


VIT

Vellore Institute of Technology

Final Assessment Test - April 2025

Course: BCSE207L - Programming for Data Science

Class NBR(s): 2033/2043/2065/2082/2087

Slot: B2

Max. Marks: 100

Time: Three Hours

- 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. Write a short note on two types of functions in R programming. Then, analyse the following code fragments, providing the expected outputs for the specified line numbers (6, 11, 17, 22, and 23) in the prescribed format:

Line No.	6	11	17	22	23
Output					

Line No.	Code
1	my_function <- function(arg1, arg2) {
2	result <- arg1 * arg2
3	return(result)
4	}
5	result1 <- my_function(5, 10)
6	print(result1)
7	my_function <- function(...) {
8	return(sum(...))
9	}
10	result2 <- my_function(3, 5, 2, 6, 1, 4)
11	print(result2)
12	my_function <- function(arg1, arg2) {
13	result <- arg1 - arg2
14	return(result)
15	}
16	result3 <- my_function(result1, result2)
17	print(result3)
18	my_function <- function(...) {
19	return(range(...))
20	}
21	result4 <- my_function(result1, result2, result3)
22	print(result4)
23	print(sort(c(result1, result2, result3), decreasing=TRUE))

2. Define a vector in R programming and provide a short note on how to modify vectors by explaining the following actions with an example: adding new elements, handling indices beyond vector length, and replacing elements based on conditions. Then

Inspect the following R code fragment and complete the missing lines by providing appropriate code and output for the specified line numbers (2, 3, 7, 8, 11, and 15) in the prescribed format

Line No.	Code/output
2	

1	#create a vector using sequence function from -1 to -10 by -2
2	v1<-_____
3	print(v1)
4	v2 <- c("Hello", "World")
5	v3 <- c("Everyone", "Universe")
6	#Concatenate v1 with v2 using appropriate vector function
7	result <- _____
8	print(result)
9	vowels <- c("a", "e", "i", "o", "u")
10	#return all vowels except "e" using vector indexing
11	all_but_e <- _____
12	print(all_but_e)
13	colors <- c("red", "green", "blue", "yellow", "purple")
14	#Retrieve red, blue and yellow from colours through indices vector
15	indices <- _____
16	extracted <- colors[indices]
17	print(extracted)

3. List the functions for inspecting a data frame (my_data) for the following descriptions and write answer in tabular form, with columns: 'Sl. No.' and 'Function' only

Sl. No.	Description	Function
1	Number of rows	
2	number of columns	
3	dimensions of (rows, columns)	
4	Names of columns	
5	Names of row	
6	First few rows (as a new data frame)	
7	Last few rows (as a new data frame)	
8	Opens the data frame in as spreadsheet-	

Table 4.1. List the functions to be listed

Then, Write an "R" program for the following steps:

Step1: Create the original `exam\_data` data frame with initial values from Table 4.2. `exam_data` and print the original data frame.

Step2: Extract specific columns (`name` and `score`) into a new data frame `result` and display it.

Step3: Append `new\_exam\_data` of Table 4.3. to Table 4.2. `exam\_data` using a data frame function and print the updated `exam\_data` with the newly added rows

Variable Name	Example Values
name	Anastasia, Dima, Katherine, James, Laura, Kevin, Michael, Jonas, Jacob, Tanya
score	12.5, 9, 16.5, 12, 9, 20, 14.5, 8, 19, 14
attempts	1, 3, 2, 3, 2, 3, 1, 3, 1, 2
qualify	yes, no, yes, no, no, yes, yes, no, yes, no

Table 4.2. `exam_data`

Variable Name	Example Values
name	Robert, Sophia
score	10.5, 9
attempts	1, 3
qualify	yes, no

Table 4.3. `new_exam_data`

4. Write about pros and cons of CSV files in "R" Programming. Then,

Write an R program for the following tasks:

Step 1: Create the original `exam_data` data frame as given in Table 4.2. `exam_data` and save it as a CSV file (`exam_data.csv`).

Step 2: Then read the file (`exam_data.csv`) back into an R data frame and print its contents. Extract specific columns (*name and qualify*) from the data frame and save the extracted data into a new data frame `result`. Display the contents of the result.

Step 3: Create a new data frame using data from Table 4.3. `new_exam_data` and save it as a CSV file (`new_exam_data.csv`). Append the rows from this file to the original `exam_data` data frame using the function. Print the updated `exam_data`.

