Title of the Mini project

A Mini-project Report submitted in partial fulfillment of the requirements for the award of the degree of

Master of Computer Applications

by

Amit Gupta

23419MCA065

Department of Computer Science

Institute of Science

Banaras Hindu University, Varanasi – 221005

December-2024

CANDIDATE'S DECLARATION

I **Amit Gupta** hereby certify that the work, which is being presented in the Mini-project report, entitled **Python Diwali Sales_Analysis,** in partial fulfillment of the requirement for the award of the Degree of **Master of Computer Applications** and submitted to the institution is an authentic record of my/our own work carried out during the period August Month-Year 2024 To December Month-Year 2024 under the supervision of Vandana Kushwaha. I also cited the reference about the text(s) /figure(s) /table(s) /equation(s) from where they have been taken.

The matter presented in this Mini-project as not been submitted elsewhere for the award of any other degree or diploma from any Institutions.

Date:06-12-2024

Signature of
the Candidate

The Viva-Voce examination of Amit Gupta, M.C.A. Student has been held on _________________.

Signature of the Supervisor

ABSTRACT

This project, Diwali Sales Analysis using Python, aims to explore and derive meaningful insights from sales data during the Diwali festival period. The project utilizes powerful Python libraries, including Pandas for data manipulation, NumPy for numerical operations, Matplotlib and Seaborn for data visualization. The primary objective is to clean, preprocess, and analyze the dataset to uncover patterns, trends, and key sales drivers. Through data cleaning techniques, missing values are handled, and outliers are addressed to ensure data integrity. The analysis focuses on various dimensions such as product categories, customer demographics, sales trends, and regional performance. Visualizations provide a comprehensive understanding of customer behavior and highlight high-performing segments. This project not only offers actionable insights for businesses to optimize their marketing strategies but also demonstrates the power of data analytic s in enhancing decision-making processes. The use of Python's rich ecosystem of libraries underscores the efficiency and effectiveness of the analysis.

Keywords: Python, Pandas, NumPy, Seaborn, Matplotlib, Data Cleaning, Data Visualization, Sales Trends, Customer Behavior, Business Insights, Data Analytics, Marketing Strategy Optimization.

TABLE OF CONTENTS

Title Page No.

Introduction

Diwali, known as the festival of lights, is a major shopping season in India. Retailers experience a significant spike in sales across various product categories, making this period highly competitive. Businesses seek to understand consumer behavior and optimize their sales strategies to maximize profits. This project leverages Python's data analysis capabilities to explore sales data, uncover patterns, and provide actionable insights. The analysis focuses on understanding customer demographics, identifying high-performing products, and optimizing inventory planning using various Python libraries such as Pandas, NumPy, Seaborn, and Matplotlib.

1.1 The Importance of Diwali Sales in Business Strategy

The Diwali season is marked by several key trends that impact business strategies:

Surge in Consumer Spending: Consumers are more willing to spend on gifts, household items, and personal indulgences during this period.

Promotional Campaigns: Retailers offer discounts, festive deals, and attractive financing options to lure customers.

Product Launches: Many brands choose Diwali to introduce new products, given the heightened consumer interest.

E-commerce Boom: Online platforms witness a massive surge in traffic and transactions, making digital sales channels critical for retailers.

Understanding these trends helps businesses make data-driven decisions to maximize their returns during the festival.

1.2 Roles Of Data Analysis in Diwali Sales

Data analysis plays a vital role in boosting Diwali sales by helping businesses predict demand, ensuring popular products are well-stocked. It enables personalized marketing, optimized pricing, and the identification of sales trends to adapt quickly to market needs. Effective inventory management prevents overstocking or shortages, while campaign analysis ensures better resource allocation. By understanding regional preferences, businesses can tailor their offerings to maximize revenue and customer satisfaction.

1.3 Python as a Tool for Sales Analysis

Python has emerged as a popular tool for data analysis due to its versatility and the availability of powerful libraries. This project utilizes the following Python libraries:

Pandas: For data manipulation and cleaning. It allows easy handling of large datasets and provides tools for grouping, filtering, and summarizing data.

NumPy: Used for numerical operations, making calculations and transformations faster and more efficient.

Matplotlib: A plotting library that helps create visual representations of data, making it easier to identify trends and patterns.

Seaborn: Built on top of Matplotlib, Seaborn provides aesthetically pleasing and informative statistical graphics, aiding in deeper insights.

1.4 Objectives of the Analysis

The primary goals of this analysis are:

Understanding Customer Demographics: Analyzing factors such as age, gender, occupation, and location to identify potential customers.

Identifying High-Performing Products: Recognizing which products and categories drive the most sales during the Diwali season.

Optimizing Inventory Planning: Using insights to help businesses stock the right products in the right quantities.

Enhancing Customer Experience: Providing insights that can help tailor marketing strategies and improve customer engagement.

1.5 Significance of This Project

This project serves as a practical demonstration of how data analytics can transform raw sales data into actionable insights. By leveraging Python's data analysis capabilities, businesses can gain a competitive edge in the marketplace. For students and professionals, this project highlights the importance of data literacy and the practical application of programming skills in real-world scenarios.

The insights generated through this project will not only help businesses optimize their strategies for the Diwali season but also provide a framework for analyzing sales data during other peak shopping periods.

2 OBJECTIVES

The success of a Diwali sales analysis project hinges on clearly defined objectives that guide the data-driven approach. This section outlines the primary goals, each of which plays a crucial role in understanding and optimizing sales performance during the festive season.

2.1 Data Cleaning and Prepossessing

Objective: Handle missing values, correct inconsistencies, and prepare the data for meaningful analysis.

