

Use Graphs Effectively

UCSF, Epi 270, Ph.D. Seminar

David V. Glidden
Professor of Biostatistics UCSF

david.glidden@ucsf.edu



Irizarry, Introduction to Data Science, Chapter 10

Broman, Creating Effective Tables and Graphs

This is my favorite part about analytics: Taking boring flat data and bringing it to life through visualization.

Quantitative Communication

1. Graphs are **engaging**
people perk up
2. Unique way to **disseminate** work
*your study may be summarized in 1 graph
use that to your advantage*
3. There is an aspect of **creativity**
*graphs involve subtle choices
maintain transparency, intellectual honest*

My career highlight: having some of my
graphs used in a plenary review talk



"When communicating results to non-technical types there is nothing better than a clear visualization to make your point"

-Tukey

Some Principles

- 1.K.I.S.S. based methods
keep it simple, stupid
- 2.Think of your audience
readable, informative, illuminating to them
- 3.Maximize ratio of information to ink
no pie charts, bar + whisker plots, or pointless 3d
- 4.Choice of scale can illuminate or deceive
include full scale, sometimes not, consider logs
- 5.Facilitate the most important comparisons
thoughtful about ordering groups

Two Way Display

Internet browser usage

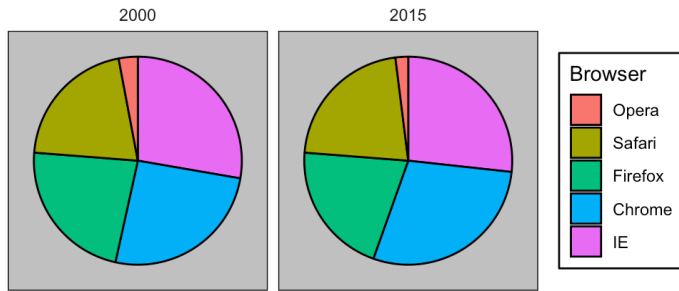
Browser	2000	2015
Opera	3	2
Safari	21	22
Firefox	23	21
Chrome	26	29
IE	28	27

Outcome: which browser (%)
unordered categorical

Predictor: 2000 v. 2015
binary

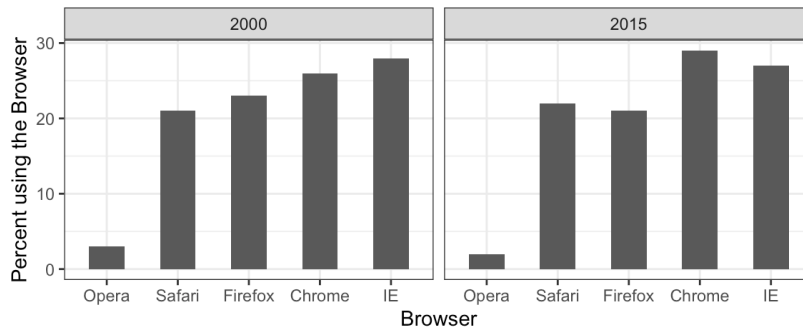
We could just stick with a table.....

No to Pie Charts



People are poor at judging area and angles (principle 2)

Better at judging position and length



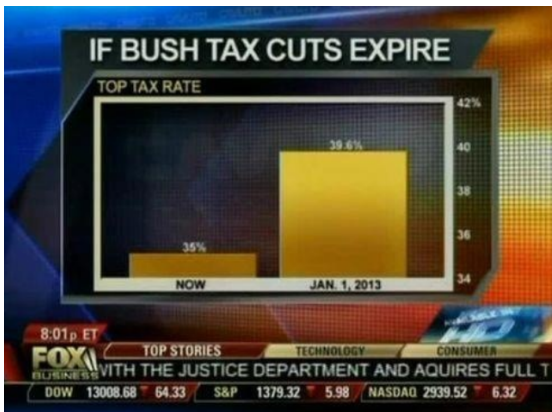
Terrible ink to information ratio (principle 3)

"There is no data that can be displayed in a pie chart, that cannot be displayed BETTER in some other type of chart"

