CompSci 190: Visualization & Graphs

Jeff Forbes

September 17, 2018

Plan For The Week (PFTW)

- Do Homework 2
- Consider different methods for visualizations of data
  - Types of charts
    - Scatter, line & bar
    - Histograms
  - Distributions
    - Categorical
    - Numerical

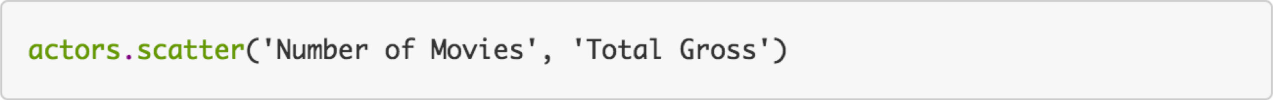
Types of Data

- Tables enforce constraints
  - All values in a column are the same type
  - Values in a column are *comparable*

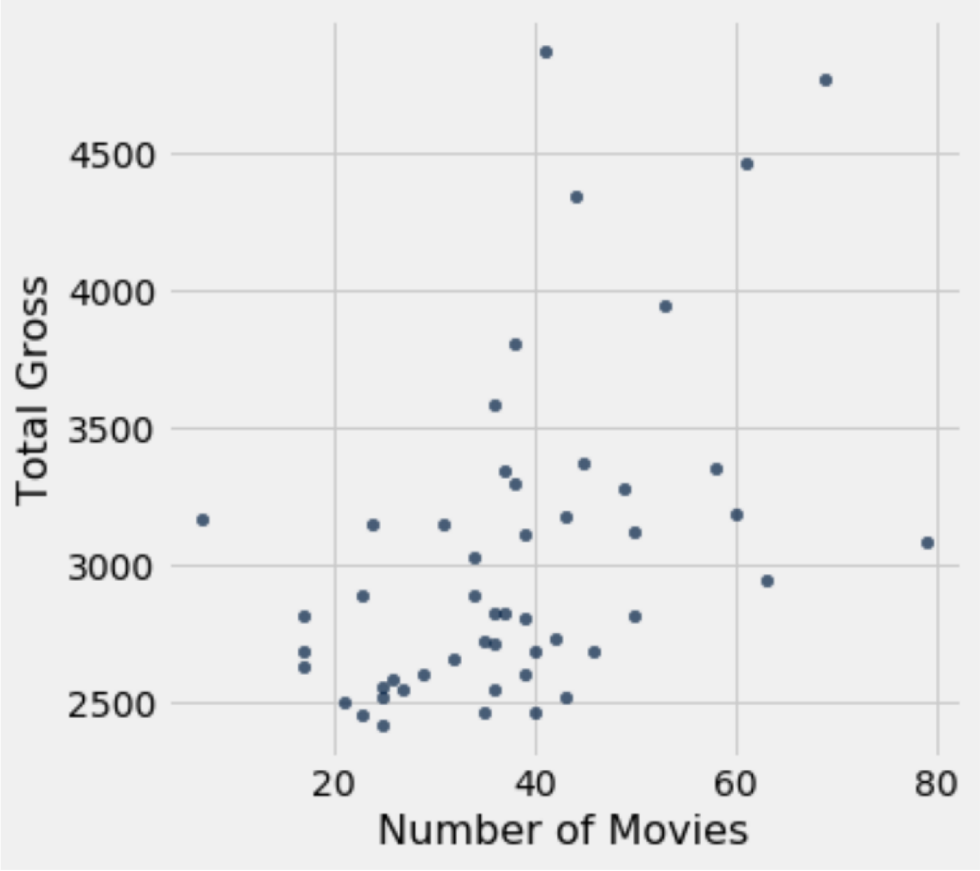
- **Numerical** — Each value is from a numerical scale
  - Numerical measurements are ordered
  - Differences are meaningful
- **Categorical** — Each value is from a fixed inventory
  - May or may not have an ordering
  - Categories can be different

