(6)

[This question paper contains 16 printed pages.]

**Your Roll No.....**

**I**

**Sr. No. of Question Paper : 1290**

Unique Paper Code : 2343012011

Name of the Paper : Data Analysis and Visualization using Python

Name of the Course : **B.Sc. (Hons.) Computer Science**

Semester : III

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. Question No. **1** is compulsory.
3. Attempt any **4** questions out of **Q. 2** to **Q.7.**
4. Answer parts of a question together.
5. Assume that the following libraries have already been imported :

   import numpy as np, pandas as pd

P.T.O.

## Section A

1. (a) What is the difference between interpolation and extrapolation? Explain with the help of an example. (6)

(b) What is the output of the following code segment? (6)

```
a = np.array([25, 35, 45, 28, 60, 22, 38, 50,
              42, 29, 52, 24]).reshape(3,4)
print(a)
print(a[:, 2])
print(a[[True, False, True], :2])
print(a[[1, 0, 2],[3, 2, 1]])
print(a[a > 25].min())
print(a + 2)
```

(c) For the following data frame: (6)

```
myDF = pd.DataFrame([[16, 32, np.NaN], [2, 4, 8],
       [np.NaN, 64, np.NaN], [128, 256, np.NaN]],
       index = ['one', 'two', 'three', 'four'],
       columns = ["Col1", "Col2", "Col3"])
```

Write Python statements to do the following :

(i) print part of myDF as follows:

| | Col3 | Col2 |
|---|---|---|
| three | NaN | 64 |
| one | NaN | 32 |
| two | 8.0 | 4 |

(ii) print row with index ' three ' using loc and iloc operators.

(iii) create a new column 'Col4' having the minimum values of the corresponding rows.

(d) Write Python code to do the following : (6)

(i) Create a data frame Employee_data containing 50 rows and three columns as mentioned below:

EmployeeID: starting from 1001 to 1050

Salary: randomly generated values ranging between 20000 and 300000.

(ii) Using Salary column of Employee_data, create 4 bins with labels and their corresponding ranges as below:

| Label | Salary Range |
|---|---|
| Beginner | 20000 – 50000 |
| Mid-level | 50000 – 100000 |
| Senior | 100000 – 200000 |
| Expert | 20000 onwards |

And display the number of employees under every label.

(e) Consider the below data df (6)

```
data = {
    "Age (years)": [2, 3, 4, 5, 6, 7, 8, 9],
    "Height (cm)": [85, 95, 105, 110, 115, 125,
                    130, 140],
    "Weight (kg)": [12, 14, 18, 20, 22, 25, 28, 30],
    "Gender": ["Boy", "Girl", "Girl", "Boy", "Girl",
                    "Boy", "Boy", "Girl"]
}
df = pd.DataFrame(data)
```

Write Python statements to create a scatter plot to analyze the relationship between a child's age and height, using the marker size to represent their weight. Label the axes appropriately. Use different colors for boys and girls.

## Section B

2. (a) Consider the dataframe, df, of a store : (8)

| | CustomerID | ItemType | Amount | Group |
|---|---|---|---|---|
| 0 | C1 | Clothing | 12000 | Working |
| 1 | C2 | Clothing | 2500 | Working |
| 2 | C3 | Electronics | 1500 | Student |
| 3 | C4 | Clothing | 5000 | Student |
| 4 | C5 | Books | 1000 | Working |
| 5 | C6 | Books | 900 | Student |
| 6 | C7 | Electronics | 1000 | Working |
| 7 | C8 | Clothing | 500 | Student |

What is the output of the following code snippet :

```
group1 = df.groupby('ItemType')['Amount'].sum()
print(group1)
group2 =
    df.groupby(['Group','ItemType'])['Amount'].sum()
print(group2)
table1 = df.pivot_table(index =['ItemType','Group'],
                values = 'Amount')
print(table1)
table2 = pd.crosstab(df['ItemType'], df['Group'])
print(table2)
```

(b) Following are some of the attributes of a dataset of employees : (3)

| Attribute Name | Description |
|---|---|
| EmployeeID | A unique identification number of the employee within the organization |
| Salary | Monthly salary of the employee |
| Designation | Can be one of the following: Intern, Analyst, Team Leader, Manager, Director |

Classify the attributes to be quantitative or categorical data. If an attribute is categorical then further classify it to be ordinal or nominal. Justify your answers.

(c) What is a boxplot and how can it be used to identify outliers? (4)

3. (a) Given the following two data frames Customer and Orders : (9)

**Customer**

| customerID | Name | City |
|---|---|---|
| 101 | Anand | Mumbai |
| 102 | Vishal | Chandigarh |
| 103 | John | Lucknow |
| 104 | Anita | Hyderabad |

**Orders**

| orderID | customerID | Product | Amount |
|---|---|---|---|
| 1 | 101 | Shirt | 600 |
| 2 | 102 | Pants | 800 |
| 3 | 101 | Kurta | 650 |
| 4 | 105 | Shoes | 1000 |

(i) What is the output of the following code statement?

```
print(pd.merge(customers, orders, how = 'outer'))
```

(ii) Write statements in Python for the following :

1) Find the customerID, name and products purchased for the customers who have purchased at least one product.

2) Display the details of the customers who have not purchased anything.

