Exploratory Data Analysis (EDA)

Data Overview

The analysis involved three datasets: **Customers**, **Products**, and **Transactions**, which were merged to create a unified dataset for further exploration. The data was checked for missing values and duplicates, and none were found, ensuring the data's completeness and accuracy.

Key Findings

1. **Category-Wise Total Sales**
   The total sales revenue for each product category is as follows (in descending order):
   - **Books**: $192,147.47
   - **Electronics**: $180,783.50
   - **Clothing**: $166,170.66
   - **Home Decor**: $150,893.93

Books contribute the most revenue, while Home Decor contributes the least.

2. **Regional Sales Analysis**
   The total sales by region were analyzed, revealing that:
   - **South America** and **Europe** generated the highest sales.
   - **Asia** and **North America** had similar sales, slightly lower than the top-performing regions.
3. **Total Value Distribution**
   - The **TotalValue** column exhibits **right skewness**, as observed from its histogram and skewness measure.
   - To address this skewness, a **log transformation** was applied, which significantly reduced the skewness and normalized the distribution.
4. **Outlier Analysis**
   - Boxplots for numerical columns (**TotalValue**, **Quantity**, and **Price**) revealed no significant outliers in the data, confirming its quality for modeling.

Visual Insights

- **Univariate Analysis**:

    - The count of customers by region shows varied distribution across regions, with some regions having significantly higher customer counts.
    - The average price by product category highlighted Electronics as having relatively higher prices compared to other categories.
- **Bivariate Analysis**:
    - A strong positive correlation between **Price** and **TotalValue** was observed, as visualized in the heatmap.
    - Regional sales analysis using bar plots revealed the performance differences across regions, aiding targeted strategies.

Data Preparation for Further Analysis

- The datasets were merged using customer and product IDs to form a comprehensive dataset.
- Duplicate entries were removed, and datetime columns were correctly formatted.
- Numerical features were scaled appropriately to prepare for machine learning tasks.