DATA VISUALIZATION

Plotting pitfalls: the data vis hall of shame

Plot critique

The grammar of graphics

The five most important plots

Enriching plots (color, faceting, labels, etc)

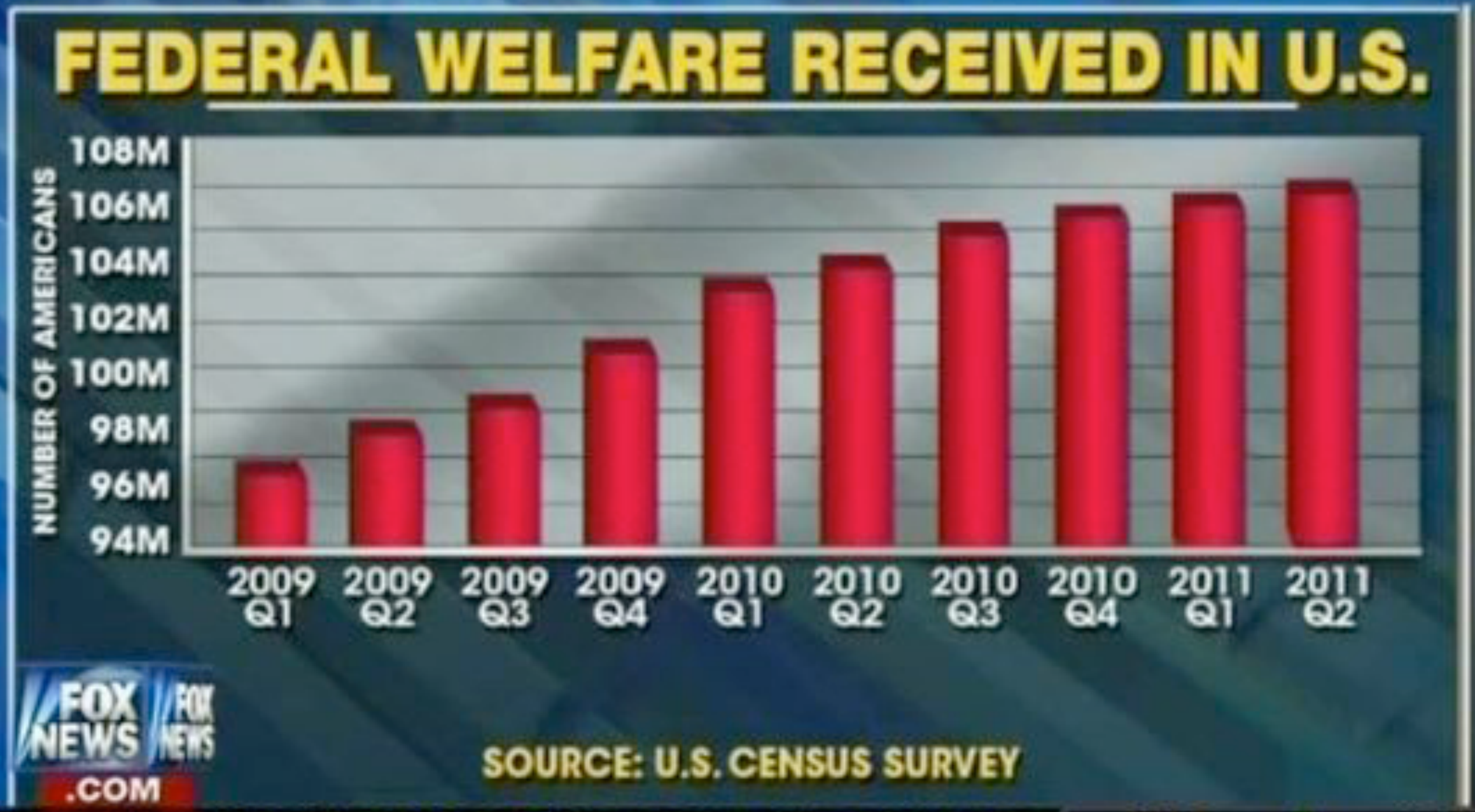
THE DATA VIS HALL OF SHAME

PLOTTING PITFALLS

- Axis trickery
- Violations of basic math
- Nearly content-free figures
- Gratuitous chartjunk
- Poorly chosen 3D graphics
- Bad design choices

PLOTTING PITFALLS

- Axis trickery (a.k.a. “little y lies”)
- Violations of basic math
- Nearly content-free figures
- Gratuitous chartjunk
- Poorly chosen 3D graphics
- Bad design choices

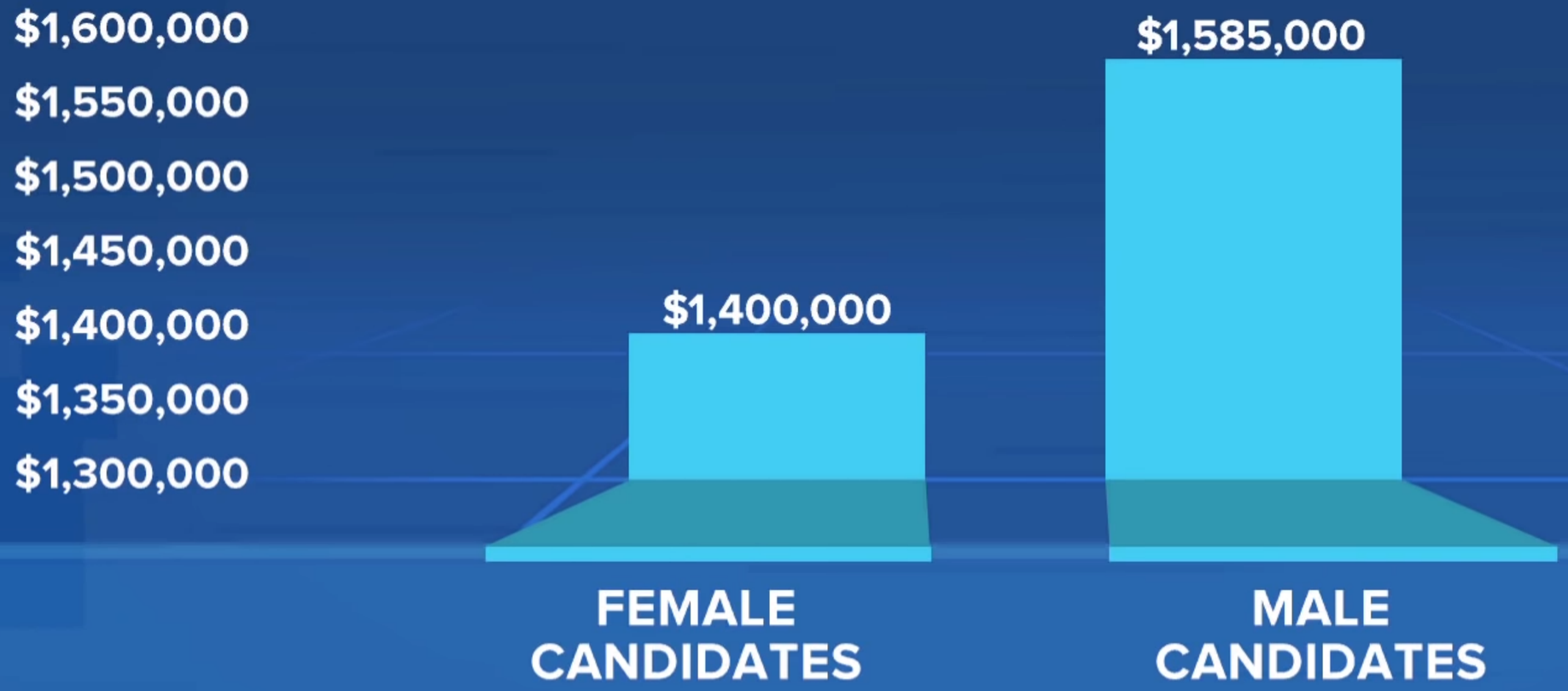


LIST TYLER HAMILTON OF HIS GOLD MEDAL FRO

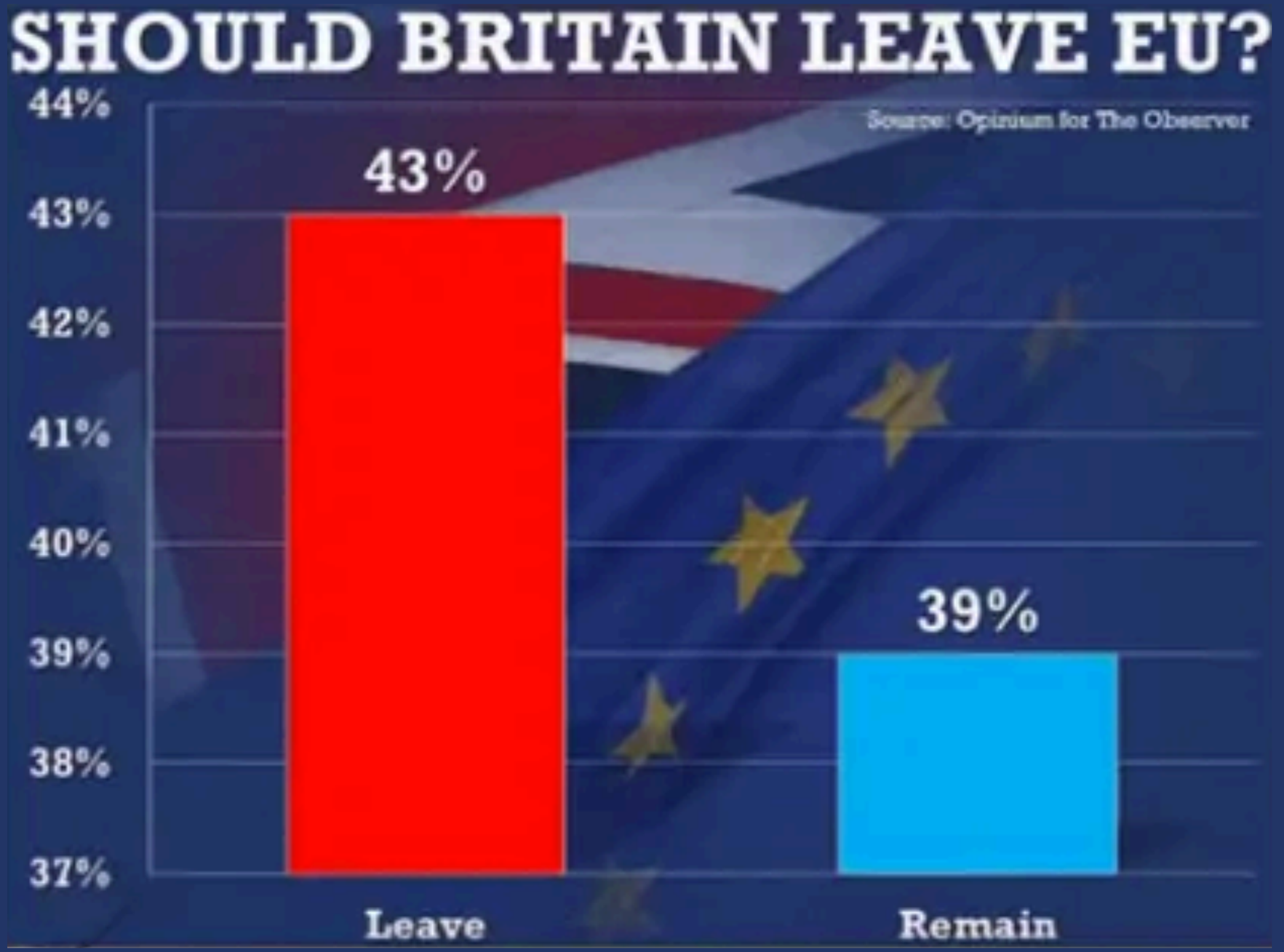
NAS ▲ 6.61

DEMOCRATIC WOMEN RAISE LESS MONEY THAN DEMOCRATIC MEN

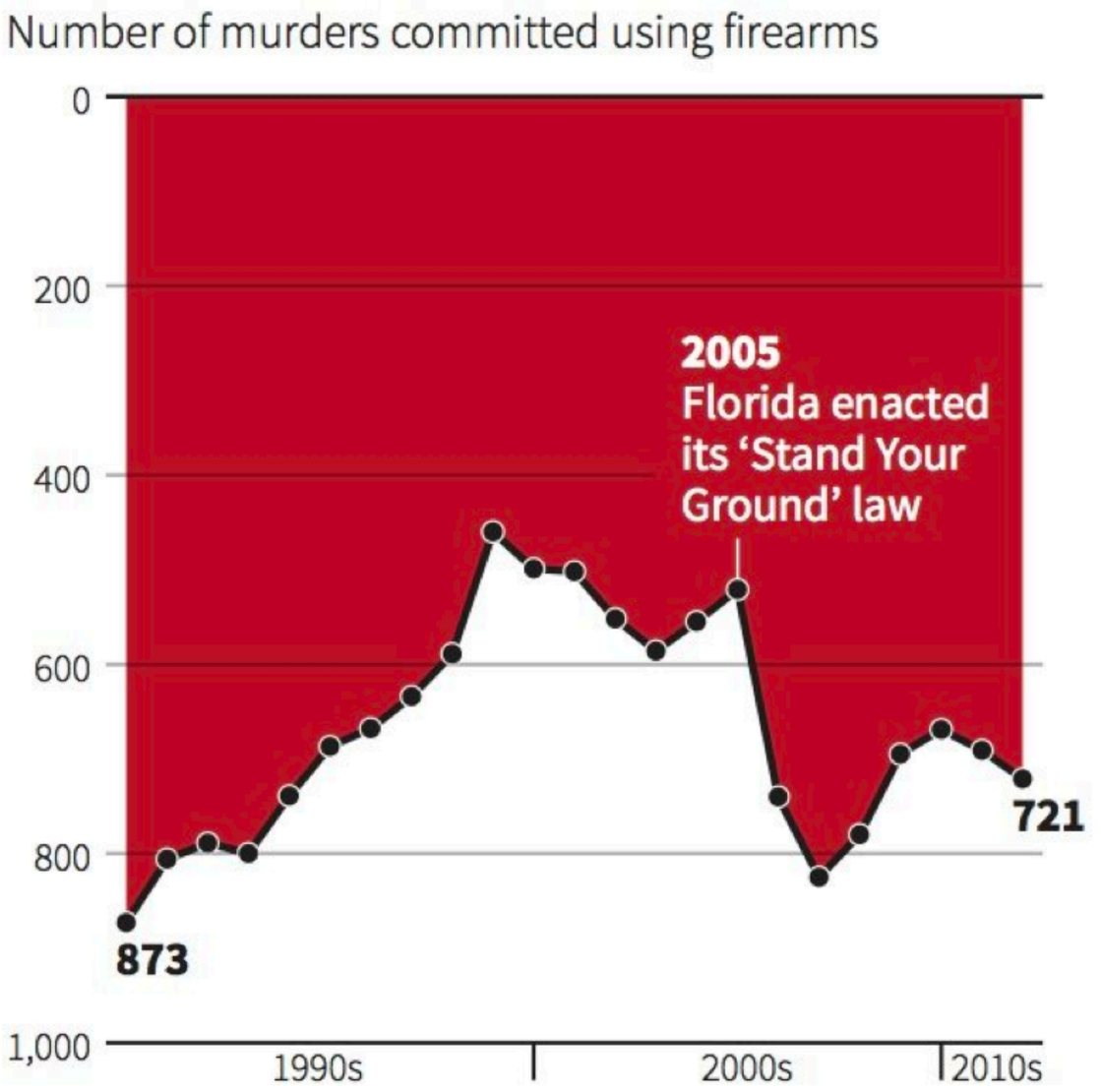
AVERAGE RAISED BY HOUSE PRIMARY WINNERS



SOURCE: CENTER FOR RESPONSIVE POLITICS


SHOULD BRITAIN LEAVE EU?
Source: Opinium for The Observer
44%
43%
42%
41%
40%
39%
38%
37%
43%
39%
Leave
Remain

Gun deaths in Florida



Source: Florida Department of Law Enforcement

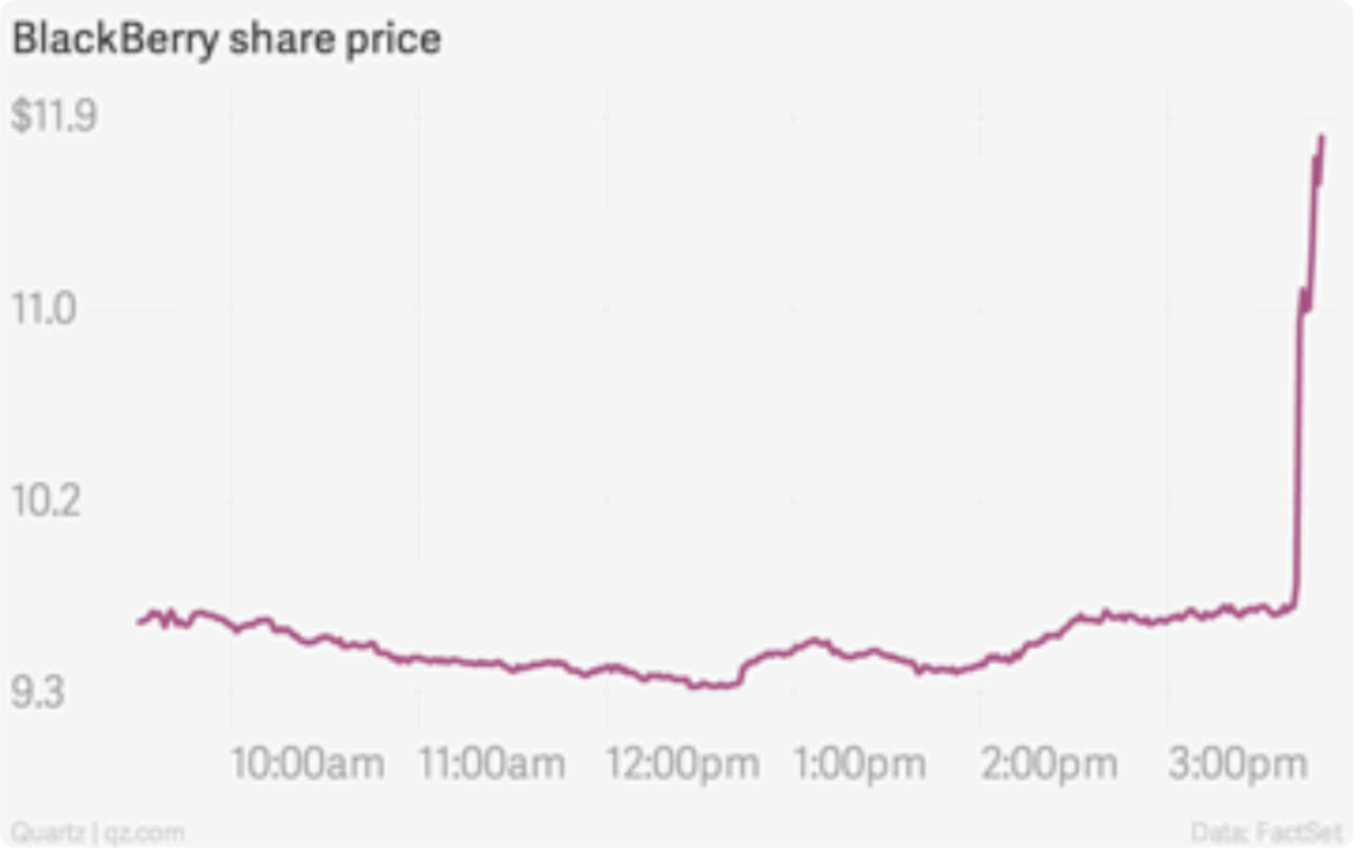
C. Chan 16/02/2014

REUTERS

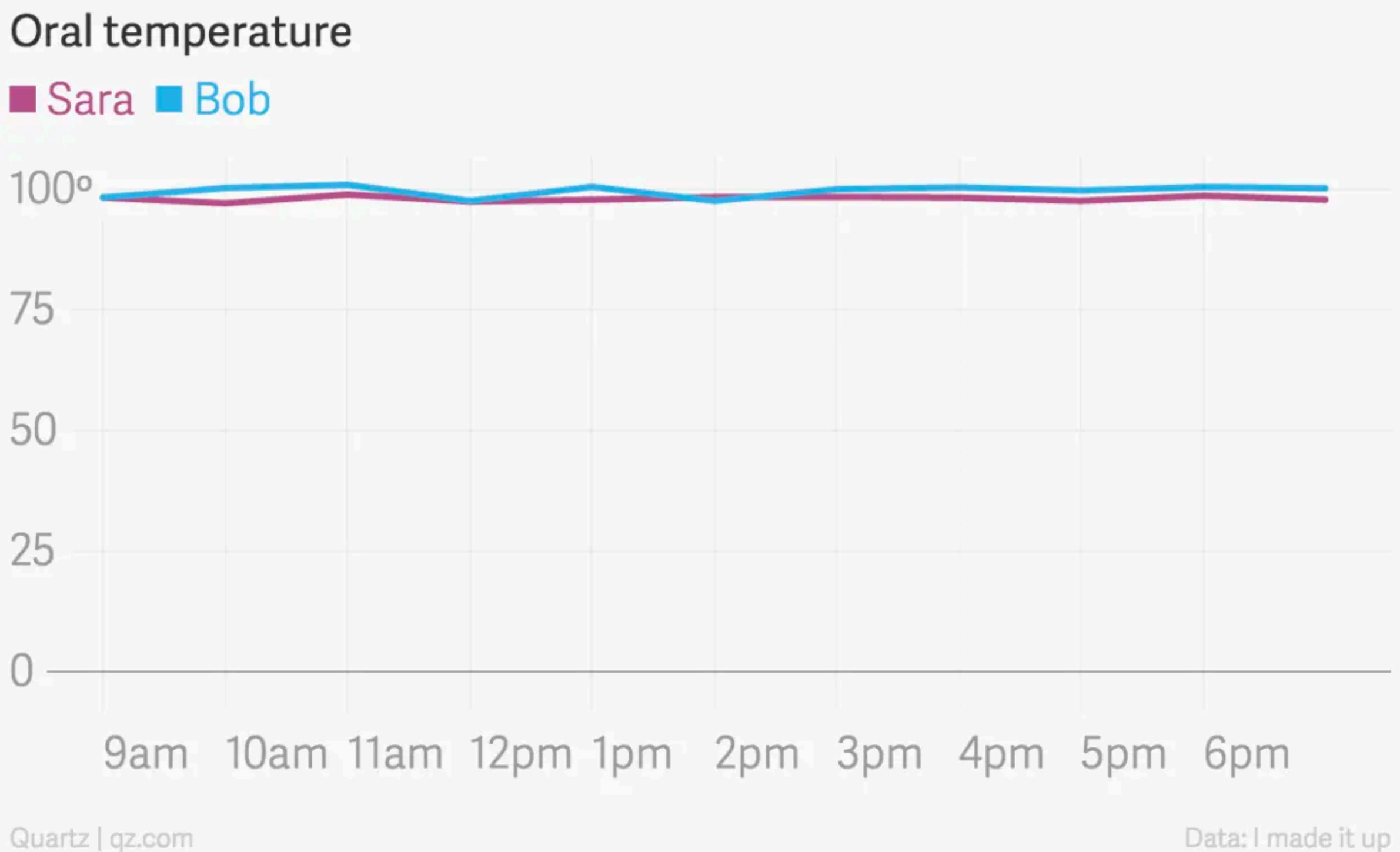
AND YET...

