PYTHON FOR DATA SCIENCE : ASSIGNMENT 1

PES2UG22EC035- BRINDA CHAUHAN

Multiple Choice Questions (MCQ)

Question 1: Given a NumPy array arr = np.array([10, 20, 30, 40, 50]), what is the output of the following code?

Python

import numpy as np

arr = np.array([10, 20, 30, 40, 50])

print(arr[1:4])

(A) [10, 20, 30]

(B) [10, 20, 30, 40]

(C) [20, 30, 40]

(D) [20, 30, 40, 50]

Answer: (C) [20, 30, 40]

Question 2: If you have a Pandas DataFrame df with a column named Age, which of the following lines of code will correctly calculate the average age, ignoring any missing values?

(A) df['Age'].sum() / len(df['Age'])

(B) df['Age'].average()

(C) df['Age'].mean()

(D) df.mean('Age')

Answer: (C) df['Age'].mean()

Subjective Coding Questions

Question 1:

You are given a string representing a CSV file of sales data for different regions. Your task is to use the Pandas library to:

1. Read this CSV data into a DataFrame.
2. Group the data by the Region column.

3. Calculate the total Sales for each region.

4. Print the resulting total sales for each region.

Data:

Region ,Sales ,Product

North,100,A

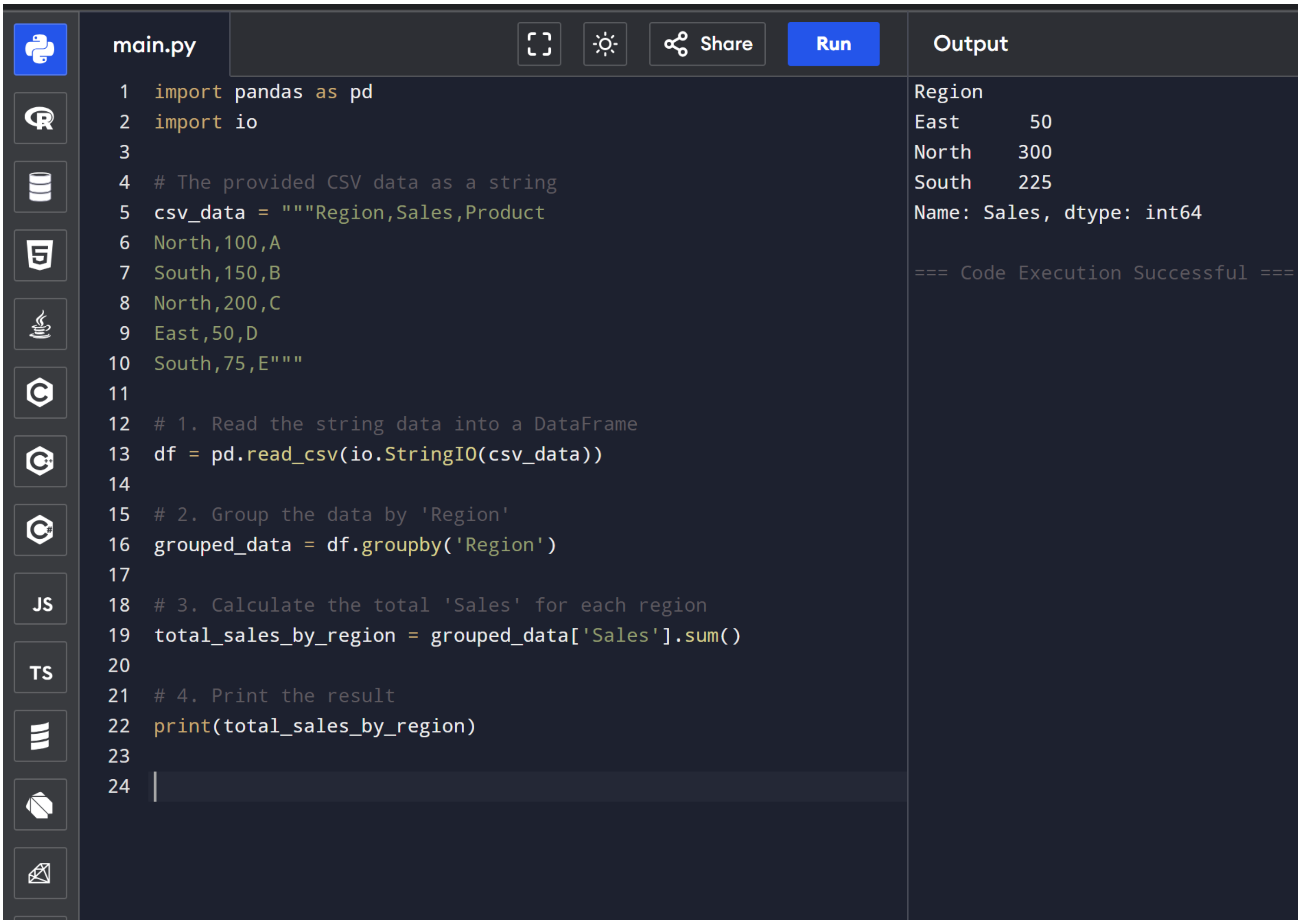
South,150,B

North,200,C

East,50,D

South,75,E

Answer:



```python
import pandas as pd
import io

# The provided CSV data as a string
csv_data = """Region,Sales,Product
North,100,A
South,150,B
North,200,C
East,50,D
South,75,E"""

# 1. Read the string data into a DataFrame
df = pd.read_csv(io.StringIO(csv_data))

# 2. Group the data by 'Region'
grouped_data = df.groupby('Region')

# 3. Calculate the total 'Sales' for each region
total_sales_by_region = grouped_data['Sales'].sum()

# 4. Print the result
print(total_sales_by_region)

