MIS2502:
Data Analytics
Principles of Data Visualization

Alvin Zuyin Zheng

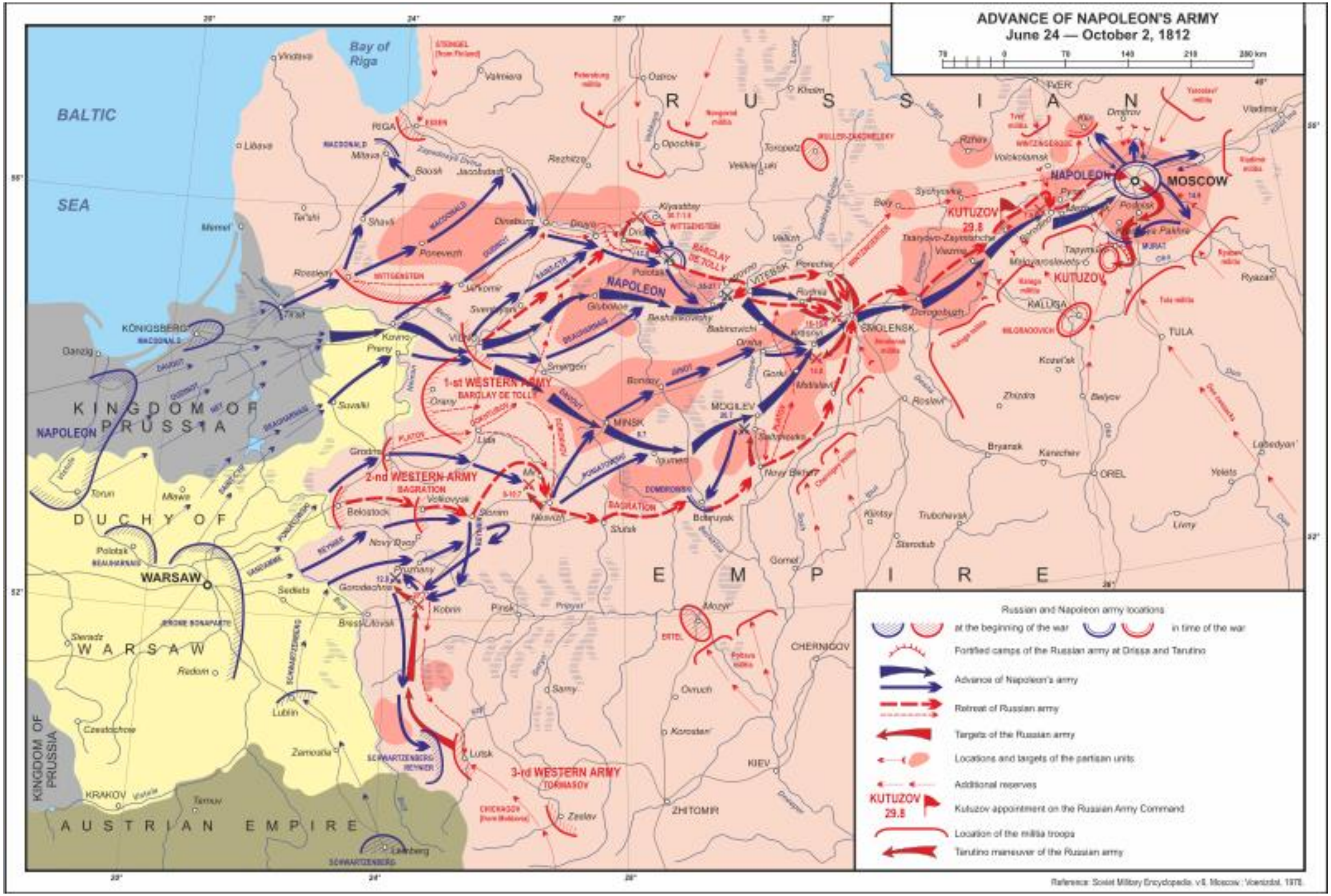
zheng@temple.edu

http://community.mis.temple.edu/zuyinzheng/

Data visualization can:

- provide clear understanding of patterns in data
- detect hidden structures in data
- condense information

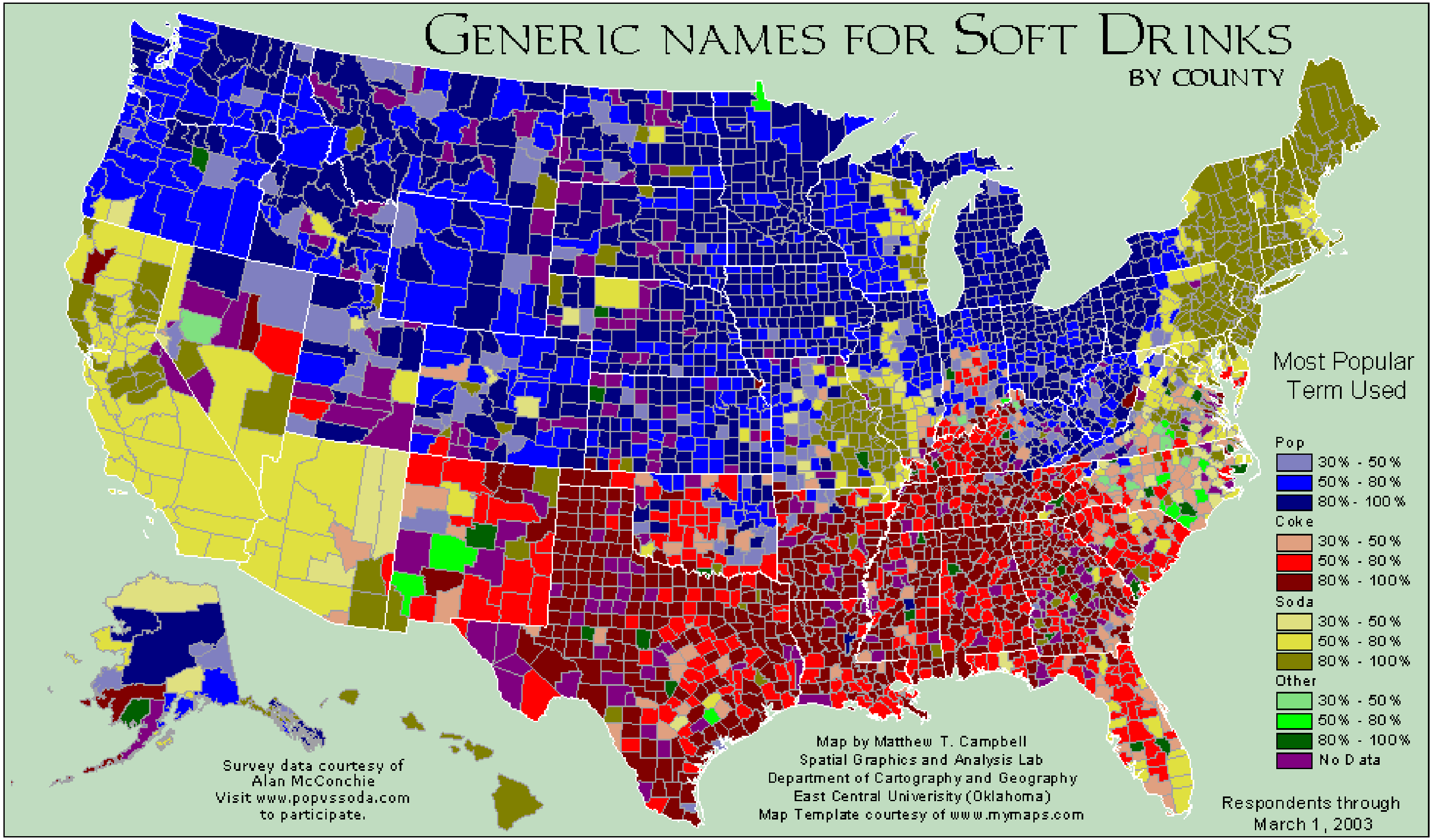
What makes a good chart?


