U.S.N.										

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

Autonomous Institute Affiliated to VTU

April 2025 Semester End Make-Up 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)		Discuss the generic steps involved in Exploratory data analysis (EDA) process.	CO1	PO 1	**06**
	b)		Imagine you are a data analyst working for an educational institute. You have been given a dataset containing information about students' performance in various subjects. The dataset includes the following columns: Student ID, Name, Gender, Age, Grade: The grade level of the student (e.g., 10th, 11th), Math Score, Science Score, English Score, Participation in Extracurricular Activities: Indicates if the student participates in extracurricular activities (e.g., Yes, No)., Parent's Education Level. Classify the types of data in the given dataset and identify their measurement scales.	CO2	PO2	**06**
	c)		Demonstrate any four types of analysis that could be done for each of the following domains using EDA techniques. Justify your answer. i. Professional Sports ii. Healthcare	CO2	PO2	**08**
			OR			
2	a)		Define Exploratory data analysis. Point out the primary aim and explain the significance of the same.	CO1	PO1	**06**
	b)		Demonstrate any three types of analysis that could be done for Marketing domain using EDA techniques. Justify your answer.	CO2	PO2	**06**
	c)		You have been provided with a dataset from a health and fitness app that tracks users' activities, dietary habits, and other lifestyle factors. The goal is to help the company understand different types of data, including their measurement scales, to improve the design of the app and offer personalized recommendations. The dataset includes the following variables: • User Demographics: Age, Gender, Height, Weight, City.	CO2	PO2	**08**

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

<table>
<tr><td></td><td></td><td></td><td>

Physical Activity Data: Daily Step Count, Hours of Sleep, Calories Burned, Exercise Frequency.
Dietary Data: Daily Water Intake, Daily Caloric Intake, Food Categories.
Health Metrics: Blood Pressure, Cholesterol Level, Blood Sugar Level.
User Feedback: Overall Health Rating, Exercise Enjoyment Rating, Sleep Quality Rating

Classify the types of data in the given dataset and identify their measurement scales.
</td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td>UNIT - II</td><td></td><td></td><td></td></tr>
<tr><td>3</td><td>a)</td><td>Discuss the process of data deduplication and the steps involved in the same with an example program.</td><td>CO1</td><td>PO1</td><td>06</td></tr>
<tr><td></td><td>b)</td><td>Write a Python program to perform the append and concatenation (with axis=1) operations on the following dataframes. Show the appropriate output.
data1 = { 'Name': ['Alice', 'Bob', 'Charlie'], 'Age': [25, 30, 35] }
data2 = { 'Name': ['David', 'Eva'], 'Age': [40, 45] }.</td><td>CO1</td><td>PO1</td><td>06</td></tr>
<tr><td></td><td>c)</td><td>A retail company has been collecting data from its sales transactions, which includes the following information.
<ol type="a">
TansactionID: Unique identifier for each transaction.
CustomerID: Identifier for the customer who made the purchase.
ProductID: Identifier for the product purchased.
Quantity: Quantity of the product purchased in the transaction.
Price: Price per unit of the product.
TotalAmount: Total amount for the transaction (Quantity * Price).

However, some of the records in the dataset are duplicates as shown below due to system errors, such as customers placing multiple orders for the same product, or the same transaction being recorded more than once.
<table>
<tr><th></th><th>TransactionID</th><th>CustomerID</th><th>ProductID</th><th>Quantity</th><th>Price</th><th>TotalAmount</th></tr>
<tr><td>0</td><td>1</td><td>101</td><td>1001</td><td>1</td><td>20</td><td>20</td></tr>
<tr><td>1</td><td>2</td><td>102</td><td>1002</td><td>2</td><td>15</td><td>30</td></tr>
<tr><td>2</td><td>3</td><td>101</td><td>1001</td><td>1</td><td>20</td><td>20</td></tr>
<tr><td>3</td><td>4</td><td>104</td><td>1003</td><td>5</td><td>30</td><td>150</td></tr>
<tr><td>4</td><td>5</td><td>103</td><td>1002</td><td>2</td><td>15</td><td>30</td></tr>
<tr><td>5</td><td>6</td><td>102</td><td>1002</td><td>3</td><td>15</td><td>45</td></tr>
<tr><td>6</td><td>7</td><td>101</td><td>1001</td><td>1</td><td>20</td><td>20</td></tr>
</table>
Write a Python program to do the following and show subsequent output after each step.
<ol type="a">
Identify duplicate rows.
Removing Duplicates
Instead of simply removing the duplicates, replace duplicate value by aggregating data such as summing up the quantities and total amount and average the prices for the duplicate records.

