Session 2: Visualizing data

Stats 60/Psych 10
Ismael Lemhadri
Summer 2020

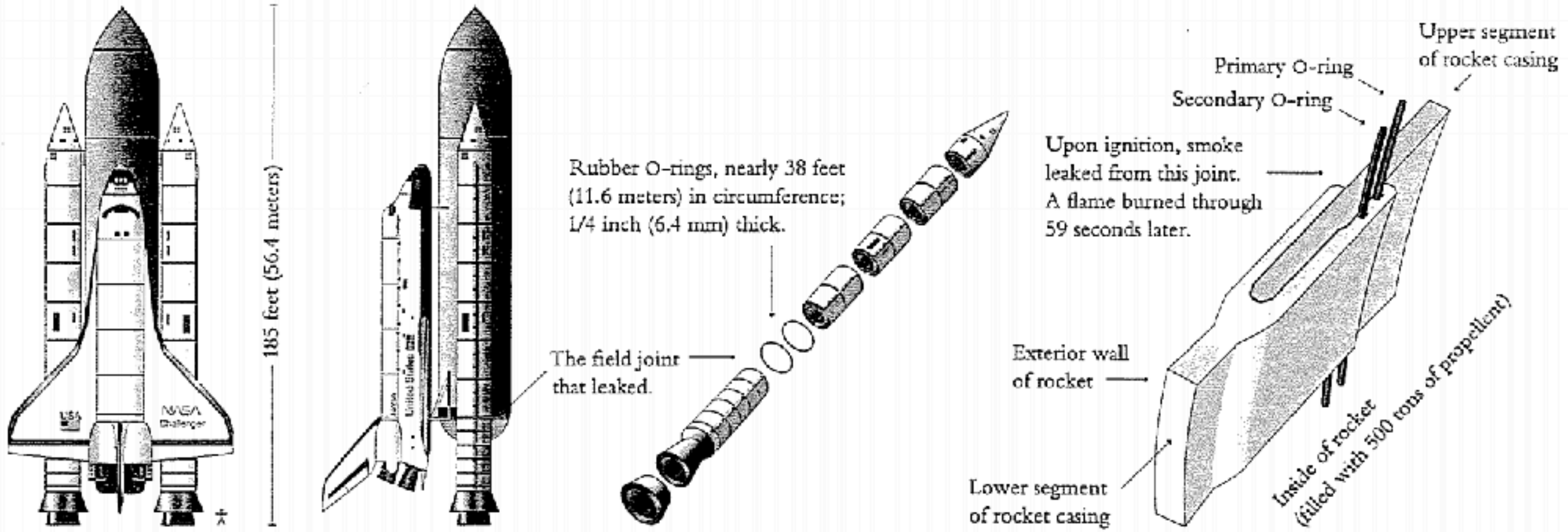
This time

- Visualizing data
  - How to spot bad graphs
  - How to create good graphs

How better data visualization could have saved 7 lives

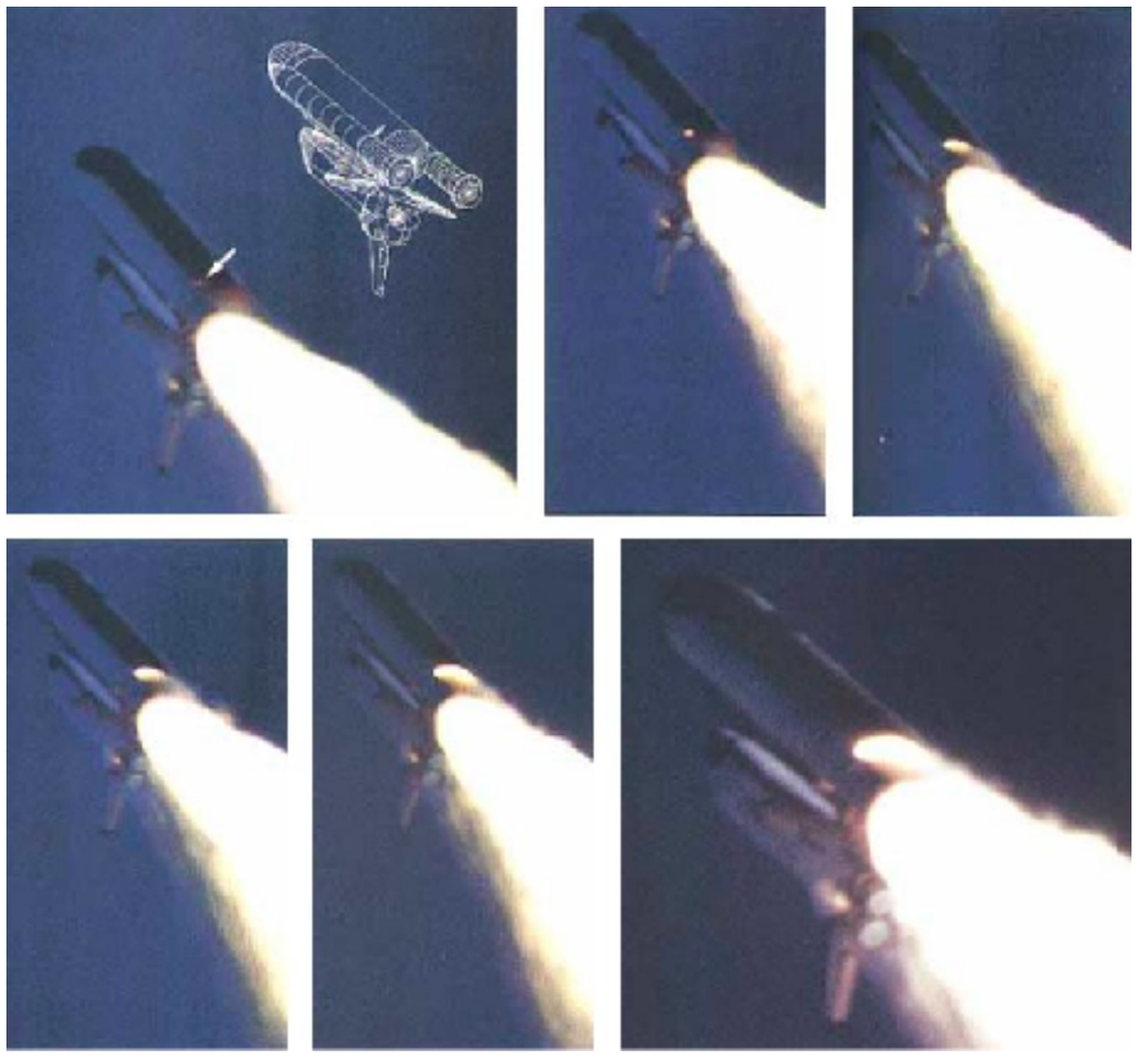
January 28, 1986

What happened?



The shuttle consists of an *orbiter* (which carries the crew and has powerful engines in the back), a large liquid-fuel *tank* for the orbiter engines, and 2 solid-fuel *booster rockets* mounted on the sides of the central tank. Segments of the booster rockets are shipped to the launch site, where they are assembled to make the solid-fuel rockets. Where these segments mate, each joint is sealed by two rubber O-rings as shown above. In the case of the Challenger accident, one of these joints leaked, and a torch-like flame burned through the side of the booster rocket.

Tufte, 1997



https://www.slideshare.net/catalyst00/truth-lies-and-orings-inside-the-space-shuttle-challenger-disaster

http://www.aerospaceweb.org/question/investigations/q0122.shtml

What does this have to do with data visualization?

- Temperatures were forecast to be very cold on Jan 28
- Engineers from the rocket contractor Morton Thiokol presented 13 charts in an attempt to convince NASA to postpone the launch due to concerns about the O-rings failing at low temperature
- They failed

Ineffective presentation of data

BLOW BY HISTORY

SRM-15 WORST BLOW-BY

- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY

- 2 CASE JOINTS (30-40°)

SRM-13A, 15, 16A, 18, 23A 24A

- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES-F)

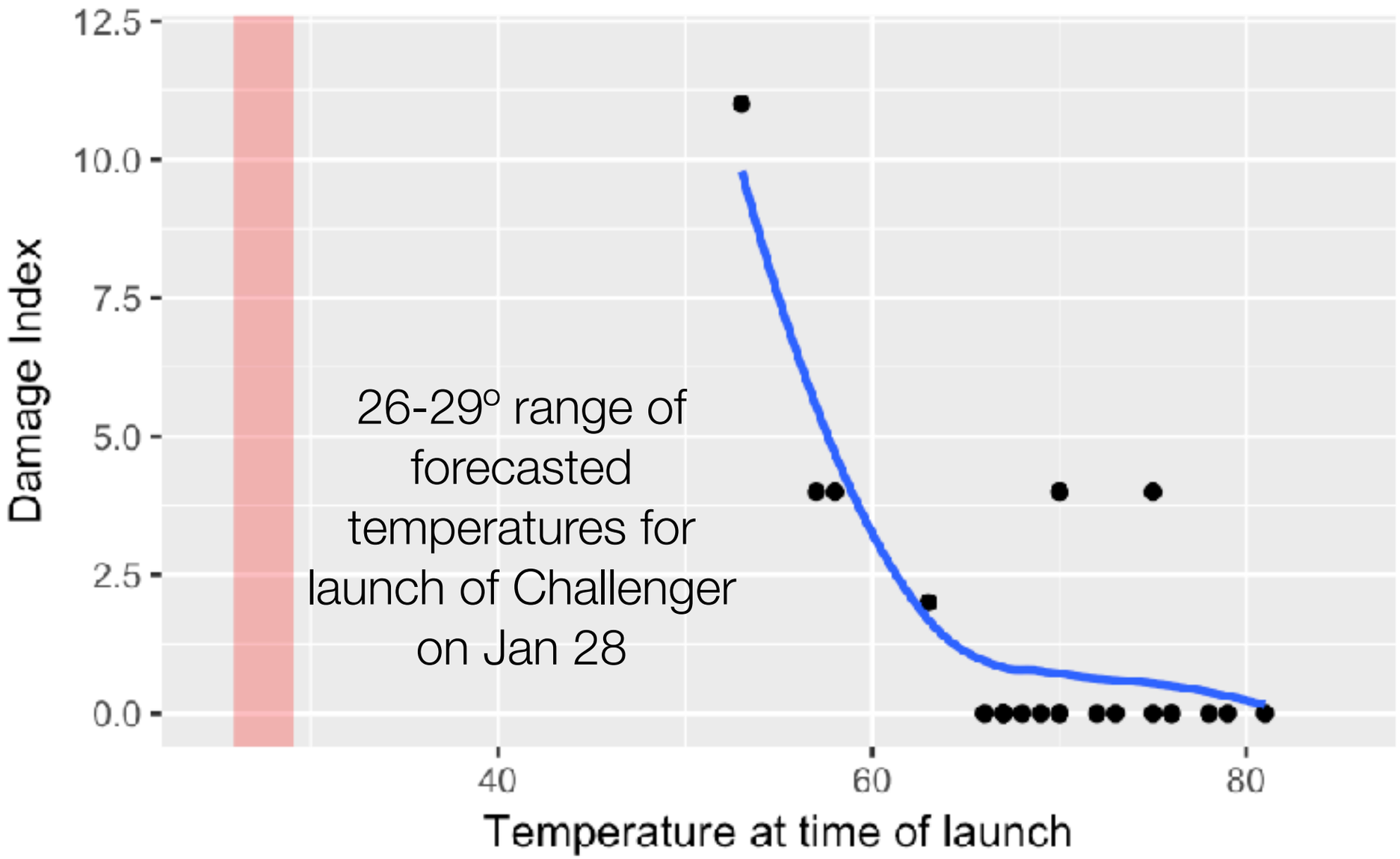
MOTOR	MGT	AMB	O-RING	WIND
DM-4	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

A more effective summary of the data

Flight	Date	Temperature °F	Erosion incidents	Blow-by incidents	Damage index	Comments
51-C	01.24.85	53°	3	2	11	Most erosion any flight; blow-by; back-up rings heated.
41-B	02.03.84	57°	1		4	Deep, extensive erosion.
61-C	01.12.86	58°	1		4	O-ring erosion on launch two weeks before Challenger.
41-C	04.06.84	63°	1		2	O-rings showed signs of heating, but no damage.
1	04.12.81	66°			0	Coolest (66°) launch without O-ring problems.
6	04.04.83	67°			0	
51-A	11.08.84	67°			0	
51-D	04.12.85	67°			0	
5	11.11.82	68°			0	
3	03.22.82	69°			0	
2	11.12.81	70°	1		4	Extent of erosion not fully known.
9	11.28.83	70°			0	
41-D	08.30.84	70°	1		4	
51-G	06.17.85	70°			0	
7	06.18.83	72°			0	
8	08.30.83	73°			0	
51-B	04.29.85	75°			0	
61-A	10.30.85	75°		2	4	No erosion. Soot found behind two primary O-rings.
51-I	08.27.85	76°			0	
61-B	11.26.85	76°			0	
41-G	10.05.84	78°			0	
51-J	10.03.85	79°			0	
4	06.27.82	80°			?	O-ring condition unknown; rocket casing lost at sea.
51-F	07.29.85	81°			0	

Tufte, 1997

An even more effective visualization of the data



What are the two important takeaway messages?

adapted from Tufte, 1997

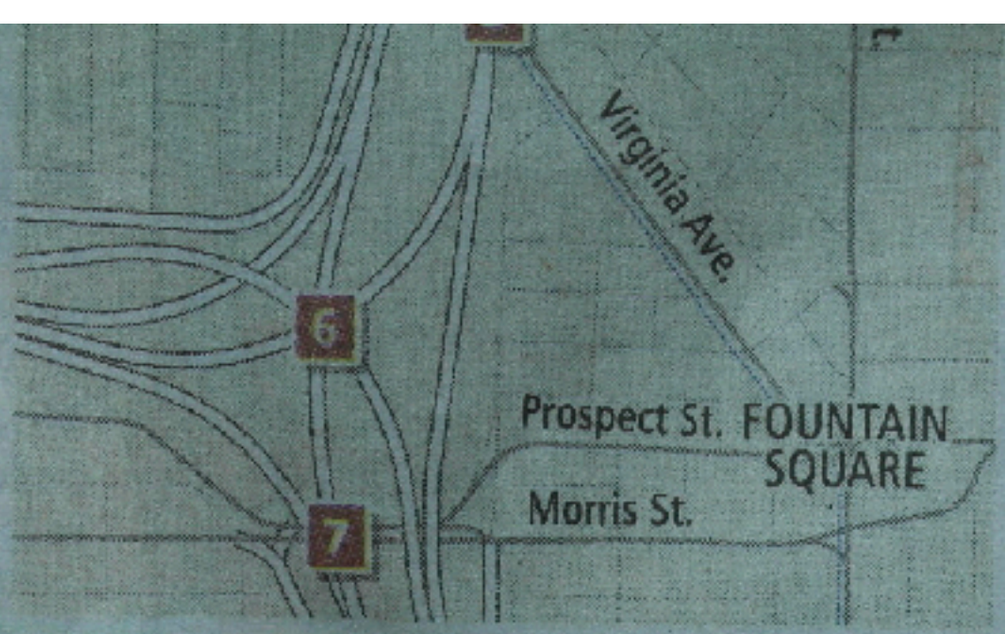
It’s very easy to find bad graphs

in late August.
9,000 drivers travel
areas daily, accord-
T numbers from 2011,
ent year available.
uncement follows a
. 22, in which an over-
struck the Virginia
dge, shutting down
I-65 and eastbound
rgency repairs over
d.
nas recorded more
nilar incidents since

to emphasize that the
structurally safe,"
sson, INDOT deputy
er. "Our goal is to de-
robability of bridge

