A10.1-R5 : Data Science Using Python

अवधि : 03 घंटे **अधिकतम अंक : 100**
DURATION : 03 Hours **MAXIMUM MARKS : 100**

ओएमआर शीट सं. : **OMR Sheet No. :**						

रोल नं. : **Roll No. :** | | | | | | |

उत्तर-पुस्तिका सं. : **Answer Sheet No. :** | | | | | | |

परीक्षार्थी का नाम : **परीक्षार्थी के हस्ताक्षर :**
Name of Candidate : _____ ; **Signature of Candidate :** _____

परीक्षार्थियों के लिए निर्देश :	**Instructions for Candidates :**
कृपया प्रश्न-पुस्तिका, ओएमआर शीट एवं उत्तर-पुस्तिका में दिये गए निर्देशों को ध्यानपूर्वक पढ़ें।	Carefully read the instructions given on Question Paper, OMR Sheet and Answer Sheet.
प्रश्न-पुस्तिका अंग्रेजी भाषा में है। परीक्षार्थी उत्तर लिखने के लिए केवल अंग्रेजी भाषा का ही प्रयोग कर सकते हैं।	Question Paper is in English language. Candidate has to answer in English language only.
इस मॉड्यूल/पेपर के **दो भाग** हैं। **भाग एक** में **चार** प्रश्न और **भाग दो** में **पाँच** प्रश्न हैं।	There are **TWO PARTS** in this Module/Paper. **PART ONE** contains **FOUR** questions and **PART TWO** contains **FIVE** questions.
भाग एक ''वैकल्पिक'' प्रकार का है जिसके कुल अंक **40** है तथा **भाग दो** ''व्यक्तिपरक'' प्रकार का है और इसके कुल अंक **60** है।	**PART ONE** is Objective type and carries **40** Marks. **PART TWO** is Subjective type and carries **60** Marks.
भाग एक के उत्तर, इस प्रश्न-पत्र के साथ दी गई **ओएमआर उत्तर-पुस्तिका** पर, उसमें दिये गए अनुदेशों के अनुसार ही दिये जाने हैं। **भाग दो** की उत्तर-पुस्तिका में **भाग एक** के उत्तर **नहीं** दिये जाने चाहिए।	**PART ONE** is to be answered in the **OMR ANSWER SHEET** only, supplied with the question paper, as per the instructions contained therein. **PART ONE** is **NOT** to be answered in the answer book for **PART TWO**.
भाग एक के लिए अधिकतम समय सीमा **एक घण्टा** निर्धारित की गई है। **भाग दो** की उत्तर-पुस्तिका, **भाग एक** की उत्तर-पुस्तिका जमा कराने के पश्चात् दी जाएगी। तथापि, निर्धारित एक घंटे से पहले **भाग एक** पूरा करने वाले परीक्षार्थी **भाग एक** की उत्तर-पुस्तिका निरीक्षक को सौंपने के तुरंत बाद, **भाग दो** की उत्तर-पुस्तिका ले सकते हैं।	Maximum time allotted for **PART ONE** is **ONE HOUR**. Answer book for **PART TWO** will be supplied at the table when the Answer Sheet for **PART ONE** is returned. However, Candidates who complete **PART ONE** earlier than one hour, can collect the answer book for **PART TWO** immediately after handing over the Answer Sheet for **PART ONE** to the Invigilator.
परीक्षार्थी, उपस्थिति-पत्रिका पर हस्ताक्षर किए बिना और अपनी उत्तर-पुस्तिका, निरीक्षक को सौंपे बिना, परीक्षा हॉल/कमरा नहीं छोड़ सकते हैं। ऐसा नहीं करने पर, परीक्षार्थी को इस मॉड्यूल/पेपर में अयोग्य घोषित कर दिया जाएगा।	**Candidate cannot leave the examination hall/room without signing on the attendance sheet and handing over his/her Answer Sheet to the invigilator. Failing in doing so, will amount to disqualification of Candidate in this Module/ Paper.**
प्रश्न-पुस्तिका को खोलने के निर्देश मिलने के पश्चात् एवं उत्तर लिखना आरम्भ करने से पहले उम्मीदवार यह जाँच कर सुनिश्चित कर लें कि प्रश्न-पुस्तिका प्रत्येक दृष्टि से संपूर्ण है।	After receiving the instruction to open the booklet and before starting to answer the questions, the candidate should ensure that the Question Booklet is complete in all respect.

