THINKING ABOUT DATA VISUALIZATION

DANIEL KERCHNER, SENIOR SOFTWARE DEVELOPER
EMILY BLUMENTHAL, DATA SERVICES LIBRARIAN
GW LIBRARIES AND ACADEMIC INNOVATION

YOUR EXPERIENCES

Share your data visualization experiences and workshop goals

1 What goals are you trying to achieve with your data visualizations?

2 What challenges have you faced while working with data visualizations?

3 What do you want to learn to do with respect to data visualizations?

4 What additional questions do you have about data visualizations?

WORKSHOP OUTLINE

Our content today is divided into four parts. Each part will be described with examples.

01 Data Visualization Process

Walking through the visualization process, from setting goals to visualizing the data.

02 Basic Design Considerations

Best practices and accessibility considerations in data visualization.

03 Solutions to Common Problems

How to resolve, or even better, avoid common problems in data visualization.

04 Visualization Resources

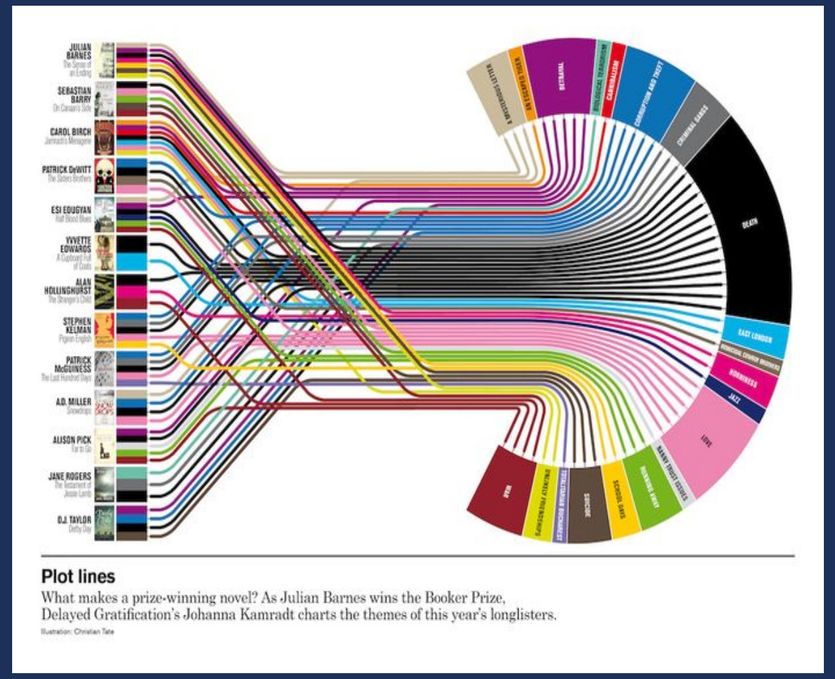
Workshops, tools, and other resources for data visualization.

WHAT'S WRONG WITH THESE VISUALIZATIONS?

Let's evaluate some data visualizations.

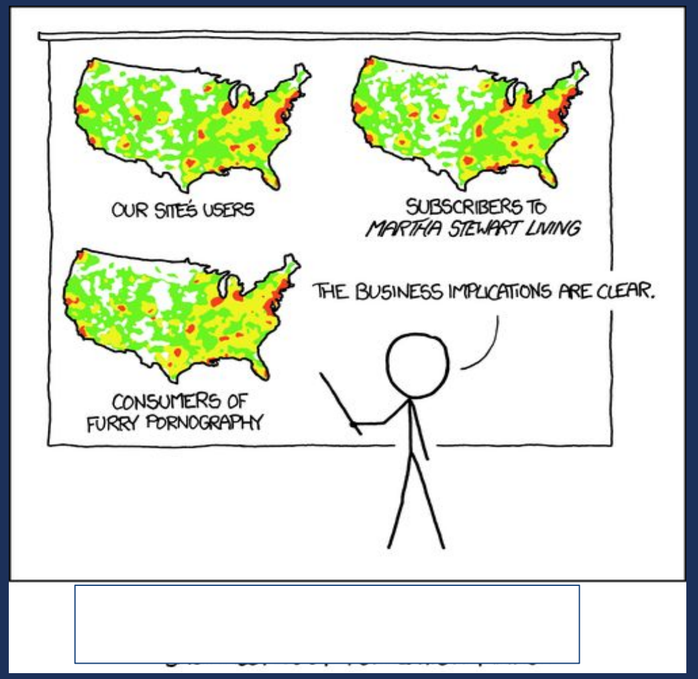
INCORRECT SCALE

https://twitter.com/indiainpixels/status/1657102345396682753



TOO BUSY

https://www.slow-journalism.com/infographics/booker-prize-the-infographics



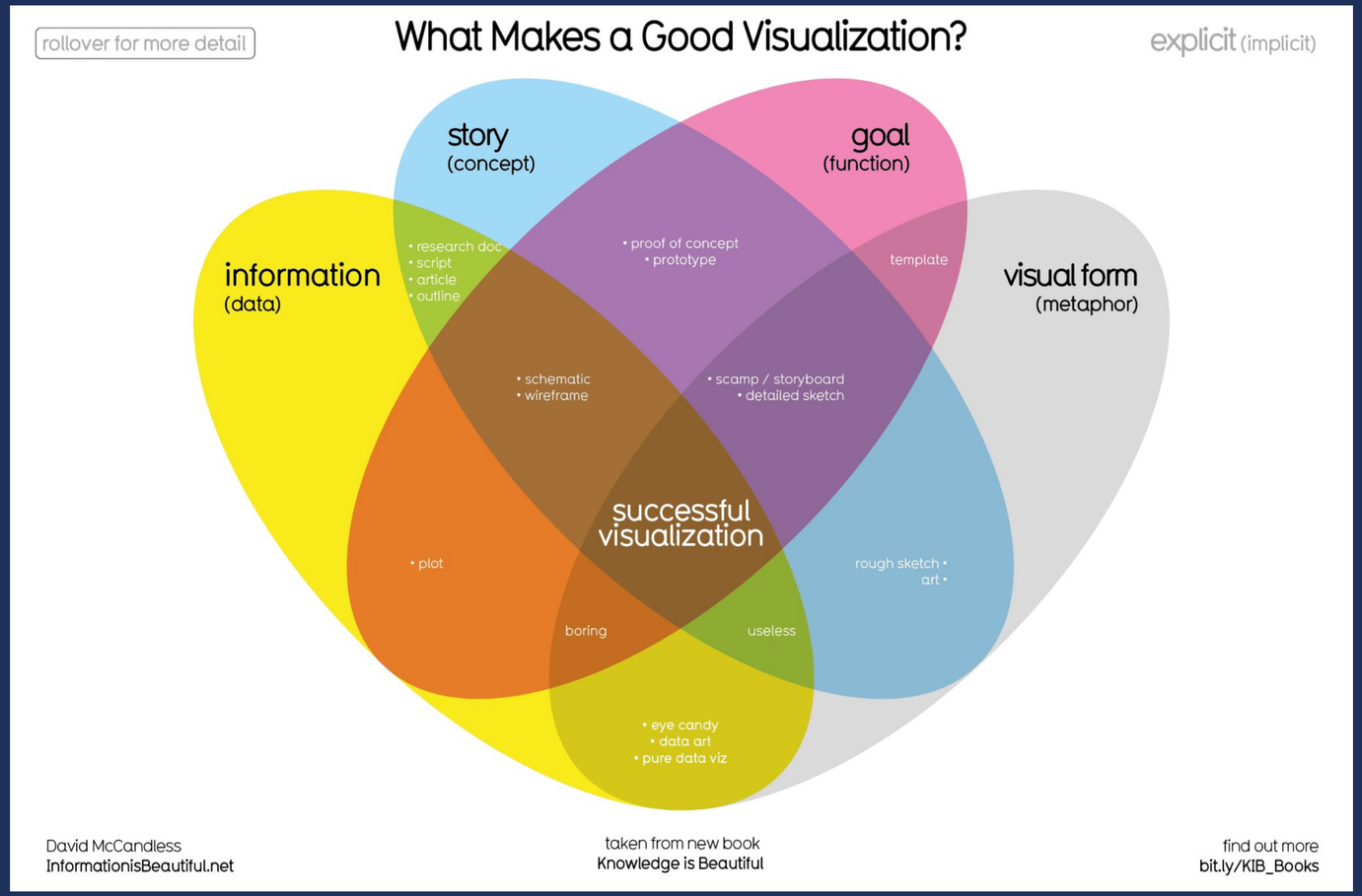
WRONG ANALYSIS

https://xkcd.com/1138

WHAT MAKES A GOOD VISUALIZATION?

A good visualization:

- has clear purpose and focus
- contains only useful and accurate information
- is structured correctly
- is formatted accessibly



WORKSHOP OUTLINE

Our content today is divided into four parts. Each part will be described with examples.

01 Data Visualization Process

Walking through the visualization process, from setting goals to visualizing the data.

02 Basic Design Considerations

Best practices and accessibility considerations in data visualization.

03 Solutions to Common Problems

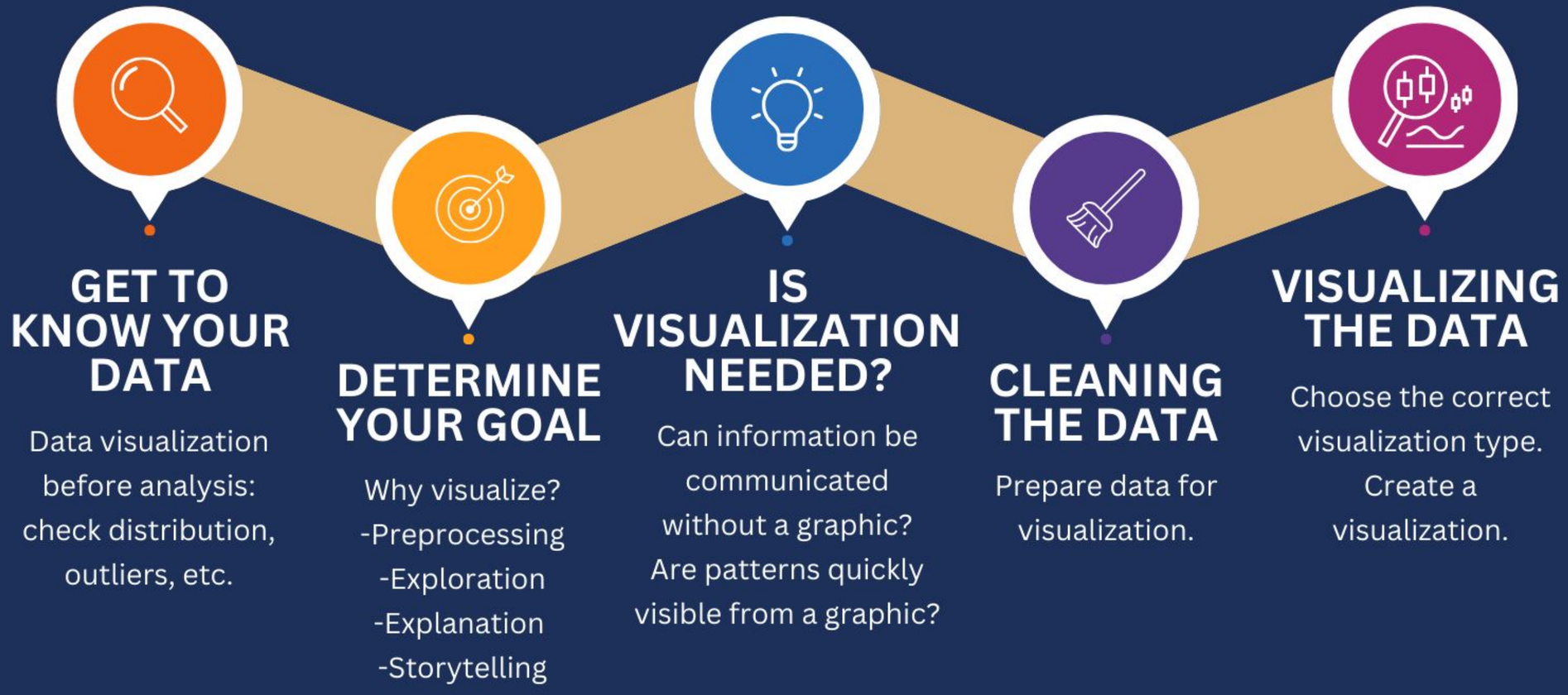
How to resolve, or even better, avoid common problems in data visualization.

04 Visualization Resources

Workshops, tools, and other resources for data visualization.

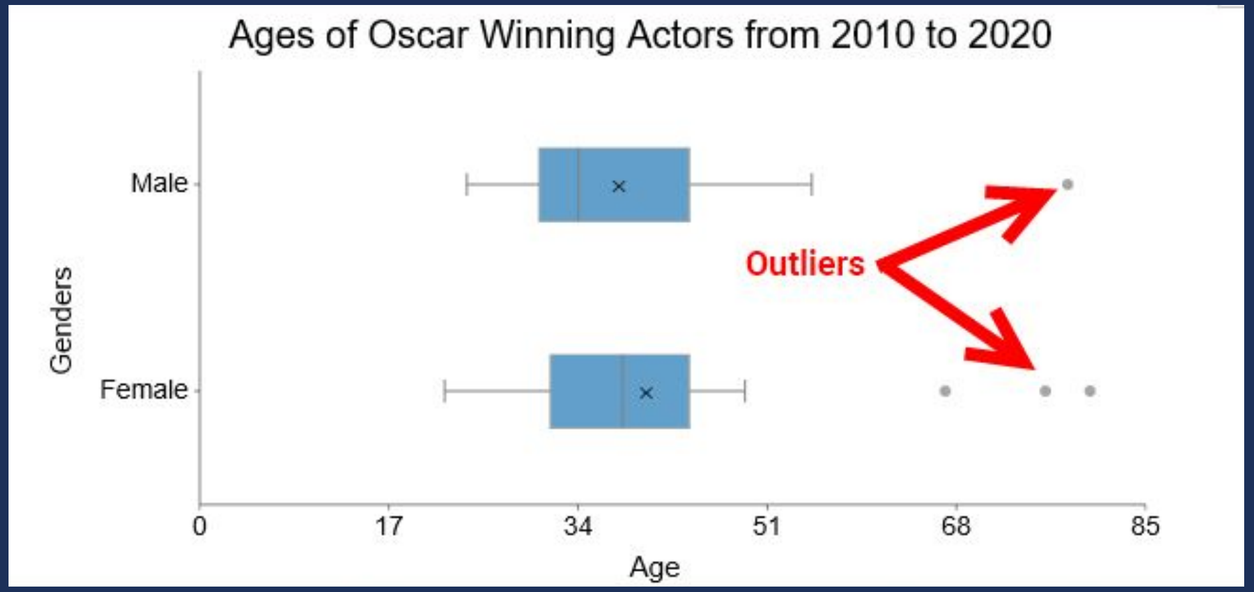
DATA VISUALIZATION PROCESS

Data visualizations from start to finish



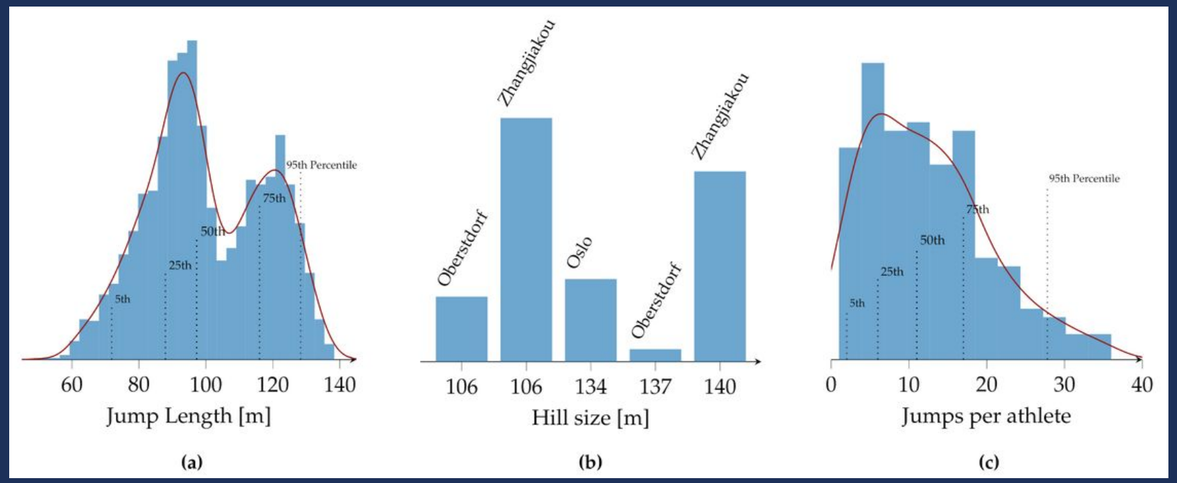
Visualization Goal

GETTING TO KNOW YOUR DATA



https://chartexpo.com/blog/box-plot-outliers#

Identifying Outliers



Checking the Distribution

Link J, Schwinn L, Pulsmeyer F, Kautz T, Eskofier BM. xLength: Predicting Expected Ski Jump Length Shortly after Take-Off Using Deep Learning. Sensors. 2022; 22(21):8474. https://doi.org/10.3390/s22218474

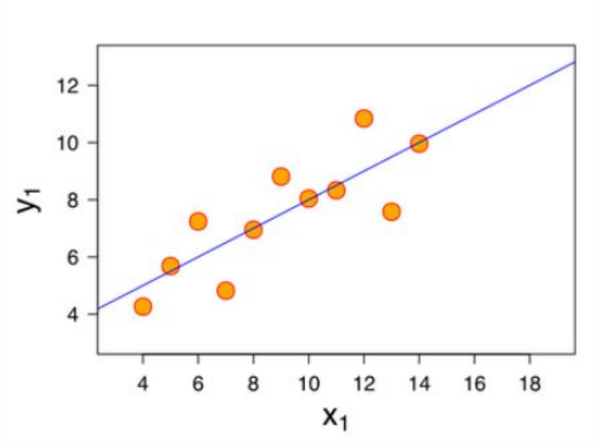
Visualization Goal

GETTING TO KNOW YOUR DATA

I	
x	*y*
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places



- Explore relationships between variables
  - Visualize trends or patterns
    - **Anscombe's Quartet**

Visualization Goal

GETTING TO KNOW YOUR DATA

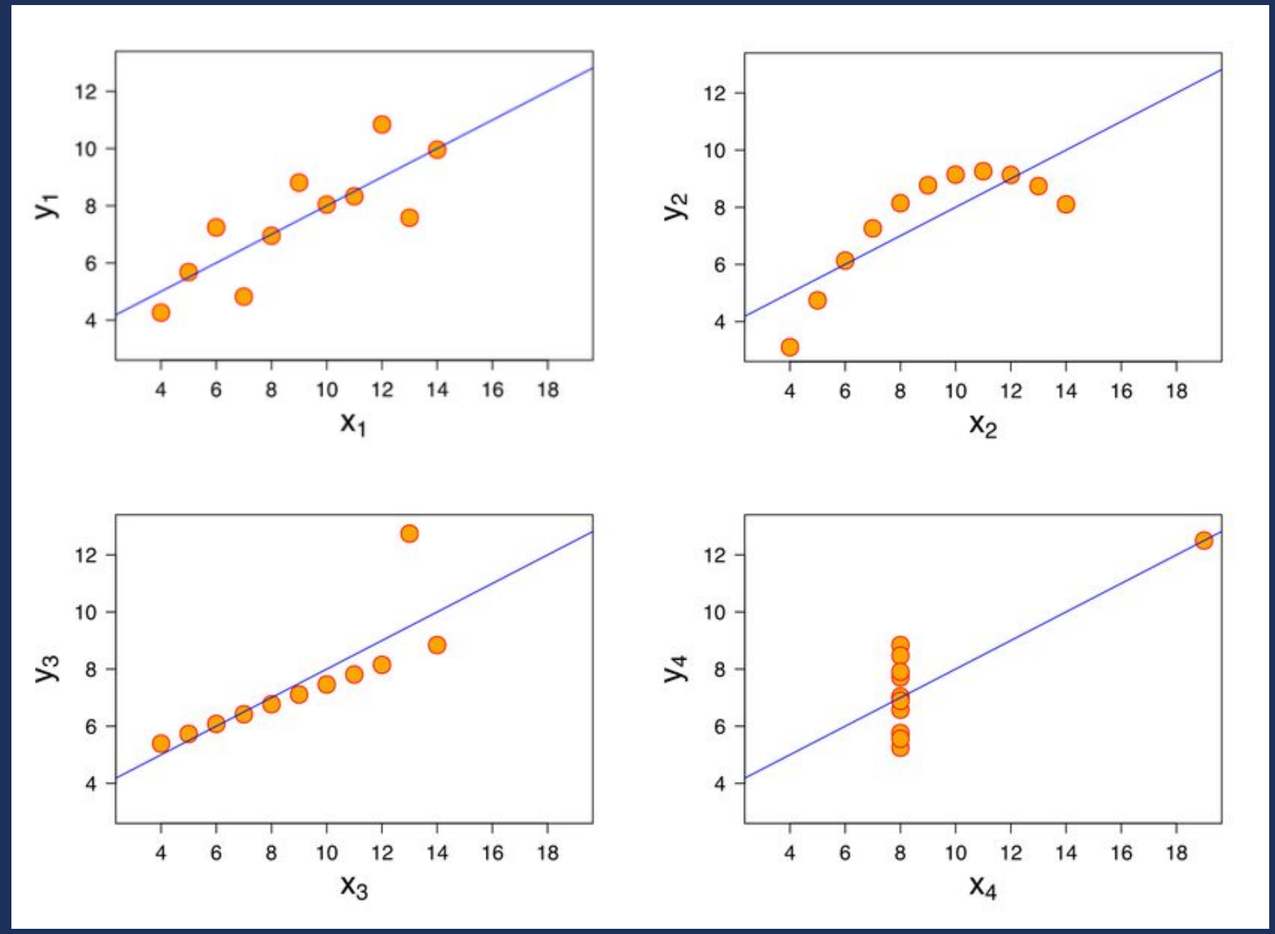
Anscombe's quartet

I		II		III		IV	
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

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

For all four datasets:

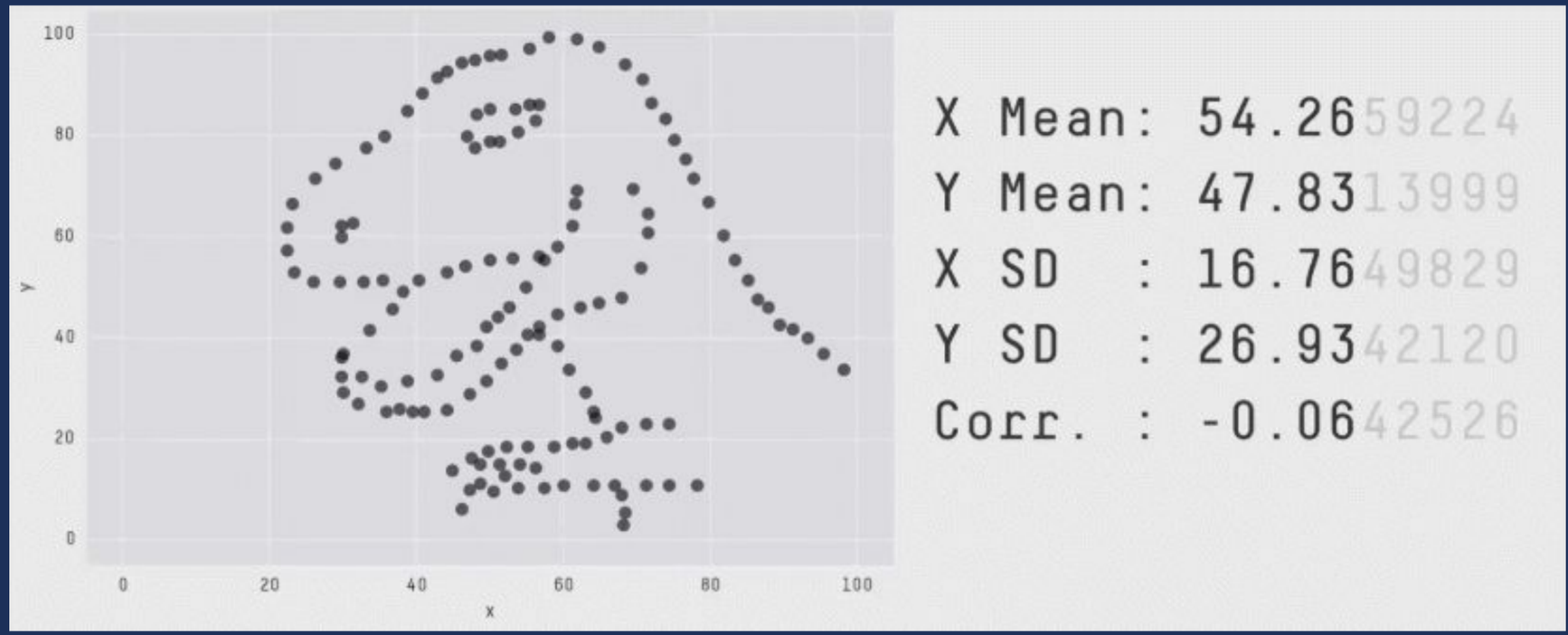
Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*: s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*: s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places



