CompSci 190: Visualization & Graphs

Jeff Forbes

January 31, 2019

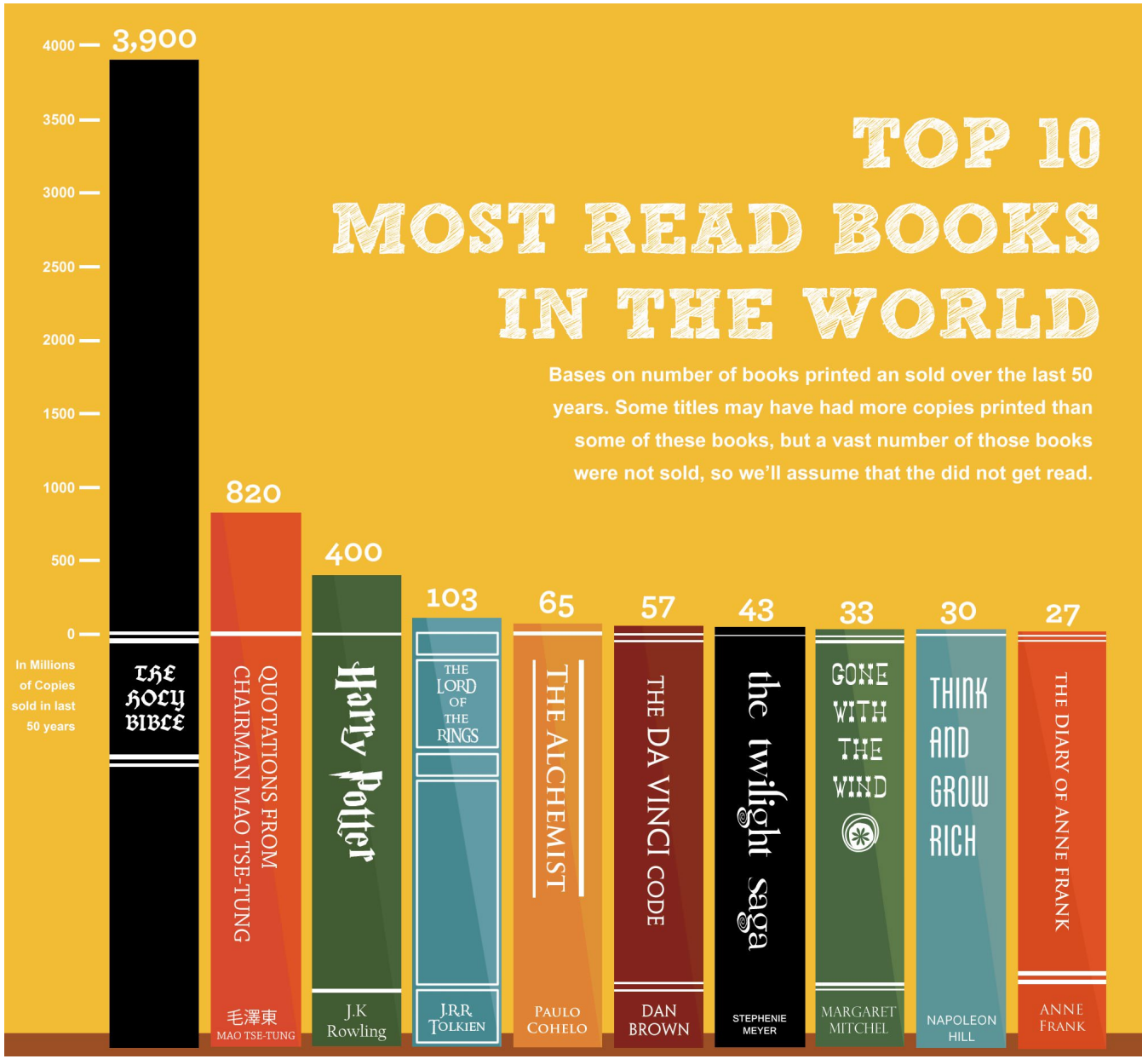
Plan For Today

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

Principles of Visualization

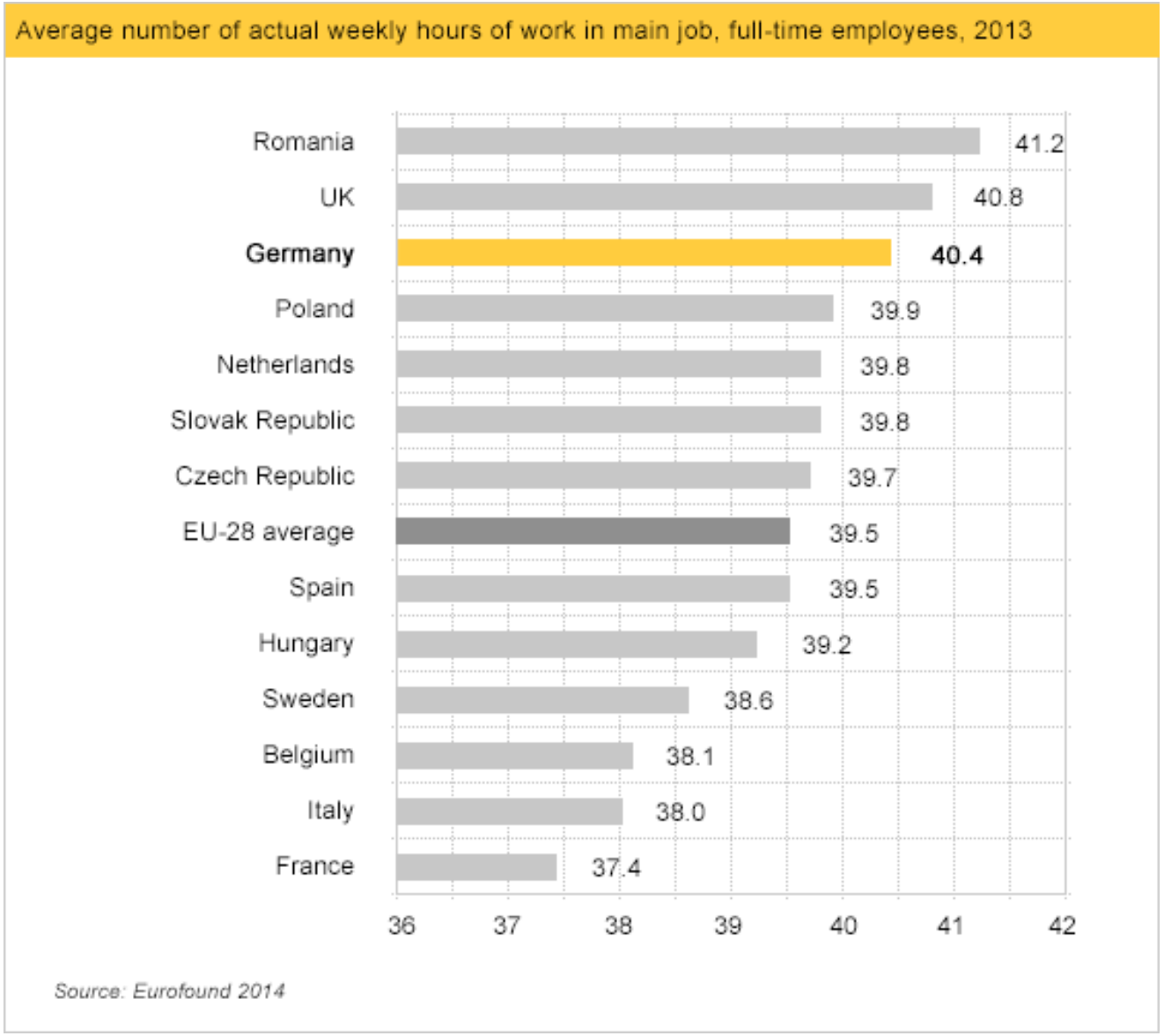
- Above all else, show the data
- Comparison rather than just description
- Maximize the data-ink ratio
- Don’t be misleading
  - Watch your axes

