DATA 8

Spring 2020

Lecture 9

Functions

Announcements

- Lab 1/2, HW1 regrade requests **due tonight**
  - Triple check okpy AND gradescape
- Midterm will be on **March 13, 7-9PM**
  - Excused conflict? Fill this form out ASAP
- HW3 due Thurday

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`

Density

(Demo)

How to Calculate Height

The [40, 65) bin contains 51 out of 200 movies

- “51 out of 200” is 25.5%
- The bin is 65 - 40 = 25 years wide

$$\text{Height of bar} = \frac{25.5 \text{ percent}}{25 \text{ years}}$$

$$= 1.02 \text{ percent per year}$$

Height Measures Density

$$\text{Height} = \frac{\%\text{ in bin}}{\text{width of bin}}$$

- The height measures the percent of data in the bin ***relative to the amount of space in the bin***.
- Height measures crowdedness, or **density**.
- Units: percent per unit on the horizontal axis

Area Measures Percent

Area of bar = % in bin = Height x width of bin

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

Charts Review

(Demo)

Bar Chart or Histogram?

To display a distribution:

Bar Chart

- Distribution of categorical variable
- Bars have arbitrary (but equal) widths and spacings
- **height (or length)** and **area** of bars proportional to the percent of individuals

Histogram

- Distribution of numerical variable
- Horizontal axis is numerical: to scale, no gaps, bins can be unequal
- **Area** of bars proportional to the percent of individuals; **height** measures density

Review: Charts

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

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 above 72º?
- Do hotter days tend to also have hotter nights?

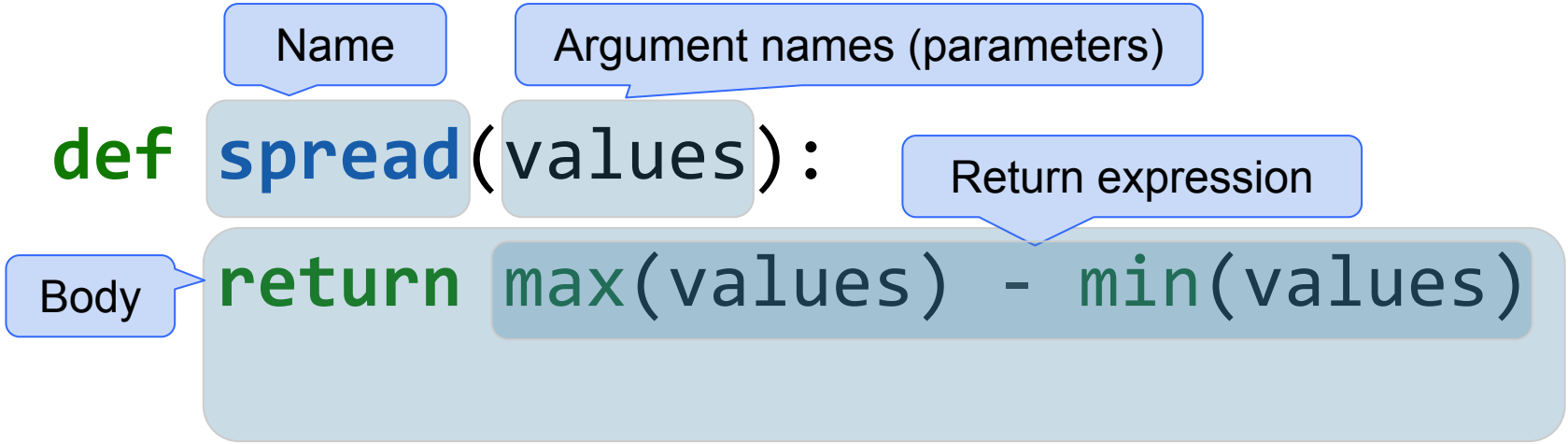
Day	High	Low	Sky condition
1	55.1	43.7	Cloudy
2	57.2	46	Sunny
3	56.8	45.9	Cloudy

... (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

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)