Statistical Graphics Considerations

Bootcamp 2018

Instructor:
Dawn Koffman, Statistical Programmer
Office of Population Research (OPR)
Princeton University

Statistical Graphics Considerations

Why this topic?

“Most of us use a computer to write,
 but we would never characterize a Nobel prize winning writer
 as being highly skilled at using a word processing tool.

Similarly, advanced skills with graphing languages/packages/tools
won’t necessarily lead to effective communication of numerical data.

You must understand the principles of effective graphs in addition to the mechanics.”

Jennifer Bryan, Associate Professor Statistics & Michael Smith Labs, Univ. of British Columbia.
http://stat545-ubc.github.io/block015_graph-dos-donts.html

"... quantitative visualization is

a core feature of scientific practice from start to finish.
All aspects of the research process from the initial exploration of data to the
effective presentation of a polished argument can benefit from good graphical habits.

... the dominant trend is toward a world where the visualization of data and results
is a routine part of what it means to do science."

But ... for some odd reason

" ... the standards for publishable graphical material vary wildy between and even within articles
– far more than the standards for data analysis, prose and argument.

Variation is to be expected, but the absence of consistency in elements as simple as
axis labeling, gridlines or legends is striking."

Kieran Healy and James Moody, *Data Visualization in Sociology*, Annu. Rev. Sociol. 2014 40:5.1-5.5.

What is a statistical graph?

“A statistical graph is a visual representation of statistical data.

The data are observations and/or functions of one or more variables.

The visual representation is a picture on a two-dimensional surface
using symbols, lines, areas and text to display possible relations between variables.”

David A Burn. *Designing Effective Statistical Graphs, Handbook of Statistics*, Vol 9. CR Rao, ed. 1993.

A statistical graph allows us to…

- see the big picture

Graphs reveal the big picture: an overview of a data set.
An overview summarizes the data’s essential characteristics, from which we can discern what’s routine vs. exceptional.

- easily and rapidly compare values

Graphs make it possible to see many values at once and easily and rapidly compare them.

- see patterns among values

Graphs make it easy to patterns formed by sets of values.
For example, patterns may describe correlations among values, how values are distributed, or how values change over time.

- compare patterns among sets of values

Graphs let us compare patterns found among different sets of values.

From Steven Few, Perceptual Edge: http://www.perceptualedge.com/blog/?p=1897 primary goals of a statistical graph

- explore and understand data by **accurately** representing it
- allow viewer to easily see **comparisons** of interest (including trends)
- communicate results in a **clear** and memorable way

how to do this is somewhat subjective ...

- few hard and fast rules
- many trade-offs
- many guidelines which some may disagree with
- **iterative process** is often helpful
  ... design and “build” multiple version of “same graph”

purpose of this session is to encourage you to **consider**

- techniques
- guidelines
- tradeoffs

and then to determine what *you* think makes the most sense for your particular case

One more thing ... a disclaimer …

Let me state clearly ... I intend no criticism of graph authors, either individually or as a group.

Shortcomings show only that we are all human, and that under the pressure of a large, intellectually demanding task like designing and building a statistical graph it is much too easy to do things imperfectly.   Additionally, many design considerations involve trade-offs, where there may be, in fact, no “best” solution.

Lastly, I have no doubt that some of the “better graphs” I show will provide “bad” examples for future viewers – I hope only that they will learn from the experience of studying them carefully.

Inspired by Brian W. Kernighan and P. J. Plauger, Preface to First Edition *of The Elements of Programming Style*, 1978.

Session Organization

Preface Why this topic?

What is a statistical graph?

I Introduction

Tables vs graphs (When)

Audience and setting (Where)

II Representing data **accurately**

III Highlighting **comparisons** of interest (How)

IV Simplicity and **clarity**

V Summary

VI Conclusions

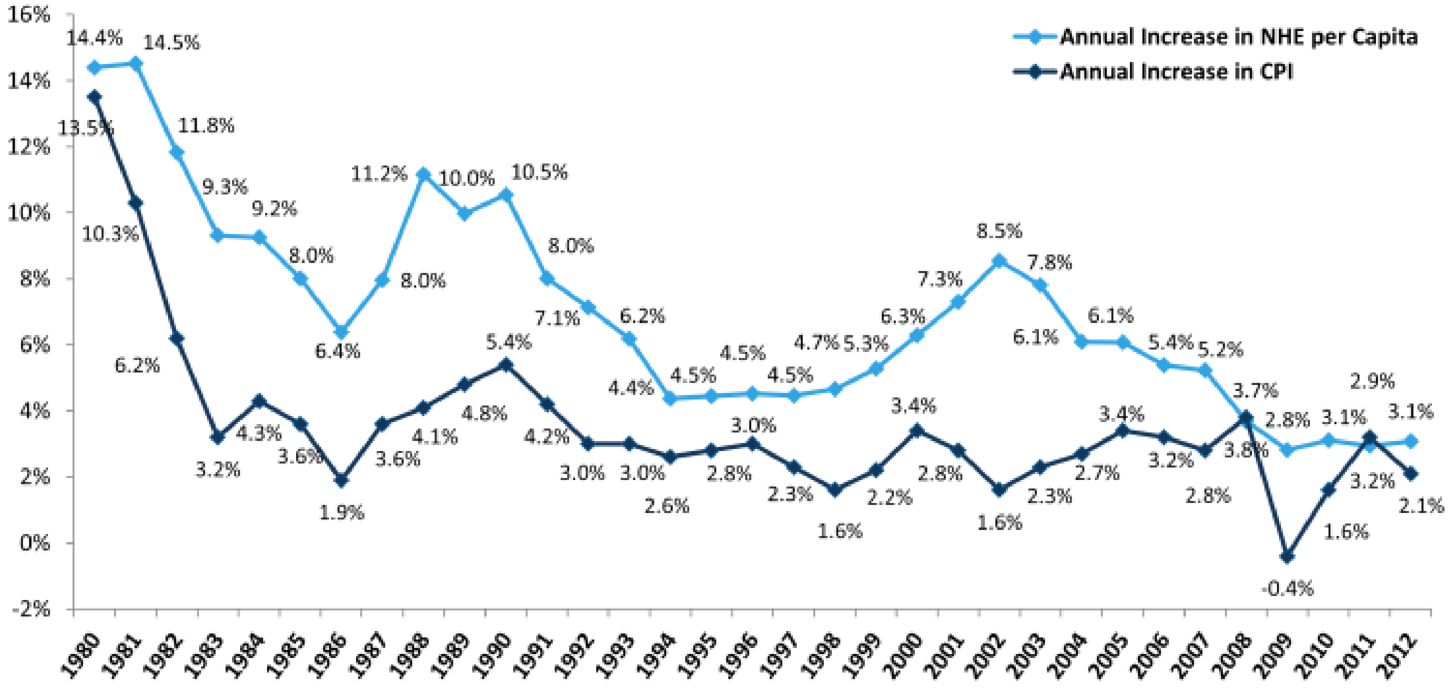
Tables vs Graphs

tables: look up individual, precise values

graphs: see overall distribution (shape, pattern) of data
make comparisons
perceive trends
often more useful when working with large sets of data

A graph trying to also serve as a look-up table ...

Percent Annual Increase in National Health Expenditures (NHE) per Capita vs. Increase in Consumer Price Index (CPI), 1980-2012

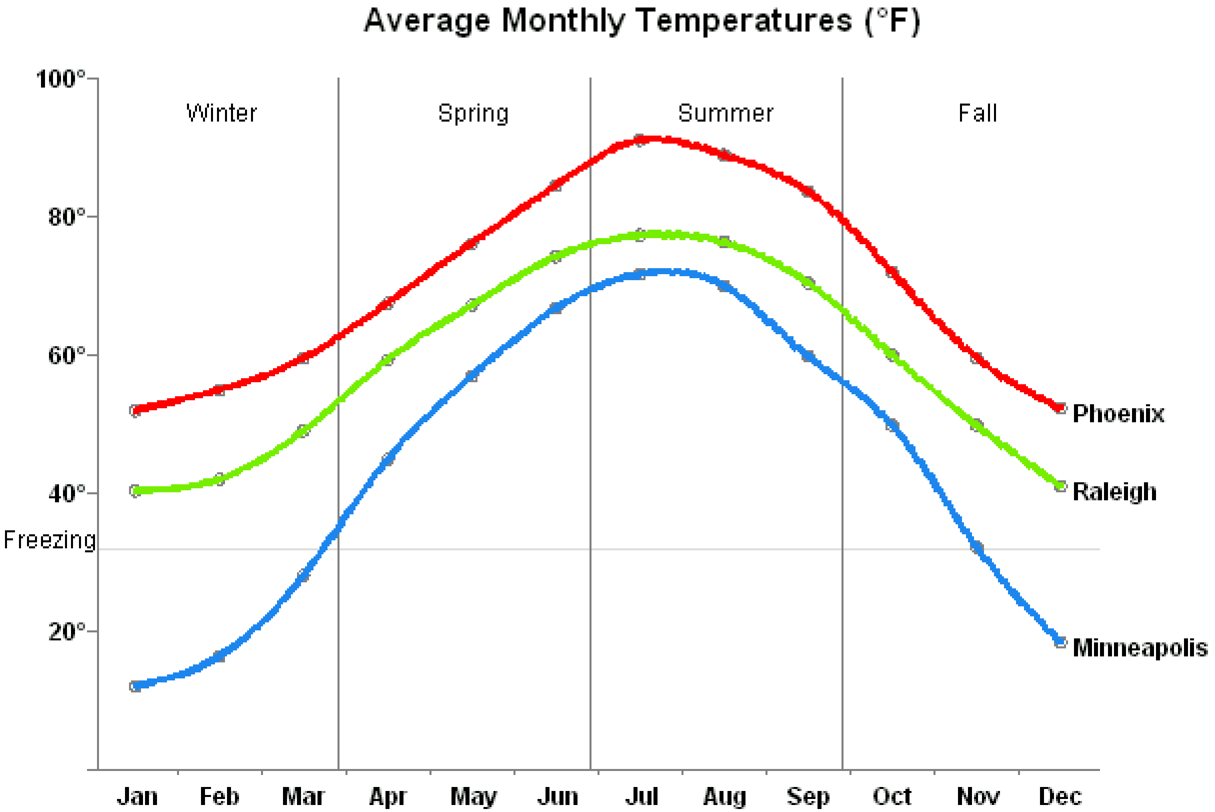


SOURCE: Kaiser Family Foundation calculations using NHE data from Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group, at http://www.cms.hhs.gov/NationalHealthExpendData/ (see Historical; National Health Expenditures by type of service and source of funds; file nhe12.zip), and CPI data from Bureau of Labor Statistics at ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt (All Urban Consumers, All Items, 1982-1984=100, Not Seasonally Adjusted, U.S. city average).

THE HENRY J. KAISER FAMILY FOUNDATION

A much nicer way to show a graph and table …

Average Monthly Temperatures (°F)



Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	
52.1	55.1	59.7	67.7	76.3	84.6	91.2	89.1	83.8	72.2	59.8	52.5	🟥 Phoenix
40.5	42.2	49.2	59.5	67.4	74.4	77.5	76.5	70.6	60.2	50.0	41.2	🟩 Raleigh
12.2	16.5	28.3	45.1	57.1	66.9	71.9	70.2	60.0	50.0	32.4	18.6	🟦 Minneapolis

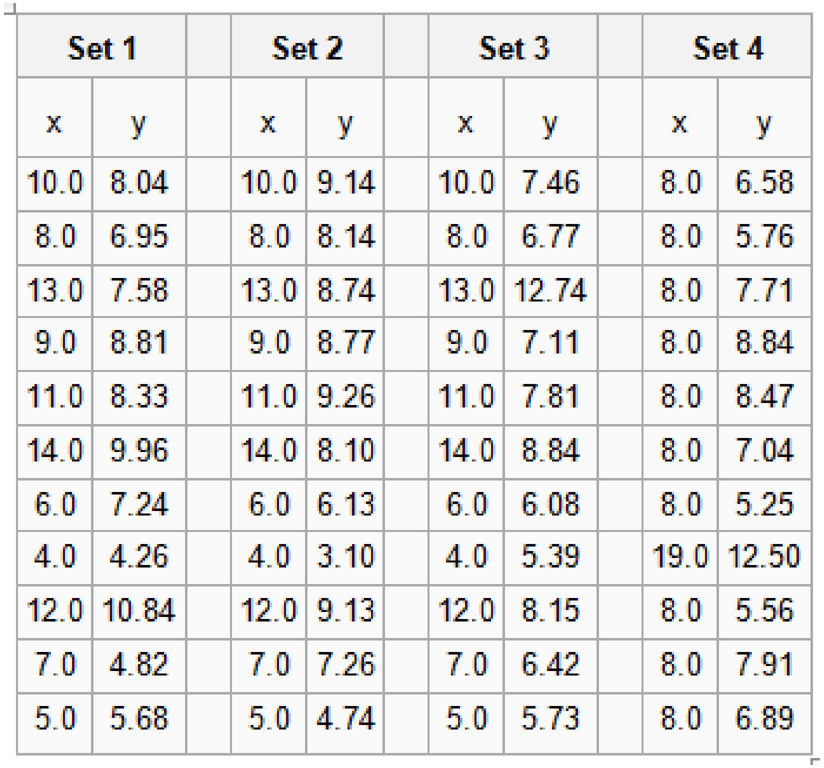
From Stephen Few, Perceptual Edge: http://www.perceptualedge.com/example2.php

Anscombe’s Quartet

4 data sets that have
nearly identical summary statistics

each has 11 non-missing pairs of values

constructed in 1973 by statistician Francis Anscombe to demonstrate importance of graphing data and effect of outliers

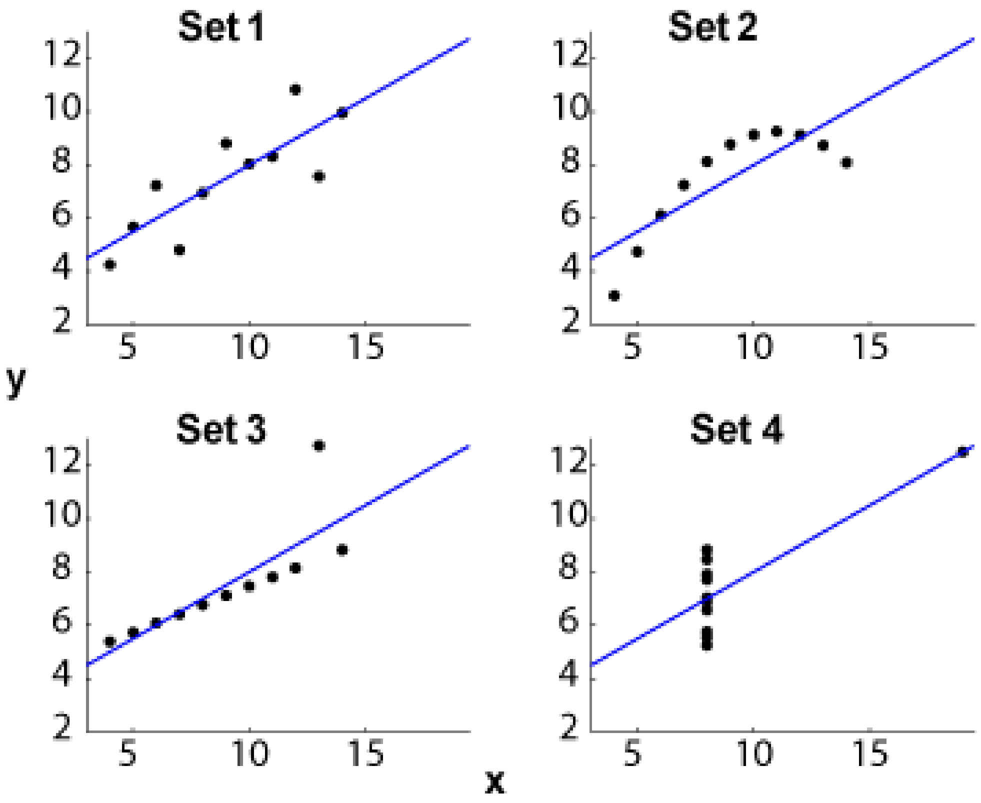


Set 1		Set 2		Set 3		Set 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

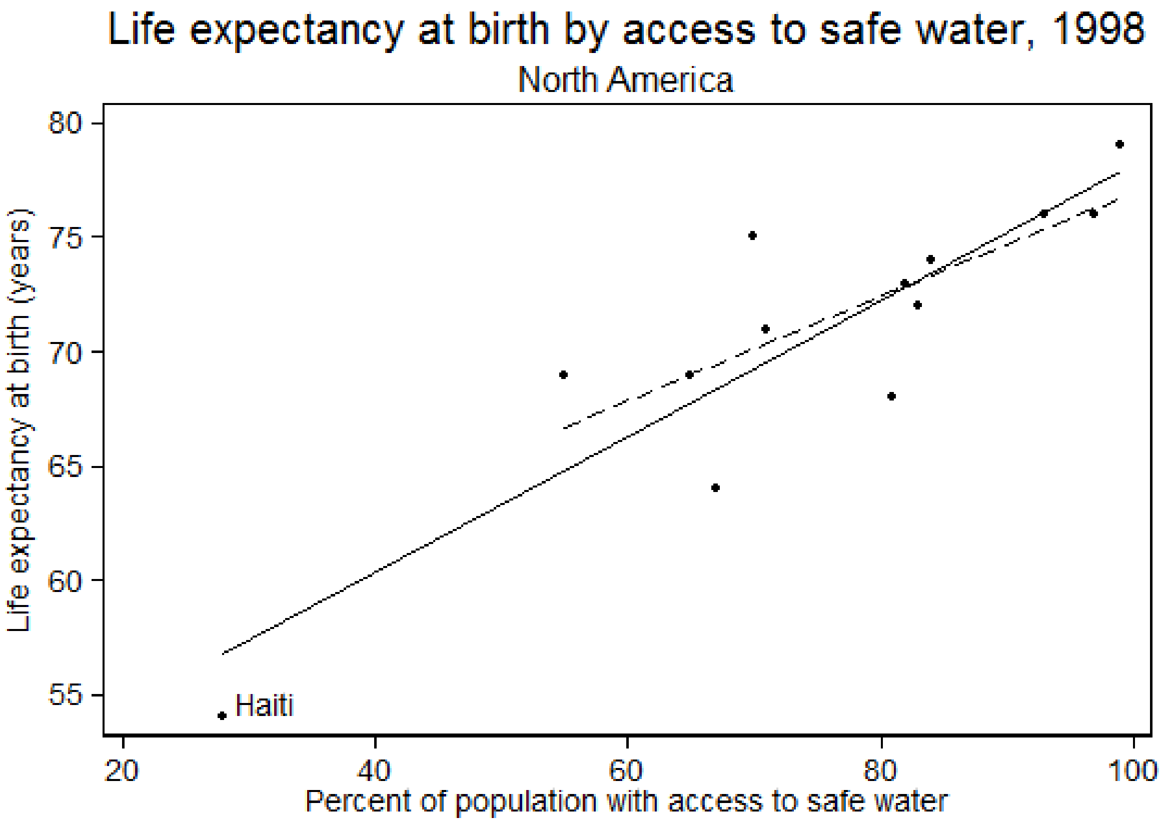
SUMMARY STATISTICS				
mean value of x	9	9	9	9
mean value of y	7.5	7.5	7.5	7.5
variance of x	11	11	11	11
variance of y	4.1	4.1	4.1	4.1
correlation between x and y	0.816	0.816	0.816	0.816
linear regression (best fit) line is:	y=0.5x+3	y=0.5x+3	y=0.5x+3	y=0.5x+3

Anscombe, FJ (1973). "Graphs in Statistical Analysis". American Statistician 27 (1): 17–21.

Anscombe's Quartet



hard to see the forest when looking at the trees

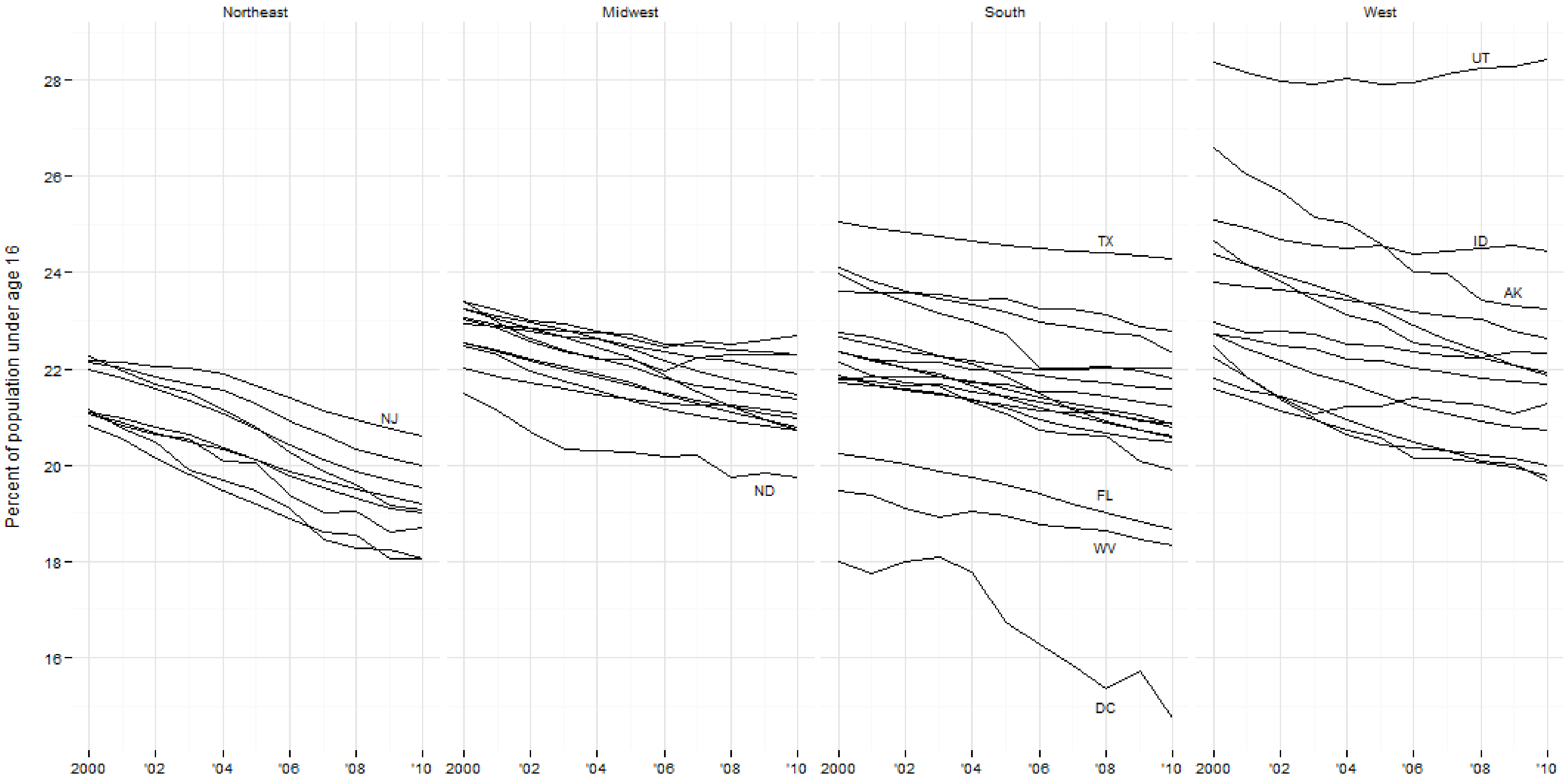


graph allows simple visual examination of effect of outlier on model summary

US States: percent of population under age 16, 2000-2010

State	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	region
Alabama	22.35	22.18	22.03	21.83	21.74	21.58	21.39	21.29	21.17	21.04	20.84	3
Alaska	26.59	26.03	25.7	25.15	25.04	24.59	24	23.97	23.44	23.32	23.25	4
Arizona	23.81	23.72	23.63	23.56	23.43	23.33	23.17	23.09	23.04	22.8	22.62	4
Arkansas	22.36	22.21	22.14	22.13	22	21.97	21.86	21.76	21.7	21.61	21.59	3
California	24.39	24.16	23.94	23.75	23.52	23.23	22.91	22.59	22.34	22.09	21.94	4
Colorado	22.72	22.64	22.49	22.43	22.21	22.17	22.03	21.92	21.8	21.74	21.67	4
Connecticut	22.13	22.02	21.83	21.67	21.57	21.27	20.9	20.64	20.33	20.13	19.99	1
Delaware	22.14	21.86	21.71	21.64	21.3	21.07	20.72	20.64	20.59	20.07	19.89	3
District of Columbia	18.01	17.74	17.98	18.1	17.78	16.75	16.29	15.85	15.34	15.71	14.74	3
Florida	20.25	20.13	20.02	19.86	19.73	19.58	19.41	19.2	19.02	18.82	18.68	3
Georgia	23.61	23.58	23.58	23.55	23.42	23.46	23.26	23.24	23.11	22.89	22.79	3
Hawaii	21.58	21.37	21.13	20.94	20.72	20.57	20.15	20.14	20.05	19.97	19.79	4
Idaho	25.1	24.92	24.7	24.58	24.5	24.58	24.37	24.45	24.51	24.58	24.44	4
Illinois	23.23	23.1	22.96	22.83	22.64	22.43	22.18	21.97	21.78	21.61	21.47	2
Indiana	22.95	22.89	22.77	22.66	22.62	22.47	22.36	22.23	22.16	22.03	21.88	2
Iowa	22.02	21.83	21.71	21.58	21.46	21.36	21.29	21.24	21.25	21.17	21.08	2
Kansas	23.39	23.21	22.99	22.95	22.79	22.62	22.44	22.56	22.51	22.59	22.7	2
Kentucky	21.71	21.66	21.59	21.5	21.35	21.25	21.12	21.09	21.1	20.92	20.89	3
Louisiana	23.97	23.65	23.39	23.16	22.96	22.72	22.01	22.01	22.05	21.95	21.79	3
Maine	20.83	20.53	20.14	19.82	19.48	19.18	18.88	18.61	18.56	18.05	18.07	1
Maryland	22.76	22.66	22.48	22.27	22.1	21.83	21.5	21.17	20.91	20.72	20.58	3
Massachusetts	21.11	20.98	20.79	20.63	20.37	20.1	19.77	19.53	19.32	19.1	19	1
Michigan	23.24	23.04	22.86	22.65	22.46	22.22	21.86	21.54	21.21	20.95	20.72	2
Minnesota	23.04	22.84	22.57	22.36	22.23	22.05	21.79	21.64	21.56	21.45	21.37	2
Mississippi	24.09	23.83	23.61	23.47	23.33	23.19	22.96	22.88	22.76	22.68	22.32	3
Missouri	22.54	22.35	22.17	21.98	21.82	21.64	21.5	21.33	21.22	21.07	20.98	2
Montana	22.23	21.83	21.38	20.98	20.65	20.43	20.36	20.31	20.08	20.02	19.68	4
Nebraska	23.05	22.91	22.86	22.77	22.76	22.71	22.5	22.49	22.38	22.34	22.3	2
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.

US state trends for percent of population under age 16, 2000-2010



Source: 2000-2010 State Characteristics Intercensal Population Estimates File, US Census Bureau, Population Division.

audience and purpose ... differences may lead to different design decisions

yourself

to check data correctness
to examine a variable distribution, outlier values, relationships with other variables
to examine model fit

others

to display distributions/relationships of variables that are important to your results (“your story”)

to most accurately and most clearly present your results

- experts vs novice in subject matter?   (use acronyms or abbreviations as axis labels?)

- experts vs novice at interpreting statistical graphs?

consider graph’s setting ...

presentation ... limited time ... CLARITY, CLARITY, CLARITY
... some audience members sitting farther away:
larger font size, higher contrast, brighter colors, axis labeling at top, etc.

class ...limited time

poster session ... more time plus possible interaction with author
(but usually little text) ... graphs need to be able to “stand on their own”

print journal article, book, report ... possibly more time plus full text ,
but may be limited by publication constraints such as
graph size, number, color and resolution
… may be able to provide more details in an appendix

web (online article, book, report or blog post) ... possibly more time plus full text
usually less limited by publication constraints
… usually can provide links to further detail

II Representing data **accurately**

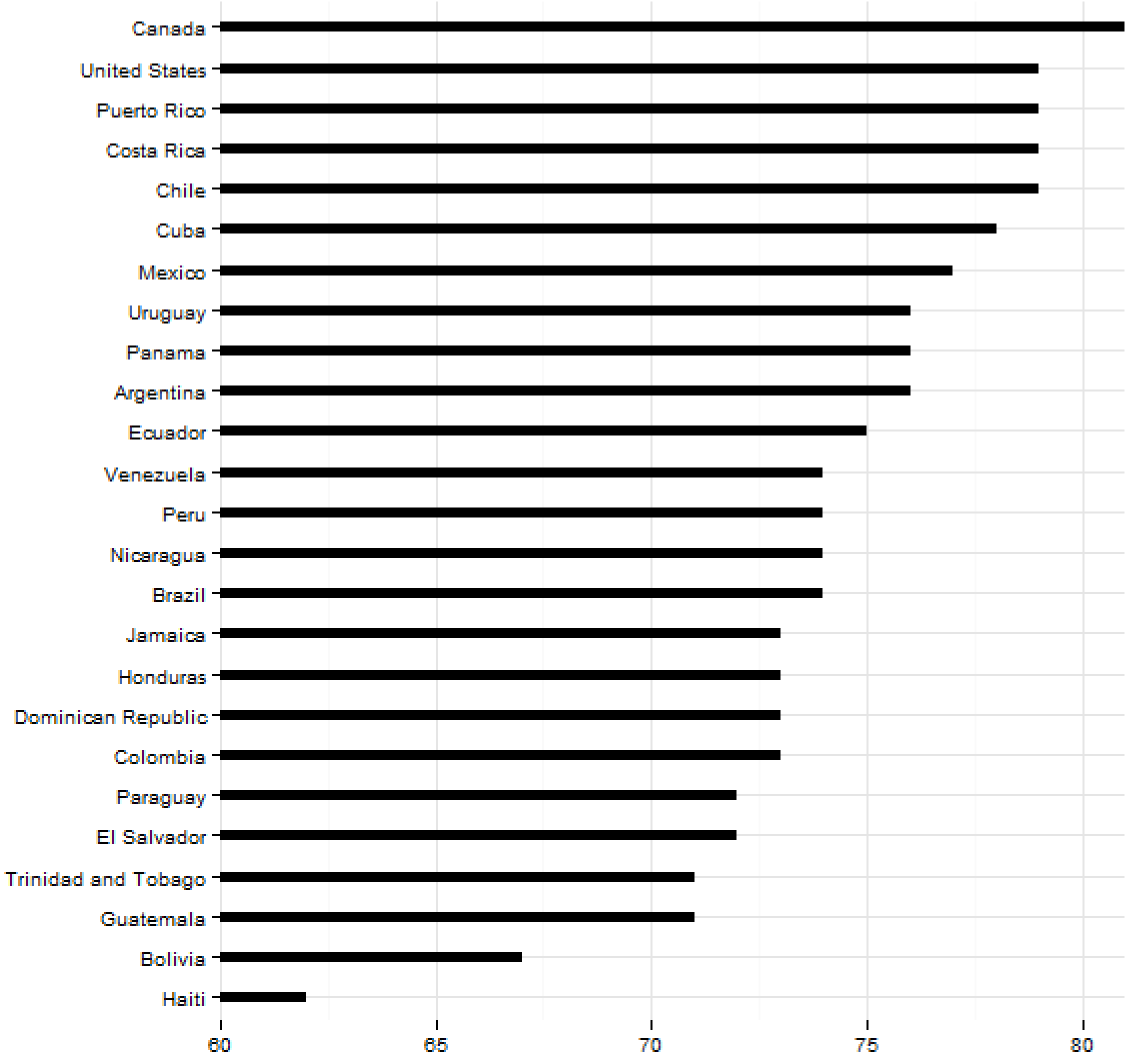
“The representation of numbers,
as physically measured on the surface of the graphic itself,
should be directly proportional to the quantities represented.”
Edward Tufte

“Visual connections should reflect real connections.”
Hadley Wickham

“Avoid distorting what the data have to say.”
Edward Tufte

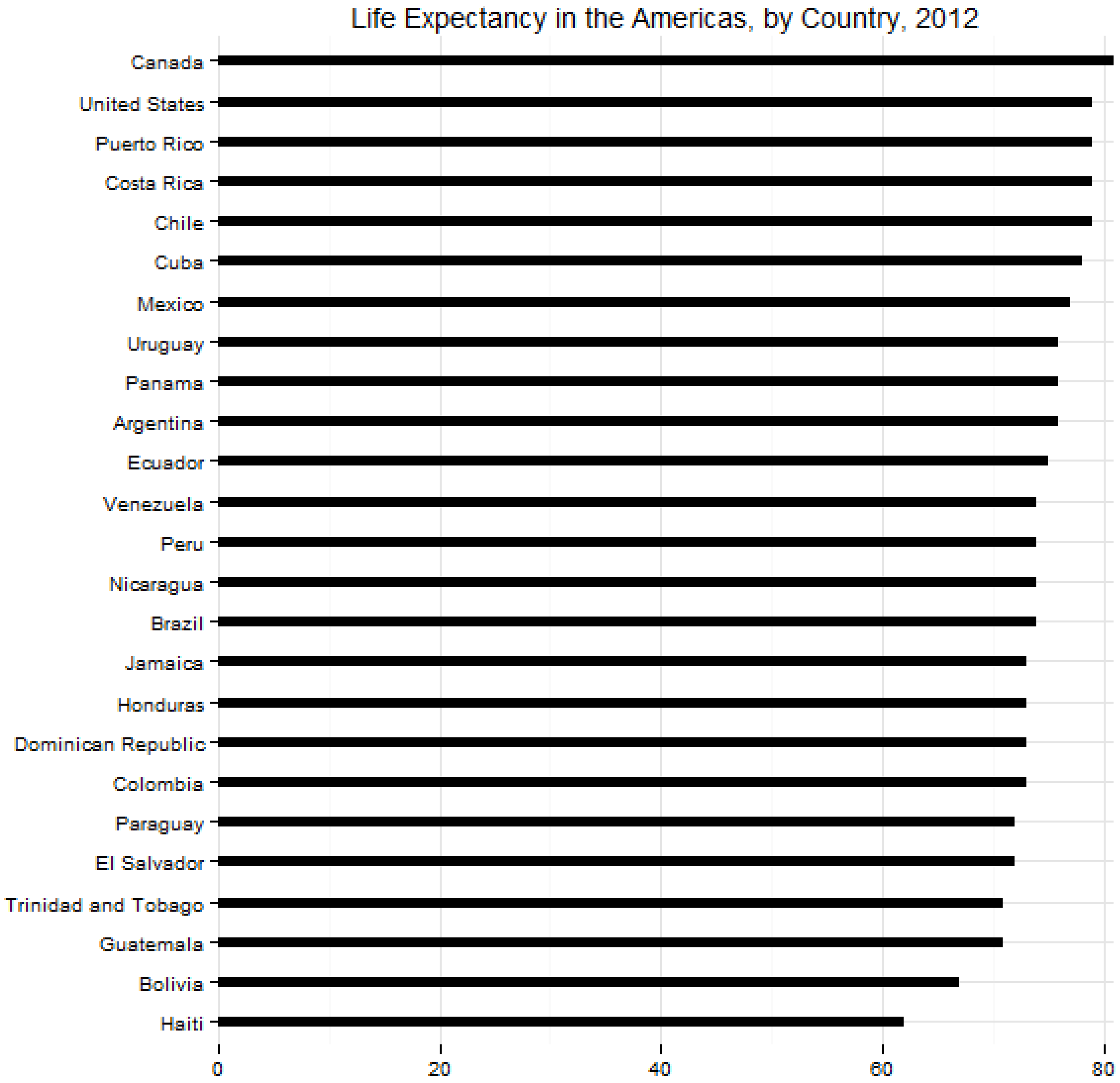
Tufte, E. The Visual Display of Quantitative Information, Second Ed. Graphics Press, Cheshire CT. 2001.

