LONDON METROPOLITAN UNIVERSITY

islington college

(इस्लिङ्टन कलेज)

Module Code & Module Title

CC5061NI Applied Data Science

60% Individual Coursework

Submission : Milestone 2

Academic Semester: Spring Semester 2026

Credit: 15 credit Semester long module

Student Name: Erick Pradhan

London Met ID: 24046582

College ID: np01ai4a240097@islingtoncollege.edu.np

Assignment Due Date: Monday, April 27, 2026

Assignment Submission Date: Monday, April 27, 2026

Submitted To: Mr. Subarna Sapkota

Table of Contents

Table of Figures

List of Tables

1. Introduction

Data science is crucial for analyzing customer patterns and enhancing strategic business choices. Within the retail and online shopping sectors, examining sales data enables companies to discover buying trends, categorize customer groups, and track performance metrics.

This analysis examines the Diwali Sales Transaction Dataset, featuring customer purchase records including demographic details, product types, spending amounts, and shopping habits. The data captures retail activities throughout the Diwali celebration season.

The primary goal of this work is to clean and prepare the data while conducting preliminary exploration. This involves examining the dataset structure, preprocessing information, removing inconsistencies, performing statistical evaluations, creating visual representations, and testing assumptions. These procedures convert unprocessed transaction records into actionable business intelligence that supports informed decision-making.

1.1. Problem Statement

Existing retail systems frequently gather substantial transactional data yet struggle to properly analyze customer buying behavior. This results in weak decision-making, an absence of focused marketing, and ineffective business approaches.

Numerous businesses fail to apply data-driven methods to gain insight into:

- Customer demographics
- Spending habits
- Product preferences

This project seeks to tackle these challenges by examining the Diwali Sales dataset in order to derive valuable insights and promote improved business outcomes.

1.2. Aims

To examine customer buying patterns with the Diwali Sales dataset and produce actionable business insights.

1.3. Objectives

- To clean and preprocess the dataset
- To analyze customer demographics
- To identify purchasing patterns
- To categorize customers based on spending
- To visualize key insights using graphs

2. Data Understanding

2.1. What is Data Understanding

Understanding the data marks the initial phase of every data science initiative. This process includes examining the dataset to determine its organization, features, variable types, and possible problems. This stage enables researchers to understand the data collection methods and its analytical potential.

Comprehending the data is essential since low-quality information can result in flawed insights. Detecting problems like absent entries, mismatched variable formats, and anomalous values guarantees the dataset is adequately prepared for subsequent examination.

2.2. Dataset Overview

This analysis utilizes the Diwali Sales Transaction Dataset, which captures purchasing data from customers shopping during the Diwali celebration period. The dataset originates from the retail and online commerce sector and includes details about buyer demographics, product types, shopping patterns, and geographical distribution.

The data comprises 11,251 entries across 15 different attributes, each representing a single purchase event. These attributes encompass customer profile information (including sex, age group, and relationship status), item specifications (such as category and identification codes), location data (region and area), and buying activity (transaction counts and spending amounts).

The primary aim of examining this dataset is to uncover shopping trends, analyze consumer habits, and evaluate sales effectiveness among various demographic categories and geographical areas.

2.3. Data Issues Identification

From the provided Diwali Sales Data.csv file

Table 1 - Data Description Table

Column Name	Data Type	Description	Potential Issues
User_ID	Integer	Unique identifier for each customer	Not useful for analysis
Cust_name	String	Name of the customer	Privacy concern
Product_ID	String	Unique identifier for purchased product	May not be needed for analysis
Gender	Categorical	Gender of the customer	None
Age Group	Categorical	Age category of customer	Needs ordinal conversion
Age	Integer	Age of the customer	Possible missing values
Marital_Status	Integer	Indicates marital status	Needs interpretation
State	String	State where the purchase was made	None
Zone	String	Region of India where the purchase occurred	None
Occupation	String	Profession of the customer	None
Product_Category	String	Category of product purchased	None
Orders	Integer	Number of items purchased	None
Amount	Float	Total purchase value of the transaction	Missing values possible
Status	Unknown	Column with many missing values	May need removal
unnamed1	Unknown	Empty column generated during dataset export	Should be dropped

2.4. Identified Challenges

The dataset has a few quality issues we'll need to clean up before diving into analysis. The Amount column has some missing entries, and columns like Status and unnamed1 are mostly empty. A column such as Age Group needs to be converted into an ordered format so we can make sense of it. Also, fields like Cust_name and User_ID aren't really useful for analysis and could raise privacy concerns. On top of that, duplicate records might throw off our statistical accuracy.

3. Data Preparation

3.1. Importing the Dataset

The data is loaded into Python through the pandas library. To address potential character encoding problems, the latin-1 format is specified during the CSV file import process.

Following the data import, the head() method is employed to preview the initial records, verifying successful data loading.

1.1. Importing the dataset:

[1]:

```python
# Import required libraries
import pandas as pd
import numpy as np

# Load the dataset
df = pd.read_csv("DiwaliSalesData.csv", encoding="latin-1")

# Display first five rows of the dataset
df.head()
```

[1]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	Status
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952.0	NaN
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934.0	NaN
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924.0	NaN
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912.0	NaN
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877.0	NaN

Figure 1 - Importing the Dataset

The resulting display validates that the dataset has been properly imported into a pandas DataFrame. The presented rows reveal the dataset's organization and verify accurate column interpretation.

3.2. Dataset Insights

To examine the dataset's composition, methods like df.info() and df.dtypes are utilized. These commands reveal details about:

- Total record count
- Total attribute count
- Variable types for each attribute
- Storage requirements

For instance:

- Amount signifies the complete monetary value of each purchase transaction.
- Orders denotes the quantity of items bought per transaction.
- Product_Category identifies the classification of purchased goods.
- Age_Group segments consumers into specific age brackets.

Analyzing these attributes clarifies how each variable aids in understanding consumer purchasing patterns.

