

A bit about visualization

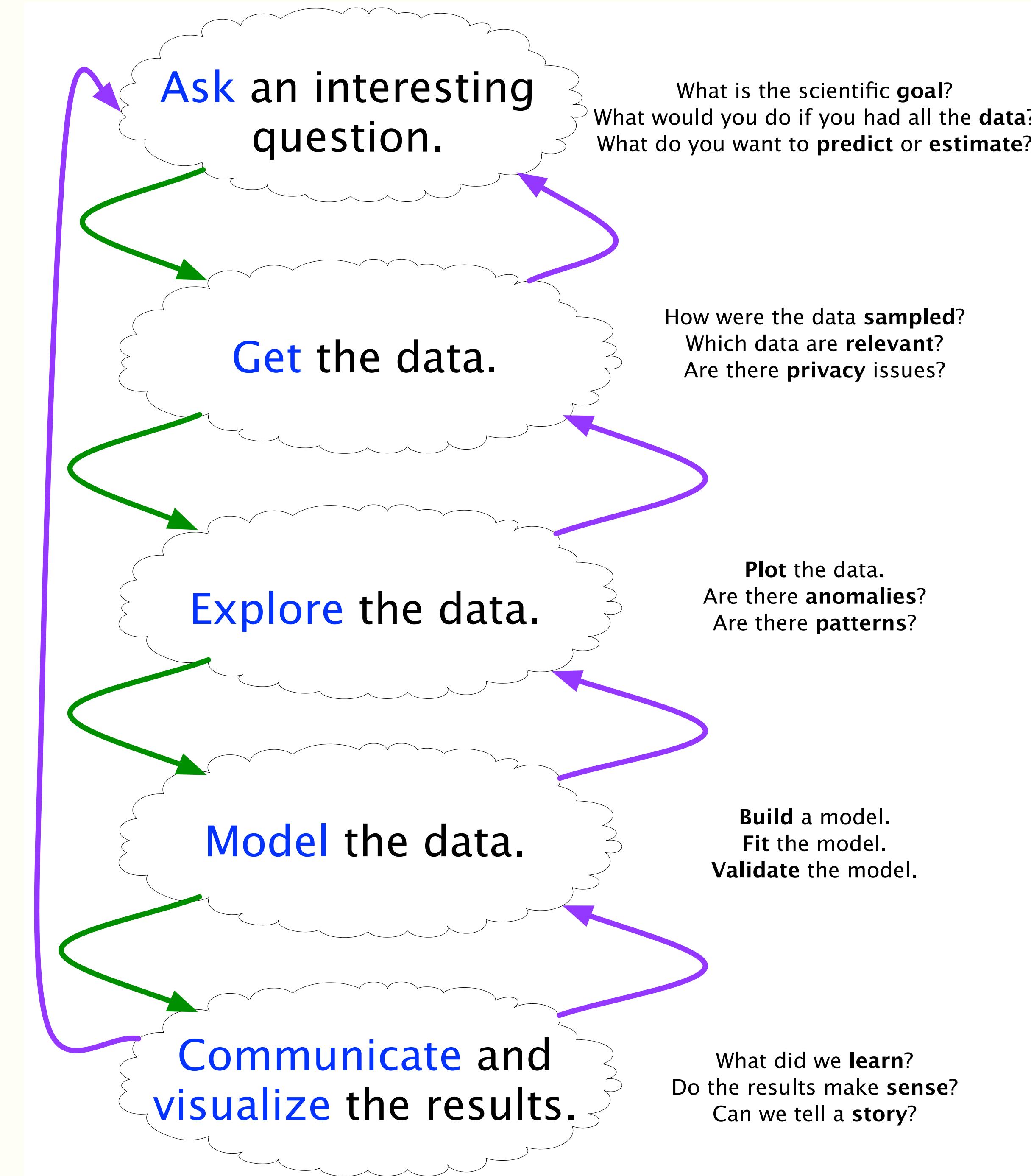
Rahul Dave

Univ.AI

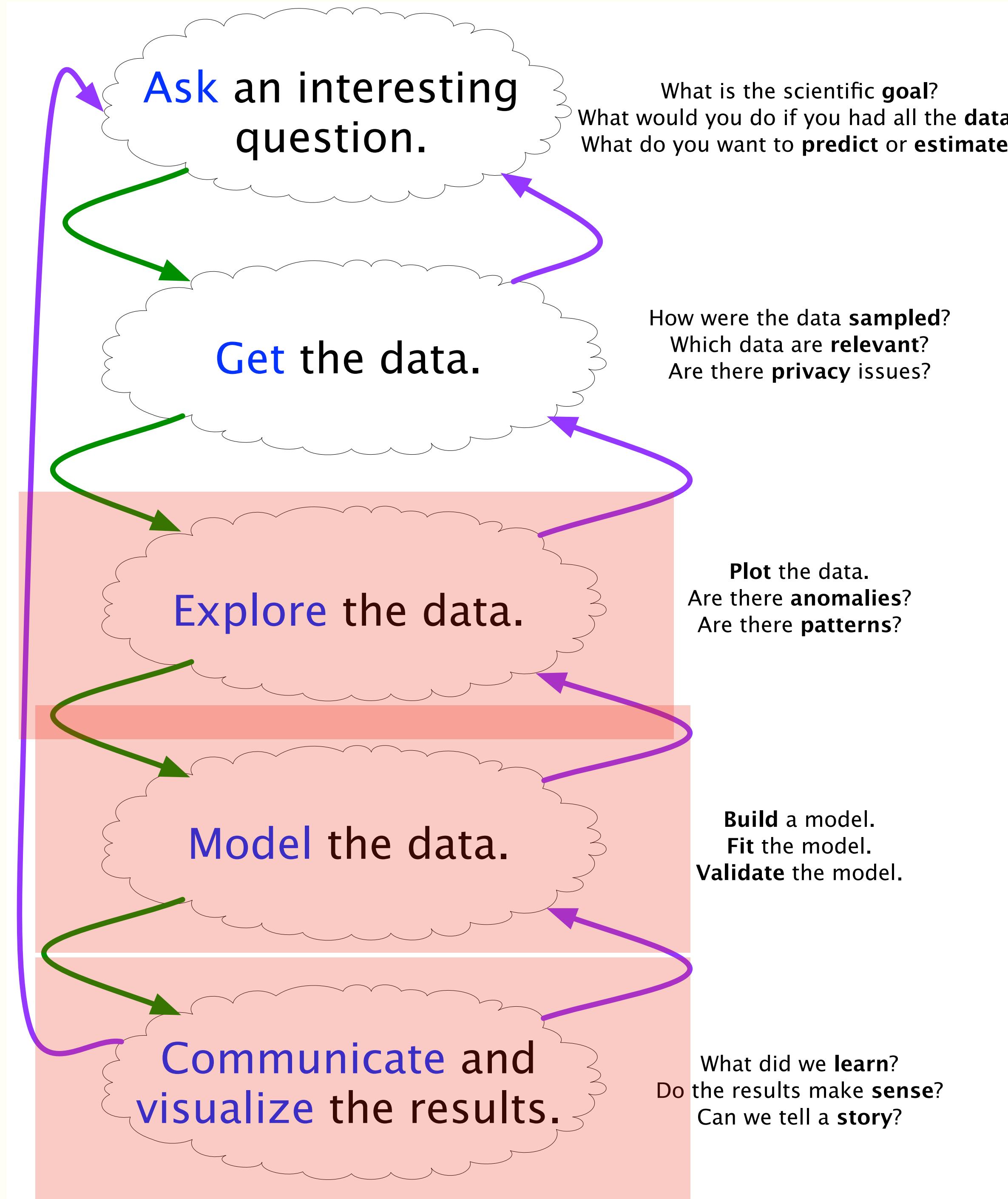
Based on cs109a: Data Science with lots of slides from Hanspeter Pfister (Harvard) and Alberto Cairo (University of Miami).

Why Visualize?

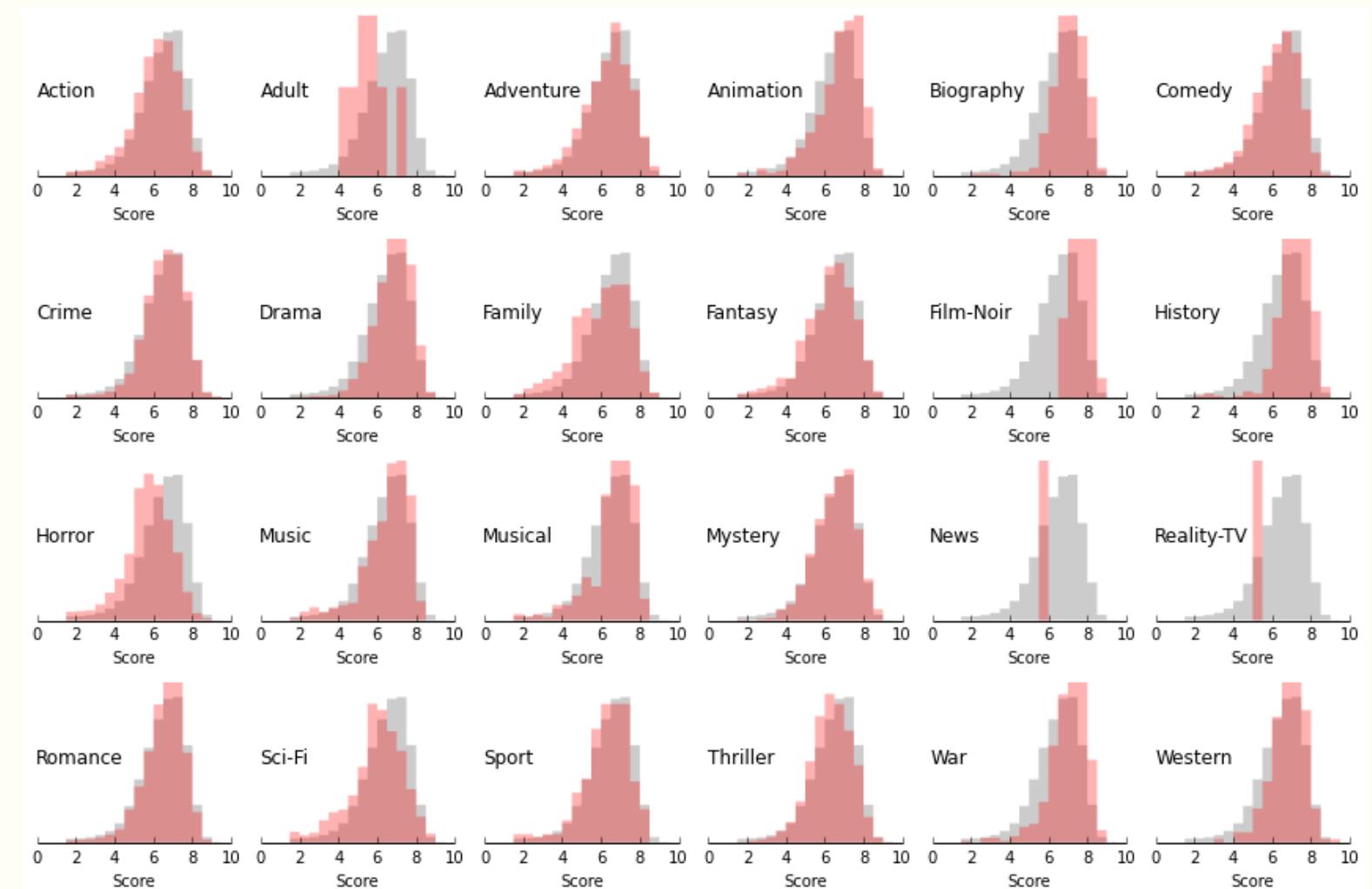
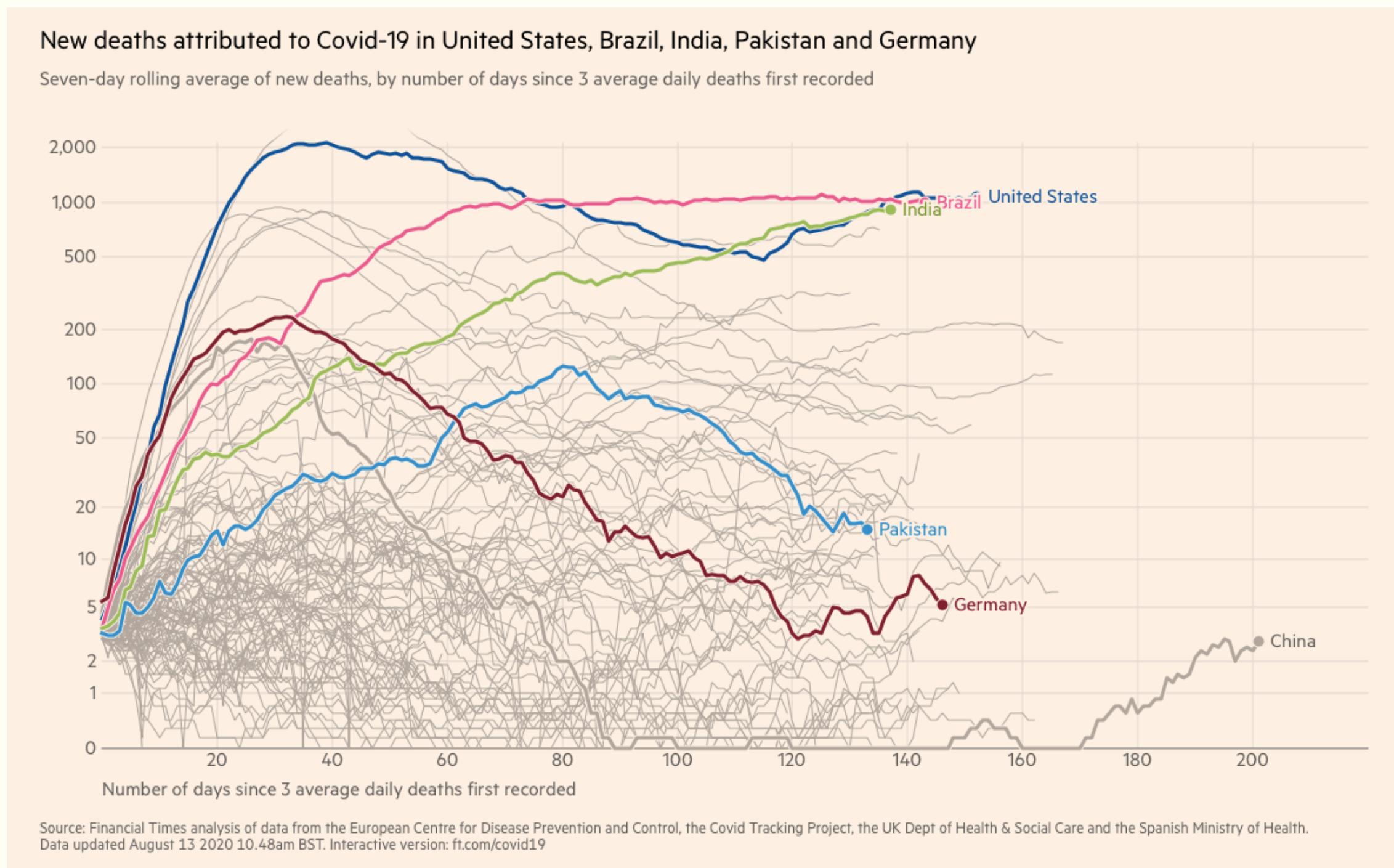
From cs109, Harvard



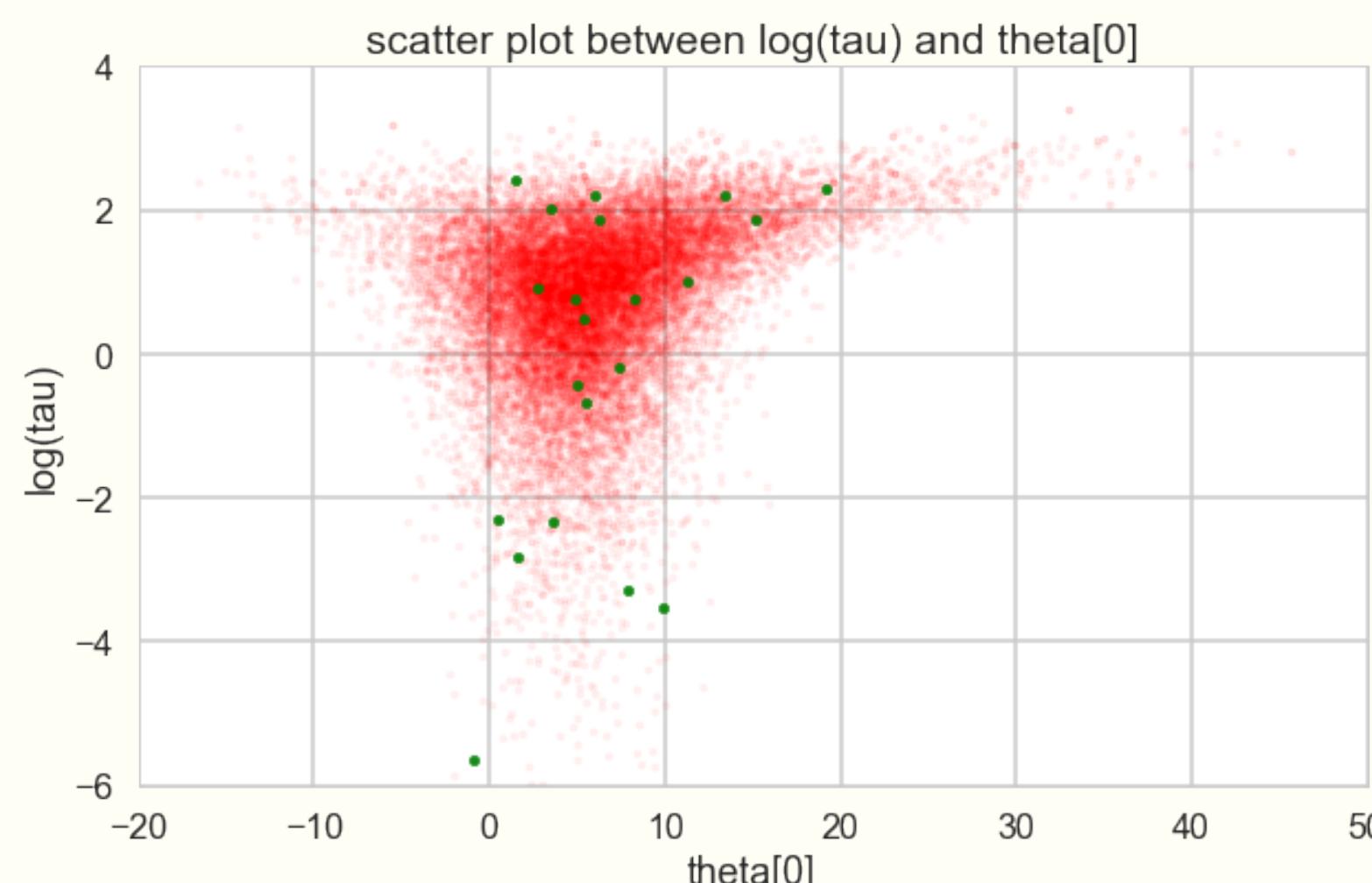
From cs109, Harvard



Communicate Results



Explore data.
Explore models.