Wickham, H. Stat 405, Effective Visualisation. http://stat405.had.co.nz/lectures/20-effective-vis.pdf

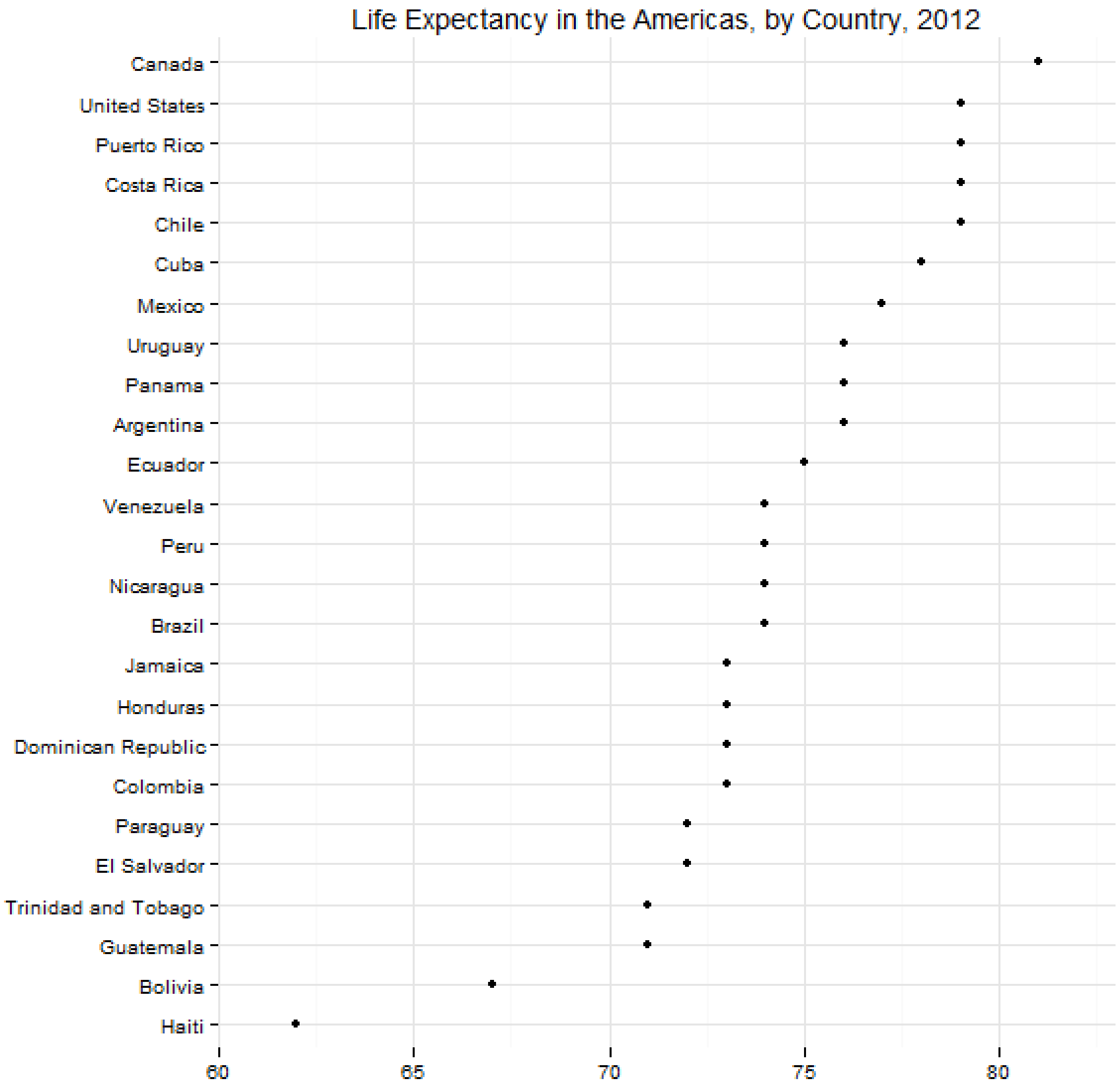


Life Expectancy in the Americas, by Country, 2012

Source: 2012 World Population Data Sheet by Pop. Ref. Bureau

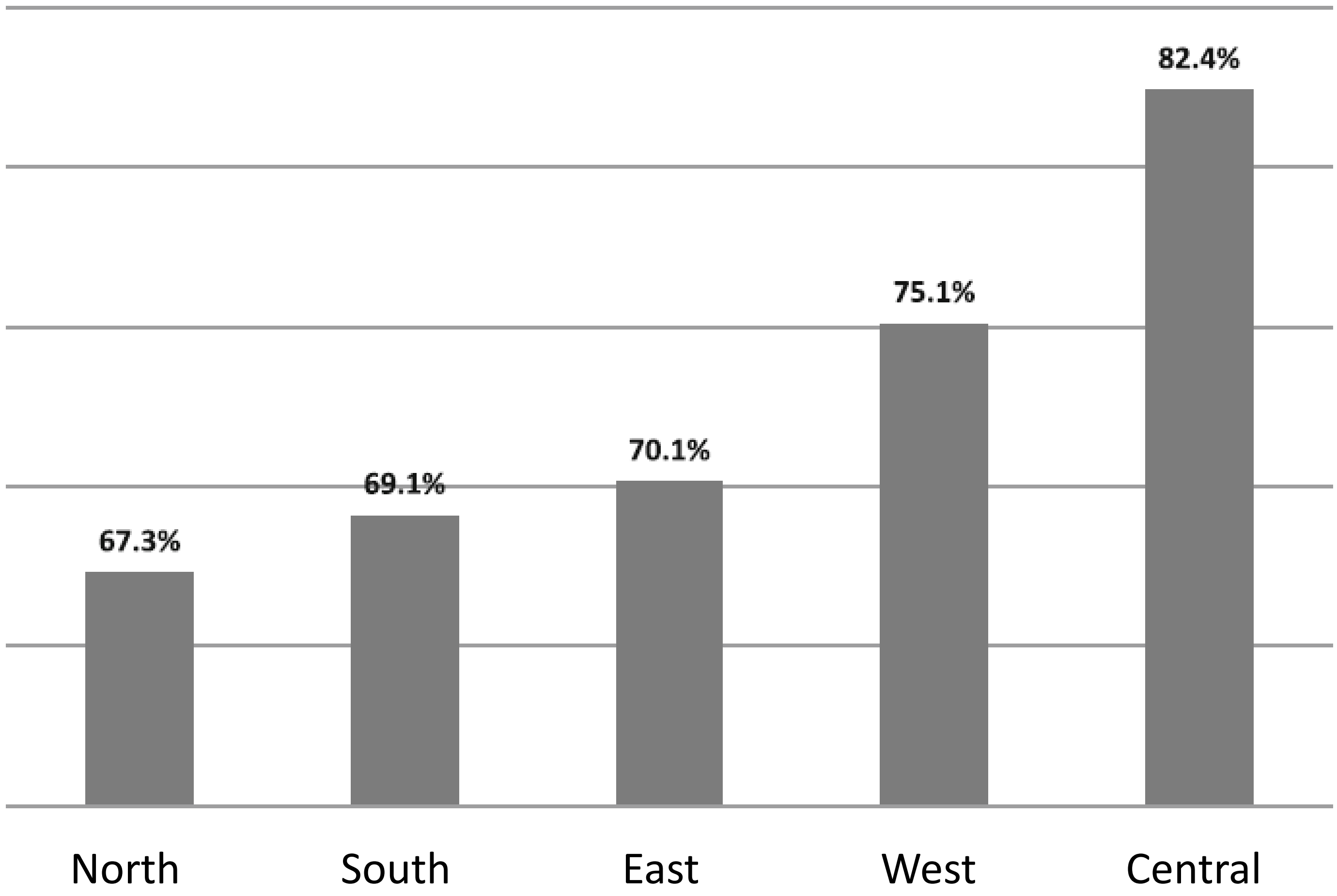


Source: 2012 World Population Data Sheet by Pop. Ref. Bureau



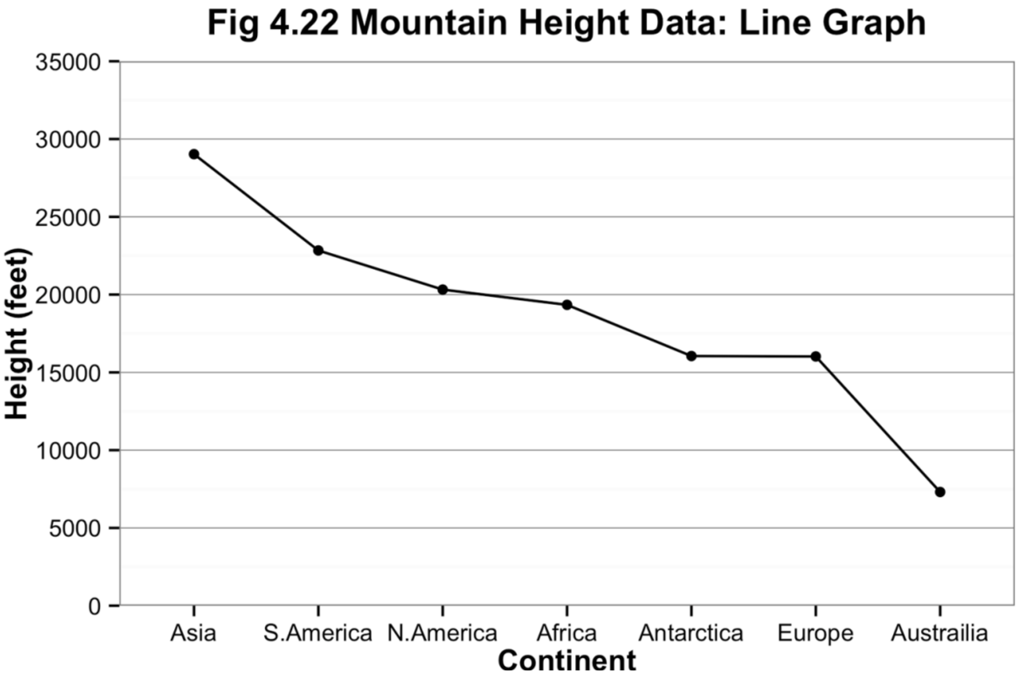
Life Expectancy in the Americas, by Country, 2012

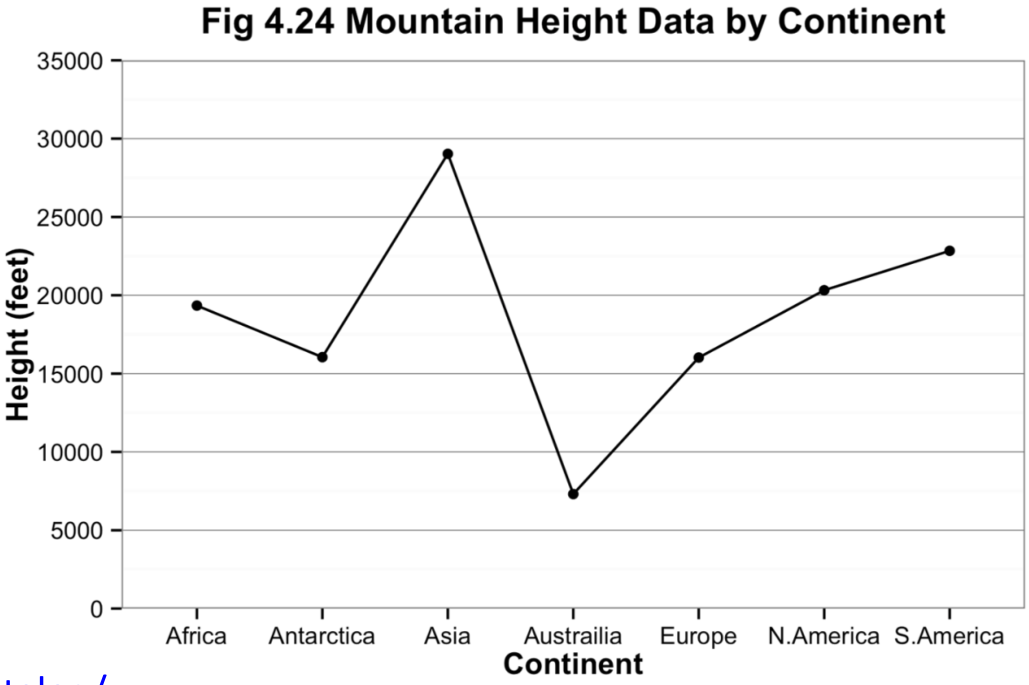
Source: 2012 World Population Data Sheet by Pop. Ref. Bureau


82.4%
75.1%
70.1%
69.1%
67.3%
North
South
East
West
Central

line graphs need interval scales for slopes to be meaningful ...

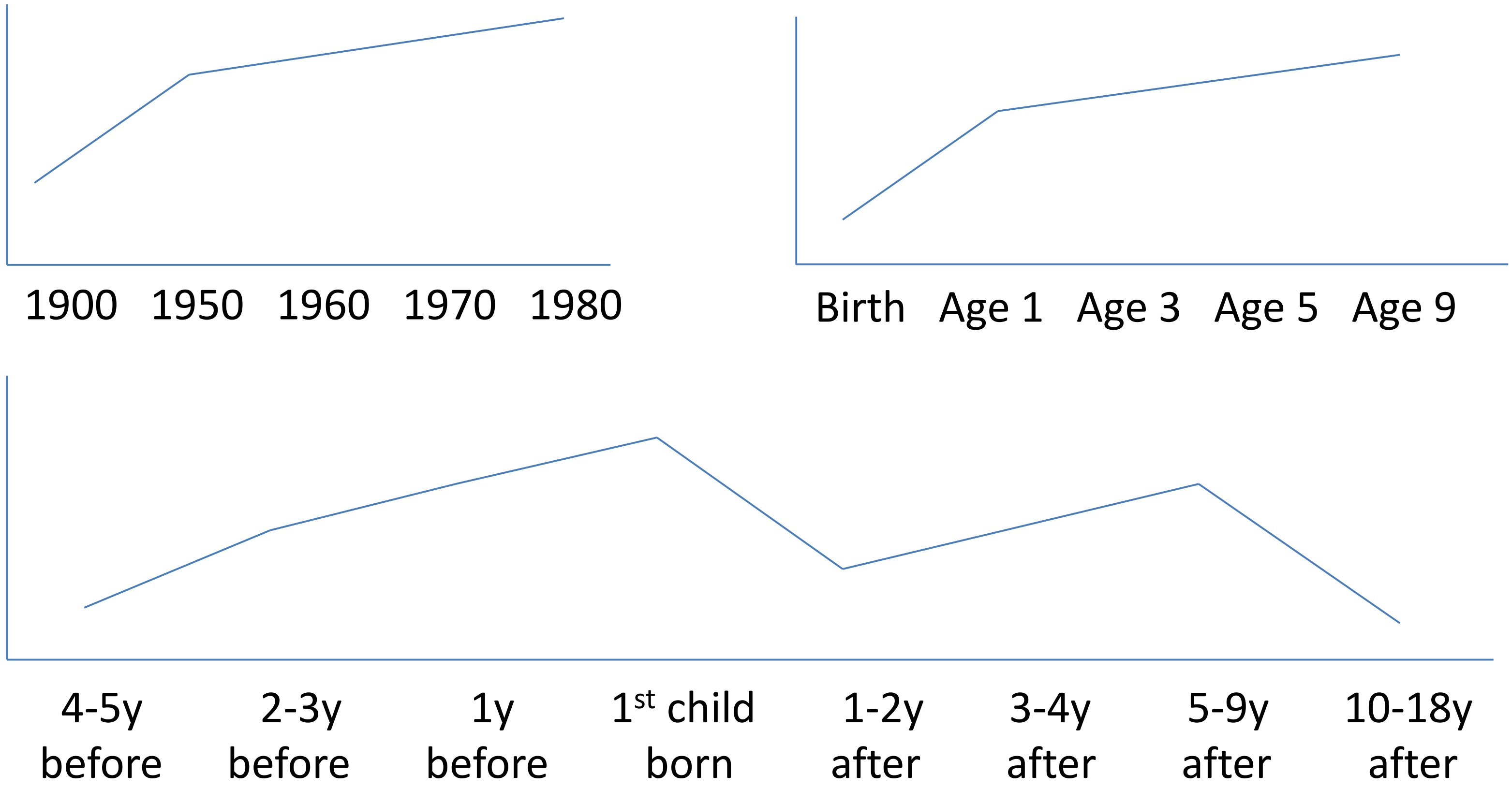
does it make sense to use a line graph when showing values for a nominal variable?





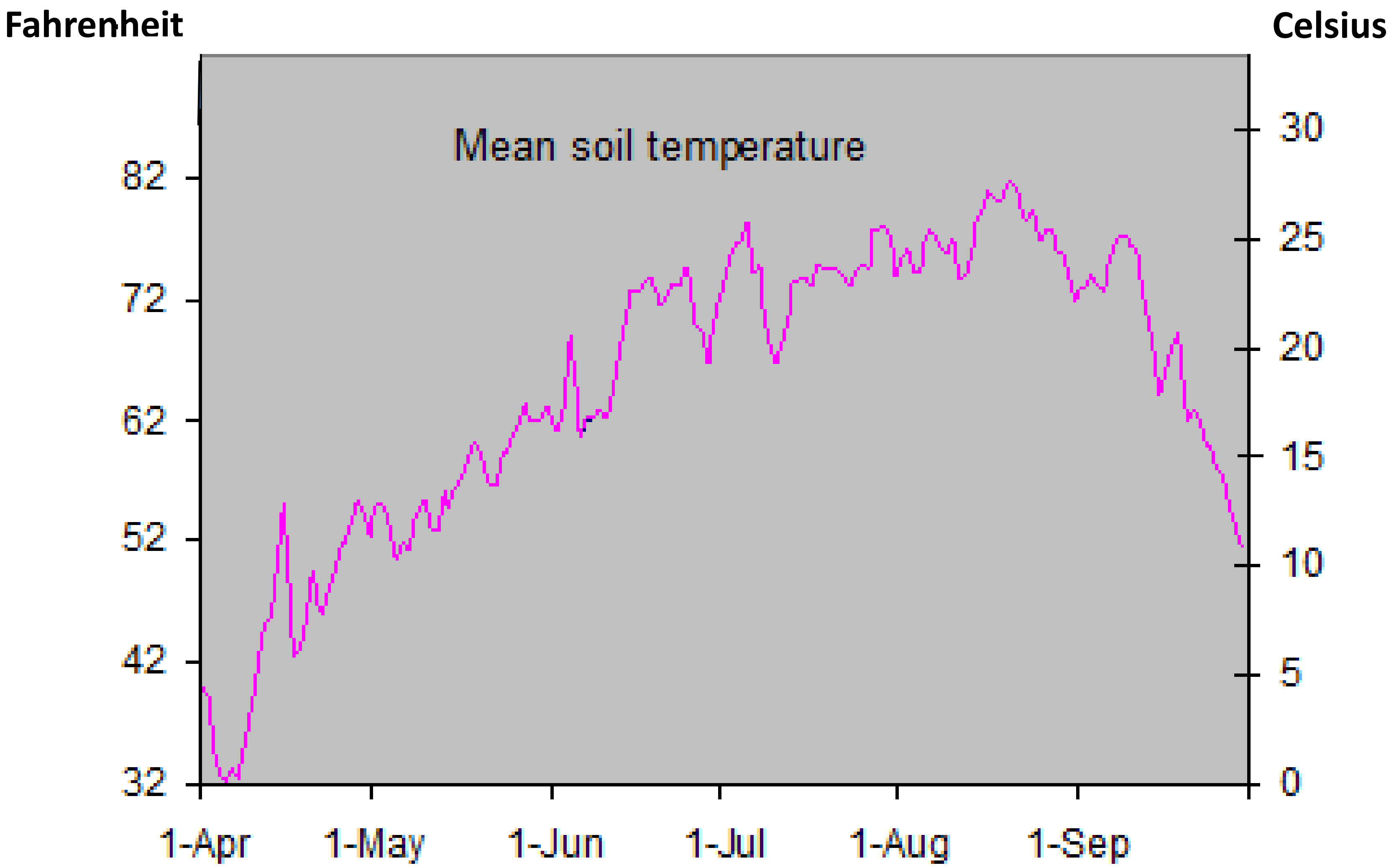
Amazing website by Jennifer Bryan
http://shinyapps.stat.ubc.ca/r-graph-catalog/

do not change scale part way along an axis:

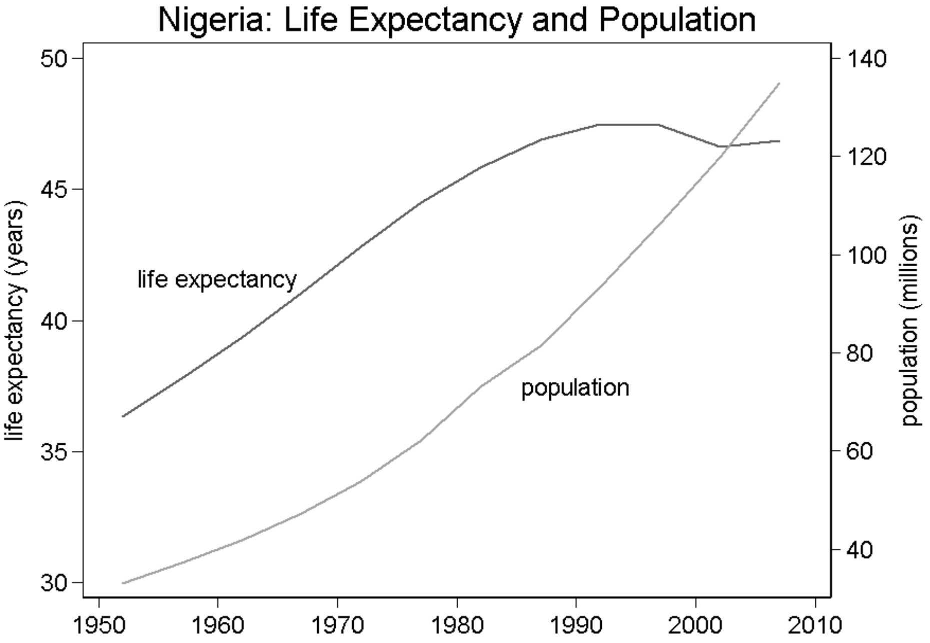


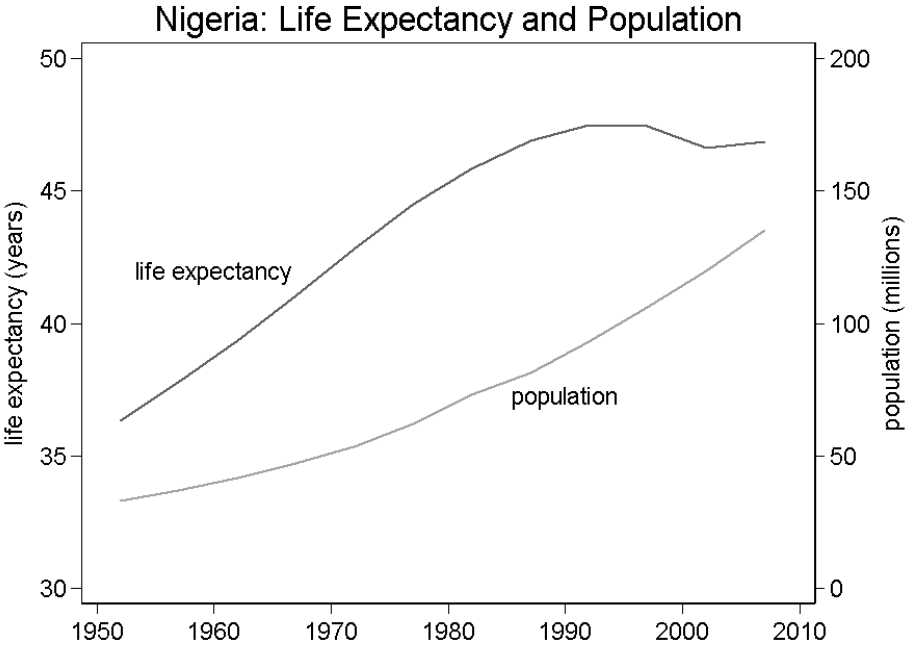
consider when to use two (dual) scales for the same axis ...

two scales measuring the same data with different labels:



two scales showing two variables on the same graph ???

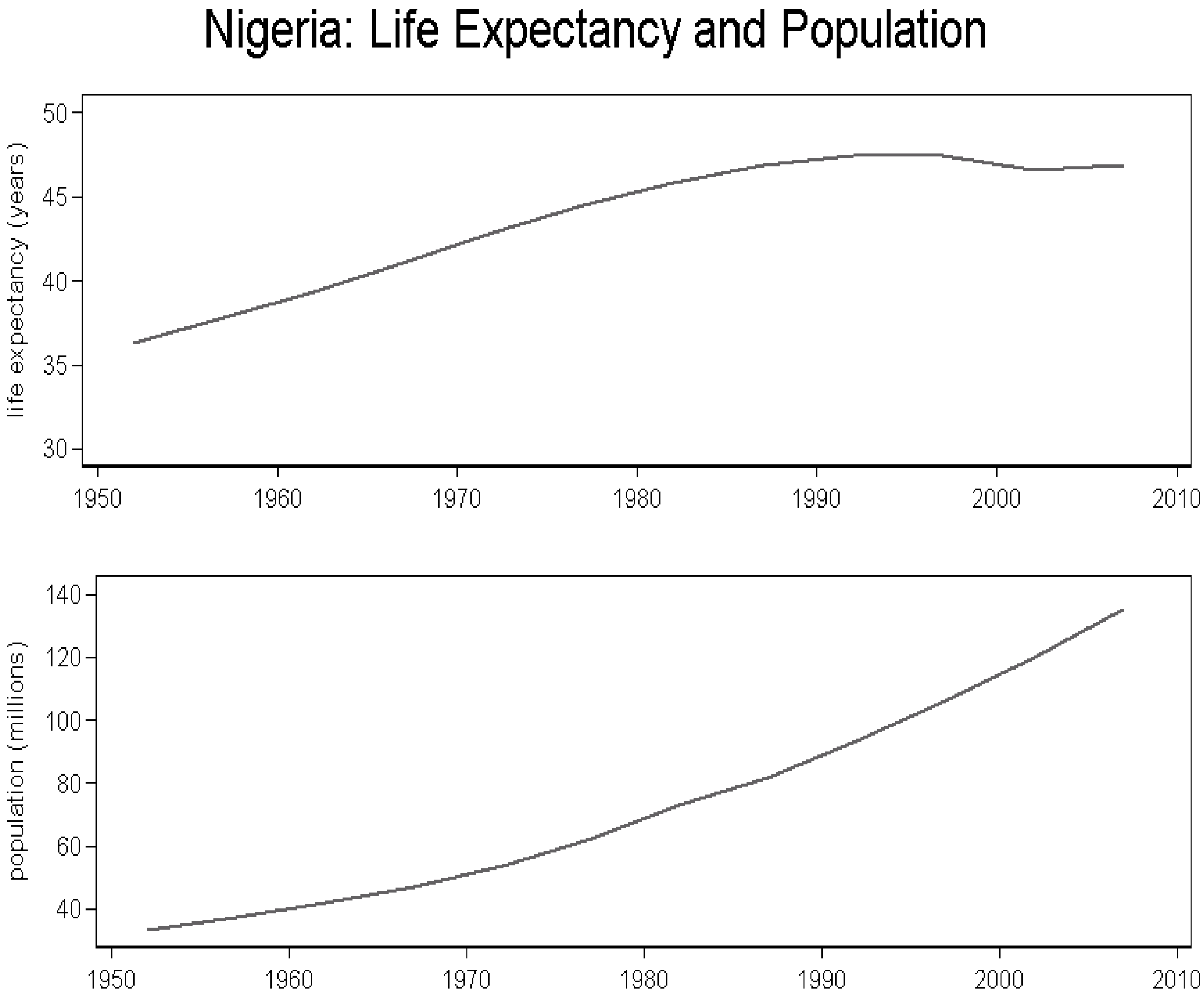




is the slope of one curve relative to the slope of the other curve meaningful?

is the intersection point meaningful?

a possible alternative?



stackoverflow.com response by Hadley Wickham:

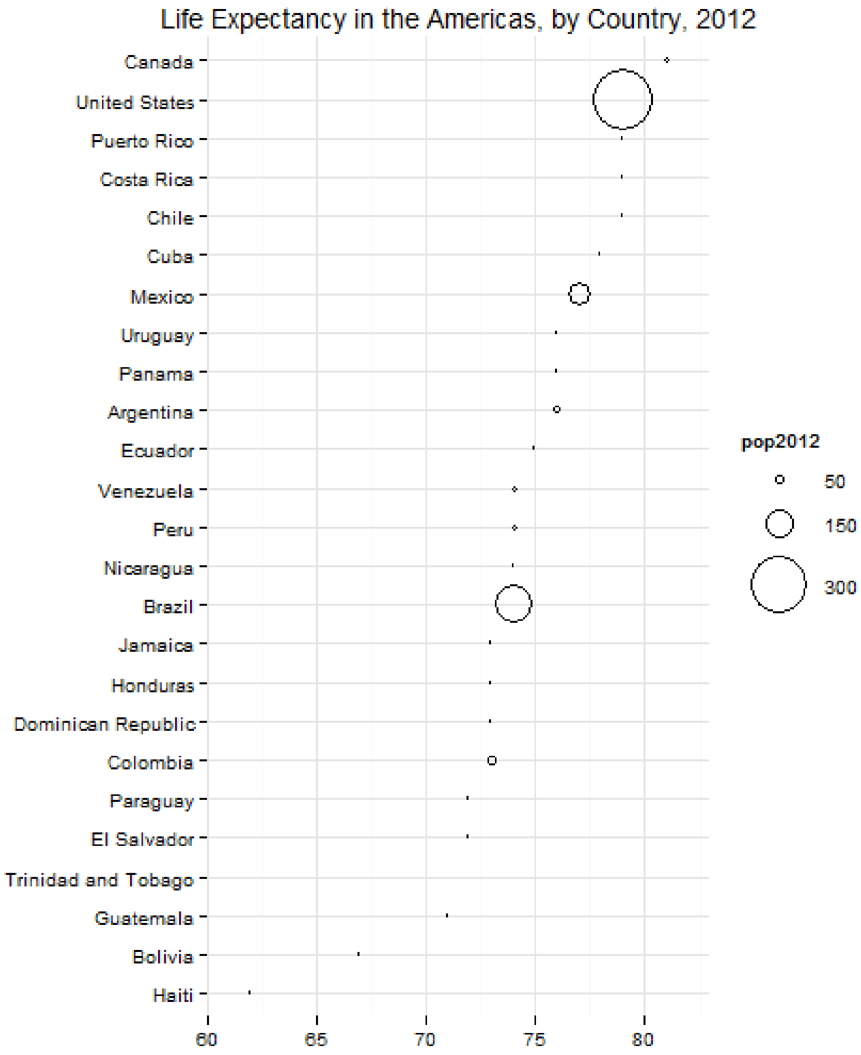
Q: How can ggplot2 be used to make plot with 2 axes, one on left, another on right?

A: “It's not possible in ggplot2 because I believe plots with separate y scales (not y scales that are transformations of each other) are fundamentally flawed.

- They are not invertible: given a point on the plot space, you can not uniquely map it back to a point in the data space.

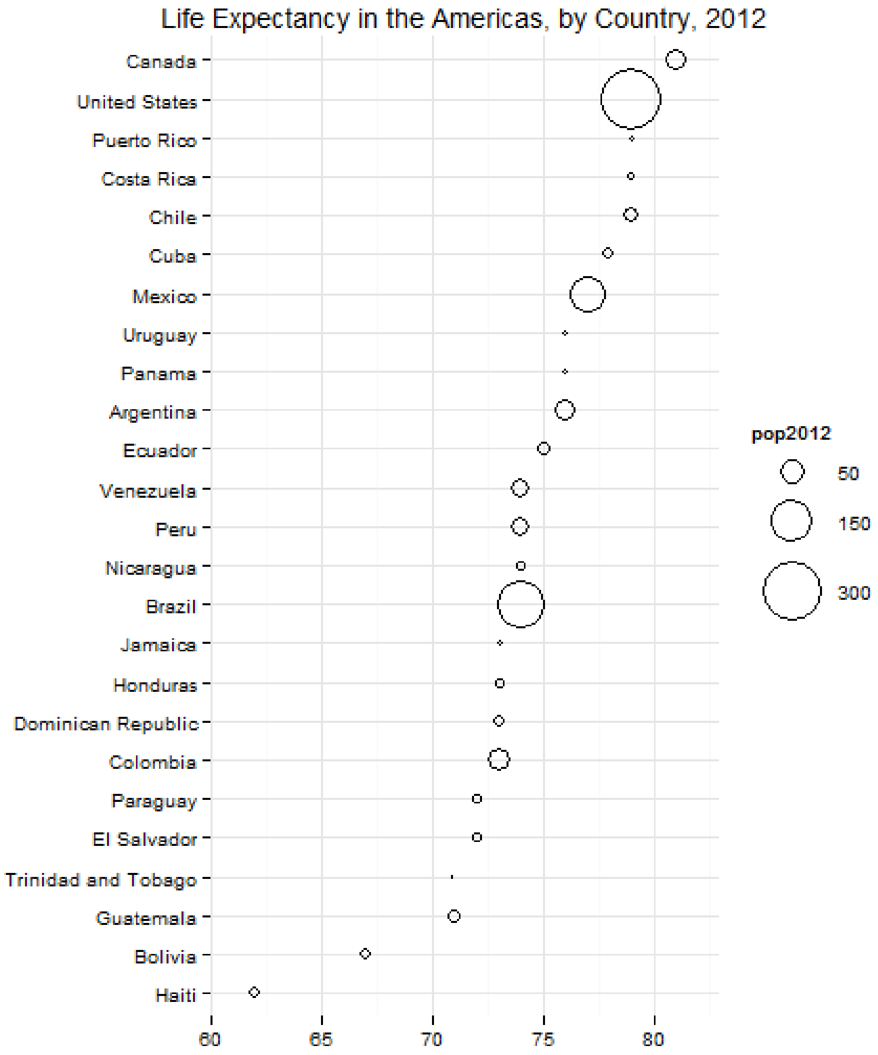
- They are relatively hard to read correctly compared to other options.

- They are easily manipulated to mislead: there is no unique way to specify the relative scales of the axes, leaving them open to manipulation.



Source: 2012 World Population Data Sheet by Pop. Ref. Bureau

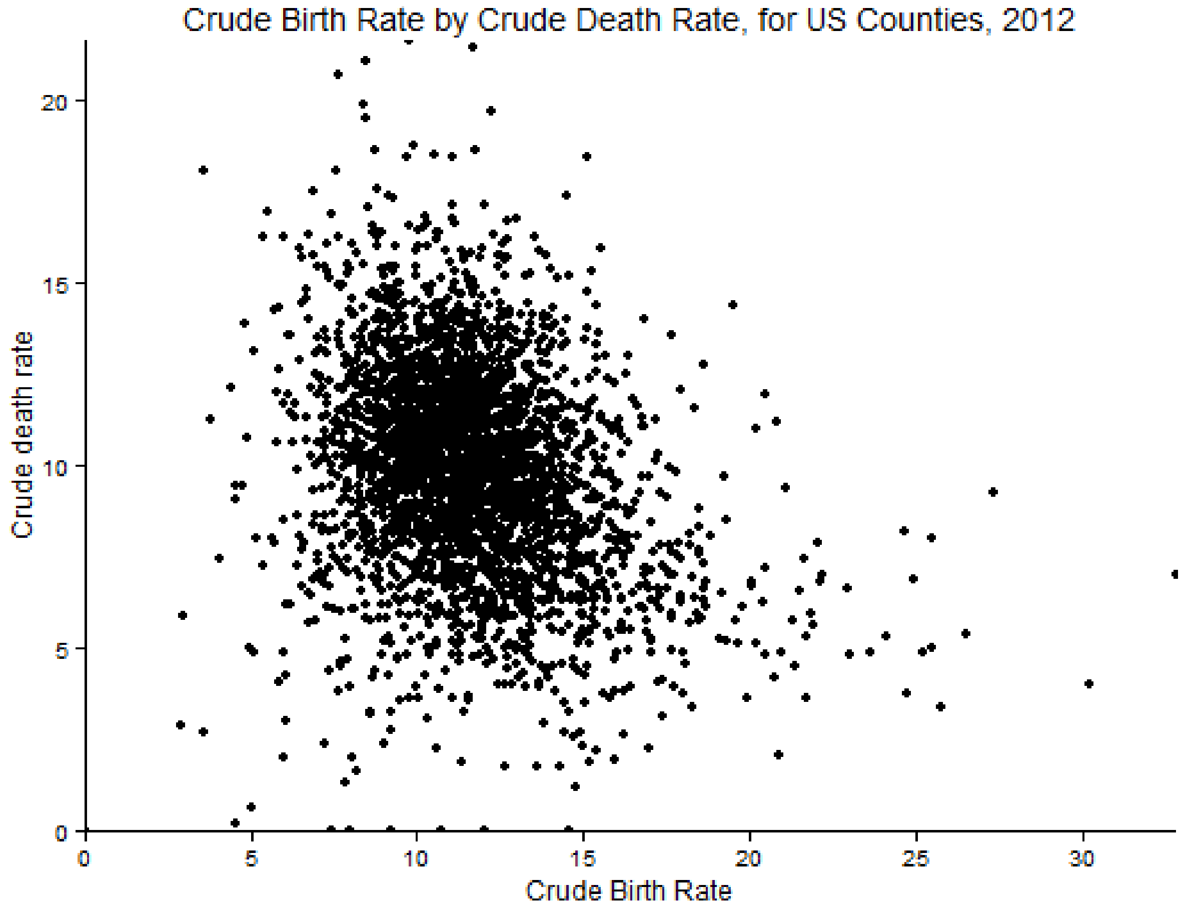
pop2012 value mapped to **radius** of bubble ... doubling value results in quadrupling area!



Source: 2012 World Population Data Sheet by Pop. Ref. Bureau

pop2012 value mapped to **area** of bubble ... Canada = 35, US=314 (about 9 times more)

over-plotting hides data points



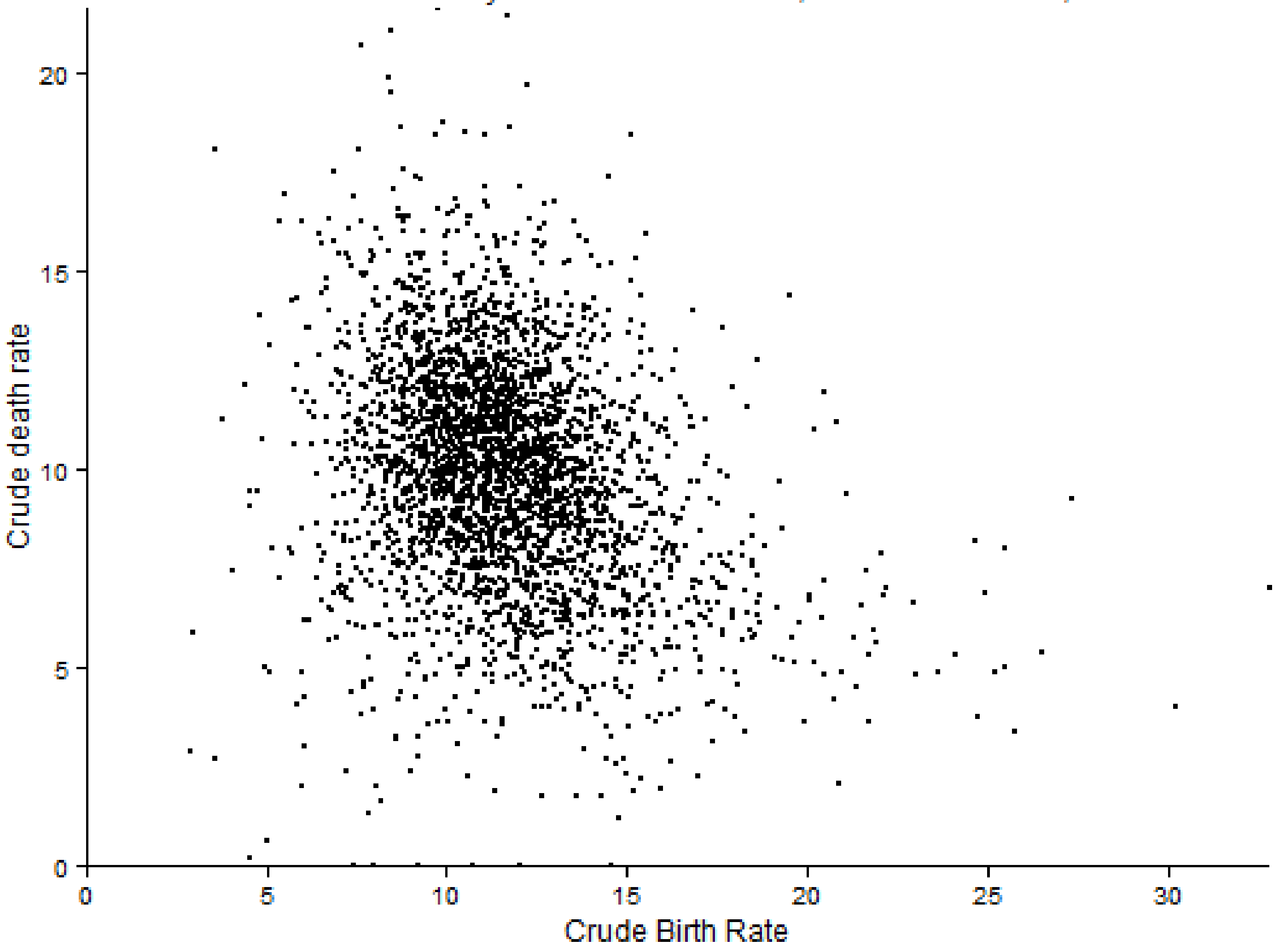
Source: CO-EST2012-alldata: US Census Bureau, Population Division

techniques to accurately display data density include:

adjust point size
adjust point fill
adjust point shape
adjust point transparency
use data stratification
use point jittering