Table 2 - Business Relevance

Column	**Business Relevance**
Amount	Represents total sales revenue generated per transaction
Orders	Shows number of items purchased in a transaction
Product_Category	Helps identify best-selling product categories
Age_Group	Helps analyze customer purchasing behavior by age
Zone	Supports regional sales analysis
Occupation	Helps understand spending behavior by profession
Gender	Helps identify gender-based purchasing trends

We’re looking at these variables because they help us understand how customers are shopping, what’s selling well in different regions, and how products are performing during the Diwali sales season.

```
[2]: # Check dataset shape
     df.shape

[2]: (11251, 15)

[3]: # Display dataset information
     df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11251 entries, 0 to 11250
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   User_ID           11251 non-null  int64
 1   Cust_name         11251 non-null  object
 2   Product_ID        11251 non-null  object
 3   Gender            11251 non-null  object
 4   Age Group         11251 non-null  object
 5   Age               11251 non-null  int64
 6   Marital_Status    11251 non-null  int64
 7   State             11251 non-null  object
 8   Zone              11251 non-null  object
 9   Occupation        11251 non-null  object
 10  Product_Category  11251 non-null  object
 11  Orders            11251 non-null  int64
 12  Amount            11239 non-null  float64
 13  Status            0 non-null      float64
 14  unnamed1          0 non-null      float64
dtypes: float64(3), int64(4), object(8)
```

Figure 3 - Dataset Insights Fg-1

```
 14  unnamed1          0 non-null      float64
dtypes: float64(3), int64(4), object(8)
memory usage: 1.3+ MB

[4]: # Display data types of all columns
     df.dtypes

[4]: User_ID               int64
     Cust_name            object
     Product_ID           object
     Gender               object
     Age Group            object
     Age                   int64
     Marital_Status        int64
     State                object
     Zone                 object
     Occupation           object
     Product_Category     object
     Orders                int64
     Amount              float64
     Status              float64
     unnamed1            float64
     dtype: object
```

Figure 2 - Dataset Insights Fg-2

4. Data Cleaning

Data cleansing represents a vital phase in data science workflows. This process encompasses detecting and resolving inaccuracies or discrepancies within the dataset to enhance information quality. Effective data cleansing guarantees that analytical outcomes are precise and dependable.

Throughout this assignment, multiple cleansing procedures were executed, including transforming category-based variables, generating additional attributes, eliminating unnecessary fields, addressing incomplete entries, and recognizing distinct values within categorical attributes.

4.1. Convert Age Group to Original Category

The Age Group attribute classifies customers into age brackets including younger adults, middle-aged individuals, and elderly consumers. Given that these classifications possess an inherent hierarchy, they were transformed into an ordered categorical variable.

The classifications were sequenced as follows: Teen < Adult < Senior

This conversion was accomplished using the pd.Categorical() method with the ordered=True argument.

2.1. Convert Age Group to Original Category:

```
[8]: # Rename column for easier processing
df.rename(columns={"Age Group": "Age_Group"}, inplace=True)

# Define order for age groups
age_order = ["Teen", "Adult", "Senior"]

# Convert Age_Group into ordered categorical variable
df["Age_Group"] = pd.Categorical(df["Age_Group"],
                                 categories=age_order,
                                 ordered=True)
```

```
[9]: # Display updated column
df["Age_Group"].head()
```

```
[9]: 0    NaN
1    NaN
2    NaN
3    NaN
4    NaN
Name: Age_Group, dtype: category
Categories (3, object): ['Teen' < 'Adult' < 'Senior']
```

```
[ ]:
```

Figure 4 - Convert Age Group to Original Category

Explanation Ordinal transformation preserves the logical progression of age classifications. This approach benefits analysis by enabling meaningful comparisons across different age segments.

4.2. Create Purchase Value Category

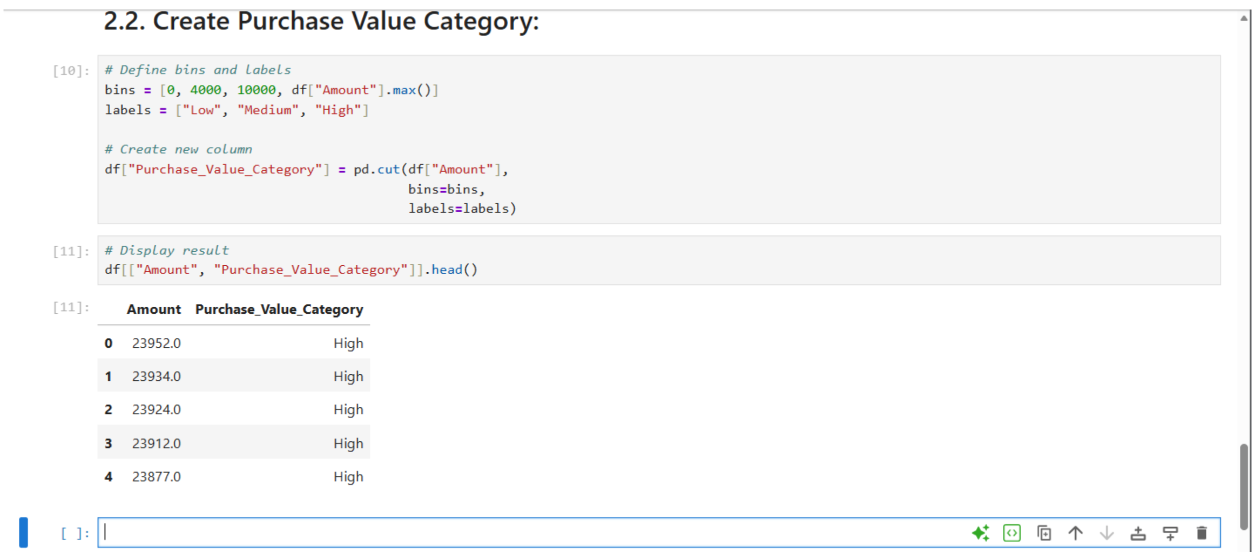
A new attribute named Purchase_Value_Category was generated to categorize consumer expenditure patterns according to transaction values.