```

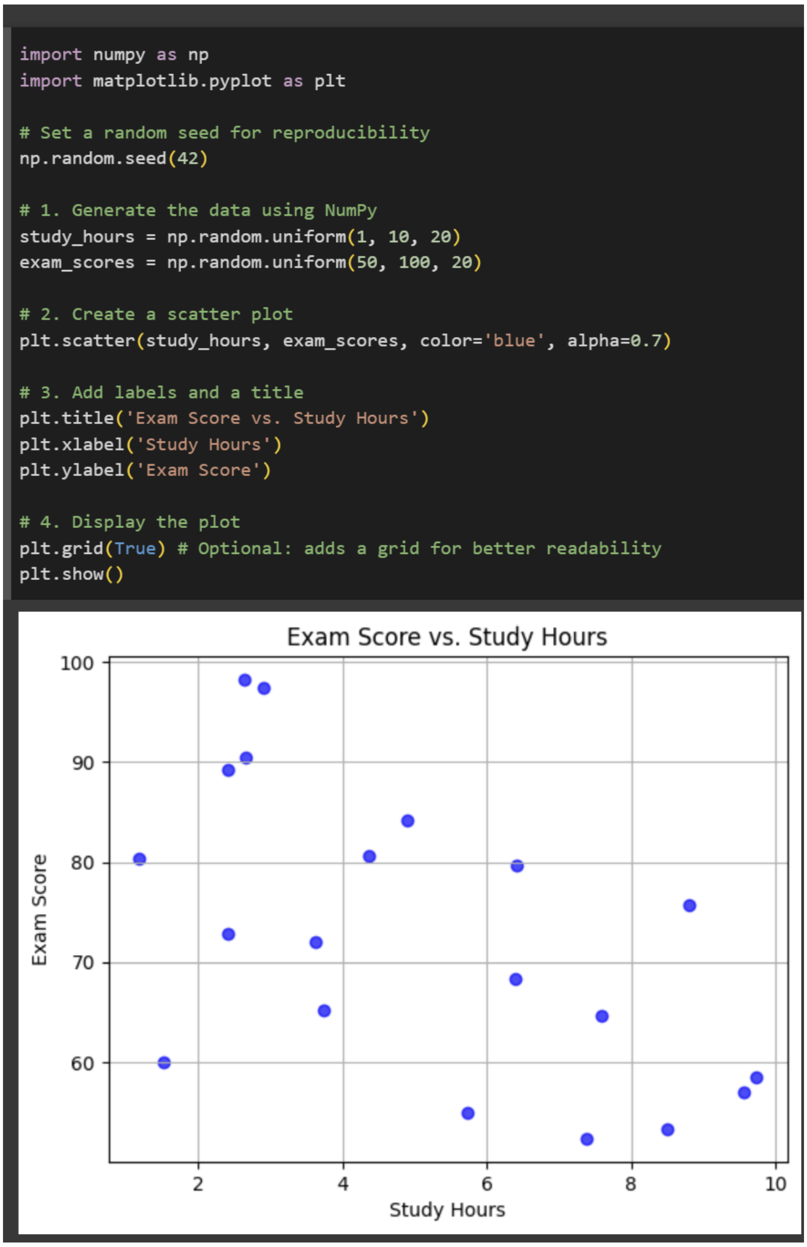
```
Region
East      50
North    300
South    225
Name: Sales, dtype: int64

=== Code Execution Successful ===
```

Question 2:

Your task is to create a scatter plot to visualize the relationship between two hypothetical variables: "Study Hours" and "Exam Score."

1. Use the NumPy library to generate two arrays:
   - study_hours: An array of 20 random numbers between 1 and 10.
   - exam_scores: An array of 20 random numbers between 50 and 100.
2. Use the Matplotlib library to create a scatter plot of exam_scores against study_hours.
3. Add appropriate labels for the x-axis ("Study Hours") and y-axis ("Exam Score"), and give the plot a title ("Exam Score vs. Study Hours").
4. Display the plot.

   ANSWER:



```python
import numpy as np
import matplotlib.pyplot as plt

# Set a random seed for reproducibility
np.random.seed(42)

# 1. Generate the data using NumPy
study_hours = np.random.uniform(1, 10, 20)
exam_scores = np.random.uniform(50, 100, 20)

# 2. Create a scatter plot
plt.scatter(study_hours, exam_scores, color='blue', alpha=0.7)

# 3. Add labels and a title
plt.title('Exam Score vs. Study Hours')
plt.xlabel('Study Hours')
plt.ylabel('Exam Score')

# 4. Display the plot
plt.grid(True) # Optional: adds a grid for better readability
plt.show()
```