INTRODUCTION TO DATA ANALYSIS WITH PYTHON AND R
(CSEN 3142)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Which of the following is a valid variable name in Python?
(a) 123John (b) John123 (c) import (d) None of the above.

(ii) The four V's associated with Big Data are:
(a) Volume, Variety, Velocity, Veracity (b) Volume, Vanity, Velocity, Vagility
(c) Volume, Variety, Vanity, Veracity (d) Volume, Velocity, Vagility, Veracity

(iii) Once data is ready for analysis you do exploratory data analysis (EDA). Which of the activity(s) is not part of this step?
(a) Inspect the data and all its properties
(b) Compute descriptive statistics to extract features and test significant variables
(c) Data visualization to identify patterns and trends
(d) Data scrubbing.

(iv) If a is given list, print([x*x for x in a if x > 2 and x < 5]) is an example of
(a) Reduction (b) List comprehension (c) Concatenation (d) Inheritance

(v) Python script will execute faster when file with .pyc suffix has newer than the file with the .py suffix. What is the reason behind it?
(a) Python interpreter will start executing the script file
(b) Python will load the byte code directly and skip the compilation
(c) None of the above options
(d) Observation is wrong; executing speed does not depend on file with .pyc suffix.

(vi) What will be output of the following code snippet?

```
questions = 30
correct_answers = 23
print(f"You got {correct_answers / questions :.2%} correct!")
```

(a) You got 76.67 correct (b) Syntax Error : FormatError
(c) You got 76.67% correct (d) You got 77% correct

(vii) What is the shape of y in the following code?

```
import numpy as np
y = np.arange(3)
```

(a) (1, 3) (b) (3, 1) (c) (3,) (d) (,3)

(viii) What will be displayed when the following piece of code is executed?

```
import numpy as np
import pandas as pd
df = pd.DataFrame([[5.1], [np.nan]], index = ['Row1','Row2'],
columns = ['Col1'])
print(df.sum())
```

(a) Col1 5.1 (b) Col1 NaN (c) Row1 5.1 (d) Row1 NaN

(ix) Rprof() function is a built-in tool that enables which of the following?
(a) Write professional level R code
(b) Offer on-demand help to programmers
(c) Determine where a program spends most of its execution time
(d) None of the above.

(x) Which of the following is invalid assignment?
(a) >c (2, 5, 7,9)-> y (b) y<- c(3, 5, 7, 3)
(c) assign("y", c(4, 6, 7, 9)) (d) None of the other choices.

Fill in the blanks with the correct word

(xi) When a child class modifies or replaces the behaviour inherited from the parent class, this is called _____.

(xii) In R a two-dimensional version of a list that is very useful for data analysis is called a _____.

(xiii) In R, missing values are denoted by _____.

(xiv) In Python 3, range(n) produces the sequence _____.

(xv) In the context of file management, buffer is a _____.

Group - B

2. (a) What will be displayed when the code below is executed, and why?
```
nested_list = [[10,15], 24, [True], [5, ['Tom', 'Harry'], 10]]
print(nested_list[3][1][1]).
```
[(CO2)(Understand/LOCQ)]

(b) What will be displayed when the code below is executed, and why?
```
d = {4: 'India', 9: 'USA', 17: 'Japan', 24: 'Australia'}
print([d[i] for i in d.keys() if i%2 == 0]).
```
[(CO2)(Analyse/IOCQ)]

(c) What will be displayed when the code below is executed, and why?
```
import re
str = "examination"
pattern = re.compile("a\w\w\w")
print(pattern.findall(str))
print(pattern.sub("yyy", str))
print(pattern.split(str))
```
[(CO1)(Analyse/LOCQ)]

4 + 4 + 4 = 12

3. (a) Abecedarian series refers to a sequence or list in which the elements appear in alphabetical order. Write a recursive function to check whether sequence sent as a parameter of that function is Abecedarian or not. What is the output you expect for 'accddmop' and 'addccnm'. *[(CO1)(Apply/IOCQ)]*

(b) Say you have defined following function:
```
def  empInfo(name, /, age, *,dept):
        print(name, age, dept)
```
What output you expect from the following calls :
```
empInfo('Pranab', age=40, dept='HRD')
empInfo(('Pranab',40, 'HRD')
empInfo(name='Pranab', age=40, dept='HRD').
```
[(CO1)(Analyse/IOCQ)]

(c) Write a function to swap the first and last element of a list using * operand and slicing. *[(CO1)(Understand/LOCQ)]*

(d) Write a function to perform binary search and write a code snippet to accept numbers from user to form the search list and call that function to check whether a number is in list or not. *[(CO1)(Understand/LOCQ)]*

2 + 3 + 3 + 4 = 12

Group - C

4. (a) Observe the following code:
```
def fun(n):
    if n == 0:
        return "0"
    elif n == 1:
        return "1"
    else:
        return fun(n//2) + fun(n%2)
```
What will be the output of print(fun(n)) if n equals the sum of the last three digits of your class roll number, and why? (For example, if a student's class roll number is 2151168, for that student, n = 1+6+8 = 15.)
What is the type of the value returned by the function in the above question for your specific value of n, and why?
[(CO2)(Understand/IOCQ)]

