Design principles for data presentation: Translating analysis into visualization

Andrew Eppig, UC Berkeley
CAIR Annual Conference
8 November 2012

Overview

- Design Principles
- Data Sample and Summary
- Why Charts Matter
- Visualization Process
  - Chart Selection
  - Layout
  - Aesthetics
  - Self-Sufficiency Check
- Institutional Examples and Applications

Guiding Design Principles

- Good visualizations start with good data and detailed analysis
  - Know your data
- Good visualizations directly answer specific, focused questions
  - Know what question(s) you are asking
- Good visualizations get out of the way of the data
  - Let the data tell its story without excess clutter or distraction

"Too often we pay more attention to 'pretty' than to the **most important element: information**."
-- Dona Wong, *The Secrets of Graphics Presentation*

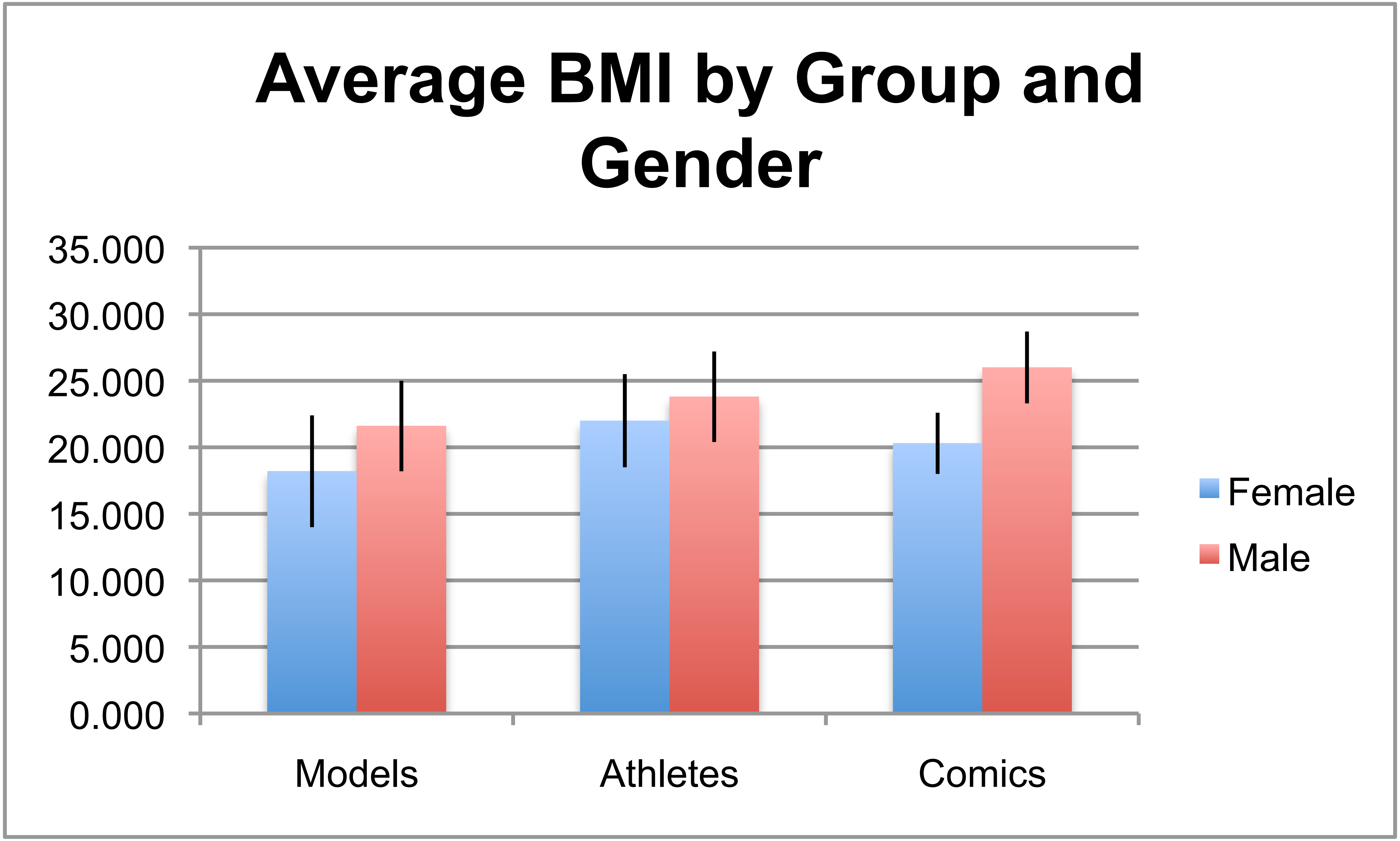
Data Sample, Summary, and Metrics

Name	Weight (lb.)	Height (in.)	BMI
Batman	210	74	27.0
Michael Phelps	165	75	20.6
Wonder Woman	130	72	17.6
Hope Solo	140	69	20.7

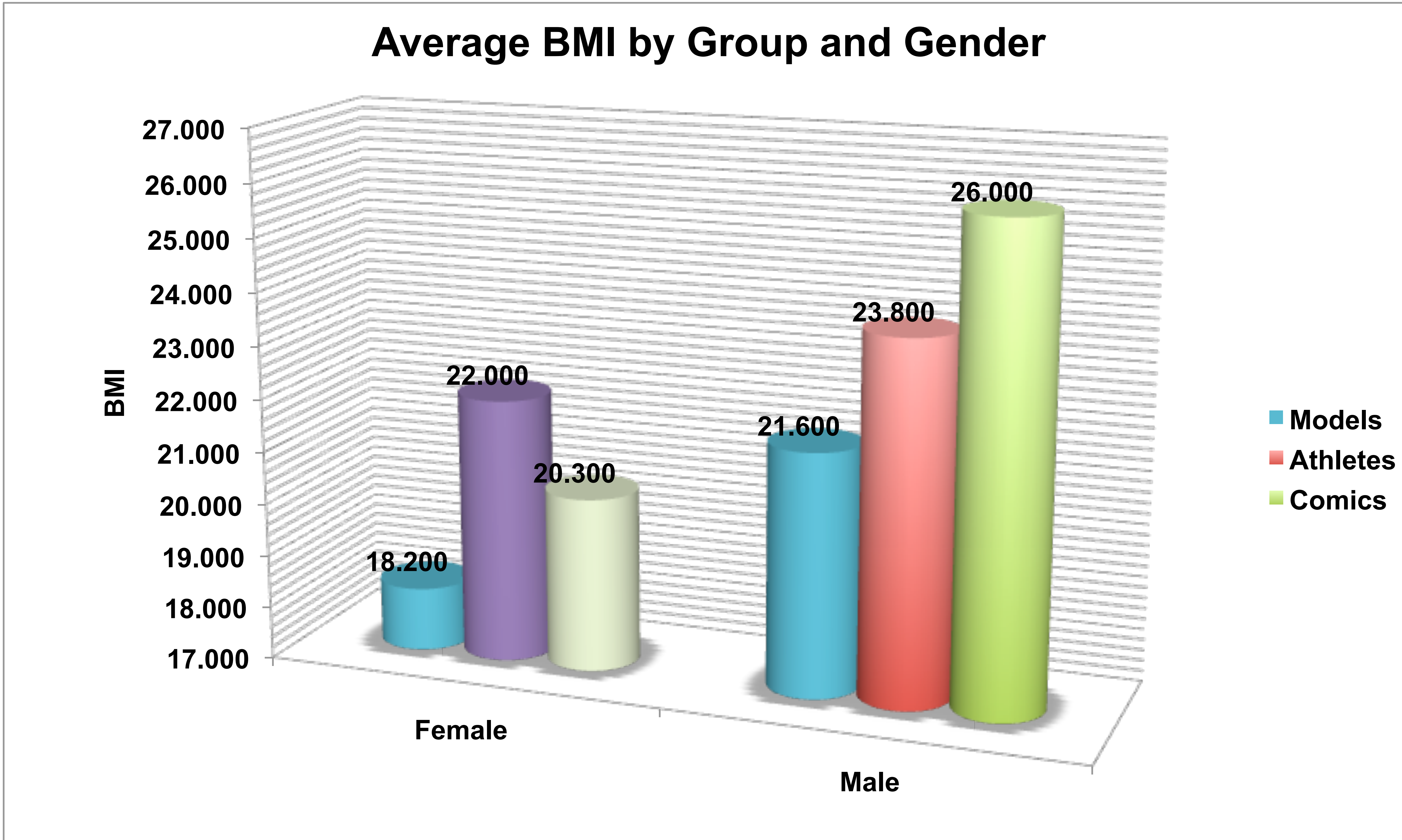
$$\text{BMI} = 703 \times \frac{\text{weight[lb]}}{\text{height[in]}^2}$$

Gender	Group	N	BMI Mean	BMI Std. Dev.
Male	Comics	1,239	26.0	4.2
	Athletes	403	23.8	3.5
	Models	493	21.6	2.3
Female	Comics	505	20.3	3.4
	Athletes	254	22.0	3.4
	Models	489	18.2	2.7

Default Excel Charting

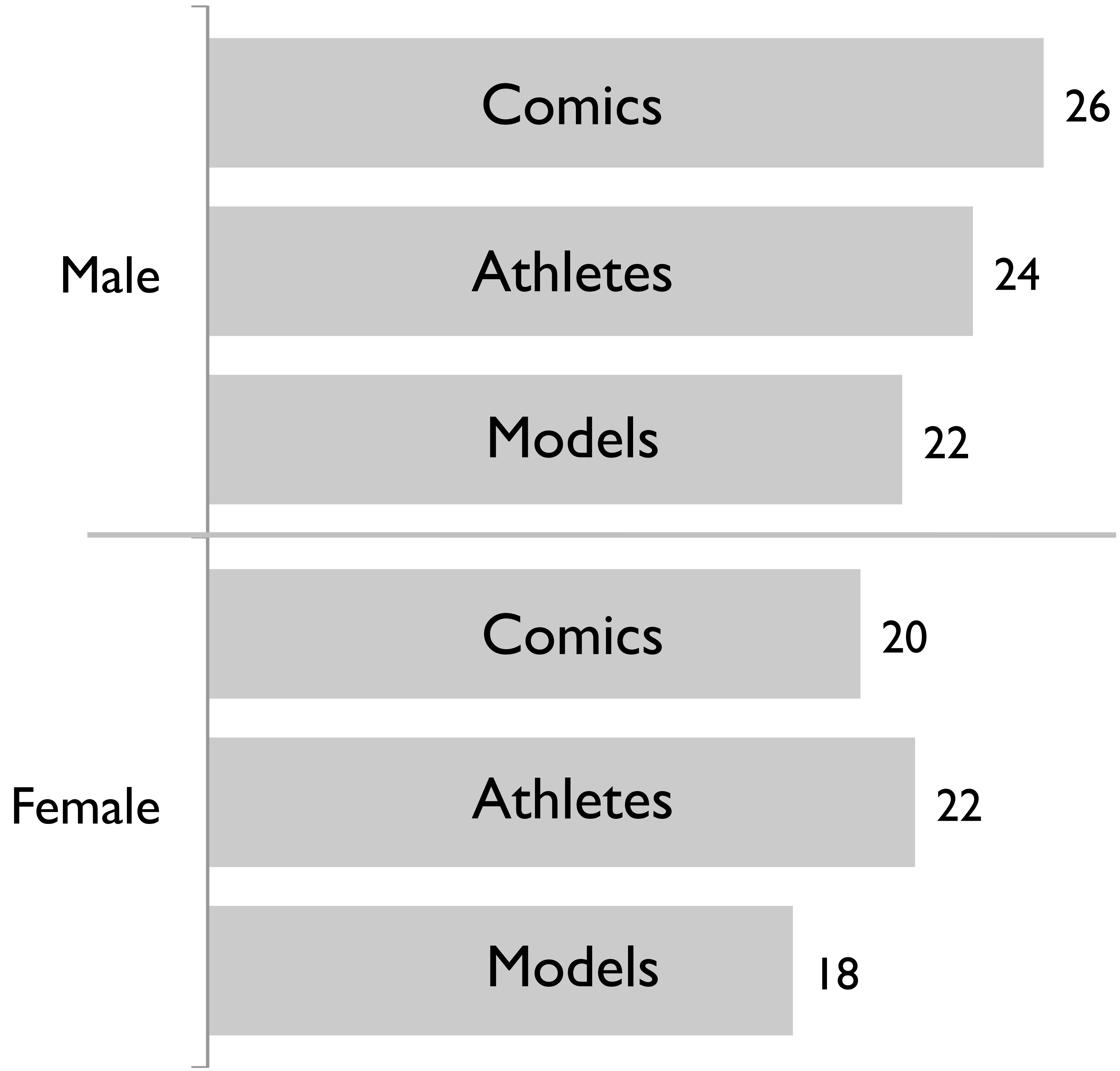


Excessive Chart Junk



Improved Excel Charting

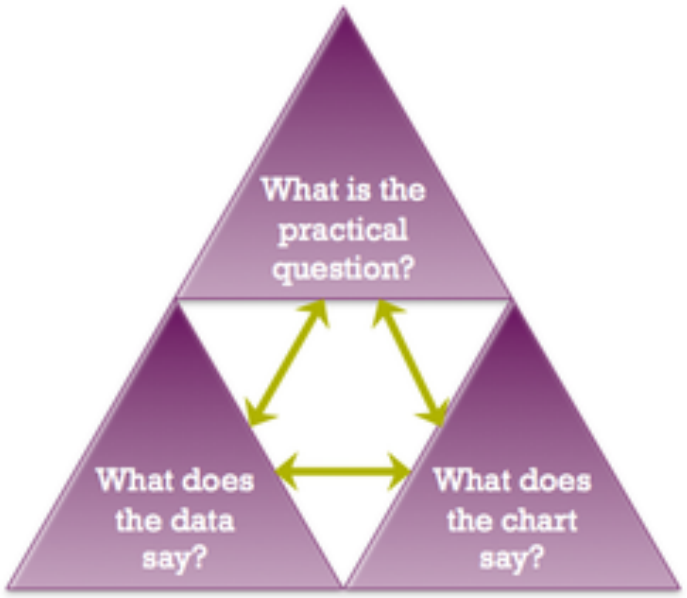
Average BMI by Group and Gender



Visualization Checklist

- What stories does the data tell? Which story do you want to tell?
- What visualization will best aid the story?
  - Who is the audience?
  - What metric should you use?
  - Which type of chart should you use?
  - What is the layout of the visualization?
  - How can details enhance the chart?
    - Font, color, lines/shading, and text

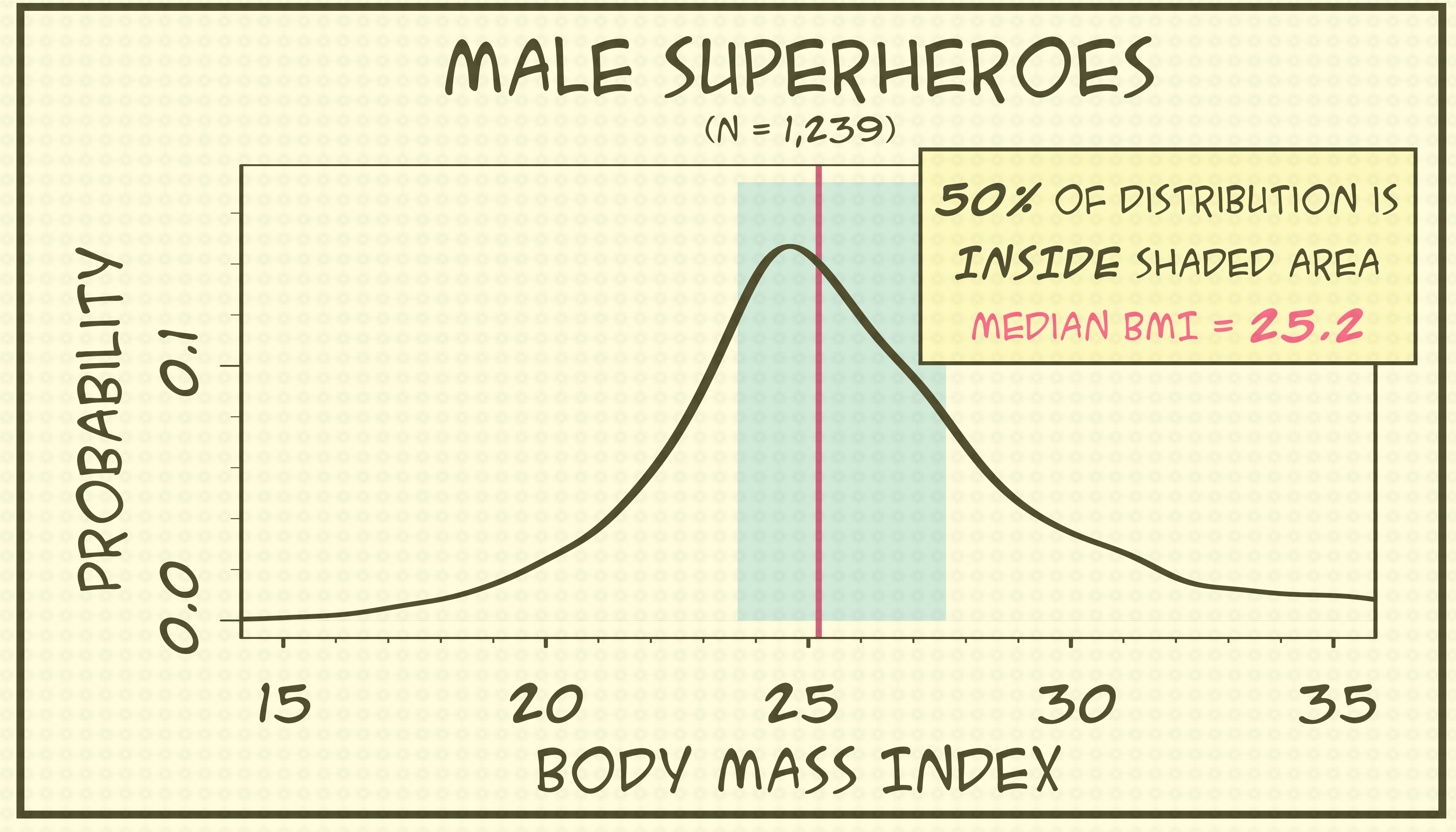
Junk Charts Trifecta Checkup



Source: Kaiser Fung, *Junk Charts*

"What are the content-reasoning tasks that this display is supposed to help with?" -- Edward Tufte, *Beautiful Evidence*

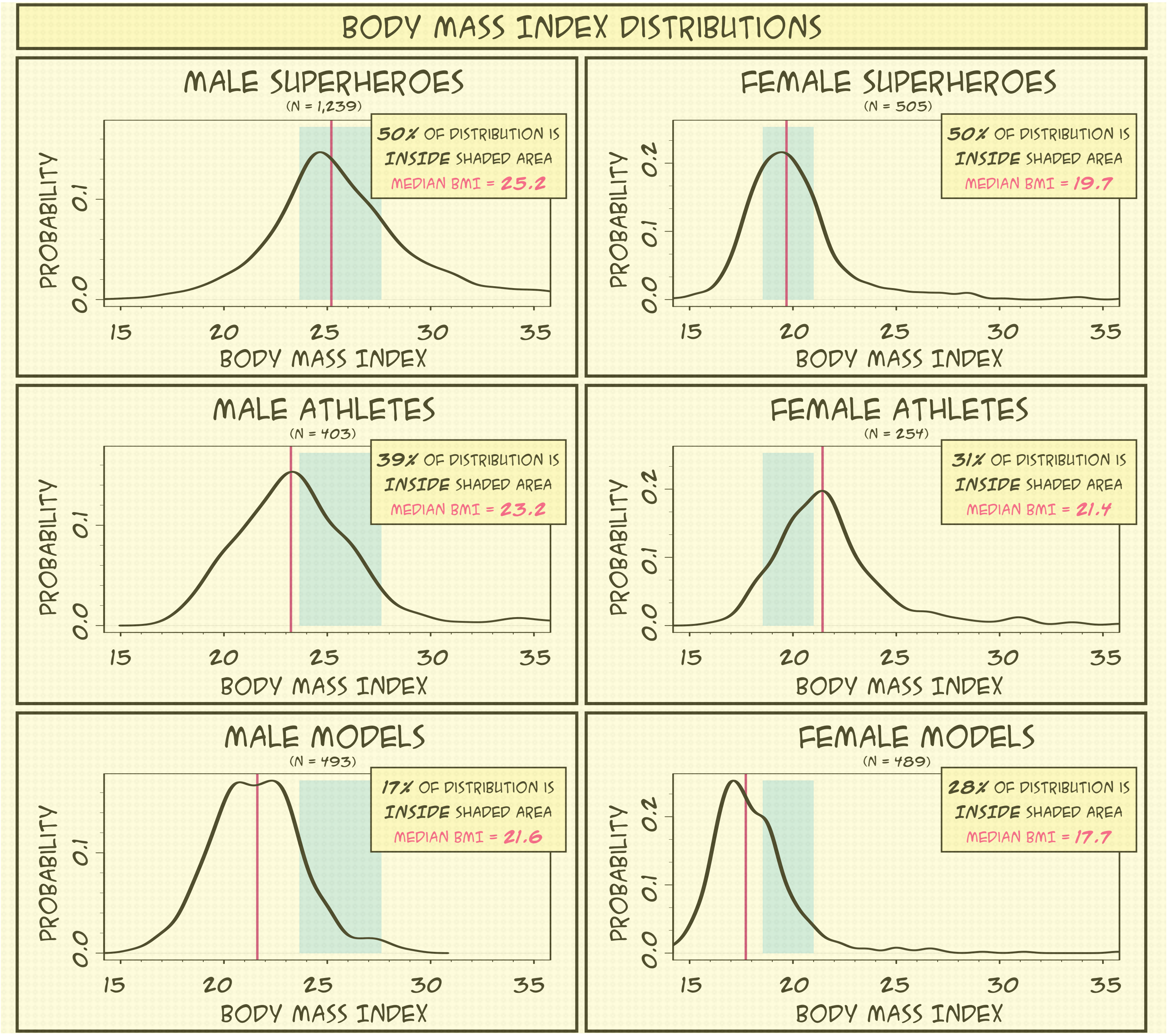
Chart Selection



"Meaningful quantitative information always involves relationships. When displayed in graphs, these relationships always boil down to one or more of eight specific relationships: time series, ranking, part-to-whole, **deviation**, **distribution**, correlation, geospatial, nominal comparison."