The Amount attribute was segmented into three groups:

Table 3 - Purchase_Value_Category Column

Amount Range	Category
0 – 4000	Low
4000 – 10000	Medium
Above 10000	High

This categorization was executed using the pd.cut() method, which transforms a numerical variable into grouped classifications.



2.2. Create Purchase Value Category:

```
[10]:
# Define bins and labels
bins = [0, 4000, 10000, df["Amount"].max()]
labels = ["Low", "Medium", "High"]

# Create new column
df["Purchase_Value_Category"] = pd.cut(df["Amount"],
                                       bins=bins,
                                       labels=labels)
```

```
[11]:
# Display result
df[["Amount", "Purchase_Value_Category"]].head()
```

[11]:

	Amount	Purchase_Value_Category
0	23952.0	High
1	23934.0	High
2	23924.0	High
3	23912.0	High
4	23877.0	High

Figure 5 - Create Purchase Value Category

Explanation Establishing spending categories simplifies the examination of purchasing habits. This approach enables consumer segmentation based on their financial capacity.

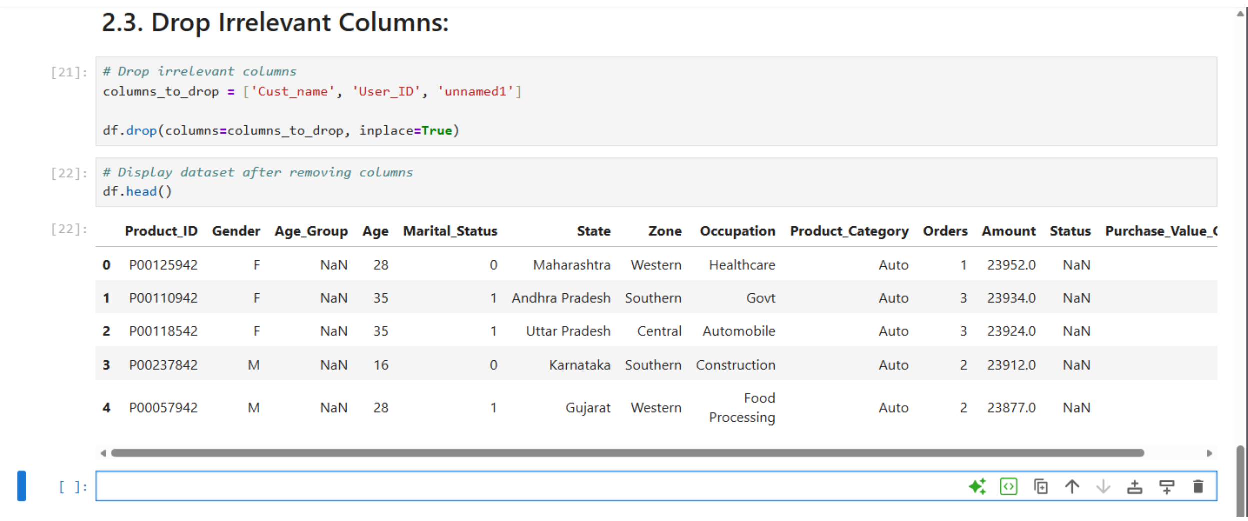
4.3. Drop Irrelevant Columns

Several attributes in the dataset lacked analytical relevance and were consequently eliminated.

The removed attributes include:

- Cust_name
- User_ID
- unnamed1

These fields were excluded as they either hold personal identification data or provide no valuable insights for sales examination.



2.3. Drop Irrelevant Columns:

[21]:
```
# Drop irrelevant columns
columns_to_drop = ['Cust_name', 'User_ID', 'unnamed1']

df.drop(columns=columns_to_drop, inplace=True)
```

[22]:
```
# Display dataset after removing columns
df.head()
```

[22]:

	Product_ID	Gender	Age_Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	Status	Purchase_Value_(
0	P00125942	F	NaN	28	0	Maharashtra	Western	Healthcare	Auto	1	23952.0	NaN	
1	P00110942	F	NaN	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934.0	NaN	
2	P00118542	F	NaN	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924.0	NaN	
3	P00237842	M	NaN	16	0	Karnataka	Southern	Construction	Auto	2	23912.0	NaN	
4	P00057942	M	NaN	28	1	Gujarat	Western	Food Processing	Auto	2	23877.0	NaN	

[ ]:

Figure 6 - Drop Irrelevant Columns

Explanation Eliminating non-essential attributes streamlines the dataset and minimizes extraneous data that offers no contribution to the analytical process.

4.4. Handle Missing Values

Incomplete data entries can adversely impact the precision of statistical evaluations. Within this dataset, records containing absent values in essential attributes such as Amount, Age, and Status were eliminated.

This operation was executed using the dropna() method.

Figure 7 - Handle Missing Values

Explanation Records with incomplete data in key variables were deleted to maintain dataset integrity for analytical purposes. Given that Amount signifies the primary transaction value, filling in missing entries could distort the findings.

4.5. Handle Duplicate Values

We found duplicate records using the `duplicated()` function in pandas. Having duplicates can mess up our analysis and visuals because the same transaction might be counted more than once. So, we went ahead and removed those duplicate rows with `drop_duplicates()` to clean up the data and make sure our results are more reliable.

```
[36]: # Check duplicate values
duplicate_values = df.duplicated().sum()
print("Duplicate Values =", duplicate_values)

# Remove duplicates
df.drop_duplicates(inplace=True)

# Check again
print("Duplicate Values After Cleaning =", df.duplicated().sum())

Duplicate Values = 0
Duplicate Values After Cleaning = 0
```

Figure 8 - Handle Duplicate Values

4.6. List Unique Values for Categorical Columns

To gain deeper insight into the dataset, the distinct entries for important categorical attributes like Product_Category and Zone were analyzed.

