



PRESIDENCY UNIVERSITY

BENGALURU

Roll No.														
----------	--	--	--	--	--	--	--	--	--	--	--	--	--	--

End - Term Examinations –May/June 2026

Date: 19-05-2026

Time: 09:30am – 12:30pm

School: SOCSE	Program: B. Tech. CSD	
Course Code: ADS2107	Course Name: Exploratory Data Analytics	
Semester: IV	Max Marks: 100	Weightage: 50%

CO - Levels	C01	C02	C03	C04	C05
Marks	4	22	22	26	26

Instructions:

- (i) Read all questions carefully and answer them accordingly.
- (ii) Do not write anything on the question paper other than roll numbers.

Part A

Answer ALL the Questions. Each question carries 2marks.

10Q x 2M=20M

1.	What is meant by hypothesis testing in statistics? Demonstrate using a small scenario	2 Marks	L1	C01
2.	What are the different types of variables found in a dataset during data analysis?	2 Marks	L1	C01
3.	Explain how correlation between two variables X and Y is interpreted using the correlation coefficient.	2 Marks	L2	C02
4.	Explain how min-max normalization is applied to a data value.	2 Marks	L2	C03
5.	Summarize the concept of interpolation, and how it helps in data analysis.	2 Marks	L2	C04
6.	Given the dataset [56,64, NaN,48,30], apply an appropriate imputation technique to handle the missing value and briefly explain the steps	2 Marks	L2	C04
7.	Explain how one-hot encoding is applied to categorical variables.	2 Marks	L2	C04

8.	What type of insights can be obtained from a histogram compared to a kernel density plot?	2 Marks	L2	C05
9.	Explain why a histogram may sometimes provide misleading insights if the bin size is not chosen appropriately.	2 Marks	L2	C05
10.	Explain how a box plot provides insights into the spread and central tendency of a dataset.	2 Marks	L2	C05

Part B

Answer the Questions.

Total Marks 80M

11.	a.	Given the datasets: $X = [10, 12, 12, 13, 12, 11, 90, 13, 12]$ $Y = [8, 9, 11, 10, 9, 8, 95, 10, 9]$ Explain the concepts of covariance and correlation, including their definitions and significance in measuring relationships between variables. Compute the covariance between X and Y, showing all intermediate steps. Calculate the Pearson correlation coefficient for the given datasets. Interpret the results obtained, discussing the strength and direction of the relationship between the variables. Comment on the effect of any extreme values (outliers) on these measures.	20 Marks	L2	C02
Or					
12.	a.	Explain the classification of Exploratory Data Analysis (EDA) into univariate, bivariate, and multivariate analysis, stating the objective of each with suitable examples. Differentiate between graphical and non-graphical techniques and describe at least two methods for each with their purpose.	10 Marks	L2	C02
	b.	Discuss appropriate EDA techniques for analyzing distributions, patterns, and relationships in datasets. Based on this, explain how suitable methods are selected for datasets containing both numerical and categorical variables.	10 Marks	L2	C02
13.	a.	Given the dataset: [21,10,13,41,100,121,1000], Apply Min-Max Scaling to transform the dataset into the range [0,1], showing all intermediate steps and final values. Apply Robust Scaling using the median and interquartile range (IQR), showing all	20 Marks	L3	C03

		intermediate steps and final values. Compare the scaled results obtained from both methods in terms of the effect of the outlier and the distribution of transformed values. Analyze how each scaling technique influences the dataset and select which method is more appropriate for this dataset. Also, state the conditions under which each technique is preferred.			
Or					
14.	a.	Given the categorical dataset ["Red", "Blue", "Green", "Blue", "Red", "Yellow"], apply both One-Hot Encoding and Label Encoding to transform the data, showing all intermediate steps and final encoded representations for each method. Compare the resulting feature representations in terms of dimensionality and information preservation and analyze how each encoding impacts the dataset. Based on your observations, justify which encoding technique is more suitable for this dataset and in what scenarios one may be preferred over the other.	20 Marks	L3	CO3
15.	a.	Given the dataset representing employee salaries [20000, 25000, 30000, 40000, 60000, 150000], which exhibits positive skewness, apply logarithmic (natural log) and square root transformations, showing all intermediate calculations and transformed values. Compare both transformed datasets with the original in terms of skewness reduction and distribution. Analyze the effectiveness of each method and choose which is more suitable for this dataset. Further, explain the conditions under which each transformation is preferred.	20 Marks	L3	CO4
Or					
16.	a.	Given the dataset [3, 4, 6, 8, 12, 50], which exhibits positive skewness, explain the concept of the Box-Cox transformation and its role in stabilizing variance and reducing skewness. Apply the transformation for $\lambda = 0$ and $\lambda = 0.5$, showing the necessary steps. Compare the effects of these values on the distribution in terms of skewness and spread and explain why Box-Cox may be preferred over logarithmic or square root transformations in certain scenarios.	20 Marks	L3	CO4
17.	a.	. Given a dataset representing the scores of students in three subjects over eight assessments: Math: [65,70,75,80,85,90,78,72] Science: [60,68,72,78,82,88,74,69] English: [55,60,65,70,75,80,68,62]	20 Marks	L4	CO5

		Explain how visualizations such as a box plot and a violin plot can be used to analyze the distribution and spread of data across subjects. Using the given dataset, analyze the distribution of scores in each subject, commenting on central tendency, variability, and any noticeable patterns. Compare the distributions across the three subjects and identify any similarities or differences in performance trends. Analyze how the chosen visualizations help in identifying patterns, variations, or anomalies in the dataset, and discuss their effectiveness in supporting data interpretation			
Or					
18.	a.	Given a dataset representing the number of website visitors across different traffic sources over six months, Month: ["Jan", "Feb", "Mar", "Apr", "May", "Jun"], Organic: [500, 550, 600, 650, 700, 750], Paid: [300, 320, 340, 360, 400, 420], Referral: [200, 210, 220, 230, 250, 270] Explain how appropriate visualizations such as stacked bar charts and area charts can be used to represent trends and relative contributions in the data. Analyze how these visualizations help in understanding growth patterns, comparing traffic sources over time, and identifying dominant contributors. Based on your analysis, discuss which type of visualization is more effective for different aspects of the data and justify your reasoning.	20 Marks	L4	C05