-- Stephen Few, *Designing Effective Tables and Graphs*

Visualization Layout



"Tufte's (1990) recommendation of '**small multiples**' [...] uses the replication in the display to **facilitate comparison** to the **implicit model of no change** between the displays."
-- Andrew Gelman, *Exploratory Data Analysis for Complex Models*

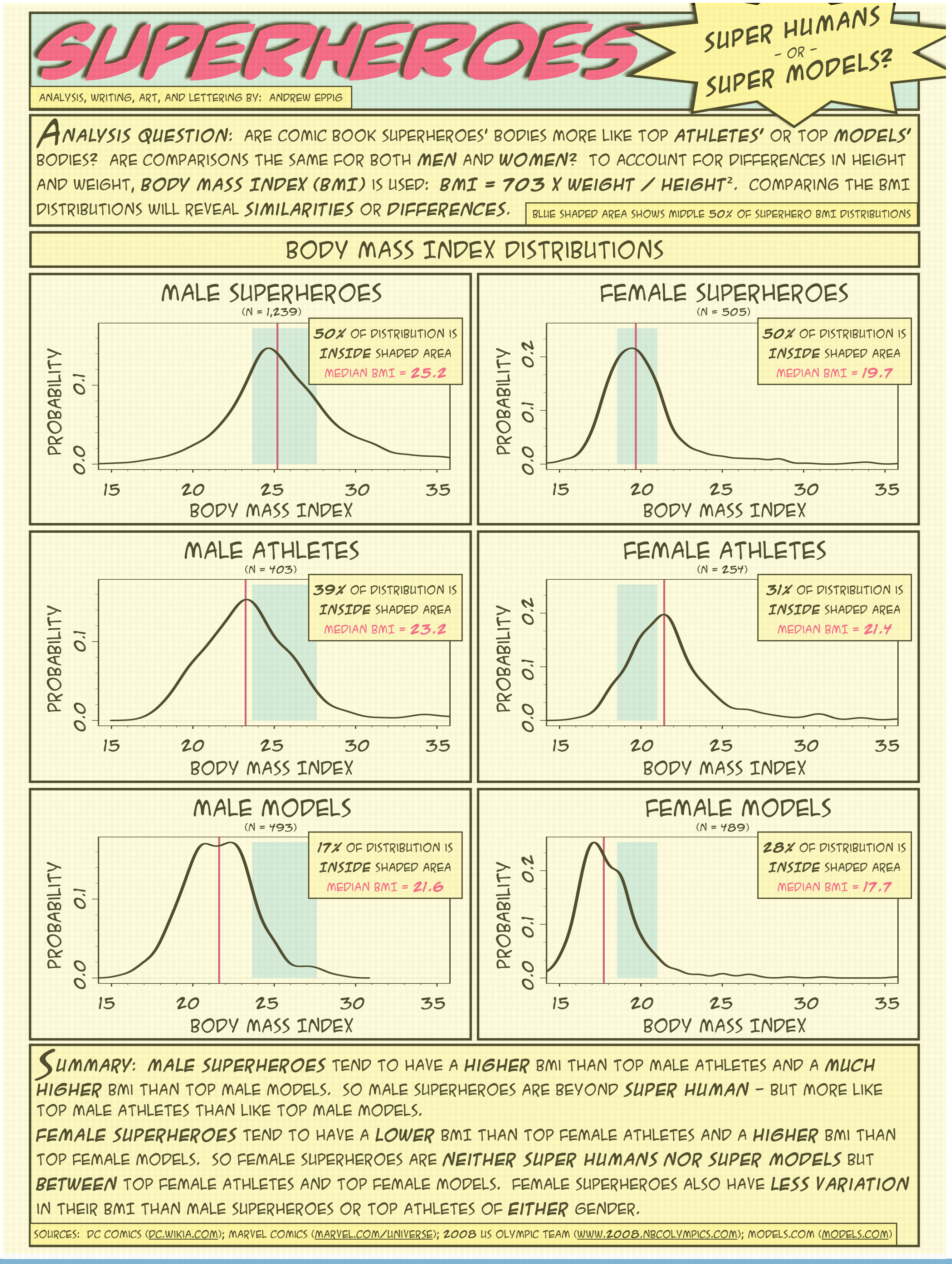
Aesthetic Considerations

- Font: How do you increase legibility and decrease distraction?
- Color: Which color palette is appropriate?
- Line/Shading: Which weight, color, and style will enhance the final product?
- Text: Can adding labels and narrative provide useful context?

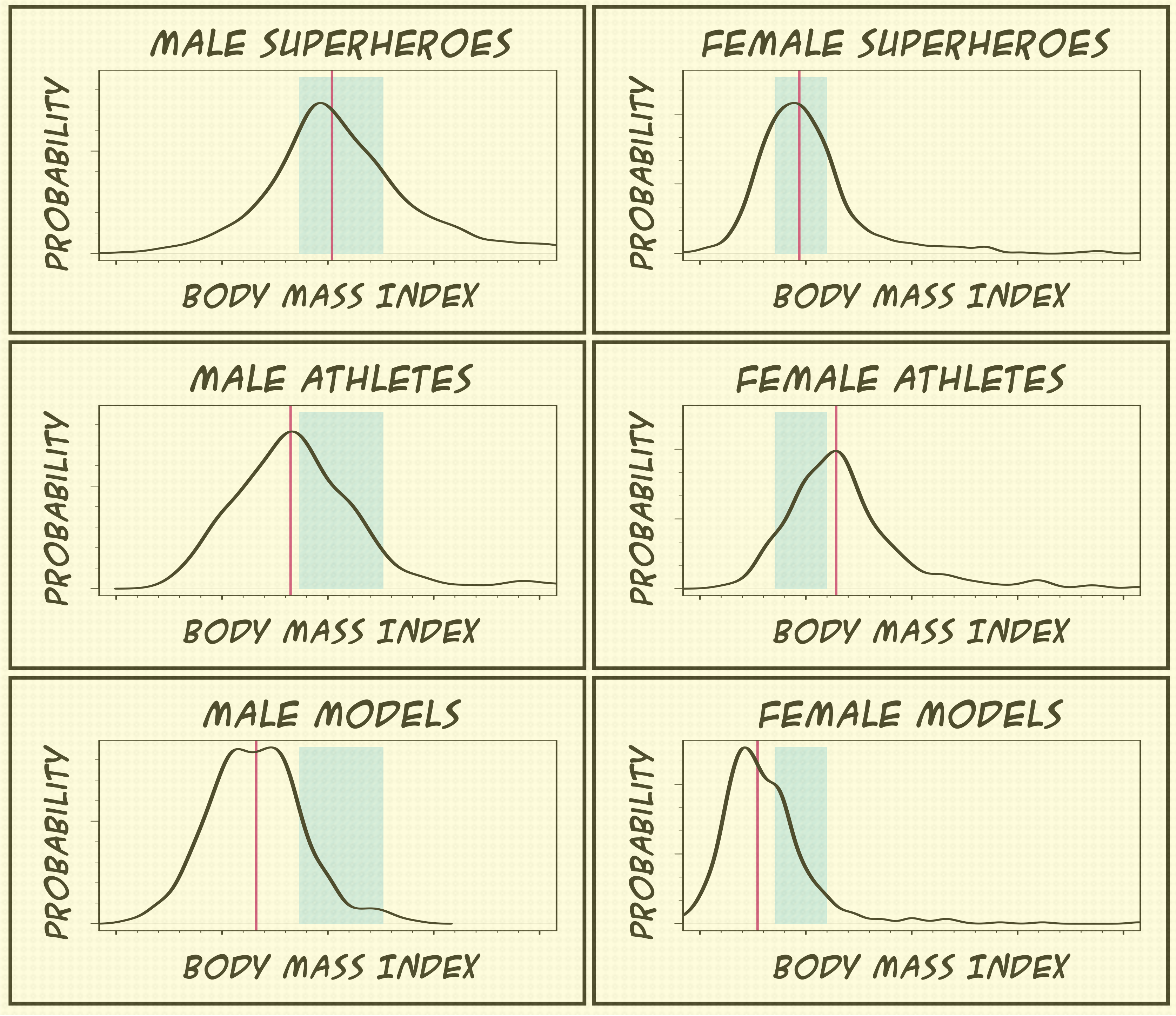
Source: Chemical Color Plate Corp.

"Hue contrast is easy to overuse to the point of visual clutter. A better approach is to use a **few high chroma colors** as color contrast in a presentation consisting **primarily of grays and muted colors**."
-- Maureen Stone, *Choosing Colors for Data Visualization*

Final Infographic



Self-Sufficiency Test



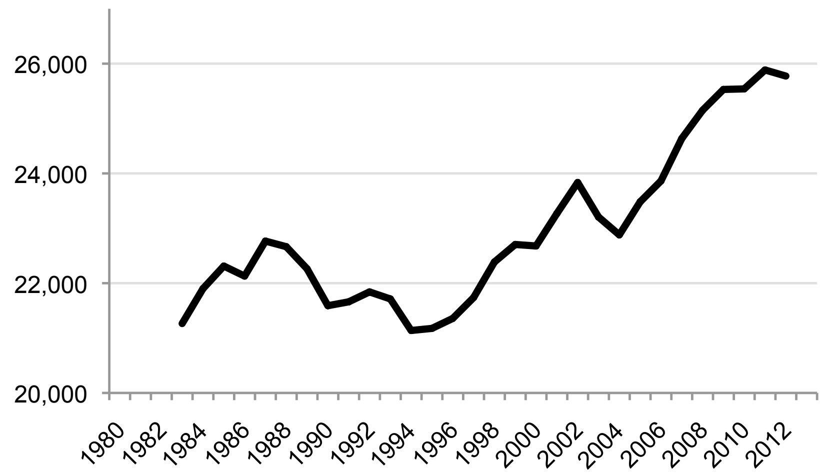
"Can the graphical elements stand on their own feet? If one **removes the numbers** from the graphic, can one **still understand the key messages**?" -- Kaiser Fung, *Junk Charts*

Chart Function and Selection

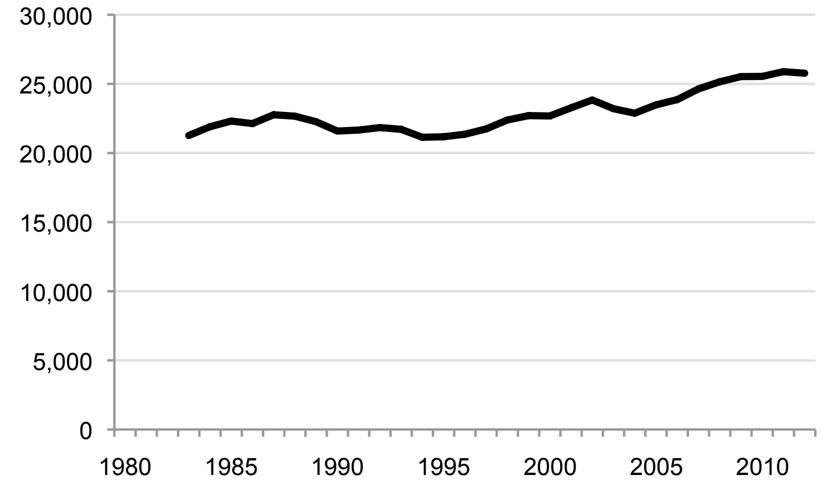
Analytical Relationship	Highlighted Feature
Time Series	Changes over time
Ranking	Relative position
Part-to-Whole	Fraction of whole
Deviation	Differences between sets
Distribution	Range and frequency
Correlation	Relationship between sets
Geospatial	Location
Nominal Comparison	Group values

Source: Stephen Few, *Show Me the Numbers* (2nd edition)

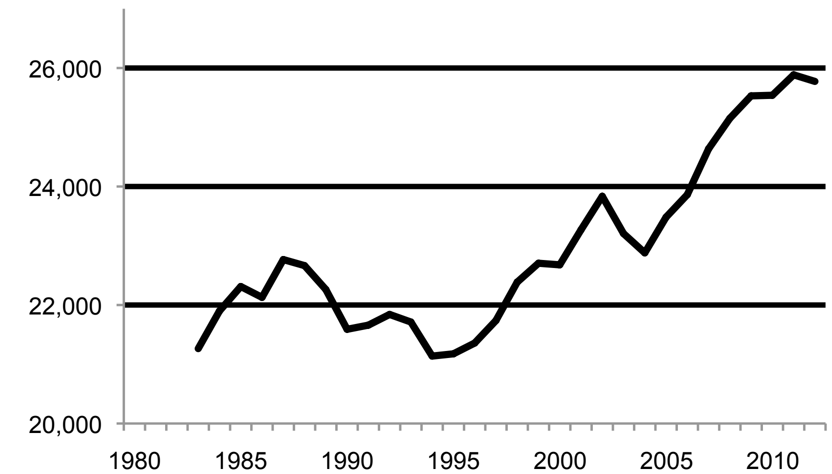
Line Charts - Time Series



✘ x-axis has too many labels and labels are slanted

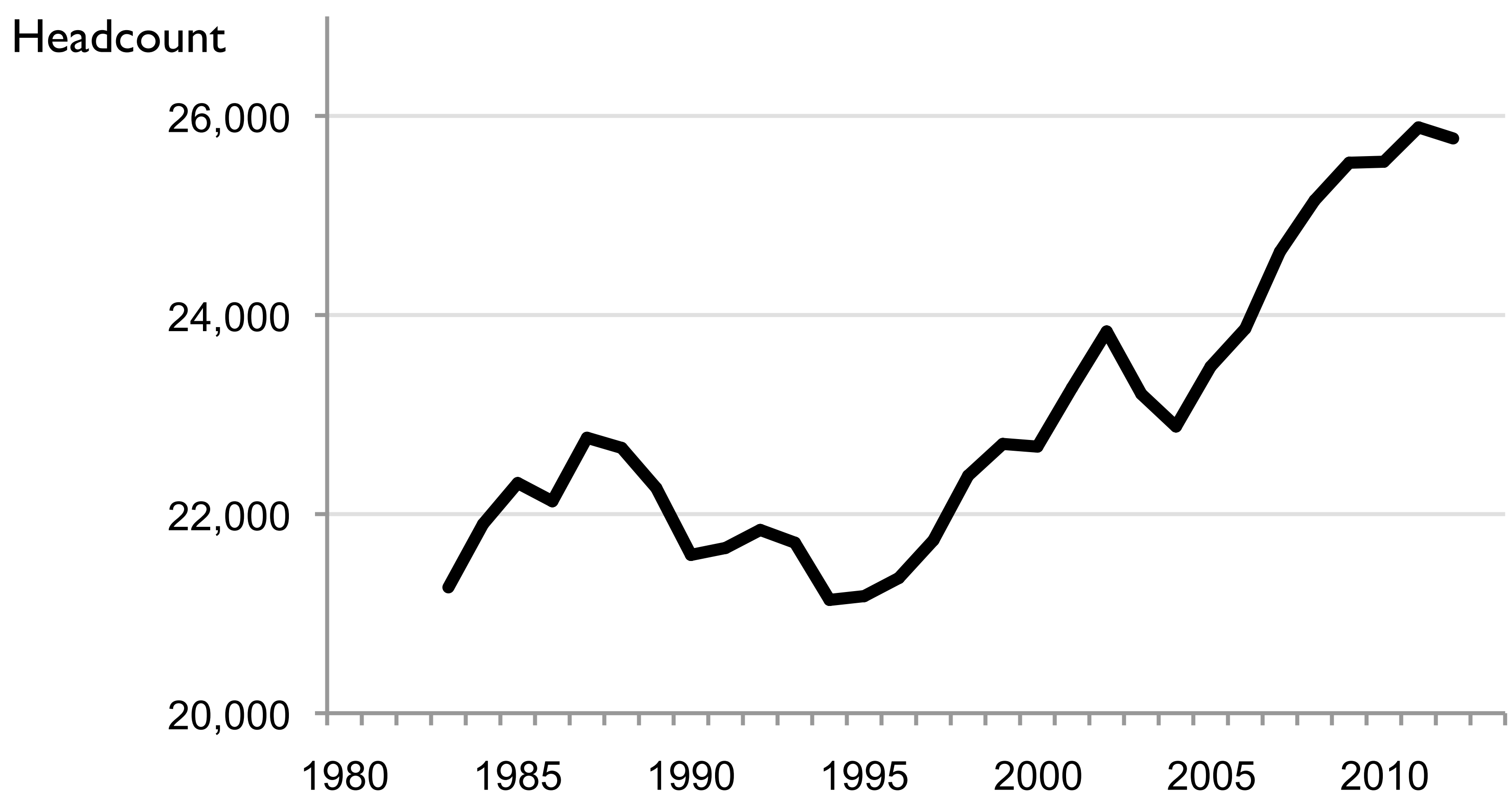


✘ y-axis scale is too large



✘ y-axis gridlines are too heavy

UC Berkeley Undergraduate Fall Enrollment, 1983-2012



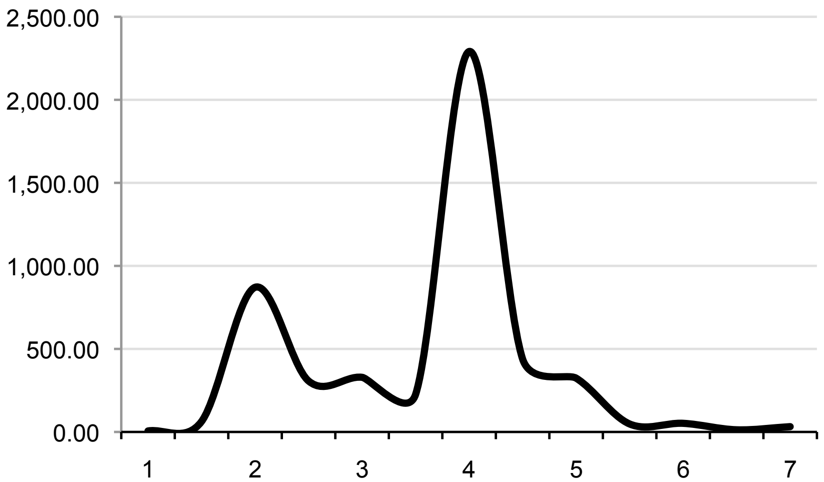
✓ x-axis labels are horizontal and legibly spaced