Visualization Goal

GETTING TO KNOW YOUR DATA

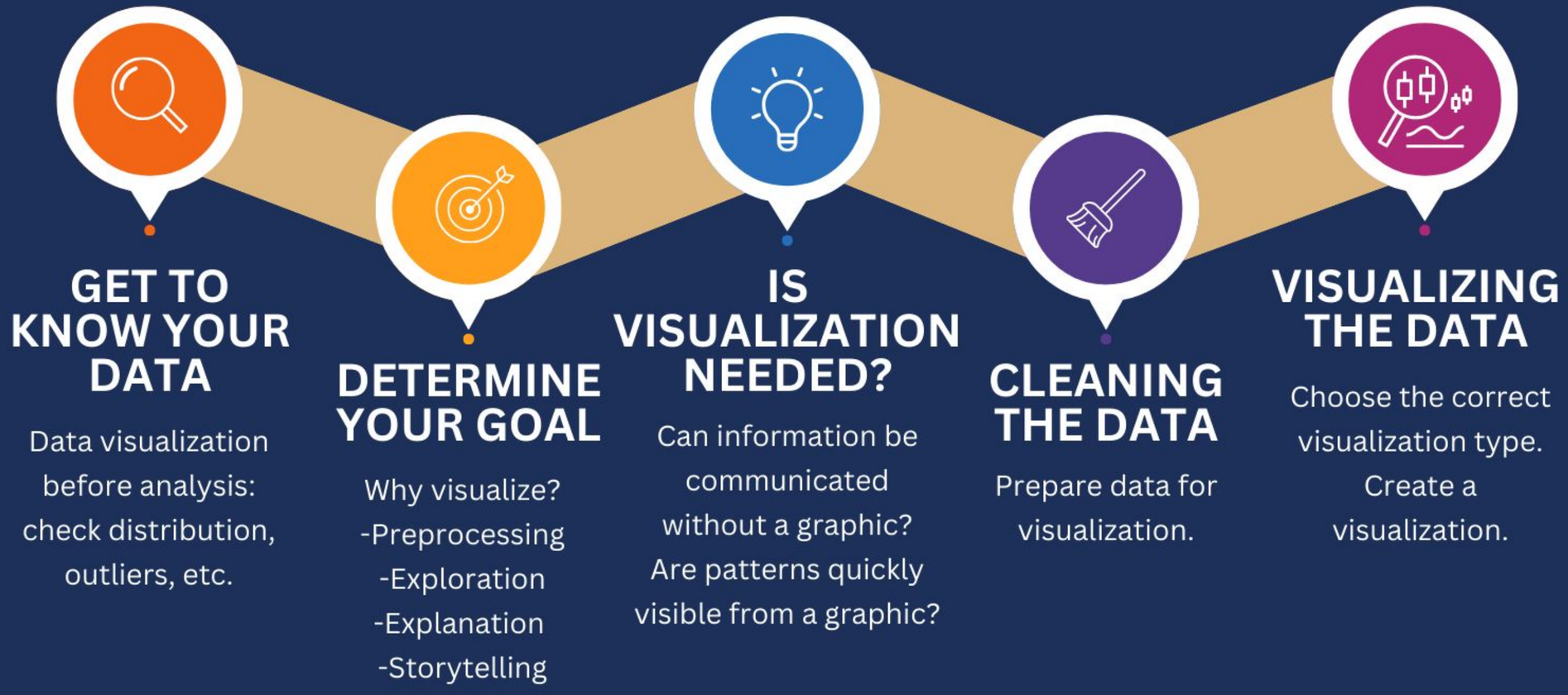
- Explore relationships between variables
  - Visualize trends or patterns
    - **Anscombe's Quartet**



http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

DATA VISUALIZATION PROCESS

Data visualizations from start to finish

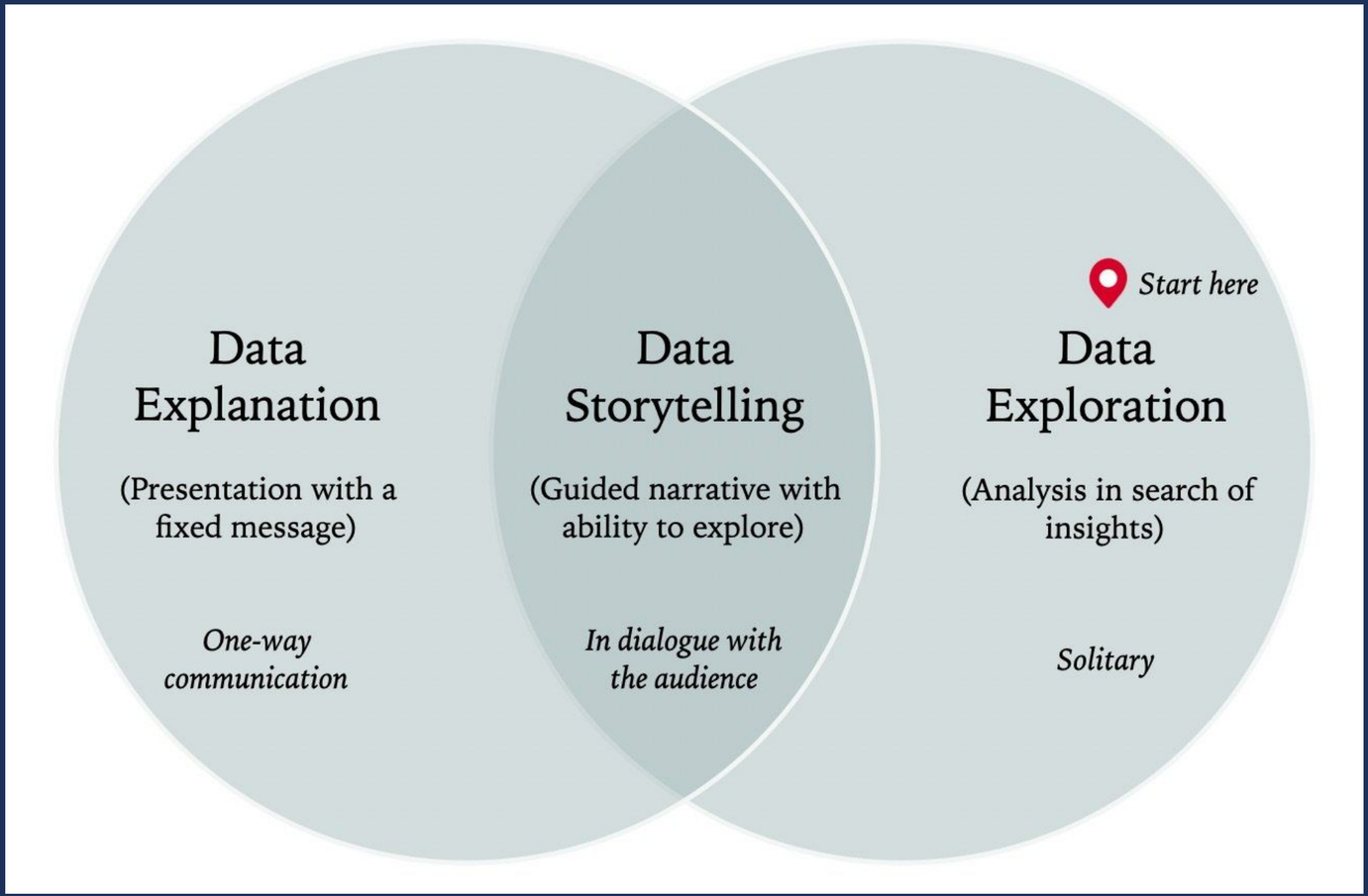


Visualization Goal

DATA ANALYSIS

- Purposes of data visualizations:
  - Exploratory
  - Explanatory
  - Storytelling

- For all purposes, consider:
  - Who is the target audience?
  - Do the data support the message?

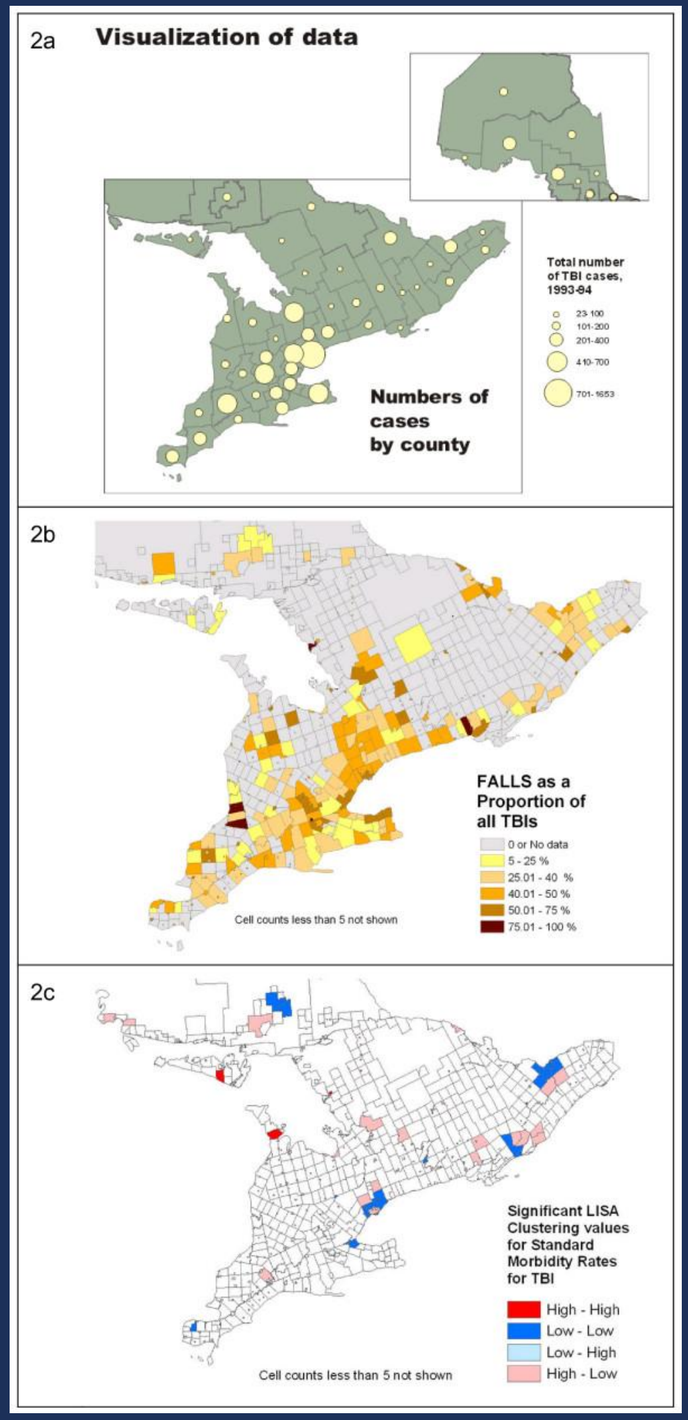


https://www.juiceanalytics.com/writing/exploratory-vs-explanatory-understanding-the-difference-between-data-analysis-and-data-presentation

Visualization Goal

DATA EXPLORATION

- Consider:
  - What features of the data are you comparing and contrasting?
  - What relationships within the data are you exploring?



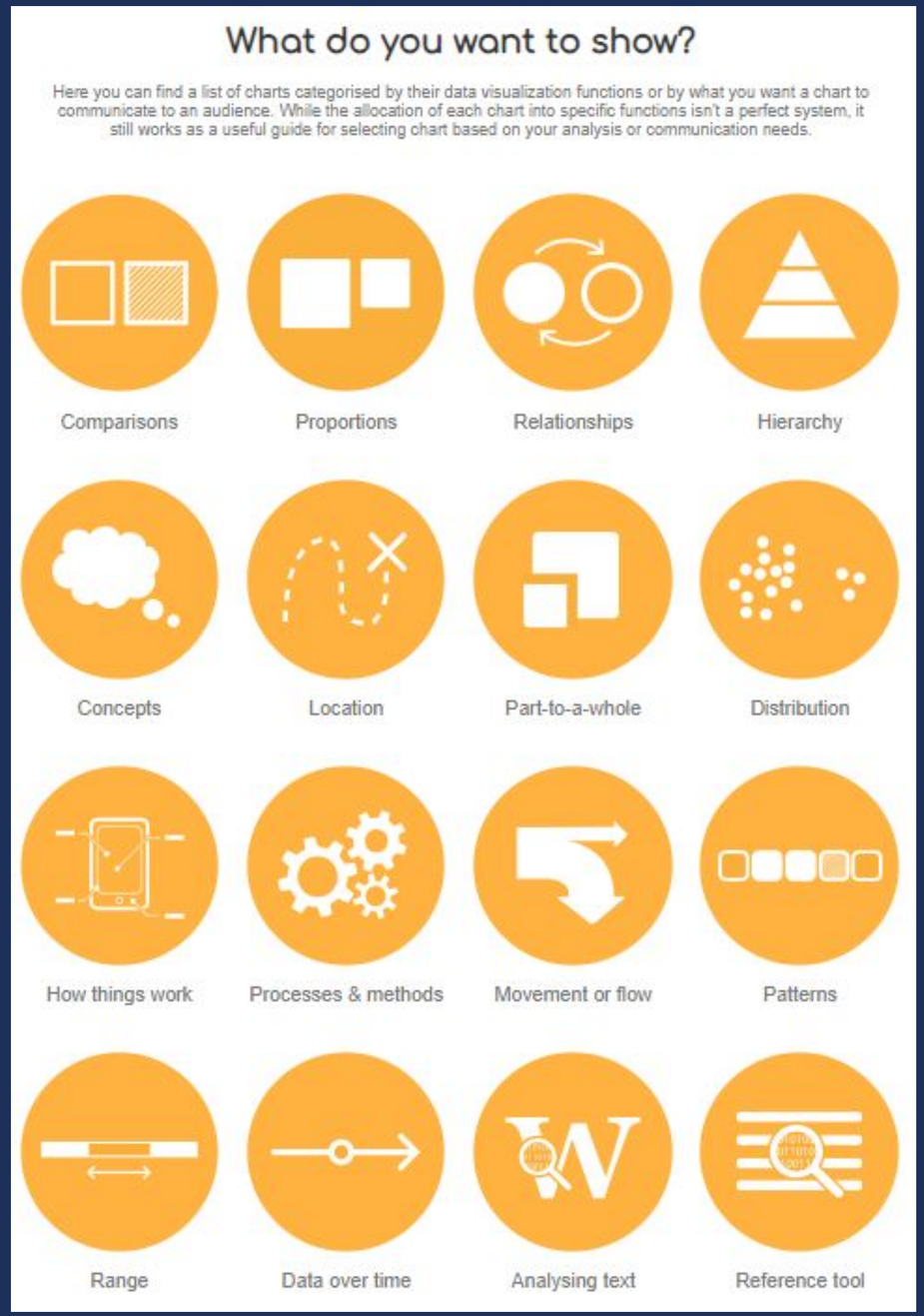
Colantonio, Angela & Moldofsky, Byron & Escobar, Michael & Vernich, Lee & Chipman, Mary & McLellan, Barry. (2011). Using geographical information systems mapping to identify areas presenting high risk for traumatic brain injury. Emerging themes in epidemiology. 8. 7. 10.1186/1742-7622-8-7.

Visualization Goal

DATA EXPLANATION

- Different types of **explanatory** data visualizations.

- Consider:
  - What comparisons or relationships do you want to show?
  - Who is your audience?
  - What visualizations are appropriate for your data type?

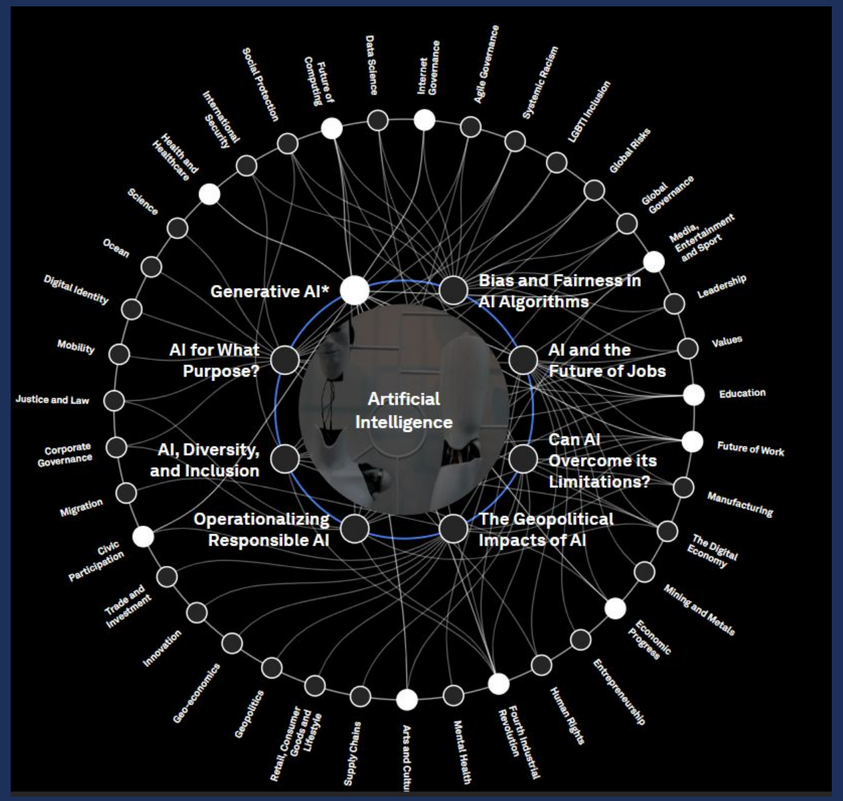


https://datavizcatalogue.com/

Visualization Goal

DATA STORYTELLING

- Consider:
  - What story are you trying to tell?
  - Who is your audience?



https://intelligence.weforum.org/topics/a1Gb0000000pTDREA2/key-issues/a1G680000000Ne9EAE

https://www.youtube.com/watch?v=hRWEteXYD_Y

Visualization Goal

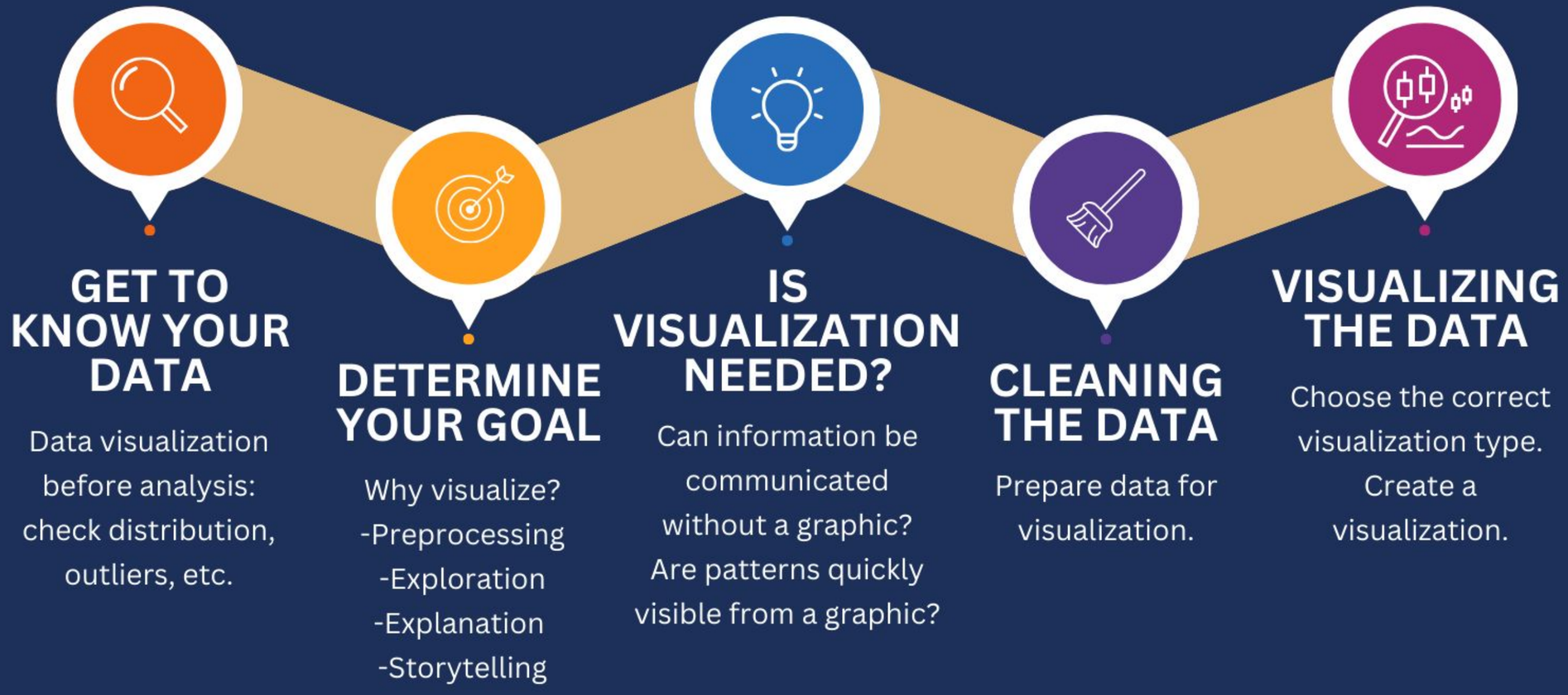
DATA STORYTELLING

- Interactive tools may aid in data storytelling

https://www.washingtonpost.com/politics/interactive/2022/redistricting-mini-golf/

DATA VISUALIZATION PROCESS

Data visualizations from start to finish



IS VISUALIZATION NEEDED?

- **Data visualizations** must be interpretable and relevant
  - Goal-oriented
  - Clearly display patterns or comparisons

WORKSHOP OUTLINE

Our content today is divided into four parts. Each part will be described with examples.

01 Data Visualization Process

Walking through the visualization process, from setting goals to visualizing the data.

02 Basic Design Considerations

Best practices and accessibility considerations in data visualization.

03 Solutions to Common Problems

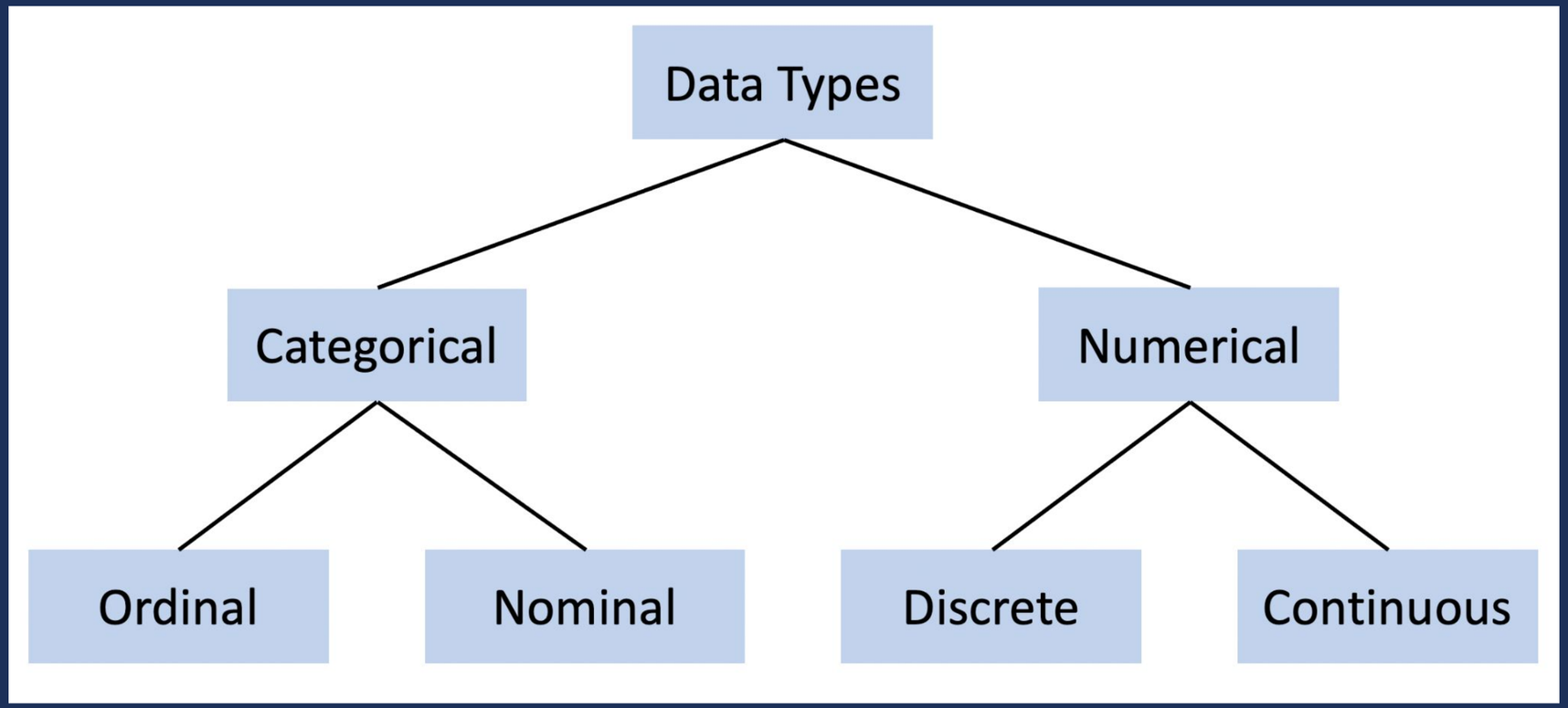
How to resolve, or even better, avoid common problems in data visualization.