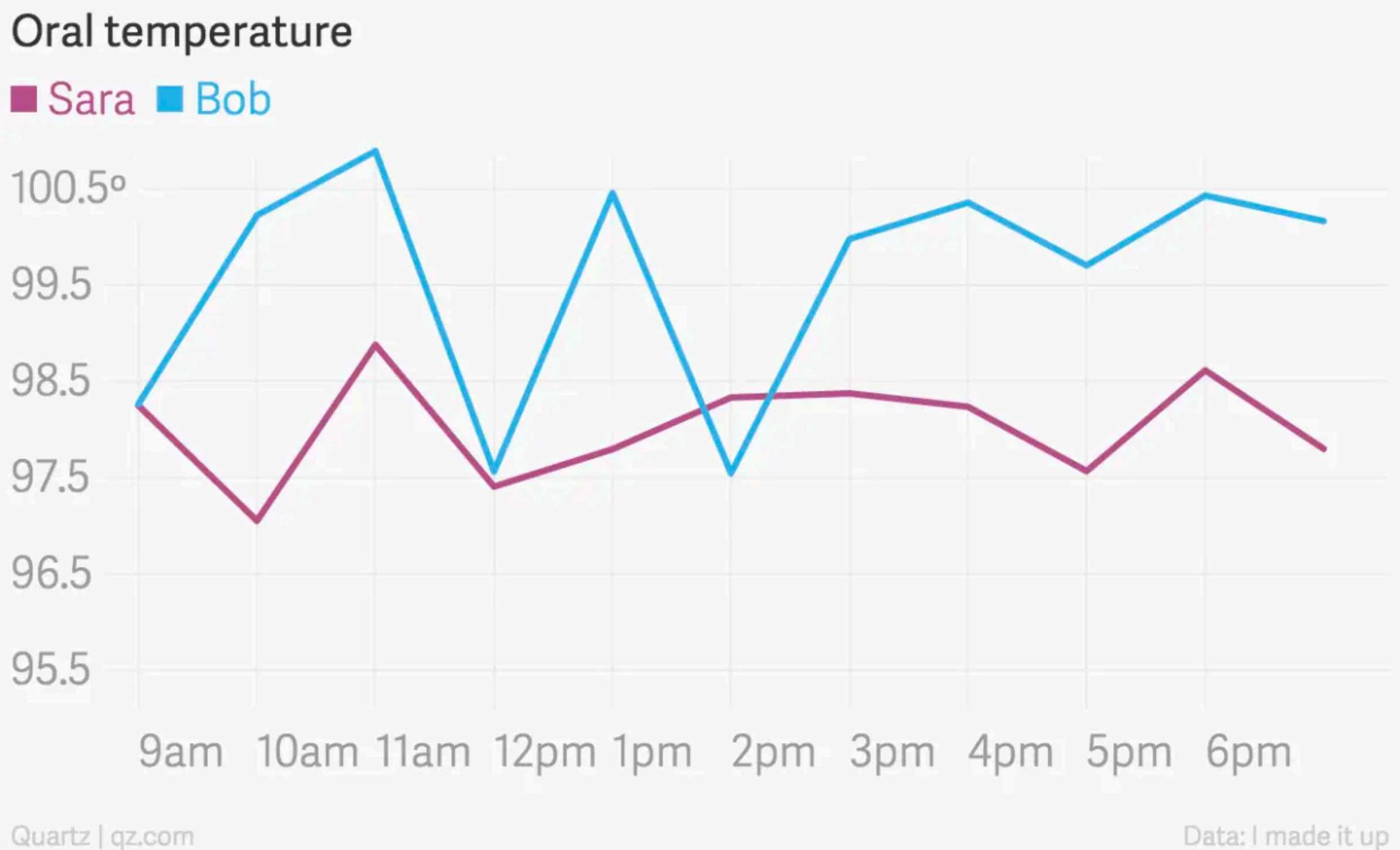
Quartz
@qz

Acquisition rumors light a fire under BlackBerry's stock.
qz.com/note/326942/



♡ 6 4:44 PM - Jan 14, 2015


Oral temperature
Sara
Bob
100°
75
50
25
0
9am
10am
11am
12pm
1pm
2pm
3pm
4pm
5pm
6pm
Quartz | qz.com
Data: I made it up


Oral temperature
Sara
Bob
100.5°
99.5
98.5
97.5
96.5
95.5
9am
10am
11am
12pm
1pm
2pm
3pm
4pm
5pm
6pm
Quartz | qz.com
Data: I made it up

TRUNCATING THE Y AXIS IS SOMETIMES OK

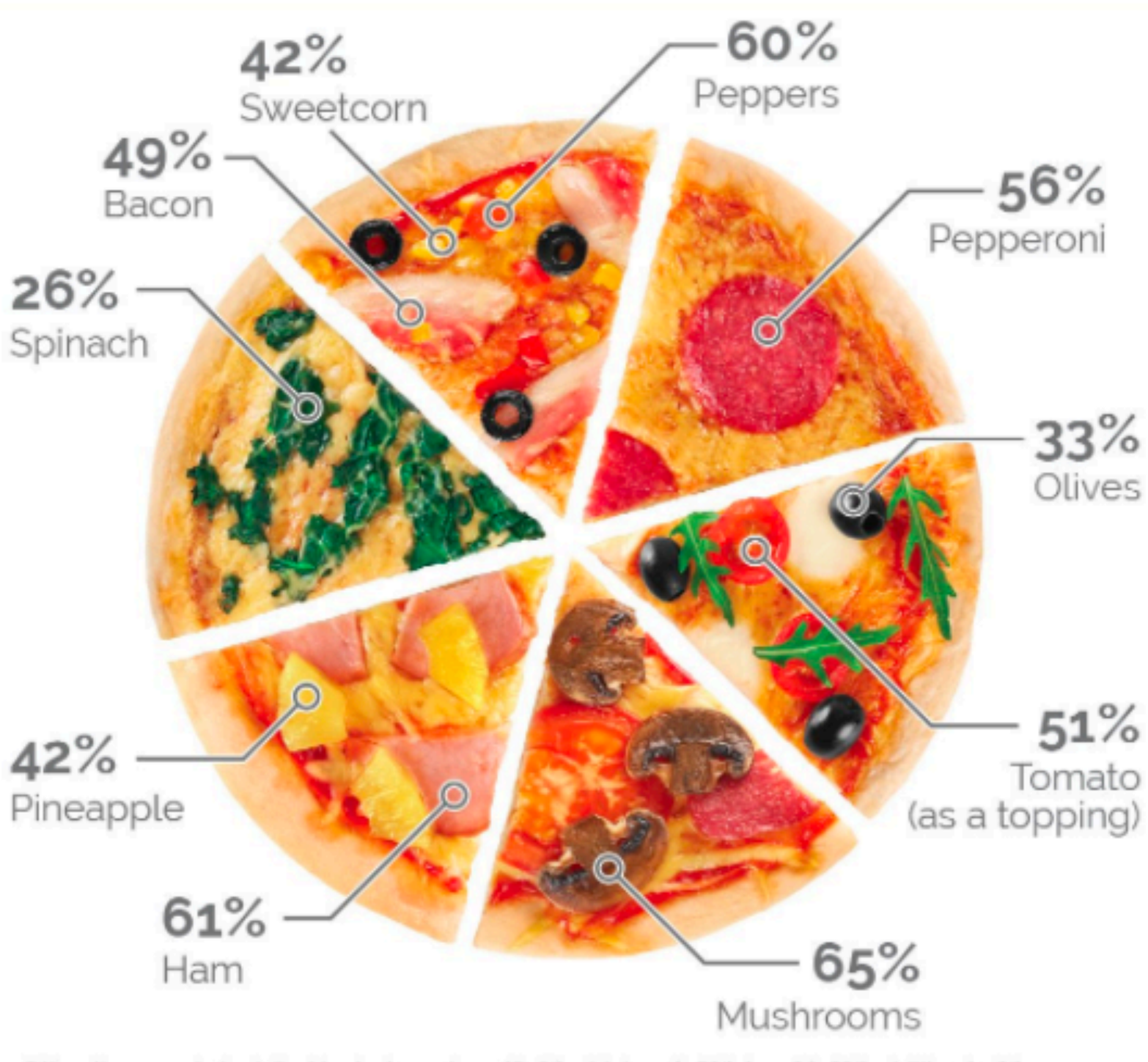
- When you're trying to emphasize change, rather than relative magnitude.
- When you're plotting data over time.
- When zero is not a sensible baseline for comparison.

Bottom line: use your judgment; don't mislead people; watch out for "little y lies."

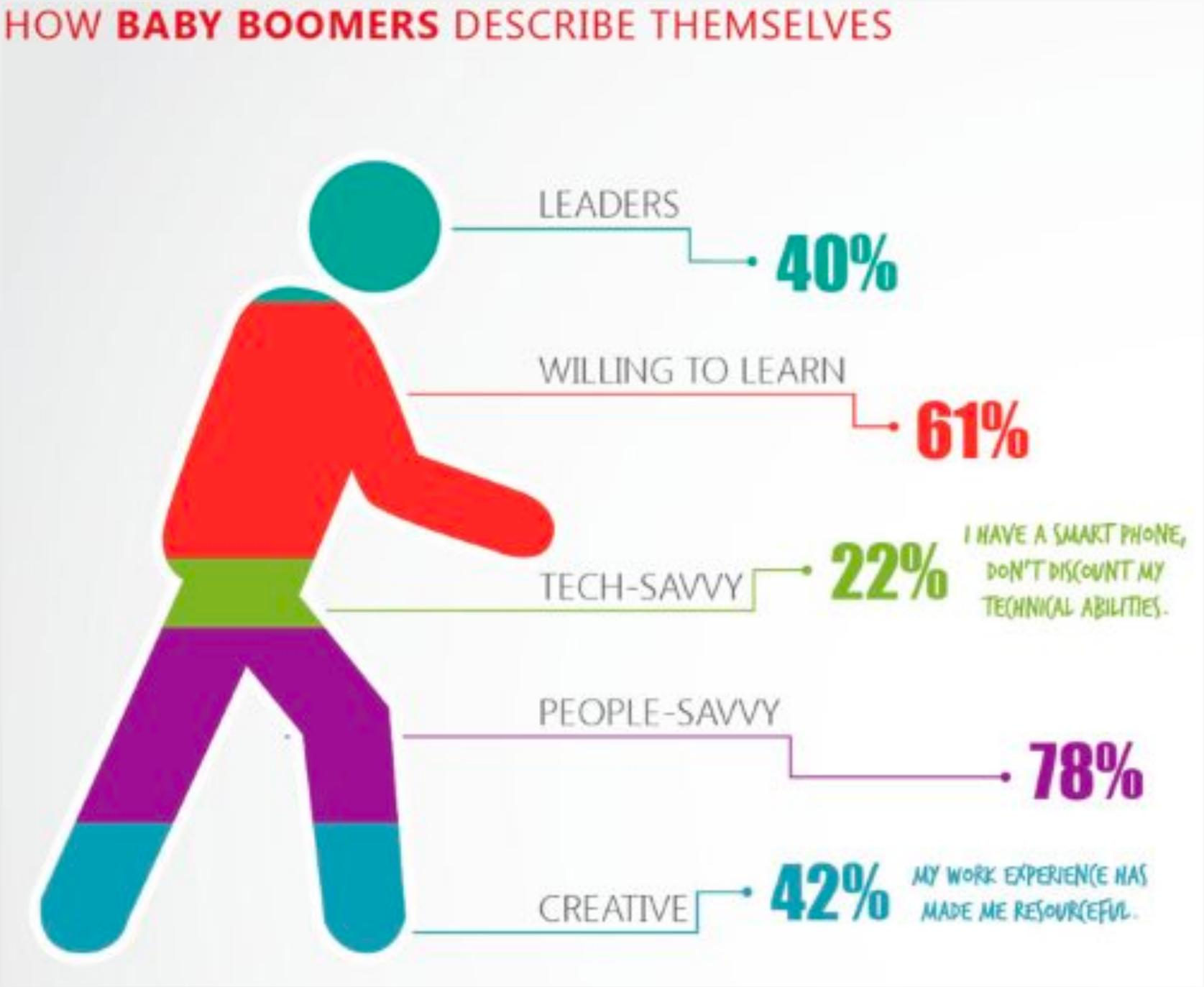
PLOTTING PITFALLS

- Axis trickery
- Violations of basic math
- Nearly content-free figures
- Gratuitous chartjunk
- Poorly chosen 3D graphics
- Bad design choices

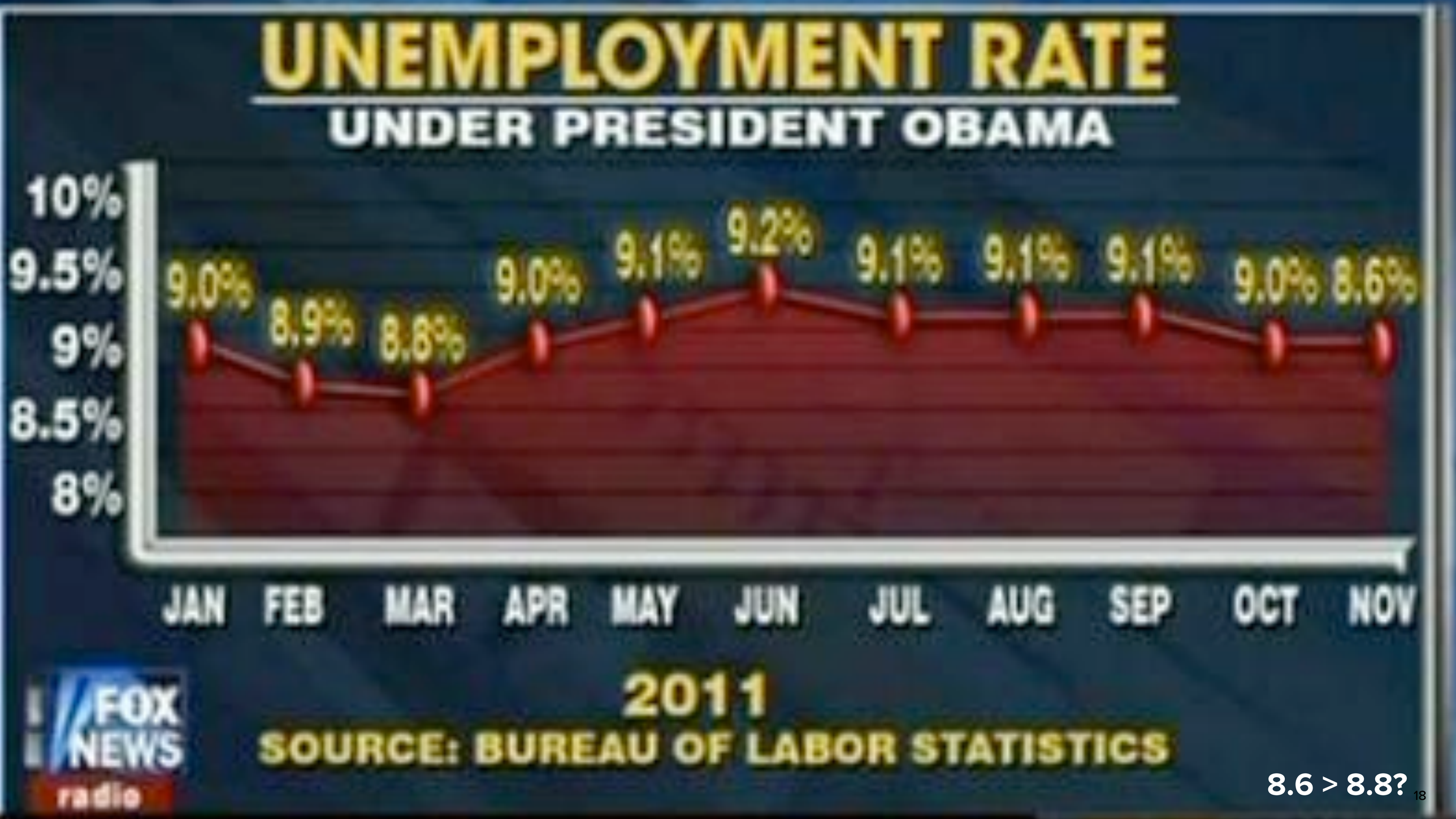

RASMUSSEN
REPORTS POLL
Did scientists falsify research to support their own theories on Global Warming?
59% SOMEWHAT LIKELY
35% VERY LIKELY
26% NOT VERY LIKELY
FOX NEWS .COM
ATE CHANGE RESEARCH
FOX NEWS
GOP
NHL TOR 6 COB 3


42%
Sweetcorn
60%
Peppers
49%
Bacon
56%
Pepperoni
26%
Spinach
33%
Olives
51%
Tomato
(as a topping)
42%
Pineapple
61%
Ham
65%
Mushrooms

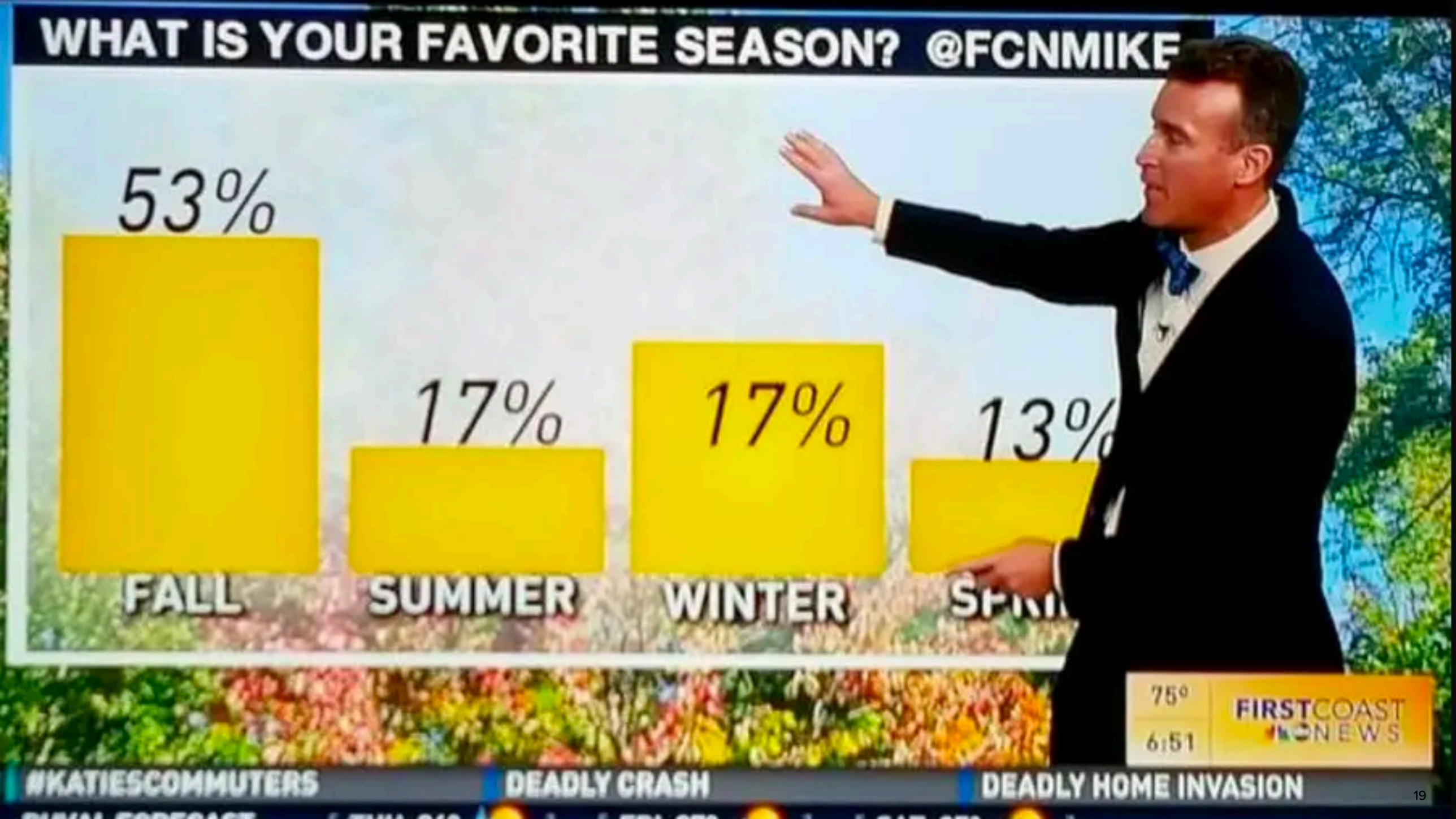
HOW **BABY BOOMERS** DESCRIBE THEMSELVES



243% Boomer


UNEMPLOYMENT RATE
UNDER PRESIDENT OBAMA
10%
9.5%
9%
8.5%
8%
9.0%
8.9%
8.8%
9.0%
9.1%
9.2%
9.1%
9.1%
9.1%
9.0%
8.6%
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
2011
SOURCE: BUREAU OF LABOR STATISTICS
FOX NEWS radio