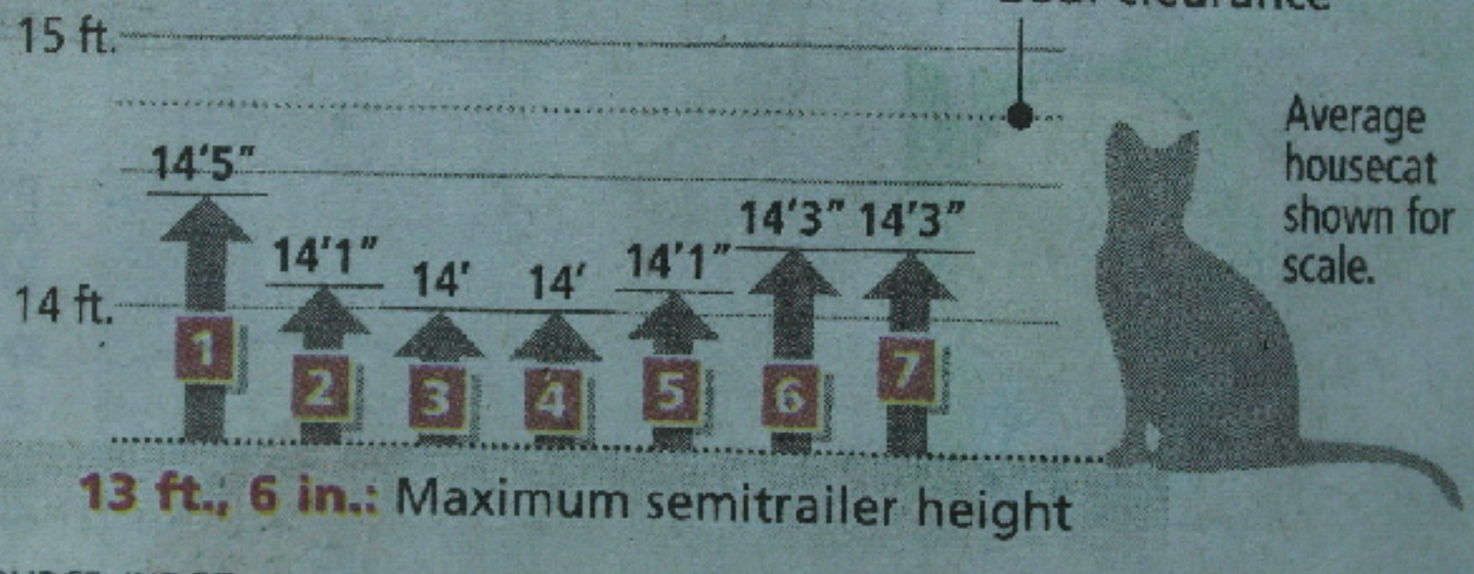
e closure, Wingfield
vement will be low-

» See SPLIT, Page B3



6. I-65 southbound ramp bridge to I-70 westbound
7. Morris Street bridge over I-65 southbound

How heights of the bridges above compare with the maximum height for semitrailers:



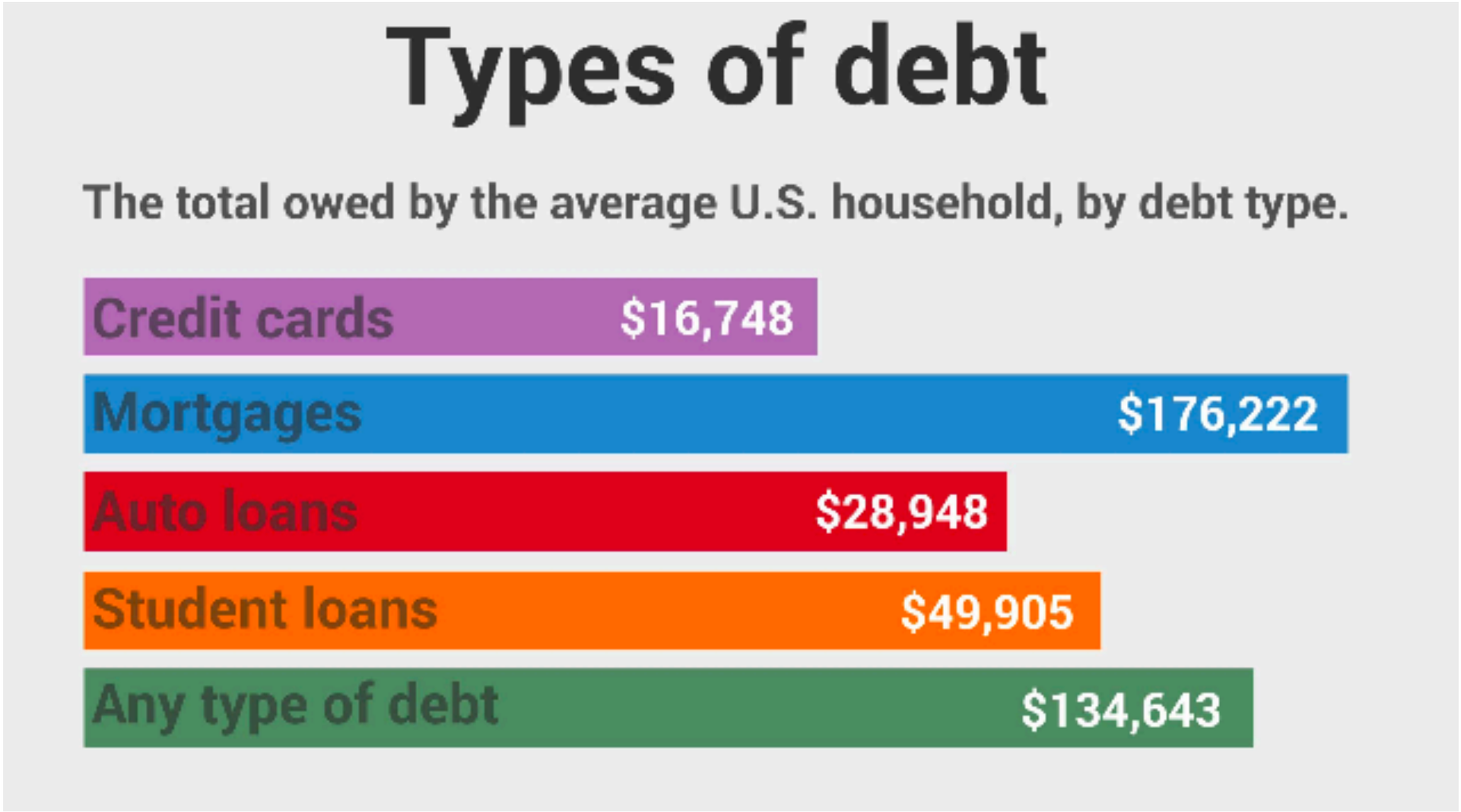
14 ft., 9 in.: Goal clearance

13 ft., 6 in.: Maximum semitrailer height

SOURCE: INDOT

STEPHEN J. BEARD / THE STAR

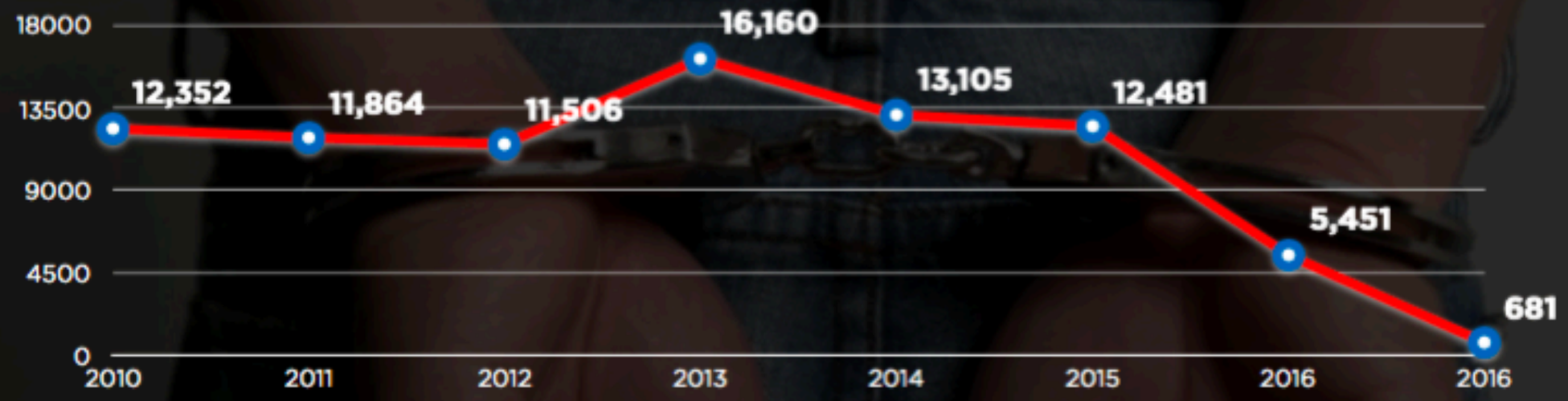
https://flowingdata.com/2013/07/15/open-thread-what-is-wrong-with-these-charts/


Types of debt
The total owed by the average U.S. household, by debt type.
Credit cards
$16,748
Mortgages
$176,222
Auto loans
$28,948
Student loans
$49,905
Any type of debt
$134,643

http://viz.wtf/

NO to DRUGS | YES to PEACE & ORDER

NUMBER OF MURDER AND HOMICIDE CASES REPORTED
(2010- AUG 3, 2016)



	2010	2011	2012	2013	2014	2015	2016 JAN-JUNE	2016 JULY 1 - AUG3
Total Murder + Homicide	12, 352	11,864	11,506	16,160	13,105	12,481	5,451	681
Average/ Daily	34	32	32	44	36	34	30	20

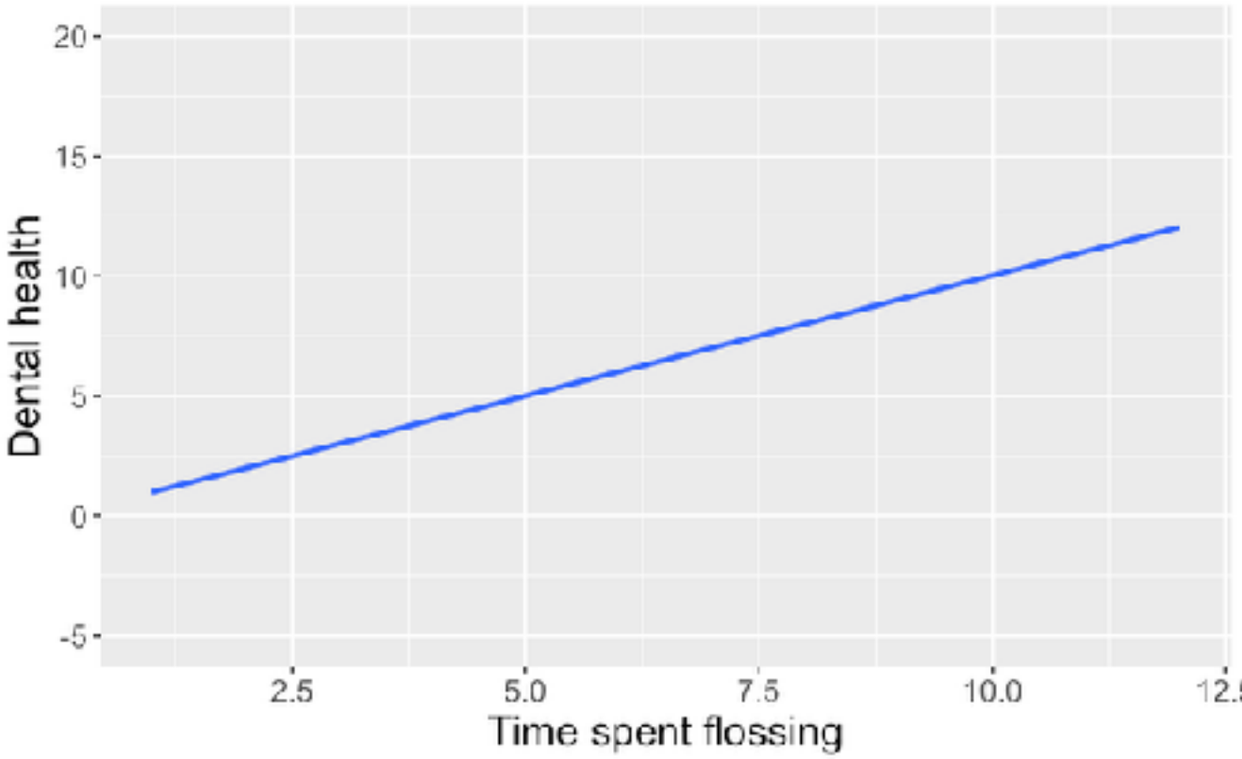
Source: PNP Directorate for investigative and detective management