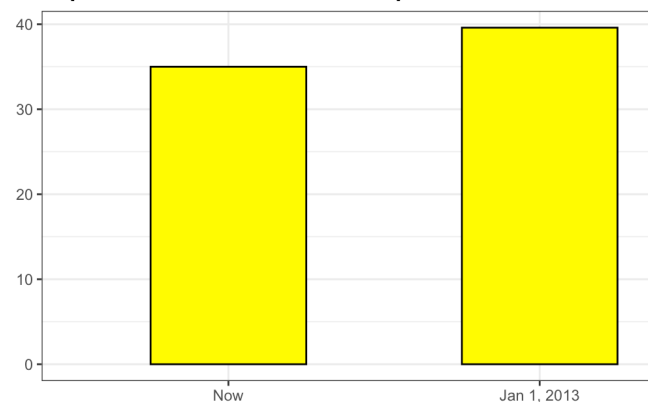
-Tukey

Know When to Include 0

Now: 35.0%; Jan 1, 2013: 39.6%



Top Tax Rate If Bush Tax Cut Expires



left bar height = $35 - 34\% = 1\%$
 right bar height = $39.6 - 34\% = 5.6\%$

Orders is correct

Magnitude is distorted

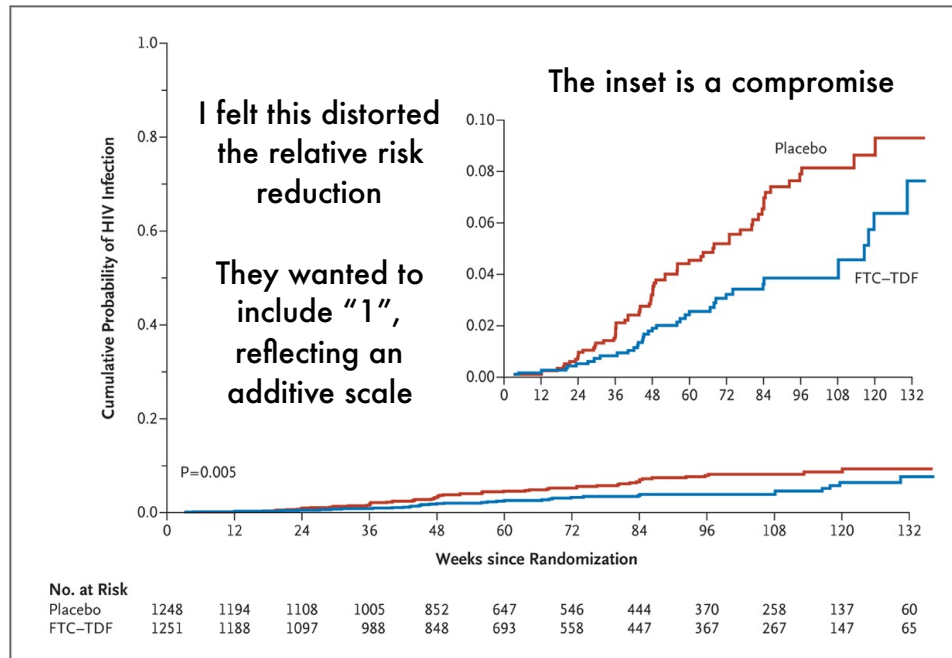
Ratio of bar height = 5.6, principle #4

"Visualization is often used for evil - twisting insignificant data changes and making them look meaningful. Don't do that crap if you want to be my friend. Present results clearly and honestly. If something isn't working - those reviewing results need to know."

