Project guidelines

Eduard Campillo-Funollet

General notes

- MATH246 students have a separate project. This guidelines only apply to MATH245 projects.
- **You only have to do one project**. There are three projects to choose from to offer you a range of topics. The three projects are similar in difficulty, but they use different algorithms and methods. My suggestion is that you choose a project that you find interesting! See more details on the project topics below.
- You can do your project in R or Python.
- I strongly recommend that you use the libraries we discussed in the module. In particular, the plotting libraries (ggplot2 or plotnine). Some of the questions are *more difficult* if they are done using other plotting libraries.

Deadline

- All projects are due on April 23rd 2024, 2pm.

Submission

- You **must** submit your projects in Quarto and PDF format.
- The PDF output of your project must not exceed 20 pages (including the question statements).
- Quarto templates for R and Python have been provided.
- Your Quarto submission **must** produce the submitted PDF, and should run in a folder exactly as the project folder in Moodle. In particular, note that all paths to files and folders must be relative to the qmd file location, and that the data files will be located exactly as in the Moodle folders.

Marking

- A detailed marking scheme will be published before the submission deadline.
- 25 marks will be allocated to coding style (comments, correct use of loops, variable names, functions). Some of this marks will be removed if the Quarto document cannot be compiled.
- Plots must be correctly labeled. Please look at Chapter 5 of the lecture notes for details.

On the project topics

- The DNA project is about the information content of DNA and a method to visualise it. Although it is "inspired" by biology, it is very a much a project about how information is encoded in a string of symbols, and how do we measure information. The data provided is the DNA sequence of one chromosome of a species of yeast (*S. Pombe*)
- The Earthquakes project is about predicting secondary earthquakes. It is a statistical analysis of real earthquake data, and how to use simple features to fit a statistical model.
- The COVID-19 project is about using ordinary differential equations to model infectious diseases. It uses data from the Omicron variant of COVID-19.

Final notes

- Questions are independent (you might do Question 2 even if you cannot complete Question 1). In some cases, extra data files are provided for that.
- "Roughly", the questions go from easy to challenging.