



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

REG.NO.:

SLOT: E2+TE2

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

Programme Name & Branch : B.Tech. and CSE
Course Code and Course Name : BACSE301 - Exploratory Data Analysis
Faculty Name(s) : Prof. Nalini N, Prof. Kathiravan S, Prof. Mehfooza, Prof. Ramesh C
Class Number(s) : VL2026270101739, 2818, 1752, 1749
Date of Examination : 13.08.2026
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:
CO1: Analyze real world data sets by selecting and applying appropriate data analysis methods.
CO2: Apply data preprocessing techniques and detect outliers to prepare data for analysis.

Q. No	Question	M	CO	BL																																										
1.	A hospital stores the following synthetic patient records. The values are illustrative and contain no real personal information. Dates are recorded in DD-MM-YYYY format. <table><tr><th>Patient_ID</th><th>Name</th><th>Address</th><th>DoB</th><th>Email</th><th>Gender</th><th>Weight (kg)</th></tr><tr><td>P001</td><td>Aarav Kumar</td><td>Bengaluru, India</td><td>30-12-1989</td><td>patient001@example.org</td><td>M</td><td>68</td></tr><tr><td>P002</td><td>Yashvi Rao</td><td>Bergen, Norway</td><td>10-07-2018</td><td>patient002@example.org</td><td>F</td><td>24</td></tr><tr><td>P003</td><td>Ananya Mukherjee</td><td>Bergen, Norway</td><td>10-12-1997</td><td>patient003@example.org</td><td>F</td><td>54</td></tr><tr><td>P004</td><td>Asha Gaire</td><td>Butwal, Nepal</td><td>30-11-1990</td><td>patient004@example.org</td><td>F</td><td>63</td></tr><tr><td>P005</td><td>Ola Nordmann</td><td>Stockholm, Sweden</td><td>12-12-1979</td><td>patient005@example.org</td><td>M</td><td>75</td></tr></table> For each of the seven attributes in the table: a) Specify an appropriate logical or database data type. [4 marks] b) Classify its level of measurement as nominal, ordinal, interval, or ratio. [4 marks] c) Briefly justify the measurement-scale classifications for DoB, Gender, and Weight (kg). [2 marks]	Patient_ID	Name	Address	DoB	Email	Gender	Weight (kg)	P001	Aarav Kumar	Bengaluru, India	30-12-1989	patient001@example.org	M	68	P002	Yashvi Rao	Bergen, Norway	10-07-2018	patient002@example.org	F	24	P003	Ananya Mukherjee	Bergen, Norway	10-12-1997	patient003@example.org	F	54	P004	Asha Gaire	Butwal, Nepal	30-11-1990	patient004@example.org	F	63	P005	Ola Nordmann	Stockholm, Sweden	12-12-1979	patient005@example.org	M	75	10	1	4
Patient_ID	Name	Address	DoB	Email	Gender	Weight (kg)																																								
P001	Aarav Kumar	Bengaluru, India	30-12-1989	patient001@example.org	M	68																																								
P002	Yashvi Rao	Bergen, Norway	10-07-2018	patient002@example.org	F	24																																								
P003	Ananya Mukherjee	Bergen, Norway	10-12-1997	patient003@example.org	F	54																																								
P004	Asha Gaire	Butwal, Nepal	30-11-1990	patient004@example.org	F	63																																								
P005	Ola Nordmann	Stockholm, Sweden	12-12-1979	patient005@example.org	M	75																																								
2.	Compare data collection through a documented Web API with data collection through web scraping. Discuss the two approaches in terms of interface stability, data accessibility, and the predictability of the obtained data structure. Support your comparison with suitable examples.	10	1	2																																										
3.	Consider the following data vector: $v = (3, 5, 5, 8, 9, 12, 12, 13, 15, 16, 17, 19, 22, 24, 25, 134)$ For each observation (v_i), calculate the following transformations: a) Min-max scaling to the interval [0, 1] [4 marks] b) Population (z)-score standardization [6 marks] Use ($n=16$) and the population standard deviation, with (n) as the denominator. Consider rounding off each final transformed value to three decimal places. Present the transformed values in the same order as the original observations.	10	2	3																																										