✓ data range is roughly 2/3 of the y-axis scale

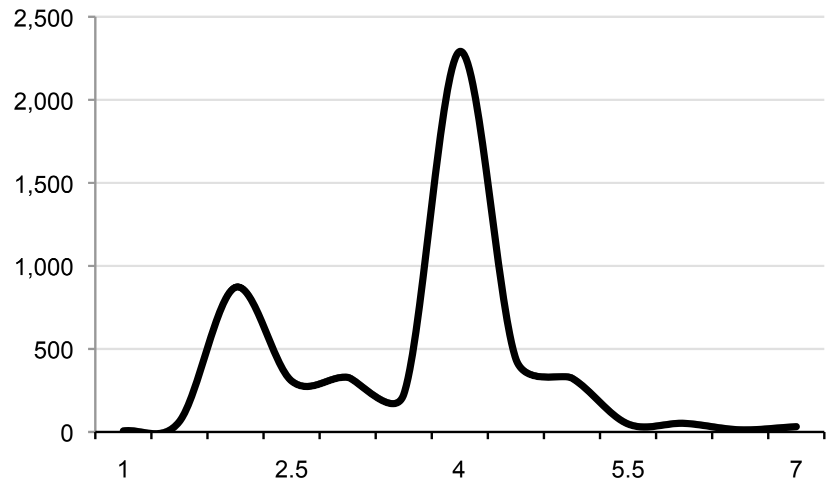
✓ y-axis gridlines are light in color and weight

Source: UC Berkeley, Cal Answers

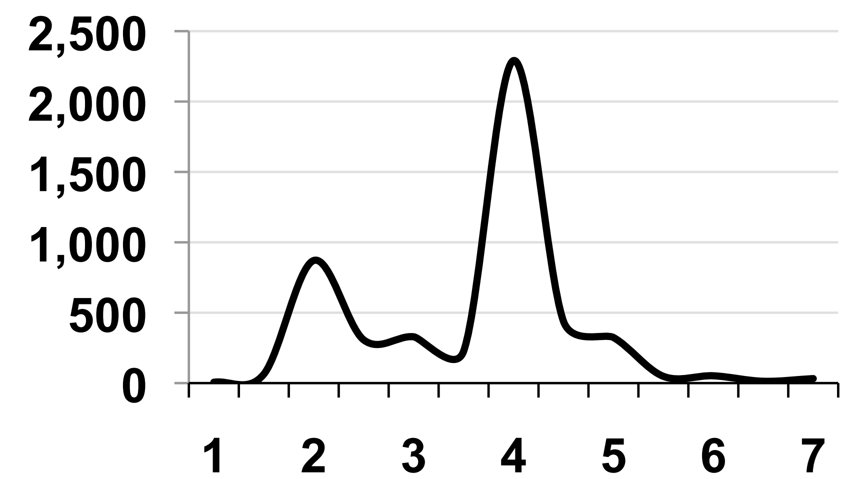
Line Charts - Distribution



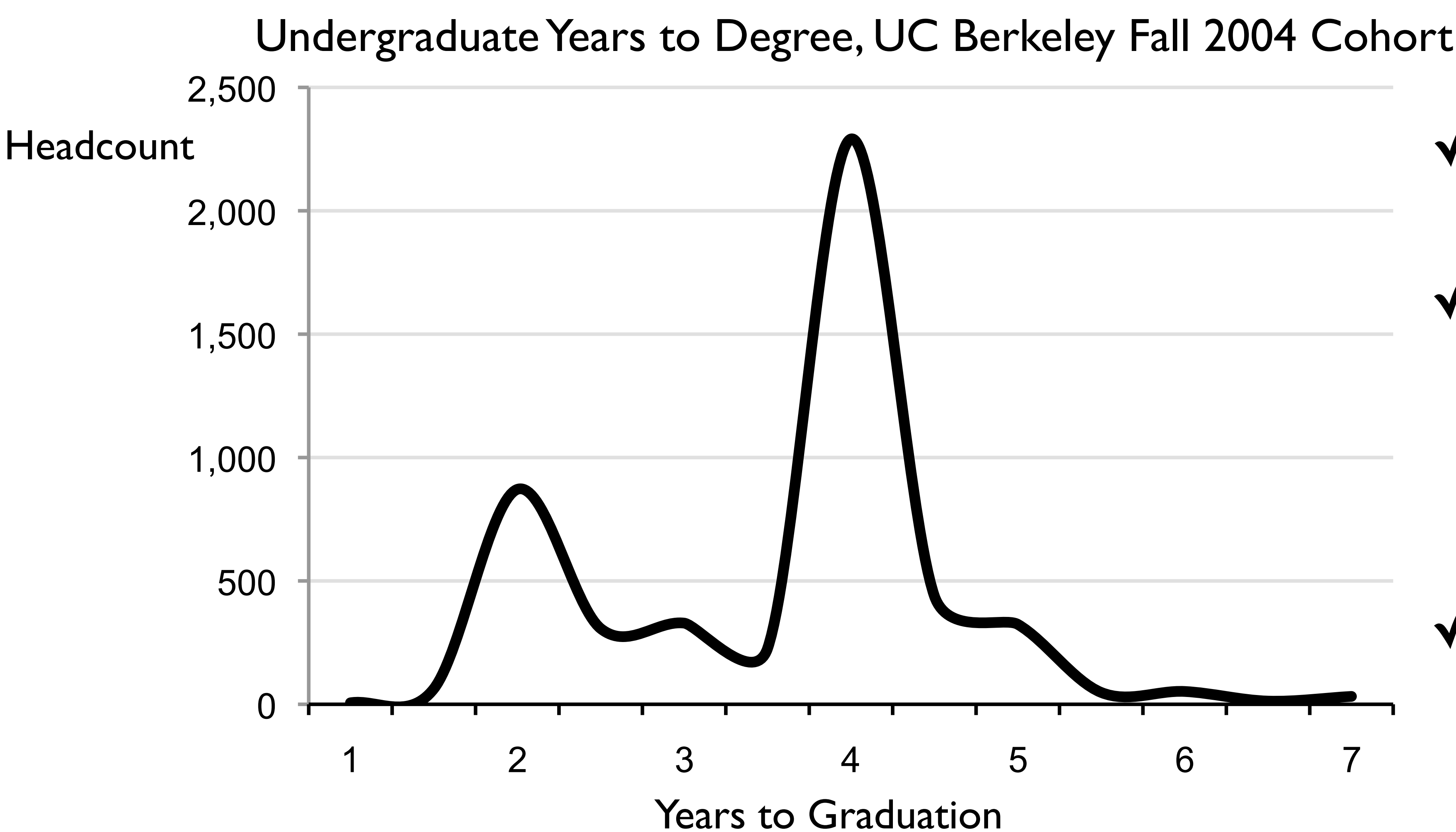
✘ y-axis labels have extraneous decimal points



✘ x-axis labels use an unintuitive interval



✘ axis labels are too heavy



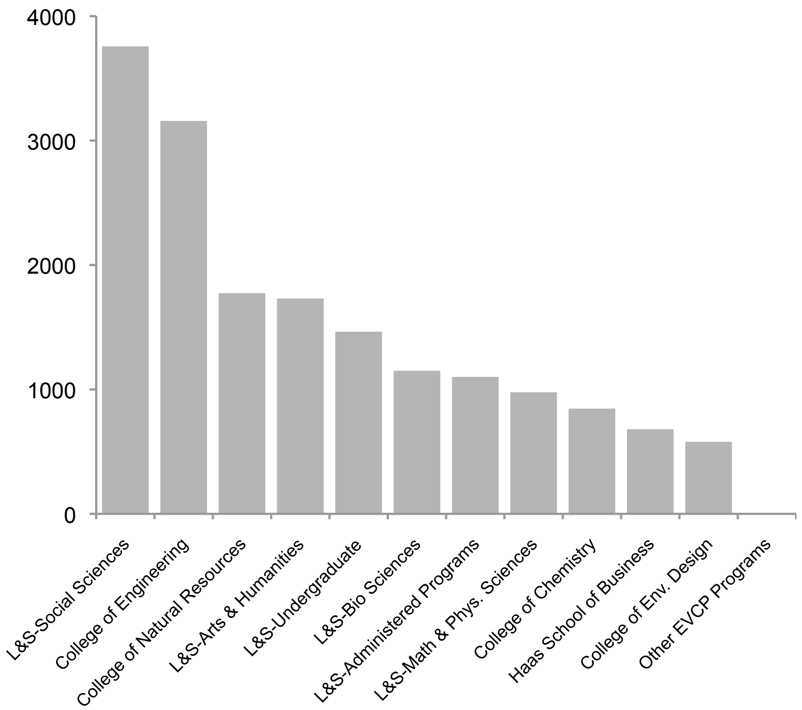
✓ y-axis labels are rounded to the nearest major increment

✓ axis labels use natural intervals
1, 2, 3...
2, 4, 6...
10, 20, 30...

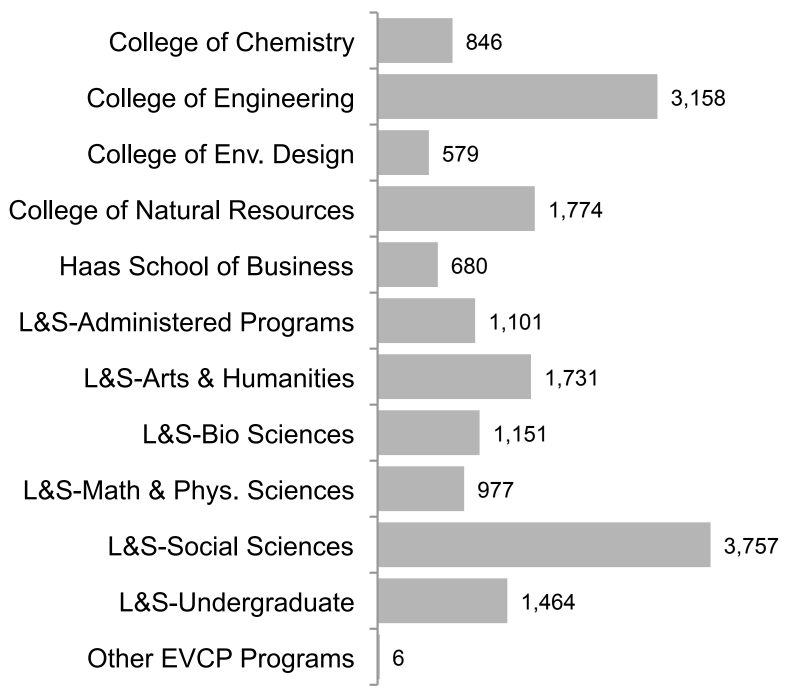
✓ x-axis labels are light in size and weight

Source: UC Berkeley, Cal Answers

Bar Charts - Ranking

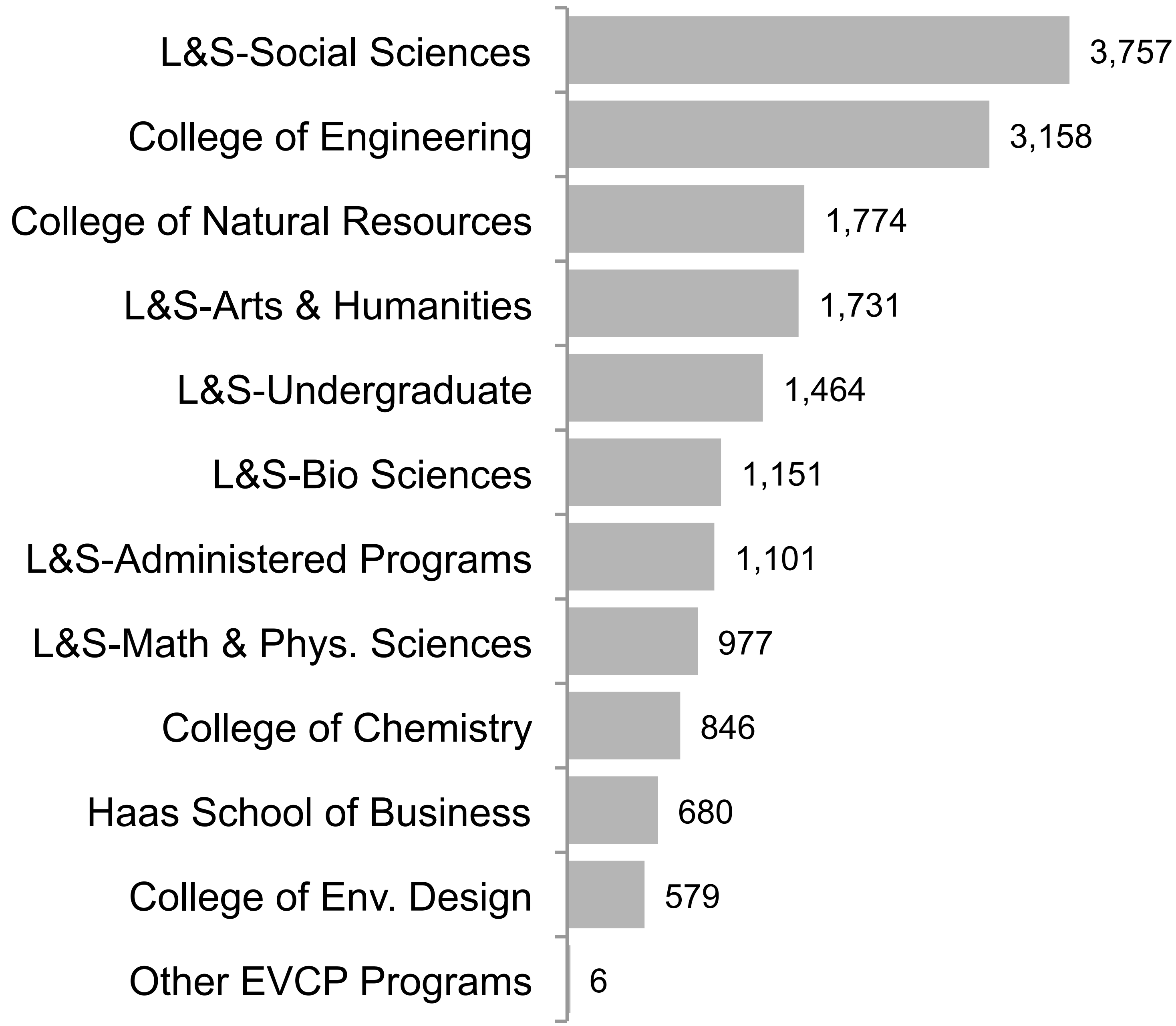


✘ x-axis labels are slanted and small



✘ units are ranked alphabetically

Undergraduate Major Headcount by Division,
UC Berkeley Fall 2012

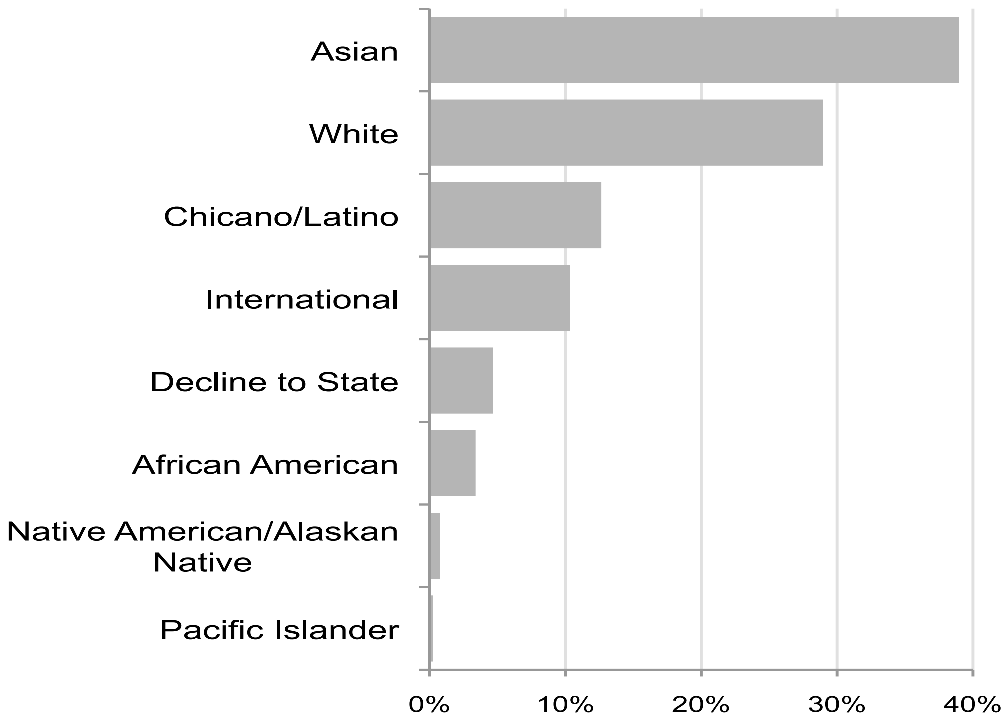


✓ labels are horizontal and easy to read

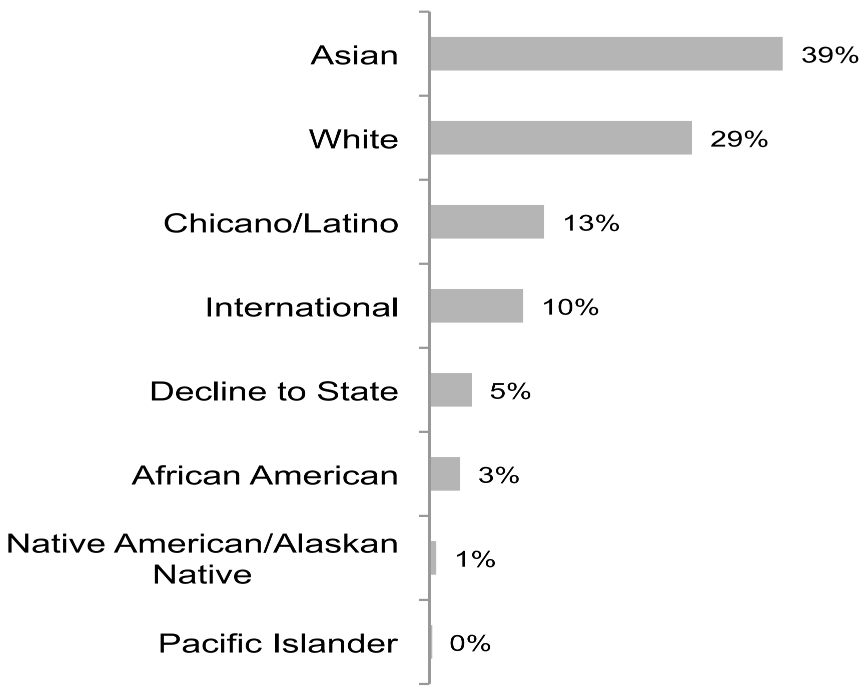
✓ units are usefully ranked by descending value

