INTRODUCTION TO R
(DSC2201)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) What is one of the main distinctions between R and Python?
(a) R is used for web development
(b) Python is mainly used for statistical analysis
(c) R is mainly used for statistical analysis
(d) Python does not support data visualization

(ii) What will the command 'typeof(3.414)' display?
(a) integer (b) double (c) character (d) logical

(iii) How can you view the structure of objects in a list or data frame in R?
(a) Using the 'inspect' function (b) Using the 'str' function
(c) Using the 'view' function (d) Using the 'show' function

(iv) What is an important special case of categorical data?
(a) Continuous data (b) Ordinal data
(c) Binary data (d) Text data

(v) What is the result of unclass(gender) if gender is a factor with levels "female" and "male"?
(a) Numeric codes that correspond to levels
(b) Character strings representing the levels
(c) Logical values corresponding to levels
(d) Integer values representing the levels

(vi) What will class(4) return?
(a) character (b) numeric (c) logical (d) factor

(vii) What does the 'traceback' tool do in R?
(a) It displays the call stack of functions called before an error occurred
(b) It allows the user to step through the code
(c) It combines the call stack with browser mode
(d) It adds a browser statement into a function

(viii) What does using a negative index in R do?
(a) Duplicates elements (b) Excludes elements
(c) Includes elements (d) Reorders elements

(ix) What operator is used in R to generate a sequence of numbers
(a) The / operator (b) The ; operator
(c) The * operator (d) The : operator

(x) How can new columns be added to a data frame in R?
(a) By merging it with another data frame
(b) By expanding it to accommodate the new values
(c) By converting it to a matrix
(d) By using the append function

Fill in the blanks with the correct word

(xi) Logistic regression provides a _____ response, while multiple linear regression gives a _____ response.

(xii) The goodness of fit of a logistic regression model is measured by the _____ value.

(xiii) The _____ function is used to calculate the mode of a vector of values.

(xiv) In ggplot2, control elements of the graph that are not related to the data itself but influence the output, such as background colour, font sizes, gridlines, and label colours are called _____.

(xv) Among the techniques covered in this course, _____ is most suitable in the context of deciding whether a loan applicant in a bank is eligible to receive a loan.

Group - B

2. (a) In R, how can you have a function return multiple objects as output? Explain with an example. *[(CO5)(Remember/LOCQ)]*

(b) (i) What will be the output of the following code?

```
X <- c(4, 5, 3, 1, 19, 2)
A <- sort(X)
print(A)
```

(ii) How should the above code be changed to sort X in the reverse order of how the above code sorts it in? *[(CO5)(Analyse/IOCQ)]*

(c) What will be displayed when the following statements are executed in R?

```
matrix(1:8, ncol = 4)
    class("HITK")
    df<- data.frame(Student = c("Priya", "Sajal", "Abhi"),
                    Physics = c(89, 91, 94), Chemistry = c(95, 84, 90),
                    Mathematics = c(90, 99, 96))
    df[2, c(TRUE, TRUE, FALSE, TRUE)]
```

[(CO3)(Apply/IOCQ)]

4 + (2 + 2) + 4 = 12

3. (a) What are the rules for naming an object in R? *[(CO5)(Remember/LOCQ)]*

(b) What will the second and third lines of the following code display?

```
die <- 1:6
die
die + 1:2
```

[(CO5)(Analyse/IOCQ)]

(c) What is the simplest type of object in R? Illustrate its use with examples. *[(CO5)(Understand/LOCQ)]*

(d) How is the factor class of data used in R? Explain with examples. *[(CO5)(Understand/LOCQ)]*

3 + 2 + 3 + 4 = 12

Group - C

4. The following questions are based on the figure below that gives information about all the columns and some rows of the mtcars data frame:

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

Write R code that does the following. Explain each line of each code component.

(i) Find out the number of rows and columns in mtcars, and the types of data in each column.

(ii) Draw a bar blot of the cars with different gears.

(iii) Find out the correlation coefficient between hp and mpg.

