Introduction to Data Visualization

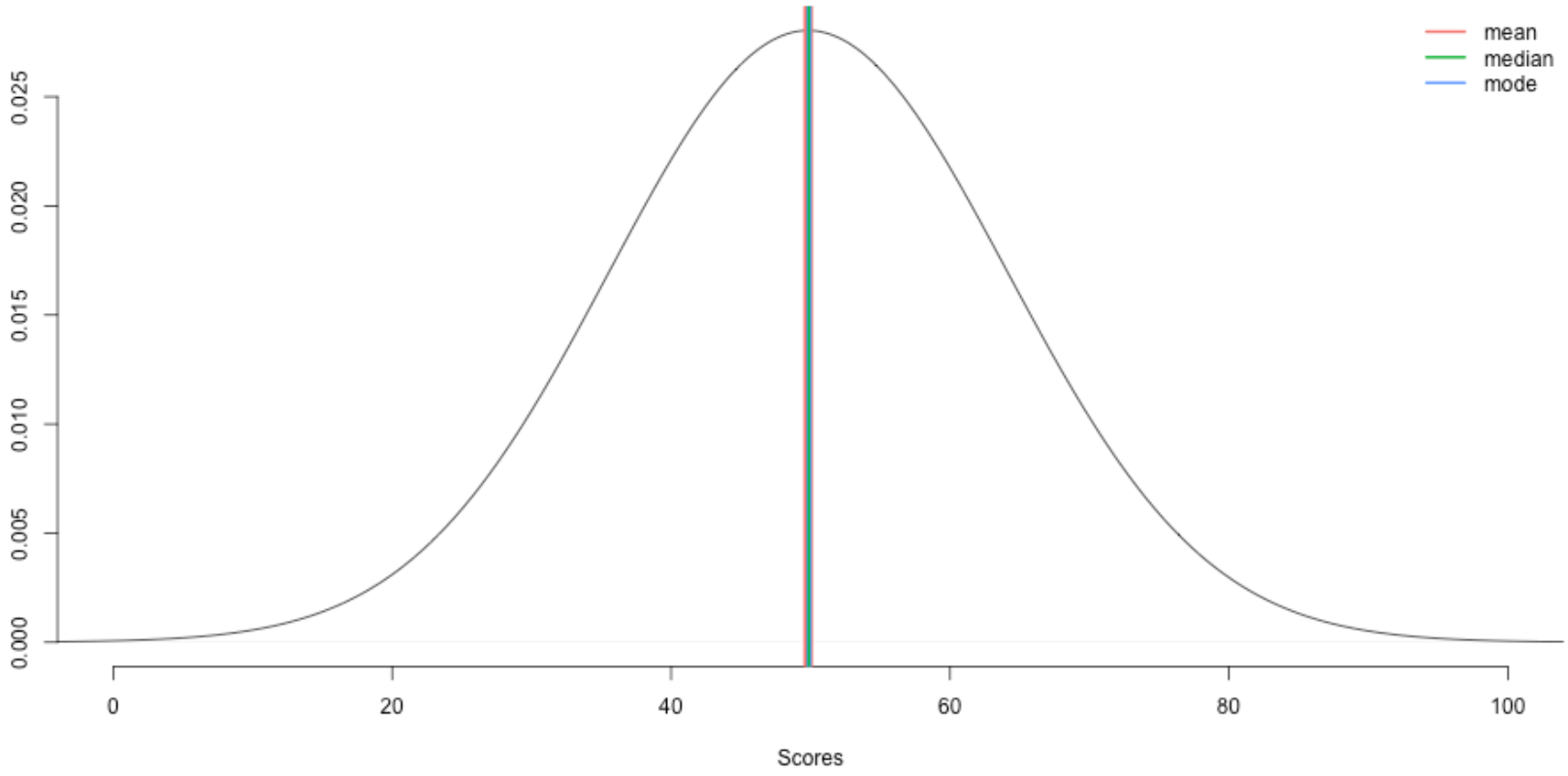
Some dos and don'ts"

Daniel Anderson

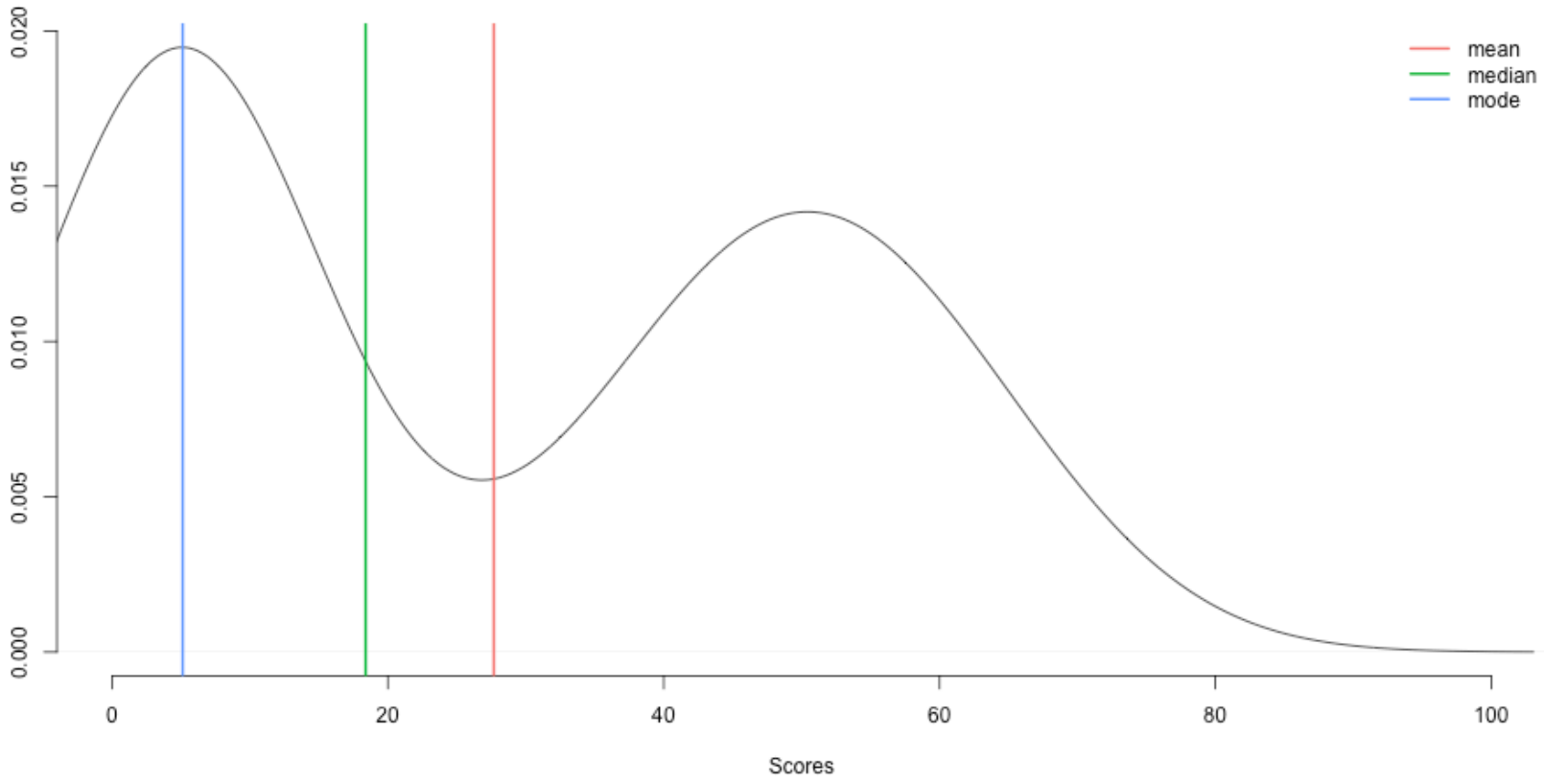
Quick stats lesson (with pictures!)

- Mean versus Median versus Mode
- Standard Deviation
- Correlation

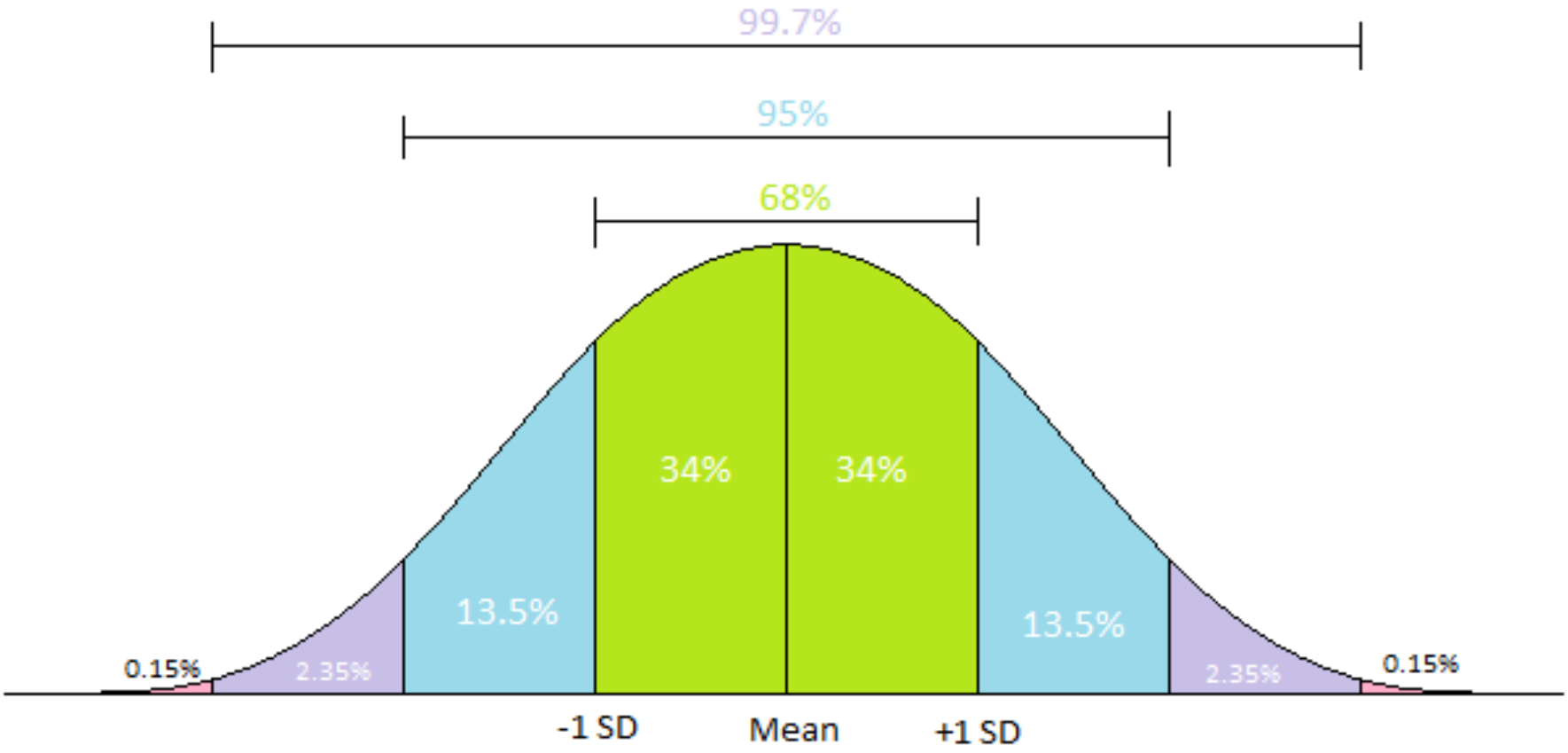
Mean versus median versus mode



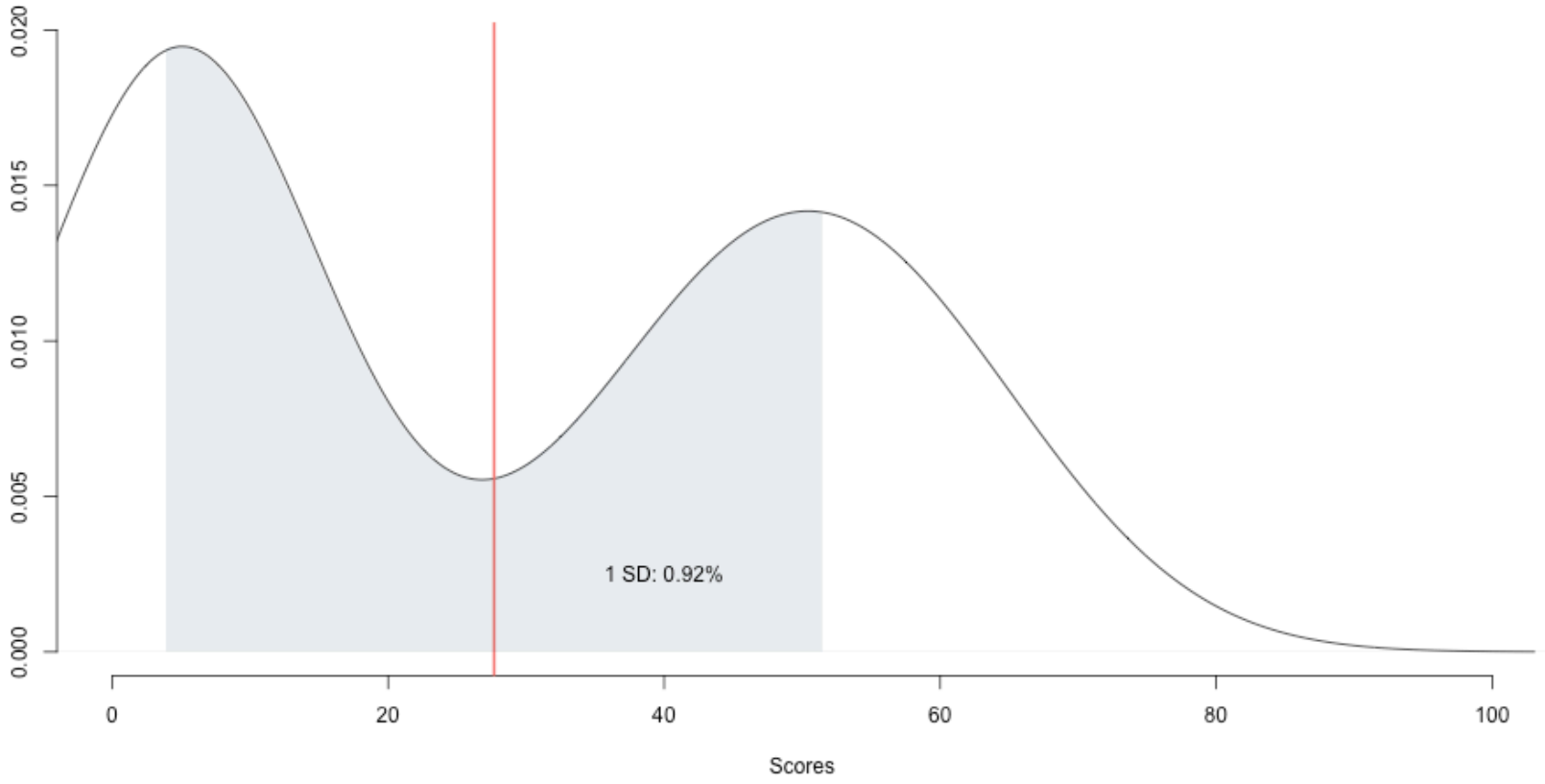
Mean versus median versus mode



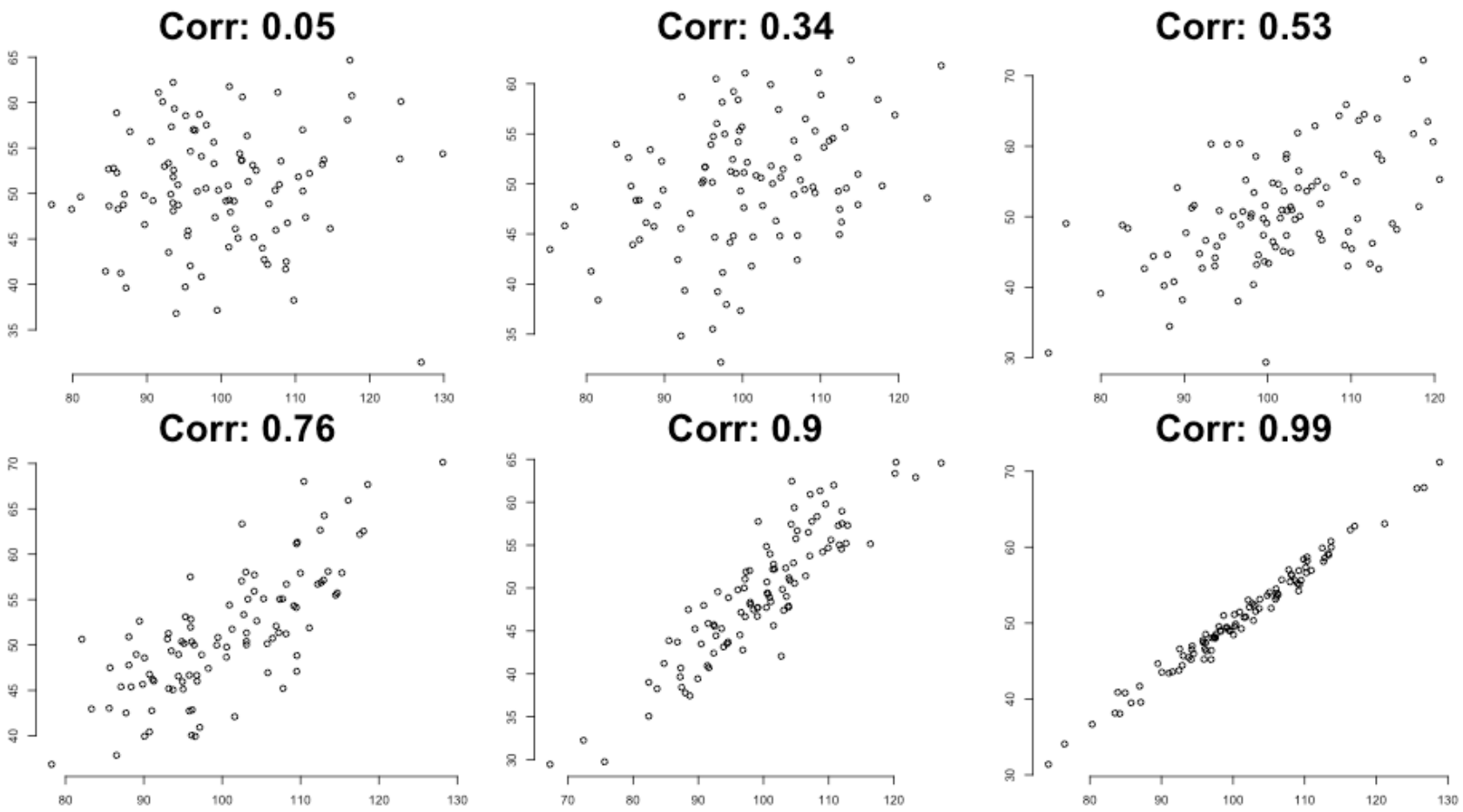
Standard deviation



Standard deviation



Correlation



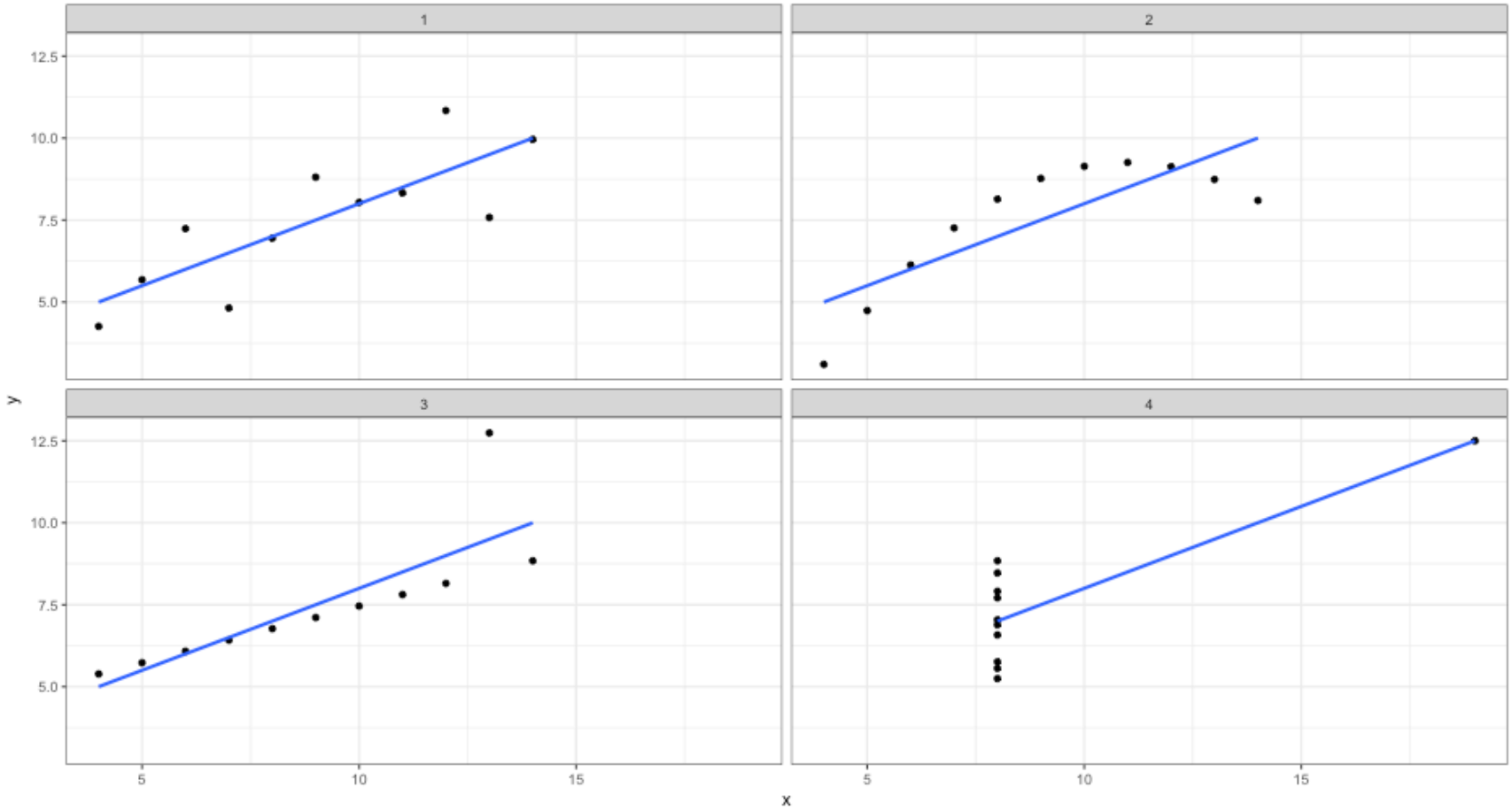
> “Above all else, show the data”

Edward Tufte

Classical example: Anscombe's Quartet

PLOT	MEAN_X	MEAN_Y	SD_X	SD_Y	COR
1	9	7.500909	3.316625	2.031568	0.8164205
2	9	7.500909	3.316625	2.031657	0.8162365
3	9	7.500000	3.316625	2.030424	0.8162867
4	9	7.500909	3.316625	2.030578	0.8165214

