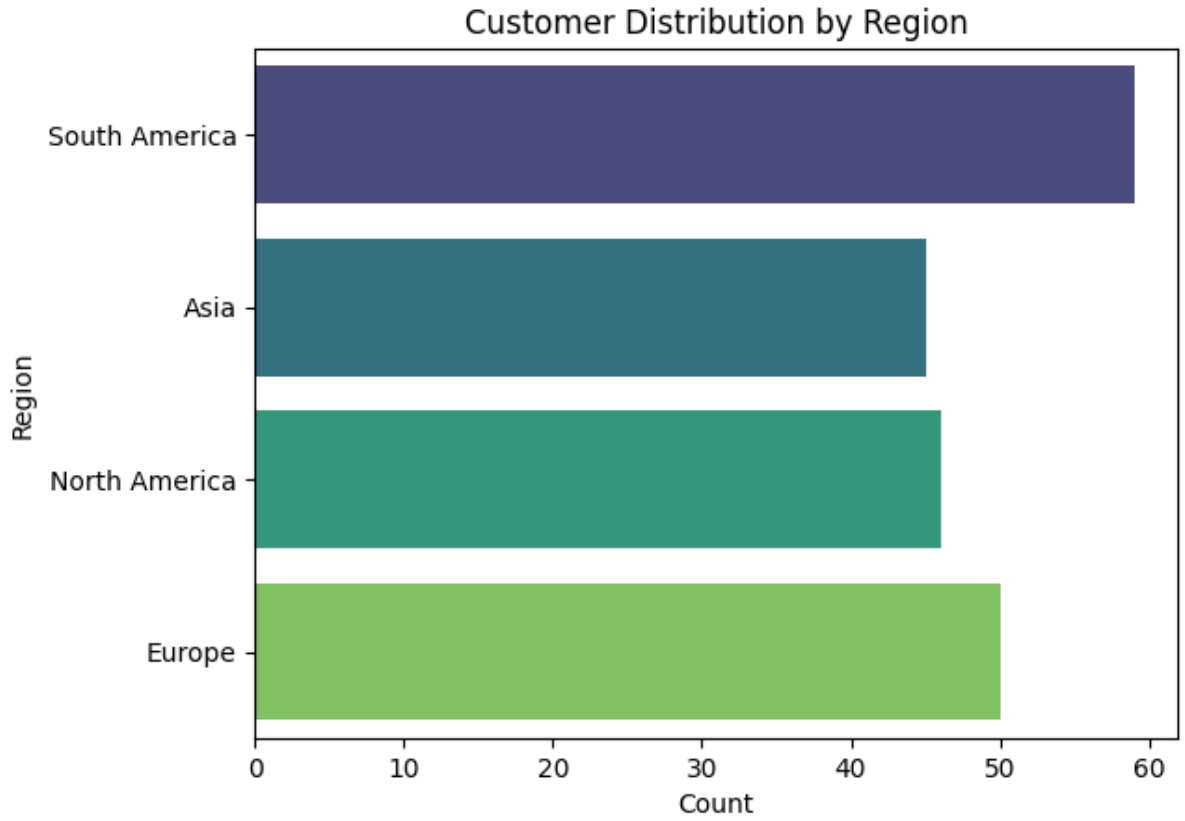
Task 1: Exploratory Data Analysis (EDA)

#load csv files into a Jupyter Notebook or Python environment.

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

# Load datasets
customers = pd.read_csv("Customers.csv")
products = pd.read_csv("Products.csv")
transactions = pd.read_csv("Transactions.csv")

# Display first few rows
print("Customers:")
print(customers.head())
print("\nProducts:")
print(products.head())
print("\nTransactions:")
print(transactions.head())

#Check for missing values, duplicates, and inconsistent formats.
# Check for missing values
print("Missing values in Customers:")
print(customers.isnull().sum())
print("\nMissing values in Products:")
print(products.isnull().sum())
print("\nMissing values in Transactions:")
print(transactions.isnull().sum())

# Check for duplicates
print("\nDuplicates in each dataset:")
print("Customers:", customers.duplicated().sum())
print("Products:", products.duplicated().sum())
print("Transactions:", transactions.duplicated().sum())