-Tukey

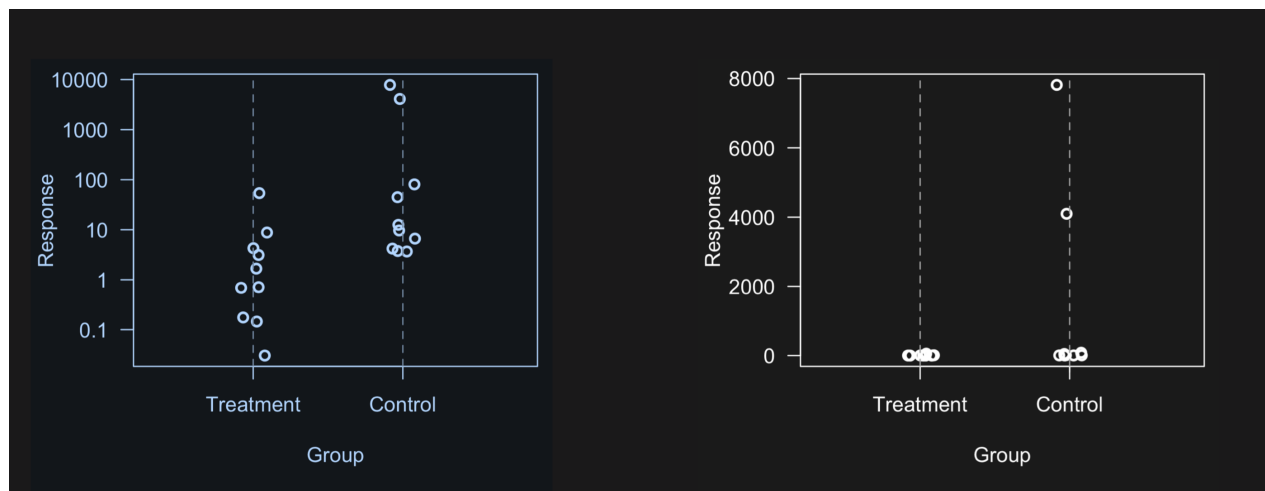
I Once Fought with NEJM

(I lost)



Grant et al 2010

The log scale



Can show more of the data

Preserve ordering

Modifies magnitude

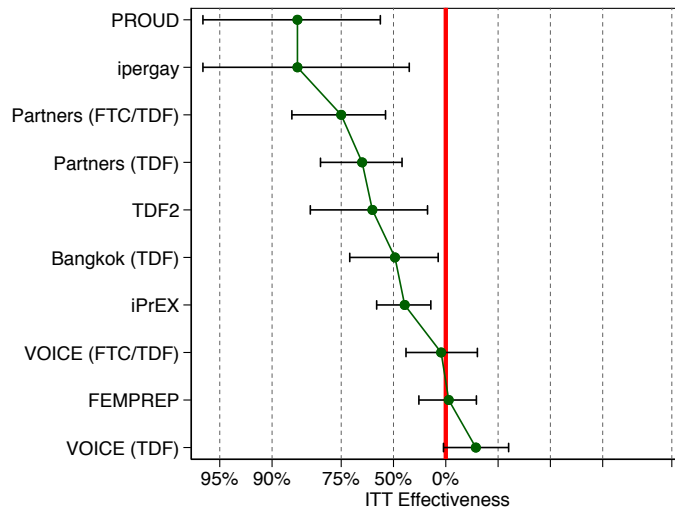
Can be particularly compelling when fold changes are of interest.

