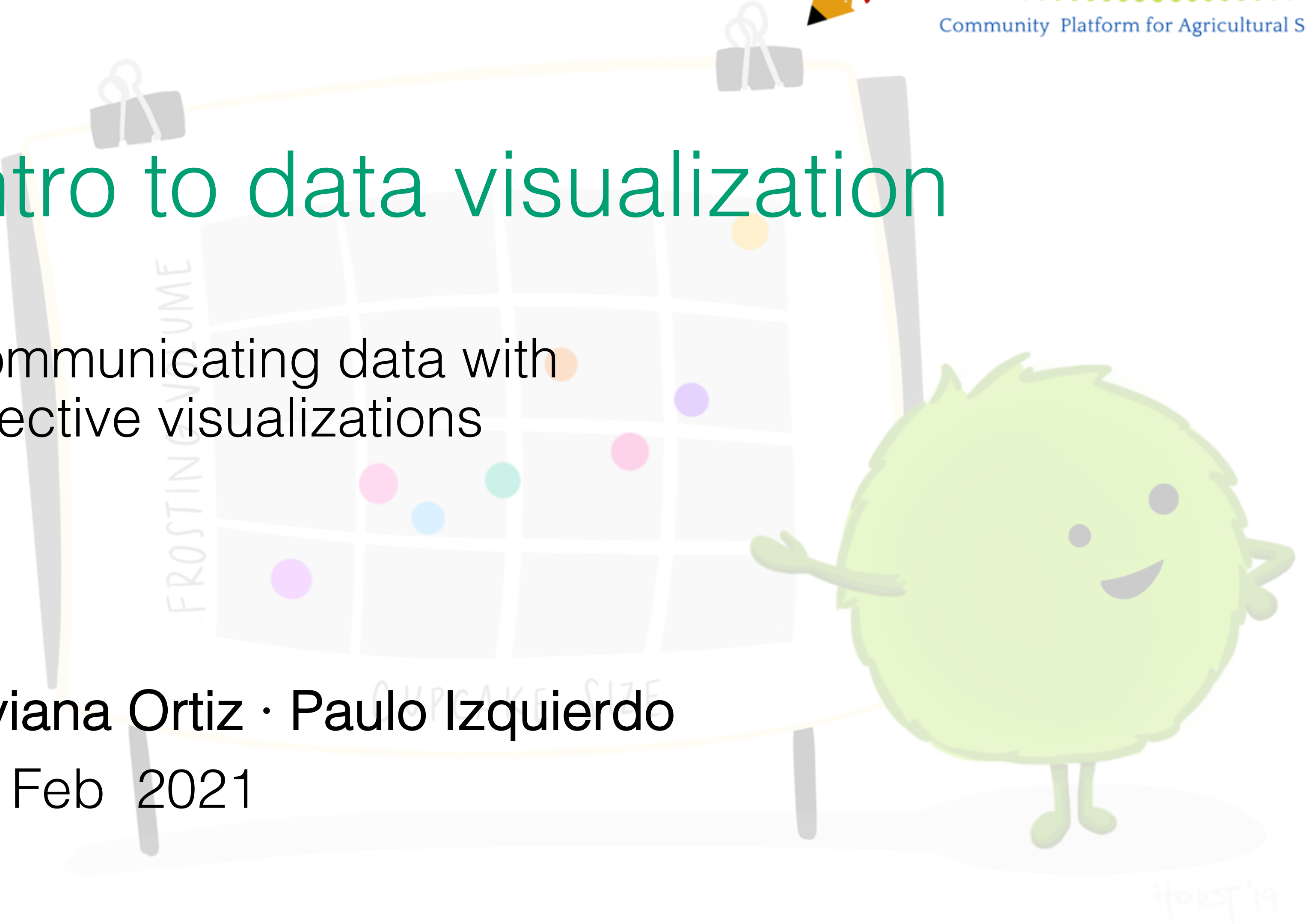
COMPASS

Community Platform for Agricultural Sciences

Intro to data visualization

Communicating data with effective visualizations



Viviana Ortiz · Paulo Izquierdo

20 Feb 2021

Artwork by Allison Horst,
some slides modified from https://www.allisonhorst.com/talk/sccwrp_dataviz_2019/

What is Data Visualization?

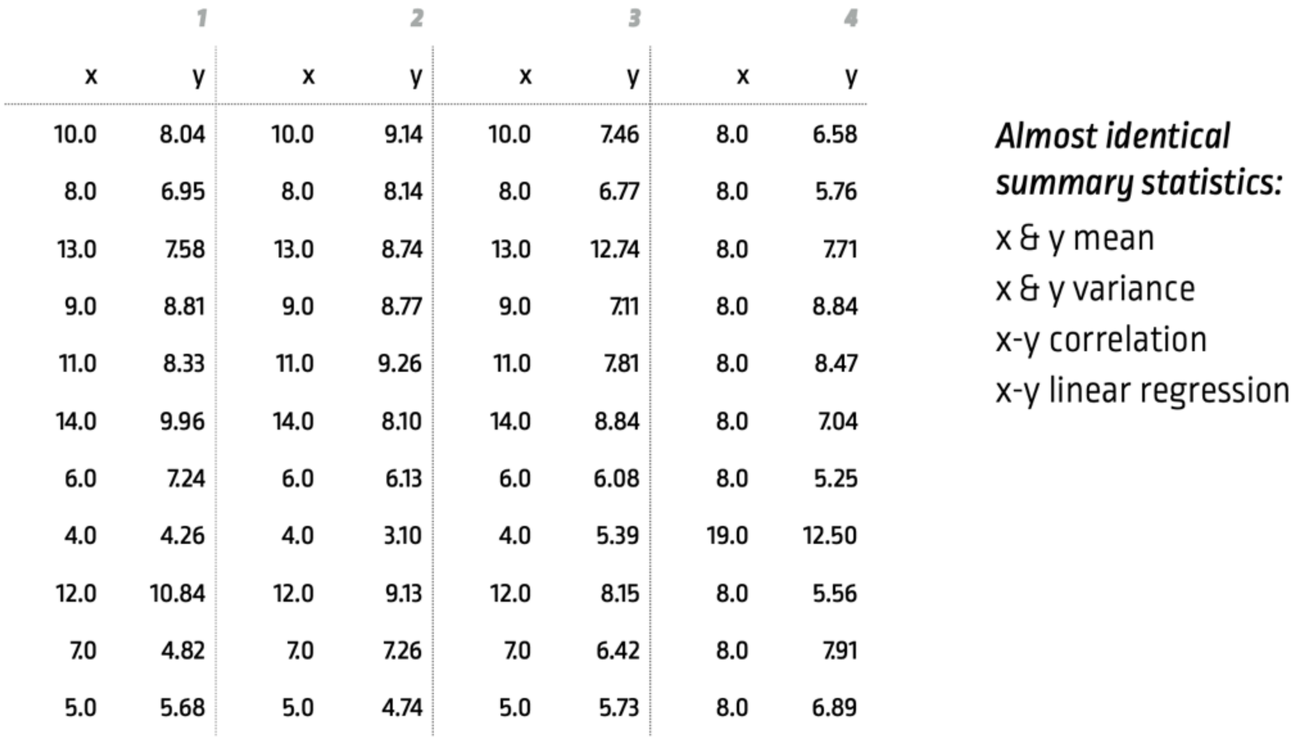
Visual representation of data

charts, graphs, maps, even just tables

http://guides.library.duke.edu/datavis

Why visualization?

Identify patterns



	1		2		3		4
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

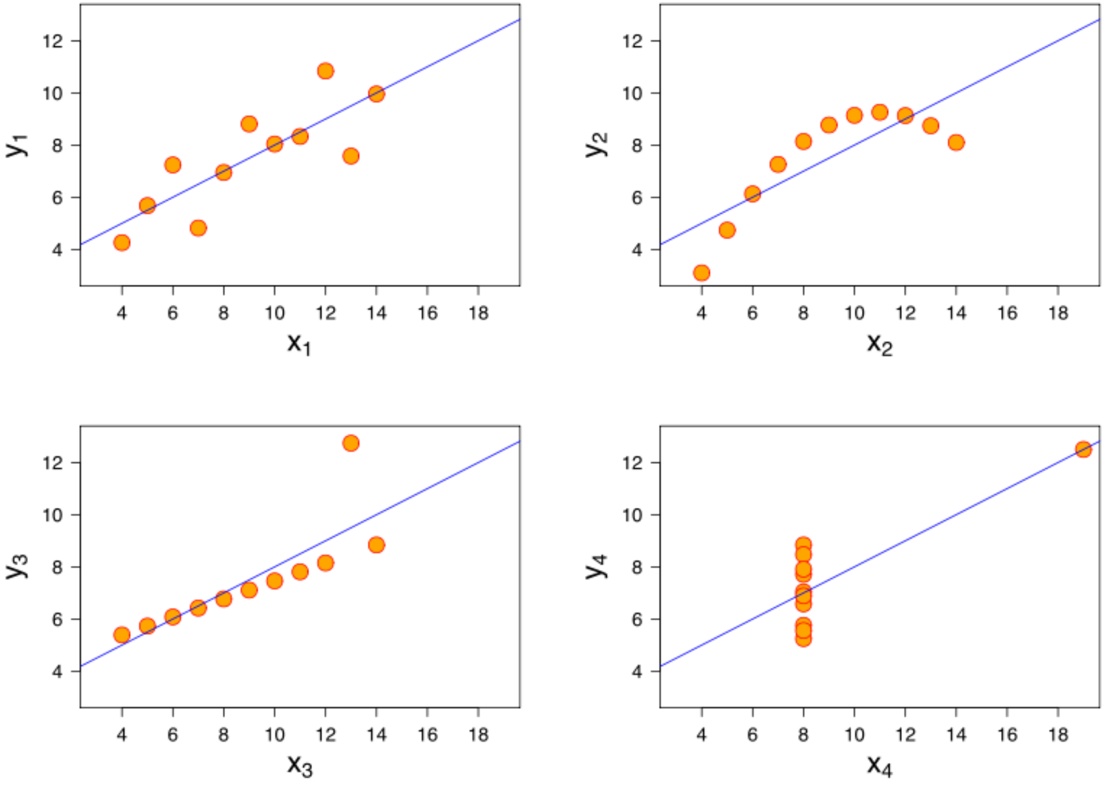
Almost identical summary statistics:

x & y mean
x & y variance
x-y correlation
x-y linear regression

Anscombe's quartet https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Why visualization?

Identify patterns



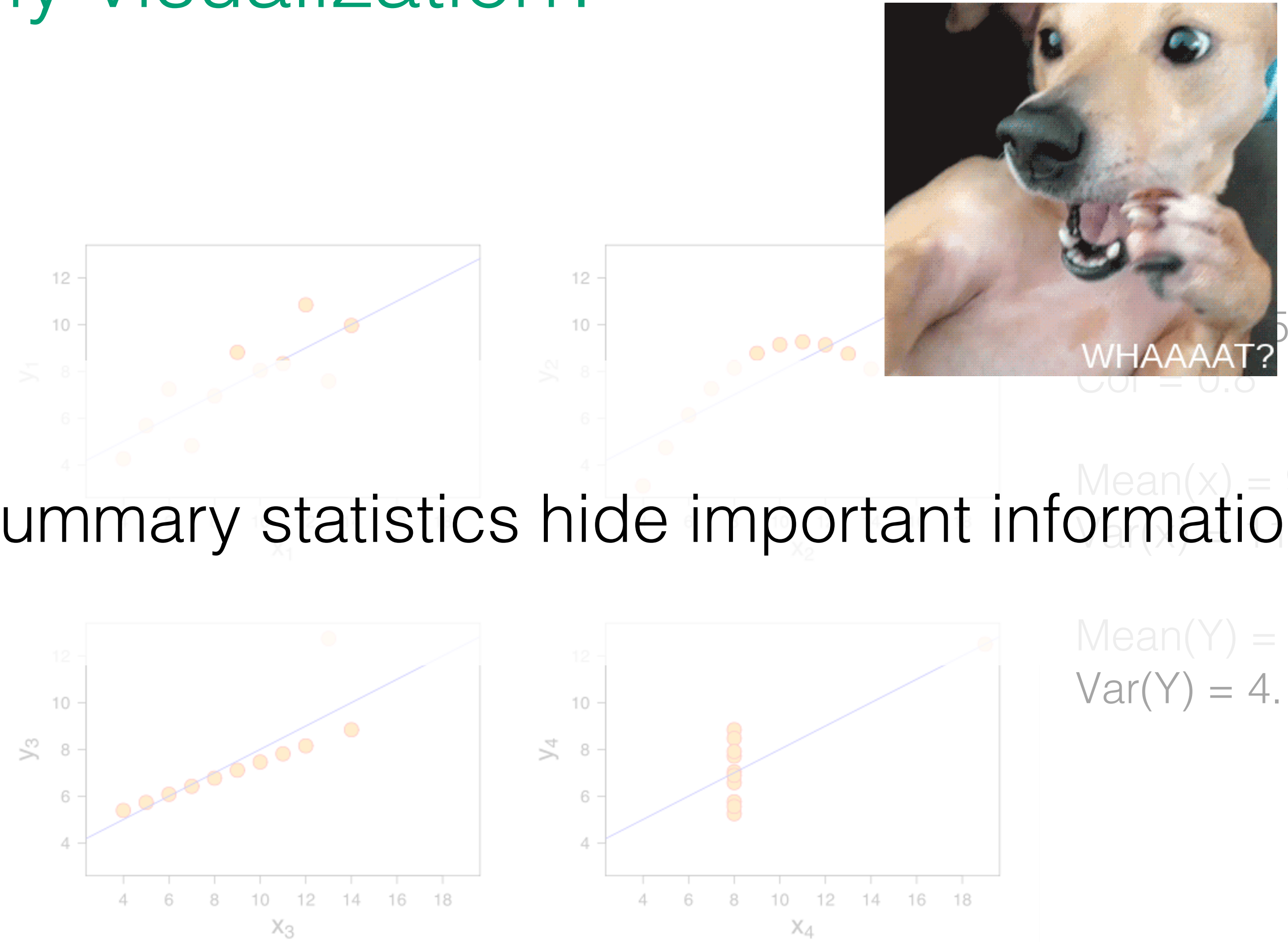
Y = 3 + 0.5x
Cor = 0.8

Mean(x) = 9
Var(x) = 11

Mean(Y) = 7.5
Var(Y) = 4.1

Anscombe's quartet https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Why visualization?



5x

Cor = 0.8

Mean(x) = 9

Mean(Y) = 7.5

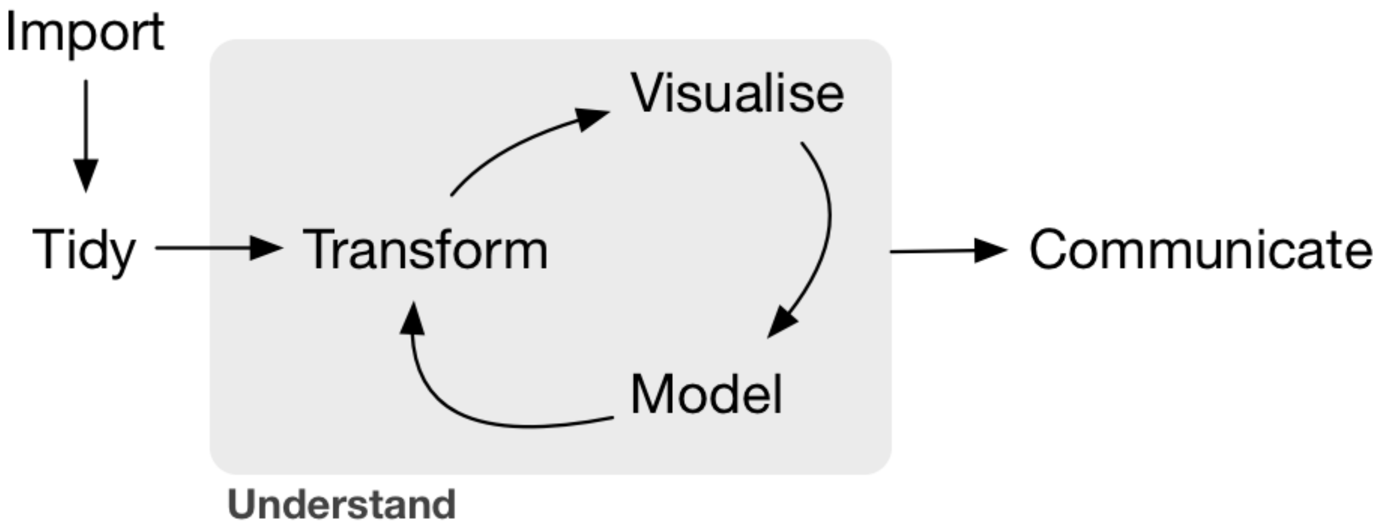
Var(Y) = 4.1

Summary statistics hide important information

Anscombe's quartet https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Why visualization?

The data science pipeline

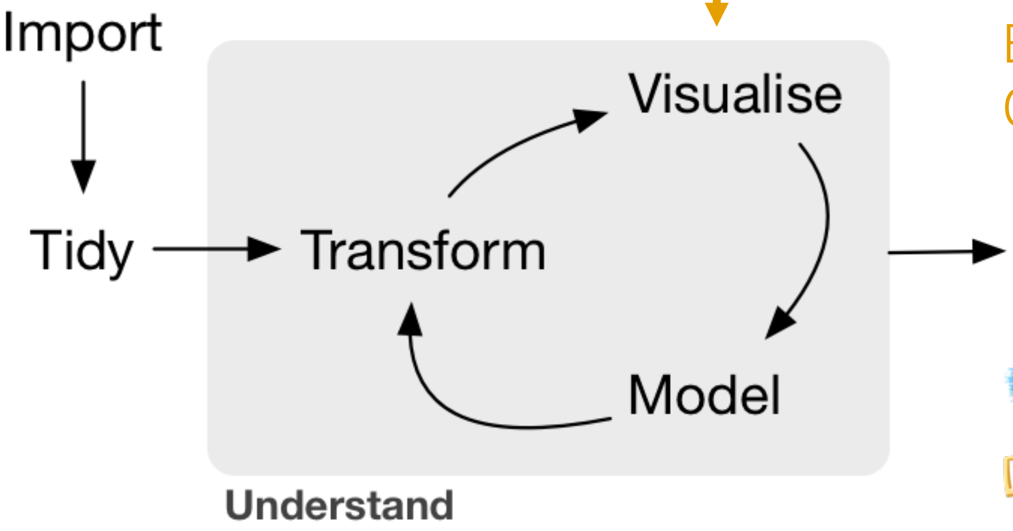


R for Data Science, Whickam et al.

Why visualization?

Visualization main purposes:

Exploration/
Analysis



Import

Tidy

Transform

Visualise

Model

Understand

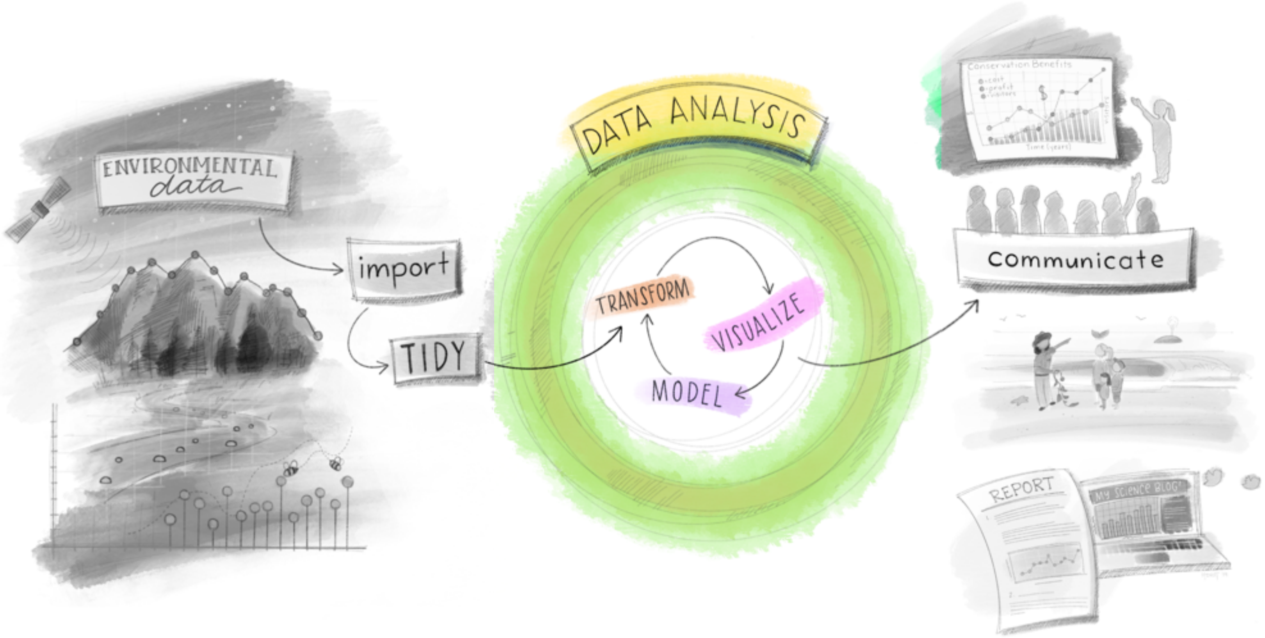
Explanation/
Communication

Communicate

Artwork by Allison Horst
R for Data Science, Whickam et al.

Exploration/Analysis

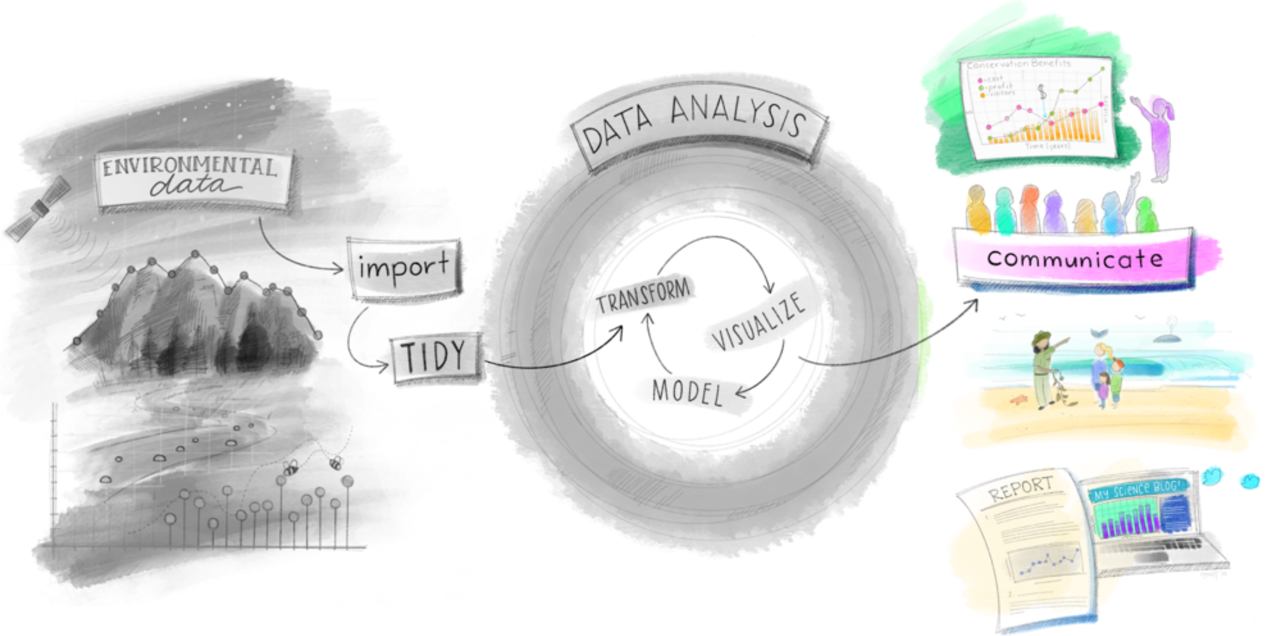
- Raise new questions
- Generate and test hypothesis
- Understand data
- Interpret results



Artwork by Allison Horst

Explanatory/Communication

- Communicate your results to others
- Illustrates important findings
- Tells a story



Artwork by Allison Horst

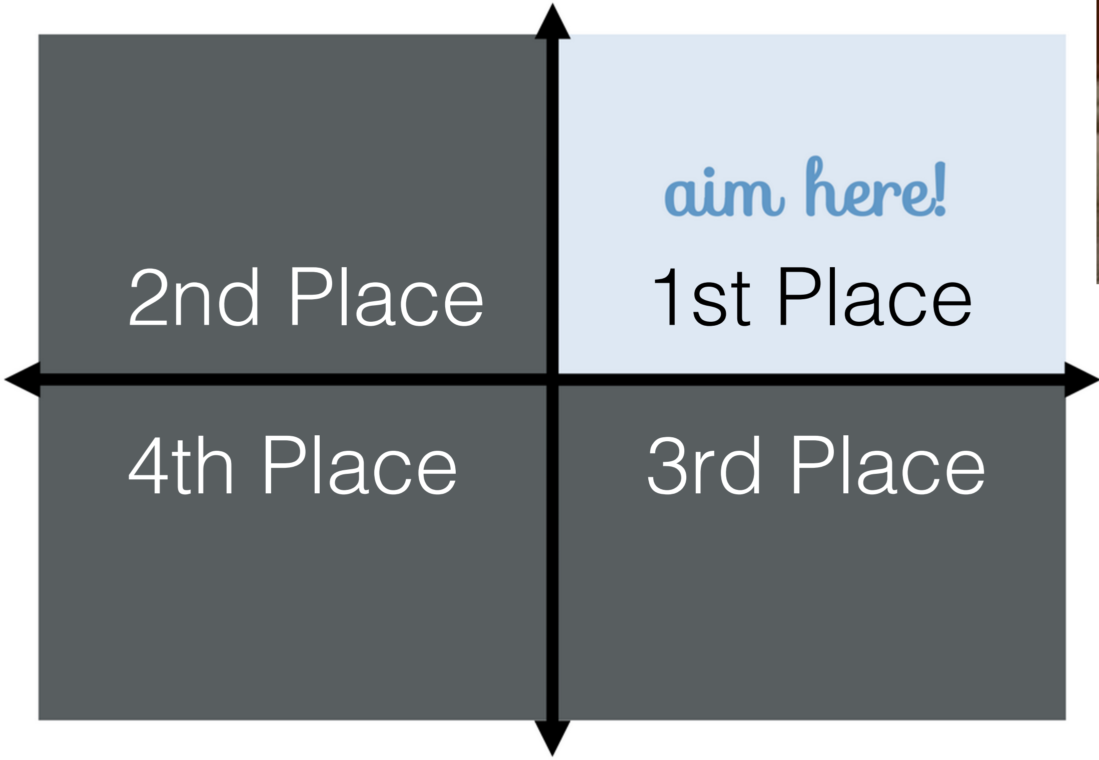
Purpose in the scientific literature…

- Immediately convey information about the study design
- Allow the reader to confirm that the statistical analysis is appropriate for the study design
- Allow the reader to critically evaluate the data

> “Design for the **right audience**, **accurately represent the data**, and keep it **clear**.”

Yan Holtz, dataviz
https://www.data-to-viz.com

DATA VIZ HIERARCHY



USEFUL

2nd Place

aim here!

1st Place

UGLY

PRETTY

4th Place

3rd Place

CRAP

Source: Jackie Wirz

via Allison Horst

What you will learn today

1. Responsible data visualization
2. Clear data viz for your audience
3. All about aesthetics

1
Responsible data visualization

Have a practical sense for why some graphs and figures work well, while others may fail to inform or actively mislead.

