Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Assessment 2080

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 answer sheets and use R studio for answering question numbers **6-10** in your laptop. R scripts and knitted HTML files (outputs with interpretation)of question number 6-10 must be saved in a folder with name/exam roll number and submit them for grading.*

Attempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Explain how to import these types of data in R using base R functions:
   a) Comma separated values text file
   b) Excel data file
   c) SPSS data file
2. Explain how you can do sub-setting with codes in R software:
   a) Define the 6x5 matrix and select last two rows
   b) Select third and fifth row with second and fourth column
   c) Add 3 new rows in this matrix
3. Explain differences of these terms with examples using R codes:
   a) Arrays and matrices
   b) List and factors
   c) Data frame and tibble
4. Explain the following concepts on working efficiently with "big data" in R software:
   a) Sample and model
   b) Chunk and pull
   c) Push compute to data
5. Explain different types of pipe operators with R codes and examples:
   a) Compound assignment operator
   b) Tee operator
   c) Exposition operator

Group B **[5 × 6 = 30]**

6. Do the followings with R script in R Studio:
   a) Define a column vector X with numbers between 1 and 30
   b) Define another column vector Y with cubes of X
   c) Combine the two column vectors in a new data frame called DF
   d) Get plot X and Y variables and decide which type of relationship is seen
   e) Show the code to make this a linear relationship

7. Create a function and do as follows:
   a) Define a function: "roll" of a fair "die" twice with random sampling with replacement as true
   b) Get the first roll and interpret the result
   c) Get the second roll and interpret the result
   d) Get the third roll and interpret the result
   e) Write a summary of the results obtained in the earlier steps with conclusion

Scanned with CamScanner

8. Do the following in R Studio with tidy verse package using R Script to knit *HTML* output:
   a) Define a tibble having country, year, cases and population variables with 10 random data each
   b) Transform this tibble to long format and interpret it carefully in terms of tidy data format
   c) Transform the cases variable as log of cases (LnCase) and population variable as log of population (LnPop)
   d) Create scatter plots of 1. Cases and population, 2. LnCase and population, 3. Cases and LnPop and 4. LnCa and LnPop
   e) Show the four graphs in a single graph window

9. Use the "mtcars" dataset of tidiverse package and do as follows with R script to knit HTML output:
   a) Plot histogram of mpg variable and interpret it carefully
   b) Refine the histogram by filling the bars with "blue" color and changing number of bins to 10
   c) Add a vertical abline at 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 "air quality" dataset of R to do following using base R to knit HTML output with R script:
   a) Create line plot of "Temp" with "Day" as the row index and interpret it carefully
   b) Create bar plot of "Temp" variable after defining class intervals systematically
   c) Create histogram of "Temp" variable and compare it with the bar plot of "Temp" variable
   d) Plot Normal Q-Q plot of "Temp" variable and interpret it carefully
   e) Create a scatter plot of "Temp" and "Wind" variables and interpret it carefully

10. Load the "igraph" package in R studio and do the basic SNA as follows with R scripts to knit HTML output:
   a) Define g as graph object with (1,2,3,4) as its elements
   b) Plot the g and interpret it carefully
   c) Define g1 as graph object with ("Sita", "Ram", "Rita", "Gita", "Gita", "Sita", "Sita", "Gita", "Anita", "Rita", "Ram", "Sita") as its elements
   d) Plot g1 with node color as green, node size as 20, link color as red and link size as 10 and interpret it
   e) Get degree, closeness and betweenness of g1 and interpret them carefully.

OR

Do as follows in R console and then to R Studio with R script to knit HTML outputs:
   a) Open R console and then go to Help and Manuals (in PDF) and open "An Introduction to R" file
   b) Save this file in the working directory and import this pdf file in R studio using "pdftools" package
   c) Perform pre-processing and create 'corpus' using "tm" package
   d) Find the most frequent terms and create histogram of the most frequent terms
   e) Create word cloud of the corpus, color it using rainbow or R Color Brewer package
   f) Perform topic modelling and interpret the result carefully

***

Scanned with CamScanner

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2081

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 1-5in the answer sheet and use laptop for answering question numbers 6-10 with R scripts. R scripts must be knitted as PDFwith the outputs/interpretation of question number 6-10 and it must be saved in a folder with the PDF file/s alongside the name/class roll number for grading.