In any real-world data-set, especially sales data, inconsistencies such as missing values, duplicate entries, and formatting issues are common. Data cleaning is a vital first step to ensure accuracy and reliability in the analysis. For Diwali sales data, this involves:

- **Handling Missing Values:** Sales records may have missing information like customer age or product details. These gaps are addressed using techniques such as imputation or by removing incomplete entries, ensuring that the data-set remains robust.
- **Correcting Inconsistencies:** Standardizing formats for attributes such as state names, product categories, and customer information is critical to avoid duplication and errors.
- **Preparing Data for Analysis**: The cleaned data-set is formatted to facilitate smooth analysis, with necessary transformations applied to numerical and categorical data.

Significance: Clean data ensures that insights derived from the analysis are accurate and actionable, reducing the risk of misleading conclusions.

2.2 Exploratory Data Analysis (EDA)

Objective: Perform statistical and visual analysis to uncover insights.

EDA is a crucial step in understanding the underlying patterns and relationships within the data-set. It involves both statistical summaries and visualizations that help in identifying trends and anomalies. In the context of Diwali sales analysis:

Statistical Analysis: Metrics like mean, median, and mode for sales amounts help in understanding overall performance.

Visual Analysis: Bar charts, histograms, and heatmaps illustrate trends such as sales distribution by region, gender, and product category.

Trend Identification: Seasonal trends, customer preferences, and high-demand periods are identified, offering insights into what drives sales during Diwali.

Significance: EDA provides a comprehensive overview of the data, enabling businesses to make informed decisions based on observed patterns.

2.3 Customer Segmentation

Objective: Identify potential customers based on demographic attributes like state, occupation, gender, and age group.

Customer segmentation allows businesses to tailor their marketing strategies to different customer groups. For Diwali sales, segmentation focuses on:

State-wise Analysis: Identifying which states generate the highest sales helps in regional marketing efforts.

Occupation and Gender Analysis: Understanding customer profiles based on occupation and gender enables targeted promotions and offers.

Age Group Insights: Age-based segmentation highlights which age groups are most likely to make purchases, guiding product offerings and marketing messages.

Significance: Effective segmentation enhances customer engagement by delivering personalized experiences, ultimately driving higher sales and customer loyalty.

2.4 Sales Optimization

Objective: Determine top-selling products and categories to assist in inventory planning.

Optimizing sales involves understanding which products and categories perform best during Diwali. This helps businesses:

Identify High-Demand Products: Knowing which items are most popular allows for better stocking decisions, reducing the risk of stockouts.

Optimize Inventory: Efficient inventory management ensures that businesses can meet customer demand without overstocking.

Plan Promotions: Insights into best-selling categories guide promotional strategies and discount offers, maximizing sales potential.

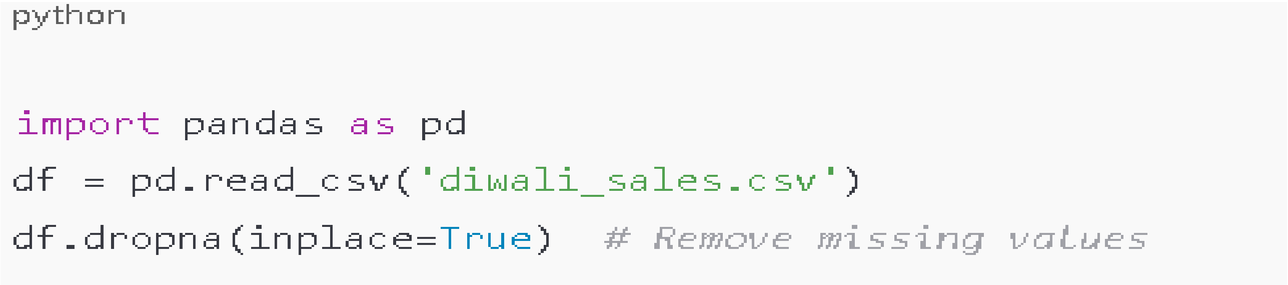
Significance: Sales optimization not only boosts revenue but also minimizes costs associated with excess inventory and lost sales opportunities.

3. Tools and Technologies

3.1 Pandas

Pandas is a powerful library for data manipulation and analysis. It provides data structures like DataFrames and Series to handle and clean data efficiently. Functions like filtering, grouping, and aggregating make it easy to prepare data for analysis.

Example:

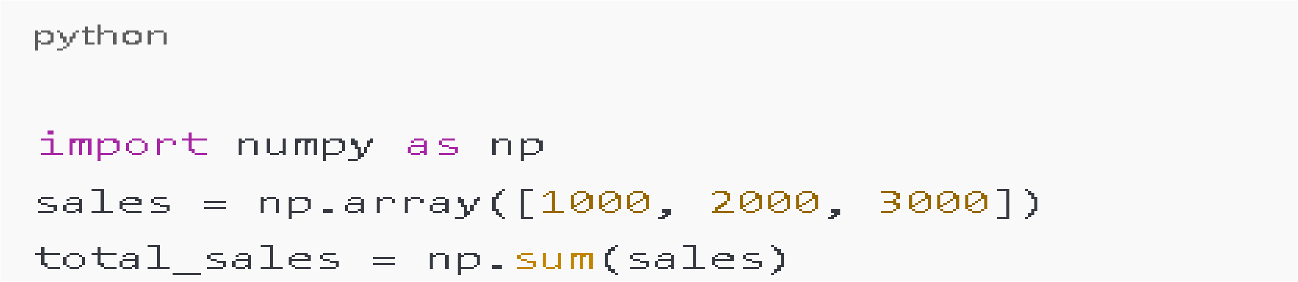


```python
import pandas as pd
df = pd.read_csv('diwali_sales.csv')
df.dropna(inplace=True)  # Remove missing values
```

3.2 NumPy

NumPy is a library used for numerical operations and handling arrays. It provides fast mathematical functions and supports multi-dimensional arrays for efficient computation.

