2	1993	Credit Card	19	0.0	Jose Collier	19	Male	0

2. EDA and Data Cleaning

```
# Details of the data
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 13 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Customer ID           250000 non-null  int64
 1   Purchase Date         250000 non-null  object
 2   Product Category      250000 non-null  object
 3   Product Price         250000 non-null  int64
 4   Quantity              250000 non-null  int64
 5   Total Purchase Amount 250000 non-null  int64
 6   Payment Method        250000 non-null  object
 7   Customer Age          250000 non-null  int64
 8   Returns               202404 non-null  float64
 9   Customer Name         250000 non-null  object
 10  Age                   250000 non-null  int64
 11  Gender                250000 non-null  object
 12  Churn                 250000 non-null  int64
dtypes: float64(1), int64(7), object(5)
```

The Purchase date column has datatype object. We should convert it to a datetime datatype for further analysis.

```
# Converting Datatype from object to datetime
data['Purchase Date']=pd.to_datetime(data['Purchase Date'],errors='coerce')
```

```
# Details of the data
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 13 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Customer ID           250000 non-null  int64
 1   Purchase Date         250000 non-null  datetime64[ns]
 2   Product Category      250000 non-null  object
 3   Product Price         250000 non-null  int64
 4   Quantity              250000 non-null  int64
 5   Total Purchase Amount 250000 non-null  int64
 6   Payment Method        250000 non-null  object
 7   Customer Age          250000 non-null  int64
 8   Returns               202404 non-null  float64
 9   Customer Name         250000 non-null  object
 10  Age                   250000 non-null  int64
 11  Gender                250000 non-null  object
 12  Churn                 250000 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(7), object(4)
memory usage: 24.8+ MB
```

```
# Checking for null values
data.isnull().sum()
```

	0
Customer ID	0
Purchase Date	0
Product Category	0
Product Price	0
Quantity	0
Total Purchase Amount	0
Payment Method	0
Customer Age	0
Returns	47596
Customer Name	0
Age	0
Gender	0
Churn	0

dtype: int64

The only 'Returns' column has 47596 null values. Filling it by mean.

```
# Filling null values by mean
data=data.fillna('mean')
```

```
data.isnull().sum()
```

	0
Customer ID	0
Purchase Date	0
Product Category	0
Product Price	0
Quantity	0
Total Purchase Amount	0
Payment Method	0
Customer Age	0
Returns	0
Customer Name	0
Age	0
Gender	0

Churn	0

dtype: int64

All null values are filled by mean.

```
# Checking duplicate values
data.duplicated().sum()
```

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

There are no duplicate values in the data.

```
# Statistical Summary
data.describe()
```

	Customer ID	Purchase Date	Product Price	Quantity	Total Purchase Amount	Customer Age	Age	Churn
count	250000.00000	250000	250000.000000	250000.000000	250000.000000	250000.000000	250000.000000	250000.000000
mean	25004.03624	2021-11-06 23:31:24.372304384	254.659512	2.998896	2725.370732	43.940528	43.940528	0.199496
min	1.00000	2020-01-01 00:15:00	10.000000	1.000000	100.000000	18.000000	18.000000	0.000000
25%	12497.75000	2020-12-02 19:33:23.249999872	132.000000	2.000000	1477.000000	31.000000	31.000000	0.000000
50%	25018.00000	2021-11-06 13:10:59	255.000000	3.000000	2724.000000	44.000000	44.000000	0.000000
75%	37506.00000	2022-10-11 03:42:32.750000128	377.000000	4.000000	3974.000000	57.000000	57.000000	0.000000
max	50000.00000	2023-09-15 12:24:08	500.000000	5.000000	5350.000000	70.000000	70.000000	1.000000
std	14428.27959	NaN	141.568577	1.414694	1442.933565	15.350246	15.350246	0.399622

➢ Conclusions:

1. Dataset Size and Coverage

- The dataset contains 250,000 transaction records, indicating a large and reliable sample size for customer behavior analysis.
- Customer IDs range from 1 to 50,000, suggesting a broad customer base with multiple transactions per customer.