(iv) Draw a scatterplot between hp and mpg. *[(CO3)(Apply/IOCQ)]*

(3 × 4) = 12

5. (a) Using R code, sample 50 values between 51 and 152 without replacement, and display the sample. Explain your code. *[(CO4)(Apply/IOCQ)]*

(b) For the data sampled in the preceding question, you need to visualize the interquartile range, the outliers, and the median. (i) Write R code to generate a plot for these purposes and explain the code. (ii) Sketch and explain a representative plot. *[(CO4)(Apply/IOCQ)]*

(c) Write a function in R to generate the Fibonacci sequence (0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, …) using recursion. The function needs to accept the argument n, and return the sequence till the n'th Fibonacci number. *[(CO4)(Apply/IOCQ)]*

4 + 4 + 4 = 12

Group - D

6. Observe the details below about a data frame (mydata) that has information about variables such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution relate to admission into graduate school, and answer the following questions.

```
## 'data.frame': 400 obs. of  4 variables:
##  $ admit: int  0 1 1 1 0 1 1 0 1 0 ...
##  $ gre  : int  380 660 800 640 520 760 560 400 540 700 ...
##  $ gpa  : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
##  $ rank : int  3 3 1 4 4 2 1 2 3 2 ...
```

(i) Write R code to generate the descriptive statistics of the dataset. *[(CO4)(Apply/IOCQ)]*

(ii) In the dataset, one variable needs to be considered as a categorical variable. Write R code to convert that variable into a categorical variable and add the converted variable to the dataset. *[(CO4)(Apply/IOCQ)]*

(iii) Write R code to develop, and run, and show the output of a logistic regression model to understand how variables influence admission to graduate school. Explain your code. *[(CO4)(Apply/IOCQ)]*

(2 + 4 + 6) = 12

7. (a) Why is Naïve Bayes classifier called "naïve"? What are the implications of the naiveness of Naïve Bayes classifier? *[(CO3)(Understand/LOCQ)]*

(b) State and explain the theorem underpinning Naïve Bayes classifier. *[(CO3)(Understand/LOCQ)]*

(c) Compare and contrast logistic regression and Naïve Bayes classifier. *[(CO3)(Understand/LOCQ)]*

(d) What are the metrics that are used to evaluate a Naïve Bayes classifier model? *[(CO3)(Understand/LOCQ)]*

3 + 3 + 3 + 3 = 12

Group - E

8. (a) How are missing values usually represented in R? Given a data frame called mydata, write R code to remove empty rows and columns from data. *[(CO5)(Remember/LOCQ)]*

(b) Which type of a plot can help detect outliers in a dataset? Explain with examples. What is the function in R for generating such a plot? *[(CO6)(Remember/LOCQ)]*

(c) You are given the runs scored in each test match innings batted by Sir Donald Bradman and Sachin Tendulkar. Using this data and all the R in-built functions and programming techniques you are aware of, design a study that will help you establish whether Bradman or Tendulkar was a better batsman. Write down the steps of the study with supporting R code. *[(CO5)(Create/HOCQ)]*

(1 + 3) + 2 + 6 = 12

9. (a) You are given the following code:

```
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152,131)
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
relation <- lm(y~x)
print(summary(relation))
```

And this is the output from running the above code:

```
Call:
lm(formula = y ~ x)
Residuals:
Min 1Q Median 3Q Max
-6.3002 -1.6629 0.0412 1.8944 3.9775
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.45509 8.04901 -4.778 0.00139 **
x 0.67461 0.05191 12.997 1.16e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.253 on 8 degrees of freedom
Multiple R-squared: 0.9548, Adjusted R-squared: 0.9491
F-statistic: 168.9 on 1 and 8 DF, p-value: 1.164e-06
```

What is the model and its output this R code represents? How well does the model fit the data? Explain your answer. *[(CO4)(Analyse/IOCQ)]*

(b) What is skewness and why is it important? For a distribution that is negatively skewed, would the mean be higher than the median, or vice versa? Explain your answer. *[(CO4)(Remember/LOCQ)]*

(2 + 4) + (2 + 2 + 2) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	39.55	54.2	6.25