Bad Ink


TOP 10
MOST READ BOOKS
IN THE WORLD
Bases on number of books printed an sold over the last 50 years. Some titles may have had more copies printed than some of these books, but a vast number of those books were not sold, so we'll assume that the did not get read.
In Millions of Copies sold in last 50 years
4000
3500
3000
2500
2000
1500
1000
500
0
3,900
THE HOLY BIBLE
820
QUOTATIONS FROM CHAIRMAN MAO TSE-TUNG
毛澤東
MAO TSE-TUNG
400
Harry Potter
J.K Rowling
103
THE LORD OF THE RINGS
J.R.R. TOLKIEN
65
THE ALCHEMIST
PAULO COHELO
57
THE DA VINCI CODE
DAN BROWN
43
the twilight saga
STEPHENIE MEYER
33
GONE WITH THE WIND
MARGARET MITCHEL
30
THINK AND GROW RICH
NAPOLEON HILL
27
THE DIARY OF ANNE FRANK
ANNE FRANK

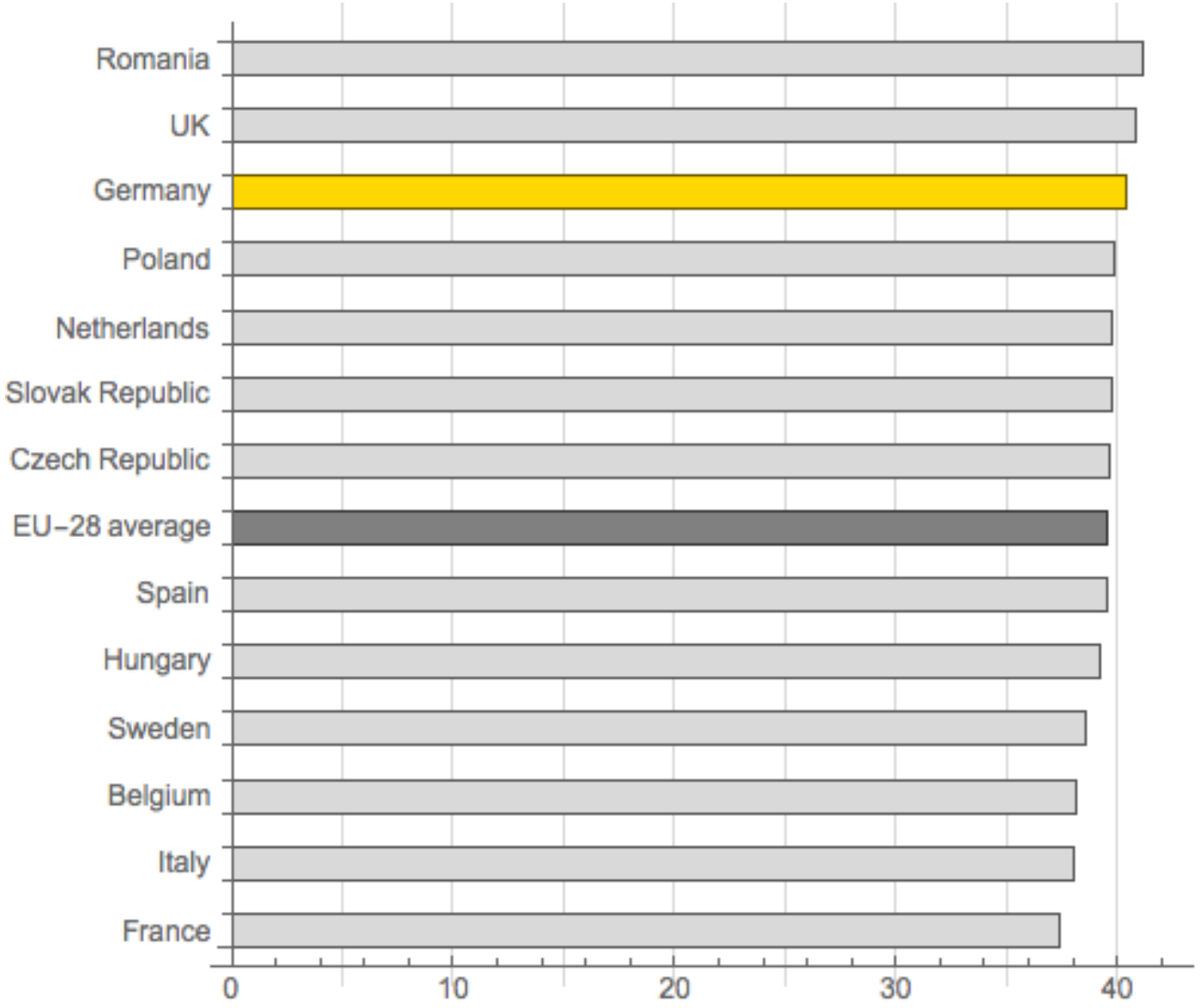
What's up with your axes?

- callingbullsh#t.org



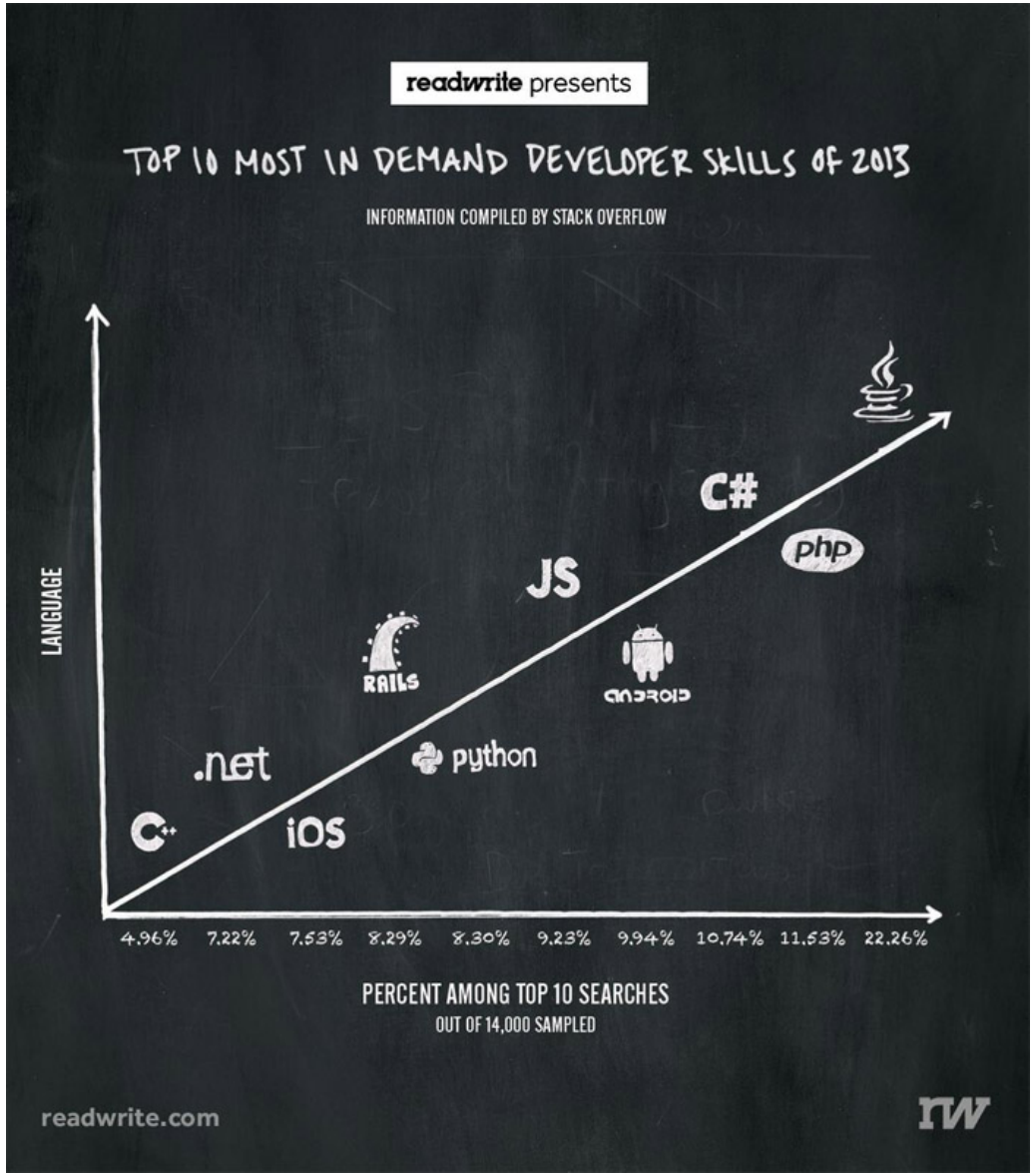
Better axes!