2. Purchase Timeline

- Transactions span from January 2020 to September 2023, covering nearly four years of customer activity.
- This wide time range makes the dataset suitable for trend analysis, retention studies, and churn analysis.

3. Pricing and Purchase Behavior

- The average product price is approximately ₹255, with prices ranging from ₹10 to ₹500, indicating a mix of low-value and moderately priced products.
- Customers purchase an average of 3 items per transaction, with quantities ranging from 1 to 5 units.
- The average total purchase amount per transaction is approximately ₹2,725, suggesting moderate transaction values.

4. Customer Spending Distribution

- The 25th percentile of total purchase amount is ₹1,477, while the 75th percentile is ₹3,974, showing noticeable variability in customer spending.
- A relatively high standard deviation in total purchase amount indicates heterogeneous customer spending behavior, which supports the need for customer segmentation (RFM analysis).

5. Customer Demographics

- The average customer age is approximately 44 years, with customers ranging from 18 to 70 years.
- This indicates a diverse age group, making age-based behavior analysis and targeted engagement strategies possible.

6. Churn Behavior

- The average churn value is approximately 0.20, indicating that around 20% of customers are churned.
- This highlights a significant churn risk, emphasizing the importance of retention strategies and churn prevention efforts.

7. Data Quality Observations

- No missing values are observed in key numerical variables, indicating good data completeness.
- Standard deviation values suggest natural variability without extreme anomalies, making the data suitable for further modeling and segmentation.

```
data.describe(include='O')
```

	Product Category	Payment Method	Returns	Customer Name	Gender
count	250000	250000	250000.0	250000	250000
unique	4	4	3.0	39920	2
top	Clothing	Credit Card	0.0	Michael Smith	Female
freq	75052	100486	101635.0	107	125560

1. Product Category Distribution

- The dataset consists of 4 distinct product categories, indicating a focused but diverse product portfolio.
- Clothing is the most frequently purchased category, accounting for the highest number of transactions.
- This suggests that clothing products are a major revenue driver and should be prioritized in promotional strategies.

2. Payment Method Preference

- There are 4 different payment methods used by customers.
- Credit Card is the most commonly used payment method, with the highest transaction frequency.
- This indicates strong customer preference for digital payment methods, highlighting the importance of ensuring a seamless and secure card payment experience.

3. Returns Behavior

- The returns variable has 3 distinct values, indicating different return states.
- The most common value is 0 (no returns), showing that the majority of transactions do not result in returns.
- This suggests overall customer satisfaction with products and relatively low return rates.

4. Customer Name Distribution

- The dataset contains a large number of unique customer names (~39,920), confirming a broad and diverse customer base.

- The most frequent customer name appears only a small number of times, indicating that the dataset is not dominated by a few customers.
- This supports fair and unbiased customer segmentation.

5. Gender Distribution

- The dataset includes 2 gender categories.
- Female customers represent the majority of transactions.
- This indicates a potential opportunity for female-focused marketing campaigns, especially in dominant product categories such as clothing.

Feature Engineering

Creating RFM Table

R = Recency, F = Frequency, M = Monetary

```
# RFM Date features
data['Year']=data['Purchase Date'].dt.year
data['Month']=data['Purchase Date'].dt.month
data['DayOfWeek']=data['Purchase Date'].dt.day_name()
```

```
# Reference date to calculate Recency
snapshot_date=data['Purchase Date'].max()+pd.Timedelta(days=1)
```

Finds the most recent purchase date in the dataset.

Adds 1 day to it.

