Reminders

Upcoming due dates
Fri Jan 30 Project review (1 per group)
Discussion lab 3
Mon Feb 02 Quiz 4
Wed Feb 04 Project Proposal

Repo invites: Click accept in email!

Exploratory Data Analysis (and map making!)

Data Science in Practice

Jason G. Fleischer, PhD Dept. of Cognitive Science UC San Diego https://jgfleischer.com

Descriptive: The goal of descriptive analysis is to understand the components of a data set, describe what they are, and explain that description to others who might want to understand the data.

- Problem: Understanding whether users are nice or mean on Youtube
- Data science question: Are the words that people use in their comments more frequently positive words (great, awesome, nice, useful) or negative words (bad, stupid, lame, awful)?
- Type of analysis: Descriptive analysis

To answer this you would calculate <u>statistics</u> about YouTube comments

Statistics: The science that deals with the collection, classification, analysis, and interpretation of numerical facts or data

A statistic: a quantity computed from a sample

For our YouTube analysis, we could take a <u>random sample</u> of comments from YouTube and calculate the following statistic: *the number of positive and the number of negative words in each review*.

Population

We want to learn something about this

Sampling

Inference

Sample

....but we can only *actually* collect data from this

1 million comments from 2020

CC licensed clip art from pixabay.com

Best sampling practices:

- Always think about what your population is
- Collect data from a sample that is representative of your population
- If you have no choice but to work with a dataset that is not collected randomly and is biased, be careful not to generalize your results to the entire population

You'd want to be sure you sample randomly across *all* YouTube comments, making sure not to get more comments from one genre over another, or one location over another, etc.

Examples of bad sampling:

- Surveying subscribers of a gun-related magazine for research on Americans' attitudes toward owning guns
- Randomly sampling Facebook users for what TV shows people like

To understand *all* YouTube comments, you wouldn't just want to sample from one YouTube channel, or videos in a single language.

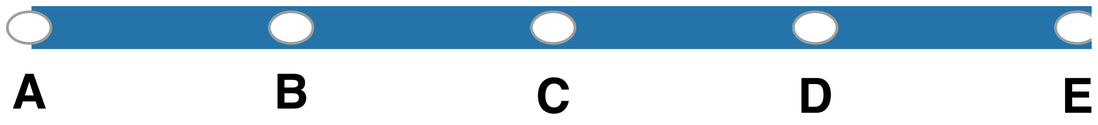
At the <u>beginning</u> of a project determine whether or not the data you have will actually help answer the question

Garbage In. Garbage Out.

Data Analysis

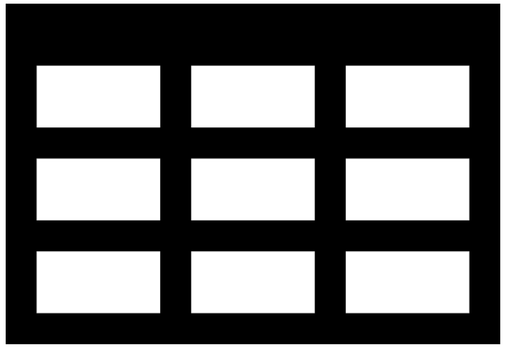
https://forms.gle/v9bn5kCUVLW5Robc6

For the survey data I collected from you all, which of the following best describes the population I could generalize findings back to.

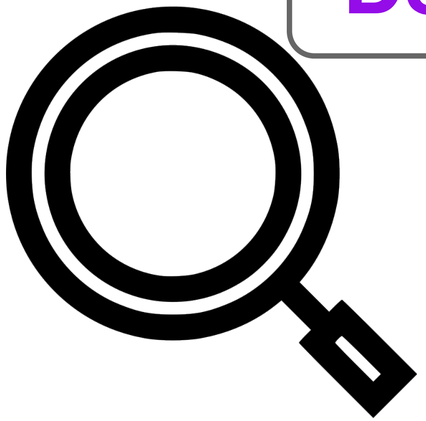


Descriptive

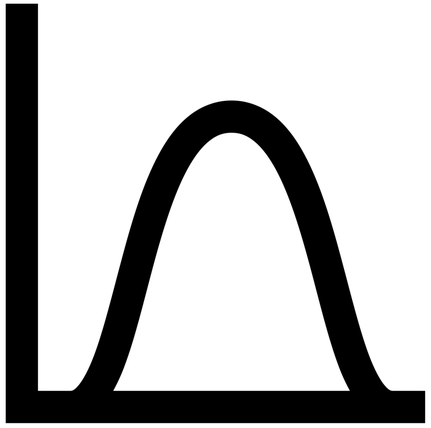
Descriptive Analysis



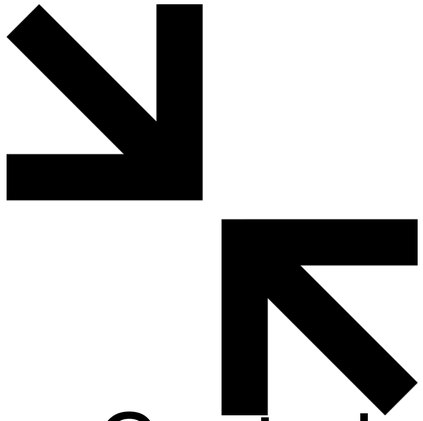
Size



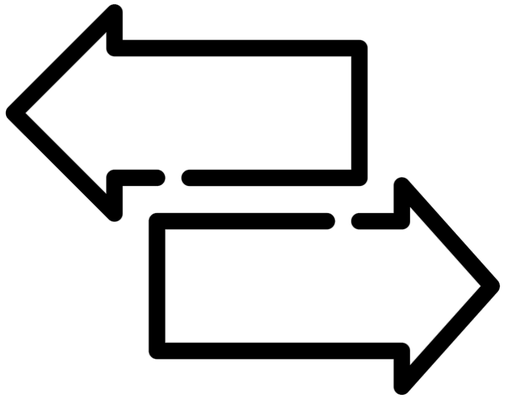
Missingness



Shape

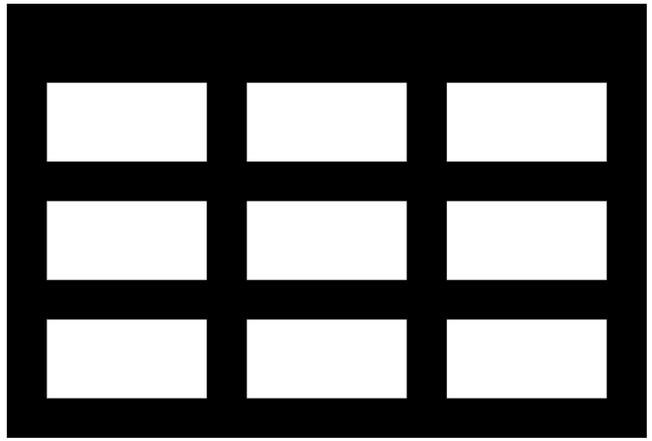


Central Tendency



Variability

Descriptive



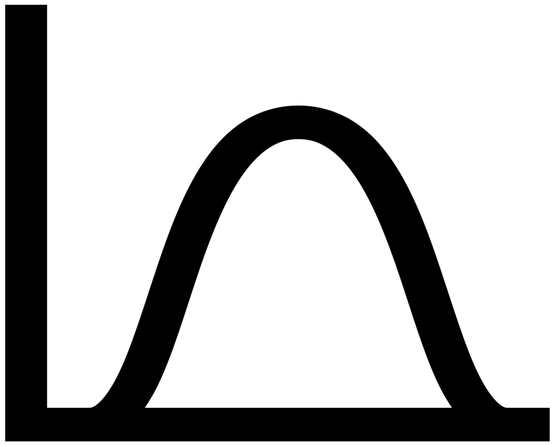
Size How many observations (rows) and variables (columns) you have is an important first step. You should always be aware of the size of your dataset

Descriptive

Missingness

It's critical to know how many observations have missing data for variables of interest in your data. Knowing *why* their missing is also important.

Descriptive



Shape

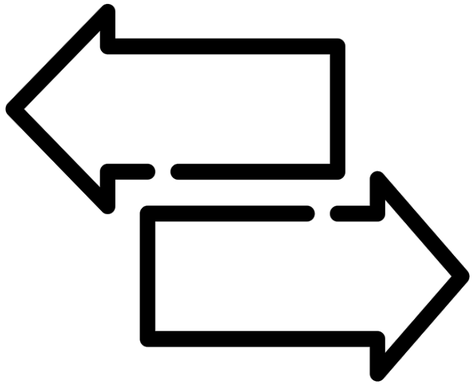
It's critical to know the distribution of the variables in your dataset. Certain statistical approaches can only be used with certain distributions.

Descriptive

Central Tendency

Knowing the mean, median, and/or mode can help you get an idea of what a typical value is for your variable(s) of interest

Descriptive



Variability

The central tendency tells you part of the story. The variability in the values in your observation helps fill in the rest.

Descriptive

Descriptive Analyses are often included as “Table 1” in academic publications

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Ranibizumab Monthly (N=301)	Bevacizumab Monthly (N=286)	Ranibizumab as Needed (N=298)	Bevacizumab as Needed (N=300)
Age — no. (%)				
50–59 yr	2 (0.7)	1 (0.3)	6 (2.0)	2 (0.7)
60–69 yr	33 (11.0)	28 (9.8)	31 (10.4)	34 (11.3)
70–79 yr	102 (33.9)	84 (29.4)	115 (38.6)	103 (34.3)
80–89 yr	142 (47.2)	150 (52.4)	126 (42.3)	142 (47.3)
≥90 yr	22 (7.3)	23 (8.0)	20 (6.7)	19 (6.3)
Mean — yr	79.2±7.4	80.1±7.3	78.4±7.8	79.3±7.6
Sex — no. (%)				
Female	183 (60.8)	180 (62.9)	185 (62.1)	184 (61.3)
Male	118 (39.2)	106 (37.1)	113 (37.9)	116 (38.7)
Race — no. (%)†				
White	297 (98.7)	281 (98.3)	296 (99.3)	294 (98.0)
Other	4 (1.3)	5 (1.7)	2 (0.7)	6 (2.0)
History of myocardial infarction — no. (%)	34 (11.3)	40 (14.0)	30 (10.1)	36 (12.0)
History of stroke — no. (%)	14 (4.7)	18 (6.3)	22 (7.4)	16 (5.3)
History of transient ischemic attack — no. (%)	12 (4.0)	25 (8.7)	12 (4.0)	19 (6.3)
Blood pressure — mm Hg				
Systolic	134±18	135±19	136±17	135±17
Diastolic	75±10	75±10	76±9	75±10
Visual-acuity score and Snellen equivalent				
68–82 letters, 20/25–40 — no. (%)	111 (36.9)	94 (32.9)	116 (38.9)	103 (34.3)
53–67 letters, 20/50–80 — no. (%)	98 (32.6)	118 (41.3)	108 (36.2)	119 (39.7)
38–52 letters, 20/100–160 — no. (%)	67 (22.3)	53 (18.5)	58 (19.5)	58 (19.3)
23–37 letters, 20/200–320 — no. (%)	25 (8.3)	21 (7.3)	16 (5.4)	20 (6.7)
Mean score	60.1±14.3	60.2±13.1	61.5±13.2	60.4±13.4
Total thickness at fovea — μm‡	458±184	463±196	458±193	461±175
Retinal thickness plus subfoveal-fluid thickness at fovea — μm	251±122	254±121	247±122	252±115
Foveal center involvement — no. (%)				
Choroidal neovascularization	176 (58.5)	153 (53.5)	176 (59.1)	183 (61.0)
Fluid	85 (28.2)	81 (28.3)	77 (25.8)	72 (24.0)
Hemorrhage	20 (6.6)	24 (8.4)	24 (8.1)	25 (8.3)
Other	18 (6.0)	20 (7.0)	15 (5.0)	18 (6.0)
No choroidal neovascularization or not possible to grade	2 (0.7)	8 (2.8)	6 (2.0)	2 (0.7)

* Plus–minus values are means ±SD.
† Race was self-reported.
‡ Total thickness at the fovea includes the retina, subretinal fluid, choroidal neovascularization, and retinal pigment epithelial elevation.