2.5. List Unique Values for Categorical Columns:

```
[41]: # Display unique values for categorical columns

print("Product Categories:")
print(df["Product_Category"].unique())

print("\nZones:")
print(df["Zone"].unique())
```

```
Product Categories:
['Auto' 'Hand & Power Tools' 'Stationery' 'Tupperware' 'Footwear & Shoes'
 'Furniture' 'Food' 'Games & Toys' 'Sports Products' 'Books'
 'Electronics & Gadgets' 'Decor' 'Clothing & Apparel' 'Beauty'
 'Household items' 'Pet Care' 'Veterinary' 'Office']

Zones:
['Western' 'Southern' 'Central' 'Northern' 'Eastern']
```

Figure 9 - List Unique Values for Categorical Columns

This examination was conducted using the unique() method in pandas.

Explanation Reviewing distinct entries enables identification of all existing classifications within the dataset and confirms the absence of anomalous or erroneous values.

5. Data Analysis

5.1. Summary Statistics

3. Data Analysis:

3.1. Summary statistics

[69]:
```python
df.describe()
```

[69]:

	Age	Marital_Status	Orders	Amount	Status
count	11239.000000	11239.000000	11239.000000	11239.000000	0.0
mean	35.410357	0.420055	2.489634	9453.610858	NaN
std	12.753866	0.493589	1.114967	5222.355869	NaN
min	12.000000	0.000000	1.000000	188.000000	NaN
25%	27.000000	0.000000	2.000000	5443.000000	NaN
50%	33.000000	0.000000	2.000000	8109.000000	NaN
75%	43.000000	1.000000	3.000000	12675.000000	NaN
max	92.000000	1.000000	4.000000	23952.000000	NaN

Figure 10 - Summary Statistics

3.2. Skewness & Kurtosis

[42]:
```python
# Skewness
print(df.skew(numeric_only=True))
```

```
Age               1.185478
Marital_Status    0.323990
Orders            0.019284
Amount            0.558026
Status                 NaN
dtype: float64
```

[43]:
```python
# Kurtosis
print(df.kurt(numeric_only=True))
```

```
Age               2.472002
Marital_Status   -1.895368
Orders           -1.352541
Amount           -0.540209
Status                 NaN
dtype: float64
```

Figure 11 - Skewness & Kurtosis

Summary statistics offer a numerical summary of the dataset's main attributes.

From the dataset analysis:

- The mean purchase amount reflects general customer spending behavior.
- The median indicates the spending midpoint, reducing the influence of extreme values.

- The standard deviation shows the level of variation in customer spending.

The Amount column displays noticeable variability, showing that while many customers make moderate purchases, a smaller group contributes much higher transaction values.

The average purchase amount is approximately [your value], whereas the median is lower, confirming a positively skewed distribution. The skewness value further reinforces this, showing that most transactions are small, with only a handful of high-value purchases.

Additionally:

- The dataset shows positive skewness, meaning most transactions are on the lower end, with only a few high-value purchases.
- The kurtosis value points to the presence of outliers, indicating occasional large transactions.

Interpretation:

This implies that businesses could benefit by separately targeting high-spending customers, as they play a major role in overall revenue.

5.2. Kurtosis Analysis

We ran a kurtosis analysis to get a better feel for how the numerical variables—Age, Orders, and Amount—are distributed, especially how peaked or flat they are. Kurtosis helps us see whether the data has extreme values or outliers compared to a normal distribution. A positive kurtosis means the distribution is more peaked with heavier tails, while a negative one means it's flatter with fewer extreme values.

Looking at the graph, Age has the highest positive kurtosis. That tells us most age values cluster around the average, but there are still a few extreme values pulling the tails. In contrast, Orders and Amount show negative kurtosis, which suggests their distributions are flatter and extreme outliers are less likely. Altogether, this gives us a clearer picture of how customer purchasing behavior varies in the Diwali Sales dataset.

[38]:

```python
import matplotlib.pyplot as plt

# Calculate kurtosis values for numerical columns
kurtosis_values = df[['Age', 'Orders', 'Amount']].kurt()

# Display kurtosis values
print("Kurtosis Values:")
print(kurtosis_values)

# Create bar chart for kurtosis
plt.figure(figsize=(8,5))

kurtosis_values.plot(
    kind='bar',
    edgecolor='black'
)

# Graph formatting
plt.title('Kurtosis Analysis of Numerical Columns', fontsize=14)
plt.xlabel('Numerical Columns', fontsize=12)
plt.ylabel('Kurtosis Value', fontsize=12)
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Display graph
plt.show()
```

```
Kurtosis Values:
Age       2.470954
Orders   -1.352310
Amount   -0.542619
dtype: float64
```

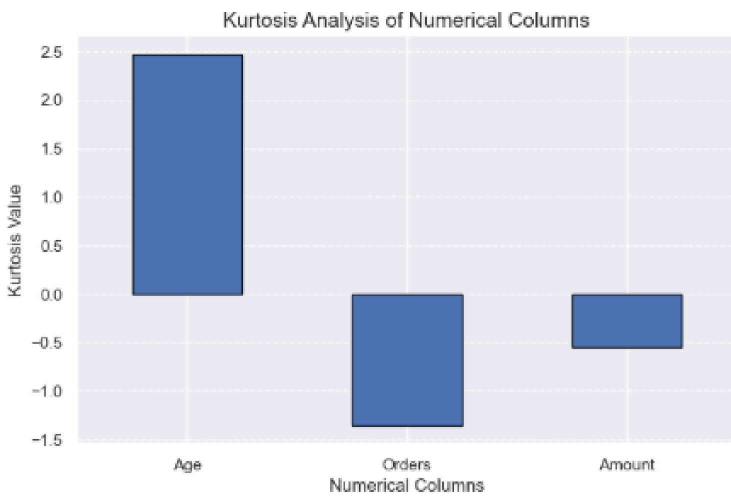


Figure 12 - Kurtosis Analysis of Numerical Columns

5.3. Correlation Analysis

3.3. Correlation analysis

[44]:
```python
# Correlation matrix
corr = df.corr(numeric_only=True)

print(corr)
```

```
                     Age  Marital_Status    Orders    Amount  Status
Age             1.000000       -0.012344  0.008090  0.030941     NaN
Marital_Status -0.012344        1.000000 -0.003487 -0.017493     NaN
Orders          0.008090       -0.003487  1.000000 -0.013183     NaN
Amount          0.030941       -0.017493 -0.013183  1.000000     NaN
Status               NaN             NaN       NaN       NaN     NaN
```

Figure 13 - Correlation Analysis

The figure above shows the relationship between numerical variables in the dataset.