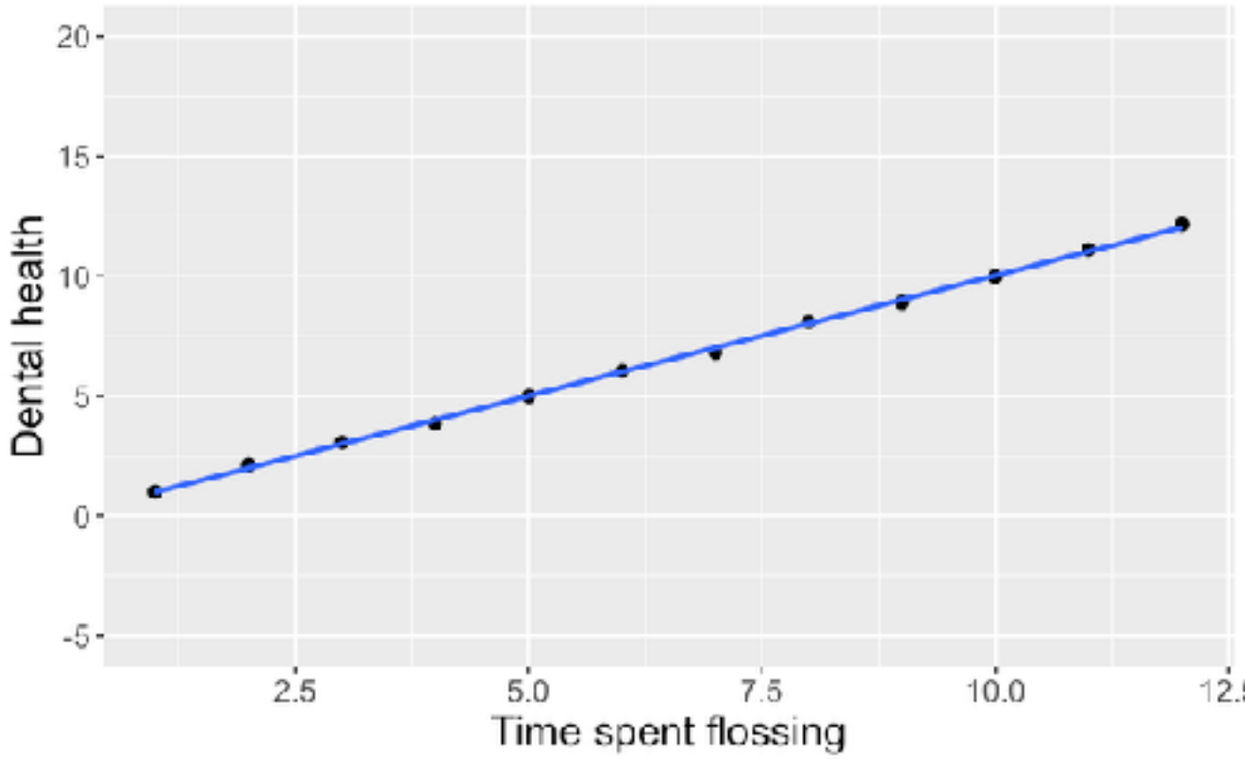
http://viz.wtf/

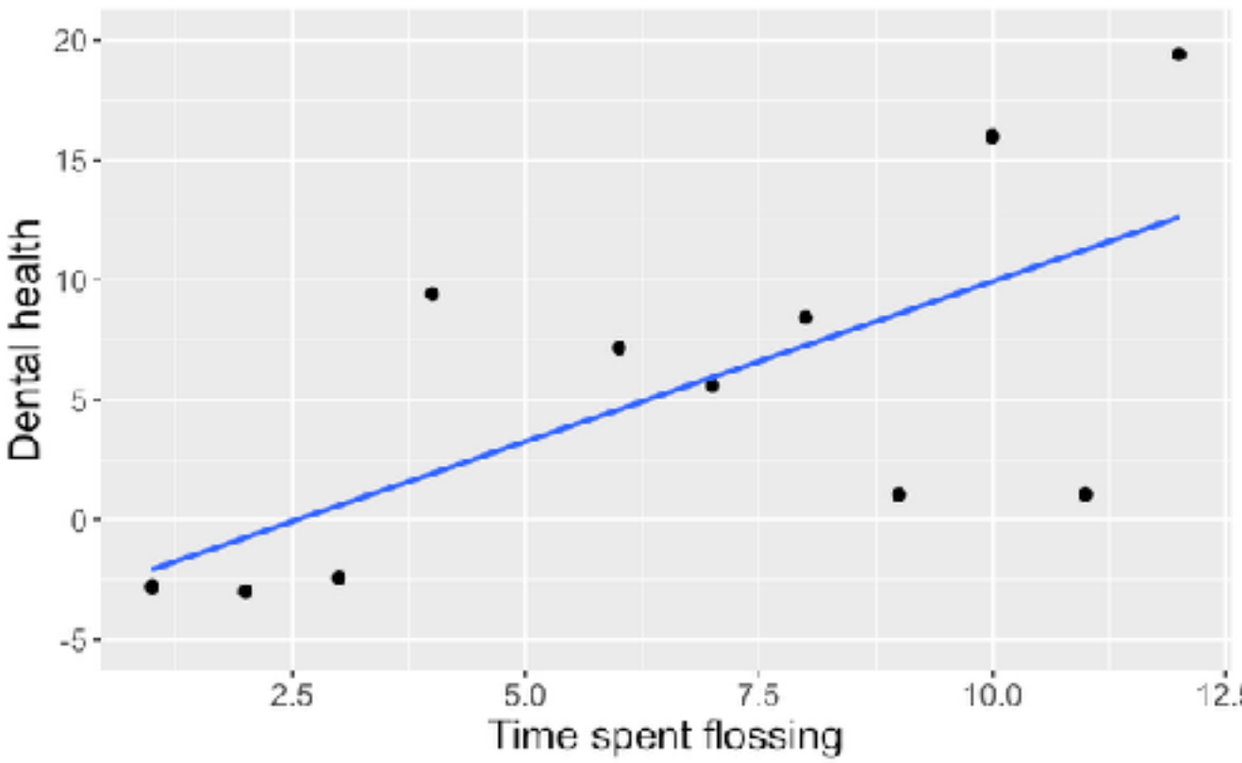
Principles of good visualizations

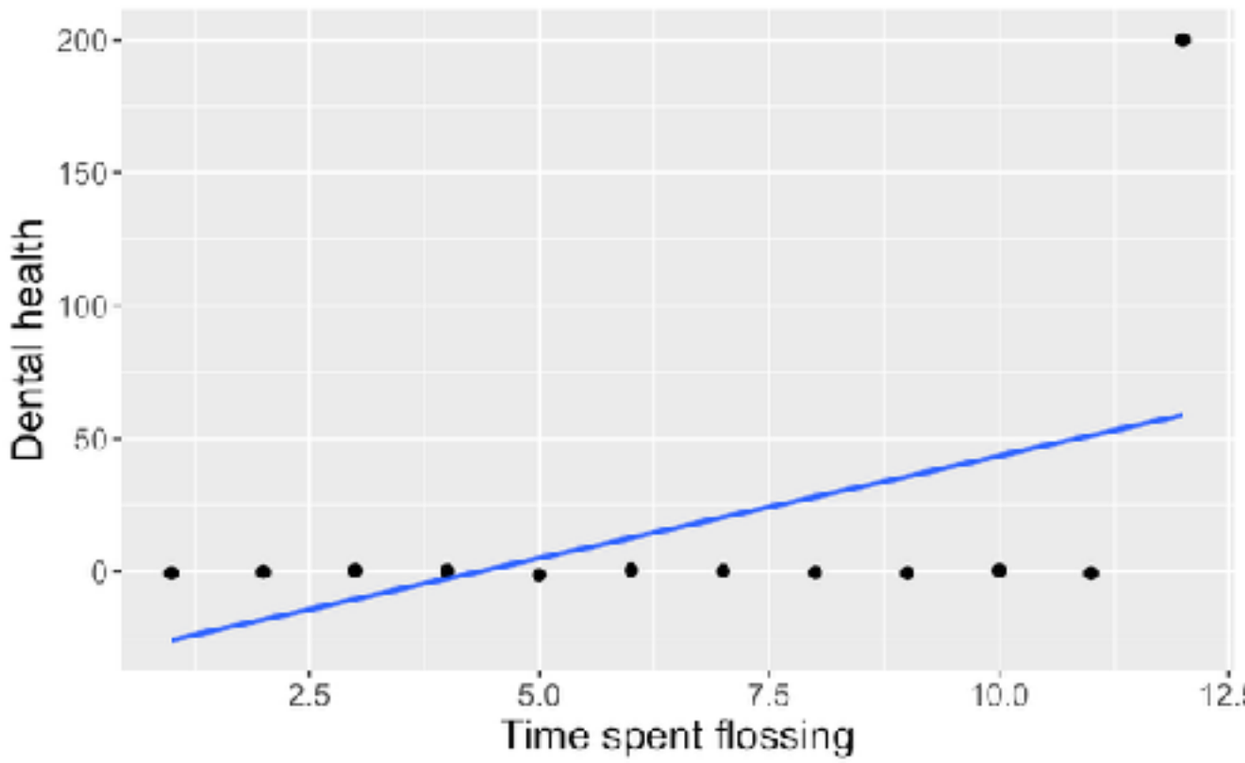
1. Show the data and make them stand out
   - Avoid clutter and chartjunk
2. Avoid distorting the data
   - Use proper scales
3. Keep human limitations in mind
4. Reveal the underlying message of the data
   - Make captions and labels clear and informative

Show us the data!



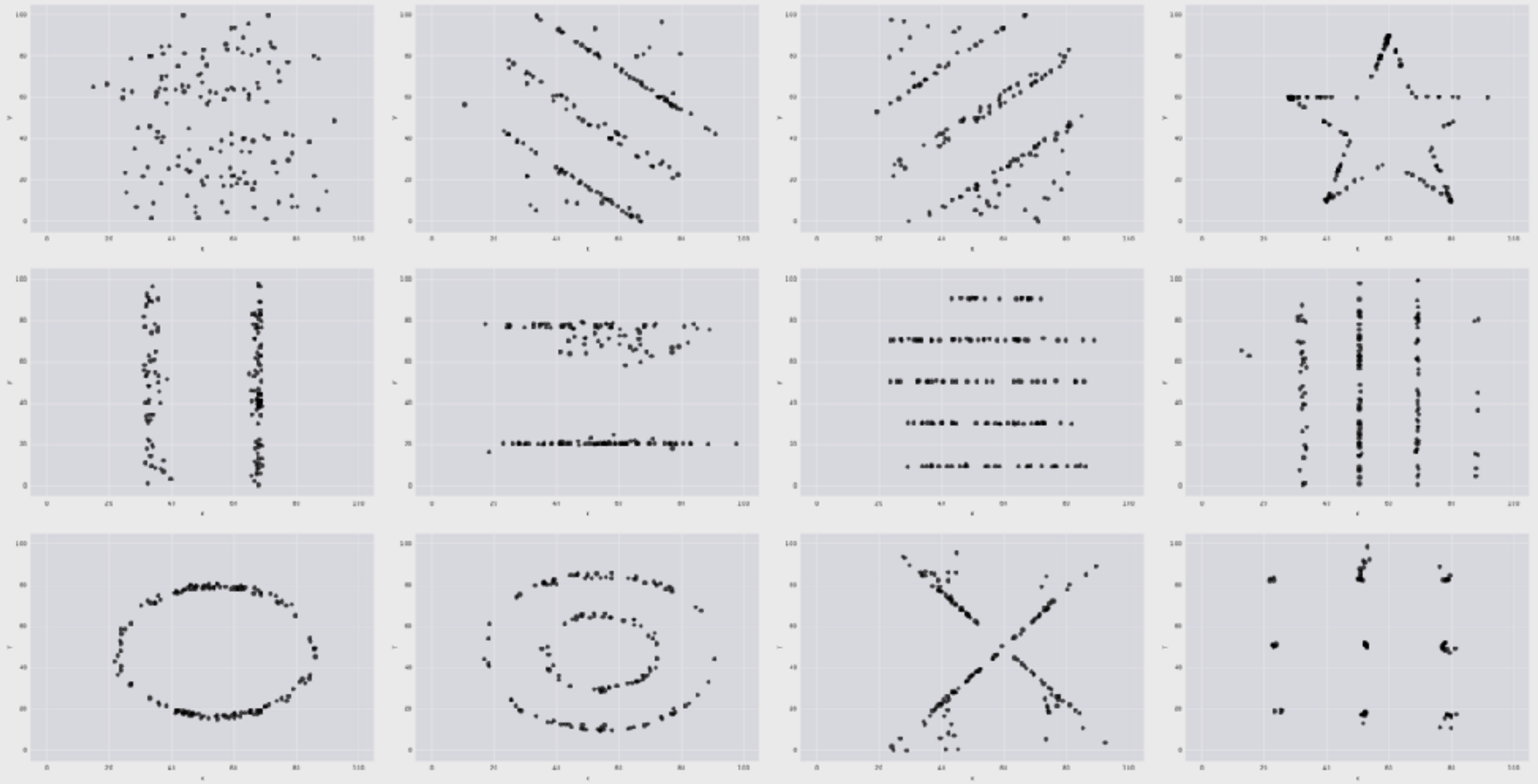






The “Datasaurus Dozen”

```
X Mean: 54.26
Y Mean: 47.83
X SD   : 16.76
Y SD   : 26.93
Corr.  : -0.06
```

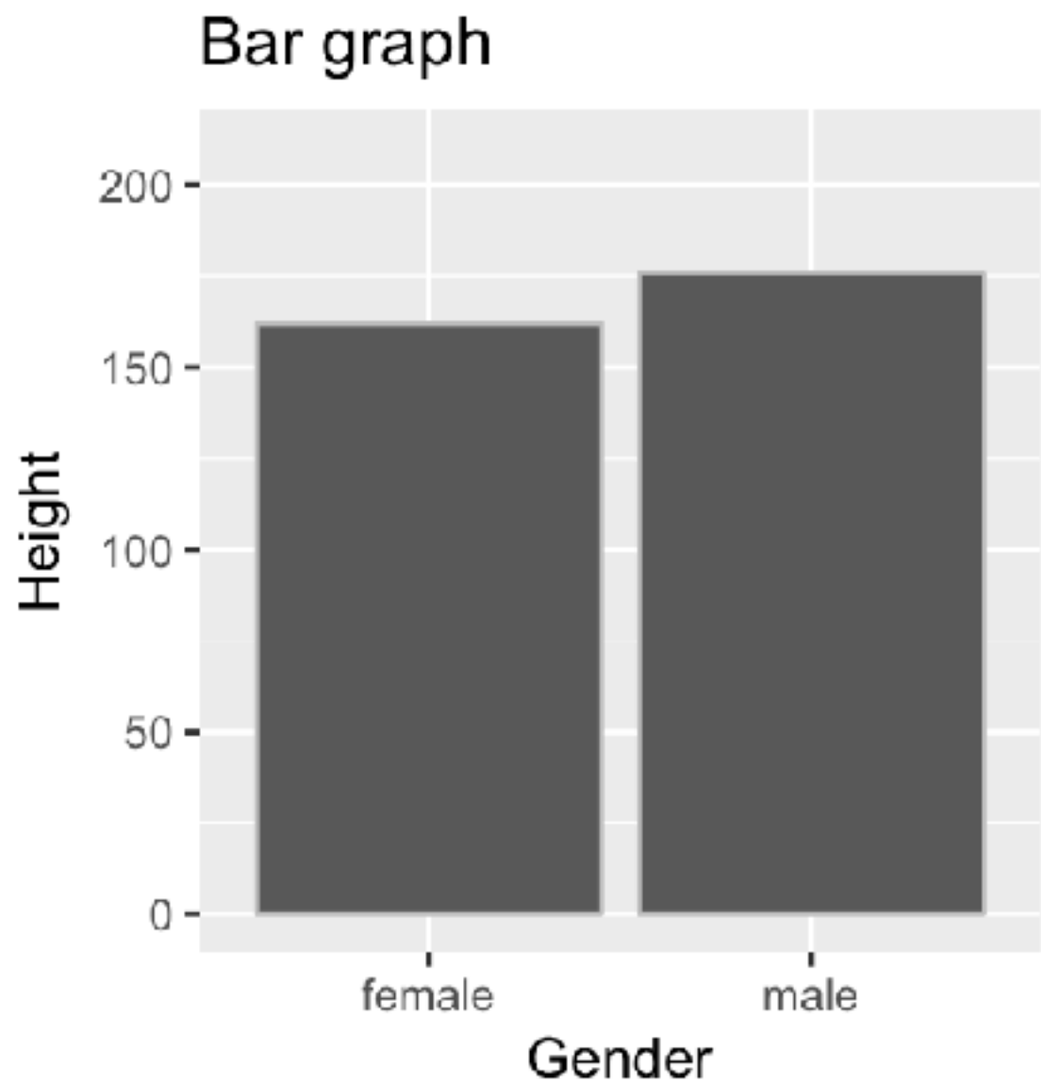


https://www.autodeskresearch.com/publications/samestats

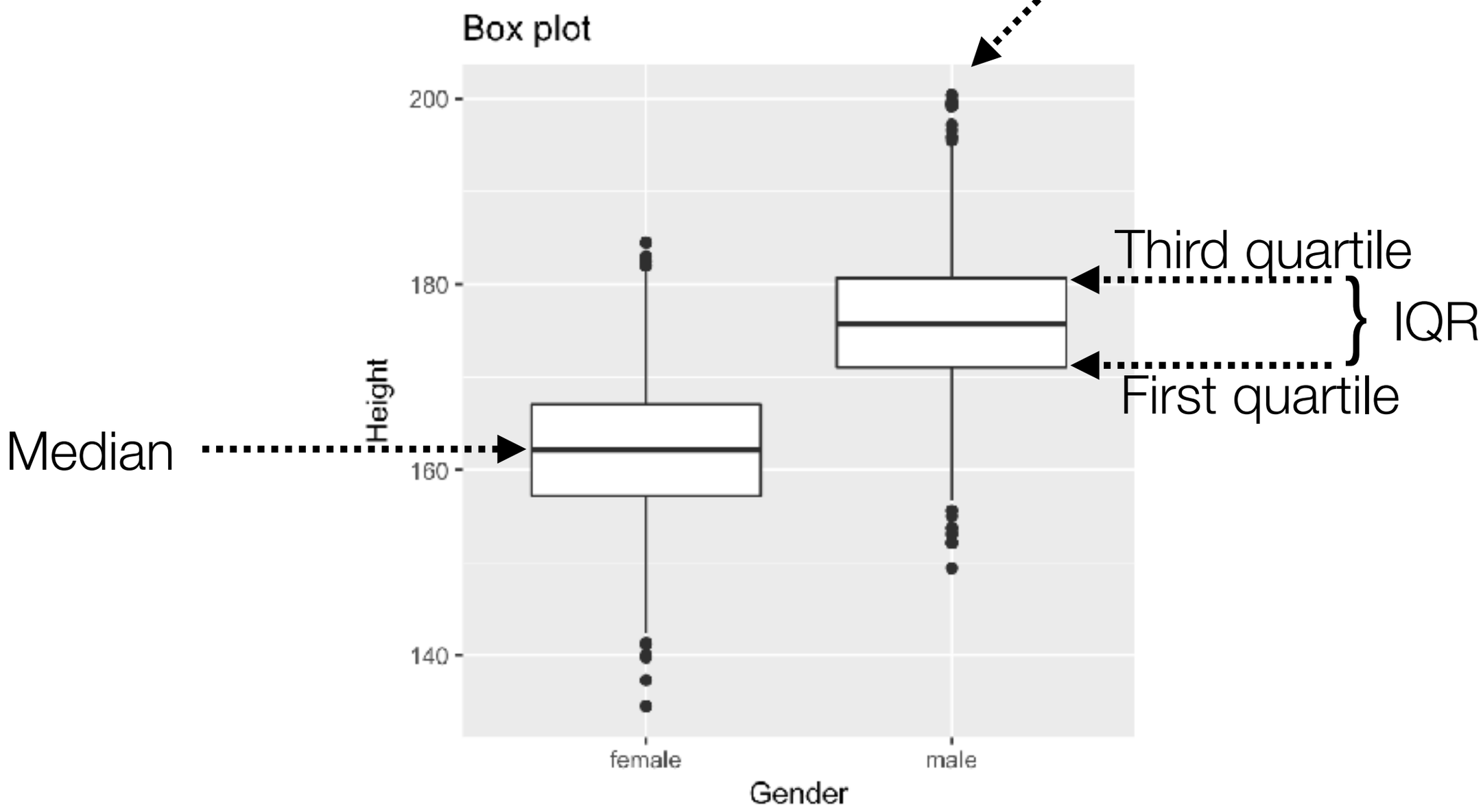
Not a very good graph

```
dfmean <- NHANES_adult %>%
    group_by(Gender) %>%
    summarise(Height=mean(Height))

ggplot(dfmean,aes(x=Gender,y=Height)) +
  geom_bar(stat="identity") +
```

