SALES DATA ANALYSIS PROJECT

Uncovering Customer Trends, Product Demand & Regional Performance from Diwali Sales Data

Introduction

This project investigates over 11,000 transactions across diverse demographics and geographies in India. The purpose is to extract insights from Diwali sales data and solve key business problems using data.

Objective:

The objective of this analysis is to perform an in-depth Exploratory Data Analysis (EDA) on Diwali sales data collected from various customers across India. The aim is to uncover patterns, trends, and actionable insights that can support business decisions in areas such as:

- Marketing strategies
- Customer segmentation
- Product demand forecasting

Business Problems Solved Through EDA

Through Exploratory Data Analysis, the following business problems are addressed:

- Identify the most profitable customer demographics (age, gender, marital status, occupation).
- Understand regional sales performance to improve localized promotions.
- Determine top-selling product categories and associated sales volumes.
- Evaluate the impact of customer behavior (orders, purchase amount) on revenue.
- Discover underperforming segments for targeted improvement.

Tools & Technology Used

- **Python:** Data processing and analysis
- **Pandas:** DataFrame operations
- **Matplotlib & Seaborn:** Visualization tools
- **Jupyter Notebook:** Interactive environment for developing insights

Project Flow

1. **Data Load & Initial Inspection:**

- Loaded the CSV dataset using Pandas
- Previewed rows, checked column types and missing values
- Summarized numeric data to get initial distribution

2. Data Cleaning:

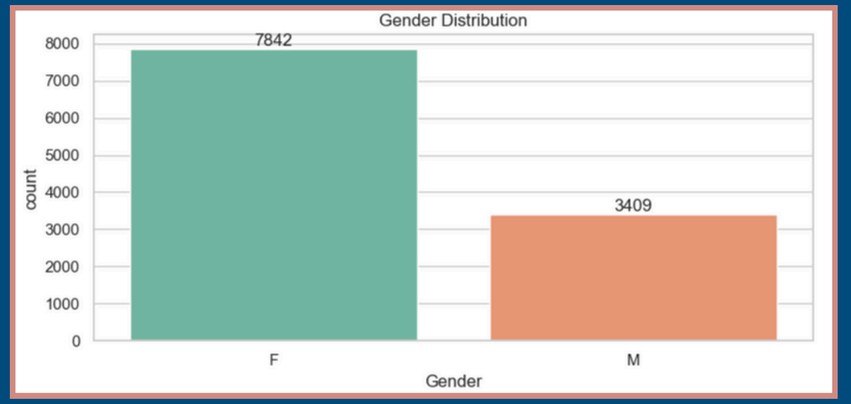
- Dropped unnecessary columns (Status, unnamed1)
- Removed rows with missing Amount values
- Ensured proper data types (e.g., Marital_Status as integer)

3. Exploratory Data Analysis (EDA)

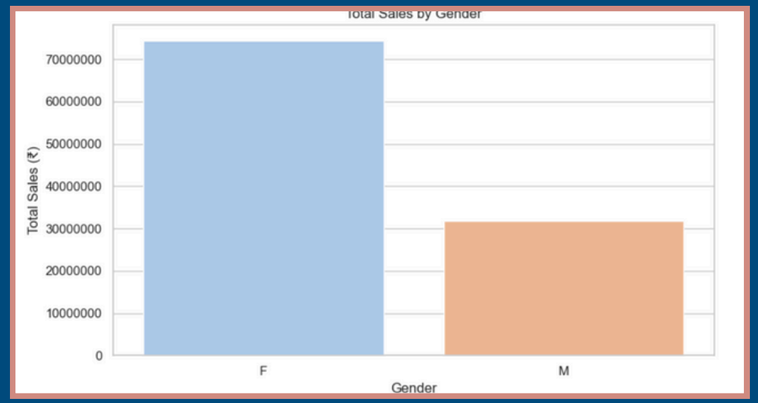
Exploratory Data Analysis (EDA)

1. Gender Distribution & Gender-wise Total Sales:

```
plt.figure(figsize=(8, 4))
ax = sns.countplot(data = df, x = 'Gender', order=df['Gender'].value_counts().index, palette='Set2')

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

plt.title("Gender Distribution")
plt.tight_layout()
plt.show()
```




```
plt.figure(figsize=(8,5))
gender_sales = df_clean.groupby('Gender')['Amount'].sum().reset_index()
sns.barplot(data=gender_sales, x='Gender', y='Amount', palette='pastel')
plt.title('Total Sales by Gender')
plt.ylabel('Total Sales (₹)')
plt.xlabel('Gender')
plt.tight_layout()
plt.gca().ticklabel_format(style='plain', axis='y')
plt.show()
```



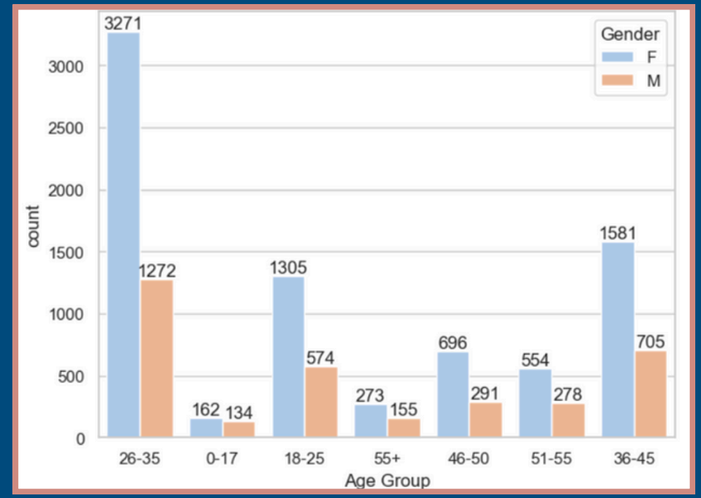
Insights:

