INTRODUCTION TO DATA ANALYSIS WITH PYTHON AND R
(CSEN 3135)

Time Allotted : 3 hrs   **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and
<u>any 5 (five)</u> from Group B to E, taking <u>at least one</u> from each group.

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

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

   (i) What will be displayed when the following piece of code is executed in Pandas?
   ```
   import functools
   def cumul_sub(x, y):
        return (x – y)
   lst = [6, 5, 4, 3, 2, 1]
   print(functools.reduce(cumul_sub, lst))
   ```
   (a) 3   (b) -21   (c) -9   (d) 9.

   (ii) Which of the following can be considered as object attribute as far as R programming language is concerned?
   (a) dimensions   (b) class   (c) length   (d) all of these.

   (iii) When an object is called in the command line, R will first look for that object in the -
   (a) current active environment
   (b) parent of the current active environment
   (c) child of the current active environment
   (d) global environment.

   (iv) Which one of the following statements is false?
   (a) Set does not maintain the elements in a particular order
   (b) Values of only mutable types can be the elements of a Python set
   (c) Duplicates are automatically removed in a set
   (d) s= {} cannot be used to create an empty set.

   (v) What do we use to define a block of code in Python language?
   (a) Key   (b) Brackets   (c) Indentation   (d) None of these

   (vi) Consider the following code snippet written in Python:
   ```
   import re
   txt = "The rain in Spain"
   ```

```
x = re.split("\s", txt, 1)
print(x)
```
What would be the output of the above code?
(a) ['The', 'rain in Spain'] (b) ['The']
(c) ['The', 'rain', 'in', 'Spain'] (d) None of these.

(vii)   What will be returned, when the statement as.logical(1) is executed in R?
(a) TRUE (b) FALSE (c) "1" (d) ASCII code of 1.

(viii) Which one of the following statements is true, when a function 'fun' is defined in Python as follows?

def fun (a, b=0):

(a) The argument b always takes the value 0
(b) If the function is called with the value of the second argument as 0, then only the function will be executed
(c) If the second argument is provided at the time of calling the function, its value is assumed to be 0
(d) If the second argument is not provided at the time of calling the function, its value is assumed to be 0.

(ix)   Which of the following statements is correct for R programming language?
(a) Empty vectors can be created with the vector () function
(b) A sequence is represented as a vector but can contain objects of different classes
(c) "raw" objects are commonly used directly in data analysis
(d) The value NaN represents undefined value.

(x)   Which of the following Python declarations is incorrect?
(a) _x = 2 (b) __x = 3 (c) __xyz__ = 5 (d) None of these.

Group – B

2. (a)   Name four categories of data types with suitable examples that Python uses. What are the six steps of the data science process?
         [(CSEN3135.1)(Remember/LOCQ)]
    (b)   Why is data cleaning required before applying any machine learning algorithm? Write down the steps of data cleaning process.
         [(CSEN3135.1)(Remember/LOCQ)]
    (c)   Why does the statement **print(int('97', 8))** give rise to error, when executed in Python? [(CSEN3135.1) (Understand/LOCQ)]

    (4 + 3) + 4 + 1 = 12

3. (a)   Write a Python program to first take an integer number (x), where $1 \leq x \leq 255$, from the user by using the prompt "Number of registers in the CPU:". Use try-except block to wait indefinitely until the given input is an integer. Check the validity of the range of the input in a user defined function 'check_range'. Now, call a function 'find_n' to find out a positive number (n) such that 2^n is the least

integer greater than or equal to x. Display the value of n with prompt "The number of bits for the Register field in the Instruction format is:"
[(CSEN3135.6)(Create, Evaluate/HOCQ)]

(b) What do you mean by Data Visualization? Why is it required?
[(CSEN3135.1)(Remember/LOC)]

(1 + 2 + 2 + 2 + 1) + (4) = 12

Group – C

4. (a) Write a Python code using Regular Expression to find out whether a string of characters input by the user is a valid floating-point number or not. If the number is valid, display the square of the number, otherwise display "Invalid input". [(CSEN3135.1, CSEN3135.2) (Apply/IOCQ)]

(b) Write a Python program, using List Comprehension method to read a text file and display the number of characters in each line. What is the problem of this technique for a very large file? Write code to solve this problem.
[(CSEN3135.1,CSEN3135.2)(Apply/IOCQ)]

(c) Write a Python code, using List comprehension method to display a list of all the triples (x, y, z) such that $0 \leq x, y, z \leq 50$ and $x^2 + y^2 = z^2$.
[(CSEN3135.1,CSEN3135.2)(Apply/IOCQ)]

4 + (2 + 1 + 3) + 2 = 12

5. (a) Write a Python program to count the number of characters in a string using dictionaries. Display the keys and their values in alphabetical order.
[(CSEN3135.1, CSEN3135.2)(Apply/IOCQ)]