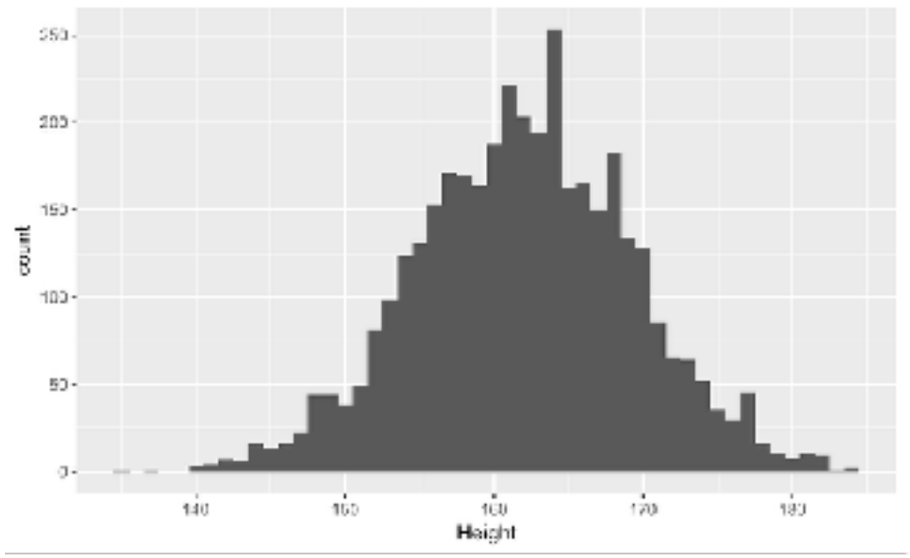


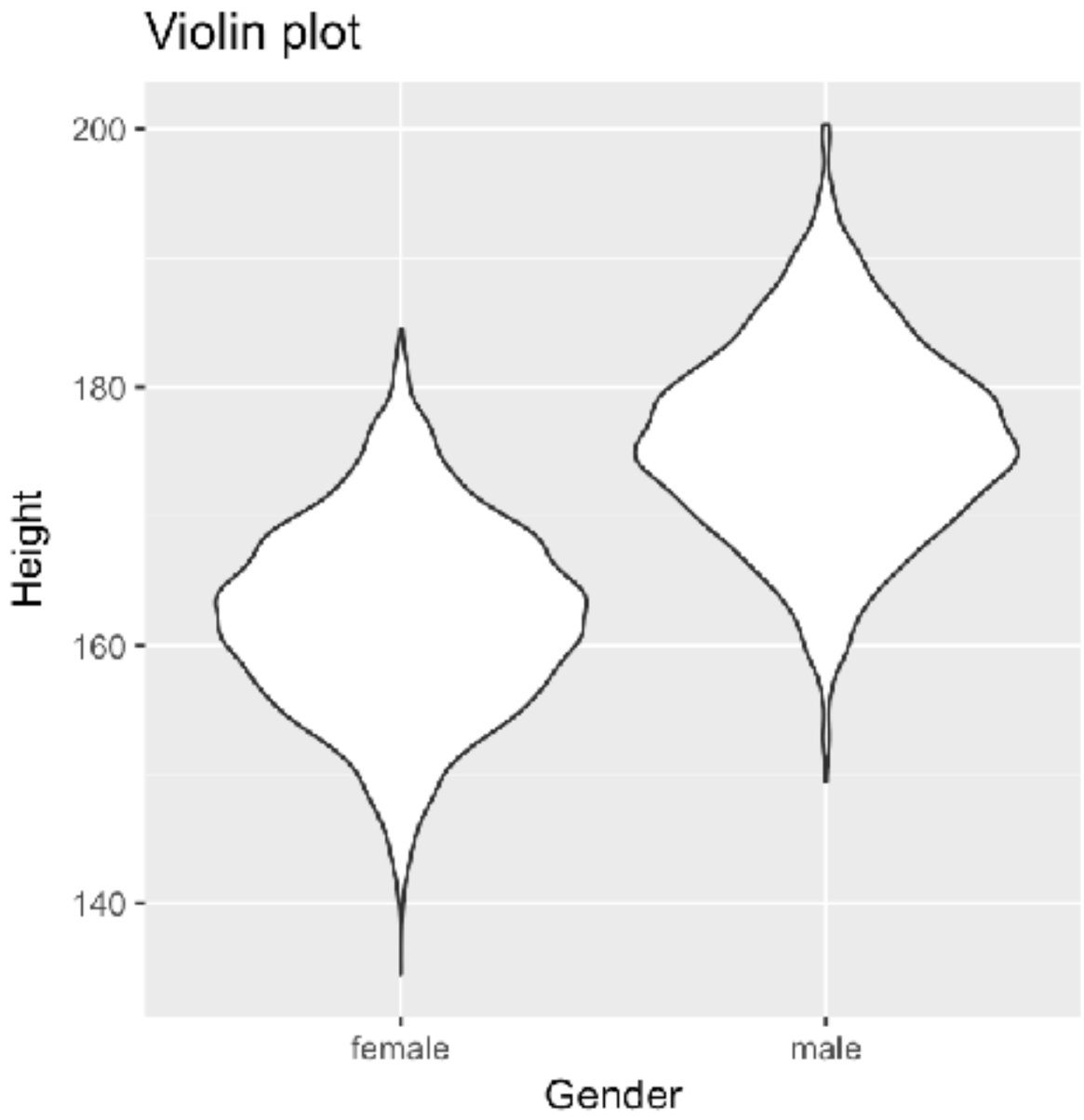
Much better: Box plot




```
ggplot(NHANES_adult,aes(x=Gender,y=Height)) +
  geom_boxplot()
```

Also great: Violin plot



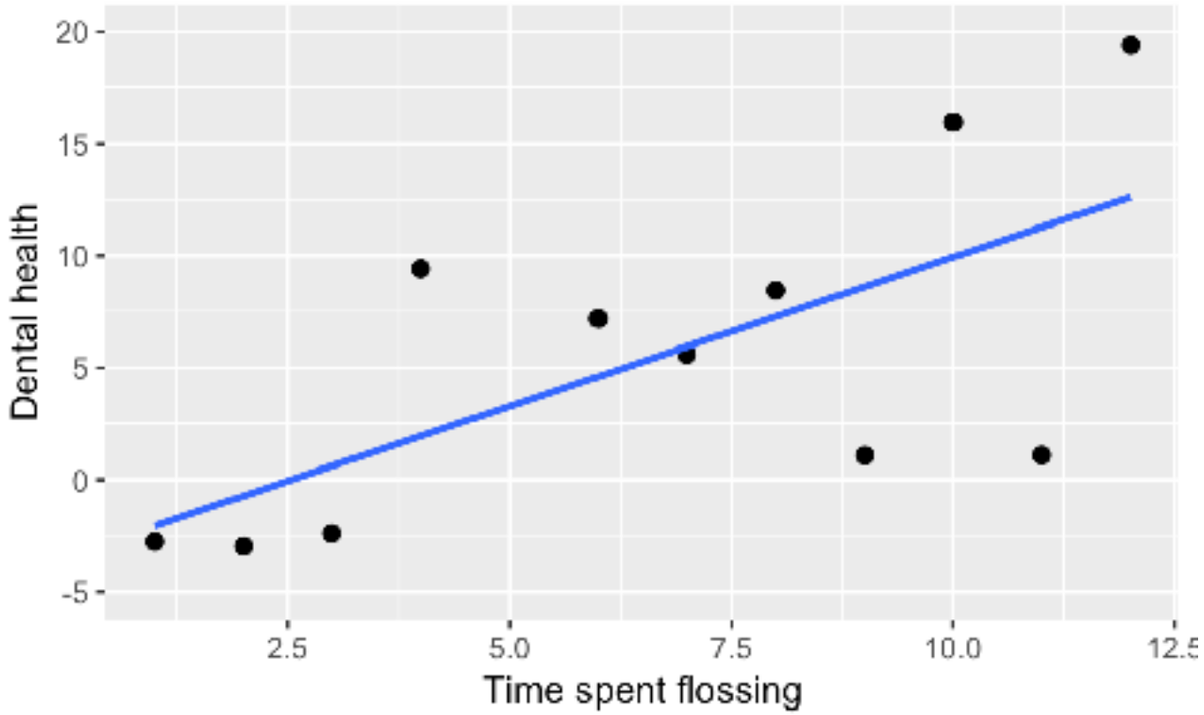


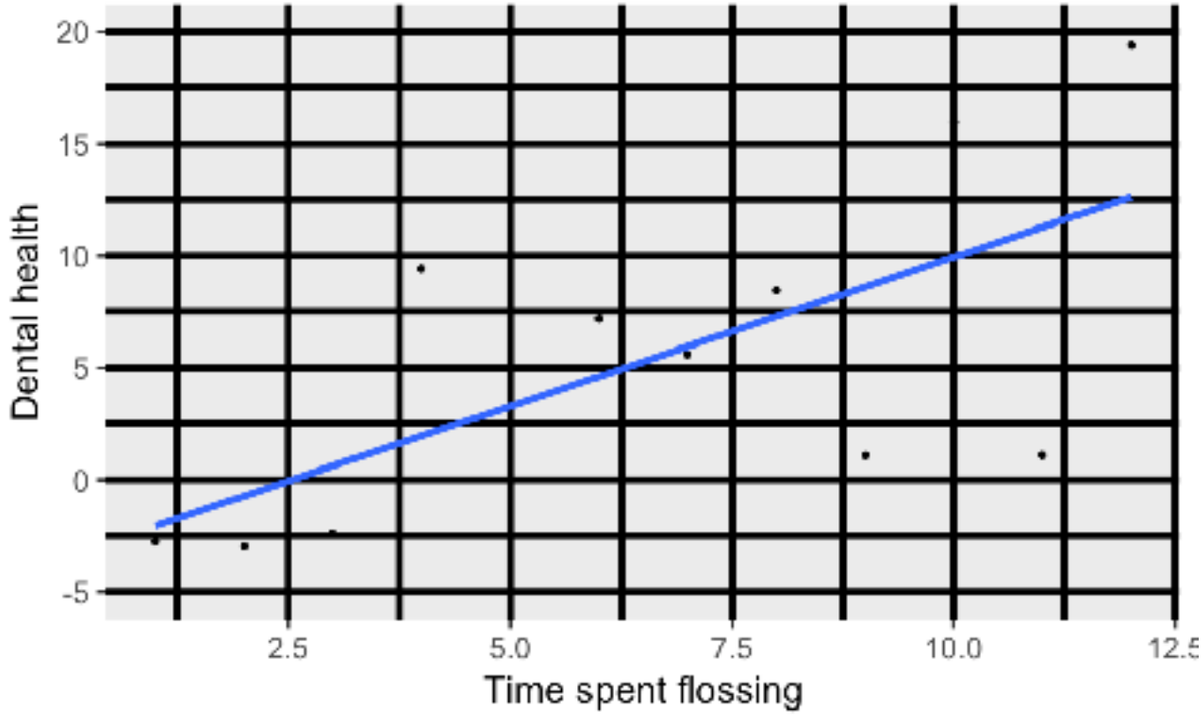

```
ggplot(NHANES_adult,aes(x=Gender,y=Height)) +
  geom_violin()
```

Maximize the data-ink ratio

$$\text{Data-ink ratio} = \frac{\text{Amount of ink used on data}}{\text{Total amount of ink}}$$