Descriptive

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Ranibizumab Monthly (N=301)	Bevacizumab Monthly (N=286)	Ranibizumab as Needed (N=298)	Bevacizumab as Needed (N=300)
Age — no. (%)				
50–59 yr	2 (0.7)	1 (0.3)	6 (2.0)	2 (0.7)
60–69 yr	33 (11.0)	28 (9.8)	31 (10.4)	34 (11.3)
70–79 yr	102 (33.9)	84 (29.4)	115 (38.6)	103 (34.3)
80–89 yr	142 (47.2)	150 (52.4)	126 (42.3)	142 (47.3)
≥90 yr	22 (7.3)	23 (8.0)	20 (6.7)	19 (6.3)
Mean — yr	79.2±7.4	80.1±7.3	78.4±7.8	79.3±7.6
Sex — no. (%)				
Female	183 (60.8)	180 (62.9)	185 (62.1)	184 (61.3)
Male	118 (39.2)	106 (37.1)	113 (37.9)	116 (38.7)
Race — no. (%)†				
White	297 (98.7)	281 (98.3)	296 (99.3)	294 (98.0)
Other	4 (1.3)	5 (1.7)	2 (0.7)	6 (2.0)

* Plus–minus values are means ±SD.
† Race was self-reported.
‡ Total thickness at the fovea includes the retina, subretinal fluid, choroidal neovascularization, and retinal pigment epithelial elevation.

Size

Shape

Central tendency

variability

Zooming in on this we see variables stratified by Age, Sex, and Race

Stratification changes results

Sample: 400 patients with index vertebral fractures

…looks like vertebroplasty was *way* worse for patients!

Vertebroplasty	Conservative care	Relative risk (95% confidence interval)
30/200 (15%)	15/200 (7.5%)	2.0 (1.1–3.6)

subsequent fractures

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/

But wait...at time of initial fracture...

	Vertebroplasty N = 200	Conservative care N = 200
Age, y, mean ± SD	78.2 ± 4.1	79.0 ± 5.2
Weight, kg, mean ± SD	54.4 ± 2.3	53.9 ± 2.1
Smoking status, No. (%)	110 (55)	16 (8)

Age and weight are similar between groups. **Smoking Status** differs vastly.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/

So...let's stratify those results real quick

Smoke			No smoke		
Vertebroplasty	Conservative	RR (95% confidence interval)	Vertebroplasty	Conservative	RR (95% confidence interval)
23/110 (21%)	3/16 (19%)	1.1 (0.4, 3.3)	7/90 (8%)	12/184(7%)	1.2 (0.5, 2.9)

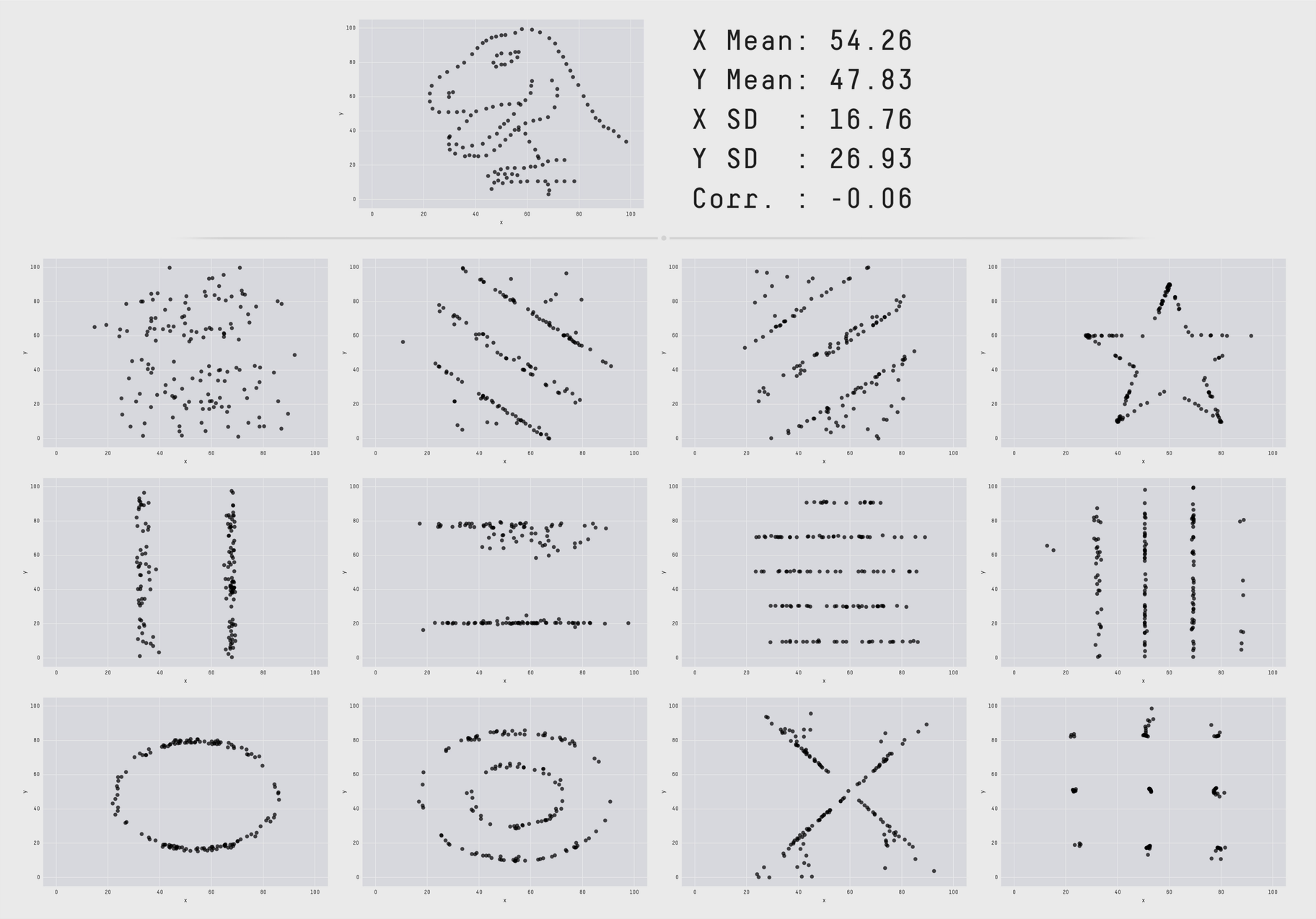
Risk of re-fracture is now similar within group

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/

Descriptive Statistics & Summary

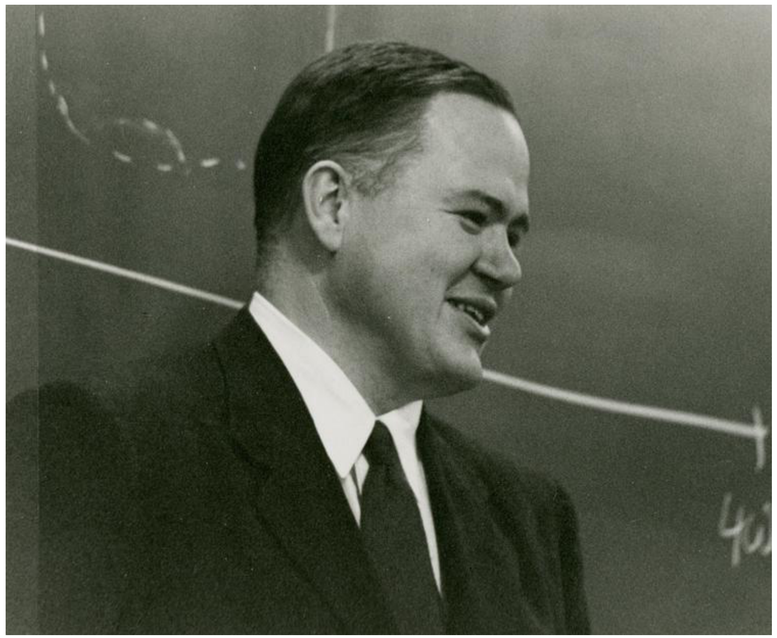
"*We must suppress some of the truth to communicate the truth... In short, the techniques of descriptive statistics are designed to match the salient features of the data set to human cognitive abilities.*"

-I.J. Good (1983)



