**Python Mini-Project Report**

Academic Year 2024-2025

Course: Python for Data Science Laboratory Course Code: DJS23SLPC303

Class: S.Y.B.Tech. Semester: III Division: S

Department: Artificial Intelligence (AI) and Data Science Batch: A1

**E-Commerce Sales Analysis**

*This Python Mini-Project was prepared by :*

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Roll No.** | **Name** | **SAP ID** |
| 1 | S012 | Dhwaj Jain | 60018230070 |
| 2 | S014 | Eric Kurissery | 60018230071 |
| 3 | S015 | Hadi Gala | 60018230085 |
| 4 | S026 | Kaustav Dedhia | 60018230009 |

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Topic** | **Page No.** |
| 1 | Project Overview | 2 |
| 2 | Tech Stack | 3 |
| 3 | Code & Output | 3 |
| 4 | Conclusion | 19 |
| 5 | References | 20 |

**Project Overview**

The E-commerce Sales Analysis project explores the Superstore USA dataset to uncover sales trends, customer behavior, and profitability drivers. By analyzing data on products, categories, sales, discounts, profits, and regions, the project aims to identify actionable insights to optimize operations and decision-making in e-commerce.

The analysis begins with cleaning and preprocessing the dataset, including handling missing values, standardizing formats, and converting date fields. Exploratory Data Analysis (EDA) and visualization reveal trends, such as seasonal sales patterns, region-wise performance, and the impact of discounts on profitability.

Approach:

Data Cleaning: Address missing values, inconsistent formats, and prepare data for analysis.

EDA: Analyze relationships among sales, discounts, profits, and regions using descriptive statistics and visualizations.

Visualization: Use charts like line graphs for sales trends, heatmaps for regional performance, and bar plots for top categories.

Key Insights:

Sales Trends: Identify peak seasons and yearly growth.

Profitability: Highlight categories impacted by discounts.

Regional Analysis: Discover high and low-performing regions.

Customer Behavior: Understand preferences across segments.

This project delivers insights to enhance e-commerce strategies and boost profitability.

**Tech Stack**

Python: Programming language used for data analysis and data visualization

libraries of Bengaluru Housing dataset.

Python Libraries:

o Pandas: To preprocess and analyze the dataset, including handling missing

values and filtering data.

o NumPy: For efficient numerical operations, such as data type conversions

and mathematical computations.

o Matplotlib & Seaborn: For creating detailed visualizations like histograms,

scatter plots, and bar plots.

Dataset used: Superstore USA from Kaggle

<https://www.kaggle.com/datasets/vivek468/superstore-dataset-final>

**Code & Output**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import io

from google.colab import files

uploaded = files.upload()

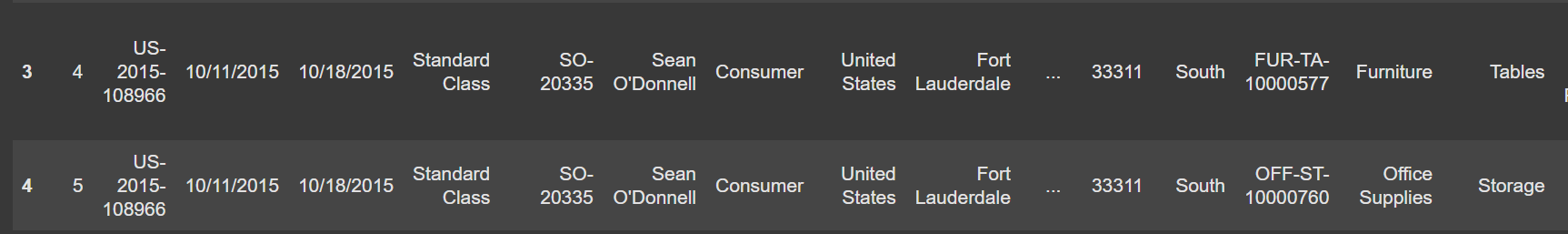
to load the csv dataset file

head used to display first 5 rows of dataset for verification

df = pd.read\_csv(io.BytesIO(uploaded['Superstore\_dataset (1).csv']), encoding='latin-1')

df.head()





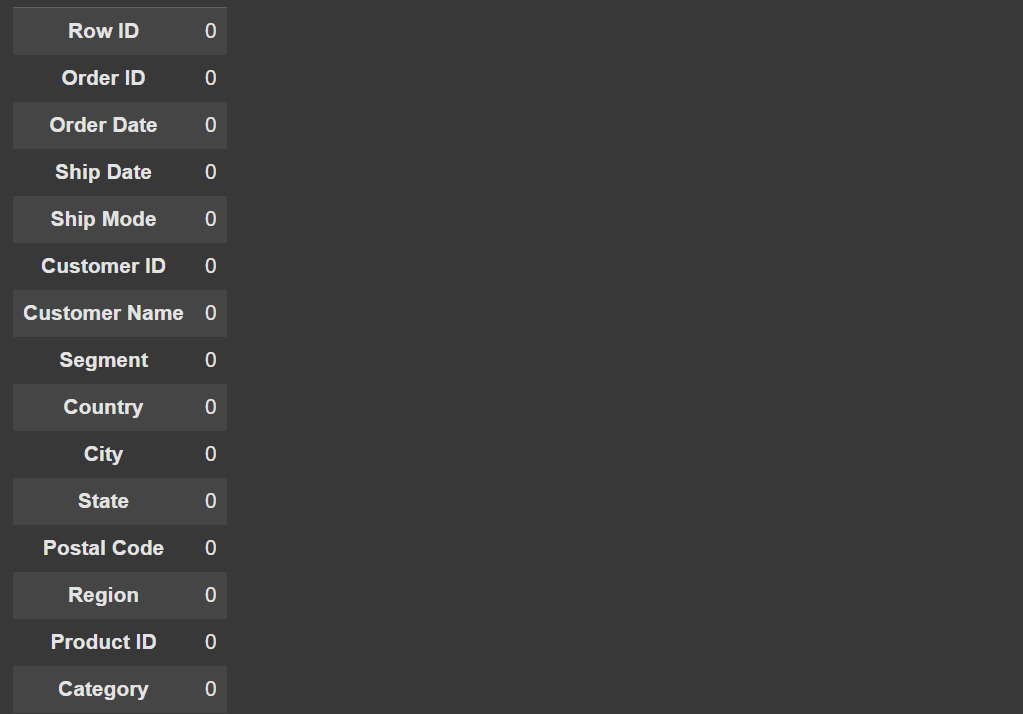
to find the size of dataset rows\*columns

df.shape

(9994, 21)

to find null values if any and in which column

df.isnull().sum()