5. You have been tasked with analyzing the starwars dataset to gain insights into the characteristics of various Star Wars characters. Analyze the starwars dataset in R to gain insights into the characteristics of Star Wars characters by performing various tasks. Specifically, you need to get appropriate package and its functions to Extract relevant columns such as name, species, height, mass, and homeworld, to extract rows with the focus on data for characters who are Human, to calculate the Body Mass Index (BMI) for each character which is calculated using the formula:

$$\text{BMI} = \frac{\text{mass (kg)}}{\left(\frac{\text{height (cm)}}{100}\right)^2}$$

and Find the average BMI for Human characters from each homeworld.

Use functions appropriately and complete code by filling missing code fragments at the following line numbers: 1,2,3,6,9,12,and 16, with description on code/Function reference to problem statement in the prescribed format

Line No.	Code	Description
1		

Line No.	Code
	#Step1: Install and Load the necessary packages
1	_____
2	_____
	#load starwars dataset which is a built-in dataset of package
3	_____
4	View(starwars)
	#Step2 Extract relevant columns such as name, species, height, mass, and homeworld.
5	starwars_data <- starwars %>%
6	_____
7	View(starwars_data)
	#Step 3: Extract rows with the focus on data for characters who are Human
8	human_data <- starwars_data %>%
9	_____
10	View(human_data)
	#Step 4: Calculate the Body Mass Index (BMI) for each character in a new column.
11	human_data <- human_data %>%
12	_____
13	View(human_data)
	#Step5: To provide summary of the average BMI aggregated grouped data for Human characters from each homeworld
14	summary_data <- human_data %>%
15	group_by(homeworld) %>%
16	_____
17	View(summary_data)

6

Explain briefly the importance of primary and foreign keys in relational databases and their role in ensuring data integrity and establishing table relationships. Then,

You are provided with a partially complete R program that interacts with a SQLite database named songs.db, which contains a table of songs. The program is designed to perform the following tasks:

1. Install the required packages and load them into the R environment.
2. Set the working directory to the location of the database file.
3. Establish a connection to the songs.db database.
4. Create a reference to the songs table within the database.
5. Write and execute a query to retrieve the titles of songs by the artist with artist_id = 11.
6. Display the constructed SQL query and retrieve the actual data as a table.
7. Disconnect from the database after the operations.

The code provided has missing fragments at the specified line numbers. Your task is to inspect the code and fill in the missing code fragments at the following line numbers: 1, 2, 3, 4, 5, 6, 9, 10, 13, 14, 17, and 19, ensuring proper functionality of the program.

Please complete the program by filling the missing code fragments in the prescribed format.

Lin. No.	Missing Code fragment
Line No.	Code
	#Step 1: Install & Load required 3 Packages and set working directory(WD)
	# To interact with databases # To use its functions on the database! # To access an database
1	_____
2	_____
3	_____
	# Load packages
4	_____
5	_____
6	_____
	#Get Working Directory details and Set Working directory path of Database
7	getwd()
8	setwd("E:\\Database")
	#Step 2: #Establish Database Connection to the database file "songs.db"
9	db_connection <- _____

	#Step 3: #Create a reference to the "songs" table in the database
10	songs_table <- _____
11	print(songs_table)
	#Construct a Query: Build a query to filter rows for a specific artist (e.g., "Queen" with artist ID=11) and select relevant columns (e.g., "title").
12	queen_songs_query <- songs_table %>%
13	_____
14	_____
15	print(queen_songs_query)
	# Display the SQL syntax stored in the query 'queen_songs_query'
16	show_query(queen_songs_query)
	# Execute the 'queen_songs_query' request, returning the *actual data* from the database
17	queen_songs_data <- _____
18	print(queen_songs_data)
	#Step 4 #Disconnect DB
19	_____

7. Write about RESTful APIs and its few standard HTTP methods. Then, Write an R script to handle JSON data using required the package(s). The tasks include

1. Install and load the library.
2. Set the working directory to a specified path (E://JSoN_Files).
3. Read and print JSON data from a file named "students.json".
4. Add a new student's details to the existing data.
5. Save the updated data to a new JSON file, "updated_students.json", ensuring proper formatting.
6. Reload and verify the updated file's contents.

new student's details	
Key Variable	Value
Name	David
Age	23
Major	Biology
Email	david@example.com

Explain any four key elements of the Grammar of Graphics in ggplot2 and how they contribute to systematic data visualization. Then,

8 A research scholar is conducting data analysis on a sample dataset to explore visualization techniques using the ggplot2 package in R. This analysis will support their thesis work by effectively presenting insights through data visualization. The dataset is as follows:

Experiment_ID	Measurement_Value	Category_Label
1	0.85	A
2	-1.12	B
3	0.43	C
4	-0.67	D
5	1.25	E

Perform the following steps

Step 1 : Install and load the necessary package

Step 2 : Create a dataframe for the given dataset

Step 3 : To visualize the data, the research scholar needs to create a scatter plot using ggplot2, where Experiment_ID is mapped to the x-axis and Measurement_Value to the y-axis. The plot should include forestgreen colored points with a size of 3 and text labels from the Category_Label column, positioned above the points in blue color. Additionally, appropriate axis labels and a title (Scatter Plot of Experimental Measurements) should be included to enhance the clarity of the visualization. Assist the research scholar by writing the R code required to generate this plot.

