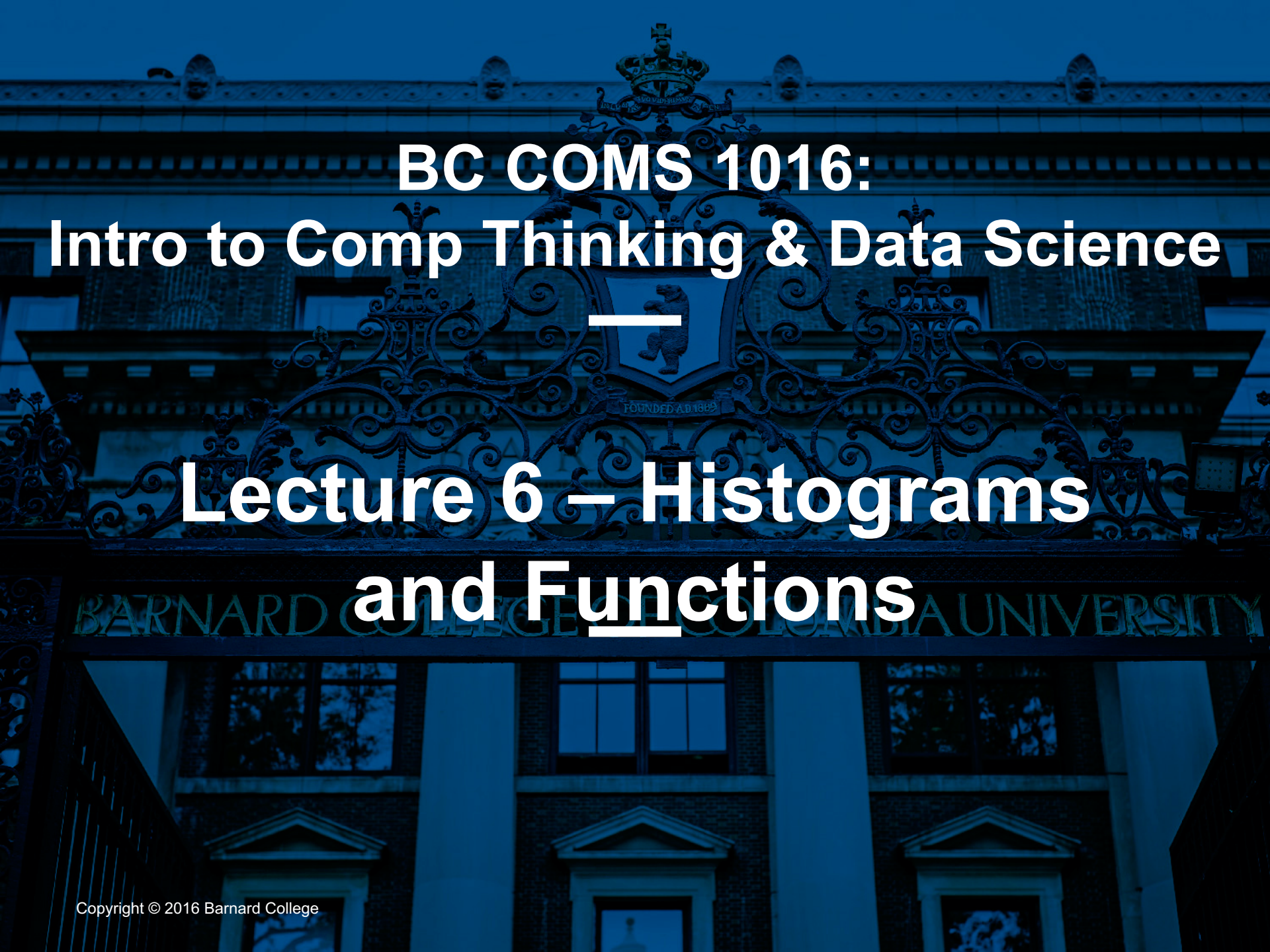


BC COMS 1016: Intro to Comp Thinking & Data Science

—

Lecture 6 – Histograms and Functions

Announcements

- Lab02 (Data Types and Arrays)
  - Due tomorrow night (Friday 11/06)
- HW02 - Table Manipulation & Visualization:
  - Due Monday (11/09)
- Checkpoint/Project 1:
  - Paired assignment that covers the previous section of the course material
  - Released Wednesday (11/11) and due Wednesday (11/18)