# Convert date columns to datetime
customers['SignupDate'] = pd.to_datetime(customers['SignupDate'])
transactions['TransactionDate'] = pd.to_datetime(transactions['TransactionDate'])

# Ensure Price consistency (Products and Transactions)
print("\nUnique Prices in Products:")
print(products['Price'].unique())
print("\nUnique Prices in Transactions:")
print(transactions['Price'].unique())
```

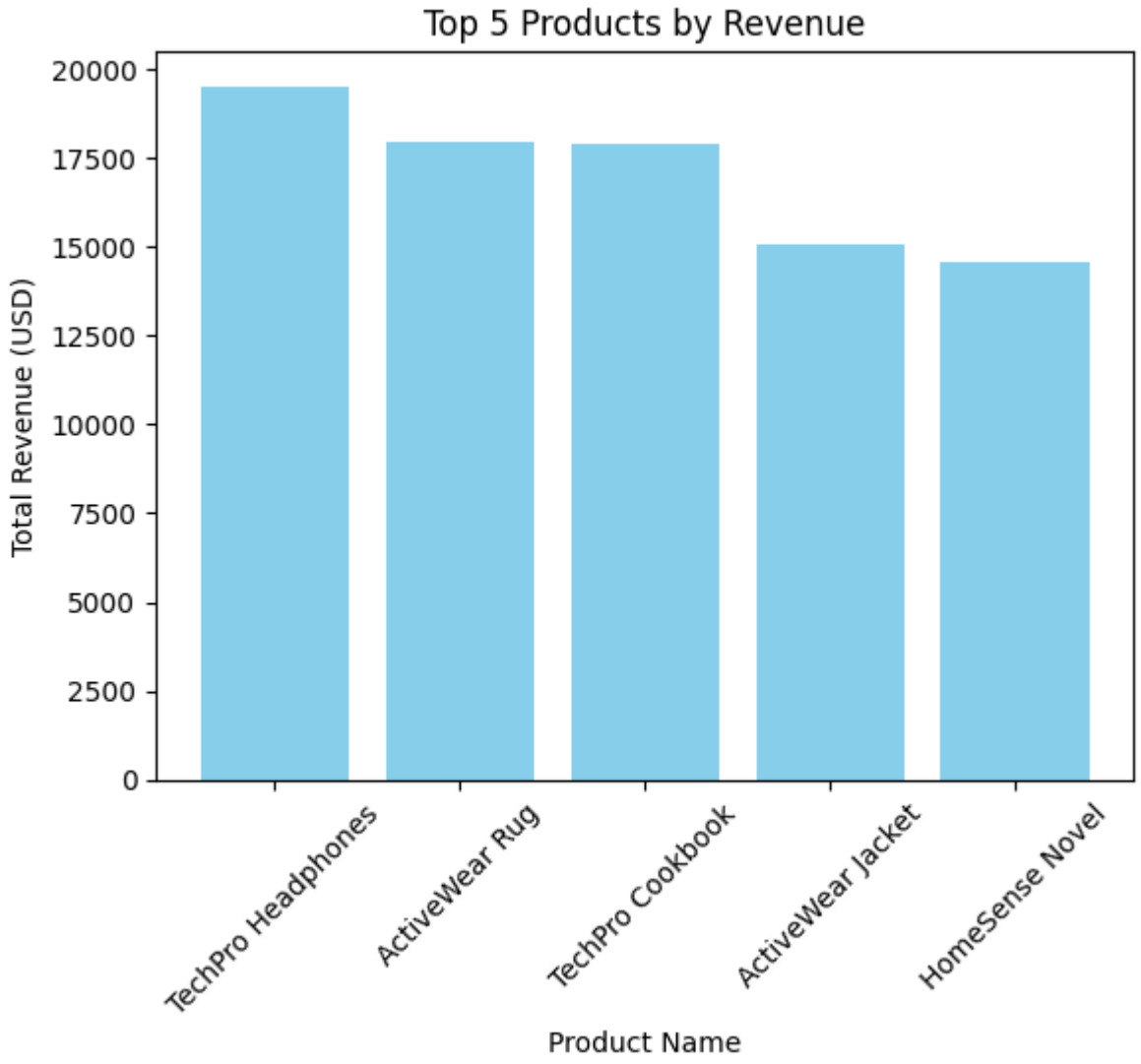
#Exploratory Visualizations
Customer distribution by region:

```
sns.countplot(y='Region', data=customers, palette='viridis')
plt.title("Customer Distribution by Region")
plt.xlabel("Count")
plt.ylabel("Region")
plt.show()
```



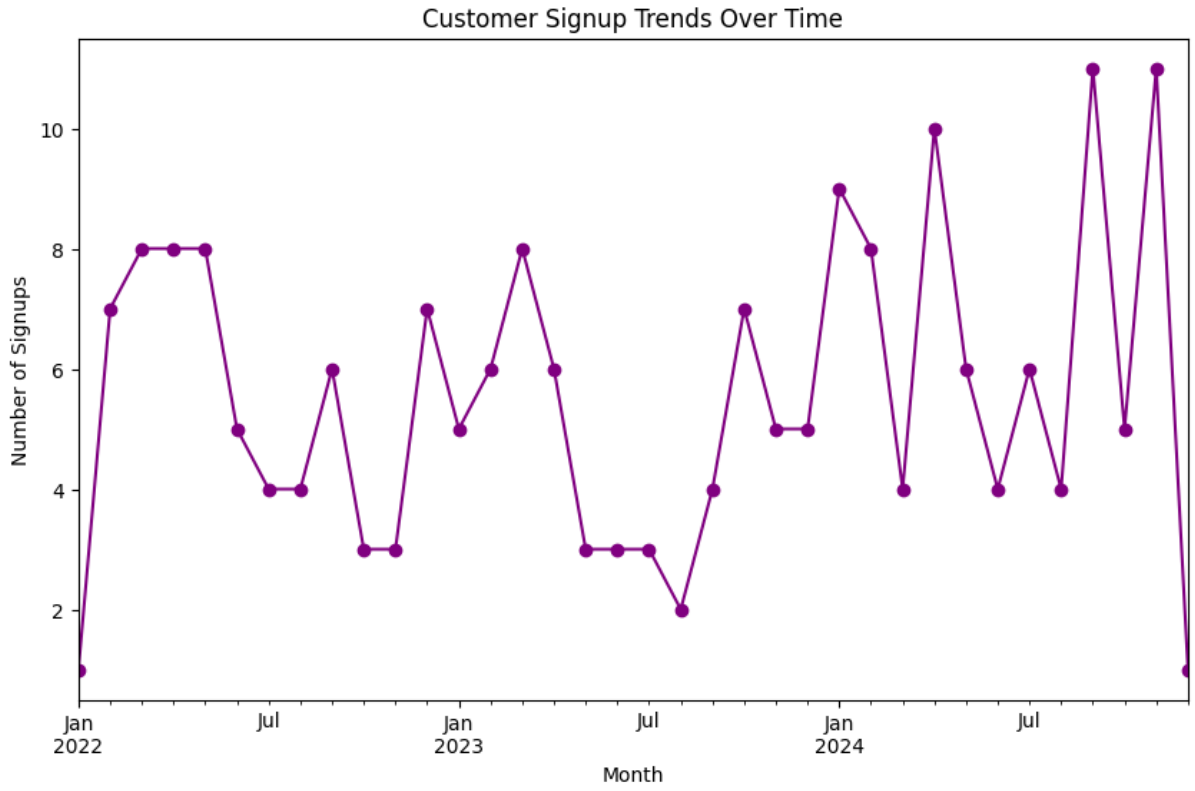
Top 5 products by revenue:

```
top_products = transactions.groupby('ProductID')['TotalValue'].sum().nlargest(5)
top_products = top_products.reset_index().merge(products, on='ProductID')

