Day : 18.06.2025

**Practice programs**

**1 .You are a data analyst working for an e-commerce company. The company has provided you with a dataset containing information about customer orders. The dataset includes columns such as Order ID, Customer ID, Product ID, Quantity, and Total Price. Your task is to use Pandas data frames to analyze and derive insights from the dataset.**

**· How would you load the dataset into a Pandas data frame?**

**· Can you show the first few rows of the data frame to get an overview of the data?**

**· Calculate the total sales by summing up the Total Price column.**

**· Find the average quantity of products ordered by customers.**

**· Identify the top-selling products based on the Quantity column.**

import pandas as pd

# Load CSV file

df = pd.read\_csv("ecommerce\_orders.csv")

# Show first 5 rows

print("Data Preview:")

print(df.head())

# Total Sales

print("Total Sales:", df['Total Price'].sum())

# Average Quantity

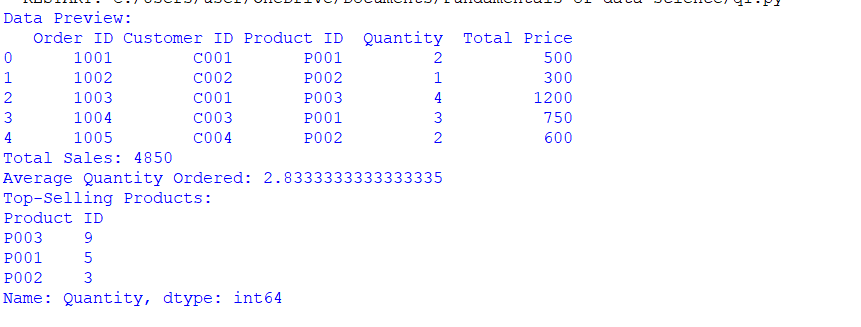
print("Average Quantity Ordered:", df['Quantity'].mean())

# Top-selling Products

top\_products = df.groupby('Product ID')['Quantity'].sum().sort\_values(ascending=False)

print("Top-Selling Products:")

print(top\_products.head())



2. **You are a data scientist working for a medical research institute. The institute is conducting a study to understand the relationship between smoking habits and the incidence of lung cancer among a group of individuals. As part of your analysis, you are tasked with calculating the correlation coefficient between smoking and lung cancer rates and creating a scatter plot to visualize the data.**

**Data = {**

**‘Smoking’: [20, 15, 5, 25, 30, 10, 18, 22, 8, 12],**

**‘LungCancer’: [5, 4, 1, 6, 8, 2, 3, 7, 1, 2]**

**}**

import pandas as pd

import matplotlib.pyplot as plt

data = {'Smoking': [20, 15, 5, 25, 30, 10, 18, 22, 8, 12],

'LungCancer': [5, 4, 1, 6, 8, 2, 3, 7, 1, 2]}

df = pd.DataFrame(data)

print("Correlation:", df['Smoking'].corr(df['LungCancer']))

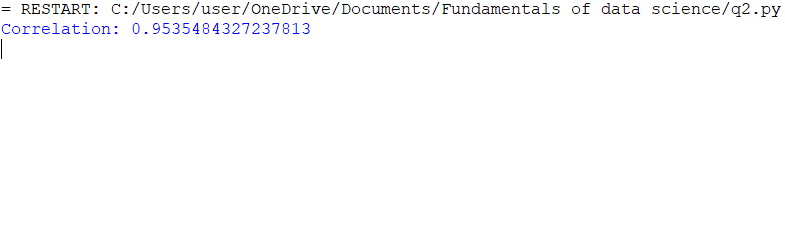
plt.scatter(df['Smoking'], df['LungCancer'])

