Programming for Data Analysis, Processing and Visualisation

Assignment 2

John O'Sullivan

Instructions:

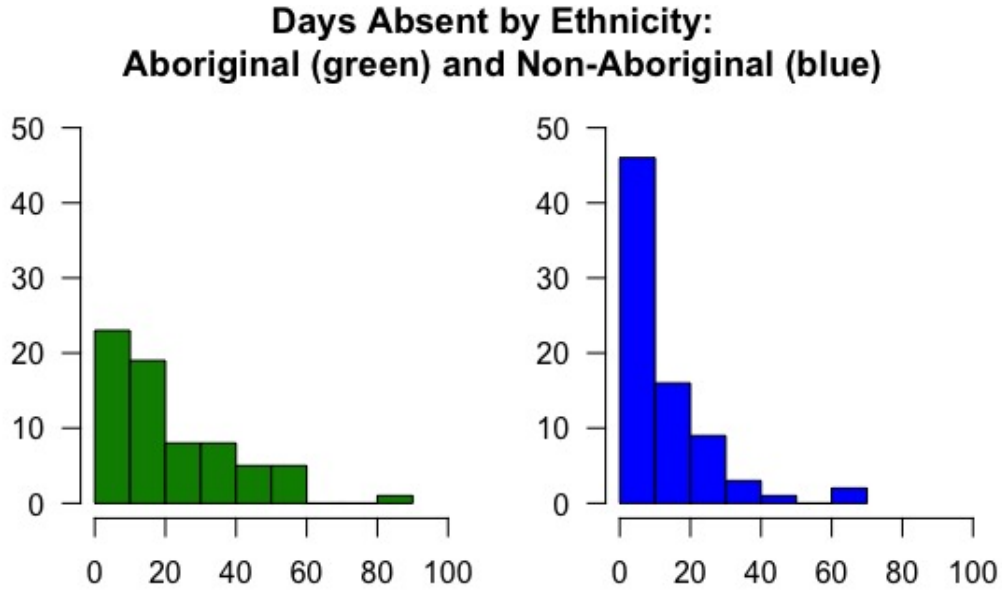
- This assignment is due on Tuesday the 18th of December at 11:59pm
- You should submit your assignment to the 'CA_TWO_(30%)' object in Moodle
- You should submit two files separately (**i.e., not a zipped folder**):
  - (i) a single .Rmd script file containing all of the commented code you used to obtain 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]

(i) Load the MASS library, and access the dataset **quine** [1 mark]

(ii) A researcher is interested in seeing if the number of days in which students were absent differs between Aboriginal and non-Aboriginal students. In order to do this, she creates the plot below, showing two histograms on the same panel. Your task is to reproduce this plot. (You can be creative and make the histograms look different if you like - but the information it displays must obviously be the same.) [7 marks]

(iii) Comment on whether you think there is any evidence of a difference between the two groups. [2 marks]

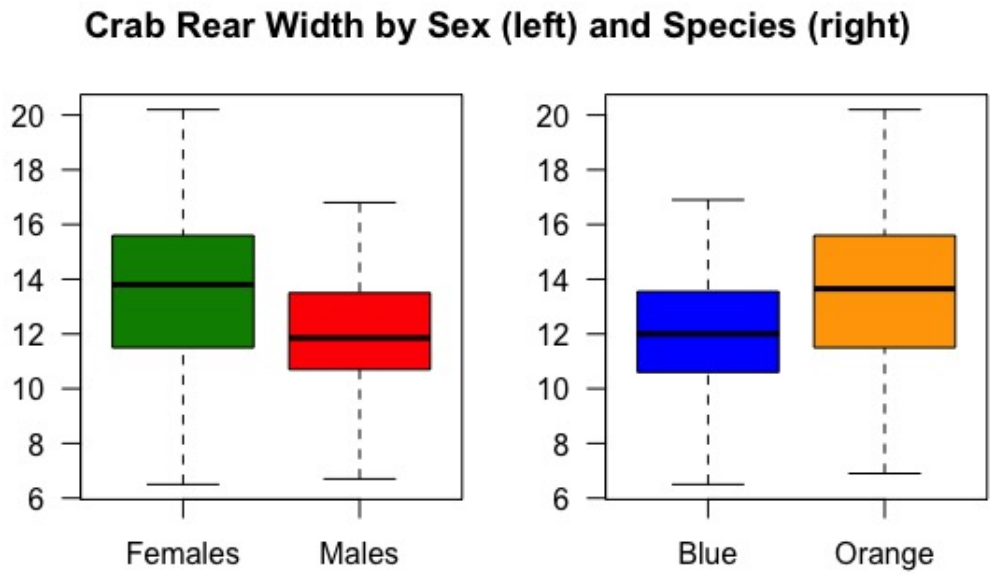


Question 2: [10 marks]

(i) Load the MASS library, and access the dataset **crabs**. Look at its structure (you may also need to load the help file to read about the dataset). [1 marks]

(ii) A marine biologist is interested in seeing what relationship the categorical variables of sex and species have with the numerical variable measuring the rear width of the crabs. He produces the plot below. Your task is to reproduce this plot. (You can be creative and make the graph look different if you like - but the information it displays must obviously be the same.) [7 marks]

(iii) Comment on the resulting plots - what can the marine biologist say about the rear width of the crabs? [2 marks]



Question 3: [10 marks]

(i) Set x to be 2, y to be 3 and z to be 4. Write a **while()** loop which prints $x + y + z$ and then doubles both x and y and adds 1 to z. The loop should stop only when x is greater than 50, or y is greater than 70, or z is greater than 10. What is the final value printed to the console? [3 marks]

(ii) Create a matrix of size 10×6. Write a double **for()** loop which fills this matrix with values equal to the sin of the row index times the cosine of the column index. e.g., the [1, 1] entry should be $\sin(1) \times \cos(1)$, the [3, 4] entry should be $\sin(3) \times \cos(4)$ etc. Print the matrix. [4 marks]

(iii) Set i to be 2. Write a **repeat()** loop which trebles i until i is greater than 200. What value is i now? [3 marks]

Question 4: [30 marks]

Download the file **imdb.txt** which contains the html webpage output of the IMDB top 250 films). Scan this in, and then answer the following questions:

(i) In how many of the top 250 films does the actress Grace Kelly appear? [5 marks]

(ii) What are the names of the films in which she appears? [5 marks]

(iii) In how many of the top 250 films does the director have B as the first initial of their first name? Are any of these directors included more than once in the list? (As always, you need to use code to answer all parts of this - not just read the answer from a table etc.) [5 marks]

(iv) How many of the top 250 films have 'A' as the first word in their title? [5 marks]

(v) How many of the top 250 films have a score greater than 8 and less than 8.4? [5 marks]

(vi) User ratings are used to define the top 250 films. What is the average number of user ratings for the 250 films? [5 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 safest way to read in the data - the **header** and **row.names** arguments will help you to read it in correctly.) [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 1841 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]