Why add 1 day? =>To avoid Recency = 0 for customers who purchased on the last date.

```
# RFM Table creation
rfm=data.groupby('Customer ID').agg({
    'Purchase Date': lambda x: (snapshot_date - x.max()).days,
    'Customer ID': 'count',
    'Total Purchase Amount': 'sum'
})
```

```
# Rename RFM columns
rfm.columns=['Recency','Frequency','Monetary']
```

Converting raw RFM values(Recency, frequecy, Monetary) into scores

```
rfm['R_Score']=pd.qcut(rfm['Recency'],4,labels=False,duplicates='drop')
rfm['F_Score']=pd.qcut(rfm['Frequency'],4,labels=False,duplicates='drop')
rfm['M_Score']=pd.qcut(rfm['Monetary'],4,labels=False,duplicates='drop')

rfm['RFM_Score']=rfm[['R_Score','F_Score','M_Score']].sum(axis=1)
```

```
rfm
```

Customer ID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score
1	58	1	3491	0	0	0	0
2	299	3	7988	2	0	0	2
3	89	8	22587	1	3	3	7
4	127	4	8715	1	1	0	2
5	171	8	12524	1	3	1	5
...	...	...	...	...	...	...	...
49996	272	4	14044	2	1	2	5
49997	50	8	22431	0	3	3	6
49998	12	4	8610	0	1	0	1
49999	420	2	6984	3	0	0	3
50000	127	2	4116	1	0	0	1

49673 rows × 7 columns

➢ Profiling Customer Segments

1. Creating RFM based segments

```
def rfm_segment(row):
    if row['RFM_Score'] >= 8:
        return 'Champions'
    elif row['RFM_Score'] >= 6:
        return 'Loyal Customers'
    elif row['RFM_Score'] >= 4:
        return 'Potential Loyalists'
    elif row['RFM_Score'] >= 2:
        return 'At-Risk'
    else:
        return 'Lost Customers'

rfm['Segment'] = rfm.apply(rfm_segment, axis=1)
```

2. Profiling Each Segment

```
segment_profile = rfm.groupby('Segment').agg({
    'Recency': 'mean',
    'Frequency': 'mean',
    'Monetary': 'mean',
    'RFM_Score': 'mean',
    'Segment': 'count'
}).rename(columns={'Segment':'Customer_Count'})

segment_profile
```

Segment	Recency	Frequency	Monetary	RFM_Score	Customer_Count
At-Risk	325.239625	3.320233	8359.329089	2.663348	15470
Champions	346.493868	7.752617	22520.446306	8.195334	3343
Lost Customers	76.052785	2.785411	6454.304509	0.671618	3770
Loyal Customers	205.449495	7.171796	20462.730271	6.379025	12672
Potential Loyalists	270.897628	4.947774	13393.089471	4.487377	14418

