Assignment Q1 2021

301113 Programming for Data Science

Supplementary Assignment

Due Sunday 16th of May 2021

Assignment description

This assignment has four questions of equal value. For each question you need to:

- Write appropriate R code; you must *only* use R to do this assignment
- *Do **not** use any R packages other than those part of default R; i.e. only use functions that are described in lectures and tutorials of this unit! You will lose marks if you don't comply*
- Include **all** R code used in your assignment
- Format the code so that it assists others in understanding it; e.g. use meaningful variable and function names, also make sensible use of white space / layout
- Include reasonable comments within the code, such that the basic operation is documented; what functions do and also comment on what tricky sub-sections of code do. If what a line of code does is reasonably obvious, it does not need comments, but the goal of a block of code probably deserves a comment
- Test that the code works correctly, that it handles the supplied data correctly and ideally handles any reasonable differences that could be expected in a new dataset. Hence your code should be reasonably robust; better that it returns an error than quietly does the unexpected.

Your solution to this assignment consists of two parts: a PDF file containing verbal descriptions, hence all other supporting material and a single R script file that contains *all* of your code.

The PDF file contains the following:

- Your name
- Student number
- Unit name
- Unit number
- The declaration shown on page 3 of this document, to be located on your second page
- You *must* sign the declaration
- Any descriptions that are beyond normal programming comments

- Any program output, for example, tables and plots; clearly tie such output with the code that produced it, e.g. include one or two lines of relevant code with the output

The R script file contains the following:

- Your name
- Student number
- Unit name
- Unit number
- The complete code you produced to answer the questions
- The code must agree with any descriptions found in the PDF file
- Clearly show which question code belongs to
- Organise the file in some sensible way
- Include comments as are appropriate in a program
- Comments must be meaningful; hence not stating the obvious, rather, declaring something that is not immediately apparent. Useful comments aid quick understanding of the code by someone you already understands R. Comments are short and to the point. Sometimes, a useful comment sums up the purpose of a block of code. Typically, a single comment consists of one or two lines. Don't have comments, or code for that matter, that requires horizontal scrolling to view it

Submit this solution view vUWS. Your solution will be checked for plagiarism. Submission must occur *prior* to midnight of the declared due date. Late submissions will receive a 10% reduction in marks for each day after the due date.

Marking criteria

This assignment is worth 40% of the unit assessment tasks. The four questions are each worth 10 marks. The marking criteria for each question is given in Table 1. When writing the solutions to each of the four questions, make sure to consult the marking criteria and check that you have covered the declared requirements. Therefore this assignment will be marked using this criteria.

Criteria	Q1	Q2	Q3	Q4
Code correctness (5 marks)				
Comments explaining code (2 marks)				
Code testing (1 mark)				
Code style and readability (2 marks)				
Total (10 marks)				

Table 1

Declaration

Before submitting this assignment, include the following declaration in a clearly visible and readable place, ideally on the second page of your assignment.

By including this statement, I the author of this work, verifies that:

- I hold a copy of this assignment and can produce it if requested by the unit co-ordinator
- I hereby certify that no part of this assignment has been copied from any other student's work or from any other source, except where due acknowledgement was made in this assignment
- No part of this assignment has been written for me by another person except where such collaboration has been authorised by the unit co-ordinator
- I am aware that this work may be reproduced and submitted to plagiarism detection software systems for the purpose of detecting possible plagiarism. NB that such systems may retain a copy for future plagiarism checking
- I hereby certify that I have read and understand what the School of Computing, Engineering and Mathematics, defines as minor and substantial breaches of conduct as outlined in the learning guide for this unit

Note: an examiner or lecturer / tutor has the right not to mark this assignment if the above declaration has not been added to this assignment and been signed.

Questions

Question 1

This question requires you to load the supplied data file called "iris_badvalues.txt". The product of loading this file must result in data that has identical structure to the *iris* data set provided in R; hence five columns, four of which are numeric, the remainder declaring species.

Create two or more functions for this task. The main function must have the following prototype

load.data <- function(method = c('scan', 'read.table', 'read.delim'),

filename = 'iris_badvalues.txt')

Therefore *load.data()* uses the specified method to load the data. Regardless of the method used, the end result must be identical. Also create the function

test.load.data<- function()

in order confirm whether the three methods produce identical data sets.

Question 2

This question involves identifying and removing obviously bad data. Review the data set produced by *load.data()* and determine what obvious issues exist. Describe how you reviewed the data set, declare why you used that approach and briefly describe your conclusions.

