

# 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**

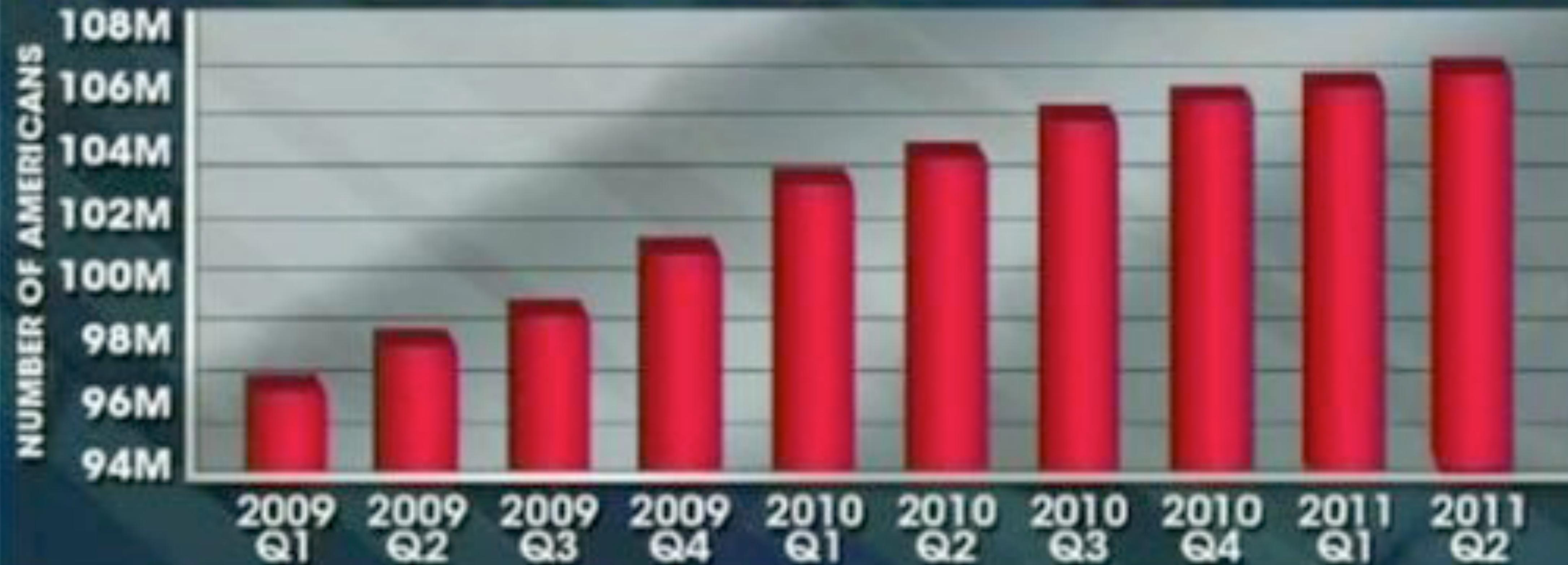
# PLOTTING PITFALLS

- **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**

# FEDERAL WELFARE RECEIVED IN U.S.



FOX NEWS FOX  
NEWS NETS  
.COM

SOURCE: U.S. CENSUS SURVEY

:LIST TYLER HAMILTON OF HIS GOLD MEDAL FRO!

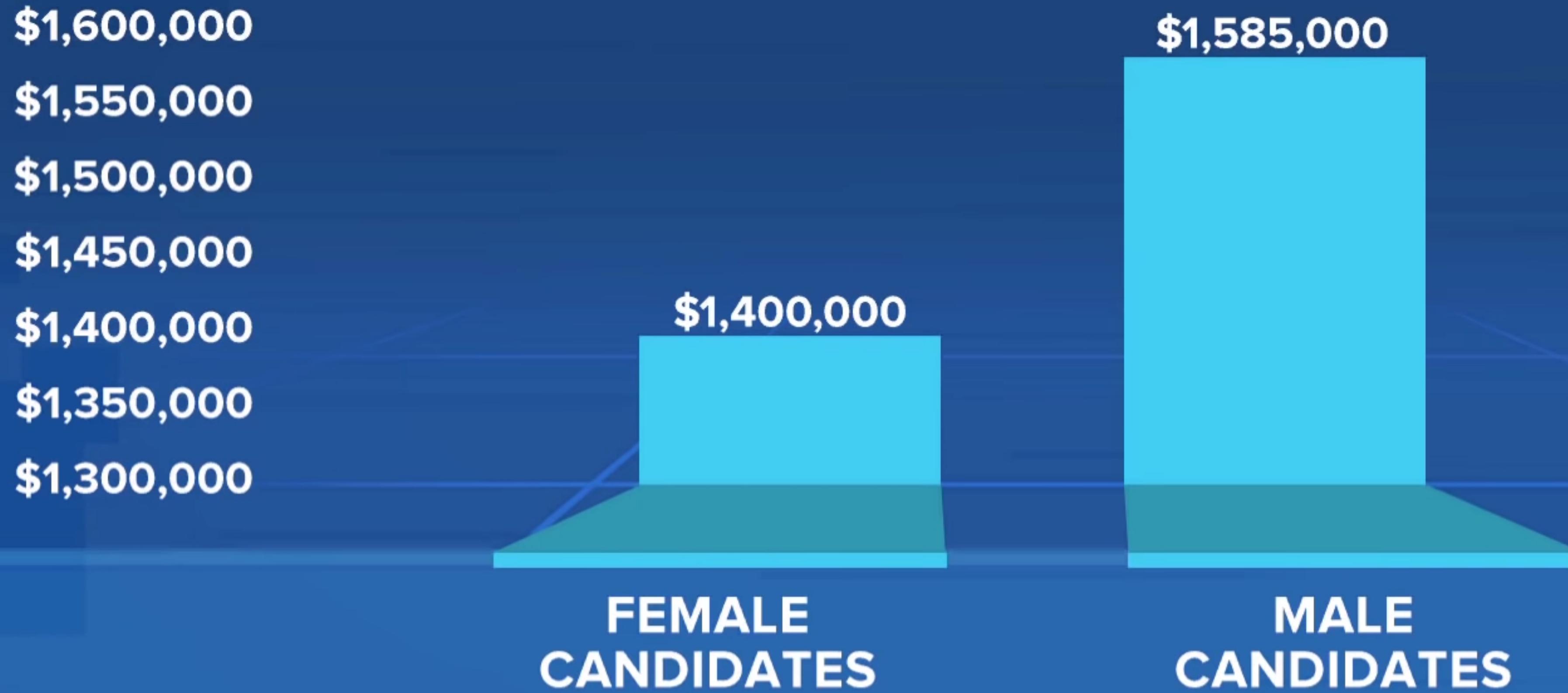
NAS



6.61

# DEMOCRATIC WOMEN RAISE LESS MONEY THAN DEMOCRATIC MEN

## AVERAGE RAISED BY HOUSE PRIMARY WINNERS



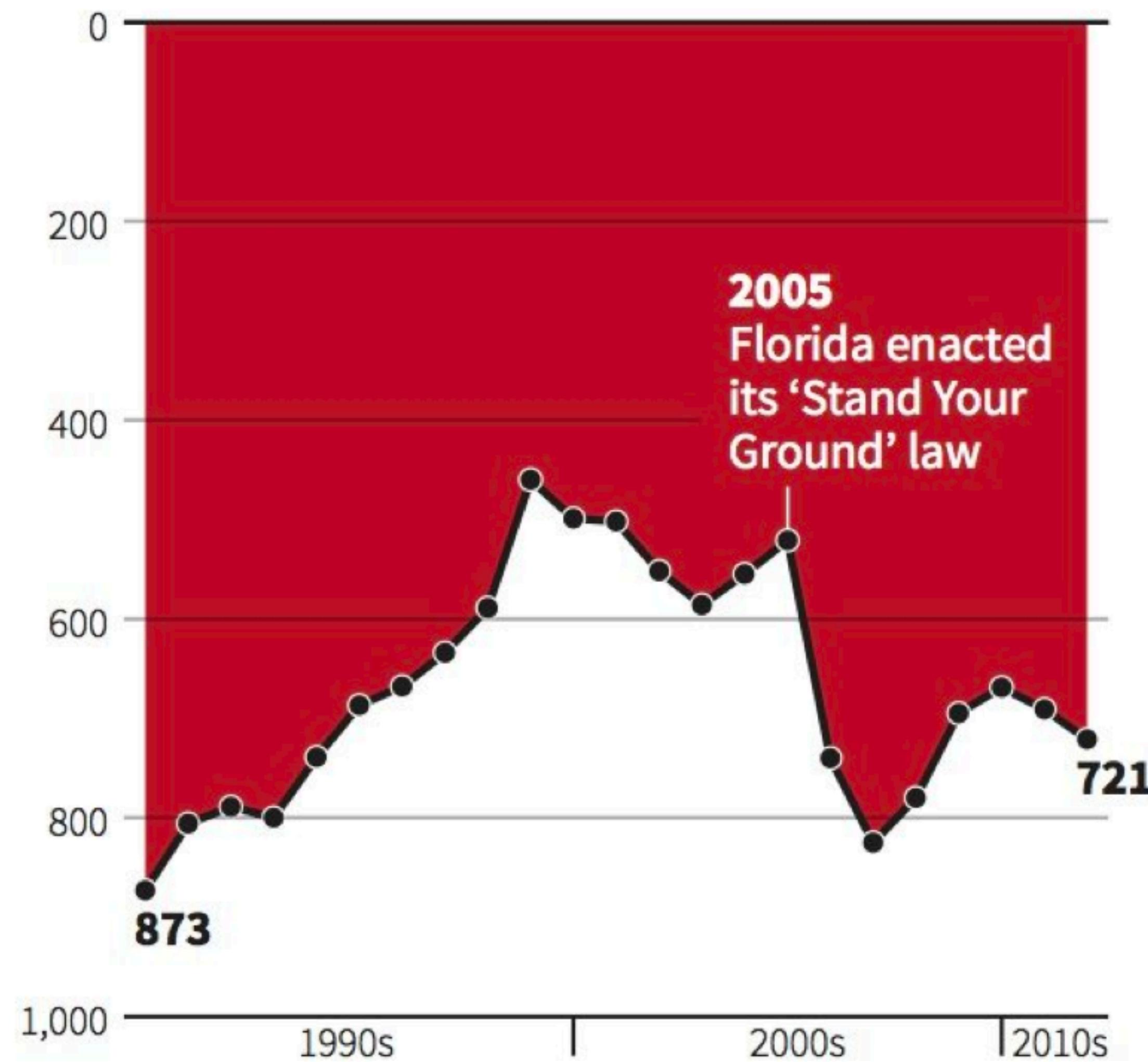
SOURCE: CENTER FOR RESPONSIVE POLITICS

# SHOULD BRITAIN LEAVE EU?



# Gun deaths in Florida

Number of murders committed using firearms



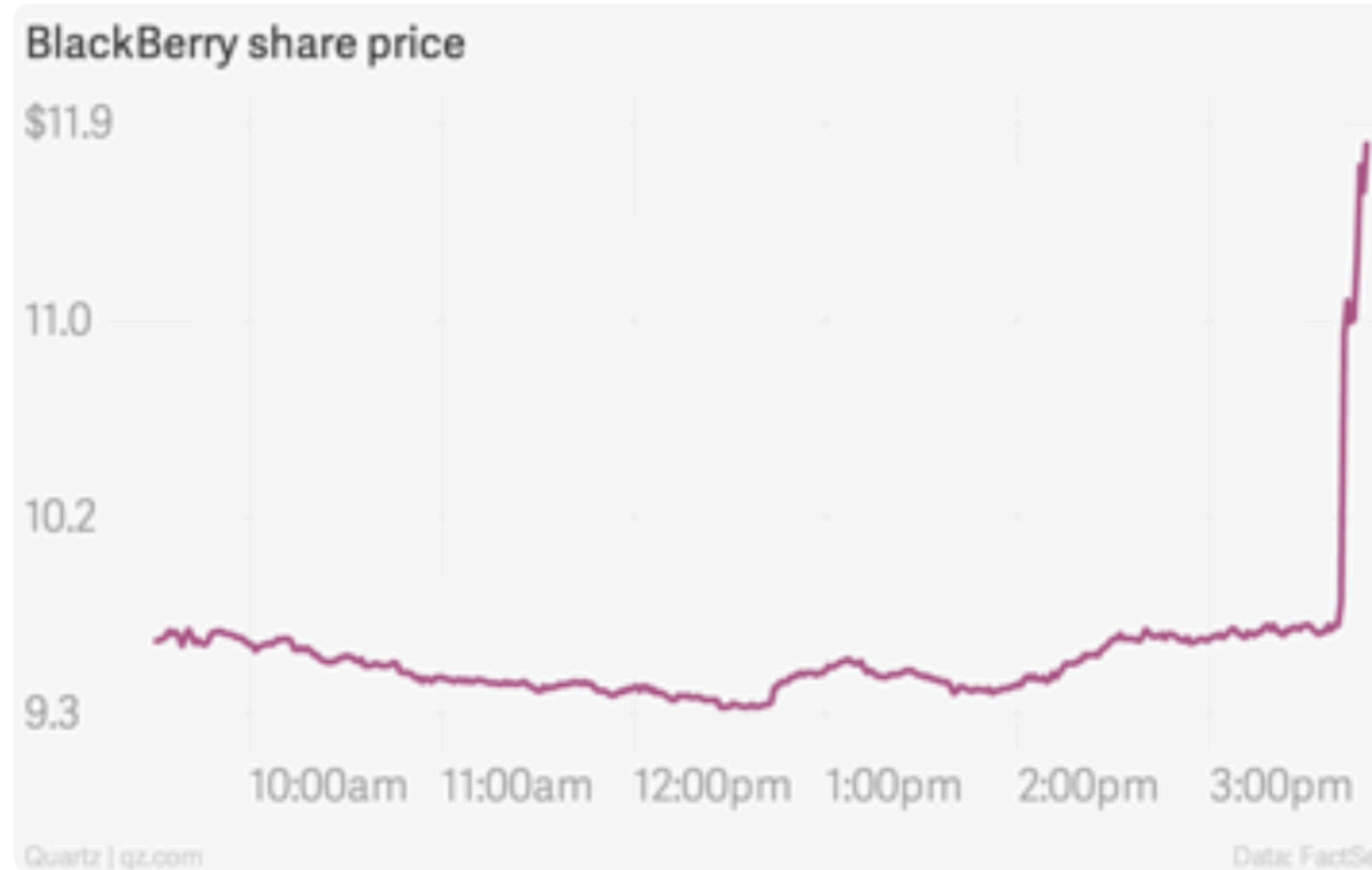
Source: Florida Department of Law Enforcement

**AND YET...**



Quartz @qz

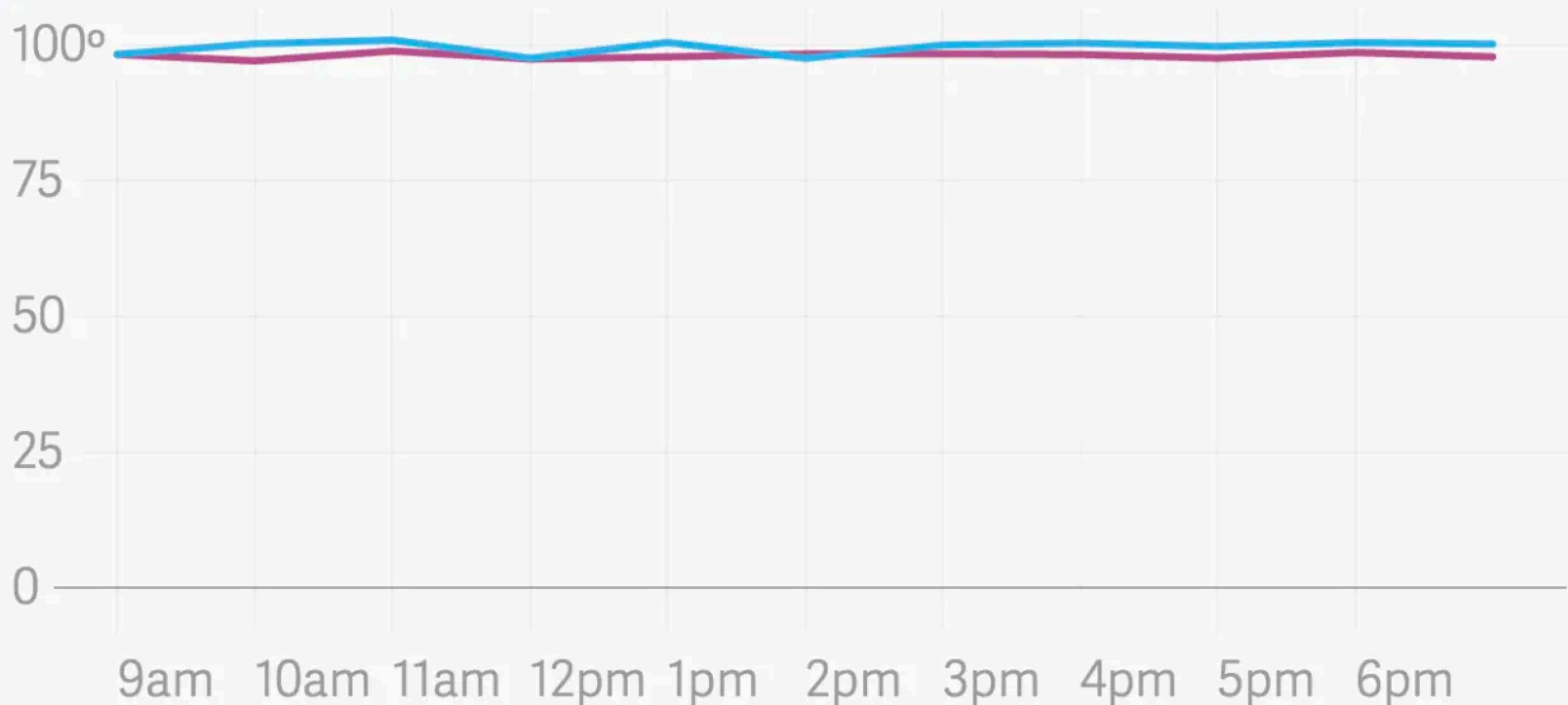
Acquisition rumors light a fire under BlackBerry's stock.  
[qz.com/note/326942/](http://qz.com/note/326942/)



