YData: Introduction to Data Science

Class 18: Maps and Intro to Statistical Inference

Overview

Very quick review of interactive heatmaps

Mapping continued

If there is time

- Introduction to Statistical Inference

Reminder: class project

The final project is a 6-10 page Jupyter notebook report where you analyze your own data to address a question that you find interesting

- A project template Jupyter notebooks is on Canvas

A polished draft of the project is due on April 9th

Focus on giving insight into some interesting questions

- You do not need to use all methods discussed in the class

Very quick review of Interactive visualizations

Interactive visualizations are useful for exploring data to find trends

We discussed several interactive visualization we can make with plotly:

```
import plotly.express as px

px.scatter(data_frame = , x = , y = , size = , color = , hover_name = )
px.line(data_frame =, x =, y = , color = , hover_name = , line_shape = )

px.sunburst(data_frame = ,  path = , values = ,  color = )
px.treemap(data_frame = , path = , values = , color =  )
px.imshow(df2)    # heatmap
```

Pivot Tables and heatmaps

Pivot tables aggregate values based on to two grouping variables, and create a table where:

- The columns are the levels of one variable
- The rows are the levels of the other variable

df2 = df.pivot_table(index = "col1", columns = "col2",
values = "col3”, aggfunc = "mean")

Once we have a 2D table, we can visualize it using:

- px.imshow(df2) # create a heatmap using plotly
- sns.heatmap(df2) # create a heatmap using seaborn

Grouping: df.groupby(["col1" col2"]).

col1 col2

Flavor	Color	count
bubblegum	pink	1
chocolate	dark brown	2
chocolate	light brown	1
strawberry	pink	2

Pivot Table: df.pivot_table()

col1

col2

Color	bubblegum	chocolate	strawberry
dark brown	0	2	0
light brown	0	1	0
pink	1	0	2

Pivot Tables and heatmaps

If we want to create a pivot table without aggregating data, we can use the .pivot() method

- rather than .pivot_table() method

df2 = df.pivot(index = "col1", columns = "col2",

values = "col3")

Note: there needs to be one value for each combination of "col1" x "col2" levels.

Grouping: df.groupby(["col1" col2"]).

col1 col2

Flavor	Color	count
bubblegum	pink	1
chocolate	dark brown	2
chocolate	light brown	1
strawberry	pink	2

Pivot Table: df.pivot_table()

col1

col2

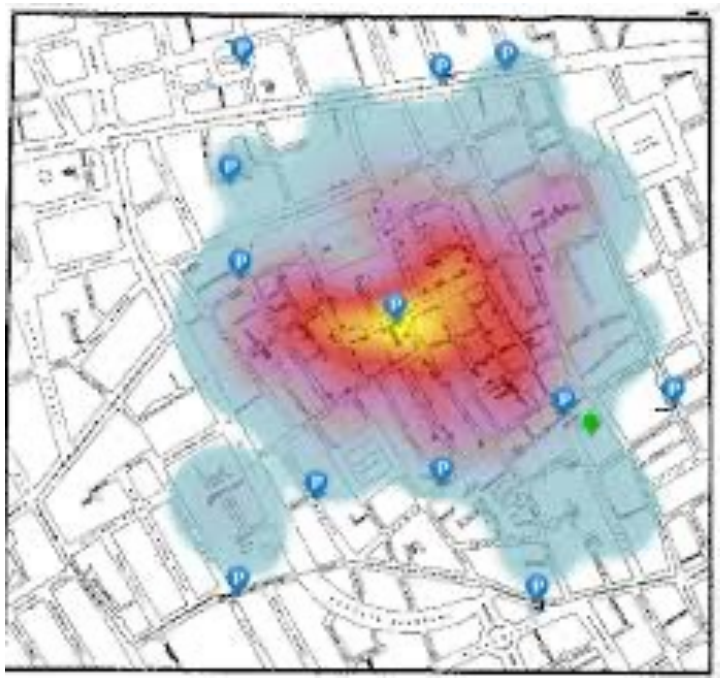
Color	bubblegum	chocolate	strawberry
dark brown	0	2	0
light brown	0	1	0
pink	1	0	2

Let's explore this in Jupyter!

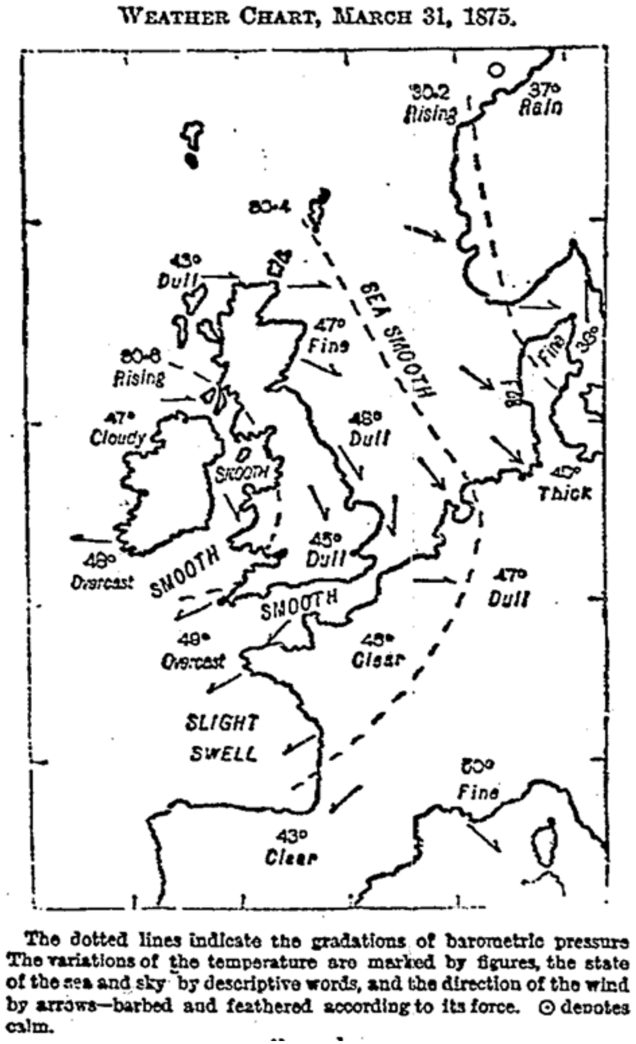
Maps

Maps

Visualizing data on a map can be a powerful way to see spatial trends



John Snow's ghost map (1854)



