No. of Printed Pages : **8**

A10.1-R5.1 : DATA SCIENCE USING PYTHON

DURATION : 03 Hours **MAXIMUM MARKS : 100**

OMR Sheet No. :					

Roll No. : □□□□□□ **Answer Sheet No. :** □□□□□□

Name of Candidate : ____________________ **; Signature of Candidate :** ____________________

INSTRUCTIONS FOR CANDIDATES :

- Carefully read the instructions given on Question Paper, OMR Sheet and Answer Sheet.
- Question Paper is in English language. Candidate has to answer in English language only.
- There are **TWO PARTS** in this Module/Paper. **PART ONE** contains **FOUR** questions and **PART TWO** contains **FIVE** questions.
- **PART ONE** is Objective type and carries **40** Marks. **PART TWO** is Subjective type and carries **60** Marks.
- **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**.
- 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.
- **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.**
- After receiving the instruction to open the booklet and before answering the questions, the candidate should ensure that the Question Booklet is complete in all respects.

DO NOT OPEN THE QUESTION BOOKLET UNTIL YOU ARE TOLD TO DO SO.

PART - ONE

(Answer all 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 the instructions therein.** **(1x10)**

1.1 What is the term for a data point that falls far from the rest of the data in a dataset ?

(A) Outlier
(B) Median
(C) Mean
(D) Variance

1.2 Which step in the Data Science process involves selecting the appropriate model and algorithm for analysis ?

(A) Data Cleaning
(B) Data Visualisation
(C) Data Collection
(D) Model Building

1.3 Which of the following allows you to find the relationship you didn't know about ?

(A) Inferential
(B) Exploratory
(C) Causal
(D) None of the options

1.4 If a related package already existed in a file then what is the correct sequence of x = np.arange(10) (x > 4) & (x < 8)

(A) [False, False, False, False, False, True, True, True, False, False]
(B) [False, False, False, False, True, True, True, True, False, False]
(C) [False, False, False, False, False, True, True, False, False, False]
(D) [True, True, True, False, False, False, False, False, False, False]

1.5 How can you create a list of squares of numbers from 1 to 10 in Python ?

(A) [i**2 for i in range(10)]
(B) [i*i for i in range(1, 11)]
(C) [i**2 for i in range(1, 10)]
(D) [i*2 for i in range(1, 11)]

1.6 What is the mean of test scores ? {70, 70, 80, 85, 85, 90, 95, 95, 100, 100}

(A) 85, 95, and 100
(B) 30
(C) 87
(D) None of the options

1.7 What does "p-value" indicate in hypothesis testing ?

(A) The probability that the null hypothesis is true
(B) The likelihood of observing data at least as extreme as the data observed
(C) The probability that the null hypothesis is false
(D) The probability that the null hypothesis might be both.

1.8 Given n indistinguishable particles and m(>n) distinguishable boxes, we place at random each particle in one of the boxes. The probability that in n preselected boxes, one and only one particle will be found is :

(A) $\frac{n!}{m^n}$

(B) $\frac{(m-1)!\,n!}{(m+n-1)!}$

(C) $\frac{1}{m^n}$

(D) $\frac{1}{m}$

1.9 Which of the following would be the leave on out cross-validation accuracy for k=5 ?

(A) 10/14
(B) 4/14
(C) 6/14
(D) 8/14

1.10 What would be the relation between the time taken by 1-NN, 2-NN, 3-NN ?

(A) 1-NN > 2-NN > 3-NN
(B) 1-NN < 2-NN < 3-NN
(C) 1-NN ~ 2-NN ~ 3-NN
(D) None of the options

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 the instructions therein. (1x10)**

2.1 Causal analysis is commonly applied to census data.

2.2 Data cleaning is not an optional step in the data science process.

2.3 The k-NN algorithm does more computation on test time rather than train time.

2.4 Fancy indexing in NumPy refers to using arrays of integers as indices to access elements.

2.5 A Pandas DataFrame is a two-dimensional data structure similar to a spreadsheet.

2.6 Regression analysis is used to determine if there is a statistical relationship between two categorical variables.

2.7 Tkinter is a built-in Python library not used for creating graphical user interfaces (GUIs).

2.8 A scatter plot is used to show the relationship between two continuous variables.

2.9 A Pandas Series is always one-dimensional.

2.10 Bar charts in Matplotlib can only display vertical bars.

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

	X		Y
3.1	Method to remove whitespace from a string	**A**	applyall()
3.2	Process of transforming raw data into a clean format	**B**	apply()
3.3	Function to generate random numbers from a normal distribution in NumPy	**C**	Histogram
3.4	Performing arithmetic on arrays of different shapes	**D**	strip()
3.5	Two-dimensional data structure in Pandas	**E**	numpy.random.normal()
3.6	Method to apply a function to each DataFrame column	**F**	String
3.7	Graph that shows the frequency of data within intervals	**G**	Broadcasting
3.8	Tkinter widget used to select multiple options from a list	**H**	Listbox widget
3.9	Techniques to handle imbalanced datasets in classification problems	**I**	Data munging
3.10	Python object used to represent an immutable sequence of Unicode characters	**J**	Series
		K	DataFrame
		L	Random Under-Sampling
		M	Random oversampling

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

A	Nan	**B**	Computational	**C**	2-Dimensional
D	Plotting	**E**	Immutable	**F**	Data Cleaning
G	Data Visualization	**H**	Machine Learning	**I**	Universal Function
J	Specific portions	**K**	loc []	**L**	Histogram
M	Pack()				

4.1 In Python, slicing a list or tuple allows you to access a __________ of the original data.

4.2 Strings in Python are __________, meaning their contents cannot be altered.

4.3 Exploratory Data Analysis (EDA) typically involves __________ and summarizing the main characteristics of a dataset.

4.4 __________ is the process of finding and fixing errors or inconsistencies in data.