For the purposes of this question only focus on determining which records / rows of the data should be discarded. Limit your investigation to missing data, numeric values that are either less than or equal to zero, or unreasonably large. Briefly explain how you determined what is unreasonably large and why you think that determination is sensible. Write one or more functions to find records that have these issues and produce a cleaned version of the original data set.

Seek to make your code modular, so that improvements or bug fixes are made easier or cleaner to implement. In particular, have the view that you will be adding functionality, later on, to repair data that was deemed to have issues.

Produce some plots to show the difference between before and after cleaning. Provide a brief interpretation of your observations and apparent improvements to data quality obtained.

Question 3

This question involves writing a modified version of the code produced in question 2, in order to repair data that was discarded by question 2.

Create a function with the following prototype

build.repair.data <- function(df)

to generate data to be used for repairing any of the measurements (Sepal.Length, Sepal.Width, etc), but on a per species basis. Therefore the result of *build.repair.data()* would be a matrix containing 12 entries. Devise a simple but reasonable approach for generating repair data; therefore a repair value should be reasonably likely to occur in that measurement for the species in question. Briefly explain the approach used and why it is considered reasonable; do this as a non-statistician, but using basic statistical knowledge.

Using copies of the functions generated in question 2, but only those that need to change and produce code to do the following. Making use of *build.repair.data()* function, extend the functionality described in question 2, that is, repairing records that were previously discarded.

Produce some plots to show the difference between the results of question 2 and the results obtained in this question. Provide a brief interpretation of your observations and apparent improvements to data quality obtained.

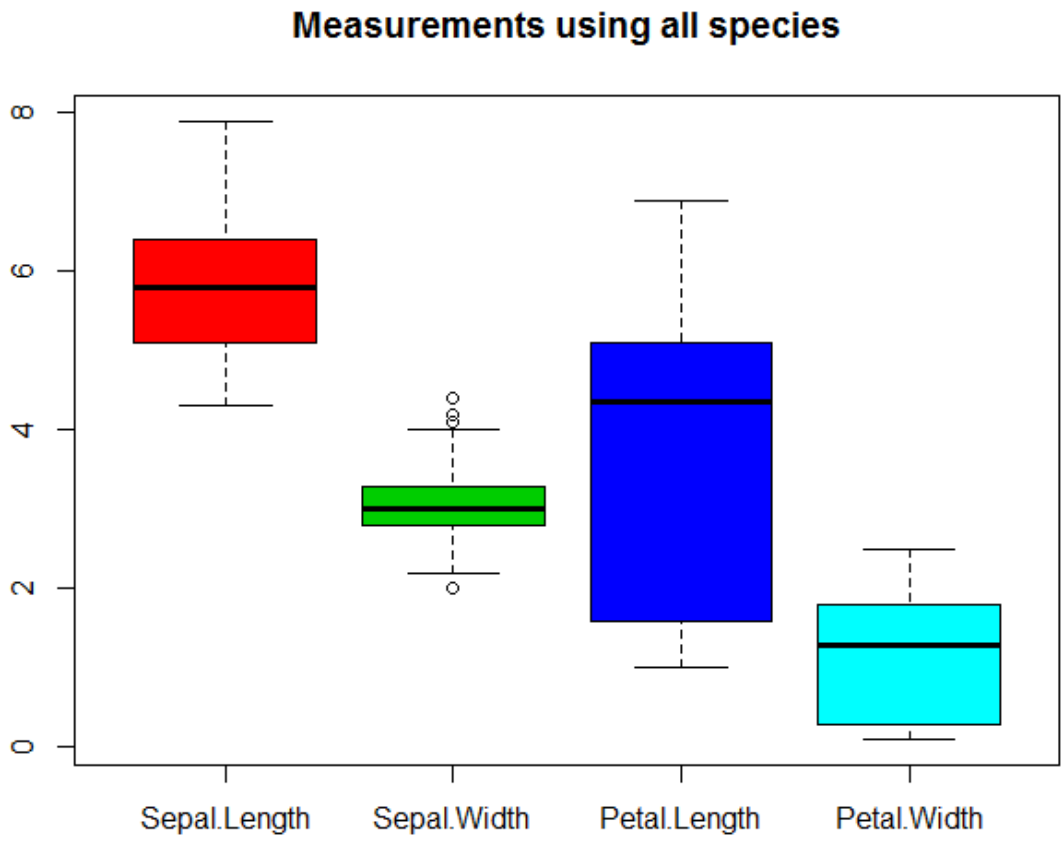
Question 4

This question consists of two sub-questions, both of equal value; hence, each receiving 50% of the advertised marks in the marking criteria. For this question you need to use the *iris* data set provided by R.

i)

Develop code to generate the two outputs shown below

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Min.	4.300000	2.000000	1.000	0.100000
1st Qu.	5.100000	2.800000	1.600	0.300000
Median	5.800000	3.000000	4.350	1.300000
Mean	5.843333	3.057333	3.758	1.199333
3rd Qu.	6.400000	3.300000	5.100	1.800000
Max.	7.900000	4.400000	6.900	2.500000



The two above outputs show the same information and were generated without considering species. Generate needed code, but *only* using the functions: cat, apply, summary and boxplot.

