U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

January / February 2025 Semester End Main Examinations

Programme: B.E. **Semester: V**
Branch: Computer Science and Engineering **Duration: 3 hrs.**
Course Code: 23CS5PCDEV **Max Marks: 100**
Course: Data Exploration and Visualization

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

		UNIT - I	CO	PO	Marks
1	a)	Explicate in brief the steps involved in Exploratory Data Analysis (EDA).	CO2	PO 1	**10**
	b)	With an example, discuss Numerical data and Categorical data types along with its various sub-classification. For the given Student Record, classify the data into Numerical or categorical. Justify your answer. STUDENT_ID = 1001 Name = REYANSH Address = Mannsverk 61, 5094, M G ROAD, BENGALURU Date of birth = 10th July 2018 Email = rey@bmsce.in Weight = 60 Gender = Male	CO1	PO2	**10**
		OR			
2	a)	Compare EDA with classical and Bayesian Analysis.	CO2	PO1	**10**
	b)	List out the different types of measurement scales described in statistics. Explain each of them with a suitable example.	CO2	PO1	**10**
		UNIT - II			
3	a)	You have the following dataset of ages (in years): [25,30,"NaN",40,35,28,"missing",22]	CO1	PO3	**10**

			i. Clean this dataset by replacing the missing values ("NaN", "missing") with the median of the available values. ii. Define data transformation for the above dataset, replace the numerical data(ages) into category such as (Youth, Gentlemen, Senior). Clearly specify the range value considered.			
		b)	What is binning in data transformation? Given data on the heights of a group of students as follows: height = [120, 122, 125, 127, 121, 123, 137, 131, 161, 145, 141, 132], convert that dataset into intervals of 118 to 125, 126 to 135, 136 to 160, and finally 160 and higher. Write the suitable python code snippet and write the possible output for the same of the above operation.	*CO1*	*PO3*	**10**
			OR			
	4	a)	Given the dataset of income values (in Rupees): [45000,52000,48000,51000,60000,52000,"error"] i. Identify and remove the erroneous value ("error") from the dataset, and calculate the average income of the cleaned dataset. ii. Apply Data transformation technique to replace the income values to categorical such as [Fresher, Experienced, HighNetWorth] Write the corresponding Python code snippet.	*CO1*	*PO3*	**10**
		b)	Demonstrate with suitable examples, how does skewness help in understanding the distribution of data, and how can positive or negative skewness be interpreted to identify potential outliers in a dataset?	*CO2*	*PO2*	**10**
			UNIT - III			
	5	a)	Illustrate cross-tabulation and Pivot table with suitable example with appropriate code snippet.	*CO1*	*PO1*	**10**
		b)	Discuss the concept of linear interpolation and how it is applied to fill missing values in a time-series dataset with an example. Examine the potential impact of outliers on the effectiveness of linear interpolation. How might extreme values influence the interpolated results?	*CO1*	*PO1*	**10**
			OR			
	6	a)	Give a case study on univariate and multivariate analysis with example.	*CO1*	*PO4*	**10**
		b)	What is central tendency and Dispersion? For the given data set, apply central tendency measures (mean, median and mode) and	*CO1*	*PO3*	**10**

			also apply Dispersion measures (Range, Variance, Standard Deviation). Dataset of daily temperatures recorded over a week: 22°C, 23°C, 21°C, 25°C, 22°C, 24°C, and 20°C.			
			UNIT - IV			
	7	a)	A Pandas DataFrame contains the exam scores of students in three subjects: Math, Science, and English. Give Python code snippet to create: i. A histogram to visualize the distribution of Math scores? ii. A density plot (kernel density estimate) to visualize the distribution of Science scores?	*CO3*	*PO3*	**10**
		b)	Interpret the purpose of a scatter plot matrix (pair plot) and a correlation matrix heatmap. How do these visualizations help in understanding the relationships between multiple quantitative variables? Write python code snippet to draw plot matrix and heatmap.	*CO3*	*PO3*	**10**
			OR			
	8	a)	Illustrate how data can be mapped onto aesthetics, scales, and coordinate systems in data visualization. Provide an example using a scatter plot where the x-axis represents a numerical variable, the y-axis represents another numerical variable, and the color represents a categorical variable.	*CO3*	*PO3*	**10**
		b)	Demonstrate how visualizations like error bars and stacked bar charts can be used to represent data uncertainty and proportions, respectively. What insights do these visualizations provide?	*CO2*	*PO4*	**10**
			UNIT - V			
	9	a)	Case study on Data wrangling: Given with a sales dataset that contains information on transactions made over the last year. The dataset has columns such as TransactionID, ProductCategory, Price, Quantity, and DateOfSale. It is noticed that some missing values in the Price and Quantity columns are seen, with some rows where both are missing, as well as a few duplicates based on TransactionID. The goal is to clean the data for a time-series analysis on monthly sales trends. How can these missing values and duplicates handled efficiently? List out the steps you would apply for the above goal. Give corresponding python code snippet.	*CO3*	*PO4*	**10**
		b)	Demonstrate using a Python code snippet to scrape the titles and URLs of all articles from the homepage of a news website.	*CO3*	*PO3*	**10**

			OR			
	10	a)	Demonstrate string manipulation in Pandas for string replacement and combining of strings on a Data frame and columns.	*CO3*	*PO3*	**10**
		b)	A Pandas DataFrame with monthly sales data for different products. i. Create a DataFrame which contains montly sales of three products in three months viz Jan, Feb and March. ii. Compute the **total sales** for each product across all months using vectorized operations? iii. Calculate the **average sales** for each product across all months using vectorized operations in Pandas?	*CO3*	*PO3*	**10**

****