ADVANCE OF NAPOLEON'S ARMY
June 24 — October 2, 1812
70 0 70 140 210 280 km
BALTIC SEA
Bay of Riga
RUSSIAN EMPIRE
KINGDOM OF PRUSSIA
DUCHY OF WARSAW
KINGDOM OF PRUSSIA
AUSTRIAN EMPIRE
MOSCOW
NAPOLEON
KUTUZOV 29.8
SMOLENSK
VITEBSK
MINSK
MOGILEV
WARSAW
RIGA
KÖNIGSBERG
KIEV
CHERNIGOV
ZHITOMIR
KALUGA
TULA
OREL
TVER
1-st WESTERN ARMY
BARCLAY DE TOLLY
2-nd WESTERN ARMY
BAGRATION
3-rd WESTERN ARMY
TORMASOV
Russian and Napoleon army locations
at the beginning of the war
in time of the war
Fortified camps of the Russian army at Drissa and Tarutino
Advance of Napoleon's army
Retreat of Russian army
Targets of the Russian army
Locations and targets of the partisan units
Additional reserves
KUTUZOV 29.8
Kutuzov appointment on the Russian Army Command
Location of the militia troops
Tarutino maneuver of the Russian army
Reference: Soviet Military Encyclopedia. v.6. Moscow: Voenizdat, 1978.

Wikipedia: Patriotic War of 1812
http://en.wikipedia.org/wiki/File:Patriotic_War_of_1812_ENG_map1.svg

Video: Napoleonic Wars in 8 Minutes

What can you learn from this map?



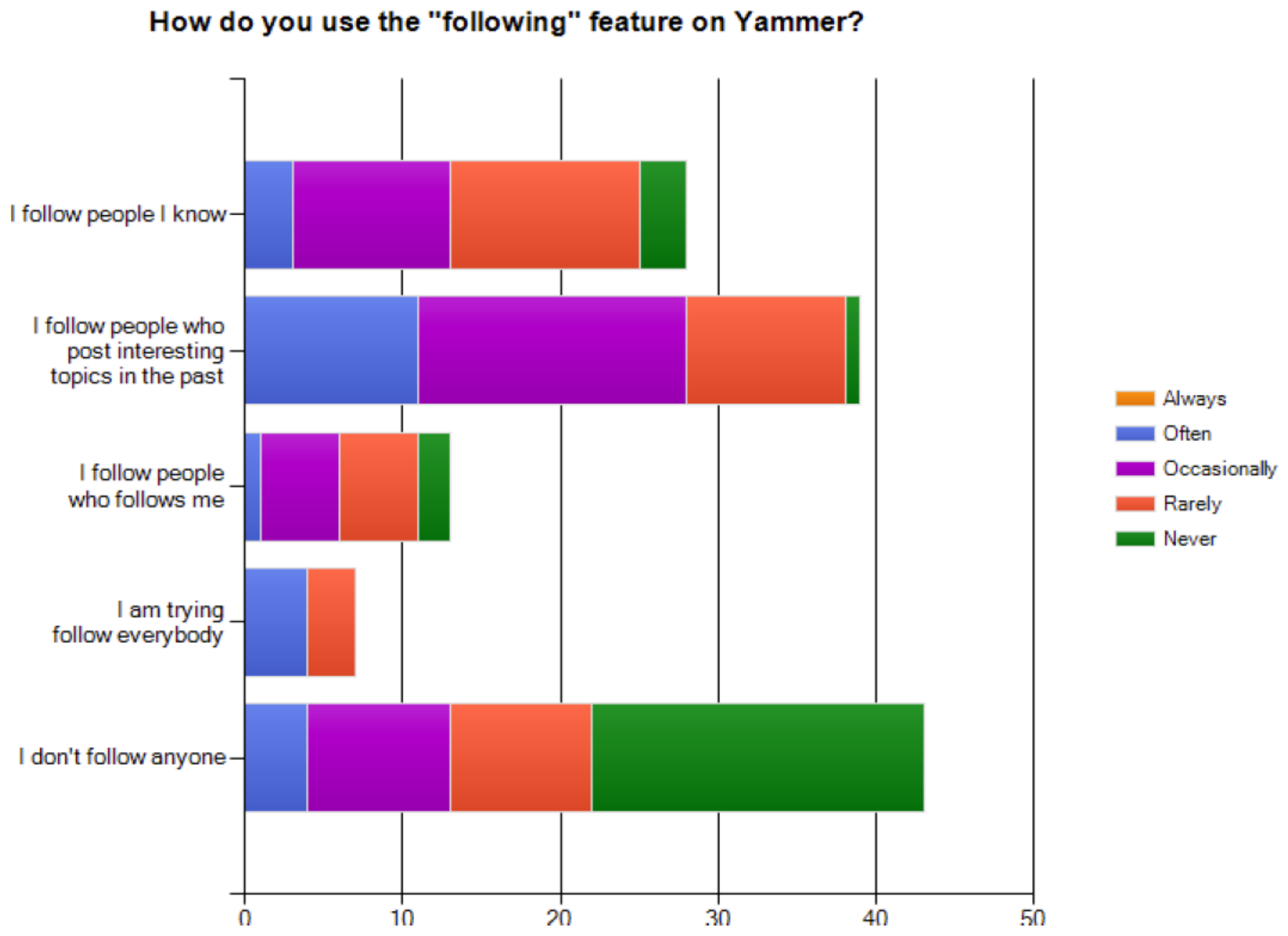
Survey data courtesy of
Alan McConchie
Visit www.popvssoda.com
to participate.

Map by Matthew T. Campbell
Spatial Graphics and Analysis Lab
Department of Cartography and Geography
East Central Univeristy (Oklahoma)
Map Template courtesy of www.mymaps.com

Respondents through
March 1, 2003

http://www.popvssoda.com/countystats/total-county.html

What makes a good chart?



This is from an academic conference paper.

What are the problems with this chart?

Zhang et al. (2010), "A case study of micro-blogging in the enterprise: use, value, and related issues," Proceedings of the 28th International Conference on Human Factors in Computing Systems.

Some basic principles (adapted from Tufte 2009)

1
- The chart should tell a story

2
- The chart should have graphical integrity

3
- The chart should minimize graphical complexity

Tufte’s fundamental principle:
Above all else show the data

Principle 1: The chart should tell a story

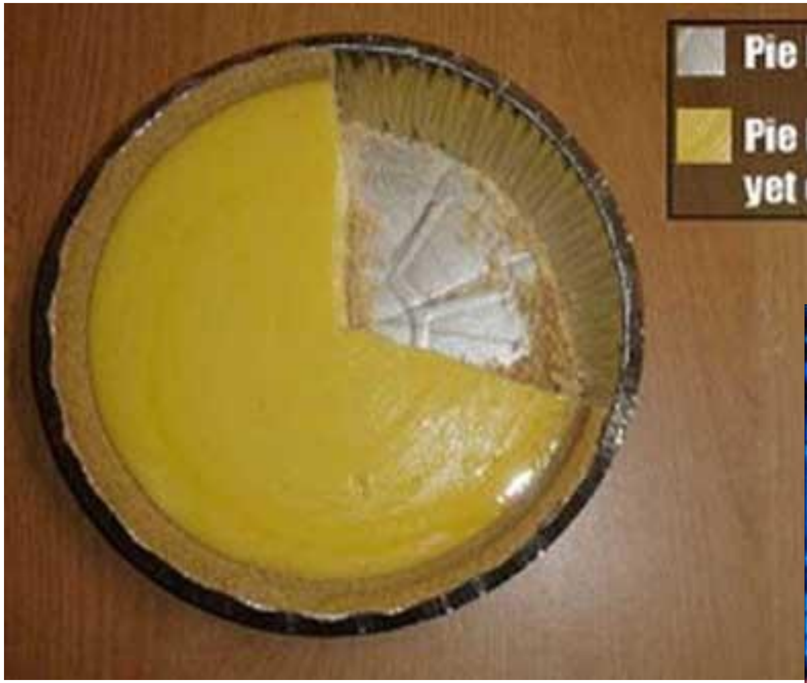
Graphics should be clear on their own

The depictions should enable meaningful comparison

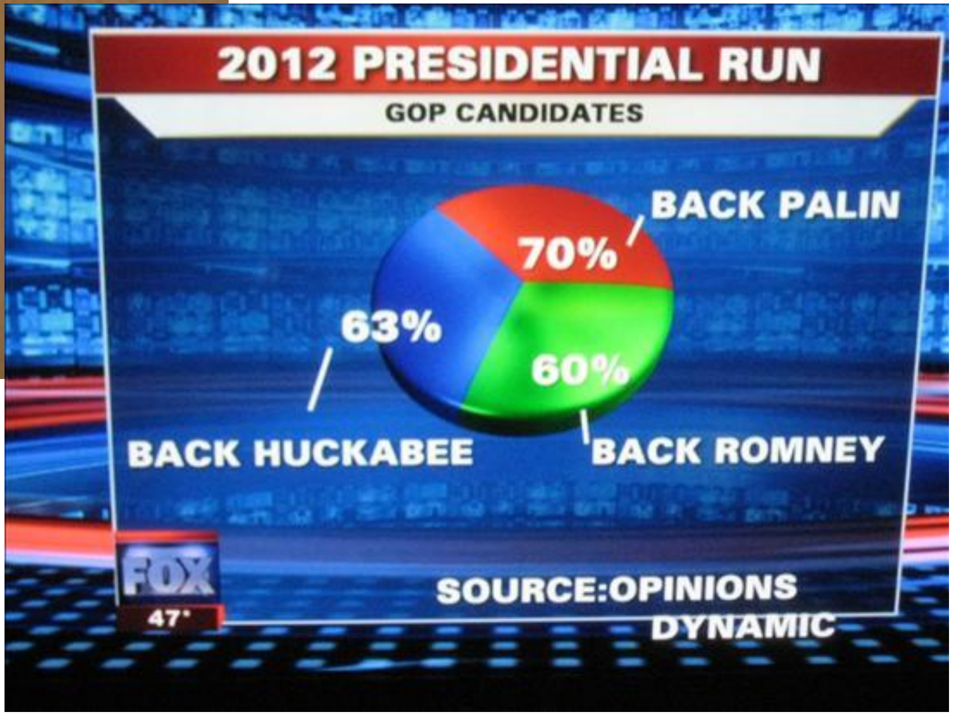
The chart should yield insight beyond the text