6 4:44 PM - Jan 14, 2015

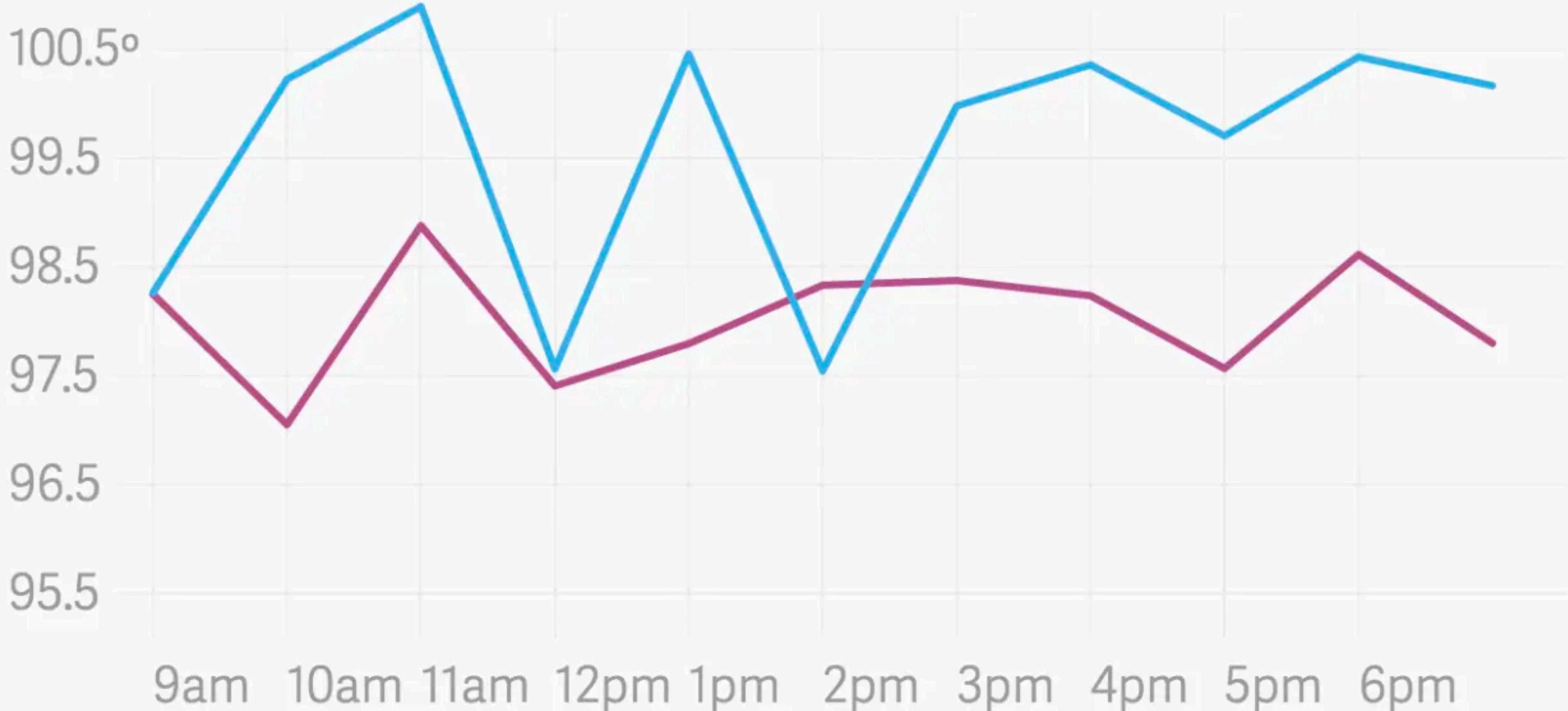
# Oral temperature

Sara Bob



# Oral temperature

Sara Bob



# 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

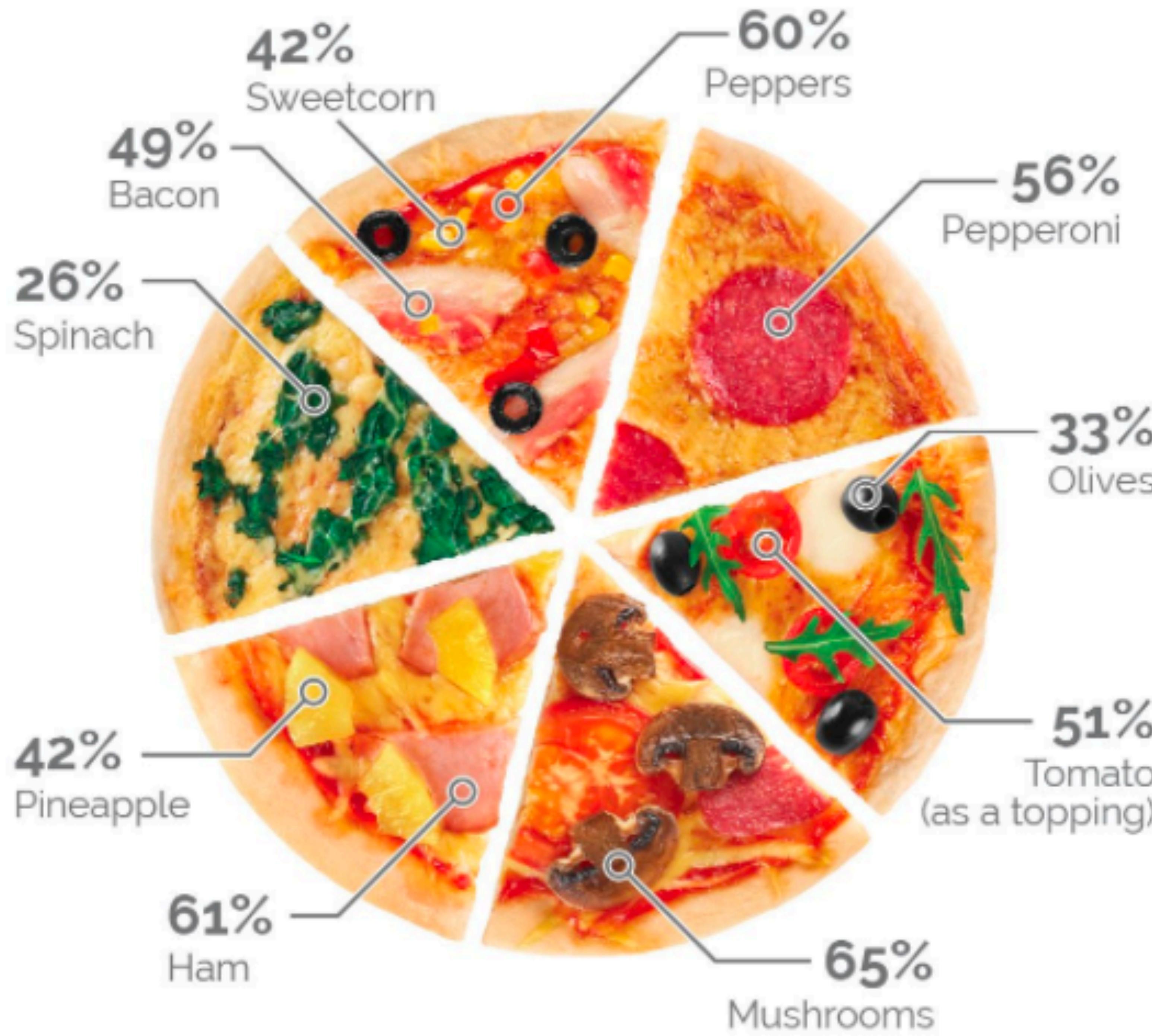
35%

VERY LIKELY

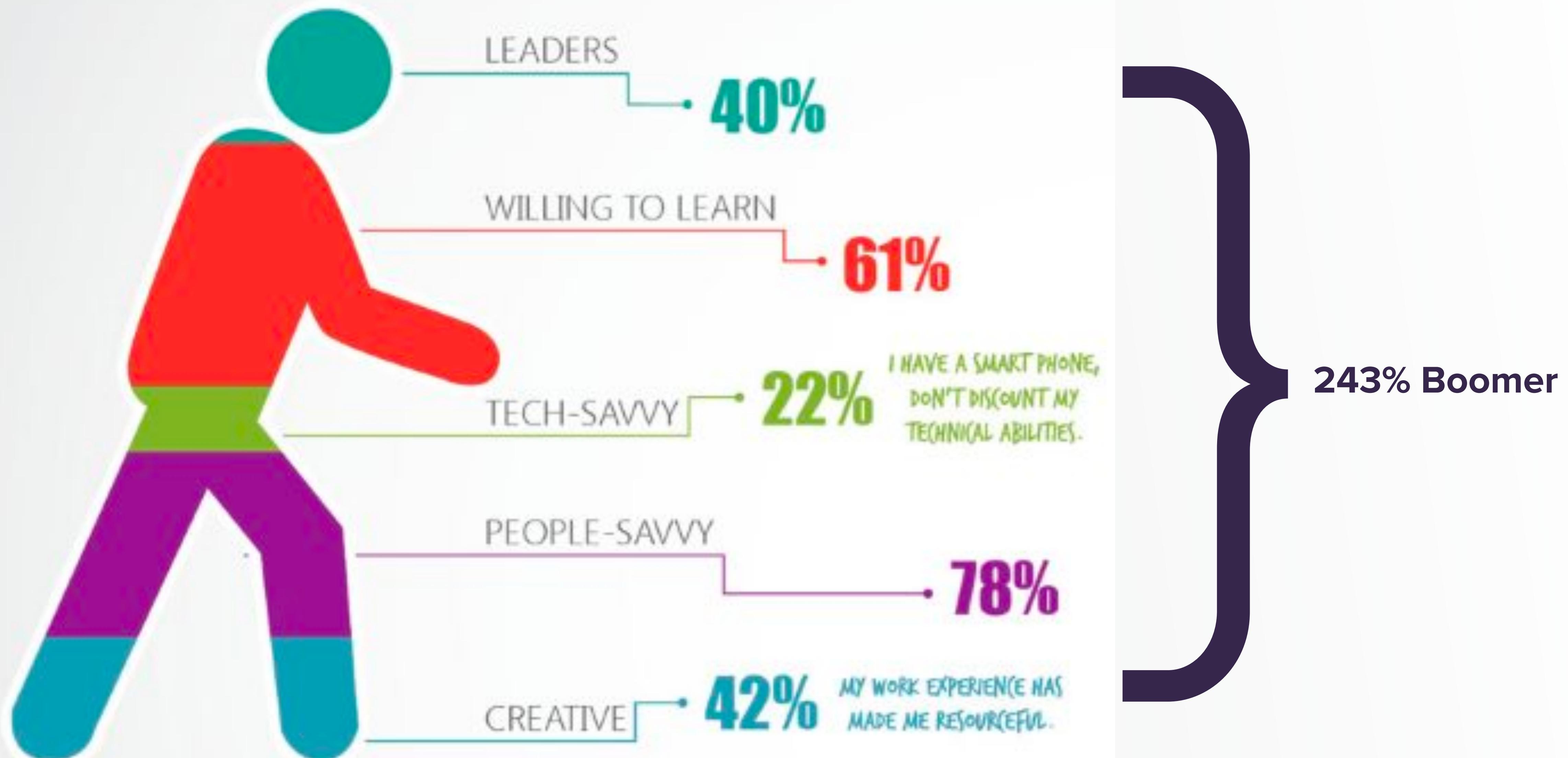
26%

NOT VERY LIKELY





## HOW BABY BOOMERS DESCRIBE THEMSELVES



# **UNEMPLOYMENT RATE UNDER PRESIDENT OBAMA**



# WHAT IS YOUR FAVORITE SEASON? @FCNMIKE

53%

FALL

17%

SUMMER

17%

WINTER

13%

SPRING



75°

6:51

FIRST COAST  
NEWS

#KATIESCOMMUTERS

DEADLY CRASH

DEADLY HOME INVASION

DAILY FORECAST

LOW: 64° / HIGH: 75°

LOW: 64° / HIGH: 75°

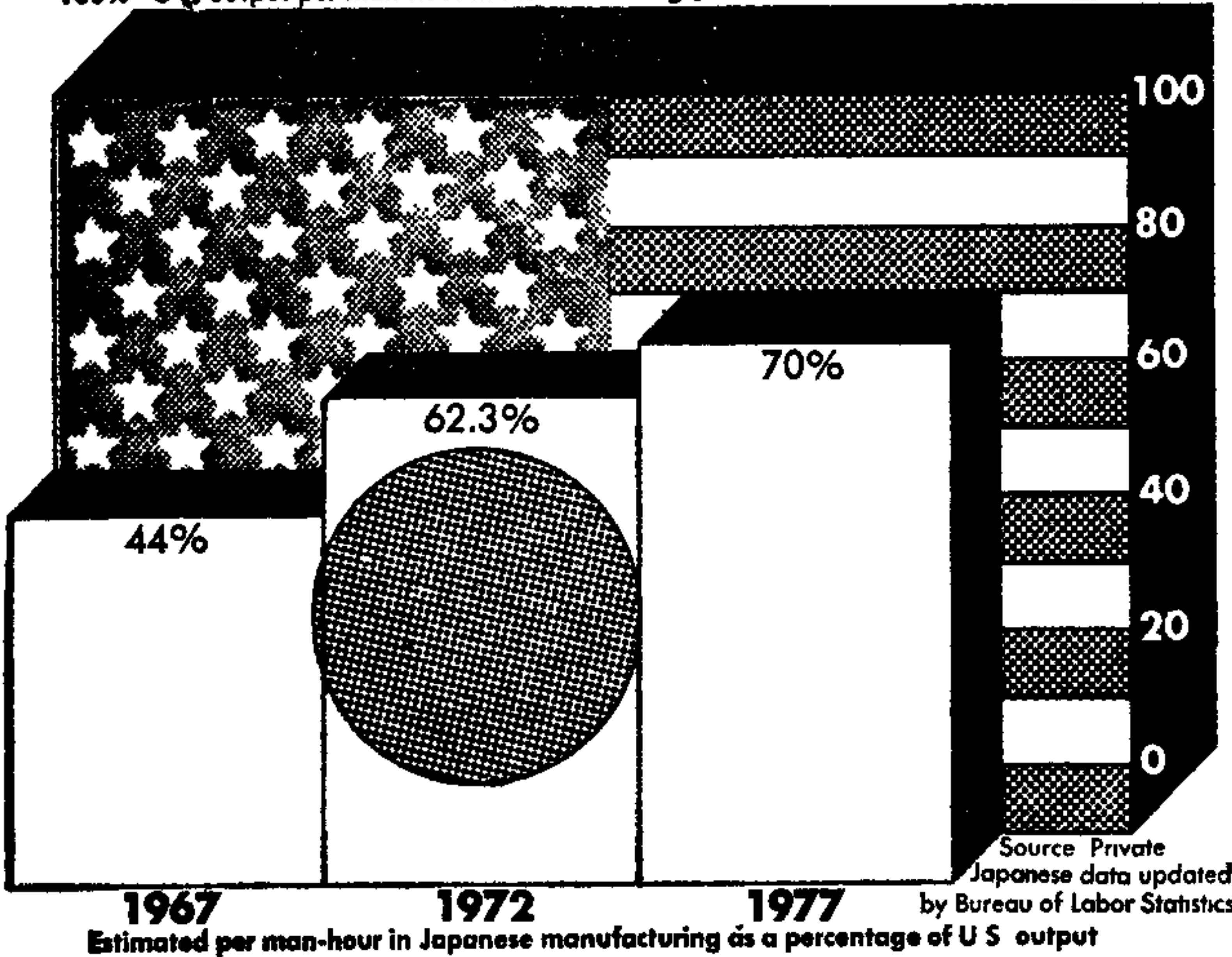
LOW: 64° / HIGH: 75°

# 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

100% = U.S. output per man-hour in manufacturing

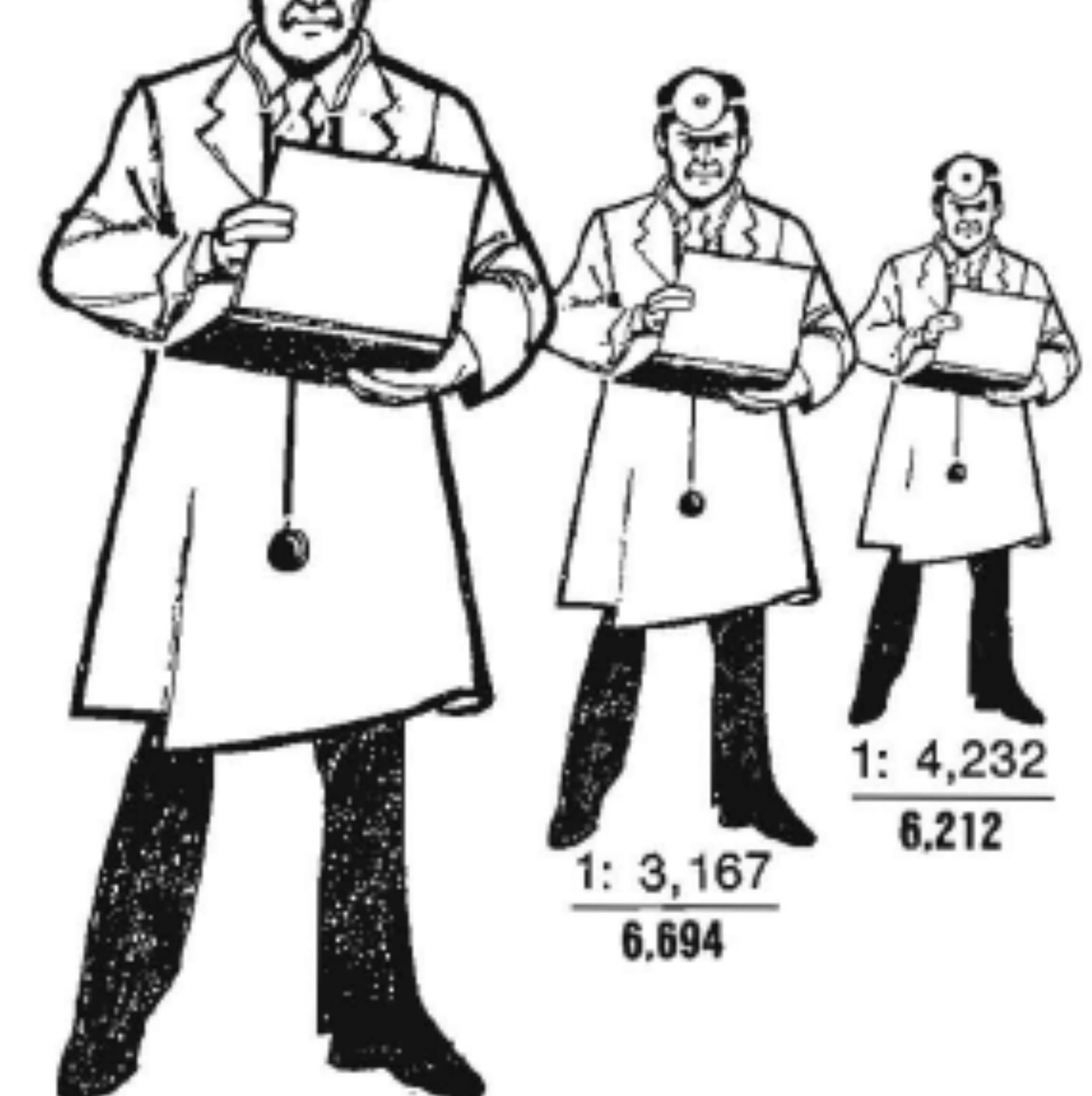


# THE SHRINKING FAMILY DOCTOR

## In California

Percentage of Doctors Devoted Solely to Family Practice