</td><td>CO2</td><td>PO2</td><td>08</td></tr>
</table>

			OR			
	4	a)	Discuss the importance of random sampling of a dataset. Differentiate between random sampling with replacement and without replacement.	*CO1*	*PO1*	**06**
		b)	A retail company has collected data on the ages of its customers to analyze purchasing behavior across different age groups. To perform a more meaningful analysis, the company wants to classify the customers into distinct age groups (bins) rather than working with individual age values. The dataset contains the following columns: • CustomerID: Unique identifier for each customer. • Age: Age of the customer. • Spending Amount: Total amount spent by the customer in the last month. Write a Python program to categorize customers into predefined age groups such as "18-25", "26-35", "36-45", etc. and find the average of their spending behavior within each group.	*CO2*	*PO2*	**06**
		c)	A company collects data from two different sources: one containing information about its employees, and another containing the details of the projects that each employee is working on. DataFrame- df_employees DataFrame- df_projects These two data sources need to be merged to gain a complete picture of each employee and their associated projects. Analyse the types of joins that needs to performed to do the following. Write necessary code along with suitable output. 1. If an employee doesn't have a corresponding project, they will be excluded from the result. 2. A join that returns all rows from the df_employees and the matching rows from the right df_projects. If there is no match, the result will contain NaN for columns from the right DataFrame.	*CO2*	*PO2*	**08**

	EmployeeID	Name	Department
0	1	Alice	HR
1	2	Bob	Finance
2	3	Charlie	IT
3	4	David	Sales
4	5	Eva	Marketing

	EmployeeID	ProjectID	ProjectName
0	1	101	Project A
1	2	102	Project B
2	4	103	Project C
3	5	104	Project D
4	6	105	Project E

			3. A join that returns all rows from the df_projects and the matching rows from df_employees. If there is no match, the result will contain NaN for columns from the left DataFrame. 4. A join that returns all rows from both DataFrames, with matching records from both sides when available. If there is no match, NaN is filled in for the columns from the DataFrame without a match.			
			UNIT - III			
	5	a)	Explain the most common measures for analyzing the distribution frequency. Demonstrate the same with a Python program.	*CO1*	*PO1*	**06**
		b)	You are a data analyst working for a retail company that tracks sales data across different regions and products. You have a dataset containing information about sales figures and profit margins for various products in different regions. **Data set**: 'Region': ['North', 'North', 'South', 'South', 'North', 'South'], 'Product': ['A', 'B', 'A', 'B', 'A', 'B'], 'Sales': [100, 150, 200, 120, 180, 90], 'Profit_Margin': [0.2, 0.3, 0.25, 0.15, 0.18, 0.12] Write a program to calculate the total sales and maximum profit margin for each product across all regions.	*CO2*	*PO2*	**06**
		c)	Write a Python program to do the following. i.Create a dataframe for the following data. 1. Date: The date of the transaction or record. 2. Category: The category of the product sold (e.g., Electronics, Clothing, Food). 3. Region: The region where the sale took place (e.g., North, South, East, West). 4. Sales: The total sales amount for each transaction. 5. Units Sold: The number of units sold in the transaction ii.Create a pivot table that shows the total Sales by Category and Region. iii.Create a cross-tabulation that shows the Units Sold by Category and Region. iv.Print the average sales per unit sold for each product category across different regions. v.Print the region that has the highest sales for the Electronics category.	*CO2*	*PO2*	**08**
			OR			
	6	a)	Illustrate the various characteristics of the time series dataset with suitable example.	*CO1*	*PO1*	**06**
		b)	As a data analyst for a retail company that wants to understand the relationship between the amount spent by customers and the time they spend on the company's website. This information will aid in optimizing the website experience and marketing strategies. 1. What type of analysis technique can be applied?	*CO2*	*PO2*	**06**

