STAT40620/STAT40730
Data Programming with R

Final Project

Dr. Isabella Gollini
Dr. John O'Sullivan

Instructions

- This assignment is due on **December 22nd 2021 at 5:00pm**.
- You should submit it to the 'Project' assignment object in Brightspace.
- You should submit the following files only:
  1. `Rmd` files detailing the commented code you used to obtain your answers.
  2. final document in either `pdf` or `Word` which should contain answers to the project questions.
     - If you created an HTML file, please convert it to pdf. You can use Google Chrome: `File` > `Print` > `Destination [Change...]` > select `Save as PDF`.
  3. Note that, you can submit a single Rmd (plus pdf/Word) for all three parts of the project, or separate files for the three parts.
- You may submit it multiple times before the deadline, but only the last version will be marked.
- There is a maximum of 50 marks for this assignment. This assignment is worth 50% of your final grade.
- The marks available for each question are shown in brackets
- Late submissions will score 0, unless a "Late Submission of Coursework" form is submitted.
- The project is broken up into three parts: analysis, R package, and functions/programming.
- You may have to discover and learn some new functions. Use `help()` and `help.search()` to find what you need.
- Complete your assignment using R Markdown, check that all the output and code are correctly shown in your final document. Knit your document frequently to fix errors. Once completed, submit the `Rmd` file(s) and the resulting `pdf` or `word` document(s) which shows all your code.
- Some tips on using R Markdown are given at the end of this document.

Final Project

The final project has three main parts: Analysis, R Package and Functions/Programming.

If you intend to use packages that have not been used throughout the module, you should explain why they were necessary to complete the assignment, and cite the packages used (hint: most packages include citation information by running the function `citation()`. E.g. `citation("ggplot2")`).

R Markdown [5]

Complete your assignment using R Markdown, check that all the output and code are correctly and nicely shown in your final document. Knit your document frequently to fix errors. Once completed, submit the `Rmd` file and the resulting `pdf` or `word` document which shows all your code.

Part 1: Analysis [15]

This task involves finding a dataset of interest to you, that contains a mix of categorical (factors) and numerical variables. As a guideline, the dataset would typically have a minimum of two categorical variables and three numerical variables; these minima are guidelines and not hard thresholds.

If you wish you can make use of the following websites to find the dataset:

- The Irish government data repository: `https://data.gov.ie/`
- Google dataset search: `https://datasetsearch.research.google.com/`

The task is to use the methods covered in this course to complete an analysis and write a report using R Markdown on the data. The analysis of the data should involve the use of graphical summaries, tables and numerical summaries of the data.

This part of the project will be assessed in terms of:

- Using the functionality and settings of the appropriate functions in R.
- Clearly annotating the code in the R Markdown file.
- Producing clear results for the data.
- Summarizing the conclusions from the analysis appropriately.

Part 2: R Package [15]

This task involves finding an existing R package, that we didn't use extensively in the course, and write a report demonstrating its use using R Markdown.

The report should demonstrate some of the key functionality of the package, but doesn't need to demonstrate all of the functions (only the main ones).

This part of the project will be assessed in terms of:

- Clearly summarising the purpose of the package.

- Clearly demonstrating the functionality of some of the main functions in the package on appropriate data.
- Clearly showing the code and output for the demonstration examples.

Part 3: Functions/Programming [15]

This task is to write an R function (or set of functions) that can be used to provide a statistical analysis of interest. The function(s) should be documented by the code having comments and a working example.

The output from the function should use S3 or S4 classes and an appropriate `print`, `summary` and `plot` methods should be developed and demonstrated with an example.

This part of the project will be assessed in terms of:

- Writing a working function to provide an analysis of interest.
- Providing appropriate `print`, `summary` and `plot` methods for the output from the function.
- Clearly commenting the code and writing a clear example.

Tips for R Markdown

- Be aware that a common error is to give the same label to two different code chunks!

  ```
  ```{r cars}
  summary(cars)
  ```


  ```{r cars}
  plot(cars)
  ```
  ```

  You can fix this by changing the label to one of them:

  ```
  ```{r cars2}
  plot(cars)
  ```
  ```

- If you want to improve the appearance of your plot in your knitted document you can set up the dimension of your figure:

  ```
  ```{r, fig.height = 10, fig.width = 7, fig.align = "center"}
  plot(Nile)
  ```
  ```

- In case of an error in your code, add the option `error = TRUE` into the R chunk to run the code, show the error message on the knitted file. For example:

  ```
  ```{r, error = TRUE}
  x <- "a"
  sum(a)
  ```
  ```

- For all the available options for the R chunk, you can see here: `https://yihui.name/knitr/options/`
- R Markdown website: `https://rmarkdown.rstudio.com/`
- R Markdown cheatsheet is available here: `https://www.rstudio.com/resources/cheatsheets/#rmarkdown`