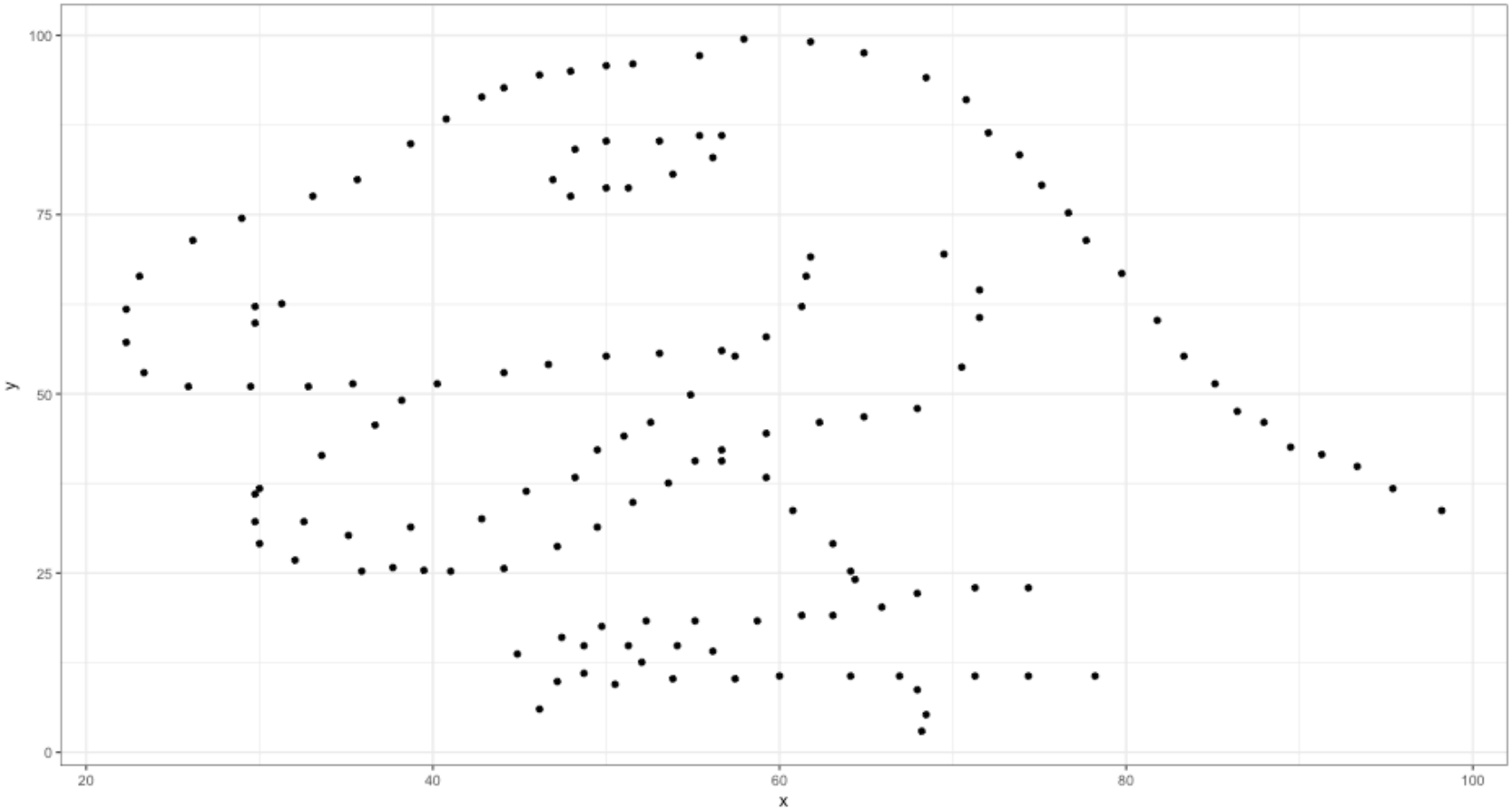
In visual form



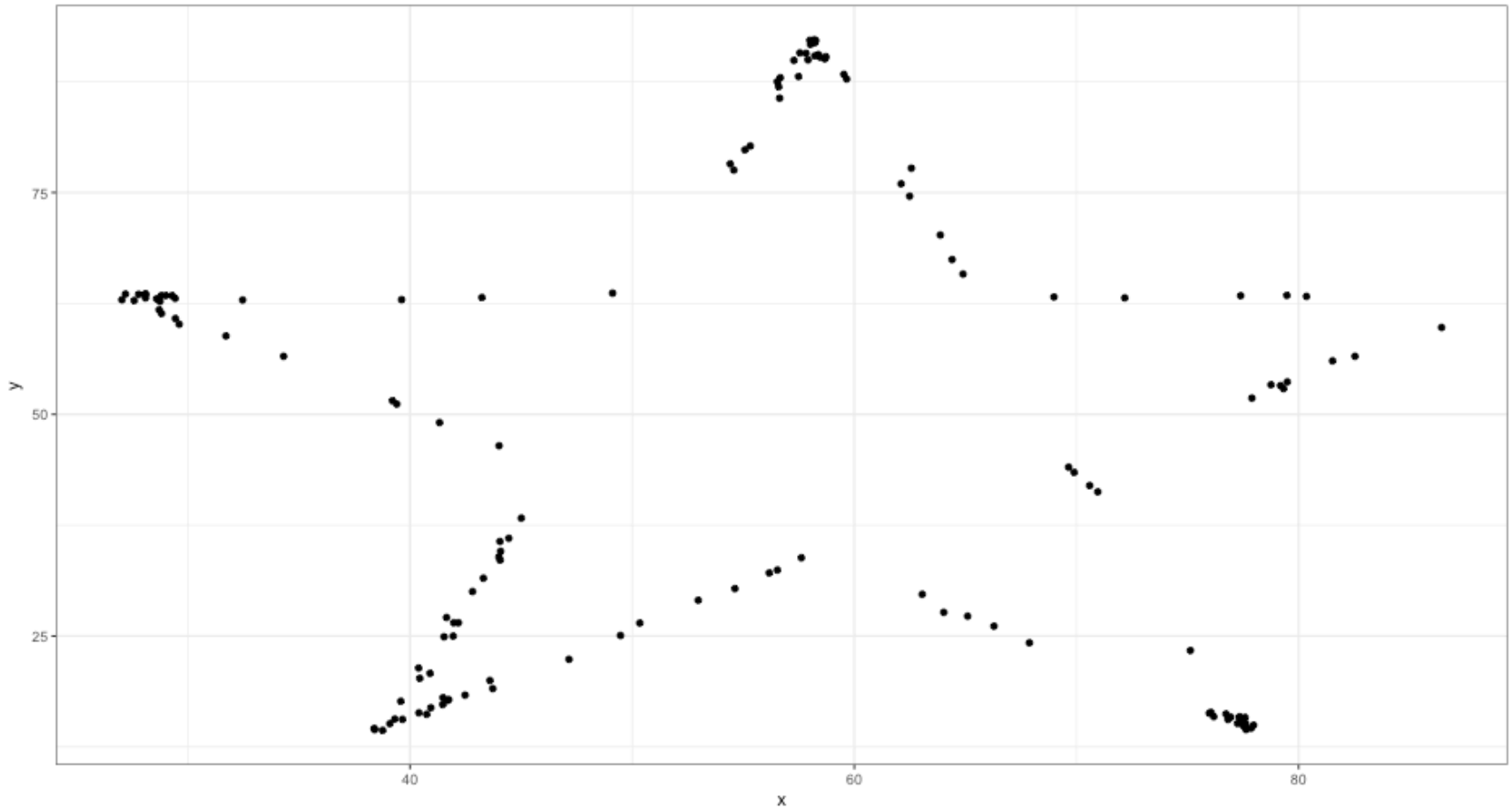
Better example

DATASET	MEAN_X	MEAN_Y	SD_X	SD_Y	COR
away	54.26610	47.83472	16.76983	26.93974	-0.0641284
bullseye	54.26873	47.83082	16.76924	26.93573	-0.0685864
circle	54.26732	47.83772	16.76001	26.93004	-0.0683434
dino	54.26327	47.83225	16.76514	26.93540	-0.0644719
dots	54.26030	47.83983	16.76774	26.93019	-0.0603414
h_lines	54.26144	47.83025	16.76590	26.93988	-0.0617148
high_lines	54.26881	47.83545	16.76670	26.94000	-0.0685042
slant_down	54.26785	47.83590	16.76676	26.93610	-0.0689797
slant_up	54.26588	47.83150	16.76885	26.93861	-0.0686092
star	54.26734	47.83955	16.76896	26.93027	-0.0629611
v_lines	54.26993	47.83699	16.76996	26.93768	-0.0694456
wide_lines	54.26692	47.83160	16.77000	26.93790	-0.0665752
x_shape	54.26015	47.83972	16.76996	26.93000	-0.0655833

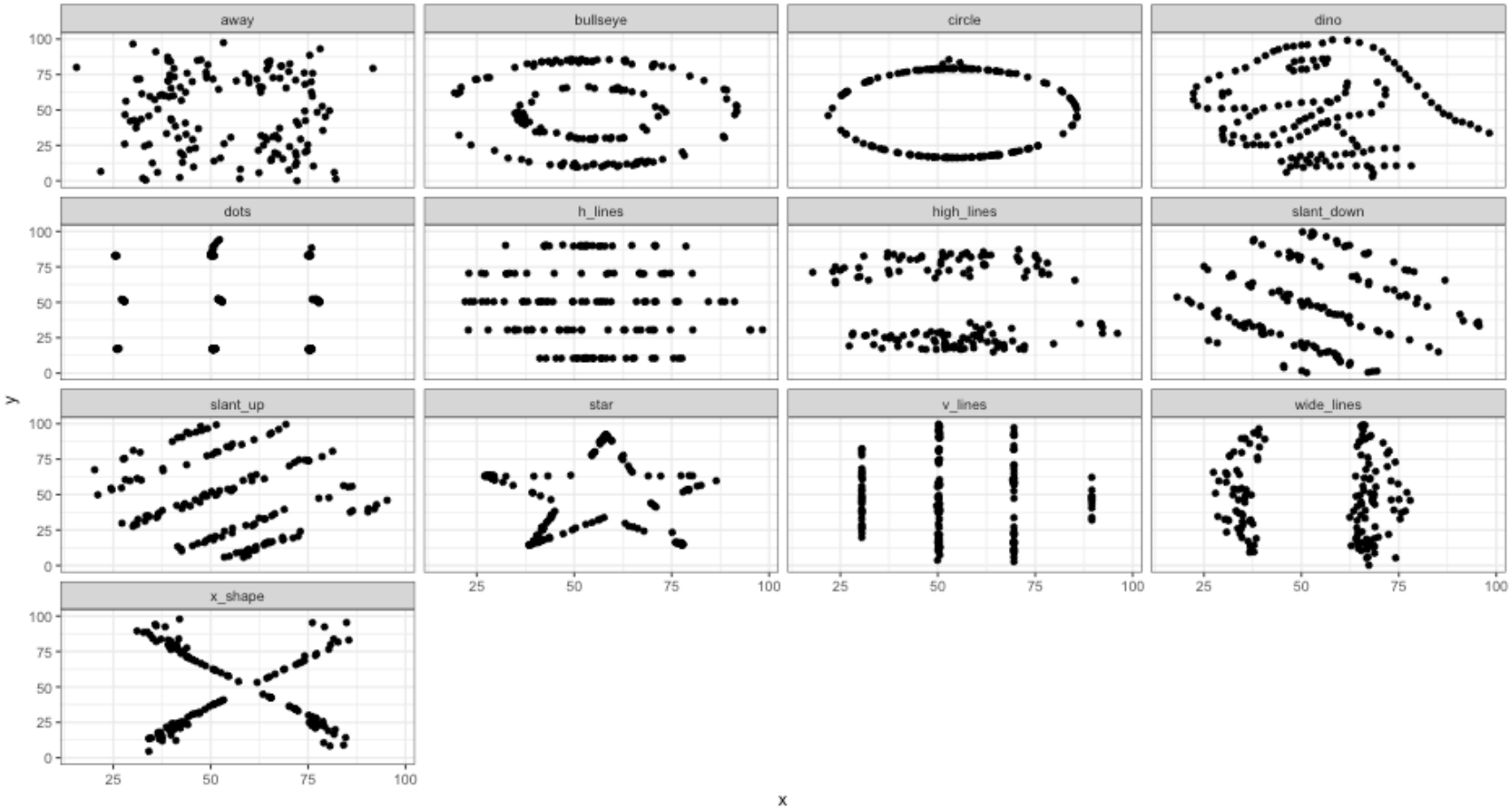
Plot one of the datasets



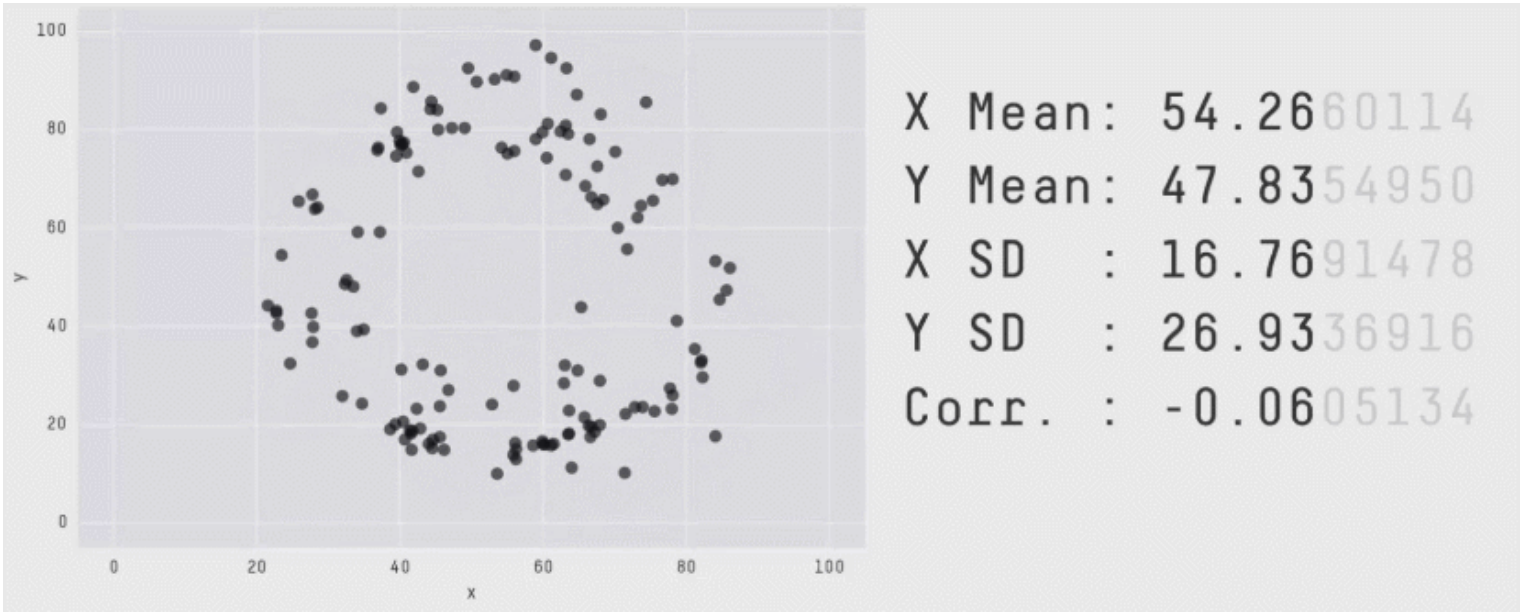
Plot another dataset



Plot all the datasets



In gif form



Matejka & Fitzmaurice (2017). *Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing.*

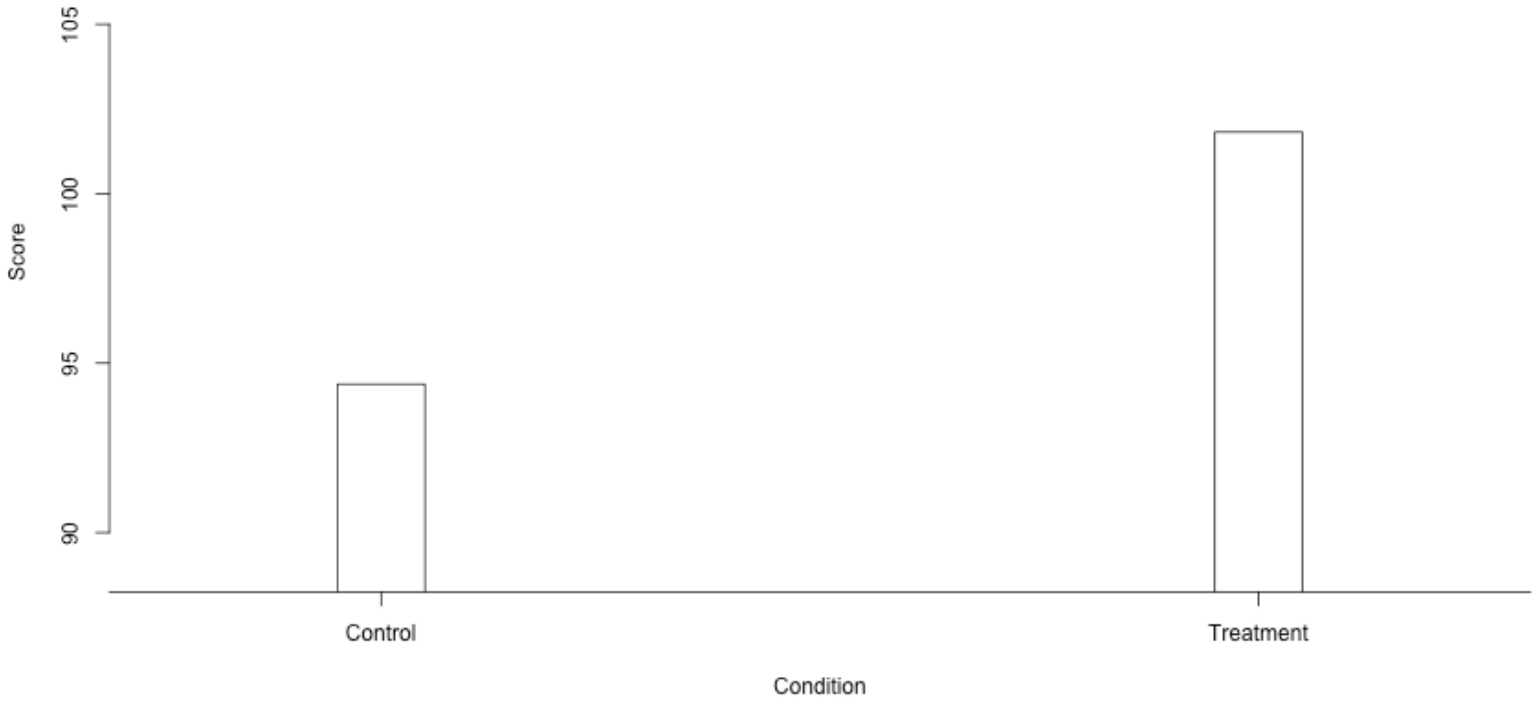
Some applied examples

Hypothetical example

- School wants to try out a new reading intervention
- Work with researchers at the UO to design a study
- Kindergarten students who are behind their peers in literacy are selected
- Randomly assign half the students to the intervention, the rest continue with "typical" instruction
- Now the study is over - how do we tell if it worked? Visualize it! (and other stuff too)

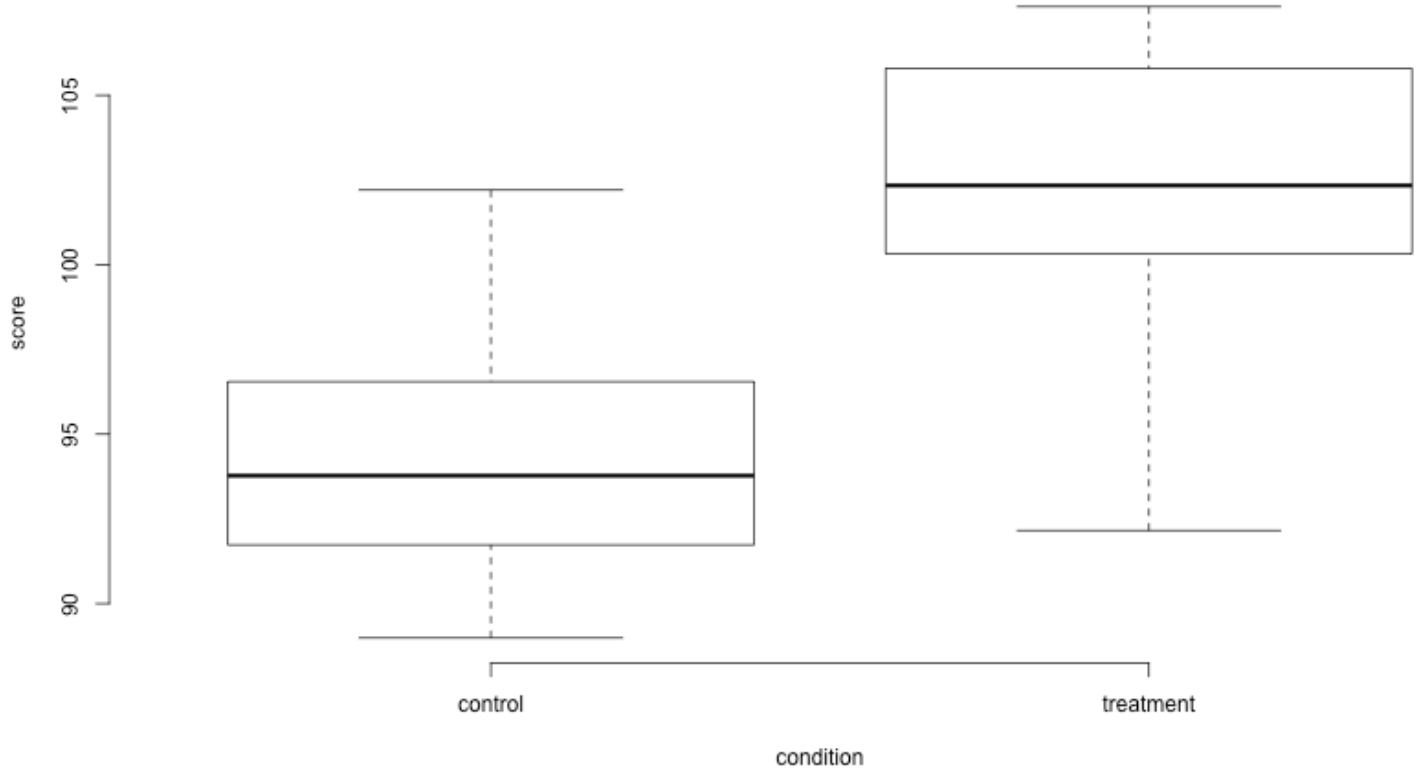
Barplots

(tried and true)



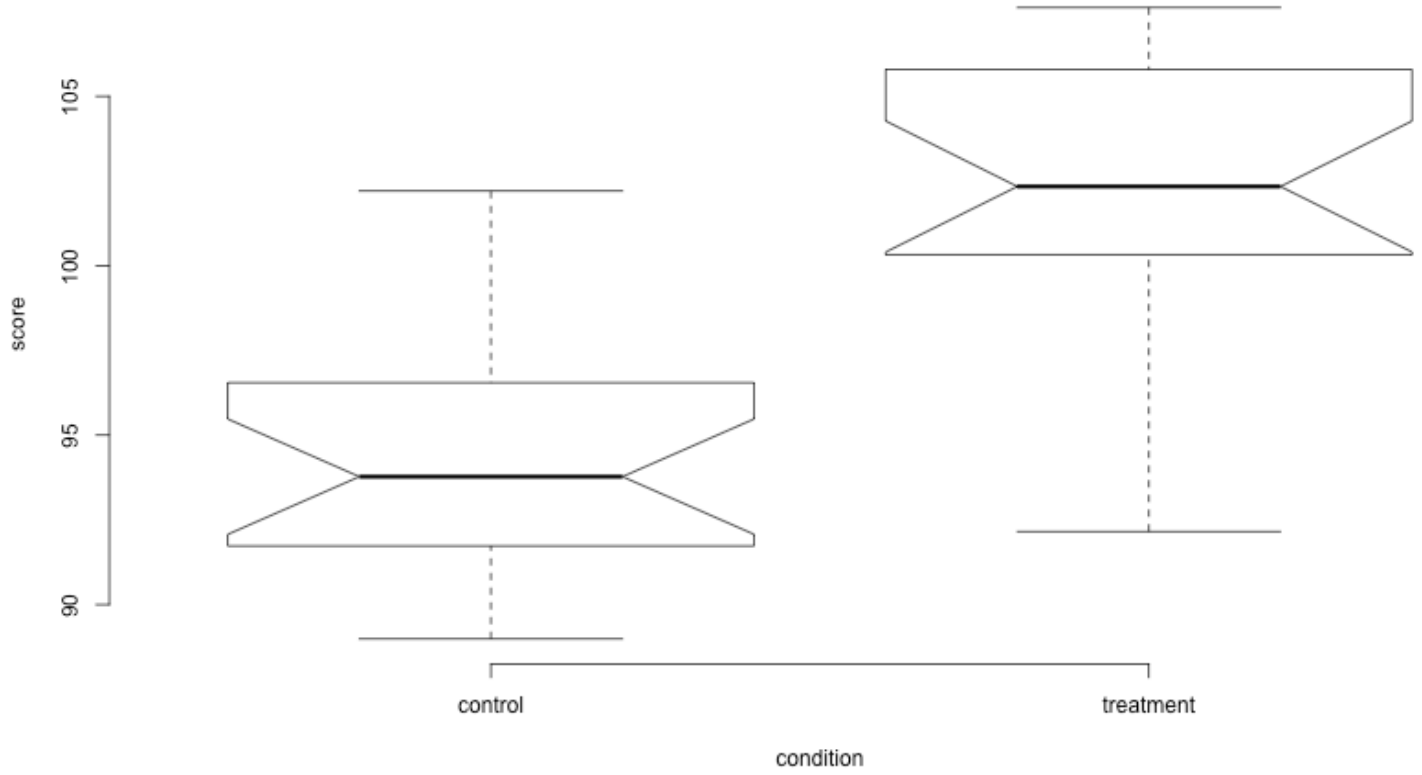
Boxplots

(tried and true)



