Question Paper

Exam Date & Time: 07-Jul-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

B.TECH END SEMESTER MAKEUP EXAMINATIONS

DATA SCIENTISTS TOOLBOX AND R PROGRAMMING [CRA 4059]

Marks: 50 **Duration: 180 mins.**

A

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed
Questions with subparts must be attempted with subparts clearly mentioned

1) Employ the R metacharacters ($,^, [], ? ,+, and *) and generate suitable regular expressions for the following test cases. Generate a regular expression (5)

A)

a. that matches any string that ends with ".csv".

b. that matches any string that starts with "A" or "B" and ends with "X" or "Y".

c. to match all strings that contain either the word "cat" or the word "dog".

d. that matches all strings that contain only letters and numbers, and no other characters.

e. that matches all strings that contain at least one uppercase letter and one digit.

B) Write a nested for loop that prints all pairs of numbers from two vectors, but skips pairs where the sum of the two numbers is greater than 10. (3)

C) Explain the difference between the sub() and gsub() functions in R, and provide an example of how to use them to replace a pattern in a character vector. (2)

2) For the data frame given below, write suitable code snippets (using dplyr packages) to get the required outputs. (5)

A)

sl name age ht school

1 Abhi 7 46 yes

2 Bhavesh 5 NA yes

3 Chaman 9 NA no

4 Dimri 16 69 no

a. Finding rows with no NA values

b. Calculating a variable x3 which is sum of height and age printing with ht and age

c. Calculating min of age

d. Arranging names according to the age.

e. Group by the variable school.

B) Suppose you have a character vector in R containing email addresses, and you want to extract the domain names from each address. Demonstrate with an example on how string manipulation functions in R to accomplish this. (3)

C) Name any 2 examples of real-world applications where qualitative, quantitative, raw, or processed data have been analyzed using R? (2)

3) Suppose you have a data frame called employee_df with columns for "name", "job_title", and "salary". You want to analyze the salaries by job title and perform the following tasks: (5)

A)

a. Calculate the average salary for each job title.

b. Calculate the difference between the highest and lowest salaries for each job title.

c. Split the data frame by job title.

d. Create a new data frame with columns for "job_title" and "total_salary", where "total_salary" is the sum of salaries for each job title.

e. Calculate the standard deviation of salaries for each job title.

Note: Use R loop functions (apply, lapply, sapply , tapply , mapply and split) accordingly.

B) What is the difference between "git pull" and "git fetch" commands in Git version control system. Explain when and why you would use each of these commands. (3)

C) What are some common data types in data science? (2)

4) What is subsetting and why is it useful? Write a line of code each to (5)

A) a. Create a vector my_vec with the values "apple", "banana", "cherry", "date".

b. Subset the vector to include only the elements that contain the letter "e".

B) A company wants to store information about its customers, including their contact information, purchase history, and preferences. What R datatype should be used to store this data? (3)

C) A weather station wants to store the temperature, humidity, and wind speed readings taken at regular intervals throughout the day. What R datatype should be used to store this data? (2)

5) Suppose you are given a dataset that contains information about the prices and features of different smartphones. The dataset includes variables such as screen size, battery capacity, camera quality, and price. You want to analyze this dataset to understand the relationship between these variables and the prices of the smartphones. (5)

A)

What are some steps you would take to prepare and analyze this dataset using R?

B) Create a 3x3 matrix in R called *my_matrix* that contains the numbers 1 to 9 in row-major order. Then, write R code to extract the element in the second row and third column of *my_matrix* and assign it to a variable called x. (3)

C) Sara takes a random sample of individuals in a population and identifies whether they smoke and if they have cancer. She observes a strong relationship between whether a person in the sample smoked or not and whether they have lung cancer. Sara claims that smoking is related to lung cancer in the larger population. She explains that the reason for this relationship is because cigarette smoke contains known carcinogens such as arsenic and benzene, which make cells in the lungs become cancerous. What sort of data analysis is Sara conducting? Justify your answer. (2)

-----End-----