Galton's first weather map (1875)

geopandas

To create maps in Python we will use the geopandas package

- import geopandas as gpd

The key object of interest is the geopandas DataFrame

- It is the same as a regular data frame but it has an extra column called "geometry" that contains geospatial shape features
  - The geometry column as "Shapely" objects used to represent geometric shapes

	key_comb_drvr	geometry
0	M11551	POINT (117.525391 34.008926)
1	M17307	POINT (86.51248 30.474344)
2	M19584	POINT (89.537415 37.157627)
3	M21761	POINT (117.526871 34.00647)
4	M22374	POINT (117.525345 34.008915)
5	U01997A	POINT (84.80533 33.719654)
6	U153601	POINT (78.24838 39.986454)
7	U159393	POINT (98.49438499999999 40.801544)
8	U722222	POINT (84.23309 33.9386)
9	U723030	POINT (83.86456 34.08479)
10	U723333	POINT (85.67151 42.83093)
11	U753333	POINT (117.498535 34.069157)
12	U760505	POINT (90.61252 41.456993)

geopandas

We can read in data as a geopandas DataFrame using

- map = gpd.read_file('my_file.geojson')

We can plot maps using the gpd.plot() function

Let's explore this in Jupyter!

Coordinate reference systems

A coordinate reference system (CRS) is a framework used to precisely measure locations on the surface of the Earth as coordinates

The goal of any coordinate reference system is to create a common reference frame in which locations can be measured precisely as coordinates, so that any recipient can identify the same location that was originally intended.

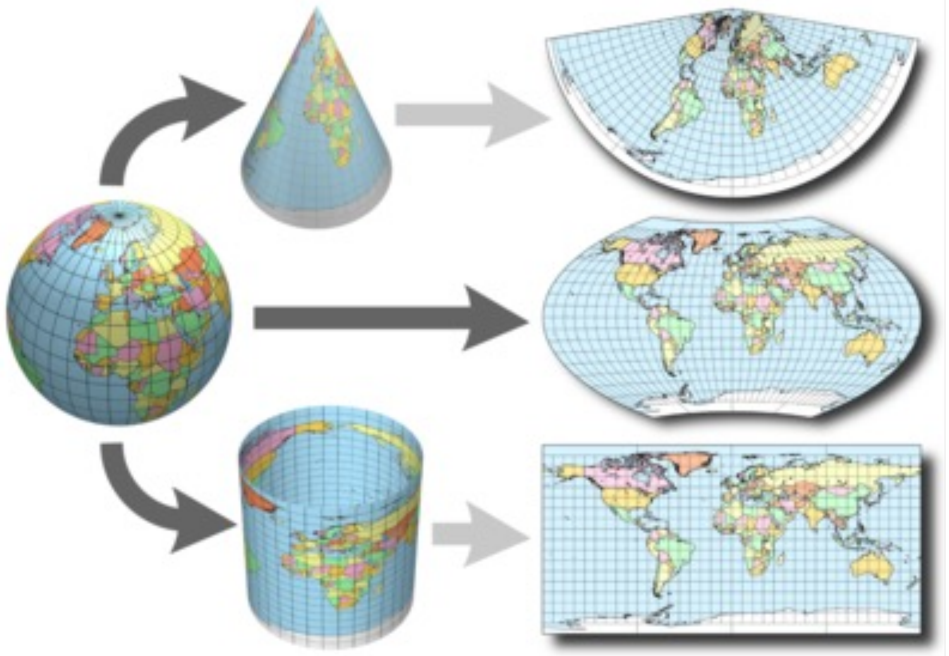
- Needed for aligning different layers on maps