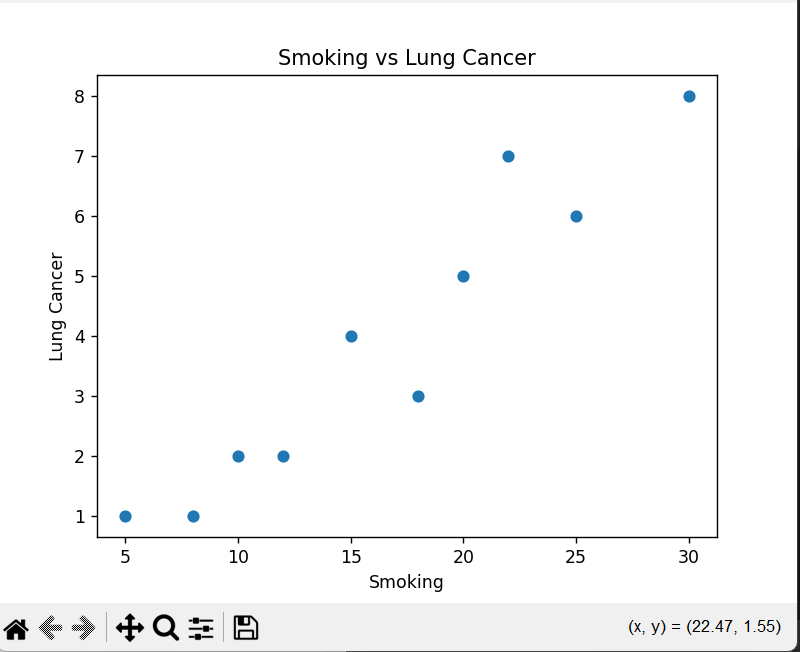
plt.xlabel('Smoking')

plt.ylabel('Lung Cancer')

plt.title('Smoking vs Lung Cancer')

plt.show()





3. **You work as a data analyst for a large e-commerce company that sells a variety of products online. Your company has collected sales data over the past year and wants to analyze and visualize this data to gain insights into sales trends, product performance, and customer behavior. To understand which product categories are most popular, create line, scatter and bar plot that displays the distribution of sales across different product categories. Each plot has to represents a category, and the height of the bar indicates the total sales amount for that category.**

import pandas as pd

import matplotlib.pyplot as plt

# Sample sales data

data = {'ProductCategory': ['Electronics', 'Clothing', 'Home Decor', 'Electronics', 'Clothing', 'Books'],

'SalesAmount': [5000, 3000, 4500, 6000, 2000, 1500]}

df = pd.DataFrame(data)

# Group by category and sum sales

sales\_by\_category = df.groupby('ProductCategory')['SalesAmount'].sum()

# Line plot

sales\_by\_category.plot(kind='line', title='Line Plot')

plt.ylabel('Total Sales')

plt.show()

# Scatter plot

sales\_by\_category.plot(style='o', title='Scatter Plot')

plt.ylabel('Total Sales')

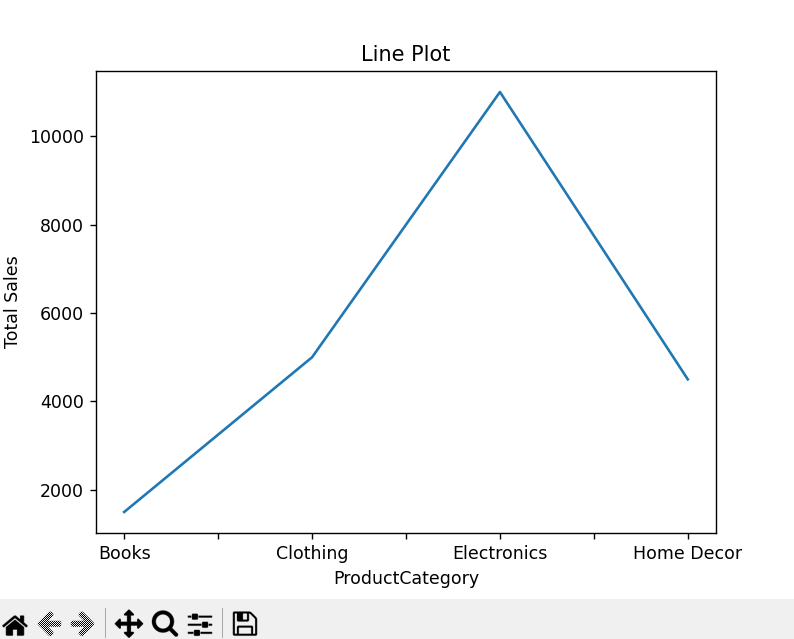
plt.show()

