Exploratory Data Analysis (EDA) with Pandas in Diwali Sales

Project Purpose:

The purpose of this project is to explore and analyze a Diwali sales dataset using the Pandas framework. The goal is to derive insights into customer behavior, sales performance, and product trends during the Diwali shopping season. The analysis will focus on identifying high-performing regions, products, and customer segments to help improve business decision-making.

Goals of the Project:

1. **Explore the Diwali Sales Dataset Using Pandas**
    - Inspect the dataset's structure (columns, data types, and summary statistics).
    - Identify and handle missing values, duplicate rows, and other data quality issues.
    - Create an overview of key sales metrics (total sales, average order value, etc.).
2. **Perform Feature Engineering**
    - Generate new features such as Total_Spending, Age Group, and Order_Category.
    - Create new features based on order quantity (e.g., "Low", "Moderate", "High" spending categories).
    - Derive meaningful columns like shipping delay and profit margin from the data.
3. **Data Visualization**
    - Use visualizations to explore distributions of key variables (e.g., purchase amount, total spending).
    - Analyze sales trends over time (monthly, quarterly).

    - Compare sales by region, product category, and customer age group.
4. **Summarize Key Findings**
    - Identify top-performing regions and product categories.
    - Understand customer behavior: identify high-spending customers, returning customers, and purchasing trends.
    - Gain insights into sales fluctuations and the effectiveness of different shipping methods.

Materials and Methods:

Dataset:
The dataset consists of sales data from a Diwali campaign and contains the following features:

- **Customer Details**: Region, age group, and order history.
- **Order Details**: Order amount, product category, and quantity.
- **Shipping Information**: Shipping delay, shipping type, and delivery details.

Libraries Used:

- **Pandas**: For data manipulation, cleaning, and exploration.
- **NumPy**: For numerical operations.
- **Seaborn & Matplotlib**: For data visualization.
- **Warnings**: To suppress irrelevant warnings.

Steps in the EDA Process:

1. **Data Exploration**
    - Load the dataset.
    - Check for missing values, duplicates, and data types.

    - Get summary statistics for numeric features.
2. **Data Cleaning**
    - Handle missing or duplicated data.
    - Create new features like Total_Spending, Age Group, and categorize orders based on order quantity.
3. **Feature Engineering**
    - Generate new columns to explore relationships (e.g., Total_Spending, Age Group, Order_Category).
    - Analyze sales performance across regions and product categories.
4. **Visualization**
    - Use various plots such as bar plots for regional sales, histograms for distribution of amounts, and time series plots for trends over time.
    - Explore customer behavior by segmenting the dataset into age groups, order categories, and shipping delays.
5. **Customer Analysis**
    - Group by User_ID to analyze customer spending and identify top customers.
    - Examine sales performance across different customer age groups and categories.

Expected Outcomes and Insights:

- **Sales Trends**: Analysis of total sales by region and product category, and how they fluctuate over time.
- **Customer Insights**: Identify high-spending and returning customers, which will help with targeted marketing and customer retention strategies.
- **Product Insights**: Identify which product categories perform the best and align marketing efforts accordingly.

- **Shipping Analysis**: Understand the impact of shipping delays on customer satisfaction and profitability.
- **Profitability**: Analysis of average spending per order and overall profitability by region and category.

Visualizations:

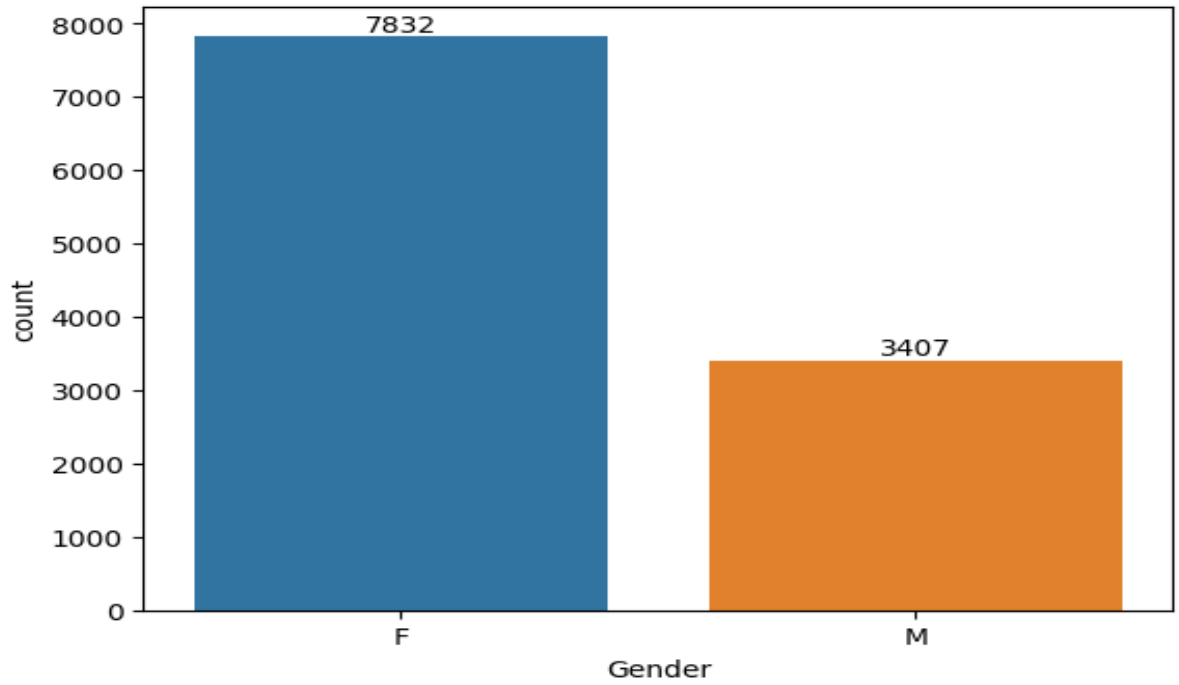
- **Sales by Region**: A bar chart to show total sales across different regions.
- **Sales by Product Category**: A bar plot highlighting which categories are generating the most revenue.
- **Customer Spending Distribution**: A boxplot to identify high-spending customers and spending patterns.
- **Sales Trends**: A line chart showing how sales evolve over the Diwali period.
- **Shipping Delay vs Sales**: A bar plot or boxplot comparing shipping types and their impact on sales.

Visualization:

Several charts created to present inside including

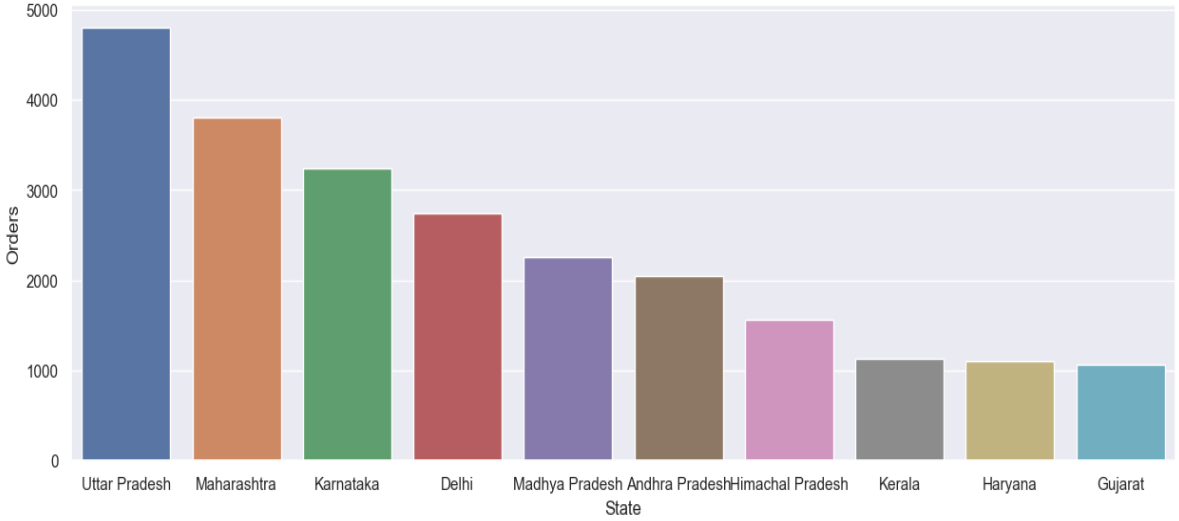
Gender:

plotting a bar chart for Gender and it's count



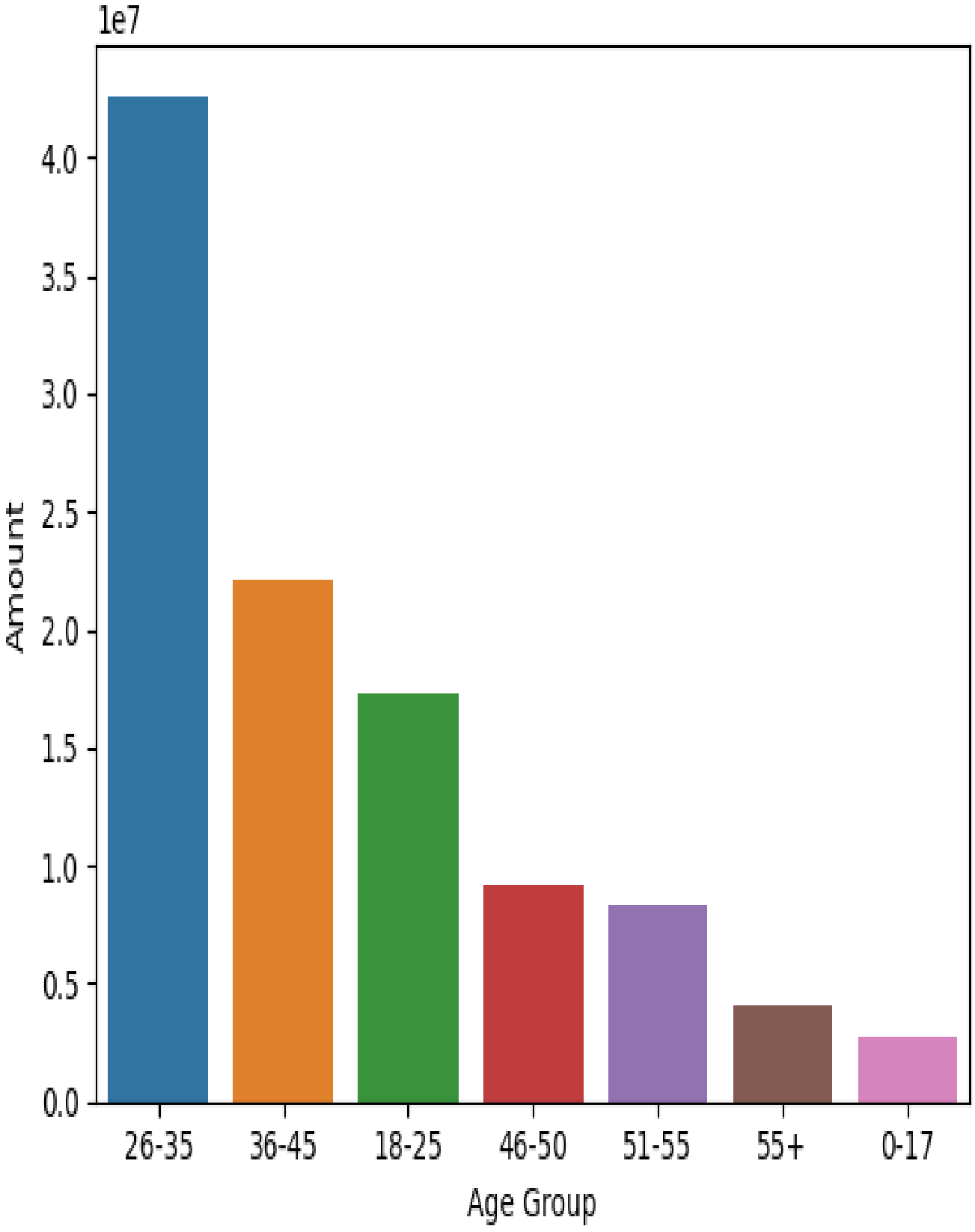
State:

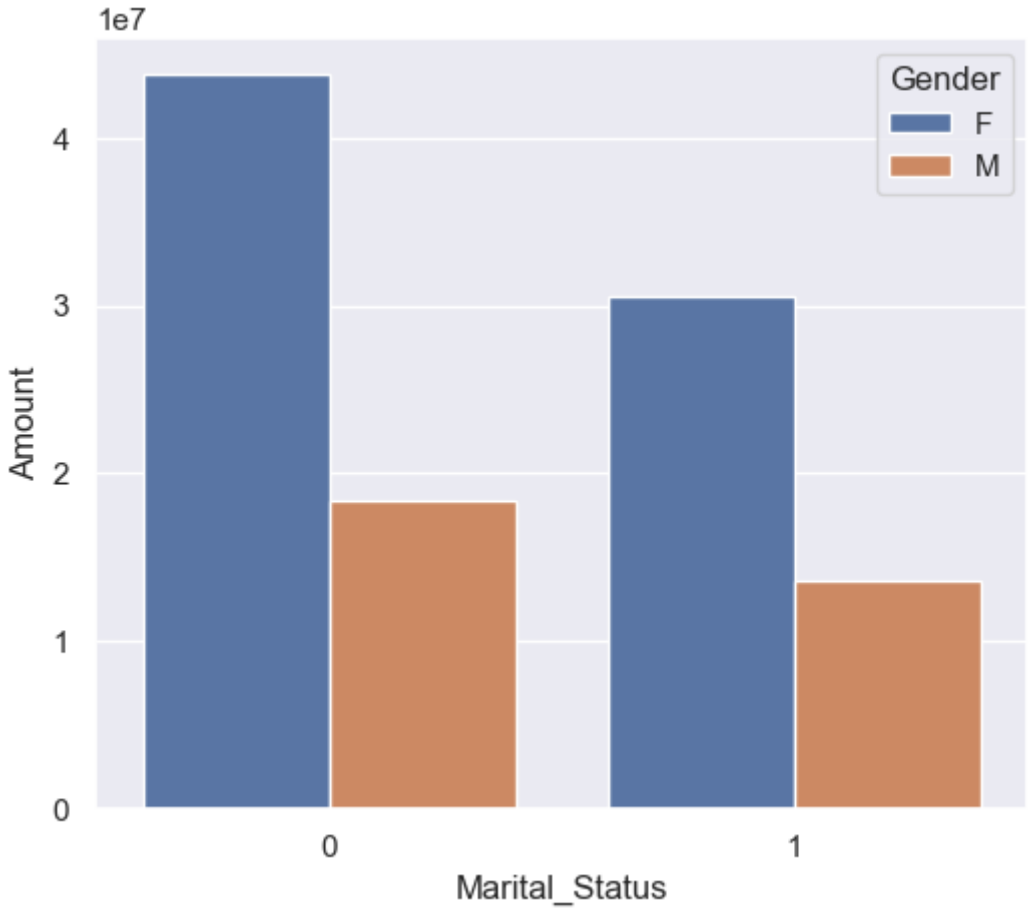
Total number of orders from top 10 states



From above graphs we can see that most of the orders & total sales are from Uttar Pradesh, Maharashtra and Karnataka respectively

Age:

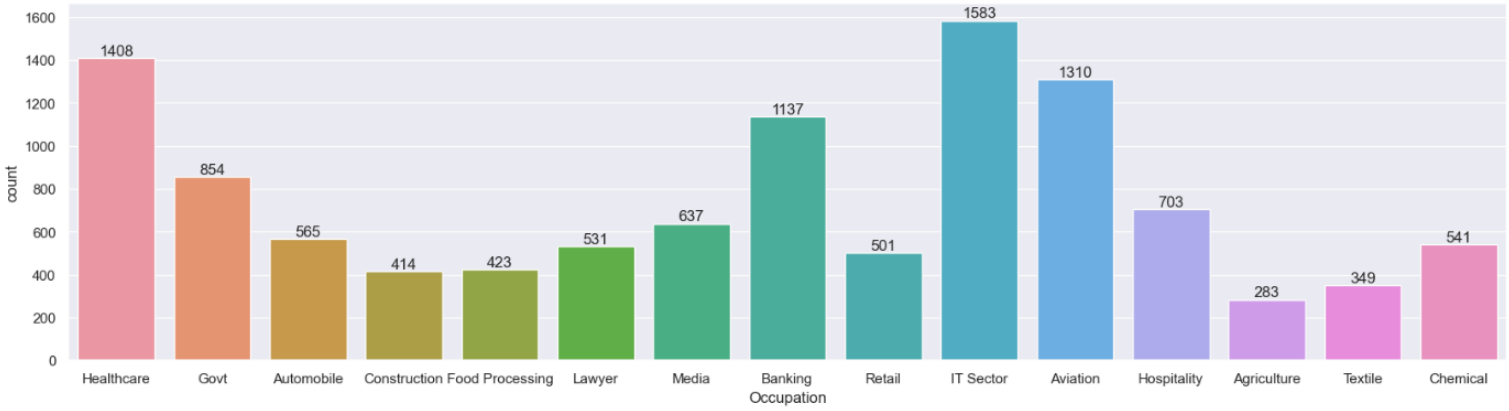

1e7
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Amount
26-35
36-45
18-25
46-50
51-55
55+
0-17
Age Group

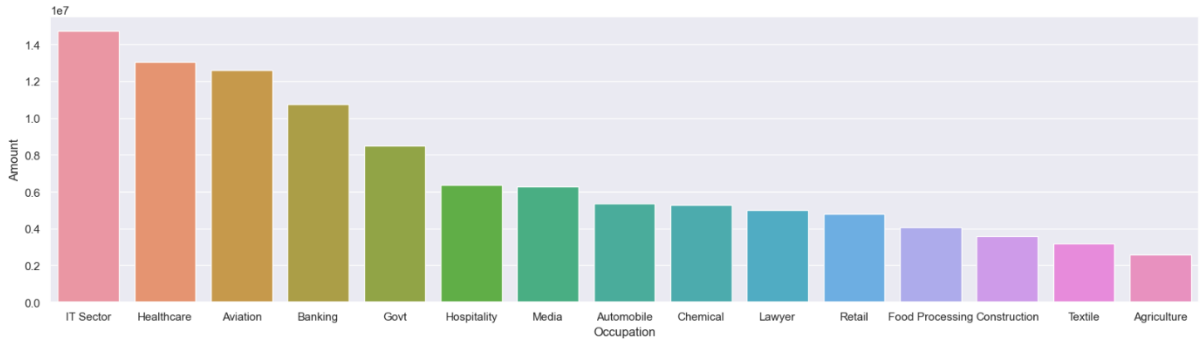


Marital States :

From above graphs we can see that most of the buyers are married (women) and they have high purchasing power.

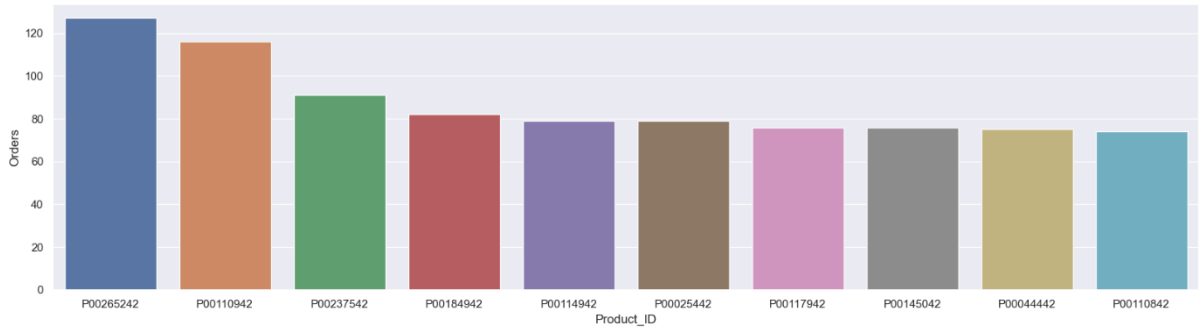
Occupation:





From above graphs we can see that most of the buyers are working in IT, Healthcare and Aviation sector.

Product_ID:



Conclusion:

Married women age group 26-35 yrs from UP and Maharastra and Karnataka working in IT, Healthcare and Aviation are more likely to buy products from Food, Clothing and Electronics category.