Scatter Plot

- Relation/association between two numerical values
- Arguments
  1. Label of column for horizontal (x) axis
  2. Label of column for vertical (y) axis

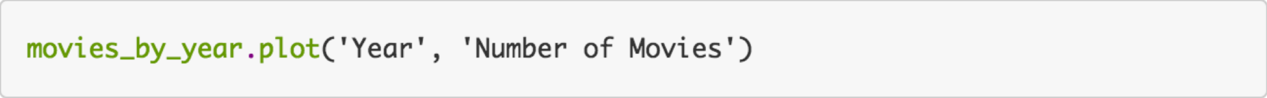


```
actors.scatter('Number of Movies', 'Total Gross')
```

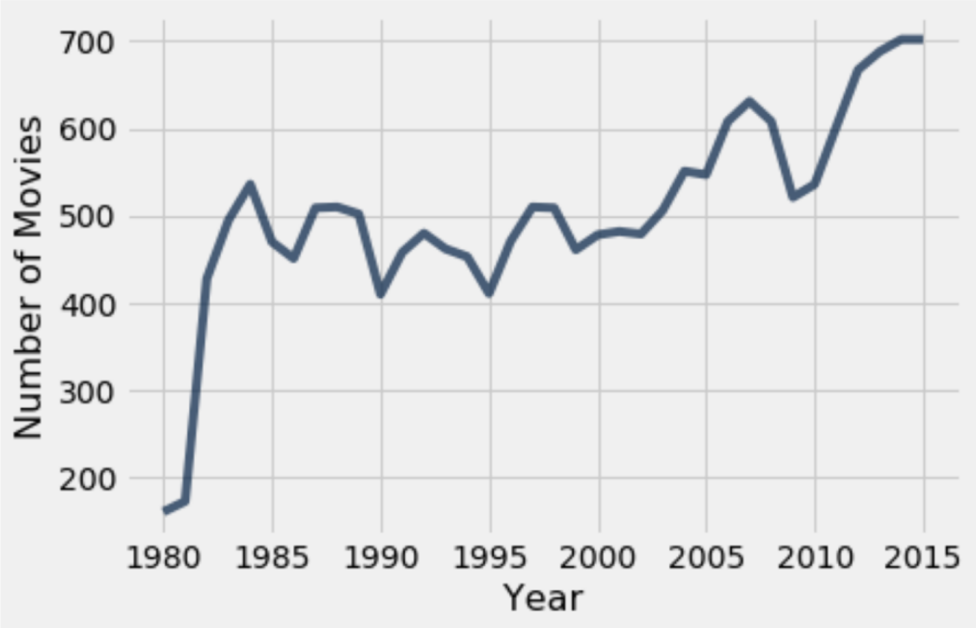


Line Graph

- **Use:** chronological trends
- Arguments
  1. Label of column for horizontal (x) axis
  2. Label of column for vertical (y) axis

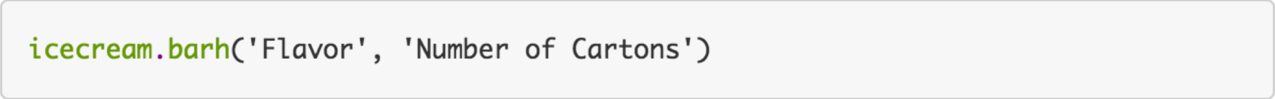


```
movies_by_year.plot('Year', 'Number of Movies')
```