Source: UC Berkeley, Cal Answers

Bar Charts - Part-to-Whole

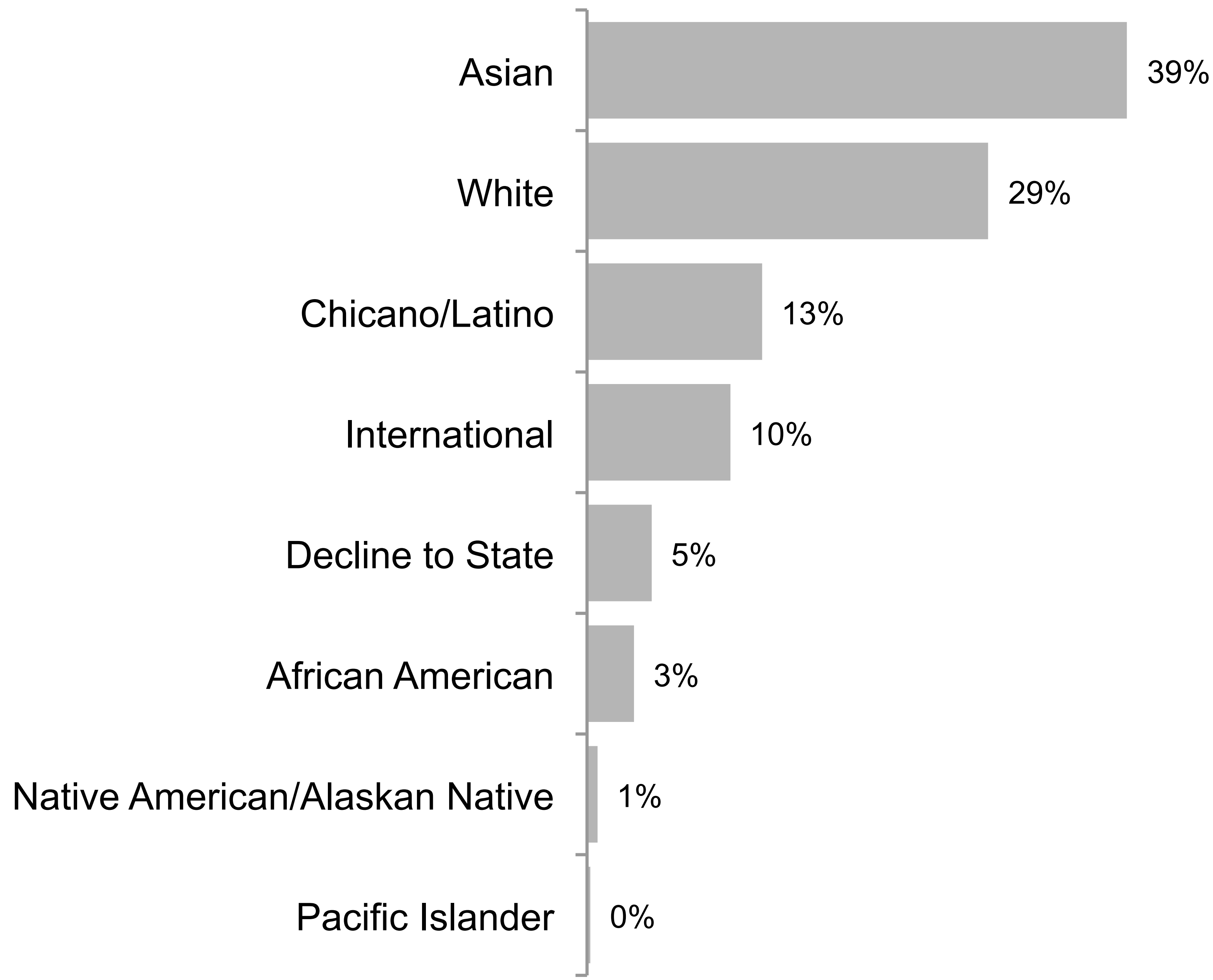


✗ values are hard to determine



✗ gaps between bars are too large

Undergraduate Demographic Shares by Race/ Ethnicity, UC Berkeley Fall 2012

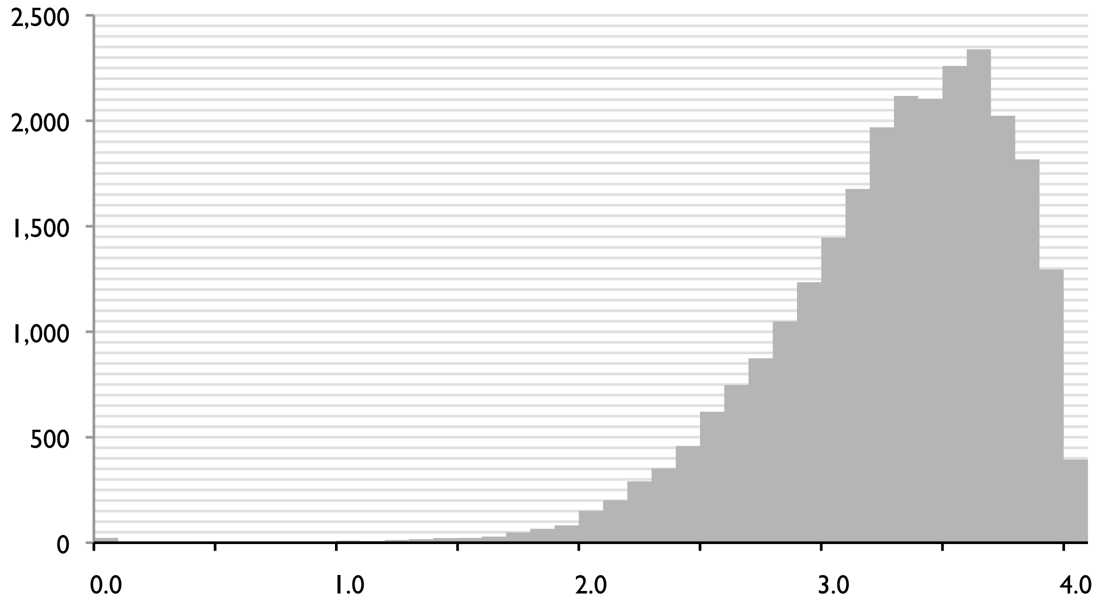


✓ values are easy to read

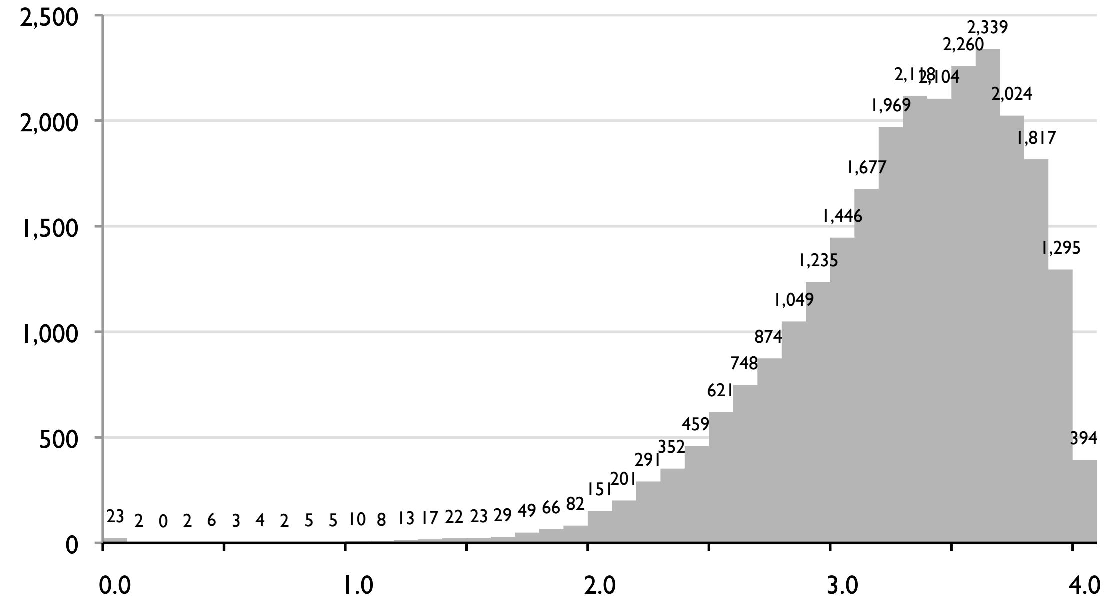
✓ gaps between bars are 25-50% of the bar width

Source: UC Berkeley, Cal Answers

Bar Charts - Histogram

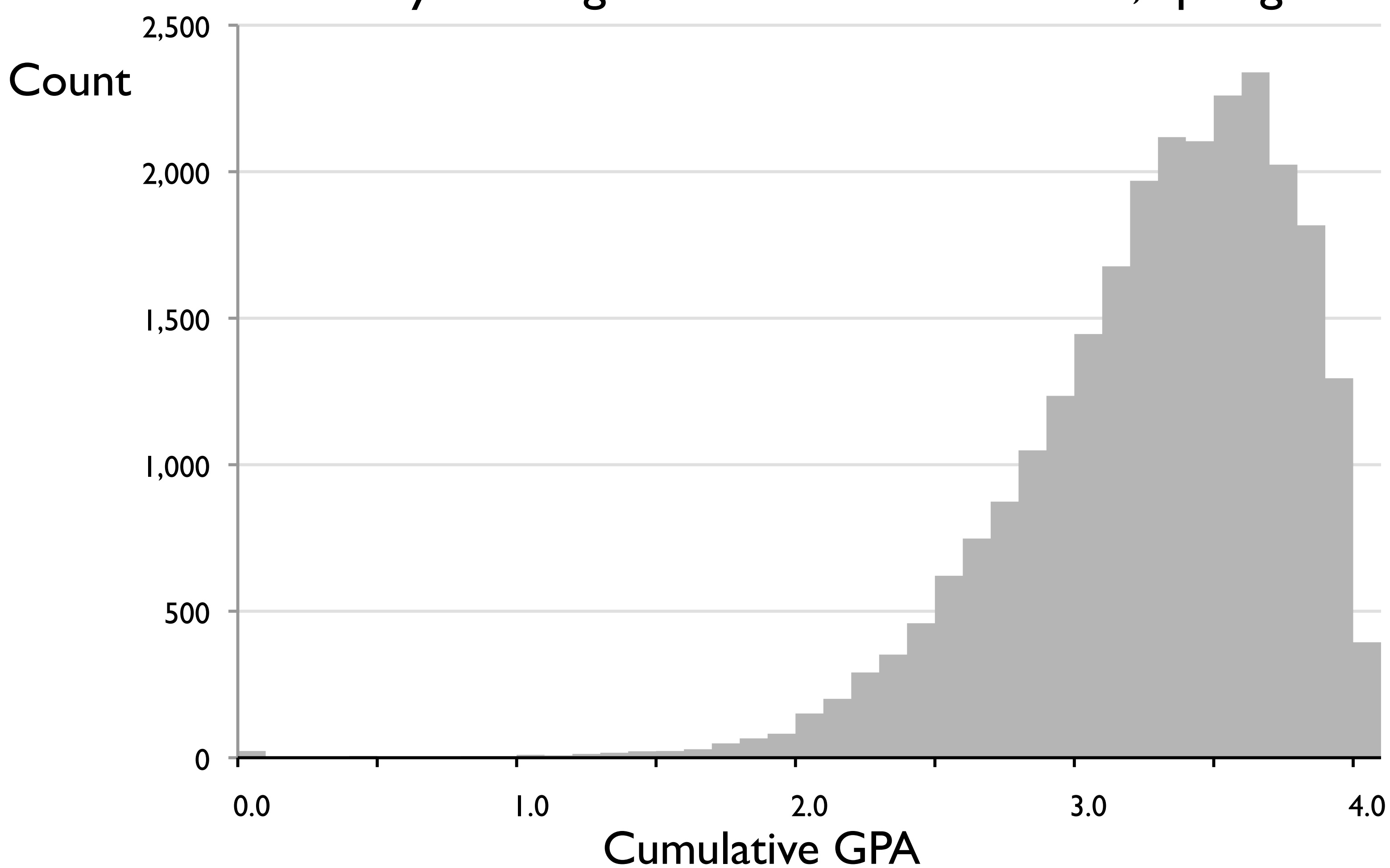


✗ y-axis gridlines are too dense



✗ data labels distract, add clutter

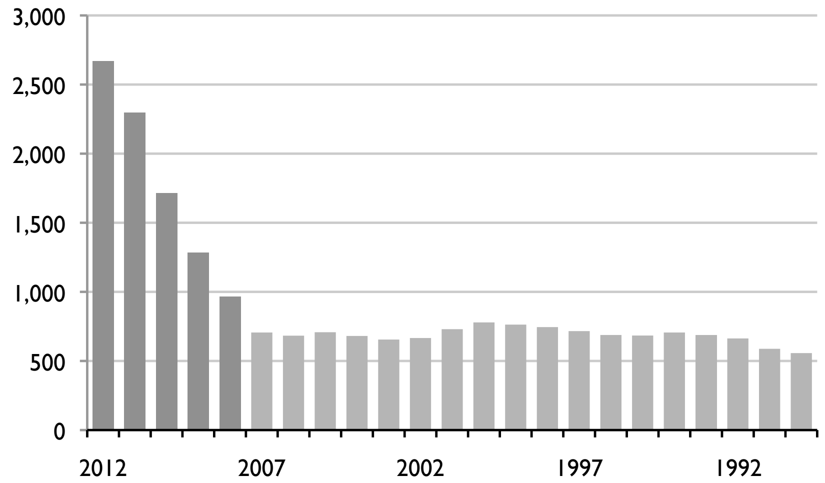
UC Berkeley Undergraduate Cumulative GPA, Spring 2012



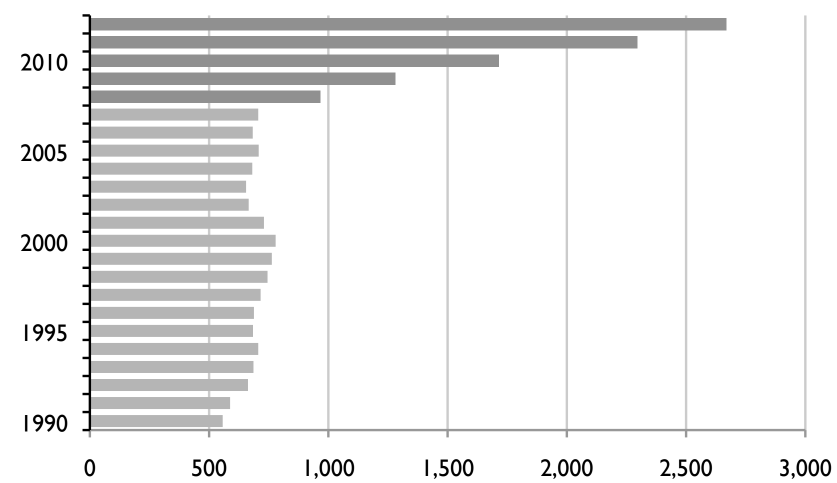
- ✓ y-axis gridlines are only on the major increments
- ✓ shape of the distribution is easily seen without distraction

Source: UC Berkeley, Cal Answers

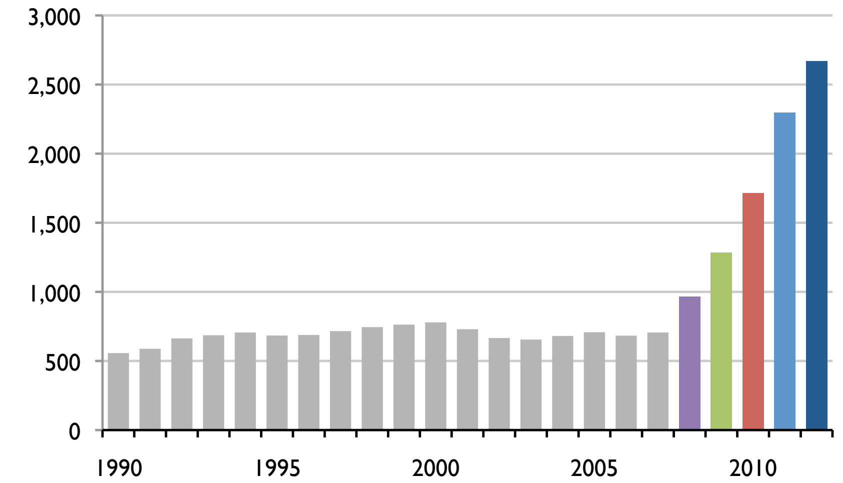
Bar Charts - Time Series



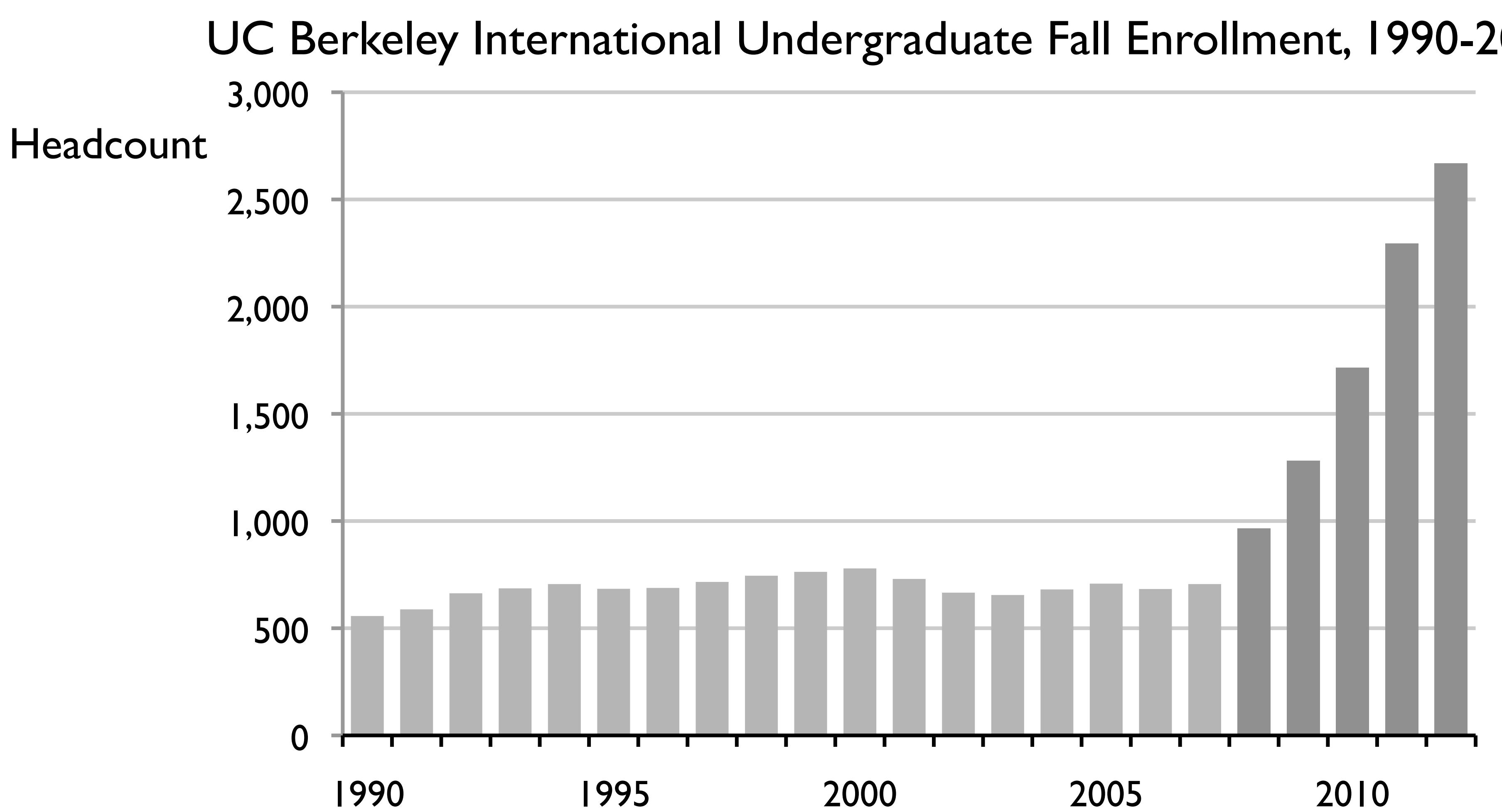
✘ time variable is shown from right to left



