Question Paper

Exam Date & Time: 06-May-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

MIT MPL and BLR - BTech VI Semester - End Semester Examination - Jan-April 2024

DATA SCIENTISTS TOOLBOX AND R PROGRAMMING [CRA 4059]

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

Descriptive

Answer all the questions.

* Answer all questions.
* Assume the missing data suitably.
* Write neatly and legibly.
* Give suitable examples wherever necessary.

1) You have to manage a Git repository for a software project. Below are commands commonly used in Git. For each command, explain what it does in the context of Git repository management: (5)

a. git init

b. git add < file>

c. git commit -m "Commit message"

d. git status

e. git checkout -b < branch_name>

2) Consider the following scenarios related to data analysis: (3)

a. A retail company has collected sales data from its stores over the past year. They want to explore this dataset to identify trends and patterns in customer purchasing behavior. The goal is to gain insights into seasonal variations in sales, popular product categories, and customer preferences.

b. A marketing team is tasked with analyzing customer feedback survey responses to understand overall satisfaction levels and identify areas for improvement. They plan to calculate summary statistics, such as mean and standard deviation, to describe the distribution of responses and visualize the results using bar charts and histograms.

c. A telecommunications company wants to develop a predictive model to forecast customer churn. They plan to analyze historical customer data, including demographics, usage patterns, and past churn events, to identify predictive factors and build a machine learning model that can predict the likelihood of a customer churning in the future.

Identify which scenario corresponds to exploratory analytics, descriptive analytics, and predictive analytics. Justify your answers.

3) Compare and contrast vectors and lists in R. (2)

4) For each subpart listed below, describe a scenario where it could lead to a loop error and how to correct them: (5)

a.Using incorrect loop indices

b.Infinite loops

c.Variable scope issues

d.Inefficient use of loops

e.Syntax errors

5) Write the output for the following code snippet given in Figure 2B. Justify your answer. (3)

```
mat <- matrix(1:6, nrow = 2)
for (i in 1:nrow(mat)) {
  for (j in 1:ncol(mat)) {
    print(mat[i, j] * 2)
  }
}
```

Figure 2B

6) Explain the rationale behind employing a loop function within the framework of R programming, and provide an example of when it might be useful. (2)

7) You have a dataset containing information about sales transactions, including variables such as "transaction_id", "customer_id", "product_id", "quantity", and "unit_price". Your task is to calculate the total sales revenue generated by each customer over the entire dataset. You also need to exclude any transactions where the quantity sold is negative. Write R code to (5)

a. Initialize a dataset in R based on the Figure 3B.

b. Calculate the total revenue generated by each customer while excluding transactions with negative quantities.

c. Display the total revenue for each customer.

transaction_id	customer_id	product_id	quantity	unit_price
1	101	201	5	10
2	102	202	-2	15
3	101	201	3	12
4	103	203	4	8
5	102	202	6	20

Figure 3B

8) Classify the different R pseudo random generators based on the probability distributions they simulate. Additionally, with appropriate syntax simulate a Linear model. (3)

9) Compare and contrast the outputs generated by cbind and rbind functions for the R code given below. (2)

```
x < - 1:5

y < - 6:10

z < - 11:15
```

a. cbind(x,y,z)

b. rbind(x,y,z)

10) Consider the R code snippet given below. (5)

```
# import dplyr package

library(dplyr)

# create a data frame

stats < - data.frame(player=c('A', 'B', 'C', 'D'),

runs=c(100, 200, 408, 19),

wickets=c(17, 20, NA, 5))
```

Provide appropriate dplyr package code snippets to obtain the following outputs.

a. Fetch the player who has scored more than 100 runs.

b. Order the data based on runs.

c. Add new column avg whose value equals run/4.

d. Display the total sum scored and the average run rate.

e. Display the player names and run scored.

11) Write appropriate syntax and package information to explain how data from the following files can be read using R programming. (3)

a. XML file

b. Excel file

c. JSON file.

12) Examine the use and importance of Profiling in R. (2)

13) Inspect the use of debugging in any programming language. For the R code snippet given below compare and contrast the outputs of the (5)

different debugging tools in R.

```
# Function 1
function_1 < - function(a){
a + 5
}
# Function 2
function_2 < - function(b) {
function_1(b)
}
# Calling function
function_2("s")
```

14) Illustrate the function of the following code snippets on the data table created using the R code given below. (3)

```
dt < - data.table(x = rnorm(9), y = rep(c(a, b, c), each = 3), z = rnorm(9))
```

a. dt[c(2, 3)]

b. dt[, list(means(x), sum(z)]

c. dt2 < - dt; dt[, y:= 2]

15) Write the outputs returned by the following date functions in R. (2)

a. format(Sys.Date(), "%a %b %d")

b. a< - as.Date(Sys.Date(), "%d%b%Y")

-----End-----