UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
STA5075Z - Data Science Masters Statistical Computing
Assignment 1 2023

Due: Tuesday 21 February 2023 @ 11:55pm

IMPORTANT!!!

The data you are going to use are in the CSV file called "bugs". Do not manipulate the Excel file. All data manipulations need to be made in R. Use the "SC-Assignment-1-template.Rmd" template provided to produce a pdf report. Knit your .Rmd file to pdf and upload both the .Rmd file and the rendered pdf file. Please include the options 'eval = T' in all of your chunk options. The marker needs to be able to run this Rmd file and get the answers that you submit.

Question 1 of 1 [20 marks]

The severity of software bugs can be classified as Low, Non trivial, Major, Critical and High. Five software packages were tested for bugs in each of their core environments before and after Beta testing of the software package. In each test the number of each kind of bug was noted.

a. Read the Bug data into `R` and print out the first 10 observations. [2]

b. During data entry there was a mistake. Counts which were meant to be 10 were mistakenly entered in as zeros. Identify these observations. [2]

c. Correct the data entry mistake mentioned in the question above. Use `summary()` to produce the five number summary for the variable (`Count`) after cleaning. [2]

d. Create a plot showing the total number of bugs identified for each software. [2]

e. What was the highest number of bugs found after Beta testing? [2]

f. For PDE software create a table showing the total number of bugs of each type identified before and after the Beta test. Clearly label your rows and columns for the table. [3]

g. Add one to all the count values and take the logarithm, i.e., compute $log(Count + 1)$. What is the mean value over all the entries? Give your answer to three decimal places. [2]

h. For each software we can estimate λ the rate of occurrence of the different kinds of bugs before and after Beta testing as

$$\lambda = \frac{k}{n}$$

where:
k number of occurrences of bug type for software.
n total number of errors for software.

Write a function `queryRate()` that can be used in general to query data sets of this type and report the error rate λ for each software type, error severity and Beta test timing combination. The function should have arguments `softwareType`, `errorSeverity` and `beta` to specify respectively the software type, bug type and Beta testing status of a particular error rate (λ) query. The function should return values for k, n and λ as output. Subsequently use the function to query the rate for Low errors in the PDE software before Beta testing. [5]

‘