Course Outline

- Exploration **Week 1 - 3**
  - Discover patterns in data
  - Articulate insights (visualizations)
- Inference **Week 3 - 5**
  - Make reliable conclusions about the world
  - Statistics is useful
- Prediction **Week 6-7**
  - Informed guesses about unseen data

Checkpoints/Projects – Paired assignments

- 3 checkpoints/projects:
  - Flexible grading scheme – weight based on scores
    - Best performing one will count as “midterm”
    - Remaining two will count as 20% in “project” from first lecture
- Exploration checkpoint/project
  - Released 11/11 (maybe earlier)
  - Due Wednesday 11/18
  - HW3 & HW4 are on the shorter side

Visualization Review

- Line plots
  - Sequential data
- Scatter plots
  - Finding associations
- Bar plots
  - Categorical distributions

Bar Plots

- Display relationship between categorical variable and a numerical variable

- Display a categorical distribution

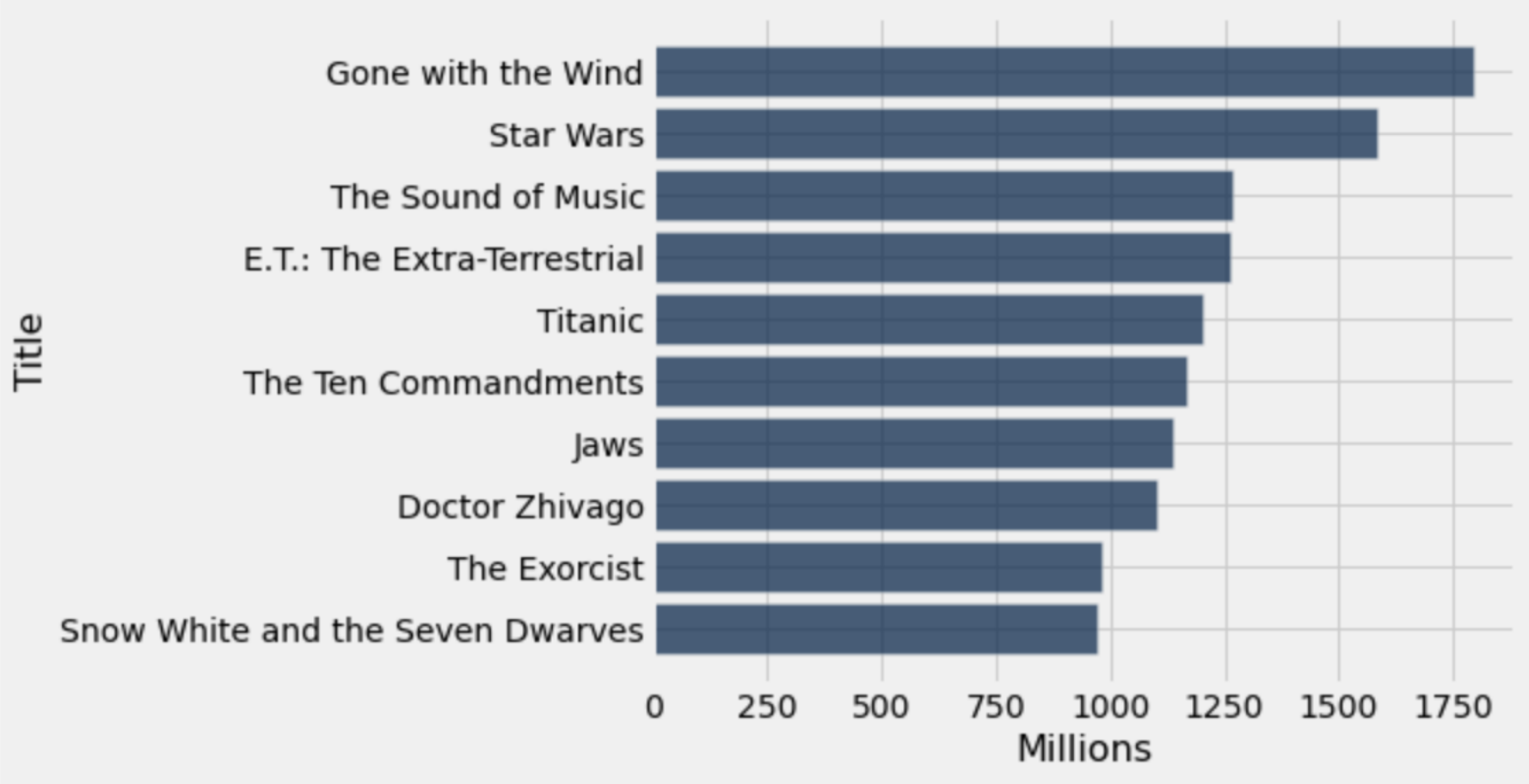
Bar Charts

```
top_movies = Table.read_table('top_movies_2017.csv')
top_movies
```