Example:

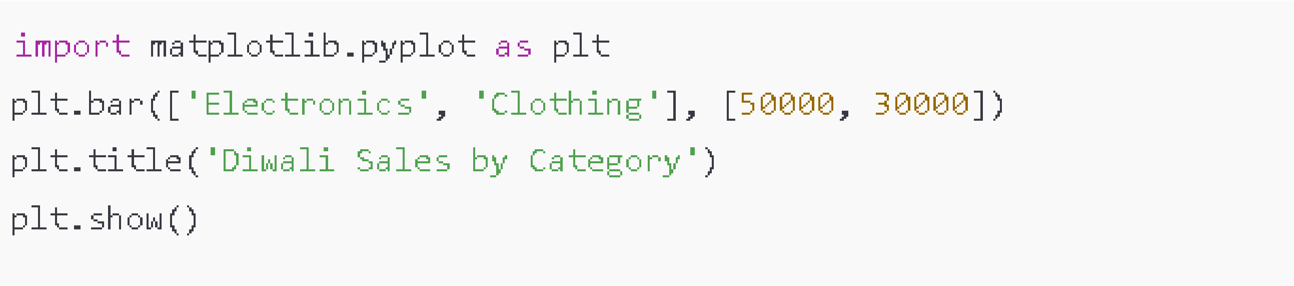


```python
import numpy as np
sales = np.array([1000, 2000, 3000])
total_sales = np.sum(sales)
```

3.3 Matplotlib

Matplotlib is a plotting library used to create static, interactive, and animated visualizations. It allows customization of plots and supports charts like line plots, bar charts, and histograms.

Example:

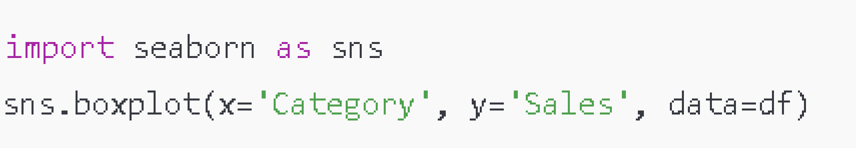


```
import matplotlib.pyplot as plt
plt.bar(['Electronics', 'Clothing'], [50000, 30000])
plt.title('Diwali Sales by Category')
plt.show()
```

3.4 Seaborn

Seaborn builds on Matplotlib, providing more aesthetically pleasing and informative visualizations. It simplifies the creation of statistical graphics such as heatmaps, box plots, and pair plots.

Example:



```
import seaborn as sns
sns.boxplot(x='Category', y='Sales', data=df)
```

Dataset Overview

The dataset utilized in this project comprises detailed information on sales transactions during the Diwali festival. It includes key attributes such as customer demographics, product categories, purchase details, and sales amounts. The dataset is extensive, containing 11,251 rows and 15 columns, offering a rich source of data for meaningful analysis. This comprehensive dataset provides valuable insights into consumer behavior, product performance, and overall sales trends during the festive season, making it ideal for exploring and optimizing sales strategies. The key attributes of the data in CSV file such as :

User ID(as a primary key)	Customer name	Product ID
Gender	Age Group	Age
State	Zone	Marital Status
Occupation	Product Category	Orders
Amount	Status	Unnamed1

Methodology

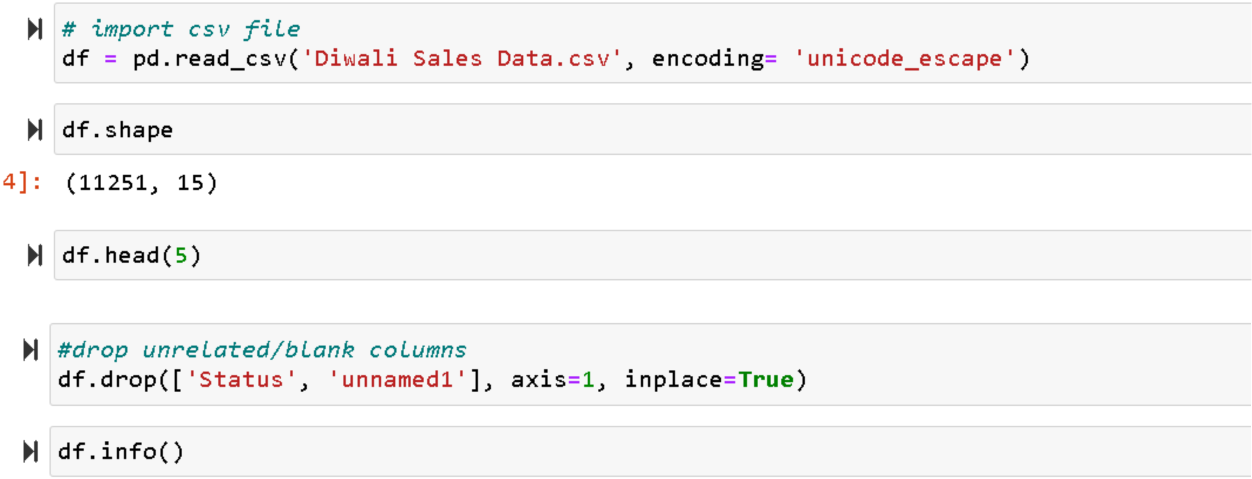
5.1 Data Cleaning and Preprocessing

Data cleaning is essential to ensure the quality and integrity of the analysis. The following steps were performed:

Handling Missing Values: Missing data can distort analysis, so it was filled with appropriate measures like the mean, median, or mode.

Removing Duplicates: Duplicate records were removed to avoid redundancy.

