# Problem

The file `SLR-Data.csv` contains data concerning many types of litter on the shores of Swiss lakes and rivers, collected by trained volunteers. The file contains a lot of information, including the lengths of the beaches in metres. The numbers of items are given for around 100 different types of litter (different codes), most of which are absent on most occasions. The most common type is cigarette butts and filters ('mégots', code G27). The goal of the analysis is to provide a model that would enable you to give a prediction interval for the number of butts on a given beach on a day in August 2023. You will need to take into account:

- beach length;

- any seasonality from month to month or variation from year to year;

- variation among different beaches;

- possible overdispersion relative to standard parametric models; and

- bizarre observations, and oddities of the model fit.

The file `hammerdirt.R` contains some basic code to read in the data, make a basic data frame, and perform a simple (and clearly very inadequate) analysis . . .

# Report

## Submission and deadlines

A short report of at most **12 pages** (excluding cover page, table of contents and bibliography) is to be **submitted on Moodle** as a PDF file (nothing else will be accepted) by midnight on **19 January 2023**. Later submission will not be accepted. Reports must be written **in pairs**; you can discuss with your classmates, but the code and writing should be your own. You should upload your code in a separate file and ensure the output is reproducible (not proper to your laptop or computer).

Your file should be named `NameA-NameB-RMProject-2023.pdf` (in alphabetical order of surname, e.g., `Hardy-Laurel-RMProject-2023.pdf`).

This will count as **40%** of the final grade for the course.

## Structure of the report

The report should be typed in English. Some notes on report-writing can be found at

`https://moodle.epfl.ch/pluginfile.php/2567584/mod_resource/content/2/SMA-Projects-Description.pdf`

and there is an example report posted on Moodle.

**Introduction**: Briefly state the purpose of the analysis, discuss the main features of the data (e.g., via exploratory data analysis), and outline what will follow.

**Analysis**: describe the model(s) fitted, using your own words. Give the key elements only: you can refer to the lecture notes and to books, but should give careful references (to pages and equations etc.). It is not enough simply to give a list of sources at the end of the work: references should be mentioned in the text, and only those mentioned in the text should be listed at the end. Use BibTeX or similar to ensure that the references appear properly; check a book or journal article to see what details should appear in the bibliography.

**Discussion** of the results in more detail. Include crucial graphs and tables only, make sure that their contents are understandable without reference to the text, and that their axis labels and captions are clear and informative; each graph and/or table should tell the reader a coherent story. Give appropriate numbers of digits for tables. The text should give detailed interpretations of the plots and tables, with more details, if they are needed, and should show where the graph/table fits into the overall picture.

**Conclusions**: the take-away message from your analysis. Convince the reader that you know what you did and are aware of its strengths and limitations. Sketch what more you might do, if you had more time.

## Discussion points

**Exploratory data analysis** : nature of covariates and distribution of response, presence of outliers or missing values, range of variables.

**Modelling** : key variables, with fitting using GLMs and/or GAMs, Test statistics and goodness-of-fit diagnostics. Interpretation of the final model on a meaningful scale.

**Discussion** : discuss the results, showing (e.g.,) histograms of fitted values, comparisons of fitted and predicted probabilities, analysis of deviance and/or AIC for model comparisons, give estimates and their standard errors, explain any disagreement between the models. Ensure that you carefully interpret the fitted models in terms of the original problem.

## Suggestions and caveats

1. We recommend that you use LaTeX.
2. Your report should be sufficiently detailed that a reader can reproduce your results after reading it.
3. Figures and tables should be numbered and have captions briefly explaining their contents. Reference should be made to each figure/table from within the text.
4. Read your report carefully before handing it in and use a spell checker to find any typos.
5. Mention any references you have used and provide a detailed bibliography. References should be made to scientific articles or books. Detailed (chapter, section, page, equation) references to books are usually needed, so that the reader does not have to figure out which page(s) of a book you are referring to.
6. Pasting plain computer output is not acceptable.
7. Due to space limitations, you should provide only relevant output. Your code can however contain exploratory data analysis and other model fits and diagnostics that are not reported in the text.
8. Your code file should be commented (but self-explanatory commands need not be commented on).
9. When writing a report, you should not answer questions directly. Instead, make sure your report covers the material discussed in each point, but structure your report as a scientific paper.
10. Common problems: many students don't give enough (or sometimes any!) interpretation of fitted models; often tables have too many digits; often figures are too small to be read properly or don't use the page layout well; often captions to tables or figures are uninformative; often the discussion section is insufficiently detailed; often the bibliography has missing details; often references to publications are inadequate; often equations are not (or are incorrectly) punctuated; often the English has persistent spelling errors.

## Marking scheme

| Correctness | Accurate, appropriate use of statistical tools | | | | | | Incorrect, many errors | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | Score _____ |

| Discussion | Thoughtful, detailed, apposite | | | | | | Banal, obvious, thin | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | Score _____ |

| Graphics and tables | Clearly labelled, well-chosen, good captions and discussion, appropriate numbers of digits | | | | | | Poorly labelled, no discussion, unmotivated, unedited output | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | Score _____ |

| Originality and scope | Wide range of tools/ideas | | | Limited range of tools/ideas | | | |
|---|---|---|---|---|---|---|---|
| | 5 | 4 | 3 | 2 | 1 | 0 | Score _____ |

| Quality of writing | Good grammar and punctuation, including mathematics | | | Poor grammar and punctuation | | | |
|---|---|---|---|---|---|---|---|
| | 5 | 4 | 3 | 2 | 1 | 0 | Score _____ |

| Referencing | Full, accurate, and detailed references given | | | Inadequate citation of sources | | | |
|---|---|---|---|---|---|---|---|
| | 5 | 4 | 3 | 2 | 1 | 0 | Score _____ |

Grand total (max 45) _____