VI SEMESTER B.TECH.

GRADE IMPROVEMENT/MAKE-UP EXAMINATIONS, AUGUST 2021

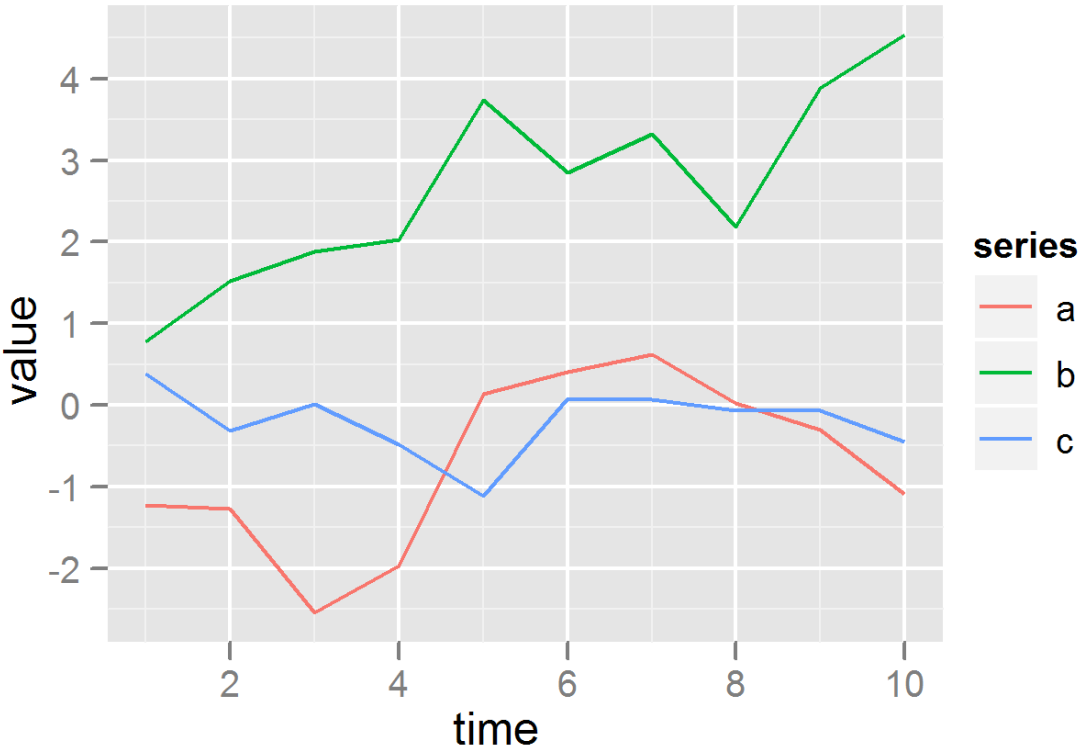
SUBJECT: INTRODUCTION TO DATA SCIENCE [CRA 4060]

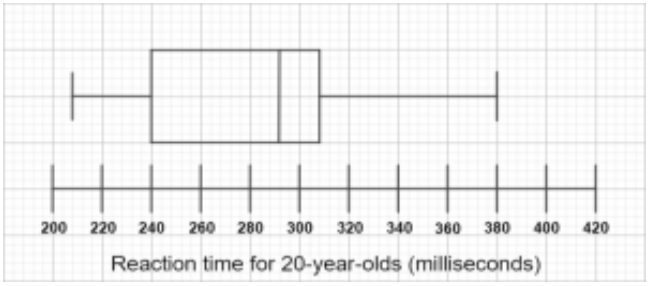
REVISED CREDIT SYSTEM
(12/08/2021)

Time: 2 Hours MAX. MARKS: 40

Instructions to Candidates:

- Answer **ANY FOUR FULL** questions.
- Missing data, if any, may be suitably assumed.

1A.	Briefly explain why we need graphs in data analysis? Briefly explain the different plots available in R to plot one dimension data. Consider the plot given below which is displayed on the current screen device. Write suitable R code to transfer this plot to a file of any format.  .	**5**
1B.	List and explain with help of R code, various functions available in dplyr package for data manipulation.	**5**
2A.	Write note on the role of statistical programming in reproducible research. What are the limitations of Sweave?	**5**
2B.	Explain the following with respect to reproducible research. (a) Replication (b) Reproducibility (c)Evidence based data analysis	**5**
3A.	Explain different lattice functions available with suitable R commands and examples.	

		5
3B.	When is it appropriate to visualize a dataset using a histogram? How is it more advantageous than box plots?. The reaction times (in milliseconds) of a group of 20-year-olds and a group of 30-year-olds were tested. The reaction times for the 20-year-olds has been shown in Fig Q. 3B below.  Fig Q. 3B: Box plot depincting reaction times of 20 year olds. The reaction times for the 30-year-olds are as follows: 220, 252, 256, 312, 332, 332, 400. Construct a boxplot for reaction times of the 30-year-olds data and write any one differences between the two groups.	**5**
4A.	Explain in detail the check list items to be followed to perform data processing in a reproducible manner.	**5**
4B.	Explain how replication helps to strengthen scientific evidence in reproducible research? Mention the challenges in doing replication. List out the data analysis files that are produced while addressing reproducible projects.	**5**
5A.	Suppose you have set of multivariate variables X_1……. Xn where X_1 = (X_{11}……X_{1m}), explain how you can find the solutions to the below listed problems? a. Find a new set of variables that are uncorrelated and explain variance possible. b. To find the one best matrix which depicts the original data with fewer variables.	**5**
5B.	Briefly explain the metrics used to find similarity between the data objects. With the help of R code, explain the following functions i. smoothScatter() ii. brewer.pal() iii. colorRampPalette()	
6A.	What is knitr? Explain its advantages. List the types of documents knitr is good at processing.	**5**
6B.	Create an R code chunk named “NewChunk” in a knitr document? List the steps involved in processing of knitr document.	**5**