plt.bar(top_products['ProductName'], top_products['TotalValue'], color='skyblue')
plt.title("Top 5 Products by Revenue")
plt.xlabel("Product Name")
plt.ylabel("Total Revenue (USD)")
plt.xticks(rotation=45)
plt.show()
```



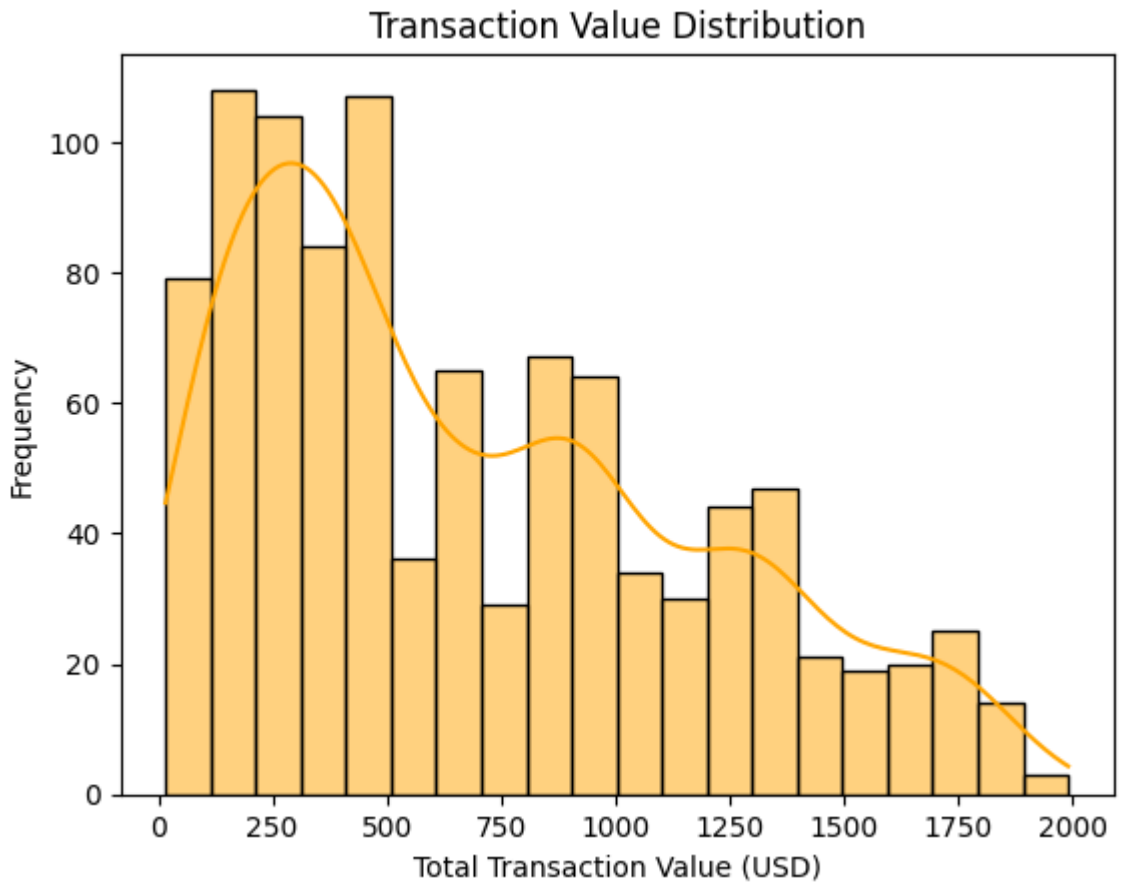
Signup trends over time:

```
signup_trends = customers['SignupDate'].dt.to_period('M').value_counts().sort_index()
signup_trends.plot(kind='line', marker='o', figsize=(10, 6), color='purple')
plt.title("Customer Signup Trends Over Time")
plt.xlabel("Month")
plt.ylabel("Number of Signups")
plt.show()
```



Transaction value distribution:

```
sns.histplot(transactions['TotalValue'], bins=20, kde=True, color='orange')
plt.title("Transaction Value Distribution")
plt.xlabel("Total Transaction Value (USD)")
plt.ylabel("Frequency")
plt.show()
```



Exploratory Data Analysis Insights

Customer Distribution by Region:

South America has the highest number of customers, followed closely by Europe. This suggests that these regions are the primary markets for the business.

Top Revenue-Generating Products:

The "TechPro Headphone" is the most sold product, generating approximately $18,500 in revenue. The "Activewear Rug" and "TechPro Cookbook" are tied for second place, each contributing around $17,500 to total revenue.

Signup Trends:

The lowest customer signups occurred in July of both 2022 and 2023. However, a significant surge in signups was recorded in 2024, marking it as the highest signup year.

Transaction Value Analysis:

Minimum Transaction Value: $2,000.

Most Frequent Transaction Range: $100–$500, with a frequency exceeding 100 transactions.

Average Transaction Value: Approximately $750–$800, with an average transaction frequency of 50–60 units.