Exploratory Data Analysis (EDA)

This project involved analyzing datasets related to customers, products, and transactions to uncover insights and trends. The task was performed independently using Python, with libraries such as pandas, seaborn, and matplotlib for data manipulation and visualization. Below is a detailed breakdown of the task:

1. Loading and Preparing the Data

- Datasets Used:
  - Customers.csv: Contains information about customers, including their region.
  - Products.csv: Contains details about various products, categorized by type.
  - Transactions.csv: Includes transaction data, such as dates and total transaction values.
- Steps:
  - Loaded all datasets using pandas.
  - Checked for missing values in each dataset using isnull() and sum(). This step ensured data completeness and highlighted potential cleaning requirements before proceeding.

2. Data Visualization

- **Customer Distribution by Region**:
  - Created a count plot to visualize the distribution of customers across different regions.
  - Used seaborn's countplot function, ordering regions based on the number of customers.
  - This visualization helped identify regions with higher or lower customer concentrations.
- **Product Distribution by Category**:
  - Visualized the number of products within each category using another count plot.

- This analysis provided insights into which product categories were more prominent in the dataset.

3. Transaction Analysis Over Time

- **Steps**:
  - Converted the TransactionDate column in the transactions dataset into a datetime format for easier manipulation.
  - Extracted the month from transaction dates and grouped the data by month to calculate total transaction values.
- **Visualization**:
  - Used a line plot to display the trend of total transaction values over time.
  - This visualization showed monthly trends, helping identify patterns, such as seasonality or spikes in sales.

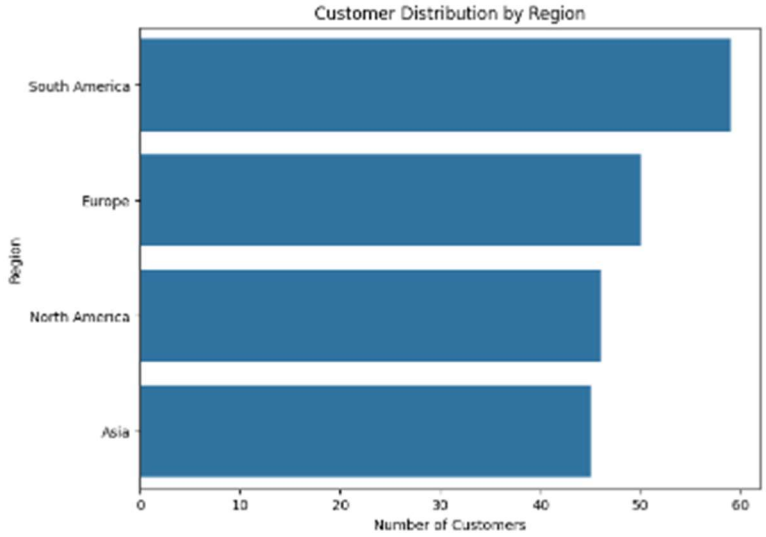
Key Learnings

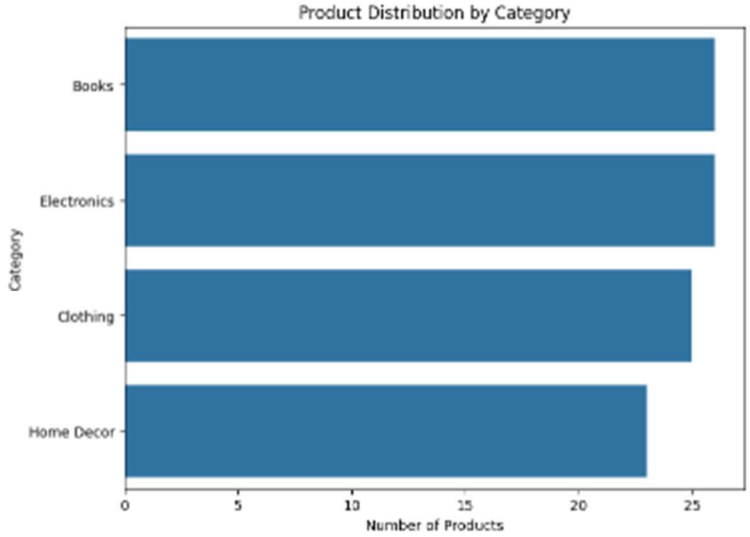
- Identifying missing values is critical for ensuring reliable analysis.
- Visualizing customer and product distributions provides a better understanding of data trends.
- Analyzing transaction trends over time reveals seasonality and growth patterns, enabling data-driven decision-making.

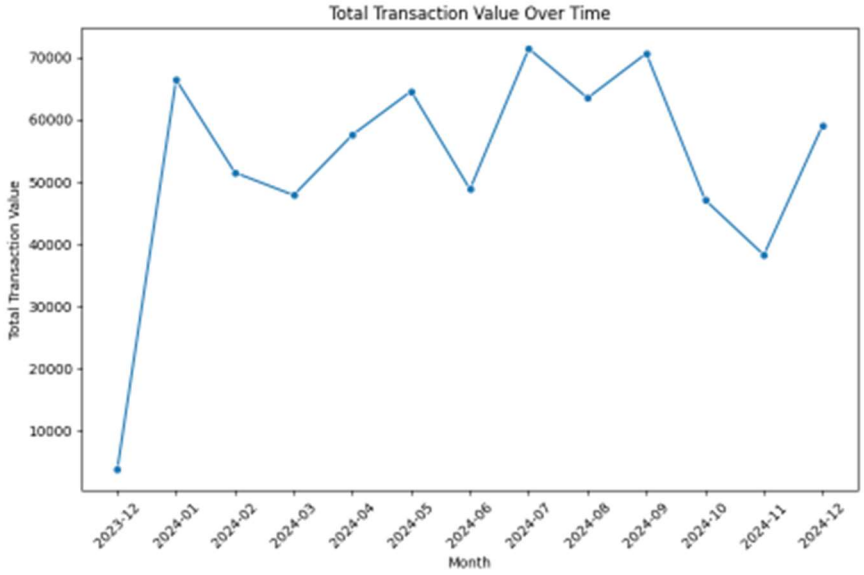
```
Missing values in customers dataset:
CustomerID      0
CustomerName    0
Region          0
SignupDate      0
dtype: int64

Missing values in products dataset:
ProductID      0
ProductName    0
Category       0
Price          0
dtype: int64

Missing values in transactions dataset:
TransactionID      0
CustomerID         0
ProductID          0
TransactionDate    0
Quantity           0
TotalValue         0
Price              0
dtype: int64
```


Customer Distribution by Region
South America
Europe
North America
Asia
Region
0
10
20
30
40
50
60
Number of Customers


Product Distribution by Category
Books
Electronics
Clothing
Home Decor
Category
0
5
10
15
20
25
Number of Products


Total Transaction Value Over Time
70000
60000
50000
40000
30000
20000
10000
Total Transaction Value
2023-12
2024-01
2024-02
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
Month