Attempt ALL Questions.

Group A [5 × 3 = 15]

1. Describe data visualization with focus on:
   a) Concept of grammar of graphics with Wilkinson's approach
   b) Layers in grammar of graphics with ggplot package's approach
   c) Statistical transformations in grammar of graphics
2. Describe followings for checking fit of the multiple linear regression model:
   a) Outliers
   b) Cook's distance
   c) Leverage
3. Describe supervised learning **classification** regression model with focus on:
   a) Model fit indices
   b) Confusion matrix with an example
   c) Prediction accuracy with ROC curve
4. Describe following with example on it use:
   a) Poisson regression
   b) Zero-inflated Poisson regression
   c) Negative binomial regression
5. Describe supervised linear regression model with focus on:
   a) Cross-validation
   b) K-fold cross-validation
   c) Repeated k-fold cross-validation

Group B [5 × 6 = 30]

6. Do the following in R Studio using ggplot2 package with R script to knit PDF output:
   a) Create a dataset with following variables: age (10-99 years), sex (male/female), educational levels (No education/Primary/Secondary/Beyond secondary), socio-economic status (Low, Middle, High) and body mass index (14 – 38)with random 200 cases of each variable. Your roll number must be used to set the random seed.
   b) Create scatter plot of age and body mass index variables using ggplot2 package and interpret the result carefully.
   c) Create classes of body mass index variable as: <18, 18-24, 25-30, 30+ and show it as pie chart using ggplot2 package and interpret it carefully
   d) Create histogram of age variable with bin size of 15 using the ggplot2 package and interpret it carefully

Do the following in R Studio using "airquality" data set of R with R script to knit PDF output:

Scanned with CamScanner

a) Perform goodness-of-fit test on Temp variable to check if it follows normal distribution or not
b) Perform goodness-of-fit test on Temp variable by Month variable to check if the variances of mpg are equ not on am variable categories
c) Discuss which independent sample test must be used to compare "Temp" variable by "Month" vari categories based on the results obtained above
d) Perform the best independent sample statistical test for this data now and interpret the results carefully

7. Do the following in R Studio using "Arrests" dataset of **car** package with R script to knit PDF output:
   a) Divide the **Arrests** data into train and test datasets with 80:20 random splits
   b) Fit a supervised logistic regression and naïve Bayes classification models on train data with "released" dependent variable and colour, age, sex, employed and citizen as independent variable
   c) Predict the released variable in the test datasets of these models and interpret the result carefully
   d) Compare and decide which classification model is better for this data

8. Do as follows using in-built "iris" dataset with R script to knit PDF output:
   a) Create a "flower scale" of first four variables of iris dataset using the Principal Component Analysis
   b) Compute the eigenvalues and interpret the PCA result carefully using Kaiser's criteria
   c) Show the Scree plot and decide on the number of components to retain with careful interpretation
   d) Revise the flower scale with 3 components using VARIMAX rotation and interpret the result carefully

OR

Do as follows using given dataset of 10 US cities in R studio with R script:

City	Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington D.C
Atlanta	0	587	1212	701	1936	604	748	2139	2182	543
Chicago	587	0	920	940	1745	1188	713	1858	1737	597
Denver	1212	920	0	879	831	1726	1631	949	1021	1494
Houston	701	940	879	0	1374	968	1420	1645	1891	1220
Los Angeles	1936	1745	831	1374	0	2339	2451	347	959	2300
Miami	604	1188	1726	968	2339	0	1092	2594	2734	923
New York	748	713	1631	1420	2451	1092	0	2571	2408	205
San Francisco	2139	1858	949	1645	347	2594	2571	0	678	2442
Seattle	2182	1737	1021	1891	959	2734	2408	678	0	2329
Washington D.C	543	597	1494	1220	2300	923	205	2442	2329	0

   a) Get dissimilarity distance as city. dissimilarity object
   b) Fit a classical multidimensional model using the city. dissimilarity object
   c) Get the summary of the model and interpret it carefully
   d) Get the bi-plot of the model and interpret it carefully

9. Use the first four variables of the "iris" data and do as follows with R Script to knit PDF output:
   a) Fit a hierarchical clustering model using single linkage and get the dendogram for this model
   b) Fit a hierarchical clustering model using complete linkage and get the dendogram for this model
   c) Fit a hierarchical clustering model using average linkage and get the dendogram for this model
   d) Find the best hierarchical clustering model model for this data and locate the number of clusters for it

