W/D/TY

VIT
Vellore Institute of Technology

Final Assessment Test – April 2025

Course: **BCSE207L - Programming for Data Science**

Class NBR(s): **2039/2047/2056/2060/2067/2079** Slot: **B1**

Time: **Three Hours** Max. Marks: **100**

- **KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

Answer ALL Questions

(10 X 10 = 100 Marks)

1. Complete the R code to define variables for a student's name and score, calculate the average score using a user-defined function, check pass/fail status based on a score above 50, and find the highest score using a built-in function. Include error handling, user input, and comments to explain the code.

2.a) Quantum computers operate at millikelvin temperatures. Create a vector of Celsius values from -273.15°C to -0.01°C, in 0.01°C steps. Convert these to Kelvin, filter temperatures below 0°C, modify values above -100°C, and access specific elements using indices. Use a switch statement to determine the quantum operation that can be performed based on specific temperature ranges (e.g., below -100°C: "Qubit Initialization", above -100°C: "Qubit Measurement"). Print the results, with comments explaining the code.
(Hint: Kelvin = Celsius + 273.15)

OR

2.b) Imagine you're developing a program to manage employee data in a company. Create a list to store various types of employee information (e.g., name, age, department, salary). Access and modify specific elements of the list using single and double brackets. Use lapply() to apply a function to calculate a 10% salary raise for each employee. Access and print sublists such as all employees in a particular department or employees with a salary greater than a certain threshold. Include comments to explain the code.

3. Describe the role of data frames in R for handling structured data. How can you work with CSV data in R using data frames? Provide an example demonstrating the following steps: (i) Loading a CSV file, (ii) Cleaning missing values, (iii) Separating or uniting columns for enrichment, (iv) Validating data by detecting duplicates or inconsistencies.

4. Explain the data generation process and the types of data encountered in data science. Discuss the significance of identifying true zero points in numerical data, differentiating between continuous and categorical data, and their impact on analysis. How can relevant data be found, and what are the key steps in interpreting data for meaningful insights?

5.a)

sid	Name	gender	age	prg	YoA	Native	State
101	Sarath	M	19	BCE	2022	Chennai	TamilNadu
101	Neha	F	18	BEE	2021	NA	Delhi
101	Naseer	M	NA	BCL	2022	Madurai	TamilNadu
102	Prabhu	M	18			Kollam	Kerala

Tasks:

i. Create the dataset in R using data frames as given in Table.

ii. Reshape the data into a long format; why is this transformation useful in data analysis?

iii. In what scenarios is the wide format is preferable? Convert that dataset back to a wide format.

iv. Create a unique student identifier by combining Admitted_year + Program + sid. (Ex: 2022BCE101). Discuss why this is important in Data Analysis?

v. Remove rows where Program **(prg)** and Admitted_year **(YoA)** are missing.

vi. Fill missing values in **"age"** using the last variable value for each student; Also fill missing values in **"Native"** by identifying a reasonable approach. Justify your choice.

OR

5.b) ***Tasks:***

i. *Create the dataset in R using data frames as given in the question 5 A.*

ii. *Convert the dataset in to a long format, where Age, Native, and State become key-value pairs under a new column Attribute. Why is this transformation useful?*

iii. Join an additional dataset that contains Student_ID and CGPA. What type of join ensures that all students are retained? Perform this operation. (use pipe operator with required functions)

iv. Find the total number of students in each program per admitted year and compute the average age per program.
(use pipe operator with required functions)

v. Convert the Program codes into their full forms using a mapping rule:
BCE → B.Tech Civil Engineering
BEE → B.Tech Electrical Engineering
BEC → B.Tech Electronics and Communication
BAI → B.Tech Artificial Intelligence
BST → B.Sc Statistics

vi. Identify students who have missing values in both Native and State and replace them with "Unknown". Why is this a good approach?

6. Compare relational and non-relational databases, discussing when relational databases are more suitable. Write an R script to generate a dataset of 60 students with suitable attributes, store it in an SQLite database, Query students with a CGPA above 8.0 and who have completed their internship. Explain each step and its role in R-database interaction.

7. Examine the role of REST APIs in data retrieval and discuss key HTTP methods with examples. Then, write an R script to parse a JSON response from a web API containing restaurant data, extract the names of the restaurants, convert nested lists into a data frame, and flatten the data for analysis.

8. Examine the principles of expressive and aesthetic data visualization, focusing on graphical encodings and layouts. Using Anscombe's dataset, demonstrate how visualization reveals patterns missed by summary statistics. Create multi-layered ggplot2 plots comparing its subsets, applying customizations like colors, themes, and annotations. Explain each customization's significance.

9. Explore the principles of data visualization for different types of attributes. Discuss approaches for univariate and multivariate visualization, explaining how to choose suitable plots for categorical, numerical, and mixed data types. Provide examples of appropriate ggplot2 visualizations for each case, justifying your choices.

10. Analyse the role of interactivity in data visualization and compare the features of Plotly, Rbokeh, and Leaflet in R. Create an interactive visualization using any one of these packages to explore a dataset of your choice, explaining how interactivity improves data interpretation.

⇔⇔⇔ **W/D/TY** ⇔⇔⇔