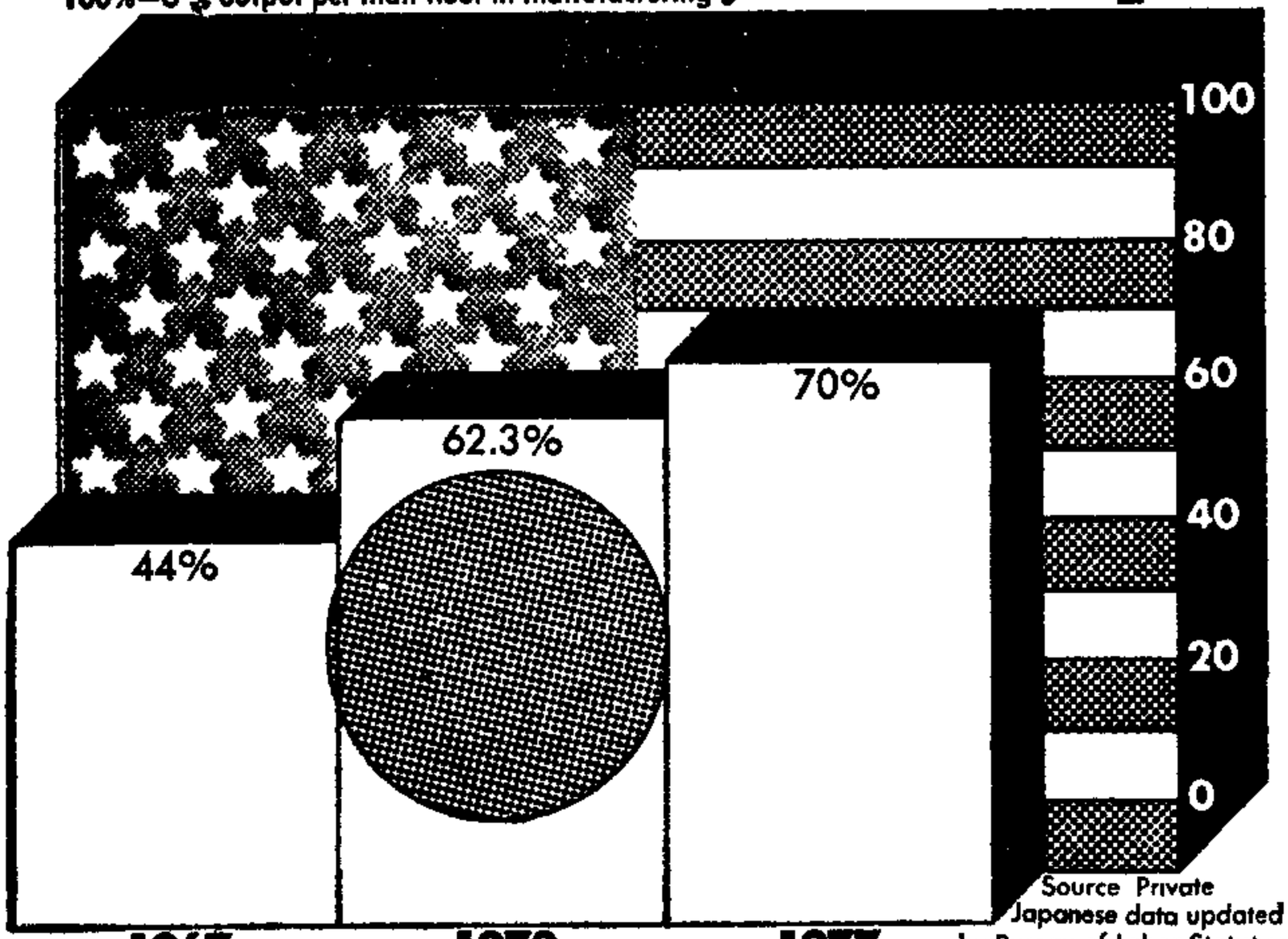
8.6 > 8.8?


WHAT IS YOUR FAVORITE SEASON? @FCNMIKE
53%
17%
17%
13%
FALL
SUMMER
WINTER
75°
6:51
FIRSTCOAST NEWS
#KATIESCOMMUTERS
DEADLY CRASH
DEADLY HOME INVASION

PLOTTING PITFALLS

- Axis trickery
- Violations of basic math
- Nearly content-free figures
- Gratuitous chartjunk
- Poorly chosen 3D graphics
- Bad design choices

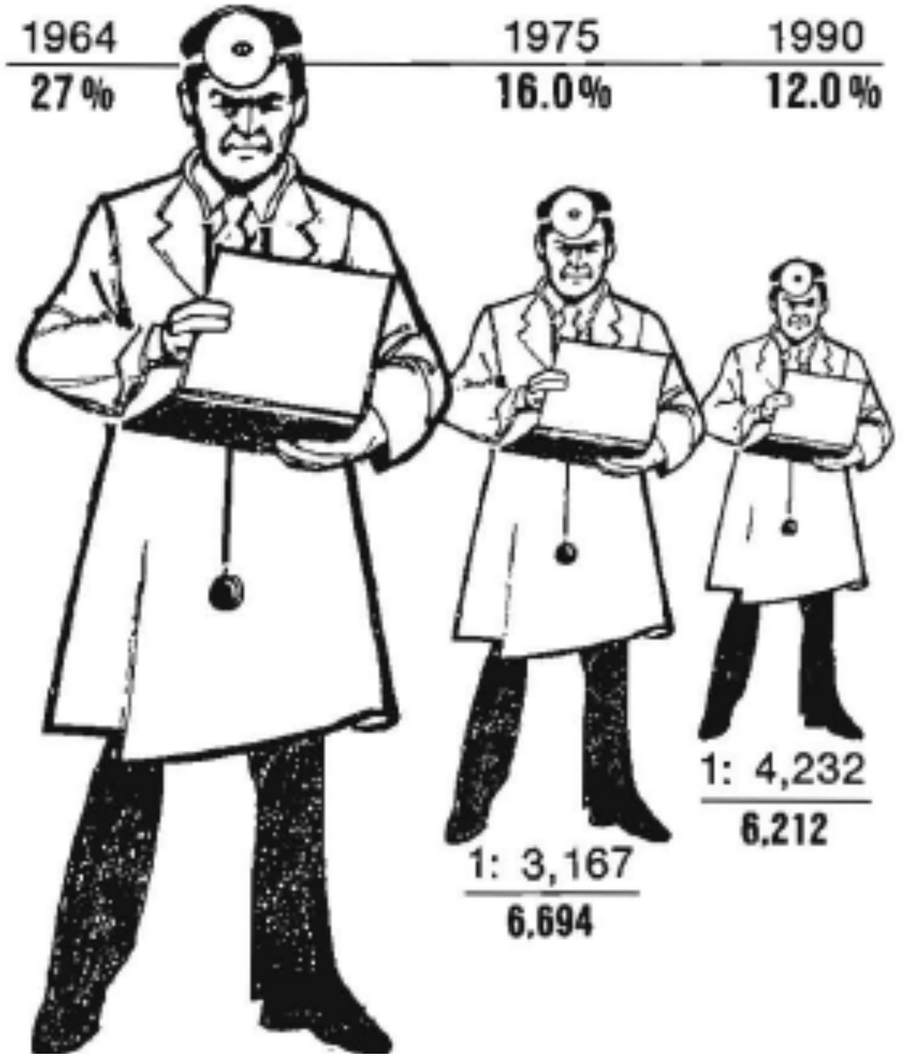
Labor Productivity: U.S. vs Japan

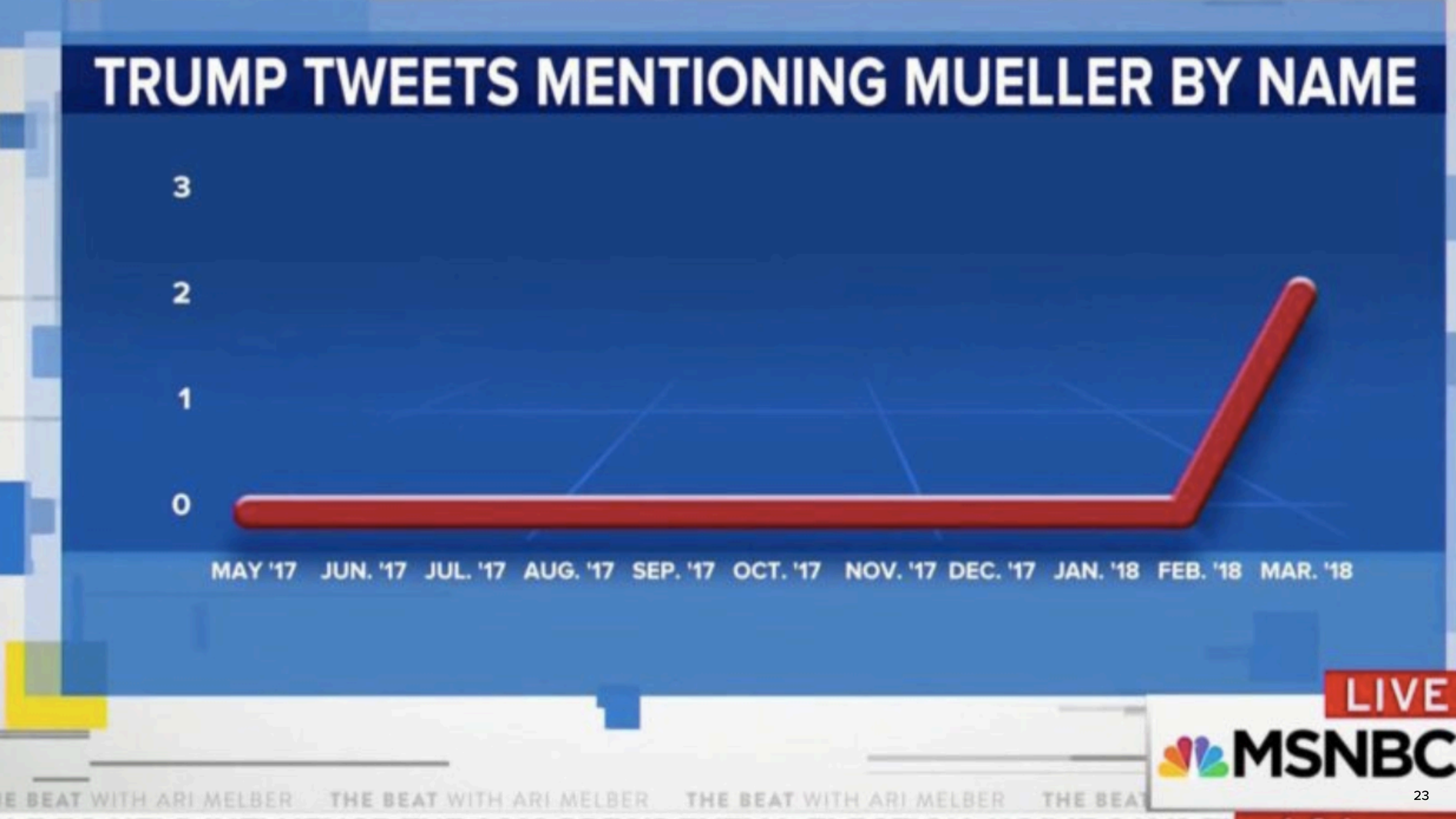


THE SHRINKING FAMILY DOCTOR

In California

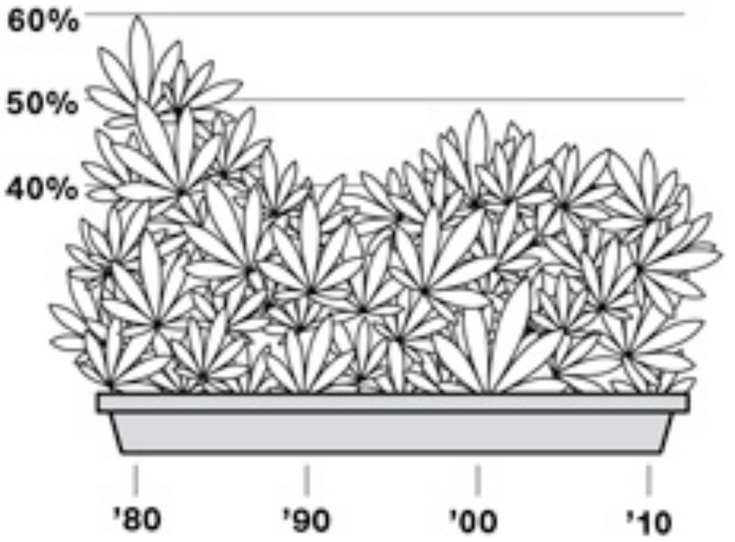
Percentage of Doctors Devoted Solely to Family Practice



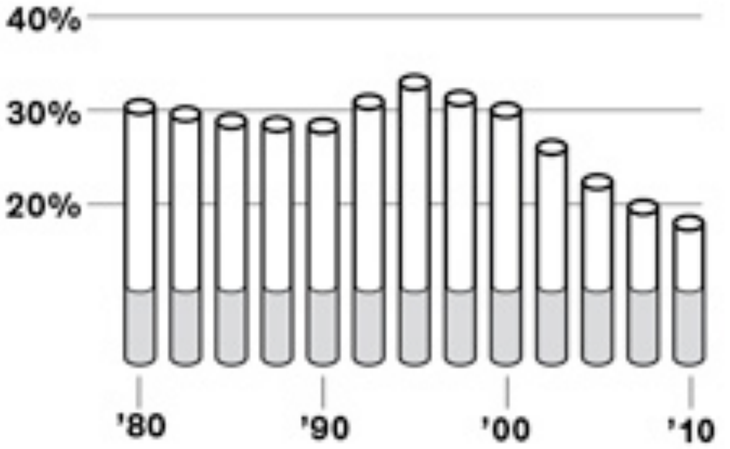

TRUMP TWEETS MENTIONING MUELLER BY NAME
3
2
1
0
MAY '17
JUN. '17
JUL. '17
AUG. '17
SEP. '17
OCT. '17
NOV. '17
DEC. '17
JAN. '18
FEB. '18
MAR. '18
LIVE
MSNBC
THE BEAT WITH ARI MELBER
THE BEAT WITH ARI MELBER
THE BEAT WITH ARI MELBER

PLOTTING PITFALLS

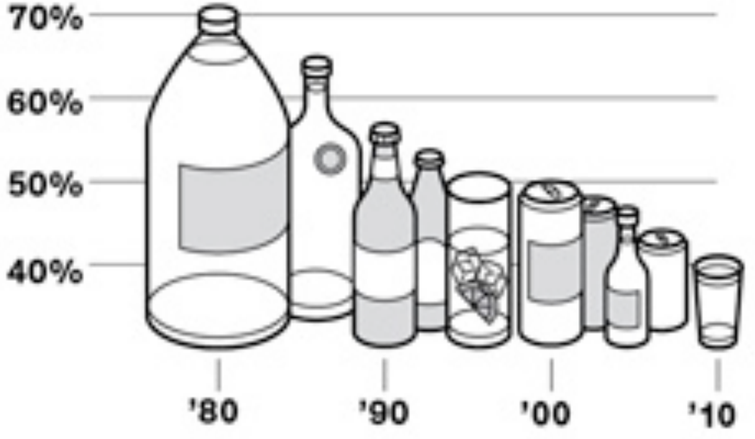
- Axis trickery
- Violations of basic math
- Nearly content-free figures
- Gratuitous chartjunk
- Poorly chosen 3D graphics
- Bad design choices



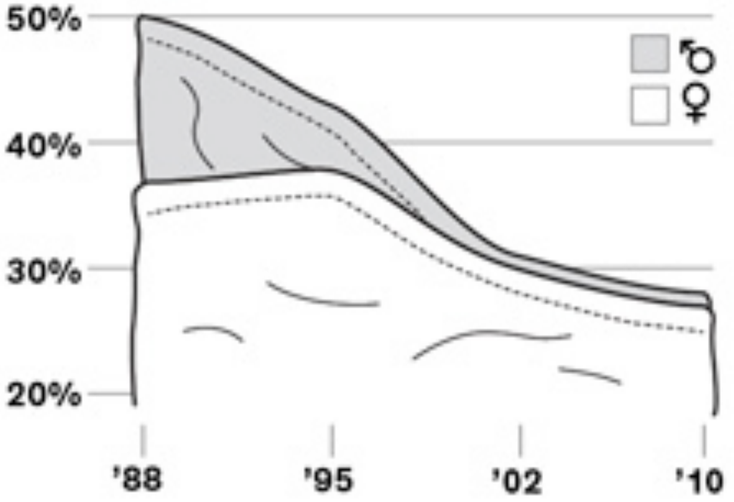
△ Percentage of high-school seniors who have ever tried pot.



□ Smoking habits of high-school seniors over time.



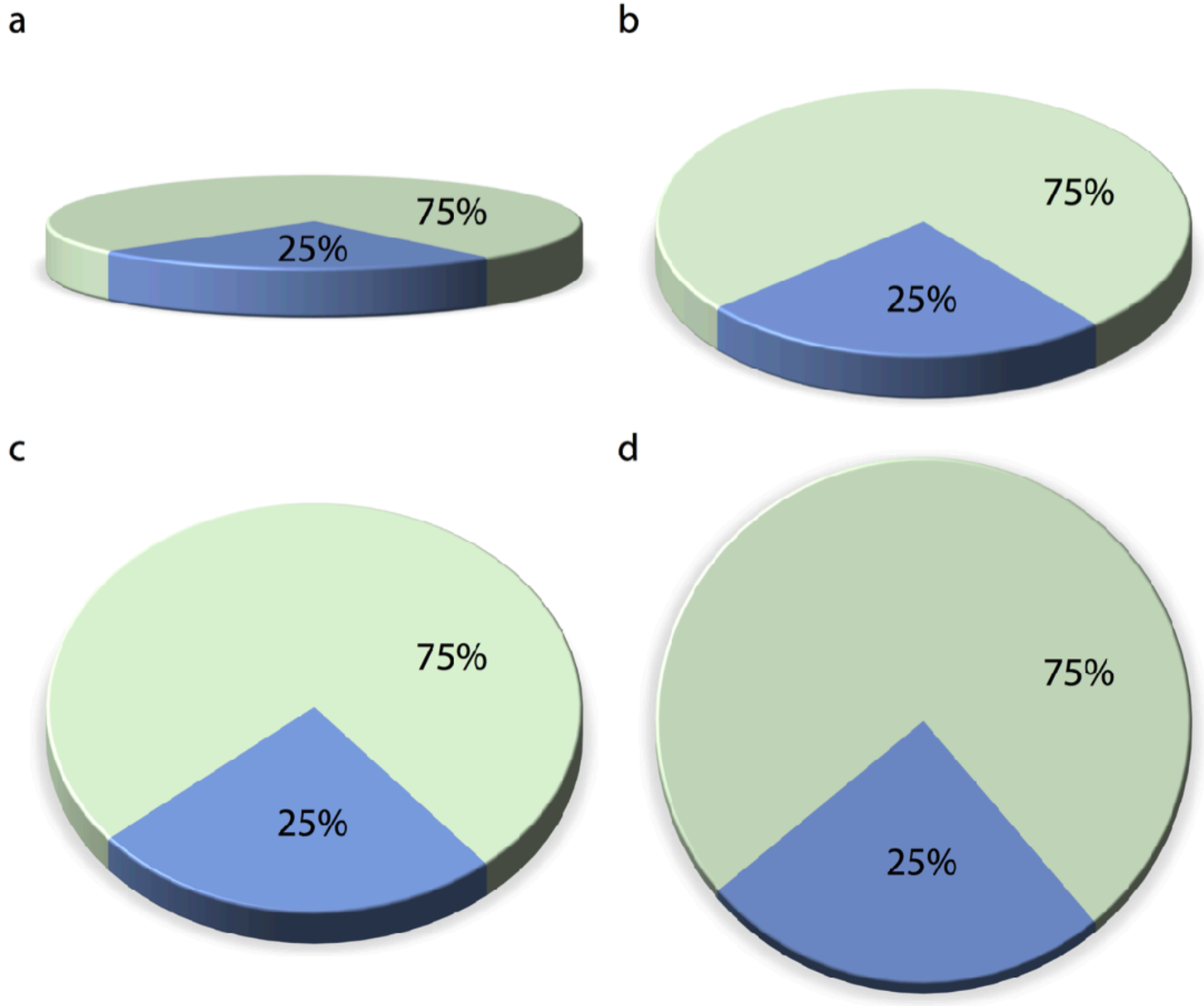
○ Alcohol use by high-school seniors over time.



/ Percentage of 15-to-17-year-olds who have had sex.

