



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
WINTER SEMESTER 2024-2025

SLOT: B2

Programme Name & Branch: : B. Tech - All
Course Code and Course Name: : BCSE207L - Programming for Data Science
Faculty Name(s): : Prof. Rajasekhara Babu M, Prof. Rajkumar R, Prof. Sathiyakumar C,
Prof. Yuvaraj N, Prof. Raparathi Yaswanth
Class Number(s): : VL2024250502033, 2043, 2065, 2082, 2087
Date of Examination: : 17/03/2025, 09:30 AM to 11:00 AM
Exam Duration: : 90 minutes
Maximum Marks: 50

Q. No	Question	M	CO	BL
1.	A data analyst working for a retail company is tasked with analyzing sales trends and discount impact. The dataset (sales_data) contains the following columns: - store_id, category, revenue, quantity, discount_applied, payment_method, transaction_date a. Explain the following dplyr functions with syntax and examples: (5 marks) - filter() to extract rows based on revenue thresholds - group_by() for grouping data by category - summarise() for aggregating total revenue per category - mutate() to compute revenue per item - arrange() to sort data by revenue b. Write an R script using dplyr to analyze the following operations: (5 marks) - Filter transactions to remove outliers by excluding revenue values below the 10th percentile and above the 95th percentile. - Group data by category and discount_applied to compute total revenue, average transaction revenue, and discount impact. - Identify key insights by determining the top 3 highest revenue-generating categories and the store with the highest total sales. - Analyze payment trends by ranking payment methods in descending order based on transaction frequency. - Sort and present the summarized results in descending order of total revenue.	10	2	3
2.	A ride-sharing company maintains a dataset tracking the performance of drivers across different cities. The dataset includes the following columns: - driver_id, city, total_rides, peak_hour_rides, off_peak_rides, cancellations, average_rating, earnings Using dplyr & tidyr, perform the following data transformations in a structured pipeline: - Reshape the dataset from wide format to long format, creating a ride_type column (peak_hour or off_peak) and a ride_count column. - Filter drivers who have a cancellation rate above the 75th percentile and an average rating below 4.0. - Group the data by city and ride_type, then compute the total rides, average earnings per ride, and standard deviation of earnings for each group. - Classify drivers into performance tiers (High Performer, Moderate, Low Performer) based on their total earnings using a conditional logic function. - Reshape the dataset back to wide format, ensuring that each ride type's average earnings per city appears as a separate column. The final dataset should provide city-wise driver performance insights, enabling comparative analysis of ride patterns, earnings, and driver reliability. Analyze each transformation step, justify its necessity, and provide the complete R code implementation.	10		2



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
WINTER SEMESTER 2024-2025

SLOT: B2

A movie streaming platform maintains a relational database with the following tables:		10	3	3
3.	A movie streaming platform maintains a relational database with the following tables: • movies (movie_id, title, genre, release_year, avg_rating) • users (user_id, name, age, region) • watch_history (user_id, movie_id, watch_time, rating) Using SQL queries in R, perform the following tasks: 1. Retrieve the top 5 highest-rated movies with their genres. 2. Identify the most-watched movie genre in each region based on unique users. 3. Compute the average rating given by users in each age group (18-25, 26-35, etc.) and sort in descending order. 4. Write R code to connect to the database, execute the queries, and return results as a data frame. Explain the SQL logic, justify each approach, and provide the complete R implementation.			
4.	A weather forecasting API provides real-time weather data for multiple cities, returning information such as temperature, humidity, wind speed, and weather conditions in JSON format. The goal is to develop an R-based automated system that fetches, processes, and analyses weather data efficiently. Analyse the below mentioned tasks using R programming. 1. Write an R function that: • Fetches real-time weather data for at least five cities from a public API. • Extracts and structures key information like city, temperature, humidity, wind speed, and weather description into a tidy data frame. 2. Implement a mechanism to store the retrieved weather data over time, ensuring it can be used for trend analysis. 3. Propose a strategy to automate API calls at fixed time intervals (e.g., every 30 minutes) and handle potential API request failures. 4. Suggest a method to visualize temperature trends across cities dynamically using an appropriate R package.	10	3	4
5.	A public restaurant API provides data on eateries in Seattle, including name, location, cuisine type, ratings, number of reviews, and descriptions. The dataset includes Cuban-fusion restaurants, non-Cuban establishments with Cuban-inspired dishes, and some locations that are misclassified. Extracting only authentic Cuban restaurants requires an advanced filtering approach. Apply R programming to perform the below mentioned tasks. (5 Marks) a. Data Extraction and Cleaning Develop a robust querying and filtering strategy to retrieve restaurant data from the API while ensuring that only authentic Cuban restaurants are included. Handle inconsistencies in restaurant categorization, apply text-based filtering on descriptions, and prioritize restaurants with high ratings and a significant number of reviews. (5 Marks) b. Neighborhood Analysis and Visualization Using the cleaned data, determine which neighborhoods in Seattle have the highest concentration of top-rated Cuban restaurants. Implement an analytical model to identify these areas and propose an effective visualization strategy to map restaurant locations. The response must include a structured API querying approach, advanced data processing techniques, and an optimized R implementation that automates the discovery of authentic Cuban cuisine in Seattle.	10	3	3
