



VIT

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

REG.NO.:

SLOT: C2

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

Programme Name & Branch : B.Tech [BDS/ BME]
Course Code and Course Name : BCSE331L - Exploratory Data Analysis
Faculty Name(s) : Prof. Prakash M, Prof. Manooov R
Class Number(s) : VL2025260101620, 1631
Date of Examination : 07-10-2025
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:
- CO3: Identify the outliers if any in the data set.
- CO4: Choose appropriate feature selection and dimensionality reduction.

Q. No	Question	M	CO	BL																																																							
1.	A data scientist working in the energy sector is tasked with supporting the national power grid authority, aims to forecast future electricity demand in order to optimize power generation and prevent shortages. A dataset <code>electricity_demand.csv</code> , containing two columns: Date (YYYY-MM-DD) representing daily records over the past 15 years, and Demand (MW) representing the electricity consumption in megawatts. Explain the key characteristics of a time series dataset, then develop a robust time series forecasting model. Load the dataset into Python and generate insights by performing time-based indexing, creating basic visualizations, grouping the data, applying resampling techniques, and plotting the differences.	10	3	3																																																							
2.	A dataset (<code>students.csv</code>) contains <code>RollNo</code>, <code>Name</code>, <code>Internal</code>, <code>External</code>, <code>TotalMarks</code> with 10 rows are given below. i) Compute 1D statistical summaries for the attribute <code>TotalMarks</code> including Mean, Median, Interquartile Range, Variance, Standard Deviation, Skewness, Minimum, Maximum, 1st Quantile. Provide the summarization in the table. (5) ii) Draw a Sunray Plot without changing the above summarization order to visually represent the statistics and explain. (5) <table border="1"> <thead> <tr> <th>RollNo</th><th>Name</th><th>Internal</th><th>External</th><th>TotalMarks</th></tr> </thead> <tbody> <tr><td>20BD01</td><td>Advik</td><td>30</td><td>48</td><td>78</td></tr> <tr><td>20BD02</td><td>Bhaves</td><td>31</td><td>54</td><td>85</td></tr> <tr><td>20BD03</td><td>Chirag</td><td>32</td><td>35</td><td>67</td></tr> <tr><td>20BD04</td><td>Daksh</td><td>31</td><td>59</td><td>90</td></tr> <tr><td>20BD05</td><td>Eklavya</td><td>34</td><td>38</td><td>72</td></tr> <tr><td>20BD06</td><td>Gaurav</td><td>28</td><td>59</td><td>88</td></tr> <tr><td>20BD07</td><td>Hrithik</td><td>20</td><td>40</td><td>60</td></tr> <tr><td>20BD08</td><td>Ishaan</td><td>36</td><td>59</td><td>95</td></tr> <tr><td>20BD09</td><td>Jayden</td><td>33</td><td>37</td><td>70</td></tr> <tr><td>20BD10</td><td>Kabir</td><td>32</td><td>48</td><td>80</td></tr> </tbody> </table>	RollNo	Name	Internal	External	TotalMarks	20BD01	Advik	30	48	78	20BD02	Bhaves	31	54	85	20BD03	Chirag	32	35	67	20BD04	Daksh	31	59	90	20BD05	Eklavya	34	38	72	20BD06	Gaurav	28	59	88	20BD07	Hrithik	20	40	60	20BD08	Ishaan	36	59	95	20BD09	Jayden	33	37	70	20BD10	Kabir	32	48	80	10	4	3
RollNo	Name	Internal	External	TotalMarks																																																							
20BD01	Advik	30	48	78																																																							
20BD02	Bhaves	31	54	85																																																							
20BD03	Chirag	32	35	67																																																							
20BD04	Daksh	31	59	90																																																							
20BD05	Eklavya	34	38	72																																																							
20BD06	Gaurav	28	59	88																																																							
20BD07	Hrithik	20	40	60																																																							
20BD08	Ishaan	36	59	95																																																							
20BD09	Jayden	33	37	70																																																							
20BD10	Kabir	32	48	80																																																							

