Data 8

Spring 2020

Lecture 10

Groups

Announcements

- HW3 due tomorrow
    - Submit tonight 11:59PM for bonus point
- **Project 1** is out on Friday
    - You can get a partner in lab
    - Checkpoint Fri. 02/21, **due Fri. 02/28**

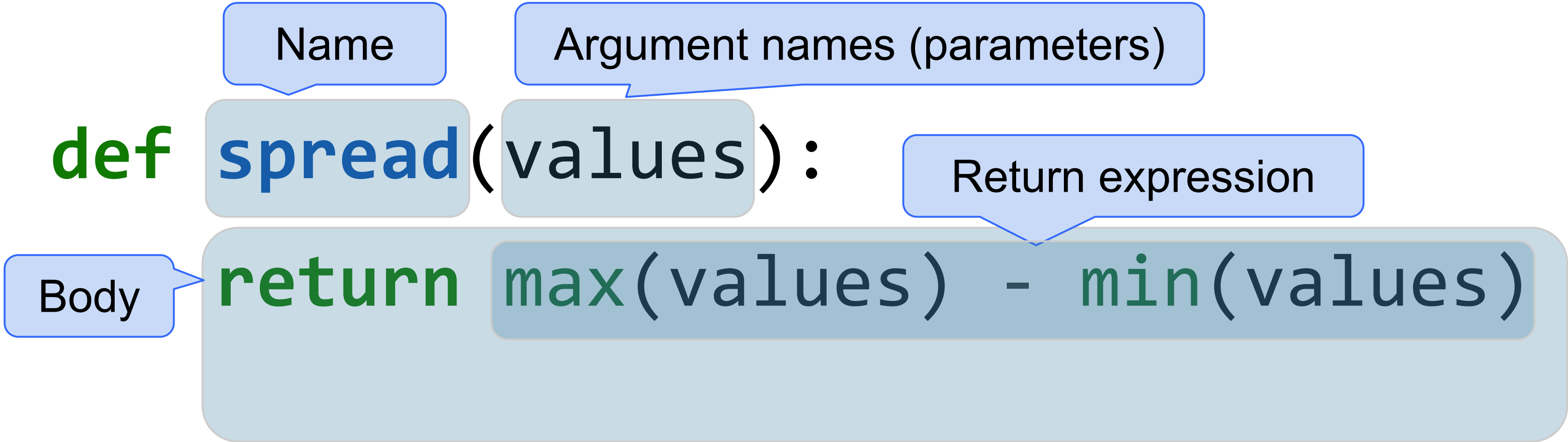
Weekly Goals

- **Today**
  - Review of graphs
  - Writing our own functions
- **Wednesday**
  - Making predictions
  - Aggregating data using `group`
- **Friday**
  - Two-way `group` and `pivot`
  - Combining tables using `join`

Defining Functions

Def Statements

User-defined functions give names to blocks of code



```
def spread(values):
    return max(values) - min(values)
```

Apply

Apply

The `apply` method creates an array by calling a function on every element in input column(s)

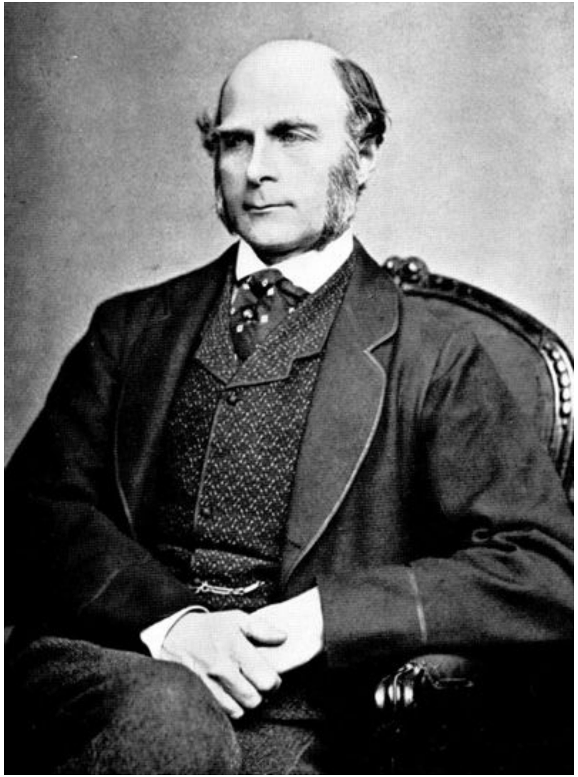
● First argument:	Function to apply
● Other arguments:	The input column(s)

```
table_name.apply(function_name, 'column_label')
```

(Demo)

Sir Francis Galton

- 1822 - 1911 (knighted in 1909)
- A pioneer in making predictions
- Particular (and troublesome) interest in heredity
- Charles Darwin's half-cousin



Prediction

(Demo)

Prediction Accuracy

(Demo)

Discussion Questions

1. How could we take gender into account when making predictions?
2. Do we make smaller errors on average when we do this?

Grouping

Grouping by One Column

The `group` method aggregates all rows with the same value for a column into a single row in the resulting table.

- First argument: Which column to group by
- Second argument: (Optional) How to combine values
  - `len` — number of grouped values (default)
  - `list` — list of all grouped values
  - `sum` — total of all grouped values

(Demo)

Lists

Lists are Generic Sequences

A list is a sequence of values (just like an array), but the values can all have different types

```
[2+3, 'four', Table().with_column('K', [3, 4])]
```

- Lists can be used to create table rows.
- If you create a table column from a list, it will be converted to an array automatically

(Demo)

Grouping By Multiple Columns

The `group` method can also aggregate all rows that share the combination of values in multiple columns

- First argument: A list of which columns to group by
- Second argument: (Optional) How to combine values

(Demo)

Pivot Tables

Pivot

- Cross-classifies according to two categorical variables
- Produces a grid of counts or aggregated values
- Two required arguments:
    - First: variable that forms column labels of grid
    - Second: variable that forms row labels of grid
- Two optional arguments (include **both** or **neither**)
    - `values`=’column_label_to_aggregate’
    - `collect`=function_to_aggregate_with

(Demo)