https://www.autodesk.com/research/publications/same-stats-different-graphs

A modern version of Anscombe's Quartet



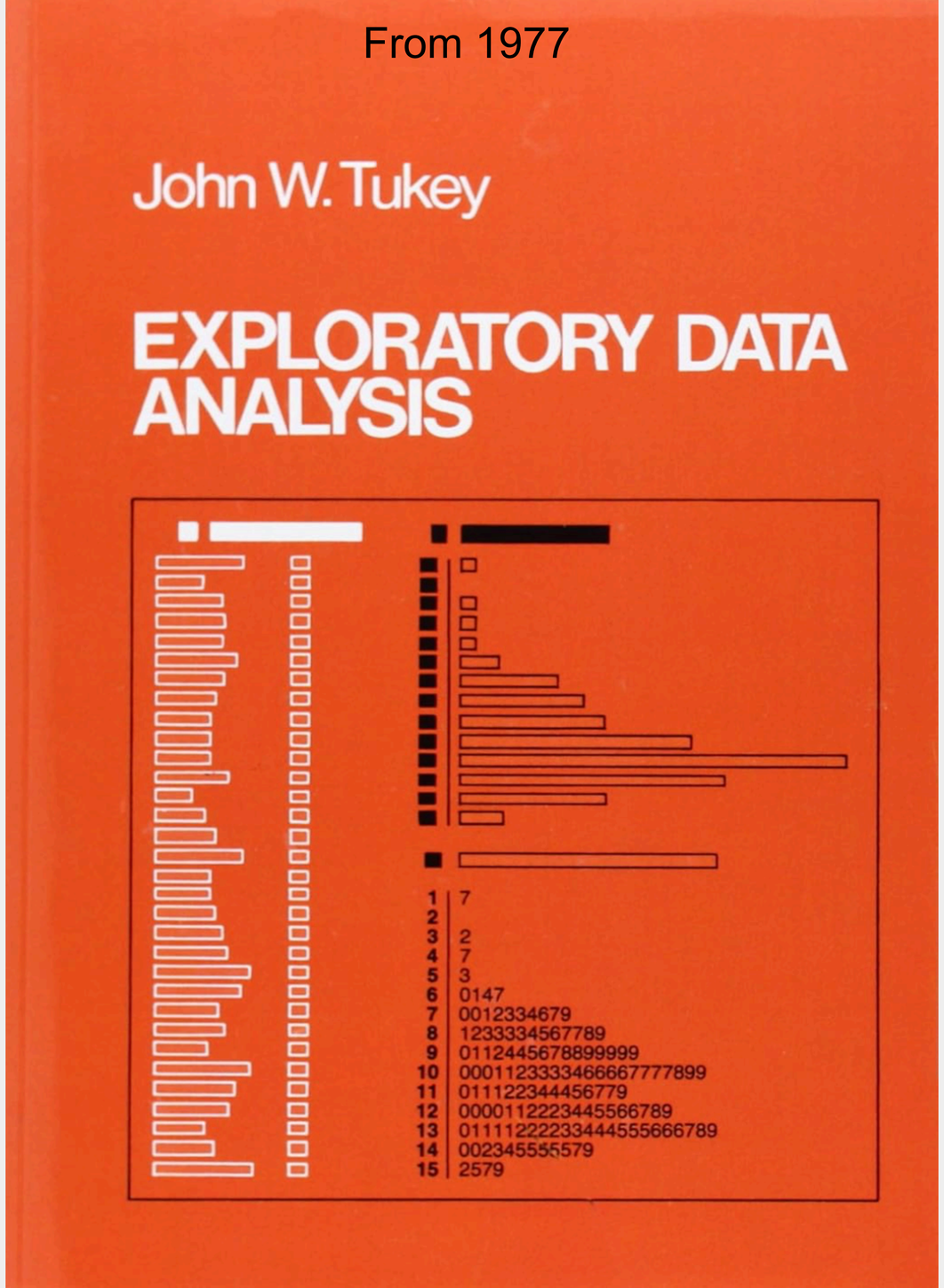
John Tukey

1915 — 2000

Was maybe the first person to be what we now call a data scientist

Cool things he did:
FFT, box plot, HSD, coined the terms bit and software

From 1977

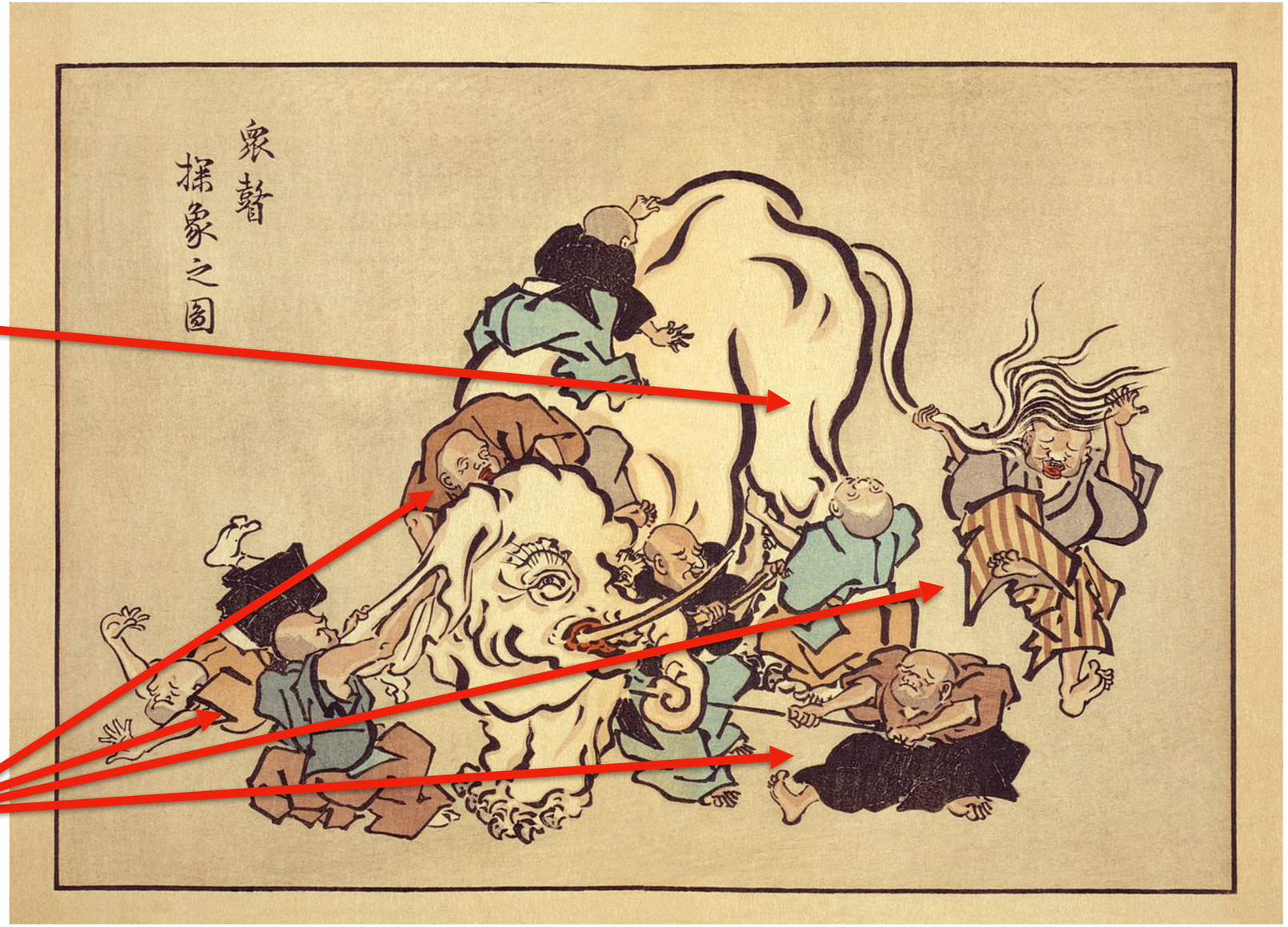


What do we get from EDA?

Exploratory Data Analysis

- Understand data properties
- Discover Patterns (You get to play detective!)
- Generate & Frame Hypothesis
- Suggest modeling strategies
- Check assumptions (sanity checks)
- Communicate results (present the data)

The dataset



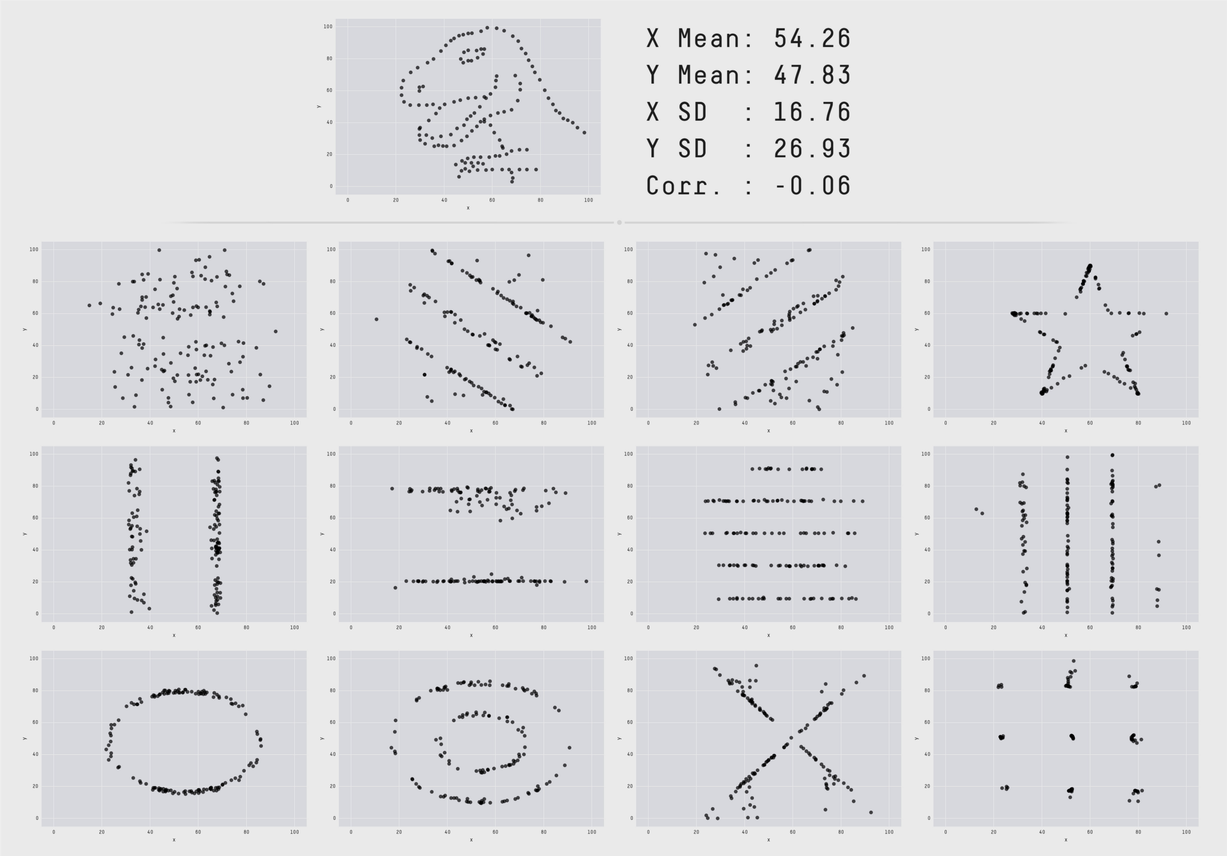
You

The general principles of EDA

- Generate plots and tables to explore relationships
- Prefer methods that do not assume a distribution (unless you've checked first that the data fit the assumption)
- If necessary, transform variables and generate more plots and tables

Start raw

- Examine raw data in the most direct way you can reasonably do so
- View a random sample of the data
- Plots, especially subsets of variables and dimensionality reduction
- Helpful for seeing weirdness, missingness, outliers, min/max/typical values

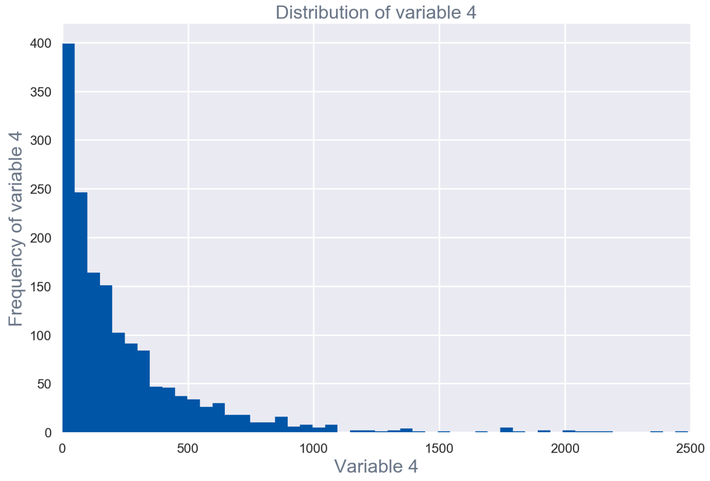


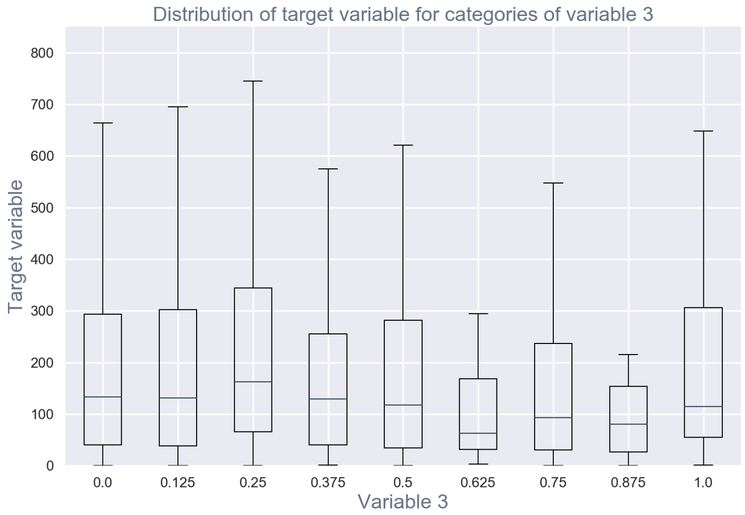
https://www.autodesk.com/research/publications/same-stats-different-graphs

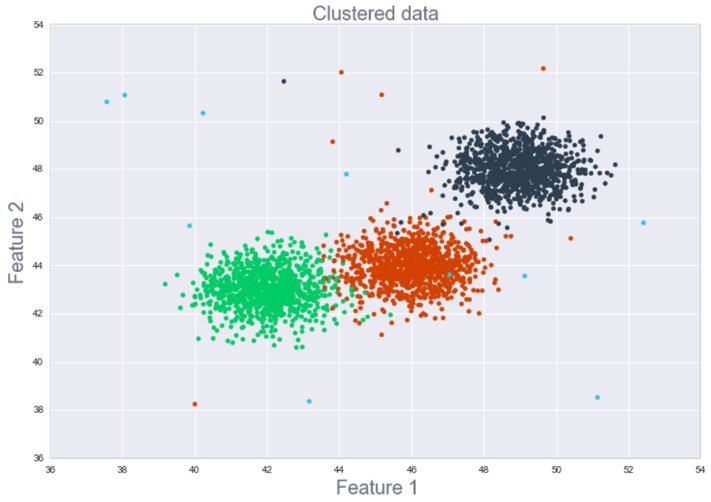
EDA Approaches to “Get a Feel for the Data”

Understanding the relationship between variables in your dataset

Exploratory







Univariate

understanding a single variable

i.e.: histogram, densityplot, barplot

Bivariate

understanding relationship between 2 variables

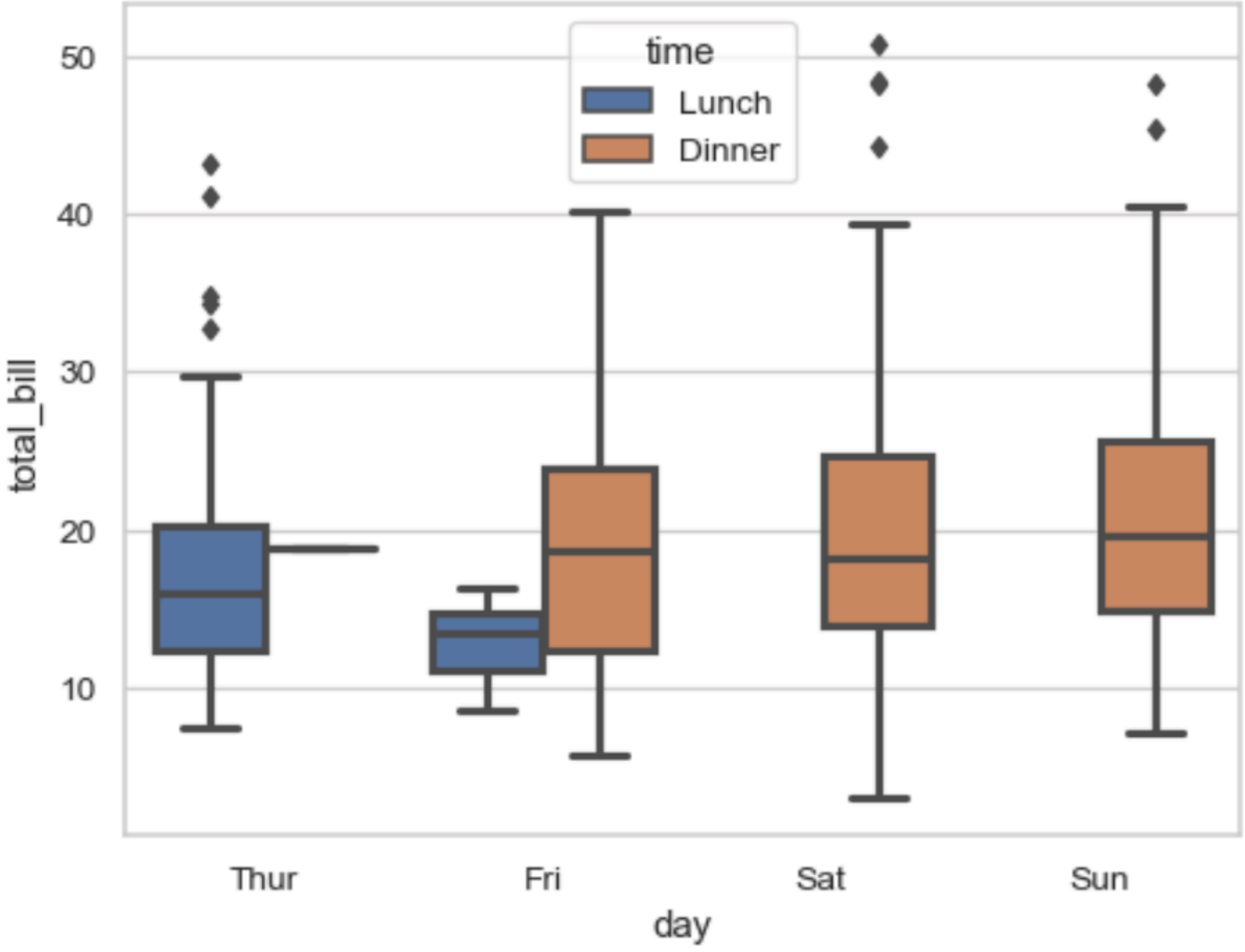
i.e.: boxplot, scatterplot, grouped barplot, boxplot