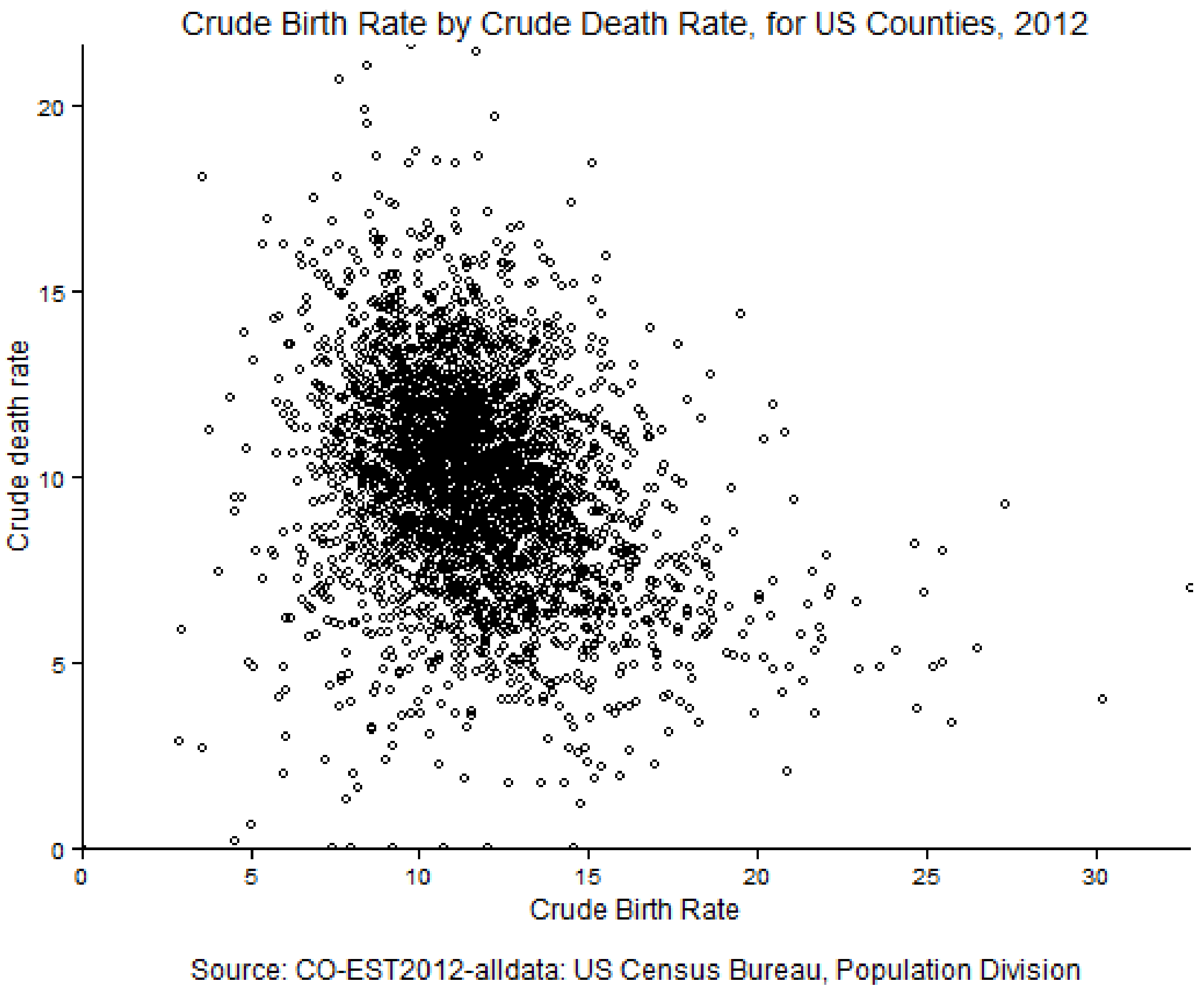
adjust point size

Crude Birth Rate by Crude Death Rate, for US Counties, 2012

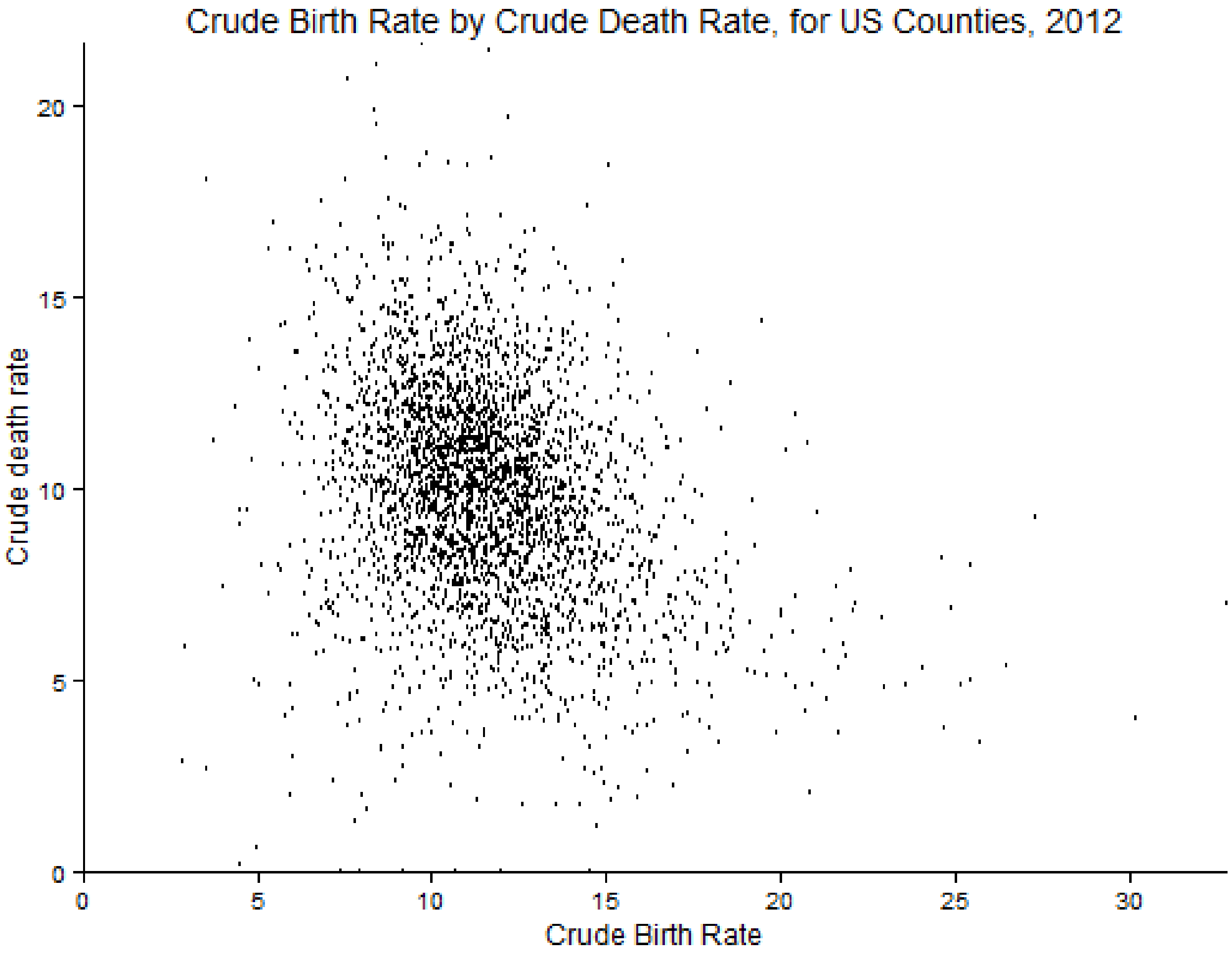


Source: CO-EST2012-alldata: US Census Bureau, Population Division

adjust point fill

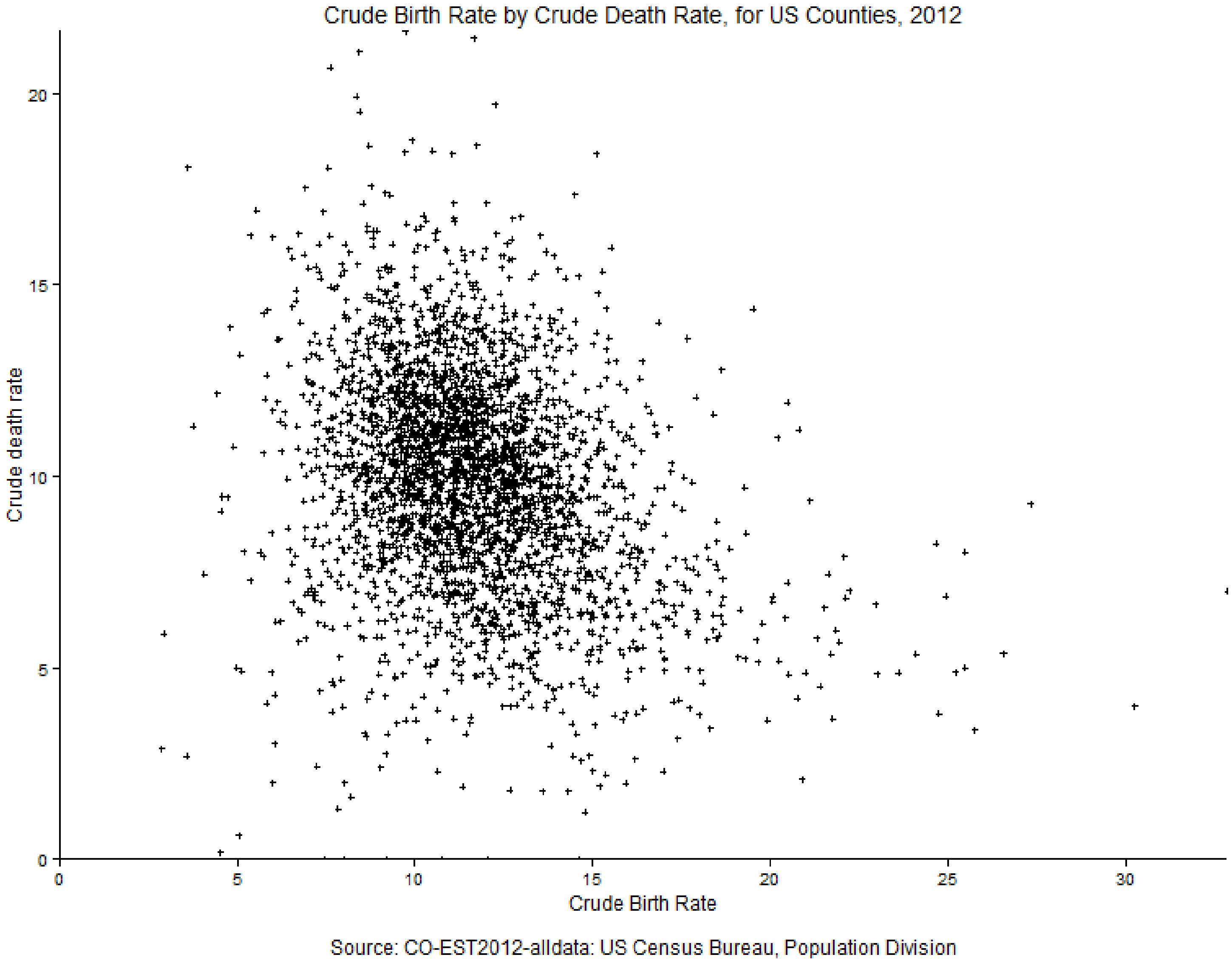


adjust point size and fill

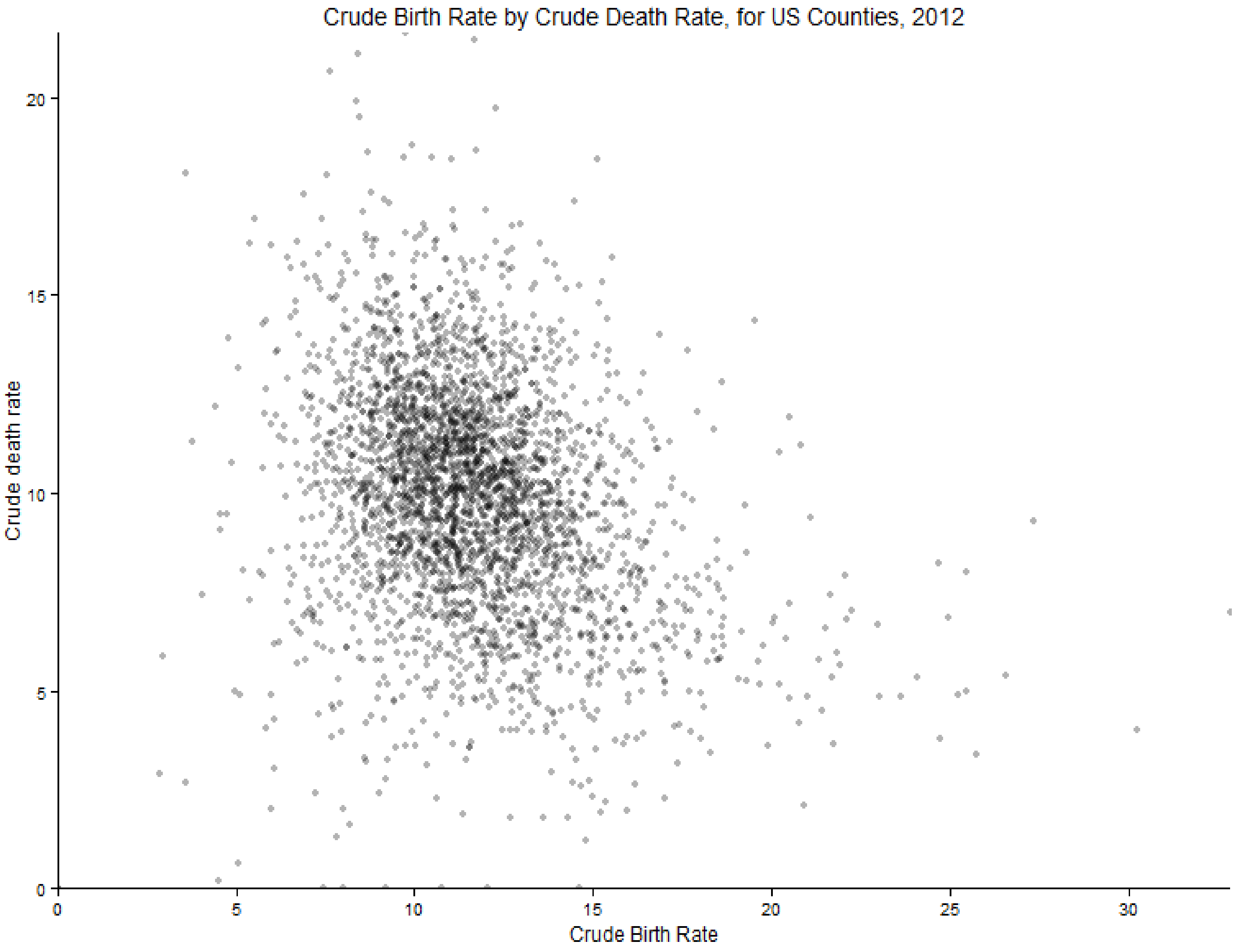


Source: CO-EST2012-alldata: US Census Bureau, Population Division

adjust point shape

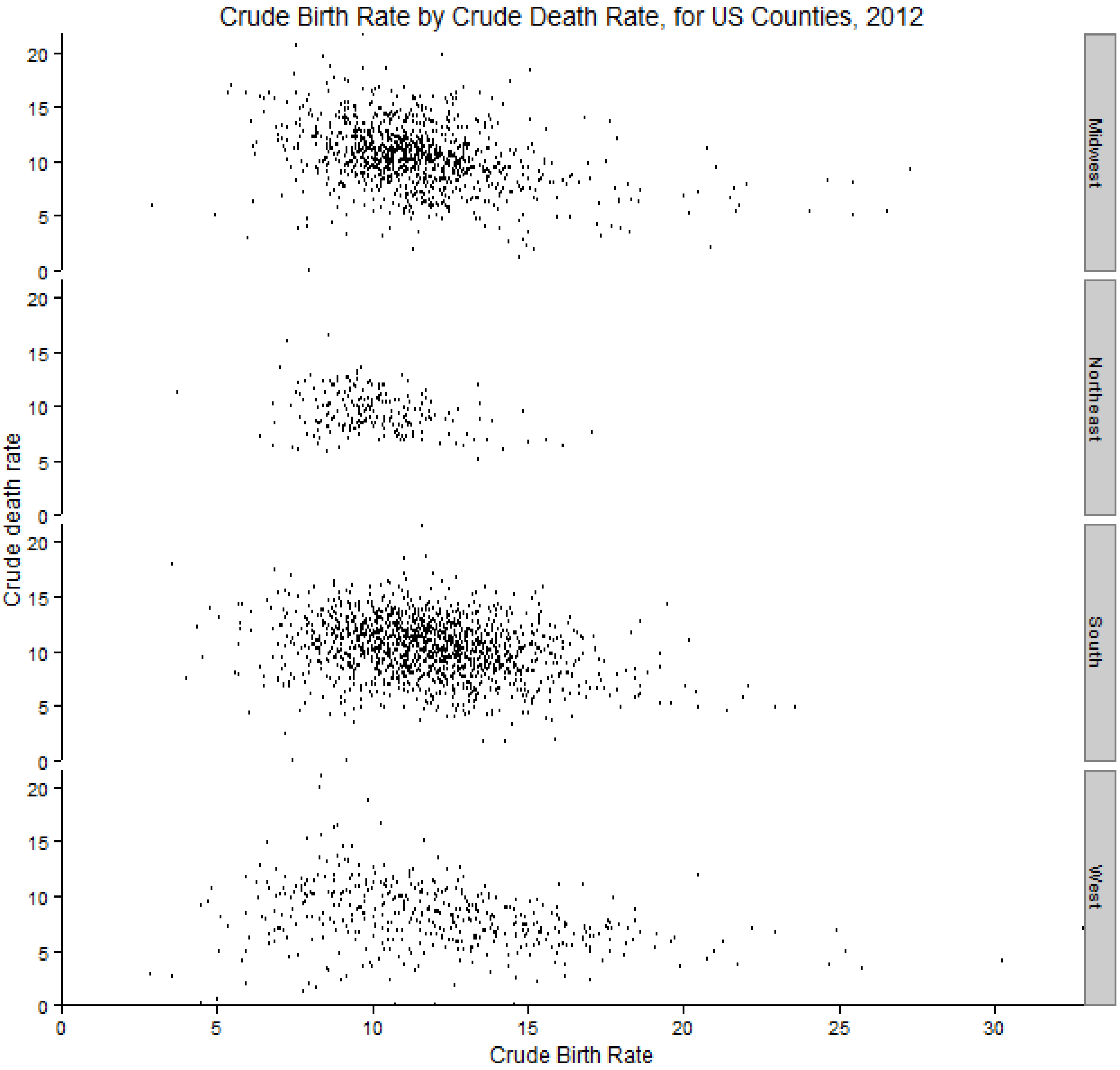


adjust point transparency



Source: CO-EST2012-alldata: US Census Bureau, Population Division

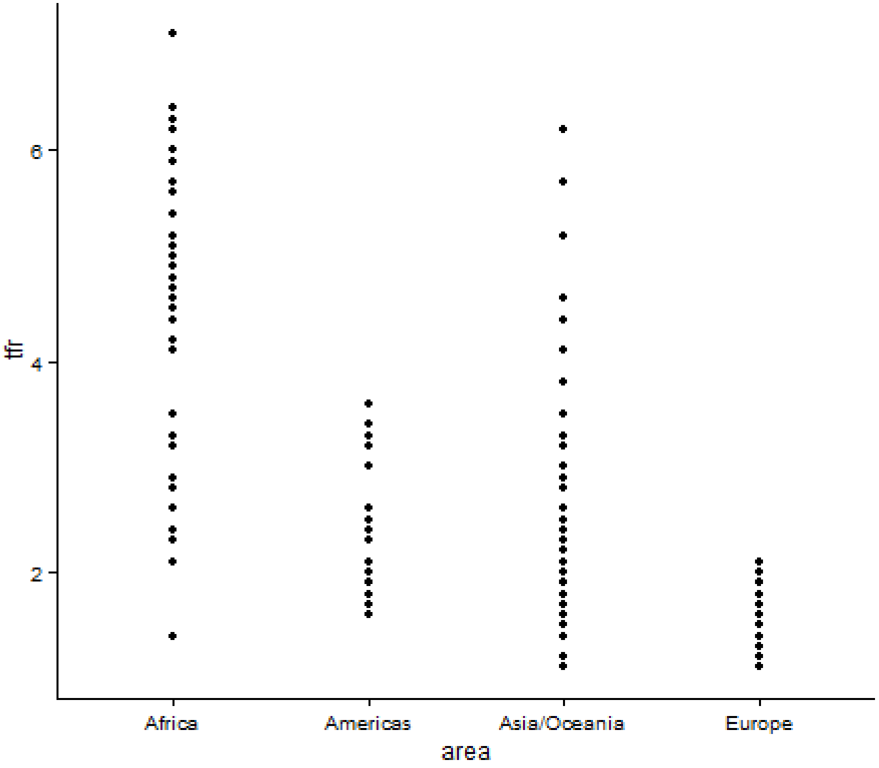
use data stratification

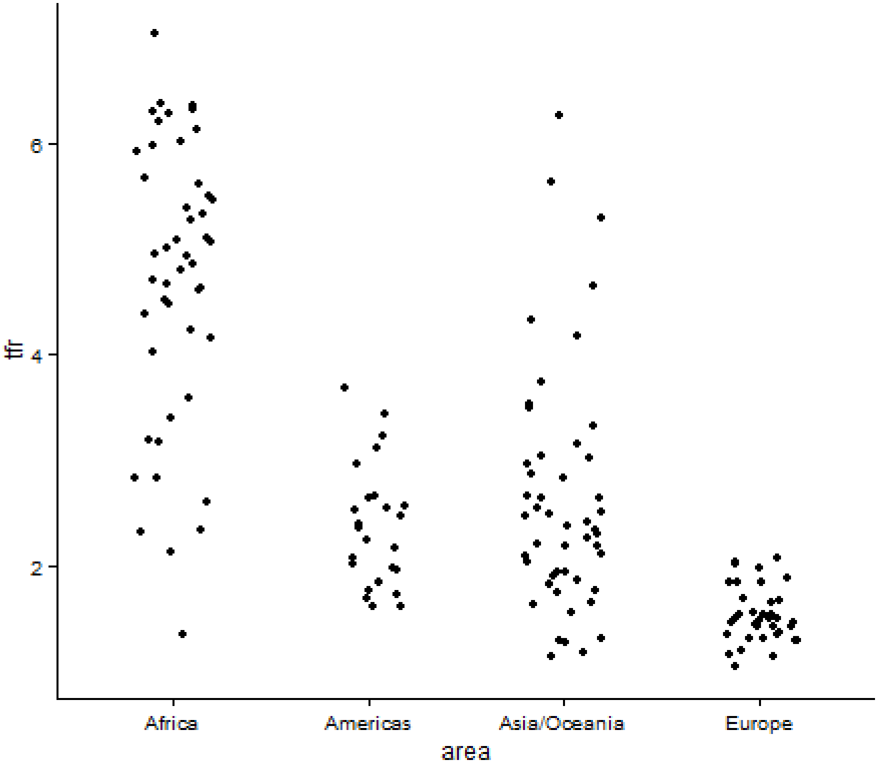


Source: CO-EST2012-alldata: US Census Bureau, Population Division

use point jittering – moving overlapping points a bit

– trade-off: sacrifice positional precision for more accurate display of data density





See Ellis & Dix, *A Taxonomy of Clutter Reduction for Information Visualisation*, IEEE Transactions on Visualization & Computer Graphics, 2007.

III Highlighting **comparisons** of interest

“At the heart of quantitative reasoning is a single question:
Compared to what? ”

- Edward Tufte

Tufte, E. Envisioning Information. Graphics Press. Cheshire, CT. 1990.

Highlight comparisons

1. Determine true quantity (or quantities) of interest, for example ...

   - magnitude of A and magnitude of B?

     or

   - difference between magnitude of A and magnitude of B?

     or

   - ratio of magnitude of A to magnitude of B?

2. Make sure the data is easily seen
   - size, contrast, not hidden by other data markers (points, lines, areas), labels, legends, tick marks, or gridlines

3. Show the data, not just summary measures, when possible

4. Involve perceptual tasks high on Cleveland's list of performing accurate judgements
   - position along a common scale
   - position along identical, non-aligned scales
   - length
   - angle, slope
   - area
   - color

5. Consider proximity, alignment and ordering

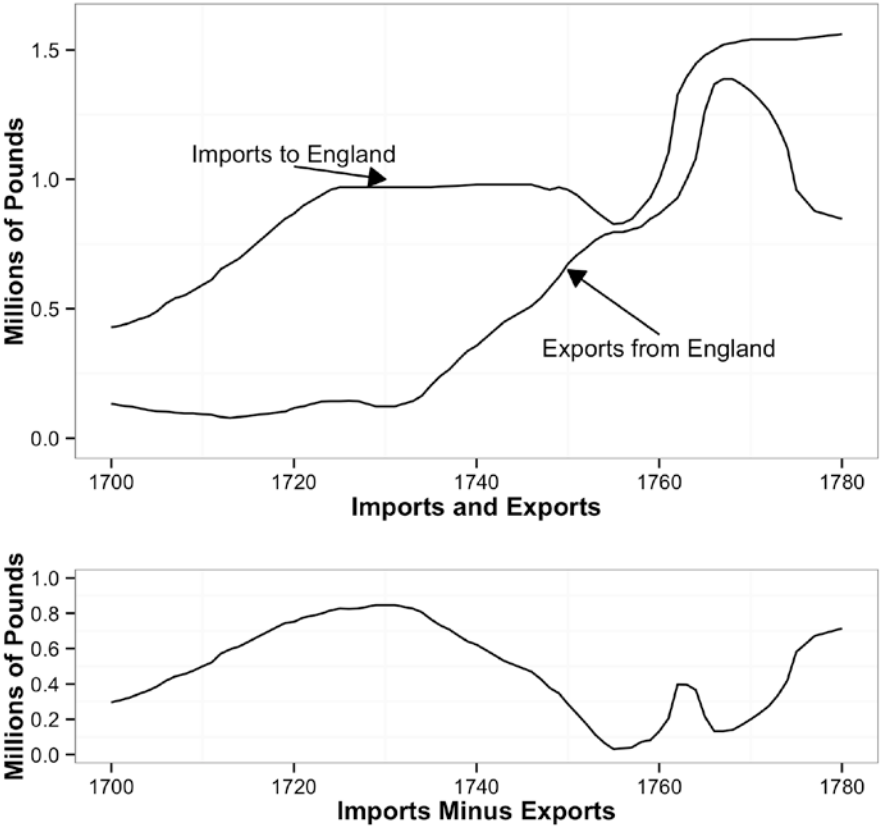
determine true quantity of interest

To show the difference between A and B, graph the difference between A and B.
You may want to graph A and B on their own too, but don’t stop there.

show imports and exports to and from England

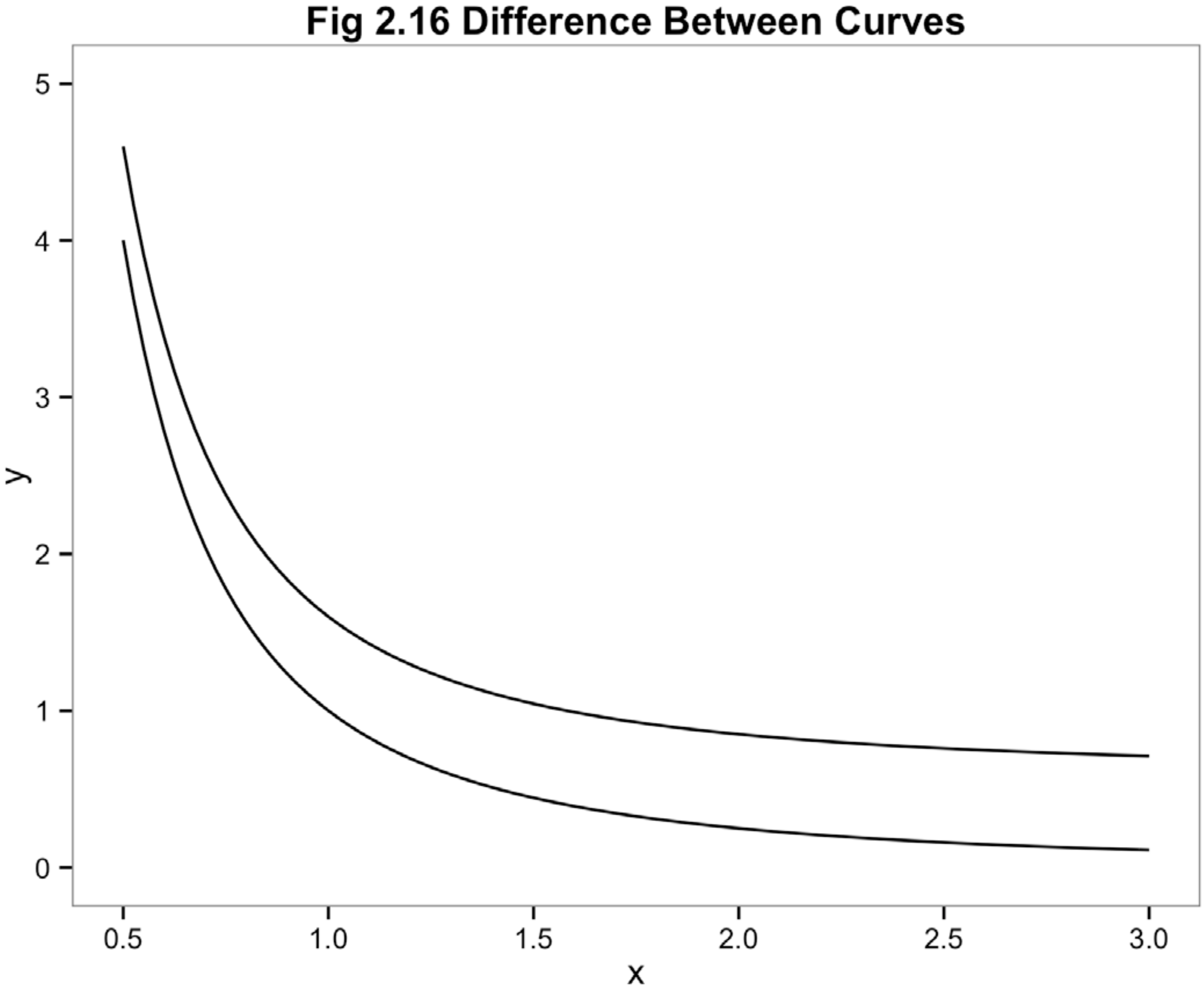
to show balance of trade, imports - exports, graph that also

Fig 2.15 Playfair's Balance-of-Trade Data: Imports Minus Exports



R graph Catalog by Jennifer Bryan http://shinyapps.stat.ubc.ca/r-graph-catalog

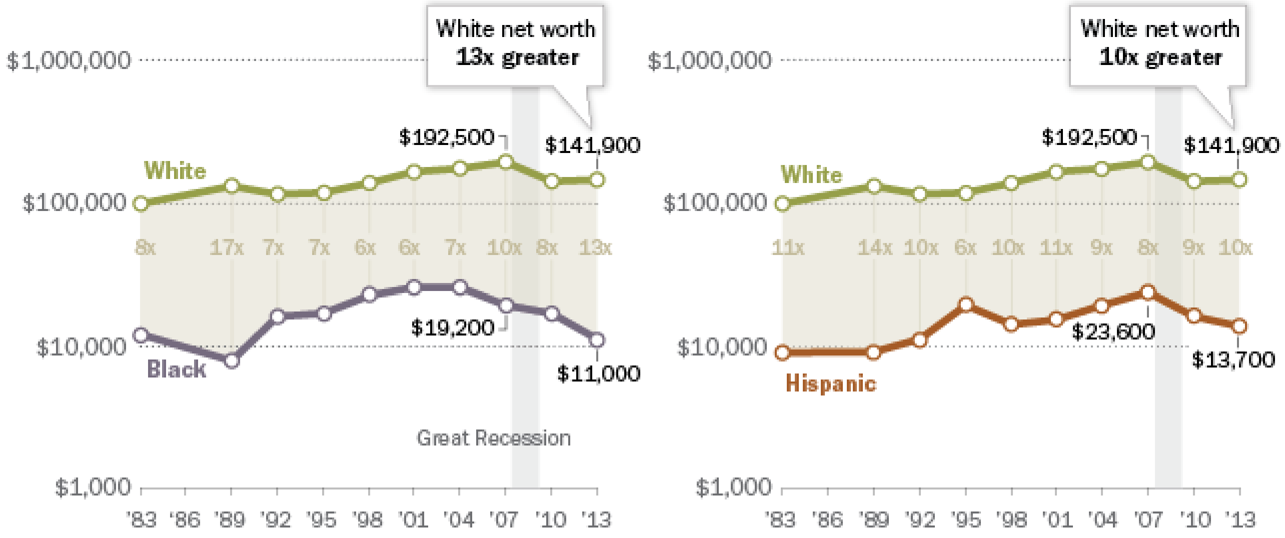
sometimes surprisingly difficult to calculate the difference between curves:



don’t ask viewers to do extra work

Racial, Ethnic Wealth Gaps Have Grown Since Great Recession

Median net worth of households, in 2013 dollars



Notes: Blacks and whites include only non-Hispanics. Hispanics are of any race. Chart scale is logarithmic; each gridline is ten times greater than the gridline below it. Great Recession began Dec. '07 and ended June '09.

Source: Pew Research Center tabulations of Survey of Consumer Finances public-use data

PEW RESEARCH CENTER

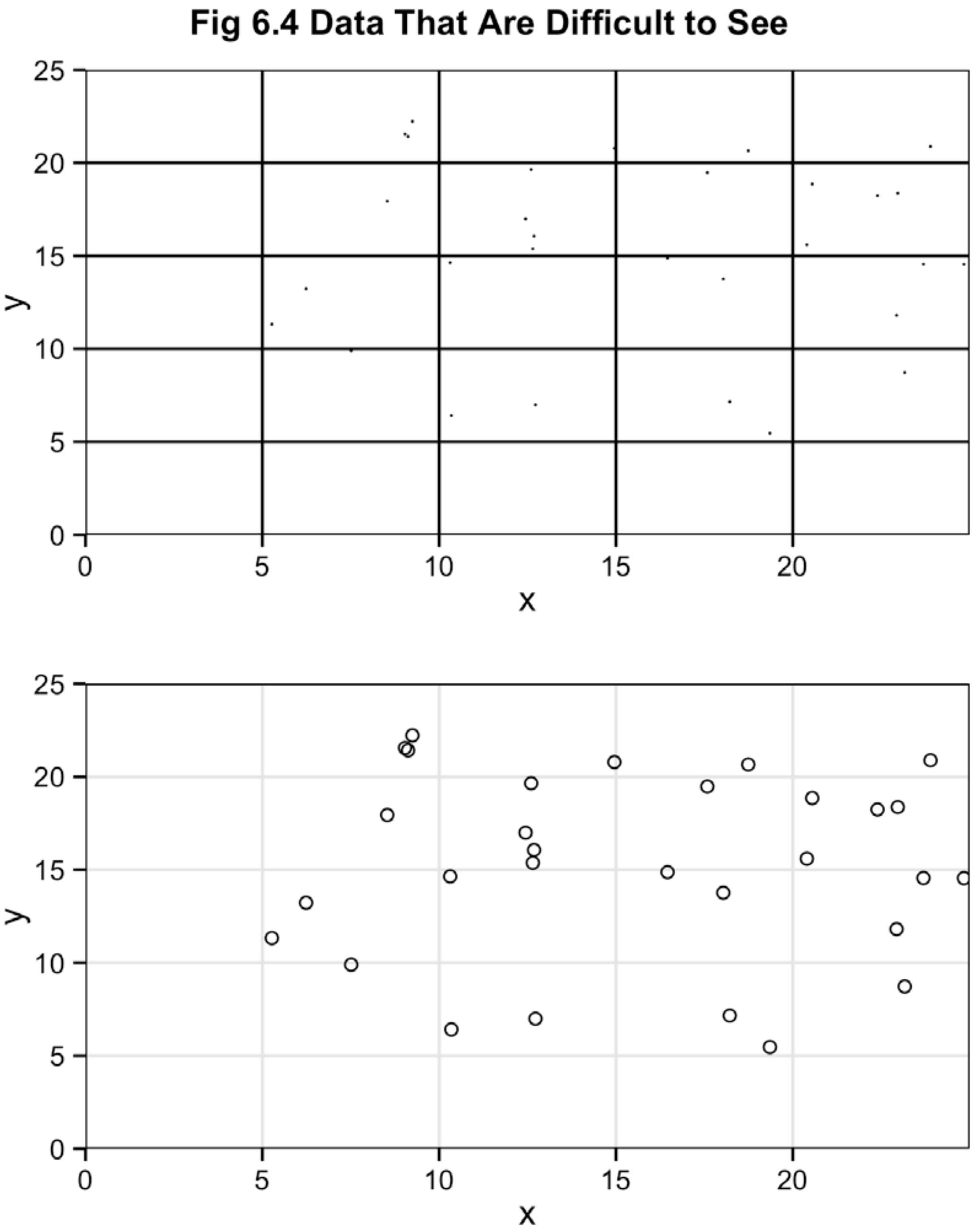
Pew Research Center Fact Tank. December 12, 2014.
Wealth inequality has widened along racial, ethnic lines since end of Great Recession.

http://www.pewresearch.org/fact-tank/2014/12/12/racial-wealth-gaps-great-recession/

show the data - make sure it can be easily seen.

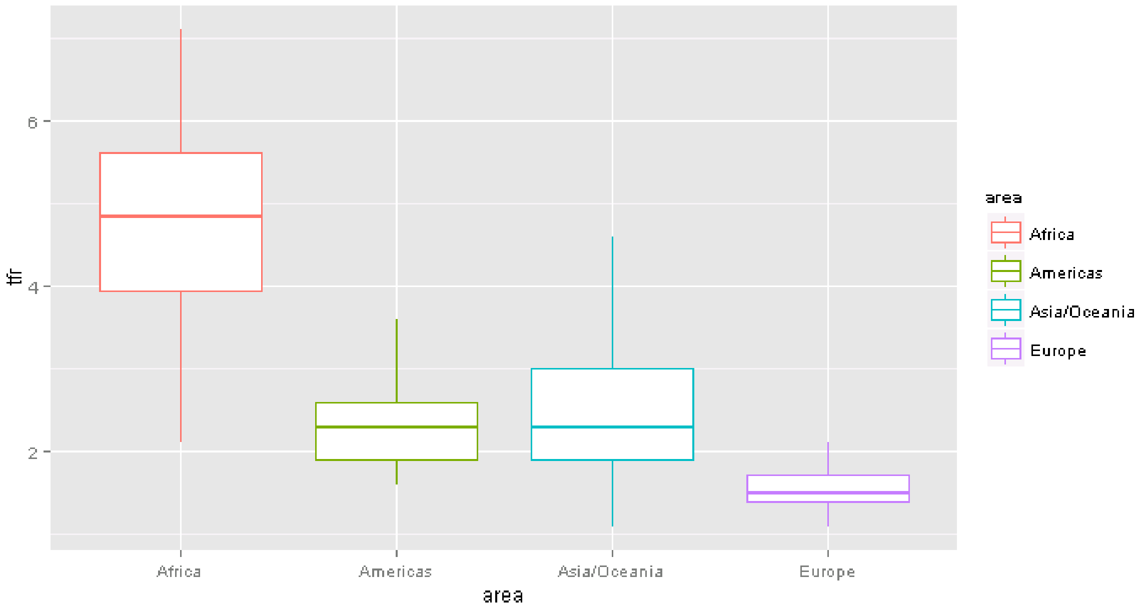
consider:

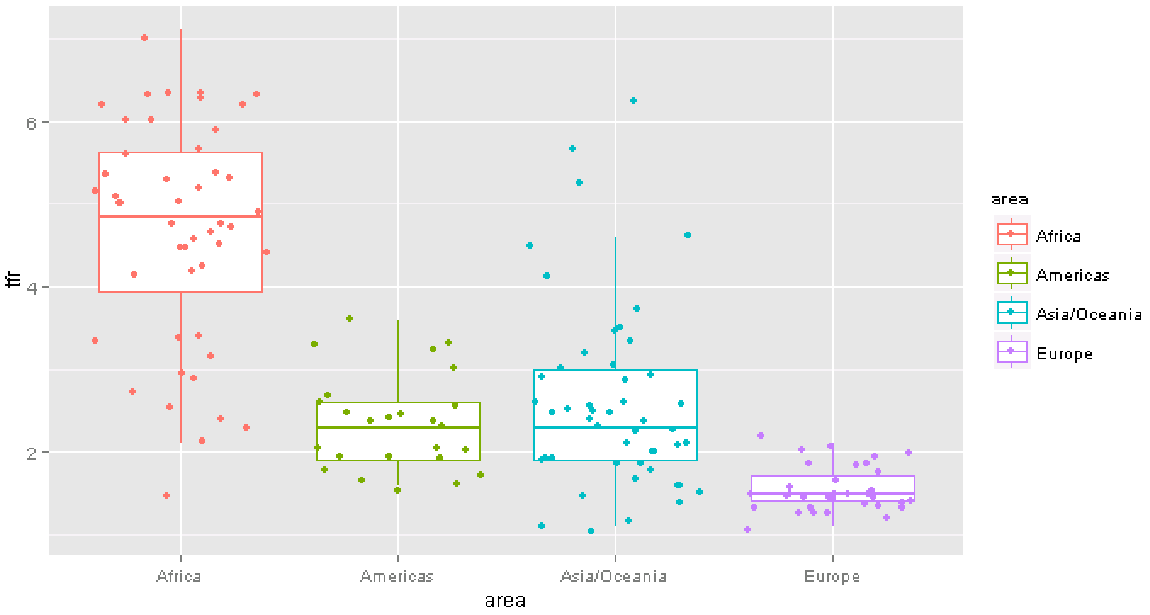
- size
- contrast
- overplotting
- hidden by tick marks, legends, labels, gridlines, reference lines, text annotations



R graph Catalog by Jennifer Bryan http://shinyapps.stat.ubc.ca/r-graph-catalog

show the data, not just summary measures, when possible





show the data - display and compare **distributions** of continuous variables.

how?

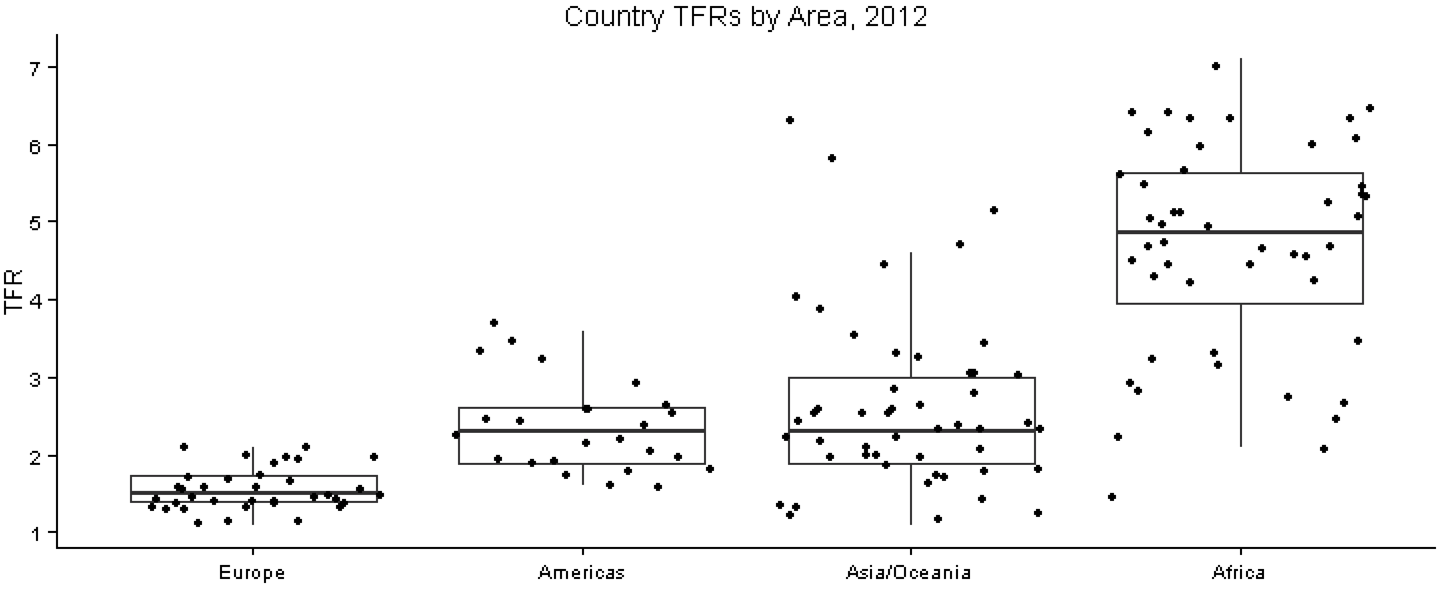
- box plot
- violin plot
- histogram
- density diagram

why?

