STATS 782, Semester One, 2022
Assignment 1

Instructions

1. Please read these instructions carefully.
2. Submit two files to Canvas. (1) Your source code — either a .R or .Rmd file that works when run in R or knitted respectively. (2) A final PDF containing your code and answers. Generally, the marker will only read (2) unless there is a problem.
3. Note the time difference between countries if you're not in New Zealand.
4. Coversheet: please make sure you do one of the following else your assignment will not be marked: (a) Sign the Cover Sheet and combine with your assignment document (pdf or Word) into a single file before submission, OR (b) Type or write for the following at the beginning of your assignment: Your name (as it appears in Canvas), your UPI, and the following statement: "I have read the declaration on the cover sheet and confirm my agreement with it."
5. Please comment on almost all of your output, especially parts that need human interpretation, or marks will be deducted. That is, you need to convince the marker that you understand what the solution is doing.
6. Comment your code if appropriate, e.g., for functions, blocks of code, and key variables.
7. You may occasionally need to look online or in R's help documentation for details (e.g., about functions) that are not found in the coursebook or lectures.
8. Your mark for this assignment will depend on getting the right answer, the elegance/efficiency of your approach, and the tidiness and documentation of your code/report. Marks will be deducted for messy code, etc.

Question 1 [15 marks]

Write R code to generate vectors containing the given sequences of values. Do not use `c()` or loops.

(a) [3 marks]

```
[1] 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
```

(b) [3 marks]

```
[1] 1 2 3 4 1 2 3 4 1 2 3 4
```

(c) [3 marks]

```
[1] "a 1" "b 2" "c 3" "d 4" "e 5" "f 6" "g 7"
```

(d) [3 marks]

```
[1] 0.01 0.10 1.00 0.10 0.01
```

(e) [3 marks]

```
[1] "AaBbCcDdEeFfGgHhIiJjKkLlMmNnOoPpQqRrSsTtUuVvWwXxYyZz"
```

Question 2 [15 marks]

Suppose the number of motorcycle accidents in a region each month follows a Poisson distribution with $\lambda = 5$. Use R's `dpois()` and similar functions to:

(a) [3 marks] Plot the probability mass function from $x = 0$ and $x = 20$.

(b) [3 marks] Calculate the probability that the number of accidents in a given month is either 8, 9, or 10, using `ppois()`.

(c) [4 marks] Redo (b) using `dpois()` and `sum()`.

(d) [5 marks] Write a small function that simulates the number of accidents in a month. By calling this function 10000 times, calculate the proportion of simulated months with either 8, 9, or 10 accidents. This is a 'Monte Carlo' method for redoing (b) and (c).

Question 3 [15 marks]

Write an R function called `analyse_text()` that takes a single string (character vector of length 1) as input, and returns a named numeric vector with three values in it: the number of words, the length of the longest word, and the mean word length.

The function should behave as follows when called:

```
> analyse_text("Hello!")
      num_words  max_word_length mean_word_length
              1                5                5
> analyse_text("The quick      brown fox jumps over the lazy dog.")
      num_words  max_word_length mean_word_length
       9.000000         5.000000         3.888889
```

Note how multiple spaces do not matter, and punctuation also does not count. **Hint**: The function `gsub()` can be used to remove certain characters, and `strsplit()` to split a string into several parts.