“If the statistics are boring, then you’ve got the wrong numbers.” (Tufte 2009)

Do these tell a story?

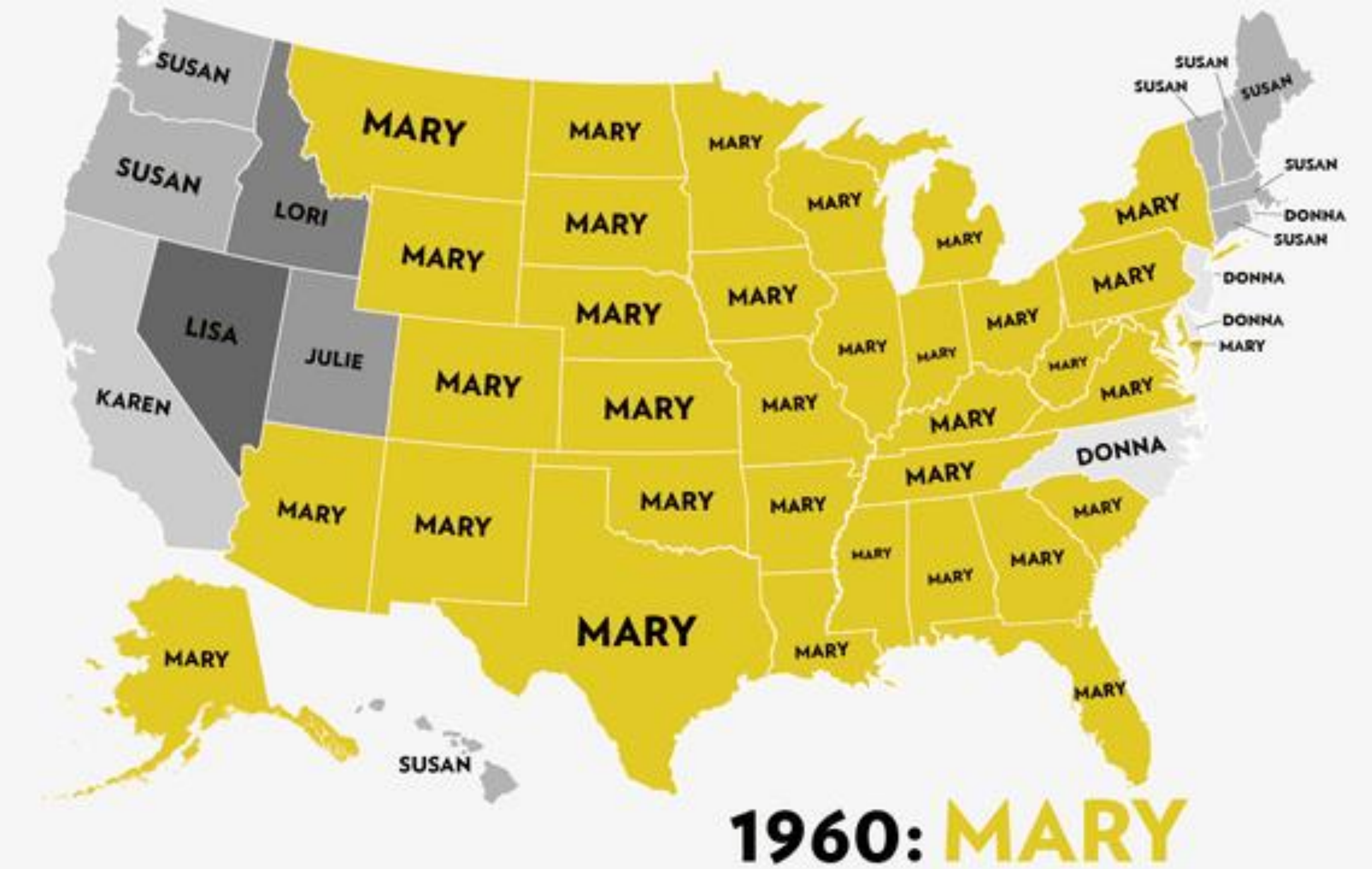


http://www.evl.uic.edu/aej/491/week03.html



http://flowingdata.com/2009/11/26/fox-news-makes-the-best-pie-chart-ever/

Most Popular Girl Names in Map



1960: MARY

by Reuben Fischer-Baum (http://jezebel.com/map-sixty-years-of-the-most-popular-names-for-girls-s-1443501909)

Principle 2: The chart should have graphical integrity

- Basically, it shouldn't "lie" (mislead the reader)

- Tufte's "Lie Factor":
  - $Lie\ Factor = \frac{size\ of\ effect\ shown\ in\ graphic}{size\ of\ effect\ in\ data}$

Should be ~ 1

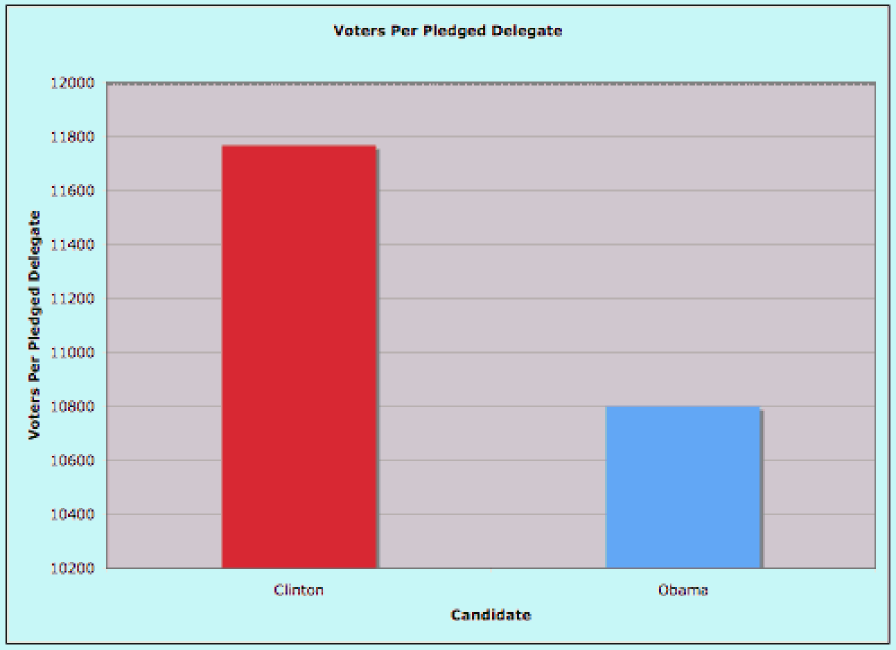
< 1 = understated effect

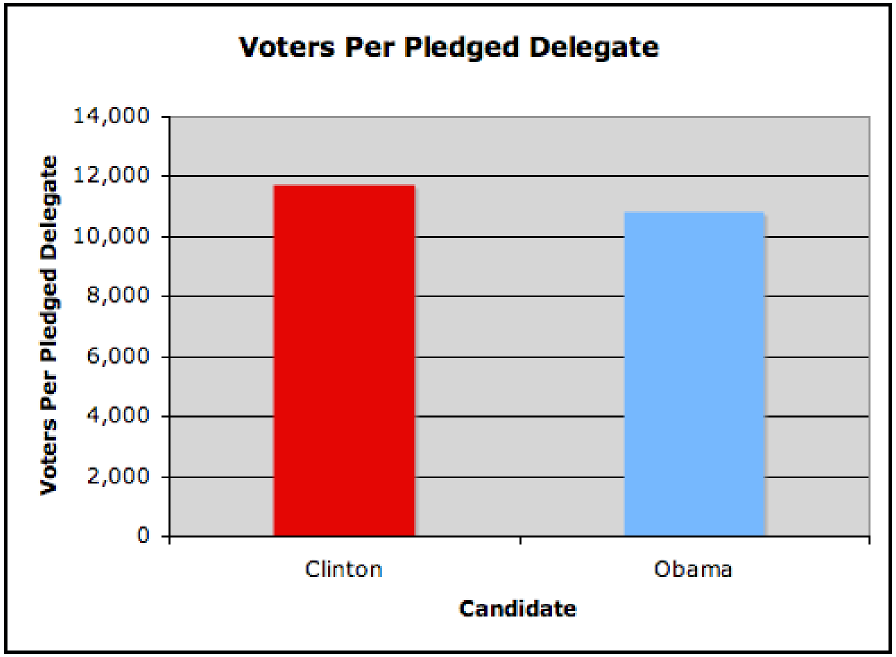
> 1 = exaggerated effect

How is this deceptive?

