# Introduction to Python for Social Science Lecture 3 - Data Structures and Pandas II

Musashi Harukawa, DPIR

3rd Week Hilary 2020



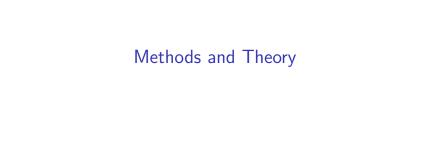
#### Last Week

- ► Graph, Tree and Tabular Data Structures
- ► Representations of information
- ▶ Introduction to pandas

#### This Week

This week we learn more advanced methods for working with data:

- Functions
- apply and vectorization
- GroupBy: Split-apply-combine
- Combining dataframes: append, concat and merge
- Long- vs wide-form data; melting data



#### **Functions**

- ▶ A function is a mapping of two sets that relates each element of the first set to exactly one element of the second set.
  - Formally, a function f is a mapping of elements of a set X to set Y defined by ordered pairs G = (x, y) such that  $x \in X$  and  $y \in Y$ .
  - ▶ X is referred to as the *domain* of f, and Y is the *codomain*.
  - $\triangleright$  y is the *image* or *value* of f applied to the argument x.
- Practically, a function is an operation that takes one or more inputs, and returns zero or more outputs.
  - For instance, the function f(a, b) defined as a + b takes two arguments, a and b, and returns a value a + b.
  - y can be the null set, in the sense that functions can return nothing.

### Functions and Vectors

There are several ways to think about applying a function to a vector  $X_i$  of i values.

### **Transformation**

The vector of all values in  $X_i$ ,  $[x_1, x_2, ...x_i]$  is in the domain of f, and a vector  $Y_i$  of equal length i is returned.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} = f \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \end{bmatrix}$$

## **Element-wise Operations**

Individual elements of  $X_i$  are in the domain of f, and f is applied to each element of  $X_i$  to return a vector of length i where the ith element is the value of  $f(x_i)$ .

$$\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_i) \end{bmatrix} = f \odot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \end{bmatrix}$$

### Summaries

A summary reduces a vector  $X_i$  of length i to a single value  $\theta$ . Thus vector  $X_i$  is within the domain of f, and  $\theta$  is value of f applied to  $X_i$ .

$$\theta = f \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \end{bmatrix}$$

## **Functions**

#### **Functions**

- ▶ A function is a structure takes one or more inputs, and gives zero or more outputs.
- ➤ You have already encountered many functions, such as sum(), which takes the sum of an series.

# Functions in Python

```
Here's a simple function that adds 1 to the input:

def add_one(x):
    """
    This function adds 1 to the input.
    """
    y = x+1
    return y
```

## def add\_one(x):

```
def add_one(x):
    """
    This function adds 1 to the input.
    """
    y = x+1
    return y
```

- ► The command def followed by a space tells Python that you are defining a function.
- This function is given the name followed by def; in this case add\_one.
- ► The *arguments* of the function are given after the function name, inside ().
- The : says that the definition line is done. The following line must be indented by four spaces.

# **Docstrings**

- A string immediately after a function definition is automatically assigned as the **docstring** for that function.
- ► The docstring is the documentation that appears when you use the func? command.
- ► This is optional, but a great way to document your code. It also helps you remember and read your code faster.
- NB: I use a triple-double quote """ to create a multiline string. This is convenient, but not necessary (you can use a simple " or ').