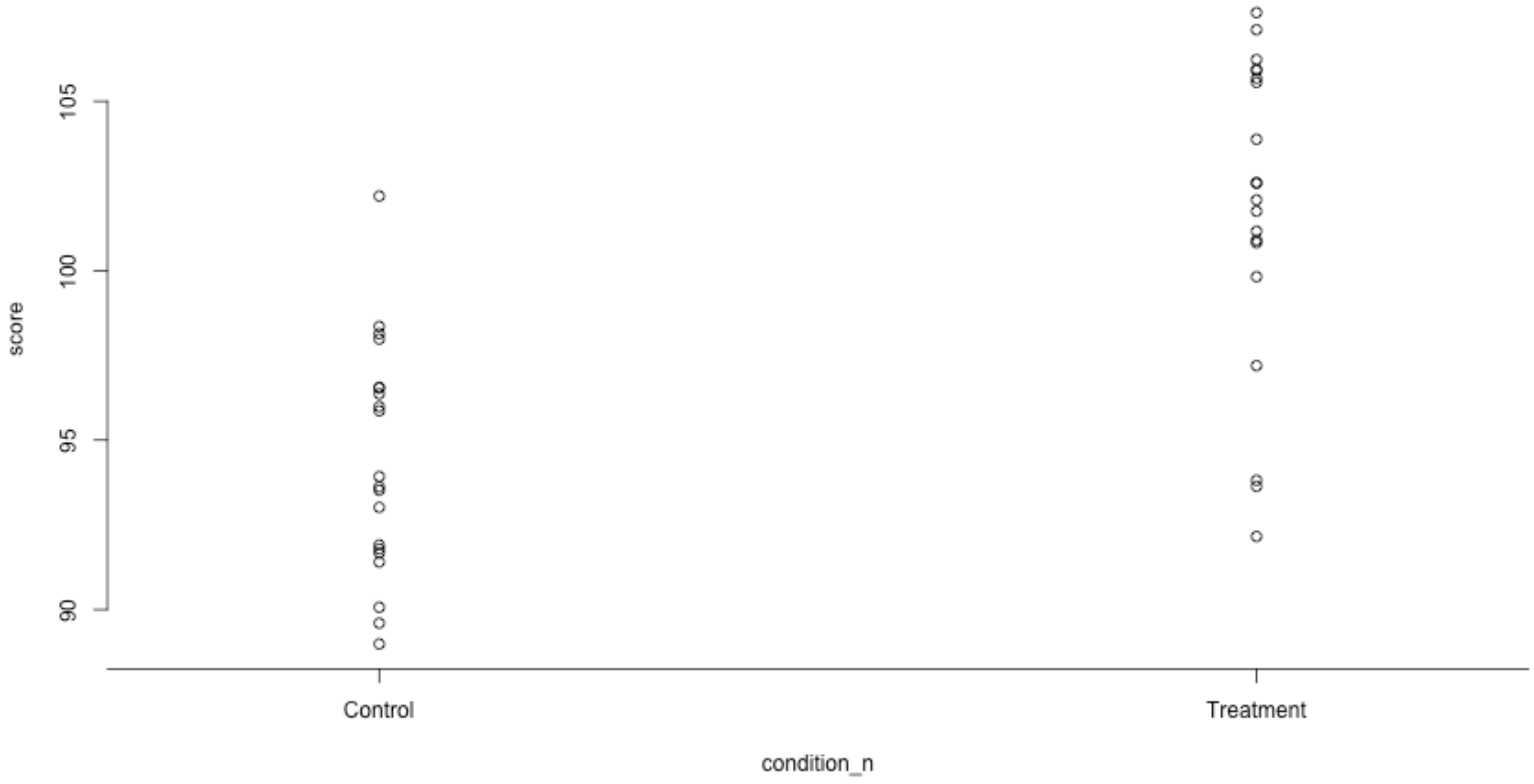
Notched boxplots

(slightly better)



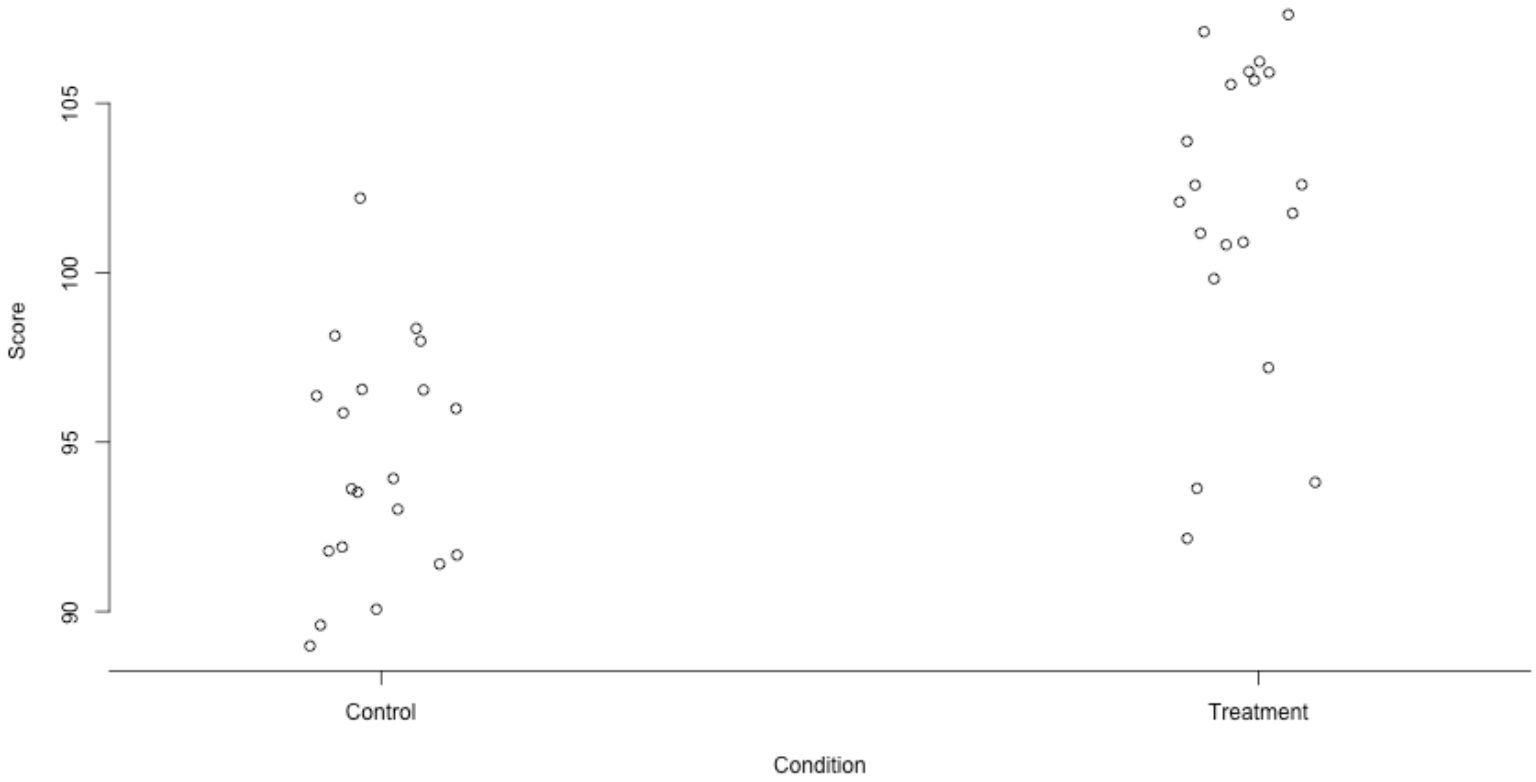
Stripcharts

Show the data!

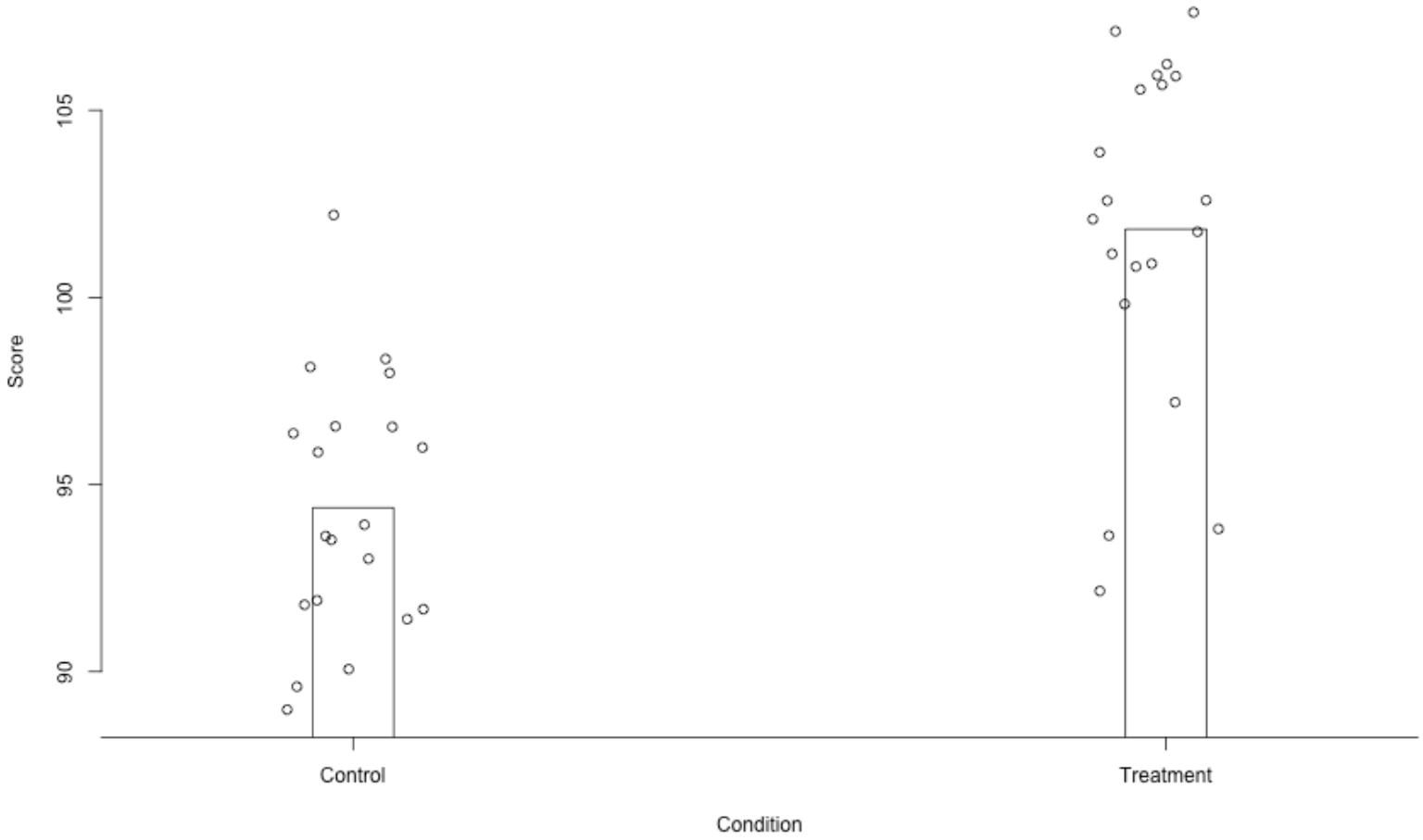


Jittered stripcharts

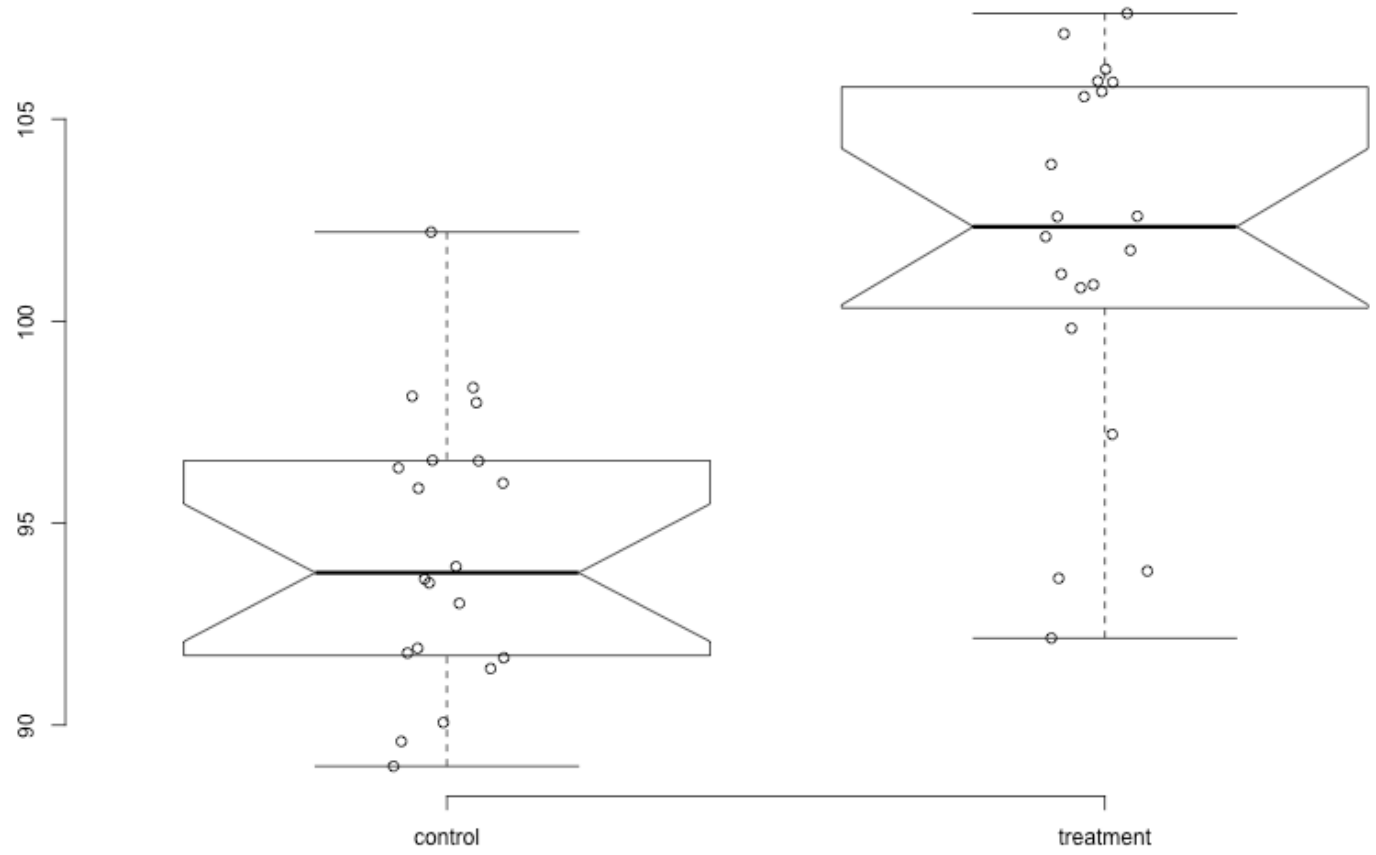
Show the data!



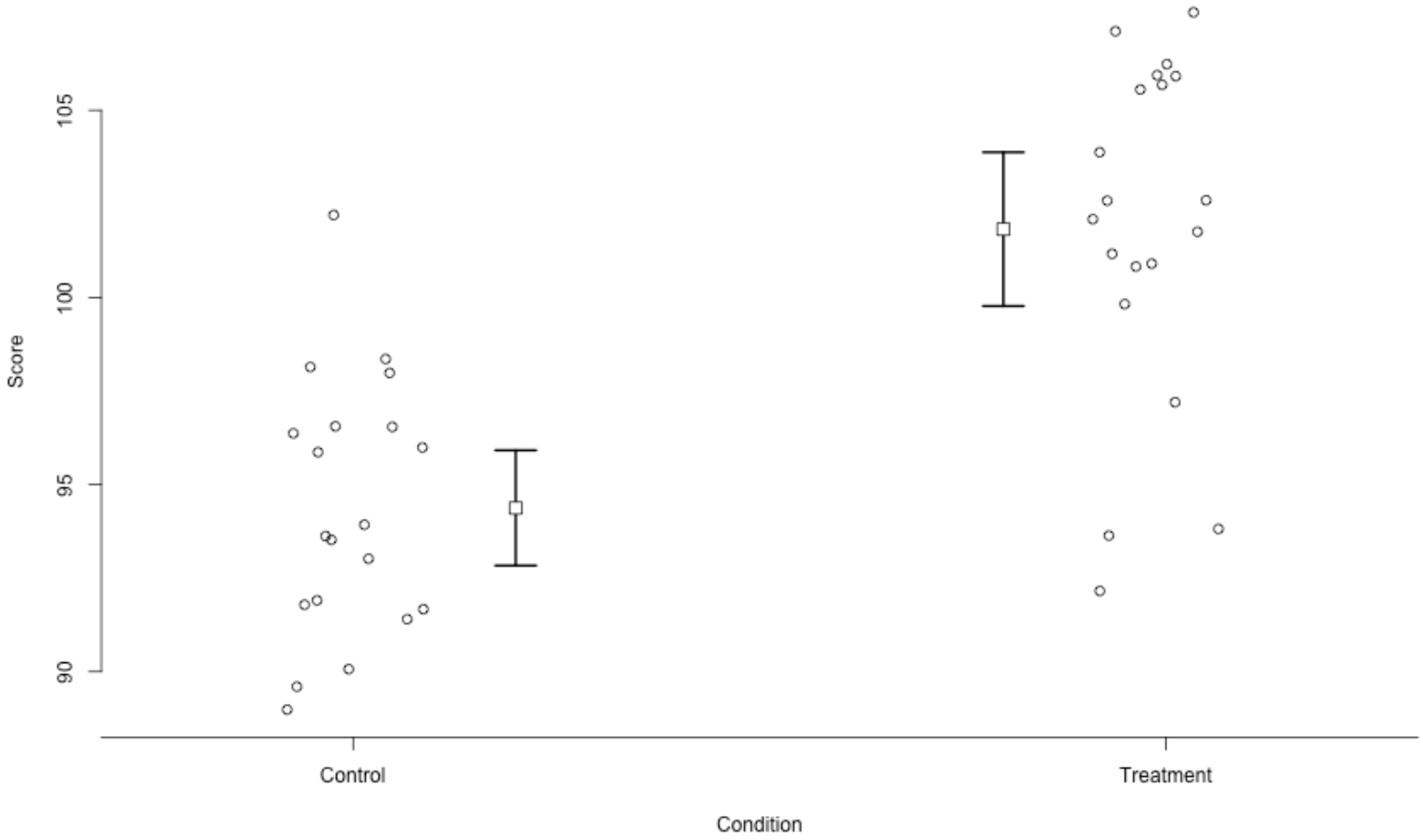
Combine barplots and jittered stripcharts



Combine boxplots and jittered stripcharts



Best?

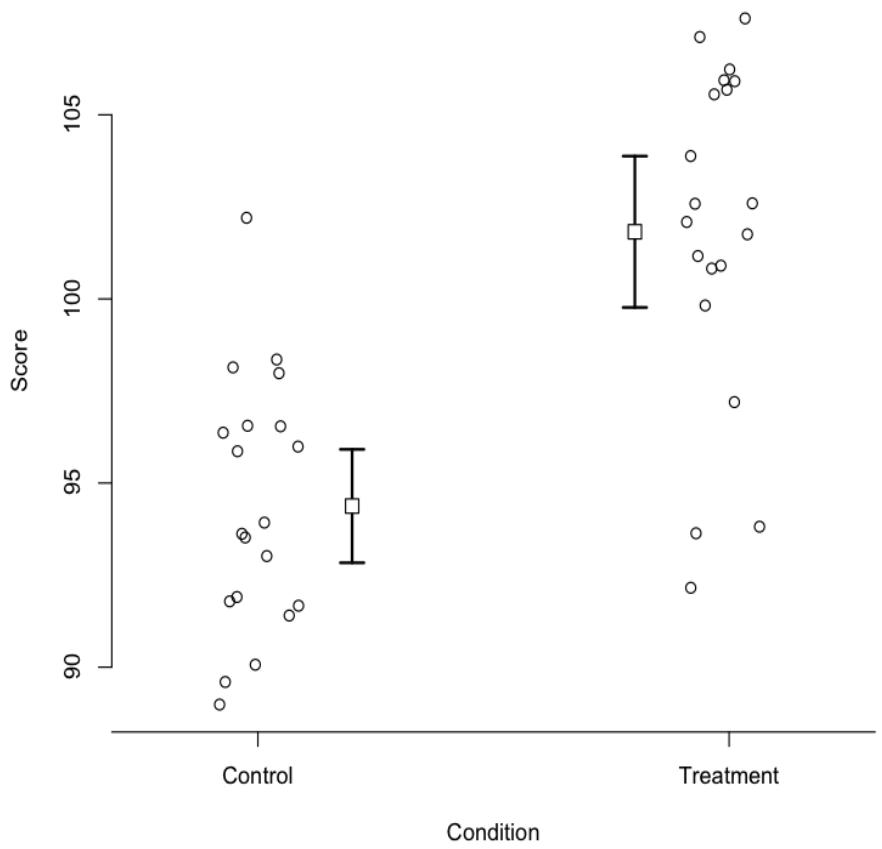


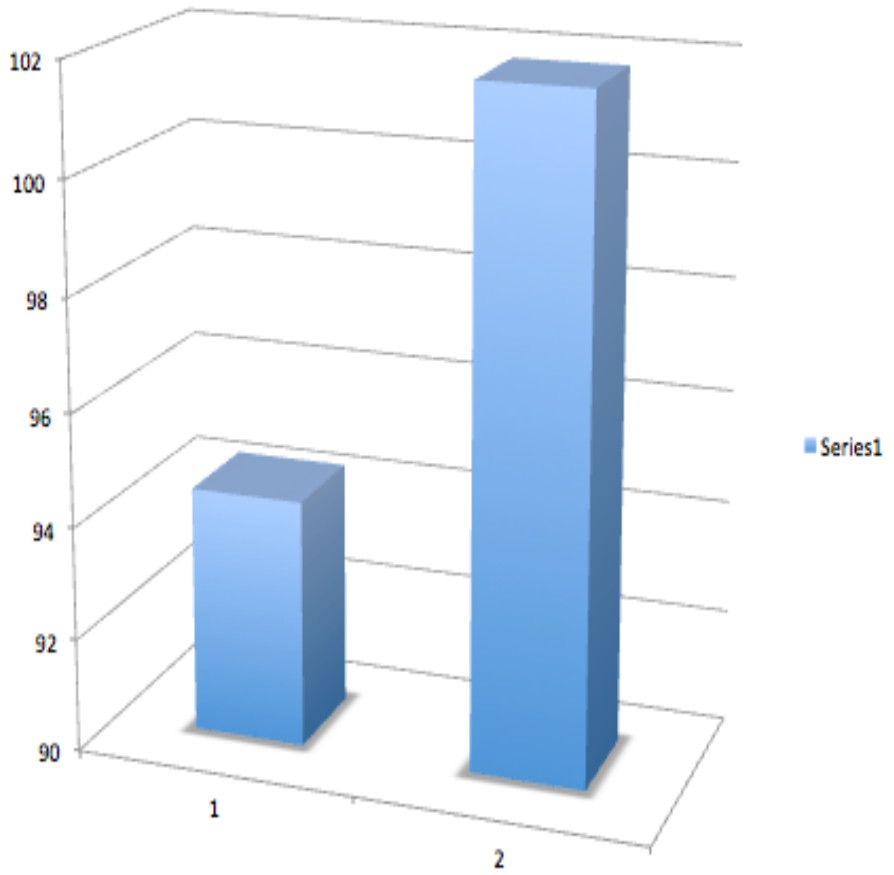
Some things to avoid

- 3D plots
- Pie charts
- Dual axes
- Restricted axes
- Unnecessary frills (colors, etc)
    - Show the data as plainly as possible. Let the data speak!