1. Champions

Low recency → purchased recently

High frequency → buy often

High monetary value → spend more

Business meaning: Most valuable customers

2. Loyal Customers

Purchase frequently but not always recently

Medium–high spending

Stable revenue contributors

3. Potential Loyalists

Recent buyers

Moderate purchase frequency

Medium spending

4. At-Risk Customers

Long time since last purchase

Previously good spenders

Declining engagement

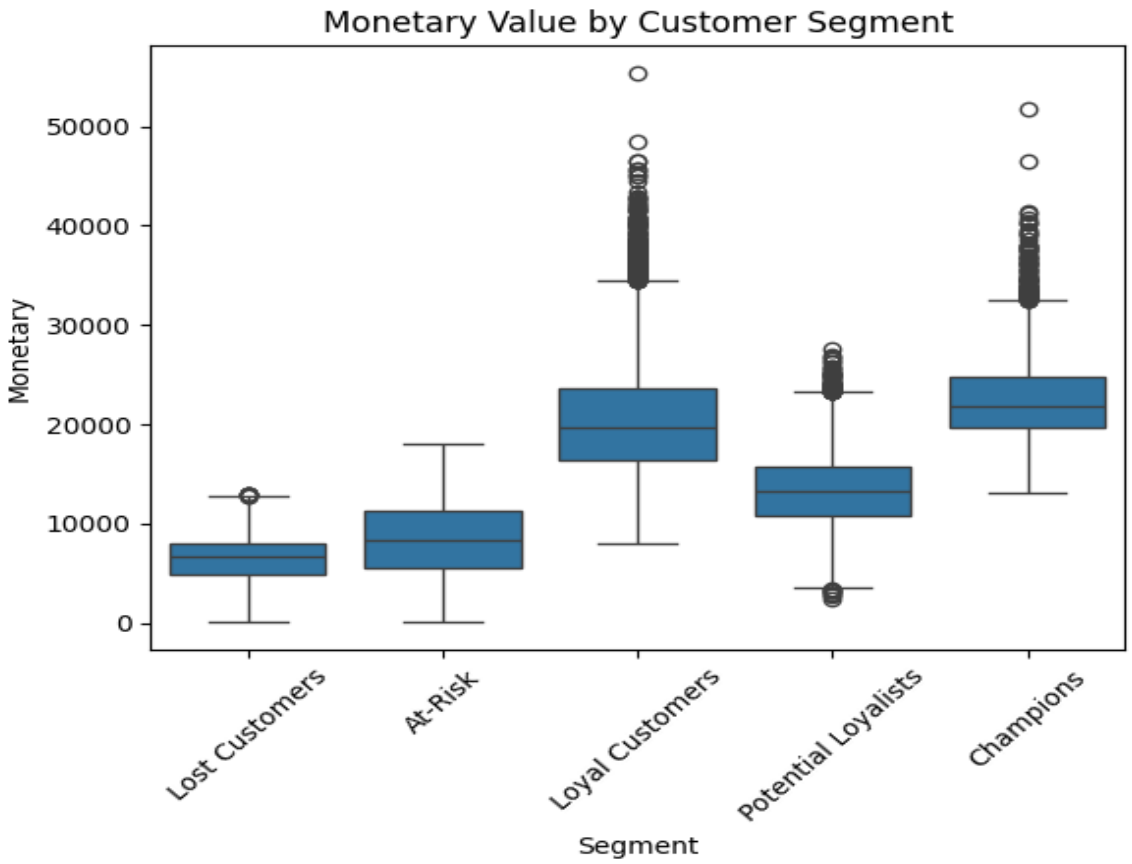
5. Lost Customers

Very high recency (inactive)

Low frequency

Low spending

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.boxplot(x='Segment', y='Monetary', data=rfm)
plt.xticks(rotation=45)
plt.title("Monetary Value by Customer Segment")
plt.show()
```



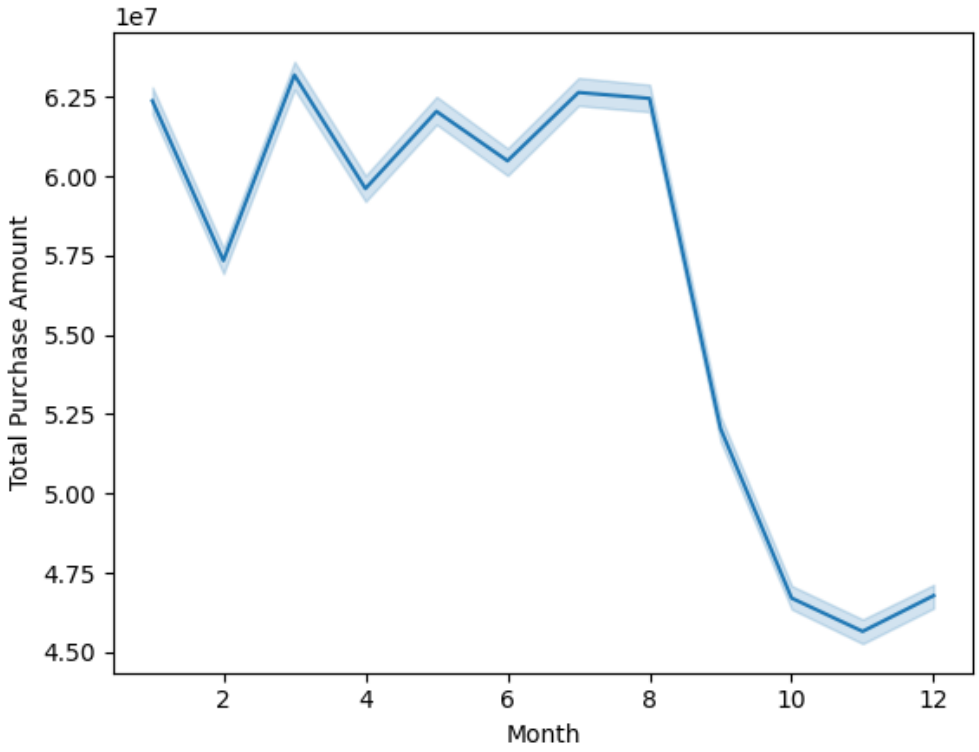
Conclusion :

The box plot shows clear differences in spending across customer segments. Champions and Loyal Customers have the highest monetary values, making them the main revenue contributors. Potential Loyalists show moderate spending with growth potential, while At-Risk and Lost Customers have the lowest spending, indicating declining or minimal engagement. This validates the effectiveness of RFM segmentation and highlights where retention and upselling efforts should focus.

➢ Data Visualization

A. Purchase Patterns