1964	1975	1990
27%	16.0%	12.0%



# TRUMP TWEETS MENTIONING MUELLER BY NAME

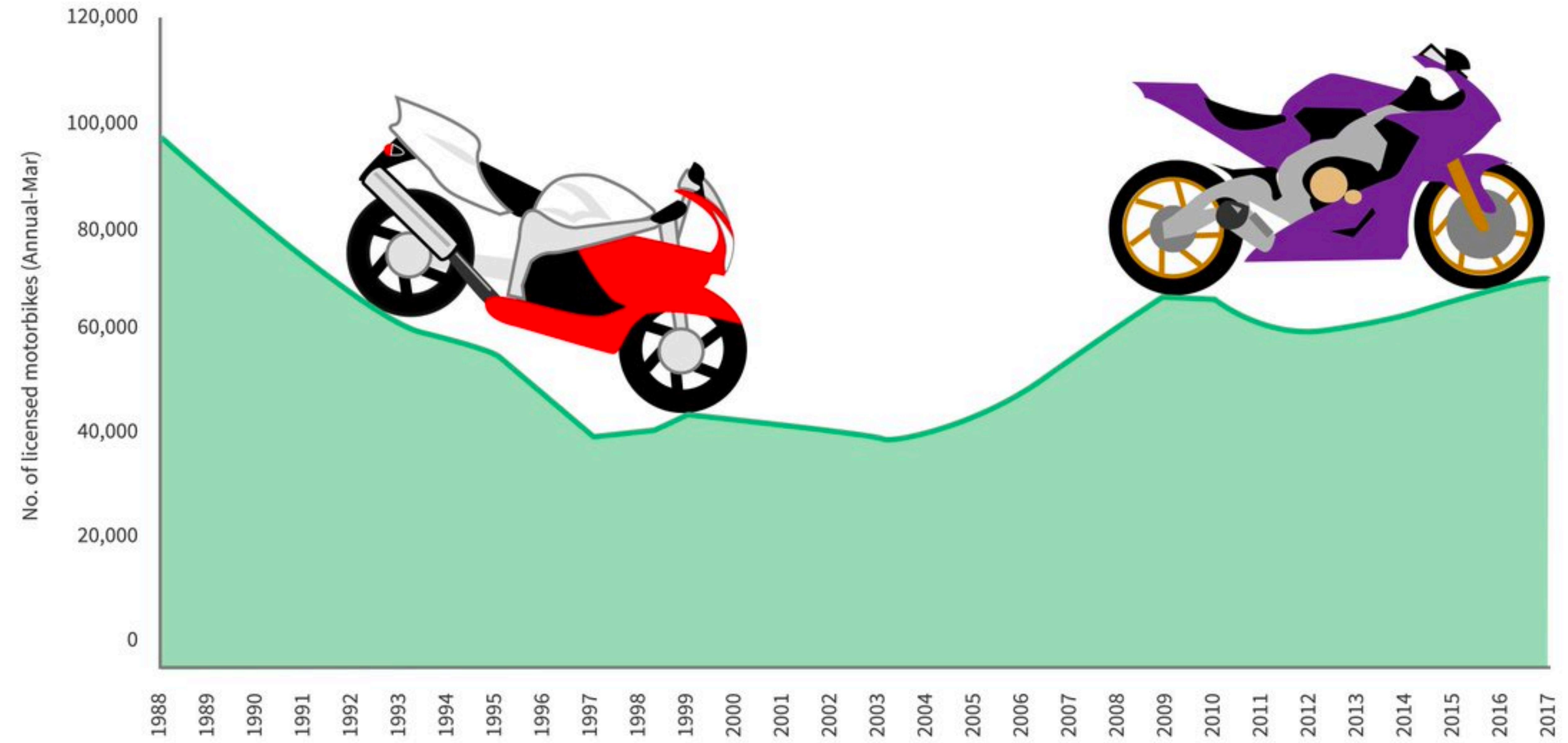


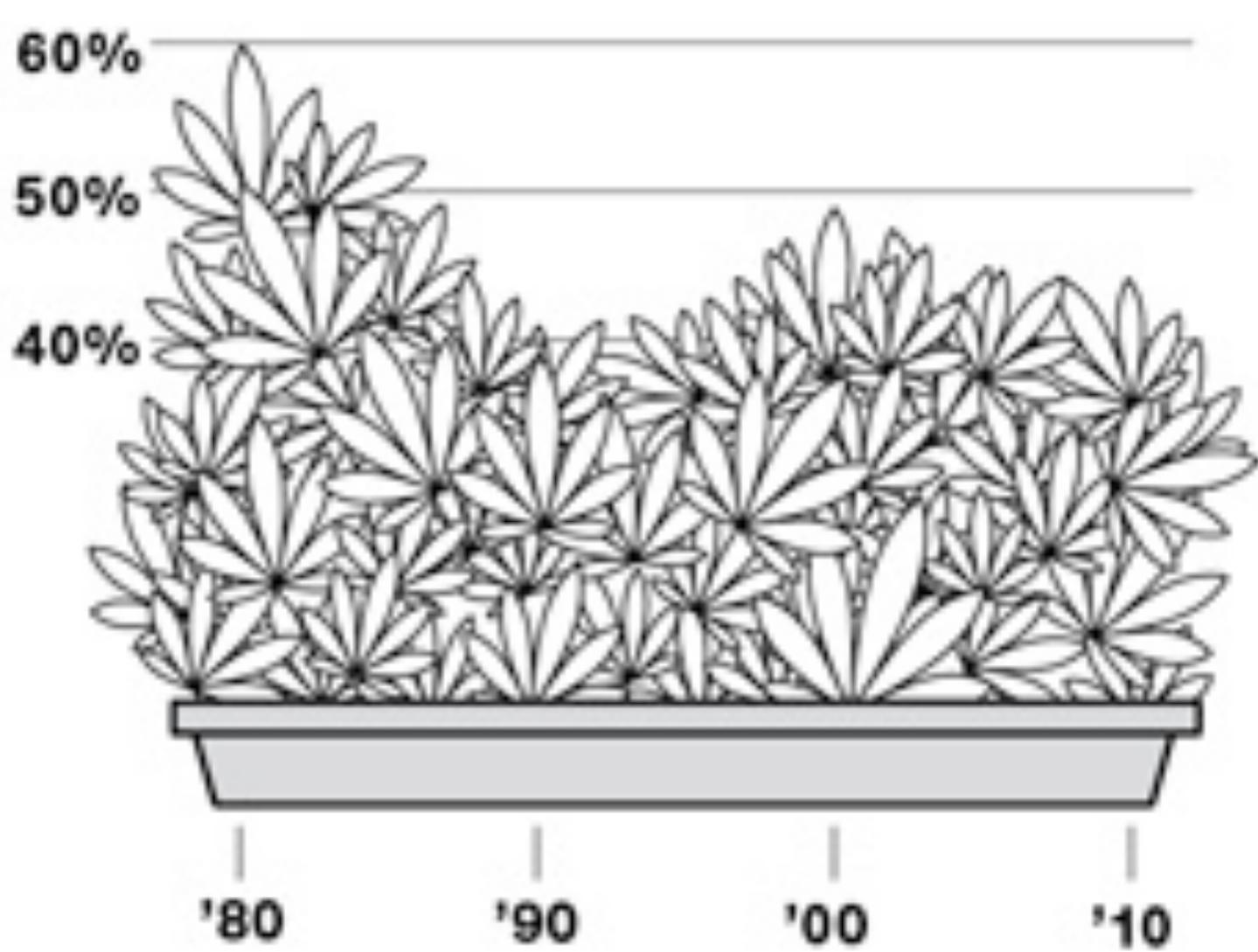
LIVE

 MSNBC

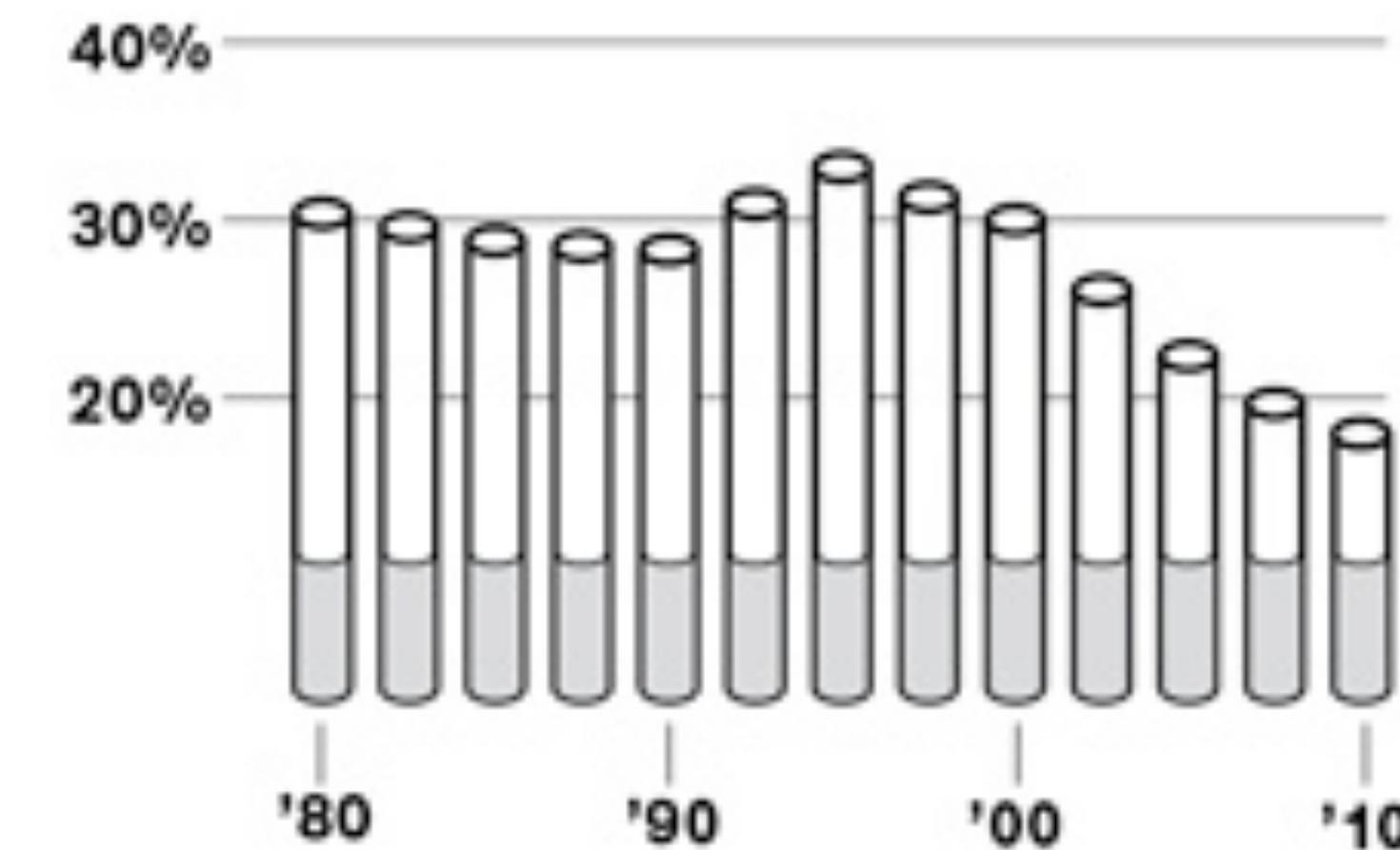
# PLOTTING PITFALLS

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

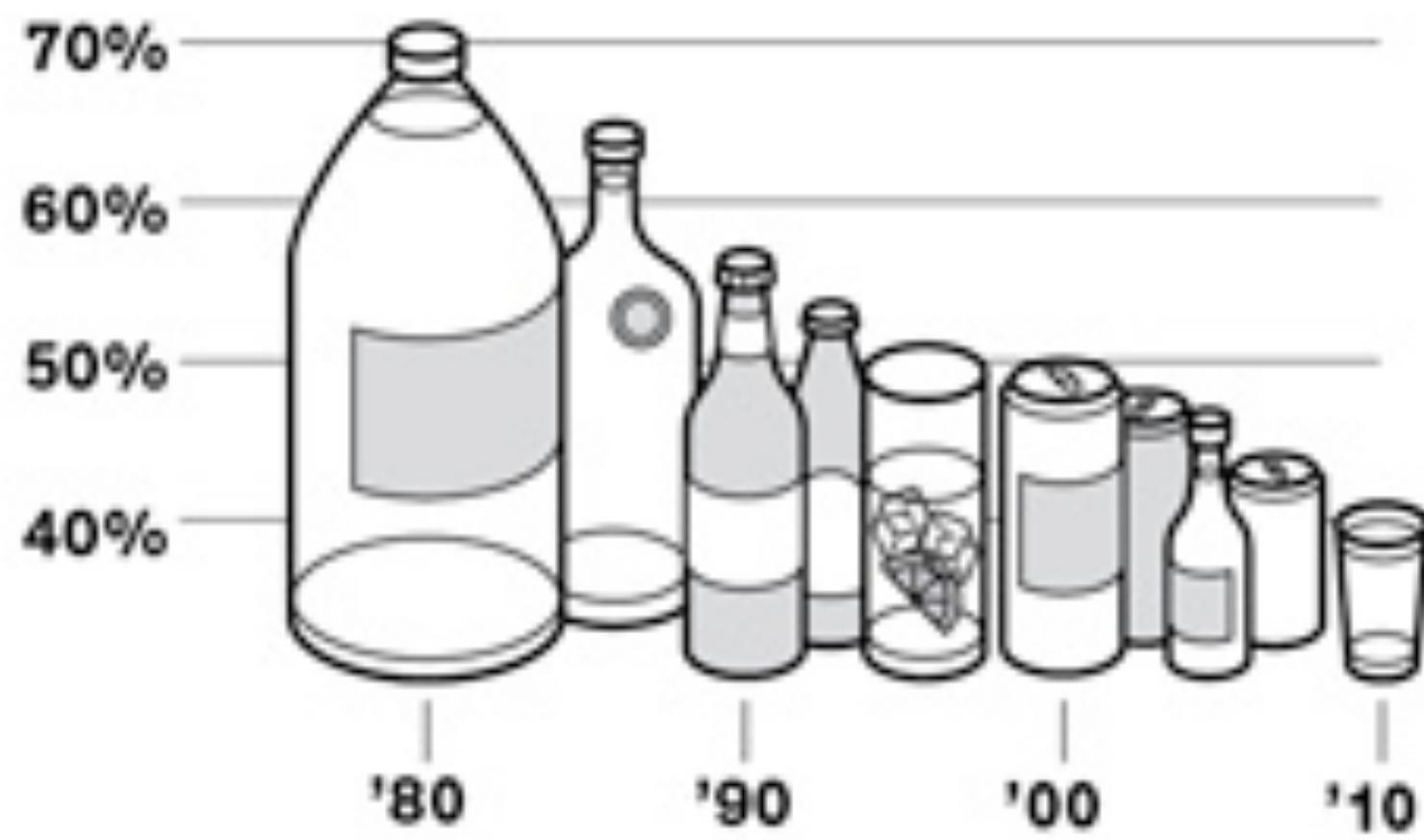




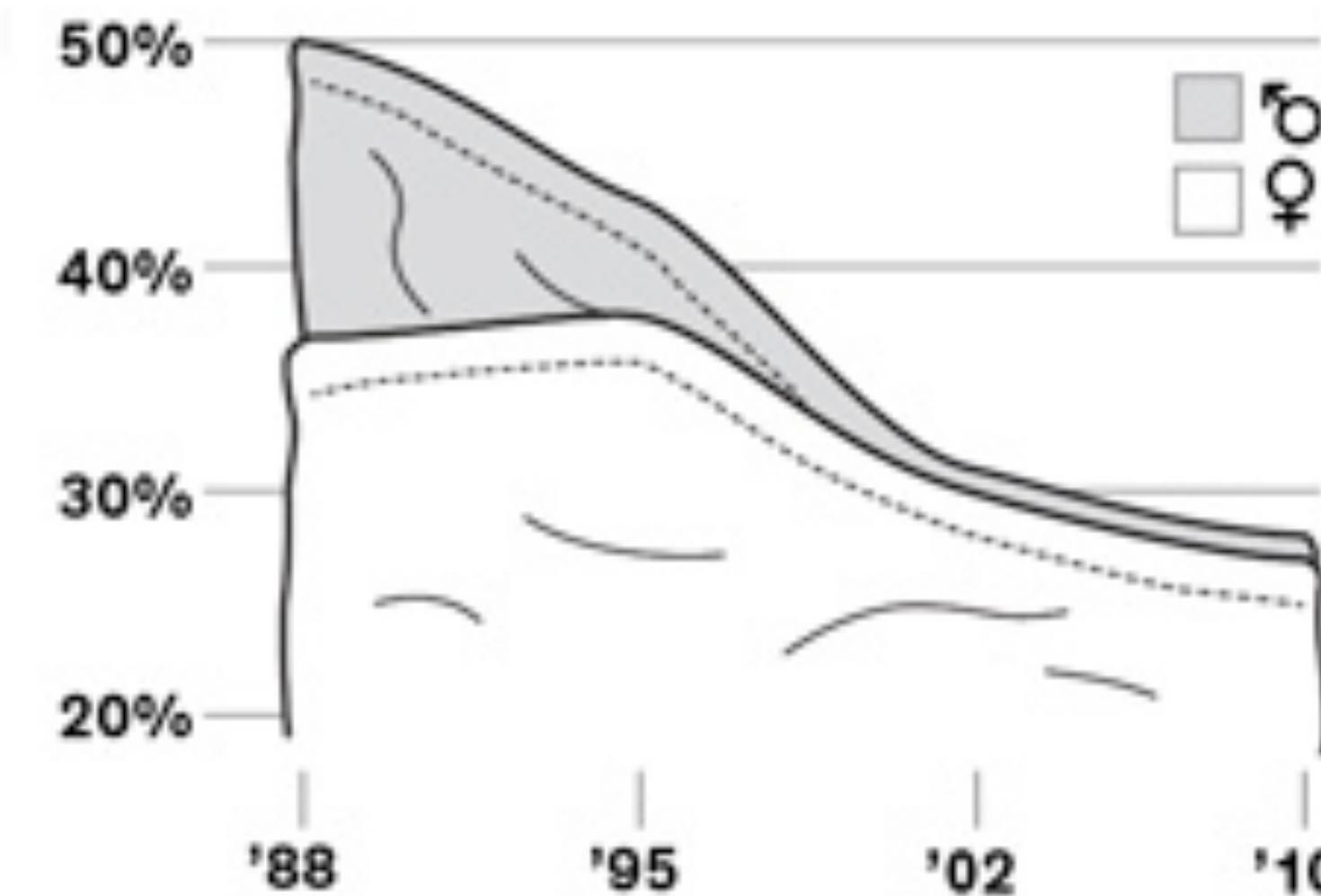
▲ Percentage of high-school seniors  
who have ever tried pot.



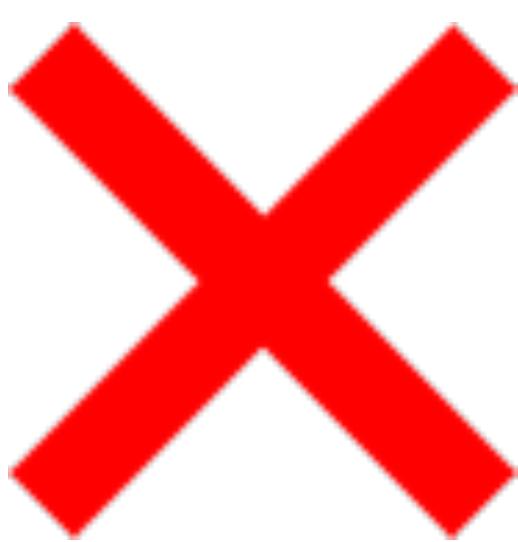
▣ Smoking habits of high-school  
seniors over time.



○ Alcohol use by high-school  
seniors over time.

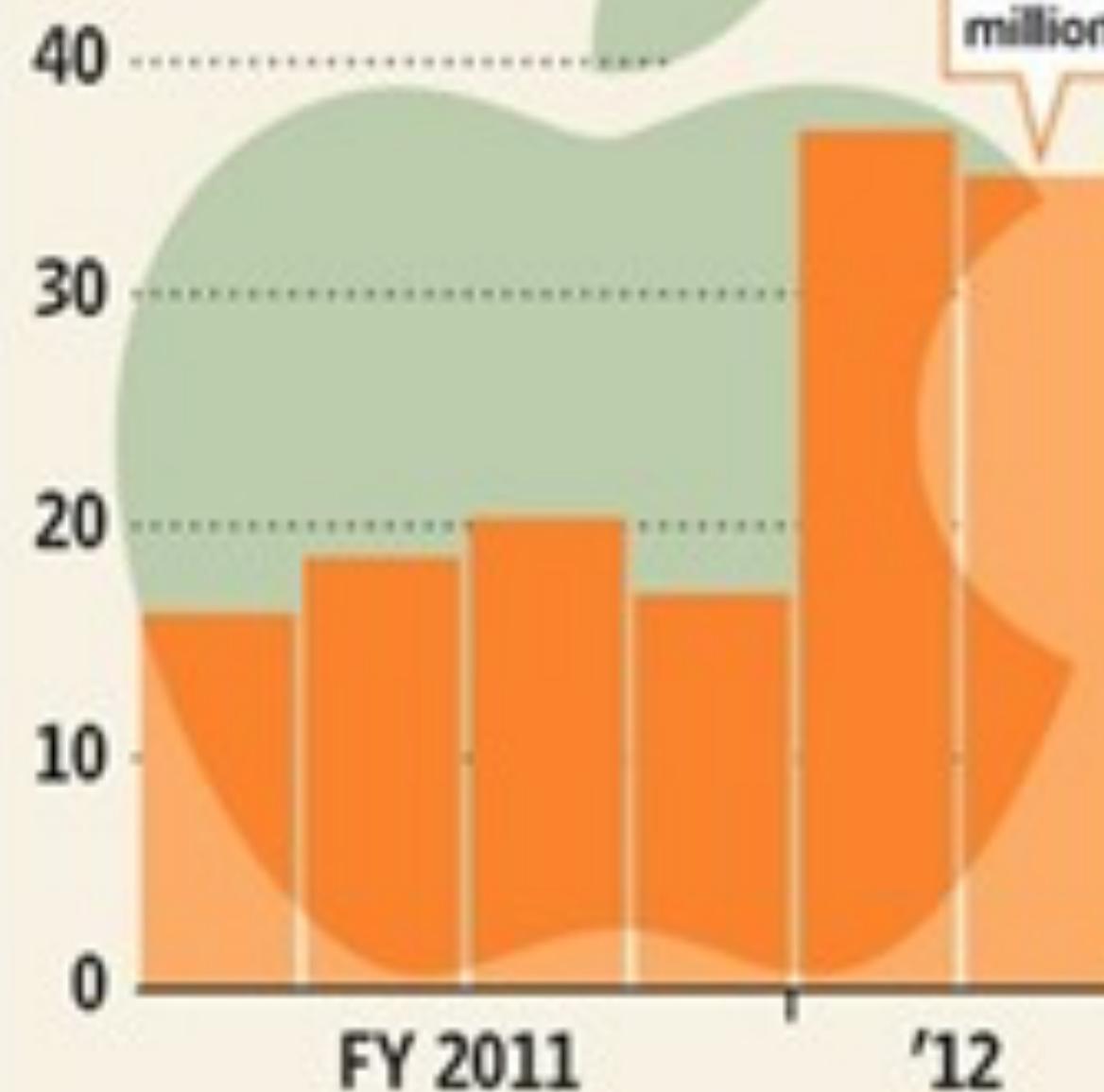


♂ Percentage of 15-to-17-year-olds  
who have had sex.



## Juicy Sales

Quarterly iPhone  
shipments, in millions



Source: the company

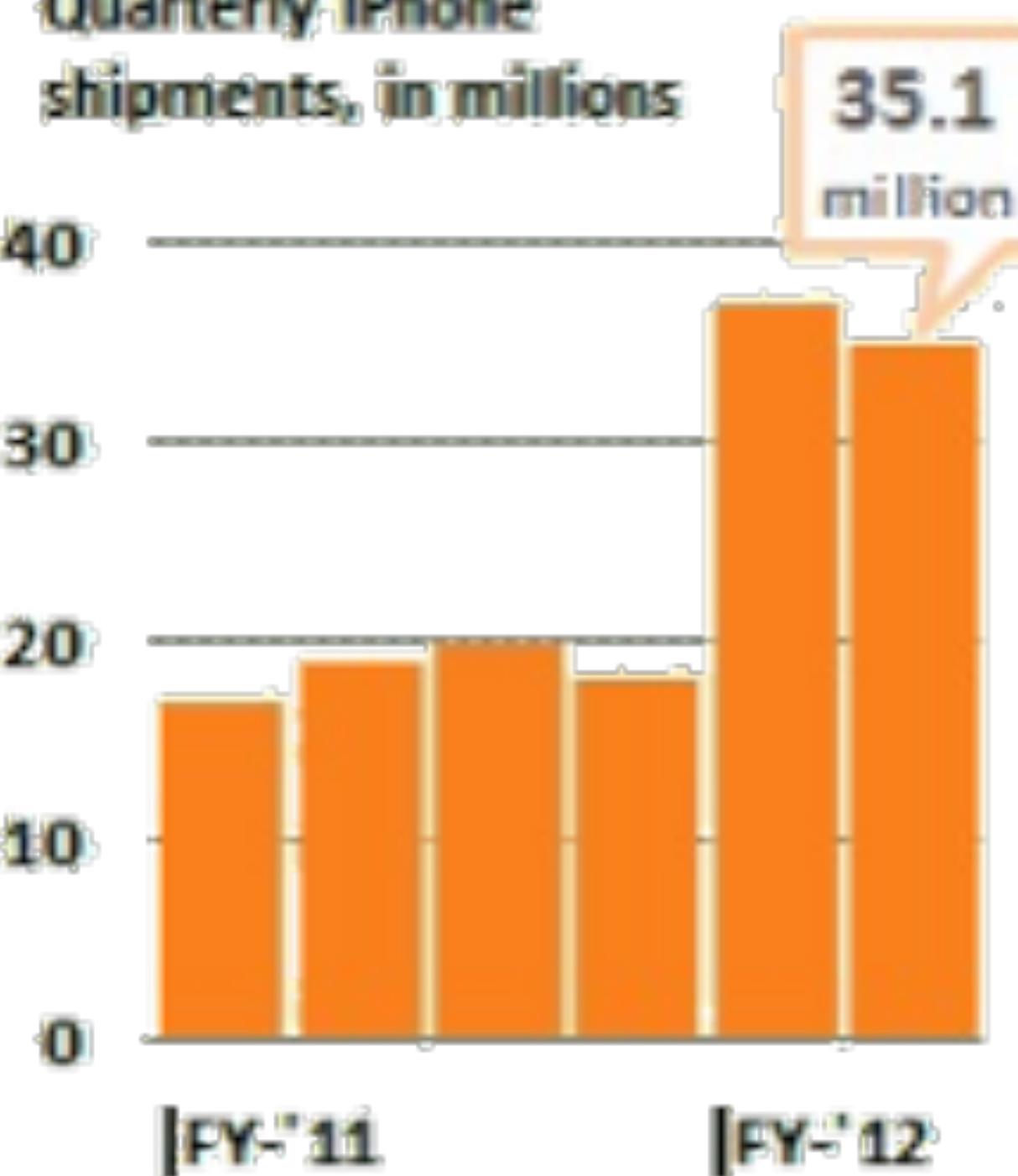
"Be **clear** first and **clever** second. If you have to throw one of those out, throw out **clever**.