Dimensionality Reduction

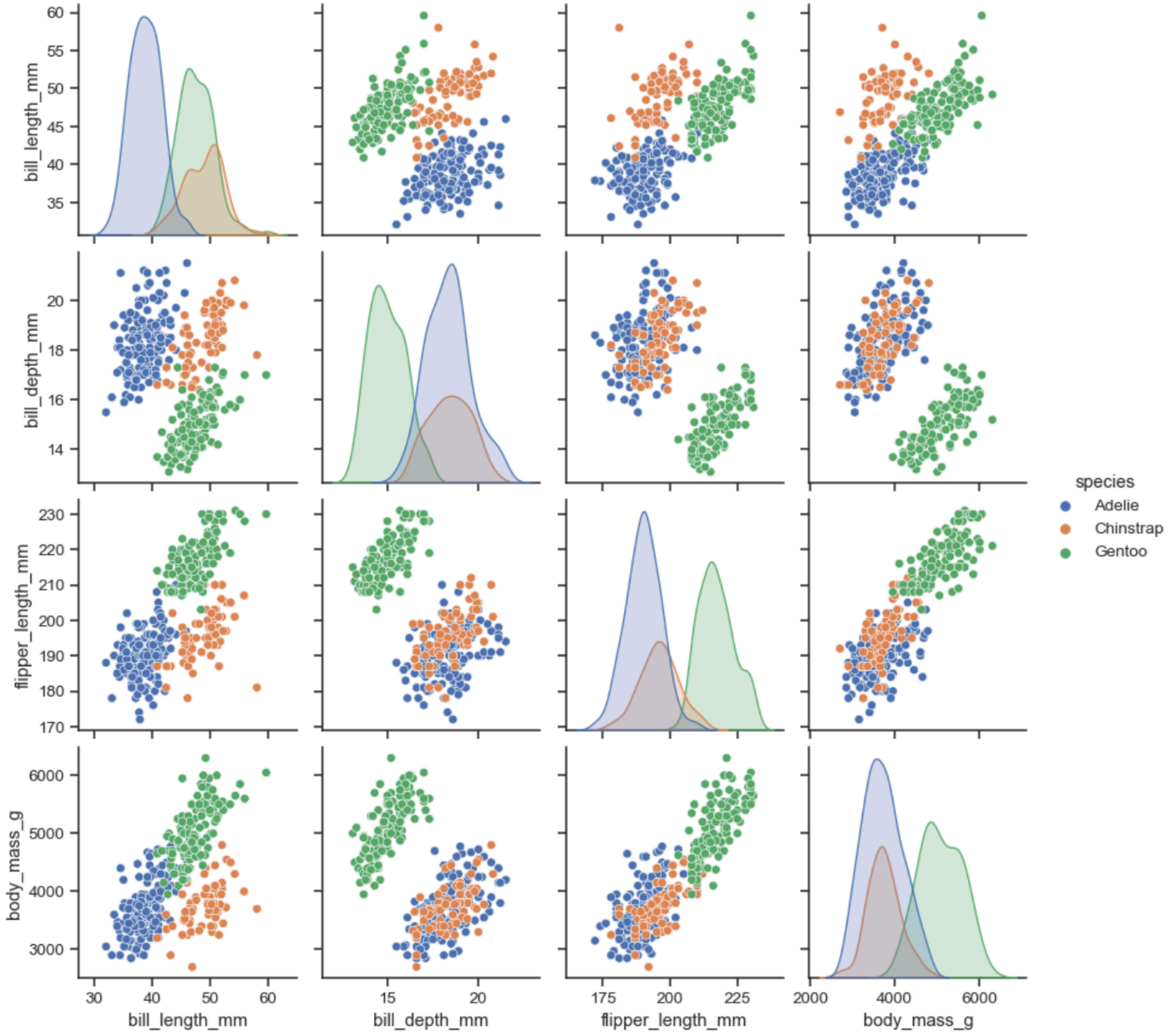
projecting high-D data into a lower-D space

i.e.: PCA, ICA, Clustering

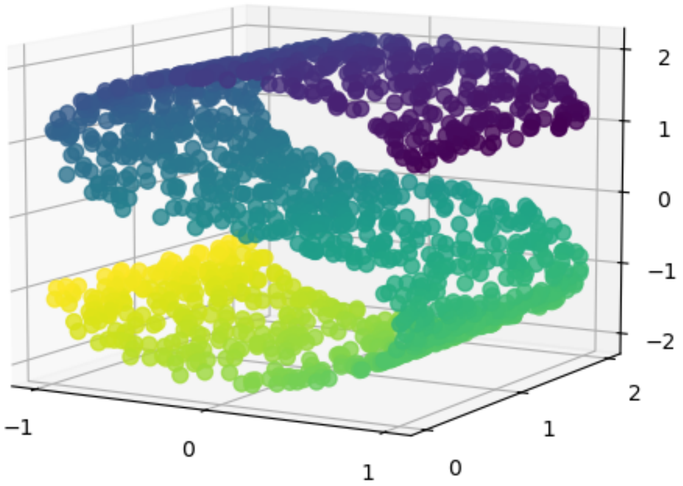
```
>>> ax = sns.boxplot(x="day", y="total_bill", hue="time",
...                  data=tips, linewidth=2.5)
```



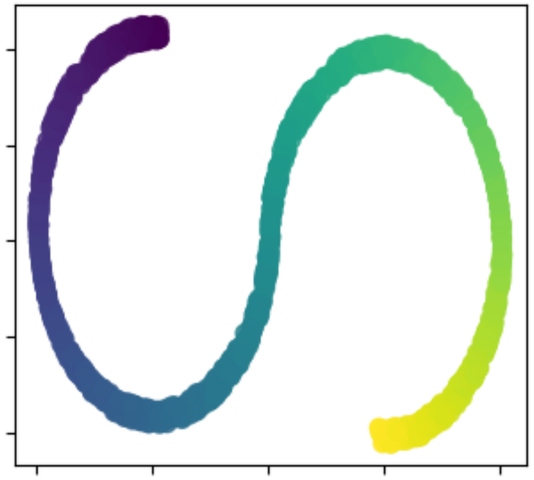
```
sns.pairplot(penguins, hue="species")
```



