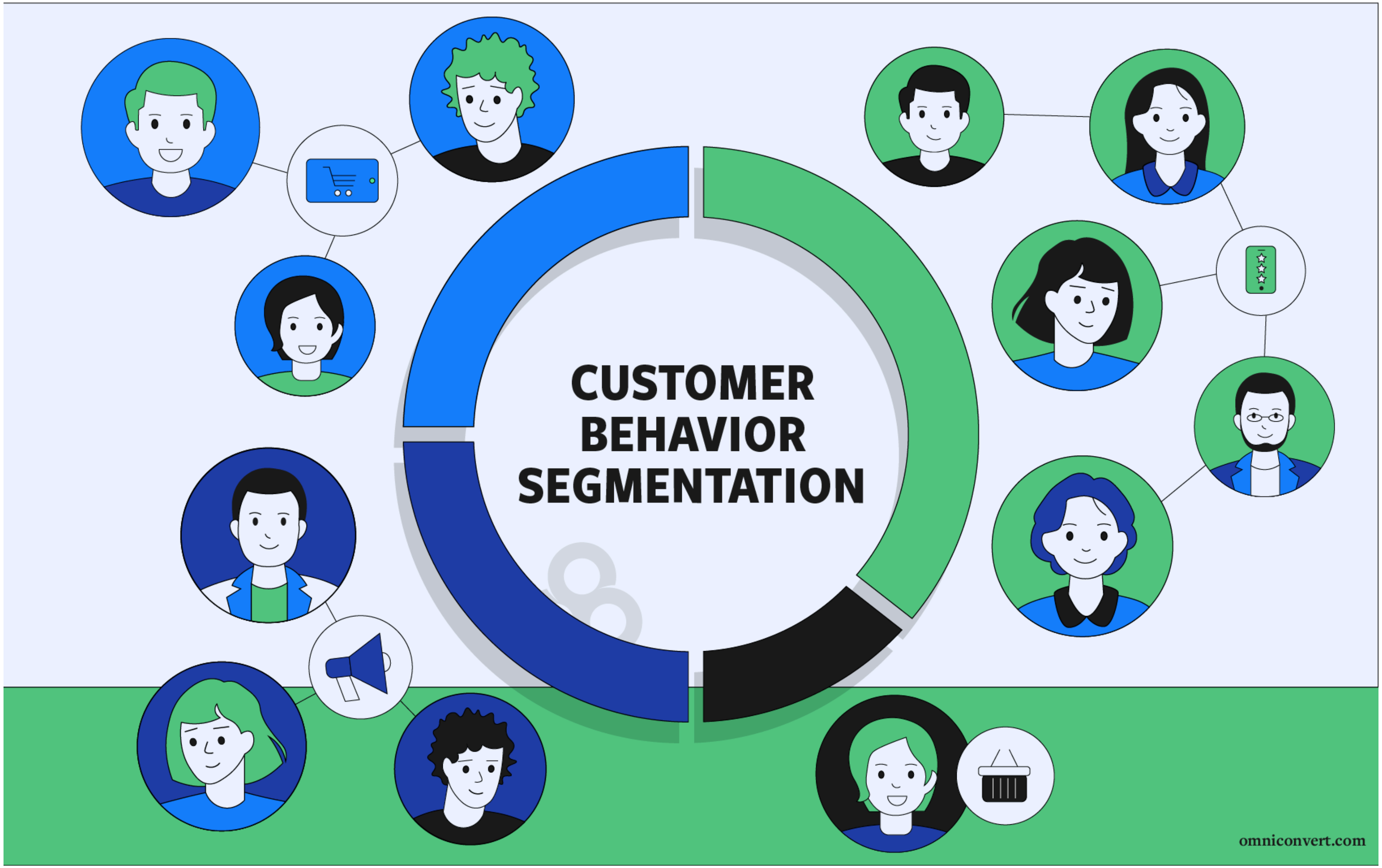
🛒 E-commerce Customer Data For Behavior Analysis



This project involves analyzing customer data from an e-commerce platform to understand user behavior, purchasing patterns, and preferences. By examining variables such as purchase history, browsing data, customer demographics, and product reviews, the goal is to uncover actionable insights that can help improve customer experience, personalize marketing strategies, and increase sales.

Explore Customer Shopping Habits, Churn, and Purchase Patterns 💳 🛒

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
df=pd.read_csv("ecommerce_customer_data_custom_ratios.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	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

In [4]:

```
df.tail()
```

Out[4]:

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
249995	33308	2023-08-10 13:39:06	Clothing	279	2	2187	PayPal	55	1.0	Michelle Flores	55	Male	1
249996	48835	2021-11-23 01:30:42	Home	27	1	3615	Credit Card	42	1.0	Jeremy Rush	42	Female	1
249997	21019	2020-07-02 14:04:48	Home	17	5	2466	Cash	41	0.0	Tina Craig	41	Male	0
249998	49234	2020-12-30 02:02:40	Books	398	2	3668	Crypto	34	0.0	Jennifer Cooper	34	Female	1
249999	16971	2021-03-13 16:28:35	Electronics	425	4	2370	Cash	36	1.0	Justin Lawson	36	Female	1

Objectives:

Identify key customer segments based on behavior and demographics

Analyze product preferences and purchasing trends

Predict future purchases using machine learning models

Improve customer retention by identifying churn indicators

Provide data-driven recommendations for business growth

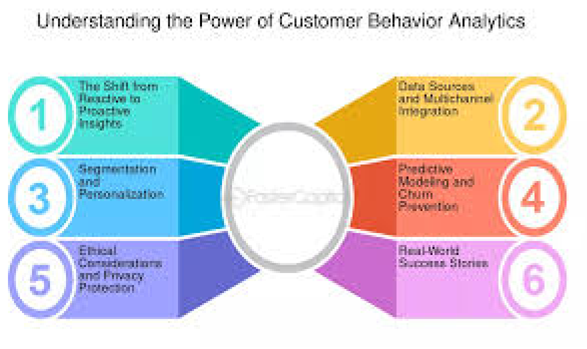
Key Features:

Data preprocessing and cleaning

Exploratory data analysis (EDA) with visualizations

Customer segmentation using clustering techniques

Predictive modeling (e.g., classification or regression)



In [5]:
```
df.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)
memory usage: 24.8+ MB
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
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
```

In [7]:
```
df['Returns'].unique()
```

Out[7]:
```
array([ 0.,  1., nan])
```

In [8]:
```
df['Product Category'].unique()
```

Out[8]:
```
array(['Electronics', 'Home', 'Clothing', 'Books'], dtype=object)
```

In [9]:
```
df['Churn'].unique()
```

Out[9]:
```
array([0, 1], dtype=int64)
```

In [10]:
```
df['Payment Method'].unique()
```

Out[10]:
```
array(['Credit Card', 'PayPal', 'Cash', 'Crypto'], dtype=object)
```

In [11]:
```
df['Returns'].fillna(0,inplace=True)
```

In [12]:
```
total_revalue=df["Total Purchase Amount"].sum()
```

In [13]:
```
total_revalue
```

Out[13]: 681342683

In [14]:
```
avo=df['Total Purchase Amount'].mean()
```

In [15]:
```
avo
```

Out[15]: 2725.370732

In [16]:
```
unique_custumer=df['Customer ID'].nunique()
```

In [17]:
```
unique_custumer
```

Out[17]: 49673

In [18]:
```
repate_custumer=df['Customer ID'].value_counts()
```

In [19]:
```
repate_custumer
```

Out[19]:
```
Customer ID
36437    17
47087    17
39817    17
5252     15
14400    15
         ..
6861      1
49276     1
40043     1
31599     1
16971     1
Name: count, Length: 49673, dtype: int64
```

Top Products and Categories

In [20]:
```
df.head(2)
```

Out[20]:

	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

In [21]:
```
df.groupby('Product Category')['Total Purchase Amount'].sum().sort_values(ascending=False).head()
```

Out[21]:
```
Product Category
Books          204939601
Clothing       204532405
Electronics    136599467
Home           135271210
Name: Total Purchase Amount, dtype: int64
```

In [22]:
```
df.groupby('Product Category')['Quantity'].sum().sort_values(ascending=False).head()
```

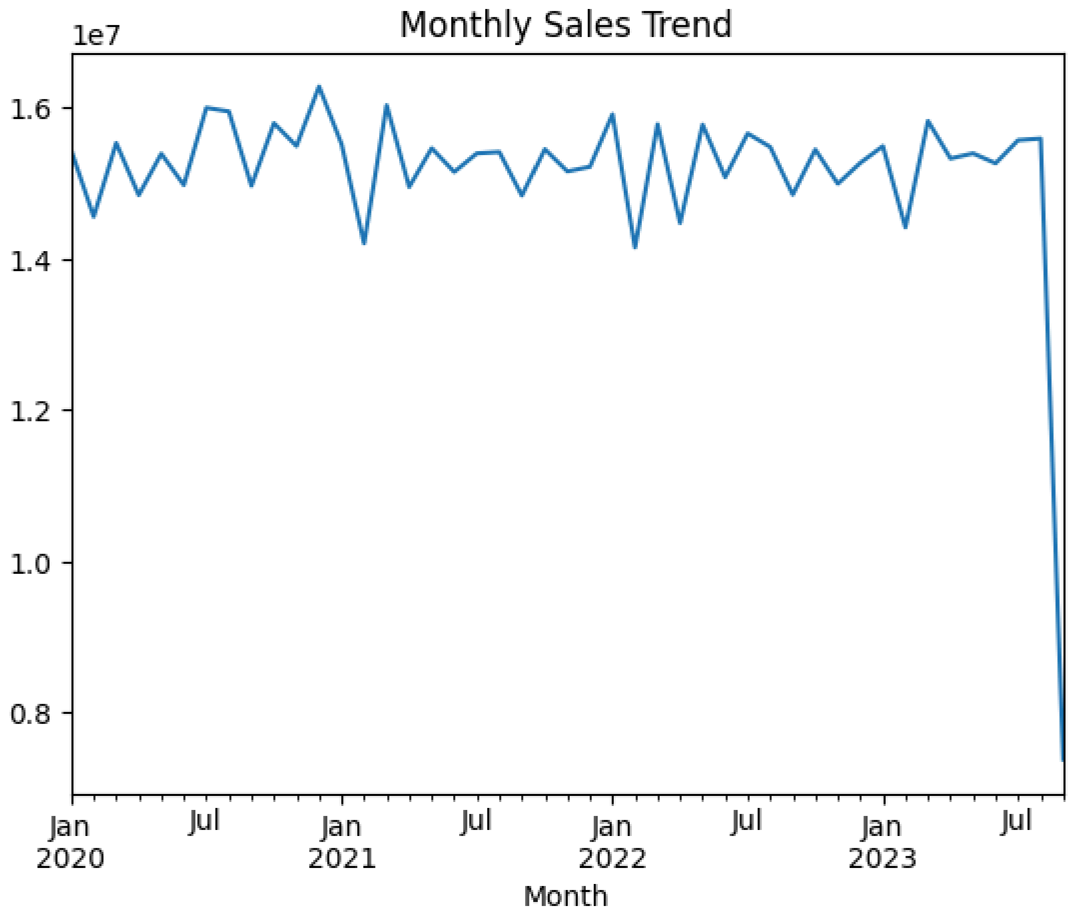
Out[22]:
```
Product Category
Clothing       225322
Books          223876
Electronics    150828
Home           149698
Name: Quantity, dtype: int64
```