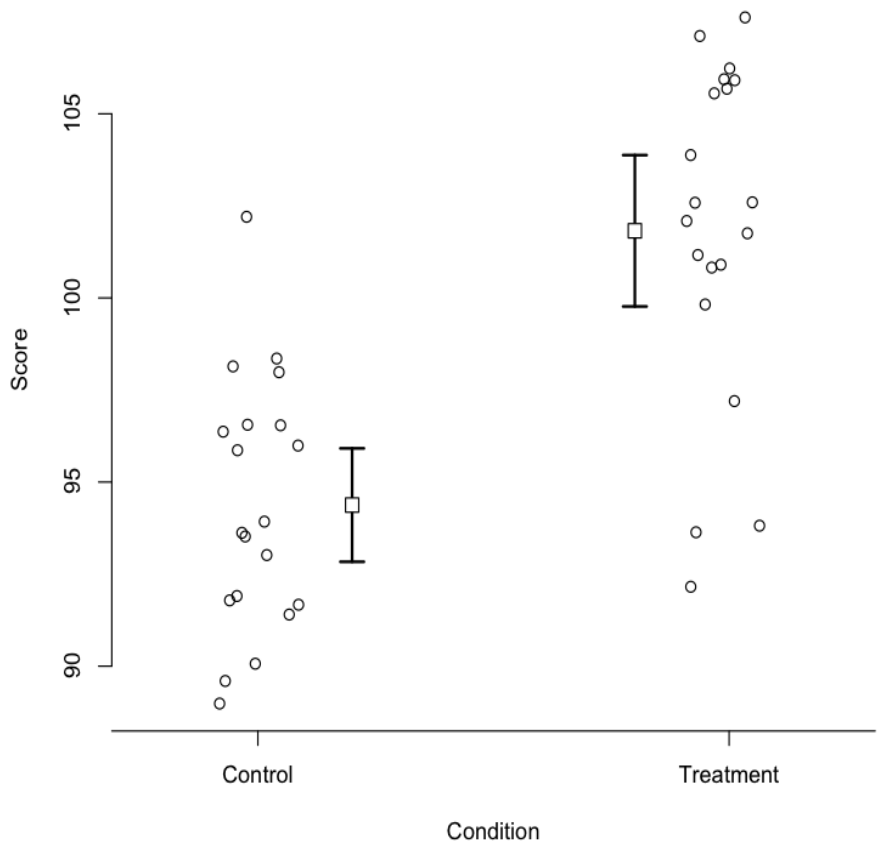
NOTE: The following 10 slides (and the previous plot) inspired/taken from Karl Broman's presentation on graphs (see here)