Keep labels on original scale

Using Log Scale

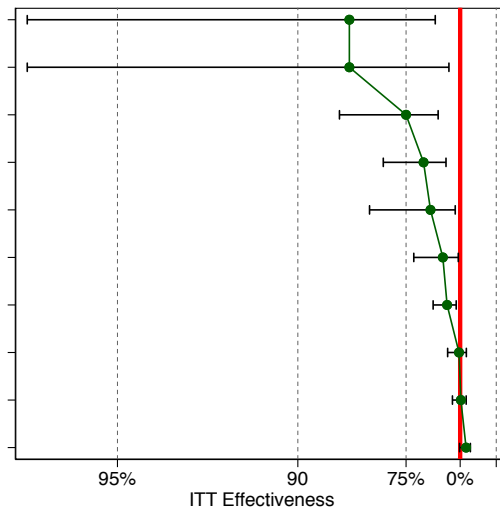
RR of
Placebo
to HIV PrEP

20 10 3 2 1 1/2



A RR based measure
compared to others
more nature on RR scale

20 10 3 2 1



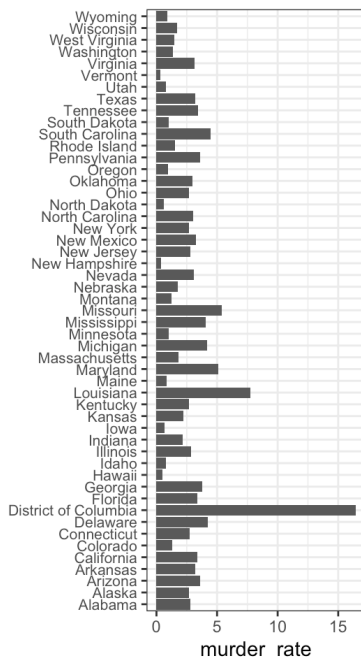
2 biggest RRs
dominate the plot

Note: % Effectiveness = $100 * (1 - 1/RR)$

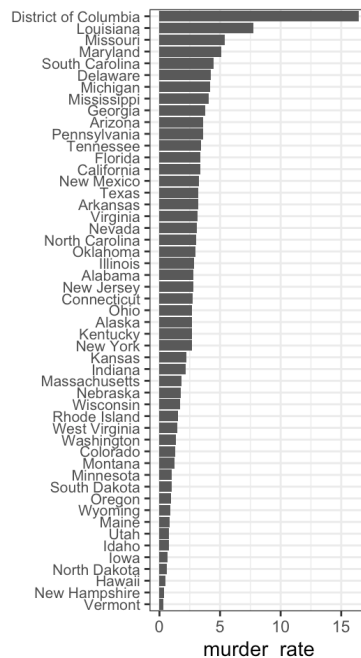
Order Categories Meaningfully

Principle 5

vertical or horizontal?



the "per" is missing!



what if scale went to 20 or 30?

We want a sense of
rank and magnitude

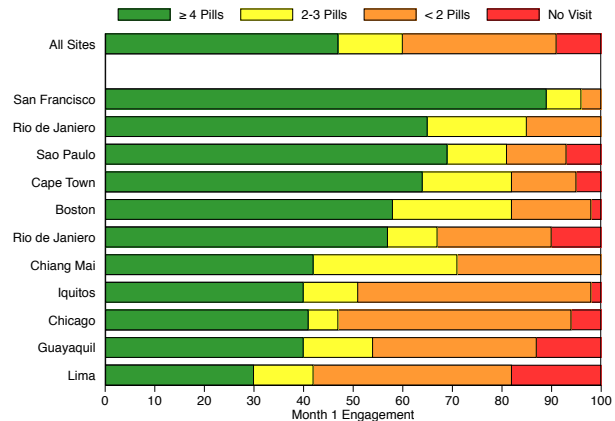
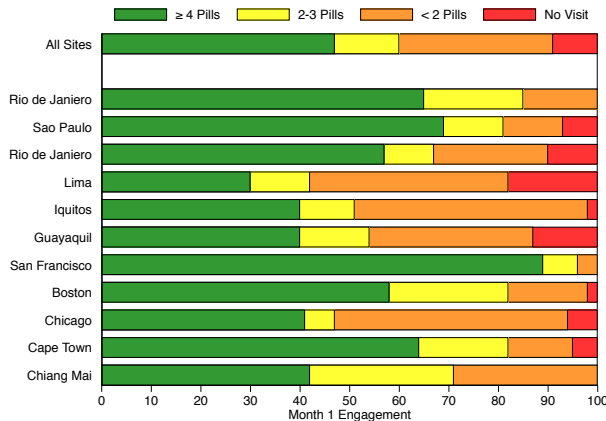
Left panel gives us that

Easiest comparison is to
nearest rate

Easy to compare highest
v. lowest rate as well

Not nearest alphabetically
(or in data code)

How is this ordering?