A. What is an appropriate graph for this data?

B. Are the data visualized responsibly?
   a. Axes issues
   b. Are you hiding the data?
   c. Have you included uncertainty?
   d. Trendline overuse & responsibilities

A. What is an appropriate graph for this type of data?

Great resources for choosing a graph type:

- From Data to Viz by Yan Holtz: “find the graphic you need”
  https://www.data-to-viz.com/

- Clause Wilke’s “Fundamentals of Data Visualization” - Ch. 5
  https://serialmentor.com/dataviz/

- The R Graph Gallery by Yan Holtz - great inspiration for graph types
  https://www.r-graph-gallery.com/

- The Data Visualization Catalogue

Choosing the appropriate graph(s) for the data

- Discrete & continuous quantities
- Proportions/percentages
- Nominal data (categories)

Visit Data to Viz for many more options & combinations!
https://www.data-to-viz.com

Discrete & continuous data

Numeric data

-Continuous data: values that can be measured, and can have any of an infinite range of values within a possible range (e.g. temperature, salinity)

-Discrete data: values, often counted, that can only exist at finite values (e.g. number of plants per row, number of leaves in a plant)

Discrete & continuous data

Numeric data

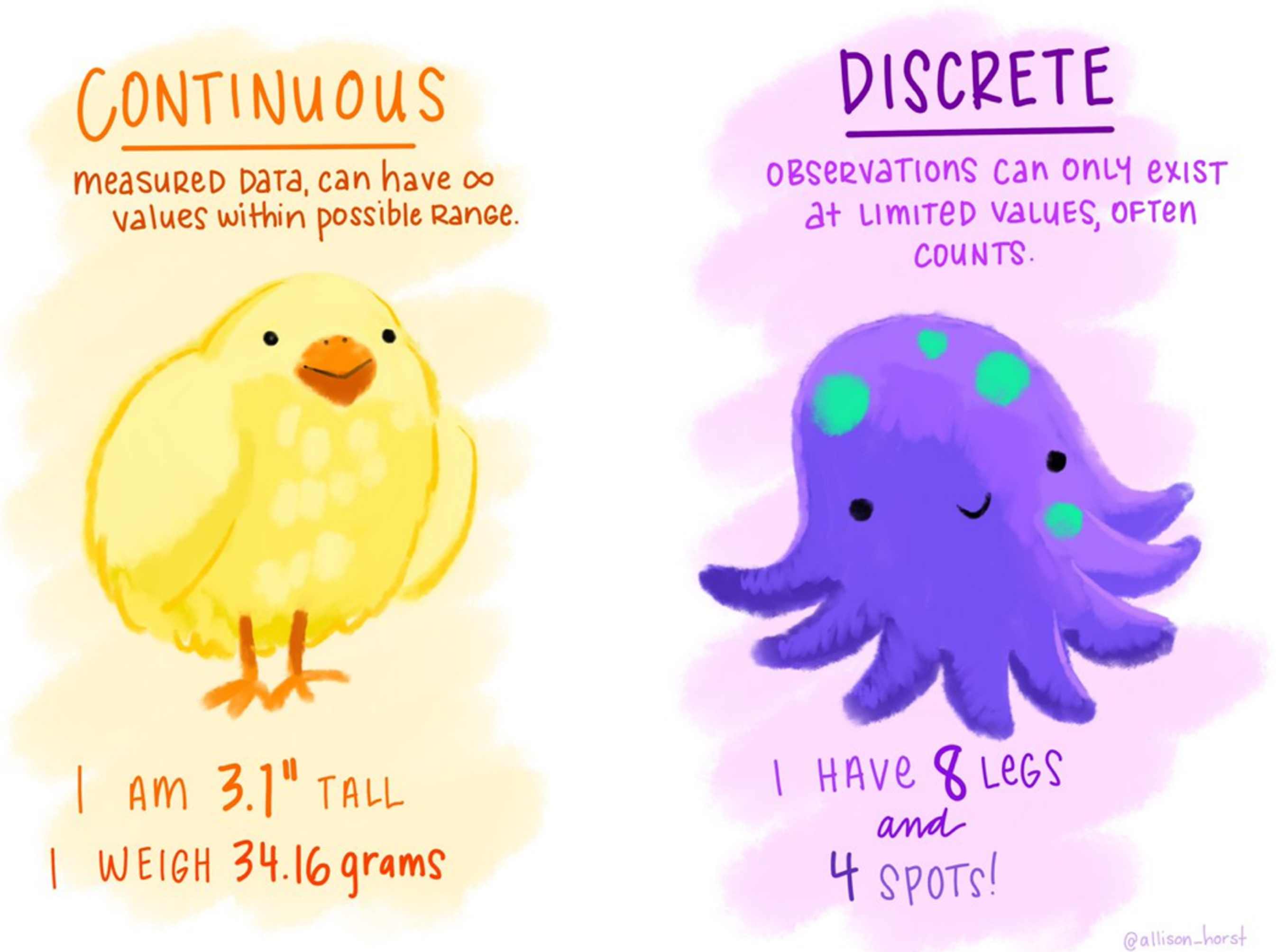
-Continuous data: values that can be measured, and can have any of an infinite range of values within a possible range (e.g. temperature, salinity)

-Discrete data: values, often counted, that can only exist at finite values (e.g. number of plants per row, number of leaves in a plant)

Categorical data: qualitative descriptions (nominal, ordinal, binary), data can take on only a specific set of values representing a set of possible categories

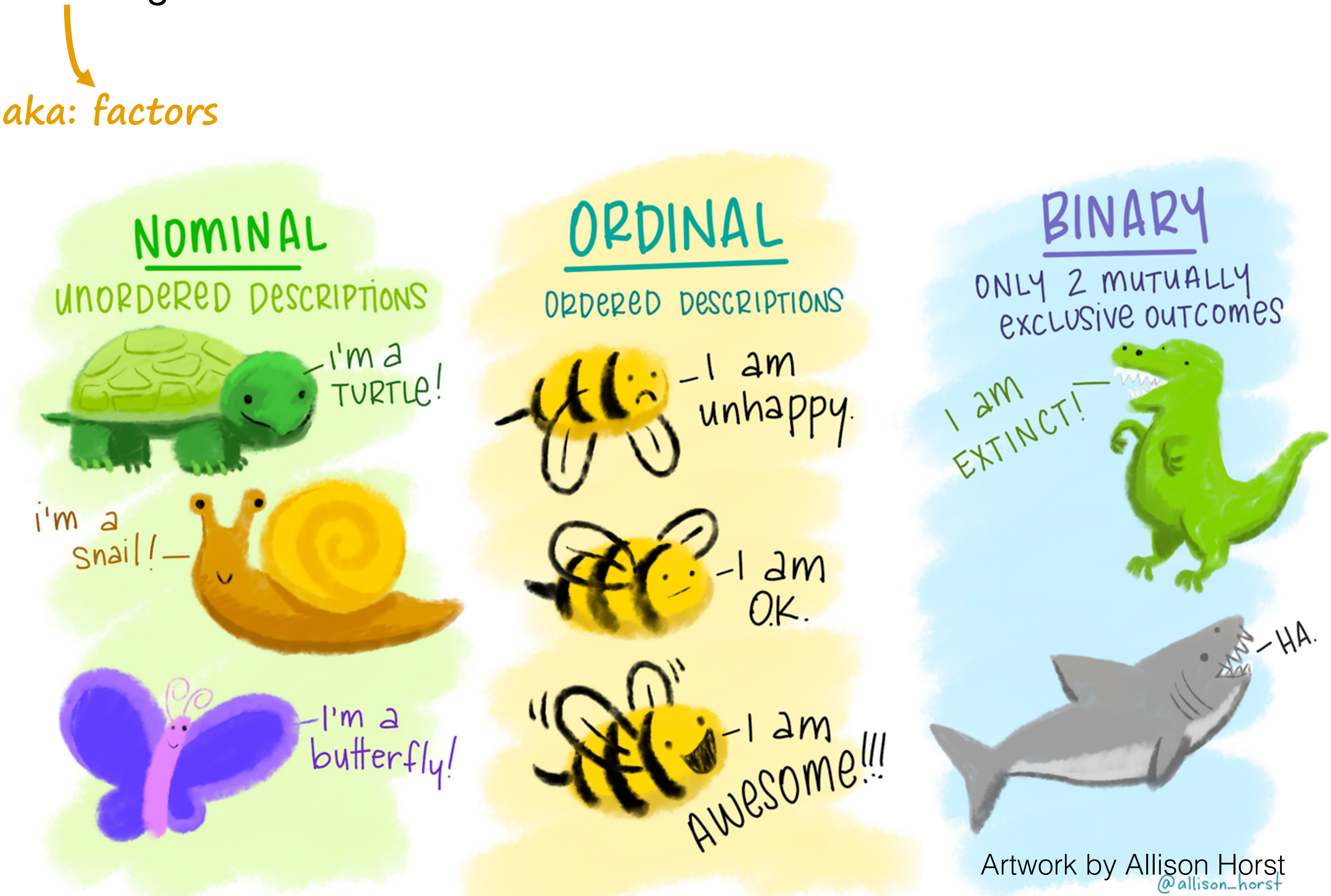
Note: sometimes low resolution continuous observations (e.g. “plant height was recorded to the nearest 0.5 cm”) can look like discrete data because values only exist at intervals.

Numeric data


CONTINUOUS
measured data, can have ∞ values within possible range.
I am 3.1" tall
I weigh 34.16 grams
DISCRETE
observations can only exist at limited values, often counts.
I have 8 legs and 4 spots!
@allison_horst

Artwork by Allison Horst

Categorical data

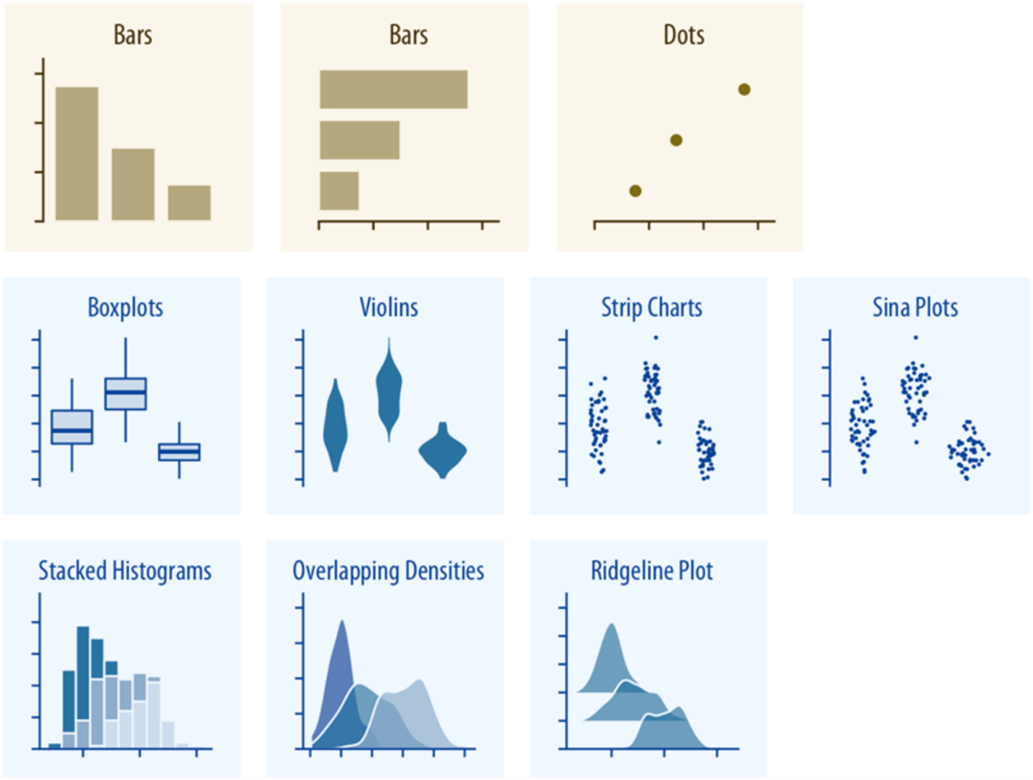


Artwork by Allison Horst

Usually represented by counts or proportions within groups

Visualizing continuous variables

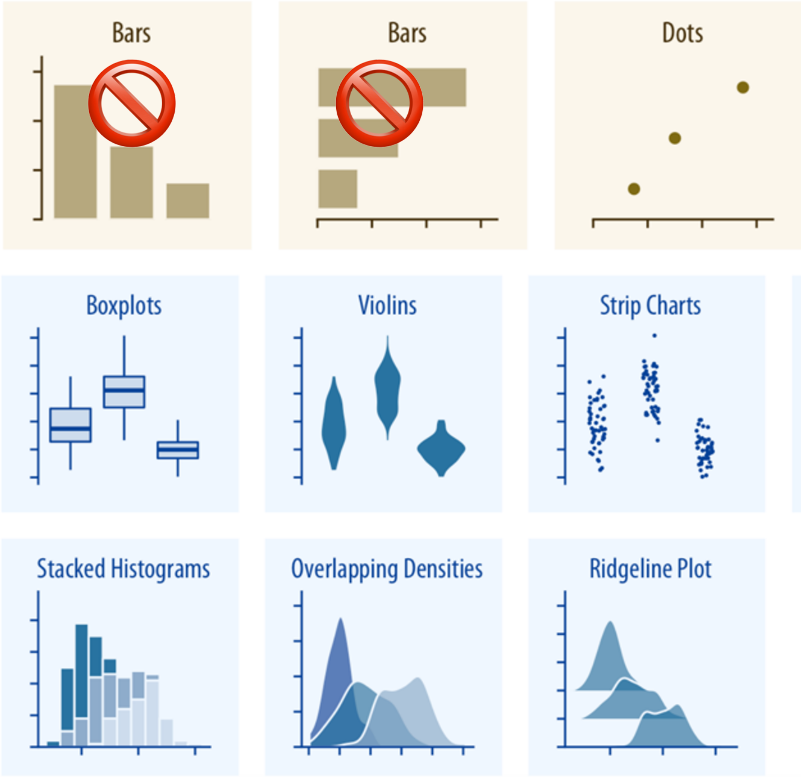
Bars, points, densities



Wilke, C.O. Ch. 5 Directory of Visualizations,
Fundamentals of Data Visualization

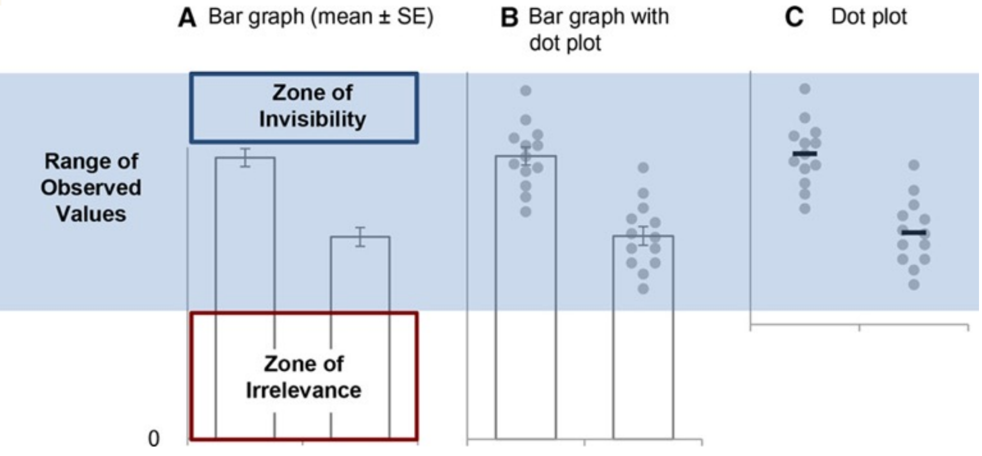
Visualizing continuous variables

~~Bars~~, points, densities

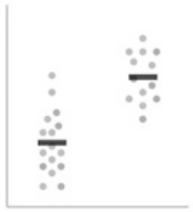
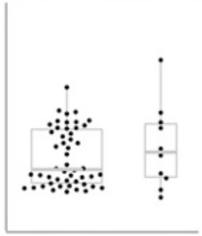
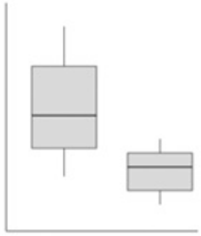
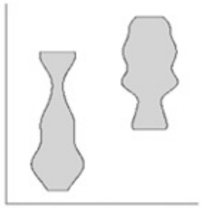
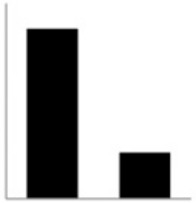


Wilke, C.O. Ch. 5 Directory of Visualizations, Fundamentals of Data Visualization

Bars not ideal for continuous data



Transforming Data Visualization to Improve Transparency, Weissgerber et al., 2019

Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	Sample size is too small to determine data distribution OR Any data distribution	• Make all data points visible - use symmetric jittering • Many groups: Increase white space between groups, emphasize summary statistics & de-emphasize points • Only add error bars if the sample size is large enough to avoid creating a false sense of certainty • Avoid "histograms with dots"
Dot plot with box plot or violin plot		Continuous	Combination of dot plot & box plot or violin plot (see descriptions above and below)	Medium	Any	• Make all data points visible (symmetric jittering) • Smaller n: Emphasize data points and de-emphasize box plot, delete box plot and show only median line for groups with very small n • Larger n: Emphasize box plot and de-emphasize points
Box plot		Continuous	Horizontal lines on box: 75th, 50th (median) and 25th percentile Whiskers: varies; often most extreme data points that are not outliers Dots above or below whiskers: outliers	Large	Do not use for bimodal data	• List sample size below group name on x-axis • Specify what whiskers represent in legend
Violin plot		Continuous	Gives an estimated outline of the data distribution. The precision of the outline increases with increasing sample size.	Large	Any	• List sample size below group name on x-axis • The violin plot should not include biologically impossible values
Bar graph		Counts or proportions	Bar height shows the value of the count or proportion	Any	Any	• **Do not use for continuous data**

Transforming Data Visualization to Improve Transparency, Weissgerber et al., 2019

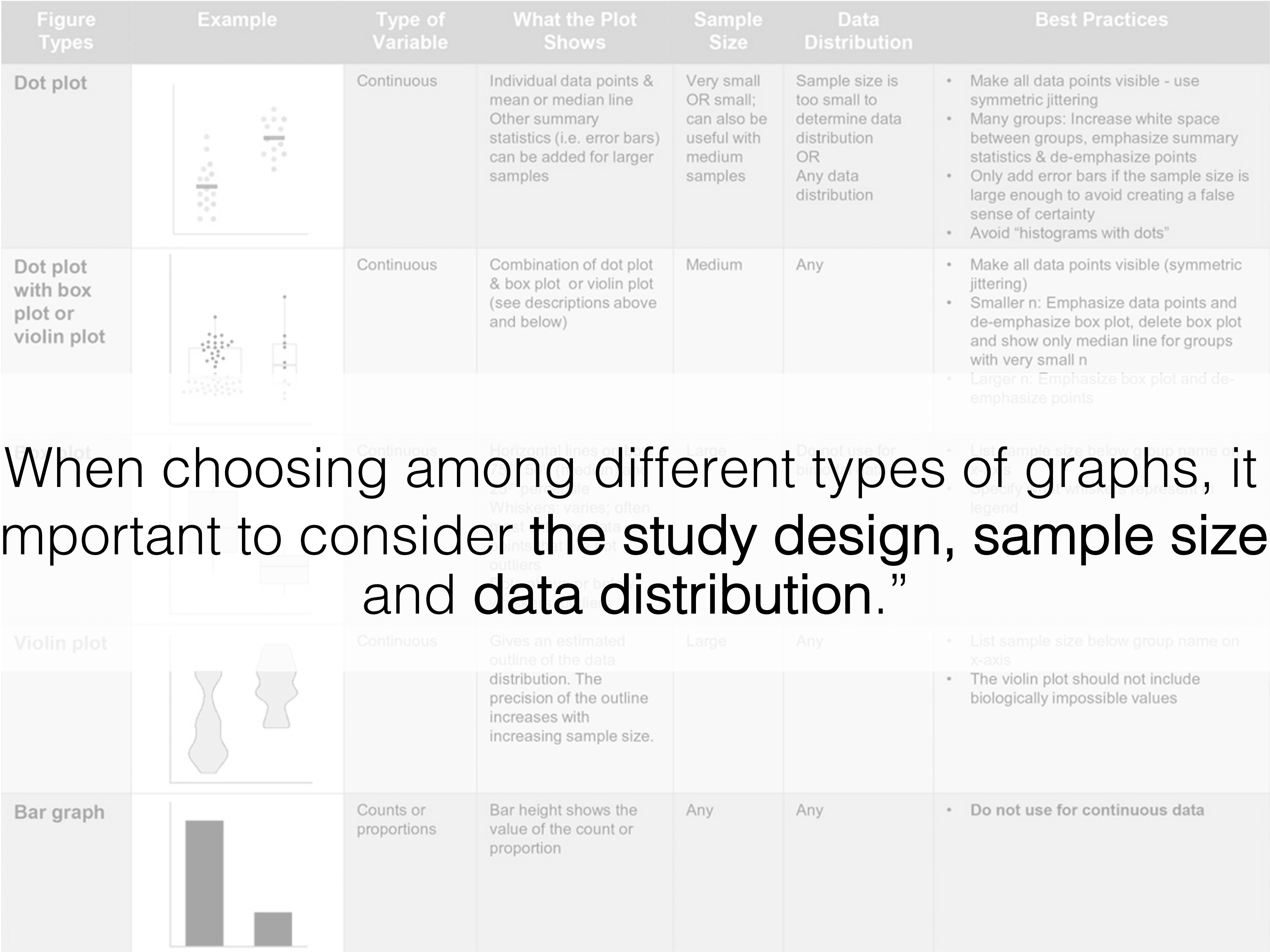


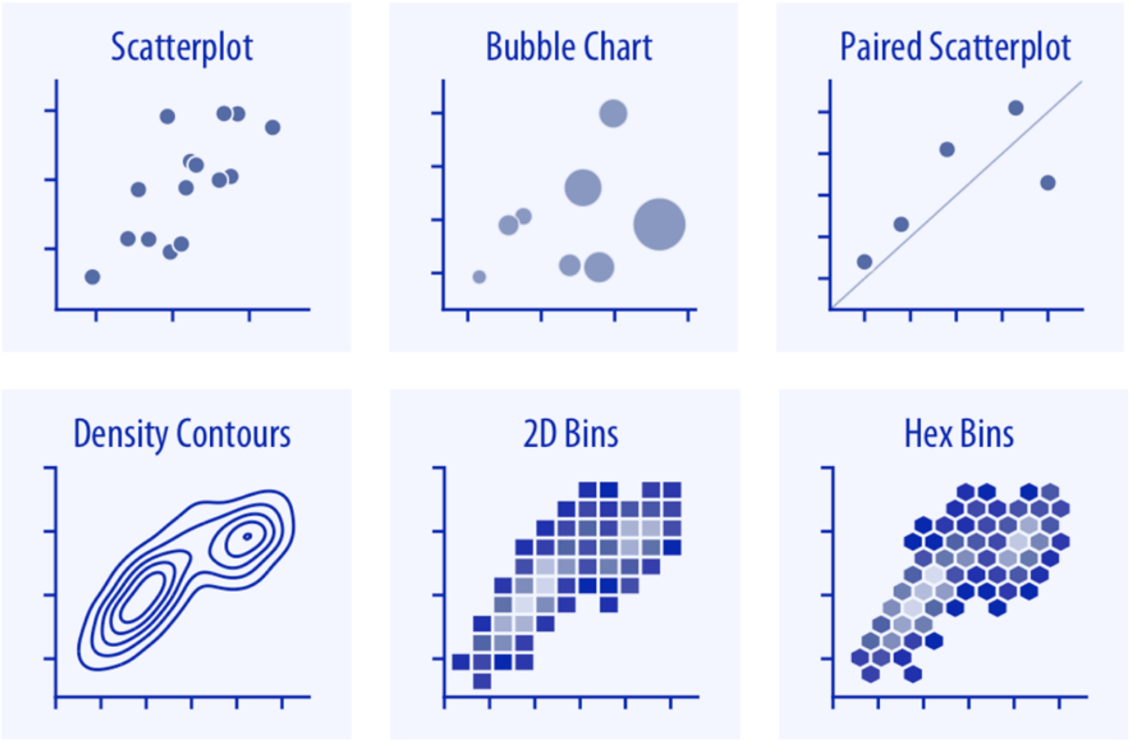
Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	Sample size is too small to determine data distribution OR Any data distribution	• Make all data points visible - use symmetric jittering • Many groups: Increase white space between groups, emphasize summary statistics & de-emphasize points • Only add error bars if the sample size is large enough to avoid creating a false sense of certainty • Avoid "histograms with dots"
Dot plot with box plot or violin plot		Continuous	Combination of dot plot & box plot or violin plot (see descriptions above and below)	Medium	Any	• Make all data points visible (symmetric jittering) • Smaller n: Emphasize data points and de-emphasize box plot, delete box plot and show only median line for groups with very small n • Larger n: Emphasize box plot and de-emphasize points
Box plot		Continuous	Horizontal lines on box: 75th, 50th (median) and 25th percentile Whiskers: varies; often most [illegible] data points [illegible] plot outliers [illegible] or below [illegible] outliers	Large	Do not use for bimodal data	• List sample size below group name on x-axis • Specify [illegible] whiskers represent in legend
Violin plot		Continuous	Gives an estimated outline of the data distribution. The precision of the outline increases with increasing sample size.	Large	Any	• List sample size below group name on x-axis • The violin plot should not include biologically impossible values
Bar graph		Counts or proportions	Bar height shows the value of the count or proportion	Any	Any	• **Do not use for continuous data**

“When choosing among different types of graphs, it is important to consider **the study design, sample size,** and **data distribution.**”

Transforming Data Visualization to Improve Transparency, Weissgerber et al., 2019

Visualizing continuous variables

2 continuous variables



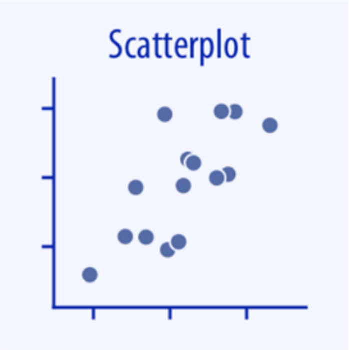
Wilke, C.O. Ch. 5 Directory of Visualizations,
Fundamentals of Data Visualization

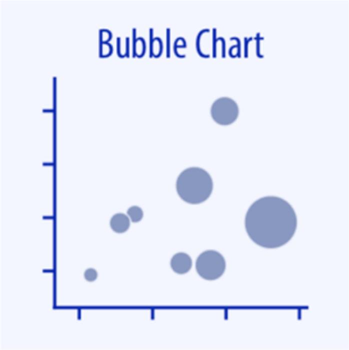
Visualizing continuous variables

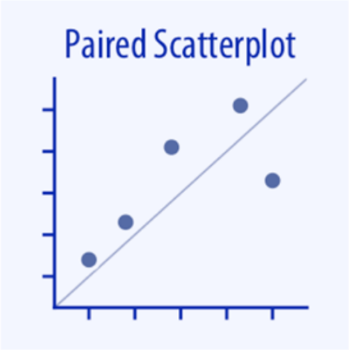
2 continuous variables

three variables,
map one onto
the dot size

one variable
relative to
another







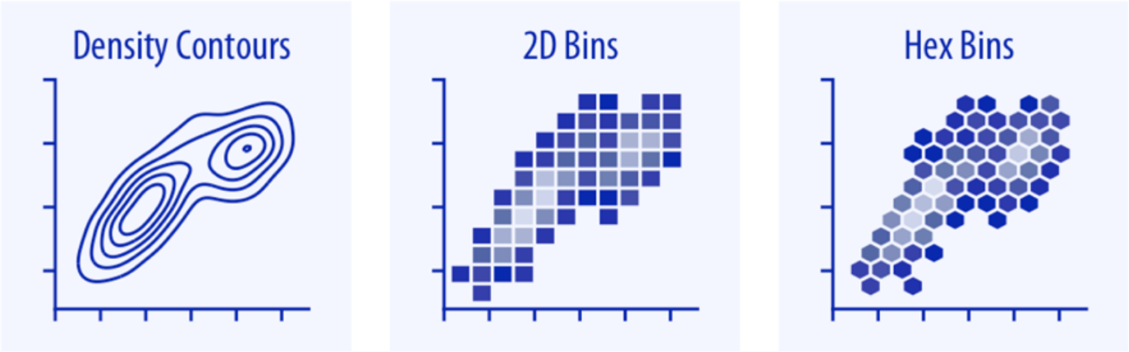
paired data,
where the
variables along
the x and
the y axes are
measured in
the same units

Wilke, C.O. Ch. 5 Directory of Visualizations,
Fundamentals of Data Visualization

Visualizing continuous variables

2 continuous variables

For large numbers of points, regular scatterplots can become uninformative due to overplotting

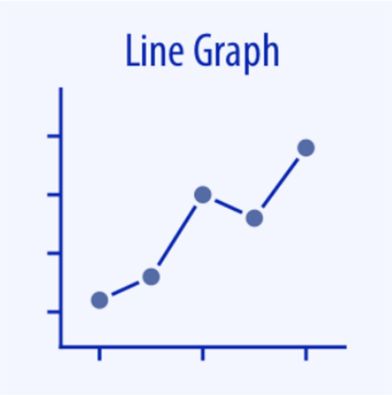


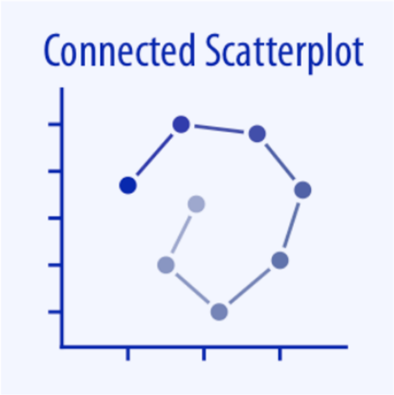
Wilke, C.O. Ch. 5 Directory of Visualizations, Fundamentals of Data Visualization

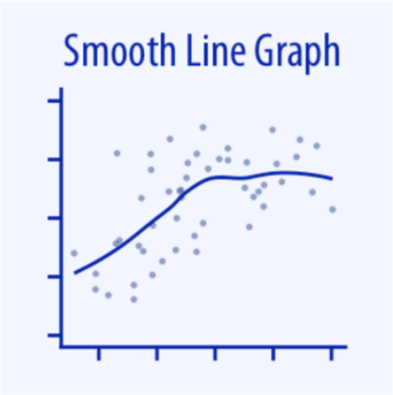
Visualizing a measured variable over time

“When **the x axis represents time** or **a strictly increasing quantity** such as a treatment dose, we commonly draw line graphs.”

