U.S.N.										

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

Autonomous Institute Affiliated to VTU

July 2024 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

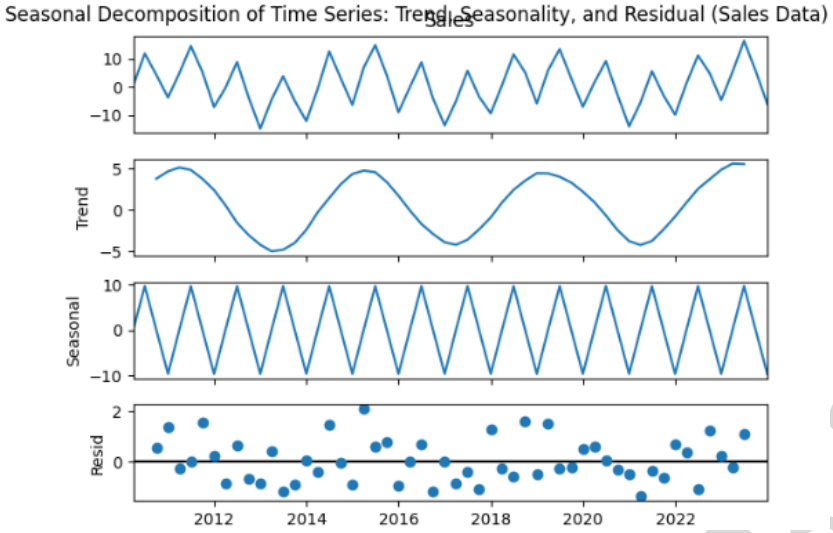
Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

		UNIT - I	CO	PO	Marks
1	b)	Detail the stages integral to Exploratory Data Analysis (EDA) and elucidate the significance of these steps in the context of data analysis.	CO1	PO1	**10**
	b)	Consider "The Employee Performance Evaluation Dataset," which includes information about employees in a company. Employee ID - Unique identifier for each employee Department - The department in which the employee works Years of Experience - Number of years of experience the employee has Performance Rating - Performance rating given to the employee (on a scale of 1 to 5) Salary (USD) - The salary of the employee Educational Qualification - Highest level of education achieved by the employee Project Assignments - Number of projects currently assigned to the employee Work Hours - Average number of work hours per week Identify and explain the types of data (Quantitative, Qualitative, and Scale) present in this dataset and suggest possible operations on these data.	CO2	PO2	**10**
		UNIT - II			
2	a)	A dataset containing information about students' performance, but the dataset has missing values across multiple columns due to various reasons. Your task is to analyze and handle these missing values using Pandas' methods for the below questions. 1. How many missing values are there in the 'Grade' column of the dataset? 2. What is the total count of missing values in the entire dataset?	CO2	PO2	**10**

			3. Can you identify and display rows where 'Absenteeism' values are missing?
			4. How many rows have complete data (non-missing) in the 'Study Hours' column?
			5. What is the sum of missing values across each column in the dataset?
			`# Sample student performance data with missing values`
			`data = {`
			`'Student_ID': [1, 2, 3, 4, 5],`
			`'Name': ['Alice', 'Bob', 'Charlie', 'David', 'Eva'],`
			`'Grade': ['A', 'B', np.nan, 'C', 'B'],`
			`'Absenteeism': [2, np.nan, 5, np.nan, 1],`
			`'Study_Hours': [4, 6, np.nan, 3, 5]`
			`}`			