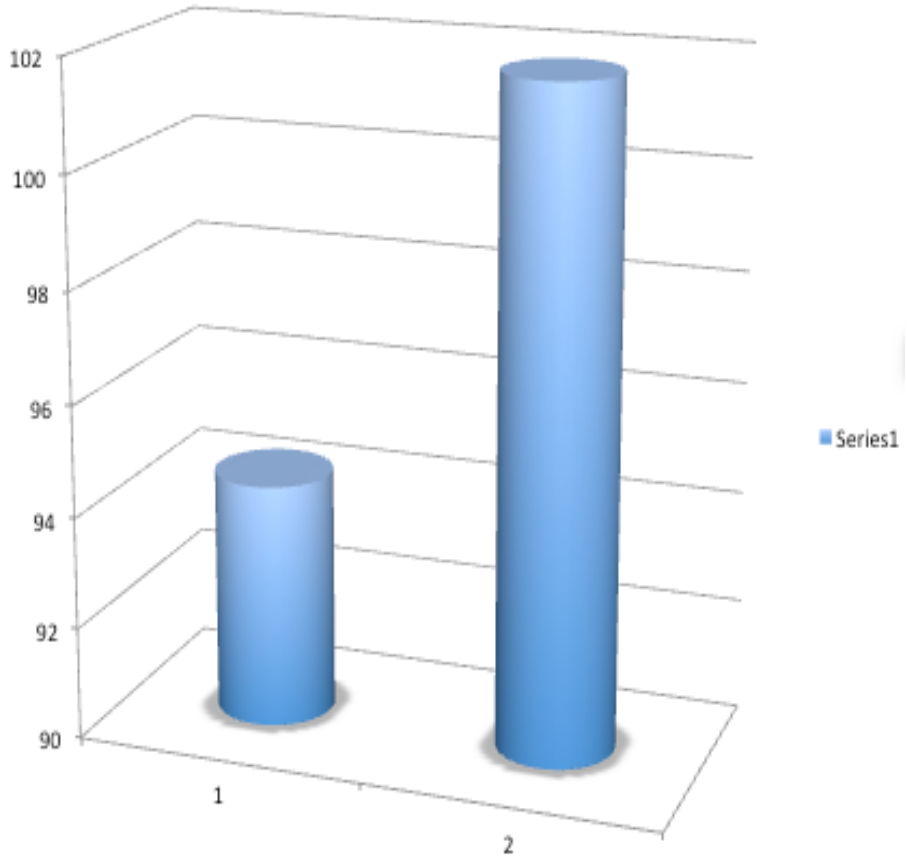
Examples


Score
105
100
95
90
Control
Treatment
Condition

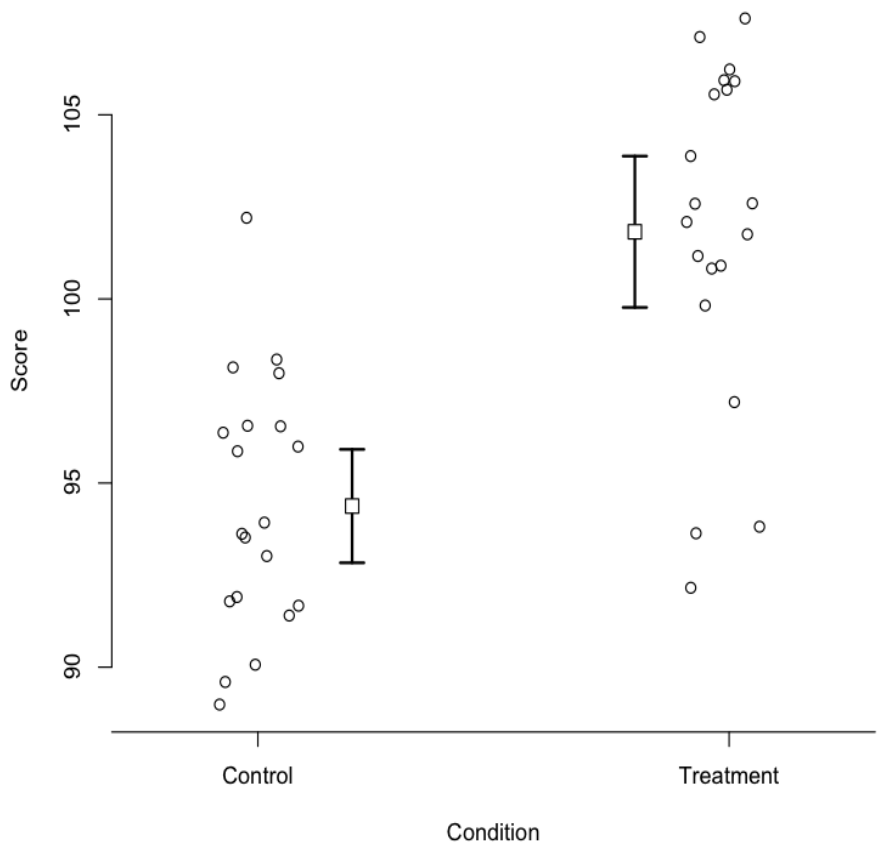

102
100
98
96
94
92
90
1
2
Series1

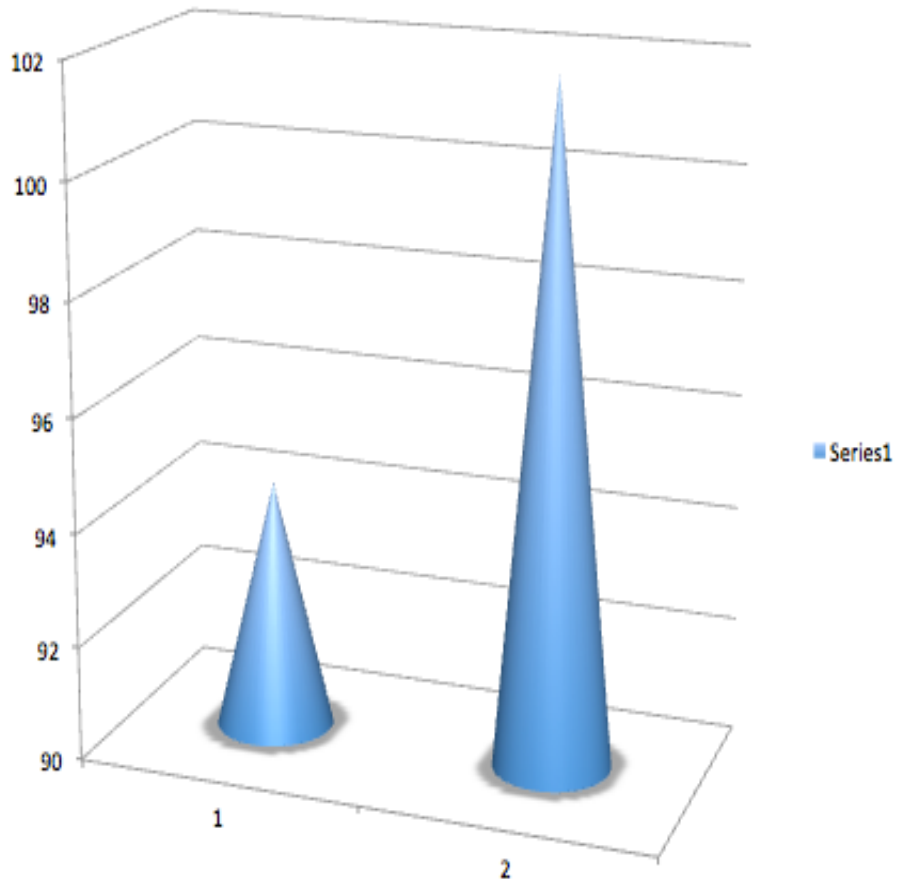
Examples


Score
105
100
95
90
Control
Treatment
Condition

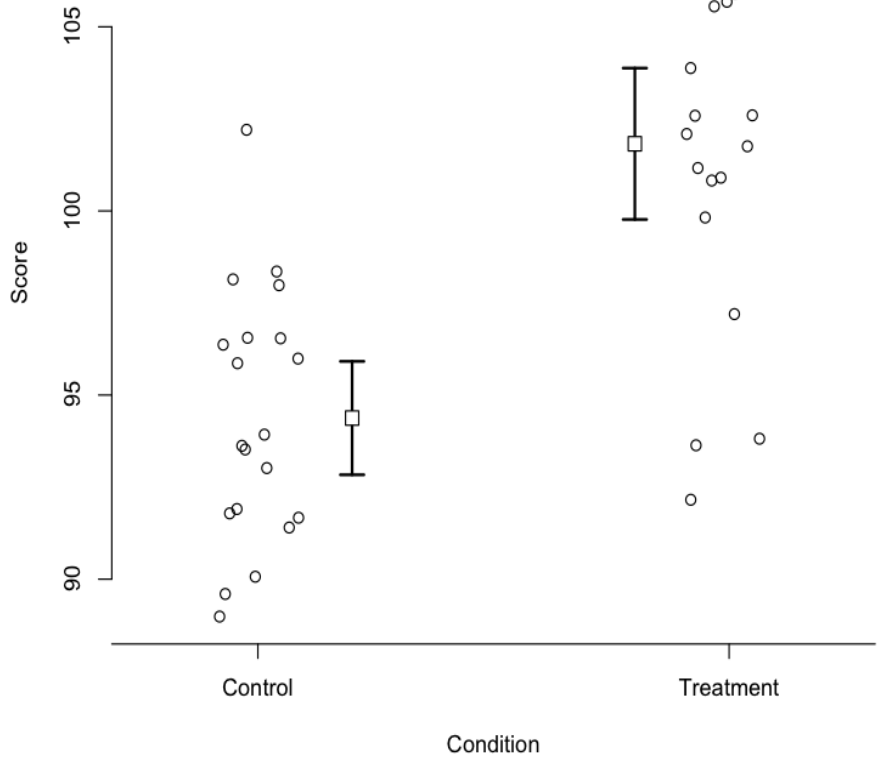

102
100
98
96
94
92
90
1
2
Series1

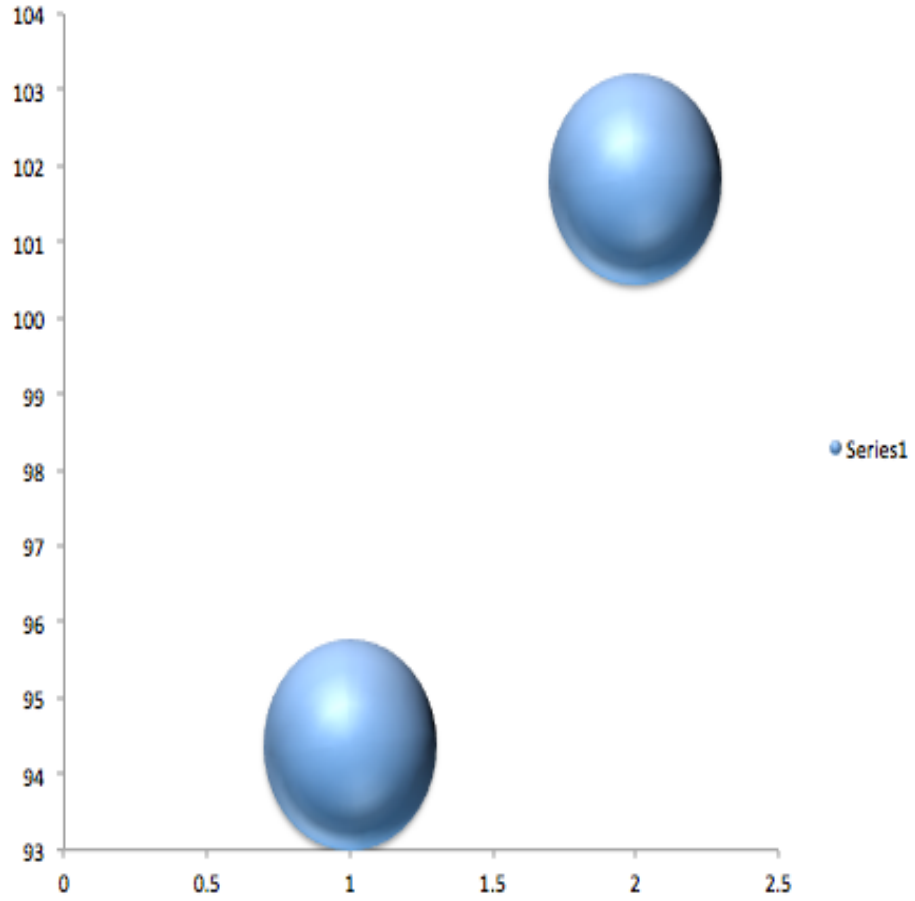
Examples


Score
105
100
95
90
Control
Treatment
Condition

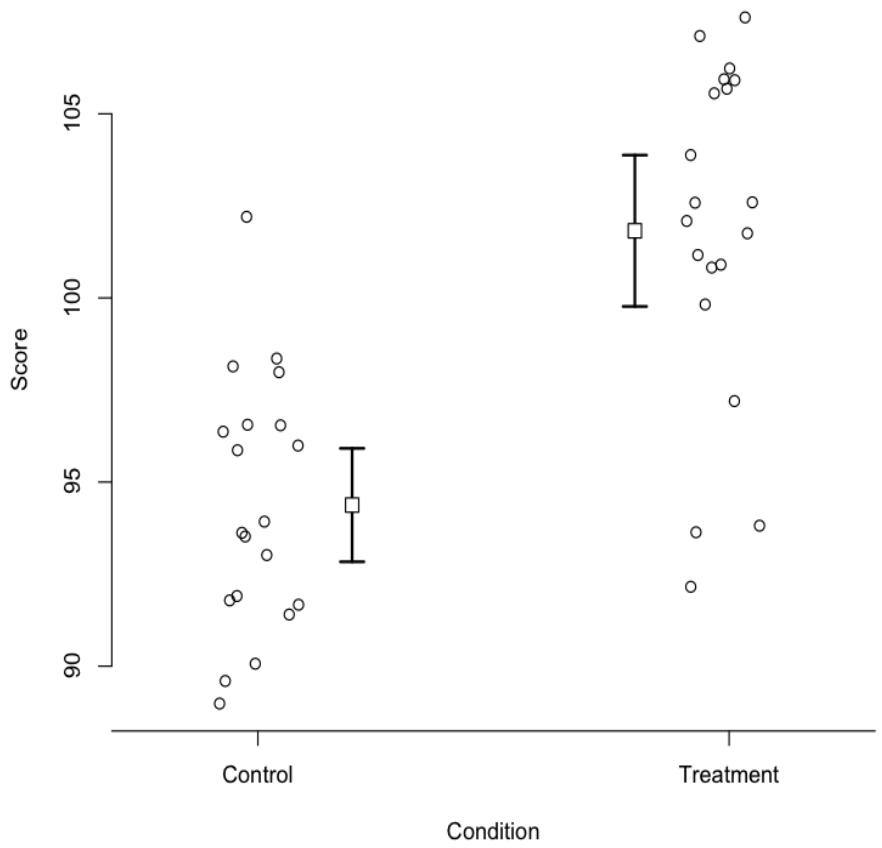

102
100
98
96
94
92
90
1
2
Series1

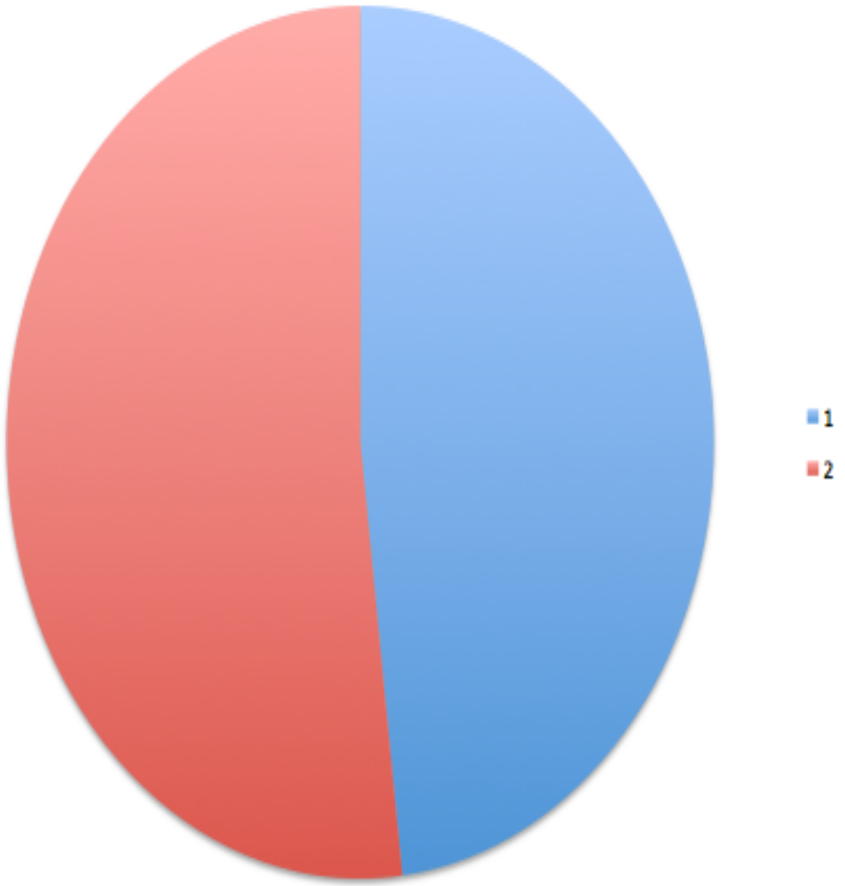
Examples


Score
105
100
95
90
Control
Treatment
Condition

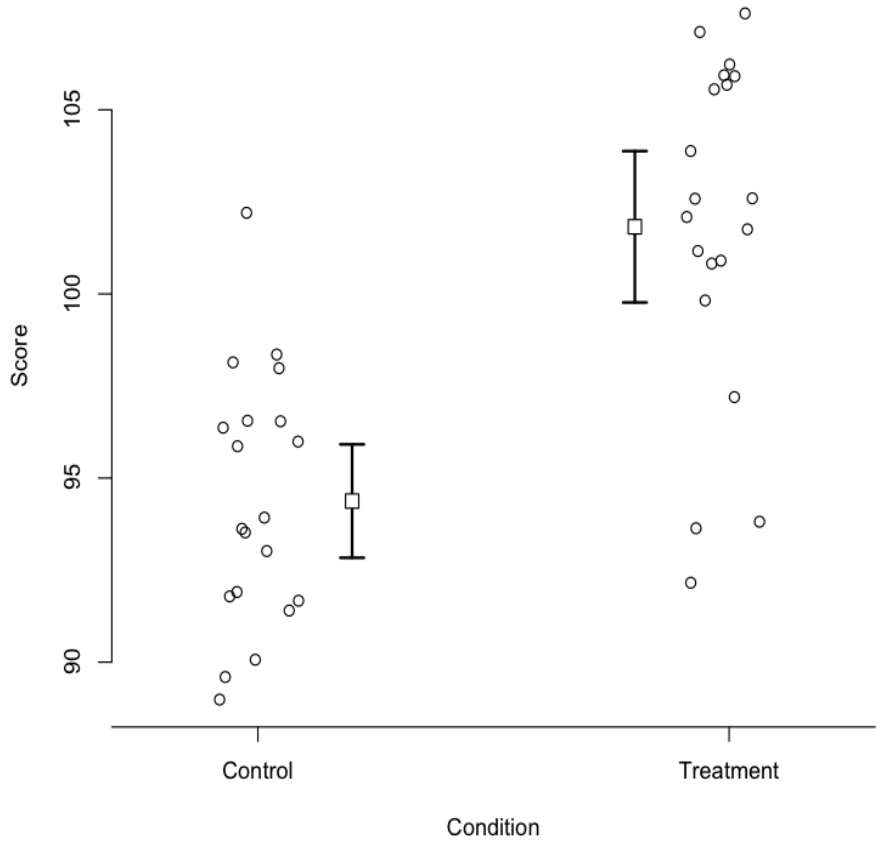

104
103
102
101
100
99
98
97
96
95
94
93
0
0.5
1
1.5
2
2.5
Series1

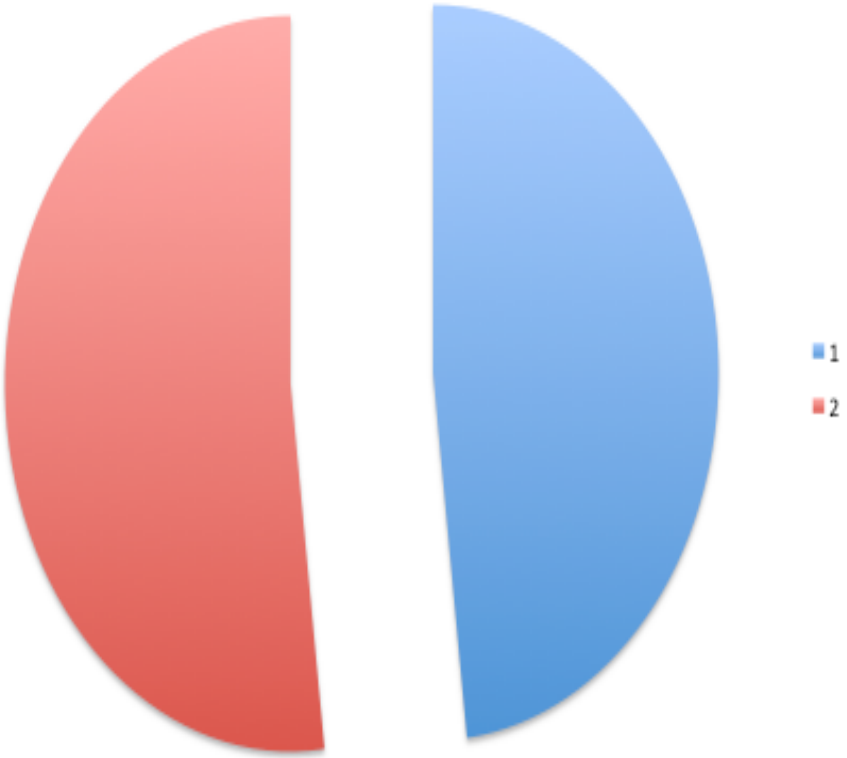
Examples



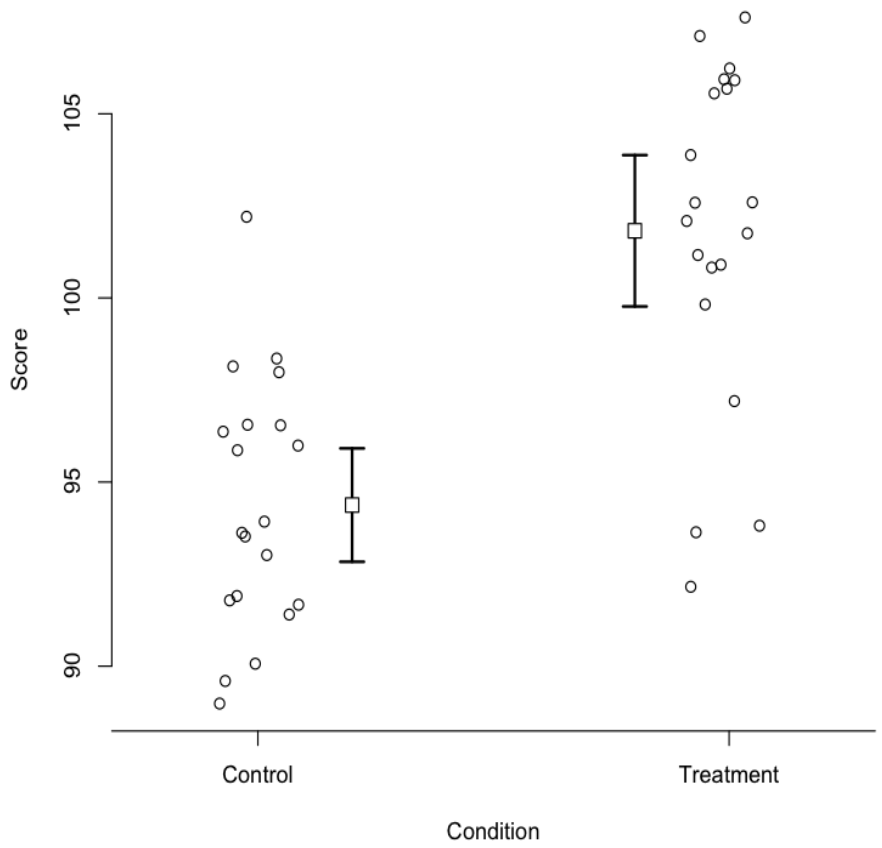


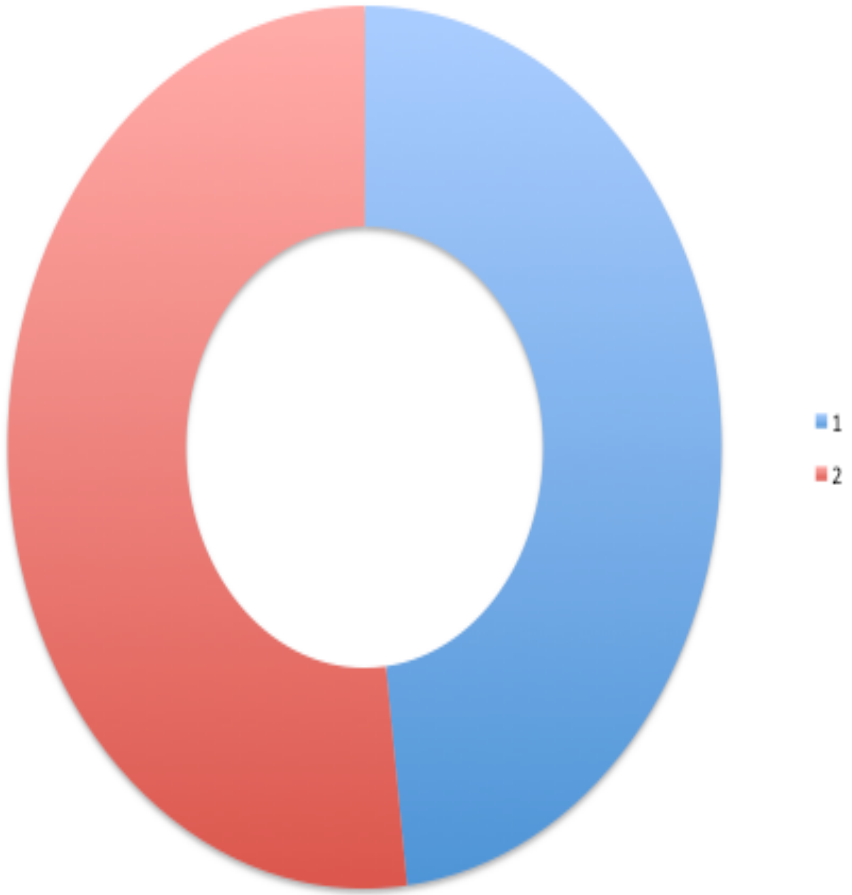
Examples



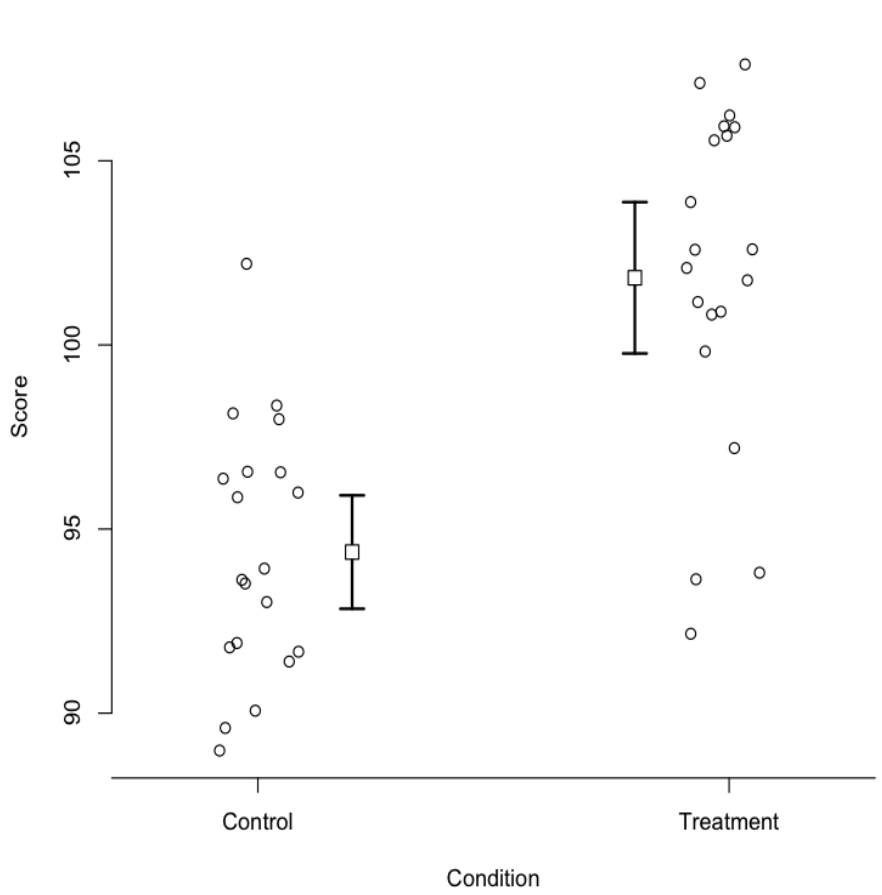


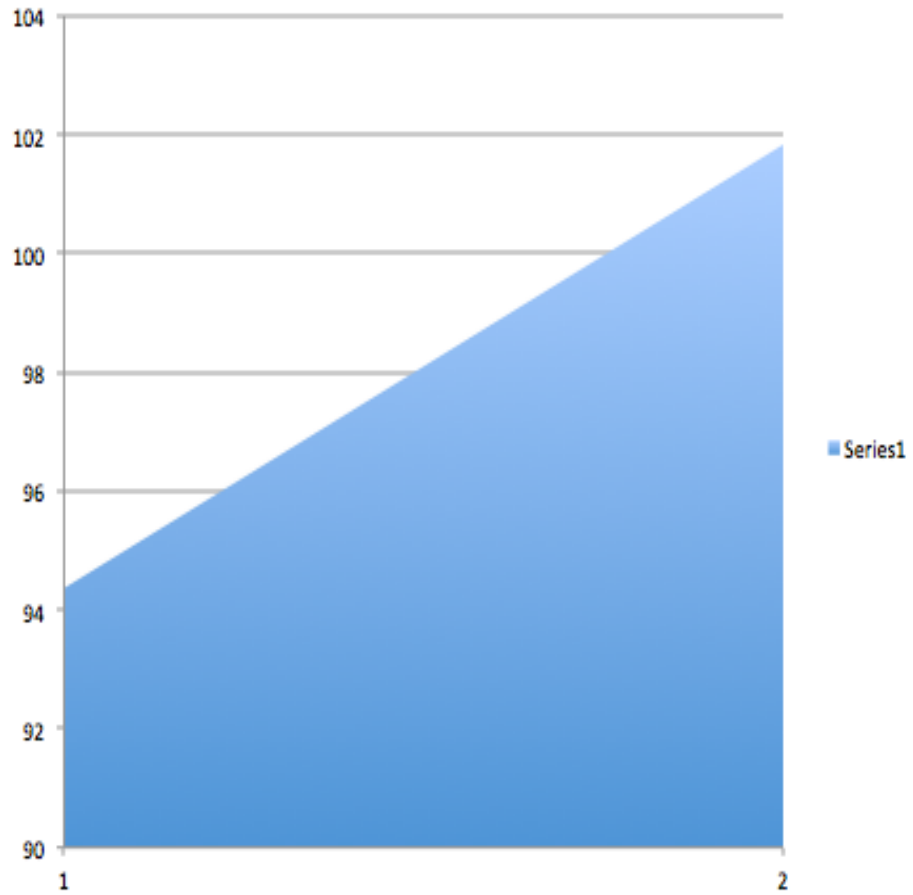
Examples



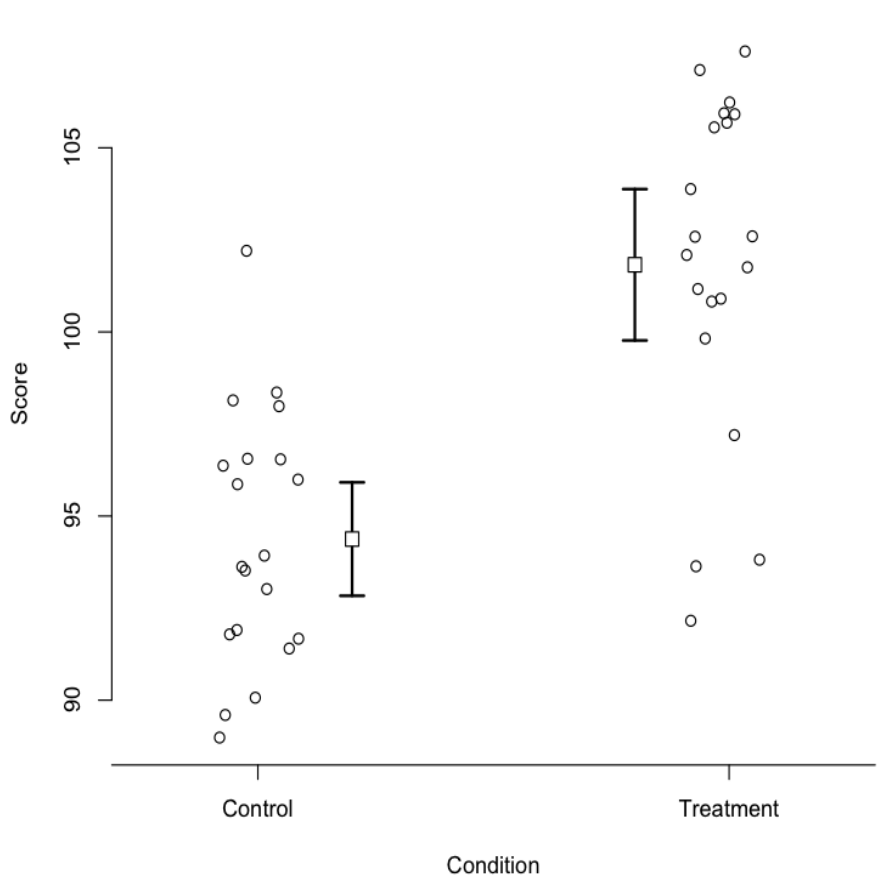


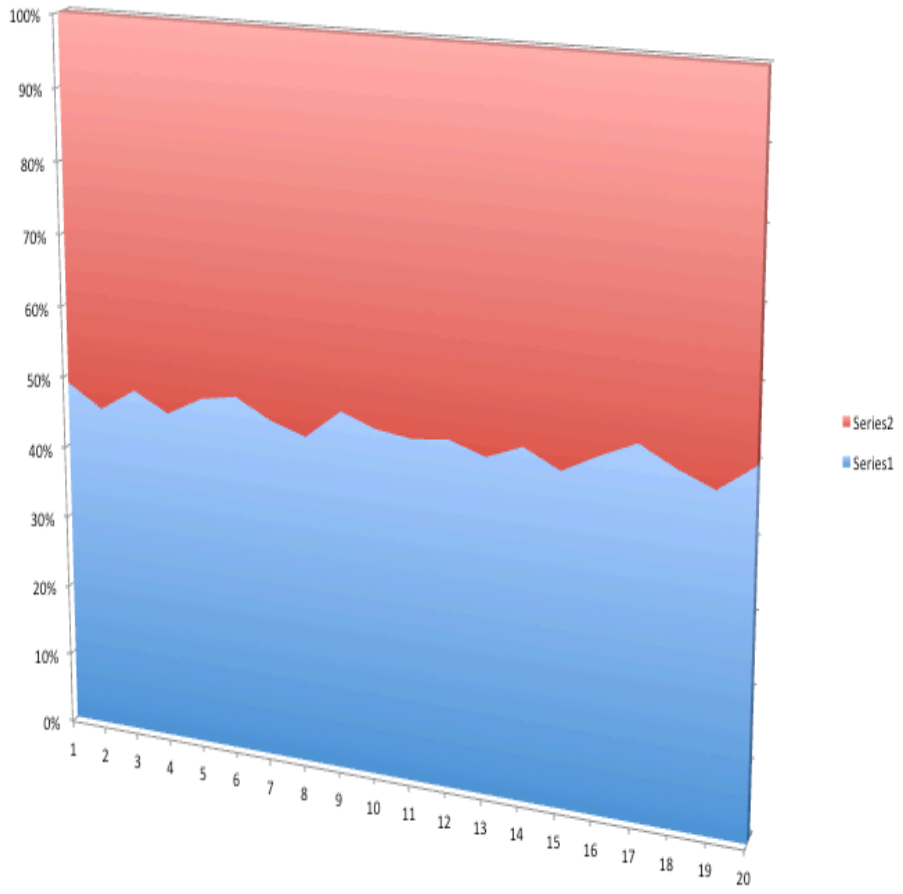
Examples


Score
105
100
95
90
Control
Treatment
Condition

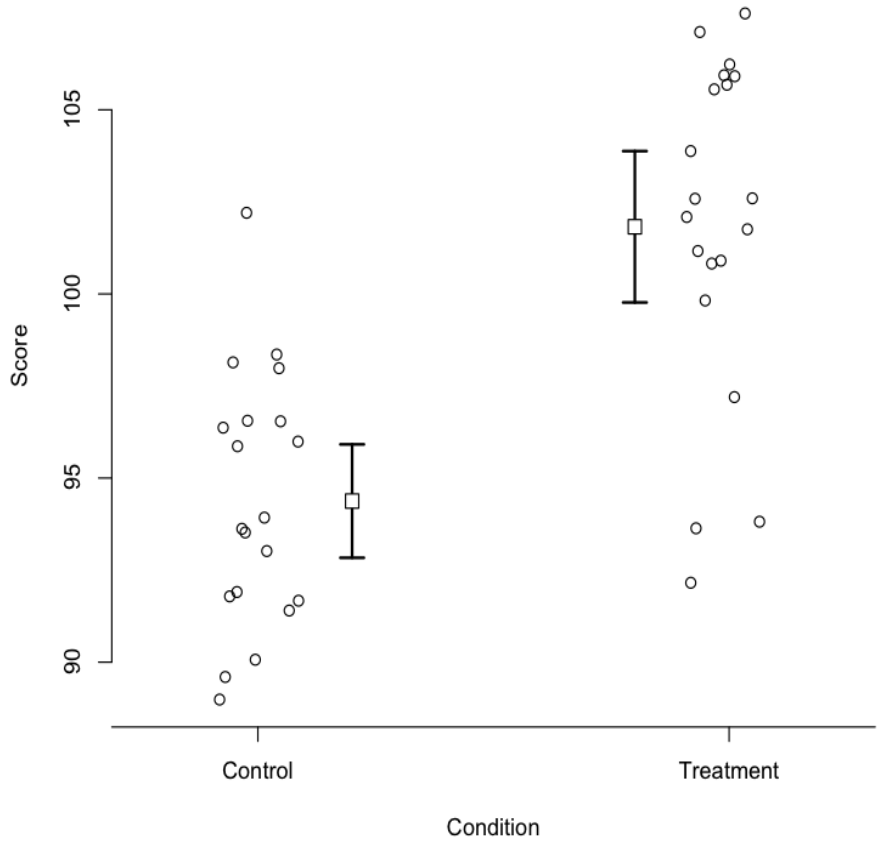

104
102
100
98
96
94
92
90
1
2
Series1

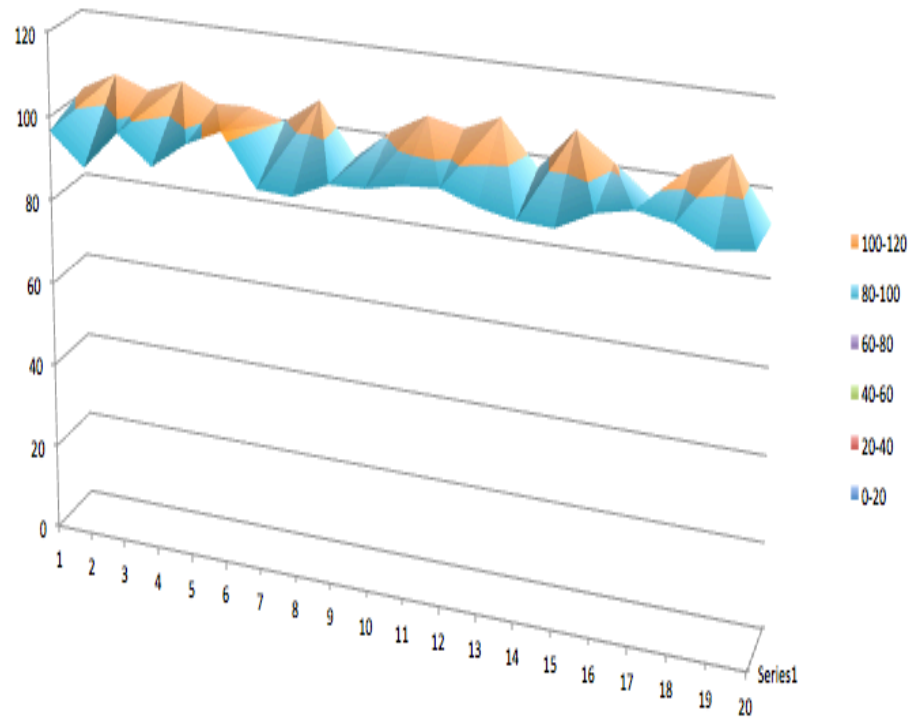
Examples


Score
105
100
95
90
Control
Treatment
Condition


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Series2
Series1

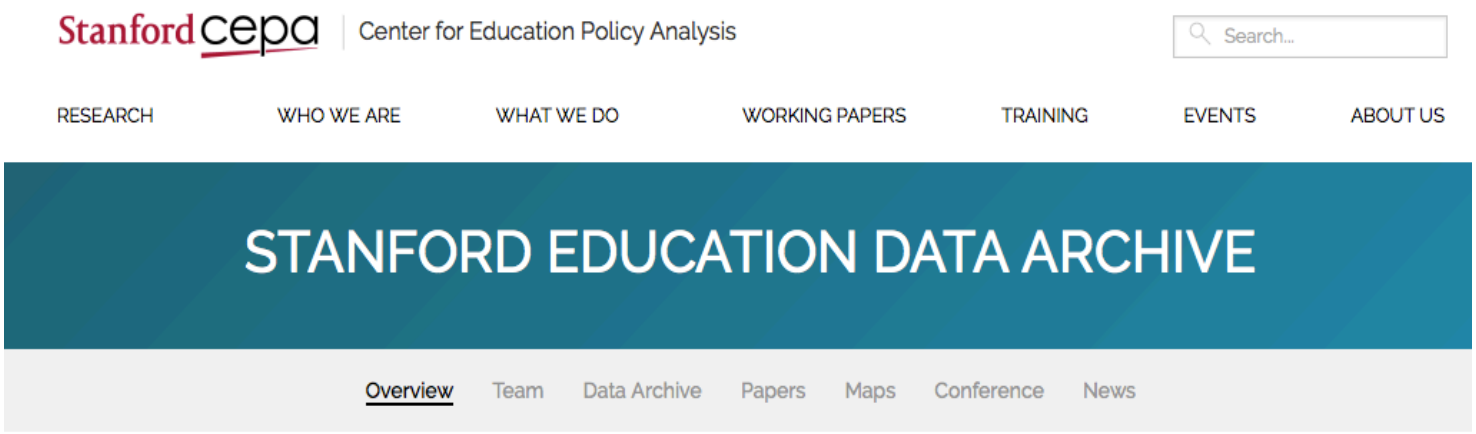
Examples


Score
105
100
95
90
Control
Treatment
Condition


120
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Series1
100-120
80-100
60-80
40-60
20-40
0-20

Some great examples: SEDA

Sean Reardon: https://cepa.stanford.edu/seda/overview

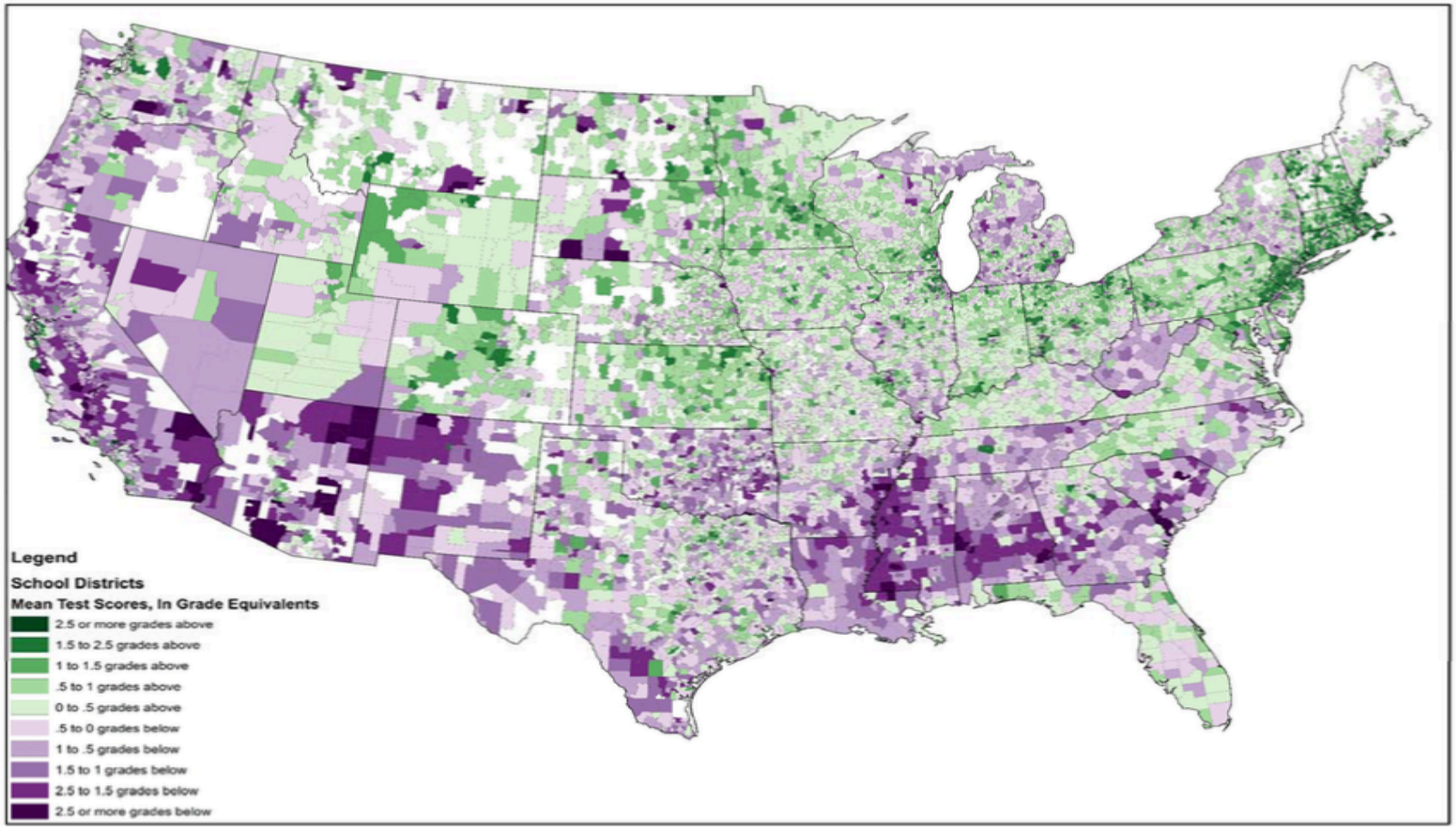


Racial, socioeconomic, and gender disparities in academic performance and educational attainment are stubborn features of the U.S. educational system. These disparities are neither inevitable nor immutable, however. They have been produced by—and so may also be reduced by—a welter of social and economic policies, social norms and patterns of interaction, and the organization of American schooling.

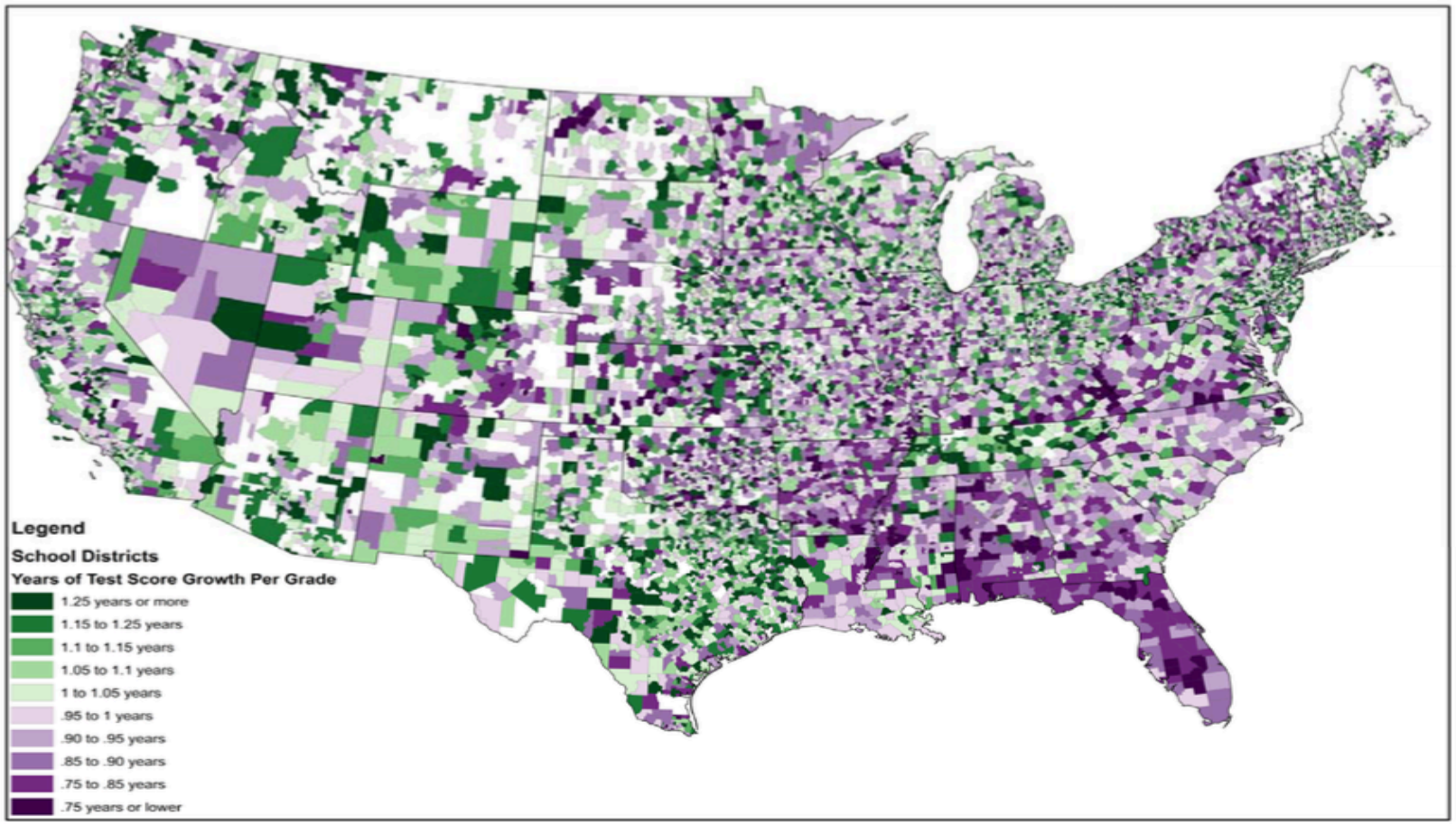
The Stanford Education Data Archive (SEDA) is an initiative aimed at harnessing data to help us—scholars, policymakers, educators, parents—learn how to improve educational opportunity for all children. We are making the data files public so that anyone who is interested can obtain detailed information about

Sean F. Reardon (Stanford University). *The Landscape of U.S. Educational Inequality.* Download presentation