- Clause O. Wilke, Fundamentals of Data Visualization





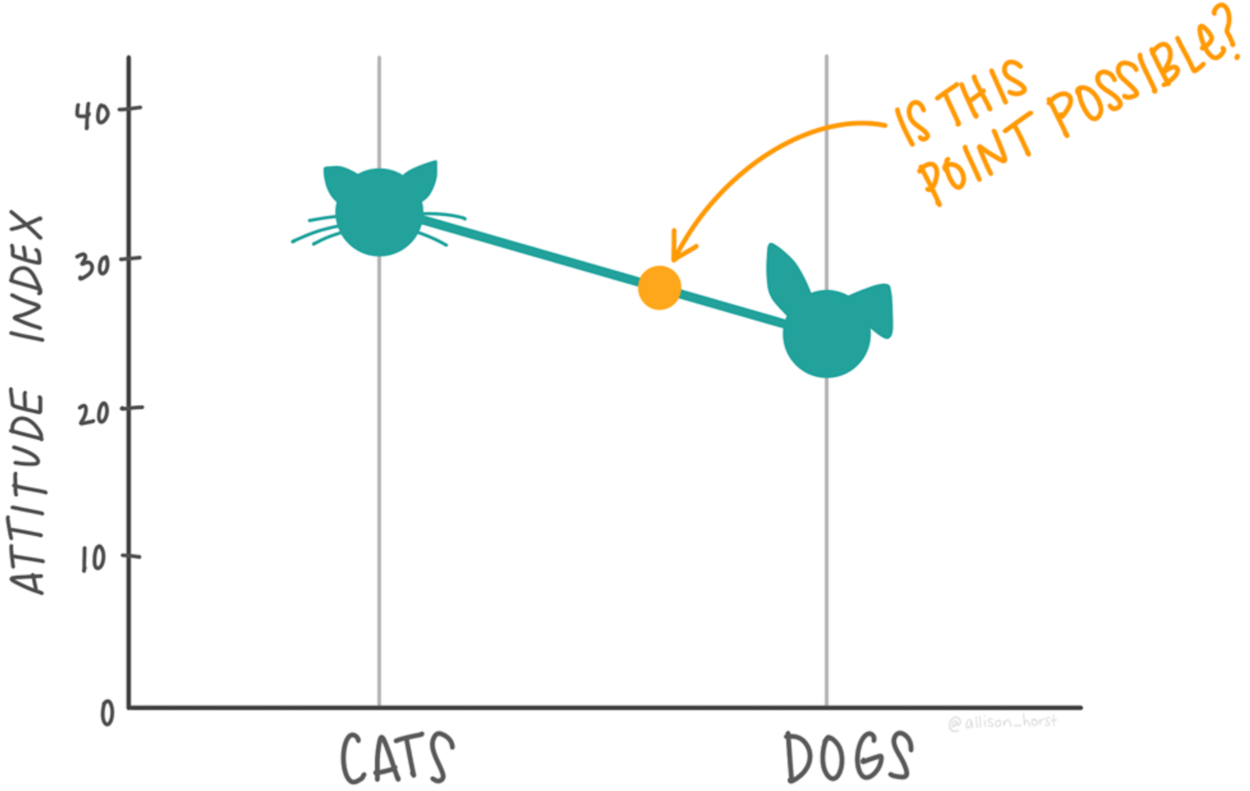


Common pitfall:

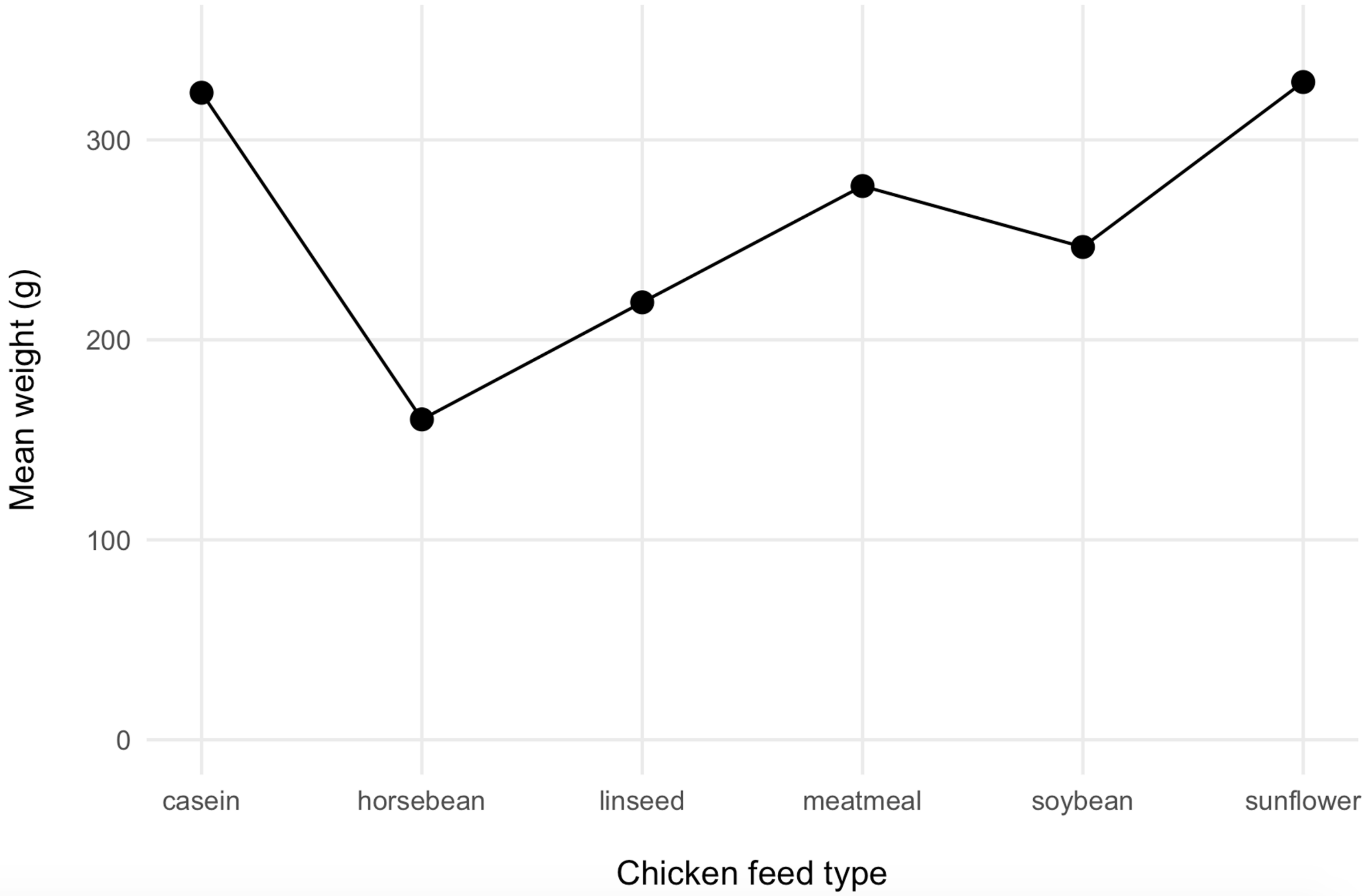
Adding a continuous element to a discrete scale (false trends)

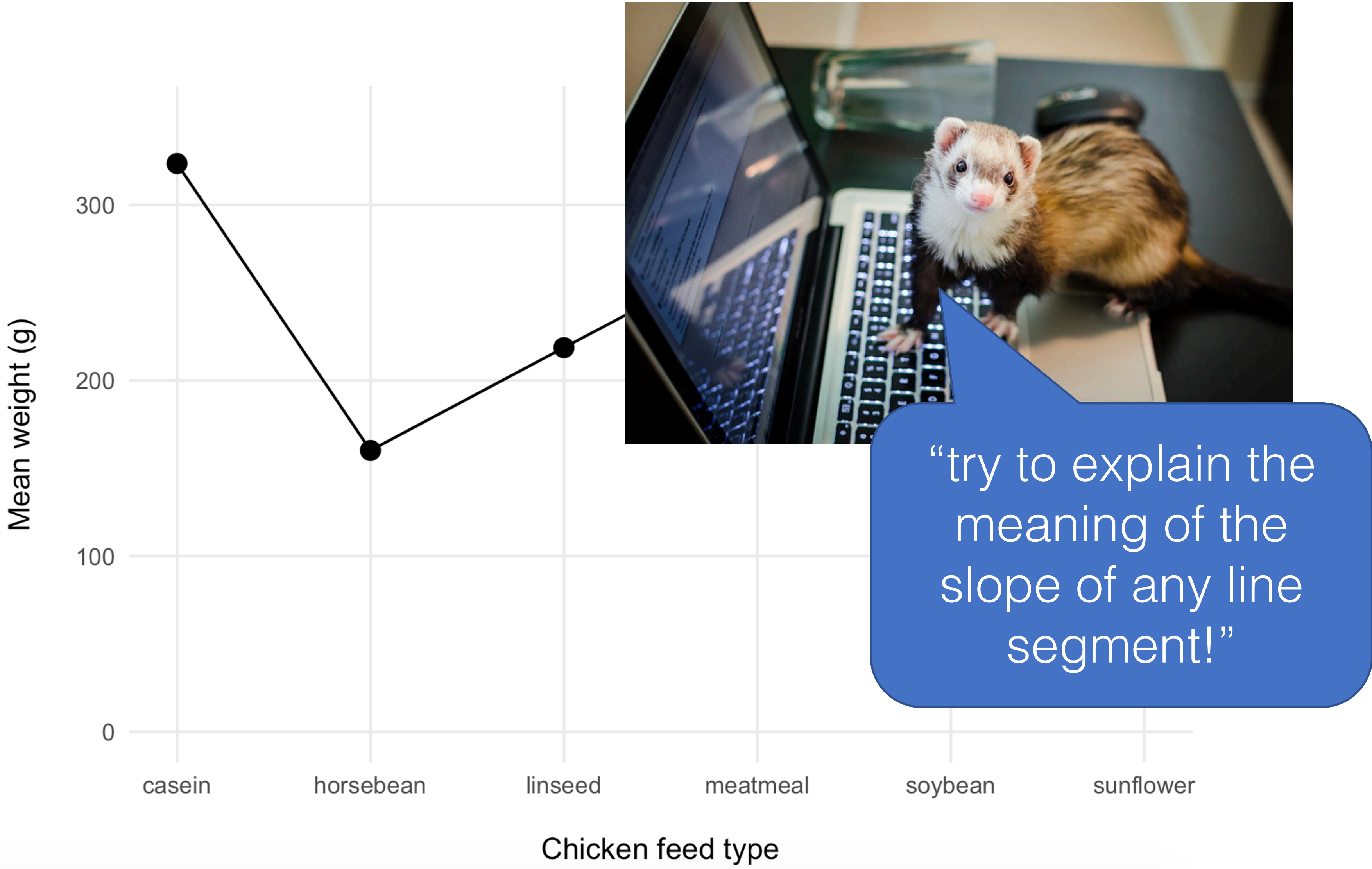
Why is this a problem? Connecting lines imply that there are possibilities that exist between nodes. That is often not the case. Avoid false trends.

The big idea:



Artwork by Allison Horst


Mean weight (g)
300
200
100
0
casein
horsebean
linseed
meatmeal
soybean
sunflower
Chicken feed type



“try to explain the meaning of the slope of any line segment!”

If you can’t do it clearly, the audience doesn’t even have a chance - and often, it will cause confusion or misinterpretation

Levels of data precision

Continuous measured

↓

Discrete / ordinal

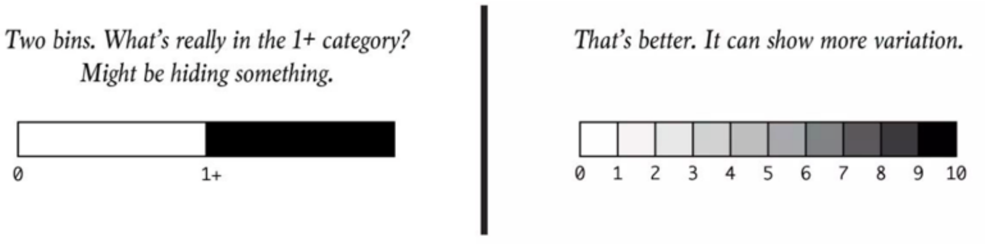
↓

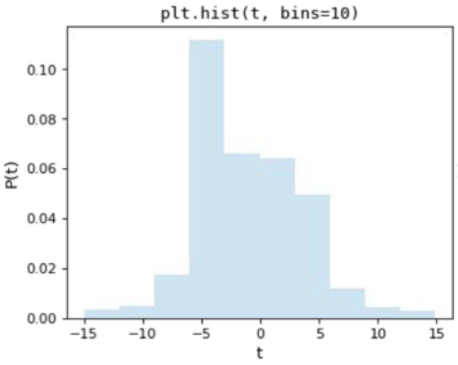
Nominal

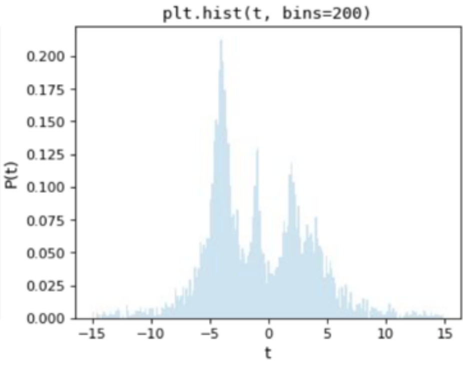
Sometimes we can carefully bin downward from higher to lower precision types...

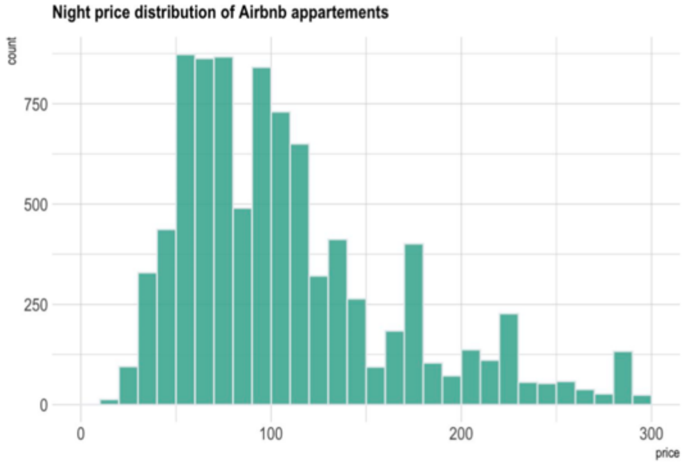
...but the other direction is usually either not possible, or highly irresponsible!

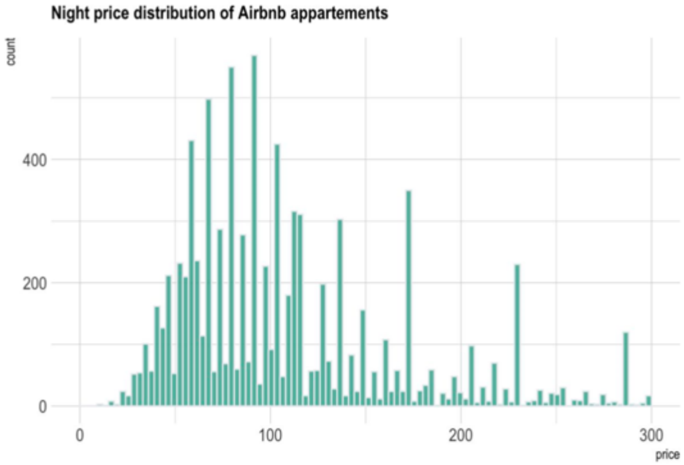
Same data, different bin widths:


Two bins. What's really in the 1+ category? Might be hiding something.
0
1+
That's better. It can show more variation.
0 1 2 3 4 5 6 7 8 9 10









https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

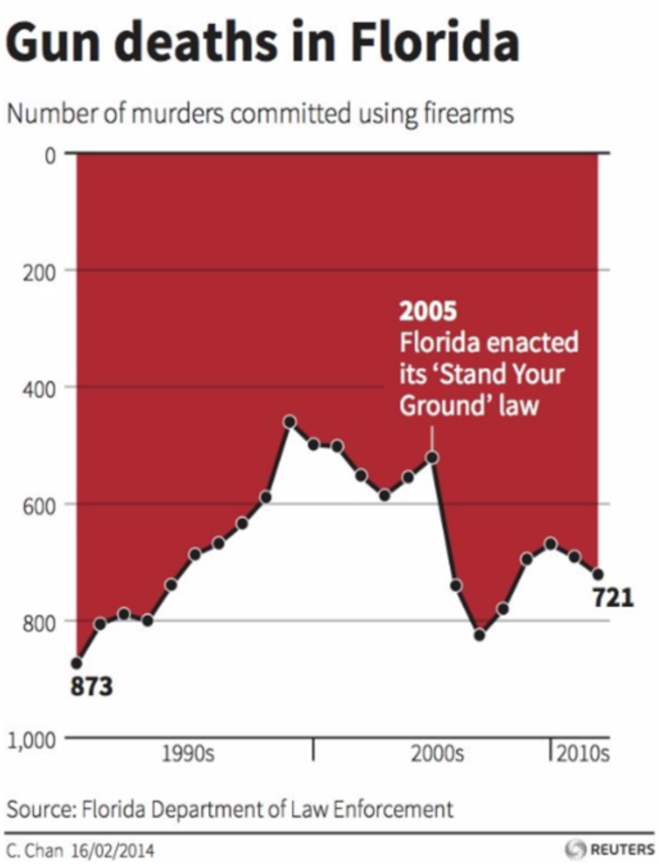
B. Are the data visualized <u>responsibly</u>?

Am I accurately representing the story that the data are telling?

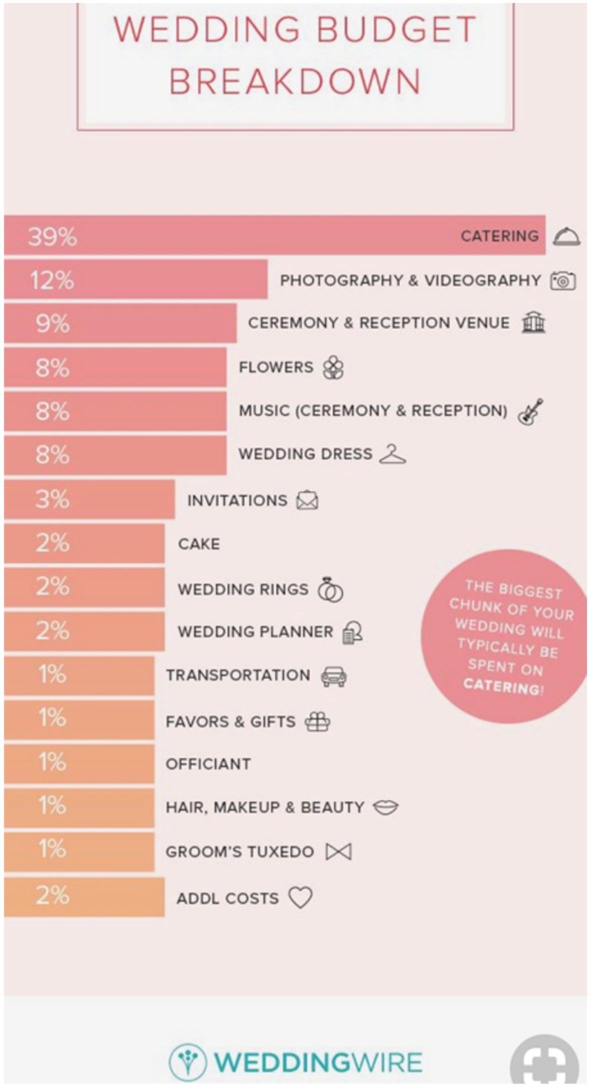
Including, not limited to:

- Reversing axes scale direction
- Scaling data without transparency
- Cropping axes scale to exaggerate differences
- Two y-axes, with intent to mislead
- Limited scope
- Unnecessary or misleading trend lines

Reversing axes scale direction

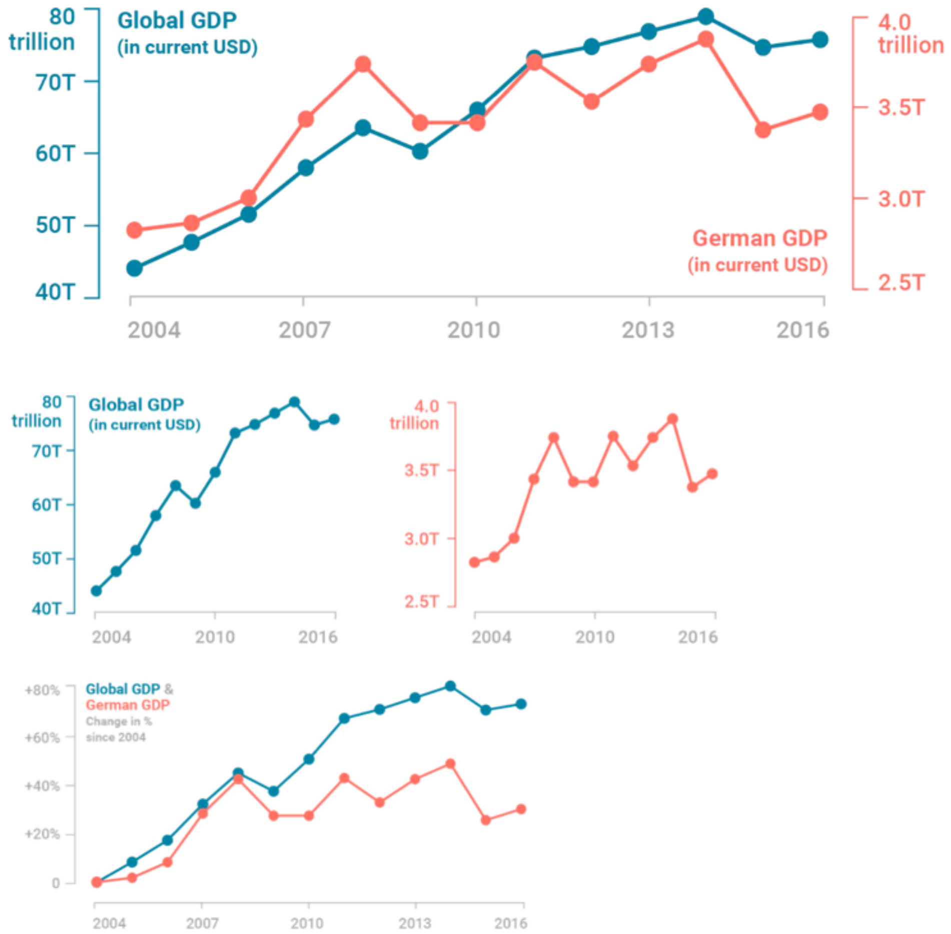


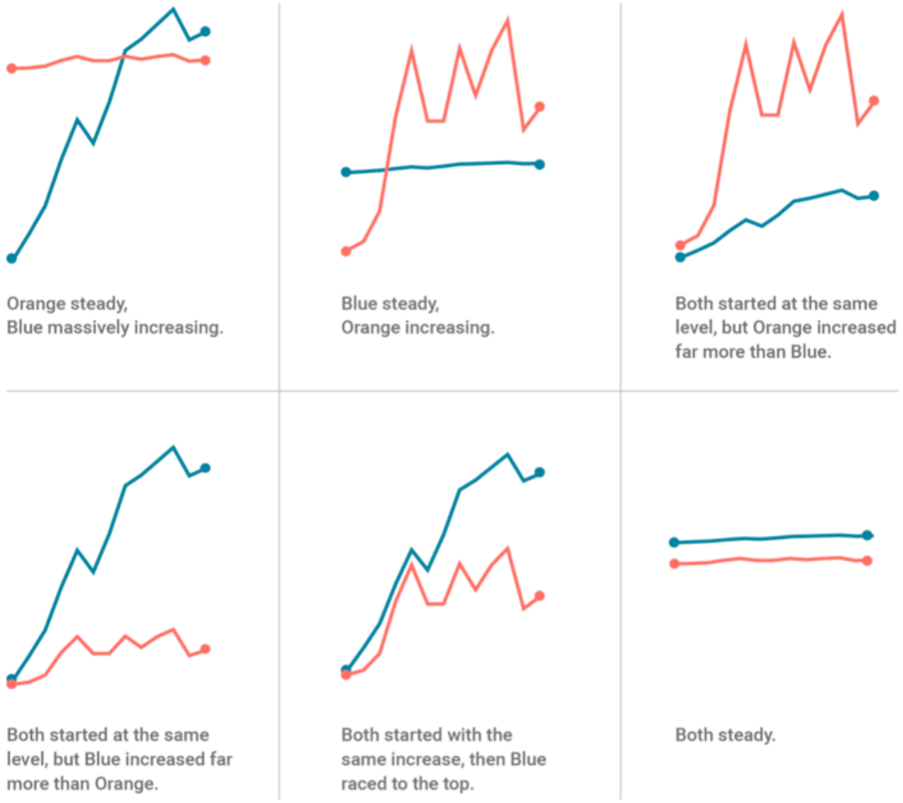
Scaling data without transparency



WTF Visualizations

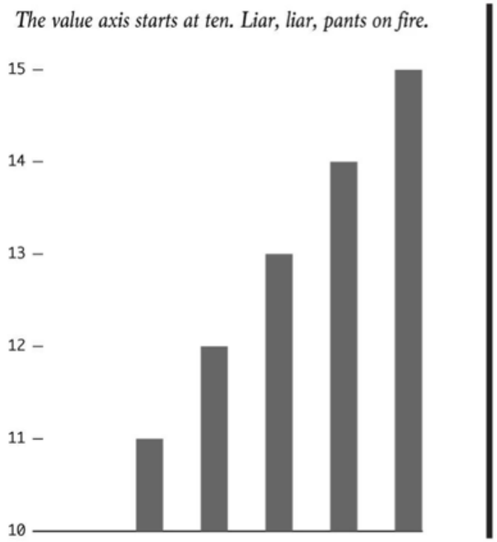
Two y-axes, with intent to mislead:

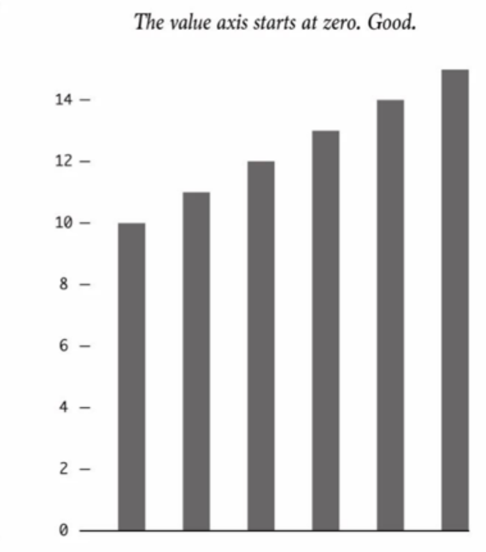


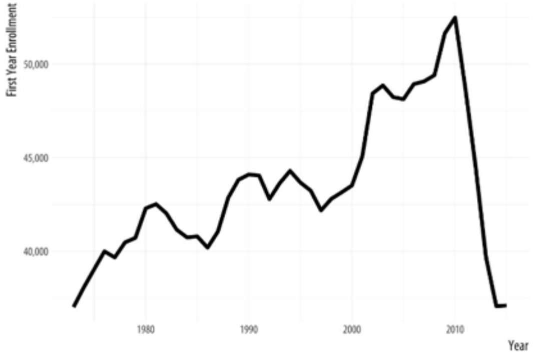


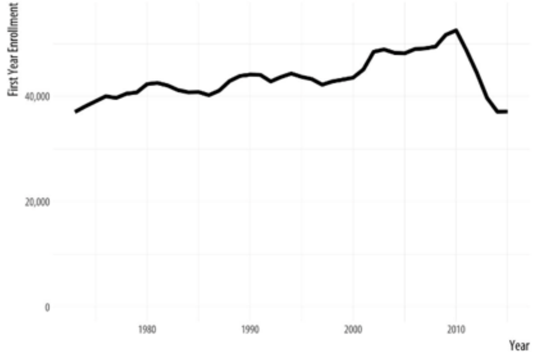
https://blog.datawrapper.de/dualaxis/

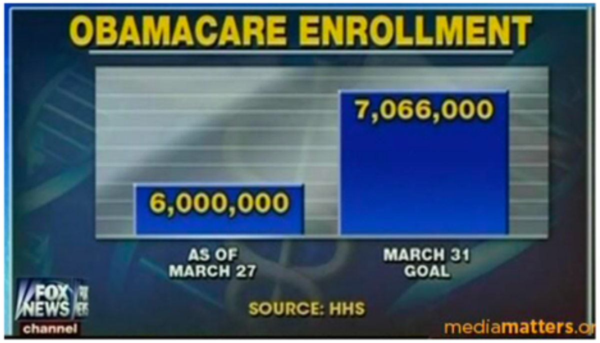
Cropping axes scale to exaggerate differences

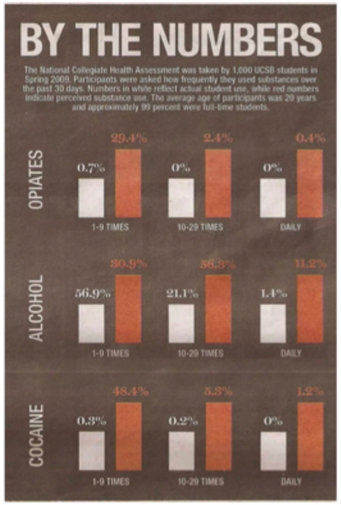








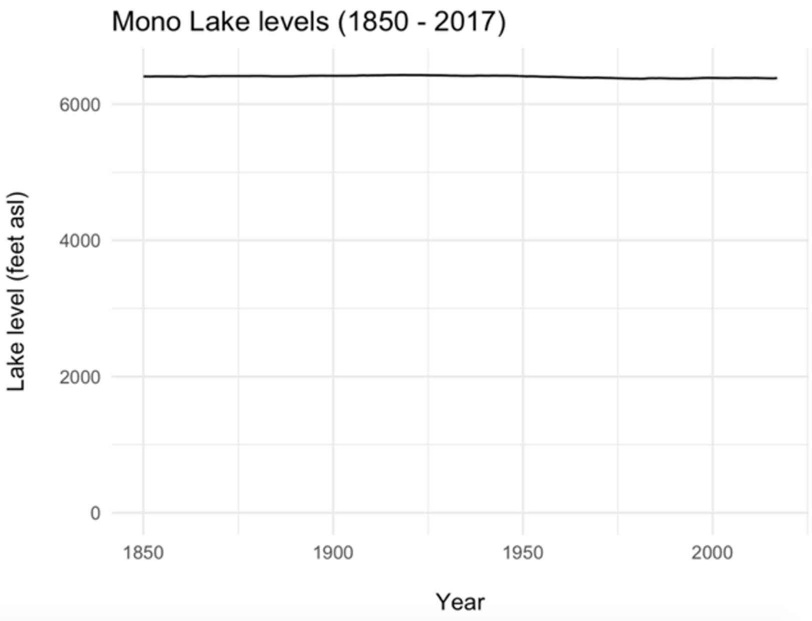


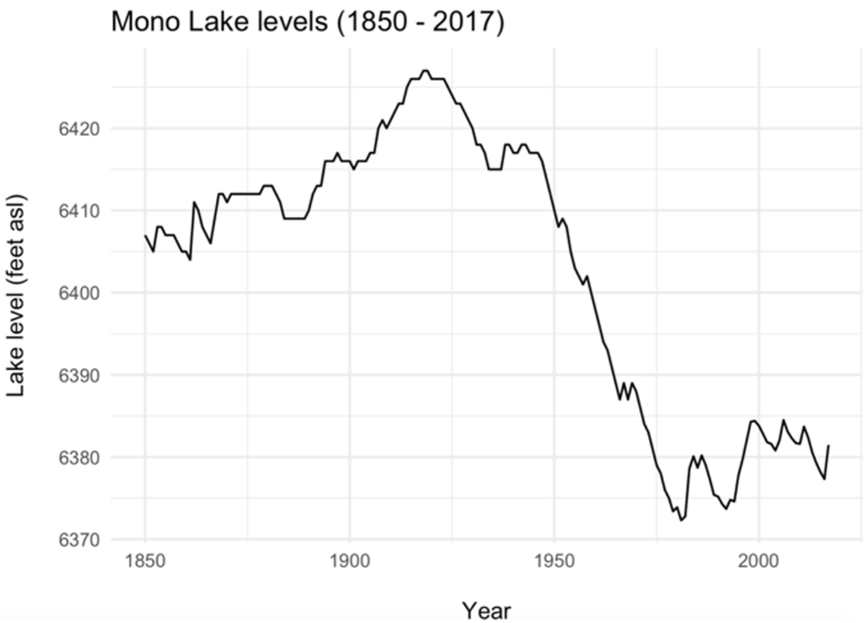

BY THE NUMBERS

The National Collegiate Health Assessment was taken by 1,000 UCSB students in Spring 2009. Participants were asked how frequently they used substances over the past 30 days. Numbers in white reflect actual student use, while red numbers indicate perceived substance use. The average age of participants was 20 years and approximately 99 percent were full-time students.

https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/
http://socviz.co/lookatdata.html

Exception: If value 0 isn’t a meaningful starting point, then it might make sense to not have a 0 starting point...

