STAT40730 Data Programming with R - Final Project

Isabella Gollini

Instructions

- This final project is due on **Wednesday 20th December 2023** at 11:59pm.
- You should submit it to the "Final Project" assignment object in Brightspace.
- You should submit 2 or 3 files only depending on your final format.
    - If you **render to pdf** you have to submit 2 files:
        1. **Qmd** file detailing the commented code you used to obtain your answers.
        2. Rendered document in **pdf** showing all your code and your answers.
    - If you **render to HTML** you have to submit 3 files:
        1. **Qmd** file detailing the commented code you used to obtain your answers.
        2. **A zip file** containing the HTML file showing all your code and your answers. (Notice that the zip file must contain only the HTML file).
        3. a **pdf** file obtained by converting the HTML to pdf. *You can use Google Chrome. File > Print > Destination [Change. . .] > select Save as PDF.*
- Remember that if you decide to produce an HTML output you must have the following on your YAML header:

```
format:
  html:
    embed-resources: true
```

- 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 final project. This final project 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 final project consists of 3 tasks: 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.
- A couple of suggestions: create an RStudio Project for this project and save the .qmd file and the data set in the same folder. Render your document frequently to fix errors.

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 this project, and cite the packages used (hint: most packages include citation information by running the function `citation()`. E.g. `citation("ggplot2")`).

Quarto [**5**]

- Complete your assignment using Quarto, check that all the code and output are correctly and nicely shown in your final document.
- Cite the source of the datasets and any new package you use.
- Render your document frequently to fix errors.

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, datasets from Kaggle or datasets from TidyTuesday.

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

- Central Statistics Office www.cso.ie
- Ireland's Open Data Portal data.gov.ie
- Met Éireann www.met.ie
- World Trade Organisation W.T.O www.wto.org
- Organisation for Economic Co-operation and Development O.E.C.D. www.oecd.org
- For sport fans: https://fbref.com/en/
- data.europa.eu - The official portal for European data data.europa.eu/data/datasets
- Google dataset search datasetsearch.research.google.com
- This has a good list of places with datasets github.com/awesomedata/awesome-public-datasets

The task is to use the methods covered in this course to complete an analysis and write a report using Quarto 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 Quarto file.
- Provide a link to your data, or upload the dataset and provide a reference to ensure we can reproduce your project.
- Producing clear results for the data.
- Quality of the graphics included.
- Exploration or interpretation of what was being produced on any graphs/table/summaries.
- Structure your analysis appropriately (e.g., a report format with subsections)
- Summarizing the conclusions from the analysis appropriately.

Some hints for this part:

- Any packages/functions seen in the lecture can be used (e.g., you could choose to use only `ggplot2` or only *Base R* graphics for your analysis) and not all functions seen need to be used (e.g., your data may not require any matrix manipulation). The choice of functions will be impacted by the data chosen.
- For your data analysis, apart from summaries, do consider other functionality e.g., remove NA rows.
- You can choose to answer some questions about the data if you wish (i.e. set questions at the beginning) or to keep it more exploratory.

Part 2: R Package [**10**]

This task involves finding an existing R package, that we didn't use in the course, and write a report demonstrating its use using Quarto. 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. *The data used here can be different from Part 1 if you wish, any dataset is fine. However you cannot provide the exact examples that can be seen in the example section of the help page (`?function name`) for a function, or on any paper or tutorial available on the web.*

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 (at least three functions should be examined).
- 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. *The data used in this part can be different from Part 1 or Part 2 if you wish, any dataset is fine.*

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. Notice that the `summary` and `print` function should be different from each other.
- Clearly commenting the code and writing a clear example.
- Clearly state the class being assigned.

END OF FINAL PROJECT