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 (CSL1542)

I.A. Marks : 15 **Exam Hours: 02**

1. The following table gives the size of the floor area (ha) and the price ($A000), for 15 houses sold in the Canberra (Australia) suburb of Aranda in 1999.
   Type these data into a data frame with column names area and sale.price.
   (a) Plot sale.price versus area.
   (b) Use the hist() command to plot a histogram of the sale prices.
   (c) Repeat (a) and (b) after taking logarithms of sale prices.
   (d) The two histograms emphasize different parts of the range of sale prices. Describe the differences.

	area	sale.price
1.	694	192.0
2.	905	215.0
3.	802	215.0
4.	1366	274.0
5.	716	112.7
6.	963	185.0
7.	821	212.0
8.	714	220.0
9.	1018	276.0
10.	887	260.0
11.	790	221.5
12.	696	255.0
13.	771	260.0
14.	1006	293.0
15.	1191	375.0

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. Use the functions mean() and range() to find the mean and range of:
   (a) the sample of 50 random normal values, that can be generated from a normal distribution with mean 0 and variance 1
   (b) the columns height and weight in the data frame "women".
   [The *datasets* package that has this data frame is by default attached when R is started.]

5. Use the 'cats' data set from MASS library. Extract Male cats data set separately. Do scatterplot for male cats. Birth weight(BWT) in X axis and Heart weight(HWT) in Y axis. Fit a linear regression model for male cats. Add fitted regression line to scatterplot of male cats data.

6. Data from 20 chemical dissolutions are collected to analyze the association between the toxicity of dissolution on one side and 3 explanatory variables on the other side. Store the data in lser.csv with variables (Suitable assumptions can be made)
   tox: toxity of dissolution
   base:ability to accept hydrogen ions
   acid: ability to liberate hydrogen ions
   colour: ability to change colour
   Create dataset called lser with data. Plot to get overview of dataset. Fit a linear regression models with tox as response to base,acid and colour as explanatory variables(tox vs base, tox vs acid, tox vs colour). Give the interpretation of the parameter estimates. Are all explanatory variables significant? Measure expected toxity for a specific solvent with specific base.

7. Pneumoconiosis among coalface workers Data were collected in order to examine the relationship between exposure time (years) and risk of disease.

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

   Store the data in coalworker1.csv. Plot the data to get overview of data. Fit the linear regression model for the data. Does the model seem to git the data

reasonably well? What happens if the exposure time doubled? Predict the danger values.