School of Computer Applications BBDU

Roll No.	1	2	2	0	2	5	8	0	8	2

BCADS Second Sessional Tests 2023-24
Fourth Semester
BCADS1402: Data Science

me: 3 Hours **Max. Marks: 60**

(Section A)

1. Attempt All Questions **(1*8)**

a) How do you save a DataFrame as a CSV file in Pandas?
b) What is the syntax for setting the figure size to (6,4) in Matplotlib?
c) How do you load an inbuilt sample dataset from Seaborn?
d) Define temporal and spatial data, provide an example for each.
e) Advantages of Numpy compared to Python lists.
f) Explain the concept of standardization.
g) What is data science and why is it important in today's world.
h) Name of data analytics methodologies that are commonly used?

Section B

2. Attempt any Two. **(6*2)**

a) Discuss the Central Limit Theorem and its importance in statistical analysis. Provide a concise definition and explain its relevance in ensuring the reliability of statistical inference and estimation.
b) Explain the procedure for accessing IBM Cloud, including account creation and navigating the dashboard.
c) What are the various domains within Data Science, and what are the typical roles associated with each domain?
d) Scenario: Imagine a system that classifies images of different types of flowers (Rose, Sunflower). The confusion matrix would look something like this:

	Predicted Rose	Predicted Sunflower
Actual Rose	104	14
Actual Sunflower	32	30

Find Accuracy, Precision, Recall and F1 score for the above Matrix.

Q3 Attempt Any Two. **(5*2)**

a) Elaborate on the methodologies frequently utilized in data analytics, highlighting their real-world applications. Describe the stages of executing data analytics projects and analyze the challenges encountered at each step.

b) Explain the fundamental principles of data science and demonstrate their relevance across different industries.

c) What are the essential skills that a data scientist must possess?

Q4 Attempt Any Two. **(5*2)**

a) Discuss the importance of data cleaning in preparing the dataset for analysis. What techniques did we employ to handle missing values and duplicates?

b) What are outliers in a dataset, and what methods can be used to analyze, estimate, and remove them?

c) Discuss the importance of Seaborn library in data visualization, highlighting its key features, advantages, and examples of plot creation.?

Q5 Attempt Any Two. **(5*2)**

a) The average score on a test is 45 with a standard deviation of 10. With a new teaching curriculum introduced it is believed that this score will change. On random testing, the score of 38 students, the mean was found to be 58. With a 0.05 significance level, is there any evidence to support this claim? **(Take critical value 1.96)**.

b) Provide Python code utilizing Pandas to (1) load a CSV file, (2) display the DataFrame, (3) compute statistics, (4) filter by 'female' gender, and (5) save the filtered data as a new CSV file.

c) Explain the concepts of linear regression and logistic regression.

Q6 Attempt Any Two **(5*2)**

a) Describe the features offered by Watson Studio in supporting data science projects.

b) Describe the decision tree classifier algorithm in machine learning, including its construction process and prediction mechanism. Analyze the significance of decision trees' interpretability.

c) Describe the importance of integrated environments in data science projects. Explain how tools like Jupyter Notebooks or Anaconda enhance collaboration and streamline the development process for data scientists?