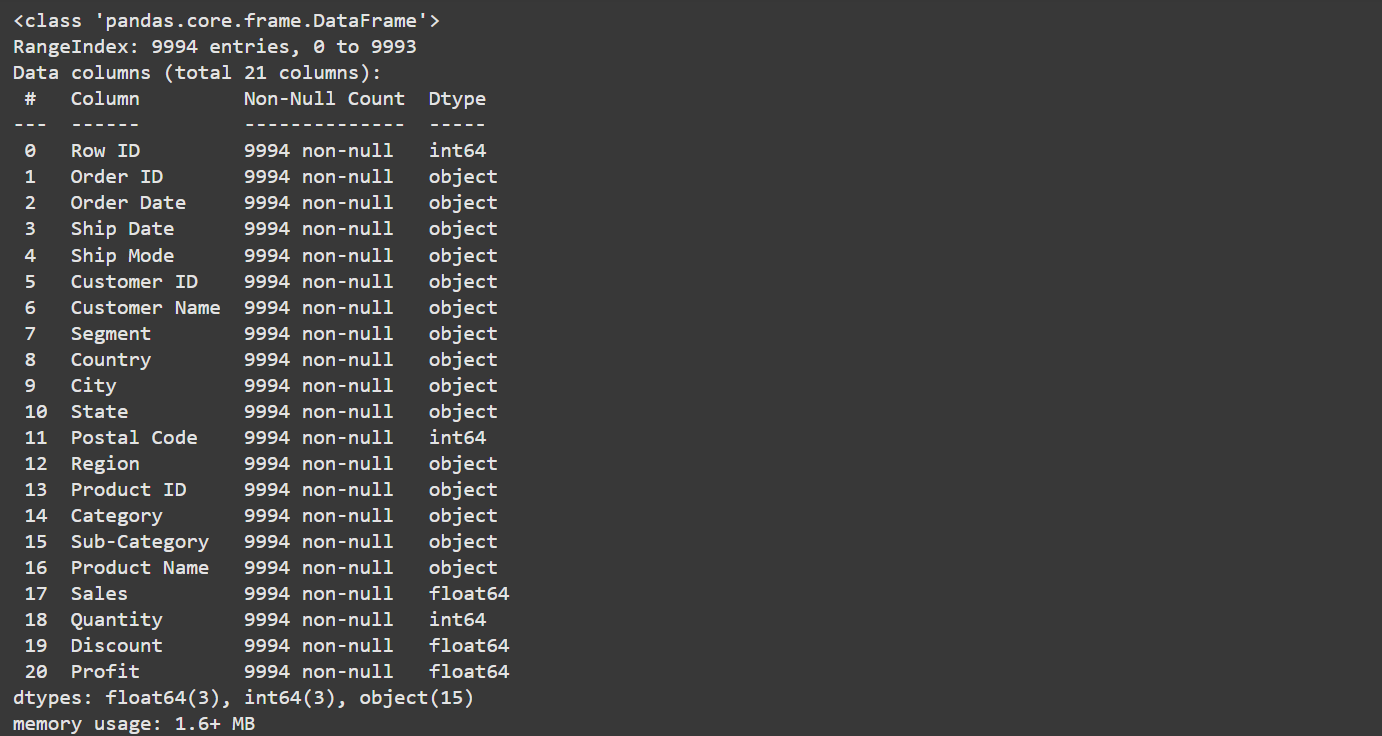




to remove null values if any use following code

df['column name'].fillna(df['column name'].mean(),inplace=True)

df.info()



df['Category'].unique()

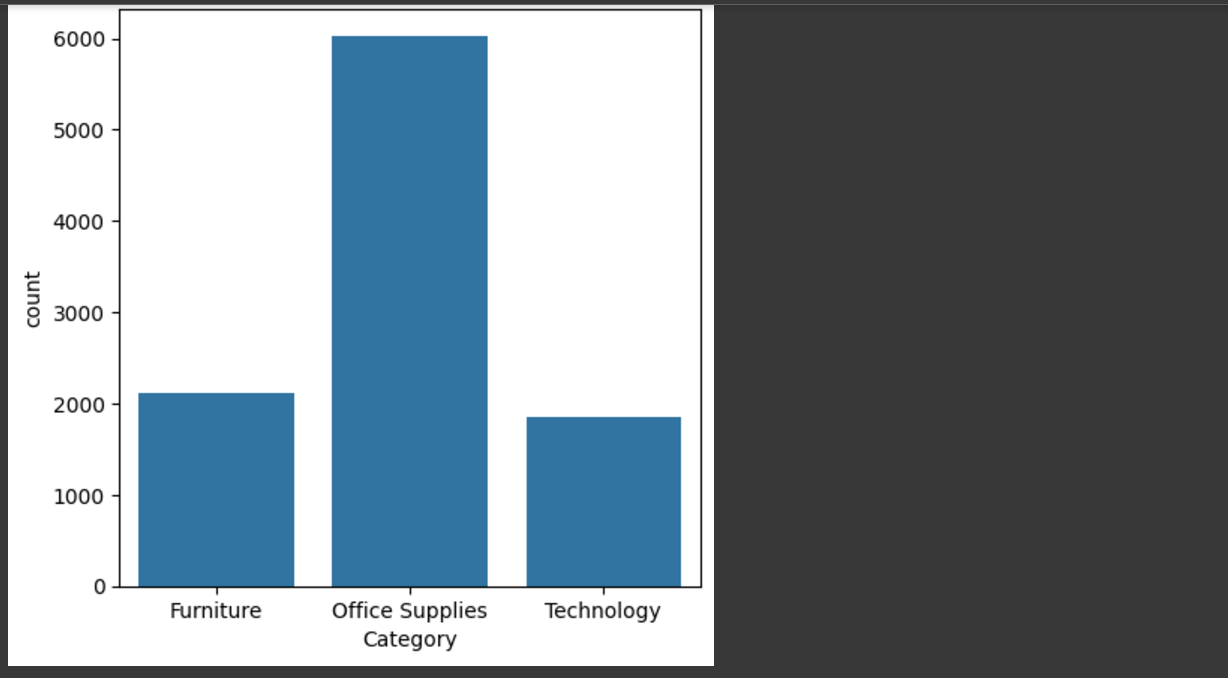
df['Category'].value\_counts()



plt.figure(figsize=(5,5))

sns.countplot(x='Category',data=df)

plt.show()