- callingbullsh#t.org



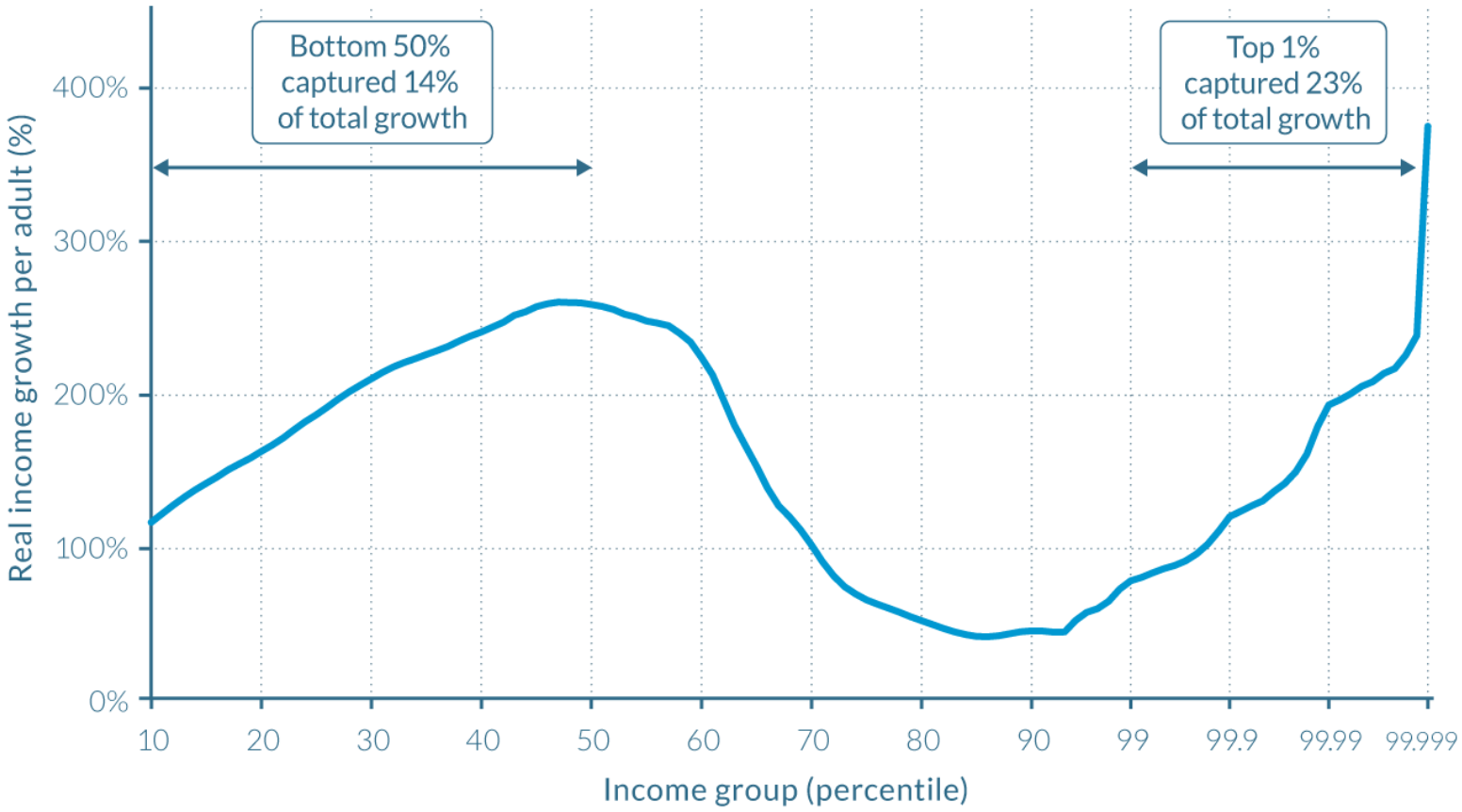
What's up with your axes?

- callingbullsh#t.org



What's up with your axes?

- callingbullsh#t.org



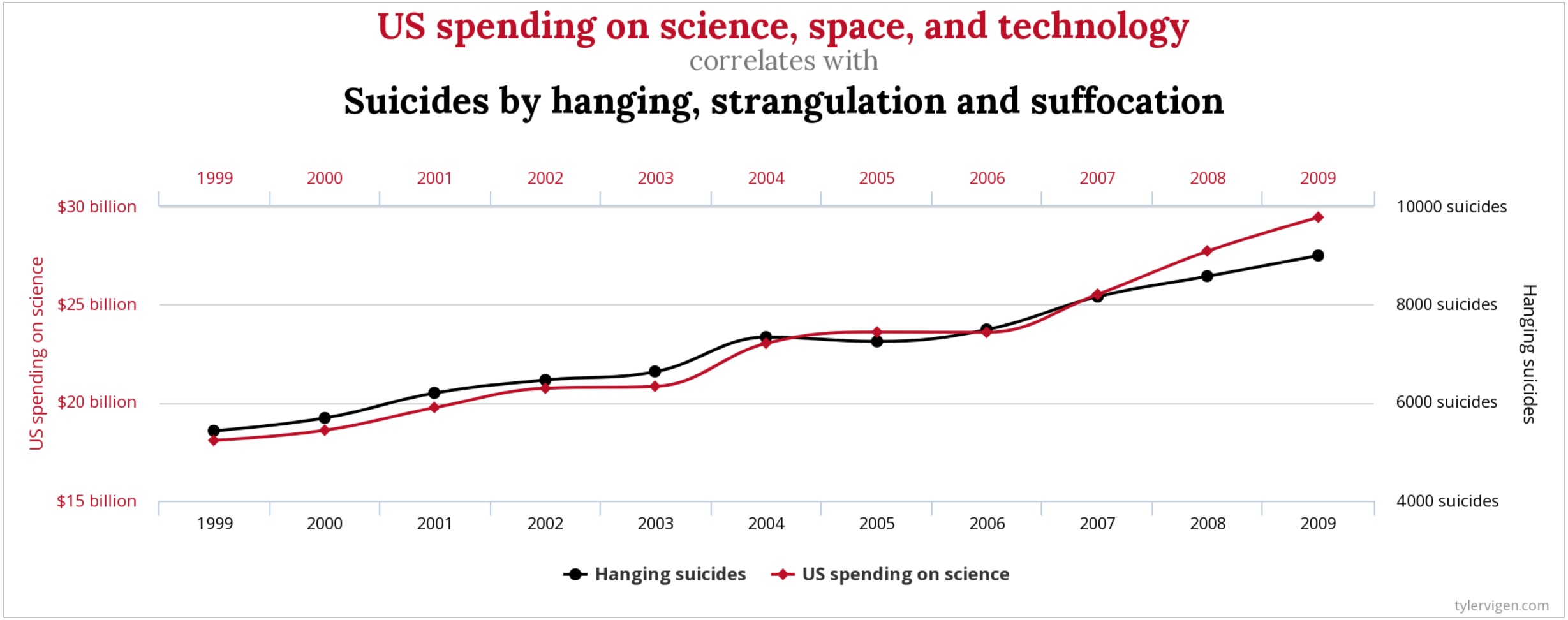
Source: WID.world (2017). See wir2018.wid.world/methodology.html for data series and notes.

