Programming for Data Analysis, Processing and Visualisation

Assignment 2

John O'Sullivan

Instructions:

- This assignment is due on Sunday the 21st of July at 11:59pm
- You should submit your assignment to the 'CA_TWO_(30%)' object in Moodle
- You should submit two files separately (**i.e., not a zipped folder**):
  - (i) a single .Rmd script file containing all of the commented code you used to obtain your answers
  - (ii) the HTML (or pdf) file which you produced from the .Rmd script
- The marks available for each question are shown in brackets
- You may need to find some new functions in order to do some of these tasks. Remember to use R's search engine, as well as checking online.
- Make sure that your file is readable and has a neat presentation and clear flow. The HTML (or pdf) output file should be a stand-alone document containing the answers to all questions and showing all necessary code.
- There are marks in this assignment for document presentation [10 marks] - your final document should be neat with a clear layout, showing a good use of RMarkdown to mix free-flowing text, code, and output
- I advise you to first create an R script with all of your answers. When you are happy with this, convert it piece-by-piece into an .Rmd file.

Question 1: [10 marks]

Load the **MASS** library, and access the **Cars93** dataset.

Your boss has asked you to produce *two* different plots, which are intended to be included in a report designed to communicate some of the main features of this dataset.

(i) Produce two different plots from the Cars93 dataset - each plot can be contained in a single panel, or a plot can be spread over two or more panels if necessary (these are called multipanel plots - e.g., side-by-side barcharts). Your plots should be neat and presentable, clearly labelled, and use colour and relevant legends where appropriate.

Your two plots must be of different types - e.g., don't produce two different pie charts as your answer to this question!

[6 marks]

(ii) Write 3 or 4 sentences for each plot to explain what it tells us about the dataset.

[4 marks]

Question 2: [10 marks]

The dataset **leafshape** is available from the **DAAG** library.

(i) Load the **DAAG** library and access the **leafshape** dataset contained in it. Load the help file and read about the dataset. In what country is the data collected? [1 mark]

(ii) A researcher is interested in seeing if there is a relationship between the blade length of the biomass and the location where it is growing. Create a plot containing boxplots of the blade length variable, grouped by location. [6 marks]

- colour the boxplots using different colours
- include sensible axis labels and titles
- include anything else you feel is relevant to improve the appearance of the graph

(iii) Comment on the resulting graph. What information does it illustrate? [3 marks]

Question 3: [10 marks]

(i) Set x to be 2 and y to be 10. Write a **while()** loop which prints $x+y$ and then squares x and doubles y. The loop should stop when either x is greater than or equal to 40, or y is greater than or equal to 40. What value is x now? [3 marks]

(ii) Create an array of size $3 \times 4 \times 5$. Write a **for()** loop which fills this array indicating whether the sum of the indexes for that entry is divisible by 3. e.g., the [2, 1, 3] entry should be labelled "Div.by.3" since $2 + 1 + 3 = 6$, and 6 is divisble by 3, whereas the [1, 1, 2] entry should be "Not.div.by.3" since $1 + 1 + 2 = 4$ and 4 is not divisble by 3. When finished, print the middle face (also known as a slice) of the array. [4 marks]

(iii) Set i to be 3. Write a **repeat()** loop which trebles i until i is greater than 100. What value is i now? [3 marks]

Question 4: [30 marks]

Using the **RCurl** package and the **getURL()** function (or otherwise), read in the code from the following webpage: https://www.imdb.com/chart/top. Then answer the following questions:

(i) In how many of the top 250 films does the actress Grace Kelly appear? What are the names of these films? [6 marks]

(ii) Extract all of the film titles into a vector of length 250, where the first element of the vector is the first film in the ranking, etc. Print the first 6 entries of this vector. How many films have 'A' as the first word of their title? [6 marks]

(iii) In how many of the top 250 films does the director have S as the first letter of their first name? [6 marks]

(iv) How many of the top 250 films have a score of 8.4 or greater? [6 marks]

(v) What is the mean number of user ratings used to define the top 250 films? [6 marks]

Question 5: [30 marks]

The file **Dublin.csv** contains census information on population figures for Dublin from 1841 to 2016. It contains three variables: total population count, the number of males, and the number of females.

(i) Read the dataset in and call it **dublin**. Assign to the **dublin** object the classes **pop.data** and **data.frame** (in that order). (The **read.csv()** function is the best way to read in the data.) [3 marks]

(ii) Write an S3 **summary** method for an object of class **pop.data** which displays the following statistical summaries for the Male and the Female variables: minimum, maximum, and mean population count. The years corresponding to the minima and maxima should also be printed for both variables. This summary should be neat and clear, and easy to read and understand. [10 marks]

(iii) Test your summary method by running the code **summary(dublin)**. [1 mark]

(iv) Create an S3 **plot** method for the class **pop.data** that produces the following plot:

- A line plot (a time series plot) containing two lines to show the population trend for males and females
- By default, the plot will draw a red line for males and a blue line for females - the user must be able to change these colours if desired
- The plot must include meaningful labels for the axis and legend
- The plot should be neat and clear and be easy to interpret - pay attention to distances between the plot edge and the plotting panel on all sides, the orientation of numbers, the position of titles, the default width of lines on the plot etc.
- The method should also include a generic title by default, but allow the user to include their own title as an argument if desired.
- Your code should not 'hard-code' numbers into it unnecessarily - e.g., if a longer or shorter dataset is supplied to it, it should be able to plot this without any errors

[10 marks]

(v) Test your plot method by running **plot(dublin)**. Include a user-specified title relevant to this dataset in your arguments. [1 mark]

(vi) The file **Mayo.csv** contains similar information on population figures from 1861 to 2016 in the county of Mayo. Load the dataset, call it **mayo**, and assign it the two classes **pop.data** and **data.frame** as before. Run **summary(mayo)** and **plot(mayo)** to test your two methods (including an appropriate title for the plot). Interpret the findings, commenting on any differences or similarites between the two summaries and two plots. [5 marks]