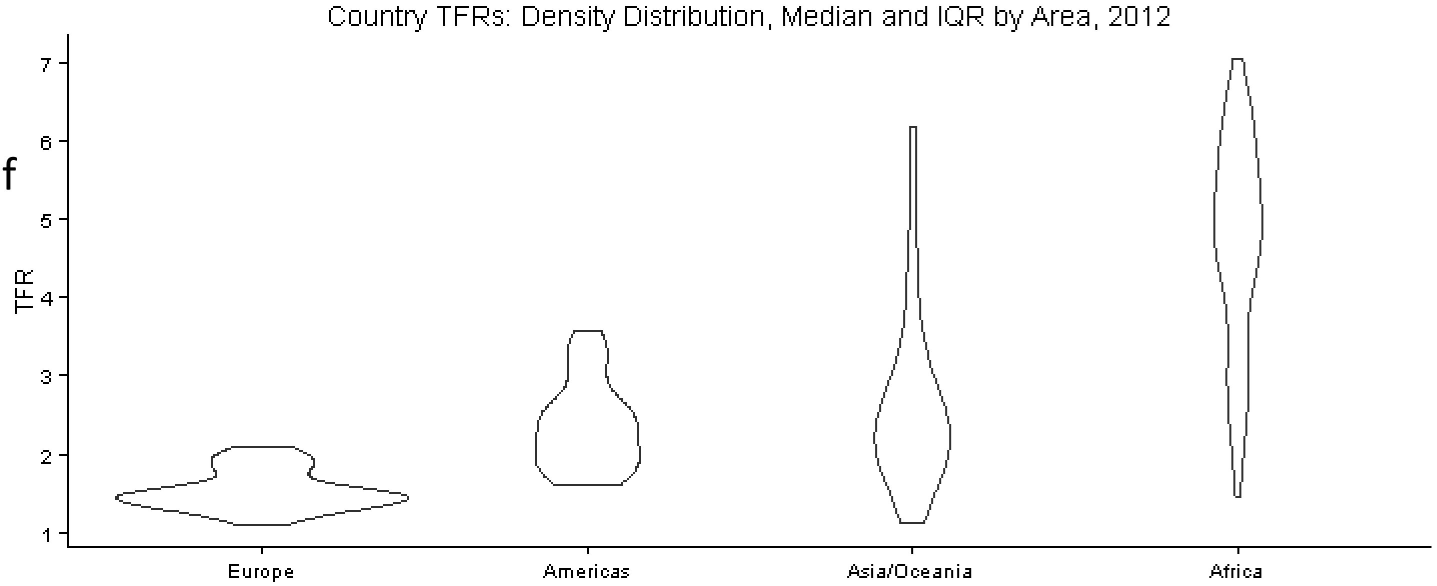
look for shape of distribution: normal, uniform, bi-modal, skewed, etc.
also look for outliers, data errors, missing data

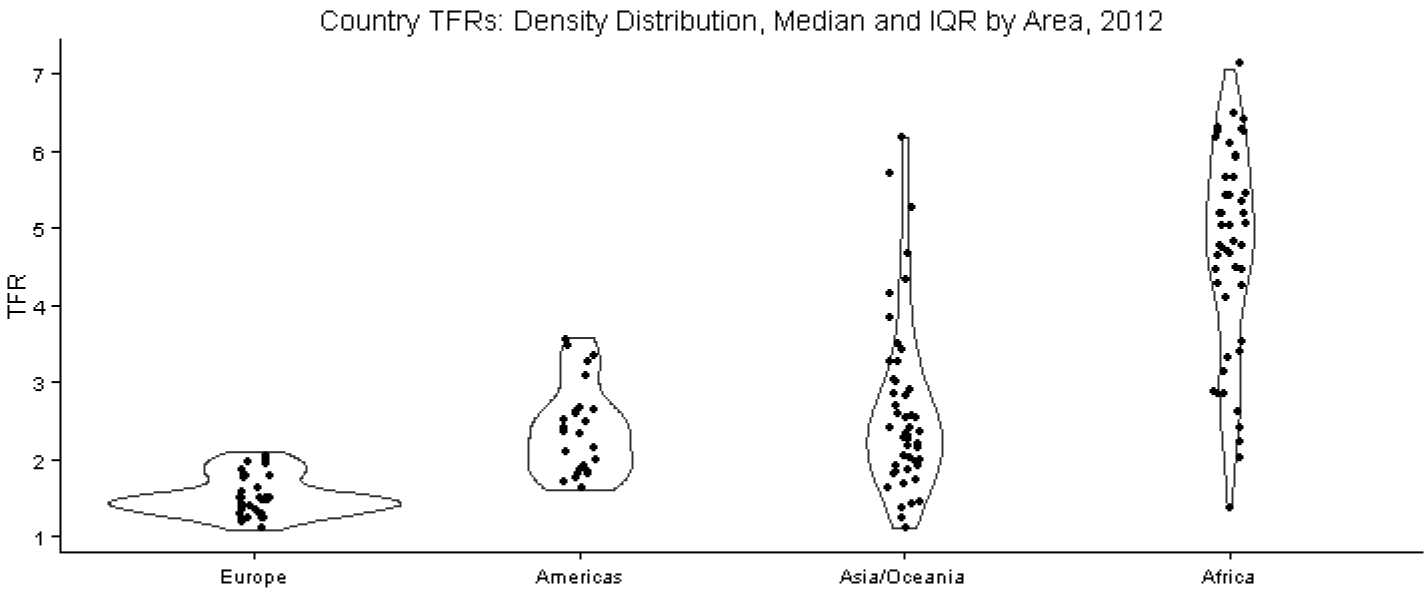
understand data before modeling

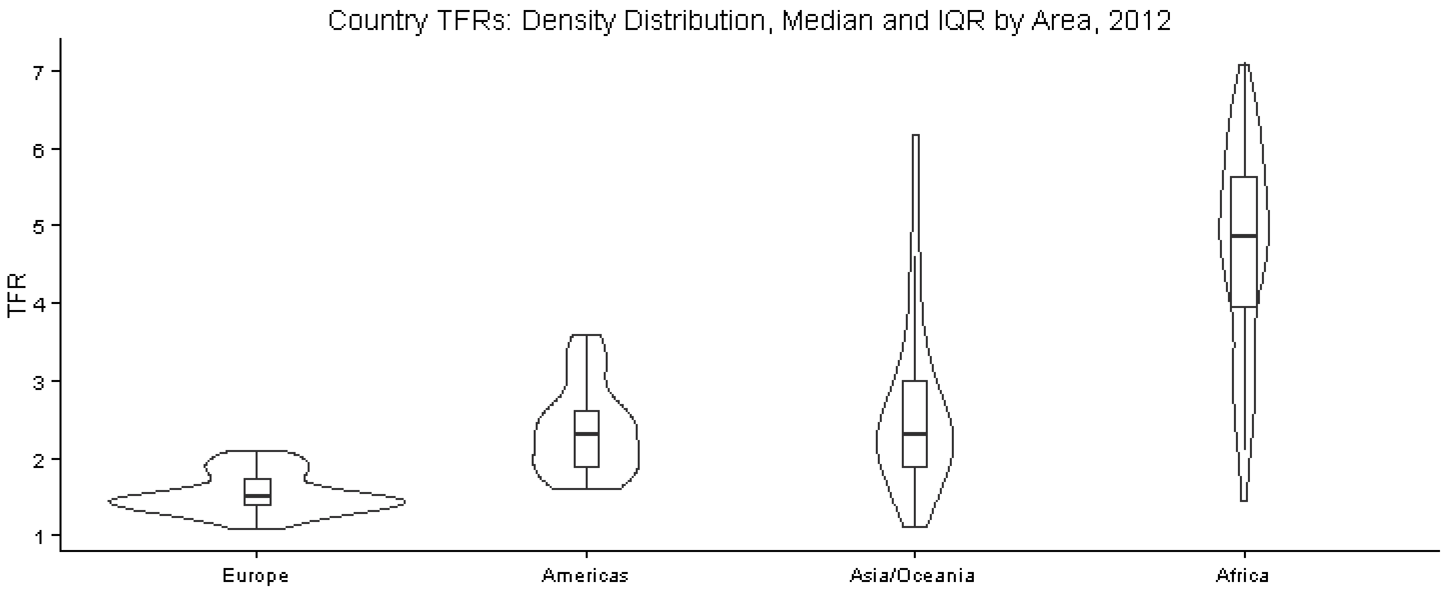
box plot:
box showing median, iqr and contiguous values up to 1.5 times upper and lower quartiles



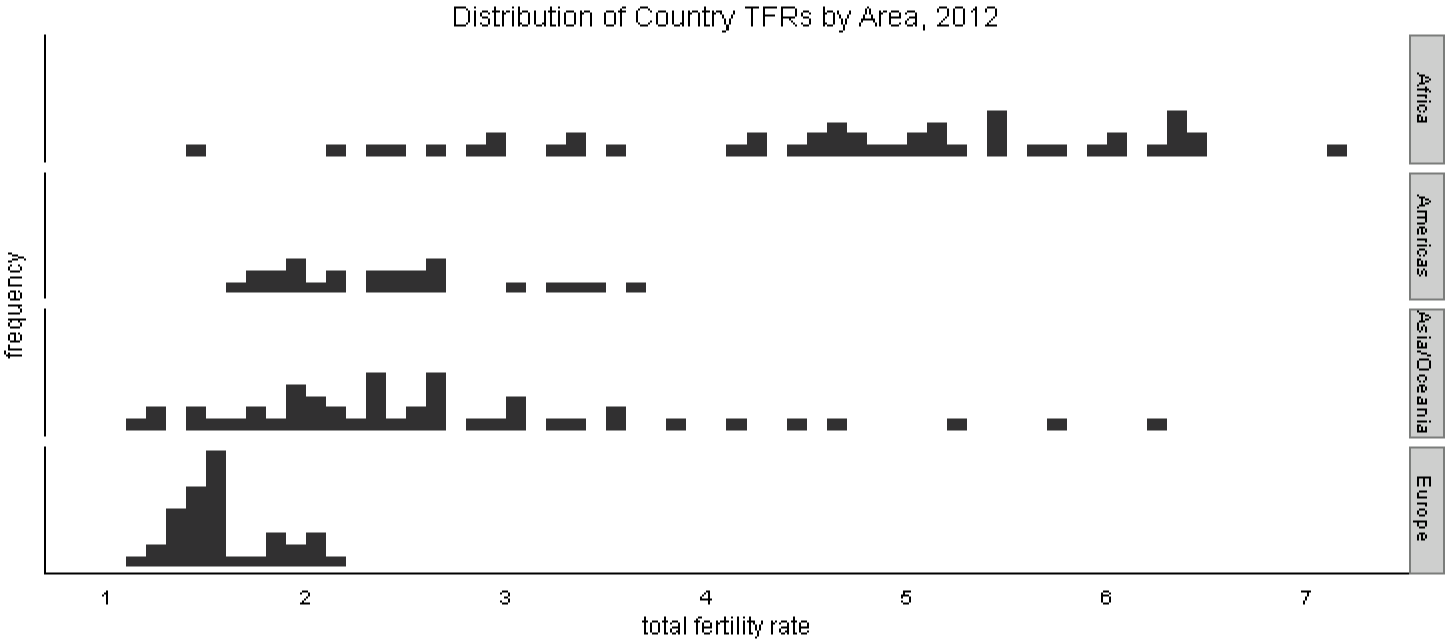
violin plot:
symmetric shape showing density of data values



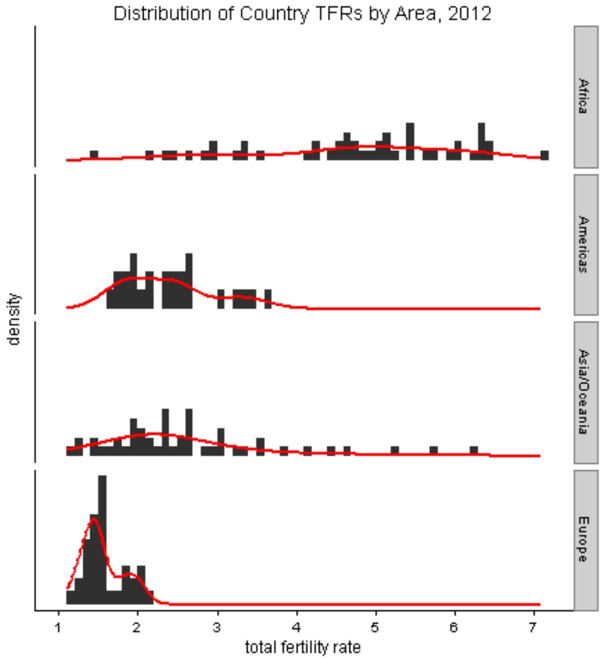

Country TFRs: Density Distribution, Median and IQR by Area, 2012
TFR
7
6
5
4
3
2
1
Europe
Americas
Asia/Oceania
Africa


Country TFRs: Density Distribution, Median and IQR by Area, 2012
TFR
7
6
5
4
3
2
1
Europe
Americas
Asia/Oceania
Africa

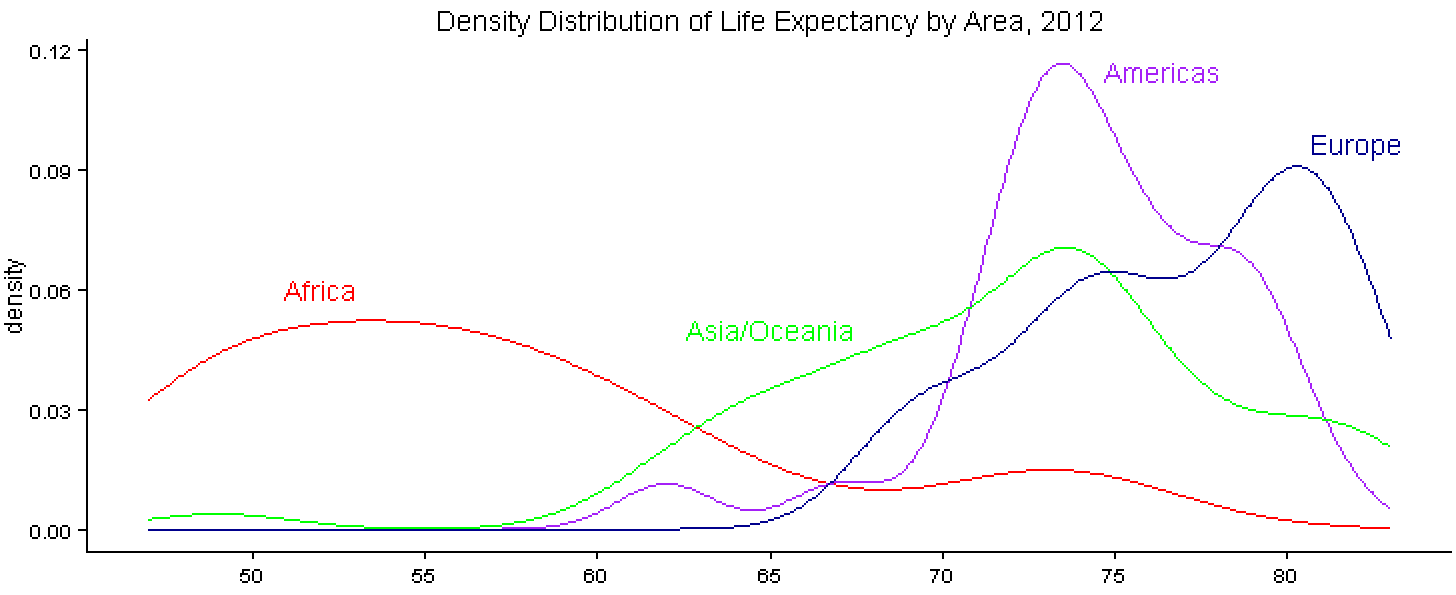
histogram: frequency shows number of values within each bin of continuous values



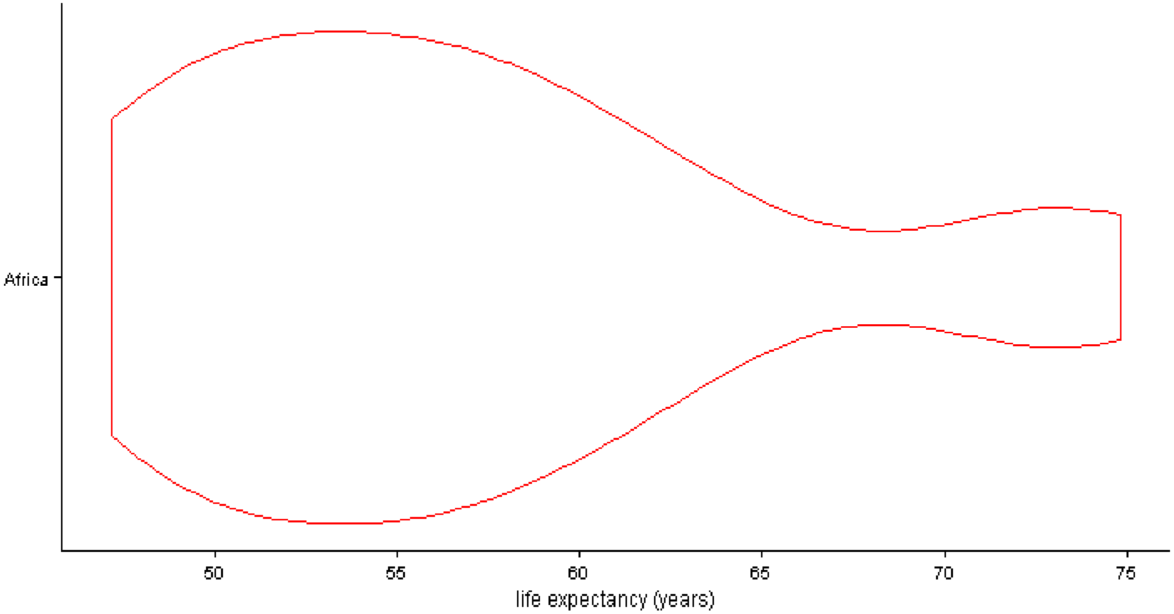
histogram may show proportion of values within each bin, rather than frequency



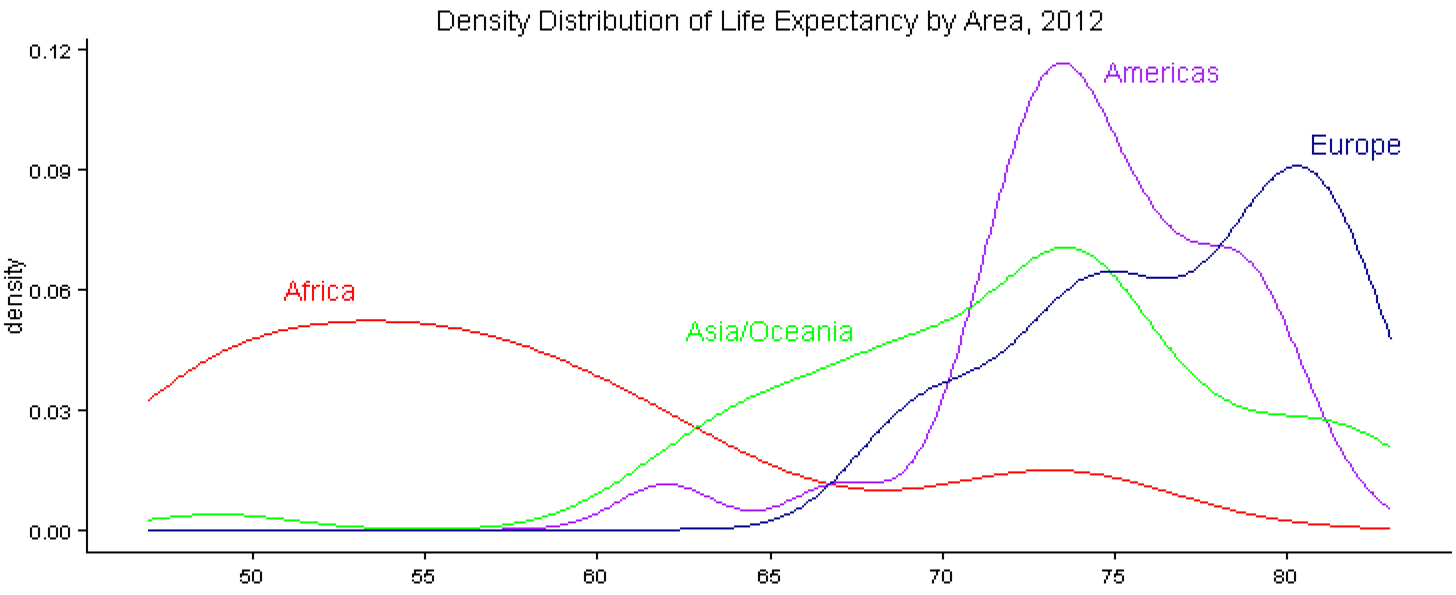
density curve: similar to histogram, but is smooth and continuous



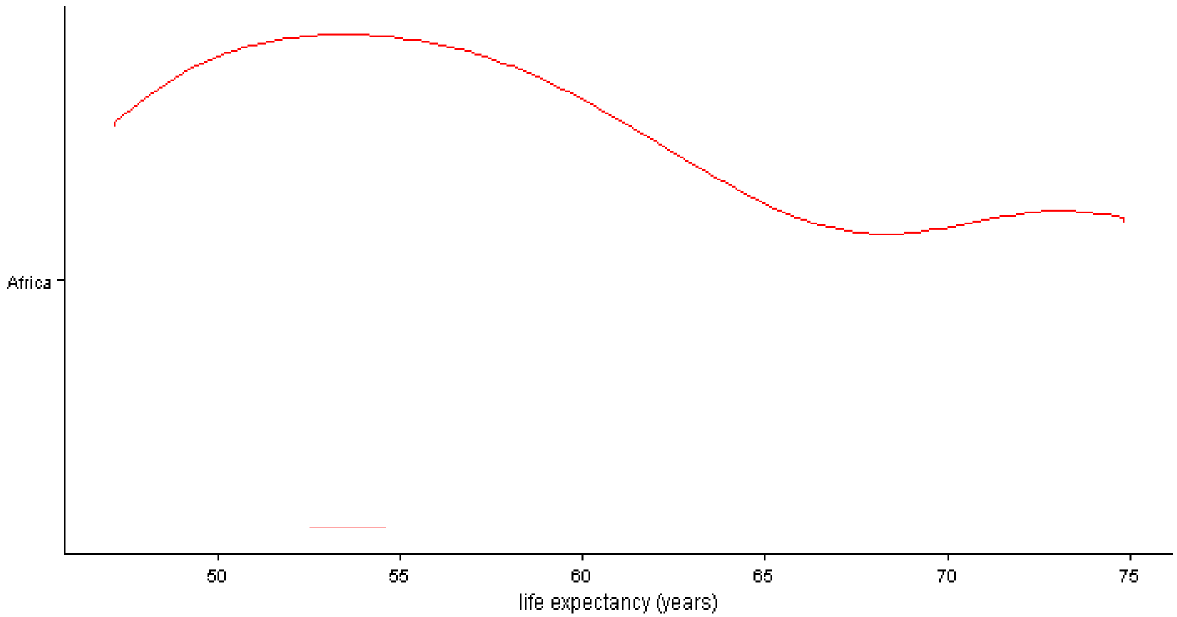
violin plot: symmetric shape showing density of data values



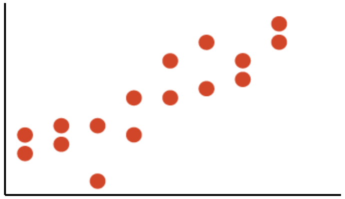
Density curve: similar to histogram, but is smooth and continuous



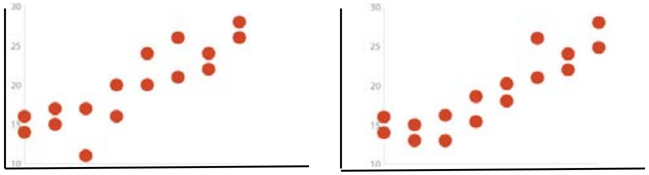
violin plot: symmetric shape showing density of data values



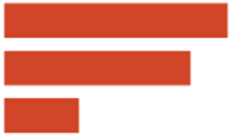
1. Position along a common scale



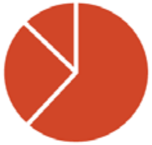
2. Position along identical, nonaligned scales



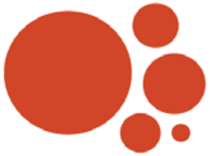
3. Length



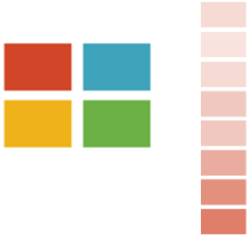
4. Angle-slope



5. Area



6. Color hue and color intensity



Involve perceptual tasks high on William Cleveland’s list of performing accurate judgements.

“Order is based on the theory of visual perception, on experiments in graphical perception, and on informal experimentation.”

- William S. Cleveland, *The Elements of Graphing Data*, 1985.

via Data + Design, https://infoactive.co/data-design/titlepage01.html

Pie charts may be the most criticized graph form, but are surprisingly common.

They encode values in angles and areas, which are hard for humans to judge.

It is easier to judge position along a common scale,
which is why many think dot plots are more effective than pie charts.

Fig 1.1 Similar Pie Wedges

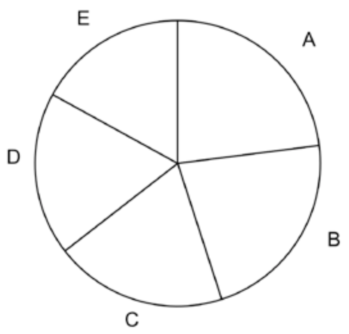


Fig 1.2 Similar Pie Wedges: Dot Plot

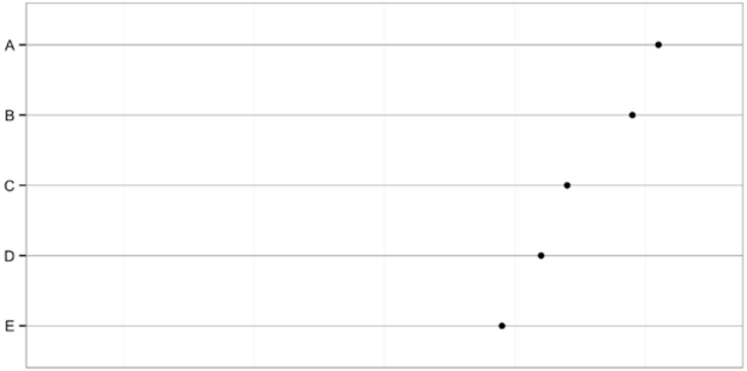
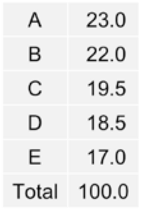


Fig 1.3 Similar Pie Wedges: Table



A	23.0
B	22.0
C	19.5
D	18.5
E	17.0
Total	100.0

angle/area

position along a common scale

R graph Catalog by Jennifer Bryan http://shinyapps.stat.ubc.ca/r-graph-catalog

Fig 2.1 Structured Data Set

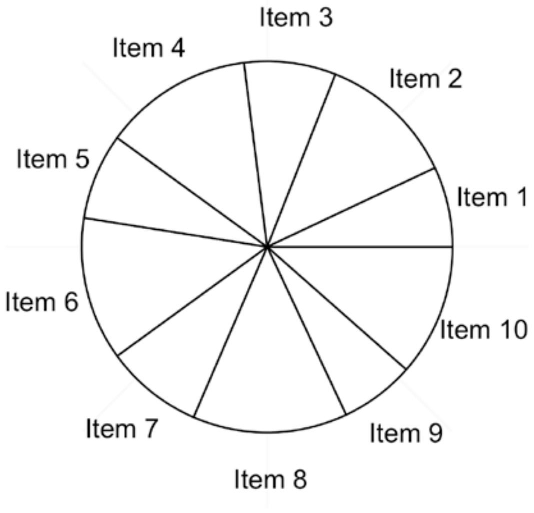
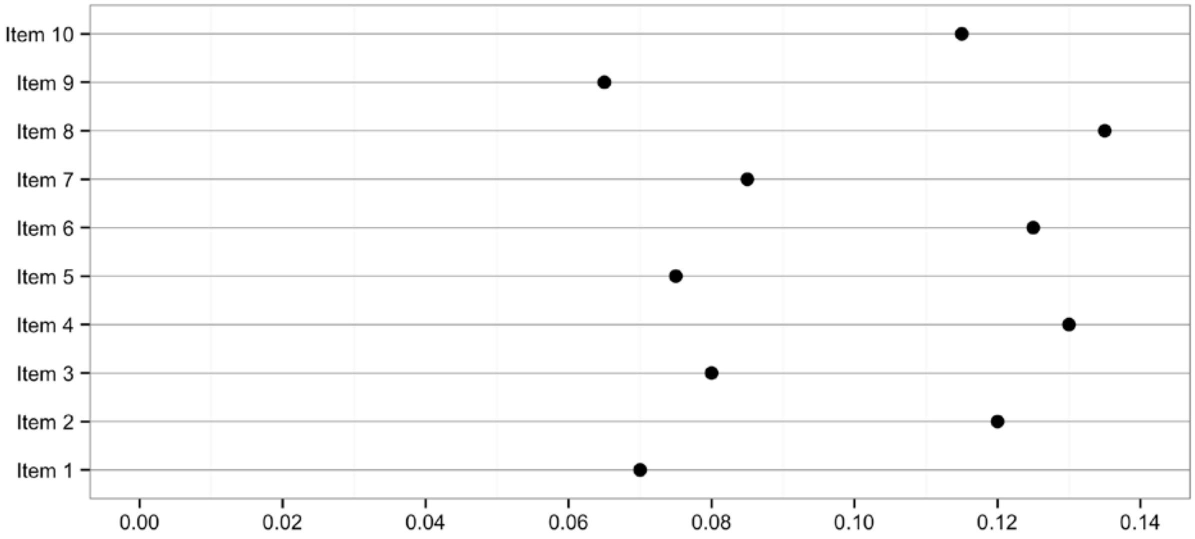
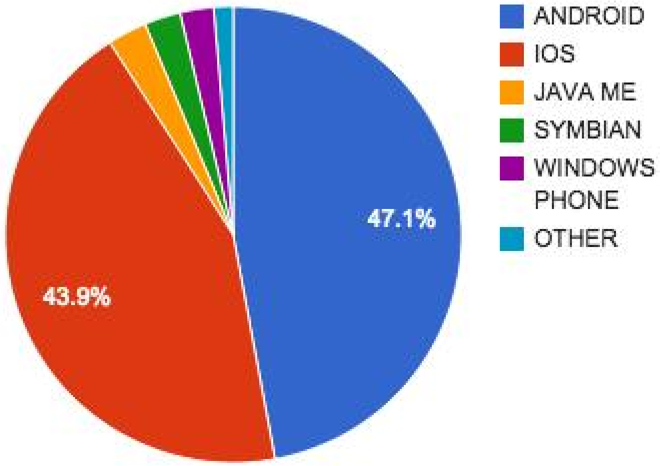


Fig 2.2 Structured Data Set: Dot Plot



Mobile OS Share (September 2014)



R graph Catalog by Jennifer Bryan http://shinyapps.stat.ubc.ca/r-graph-catalog

“A table is nearly always better than a dumb pie chart; the only worse design than a pie chart is several of them for then the viewer is asked to compare quantities located in spatial disarray both within and between pies . … Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.”

- Tufte, *The Visual Display of Quantitative Information*,1983, page 178.

“Pie charts have severe perceptual problems. Experiments in graphical perception have shown that compared with dot charts, they convey information far less reliably. But if you want to display some data , and perceiving the information is not so important, then a pie chart is fine.”

- Becker and Cleveland, *S-Plus Trellis Graphics User’s Manual.* 1996.

Common misunderstanding

Pie charts are bad! **Die pie chart, DIE**

Pie charts are bad when you want to accurately compare two numbers

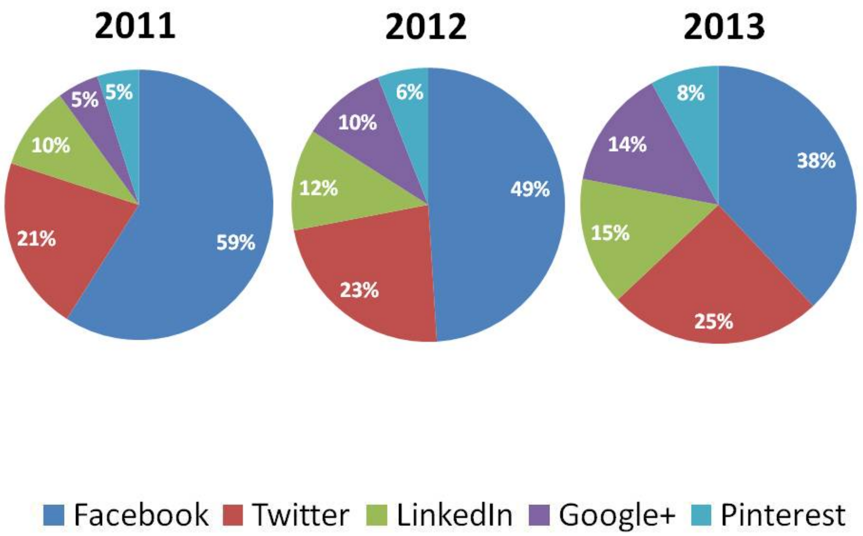
But:

As good as bars for estimating percentage of whole.
Better than bars for comparing compound proportions (A + B vs C + D)

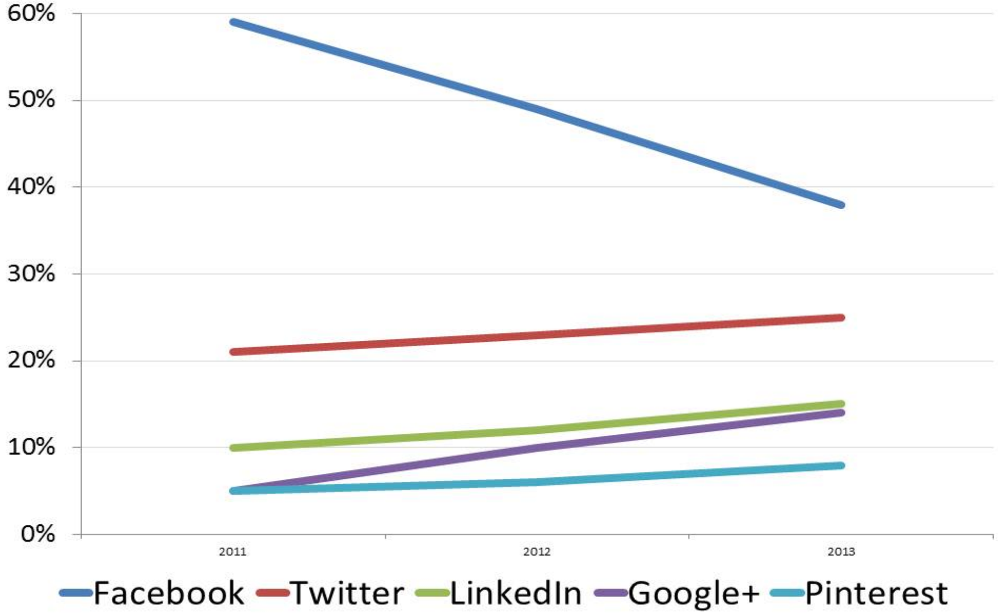
I. Spence. No Humble Pie: the Origins and Usage of a Statistical Chart. *Journal of Educational and Behavioral Statistics*, 30:353-368, 2005.

Hadley Wickham, Creating Effective Visualisations, slide presentation June 2012: http://courses.had.co.nz/12-effective-vis/

Favourite social media channel



Favourite social media channel



perceptual tasks – using length vs position along a common scale

Fig 8.12 Car Production: Stacked Bar Chart

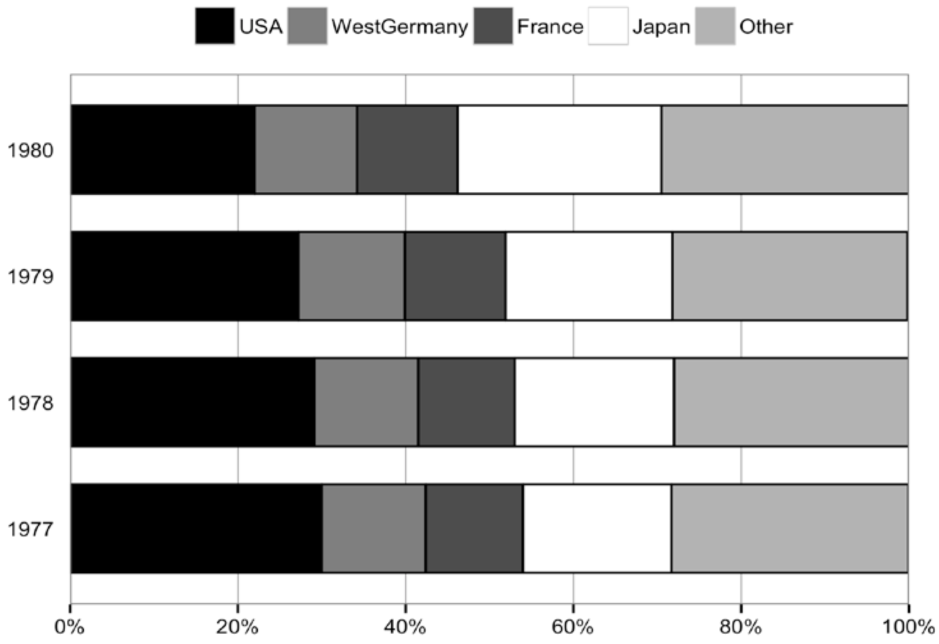
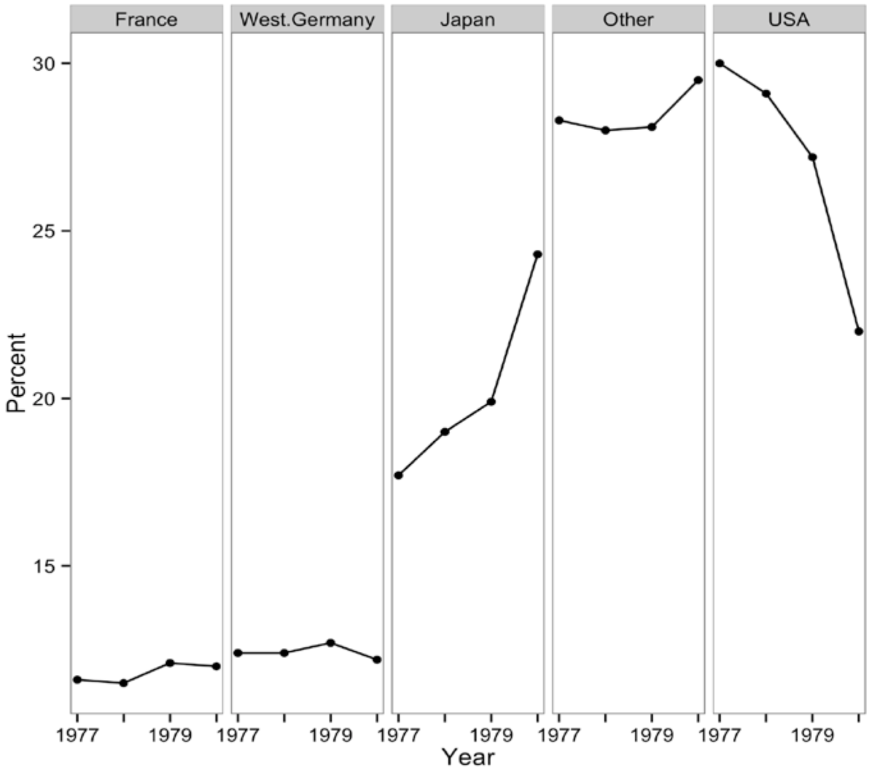


Fig 8.13 Car Production: Trellis Panels

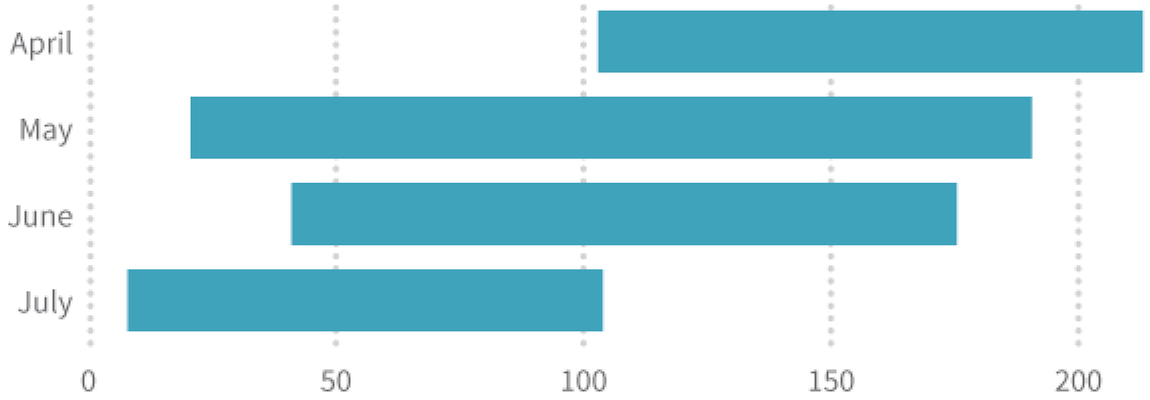


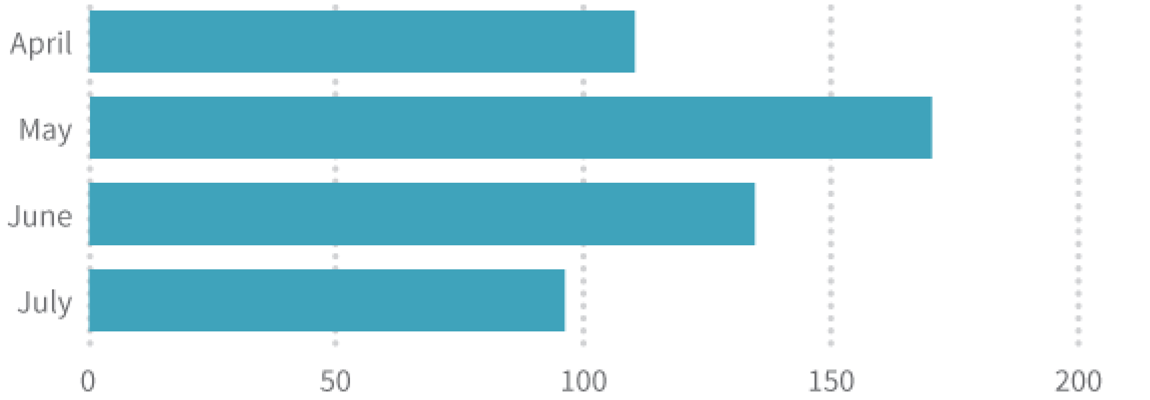
trends for categories on left and right are easy to see,
but trends for categories in middle are hard to judge

stacked bar charts: difficult to decode because they lack a common baseline for judging length

R graph Catalog by Jennifer Bryan http://shinyapps.stat.ubc.ca/r-graph-catalog

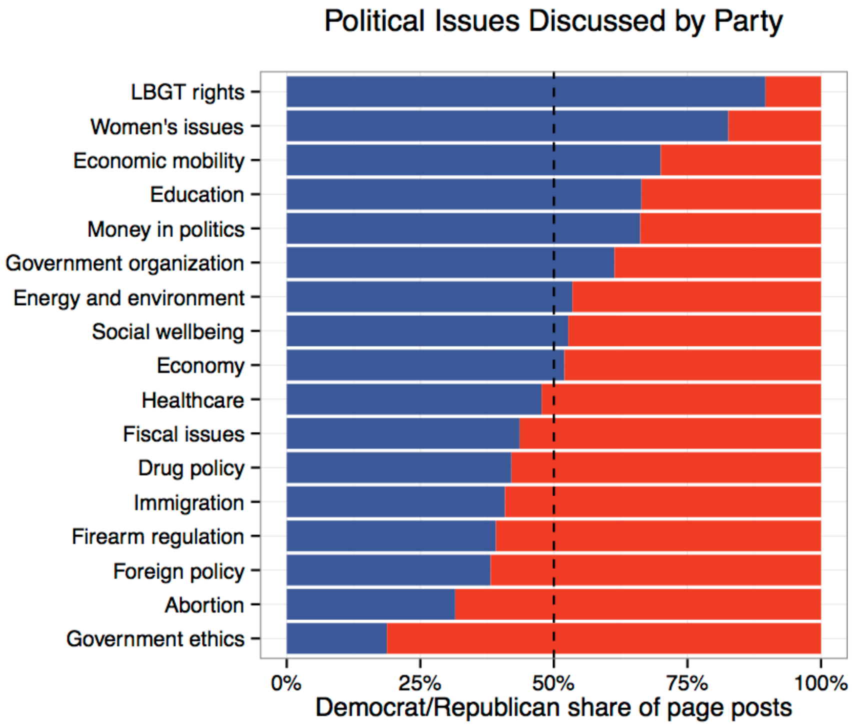
Humans are fairly good at comparing differences in length, but only when things share a common reference point. (Cleveland, William S. and Robert McGill. “Graphical Perception and Graphical Methods for Analyzing Scientific Data.” Science 229.4716 (1985): 828-833.

