MAST7866 — Foundations of Data Science

Lecturer: Professor Jian Zhang Room: CS115A

Email: j.zhang-79@kent.ac.uk

Deliver: Lectures + PC classes+ videos

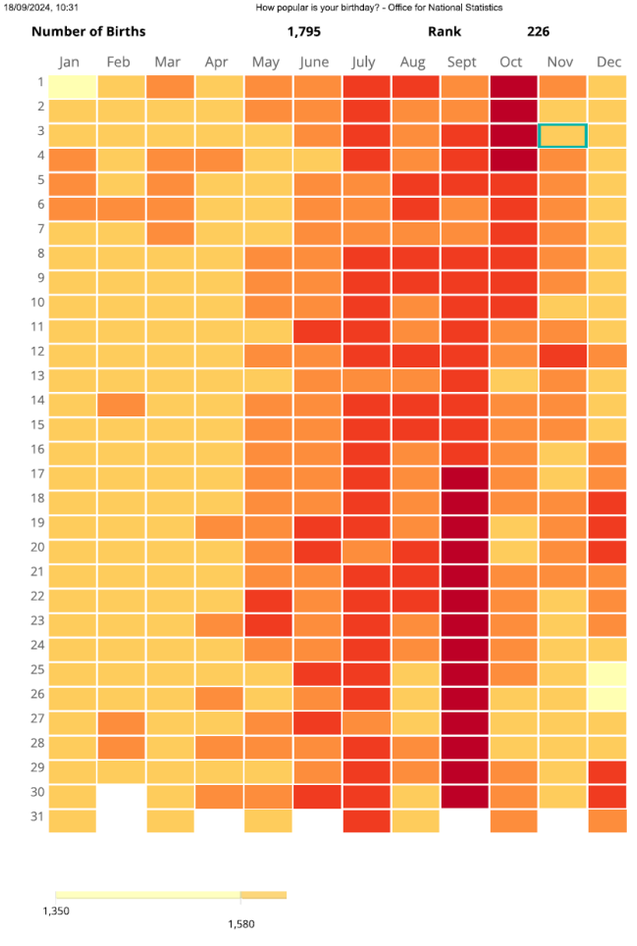
Assessments: **Exercises plus coding** (20%), **Group project report** (50%) and **Individual presentation** (30%).
Policy: Following the new late submission policy to students, academics have no leeway to accept late submissions.
Syllabus: Axioms of probability, discrete and continuous random variables, expectation and variance, common distributions, testing a hypothesis, data visualisation, programming with R, and linear regressions.
Resources: R tidyverse http://www.tidyverse.org/

R for Data Science http://r4d5.hadley.nz/

Introductory Examples

In data science, an **experiment** is designed and undertaken which yields data. We want to extract information from the data using probabilistic models.

For example: Sample of birthdays in the class.



Source: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths

Random Sampling

Newspaper and television reports often refer to 'random samples', which has a very precise meaning, namely

'select a sample in such a way that all possible samples have exactly the same probability of being selected'.

This forms a probabilistic argument for **equally likely outcomes**.

Random Sampling

A marine biologist studying the fish in a lake may use relative frequency in a process called capture-recapture which helps them to estimate the total number N of fish in the lake.

This method works by assuming that each fish has an equal chance of being captured. For example, a sample of 80 fish could be taken, which are all then tagged. Later a second sample of 30 fish is taken and the number of them that are tagged is recorded. In this case, two of the fish are tagged.

The relative frequency of picking a tagged fish is $2/30$. which is the same as the original $80/N$. To find the population of fish in the lake the formula $N = 30 \times 80/2 = 1200$. The number of fish in the lake is estimated to be 1200