📅 Time Series Analysis

In [23]:
```
df['Purchase Date'] = pd.to_datetime(df['Purchase Date'], errors='coerce')
```

In [24]:
```
df['Month'] = df['Purchase Date'].dt.to_period('M')
monthly_sales = df.groupby('Month')['Total Purchase Amount'].sum()
monthly_sales.plot(kind='line', title='Monthly Sales Trend')
```

Out[24]:
```
<Axes: title={'center': 'Monthly Sales Trend'}, xlabel='Month'>
```



Customer Segmentation (RFM Analysis)

In [25]:
```
import datetime as dt
snapshot_date = df['Purchase Date'].max() + pd.Timedelta(days=1)
```

In [26]:
```
rfm = df.groupby('Customer ID').agg({
    'Purchase Date': lambda x: (snapshot_date - x.max()).days,
    'Customer ID': 'count',
    'Total Purchase Amount': 'sum'
}).rename(columns={
```

```
    'Purchase Date': 'Recency',
    'Customer ID': 'Frequency',
    'Total Purchase Amount': 'Monetary'
})
```

Return Behavior

In [27]:
```
returns=df[df['Returns']>0]
returns_by_category=returns.groupby('Product Category')['Returns'].sum()
```

Churn Analysis

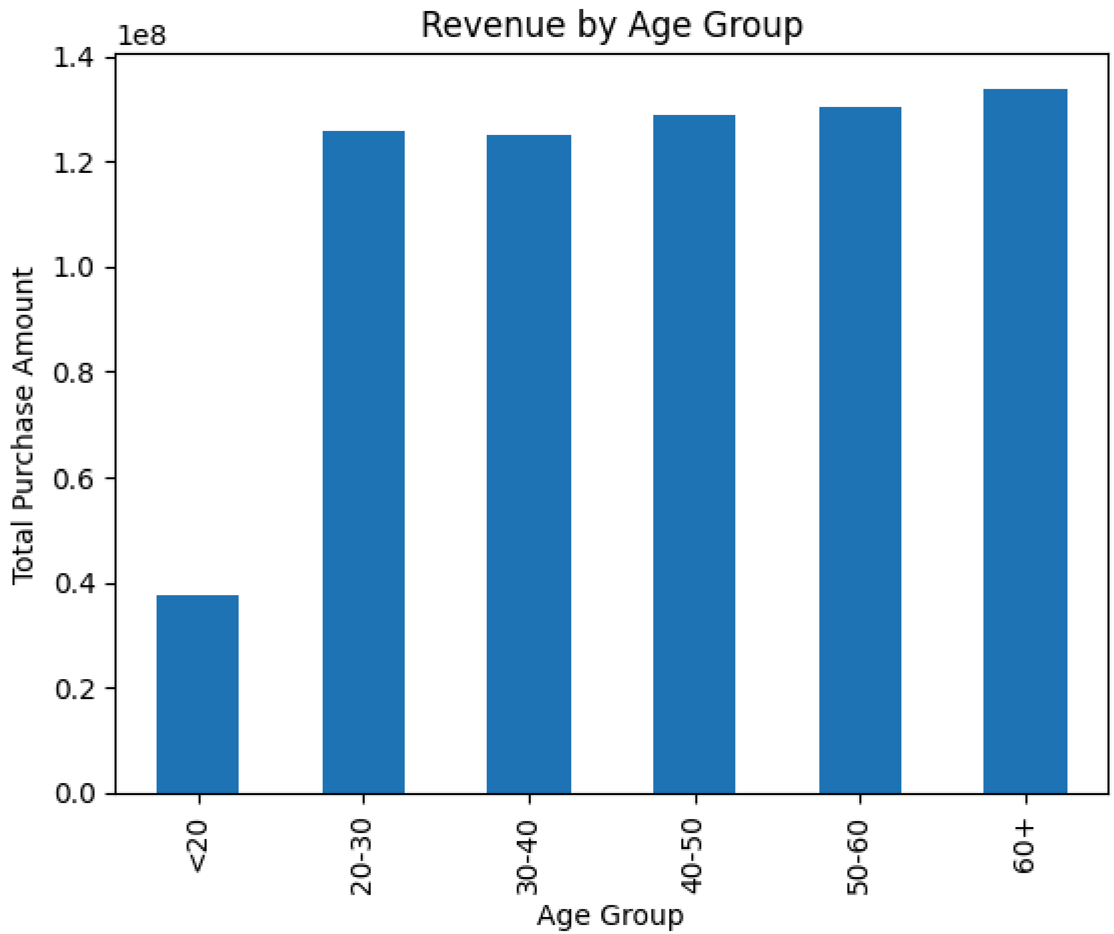
In [28]:
```
churned=[df['Churn']==0]
churned=[df['Churn']==1]
```

In [29]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Gender-based spend
sns.barplot(x='Gender', y='Total Purchase Amount', data=df)