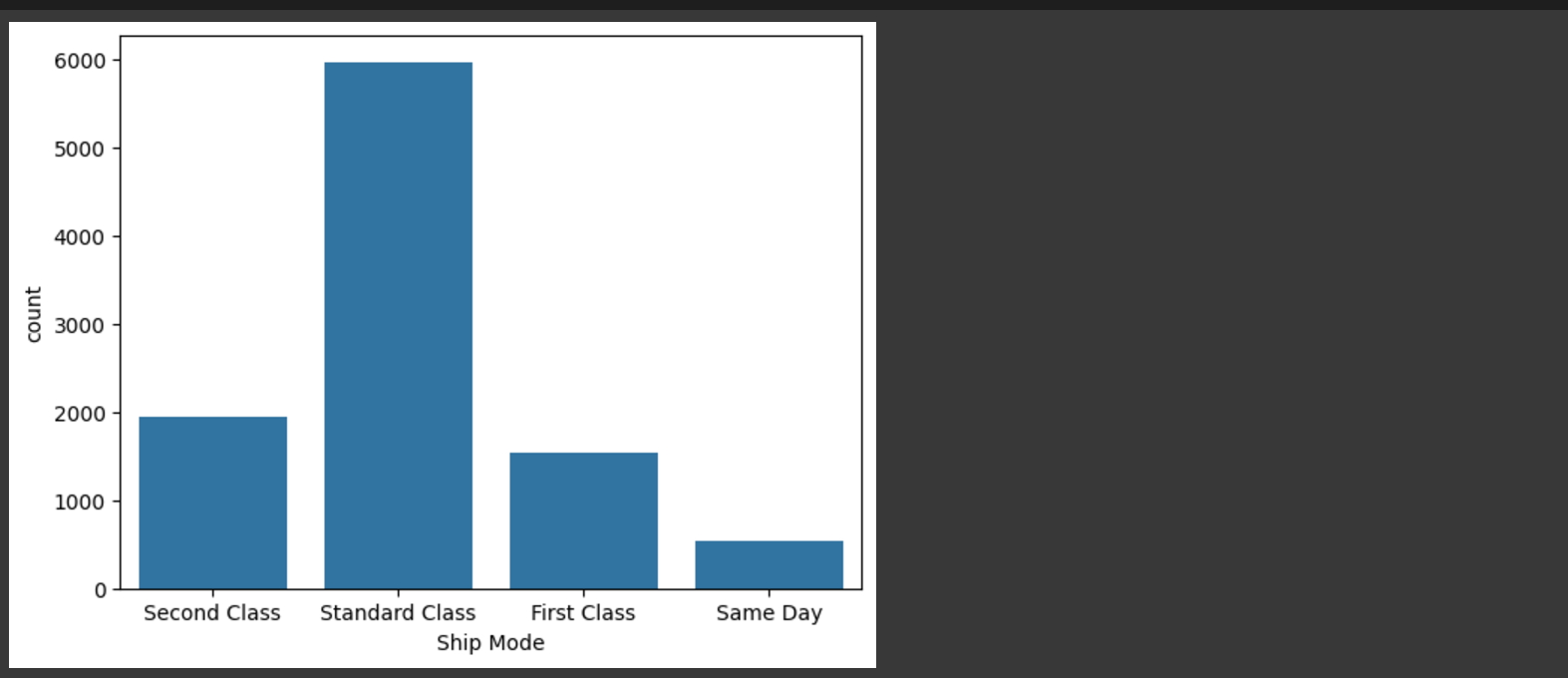


df['Ship Mode'].unique()

df['Ship Mode'].value\_counts()

sns.countplot(x='Ship Mode',data=df)

plt.show()



df['Country'].unique()

df['Country'].value\_counts()

df['State'].unique()

df['State'].value\_counts()



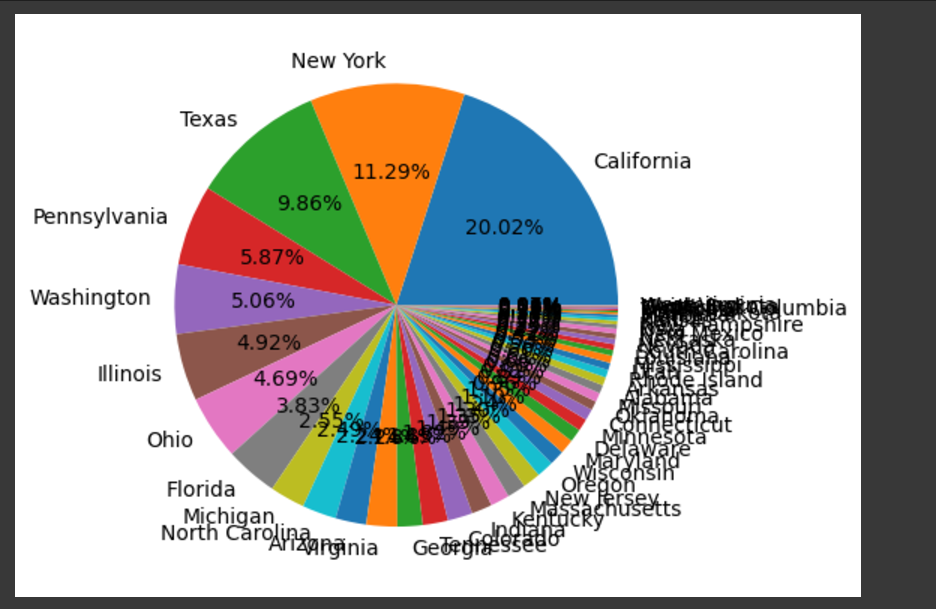


x=df['State'].value\_counts().index

y=df['State'].value\_counts().values

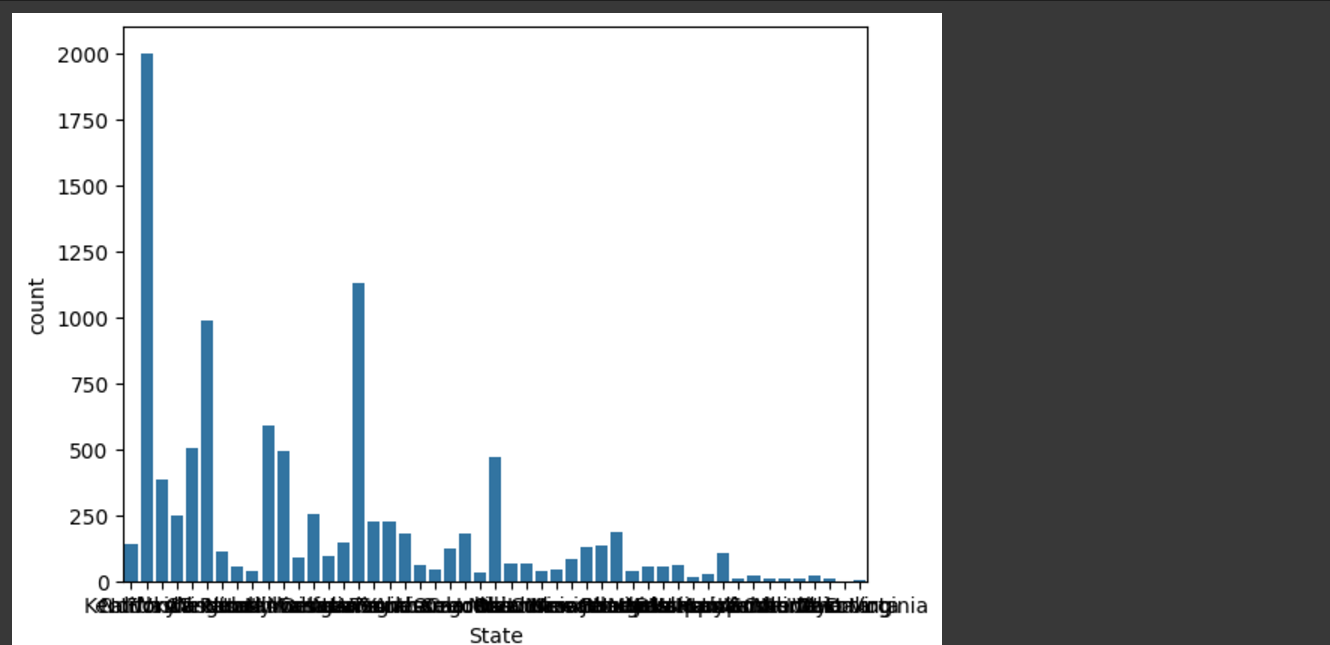
plt.pie(y,labels=x,autopct="%0.2f%%")

plt.show()



sns.countplot(x='State',data=df)

plt.show()



bivariate analysis of each category and their respective shipping modes

sns.countplot(x='Ship Mode',data=df,hue="Category")

plt.show()

plt.figure(figsize=(5,4))

sns.countplot(x='Segment',data=df)

