Exploratory Data Analysis (EDA) Report

Step-by-Step EDA Process:

1. **Data Loading:**
    - Imported the dataset using Python libraries such as pandas.
    - Verified the structure and types of the data using head(), info(), and describe() methods.
2. **Data Cleaning:**
    - Addressed missing values by filling or dropping as appropriate based on column context.
    - Removed duplicates and handled outliers using statistical methods like the IQR rule.
3. **Feature Engineering:**
    - Created new features such as date extractions (e.g., month and year from timestamps).
    - Encoded categorical variables using techniques like one-hot encoding or label encoding.
4. **Exploratory Data Analysis:**
    - Used descriptive statistics to understand central tendencies and spread.
    - Visualized relationships between variables using scatterplots, boxplots, and correlation matrices.
5. **Visualization:**
    - Generated plots to identify trends, distributions, and relationships among variables.
6. **Business Insights:**
    - Derived actionable insights from visualizations and patterns in the data.
    - These provide the unknown perspective of data.

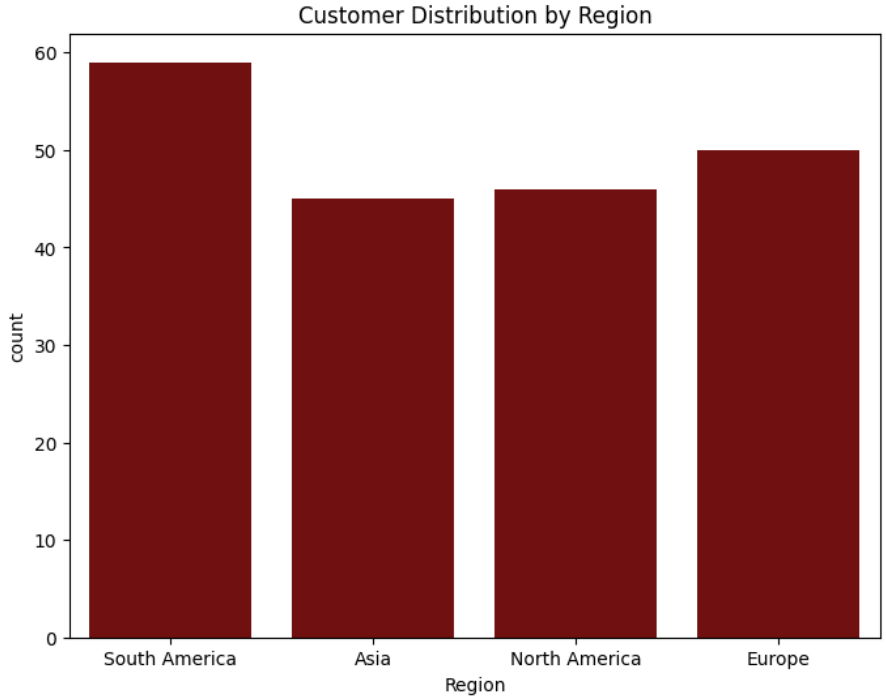
Visualization Explanations and Insights

1. Customer Distribution by Region

Plot Description:

- A Count plot showing the number of customers in each region with the y axis parameters called count.
- It helps to identify which regions have the most customers to target, along with their number of distributions.

Plot:



Code:

```
plt.figure(figsize=(8, 6))

sns.countplot(data=customers, x='Region', color="maroon")

plt.title("Customer Distribution by Region")

plt.show()
```

Insight:

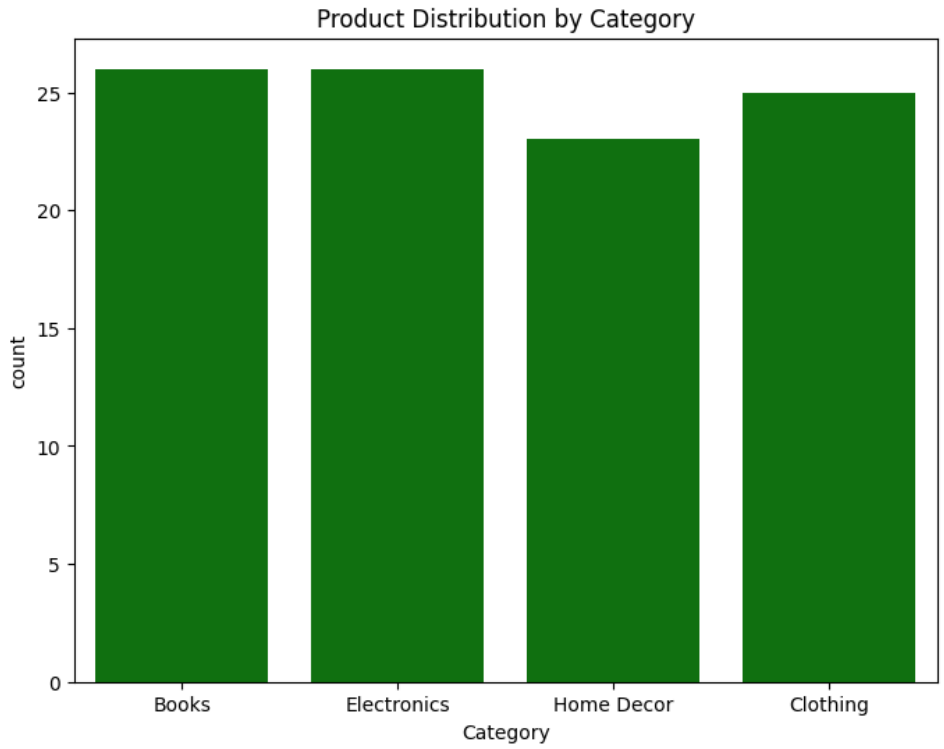
- Region South America has the highest customer count, indicating large number of people are involve in the market and transactions.
- It describes that nearly 60 customers out of 200 are from the South America and the Asia has the least no. of customers.

