M.S.Ramaiah Institute of Technology

(Autonomous Institute, Affiliated to VTU)

Department of Computer Science & Engineering

QUESTION BANK FOR VII SEMESTER (Term: Aug-Dec 2016)

Data Analytics Laboratory (CSL717)

I.A. Marks : 50	**Exam Hours: 03**
Credits: 0:0:1	**Exam Marks: 50**

Instructions: Student has to execute one program from Part-A and one program from Part-B completely.

Part – A

1. Create a vector x with values (1,7,3,2,5,0.5,9,10)
    a. Write a function that calculates the mean of a numeric vector x, ignoring the smallest (s) and largest (l) values (this is a trimmed mean).
    b. Using a for loop and your function from part **(a)**, create a vector whose elements are the trimmed means of the vectors in list.random, taking s = 5 and l = 5
    c. Calculate the un-trimmed means for each of the vectors in the list. How do these compare to the trimmed means you calculated in part (b)? Explain your findings.
    d. Repeat part **(b)**, using the lapply and sapply functions instead of a for loop. Your lapply command should return a list of trimmed means, and your sapply command should return a vector of trimmed means.

2. Create a list of vectors of varying length:
    a) Create a vector "vector1" of 100 random numbers from 10 to 20.
    b) Create a list with 100 vectors containing random numbers from uniform distribution of size given by "vector1"
    c) Use a for loop to find the lengths of the vectors in the list. First make a vector for storing the lengths.
    d) Repeat c) using sapply
    e) Repeat c) using lapply

3. Examine the built in ChickWeight data (the help gives background about the data).
    (a) Construct a plot of weight against time for chick number 34.
    (b) For chicks in diet group 4, display box plots for each time point.
    (c) Compute the mean weight for chicks in group 4, for each time point. Plot this mean value against time.
    (d) Repeat the previous computation for group 2. Add the mean for group 2 to the existing plot.
    (e) Add a legend and a title.

4. Consider a vector x=(0,1,1,0,1) and perform the following.
    a. Write a function called isBinary that accepts a single argument called x. Your function should return TRUE if x is binary and FALSE otherwise.
    b. Rewrite your isBinary function to now have arguments (x, allow.NA = FALSE) that has the following behaviour:

   i. If allow.NA is TRUE, your function should return TRUE if the elements of x are 0, 1 or NA. It should return FALSE otherwise.
   ii. If allow.NA is FALSE your function should behave exactly as in (a). i.e., the function should return TRUE if and only if the elements of x are 0 or 1.

  c. Write a function called calculateBinarySummary()that returns the proportion of values in x that are equal to 1, out of the total number of values that are 0 or 1. (NAs should not count toward the denominator in calculating your proportion.)

5. Load the in-built dataset mtcars() and perform the following.
   a. Dot plot of mpg for each car model
   b. Create a colored histogram of 12 bins with x-axis as 'Miles per gallon' and y-axis as 'frequency'.
   c. Create kernel density plots of mpg by number of cylinders with legends as 4 cylinders, 6 cylinders and 8 cylinders. Interpret the results obtained in (a) & (b).
   d. Generate a box plot of car mileage versus transmission type and number of cylinders.

6. Create the following patterned matrices. In each case, your solution should make use of the special form of the matrix—this means that the solution should easily generalise to creating a larger matrix with the same structure and should not involve typing in all the entries in the matrix.

$$\textbf{(a)} \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 & 0 \\ 2 & 3 & 4 & 0 & 1 \\ 3 & 4 & 0 & 1 & 2 \\ 4 & 0 & 1 & 2 & 3 \end{pmatrix} \qquad \textbf{(b)} \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 8 & 9 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 9 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{pmatrix}$$

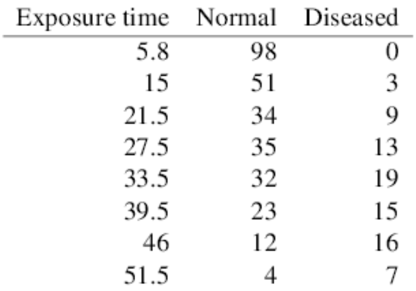
$$\textbf{(c)} \begin{pmatrix} 0 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \\ 1 & 0 & 8 & 7 & 6 & 5 & 4 & 3 & 2 \\ 2 & 1 & 0 & 8 & 7 & 6 & 5 & 4 & 3 \\ 3 & 2 & 1 & 0 & 8 & 7 & 6 & 5 & 4 \\ 4 & 3 & 2 & 1 & 0 & 8 & 7 & 6 & 5 \\ 5 & 4 & 3 & 2 & 1 & 0 & 8 & 7 & 6 \\ 6 & 5 & 4 & 3 & 2 & 1 & 0 & 8 & 7 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 & 8 \\ 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 \end{pmatrix}$$

7. Consider a matrix with random-normally distributed values with some occurrences of NA values and perform the following.
   a. Using the matrix initialized write a function which takes a single argument that picks out the submatrix which consists of all columns which contain no occurrence of NA.
   b. Write a function which takes a single argument and returns the submatrix which is obtained by deleting every row and column from the input matrix which contains an NA.

