Post-Strike Recap

Max Turgeon

DATA 2010–Tools and Techniques in Data Science

Changes to the assessment schedule i

- We have 1 assignment and 2 quizzes left. We also have the final project and the final exam.
- I'll provide flexibility: let me know if you need extensions.
- I also don't expect you to work during the Holiday break.

Assessment	New Date
Lab Quiz 4	Dec 20
Progress report	Dec 23
Lab Quiz 5	Jan 17
Assignment 3	Jan 19
Final Report	Jan 19

Changes to the assessment schedule ii

- Some comments:
  - Lab Quiz 4 will last 30 minutes, but it will open for a period of 24 hours (you can start whenever you want).
  - Assignment 3 will cover material up to next Friday. It will be posted this week, and feel free to submit before Christmas if you want.
  - A new final exam schedule will be provided soon.

Lectures in December

- I understand that some of you will be writing exams in the next few weeks.
    - And attending lectures will be difficult.
- I'm aiming to provide slides and video recordings of all December lectures by **next Monday**.
- I will also post Assignment 3 later this week.

Recap

Probability and Statistics

- We started with a review of probability and statistics.
- We focused on *conditional* probabilities.
    - If I have some information about a random variable, what does it tell me about another random variable?
    - Bayes Theorem
- Discrete vs Continuous distributions, etc.

Data Wrangling

- We spent several lectures (and labs!) on data manipulation.
    - Summarize (by groups or not), create new variables, filter data.
    - Joining multiple datasets
    - Tidy data
    - Dates, regular expressions
    - Pandas
- Analysts spend a lot of time cleaning data.

Data Visualization

- Visualizations in R and Python
    - Histograms, bar plots, box plots, scatter plots, etc.
- We also talk about principles of effective data visualization.
    - What visual cues are you using? Are they effective?
    - How can I best highlight important comparisons?

Modelling

- The last few modules focused on data analysis and modelling.
  - We started by discussing correlation, distributions and significance.
  - We then discussed scores and rankings.
  - We just started discussing how to build models.