Data Analytics Lab

List of Experiments

1. Install R and then install RStudio. Get yourself acquainted with the GUI of various working windows of RStudio.
2. Create the following objects in R and then check their class
   (a). A vector of strings
   (b). A vector consisting of factor type data. For instance, a vector consisting of hair color of a few individuals.
   (c). A list data type consisting of vectors of names of five students and a matrix of the marks of students in four courses.
   (d). A data frame consisting of names of students, their age, total marks, and grades awarded
3. Apply str and summary commands to the object created in assignment 2. Interpret the output.
4. Check and justify the outcome of the following expressions:
   (a). sqrt(3) ^2 ==3
   (b). near(sqrt(3)^2,3)
5. Install, load package *'stringdist'* and run the following code:

```
my_string = c("Viraj", "Viraj", "Viraj", "Vikraj", "Viraji", "Viroj", "Vroj", "Siroji")
name "Viraj"
matched = (stringdist(my_string, name) ==0)
matched = (stringdist(my_string, name) ==1)
matched = (stringdist(my_string, name) ==2)
```

   Interpret the output. [Hint: '*stringdist*' is a package to find the distance between strings in term of replacement, insertion and declaration of letters.

6. Apply summary command to iris dataset of the *'datasets'* package and interprets the output.
7. Use plot(iris) function and interpret the output. Write down your finding about the dataset.
8. Install and load the MASS package and access the *Boston* dataset. Study the dataset from the resources available on the internet and write what you can find relevant to the dataset.
9. Write a script file to compute the following of the numeric variable in *Boston* dataset.
   (a). Sum
   (b). Range
   (c). Mean
   (d). Standard deviation

10. Create a vector x of all those values from 1:100 that are divisible by 5 and do the following operations on the vector:

(a). Find the length of vector x.
(b). Print the values stored at the fifth, tenth, and fifteenth location of vector x.
(c). Find the sum mean range median and standard deviation of vector x.
(d). Replace the fifth and tenth values with NA and NaN values, Respectively and find the mean of modified vector.
(e). Check if x contains any NA values and print the indices of NA values in vector x.
(f). Remove NA values from vector x and use summary command on it.
(g). Print the values of first and third quartile of vector x from the output of the summary command.

11. Assume the given vectors and do the following operations:

x = 1:12; y = 13:24; z = 1:6; a = 1:12;
b = (13, 15, 17, 19, 20, 21, 23, 25, 27, 29, 31, 24); c = (5, 10, 15, NA, 25, NaN);
v = (26, 21, 87, 56, 72, 60); k = (0, 2, 4, 6, 8, 16, 32)

(a). Find x × y and x × y × z and interpret the output.
(b). Do an element-wise comparison between x and a, and y and b.
(c). Find all the elements that are greater than 6 of vector x and store these elements into another vector p.
(d). Check for NA and NaN values in vectors b and c.
(e). Check if overall vector x is equal to vector a and vector b.
(f). Why does *identical(x, z)* evaluate to FALSE?
(g). What is the difference between *all()* and *all.equal()* functions? Illustrate with the help of an example.
(h). Run any(x, z) function and interpret the out.
(i). Create a new vector of the non-NA values of vector c using a single line code.
(j). Sort vector v in descending order and output the original indices in order of the sorted elements. Find log to the base 2 of vector k.

12. Assuming the character vector cv = c("sunita", "bimla", "kavita", "geeta", "anu", "dikshita", "susmita", "seema"):

(a). Find the character count in each name.
(b). Find the geeta exist in vector cv

13. Output the indices of the names that contain substring ee in vector cv of assignment 12.

14. Find out how many strings end with the letters ta in vector cv of in assignment 12.

15. Create a vector of factor type data for the hair colors of ten people where values for hair colors are black, dark brown, grey and blond.

(a). Display the levels of factor data.
(b). Find the modal value in vector of hair colors.

16. Create a vector to store the grades of 20 students for the first minor exam. Grades are given at four levels (A, B, C, D). Compare the grades for the two exams. Count the number of students who have got a higher grade in second minor.