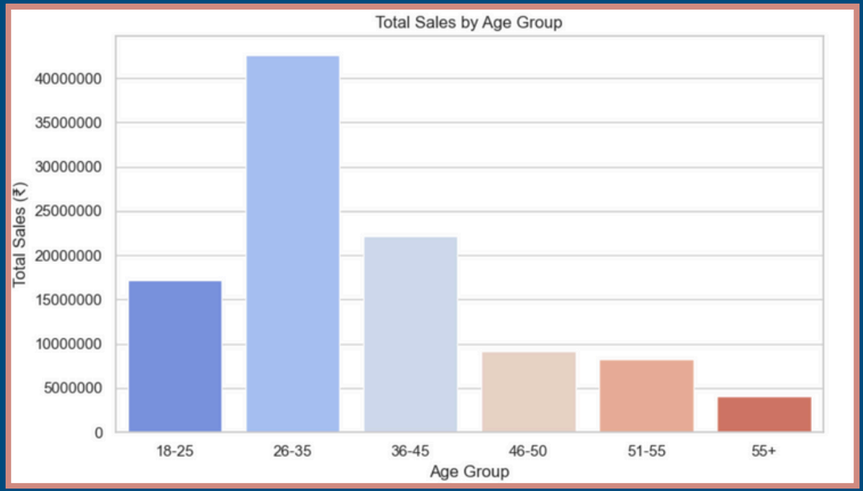
- Female customers show comparable purchase volume but slightly lower spending per transaction than males.
- Indicates potential to target female audiences with bundled or value-based offers.

Exploratory Data Analysis (EDA)

2. Age Group & Age Group-wise Sales:

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

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

plt.tight_layout()
plt.show()
```






```
plt.figure(figsize=(8,5))
age_sales = df_clean.groupby('Age Group')['Amount'].sum().reset_index()
order = ['18-25', '26-35', '36-45', '46-50', '51-55', '55+']
sns.barplot(data=age_sales, x='Age Group', y='Amount', order=order, palette='coolwarm')
plt.title('Total Sales by Age Group')
plt.ylabel('Total Sales (₹)')
plt.xlabel('Age Group')
plt.tight_layout()
plt.gca().ticklabel_format(style='plain', axis='y')
plt.show()
```

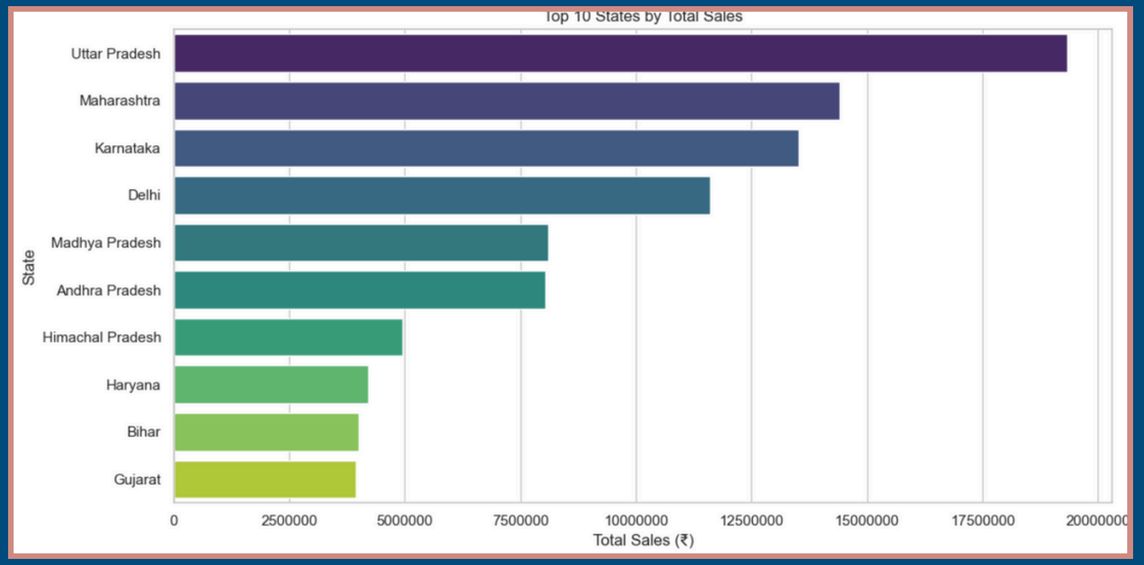
Insights:

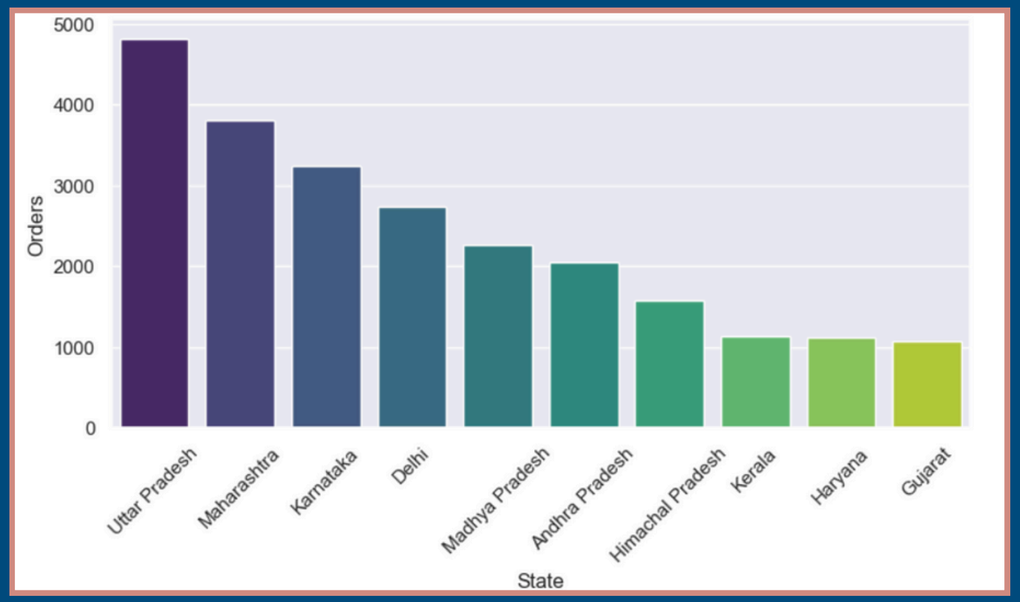
- 26-35 age group is the most active in both orders and spending, followed by 18-25.
- Suggests campaigns should focus on young working professionals.

Exploratory Data Analysis (EDA)

3. Top 10 States by Order & Sales:

```
plt.figure(figsize=(12,6))
state_sales = df_clean.groupby('State')['Amount'].sum().sort_values(ascending=False).head(10).reset_index()
sns.barplot(data=state_sales, y='State', x='Amount', palette='viridis')
plt.title('Top 10 States by Total Sales')
plt.xlabel('Total Sales (₹)')
plt.ylabel('State')
plt.gca().ticklabel_format(style='plain', axis='x')
plt.tight_layout()
plt.show()
```






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

sns.set(rc={'figure.figsize':(8,5)})
sns.barplot(data = sales_state, x = 'State',y= 'Orders', palette='viridis')
plt.xticks(rotation=45)
plt.tight_layout()
```

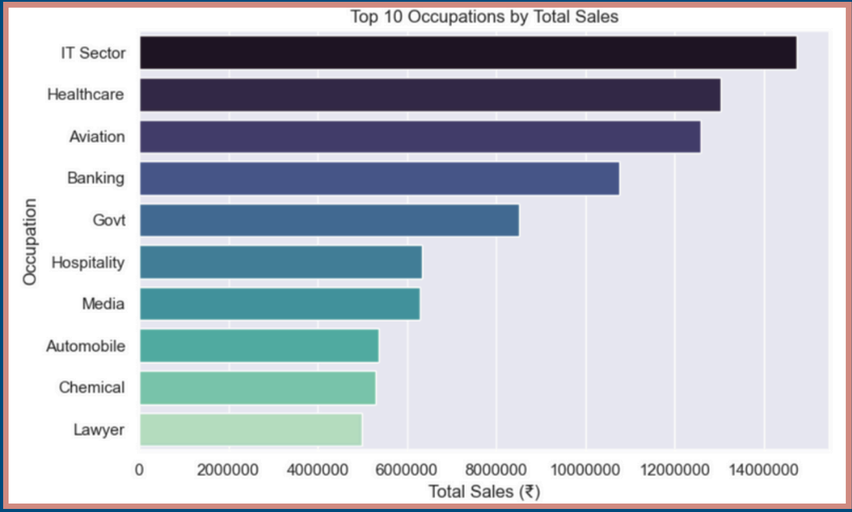
Insights:

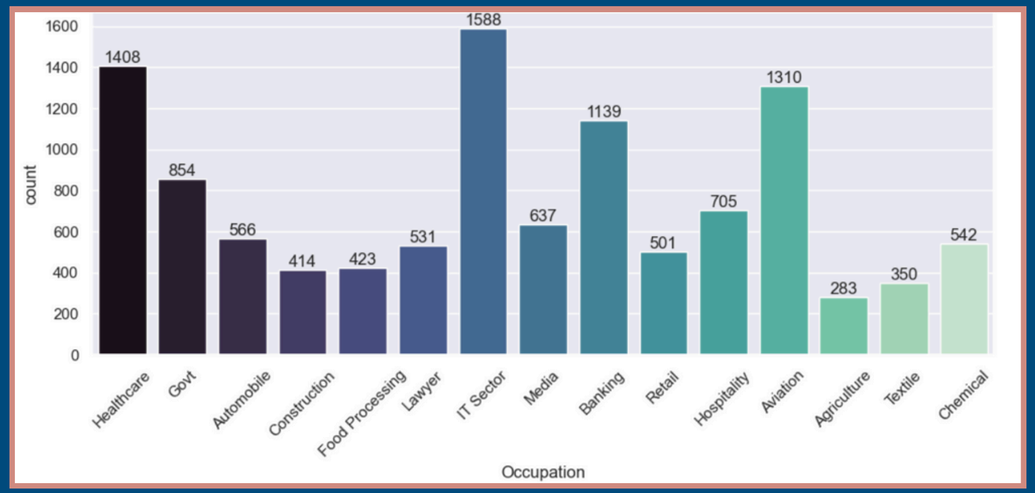
- Uttar Pradesh, Maharashtra, and Karnataka are consistently the top three states in terms of both orders placed and total sales amount. This indicates they are the most significant markets and should be prioritized for marketing campaigns and logistical support.
- The high volume of sales and orders in these states suggests a strong consumer base and effective distribution channels.

Exploratory Data Analysis (EDA)

4. Occupation-wise Sales (Top 10):

```
sns.set(rc={'figure.figsize':(10,5)})
ax = sns.countplot(data = df, x = 'Occupation', palette='mako')

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

plt.xticks(rotation=45)
plt.tight_layout()
```






```
plt.figure(figsize=(8,5))
occupation_sales = df_clean.groupby('Occupation')['Amount'].sum().sort_values(ascending=False).head(10).reset_index()
sns.barplot(data=occupation_sales, y='Occupation', x='Amount', palette='mako')
plt.title('Top 10 Occupations by Total Sales')
plt.xlabel('Total Sales (₹)')
plt.ylabel('Occupation')
plt.tight_layout()
plt.gca().ticklabel_format(style='plain', axis='x')
plt.show()
```

Insights:

• Customers employed in the IT Sector, Healthcare, and Aviation demonstrate the highest spending power and contribute the most to overall sales.

• Targeted marketing campaigns and product offerings tailored to these professional groups could yield significant returns due to their high purchasing capacity.

Exploratory Data Analysis (EDA)

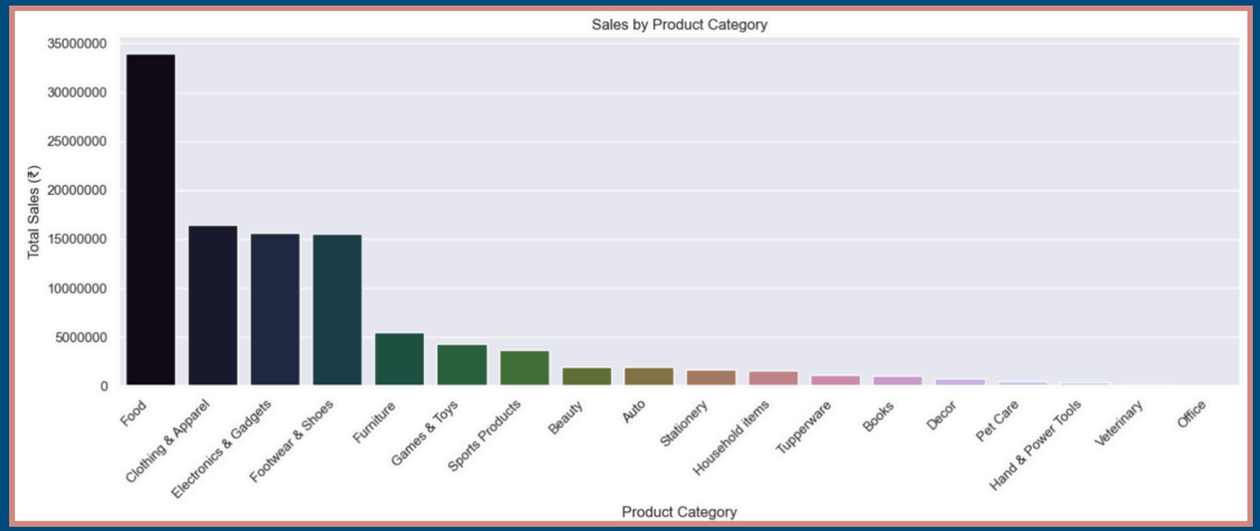
5. Product Category Sales:

```
plt.figure(figsize=(14,6))
category_sales = df_clean.groupby('Product_Category')['Amount'].sum().sort_values(ascending=False).reset_index()
sns.barplot(data=category_sales, x='Product_Category', y='Amount', palette='cubehelix')
plt.title('Sales by Product Category')
plt.ylabel('Total Sales (₹)')
plt.xlabel('Product Category')
plt.xticks(rotation=45, ha='right')
plt.gca().ticklabel_format(style='plain', axis='y')
plt.tight_layout()
plt.show()
```