Map projections

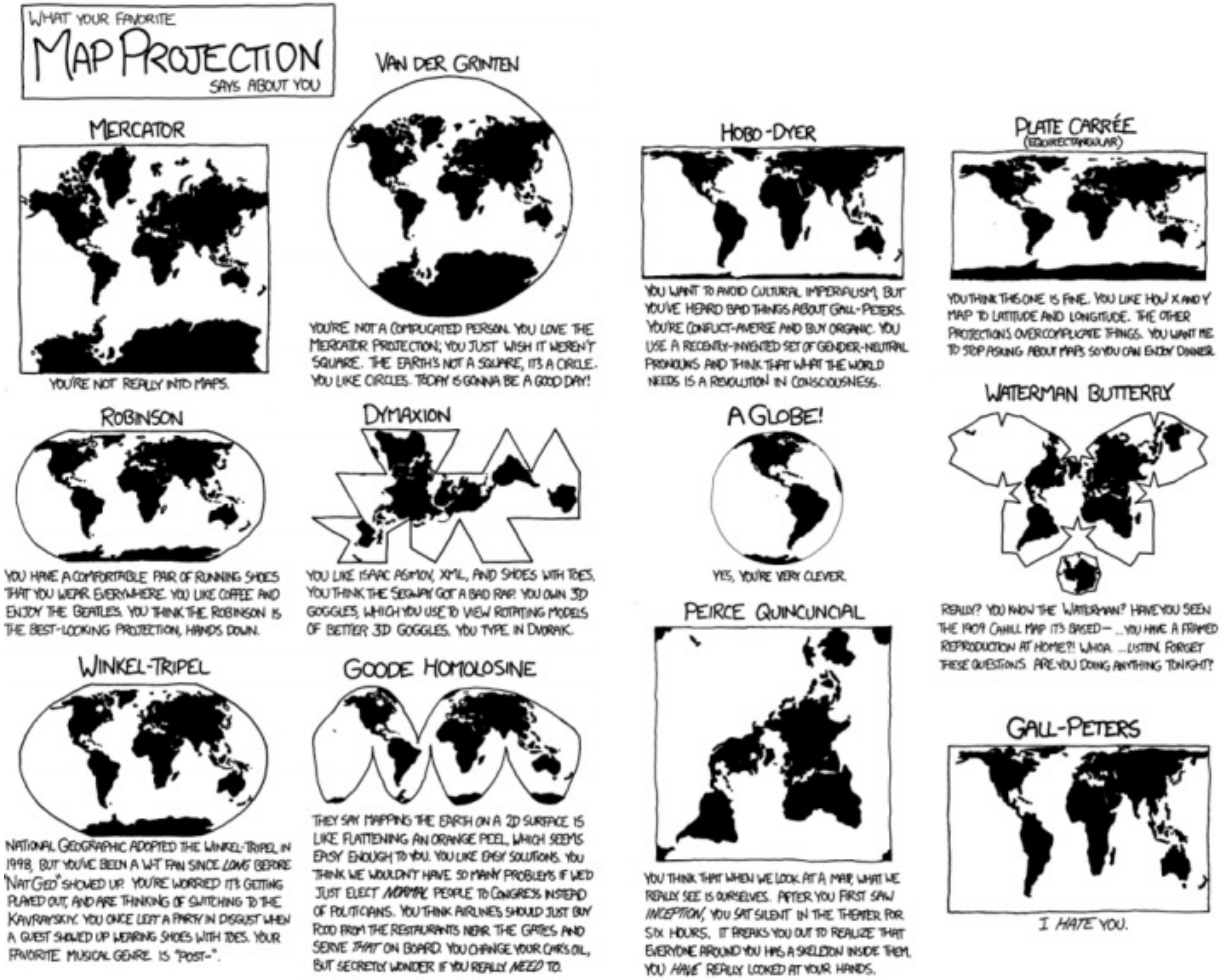
Since the earth is a 3D structure, coordinate systems have to project their data onto a 2D maps

Different projects preserve different properties

- **Mercator projection** keeps angles intact
  - Useful for navigation
- **Eckert IV projection** keeps the size of land areas intact



Let's explore this in Jupyter!