✘ y-axis is used for time variable



✘ multiple hues are used for the same kind of data



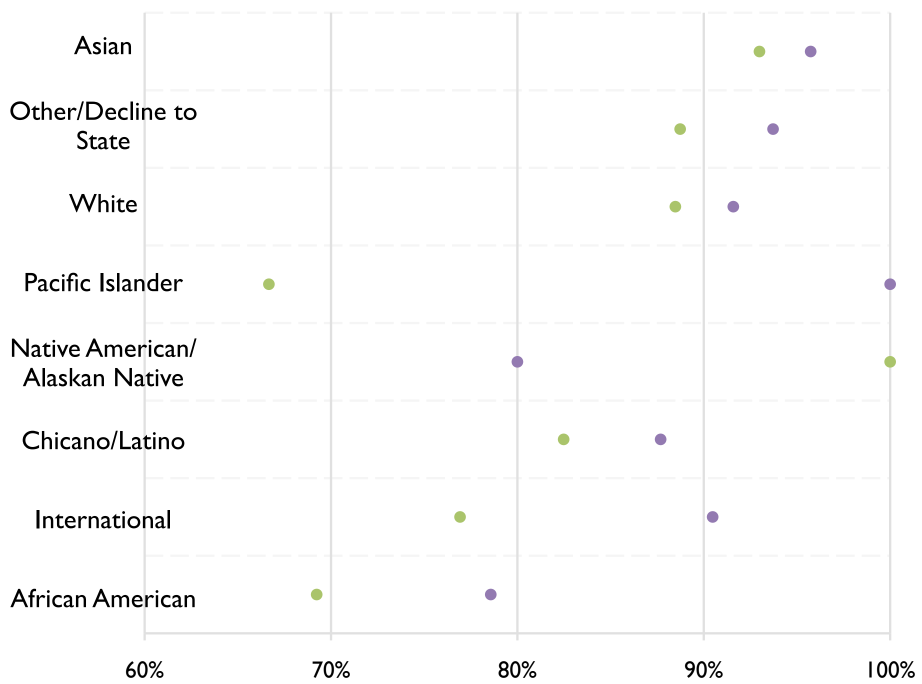
✓ years increase from left to right

✓ years are plotted on x-axis

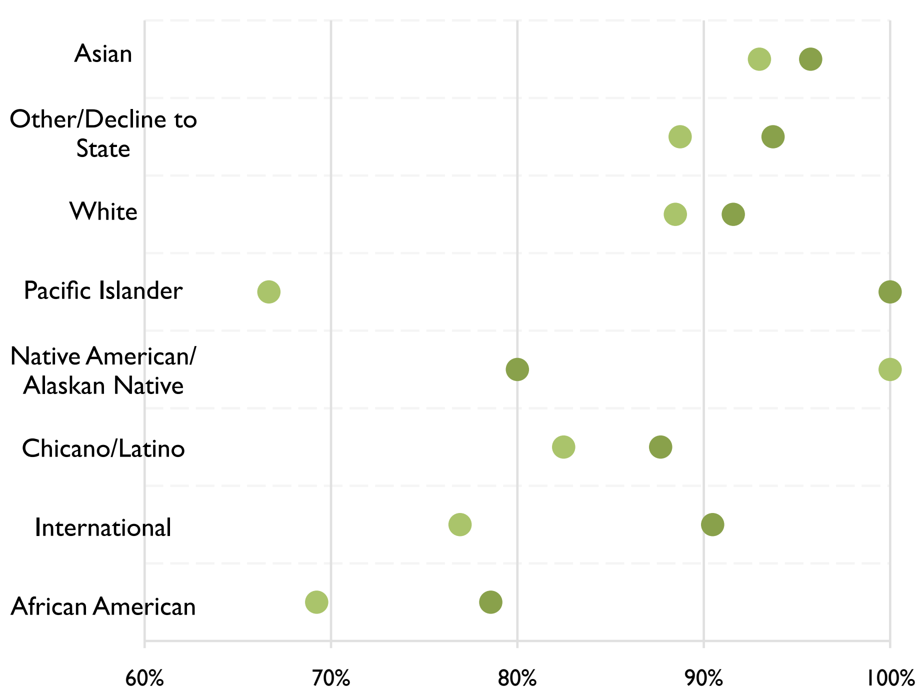
✓ bars are colored using shades of a single hue

Source: UC Berkeley, Cal Answers

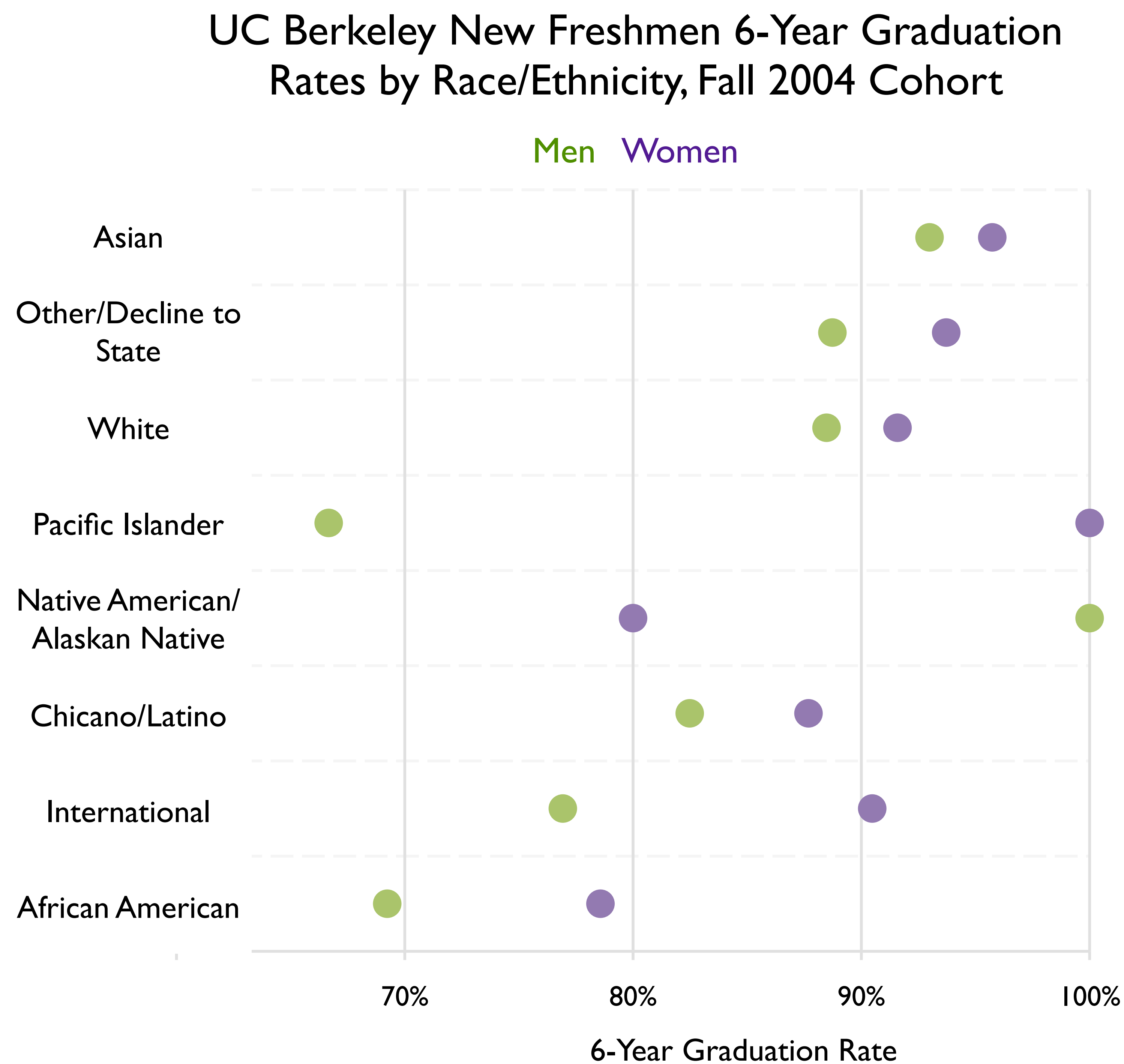
Dot Plots - Deviation



✘ points are too small



✘ group colors are too similar

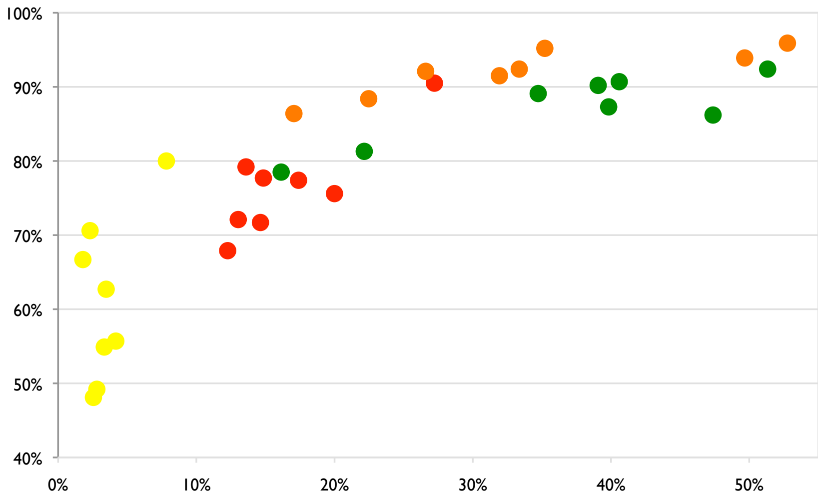


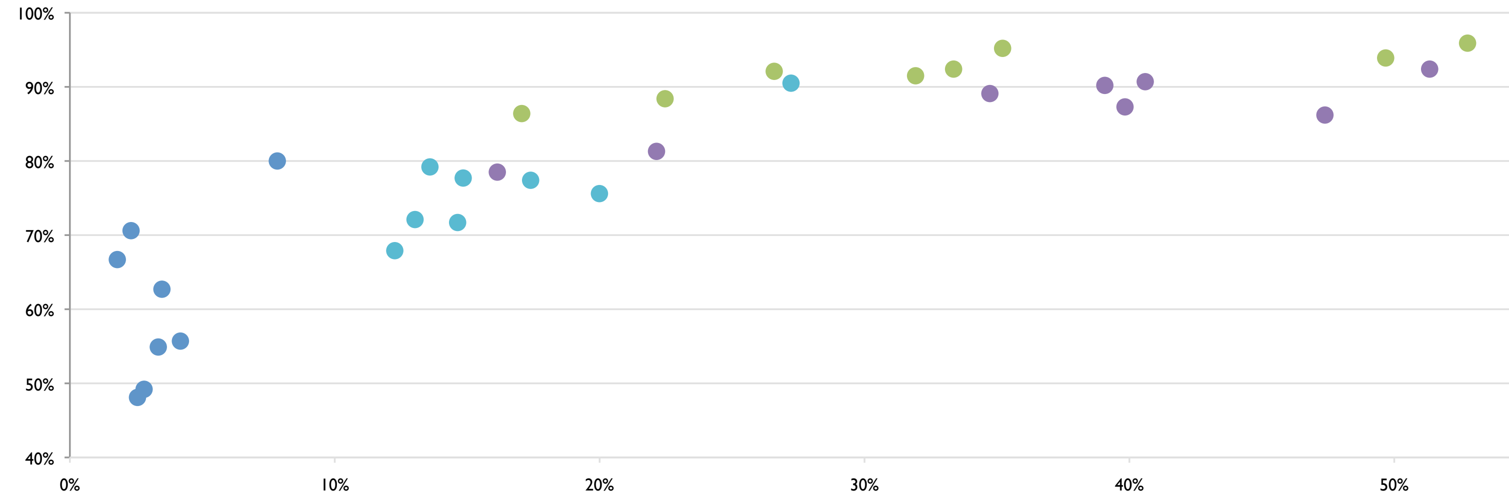
✓ dots' size makes them easy to see

✓ group colors are complementary

Source: UC Berkeley, Cal Answers

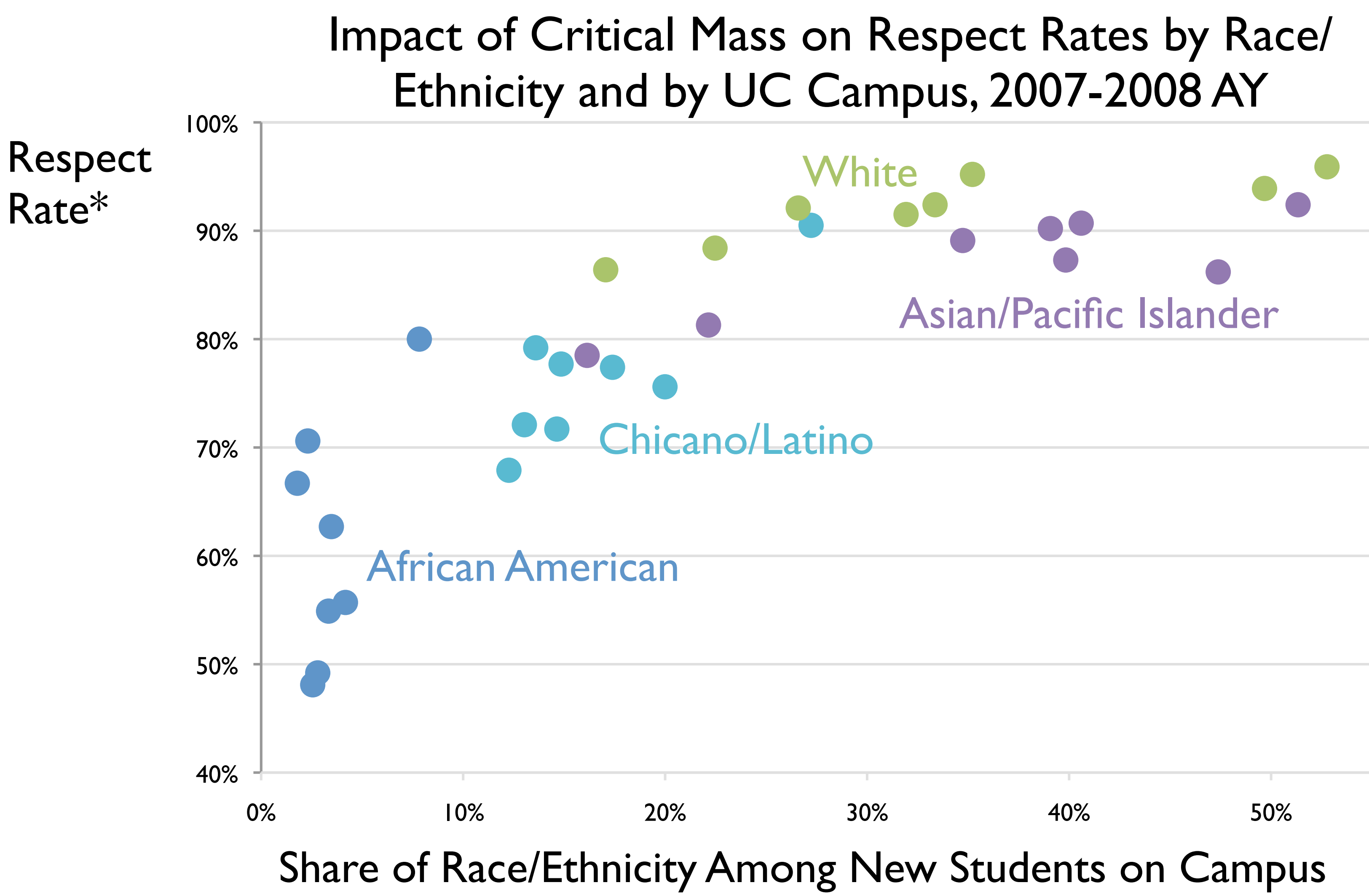
Scatter Plots - Correlation





✘ too many bright hues are used to mark groups

✘ chart length distorts the correlation of the data



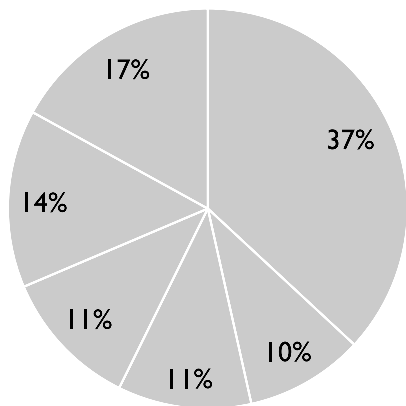
✓ groups are marked using muted hues

✓ chart dimensions are close to the golden ratio (1:1.618)

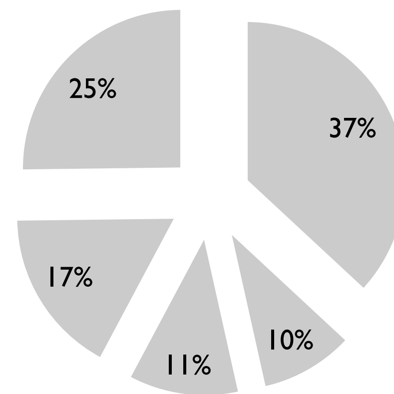
* Respect Rate = percentage of students of a given race/ethnicity who responded *strongly agree*, *agree*, or *somewhat agree* to the prompt "students of my race/ethnicity are respected at this campus" on UCUES.