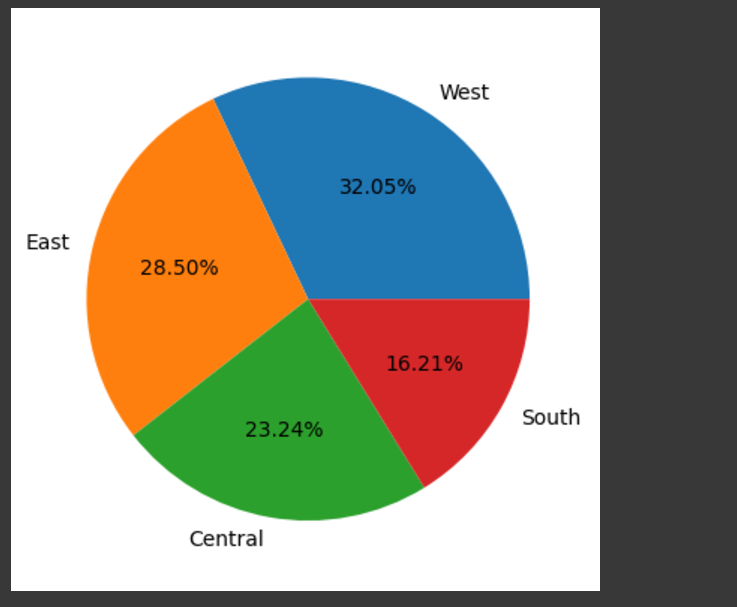
plt.show()

x=df['Region'].value\_counts().index

y=df['Region'].value\_counts().values

plt.pie(y,labels=x,autopct="%0.2f%%")

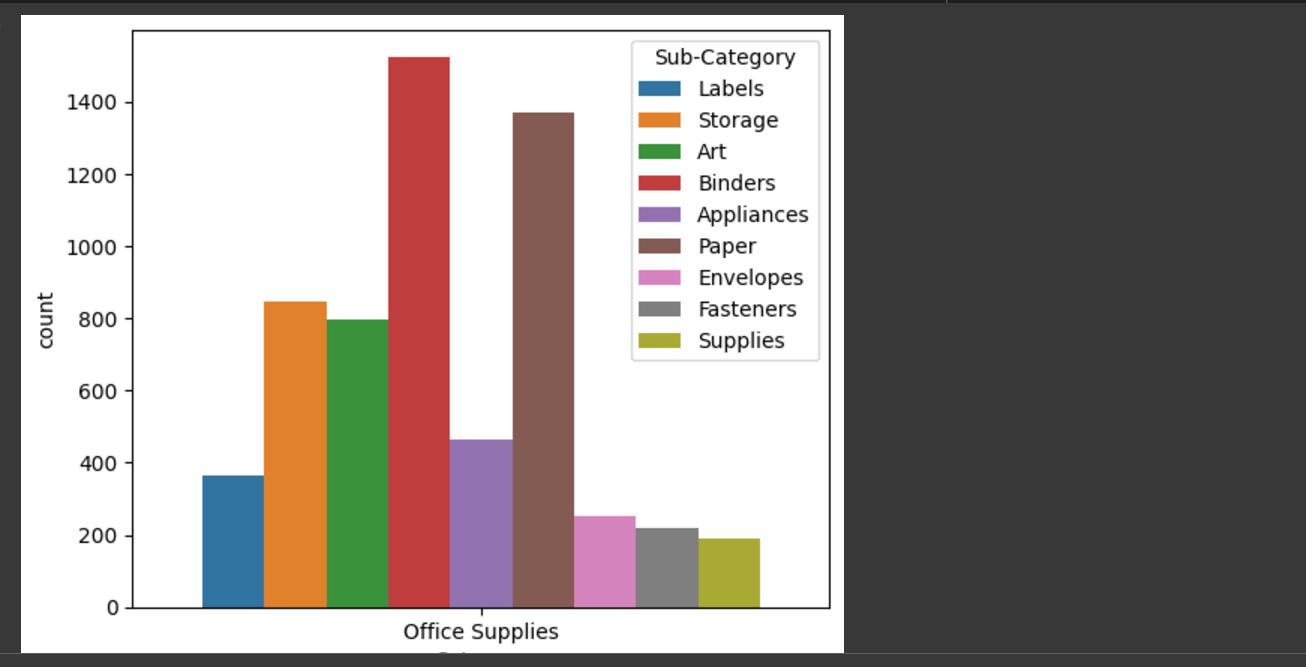
plt.show()



plt.figure(figsize=(6,5))

sns.countplot(x='Category',data=df[df['Category']=="Office Supplies"],hue="Sub-Category")

plt.show()



plt.figure(figsize=(5,5))

sns.countplot(x='Category',data=df[df['Category']=="Furniture"],hue="Sub-Category")

plt.show()

plt.figure(figsize=(5,5))

sns.countplot(x='Category',data=df[df['Category']=="Technology"],hue="Sub-Category")

plt.show()

plt.figure(figsize=(5,5))

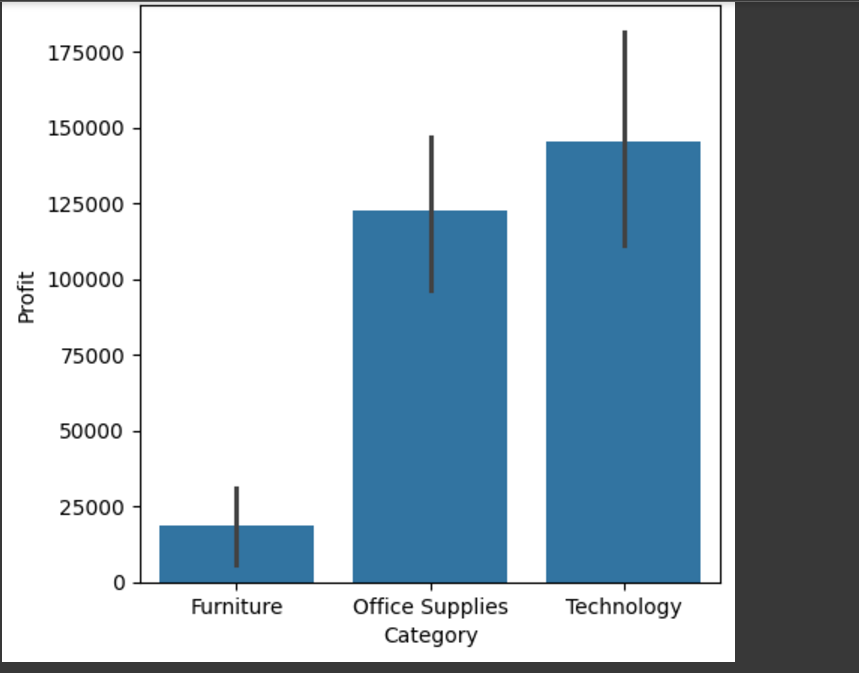
sns.barplot(x='Category',y='Sales',data=df)

plt.show()

plt.figure(figsize=(5,5))

sns.barplot(x='Category',y='Profit',data=df,estimator='sum')

plt.show()



# Convert 'Order Date' to datetime objects

df['Order Date'] = pd.to\_datetime(df['Order Date'])

df['Year'] = df['Order Date'].dt.year

yearly\_sales = df.groupby('Year')['Sales'].sum()

print(yearly\_sales)

plt.figure(figsize=(10, 6))