Data Standardization: Column names and data types were standardized for consistency.



```
# import csv file
df = pd.read_csv('Diwali Sales Data.csv', encoding= 'unicode_escape')
```

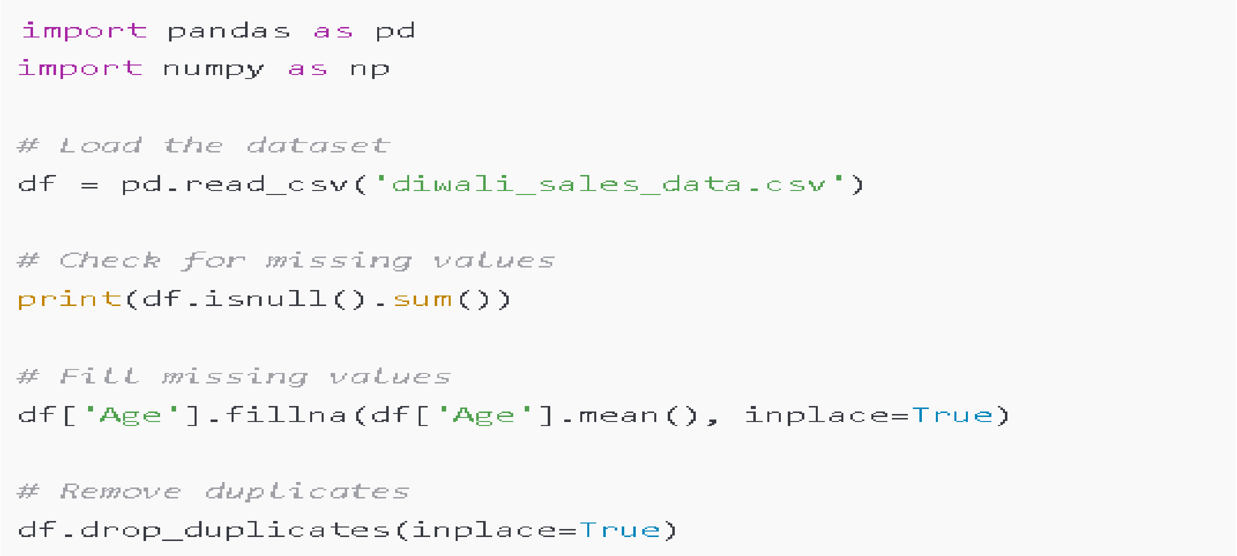
```
df.shape
```

```
4]: (11251, 15)
```

```
df.head(5)
```

```
#drop unrelated/blank columns
df.drop(['Status', 'unnamed1'], axis=1, inplace=True)
```

```
df.info()
```



```
import pandas as pd
import numpy as np

# Load the dataset
df = pd.read_csv('diwali_sales_data.csv')

# Check for missing values
print(df.isnull().sum())

# Fill missing values
df['Age'].fillna(df['Age'].mean(), inplace=True)

# Remove duplicates
df.drop_duplicates(inplace=True)
```

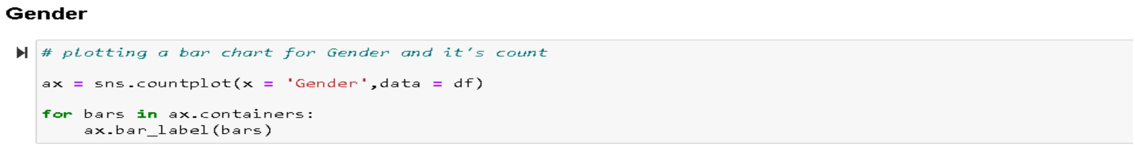
Output: After data cleaning, the dataset was ready for analysis

5.2 Exploratory Data Analysis (EDA)

EDA is performed to summarize the main characteristics of the data and visualize patterns.

5.2.1 Gender-wise Sales Analysis

Analyzing sales by gender helps in understanding purchasing patterns across males and females..

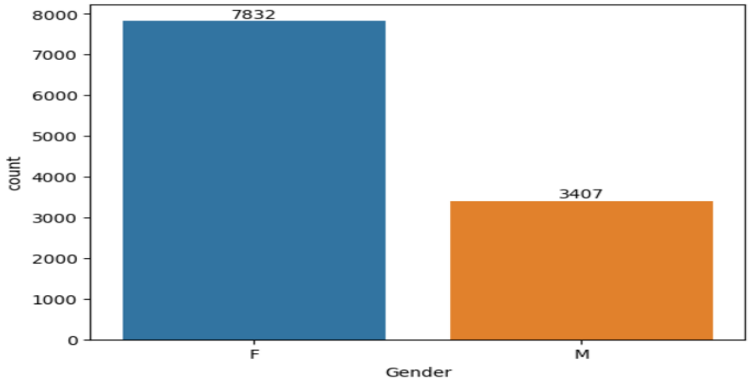


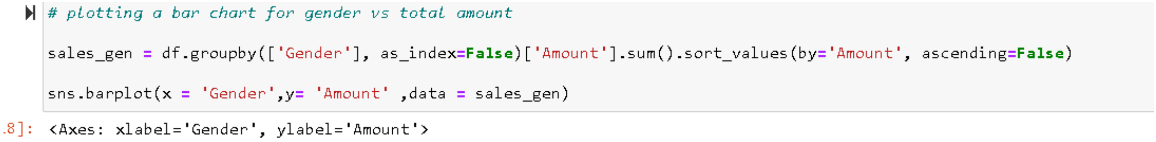
Gender

```
# plotting a bar chart for Gender and it’s count

ax = sns.countplot(x = 'Gender',data = df)

for bars in ax.containers:
    ax.bar_label(bars)
```