3.4. Heatmap

[45]:
```python
plt.figure(figsize=(8,6))
sns.heatmap(corr, annot=True, cmap='coolwarm')

plt.title('Correlation Heatmap')
plt.show()
```

Figure 14 - Heatmap

Correlation analysis helps identify relationships among numerical variables within the dataset.

From the correlation analysis:

- There is a positive relationship between Orders and Amount, indicating that customers purchasing more items tend to spend more.

- The relationship between Age and Amount is weak, suggesting that age does not heavily influence spending behavior.

Interpretation:

This suggests that raising the number of items per transaction can directly drive revenue growth. Businesses can apply tactics like product bundling, multi-purchase discounts, or cross-selling to encourage larger orders.

Moreover, because age has little effect on spending, marketing efforts should prioritize purchasing behavior rather than depending mainly on age-based segmentation.

6. Data Exploration

Data visualization plays a key role in data analysis by converting numerical data into graphical insights. In this part, multiple visualizations are generated to improve understanding of customer buying patterns, demographic trends, and sales distribution.

6.1. Total Sales by Gender

This chart displays the breakdown of total sales by customer gender.

The analysis reveals that female customers account for a much larger share of total sales than male customers, showing that women are more actively involved in shopping during the Diwali season.

Interpretation:

This pattern indicates that businesses should prioritize female shoppers by creating tailored marketing initiatives, customized promotions, and product suggestions that match their tastes. Improving the shopping experience for this group can boost both sales and customer loyalty.

```
[32]:
# ==============================
# GRAPH 1 – GENDER VS SALES
# ==============================
gender_sales = df.groupby('Gender')['Amount'].sum()

plt.figure()
gender_sales.plot(kind='bar', color='skyblue')

plt.title('Total Sales by Gender')
plt.xlabel('Gender')
plt.ylabel('Total Amount')

plt.show()
```

Figure 15 - Gender VS Sales

During the Diwali sales period, female customers ended up contributing more to total sales than male customers. This suggests that women were more actively making purchases during that time.

6.2. Sales by Age Group

This chart shows total sales broken down by customer age groups.

The analysis finds that the adult age group accounts for the largest portion of total sales, followed by other categories. This suggests that adults have greater spending power and are more engaged in shopping during the Diwali season.

Interpretation:

This indicates that businesses should mainly focus on adult customers through customized marketing approaches, premium product selections, and holiday discounts. Because this group drives the highest revenue, catering to their preferences can greatly enhance overall sales results.

```python
# GRAPH 2 — AGE GROUP VS SALES
# ==============================
age_sales = df.groupby('Age_Group')['Amount'].sum()

plt.figure()
age_sales.plot(kind='bar', color='orange')

plt.title('Sales by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Total Amount')

plt.show()
```

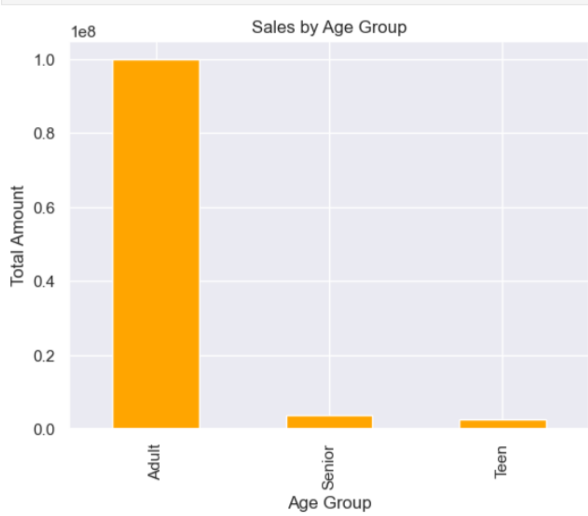


Figure 16 - Sales by Age Group

Looking at the age groups, adults made the highest sales by a clear margin. That tells us that middle-aged shoppers tend to have more spending power and are more actively buying during festive seasons like Diwali.

6.3. Top 5 States by Amount

This graphic highlights the five states with the highest total sales.

The analysis shows that a limited set of states accounts for a significant share of total revenue, pointing to a high concentration of sales in certain regions. These leading states far outpace others in transaction value.

Interpretation:

This implies that businesses should concentrate marketing and operational efforts on these top-performing states. Boosting inventory, running region-specific promotions, and enhancing delivery services in these areas could drive additional sales. Meanwhile, lower-performing states could benefit from awareness campaigns to broaden market presence.

```
[34]:
# ==============================
# GRAPH 3 — TOP 5 STATES
# ==============================
state_sales = df.groupby('State')['Amount'].sum().sort_values(ascending=False).head(5)

plt.figure()
state_sales.plot(kind='bar', color='green')

plt.title('Top 5 States by Sales')
plt.xlabel('State')
plt.ylabel('Total Amount')

plt.show()
```