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

4.	A university student information system contains the following data. During data entry, the Email field was requested only when Scholarship = Yes. <table><thead><tr><th>Stu_ID</th><th>Dept</th><th>Gender</th><th>CGPA</th><th>Scholarship</th><th>Email</th></tr></thead><tbody><tr><td>101</td><td>CSE</td><td>Female</td><td>8.5</td><td>Yes</td><td>student101@univ.edu</td></tr><tr><td>102</td><td>ECE</td><td>Male</td><td>7.8</td><td>No</td><td>NULL</td></tr><tr><td>103</td><td>CSE</td><td>Male</td><td>9.1</td><td>Yes</td><td>student103@univ.edu</td></tr><tr><td>104</td><td>MECH</td><td>Female</td><td>6.9</td><td>No</td><td>NULL</td></tr><tr><td>105</td><td>ECE</td><td>Female</td><td>8.2</td><td>Yes</td><td>student105@univ.edu</td></tr><tr><td>106</td><td>CSE</td><td>Male</td><td>7.5</td><td>No</td><td>NULL</td></tr><tr><td>107</td><td>MECH</td><td>Male</td><td>8.0</td><td>Yes</td><td>student107@univ.edu</td></tr><tr><td>108</td><td>EEE</td><td>Female</td><td>7.2</td><td>No</td><td>NULL</td></tr></tbody></table> a) Define the missingness indicator R_{Email} (1 = observed, 0 = missing). Using the stated data-entry rule, classify the Email missingness mechanism as MCAR, MAR, or MNAR and justify the classification in terms of observed and unobserved variables. [6 marks] b) Discuss whether imputation is appropriate for the missing Email values. Justify your answer with reference to the nature of the Email variable, the data-entry rule, and the available observed data. [4 marks]	Stu_ID	Dept	Gender	CGPA	Scholarship	Email	101	CSE	Female	8.5	Yes	student101@univ.edu	102	ECE	Male	7.8	No	NULL	103	CSE	Male	9.1	Yes	student103@univ.edu	104	MECH	Female	6.9	No	NULL	105	ECE	Female	8.2	Yes	student105@univ.edu	106	CSE	Male	7.5	No	NULL	107	MECH	Male	8.0	Yes	student107@univ.edu	108	EEE	Female	7.2	No	NULL	10	2	3
Stu_ID	Dept	Gender	CGPA	Scholarship	Email																																																					
101	CSE	Female	8.5	Yes	student101@univ.edu																																																					
102	ECE	Male	7.8	No	NULL																																																					
103	CSE	Male	9.1	Yes	student103@univ.edu																																																					
104	MECH	Female	6.9	No	NULL																																																					
105	ECE	Female	8.2	Yes	student105@univ.edu																																																					
106	CSE	Male	7.5	No	NULL																																																					
107	MECH	Male	8.0	Yes	student107@univ.edu																																																					
108	EEE	Female	7.2	No	NULL																																																					
5.	An online shopping company records the following transaction amounts, expressed in Indian rupees (INR): 850, 1100, 1250, 1500, 1750, 2000, 2250, 2500, 2800, 3100, 3400, 3700, 4000, 4300 4600, 4900, 6500, 12000, 520,000, 750,000. a) State the direction of skewness of the original sample distribution. [2 marks] b) For every transaction amount x_i, calculate the following skewness-reducing transformations and present the results in one table: (i) Natural-log transformation: $y_i = \ln(x_i)$ (ii) Square-root transformation: $y_i = \sqrt{x_i}$ (iii) Fourth-root power transformation: $y_i = \sqrt[4]{x_i}$ [6 marks] c) Using the transformed values obtained in 5.(b), compare the natural-log, square-root, and fourth-root transformations in terms of their effectiveness in reducing skewness. Identify the most effective transformation and briefly justify your answer. [2 marks]	10	2	3																																																						