(ii) a.rank(method = 'first')

(iii) a.rank(ascending = False)

[This question paper contains 16 printed pages.]

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

Sr. No. of Question Paper : 2012   **F**

Unique Paper Code   : 2344001201

Name of the Paper   : Data Analysis and Visualization Using Python

Name of the Course   : **Computer Science: 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. This question paper has **two** sections **A** and **B**.

3. Question **1** in **Section A** is compulsory.

4. Attempt any **4** questions from **Section B**.

5. Parts of a question must be attempted together.

6. **Section A** carries **30** marks and each question in **Section B** carries **15** marks.

7. Use of Calculator is not allowed.

     P.T.O.
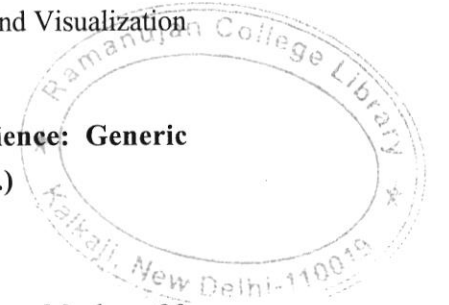
Section A

Assume numpy has been imported as np and pandas has been imported as pd.

1.   (a) Explain unimodal, bimodal and multimodal distribution with the help of examples.   (5)

     (b) Consider the DataFrames First and Second given below :   (5)

One	Two
0	'A'
2	'B'
5	'D'
6	'C'

First

One	Two
0	'B'
1	'C'
5	'E'
2	'A'

Second

Consider the following python code segment :

```
right = pd.merge(first, second, how='right', on='One')
left = pd.merge(first, second, how='inner', on='Two')
```

Show the content of the new DataFrames right and left.

     (b) Give the output of the following code segment :   (4)

```
arr = np.array([89, 54, 76, 32, 47, 21, 92, 39, 82])

arr1 = arr[5:9]

arr2 = arr[5:9].copy()

arr1 = 36

arr2 = 7

print(arr)

print(arr1)

print(arr2)
```

     (c) Consider the series a given below and give the output of the following commands :   (3)

```
a = pd.Series([4, 1, 7, 1, 8, 9, 0, 8, 2, 3, 9])
```

      (i) a.rank()

7. (a) Consider the DataFrame df given below : (8)

EmployeeID	Department	Salary	Age
1001	English	1000	23
1002	English	1002	34
1003	English	1004	39
1004	English	1005	43
1003	Maths	1004	34
1004	Maths	1005	43
1001	Maths	1006	53
1002	Maths	1002	43

Write the python code to perform the following operations :

   (i) Create a hierarchical index on Department and Employee ID.

   (ii) Give the summary level statistics for each column.

   (iii) Give the output for the following :

     1. df.stack()

     2. df.unstack()

(c) Write python commands to create a figure object using matplotlib. The Figure object has one subplot that contains 3 line graphs. Define legend and chart title of the graph. Define a different style and colour for each line in the subplot. Import appropriate libraries. (5)

(d) List and describe the steps involved in process of Data Analysis. (5)

(e) Give the output of the following code snippets:

(4)

   (i) y=np. arange(12).reshape(4,3)

     print(y)

     y[(y > 5)] = −1

     print(y)

   (ii) x = np.array ([[2, 4], [5,1]])

     z=np.ones_like(x)

     print(z)

     w=np.eye(2) * x

     print(w)

(f) Consider the series S1 and S2 given below : (6)

S1		S2	
A	1	A	5
B	2	B	6
C	3	D	7
D	4	E	8

Give the output of the following python pandas commands :

(i) S1 [: 3] * 10

(ii) S1 + S2

(iii) S2 [: : −1] * 5

Section B

2. (a) Consider the DataFrame Frame given below : (7)

Name	Age	Weight	Height
Ram	15	45.6	140
Ravi	23	34.9	160
Reena	32	45.6	145
Rita	20	60.7	155
Rishi	33	54.7	170
Romi	21	34.6	144

(ii) student [student ['Age'] >20]

(iii) student [student ['Age'] >20] ['Name']

(iv) avg_marks = np.mean (student.Marks)

student[student['Marks']>avg_marks]

(v) first = student [student ['Year'] ==1]['Marks']

np.mean(first)

(b) Consider the following list 11. (5)

11 = [10, 10, 20, 40, 50, 60, 70, 80, 90, 90]

Discretise the 11 into 4 bins using cut() and qcut(). Give the names ['first', 'second', 'third', 'fourth'] to the bins. What type of object is returned by the pandas after binning? What output is generated by attributes codes and categories of binning object?