PLOTTING PITFALLS

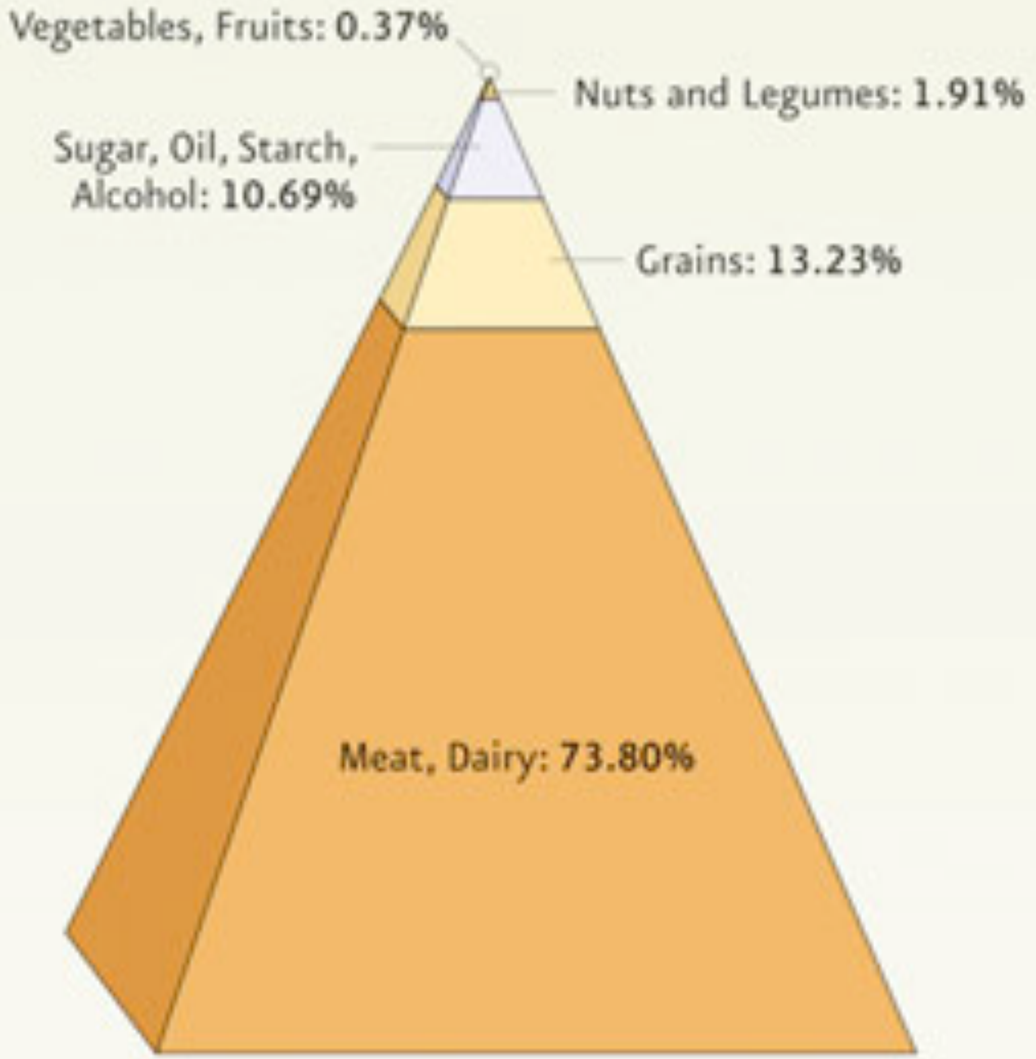
- Axis trickery
- Violations of basic math
- Nearly content-free figures
- Gratuitous chartjunk
- Poorly chosen 3D graphics
- Bad design choices



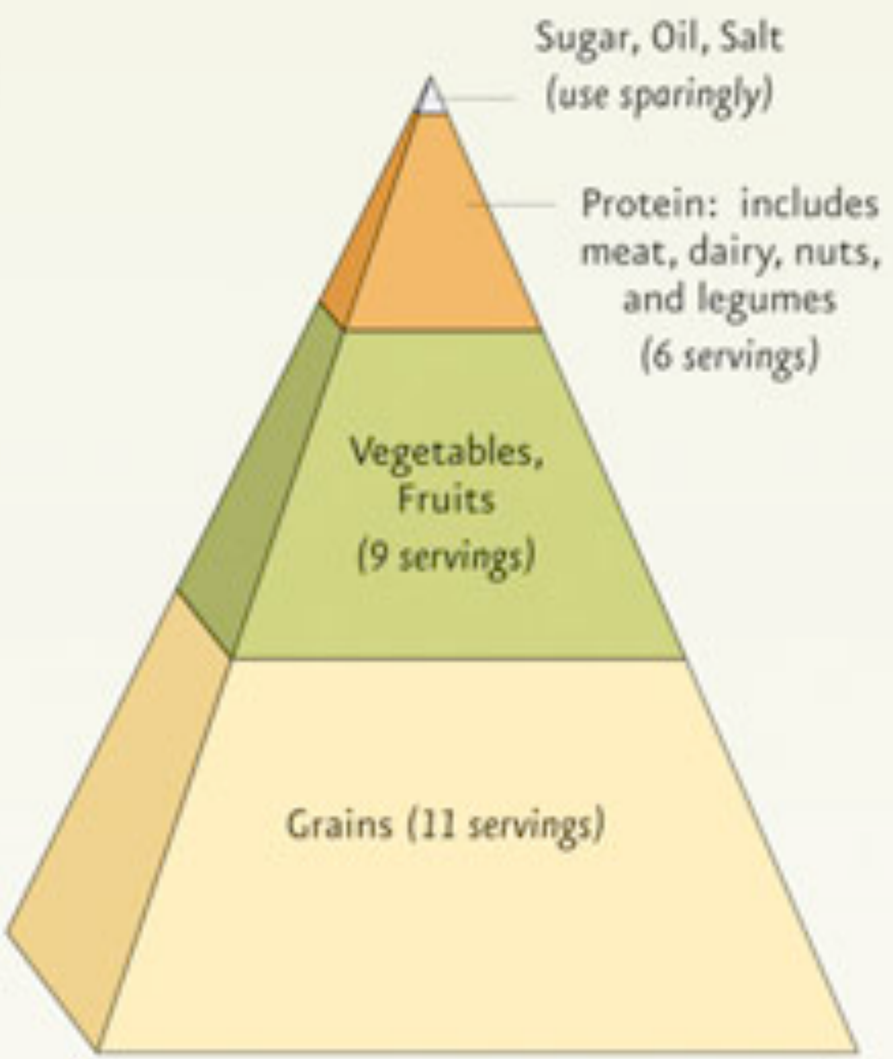
Same chart, four angles

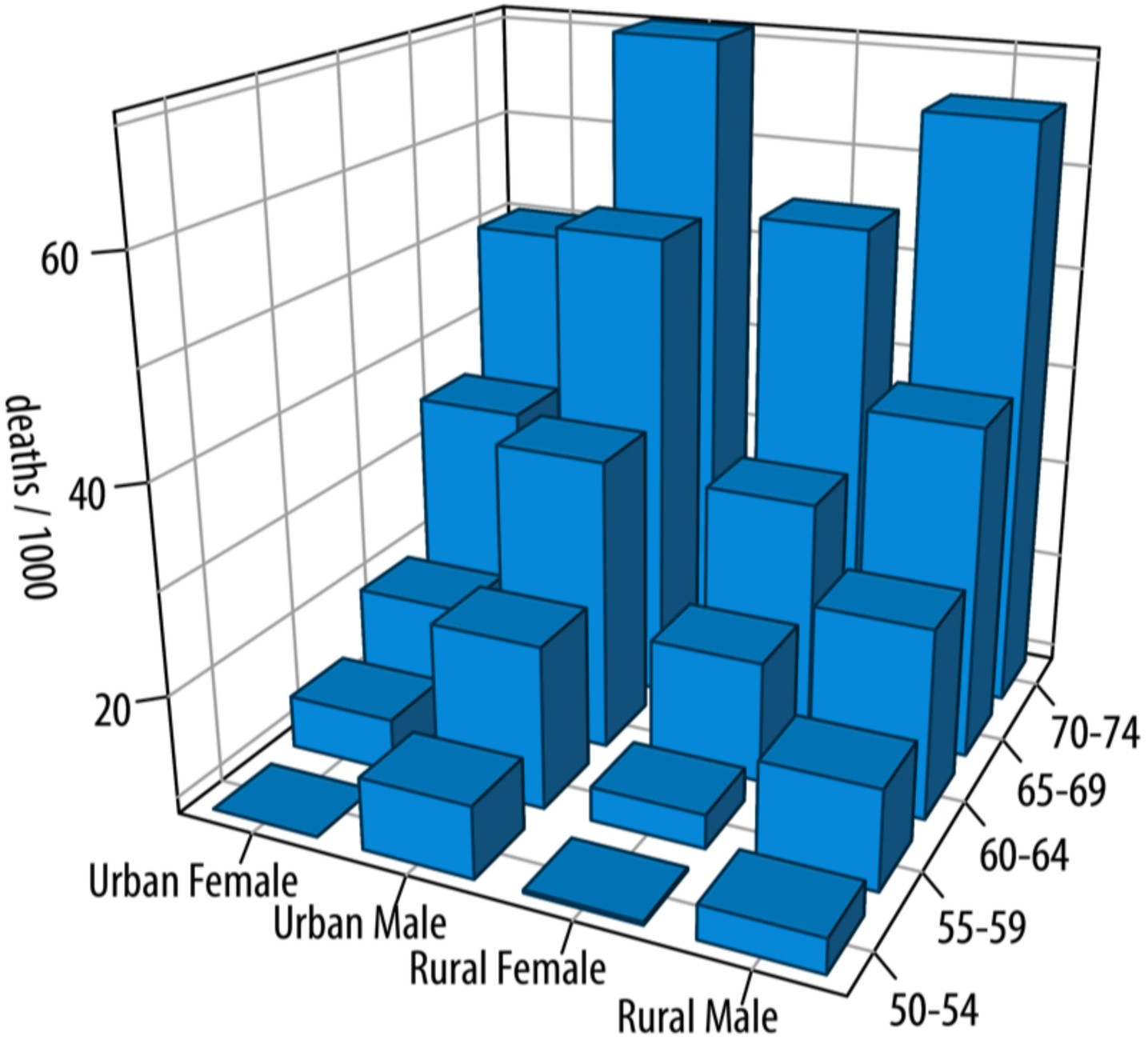
Why Does a Salad Cost More Than a Big Mac?

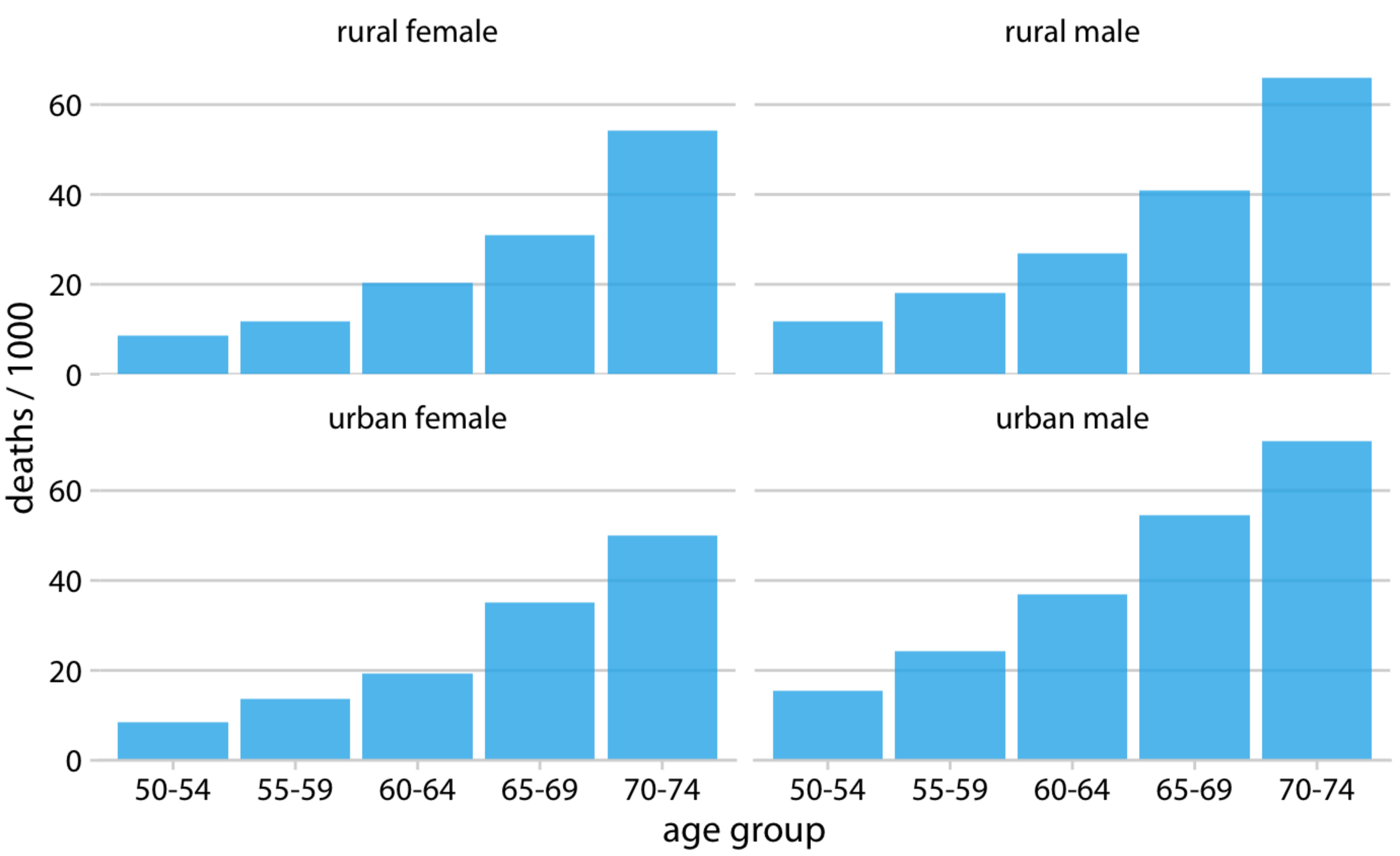
Federal Subsidies for Food Production, 1995-2005*



Federal Nutrition Recommendations

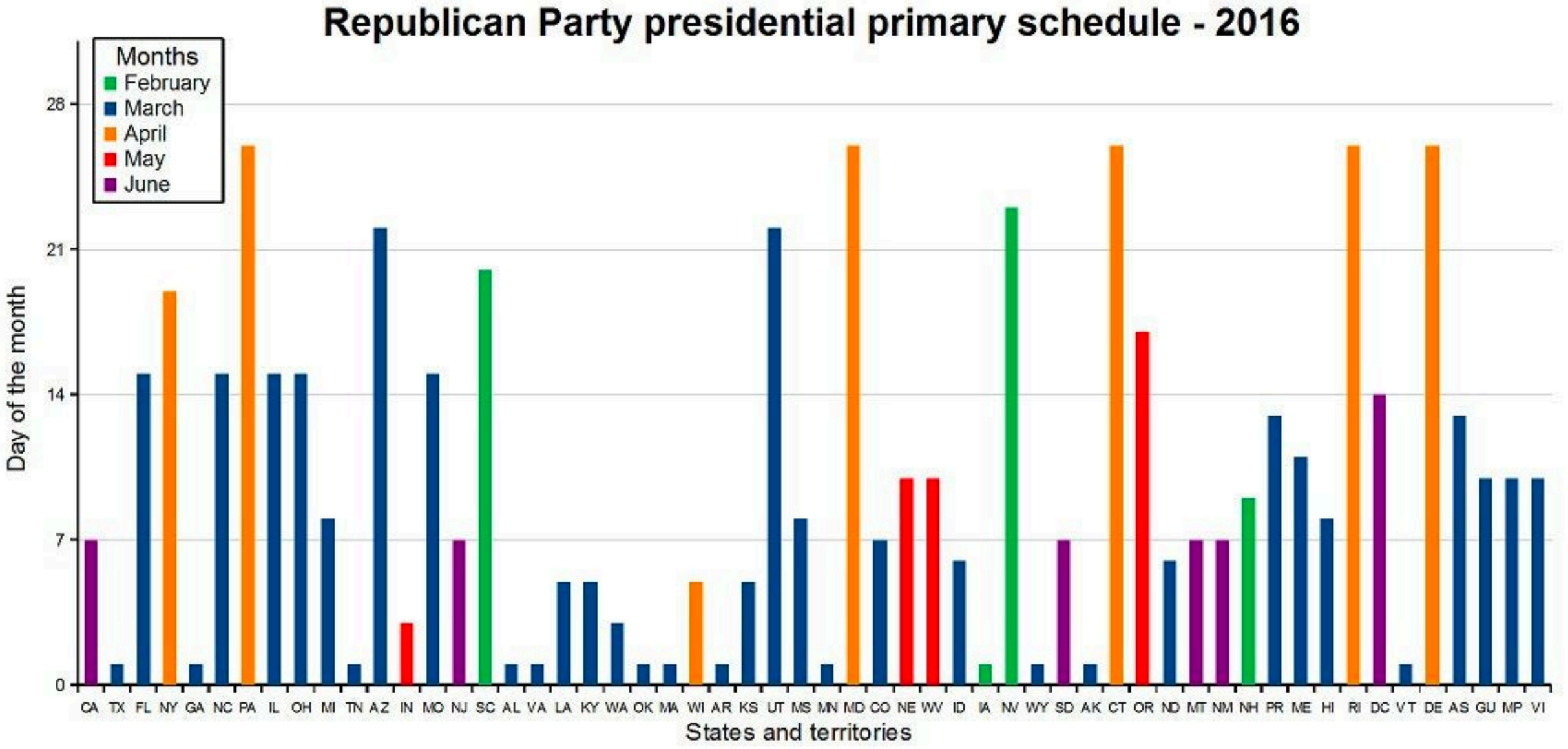



60
40
20
deaths / 1000
Urban Female
Urban Male
Rural Female
Rural Male
70-74
65-69
60-64
55-59
50-54


rural female
rural male
urban female
urban male
deaths / 1000
0
20
40
60
50-54
55-59
60-64
65-69
70-74
age group

PLOTTING PITFALLS

- Axis trickery
- Violations of basic math
- Nearly content-free figures
- Gratuitous chartjunk
- Poorly chosen 3D graphics
- Bad design choices



May we suggest... a calendar?

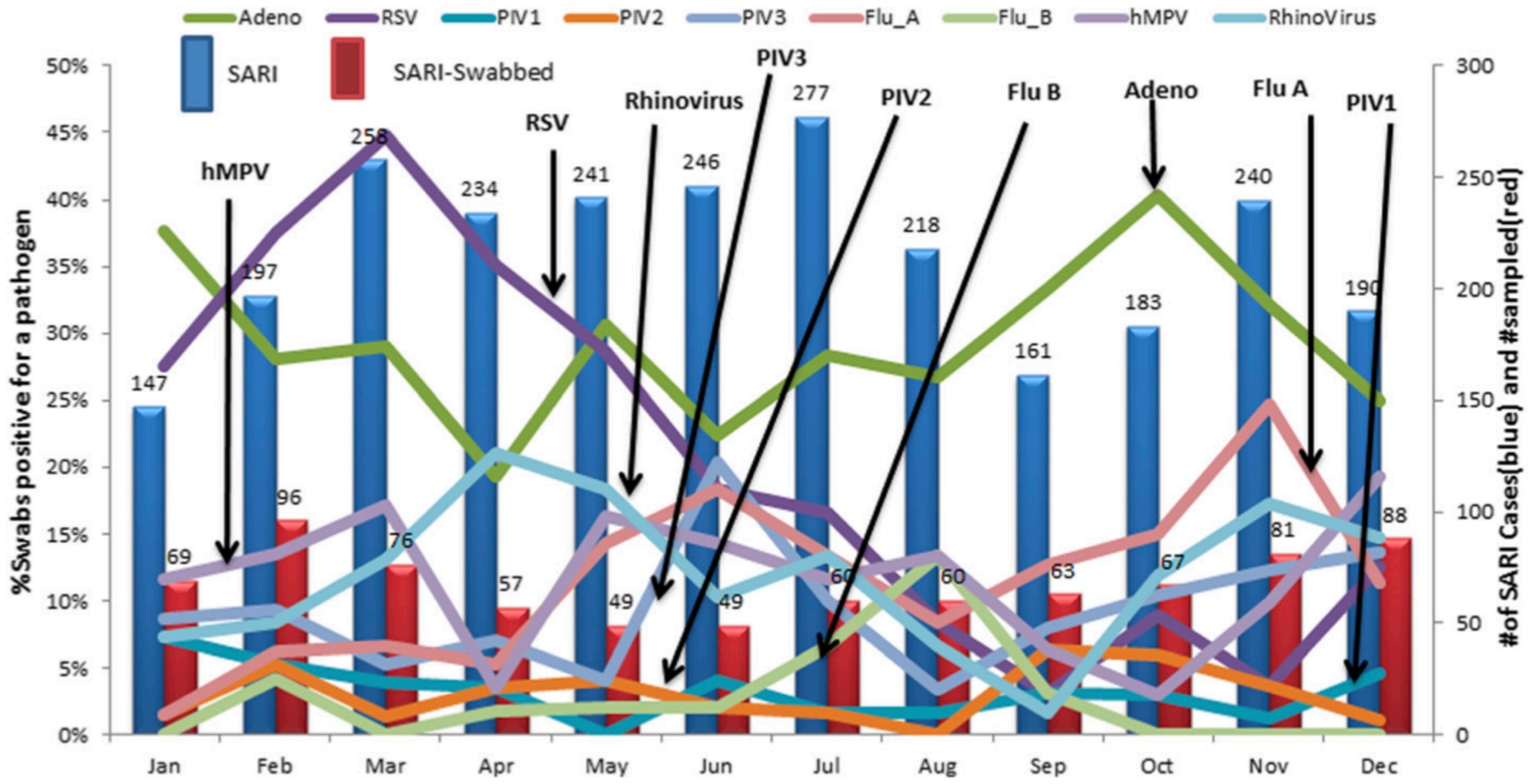
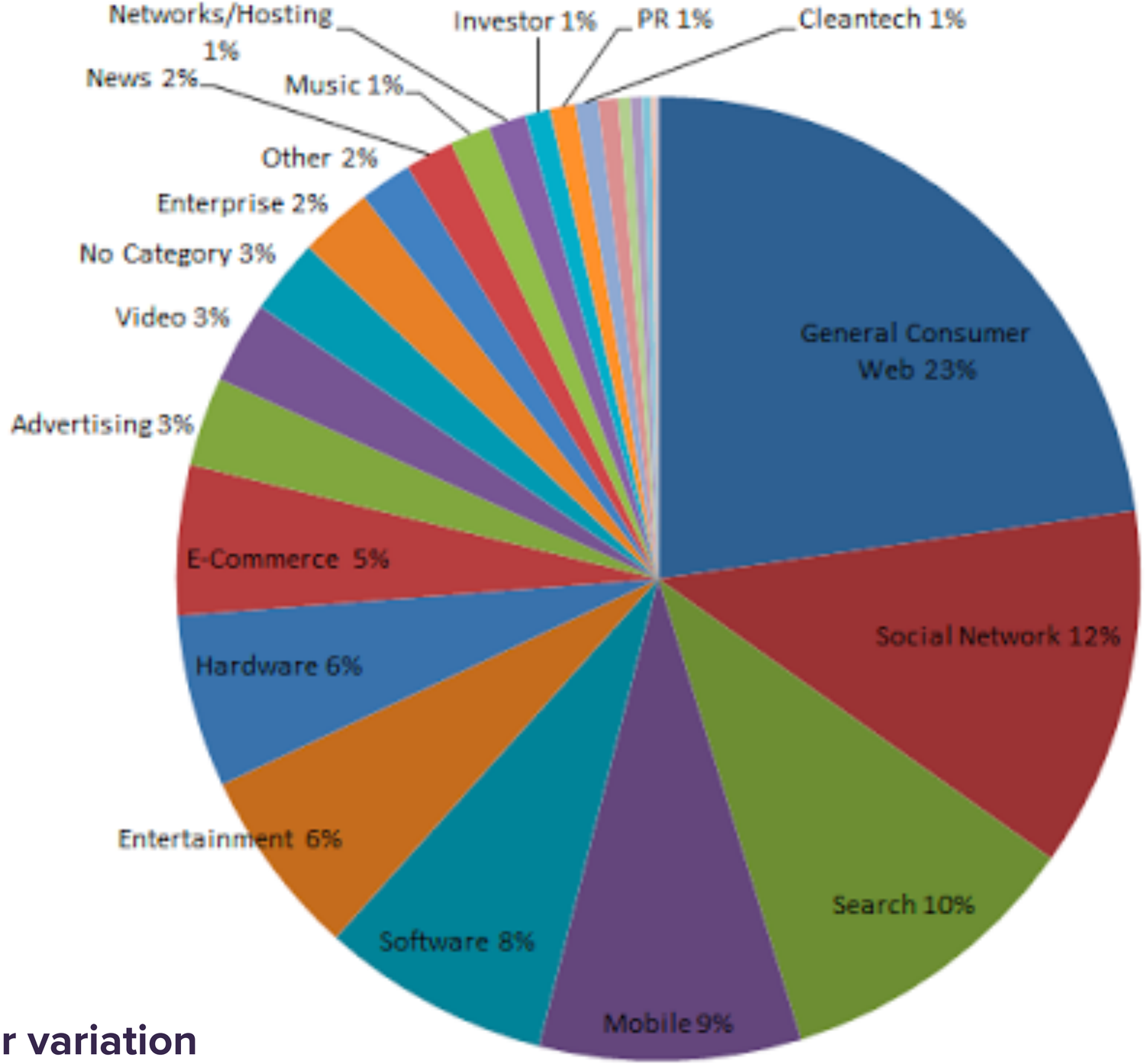
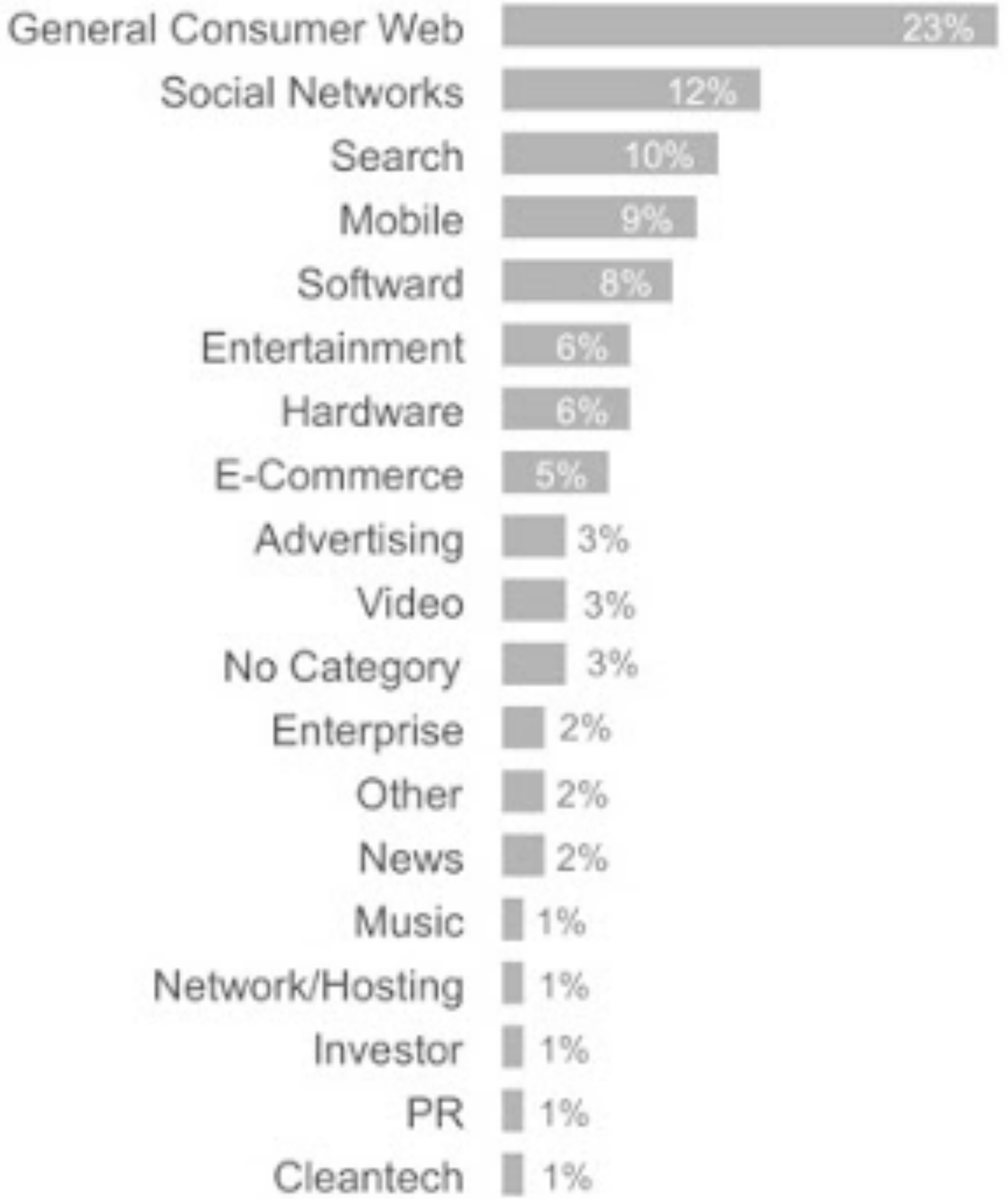