Title	Studio	Gross	Gross (Adjusted)	Year
Gone with the Wind	MGM	198676459	1796176700	1939
Star Wars	Fox	460998007	1583483200	1977
The Sound of Music	Fox	158671368	1266072700	1965
E.T.: The Extra-Terrestrial	Universal	435110554	1261085000	1982
Titanic	Paramount	658672302	1204368000	1997
The Ten Commandments	Paramount	65500000	1164590000	1956
Jaws	Universal	260000000	1138620700	1975
Doctor Zhivago	MGM	111721910	1103564200	1965
The Exorcist	Warner Brothers	232906145	983226600	1973
Snow White and the Seven Dwarves	Disney	184925486	969010000	1937

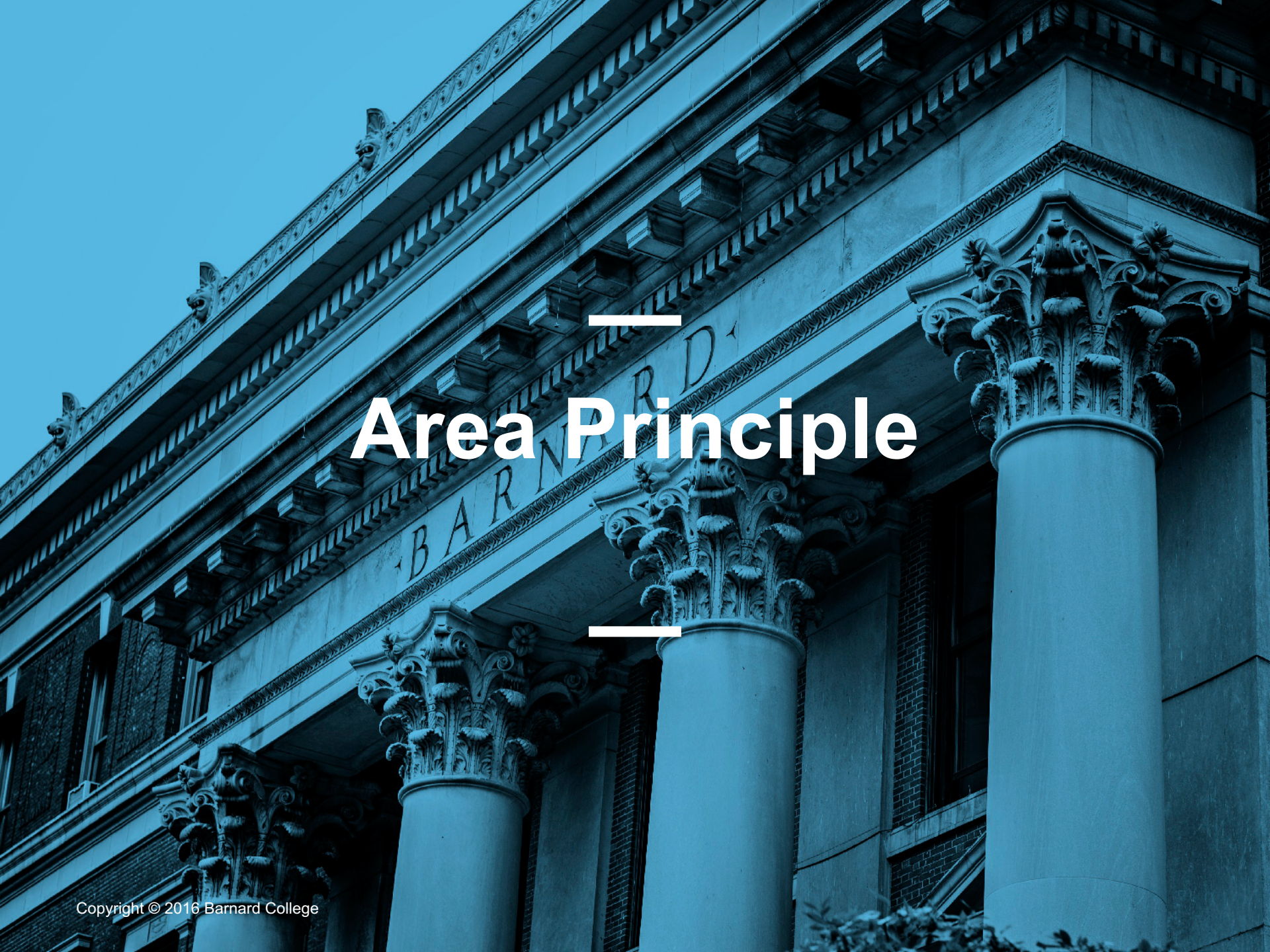
Bar plot

```
: top10_adjusted.barh('Title', 'Millions')
```



