Introduction

Exploratory Data Analysis (EDA) is a critical step in the data analysis pipeline. Its purpose is to understand the underlying patterns, trends, and relationships within the datasets, while identifying any anomalies, missing values, or duplicates. For this analysis, we have utilized three datasets:

- **Customer Information:** Contains demographic and unique customer-related data.
- **Product Information:** Details of products available, including categories and pricing.
- **Transaction History:** Records of purchases made by customers, linking them to specific products.

By combining these datasets, we aim to uncover actionable insights related to customer behavior, product performance, and overall business trends.

Code

```
# Import Libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

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

# Check for Missing Values
print(customers.isnull().sum())
print(products.isnull().sum())
print(transactions.isnull().sum())

# Check for Duplicates
print(customers.duplicated().sum())
print(products.duplicated().sum())
print(transactions.duplicated().sum())
```

output:

```
0

0

0
```

```
# Basic Statistics

print(customers.describe())

print(products.describe())

print(transactions.describe())
```

output:

```
CustomerID     CustomerName        Region  SignupDate

count        200             200           200        200

unique       200             200             4        179

top        C0001  Lawrence Carroll  South America  2024-11-11

freq           1               1            59          3

           Price

count  100.000000

mean   267.551700

std    143.219383

min     16.080000

25%    147.767500

50%    292.875000

75%    397.090000

max    497.760000

                  TransactionDate     Quantity   TotalValue      Price

count                        1000  1000.000000  1000.000000  1000.00000
```

```
mean   2024-06-23 15:33:02.768999936    2.537000   689.995560   272.55407
min            2023-12-30 15:29:12    1.000000    16.080000    16.08000
25%      2024-03-25 22:05:34.500000    2.000000   295.295000   147.95000
50%      2024-06-26 17:21:52.500000    3.000000   588.880000   299.93000
75%            2024-09-19 14:19:57    4.000000  1011.660000   404.40000
max            2024-12-28 11:00:00    4.000000  1991.040000   497.76000
std                            NaN    1.117981   493.144478   140.73639
```

```
# Check for Missing Values
print(customers.isnull().sum())
print(products.isnull().sum())
print(transactions.isnull().sum())
```

output:

```
CustomerID     0
CustomerName    0
Region         0
SignupDate     0
dtype: int64
ProductID     0
ProductName    0
Category      0
Price         0
dtype: int64
TransactionID     0
CustomerID        0
ProductID         0
TransactionDate    0
Quantity          0
TotalValue        0
```

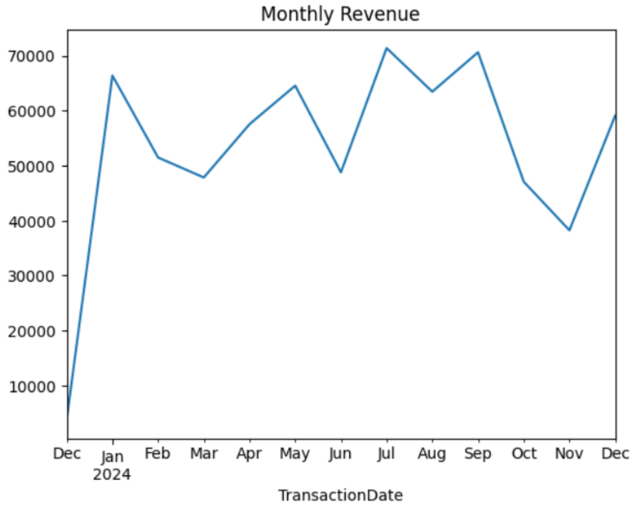
```
Price           0
dtype: int64
```

```
# Merge Datasets
transactions["TransactionDate"] = pd.to_datetime(transactions["TransactionDate"])
merged = transactions.merge(customers, on="CustomerID").merge(products, on="ProductID")

# Save the merged dataset to a CSV file
merged.to_csv("merged_dataset.csv", index=False)

# Download the file
from google.colab import files
files.download("merged_dataset.csv")


# Revenue by Month
monthly_revenue = merged.groupby(merged["TransactionDate"].dt.to_period("M"))["TotalValue"].sum()
monthly_revenue.plot(kind="line", title="Monthly Revenue")
plt.show()
```