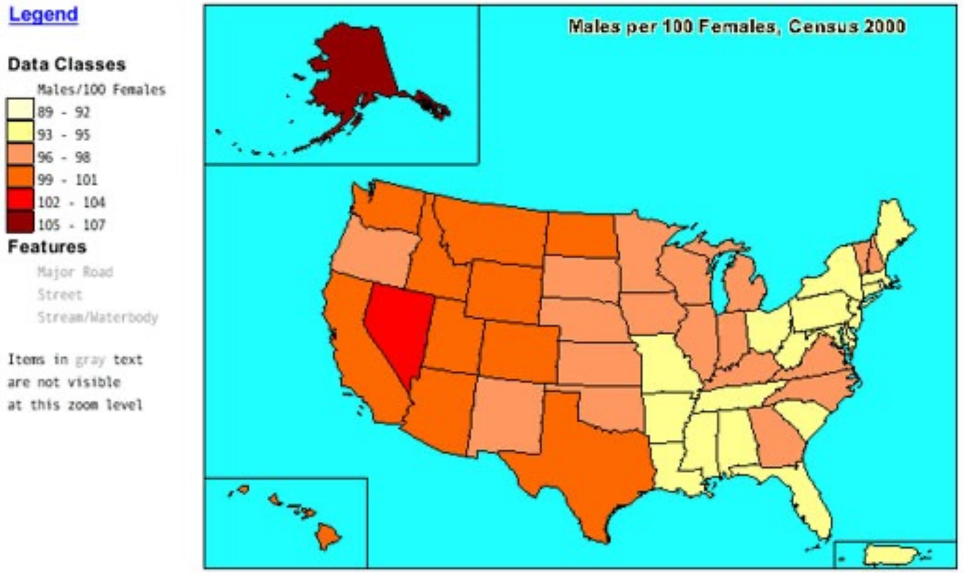

WHAT YOUR FAVORITE
MAP PROJECTION
SAYS ABOUT YOU
MERCATOR
YOU'RE NOT REALLY INTO MAPS.
VAN DER GRINTEN
YOU'RE NOT A COMPLICATED PERSON. YOU LOVE THE MERCATOR PROJECTION; YOU JUST WISH IT WEREN'T SQUARE. THE EARTH'S NOT A SQUARE, IT'S A CIRCLE. YOU LIKE CIRCLES. TODAY IS GONNA BE A GOOD DAY!
HOBO-DYER
YOU WANT TO AVOID CULTURAL IMPERIALISM, BUT YOU'VE HEARD BAD THINGS ABOUT GALL-PETERS. YOU'RE CONFLICT-AVERSE AND BUY ORGANIC. YOU USE A RECENTLY-INVENTED SET OF GENDER-NEUTRAL PRONOUNS AND THINK THAT WHAT THE WORLD NEEDS IS A REVOLUTION IN CONSCIOUSNESS.
PLATE CARRÉE
(EQUIRECTANGULAR)
YOU THINK THIS ONE IS FINE. YOU LIKE HOW X AND Y MAP TO LATITUDE AND LONGITUDE. THE OTHER PROJECTIONS OVERCOMPLICATE THINGS. YOU WANT ME TO STOP ASKING ABOUT MAPS SO YOU CAN ENJOY DINNER.
ROBINSON
YOU HAVE A COMFORTABLE PAIR OF RUNNING SHOES THAT YOU WEAR EVERYWHERE. YOU LIKE COFFEE AND ENJOY THE BEATLES. YOU THINK THE ROBINSON IS THE BEST-LOOKING PROJECTION, HANDS DOWN.
DYMAXION
YOU LIKE ISAAC ASIMOV, XML, AND SHOES WITH TOES. YOU THINK THE SEGWAY GOT A BAD RAP. YOU OWN 3D GOGGLES, WHICH YOU USE TO VIEW ROTATING MODELS OF BETTER 3D GOGGLES. YOU TYPE IN DVORAK.
A GLOBE!
YES, YOU'RE VERY CLEVER.
WATERMAN BUTTERFLY
REALLY? YOU KNOW THE WATERMAN? HAVE YOU SEEN THE 1909 CAHILL MAP IT'S BASED— ...YOU HAVE A FRAMED REPRODUCTION AT HOME?! WHOA. ...LISTEN, FORGET THESE QUESTIONS. ARE YOU DOING ANYTHING TONIGHT?
WINKEL-TRIPEL
NATIONAL GEOGRAPHIC ADOPTED THE WINKEL-TRIPEL IN 1998, BUT YOU'VE BEEN A W-T FAN SINCE LONG BEFORE "NAT GEO" SHOWED UP. YOU'RE WORRIED IT'S GETTING PLAYED OUT, AND ARE THINKING OF SWITCHING TO THE KAVRAYSKIY. YOU ONCE LEFT A PARTY IN DISGUST WHEN A GUEST SHOWED UP WEARING SHOES WITH TOES. YOUR FAVORITE MUSICAL GENRE IS "POST-".
GOODE HOMOLOSINE
THEY SAY MAPPING THE EARTH ON A 2D SURFACE IS LIKE FLATTENING AN ORANGE PEEL, WHICH SEEMS EASY ENOUGH TO YOU. YOU LIKE EASY SOLUTIONS. YOU THINK WE WOULDN'T HAVE SO MANY PROBLEMS IF WE'D JUST ELECT NORMAL PEOPLE TO CONGRESS INSTEAD OF POLITICIANS. YOU THINK AIRLINES SHOULD JUST BUY FOOD FROM THE RESTAURANTS NEAR THE GATES AND SERVE THAT ON BOARD. YOU CHANGE YOUR CAR'S OIL, BUT SECRETLY WONDER IF YOU REALLY NEED TO.
PEIRCE QUINCUNCIAL
YOU THINK THAT WHEN WE LOOK AT A MAP, WHAT WE REALLY SEE IS OURSELVES. AFTER YOU FIRST SAW INCEPTION, YOU SAT SILENT IN THE THEATER FOR SIX HOURS. IT FREAKS YOU OUT TO REALIZE THAT EVERYONE AROUND YOU HAS A SKELETON INSIDE THEM. YOU HAVE REALLY LOOKED AT YOUR HANDS.
GALL-PETERS
I HATE YOU.

Maps

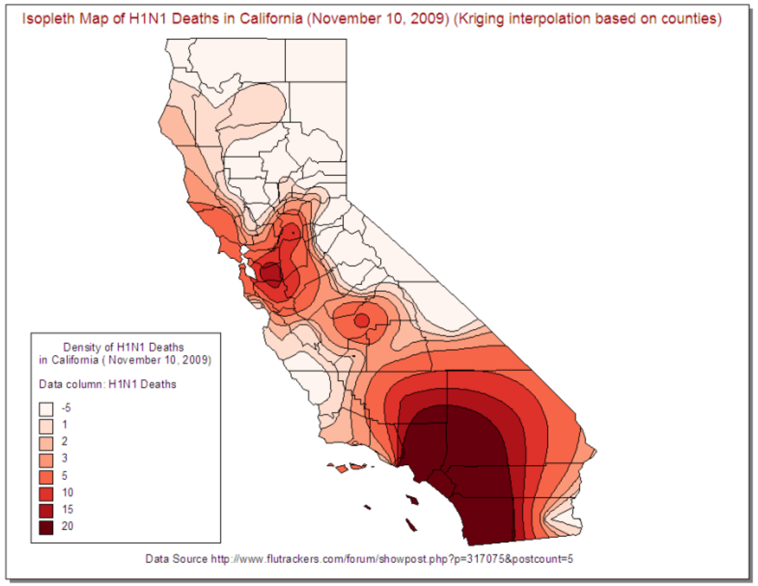
Choropleth maps: shades/colors in predefined areas based on properties of a variable