Figure 1 Monthly (aggregate) distribution of viral pathogens, March 1, 2007-Feb 28, 2011.



Meaningless color variation


General Consumer Web 23%
Social Networks 12%
Search 10%
Mobile 9%
Softward 8%
Entertainment 6%
Hardware 6%
E-Commerce 5%
Advertising 3%
Video 3%
No Category 3%
Enterprise 2%
Other 2%
News 2%
Music 1%
Network/Hosting 1%
Investor 1%
PR 1%
Cleantech 1%

As a bar plot

TAKE-HOME LESSONS

- Make relevant comparisons and avoid irrelevant ones.
- Be truthful about magnitude.
- Minimize distractions.
- Stay 2D (unless visualizing an actual 3D object, like a mountain or a building).
- Avoid meaningless variation (e.g. color differences that don’t encode information).
- Have clear axis labels and informative titles/annotations.
- Be sensitive to the knowledge and limitations of the viewer.

KEEP IT SIMPLE...

- Simple and effective plots:
  - Scatter plots
  - Line graphs
  - Boxplots
  - Histograms
  - Bar plots
- Simple and effective strategies for enriching these plots:
  - Variation in size, shape, and/or color (but be colorblind-friendly and don’t go crazy)
  - Faceting: showing the same basic plot across multiple conditions
  - Labels

THE GRAMMAR OF GRAPHICS

- Consider sentences that an English learner (e.g. a toddler) might say:

  “Squirrel climbs tree.”

  “Anna eats cake.”

  “Jimmy needs potty.”

- These sentences obey a simple, consistent grammar with three core elements:

  a *subject* that performs an action (squirrel, Anna).

  an *object* that receives the action (tree, cake).

  a *verb*, i.e. the action itself (climbs, eats).

- If you swap out different subjects/objects/verbs in the basic grammar, you can generate lots of different, fully comprehensible sentences.

THE GRAMMAR OF GRAPHICS

- Statistical plots are just the same.
- They obey a simple, consistent grammar with three basic elements

  Variables in a data set (like the weight or engine size of a vehicle). These are like the subject of a sentence.

  Objects: specifically, geometric objects (like dot or lines or bars). These are like the object of a sentence.

  Mappings from the data variables to aesthetic properties of the geometric objects (like their size, location, or color). These are like the verb in a sentence.

- If you swap out different variables/objects/mappings in the basic grammar, you can generate lots of different plots.

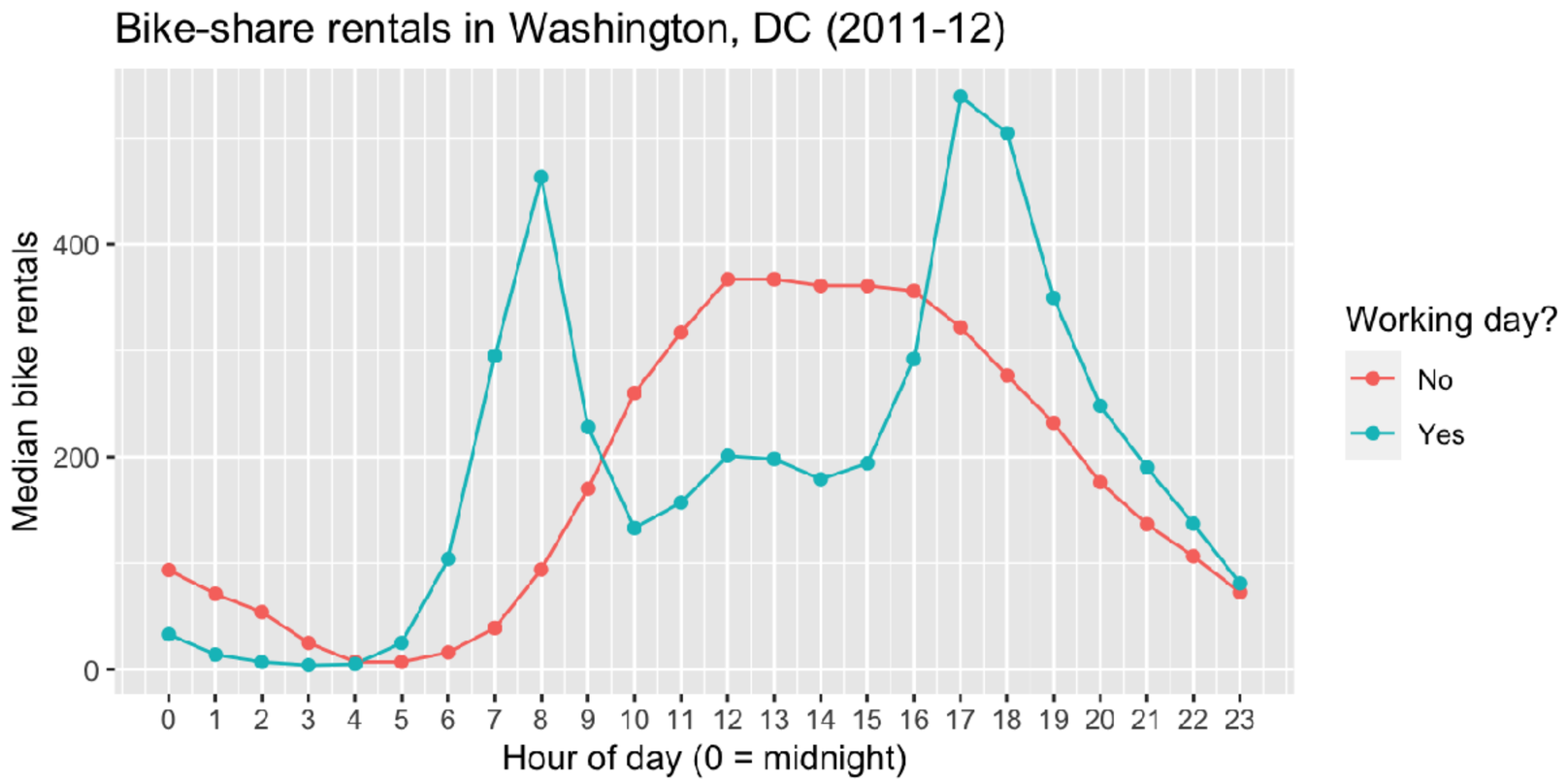
EXAMPLE: BIKE SHARE DATA

Here you see the first several rows of a data set on median hourly demand for bike-share rentals in Washington, DC's Capital Bikeshare program, stratified by whether it's a working day (1) or not (0).

hr	workingday	median_rentals
0	0	94
0	1	33
1	0	71
1	1	14
2	0	54
2	1	7

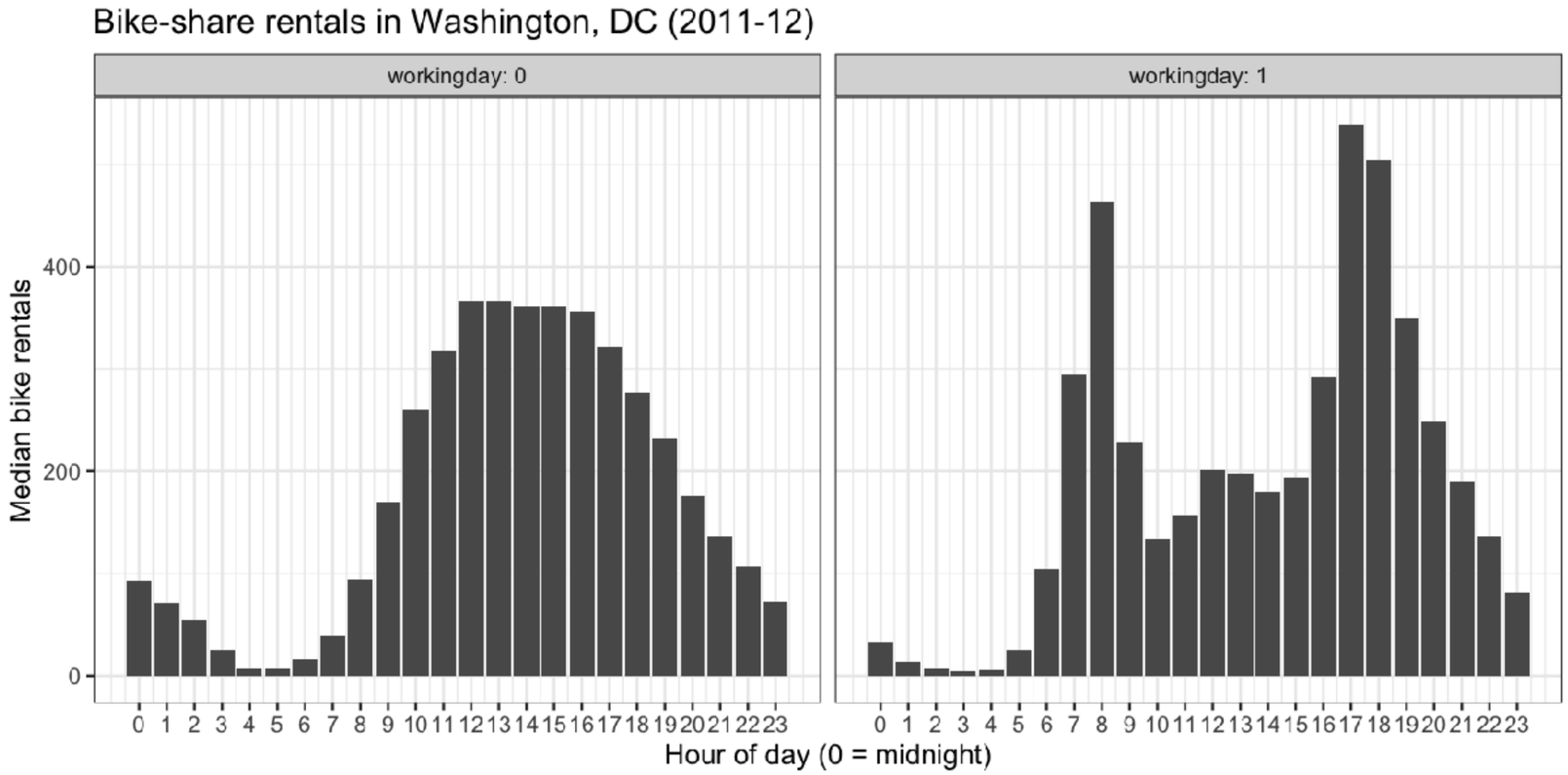
EXAMPLE: BIKE SHARE DATA

- **Variables:**
  - Hour
  - Median rentals
  - Working day
- **Objects:**
  - Points/lines
- **Mappings:**
  - Hour –> x
  - Median rentals —> y
  - Working day —> color



EXAMPLE: BIKE SHARE DATA

- **Variables:**
  - Hour
  - Median rentals
  - Working day
- **Objects:**
  - The two panels
  - The bars in each panel
- **Mappings:**
  - Hour –> location of bar
  - Median rentals —> height of bar
  - Working day —> left vs. right panel



Same variables, but different objects/ mappings, and so a different plot.

A plot is a mapping of data variables to the aesthetic properties of geometric objects.

THE FIVE MOST IMPORTANT PLOTS

If you have…	Consider a…	Using…
Numerical y and numerical x	Scatter plot	geom_point
Numerical y and sequential x	Line graph	geom_line
Numerical y alone	Histogram	geom_histogram
Numerical y and categorical x	Boxplot	geom_boxplot
Summary statistics by group	Bar plot	geom_col

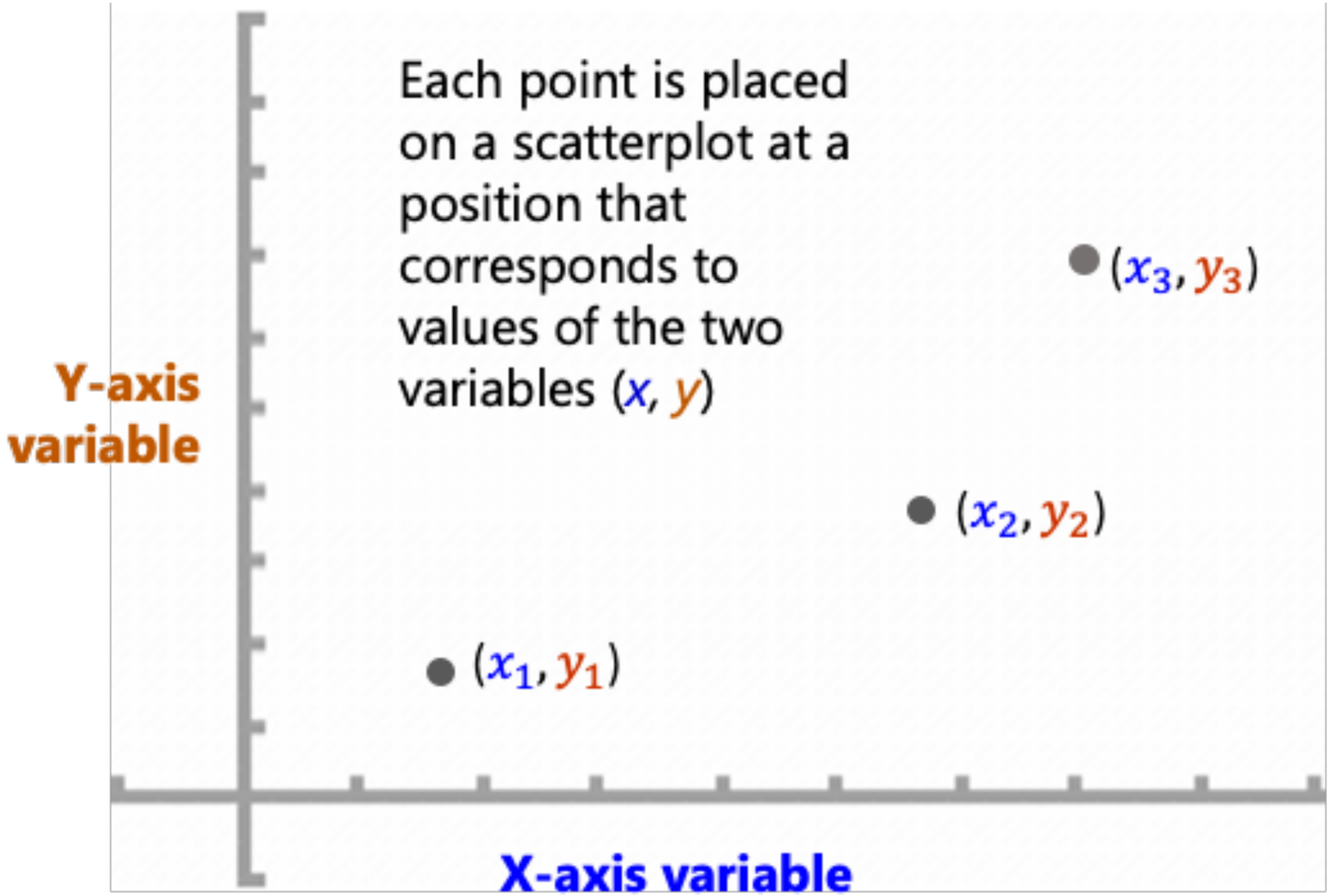
In the table, x refers to an explanatory variable and y refers to a response variable (i.e. the variable whose behavior is being "explained" by x.