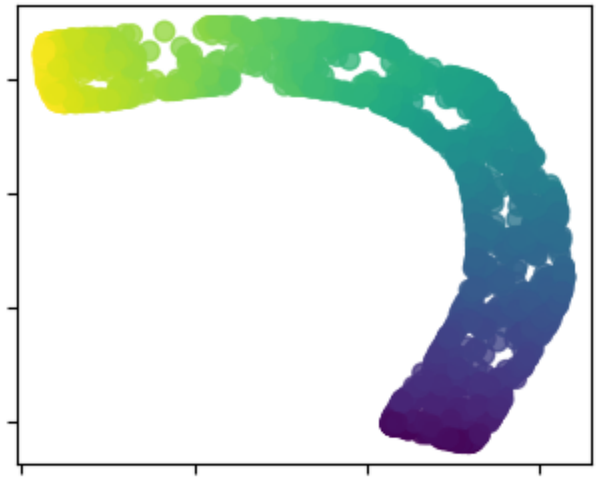
3D S-curve dataset



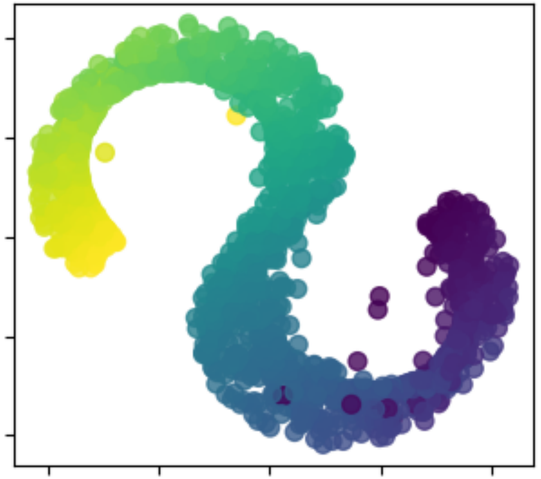
PCA

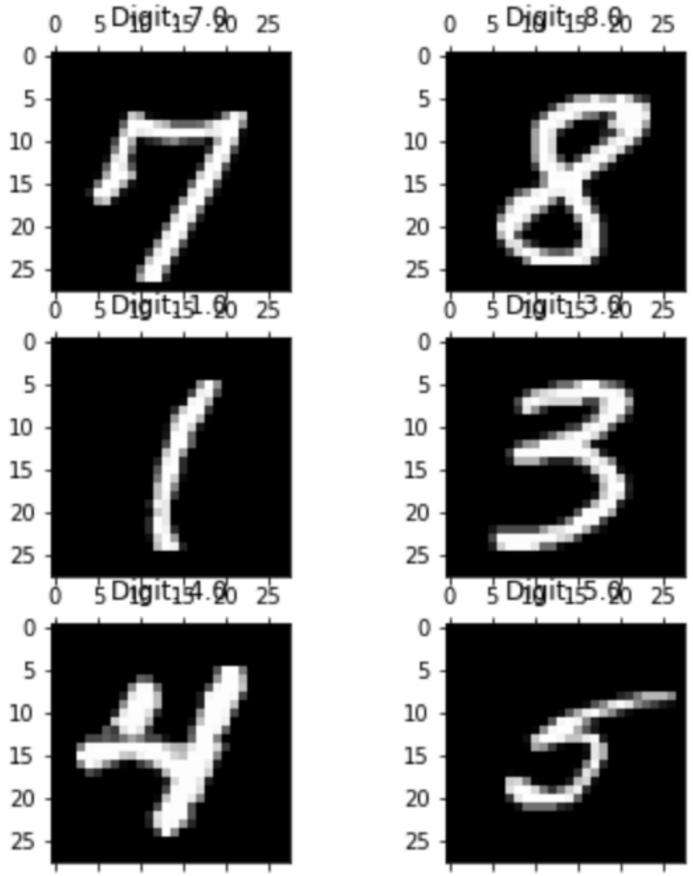


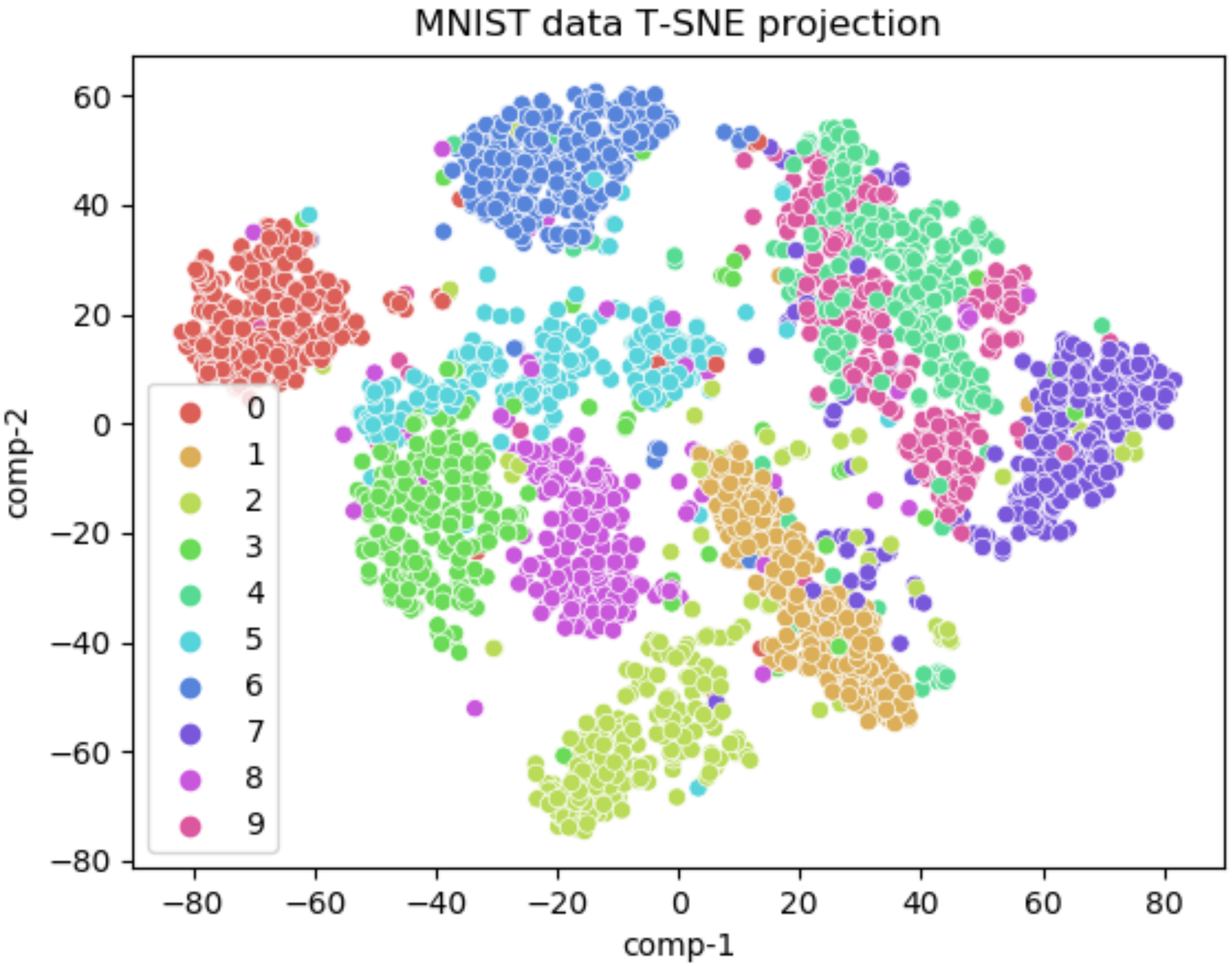
T-distributed Stochastic Neighbor Embedding



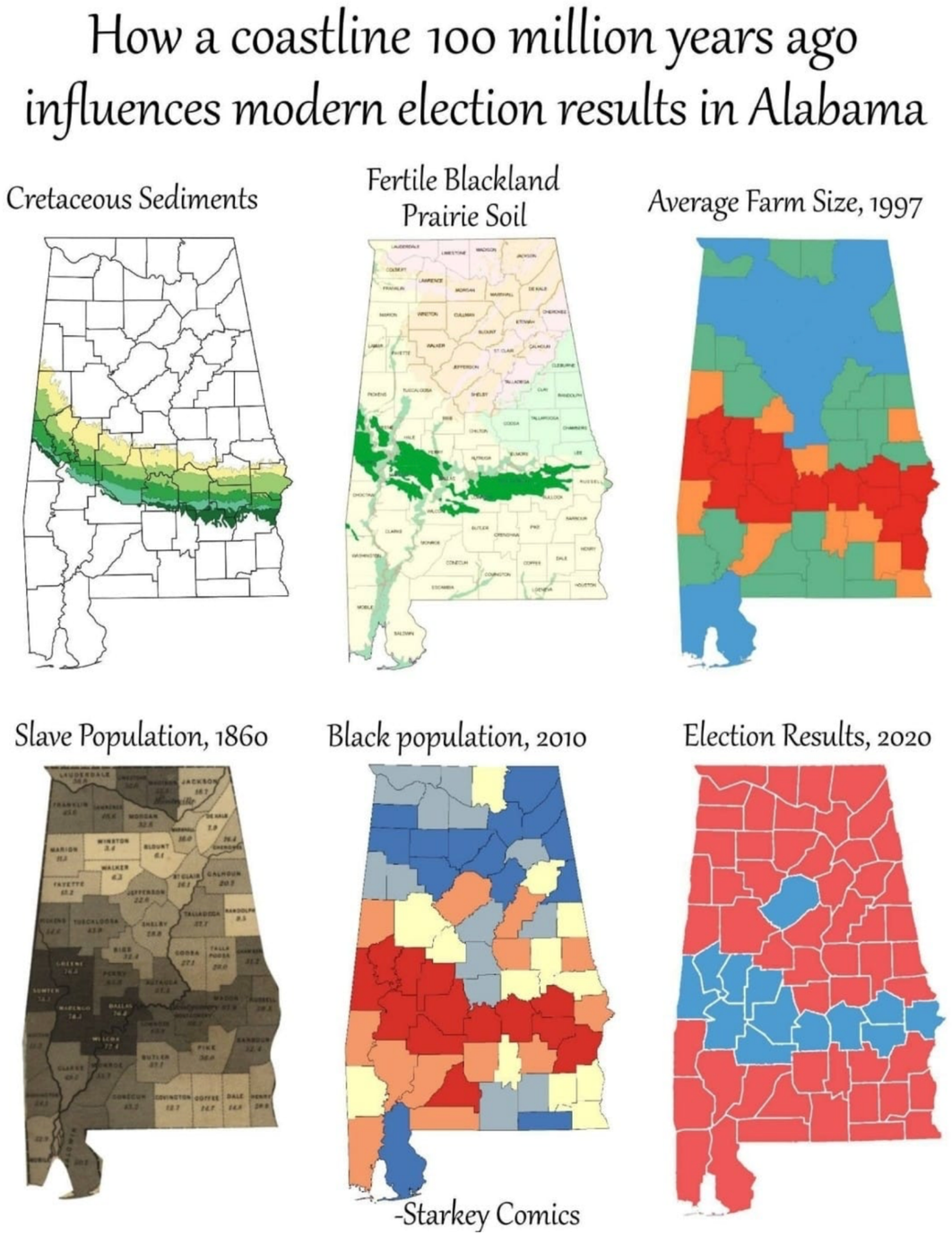
Multidimensional scaling




Digit: 7.0
Digit: 8.0
Digit: 1.0
Digit: 3.0
Digit: 4.0
Digit: 5.0


MNIST data T-SNE projection
comp-1
comp-2
0
1
2
3
4
5
6
7
8
9

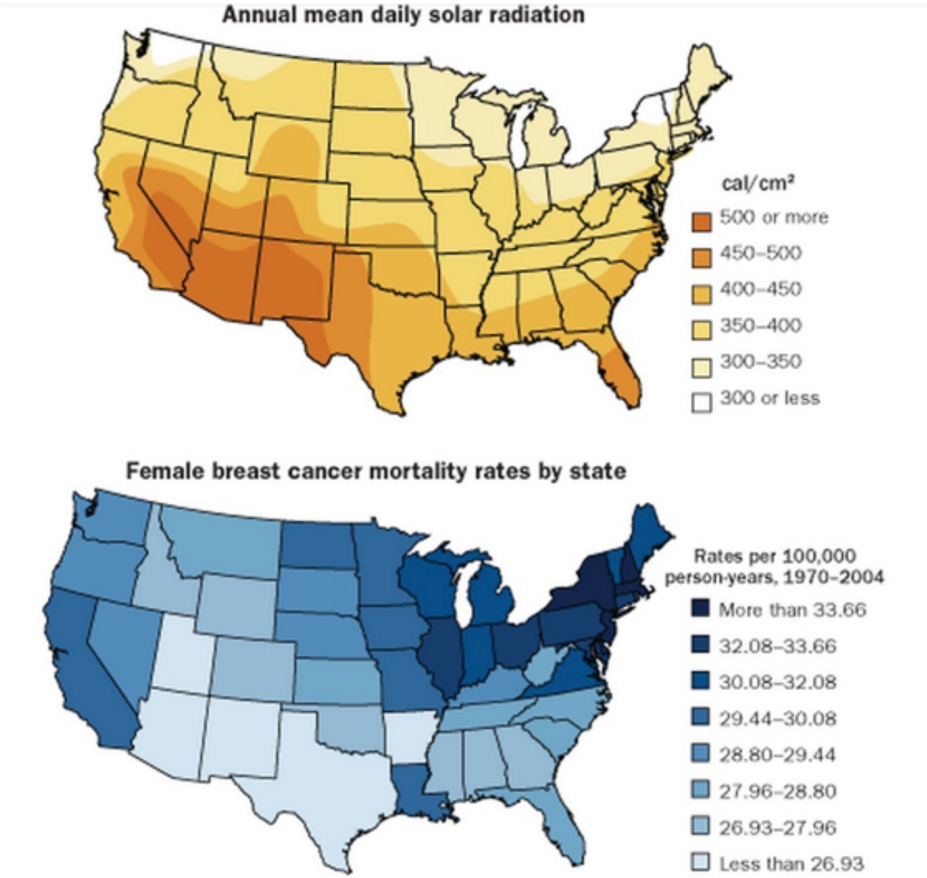
Geospatial viz as EDA



Why Geospatial Analysis?

‘Everything is related to everything else, but near things are more related than distant things.”
(Space > Time)
—Tobler 1979

Why Geospatial Analysis?



ON THE MAP Scientists who study vitamin D can't help but notice that a host of diseases seem to vary with latitude. Type 1 diabetes, multiple sclerosis and even some cancers appear to be more common in areas that get less sun -- meaning less opportunity for the body to produce vitamin D. The maps above illustrate the apparent link between solar radiation and breast cancer mortality rates.

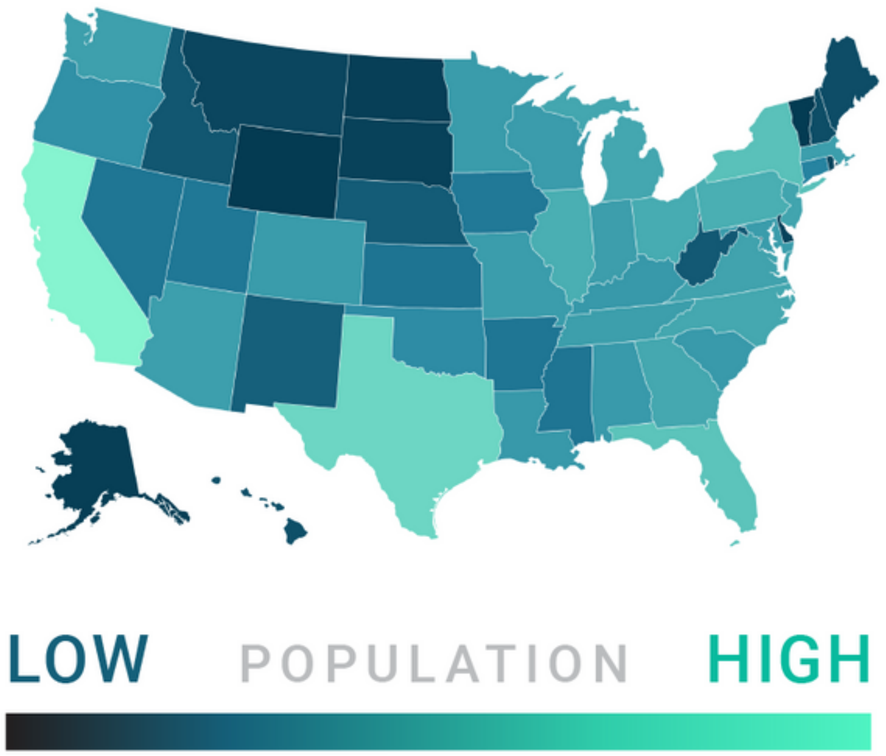
SOURCE, FROM TOP: D. M. HARRIS AND V.L.W. GO / /J. OF NUTRITION 2004; NATIONAL CANCER INSTITUTE

"...the purpose of geographic inquiry is to examine relationships between geographic features collectively and to use the relationships to describe the real-world phenomena that map features represent"