Isopleth maps: creates regions based on constant values

Choropleth map



Isopleth map



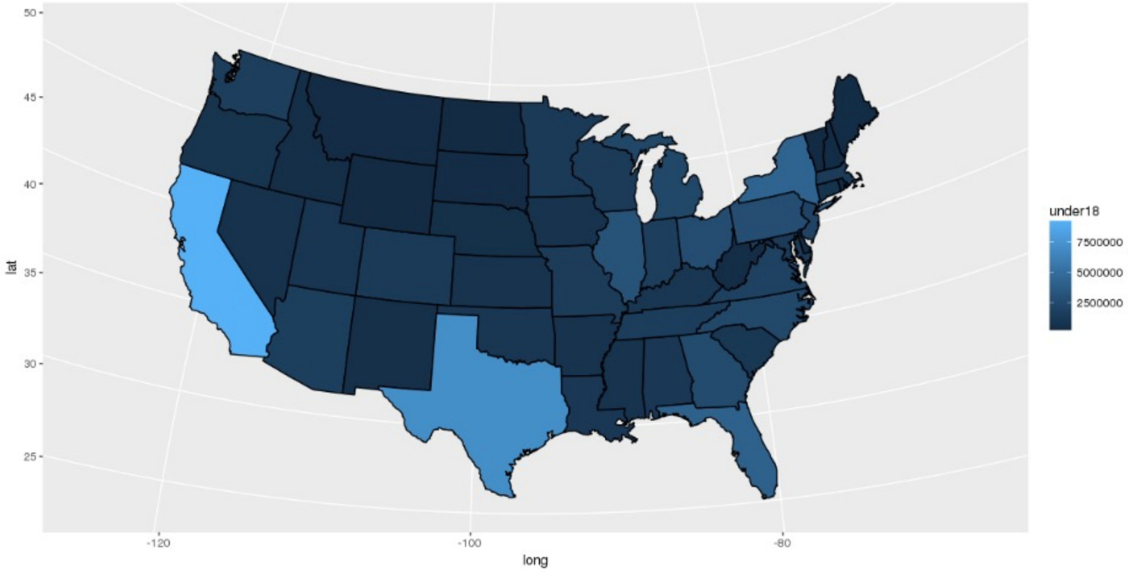
Choropleth maps

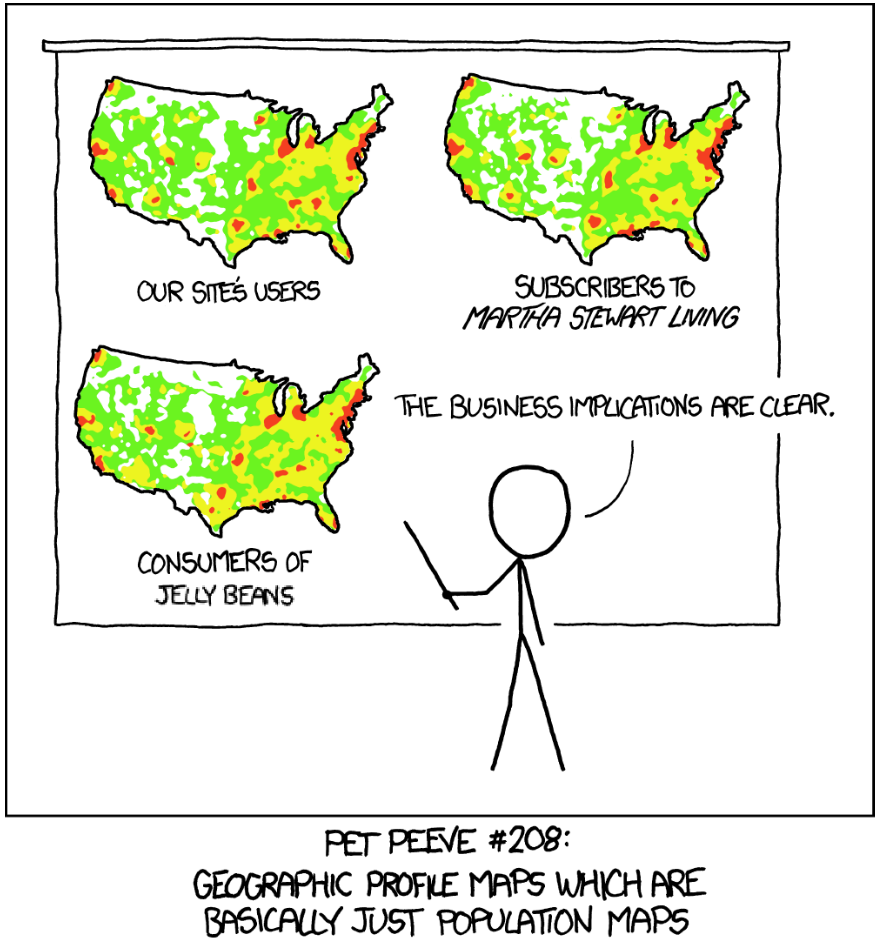
We can create choropleth maps using geopandas by joining region information on to a geopandas DataFrame that has a map

We can then use the `gpd.plot(column = )` method to visualize the map

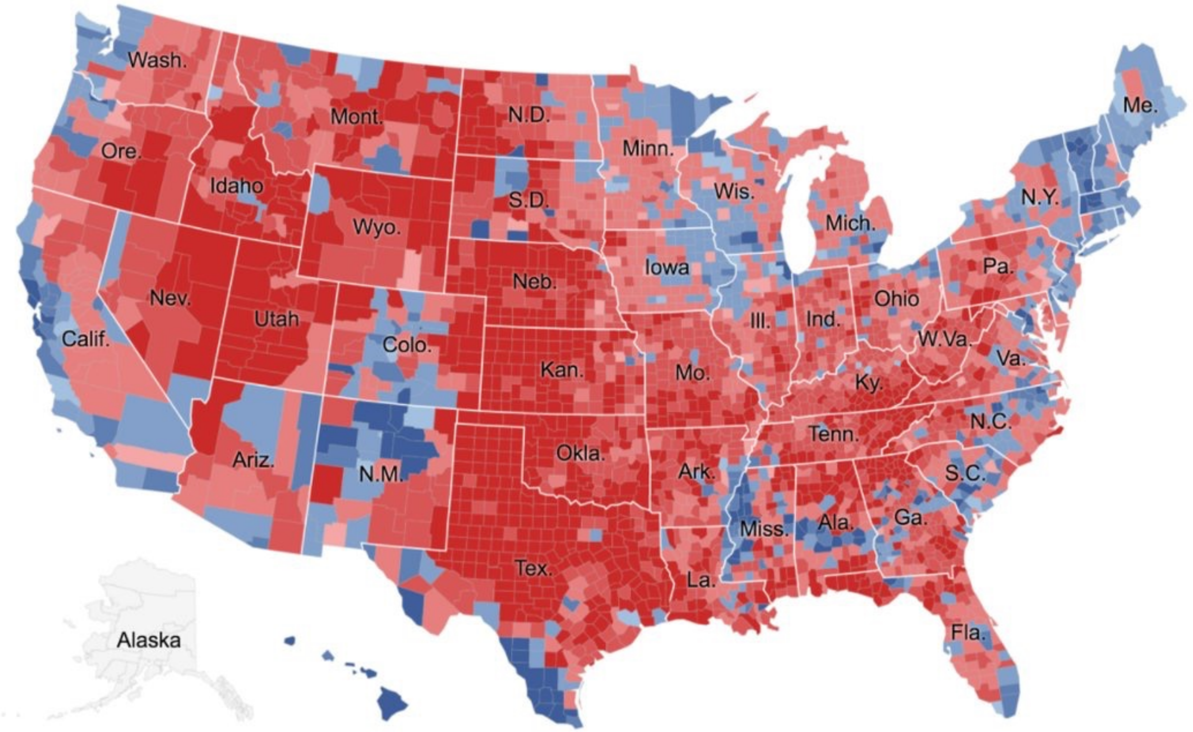
Let's explore this in Jupyter!

