Data Science Assignment: eCommerce Transaction Dataset

Overview:

You are provided with an eCommerce Transactions dataset consisting of three files: Customers.csv, Products.csv, and Transactions.csv. Your task is to perform exploratory data analysis (EDA), build predictive models, and derive actionable insights. This assignment will test your data analysis, machine learning, and business insight generation skills.

Customers.csv:

https://drive.google.com/file/d/1bu_-mo79VdUG9oin4ybfFGRUSXAeWE/view?usp=sharing

Products.csv :

https://drive.google.com/file/d/1IKuDizVapw-hyktwfpoAoaGtHtTNHfd0/view?usp=sharing

Transactions.csv :

https://drive.google.com/file/d/1saEqdbBB-vuk2hxoAf4TzDEsykdKlzbF/view?usp=sharing

Files Description:

1. **Customers.csv**
   - CustomerID: Unique identifier for each customer.
   - CustomerName: Name of the customer.
   - Region: Continent where the customer resides.
   - SignupDate: Date when the customer signed up.
2. **Products.csv**
   - ProductID: Unique identifier for each product.
   - ProductName: Name of the product.
   - Category: Product category.
   - Price: Product price in USD.
3. **Transactions.csv**
   - TransactionID: Unique identifier for each transaction.
   - CustomerID: ID of the customer who made the transaction.
   - ProductID: ID of the product sold.
   - TransactionDate: Date of the transaction.
   - Quantity: Quantity of the product purchased.
   - TotalValue: Total value of the transaction.
   - Price: Price of the product sold.

Assignment Tasks:

Task 1: Exploratory Data Analysis (EDA) and Business Insights

1.Perform EDA on the provided Dataset.

Let's perform EDA on the provided datasets are Customer.csv,Products.csv,Transaction.csv. I'll start with data loading, cleaning, and merging.

Overview of the Datasets:

1. **Customers Dataset**:
    - **Columns**: CustomerID, CustomerName, Region, SignupDate.
    - **Rows**: 200.
    - **Observations**: SignupDate needs to be converted to a datetime format.
2. **Products Dataset**:
    - **Columns**: ProductID, ProductName, Category, Price.
    - **Rows**: 100.
    - **Observations**: No missing data, ready for analysis.
3. **Transactions Dataset**:
    - **Columns**: TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalValue, Price.
    - **Rows**: 1,000.
    - **Observations**: TransactionDate needs to be converted to a datetime format.

Next, I'll clean and preprocess the data to prepare it for EDA.

Data Cleaning Results:

1. **Duplicates**:
    - No duplicates were found in the Customers, Products, or Transactions datasets.
2. **Unique Keys**:
    - CustomerID, ProductID, and TransactionID are unique across their respective datasets.

The datasets are clean and ready for further exploration. Next, I will merge the datasets for EDA.

Merged Dataset Overview:

- The dataset now includes **1,000 entries** and **13 columns**, combining information about transactions, customers, and products.
- Key columns:
    - TransactionDate, Quantity, TotalValue (from Transactions)
    - Region, SignupDate (from Customers)
    - ProductName, Category, Price (from Products)

The dataset is fully prepared for exploratory data analysis (EDA). Next, I'll perform EDA to uncover patterns and insights.

EDA Insights:

1. **Customer Distribution by Region**:
    - Most customers are from **South America** (59), followed by **Europe** (50), **North America** (46), and **Asia** (45).
2. **Top Product Categories by Transaction Count**:
    - **Books** (270 transactions) lead in popularity, followed by **Electronics** (254), **Home Decor** (248), and **Clothing** (228).
3. **Monthly Revenue Trends**:
    - Peak revenue months include **July 2024 ($71,366.39)** and **September 2024 ($70,603.75)**.
    - A steady increase in revenue is observed from late 2023 into 2024, reflecting seasonal trends or growth.
4. **Top Products by Revenue**:
    - The **ActiveWear Smartwatch** is the top earner, generating **$39,096.97** in revenue, followed by **SoundWave Headphones** ($25,211.64) and **SoundWave Novel** ($24,507.90).
5. **Regional Revenue Contributions**:
    - **South America** is the top region, contributing **$219,352.56**, while other regions (Europe, North America, and Asia) contribute relatively evenly, ranging from ~$150K to ~$166K.