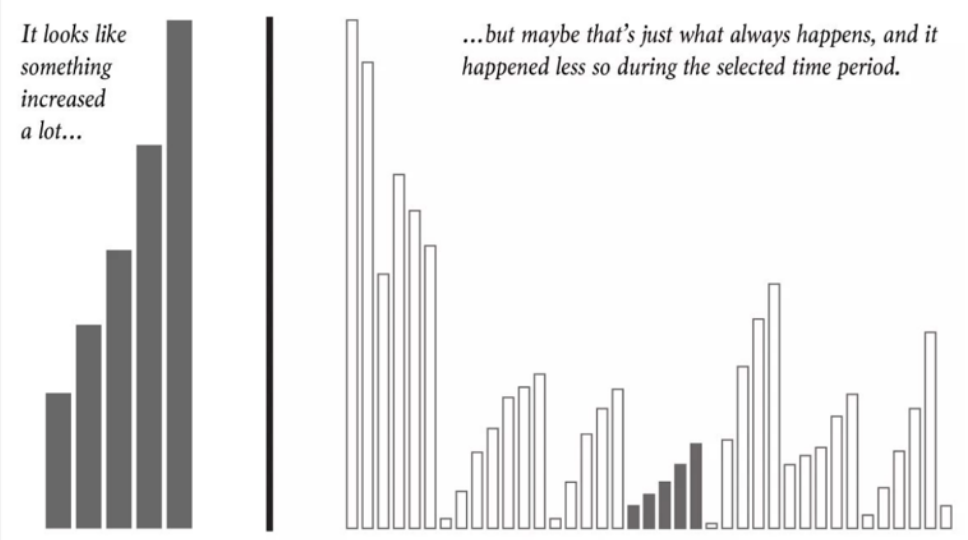


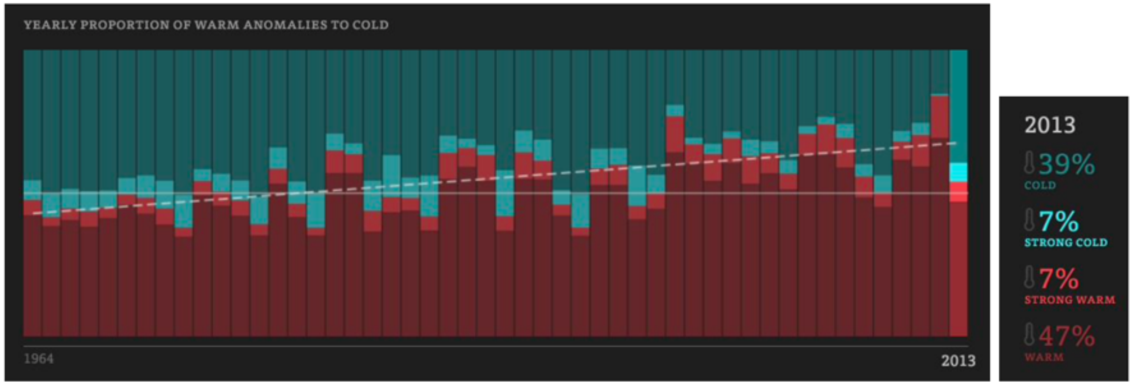


Data: 1912-1979 from LADWP and USGS compilations. 1979-present from Los Angeles Aqueduct Daily Reports, and observations by the Mono Lake Committee. Compiled by the Mono Lake Committee. Accessed from Mono Basin Clearinghouse.

Limited scope

Limiting variable ranges (especially time) in order to mislead audiences about trends / comparisons





https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

When it comes to axes and scales:

- Start at 0 unless you have a good reason not to (e.g., 0 is not part of the possible scale for that variable), and you've thought really hard about the possible misinterpretation / misrepresentation of your data that can result.

- Avoid dual axes. Again, avoid dual axes. I you decide you must use dual axes, be extremely cautious about bias and misrepresentation.

- Avoid scaling / transforming data. If you have to, make sure you're transparent in how it's been transformed.

are your summary statistics hiding something interesting?
@allison_horst

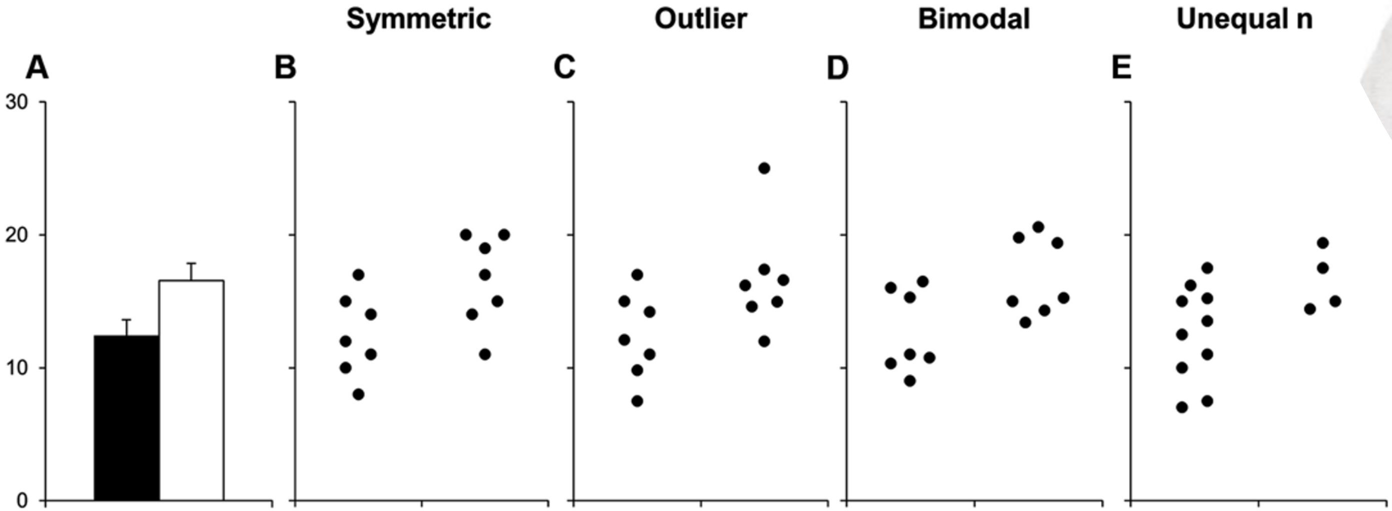
Artwork by Allison Horst

Are you hiding the true story of the data?

- Only showing summary statistics?
- If you are, are you clearly showing spread/uncertainty?
- Irresponsible trend lines?
- Reflecting study design?

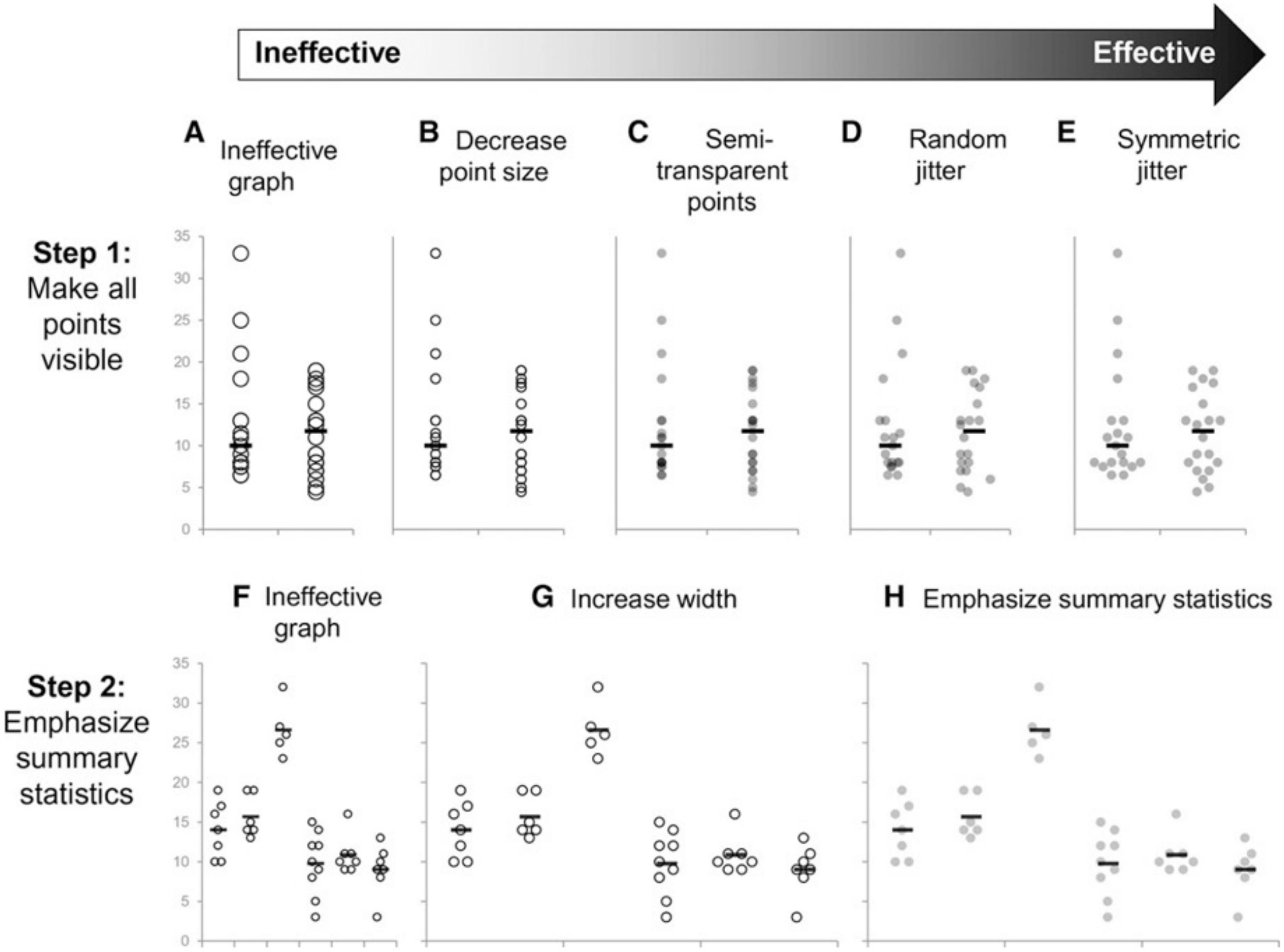
The problem with bar graphs

Underlying data
is inscrutable!



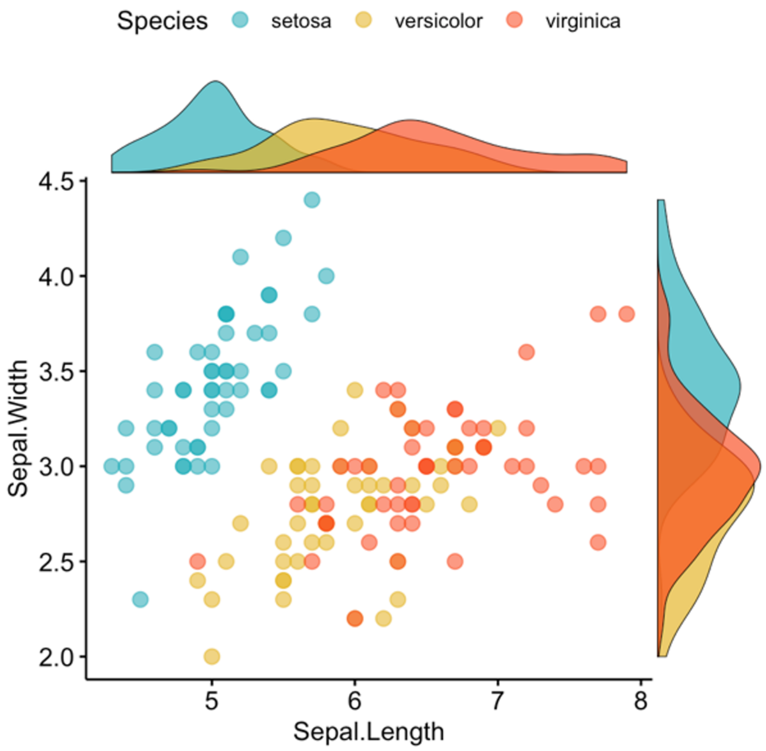
Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) ***Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm***. PLoS Biol 13(4): e1002128. https://doi.org/10.1371/journal.pbio.1002128

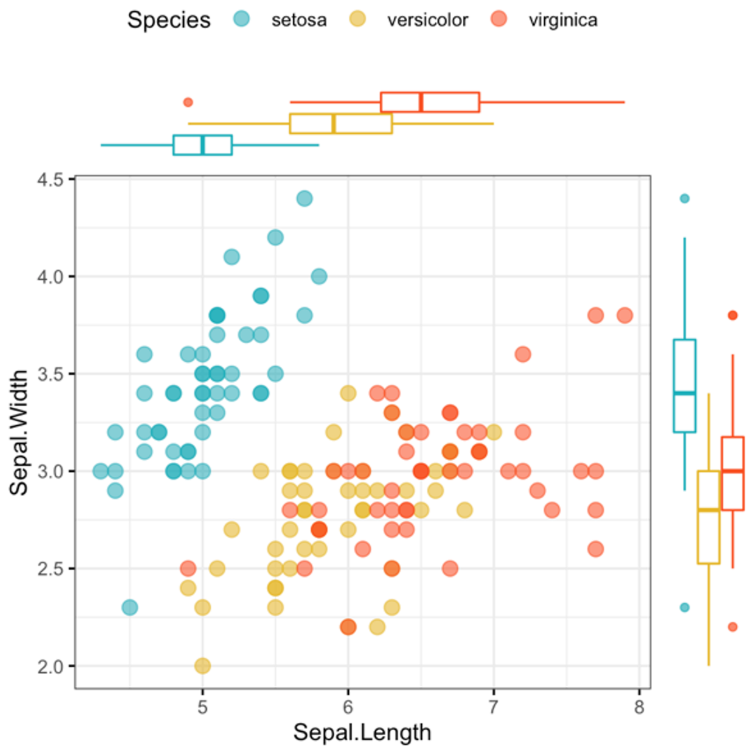
Show the data structure



Transforming Data Visualization to Improve Transparency, Weissgerber et al., 2019

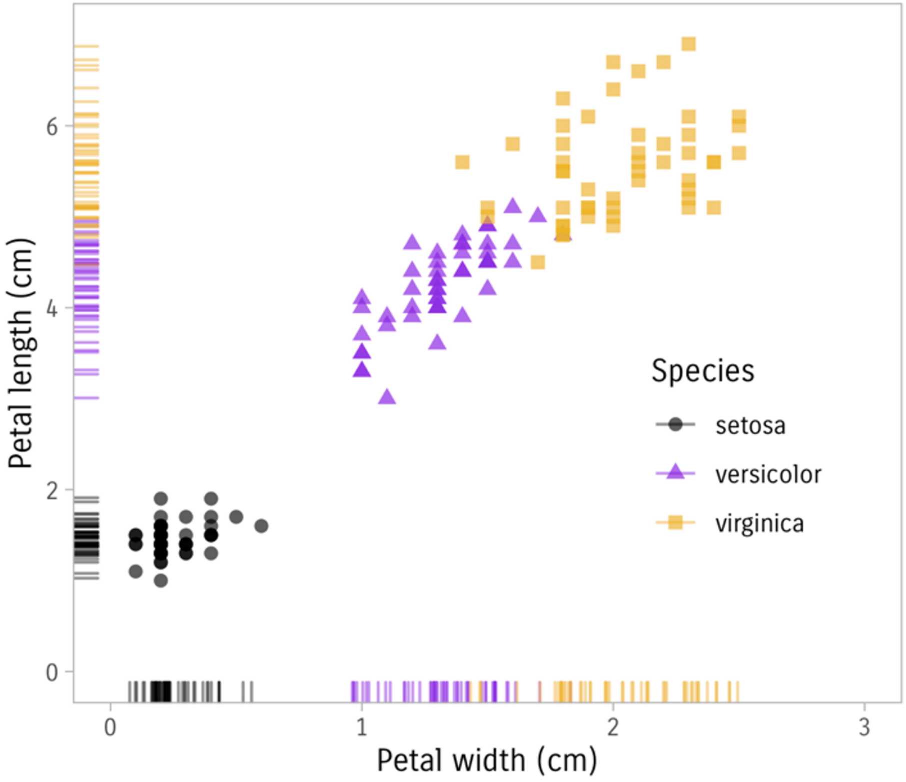
Another option to show data + summary: Marginal plots



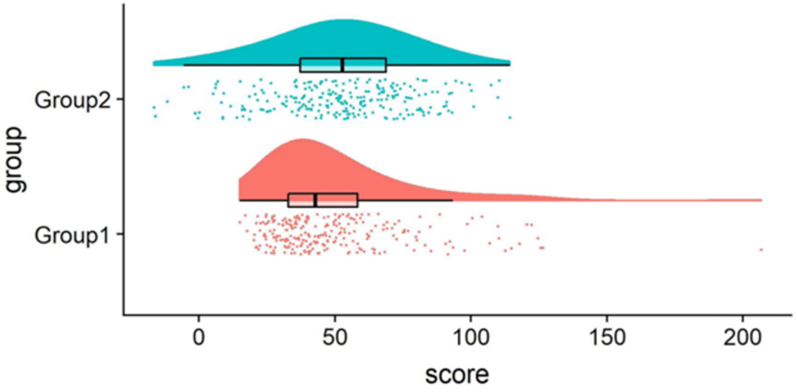


Datanovia.com ggplot Examples Reference

Rug plots



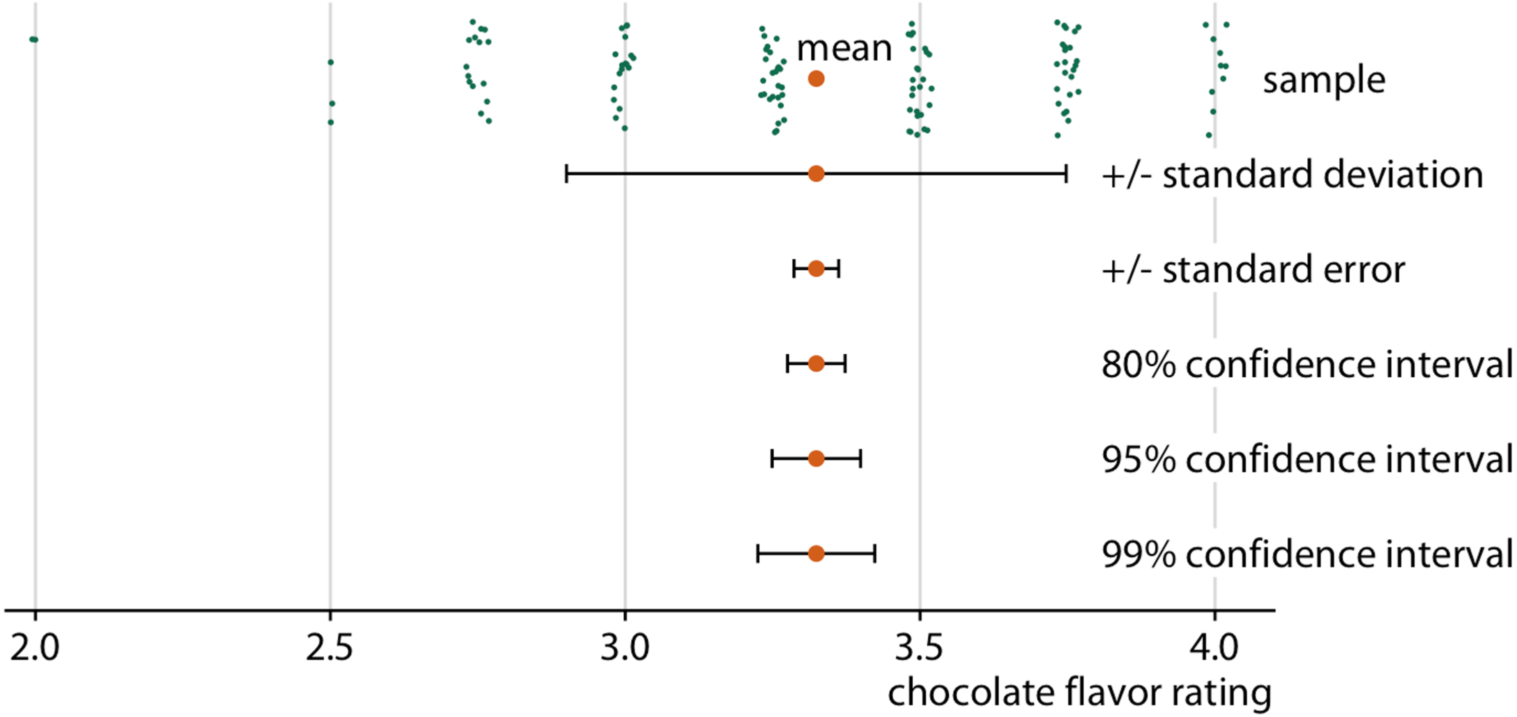
Raincloud plots



“The raincloud plot combines an illustration of data distribution (the ‘cloud’), with jittered raw data (the ‘rain’). This can further be supplemented by adding boxplots or other standard measures of central tendency and error.”

Allen M, Poggiali D, Whitaker K *et al.* Raincloud plots: a multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. *Wellcome Open Res* 2019, **4**:63 (https://doi.org/10.12688/wellcomeopenres.15191.1)

IF showing a summary statistic, ALSO show uncertainty:



Ch. 16, Fundamentals of Data Visualization by Claus O. Wilke

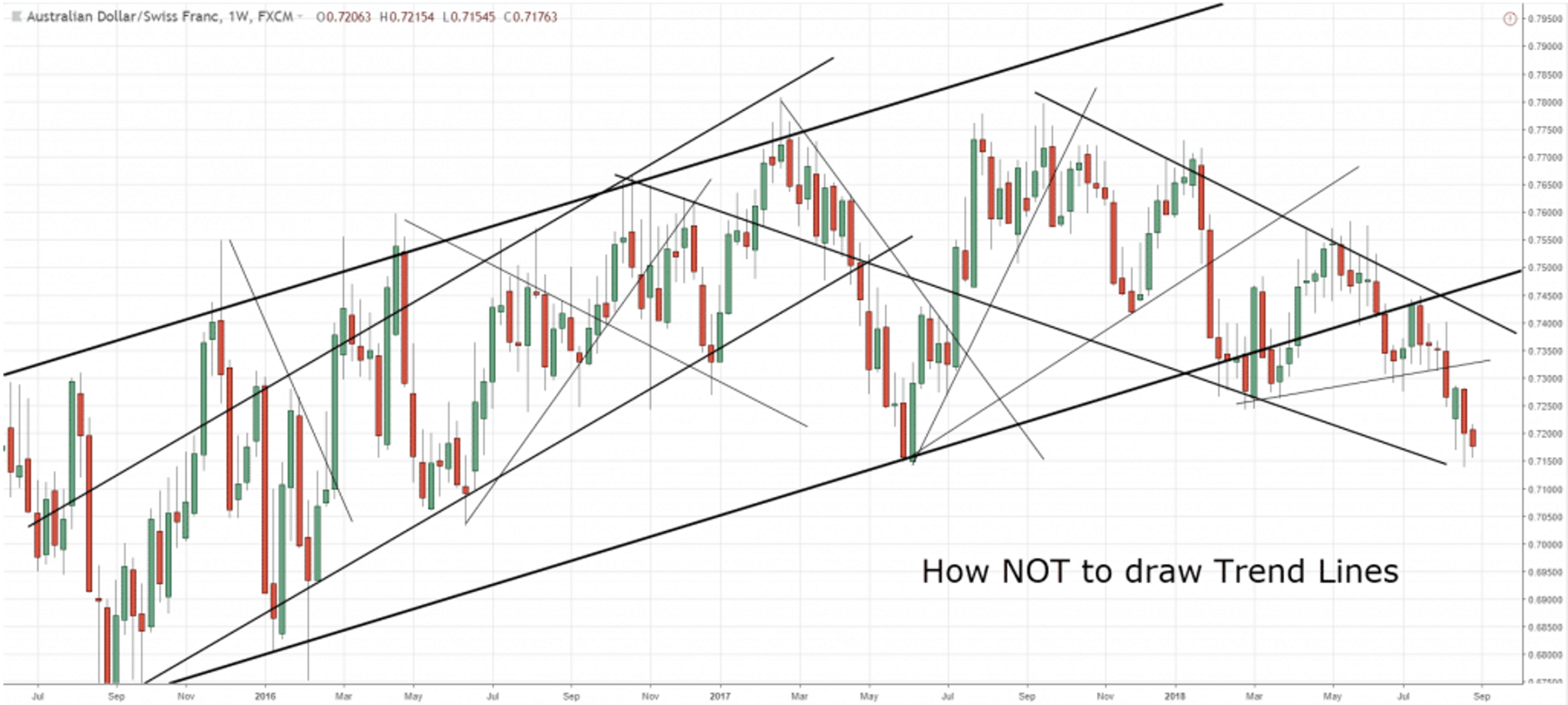
Trend lines

- **Trend lines are not data.** Consider the assumptions that go into adding trend lines: model, parameters, algorithm, ranges included, extrapolation, appearance, etc.

Trend lines

- **Trend lines are not data**. Consider the assumptions that go into adding trend lines: model, parameters, algorithm, ranges included, extrapolation, appearance, etc.

- **Trend lines can irresponsibly imply patterns** and stories that the data itself do not actually show themselves.



Trading with Rayner, Guide to Trendline Trading

BUT sometimes smoothing / trend lines are useful

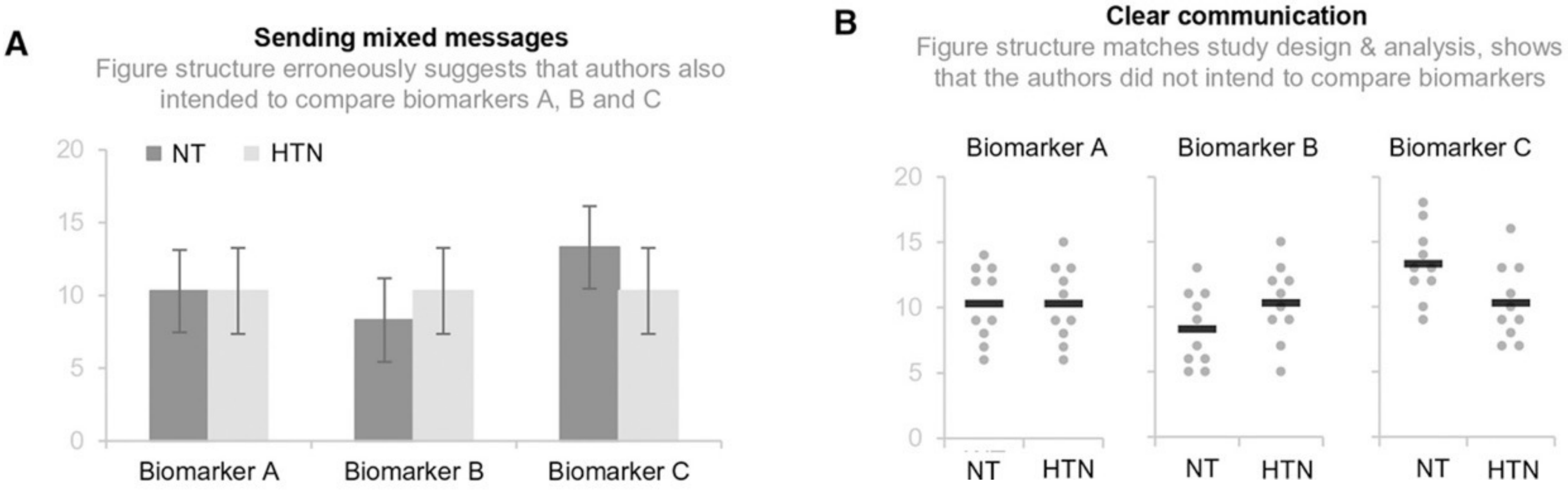
- When noise makes it hard **to see the actual patterns** that do exist in the data itself
- When it is valuable **to describe relationships between variables mathematically**, when you have done all necessary work to choose the appropriate model

Hey! What about model fit?!

Keep in mind: A equation and R^2 value is not a complete analysis or report of a model

Reflecting study design

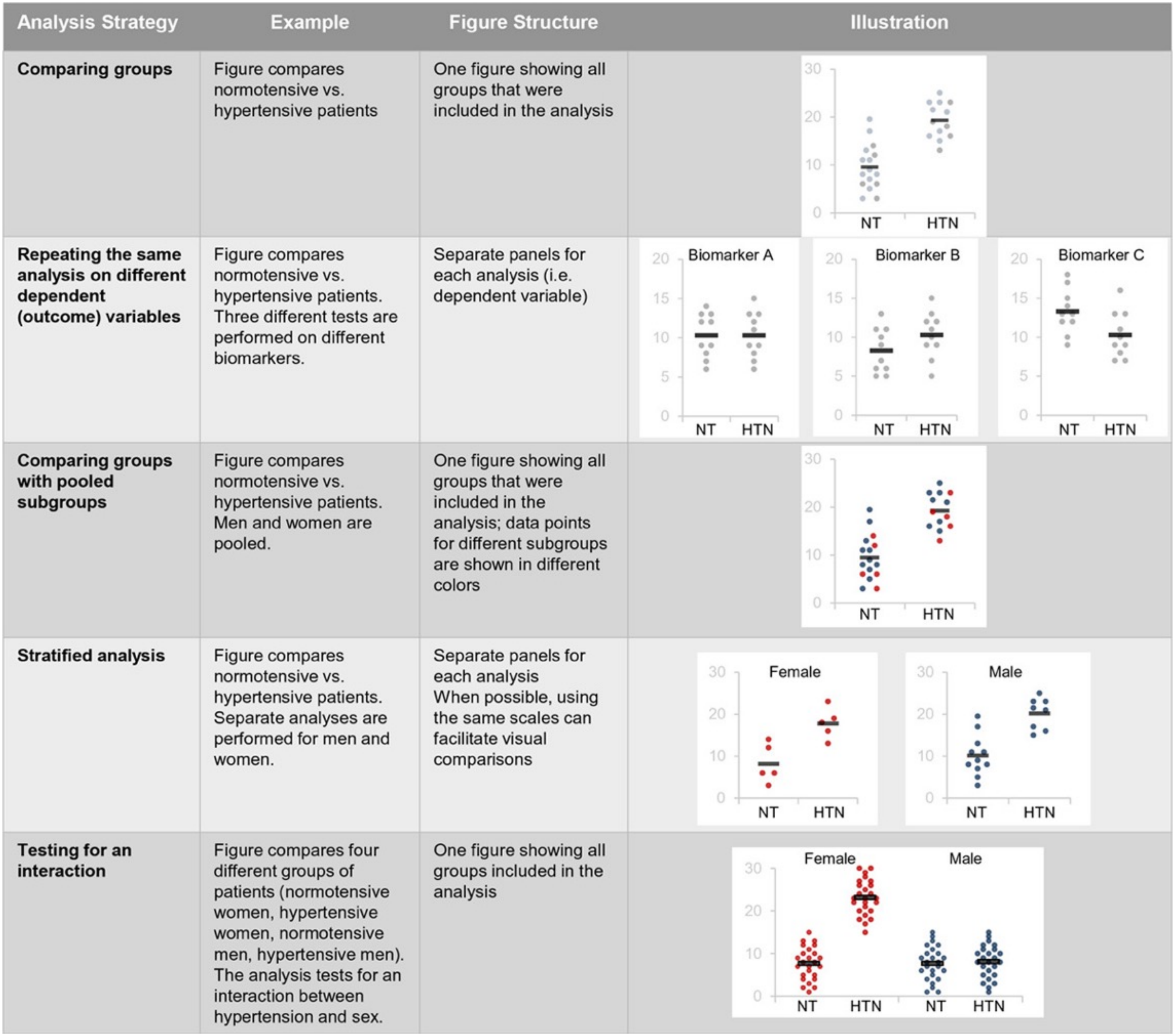
Experimental goal: Compare normotensive (NT) vs. hypertensive (HTN) patients
Statistical analysis: t-tests were used to compare values for each dependent variable (biomarker A, B and C)



A **Sending mixed messages**
Figure structure erroneously suggests that authors also intended to compare biomarkers A, B and C

B **Clear communication**
Figure structure matches study design & analysis, shows that the authors did not intend to compare biomarkers

Transforming Data Visualization to Improve Transparency, Weissgerber et al., 2019



Analysis Strategy	Example	Figure Structure	Illustration
Comparing groups	Figure compares normotensive vs. hypertensive patients	One figure showing all groups that were included in the analysis	
Repeating the same analysis on different dependent (outcome) variables	Figure compares normotensive vs. hypertensive patients. Three different tests are performed on different biomarkers.	Separate panels for each analysis (i.e. dependent variable)	
Comparing groups with pooled subgroups	Figure compares normotensive vs. hypertensive patients. Men and women are pooled.	One figure showing all groups that were included in the analysis; data points for different subgroups are shown in different colors	
Stratified analysis	Figure compares normotensive vs. hypertensive patients. Separate analyses are performed for men and women.	Separate panels for each analysis When possible, using the same scales can facilitate visual comparisons	
Testing for an interaction	Figure compares four different groups of patients (normotensive women, hypertensive women, normotensive men, hypertensive men). The analysis tests for an interaction between hypertension and sex.	One figure showing all groups included in the analysis	

Transforming Data Visualization to Improve Transparency, Weissgerber et al., 2019

Quick break!

