Data Prog with R - Autumn Trimester 2022/23

Lecturer Dr Emma Howard

Instructions

This project is due on Wednesday December 21st 2022 at **11.59pm**.

- You should submit it to the 'Project' assignment object in Brightspace.
- You should submit the following files only:
    1. Rmd file detailing the commented code you used to obtain your answers.
    2. final document in pdf format 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. If you have created a word document, please convert it to pdf by saving it as a .pdf file}

    3. The dataset that you are working with (unless the code directly downloads the file from online)
- 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 an extenuating circumstance form has been submitted and approved.
- 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.

Plagiarism

While you are encouraged to ask about the module material, this project should be completed individually. Any student who plagiarises will receive a 0 mark. Projects will be reviewed by the UCD plagiarism software. If you are unsure whether a question about the project would be considered as plagiarism, please email the question to the lecturer rather than posting on the discussion forums. The UCD Plagiarism Policy applies to all students. This can be consulted at the following *link*.

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 [20]

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 miminum criteria are guidelines and not hard thresholds. Do not use an in-built R dataset or a dataset from kaggle.

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.
- Quality of the graphics included.
- Summarizing the conclusions from the analysis appropriately.

Part 2: R Package [10]

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). The full list of CRAN packages is available at this *link*. Alternatively, you can choose a package on Github.

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.