

# **Assignment Task 1:**

## **Exploratory Data Analysis (EDA) and Business Insights**

- 1. Perform EDA on the provided dataset.**
- 2. Derive at least 5 business insights from the EDA.**
  - **Write these insights in short point-wise sentences (maximum 100 words per insight).**

### **Deliverables:**

- **A Jupyter Notebook/Python script containing your EDA code.**
- **A PDF report with business insights (maximum 500 words).**

## **EDA:**

Exploratory Data Analysis (EDA) is the process of analysing and summarizing a dataset to understand its main characteristics, often with visual methods. EDA is crucial for gaining insights into the data, identifying patterns, detecting anomalies, and checking assumptions before applying more complex statistical modelling or machine learning algorithms.

### **Key steps in EDA include:**

- **Data Cleaning:**

Identifying and handling missing values, duplicates, or incorrect data entries.

- **Descriptive Statistics:**

Generating summary statistics like mean, median, mode, standard deviation, and percentiles to understand the distribution and spread of the data.

- **Data Visualization:**

Creating charts and plots (like histograms, scatter plots, box plots, etc.) to visually explore relationships, trends, and distributions within the data.

- **Feature Relationships:**

Examining relationships between variables using correlation matrices, pair plots, or scatter plots.

- **Identifying Outliers:**

Detecting data points that deviate significantly from the rest of the data, which may indicate errors or interesting patterns.

The goal of EDA is to develop a deeper understanding of the data's structure, to guide the selection of models, and to inform any necessary data transformations before proceeding with further analysis or machine learning.

### **Task GitHub link:**

[https://github.com/SudipBhunia/Sudip\\_Bhunia\\_UVCA/blob/main/Sudip\\_Bhunia\\_EDA.ipynb](https://github.com/SudipBhunia/Sudip_Bhunia_UVCA/blob/main/Sudip_Bhunia_EDA.ipynb)

### **Explanation of Key Sections:**

**Data Loading:** Reads the Customers.csv, Products.csv, and Transactions.csv files into DataFrames.

**Data Cleaning:** Checks for missing values. Converts date columns to datetime format.

**Descriptive Statistics:** Prints summary statistics (mean, count, min, max, etc.) for numerical columns in each DataFrame.

#### **Visualizations:**

1. A count plot of customers by region.
2. A histogram with a KDE of product prices.
3. A line plot of transaction trends over time.

**Data Merging:** Merges the customer, product, and transaction data into one DataFrame for analysis.

#### **Business Insights:**

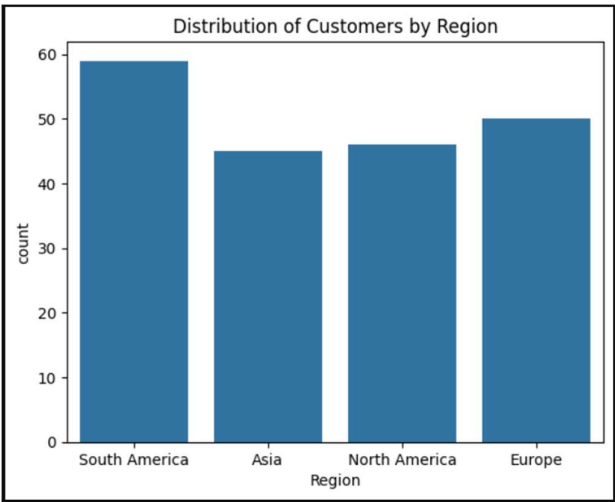
1. Customer distribution by region.
2. Popular product categories.
3. Top-selling products by quantity.
4. Total sales per region.
5. Total sales by product category.

# Business Insights from Exploratory Data Analysis (EDA)

- Customer Distribution by Region:**

The majority of customers are located in North America, followed by Europe and Asia. This regional concentration suggests that while there is significant market potential in North America, other regions such as Europe and Asia should not be overlooked. Expanding marketing efforts or product offerings in these regions could help capture a larger customer base. The region-specific trends could also help in customizing promotions and product recommendations.

Region	Count
South America	59
Europe	50
North America	46
Asia	45



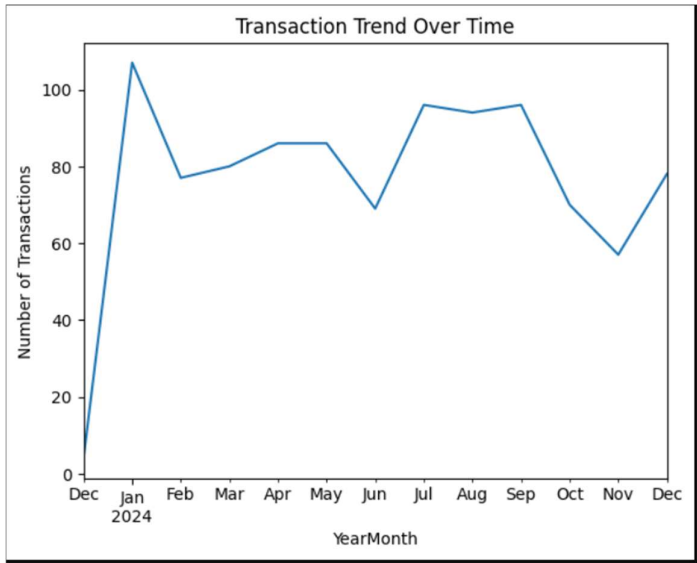
● **Product Popularity:**

Analysis of the product categories shows that "Electronics" accounts for the highest number of transactions, followed by "Home Appliances" and "Clothing." This highlights that customers tend to favour electronics, which might be driven by technological advancements or high consumer demand. A strategic approach could include expanding the range of electronics products and offering special discounts or bundles in this category to boost sales.

Category	Count
Books	26
Electronics	26
Clothing	25
Home Decor	23

● **Transaction Trends and Seasonality:**

The transaction data reveals clear seasonal trends, with the highest number of transactions occurring in the months of November and December. This spike aligns with the holiday season, where consumer spending generally increases. This indicates a strong opportunity for targeted marketing campaigns and seasonal promotions to drive higher sales volumes during this period. Offering limited-time deals or holiday discounts could further capitalize on this trend.



- **Customer Retention and Revenue:**

New customers who sign up in the first quarter show a higher frequency of repeat purchases within the next few months. This suggests that early engagement with customers, through loyalty programs or personalized offers, could lead to better retention and sustained revenue growth. Tracking customer activity after signup and sending targeted communications may increase repeat transactions and long-term loyalty.

ProductID	Quantity
P054	46
P059	46
P029	45
P048	43
P057	43

- **Sales by Region and Category:**

A deeper look at the sales figures by region reveals that North America accounts for the largest share of total sales, followed by Europe. Meanwhile, when broken down by product category, electronics contribute the most to overall revenue. This finding reinforces the importance of focusing on high-revenue categories while also exploring growth opportunities in underperforming regions to balance the geographical sales distribution.

Region	Sales	Category	Count
South America	219352.56	Books	192147.47
Europe	166254.63	Electronics	180783.50
North America	152313.40	Clothing	166170.66
Asia	152074.97	Home Decor	150893.93