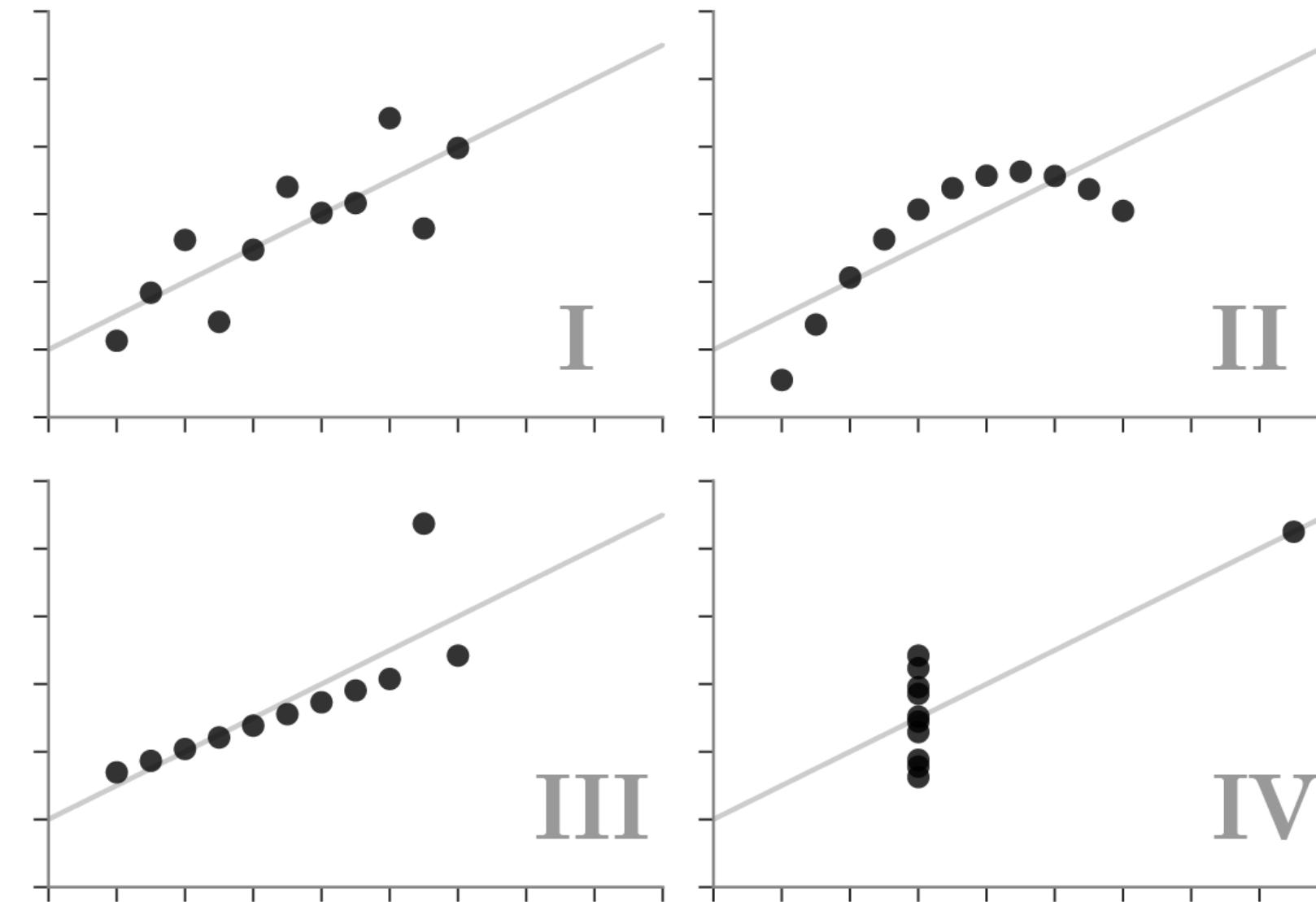


Make sure statistics are not fooling you

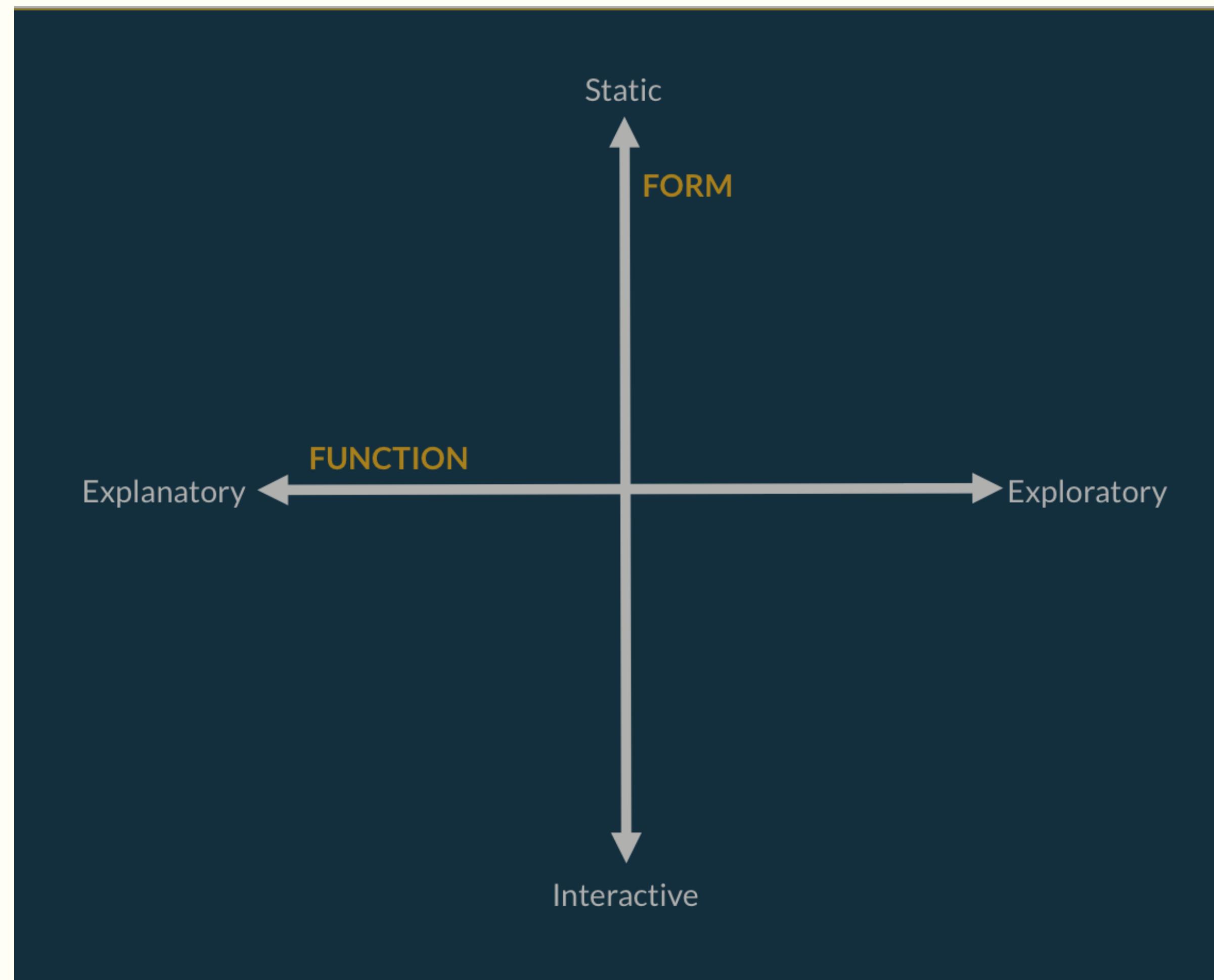


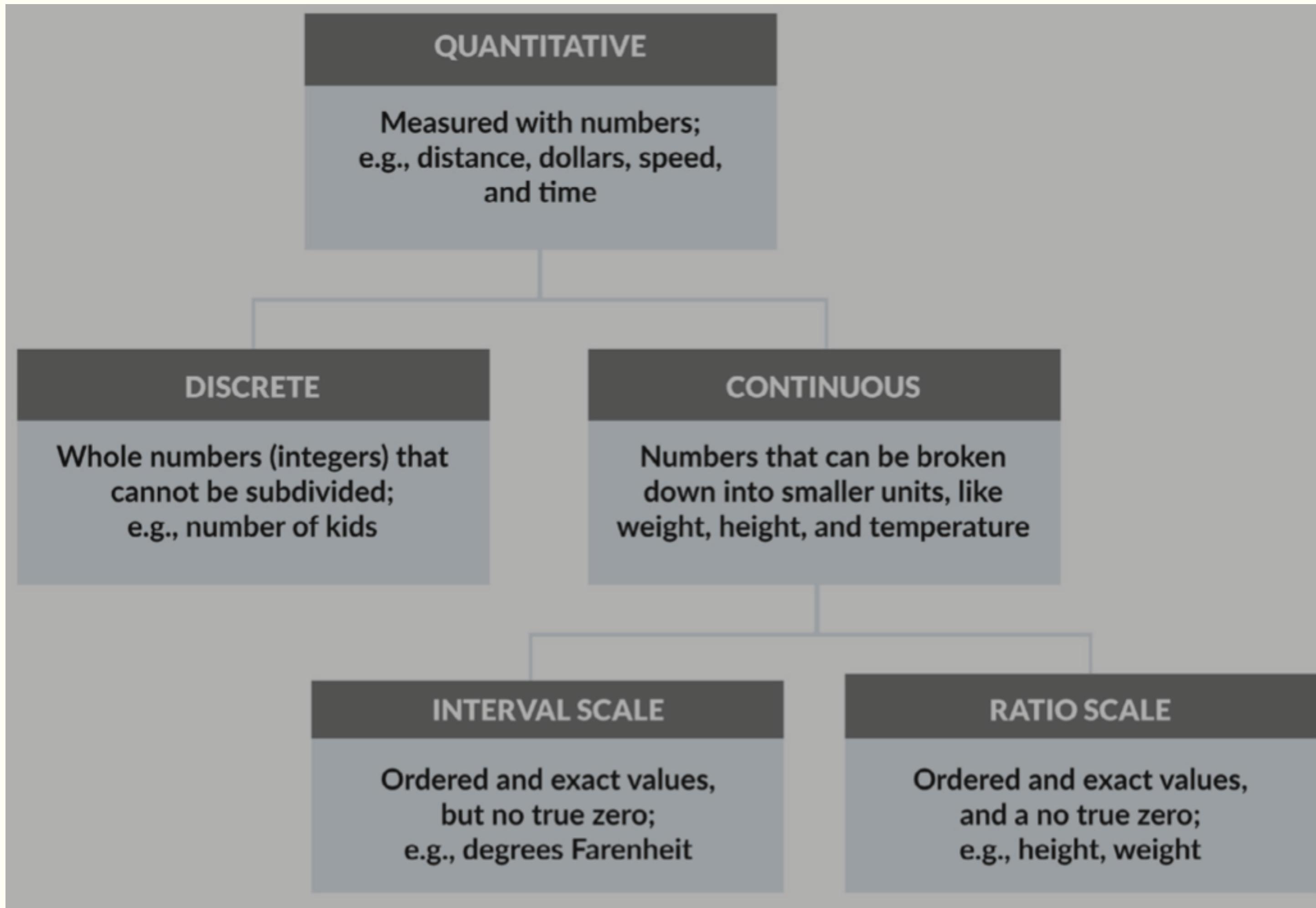
Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



This talk: Static Exploratory



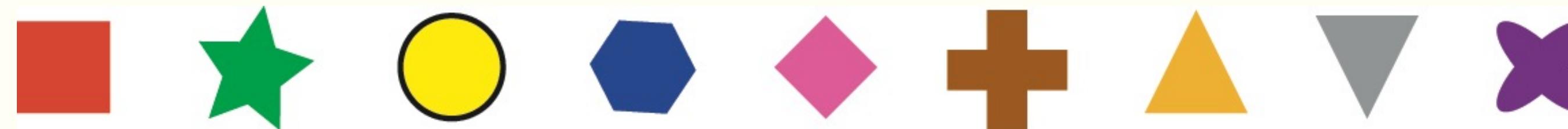


Data Types

Use Color Sensibly

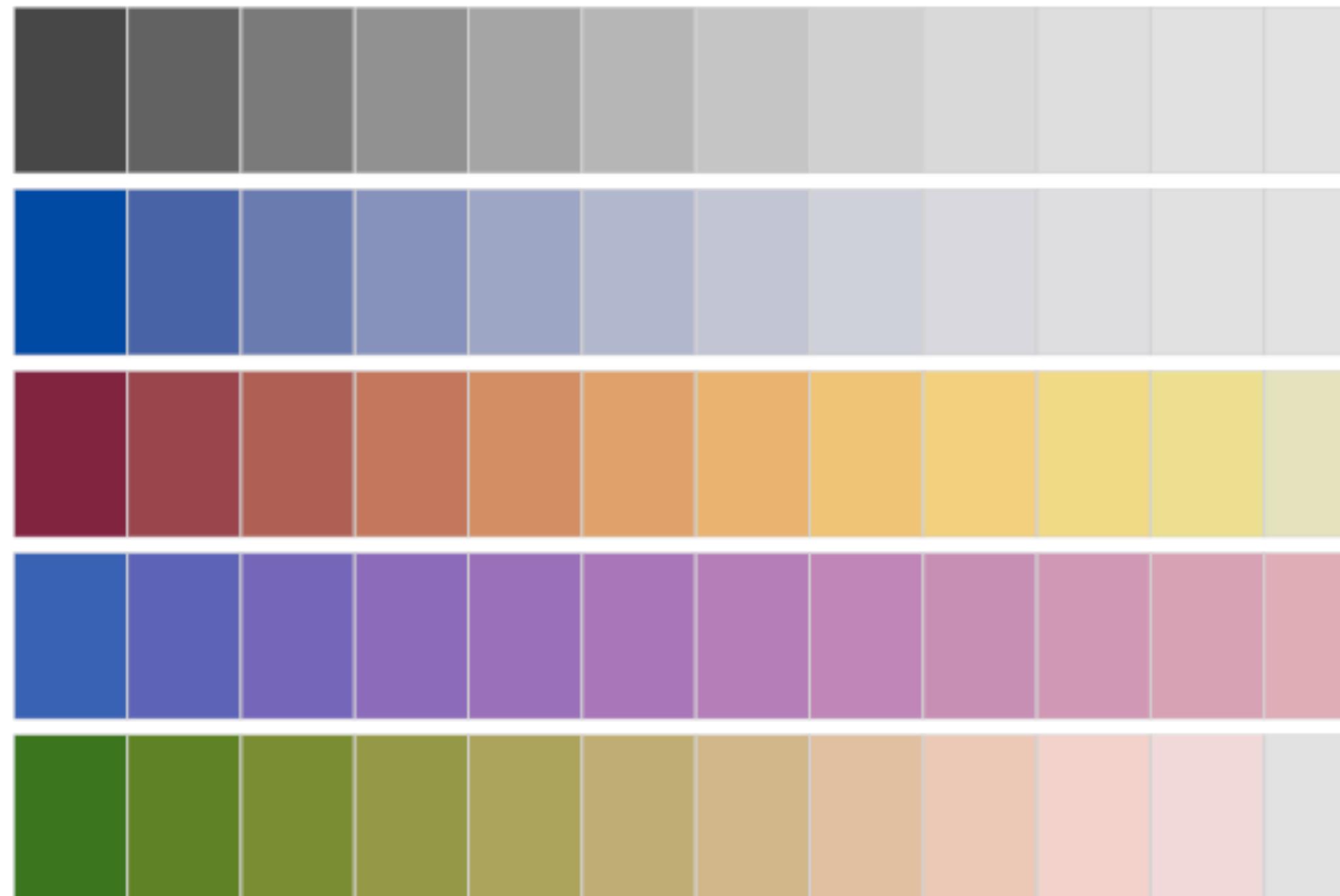
Colors for Categories

Do not use more than 5-8 colors at once



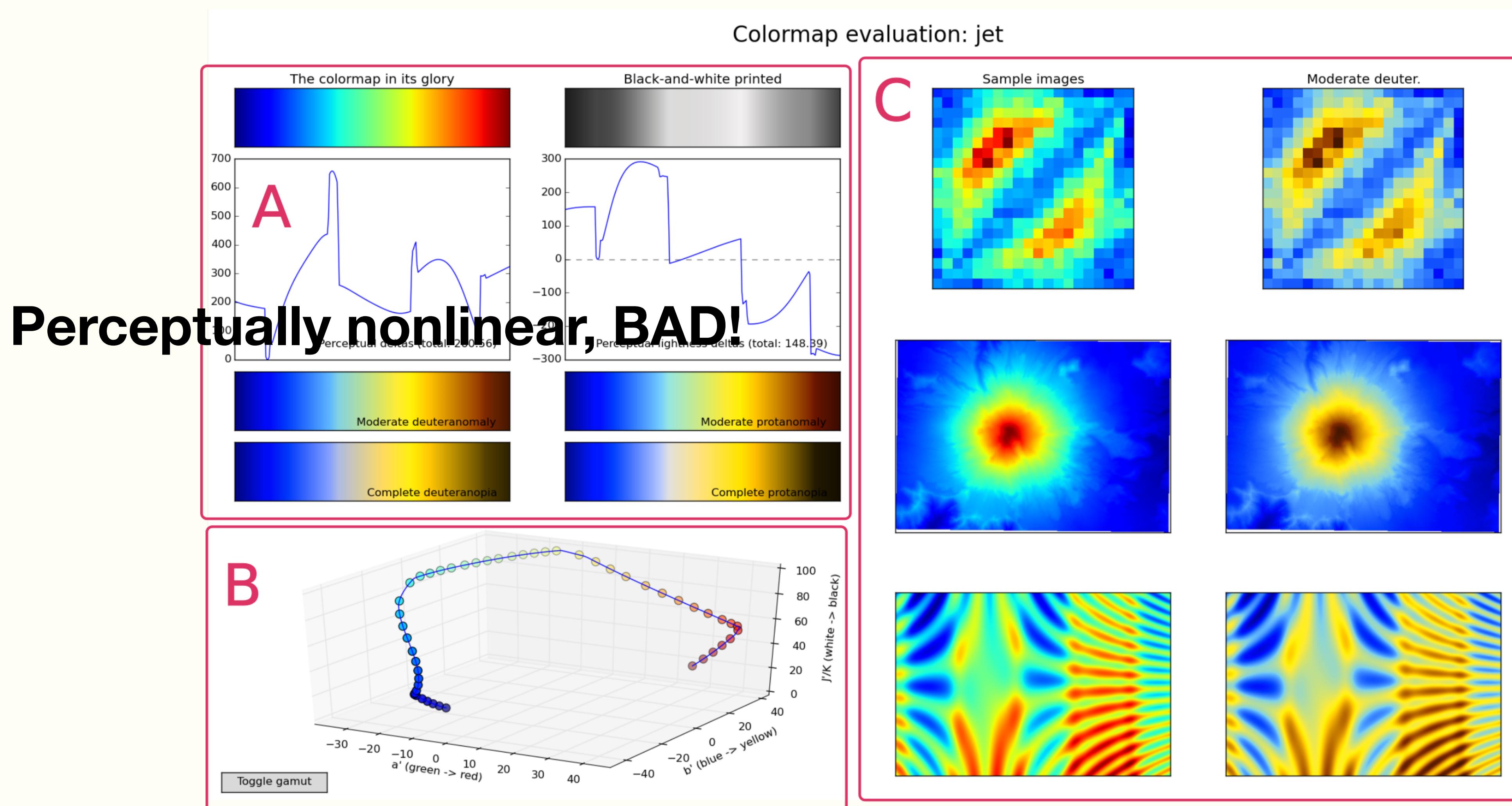
Colors for Ordinal Data

Vary luminance and saturation (see HSL Scale)

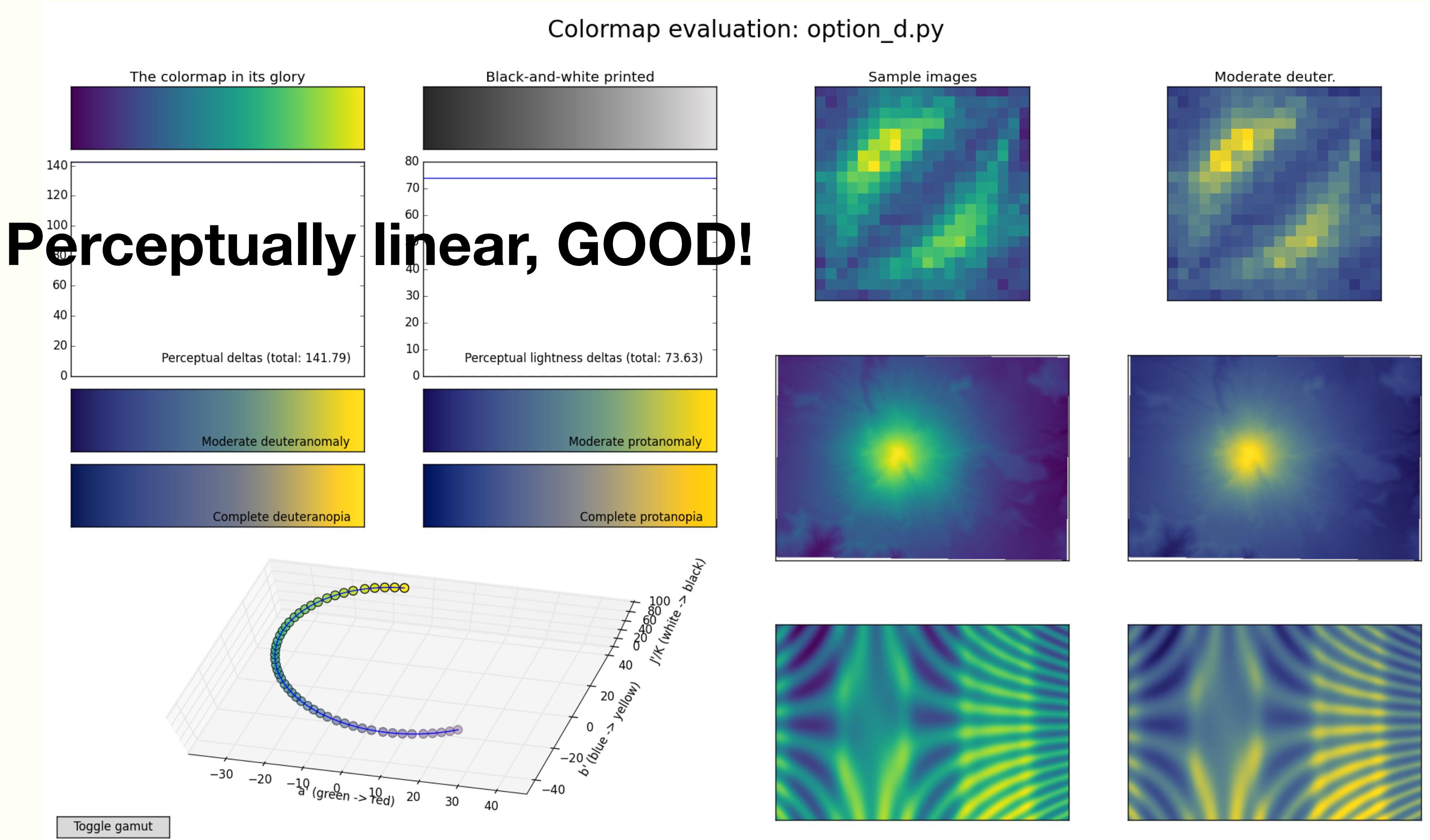


Zeilis et al, 2009, "Escaping RGBland:
Selecting Colors for Statistical Graphics"