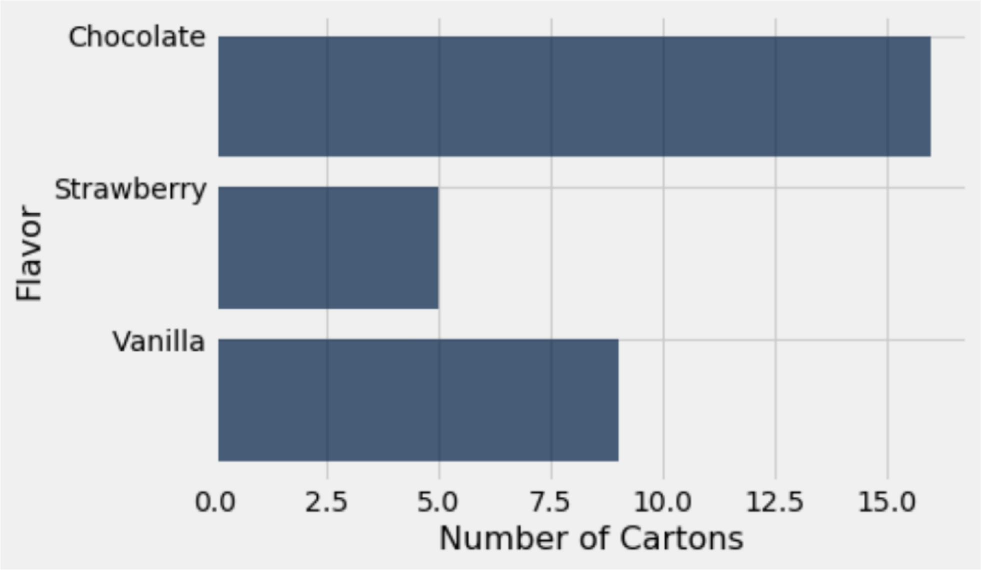


Bar Chart

- Categorical distributions
  - Implications?
    - Width of bars
    - Ordering of categories
- Arguments
  1. Label of column for categories
  2. Label of column for frequencies

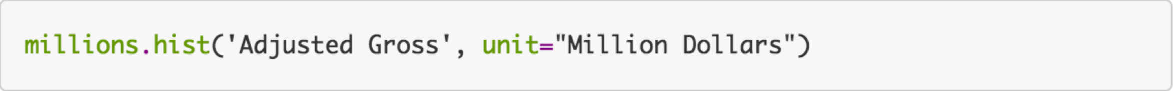


```
icecream.barh('Flavor', 'Number of Cartons')
```

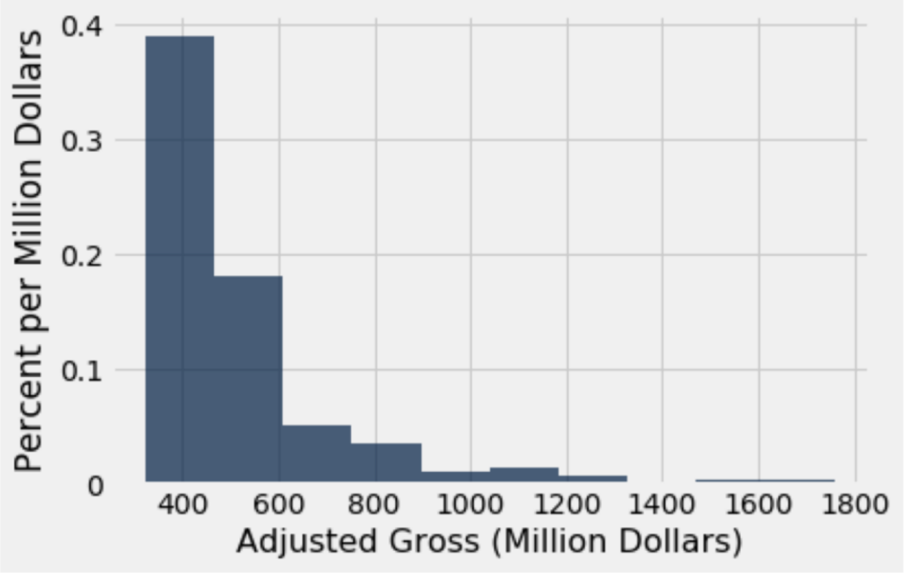


Histograms

- Numerical distributions
  - Implications?
    - Width of bars
- Arguments
  1. Values to display
- Optional arguments
  - `unit`: label for axes
  - `bins`: endpoints for buckets
  - `normed`: display proportion instead of counts