On the horizontal axis, the world population is divided into a hundred groups of equal population size and sorted in ascending order from left to right, according to each group's income level. The Top 1% group is divided into ten groups, the richest of these groups is also divided into ten groups, and the very top group is again divided into ten groups of equal population size. The vertical axis shows the total income growth of an average individual in each group between 1980 and 2016. For percentile group p99p99.1 (the poorest 10% among the world's richest 1%), growth was 77% between 1980 and 2016. The Top 1% captured 23% of total growth over this period. Income estimates account for differences in the cost of living between countries. Values are net of inflation.

Bad Axes?

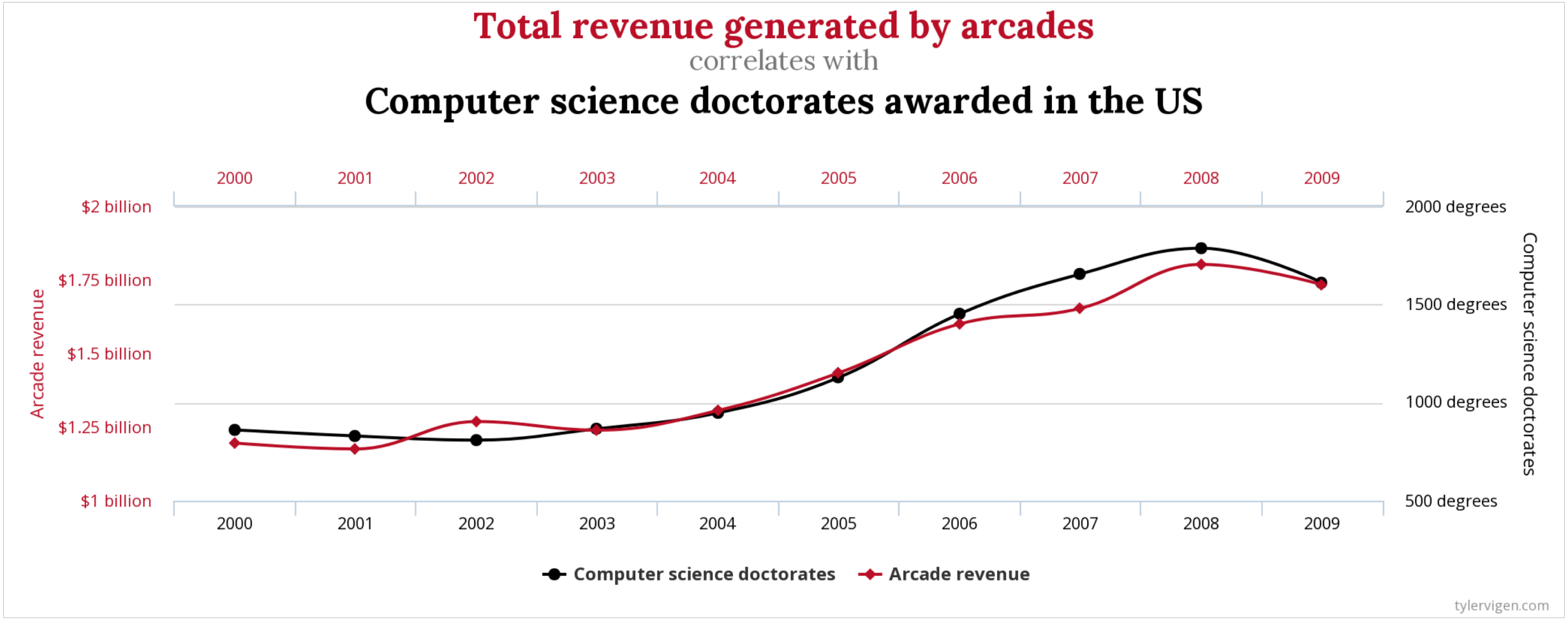

FRED
Employment-Population Ratio: Men (left)
Employment-Population Ratio: Women (right)
Percent
85
80
75
70
65
60
64
56
48
40
32
24
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
Shaded areas indicate U.S. recessions
Source: U.S. Bureau of Labor Statistics
fred.stlouisfed.org