```
# Line Plot
import seaborn as sns
sns.lineplot(data=data, x='Month', y='Total Purchase Amount', estimator='sum')
```

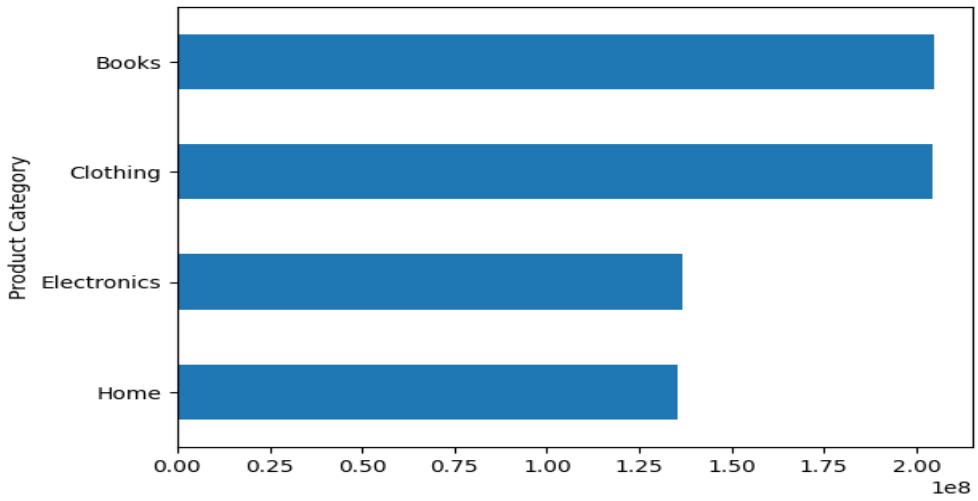


Conclusion :

The line plot shows that total purchase amount remains relatively high and stable during the first eight months, with noticeable peaks around mid-year. A sharp decline is observed from September onwards, reaching the lowest levels in October and November, followed by a slight recovery in December. This indicates seasonal purchasing behavior, suggesting the need for targeted promotions and retention strategies during the later months of the year to maintain revenue.

B. Category Preferences

