University of Sri Jayewardenepura

Department of Statistics

M.Sc. in Applied Statistics Program

First Year First Semester Mid Course Unit Examination

October 2020

STA 517 3.0 Programming and Statistical Computing with R

Time allowed: **Two (2) hours**

Answer **ALL** questions.

Instructions to the Candidate:

1. This is an open book examination.
2. This question paper consists of 2 main questions on 3 pages. If any questions or any page is missing, inform the supervisor immediately at the beginning of the exam.
3. Create a new project to save your examination work. Name the project by your index number.

Question 1: 50 marks

In each case below, write an R code. Write all the necessary functions in one script file. Name your script file using your index number. Separate the answers to different parts by `# -------------------`

For example `# ------------ Q1 - Part i`

Part - i [*3 marks*]

To create the vector $(2, 2, ..2, 4, 4, \ldots 4, 6, 6, \ldots 6)$, where there are 10 occurrences of 2, 20 occurrences of 4 and 30 occurrences of 6.

Part - ii [*3 marks*]

To generate 5 random numbers from the Binomial distribution with $n = 10$ and $p = 0.8$.

Part - iii [*5 marks*]

To calculate the following

$$\sum_{j=1}^{100}(j^3 + 4j^2)$$

Part - iv [*9 marks*]

A continuous random variable X is said to have the Normal distribution with mean (μ), 5 and variance (σ^2), 9. Let $f_X(x)$ and $F_X(x)$ denote the probability density function and cumulative distribution function of X respectively. Write R codes to find

(a) $f_X(4)$

(b) $F_X(4)$

(c) $F_X^{-1}(0.5)$

Part - v [*5 marks*]

To extract non-missing values in the following vector. The name of the vector used to create the following output is "a".

```
 [1]  1.32988645 -1.33590297  0.50010472 -0.31939456  0.24308697 -0.80593682
 [7]  0.37143153  1.01947632  0.40602012 -0.55657409          NA          NA
[13] 10.00000000          NA  0.10108128  0.80321539 -0.25749675  0.64259455
[19] -1.56117411 -1.04678011  0.42701164 -1.59572963 -0.12451915  0.24530244
[25] -0.30863440  0.57673110  0.34545189 -0.20114663  0.07097780  0.08831212
[31] -1.31271112  0.72888941  2.55130940 -1.05196501          NA          NA
[37]          NA          NA          NA
```

Part - vi [*3 marks*]

Simplify the following code using the pipe operator.

```
head (extract(mtcars, 1:4), 10)
```

Note: `extract` command is from `magrittr` package in R.

Part - vii [*10 marks*]

Mr. Perera who lives in Soratha Mawatha - Wijerama wants to sell his house. He wants to decide a price for his house to list it in the market. He believes that the size of the house is one likely determinant of price. He asked from 10 homes in the neighbourhood, "what price should you ask for your home?" and the house size (in square feet). The collected data are shown below

```
   size_x price_y
1    1000     810
2    1500    1210
3    1800    1450
4    2000    1610
5    2100    1690
6    2500    2010
7    1850    1490
8    2100    1690
9    2350    1890
10   3000    2410
```

(a) Write an R code to input `size_x` and `price_y` as two separate vectors.

(b) Mr. Perera wants to compute the least squares estimates of the model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$. Write an R code to compute $\hat{\beta}_0$ and $\hat{\beta}_1$.

Help: The least squares estimates can be computed as follows.

$\hat{\beta} = (X^T X)^{-1} X^T Y$.

Where,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_n \end{pmatrix} \text{ and } X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ . & \\ . & \\ . & \\ 1 & X_n \end{pmatrix}$$

Part - viii [*12 marks*]

Write an R function to compute the sample excess kurtosis given by

$$g = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3$$

for a give vector $x = (x_1, x_2, ...x_n)$, where $\bar{x}$ is the mean of the elements of x.

Test your function for $x = 1 : 100$.

Question 2: 50 marks

For this question the outputs generated in R along with **interpretations** should be stored in a word document. Name the word document by using your index number.

Part a

Perform an exploratory data analysis using R on the `diamonds` dataset to understand

i. What are the general relationships between variable in the diamond dataset?

ii. What are the general relationships of each variable with the price of the diamonds?

iii. What variable in the diamonds dataset is most important for predicting the price of a diamond?

The data is in a particular library in R called ggplot2. You can use the following command to load the dataset.

```r
library(ggplot2)
data(diamonds)
```

To find out what information there is about the dataset, you can run the command:

```r
help(diamonds)
```

Part b

Write an R code to classify price as high ($price > 5000$) and low ($price \leq 5000$).

END