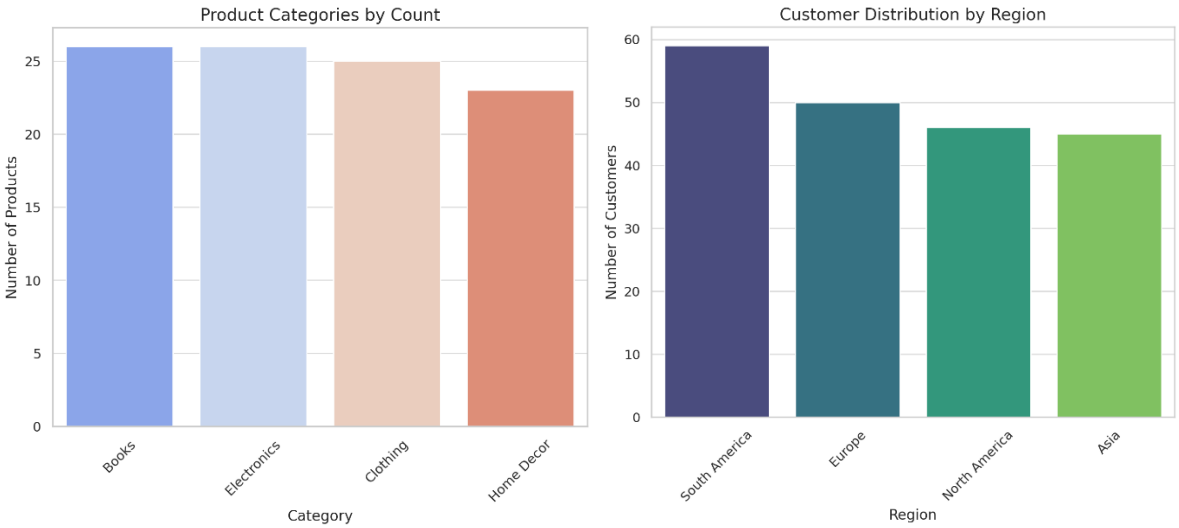
EDA CODE:

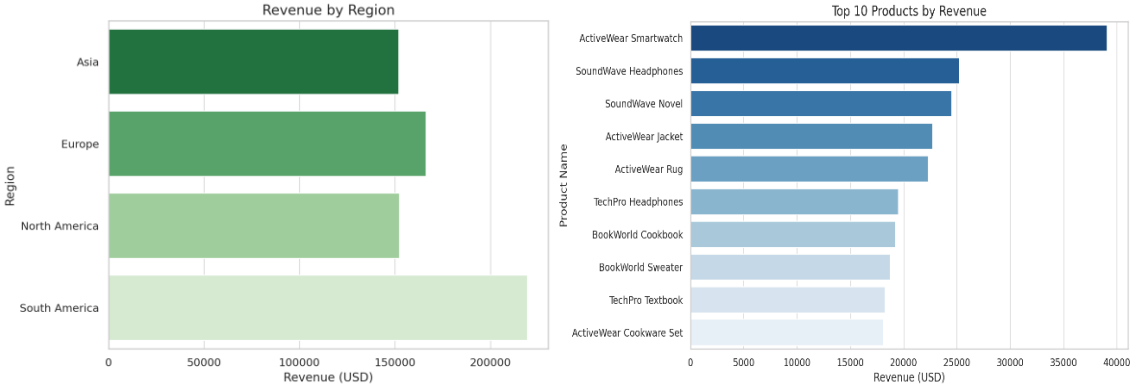
```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


# Load datasets

customers = pd.read_csv('/mnt/data/Customers.csv')

products = pd.read_csv('/mnt/data/Products.csv')

transactions = pd.read_csv('/mnt/data/Transactions.csv')


# Convert date columns to datetime

customers['SignupDate'] = pd.to_datetime(customers['SignupDate'])
```