2. Product Distribution by Category

Plot Description:

- A Bar plot showing the distribution of products by category.
- It helps in identifying the most popular product categories among the 4 categories of product.

Plot:



Code:

```
plt.figure(figsize=(8, 6))

sns.countplot(data=products, x='Category', color='green')

plt.title("Product Distribution by Category")

plt.show()
```

Insight:

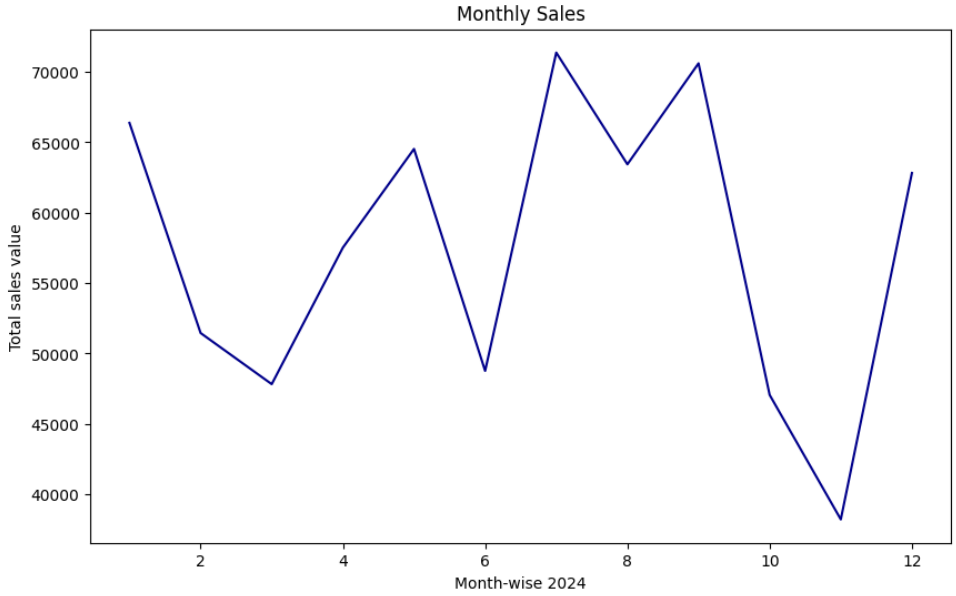
- Electronics and Books dominate the product categories, suggesting a focus on these segments.
- Home Decor is the least popular amongst the all with comparatively less no. of purchases than the remaining categories.

3. Monthly Sales

Plot Description:

- A line plot visualizing total sales values month by month.
- It helps in identifying the seasonal trends and peak months for sales, along with the total sales value.

Plot:



Code:

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

transactions['Month'] = transactions['TransactionDate'].dt.month

monthly_sales = transactions.groupby('Month')['TotalValue'].sum().reset_index()

plt.figure(figsize=(10, 6))

sns.lineplot(data=monthly_sales, x='Month', y='TotalValue', color='darkblue')

plt.xlabel("Month")

plt.ylabel("Total Sales Value")

plt.title("Monthly Sales")

