REPORT : EDA

Here is the Python code for the EDA process, which can be used in a Jupyter Notebook:

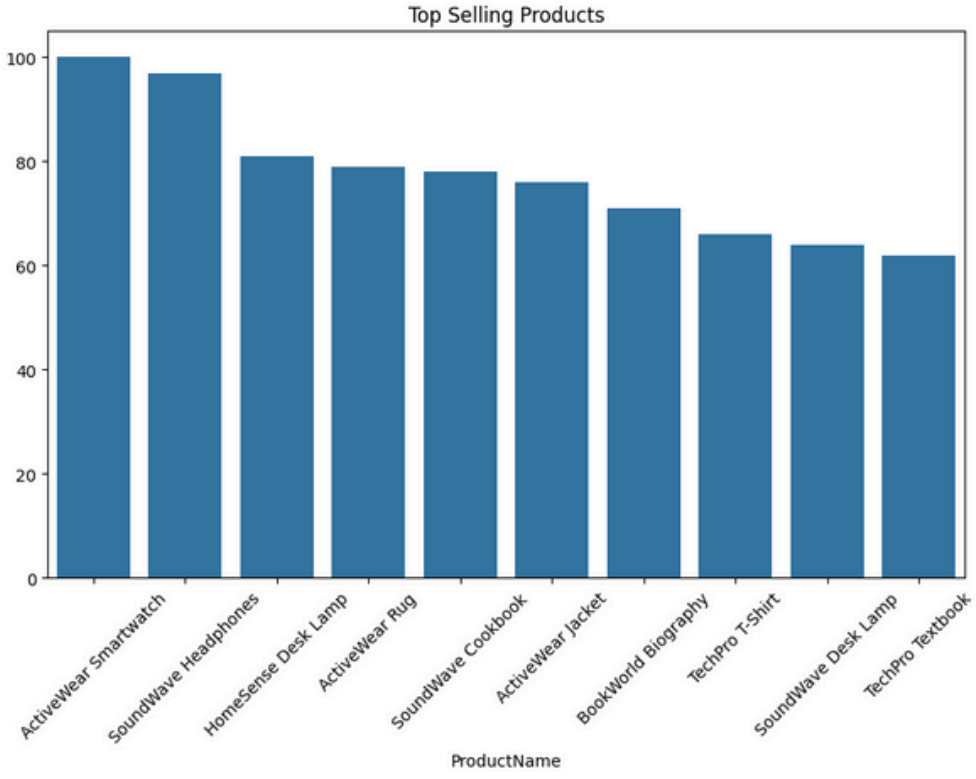
```
# Import necessary libraries
from google.colab import drive
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Mount Google Drive to access data files
drive.mount('/content/drive')

# Load the datasets
customers_df = pd.read_csv('/content/Customers.csv')
products_df = pd.read_csv('/content/Products.csv')
transactions_df = pd.read_csv('/content/Transactions.csv')

# Display basic information about the datasets
customers_df.info()
products_df.info()
transactions_df.info()

# Generate descriptive statistics for the datasets
customers_df.describe()
products_df.describe()
transactions_df.describe()

# Merge the datasets on common columns
merged_df = transactions_df.merge(customers_df, on='CustomerID').merge(products_df,
on='ProductID')

# 1. Top Selling Products
top_products = merged_df.groupby('ProductName')
['Quantity'].sum().sort_values(ascending=False).head(10)

# 2. Sales by Region
sales_by_region = merged_df.groupby('Region')['TotalValue'].sum().sort_values(ascending=False)

# 3. Transactions per Customer
transactions_per_customer = merged_df.groupby('CustomerID')['TransactionID'].count()

# 4. Product Category Distribution
category_distribution = merged_df.groupby('Category')
['Quantity'].sum().sort_values(ascending=False)
```