Source: UC Accountability Report, 2011

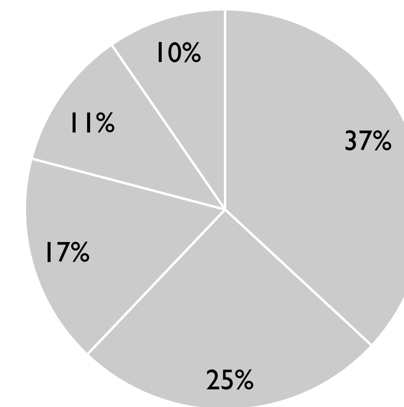
Pie Charts - Part-to-Whole



✘ more than five groups are shown

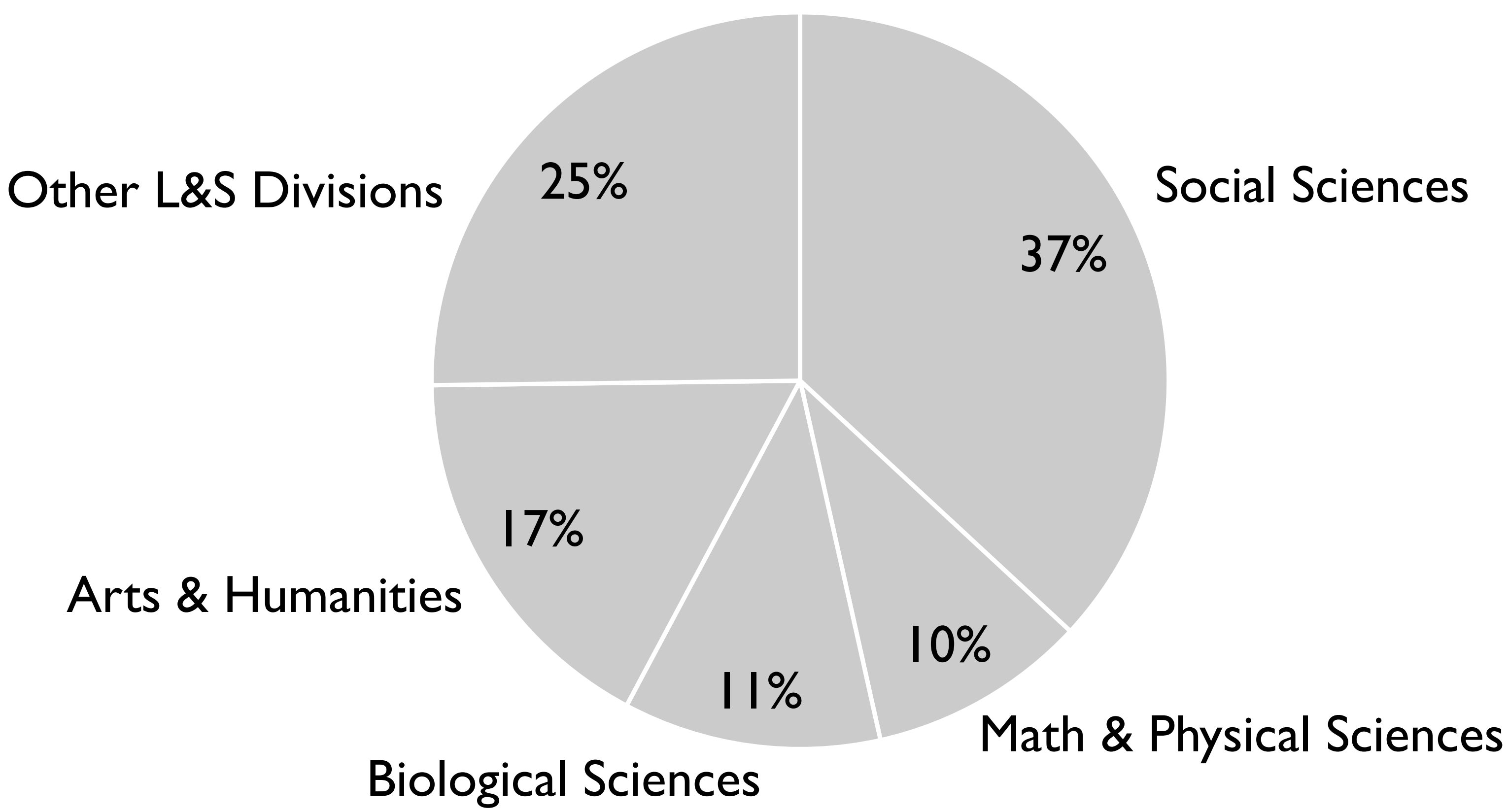


✘ slices are blown out of the pie



✘ smallest slices are given too prominent location

UC Berkeley College of Letters & Sciences
Enrollment Shares, Fall 2012



✓ Only five groups are shown with the rest aggregated together

✓ center of the pie is shown making angles visible

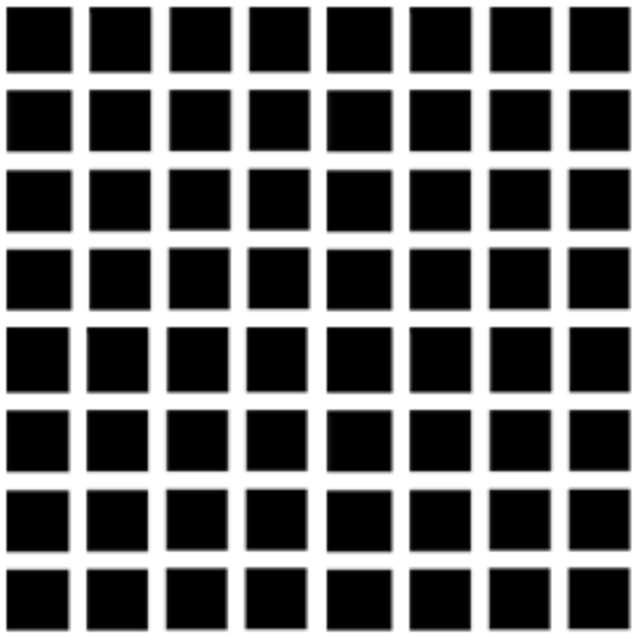
✓ groups are ordered usefully with the largest slices at the top

Source: UC Berkeley, Cal Answers

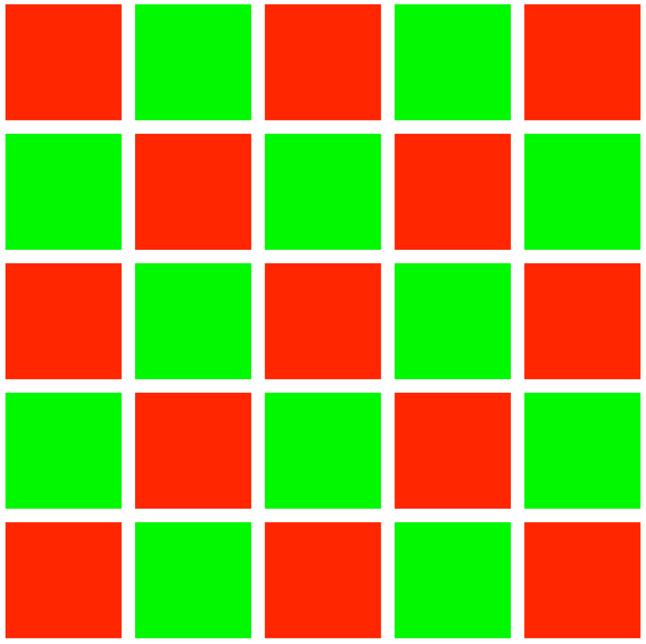
Visualization Layout: Attention Areas

High Visual Focus Good for primary content	Medium Visual Focus Good for secondary content
Medium Visual Focus Good for secondary content	Low Visual Focus Good for tertiary content

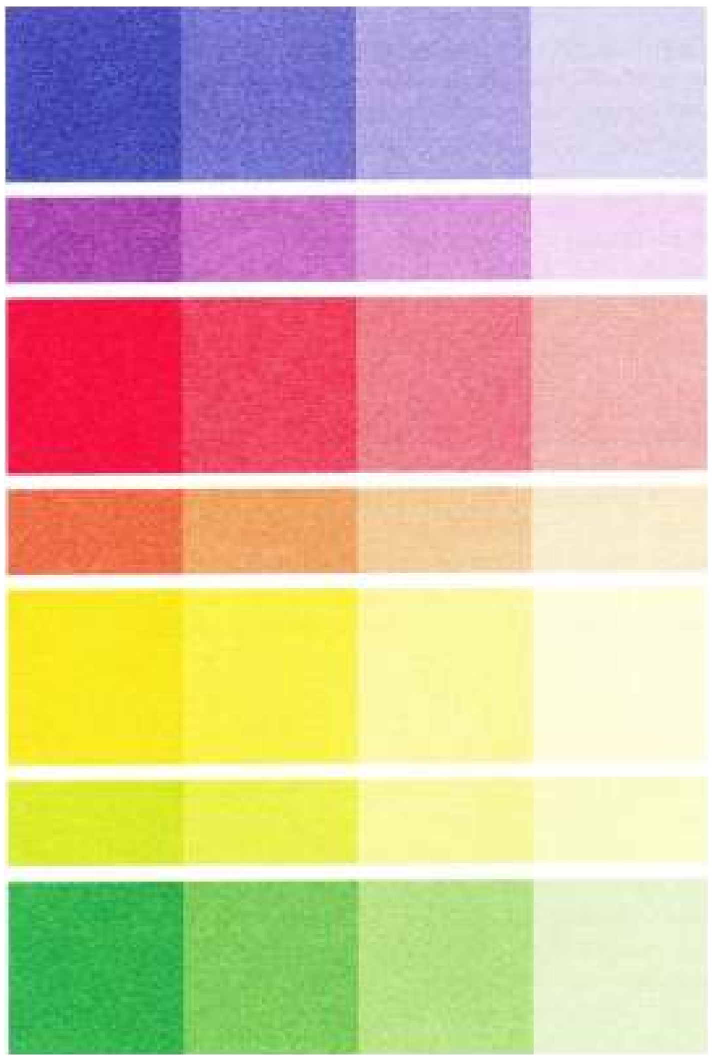
Visualization Aesthetics: Color



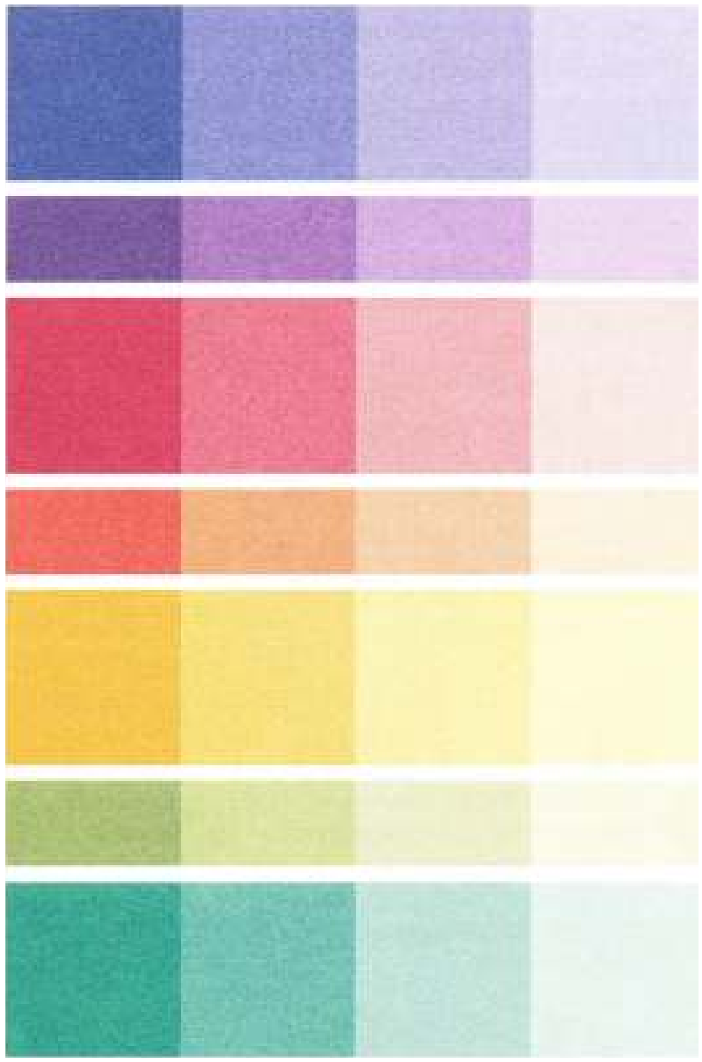
✘ avoid alternating high contrast hues



✘ avoid using more than one high chroma hue



Bright (high chroma)



Muted

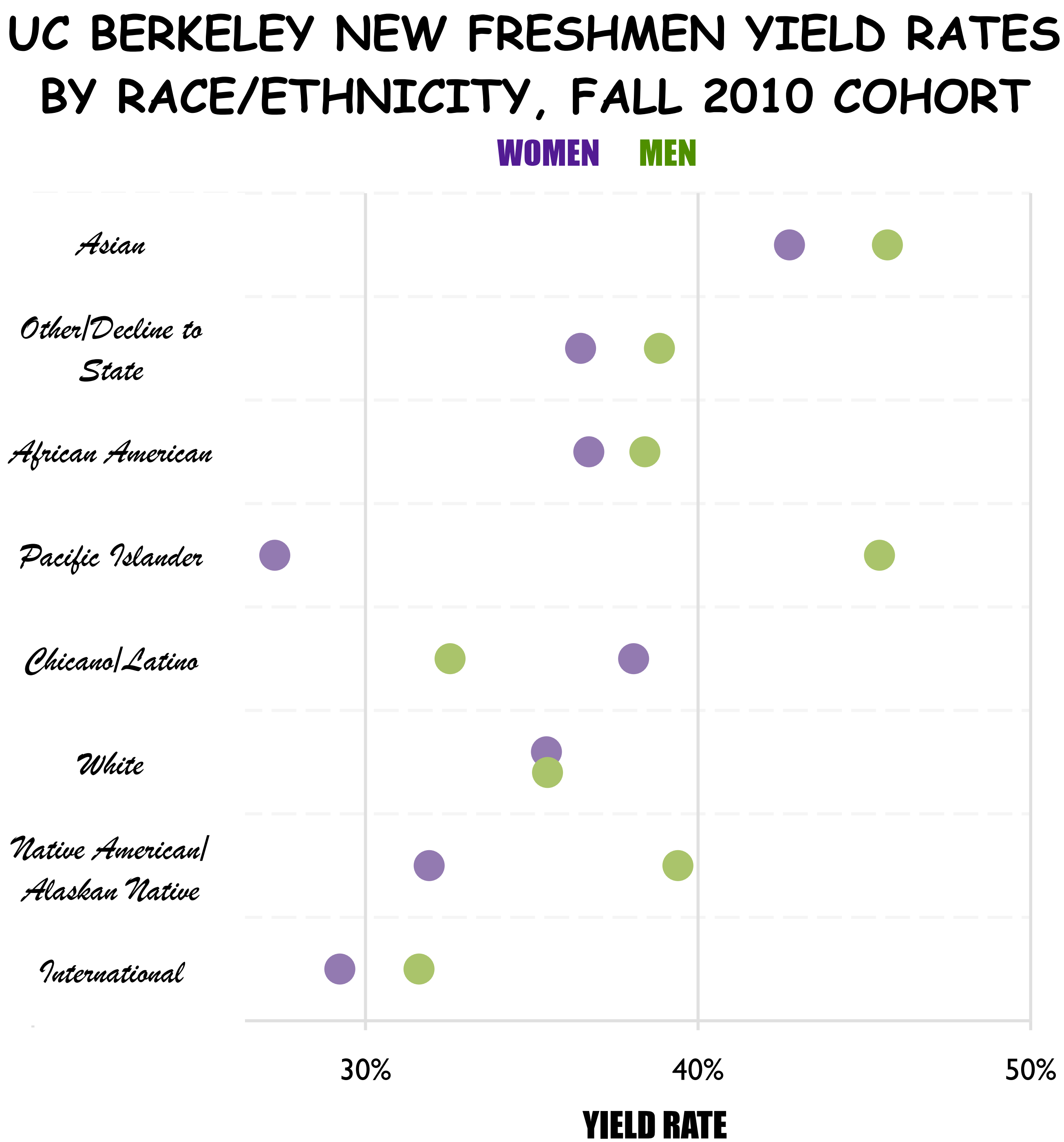
Source: Dona Wong, *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*

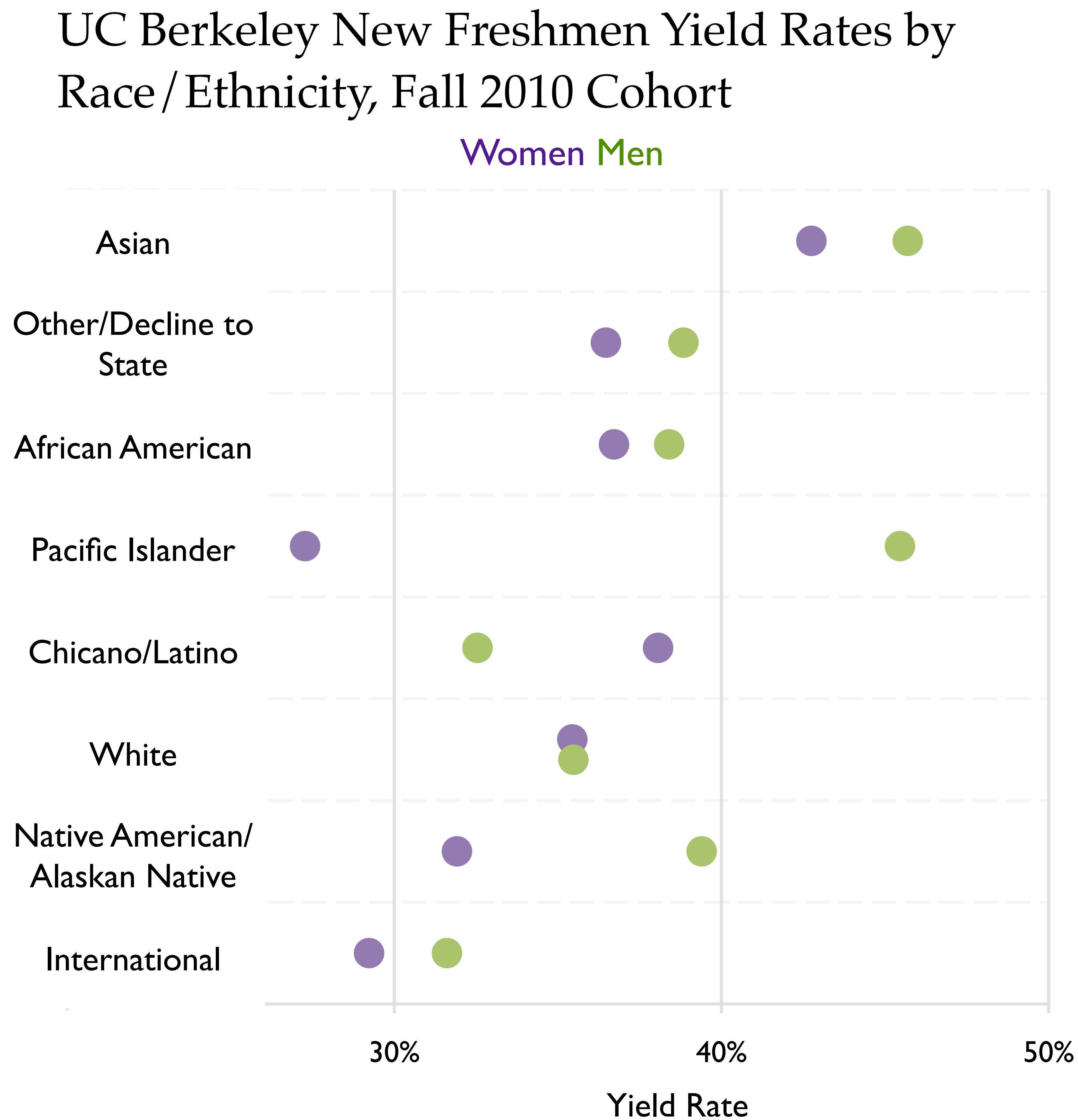
✓ use a palette mostly of grays and muted hues

