Gender	Role	Experience	Salary
Male	Data Analyst	1	48000
Male	Data Analyst	1	42000
Male	Data Analyst	3	51000
Male	Data Scientist	5	62000
Female	Data Scientist	6	71000
Female	Data Scientist	8	73000
Male	Manager	10	82000
Female	Manager	11	87000
Female	Manager	12	91000

Write Python statement(s) to do the following. (Make use of appropriate libraries.) :

(a) Read data from the given CSV file 'employee.csv' into a dataframe empData.

(b) Calculate and display the total salary for each role.

(c) Display the total number of females along with their average salary.

(d) Compare the highest and lowest salary for each gender using bar plot.

(e) Delete records with salary less than the average salary of all employees. (15)

[This question paper contains 12 printed pages.]

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

Sr. No. of Question Paper : 1673 **G**

Unique Paper Code : 2343012002

Name of the Paper : Data Analysis and Visualization

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. **Section A** is compulsory.
3. Attempt any **four** questions from **Section B.**
4. Parts of a question must be answered together.

Section A

Assume that the following libraries have already been imported :

```
import numpy as np
import pandas as pd
```

1. (a) Given rainfall = [5, 2, 7, 8, 2] captured for 5 days of a month, days = [1, 3, 5, 7, 9]. Write code in Python to plot a line with days and rainfall as x and y axis respectively. Mark each point with a red circle of size 20. Add a title to the graph. (Make use of appropriate libraries.) (5)

(b) Consider the following dataframe, company, having details of employees of an organization : (5)

```
       Name  Age
0  Sangeeta   18
1    Sarika   30
2  Sangeeta   45
3    Babita   50
4    Sarika   32
```

Using appropriate libraries, write Python statements to do the following and also show the output :

(i) Display the total number of distinct names of the employees.

(ii) Compute the average age of the employees with the same name.

(c) Consider the following tables named section1 and section2, each having details viz. RollNo and Name of students in each class : (5)

(iii)
```
df1 = pd.DataFrame({'A': [21, 32],
                    'B': [27, 30]})
df2 = pd.DataFrame({'A': [23, 41]})
print(df1)
print(df2)
df2['A'][1] = df1['A'][1] + 10
print(df2)
print(df2 > df1['B'].min())
```
(5)

(b) Consider an array, ages, consisting of age of 12 people [20, 22, 25, 27, 21, 23, 37, 31, 61, 45, 41, 32].

Using appropriate libraries, write code to :

(i) Create four bins of the array ages, using right side closed intervals (18–25], (25–35], (35-60], (60-100]. Name the categories as 'Youth', 'YoungAdult', 'MiddleAged' and 'Senior' respectively. Display the number of values in each category.

(ii) Create four equal-sized categories of the array ages. (5)

7. Given the following CSV file 'employee.csv' consisting of details of employees :

(ii) Create a figure and add two subplots in it. In the first subplot, create a scatter plot between Salary and Age. Give labels to the x-axis as Salary and the y-axis as Age. Also, give a title to this plot. Discretize Salary into 3 equal bins. In the second subplot, draw a figure to visualize the count of the number of employees in each of these bins.

(iii) Save the plotted figure to a file named 'Employees.png'.

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

(i) (2)

```
s1 = pd.Series([5, 0, -4, 8])
print(s1)
print(s1.rank())
```

(ii) (3)

```
data1 = pd.DataFrame({
'One': ['a', 'b'] * 2 + ['b'],
'Two': [21, 22, 21, 23, 24]})
print(data1)
data2 = data1.drop_duplicates(['One', 'Two'],
keep='last')
print(data1)
print(data2)
```

RollNo	Name
1	Abhav
2	Vihaan
3	Chitra
4	Devansh

section1

RollNo	Name
1	Roni
5	Kabeer
3	Ishani
2	Vihaan

section2

Write Python statements to do the following :

(i) Create a dataframe named section1 for the table section 1.

(ii) Display details of all students of section 2 along with details of students of section 1 with the same Name.

(iii) Display details of students with the same Name and RollNo in both sections.

(d) Find the output that will be produced on the execution of the following code snippet: (5)

```
a1 = np.zeros((2, 3))
print(a1)
a2 = [[3, 4, 5],[7, 8, 9]]
print(np.add(a1, a2))
a1 = np.append(a1, a2, axis = 0)
print(a1)
print('Shape of array:', a1.shape)
```

(e) Consider a NumPy array, empSalary, containing salary of 10 employees. Write Python statements to do the following : (5)

(i) Find total number of employees earning salary > 5000.

(ii) Create a new array, incentive, to store incentives given to each employee where incentive is 10% of the salary.

