**[Subject Code and Subject Name]**

|  |  |
| --- | --- |
| **Assessment Task** | [1] |
| **College** | [College of Science and Engineering] |

**Student:** Please sign, date, and attach this cover sheet to the front of your assessment task for all hard copy submissions.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Student Family Name** | | **Student Given Name** | **JCU Student Number** | | | | | | | | |
| LI | | YANGYANG | 1 | 3 | | 7 | 6 | 3 | 3 | 3 | 0 |
| **Assessment Title** | | R Fundamentals | | | | | | | | | |
| **Due Date** | | 21/08/2020 | | | | | | | | | |
| **Lecturer Name** | | Sourav Das | | | | | | | | | |
| **Tutor Name** | |  | | | | | | | | | |
| **Student Declaration**   1. This assignment is my original work and no part has been copied/ reproduced from any other person’s work or from any other source, except where acknowledgement has been made (see *Learning, Teaching and Assessment Policy 5.1*). 2. This work has not been submitted previously for assessment and received a grade OR concurrently for assessment, either in whole or part, for this subject (unless part of integrated assessment design/approved by the Subject Coordinator), any other subject or any other course (see *Learning, Teaching and Assessment Policy 5.9*). 3. This assignment has not been written for me. 4. We hold a copy of this assignment and can produce a copy if requested. 5. This work may be used for the purposes of moderation and identifying plagiarism. 6. We give permission for a copy of this marked assignment to be retained by the College for benchmarking and course review and accreditation purposes.   [Learning, Teaching and Assessment Policy](https://www.jcu.edu.au/policy/learning-and-teaching/learning-teaching-and-assessment-policy) 5.1. A student who submits work containing plagiarised material for assessment will be subject to the provisions of the [Student Academic Misconduct Requirements Policy](https://www.jcu.edu.au/policy/student-services/student-academic-misconduct-requirements-policy).  **Note the definition of plagiarism and self plagiarism in the Learning, Teaching and Assessment Policy:**  **Plagiarism:** reproduction without acknowledgement of another person’s words, work or expressed thoughts from any source. The definition of words, works and thoughts includes such representations as diagrams, drawings, sketches, pictures, objects, text, lecture hand-outs, artistic works and other such expressions of ideas, but hereafter the term ‘work’ is used to embrace all of these. Plagiarism comprises not only direct copying of aspects of another person’s work but also the reproduction, even if slightly rewritten or adapted, of someone else’s ideas. In both cases, someone else’s work is presented as the student’s own. Under the Australian Copyright Act 1968 a copyright owner can take legal action in the courts against a party who has infringed their copyright.  **Self Plagiarism:** the use of one’s own previously assessed material being resubmitted without acknowledgement or citing of the original. | | | | | | | | | | | |
| **Student Signature** |  | | | | Submission Date: 21/08/2020 | | | | | | |

Assessment 1: R Fundamentals – Total marks 40

This assessment consists of 8 questions with several sub-questions. Please write your answers in the box provided. Answers outside the box won’t be marked. The questions are based on the contents of Weeks 1-4 and include fundamentals of Base R software and Measures of proximity.

Objective: This assessment has been prepared to test your knowledge on-

* Basic data types in base R.
* Preliminary operations and functions in base R.
* Computation of proximity measures using base R.
* You must write the code/use an operator to reflect the exact statement of the question. Wherever applicable you would be marked both for output and specific syntaxes.

These are based on the topic contents in Weeks 1 and 2.

1. Write the code for identifying and replacing all missing values with the mean of the known values, in an R Data Set. Marks (2)

x[is.na(x)] = mean(x, na.rm=T)

1. List the atomic vectors in R. Marks (2)

logical, integer, double, complex, character, raw

1. Write down the values for the R output for
   1. >log(-20). Mark(1)

Error: unexpected '>' in ">"

* 1. > z <- log(20)/0 ==0;. Mark (1)

FALSE

1. A. Highlight the main difference between a matrix and data.frame. Mark(1)

A data frame can contain different types of data, but matrix only have the same type of data.

