AUCKLAND UNIVERSITY OF TECHNOLOGY

STAT500: APPLIED STATISTICS - SEMESTER 1 2023

STAT500 **Assignment 1** **Due: 11:59pm on Thursday 6 April 2023**

Outline: This assignment has four questions and it's worth 10% of your final grade.

Purpose: To assess your analytical and computing skills on the material covered.

Total possible marks: 100 marks.

Instructions:

1. **Use the .Rmd file provided to write your answers**. Enter your name and student id in the 'author' field at the top of the file.
2. Save the file on disk. The filename must include 1) your lastname, 2) your firstname, and 3) your student id, e.g., if Jane Doe submits her assignment, her file must be named "Doe_Jane_123456789".
3. Finally, knit your .Rmd answer file to .pdf and submit to Canvas.
4. Fill in and sign an assessment cover-sheet which must be the very first page in the PDF. Use, e.g., Adobe Acrobat Pro on Uni computers. You can submit this cover sheet as a separate file.

IMPORTANT: Any `R` code employed to complete this assignment must be self-explanatory and must be embedded in your answer using R Markdown code chunks. Screenshots and other images will not be allowed and will be penalized. Make sure you make comments on your code for a full understanding of your answer.

Late submissions: There is a lateness penalty of 5 marks/day (or part of a day thereof), up to a maximum of 4 days, unless the student gets an approved SCA by NO later than Monday 10 April 2023. See note below.

Note: If you need an extension with no lateness penalty because, e.g., your performance has been impacted by some extenuating, unexpected, circumstances, you can submit and SCA along with relevant evidence using the submission link from the STAT500 Home page. **Bear in mind that SCA processing may take up to 5 working days**. If you have questions, contact `victor.miranda@aut.ac.nz` or `nuttanan.wichitaksorn@aut.ac.nz`.

QUESTION 1: Variables (20 marks)

(a) [5 marks] Create a variable from your student id using the following code and name your variable as 'sid'. For example, Jane Doe with the student id 123456789 will have:

```
sid <- c(1,2,3,4,5,6,7,8,9)
```

(b) [3 marks] Find the length of this variable.

(c) [2 marks] What type is this variable? Explain.

(d) [5 marks] Using R functions, find mean, mode, median, standard deviation, and variance of this variable.

(e)) [5 marks] Interpret the median and the mode. Use 2 - 3 sentences.

QUESTION 2: Working with Data (25 marks)

Run the code below, with the 'sid' variable defined as above.


```
set.seed(sum(sid))
d1 <-  data.frame(percent_cured = runif(100, min = 0, max = 100),
                  age = rep(c('child', 'adult'), each = 50))
```


(a) [5 marks] Using R, Create and add a variable called 'dose' to the data frame d1. Use $c(100, 200, 100, 200)$ with every value repeated 25 times so that the length of 'dose' matches the other two variables.

(b) [5 marks] Using R, find the summary statistics of 'percent_cured' for child and adult.

(c) [5 marks] In 2-3 sentences, describe the mean, min and max of 'percent_cured' for child and adult. You must consider your answer to (b),

(d) [10 marks] Plot 3 graphs: (1) a histogram of the variable 'percent_cured', (2) a box plot of 'percent_cured' according to 'age', and (3) a box plot of 'percent_cured' according to 'dose'. What's the median of 'percent_cured' according to 'dose'?

QUESTION 3: Normal Distribution (25 marks)

(a) [5 marks] Generate 100 random numbers from normal distribution with the mean of the 'sid' variable in Question 1 and name it as the 'sid.norm1' variable and find its mean.

(b) [5 marks] How can we assess that the 'sid.norm1' is from a normal distribution?

(c) [5 marks] Increase the random numbers to 1,000 and 10,000 with the same mean as above, and respectively call the variables, 'sid.norm2' and 'sid.norm3'.

(d) [10 marks] Compare the mean of all three variables and discuss widely. You must write a short paragraph of 4 - 5 sentences for this.

QUESTION 4: Importing Data and Data Analysis (30 marks)

First, complete the following tasks:

1. Go to the COVID-19 data portal by Statistics New Zealand website at

   `https://www.stats.govt.nz/experimental/covid-19-data-portal`

and click at the orange "DOWNLOAD DATA" button (next to ABOUT) around the middle of the page.

2. Choose **any two** indicators of your own choice. You can select one at at time and download its data.

3. Once you successfully download each indicator, delete the "metadata" sheet. Then, save the file in an appropriate folder.

4. Import the dataset for each indicator into R.

Questions:

(a) [5 marks] For the completion of the four steps above

(b) [10 marks] **For each indicator**, write down the variable type AND explain why you chose it (5 marks p/ indicator).

(c) [10 marks] Use an appropriate plot to visualize each indicator/variable Explain the selected plot/s in 2 - 3 sentences (5 marks p/ indicator including explanation).

(d) [5 marks] Choose one of your indicators. Make an appropriate plot to see its distribution. Then discuss this distribution in 3 - 4 sentences.

** END OF ASSIGNMENT 1 **