DATA 8

Fall 2016

# Lecture 7, September 9

## Visualization

Slides created by Ani Adhikari and John DeNero

# Announcements

- **Waitlisted students:** I have a meeting today about enrollment. I will email all waitlisted students after that.
- No late work. If you joined the class late, please do current work. We'll prorate based on when you joined.
- **Technical problems? Questions about hw/lab credit?** Please email your GSI, not me. Addresses are on the Staff Contact page in data8.org.
- **Concurrent Enrollment:** The class will fill up with registered students. Please try CS 10 or other courses.
- **Auditors:** data8.org and textbook are public. Lecture video needs a Berkeley email account. No other materials; sorry.

# Methods involving rows

Each of these methods creates a new table, containing:

- all of the rows, arranged in increasing or decreasing order of the values in one column
    - `sort`
- a specified set of rows
    - `take`
- all rows that satisfy a condition
    - `where`

# Quick check

The table bubble_tea has 19 rows, one for each tea on a cafe’s menu. The columns are Flavor and Price, in that order. One of the flavors is Garlic. Write one line of code that evaluates to:

a) A table of all the rows with any flavor but Garlic
b) A flavor that has the lowest of all the prices
c) A table consisting of rows 3, 7, 11, ...

Answers:

```
bubble_tea.where(‘Flavor’, are.not_equal_to(‘Garlic’))
bubble_tea.sort(‘Price’).column(‘Flavor’).item(0)
bubble_tea.take(np.arange(3, 27, 4))
```

A number > 15

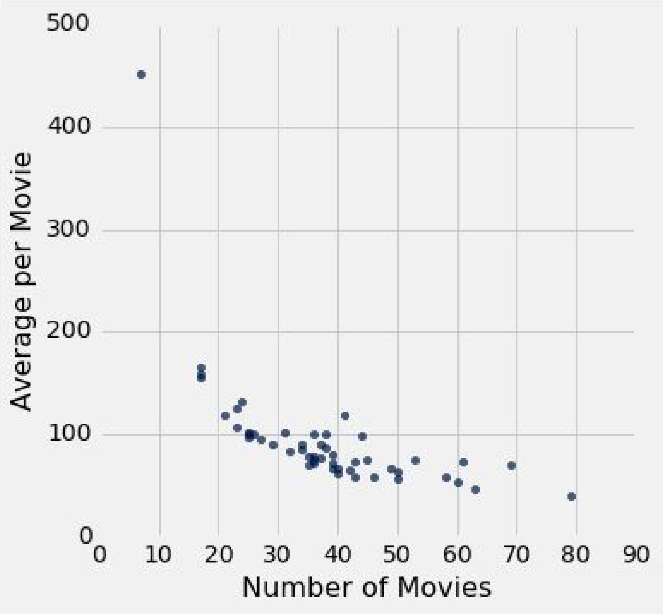
# Visualization

A picture is worth a thousand numbers.

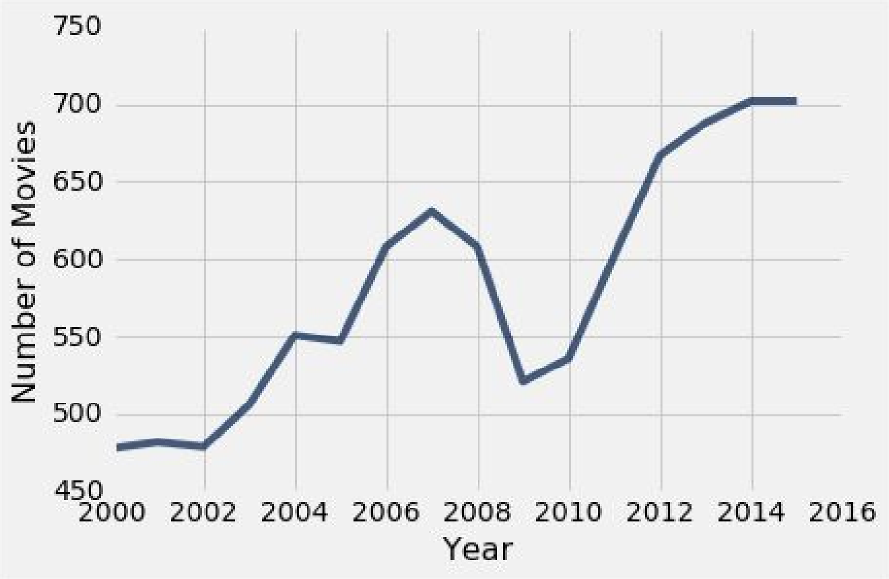
(Demo)

# Plotting Two Numerical Variables

Scatter plot: `scatter`



Line graph: `plot`

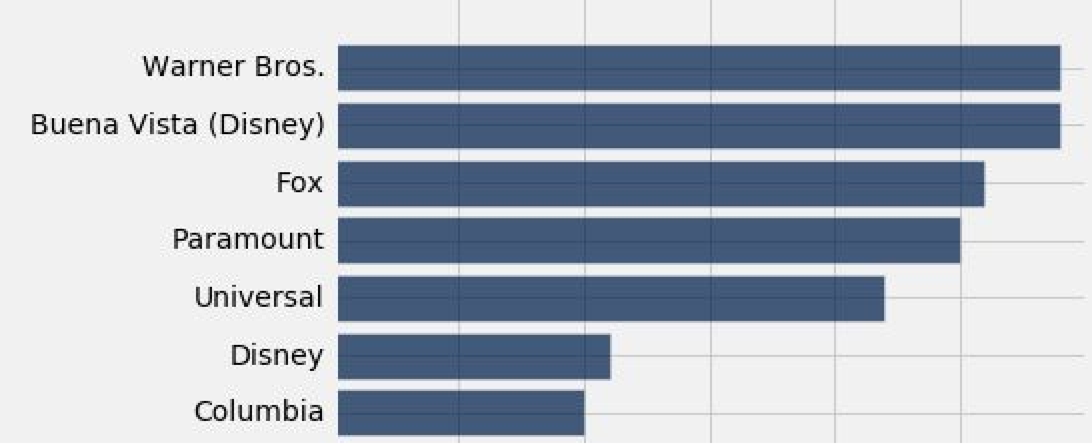


# Terminology

- **Individuals**: those whose features are recorded
- **Variables**: features; these vary across individuals
- Variables have different **values**
- Values can be **numerical**, or **categorical**, or of many other types
- **Distribution**: For each different value of the variable, the frequency of individuals that have that value
- Frequency is measured in counts. Later we will use proportions or percents.

# Categorical Distributions

bar chart: `barh`



Displays a categorical distribution

(But when the values of the variable have a rank ordering, or fixed sizes relative to each other, more care might be needed.)