Pet Peeve #208





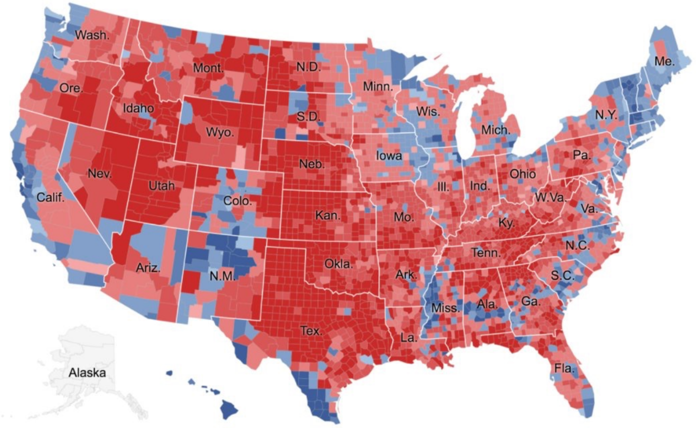
Question: in what way could this map be misleading?



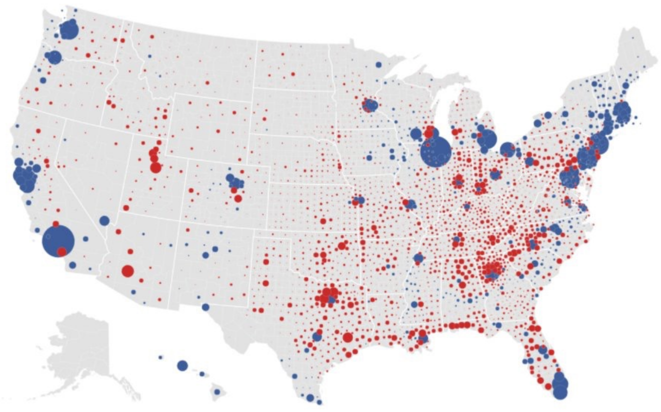
Darker red: county had higher % Trump vote

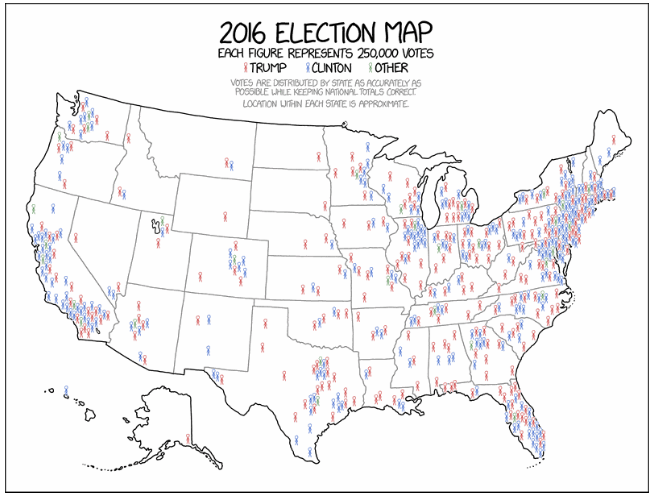
Darker blue: county had higher % Clinton vote

Cloropleth maps can be misleading



Looks like most of the country voted republican



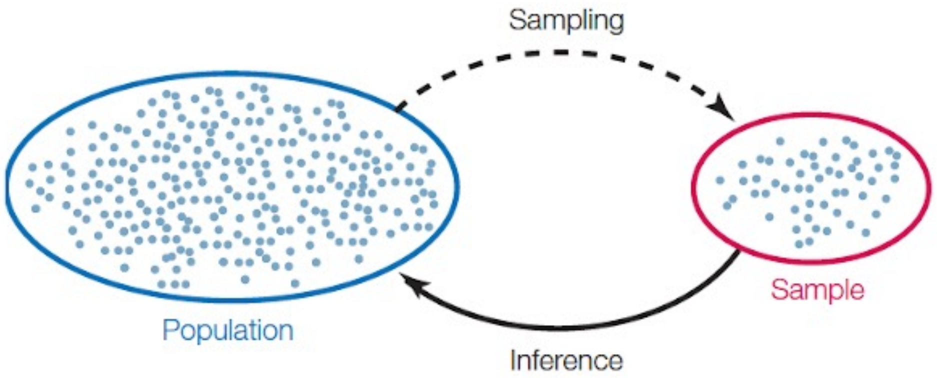


Statistical Inference

Inference

Statistical Inference: Making conclusions about a population based on data in a random sample

This usually involves using data in a sample to estimate the value of a fixed unknown number



Example:

- Estimating the average height of all humans on Earth from a random sample of 1,000 humans
  - Our estimate will vary from sample to sample