Adherence to HIV PrEP determined by drug level tests by study clinic:
order category (adherence: pills per week) by unordered category (site)

Left one order by site codes

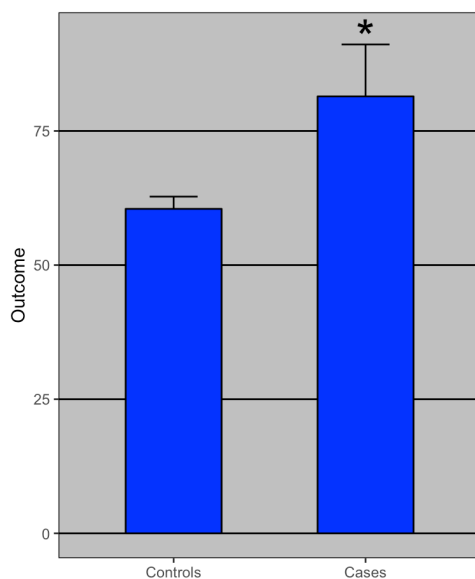
Right one order by overall adherence

Colleagues asked me to use the one on the left

why?

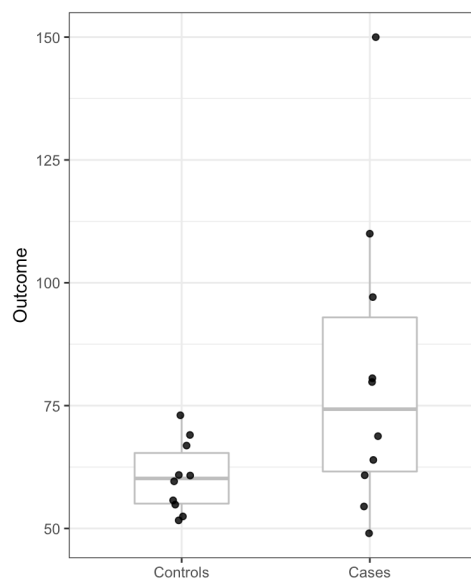
Show Distributions

Dynamite plots must die, Irizarry



Plot based on 4 numbers

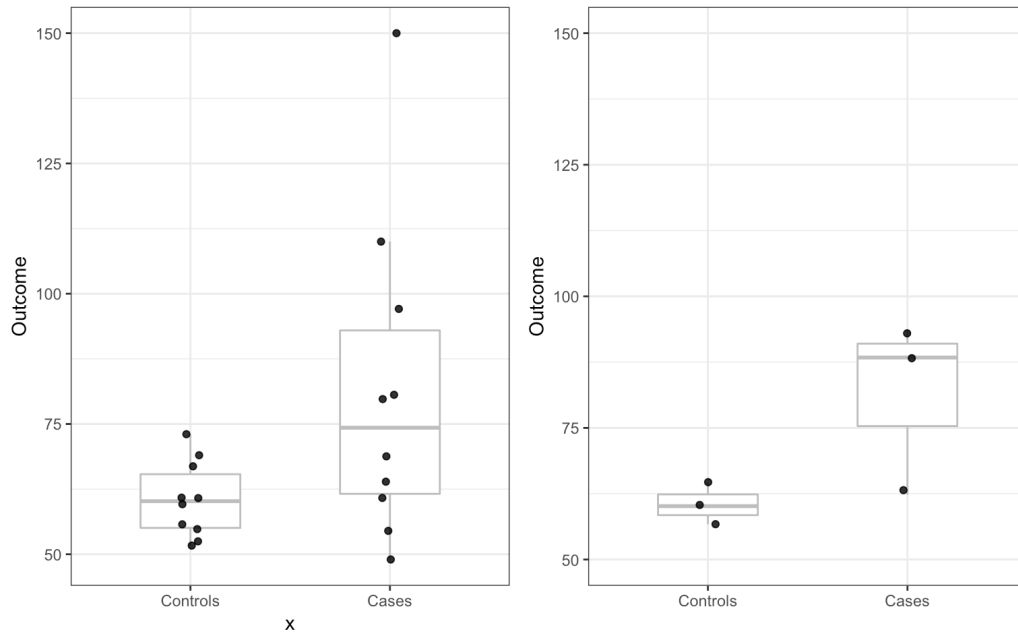
##		x	average	se
## 1	Controls	60	2.3	
## 2	Cases	81	9.7	



Ink to Information much higher
Includes "robust" summaries
(mean, quartiles)