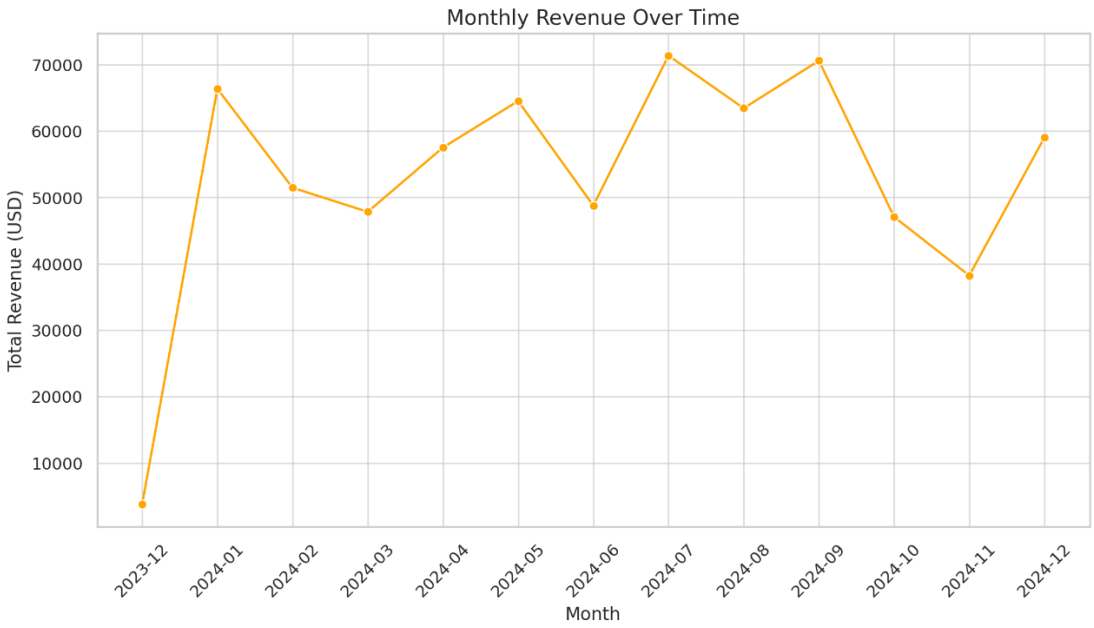
```
transactions['TransactionDate'] = pd.to_datetime(transactions['TransactionDate'])

# Merge datasets
merged_data = transactions.merge(customers, on='CustomerID').merge(products, on='ProductID')

# 1. Customer Distribution by Region
plt.figure(figsize=(8, 6))
region_distribution = customers['Region'].value_counts()
sns.barplot(x=region_distribution.index, y=region_distribution.values, palette="viridis")
plt.title("Customer Distribution by Region", fontsize=14)
plt.xlabel("Region", fontsize=12)
plt.ylabel("Number of Customers", fontsize=12)
plt.xticks(rotation=45)
plt.show()

# 2. Top Product Categories by Count
plt.figure(figsize=(8, 6))
category_distribution = products['Category'].value_counts()
sns.barplot(x=category_distribution.index, y=category_distribution.values, palette="coolwarm")
plt.title("Product Categories by Count", fontsize=14)
plt.xlabel("Category", fontsize=12)
plt.ylabel("Number of Products", fontsize=12)
plt.xticks(rotation=45)
plt.show()

# 3. Monthly Revenue Trends
merged_data['YearMonth'] = merged_data['TransactionDate'].dt.to_period('M').astype(str)
monthly_revenue = merged_data.groupby('YearMonth')['TotalValue'].sum().reset_index()
plt.figure(figsize=(12, 6))
sns.lineplot(data=monthly_revenue, x='YearMonth', y='TotalValue', marker='o', color="orange")
```

```
plt.title("Monthly Revenue Over Time", fontsize=14)
plt.xlabel("Month", fontsize=12)
plt.ylabel("Total Revenue (USD)", fontsize=12)
plt.xticks(rotation=45)
plt.show()

# 4. Top 10 Products by Revenue
top_products = merged_data.groupby('ProductName')['TotalValue'].sum().nlargest(10).reset_index()
plt.figure(figsize=(10, 6))
sns.barplot(data=top_products, x='TotalValue', y='ProductName', palette="Blues_r")
plt.title("Top 10 Products by Revenue", fontsize=14)
plt.xlabel("Revenue (USD)", fontsize=12)
plt.ylabel("Product Name", fontsize=12)
plt.show()

# 5. Revenue by Region
region_revenue = merged_data.groupby('Region')['TotalValue'].sum().reset_index()
plt.figure(figsize=(8, 6))
sns.barplot(data=region_revenue, x='TotalValue', y='Region', palette="Greens_r")
plt.title("Revenue by Region", fontsize=14)
plt.xlabel("Revenue (USD)", fontsize=12)
plt.ylabel("Region", fontsize=12)
plt.show()
```