	b)	Suppose you have two datasets related to employee information: employee_data_basic containing basic details such as employee ID, name, and department. employee_data_additional providing additional information like salary and job title. Write a python code using Panda's library to perform following tasks: 1. Combine both datasets vertically to create a comprehensive employee dataset. 2. Concatenate the employee data along columns for a more detailed analysis. 3. Merge the datasets based on a common key (employee ID) to enrich the employee information. 4. Join the datasets based on the employee names to analyze employee details by names.	*CO2*	*PO2*	**10**
		UNIT – III			
3	a)	Imagine you are an HR analyst at a mid-sized company, and you have been tasked with analyzing the salaries of the employees to understand the company's payroll distribution. You have a dataset containing the salaries of 10 employees: $40,000, $45,000, $50,000, $55,000, $60,000, $65,000, $70,000, $75,000, $80,000, $85,000 1. Calculate Measures of Central Tendency: Calculate the mean, median, and mode for the employee salaries. Determine which measure best represents the typical salary for these employees. 2. Interpretation and Recommendation: Based on the calculated measures of central tendency, provide an interpretation of the salary distribution. Which measure would you recommend using to describe the typical salary, and why?	*CO2*	*PO2*	**5**
	b)	An e-commerce platform that sells various products across different categories. A dataset containing information about purchases made on your platform, including product categories, purchase amounts, and customer IDs.	*CO3*	*PO3*	**5**

			Dataset Example: 'Customer_ID': [1, 2, 3, 1, 2, 3, 1, 2, 3], 'Product_Category': ['A', 'B', 'A', 'B', 'A', 'B', 'A', 'B', 'A'], 'Purchase_Amount': [100, 150, 200, 120, 180, 90, 210, 160, 80] **Write Python code (Pandas Library to:** 1. Calculate the average purchase amount per product category. 2. Identify the top-spending customer in each product category based on their total purchase amount.			
	c)	Examine the concept of kurtosis with a clear explanation of each type, supported by illustrative diagrams. Discuss how these different forms of kurtosis manifest in the shape of distributions and the implications they carry for statistical analysis.	*CO1*	*PO1*	**10**
		UNIT - IV			
4	a)	In the context of data analysis, explore the different types of analysis that can be applied to gain insights from datasets. Provide a comprehensive overview and examples of each type.	*CO1*	*PO1*	**10**
	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? 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.	*CO3*	*PO3*	**5**
	c)	Analyze the below given python code and mention the types of analysis and interpret output of the plot. sns.pairplot(df,vars = ['normalized-losses', 'price','horsepower'], kind="reg") plt.show()	*CO2*	*PO2*	**5**
		OR			
5	a)	Illustrate the various characteristics of the time series dataset with suitable example.	*CO2*	*PO2*	**10**
	b)	A dataset representing quarterly sales data for a retail business over the past decade. Your objective is to analyze the time series data and identify the trend, seasonal pattern, and any potential cyclical behavior. Explain below with panda's code. 1. Load the dataset and convert the "Date" column to the datetime format. 2. Set the "Date" column as the index of the DataFrame. 3. Plot the time series to visualize the overall pattern in the sales data. 4. Apply a suitable method (e.g., moving averages or seasonal decomposition) to identify and separate the trend, seasonal, and residual components.	*CO3*	*PO3*	**10**

			5. Interpret the identified trend, seasonal pattern, and any observed cyclical behavior. 			
			UNIT - V			
	6	a)	Differentiate between aesthetics that can represent both continuous and discrete data. Provide examples for each category.	*CO1*	*PO1*	**5**
		b)	Consider a dataset representing the monthly average temperatures in a location over several years. Explain how polar coordinates could be employed to visualize the temperature variations throughout the year. Specify the advantages and potential drawbacks of using polar coordinates in this context.	*CO2*	*PO2*	**5**
		c)	Explore the potential challenges of representing large datasets using a grouped bar plot. Discuss strategies to address issues related to overcrowding, especially when dealing with numerous categories or groups. Consider the balance between displaying detailed information and maintaining a clear and uncluttered visualization.	*CO1*	*PO1*	**10**
			OR			
	7	a)	Describe how Empirical Cumulative Distribution Function (ECDF) can be used to compare distributions of two or more datasets. What visual cues from the ECDF plot can indicate similarities or differences between distributions?	*CO1*	*PO1*	**5**
		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*	**5**
		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?	*CO2*	*PO2*	**10**

****