Exploratory Data Analysis (EDA) on E-commerce Dataset

Introduction

Exploratory Data Analysis (EDA) is a crucial step in understanding the structure, relationships, and trends within a dataset. This analysis was conducted on an **E-commerce dataset (Superstore_USA.xlsx)** using Python libraries such as **pandas, NumPy, Matplotlib, and Seaborn**. The dataset contains information about orders, sales, customers, product categories, regions, and order priorities.

The EDA process involved:

- Data Loading & Initial Exploration
- Missing Value Analysis & Handling
- Statistical Summary
- Data Visualization
- Key Business Insights

1️⃣ Data Loading & Initial Exploration

Dataset Overview

The dataset was imported using `pandas` from an **Excel file** and examined using:

- `df.head()`: Displayed the first few rows.
- `df.info()`: Checked column data types and null values.
- `df.describe()`: Generated summary statistics for numerical columns.

The dataset comprises multiple fields such as:

- **Order ID, Product Name, Category, Sales, Profit, Order Priority, Region, Customer Segment, Discount, and Quantity.**
- The goal was to extract insights about sales patterns, profitability, and customer purchasing behavior.

2️⃣ Missing Value Analysis & Handling

Identifying Missing Data

The presence of missing values was checked using:

```
print(df.isnull().sum())
```

Several columns had missing values, which needed to be addressed to maintain data integrity.

Handling Missing Values

Approaches used:

- **Numerical Data:** Missing values were filled using **mean/median** imputation.
- **Categorical Data:** Mode was used for replacement or missing entries were dropped if they were insignificant.
- `df.fillna()` and `df.dropna()` were applied as necessary.

Post-cleaning, the dataset was re-examined to ensure there were no missing values left.

3️⃣ Exploratory Data Analysis (EDA) & Key Insights

A. Order Priority Distribution

- A **bar plot** was used to visualize the frequency of different order priorities (Critical, High, Medium, Low).
- Critical orders formed a significant portion, indicating urgent customer demand.

```
sns.countplot(x=df['Order Priority'])
plt.title("Distribution of Order Priorities")
plt.show()
```

B. Sales & Profit Trends

- **Line Charts & Histograms** were used to analyze sales performance over time.
- Monthly and yearly trends were identified.
- **Profitability across different product categories was examined.**
- The **top-performing categories** were identified based on total revenue.

```
plt.figure(figsize=(10,5))
df.groupby('Order Date')['Sales'].sum().plot()
plt.title("Sales Over Time")
plt.show()
```

C. Category & Region-Wise Sales Analysis

- A **bar chart** displayed sales per category.
- A **pie chart** illustrated revenue contribution by region.
- Some regions performed significantly better, guiding potential expansion strategies.

```
df.groupby('Category')['Sales'].sum().plot(kind='bar')
plt.title("Sales by Category")
plt.show()
```

D. Correlation Between Variables

- A **heatmap** (`sns.heatmap(df.corr())`) was used to study correlations.
- Key findings:
  - Strong positive correlation between **Sales and Profit**.
  - Negative correlation between **Discount and Profit**, suggesting high discounts reduce profitability.

```
plt.figure(figsize=(10,5))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```

4️⃣Data Visualization Techniques Used

To present insights effectively, various visualization methods were employed:

- 📊 **Bar Charts** – Order priorities, category-wise sales.
- 📈 **Line Charts** – Sales trends over time.
- 🔥 **Heatmaps** – Feature correlations.
- 🏆 **Pie Charts** – Regional sales contribution.
- 📉 **Histograms** – Profit distribution.

5️⃣Key Takeaways

✅ Order Priority Trends

- Most orders were classified as **High and Critical**, highlighting urgent demand.

✅ Profitability Insights

- Some categories with **lower sales volume** still achieved high profitability.

- Discounting strategies needed improvement as excessive discounts reduced overall profit.

✅ Regional Analysis

- Certain states/regions contributed **higher revenue**, making them ideal targets for future expansion.

✅ Discount Impact

- Higher discounts had a **negative impact on profit margins**.
- A **dynamic pricing strategy** could help optimize profits.

Conclusion

This EDA provided **valuable business insights** by uncovering patterns in order priority, sales performance, regional contributions, and discount strategies. The findings can be leveraged to **optimize pricing, enhance marketing strategies, and improve inventory management** for better profitability and efficiency.

The next steps involve deploying **predictive analytics models** to forecast sales and demand, enabling **data-driven decision-making** for the business.

This analysis serves as a foundation for strategic improvements in the e-commerce domain! 🚀