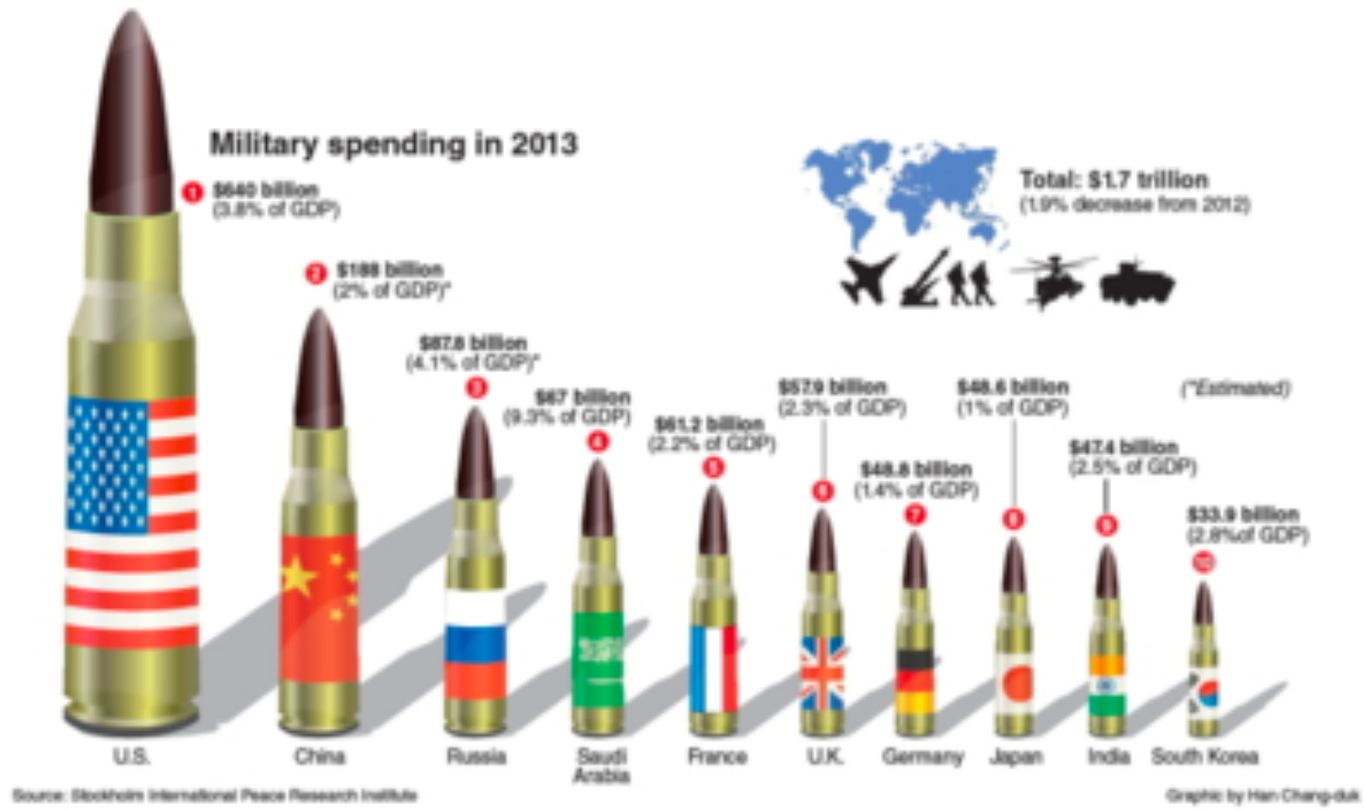
Maximizing the data-ink ratio



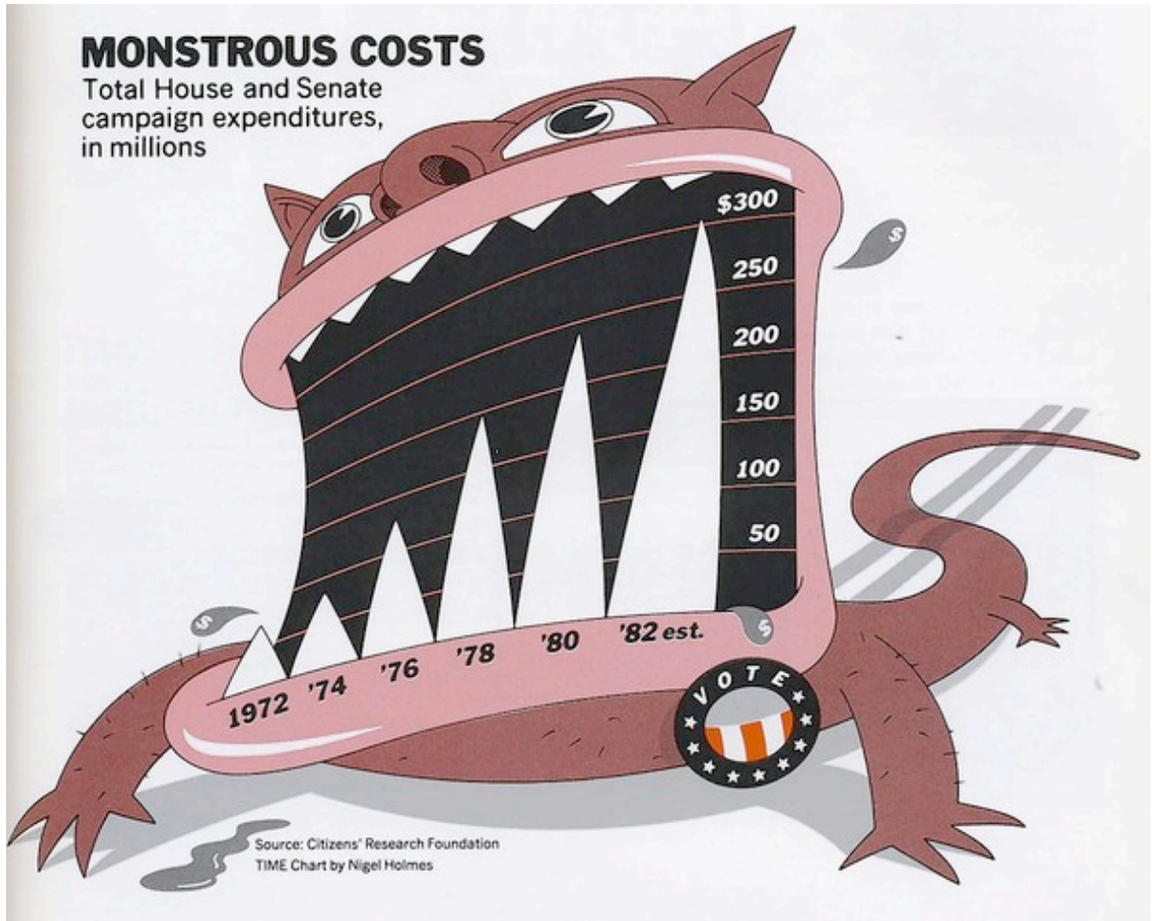


Avoid “chartjunk”

- Extraneous visual elements

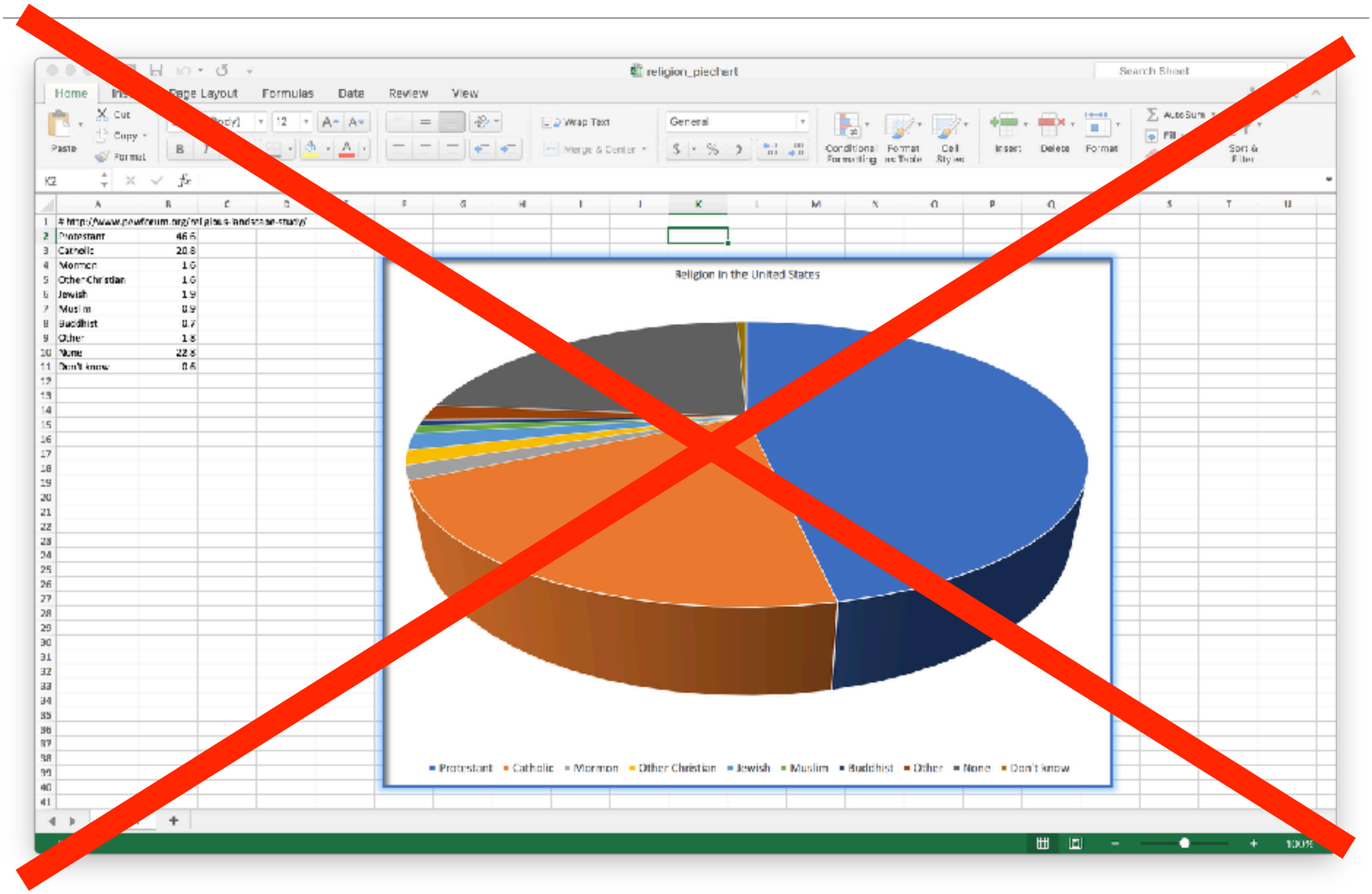


http://junkcharts.typepad.com/junk_charts/2014/10/index.html



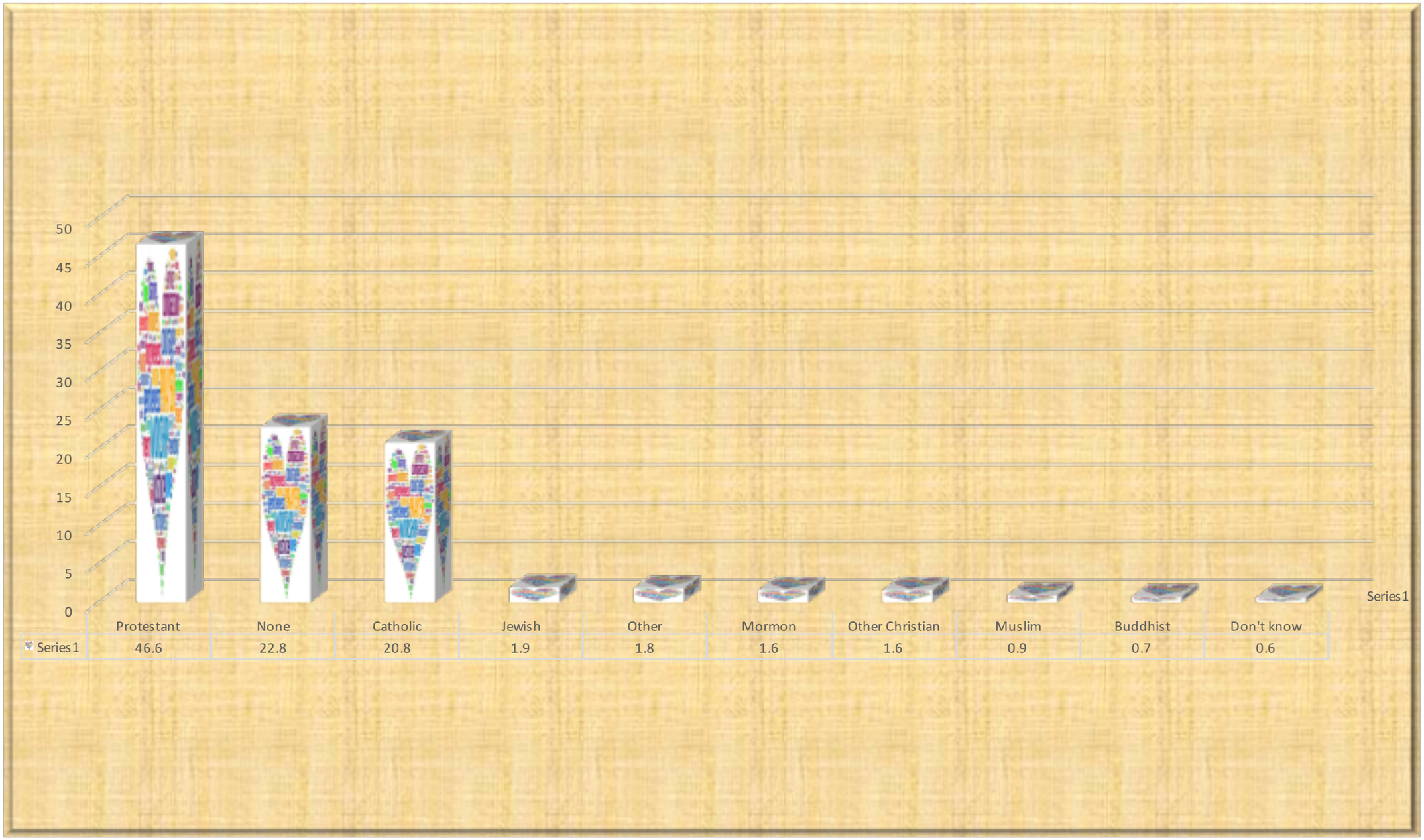
http://classes.engr.oregonstate.edu/eecs/spring2015/cs419-001/Slides/tufteDesign.pdf

Rule #1 for avoiding bad visualizations: Don't use Microsoft Office to generate them



Avoiding chartjunk

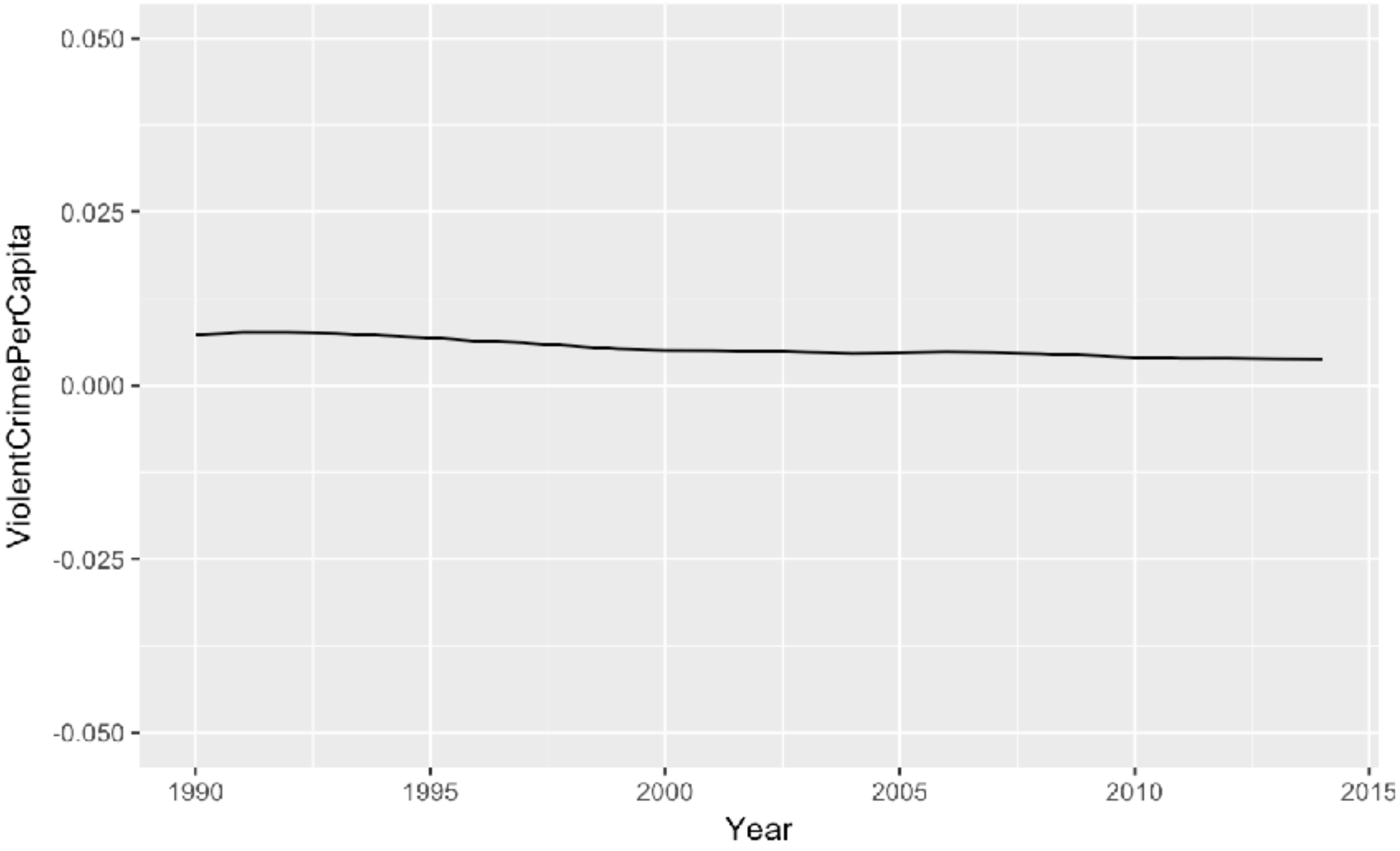
- Avoid textures and images in plots



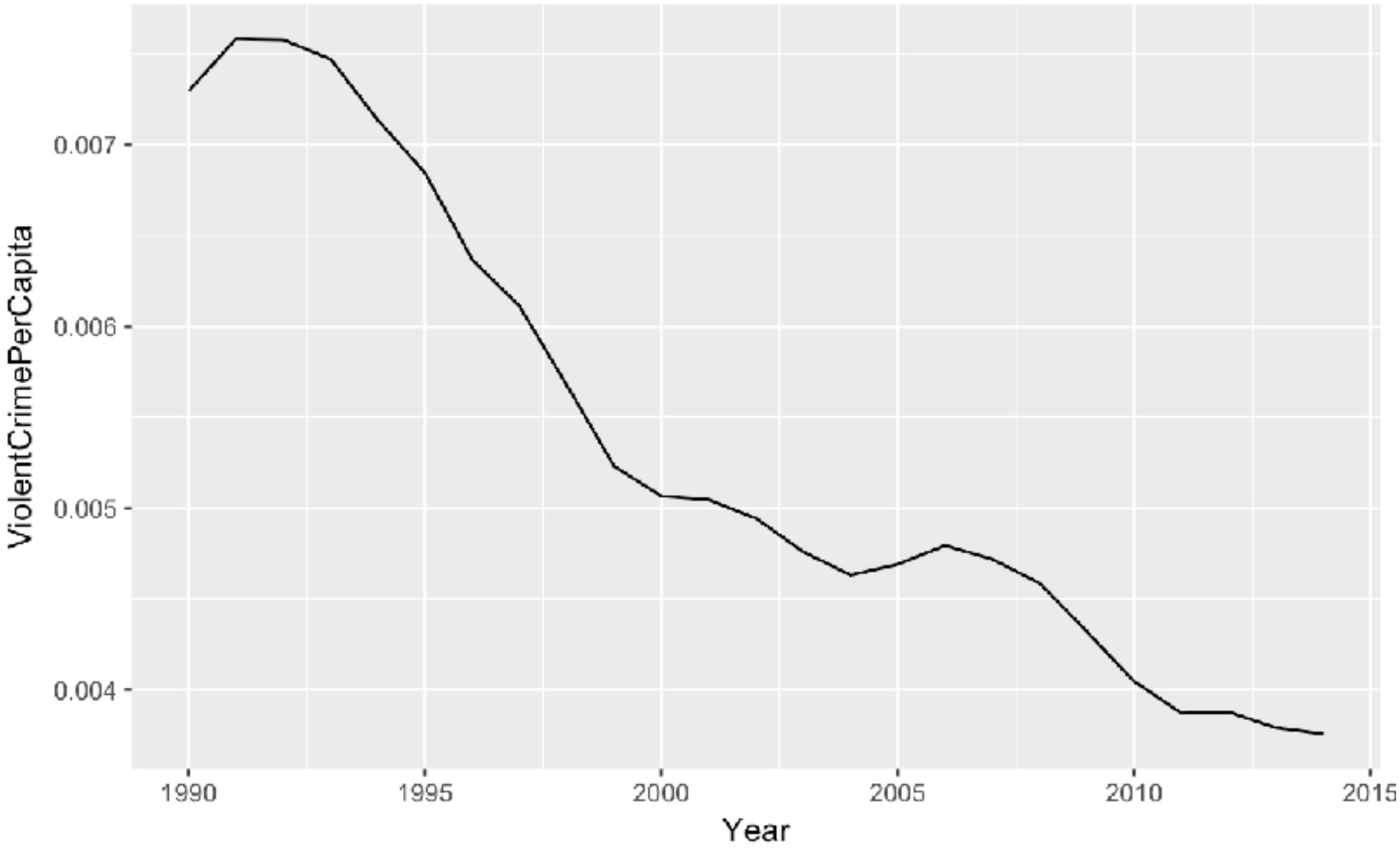
Avoid distorting the data

- Use appropriate scales for the Y axis
- Beware of effects that distort the data