2

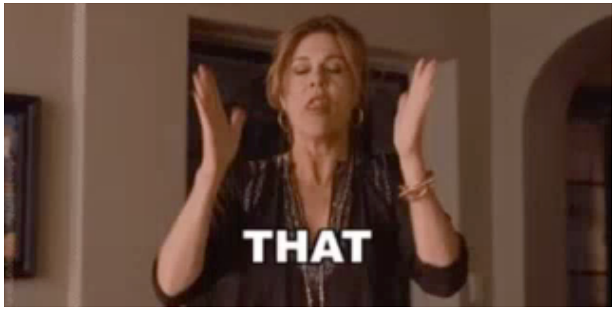
Clear data viz for your audience

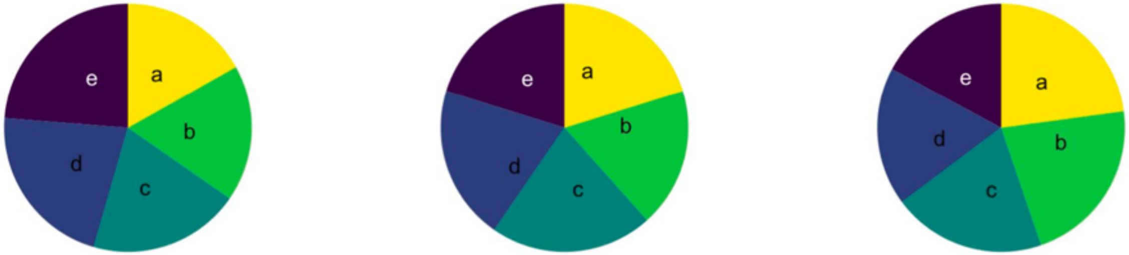
Understand the basic principles behind effective data visualization.

Clear data viz for your audience

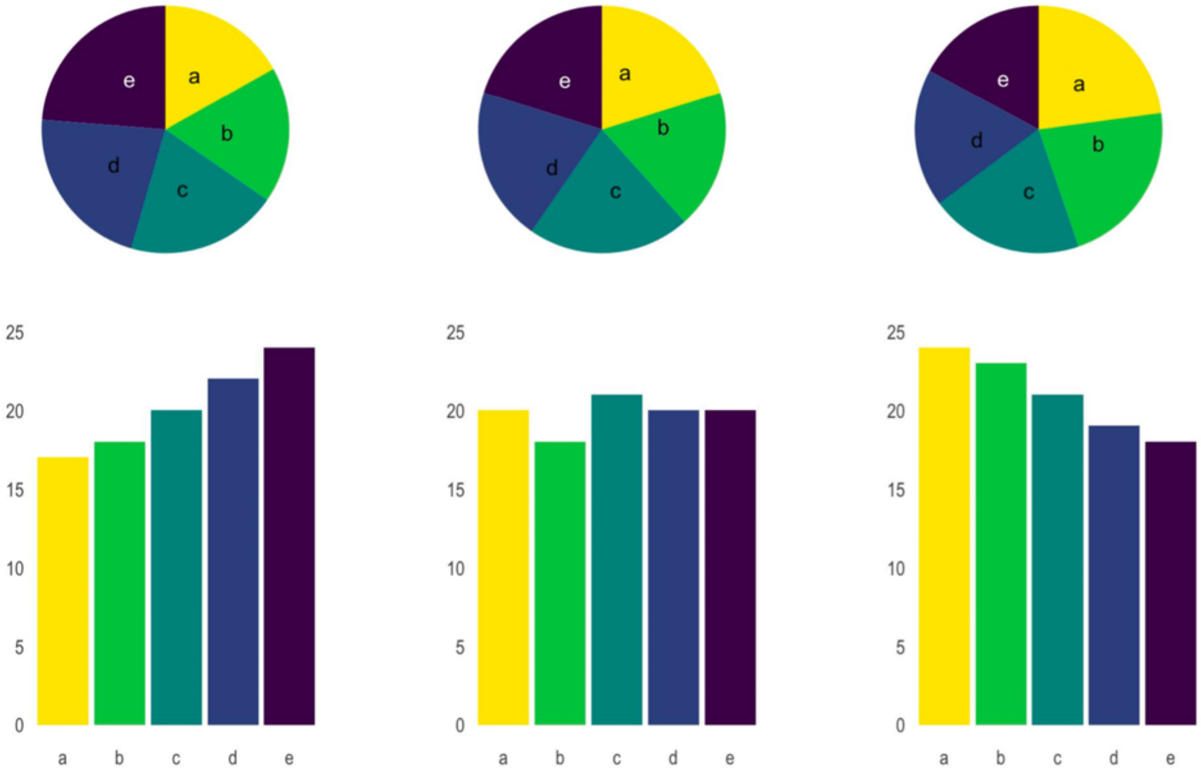
a. Pie charts – almost never a good idea
b. Principles for effective visualizations
c. Audience-centric data viz considerations

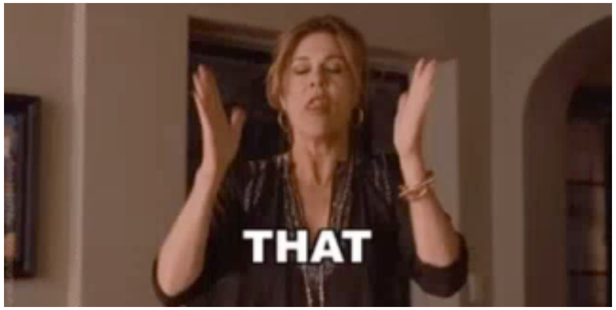
Pie charts?





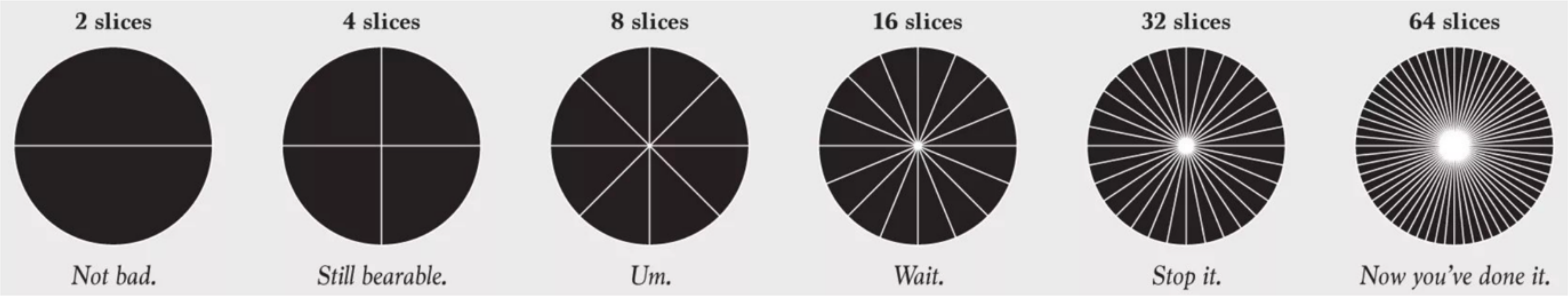
Pie charts?

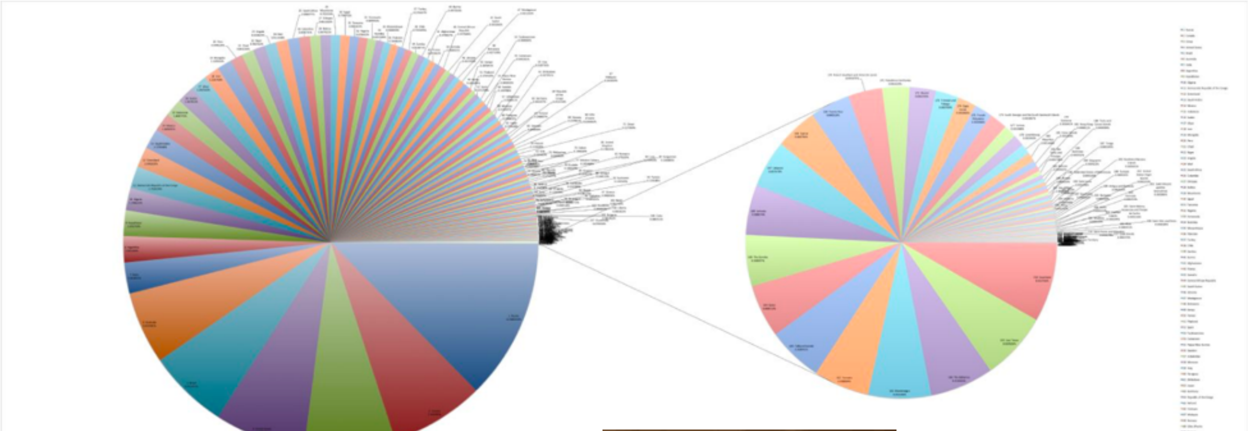




– almost never a good idea

https://www.data-to-viz.com/caveat/



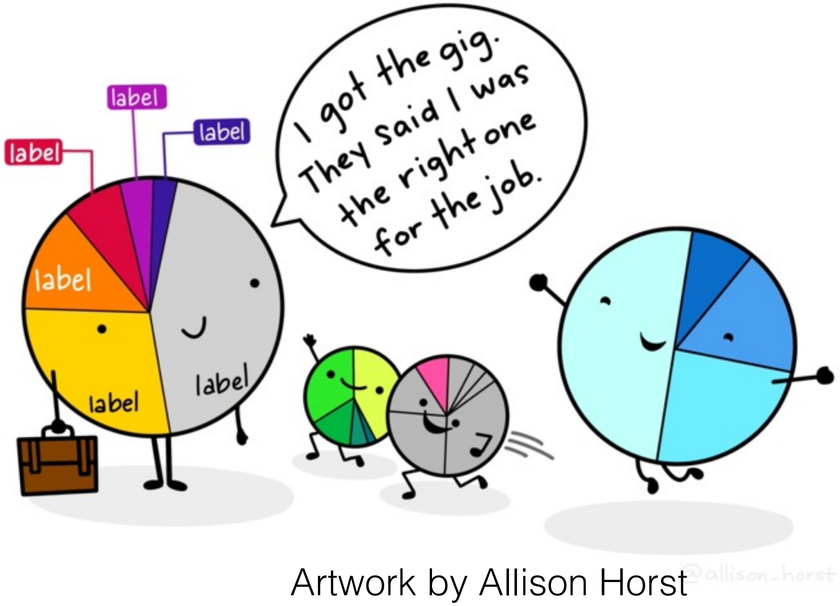


https://flowingdata.com/2015/08/11/real-chart-rules-to-follow/

https://commons.wikimedia.org/wiki/File:Pie_chart_of_countries_by_area.png

IF you decide a pie chart is a good option:

- Are proportions different enough to notice quickly & easily?
- Avoid a ton of wedges (> 7 too many?)
- Emphasize one by highlighting or having it “pop-out”
- Label directly
- Always **compare to a bar chart** version to see which makes the data story clearer for your audience



Artwork by Allison Horst

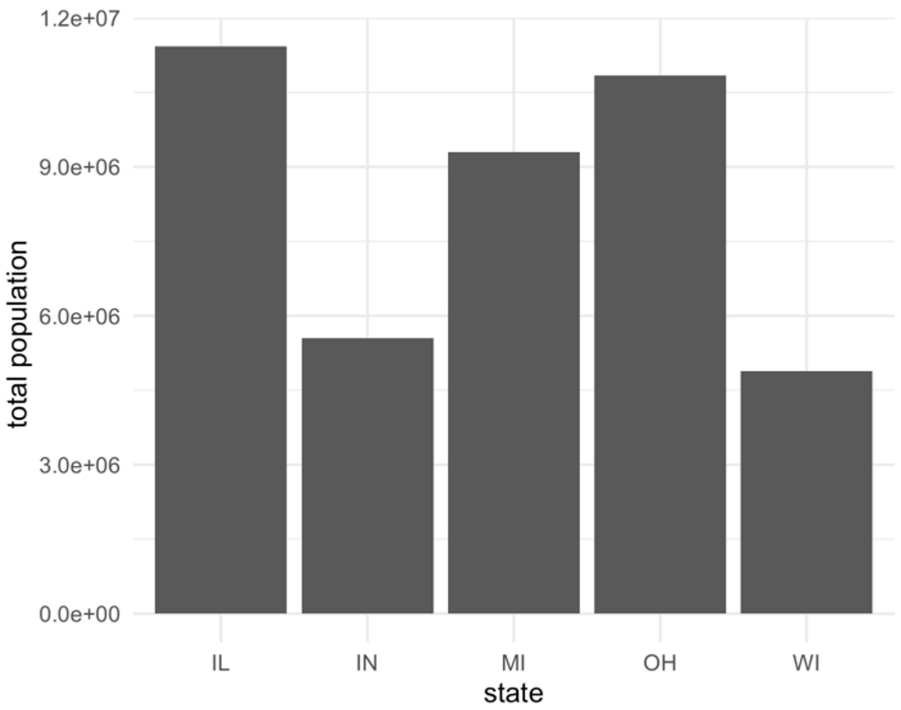
For clearer data viz:

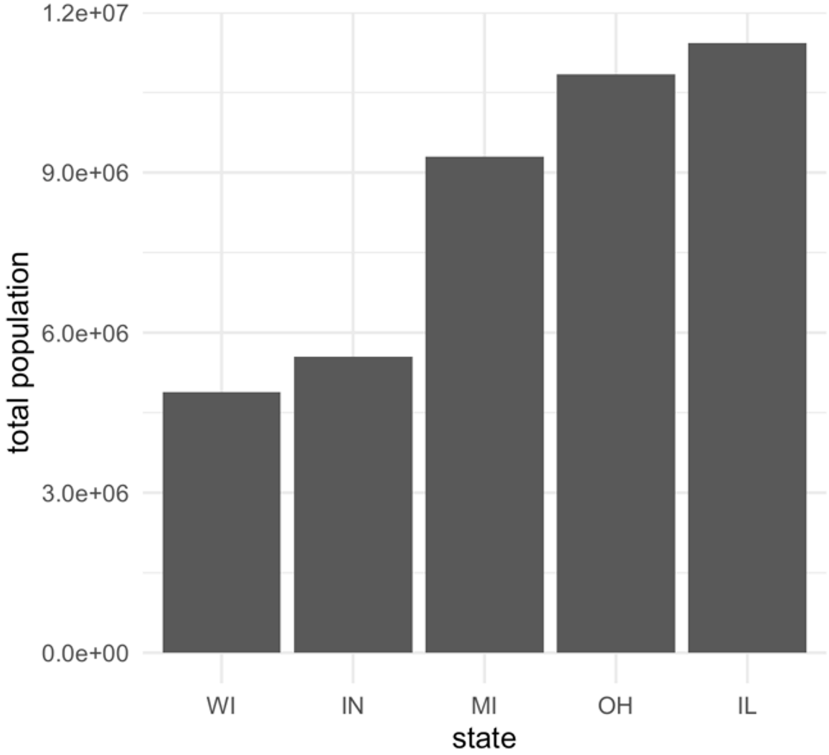
- Label axes
- Remove distractions
- Emphasize as useful / relevant
- Simplify. Facet? Smaller pieces > one giant beast graph
- Put things in meaningful order
- Customize legends (or remove & label instead)
- Add context (labels, annotation, etc.)
- Avoid abbreviations
- Careful with data transform (e.g. log, semi-log, etc.)

Principles for effective visualizations

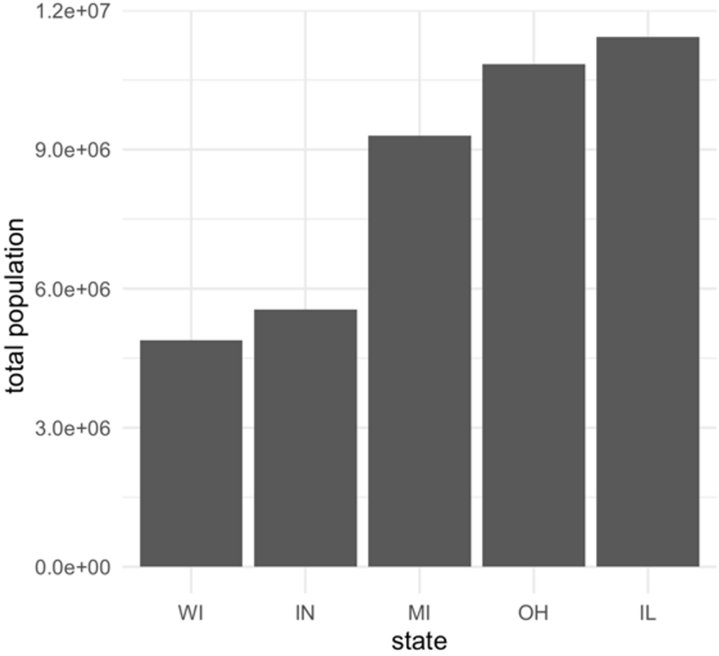
1. Put things in meaningful **order**
2. **Tell a story**: emphasis / highlight / faceting
3. **Use text to clarify**: direct labeling
4. **Keep it simple**, less is more!

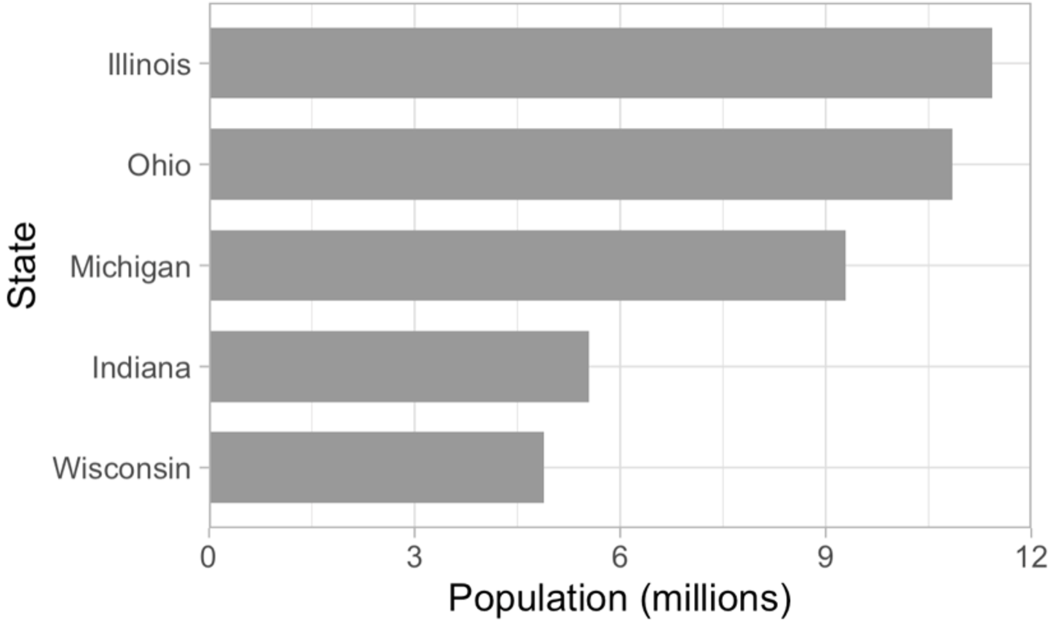
Ordering things makes graphs feel less overwhelming





Taking it a bit further (& aesthetics preview):

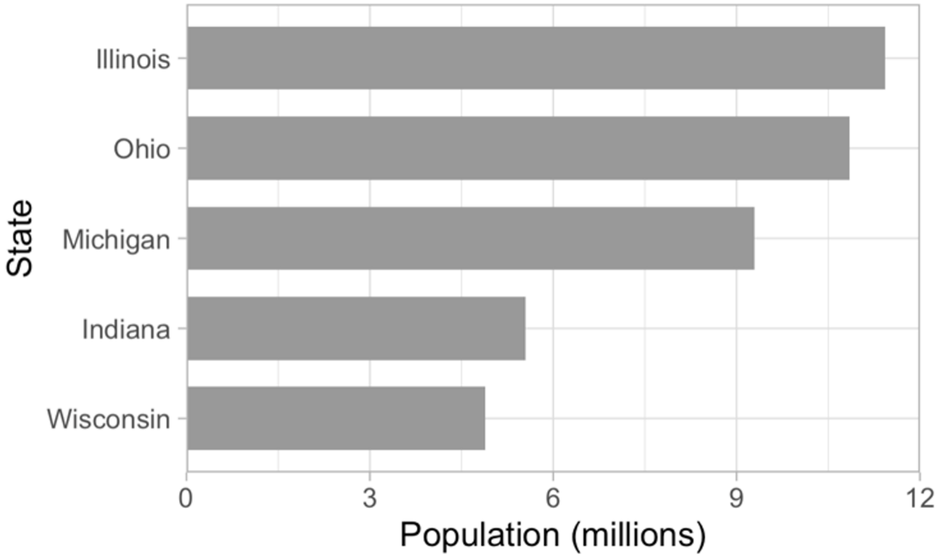


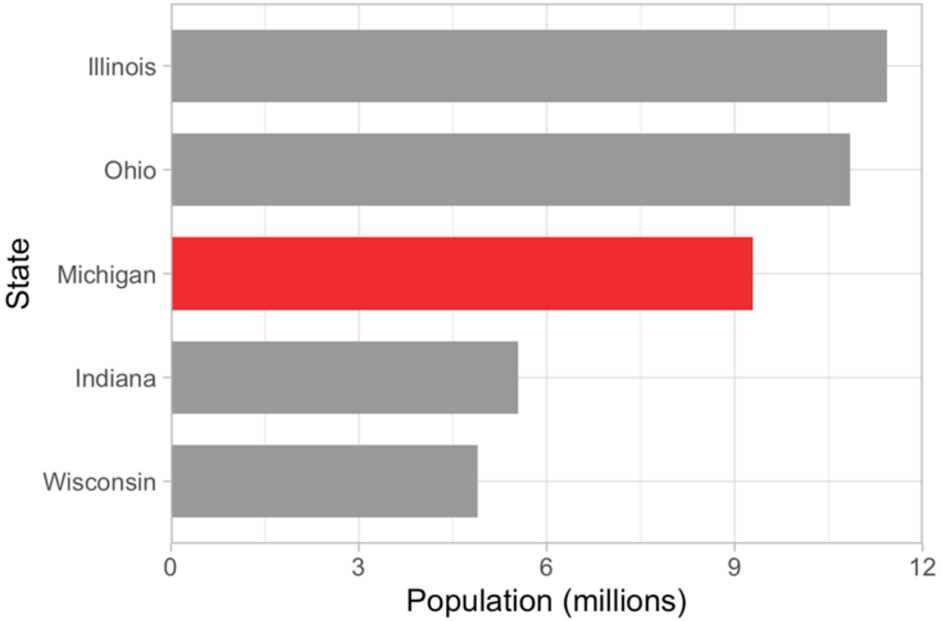


Put long categories on y-axis
Axis labels:

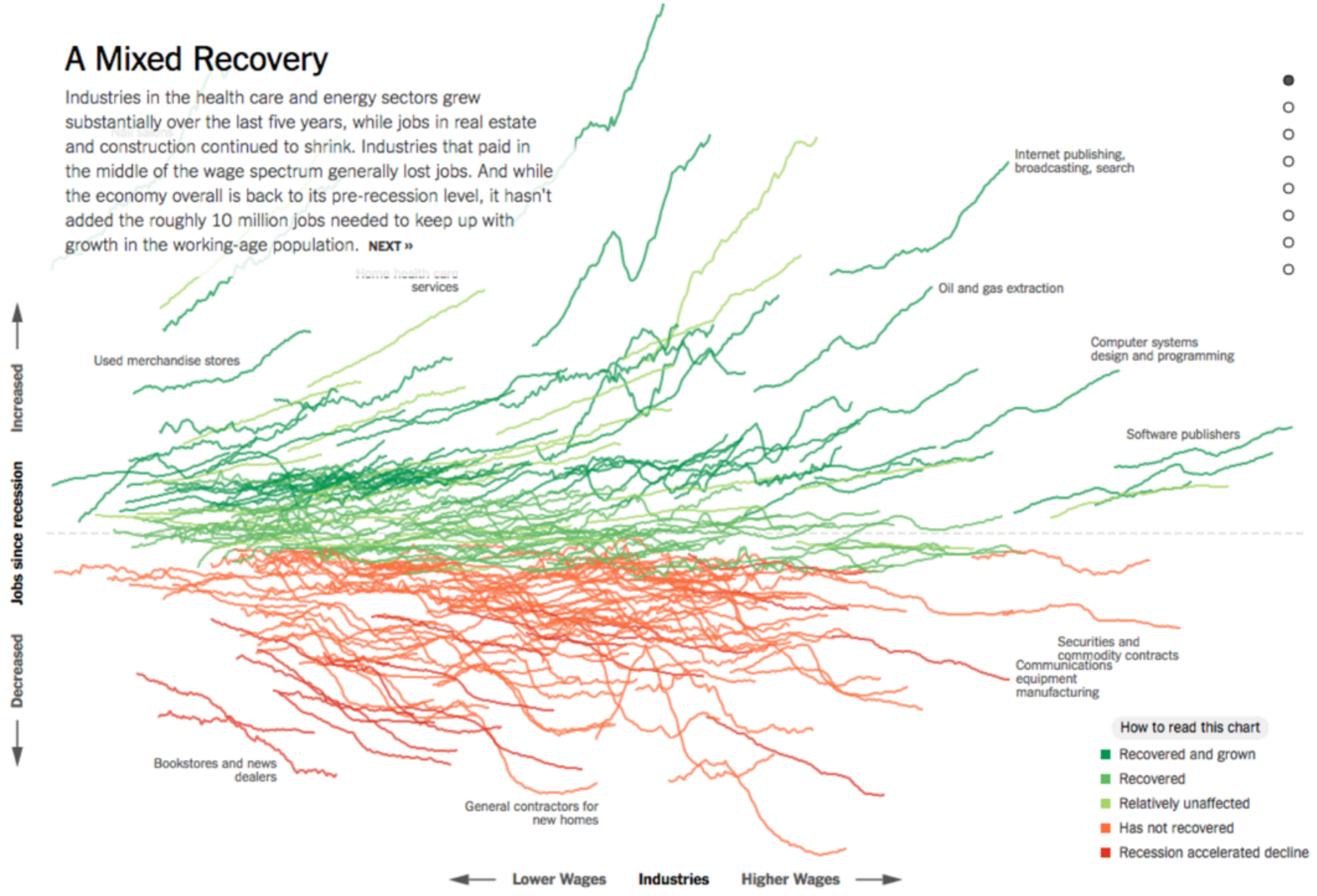
- Briefly describe variable
- Need units as relevant
- Perfect notation, case
- Avoid abbreviations

Once things are in order, **highlight** the series / levels that you want the audience to focus on





All the data doesn’t tell a story



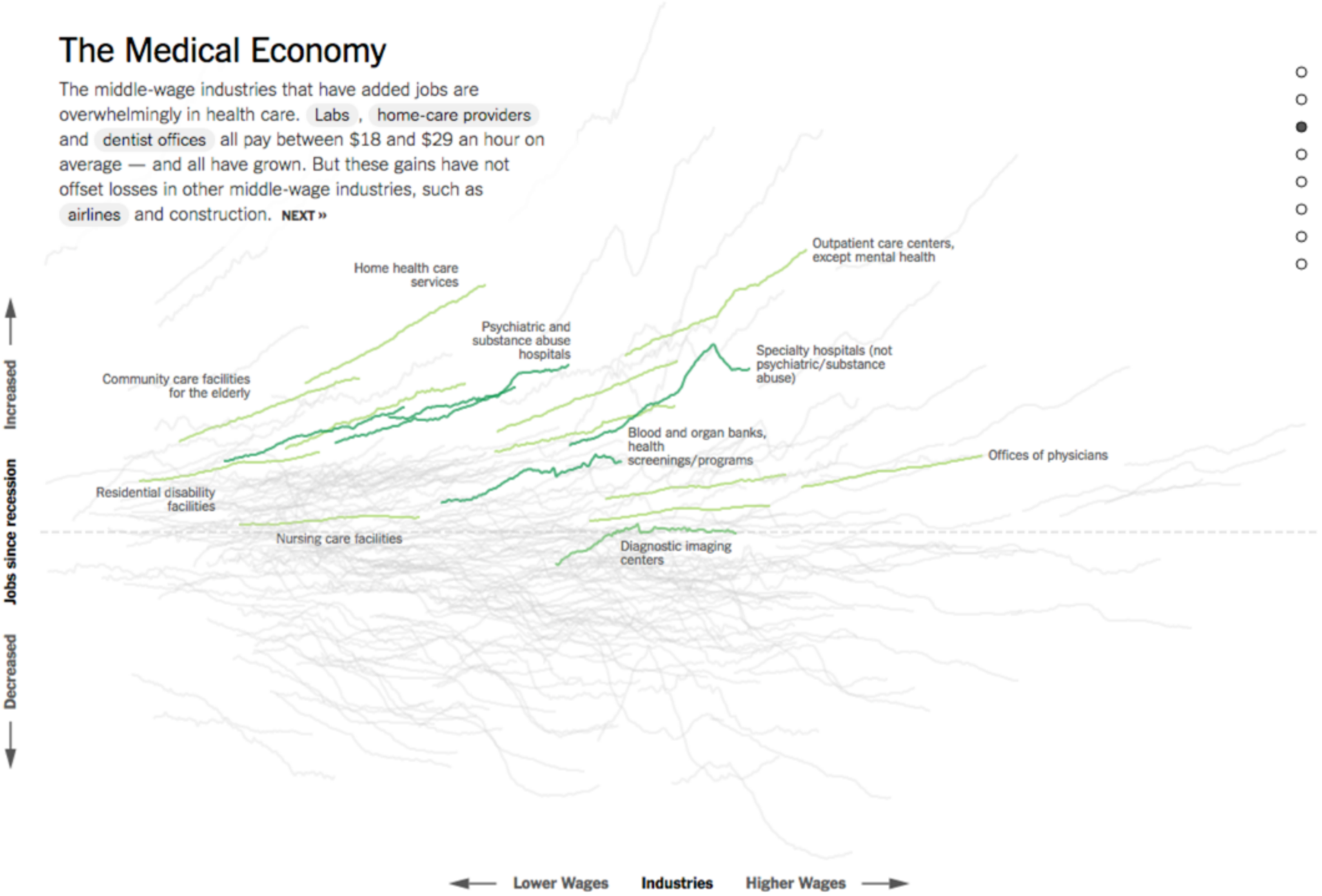
A Mixed Recovery

Industries in the health care and energy sectors grew substantially over the last five years, while jobs in real estate and construction continued to shrink. Industries that paid in the middle of the wage spectrum generally lost jobs. And while the economy overall is back to its pre-recession level, it hasn't added the roughly 10 million jobs needed to keep up with growth in the working-age population. **NEXT »**