OR

Use the first four variables of "iris" data file into R Studio and do as follows with R script to knit PDF output:

a) Fit a k-means clustering model in the data with k=2 and k=3
b) Plot the clusters formed with k=3 in the single graph and interpret them carefully
c) Add cluster centers for the plot of clusters formed with k=3 and interpret it carefully
d) Compare the k=3 clusters with Species variable using confusion matrix and interpret the result carefully

***

Scanned with CamScanner

Here are the sample questions for first assessment of MDS 503 course:

Five questions of 3 marks like this one (to be done in paper):

1. Explain how can you import following types of text files into the R software with simple codes:

a) Tab separated text file
b) Comma separated value text file
c) Semi colon separated text file

Five questions of 6 marks like this one (to be done in computer):

6. Import the "covid_tbl.csv" file in R studio as data frame and do as follows with R script there:

a) Check the structure of the data frame
b) View the data frame: remove the first row and last column
c) Change column names by adding underscore for the spaces
d) Remove "+" and "%" from the columns where they appear
e) Change attributes of the number variables from characters to numbers

The 6 marks questions will have TWO extra/or questions for TWO selected questions.

Best wishes!

Scanned with CamScanner

03-2080✡

Tribhuvan University
Institute of Science and Technology
2080

✡

bject: **Statistical Computing with R** — *Full Marks*: 45
urse No: **MDS 503** — *Pass Marks*: 22.5
el: **MDS /I Year /I Semester** — *Time*: 2hrs

ndidates are required to write answers with examples for answering question numbers 1-5in the answer sheet and use top foranswering question numbers 6-10 with R scripts. R scripts must be knitted as PDFwith the puts/interpretation of question number 6-10 and it must be saved in a folder with the R script and the name/exam roll nber for grading.

tempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Explain these terms with examples for R:
   a) Getting multi-way table with array
   b) Creating class intervals of continuous variable
   c) Missingness vs nothingness

2. Explain following concepts with focus on R software:
   a) Raw data
   b) Data wrangling
   c) Tidy data

3. Explain the followings with examples for R:
   a) Reference range based on mean
   b) Reference range based on median
   c) Outliers and extreme values

4. Explain the following concepts with focus on R software:
   a) Test of normality
   b) Parametric tests
   c) Residual analysis

5. Describe decision tree classification model with focus on:
   a) Bagging
   b) Improved bagging
   c) Boosting

Group B **[5 × 6 = 30]**

6. Do the following in R Studio with R script:
   a) Create a dataset with following variables: age (18-99 years), sex (male/female), educational levels (No education/Primary/Secondary/Beyond secondary), socio-economic status (Low, Middle, High) and body mass index (14 – 38) with 150 random cases of each variable. Your exam roll number must be used to set the random seed.
   b) Show a sub-divided bar diagram of body mass index variable by sex and socio-economic variables separately with interpretations.

Scanned with CamScanner

c) Show multiple bar diagram of age variable withsexand educational level variablesand interpret carefully
d) Show boxplots of age and body mass index variable separately and interpret the results carefully
e) Create histogram of age and body mass index variable separately and interpret the results carefully

7. Do the following in R studio and with R script to knit HTML output:
a) Define an object "rating" with 9, 2, 5, 8, 6, 1, 3, 2, 8, 4, 6, 8, 7, 1, 2, 6, 10, 5, 6, 9, 6, 2, 4, 7 values.
b) Replicate the given table obtained from SPSS software for the rating object in R.

rating

Valid	Frequency	Percent	Valid Percent	Cumulative Percent
1	2	8.3	8.3	8.3
2	4	16.7	16.7	25.0
3	1	4.2	4.2	29.2
4	2	8.3	8.3	37.5
5	2	8.3	8.3	45.8
6	5	20.8	20.8	66.7
7	2	8.3	8.3	75.0
8	3	12.5	12.5	87.5
9	2	8.3	8.3	95.8
10	1	4.2	4.2	100.0
Total	24	100.0	100.0	

8. Do the following in R Studio with R script:
a) Create a dataset with following variables: age (18-99 years), sex (male/female), educational levels (No education/Primary/Secondary/Beyond secondary), socio-economic status (Low, Middle, High) and body mass index (14 – 38) with random 250 cases of each variable. Your exam roll number must be used to set the random seed.
b) Create scatterplot of age and body mass index variable and interpret it carefully.
c) Which correlation coefficient must be used based on the interpretation of the scatterplot? Why?
d) Compute the best correlation coefficient identified from the scatterplot and interpret it carefully.
e) Test whether this correlation coefficient is statistically valid or not and justify its value.