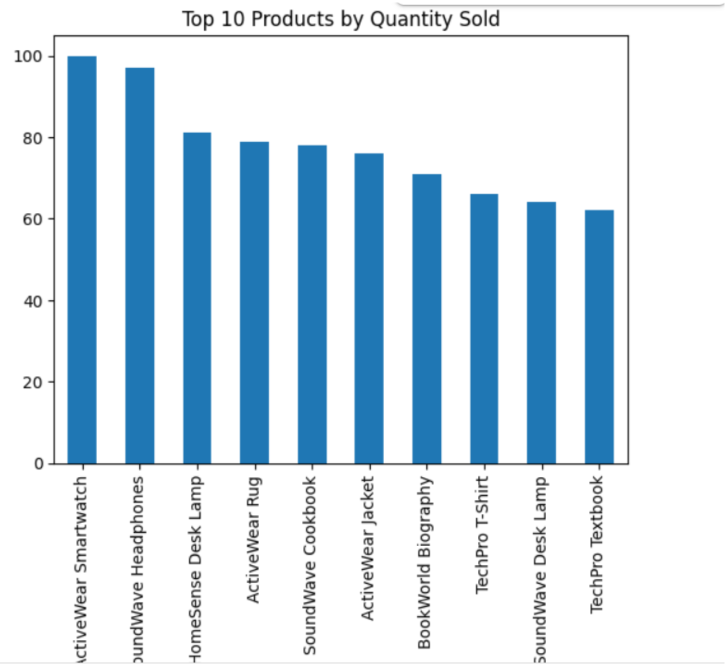


```
# Top-Selling Products

top_products = merged.groupby("ProductName")["Quantity"].sum().nlargest(10)

top_products.plot(kind="bar", title="Top 10 Products by Quantity Sold")

plt.show()
```

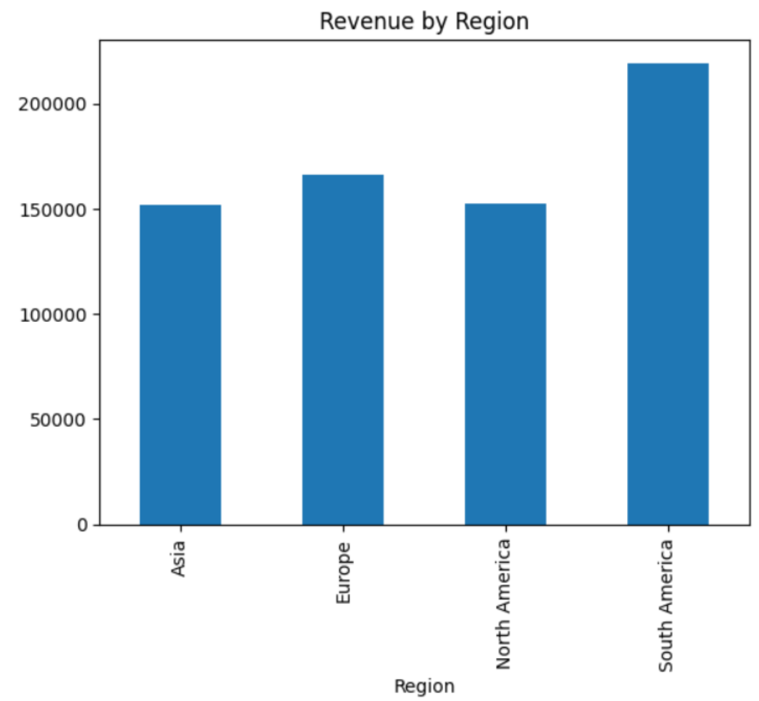


```
# Revenue by Region

revenue_by_region = merged.groupby("Region")["TotalValue"].sum()

revenue_by_region.plot(kind="bar", title="Revenue by Region")

plt.show()
```



Key Findings

Data Overview

- **Number of Entries:**
  - Total Customers: 200
  - Price: 100
  - Total Transactions: 1000
- **Data Quality:**
  - Missing Values: No Missing values found.
  - Duplicate Records: No duplicates were identified.

Trends

- **Monthly Revenue Trends:**
  - Revenue peaks in months July and September, indicating possible seasonality.
  - A decline observed during months June and Novmber.
- **Top-Selling Products:**
  - Products Active Ware Smartwatch, SoundWave Headphones were the most frequently purchased.
- **Regional Insights:**

- South America generated the highest revenue
- Europe show potential for growth with targeted marketing strategies.

Visualizations

1. **Monthly Revenue Trend:** A line chart depicting revenue fluctuations across the year.
2. **Top 10 Products:** A bar chart showing the quantity sold for the highest-performing products.
3. **Revenue by Region:** A bar chart highlighting revenue contributions by different regions.