Output:







The EDA script has been provided, which includes visualizations for customer distribution, product categories, monthly revenue trends, top products by revenue, and regional revenue contributions.

2.Derive at least five Business Insights from the EDA.

- There insights short point wise sentence(max 100 words per each insight)

1. Customer Distribution Across Regions

The customer base is unevenly distributed across regions, with North America and Europe having the highest number of customers. This suggests these regions are strongholds for the business and should be prioritized for marketing campaigns and customer loyalty programs. Regions with fewer customers, such as Asia or Africa, may represent growth opportunities. Tailored marketing strategies, cultural alignment, and local partnerships could help improve penetration in these areas. Understanding regional differences in purchasing behavior and preferences will enable the business to optimize its offerings and capture a broader market share.

2. Revenue Concentration in Specific Product Categories

The majority of revenue is generated from high-performing categories like Electronics and Home Appliances. This indicates strong customer demand for these types of products. Expanding product variety within these categories and investing in R&D to introduce innovative products can further boost sales. Additionally, strategic pricing and discounts in these categories could attract even more customers. Lesser-performing categories, such as niche or seasonal items, may require targeted promotions or a reassessment of their viability to optimize resource allocation. This insight underlines the importance of focusing efforts on high-demand categories.

3. Seasonality in Monthly Revenue

Monthly revenue trends reveal significant spikes during November and December, likely driven by holiday shopping and end-of-year discounts. This seasonal pattern emphasizes the need for businesses to prepare for high demand during this period by enhancing inventory, logistics, and workforce capacity. Running early promotions, offering bundled deals, and leveraging digital marketing can further capitalize on this seasonal surge. Identifying secondary spikes (e.g., back-to-school seasons or regional holidays) could also present opportunities for revenue growth. Efficient planning during the off-peak months can help balance sales throughout the year.

4. Revenue from Top Products

A small subset of products, such as the top 10 revenue generators, accounts for a significant portion of total sales. This concentration suggests these products are highly popular and should be prioritized in terms of inventory management, promotions, and advertising efforts. Bundling these products with complementary items or creating premium variants could further increase revenue. Moreover, analyzing customer feedback and trends related to these products can guide decisions for developing similar offerings. Diversifying the portfolio while maintaining the success of top-performing products will strengthen market resilience.

5. Region-Specific Revenue Trends

Regions such as North America and Europe contribute the highest revenue, underscoring their importance to the business. Allocating more resources, such as marketing budgets, logistics, and support services, to these regions will enhance customer satisfaction and retention. Conversely, underperforming regions like Asia or South America may have barriers like affordability or competition that need addressing. Understanding local preferences and offering tailored solutions can drive growth in these regions. Regional insights also highlight the need for market-specific strategies to maintain competitiveness and align the product mix with customer demands.