OR

Do the following in R Studio with R script:
a) Create a dataset with following variables: age (18-99 years), sex (male/female), educational levels (No education/Primary/Secondary/Beyond secondary), socio-economic status (Low, Middle, High) and body mass index (14 – 38) with random 250 cases of each variable. Your exam roll number must be used to set the random seed.
b) Check if body mass index variable follows normal distribution using suggestive plot and confirmative tests and interpret the results carefully.
c) Check if body mass index variables have equal variance for sex variable using suggestive plot and confirmatory test and interpret the results carefully.
d) Which independent sample t-test must be used to compare body mass index by sex? Why?
e) Perform the independent sample t-test identified above and interpret it carefully.

Scanned with CamScanner

9. Do the following in R Studio using "mtcars" dataset with R script:
   a) Divide the mtcars data into train and test datasets with 70:30 random splits.
   b) Fit a supervised logistic regression model and naïve bayes classification models on train data with transmission (am) as dependent variable and miles per gallon, displacement (disp), horse power (hp) and weight (wt) as independent variable.
   c) Predict the transmission (am) variable in the test data for both the models and interpret the result carefully
   d) Get the confusion matrix, sensitivity, specificity of both the models using predicted transmission variable on test data and interpret them carefully.
   e) Which supervised classification model is the best for doing prediction? Why?

10. Do as follows using given dataset of 10 US cities in R studio with R script:

City	Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington D.C
Atlanta	0	587	1212	701	1936	604	748	2139	2182	543
Chicago	587	0	920	940	1745	1188	713	1858	1737	597
Denver	1212	920	0	879	831	1726	1631	949	1021	1494
Houston	701	940	879	0	1374	968	1420	1645	1891	1220
Los Angeles	1936	1745	831	1374	0	2339	2451	347	959	2300
Miami	604	1188	1726	968	2339	0	1092	2594	2734	923
New York	748	713	1631	1420	2451	1092	0	2571	2408	205
San Francisco	2139	1858	949	1645	347	2594	2571	0	678	2442
Seattle	2182	1737	1021	1891	959	2734	2408	678	0	2329
Washington D.C	543	597	1494	1220	2300	923	205	2442	2329	0

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

OR

Use the first four variablesof "iris" data file into R Studio and do as follows with R script:

   a) Fit a k-means clustering model in the data with k=2 and k=3.
   b) Plot the clusters formed with k=3 in the single graph and interpret them carefully.
   c) Add cluster centers for the plot of clusters formed with k=3 above and interpret it carefully.
   d) Compare the k=3 cluster variable with Species variable of iris data using confusion matrix and interpre the result carefully.

Scanned with CamScanner

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

Scanned with CamScanner

Group B [5 × 6 = 30]

following in R Studio with R script so that it can be knitted as PDF:

epare a column vector of miles per gallon (mpg) variable with random range between 10 to 50 of 500 alues, **do not forget to use your exam roll number as random seed to replicate the result**

Plot histogram of this "mpg" variable and interpret it carefully

Refine the histogram by filling the bars with "blue" color and changing number of bins to 8

Add a vertical abline at the arithmetic mean of the mpg variable

Plot Q-Q plot of mpg variable, add normal Q-Q line of red color on it and interpret it carefully

Plot density plot of mpg variable without the border, fill it with yellow color and interpret it

OR

e the "ggplot2" package and do as follow in R studio:

Define first layer of the ggplot object with diamond data, carat as x-axis and price as y-axis

Add layer with geometric aesthetic as "point", statistics and position as "identity"

Add layers with scale of y and x variables as continuous

Add layer with coordinate system as Cartesian

Add layer with appropriate title and interpret the resulting graph carefully

o the following in R Studio with R script so that it can be knitted as PDF:

) 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**

) Perform goodness-of-fit test on miles per gallon (mpg) variable to check if it follows normal distribution or not

) 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

l) Perform the best 1-way analysis of variance test based on goodness-of-fit results with justification.