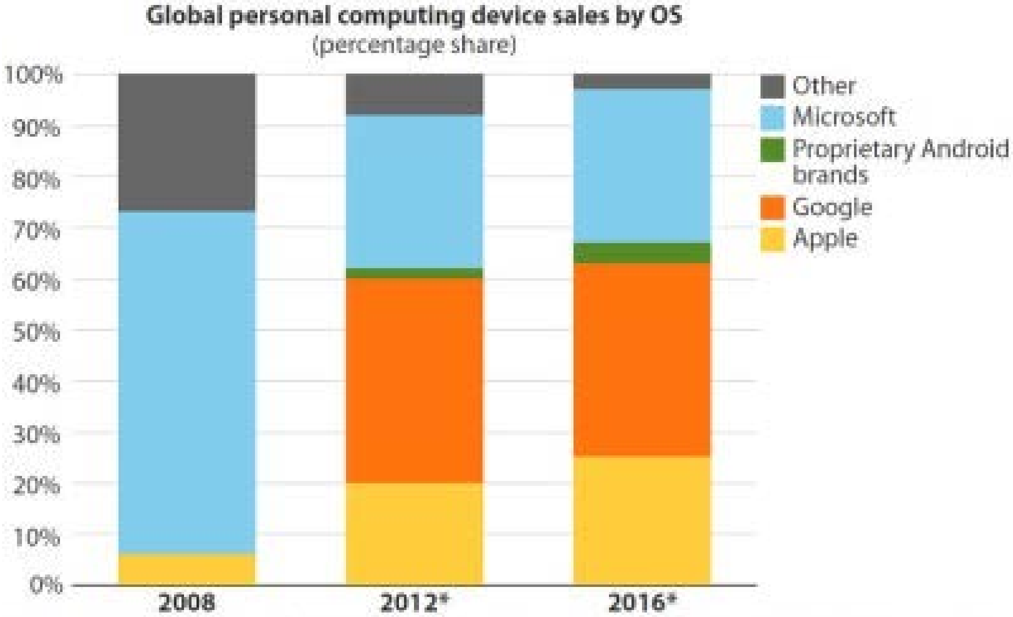




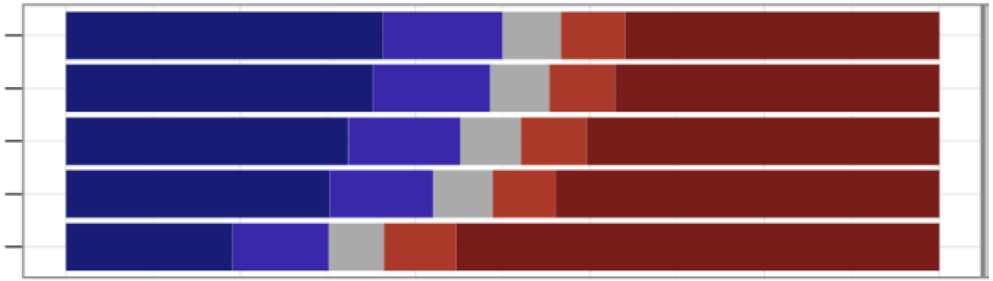
via Data + Design, https://infoactive.co/data-design/titlepage01.html

when to use, or not use, stacked bar charts? importance of a common baseline for comparisons.



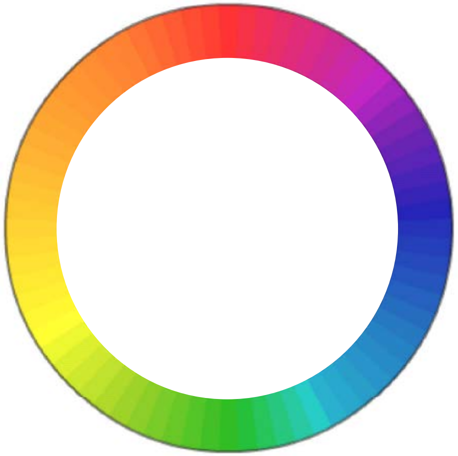


cludes annual estimates of global IT and consumer purchased devices in 62 leading countries
Research: company reports.

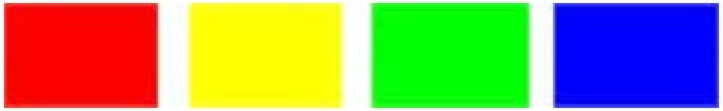


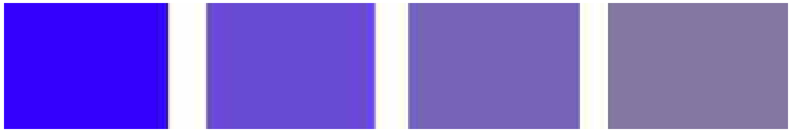
Solomon Messing Blog:
https://solomonmessing.wordpress.com/2014/10/11/when-to-use-stacked-barcharts/

color dimensions: hue, chroma, luminance (hcl)

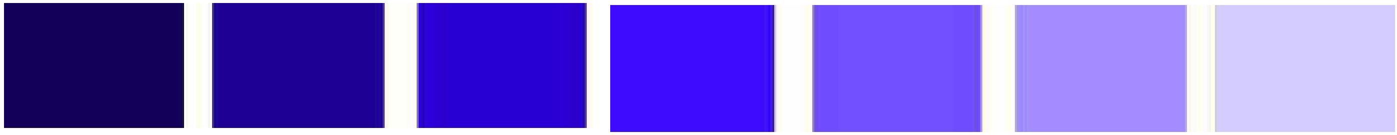


hue: unordered (position along color wheel)





chroma (purity): ordered
~ how much gray is added to pure color



luminance (lightness) : ordered
~ how much black or white is added to pure color

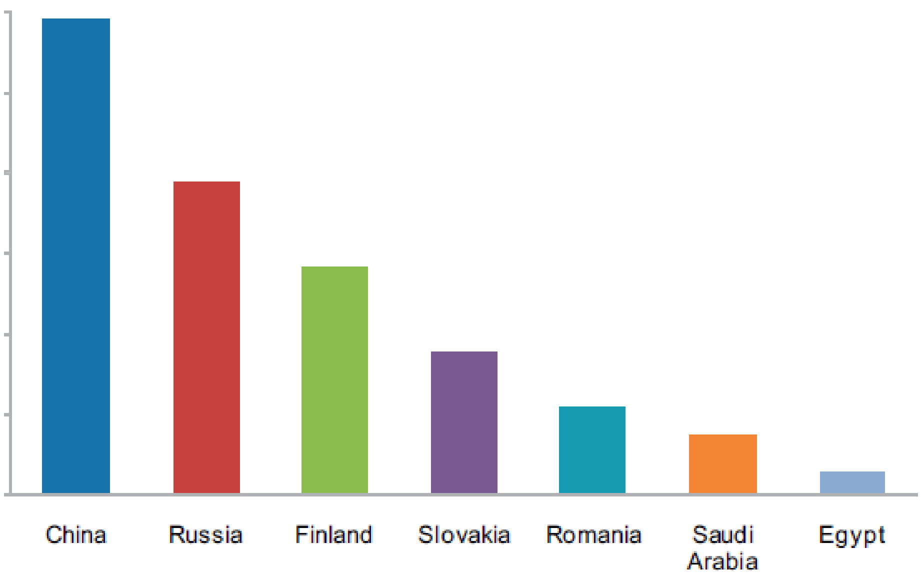
Maureen Stone, Choosing Colors for Data Visualization, 2006.

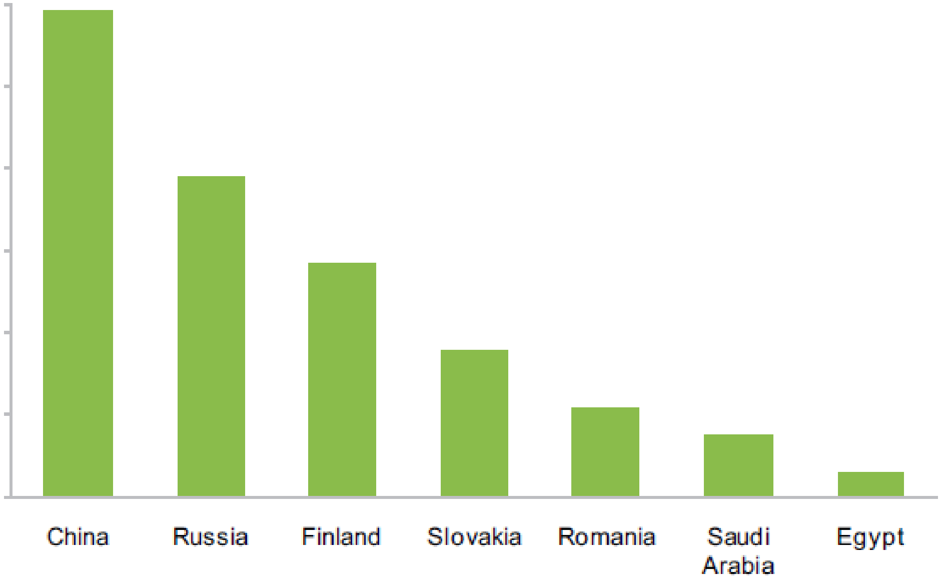
http://www.perceptualedge.com/articles/b-eye/choosing_colors.pdf

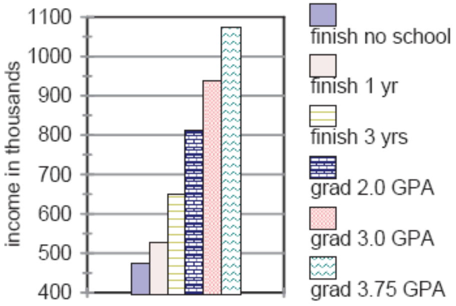
use color dimensions

- to distinguish groups
- to highlight particular data
- to encode quantitative values

only vary color for a reason







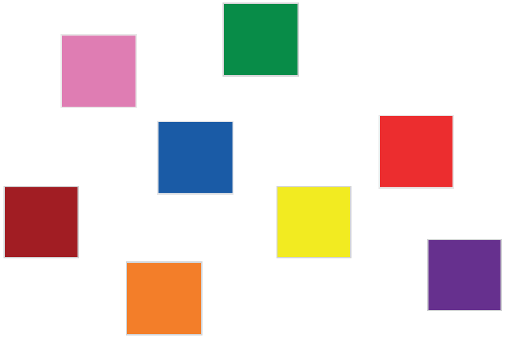
Source: Stinebrickner and Stinebrickner (2013).

varying color and pattern and length

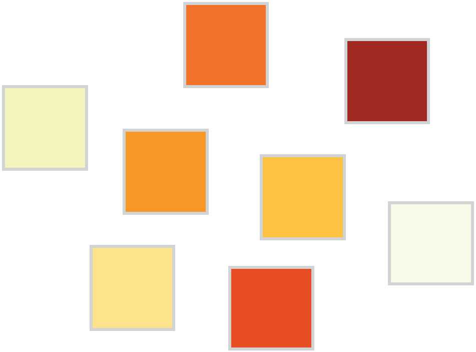
From Stephen Few, Perceptual Edge:
http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf

hue is unordered and not perceived quantitatively ... usually a poor choice for indicating magnitude

how to order these colors from smallest to largest??



How to order **these** colors from smallest to largest?



Two ways to encode quantitative data using color

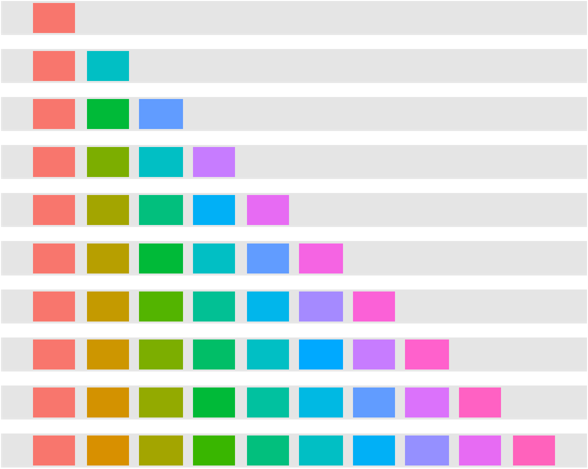
- sequential scale:
  single hue where color varies from light to dark
  or
  single hue where color varies from pale to pure

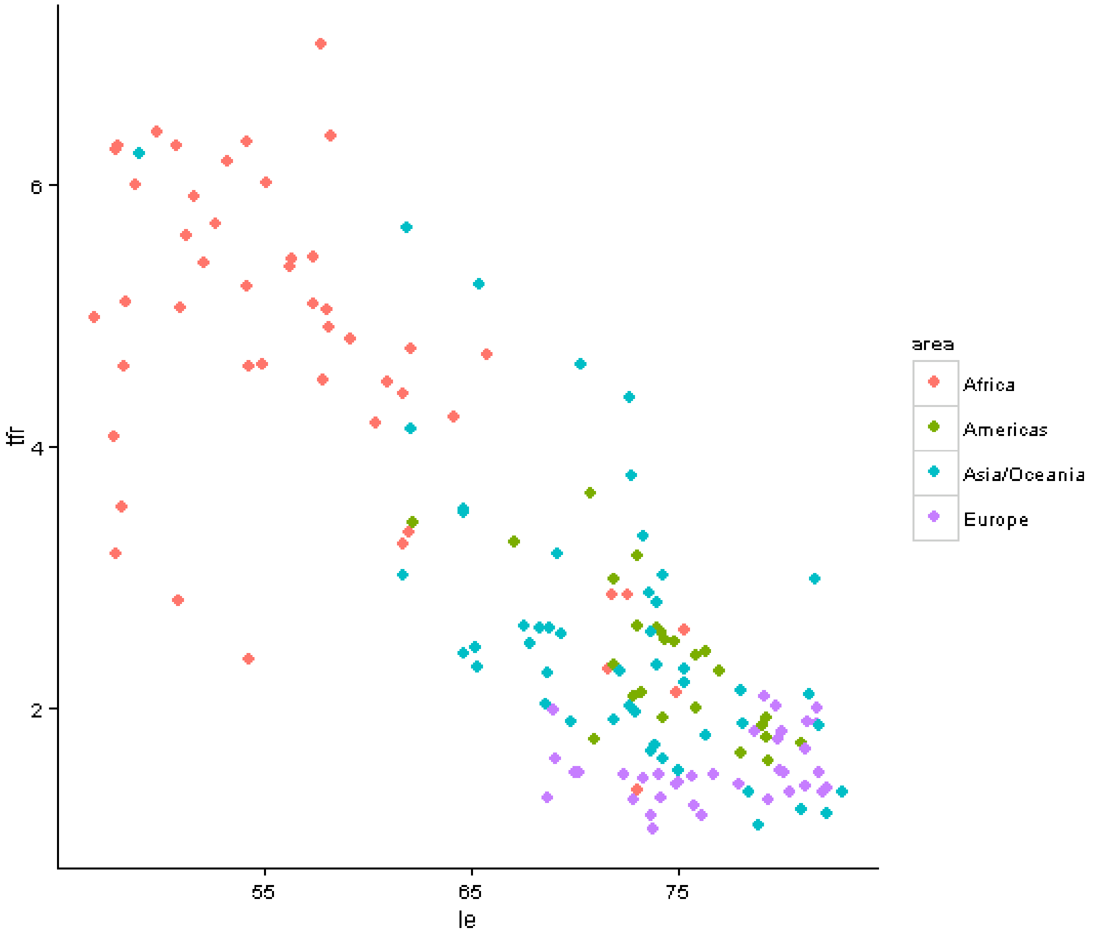
- diverging scale: two hues with a neutral color in between, where each hue varies from light to dark or pale to pure

From Stephen Few, Perceptual Edge:
http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf

use hue to distinguish groups

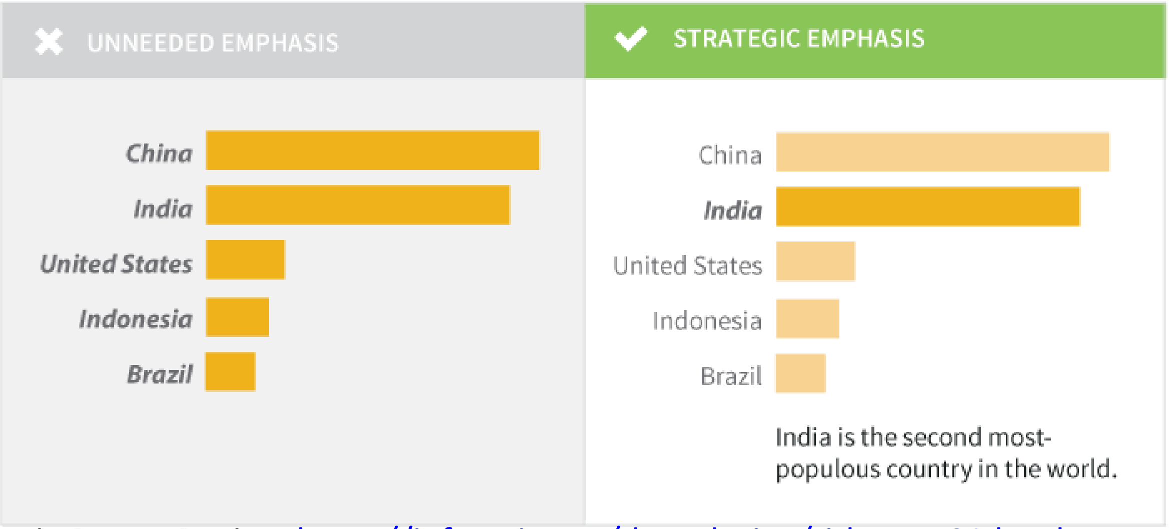
use equally spaced
hues along color wheel,
for example:





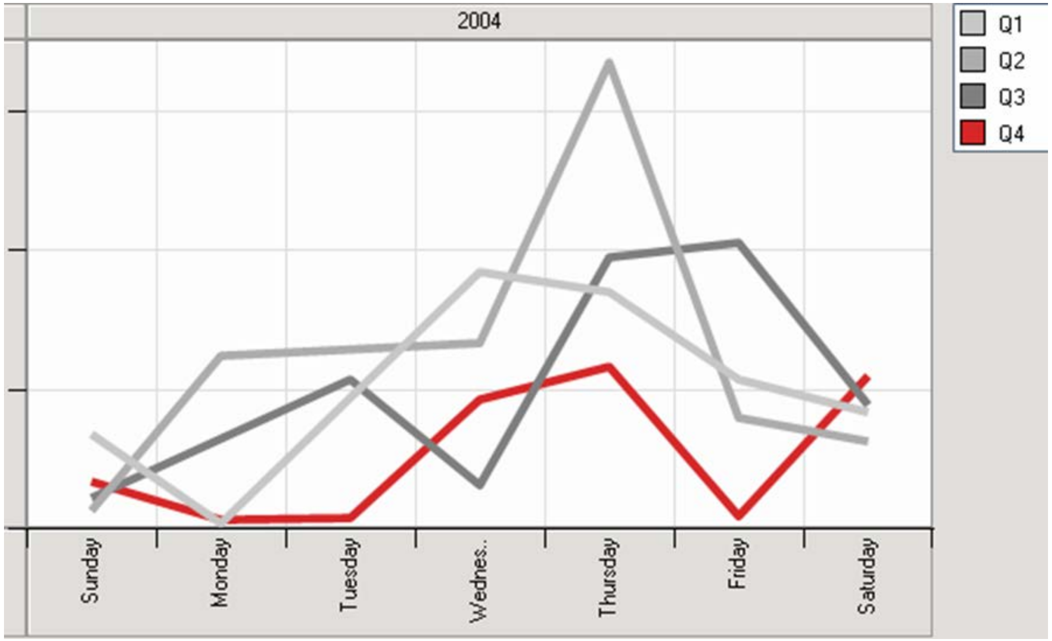
use color to highlight

use soft colors to display most information and bright and/or dark colors for emphasis



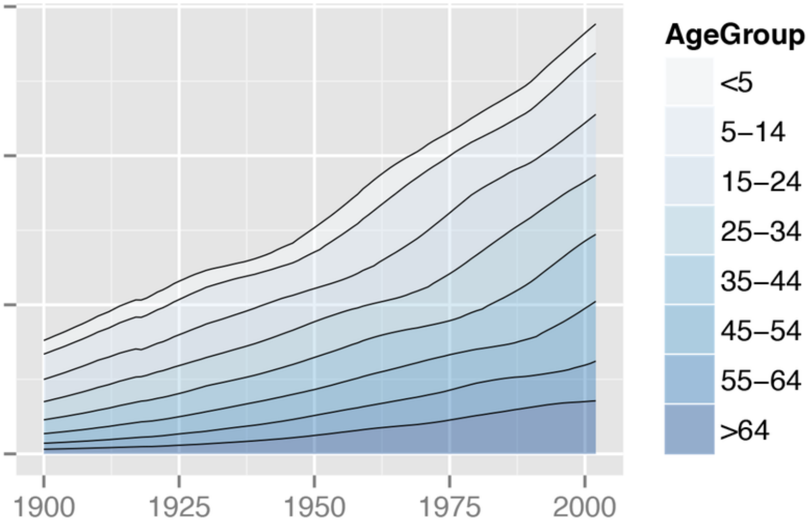
India is the second most-populous country in the world.

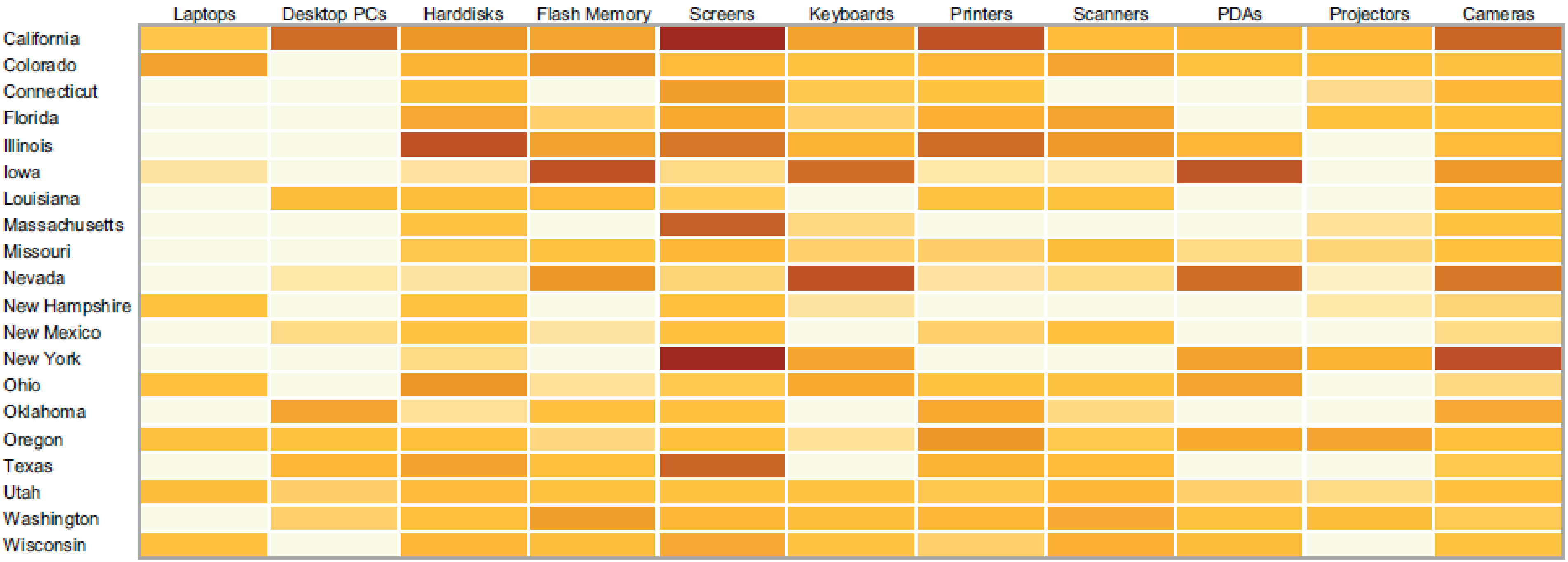
via Data + Design, https://infoactive.co/data-design/titlepage01.html



http://www.perceptualedge.com/articles/b-eye/choosing_colors.pdf use color to encode quantitative information

use a single hue
where color varies
from light to dark
or pale to pure

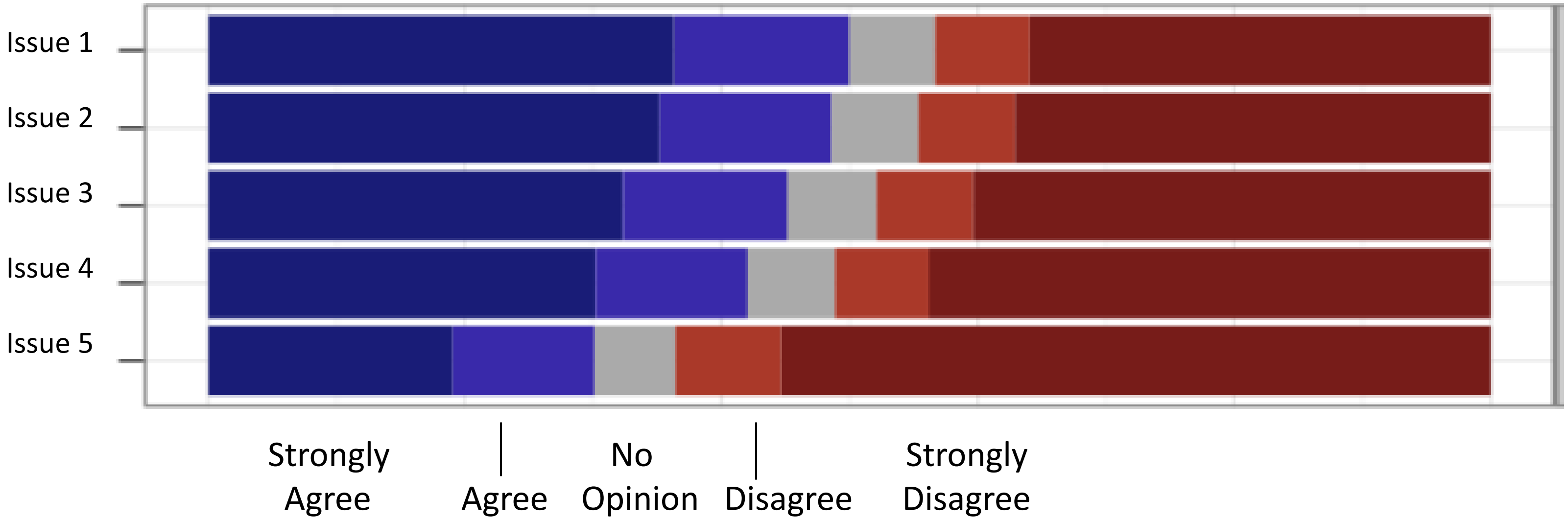




http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf

use color to encode quantitative information

use two hues with a neutral color in between,
where each hue varies from light to dark or pale to pure to create a diverging scale

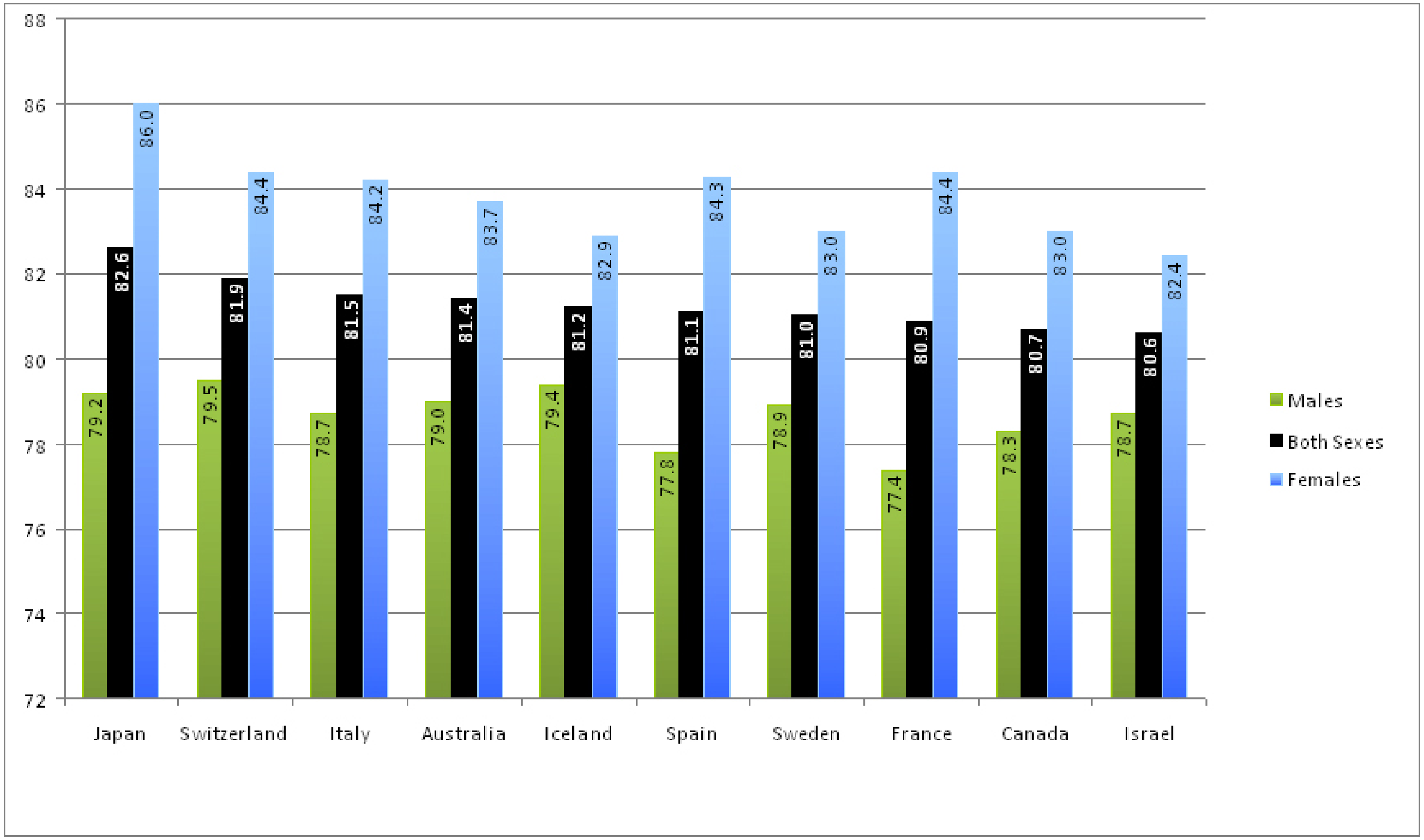


Based on Solomon Messing Blog: https://solomonmessing.wordpress.com/2014/10/11/when-to-use-stacked-barcharts/

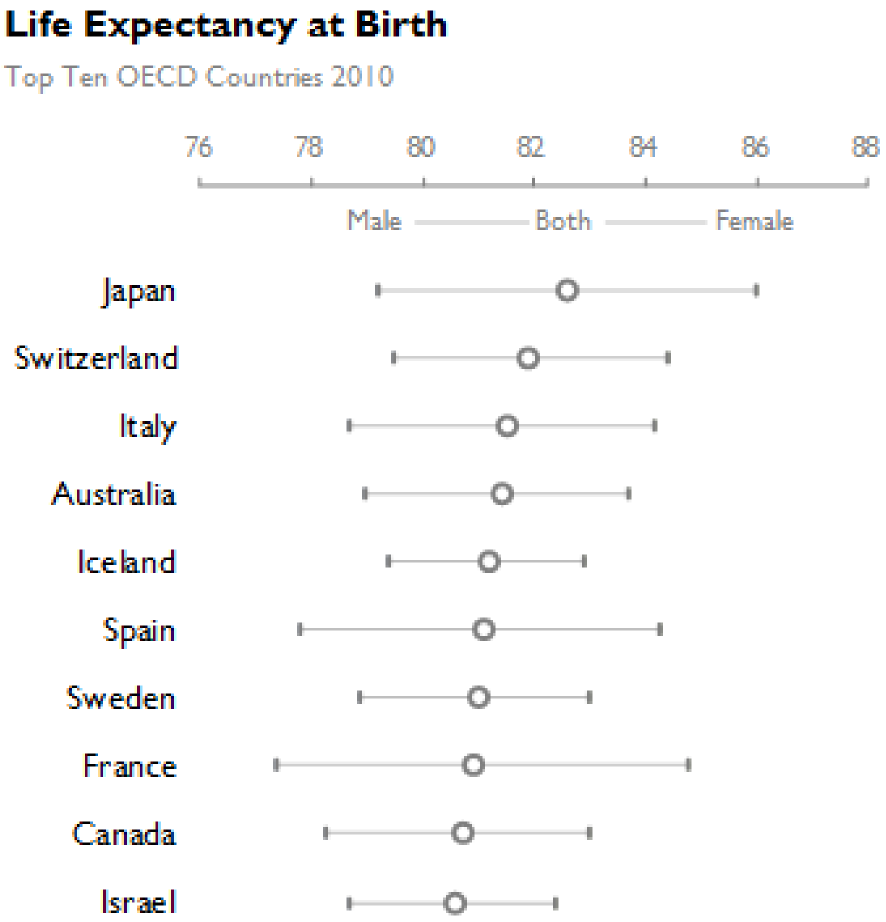
proximity - grouped bar charts are difficult because it’s hard to make comparisons between values that aren’t near each other ... **try to put values to be compared near each other**

- hard to compare data for each group (males, both, females) across countries,
  because other bars get in the way
- non-zero baseline (again)

Life expectancy at birth, top 10 OECD countries



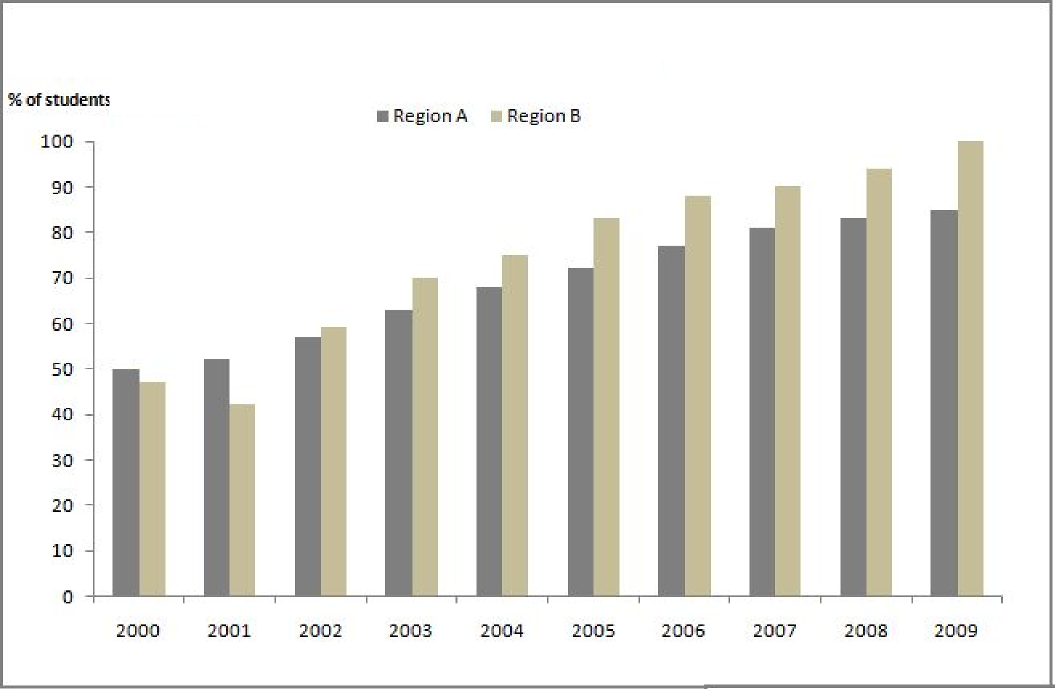
Darkhorse Analytics Blog: http://darkhorseanalytics.com/blog/too-many-bars/



