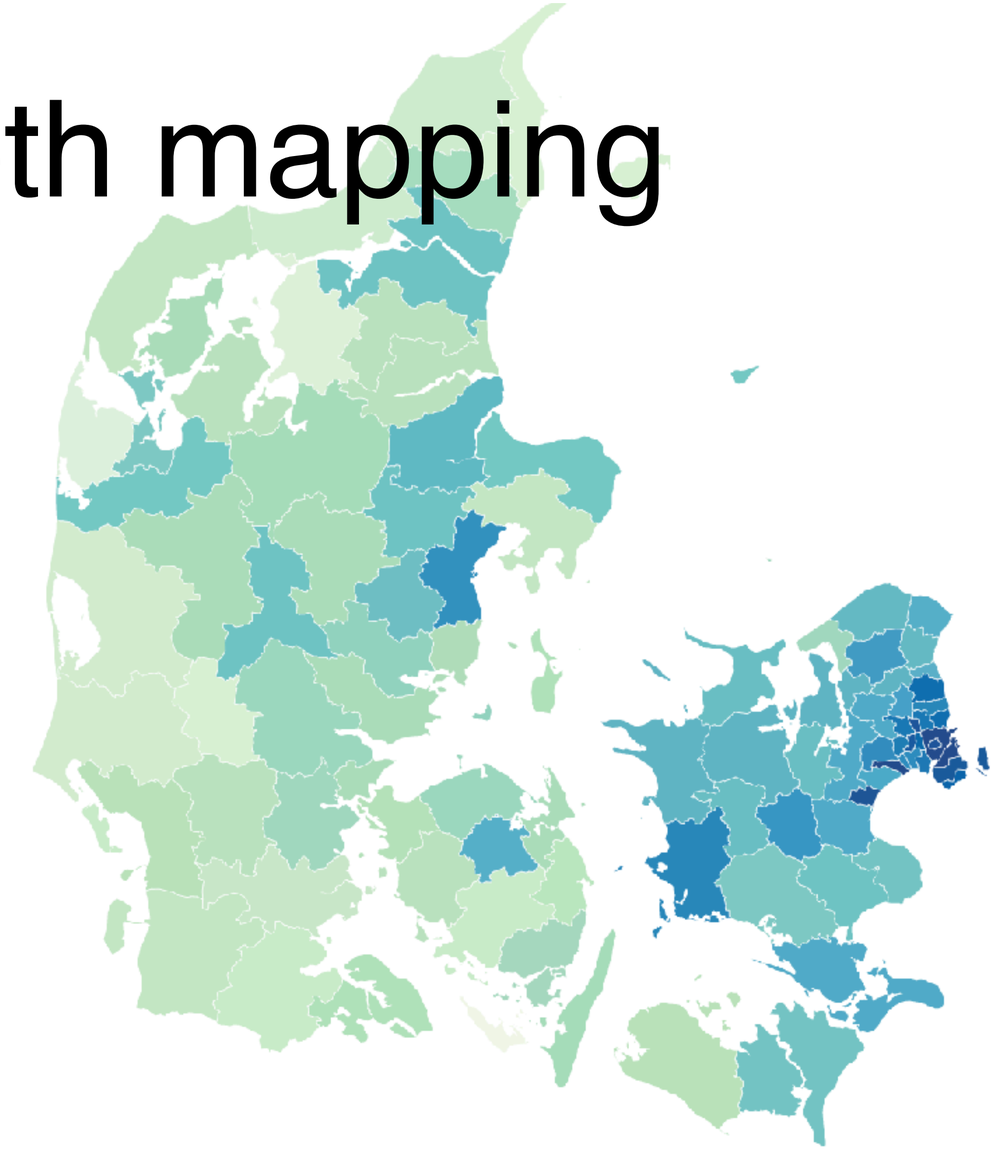
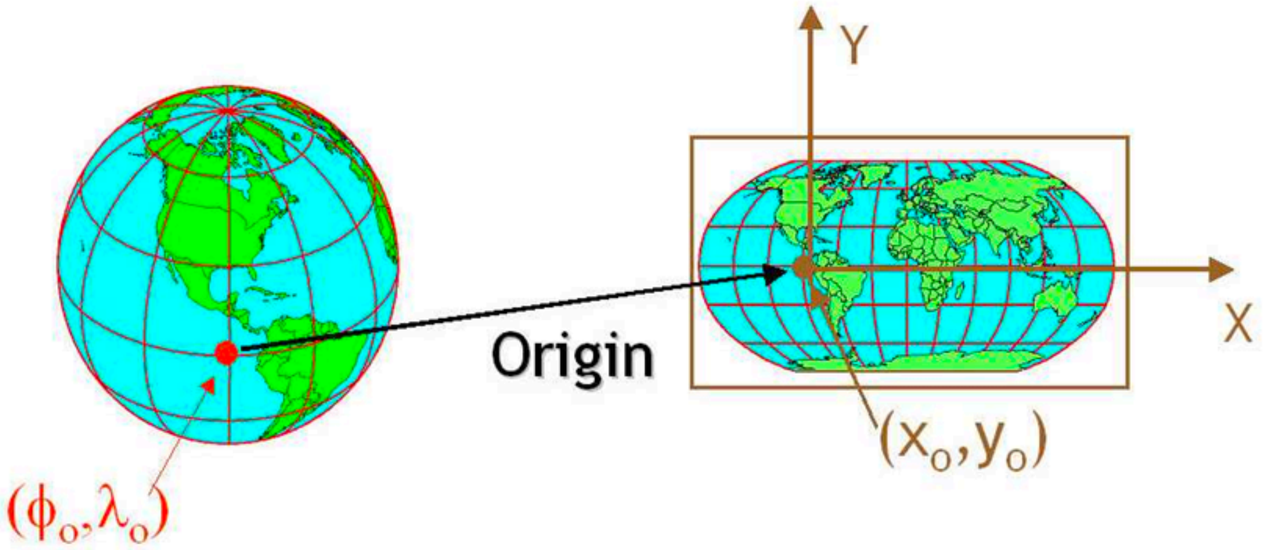
Lecture 3: Choropleth mapping