— Jason Fried



## Apple's Juicy Sales

Quarterly iPhone  
shipments, in millions

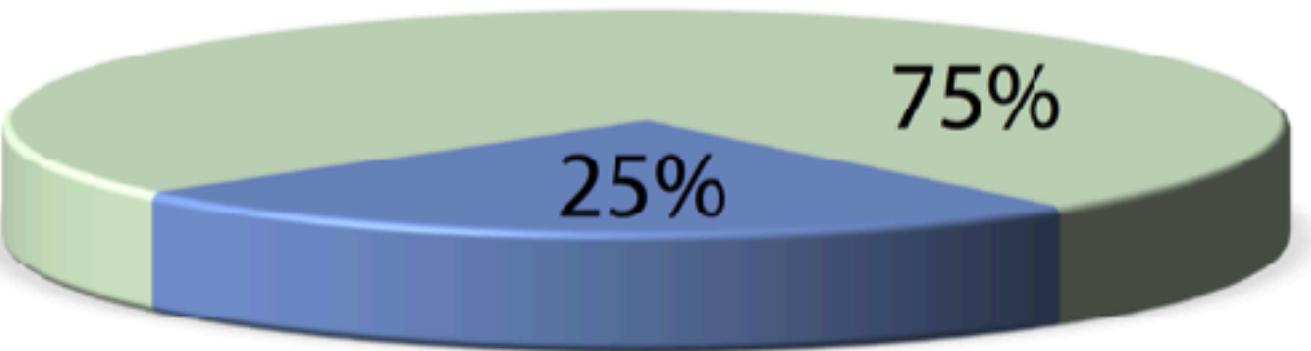


Source: the company

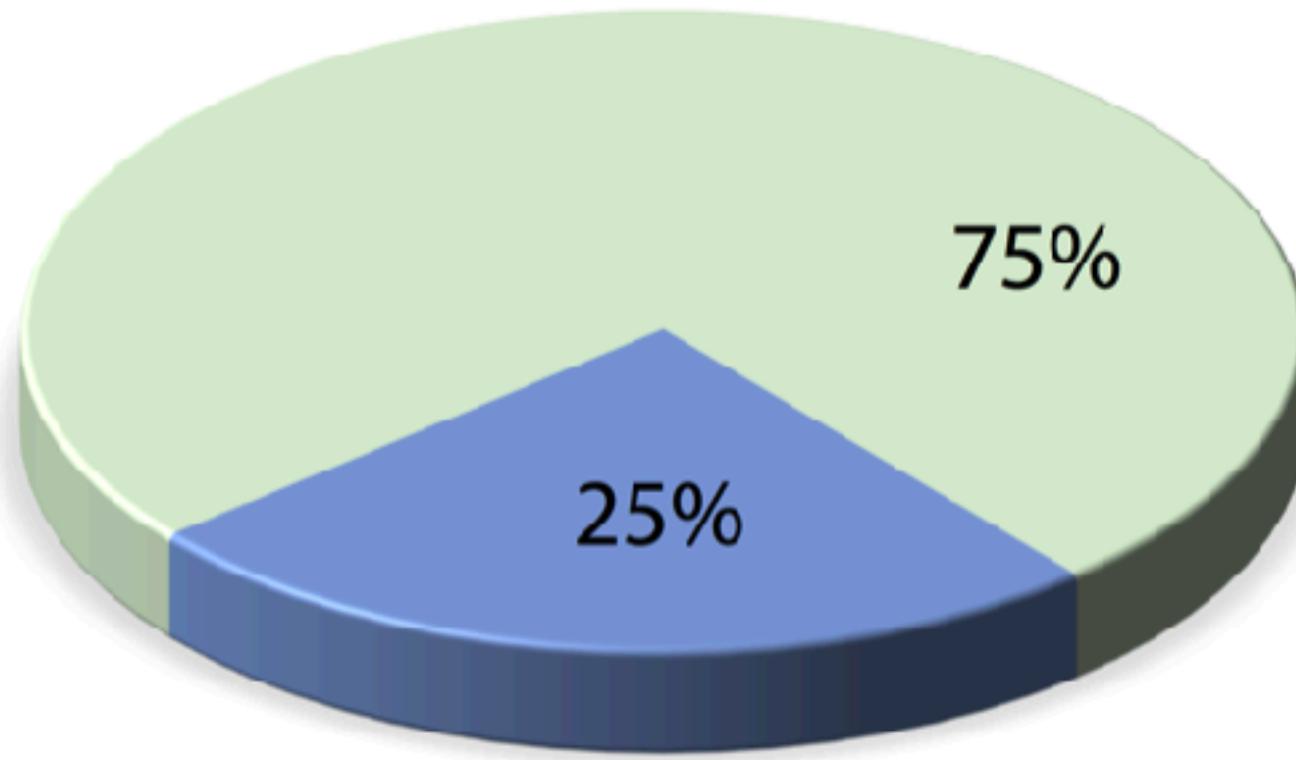
# PLOTTING PITFALLS

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

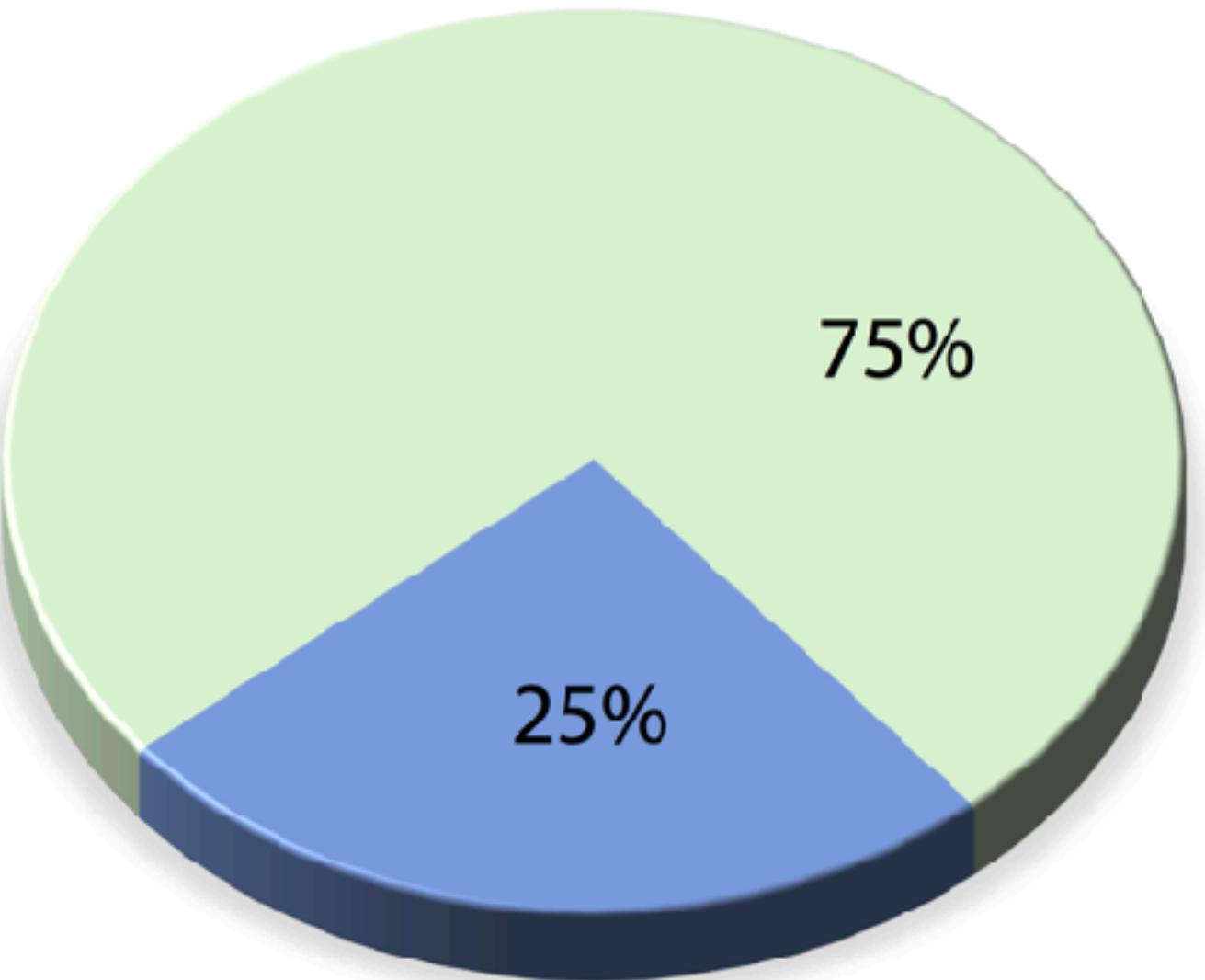
a



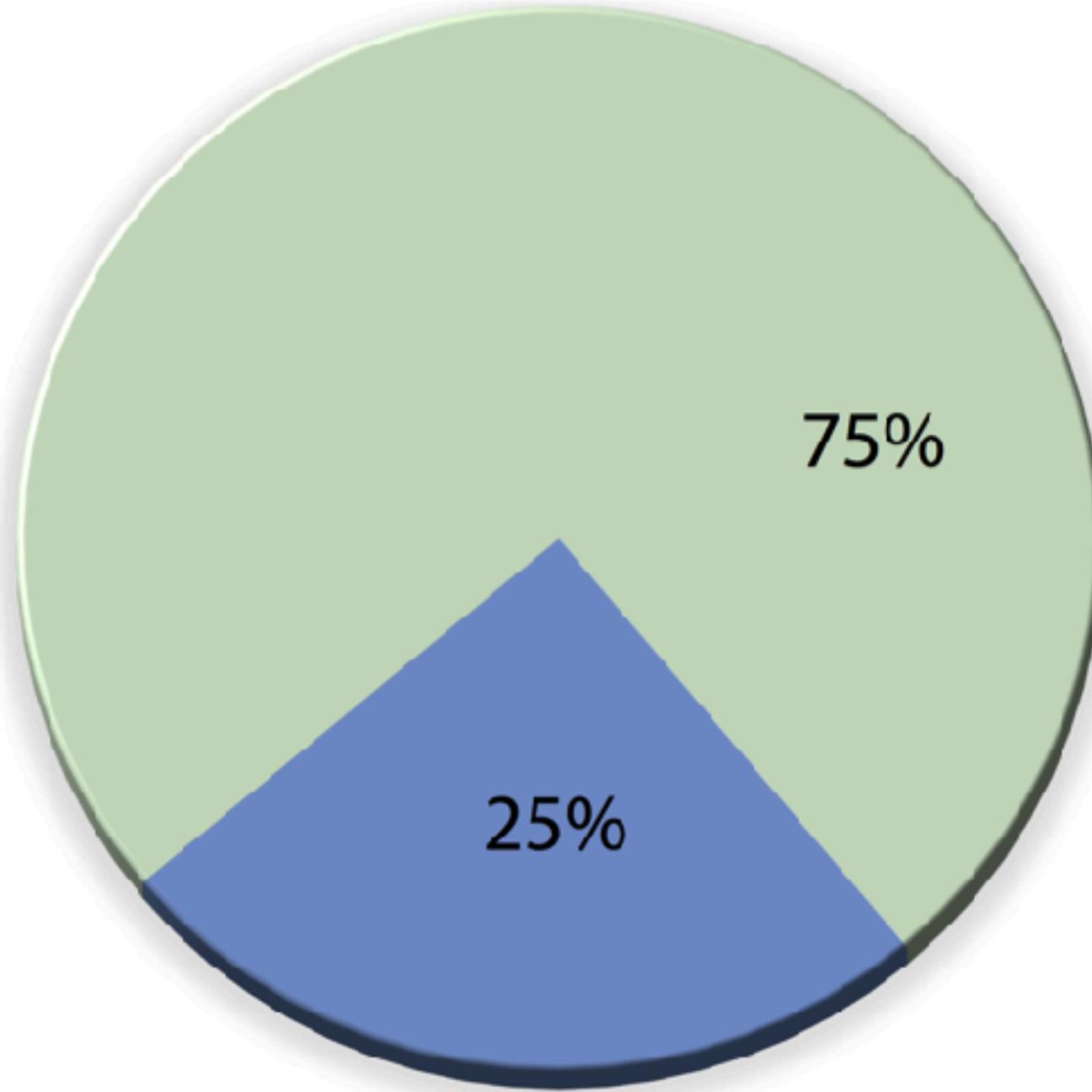
b



c



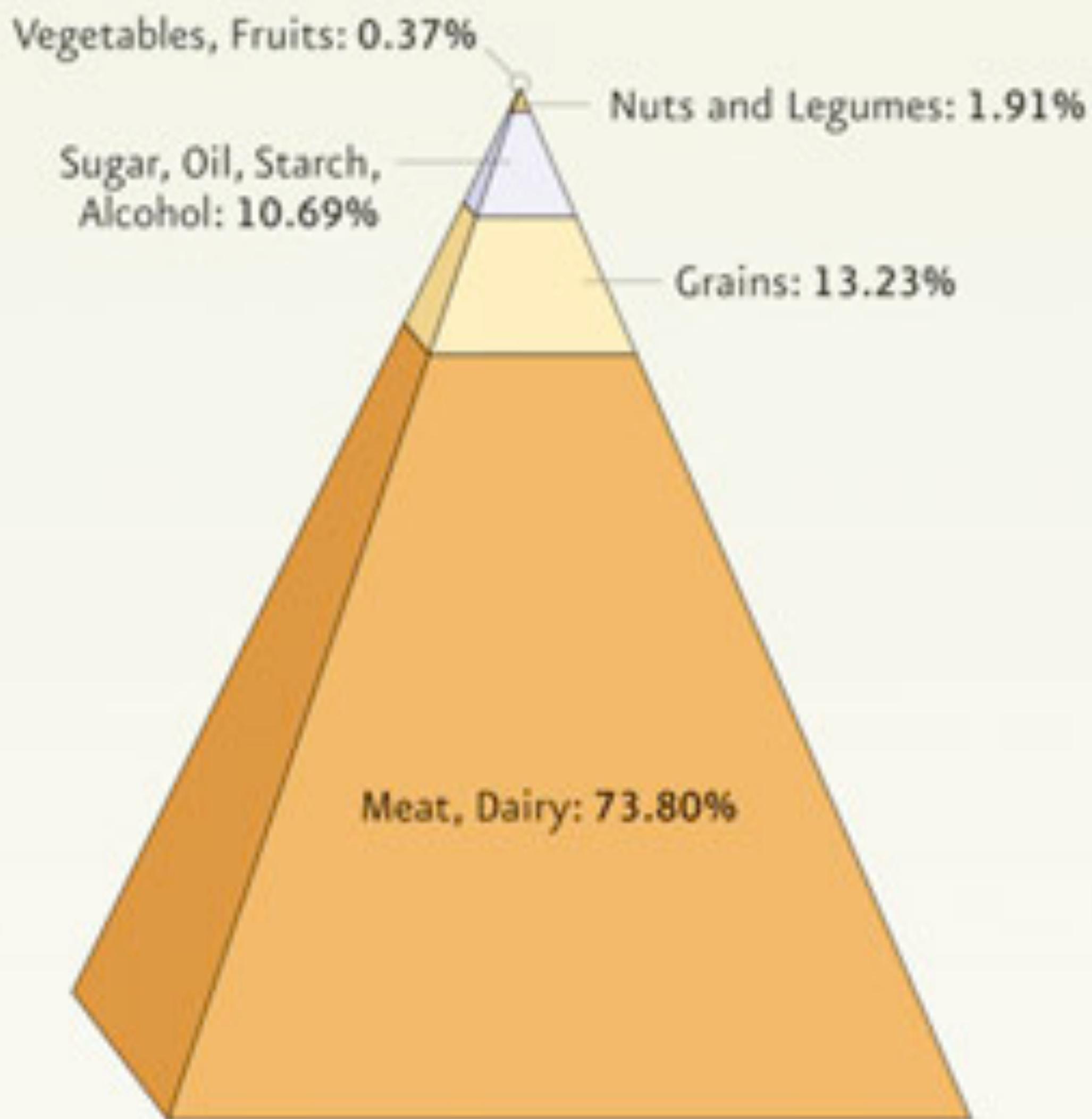
d



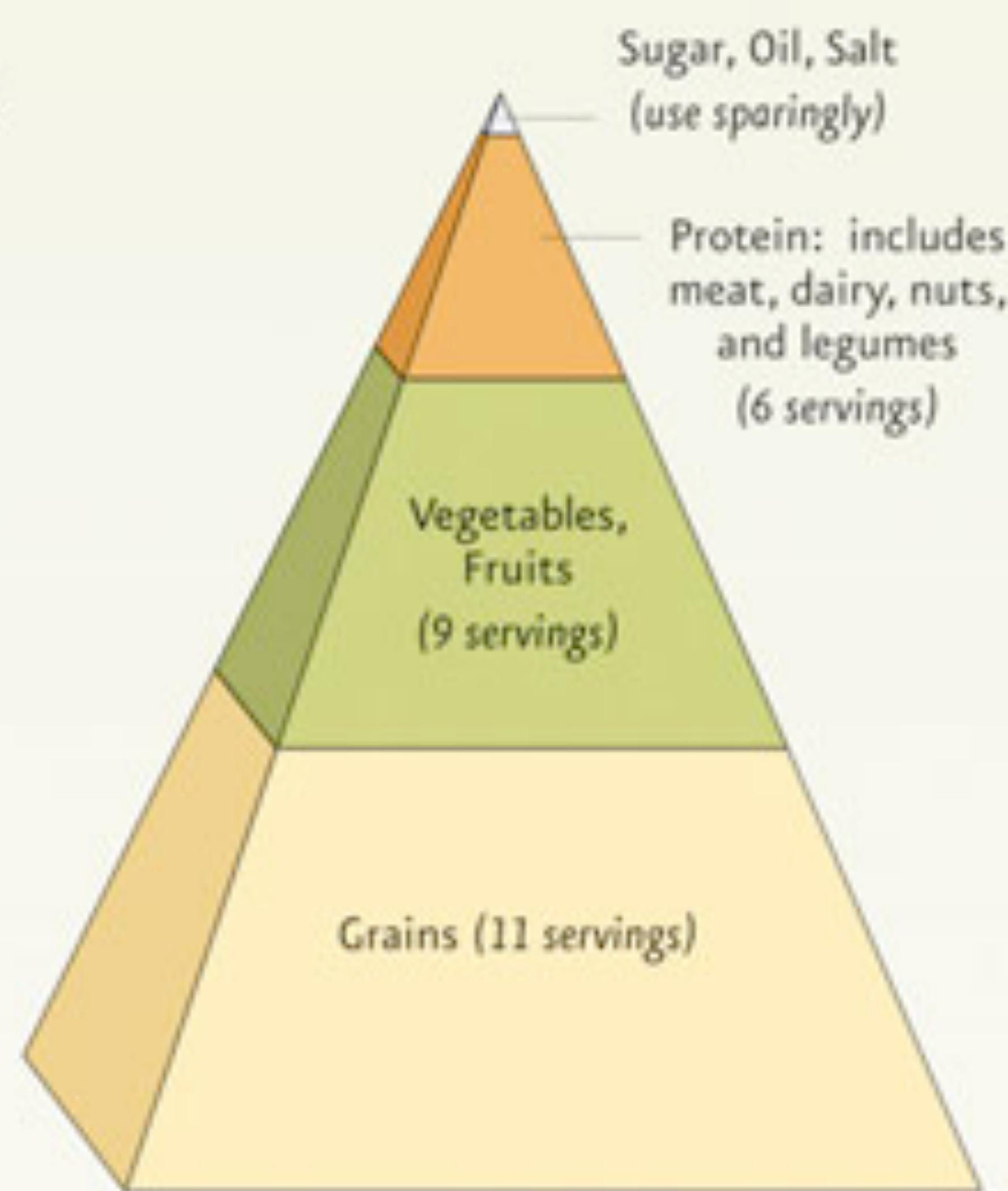
**Same chart, four angles**

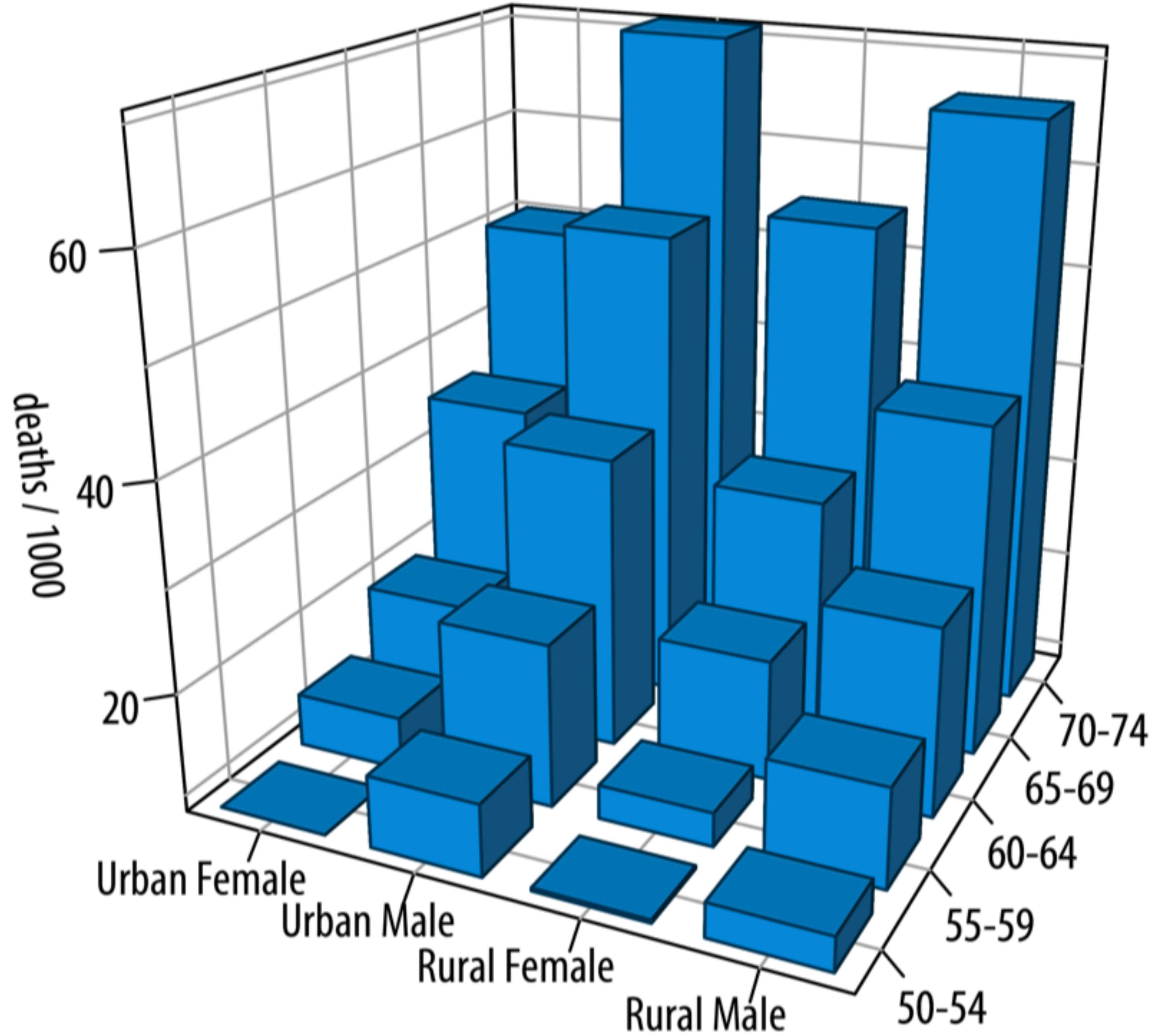
# Why Does a Salad Cost More Than a Big Mac?

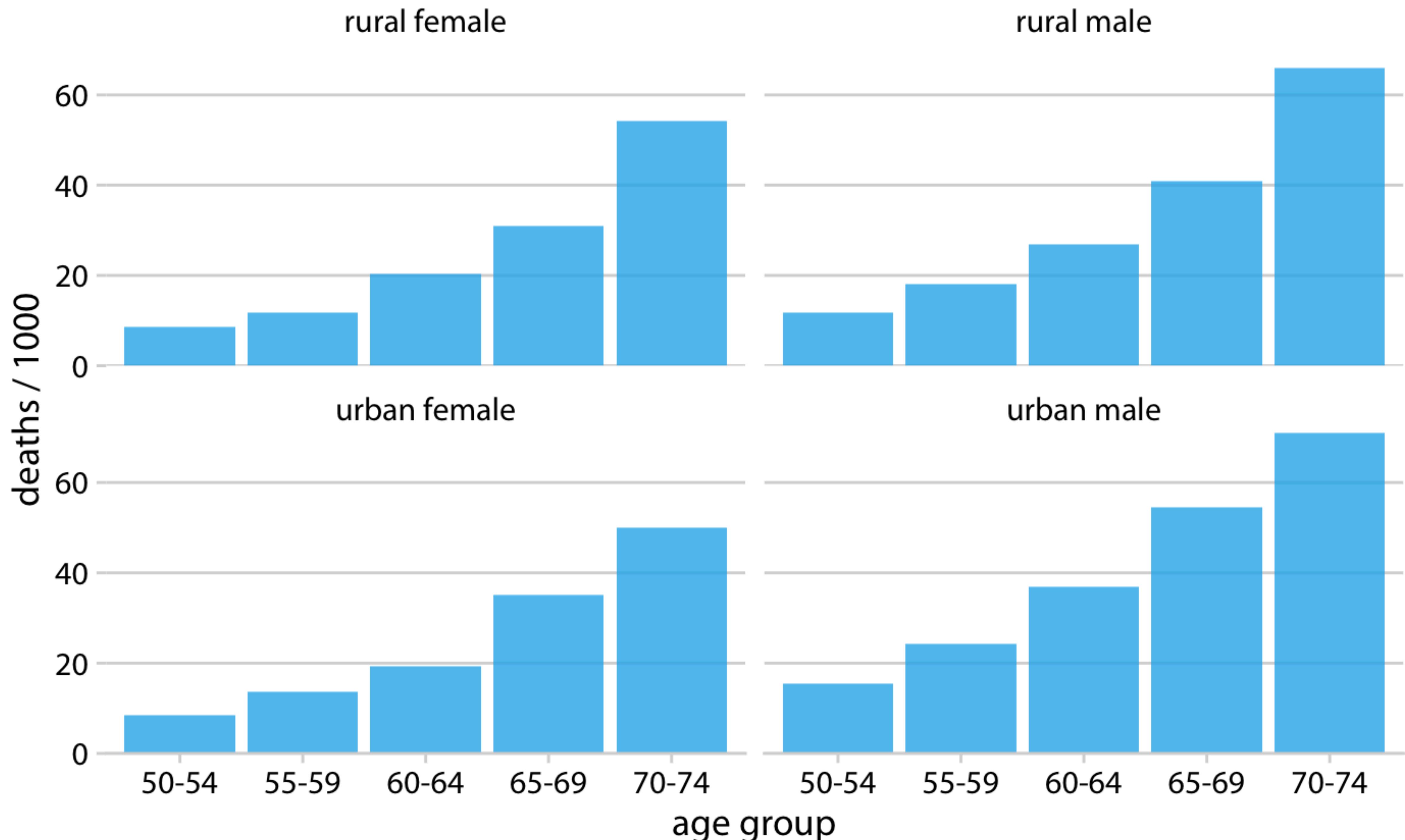
Federal Subsidies for Food Production, 1995-2005<sup>®</sup>



Federal Nutrition Recommendations



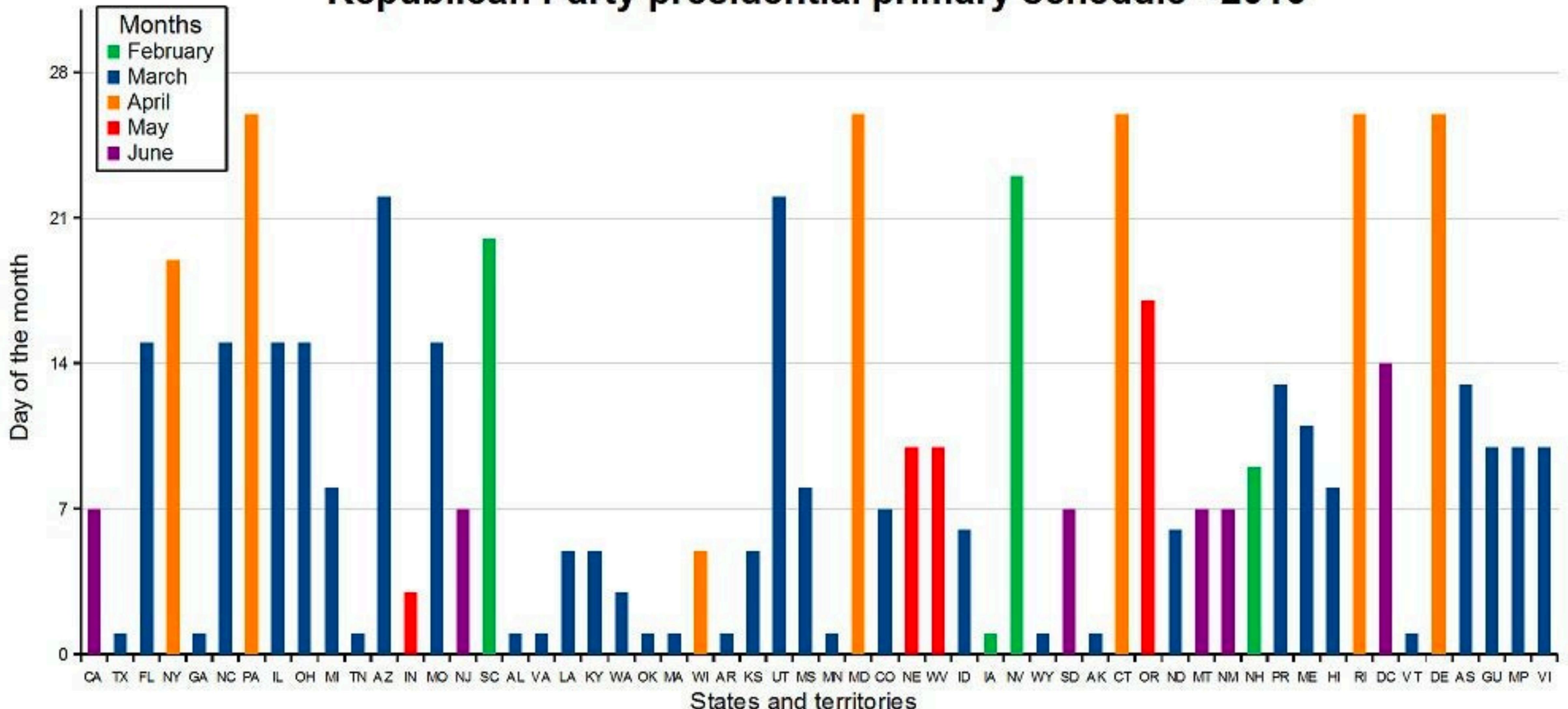




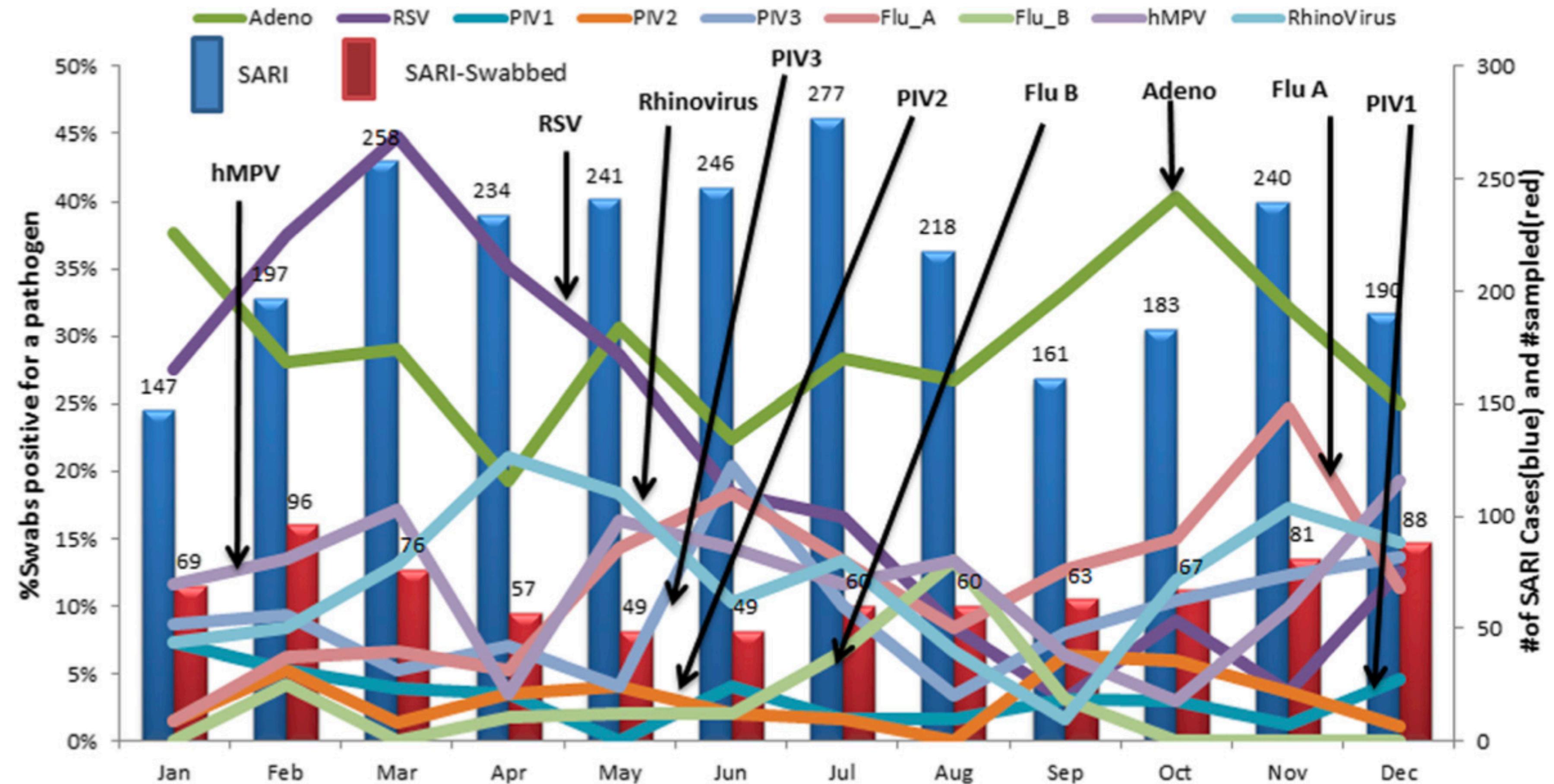
# PLOTTING PITFALLS

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

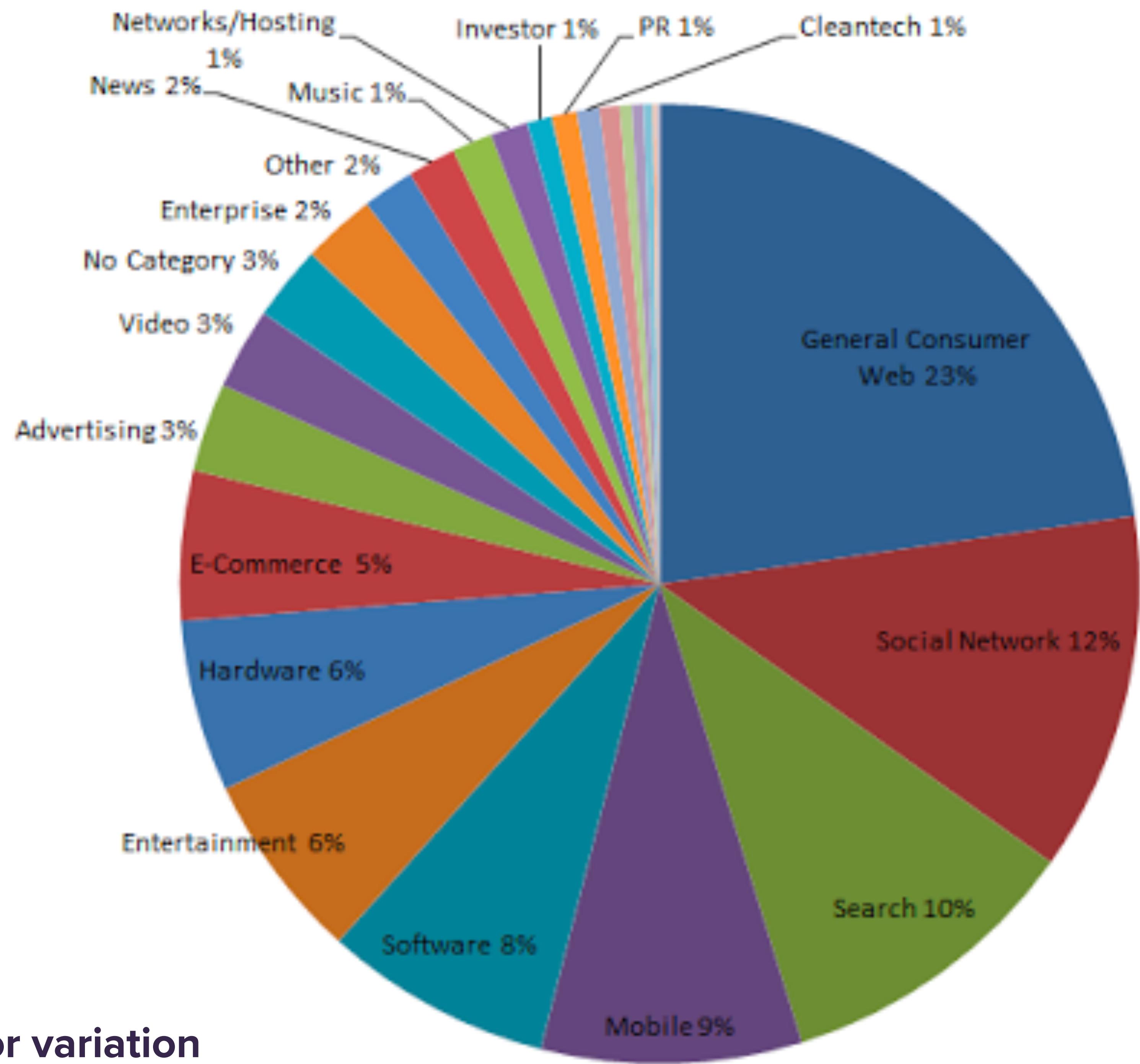
# Republican Party presidential primary schedule - 2016



