|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Roll No. |  |  |  |  |  |  |  |  |  |  |  |  |



 **PRESIDENCY UNIVERSITY**

  **Bengaluru**

|  |
| --- |
| **End - Term Examinations – JANUARY 2025** |
| **Date:** 17 – 01- 2025 **Time:** 01:00 pm – 04:00 pm |

|  |  |
| --- | --- |
| **School:** SOCSE | **Program:** B. Tech-(CCS/CIT/CSG/CST/CDV/CBC/CSD/CBD/CSN/CSE/IST/CSI/IST-(AI &DS) |
| **Course Code:** CSE3190 | **Course Name:** Fundamentals of Data Analytics |
| **Semester**: III | **Max Marks**: 100 | **Weightage**: 50% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CO - Levels** | **CO1** | **CO2** | **CO3** | **CO4** | **CO5** |
| **Marks** | **16** |  **41** | **19** | **24** | **00** |

**Instructions:**

1. *Read all questions carefully and answer accordingly.*
2. *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** | What is Data analysis? And what is the purpose of Data Analysis. | **2M** | **L1** | **CO1** |
| **2** | List the different types of digital data with examples. | **2M** | **L1** | **CO1** |
| **3** | List the following variables based on their scale of measurement: Movie ratings, Political parties, Profit, Time. | **2M** | **L1** | **CO1** |
| **4** | Which are some common causes of Missing Data. | **2M** | **L1** | **CO2** |
| **5** | What are Data Classes? List the different One-dimensional data classes. | **2M** | **L1** | **CO2** |
| **6** | List the key Data Manipulation Techniques in R | **2M** | **L1** | **CO2** |
| **7** | What is a Proportion Test, and list some Application of proportion test. | **2M** | **L1** | **CO3** |
| **8** | What is a Regression Line? | **2M** | **L1** | **CO3** |
| **9** | Which is the equation which defines Line f Best Fit in Analysis. | **2M** | **L1** | **CO4** |
| **10** | Define the different types of linear regression. | **2M** | **L1** | **CO4** |

 **Part B**

|  |
| --- |
| **Answer the Questions Total 80 Marks** |
| **11.** | **a.** | Explain the different V’s of data with examples. | **10M**  | **L2** | **CO1** |
| **or** |
| **12.** | **a.** | Identify the different data sources and recognize the importance of data cleaning. | **10M** | **L3** | **CO1** |
|  |  |  |  |  |  |
| **13.** | **a.** | Demonstrate an R script to create a function named calculator that takes two numbers and an operator as input. The function should perform the specified operation and return the result. The valid operators should be "+", "-", "\*", and "sqrt". | **10M** | **L2** | **CO2** |
| **or** |
| **14.** | **a.** | Student Performance Analysis: The school administration is concerned about the declining overall academic performance of students. They want to analyze student data to identify areas for improvement and implement targeted interventions.Explain how the four types of data analysis will be used to analyze school performance. | **10M** | **L2** | **CO2** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **15.** | **a.** | Explain Recoding in R, also explaini. what is the Purpose of Recodingii. Use cases of Recodingiii. List the different functions in R for recoding with an example. | **10M** | **L2** | **CO2** |
| **Or** |
| **16.** | **a.** | **Summarize Data manipulation in R,** also explaini. Different Data Manipulation Techniques.ii. List the different functions used in Reshaping and merging in R | **10M** | **L2** | **CO2** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **17.** | **a.** | Summarize different types of Missing Data?i. List the different functions used in identifying missing data.ii. List the different steps in handling missing values in R. | **15M** | **L2** | **CO2** |
| **Or** |
| **18.** | **a.** | A survey was conducted to **Identify** if more than 70% of employees in a company prefer remote work. Out of 200 employees surveyed, 150 preferred remote work.1. **One-Sample Proportion Test**:
	* Using a significance level of 0.05, conduct a one-sample proportion test to determine if the proportion of employees preferring remote work is significantly different from 70%.
	* The critical value for the Z-test at α = 0.05 is ±1.96. If the p-value obtained is 0.015, interpret the p-value in the context of this test.
2. **Correlation Test**:
	* The company also collected data on the number of years employees have been with the company and their productivity scores. They want to determine if there is a significant correlation between years of service and productivity scores.
	* If the correlation coefficient is 0.45 and the p-value obtained is 0.005, interpret the results in the context of this test.
 | **15M** | **L3** | **CO2** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **19.** | **a.** | A researcher wants to **Experiment** whether there are significant differences in the effectiveness of three different treatments on patients' recovery times. The researcher collects recovery time data for the three treatments:* **Treatment A**: [5, 6, 7, 8]
* **Treatment B**: [7, 9, 10, 11]
* **Treatment C**: [4, 5, 6, 5]

To analyze the data, the researcher decides to use both the Wilcoxon Rank Sum Test and One-Way ANOVA.1. **Wilcoxon Rank Sum Test**:
	* How can the researcher use the Wilcoxon Rank Sum Test to determine if there is a significant difference in recovery times between Treatment A and Treatment B?
	* The critical value for the Wilcoxon Rank Sum Test at α = 0.05 is 8. Compare the test statistic to the critical value and state whether to reject or fail to reject the null hypothesis.
	* If the p-value obtained from the Wilcoxon Rank Sum Test is 0.04, interpret the p-value in the context of this study.
2. **One-Way ANOVA**:
	* How can the researcher use One-Way ANOVA to determine if there are statistically significant differences in recovery times among all three treatments?
	* Calculate the F-statistic and compare it to the critical value. If the critical value for F(2, 9) at α = 0.05 is 4.26, compare the F-statistic to the critical value and state whether to reject or fail to reject the null hypothesis.
	* If the p-value obtained from the One-Way ANOVA is 0.03, interpret the p-value in the context of this study.
 | **15M** | **L3** | **CO3** |
| **Or** |
| **20.** | **a.** | Explain the different Evaluation metrics for regression model in detail with their advantages and disadvantages**.** | **15M** | **L2** | **CO3** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **21.** | **a.** | Develop an R program to analyze a dataset containing information on 10 students, where each student has the following attributes:• Student: A numeric identifier for each student (1 to 10).• Hours\_Studied: The number of hours each student studied (ranging from 2 to 10).• Exam\_Score: The exam score obtained by each student (ranging from 42 to 95).Perform the following tasks:1. Calculate the mean, median, and standard deviation for both Hours\_Studied and Exam\_Score.2. Create the following visualizations using ggplot2:• Scatter plot of Hours\_Studied vs Exam\_Score with a regression line.• Histogram for Hours\_Studied.• Histogram for Exam\_Score. | **20M** | **L3** | **CO4** |
| **Or** |
| **22.** | **a.** | **Develop** an R program to analyze temperature data for six cities (New York, Los Angeles, Chicago, Houston, Phoenix, Philadelphia) across four seasons (Winter, Spring, Summer, Fall). The dataset contains 600 temperature observations. Perform the following tasks:1. Generate the dataset with:• Cities repeated 100 times each.•Seasons distributed equally across the dataset.•Temperature values randomly generated with a mean of 60 and a standard deviation of 102.Create the following visualizations using ggplot2:•A line plot showing the temperature trend for each city across seasons.•A bar chart showing the temperature for each city.* A box plot comparing the temperatures across cities.
 | **20M** | **L3** | **CO4** |

**\*\*\*\*\* BEST WISHES \*\*\*\*\***