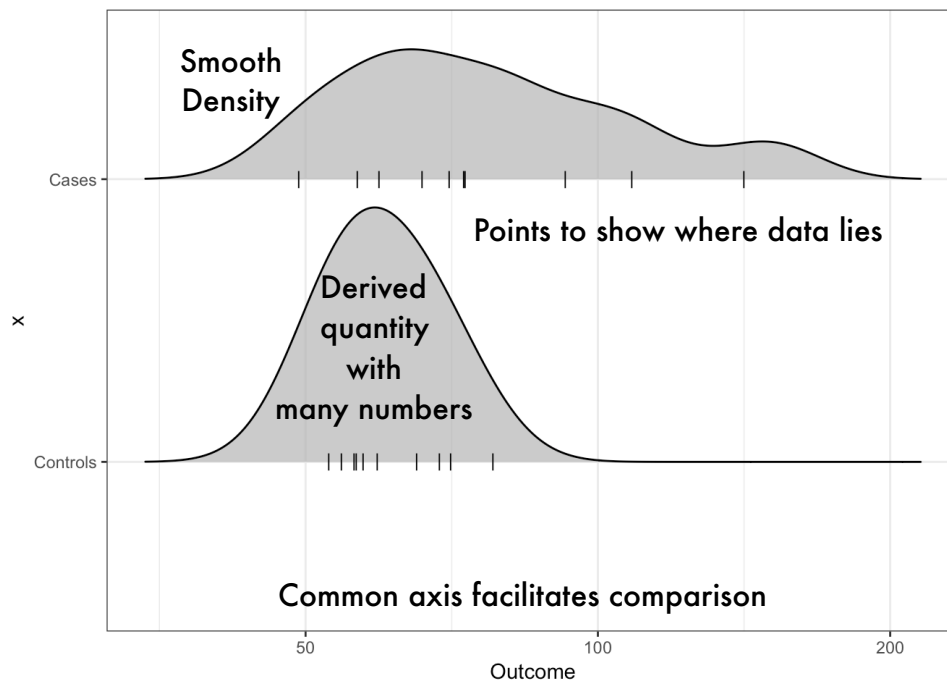
Different Distributions

Same dynamite plot



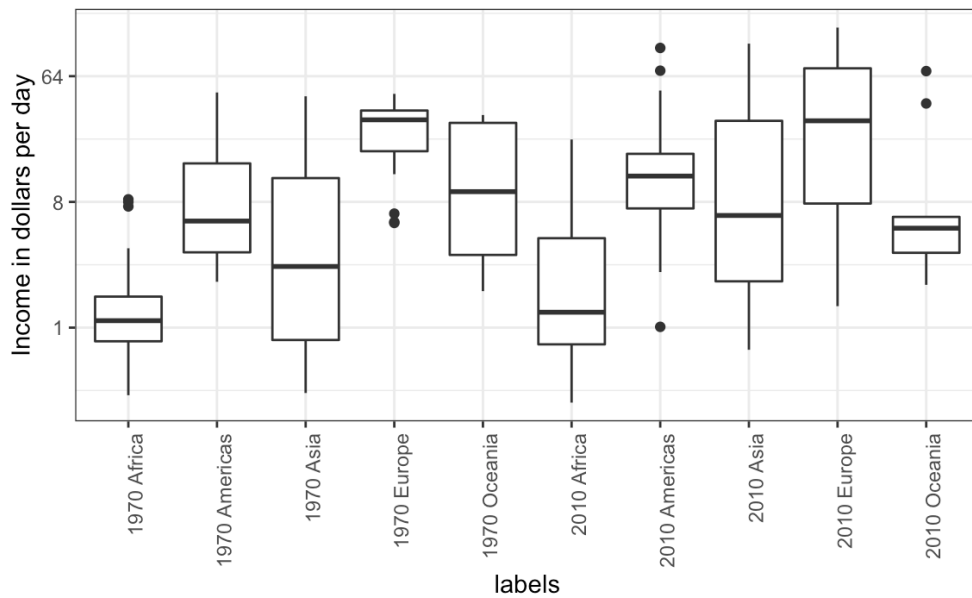
Density Estimates

High Information to ink ratio



Income by Continent

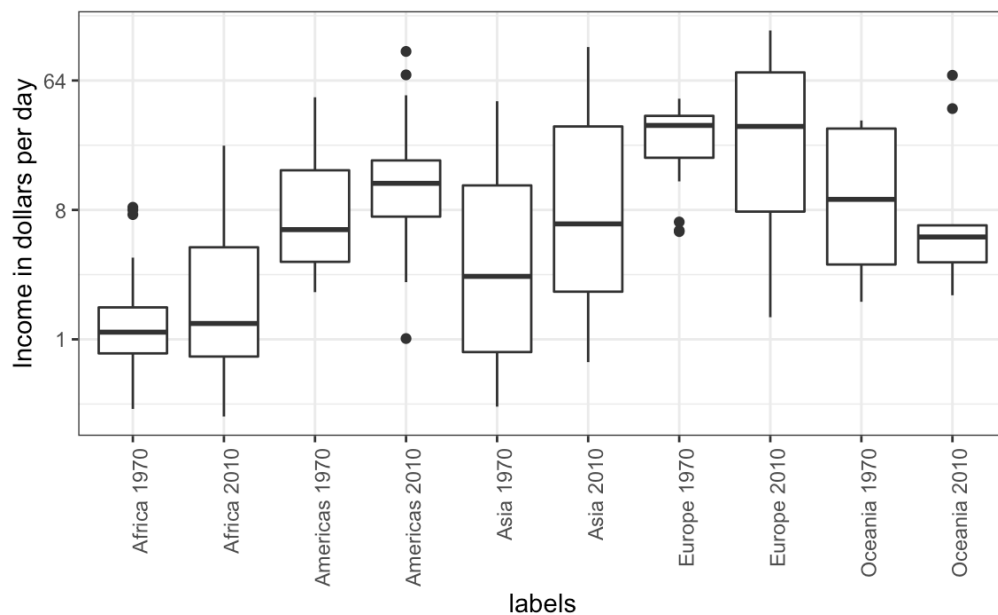
income in countries by continent by year



How has income change from 1970 to 2010?