जब तक आपसे कहा न जाए, तब तक प्रश्न-पुस्तिका न खोलें।
DO NOT OPEN THE QUESTION BOOKLET UNTIL YOU ARE TOLD TO DO SO.

PART ONE

(Answer all the questions; each question carries ONE mark)

1. Each question below gives a multiple choice of answers. Choose the most appropriate one and enter in the "OMR" answer sheet supplied with the question paper, following instructions therein. (1x10)

1.1 What is the output of following code ?

```
import pandas as pd
import numpy as np
s = pd.Series(np.random.randn(6))
print s.ndim
```

(A) -1

(B) 1

(C) 0

(D) 6

1.2 What is the output of following code ?

```
import numpy as np
a = np.array([1,2,3])
print a
```

(A) [[1, 2, 3]]

(B) [1]

(C) [1, 2, 3]

(D) Error

1.3 Point out the **correct** statement.

(A) Raw data is original source of data

(B) Pre-processed data is original source of data

(C) Raw data is the data obtained after processing steps

(D) None of the mentioned

1.4 Which of the following is one of the key data science skills ?

(A) Statistics

(B) Machine Learning

(C) Data Visualization

(D) All of the mentioned

1.5 Point out the **correct** statement.

(A) The mean is a measure of central tendency of the data

(B) Empirical mean is related to "centering" the random variables

(C) The empirical standard deviation is a measure of spread

(D) All of the mentioned

1.6 Point out the **wrong** statement.

(A) Regression through the origin yields an equivalent slope if you center the data first

(B) Normalizing variables results in the slope being the correlation

(C) Least squares is not an estimation tool

(D) None of the mentioned

1.7 Which of the following is contained in NumPy library ?

(A) n-dimensional array object

(B) tools for integrating C/C++ and Fortran code

(C) Fourier transform

(D) all of the mentioned

1.8 Which of the following function stacks 1D arrays as columns into a 2D array ?

(A) row_stack

(B) column_stack

(C) com_stack

(D) all of the mentioned

1.9 Which of the following function take only single value as input ?

(A) is complex

(B) minimum

(C) fmin

(D) all of the mentioned

1.10 Which of the following is used for machine learning in python?

(A) scikit-learn

(B) seaborn-learn

(C) stats-learn

(D) none of the mentioned

2. **Each statement below is either TRUE or FALSE. Choose the most appropriate one and enter your choice in the "OMR" answer sheet supplied with the question paper, following instructions therein. (1x10)**

2.1 Statistical inference is the process of drawing formal conclusions from data.

2.2 Spyder can introspect and display Pandas DataFrames.

2.3 Statsmodels provides powerful statistics, econometrics, analysis and modeling functionality that is out of panda's scope.

2.4 Panel is generally 2D labelled size-mutable array.

2.5 rPy provides a lot of scientific routines that work on top of NumPy.

2.6 Raw data should be processed only one time.

2.7 Data visualization is the organization of information according to preset specifications.

2.8 ndarray is also known as the alias array.

2.9 In Numpy, universal functions are instances of the numpy.ufunction class.

2.10 Mutable objects can change their state or contents and immutable objects can't change their state or content.

3. Match words and phrases in column X with the closest related meaning / words(s) / phrase(s) in column Y. Enter your selection in the "OMR" answer sheet supplied with the question paper, following instructions therein. (1x10)

Column X		Column Y	
3.1	In Numpy, dimensions are called	A.	Minimize errors
3.2	Import pandas as	B.	maps
3.3	Multiple testing in statistical inference	C.	hypothesis testing
3.4	Data visualization	D.	import panda as pd
3.5	Data can be visualized using	E.	import pandas as py
3.6	This is not a part of data science process	F.	Chebyshev
3.7	Non-Statistical Analysis	G.	Regression analysis
3.8	This inequality is useful for interpreting variances	H.	Applied statistics
3.9	Answering yes/no questions about the data	I.	Axes
3.10	Modeling relationships within the data	J.	Quantitative Analysis
		K.	Qualitative Analysis
		L.	Communication Building
		M.	Visual art