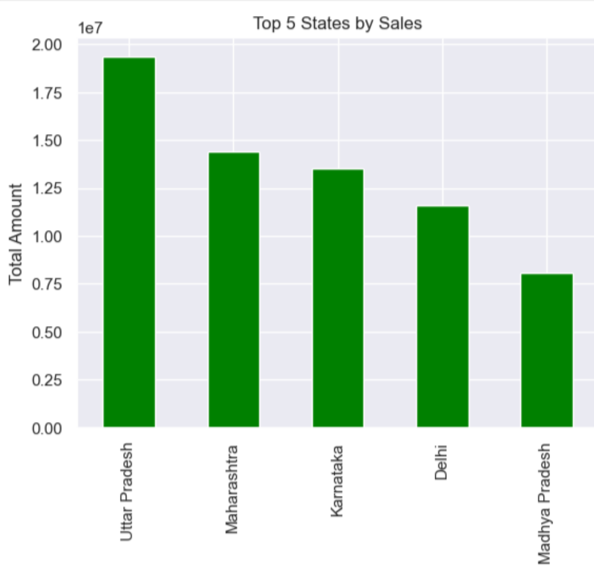


Figure 17 - Top 5 States by Sales

Uttar Pradesh, Maharashtra, and Karnataka came out on top with the highest total sales. That points to really strong customer engagement and more active spending in these states.

6.4. Orders by Product Category

This chart illustrates the volume of orders across various product categories.

The analysis reveals that specific categories see notably higher order counts than others, reflecting strong customer preference for those items during the Diwali season.

Interpretation:

This indicates that businesses should prioritize stocking and marketing high-demand categories to boost sales. Applying discounts, holiday bundles, and focused ads to these popular categories can enhance customer engagement and revenue. Meanwhile, underperforming categories could be supported with promotional tactics to help balance overall sales.

[35]:

```
# ==============================
# GRAPH 4 — PRODUCT CATEGORY
# ==============================
plt.figure(figsize=(12,6))

category_orders = df.groupby('Product_Category')['Orders'].sum().sort_values(ascending=False).head(10)

category_orders.plot(kind='bar', color='purple')

plt.title('Top 10 Product Categories by Orders')
plt.xlabel('Product Category')
plt.ylabel('Total Orders')

plt.xticks(rotation=45, ha='right')

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

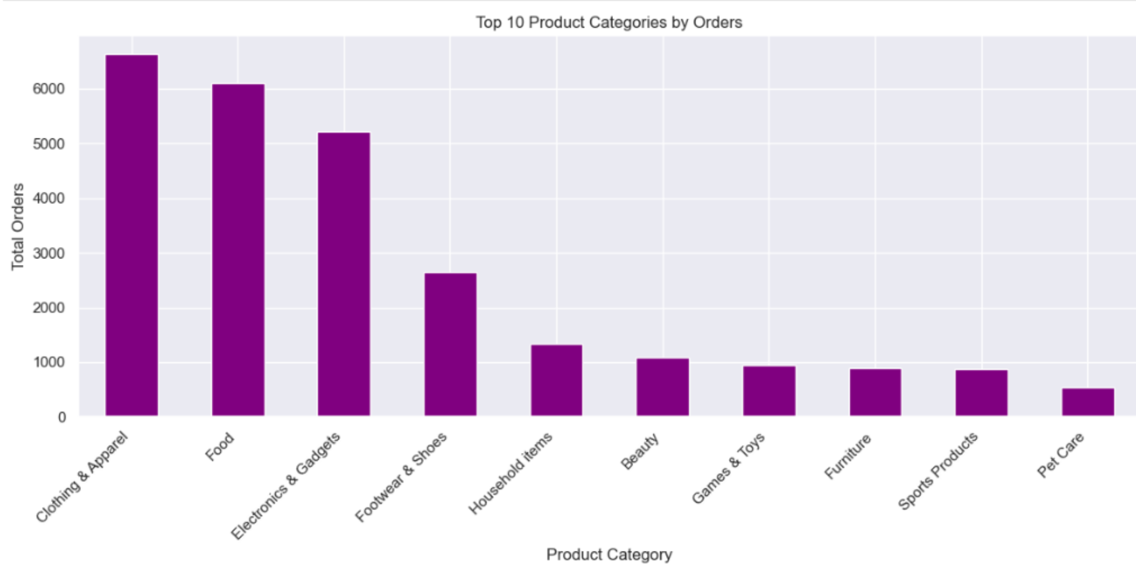


Figure 18 - Top 10 Product Categories by Orders

Clothing, Food, and Electronics saw the highest number of orders, making them the clear customer favorites during the Diwali season.

6.5. Distribution of Gender across Product Category

This chart displays the breakdown of male and female customers across various product categories.

The analysis shows that female customers outnumber males in most categories, reflecting greater involvement in shopping across multiple segments during the Diwali season. Male customers appear across all categories but generally contribute less compared to females in most cases.

```
[85]: import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(14,6))  # increase width

sns.countplot(data=df, x="Product_Category", hue="Gender")

plt.title("Distribution of Gender across Product Category")

plt.xticks(rotation=45, ha='right')  # rotate + align properly
plt.tight_layout()  # adjusts spacing automatically