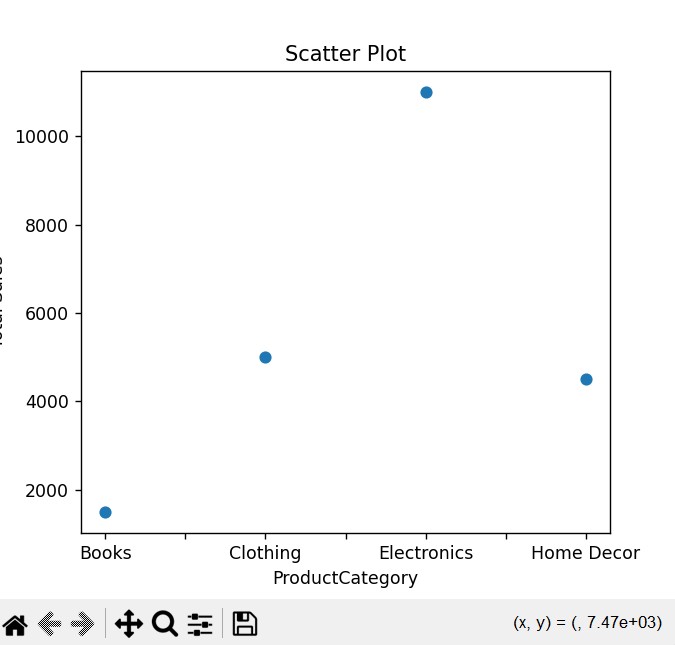
# Bar plot

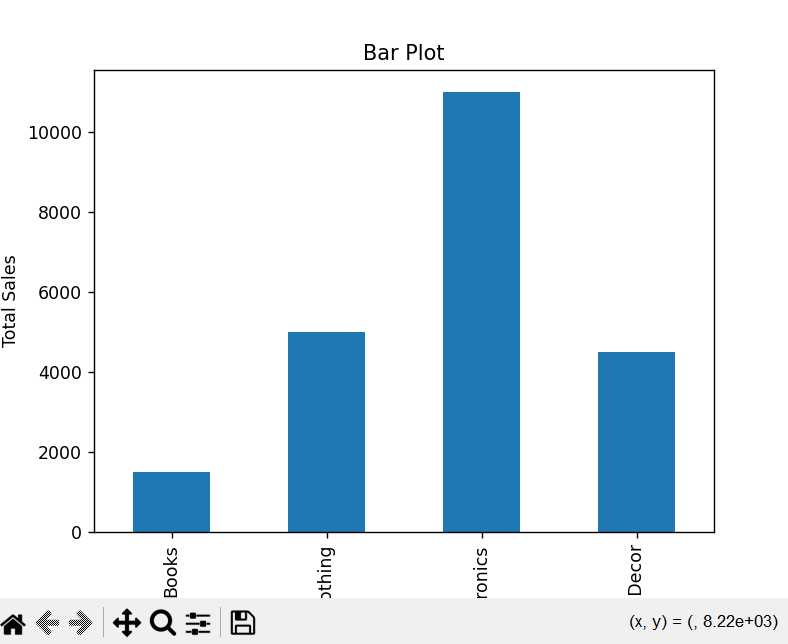
sales\_by\_category.plot(kind='bar', title='Bar Plot')

plt.ylabel('Total Sales')

plt.show()







4. **Scenario: You are a data analyst working for an e-commerce company. The company has provided you with a dataset containing information about customer orders. The dataset includes columns such as OrderID, CustomerID, ProductID, Quantity, and TotalPrice. Your task is to use Pandas data frames to analyze and derive insights from the dataset.**

**● How would you load the dataset into a Pandas data frame?**

**● Create a new column called 'Order Date' and extract the date information from the Order ID.**

**● Filter the data to show only orders placed by a specific customer (choose a Customer ID).**

**● Group the data by Customer ID and calculate the total amount spent by each customer.**

**● Visualize the distribution of Total Price using a histogram**

import pandas as pd

import matplotlib.pyplot as plt

# Sample dataset

data = {

'OrderID': ['20240101-001', '20240102-002', '20240101-003', '20240103-001'],

'CustomerID': ['C1', 'C2', 'C1', 'C3'],

'ProductID': ['P1', 'P2', 'P3', 'P1'],

'Quantity': [2, 1, 3, 4],

'TotalPrice': [200, 150, 300, 400]

}

df = pd.DataFrame(data)

# Extract 'Order Date' from OrderID (first 8 digits assumed to be YYYYMMDD)

df['Order Date'] = pd.to\_datetime(df['OrderID'].str[:8])

# Filter for a specific customer

customer\_orders = df[df['CustomerID'] == 'C1']

# Total spent by each customer

total\_by\_customer = df.groupby('CustomerID')['TotalPrice'].sum()

# Histogram of Total Price

plt.hist(df['TotalPrice'], bins=5)

plt.title('Distribution of Total Price')

plt.xlabel('Total Price')

plt.ylabel('Frequency')

plt.show()

