STAT40620
Data Programming with R
Assignment 2

Instructions

- This assignment is due on Wednesday 14th November 2018 at 11:59pm.
- You should submit your assignment to the 'Assignment 2' assignment object in Blackboard.
- You should submit two files only:
  1. `Rmd` file or a single `R` script file detailing the commented code you used to obtain your answers.
  2. final document in either `pdf` or `Word` which should contain answers to the questions below.
     - If you created an HTML file, please convert it to pdf. You can use Google Chrome: `File` > `Print` > `Destination [Change...]` > select `Save as PDF`.
- To get one bonus mark you must submit the `Rmd` file and the resulting document which shows all your code.
- The assignment grades are capped at a maximum of 10 marks, bonus included.
- The marks available for each question are shown in brackets.
- Assignment 2 is broken up into 3 tasks: analysis, writing your own function, and writing S3 methods.
- You may have to learn and discover some new functions. Use `help()` and `help.search()` to find what you need.

Task 1: Analysis

Download the `Lawyers.csv` data set from Blackboard. The data set contains seven variables recorded on 71 lawyers in a northeastern American law firm. Details on the seven variables, and their associated levels for the categorical variables, are given here:

Variable	Levels
Seniority	Associate
	Partner
Gender	Female
	Male
Office	Boston
	Harvard
	Providence
Years (Years in the firm)	–
Age	–
Practice	Corporate
	Litigation
School	Harvard or Yale
	Other
	University of Connecticut

1. Load the lawyers' data into R. What proportion of the lawyers practices litigation law? (Give your answer to 2 decimal places.) [0.7]

2. Is the proportion of lawyers in the Boston office that practice corporate law higher than the proportion of lawyers in the Providence office that practice corporate law? [0.5]

3. Use the `aggregate` function to compute the average age of lawyers who practice corporate law and of lawyers who practice litigation law, across the different levels of seniority. Label the columns of the resulting data frame appropriately. [0.7]

4. Which office has the youngest median age? [0.6]

Task 2: Writing your own function

Rosenbrock banana function has a multivariate input $\mathbf{x}$ and a scalar output:

$$f(\mathbf{x}) = \sum_{i=1}^{N-1} 100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2$$

Add a comment to each line of code to explain what it does exactly.

1. Write a function which compute the Rosenbrock banana function using a loop. Test the function on the vectors $\mathbf{x} = (.2, .5)$ and $\mathbf{x} = (.2, .5, .1, .6)$ [1]

2. Propose an alternative function that does not use any loop. Test the function on the same two vectors. [1]

3. Compare the timings you obtain by repeating the function calls 100 times using the vector $\mathbf{x} = (.2, .5, .1, .6)$ as input. [0.5]

Task 3: Writing S3 methods

The file `2018_09_Dublin_Airport.csv` contains the Historical Data recorded at the Dublin Airport Met Éireann Weather Observing Station in September 2018[1]. The data set contains seven variables: `date`: Date (dd-mmm-yy), `rain`: Precipitation Amount (mm), `maxtp`: Maximum Air Temperature (C), `mintp`: Minimum Air Temperature (C).

1. Load in the data as an object called `DublinAirport`. Assign to the `DublinAiport` object the classes `WeatherData` and `data.frame`. [0.2]

2. Write an S3 summary method for an object of class `WeatherData` which produces the following statistical summaries for the `rain`, `maxtp`, `mintp` variables: mean, standard deviation, minimum, maximum. [1]

3. Download the new data set `2018_09_Cork_Airport.csv` from Blackboard, assign the classes `WeatherData` and `data.frame` to the object containing the Cork data, and test your function on it. Interpret your findings for Dublin and Cork Airports. [0.5]

4. Create an S3 plot method for the class `WeatherData` that produces the following plots.

   - Two plots must be on a single panel, one above the other. Only the plot on the top panel will contain a main title. [0.3]
   - The plot on the top is about the daily Air Temperature (C). It must include the following: [1.5]
     - lines plot to show the daily air temperatures
     - by default the plot will draw a red line for the maximum temperatures and a blue line for the minumum temperatures. The user must be able to change these colors
     - the plot must include meaningful labels for the axis and legend
     - the plot must include a grey vertical dotted line for each day to clearly identify the day corresponding to each couple of points. (hint: see the `abline` function)
     - the plot by default should allow the user to identify clearly the noteworthy points by adding a point character over the value of the highest maximum temperature registered and a point character over the value of the lowest minimum temperature registered. The user must be able to decide to avoid to add the point characters to the plot.
   - The plot on the bottom is about the daily Precipitation Amount (mm). It must include the following: [1]
     - vertical-line plot to show the daily precipitation amount (hint: look at the help file of `plot` to see which `type` of plot you want to draw)
     - by default the plot will draw the vertical bar for the day with the highest amount of rain in red. The user must be able to change the color to be used.
   - Test your function on the Dublin and Cork airport data set, and set different meaningful titles for the two cases. [0.5]

[1]Source: `https://www.met.ie/climate/available-data/historical-data`