(iv) Delete the column Longitude from data.

(v) Save data as a csv with separator as ';'.

(b) Write a python code to create a figure and a subplot using matplotlib functions. Plot a rectangle of size 3.5 x 8.5 at point (2.0, 7.0), a circle of radius 2.5 at point (7.0, 2.0) as patches in the subplot, functions for plotting. Set the colour of rectangle as 'Green' and color of circle as 'Blue'. Set the x-scale and y-scale to 1-10. Import appropriate libraries.   (5)

6.   (a) Consider the following dataset student.   (10)

Year	Name	Roll No	Marks	Age
1	Rani	23	70	18
2	Rita	24	75	20
3	Raj	25	80	22
1	Rahul	26	65	25
2	Rohit	27	80	28

Give the output of the following python commands :

(i) student [['Roll No ',' Name ']] [2 : 4]

Write python commands to perform the following operations :

(i) Compute the correlation of Age with both Weight and Height.

(ii) Sort Frame in descending order of Age.

(iii) To find the index for the row with minimum Age.

(iv) Calculate cumulative sum for Weight for all Students.

(v) To set height of 'Rita' and 'Romi' to NA.

(vi) Replace the value 32 with 18 and 33 with 19 in Age column.

(vii) Define map function to convert values of Name column to upper case.

P.T.O.

(b) Refer to the DataFrame Frame given in question 2 (a), Write a python program to perform the following operations in the given dataset with columns Name, Age, Weight, Height.   (8)

(i) Create a figure and include 2 subplots in it.

(ii) In the first subplot create a scatter plot between two variables Age and Height.

(iii) In the second subplot draw a horizontal bar plot between Name and Weight.

(iv) Set the title for the figure as 'Data Analysis'.

(v) Give appropriate labels for x and y axis.

(vi) Save the figure to file with name 'analysis.png'.

(vi) Multiply each element in mat with 25.

5. (a) Give the python commands to create a dictionary with 5 keys – 'A', 'B', 'C', 'D', 'E' and value as follows.   (10)

Key	Value
A	List of numbers from 1 to 10 skipping 2 at a time.
B	List of Strings from A to E.
C	List of 5 numbers obtained using random normal distribution function.
D	List of 5 random integers from 20 to 30.
E	Square root of 5 random numbers from 50 to 70.

Give python commands to perform the following operations :

(i) Create DataFrame data using the above dictionary.

(ii) Convert Column A to index.

(iii) Rename the rest of the columns as Area, Temperature, Latitude and Longitude.

(iv) Replace all null values by the last known valid observation.

(b) What are outliers? How can you detect outliers using boxplots?   (5)

(c) Consider the given numpy array mat :   (6)

mat = np.array([[[−1,2], [3,4]], [[−5,6], [7,8]]])

Write numpy commands to perform the following operations :

(i) Create an array of zeros with the same shape as mat.

(ii) Print the shape of the mat.

(iii) Print the datatype of the elements in mat.

(iv) Print the elements which are greater than 6 in mat.

(v) Convert all the elements in mat as float type.

3. (a) Consider the following numpy array matrix :

(10)

$$[[5,10,20],$$

$$[20,13,43],$$

$$[34,27,67],$$

$$[12,46,77]]$$

Give the output of the following numpy commands :

(i) matrix.T

(ii) matrix[:1,1:]

(iii) matrix[[1,3,0],[2,1,0]]

(iv) matrix[[−2,−4]]

(v) matrix[[True, False, False, True]]

(vi) matrix[3] [:2]

(vii) matrix[::−1]

(viii) matrix.ndim

(ix) np.swapaxes(matrix, 1, 0)

(x) matrix+10

(b) Consider the following DataFrame df.   (5)

Items	Sugar Type	Price
Yogurt	Low Fat	45
Chips	Regular	30
Soda	Low Fat	50
Yogurt	High Fat	70
Cake	Regular	140
Chips	Low Fat	40
Yogurt	Regular	50

Give commands to perform the following operations:

(i) List the name of unique items sold.

(ii) Count the number of times each value in items is stored.

(iii) Delete the rows which have duplicate values of Items.

(iv) Give the average price of all Low Fat items.

(v) Check if 'Juice' ims one of the items sold.

4. (a) Consider the DataFrame data given below.   (4)

One	Two	Three	Four	Five
1	14	34	NaN	NaN
34	21	NaN	12	NaN
NaN	23	NaN	2	NaN
34	21	32	33	NaN

Write python commands to perform the following operations :

(i) Drop columns with any null values.

(ii) Replace the null values with the mean of each column.

(iii) Drop the null values where there are at least 2 null values in a row.