(f) Find the output that will be produced on the execution of the following code snippet : (5)

```
data = pd.DataFrame([[2, 4, 6],[np.NaN, 8, 10],
     [np.NaN, 12, np.NaN], [np.NaN, np.NaN, np.NaN]])
print(data)
print(data.dropna(thresh = 2))
print(data.fillna(method = 'ffill', limit = 2))
```

Section B

Assumé that the following libraries have already been imported

import numpy as np

import pandas as pd

2. (a) Consider the following dataframe, df : (6)

```
c1 = np.arange(0, 24, 2)
c2 = c1.reshape((2, 6))
print(c1, c2, sep = '\n')
print(c2.reshape((3,4)))
arr2[:3, 3:] = 0
print(c2)
print(c1 * 2)
```

(b) Assume that the data given below is saved in an excel file 'data.xlsx' (with 4 columns Employee_id, Department, Salary and Age) : (10)

Employee_id	**Department**	**Salary**	**Age**
101	Computer Science	2000	23
102	Computer Science	2002	34
103	Computer Science	2040	39
104	English	2045	43
105	English	2030	34
106	English	2006	53

Write Python statements to do the following (Make use of appropriate libraries.) :

(i) Read data from the given excel file 'data.xlsx' into a dataframe, df1. Set Employee_id as the index of the df1.

```
  person  sales  quarter country
0      A   1000        1      US
1      B    300        1   Japan
2      C    400        1  Brazil
3      D    500        1      UK
4      E    800        1      US
5      A   1000        2  Brazil
6      B    500        2   Japan
7      C    700        2  Brazil
8      D     50        2      US
```

Write Python statements to do the following :

(i) Find the maximum and minimum sales for Brazil.

(ii) Display total sales for each country.

(iii) Display the name of the salesperson with maximum average sales.

(iv) Display statistical summary of the numerical attributes only.

(v) Draw a boxplot of the sales.

5. (a) Find the output that will be produced on the execution of the following code snippet: (5)

```
df = pd.DataFrame(np.arange(12).reshape(4, 3),
     index = [['North', ' North', 'South', 'South'],
              [1, 2, 1, 2]],
     columns = [['Delhi', 'Delhi', 'Chandigarh'],
                ['Green', 'Red', 'Green']])
```

Find the output that will be produced on the execution of the following code snippet :

(i) print(df)

(ii) df.index.names = [‘key1’, ‘key2’]
print(df)
df1=df.swaplevel(‘key1’, ‘key2’)
print(df1)

(iii) df 2=df1.sort_index (level=0)
print(df2)

(b) Construct a NumPy array, markSheet, to store marks obtained by 2 students in 3 subjects, where marks are between 60 and 100. Write Python statements to display the data type, shape and dimension of markSheet. (5)

(c) Consider the following dataframe, itemRate : (4)

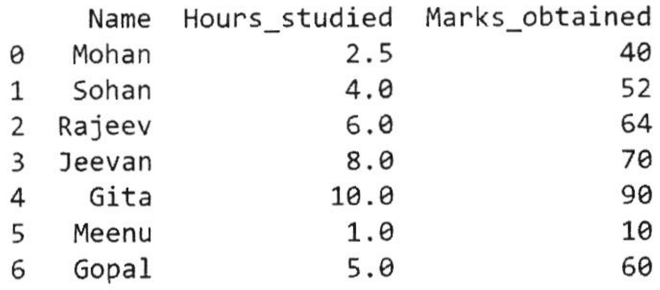
```
      Item  Rate
0   Apples   220
1  Oranges    90
```

Write Python statements to do the following :

(i) Double the value of the column Rate of each item.

(ii) Display the type of item with minimum rate.

3. (a) Consider the following dataframe, df Student, consisting of student details : (8)



```
     Name  Hours_studied  Marks_obtained
0   Mohan            2.5              40
1   Sohan            4.0              52
2  Rajeev            6.0              64
3  Jeevan            8.0              70
4    Gita           10.0              90
5   Meenu            1.0              10
6   Gopal            5.0              60
```

Write Python code to answer the following. (Make use of appropriate libraries.) :

(i) Find names of students who got maximum marks.

(ii) Find the average number of hours studied by the students.

(iii) Compute the correlation and covariance between Hours_studied and Marks_obtained.

(iv) Plot the heatmap of columns Hours_studied and Marks_obtained of the dataframe Student.

(b) Find the output on the execution of the following code snippet : (7)

```
b1 = np.arange(6)
b2 = np.array([[1, 2, 3],[4, 6, 8]])
print('i.\n', b1)
print('ii.\n', b2)
print('iii.\n', 2/b2)
print('iv.\n', b1[1], b2[1])
print('v.\n', b1[:1], b2[::2])
```

4. (a) Using diagrams give example of each of the following data distributions : (6)

(i) unimodal

(ii) bimodal

(iii) multimodal

(b) Consider the following dataframe, company, showing details of sales done by salespersons in two quarters : (9)

P.T.O.