Introduction to Python for Social Science

Lecture 2 - Data Structures and Pandas I

Musashi Harukawa, DPIR

2nd Week Hilary 2020

Recap

Last week we covered the following topics:

- What is Python, and what can I use it for?
- What tools do I have to write, test and run Python code?
  - Opening up JupyterLab/Notebooks/IPython
  - Executing code in Jupyter, etc.
- Writing your first Python script and notebook.
  - Installing libraries
  - Importing libraries
  - `print`, `+`, `-`, `*`, `/`, `list`, `for`...

This week

This week we will learn about **data**:

- Thinking about (tabular) data
- I/O
- Slicing, indexing
- Summarising

And we will do all of these things with the `pandas` library.

The Hard Truth About Data Science...

- Analysis usually takes $<$30% of your time.
- $>$50% of your time will be spent reading, cleaning, checking, storing, and cursing your data.

The Good News:

- Data cleaning is meticulous work, but that doesn't mean you can't be efficient.

Thinking About Data

Without getting into information theory, we can think about two properties of data:

1. Value
2. Relation (to other values)

Data Structures

Three Ways of Structuring Data:

- Graph (Relational)
- Hierarchical
- **Tabular**

Data I/O

Pandas comes with functions for reading and writing to all kinds of data formats. A quick list can be viewed using tab completion:

```
In [1]: import pandas as pd

In [2]: pd.read_<TAB>
 read_clipboard() read_hdf()       read_sas()
 read_csv         read_html()      read_sql()
 read_excel()     read_json()      read_sql_query()
 read_feather()   read_msgpack()   read_sql_table()
 read_fwf()       read_parquet()   read_stata()
 read_gbq()       read_pickle()    read_table
```

csv format

csv, also known as comma-separated-values (or as I prefer to call it, character-separated values), is a standard *plaintext* tabular data storage format.

Some reasons to use csv:

- lightweight
- human-readable
- optional header (first row)
- *fairly* portable between systems

Some limitations of csv:

- unpredictable behaviour of separator is common character (e.g. ,)
- fixed number of rows (strictly tabular)
- not very durable