Quantitative data colors: Rainbow Colormap



Viridis

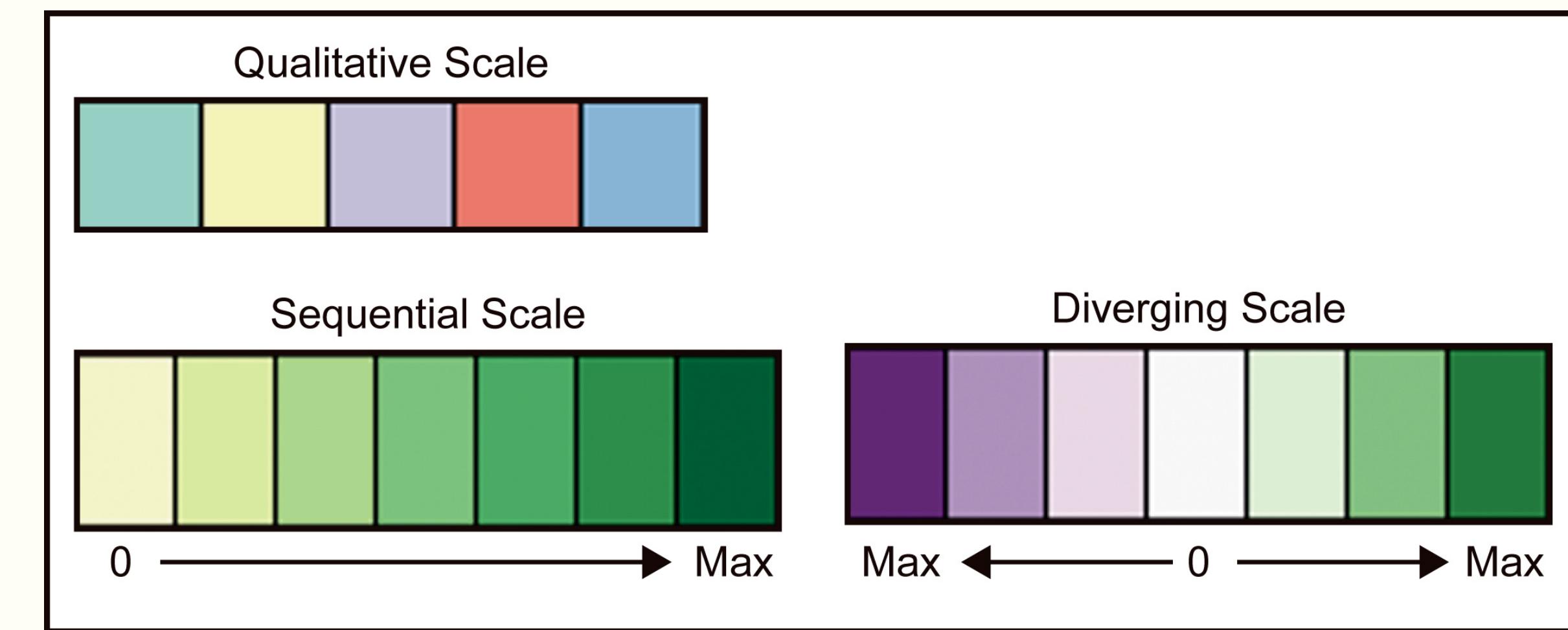


Color Scales

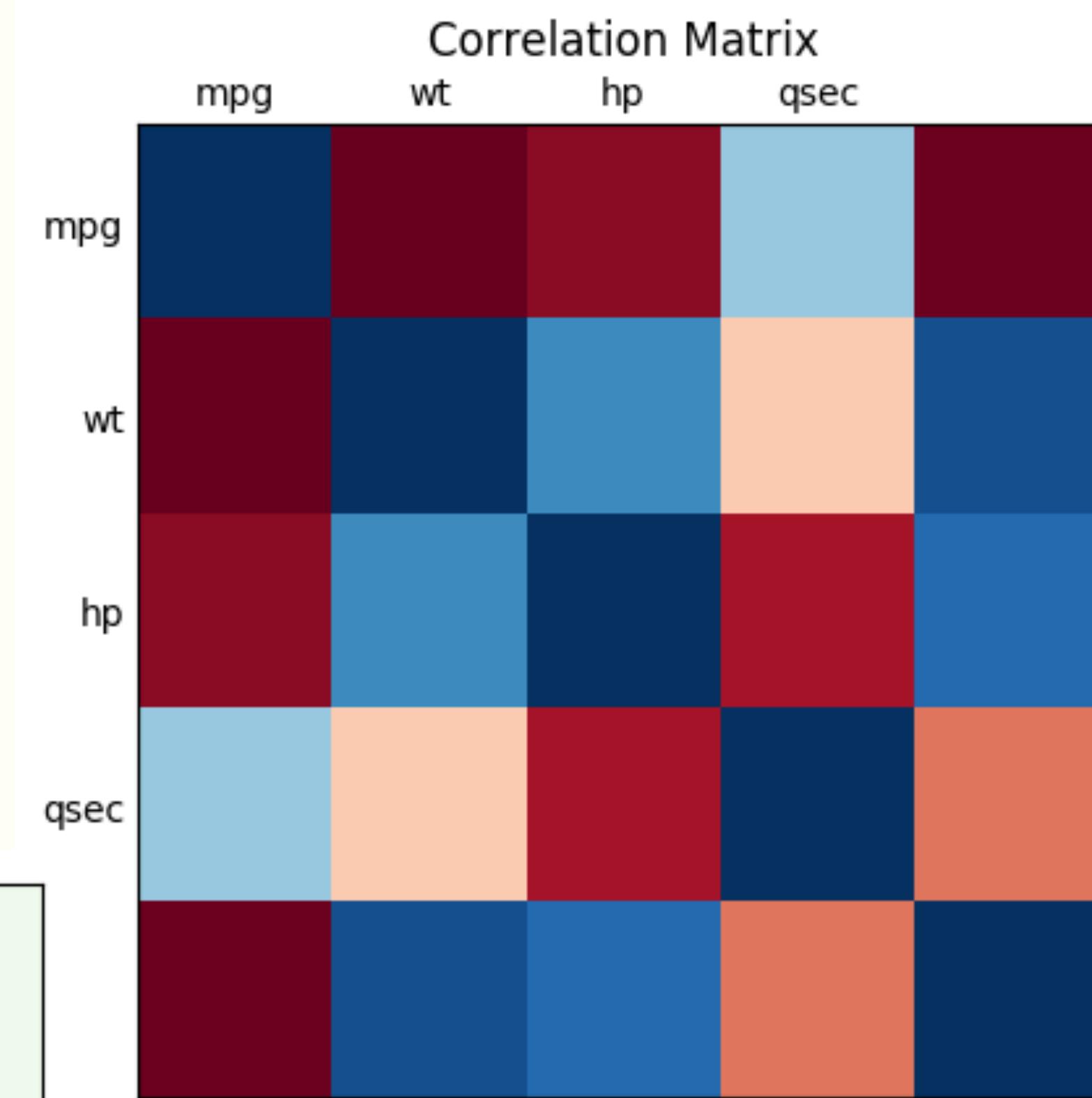
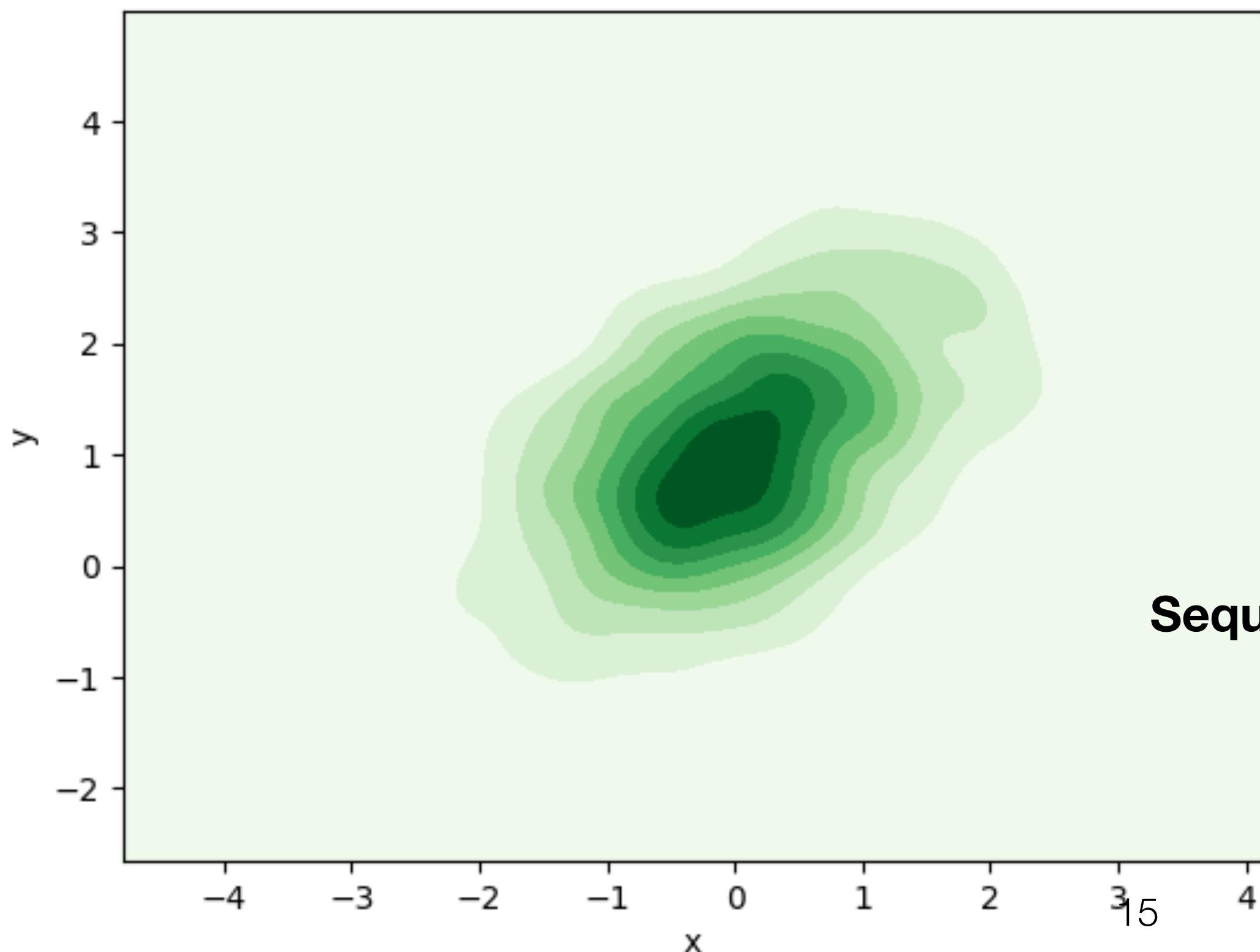
See colorbrewer2.org

Nominal
Ordinal or Quantitative

Qualitative palettes are used for multiple curves in a plot or categoricals



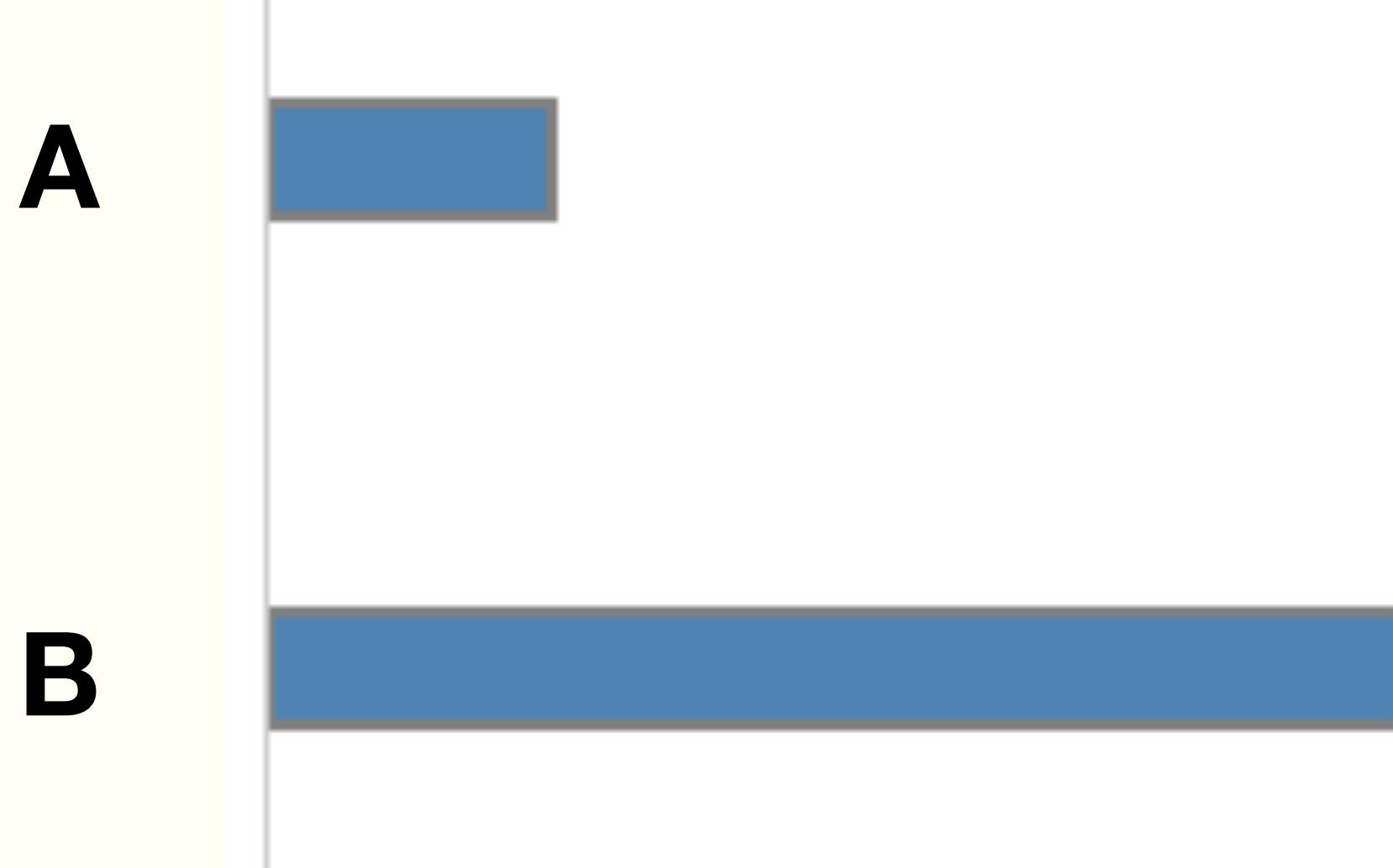
Diverging Palette for Correlations



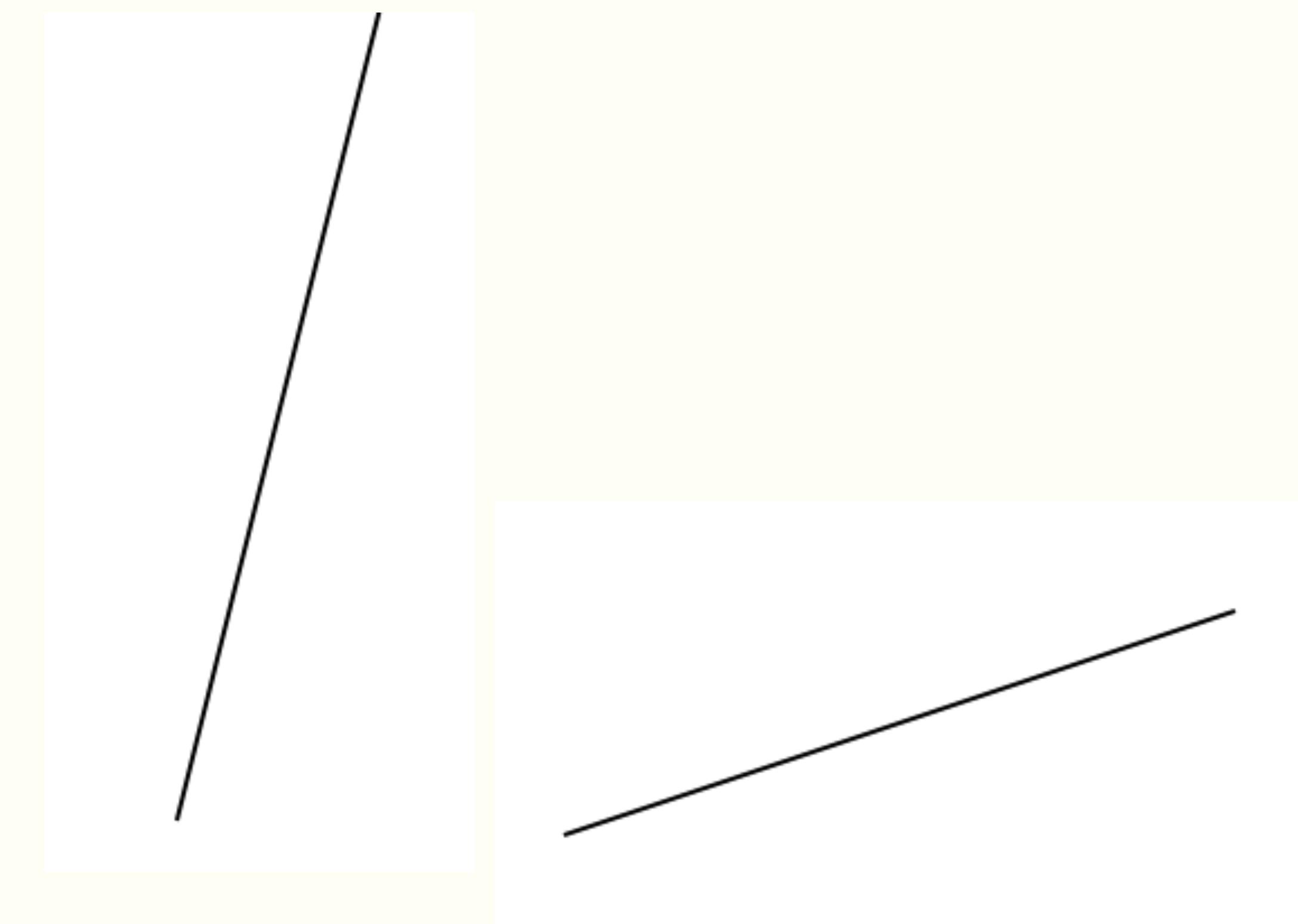
Sequential Palette for Densities (all positive)

Perceptual Effectiveness

How much longer?



How much steeper the slope?



How much longer?

A

B

4x

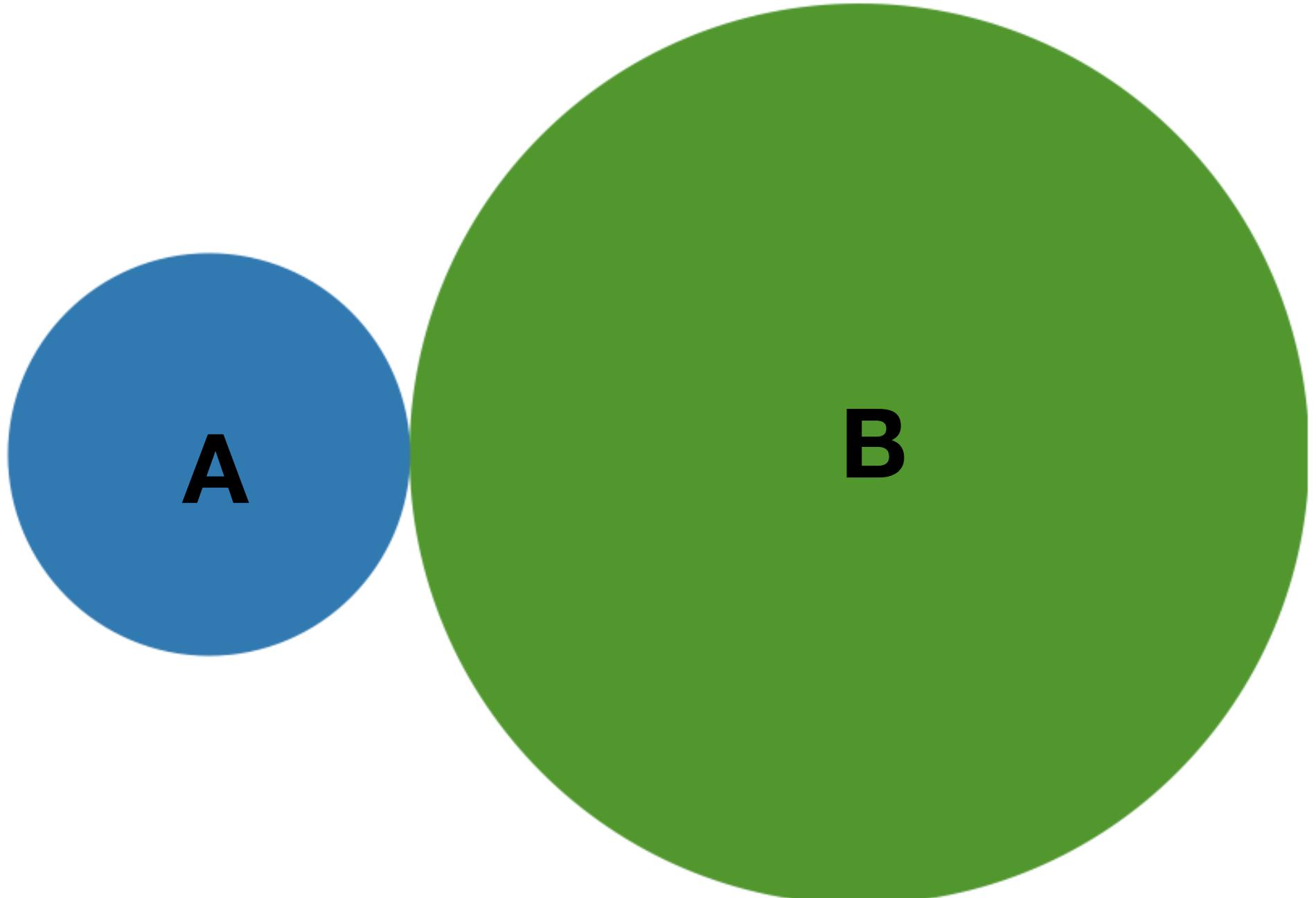
How much steeper the slope?

A

B

4x

How much larger area?