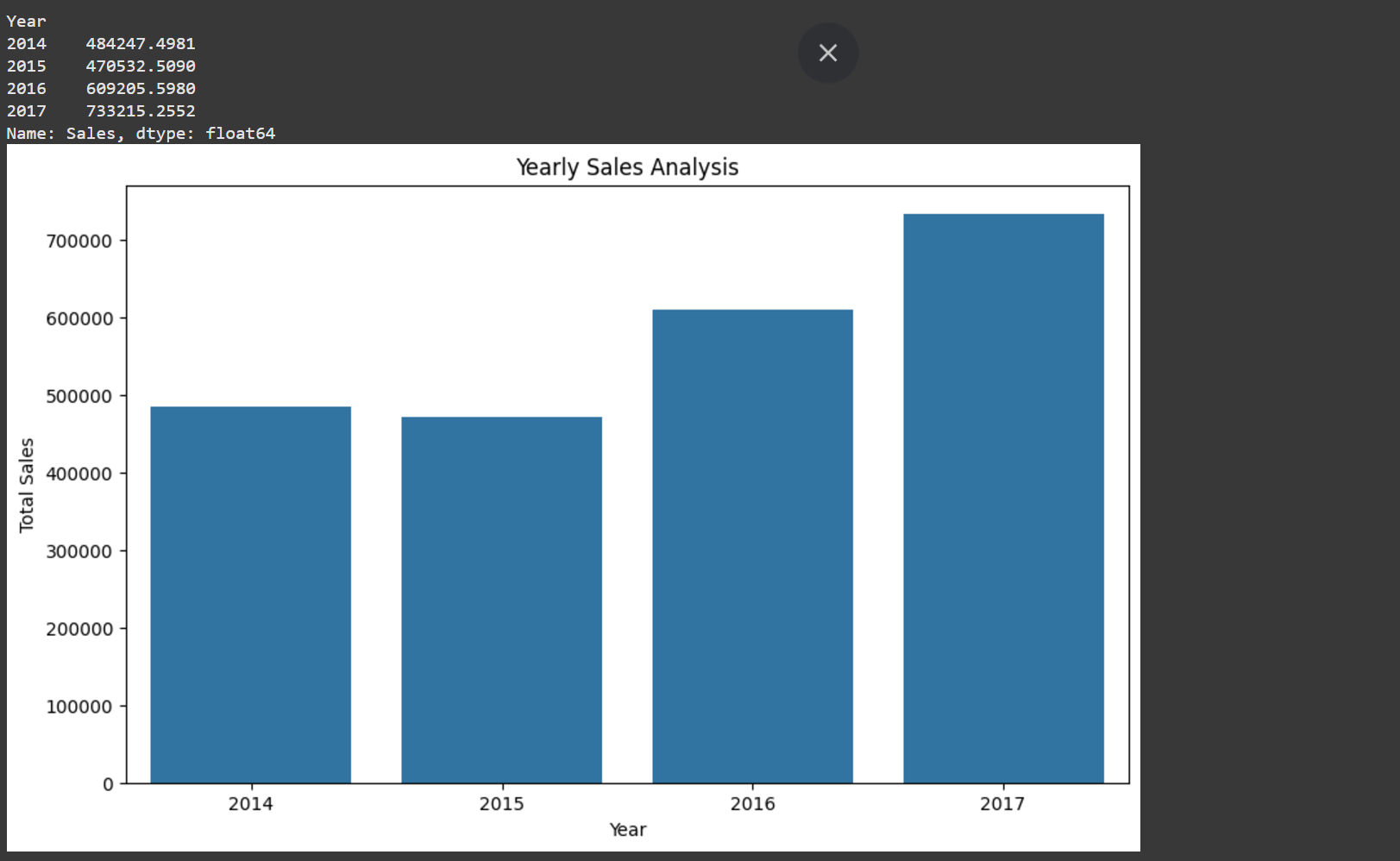
sns.barplot(x=yearly\_sales.index, y=yearly\_sales.values)

plt.xlabel('Year')

plt.ylabel('Total Sales')

plt.title('Yearly Sales Analysis')

plt.show()



# Correlation Analysis

numerical\_features = df.select\_dtypes(include=np.number).columns

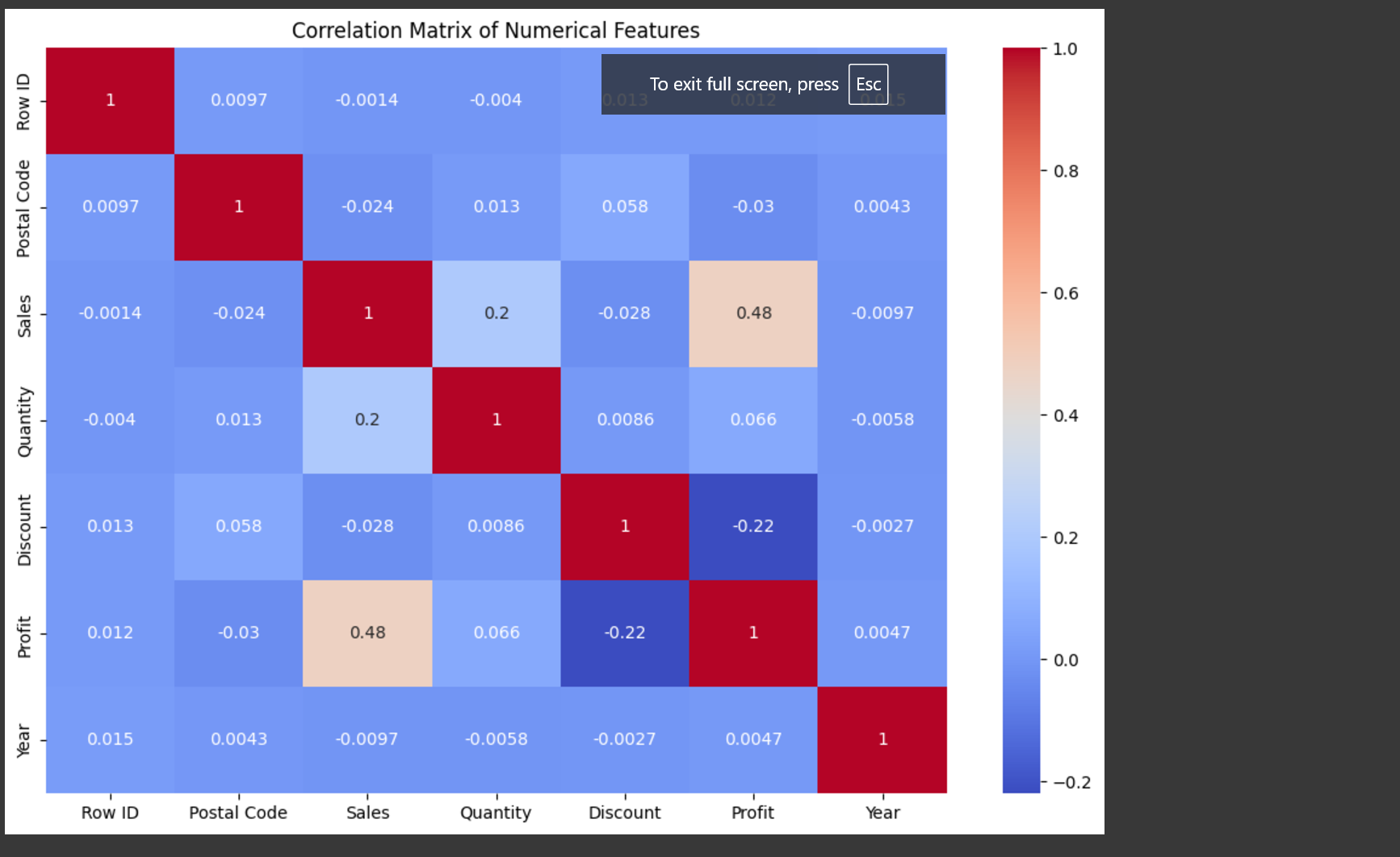
correlation\_matrix = df[numerical\_features].corr()

plt.figure(figsize=(12, 8))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')

plt.title('Correlation Matrix of Numerical Features')

plt.show()



