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: 22CS5PEDEV **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)	Suppose you have two datasets related to employee information: employee_data_1 contains information about employees' basic details like Employee_ID, Name, and Department. employee_data_2 contains additional details such as Employee_ID, Name,Salary, and Joining_Date. Write a python code for the following tasks i. Combine both datasets vertically to create a comprehensive employee dataset. ii. Concatenate the employee data along columns for a more detailed analysis. iii. Merge the datasets based on a common key (Employee_ID) to enrich the employee information. iv. Join the datasets based on the employee names to analyze employee details by names.	*CO2*	*PO2*	**10**
	b)	Product Reviews Analysis: You collected product reviews where customers rated the products on a scale of "Poor," "Average," and "Excellent." Analyze the reviews using the ordinal scale. What insights can you draw about the overall satisfaction of customers with different products? Distance Traveled: You recorded the daily distances traveled by individuals for a fitness challenge. Analyze the data using the ratio scale. How does the presence of a true zero point impact the interpretation of the distances covered by participants? Job Satisfaction Levels: Employees in a company provided job satisfaction ratings on a scale from 1 to 7. Analyze the ratings and discuss the limitations of using a ratio scale versus an ordinal scale for measuring job satisfaction.	*CO2*	***PO2***	**10**

			Social Media Likes: Compare the number of likes received by posts on two different social media platforms. Analyze the data using the ratio scale and discuss how the true zero point influences the interpretation of likes. Customer Loyalty Ratings: Customers provided loyalty ratings to a company on a scale from "Not at all likely" to "Extremely likely." Analyze the data using the ordinal scale and discuss the challenges of quantifying the exact differences in loyalty levels between different responses.			
			OR			
	2	a)	Explain the significance of Exploratory Data Analysis (EDA) in the data science process. Provide examples of how EDA can uncover patterns and outliers in a dataset	*CO1*	*PO1*	**10**
		b)	Discuss the four types of measurement scales (Nominal, Ordinal, Interval, and Ratio) with examples. Highlight the key differences between them.	*CO1*	*PO1*	**10**
			UNIT - II			

	3	a)	Sample sales dataset with missing values in the 'Sales' column
			data_sales = {
			'Product_ID': [101, 102, 103, 104, 105],
			'Region': ['North', 'South', np.nan, 'East', 'West'],
			'Sales': [500, 600, np.nan, 700, 800],
			}
			Write a Python code for each of the following task:
			1. How many missing values are there in the 'Sales' column?
			2. What is the total count of missing values in the entire dataset?
			3. identify and display rows where 'Sales' values are missing
			4. How many rows have complete data (non-missing) in the 'Sales' column?
			5. What is the sum of missing values across each column in the dataset?
				CO3	*PO3*	**10**

		b)	Consider the dataset containing the ages of individuals, and you want to categorize them into age groups. Implement binning to discretize the 'Age' column into three equal-width bins. Consider the following sample dataset data = {'Age': [22, 35, 42, 28, 55, 20, 38, 45, 30, 50]}	*CO2*	*PO2*	**5**
		c)	Consider the following dataset: data_customers = { 'Customer_ID': [101, 102, 103, 104, 101, 105, 106, 102, 107, 108], 'Name': ['Alice', 'Bob', 'Charlie', 'David', 'Alice', 'Eva', 'Frank', 'Bob', 'Grace', 'Henry'], 'Email': ['alice@example.com', 'bob@example.com', 'charlie@example.com', 'david@example.com', 'alice@example.com', 'eva@example.com', 'frank@example.com', 'bob@example.com',	*CO3*	*PO3*	**5**

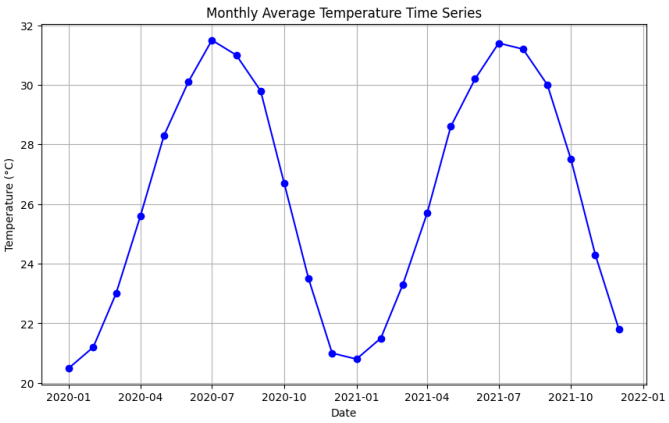
			'grace@example.com', 'henry@example.com'], 'Phone': ['123-456-7890', '234-567-8901', '345-678-9012', '456-789-0123','123-456-7890', '567-890-1234', '678-901-2345', '234-567-8901','789-012-3456', '890-123-4567'] } Write a python code to display original data and duplicated data.			
			OR			
	4	a)	Explain the concept of data deduplication in data preprocessing. How can you identify and remove duplicate records from a dataset? Provide a Python example using pandas to demonstrate how to identify and remove duplicates based on specific columns.	*CO1*	*PO1*	**10**
		b)	Explain the concepts of discretization and binning in data preprocessing. How are continuous variables transformed into categorical variables using binning? Provide an example in Python to perform equal-width binning and equal-frequency binning.	*CO1*	*PO1*	**10**
			UNIT - III			
	5	a)	Describe the various forms of skewness in data using clear diagrams and explore the effects of negative skewness on the distribution's shape and the mean's position. Discuss the implications of these effects on the interpretation of the data.	*CO1*	*PO1*	**10**