Terminology

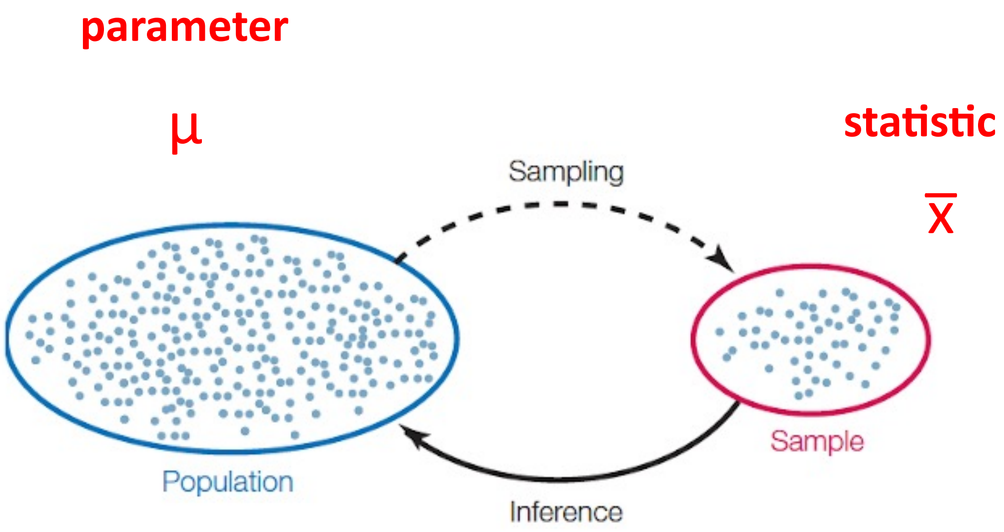
A parameter is number associated with the population

- e.g., population mean μ
- e.g., average height of all humans

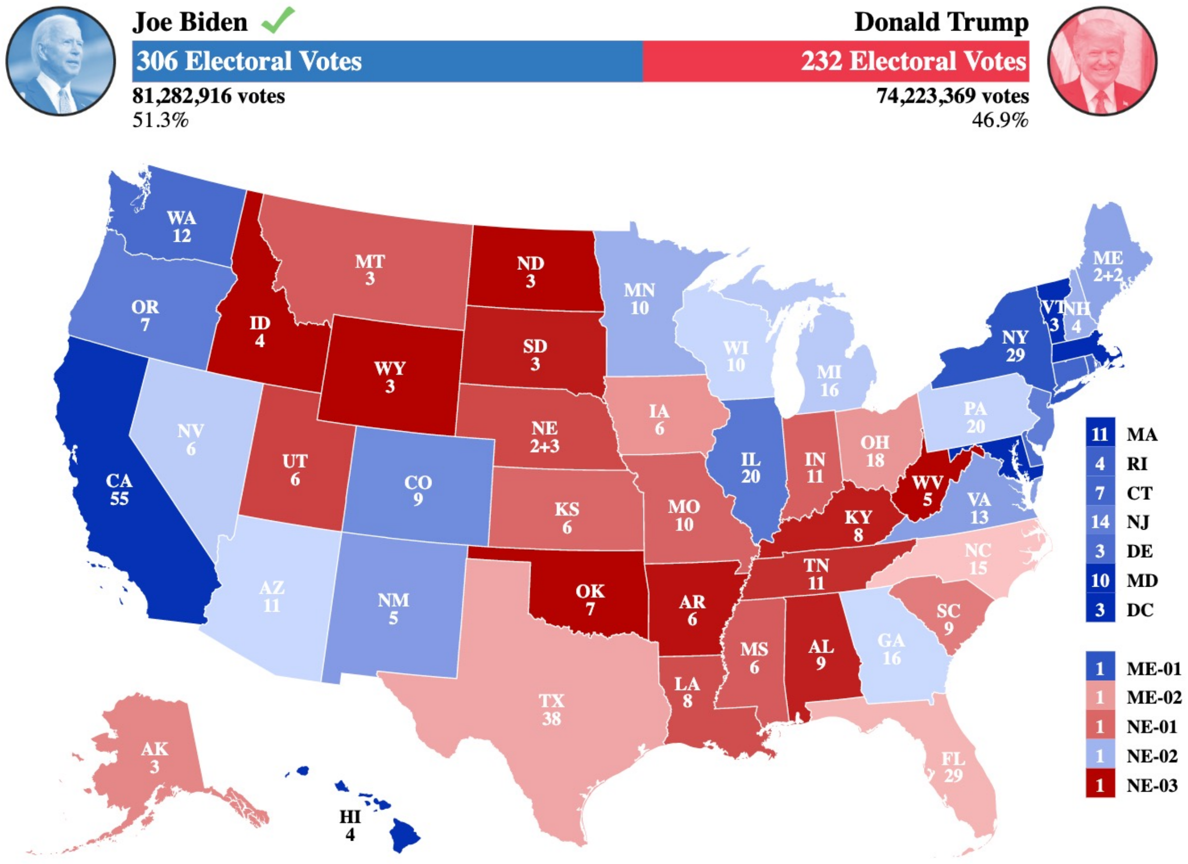
A **statistic** is number calculated from the sample

- e.g., sample mean $\bar{x}$
- e.g., average height of 1,000 people in our sample