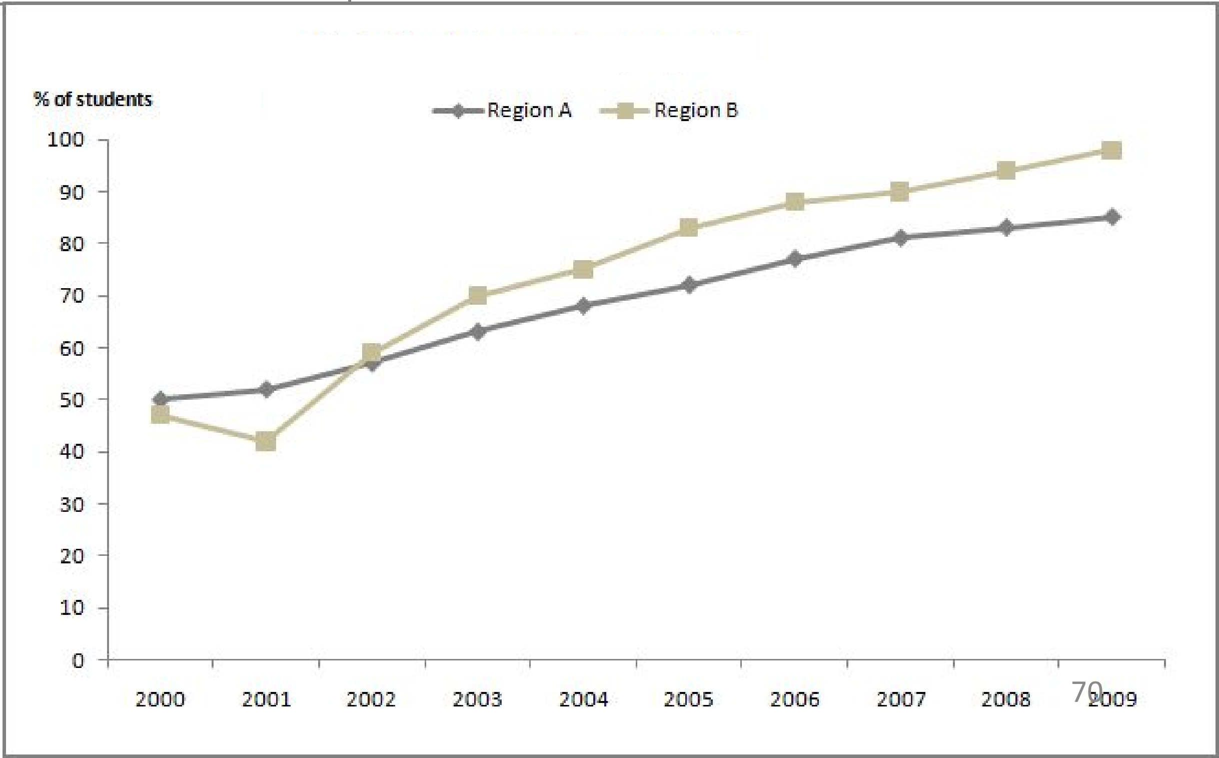
Life Expectancy at Birth

Top Ten OECD Countries 2010

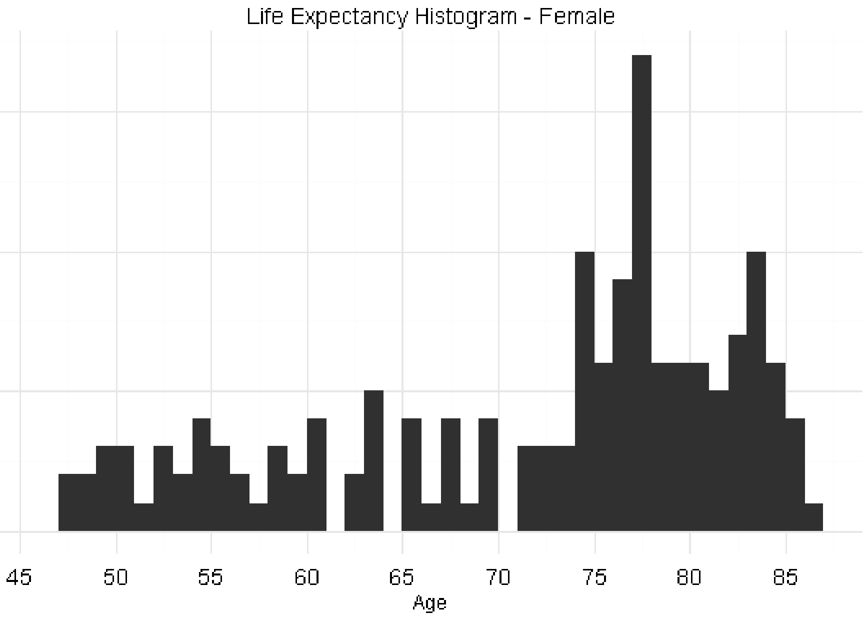
Darkhorse Analytics Blog: http://darkhorseanalytics.com/blog/too-many-bars/

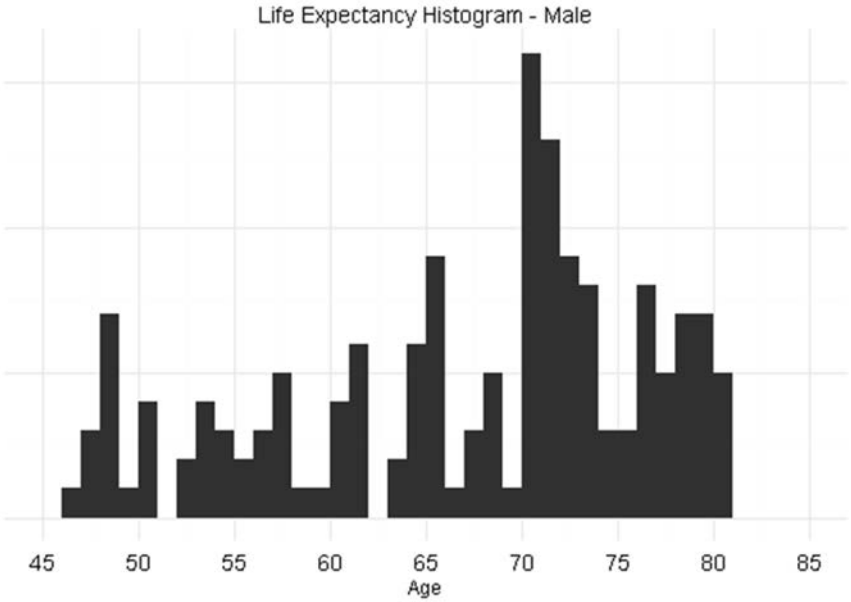


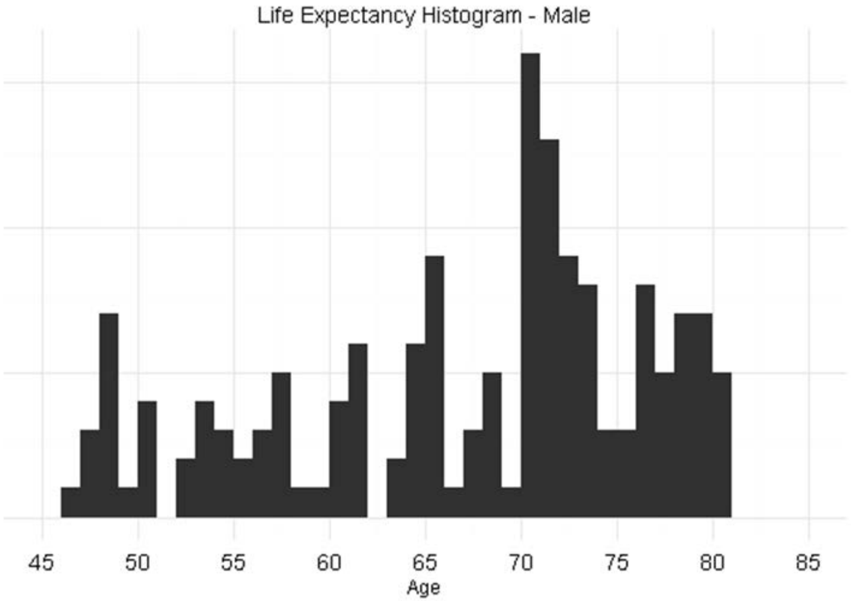
Grouped Bar Graph vs Line Graph



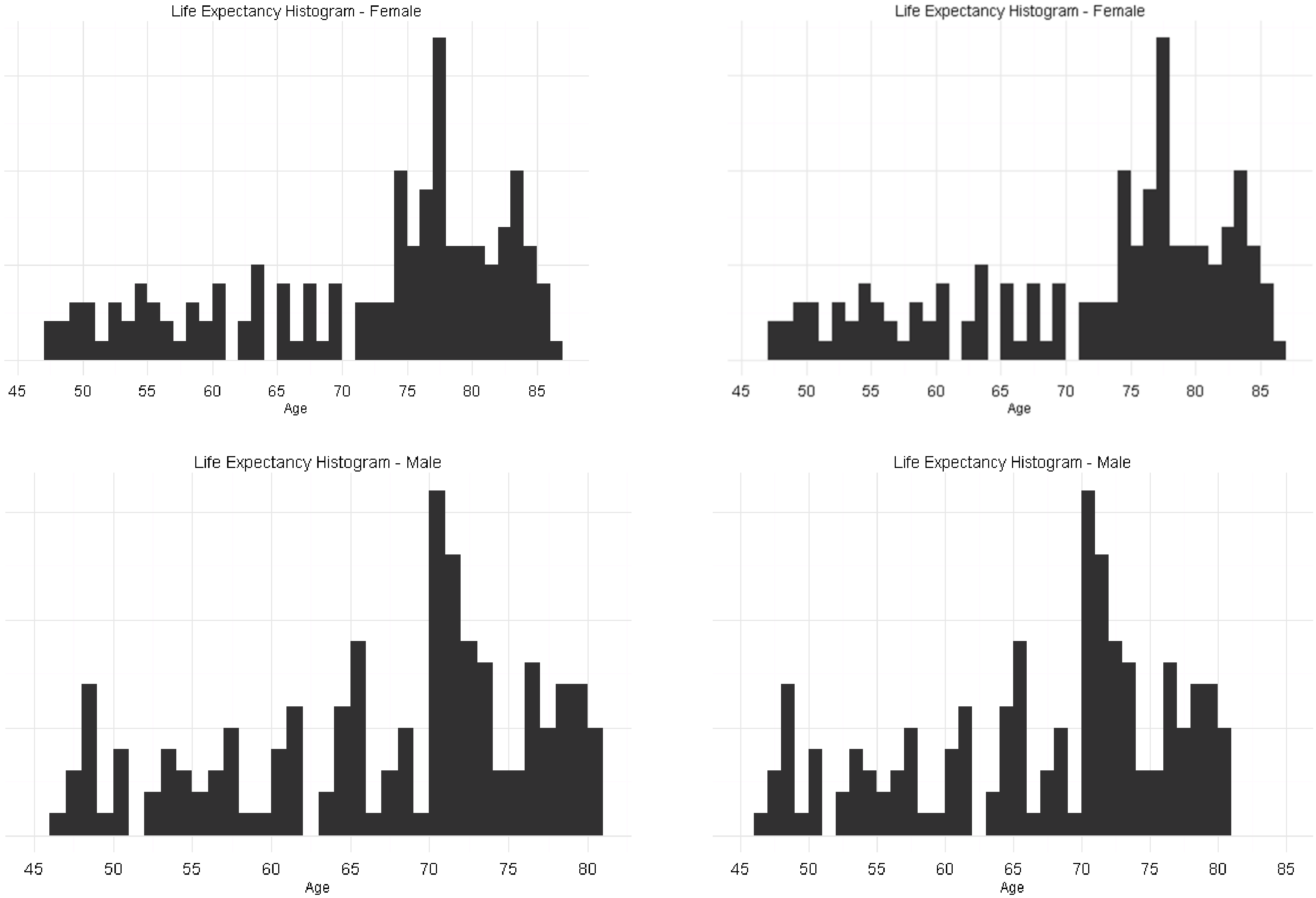
ease comparisons - align things vertically







ease comparisons – use common axes



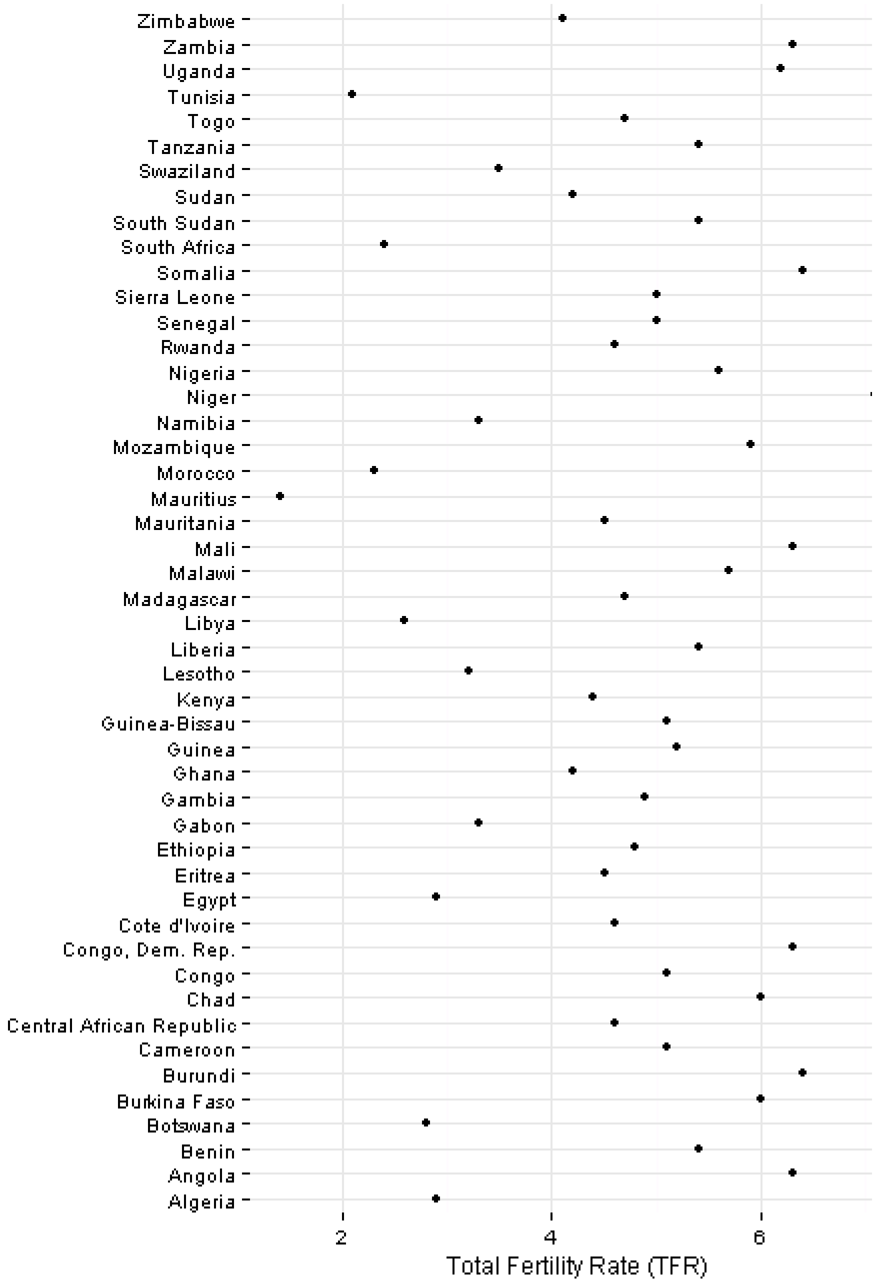
Life Expectancy Histogram - Female

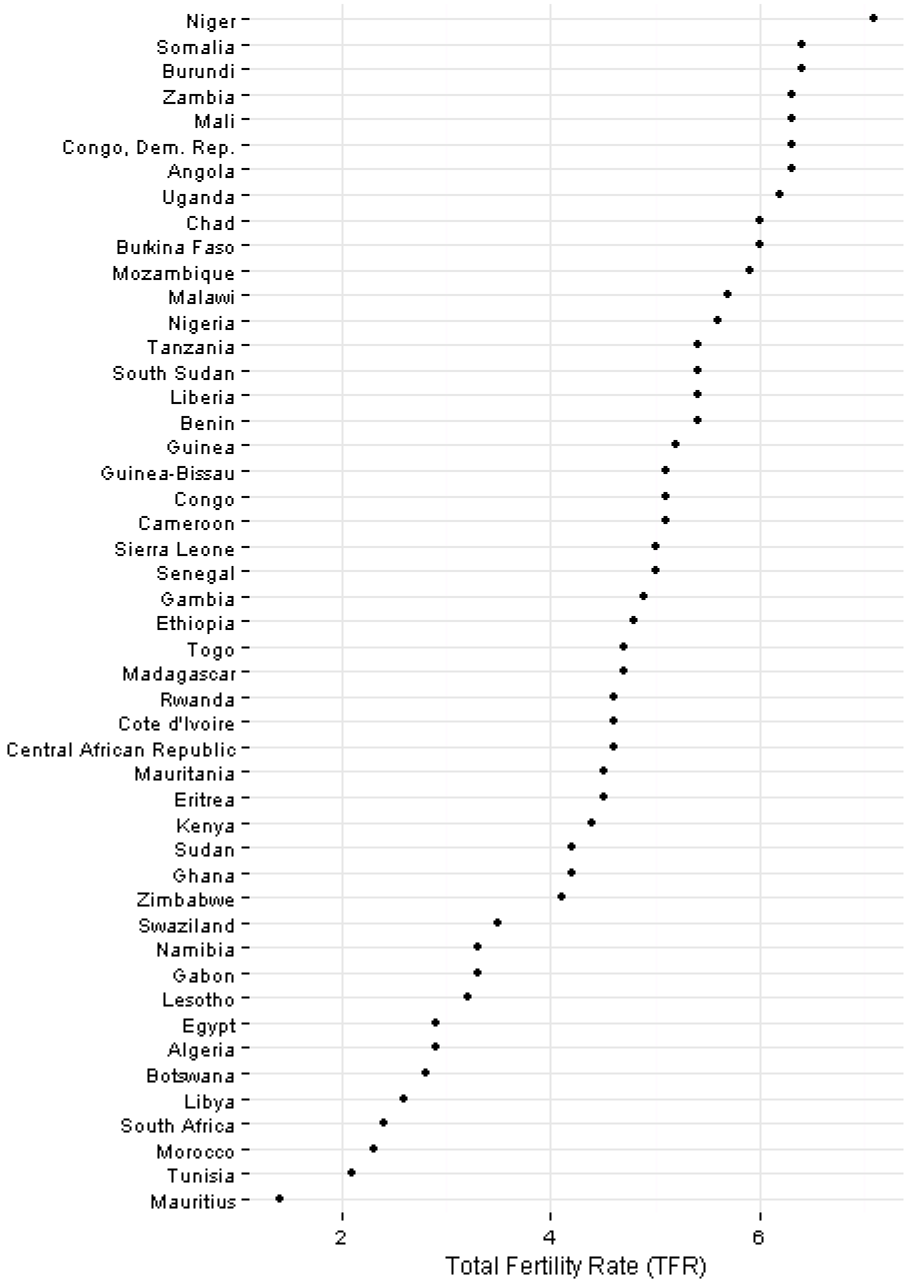
Life Expectancy Histogram - Female

Life Expectancy Histogram - Male

Life Expectancy Histogram - Male

ordering – don’t sort alphabetically





relationships among subsets

consider having two continuous variables, plus a third categorical variable.
how to compare the relationships both within and between categories ?

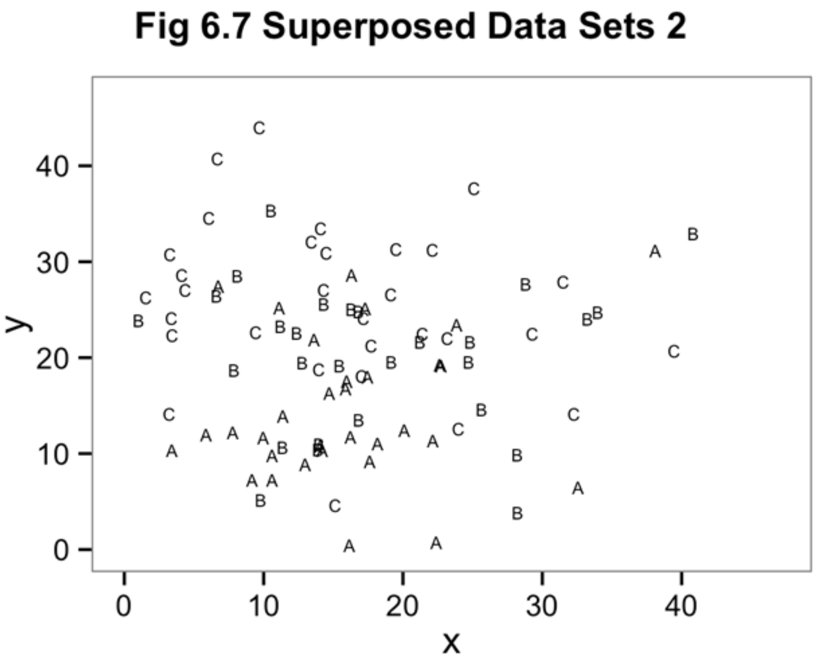
small multiples: "a series of graphics, showing the same combination of variables, indexed by changes in another variable."

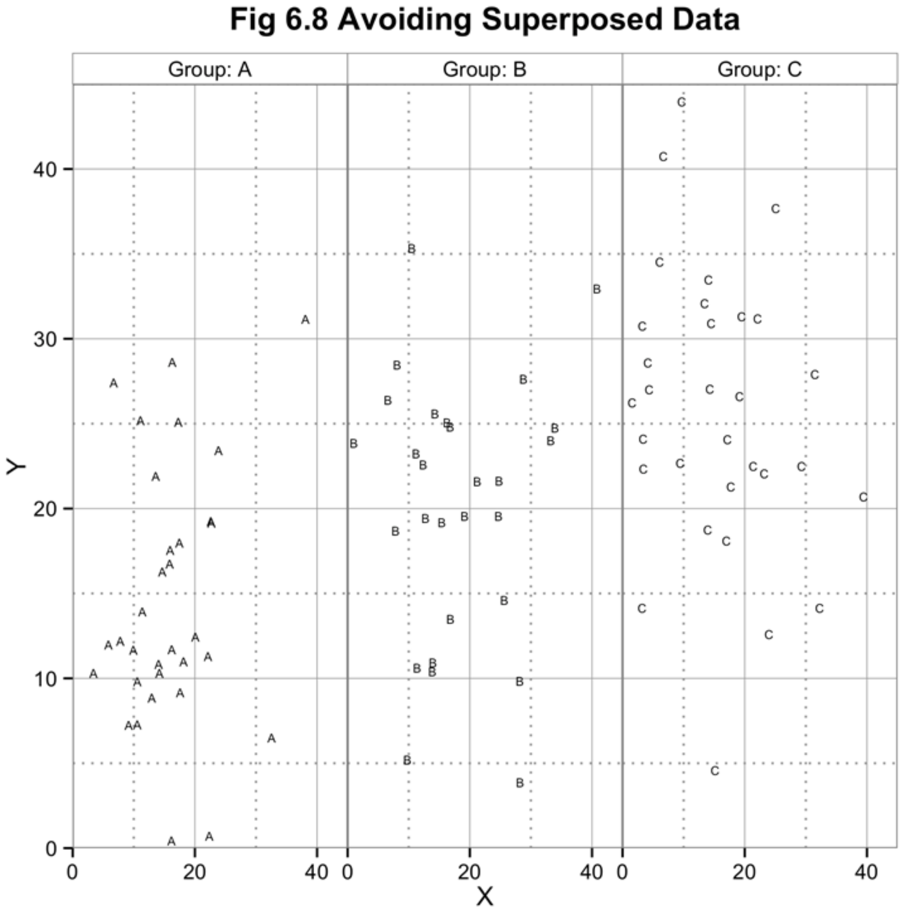
- Edward Tufte

Tufte, E. Envisioning Information. Graphics Press. Cheshire, CT. 1990.

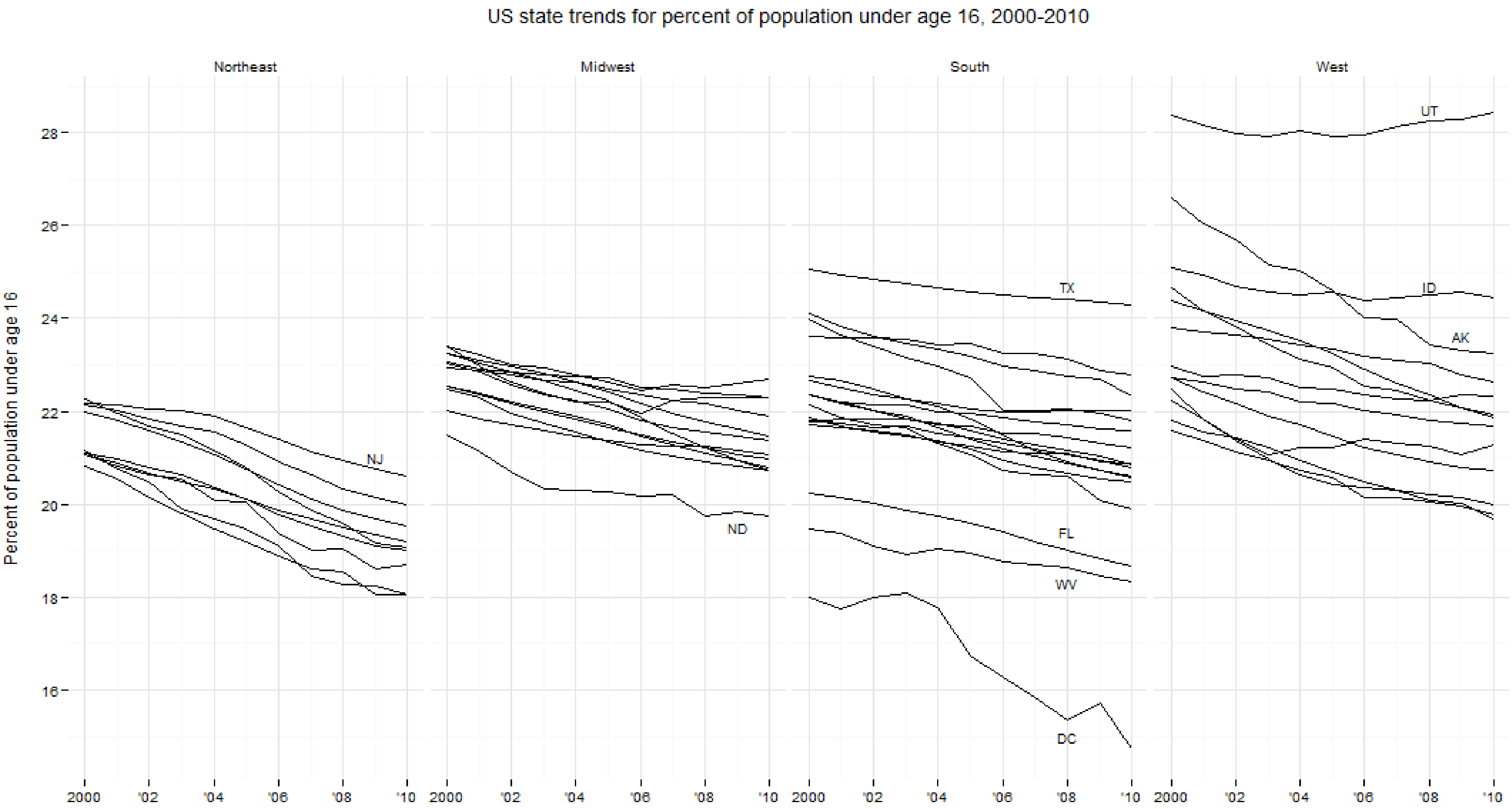
a graph showing **small multiples** (sometimes called a *trellis*, *lattice*, *grid*, *panel or facet graph*) shows a series of small graphs displaying the same relationships for different subsets of data.

"Small multiple designs, multivariate and data bountiful, answer directly by visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives.
For a wide range of problems in data presentation, small multiples are the best design solution."
- Edward Tufte, Envisioning Information, p. 67.





R graph Catalog by Jennifer Bryan: http://shinyapps.stat.ubc.ca/r-graph-catalog



Source: 2000-2010 State Characteristics Intercensal Population Estimates File, US Census Bureau, Population Division.

Using the same axis scale and keeping proximity, alignment and ordering in mind, allows for easy comparisons both within and between the small multiples.

small multiples

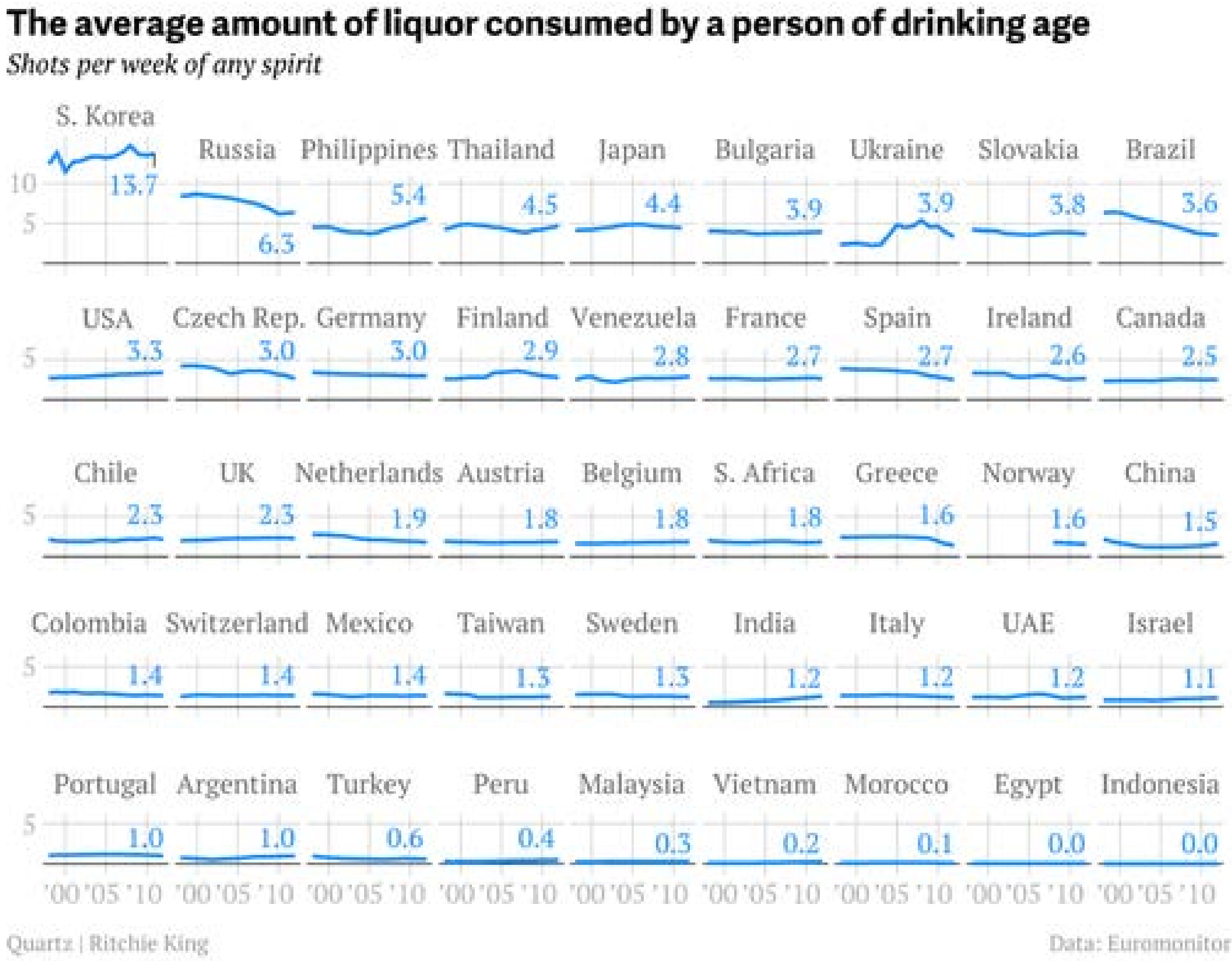
- countries ordered by most recent data point rather than alphabetically

- scale labels on outer edges only, rather than one set per panel

- only used three labels for the 11 years on the plot

- did not overdo the vertical scale either

-extra large scale used for top row. This doesn't follow principle of small multiples but draws attention to the top left corner.



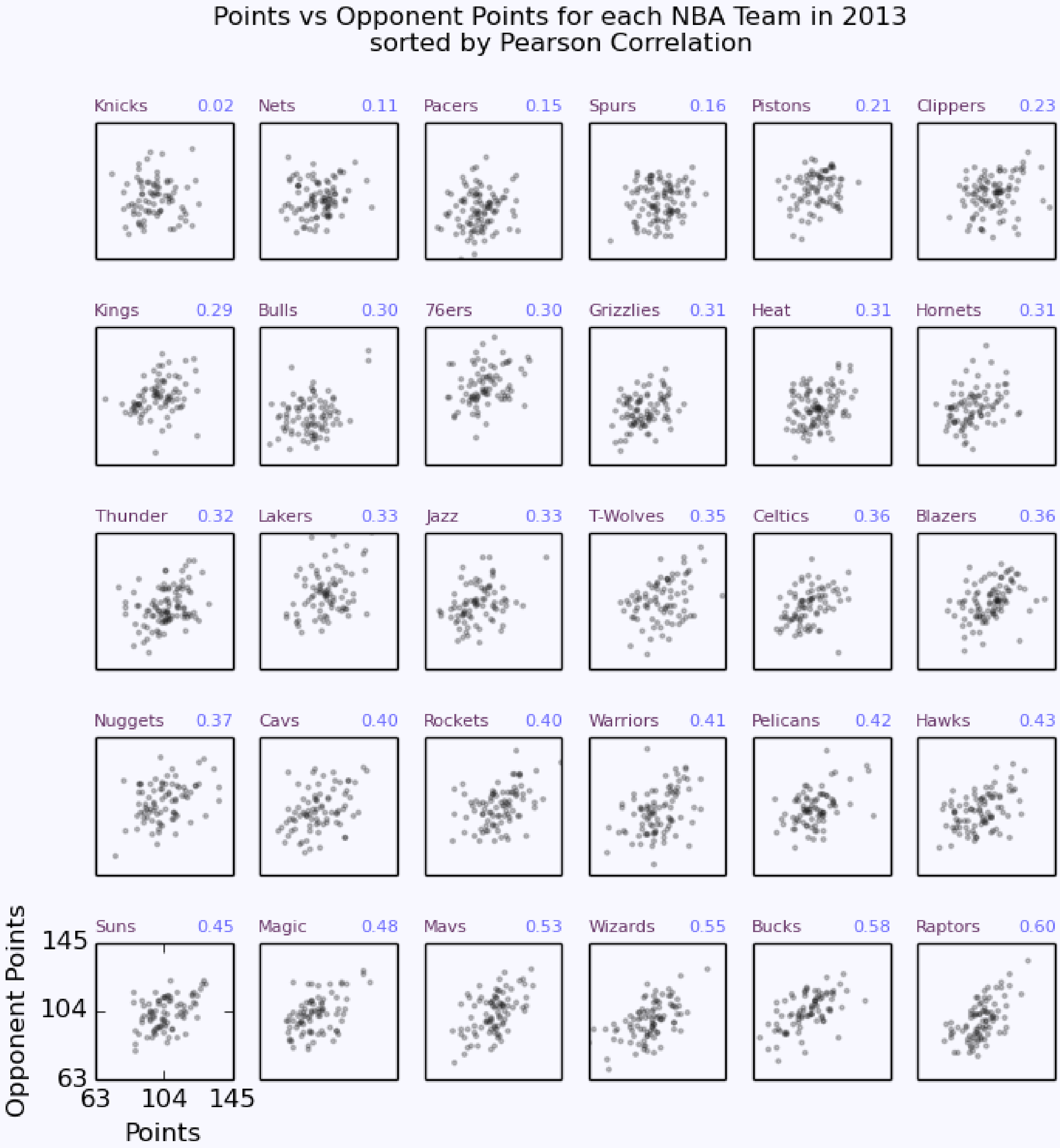
The average amount of liquor consumed by a person of drinking age

Shots per week of any spirit

Quartz | Ritchie King

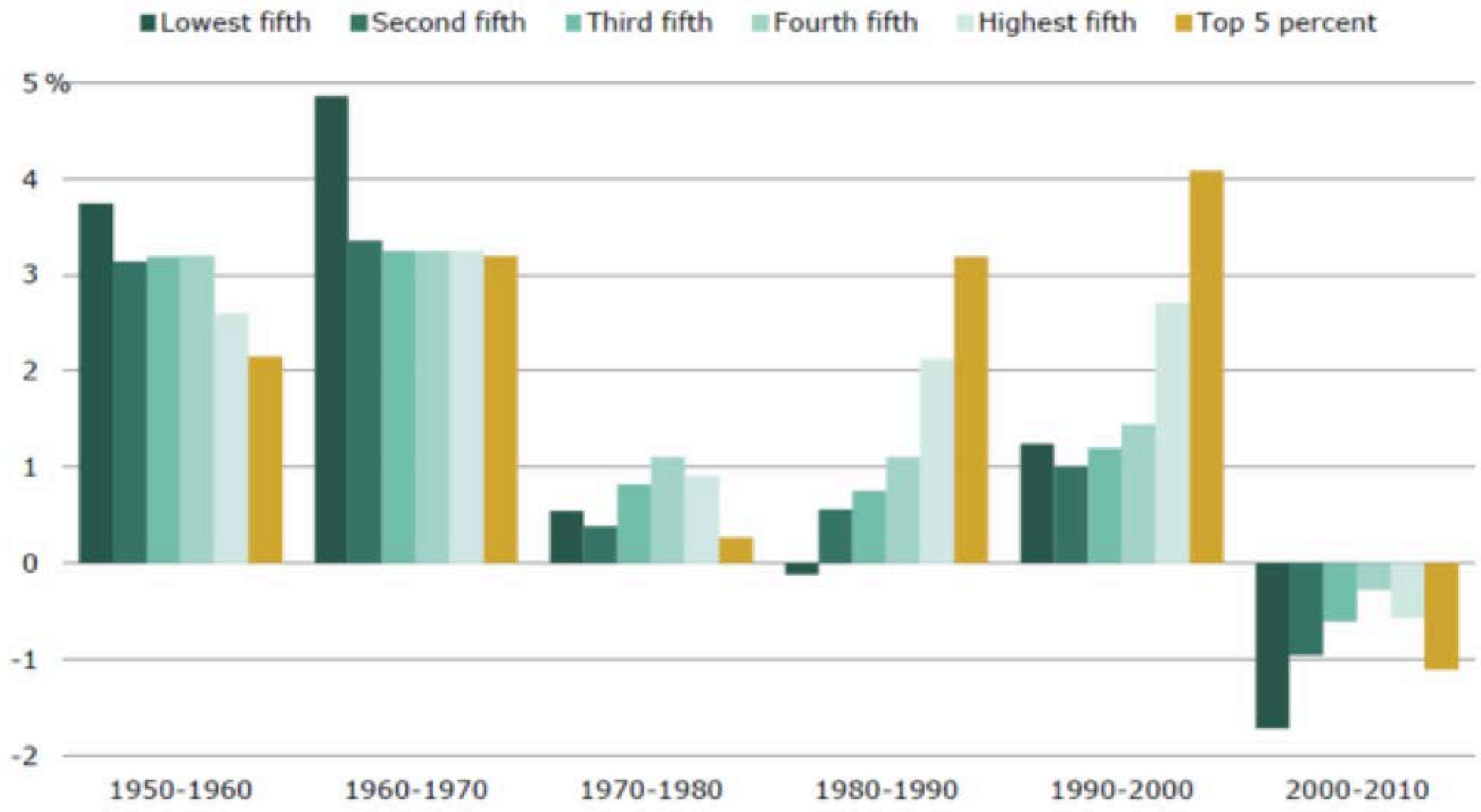
Data: Euromonitor

Junk Charts Blog: http://junkcharts.typepad.com/junk_charts/2014/02/small-multiples-with-simple-axes.html



http://viz.sdql.com/blog/2014/11/11/nba-points-vs-opp-points-for-each-team-in-2013

Average Annual Change in Mean Family Income, 1950-2010, by Quintile and for the Top 5 Percent

