[This question paper contains 4 printed pages.]

Your Roll No.....

Sr. No. of Question Paper : 9082 **K**

Unique Paper Code : 6203450015

Name of the Paper : Data Analysis and Visualization

Name of the Course : **B.Voc Software Development**

Semester : VII

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. The paper has **two** sections. **Section A** is compulsory. Each question is of **5** marks.
3. Attempt any **four** questions from **Section B**. Each question is of **15** marks.

Section A

1. (a) Given the following data of students' test scores: (5)

```
import numpy as np
scores = np.array([45, 56, 67, 89, 90, 91, 92, 94, 95, 96, 99, 100, 45, 47, 48])
```

Write Python commands to:

i) Compute Q1, Q3, and IQR.
ii) Identify the lower and upper bounds for detecting outliers using the $1.5 \times$ IQR rule.
iii) Display the outlier values (if any).

(b) Consider the following data of the number of students enrolled in different B.Voc courses: (5)

```
courses = ['Web Design', 'Data Analytics', 'Banking', 'Retail Management',
'Software Development']
students = [120, 90, 80, 100, 110]
```

Write Python commands to:

(i) Create a horizontal bar chart for student enrollment.
(ii) Annotate each bar with its respective value.
(iii) Change the figure size to (8, 5).
(iv) Assign the title "Student Enrollment in B.Voc Courses".

P.T.O.

(c) What is Sampling? Discuss the three types of Sampling (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': ['A', 'B', 'C', 'D', 'E', 'F'],
  'Salary': [45000, 55000, 60000, 65000, 50000, 52000]
})
```

Write Python commands to:

i) Group data by Department and calculate the average salary of each department.

ii) Display the maximum salary from each department.

iii) Rename the grouped column as 'Avg_Salary'.

Section B

2. (a) Write Python commands to: (7)

i) Import a CSV file named **employee.csv** into a Pandas DataFrame called **employee_df**.

ii) Display the **first five records** of the dataframe.

iii) Compute and display the **median of each numeric column** in the DataFrame. (8)

(b) Consider the following code:

```
df = pd.DataFrame({
  'Category': ['A', 'A', 'B', 'B', 'C', 'C'],
  'Value': [10, 20, 30, 25, 40, 35],
  'Discount': [2, 3, 5, 4, 6, 5]
})
```

Write Python commands to:

i) Group the data by **Category** and calculate the **sum** and **mean** of both columns.

ii) Use the agg() function to compute **min** of Value and **max** of Discount.

iii) Display the final grouped DataFrame with appropriate column labels.

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 5 rows and 4 columns containing random numbers in the range 1 to 50.

ii) Assign custom row labels as ['L1', 'L2', 'L3', 'L4', 'L5'] and column names as ['Col_A', 'Col_B', 'Col_C', 'Col_D'].

iii) Display the DataFrame.

4. (a) What is normalization and standardization? Why are they required in data preprocessing? (7)

(b) Consider the following NumPy arrays: (8)

```
import numpy as np

a1 = np.array([[2, 5, 7], [9, 11, 13]])
a2 = np.array([[1, 4, 6], [8, 10, 12], [3, 5, 9]])
```

Give the **output** of the following commands:

i) a2[1, :]

ii) a1[:, -1]

iii) a1 * 2

iv) a2 < 8

v) a2[0, 1] = 99 — What change will occur in array a2 after executing this statement?

5. (a) Consider the following Pandas Series showing the sales (in lakhs) of different shops: (7)

```
Shop_A  85
Shop_B  60
Shop_C  95
Shop_D  70
Shop_E  100
```

Write Python commands to:

i) Display the shops having sales greater than 80.

ii) Display the index labels.

iii) Assign the name 'Shop_Name' to the index.

(b) Create a DataFrame with **Employee Name**, **Department**, and **Salary**. (8)

Perform the following operations:

i) Sort employees by **Salary** in descending order.
ii) Add a new column **"Bonus"** (Bonus = 10% of Salary).
iii) Drop the column **"Department"**.
iv) Rename the columns:
'Employee Name' → **'Employee_Name'**,
'Salary' → **'Monthly_Salary'**.

6. (a) Consider a CSV file named **inventory.csv** that contains **product data** with **columns**: (7)
Product_ID, Product_Name, Category, Price, and Quantity (60 rows).

Write Python commands to:
(i) Read the file **inventory.csv** into a DataFrame named **items**.
(ii) Display the **first 8 rows** of the dataset.
(iii) Generate the **summary statistics** (like mean, std, min, max) for all numeric columns.
(iv) Delete all rows where **all column values are missing**.
(v) Detect and display any **duplicate product entries**.

(b) Write Python code to: (8)
(i) Generate a NumPy array data containing **60 floating-point values** uniformly
distributed between **10 and 50**.
(ii) Divide the data into **5 equal-width bins**.
(iii) Assign labels to the bins as:
['Very Low', 'Low', 'Medium', 'High', 'Very High'].
(iv) Round the floating values to **2 decimal places** and display both the data and their assigned bin labels.

7. (a) Write Python code using NumPy to: (7)
(i) Create a **3D NumPy array** data of size **2 × 3 ×4** containing **random numbers** between **1 and 50**.
(ii) **Transpose** the array.
(iii) **Swap axis 1 with axis 2** and print the new shape of the array.
(iv) Display the The **data type** of array elements in the array

(b) You have a file sales.txt where values are separated by **semicolons** (;): (8)
product; price; quantity
pen;10;100
pencil;5;200
Write code to read it correctly and display the DataFrame.

(200)