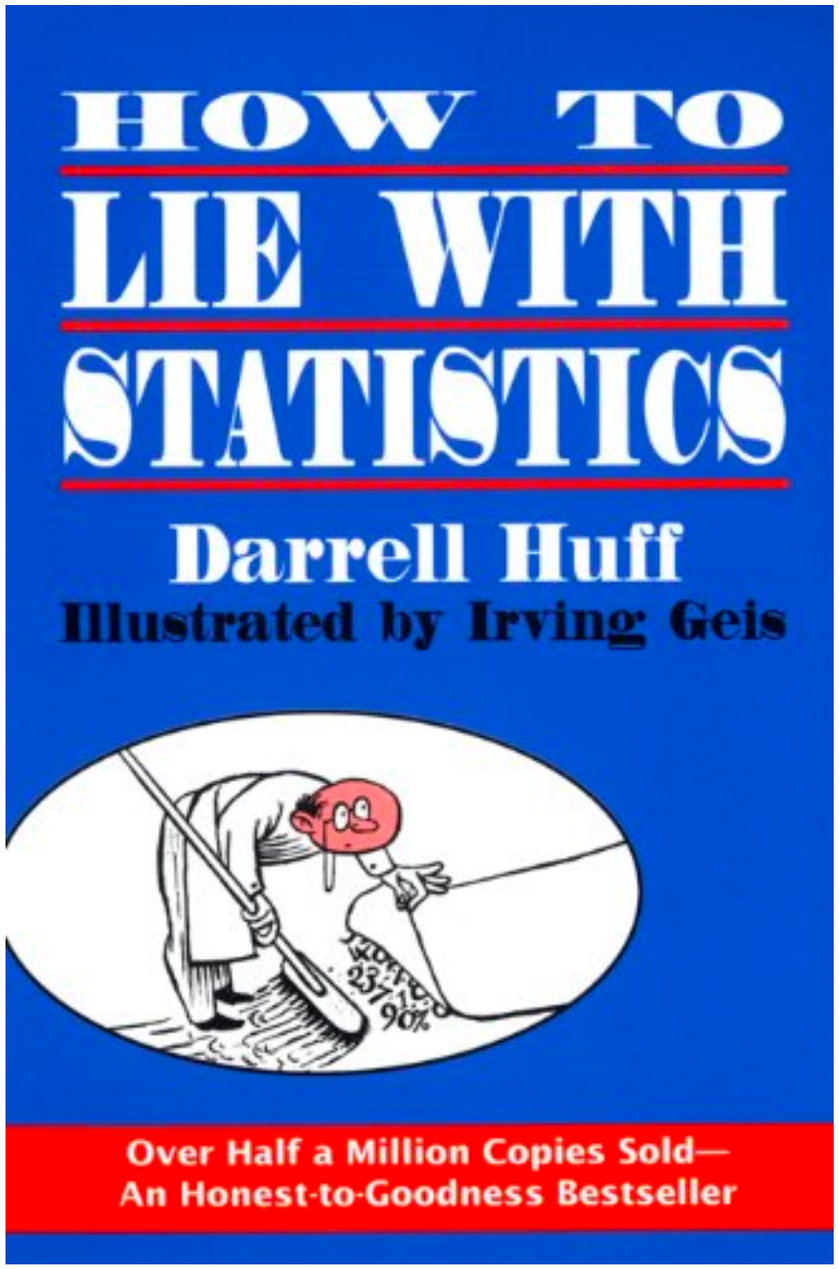
Violent crime was flat from 1990-2014

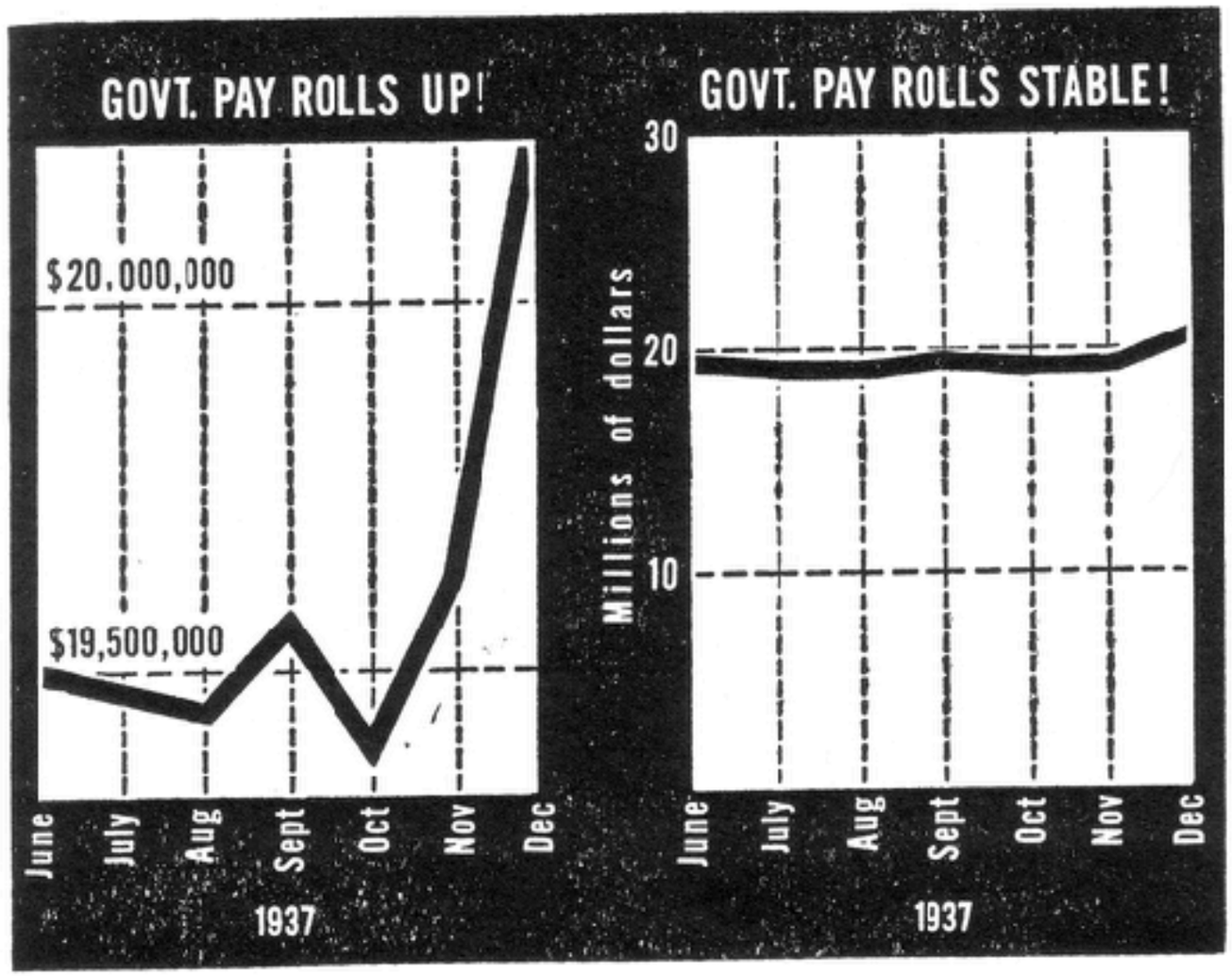


Wait… Violent crime has plummeted since 1990!

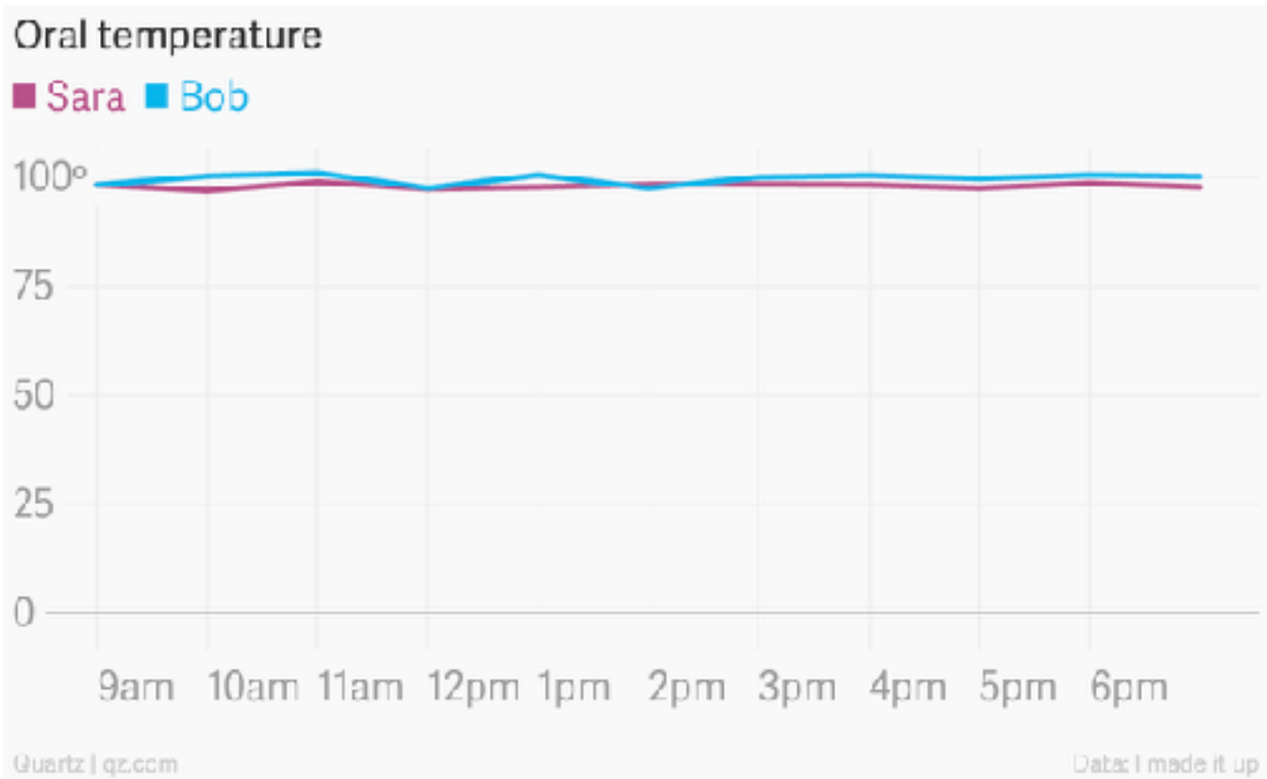


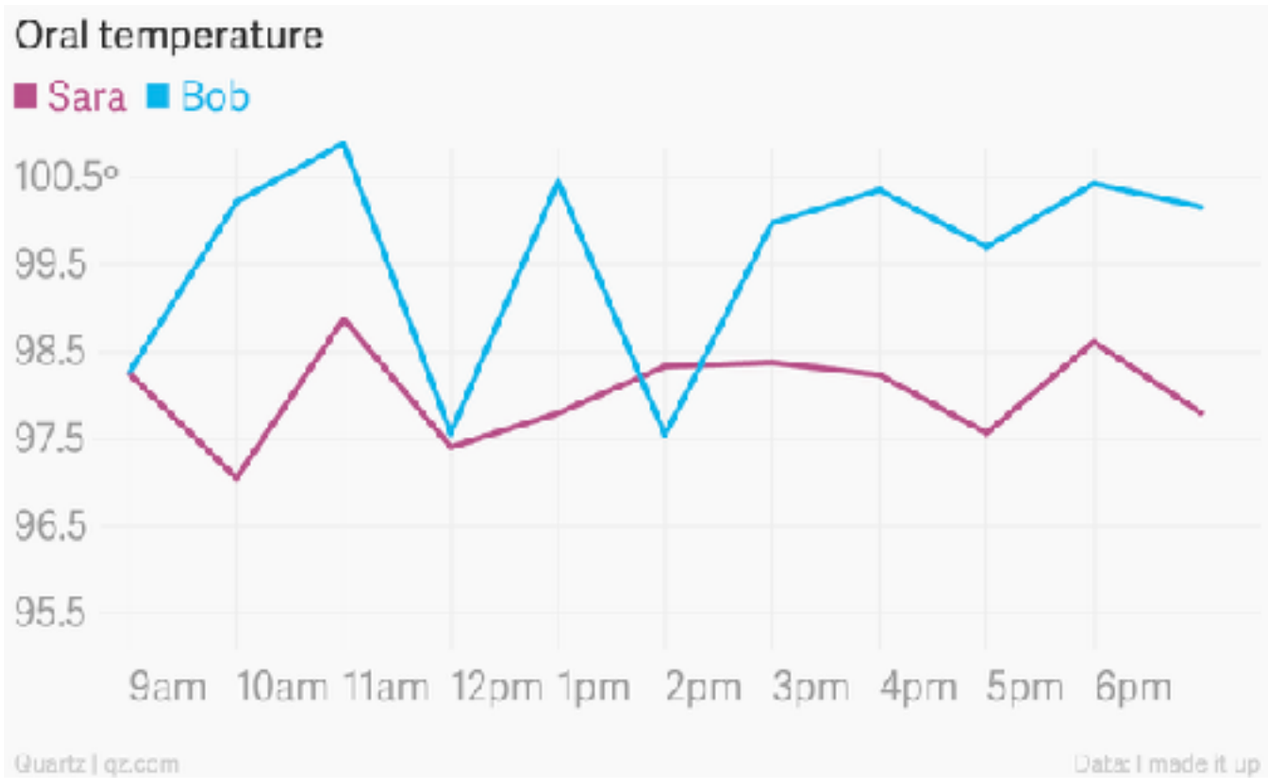
Should you always include zero in the y axis?





