

SL No of QP :10348

Unique paper code: 6203450015

Name of Paper: Data Analysis and Visualization using Python

Name of the Course: B.Voc Software Development

Semester: 4

Duration: 3 Hours

Maximum Marks: 90

Instructions for Candidates:

- Write your Roll No. on the top immediately on receipt of question paper.
- The paper has two sections. Section A is compulsory. Each question is of 5 marks.
- Attempt any four questions from Section B. Each question is of 15 marks.

Section A

- 1 a) Given the following data representing the marks obtained by students in a semester examination: 5
import numpy as np

```
marks = np.array([32, 45, 51, 60, 65, 68, 70, 72, 75, 78, 80, 85, 90, 92, 95])
```

Write Python commands to:

(i) Calculate the first quartile (Q1), third quartile (Q3), and the Interquartile Range (IQR)

(ii) Determine the lower limit and upper limit for detecting outliers using the $1.5 \times \text{IQR}$ method.

(iii) Identify and display the outlier values, if present in the dataset.

- b) Consider the following data showing the number of employees in different departments of an organization: 5

```
departments = ['HR', 'Finance', 'IT', 'Marketing', 'Operations']
```

```
employees = [45, 60, 80, 55, 70]
```

Write Python commands to:

(i) Create a horizontal bar chart representing the number of employees in each department.

(ii) Display the employee count on each bar.

(iii) Set the figure size to (8, 5).

(iv) Add the title "Department-wise Employee Strength".

- c) What is data cleaning? List and explain any three common data cleaning techniques. 5
- d) Create two NumPy arrays arr1 and arr2 of size 3×3 each. Perform element-wise addition, multiplication, and division. 5
- e) What do you understand by multi-level indexing in Pandas? Explain with example 5
- f) Consider the following DataFrame: 5
import pandas as pd

```
data = pd.DataFrame({
```

```
'Department': ['HR', 'HR', 'IT', 'IT', 'Sales', 'Sales'],
'Employee': ['Amit', 'Riya', 'Karan', 'Neha', 'Pooja', 'Rahul'],
'Salary': [45000, 50000, 60000, 65000, 48000, 52000]
})
```

Write Python commands to:

- (i) Group the data by Department and calculate the average salary of each department.
- (ii) Display the maximum salary paid in each department.
- (iii) Rename the grouped column as 'Average_Salary'.

Section B

- 2 a) Consider the following Pandas Series: 7
- ```
import pandas as pd
import numpy as np
```

```
scores = pd.Series([45, 60, np.nan, 72, 88, np.nan, 50])
```

Write Python commands to:

- (i) Replace all missing (NaN) values with the mean of the Series.
- (ii) Display all elements having values greater than 60.

- b) Consider the following code: 8

```
import pandas as pd
```

```
df = pd.DataFrame({
 'Department': ['HR', 'HR', 'IT', 'IT', 'Sales', 'Sales'],
 'Salary': [45000, 52000, 60000, 65000, 48000, 50000],
 'Bonus': [5000, 6000, 7000, 8000, 5500, 5800]
})
```

Write Python commands to:

- (i) Group the data by Department and calculate the sum and mean of both numeric columns.
- (ii) Use the agg() function to compute the minimum Salary and maximum Bonus for each department.
- (iii) Display the final grouped DataFrame with appropriate column names.

- 3 a) Explain the difference between variance and standard deviation with an example using NumPy arrays. 7

- b) Write a Python program using Pandas and NumPy to: 8

- (i) Create a DataFrame of 6 rows and 3 columns containing random integers between 10 and 100.
- (ii) Assign custom row labels as ['R1', 'R2', 'R3', 'R4', 'R5', 'R6'] and column names as ['Maths', 'Science', 'Computer'].
- (iii) Display the DataFrame.

- 4 a) Differentiate between Min-Max scaling and Z-score scaling. Where is each method preferred? 7

- b) Consider the following NumPy arrays: 8

```
import numpy as np
```

```
x = np.array([[4, 6, 8],
 [10, 12, 14]])
```

```
y = np.array([[5, 7, 9],
 [11, 13, 15],
 [2, 4, 6]])
```

Give the output of the following commands:

- (i) `y[2, :]`
- (ii) `x[:, 1]`
- (iii) `x + 5`
- (iv) `y > 10`
- (v) `y[1, 0] = 50` — What change will occur in array `y` after executing this statement?

- 5 a) Consider the following Pandas Series showing the monthly profit (in lakhs) of different companies: 7

```
Company_A 120
Company_B 75
Company_C 150
Company_D 90
Company_E 110
```

Write Python commands to:

- (i) Display the companies having profit greater than 100.
  - (ii) Display the index labels of the Series.
  - (iii) Assign the name 'Company\_Name' to the index.
- b) Create a DataFrame with Student Name, Course, and Marks. 8

Perform the following operations:

- (i) Sort the students by Marks in ascending order.
  - (ii) Add a new column "Grade" based on Marks  
(Grade = "Pass" if Marks  $\geq$  40, else "Fail").
  - (iii) Remove the column "Course".
  - (iv) Rename the columns:  
    'Student Name'  $\rightarrow$  Student\_Name,  
    'Marks'  $\rightarrow$  Final\_Marks.
- 6 a) Consider a CSV file named `employees.csv` that contains employee information with the following columns: 7  
Emp\_ID, Name, Department, Salary, Experience (80 rows).

Write Python commands to:

- (i) Read the file `employees.csv` into a DataFrame named `emp_data`.
  - (ii) Display the first 10 rows of the dataset.
  - (iii) Generate the summary statistics (mean, standard deviation, minimum, maximum) for all numeric columns.
  - (iv) Delete all rows where all column values are missing.
  - (v) Detect and display any duplicate employee entries.
- b) Write Python code to: 8
- (i) Create a NumPy array `scores` containing 50 floating-point values uniformly distributed between 0 and 100.
  - (ii) Divide the data into 4 equal-width bins.
  - (iii) Assign labels to the bins as:  
    ['Poor', 'Average', 'Good', 'Excellent'].
  - (iv) Round the values to 1 decimal place and display both the values and their corresponding bin labels.

- 7 a) Write Python code using NumPy to: 7
- (i) Create a 3D NumPy array arr of size  $3 \times 2 \times 4$  containing random integers between 10 and 100.
  - (ii) Display the transpose of the array.
  - (iii) Swap axis 0 with axis 2 and print the shape of the resulting array.
  - (iv) Display the data type of the elements present in the array.
- b) You have a file products.txt where values are separated by commas (,): 8

item, cost, stock  
notebook,50,120  
eraser,5,300

Write Python code to read the file correctly into a Pandas DataFrame and display the DataFrame.