```
# Bar Plot
data.groupby('Product Category')['Total Purchase Amount'].sum().sort_values().plot(kind='barh')
```



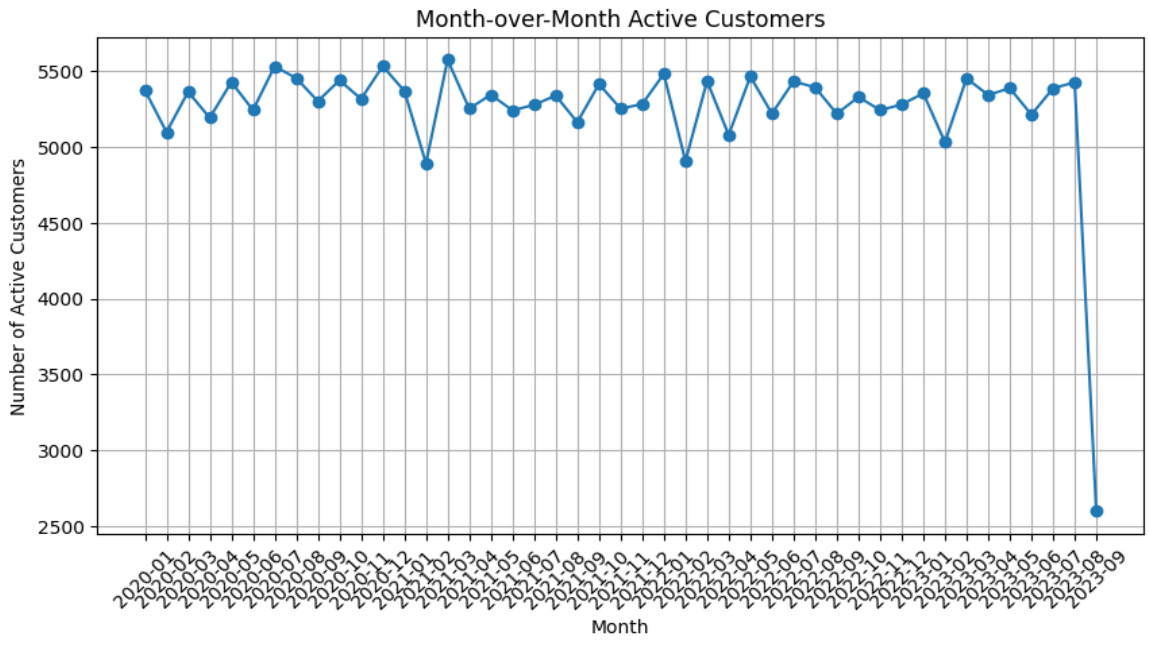
Conclusion :

The bar plot indicates that Books and Clothing generate the highest total purchase amounts, making them the top-performing product categories. Electronics and Home categories contribute comparatively lower sales. This suggests that revenue is primarily driven by Books and Clothing, and these categories should be prioritized for inventory planning, promotions, and targeted marketing, while growth opportunities exist for Electronics and Home through strategic offers and product positioning.

C. Month-over-Month Repeat Buyers (Retention Trend)

```
# Creating Year-Month column
data['YearMonth'] = data['Purchase Date'].dt.to_period('M')

# Count unique customers per month
monthly_customers = (
    data.groupby('YearMonth')['Customer ID']
    .nunique()
    .reset_index()
)

monthly_customers['YearMonth'] = monthly_customers['YearMonth'].astype(str)
```

```
import matplotlib.pyplot as plt

plt.figure(figsize=(10,5))
plt.plot(monthly_customers['YearMonth'], monthly_customers['Customer ID'], marker='o')
plt.xticks(rotation=45)
plt.title('Month-over-Month Active Customers')
plt.xlabel('Month')
plt.ylabel('Number of Active Customers')
plt.grid(True)
plt.show()
```



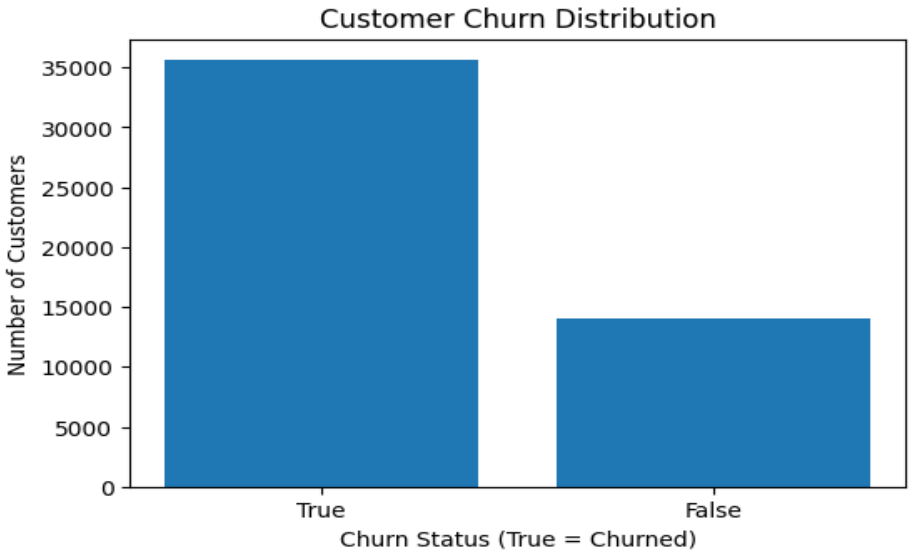
Conclusion :

The scatter plot shows that the number of active customers remains largely stable over time, fluctuating around 5,200–5,500 customers per month, indicating consistent customer engagement. Occasional dips suggest short-term inactivity or seasonal effects. The sharp decline in the final month likely reflects incomplete data or a reporting cutoff rather than a true drop-in customer activity and should be interpreted with caution.

D. Churn Flag – No Purchase in Last X Months