04 Visualization Resources

Workshops, tools, and other resources for data visualization.

CONSIDERATIONS AND CONCEPTS

A review of variable types

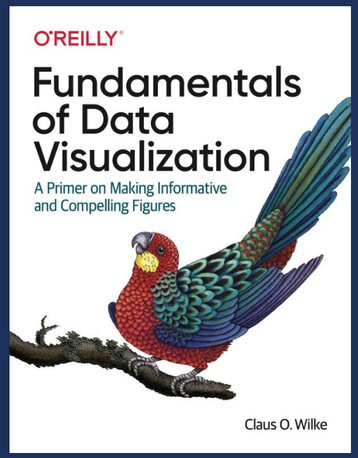


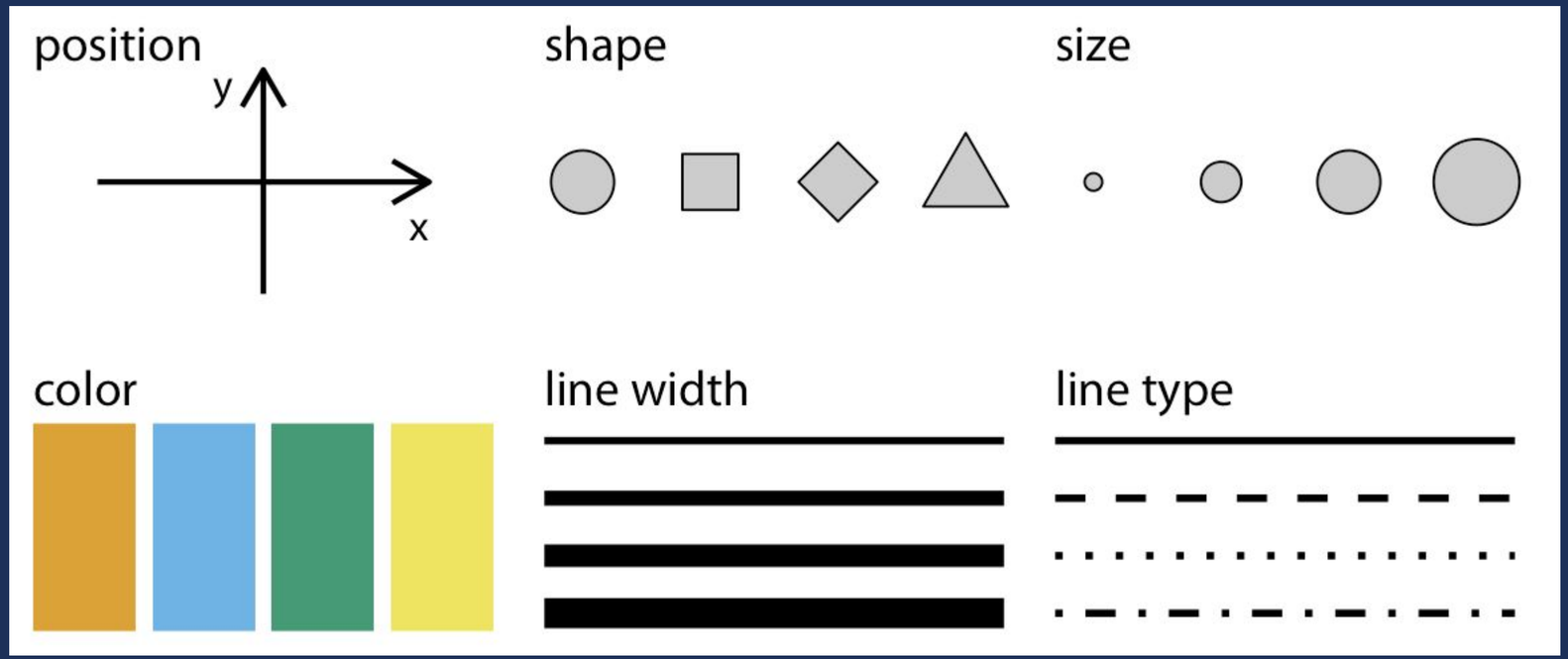
https://bookdown.org/ejvanholm/Textbook/displaying-data.html

Aesthetic properties

Acknowledgement:

Many of the upcoming images are from this book, available online at clauswilke.com/dataviz

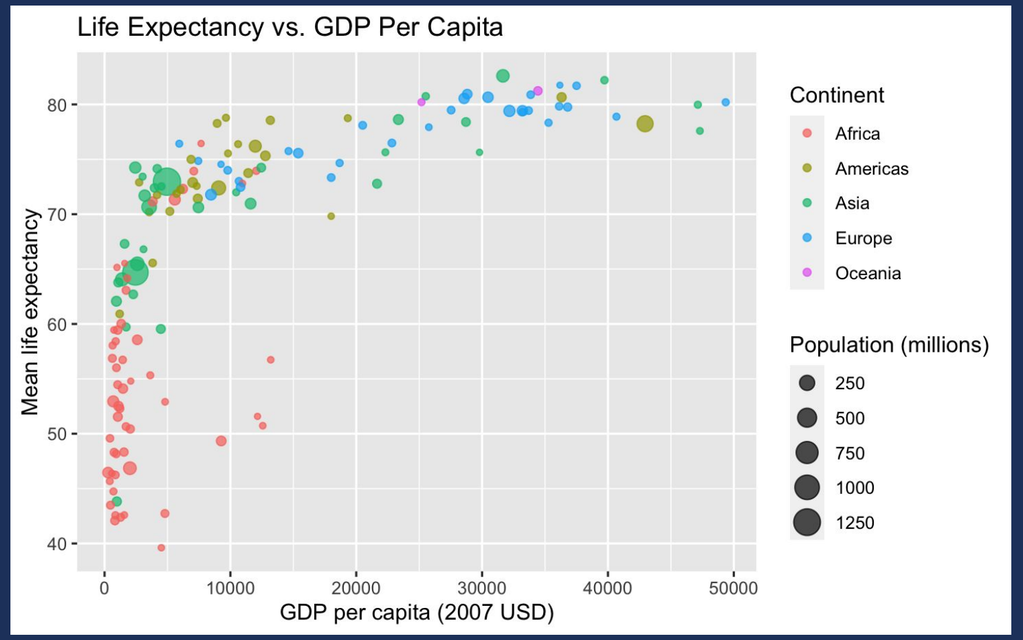




Variable → Aesthetic property

Size Color Y X

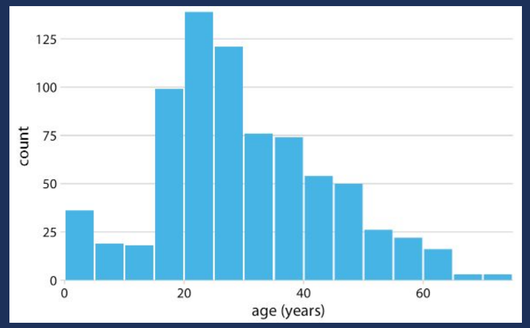
country	pop	continent	lifeExp	gdpPercap	gdp
Afghanistan	31889923	Asia	43.828	974.5803	3.107929e+10
Albania	3600523	Europe	76.423	5937.0295	2.137641e+10
Algeria	33333216	Africa	72.301	6223.3675	2.074449e+11
Angola	12420476	Africa	42.731	4797.2313	5.958390e+10
Argentina	40301927	Americas	75.320	12779.3796	5.150336e+11
Australia	20434176	Oceania	81.235	34435.3674	7.036584e+11
Austria	8199783	Europe	79.829	36126.4927	2.962294e+11
Bahrain	708573	Asia	75.635	29796.0483	2.111268e+10
Bangladesh	150448339	Asia	64.062	1391.2538	2.093118e+11
Belgium	10392226	Europe	79.441	33692.6051	3.501412e+11



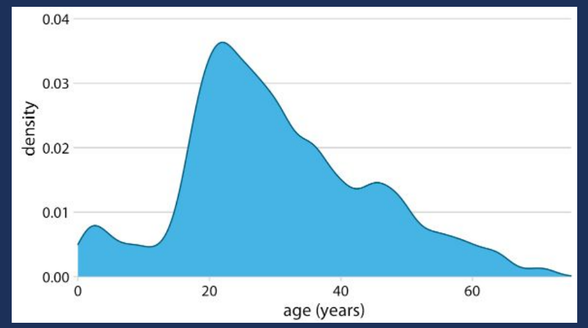
Describing data with visualizations

Single variable

Numeric - Continuous

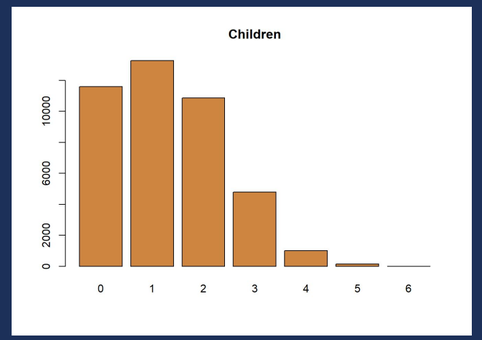


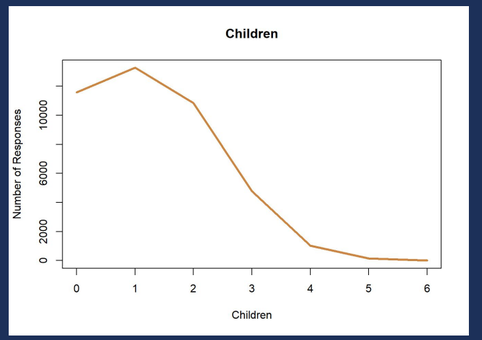
Histogram



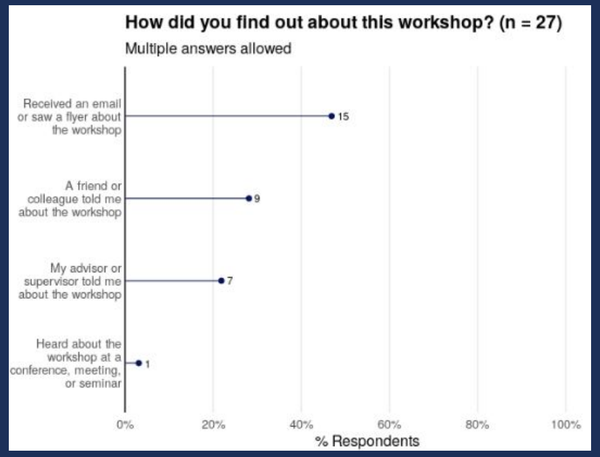
Kernel Density Estimate

Numeric - Discrete





Categorical - Nominal

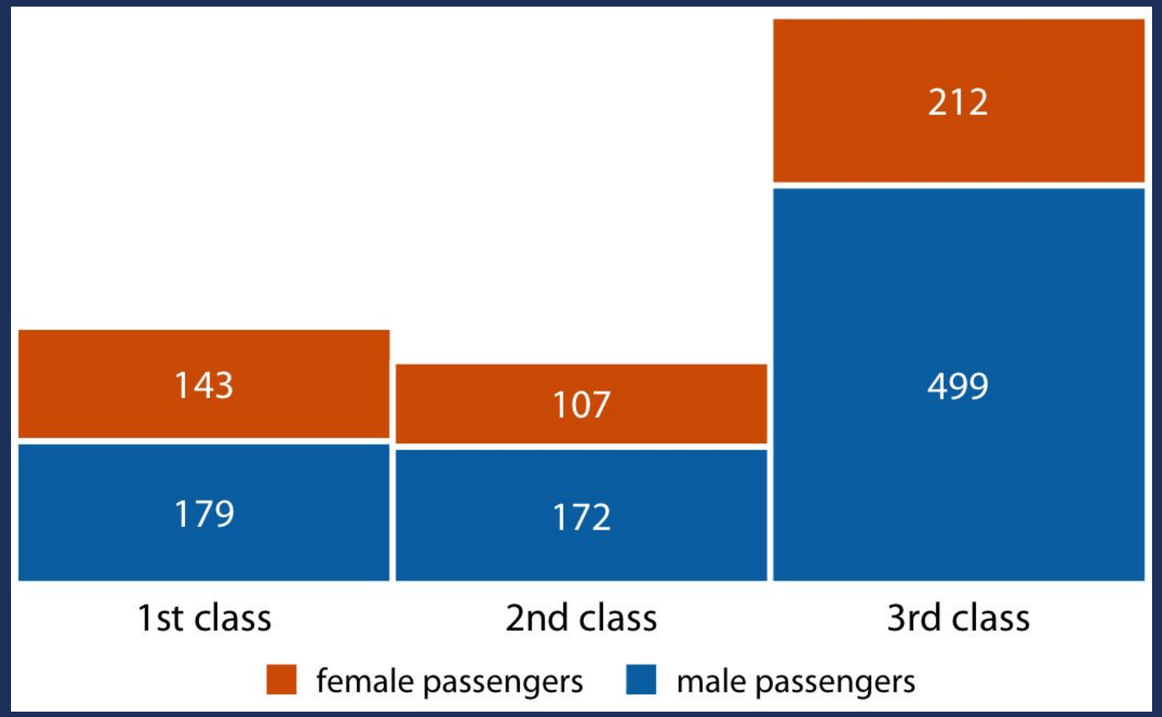


Categorical - Ordinal

Describing data with visualizations

Relationship between categorical variables

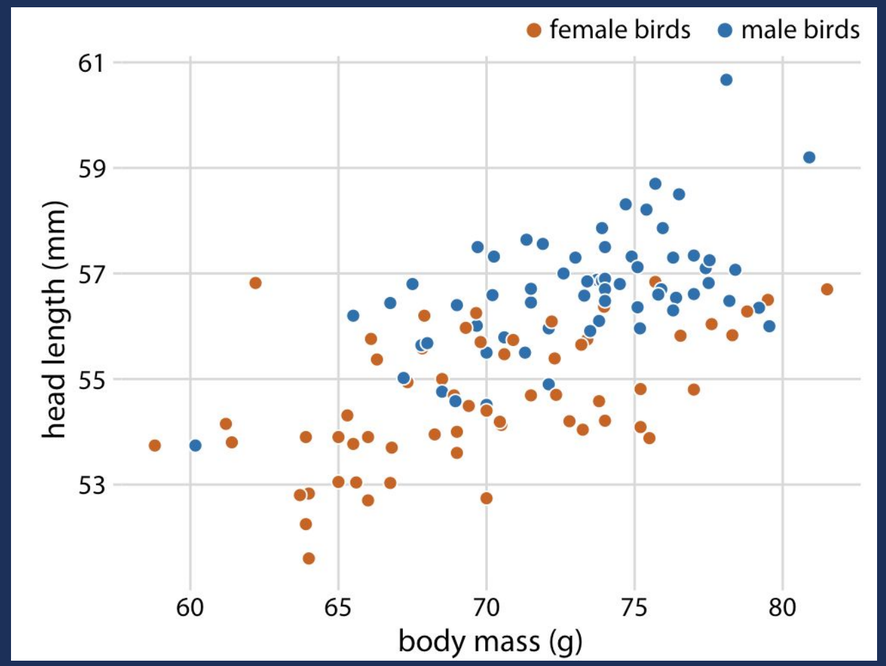
Categorical ~ Categorical

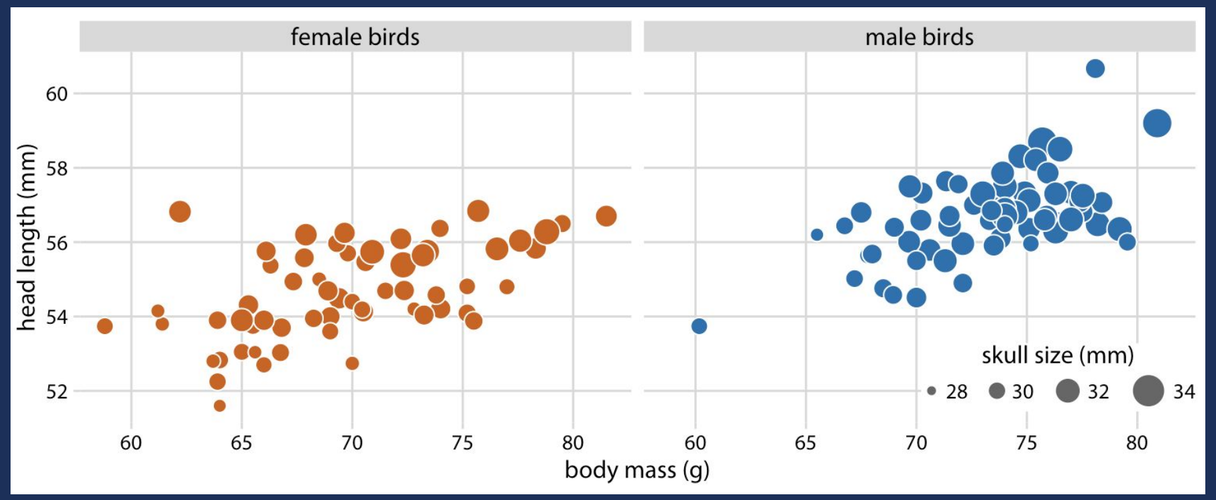


Describing data with visualizations

Relationship between continuous variables

Numeric/Continuous ~ Numeric/Continuous



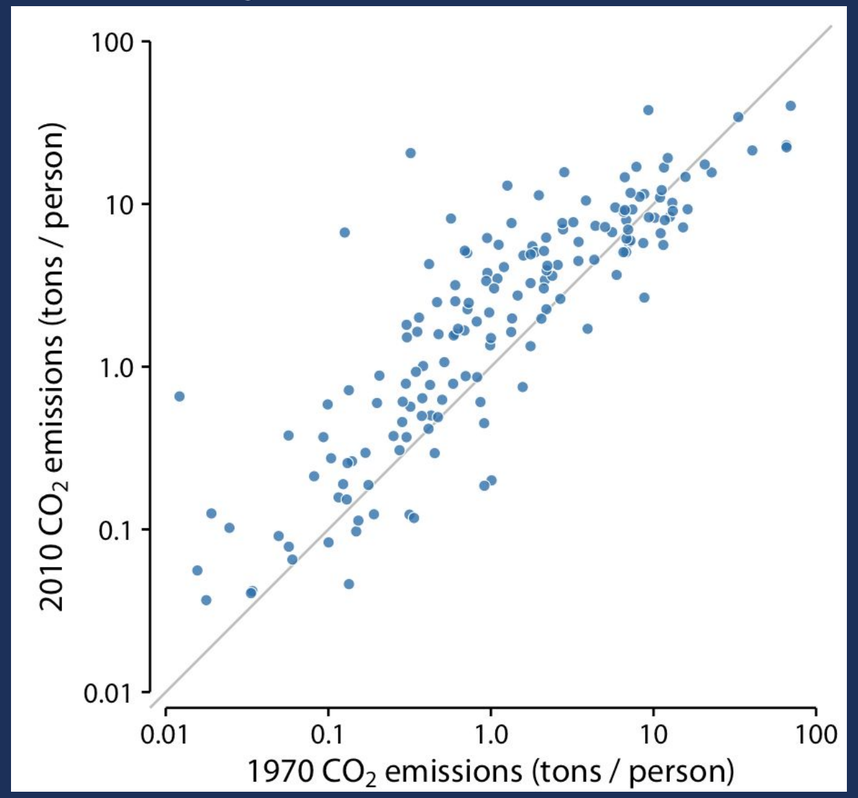


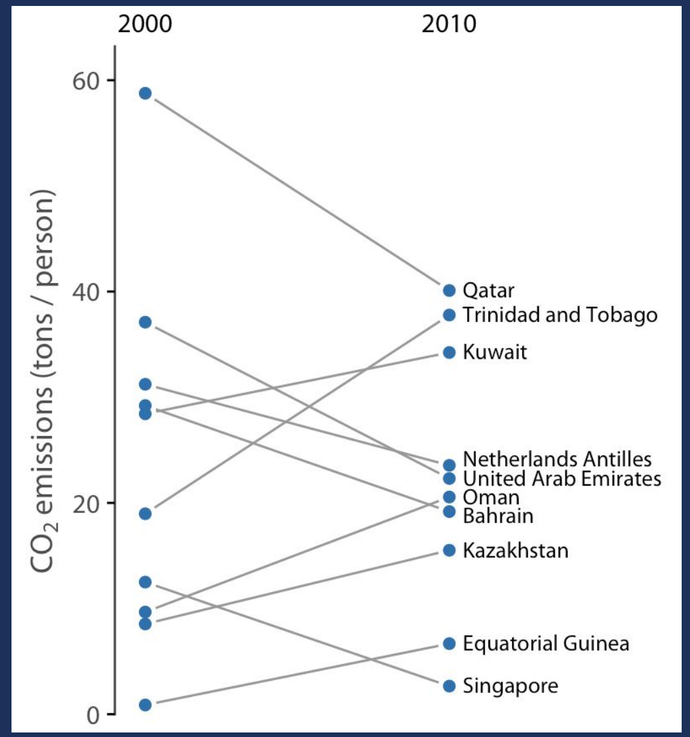
Bubble plot + faceting

Describing data with visualizations

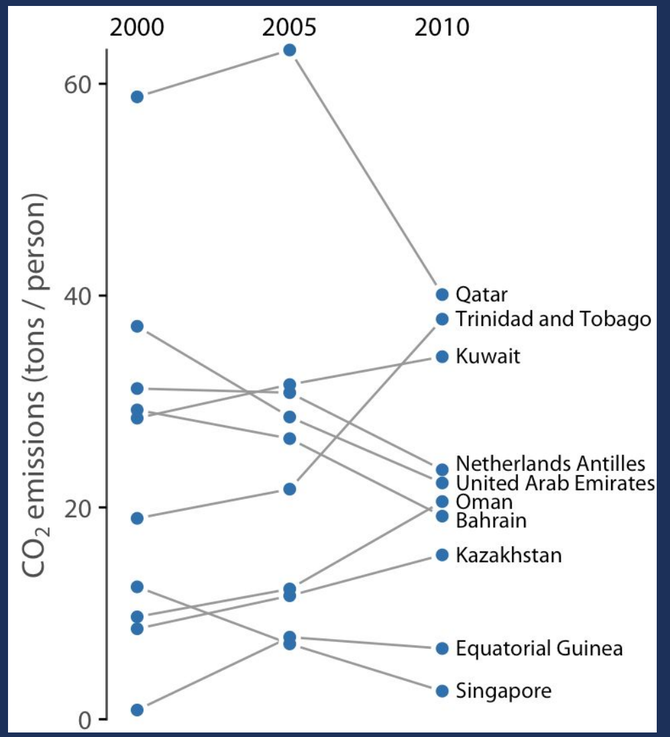
Relationship between "paired" continuous variables