# Age groups
df['Age Group'] = pd.cut(df['Age'], bins=[0, 20, 30, 40, 50, 60, 100], labels=['<20','20-30','30-40','40-50','50-60','60+'])
age_group_revenue = df.groupby('Age Group')['Total Purchase Amount'].sum()
age_group_revenue.plot(kind='bar', title='Revenue by Age Group')
```

Out[29]:
```
<Axes: title={'center': 'Revenue by Age Group'}, xlabel='Age Group', ylabel='Total Purchase Amount'>
```



In [30]:
```
df.columns
```

Out[30]:
```
Index(['Customer ID', 'Purchase Date', 'Product Category', 'Product Price',
       'Quantity', 'Total Purchase Amount', 'Payment Method', 'Customer Age',
       'Returns', 'Customer Name', 'Age', 'Gender', 'Churn', 'Month',
       'Age Group'],
      dtype='object')
```

In [31]:
```
sns.set(style="whitegrid")
plt.rcParams['figure.figsize']=(10,6)
```

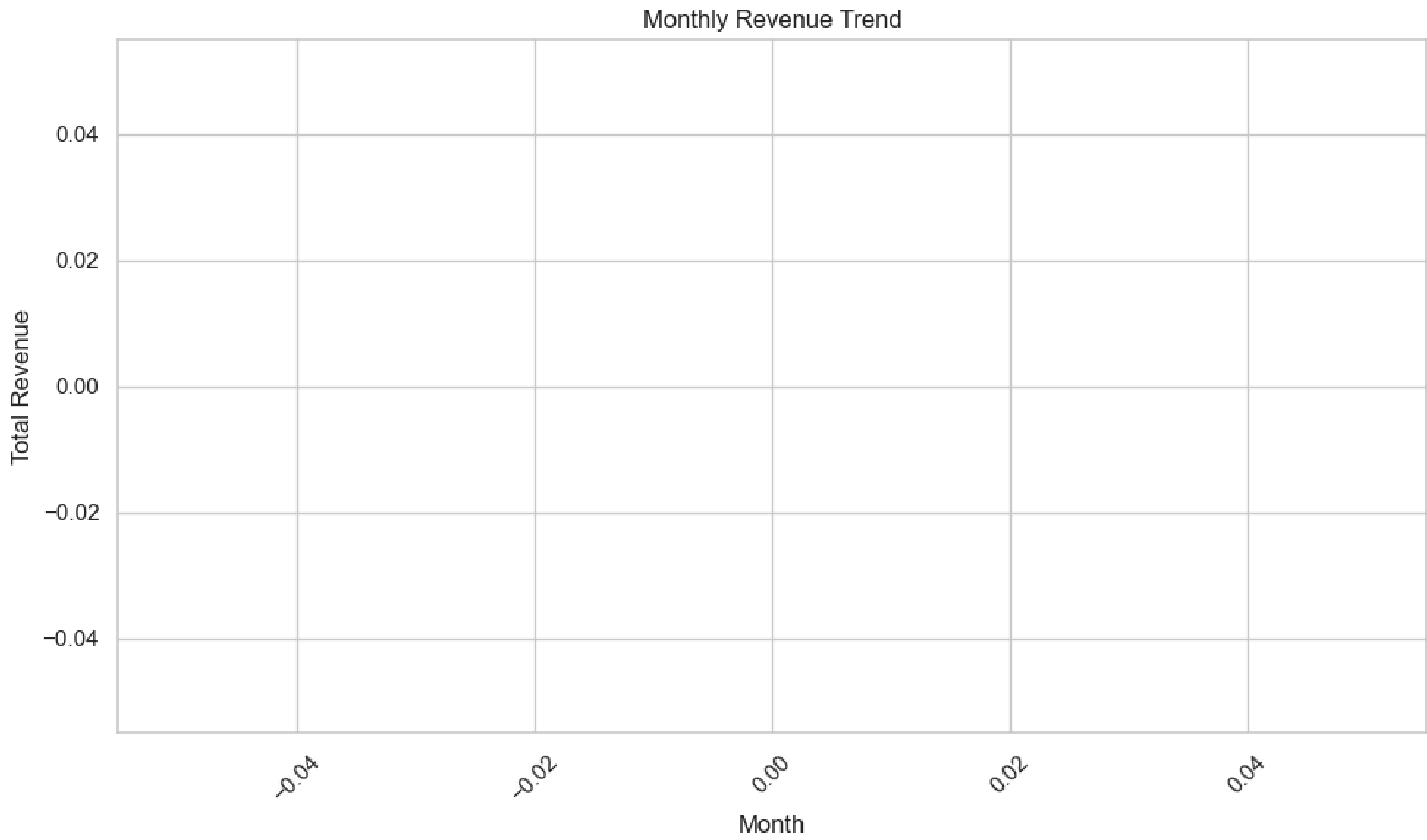
In [32]:
```
df['Purchase Date']=df['Purchase Date'].dt.to_period('M')
```

In [33]:
```
# Ensure Purchase Date is datetime
df['Purchase Date'] = pd.to_datetime(df['Purchase Date'], errors='coerce')