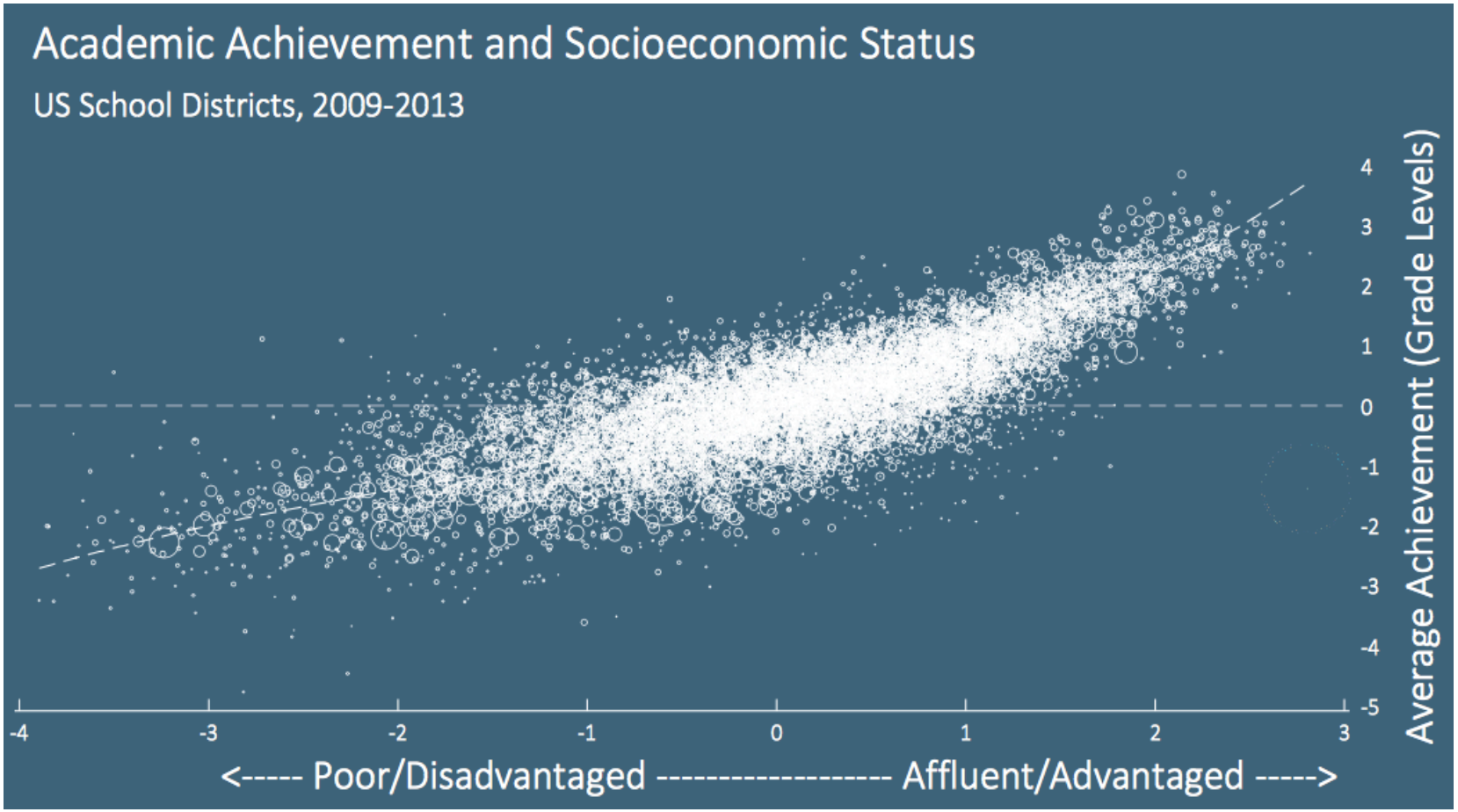
Means by district



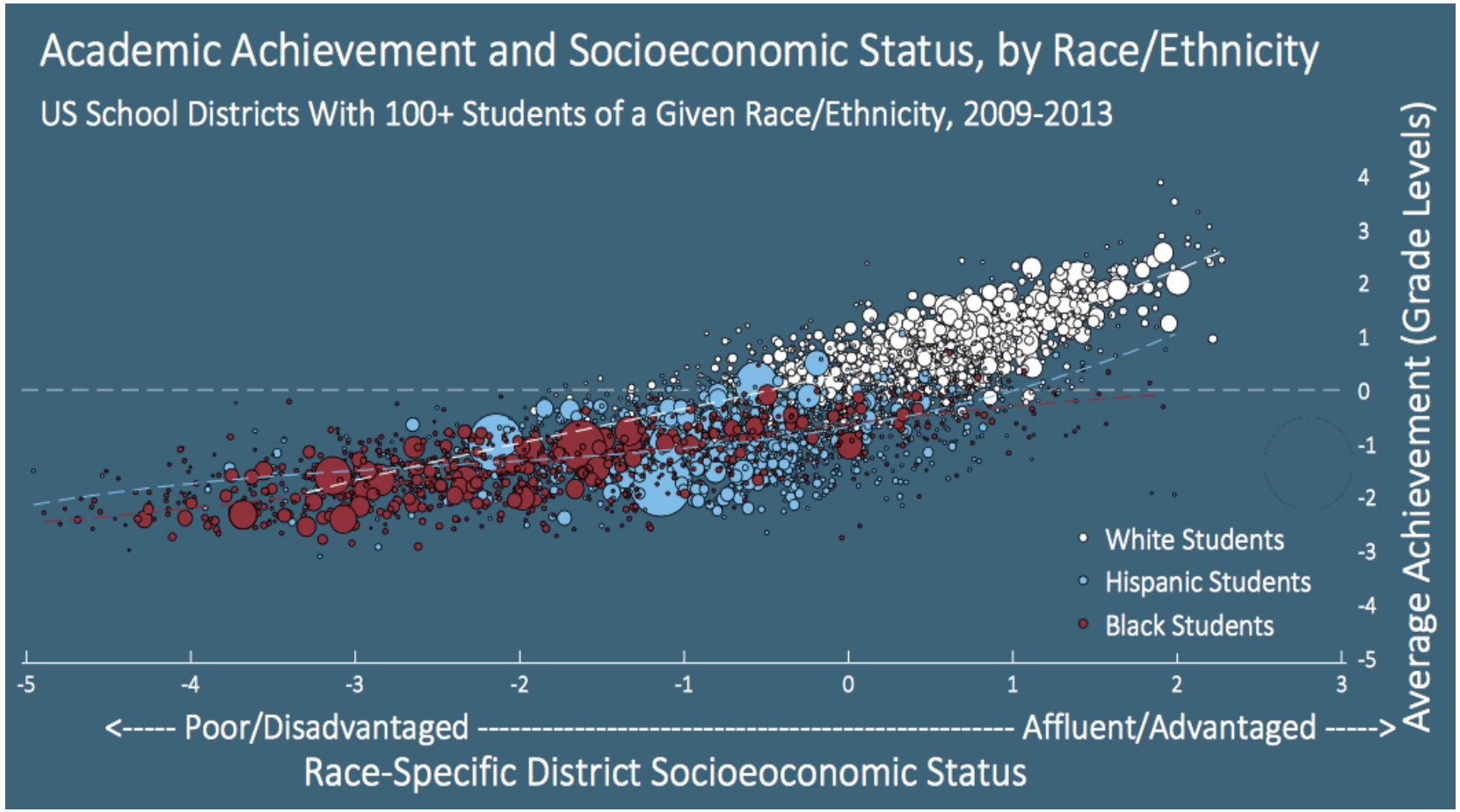
Average gains by district



Mean scores and SES



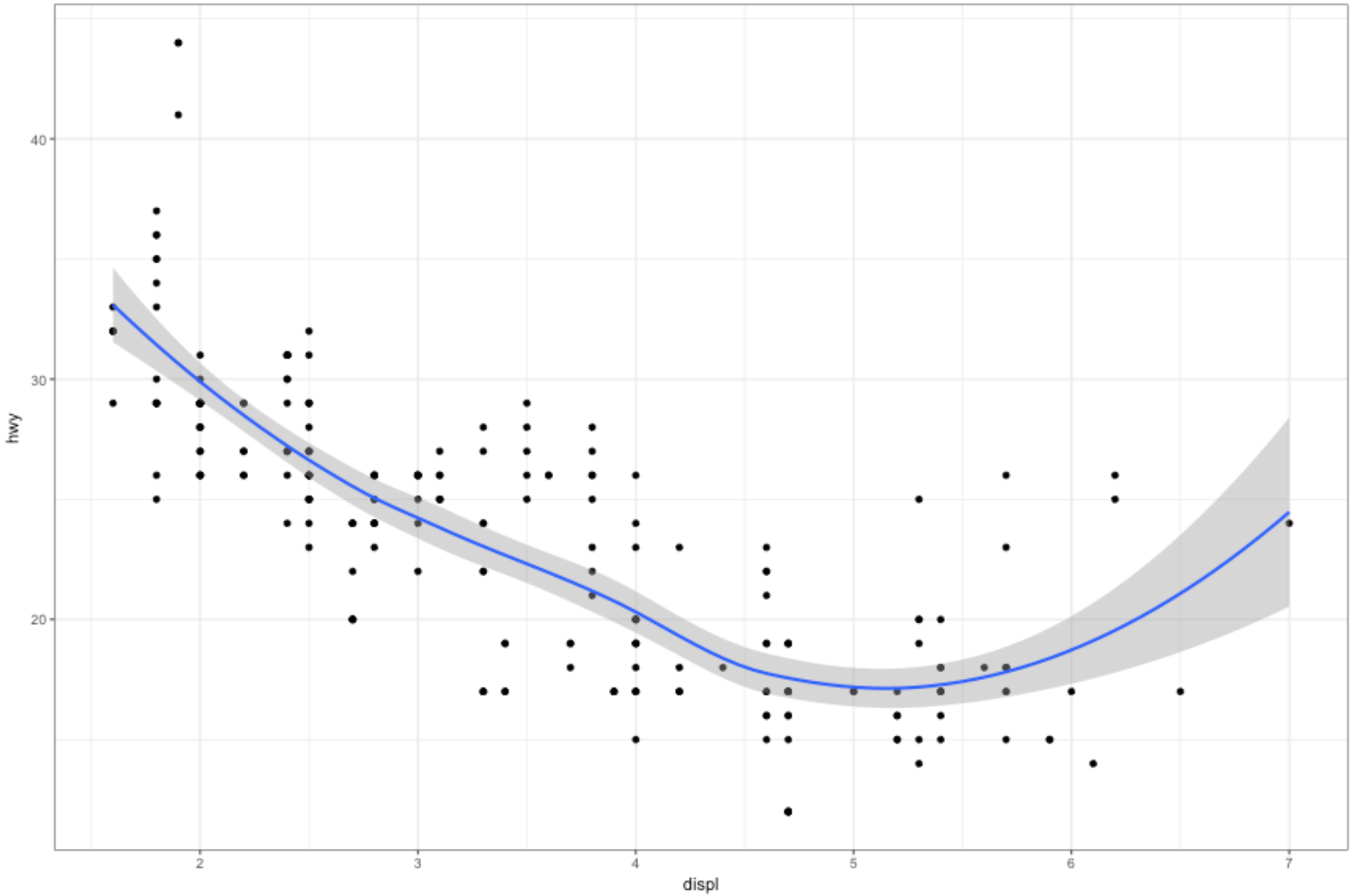
Mean scores and SES by Race/Ethniciy



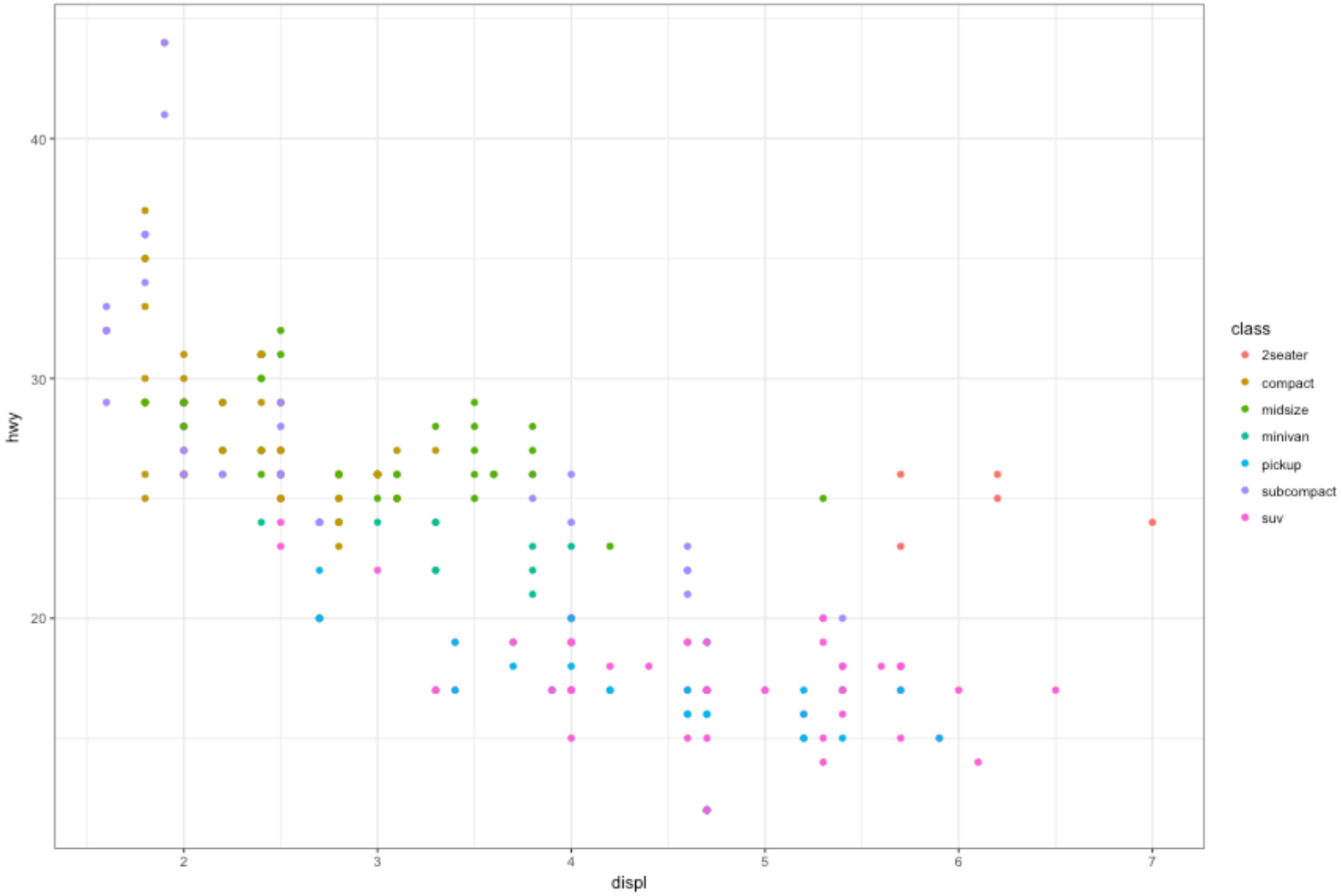
Other examples: Visualizing scale

- Space stuff: http://imgur.com/a/lGabv
- Time: https://www.businessinsider.com.au/animated-timeline-earth-history-2015-11

