

1. Summary of overall learning outcome of the internship

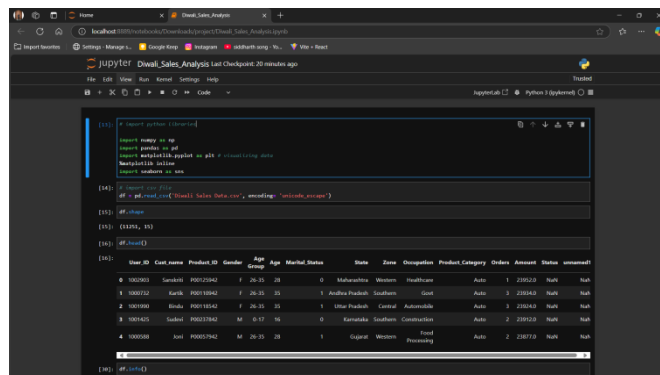
The Diwali Sales Analysis project aims to understand customer behavior during the Diwali season by analyzing real sales data. Using Python's powerful libraries like Pandas, NumPy, Matplotlib, and Seaborn, the project covers:

- Data cleaning
- Data transformation
- Exploratory Data Analysis (EDA)
- Visual storytelling through graphs and charts.

The goal is to find insights about who buys the most, what sells the most, and which groups have the highest purchasing power during Diwali.

.1.1 Step-by-Step Code Explanation

🔧 Importing Python Libraries



```
[1]: Import python Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

[2]: Import raw data
df = pd.read_csv('Diwali_Sales_Data.csv', encoding='unicode_escape')

[3]: df.shape
(11251, 15)

[4]: df.head(5)
```

User ID	Customer	Product ID	Gender	Age Group	Married Status	State	Zone	Occupation	Product Category	Order	Amount	Status	Unnamed: 0
0	1000001	Product1	F	20-25	0	Madhya Pradesh	Western	Healthcare	Auto	1	230120	NaN	NaN
1	1000002	Product2	F	26-35	0	Andhra Pradesh	Southern	Health	Auto	1	230140	NaN	NaN
2	1000003	Product3	F	26-35	0	Uttar Pradesh	Central	Automobile	Auto	1	230140	NaN	NaN
3	1000004	Product4	M	0-17	0	Karnataka	Southern	Construction	Auto	2	230120	NaN	NaN
4	1000005	Product5	M	26-35	0	Gujarat	Western	Food Processing	Auto	2	230170	NaN	NaN

🔧 Loading the Data

- Reading CSV: Loads the Diwali sales data into a DataFrame called df.
- Encoding: Some special characters in names/states need 'unicode_escape' encoding to avoid errors.
- df.shape: Tells us the number of rows and columns (11251 rows × 15 columns). Purpose: Bring raw data into Python for analysis.

🔧 Initial Data Inspection

- head(): Displays the first 5 rows to quickly inspect the structure and content.
- info(): Provides the datatypes and null value counts for each column.

Findings:

- 15 columns exist, but 2 (Status, unnamed1) are blank or unused.
- Amount column has some null values.

```

import pandas as pd

df = pd.read_csv('path/to/data.csv')

df.head()

```

User_ID	Gender	Product_ID	Gender	Age_Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	
1000700	Female	P00110042	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	239032	
1000702	Female	P00110042	F	26-35	28	1	Andhra Pradesh	Southern	Food	Auto	2	239034	
1000700	Male	P00110042	F	26-35	28	1	Tamil Nadu	Central	Automobile	Auto	3	239040	
1000702	Male	P00110042	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	239110	
1000700	Male	P00110042	M	26-35	28	1	Gujarat	Western	Food	Prescription	Auto	2	239710

```

df.info()

```

```

<class 'pandas.core.frame.DataFrame'>
Int64Index: 8, dtype: int64
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype  
---  -
 0   User_ID           8 non-null      int64   
 1   Gender            8 non-null      object  
 2   Product_ID        8 non-null      int64   
 3   Gender            8 non-null      object  
 4   Age_Group         8 non-null      object  
 5   Age              8 non-null      int64   
 6   Marital_Status    8 non-null      int64   
 7   State            8 non-null      object  
 8   Zone             8 non-null      object  
 9   Occupation        8 non-null      object  
10   Product_Category  8 non-null      object  
11   Orders            8 non-null      int64   
12   Amount           8 non-null      int64   
dtypes: int64(8), object(5)
memory usage: 3.1+ MB