The original graphic from Real Clear Politics, 2008.
(Look at the y-axis)

The adjusted graphic.

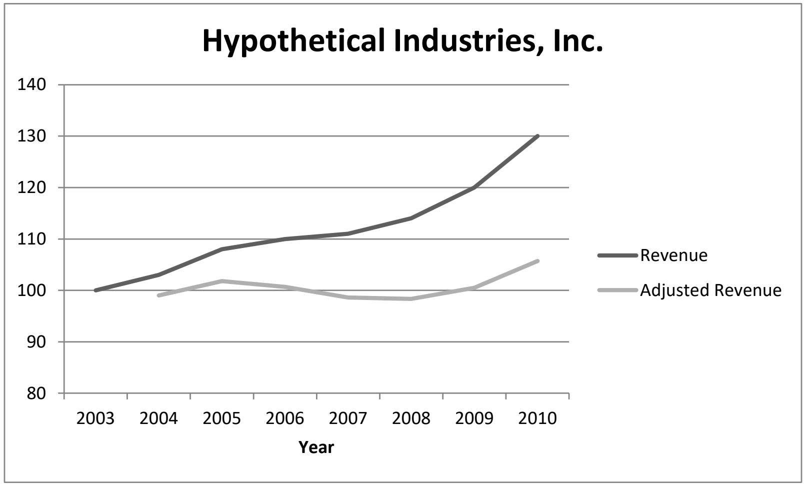




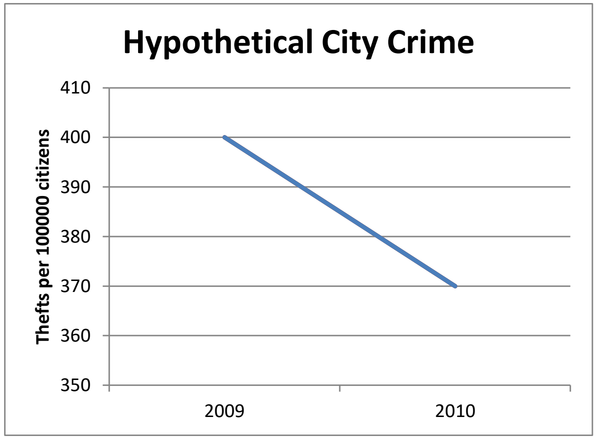
http://20bits.com/articles/politics-and-tuftes-lie-factor/

Other tips to avoid “lying”

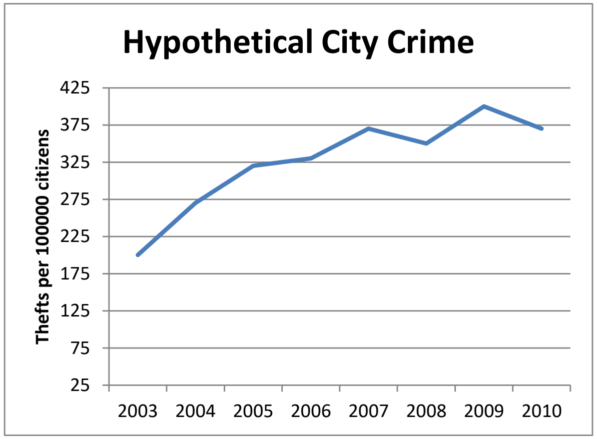
Adjust for inflation



Make sure the context is presented

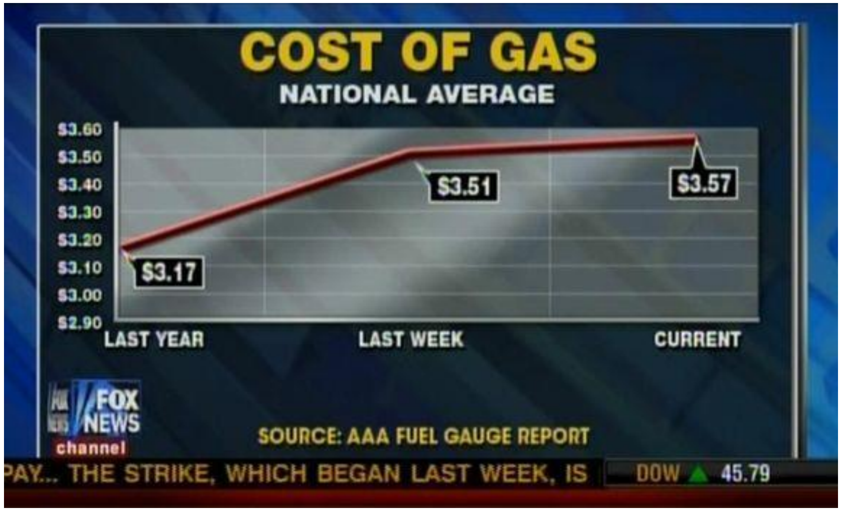


vs.

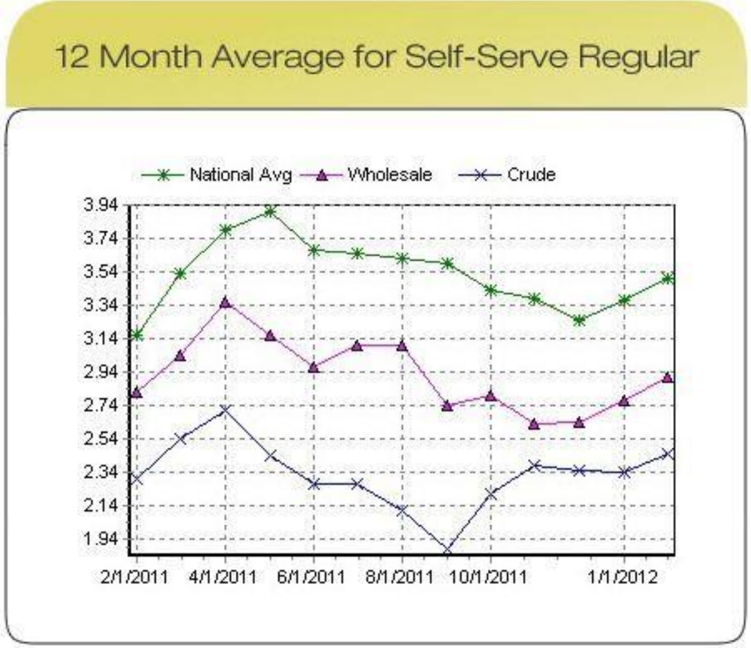


Present data in context

The original graphic from Fox News, Feb 2012.