			2. Write a panda's code to explore the relationship between the time spent on the website (in minutes) and the amount spent by customers (in dollars) to derive insights for the retail company.			
		c)	We have a dataset with information about employees in an organization. The dataset includes **Employee ID**, **Department**, **Training Hours** (hours spent in training during the last quarter), **Job Performance Score** (a composite score based on efficiency, quality of work, and job satisfaction), and **Job Satisfaction** (on a scale from 1 to 10). Employee ID | Department | Training Hours | Job Performance Scores | Job Satisfaction 101 | Sales | 40 | 85 | 8 102 | Marketing | 50 | 90 | 9 103 | HR | 20 | 75 | 7 104 | IT | 60 | 92 | 9 a. Write a Python program to do the following. i. Create a data frame for the data shown. ii. Find the correlation between Training Hours and Job Performance Score for each department. iii. Find the overall correlation between Training Hours and Job Performance Score across all departments. b. Interpret the output for the given queries. i. Find the correlation between Training Hours and Job Performance Score for each department. ii. Find the overall correlation between Training Hours and Job Performance Score across all departments.	*CO2*	*PO2*	**08**
			UNIT - IV			
	7	a)	Illustrate the linear scale and logarithmic scale with respect to axes representation. Explain how the following data points could be represented using both the scales. 1,3.16,10,31.6,100	*CO1*	*PO1*	**06**

		b)	Consider the following monthly sales dataset for 4 regions (North, South, East, West) across 6 months (January to June). The dataset contains the total sales in dollars for each region each month. Identify the various plots and charts that are could be used to visualize the given data under following circumstances. Justify your answer. i. To compare the total sales in each region across the months. ii. To show individual data points for each month's sales across regions. iii. To visualize the sales performance across regions and months in a more holistic way, where colour intensity represents sales volume.	*CO2*	*PO2*	**06**
		c)	Explore the various plots and charts used to visualize associations among two or more quantitative variables. Analyze the visualization strategies that could be adopted to address the following relationships. i. Relationship of two variables. ii. Same types of variables with two different types of scales, position and size. iii. More than three to four quantitative variables.	*CO2*	*PO2*	**08**
			OR			
	8	a)	List the limitations of Histogram and Density Plot. Explain how these limitations are handled using Empirical Cumulative Distribution Functions with an example.	*1*	*1*	**06**
		b)	Consider a business scenario where a company wants to analyze the revenue composition across different product categories. Propose a specific visualization method that breaks down revenue proportions, highlighting the parts of the total. Justify your choice and explain how this visualization could guide decision-making within the company.	*CO2*	*PO2*	**06**
		c)	Describe how confidence intervals are used to convey uncertainty in curve fits. What information do confidence intervals provide, and how can they be visually integrated into a curve fit visualization?	*1*	*1*	**08**
			UNIT - V			
	9	a)	Write a Python program to sort an array of words from longest to shortest length.	*1*	*1*	**06**
		b)	Consider the following data set.	*CO2*	*PO2*	**06**

Month	North	South	East	West
January	50,000	55,000	40,000	45,000
February	60,000	50,000	45,000	50,000
March	55,000	52,000	47,000	55,000
April	70,000	65,000	60,000	65,000
May	80,000	75,000	70,000	72,000
June	85,000	80,000	78,000	78,000

			data = { 'Name': ['Alice', 'Bob', 'Charlie', 'David', 'Eve'], 'Age': [24, 27, 22, 32, 29], 'City': ['New York', 'Los Angeles', 'Chicago', 'Houston'], 'Score': [85, 90, 78, 88, 92] } Write python code by creating a dataframe of the above dataset to do the following. i. Access the first name ii. Select Name and City from first three rows. iii. Print rows where age>25			
		c)	Write a Python program that demonstrates how to create an Employee database with ID, Name, Age, Department as attributes using SQLite, insert data into it and then transfer the data into a Pandas dataframe.	*CO2*	*PO2*	**08**
			OR			
	10	a)	Write a Python program demonstrating how to reshape the data using melt() and pivot() methods. Illustrate the purpose of the same.	*CO1*	*PO1*	**06**
		b)	Write a Python program demonstrating the use of Web APIs and processing the retrieved data using pandas.	*CO2*	*PO2*	**06**
		c)	Write a Python program to create a Student dataframe with ID, Name, Age, Grade and Score. Find the missing values if any and fill the same using mean, mode and median for whichever data applicable. Justify your choice of filling the missing values.	*CO2*	*PO2*	**08**

****