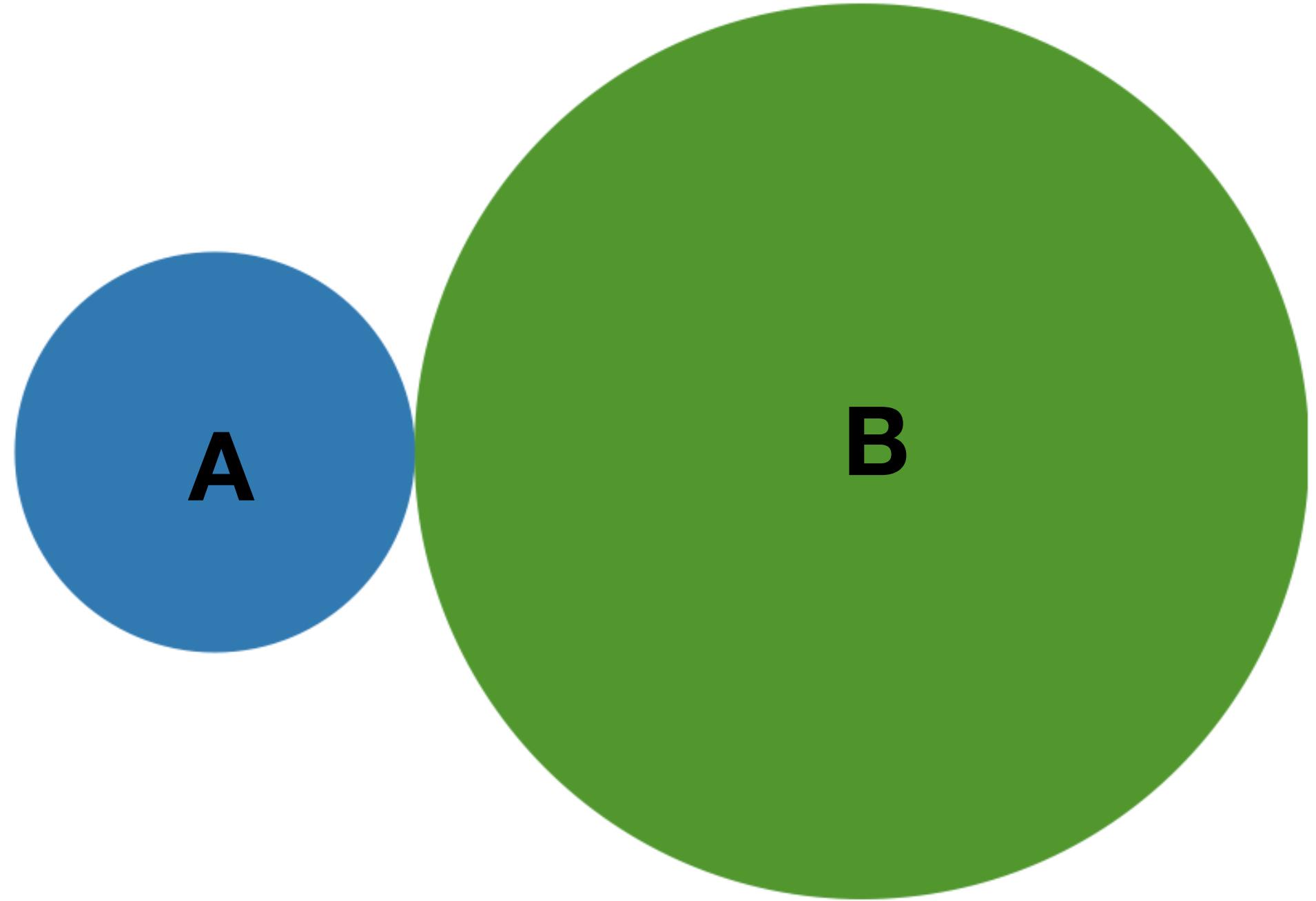


A B

How much Darker?



How much larger area?

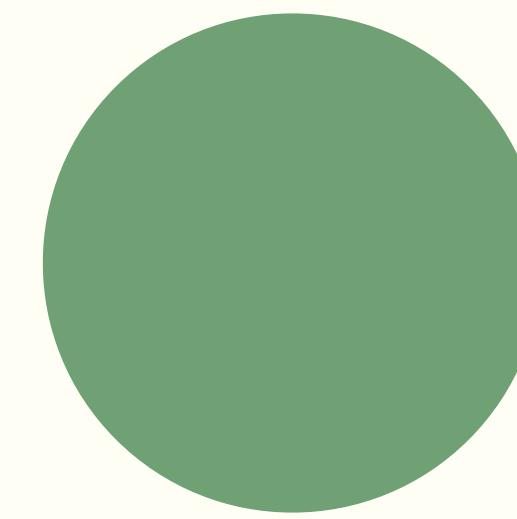


10x

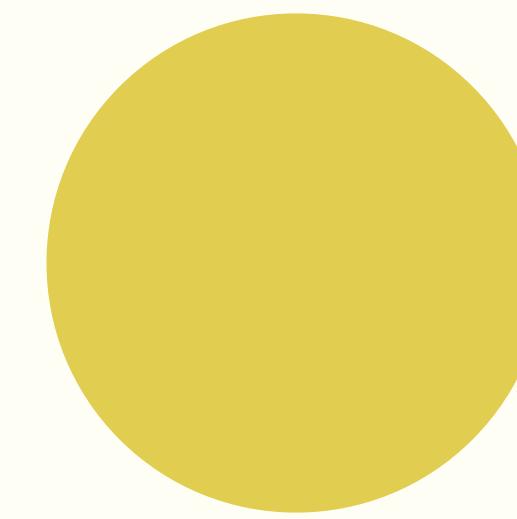


How much Darker?

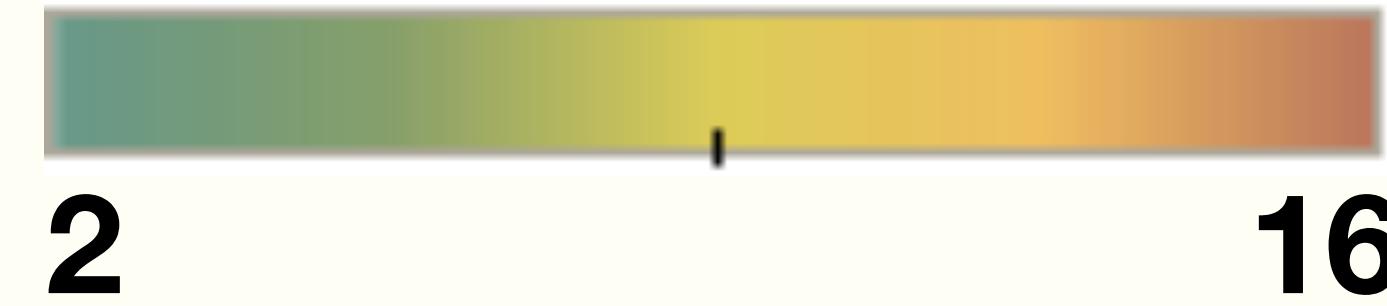
How much bigger value?



A



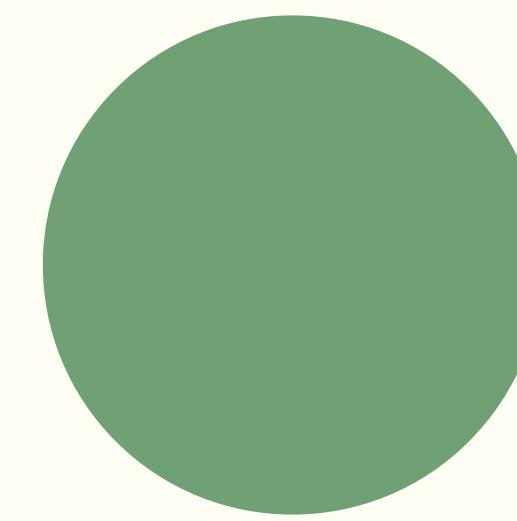
B



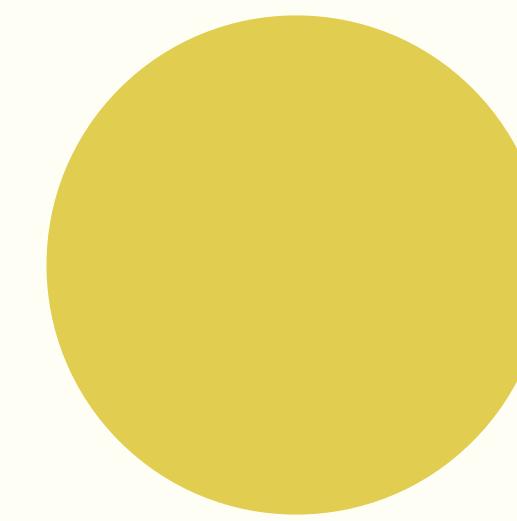
2

16

How much bigger value?

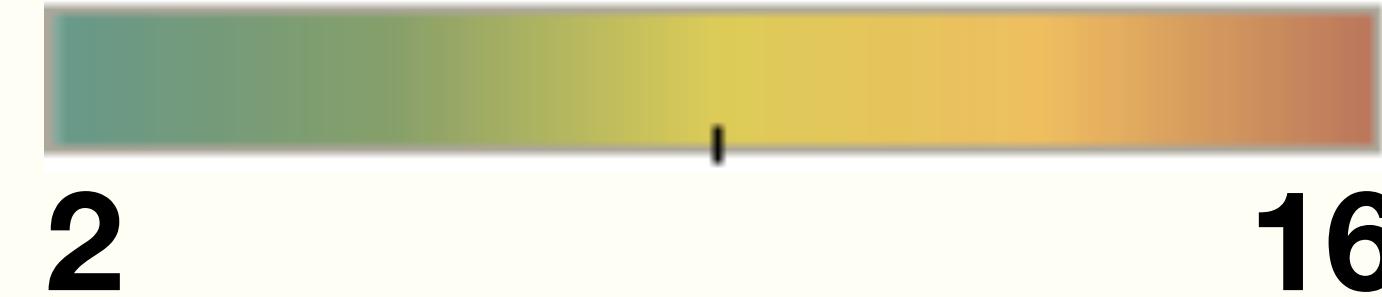


A



B

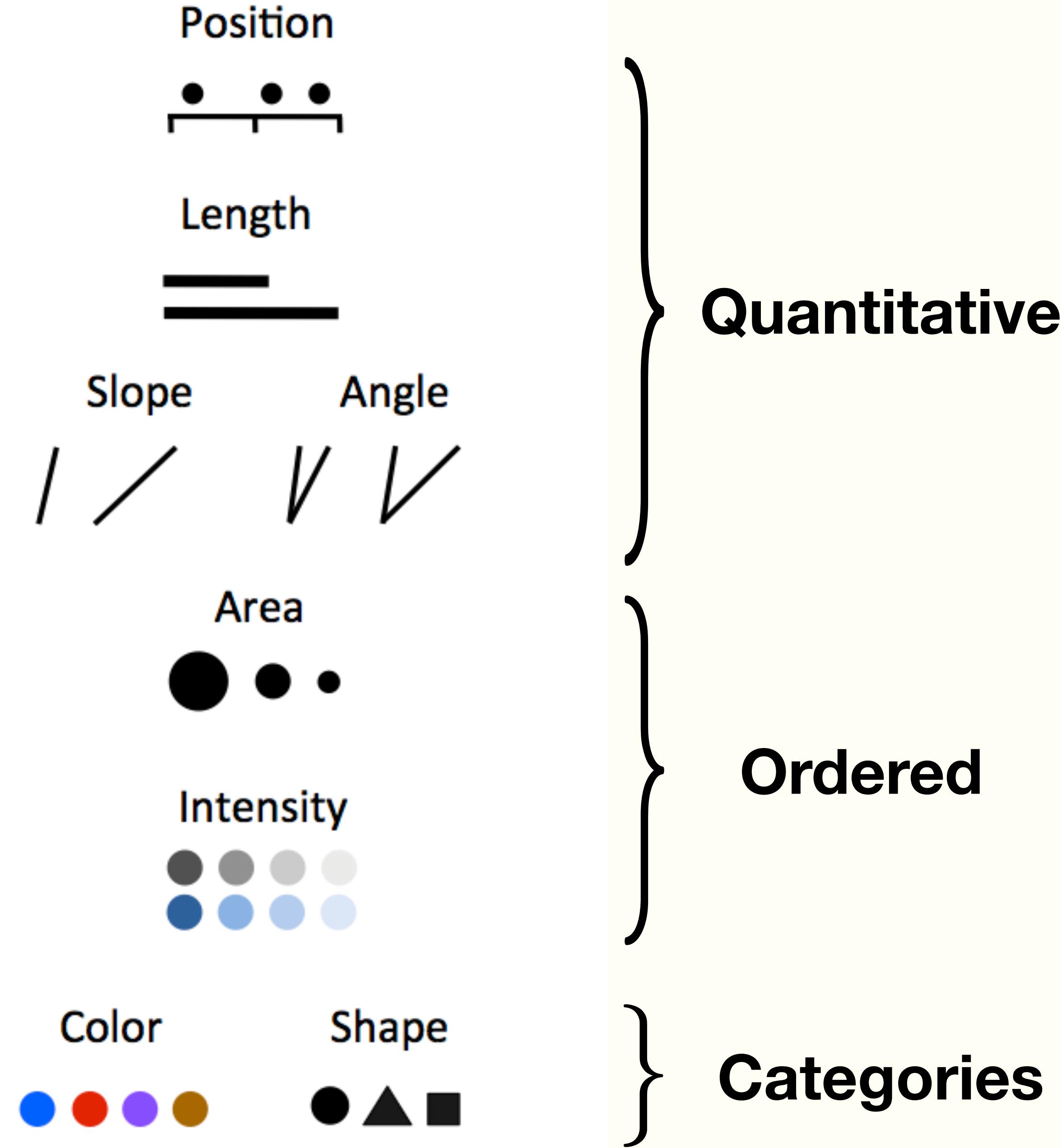
4x



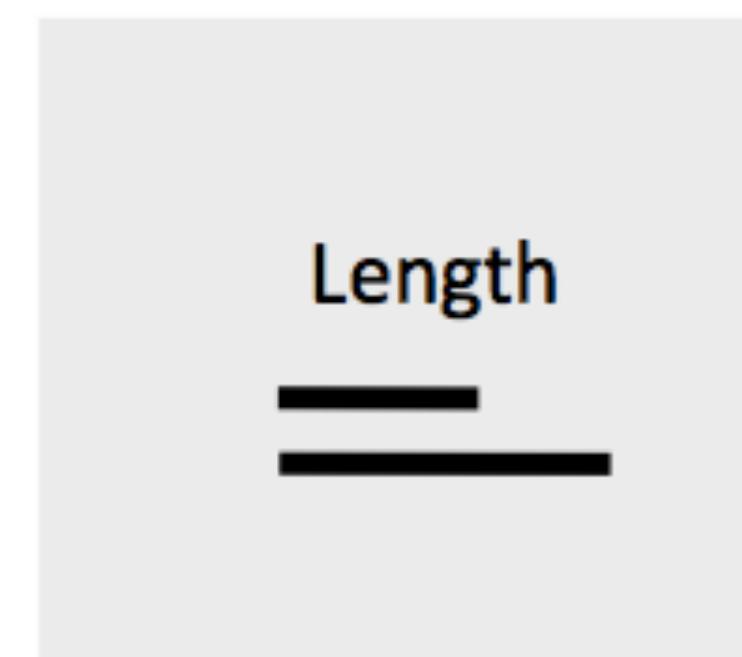
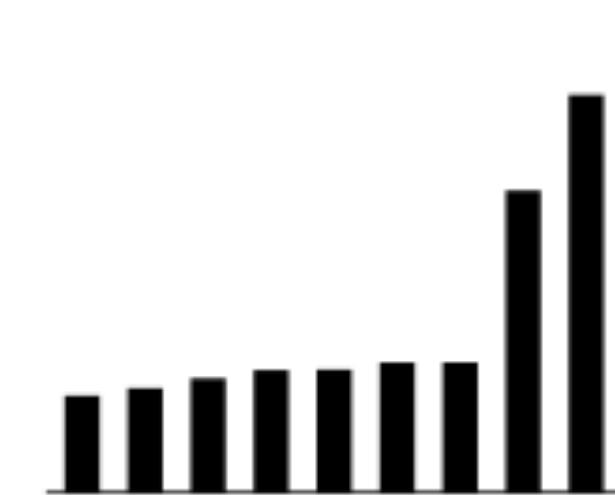
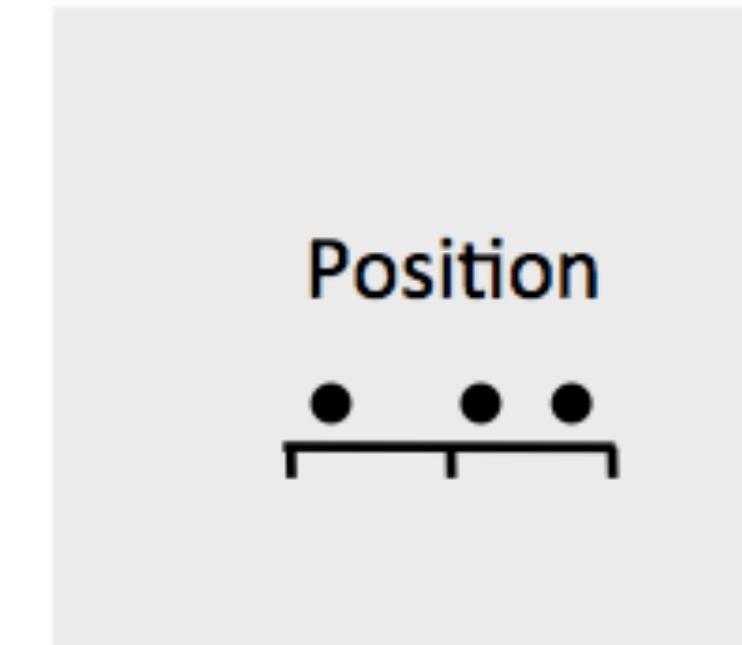
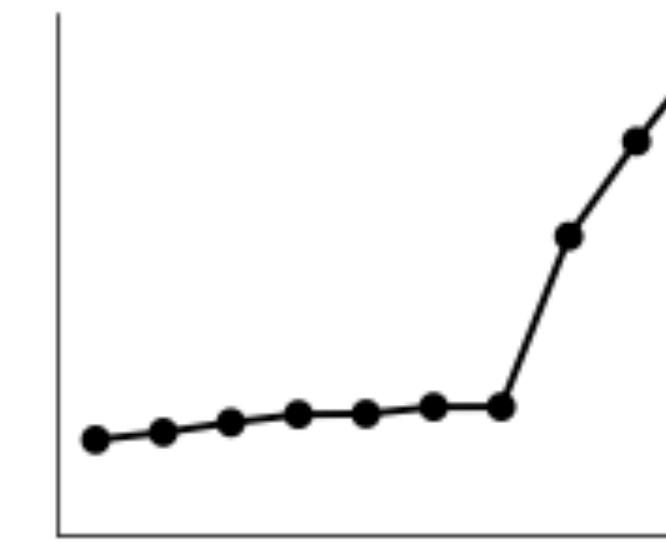
**Most
Efficient**



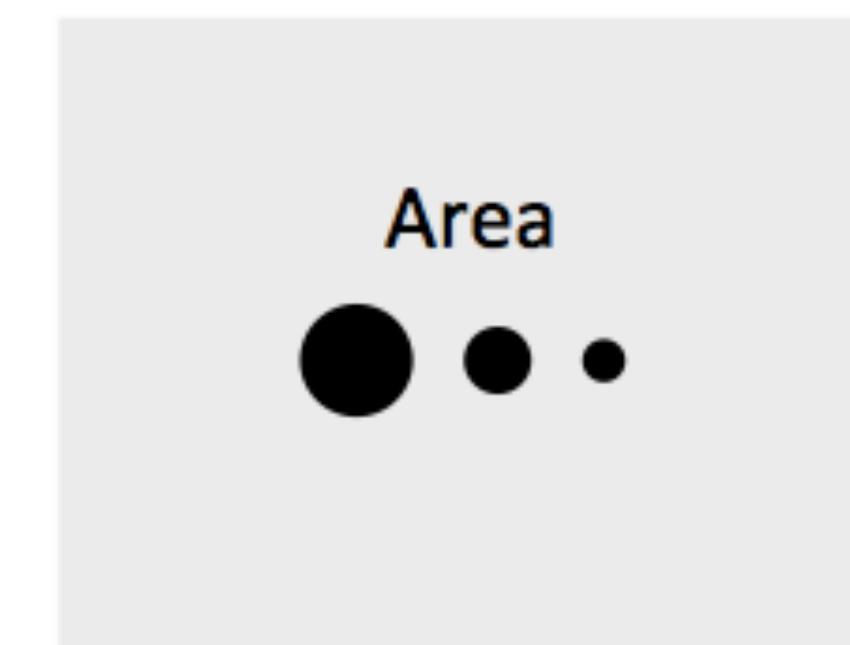
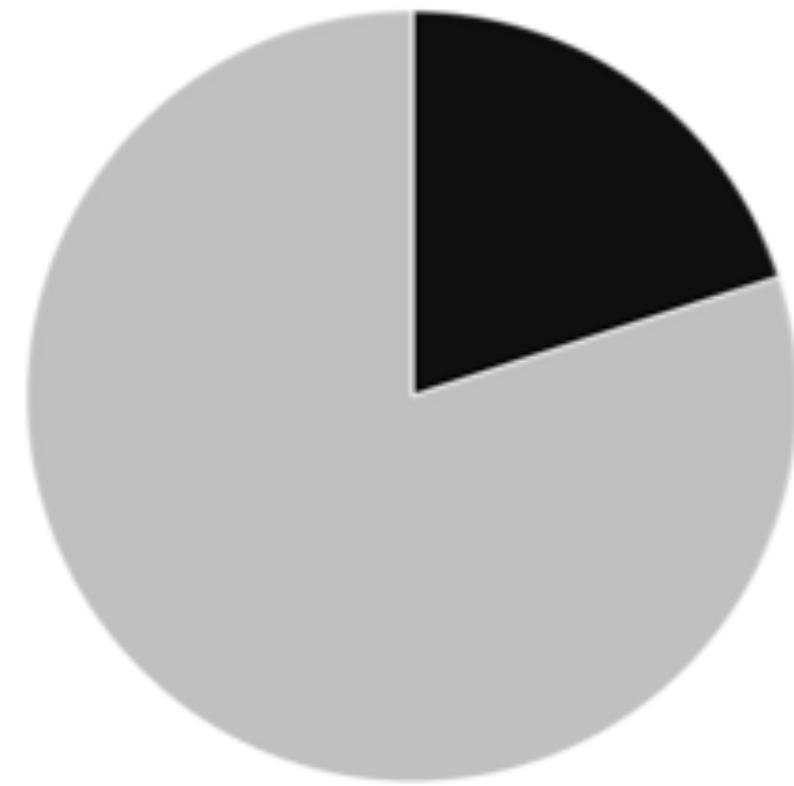
**Least
Efficient**



Most Effective



Less Effective



EDA: Which plots when?