In Reality…



http://mediamatters.org/research/2012/10/01/a-history-of-dishonest-fox-charts/190225

Principle 3: The chart should minimize graphical complexity

Generally, the simpler the better…

Key concepts

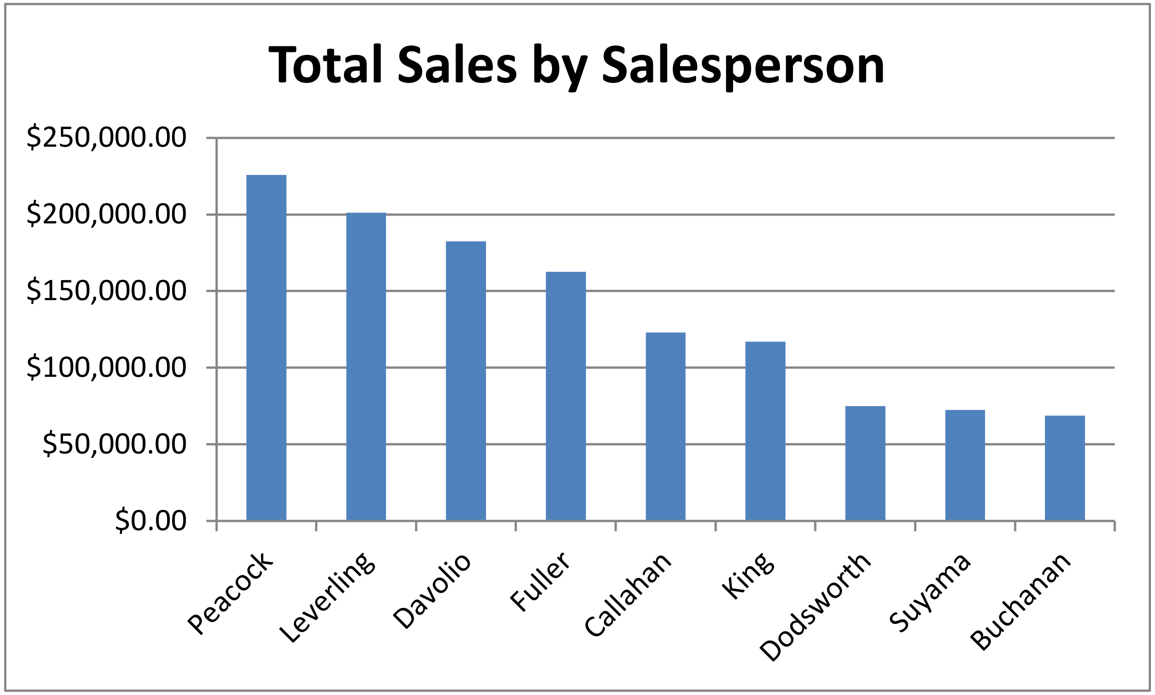
Sometimes a table is better

Data-ink

Chartjunk

When a table is better than a chart

For a few data points, a table can do just as well…

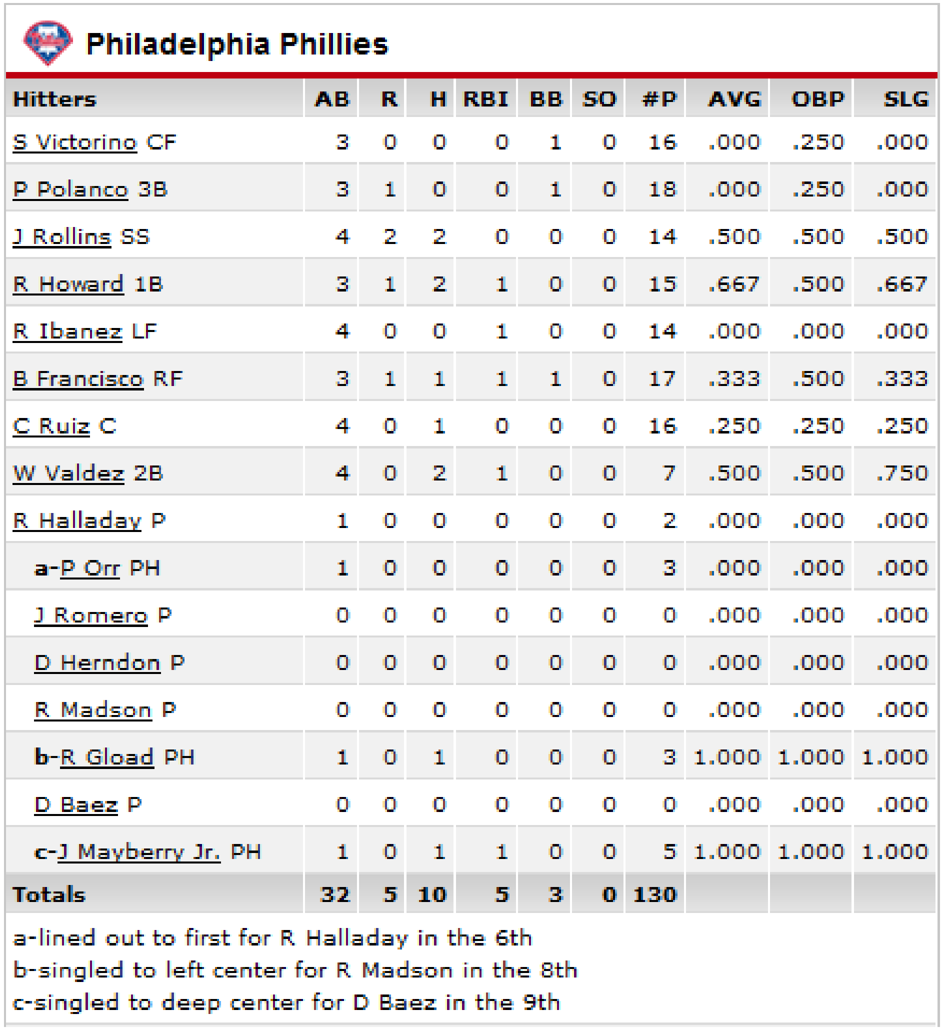


Salesperson	Total Sales
Peacock	$225,763.68
Leverling	$201,196.27
Davolio	$182,500.09
Fuller	$162,503.78
Callahan	$123,032.67
King	$116,962.99
Dodsworth	$75,048.04
Suyama	$72,527.63
Buchanan	$68,792.25

The table carries more information in less space and is more precise.

The Ultimate Table: The Box Score

- Large amount of information in a very small space
- So why does this work?
    - Depends on the reader’s knowledge of the data



Philadelphia Phillies

Hitters	AB	R	H	RBI	BB	SO	#P	AVG	OBP	SLG
S Victorino CF	3	0	0	0	1	0	16	.000	.250	.000
P Polanco 3B	3	1	0	0	1	0	18	.000	.250	.000
J Rollins SS	4	2	2	0	0	0	14	.500	.500	.500
R Howard 1B	3	1	2	1	0	0	15	.667	.500	.667
R Ibanez LF	4	0	0	1	0	0	14	.000	.000	.000
B Francisco RF	3	1	1	1	1	0	17	.333	.500	.333
C Ruiz C	4	0	1	0	0	0	16	.250	.250	.250
W Valdez 2B	4	0	2	1	0	0	7	.500	.500	.750
R Halladay P	1	0	0	0	0	0	2	.000	.000	.000
a-P Orr PH	1	0	0	0	0	0	3	.000	.000	.000
J Romero P	0	0	0	0	0	0	0	.000	.000	.000
D Herndon P	0	0	0	0	0	0	0	.000	.000	.000
R Madson P	0	0	0	0	0	0	0	.000	.000	.000
b-R Gload PH	1	0	1	0	0	0	3	1.000	1.000	1.000
D Baez P	0	0	0	0	0	0	0	.000	.000	.000
c-J Mayberry Jr. PH	1	0	1	1	0	0	5	1.000	1.000	1.000
Totals	**32**	**5**	**10**	**5**	**3**	**0**	**130**			

a-lined out to first for R Halladay in the 6th
b-singled to left center for R Madson in the 8th
c-singled to deep center for D Baez in the 9th

Data Ink

- The amount of “ink” devoted to data in a chart
- Tufte’s Data-Ink ratio:
  - $Data - ink\ ratio = \dfrac{data{-}ink}{total\ ink\ used\ in\ graphic}$