Insights:

• "Clothing & Apparel," "Food," and "Electronics & Gadgets" are the top-performing product categories.

• These categories are key revenue drivers, suggesting that businesses should focus on maintaining diverse inventory, competitive pricing, and strong promotions for these items to maximize sales.



Exploratory Data Analysis (EDA)

6. Orders vs Amount Scatter Plot:

```
plt.figure(figsize=(8,6))
sns.scatterplot(data=df_clean, x='Orders', y='Amount', hue='Gender', palette='Set2', alpha=0.7)
plt.title('Orders vs. Amount by Gender')
plt.xlabel('Number of Orders')
plt.ylabel('Amount (₹)')
plt.legend(title='Gender')
plt.show()
```

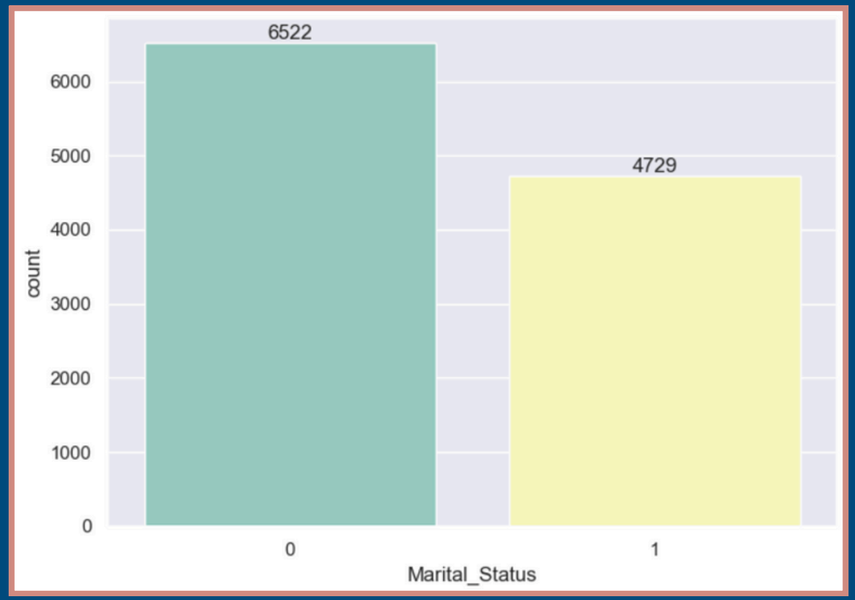
Insights:

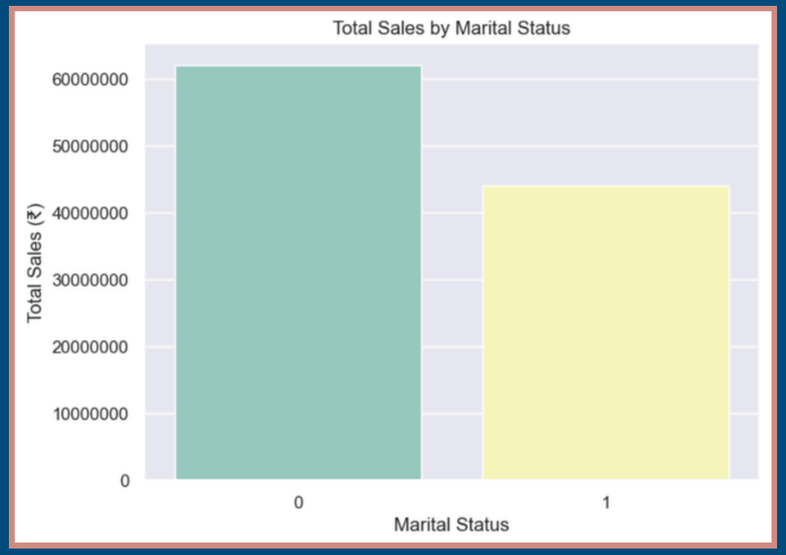
• The scatter plot likely indicates a positive correlation between the number of orders and the total amount spent, implying that customers who place more orders also tend to spend more. This could suggest that encouraging repeat purchases or higher order frequency would directly lead to increased revenue.