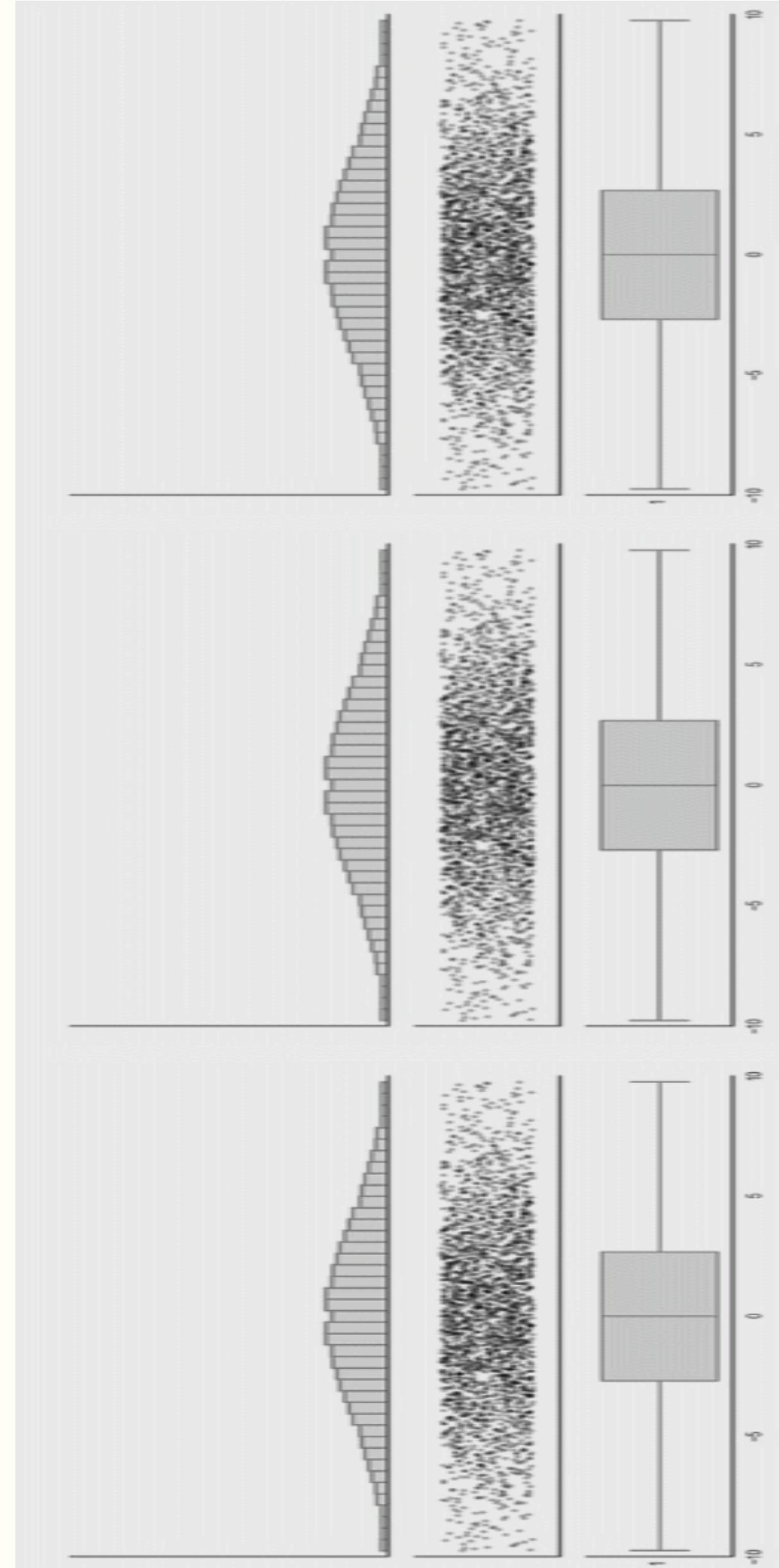
EDA Process

1. Visualize and Summarize individual features
2. Viz and summarize combination feature and target distributions
3. Transform targets and features to make them more amenable to modeling.

EDA Rubric

1. Explore **global properties**. Use histograms, scatter plots, parallelism, and aggregation functions to summarize the data.
2. Explore **group properties**. Use groupby and faceting and small multiples to compare subsets of the data.
3. Transform targets and features to make them more amenable to modeling.

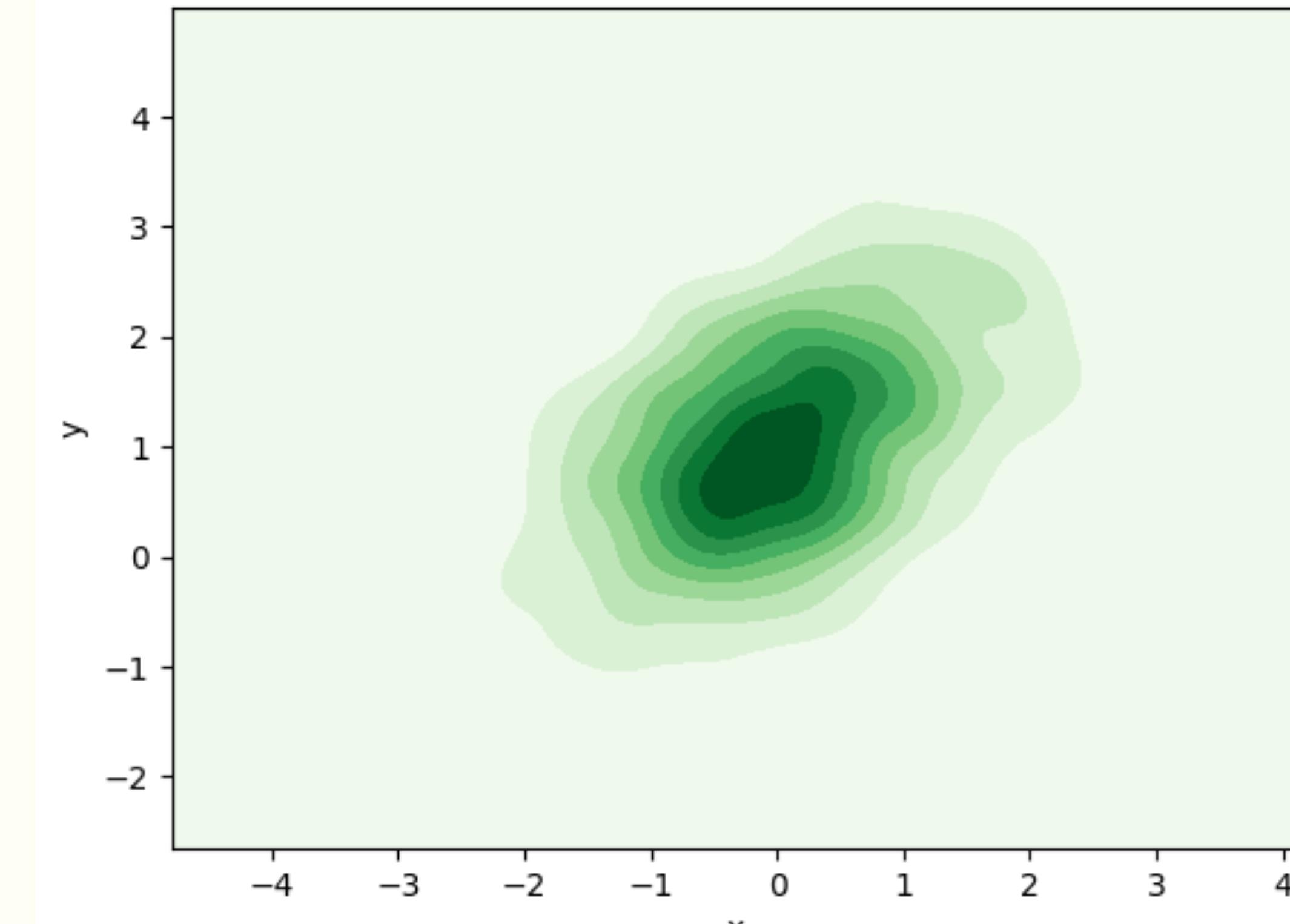
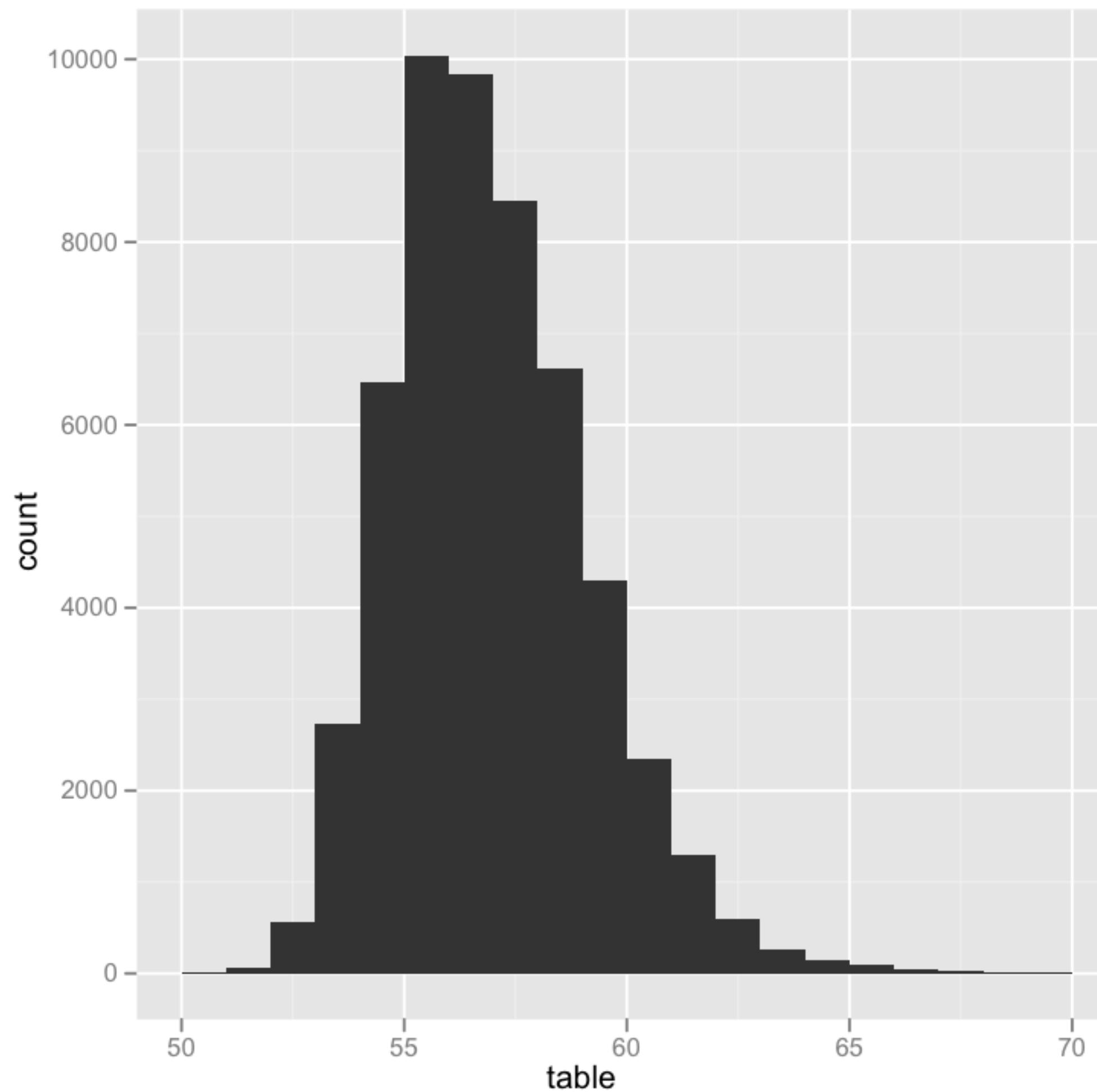
Simple Summaries: Box Plots



[https://www.autodeskresearch.com/
publications/samestats](https://www.autodeskresearch.com/publications/samestats)

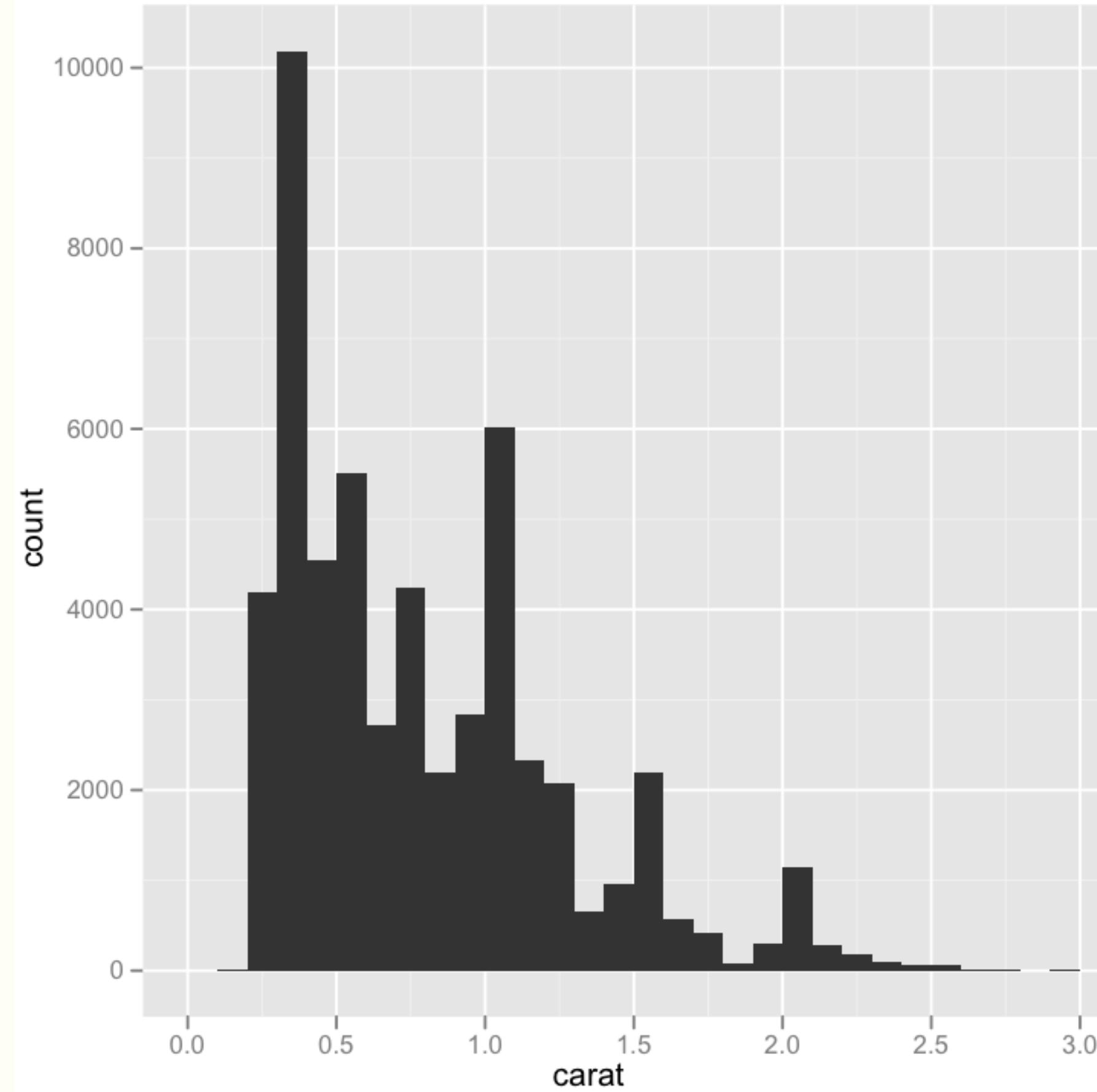
From Summaries to Distributions

Histogram

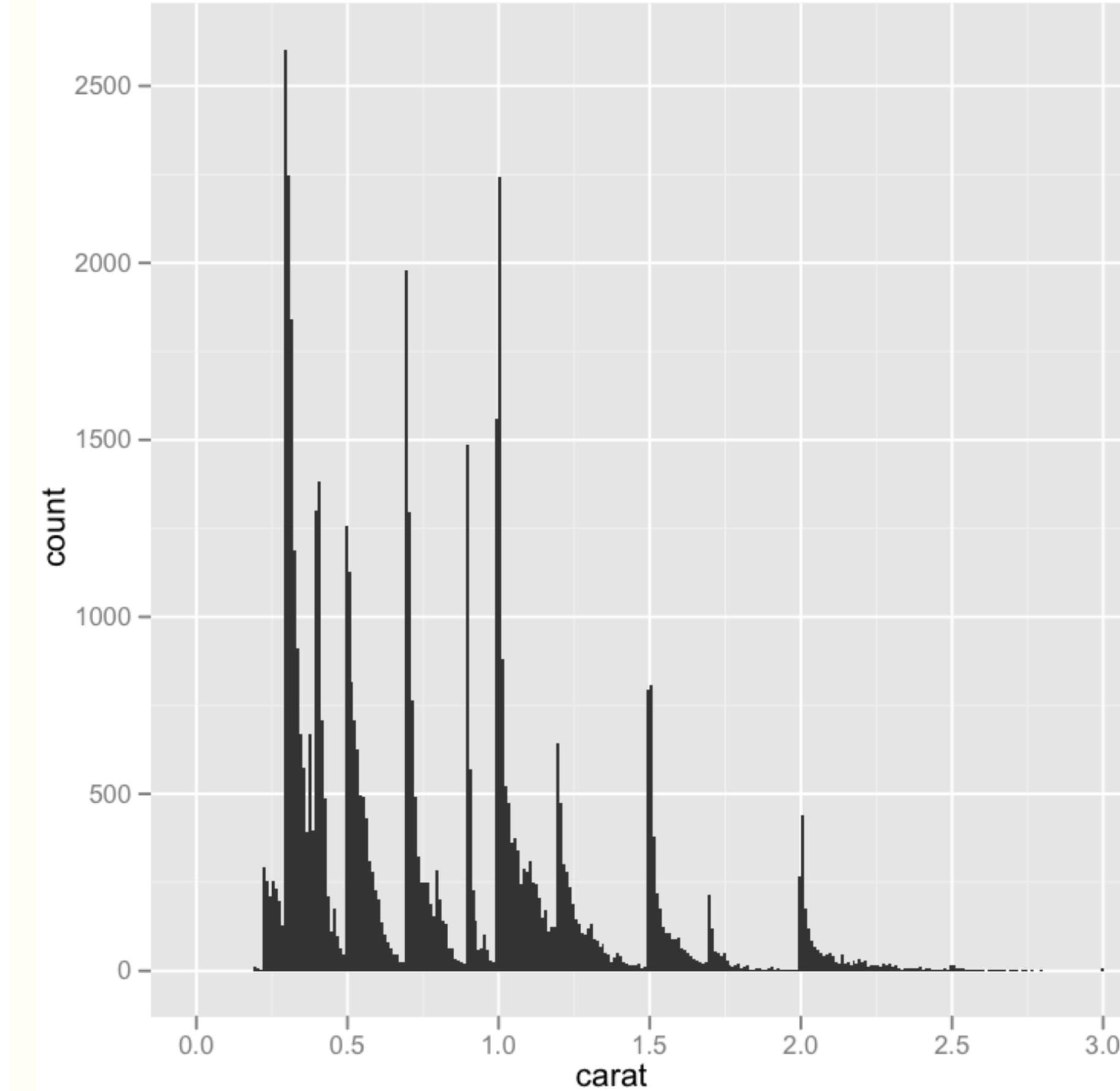


2-D Kernel Density Estimate (KDE)