x & y variables have the same scale,
so an x=y line is useful



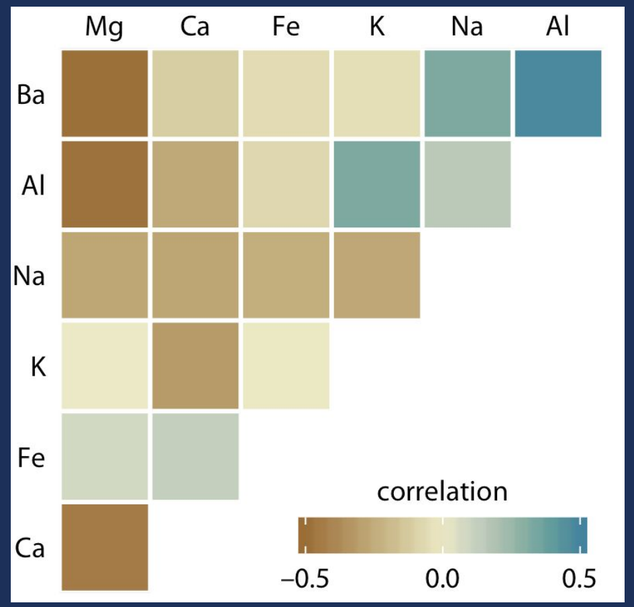


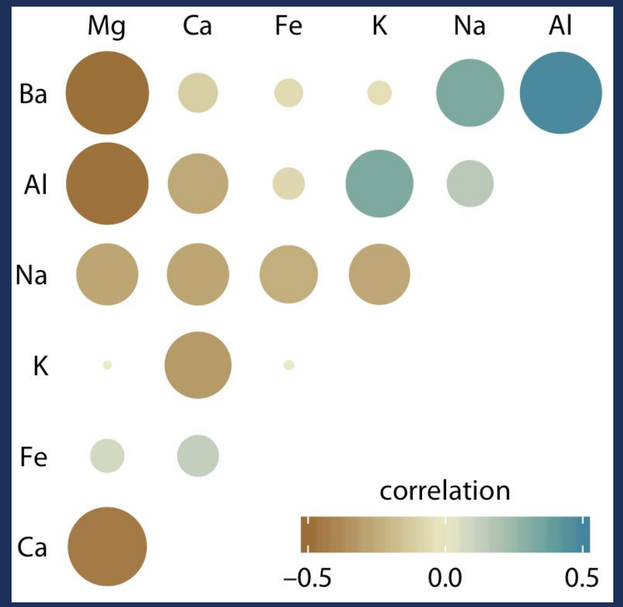
Slopegraph



Describing data with visualizations

Correlation between continuous variables



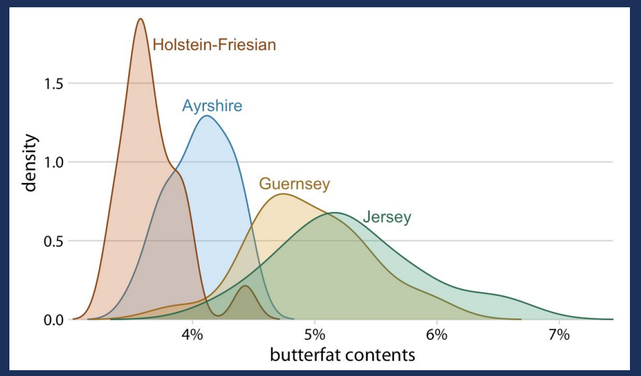


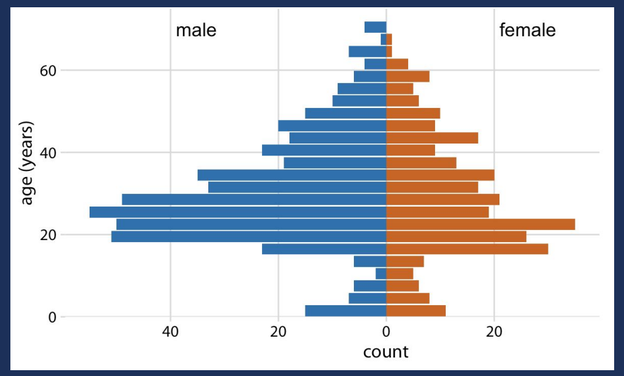
Correlogram

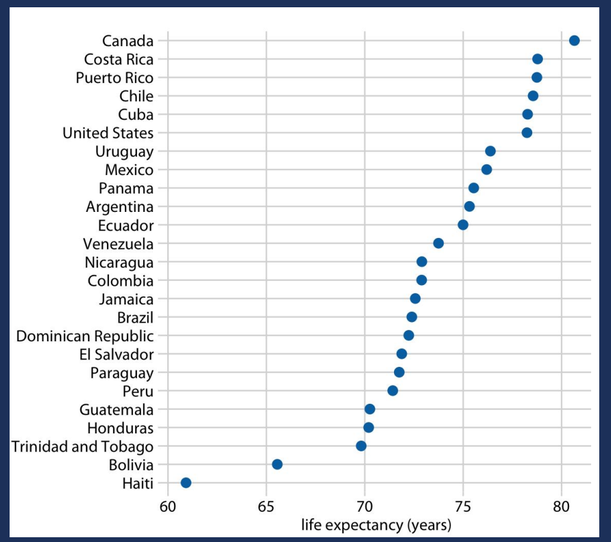
Describing data with visualizations

Relationship between continuous & categorical

Numeric ~ Categorical





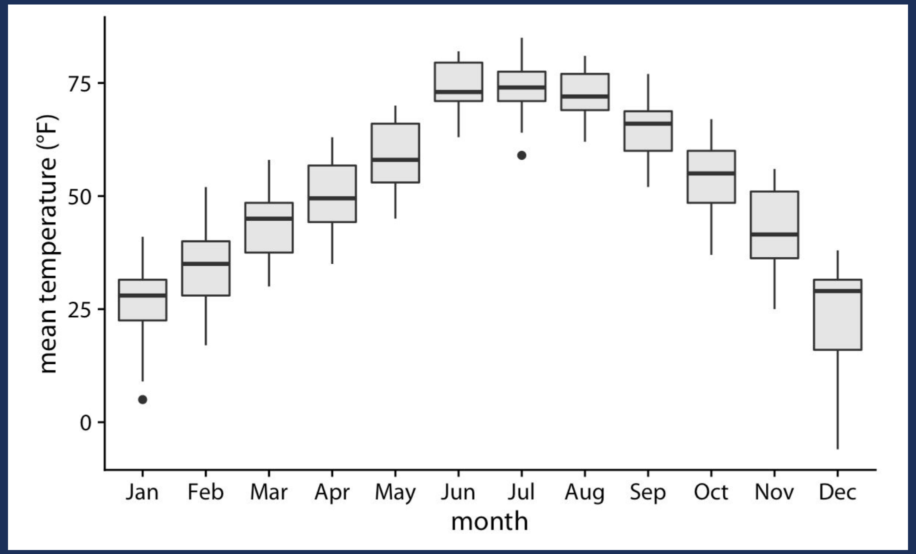


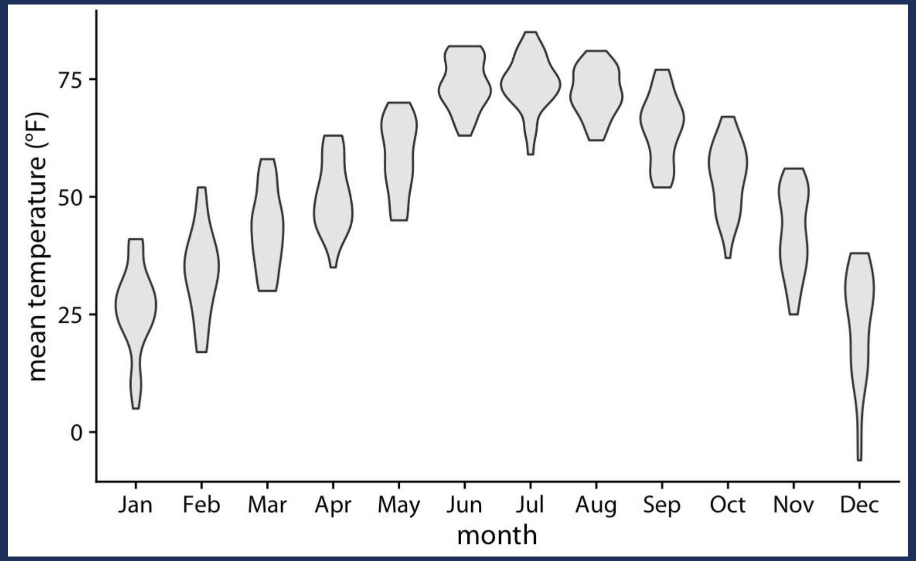
Dot plot

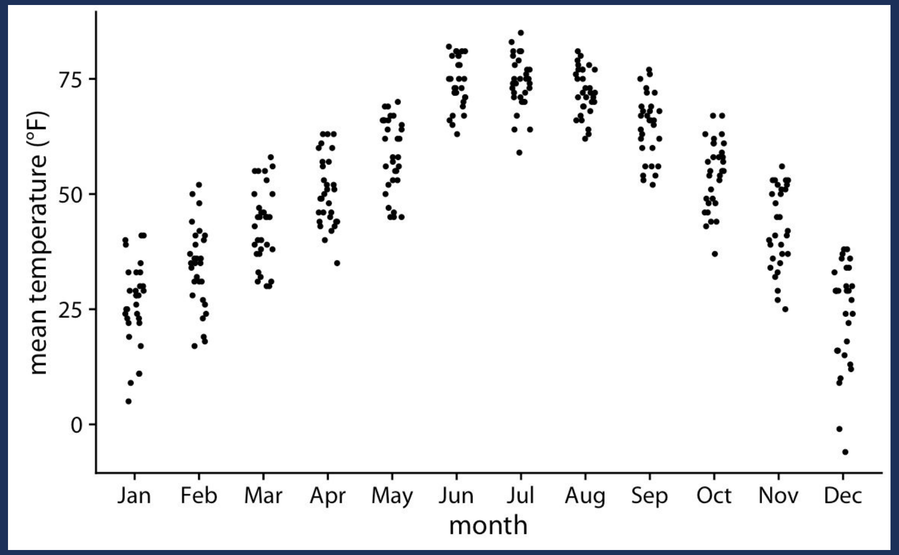
Describing data with visualizations

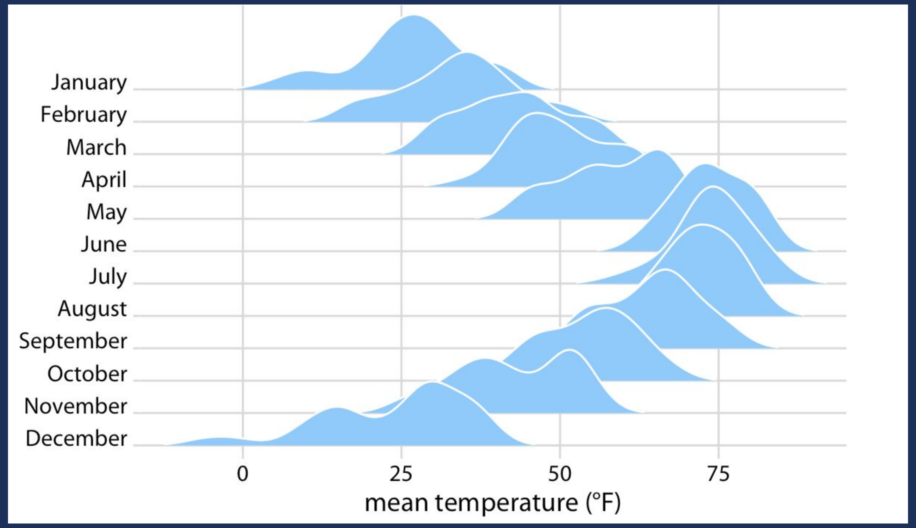
Relationship between two variables

Numeric ~ Categorical



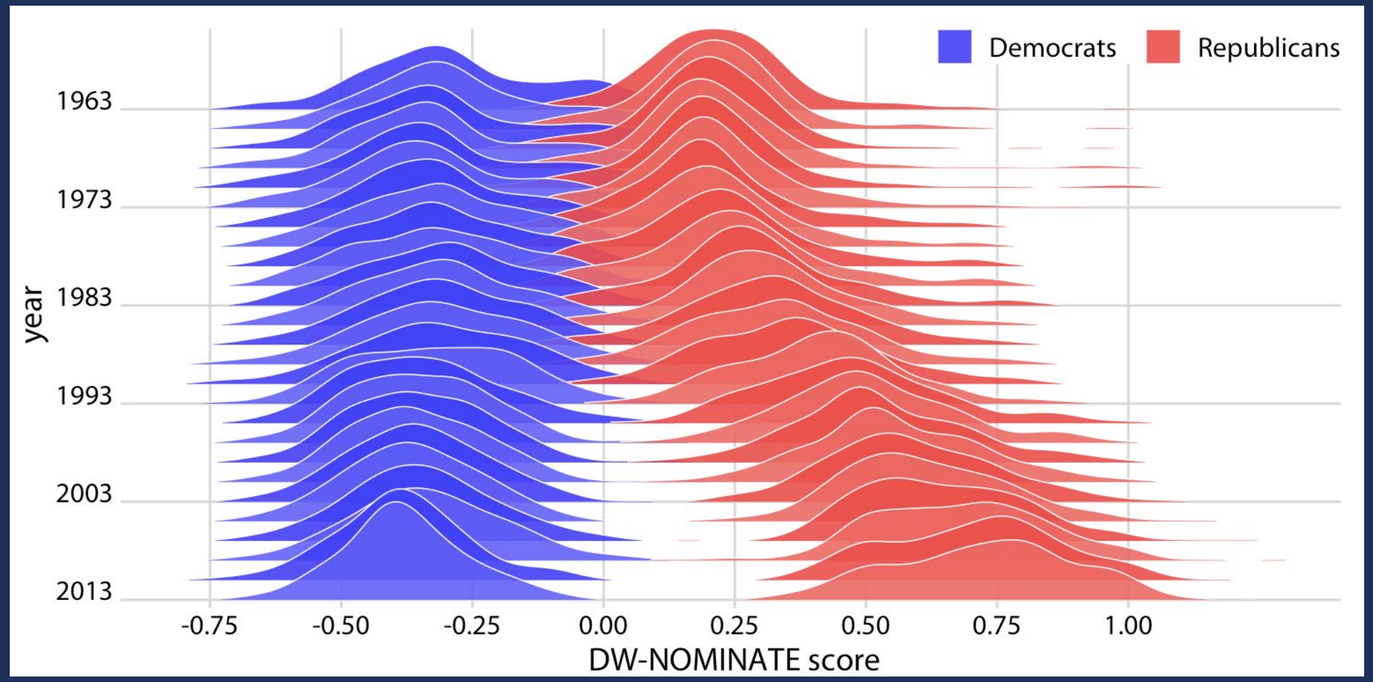






Describing data with visualizations

Which types of relationships are these visualizations showing?

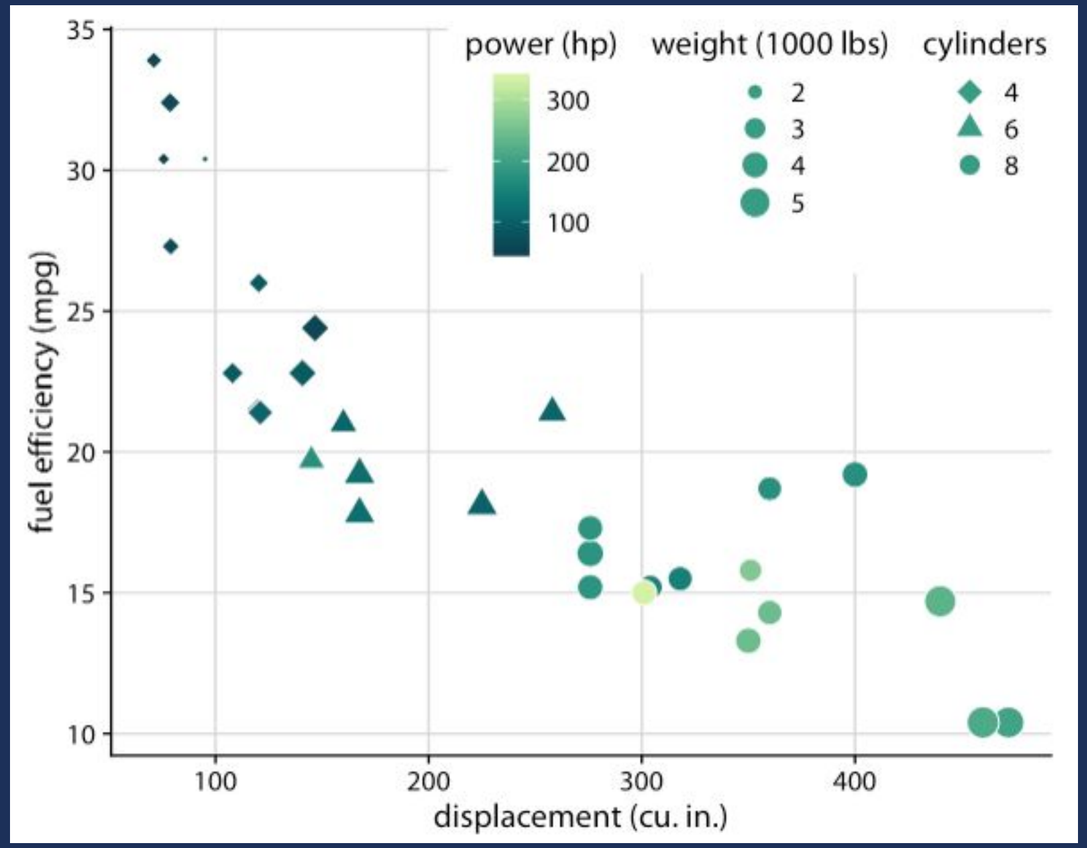


Ridgeline Plot

Describing data with visualizations

Relationship between three or more variables

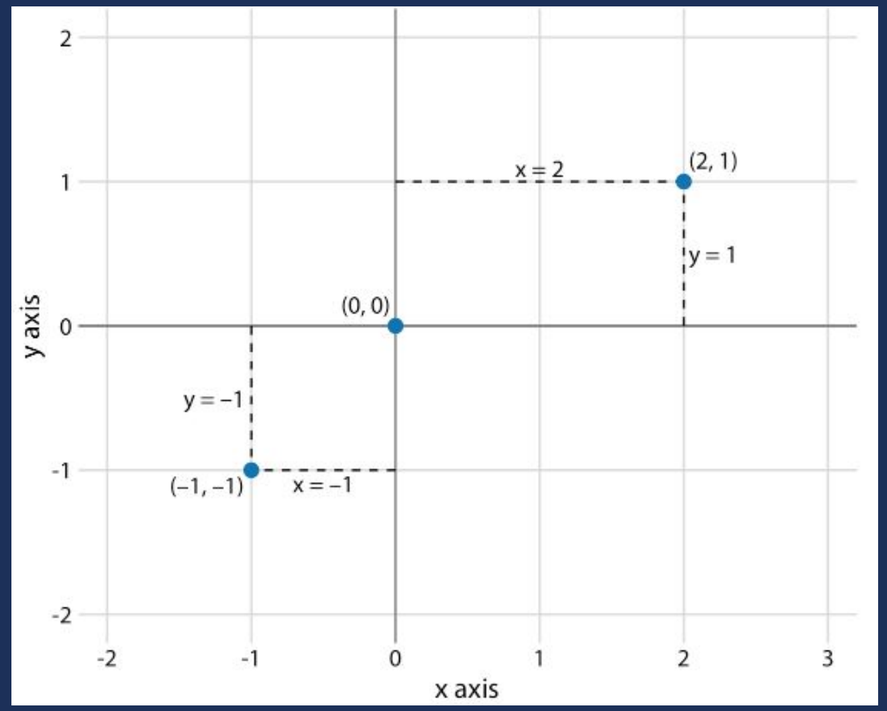
How many variables do we have here, which types, and which aesthetics are mapped to them?



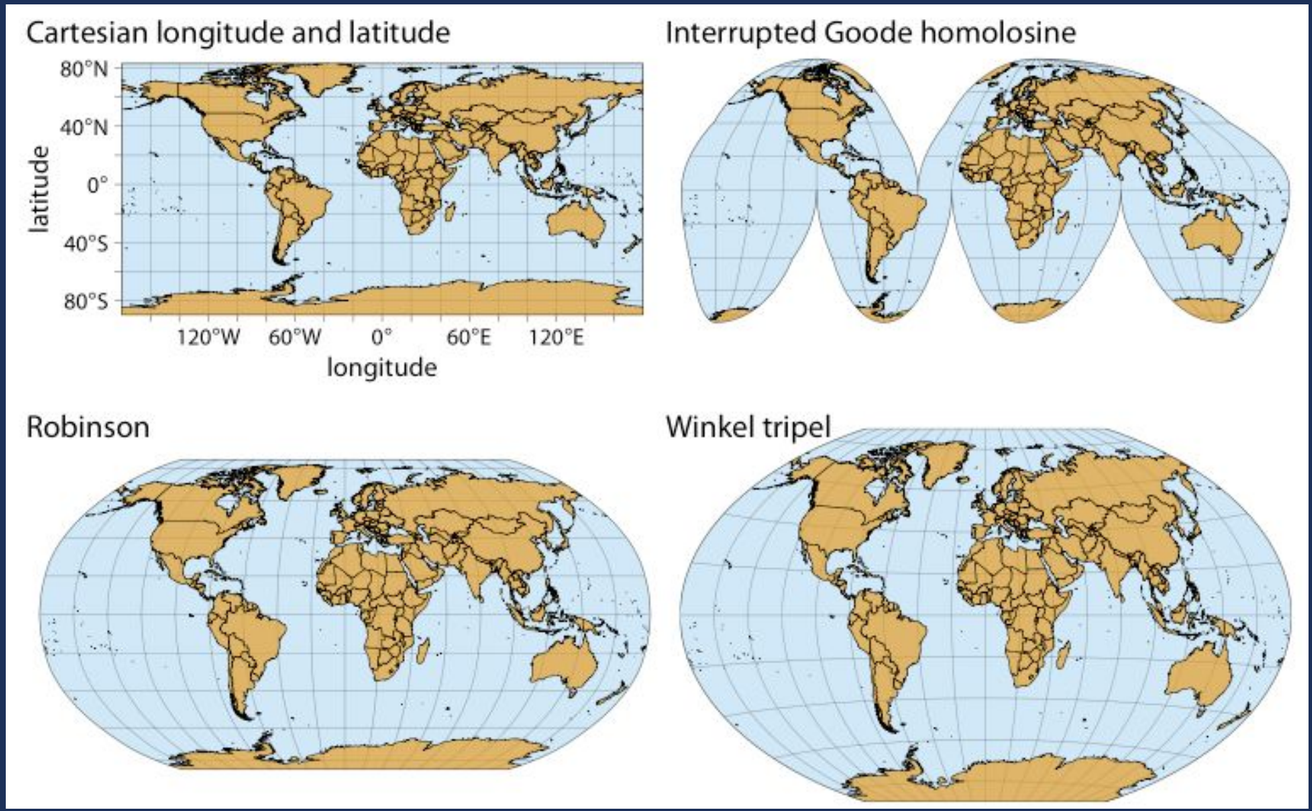
Position

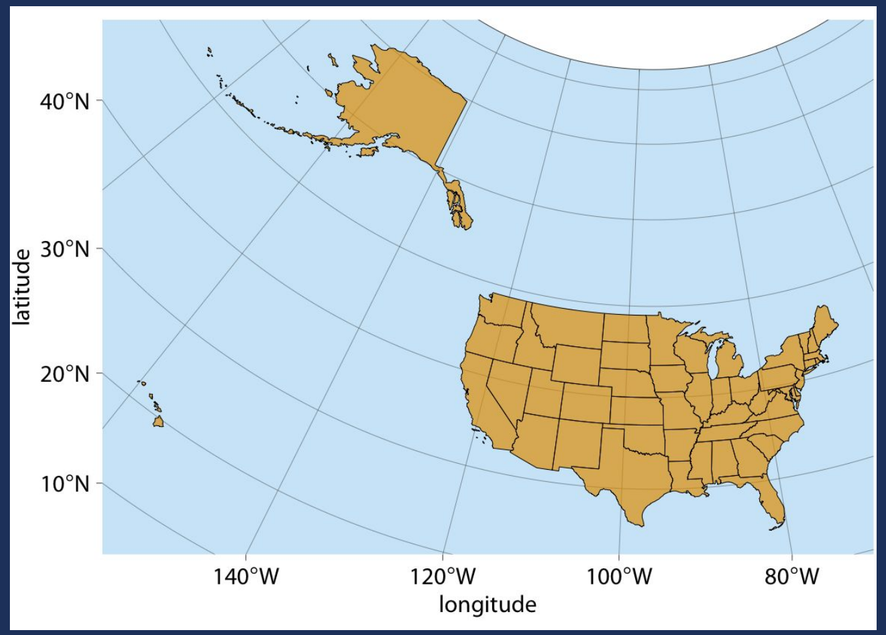
Types of position:

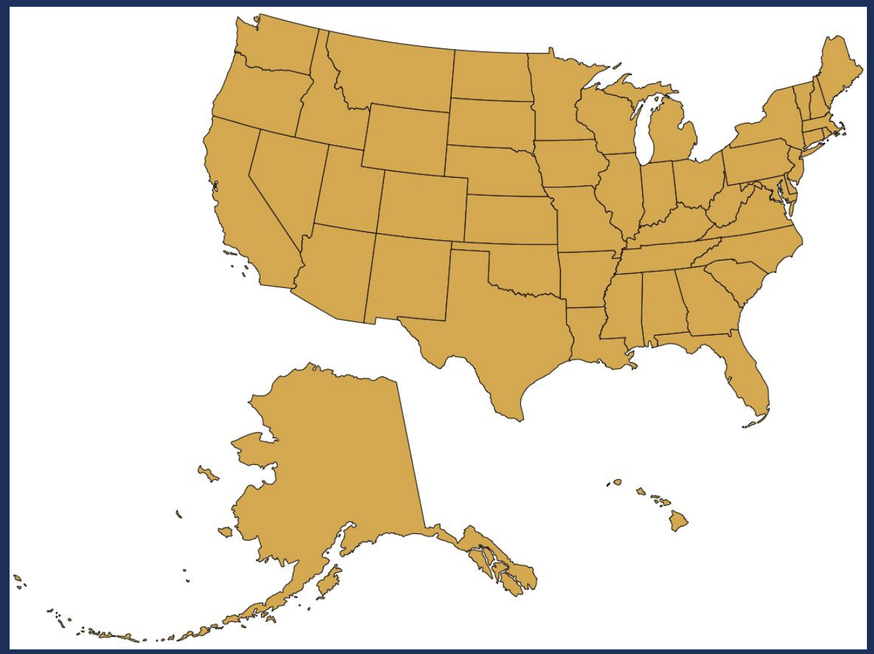
- Cartesian (x/y)
- Radial/Polar (r, theta)
- Geographic (lat/long)



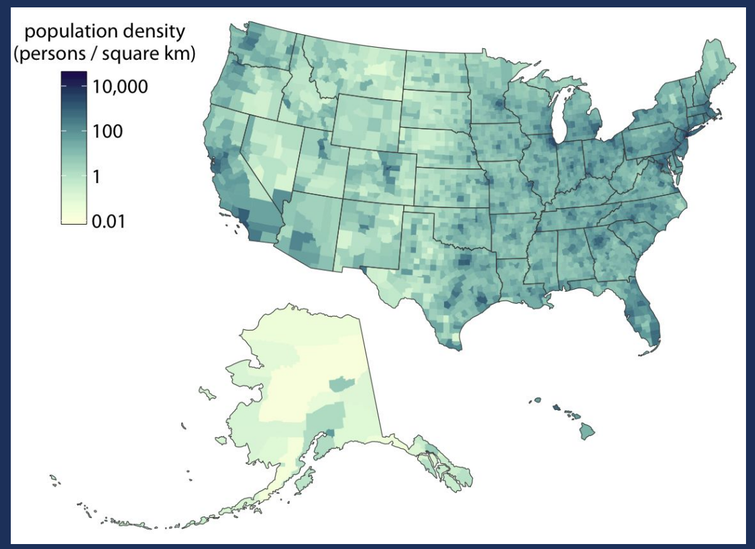
Geospatial projections





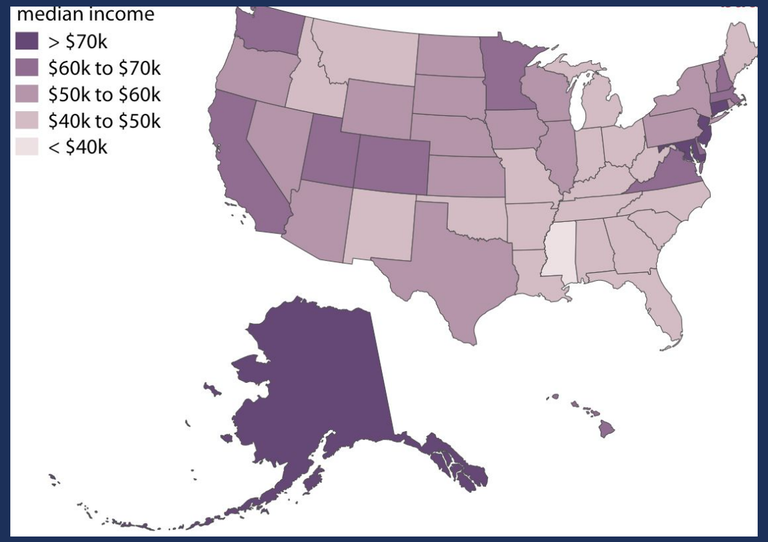


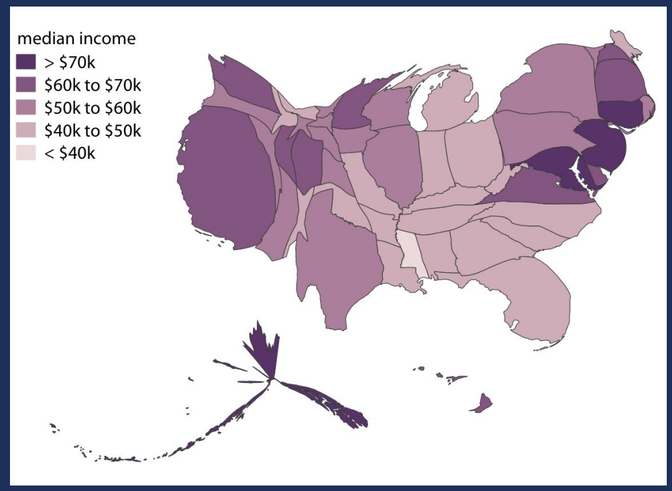
Geospatial visualizations



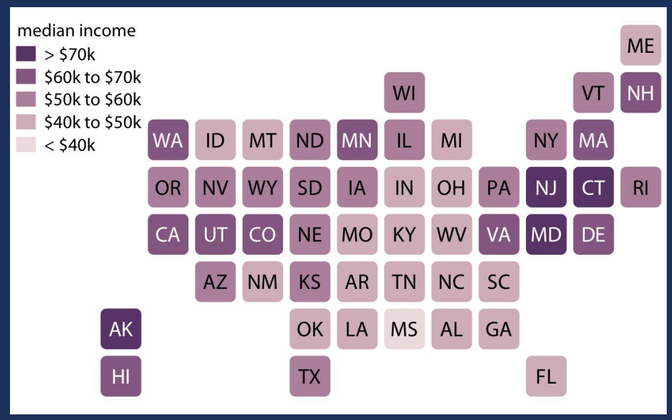
Choropleth

Geospatial visualizations





cartogram

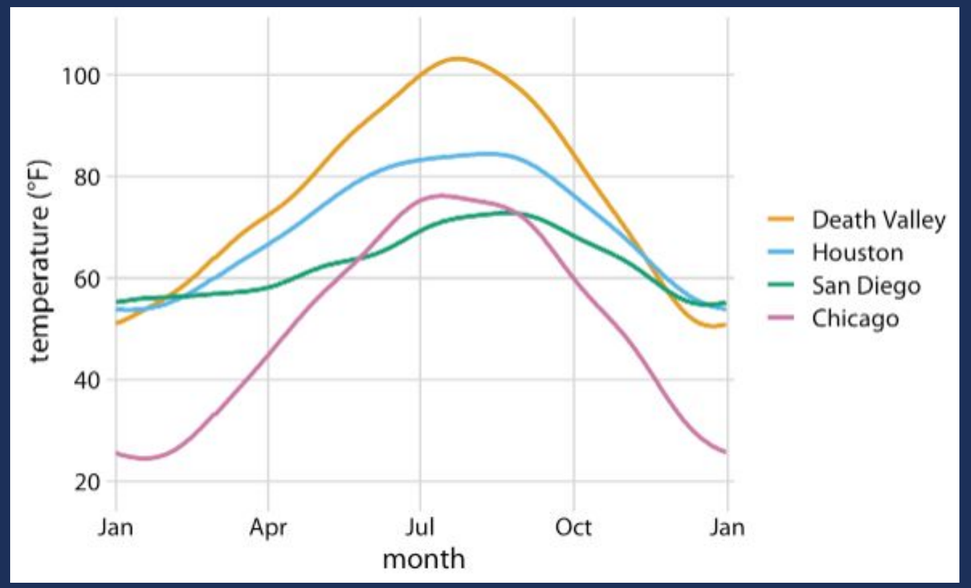


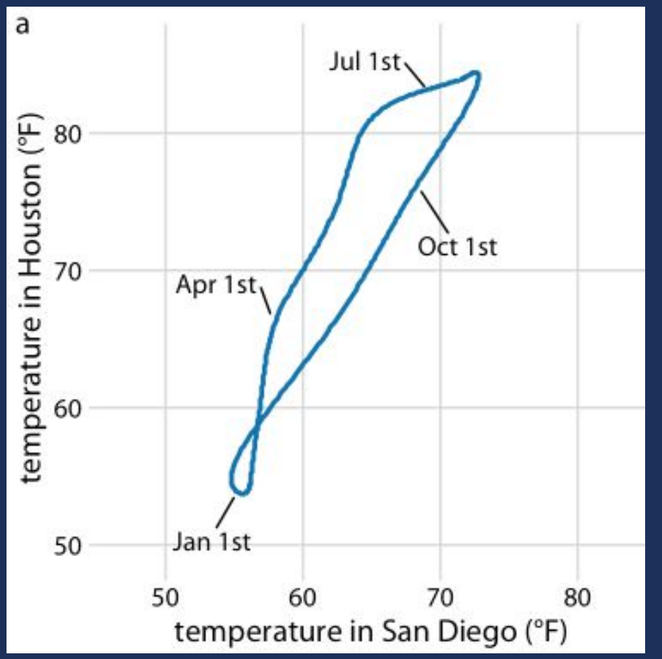
cartogram heatmap

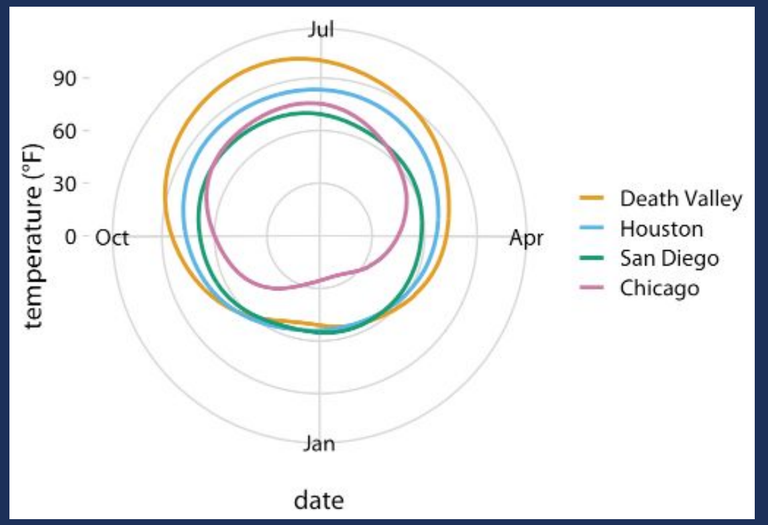
How might you represent this data?

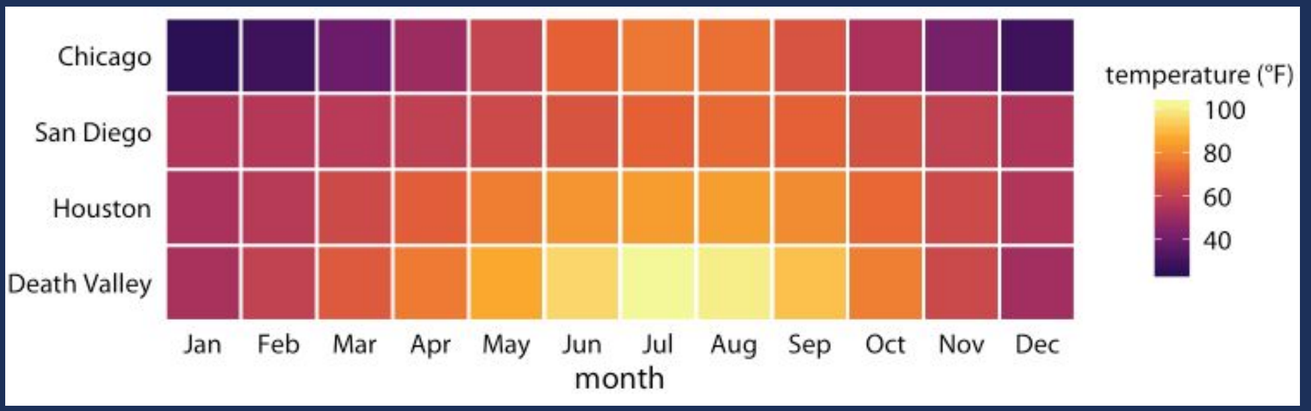
Table 2.2: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.

Month	Day	Location	Station ID	Temperature
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319	51.2
Jan	3	Chicago	USW00014819	25.3
Jan	3	San Diego	USW00093107	55.3
Jan	3	Death Valley	USC00042319	51.3
Jan	3	Houston	USW00012918	53.8


temperature (°F)
100
80
60
40
20
Jan
Apr
Jul
Oct
Jan
month
Death Valley
Houston
San Diego
Chicago


a
Jul 1st
Oct 1st
Apr 1st
Jan 1st
temperature in Houston (°F)
80
70
60
50
50
60
70
80
temperature in San Diego (°F)

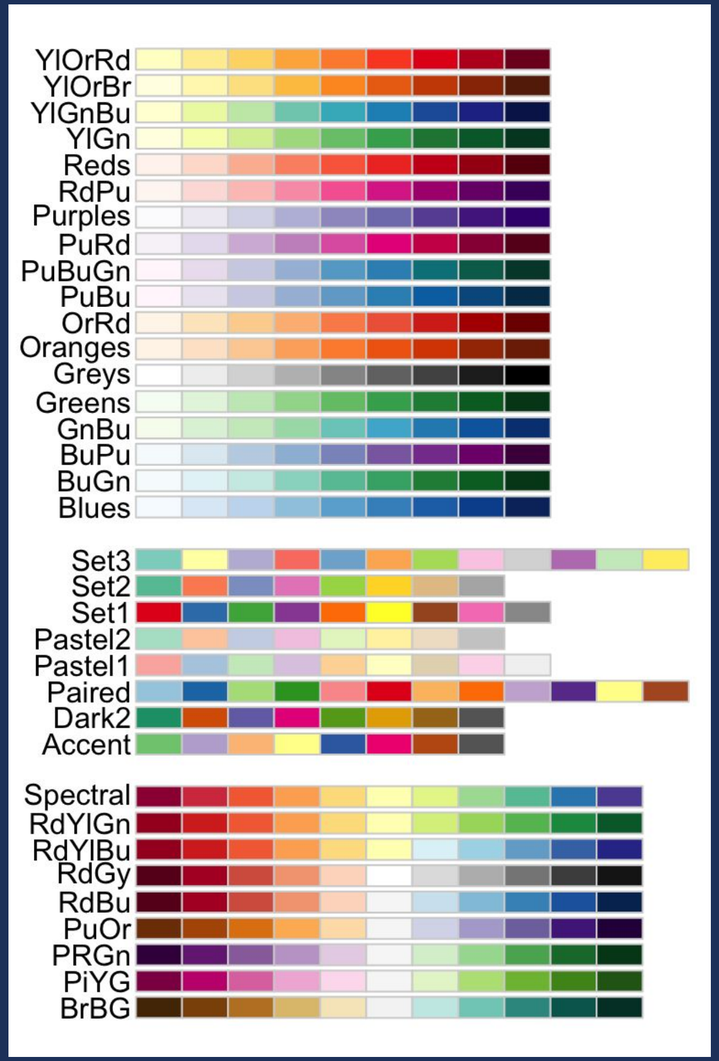

Jul
90
60
30
0
Oct
Apr
Jan
temperature (°F)
date
Death Valley
Houston
San Diego
Chicago


Chicago
San Diego
Houston
Death Valley
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month
temperature (°F)
100
80
60
40

Color

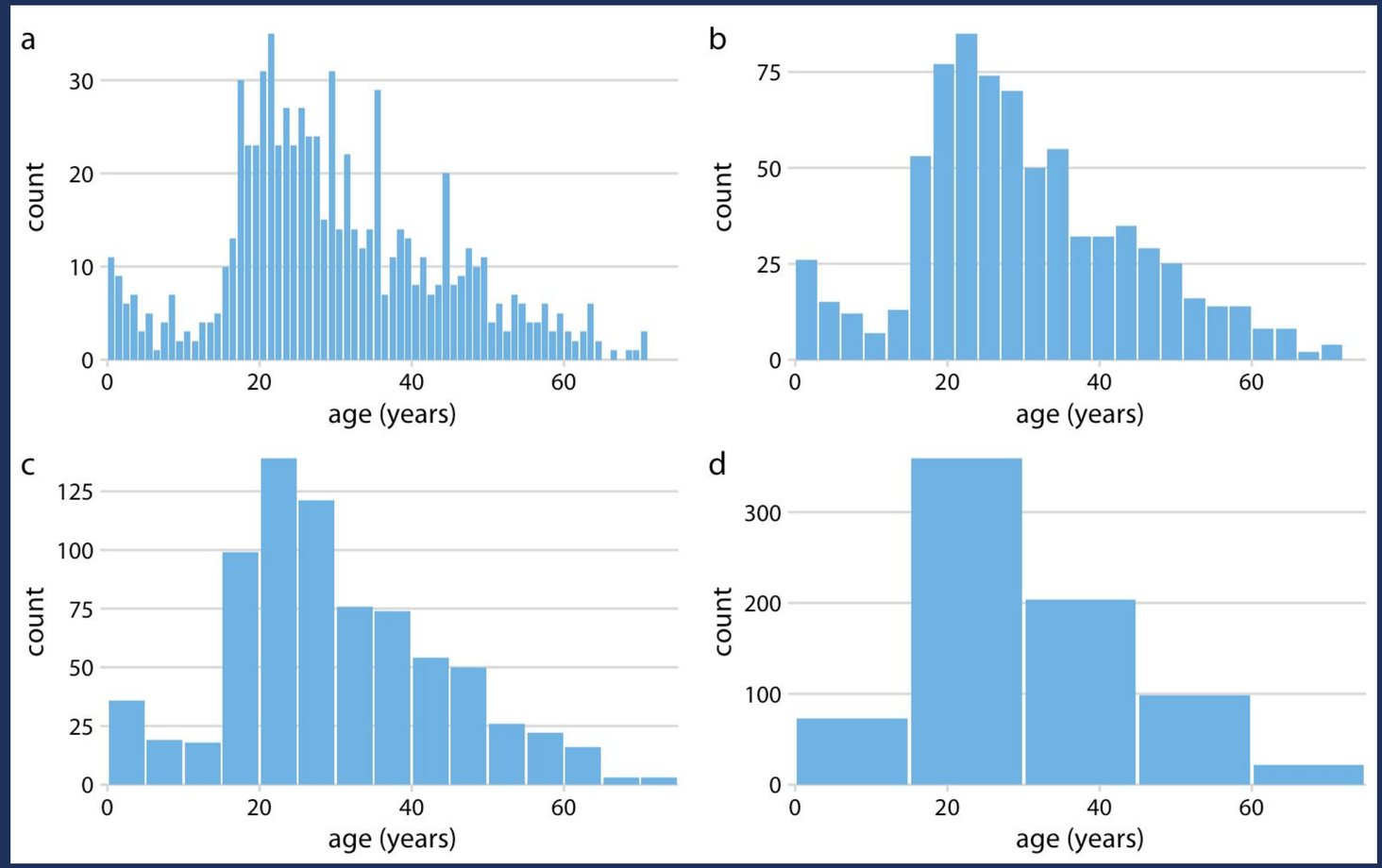
Think about what colors you might use to represent:

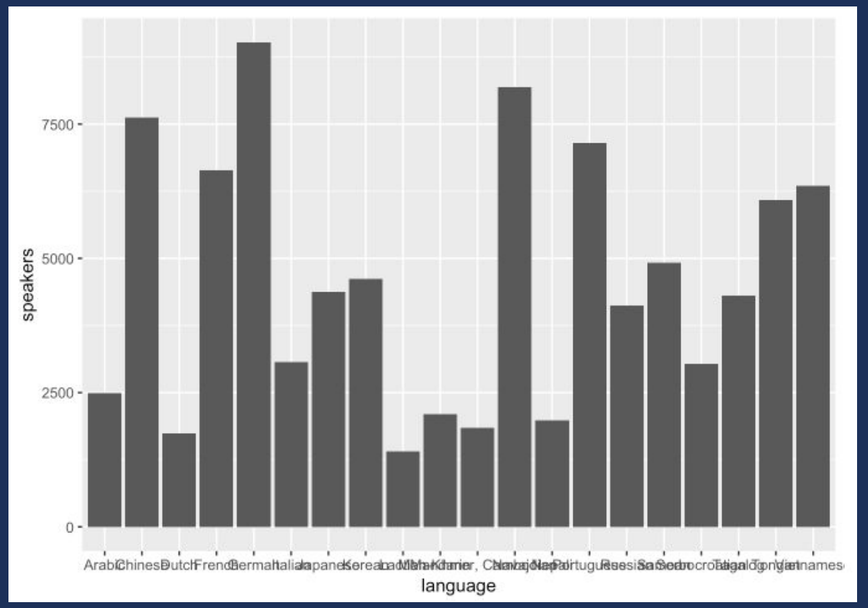
- A numeric variable (whether continuous or discrete)
- A nominal (unordered) categorical variable
- An ordinal categorical variable

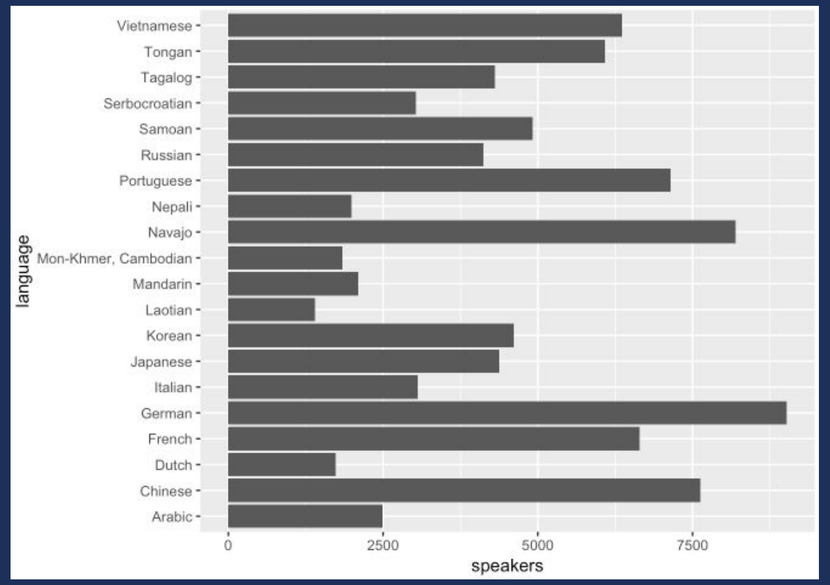


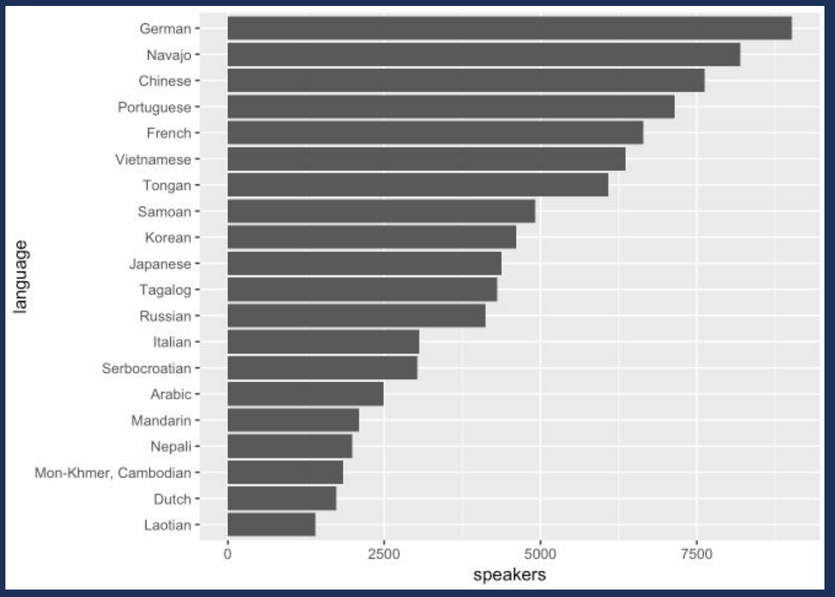
WORKING THROUGH EXAMPLES

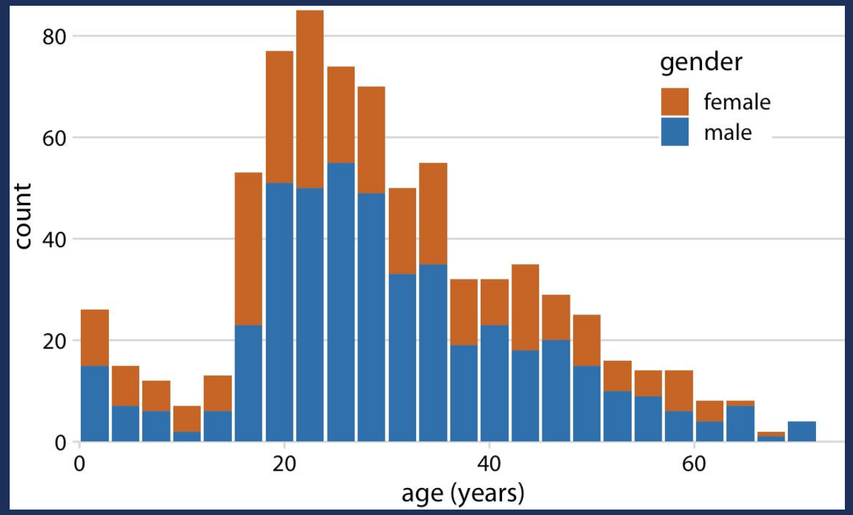
Good, Bad, or Ugly?

