



VIT

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

REG.NO.: 23808 0071

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - I
FALL SEMESTER 2026-2027

SLOT: A2

Programme Name & Branch : B.Tech & Computer Science and Engineering
Course Code and Course Name : BCSE331L and Exploratory Data Analysis
Faculty Name(s) : Prof. Gunavathi C, Prof. M. Prakash and Prof. G. N. Balaji
Class Number(s) : VL2026270102003, 2445, 2447
Date of Examination : 09.08.2026
Exam Duration : 90 minutes Maximum Marks: 50

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)
- CO1: Handle missing data in the real world data sets by choosing appropriate methods.
- CO2: Summarize the data using basic statistics. Visualize the data using basic graphs and plots.
- CO3: Identify the outliers if any in the data set.

Q. No	Question	M	CO	BL
1.	A hospital wants to analyze patient records to improve healthcare services. The hospital.csv from Table1 is collected from the outpatient department. a) Explain the steps involved in performing Exploratory Data Analysis (EDA) for the given dataset. Clearly describe how each step helps in preparing the data for analysis. (5 Marks) b) Identify the data quality issues present in the dataset and explain how EDA can help derive meaningful insights for decision-making. (5 Marks)	10	1	2
2.	Consider the dataset from Table 1, perform data normalization on the Treatment_Cost (₹) attribute after handling the missing values and invalid entries. a) Apply the following normalization techniques to the Treatment_Cost (₹) attribute. Show all intermediate calculations and the normalized values for each record. (7 Marks) • Min-Max Normalization • Z-score Normalization (using Mean and Mean Absolute Deviation (MAD)) • Normalization by Decimal Scaling b) Compare the three normalization techniques based on their computation, output range, and suitable applications. Justify which normalization technique is most appropriate for the given dataset. (3 Marks)	10	2	3
3.	Using hospital.csv dataset from Table 1, write a Python program to perform the following preprocessing tasks. a) Identify and remove duplicate records. Also detect and handle missing values represented as NaN, ?, and NULL using appropriate preprocessing techniques. Display the updated dataset. (5 Marks) b) Identify and handle invalid data entries (such as unrealistic age, negative treatment cost, and inconsistent categorical values). Display the final cleaned dataset and report the total number of records before and after preprocessing. (5 Marks)	10	2	3

4.	Using the hospital.csv dataset given in Table 1 after cleaning, identify the outlier(s) in the Treatment_Cost (₹) attribute using the Box Plot (IQR) method. a) For the Treatment_Cost (₹) values in ascending order and compute the following: - First Quartile - Median - Third Quartile - Interquartile Range (5 Marks) b) Determine the Lower Fence and Upper Fence using the IQR method. Identify the outlier(s) in the dataset and explain how they may affect data analysis. Suggest one suitable technique to handle the identified outlier(s). (5 Marks)	10	2	3
5.	Consider the hospital.csv dataset from Table 1 after cleaning, perform a suitable univariate and bivariate analysis to support the hospital management in understanding patient characteristics and treatment patterns. a) Select one numerical attribute and one categorical attribute from the dataset. Recommend appropriate univariate analysis techniques for each attribute and justify why they are suitable for the selected data. Interpret the expected insights that can be obtained. (5 Marks) b) Select two relevant attributes from the dataset and recommend a suitable bivariate analysis technique to study their relationship. Justify your choice of visualization or statistical method, explain the type of relationship you expect to observe, and discuss how the results could support hospital decision-making. (5 Marks)	10	3	4

Table1: hospital.csv

Patient_ID	Age	Disease	Treatment_Cost (₹)	Outcome
P101	35	Diabetes	12000	Recovered
P102	NaN	Hypertension	18000	Recovered
P103	42	Diabetes	?	Under Treatment
P104	58	Heart Disease	25000	Recovered
P105	27	Fever	3500	Recovered
P106	150	Diabetes	15000	Recovered
P107	45	NULL	22000	Under Treatment
P108	33	Asthma	-9999	Recovered
P101	35	Diabetes	12000	Recovered
P109	?	Hypertension	98000	Unknown