Bin Width is Important

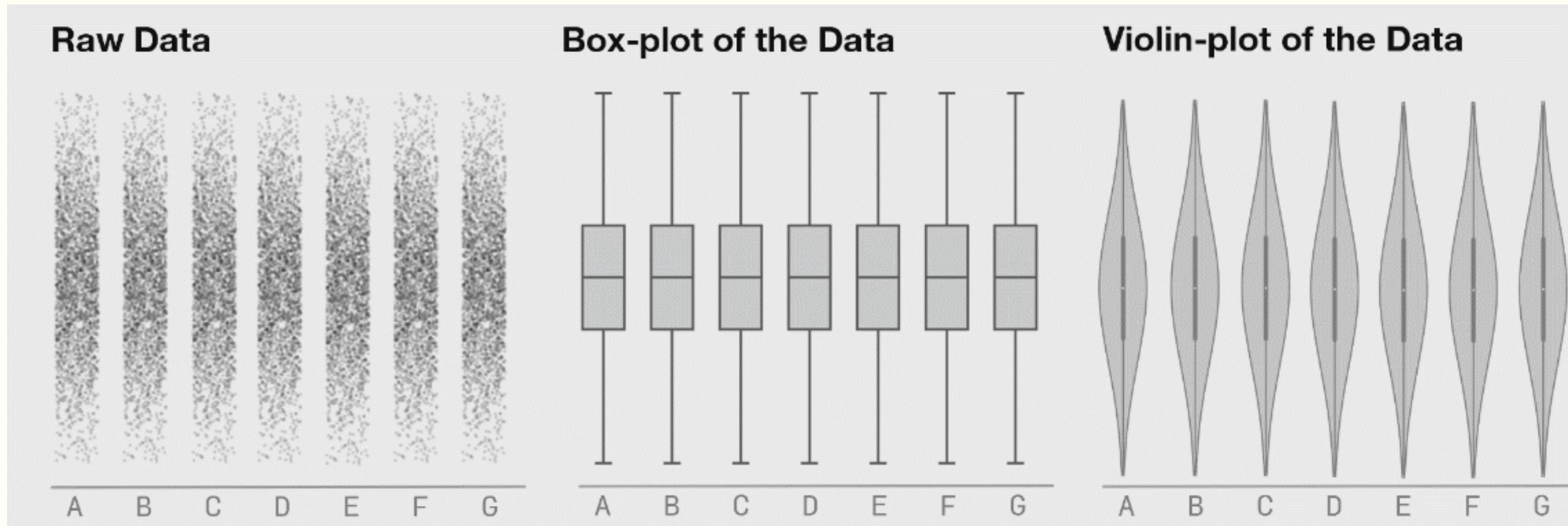


binwidth = 0.1



binwidth = 0.01

Summary Distributions: Density Plots

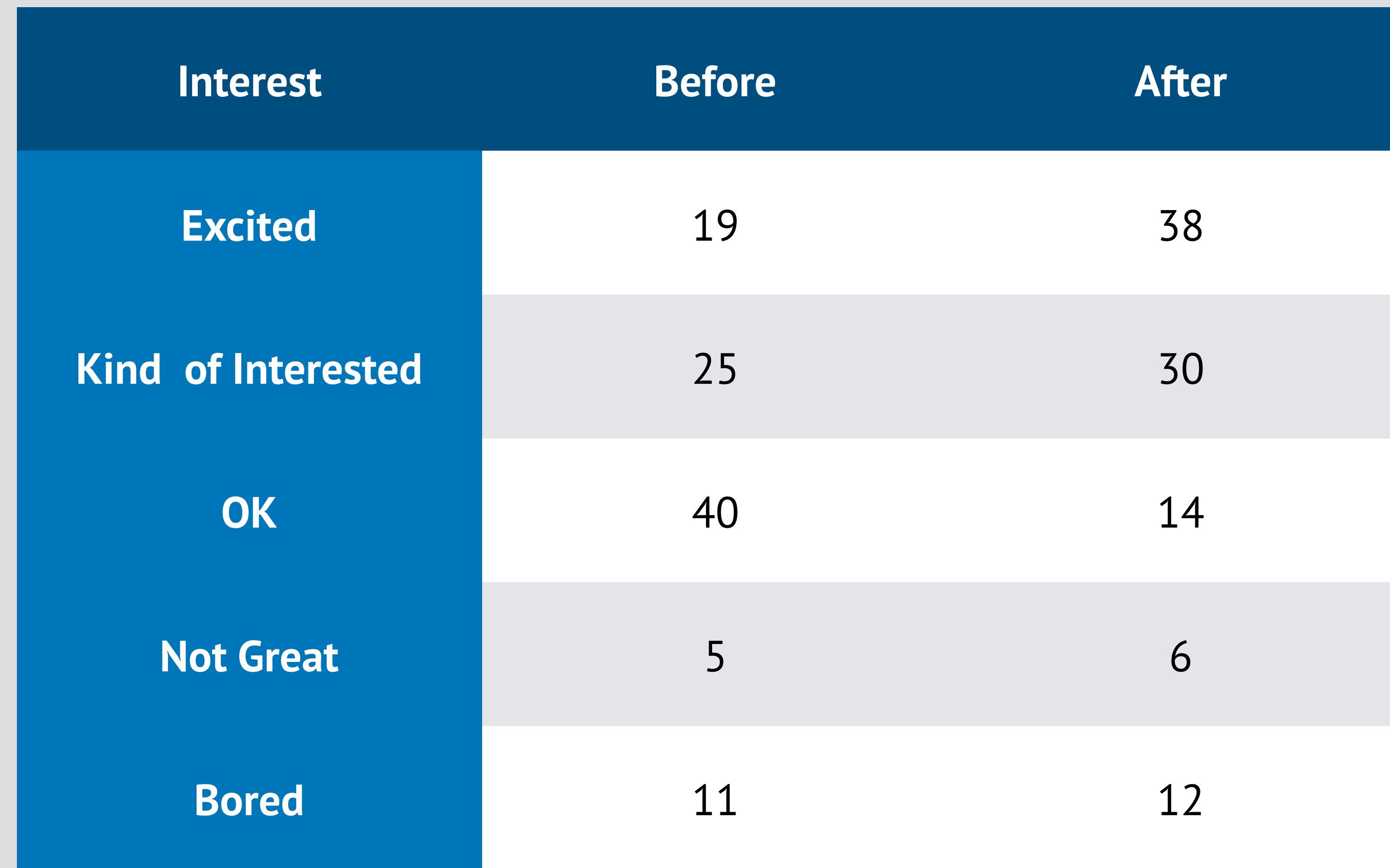


<https://www.autodeskresearch.com/publications/samestats>

Design Exercise

Hands-On Exercise

How do you feel about doing science?

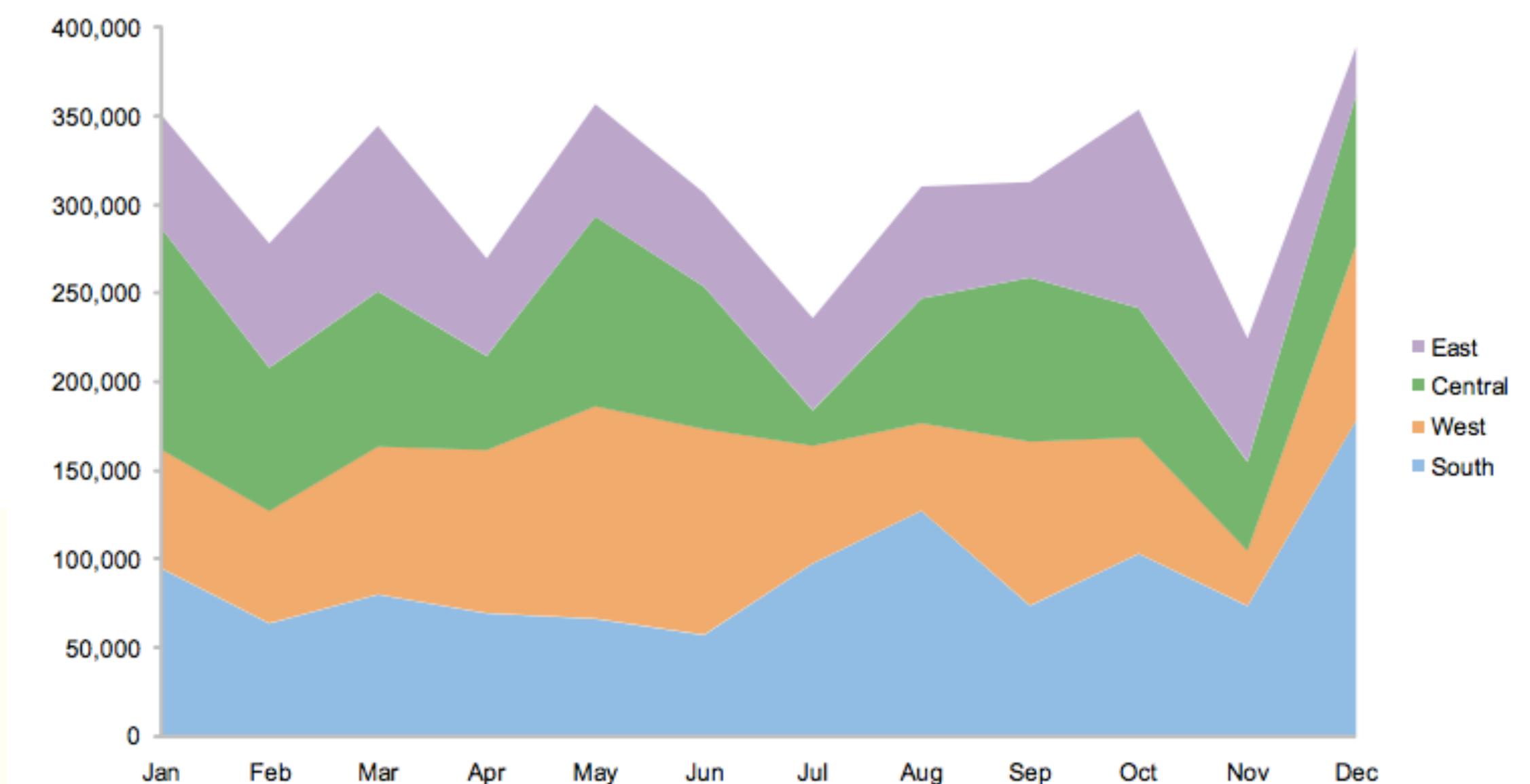
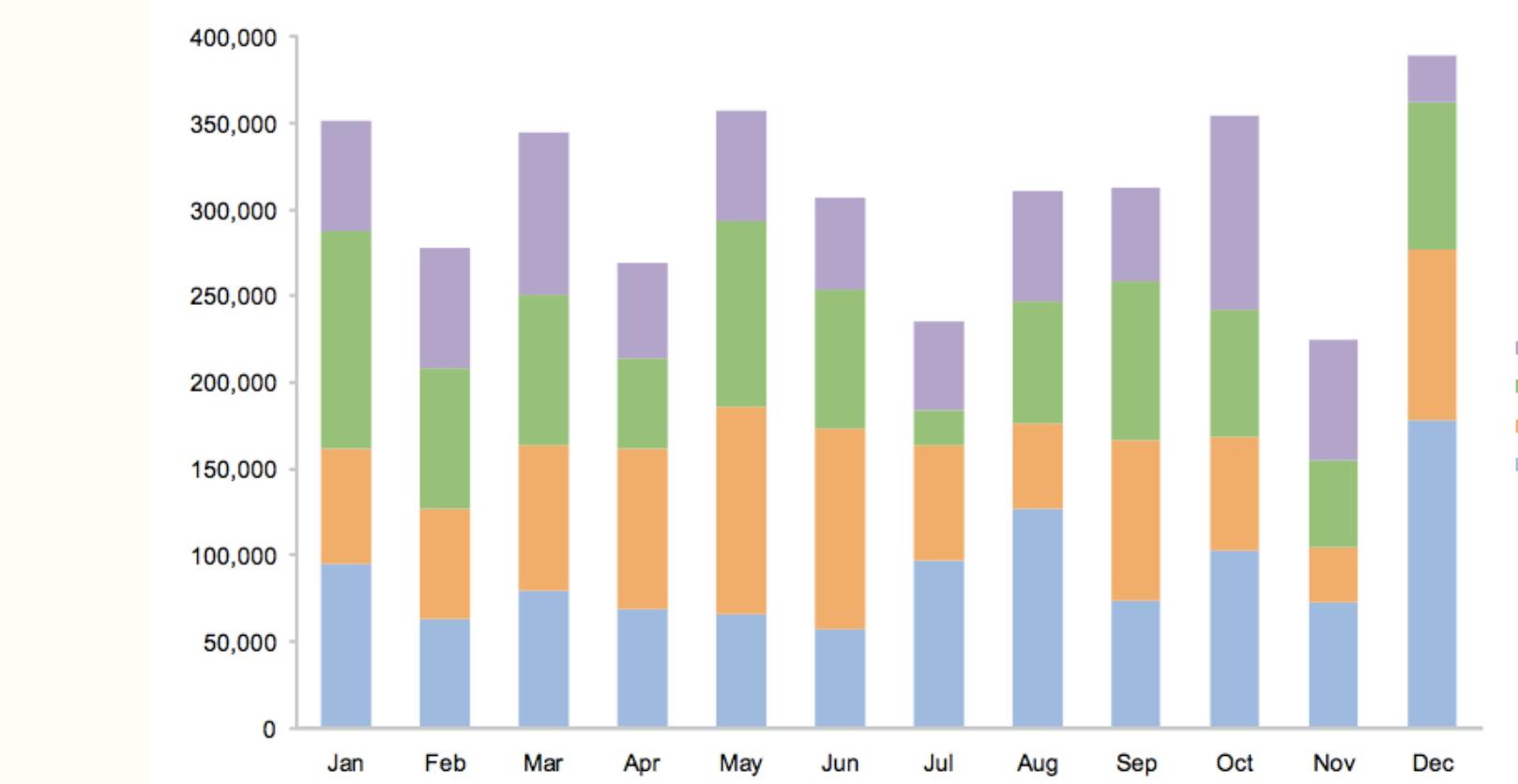
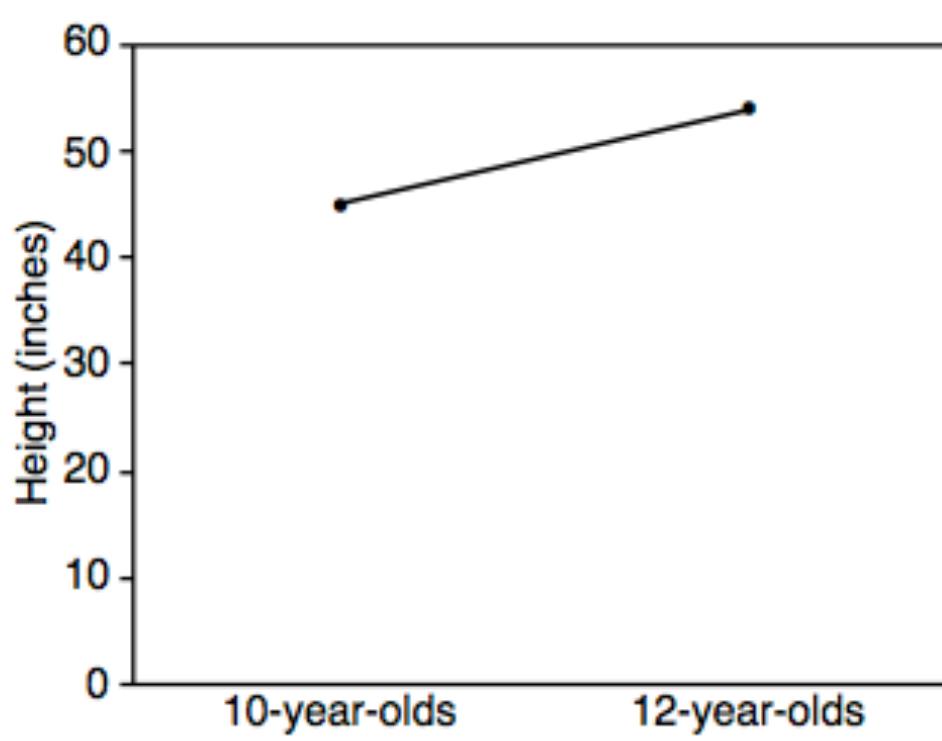
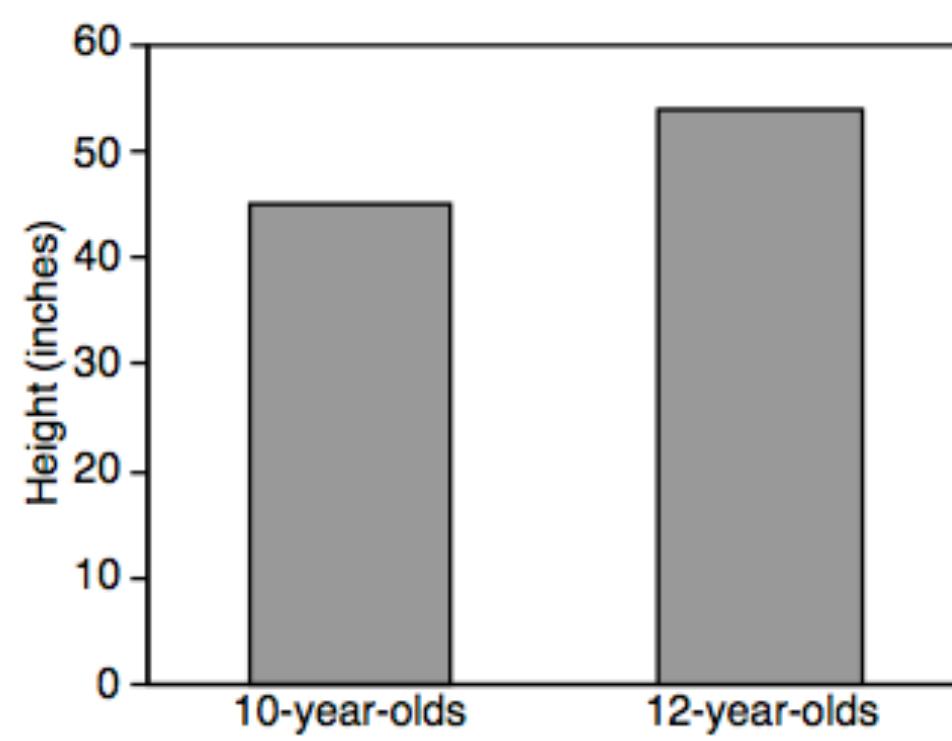
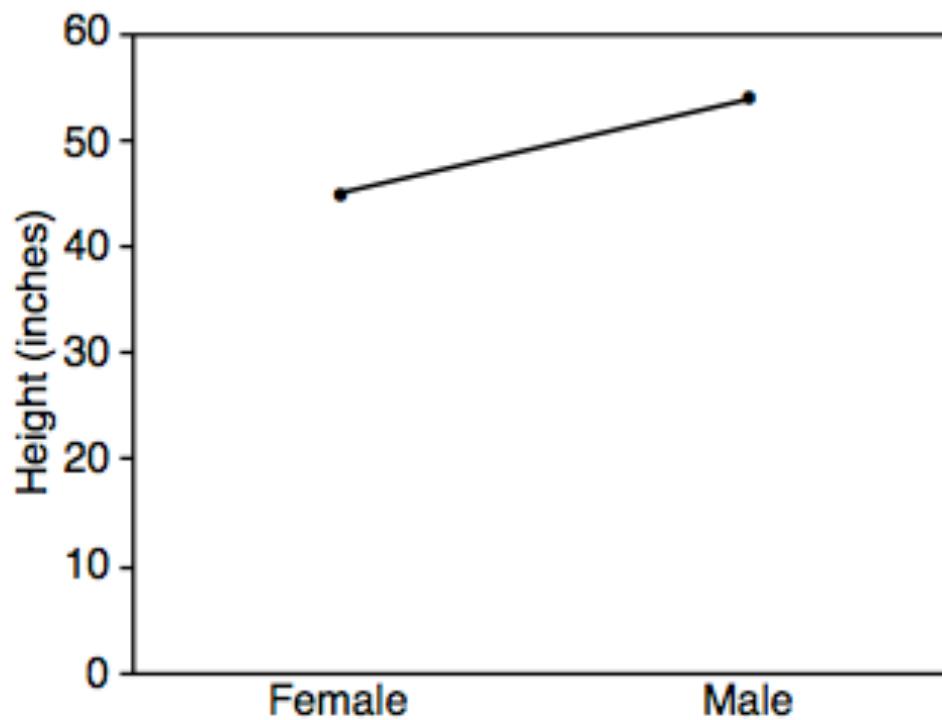
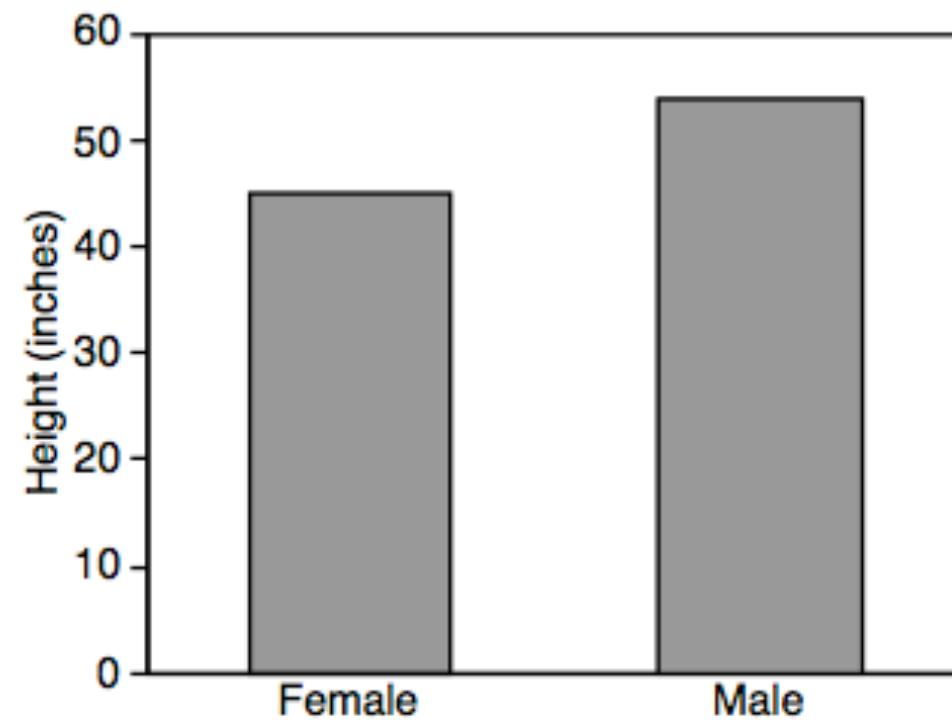


