190F Foundations of Data Science

Spring 2020

Lecture 6

Functions

Announcements

Overlaid Graphs

For visually comparing two populations

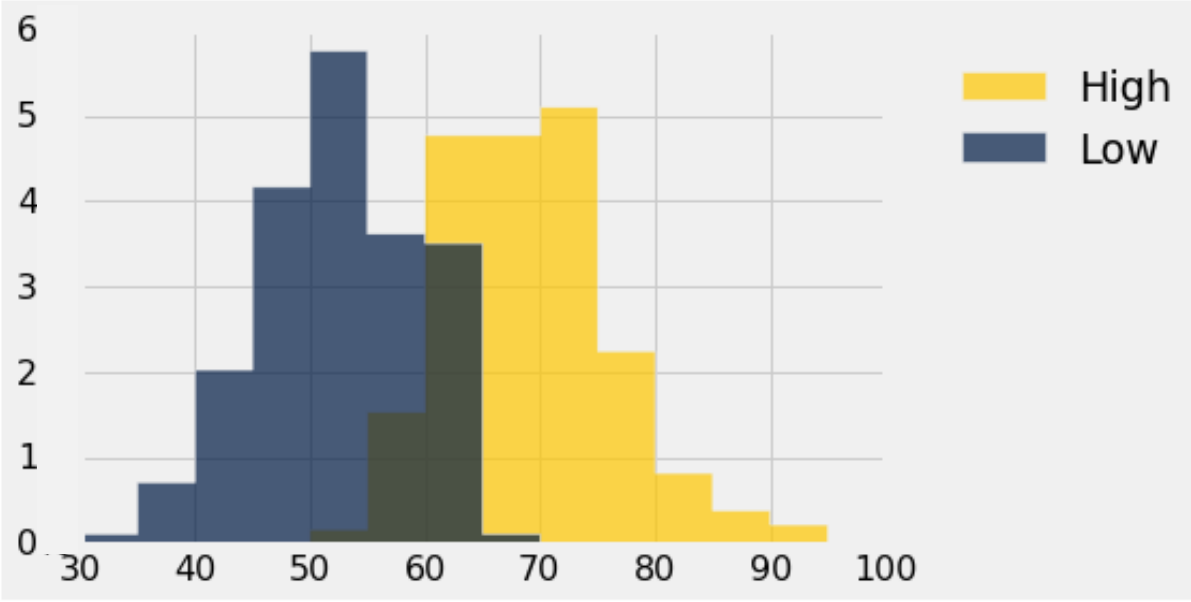
(Demo)

Discussion Question

This histogram describes a **year** of daily temperatures

Try to answer these questions:

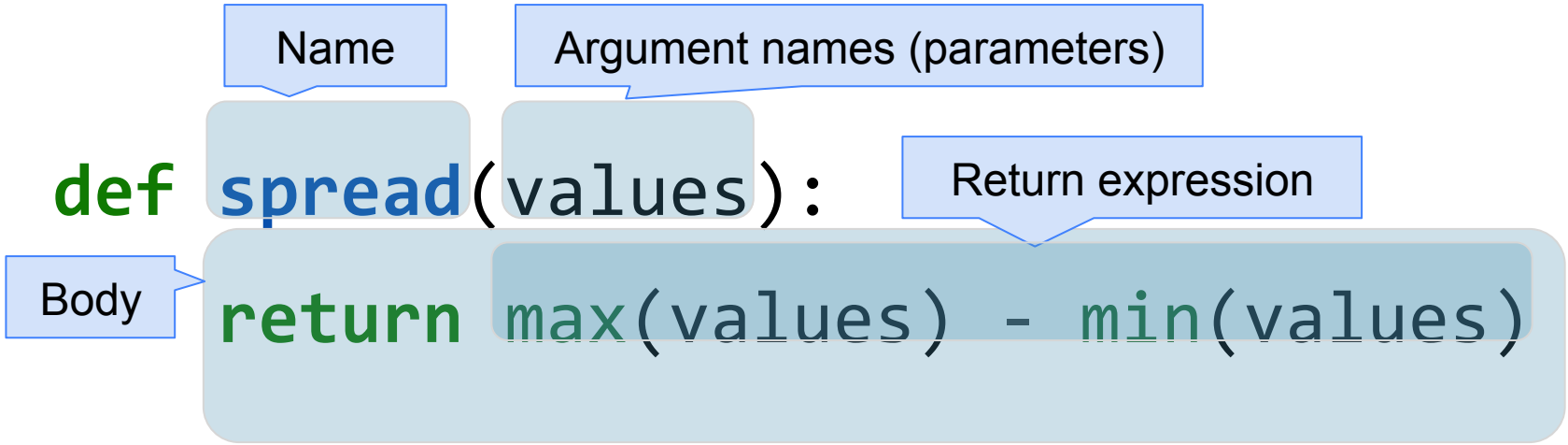
- What proportion of days had a high temp in the range 60-69?
- What proportion had a low of 45 or more?
- How many days had a difference of more than 20 degrees between their high & low temperatures?



