Q. Code: 708400

B.E./B.TECH. Degree Examination, December 2020

Fifth Semester

IT18015 – Statistical Analysis Using R Programming

(Regulation 2018)

Time: Three hours Maximum : 80 Marks

Answer **ALL** questions

PART A - (8 X 2 = 16 marks)

1. What is the correct statement of creating data frame with three rows and two columns
   a) data.frame(1:3,c("A","B"))
   b) data.frame(No=1:3,Name=c("A","B","C"))
   c) data.frame(No=1:2,Name=c("A","B","C"))
   d) data.frame(1:3,c("A","B","C"))
2. Select the correct code to redirect output to a file from both print and cat function.
   a) cat("XYZ", "\n", file="filename")
   b) print("XYZ", "\n",file="filename")
   c) sink("filename")
      source("output.R")
      sink( )
   d) sink("filename")
      cat("XYZ", "\n", file="filename")
      print("ABC", "\n",file="filename")
      sink()
3. A box contains 10 green balls and 8 red balls. Two balls are drawn with replacement. What is the probability of getting both red balls?
   a) 2/16
   b) 16/81
   c) 64/100
   d) 20/81
4. Rice crops damaged badly on account of heavy rain is
   a) Cyclical movement
   b) Random movement
   c) Secular trend
   d) Seasonal movement
5. Write R Code to create a first vector containing 8 elements starting with 1 and ending with 4. Create a second vector containing 6 elements starting with 1 and ending with 6. Add these two vectors and print the created vectors and their sum.

6. Write R code to create pie chart for the following data and use appropriate labels and title.

   Maths -------190

   Science --------200

   Social ---------170

   English -------- 180

   Others ---------180

7. If you want to know all the values in c (1, 3, 5, 7, 10) that are not in c (1, 5, 10, 12, 14), which built-in function in R can be used? Also, how this can be achieved without using the built-in function?

8. Write R code to create the logistic regression model between the columns "am" and 3 other columns - hp, wt and cyl from mtcars dataset. Predict the class of "am" when cyl=6, hp=115, wt=3.315.

PART B - (4 X16 = 64 marks)

09. (a) (i) Write a function in R to create a vector from 1 to 100. Multiply the elements which are smaller than 5 and larger than 90 with 10 and compute the sum of multiplied values. Multiply other elements with 1.5 and compute the product of multiplied values. **(8)**

    (ii) Create a 5x5 matrix of integers from 1 to 25, filling one row at a time. Create another 5x5 matrix from vectors v1 and v2. Multiply the created matrices. Square all elements of the resultant matrix and divide by 10. Implement the same with R code. **(8)**

(OR)

(b) (i) Write R code to create employee data frame with eno, ename and designation. Add a salary column to the employee data frame. Extract all employees who are earning more than 10000. Extract the designation of 3rd employee. **(8)**

(ii) Write R code to find the levels of factor of a vector c(1,3,7,4,5,10,9,8). Convert a given vector to an ordered factor. **(8)**

10. (a) (i) Write R code to roll 3 dies. **(8)**

    a) What is the probability of sum of all the outcomes will equal to 10?

    b) What is the probability of outcome of second die is greater than the first die?

(ii) A coin is tossed three times, where **(8)**

a) A : head on third toss, B : heads on first two tosses

b) A : at least two heads, B : at most two heads

Determine P(A|B), P(B|A) and implement same with R code.

(OR)

(b) (i) **(8)**

	Cigar Smoker	Not a Cigar Smoker
Male	4845	46155
Female	833	48167

What is the probability of males given that they smoke cigars? Write R code for computing the probability.

(ii) The discrete random variable X has the probability distribution **(8)**

P(x) = $\frac{x}{36}$ $for\ x = 1,2,3,4,5,6,7,8$. Compute the Mean, Variance, Standard deviation of X and also write R code.

11. (a) (i) Fit a straight line to the following data. **(8)**

X	12	17	19	25	32	38	43
Y	65	78	82	92	90	97	100

Estimate the value of Y, when X=35 and write R code for model and prediction.

(ii) Ram is skeptical of his friend's claim that Gerald's Cafe has much stronger coffee than Sabine's Beans does. So Ram takes a random sample of large coffees from both shops, and measures the amount of caffeine content in each coffee. Here is a summary of the results: **(8)**

	Sabine's Beans	Gerald's Cafe
Mean	164 mg	170 mg
Standard Deviation	5.1 mg	3.1 mg
Number of cups	37	35

Analyze the above and write a suitable R code for the type of test that could be applied. Justify your answer.

(OR)

(b) Construct the decision tree for the following data by considering Entropy and Information Gain as measures. Also write R code for implementing the same. **(16)**

Color	Type	Doors	Tires	Class
Red	SUV	2	Whitewall	+
Blue	Minivan	4	Whitewall	-
Green	Car	4	Whitewall	-
Red	Minivan	4	Blackwall	-
Green	Car	2	Blackwall	+
Green	SUV	4	Blackwall	-
Blue	SUV	2	Blackwall	-
Blue	Car	2	Whitewall	+
Red	SUV	2	Blackwall	-
Blue	Car	4	Blackwall	-
Green	SUV	4	Whitewall	+
Red	Car	2	Blackwall	+
Green	SUV	2	Blackwall	-
Green	Minivan	4	Whitewall	-

12. (a) Create a student data frame for 10 students with missing data. Write R code to perform the following operations. **(16)**

i) To remove na, nan and infinite values

ii) To find the sum of missing values in marks column

iii) To compute the percentage of missing values in entire dataset.

iv) Tabulate missing data and explore the missing data visually.

(OR)

(b) (i) How would you make multiple plots on a single page in R? Illustrate with an example. **(8)**

(ii) How will you use continuous value as the conditioning variable in R? How will you facet the data in ggplot2 package? Illustrate with examples. **(8)**