4. Each statement below has a blank space to fit one of the word(s) or phrase(s) in the list below. Choose the most appropriate option, enter your choice in the "OMR" answer sheet supplied with the question paper, following instructions therein. (1x10)

A	to_sparse	B	regression	C	NaN	D	Series
E	Cor(X, Y) = 0	F	View	G	Mutable	H	Poor
I	Immutable	J	Cor(X, Y) = 2	K	Best	L	Pandas
M	ndarray						

4.1 A tuple is a sequence of _____ Python objects

4.2 In context of Python, lists are _____.

4.3 _____ method creates a new array object that looks at the same data

4.4 Numpy array class is called _____.

4.5 Pandas consist of static and moving window linear and panel _____.

4.6 All of the standard pandas data structures have a _____ method

4.7 _____ is the standard missing data marker used in pandas

4.8 A _____ is like a fixed-size dict in that you can get and set values by index label

4.9 _____ implies no relationship with respect to correlation

4.10 Residuals are useful for investigating _____ model fit.

PART -TWO
(Attempt any four question)

5. (a) Briefly discuss Matplotlib. Write command to install Matplotlib package and import Matplotlib. Write program(s) to display following plots: Line plot, Bar plot, and Scatter plot for X = [5, 2, 9, 4, 7] and Y = [10, 5, 8, 4, 2]. Also write program to display Histogram for Y = [10, 5, 8, 4, 2]. Draw these plots (Line plot, Bar plot, and Scatter plot) for the given X and Y and Histogram for given Y.

(b) In context of Python NumPy, briefly discuss sort(), argsort(), and lexsort() functions. **[10 + 5]**

6. (a) Python has a set of built-in methods that you can use on strings. List down these methods and give brief description of these methods.

(b) What will be the output displayed by the snippet of following Python program

```
import numpy as np
arr = np.array ([[1, 2, 3, 4], [5, 2, 4, 2],
      [1, 2, 0, 1]])
newarr = arr.reshape(2, 2, 3)
print ("\nOriginal array:\n", arr)
print ("Reshaped array:\n", newarr)
```

[9 + 6]

7. (a) There are number of widgets which we can put in our tkinter application. Briefly discuss the usages and syntax of following widgets: Scale, Canvas, Frame, Listbox, and RadioButton. Further, there are number of options which are used to change the format of these widgets, where number of options can be passed as parameters separated by commas. Highlight some of the available options with above listed widgets. Also write an example syntax depicting some of these options.

(b) Using suitable examples, discuss array slicing in Python. **[10 + 5]**

8. (a) Some of the statistical functions are used in following program written in Python. Presenting the output of the following program, briefly discuss the statistical functions used in this program.

```
import statistics
arr = [1, 2, 2, 3, 3, 3]
print (statistics.median(arr))
print (statistics.median_low(arr))
print (statistics.median_high(arr))
```

(b) Let us consider the data frame, "df" having the information/data presented in following table having five columns (a, b, c, d, and e) and 10 rows (0 to 9). Here, some entries in the table are missing and represented as "NaN".

	a	b	c	d	e
0	0.093748	London	True	3.0	1
1	0.835929	Paris	True	4.0	4
2	0.166490	New York	True	5.0	5
3	0.439057	Istanbul	False	1.0	3
4	0.077856	Liverpool	False	5.0	3
5	0.669849	Berlin	NaN	2.0	3
6	0.539958	NaN	NaN	2.0	3
7	0.323087	Madrid	False	NaN	8
8	0.367877	Rome	True	NaN	8
9	0.281026	NaN	True	0.0	4

With reference to given data frame (df), write the output displayed by following commands :

(i) df.iloc[1] (ii) df.iloc[0,1]
(iii) df.iloc[:,2] (iv) df.iloc[:2,1]
(v) df.loc[:2,'b'] (vi) df.loc[:2, :'b']
(vii) df.loc[2, :'b']

[8 + 7]

9. (a) Using suitable examples, explain Merge, Join, and Concatenate Data Frames using Pandas.

(b) In context of machine learning, briefly discuss supervised learning, unsupervised learning, reinforcement learning, and data partitioning. **[9 + 6]**

SPACE FOR ROUGH WORK

SPACE FOR ROUGH WORK