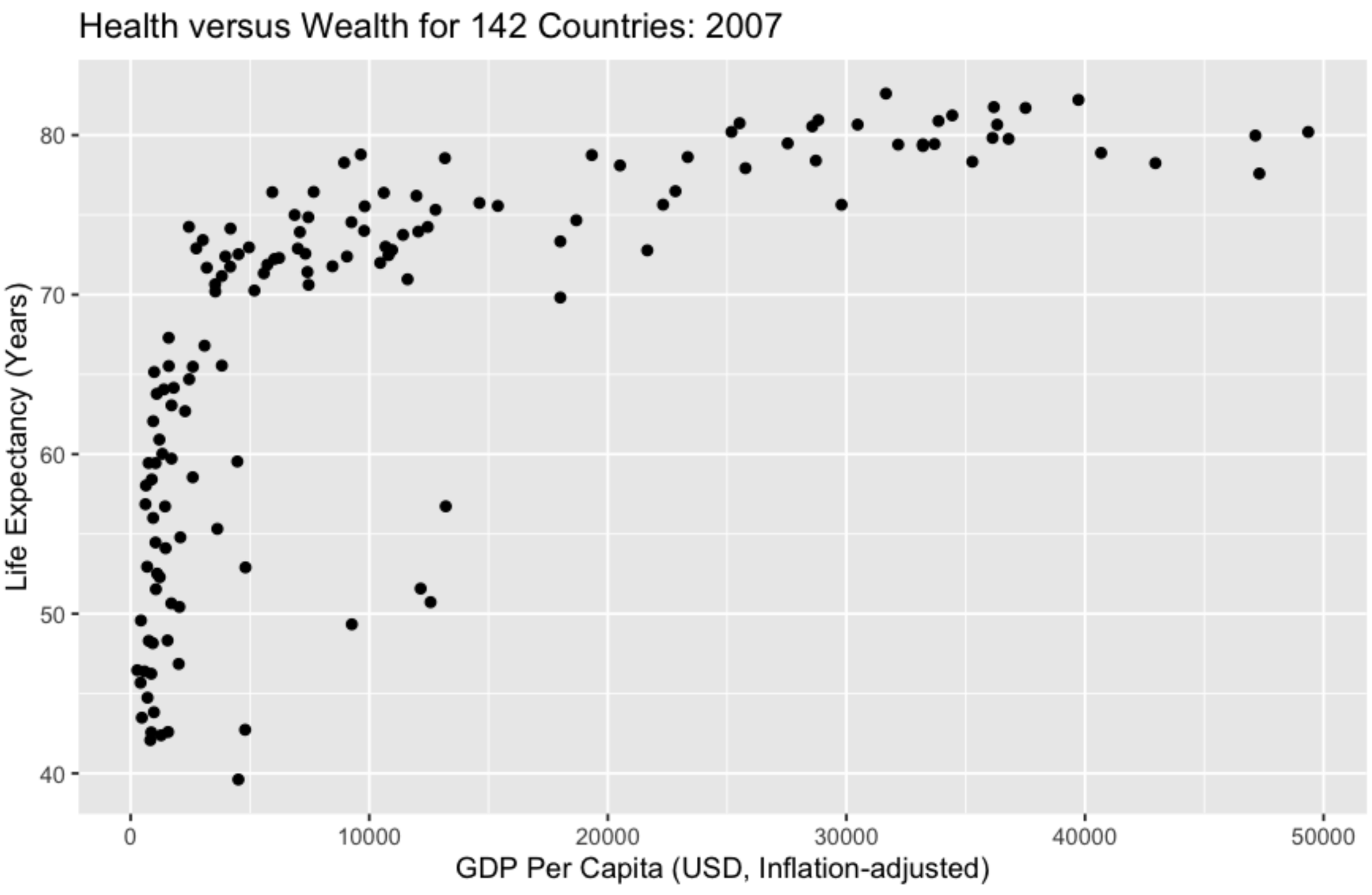
THE FIVE MOST IMPORTANT PLOTS

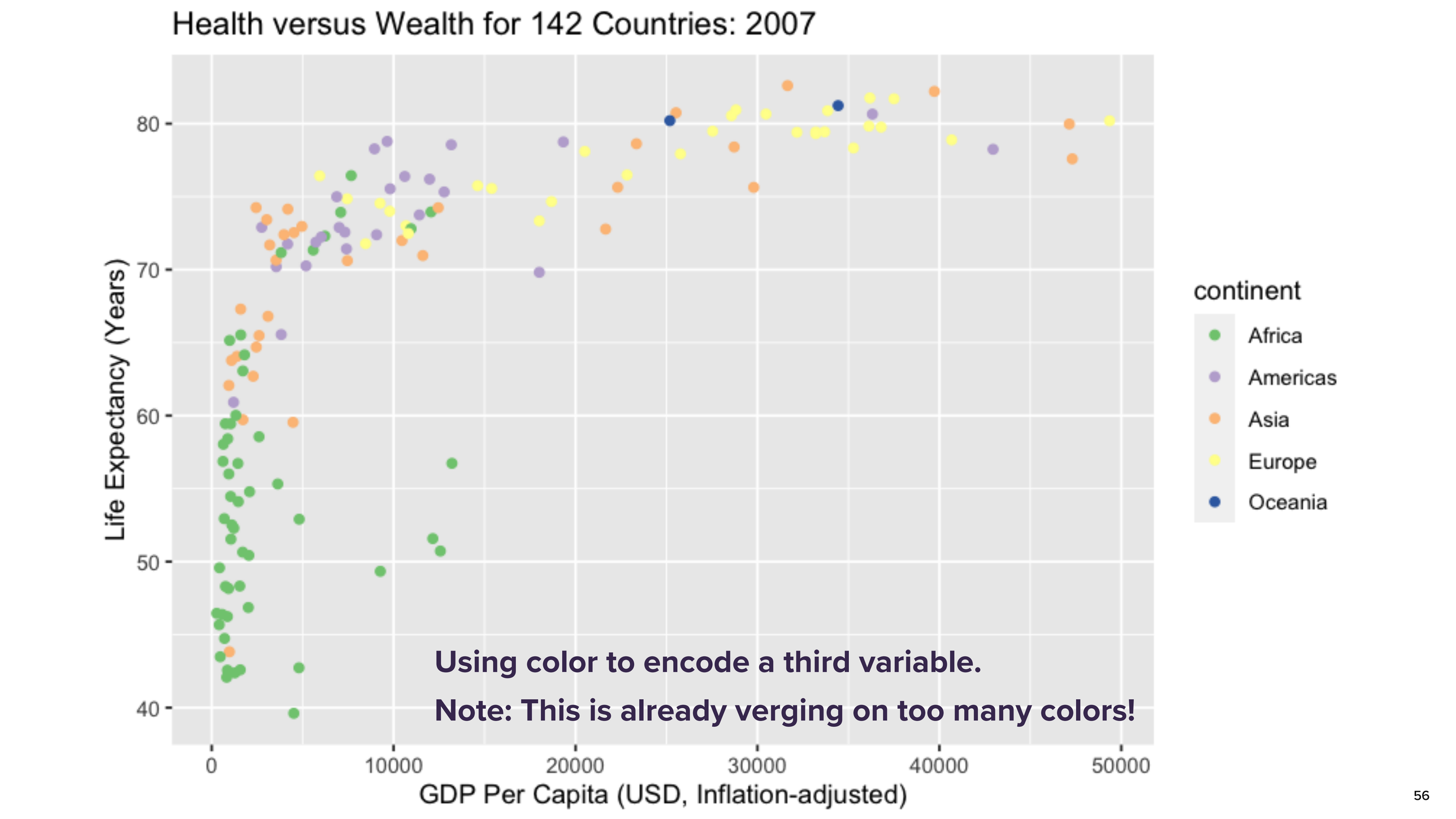
If you have…	Consider a…	Using…
Numerical y and numerical x	Scatter plot	geom_point
Numerical y and sequential x	Line graph	geom_line
Numerical y alone	Histogram	geom_histogram
Numerical y and categorical x	Boxplot	geom_boxplot
Summary statistics by group	Bar plot	geom_col

In the table, x refers to an explanatory variable and y refers to a response variable (i.e. the variable whose behavior is being "explained" by x.

SCATTER PLOTS




Health versus Wealth for 142 Countries: 2007
Life Expectancy (Years)
80
70
60
50
40
0
10000
20000
30000
40000
50000
GDP Per Capita (USD, Inflation-adjusted)


Health versus Wealth for 142 Countries: 2007
Life Expectancy (Years)
80
70
60
50
40
0
10000
20000
30000
40000
50000
GDP Per Capita (USD, Inflation-adjusted)
continent
Africa
Americas
Asia
Europe
Oceania
Using color to encode a third variable.
Note: This is already verging on too many colors!

THE FIVE MOST IMPORTANT PLOTS

If you have…	Consider a…	Using…
Numerical y and numerical x	Scatter plot	geom_point
Numerical y and sequential x	Line graph	geom_line
Numerical y alone	Histogram	geom_histogram
Numerical y and categorical x	Boxplot	geom_boxplot
Summary statistics by group	Bar plot	geom_col

In the table, x refers to an explanatory variable and y refers to a response variable (i.e. the variable whose behavior is being "explained" by x.

LINE GRAPHS

- Are used to show how a numerical variable (y) changes as a function of another numerical variable (x).
- Are like a scatter plot, but where you connect the dots in order of the x variable.
- Only make sense where there is an obvious sequential ordering to the x variable.

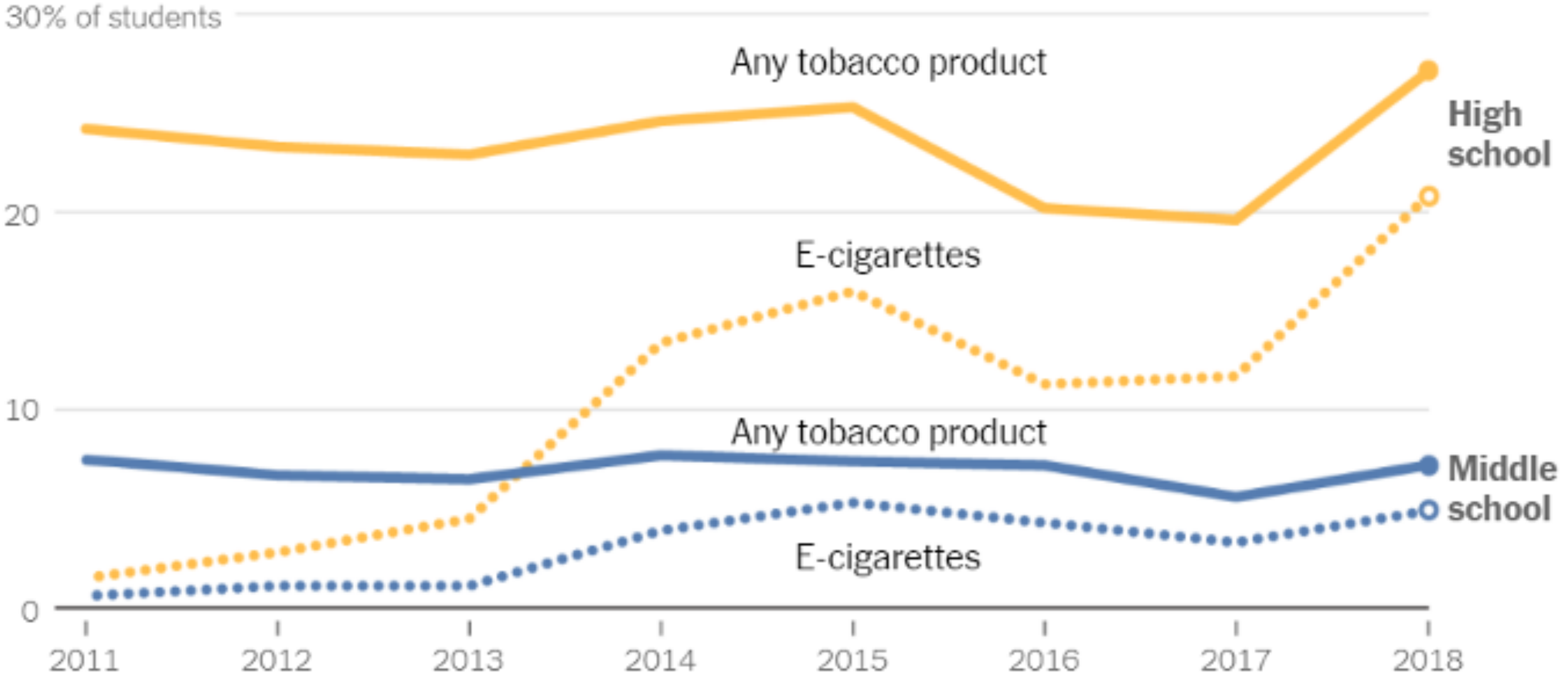
  Time

  Dosage levels of a drug

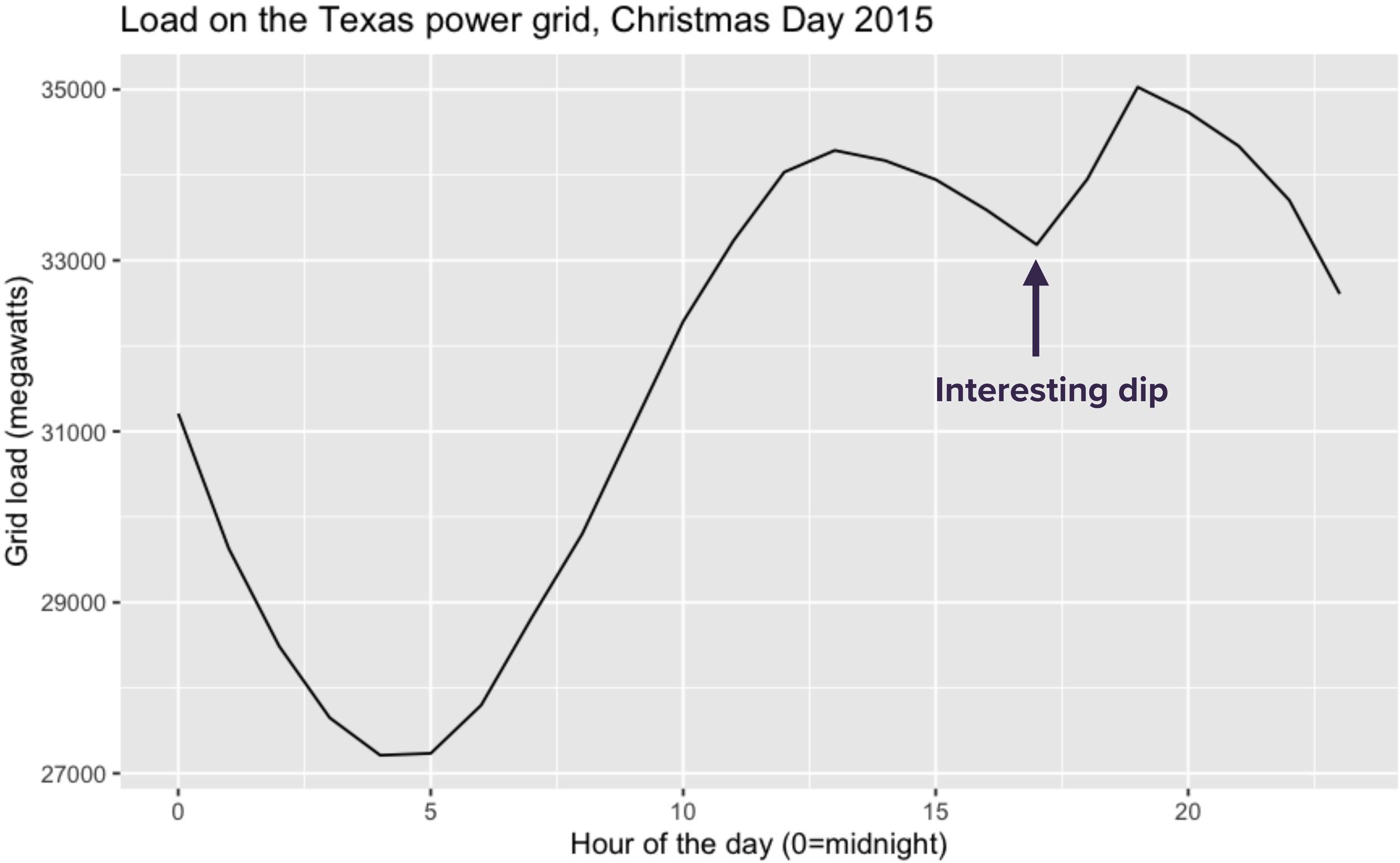
  Distance

  Etc.

Tobacco Consumption Among Students



Source: Centers for Disease Control and Prevention | By The New York Times



We'll make this line graph ourselves.

Data from ERCOT

THE FIVE MOST IMPORTANT PLOTS

If you have…	Consider a…	Using…
Numerical y and numerical x	Scatter plot	geom_point
Numerical y and sequential x	Line graph	geom_line
Numerical y alone	Histogram	geom_histogram
Numerical y and categorical x	Boxplot	geom_boxplot
Summary statistics by group	Bar plot	geom_col

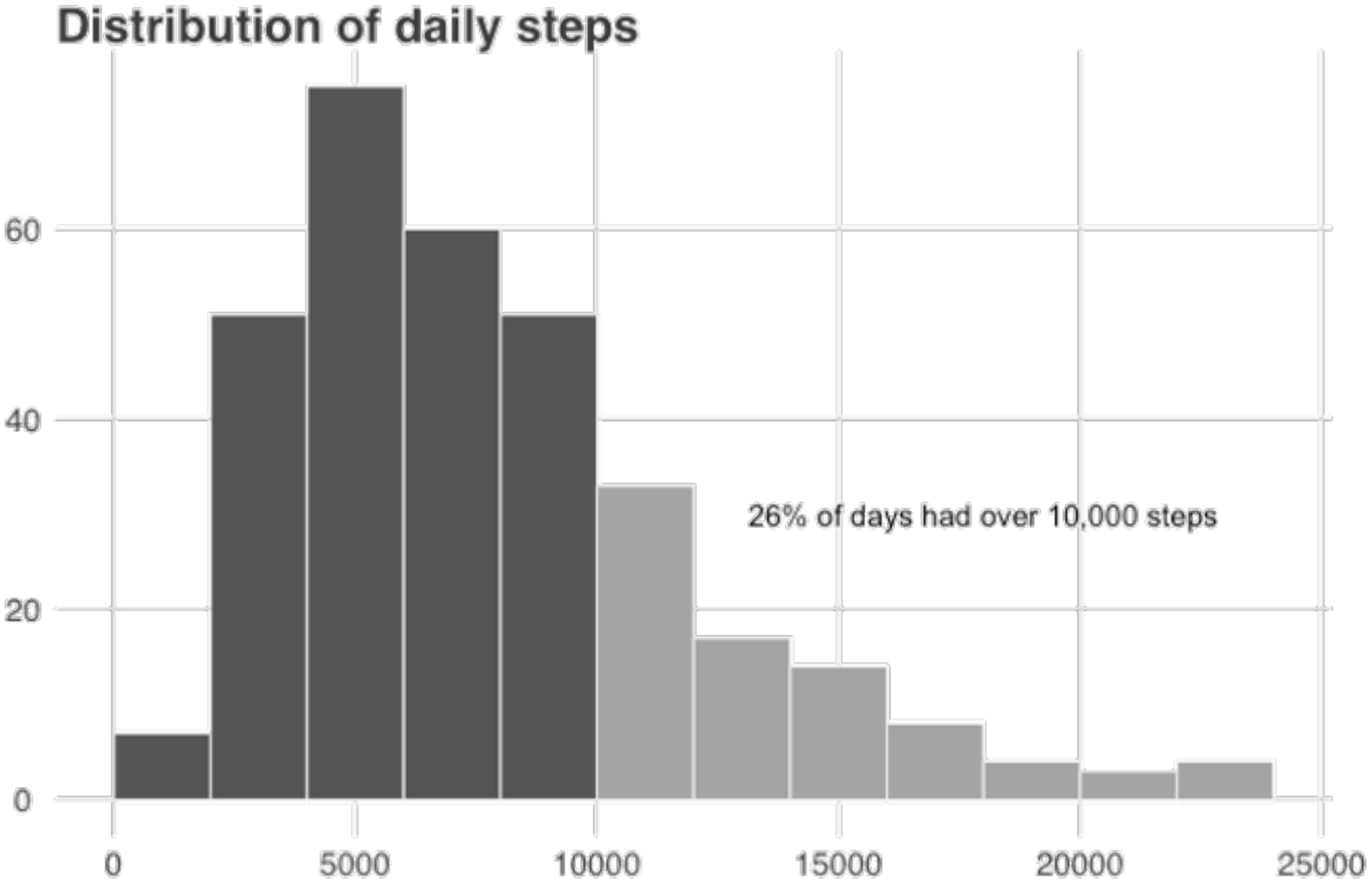
In the table, x refers to an explanatory variable and y refers to a response variable (i.e. the variable whose behavior is being "explained" by x.