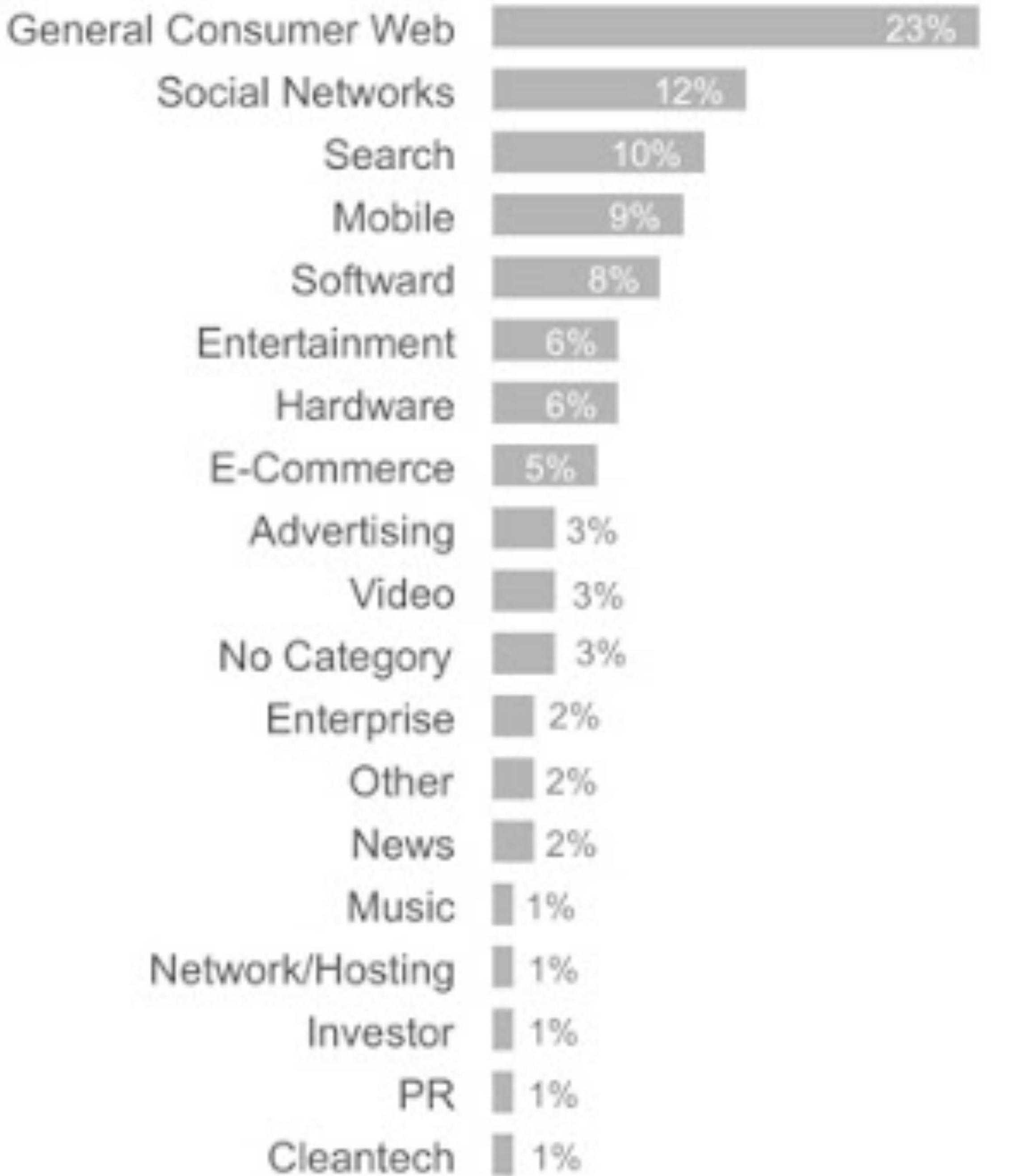
May we suggest... a calendar<sup>35</sup>



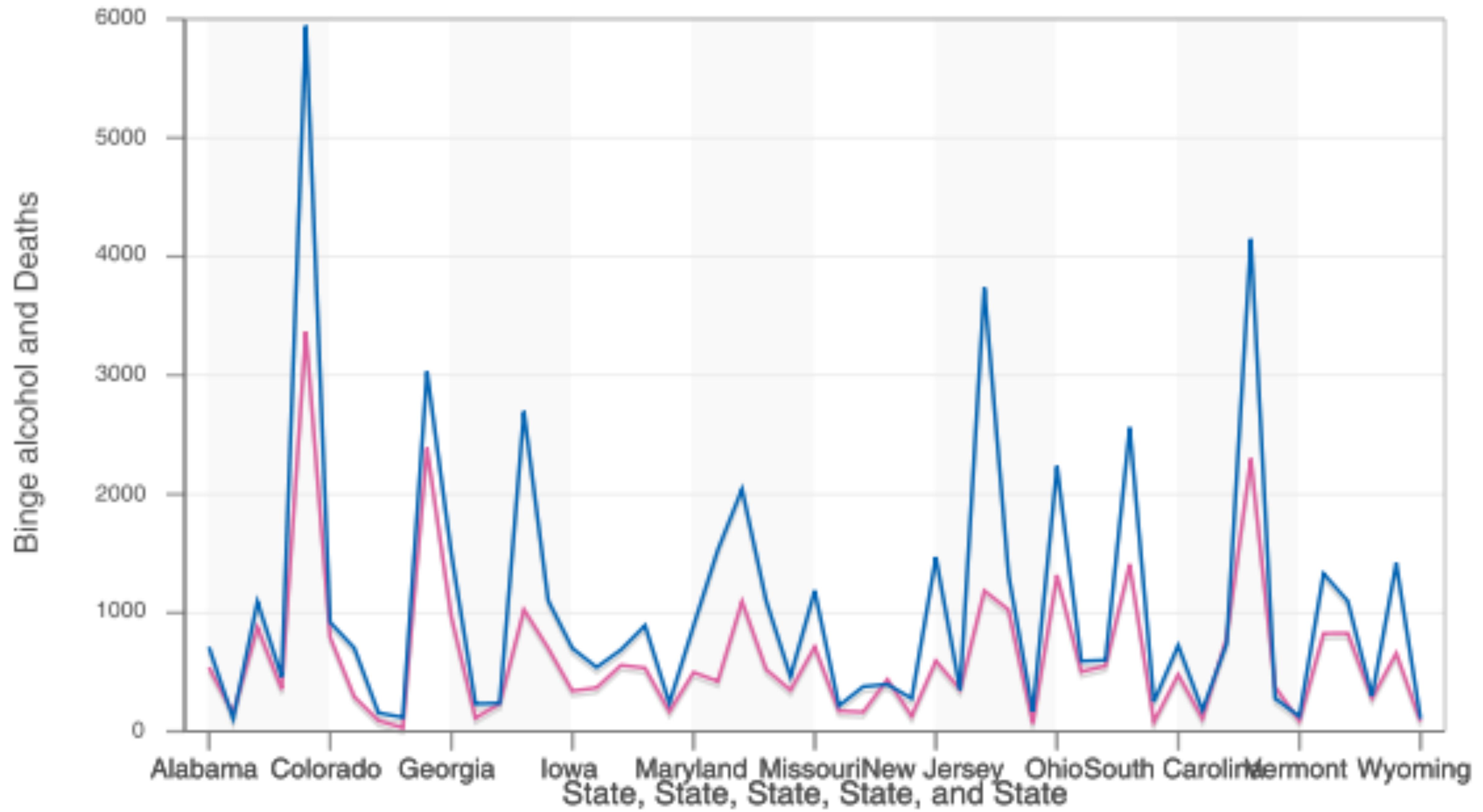
**Figure 1** Monthly (aggregate) distribution of viral pathogens, March 1, 2007-Feb 28, 2011.



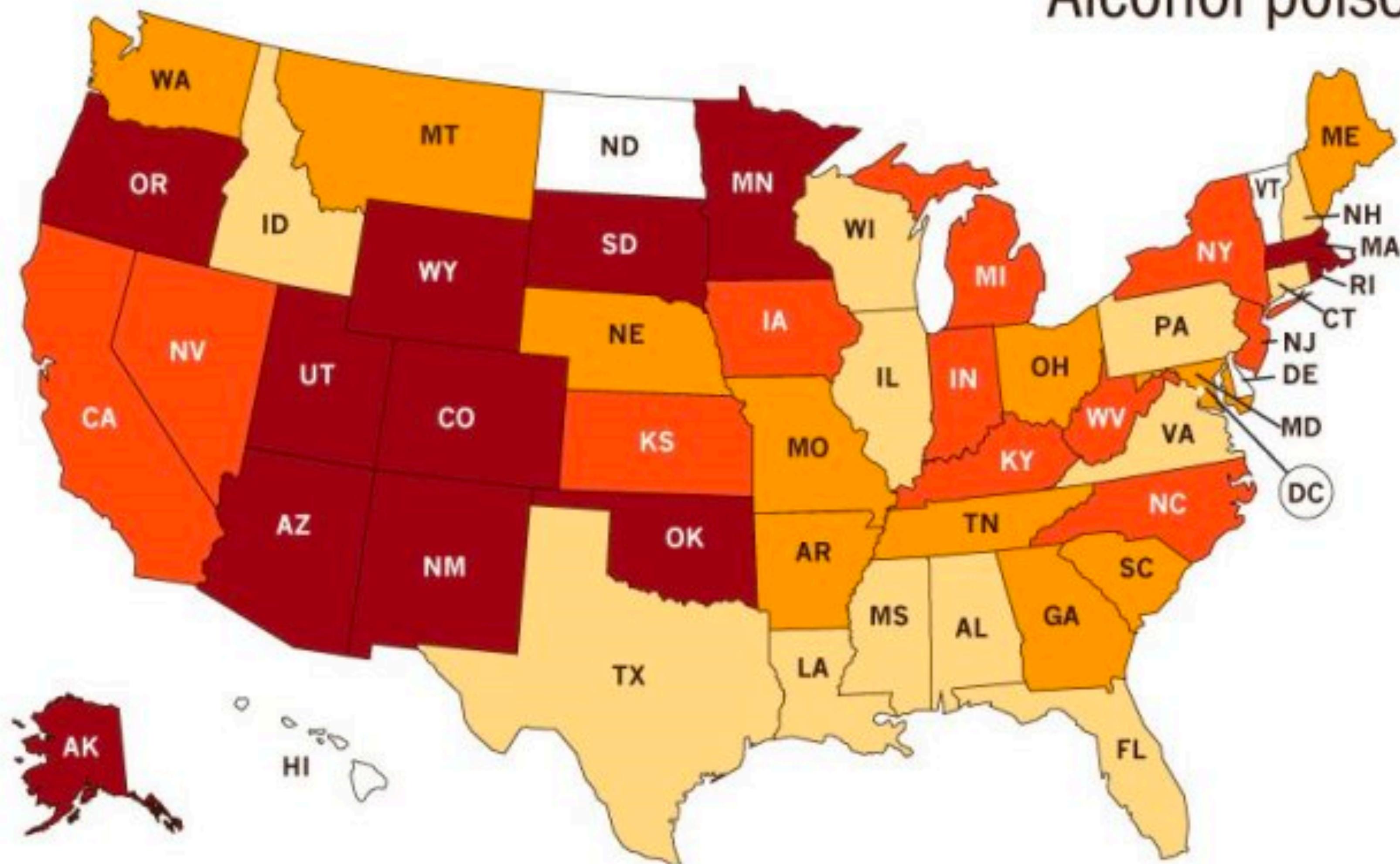
Meaningless color variation



As a bar plot



## Alcohol poisoning deaths vary by state



Deaths by state per million people	
	<b>11.9 - 46.5</b>
	<b>8.6 - 11.8</b>
	<b>6.8 - 8.5</b>
	<b>5.3 - 6.7</b>
	<b>Not calculated</b>

SOURCE: National Center for Health Statistics Mortality Multiple Cause Files, 2010-2012.

# 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

# YOUR TURN!

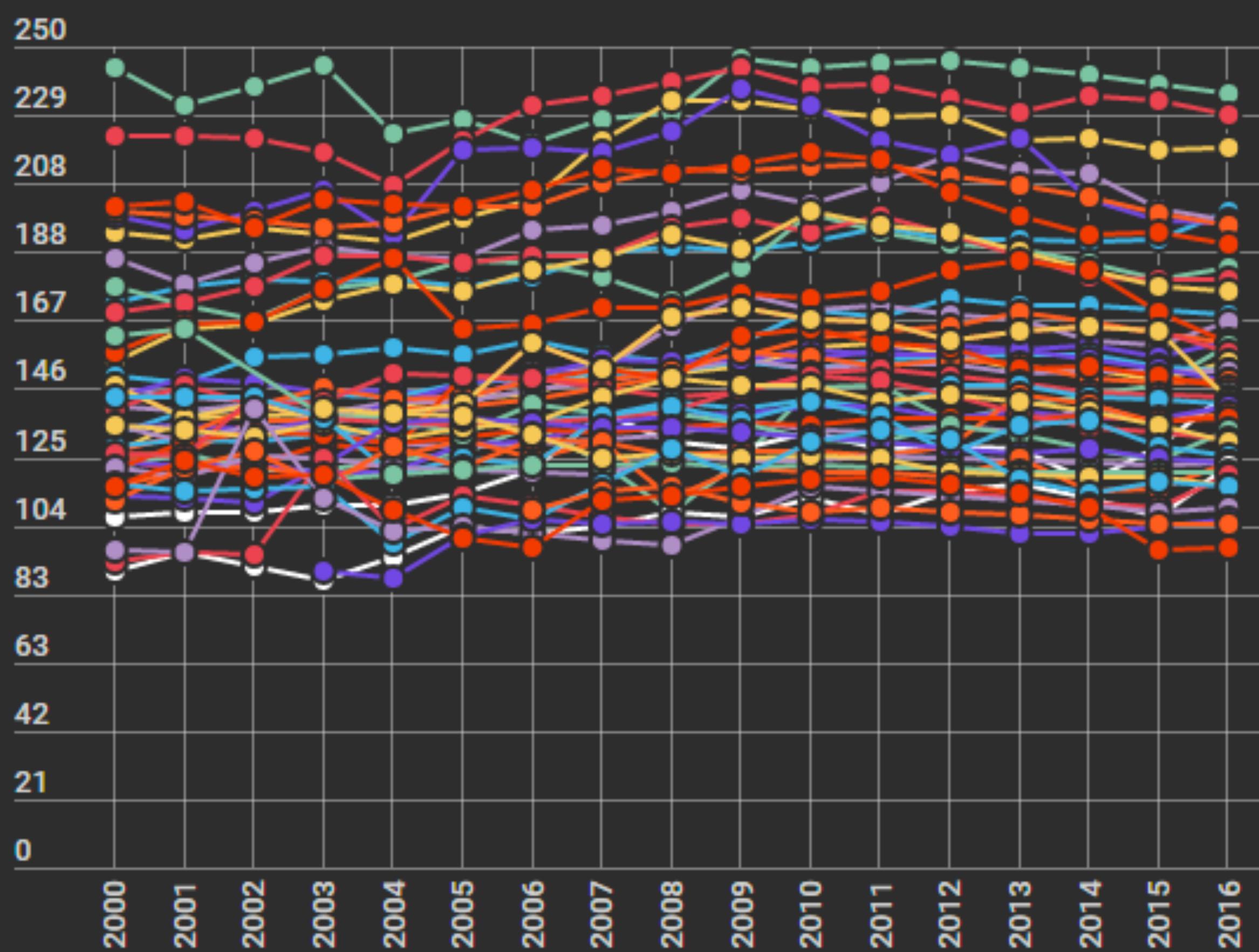
- Let's look at a plot.
- Try to answer three questions:

**What message(s) is the plot trying to convey?**

**Is the plot successful? If so, what choices make it successful? If not, what's wrong?**

**Any changes you'd suggest?**

# Canada's most heavily-policed cities (x per 100,000)



- Victoria
- Montreal
- Halifax
- Thunder Bay
- Winnipeg
- Vancouver
- Windsor
- Toronto
- Edmonton
- Regina
- Saskatoon
- Calgary
- St. John's
- Greater Sudbury
- Barrie
- Lethbridge Region
- Red Deer
- Niagara Region
- Kingston
- Chatham-Kent
- Hamilton
- London
- Guelph
- Abbotsford
- Surrey
- Peel Region
- Delta
- Waterloo Region
- Laval
- York Region
- Kelowna
- S

# 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

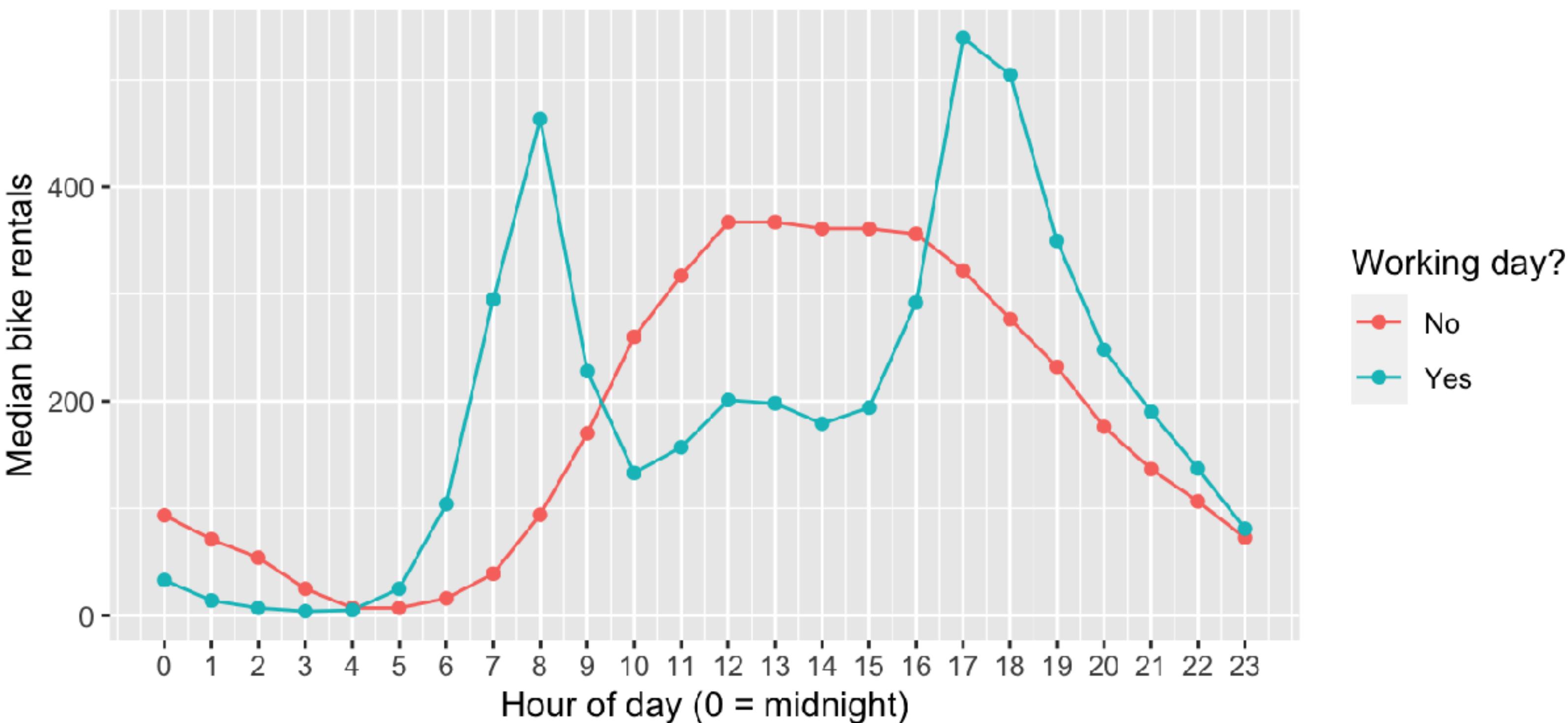
- Mappings:

- Hour → x

- Median rentals → y

- Working day → color

Bike-share rentals in Washington, DC (2011-12)



# EXAMPLE: BIKE SHARE DATA

- Variables:

- Hour

- Median rentals

- Working day

- Objects:

- The two panels

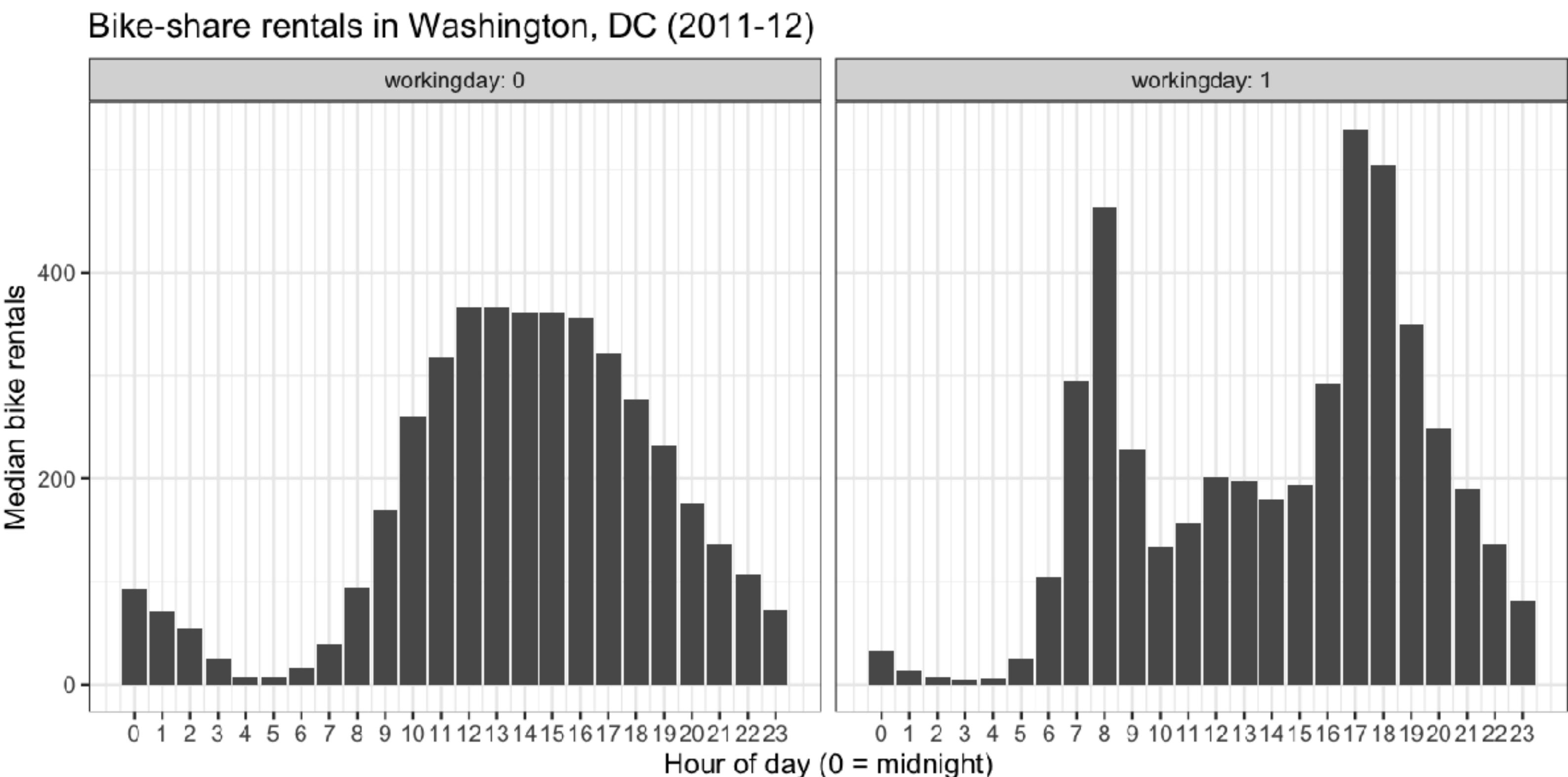
- The bars in each panel

- 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

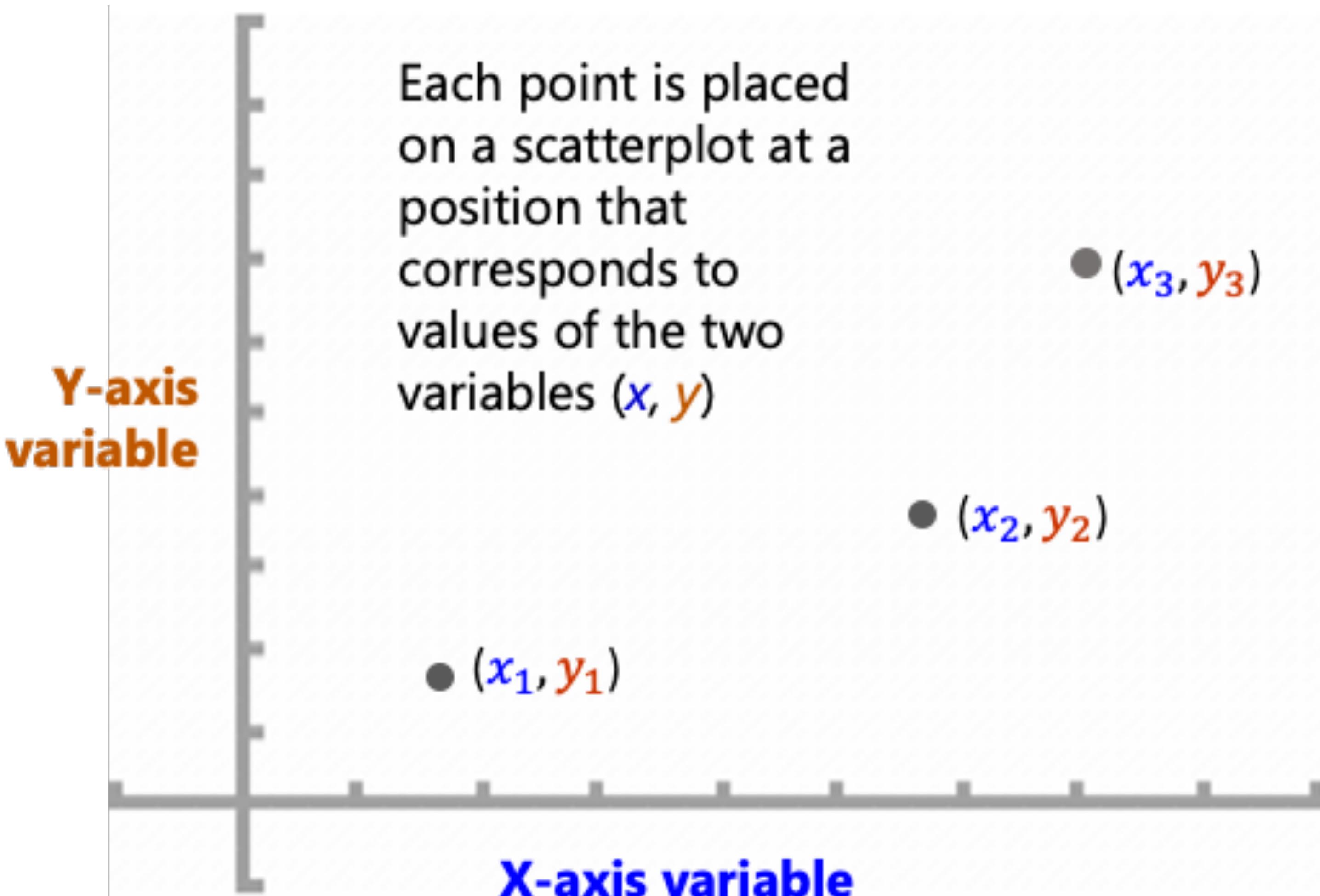
In the table, x refers to an explanatory variable and y refers to a response variable (i.e. the variable whose behavior is being “explained” by x).