Home health care services
Used merchandise stores
Internet publishing, broadcasting, search
Oil and gas extraction
Computer systems design and programming
Software publishers
Securities and commodity contracts
Communications equipment manufacturing
Bookstores and news dealers
General contractors for new homes
Increased
Jobs since recession
Decreased
Lower Wages
Industries
Higher Wages
How to read this chart
■ Recovered and grown
■ Recovered
■ Relatively unaffected
■ Has not recovered
■ Recession accelerated decline

http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

All the data doesn't tell a story

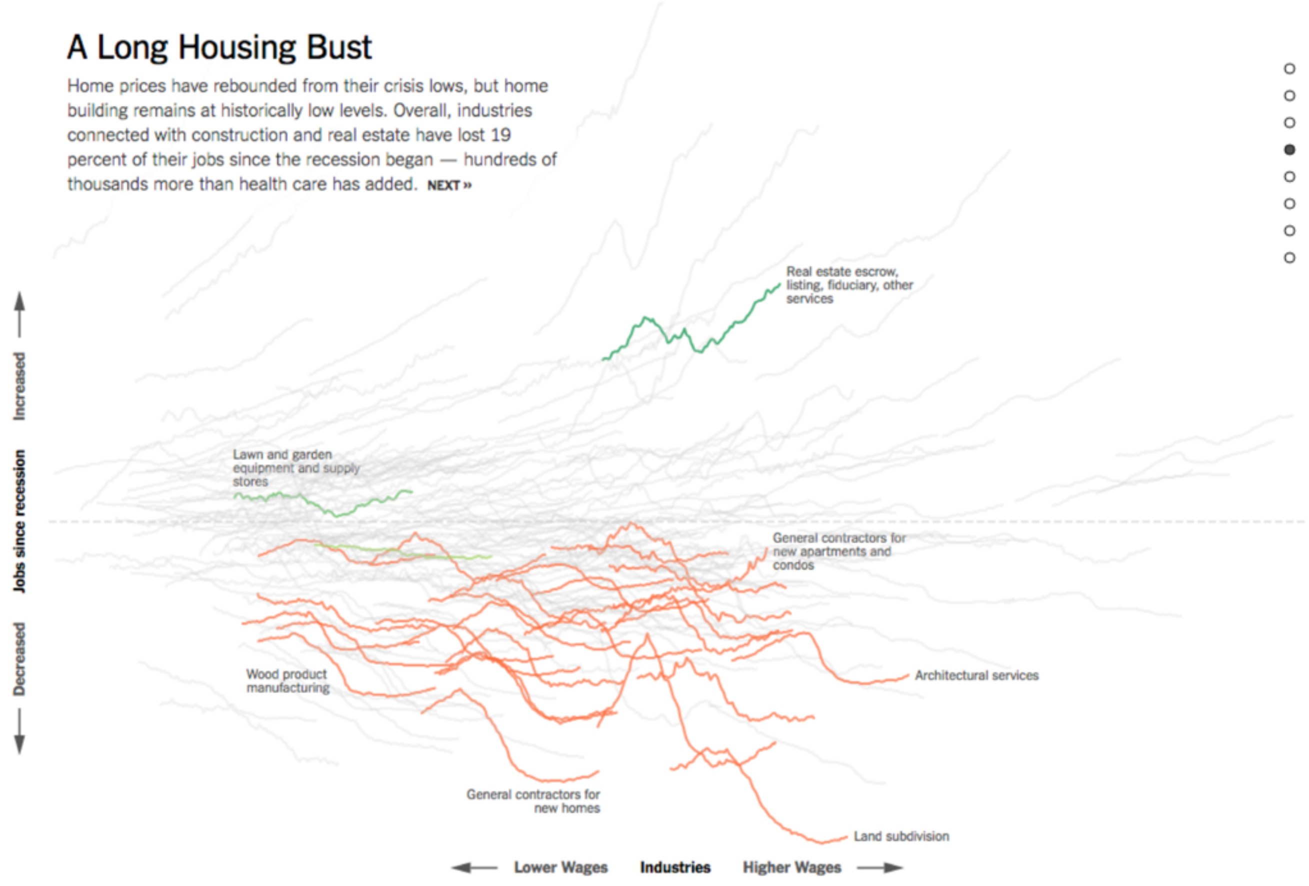


The Medical Economy

The middle-wage industries that have added jobs are overwhelmingly in health care. Labs , home-care providers and dentist offices all pay between $18 and $29 an hour on average — and all have grown. But these gains have not offset losses in other middle-wage industries, such as airlines and construction. **NEXT »**

http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

All the data doesn't tell a story



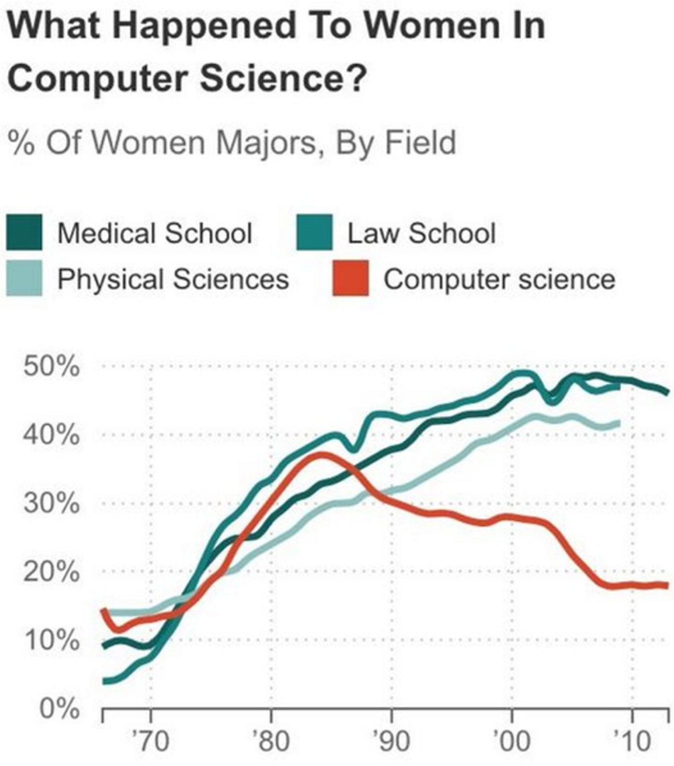
A Long Housing Bust

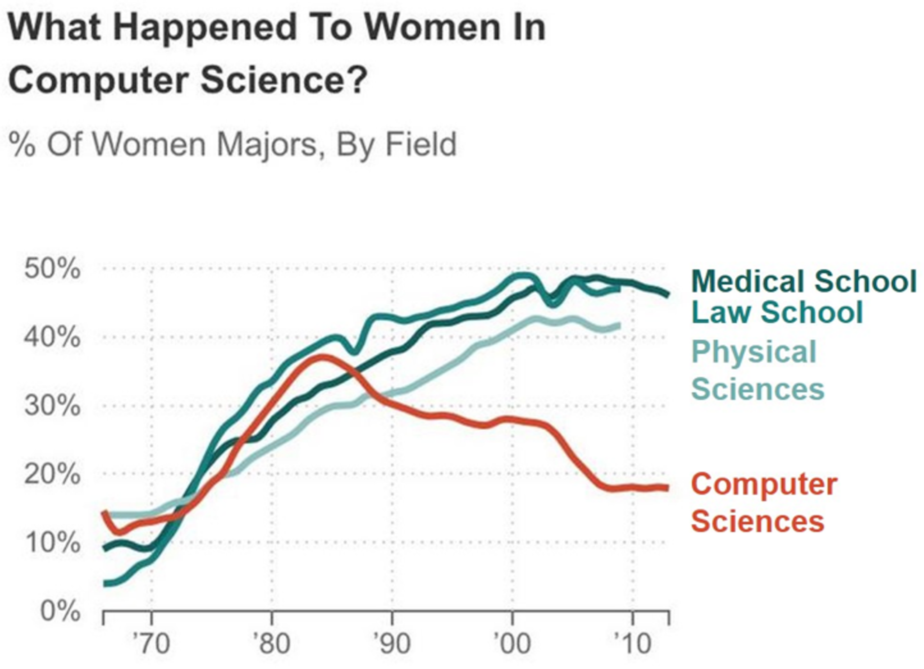
Home prices have rebounded from their crisis lows, but home building remains at historically low levels. Overall, industries connected with construction and real estate have lost 19 percent of their jobs since the recession began — hundreds of thousands more than health care has added. **NEXT »**

http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

Annotation: beyond legends

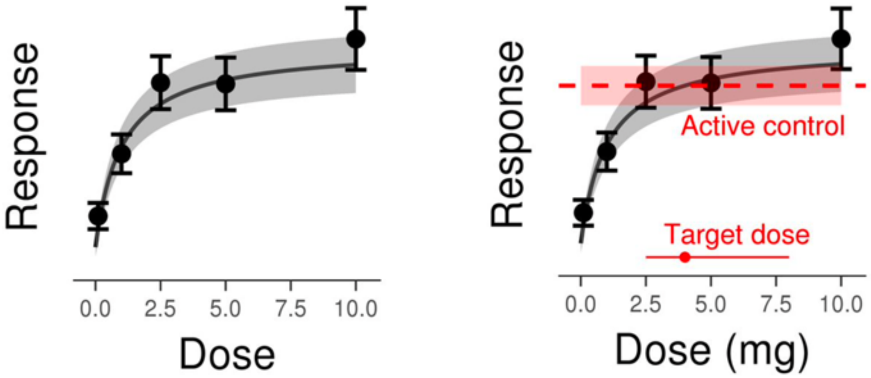
- Labels aren’t just for legend replacement
- Use labels & annotation strategically in graphs
- Use descriptive titles





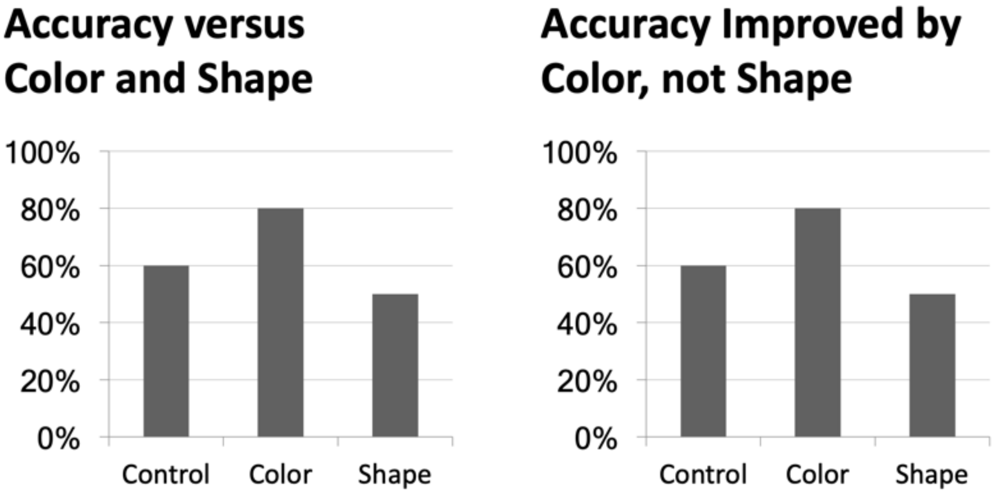
Depict Data Studio *Directly Labeling Your Line Graphs*

Informative labels and annotations to support the message



https://graphicsprinciples.github.io/

Active tittles to summarize your message



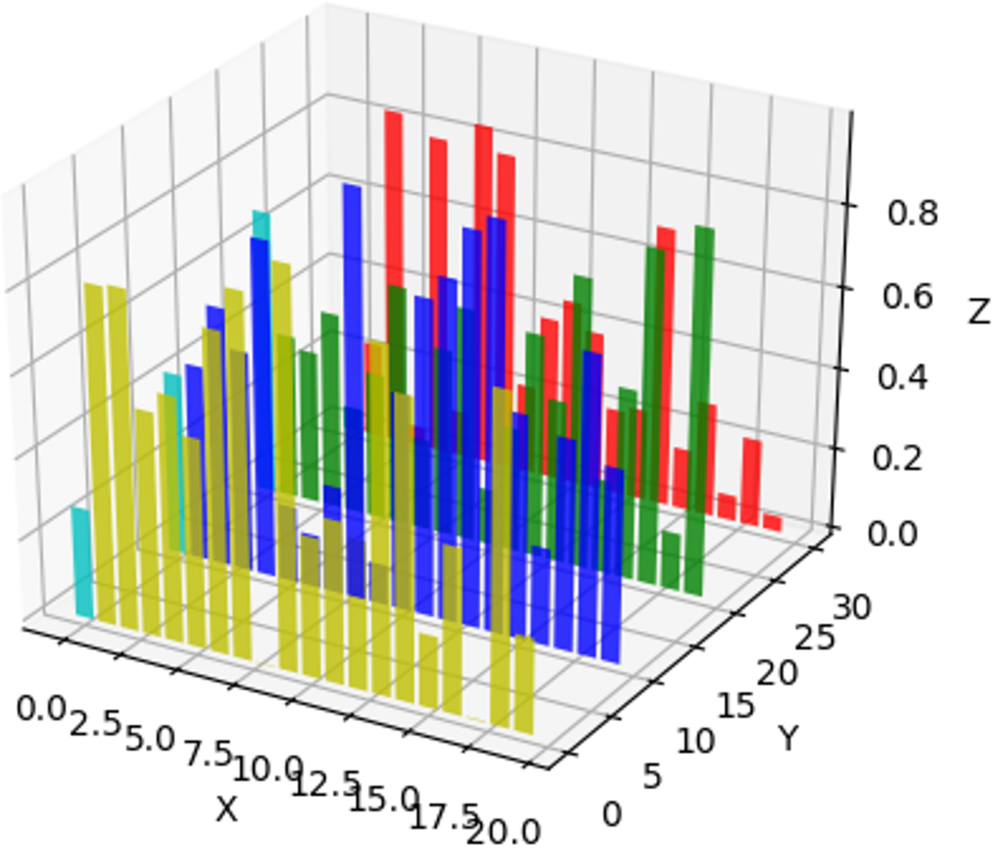
Accuracy versus Color and Shape

Accuracy Improved by Color, not Shape

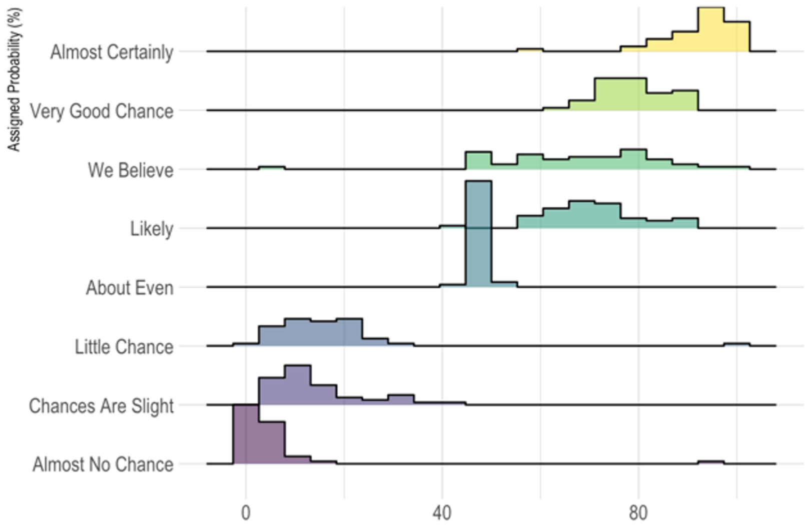
Keep it simple!

3D plots? No*

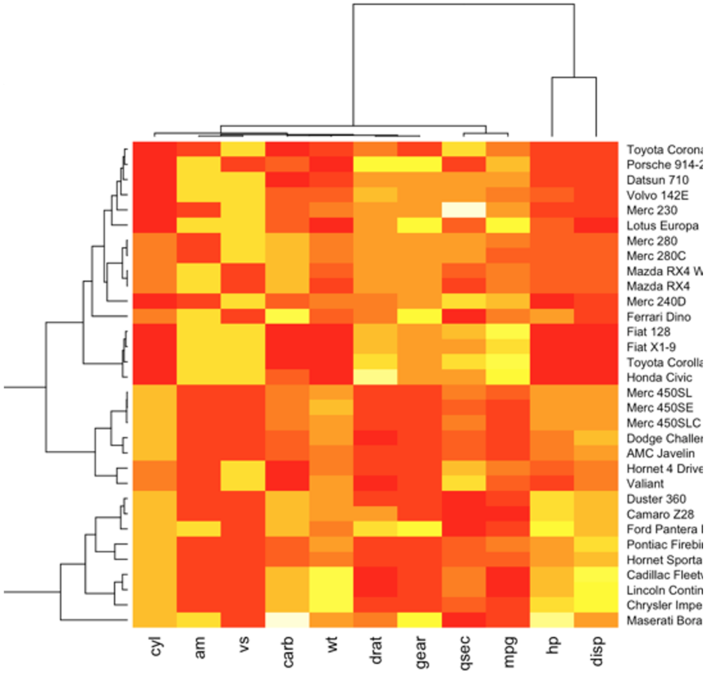
*With rare exception



Alternatives for 3-variable viz:
Ridgeline plots, heatmaps, or just facet

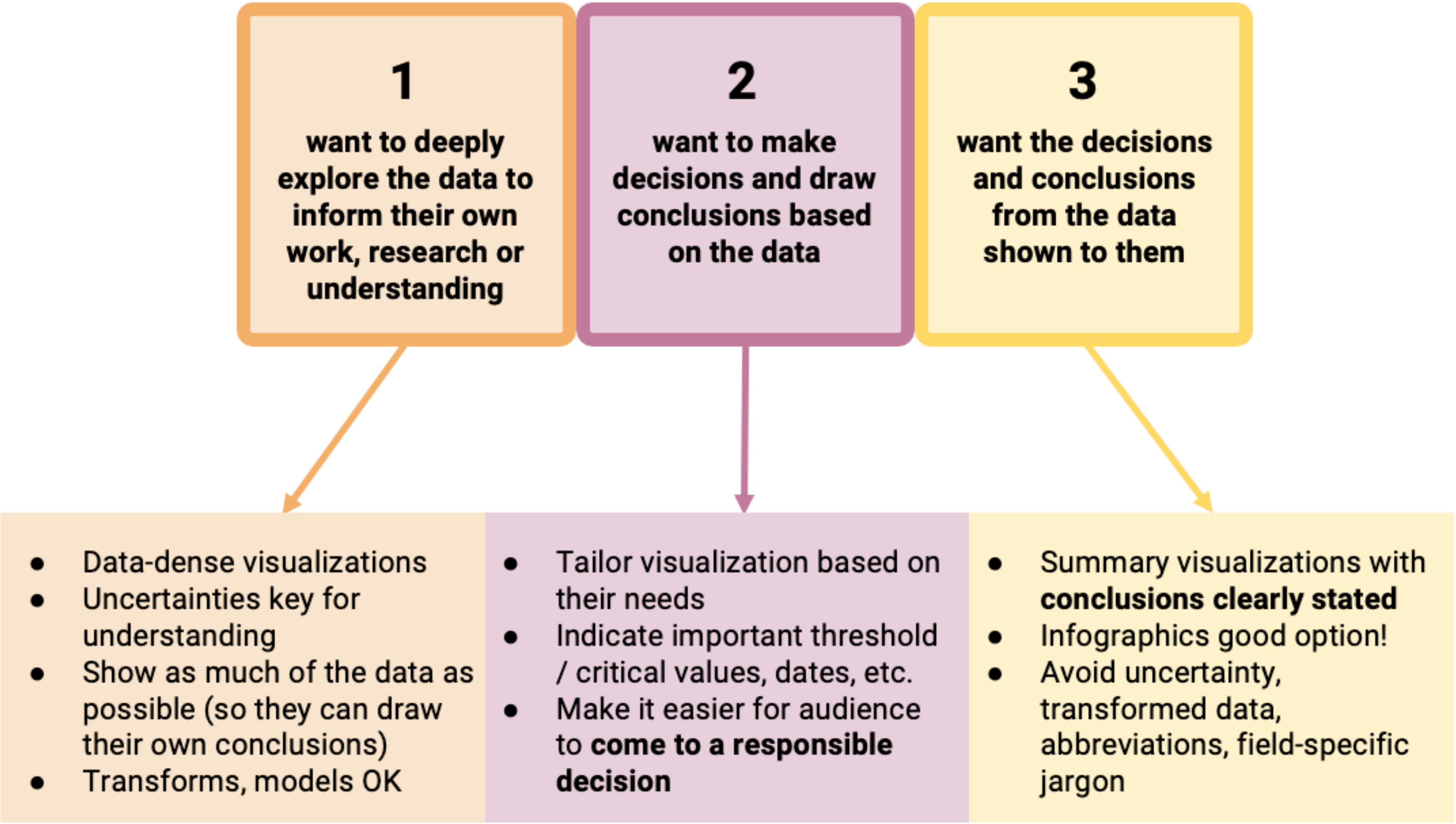


[Data to Viz]



[R Graph Gallery]

Audiences who...



1

want to deeply explore the data to inform their own work, research or understanding

- Data-dense visualizations
- Uncertainties key for understanding
- Show as much of the data as possible (so they can draw their own conclusions)
- Transforms, models OK

2

want to make decisions and draw conclusions based on the data

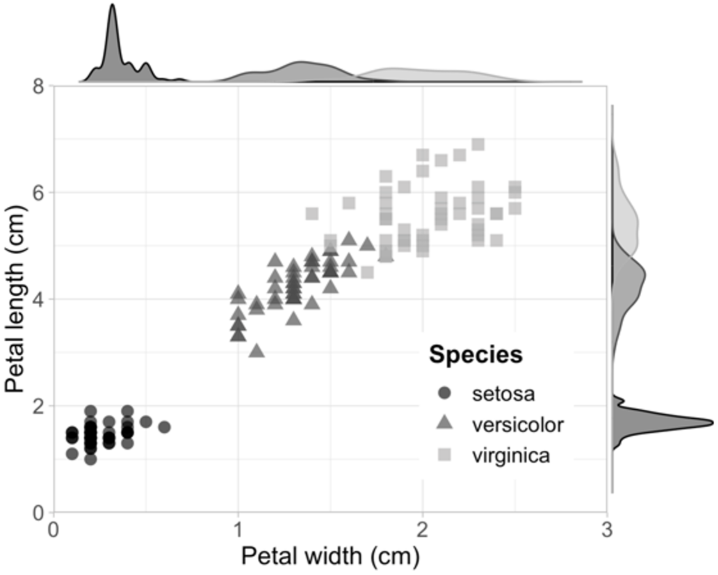
- Tailor visualization based on their needs
- Indicate important threshold / critical values, dates, etc.
- Make it easier for audience to **come to a responsible decision**

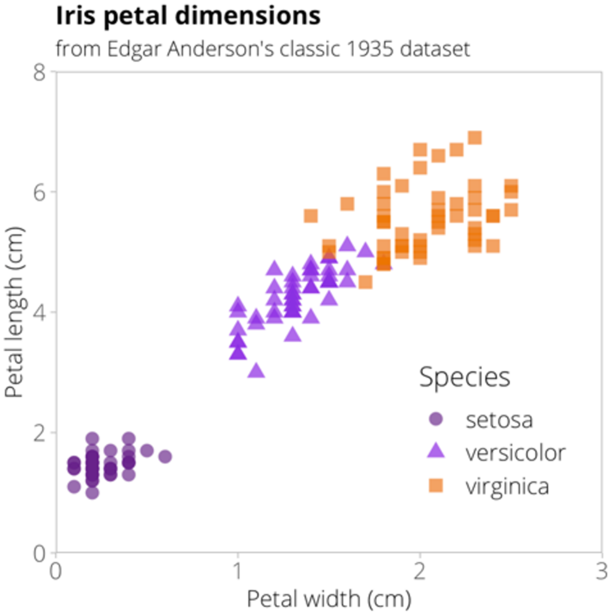
3

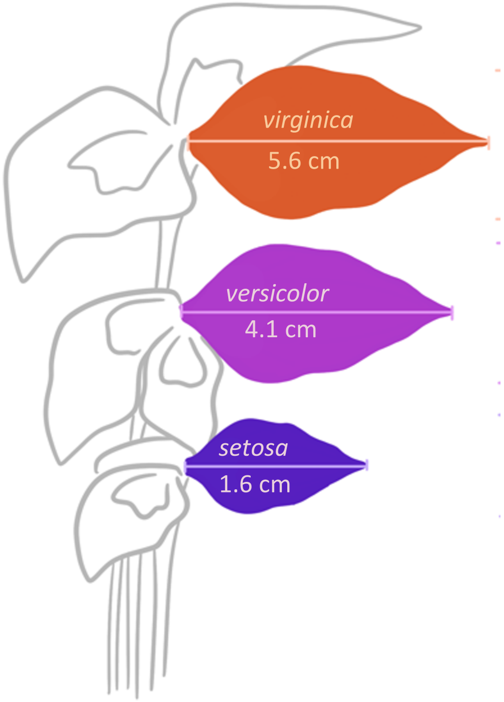
want the decisions and conclusions from the data shown to them

- Summary visualizations with **conclusions clearly stated**
- Infographics good option!
- Avoid uncertainty, transformed data, abbreviations, field-specific jargon

source Allison Horst



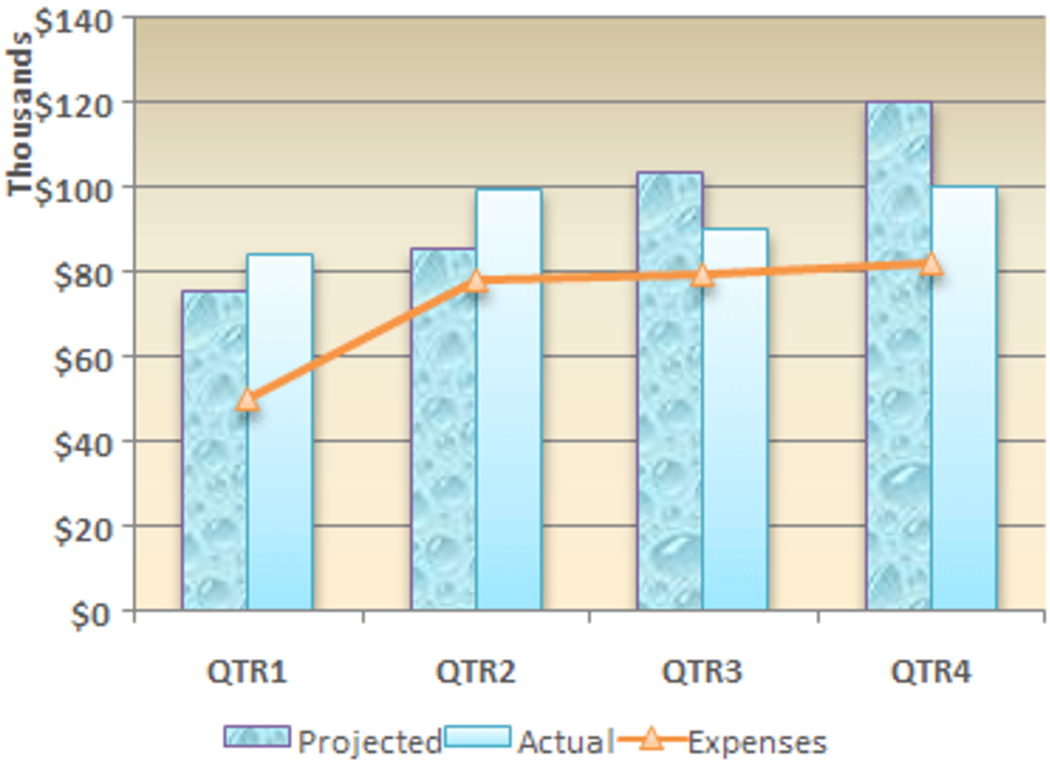




Artwork by Allison Horst

3
All about aesthetics

- Decluttering graphs
- Thoughtful color schemes
- Consistency matters
- Do the details



Avoid:

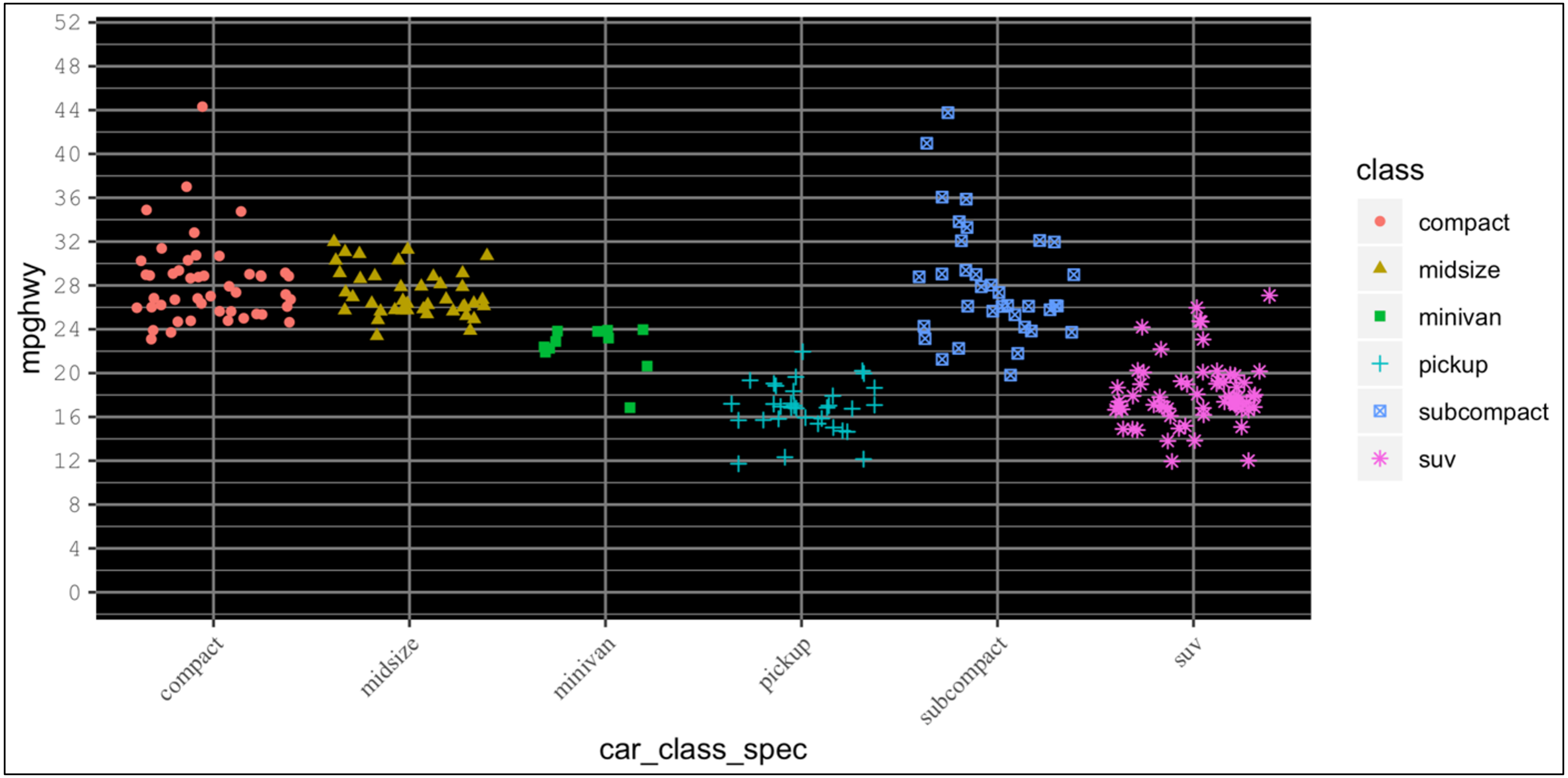
- Shadows
- Within-element gradients
- Basically any patterns
- Unnecessary symbols

From: *Change the shape fill, outline, or effects of chart elements.* Thanks, Office Support!