# Sales and Profit Distribution Analysis

plt.figure(figsize=(10, 6))

sns.histplot(df['Sales'], kde=True)

plt.title('Distribution of Sales')

plt.show()

plt.figure(figsize=(10, 6))

sns.histplot(df['Profit'], kde=True)

plt.title('Distribution of Profit')

plt.show()

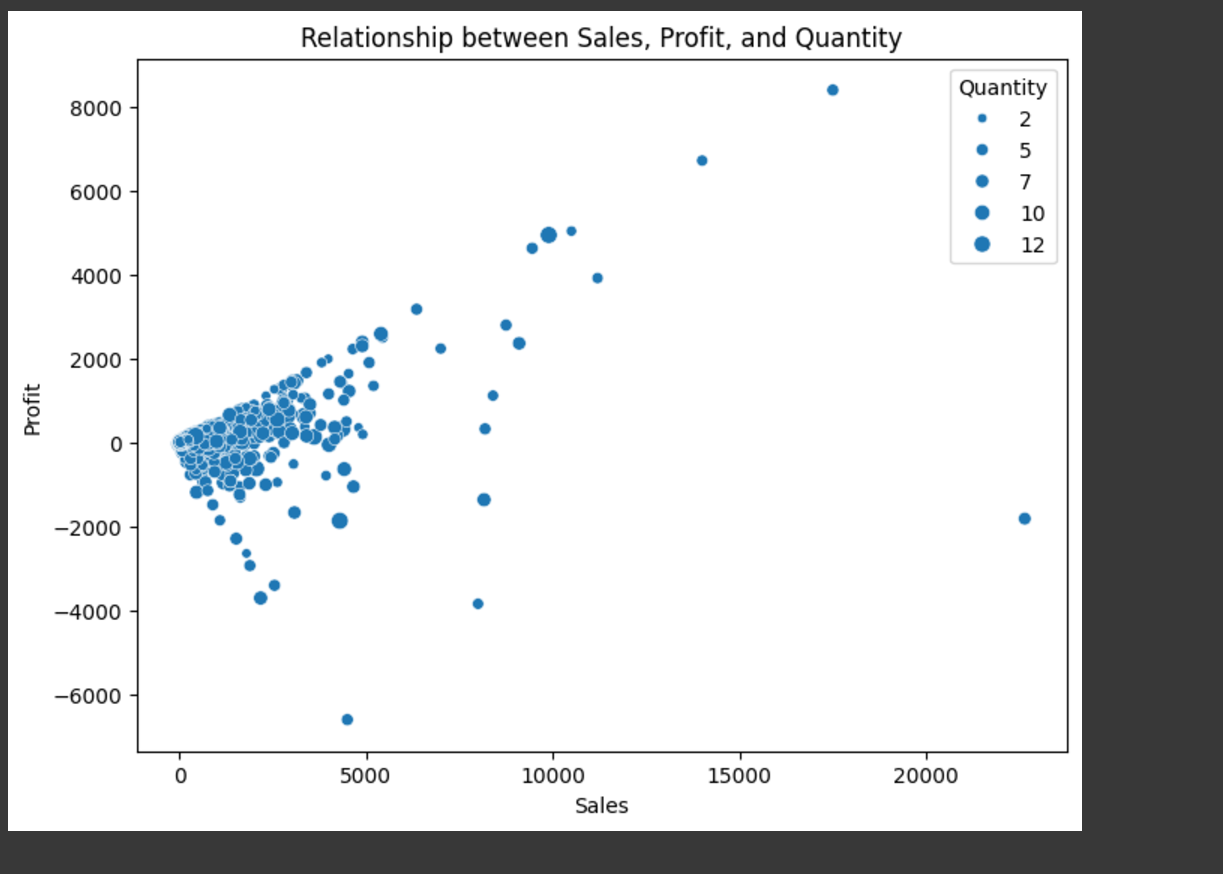
# Relationship between Sales, Profit, and Quantity

plt.figure(figsize=(8, 6))

sns.scatterplot(x='Sales', y='Profit', size='Quantity', data=df)

plt.title('Relationship between Sales, Profit, and Quantity')

plt.show()



# Sales and Profit by Sub-Category

plt.figure(figsize=(14, 8))

sns.barplot(x='Sub-Category', y='Sales', data=df, estimator=sum)

plt.title('Total Sales by Sub-Category')

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

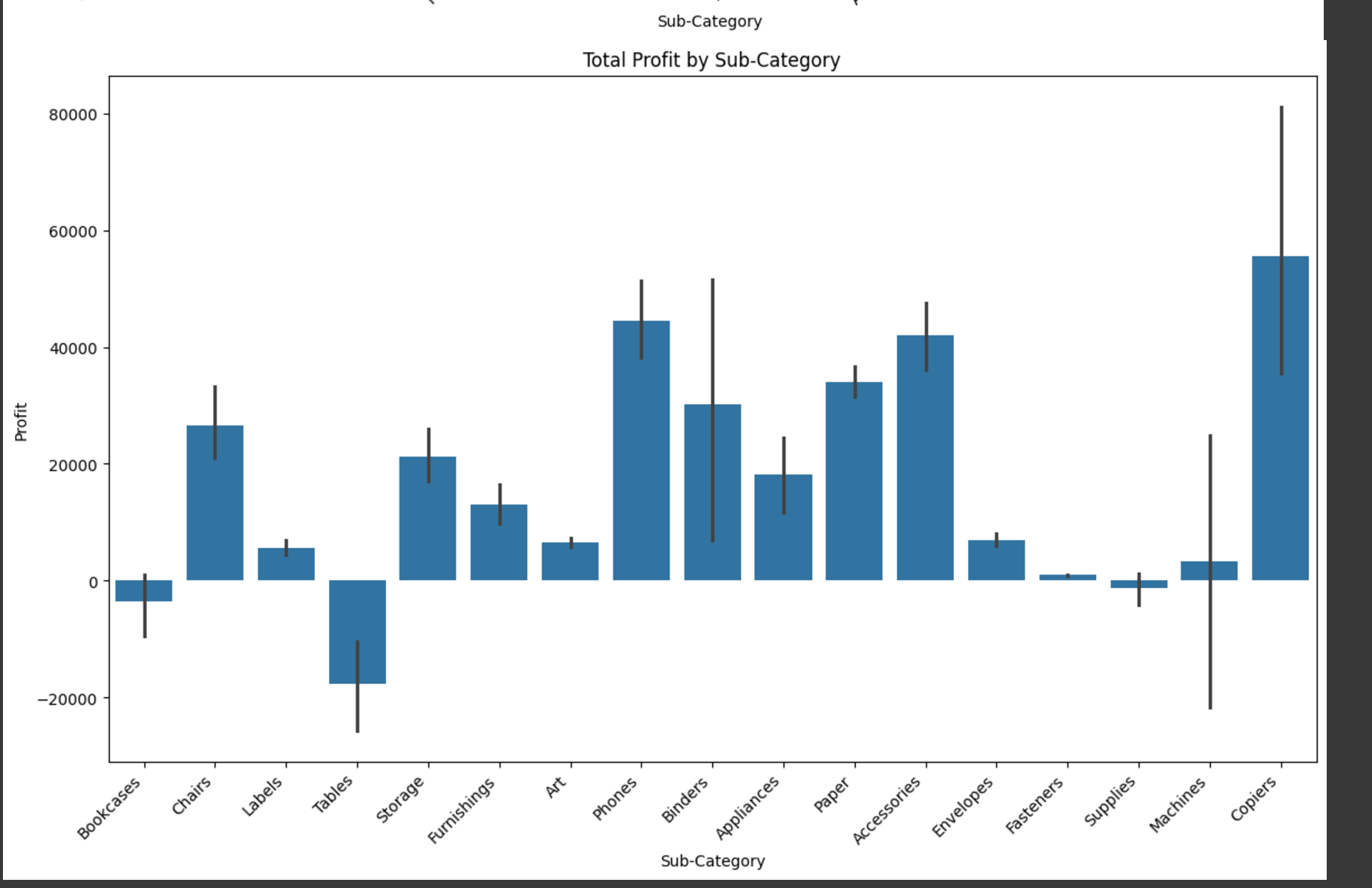
plt.figure(figsize=(14, 8))