Bad Associations



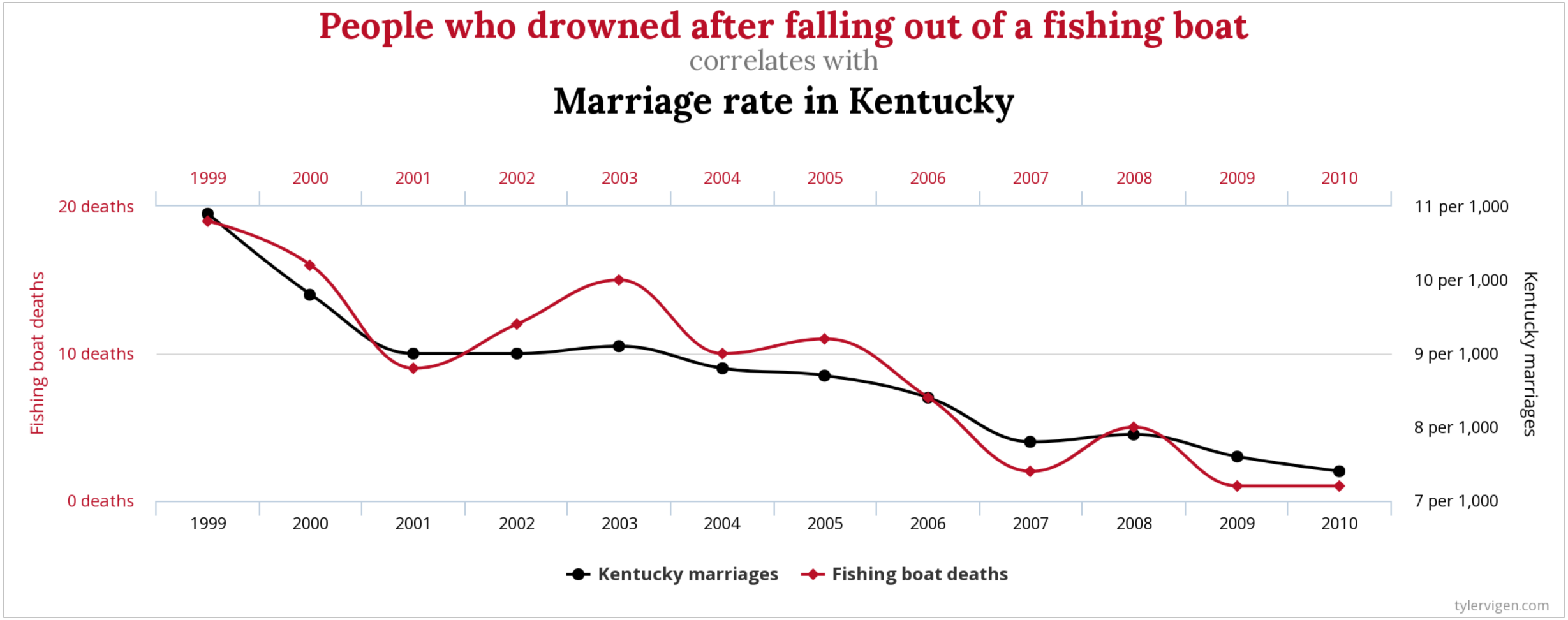
-- Tyler Vigen

Bad Associations



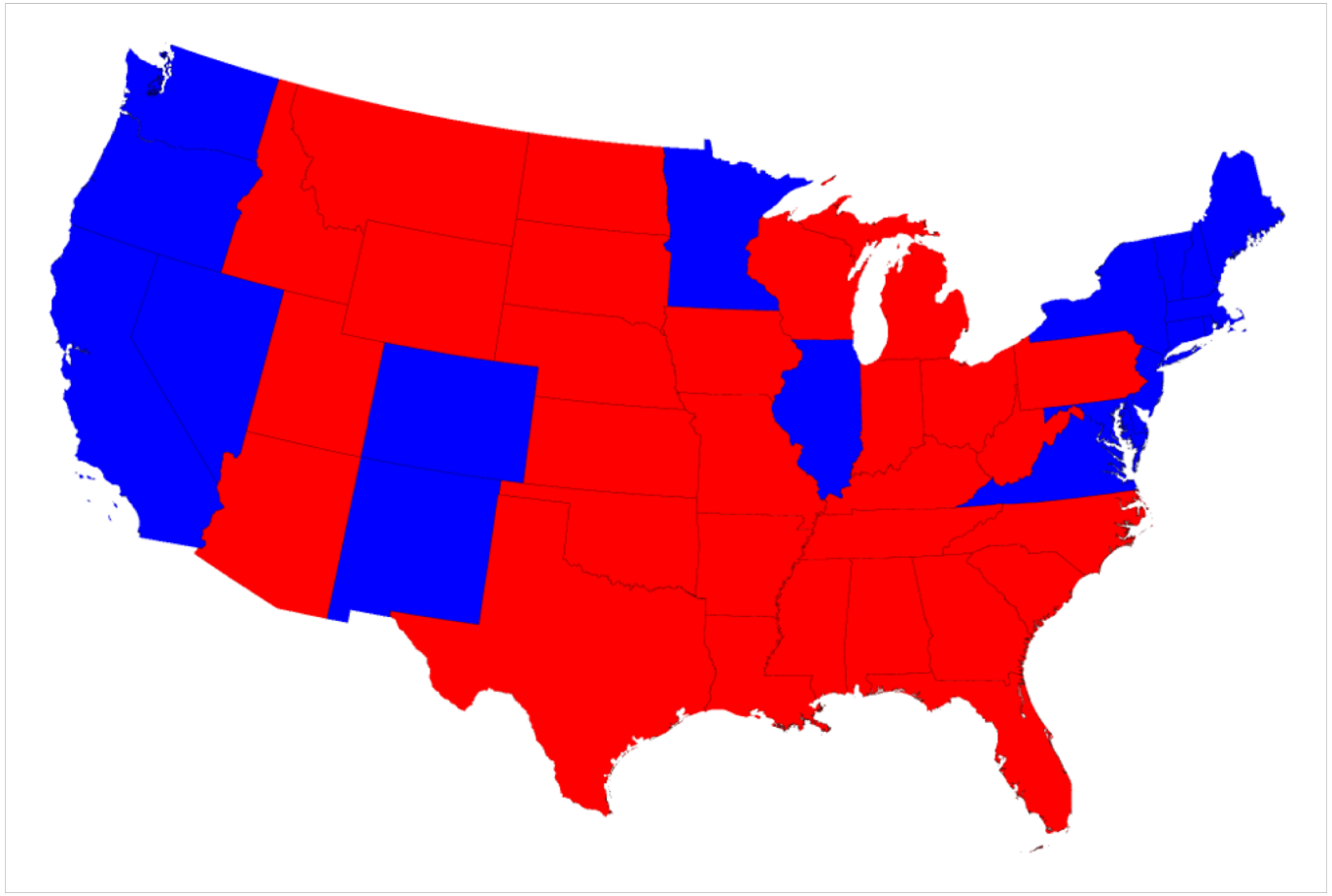
-- Tyler Vigen

Bad Associations



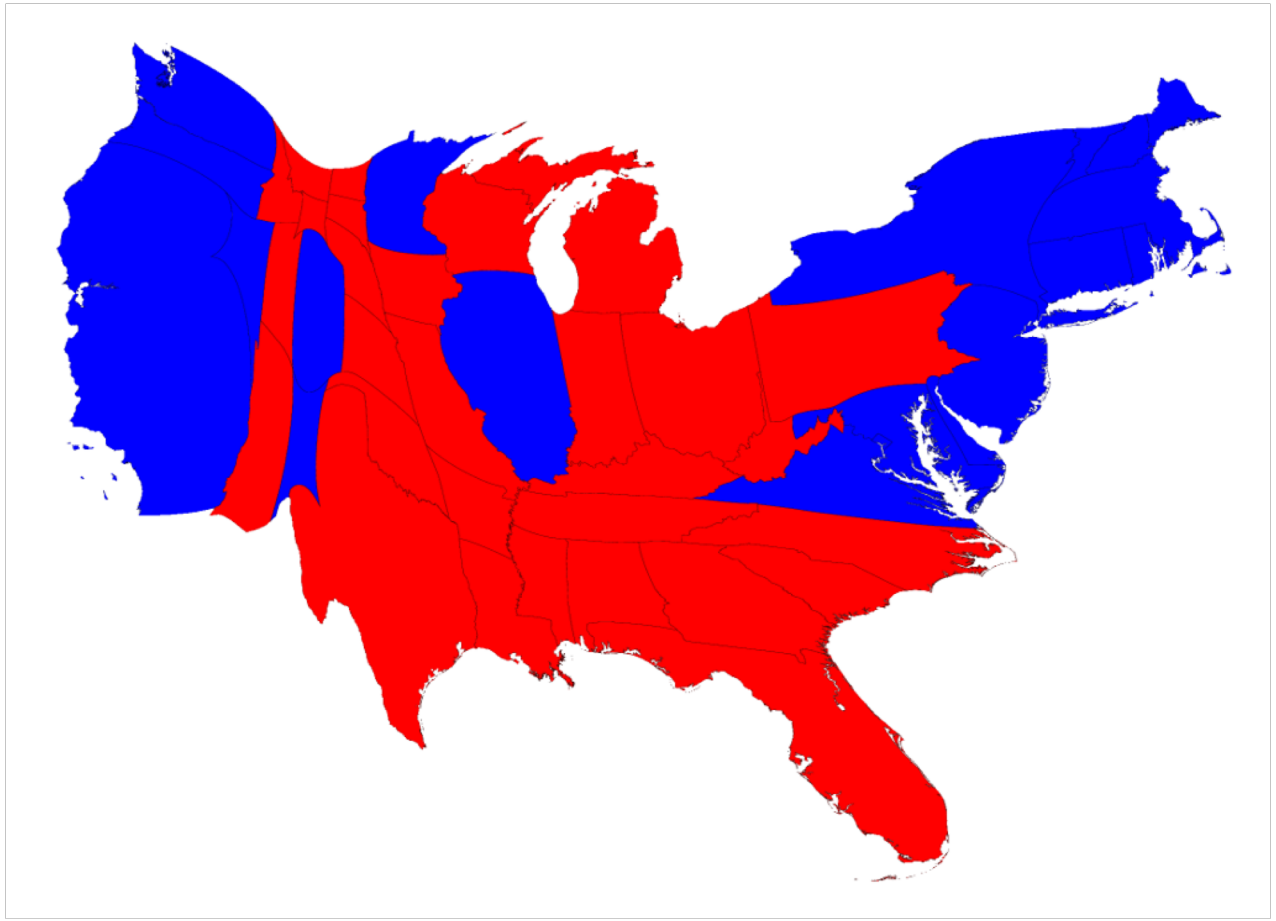
-- Tyler Vigen

2016 Presidential Election Results



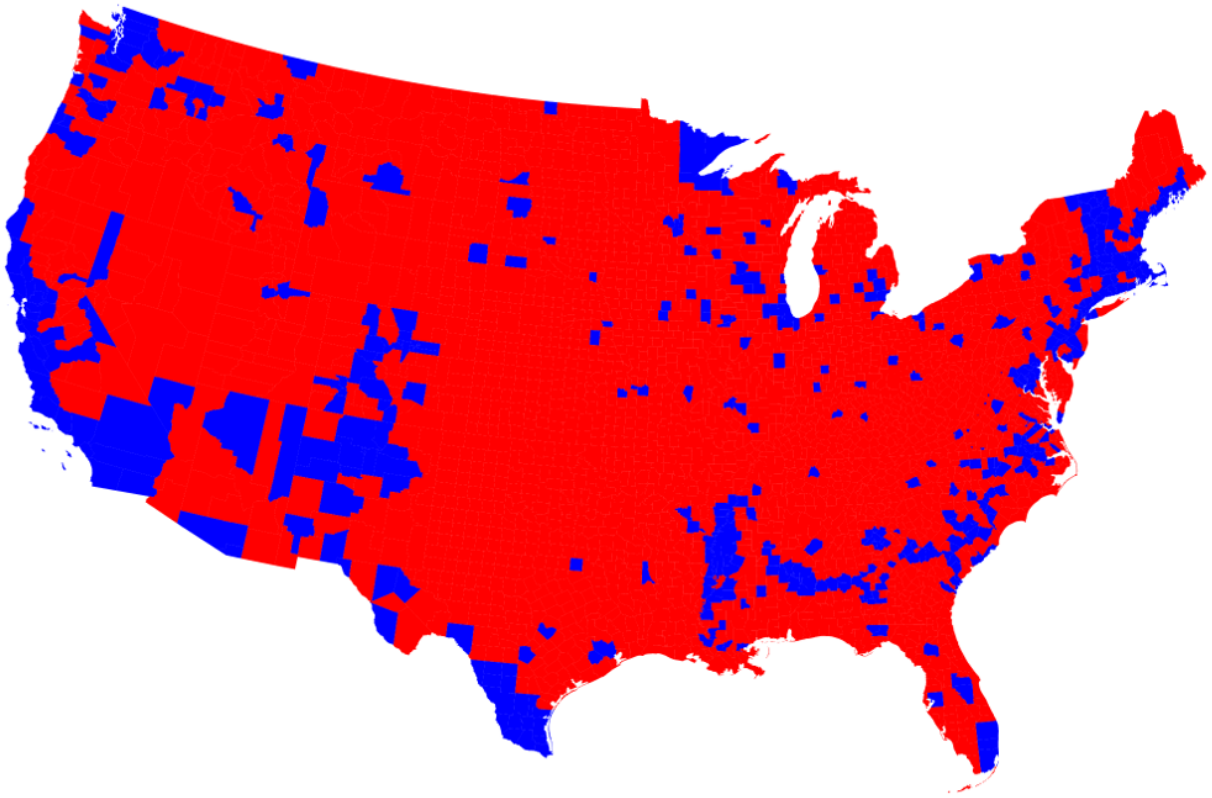
From Mark Newman, U of Michigan

2016 Presidential Election Results



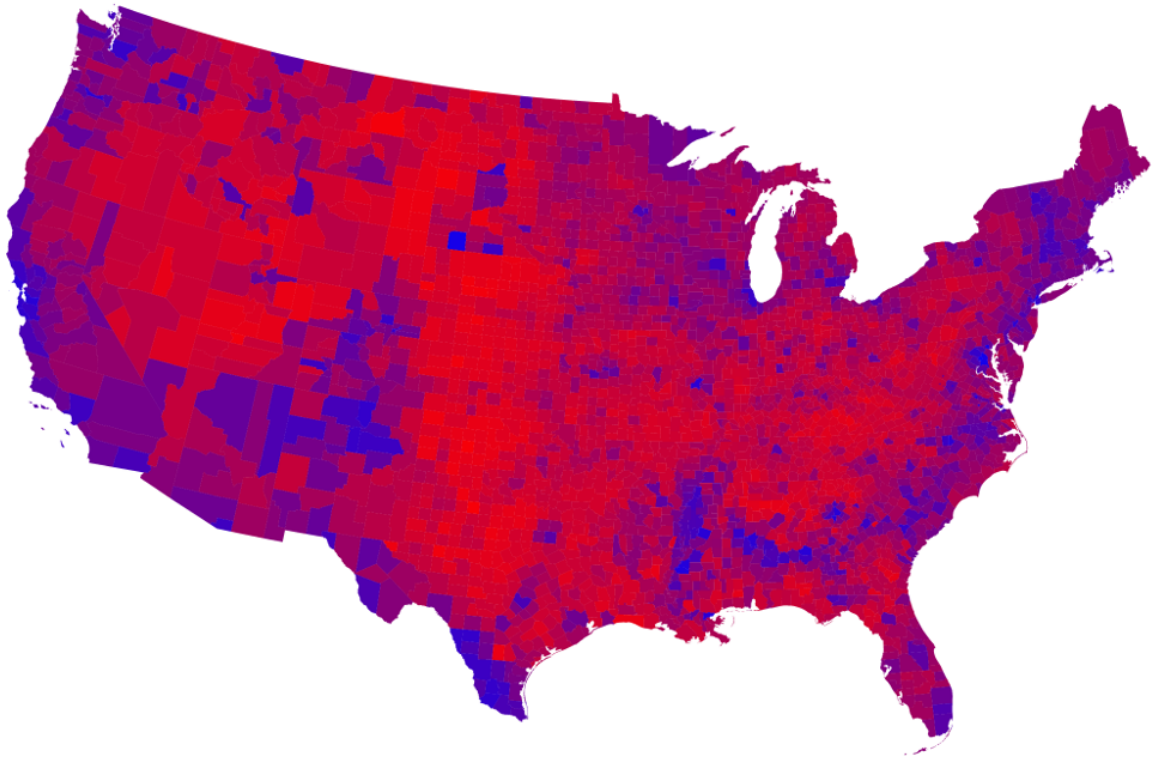