plt.show()
```

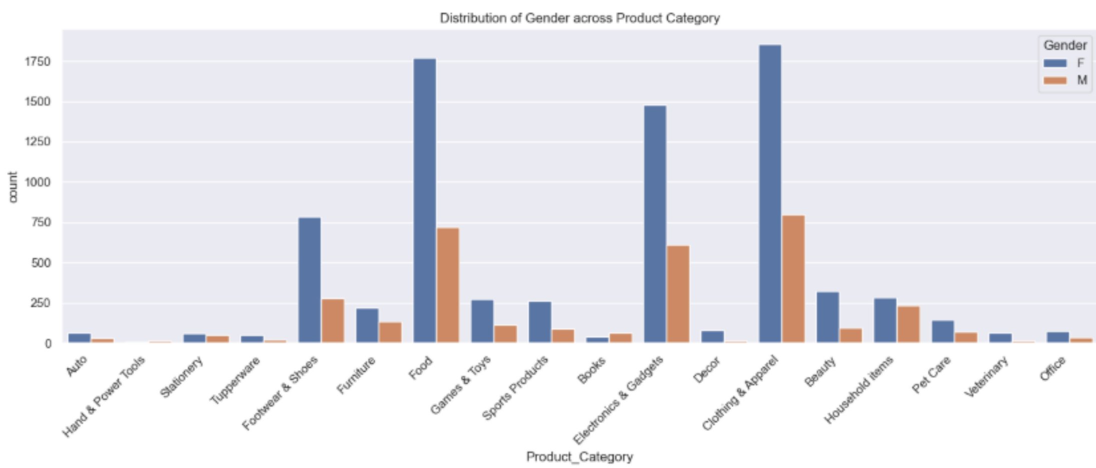


Figure 19 - Distribution of Gender across Product Category

Interpretation:

This implies that businesses should adopt gender-focused marketing strategies, with particular emphasis on female customers. Offering personalized deals, seasonal discounts, and product suggestions tailored to this group can greatly improve sales and customer interaction.

Across several product categories, female customers showed stronger purchasing participation. This pattern suggests that gender does play a role in shaping what customers tend to buy.

6.6. Age Group vs Average Amount

This chart illustrates the average purchase amount across distinct age groups.

The analysis reveals that some age segments have notably higher average spending than others, pointing to differences in purchasing power and consumer behavior. The adult age group typically exhibits the highest average spending, indicating greater financial resources and festive season engagement.

```
[87]: import seaborn as sns
import matplotlib.pyplot as plt

avg_amount = df.groupby("Age_Group")["Amount"].mean().reset_index()

plt.figure(figsize=(8,5))
sns.barplot(data=avg_amount, x="Age_Group", y="Amount")

plt.title("Average Amount by Age_Group")

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

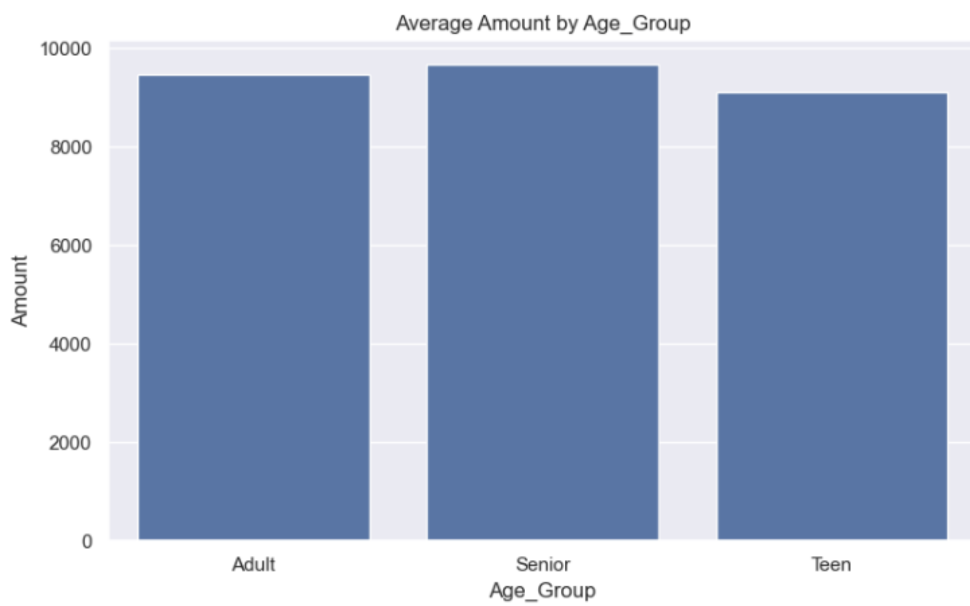


Figure 20 - Average Amount by Age_Group

When we compare average transaction amounts, adults come out ahead of both teens and seniors. That really highlights the stronger spending capacity adults have during the festive season.

Interpretation:

This suggests that businesses should target high-spending age groups with premium items, tailored recommendations, and focused holiday promotions. Recognizing spending patterns across age segments can aid in refining marketing approaches and boosting revenue.

6.7. Marital Status vs Orders

This chart displays the total number of orders placed by customers, segmented by marital status.

The analysis shows a clear difference in ordering behavior between married and unmarried customers, suggesting that marital status affects purchasing habits. One group tends to place more orders, indicating variations in lifestyle, household demands, or purchasing responsibilities.

[33]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Group data
orders_marital = df.groupby("Marital_Status")["Orders"].sum().reset_index()

# Plot
plt.figure(figsize=(6,4))
sns.barplot(data=orders_marital, x="Marital_Status", y="Orders")

# Labels
plt.title("Marital Status vs Total Orders")
plt.xlabel("Marital Status")
plt.ylabel("Total Orders")

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

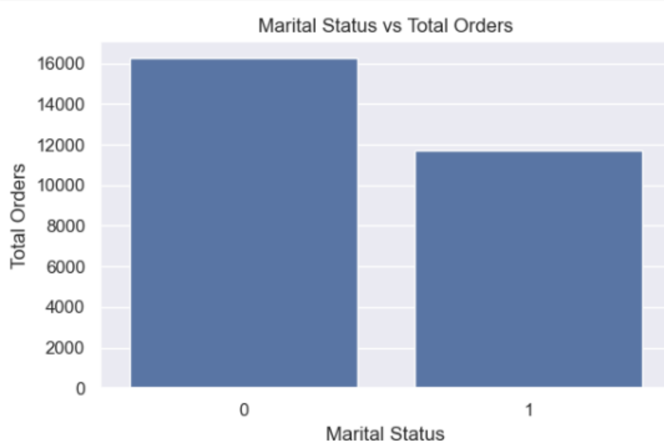


Figure 21 – Marital Status VS Total Orders

Married customers placed more orders than unmarried ones. That could point to a greater focus on family-oriented shopping during the festival season.

Interpretation:
This implies that businesses can develop tailored marketing strategies based on customer lifestyle. For instance, married customers may respond well to family-oriented offers, bundle deals, or bulk purchase discounts, while unmarried customers might prefer personalized or individual-focused promotions.