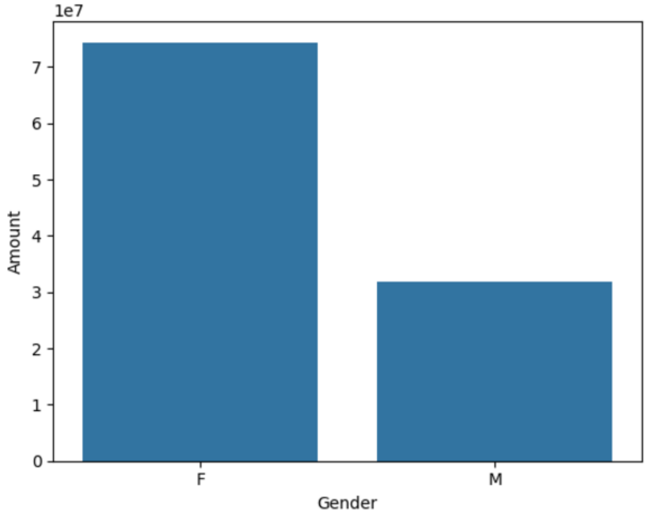


```
# plotting a bar chart for gender vs total amount

sales_gen = df.groupby(['Gender'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascending=False)

sns.barplot(x = 'Gender',y= 'Amount' ,data = sales_gen)
```

```
.8]: <Axes: xlabel='Gender', ylabel='Amount'>
```

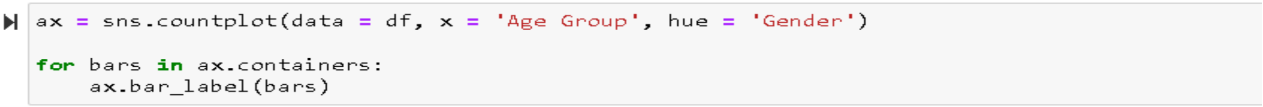


From above graphs we can see that most of the buyers are females and even the purchasing power of females are greater than men

5.2.2 Age-wise Sales Analysis

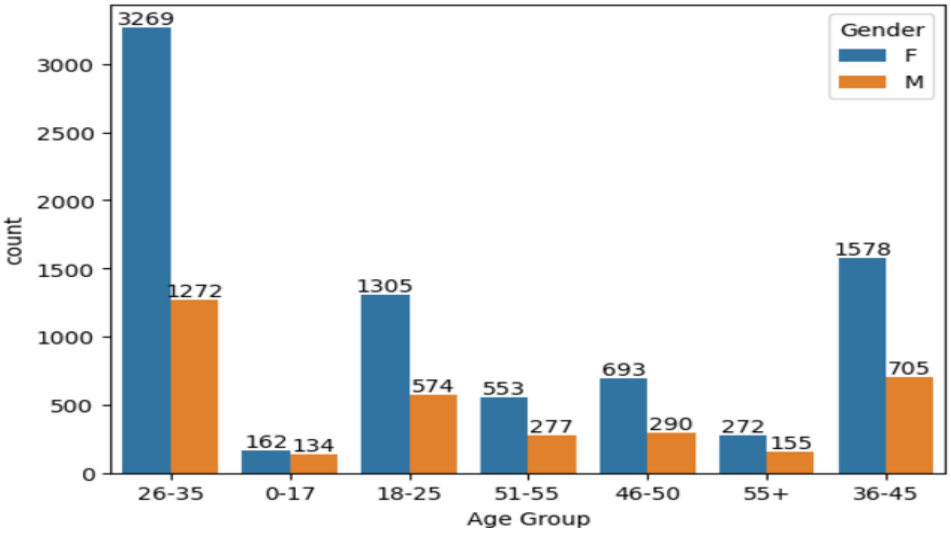
Identifying which age groups contribute the most to sales.

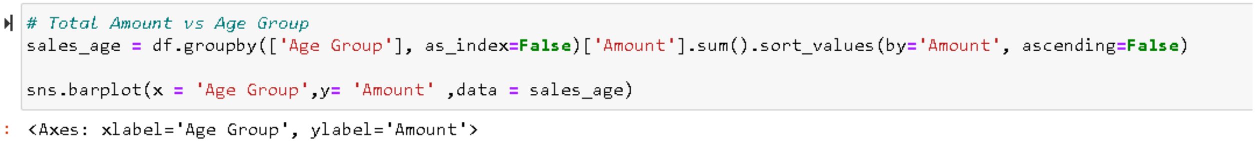
Age



```
ax = sns.countplot(data = df, x = 'Age Group', hue = 'Gender')

for bars in ax.containers:
    ax.bar_label(bars)
```

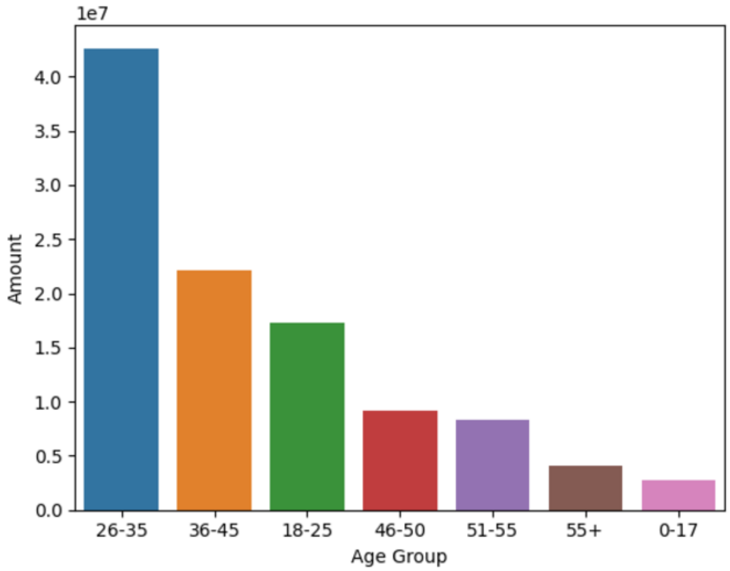




```
# Total Amount vs Age Group
sales_age = df.groupby(['Age Group'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascending=False)

sns.barplot(x = 'Age Group',y= 'Amount' ,data = sales_age)
```