sns.barplot(x='Sub-Category', y='Profit', data=df, estimator=sum)

plt.title('Total Profit by Sub-Category')

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

plt.show()



# Sales and Profit Trends over Time

df['Order Date'] = pd.to\_datetime(df['Order Date'])

df = df.set\_index('Order Date')

monthly\_sales = df['Sales'].resample('M').sum()

monthly\_profit = df['Profit'].resample('M').sum()

plt.figure(figsize=(14, 6))

plt.plot(monthly\_sales.index, monthly\_sales.values, label='Monthly Sales')

plt.plot(monthly\_profit.index, monthly\_profit.values, label='Monthly Profit')

plt.xlabel('Date')

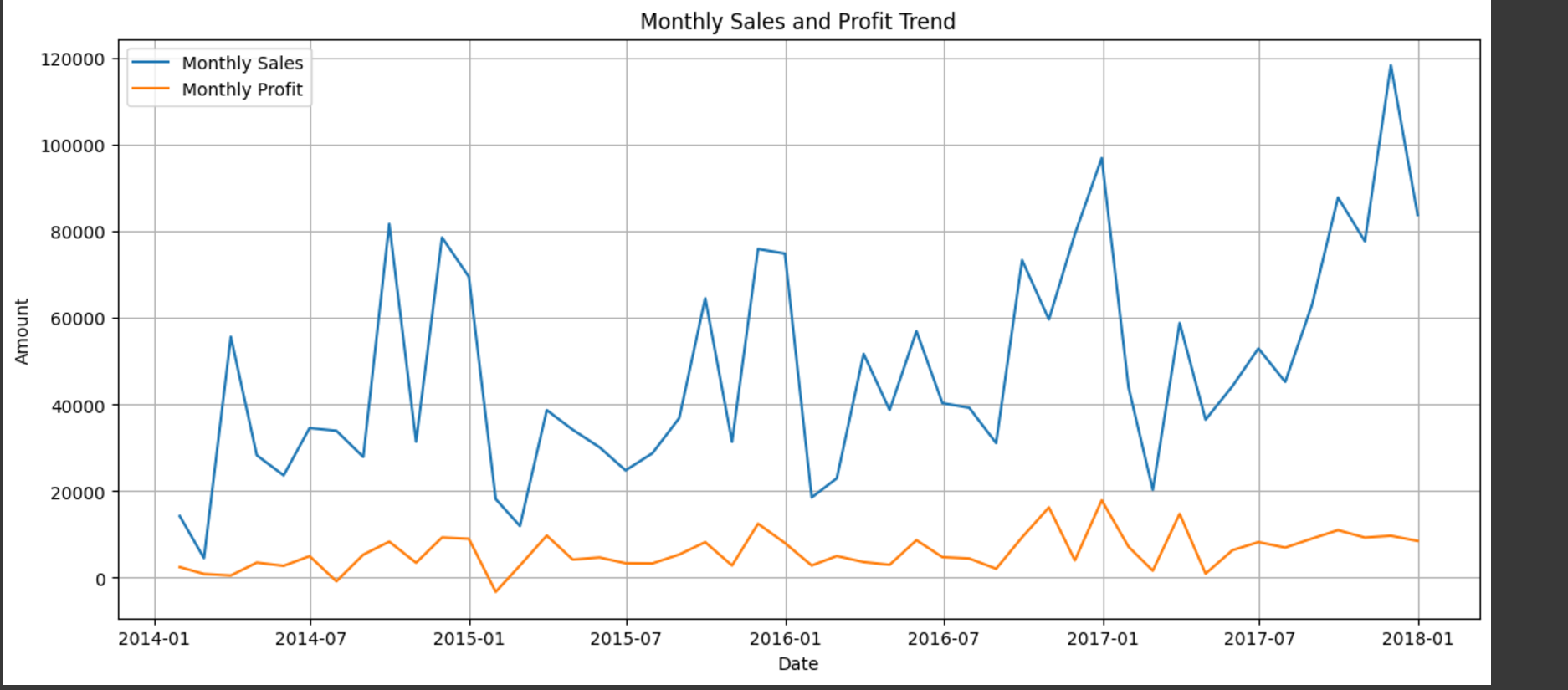
plt.ylabel('Amount')

plt.title('Monthly Sales and Profit Trend')

plt.legend()

plt.grid(True)

plt.show()