Displaying a Categorical Distribution

- Distribution of a variable describes the frequencies of the values
- The group method counts the number of values in the column
- Bar chart displays the distribution of categorical variable:
  - One bar per category/value
  - Length of bar is the count of individuals in that category



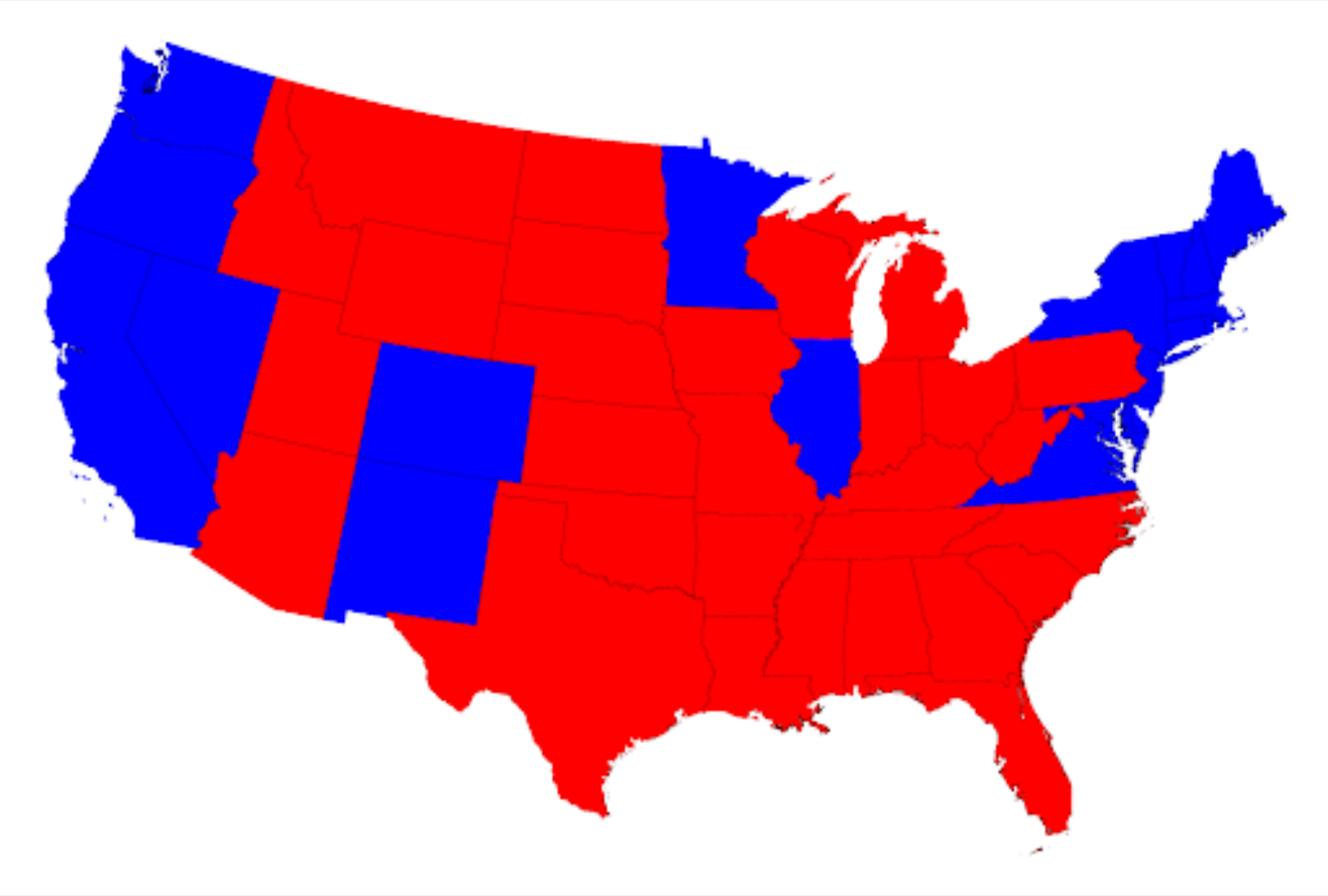
Area Principle

Area Principle

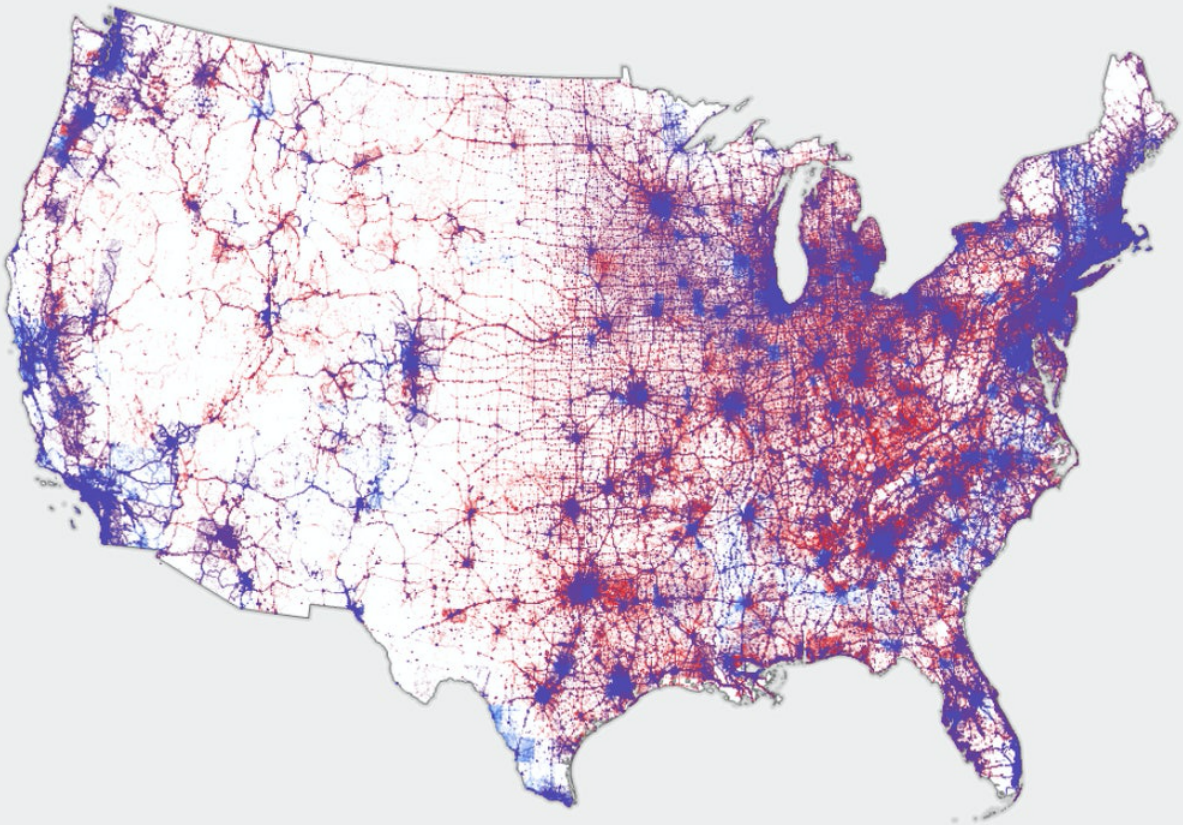
Areas should be proportional to values they represent