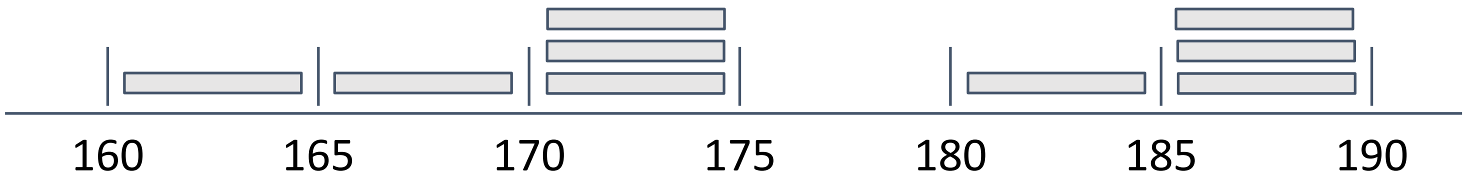
```
millions.hist('Adjusted Gross', unit="Million Dollars")
```



Binning numerical values

- Binning: # of numerical values that lie within ranges (bins)
    - Bins are defined by their lower bounds (inclusive)
    - The upper bound is the lower bound of the next bin

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



The [185,190) bin

Histogram Axes

By default, `hist` uses a scale (`normed=True`) that ensures the area of the chart sums to 100%

- The horizontal axis is a number line (e.g., years)
- The vertical axis is a rate (e.g., percent per year)
- The area of a bar is a percentage of the whole