9. a) Discuss Mackinlay's Expressiveness Criteria in data visualization. How can issues like occlusion and overlapping data points be addressed to enhance expressiveness? Provide two techniques used for this purpose. Then,

Given the following dataset containing information about car manufacturers, production years, and horsepower:

Manufacturer	Year	Horsepower
Ford	2015	150
Toyota	2016	180
Honda	2017	200
BMW	2018	250
Audi	2019	300
Nissan	2020	220
Chevrolet	2021	275

Write R code using ggplot2 to:

- Create a line plot showing the trend of horsepower over the years, with points highlighting individual data values.
- Generate a histogram to visualize the distribution of horsepower, using an appropriate bin width and color scheme for clarity.

OR

9. b) Enumerate any two anonymous functions in R apart from `lapply()`. Provide a brief note of each function, including its purpose, and key features. Then,

The proctor at Vellore Institute of Technology (VIT) is analysing the Continuous Assessment Test (CAT)-2 performance of the following four students across five subjects to identify those who are scoring low marks and guide them for improvement in the Final Assessment Test (FAT). The marks obtained by the students are represented as a list of numeric vectors, as shown below:

Student	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Student1	19	24	21	32	35
Student2	25	30	28	35	20
Student3	35	40	42	23	43
Student4	40	42	43	44	45

Complete the missing R code for the following steps by providing appropriate code for the step numbers (1, 3, and 5).

Step1:	#Create a list that stores the marks of four students list_of_vectors <-
Step 2:	print(list_of_vectors)
Step 3:	#Scale each vector element in the list by dividing by maximum marks (50) and multiply with 15 using lapply()
Step 4:	print(scaled_list)
Step 5:	#Filter elements of scaled list greater than 7.5 using lapply()
Step 6:	print(filtered_list)

10. a) List three R packages used for interactive visualization and describe their purpose. Then A dataset contains information on different machine learning models, including their accuracy percentages and training times in seconds. The dataset is as follows:

Model	Accuracy (%)	Training Time (seconds)
Model A	89.5	120
Model B	91.2	150
Model C	87.8	100
Model D	92.3	180
Model E	90.7	140

Using the Plotly package in R, create an interactive bar.

Step1: Install and Load the plotly Package

Step2: Create a Sample Dataset

Step3: Create an interactive bar chart with the x-axis as "Model" and the y-axis as "Accuracy". The training time is displayed as hover text using `paste()`, and bars are colored "steelblue".

OR

Identify appropriate package for the given list of functions and write short brief note for each function in the prescribed table format

Sl. No.	Function	Package	Description
1	mutate()		
2	group_by()		
3	gather()		
4	spread()		
5	geom_point()		
6	fromJSON()		
7	dbConnect()		
8	GET()		

Then,

Develop an "R" program to perform join operation on given datasets in the following structure

Step 1: Install and Load the Required Library

Step 2: Create the Donors Dataset

Step 3: Create the Donations Dataset

Step 4: Perform a Left Join

Step 5: Perform a Right Join

Step 6: Perform an Inner Join

Step 7: Perform an Outer Join (Full Join)

Donors		
	DonorName	Email
1	Alice Walton	alice.walton@gmail.com
2	Jacqueline Mars	jacqueline.mars@gmail.com
3	Maria Franca Fissolo	maria.franca.fissolo@gmail.com
4	Susanne Klatten	susanne.klatten@gmail.com
5	Laurene Powell Jobs	laurene.powell.jobs@gmail.com
6	Francoise Bettencourt Meyers	francoise.bettencourt.meyers@gmail.com

Donations			
	DonorName	Amount	Date
1	Maria Franca Fissolo	100	15-02-2018
2	Yang Huiyan	50	15-02-2018
3	Maria Franca Fissolo	75	15-02-2018
4	Alice Walton	25	16-02-2018
5	Susanne Klatten	100	17-02-2018
6	Yang Huiyan	150	18-02-2018

⇒⇒⇒ V/D/TY ⇒⇒⇒