HISTOGRAMS

- Are used to visualize the data distribution of a single numerical variable.
- Show all the data, regardless of the number of cases.
- Allow you to see the location, spread, number of peaks, and symmetry/skewness of the data distribution.
- Allow you to compare data distributions across conditions, via *faceting*.
- Come in two flavors:

  Count histograms: the height of each bar represents a count

  Density histograms: the height of each bar is rescaled so that the total area under the curve is 1

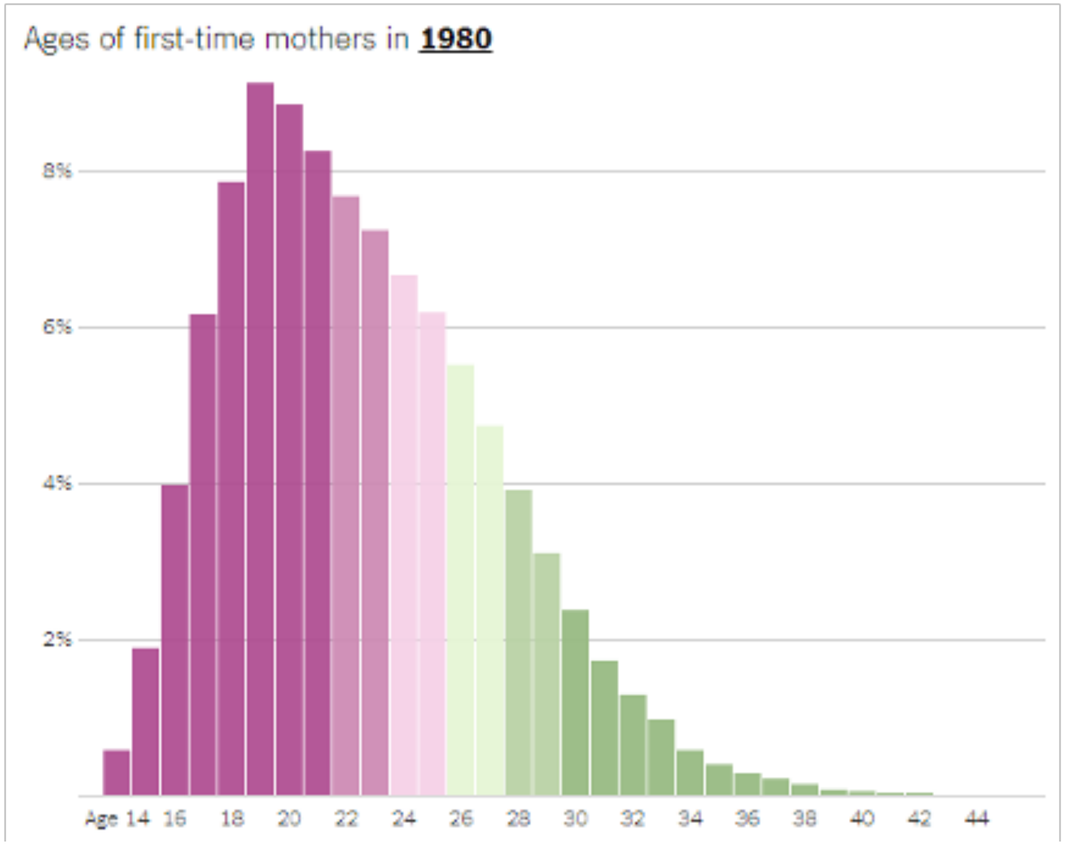

Distribution of daily steps
60
40
20
0
26% of days had over 10,000 steps
0
5000
10000
15000
20000
25000

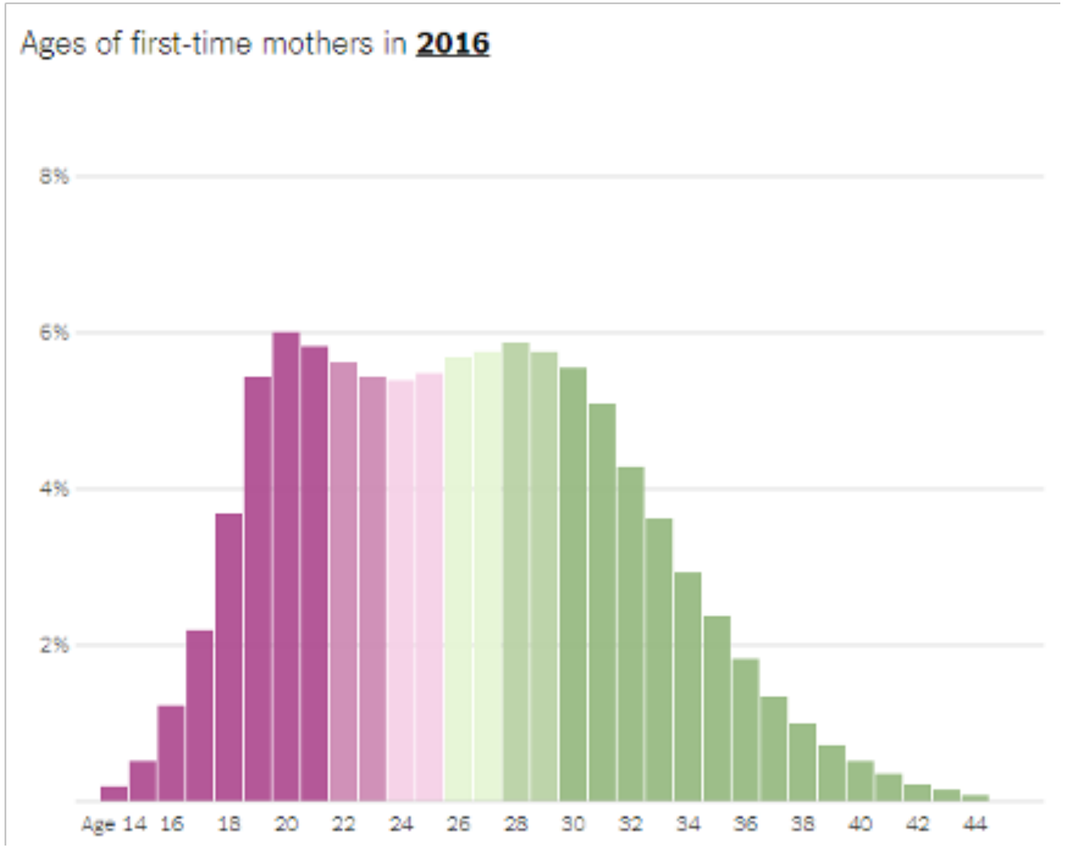
The Age That Women Have Babies: How a Gap Divides America

By QUOCTRUNG BUI and CLAIRE CAIN MILLER AUG. 4, 2018

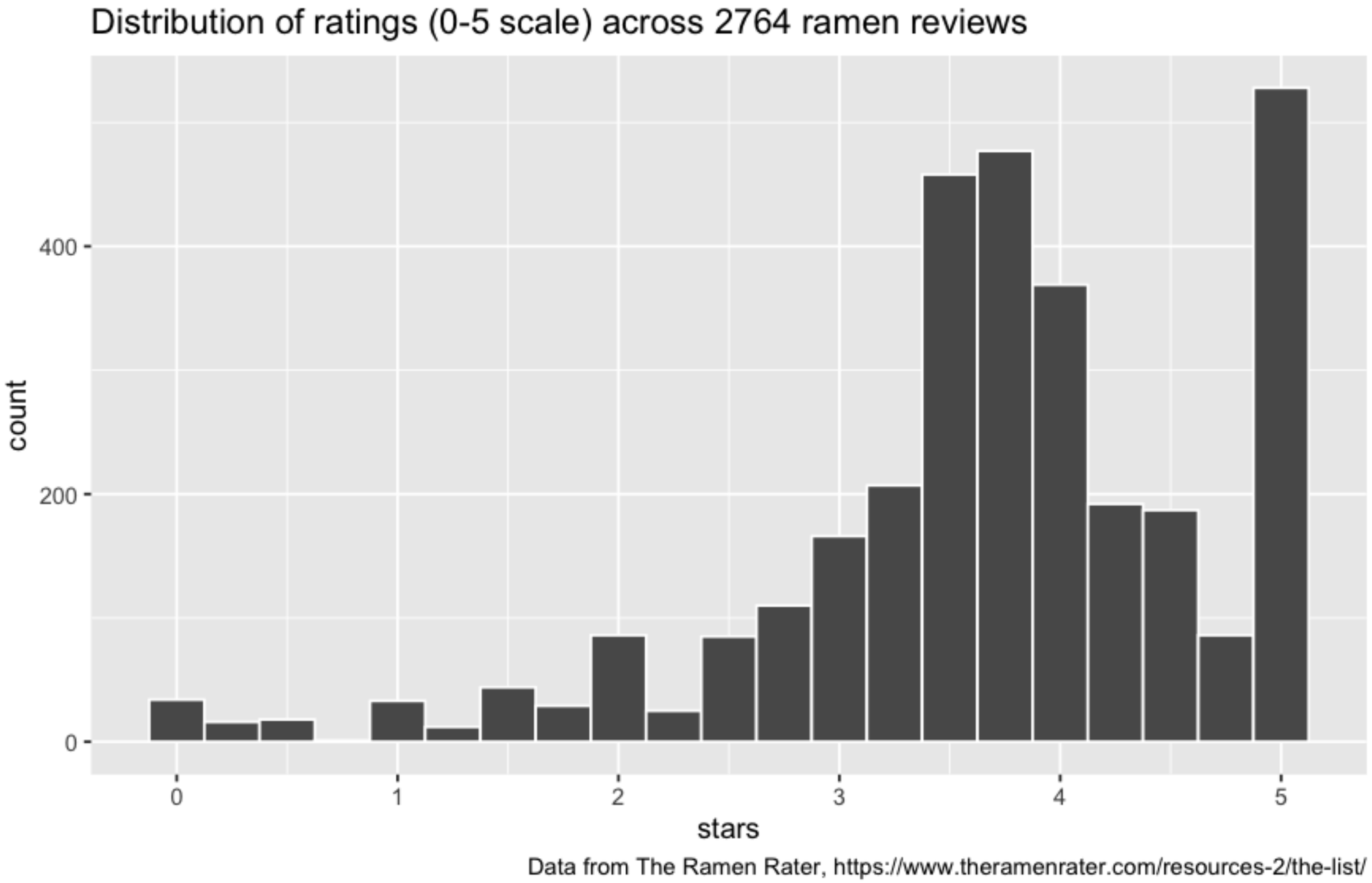
Incorporate a categorical variable (here, two "snapshots" of 1980 and 2016) with side-by-side *faceted histograms*.

The colors allow you to compare specific age ranges across the two years. The proportion of 30+ year-old moms really stands out this way.

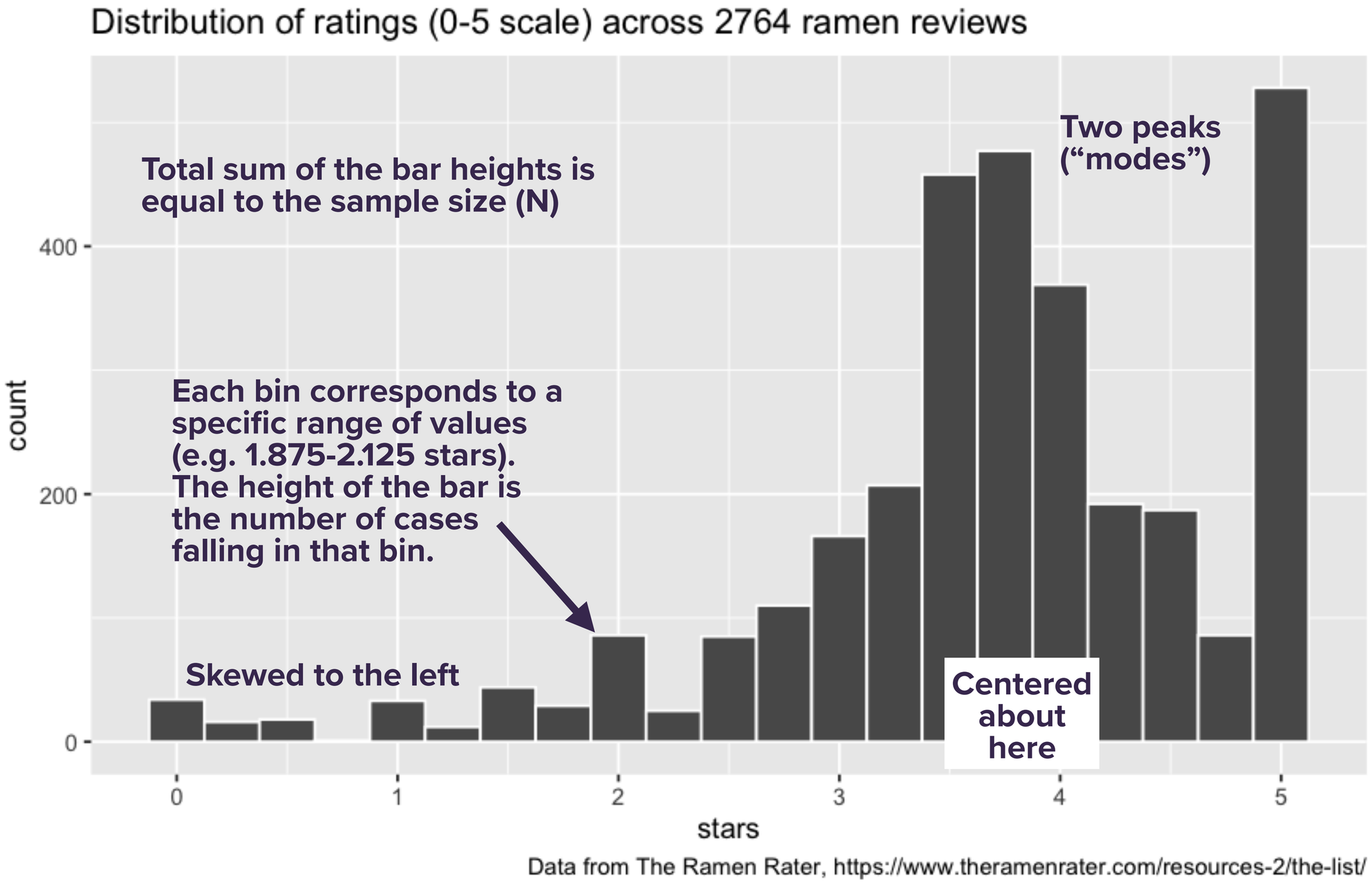




These are density histograms; the total area under each curve is 1 (100%).

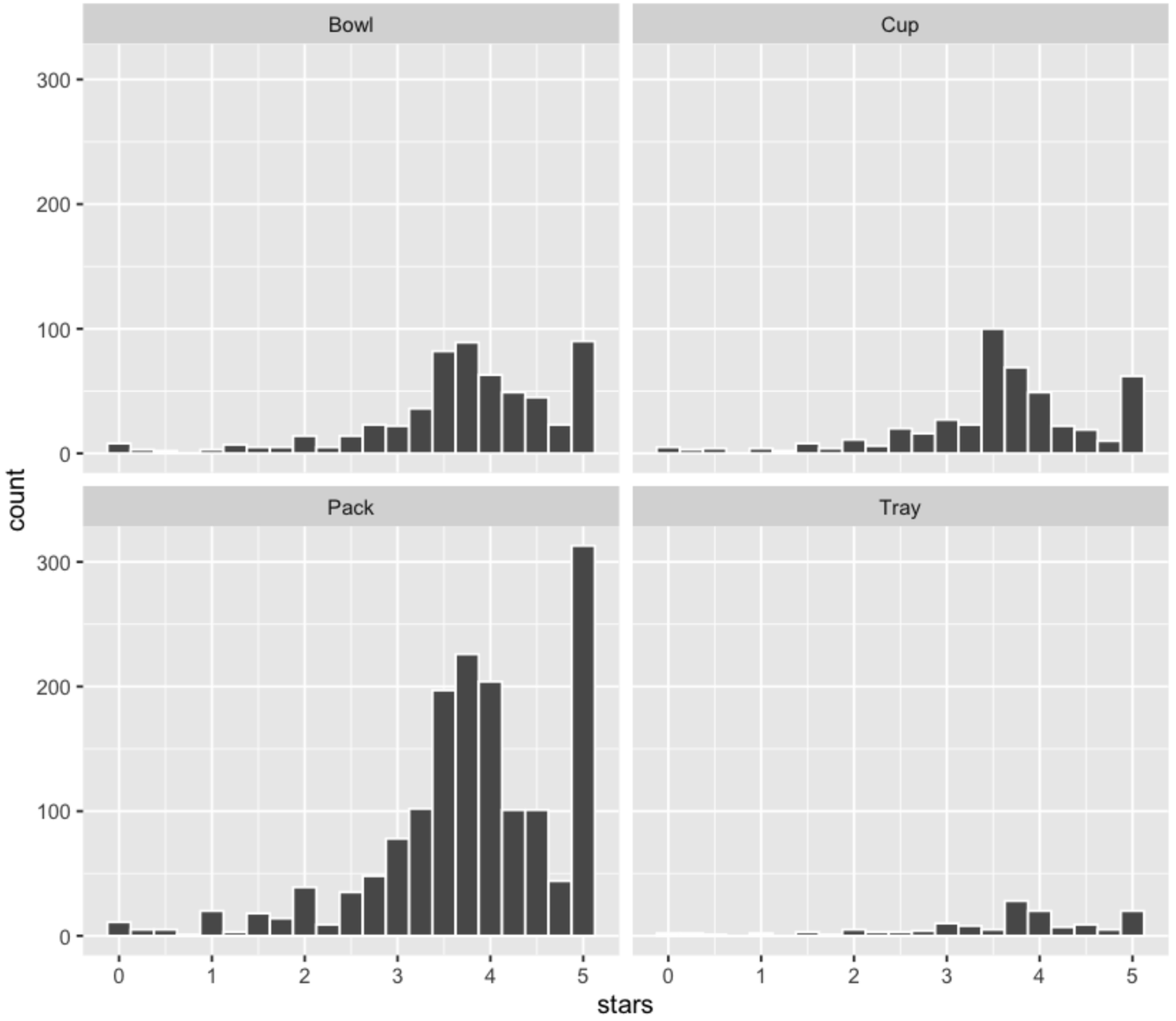


We'll make the next several histograms ourselves.


Distribution of ratings (0-5 scale) across 2764 ramen reviews
Total sum of the bar heights is equal to the sample size (N)
Two peaks ("modes")
Each bin corresponds to a specific range of values (e.g. 1.875-2.125 stars). The height of the bar is the number of cases falling in that bin.
Skewed to the left
Centered about here
count
400
200
0
0
1
2
3
4
5
stars
Data from The Ramen Rater, https://www.theramenrater.com/resources-2/the-list/

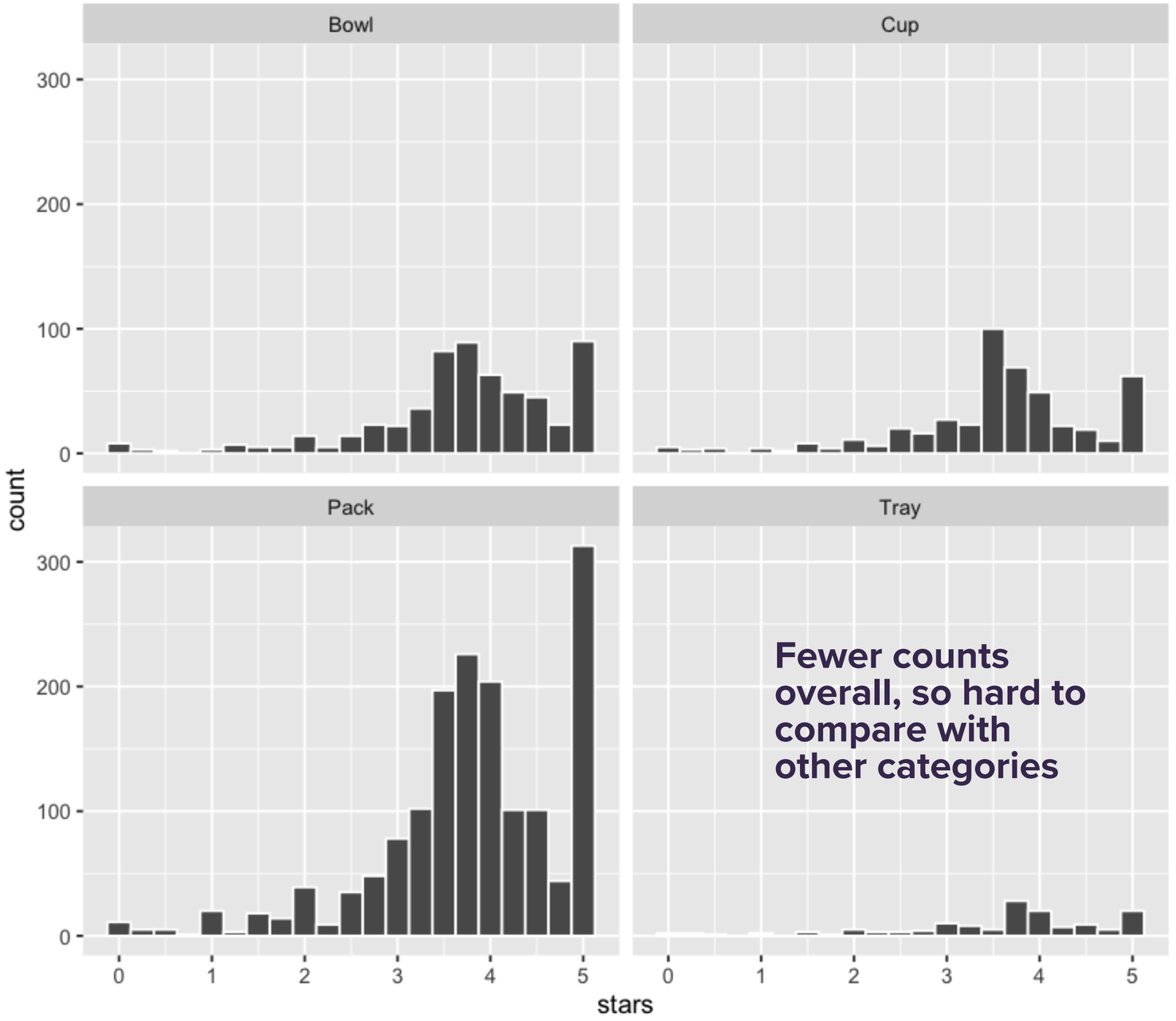
FACETING

- *Faceting* is used when we'd like to split a particular visualization by the values of another variable.
- We create multiple copies/ panels of the same type of plot with matching x and y axes, but with different content.
- Here, our faceting variable is the style of container (bowl, cup, pack, tray).



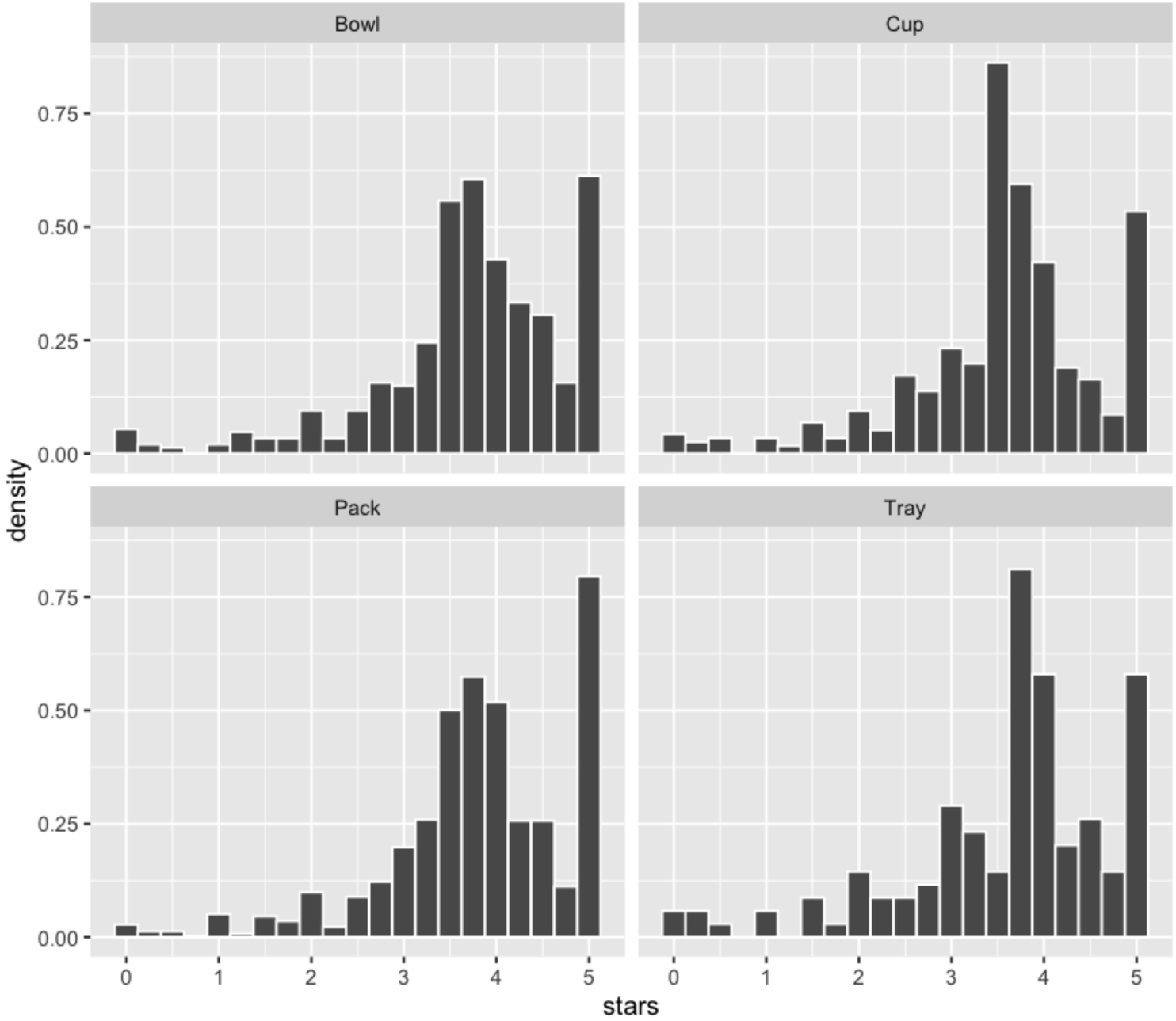
FACETING

- *Faceting* is used when we'd like to split a particular visualization by the values of another variable.
- We create multiple copies/ panels of the same type of plot with matching x and y axes, but with different content.
- Here, our faceting variable is the style of container (bowl, cup, pack, tray).