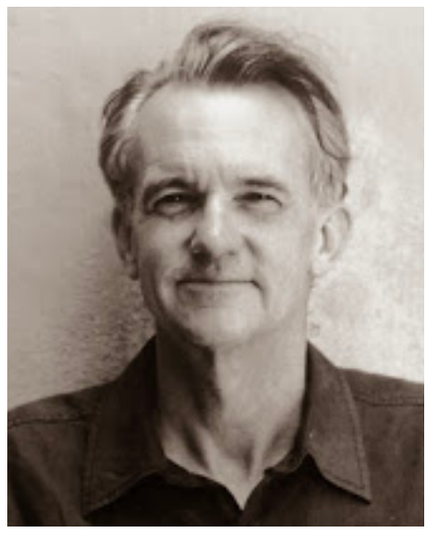
Using zero as the basis often makes no sense

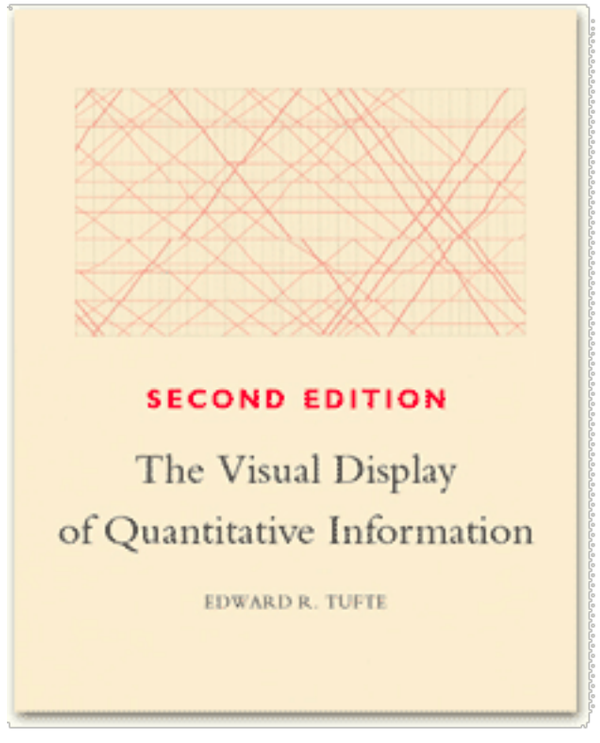




It's ok not to start your Y axis at zero

"In general, in a time-series, use a baseline that shows the data not the zero point; don't spend a lot of empty vertical space trying to reach down to the zero point at the cost of hiding what is going on in the data line itself." Edward Tufte



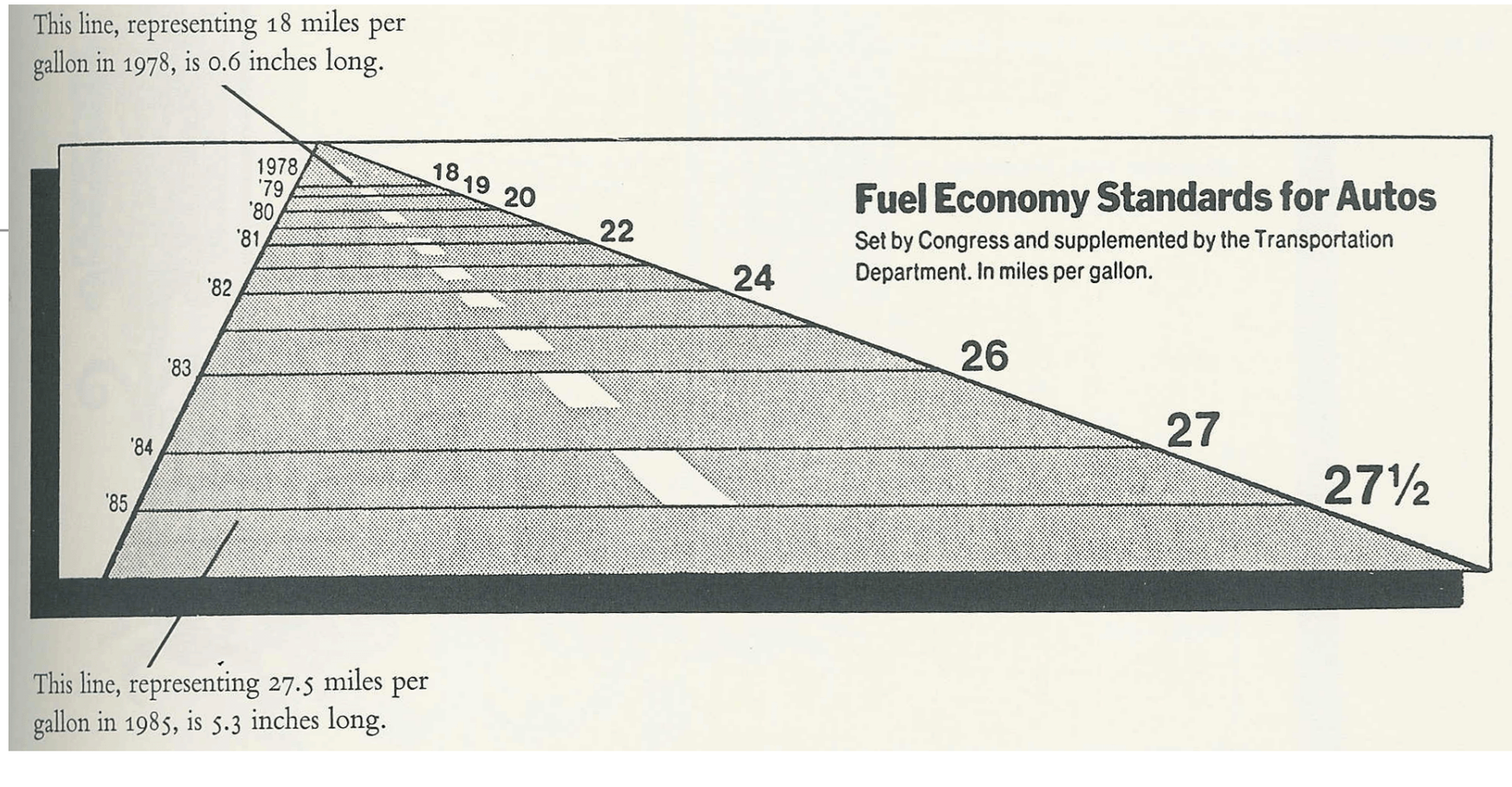


https://qz.com/418083/its-ok-not-to-start-your-y-axis-at-zero/

The “Lie Factor”

- Tufte, 1983
- The size of the effect on the physical graphic, relative to the size of the effect in the data
- A lie factor of about 1 is good

The Lie Factor

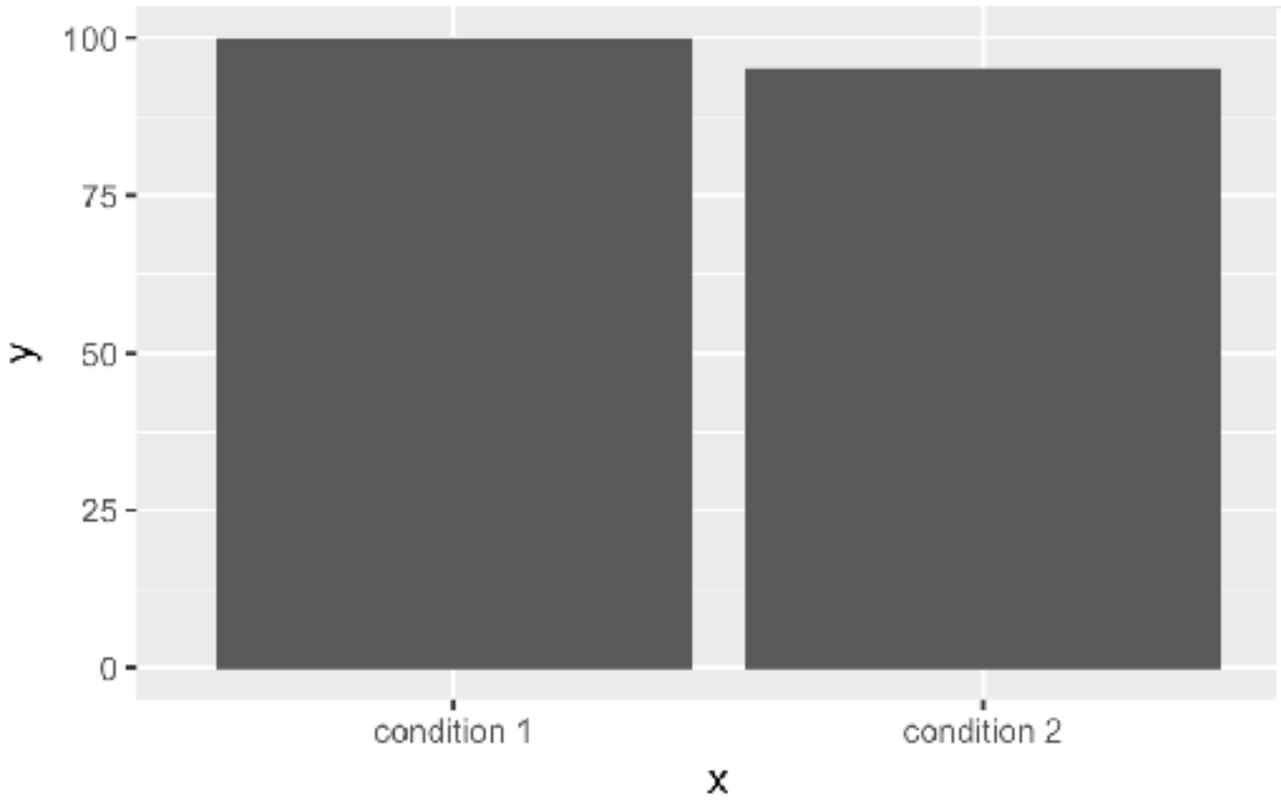


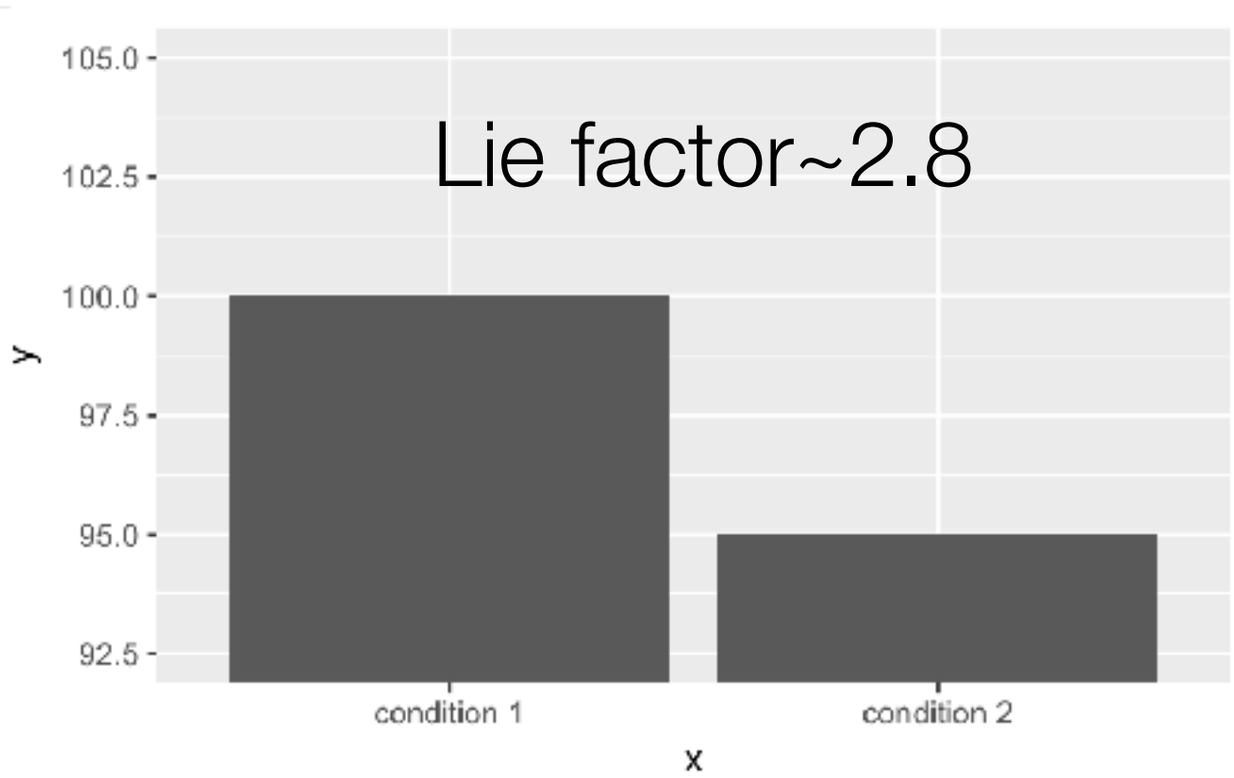
- Change in fuel economy from 1978-1985 = 53% (0.53)
- Change in graphic = change from 0.6” to 5.3”
- (5.3 - 0.6)/0.6 = 7.83 = 783%
- Lie Factor = 7.83/0.53 = 14.8 -- almost 15 times reality

