

```
ax.add_patch(triangle)
ax.set_aspect('equal', 'box')
ax.set_xlim(0, 1)
ax.set_ylim(0, 1)
plt.show()
```

(b) State differences between following by giving appropriate examples : (9)

- (i) Simple random sampling and Stratified sampling
- (ii) rand(), randint() functions in NumPy.
- (iii) Deep and shallow copy in NumPy array.

[This question paper contains 12 printed pages.]

Your Roll No.....

Sr. No. of Question Paper : 5325

J

Unique Paper Code : 2344001201

Name of the Paper : Data Analysis and Visualization using Python

Name of the Course : **Generic Elective (G.E.)**

Semester : II

Duration : 3 Hours

Maximum Marks : 90

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. Paper has **two** sections. All the questions in **Section A** are compulsory.
3. Answer any **four** questions from **Section B**.
4. Parts of question must be answered together.
5. Use of simple calculator is permitted.

Section A
(Compulsory)

1. (a) What is the purpose of using Matplotlib in Python?
Describe briefly any three types of plots that can be created using Matplotlib. (1+3=4)

- (b) What is the output of following code, give reasons :
(2+2=4)

```
import pandas as pd
import numpy as np
df = pd.DataFrame({
    'key1': ['a', 'a', 'b', 'b', 'a'],
    'key2': ['one', 'two', 'one', 'two', 'one'],
    'data1': np.arange(5),
    'data2': np.random.randn(5)
})
print(df)

mean_data = df['data1'].groupby([df['key1'],
df['key2']]).mean()
print(mean_data)
```

- (iv) Using `pd.crosstab()`, display a frequency table showing the count of students for each course and device combination. Add totals using `margins=True`.

7. (a) Find the output of the following Python code. Illustrate the shapes along with their dimensions and colour that will be displayed on the plot. (6)

```
import matplotlib.pyplot as plt
import matplotlib.patches as patches
fig, ax = plt.subplots()
rectangle = patches.Rectangle((0.1, 0.1), 0.3, 0.4,
linewidth=2,
edgecolor='blue', facecolor='lightblue')
ax.add_patch(rectangle)
circle = patches.Circle((0.7, 0.7), 0.2, linewidth=2,
edgecolor='green',
facecolor='lightgreen')
ax.add_patch(circle)
triangle = patches.Polygon([(0.5, 0.1), (0.9, 0.5),
(0.1, 0.5)],
closed=True, linewidth=2, edgecolor='red',
facecolor='lightcoral')
```

student_id	course	gender	day	device	score	attempts
S101	Python	Female	Monday	Mobile	78	2
S102	Java	Male	Tuesday	Laptop	85	1
S103	Python	Male	Monday	Laptop	74	3
S104	DataSci	Female	Wednesday	Mobile	90	2
S105	Java	Male	Monday	Tablet	68	1
S106	Python	Female	Tuesday	Laptop	88	2
S107	DataSci	Male	Wednesday	Mobile	76	2
S108	Java	Female	Monday	Laptop	92	1

- (i) Create a DataFrame in Python using the above data.
- (ii) Add a new column score_pct to store the percentage score (i.e., score / 100).
- (iii) Using Pandas pivot_table() method:
 - (a) Display the average score_pct and average attempts
 - (b) Use gender and day as row indices and device as columns
 - (c) Use margins = True to include row and column totals
 - (d) Round the result to 2 decimal places

- (c) A data analyst is working with a survey dataset that includes fields like **gender**, **age**, **income**, **education level**, and **satisfaction rating**.

Classify each field as either qualitative or quantitative, and identify the type of data (Nominal, Ordinal, Interval or Ratio) associated with each.

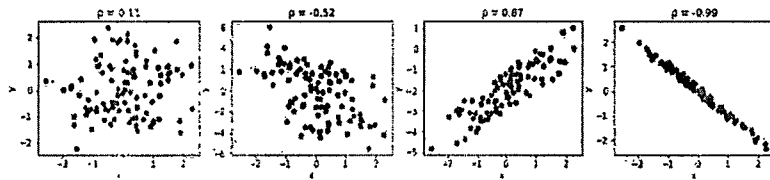
(5)

- (d) What is a NumPy ndarray? How does it differ from a Python list? Consider the following code snippet. Predict the output and explain the reason behind the result:

```
import numpy as np
arr = np.arange(10)
sliced = arr[2:6]
sliced:] = 99
print(arr)
```

(1+2+2=5)

- (e) What is correlation? How does it help in understanding relationships between two variables? Based on the scatter plots shown below, estimate the type and strength of correlation for each plot (e.g., strong positive, weak negative, no correlation, etc.). Justify your answer. (1+1+4=6)



(f) Create a Pandas DataFrame that holds employee data with the columns : (2+2+2=6)

Name, Age, and Salary, including some missing values.