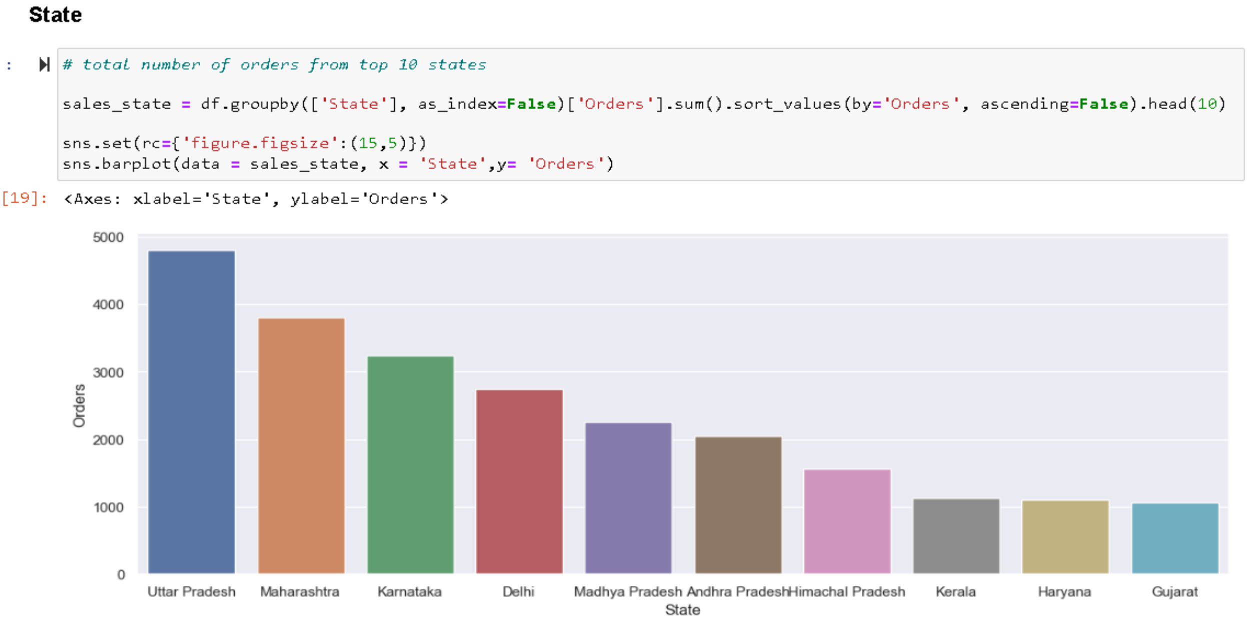
```
: <Axes: xlabel='Age Group', ylabel='Amount'>
```



From above graphs we can see that most of the buyers are of age group between 26-35 yrs female

5.2.3 State-wise Sales Analysis

Analyzing sales performance across different states to identify top-performing regions.



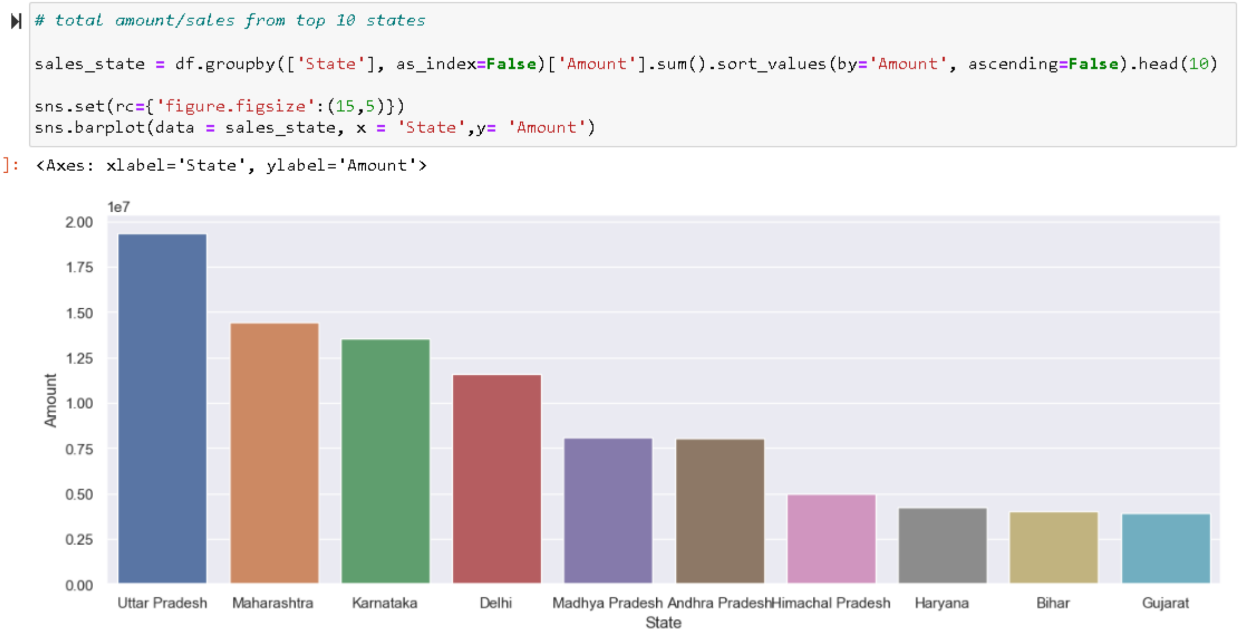
State

```
# total number of orders from top 10 states

sales_state = df.groupby(['State'], as_index=False)['Orders'].sum().sort_values(by='Orders', ascending=False).head(10)

sns.set(rc={'figure.figsize':(15,5)})
sns.barplot(data = sales_state, x = 'State',y= 'Orders')
```

```
[19]: <Axes: xlabel='State', ylabel='Orders'>
```



```
# total amount/sales from top 10 states

sales_state = df.groupby(['State'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascending=False).head(10)

sns.set(rc={'figure.figsize':(15,5)})
sns.barplot(data = sales_state, x = 'State',y= 'Amount')
```