```

Cleaning the Data

```

# Drop irrelevant columns
df.drop(['Gender', 'Product_ID', 'Gender'], axis=1, inplace=True)

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

# Drop rows where Amount is missing
df.dropna(inplace=True)

# Rename columns
df.rename(columns={'Marital_Status': 'Marital'})

df.head()

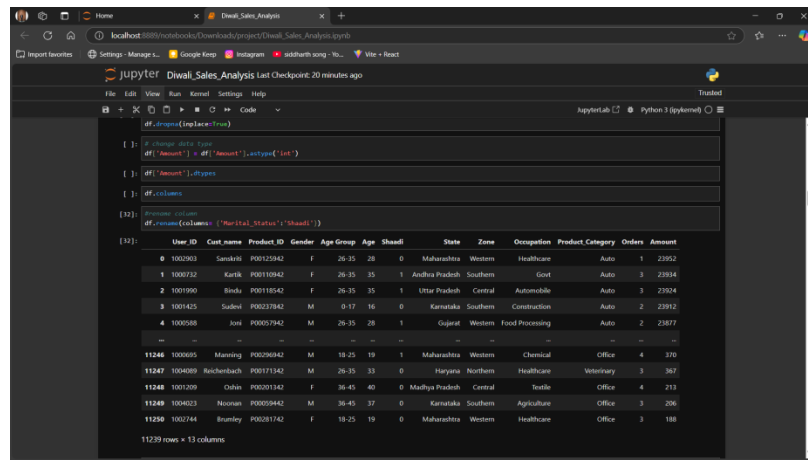
```

User_ID	Gender	Product_ID	Age_Group	Age	Marital	State	Zone	Occupation	Product_Category	Orders	Amount	
1000700	Female	P00110042	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	239032
1000702	Female	P00110042	F	26-35	28	1	Andhra Pradesh	Southern	Food	Auto	2	239034

- Removes the irrelevant columns to clean up the dataset.
- Checks for missing values.
- Drops any rows where Amount is missing.

Renaming Columns

- Marital_Status is renamed internally for better readability.



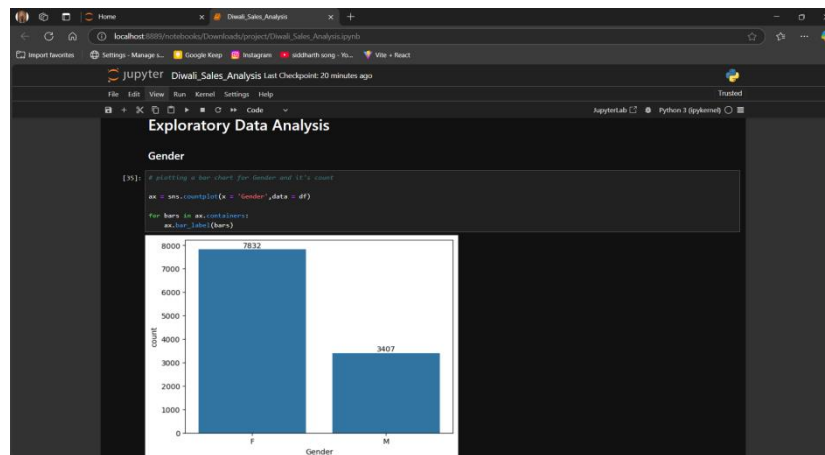
Descriptive Statistics

- Calculates mean, standard deviation, minimum, maximum, and quartiles.
- Focuses on numeric columns like Age, Orders, and Amount.

1.2 Exploratory Data Analysis (EDA)

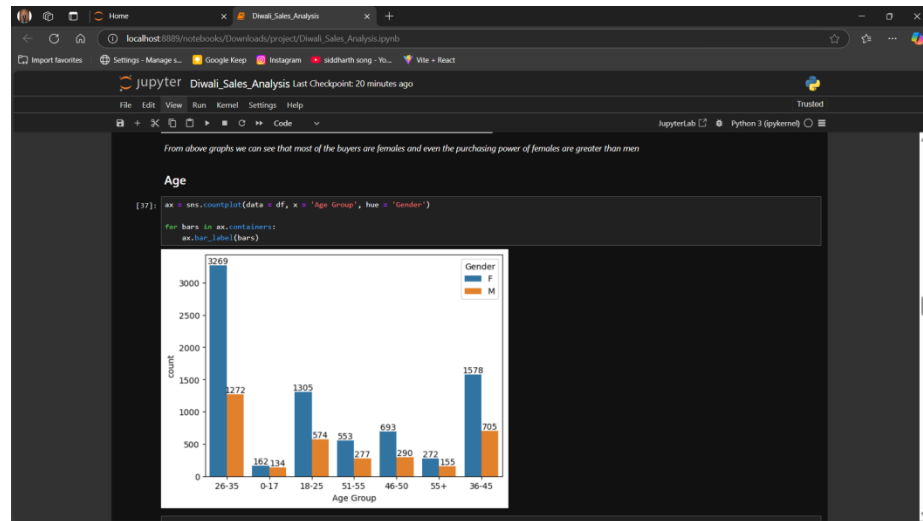
Gender Analysis

- Gender Count Plot
- Creates a bar chart showing the count of male and female customers.
- Shows which gender spent more money.