Exploratory Data Analysis (EDA)

7. Marital Status Distribution & Marital Status Sales:

```
ax = sns.countplot(data = df, x = 'Marital_Status', palette='Set3')
plt.tight_layout()

sns.set(rc={'figure.figsize':(7,5)})
for bars in ax.containers:
    ax.bar_label(bars)
```






```
plt.figure(figsize=(7,5))
marital_sales = df_clean.groupby('Marital_Status')['Amount'].sum().reset_index()
sns.barplot(data=marital_sales, x='Marital_Status', y='Amount', palette='Set3')
plt.title('Total Sales by Marital Status')
plt.ylabel('Total Sales (₹)')
plt.xlabel('Marital Status')
plt.gca().ticklabel_format(style='plain', axis='y')
plt.show()
```

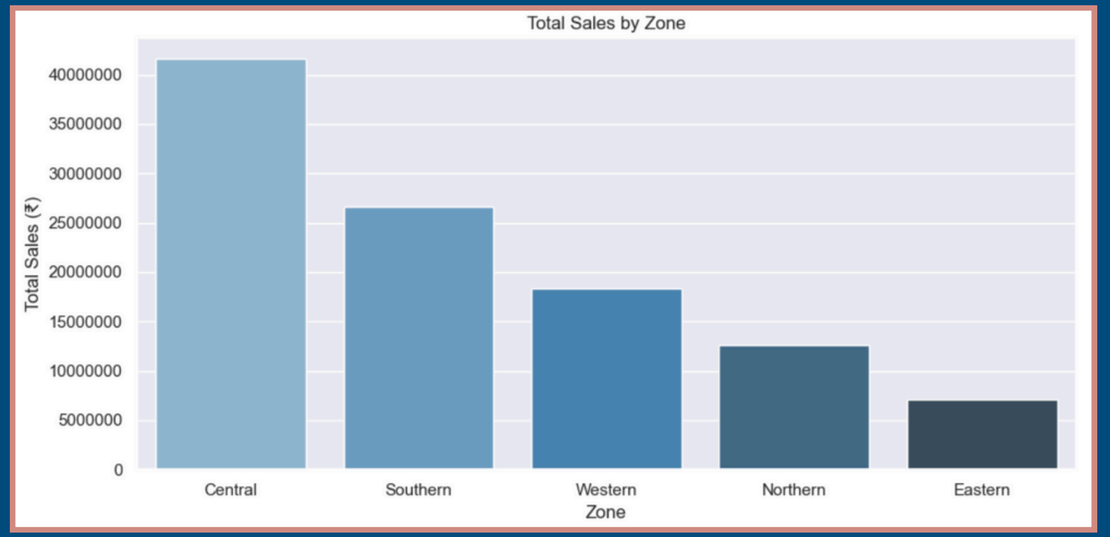
Insights:

- Married individuals contribute significantly more to sales compared to unmarried individuals. This trend is consistent across both notebooks.
- This suggests that a substantial portion of the sales revenue comes from married households.

Exploratory Data Analysis (EDA)

8. Zone-wise Sales:

```
plt.figure(figsize=(10,5))
zone_sales = df_clean.groupby('Zone')['Amount'].sum().sort_values(ascending=False).reset_index()
sns.barplot(data=zone_sales, x='Zone', y='Amount', palette='Blues_d')
plt.title('Total Sales by Zone')
plt.ylabel('Total Sales (₹)')
plt.xlabel('Zone')
plt.gca().ticklabel_format(style='plain', axis='y')
plt.tight_layout()
plt.show()
```



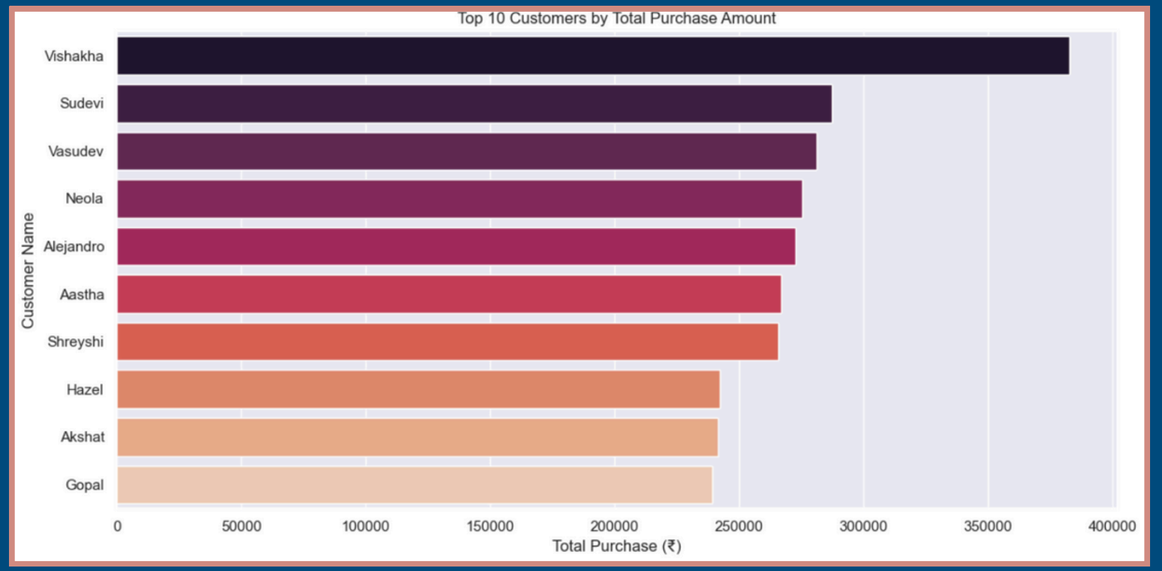
Insights:

- Central and Southern zones contribute the highest revenue.
- Indicates strong logistical networks or better promotional response in these zones.

Exploratory Data Analysis (EDA)

9. Top 10 Customers by Sales:

```
plt.figure(figsize=(12,6))
top_customers = df_clean.groupby('Cust_name')['Amount'].sum().sort_values(ascending=False).head(10).reset_index()
sns.barplot(data=top_customers, y='Cust_name', x='Amount', palette='rocket')
plt.title('Top 10 Customers by Total Purchase Amount')
plt.xlabel('Total Purchase (₹)')
plt.ylabel('Customer Name')
plt.gca().ticklabel_format(style='plain', axis='x')
plt.tight_layout()
plt.show()
```



Insights:

- Identifying the top 10 customers by sales allows businesses to recognize and reward their most valuable clients.

- Insights from these customers can be used to develop loyalty programs, personalized offers, and premium services to foster continued engagement and retention.