Some *ggplot* examples

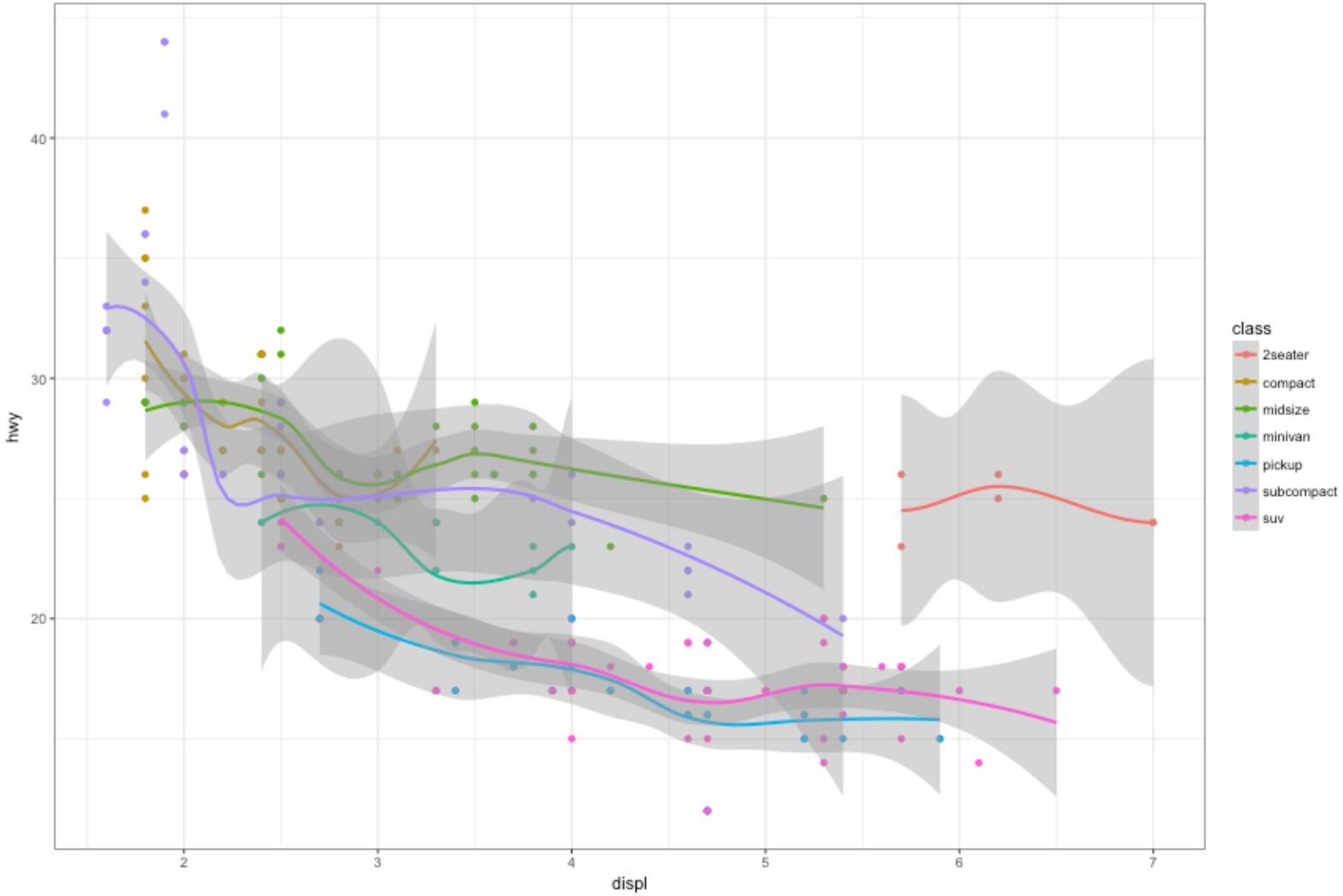


Add an additional aesthetic

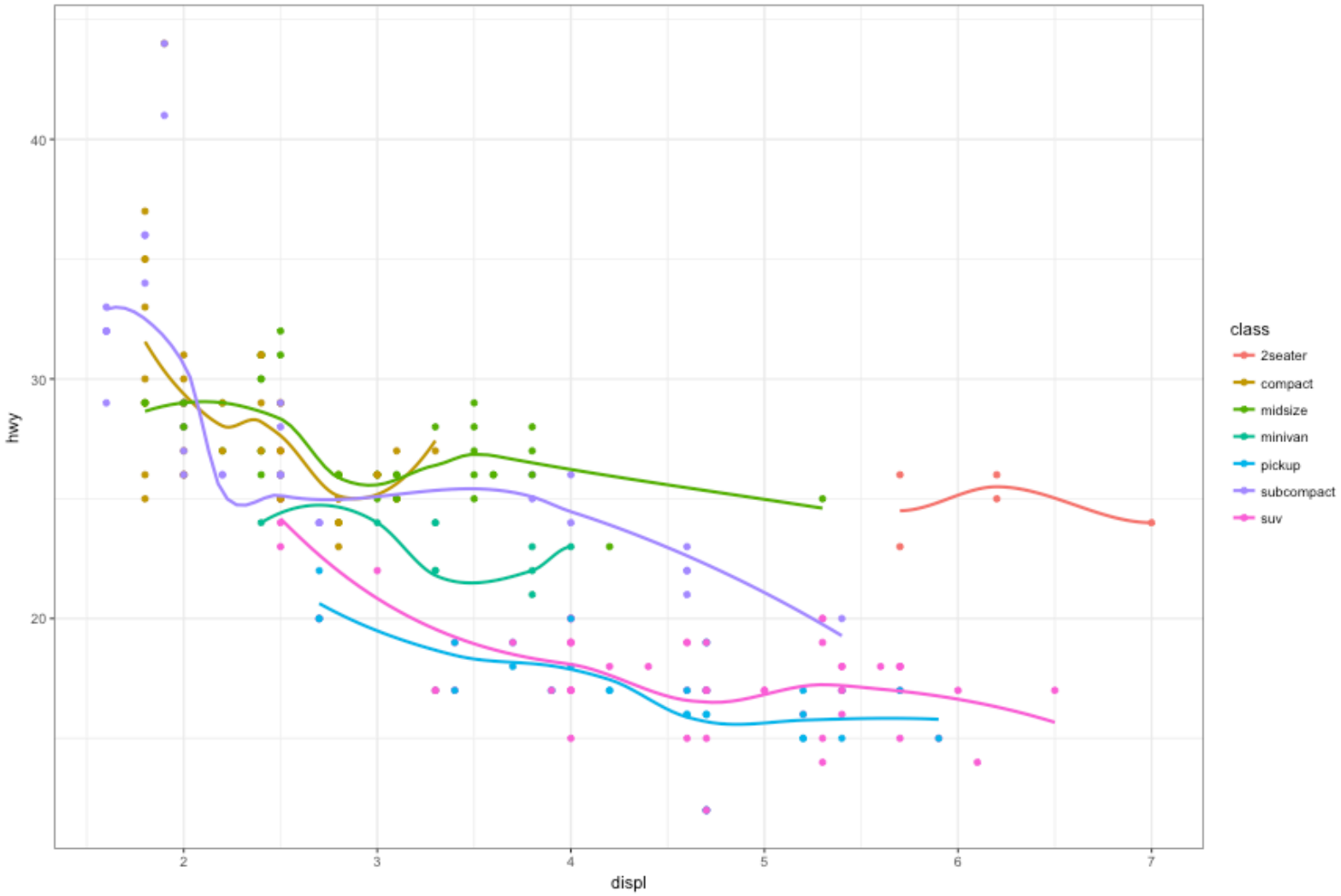


Add smooth line for each class

Too busy

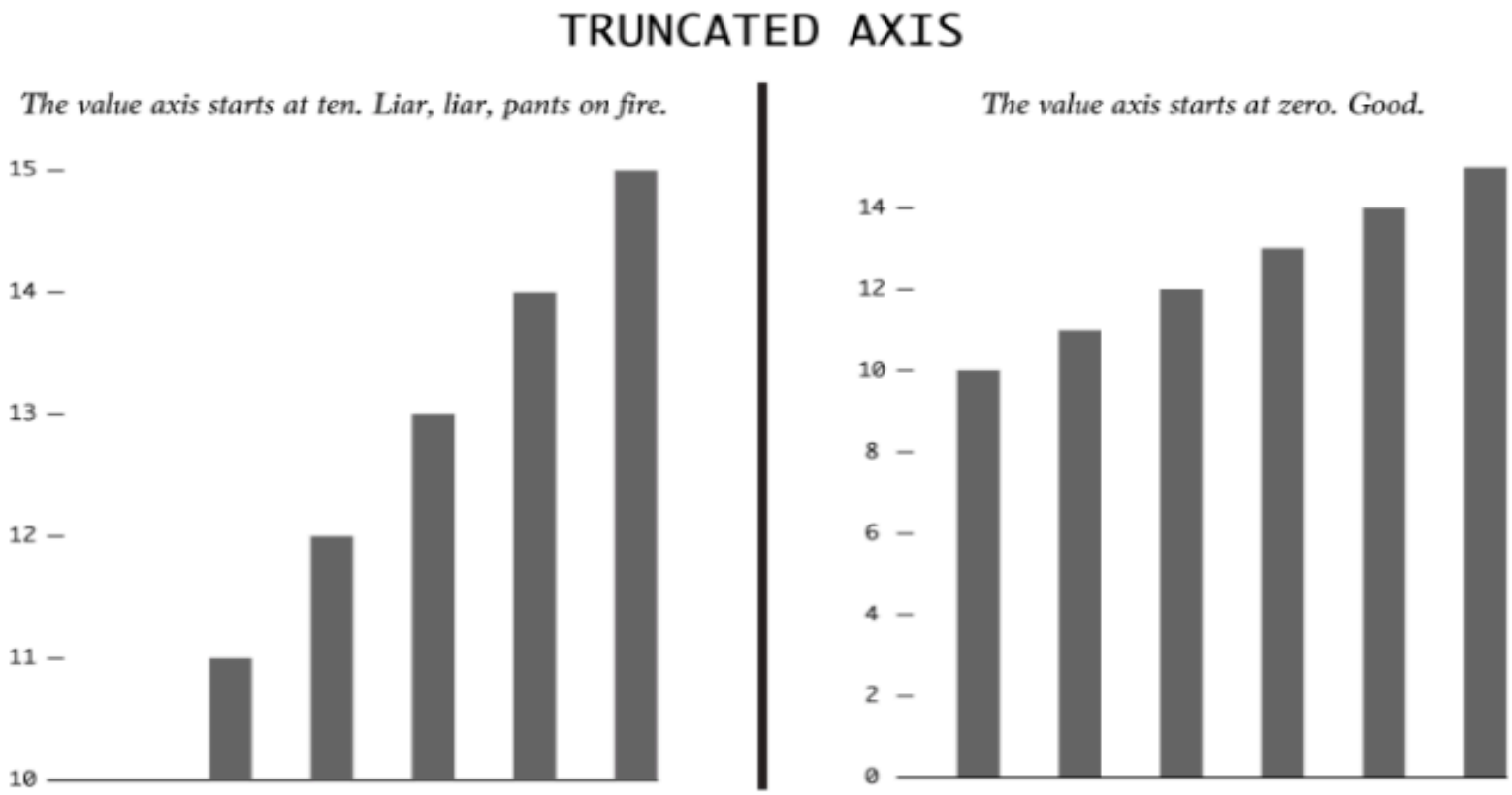


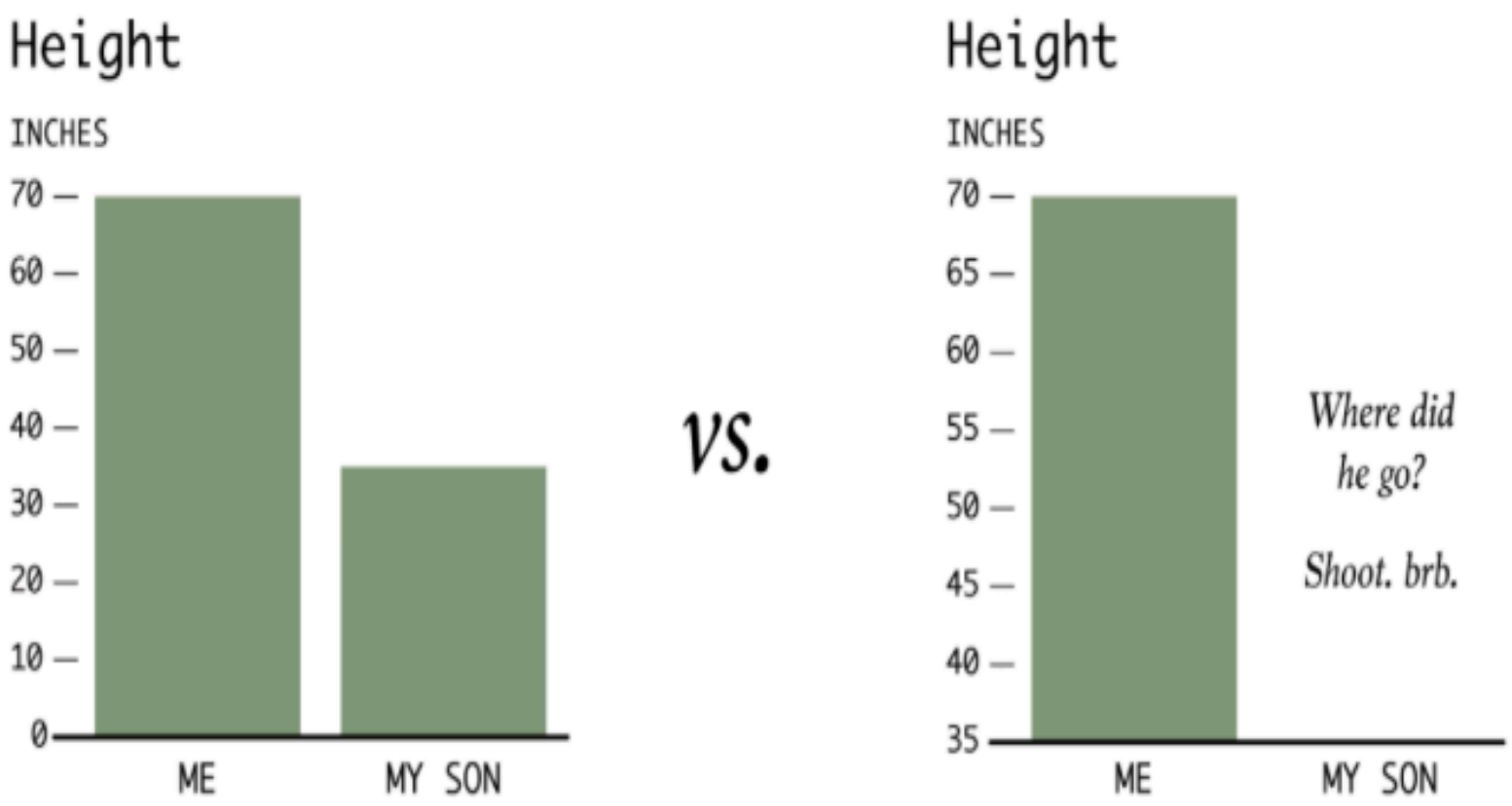
Remove SE


hwy
40
30
20
2
3
4
5
6
7
displ
class
2seater
compact
midsize
minivan
pickup
subcompact
suv

Some things to avoid

Truncated axes



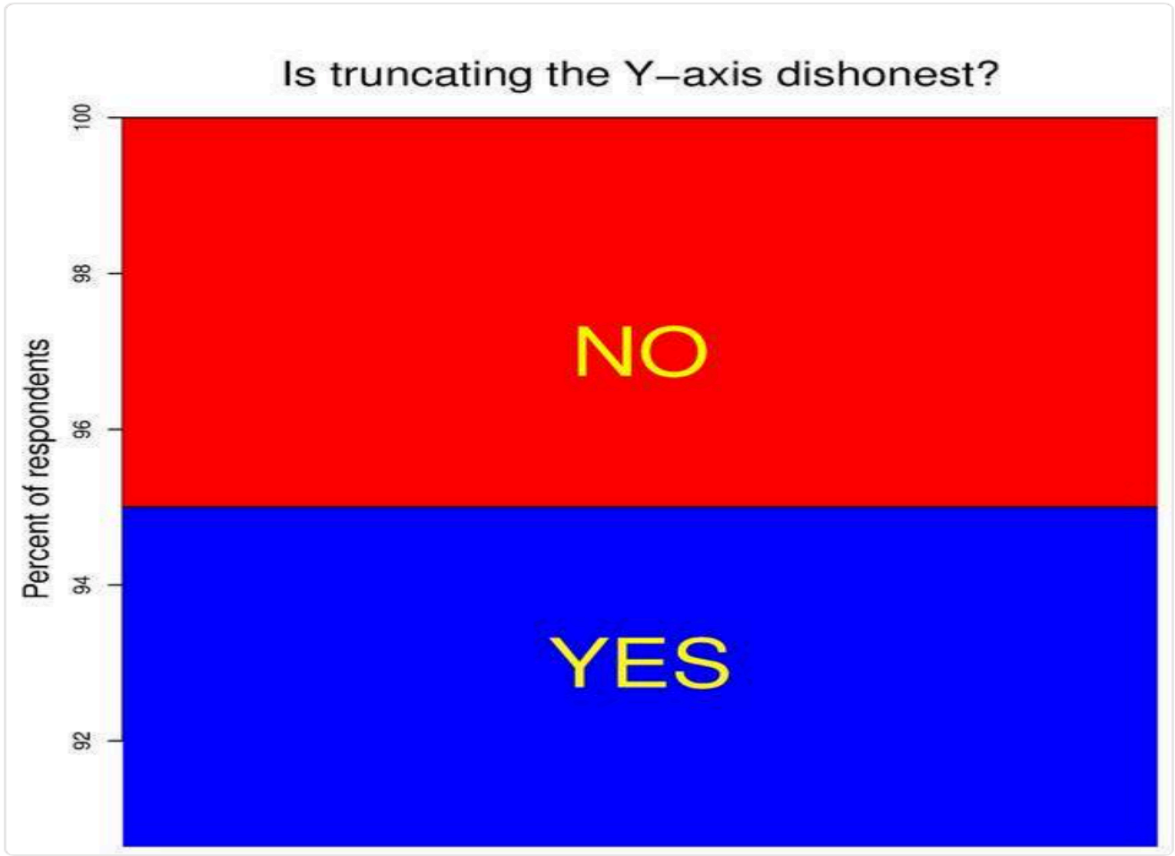

Height
INCHES
70
60
50
40
30
20
10
0
ME
MY SON
vs.
Height
INCHES
70
65
60
55
50
45
40
35
ME
MY SON
Where did he go?
Shoot. brb.

richard shotton
@rshotton

Follow

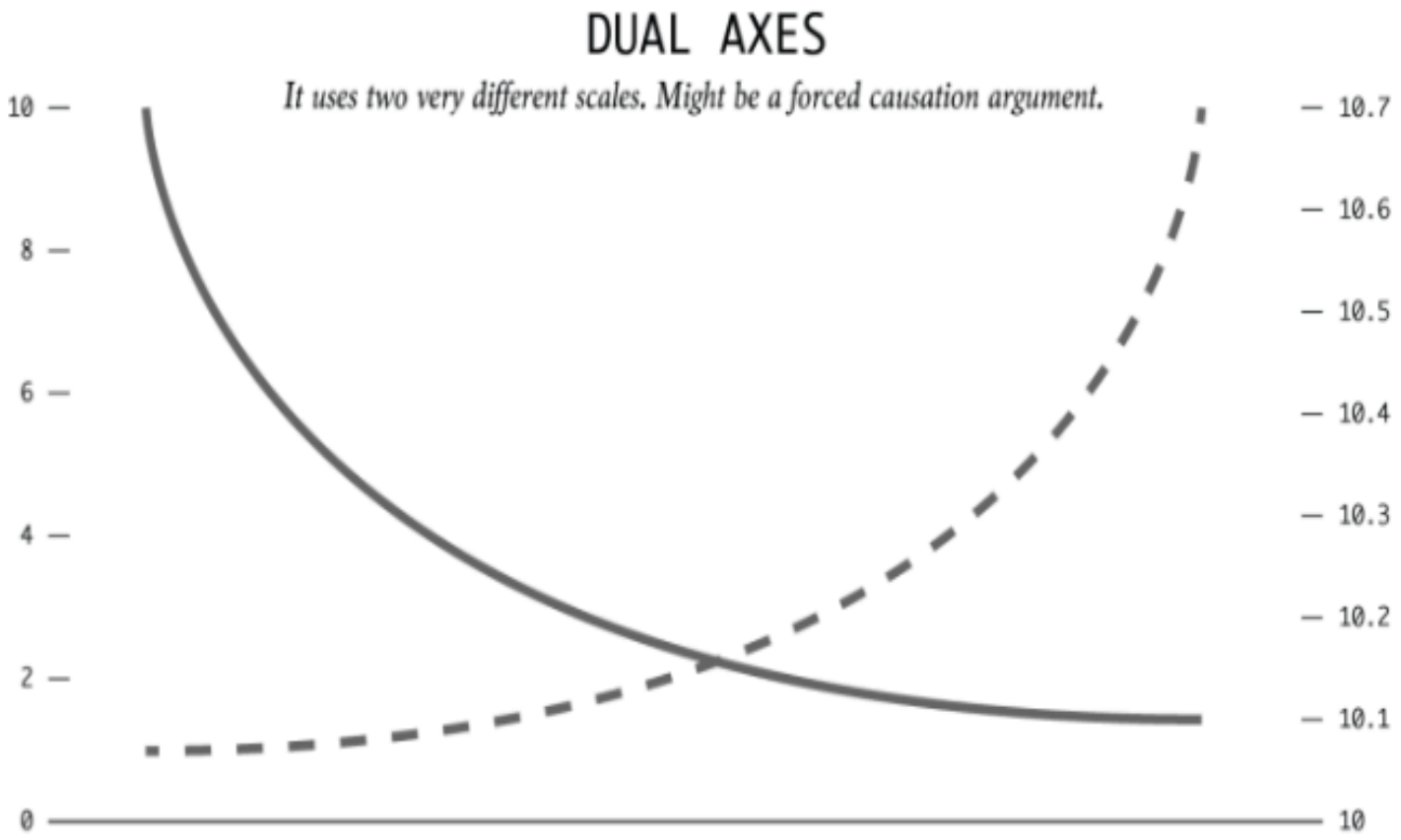
Is truncating the y-axis dishonest?

By @bill_easterly



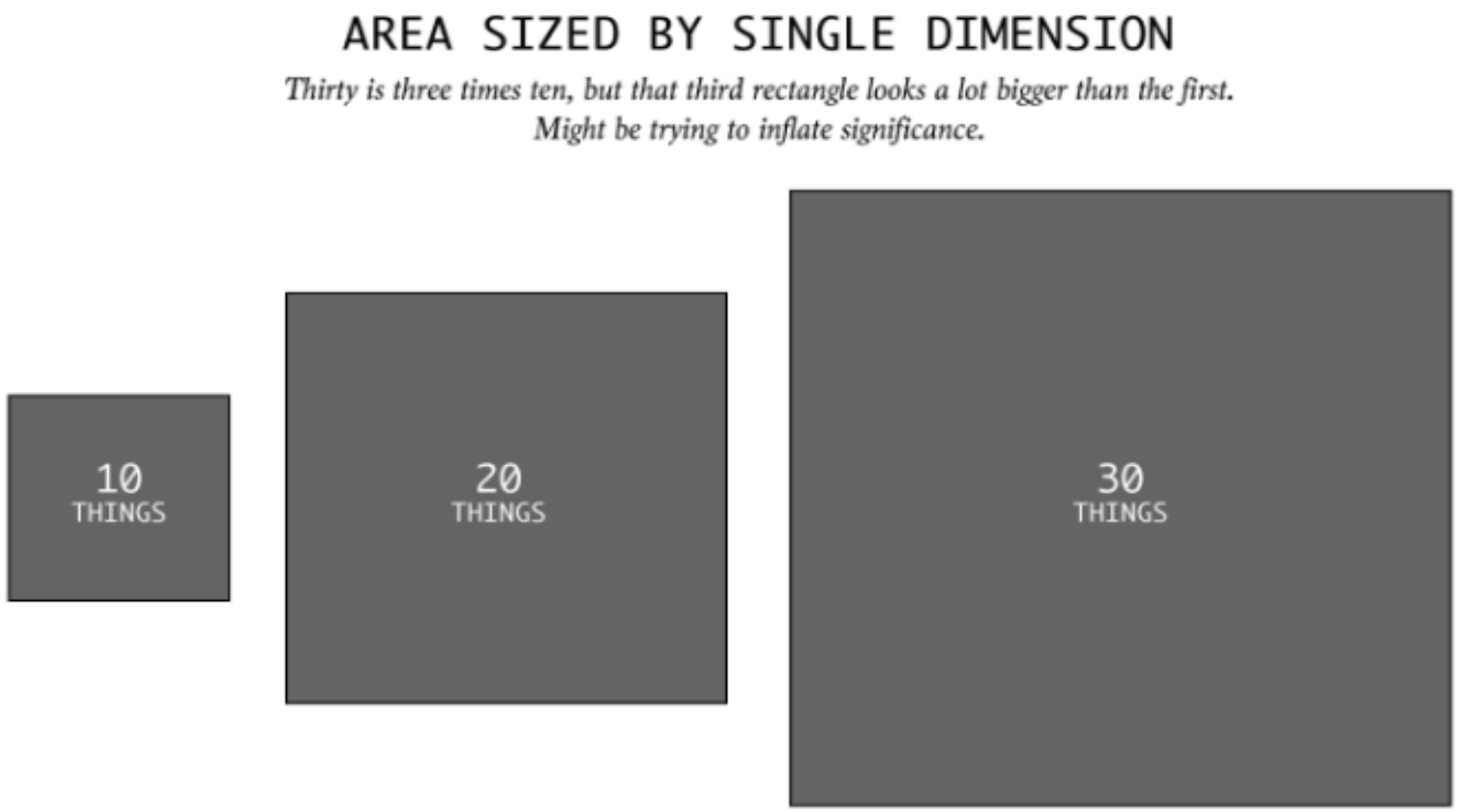
8:26 AM - 20 May 2017

Dual axes



Proof dual axes are bad. But, some lively discussion.

Scaling issues



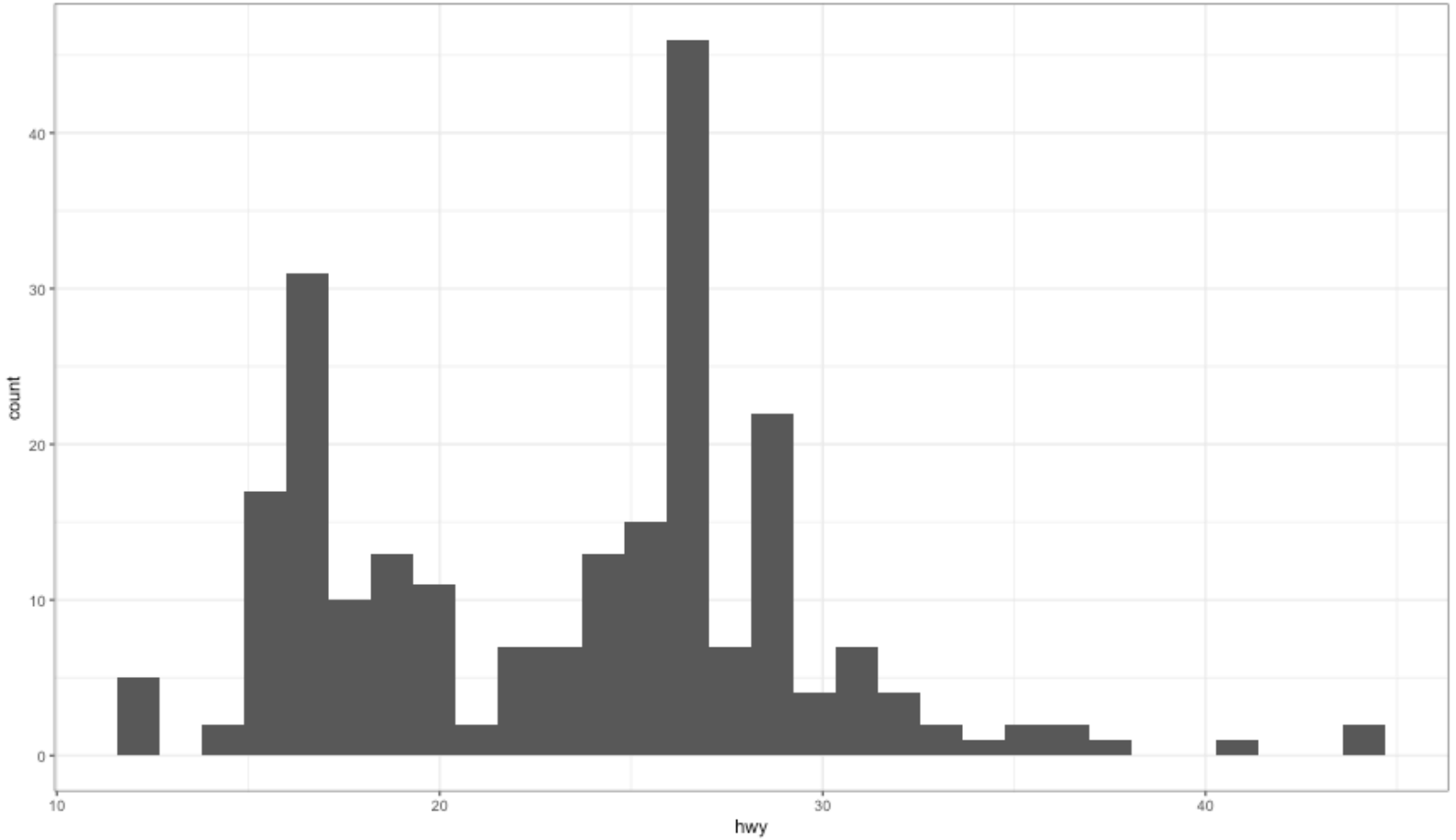
Poor binning choices

Some general advice

- Consider the purpose of the plot.
  - Relation? Scatterplots
  - Distribution? Histogram or density plot
  - Trend? Line plot, scatterplot with smoother, etc.
- How many variables? What type?
  - One continuous variable: histogram, density plot, or similar
  - Two continuous: Scatterplot (if you have lots of data, consider binning)
  - One categorical one continuous: boxplots, violin plots, bar plots
  - Two categorical variable? Mosaic plot