- If you represent 20% by 
- 40% should be represented by 
- and not by 

Area Principle – 2016 Election Map



Area Principle – 2016 Election Map



https://www.wired.com/story/is-us-leaning-red-or-blue-election-maps/

Plotting Numerical Distributions

- **Binning** coverts a numerical distribution to a categorical distribution

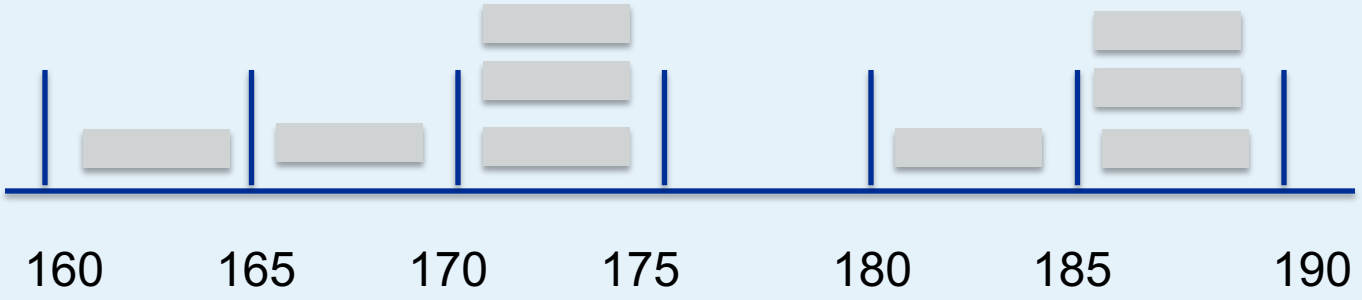
- **Binning** counts the number of numerical values that lie within a range, aka a bin

- Bins contain:
  - A lower bound (inclusive)
  - An upper bound (exclusive)

Bins - Example

- Bins contain:
  - A lower bound (inclusive)
  - An upper bound (exclusive)

188, 170, 189, 163, 183, 171, 185, 168, 173, ...



Histogram

Chart that displays the distribution of a numerical variable

Uses bins; there is one bar corresponding to each bin

Uses the area principle:

- The ***area*** of each bar is the percent of individuals in the corresponding bin

Understanding Histograms

- Axes
- Height
- Area

Histogram Axis

- By default, hist uses a scale (normed=True) that ensures the area of the chart sums to 100%
- The **area** of each bar is a percentage of the whole
- The horizontal (x-) axis is a number line (e.g., years), and the bins sizes don't have to be equal to each other
- The vertical axis is a rate (e.g., percent per year)

Histogram Height (of a bin)

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

- Height measures density
- the percent of data in the bin ***relative to the amount of space in the bin***
- Units: percent per unit on the horizontal axis

Histogram Area (of a bar)

- Area tells us what percent of all data is in a bin
- Area of a bar = Height times width of a bin
  - *"How many individuals in the bin?"* Use area.
  - *"How crowded is the bin?"* Use height

Bar Chart or Histogram?