# Extract Month
df['Month'] = df['Purchase Date'].dt.to_period('M')

# Revenue per Month
monthly_revenue = df.groupby('Month')['Total Purchase Amount'].sum().reset_index()

# Plot
sns.lineplot(data=monthly_revenue, x='Month', y='Total Purchase Amount', marker='o')
plt.title('Monthly Revenue Trend')
plt.xticks(rotation=45)
plt.ylabel('Total Revenue')
plt.xlabel('Month')
plt.tight_layout()
plt.show()
```

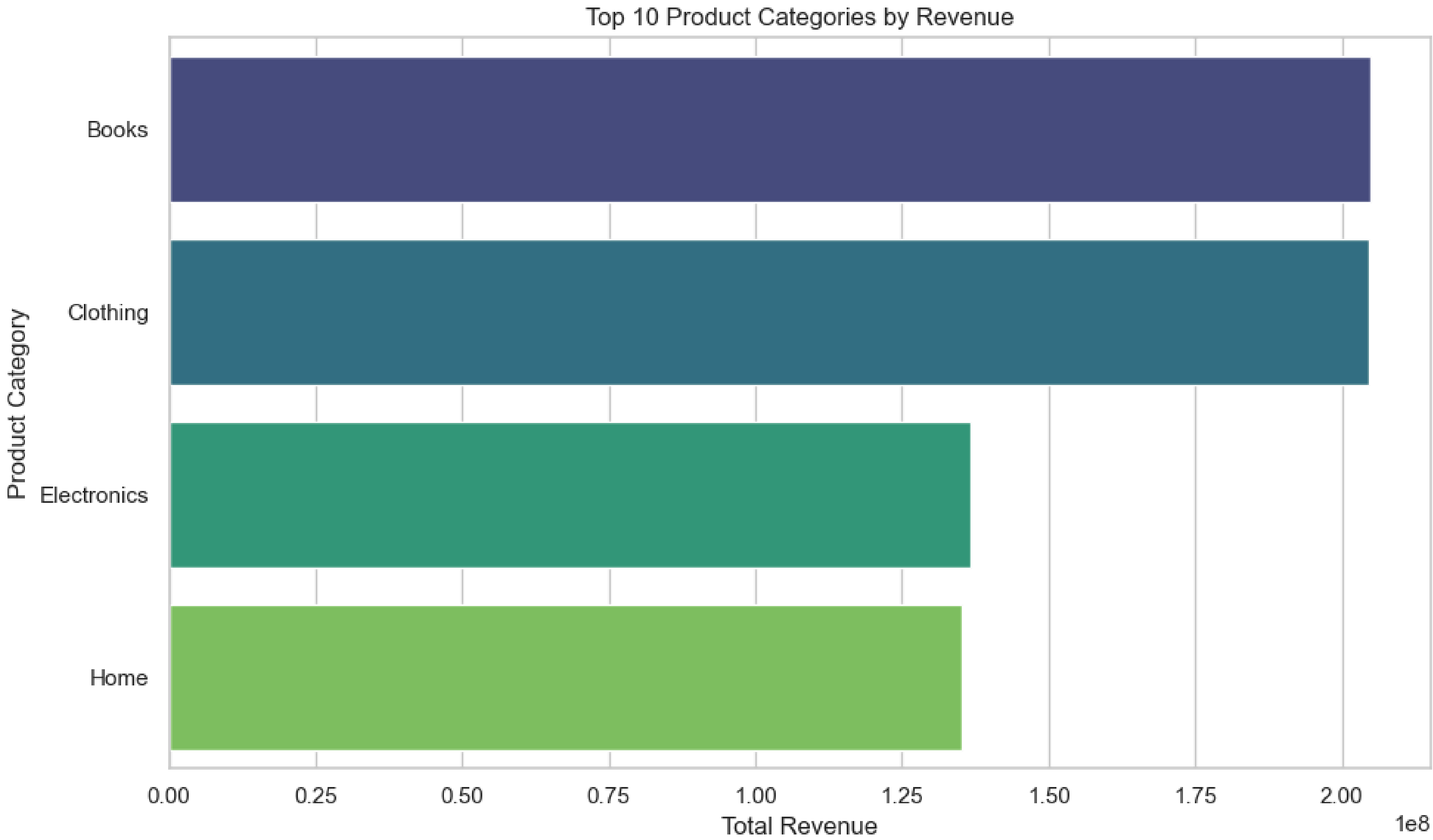


Top 10 Product Categories by Revenue

In [34]:
```
top_categories = df.groupby('Product Category')['Total Purchase Amount'].sum().sort_values(ascending=False).head(10)