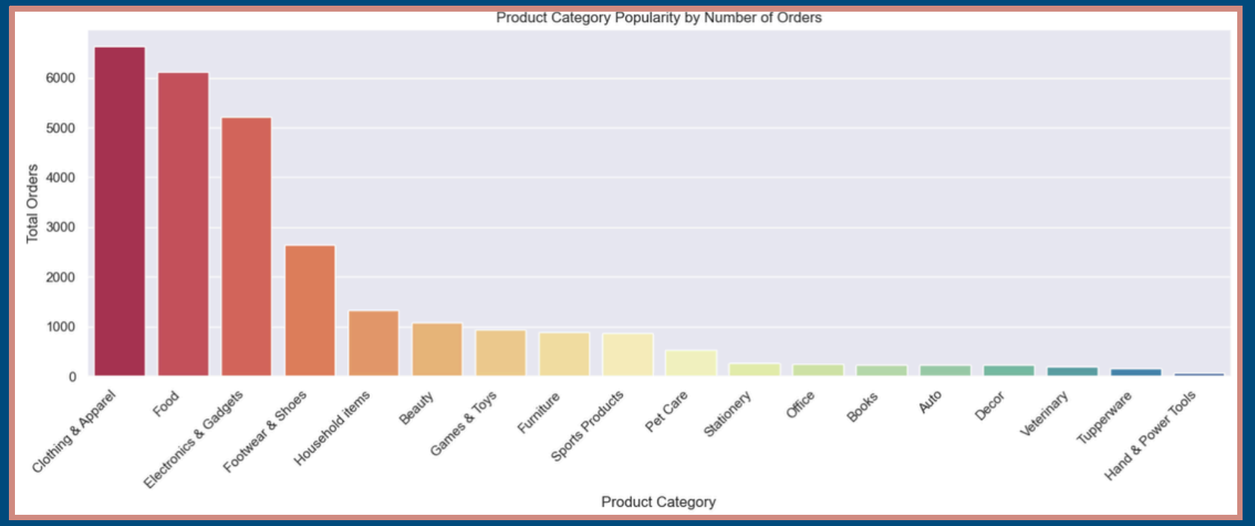
Exploratory Data Analysis (EDA)

10. Product Category Popularity by Orders:

```
plt.figure(figsize=(14,6))
category_orders = df_clean.groupby('Product_Category')['Orders'].sum().sort_values(ascending=False).reset_index()
sns.barplot(data=category_orders, x='Product_Category', y='Orders', palette='Spectral')
plt.title('Product Category Popularity by Number of Orders')
plt.ylabel('Total Orders')
plt.xlabel('Product Category')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

Insights:

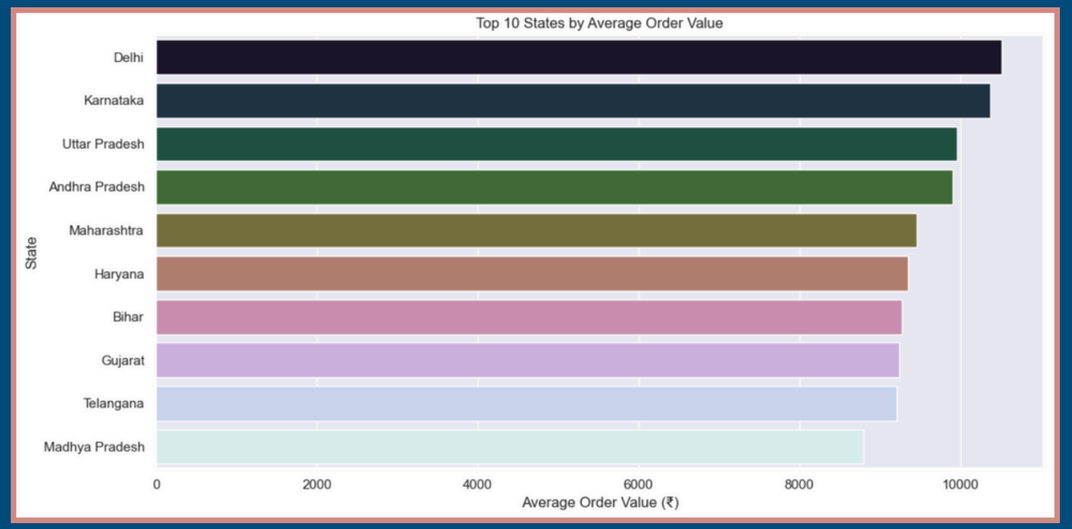
- Categories like Clothing & Apparel, Food and Electronics dominate sales.
- Brands in these categories can leverage the momentum during seasonal events.



Exploratory Data Analysis (EDA)

11. Average Order Value by State (Top 10):

```
plt.figure(figsize=(12,6))
state_avg_order = df_clean.groupby('State')['Amount'].mean().sort_values(ascending=False).head(10).reset_index()
sns.barplot(data=state_avg_order, y='State', x='Amount', palette='cubehelix')
plt.title('Top 10 States by Average Order Value')
plt.xlabel('Average Order Value (₹)')
plt.ylabel('State')
plt.tight_layout()
plt.show()
```



Insights:

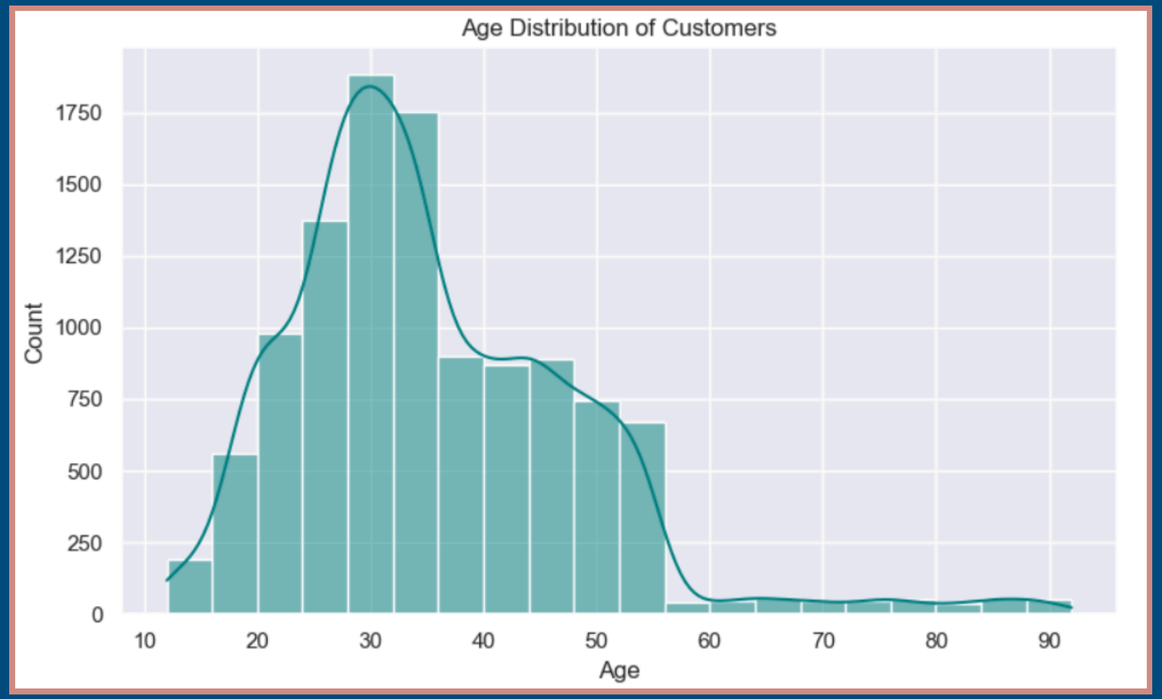
- Understanding the average order value (AOV) per state can highlight regions where customers spend more per transaction, even if the total order volume is not the highest.

- This insight can help in optimizing product bundling strategies and promotional efforts to increase the AOV in specific states.

Exploratory Data Analysis (EDA)

12. Age Distribution Histogram:

```
plt.figure(figsize=(8,5))
sns.histplot(df_clean['Age'], bins=20, kde=True, color='teal')
plt.title('Age Distribution of Customers')
plt.xlabel('Age')
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```



Insights:

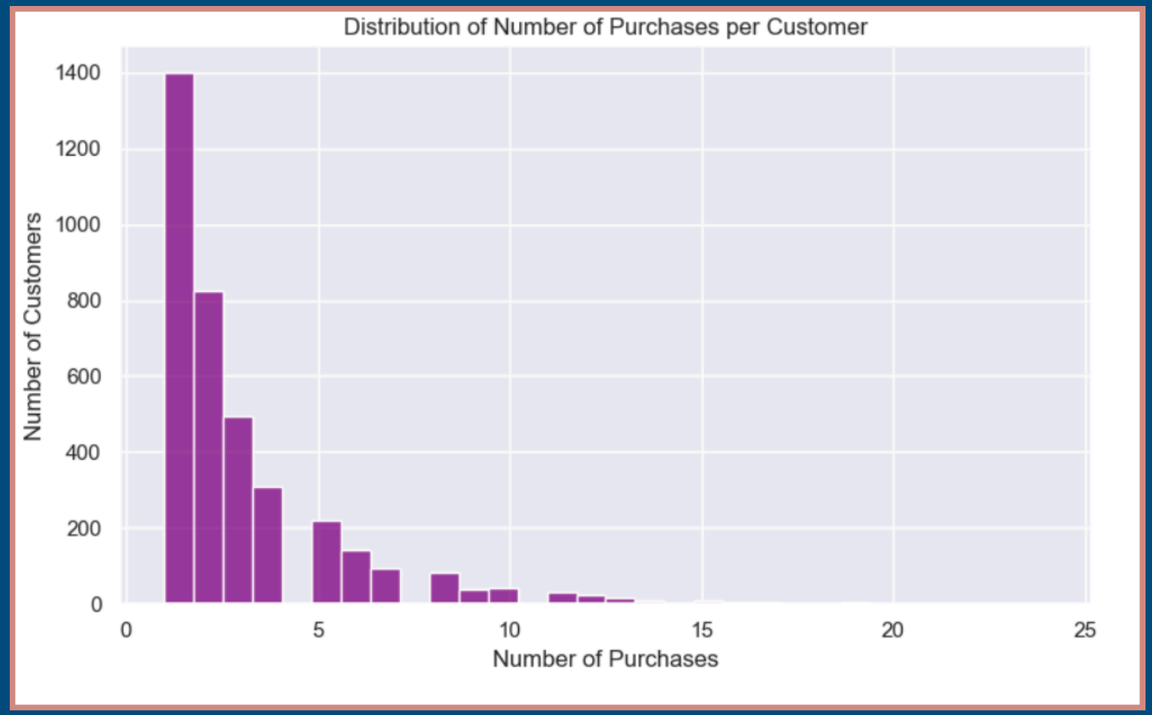
• The histogram of age distribution is likely to reveal that the 26-35 age group is the most dominant in terms of both customer count and sales contribution.

• This reinforces the importance of tailoring marketing and product strategies to this demographic, considering their preferences and purchasing habits.

Exploratory Data Analysis (EDA)

13. Repeat Customers Count Distribution:

```
plt.figure(figsize=(8,5))
repeat_counts = df_clean['User_ID'].value_counts()
sns.histplot(repeat_counts, bins=30, color='purple')
plt.title('Distribution of Number of Purchases per Customer')
plt.xlabel('Number of Purchases')
plt.ylabel('Number of Customers')
plt.tight_layout()
plt.show()
```



Insights:

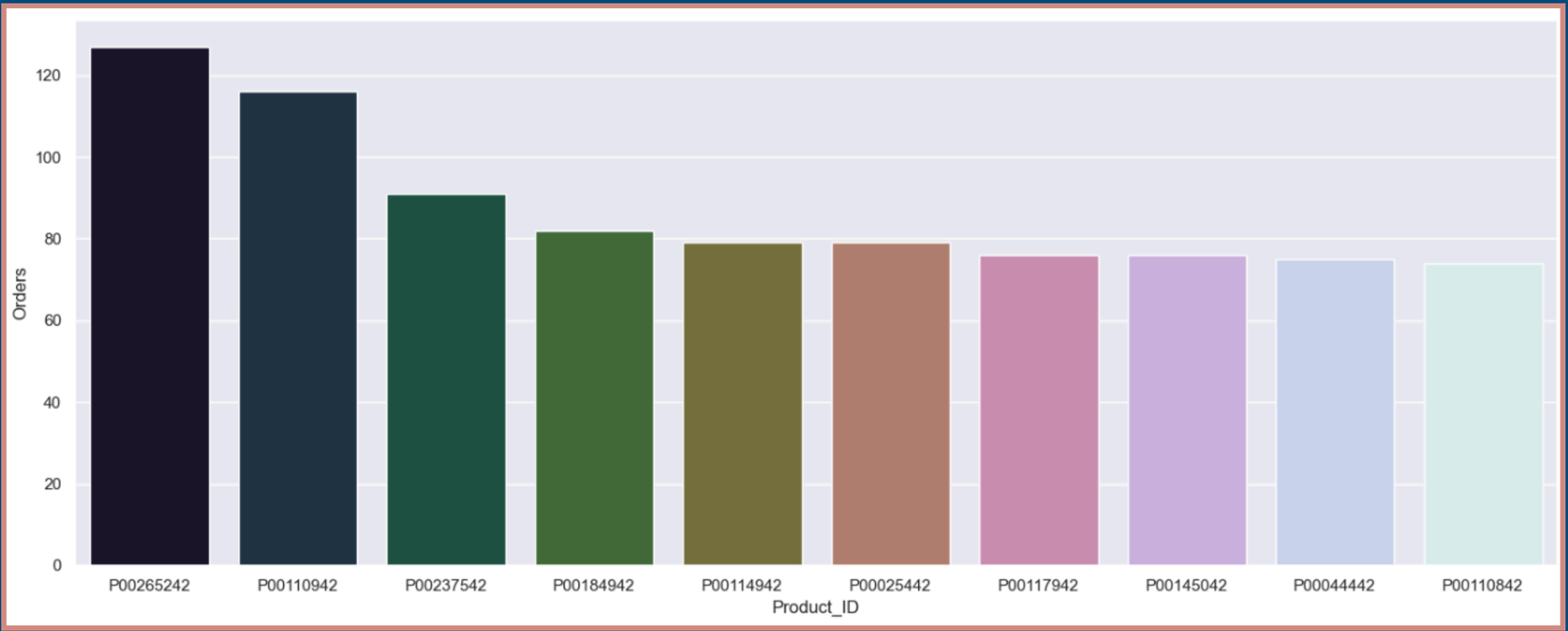
• Insights from this visualization would show the proportion of customers who make repeat purchases. A high percentage of repeat customers indicates strong customer loyalty and satisfaction.