17. Create a matrix m of five rows in a row-major order of number from 1 to 1000 incremented by step of 5 units:

(a). Find row and column-wise mean of matrix m.
(b). Find the minimum value for each row and column.
(c). Find the transpose and sort the values in each column in decreasing order.
(d). Assign the row names as R1 to R5 and column names as C1 to C4.
(e). Display all the elements of the second and forth column without using indices.
(f). Display all the elements of the first and third row without using indices.
(g). Create new matrix by deleting the second and fourth column of matrix m using indices and column names.
(h). Replace elements at indices (2, 3), (2, 4), (3, 3), and (3, 4) with NA values.
(i). Replace elements at index (1, 3) with NaN.
(j). Check if matrix m contains any NA or NaN values and interpret the output.
(k). Create two new matrices rm and cm by concatenating matrix m row-wise and column-wise with itself.

18. Interpret the output of the following commands:

(a). n = matrix(rep(m, 2), nrow = ncol(m), byrow = FLASE)
(b). n = matrix(rep(m,2), nrow = nrow(m), byrow = FALSE)
(c). n = matrix(rep(m,2), nrow = ncol(m), byrow = TRUE)
(d). m1 = do.call(rbind, replicate(2, m, simplify = FALSE))
(e). m2 = do.call(cbind, replicate(2, m, simplify = FALSE))
(f). Rename row and column names as per the requirements of matrix m1 and m2.

19. Create a matrix p of 40 rows and 16 columns of randomly assigned binary numbers.

(a). Convert each binary number to its corresponding decimal number.
(b). Scale the decimal numbers into a range of 4 to 4.
(c). Append the vector of scaled values as the last column in pop matrix.
(d). Create a new vector by evaluating the function f(x) = x^2 on the scaled values.
(e). Append the newly create vector in modified p matrix.
(f). Sort the modified p matrix on the last appended column.

20. Create a 4x3 matrix A of normally distributed random numbers with means 100 and standard deviation 10. Create another 3x4 matrix B with normally distributed random numbers with mean 10 and standard deviation 1. Perform matrix multiplication of the two matrix and store the result in third matrix C rounded up to two decimal places.

21. Create a 4x3 matrix A of uniformly distributed random integer numbers between 1 to 100. Create another 3x4 matrix B with uniformly distributed random integer number between 1 to 10. Perform matrix multiplication of the two matrices and score the result in a third matrix C.

22. Replicate the resulting matrix C obtained in Exercise 21 twice vertically.

23. Create two separate lists. The first list contains names of the five people and their age. The second contains information regarding whether they are employed or not (Y, N) and colors of their eyes. Concatenate the two lists into third list. Further, create a new list of the first and fourth elements of the third list.

24. Do the following with my_iris_data data frame created from iris data:

25. Find the dimension of the data frame.

26. Find the number of rows and columns in the data frame.

- **(a).** Display the column names of the data frame.
- **(b).** Apply summary command to the data frame and interpret the output.
- **(c).** Reshuffle the row of the data frame in random order and store the data into new data frame.
- **(d).** Find the indices of all iris flowers whose sepal length is greater than the median sepal length in the dataset.
- **(e).** Create new data frame that contains iris flowers, whose sepal length is greater than the median sepal length in the data.
- **(f).** Check the row names of the data frame created in Part (g). These are not in sequence. Reassign the row names to the data frame that is in a sequence.
- **(g).** Create three separate data frame for the three species of flowers.
- **(h).** Merge the first two species to create a new data frame.
- **(i).** From the merged dataset, create a new data frame that contains only the last three columns of the data in the data frame.
- **(j).** Check the class of last column of data frame my_iris_dat. Replace the three species ("setosa", "versicolor" and "virginica") with levels 1, 2 and 3.
- **(k).** Sort the data frame on petal length attribute, store it into a new data frame and rename the rows in a sequence.
- **(l).** Find species-wise mean of attributes of the data frame my data [Hint: Explore by() and aggregate() functions]

27. Download Mushroom dataset from the UCI repository online. Convert the file into csv format in MS Excel. Read the file into a data frame. Check the dimension of the data frame. NA values are marked by a "?" in the file. Remove all the rows of mushroom dataset that contains one or more missing values. Store the file into new data frame. Check the dimension of the new data frame. [Hint: Explore *read.csv()* function to load the file].

28. *Mushroom* data contains two kinds of mushroom (edible and poisonous). Display the number of examples of both types of mushrooms in the dataset.

29. Create a contingency table of 5^{th} and 23^{th} attribute of the mushroom dataset.

Prepared from the book Essentials of R for data Analytics, Author: Dr. Saroj Dahiya, Published: Wiley Publication