3.	i) A school administrator is evaluating the improvement in students' test scores after a new tutoring program. The score increases (in points) for 8 students are recorded as follows: 12.4, 15.1, 9.5, 14.2, 11.5, 13.0, 10.4, 16.5 Calculate the mean difference (Δ) score improvement for these students and interpret what this average suggests about the effectiveness of the tutoring program. (5) ii) A language instructor is assessing the proficiency levels of five students in reading, writing, and speaking skills. The chart below shows the percentage distribution of each skill in their overall language competency: <table><tr><th>Student</th><th>Reading (%)</th><th>Writing (%)</th><th>Speaking (%)</th></tr><tr><td>P</td><td>50</td><td>30</td><td>20</td></tr><tr><td>Q</td><td>40</td><td>35</td><td>25</td></tr><tr><td>R</td><td>30</td><td>40</td><td>30</td></tr><tr><td>S</td><td>60</td><td>25</td><td>15</td></tr><tr><td>T</td><td>45</td><td>30</td><td>25</td></tr></table> Plot these students' skill profiles on a ternary diagram to visualize the distribution of reading, writing, and speaking skills among the students. (5)	Student	Reading (%)	Writing (%)	Speaking (%)	P	50	30	20	Q	40	35	25	R	30	40	30	S	60	25	15	T	45	30	25	10	4	3																									
Student	Reading (%)	Writing (%)	Speaking (%)																																																		
P	50	30	20																																																		
Q	40	35	25																																																		
R	30	40	30																																																		
S	60	25	15																																																		
T	45	30	25																																																		
4.	Given the distance matrix among six points data points $P1$ to $P6$ is given in the below table <table><tr><th></th><th>p1</th><th>p2</th><th>p3</th><th>p4</th><th>p5</th><th>p6</th></tr><tr><td>p1</td><td>0.00</td><td>0.24</td><td>0.22</td><td>0.37</td><td>0.34</td><td>0.23</td></tr><tr><td>p2</td><td>0.24</td><td>0.00</td><td>0.15</td><td>0.20</td><td>0.14</td><td>0.25</td></tr><tr><td>p3</td><td>0.22</td><td>0.15</td><td>0.00</td><td>0.15</td><td>0.28</td><td>0.11</td></tr><tr><td>p4</td><td>0.37</td><td>0.20</td><td>0.15</td><td>0.00</td><td>0.29</td><td>0.22</td></tr><tr><td>p5</td><td>0.34</td><td>0.14</td><td>0.28</td><td>0.29</td><td>0.00</td><td>0.39</td></tr><tr><td>p6</td><td>0.23</td><td>0.25</td><td>0.11</td><td>0.22</td><td>0.39</td><td>0.00</td></tr></table> Perform Agglomerative Hierarchical Clustering and draw the corresponding dendrograms for the final cluster partition $k=3$ using 1. Complete Linkage (5) 2. Average Linkage (5)		p1	p2	p3	p4	p5	p6	p1	0.00	0.24	0.22	0.37	0.34	0.23	p2	0.24	0.00	0.15	0.20	0.14	0.25	p3	0.22	0.15	0.00	0.15	0.28	0.11	p4	0.37	0.20	0.15	0.00	0.29	0.22	p5	0.34	0.14	0.28	0.29	0.00	0.39	p6	0.23	0.25	0.11	0.22	0.39	0.00	10	5	4
	p1	p2	p3	p4	p5	p6																																															
p1	0.00	0.24	0.22	0.37	0.34	0.23																																															
p2	0.24	0.00	0.15	0.20	0.14	0.25																																															
p3	0.22	0.15	0.00	0.15	0.28	0.11																																															
p4	0.37	0.20	0.15	0.00	0.29	0.22																																															
p5	0.34	0.14	0.28	0.29	0.00	0.39																																															
p6	0.23	0.25	0.11	0.22	0.39	0.00																																															
5.	Using Vector Space Model, find the cosine similarity between the documents and the query below. Document 1: "The quick brown fox jumps over the lazy dog." Document 2: "A brown dog chased the fox." Document 3: "The dog is lazy." Document 4: "The lazy fox befriended a brown dog." Query : "brown lazy dog" Arrange the documents as per the similarity ranking.	10	5	4																																																	