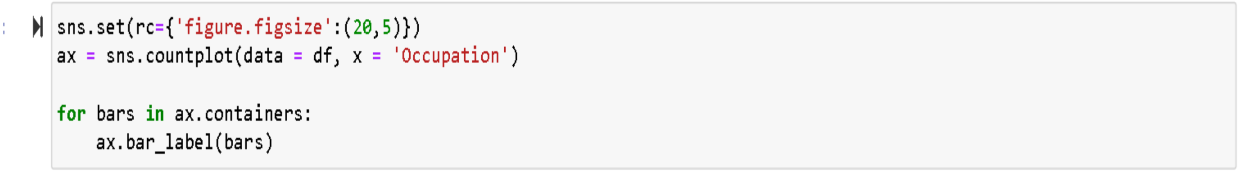
```
]: <Axes: xlabel='State', ylabel='Amount'>
```

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

5.2.4 Occupation-wise Sales Analysis

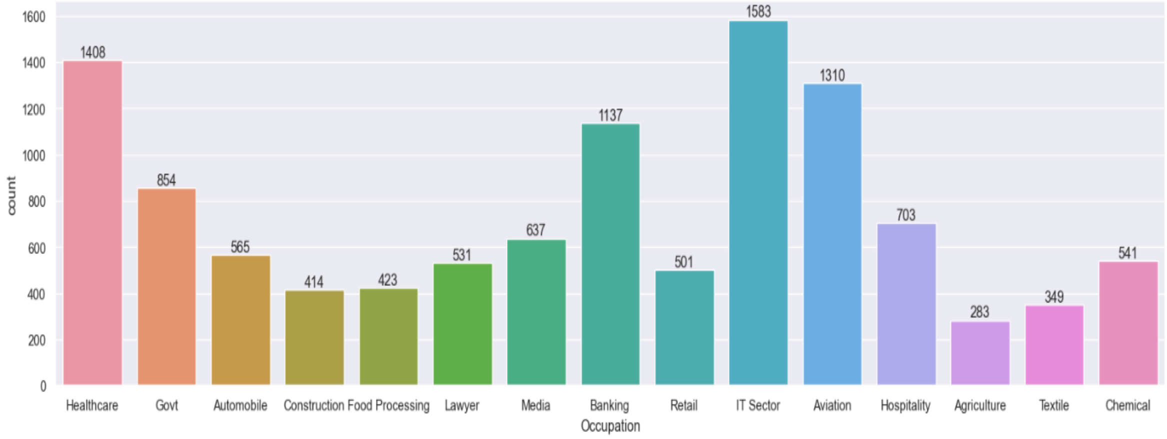
Identify the occupations that contribute the most to Diwali sales.

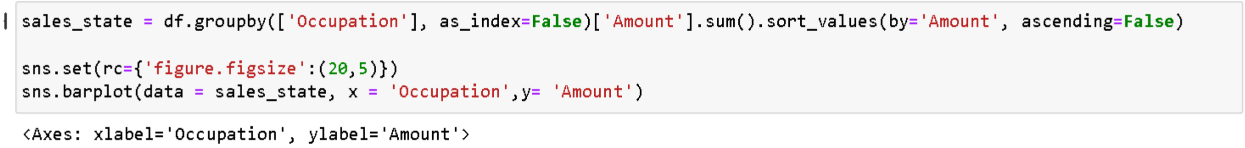
Occupation



```
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Occupation')

for bars in ax.containers:
    ax.bar_label(bars)
```

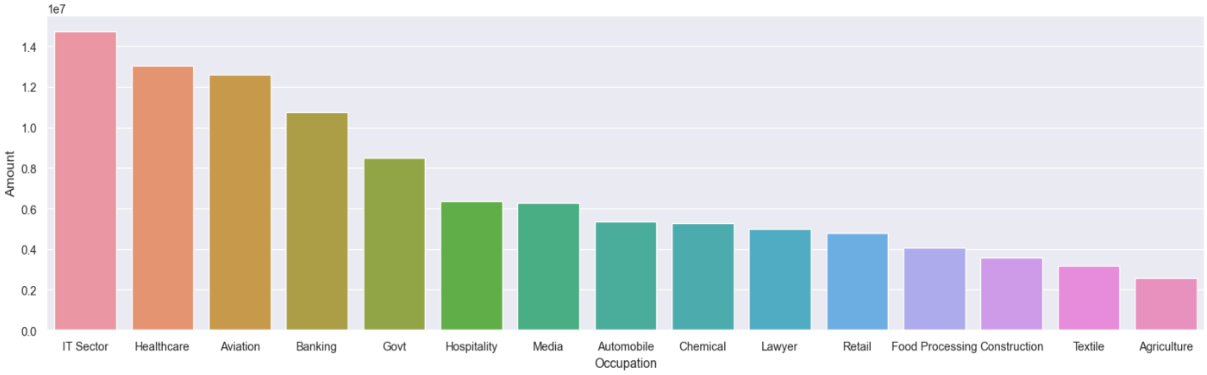




```
sales_state = df.groupby(['Occupation'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascending=False)

sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = sales_state, x = 'Occupation',y= 'Amount')
```