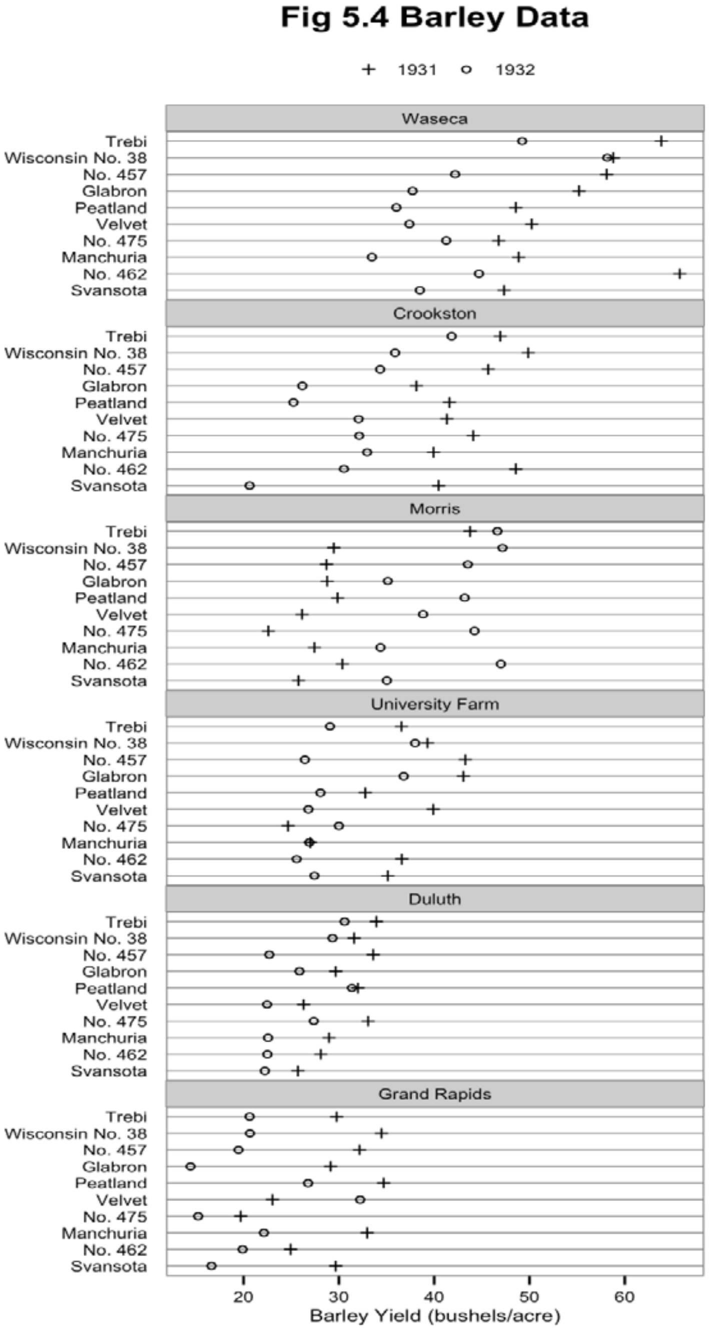


Source: U.S. Census Bureau, Historical Income Tables, Table F-3 for 1966 to 2010, and derived from Tables F-2 and F-7 for 1950 to 1965. Downloaded from http://www.census.gov/hhes/www/income/data/historical/families/ on July 11, 2012

PEW RESEARCH CENTER

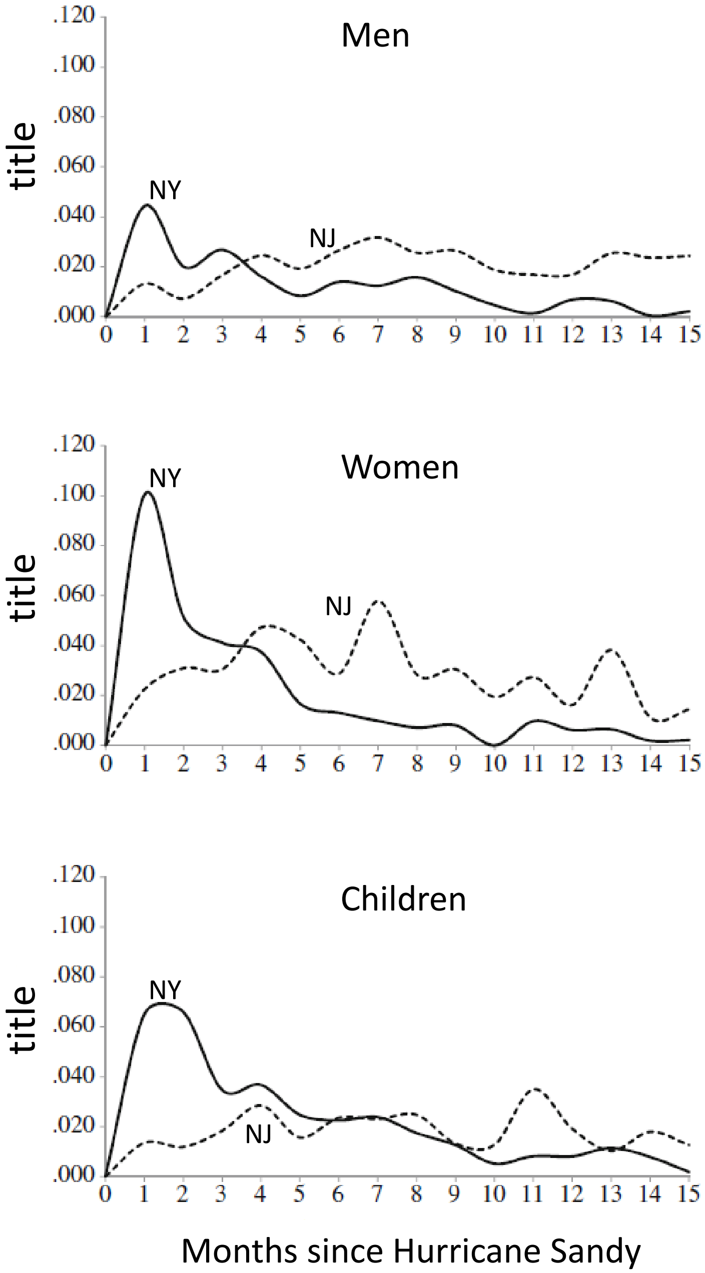
Pew Research Center, Social & Demographic Trends. Fewer, Poorer, Gloomier – The Lost Decade of the Middle Class. August 2012: http://www.pewsocialtrends.org/files/2012/08/pew-social-trends-lost-decade-of-the-middle-class.pdf

classic example:
Cleveland's use of small multiples allowed easy detection of data entry error

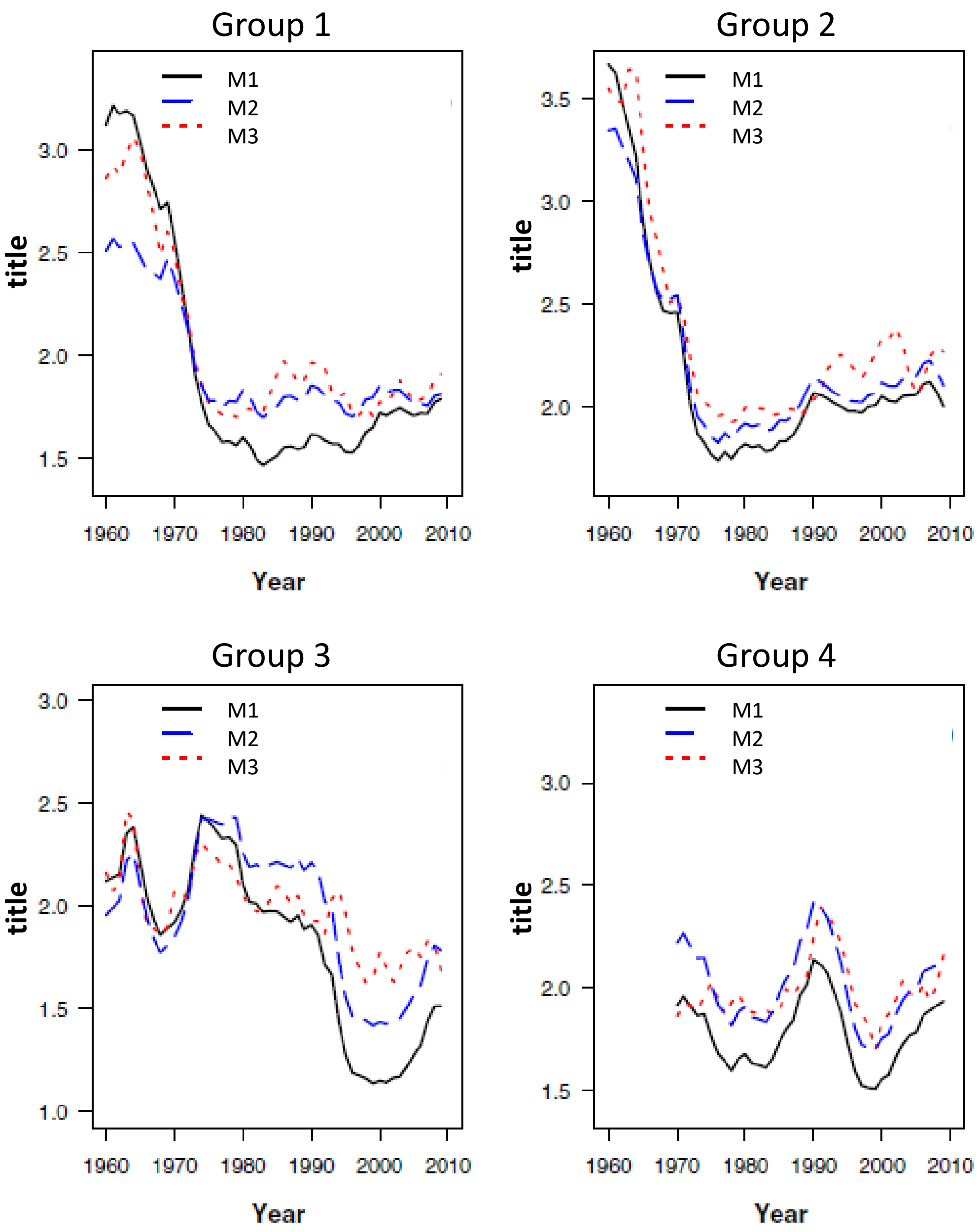


R graph Catalog by Jennifer Bryan: http://shinyapps.stat.ubc.ca/r-graph-catalog

4 variables shown –
month number (x)
measure (y)
state
group



use **common scales** to facilitate comparison across small multiples



IV Simplicity and **clarity**

"Chart junk can turn bores into disasters, but it can never rescue a thin data set."

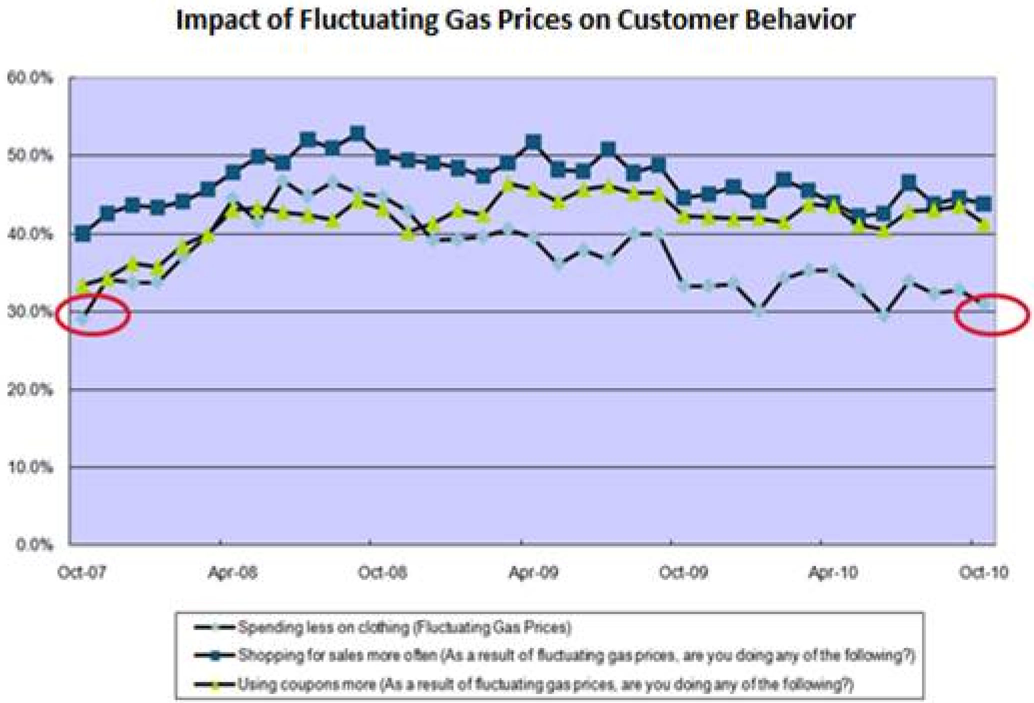
- Edward Tufte

"Non-data components of tables and graphs should be displayed just visibly enough to perform their role, but no more so, for excessive salience could cause them to distract attention from the data."

- Stephen Few

Tufte, E. The Visual Display of Quantitative Information, Second Ed. Graphics Press, Cheshire CT. 2001.

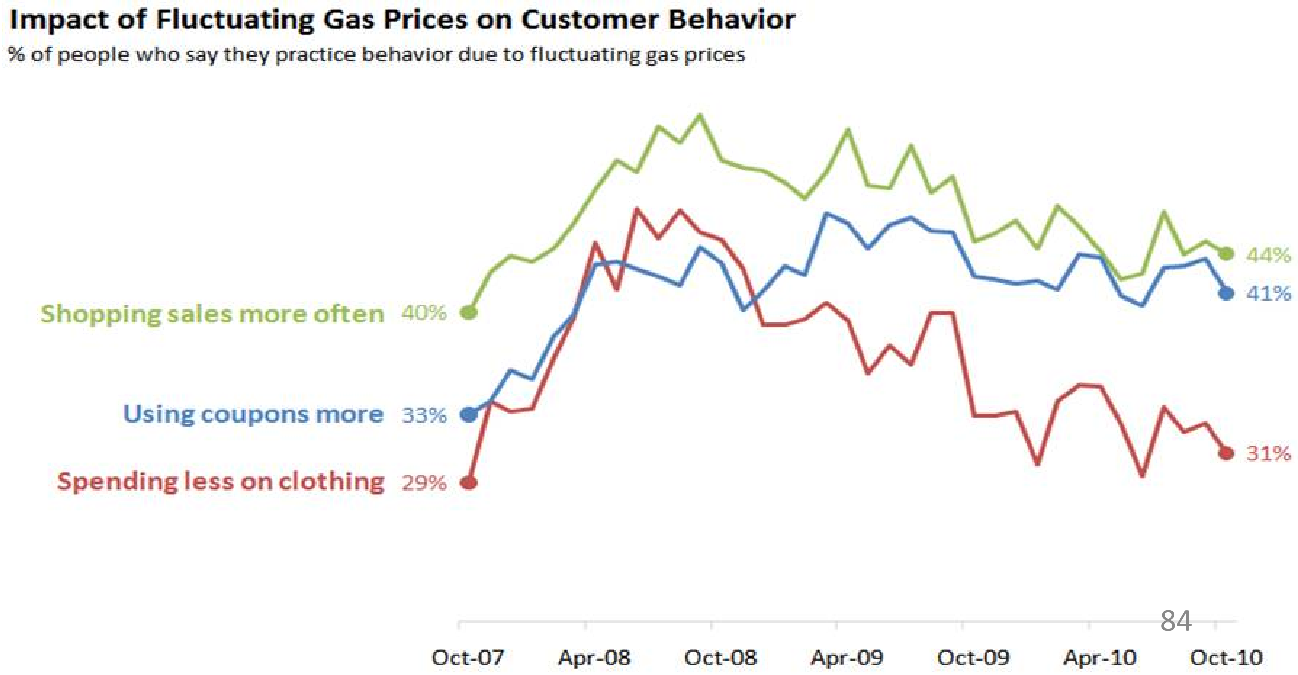
Few, S. http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf



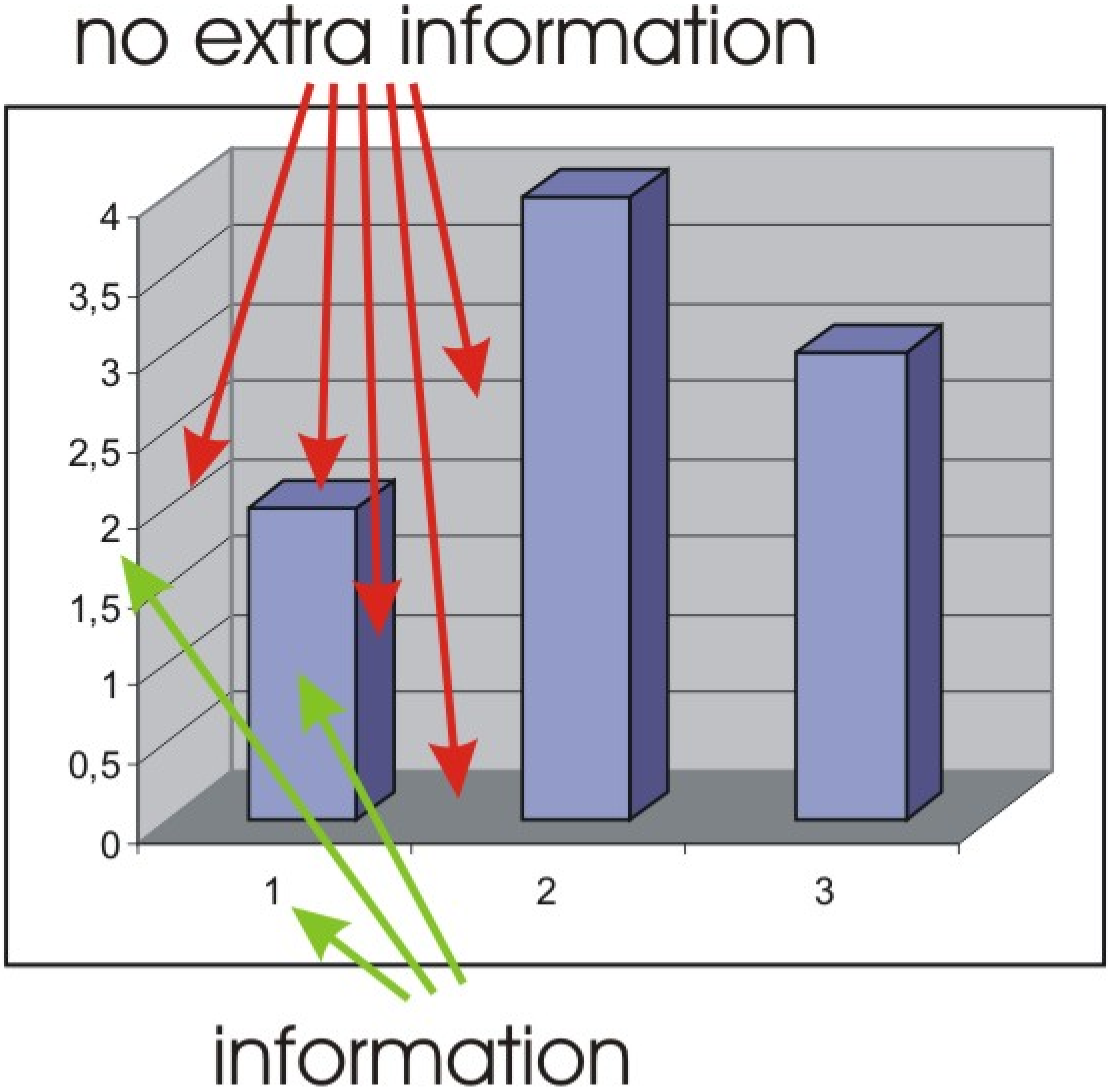
to simplify and increase clarity, consider:

eliminating legend and labeling directly
eliminating colored shapes on lines, but
keeping endpoints that are key to the message
eliminating distractions

- background shading
- gridlines
- border



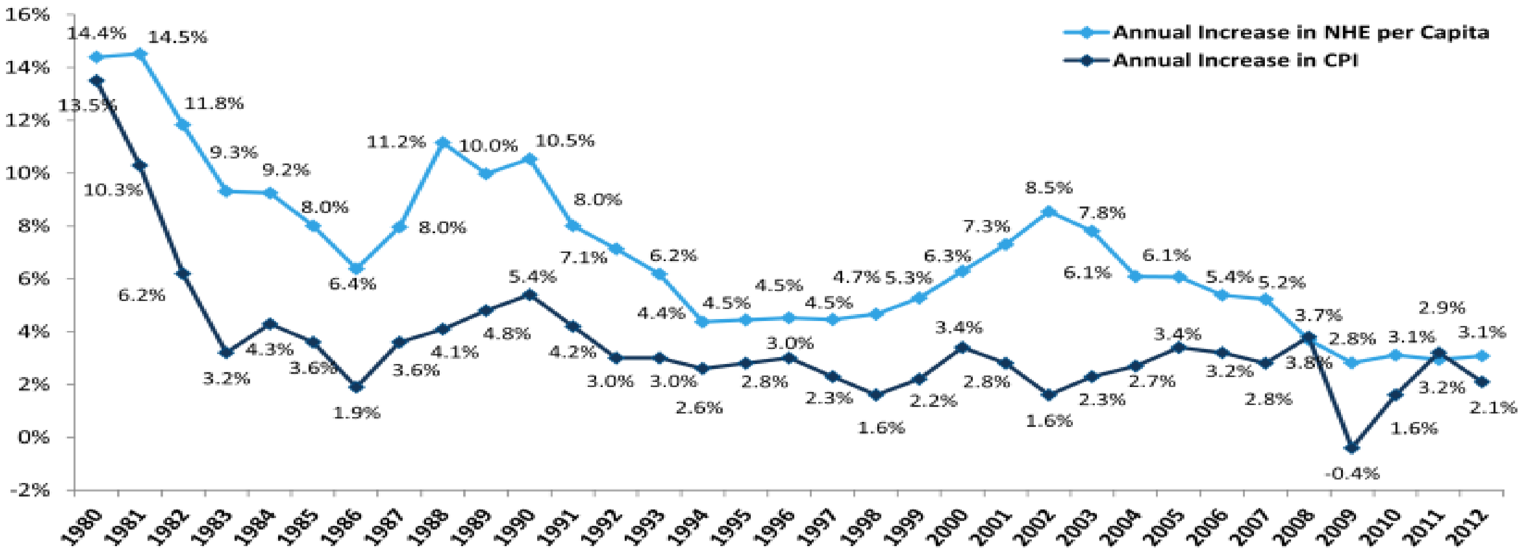
http://www.storytellingwithdata.com/2011/01/new-years-resolution-declutter-your.html



also, align labels on y axis ...
show same number of places
to right of decimal point

Info vis wiki: http://www.infovis-wiki.net/index.php?title=File:3Dbars02.jpg

Percent Annual Increase in National Health Expenditures (NHE) per Capita vs. Increase in Consumer Price Index (CPI), 1980-2012

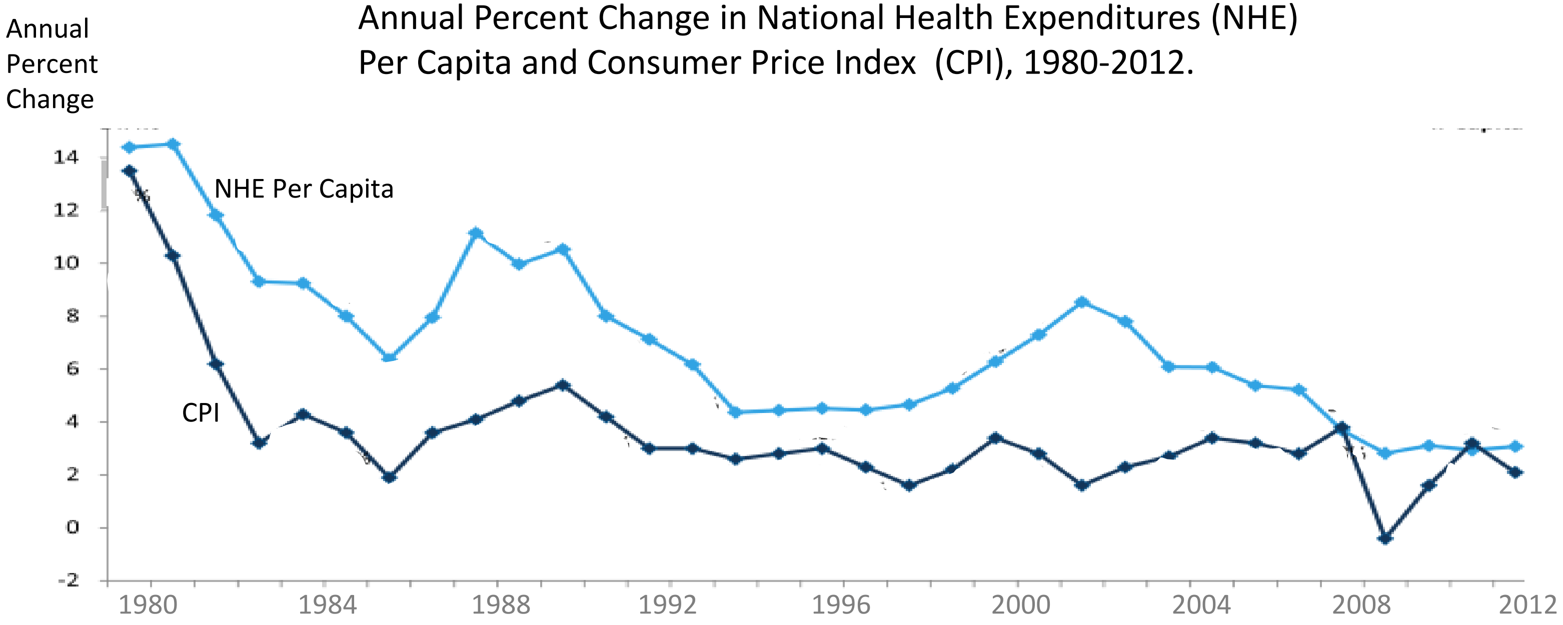


SOURCE: Kaiser Family Foundation calculations using NHE data from Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group, at http://www.cms.hhs.gov/NationalHealthExpendData/ (see Historical; National Health Expenditures by type of service and source of funds; file nhe12.zip), and CPI data from Bureau of Labor Statistics at ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt (All Urban Consumers, All Items, 1982-1984=100, Not Seasonally Adjusted, U.S. city average).

THE HENRY J. KAISER FAMILY FOUNDATION

- when possible, use direct labeling rather than legends
- don’t use too many tick marks or axis labels
- show x axis labels in upright position
- don’t repeat dollar signs or percent signs

Annual Percent Change in National Health Expenditures (NHE) Per Capita and Consumer Price Index (CPI), 1980-2012.



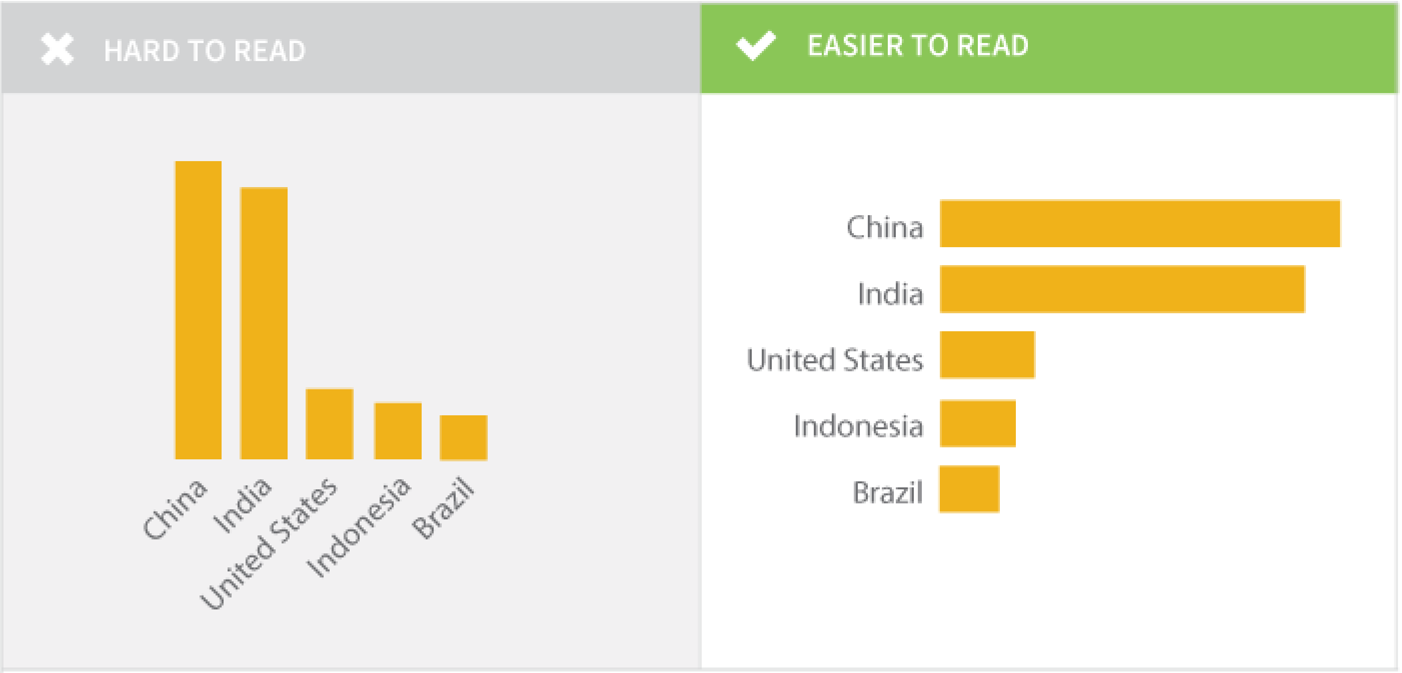
SOURCE: Kaiser Family Foundation calculations using NHE data from Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group, at http://www.cms.hhs.gov/NationalHealthExpendData/ (see Historical; National Health Expenditures by type of service and source of funds; file nhe12.zip), and CPI data from Bureau of Labor Statistics at ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt (All Urban Consumers, All Items, 1982-1984=100, Not Seasonally Adjusted, U.S. city average).

THE HENRY J. KAISER FAMILY FOUNDATION

- when possible, use direct labeling rather than legends
- don’t use too many tick marks or axis labels
- show x axis labels in upright position
- don’t repeat dollar signs or percent signs

typically, put outcome variable on vertical (y axis)

but, consider putting outcome variable on x axis if that makes long labels easier to read

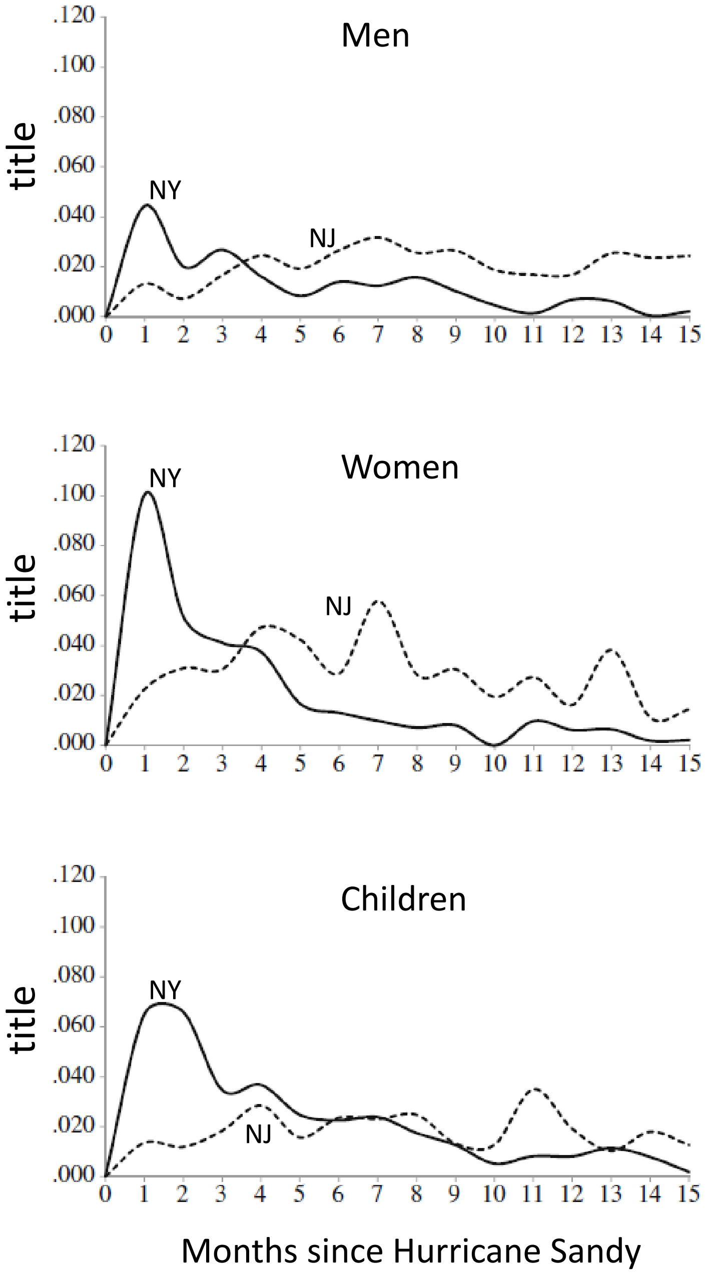


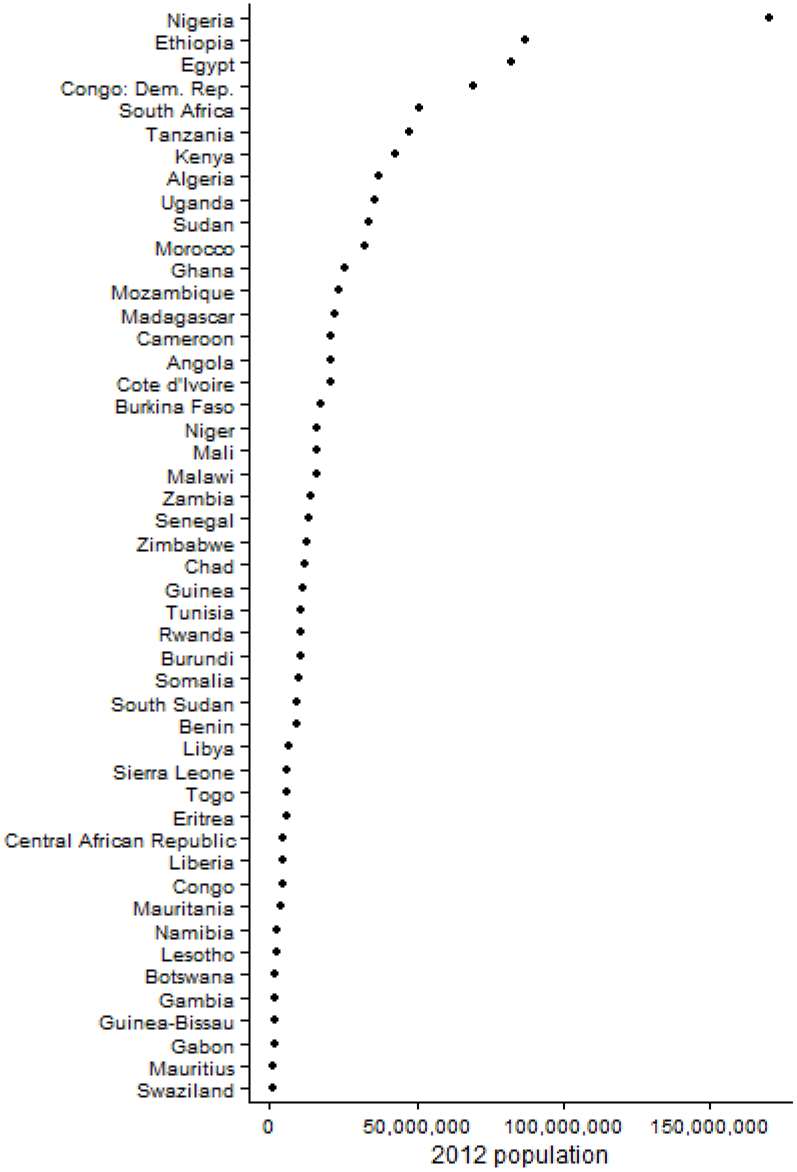
via Data + Design, https://infoactive.co/data-design/titlepage01.html

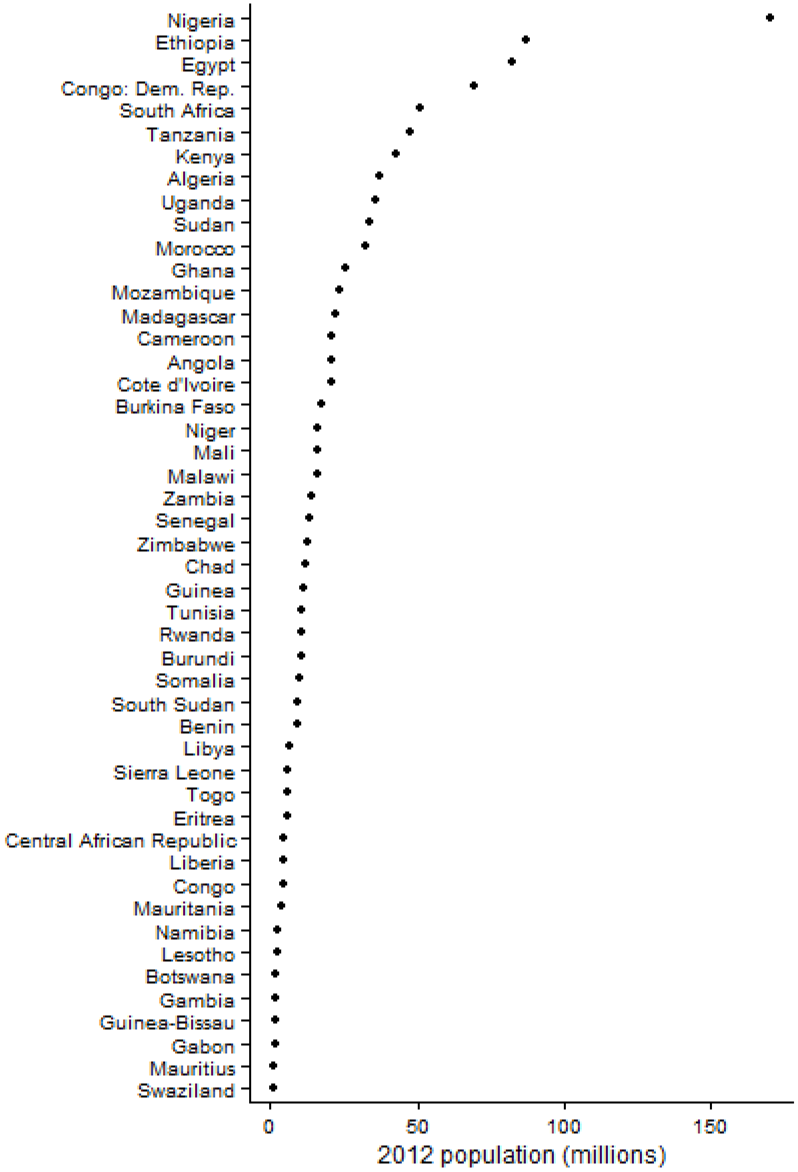
don’t show more decimal places than necessary in axis labels

why show 3 decimal places when only 2 are used?

consideration even more important when displaying small multiples where axis labels are repeated



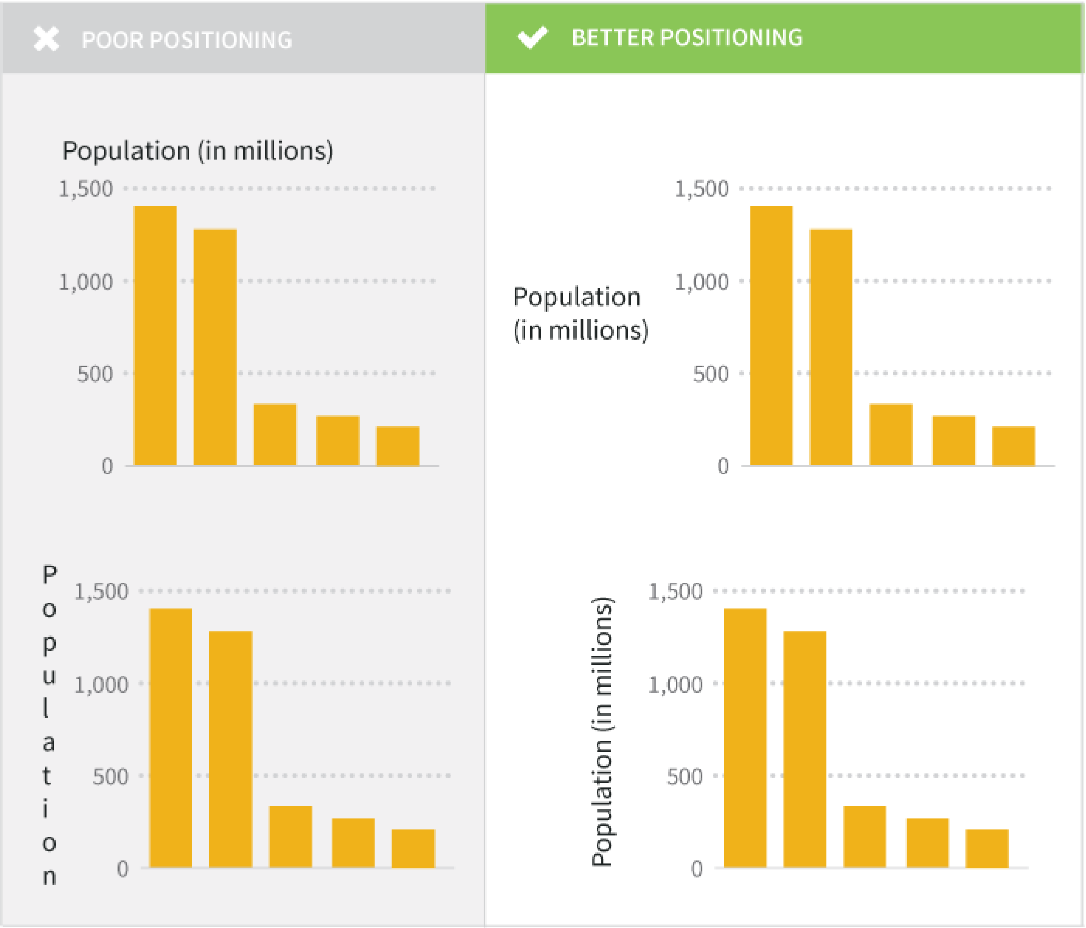




- use axis labels with units in thousands or millions rather than axis labels having long strings of zeros
- use axis title to indicate units
- is y axis title (country) needed??

axis titles give reader a summary description of variable being displayed

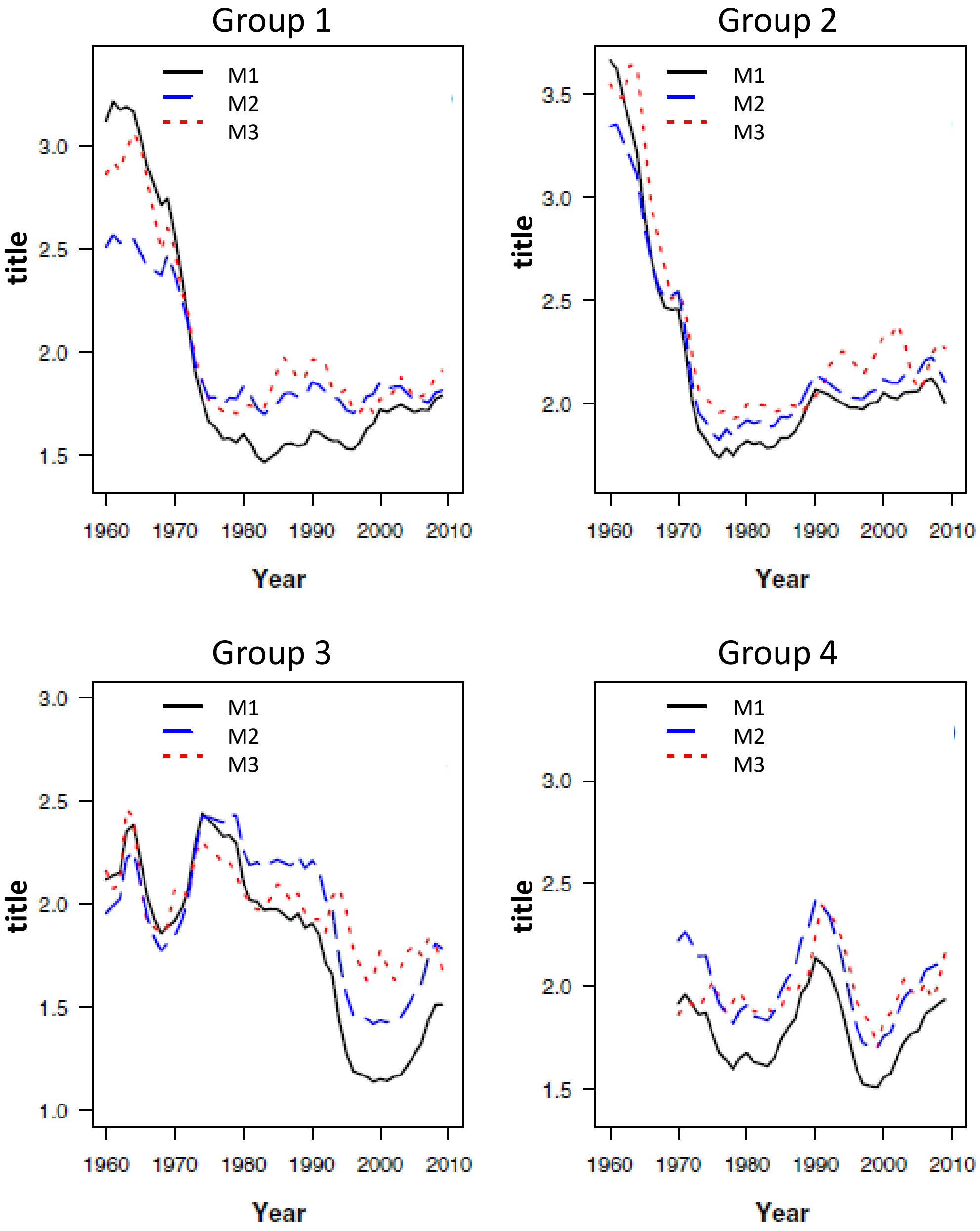
- include unit (percent, dollars, thousands, millions)



via Data + Design, https://infoactive.co/data-design/titlepage01.html

reduce clutter ...