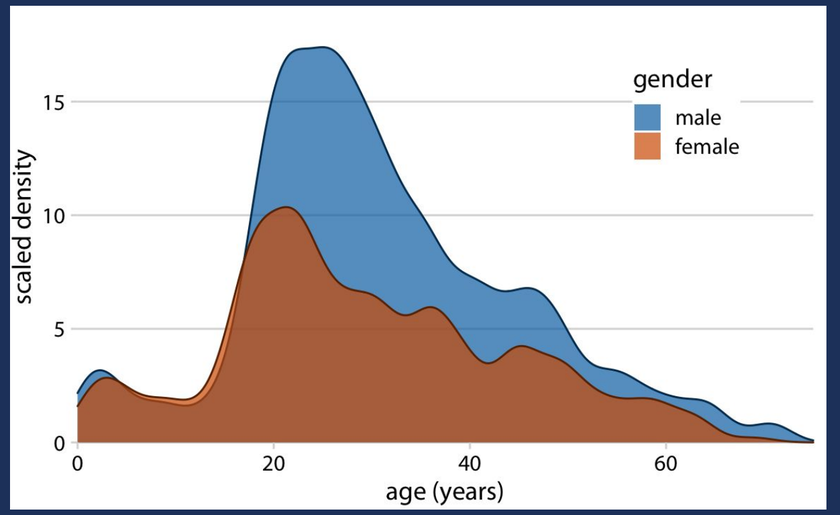


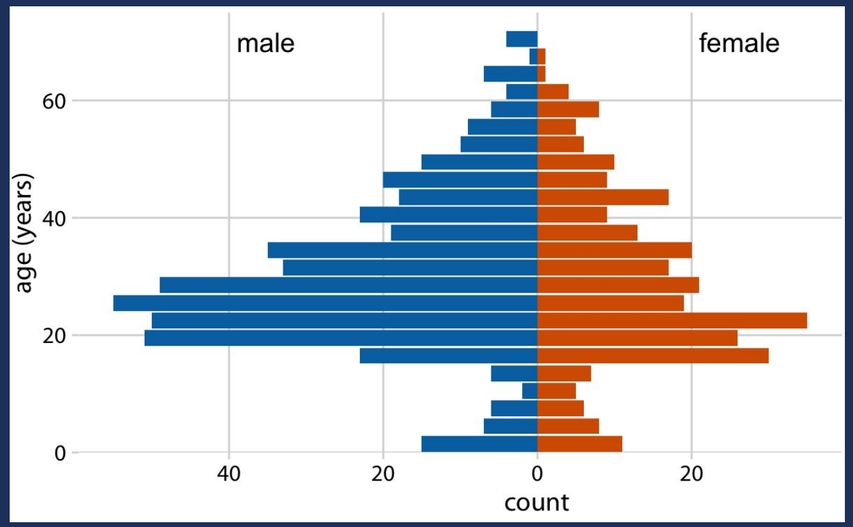

speakers
7500
5000
2500
0
language

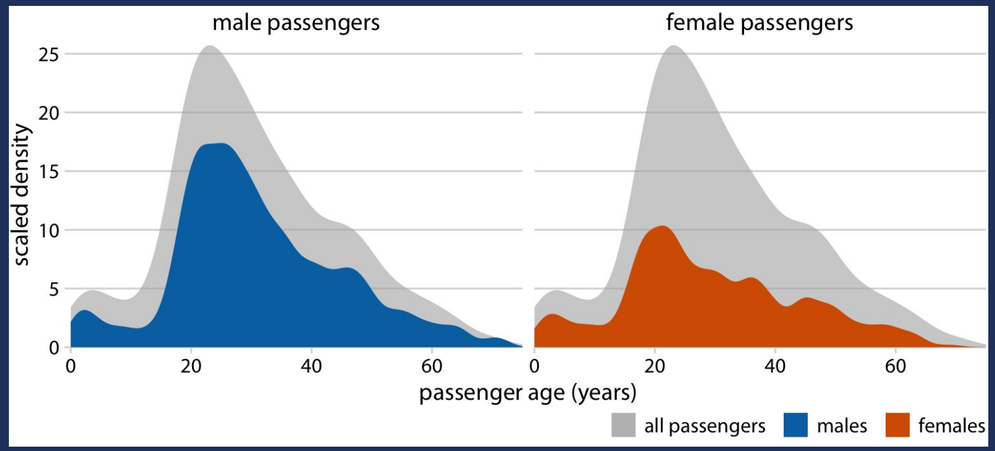

language
Vietnamese
Tongan
Tagalog
Serbocroatian
Samoan
Russian
Portuguese
Nepali
Navajo
Mon-Khmer, Cambodian
Mandarin
Laotian
Korean
Japanese
Italian
German
French
Dutch
Chinese
Arabic
0
2500
5000
7500
speakers

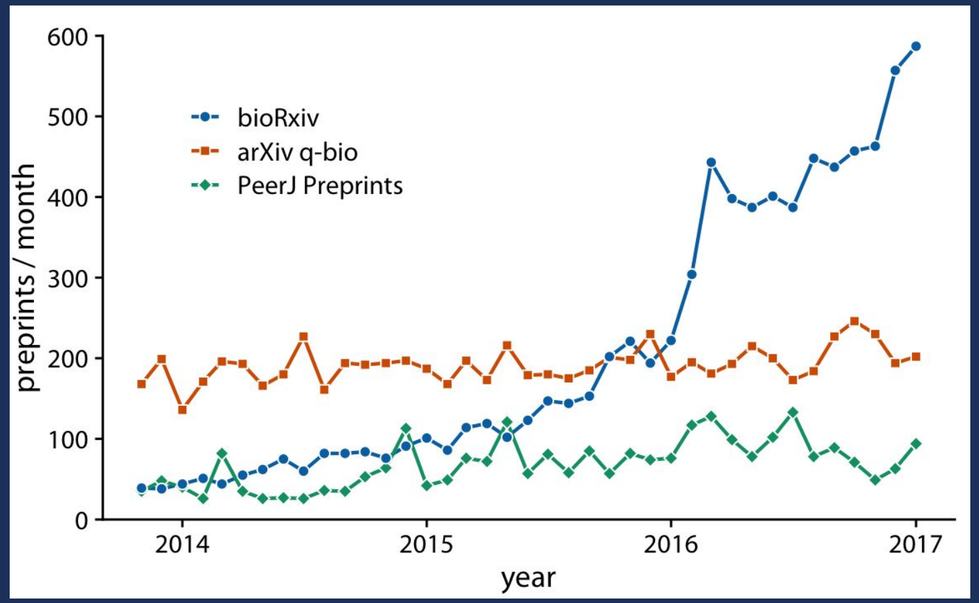

language
German
Navajo
Chinese
Portuguese
French
Vietnamese
Tongan
Samoan
Korean
Japanese
Tagalog
Russian
Italian
Serbocroatian
Arabic
Mandarin
Nepali
Mon-Khmer, Cambodian
Dutch
Laotian
0
2500
5000
7500
speakers

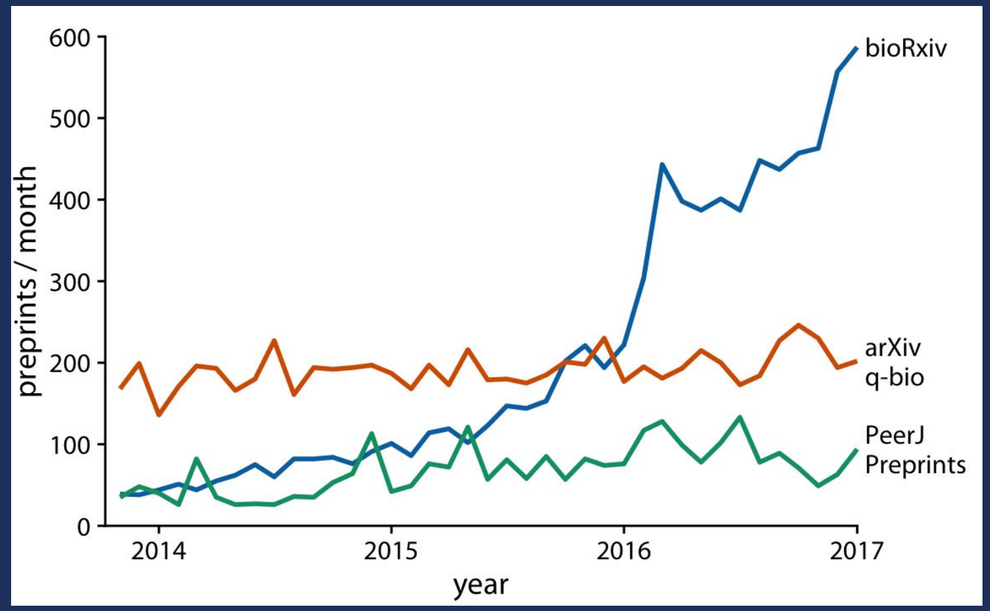

gender
female
male
count
age (years)
0
20
40
60
80

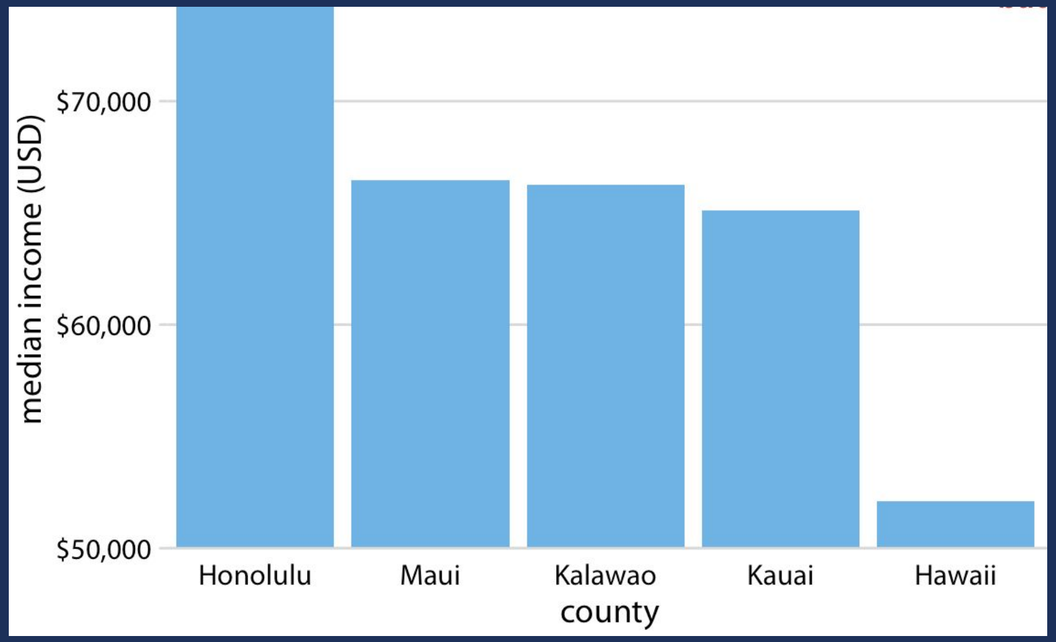

gender
male
female
scaled density
age (years)
0
5
10
15
20
40
60

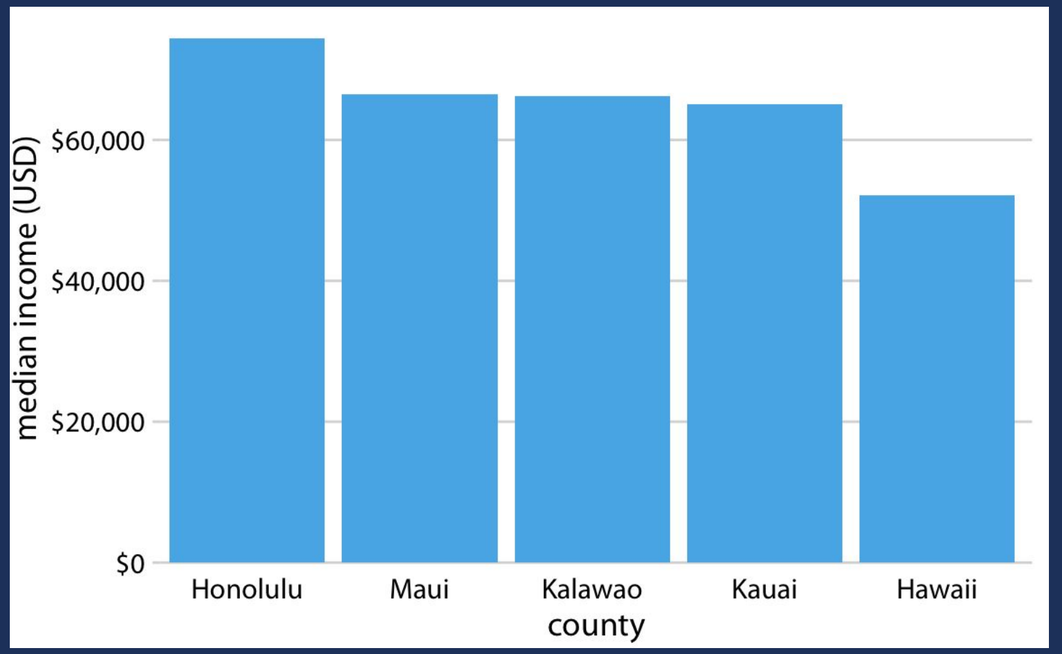

male
female
age (years)
count
0
20
40
60
40
20
0
20


male passengers
female passengers
scaled density
passenger age (years)
all passengers
males
females
0
5
10
15
20
25
20
40
60


600
500
400
300
200
100
0
preprints / month
bioRxiv
arXiv q-bio
PeerJ Preprints
2014
2015
2016
2017
year


600
500
400
300
200
100
0
preprints / month
bioRxiv
arXiv
q-bio
PeerJ
Preprints
2014
2015
2016
2017
year


median income (USD)
$70,000
$60,000
$50,000
Honolulu
Maui
Kalawao
Kauai
Hawaii
county


median income (USD)
$60,000
$40,000
$20,000
$0
Honolulu
Maui
Kalawao
Kauai
Hawaii
county


stock price (USD)
135
130
125
120
115
110
Nov 1, 2016
Dec 1, 2016
Jan 1, 2017


stock price (USD)
150
100
50
0
Nov 1, 2016
Dec 1, 2016
Jan 1, 2017

A WORD ABOUT FILE FORMATS

Bitmap (raster) formats:

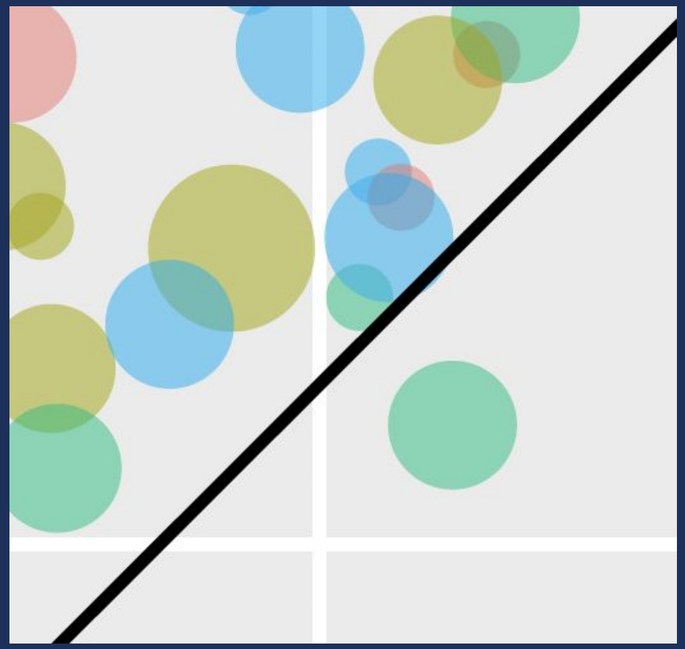
- JPEG, GIF, TIFF, BMP, …

Vector formats:

- PDF, SVG, EPS

Be deliberate about height/width (aspect ratio)!

JPEG vs. PDF



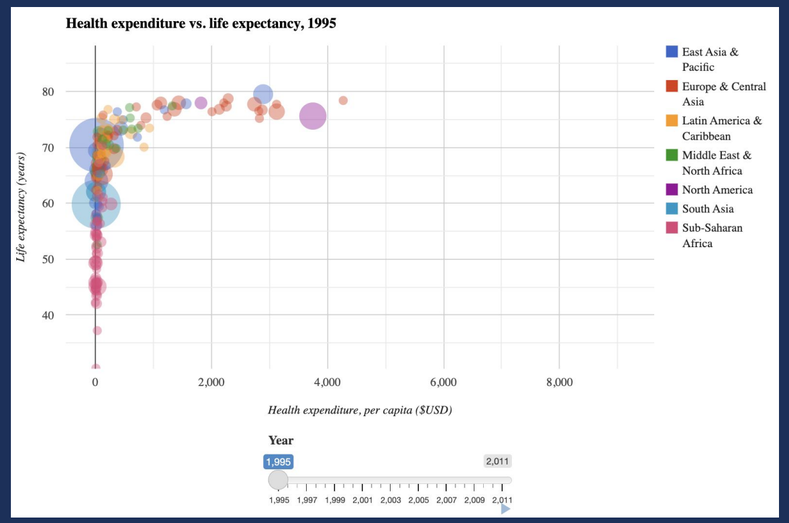
WHICH VISUALIZATION SOFTWARE TO USE?

Consider:

- Reproducibility
  - Same data + same transformations → Substantially similar visualization (conveys the same message)
- Repeatability
  - Same data + same transformations, EXACTLY the same visualization

INTERACTIVE VISUALIZATIONS

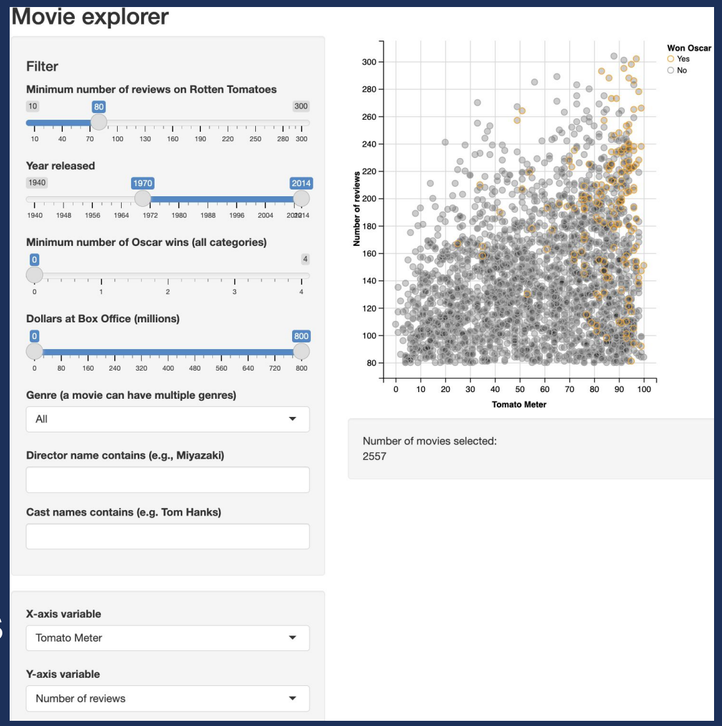
Users can engage with data to manipulate & explore data



Slide through time

Filter (declutter)

Explore different relationships



WORKSHOP OUTLINE

Our content today is divided into four parts. Each part will be described with examples.

01 Data Visualization Process

Walking through the visualization process, from setting goals to visualizing the data.

02 Basic Design Considerations

Best practices and accessibility considerations in data visualization.

03 Solutions to Common Problems

How to resolve, or even better, avoid common problems in data visualization.

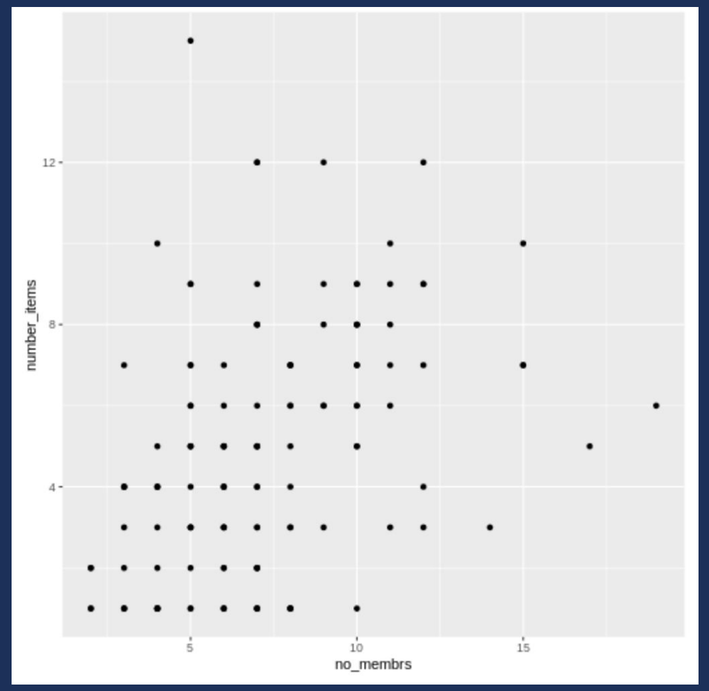
04 Visualization Resources

Workshops, tools, and other resources for data visualization.