:) Can you use this test for this data? Interpret the result carefully, if applicable.

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

Scanned with CamScanner

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

***

Scanned with CamScanner

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 1-5m 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 replace the outliers as missing value → df$x[df$x > mean(df$x) + 3×sd(df$x)] ← NA
   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

Scanned with CamScanner

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

Scanned with CamScanner

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 rand range between 10 to 50, transmission (am) as random binary variable (0 automatic, 1 Manual), we (wt) with random range of 1 to 10 and horse power (hp) with random range of 125 and 400, do not forg 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 random seed for replication

c) Fit a supervised logistic regression model on train data with transmission (am) as dependent variable ar 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 numeric 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

***

Scanned with CamScanner

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Reassessment 2081

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 answer sheets and use R studio for answering question numbers 6-10 in your laptop. R scripts and knitted HTML files (outputs with interpretation) of question number 6-10 must be saved in a folder with name/exam roll number and submit them 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 following data types in R with examples and R codes:
   a) Numeric and integer
   b) Categorical and factor
   c) Data and Date as well as time

3. Explain data mining in data science with focus and examples on:
   a) Descriptive mining
   b) Predictive mining
   c) Prescriptive mining

4. Explain how to work efficiently with "big data" in R software in relation to the:
   d) Sub setting with base R and dplyr packages
   e) ff, ffbase and ffbase2 packages
   f) 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 R or R studio software and do the followings with R script to produce HTML output:
   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 list using "This" "is" "my" "first" "programming" "in" "R"and save it as L

7. Import the "air quality" data as "aq" object in R studio and do as follows in R script for HTML output:
   a) Check the structure of the aqdataset and explain class of each variable
   b) Explain how to handle missing value of two variables of aq dataset
   c) Get the summary of all the variables and interpret them carefully
   d) Get summary statistics of "temp" variables by "Month" categories and interpret it carefully

8. Use the "aq"data file in R studio and do as follows with R script to produce HTML outputs:
   a) Create bar plot of "Month" variable and interpret it carefully
   b) Create histogram of "Temp" variable and interpret it carefully
   c) Create line plot of "Temp" and "Day" variables and interpret it carefully
   d) Create scatterplot of "Ozone" and "Solar.R" variables and interpret it carefully

OR

Use the "mtcars" dataset of tidiverse package and do as follows with R script to knit HTML output:
   a) Plot histogram of mpg variable and interpret it carefully
   b) Add a vertical abline at mean and ± 2 standard deviation of mpg variable and interpret it carefully
   c) Locate mode of mpg variable graphically and interpret it carefully
   d) Locate median of mpg variable graphical and interpret it carefully

9. Do the following in R Studio with tidyverse package using R Script to knit HTML output:
   a) Define a tibble having country, year, cases and population variables with **20 random data** each
   b) Transform the cases variable as log of cases (LnCase) and population variable as log of population (LnPop)
   c) Create scatterplots of 1. Cases and population, 2. LnCase and population, 3. Cases and LnPop and 4. LnCase and LnPop
   d) Show the four graphs in a single graph window

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

OR

Do as follows in R console and then to R Studio with R script to knit HTML outputs:
   a) Open R console and then go to Help and Manuals (in PDF) and open "An Introduction to R" file
   b) Save this file in the working directory and import this pdf file in R studio using "pdftools" package
   c) Perform pre-processing and create 'corpus' using "tm" package
   d) Find the most frequent terms and create histogram of the most frequent terms

Note: Save the R scripts and all the results of Group B questions in a folder with your name/roll number!

***

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Reassessment 2081

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**. R scripts must be knitted as PDF with the outputs interpretation of question number 6-10 and it must be saved in a folder with the PDF file/s alongside the name/class roll number for grading.*

Attempt ALL Questions.