How to Calculate Height

The [20, 40) bin contains 59 out of 200 movies

- “59 out of 200” is 29.5%
- The bin is 40 - 20 = 20 years wide

$$\text{Height of bar} = \frac{29.5 \text{ percent}}{20 \text{ years}}$$

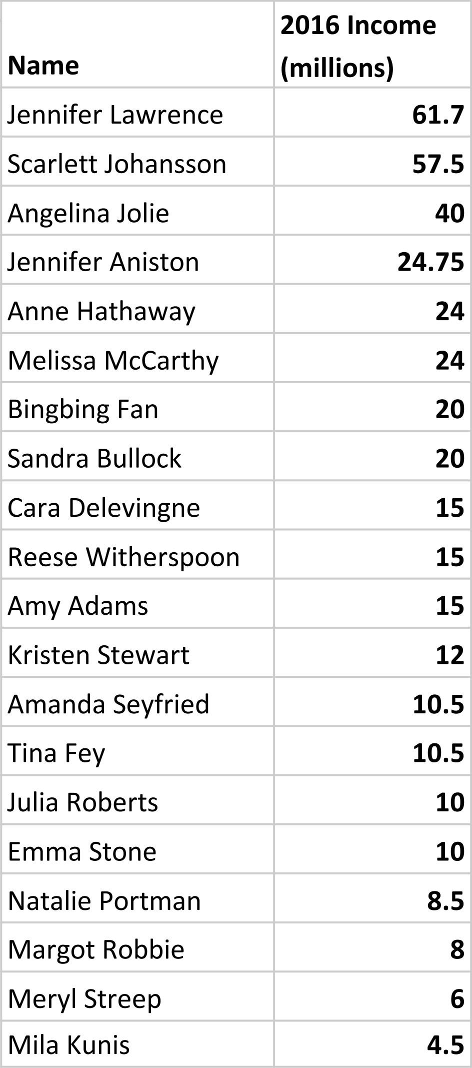
$$= 1.475 \text{ percent per year}$$

Area Measures Percent

Area = % in bin = Height x width of bin

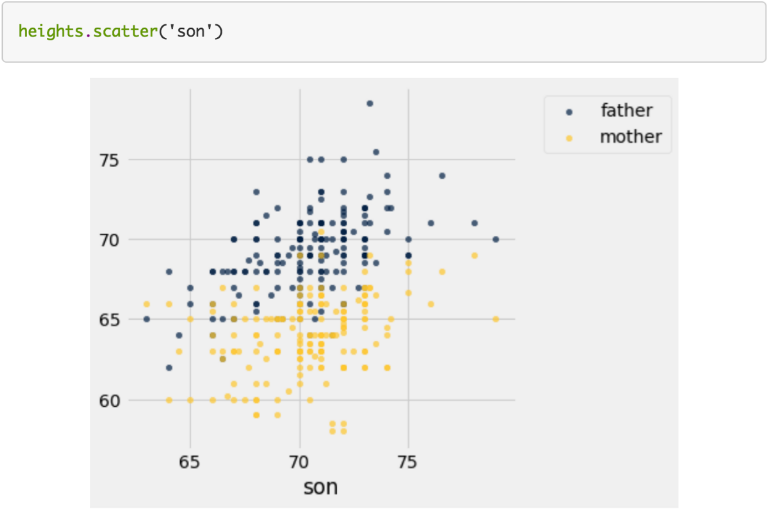
- "How many individuals in the bin?" Use area.
- "How crowded is the bin?" Use height.

- What would the y-axis of a histogram of this table be?
- http://bit.ly/FoDS-f18-0917-1

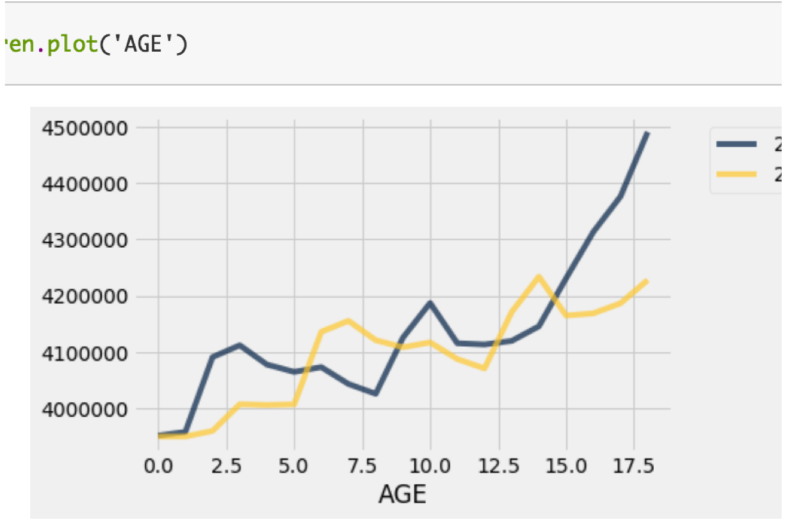


Name	2016 Income (millions)
Jennifer Lawrence	61.7
Scarlett Johansson	57.5
Angelina Jolie	40
Jennifer Aniston	24.75
Anne Hathaway	24
Melissa McCarthy	24
Bingbing Fan	20
Sandra Bullock	20
Cara Delevingne	15
Reese Witherspoon	15
Amy Adams	15
Kristen Stewart	12
Amanda Seyfried	10.5
Tina Fey	10.5
Julia Roberts	10
Emma Stone	10
Natalie Portman	8.5
Margot Robbie	8
Meryl Streep	6
Mila Kunis	4.5