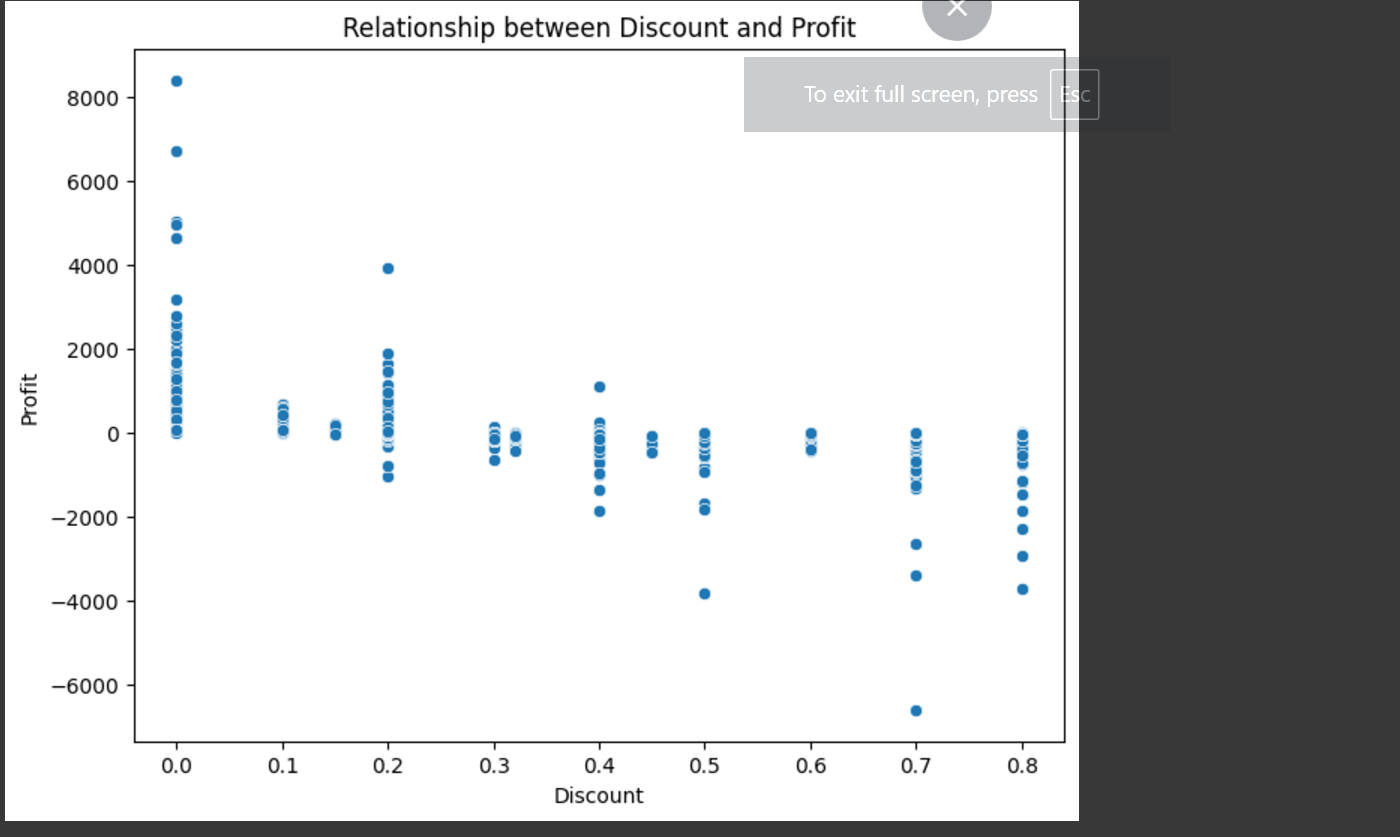
# Explore the relationship between discount and profit

plt.figure(figsize=(8, 6))

sns.scatterplot(x='Discount', y='Profit', data=df)

plt.title('Relationship between Discount and Profit')

plt.show()



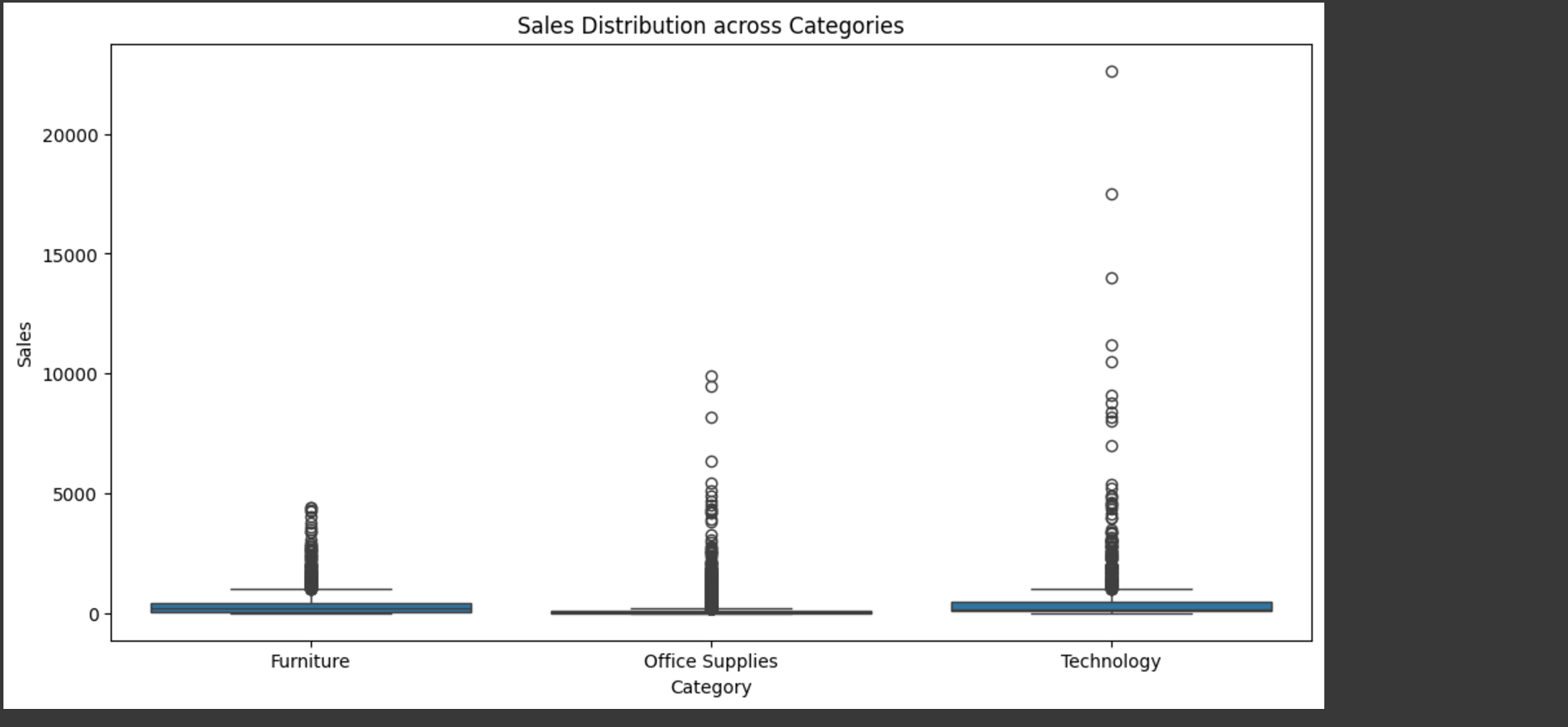
# Box plots to see the distribution of sales and profits across different categories and sub-categories

plt.figure(figsize=(12, 6))

sns.boxplot(x='Category', y='Sales', data=df)

plt.title('Sales Distribution across Categories')

plt.show()



**Conclusion**

The E-commerce Sales Analysis provides valuable insights into the Superstore USA dataset, highlighting key trends and drivers of performance in the e-commerce sector.

* Sales Trends: Sales showed notable growth over time, with peak seasons and product categories like Technology outperforming others.
* Profitability: Discounts significantly influenced profitability, with some subcategories achieving high sales at the expense of reduced profit margins.
* Regional Insights: Certain states and regions contributed disproportionately to sales and profits, suggesting the need for targeted marketing efforts.
* Customer Preferences: Office Supplies dominated in volume, while Technology yielded higher revenues and profits.

Overall, the analysis underscores the importance of data-driven strategies in optimizing operations, inventory, and marketing, enabling better decision-making and improved profitability in e-commerce.

**References**

Dataset: <https://www.kaggle.com/datasets/vivek468/superstore-dataset-final>

 Python Libraries & Documentations:

 Pandas: https://pandas.pydata.org/

 NumPy: https://numpy.org/

 Matplotlib: https://matplotlib.org/

 Seaborn: https://seaborn.pydata.org/