(b) Write a Python code to define a class 'Point' on a 2-D plane with coordinates (x, y). Define the constructor in such a fashion that the default coordinates of the point are at the origin. Define a function to find the distance of the point from the origin. Define appropriate function such that a user can display the coordinates of the point in the format (a, b) where a is the abscissa and b is the ordinate. [(CSEN3135.1,CSEN3135.2)(Apply/IOCQ)]

6 + (2 + 2 + 2) = 12

Group – D

6. (a) Using NumPy, create a 3 × 2 array containing all elements as 1.0 and a 1-D array of 3 numeric elements [0, 1, 2]. Show step-by-step in details, how Broadcasting Rules are applied to add the above two arrays. Finally comment on whether addition is possible or not. If not, why? [(CSEN3135.3)(Apply, Analyze/IOCQ)]

(b) Mention two important reasons why NumPy arrays are faster than 'for' loops over lists in Python. [(CSEN3135.3)(Remember/LOCQ)]

(c) Given a point (x, y), write a Python program to find whether it lies in the first, second, third or fourth quadrant of x - y plane. [(CSEN3135.2) (Apply/IOCQ)]

(1 + 1 + 3 + 1) + 2 + 4 = 12

7. (a) Create a data frame in Pandas containing 4 rows and 3 columns with some missing values. How are the missing data filtered and filled in Pandas? Give two different examples in each case. [(CSEN3135.3) (Remember/LOCQ)]

   (b) Explain with an example how hierarchical indexing is done for a 'Series' data structure in Pandas. How hierarchical indexing allows rearrangement of data into a Data Frame? Give example. [(CSEN3135.3) (Understand/LOCQ)]

   (c) (i) Using Pandas data structure, create a data frame from a dictionary of population (in crores) in the year 2020 and 2019 of three cities in India, as given below:

City	2020	2019
Delhi	3.029	2.94
Kolkata	1.485	1.4755
Mumbai	2.0411	2.0185

   (ii) Write the statement to add another column showing the population in the year 2018 as 2.85, 1.468, 1.998 crores respectively

   (iii) Write the statement to find the average population of each city in the 3 years and show it in a new column 'Average'.
   [(CSEN3135.3) (Analyze/IOCQ)]

   (2 × 2) + (2 + 2) + (2 + 1 + 1) = 12

Group – E

8. (a) What are the different types of vectors in R programming language? [(CSEN3135.4)(Remember/LOCQ)]

   (b) Write a program in R to simulate the probability of getting 8 as the sum of two dice rolled. Explain the code with brief comments along with each statement. Comment on the difference between the result of the code and the theoretical probability (if any). [(CSEN3135.6)(Create, Evaluate/HOCQ)]

   (c) Distinguish between sample(1:6, size = 2) and sample(1:6, size = 2, replace = TRUE) as far as R programming language is concerned.
   [(CSEN3135.4)(Analyze/IOCQ)]

   (d) What are the features of data frames in R programming language?
   [(CSEN3135.5)(Remember/LOCQ)]

   4 + 4 + 2 + 2 = 12

9. (a) Explain with suitable example why $ notation is used in R. What is the difference between $ notation and double brackets [[]]?
   [(CSEN3135.4)(Remember/LOCQ)]

   (b) What will be displayed when the following statements are executed in R?
   (i) matrix(1:8, ncol = 4)
   (ii) class("HITK") [(CSEN3135.4)(Understand/LOCQ)]

   (c) Write a function 'prime_numbers' in R to get all prime number upto a given number (based on the sieve of Eratosthenes). Call this function to display the list of prime numbers upto 25. [(CSEN3135.4,CSEN3135.5)(Apply/IOCQ)]

   (d) Write a code in R to display
   (i) the last 10 English letters in upper case
   (ii) the names of the 5th to 9th calendar months in abbreviated form.

[(CSEN3135.4) (Remember/LOCQ)]

(2 + 2) + (1 + 1) + 4 + (1 + 1) = 12

_____

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	50%	41.66%	8.33%

Course Outcome (CO):
After the completion of the course students will be able to
CSEN3135.1. Learn and understand the basics of the Python Programming Language.
CSEN3135.2. Learn about basic Python data structures.
CSEN3135.3. Learn about the NumPy and Pandas libraries in Python.
CSEN3135.4. Learn and understand the basics of the R Programming Language.
CSEN3135.5. Learn about R data structures.
CSEN3135.6. Learn how to apply Python and R in building solutions to basic data analysis problems.

*LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question

Department & Section	Submission Link
CSE - A + B1	https://classroom.google.com/c/NDA1NjY4MzgzMjcx/a/NDYzOTcxNDEzOTYy/details
CSE - B2 + C	https://classroom.google.com/u/0/c/NDA2ODEzNTg5ODAy/a/NDQzODYwMzUyMjI5/details