Tribhuvan University
Institute of Sciences and Technology
SCHOOL OF MATHEMATICAL SCIENCES
First Assessment 2078

Subject: **Statistical Computing with R** *Full Marks:* **45**

Course No: **MDS 503** *Pass Marks:* **22.5**

Level: **MDS /I Year /I Semester** *Time:* **2hrs**

*Candidates are required to write answers with examples for answering question numbers **1-5**in answer sheets and use laptop for answering question numbers **6-10**. R scripts and outputs interpretation of question number 6-10 must be saved in a folder with name exam roll number and submitted for grading.*

Attempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Explain how can you import following types of data into the R software with simple examples/codes:
   a) a text file saved in the local computer
   b) a table embedded in any webpage
   c) json file with web API
2. Explain the logic behind extraction of the following subsets from a 5x5 data frame in R software:
   a) First two rows
   b) Third and fifth row with second and fourth column
   c) Add 5 new rows in this data frame
3. Explain data mining in data science with focus and examples on:
   a) Tasks
   b) Analytics
   c) Learning`s
4. Explain how to work efficiently with "big data" in R software in relation to the:
   a) Subsetting with base R and dplyr packages
   b) ff, ffbase and ffbase2 packages
   c) data. table package
5. Explain social network analysis and describe its use in a real-life situation with:
   a) Nodes
   b) Links
   c) Attributes

Group B **[5 × 6 = 30]**

6. Open the Ror R studio software and do the followings with R script:
   a) Define integers from 1 to 15 using three different coding approaches in R
   b) Define these five numbers: 1.1, 2.2, 3.3, 4.4 and 5.5 and save it as column vector N
   c) Add, subtract, multiply and divide vector R from vector N and interpret the results carefully
   d) Define a listusing "This" "is" "my" "first" "programming" "in" "R"and save it as L
   e) Transform these list elements as characters of UL object.
7. Import the "pollution.csv" file into R studio and do as follows with R script:
   a) Check the structure of the data and explain class of each variable
   b) Change the attributes of "particulate matter", "date time" and "value" variables
   c) Get the summary of all the variables and repiace the outliers as missing value
   d) Get summary statistics of "value" variables by "particulate matter" variable categories
   e) Write a summary of the results obtained in the earlier steps with interpretation and conclusion

8. Use the "pollution csv" file imported and cleaned in R studio and do as follows with R script
   a) Create bar plot of "particulate matter" variable
   b) Create histogram of "value" variable
   c) Create line plot of "date time" and "value" variables
   d) Create histogram of "value" variable by particulate matter categories
   e) Write a summary of the results obtained in the earlier steps with interpretation and conclusion

9. Load the "term Doc Matrix. R data" file into R studio and do as follows with R script
   a) Define the term document matrix data object as matrix and store itas "m" object
   b) Define the frequencies of the terms using "row Sums" function and get the term frequencies
   c) Create a histogram of the term frequencies using ggplot2 package
   d) Create a histogram of the terms with 10 or more frequencies using ggplot2 package
   e) Create word cloud of term frequencies using word cloud package and interpret it carefully

OR

Load the "rdm Tweets. rdata" file in R studio and do as follows with "tm" and "tweetR" packages:

   a) Convert twitter list as data frame and assign it as "df" object
   b) Create corpus using the "text" column of the data frame
   c) Perform pre-processing to clean the corpus for text mining
   d) Create term document matrix using the cleaned corpus
   e) Find the most frequent terms using the term document matrix
   f) Find the co-occurrence of the term "r" with filter of 0.1 and above.

10. Load the "igraph" package in R studio and do the basic SNA as follows with R scripts to:
   a) Define g as graph object with (1,2) as its elements
   b) Plot the g and interpret it carefully
   c) Define g1 as graph object with ("S", "R", "R", "G", "G", "S", "S", "G", "A", "R") as its elements
   d) Plot g1 with node color as green, node size as 30, link color as red and link size as 5 and interpret it
   e) Get degree, closeness and betweenness of g1 and interpret them carefully.

OR

Load the "term Doc Matrix. R data" file into R Studio and do as follows with R script:

   a) Define term Doc Matrix as matrix m
   b) Transform it into adjacency matrix
   c) Build an undirected SNA graph with the adjacency matrix data
   d) Remove loops and plot the SNA graph again
   e) Interpret all the results carefully

***

Tribhuvan University

Institute of Science and Technology

Final Examination 2078

Subject: **Statistical Computing with R**
Course No: **MDS 503**
Level: **MDS /I Year /I Semester**

Full Marks: 45
Pass Marks: **22.5**
Time: 2hrs

*Candidates are required to write answers with examples for answering question numbers **1-5** in the answer sheet and use laptop for answering question numbers **6-10 with R scripts in R Notebook**. R scripts must be knitted as HTML or PDF with the outputs/interpretation of question number 6-10 and it must be saved in a folder with the R notebook and knitted HTML/PDF file with your exam roll number for grading.*

Attempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Describe the following concepts with focus on R software:
   a) Loops
   b) Function
   c) Pipe
2. Explain following concepts with examples focusing on R software:
   a) Big data
   b) Data wrangling
   c) Tidy data
3. Explain the following concept with examples focusing on R software:
   a) Measures of central tendency
   b) Measures of dispersion
   c) Measures of relative position
4. Explain the following concepts with examples focusing on R software:
   a) Correlation
   b) Parametric tests
   c) Non-parametric tests
5. Compare following model with focus on R software:
   a) Naïve Bayes and Support Vector Machine
   b) Decision Tree and Random Forest
   c) Feed-forward and feed-backward neural network

Group B [5 × 6 = 30]

6. Do the following in R Studio with R script so that it can be knitted as PDF:
   a) Prepare a column vector of miles per gallon (mpg) variable with random range between 10 to 50 of 500 values, **do not forget to use your exam roll number as random seed to replicate the result**
   a) Plot histogram of this "mpg" variable and interpret it carefully
   b) Refine the histogram by filling the bars with "blue" color and changing number of bins to 8
   c) Add a vertical abline at the arithmetic mean of the mpg variable
   d) Plot Q-Q plot of mpg variable, add normal Q-Q line of red color on it and interpret it carefully
   e) Plot density plot of mpg variable without the border, fill it with yellow color and interpret it

OR

Use the "ggplot2" package and do as follow in R studio:
   a) Define first layer of the ggplot object with diamond data, carat as x-axis and price as y-axis
   b) Add layer with geometric aesthetic as "point", statistics and position as "identity"
   c) Add layers with scale of y and x variables as continuous
   d) Add layer with coordinate system as Cartesian
   e) Add layer with appropriate title and interpret the resulting graph carefully

7. Do the following in R Studio with R script so that it can be knitted as PDF:
   a) Prepare a data with 100 random observations and two variables: miles per gallon (mpg) with random range between 10 to 50 and transmission gears (gear) as random binary variable (3=3 gear, 4=four gear and 5=five gears), **do not forget to use your class roll number as random seed to replicate the result**
   b) Perform goodness-of-fit test on miles per gallon (mpg) variable to check if it follows normal distribution or not
   c) Perform goodness-of-fit test on miles per gallon (mpg) variable to check if the variances of mpg are equal or not on gears variable categories
   d) Perform the best 1-way analysis of variance test based on goodness-of-fit results with justification.
   e) Can you use this test for this data? Interpret the result carefully, if applicable.

8. Do the followings in R Studio using R script so that it can be knitted as PDF:
   b) Prepare a data with 200 random observations and four variables: miles per gallon (mpg) with random range between 10 to 50; transmission (am) as random binary variable (0=automatic, 1=Manual), weight (wt) with random range of 1 to 10 and horse power (hp) with random range of 125 and 400, **do not forget to use your exam roll number as random seed to replicate the result**
   c) Divide this data into train and test datasets with 70:30 random splits **with your exam roll number as random seed for replication**
   d) Fit a supervised linear regression model for the train data
   e) Explain the model fit and BLUE coefficients for the fitted model
   f) Predict the mpg variable in the test data, get fit indices and interpret them carefully

9. Do the following in R Studio with R script so that it can be knitted as PDF
   a) Prepare a data with four random variables and 300 observations: miles per gallon (mpg) with random range between 10 to 50; transmission (am) as random binary variable (0 automatic, 1 Manual), weight (wt) with random range of 1 to 10 and horse power (hp) with random range of 125 and 400, **do not forget to use your exam roll number as random seed to replicate the result**
   b) Divide this data into train and test datasets with 80:20 random splits **with your exam roll number as random seed for replication**
   c) Fit a supervised logistic regression model on train data with transmission (am) as dependent variable and miles per gallon (mpg), horse power (hp) and weight (wt) as independent variable
   d) Predict the transmission variable in the test data and interpret the predicted result carefully
   e) Get the confusion matrix, sensitivity, specificity of the predicted model and interpret them carefully

10. Do as follows using "mtcars" dataset in R studio with R script so that it can be knitted as PDF:
   a) Check the head and the structure of the dataset
   b) Create a "cars scale" using the Principal Component Analysis (PCA) model based on nine numerical variables with centering and scaling of the variables
   c) Based on the PCA summary result, how many components must be extracted? Why?
   d) Get the bi-plot of the fitted model and interpret it carefully
   e) Improve the fitted model with VARIMAX process and interpret the results carefully

OR

Do as follows using "USArrests" dataset in R studio with R script so that it can be knitted as PDF:

   a) Get dissimilarity distance as state.dissimilarity object
   b) Fit a classical multidimensional model using the state.dissimilarity object
   c) Get the summary of the model and interpret it carefully
   d) Get the plot of the model and interpret it carefully
   e) Compare this model with the first two components from principal component analysis in this data

***