DATA 8

Spring 2022

Lecture 9

Functions

Announcements

- **HW3** due Thursday 2/10
  - A bit of extra credit for turning it in by 2/9
- HW1 and Lab 2 regrade requests are due by Wednesday 2/9

Why Data Visualizations?

Charts can convey a lot of info in an interpretable manner

People are good at noticing patterns in visual media

The last two weeks:

- How to generate and correctly interpret visualizations
- Isolating the part of a table that you want to visualize

This week:

- Manipulating the data you have into the data you wish to visualize
- Complete your toolset of table operations

Histogram Heights

Area Measures Percent

Area of Bar = Percent in Bin

= Height x Bin Width

- “How many individuals in the bin?” Use area.
- “How crowded (dense) is the bin?” Use height.

Discussion Questions

What is the height of each bar in this histogram?

```
my_bins = make_array(0, 25, 30, 60)
incomes.hist('Income (millions)',
  bins = my_bins)
```

What are the vertical axis units?

`incomes:`

Rank	Name	Income (millions)
1	Scarlett Johansson	56
2	Sofia Vergara	43
3	Angelina Jolie	35.5
4	Reese Witherspoon	35
5	Gal Gadot	31.5
6	Julia Roberts	30
7	Jennifer Lawrence	28
8	Jennifer Aniston	28
9	Melissa McCarthy	25
10	Kaley Cuoco	25
11	Meryl Streep	24
12	Margot Robbie	23.5
13	Charlize Theron	23
14	Emily Blunt	22.5
15	Nicole Kidman	22
16	Ellen Pompeo	19
17	Mila Kunis	16
18	Elizabeth Moss	16
19	Viola Davis	15.5
20	Cate Blanchett	12.5

Answers

Vertical axis units: Percent per million $

```
my_bins = make_array(0,25,30,60)
[0, 25): (50%)/(25 million)
         = 2 % per million
[25, 30): (20%)/(5 million)
         = 4 % per million
[30, 60): (30%)/(30 million)
         = 1 % per million
```

(Demo)

incomes:

Rank	Name	Income (millions)
1	Scarlett Johansson	56
2	Sofia Vergara	43
3	Angelina Jolie	35.5
4	Reese Witherspoon	35
5	Gal Gadot	31.5
6	Julia Roberts	30
7	Jennifer Lawrence	28
8	Jennifer Aniston	28
9	Melissa McCarthy	25
10	Kaley Cuoco	25
11	Meryl Streep	24
12	Margot Robbie	23.5
13	Charlize Theron	23
14	Emily Blunt	22.5
15	Nicole Kidman	22
16	Ellen Pompeo	19
17	Mila Kunis	16
18	Elizabeth Moss	16
19	Viola Davis	15.5
20	Cate Blanchett	12.5

Summary: Charts

- **Line graph**: sequential data (over time, etc.)
- **Scatter plot**: relation between two numerical variables
- **Bar chart**: distribution of one categorical variable *or* relation between a categorical and a numerical variable
- **Histogram**: distribution of one numerical variable

Discussion Question

You have data about daily temperatures as shown. Which type of chart would show the answer to each question?

- Are there more cloudy than sunny days?
- What percentage of days have a high at least 72º?
- Do days with hotter highs tend to have hotter lows?

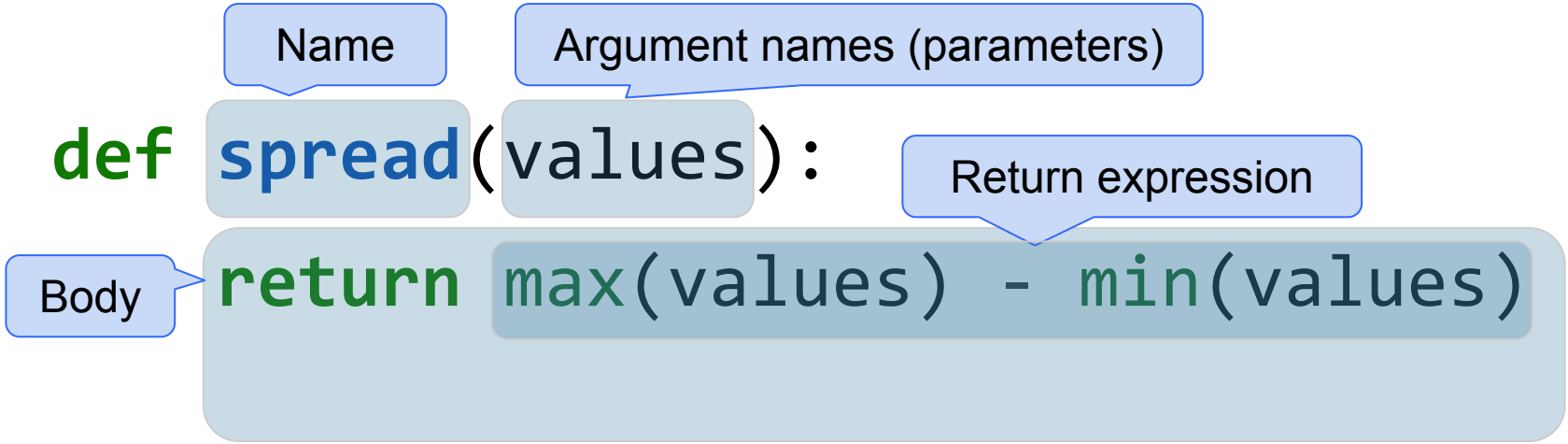
Day	High	Low	Sky condition
1	55.1	43.7	Cloudy
2	57.2	46	Sunny
3	56.8	45.9	Cloudy

... (362 rows omitted)

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

`apply`

1. Calls a function on every element in the input column(s)
2. Produces an array containing the output of the function on each input column element.

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

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

(Demo)