Adjacency

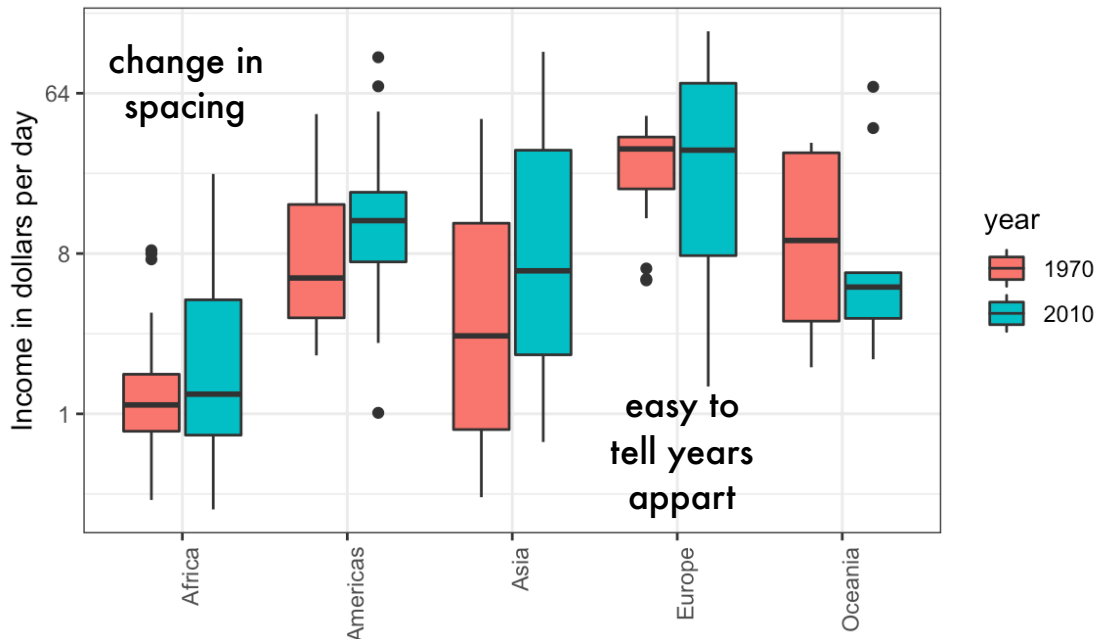
Principle 5



Like things near each other

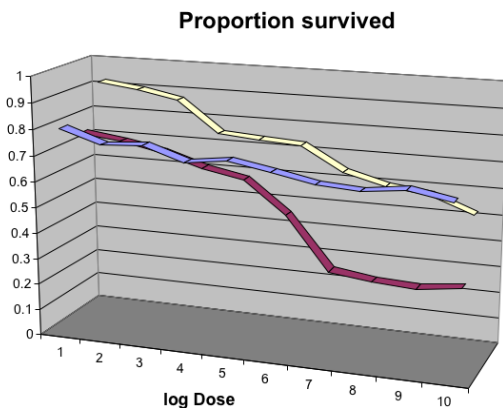
Colors and Spacing

seamlessly invites comparisons



Use of principle 5

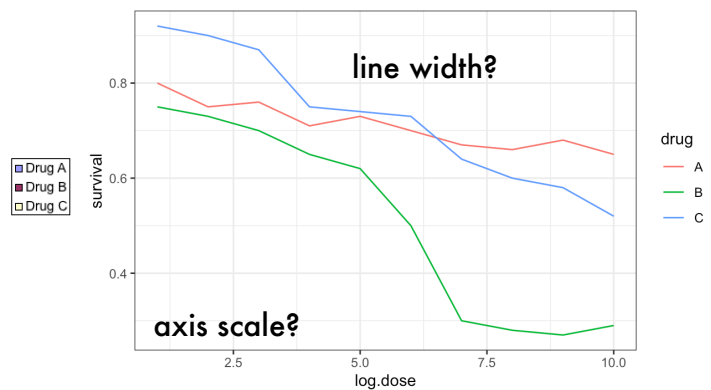
Reject Pointless 3d



low ink to information

hard to read values

eye is drawn away from data



high ink to information

easier to read off values

uncluttered, crisp contrasts

Avoid Excessive Precision

Principle #3

Too many digits
is just cluster

state	year	Measles	Pertussis	Polio
California	1940	37.8826320	18.3397861	0.8266512
California	1950	13.9124205	4.7467350	1.9742639
California	1960	14.1386471	NA	0.2640419
California	1970	0.9767889	NA	NA
California	1980	0.3743467	0.0515466	NA

Cleaner on
the eye

I round my results
quite a lot.

RR typically 1 decimal

state	year	Measles	Pertussis	Polio
California	1940	37.9	18.3	0.8
California	1950	13.9	4.7	2.0
California	1960	14.1	NA	0.3
California	1970	1.0	NA	NA
California	1980	0.4	0.1	NA

are we comparing diseases?

Applying Adjacency

Principle 5

Transpose
Columns and
Row

state	year	Measles	Pertussis	Polio
California	1940	37.9	18.3	0.8
California	1950	13.9	4.7	2.0
California	1960	14.1	NA	0.3
California	1970	1.0	NA	NA
California	1980	0.4	0.1	NA

We don't
need the state

disease	1940	1950	1960	1970	1980
Measles	37.9	13.9	14.1	1.0	0.4
Pertussis	18.3	4.7	NA	NA	0.1
Polio	0.8	2.0	0.3	NA	NA

We can almost see the plot

Think of Audience

- Graphs should be clear and visually compelling
clear comparisons, relevant, labelled, ink-to-information
- Should have a clear message
*you're try to say things you couldn't in other ways
they have a point of view
without using distortion*
- The details matter. People notice unconsciously
- A graph should be self contained.
footnotes for acronyms or important details.

*I know of no person or group
that is taking nearly adequate
advantage of the graphical
potentialities of the computer.*

- Tukey

Resources

Reading

Irizarry, Introduction to Data Science, Chapter 8-10

free online

Broman, Creating Effective Tables and Graphs

Dynamite plots must die, Irizarry

Code

<https://github.com/rafalab/dsbook>

https://github.com/kbroman/Talk_Graphs/tree/MDPhD2016