Instructor: Ane Rahbek Vierø
Feb 13, 2023



Recap from last week: Map projections & GeoPandas



GeoPandas

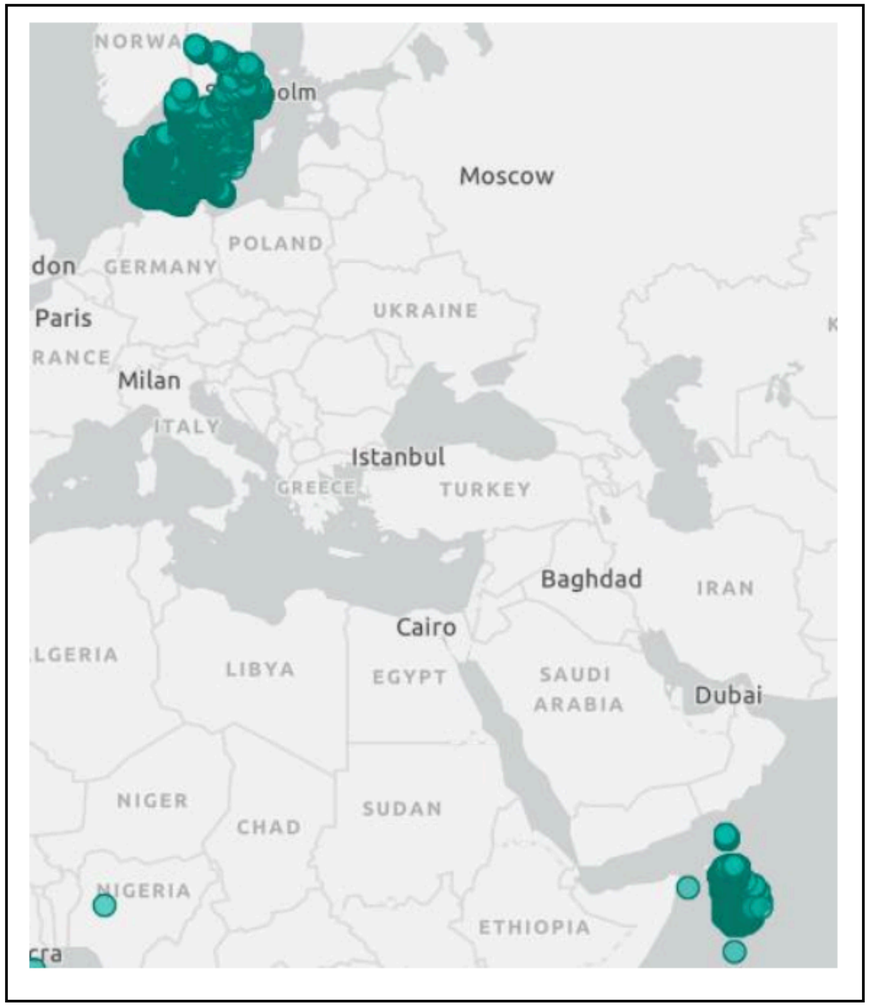
Errors in conversion of coordinates can have big effects!

