Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

End - Term Examinations – December 2025

Date: 24 – 12- 2025 **Time:** 01:00pm – 04:00pm

School: SOCSE	**Program:** Data Science	
Course Code: ADS2007	**Course Name:** Exploratory Data Analysis	
Semester: V	**Max Marks**: 100	**Weightage**: 50%

CO - Levels	CO1	CO2	CO3	CO4	CO5
Marks	27	27	23	23	

Instructions:

(i) Read all questions carefully and answer accordingly.

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

Part A

Answer ALL the Questions. Each question carries 2marks. **10Q x 2M=20M**

1.	Compare between supervised and unsupervised machine learning?	**2 Marks**	**L2**	**CO1**
2.	Why is exploratory data analysis used in data science and machine learning?	**2 Marks**	**L1**	**CO1**
3.	Recall the reason why to use KNN imputation in EDA?	**2 Marks**	**L1**	**CO2**
4.	Define Inter Quartile Range with appropriate example?	**2 Marks**	**L1**	**CO2**
5.	List out the reasons to use data transformation in EDA?	**2 Marks**	**L1**	**CO3**
6.	Infer robust scaling technique with appropriate example?	**2 Marks**	**L2**	**CO3**
7.	List and explain advantages and disadvantages of data visualization in EDA?	**2 Marks**	**L1**	**CO4**
8.	Outline any 2 categorical plots in seaborn library?	**2 Marks**	**L2**	**CO4**
9.	What is hypothesis testing with it's types in EDA?	**2 Marks**	**L1**	**CO1**
10.	Summarize advantages and disadvantages of isolation forest algorithm?	**2 Marks**	**L2**	**CO2**

Part B

Answer the Questions. **Total Marks 80M**

11.	**a.**	Shop-Ease, a growing e-commerce platform, wants to build an end-to-end machine learning pipeline to solve the problem of predicting whether a customer will complete a purchase after adding an item to the cart. The company collects various types of data such as user browsing behavior, product views, cart history, past purchase records, time spent on pages, device type, and demographic details. As the data is noisy and comes from multiple sources, Shop-Ease wants a clear pipeline that includes how data should be collected, cleaned, preprocessed, and transformed, what features should be engineered, which machine learning model(s) could be suitable, how the model should be evaluated, and how the final system can be deployed and monitored in a real-time e-commerce environment. Explain the complete machine learning pipeline for this scenario.	**10 Marks**	**L3**	**CO1**
	b.	Demonstrate types of machine learning techniques with respect to EDA with real-time examples?	**10 Marks**	**L3**	**CO1**
		Or			
12.	**a.**	Discoverthe following terms in detail with examples? Also identify on which scenarios each of the data types can be used as a part of data analysis? 1.Nominal data 2.Ordinal data 3.Ratio data 4. Interval data	**10 Marks**	**L3**	**CO1**
	b.	A retail company, "ShopSmart," has collected a large dataset containing customer purchase histories, product details, transaction timestamps, and feedback ratings. The company wants to build a predictive model to forecast customer spending and identify factors influencing high-value customers. Before starting with the modelling process, the data science team decides to perform Exploratory Data Analysis (EDA). As a data analyst, explain in detail why performing EDA is essential in this scenario. Discuss at least five reasons, with suitable justification from the given context.	**10 Marks**	**L3**	**CO1**
13.	**a.**	A fintech company, "FinPay Analytics," wants to build a machine learning model to detect fraudulent transactions. The dataset contains transaction IDs, timestamps, customer demographics, payment modes, transaction amounts, location details, and device information. The raw data also includes missing values,	**10 Marks**	**L3**	**CO2**

		duplicate entries, inconsistent formats, outliers, mixed data types, and noisy text inputs. As a data scientist working on this project, explain the steps involved in data preprocessing that you would perform before building the fraud detection model. Describe each step clearly and justify why it is important in the given scenario.			
	b.	Construct outlier detection in EDA? Interpret python code for Z-Score method and IQR method with appropriate inference.	**10 Marks**	**L3**	**CO2**
		Or			
14.	**a.**	Identify the following terms in detail with appropriate examples? Also list out the advantages & disadvantages of the same. 1.Mean imputation 2.Median imputation 3.Constant imputation 4.KNN imputation 5. Binary imputation	**10 Marks**	**L3**	**CO2**
	b.	Experiment with the python code snippet for the below categorical encoding techniques with appropriate outputs? 1.Label encoding 2.One-hot encoding 3.Ordinal encoding 4. Binary encoding	**10 Marks**	**L3**	**CO2**

15.	**a.**	Identify the main reasons to use data transformation techniques in exploratory data analysis with appropriate examples?	**10 Marks**	**L3**	**CO3**
	b.	Calculate and infer the below given dataset with respect to the following data transformation techniques? 1.Square transformation 2.Square root transformation 3.Cube transformation 4.Log transformation 5. Robust scaling Dataset=[100,200,300,400,500,600,700,800,900,1000,1500,1800,2000,2500,3000]	**10 Marks**	**L3**	**CO3**
		Or			

16.	**a.**	Discover the feature engineering with respect to the given below question? 1.What is feature engineering? 2.Why feature engineering? 3.Key aspects of feature engineering? 4. Explain any 2 key features in detail? 5. Pros and Cons of feature engineering?	**10 Marks**	**L3**	**CO3**
	b.	Illustrate the principal component analysis in detail?	**10 Marks**	**L3**	**CO3**

17.	**a.**	Model the following terms with respect to data visualization? 1.Univariate data visualization 2.Bivariate data visualization 3.Multivariate data visualization 4.Time series data visualization 5.Spatial data visualization	**10 Marks**	**L3**	**CO4**
	b.	Interpret the python code with respect to the matplotlib library and explain pros and cons of each. 1.Scatter plot 2.Boxplot 3.Violin plot 4.Stack plot 5.Pie chart	**10 Marks**	**L3**	**CO4**
		Or			
18.	**a.**	Build any 5 categorical plots in seaborn library? Also provide interpretation for each.	**10 Marks**	**L3**	**CO4**
	b.	Analyze the following terms with respect to data visualization in EDA? Also provide pros and cons of each. 1.Plotly library 2.Seaborn library 3.Matplotlib library	**10 Marks**	**L3**	**CO4**