From Mark Newman, U of Michigan

2016 Presidential Election Results by County



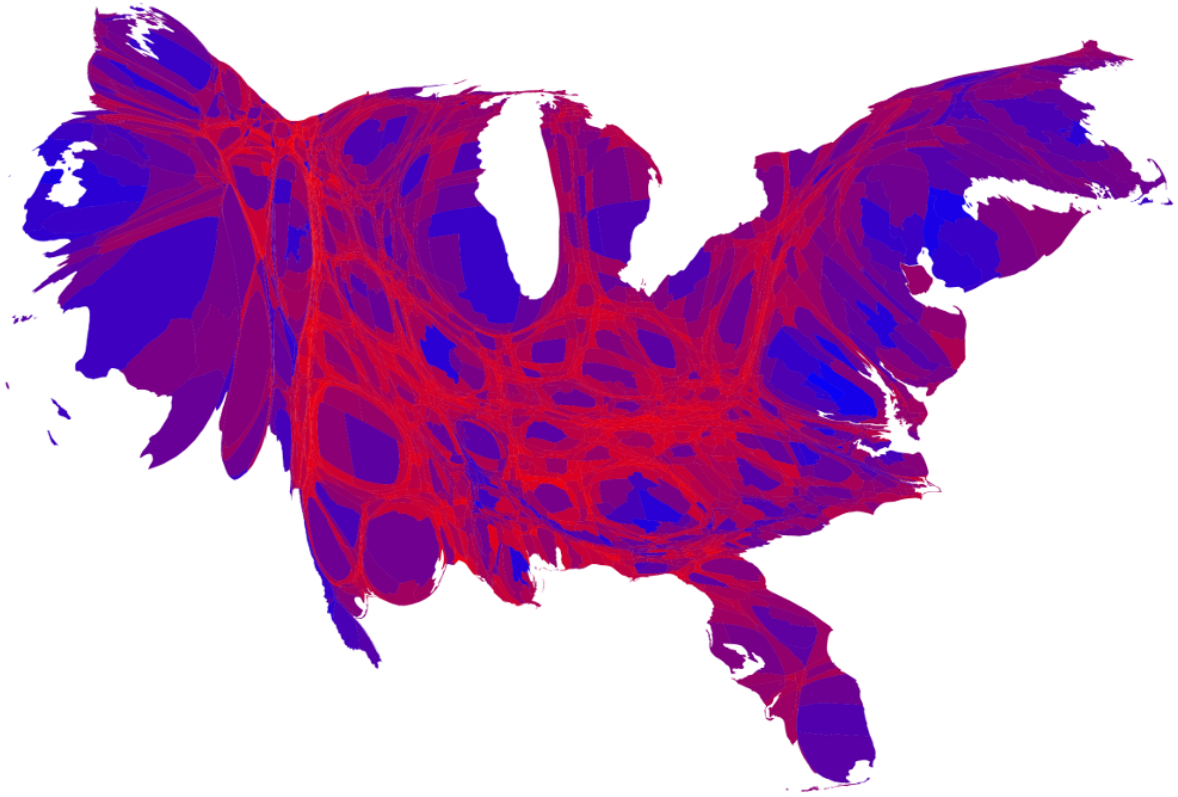
From Mark Newman, U of Michigan

2016 Presidential Election Results by County



From Mark Newman, U of Michigan

2016 Presidential Election Results by County



From Mark Newman, U of Michigan

Types of Data

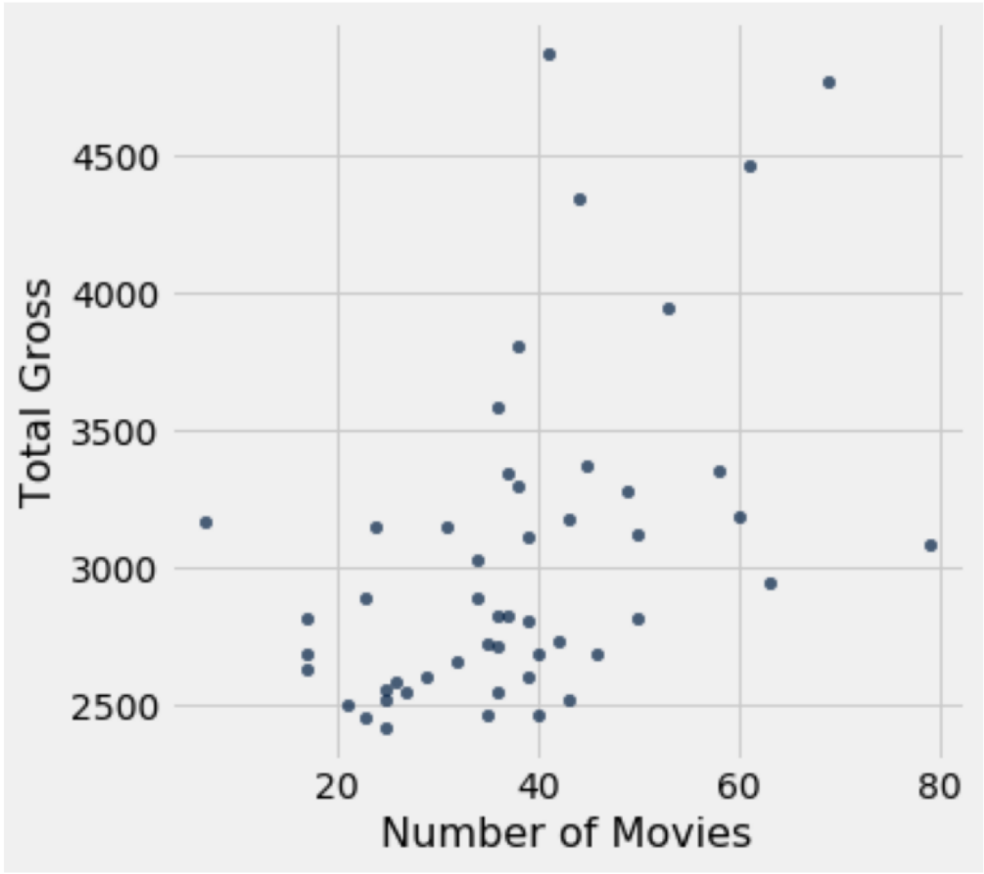
- Tables enforce constraints
