Python data science

Hamish Gibbs

Today: Python data science

- Introduction to Python data science tools.
- Introduction to a basic data science workflow.
- This afternoon: collaborating on a data science project.

Data science

- Definition of data science:
  - *"Extracting meaningful insights from data."*
- *Meaningful* is important.
  - Use the tools of programming / statistics to create meaning from your data.
- Usually, there is no "right" answer, just "better" and "worse" answers.
  - You exercise a lot of judgement.

Data science workflow

- Data science is **not** just machine learning.
  - Most data science work is:
    - Data preparation
    - Data transformation
    - Method selection
      - *Statistics / machine learning*
    - Communicating results

Python data science tools

- Today, we will learn about the most popular Python data science “stack”
  - Data preparation / data transformation
    - `pandas`, `numpy`
  - Statistics / machine learning
    - `sklearn`
  - Communicating results
    - `matplotlib`

R equivalents

- Python libraries mostly have their R equivalents:
  - `pandas`:`dplyr`
  - `sklearn`:`caret`?
  - `matplotlib`:`ggplot2`
- See what you prefer, I use both languages very regularly!

Diving deeper

- Python has many other options for data science tools.
- Alternatives to `pandas`:
  - `polars` *(Like Python's version of* `data.table`*)*
  - `dask`
- Alternatives to `sklearn`:
  - …?
- Alternatives to `matplotlib`:
  - `seaborn`
  - `plotnine` *(R users might like this one!)*

Tutorial #1: pandas

- 10 minutes to pandas
- Core concepts:
  - TODO

Tutorial #2: pandas / matplotlib

- Chart visualization
- Core concepts:
  - TODO