**Assignment: 8**

**Title: Using R for data preprocessing, exploratory analysis, visualization.**

**Problem Statement:** Perform following exercises to demonstrate data preprocessing, exploratory analysis, visualization using R**.**

**Prerequisites: install.packages(“tidyverse”)**

**( A )TO DEMONSTRATE DATA PREPROCESSING**

Exercise 1

**Read a table from a spreadsheet from read.table() method**

Exercise 2

**Create a dataframe using data.frame() method**

Exercise 3

**Find length of the object using length() method (hint create two list with 5 and 6 elements respectively, then create a data frame using these two lists)**

Exercise 4

**Name the columns and rows using name() method in a dataframe**

Exercise 5

**Convert the dataframe using stack**

Exercise 6

**Use omit function to omit na values using na.omit()**

Exercise 7

**Use rep method to repeat occurrences of values**

**Exercise 8**

Creating and setting factor data

Use str() to find structure of dataset

Use as.factor() to convert vector to a factor

Use level() to see splitup of the factor

Finding the number of levels using nlevels()

**Exercise 9**

Create a vector using gl(2,5,labels=c(‘A’, ‘B’))

**Exercise 10**

**Adding rows and columns to existing dataframe using $**

**( B )DATA EXPLORATION using dplyr and tidyverse**

* Select certain columns in a data frame with the **dplyr** function select.
* Select certain rows in a data frame according to filtering conditions with the **dplyr** function filter .
* Link the output of one **dplyr** function to the input of another function with the ‘pipe’ operator %>%.
* Add new columns to a data frame that are functions of existing columns with mutate.
* Use the split-apply-combine concept for data analysis.
* Use summarize, group\_by, and count to split a data frame into groups of observations, apply summary statistics for each group, and then combine the results.
* Describe the concept of a wide and a long table format and for which purpose those formats are useful.
* Describe what key-value pairs are.
* Reshape a data frame from long to wide format and back with the spread and gather commands from the **tidyr** package.
* Export a data frame to a .csv file.

**( C ) DATA VISUALIZATION using ggplot2**

Exercise 1

Create a variable “x” and attach to it the input attributes of the “iris” dataset. HINT: Use columns 1 to 4.

Exercise 2

Create a variable “y” and attach to it the output attribute of the “iris” dataset. HINT: Use column 5.

Exercise 3

Create a whisker plot (boxplot) for the variable of the first column of the “iris” dataset. HINT: Use boxplot().

Exercise 4

Now create a whisker plot for each one of the four input variables of the “iris” dataset in one image. HINT: Use par().

Exercise 5

Create a barplot to breakdown your output attribute. HINT: Use plot().

Exercise 6

Create a scatterplot matrix of the “iris” dataset using the “x” and “y” variables. HINT: Use scatterplot().

Exercise 7

Create a scatterplot matrix with ellipses around each separated group. HINT: Use plot="ellipse".

Exercise 8

Create box and whisker plots of each input variable again, but this time broken down into separated plots for each class. HINT: Use plot="box".

Exercise 9

Create a list named “scales” that includes the “x” and “y” variables and set relation to “free” for both of them. HINT: Use list()

Exercise 10

Create a density plot matrix for each attribute by class value. HINT: Use featurePlot().

**Output:**

**Refer to assignment 8 solution.pdf**

**Useful links for further reading:**

[1]Book, Mark gardner, Beginning R the statistical programming language

[2] <https://datacarpentry.org/R-ecology-lesson/03-dplyr.html>

[3]<http://www.r-exercises.com/2017/09/08/visualizing-dataset-to-apply-machine-learning-exercises-solutions/>

[4]http://www.r-exercises.com/2017/08/25/machine-learning-tutorial/