  - All values in a column are the same type
  - Values in a column are *comparable*

- **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

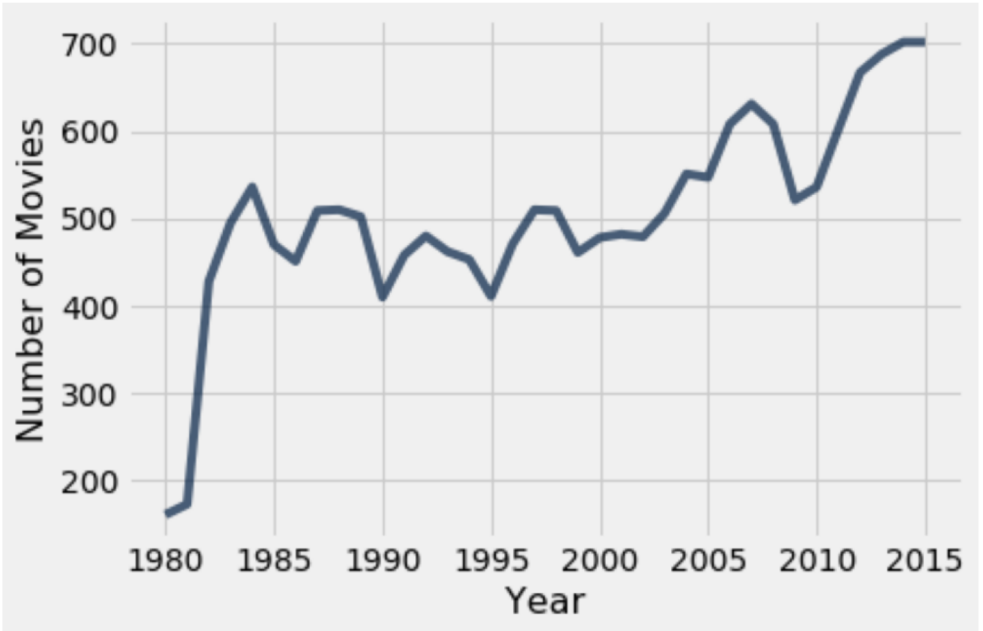
```
actors.scatter('Number of Movies', 'Total Gross')
```



Line Graph

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

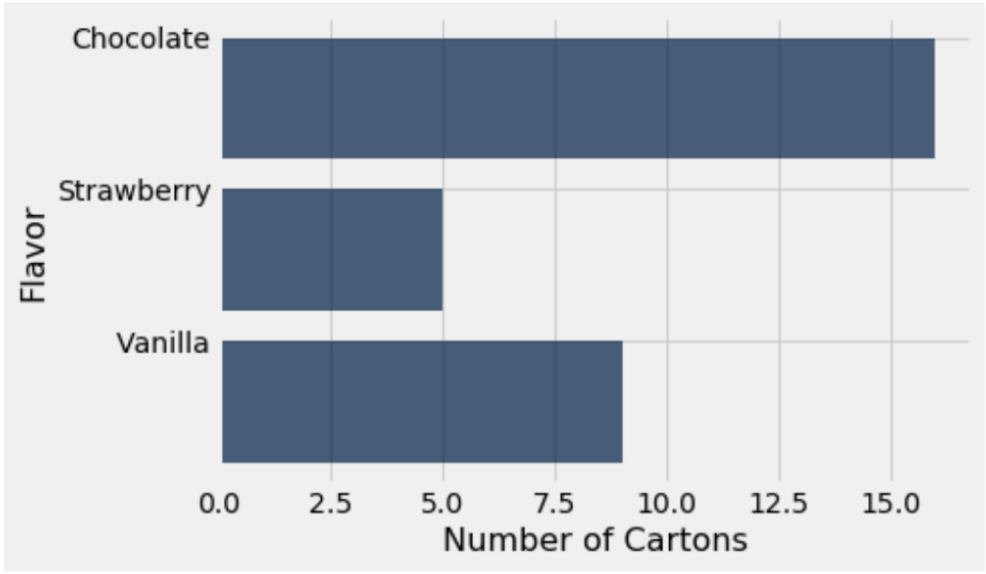
```
movies_by_year.plot('Year', 'Number of Movies')
```