Data courtesy of Cole Nussbaumer

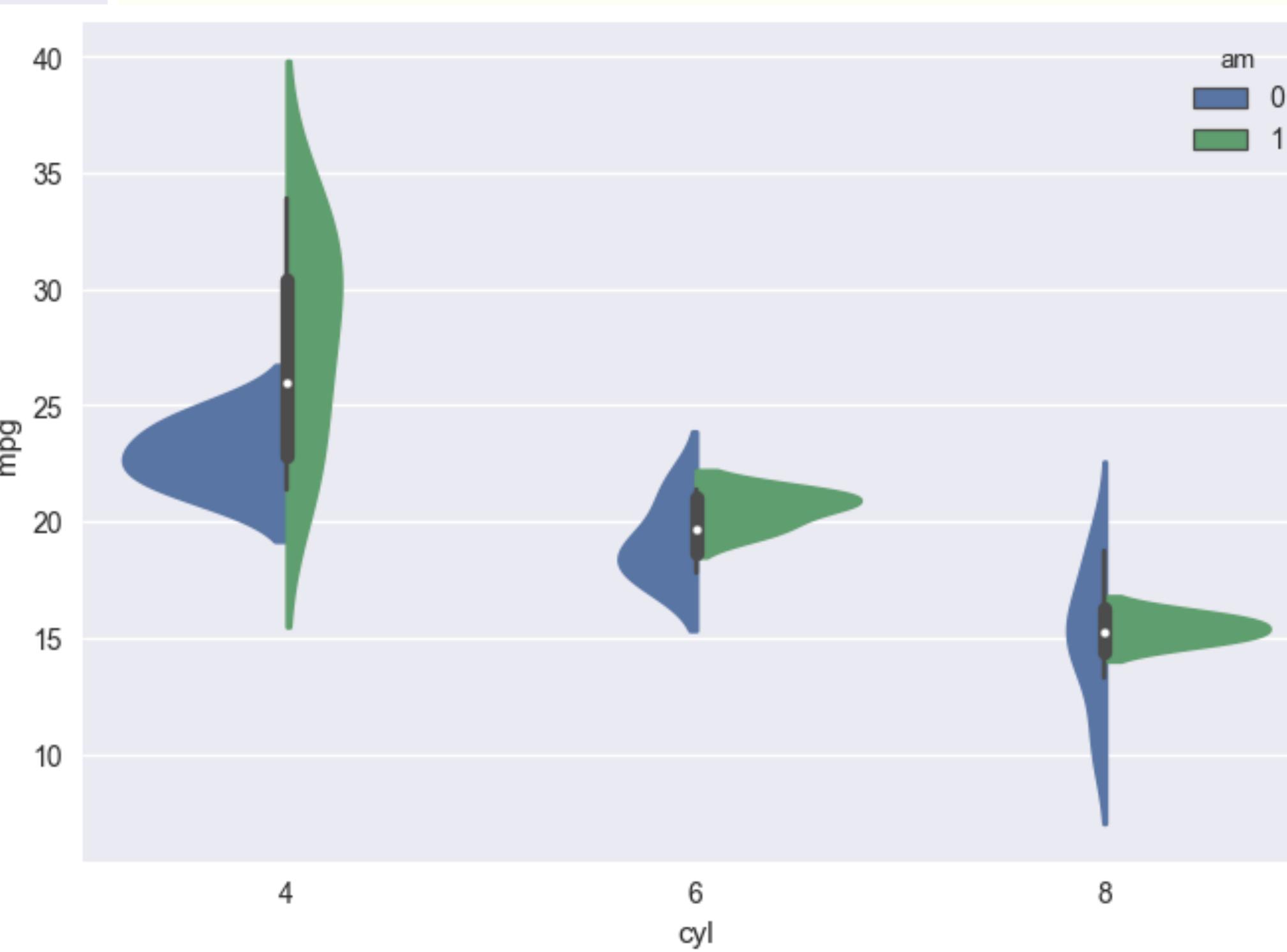
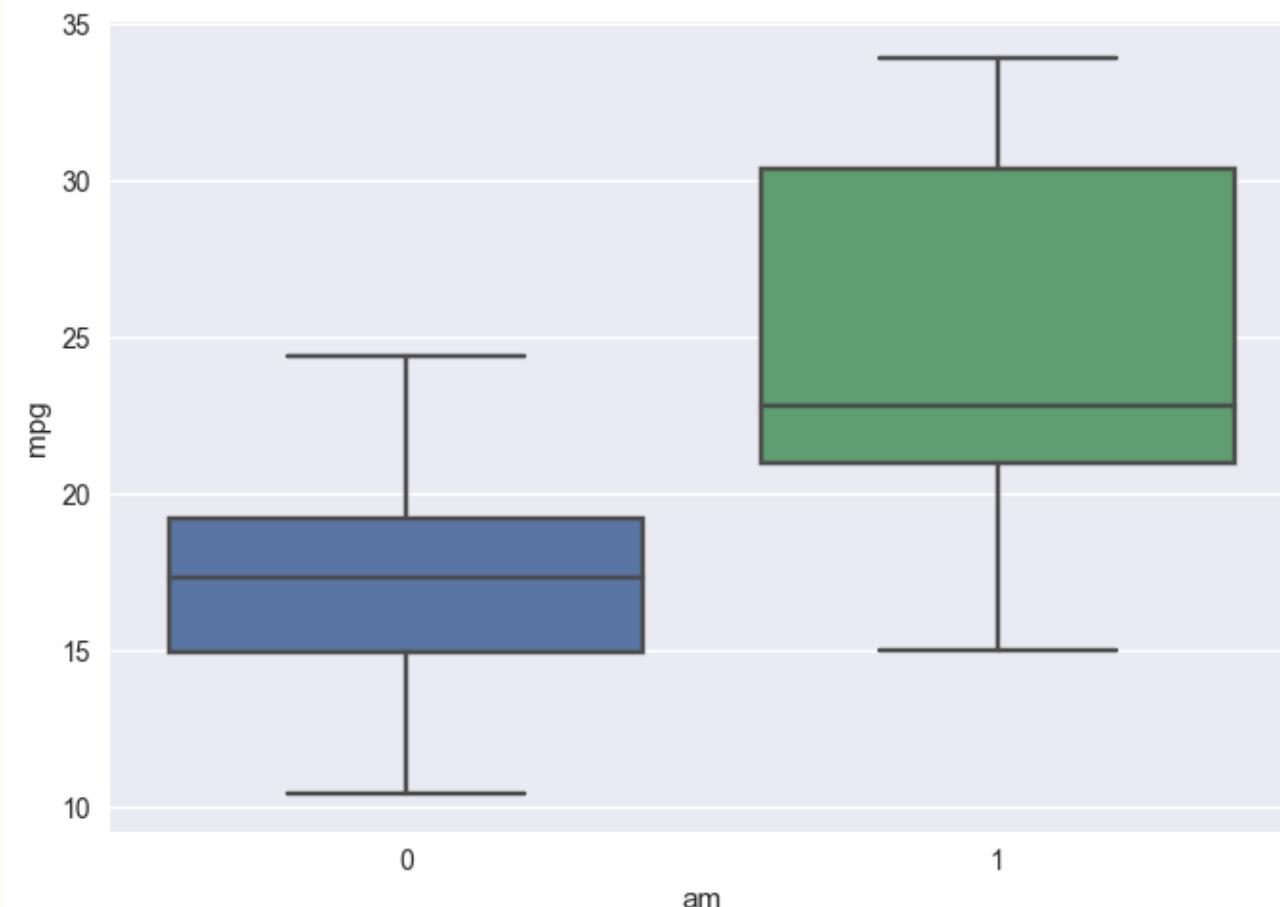
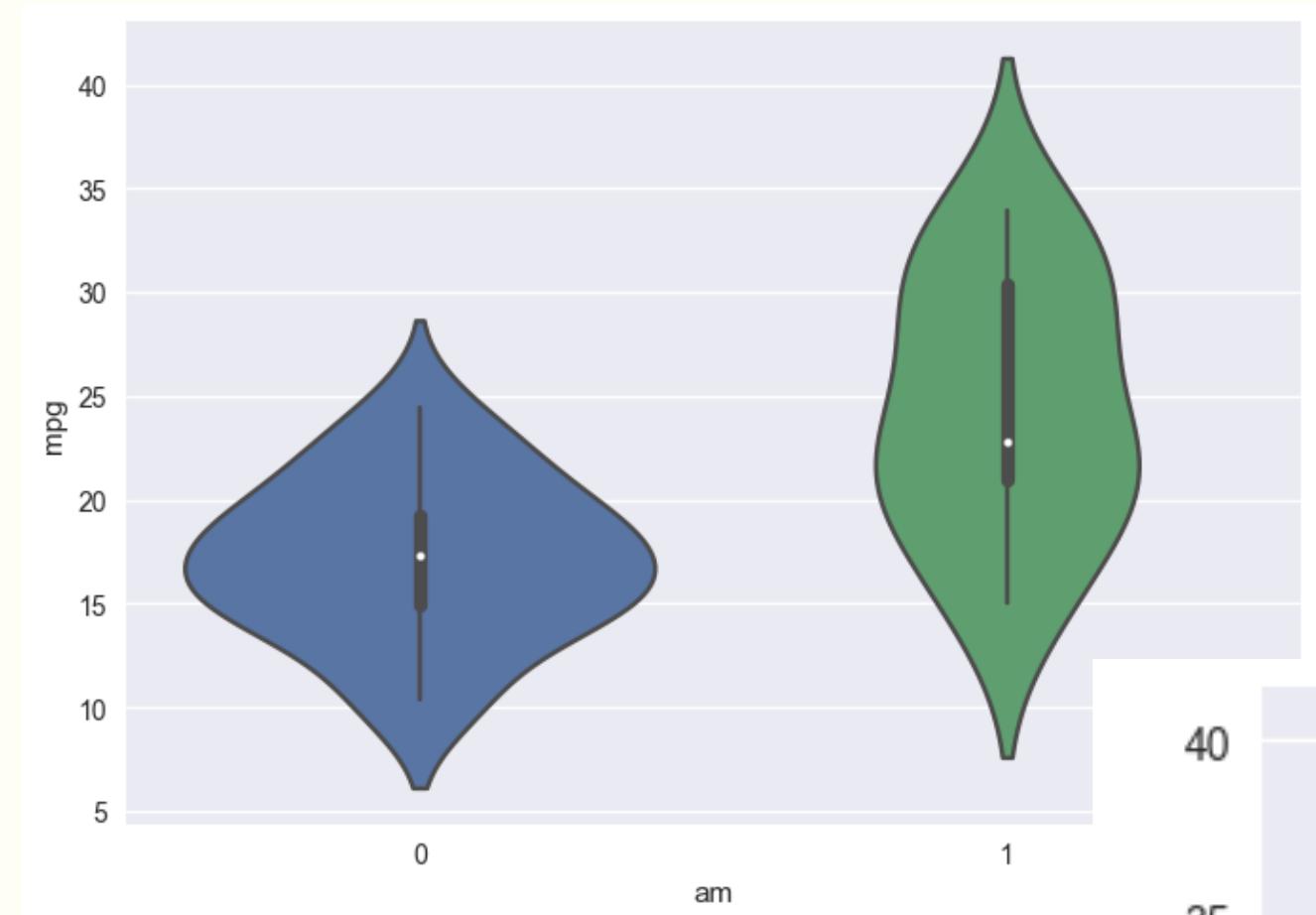
Come up with multiple visualizations.

Pen and Paper Only. Pies, Bars, What else?

Grouping

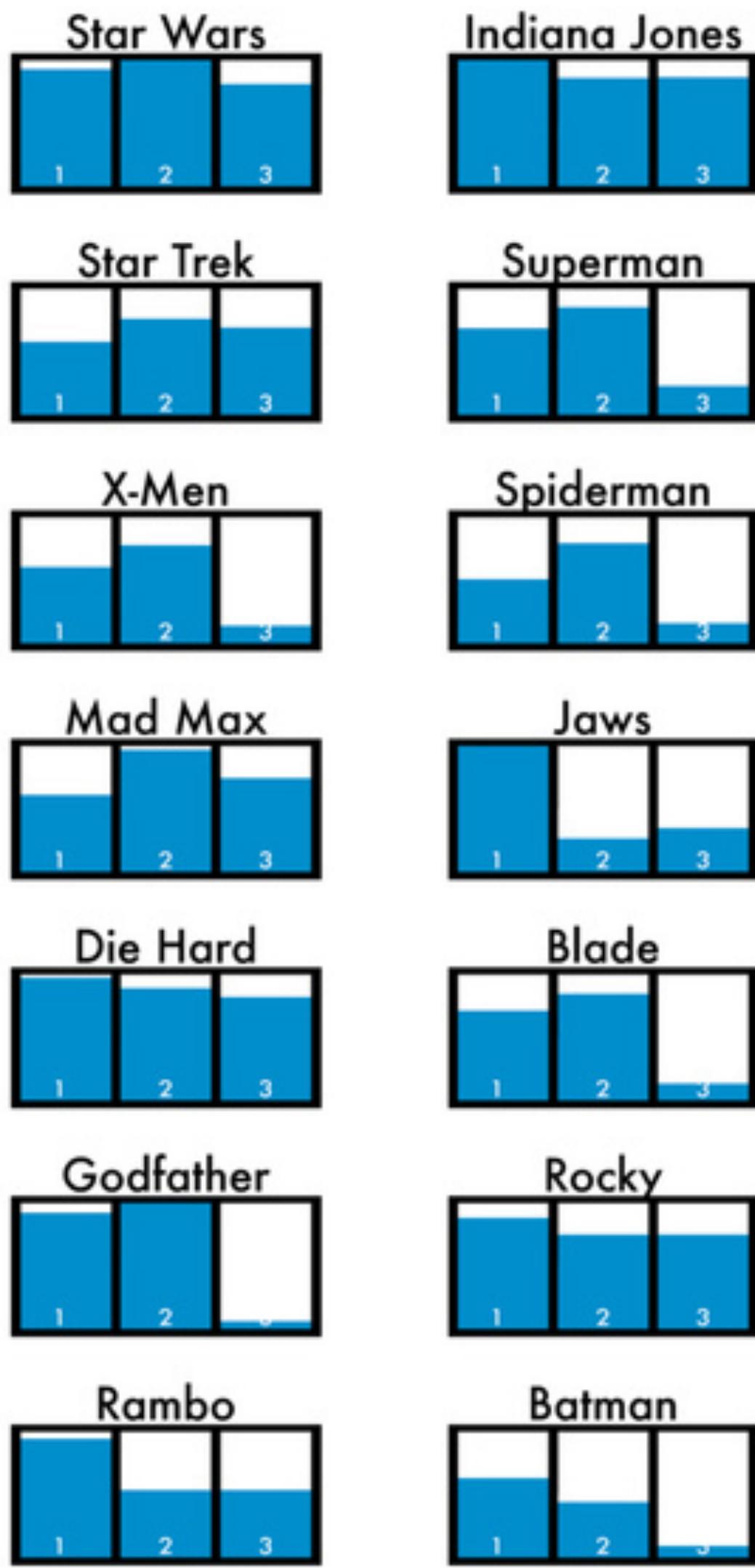


GROUP



getting complex...

THE TRILOGY METER

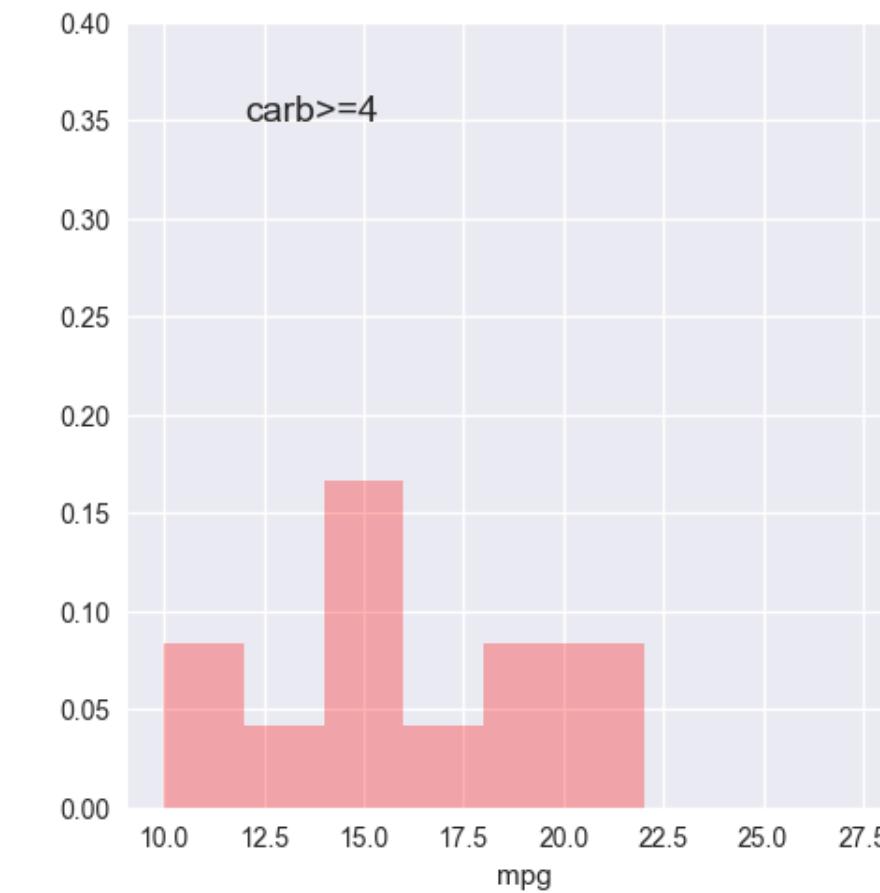
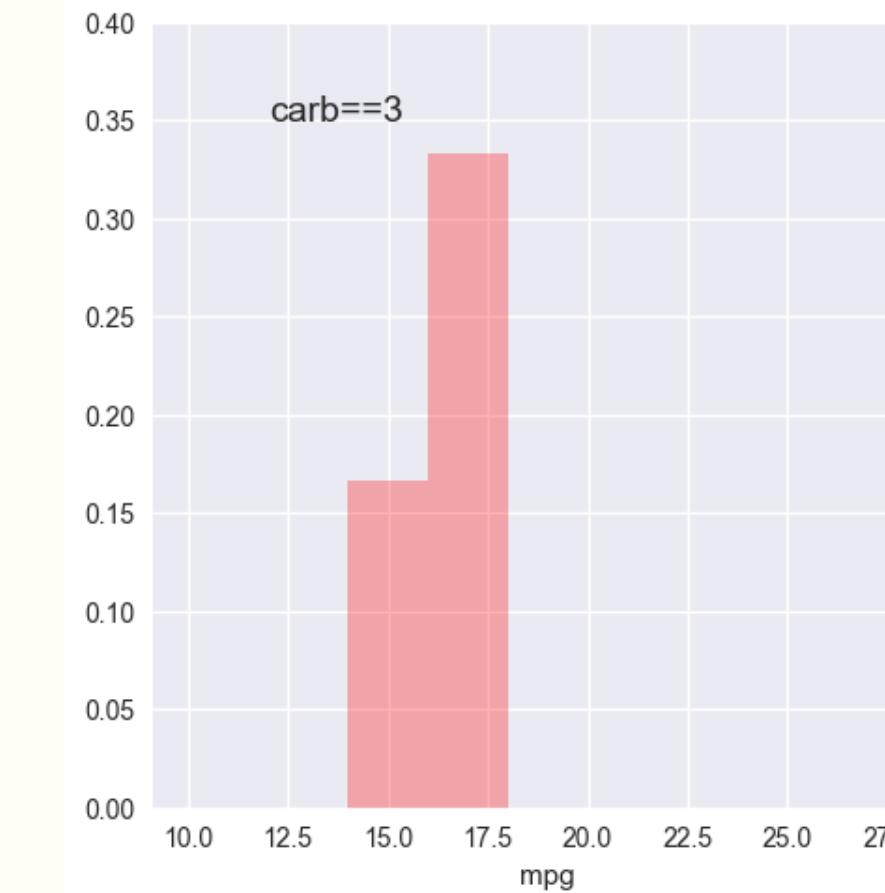
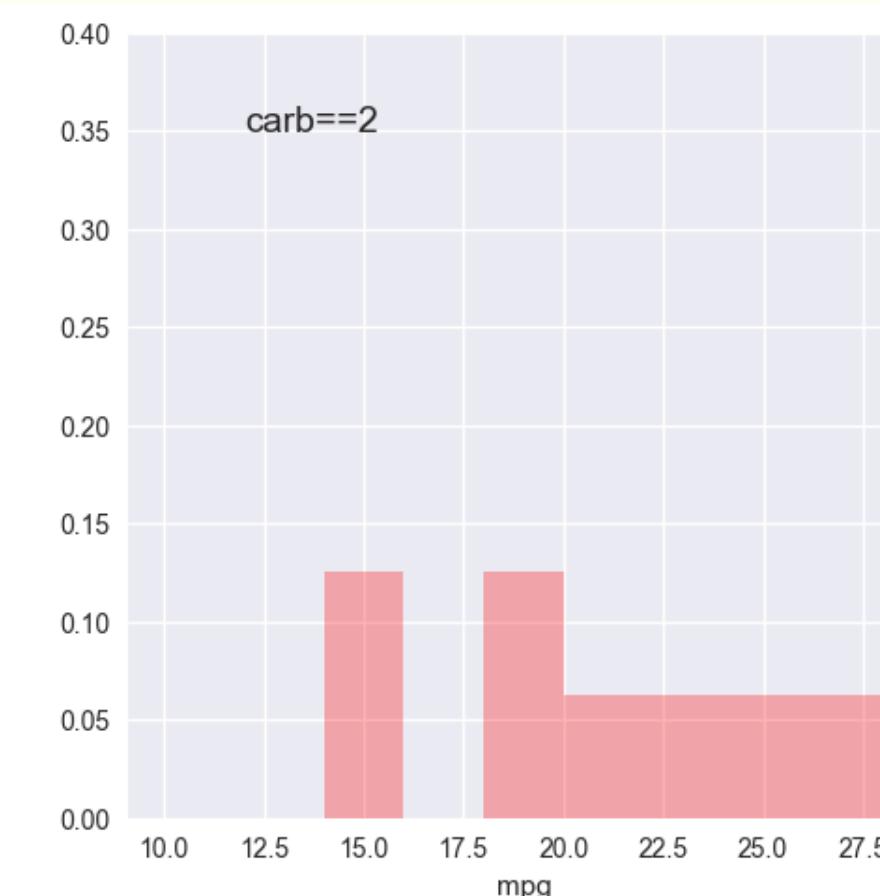
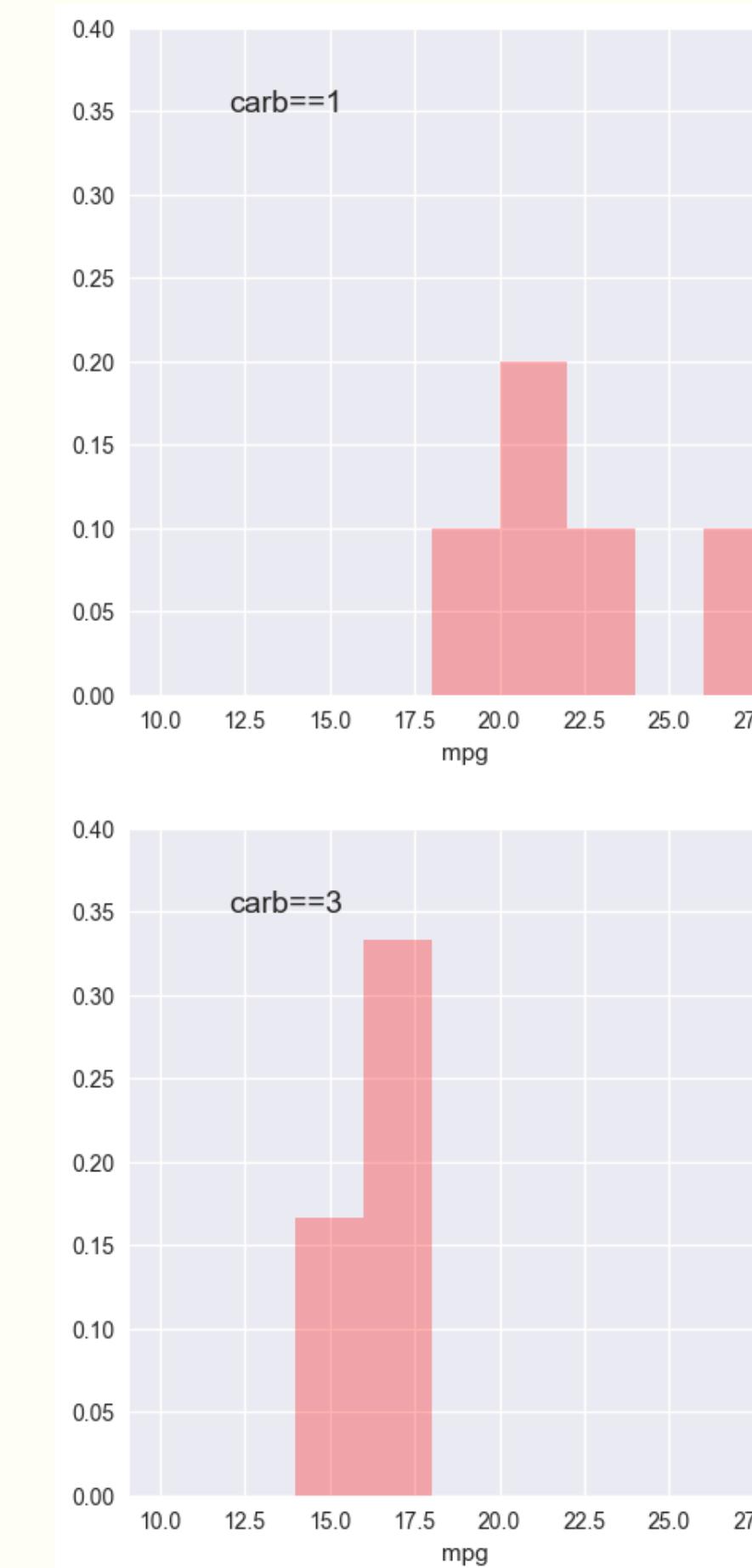


