Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

Mid - Term Examinations – October 2025

Date: 11-10-2025 **Time:** 02.00pm to 03.30pm

School: SOCSE	**Program**: Data Science	
Course Code: ADS2007	**Course Name:** Exploratory Data Analysis	
Semester: V	**Max Marks**: 50	**Weightage**: 25%

CO - Levels	**CO1**	**CO2**	**CO3**	**CO4**	**CO5**
Marks	**24**	**26**			

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

SET-B

Part A

Answer ALL the Questions. Each question carries 2marks. **5Q x 2M=10M**

1	Compare discrete variables & continuous variables with an example?	**2 Marks**	**L1**	**CO1**
2	Outline the following: a. Identification of duplicate data b. Removal of duplicate data	**2 Marks**	**L2**	**CO2**
3	Define isolation forest method with an example?	**2 Marks**	**L2**	**CO2**
4	Summarize the difference between exploratory data analysis & inferential statistics with an example?	**2 Marks**	**L1**	**CO1**
5	Explain advantages and disadvantages of one-hot encoding?	**2 Marks**	**L2**	**CO2**

Part B

Answer the Questions. Total Marks 40M

6.	**a.**	**A data analytics team is working with a healthcare dataset containing patient details such as age, gender, lifestyle habits, medical history, and diagnostic test results. The aim is to perform Exploratory Data Analysis (EDA) before building a predictive model.** **Answer the following based on this scenario:** **a) How would you carry out Data Understanding to check the types of variables, data quality, and structure of this healthcare dataset?** **b) During Correlation Analysis, the team finds that "smoking frequency" and "lung cancer risk score" have a correlation of +0.82. Interpret this result.** **c) If the team applies Clustering on patient lifestyle habits, what steps should they follow to form meaningful clusters?** **d) The dataset contains some abnormal values, e.g., patient ages recorded as -5 or 150 years. Explain why these are considered outliers and how you would handle them.** **e) A researcher assumes that *"There is no significant difference in average cholesterol levels between smokers and non-smokers"*. Formulate the Null Hypothesis (H_0) and Alternative Hypothesis (H_1) for this study.**	**10 Marks**	**L1**	**CO 1**
Or					
7.	**a.**	Illustrate various types of exploratory data analysis in detail?	**10 Marks**	**L1**	**CO 1**

8.	**a.**	Apply the python code for the following: a. Complete Case Analysis b. Constant Imputation c. Mode Imputation	**10 Marks**	**L2**	**CO 2**
Or					
9.	**a.**	**A university wants to analyze student exam performance. The** average score in a statistics exam is **70**, with a **standard deviation of 10**. A student named Sadiq scored **85 marks**. **a.** Calculate the **Z-score** of Sadiq marks. **b.** Interpret the Z-score. **c.** If another student scored **60 marks**, find their Z-score and compare it with Sadiq's performance.	**10 Marks**	**L2**	**CO 2**

10.	**a.**	Demonstrate in detail what are reasons to use the exploratory data analysis in data science and machine learning?	**10 Marks**	**L1**	**CO 1**
Or					
11.	**a.**	A financial company is building a **machine learning model** to predict whether a loan applicant will **default or not**. The dataset contains the following raw features**:** • **Applicant's age** • **Annual income** • **Loan amount** • **Employment history (text data: e.g., *"2 years"*, *"5 years"*, *"<1 year"*)** • **Credit score** • **Date of loan application** • **Loan repayment status (target variable: *Default / No Default*)** **Answer the following:** **a)** How would you convert the **employment history** feature into a numerical format suitable for the model? **b**) From the **date of loan application**, what new features could you engineer to improve model performance? **c**) If **annual income** and **loan amount** are given, suggest one derived feature that might better capture financial risk. **d**) How would you handle **categorical variables** such as "employment type" (*Salaried, Self-employed, Unemployed*) during feature engineering? **e**) Why is **feature scaling (normalization/standardization)** important in this dataset, and which features would require it?	**10 Marks**	**L1**	**CO 1**

12.	**a.**	Model the working of isolation forest algorithm with python code snippet in detail?	**10 Marks**	**L2**	**CO 2**
Or					
13.	**a.**	Construct the following with respect to categorial encoding in exploratory data analysis: **a**. Binary Encoding **b.** Ordinal Encoding **c.** Label Encoding	**10 Marks**	**L2**	**CO 2**