STAT40950 Assignment 1 2024/2025

Isabella Gollini

The aim of this assignment is to write a scientific report (20 pages) in which you use the Bayesian workflow, Bayesian Hypothesis testing, Bayesian Model Selection, and Bayesian GLM to answer scientific questions related to a given dataset.

Quarto (or RMarkdown) must be used, and all the analysis must be reproducible given your `.Rmd` or `Qmd` file and the original dataset `weather_bikes_2025.csv`. The maximum number of pages of the resulting `pdf` or `word` document is 25. (i.e. we'll stop marking after page 25).

Instructions

- This assignment is due on **Tue 17th June 2025** at 11:59pm.
- You should submit it to the *Assignment 1* assessment object in Brightspace.
- You should submit two files only:
  - `.Qmd` (or `.Rmd`) file originating the report (all the analysis must be reproducible given your `.Qmd` (or `.Rmd`) file and the original dataset `weather_bikes_2025.csv`)
  - final document in either `pdf` or `Word` which should contain answers to the questions below, you must show all the R code[1]. **The maximum number of pages of the resulting `pdf` or `word` document is 25. (We'll stop marking after page 25)**
- You may submit it multiple times before the deadline, but only the last version will be marked.
- This assignment is worth 50% of your final grade.
- Late submissions will score 0, unless a "Late Submission of Coursework" form is submitted.

Data

The dataset `weather_bikes_2025.csv` contains variable concerning bike traffic and weather conditions in Dublin from January 1st until April 30th 2025.

```
library(readr)
bikes <- read_csv("weather_bikes_2025.csv")
```

There are three variables concerning bicycle traffic volumes from cycle counters in three locations Dublin city: Clontarf, Griffith Avenue, and Grove Road. Passing cyclists are counted and logged every hour, 24 hours per day, 7 days per week. Data provided by Dublin City Council and the NTA[2] from January 1st until April 30th 2025.

The other variables concern weather condition, and they have been downloaded from Met Éireann.

In detail the dataset `weather_bikes_2025.csv` consists of the following variables:

- `Date` and `Hour` timestamp for the data collected
- `Day` day of the week
- `Clontarf` hourly bicycle traffic volumes from cycle counters in Clontarf (Pebble Beach Carpark)
- `Griffith Avenue` hourly bicycle traffic volumes from cycle counters in Griffith Avenue (Clare Rd Side)

[1] Code needed to set things up, e.g. loading packages etc. can be hidden to make the final rendered document neater, and you can suppress output messages not needed. Select carefully which output, or parts of output to display.

[2] The Bicycle traffic data have been downloaded from data.smartdublin.ie

- `Grove Road` hourly bicycle traffic volumes from cycle counters in Grove Road Totem
- The variable relating to the weather are self-explanatory: `Precipitation Amount (mm)`, `Air Temperature (°C)`, `Mean Hourly Wind Speed (kt)`, `Visibility (m)`.

Questions

Write a scientific report in which you use the Bayesian workflow, Bayesian Hypothesis testing, Bayesian Model Selection, and Bayesian GLM to answer the following scientific questions related to the `weather_bikes_2025.csv` dataset.

1. Is the average temperature in January and February greater than 5°C? (You can assume know variance $\sigma^2 = 9$)
2. Is there a correlation or association between the number of cyclists passing on different locations on the same date?[3]
3. Is there a relationship between the number of cyclists and the weather?
4. Based on this dataset, choose another scientific question and perform at least two appropriate Bayesian GLM.
5. What's the best model between the ones proposed in question 5?

You must consider all of these questions as *research questions* and it's up to you to translate them into *research hypothesis/statistical models* (you don't have to use all the variables in the dataset, just the ones you deem appropriate to answer the questions).

Notice that this is an open ended assignment, there is no correct answer *a priori*, but you must justify all the choices you make, and describe clearly all the steps of the Bayesian workflow you are following.

Notes:

- If you wish, you can use a subset of the dataset for your regression, and another part to help you choose the priors.
- If you wish, you can create new variables for your analysis such as weekday/weekend, rush hours, national holidays.

Marking scheme:

There are 50 possible points:

- *10 points* for the overall assignment format:
  - *7 points* for clear well written report
  - *3 points* for good use of rmarkdown
  - notice that we'll stop marking after page 25.
- *10 points* for Bayesian Workflow throughout the report:
  - *3 points* Exploratory data analysis for model and prior specification
  - *4 points* Validation of computation and posterior predictive checks
  - *3 points* model comparison/selection when appropriate
- *30 points* for correctly approaching and answer the questions (*6 points* for each question)

[3] You can pick any location as response variable, you can work on hourly data, or aggregate them into parts of the day, or daily data.