(b) The following function returns the number of pieces of a circular pizza obtained by running a knife n times in straight lines across the pizza. The lines may or may not intersect one another. All lines may not necessarily pass through the centre of the pizza.
```
def pizza_pieces(n):
    if n==0:
        return 1
    else:
        return pizza_pieces (n-1) + n
```
How many pizza pieces will be obtained by running the knife nine(9) times and why? How can the given code be minimally changed so that the function calculates the factorial of n, and why? *[(CO2)(Apply/IOCQ)]*

(4 + 2) + (4 + 2) = 12

5. (a) If d is a Python dictionary, what do the following functions do?
(i) d.values()
(ii) d.items() *[(CO2)(Understand/LOCQ)]*

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

```
def some_function(p):
    dp ={}
    for order in p:
        if not order==0:
            dp[order-1] = order*p[order]
    return dp
print(some_function({0:-3,3:2,5:-1}))
```

[(CO2)(Understand/LOCQ)]

(c) What is a tuple? Explain with examples. *[(CO2)(Apply/IOCQ)]*

(d) How do you define a one element tuple? *[(CO2)(Remember/LOCQ)]*

(e) What are regular expressions and how are they useful? Explain with examples. *[(CO1)(Understand/LOCQ)]*

3 + 4 + 2 + 1 + 2 = 12

Group - D

6. (a) (i) Using Pandas data structure, develop a code to create a data frame from a dictionary of population (in crores) in the year 2021 and 2020 of three cities in India, as given below:

City	2021	2020
Delhi	3.118	3.029
Kolkata	1.497	1.485
Mumbai	1.497	2.041

(ii) Then add another column showing the population in the year 2019 as 2.94, 1.476, 2.019 crores respectively.

(iii) Find and display the average population of each city in the 3 years and show it in a new column 'Average'.

[(CO4,C06)(Appy/IOCQ)]

(b) What will be displayed when the following code is executed:

```
import pandas as pd
marks = {'Ajay': 96, 'Usha': 90, 'Deep': 94, 'Ria': 89}
obj1 = pd.Series(marks)
print(obj1)
student = ['Usha', 'Vijay', 'Ria', 'Ajay']
obj2 = pd.Series(marks, index = student)
print(obj2)
print(pd.isnull(obj2))
```

[(CO4)(Analyse/IOCQ)]

(3 + 2 + 2) + (2 + 2 + 1) = 12

7. (i) Using Pandas data structure, create a data frame from a dictionary of marks in Physics, Chemistry and Mathematics of four students, as given below:

Name	Physics	Chemistry	Mathematics
Abhijit	85	90	92
Lata	80	92	99
Asha	89	80	82
Preetam	88	93	95

(ii) Add another column showing the marks in Biology as 88, 82, 80, 79 respectively.

(iii) Find the Total marks of each student and show it in a new column 'Total'.

(iv) Display in tabular form, the descriptive statistics of all the four subjects and the total. *[(CO4, C06)(Understand/LOCQ)]*

(4 × 3) = 12

Group - E

8. (a) Is R an interpreted language or a compiled language? What are the specific advantages and disadvantages of the way R code is executed? *[(CO5)(Apply/IOCQ)]*

(b) Which of the following variable names are not valid in R, and why?

(i) _my_variable
(ii) my_variable
(iii) my_variale?
(iv) 5my_variable. *[(CO5)(Understand/LOCQ)]*

(c) What are the outputs of the following code snippets?

(i)
```
fruits <- c("apple","orange","guava")
print(class(fruits))
```

(ii)
```
A <- matrix(c(5:16), nrow = 4,ncol=3)
B <- matrix(c(1:12), nrow = 4,ncol=3)
sum <- A+B
print(sum)
```

[(CO5)(Analyse/IOCQ)]

(1 + 3) + (1 + 1 + 1 + 1) + (2 + 2) = 12

9. (a) Write a R program to create a data frame "Employees" using the data given below and perform the following actions on the data frame:

(i) Extract and display only the weight of the Employees data frame.

(ii) Convert the Gender of the Employees into factors and convert them into numeric values.
(iii) Obtain unique values of the column Age.
(iv) Obtain the sorted unique values of the column age.
(v) Delete the Height column.

Age	Height	Weight	Gender
23	76	50	Female
21	62	52	Female
34	63	80	Male
44	69	65	Male
32	72	70	Female

[(CO5, C06)(Create/HOCQ)]

(b) Write a R program to create a sequence of numbers from 20 to 50 and find the mean of numbers from 20 to 60 and sum of numbers from 51 to 91.

[(CO5)(Apply/IOCQ)]

(2 + 2 + 2 + 2 + 2) + 2 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	42.71	46.87	10.42