Defining Functions

Def Statements

User-defined functions give names to blocks of code



(Demo)

Discussion Question

What does this function do? What kind of input does it take? What output will it give? What's a reasonable name?

```
def f(s):
        return np.round(s / sum(s) * 100, 2)
```

(Demo)

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')
```

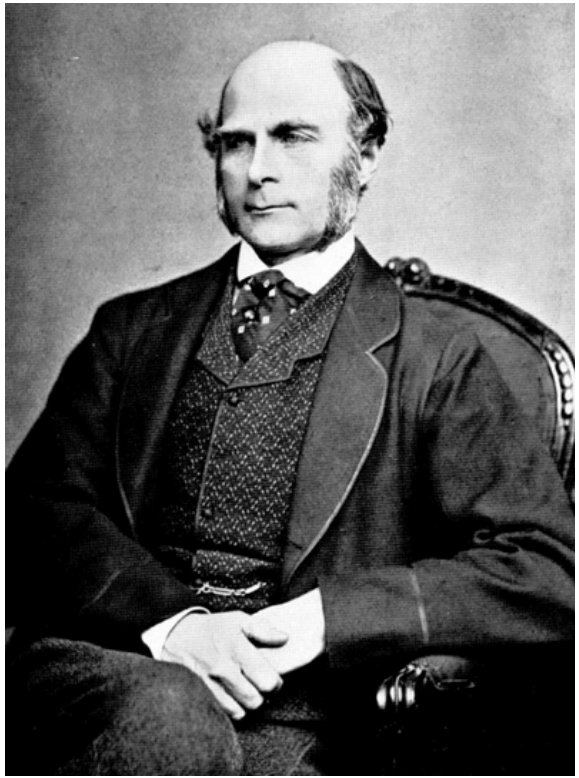
(Demo)

Example: Prediction

Sir Francis Galton

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

(Demo)



Apply

The `apply` method creates an array by calling a function on every element in one or more input columns

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

```
table_name.apply(one_arg_function, 'column_label')
table_name.apply(two_arg_function,
                 'column_label_for_first_arg',
                 'column_label_for_second_arg')
```

`apply` called with only a function applies it to each row

(Demo)

Grouping Rows

Group

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

- First argument: Which column to group by
- Second argument: (Optional) How to combine values

  - `len` — number of grouped values (default)
  - `sum` — total of all grouped values
  - `list` — list of all grouped 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_with_which_to_aggregate

(Demo)

Challenge Question

Which NBA teams spent the most on their starters in 2016?

- Each team has one *starter* per position
- Assume the starter for a team & position is the player with the highest salary on that team in that position

PLAYER	POSITION	TEAM	SALARY
Paul Millsap	PF	Atlanta Hawks	18.6717
Al Horford	C	Atlanta Hawks	12
Tiago Splitter	C	Atlanta Hawks	9.75625