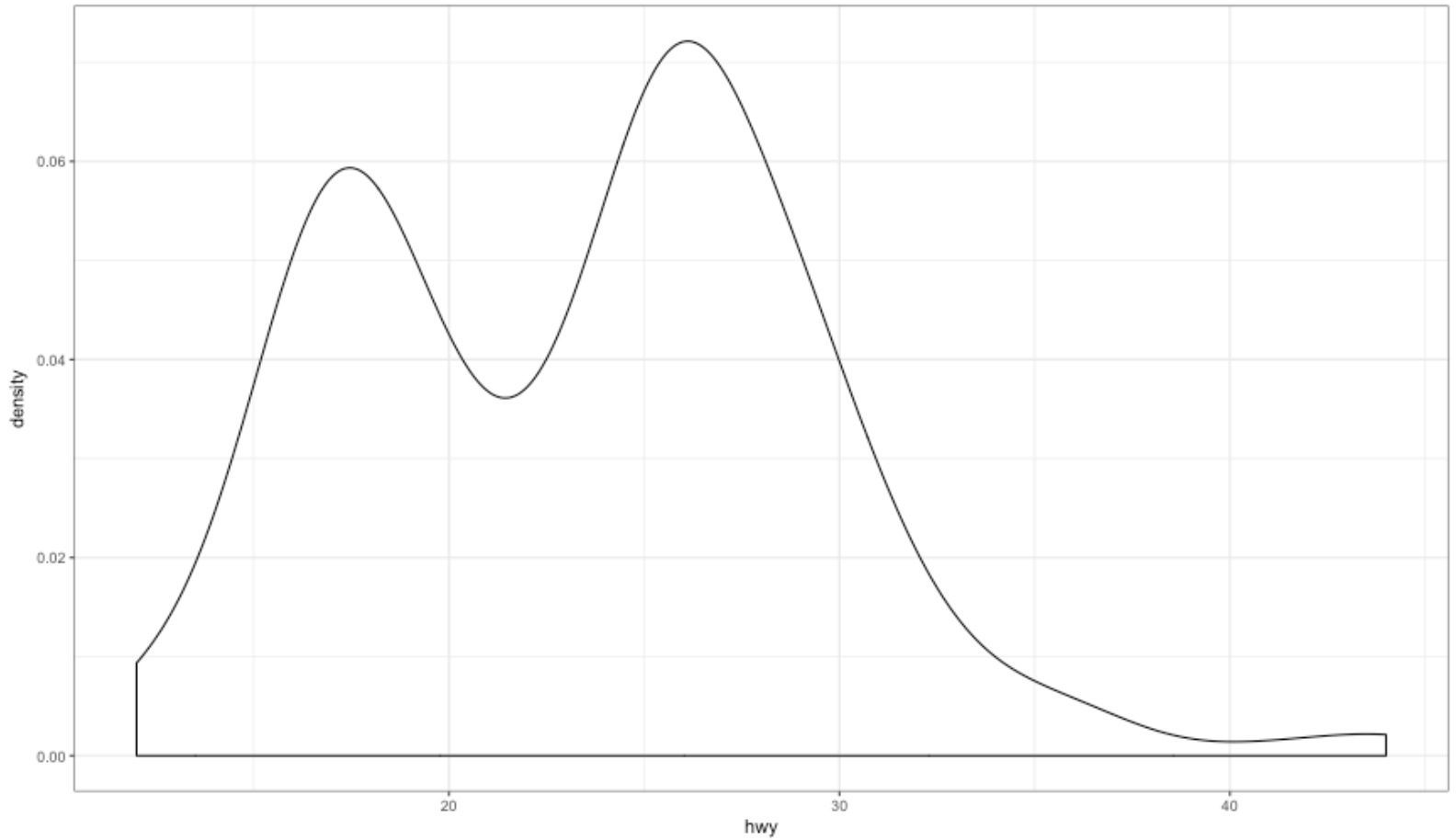
One continuous variable

Histogram



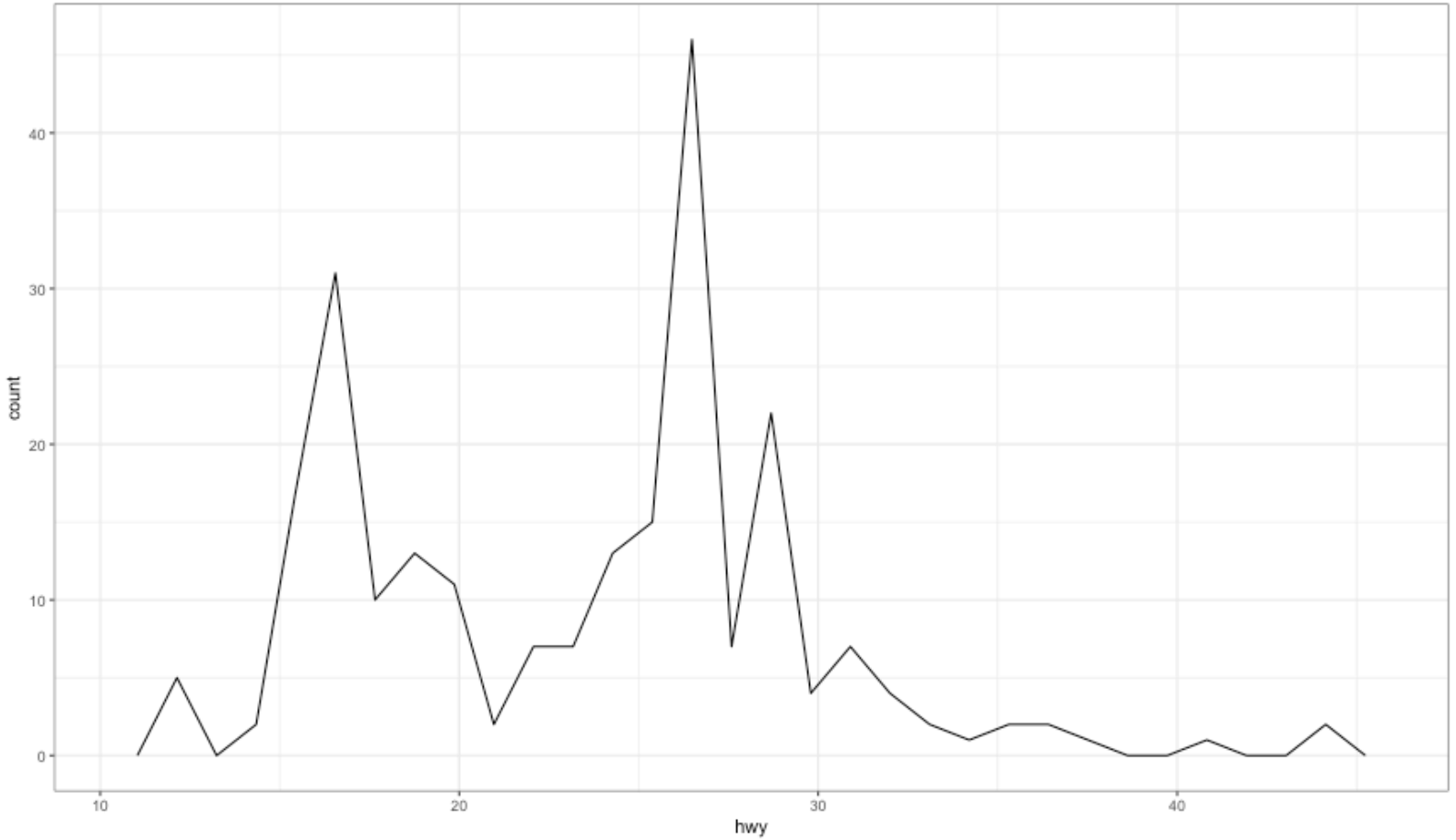
One continuous variable

Density plot

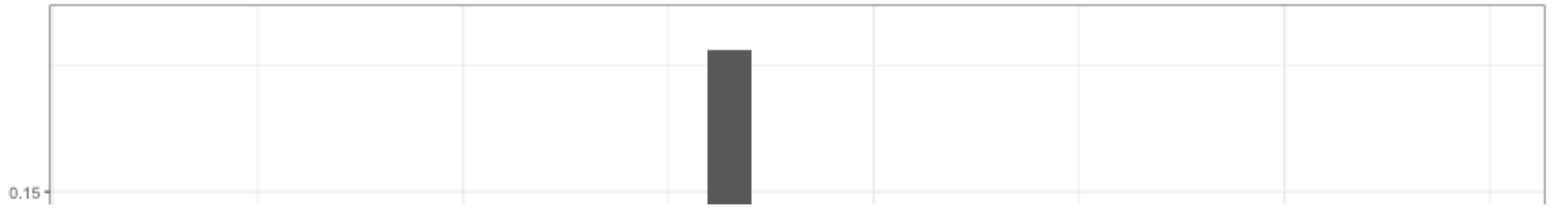


One continuous variable

Frequency polygon

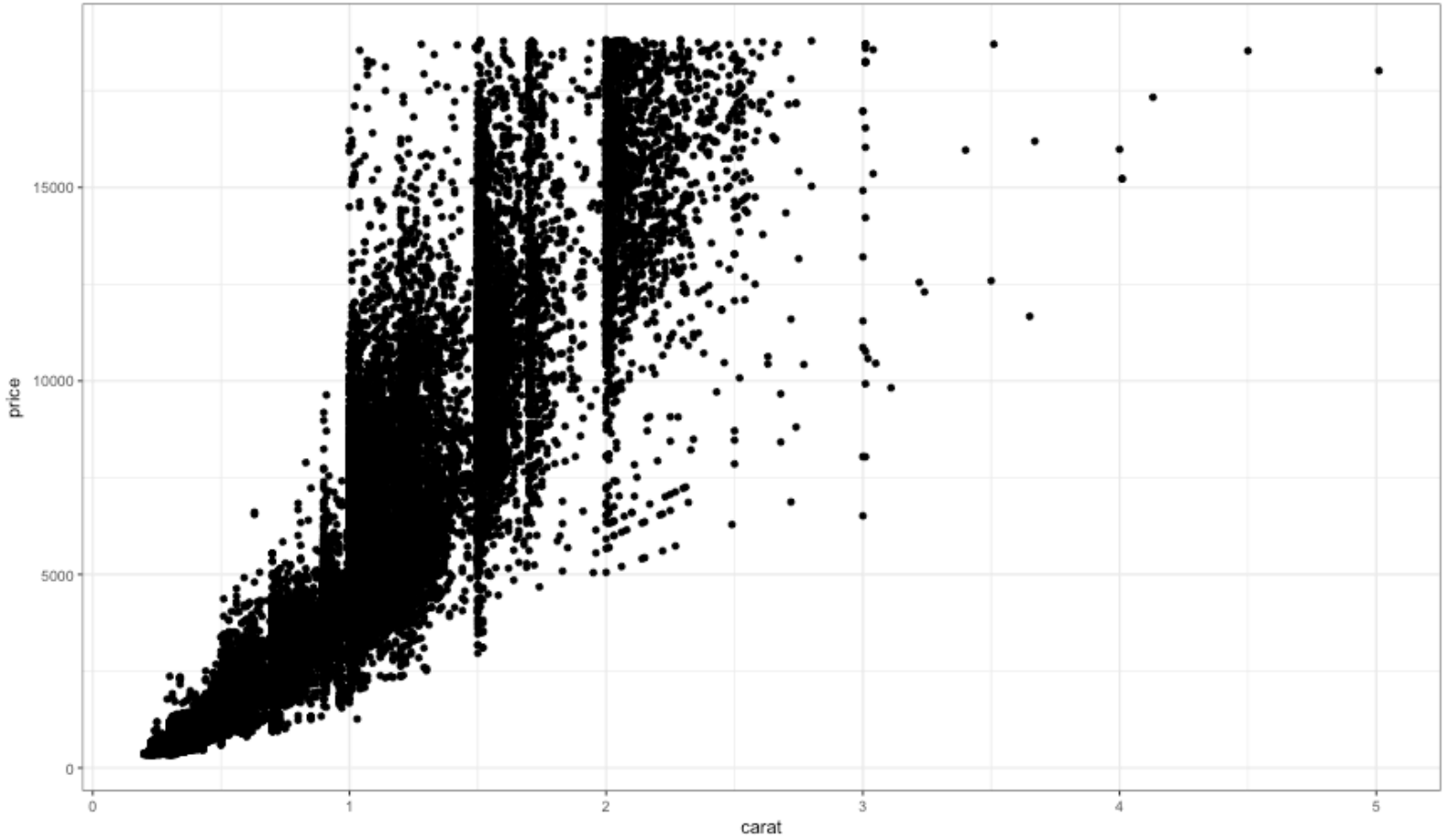


Consider overlays



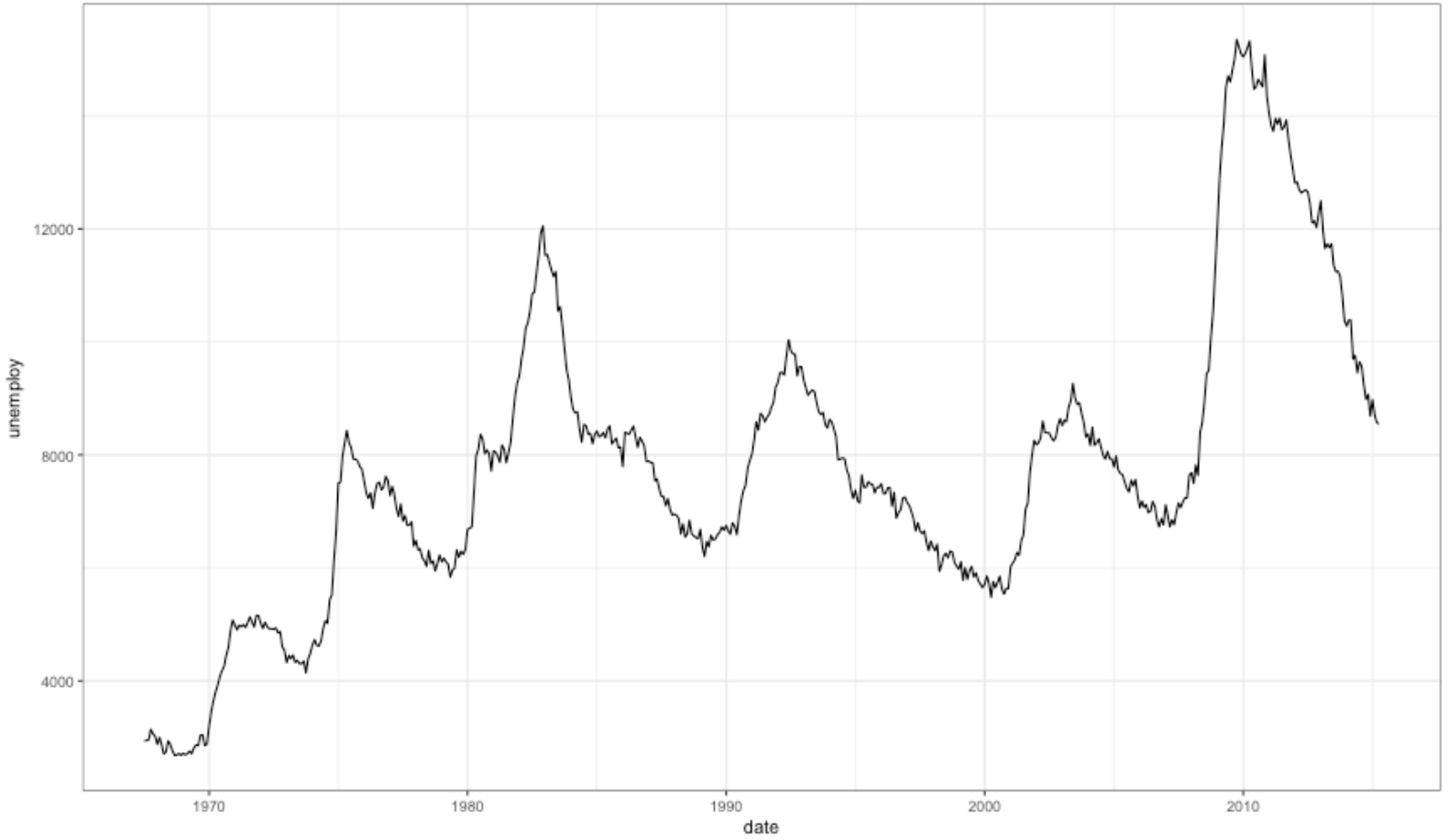
Two continuous variables

Scatterplot

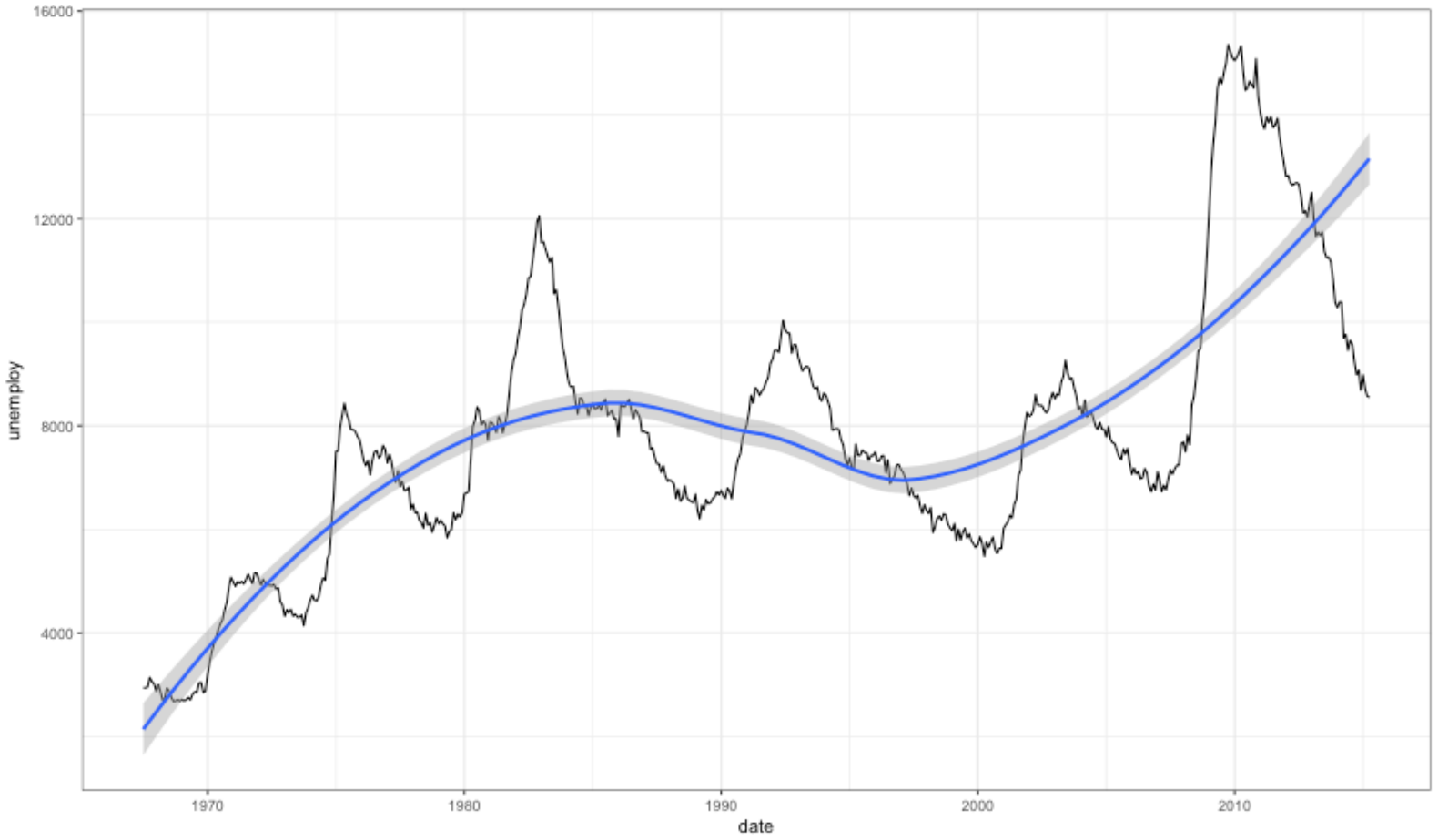


Trend

Line plot (often with date or time on x-axis)

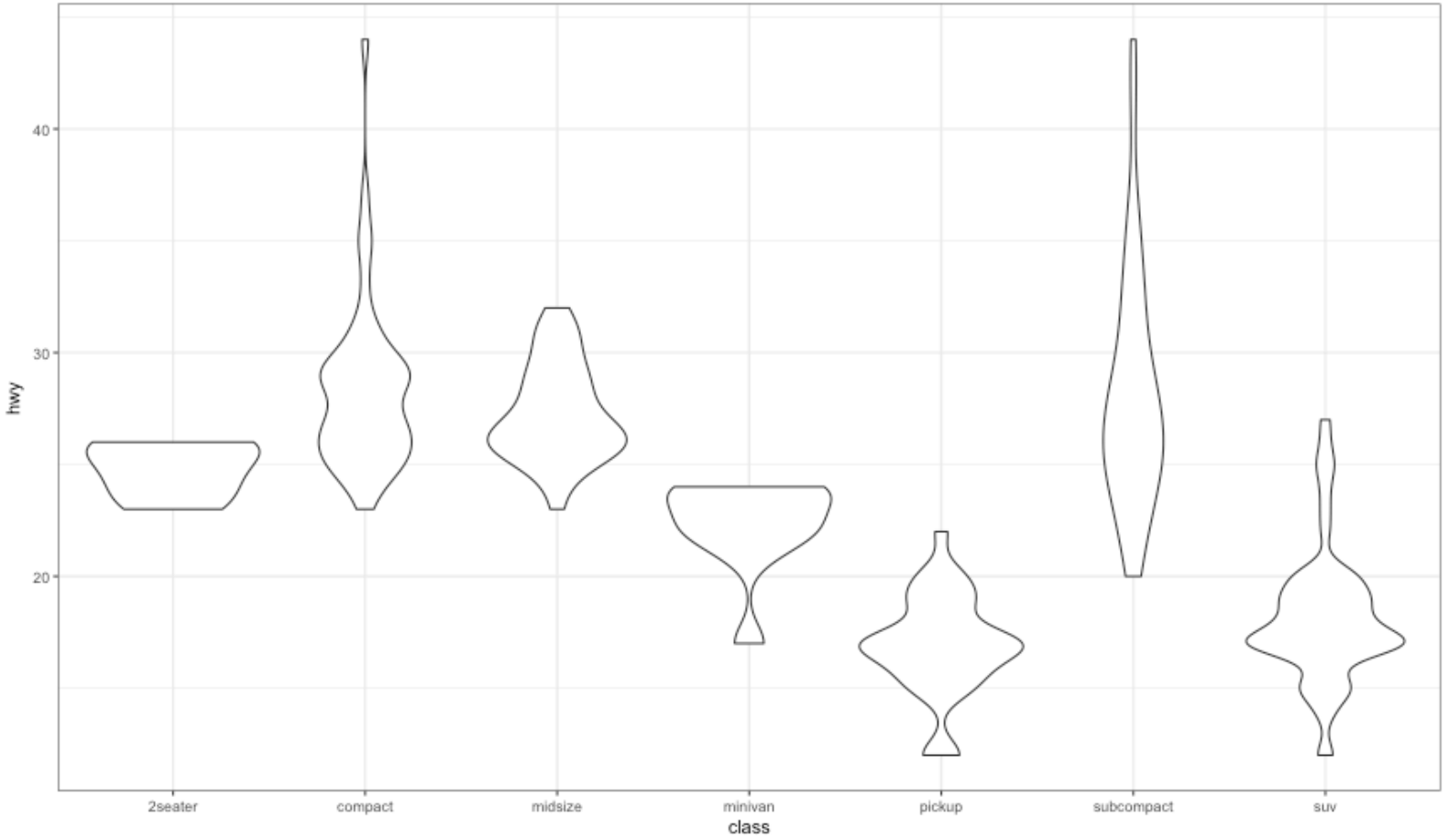


Trend w/smoother



Categorical & Continuous

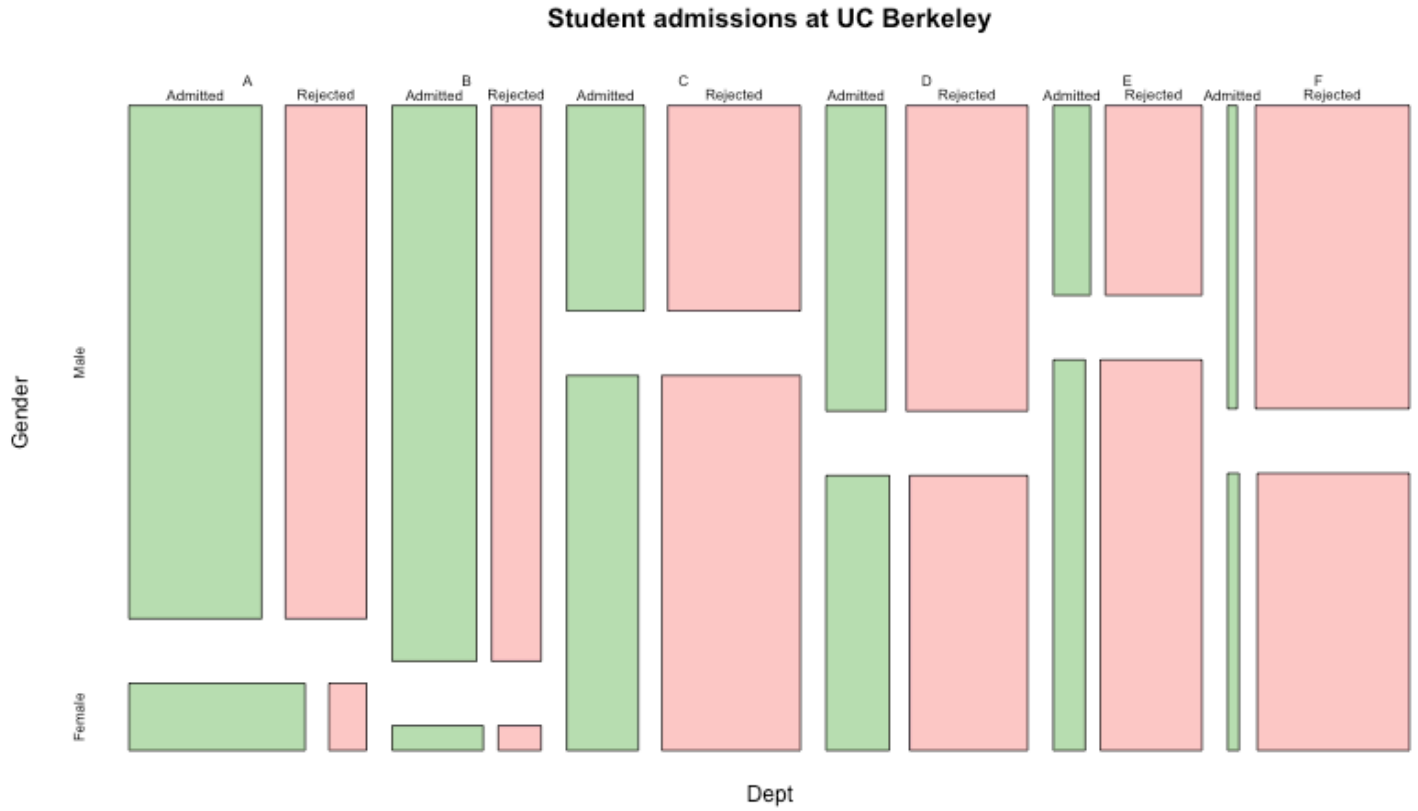
Violin plots



Overlay data

Two categorical variables

Mosaic plot



Don't end up in a blog for wrong reasons

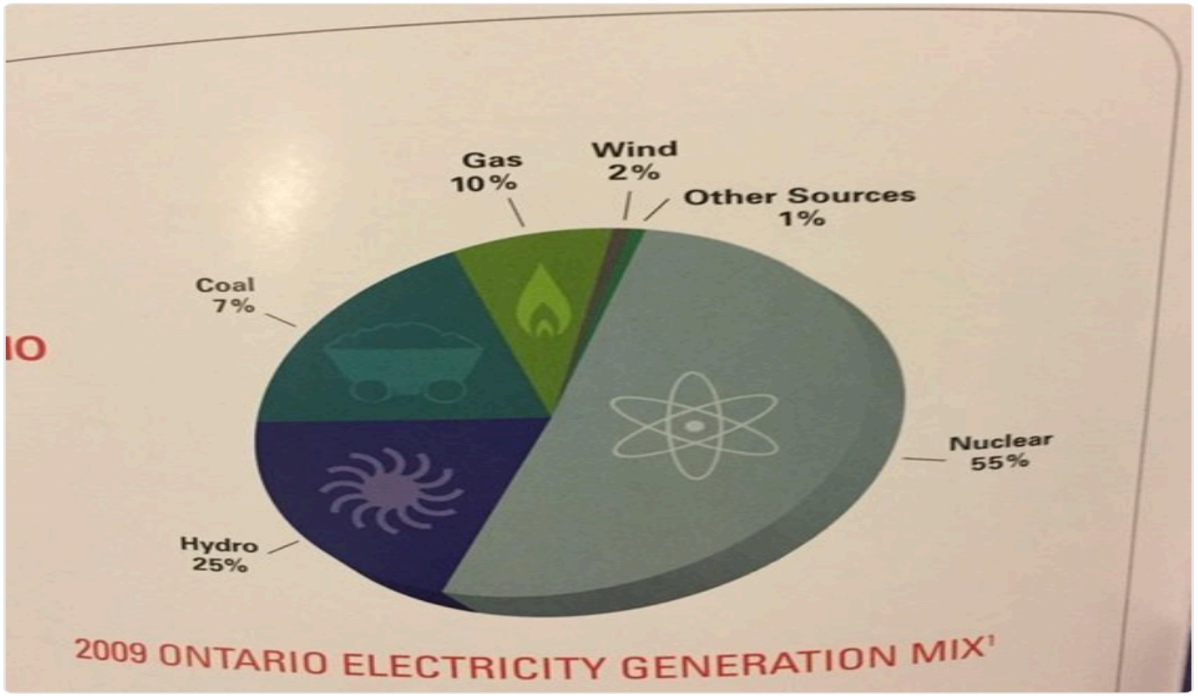
- https://flowingdata.com/2010/05/14/wait-something-isnt-right-here/
- https://flowingdata.com/2009/11/26/fox-news-makes-the-best-pie-chart-ever/

One more example

Bill the Lizard
@lizardbill

Follow

"Can we make the coal icon smaller?"
"nope"
"Let's just make that pie wedge bigger then."
"cool"



1:15 PM - 23 May 2017

Conclusions

- Essentially never
  - Use pie charts (use bar charts instead)
  - Use dual axes (produce separate plots instead)
  - Use 3D unnecessarily
  - Add color for color's sake (this isn't sales)
- Rarely
  - Truncate axes
- Do
  - Show the data
  - Be as clear as possible
  - Let the data tell the story

Last pitch

- Plotting your data can often lead to surprises.
- Good data visualization can often be just as powerful for inference as complex modeling.
- Ideally, use it for more than just communicating what you already know! (I want to help build tools to make it easier for you to do so)