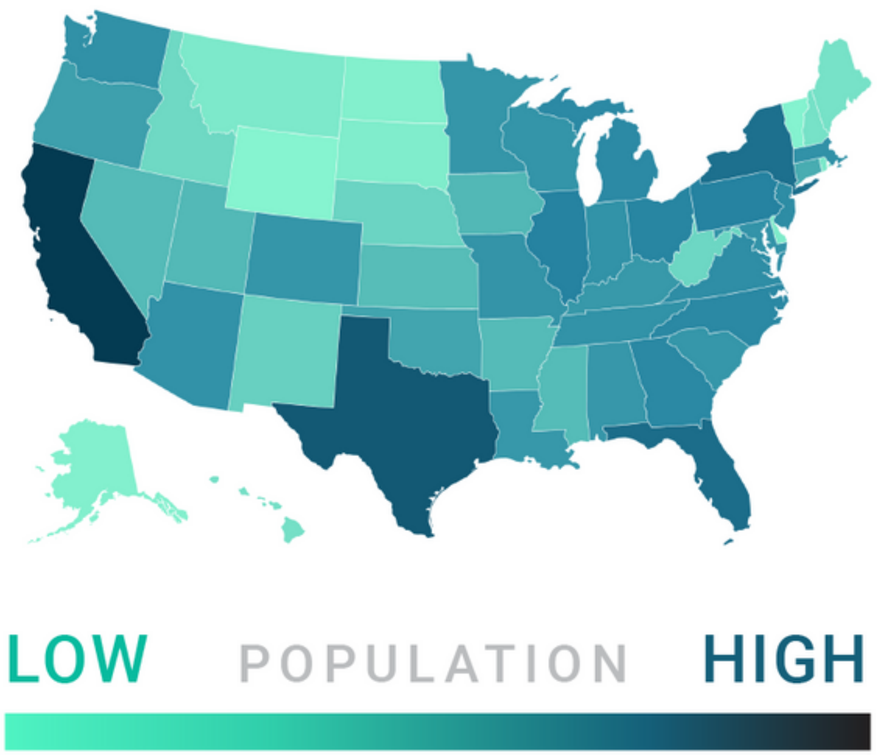
—Clarke 2001

Choropleth maps are useful for visualizing *clear regional patterns* in the data

NOT IDEAL

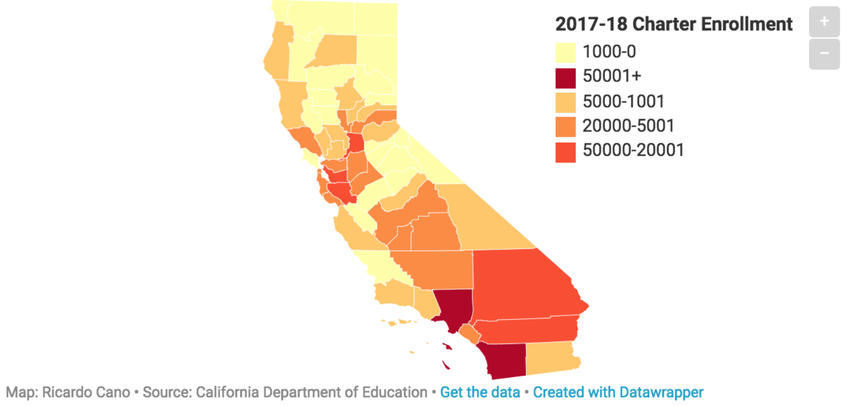


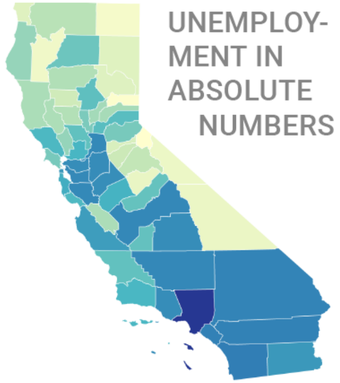
BETTER

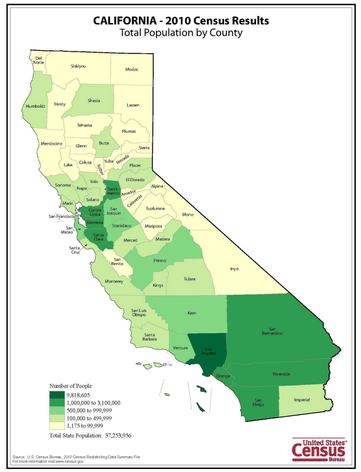


Use light colors for low values. Dark colors for high values.

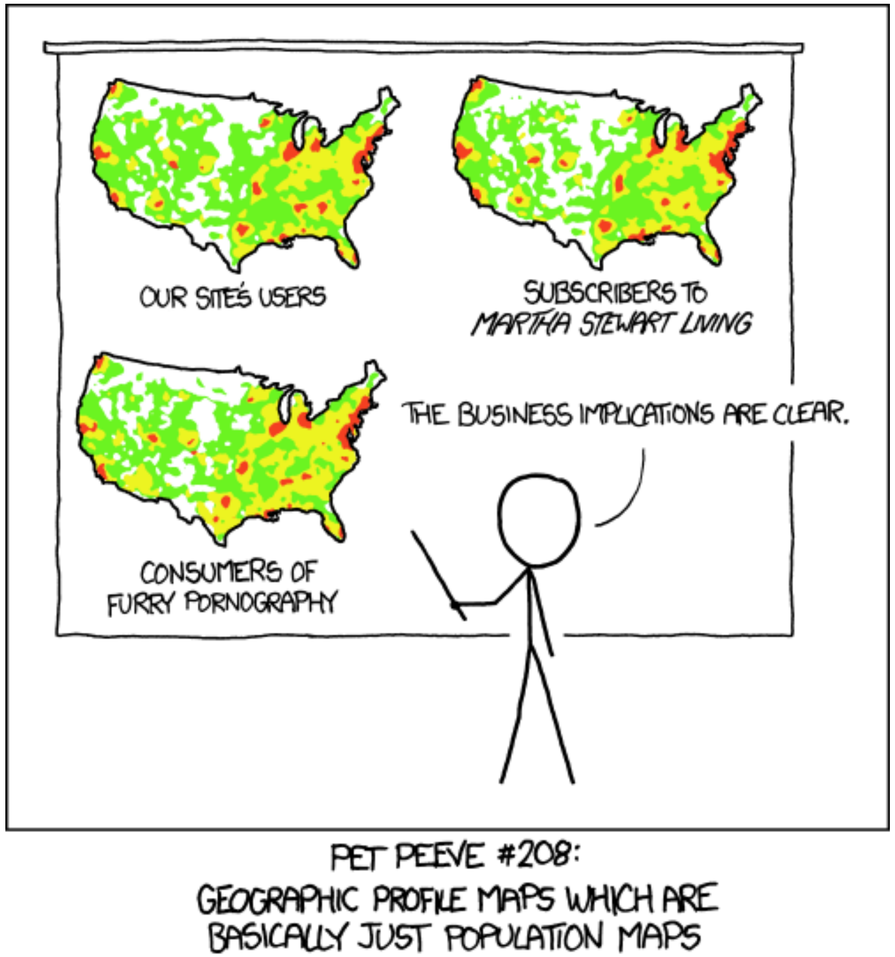
https://blog.datawrapper.de/colors/







https://xkcd.com/1138/

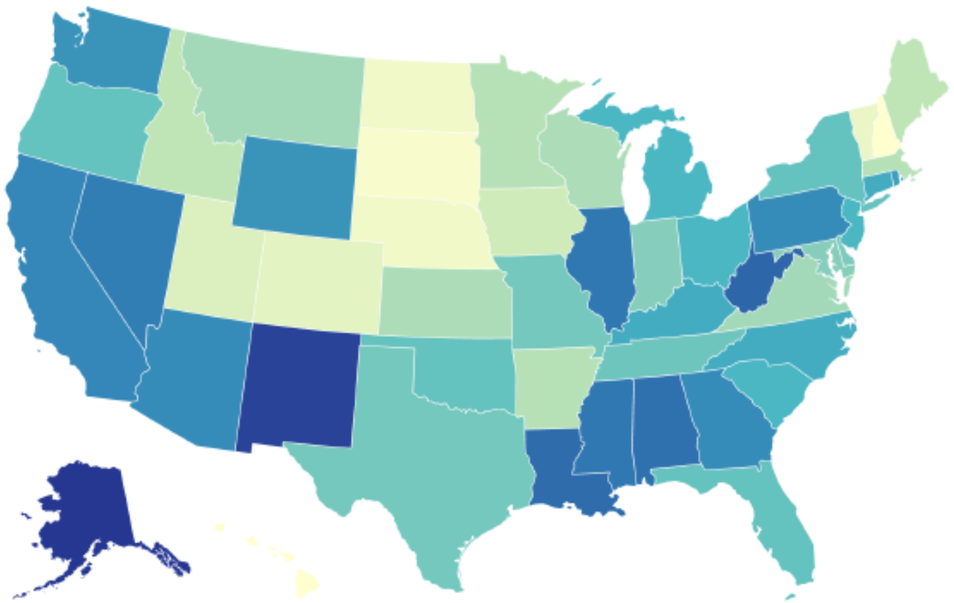


If you don't normalize the statistic (divide by population or by area) choropleth maps can be misleading

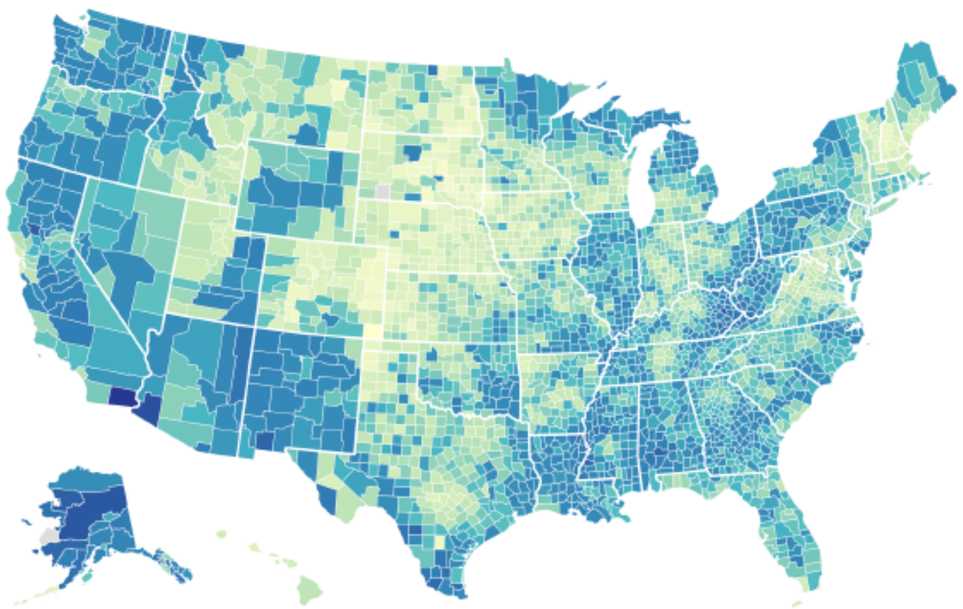
Consider using the smallest unit possible

(but there are exceptions!)