Group A [5 × 3 = 15]

1. Compare these two concept of grammar of graphics with examples for:
   a) Wilkinson's approach and ggplot2 package's approach
   b) Cartesian and polar coordinates in ggplot2 package

2. Describe followings residual diagnosis for linear regression mode:
   a) Hat values
   b) Cook's distance

3. Describe supervised learning **classification** regression model with:
   a) Confusion matrix, accuracy and misclassification error with an example
   b) Area Under Curve as accuracy with ROC curve with the same example
4. Compare the two concepts following with a clear example:
   a) Simple linear correlation and simple linear regression
   b) BLUE and LINE tests

5. Compare supervised linear regression models with focus on:
   a) K-nearest neighborhood and decision tree models
   b) Support vector machine and ANN MLP models

Group B [5 × 6 = 30]

6. Do the following in R Studio using ggplot2 package with R script to knit PDF output:
   a) Create a dataset with following variables: age (20-59 years), height (110 – 190 centimeters), weight (40-90 kg) with random 150 cases of each variable. Your roll number must be used to set the random seed.
   b) Compute body mass index (BMI) variable as: BMI = [(weight in kg) / (height in centimeter **squared**)]
   c) Create body mass index categories: <18, 18-24, 25-30, 30+ and label them as "underweight", "normal", "overweight" and "obese" respectively using dplyr package
   d) Show the percentage distribution of labelled BMI variable with pie chart using ggplot2 package
7. Do the following in R Studio using "airquality" dataset of R with R script to knit PDF output:
   a) Perform test of normality of "Wind" variable by each categories of "Month" variable interpret the results carefully
   b) Perform test of equality of variance on "Wind" variable by each categories of Month variable and interpret the result carefully
   c) Perform the best independent sample statistical test for this data now and interpret the result carefully
   d) Perform the post-hoc test for this data and interpret the results carefully

8. Do the following in R Studio using "Arrests" dataset of **car** package with R script to knit PDF output:
   a) Divide the **Arrests** data into train and test datasets with 70:30 random splits
   b) Fit a supervised logistic regression and naïve Bayes classification models on train data with "released" as dependent variable and all other variables as independent variable
   c) Predict the released variable in the test datasets of **both the models** and interpret the result carefully
   d) Compare and decide which classification model is better for this data

9. Do as follows using in-built "USArrests" dataset with R script to knit PDF output:
   a) Create a "criminality scale" of four variables of this dataset using the Principal Component Analysis
   b) Compute the eigenvalues and interpret the PCA result carefully using Kaiser's criteria
   c) Show the Scree plot and decide on the number of components to retain with careful interpretation
   d) Revise the criminality scale with 2 components using VARIMAX rotation and interpret the result carefully

OR

Do as follows using given dataset of 10 US cities in R studio with R script:

City	Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington D.C
Atlanta	0	587	1212	701	1936	604	748	2139	2182	543
Chicago	587	0	920	940	1745	1188	713	1858	1737	597
Denver	1212	920	0	879	831	1726	1631	949	1021	1494
Houston	701	940	879	0	1374	968	1420	1645	1891	1220
Los Angeles	1936	1745	831	1374	0	2339	2451	347	959	2300
Miami	604	1188	1726	968	2339	0	1092	2594	2734	923
New York	748	713	1631	1420	2451	1092	0	2571	2408	205
San Francisco	2139	1858	949	1645	347	2594	2571	0	678	2442
Seattle	2182	1737	1021	1891	959	2734	2408	678	0	2329
Washington D.C	543	597	1494	1220	2300	923	205	2442	2329	0

   a) Get dissimilarity distance as city.dissimilarity object
   b) Fit a classical multidimensional model using the city.dissimilarity object
   c) Get the summary of the model and interpret it carefully
   d) Get the bi-plot of the model and interpret it carefully

10. Use the four variables of the "USArrests" data and do as follows with R Script to knit PDF output:
   a) Fit a hierarchical clustering model using single linkage and get the dendogram for this model
   b) Fit a hierarchical clustering model using complete linkage and get the dendogram for this model
   c) Fit a hierarchical clustering model using average linkage and get the dendogram for this model
   d) Find the best hierarchical clustering model model for this data and locate the number of clusters for it

OR

Use the four variables of "USArrests" data file into R Studio and do as follows with R script to knit PDF output:
   a) Fit a k-means clustering model in the data with k=2, plot it in the single graph and interpret it carefully
   b) Add cluster centers for the plot of clusters formed with k=2 and interpret it carefully
   c) Fit a k-means clustering model in the data with k=3, plot it in the single graph and interpret it carefully
   d) Add cluster centers for the plot of clusters formed with k=3 and interpret it carefully

***