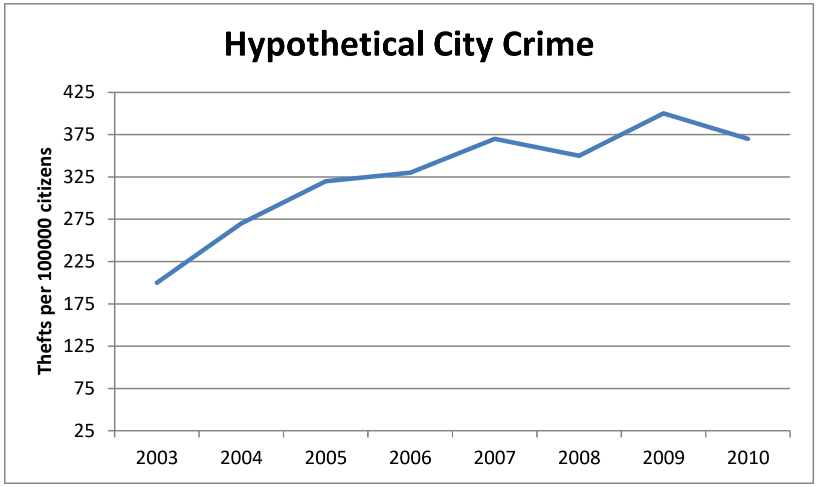
Should be ~ 1

< 1 = more non-data related ink in graphic

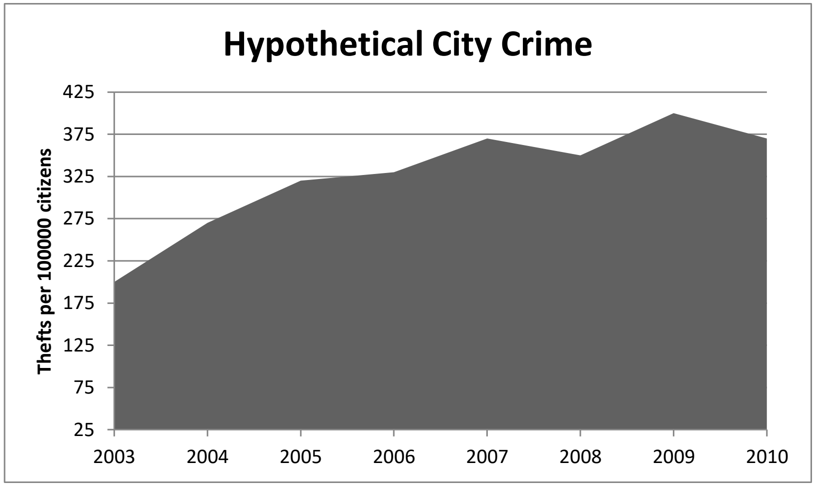
= 1 implies all ink devoted to data

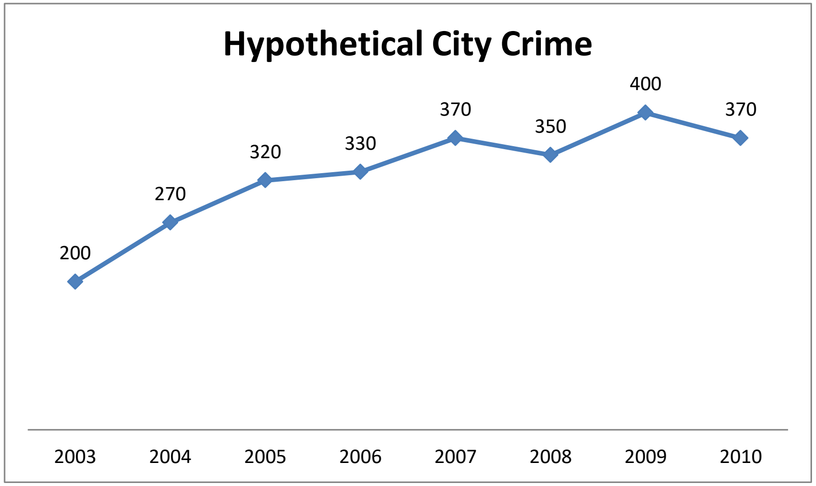
Tufte’s principle:
Erase ink whenever possible

Being conscious of data ink



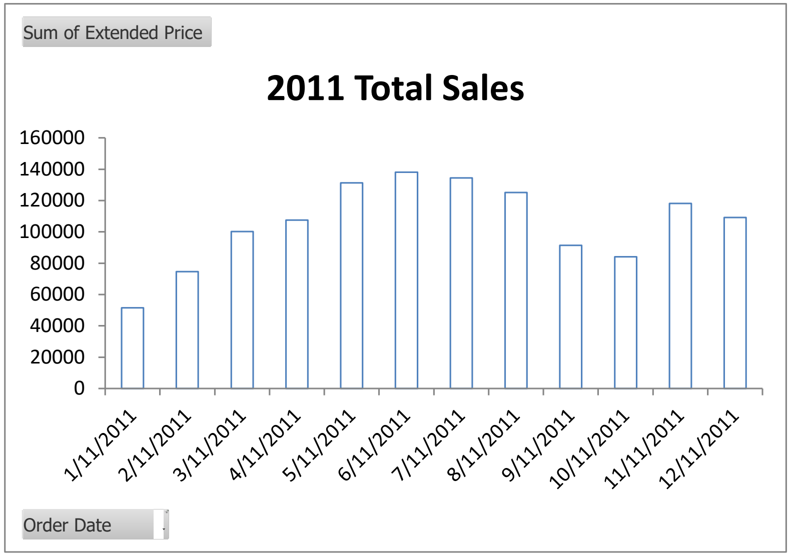
Lower data-ink ratio
(worse)

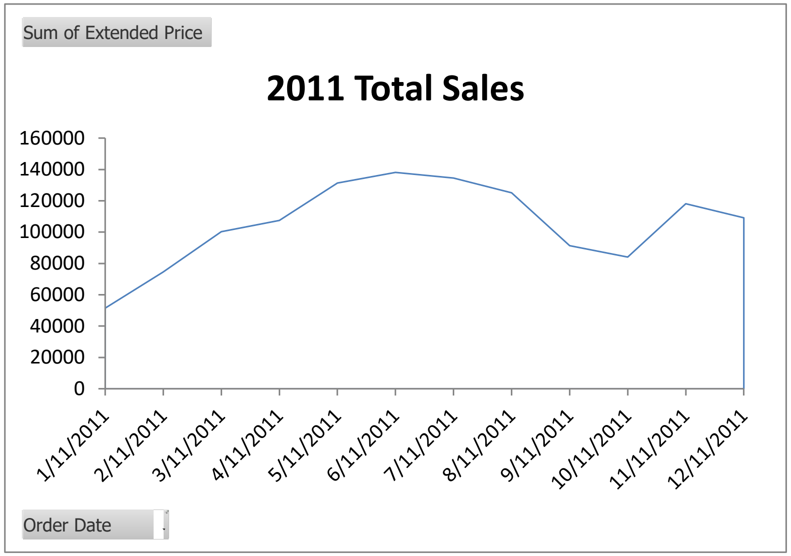




Higher data-ink ratio
(better)

What makes a good chart?



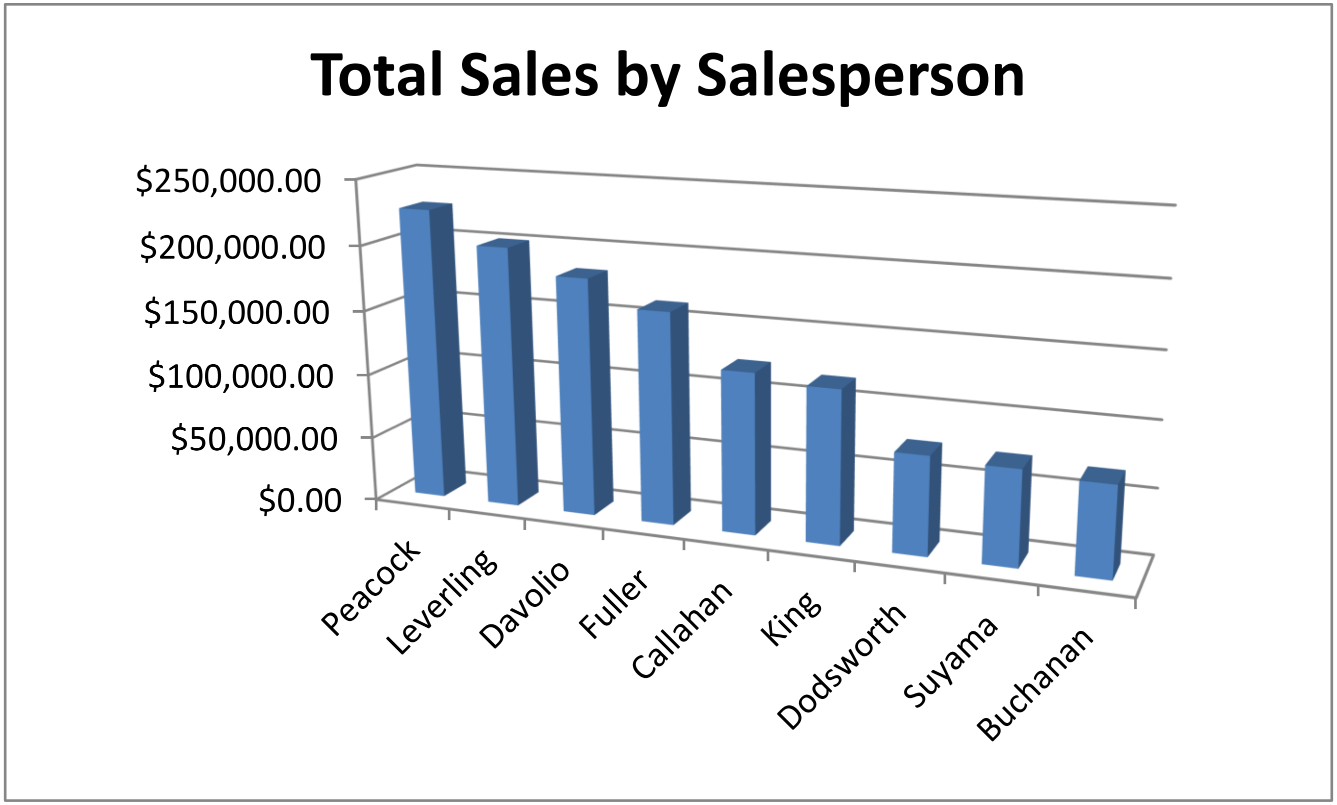


Sometimes it’s really a matter of preference.