NOT IDEAL

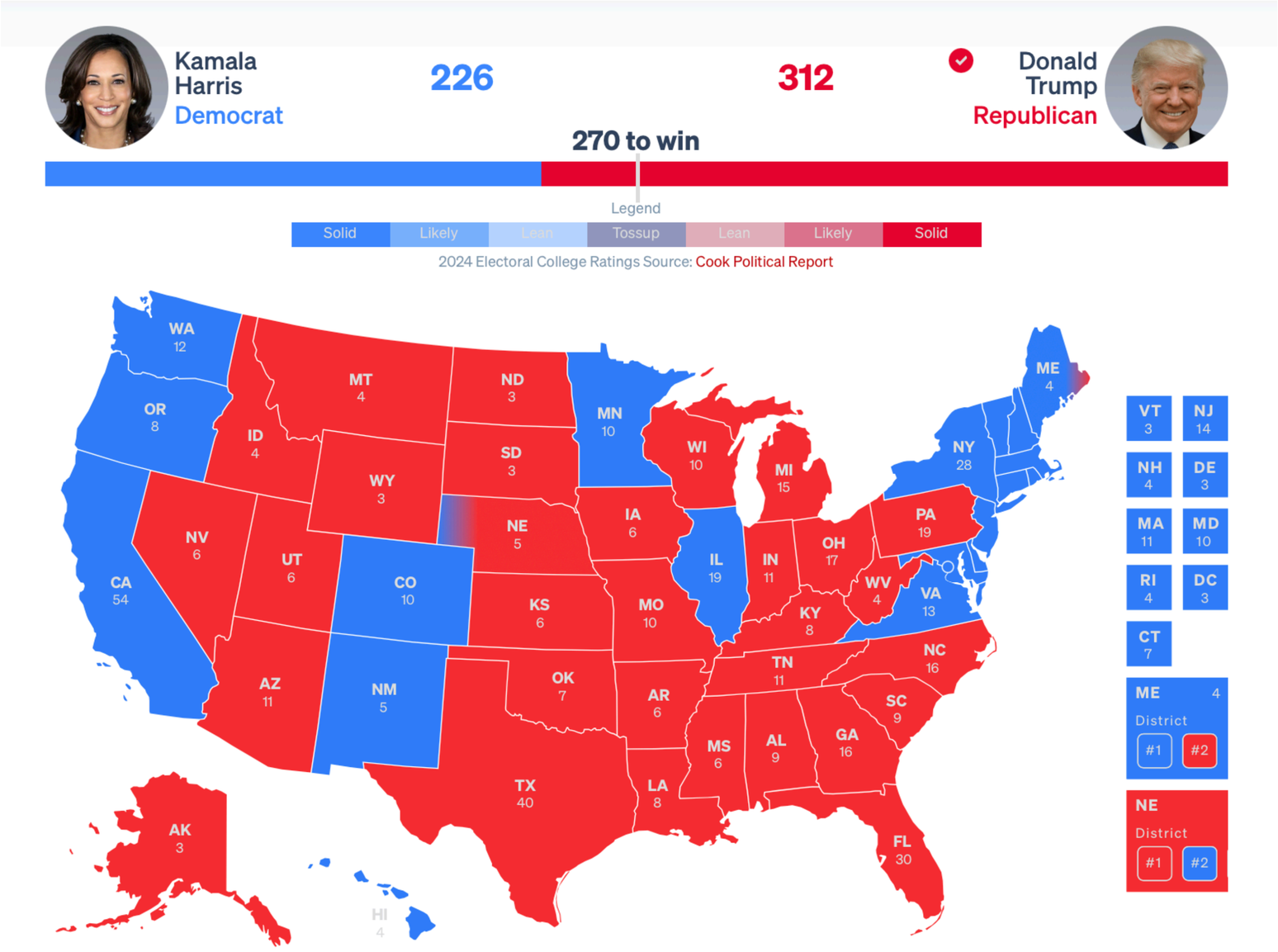


BETTER



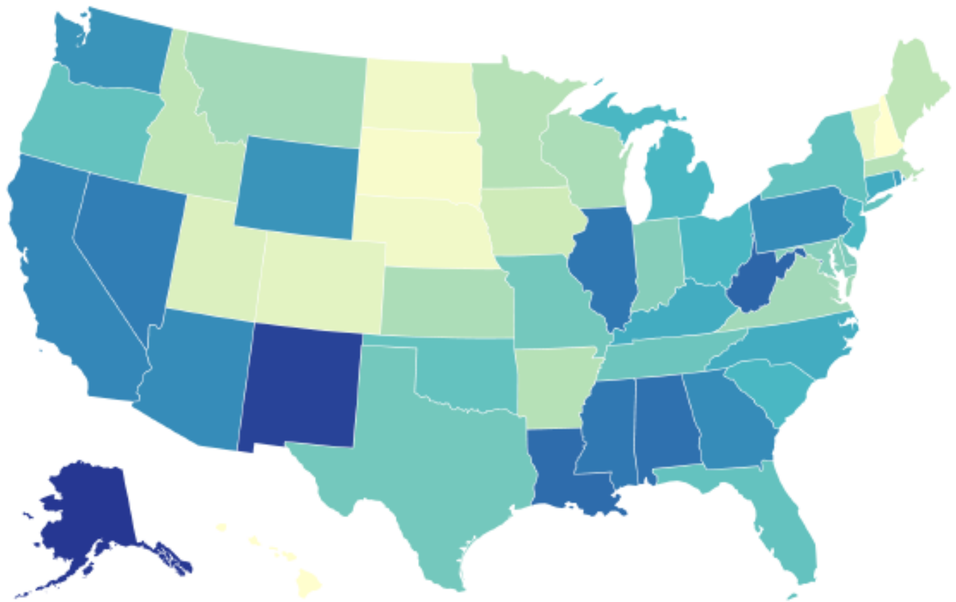
https://blog.datawrapper.de/choroplethmaps/

Sometimes summarizing at the state level is OK



Cartograms should be considered when displaying how many people were affected

NOT IDEAL

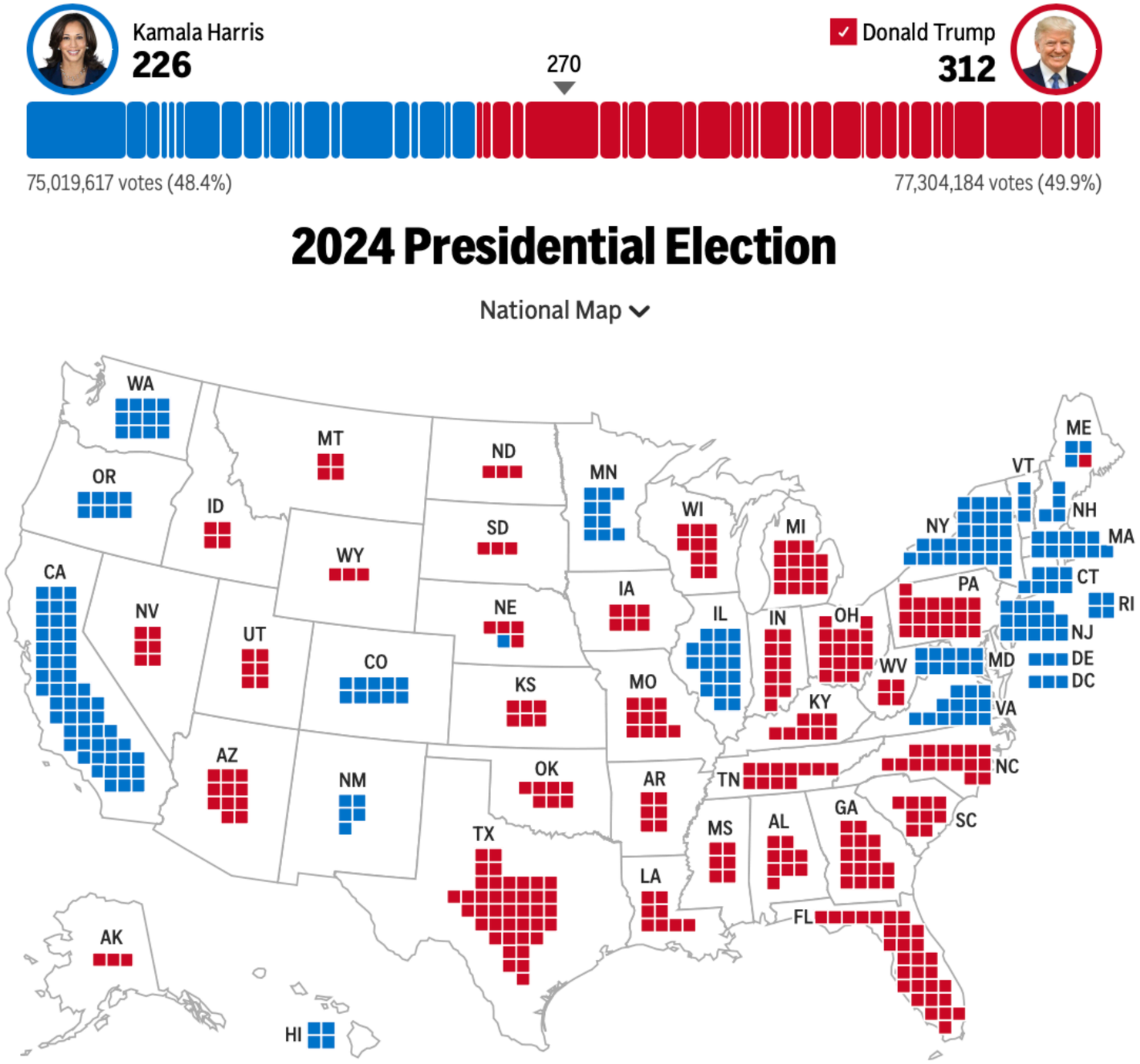


Choropleths answer “How much area was affected?”

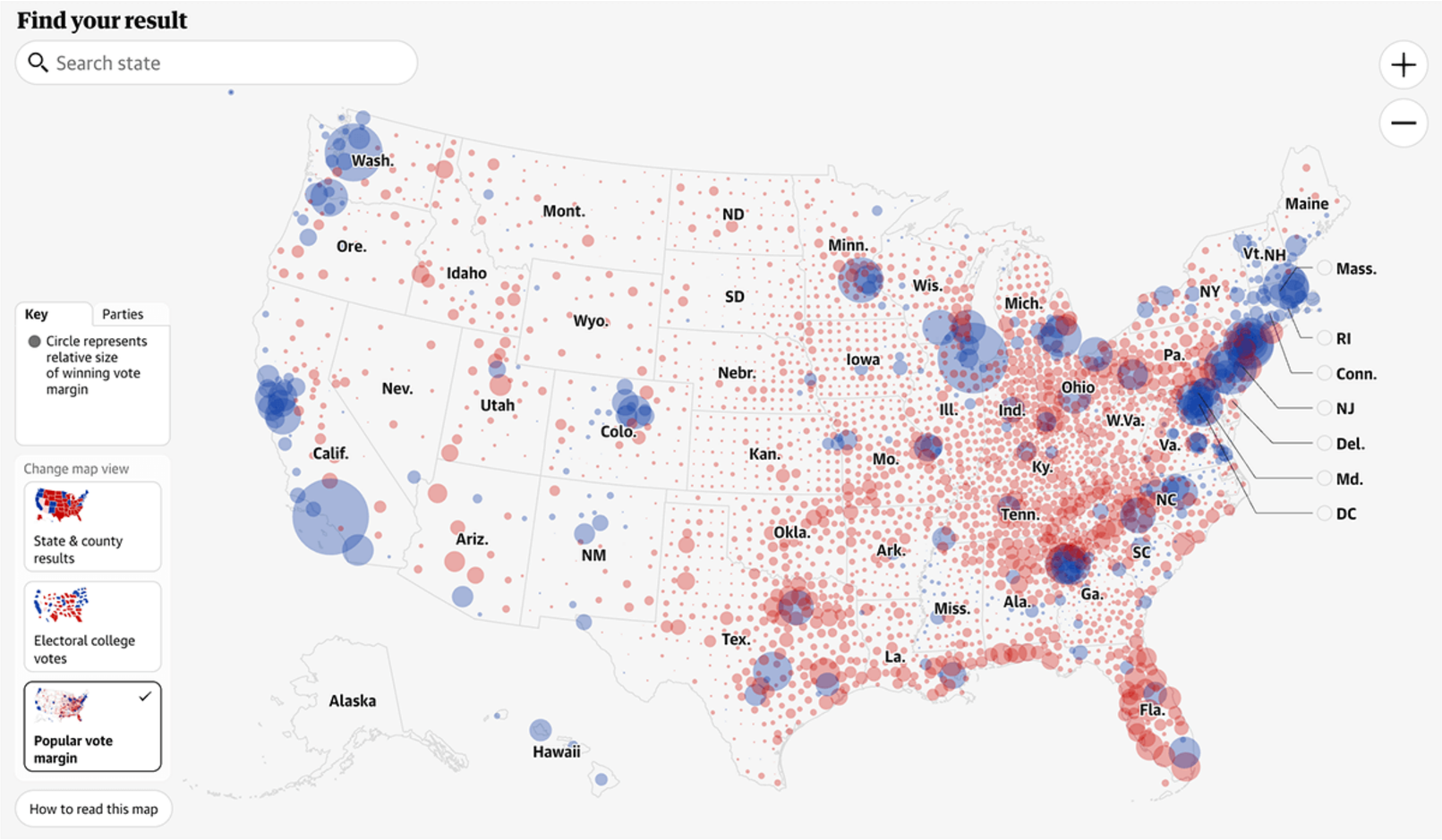
BETTER

Cartograms answer “How many people were affected?”

https://blog.datawrapper.de/choroplethmaps/



This **cartogram** more accurately shows how the populace is split since the number of tiles reflects state populations