✓ choose a few high chroma colors for contrast

✓ use shades and tints to ensure that a black-and-white copy will still be coherent

Visualization Aesthetics: Font



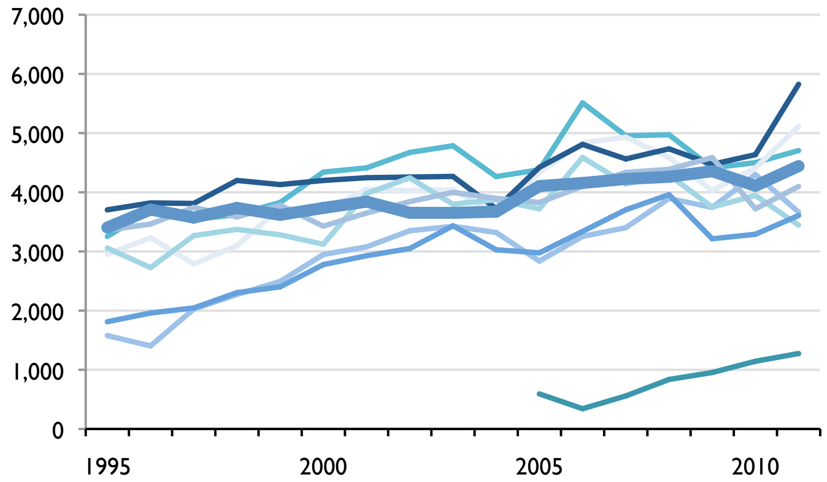


- ✗ bold and condensed fonts confuse the viewer
- ✗ multiplicity of fonts deters legibility

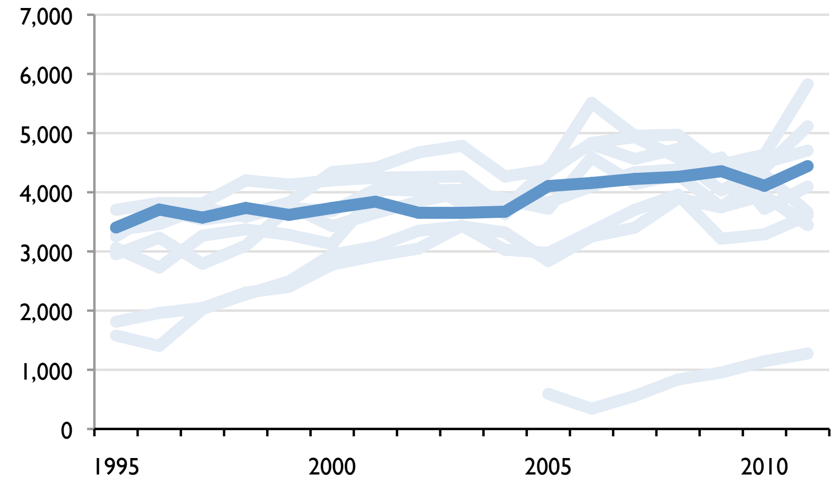
- ✓ font choice, weight, and spacing aid clarity
- ✓ single font used for labels -- second font only used for the title

Source: UC Berkeley, Cal Answers

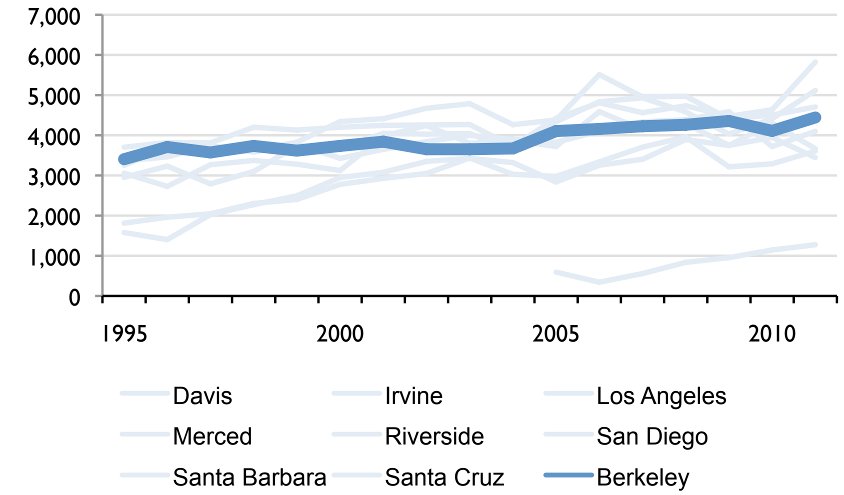
Visualization Aesthetics: Lines/Shading



✘ more than four groups are identified in one chart

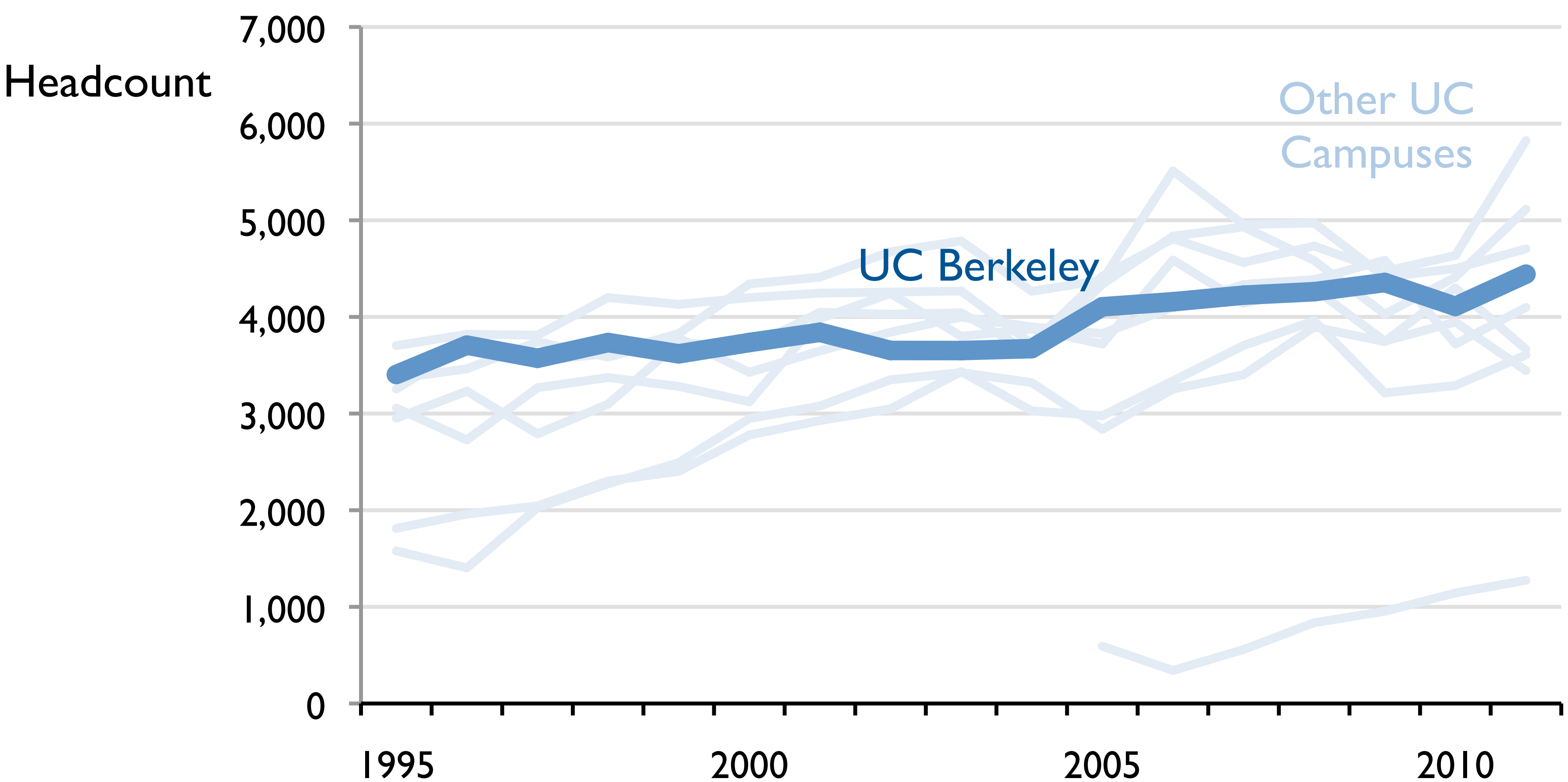


✘ weight of lines blurs trend details



✘ label position makes identification hard

UC New Fall Undergraduate Enrollment by Campus, 1995-2011



✓ only two groups are identified

✓ line weights are used for emphasis

✓ lines are directly labeled

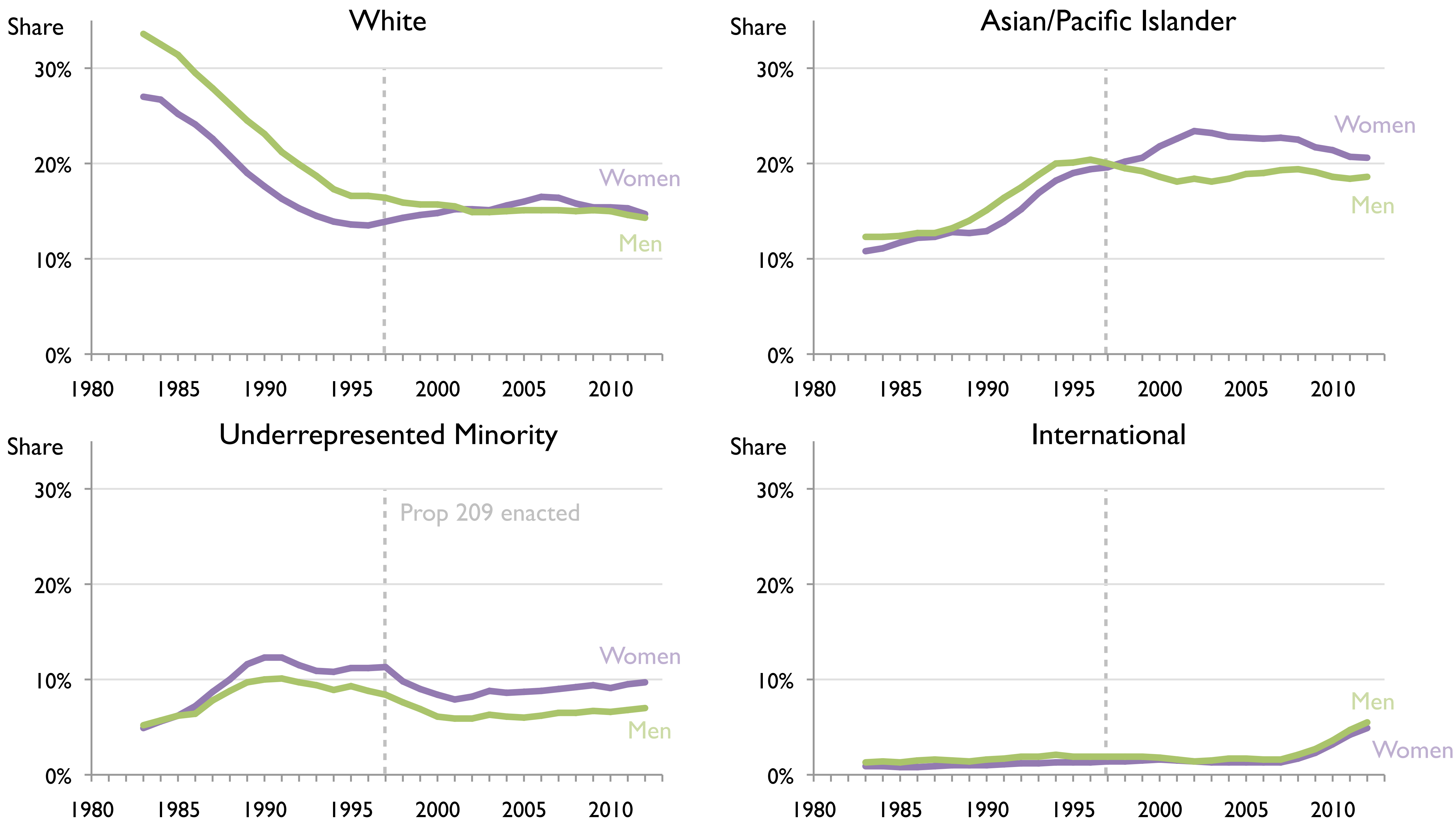
Source: UC Accountability Report, 2011

Visualization Aesthetics: Labels/Text

UC Berkeley Undergraduate New Enrollment Shares by Gender and Race/ Ethnicity, 1983-2012

Prop 209 banned affirmative action in 1997, precipitating a sharp decline in underrepresented minority (URM) students shares, which have yet to recover.

The overall gender gap with women outnumbering men is driven by Asian and URM students where the gender gaps are largest.



Source: UC Berkeley, Cal Answers

Summary

- Know what question you are asking a visualization to answer
  - Choose the best metric for your analysis and your audience
  - Choose your chart to fit your question rather than your question to fit your chart
- Let the data tell its story without excess clutter or distraction
  - Keep the focus of the visualization on the data
  - Make sure all use of font, color, shading, and text enhance rather than distract
  - Provide narrative to contextualize the highlights of the data

Contact Information

Please feel free to contact me with questions or comments

Andrew Eppig

Research Analyst

Equity & Inclusion

UC Berkeley

104 California Hall #1500

Berkeley, CA 94720-1500

aeppig@berkeley.edu

Web Resources

- *Junk Charts* -- Kaiser Fung
  - http://junkcharts.typepad.com
- *Flowing Data* -- Nathan Yau
  - http://flowingdata.com/
- *Charts 'n' Things* -- NY Times Graphics Department
  - http://chartsnthings.tumblr.com/
- *Perceptual Edge* -- Stephen Few
  - http://www.perceptualedge.com

Print Resources

Edward Tufte

- *The Visual Display of Quantitative Information*, 1983, Cheshire, CT: Graphics Press
- *Visual Explanations: Images and Quantities, Evidence and Narrative*, 1997, Cheshire, CT: Graphics Press

William Cleveland

- *The Elements of Graphing Data*, 1994, revised ed., Murray Hill, NJ: AT&T Bell Laboratories

Dona Wong

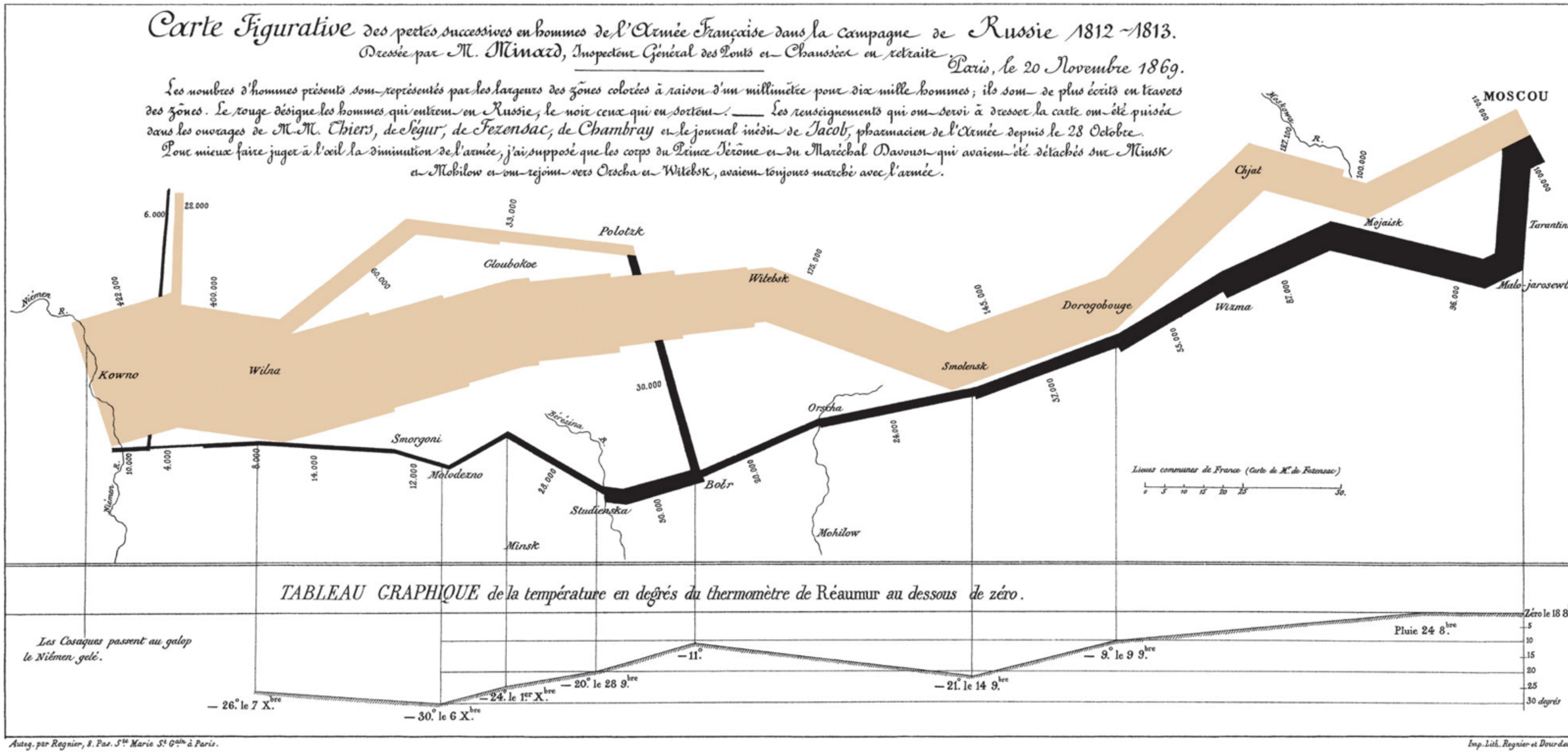
- *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*, 2010, New York: W.W. Norton and Co.