# THE FIVE MOST IMPORTANT PLOTS

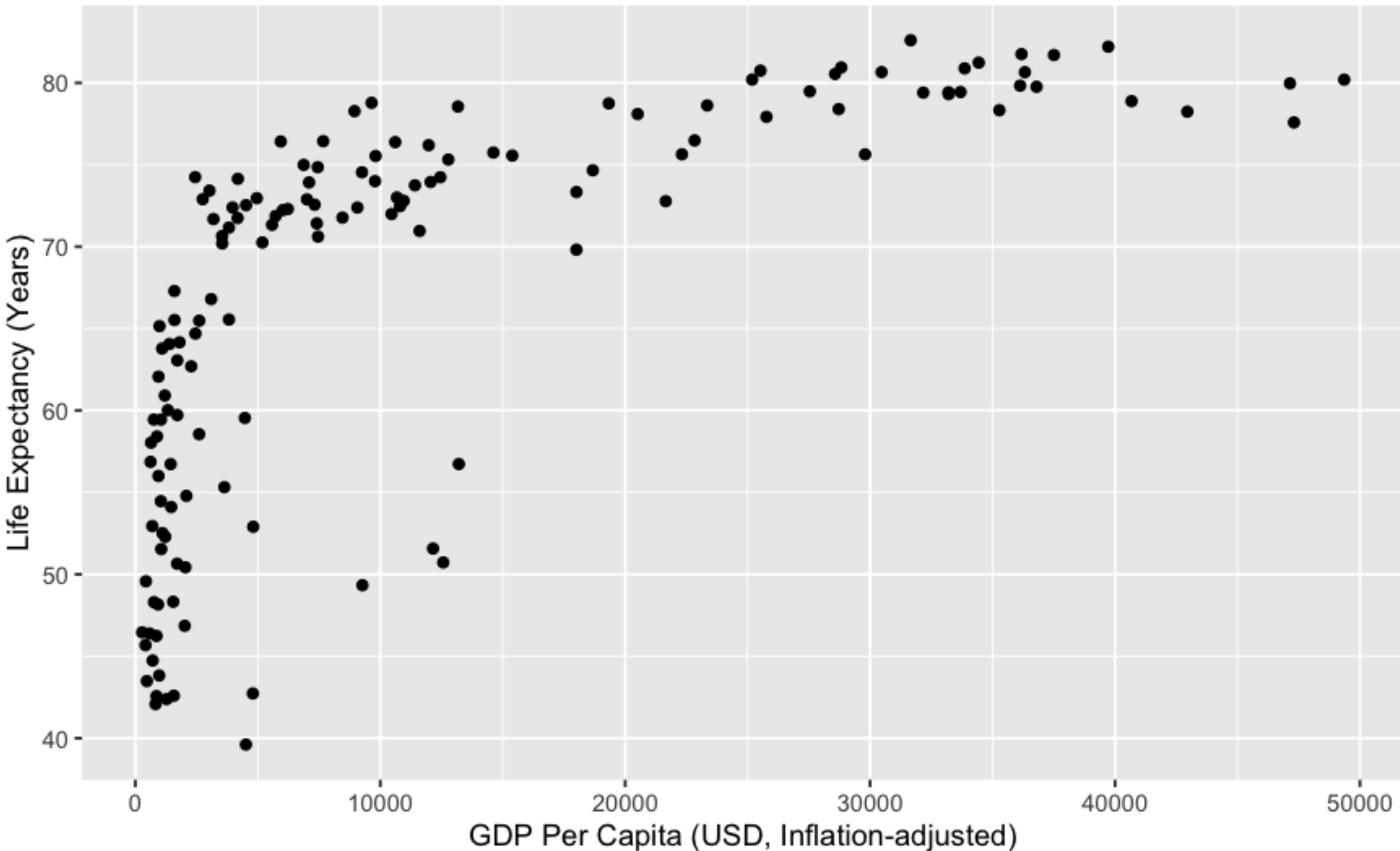
If you have...	Consider a...	Using...
Numerical y and numerical x	Scatter plot	geom_point
Numerical y and sequential x	Line graph	geom_line
Numerical y alone	Histogram	geom_histogram
Numerical y and categorical x	Boxplot	geom_boxplot
Summary statistics by group	Bar plot	geom_col

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).

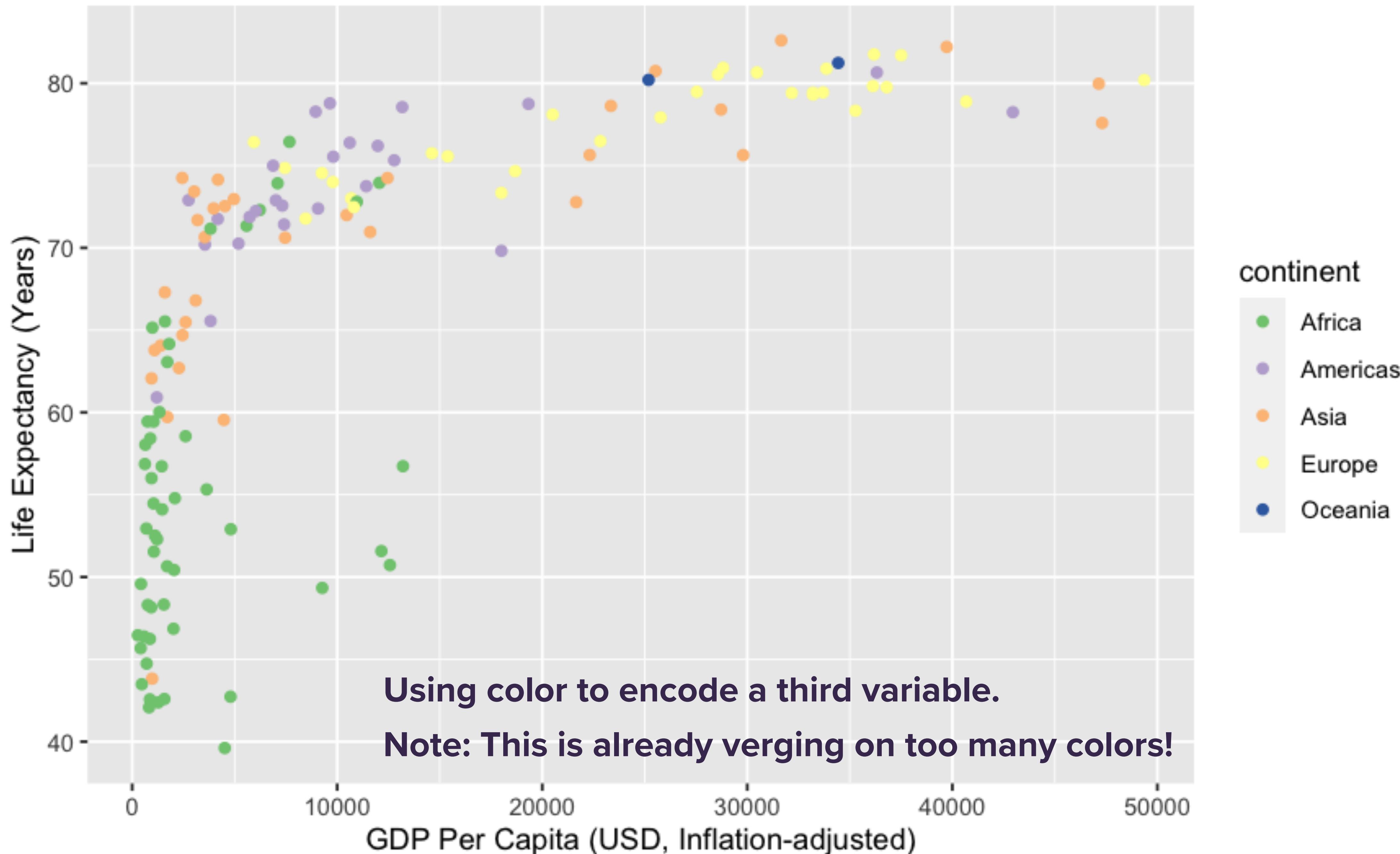
# SCATTER PLOTS



Health versus Wealth for 142 Countries: 2007



# Health versus Wealth for 142 Countries: 2007



# 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.

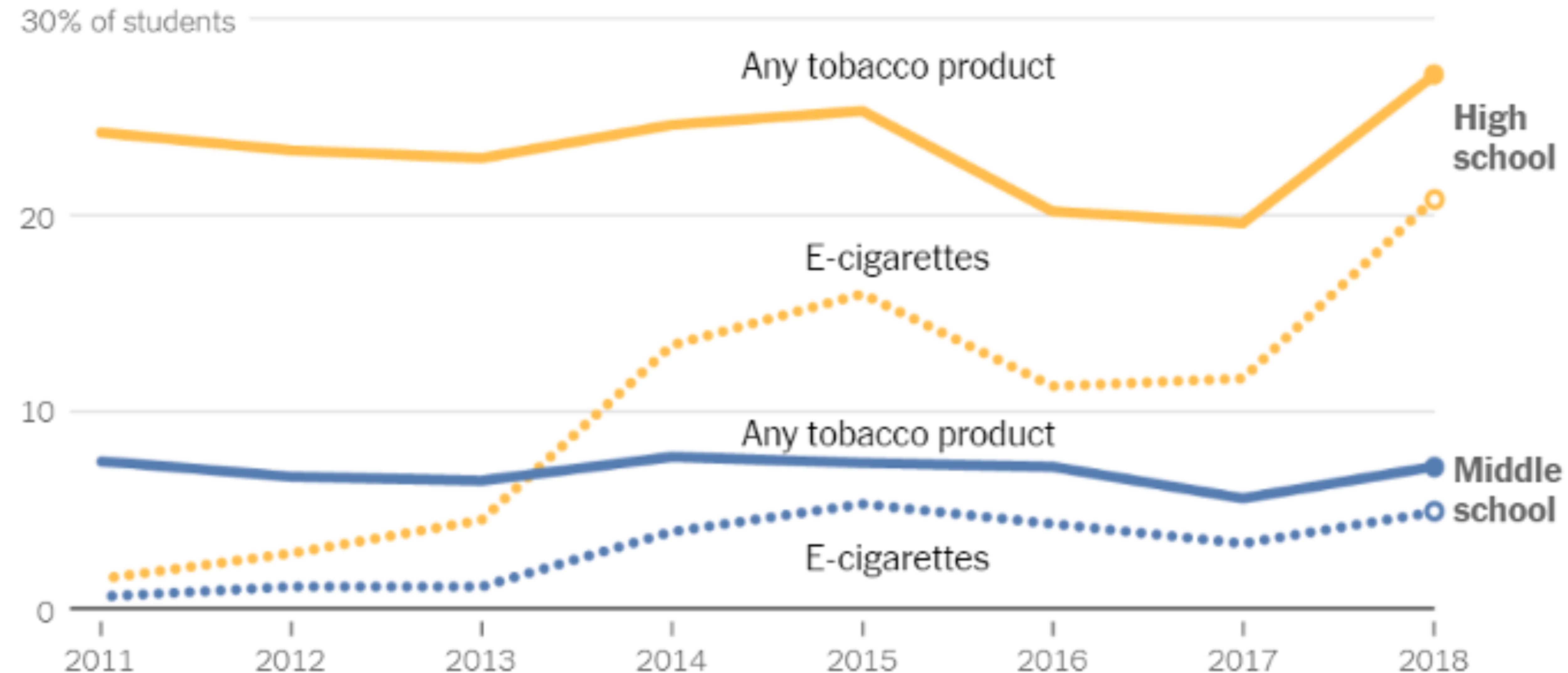
Time

Dosage levels of a drug

Distance

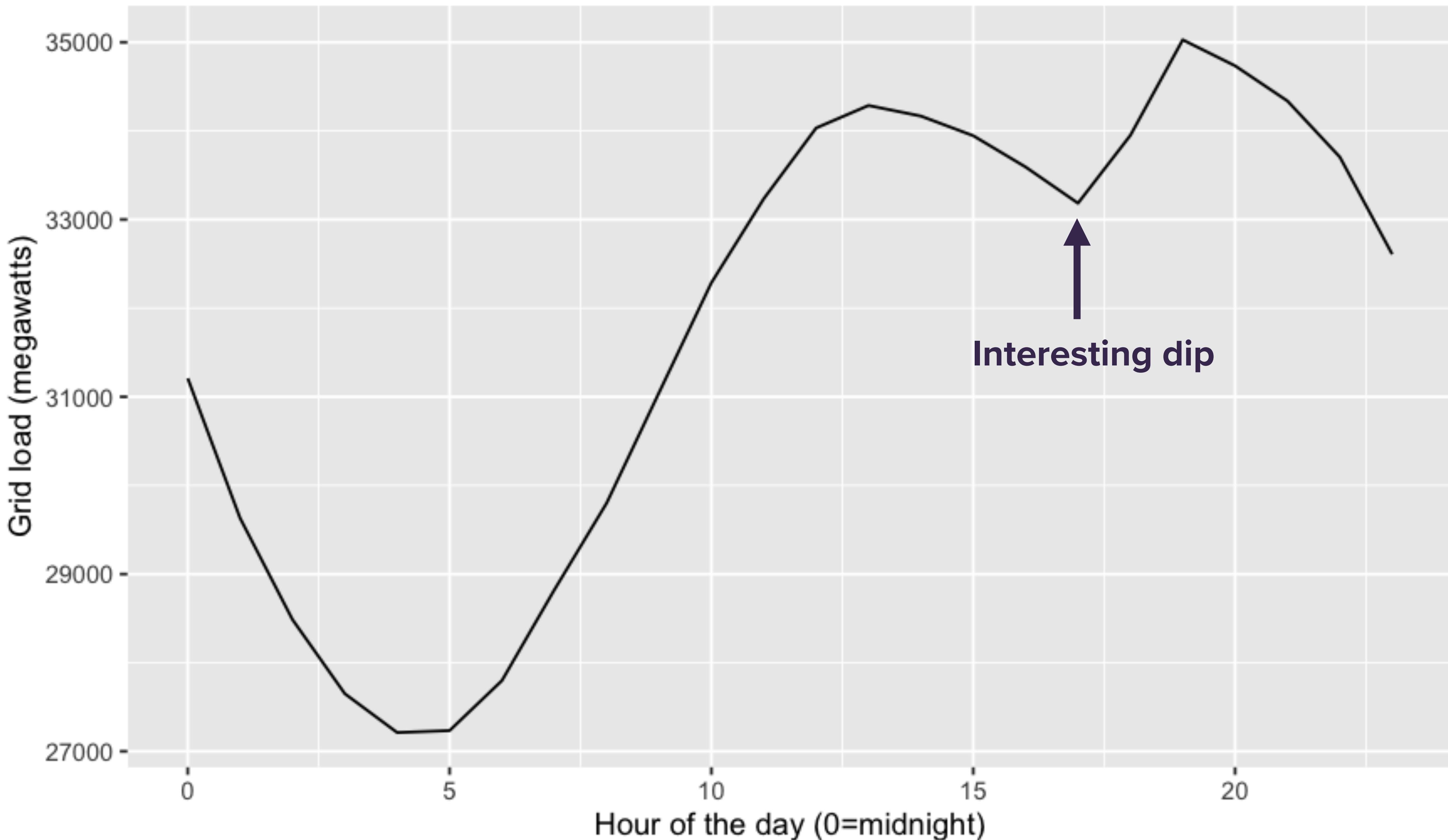
Etc.

# Tobacco Consumption Among Students



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

# Load on the Texas power grid, Christmas Day 2015



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).

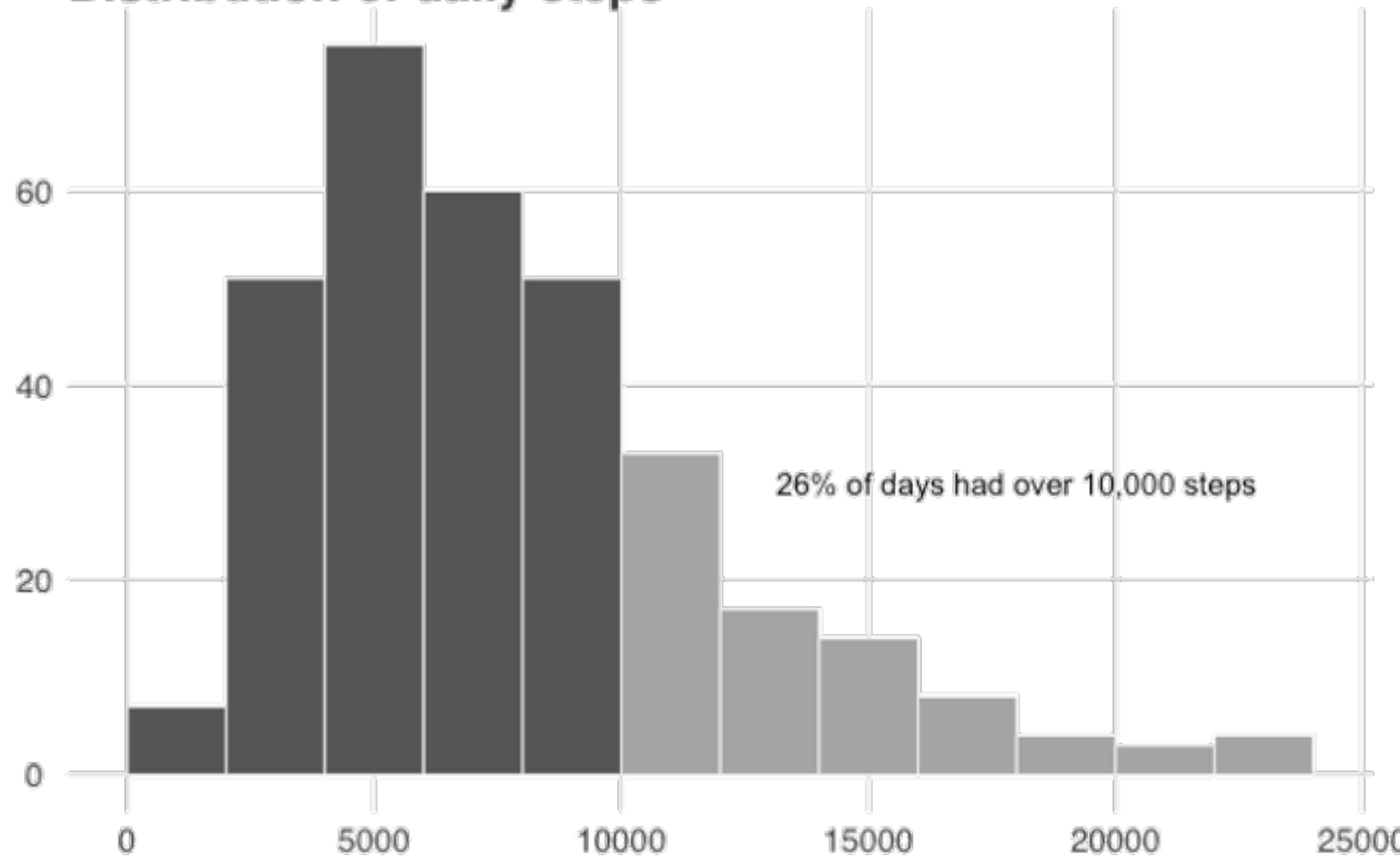
# HISTOGRAMS

- 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**

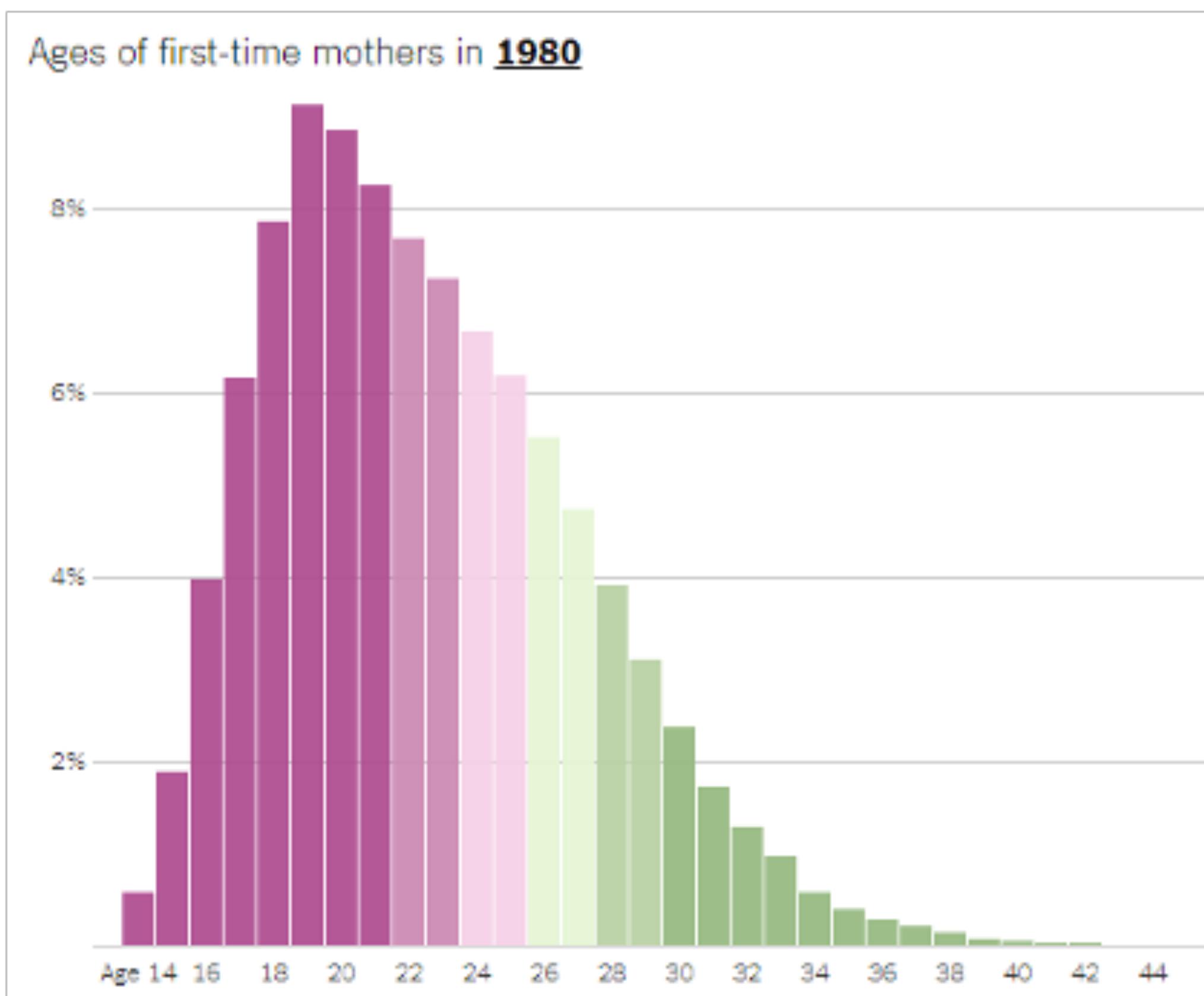
**Density histograms: the height of each bar is rescaled so that the total area under the curve is 1**