```
snapshot_date = data['Purchase Date'].max()
churn_months = 3

# Last Purchase per customer
last_purchase = data.groupby('Customer ID')['Purchase Date'].max().reset_index()

# Create churn flag
last_purchase['Months_Since_Last_Purchase'] = (
    (snapshot_date - last_purchase['Purchase Date']) / pd.Timedelta(days=30)
)

last_purchase['Churn_Flag'] = last_purchase['Months_Since_Last_Purchase'] > churn_months

# Count churned vs active customers
churn_counts = last_purchase['Churn_Flag'].value_counts().reset_index()
churn_counts.columns = ['Churn_Flag', 'Customer_Count']
```

```
# Plot churn distribution
plt.figure(figsize=(6,4))
plt.bar(churn_counts['Churn_Flag'].astype(str), churn_counts['Customer_Count'])
plt.title('Customer Churn Distribution')
plt.xlabel('Churn Status (True = Churned)')
plt.ylabel('Number of Customers')
plt.show()
```



Conclusion :

The bar plot indicates that a larger proportion of customers have churned compared to those who remain active. This suggests a significant retention challenge, highlighting the need for targeted customer retention strategies such as loyalty programs, personalized offers, and proactive engagement to reduce future churn.

Retention & Loyalty Analysis

```
last_purchase = data.groupby("Customer ID")['Purchase Date'].max().reset_index()
last_purchase['Churn'] = last_purchase['Purchase Date'] < pd.to_datetime("2025-01-01")
```

```
np.True_
```

Gives Customer churn or not by providing Customer ID.

➢ Business Insights :

1. High-Value Customers Drive Revenue
   Champions and Loyal Customers contribute the maximum monetary value, indicating that retaining these segments is crucial for sustained revenue growth.
2. Retention Is More Cost-Effective Than Acquisition
   A significant proportion of customers have churned, highlighting the need for focused retention strategies rather than relying only on new customer acquisition.
3. Early Warning for Churn Exists
   At-Risk and Potential Loyalist segments exhibit declining spending patterns, creating an opportunity to offer timely promotions and personalize engagement to prevent churn.
4. Product Category Performance Is Uneven
   Clothing and Books generate the highest sales, suggesting strong demand, while Electronics and Home categories may require better pricing, promotions, or assortment strategies.
5. Seasonal or Period-End Sales Drop
   The decline in total purchase amount and active customers in later months indicates possible seasonality or customer disengagement that should be addressed with targeted campaigns.
6. Payment and Purchase Behavior Is Concentrated
   Certain payment methods and product categories dominate customer choices, allowing the business to optimize checkout experience and partnerships around popular options.
7. Customer Age Group Is a Key Target
   The average customer age falls in the mid-40s, helping the business tailor marketing messages, product offerings, and channels toward this dominant demographic.

8. RFM Segmentation Enables Personalization
   Segment-wise profiling supports personalized marketing, such as loyalty rewards for champions, discounts for at-risk customers, and onboarding incentives for new buyers.

9. Returns and Churn Are Interlinked
   Customers with higher return behavior may contribute to churn, suggesting the need to improve product quality, descriptions, and post-purchase support.

10. Data-Driven Strategy Improves Decision-Making
    Combining RFM analysis, churn trends, and sales patterns enables the business to make informed decisions on marketing spend, inventory planning, and customer relationship management.