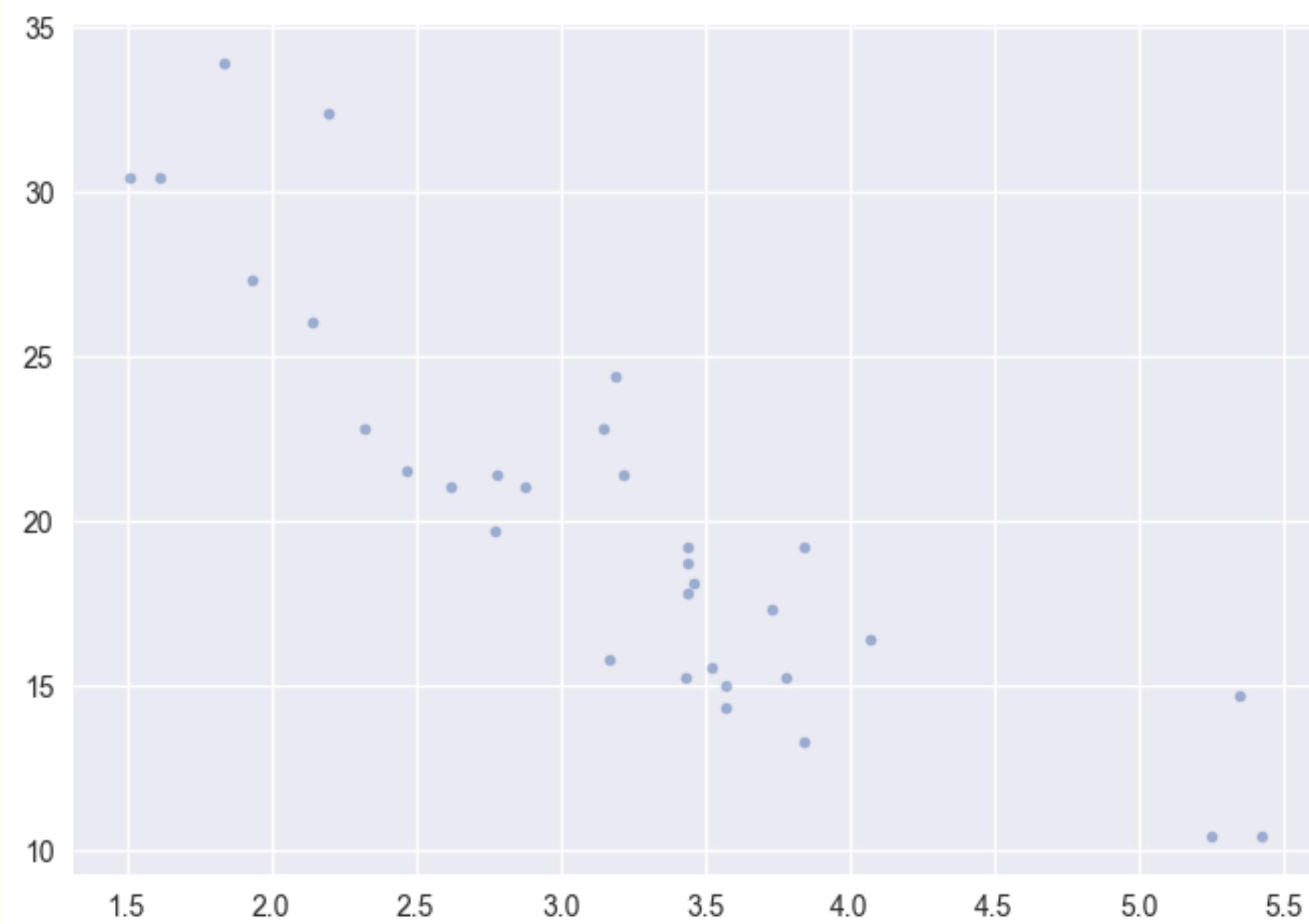
#1 In A Series of Pop Cultural Charts

DANMETH.COM

Small multiples

Use seaborn or
multiple plots in matplotlib

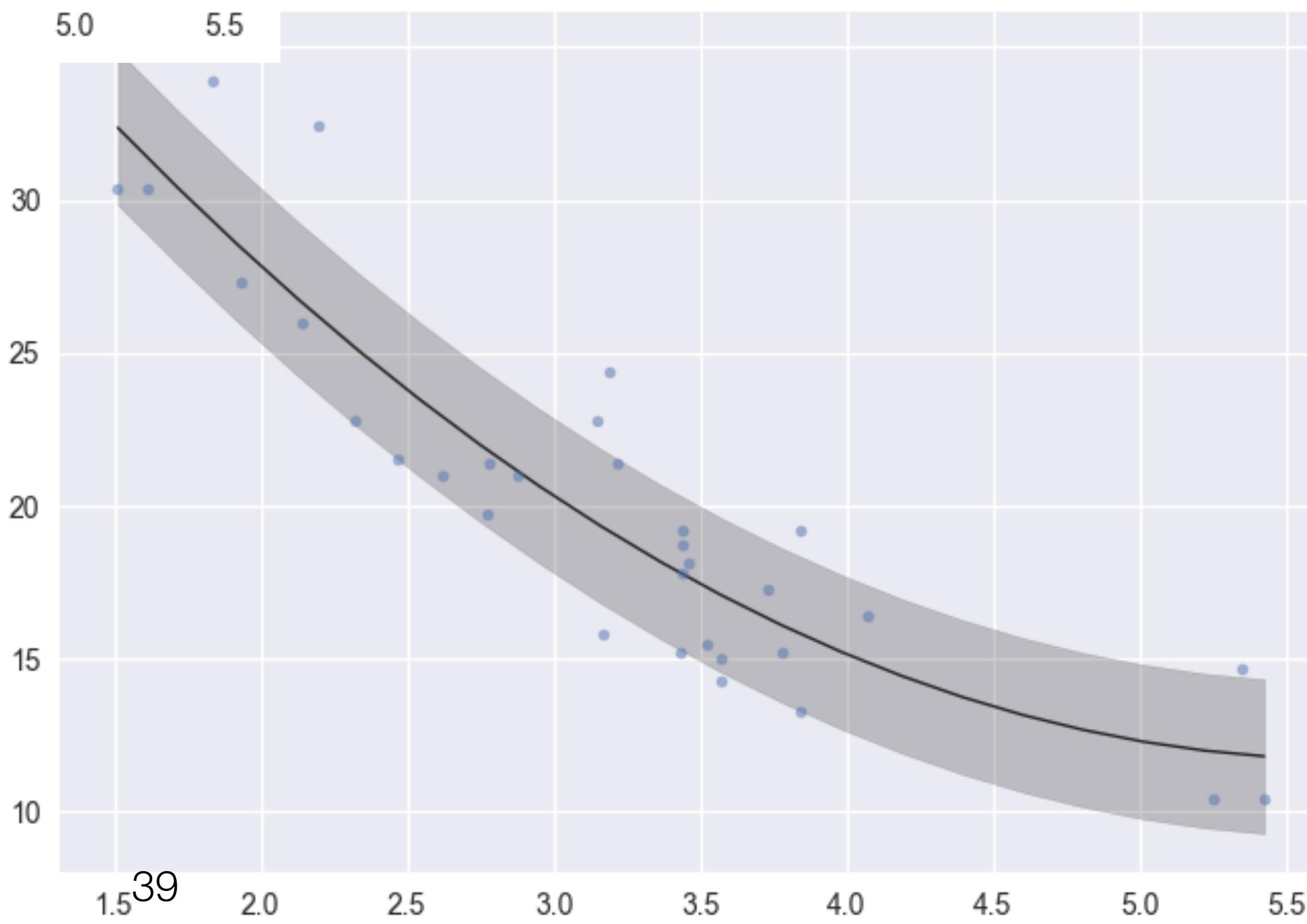




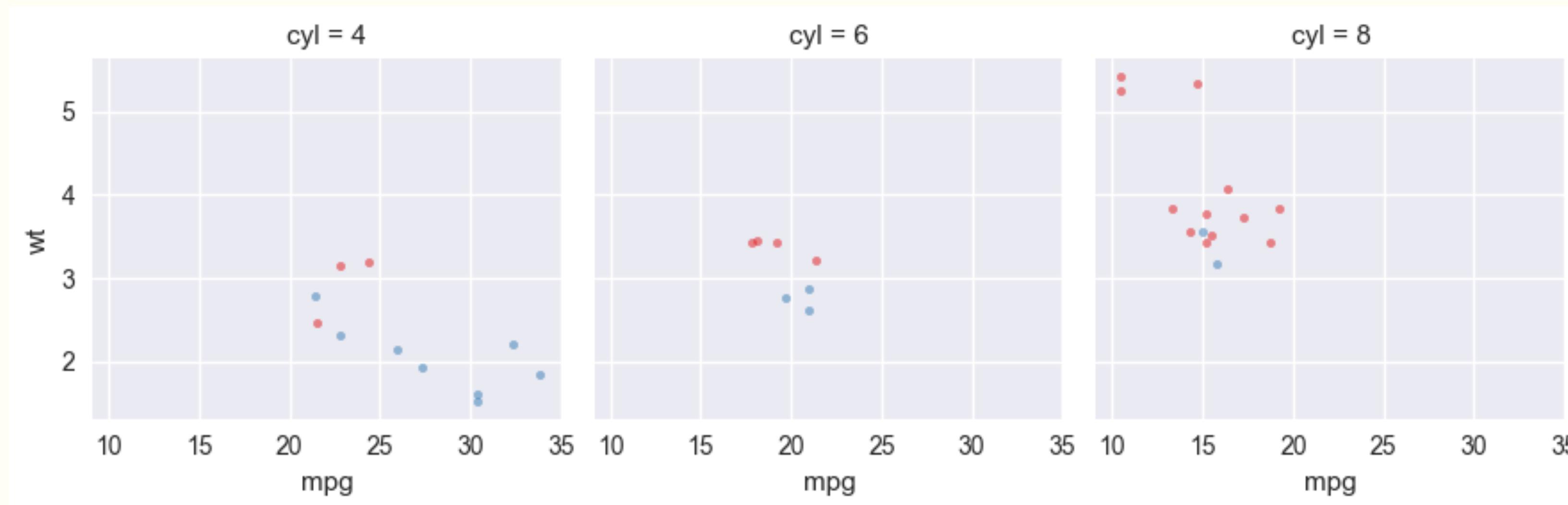
Scatter Plots to discover relationships.

Relationships

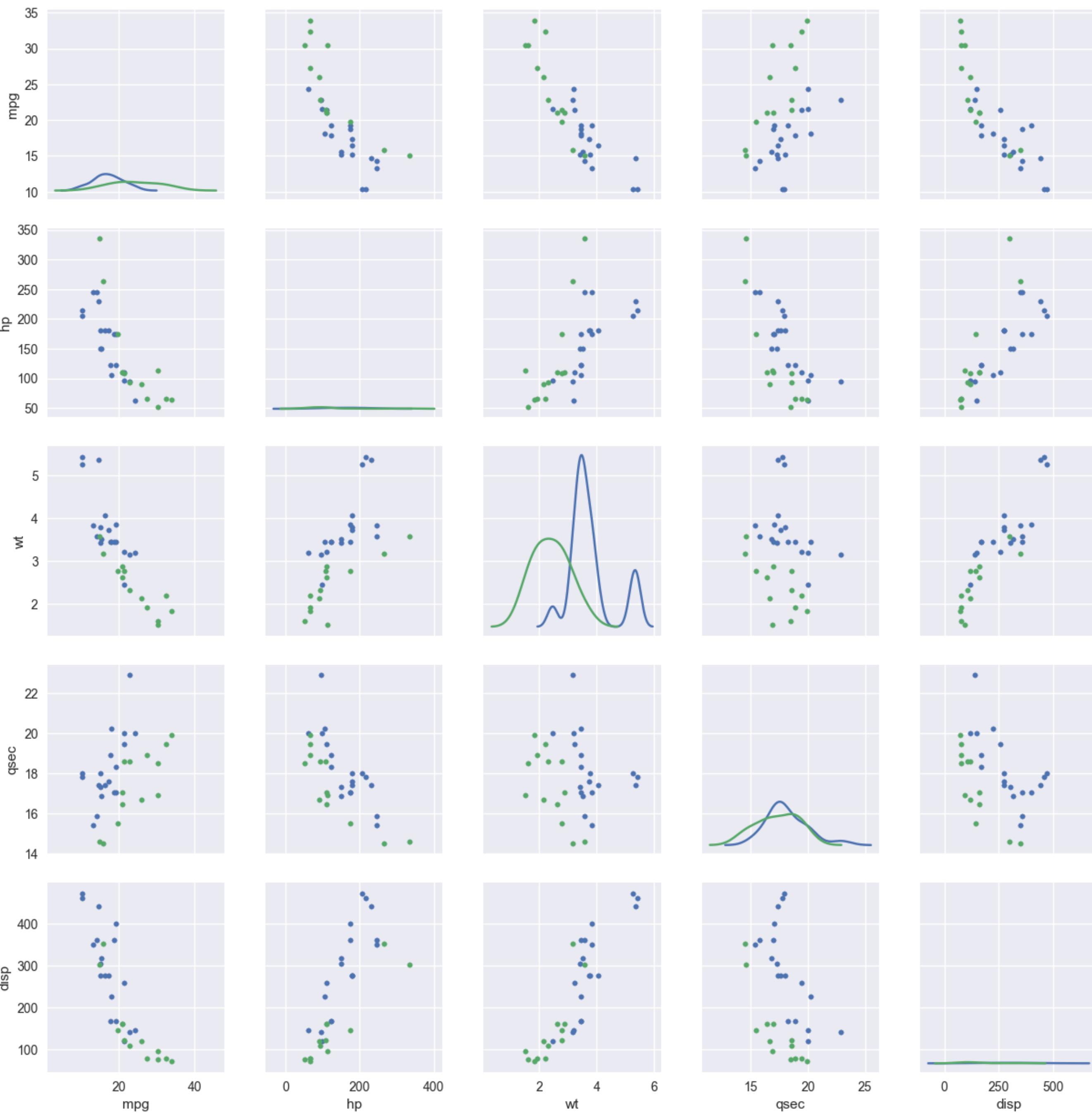
**Use lines when adjoining points involve some continuity
e.g. time**



Relationships with grouping

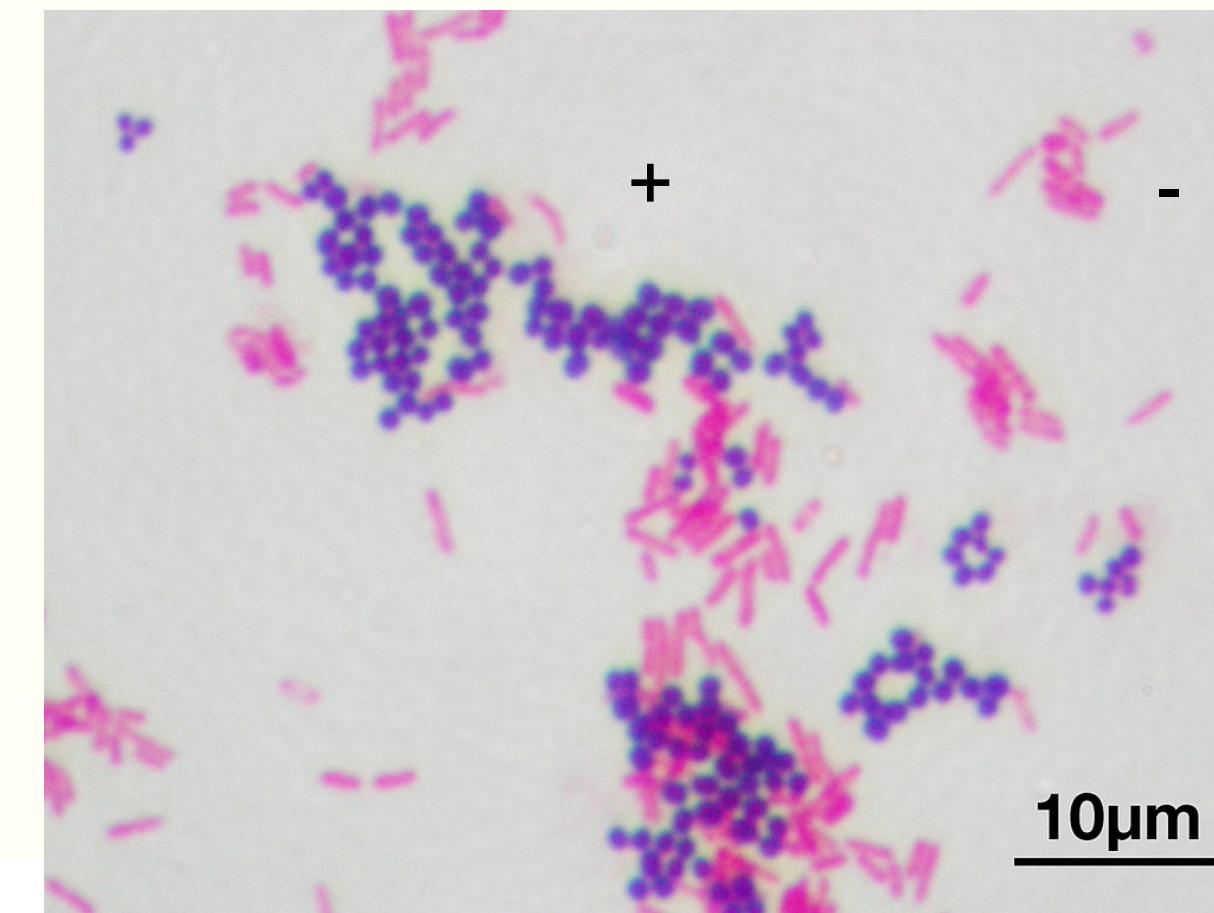


SPLOM



Example: Antibiotics

Will Burtin, 1951



Genus, Species

Table 1. Burtin's data.

Bacteria	Min. Inhibitory Concentration [ml/g]	Antibiotic Penicillin	Streptomycin	Neomycin	Gram Staining
<i>Aerobacter aerogenes</i>	870	1	1.6		negative
<i>Brucella abortus</i>	1	2	0.02		negative
<i>Brucella anthracis</i>	0.001	0.01	0.007		positive
<i>Diplococcus pneumoniae</i>	0.005	11	10		positive
<i>Escherichia coli</i>	100	0.4	0.1		negative
<i>Klebsiella pneumoniae</i>	850	1.2	1		negative
<i>Mycobacterium tuberculosis</i>	800	5	2		negative
<i>Proteus vulgaris</i>	3	0.1	0.1		negative
<i>Pseudomonas aeruginosa</i>	850	2	0.4		negative
<i>Salmonella (Eberthella) typhosa</i>	1	0.4	0.008		negative
<i>Salmonella schottmuelleri</i>	10	0.8	0.09		negative
<i>Staphylococcus albus</i>	0.007	0.1	0.001		positive
<i>Staphylococcus aureus</i>	0.03	0.03	0.001		positive
<i>Streptococcus fecalis</i>	1	1	0.1		positive
<i>Streptococcus hemolyticus</i>	0.001	14	10		positive
<i>Streptococcus viridans</i>	0.005	10	40		positive

How effective are the drugs?

Gram
Positive

Gram
Negative

If bacteria is gram positive,
Penicillin & Neomycin are
most effective

If bacteria is gram negative,
Neomycin is most effective

M. Bostock, Protopis
after W. Burtin, 1951

Gram-negative
Gram-positive