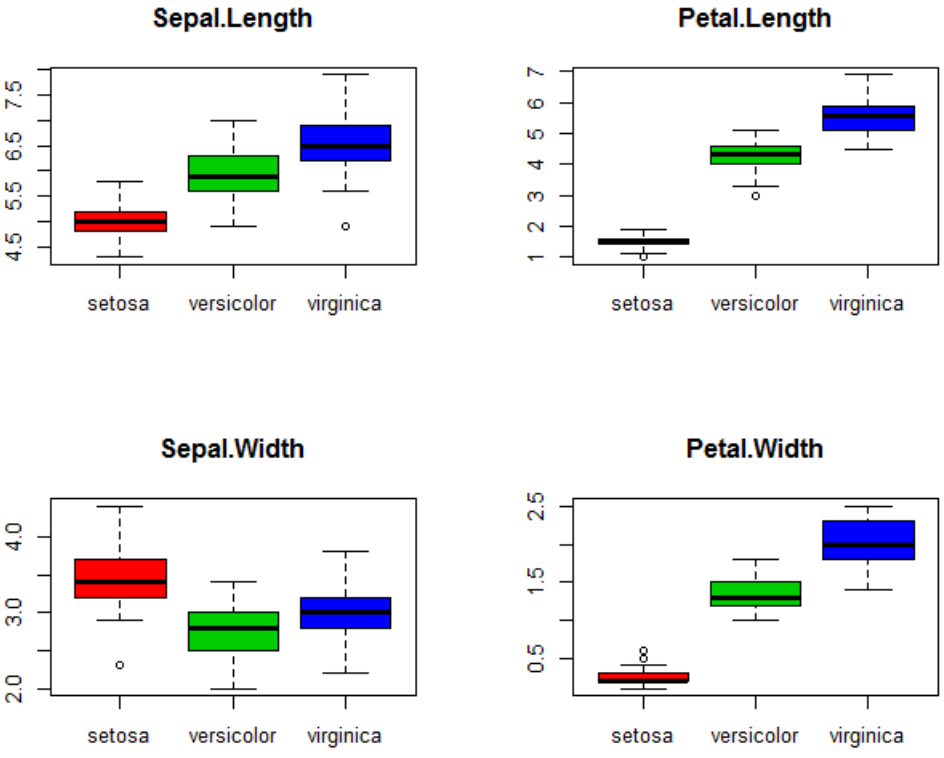
ii)

Develop code to generate the two outputs shown below

```
Species = setosa
       Sepal.Length Sepal.Width Petal.Length Petal.Width
Min.          4.300       2.300        1.000       0.100
1st Qu.       4.800       3.200        1.400       0.200
Median        5.000       3.400        1.500       0.200
Mean          5.006       3.428        1.462       0.246
3rd Qu.       5.200       3.675        1.575       0.300
Max.          5.800       4.400        1.900       0.600

Species = versicolor
       Sepal.Length Sepal.Width Petal.Length Petal.Width
Min.          4.900       2.000         3.00       1.000
1st Qu.       5.600       2.525         4.00       1.200
Median        5.900       2.800         4.35       1.300
Mean          5.936       2.770         4.26       1.326
3rd Qu.       6.300       3.000         4.60       1.500
Max.          7.000       3.400         5.10       1.800

Species = virginica
       Sepal.Length Sepal.Width Petal.Length Petal.Width
Min.          4.900       2.200        4.500       1.400
1st Qu.       6.225       2.800        5.100       1.800
Median        6.500       3.000        5.550       2.000
Mean          6.588       2.974        5.552       2.026
3rd Qu.       6.900       3.175        5.875       2.300
Max.          7.900       3.800        6.900       2.500
```



The key difference between this and *i)* is the use of data subsets corresponding to species. Structure your code so that each of the above outputs is produced by a single function; therefore two functions in total, one for the textual output and one for the graphical output. Include in your answer, the code used to execute your functions.