Then, perform the following actions :

(i) Drop all rows that contain missing values.

(ii) Filter and display the employee records where the Salary is greater than ₹50,000.

Section B

2. (a) Describe the key stages in the data analysis process. For each step, explain its purpose and significance using relevant examples. (5)

(ii) import numpy as np

```
arr = np.arange(1, 10).reshape(3, 3)
```

```
print(arr)
```

```
print(arr[1:, 1:])
```

(iii) import pandas as pd

```
data = [1, 5, 9, 12, 15, 18, 22, 25, 30]
```

```
bins = [0, 10, 20, 30]
```

```
output = pd.cut(data, bins)
```

```
print(output)
```

6. (a) Given the dataset [12, 18, 20, 25, 30], perform Min-Max scaling to the range [0, 1]. Show the minimum and maximum values, the scaling formula used, and the resulting scaled dataset. (3)

- (b) You are given the following dataset of student enrollments in various online courses:

(2+2+4+4=12)

5. (a) Write a Python program to plot a quadratic function using NumPy and Matplotlib.

$$(1+1+1+2+1=6)$$

- (i) Generate a list of integer values from -10 to 10 using a suitable NumPy function
- (ii) Compute the corresponding values of the function $y = x^2 - 4x + 3$
- (iii) Plot the function using a blue dotted line
- (iv) Add suitable labels for the X and Y axes and a meaningful title
- (v) Display the plot

- (b) What will be the output of the following? Explain your answer. (9)

```
(i) import numpy as np
a = np.array([1, 2, 3, 4])
b = np.array([5, 6, 7, 8])
print(a + b)
print(a * b)
```

- (b) The IRIS dataset is a well-known multivariate dataset that contains 150 records of iris flowers, with the following four features (attributes):

$$(2+2+3+3=10)$$

sepal length (cm), sepal width (cm), petal length (cm), petal width (cm).

Each flower sample belongs to one of three species (class labels): *Setosa*, *Versicolor*, or *Virginica*.

Write a Python program to perform the following operations on the IRIS dataset:

- (i) Load the IRIS dataset using `sklearn.datasets`.
- (ii) Display the feature names and target (class) names.
- (iii) Compute and display the mean, median, and standard deviation for each feature.
- (iv) Plot a correlation heatmap for the features.

3. (a) Explain what hierarchical indexing is in pandas. How do you create and access data using hierarchical indexing? (2+3=5)

(b) You are given Sales Data as following :

(2+2+2+2+2=10)

```
import pandas as pd
```

```
data = {'Month': ['Jan', 'Feb', 'Mar', 'Apr',
                  'May', 'Jun'], 'Sales_A': [100, 120, 130, 90, 110,
                  150], 'Sales_B': [80, 95, 100, 70, 85, 120]}
}
```

```
df = pd.DataFrame(data)
```

From above dataset containing monthly sales data for two products —

Product A and Product B, write the code for the following :

- (i) Create a line plot showing monthly sales of both products on the same chart.
- (ii) Add labels for the x-axis (Month) and y-axis (Sales).
- (iii) Add a legend to distinguish between Product A and Product B.

(iv) Create a bar chart to compare total sales for each product across the 6-month period.

(v) Add a title to each plot and display them clearly.

4. Write a Python program using NumPy to perform the following tasks : (2+2+1+2+2+6=15)

- (i) Create a 3×4 NumPy array containing random integers between 10 and 50.
- (ii) Slice and display the first two rows of the array.
- (iii) Replace the last column of the array with zeros.
- (iv) Transpose the modified array and display its shape.
- (v) Compute and display the mean and standard deviation along each column of the original (modified) array.
- (vi) Explain the use of the following ndarray attributes with suitable Python code examples:
.shape, .ndim and .dtype