(b) Find the output that will be produced on the execution of the following code snippet: (6)

```
df = pd.DataFrame(np.arange(12).reshape((4, 3)),
     index =[['MP', 'PB', 'MP', 'PB' ],['Wheat',
                'Wheat', 'Rice', 'Rice']],
     columns = ['C1', 'C2', 'C3'])
print(df)
print(df.sort_index(level = 0))
```

4. (a) Consider the below data (8)

```
data = pd.DataFrame({'A': [1, 2, np.nan],
                     'B': [4, np.nan, np.nan],
                     'C': [7, 8, 9]})
```

Write Python code for the following :

(i) Count the total number of missing values in the entire data frame.

(ii) Drop rows where more than 50% of the values are missing.

(iii) Replace missing values of column 'A', 'B', 'C' with 1, 2, 3 respectively.

(iv) Replace every value of data by its square using apply function.

(b) Consider the following data frame df containing the heights of individuals (7)

```
df = pd.DataFrame({
    'Gender': ['Male', 'Female', 'Male', 'Female',
               'Male', 'Female', 'Male', 'Female',
               'Male', 'Female'],
    'Height': [175, 160, 180, 155, 170, 165, 185,
                    150, 178, 162]})
```

Write Python code to plot histograms of heights of males and females separately. Add appropriate title, x-axis and y-axis labels to the graph. Save graph to 'heights.jpeg'.

5. (a) Consider a CSV file named student_data.csv which has the marks of the student in five subjects (P, C, M, B and E) as shown below : (2×5=10)

```
studentID;Name;P;C;M;B;E
101;Rohan;93;98;90;96;92
102;Mike;83;78;82;90;86
103;Sagar;73;70;54;78;79
104;Lalit;56;65;72;74;60
105;Sonal;84;83;81;87;95
```

Write Python code to do the following :

(i) Use Pandas to read the student_data.csv file with studentID column as the index of the data frame.

(ii) Add another column 'Rank in the Class' which has the rank of the student as per the total marks (of all the five subjects) obtained by him/her.

(iii) Change the names of the columns as mentioned below :

P to Physics

C to Chemistry

M to Mathematics

B to Biology

E to English

(iv) Display the details of the student who scored highest marks in English.

(v) Draw a stacked bar graph of the marks obtained by a student in 5 subjects, with studentID on the x-axis.

(b) What will be the output of the following code segment? (5)

```
a = np.array([[[54, 46], [92, 38]], [[98, 91],
                        [29, 83]]])
print(a.shape)
b = a.swapaxes(2, 1)
print(b)
print(b.shape)
```

6. Consider the following dataframe expenseDF where each row represents a customer transaction, including customer age, transaction amount, and region.

```
expenseDF = pd.DataFrame({
    'CustomerID': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    'Age': [25, 35, 45, 28, 60, 22, 38, 50, 42, 29],
    'Transaction_Amount': [200, 450, 350, 300, 500,
              250, 600, 400, 550, 300],
    'Region': ['North', 'South', 'East', 'West',
               'North', 'South', 'East', 'West',
               'North', 'South']
```

Write Python statements for the following :

(a) How will you determine whether there is any relationship between age and the expenditure done by a person? Write the Python statement to determine the same for expenseDF. How will you interpret the result obtained after executing the Python statement? (3)

(b) Determine the region which has the maximum number of spenders. If there is a tie for the same, then it should display all those regions. (3)

(c) Determine region-wise total expenditure done by customers along with the number of customers in the region. (3)

(d) Create a new column 'Age_Group' and using binning assign a label out of 'Under 30', '30-40', '40-50', and '50+' depending upon age of the person. (3)

(e) Create a pivot table to calculate the average transaction amount for each of the Region under each Age Group. (3)

7. (a) What will the output of the following code segment? (5)

```
a = np.array([[52, 28, 91], [37, 72, 18],
              [65, 42, 87], [6, 21, 95]])
print(a)
b = a[1:3, :2]
print(b)
s = str(b[0, 1])
b[0, 1] = int(s[::-1])
print(b)
print(a)
```

(b) Write Numpy/Pandas statements to : (10)

(i) create a 3-dimension ndarray arr1 of size 4 x 2 x 3 filled with random integers between 1 and 100.

(ii) create an arr2 from the sum of the elements of arr1 [0] and arr1[2]. What will be the shape of arr2?

(iii) create a data frame df1 for the following table with Name, Gender and Salary as column names :

| Name | Gender | Salary |
|---|---|---|
| Arnav | Male | 15800 |
| Ruhi | Female | 16000 |
| Sandesh | Male | 18000 |
| Sahil | Male | 16500 |
| Shuchi | Female | 16200 |

(iv) create a series S1 from df1 with Name as index and Salary as values.

(v) create a data frame employees from the following lists such that department is the outer index along the column and month is the next level index:

```
employee = ["Sam", "Tom", "Sam", "Tom", "Anna",
                "Anna"]
department = ['Sales', 'Marketing', 'Sales',
                 'Marketing', 'HR', 'HR']
sales = [5000, 7000, 6000, 8000, 12000, 11000]
month = ["Jan", "Jan", "Feb", "Feb", "Jan",
            "Feb"]
```

9/10/25 (M)