# Plot
sns.barplot(x=top_categories.values, y=top_categories.index, palette='viridis')
```

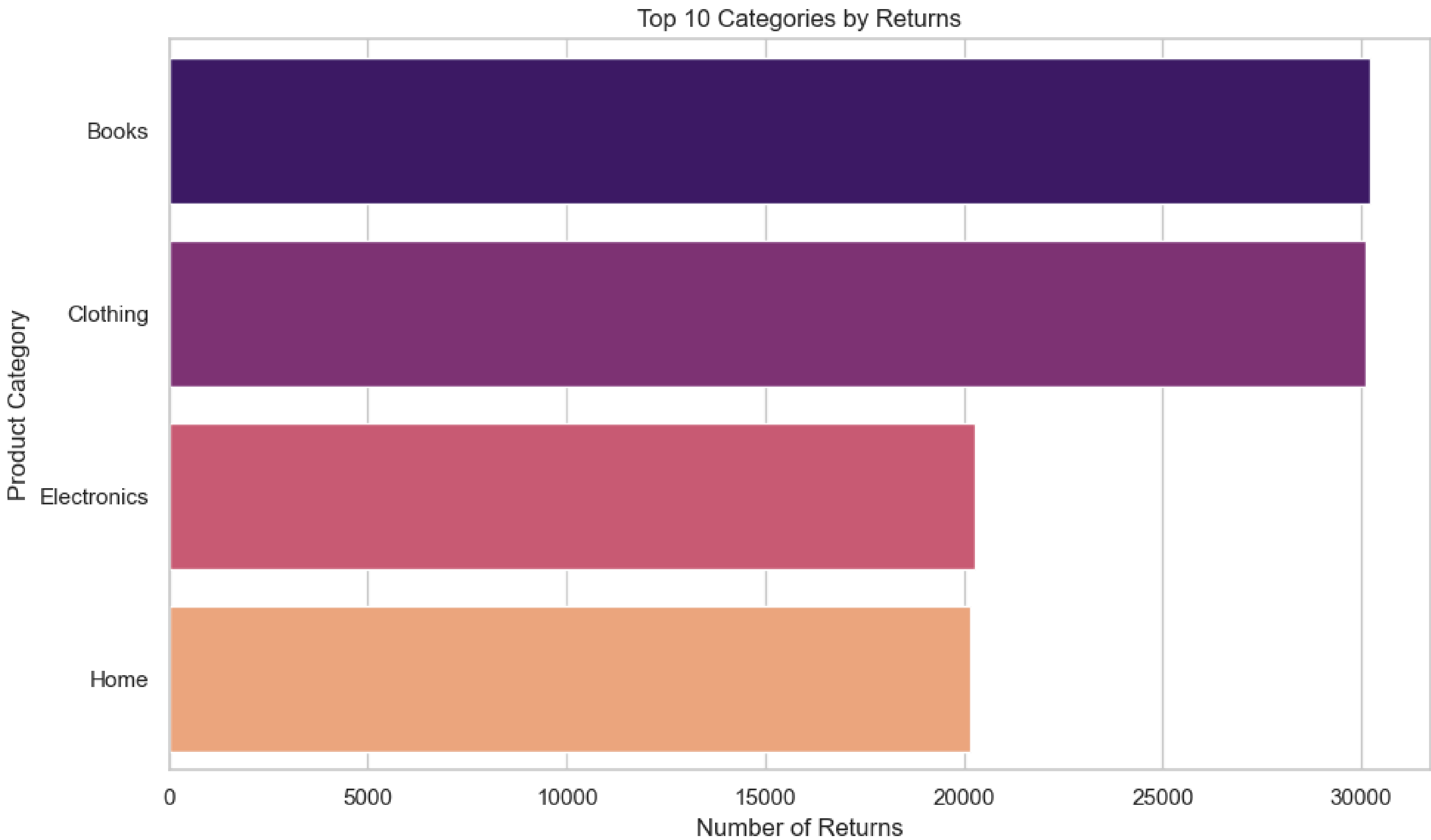
```
plt.title('Top 10 Product Categories by Revenue')
plt.xlabel('Total Revenue')
plt.tight_layout()
plt.show()
```



📦 Return Rates by Product Category

In [35]:
```
returns_by_category = df.groupby('Product Category')['Returns'].sum().sort_values(ascending=False).head(10)

sns.barplot(x=returns_by_category.values, y=returns_by_category.index, palette='magma')
plt.title('Top 10 Categories by Returns')
plt.xlabel('Number of Returns')
plt.tight_layout()
plt.show()
```

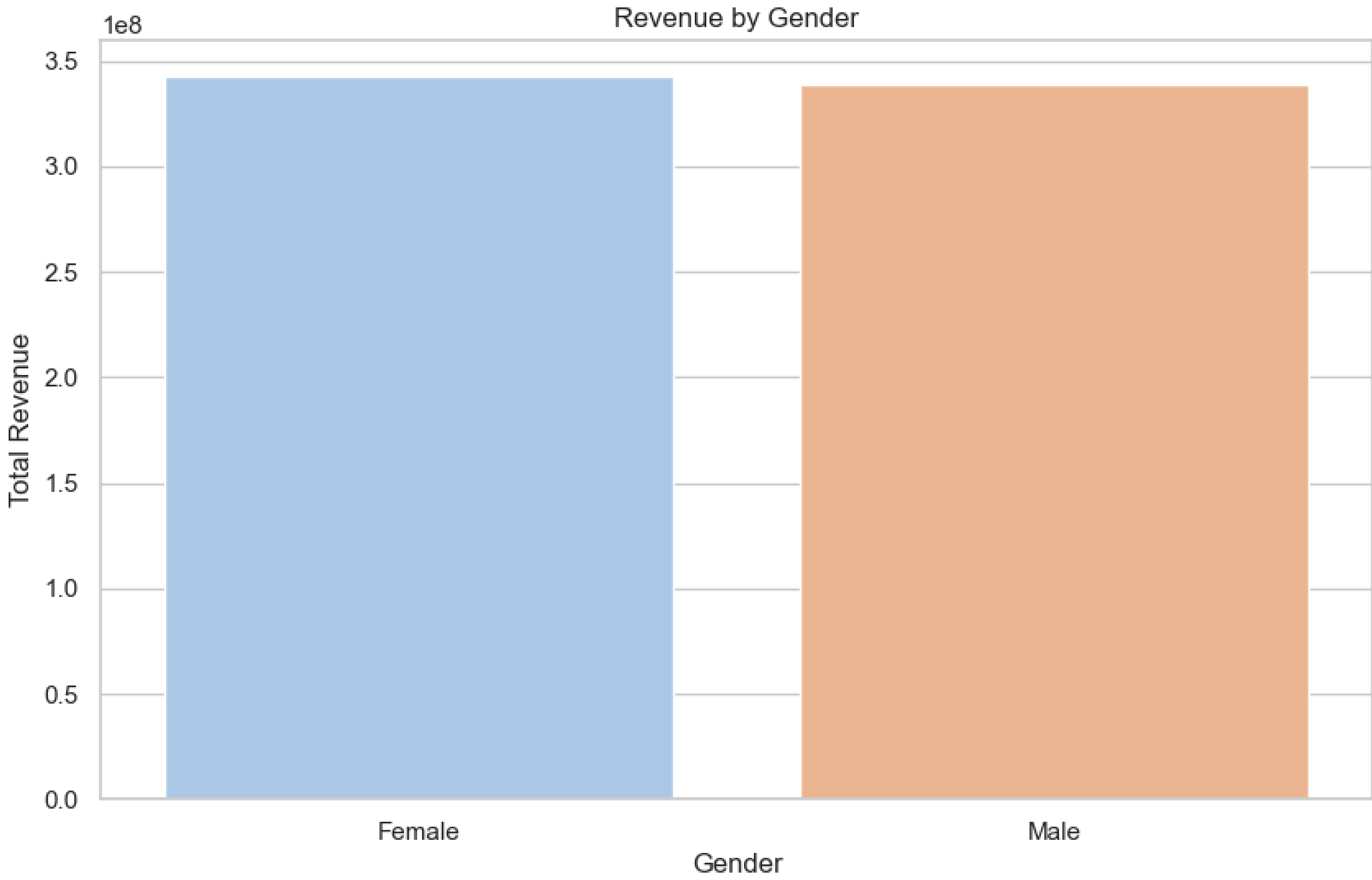


Demographic: Revenue by Gender

In [36]:

```
gender_revenue = df.groupby('Gender')['Total Purchase Amount'].sum().reset_index()

sns.barplot(data=gender_revenue, x='Gender', y='Total Purchase Amount', palette='pastel')
plt.title('Revenue by Gender')
plt.ylabel('Total Revenue')
plt.show()
```

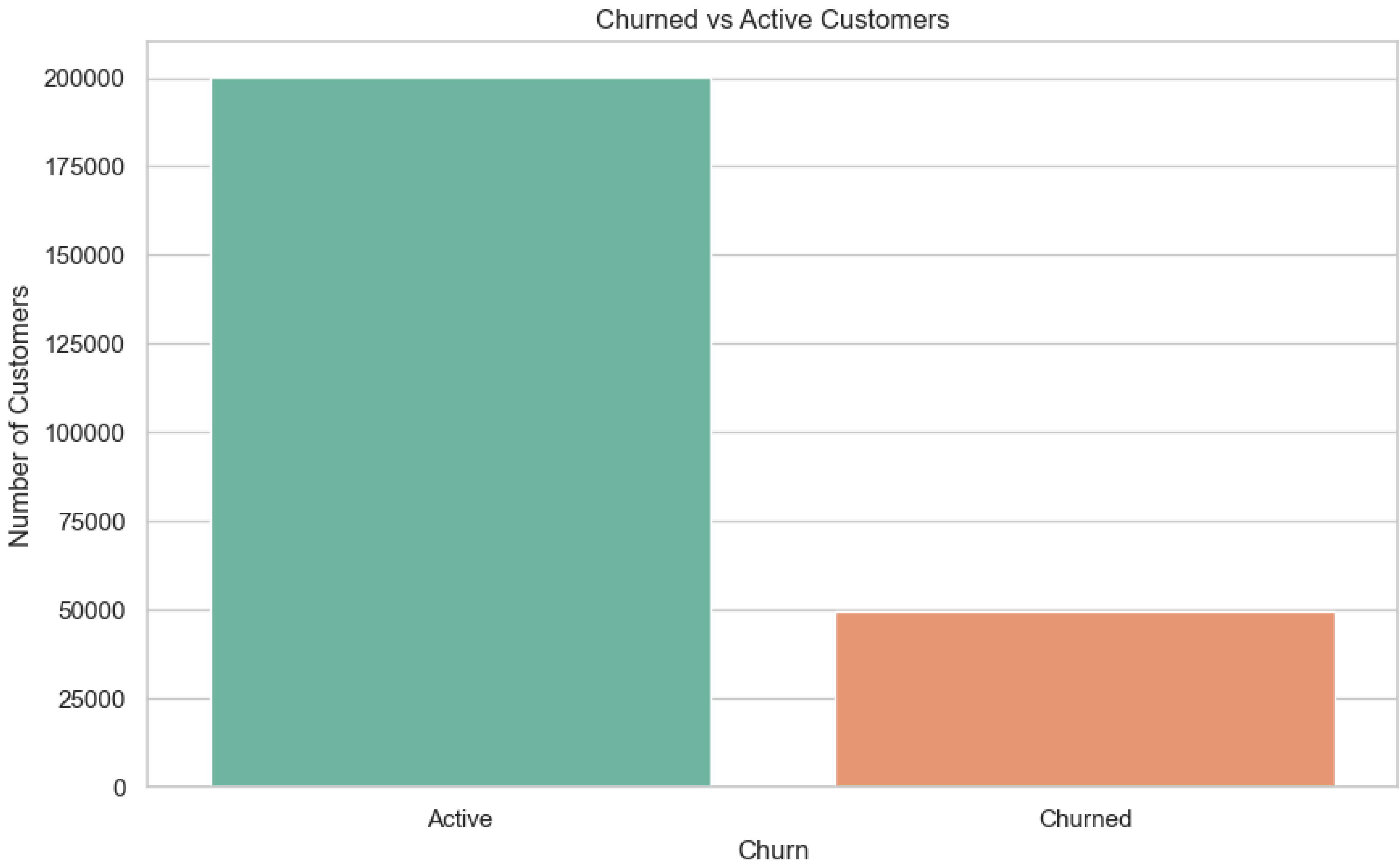


Churn vs Active Customers

In [37]:
```
churn_counts = df['Churn'].value_counts().rename({0: 'Active', 1: 'Churned'})