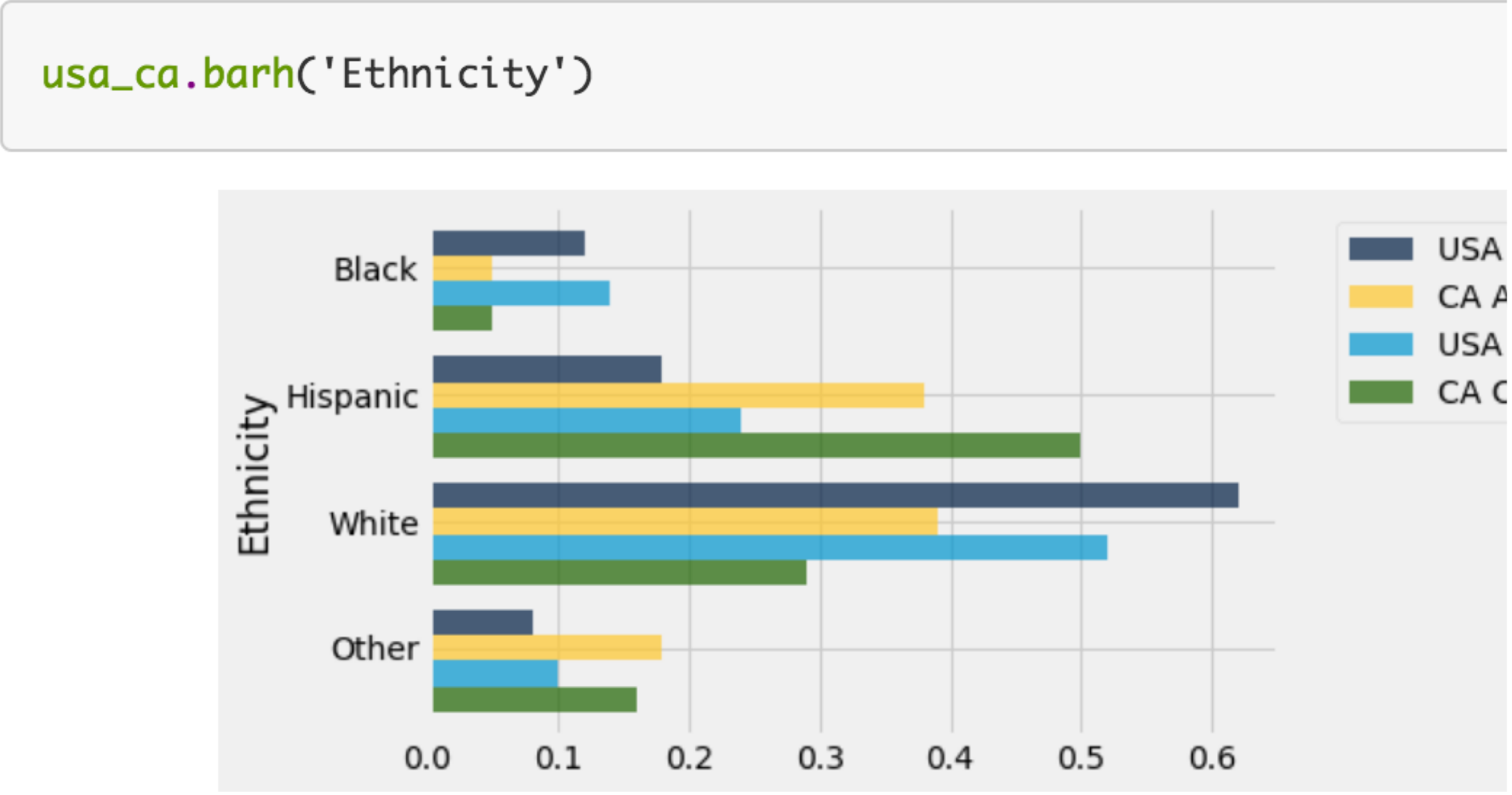
Overlaid Graphs



```
heights.scatter('son')
```



```
en.plot('AGE')
```



```
usa_ca.barh('Ethnicity')
```

What’s next?

- Read Chapter 8 of *Computational and Inferential Thinking*

- Start working on Homework 2 (out tonight)