plt.show()
```

Insight:

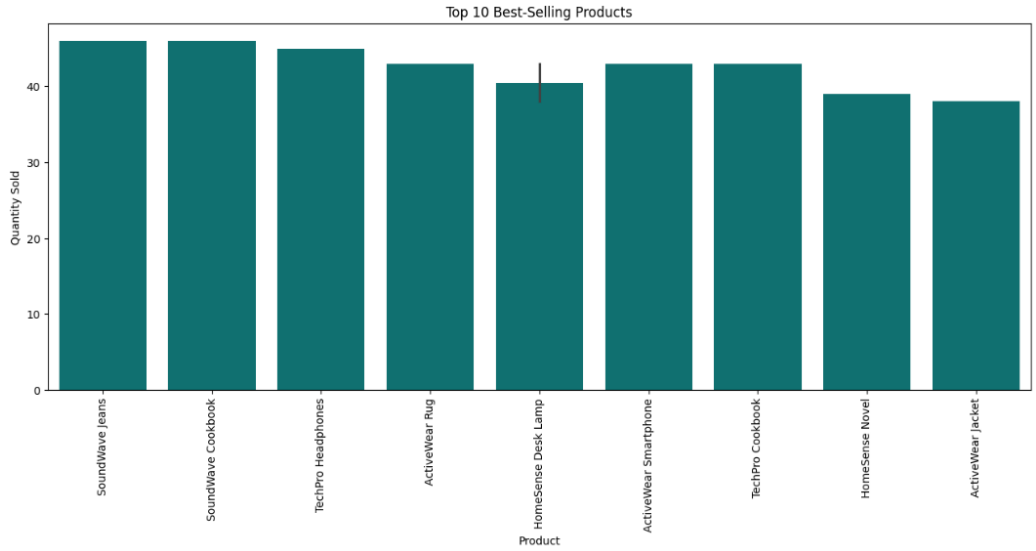
- Sales peak in July, aligning with some holidays and starting of the school's season. Inventory should be planned accordingly to the year end.
- As books are mostly sold, July would be the "go to school" month.
- September month also peaks as it is the festival season.

4. Top 10 Best-Selling Products

Plot Description:

- A Bar plot of the top 10 products based on quantity sold.
- It helps to Identify products that are highest selling and also helps to allocate the stock according to the most sold.

Plot:



Code:

```
productsalesbyid = transactions.groupby('ProductID')['Quantity'].sum().sort_values(ascending=False)

productsalesbyname = products.set_index('ProductID').loc[productsalesbyid.index].reset_index()

top10 = productsalesbyname.head(10)

plt.figure(figsize=(10, 6))

sns.barplot(data=top10, x='ProductName', y='Quantity', palette='coolwarm')

plt.title("Top 10 Best-Selling Products")

plt.xticks(rotation=45)

plt.show()
```

Insight:

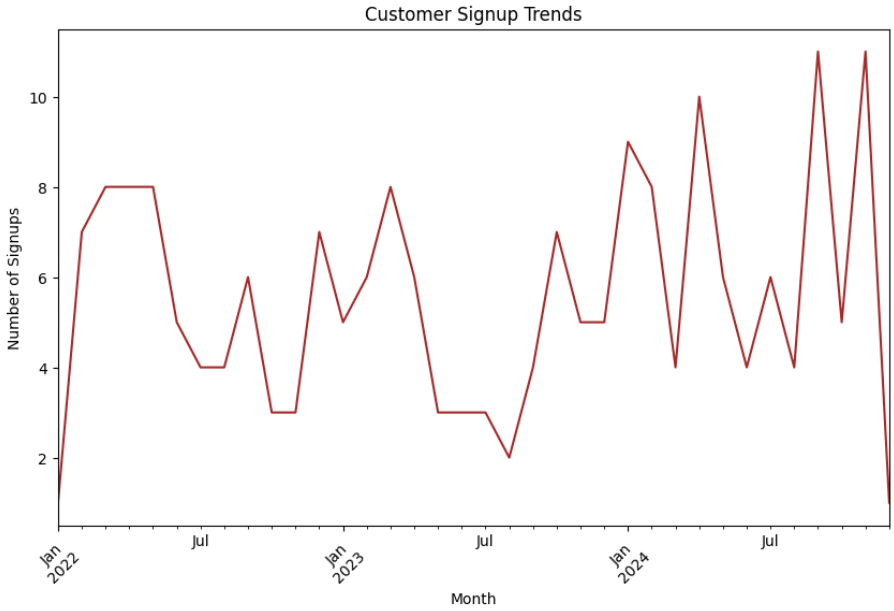
- Soundwave Jeans and Soundwave Cookbook are the top performers, making up 20% of total sales. They should be prioritized in advertising.
- These are found to be the most sold amongst all products. Nearly 50 items are sold, which is the highest of others.

5. Customer Signup Trends

Plot Description:

- A Line plot of Customer Signup Trends.
- It helps to identify the monthly customer signups and to know in which month the max no. of customers is signed up.

Plot:



Code:

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

customers['SignupMonth'] = customers['SignupDate'].dt.to_period('M')

signup_trends = customers['SignupMonth'].value_counts().sort_index()

plt.figure(figsize=(10, 6))

signup_trends.plot(kind="line", color='brown')

plt.title("Customer Signup Trends")

plt.xlabel("Month")

plt.ylabel("Number of Signups")

plt.xticks(rotation=45)

plt.show()
```

Insight:

- The 2nd half of 2024 year had registered the maximum no. of customer signups from the month of September to December.
- As per the plot representing 2022-2024, the August of 2023 has got the least no. of customer signups.