DENSITY HISTOGRAM

- At the right, we see a density histogram, rather than a histogram of raw counts.
- This entails renormalizing the vertical scale of each panel so that the total area of the bars sums to 1.
- Now it's easier to compare distributions across categories with disparate sample sizes.



THE FIVE MOST IMPORTANT PLOTS

If you have…	Consider a…	Using…
Numerical y and numerical x	Scatter plot	geom_point
Numerical y and sequential x	Line graph	geom_line
Numerical y alone	Histogram	geom_histogram
Numerical y and categorical x	Boxplot	geom_boxplot
Summary statistics by group	Bar plot	geom_col

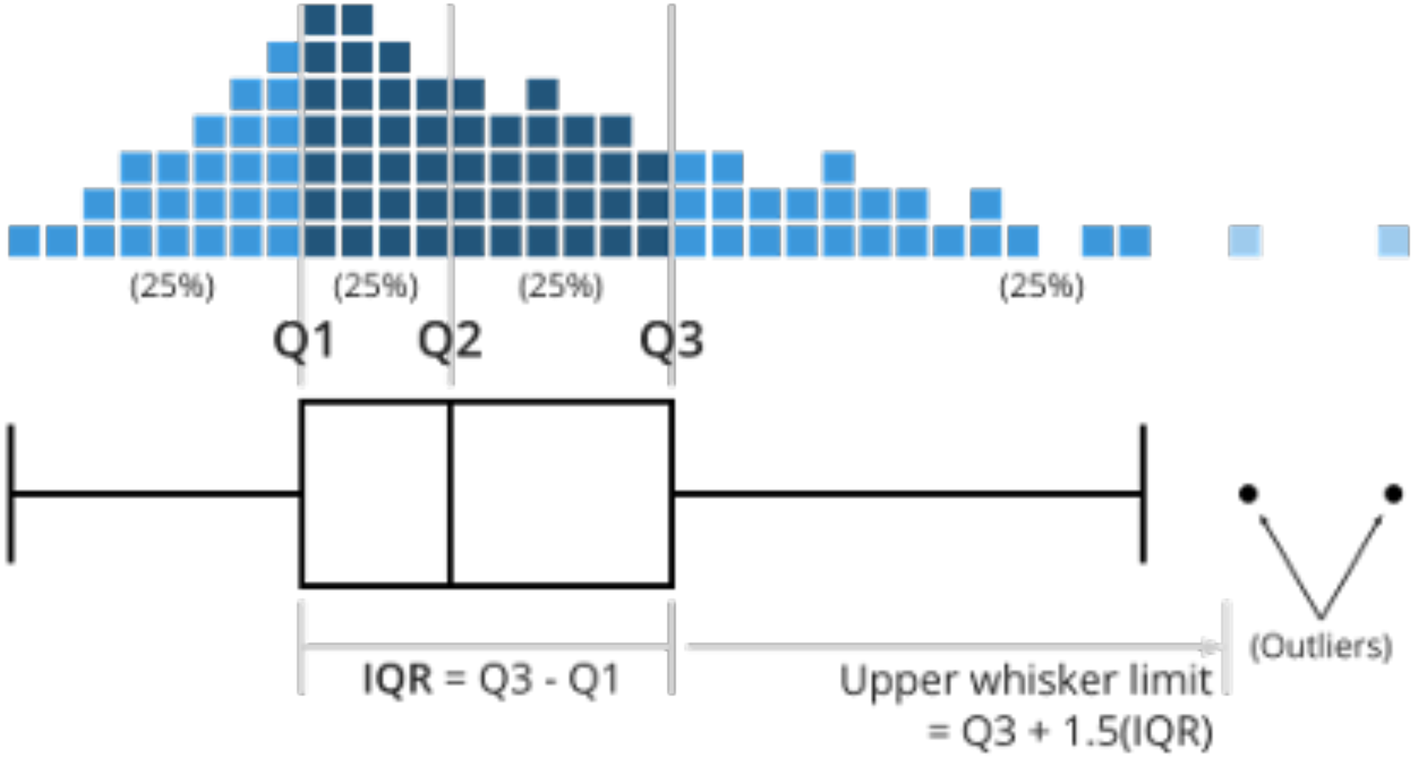
In the table, x refers to an explanatory variable and y refers to a response variable (i.e. the variable whose behavior is being "explained" by x.

BOXPLOTS

- Are used to compare the distribution of a numerical variable across groups.
- Show a five-number summary of the data distribution within each group
    - Median
    - 25th and 75th percentiles
    - Lower and upper extremes
- Are similar in spirit to faceted histograms, but don’t show the entire data distribution.

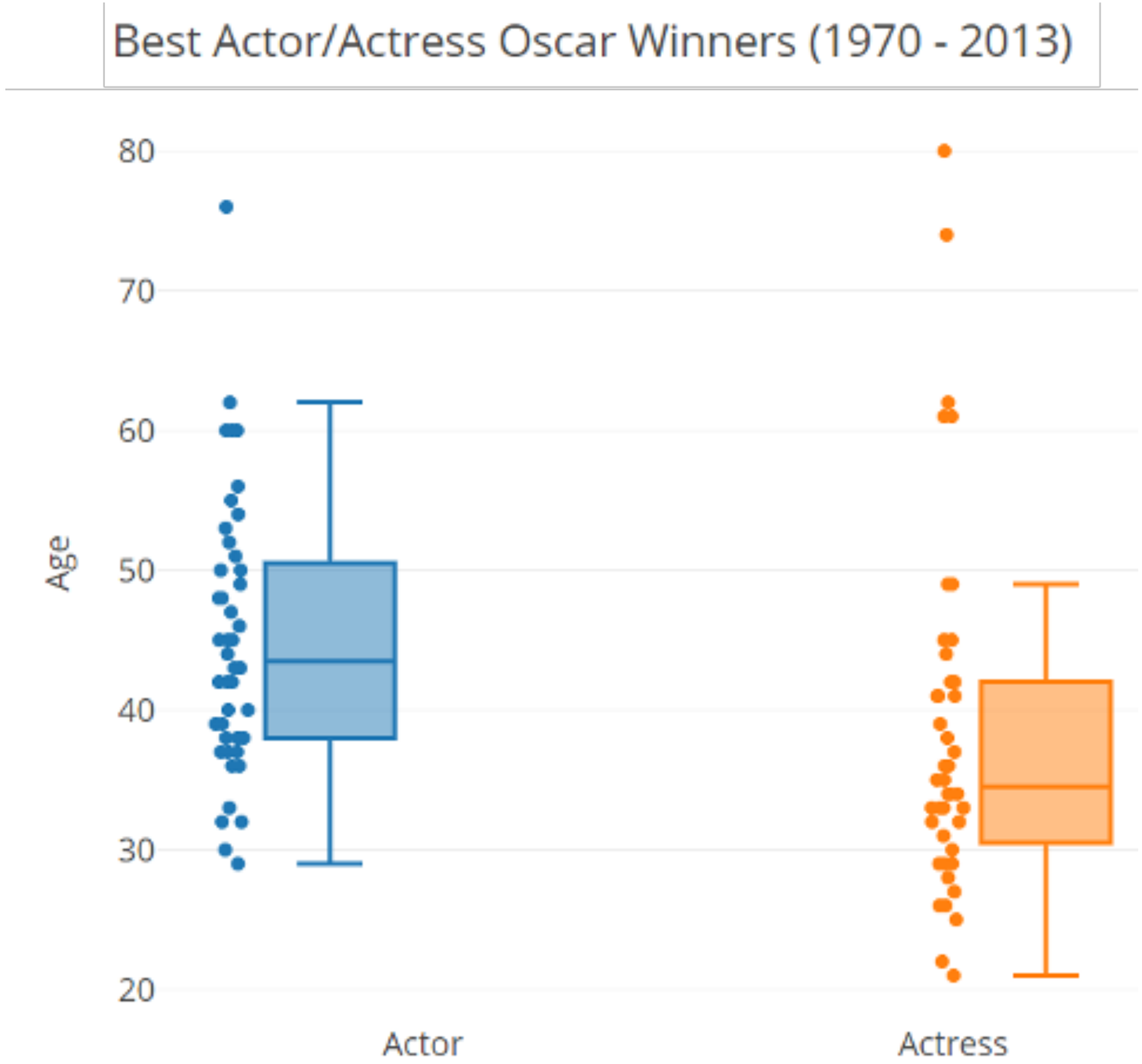
BOXPLOTS

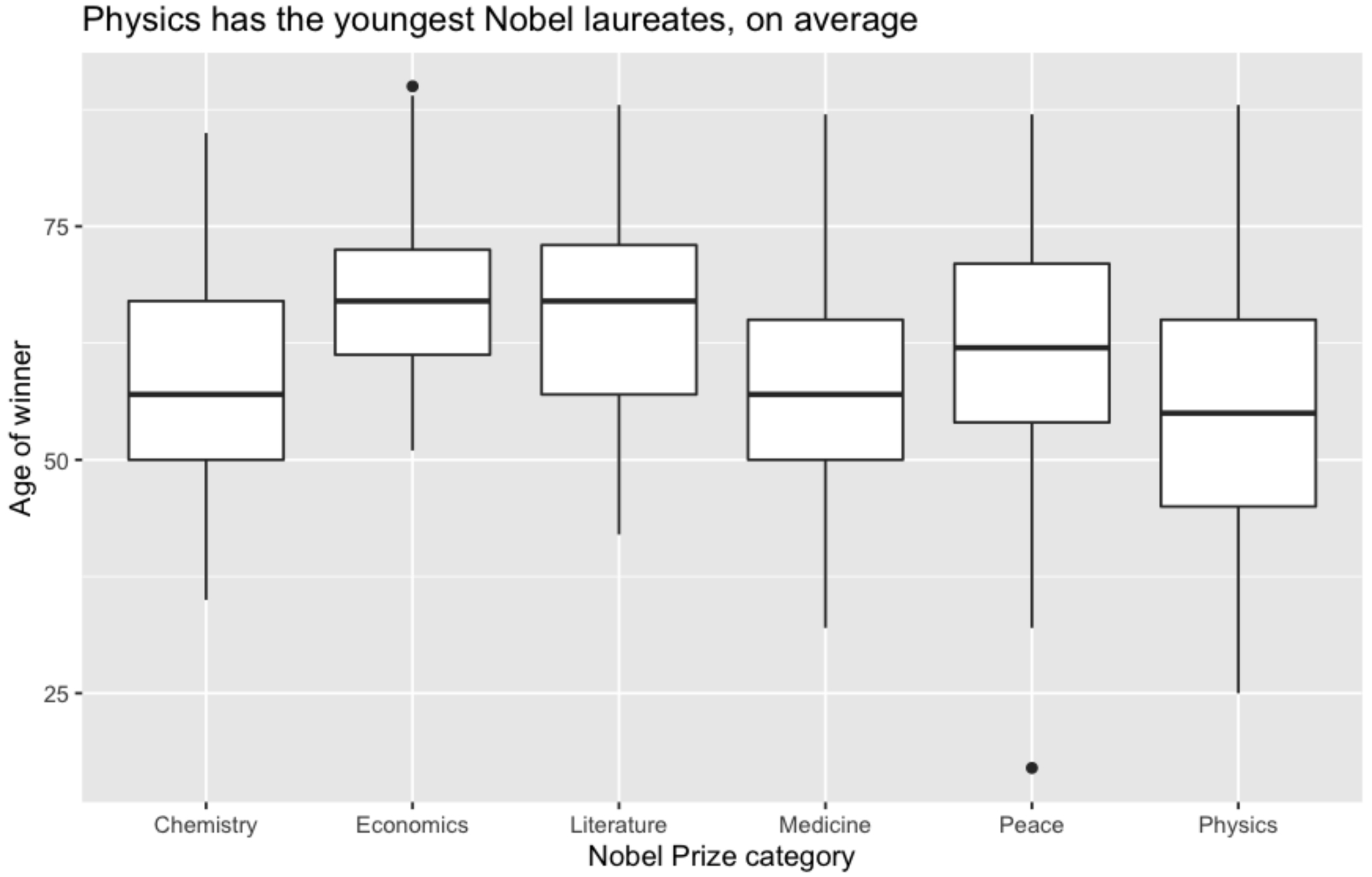
- The central box shows the middle half of the data: 25th to 75th percentiles, the interquartile range (IQR).
- If the median is roughly centered in the box, the middle half of the data is roughly symmetric. Here we see some mild skewness.
- The whiskers also show skewness based on comparing their lengths.
- A common (though not universal) convention is to cut the whiskers off at 1.5 times the IQR, rather than min/max.



BOXPLOTS

- Use side-by-side boxplots to compare distributions with respect to some categorical variable
- Here we see a boxplot of ages of male vs. female winners of the Best Actor/Actress Oscar.
- The data points are shown individually to give you some intuition for how the boxes relate to the raw data. For smaller data sets, this juxtaposition can be effective.
- The color is redundant here; it doesn't encode any extra information.




Physics has the youngest Nobel laureates, on average
Age of winner
75
50
25
Chemistry
Economics
Literature
Medicine
Peace
Physics
Nobel Prize category

THE FIVE MOST IMPORTANT PLOTS

If you have…	Consider a…	Using…
Numerical y and numerical x	Scatter plot	geom_point
Numerical y and sequential x	Line graph	geom_line
Numerical y alone	Histogram	geom_histogram
Numerical y and categorical x	Boxplot	geom_boxplot
Summary statistics by group	Bar plot	geom_col

In the table, x refers to an explanatory variable and y refers to a response variable (i.e. the variable whose behavior is being "explained" by x.

BAR PLOTS

- Are typically used to compare summary statistics across groups.

  Means or totals (for numerical variables)

  Counts or proportions (for categorical variables)

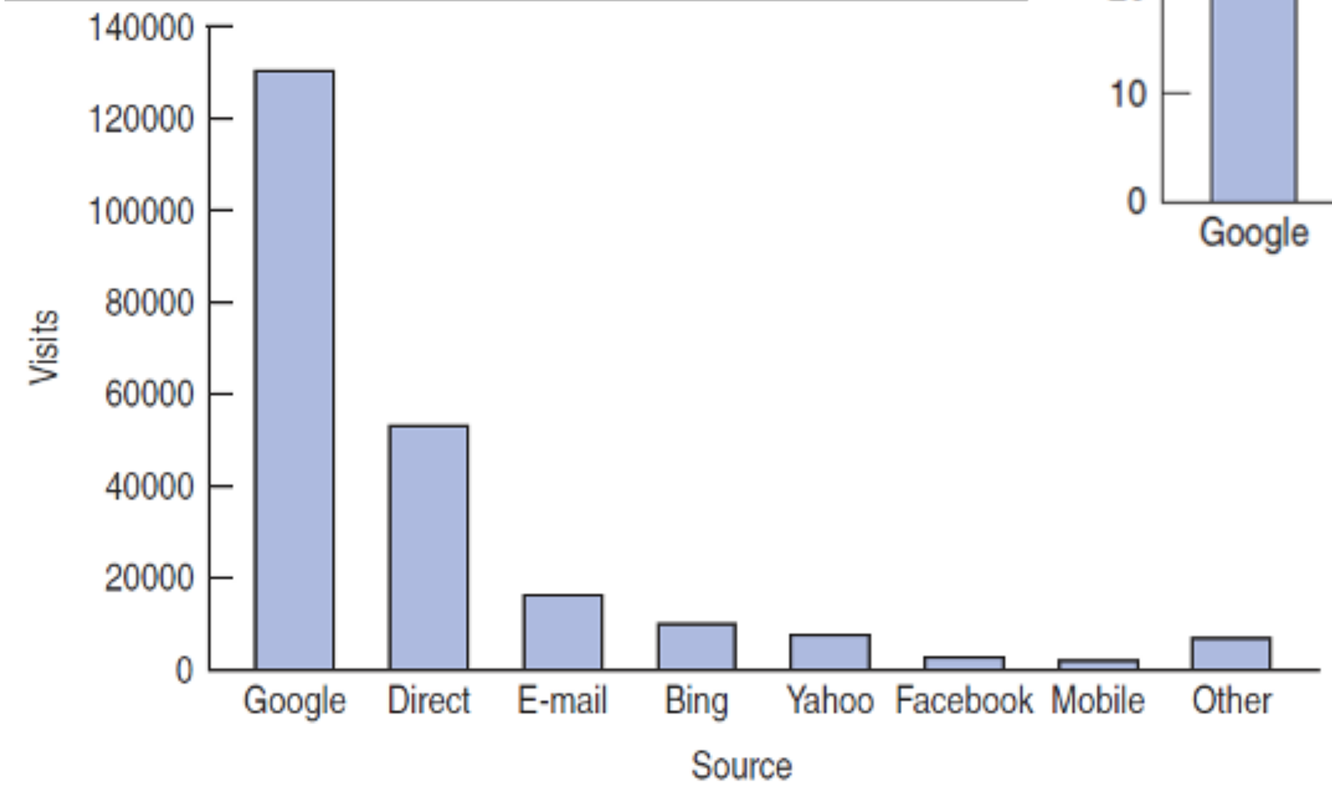
- Should always have a y axis that contains 0
- Generally require a two-stage process to make:

  (1) Summary stage: split your data set into subgroups and calculate summary statistics for each subgroup.

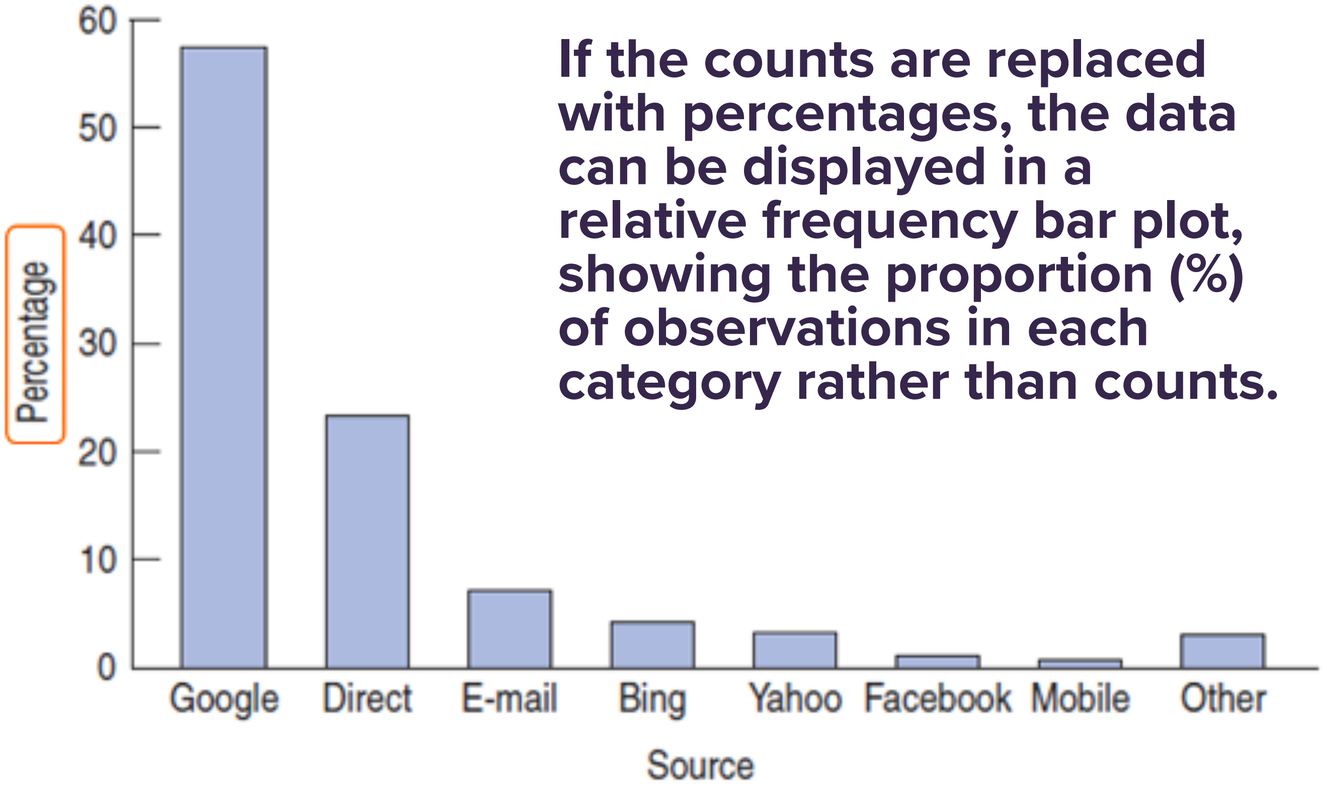
  (2) Plotting stage: make a bar plot of those summary statistics, one bar per group.

- Here we assume that stage 1 (calculating summaries), has been accomplished, and we start from stage 2 (plotting). Stage 1 comes later, in the lesson on Data Wrangling.

A bar plot displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.



If the counts are replaced with percentages, the data can be displayed in a relative frequency bar plot, showing the proportion (%) of observations in each category rather than counts.



Source	Visits	Visits by %
Google	130,158	57.36
Direct	52,969	23.34
E-mail	16,084	7.09
Bing	9,581	4.22
Yahoo	7,439	3.28
Facebook	2,253	0.99
Mobile	1,701	0.75
Other	6,740	2.97
Total	**226,925**	**100.00**