# Distribution of daily steps



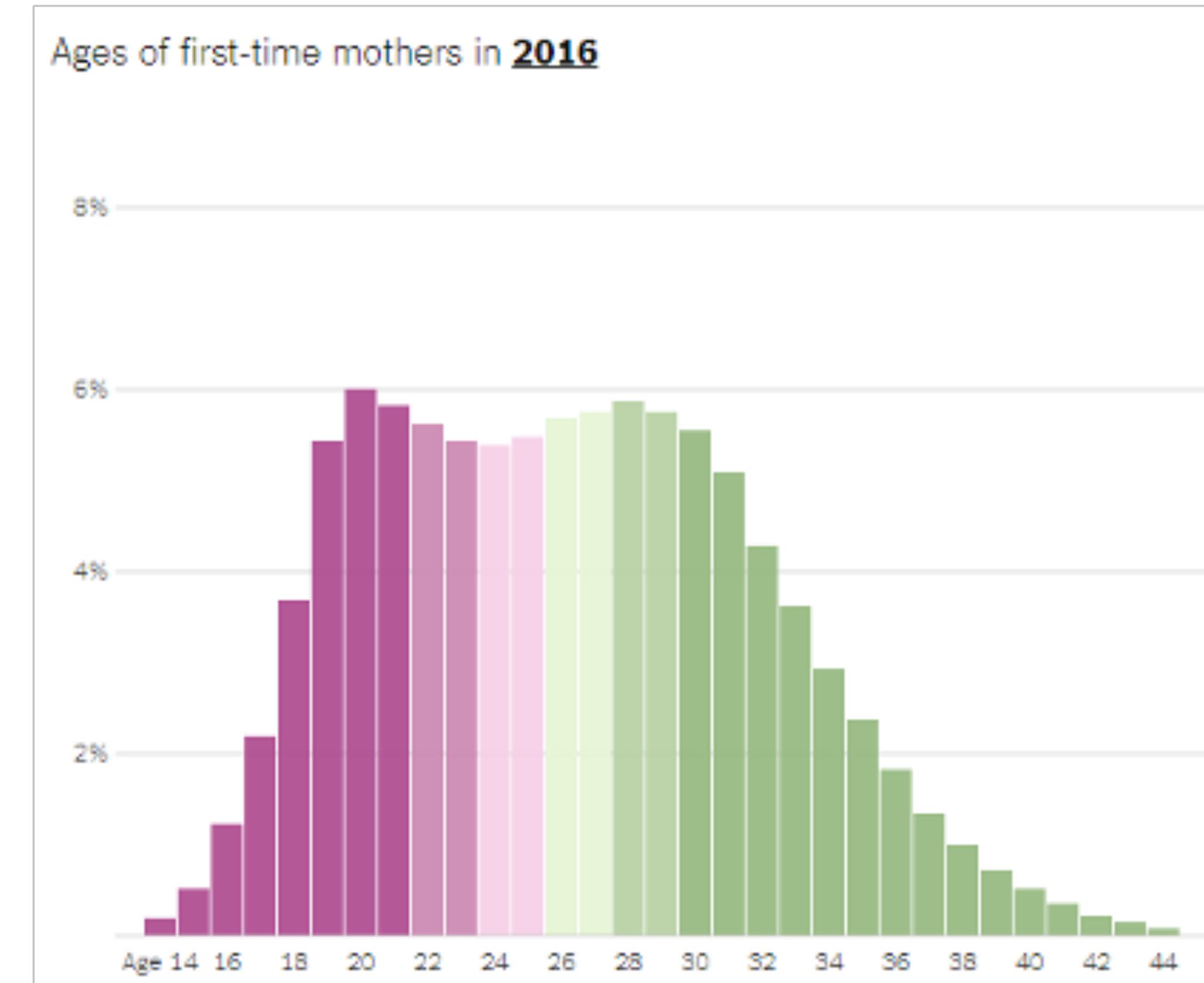
# The Age That Women Have Babies: How a Gap Divides America

By QUOC TRUNG BUI and CLAIRE CAIN MILLER AUG. 4, 2018



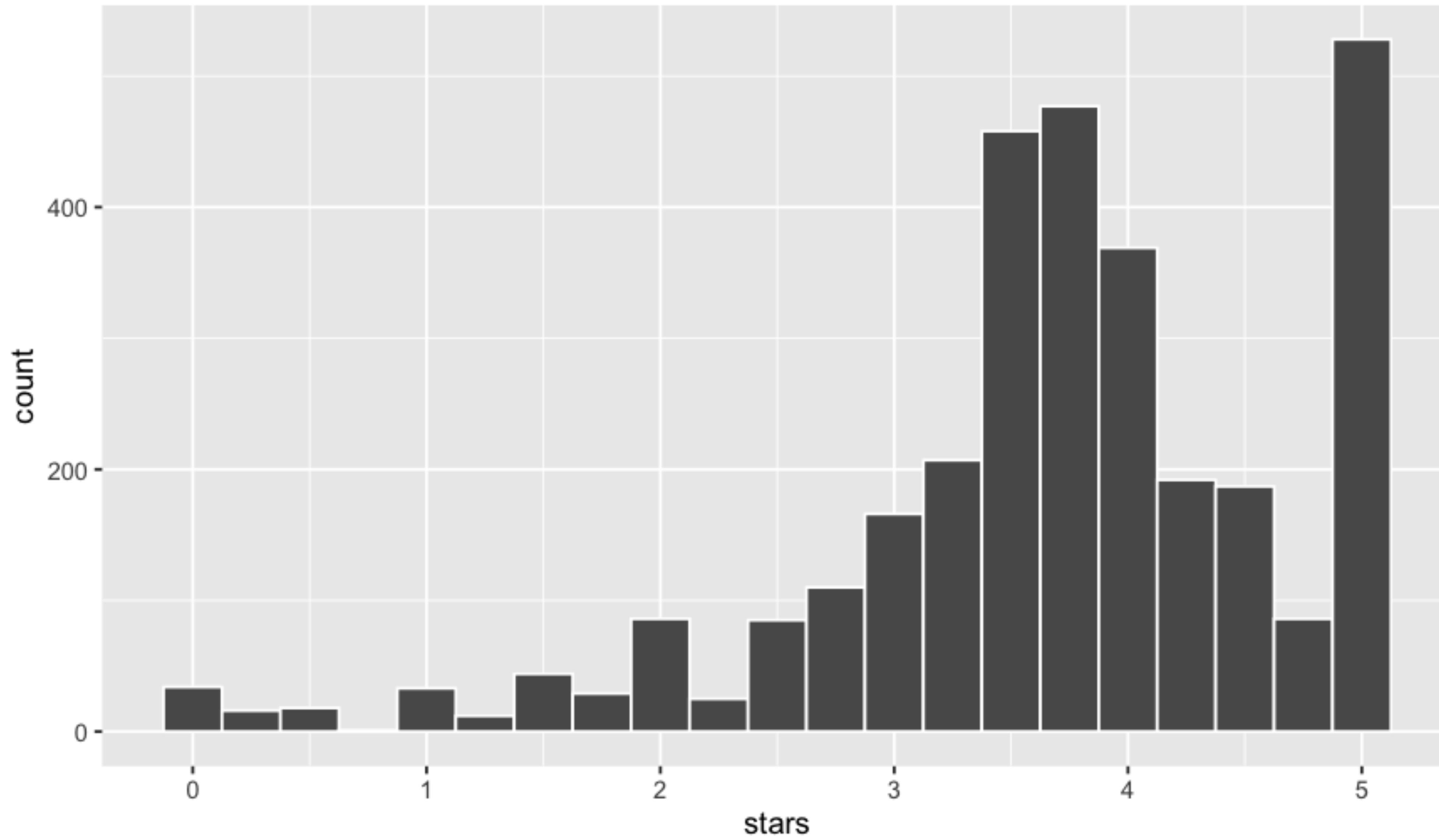
Incorporate a categorical variable (here, two “snapshots” of 1980 and 2016) with side-by-side *faceted histograms*.

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%).

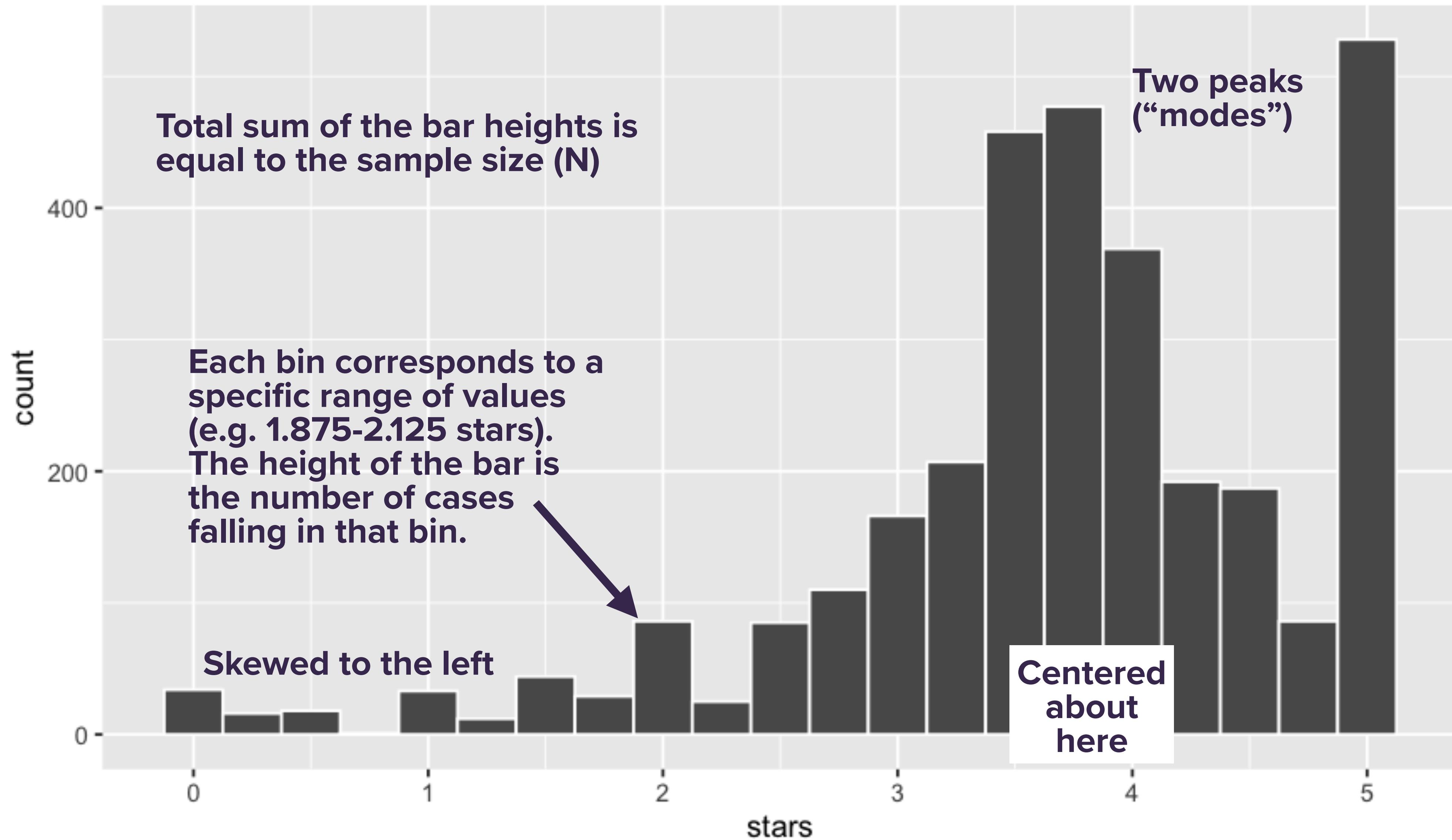
## Distribution of ratings (0-5 scale) across 2764 ramen reviews



Data from The Ramen Rater, <https://www.theramenrater.com/resources-2/the-list/>

We'll make the next several histograms ourselves.

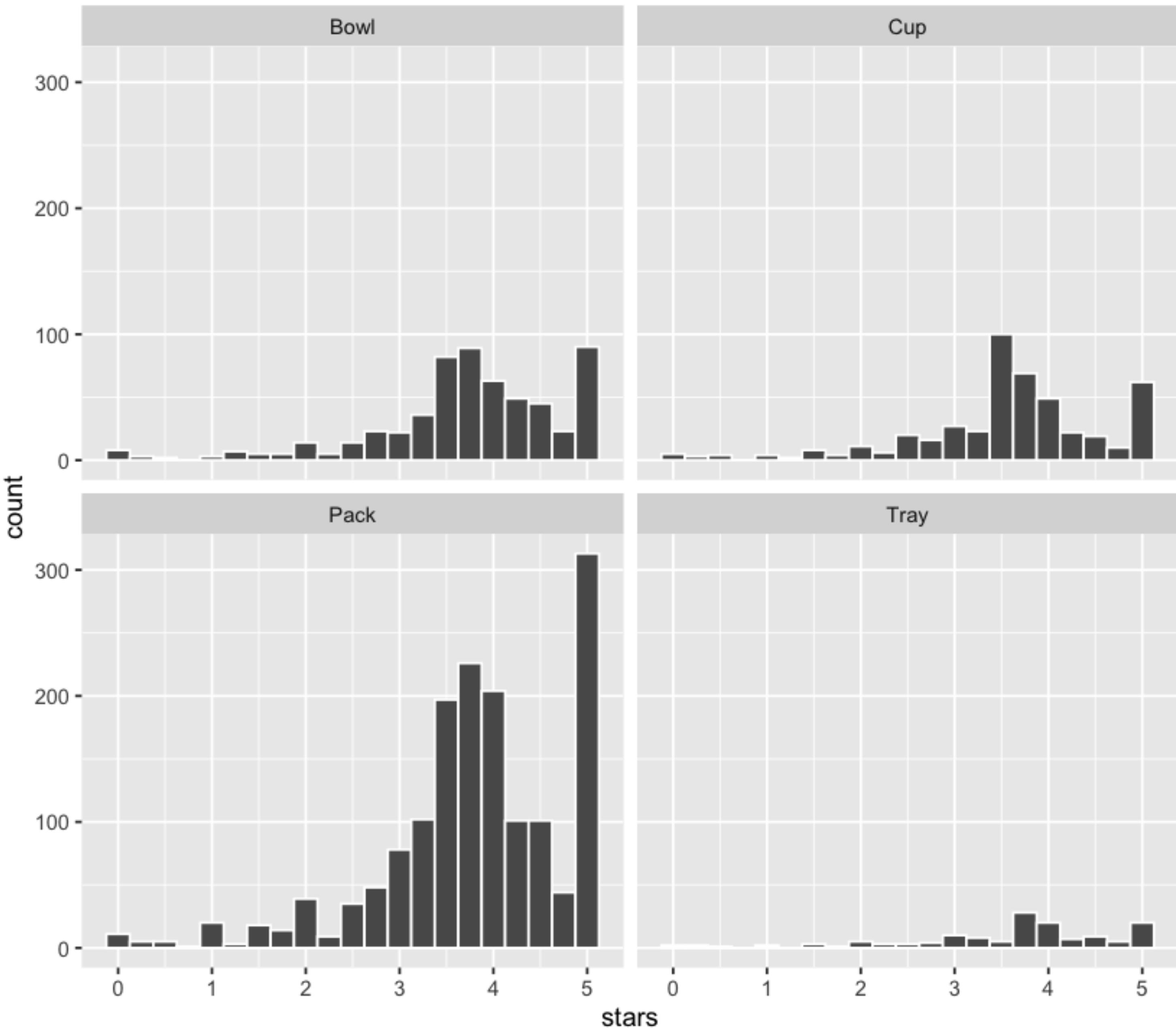
## Distribution of ratings (0-5 scale) across 2764 ramen reviews



Data from The Ramen Rater, <https://www.theramenrater.com/resources-2/the-list/>

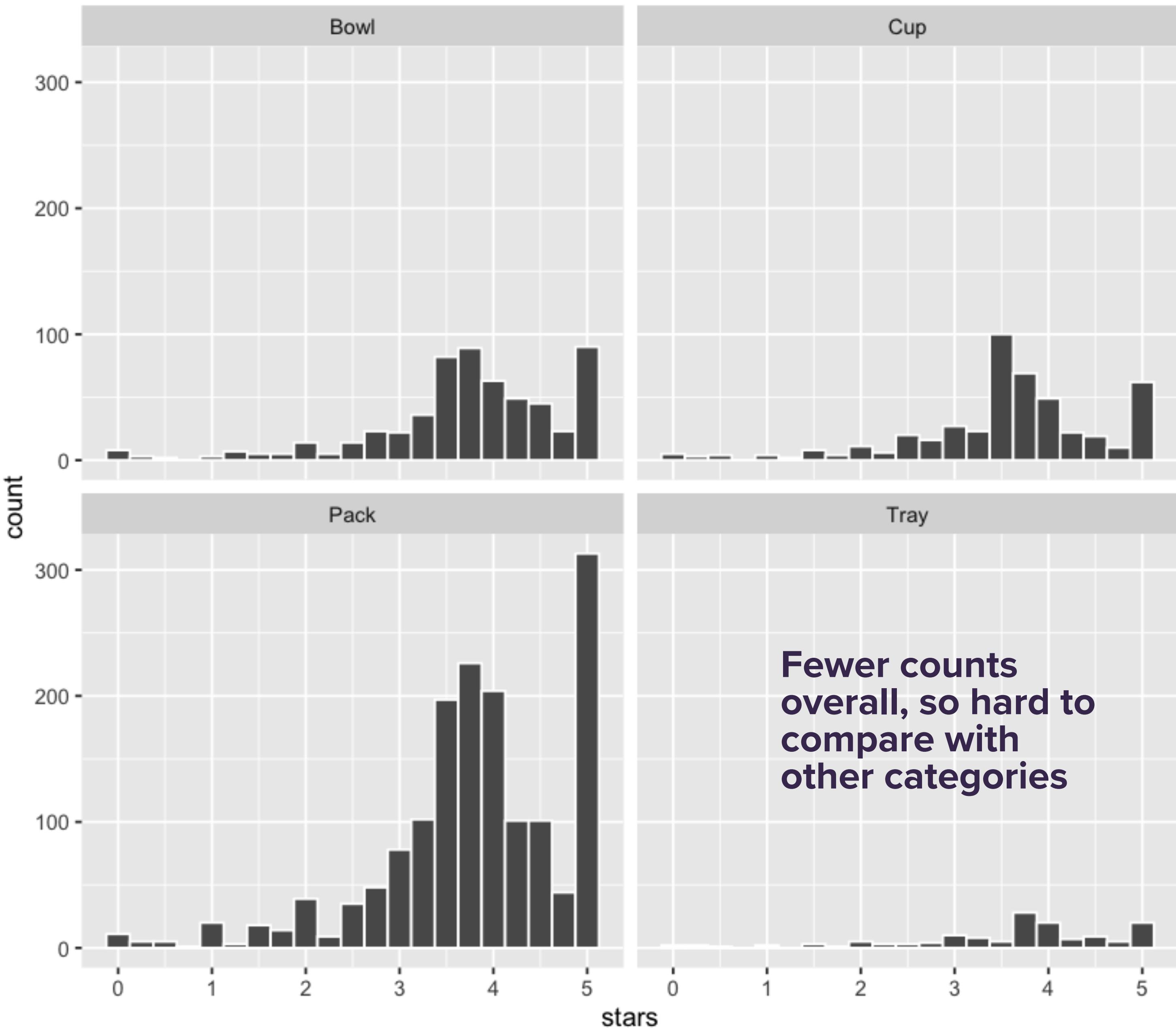
# FACETING

- **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).



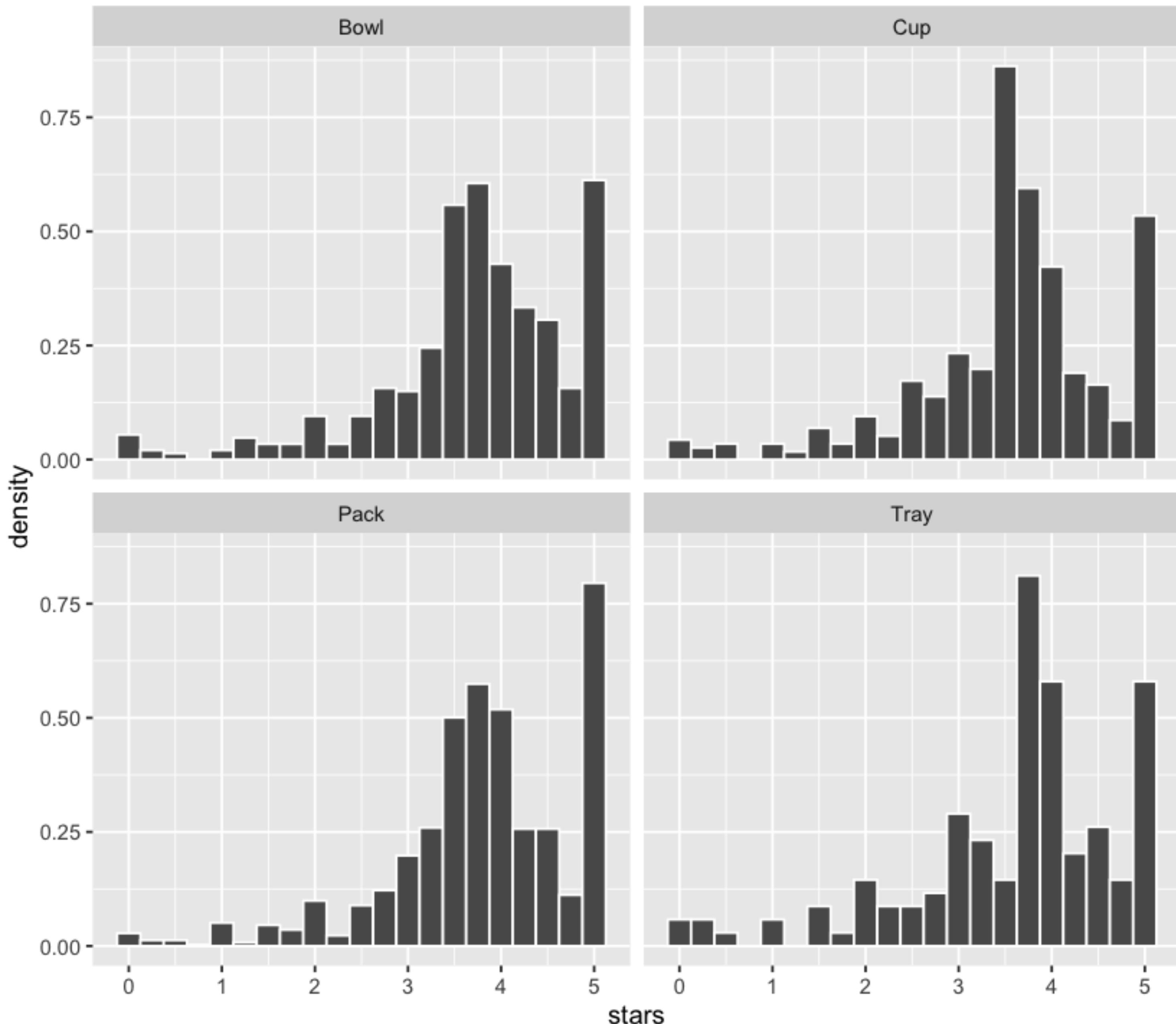
# FACETING

- **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**

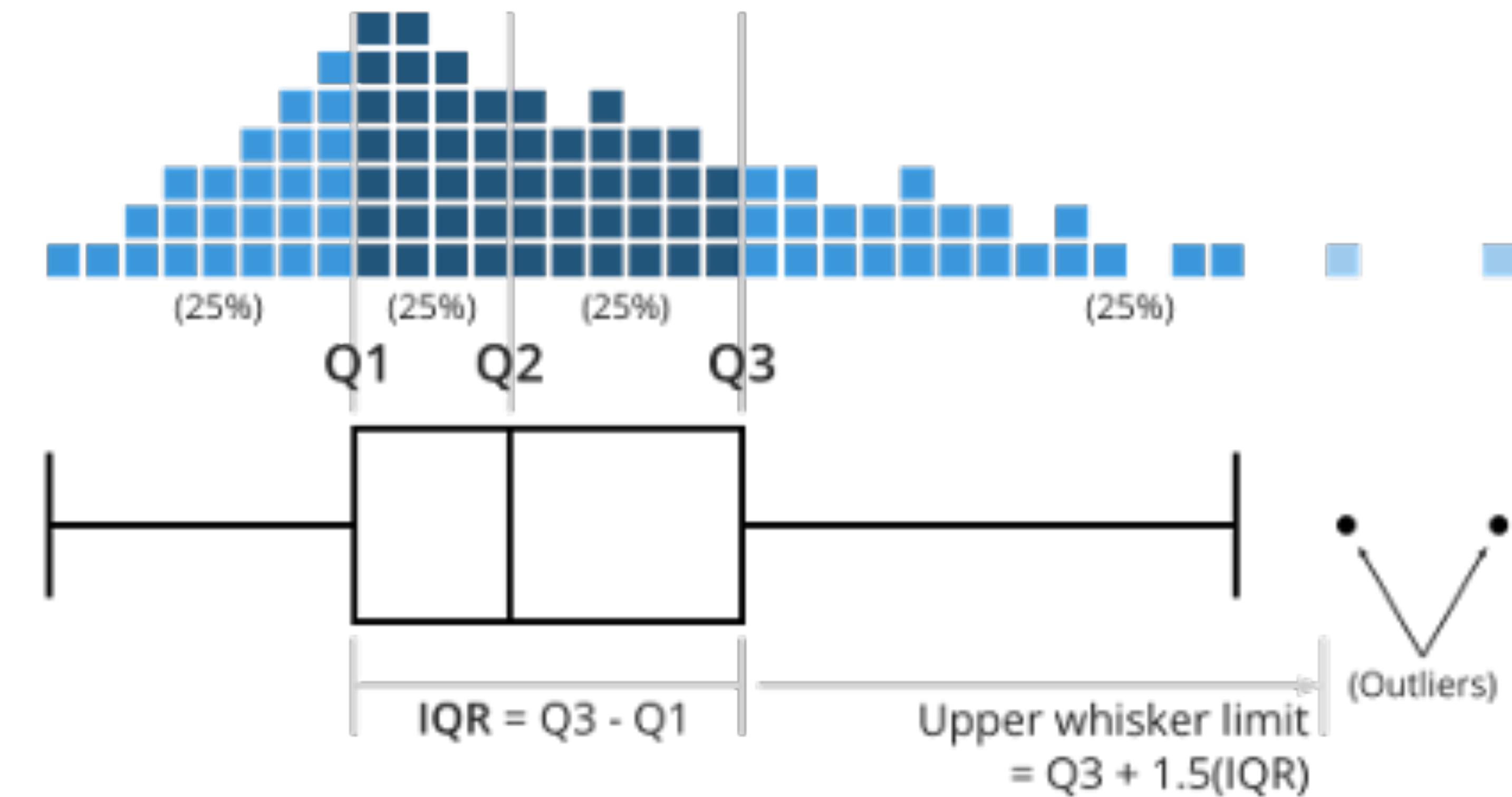
**25th and 75th percentiles**

**Lower and upper extremes**

- Are similar in spirit to faceted histograms, but don't show the entire data distribution.