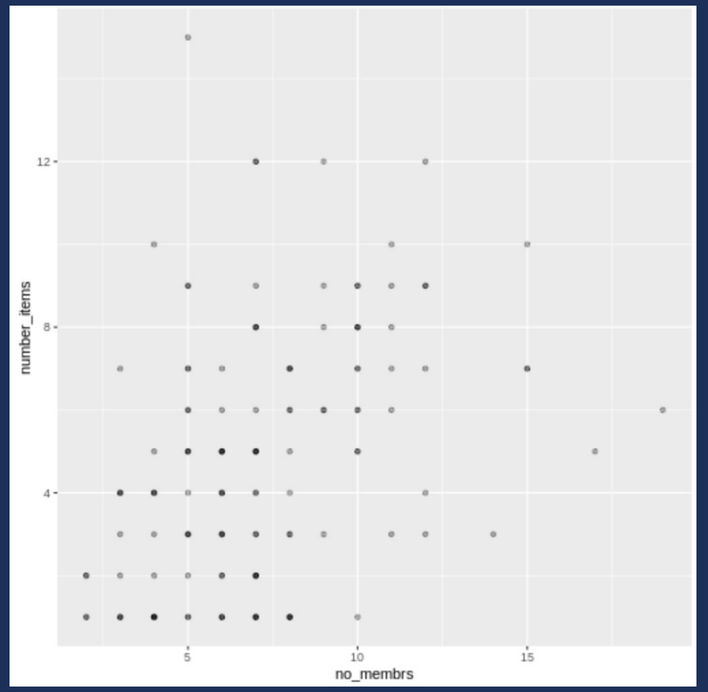
SOLUTIONS TO SOME PROBLEMS

Problem #1: Overlapping data points

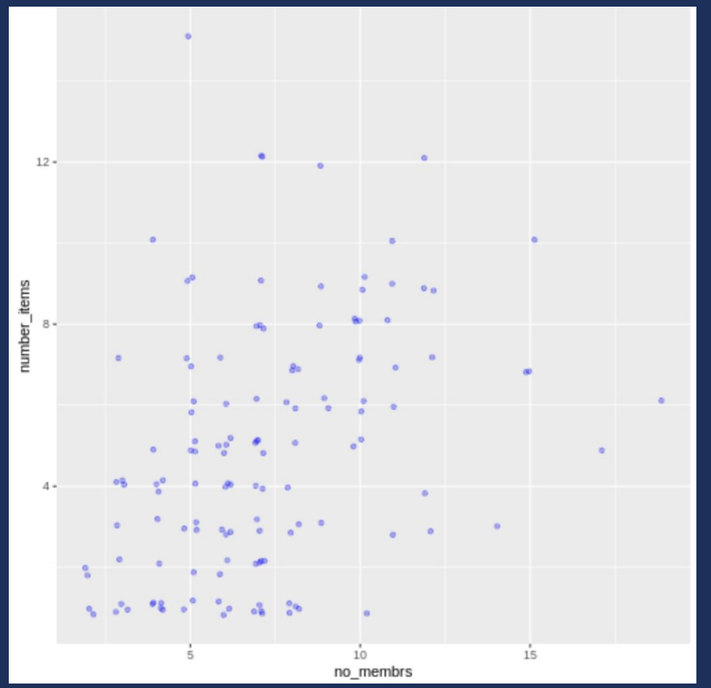
Problem: Overlapping Data Points



Partial Transparency
(a little better)

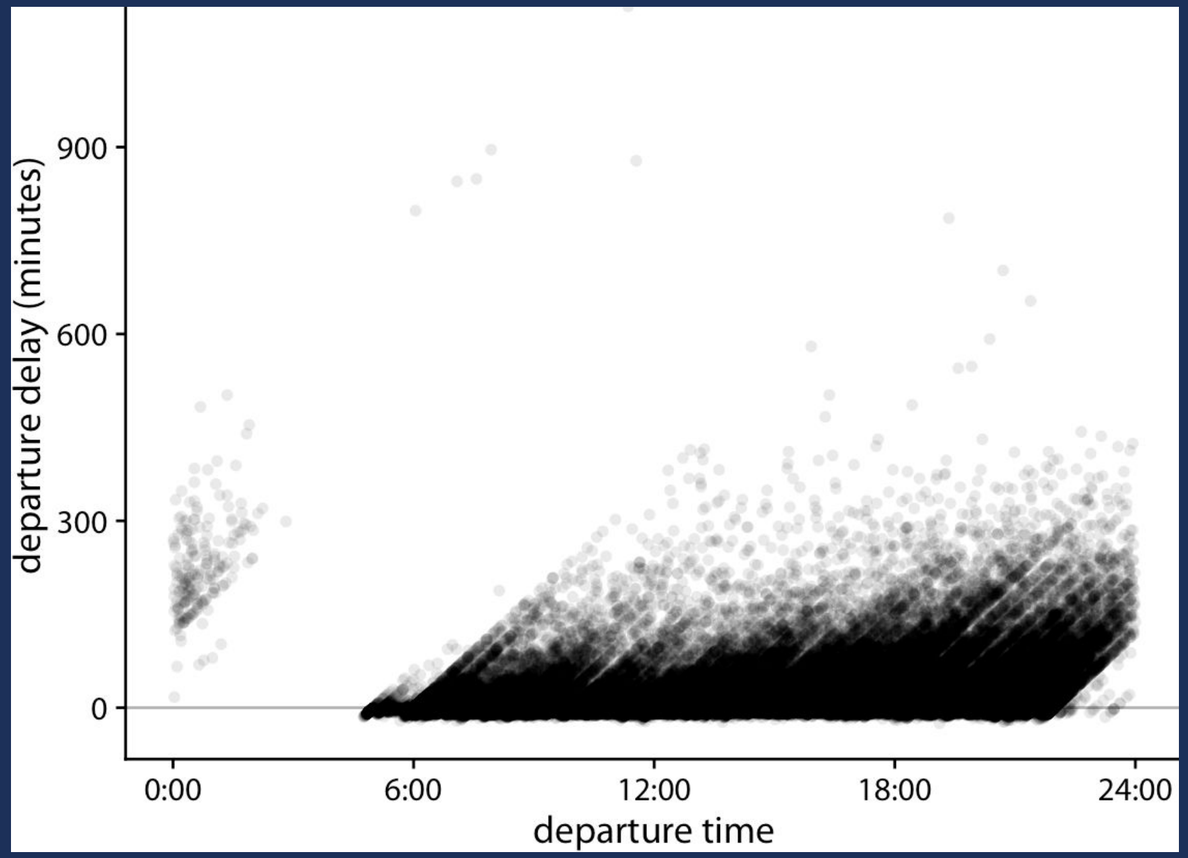


Partial Transparency
Plus Jitter (much better)

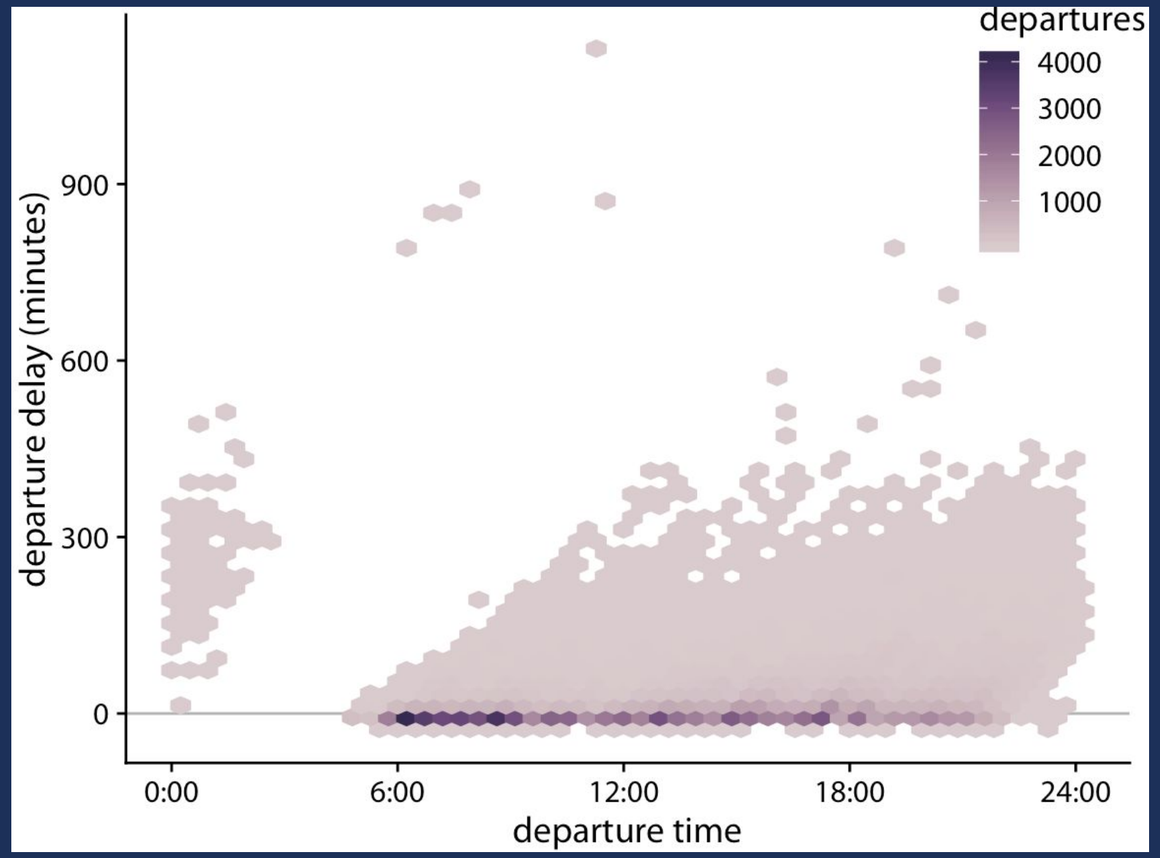


Problem: Overlapping Data Points

but with a LOT of data

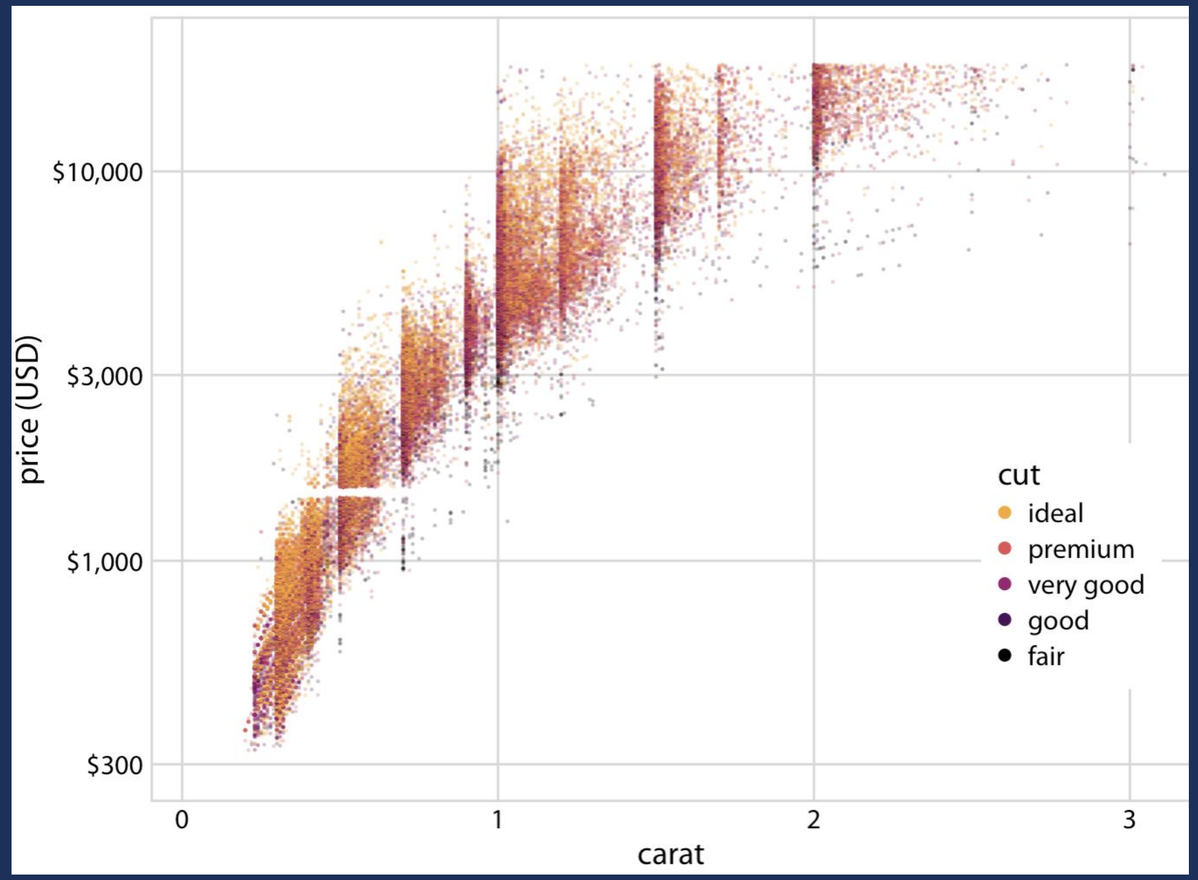


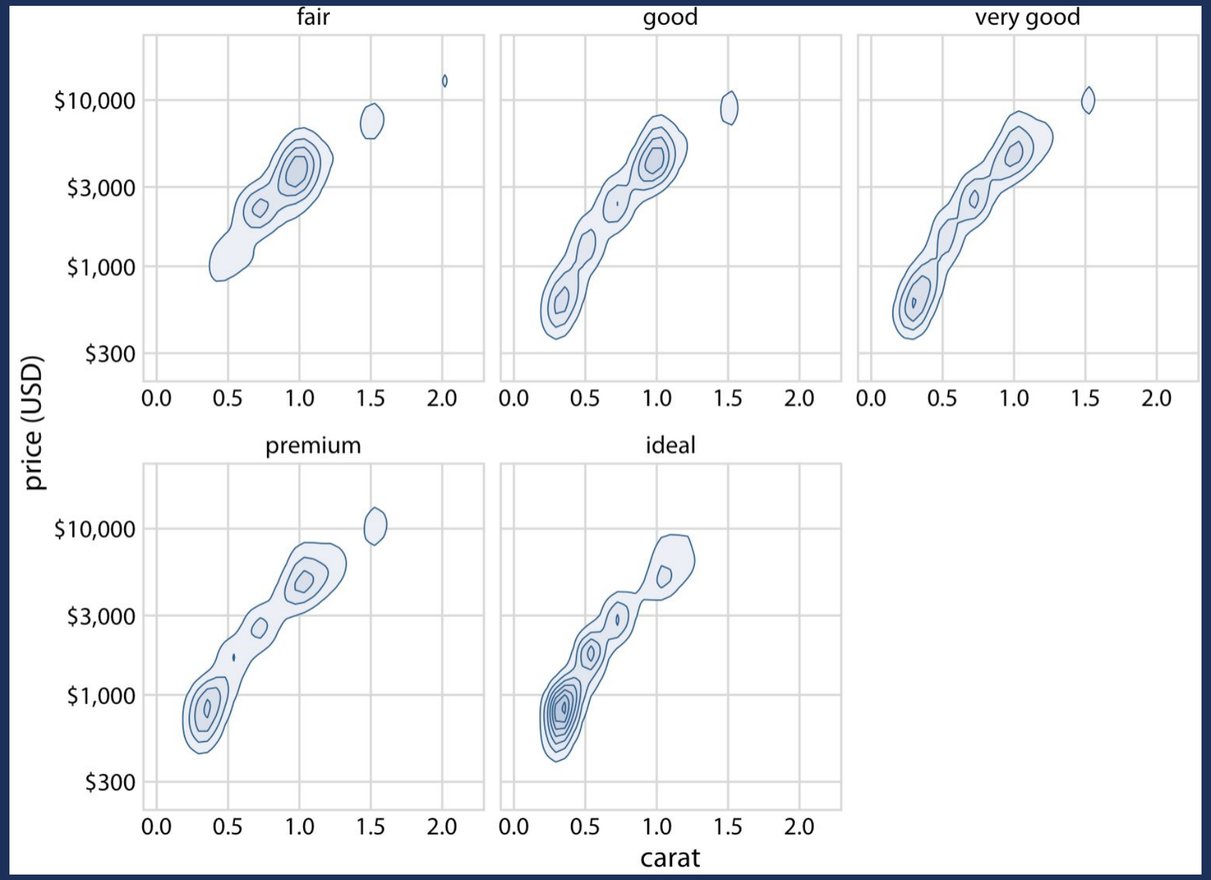
with 2-D binning



Problem: Overlapping Data Points

but with a LOT of data

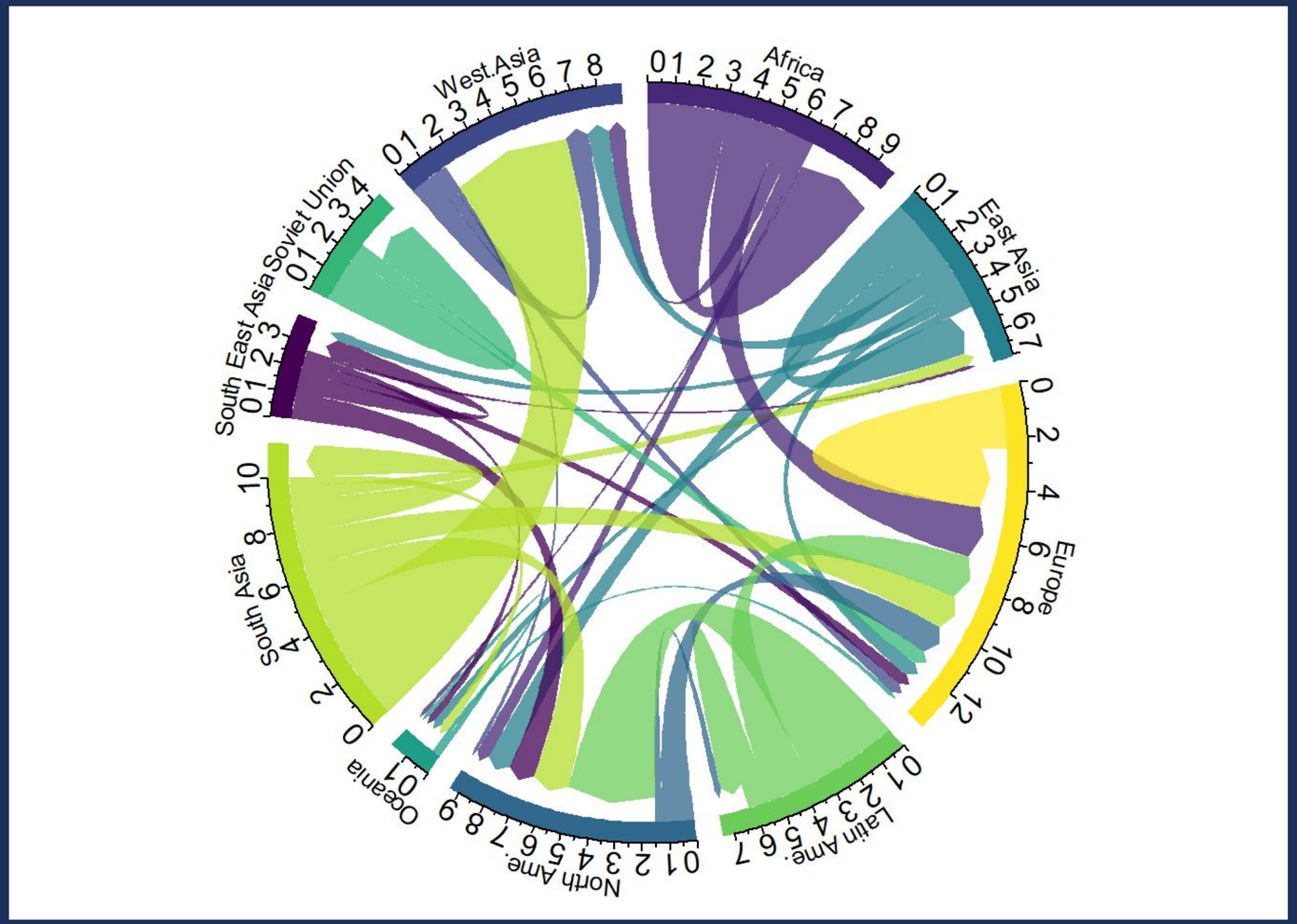




Using density contour lines + faceting

	Africa	East Asia	Europe	Latin Ame.	North Ame.	Oceania	South Asia
Africa	3.142471	0.000000	2.107883	0.000000	0.540887	0.155988	0.0000
East Asia	0.000000	1.630997	0.601265	0.000000	0.973060	0.333608	0.0000
Europe	0.000000	0.000000	2.401476	0.000000	0.000000	0.000000	0.0000
Latin Ame.	0.000000	0.000000	1.762587	0.879198	3.627847	0.000000	0.0000
North Ame.	0.000000	0.000000	1.215929	0.276908	0.000000	0.000000	0.0000
Oceania	0.000000	0.000000	0.170370	0.000000	0.000000	0.190706	0.0000
South Asia	0.000000	0.525881	1.390272	0.000000	1.508008	0.347420	1.3079
South East Asia	0.000000	0.145264	0.468762	0.000000	1.057904	0.278746	0.0000
Soviet Union	0.000000	0.000000	0.609230	0.000000	0.000000	0.000000	0.0000
West.Asia	0.000000	0.000000	0.449623	0.000000	0.169274	0.000000	0.0000

rowname	key	value
Oceania	Latin Ame.	0.000000
South Asia	Latin Ame.	0.000000
South East Asia	Latin Ame.	0.000000
Soviet Union	Latin Ame.	0.000000
West.Asia	Latin Ame.	0.000000
Africa	North Ame.	0.540887
East Asia	North Ame.	0.973060
Europe	North Ame.	0.000000
Latin Ame.	North Ame.	3.627847
North Ame.	North Ame.	0.000000
Oceania	North Ame.	0.000000
South Asia	North Ame.	1.508008
South East Asia	North Ame.	1.057904
Soviet Union	North Ame.	0.000000
West.Asia	North Ame.	0.169274
Africa	Oceania	0.155988
East Asia	Oceania	0.333608

