Pandas(1)

Akitaka Matsuo
Essex IADS

Pandas: Overview

- We learn pandas package in Python
- There are two main subjects to learn
  1. Pandas data input and output (Lecture 02)
  2. Pandas data manipulation
     - We will start with Series
     - Then DataFrames
       - Creation/indexing
       - Select rows and columns
       - Merging DataFrames
       - Get descriptive statistics
       - Pivoting (“Long” to/from “Wide”)

PANDAS DATA INPUT/OUTPUT

Pandas

- Data storage and manipulation solution for Python
  - Using *DataFrame* object
  - Read/write data
    - Understand most of common data formats
    - Read them into DataFrame (then do this and that in Pandas)
  - Data manipulation
    - Sophisticated indexing
    - Data wrangling
      - Reshaping/pivoting (wide-to-long conversion)
      - Missing data handling
      - Combining/merging datasets

Various data formats

- There are various data formats you have to work with
- Typical examples:
  - csv: comma-separated value
    - Text-based, the most common data format for distributing data
  - xlsx: Excel file
  - Statistical software specific data files
    - sps (SPSS), dta (STATA), rda (R)
  - Common data formats on the Internet
    - html, xml, json
  - Database
    - SQL
- Most of them can be read by Python via pandas, using read_**()

Getting the data files on Colab

- There are multiple ways to do that
  - Using terminal command
    - `!wget`
      - Direct download from the web
    - `!git clone`
      - Get the data placed on github repository
  - Using Google Drive
    - Colab has access to Google Drive
      1. Download file to your computer
      2. Upload on Google Drive
    - Simplest, but sometimes a bit inefficient

PANDAS SERIES

Series in Pandas

- Pandas object, similar to 1d NumPy array
- Difference: Indexes
  - Indexes can be numbers or strings
- Construction:
  - 2 lists (or NumPy arrays)
    - One for values and one for indexes
  - Dictionary
- Slicing can be done with indexes
  - Similar to the one for NumPy arrays, but more sophisticated