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

Functions

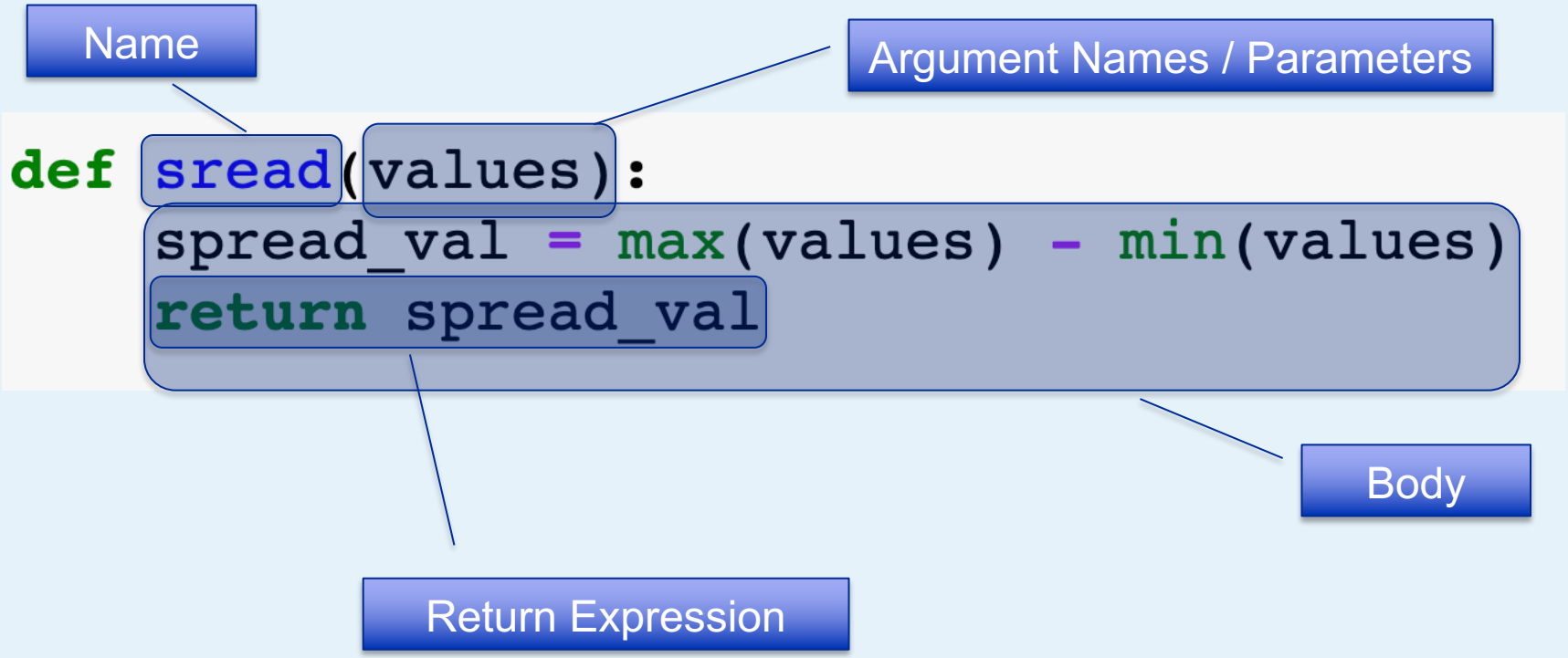
Anatomy of a Function

- Name
- Parameters / Argument Names
- Body
- Return Expression

Example Function

```
def sread(values):
    spread_val = max(values) - min(values)
    return spread_val
```

Example Function



What does this function do?

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

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