Things that are almost always bad:

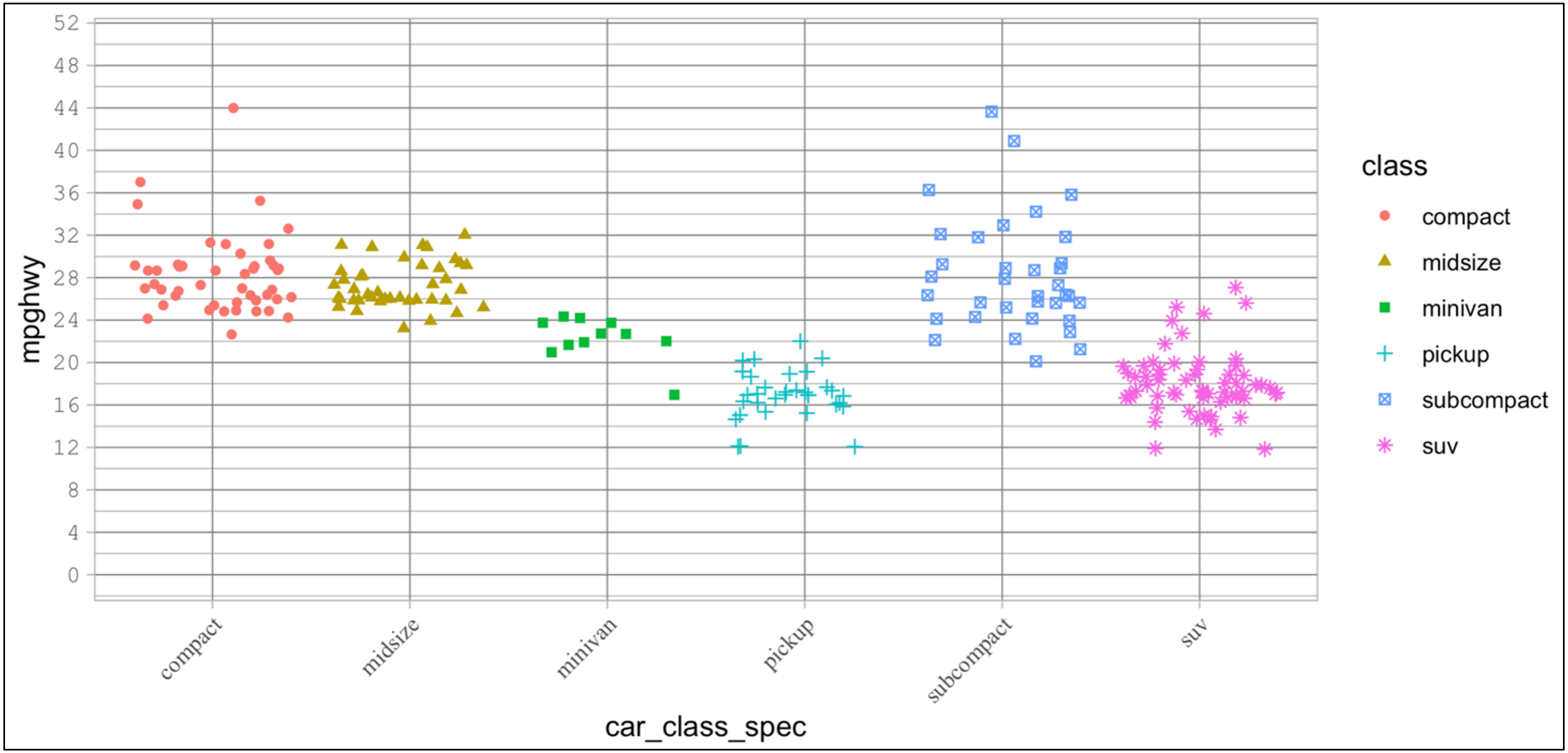
- Panel background colors (in general, but especially in viz for print)
- Outer borders around entire data viz area
- Angled text (besides 0° and 90°)
- Unnecessary / thoughtless color and or symbol and or line type use
- Excessive / unhelpful gridlines
- Far from 4:6, 3:5 or square aspect ratios
- Really creative fonts

Here is our starting point. In particular, we want to know how pickups compare for highway fuel efficiency (compared to other car types).



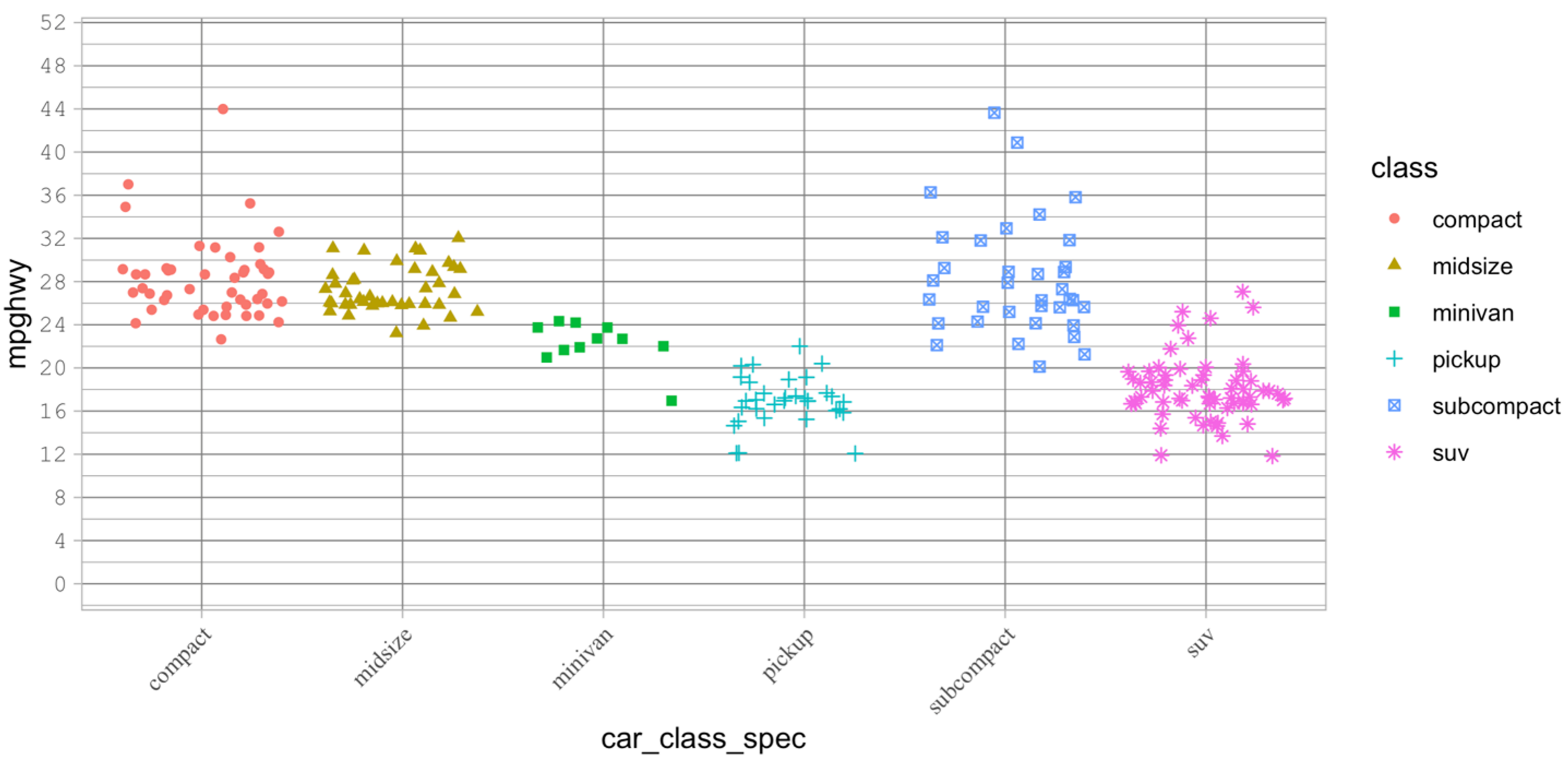
Things that are almost always bad:

- Panel background colors (in general, but especially in viz for print) ✓
- Outer borders around entire data viz area
- Angled text (besides 0° and 90°)
- Unnecessary / thoughtless color and or symbol and or line type use
- Excessive / unhelpful gridlines
- Far from 4:6, 3:5 or square aspect ratios
- Really creative fonts


mpghwy
52
48
44
40
36
32
28
24
20
16
12
8
4
0
compact
midsize
minivan
pickup
subcompact
suv
car_class_spec
class
● compact
▲ midsize
■ minivan
+ pickup
☒ subcompact
✳ suv

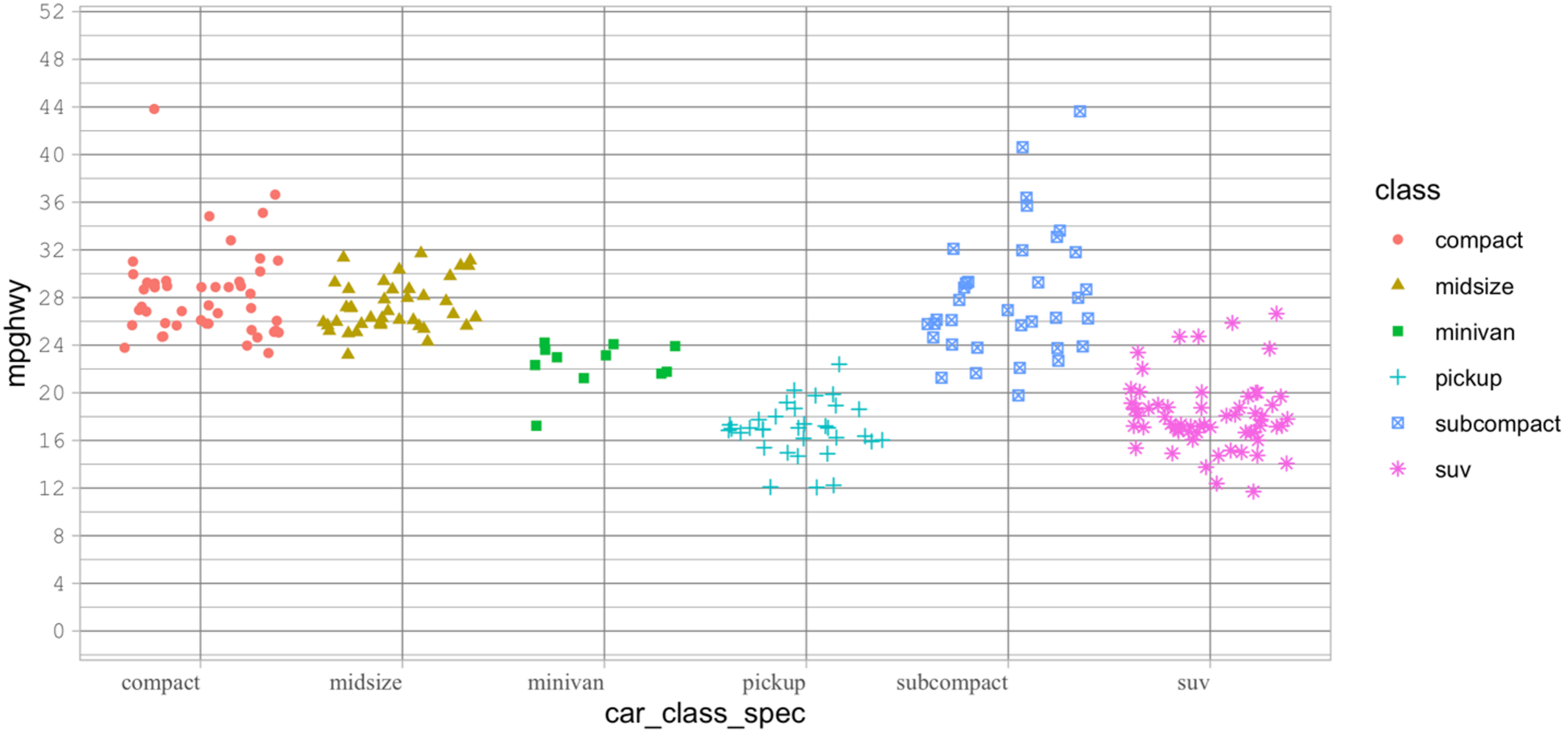
Things that are almost always bad:

- Panel background colors (in general, but especially in viz for print) ✓
- Outer borders around entire data viz area ✓
- Angled text (besides 0° and 90°)
- Unnecessary / thoughtless color and or symbol and or line type use
- Excessive / unhelpful gridlines
- Far from 4:6, 3:5 or square aspect ratios
- Really creative fonts


52
48
44
40
36
32
28
24
20
16
12
8
4
0
mpghwy
compact
midsize
minivan
pickup
subcompact
suv
car_class_spec
class
● compact
▲ midsize
■ minivan
+ pickup
☒ subcompact
✳ suv

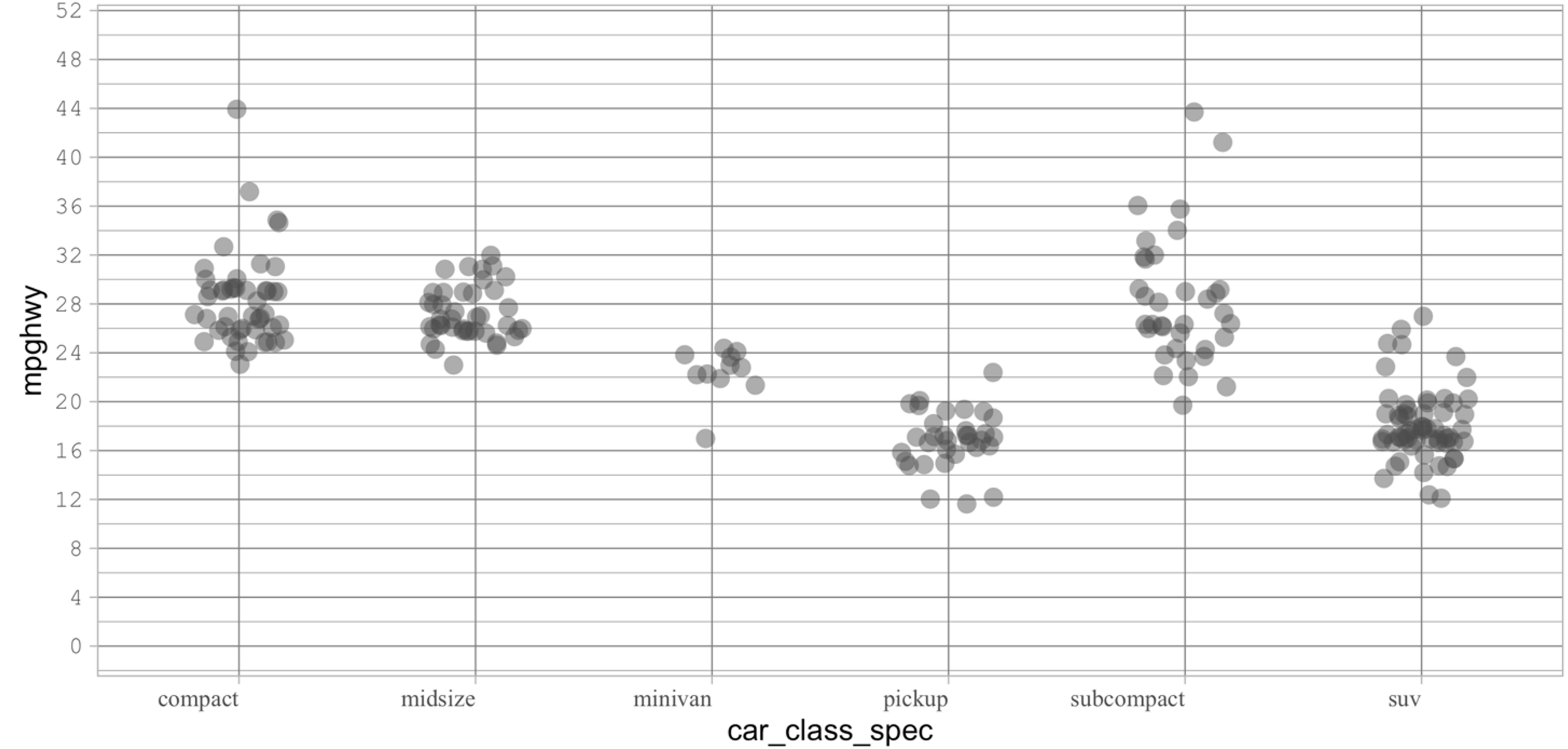
Things that are almost always bad:

- Panel background colors (in general, but especially in viz for print) ✓
- Outer borders around entire data viz area ✓
- Angled text (besides 0° and 90°) ✓
- Unnecessary / thoughtless color and or symbol and or line type use
- Excessive / unhelpful gridlines
- Far from 4:6, 3:5 or square aspect ratios
- Really creative fonts


52
48
44
40
36
32
28
24
20
16
12
8
4
0
mpghwy
compact
midsize
minivan
pickup
subcompact
suv
car_class_spec
class
compact
midsize
minivan
pickup
subcompact
suv

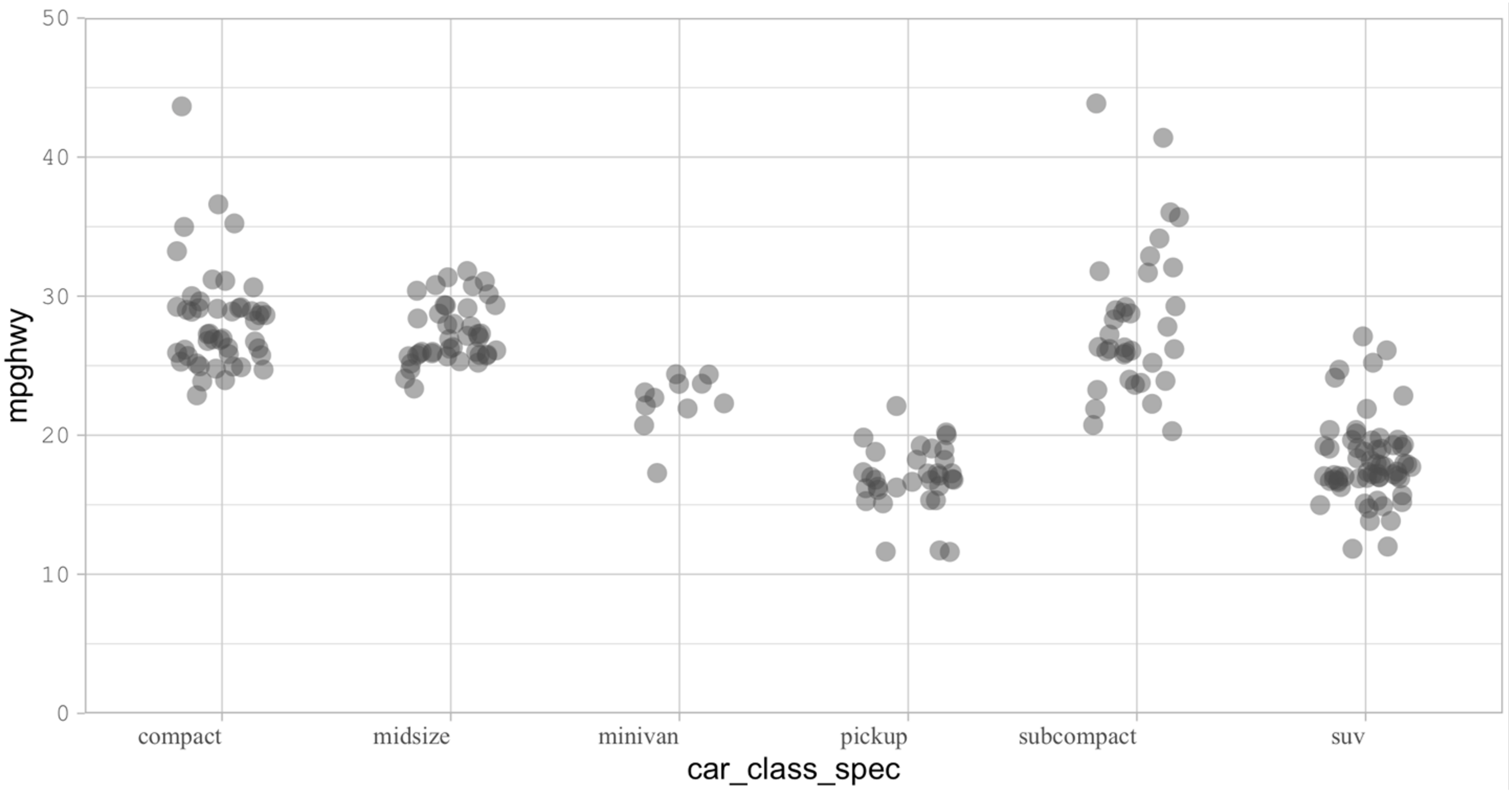
Things that are almost always bad:

- Panel background colors (in general, but especially in viz for print) ✓
- Outer borders around entire data viz area ✓
- Angled text (besides 0° and 90°) ✓
- Unnecessary / thoughtless color and or symbol and or line type use ✓
- Excessive / unhelpful gridlines
- Far from 4:6, 3:5 or square aspect ratios
- Really creative fonts


52
48
44
40
36
32
28
24
20
16
12
8
4
0
mpghwy
compact
midsize
minivan
pickup
subcompact
suv
car_class_spec

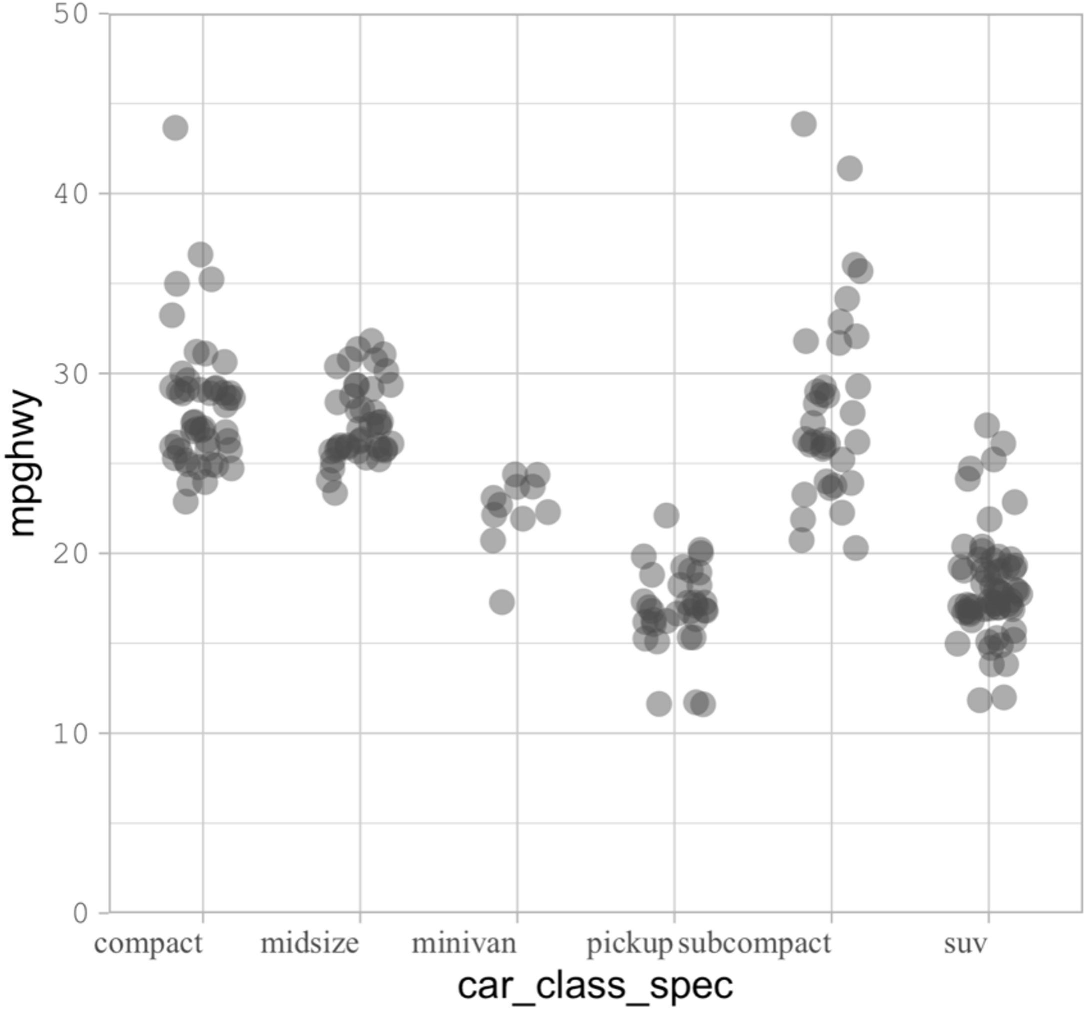
Things that are almost always bad:

- Panel background colors (in general, but especially in viz for print) ✓
- Outer borders around entire data viz area ✓
- Angled text (besides 0° and 90°) ✓
- Unnecessary / thoughtless color and or symbol and or line type use ✓
- Excessive / unhelpful gridlines ✓
- Far from 4:6, 3:5 or square aspect ratios
- Really creative fonts


50
40
30
20
10
0
mpghwy
compact
midsize
minivan
pickup
subcompact
suv
car_class_spec

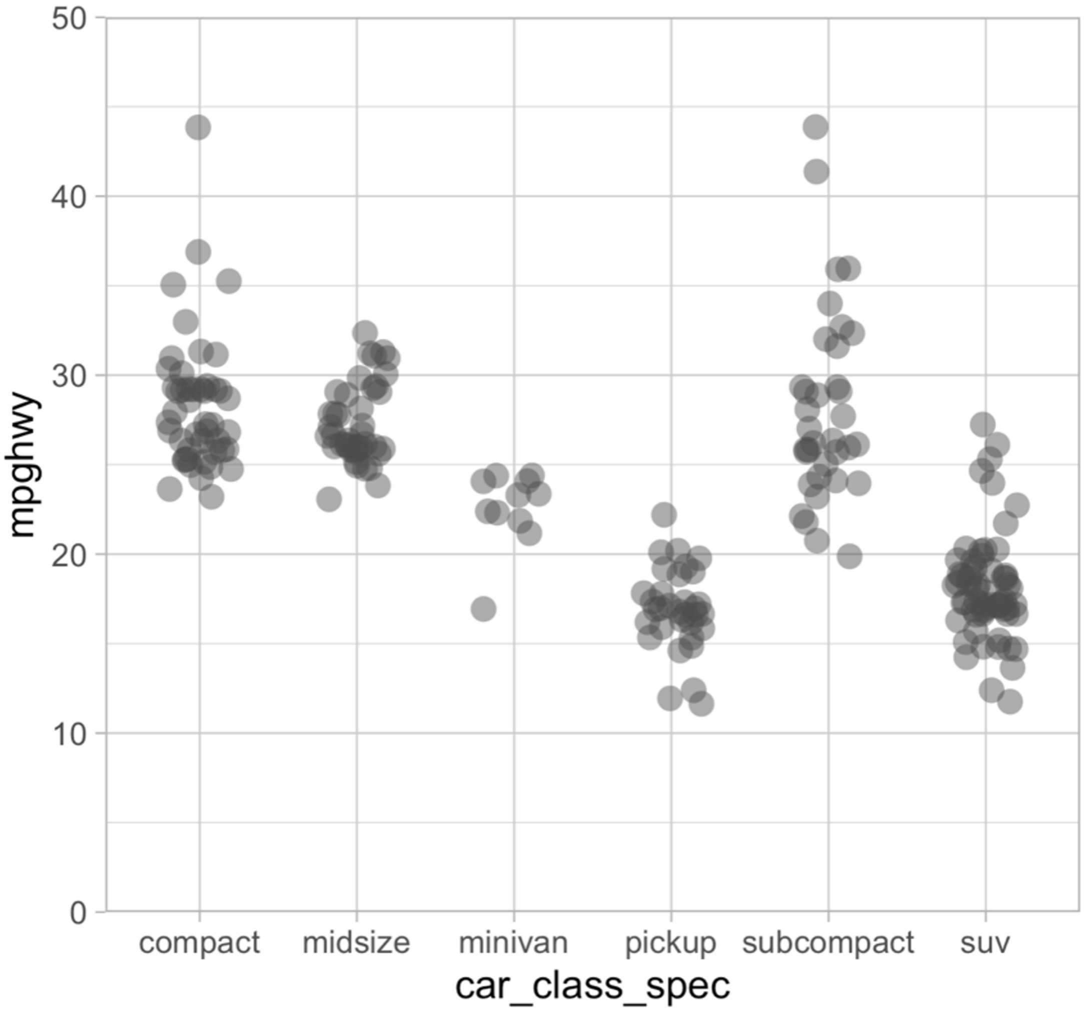
Things that are almost always bad:

- Panel background colors (in general, but especially in viz for print) ✓
- Outer borders around entire data viz area ✓
- Angled text (besides 0° and 90°) ✓
- Unnecessary / thoughtless color and or symbol and or line type use ✓
- Excessive / unhelpful gridlines ✓
- Far from 4:6, 3:5 or square aspect ratios ✓
- Really creative fonts

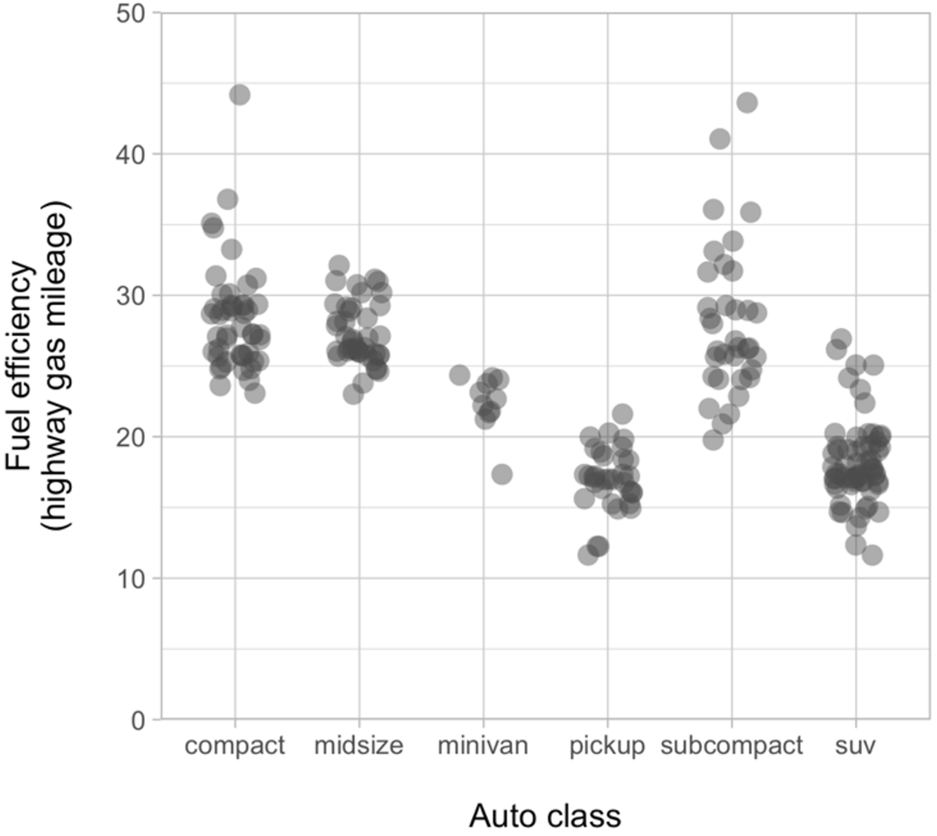

50
40
30
20
10
0
mpghwy
compact
midsize
minivan
pickup
subcompact
suv
car_class_spec

Things that are almost always bad:

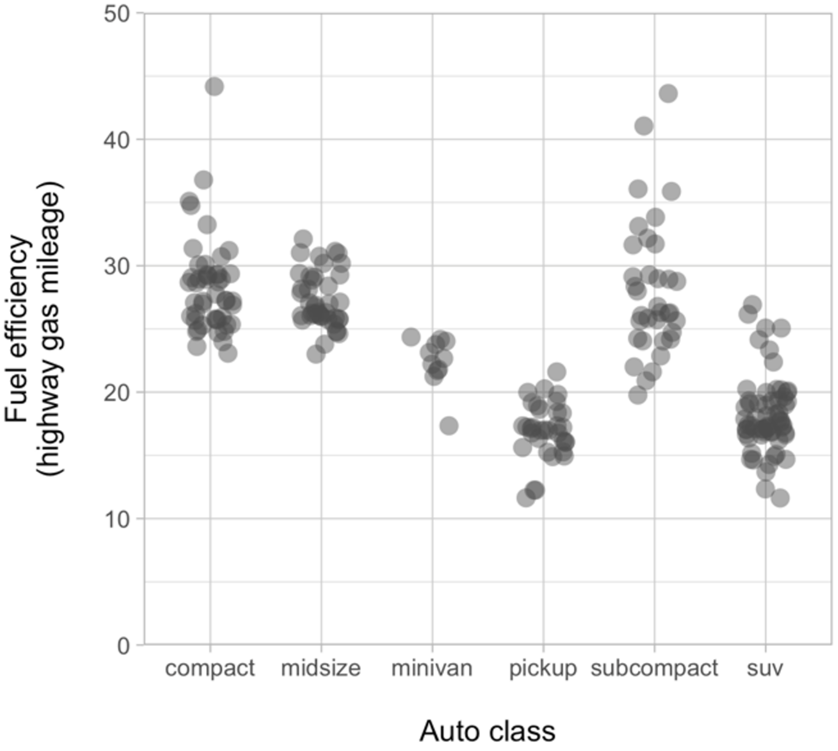
- Panel background colors (in general, but especially in viz for print) ✓
- Outer borders around entire data viz area ✓
- Angled text (besides 0° and 90°) ✓
- Unnecessary / thoughtless color and or symbol and or line type use ✓
- Excessive / unhelpful gridlines ✓
- Far from 4:6, 3:5 or square aspect ratios ✓
- Really creative fonts ✓

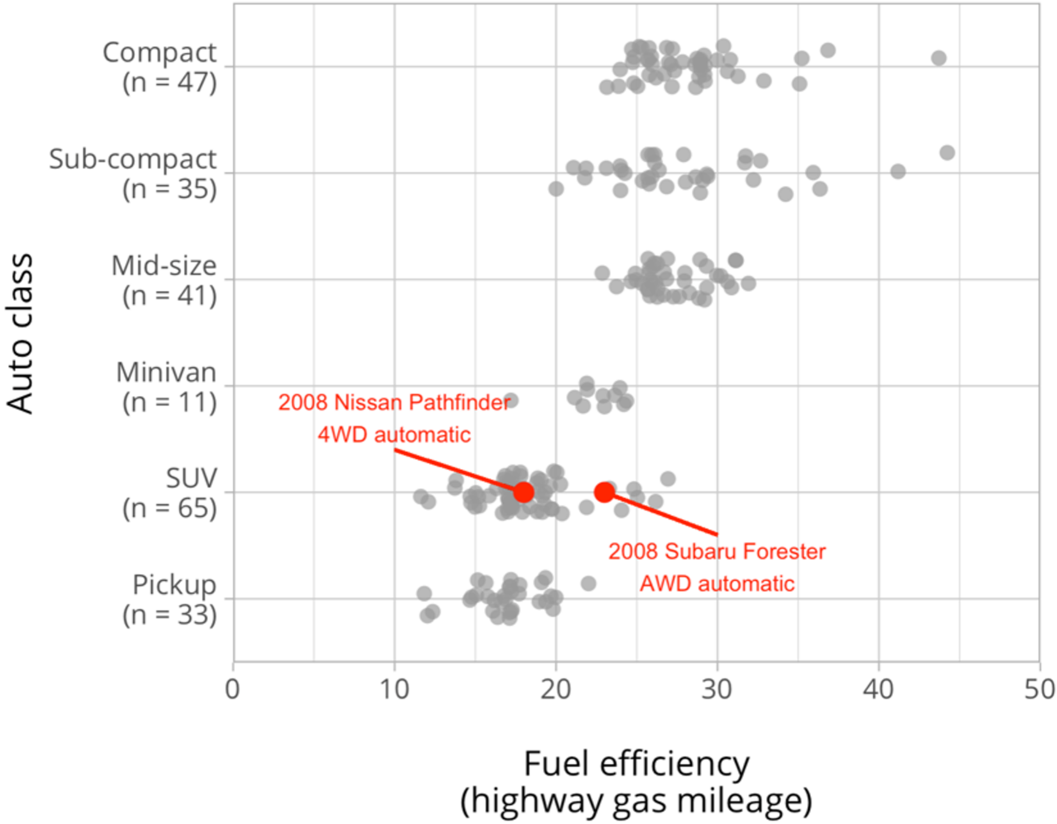

50
40
30
20
10
0
mpghwy
compact
midsize
minivan
pickup
subcompact
suv
car_class_spec

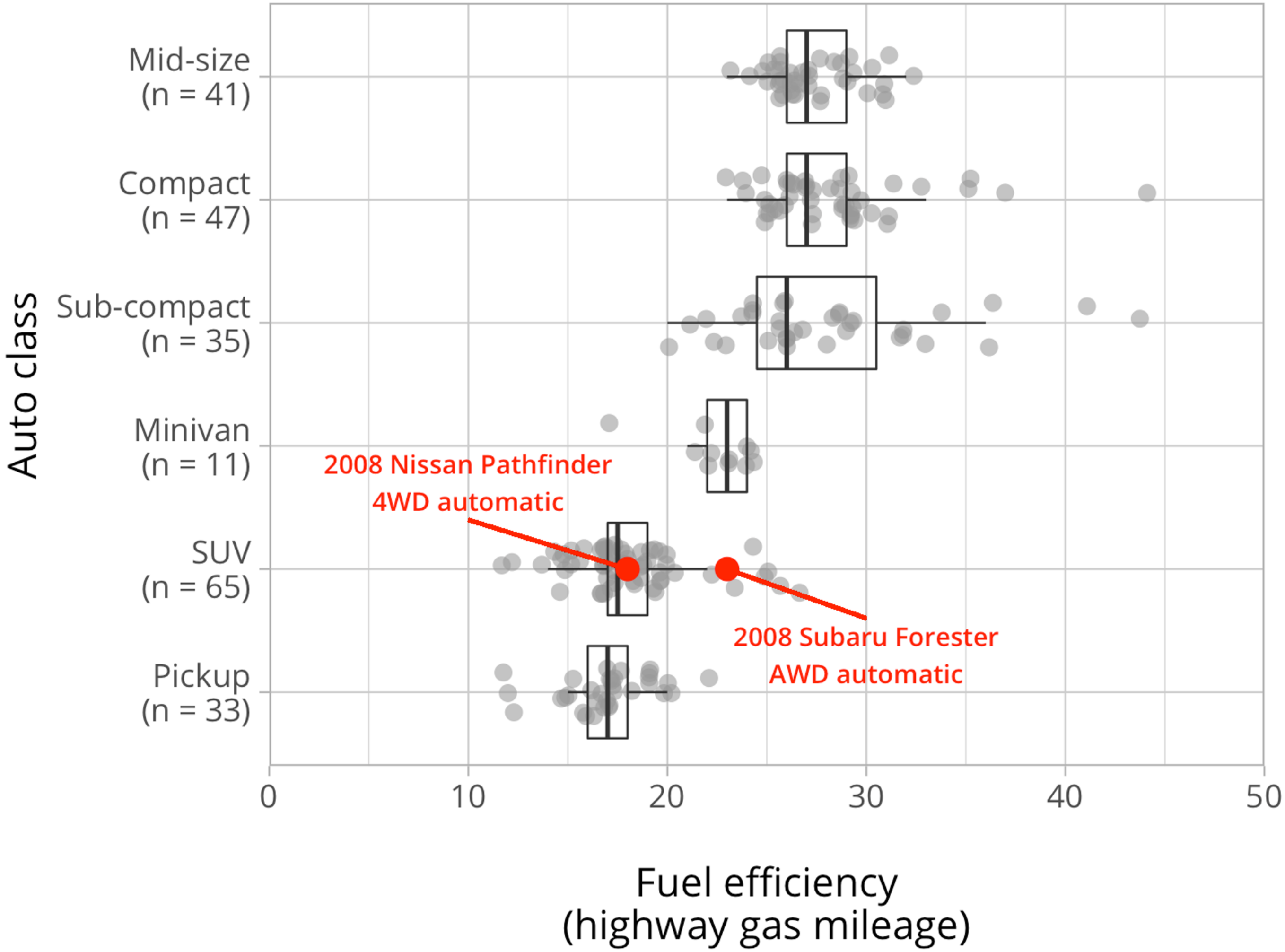
Update axis labels, then we'll call this neutral.
How can we make it good?



- Put in a meaningful order
- Flip coordinates
- Finalize axis tick mark labels
- Highlight car of interest, and add label / annotation
- Pick a single, professional font






Mid-size
(n = 41)
Compact
(n = 47)
Sub-compact
(n = 35)
Minivan
(n = 11)
SUV
(n = 65)
Pickup
(n = 33)
Auto class
2008 Nissan Pathfinder
4WD automatic
2008 Subaru Forester
AWD automatic
0
10
20
30
40
50
Fuel efficiency
(highway gas mileage)

Choose thoughtful color schemes

- Colors should help an audience member learn / retain something from the data visualization by clarifying groups, values or foci

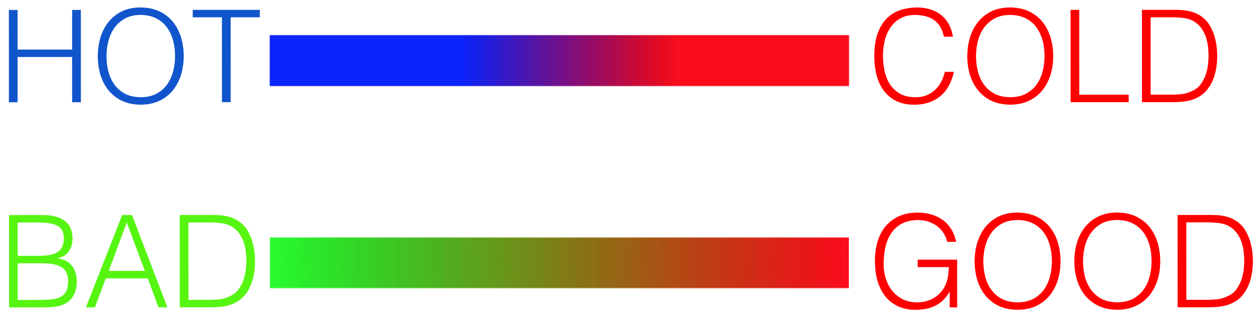
- If it also looks awesome and clarifies the data, fantastic!

- If it looks awesome but reduces clarity of the data, stop!

We naturally think about some things in color!

The colors we choose for our data viz should reflect those.

Word / color mismatch anxiety:

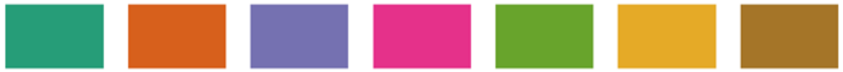


HOT COLD

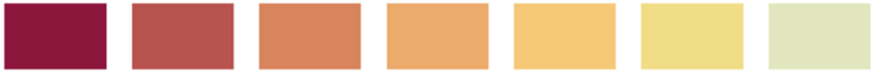
BAD GOOD

More interesting word / color associations & psychology:
https://zevendesign.com/color-association/

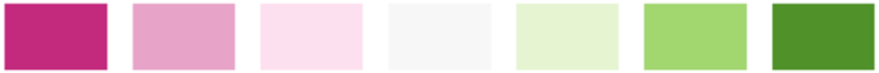
Qualitative scales: to distinguish between groups



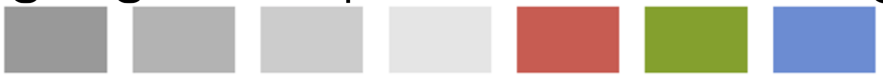
Sequential scales: to indicate values or value order



Diverging scales: when there's an obvious "mid" point, and you want to show how much higher or lower things are from it

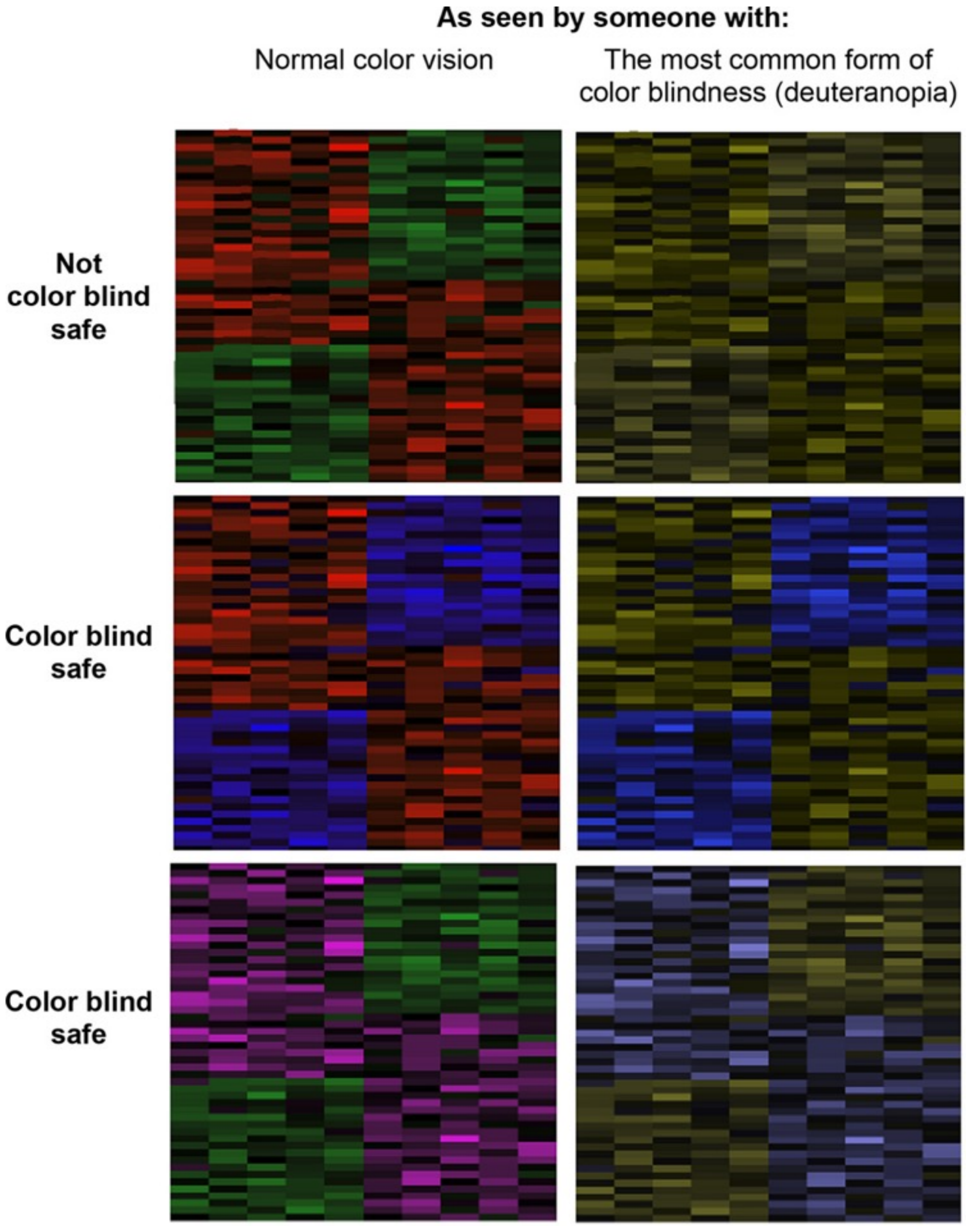


Highlights: to point out something of interest



Example palettes from *Fundamentals of Data Visualization* by Claus O. Wilke

Select color blind safe colors



Some tools & good options for color:

- Free (mac, PC) color checker: http://www.colororacle.org/

- {viridis} package (by Simon Garnier) “provides color palettes to make beautiful plots that are: printer-friendly, perceptually uniform and easy to read by those with colorblindness.” - Datanovia, Top R Color Palettes

- {RColorBrewer} package, to check for colorblind friendly palettes run:

```
display.brewer.all(colorblindFriendly = TRUE)
```

Consistency

Every time you change something stylistically, you ask the audience to adjust to something new. That means more effort on their end.

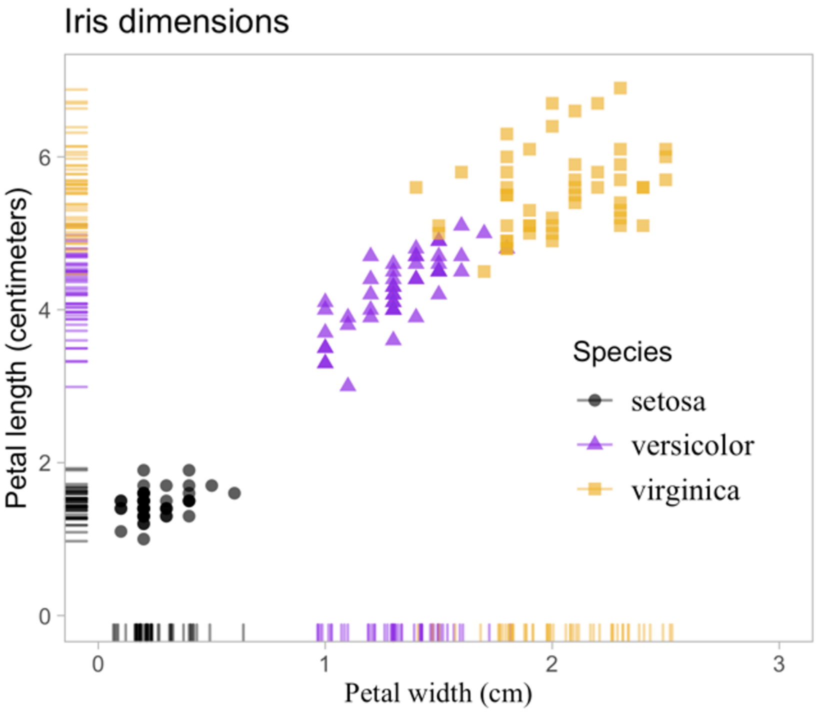
So be hyper consistent…

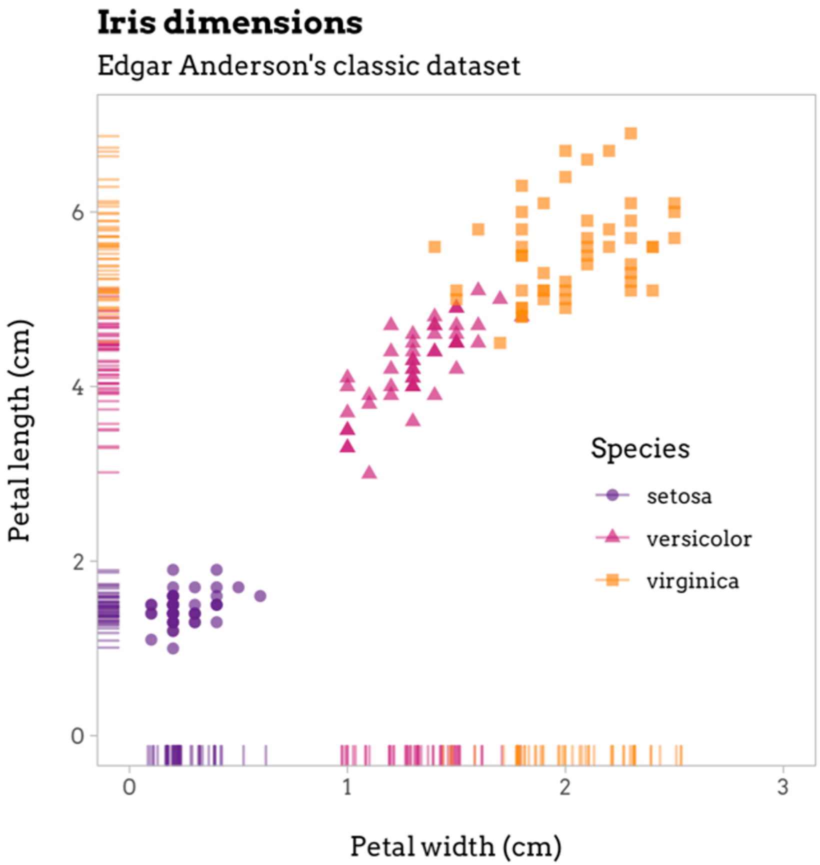
- Within single data visualizations
- Across multiple data visualizations
- Between report / presentation styles and data visualizations

That includes: Fonts, color schemes, themes, point styles, shapes / aspect ratios, overall formats (titles? captions?), and beyond.

Consistency within graphs

Max 2 fonts (good option: same font, update face, spacing, etc.)

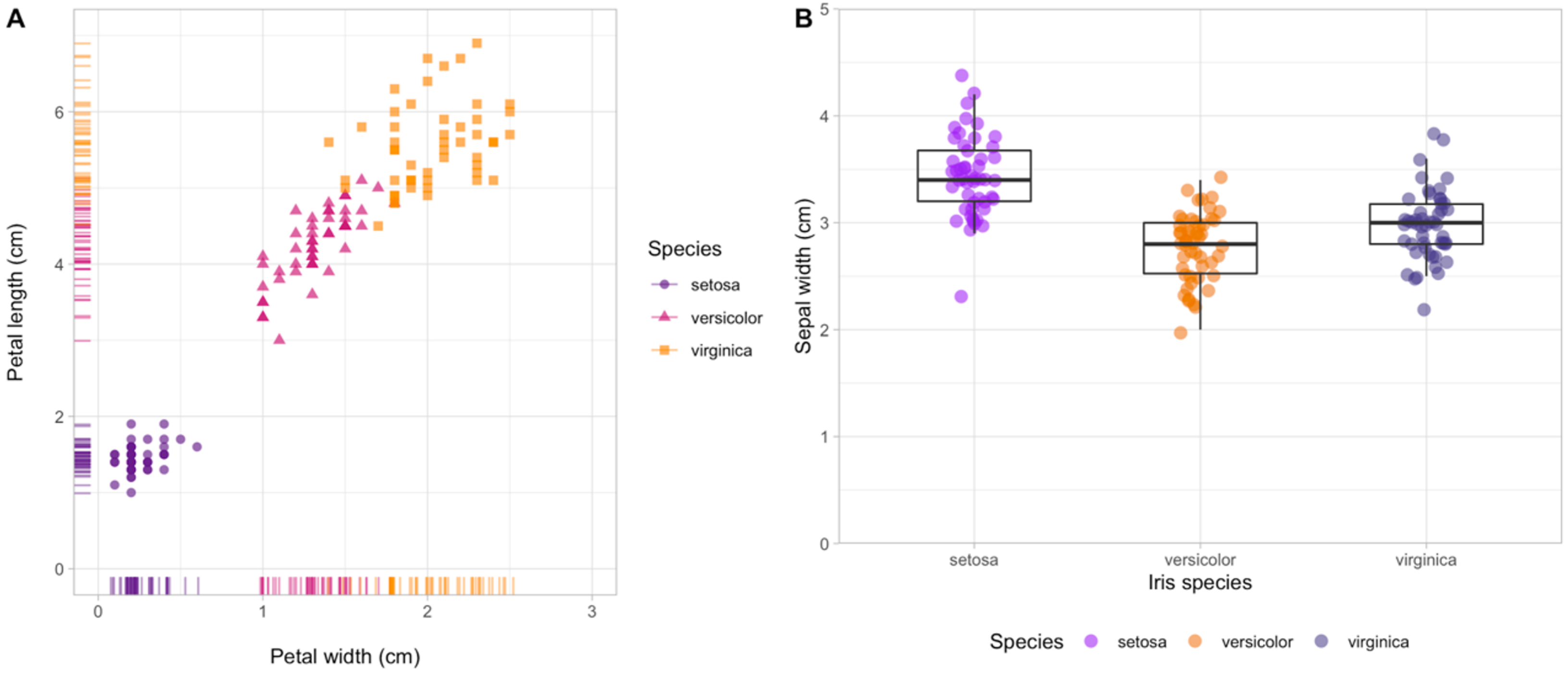




Consistency across (especially for compound figures)

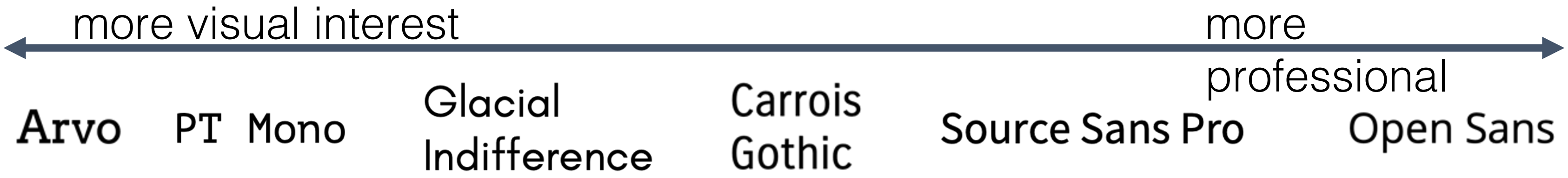
Including: color scheme, point styles, order, spacing & alignment, fonts, etc.

Yikes:



Do the details

- Use superscripts / subscripts and correct symbols
  - 3.4E+4 vs. 3.4×10^4
  - km^2 vs. km^2
- Use symbols (don’t cut corners - it’s worth the effort!)
  - Deg C
  - °C
- Be thoughtful about significant figures
- RESOLUTION MATTERS
- Spend some time with fonts (and ASK / READ / LEARN)
  - Some of my favorites (this changes):


more visual interest

more professional

Arvo PT Mono Glacial Indifference Carrois Gothic Source Sans Pro Open Sans

Resources to keep learning about data visualization

Open / free books & websites:

- Wilke, Claus O. Fundamentals of Data Visualization
- The Data Visualization Society
- The R-Graph-Gallery
- Data-to-Vizz
- The Data Visualization Catalogue
- Information is Beautiful

Other books & resources:

- Healy, Kieran Data Visualization: A Practical Guide
- Edward Tufte's books on Data Visualization
- Alberto Cairo How Charts Lie and The Truthful Art

Follow on twitter:

- @nadiahbremer
- @DataVizSociety
- @AlbertoCairo
- @alyssafowers
- @dataviz_catalogue
- @sdbernard
- @Elijah_Meeks
- @kjhealy

Inspiration and slides for this talk
Thanks!

Open content & slides:

- Allison Horst. https://www.allisonhorst.com/talk/sccwrp_dataviz_2019/
- Jessica Minnier · Meike Niederhausen. bit.ly/berd_ggplot
- Angela Zoss · Eric Monson. http://bit.ly/STA112FSVisFall2017

Thank you!

Artwork by Allison Horst