H

Roll No. .....

MB–201 (BA1)

M. B. A. (SECOND SEMESTER)

END SEMESTER

EXAMINATION, June, 2023

DATA SCIENCE USING R

Time : Three Hours

Maximum Marks : 100

Note : (i) This questions paper contains two Sections—Section A and Section B.

(ii) Both sections are compulsory.

(iii) Answer any *two* sub-questions among (a), (b) and (c) in each main question of Section A. Each sub-question carries 10 marks.

(iv) Section B consisting of case study in compulsory. Section B is of 20 marks.

P. T. O.

Section—A

1. (a) Define data science and state its benefits for business. Explain the stages in data science. (CO1)
   (b) Elaborate upon the basic Data management process in R Studio. (CO1)
   (c) Describe the use of Packages in R. Explain few important packages required in Data Visualization. (CO1)
2. (a) Summarize in detail regarding data frame and arrays with an example R code. (CO2)
   (b) What are the different categories of functions used in R ? With the help of R code create a function to print squares of numbers in a sequence. (CO2)
   (c) Interprete the usage of conditional statement in decision making. With relevant R code implement "if" and "if-else" conditional statement in R Studio. (CO2)
3. (a) With relevant R code explain the process of loading a "csv" file and converting a data frame to "csv" file as output. (CO3)

   (b) What is the use of working directory ? With R code demonstrate how to set up a new working directory. (CO3)
   (c) Explain the process of error handling in R. (CO3)
4. (a) Illustrate the importance of descriptive statistics ? Write R code to obtain descriptive statistics for a csv file named "data". (CO4)
   (b) What is correlational analysis ? With relevant code explain the steps of correlation hypothesis testing in R. (CO4)
   (c) What is cluster analysis ? Explain K-means cluster and Hierarchical cluster analysis. (CO4)

Section—B

5. **Case Study :**

   Table 1 presents a subset (named data saved in "csv" format) of "diamonds" dataset available with ggplot2 package. (CO5)

Table 1

carat	cut	colour	clarity	depth	table	price
0.23	Ideal	E	SI2	61.5	55	326
0.21	Premium	E	SI1	59.8	61	326
0.23	Good	E	VS1	56.9	65	327
0.29	Premium	I	VS2	62.4	58	334
0.31	Good	J	SI2	63.3	58	335

"data" file contains information about 53,940 round-cut diamonds with 7 columns; cut, color and clarity are categorical variables, while the remaining follow a numeric structure.

Table 2 contains the description regarding the variables :

Table 2

Variable	Description	Values
price	price in US dollars	$326-$18,823
carat	weight of the diamond	0.2-5.01
cut	quality of the cut	Fair, Good, Very Good, Premium, Ideal
color	diamond color	J (worst) to D (best)

clarity	measurement of how clear the diamond is	I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, 1F (best)
depth	total depth percentage	43-79
table	width of top of diamond relative to widest point	43-95

Based upon the above data answer the following questions :

(i) Which graph type you feel should be used for visualizing "cut" variable ?

(ii) Can the same graph type be extended to visualize "color" and "price" variable. Justify your answers.

(iii) With R code formulate a linear regression model to check the effect of "carat" variable on "price" variable.

(iv) How will you test the hypothesis that "carat" has no effect on the "price" of a diamond.