Question Paper

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

MANIPAL ACADEMY OF HIGHER EDUCATION

MIT VI SEMESTER B.TECH
END SEMESTER EXAMINATIONS
CRA 4059 : Data Scientists Toolbox and R programming
MAY 2023

DATA SCIENTISTS TOOLBOX AND R PROGRAMMING [CRA 4059]

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

A

Answer all the questions.

Instructions to Candidates:
1. Answer ALL questions
2. Missing data may be suitably assumed
3. For questions with subparts-indicate the subpart before the answer

1) Formulate suitable functions that perform the operations mentioned below using suitable R functions (5)

A) (substr(), grepl(), grep(), toupper(), tolower(), strsplit(), sub(), gsub()).

a. That takes a character vector and a regular expression as input, and returns a vector of indices where the regular expression is matched.

b. That takes a character vector, a pattern to search for, and a new value as input, and returns a new vector where all occurrences of the pattern are replaced with the new value.

c. That takes a character vector and a delimiter as input, and returns a list of substrings obtained by splitting the original vector at each occurrence of the delimiter.

d. That takes a character vector and a regular expression as input, and returns a logical vector indicating which elements of the vector match the regular expression.

e. That takes a character vector, a starting index, and a length as input, and returns a substring of the original vector starting at the given index and of the specified length.

B) Write a while loop in R that repeatedly prompts the user to guess a random number between 1 and 100, and provides feedback to the user whether their guess was too high or too low until the correct number is guessed. (3)

C) With appropriate syntax, write any two commands which are commonly used in R while dealing with directories. (2)

2) What are the main functions of dplyr, and how can they be used to manipulate and summarize data in R? Explain with appropriate examples. (5)

A)

B) What are metacharacters in R? Provide examples of metacharacters and explain their significance in regular expressions. (3)

C) What are the differences between qualitative and quantitative data, and how can raw data be (2)

processed in R to extract meaningful insights?

3) A) Consider that you have two data frames, one containing information about customer orders and another containing information about products. The customer orders data frame has columns for "customer_name", "product_name", "quantity", and "price_per_unit". The products data frame has columns for "product_name" and "category". Perform the following operations: (5)

a. Merge the customer orders and products data frames based on the "product_name" column.

b. Remove any rows where the quantity is zero or the price per unit is negative.

c. Create a new column called "order_value" that calculates the total value of each order by multiplying the quantity by the price per unit.

d. Group the resulting data frame by category.

e. Calculate the average order value for each category.

B) A music streaming service wants to store information about its users' listening habits, including the songs they listen to, the artists they follow, and the playlists they create. What R datatype should be used to store this data and how? (3)

C) Compare lexical and dynamic scoping. (2)

4) A) What Git command should be used to achieve each of the following functionality? Include relevant options and examples in your answer: (5)

a. Create a copy of a remote repository on your local machine

b. Stage changes made to files in the working directory, preparing them to be committed to the repository

c. Record changes made to the repository, creating a new commit with a message describing the changes

d. Upload local commits to a remote repository, updating it with the latest changes

e. Retrieve changes made to the remote repository and merge them with the local repository

B) What is the process for removing missing values from vectors, matrices, and data frames in data analysis? (3)

C) With suitable syntax, demonstrate how dates and times are added in R. (2)

5) A) Write a line of code for each of the below instructions: (5)

a. Create a vector my_vec with the values 2, 4, NA, 6, 8, NA.

b. Remove the missing values from the vector my_vec.

c. Create a 1x 6 matrix my_matrix with the following values: 1, 2, 3, 4, 5, 6.

d. Convert my_matrix to a 2 x 3 matrix.

e. Subset my_matrix to include only the second row.

B) Write a R function called matrix_multiplication that takes two matrices as input and returns their product. Your function should include error handling to ensure that the input matrices have compatible dimensions for multiplication. You may assume that the input matrices are represented as numeric matrices in R. (3)

C) Manav has a dataset of customer purchases and aims to predict whether a new customer will make a purchase or not. He identifies several relevant features, including the customer's age, gender, (2)

and browsing history on the website. Manav trains a machine learning model using a classification algorithm on the dataset and assesses its performance on a holdout set. He observes that the model achieves high accuracy in predicting customer purchase behavior. Manav claims that the identified features are predictive of customer purchases on the website. He explains that this is because the model has learned patterns in the data that are indicative of customer purchase behavior.

What type of data analysis is demonstrated in the described example? Provide a rationale for your answer.

-----End-----