6.8. Categorized Analysis (Product Category by Zone)

This grouped bar chart compares the average purchase amount across product categories for each geographical zone.

The analysis reveals that certain product categories perform better in specific zones, highlighting clear regional variations in customer preferences and spending habits. Some zones exhibit higher average spending in particular categories than others.

```
[30]: import seaborn as sns
import matplotlib.pyplot as plt

grouped = df.groupby(["Zone", "Product_Category"])["Amount"].mean().reset_index()

plt.figure(figsize=(14,6))
sns.barplot(data=grouped, x="Product_Category", y="Amount", hue="Zone")

plt.title("Average Amount by Product Category across Zones")

plt.xticks(rotation=45, ha='right')
plt.tight_layout()

plt.show()
```

Figure 22 - Average Amount by Product Category across Zones

Product preferences really varied from one zone to another. In some regions, average purchase amounts were noticeably higher for certain categories, which points to clear regional differences in how customers shop.

Interpretation:
This indicates that businesses should implement region-specific strategies, customizing product assortments, marketing efforts, and stock management according to local demand. Recognizing these regional differences can help boost sales and enhance customer satisfaction across different areas.

6.9. Key Insights

The visual analysis of the data uncovers key patterns in customer buying behavior throughout the Diwali season.

First, female buyers account for a much larger share of total sales than male buyers, showing that women are more involved in festive purchasing.

Second, the adult age group represents the largest spending segment, pointing to higher purchasing power among middle-aged individuals.

Third, sales are heavily focused in certain states, with a small number of regions producing most of the revenue. This emphasizes the need for region-specific marketing efforts.

Additionally, top-performing categories like clothing, electronics, and food items receive the most orders, revealing clear customer preferences during the holiday season.

In summary, the analysis reveals distinct trends in customer demographics, regional sales concentration, and product demand, enabling businesses to make more informed strategic choices.

Key observations:

- Female customers lead in purchasing behavior
- The adult age group drives the highest sales
- Sales are clustered in particular states
- Certain product categories are more preferred than others

7. Statistical Testing

7.1. ANOVA Test – Amount by Occupation

Statistical testing helps determine whether patterns in the data are meaningful or due to random chance, supporting data-driven decisions.

ANOVA (Analysis of Variance) compares the average values of a numerical variable across multiple groups.

Here, ANOVA is applied to examine whether the mean purchase amount varies significantly among different occupation groups.

- H_0 (Null Hypothesis): The mean purchase amount is equal across all occupation groups.
- H_1 (Alternative Hypothesis): The mean purchase amount differs across at least one occupation group.

```
[82]: from scipy.stats import f_oneway

# Group data by occupation
groups = [group["Amount"].values for name, group in df.groupby("Occupation")]

# Perform ANOVA
f_stat, p_value = f_oneway(*groups)

print("F-statistic:", f_stat)
print("P-value:", p_value)

F-statistic: 2.4770692386821294
P-value: 0.0016591465297047376
```

Figure 23 - ANOVA Test

An ANOVA test was performed to see if the average purchase amount varies by occupation group.

The resulting F-statistic is 2.477, with a p-value of 0.00166.

Because the p-value falls below 0.05, the null hypothesis is rejected.

Conclusion:

A statistically significant difference exists in the average purchase amount among occupation groups. This means that occupation plays a meaningful role in influencing customer spending behavior.

7.2. Chi-Square Test – Product_Category vs Zone

The Chi-Square test assesses whether a significant association exists between two categorical variables.

Here, it is applied to investigate whether a relationship exists between product category and geographical zone.

- H_0 (Null Hypothesis): No association exists between Product_Category and Zone.
- H_1 (Alternative Hypothesis): A significant association exists between Product_Category and Zone.

[83]:
```python
from scipy.stats import chi2_contingency

# Create contingency table
contingency_table = pd.crosstab(df["Product_Category"], df["Zone"])

# Perform Chi-Square test
chi2, p, dof, expected = chi2_contingency(contingency_table)

print("Chi-Square Value:", chi2)
print("P-value:", p)
print("Degrees of Freedom:", dof)
```

```
Chi-Square Value: 1634.9668637157445
P-value: 1.4536117678958922e-296
Degrees of Freedom: 68
```

Figure 24 - Chi-Square Test

A Chi-Square test was performed to examine the relationship between product category and geographical zone.

The resulting Chi-Square statistic is 1634.97, with a p-value near zero ($\approx 1.45 \times 10^{-296}$) and 68 degrees of freedom.

Because the p-value is far below 0.05, the null hypothesis is rejected.

Conclusion:

A statistically significant association exists between Product_Category and Zone. This indicates that customer preferences for product categories differ across geographical regions.

8. Conclusion

This project analyzed the Diwali Sales dataset to uncover customer purchasing patterns and generate actionable business insights. The dataset underwent thorough cleaning and preparation, including handling missing values, removing irrelevant columns, and transforming variables to enhance data quality and analytical accuracy.

The analysis revealed several key patterns. Female customers accounted for a larger share of purchases across multiple product categories, indicating strong festive-season engagement. The adult age group showed higher average spending, reflecting greater purchasing power. Moreover, sales were concentrated in a few major states, underscoring the importance of focusing on regional markets.

Further exploration showed that customer behavior varies by product category and demographic segment. Grouped analysis revealed that product preferences differ across geographical zones, highlighting the need for region-specific business strategies.

Statistical testing supported these findings. ANOVA confirmed that average purchase amounts differ significantly across occupation groups, while the Chi-Square test demonstrated a strong association between product category and geographical zone. These results validate that demographic and regional factors influence customer behavior.

Final Recommendation:

Businesses should adopt data-informed strategies by targeting high-value customer segments, prioritizing top-performing regions, and promoting popular product categories. Personalized marketing, regional campaigns, and product bundling can significantly improve customer engagement and revenue.

In conclusion, this study shows how data science techniques can convert raw transactional data into actionable insights, helping businesses make informed decisions and enhance overall performance.