B. Write down a code to construct a 50 x 20 R matrix with the first 1,000 integers. Marks (3)

M = matrix(1:1000,nrow = 50,ncol = 20)

C. Provide the code to convert this matrix to an R data.frame. Mark(1).

Mdf = as.data.frame(M)

1. Provide the following codes (not output) to-
2. Create a binary factor variable, “Smoker” with 3 “No” and 2 “Yes”, such that “No” is ordered higher than “Yes” . Marks (2)

Smoker = factor(c("No","No","No","Yes","Yes"),levels = c("No","Yes"), ordered = T)

1. Check if the value of an R numeric variable, X, is not equal to zero. Mark (0.5)

X != 0

1. Check if the value of an R character variable, X, is not equal to a vowel of the English language. Marks (2)

X != c("a","e","i","o","u")

1. Write the code to obtain the remainder when we divide a numeric vector ‘X’ by ‘Y’. Mark (1)

X%%Y

1. Multiply two R matrices, ‘X’ and ‘Y’. Mark (0.5)

X%\*%Y

1. Provide details of the following
   1. Write the code for storing the first six integers in **three different ways** in a numeric R vector called ‘Int’. Marks (3)

Int = c(1,2,3,4,5,6)

Int = (1:6)

Int = c(1L,2L,3L,4L,5L,6L)

* 1. Write the code to create an R vector named “Y” containing **each of** the first six integers 6 times. Mark (1)
     1. Write the code and output to check the length of your vector. Mark (1).

Code: Y = rep(1:6,times=5);

length(Y)

Output: 36

* 1. The following code has been compiled while entering data in R.

X <- c(“1”, “2”, “3”, “4”, “5”, “6”). Marks (2)

* + 1. Write the code and output to check if X is a factor variable
    2. Write the code to convert “X” into a numeric vector.

1. Code: is.factor(X)

Output: FALSE

1. as.numeric(X)
2. A. Write the code to enter the following matrix in R as a matrix object named “X”. Marks (2)

y = matrix(1:9, nrow=3, byrow=T)

X = y

X[2,]= y[3,]

X[3,]= y[2,]

X

B. write the code and output to print the diagonal elements of matrix X. Mark (1)

Code: diag(X)

Output: 1 8 6

C. write the code and output to print the range of elements for this matrix in R.

Range is the difference of minimum and maximum values. Marks (2)

Code: range(x)

Output: 1 9

D. write the code and output to swap rows 2 and 3 **of matrix X** and save the new matrix in an object named Y. Marks (2)

Code: Y = X

Y[2,]= X[3,]

Y[3,]= X[2,]

Y

Output: [,1] [,2] [,3]

[1,] 1 2 3

[2,] 4 5 6

[3,] 7 8 9

1. Dissimilarity measure. In this exercise you would investigate the dissimilarity between objects in a dataset.

a. Install the package “cluster” and call the data “flower” into your R session. Submit your code in the box below. Mark (1)

install.packages("cluster")

library(cluster)

flower1 = flower

* + 1. Print the dimension and the last six rows of the data.

Submit your R code and output in the box below. Mark (2)

Code: dim(flower1); tail(flower1,6)

Output: 18 8

V1 V2 V3 V4 V5 V6 V7 V8

13 1 1 0 1 2 6 20 10

14 1 1 1 4 2 11 80 30

15 1 0 0 3 2 10 40 20

16 1 0 0 4 2 18 200 60

17 1 0 0 2 2 17 150 60

18 0 0 1 2 1 5 25 10

* 1. Write the code and output to show all variable types present in the data. Mark (1)

Code: str(flower1)

Output:

'data.frame': 18 obs. of 8 variables:

$ V1: Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 2 ...

$ V2: Factor w/ 2 levels "0","1": 2 1 2 1 2 2 1 1 2 2 ...

$ V3: Factor w/ 2 levels "0","1": 2 1 1 2 1 1 1 2 1 1 ...

$ V4: Factor w/ 5 levels "1","2","3","4",..: 4 2 3 4 5 4 4 2 3 5 ...

$ V5: Ord.factor w/ 3 levels "1"<"2"<"3": 3 1 3 2 2 3 3 2 1 2 ...

$ V6: Ord.factor w/ 18 levels "1"<"2"<"3"<"4"<..: 15 3 1 16 2 12 13 7 4 14 ...

$ V7: num 25 150 150 125 20 50 40 100 25 100 ...

$ V8: num 15 50 50 50 15 40 20 15 15 60 ...

* 1. Use an appropriate R function to compute dissimilarity coefficient for this data.
     1. Justify your choice. Mark (1)

**Answer:** Daisy fuction, since the data set has different types of data.

* + 1. Submit the code for the function with the dissimilarity coefficient that you chose. Mark(1)

daisy(flower1,metric = "gower")

* + 1. How many distinct pairwise dissimilarities are computed for this data? Mark (1)

153

* + 1. What is the dissimilarity coefficient value between Flowers 2 and 13? Marks (2)

0.5248366