4.5 Data Science combines statistical analysis and __________ methods to derive insights.

4.6 __________ is the term used for representing missing data in datasets.

4.7 Tkinter's __________ geometry manager is used to stack widgets on top of each other.

4.8 A __________ plot is used to represent the distribution of data using bins.

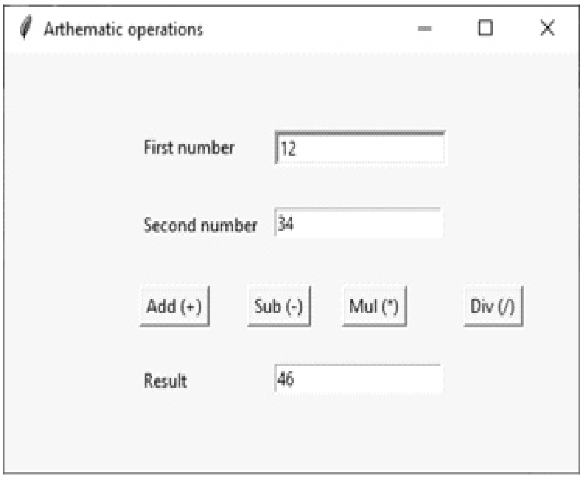
4.9 A Pandas DataFrame is a _____________ labeled data structure.

4.10 To filter rows in a Pandas DataFrame, we can use the __________ method.

PART - TWO

(Answer any FOUR questions)

5. (a) Design a simple Python Tkinter GUI application to perform arithmetic operations shown in Figure. The user inputs two numbers, and the program should display buttons for "**Add (+)**", "**Sub (-)**", "**Mul (*)**", **and** "**Div (/)**". The corresponding result should be displayed in the result field after clicking a button. Implement the logic for each operation. Also ensure error handling for invalid inputs and division by zero.



Arthematic operations

First number 12

Second number 34

Result 46

(b) List the different widgets available in Tkinter. Explain the purpose of any two widgets from the following: Scale, Canvas, Frame, Listbox, and RadioButton. Provide examples to demonstrate their usage.

(10+5)

6. (a) A data frame is given with the following data figure :

S. No.	Item Name	Color	Price
1	Gel	Red	22.5
2	Ball Pen	Black	20
3	Pencil	Blue	6.5
4	Ball Pen	Green	12.5
5	Gel Pen	Green	13
6	Notebook	Red	14.5
7	Ball Pen	Green	18.5
8	Highlighter	Blue	28.5
9	Gel	Red	22.5
10	P Maker	Blue	18.6
11	Pencil	Green	16.5
12	Ball Pen	Green	19.5
13	Pencil	Red	4.5
14	Notebook	Blue	25.5

Answer the following questions:

(i) Group the items based on their color and find the total number of items and their total price for each color.

(ii) Write the average price of each item category (ItemName) concerning its color.

(iii) Identify and write all unique item names along with the distinct colors they come in.

(iv) Calculate the total sum of prices for each color group.

(b) Define Exploratory Data Analysis (EDA) and outline the key steps involved in the process. Briefly describe each step.

(10+5)

7. (a) Explain the difference between correlation and causation. Why is it important to distinguish between the two when interpreting the results of statistical analysis ?

(b) What is the statistics module in Python ? List out some of the key functions provided by this module.

(c) Calculate the normal distribution probability density function using the following data. x = 3, μ = 4 and σ = 2.

(5+5+5)

8. (a) What is regression analysis, and how does it differ from correlation ? Describe the basic idea behind linear regression and its use in predicting outcomes.

(b) Differentiate between positive, negative, and zero correlation. Provide examples of situations where each type of correlation might occur.

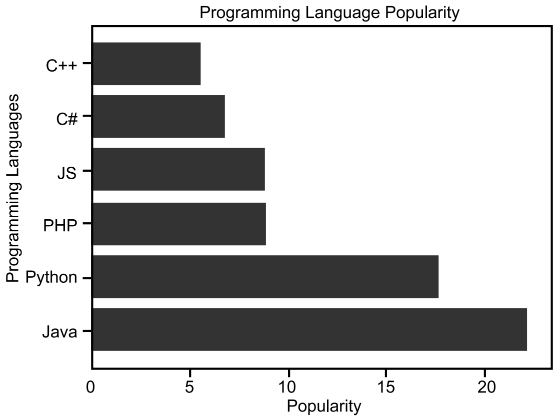
(6+9)

9. (a) What is np.loadtxt() and np.genfromtxt(), and how do they differ in loading data into NumPy arrays ?

(b) Suppose you have an arr = np.array([1, 2, 3, 2, 4, 1, 5, 4, 6]), and find the unique values and their counts in a 1D array.

(c) Mr. Pal has successfully plotted the following horizontal bar graph, which displays the popularity of different programming languages : **X-axis** : Popularity (22.2, 17.6, 8.8, 8, 7.7, 6.7) **Y-axis** : Programming languages (Java, Python, PHP, JS, C#, C++)

The bars are colored green.



Answer the following questions based on the above information :

(i) What method did Mr. Pal use to plot horizontal bars in Python ?

(ii) Write a Python code snippet that Mr. Pal could have used to create this horizontal bar graph with green bars.

(iii) How can the code be modified to display the title "Programming Language Popularity" above the graph ?

(iv) If Mr. Pal wants to display the percentage value at the top of each bar, what changes should be made in the code to achieve this ? How to determine the value of k in k-means clustering.

(4+3+8)

- o 0 o -

SPACE FOR ROUGH WORK