Stephen Few

- *Information Dashboard Design: The Effective Visual Communication of Data*, 2006, Oakland, CA: Analytics Press
- *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, 2009, Oakland, CA: Analytics Press
- *Show Me the Numbers: Designing Tables and Graphs to Enlighten,* 2012, second ed., Oakland, CA: Analytics Press

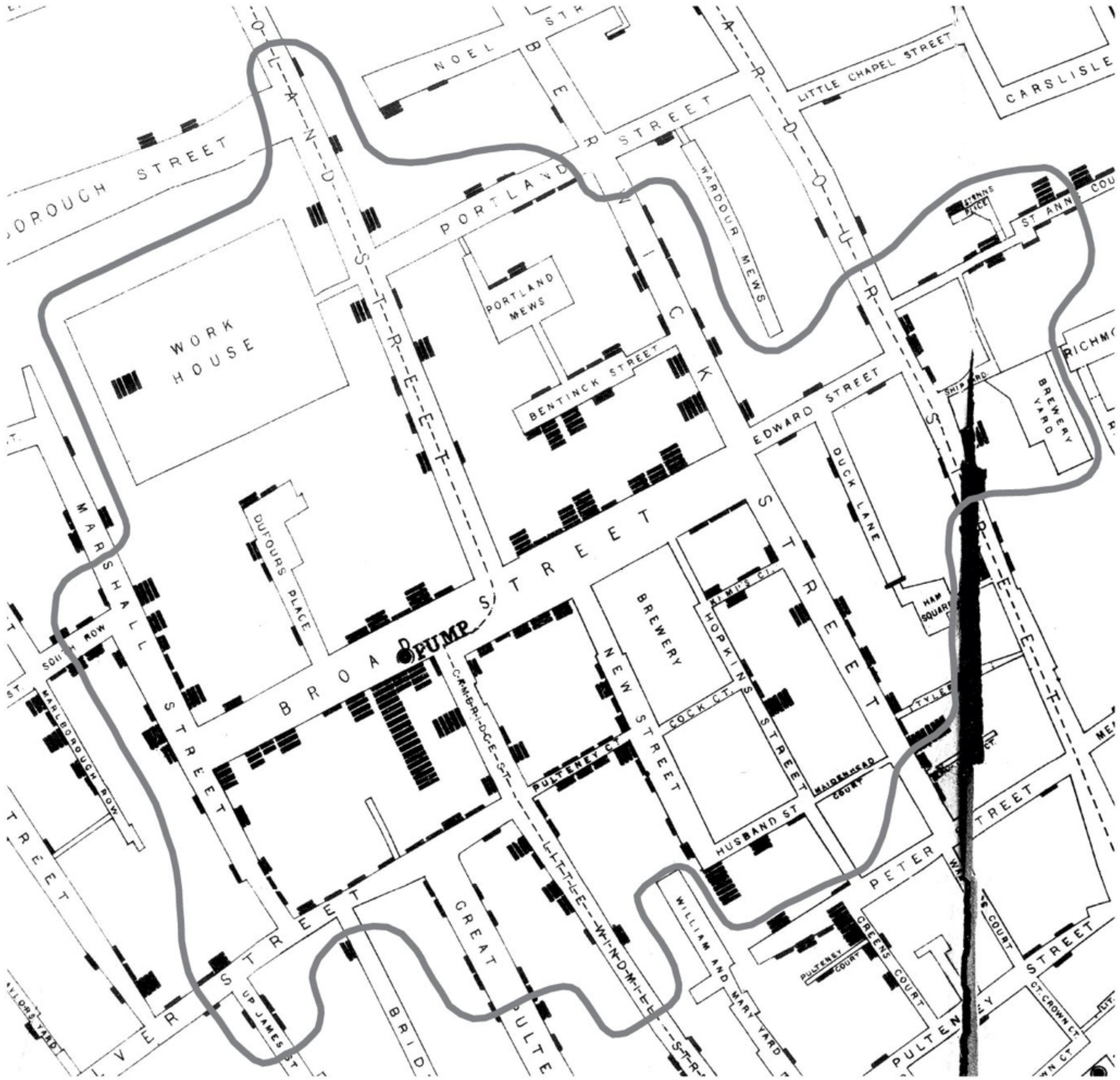
Appendices

Classic Charts



Charles Minard's 1869 chart showing the number of men in Napoleon's 1812 Russian campaign army, their movements, as well as the temperature they encountered on the return path. Lithograph, 62 x 30 cm

Classic Charts



Detail from John Snow's spot map of the Golden Square outbreak [1854 London cholera outbreak] showing area enclosed within the Voronoi network diagram. Snow's original dotted line to denote equidistance between the Broad Street pump and the nearest alternative pump for procuring water has been replaced by a solid line for legibility. Fold lines and tear in original (adapted from CIC, between 106 and 07).

Bad Chart Examples

The problem:

- The 1978 dollar should be roughly half as big as the 1958 dollar ($0.44 vs $1.00) instead of the roughly one quarter as big

How the problem occurred:

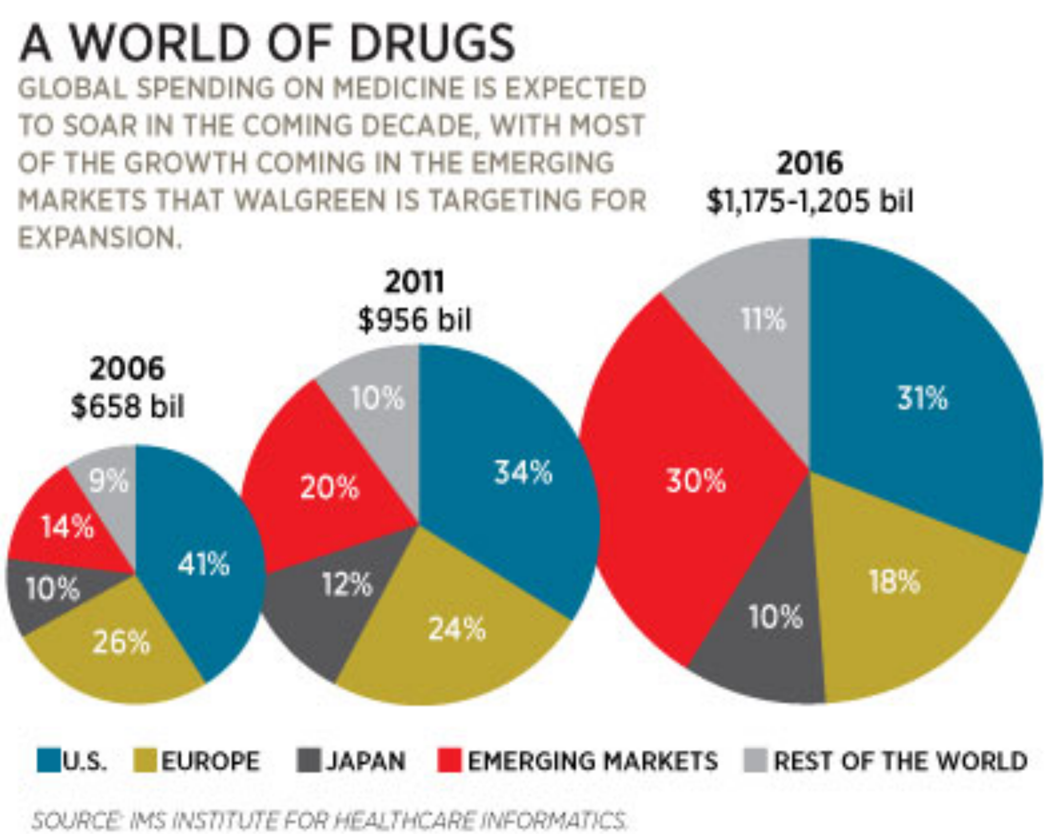
- The chart uses 2-D graphics (i.e., representations of dollar bills with length and width), and both the length and the height were scaled by 1/2 -- resulting in the area being scaled by 1/4 (1/2 x 1/2)

The fix:

- When dealing with 2-D area representations (never use 3-D), remember to scale the area rather than scaling each dimension separately

Source: Tufte, 1983

Bad Chart Examples



The problem:

- The message (growth of medical spending in emerging markets) is obfuscated and exaggerated

How the problem occurred:

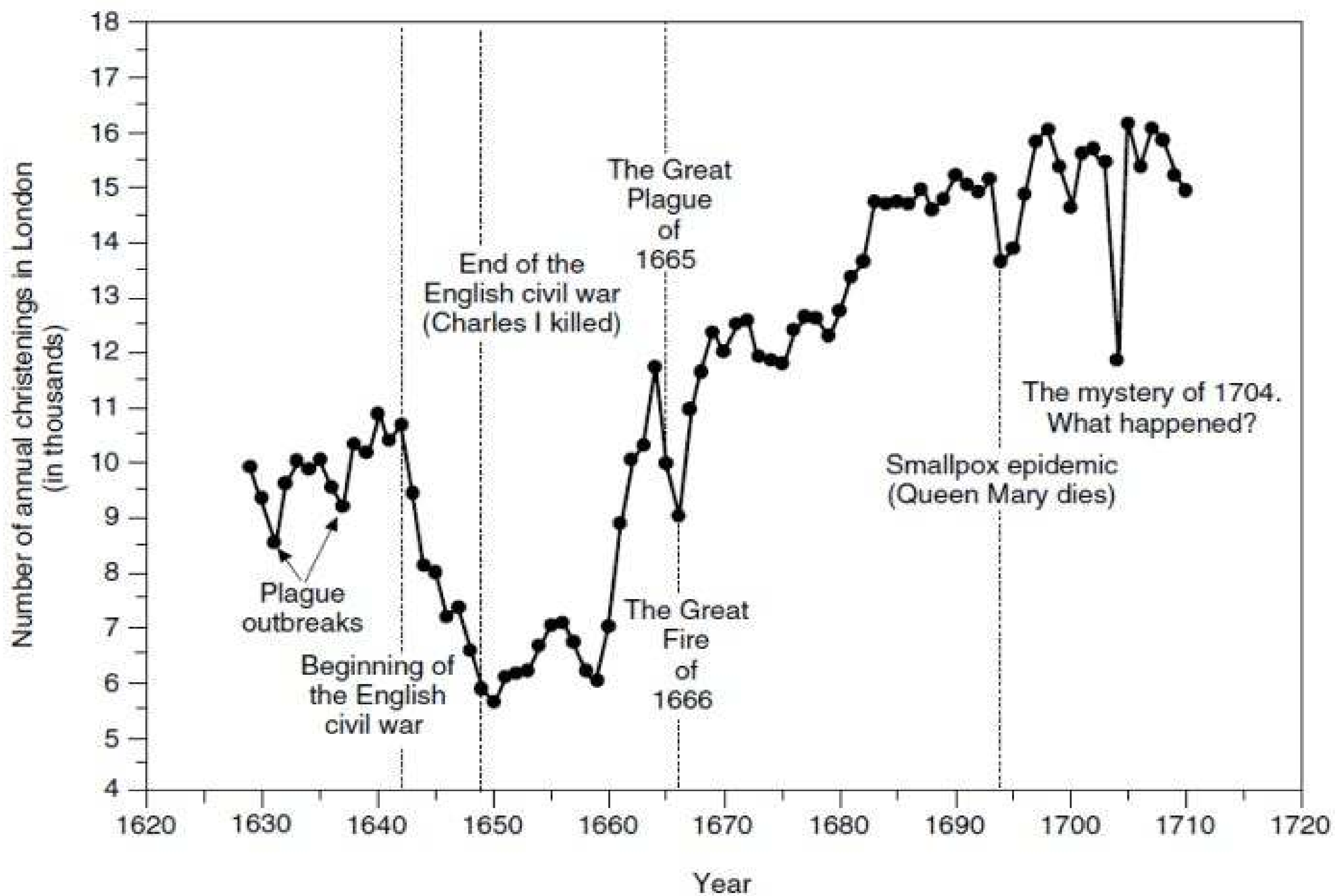
- The chart uses too many bold colors, which creates visual confusion
- The chart uses pie charts for each year, which makes it hard to see trends
- The chart scales the pie charts incorrectly by scaling only the radius opposed to the area which distorts the changes

The fix:

- When dealing with trend data, time series using line charts are the best choice

Source: “Expanding Circles of Error”, *Junk Charts*

Data Exploration via Visualization



Howard Wainer's visualization of John Arbuthnot's 1710 analysis of London *Bills of Mortality* not only depicts historical incidents, it also provides a check for data quality. The 1704 spike is not associated with any historical incident. A check of the data reveals a transcription error by Arbuthnot where the 1674 data point was mistakenly labeled as 1704.

Source: Wainer, 2009

Infographic Creation Details

Data Preparation Steps

- Source identification
- Data collection
- Data scrubbing
- Data analysis

Infographic Source Identification

- Super heroes and villains: DC and Marvel
  - http://dc.wikia.com/
  - http://marvel.com/universe/Main_Page
- Top athletes: 2008 US Olympic Team
  - http://www.2008.nbcolympics.com/athletes/index.html
- Top models: models.com listings
  - http://models.com/

Infographic Data Collection

- Create Python web scraper
  - Crawl web sites
  - Download web pages
  - Extract height, weight, and gender data
  - Save data to file

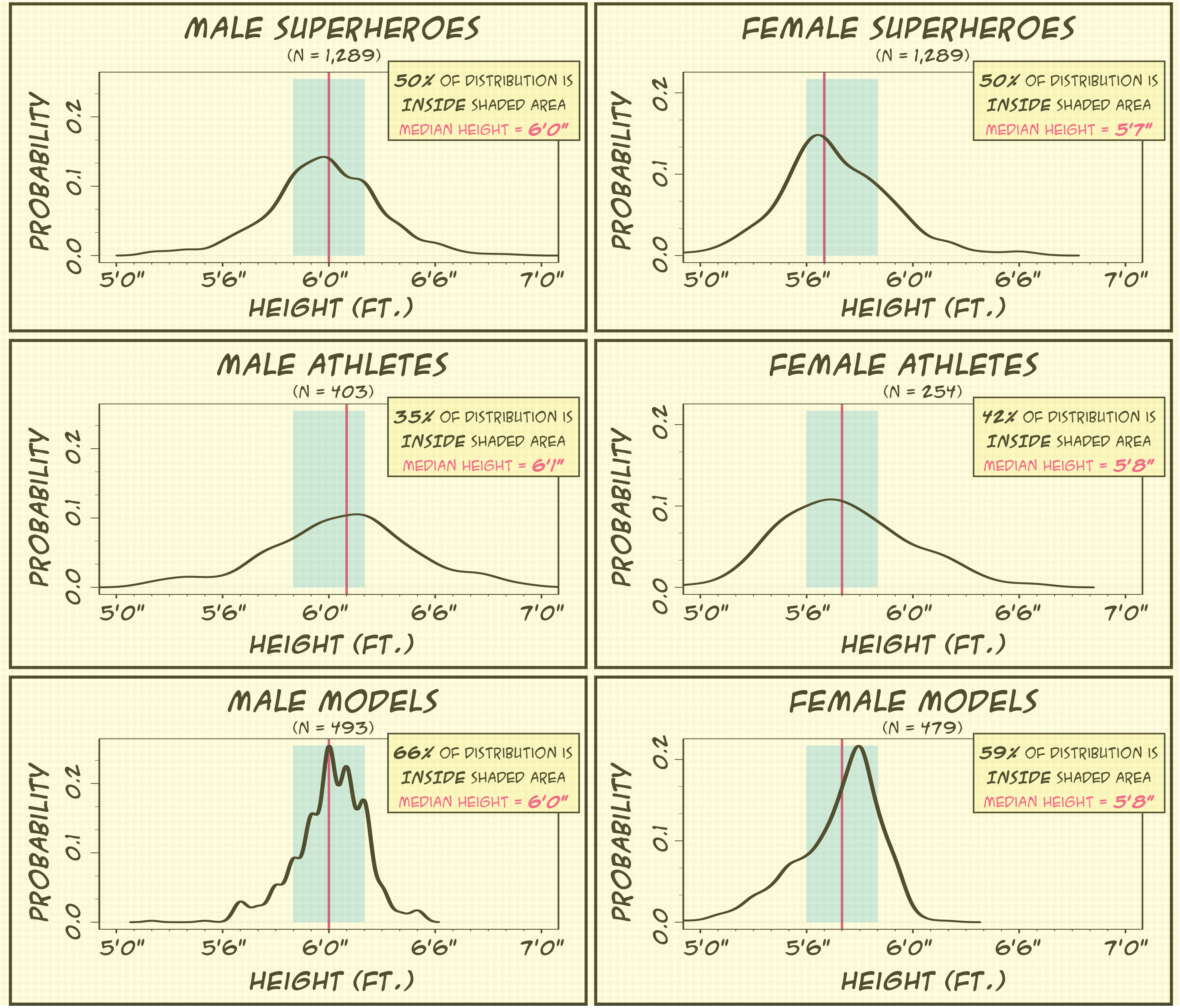
Infographic Data Scrubbing

- Check data quality
  - Did extraction get correct height and weight?
  - Are there duplicate entries?
- Remove super hero and super villain outliers
  - Define height window based on athlete and model data
  - Define weight window based on athlete and model data

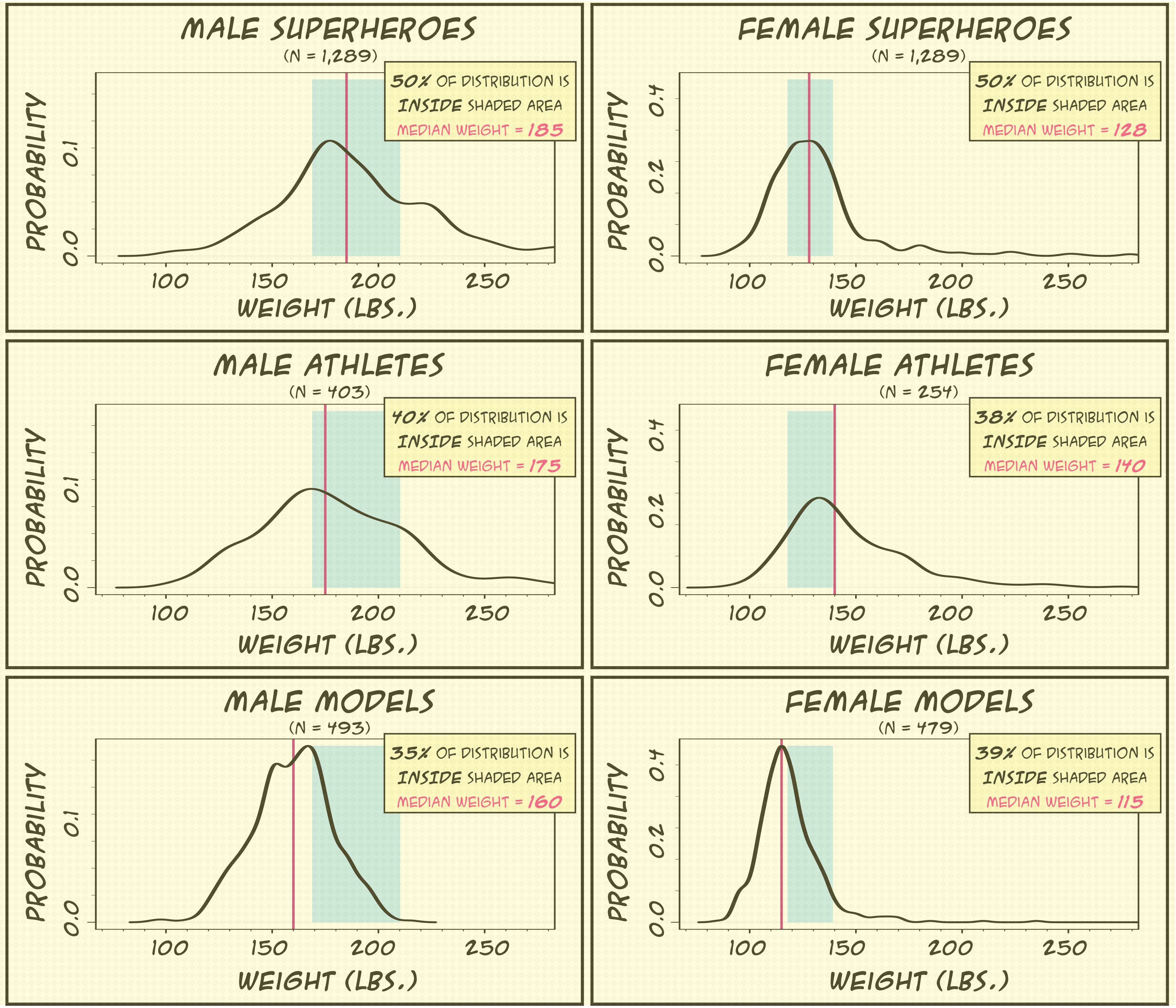
Infographic Data Analysis

- Combine all data in R
  - Super heroes and villains, athletes, and models
  - Create dummy variables
    - Gender: male, female
    - Source: super hero/villain, athlete, model
- Calculate BMI for each record
- Check summary statistics
  - Data ranges, mean, standard deviation
  - Run t-tests between groups

Height Distributions



Weight Distributions



With Revised Data