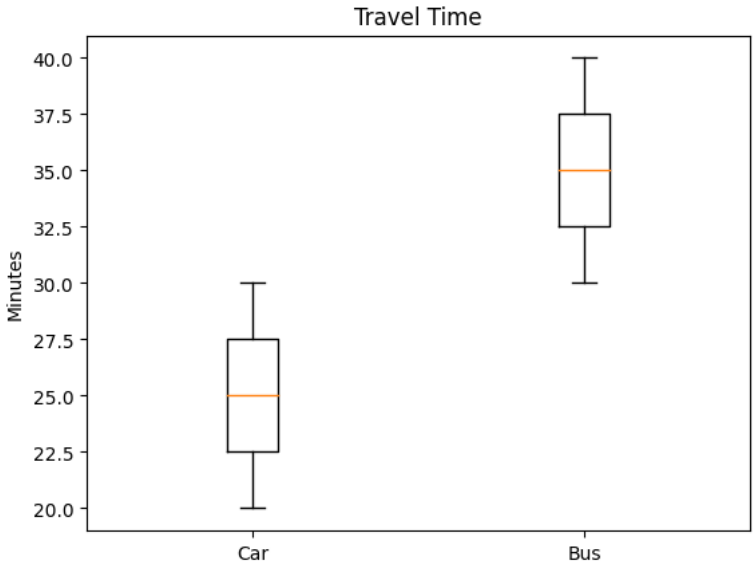
		b)	Consider the dataset with following information about students including their grades in different subjects. Write a python code using groupby function to calculate the average grade for each student.
			
			data_students = {
			
			'Student_ID': [1, 2, 1, 2, 3],
			'Subject': ['Math', 'Math', 'Physics', 'Physics', 'Math'],
			'Grade': [90, 85, 75, 80, 95]
			}
				CO2	*PO2*	**5**

		c)	Consider the dataset with information about employees, including their departments and salaries. Write a python code and use the group by function to aggregate data and find the total number of employees and the average salary for each department.	*CO3*	*PO3*	**5**
			OR			
	6	a)	Explain the importance of statistics in data analysis. Discuss the three key measures of central tendency (mean, median, and mode) and explain when each measure is most appropriate to use. Provide examples where each measure would be preferred.	*CO1*	*PO1*	**10**
		b)	What are pivot tables and crosstabulations in pandas? Discuss their uses in summarizing and analyzing data. Provide a Python example demonstrating how to create a pivot table and a crosstab, and explain the differences between the two	*CO1*	*PO1*	**10**

			UNIT - IV			
	7	a)	Define time series data and explain its characteristics in detail.	*CO1*	*PO1*	**10**

		b)	# Sample data
			data = {
			'Date': ['2022-01-01', '2022-01-01', '2022-01-02', '2022-01-02', '2022-01-03', '2022-01-03'],
			'Category': ['A', 'B', 'A', 'B', 'A', 'B'],
			'Value': [10, 15, 20, 25, 30, 35]
			}
			# Create DataFrame
			df = pd.DataFrame(data)
			# Convert 'Date' column to datetime format
			df['Date'] = pd.to_datetime(df['Date'])
			
			# Set 'Date' column as index
			df.set_index('Date', inplace=True)
			
			1. Complete the code to Group by 'Category' and resample by day and aggregating with sum.
			
			2. print(resampled_data).
			
			What will be the output of the above statement in the code assuming resampled data is stored in a variable ”resampled_data.”
				CO3	*PO3*	**10**

			OR			
	8	a)	 1. Is there a noticeable trend in the monthly average temperature over the two-year period? 2. Are there any recurring seasonal patterns in the temperature data? If so, what are the peak seasons? 3. Is there a correlation between temperature and month of the year?In which months do we typically observe higher temperatures? 4. How does the temperature data for 2021 compare to that of 2020? Are there any noticeable differences in temperature patterns between the two years? 5. Are there any abnormal fluctuations in temperature that might indicate anomalies or unusual events?	*CO3*	*PO3*	**10**

		b)	Discuss univariate and multivariate analysis in detail.	*CO1*	*PO1*	**10**
			UNIT - V			
	9	a)	Explain how scales map data values onto aesthetics with an example.	*CO1*	*PO1*	**10**
		b)	With an example describe each of the following: i. Barplots ii. Grouped and Stacked Bar iii. Dots and Heat Maps iv. Histograms	*CO1*	*PO1*	**10**
			OR			
	10	a)	Analyze the following boxplot and answer the questions: • Which mode of transportation has a higher median travel time? • Calculate the interquartile range (IQR) for both Car and Bus. • Identify if there are any outliers in the travel time of the Bus. • Compare the variability in travel time between Car and Bus 	*CO2*	*PO2*	**10**
		b)	What are the different methods to handle overlapping points in data visualization?	*CO1*	*PO1*	**10**

****