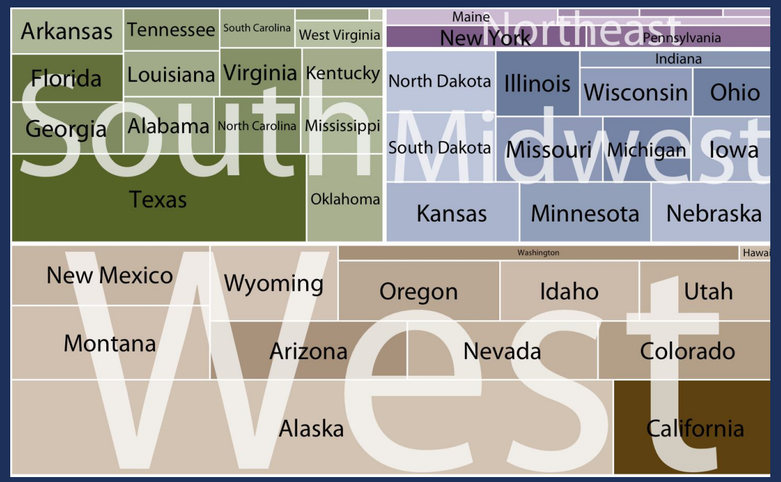


"R Gallery Book" bookdown.org/content/b298e479-b1ab-49fa-b83d-a57c2b034d49/

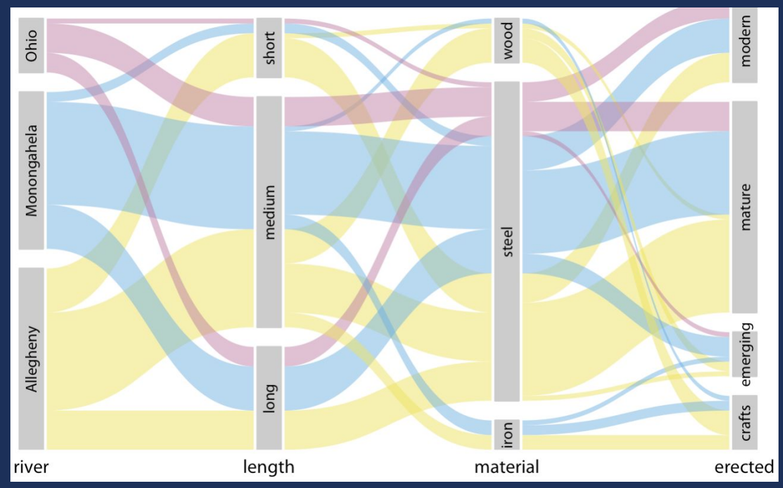
SOLUTIONS TO SOME PROBLEMS

Problem #2: Categories with subcategories

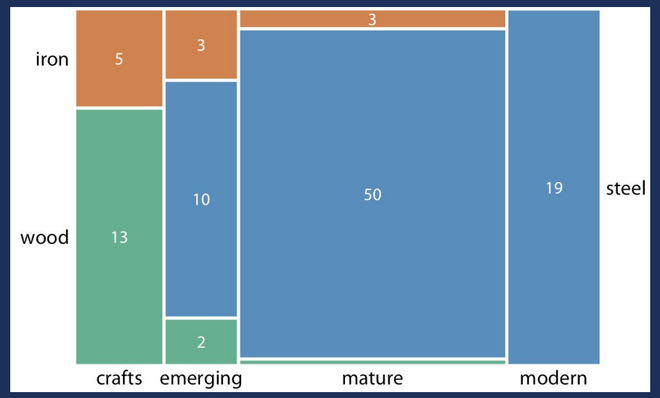
Problem: Categories with subcategories



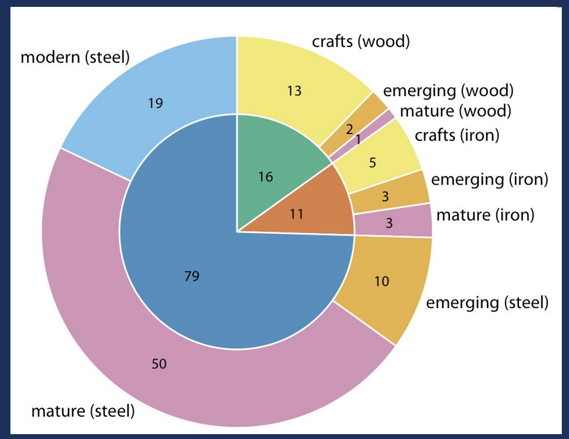
Treemap



Parallel Sets



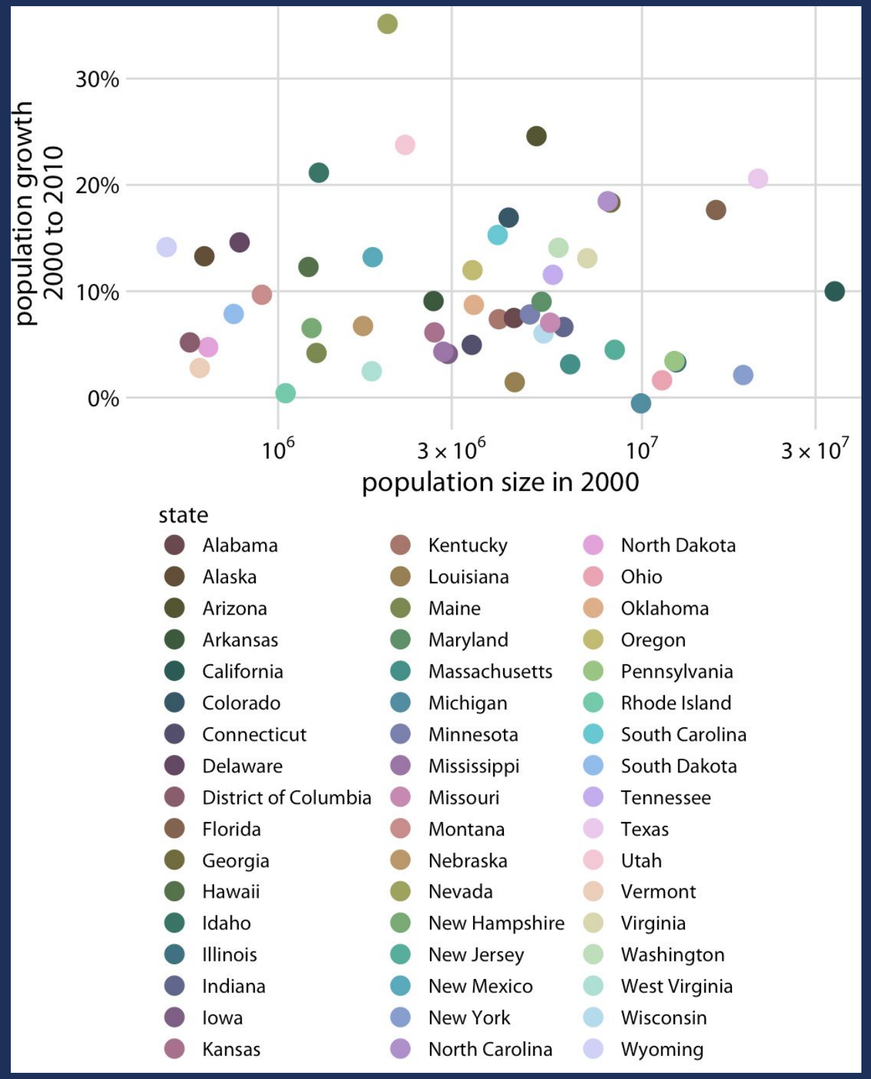
Mosaic Plot

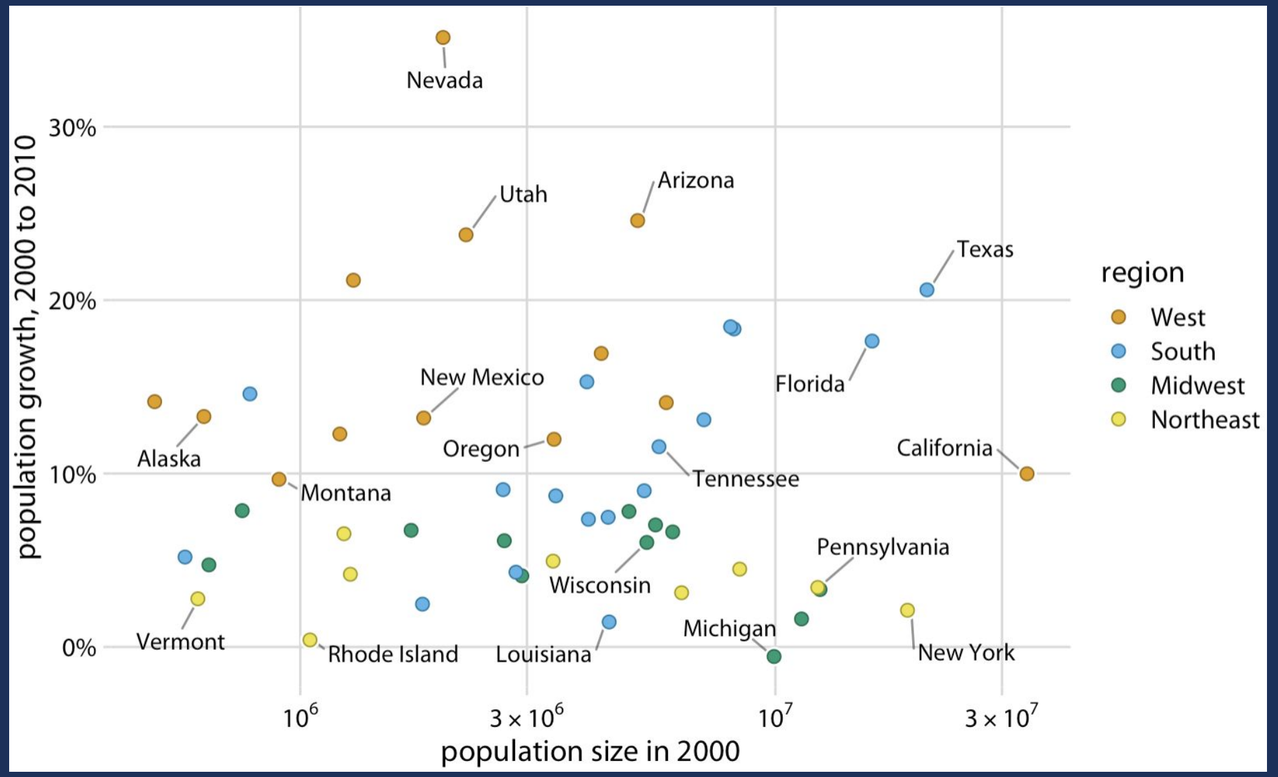


Nested Pie Chart

SOLUTIONS TO SOME PROBLEMS

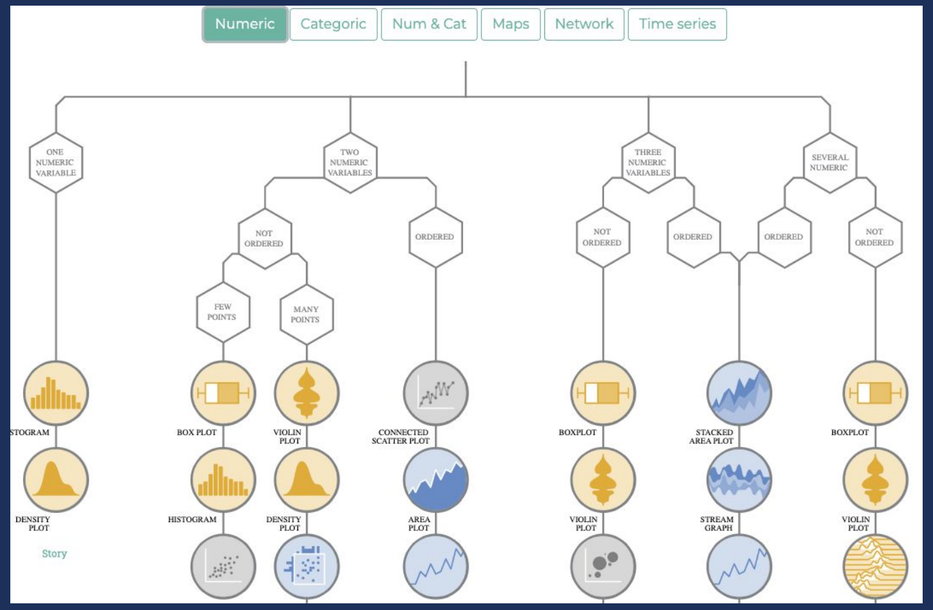
Problem #3: Too many categories/colors


population growth
2000 to 2010
30%
20%
10%
0%
10^6
3×10^6
10^7
3×10^7
population size in 2000
state
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

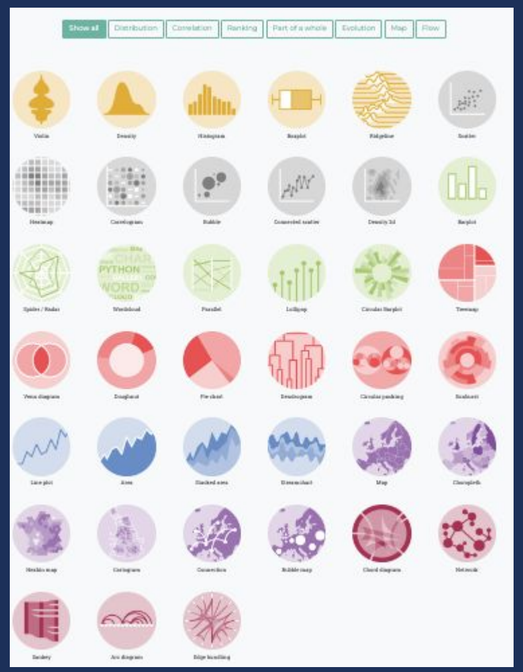

Nevada
Utah
Arizona
Texas
Florida
New Mexico
Oregon
Alaska
Montana
Tennessee
California
Pennsylvania
Wisconsin
Vermont
Rhode Island
Louisiana
Michigan
New York
population growth, 2000 to 2010
30%
20%
10%
0%
10^6
3×10^6
10^7
3×10^7
population size in 2000
region
West
South
Midwest
Northeast

from Data to Viz ~ data-to-viz.com

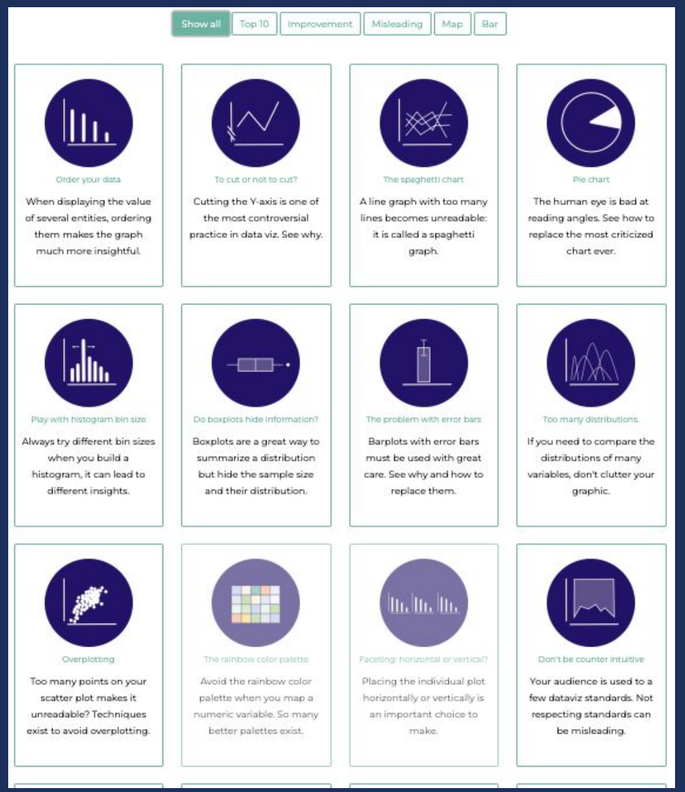
Decision Tree - suggested viz types



library of viz types



caveats



181 Beautiful Data Viz'es - dataviz-inspiration.com

WORKSHOP OUTLINE

Our content today is divided into four parts. Each part will be described with examples.

01 Data Visualization Process

Walking through the visualization process, from setting goals to visualizing the data.

02 Basic Design Considerations

Best practices and accessibility considerations in data visualization.

03 Solutions to Common Problems

How to resolve, or even better, avoid common problems in data visualization.

04 Visualization Resources

Workshops, tools, and other resources for data visualization.

RECOMMENDED RESOURCES

Click on titles to link to the library's holdings for each resource

Books

Data Points by Nathan Yau
ISBN: 9781118462195
Publication Date: 2013-04-15

Fundamentals of Data Visualization by Claus Wilke
ISBN: 1492031089
Publication Date: 2019-05-14

The Visual Display of Quantitative Information by Edward R. Tufte
ISBN: 0961392142
Publication Date: 2001-07-09

Information Visualization by Colin Ware
ISBN: 9780123814654
Publication Date: 2012-05-21

R

Data Visualisation with R by Thomas Rahlf
ISBN: 9783319497518
Publication Date: 2017-03-22

Ggplot2 by Hadley Wickham
ISBN: 9783319242774
Publication Date: 2016-06-08

Python

Python Data Visualization Cookbook - Second Edition by Igor Milovanovic; Dimitry Foures; Giuseppe Vettigli
ISBN: 9781784394943
Publication Date: 2015-11-30

CONSULTATIONS @ GWU

calendly.com/gwul-coding

- Python
- R
- HTML/CSS/JavaScript
- General coding questions

calendly.com/data-consultation

- General analysis and visualization questions
- Data management
- R, GIS, SAS, SPSS, Excel, Tableau

go.gwu.edu/dataconsulting

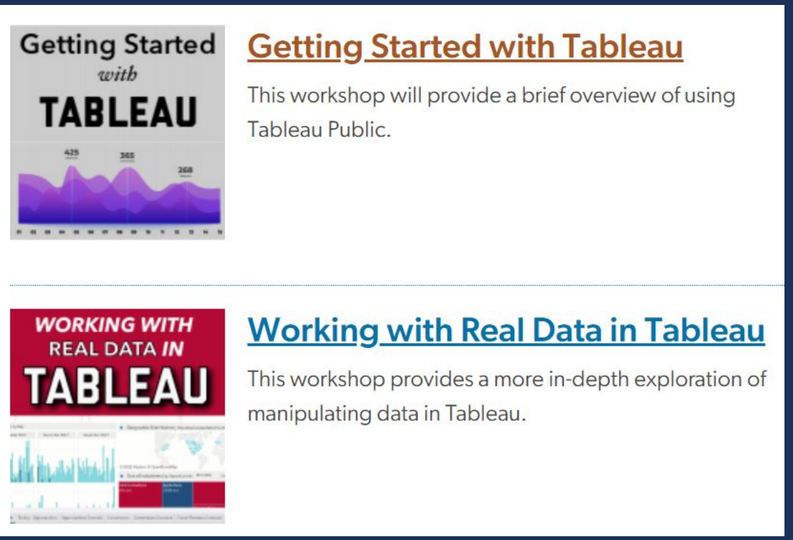
Graduate Student Data Consultants

- R, Python, SAS, SPSS, STATA, Excel

OTHER RESOURCES @ GWU

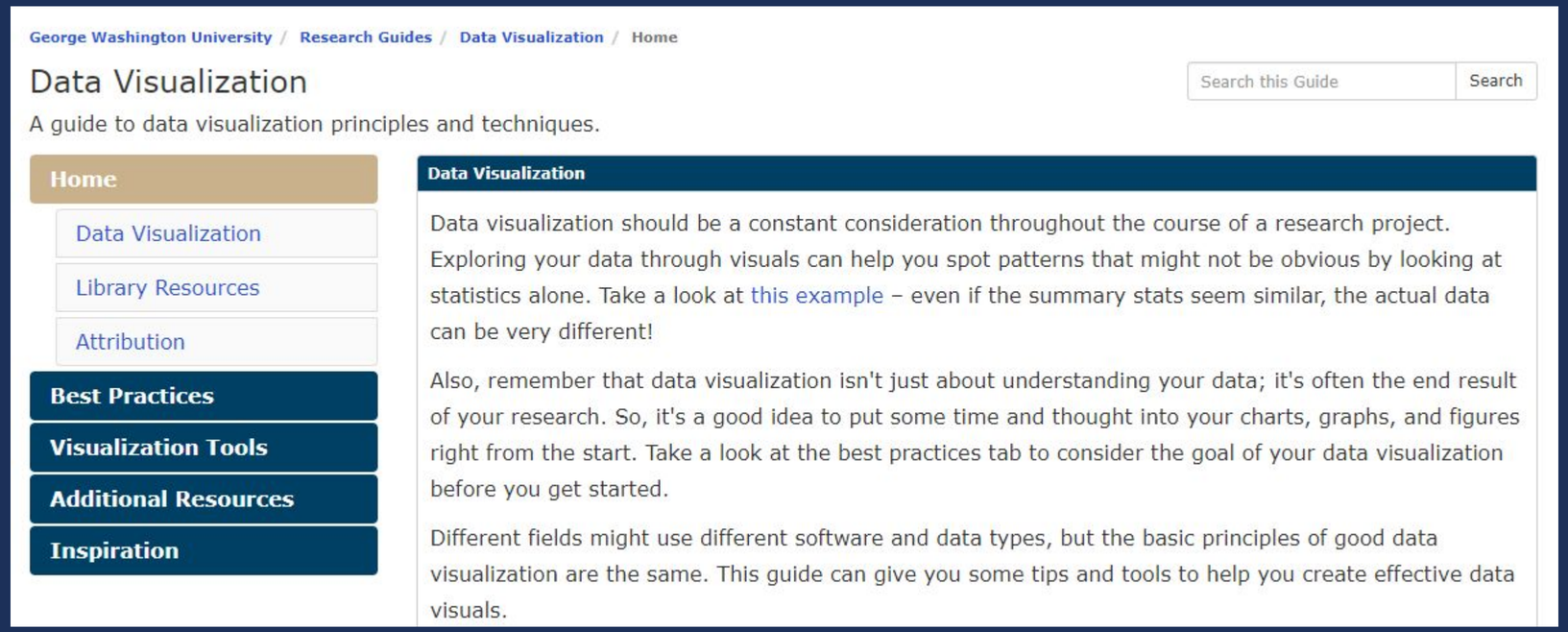
Library Workshops

https://library.gwu.edu/events

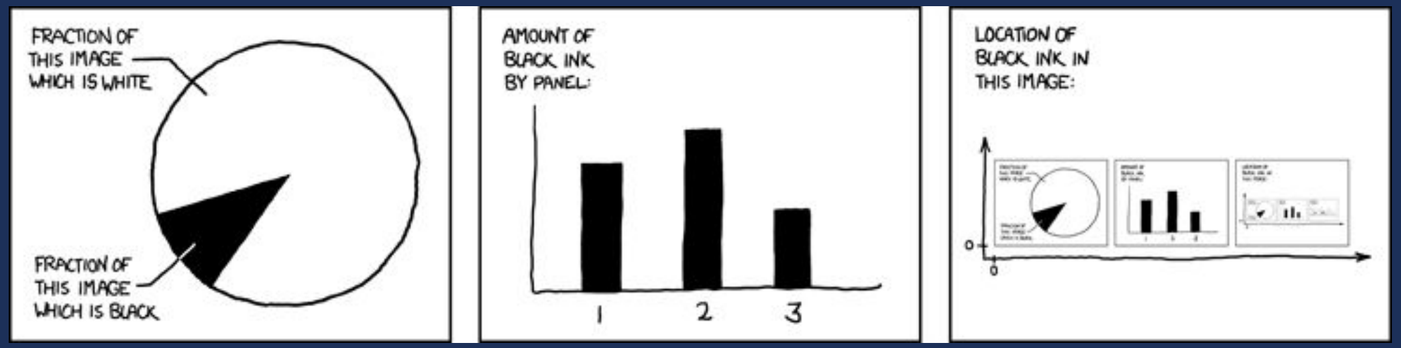


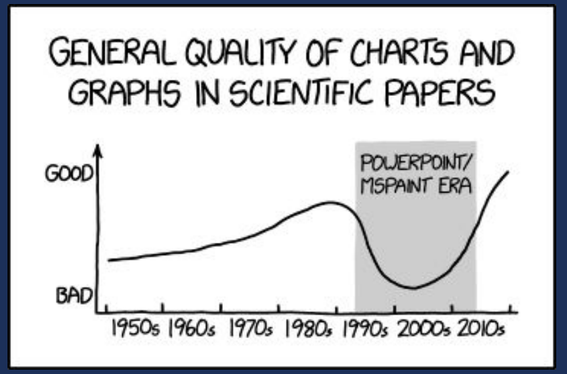
Library Research Guide

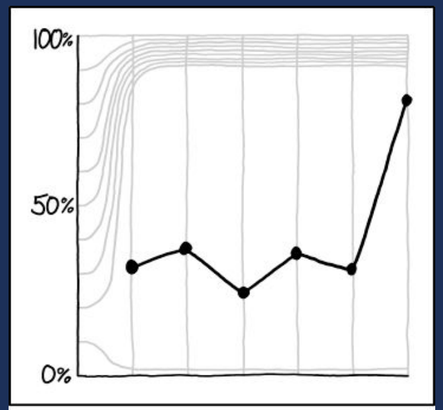
https://libguides.gwu.edu/c.php?g=1354705



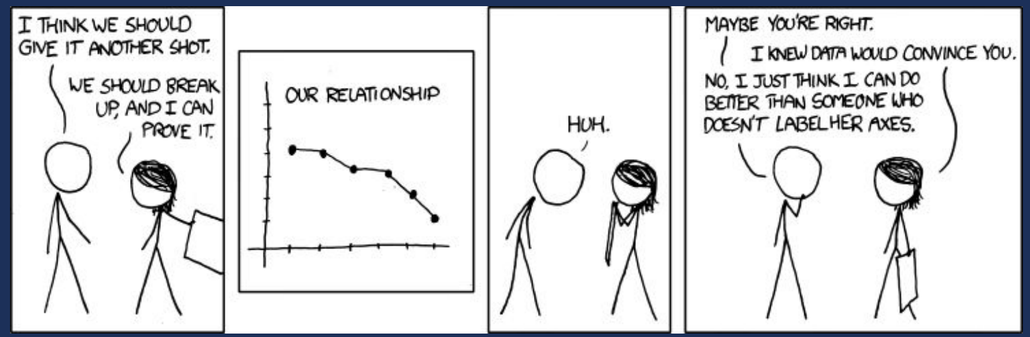
THANK YOU!

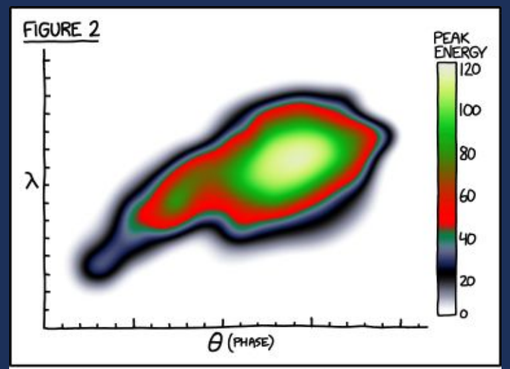




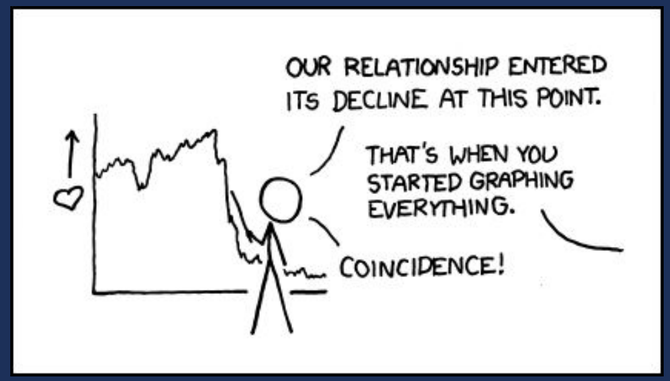


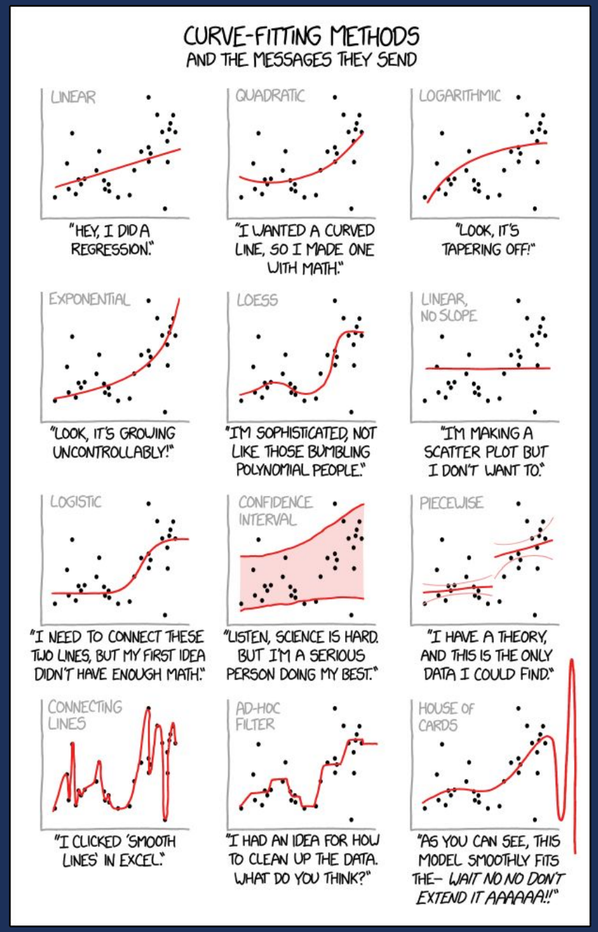
PEOPLE HAVE WISED UP TO THE "CAREFULLY CHOSEN Y-AXIS RANGE" TRICK, SO MISLEADING GRAPH MAKERS HAVE HAD TO GET CREATIVE.





EVERY YEAR, DISGRUNTLED SCIENTISTS COMPETE FOR THE PAINBOW AWARD FOR WORST COLOR SCALE.





Images are all from xkcd.com