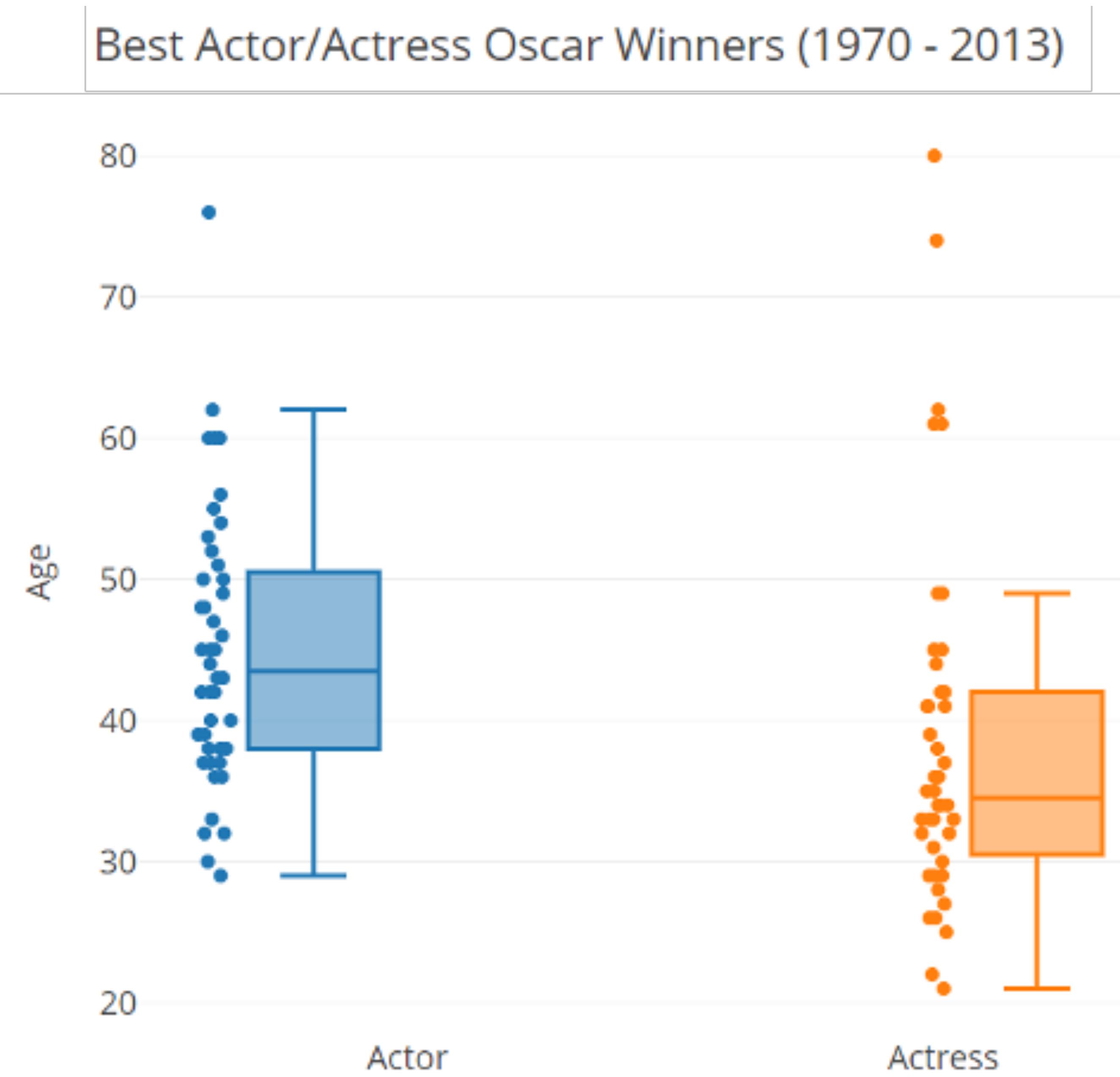
# BOXPLOTS

- 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.

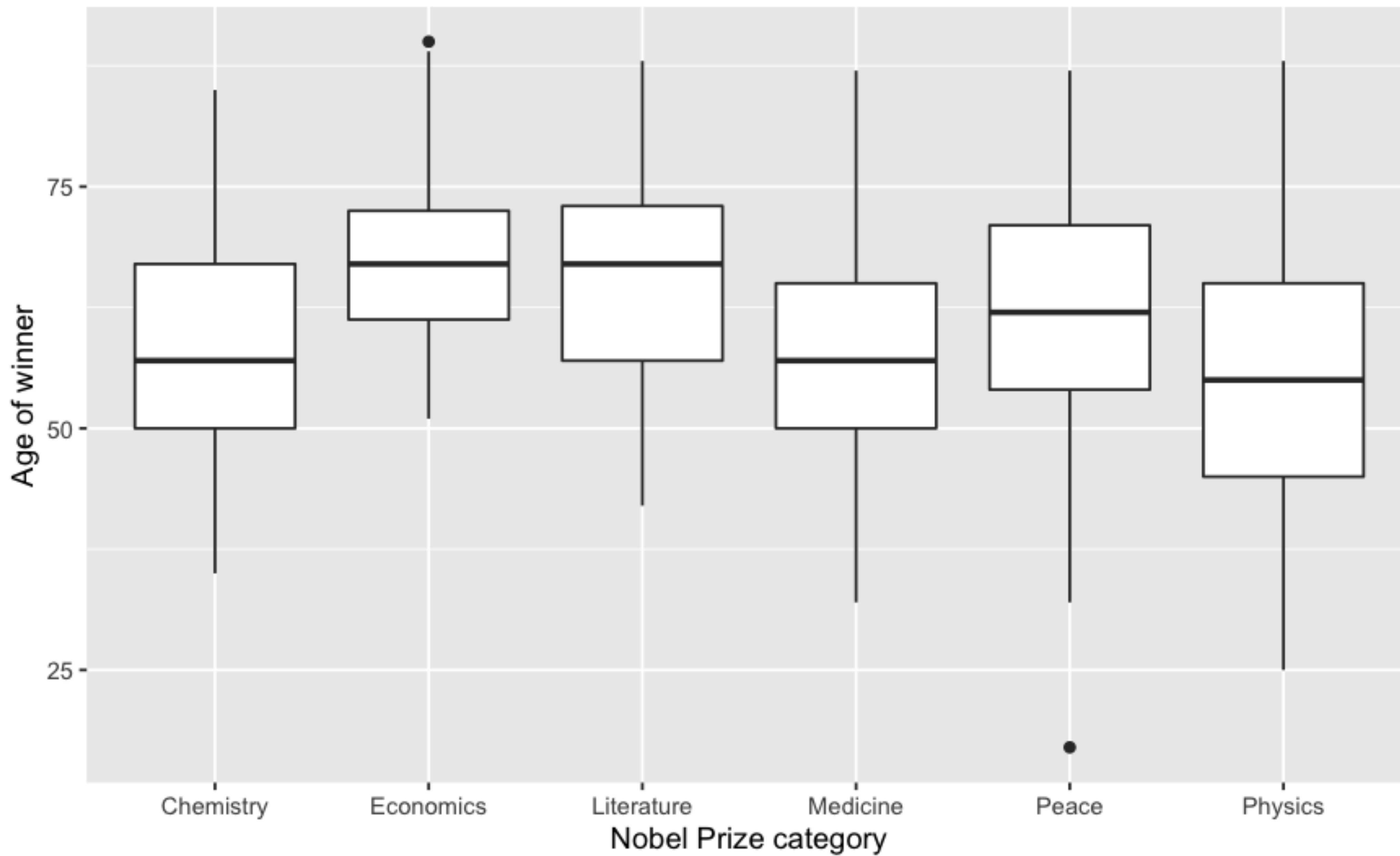


# BOXPLOTS

- 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



# 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

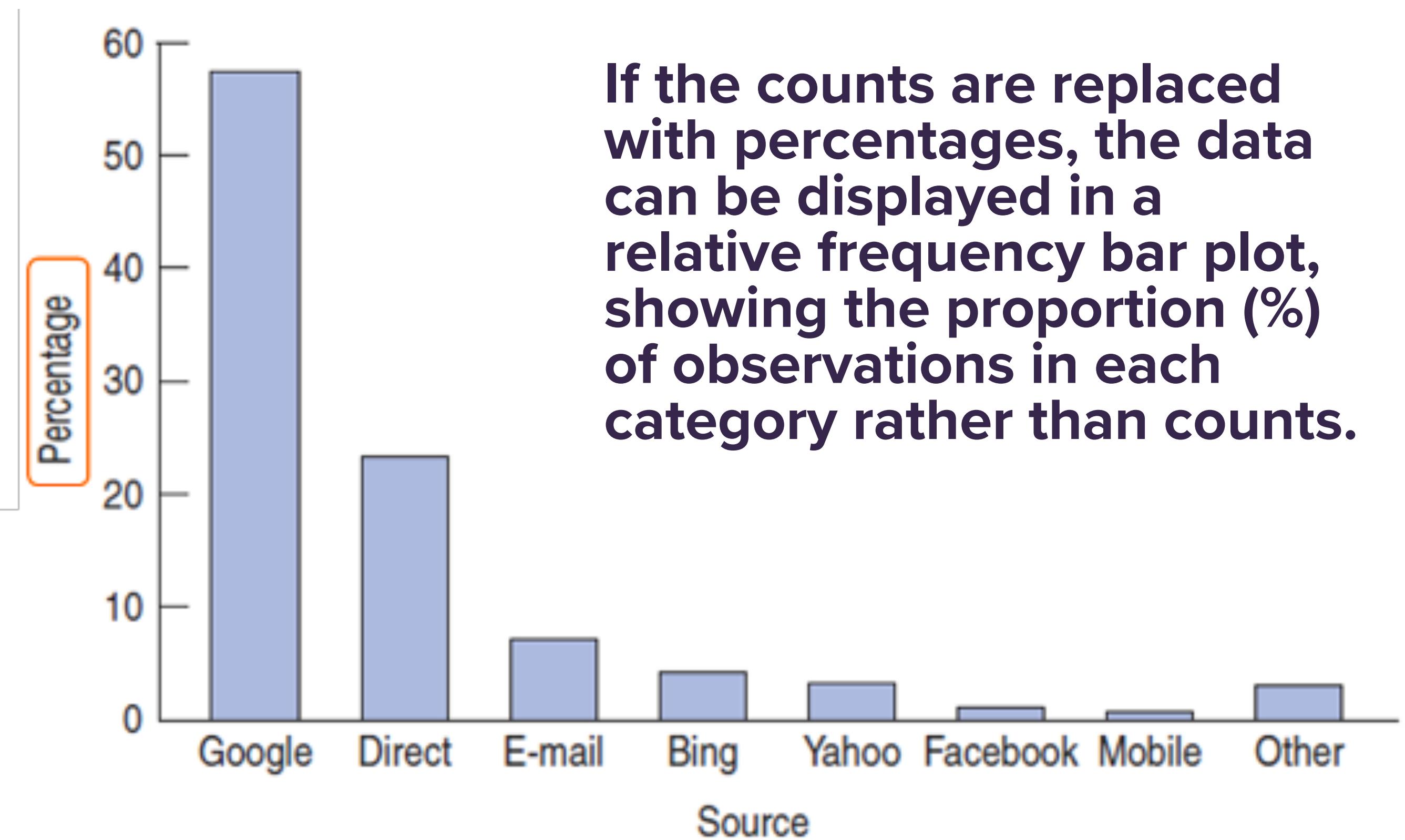
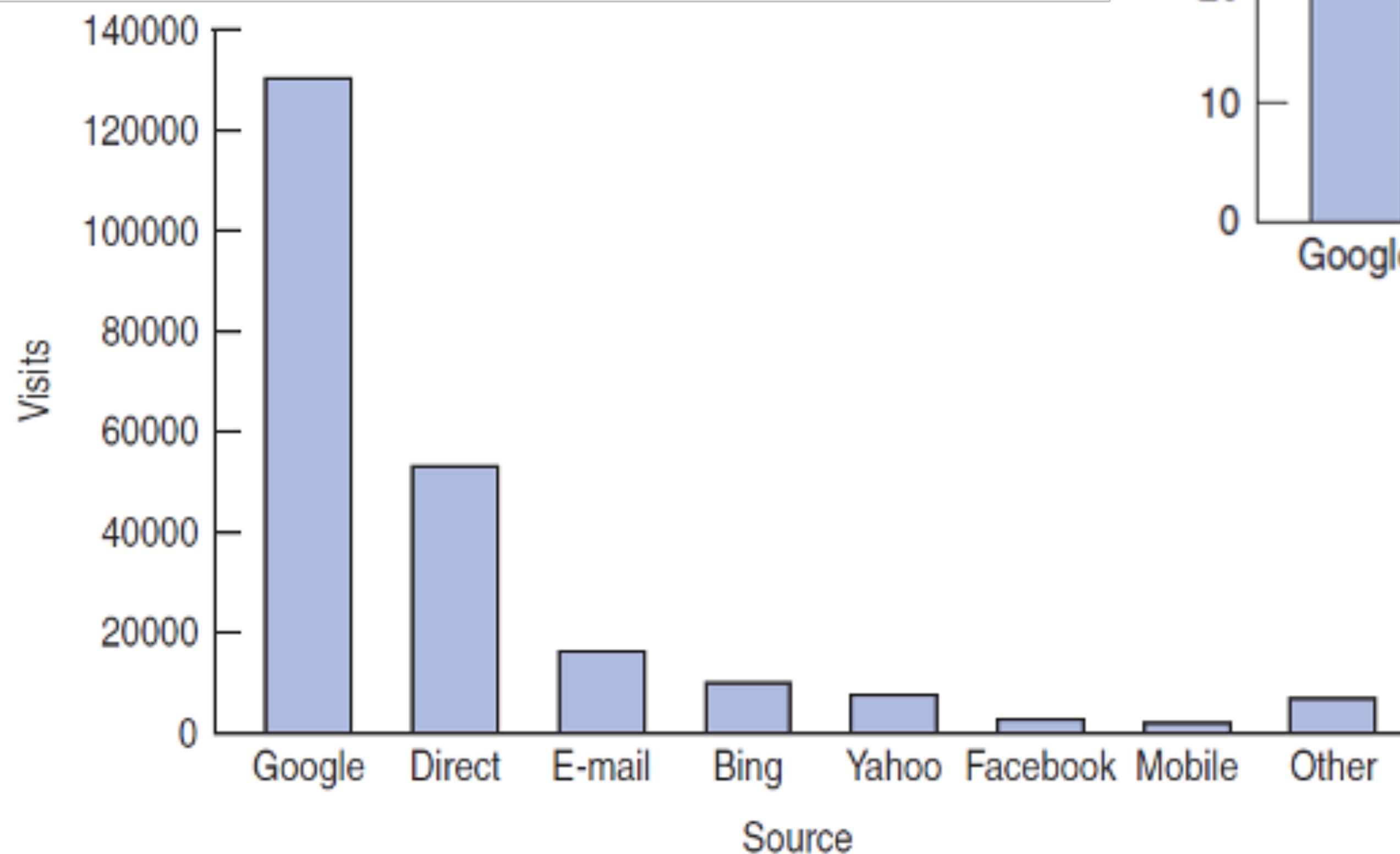
- Are typically used to compare summary statistics across groups.

**Means or totals (for numerical variables)**

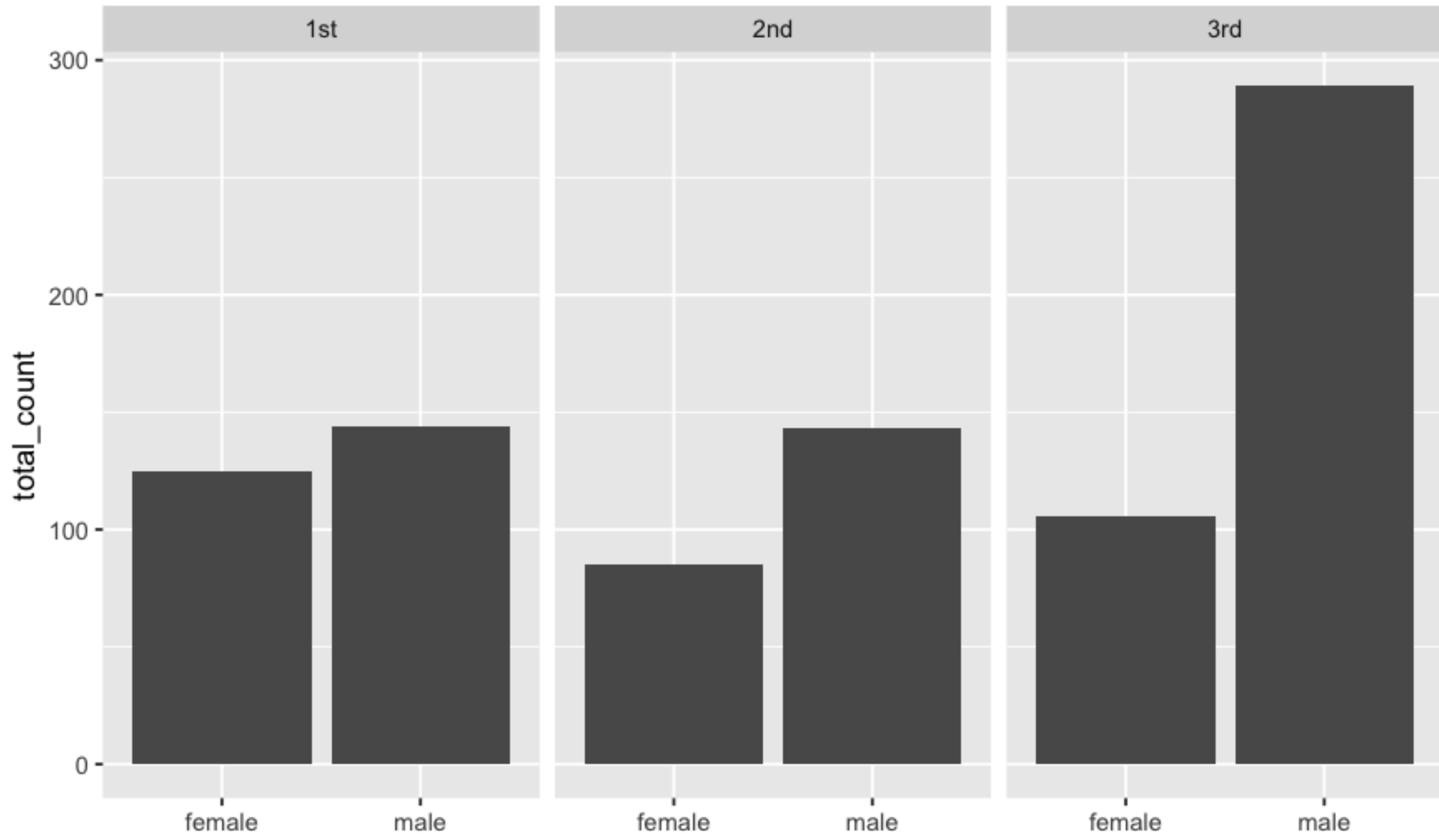
**Counts or proportions (for categorical variables)**

- 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.**

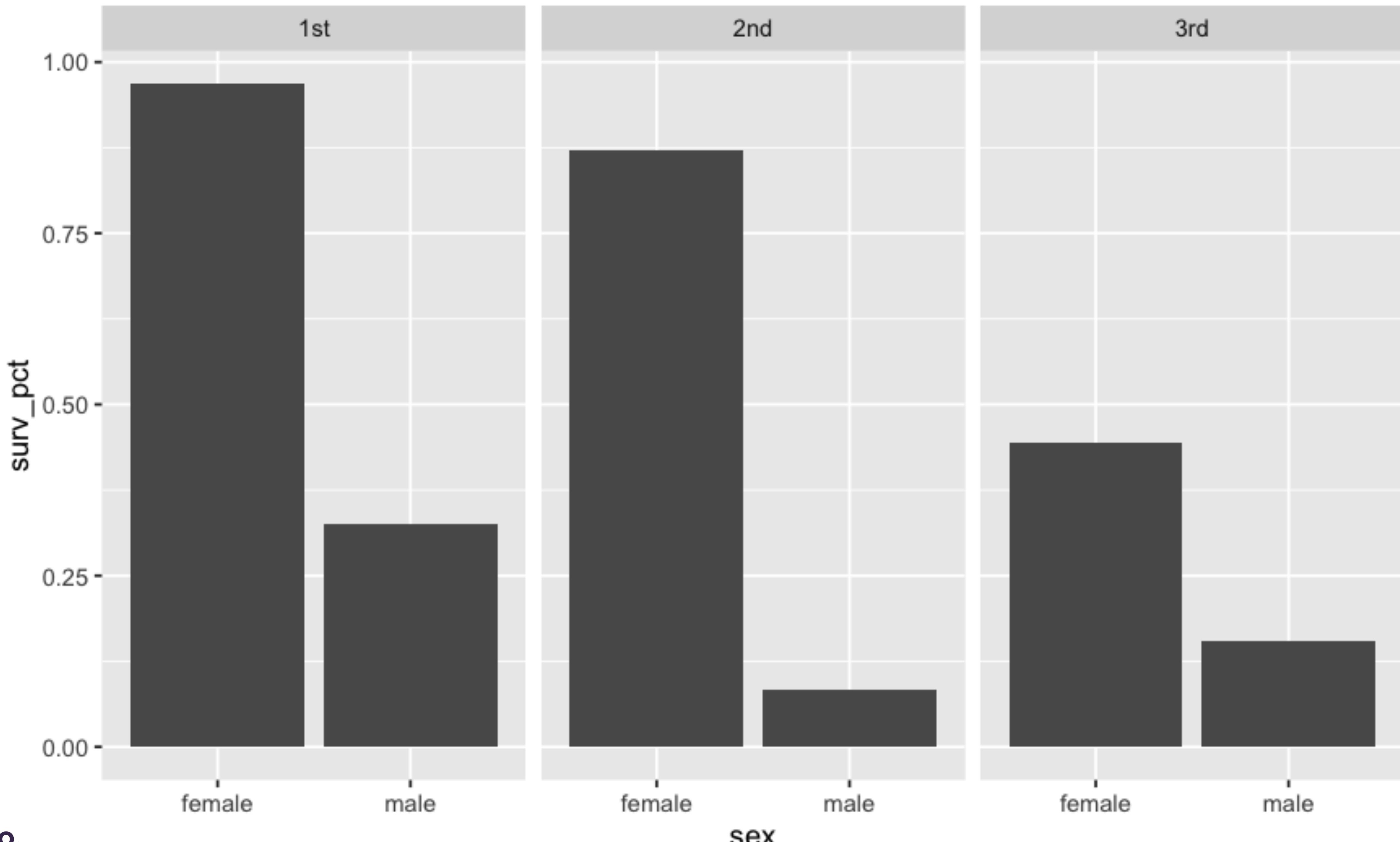


# Who sailed the Titanic? Passenger count by class and sex



Here's a bar plot we'll make ourselves.

# Who survived the Titanic? Survival percentages by class and sex



This too.

# REVIEW: START WITH WHAT'S SIMPLE AND EFFECTIVE

- 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

# CODING SESSION: OLYMPICS DATA

- The data in `olympics_top20.csv` contains information on every Olympic medalist in the top 20 sports by participant count, all the way back to 1896.
- That's 23,850 medals!
- We'll use plots to examine two example questions:
  - (1) How do athletes' heights compare across the summer vs. winter games? Does one season have taller athletes, on average?
  - (2) How has the average age of medalists changed through the years? Is the answer different for men and women?

**datavis\_intro.R**