These both minimize data ink.

Why isn’t a table better here?

3-D Charts



Evaluate this from a data-ink perspective.
How does it affect the clarity of the chart?

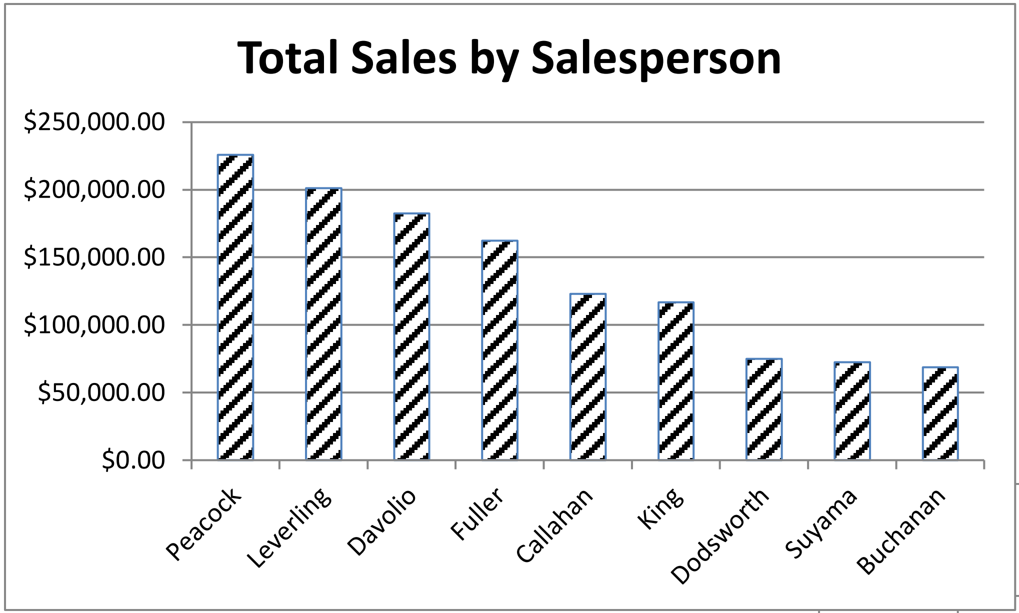
Chartjunk: Data Ink “gone wild”

Unnecessary visual clutter that doesn’t provide additional insight

Distraction from the story the chart is supposed to convey

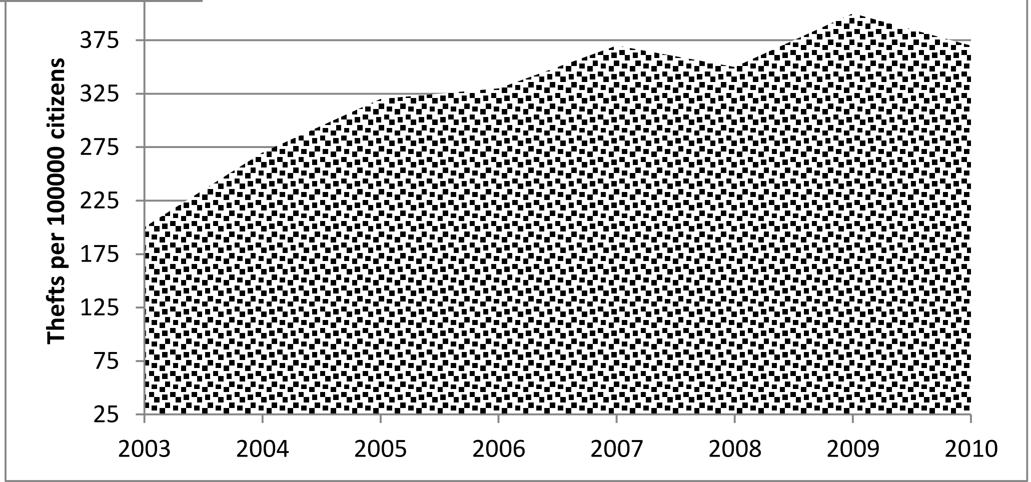
When the data-ink ratio is low, chartjunk is likely to be high

Example: Moiré effects (Tufte 2009)

