VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

REG.NO.:

SLOT: C1

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

Programme Name & Branch	: B.Tech [BDS, BEC, BME]
Course Code and Course Name	: BCSE331L Exploratory Data Analysis
Faculty Name(s)	: Dr. Manoov R, Dr. Nalini, Dr. Prakash M
Exam Duration	: 90 minutes **Maximum Marks: 50**

General instruction(s):

- Answer All Questions
- M - Max mark; CO – Course Outcome; BL – Blooms Taxonomy Level (1 – Remember, 2 – Understand, 3 – Apply, 4 – Analyse, 5 – Evaluate, 6 – Create)
- Course Outcomes
  CO-2: Summarize the data using basic statistics. Visualize the data using basic graphs and plots.
  CO-3: Identify the outliers if any in the data set.
  CO-5: Apply Techniques for handling multi-dimensional data.

Q. No	Question	Module	Marks	CO	BL
1.	A retail company tracks its daily sales data and wants to analyse the relationship between its sales and daily advertising spend. The dataset includes two time series variables: *Daily_Sales* and *Daily_Ad_Spend*. The company wants to understand if there is a correlation between these two variables on a weekly basis, as their advertising campaigns run on a weekly cycle. Explain why it would be beneficial to resample the daily data to a weekly frequency for this specific analysis. Describe the steps you would take to perform this resampling.	3	10	5	4
2.	A data scientist is analysing the monthly average temperature data for a city over several decades. They plot the time series and observe that the mean temperature seems to be gradually increasing over time. The variance of the fluctuations also appears to grow as the temperatures rise. Based on the data scientist's observations, is the time series likely stationary or non-stationary? Describe one commonly used statistical test for assessing stationarity and justify your answer.	3	10	5	3
3.	As a data analyst you are analysing a dataset from an e-commerce company containing two variables for a sample of customers: *'Annual Spending (USD)'* and *'Age'*. a) Which single statistical summary measure (for each variable) would be most appropriate to describe the central tendency if the 'Annual Spending' data is heavily right-skewed (due to a few high spenders)? Justify your choice for both variables. (5) b) Identify the best plot for two continuous variables in the above context, justify your selection, and explain the pattern to analyze.(5)	4	10	2	4

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

REG.NO.:

SLOT: C1

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

4.	You are a data analyst working with a geographical dataset of soil samples from a large agricultural field. Each sample has been analysed for four key features: *pH Level, Nitrogen Content, Moisture Retention,* and *Clay Content.* Your primary goal is to use this data to identify and group similar soil samples into distinct *soil type zones.* This information will be used by the agricultural team to implement targeted, more efficient fertilization strategies. **Sample_ ID** / **PH_ Level** / **Nitrogen_ Content** / **Moisture_ Retention** / **Clay_ Content** 1 / 6.2 / 25.1 / 45.5 / 18.2 2 / 7.1 / 18.3 / 32.1 / 35.8 3 / 5.8 / 30.5 / 55.2 / 12.1 4 / 7 / 17.8 / 30.5 / 38.9 5 / 6 / 26.8 / 48.1 / 15 Apply Spectral Clustering to determine the optimal number of clusters for the dataset. *(Your final output should be a detailed, step-by-step explanation of your approach.)*	5	10	5	4
5.	A retail company wants to group its customers based on two features: Annual Spending (in USD) and Number of Items Purchased. **CustomerID** / **Annual Spending (USD)** / **Number of Items Purchased** C01 / 1.2 / 45 C02 / 1.5 / 50 C03 / 8.5 / 15 C04 / 9 / 18 C05 / 20.5 / 5 The company uses Hierarchical Agglomerative Clustering, starting with each customer as a separate cluster. Which two customers will be merged first? Justify your answer with the calculation.	5	10	3	5

****