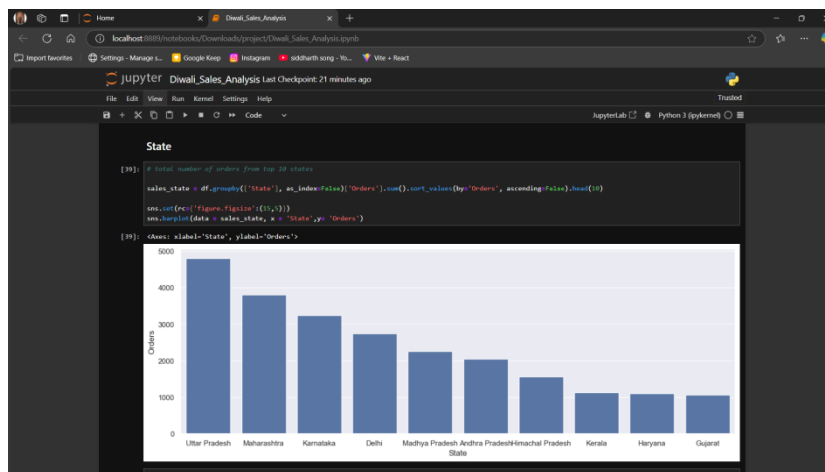


Gender vs. Total Amount

- Shows which gender spent more money.
- Females made more purchases and spent more overall compared to males



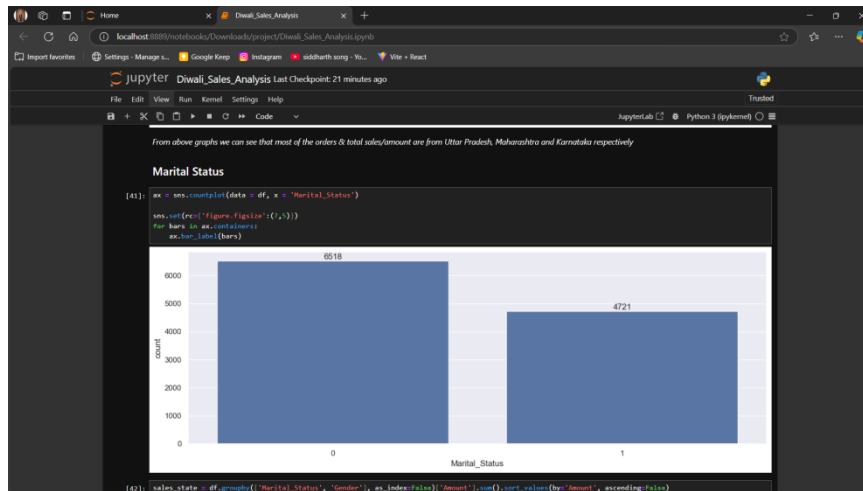
- State-wise Analysis
- Top States by Orders
- Top States by Total Amount
 - Insights:
- Uttar Pradesh, Maharashtra, and Karnataka are the top contributors in terms of both orders and revenue.



Marital Status Analysis

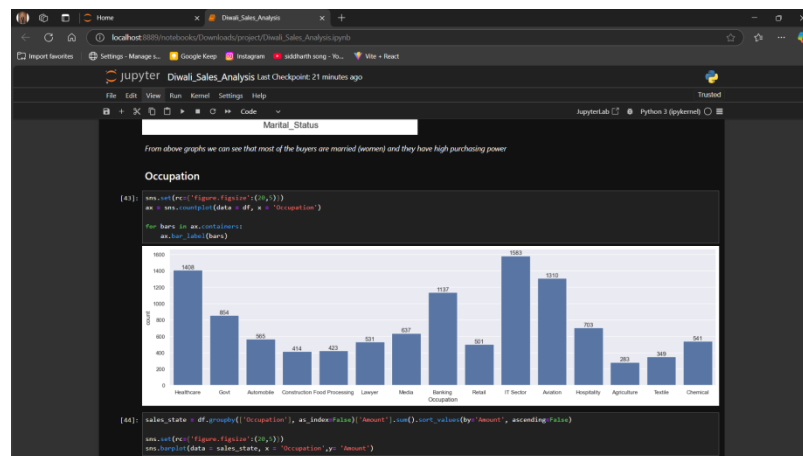
- Marital Status Count Plot
- Marital Status vs. Amount
- Insights:
- Married women spent the most.