Tufte, 1983/R. Smith

Always use zero as the basis for bar/column charts

- Doing otherwise introduces a potential lie factor

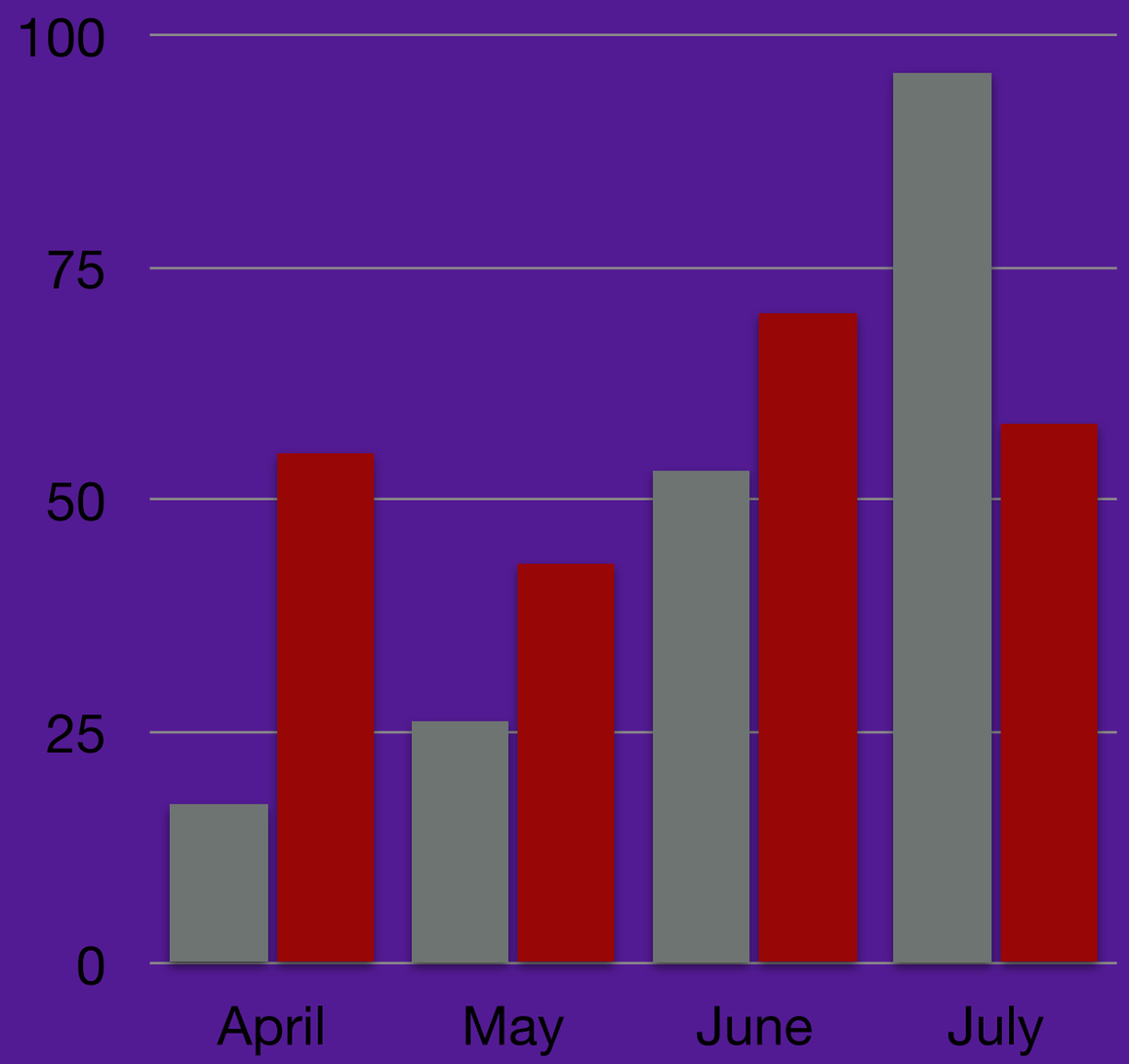




Remember human limitations

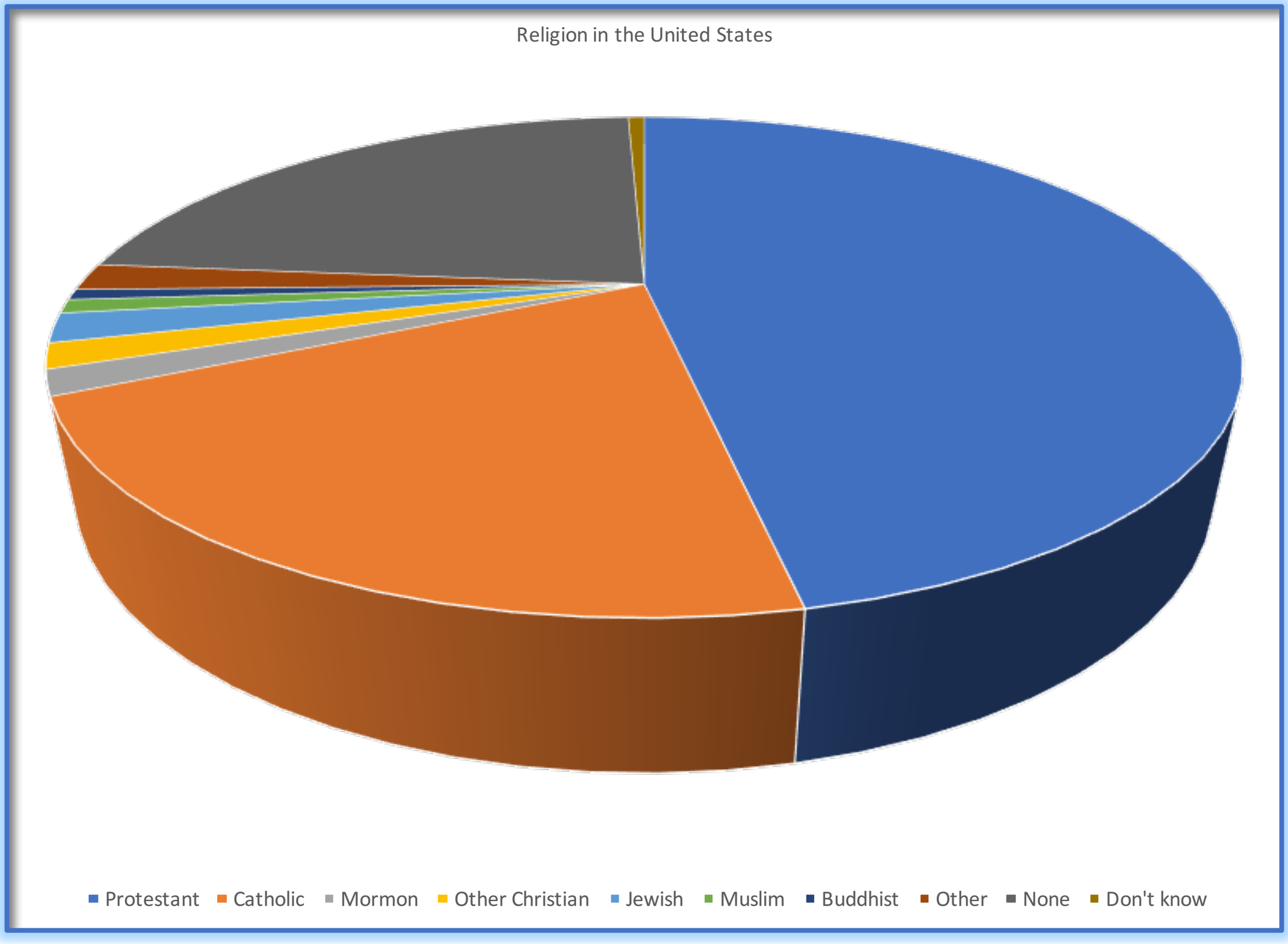
- Perceptual limitations
  - Many people have problematic color vision
  - Volume/area is harder to perceive than length
- Cognitive limitations
  - We have limited working memory capacity
  - Don’t make the viewer remember too much

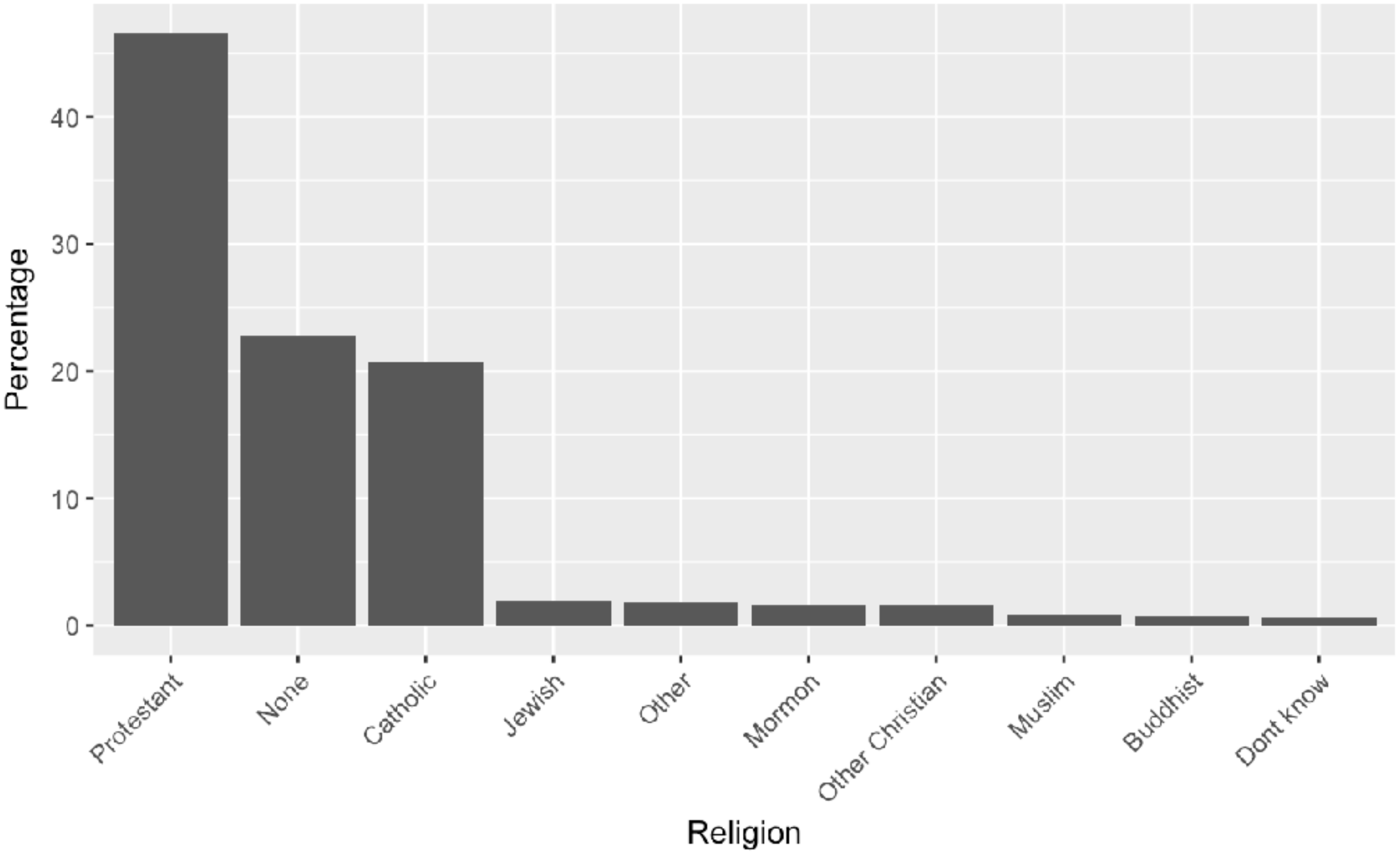
Always use brightness contrast in addition to color



Volume can be very hard to distinguish visually

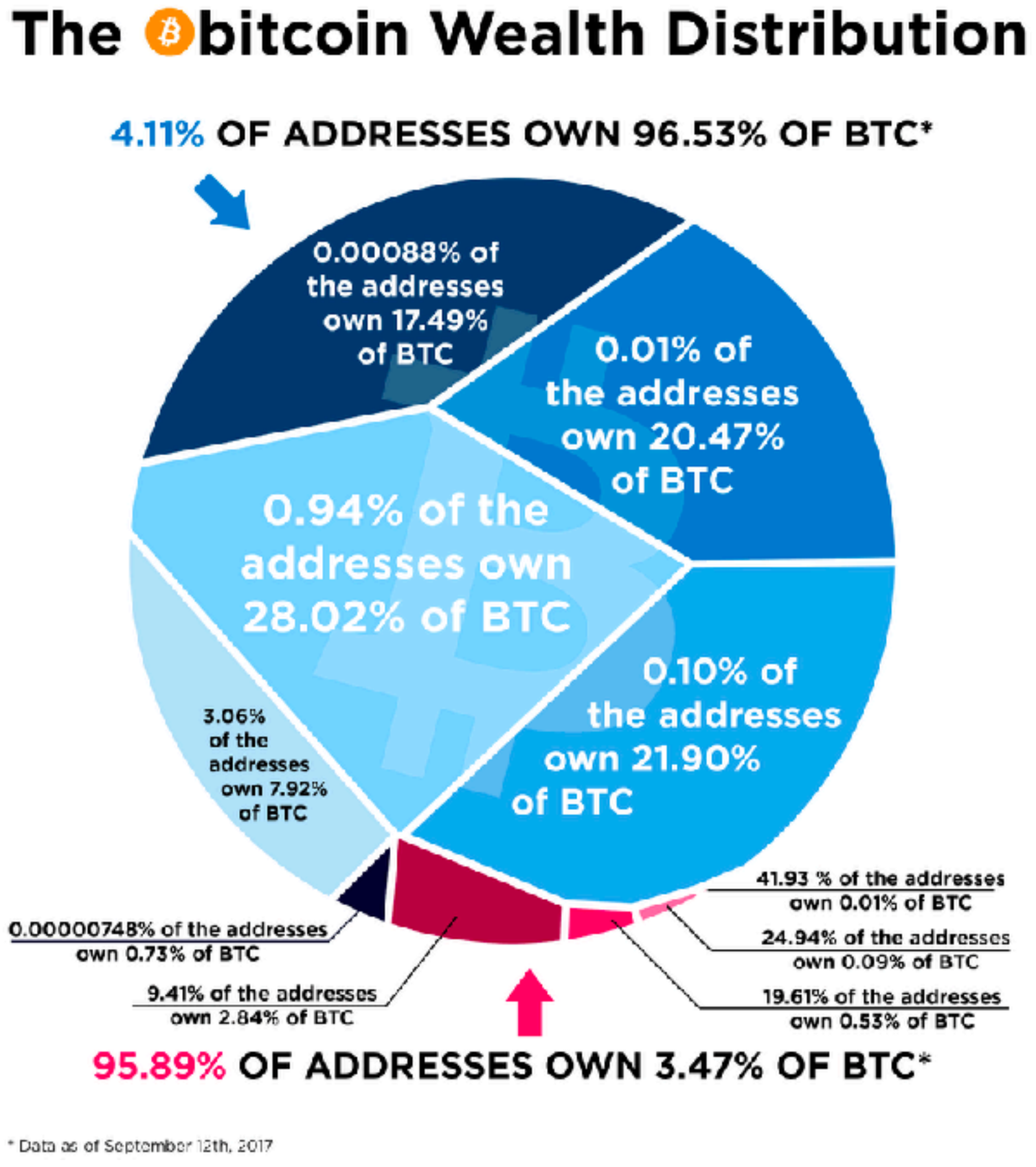
Don’t make your viewer remember too much




Percentage
40
30
20
10
0
Protestant
None
Catholic
Jewish
Other
Mormon
Other Christian
Muslim
Buddhist
Dont know
Religion

Group exercise

- What is the message of this visualization?
- How could that message be better conveyed?

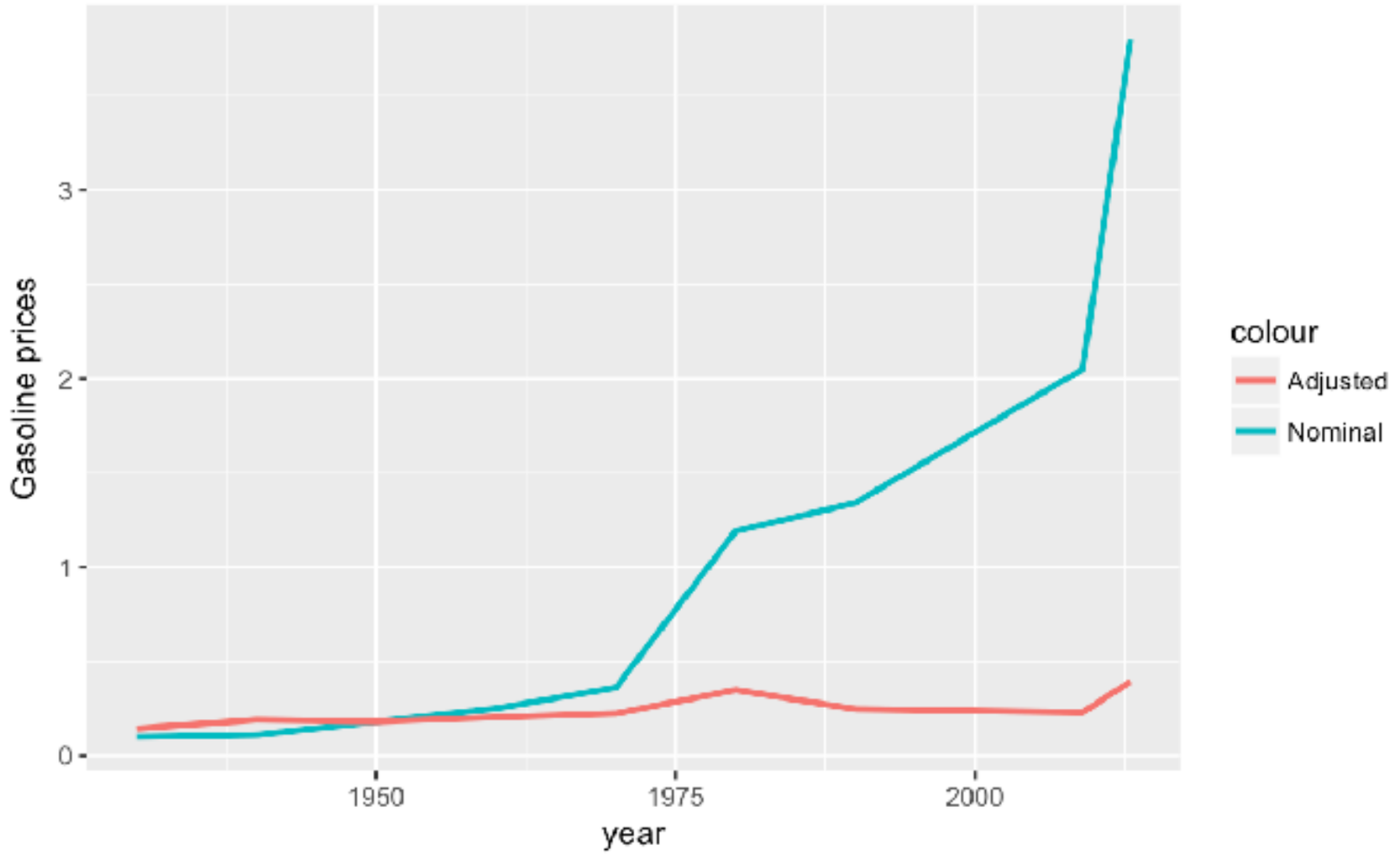


https://howmuch.net/articles/bitcoin-wealth-distribution

Correcting for other factors

- Inflation
- Population size
- Seasonal adjustment

Gasoline prices, with and without adjustment for inflation (using CPI)



Recap

- Focus on showing the data and revealing its story
- Don’t misrepresent the data through graphics