• Conversely, a low percentage might suggest a need for improved post-purchase engagement or loyalty programs to encourage repeat business.

Exploratory Data Analysis (EDA)

14. Top 10 most sold products:

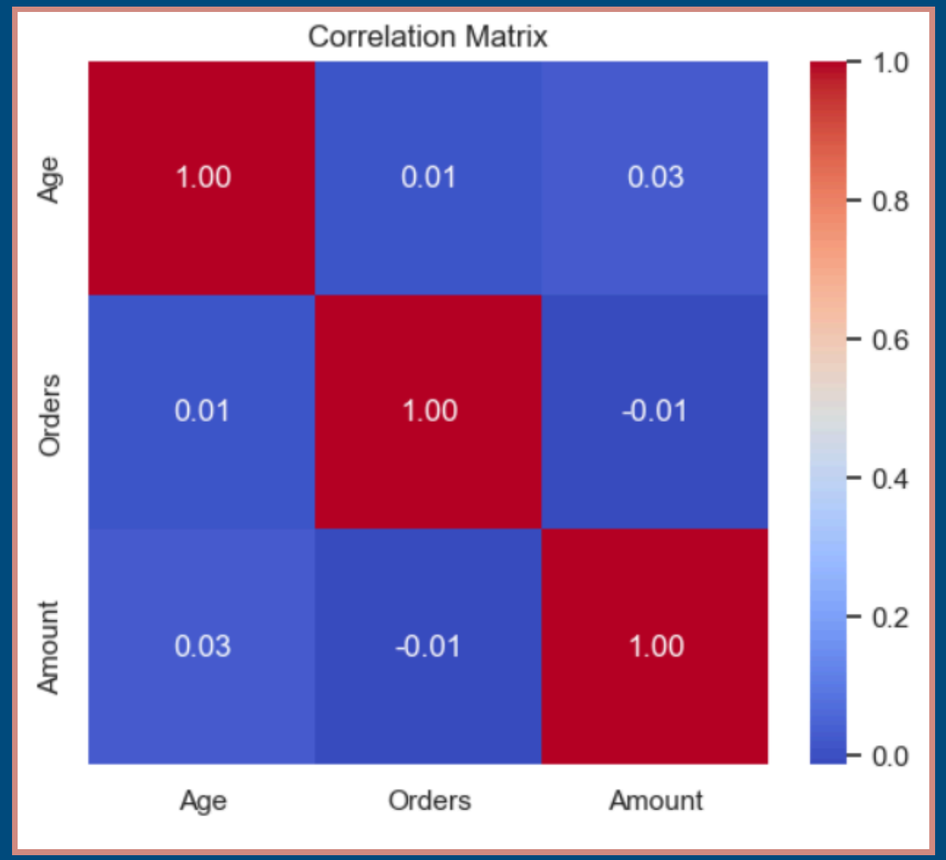
```
sales_state = df.groupby(['Product_ID'], as_index=False)['Orders'].sum().sort_values(by='Orders', ascending=False).head(10)

sns.set(rc={'figure.figsize':(15, 6)})
sns.barplot(data = sales_state, x = 'Product_ID',y= 'Orders', palette='cubehelix')
plt.tight_layout()
```



Exploratory Data Analysis (EDA)

15. Correlation Heatmap (Age, Orders, Amount):

```
plt.figure(figsize=(6,5))
corr = df_clean[['Age', 'Orders', 'Amount']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```



Insights:

• A correlation heatmap would provide a visual representation of how age, number of orders, and total amount spent are related.

• For example, a strong positive correlation between age and amount spent might indicate that older demographics have higher purchasing power, guiding specific luxury or higher-value product marketing.

• A positive correlation between orders and amount would confirm that more frequent buyers also spend more.

Key Challenges and Solutions

Key Challenges:

- Missing values in critical columns
- Unnecessary/empty columns
- Uneven sales distribution across regions

Solutions:

- Removed incomplete records
- Dropped non-contributive columns
- Focused on comparative metrics, not just raw totals

Key Insights

- Central region is a revenue leader.
- Female shoppers aged 26–35 are the most active and profitable.
- Food and fashion dominate product sales.
- Sales vary significantly by occupation and region.
- Top customers and repeat buyers drive a large share of revenue.

Key Outcomes

- Data confirms who the core customers are.
- Regional and product strategies can be refined for better ROI.
- Business can target weak areas like Eastern zone or under-30 males.
- A structured EDA process has built a foundation for predictive modeling.

Conclusion

This project successfully leveraged EDA to surface powerful insights from Diwali sales data. By aligning future actions with these insights — customer targeting, inventory allocation, regional marketing — the business can make smarter, data-backed decisions and grow sustainably.

THANK YOU

bhaskarpal.official@gmail.com