College of Engineering Pune
Probability and Statistics for Engineers(MA-21001)
T.Y.B. Tech. Semester (Electrical/Non-electrical branches) 2024-25 Autumn Semester
Assignment on R software and R Studio
by Dr. Yogita Mahesh Mahatekar

Introduction to 'R' and R Studio

1. Install latest versions of R and R studio on your system as per the type of operating system you are using.
2. Install packages namely 'plyr', 'MASS', 'ggplot2', 'dplyr' etc.
3. Find answers to log2(2^5) and log(exp(1)*exp(1)) - explain
4. Explore the **Help** tab in RStudio:
   Change the labels, title, etc. while plotting a graph of a function. Hint: try plot(x,y,xlab="x-axis") Copy paste a few examples provided in Help and see what they do
5. Find out how to calculate median - using help button and web search
6. Calculate mean and median of built in data set viz. CO2.
7. Take the help of R to see all commonds related to data frame operations.
8. Import a data set from the Internet directly or through a file.
9. In R Studio, identify source window, plot and files window, workspace and console window.
10. Create two R-scripts and save it on Desktop as testscript1.R and testscript2.R
11. Load testscript1.R in testscript2.R
12. Create a new folder on your computer and make it your working directory.
13. Create a subset from captaincy data frame with the captains who have played > 20 matches and lost < 14 matches. **Note that 'captaincy' data frame is uploaded on Moodle.**
14. Using built-in dataset `iris`, implement the functions like: Summary, class, typeof, head, tail, str, Merge. Also try importing data frames in different formats.
15. Using built-in dataset `iris`, find out the categorical variables.
16. Can you find a variable which is categorical, but R reads it as numeric?
    If yes, change it to categorical.
17. Create a numeric vector c(1:5) and a 5 by 3 matrix with elements from 1 to 15.
18. Create a named list with vector, matrix and `iris` data set.

19. Retrieve `iris` data set from the list using dollar operator and indexing method.

20. State the difference between the results obtained by using dollar operator and indexing method of accessing `iris`.

21. Read the file $moviesData.csv$
    Create a histogram of the object named $imdb\_num\_votes$ in this file.

22. Create a pie chart of the object $mpaa\_rating$ and save the plot.

23. Read the file $moviesData.csv$
    Create a bar chart of $critics\_score$ for the first 10 movies.

24. Create a scatter plot of $imdb\_rating$ and $imdb\_num\_votes$ to see their relation. Also save the plot.

25. Consider the built-in data set $mtcars$. Find the numerical variables in this data set.

26. Make a scatter plot from the objects named mpg and wt in this data set.Save the plot in $.jpeg$ format.

27. Using built-in data set $mtcars$, draw a bar chart from the object cyl. Add suitable labels to this bar chart.

28. Consider the built-in data set $mtcars$. Find the cars with hp greater than 100 and cyl equal to 3.

29. Arrange the $mtcars$ data set based on mpg variable.

30. Use the built-in data set $airquality$. Using 'select' function, select the variables $Ozone$, $Wind$, and $Temp$ in this data set.

31. Use the built-in data set $mtcars$. Rename the variables mpg and cyl with $MilesPerGallon$ and $Cylinder$, respectively.

32. Use the built-in data set iris. Using the pipe operator, group the flowers by their species and summarise the grouped data by the mean of Sepal.Length and Sepal.Width.

33. Use the built-in data set iris. Find the Species, in which Sepal.Length is greater than Petal.Length. Also Count all such Species.

34. Create a function which computes combination of two numbers.

35. Create a function which takes a natural number as an argument, and prints Fibonacci series. For example, consider fibonacci(5). It should print the first 5 elements of Fibonacci series, i.e. 1, 1, 2, 3, 5.

Additional Assignment Problems

36. Find mean, median, mode, maximum, minimum and range of the data set $\{2.3, 4.4, 5, 5, 6, 6.5, 6.5, 6.5, 6.5, 8.6, 9.2, 4.4, 4.4, 5.4, 5, 2, 4.4\}$.

37. Print the sequence from 1 to 100 with an increment 0.01. Count length of the sequence. Also find mean, median, mode, quartiles and 80th percentile of the data formed.

38. Consider eight hypothetical observations recorded under two different circumstances given as:$v_1 = \{2.3, 5.4, 3.3, 4.2, 4.2, 7.9, 4.2, 9.8\}$ and $v_2 = \{4.3, 6.4, 8.3, 6.2, 6.2, 6.9, 6.2, 8.8\}$. Find variances and standard deviations of data sets v_1 and v_2.

39. Using built in data frame CO2, compute the 10th, 30th, and 90th percentiles of all the uptakes and then use tapply to determine which treatment(chilled or non-chilled) is associated with the highest uptake.

40. What is the relative frequency of number 3 in the following set of numbers: 1,6,3,7,3,3,7,3,7,8,3,3? Use R to find it.

41. The concept of loop in any programming language is used for :

    (a) repetition of any set of arguments

    (b) formatting of conditional statements

    (c) commenting the statements for non-execution

    (d) saving the output with formatting

42. Using R studio, find arithmetic mean of the following set of observations on the age of 10 people in complete years: $19, 15, 27, 28, 14, 9, 30, 29, 20, 25$.

43. Find mode of the following set of observations on the no. of courses passed by 10 students: $10, 7, 12, 8, 11, 10, 8, 6, 9, 10$.

44. Find median of the following set of observations:$23, 22, 30, 25, 21, 32, 26, 37, 40, 30, 29$.

45. Using R, find geometric mean of $1, 25, 40$.

46. The variance of the numbers:$4, 4, 4, 4, 4, 4, 4, 4, 4$ is .....

47. Suppose that the matrix A is invertible and recalling commands to find inverse of a matrix in R and how to multiply two matrices. Write solution of the linear system $AX = B$ in R language.

48. What is the major difference between an array and a matrix?

———THE END———-