HEALTH DATA SCIENCE (7HMNT032W)

SEMESTER 1 (2024-25)

COURSEWORK GUIDELINES

Assessment 1: Data analysis (R programming)

General Instructions:

1. You should use R programming to solve this assignment. The submission should be a **single-word file that includes** accompanying narrative explanations. Don't forget to number your solutions [e.g. - Ans. 1a, Ans. 1b, Ans. 2a, etc.].
2. Read each question carefully to clearly understand what it demands.
3. You can use available functions in the R base package.
4. You **should not use** any external library and its functions unless it has been **explicitly** asked for a particular question.
5. **Refrain** from using any **AI Generative application** (for example, ChatGPT, RTutor, RCodePal, AskCody, etc.) to solve any part of this coursework. If a solution is found to have been generated by an AI application, you will be reported for **academic misconduct**.
6. Assignment and associated datasets are downloadable links on the Health Data Science Blackboard site -> Assessment -> Assessment1-CourseworkAssignment.
7. Refer to the **general assignment preparation and submission guidelines:** Blackboard site -> Assessment -> HDS-Assessment1_Guidelines.pdf.
8. Submission Deadline: **20th November 2024, 13:00 hrs** (GMT).

Good luck and all the best.

Q.1) When the given code snippet(s) are executed in R, will the given variable object(s) store the assigned values? If not, provide the amended code with the new output and explain your changes. **[3*4 = 12 pts.]**

a. var.a <- sum(c(1, 2, 3) + c(4, 5, "6"))
b. var_raq = chartoRaw("Welcome")
c. .1a <- "I am learning R programming."
d. Var-z <- matrix(c(1:4, c(1, 2)), nrow = 2, byrow = TRUE)

NOTE: Include comments in your program. You should not use any external R packages (libraries).

Q.2) Write a program in R (for each sub-part) to explain the usage of following functions in R Programming: [**4*5 = 20 pts]**

a. length()
b. lapply()
c. summary()
d. read.csv()
e. rbind()

NOTE: Include comments in your program. You should not use any external R packages (libraries).

Q.3) In the given code, a user enters the following values: var1 = 250, var2 = 35, var3 = 180, expecting the return value to be 286. However, it returns a value of 501. Modify this code to return the value as expected by the user. Explain your changes. **[8 pts.]**

```
var1 <- as.integer(readline("Enter a value:"))
var2 <- as.integer(readline("Enter a value:"))
var3 <- as.integer(readline("Enter a value:"))
        func.1 <- function(var1) {
                func.2 <- function(var2) {
                                var2 + var3
                                }
                var3 = 1
                var1 + func.2(var1)
                }
func.1(var1)
```

NOTE: Include comments in your program. You should not use any external R packages (libraries).

Q.4) You are part of a Data Science team in a Marine Analytics Organisation. As a consultant, you are helping a client company (works in marine conservation and are presently focusing on Abalones) to analyse an Abalone dataset to gain new insights.

[*Information*: Abalone are marine snails, considered as white gold. They are in high demand in the Asian food markets, with a plate of it costing ~ £400-500. This is a leading cause of its poaching in Africa and various other countries. You may read more about it here and here.]

Write a program which helps in the following tasks:

[2 + 4 + 4 + 4 + 4 + 2 = 20 pts]

a. Allow the user to import the dataset and add the column (attributes/feature) names.

b. Allow the user to group the data based on gender and give the count of male and female Abalones. Give the data summary of continuous attributes for the two groups.
c. Allow the user to calculate the Age for each Abalone and store these values in a new column.
d. Allow the user to discover whether the average age for male Abalone's is greater than female Abalone's or not.
e. Allow the user to discover all the Abalone's for whom, the sum of Shucked weight, Viscera weight and Shell weight is less than the Whole weight of Abalone.
f. Allow the user to find out all the Abalone's whose Diameter is greater than 0.500mm and Height is less than 0.200mm.

NOTE: Use the Abalone dataset from the UCI repository (https://archive.ics.uci.edu/ml/datasets/Abalone). Include comments in your program. If required, you may **use** external R packages (libraries).

Q.5) Write a function-based program in R, having an outer function with three arguments x, y and z. Create an inner function within the outer function, which has an argument with a constant value (3.56). The inner function should calculate product of arguments x and y, and sum the result with this constant value. The outer function should check whether the result of inner function is a prime number. If "TRUE" should multiply the result with argument z and return this value, else should return the result as derived from the inner function. **[20 pts.]**
NOTE: Include comments in your program. You should not use any external R packages (libraries).

Q.6) Being a part of a Data Science team in a Health Analytics Organisation, you have been assigned the following tasks:

a. Develop a R program which read two datasets (Q6-File1.csv and Q6-File2.csv): the first file includes clinical data, while the second file contains protein expression data. For each dataset, the program should do the following: **[2*3 = 6 pts.]**
   i. provide the number of dimensions, summarising the number of patients and the names of variables for any chosen dataset.
   ii. generate box plots for user-specified continuous variable(s).

b. Develop a R program, which matches the patient-ids from the first dataset with the reference-ids from the second dataset (reference id in the second file is combination of patient id with sample number). For each patient, calculate the arithmetic mean of expression values for each protein.

Thereafter, based on user specified threshold value, highlight patients whose mean protein expression values are above this threshold. **[14 pts.]**

NOTE: Include comments in your program. You should not use any external R packages (libraries).