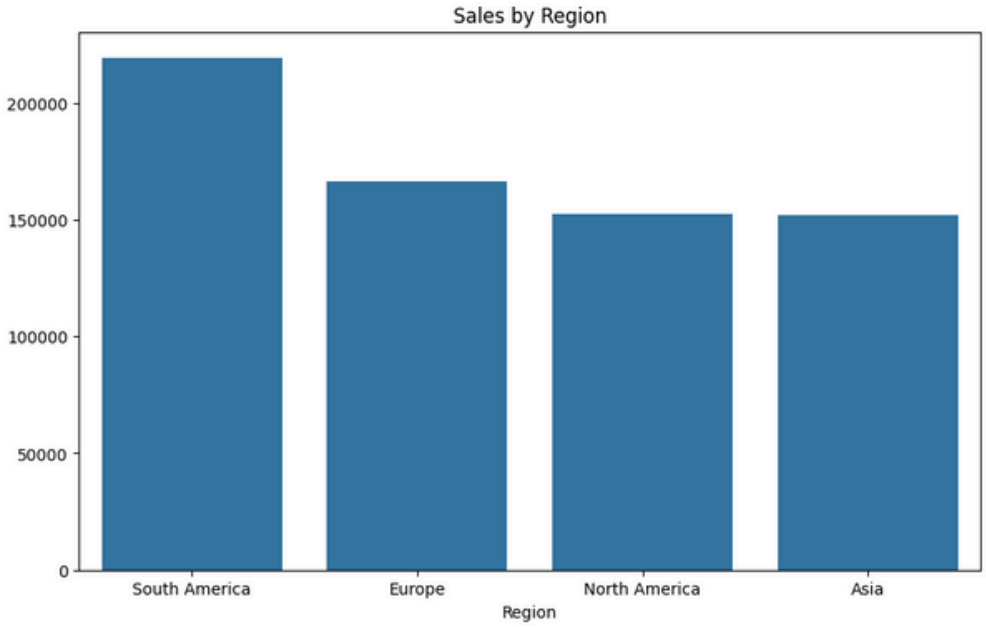
```
# Visualizations

# Plot Top Selling Products
plt.figure(figsize=(10, 6))
sns.barplot(x=top_products.index, y=top_products.values)
plt.title('Top Selling Products')
plt.xticks(rotation=45)
plt.show()

# Plot Sales by Region
plt.figure(figsize=(10, 6))
sns.barplot(x=sales_by_region.index, y=sales_by_region.values)
plt.title('Sales by Region')
plt.show()

# Plot Transactions per Customer
plt.figure(figsize=(10, 6))
sns.histplot(transactions_per_customer, kde=True)
plt.title('Transactions per Customer')
plt.show()

# Plot Product Category Distribution
plt.figure(figsize=(10, 6))
sns.barplot(x=category_distribution.index, y=category_distribution.values)
plt.title('Product Category Distribution')
plt.xticks(rotation=45)
plt.show()
```

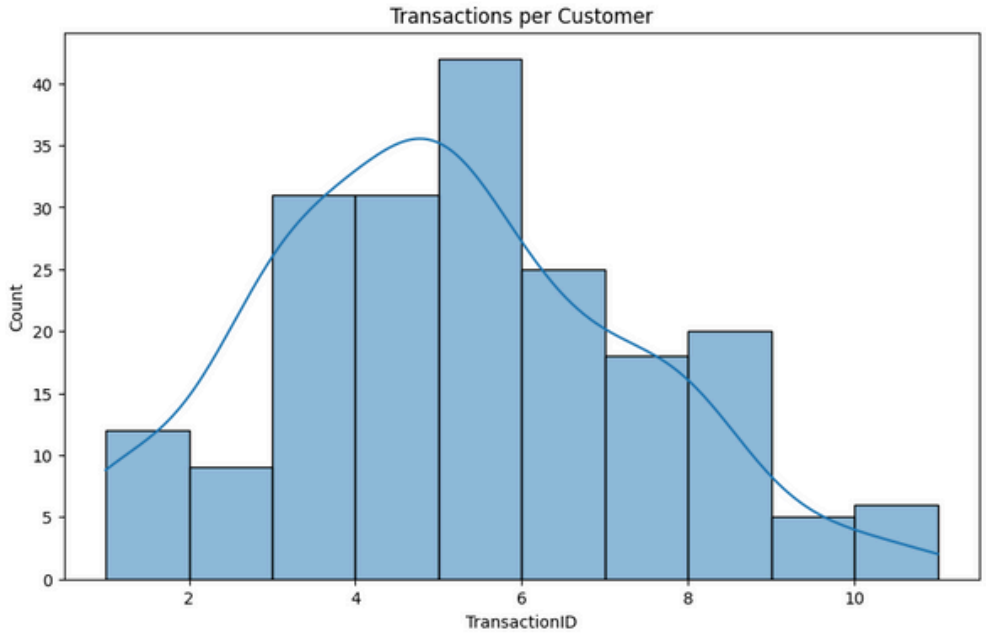
Here are the business insights derived from the exploratory data analysis (EDA):