- remove x axis title “Year”
- one legend (upper left)
- two y axis titles
 (left graphs only?)



reduce clutter ...

- remove x axis title “Year”
- one legend (upper left)
- two y axis titles (left graphs only?)

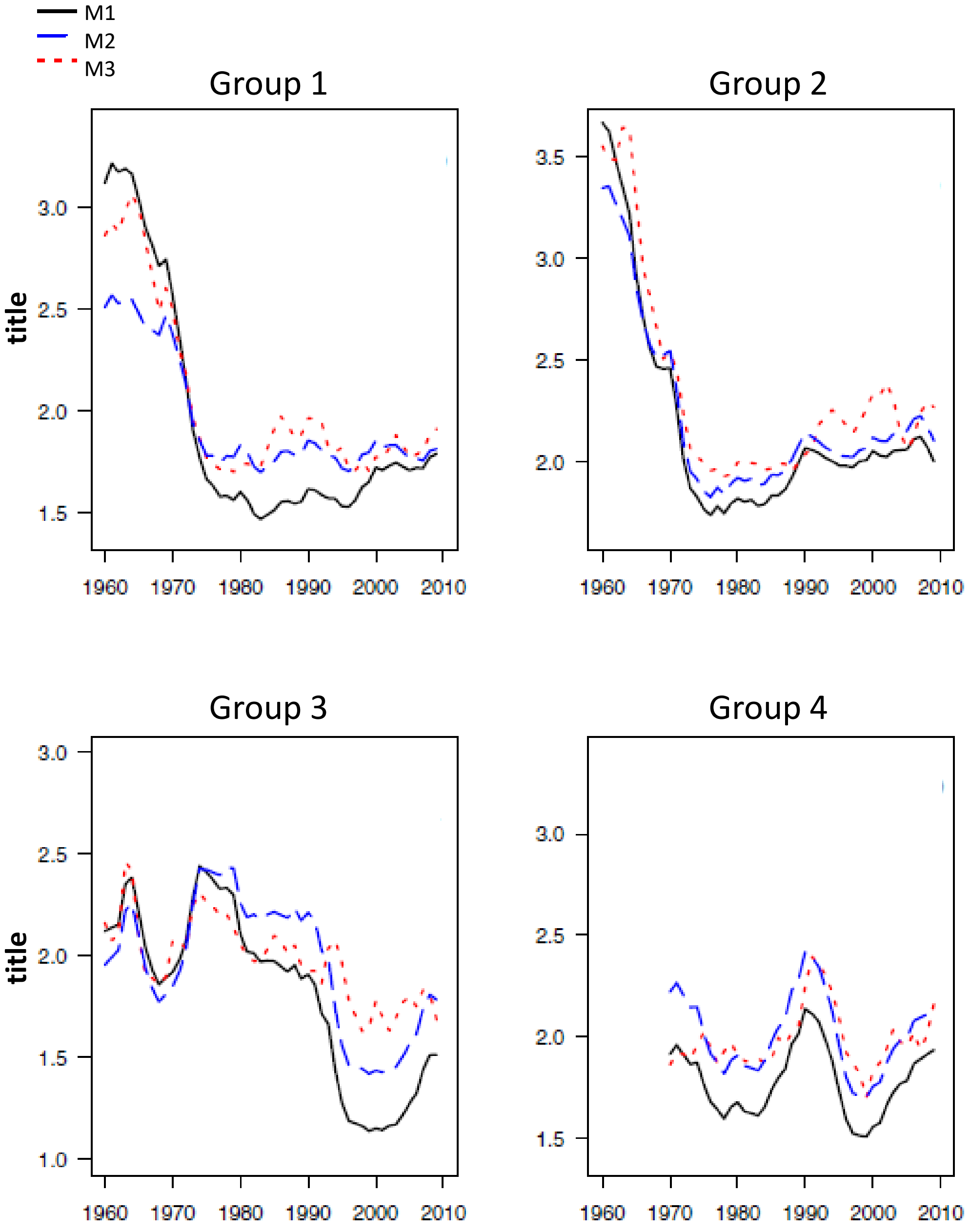
- why change color **and** pattern?

⇩

don’t vary two visual dimensions, when varying one will do

“The number of ‘information-carrying’ visual dimensions should not exceed the number of dimensions in the data.”

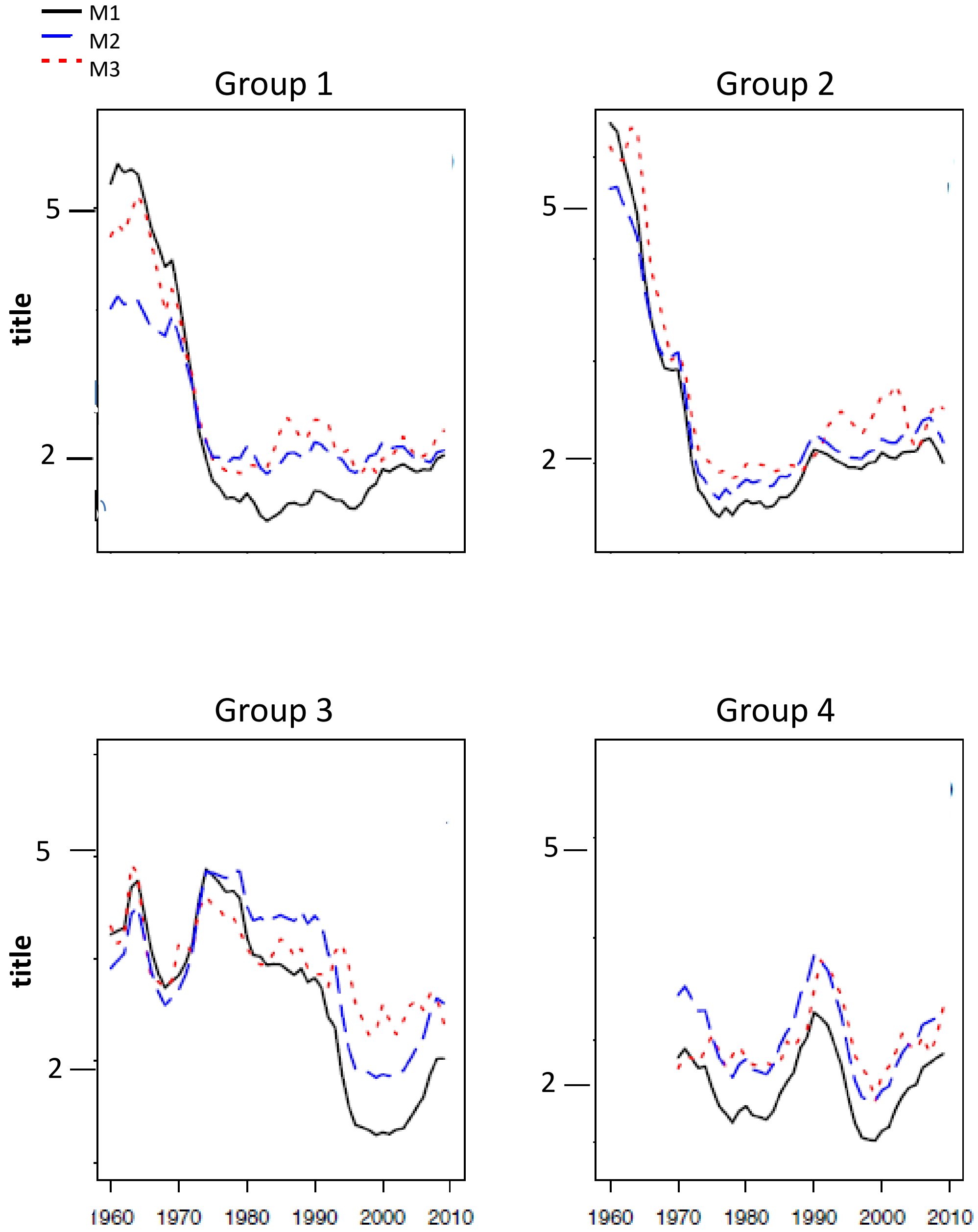
- Edward Tufte



reduce clutter ...

- remove x axis title “Year”
- one legend (upper left)
- two y axis titles (left graphs only?)

finally ... use common x and y scales to enable comparisons both within **and between** small multiples



spell out uncommon abbreviations

but what is uncommon? depends on the viewer

US
UK
HIV
STEM
URM
API
TFR
IMR
FSW
DU
MSM
SD
SA
BA
HI
.
.
.

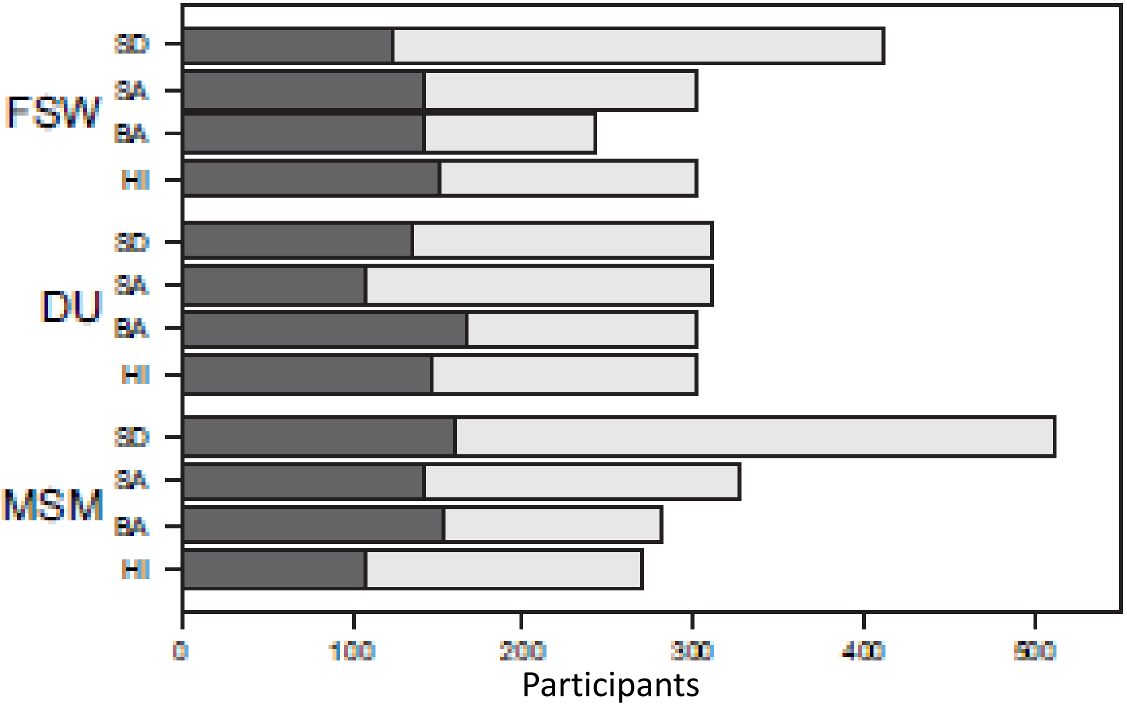
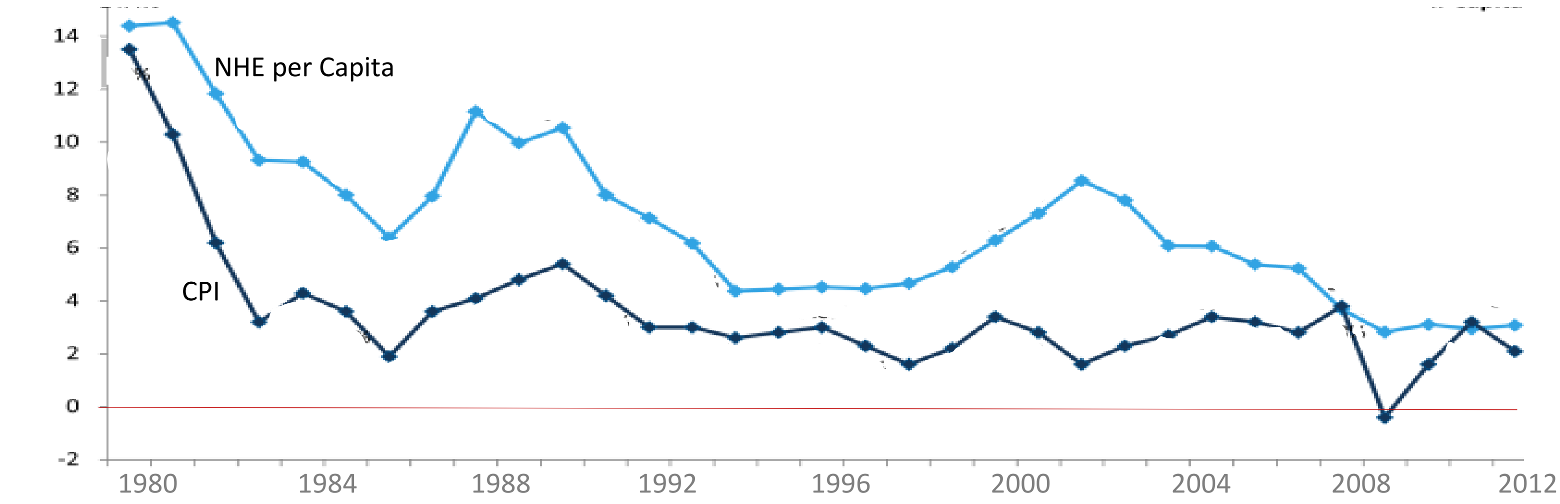


Fig. 2. Sample sizes from the 12 studies (in total, 3866 people participated, of whom 1677 (43%) completed a follow-up survey): ■ initial and follow-up □, initial only

Gile, K.G., Johnston, L.G., Salganik, M.J. 2014. Diagnostics for respondent-driven sampling." Journal of the Royal Statistical Society, Series A (Statistics in Society)

Annual Percent Change in National Health Expenditures (NHE) per Capita and Consumer Price Index (CPI), 1980-2012



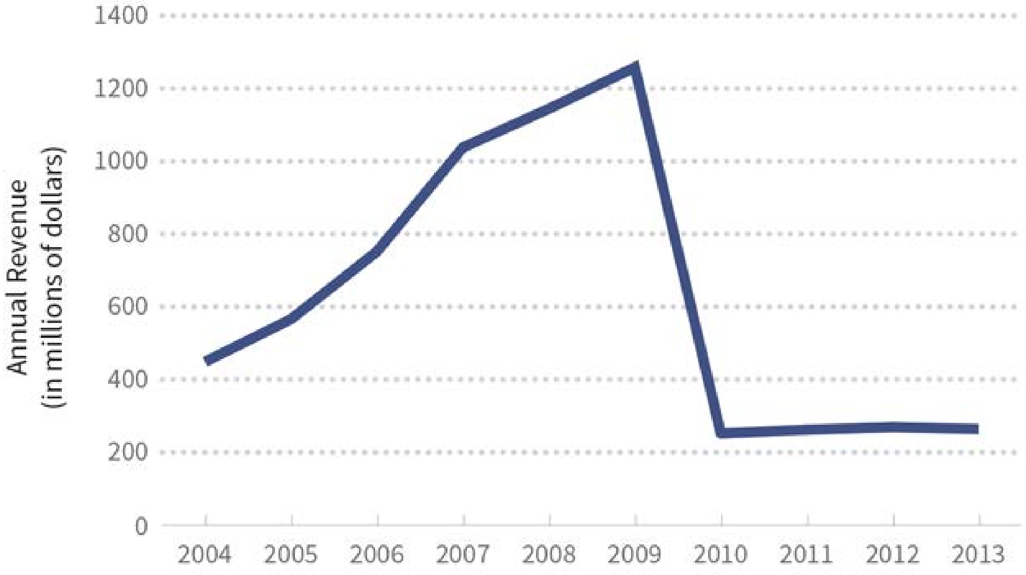
SOURCE: Kaiser Family Foundation calculations using NHE data from Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group, at http://www.cms.hhs.gov/NationalHealthExpendData/ (see Historical; National Health Expenditures by type of service and source of funds; file nhe12.zip), and CPI data from Bureau of Labor Statistics at ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt (All Urban Consumers, All Items, 1982-1984=100, Not Seasonally Adjusted, U.S. city average).

THE HENRY J. KAISER FAMILY FOUNDATION

- consider using a reference line to show an important value across an entire graph if it won’t interfere with displaying the data

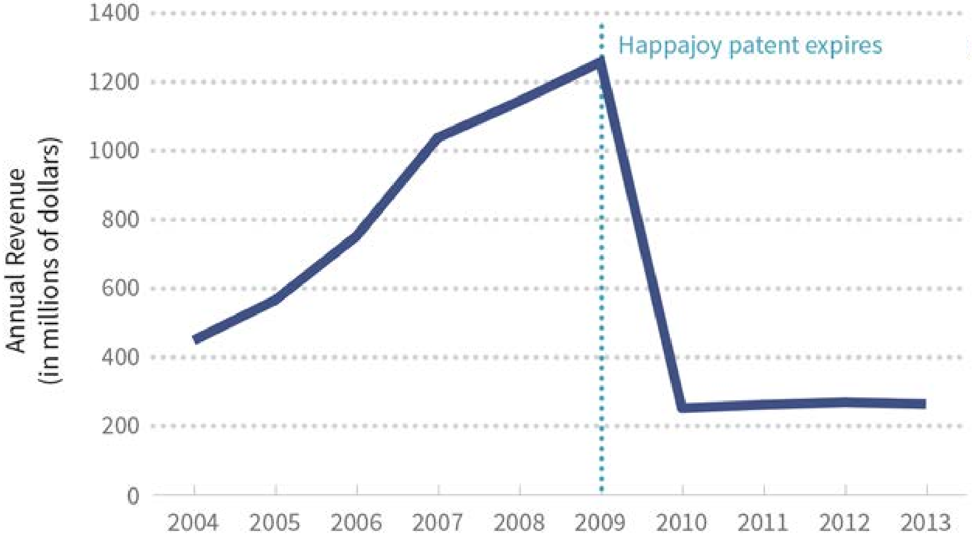
- is x axis title (year) needed ?

annotate with text



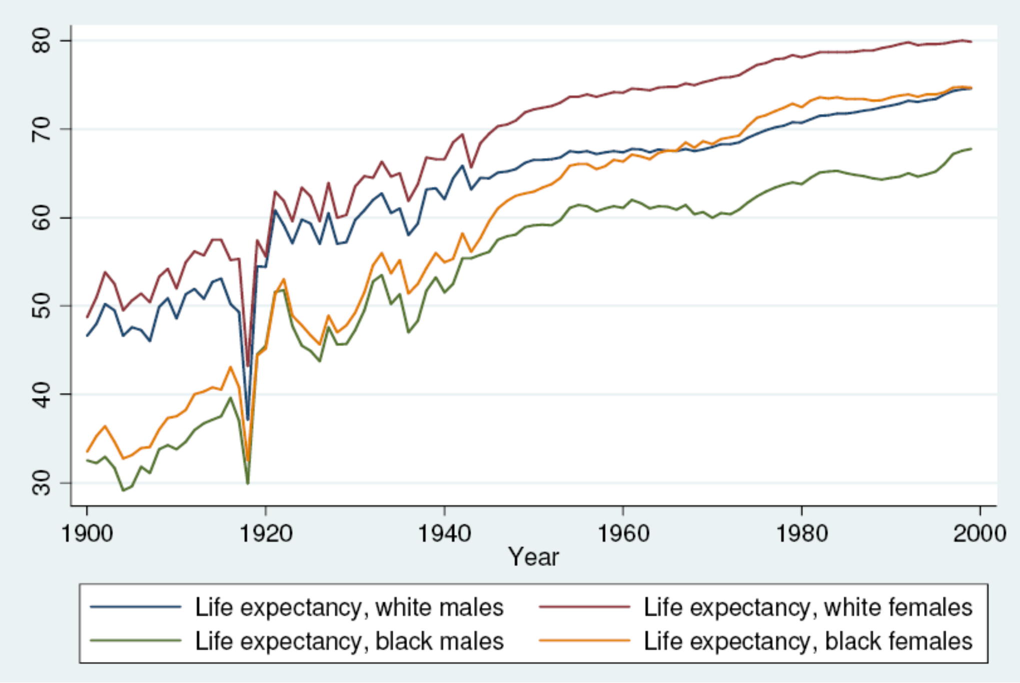
sometimes a few words are helpful

use text to highlight or explain

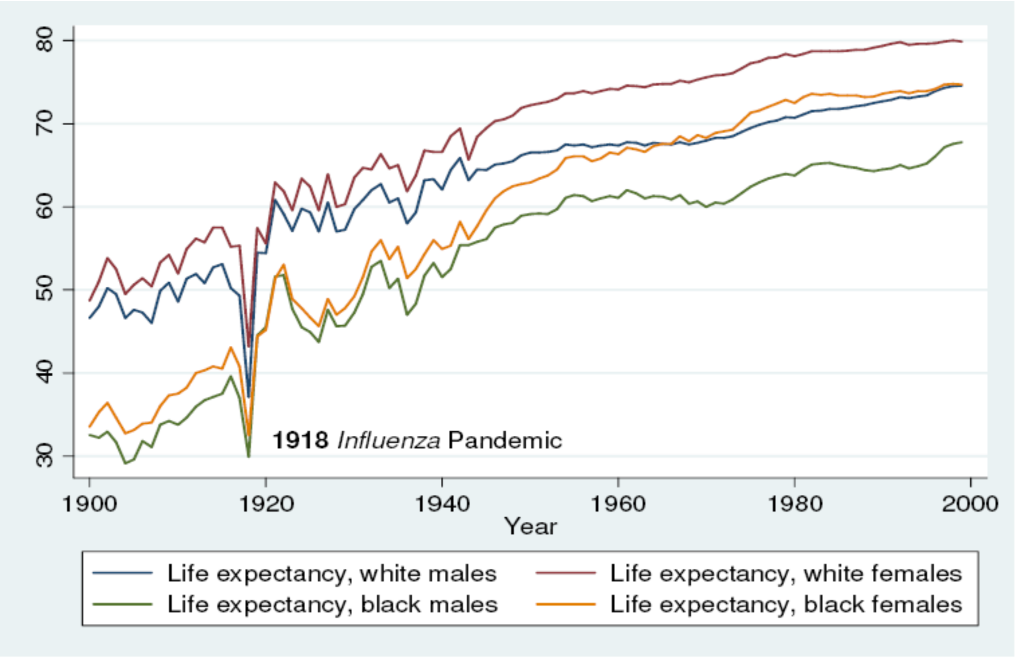


via Data + Design, https://infoactive.co/data-design/titlepage01.html

annotate with text

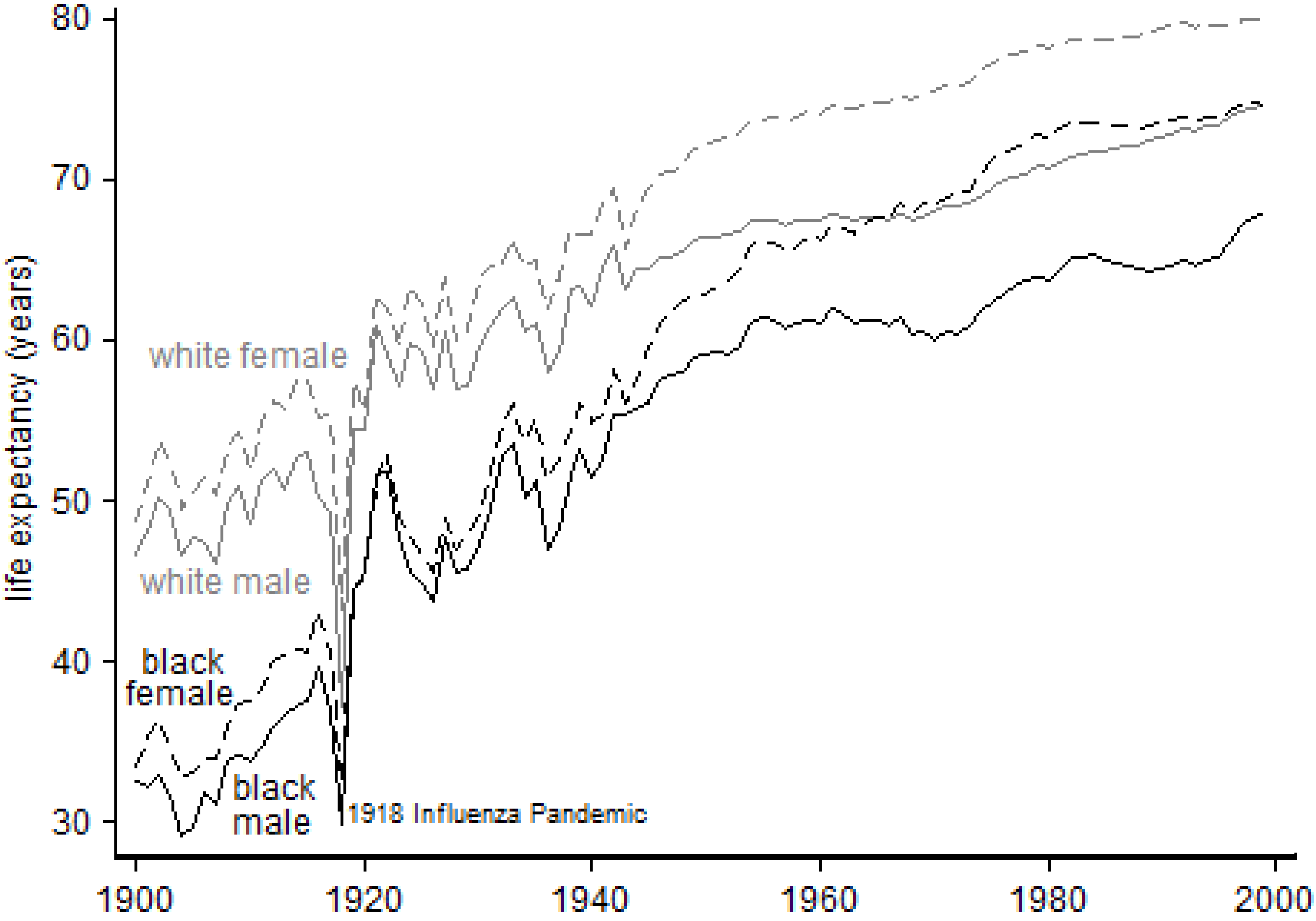


why are y axis labels laying on their sides??



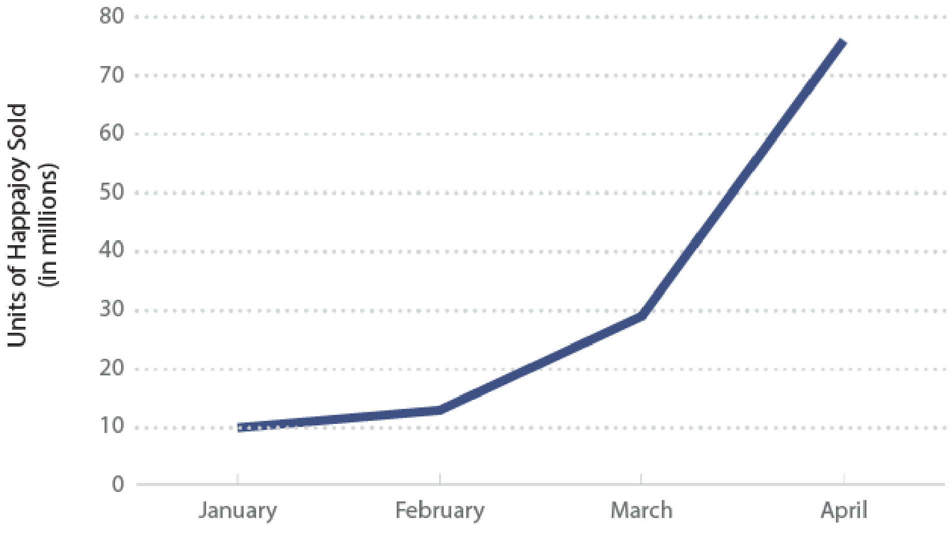
“upright “ y axis labels (orientation)

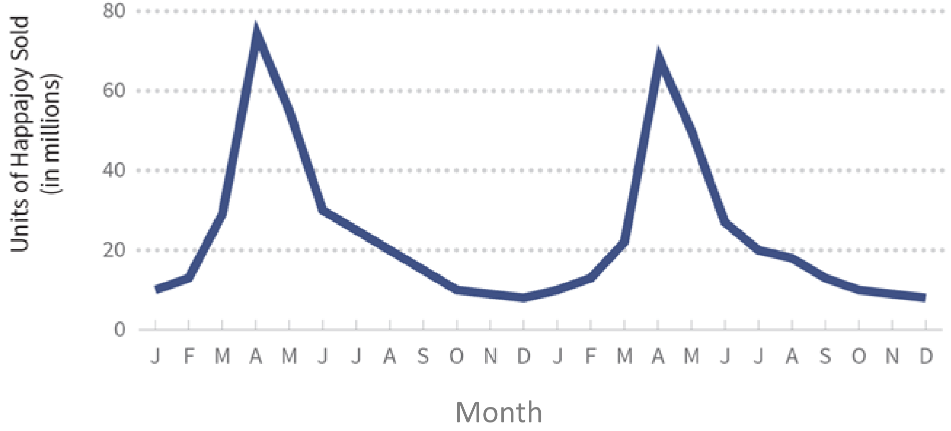
in general,
consider modifying tool’s default settings



provide context

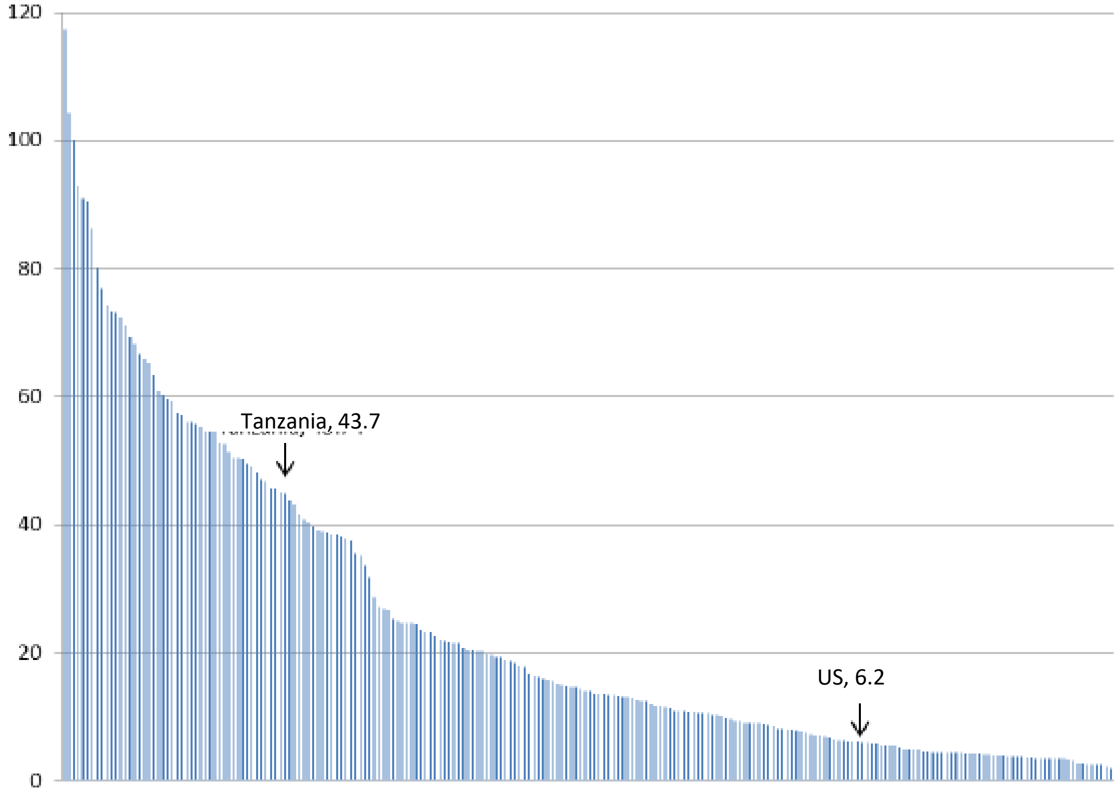
showing too few data points can sometimes give the wrong impression ("two data points do not define a trend"), especially for measures that vary cyclically





via Data + Design, https://infoactive.co/data-design/titlepage01.html provide context

National Infant Mortality Rate (IMR) per 1,000 Live Births, 2014



Lauren Gaydosh, Sociology PhD candidate, OPR Notestein seminar, September 2014.

lastly, consistency …

within and between graphs

be careful when there are multiple authors and multiple tools are used

axis label and axis title fonts, color, axis label orientation, …

let the data stand out by eliminating decoration

animation created by Darkhorse Analytics shows how communication can be greatly enhanced by eliminating clutter and de-emphasizing supporting elements.
Every aspect of a figure should be there on a “need to have it” basis.
https://speakerdeck.com/player/87bb9f00ec1e01308020727faa1f9e72#

Remove
to improve
(the **data-ink** ratio)

Created by Darkhorse Analytics www.darkhorseanalytics.com

V Summary

Summary of Statistical Graphics Considerations

Choice of table vs graph

- focus on the forest or the trees

Audience and Setting

Correctness

- scale considerations
  - for bar graphs, include natural baseline (usually 0), because relative length of bars is basis for comparison
  - for line graphs, need an interval scale, otherwise relative line segment slopes are meaningless
  - graphs with dual scales show meaningless points of intersection and meaningless relative slopes, unless two scales are simply two names for the same value
  - for circles (bubbles), use area not radius to represent magnitude of values
- data density and data hiding considerations (overplotting)
  - consider point size, fill, shape, transparency, and data stratification
  - use jittering to sacrifice positional precision for more accurate display of data density

Summary of Statistical Graphics Considerations

Comparisons

- determine **true** quantity of interest
  - difference/ratio between A and B rather than A and B
    - absolute difference or percent difference

- make sure data is easily seen
  - size of data markers and contrast against background
  - not hidden by other data markers (points, lines, areas), labels, legends, tick marks, or gridlines

- show the data, not just summary measures, when possible
  - for continuous data, consider box plots, violin plots, histograms or density plots to compare groups

- involve perceptual tasks high on Cleveland’s list of performing accurate judgments
  - position along a common scale
  - position along identical, non-aligned scales
  - length
  - angle, slope
  - area
  - color

- consider proximity, alignment and ordering

Summary of Statistical Graphics Considerations

Comparisons

- dot plots rely on judgment of position along a common scale
- small multiples rely on judgment of position along identical, non-aligned scales
- bar charts rely on judgment of length
    - stacked bar charts … difficult without a common baseline
    - grouped bar charts … difficult when bars are not adjacent (proximity)
- pie charts rely on judgment of angle
- bubble charts rely on judgment of area
- use color for a reason:
    - to indicate groups (via hue)
    - to highlight particular data (via hue or intensity)
    - to encode quantitative data (via intensity)

Summary of Statistical Graphics Considerations

Simplicity:

- don’t use too many tick marks or axis labels
- don’t repeat dollar signs or percent signs in axis labels
- don’t display more decimal places than necessary in axis labels
- don’t use long strings of zero’s in axis labels
- when possible, use direct labeling rather than legends
- consider whether axis titles are necessary
- in graphs showing small multiples, repeat axis labels, axis titles and legends only where needed
- don’t use background shading
- consider whether borders and gridlines are needed
- don’t vary multiple visual dimensions, when varying one will do
- beware of default decoration provided by packages

Clarity:

- don’t let data labels obscure data
- show axis labels in upright position
- include units in axis titles
- consider whether abbreviations and acronyms are familiar to audience
- consider using reference lines
- annotate with text to highlight or explain
- provide context for data
- be consistent both within and between graphs

Above all else: let the data stand out

VI Conclusions

"Graphical excellence is the well-designed presentation of interesting data. It consists of clarity, precision and efficiency."

- Edward R. Tufte

"Pair the depth and clarity of your data, models and writing with visualizations that are just as clear and compelling."

- Jonathan A. Schwabish

"Don't skimp on graphs, they're worth the investment."

- Matthew J. Salganik

(verbal communication)

Tufte, E. The Visual Display of Quantitative Information, Second Ed. Graphics Press, Cheshire CT. 2001.
Schwabish, J. An Economist's Guide to Visualizing Data, Journal of Economic Perspectives, Winter 2014.