Bar Chart

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

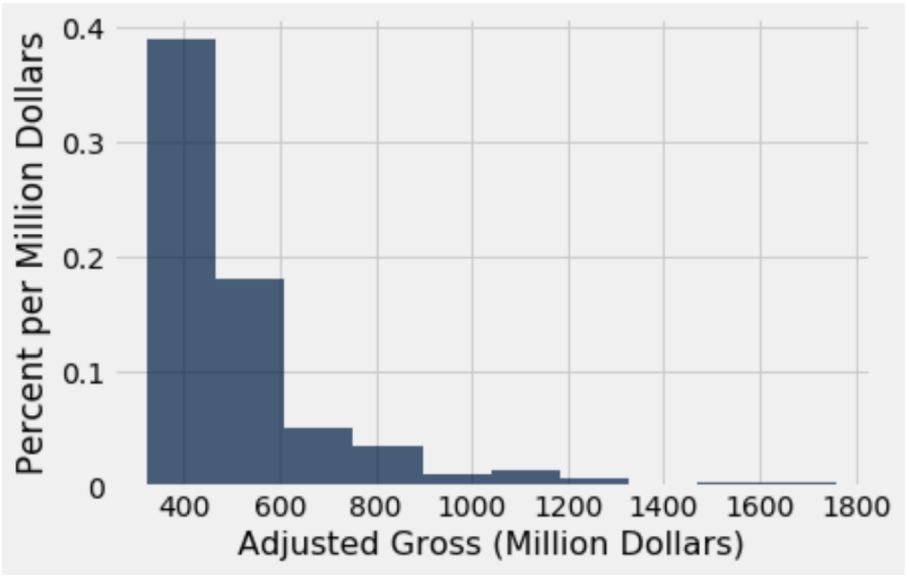
```
icecream.barh('Flavor', 'Number of Cartons')
```



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

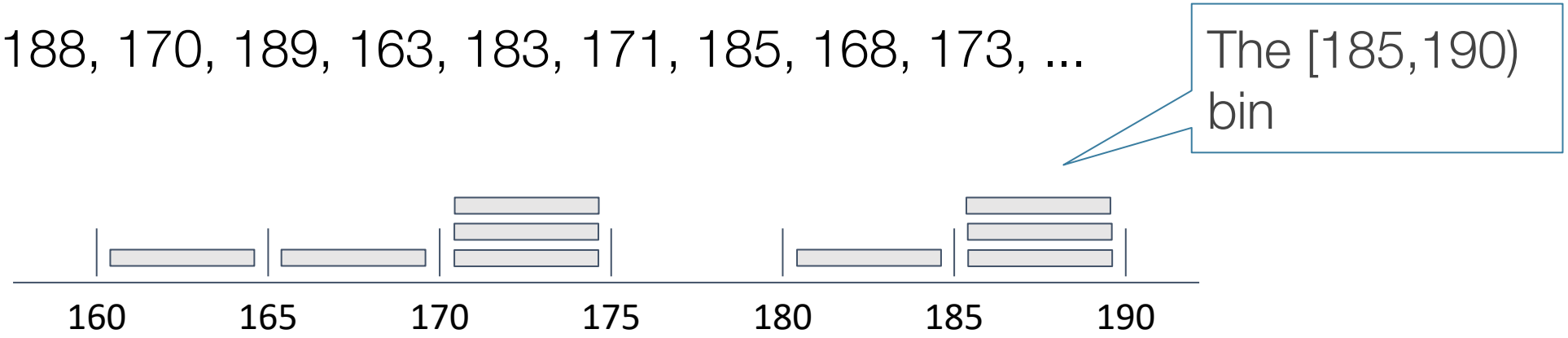
```
millions.hist('Adjusted Gross', unit="Million Dollars")
```



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, ...



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

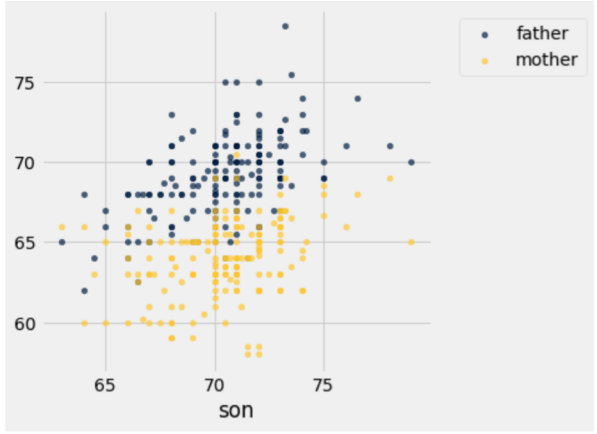
Area = % in bin = Height x width of bin

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

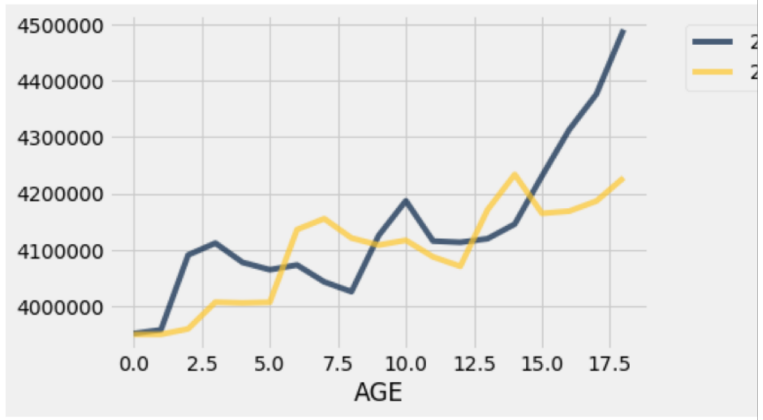
- What would the y-axis of a histogram of this table be?
- http://bit.ly/FoDS-s19-0131-1

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

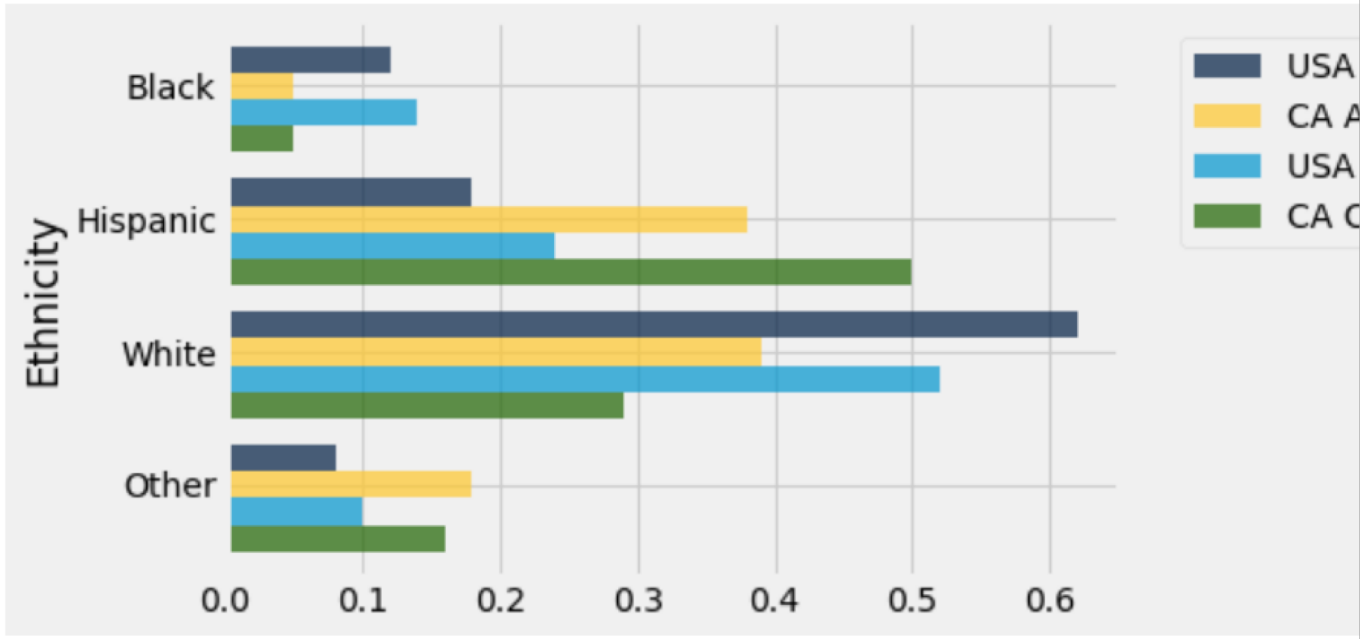
```
heights.scatter('son')
```




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




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



Overlaid Graphs

What's next?

- Read Chapter 8 of *Computational and Inferential Thinking*

- Start working on Homework 2