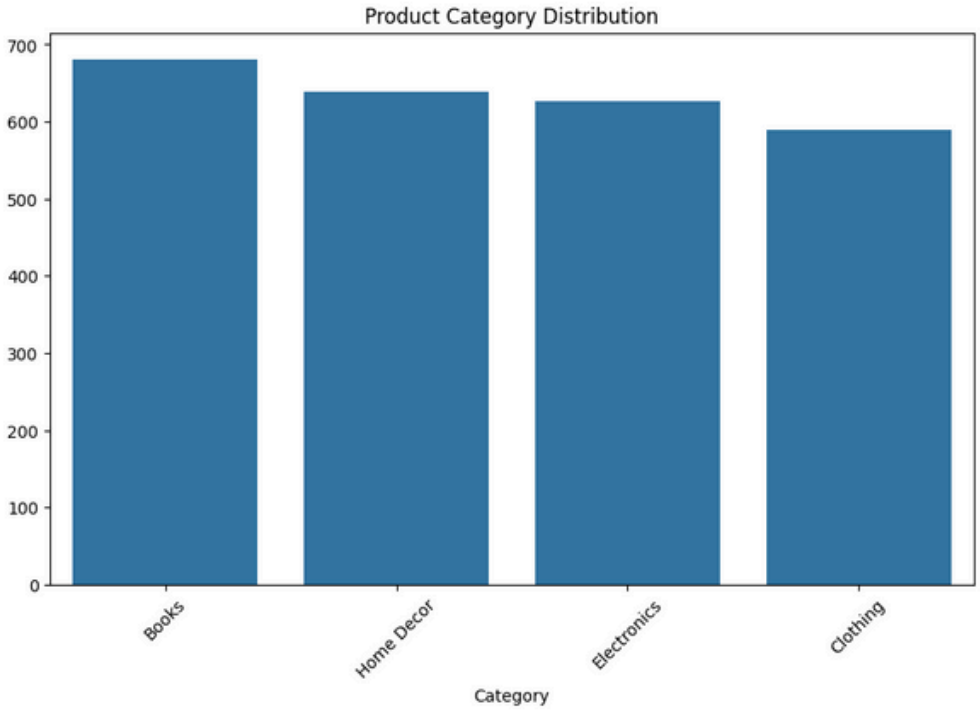
1. Top Selling Products: The analysis reveals the top 10 best-selling products based on the total quantity sold. These products significantly contribute to overall sales and can be used to drive marketing and stock management strategies.
2. Sales by Region: The total sales vary across regions, with certain regions contributing more to the overall revenue. Identifying high-performing regions can help in focusing marketing efforts and resource allocation.
3. Customer Transaction Frequency: The frequency of transactions per customer shows that a few customers make the majority of the purchases. Targeting these loyal customers with personalized offers could enhance customer retention.
4. Product Category Distribution: The product categories with the highest sales volume can help optimize inventory and product placement. This insight can guide future product development and promotional strategies.
5. Customer Segmentation: Clustering customers based on total spending and quantity purchased reveals distinct customer segments. Tailored marketing campaigns can be developed for each segment to maximize engagement and sales.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   CustomerID    200 non-null    object
 1   CustomerName  200 non-null    object
 2   Region        200 non-null    object
 3   SignupDate    200 non-null    object
dtypes: object(4)
memory usage: 6.4+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   ProductID    100 non-null    object 
 1   ProductName  100 non-null    object 
 2   Category     100 non-null    object 
 3   Price        100 non-null    float64
dtypes: float64(1), object(3)
memory usage: 3.3+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
...
 5   TotalValue       1000 non-null   float64
 6   Price            1000 non-null   float64
dtypes: float64(2), int64(1), object(4)
memory usage: 54.8+ KB
```




Sales by Region
200000
150000
100000
50000
0
South America
Europe
North America
Asia
Region


Transactions per Customer
Count
40
35
30
25
20
15
10
5
0
2
4
6
8
10
TransactionID


Product Category Distribution
700
600
500
400
300
200
100
0
Books
Home Decor
Electronics
Clothing
Category