sns.barplot(x=churn_counts.index, y=churn_counts.values, palette='Set2')
plt.title('Churned vs Active Customers')
```

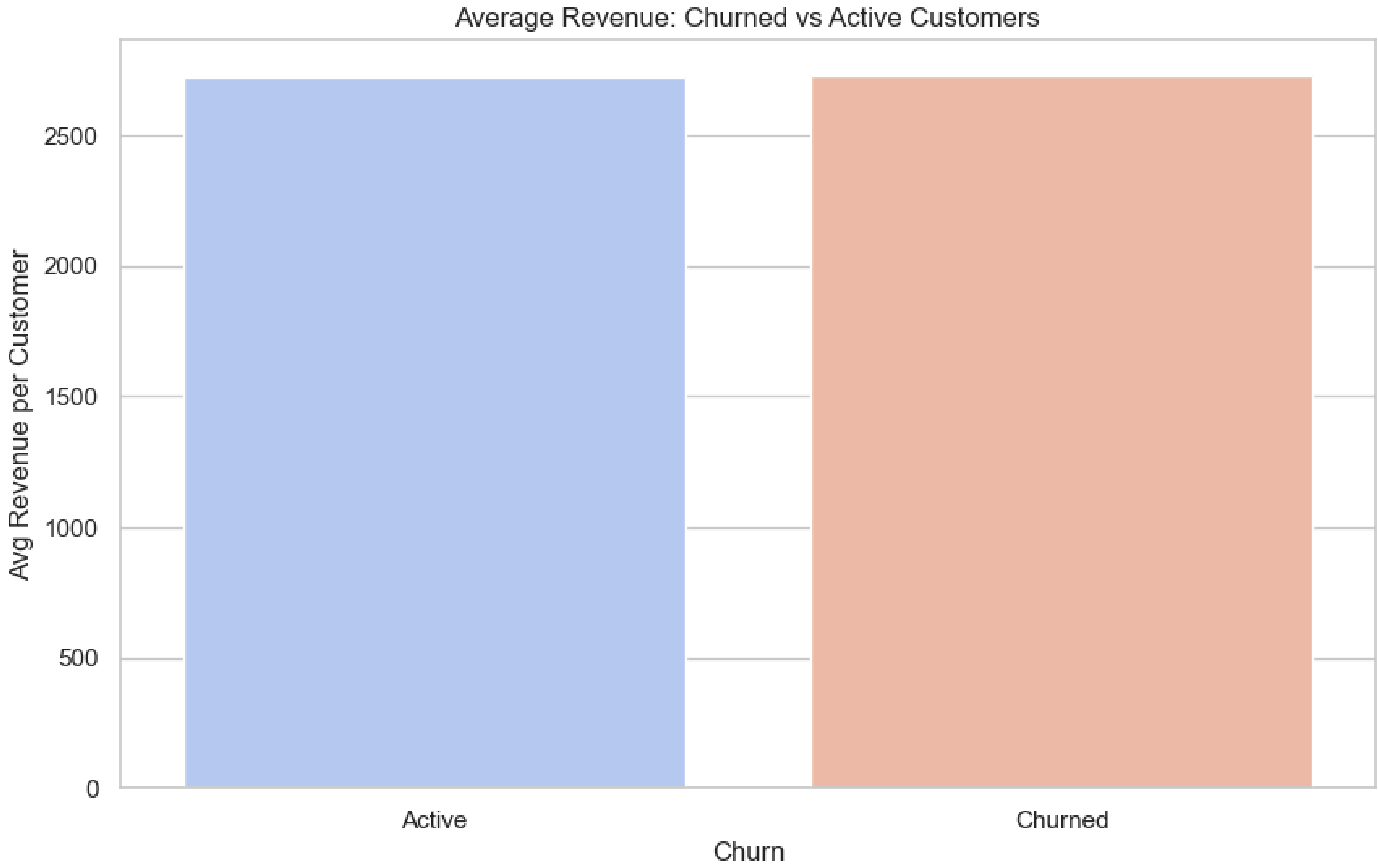
```
plt.ylabel('Number of Customers')
plt.show()
```



Churned vs Active: Revenue Comparison

In [38]:
```
rev_by_churn = df.groupby('Churn')['Total Purchase Amount'].mean().reset_index()
rev_by_churn['Churn'] = rev_by_churn['Churn'].map({0: 'Active', 1: 'Churned'})

sns.barplot(data=rev_by_churn, x='Churn', y='Total Purchase Amount', palette='coolwarm')
plt.title('Average Revenue: Churned vs Active Customers')
plt.ylabel('Avg Revenue per Customer')
plt.show()
```


Average Revenue: Churned vs Active Customers
Avg Revenue per Customer
0
500
1000
1500
2000
2500
Active
Churned
Churn

In [39]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 15 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   Customer ID            250000 non-null  int64
 1   Purchase Date          0 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                250000 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
 13  Month                  0 non-null       period[M]
 14  Age Group              250000 non-null  category
dtypes: category(1), datetime64[ns](1), float64(1), int64(7), object(4), period[M](1)
memory usage: 26.9+ MB
```

✅ Final Conclusion (Short Version)**

This project analyzed customer purchase behavior using e-commerce data. Key insights include:

- Revenue peaked seasonally, showing trends useful for sales planning.
- Top product categories generated most of the revenue—ideal for promotions.
- High return rates in certain categories suggest quality or listing issues.
- Age group 30–40 contributed the most revenue—prime target for marketing.
- Churned customers had lower spending and more returns, indicating the need for retention strategies.

Recommendation: Focus on top categories, reduce return rates, and run targeted campaigns for high-value customers to increase retention and sales.