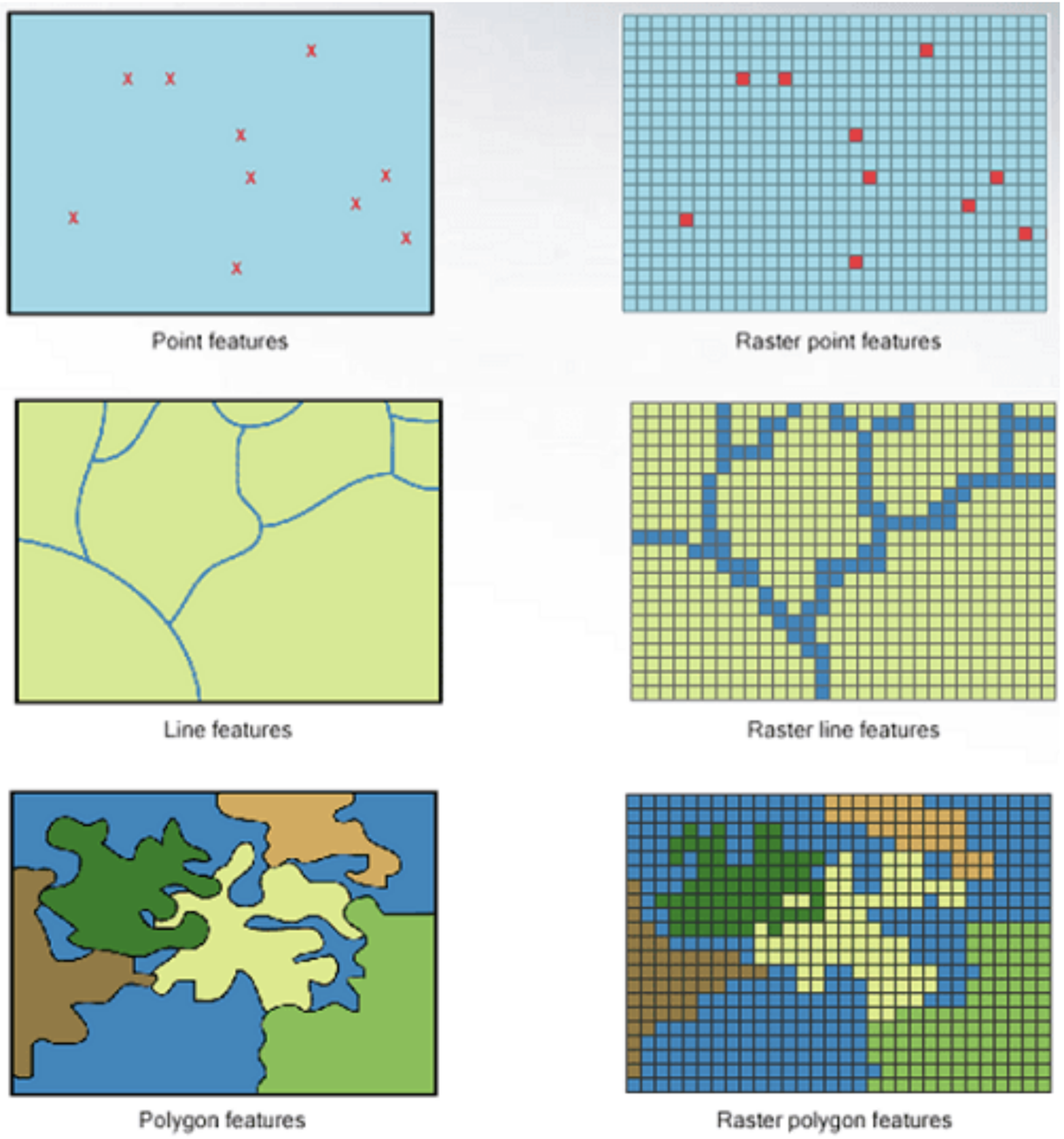
This county level **bubble graph** more accurately tells the full story, since the size of the bubbles reflects population while color shows the political split in that county

Representing spatial data and making maps like these

Representing shapes on maps

- Vector data
  - Points, lines, polygons
- Raster data
  - Encodes the world as a continuous surface represented by a grid, such as the pixels of an image.

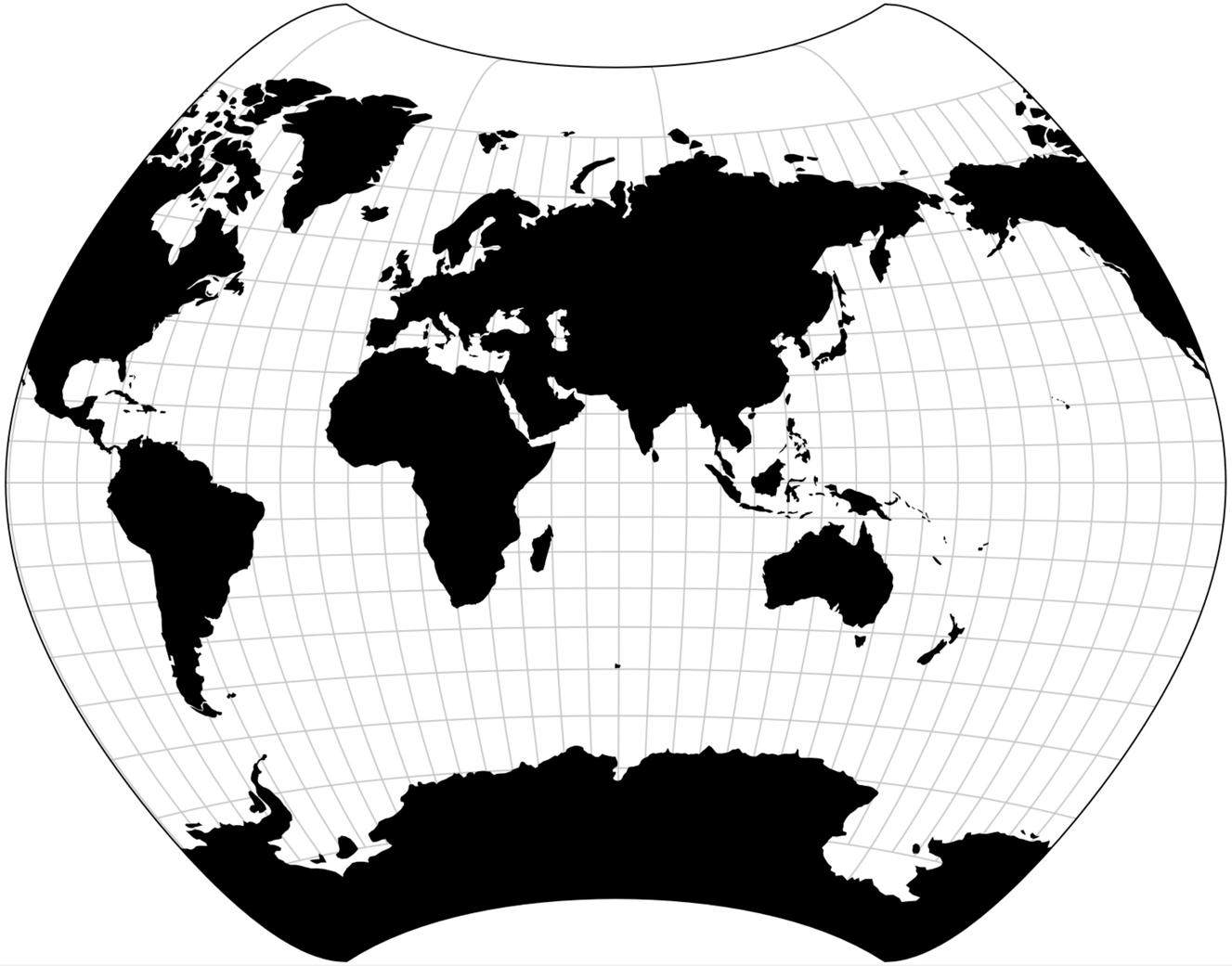
Source: Humboldt State University



Projection Transitions

This notebook interpolates smoothly between projections; this is easiest when both projections are well-defined over the given viewport (here, the world).

Laskowski tri-optimal *projection*



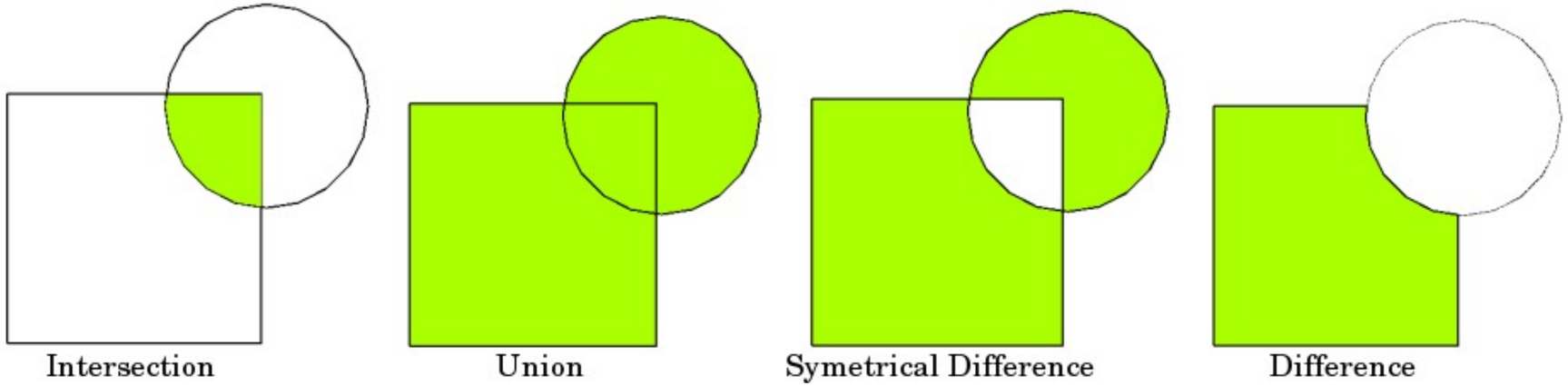
Source: https://observablehq.com/@d3/projection-transitions

Geopandas

- Extends pandas, adds support for geospatial data
- Uses shapely library to represent vector geometry (e.g., POINT, LINE, POLYGON)
- Writes/reads shape files, GeoJSON and other formats
- Uses Coordinate Reference Systems (CRS) to represent how data aligns with the real world
- Different types of CRS: Geographic Coordinate Systems and Projected Coordinate Systems.

Geopandas skills

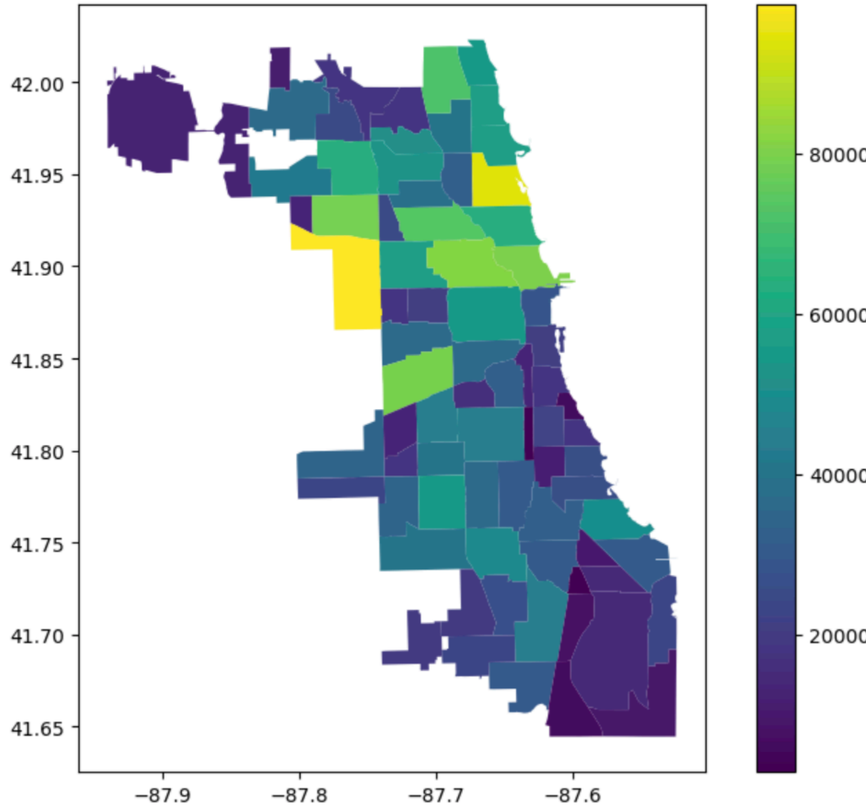
- Overlay spatial datasets: e.g. calculate what regions of Chicago are within 1km of a grocery store
- Join dataframes on shape information
- Aggregate across different subregions to get measurements at a more granular regional level



Geopandas tutorials

```
# Plot population estimates with an accurate legend
In [7]: chicago.plot(column='POP2010', legend=True);
```

- https://geopandas.org/en/stable/getting_started/introduction.html
- https://geopandas.org/en/stable/docs/user_guide/mapping.html
- https://geopandas.org/en/stable/docs/user_guide/set_operations.html
- https://geopandas.org/en/stable/gallery/index.html



Visualizing Geospatial Data

You want to visualize how many people have been affected by COVID19 worldwide.

Best approach to visualize these data?

https://forms.gle/8iSWt8MdJdfM7PC69