- Marital status greatly impacts purchasing behavior.



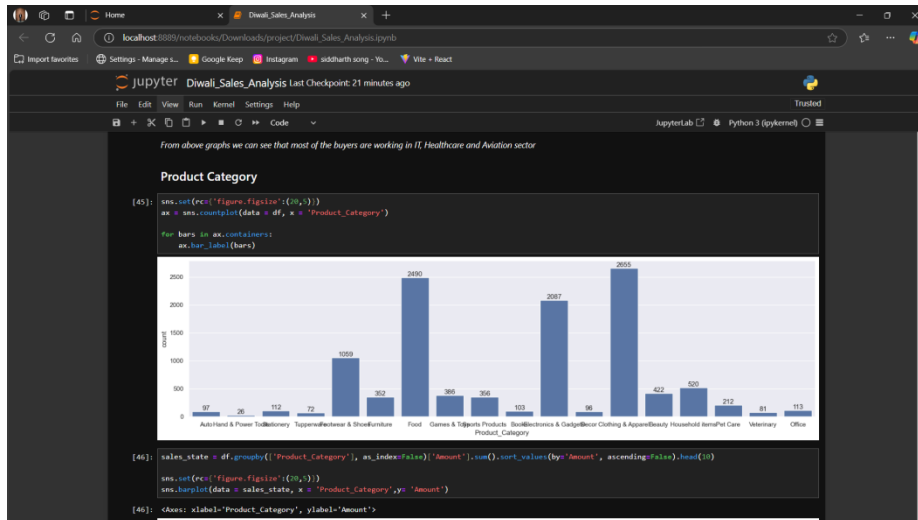
🌈 Occupation Analysis

- People working in IT, Healthcare, and Aviation sectors spend the most during Diwali.



🌈 Product Category Analysis

- Product Category Count
- Product Category by Sales Amount



Product Category Count Plot

- Purpose: Understand what categories customers are buying.
- Observation:
Food, Clothing, and Electronics are the most popular.

Product Category vs. Amount Plot

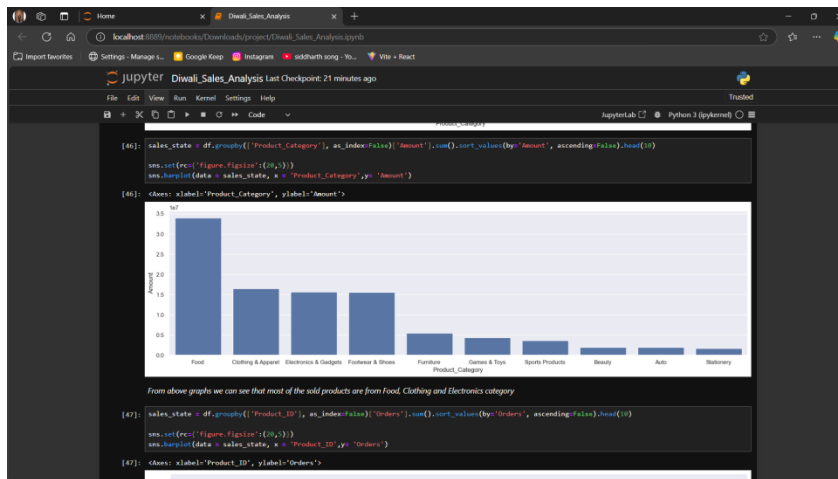
- Purpose: Understand revenue generation per category.
- Observation:
Electronics generate a huge share of total revenue.

Business Application:

- Increase inventory and promotions on Food, Clothing, and Electronics categories during Diwali.
- Bundle offers (e.g., "Buy Food + Electronics and get a discount").

Possible Improvements:

- Study profitability per product category (not just revenue).



1.3 Top Selling Products (Expanded)

- Purpose: Identify which specific products are bestsellers.
- Observation:
Certain product IDs have extremely high order counts.
- Business Application:
 - Stock more of these top-selling products.
 - Highlight them in advertisements and online platforms.
- Possible Improvements:
 - Find cross-sell patterns (what people often buy together).
 - Study return rates for top-selling products.