A statistic can be used as an estimate of a parameter



Example: The 2020 US Presidential Election



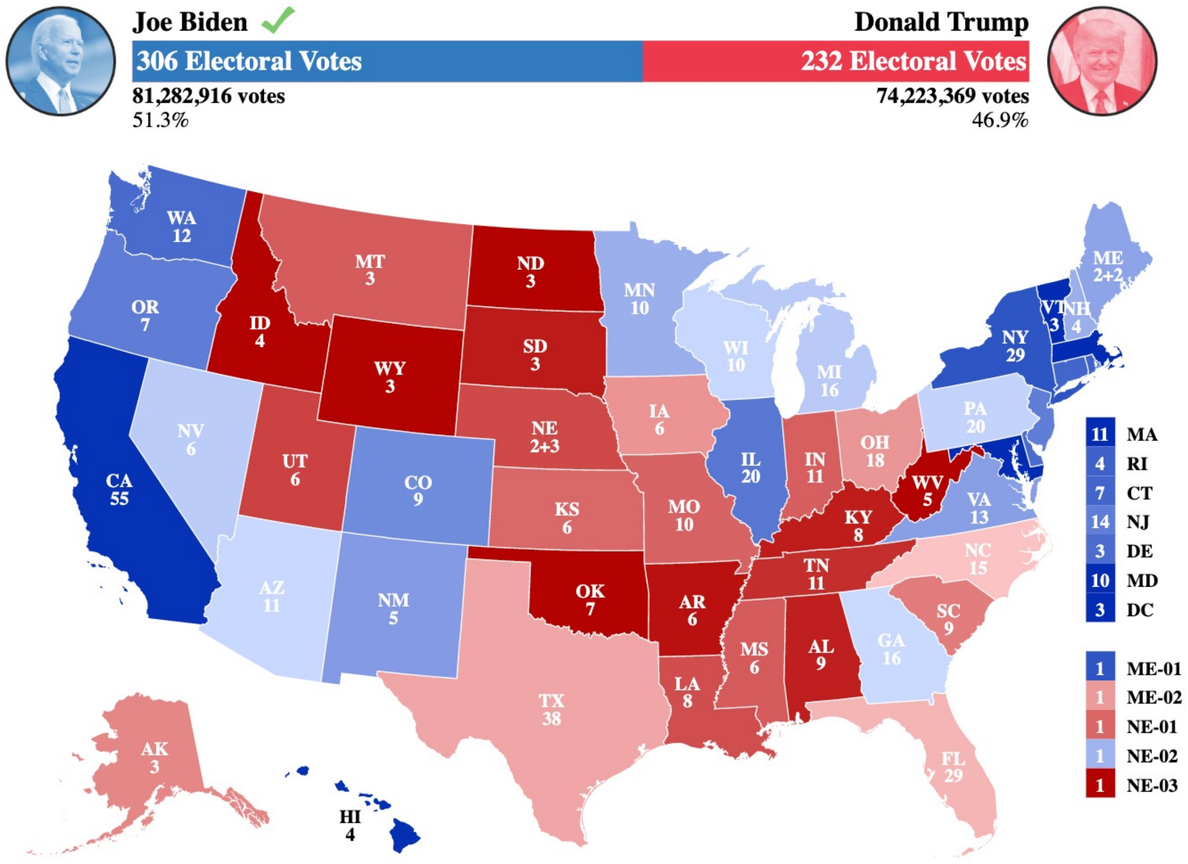
According to The Cook Political Report, the voting outcome in Georgia was

- Trump = 2,461,854
- Biden = 2,473,633

We can denote the proportion of the vote that Biden got using π_{Biden}

- Q: what is the value of π_{Biden} ?

Example: The 2020 US Presidential Election



If 1,000 voters were randomly sampled, we could denote the proportion in the sample that voted for Biden using: $\hat{p}_{Biden}$

Would we expect $\hat{p}_{Biden}$ to be equal to π_{Biden}?

If we repeated the process of sampling another 1,000 random voters, would we expect to get the same$\hat{\ }_{Biden}$?

Let's explore this in Jupyter!

Probability distribution of a statistic

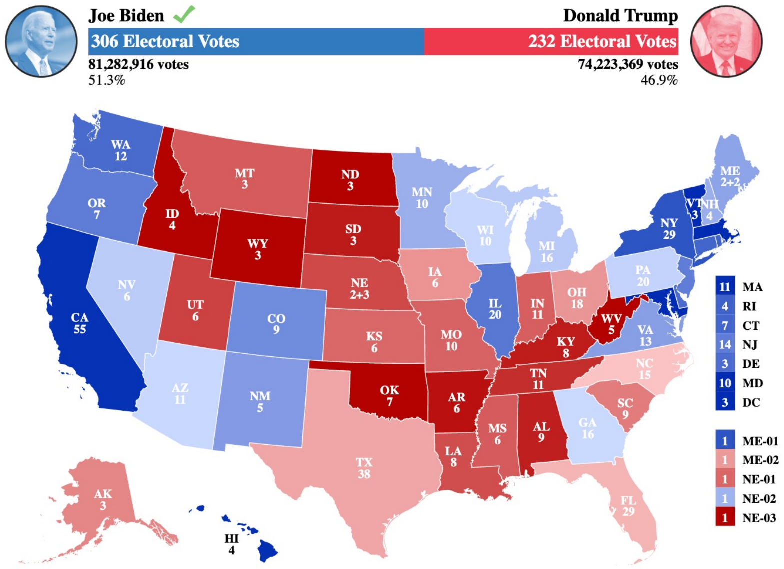
Values of a statistic vary because random samples vary

A **sampling distribution** is a probability distribution of *statistics*

- All possible values of the statistic and all the corresponding probabilities
- We can approximate a sampling distribution by a simulated statistics

π_{Biden}

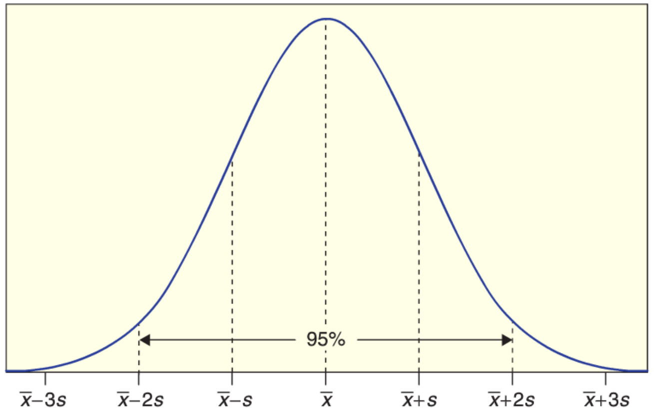
$n = 1{,}000$



$\hat{p}_{Biden}$

$\hat{p}_{Biden}$

$\hat{p}_{Biden}$



Sampling distribution!

Let's explore this in Jupyter!