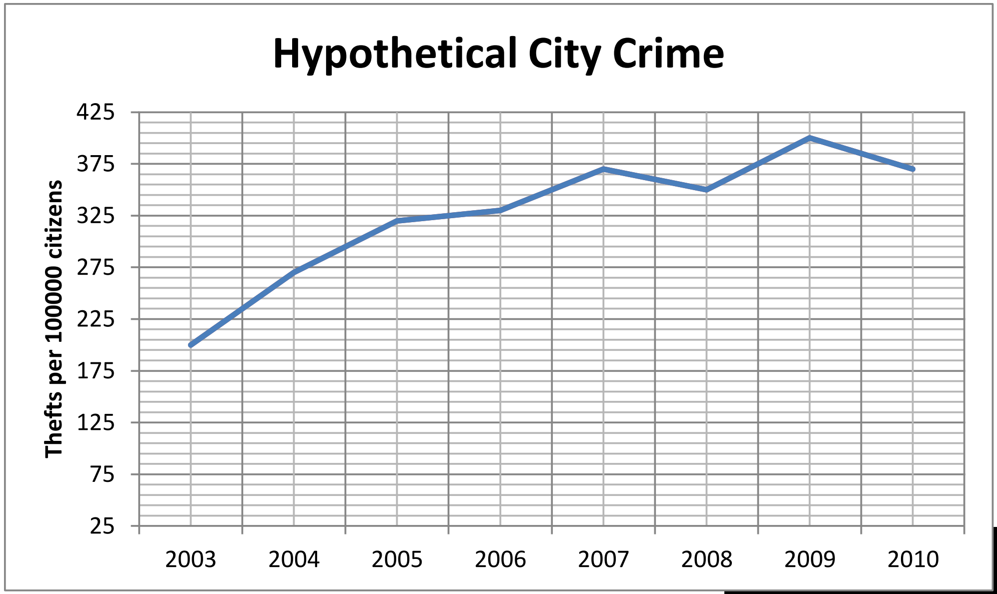


Creates illusion of movement

Stands out, in a bad way

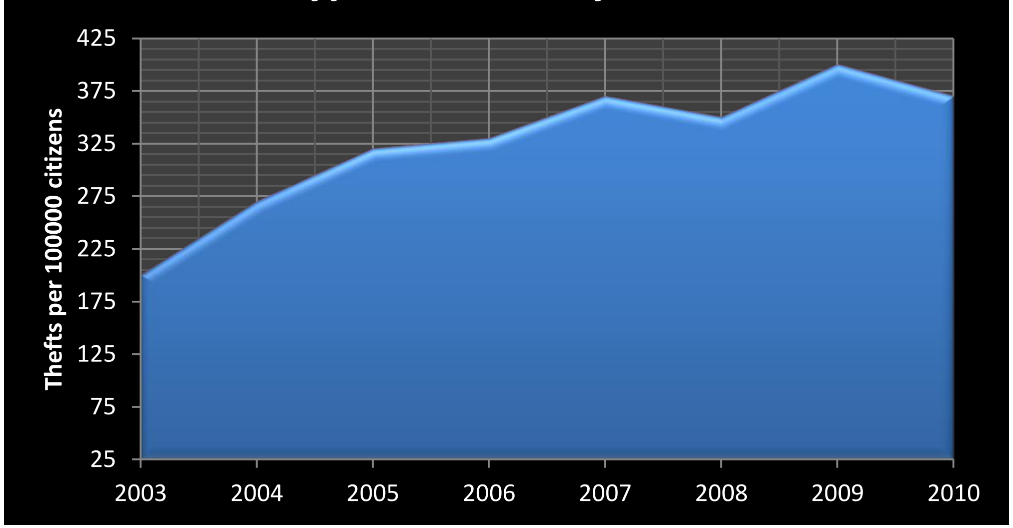


Example: The Grid

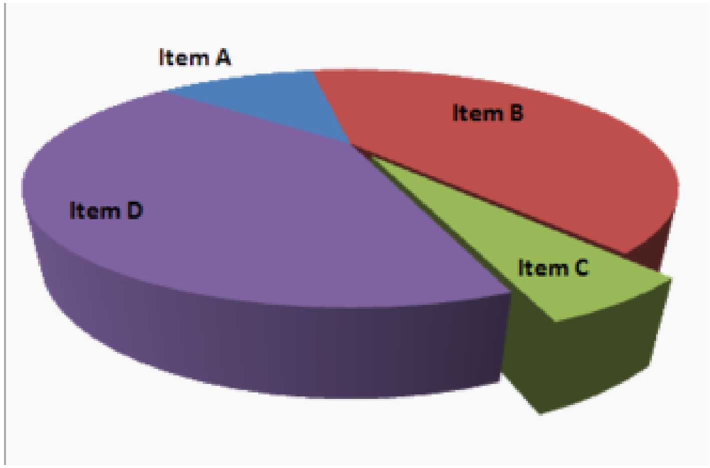


Why are these examples of chartjunk?

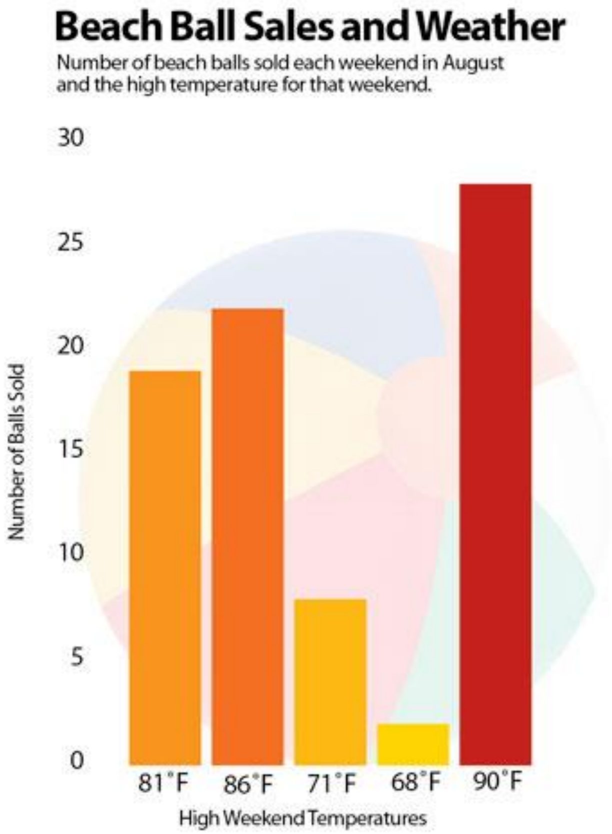
What could you do to remedy it?



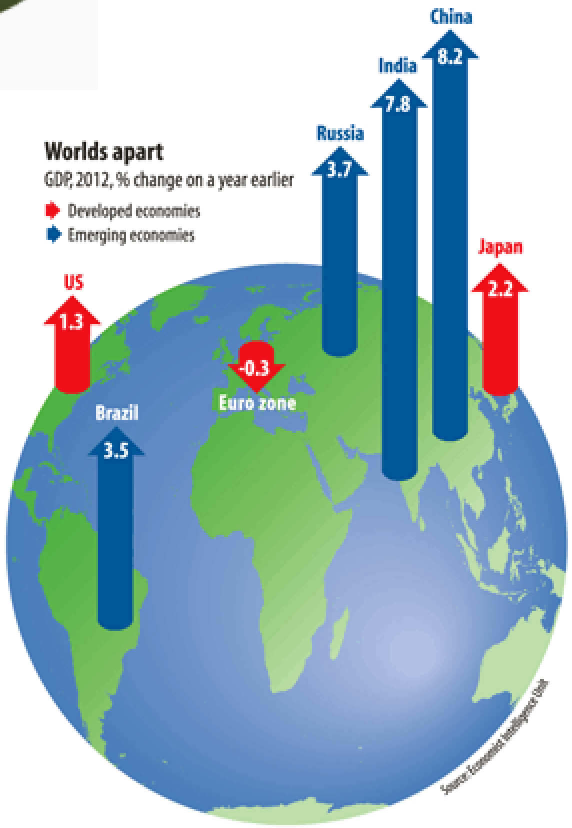
Review: What do you think of these?



https://www.boundless.com/statistics/frequency-distributions/frequency-distributions-for-qualitative-data/interpreting-distributions-constructed-by-others/images/3-d-pie-chart/

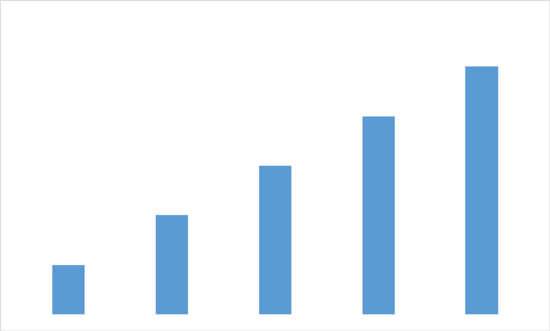


http://images.macworld.com/images/howto/graphics/134708-create-charts-good_376.jpg

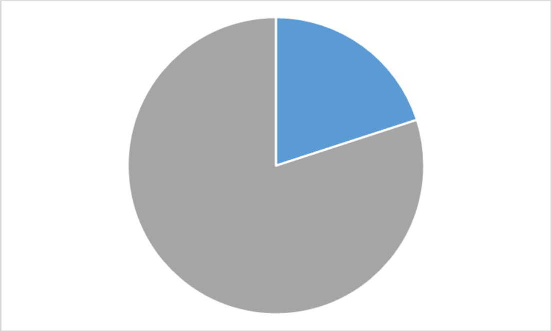


http://www.economist.com/node/21537909

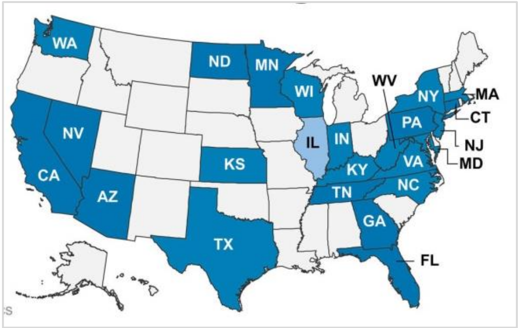
Common Chart Types



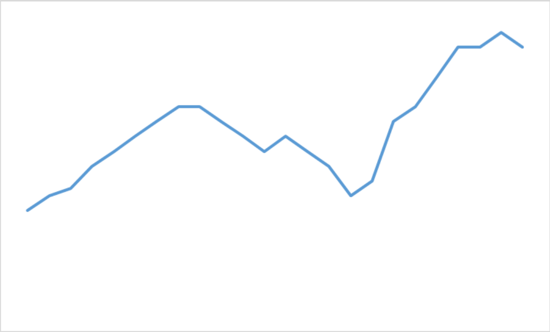
Bars
(For Comparison)



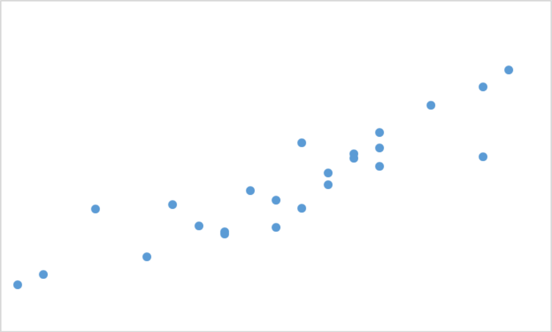
Pie
(For Composition)



Map
(For Spatial Comparison)



Line
(For Evolution)



Scatterplot
(Relationship)

Some Visualization Tools

- Excel (as always)
- R, Stata, Tableau, SAS (useful for Statistical Plots).
- Google Charts, FusionCharts (simple graphs as well as maps)
- Piktochart (infographics)
- Adobe Photoshop, Illustrator, etc (for graphical design)

Summary

- Use data visualization principles to assess a visualization
  - Tell a story
  - Graphical integrity (lie factor)
  - Minimize graphical complexity (data ink, chartjunk)
- Explain how a visualization can be improved based on those principles
- Types of visualization