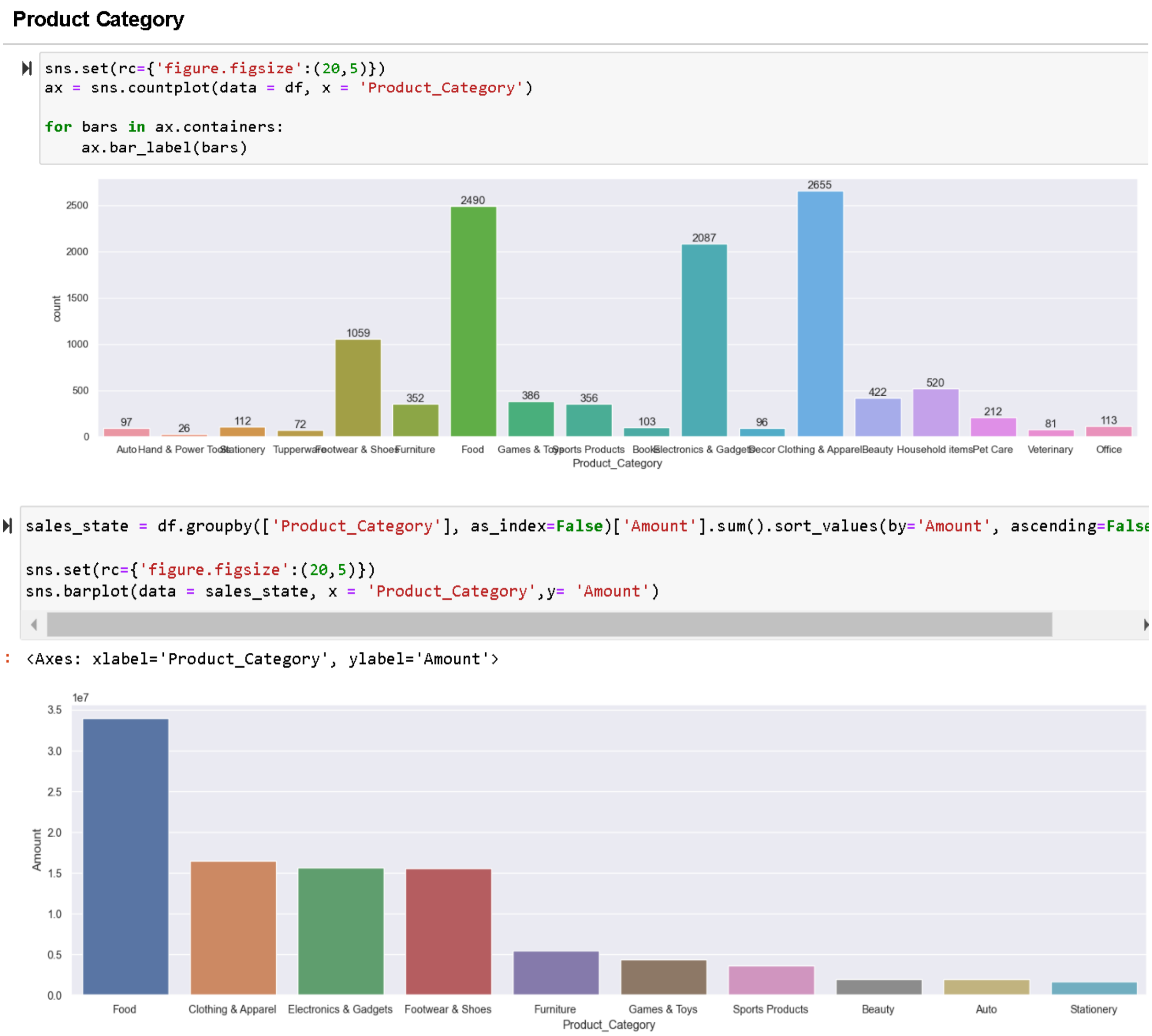
```
<Axes: xlabel='Occupation', ylabel='Amount'>
```



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

5.3 Product and Category Analysis

Identifying the top-selling product categories and products helps in inventory planning and demand forecasting.



Product Category

```
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Product_Category')

for bars in ax.containers:
    ax.bar_label(bars)
```

```
sales_state = df.groupby(['Product_Category'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascending=False

sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = sales_state, x = 'Product_Category',y= 'Amount')
```

```
: <Axes: xlabel='Product_Category', ylabel='Amount'>
```

From above graphs we can see that most of the sold products are from Food, Clothing and Electronics category

Conclusion

The Diwali Sales Analysis project highlights the transformative power of data analytics in driving strategic business decisions. Through in-depth analysis, we identified a key customer segment: married women aged 26-35 from Uttar Pradesh, Maharashtra, and Karnataka, working in IT, healthcare, and aviation, who are more likely to purchase products from the food, clothing, and electronics categories.

By leveraging Python libraries for data cleaning and analysis, we refined inventory planning to meet demand accurately and enhanced the overall customer experience. These actionable insights empower businesses to maximize profits during festive seasons, optimize resource allocation, and develop personalized marketing strategies tailored to high-potential customers.

Ultimately, this project underscores the value of data-driven decision-making in fostering long-term growth, improving customer loyalty, and driving business success.

Future Scope of the Project

Integration with Real-Time Data: Future iterations can integrate real-time sales data for dynamic insights and faster decision-making.

Machine Learning Models: Implement predictive models to forecast sales trends and customer preferences.

Customer Sentiment Analysis: Use natural language processing (NLP) techniques to analyze customer feedback and reviews for deeper insights.

Recommendation Systems: Develop personalized product recommendations based on customer purchase history.

Automation: Automate data cleaning, analysis, and reporting processes to enhance efficiency.

Scalability: Extend the analysis to include multiple festive seasons and across different regions for broader business insights.

8 Reference

1. https://www.scribd.com/document/709808482/Diwali-Sales-Analysis-EDA-1696347982

2. https://www.academia.edu/RegisterToDownload/primaryOccupation