Part – B

1. From the library MASS, use ‘cats’ data and perform the following.
    a. Extract Male cats data set separately.
    b. Display a scatterplot for male cats. Interpret the dependent and independent variables of the data.
    c. Fit a linear regression model for male cats. Add fitted regression line to scatterplot of male cats data.

2. Consider the data of Pneumoconiosis among coalface workers as shown in table and perform the following.
    a. Create a data frame with three columns ‘exposure time’, ‘normal and ‘diseased’ Data were collected. Examine the relationship between exposure time (years) and risk of disease.
    b. Display a scatterplot for male cats. Interpret the dependent and independent variables of the data.
    c. Fit the linear regression model for the data. Does the model seem to fit the data reasonably well?
    d. Predict the danger values if the exposure time doubled.



Exposure time	Normal	Diseased
5.8	98	0
15	51	3
21.5	34	9
27.5	35	13
33.5	32	19
39.5	23	15
46	12	16
51.5	4	7

3. From the library MASS, use Cars93 data and perform the following.
    a. Using the Cars93 data and the `t.test()` function, run a t-test to see if average `MPG.highway` is different between US and non-US vehicles.
    b. What is the confidence interval for the difference? Interpret this confidence interval.
    c. Repeat part (a) using the `wilcox.test()` function.
    d. Are your results for (a) and (c) very different?

4. Consider the data as shown in the table and perform the following.
    a. Create a table by name ‘smoking’ with rbind() and list() function.
    b. Use `fisher.test()` to test if there's an association between smoking and lung cancer. Interpret the results of the same.
    c. What is the odds ratio? Interpret this quantity.
    d. Write an inline code chunk that determines whether your findings are statistically significant?

has.smoked	lung.cancer	Freq
yes	yes	688
no	yes	21
yes	no	650
no	no	59

HOD, Dept. of CSE

5. Create a data frame based on the data shown in the table and perform the following.
    a. Plot a single scatterplot() that describes the relationships between all the variables in the dataset. What do you infer from the plot i.e. dependent variable?
    b. Apply linear regression model for the same using the dependent variable as income and age,education, gender as the independent variables. Interpret the results of the model? Is it over-fitted?
    c. If the model is overfit in (b), apply linear regression model once again with suitable independent variables.

ID	Income	Age	Gender
1	113	69	1
2	91	52	0
3	121	65	0
4	81	58	1
5	68	31	1

6. From the library MASS, use birthwt data and perform the following.
    a. Use with() along with the tapply() function to produce a table showing the % of babies born weighing under 2500g within each combination of mother's race and smoking status.
    b. Use the tapply() function to produce a table showing the proportion of babies born weighing under 2500g, broken down by race, smoking status, and hypertension.
    c. Repeat part (b) using the aggregate() function.

7. Consider the dataset 'bikes.csv' that contains daily bikeshare counts, along with daily measurements on environmental and seasonal information that may affect the bikesharing. Perform the following tasks on the dataset.
    a. Transform temp and atemp to degrees C instead of [0,1] scale
    b. Transform humidity to %
    c. Transform wind speed (multiply by 67, the normalizing value)
    d. Build a regression model with dependent variable as 'cnt' and independent variables as 'yr, temp, hum, season'. Modify the model with 'season' as factor variable. What is the difference between these two models?
    e. What is the interpretation the coefficient(s) of season in each of the models in part (a)?
    f. Use ggplot2 graphics to construct boxplots of count for each season. Based on what you see from these plots, which model do you think makes more sense? Explain.
    g. Using ggplot2 graphics, construct a scatterplot of cnt (bikeshare count) across mnth (month of the year). Describe what you see. Would a linear model be a good way of modeling how bikeshare count varies with month?

Note : Execution : 35 Marks (Part-A: 15 marks Part-B: 20 marks)
Write up : 8 marks (Part-A: 4 marks Part-B: 4 marks)
Viva : 7 marks
Change of either question : -5 marks

HOD, Dept. of CSE