Krimi

Politiet finder nye alvorlige fejl - stop for brug af teledata som bevis

Undersøgelse af Rigspolitiets håndtering af historiske teledata

1. oktober 2019



10,000 court cases had to be reviewed

…“has identified errors in connection with the conversion of geographic coordinates for tele communication towers…”

GDS ENV

If you use the Docker solution:

- Navigate to the folder with your notebooks and files before running docker

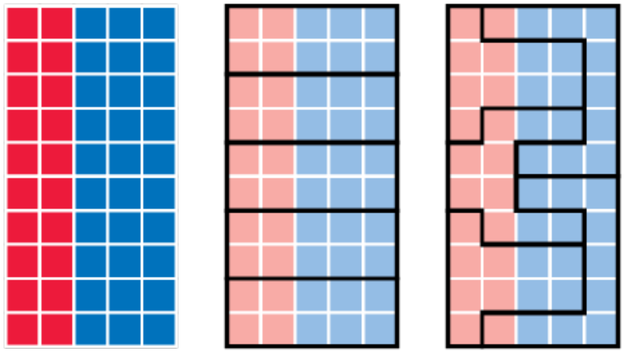
- Do not change the file path when running Docker (you *can* change PWD if you want)

- Avoid spaces in file paths if at all possible

```
docker run --rm -ti -p 8888:8888 -v ${PWD}:/home/jovyan/work darribas/gds_py:8.0
```

Today you will learn about MAUP and Choropleths

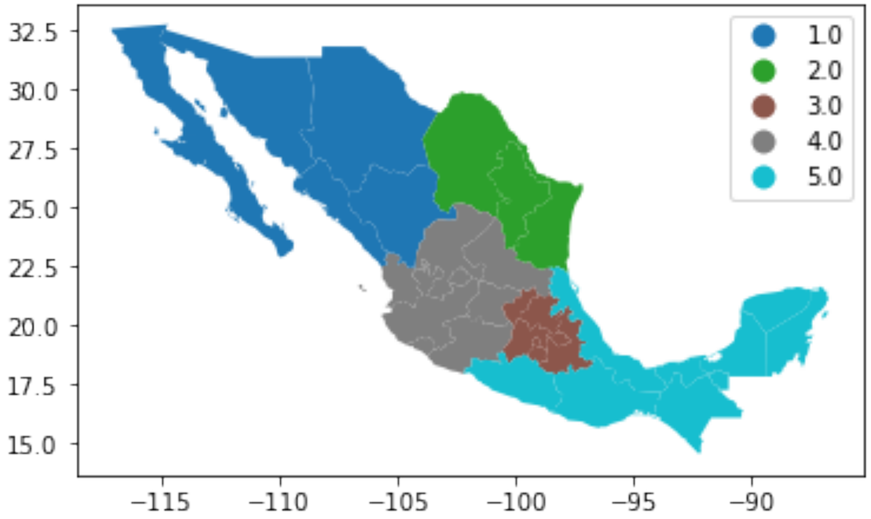
Biases from spatial aggregation (MAUP)



Classification schemes

$$c_j < y_i \leq c_{j+1} \ \forall y_i \in C_j$$

Choropleths in Python



Why do we make maps?

The most common steps in data preprocessing are:

Aggregation

Sampling

Dimensionality reduction

Discretization / Classification

Variable transformation

The most common steps in data preprocessing are:

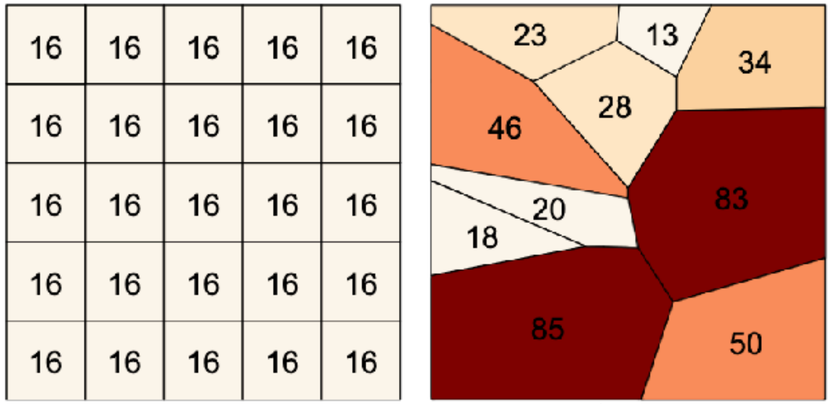
Aggregation

Sampling

Dimensionality reduction

Discretization / Classification

Variable transformation



Aggregation = Combining objects into a single one

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

NULL	Non-Freshman	3.375

Aggregation = Combining objects into a single one

Examples:

GPS coordinate → Zip Code → City → Country

Second → Minute → Hour → Day → Week → Month → Year

Advantages: Data reduction, easier to process, high-level view, smaller statistical fluctuations

Disadvantages: Loss of details, introducing biases

Choropleth maps visualize geospatial data with colors

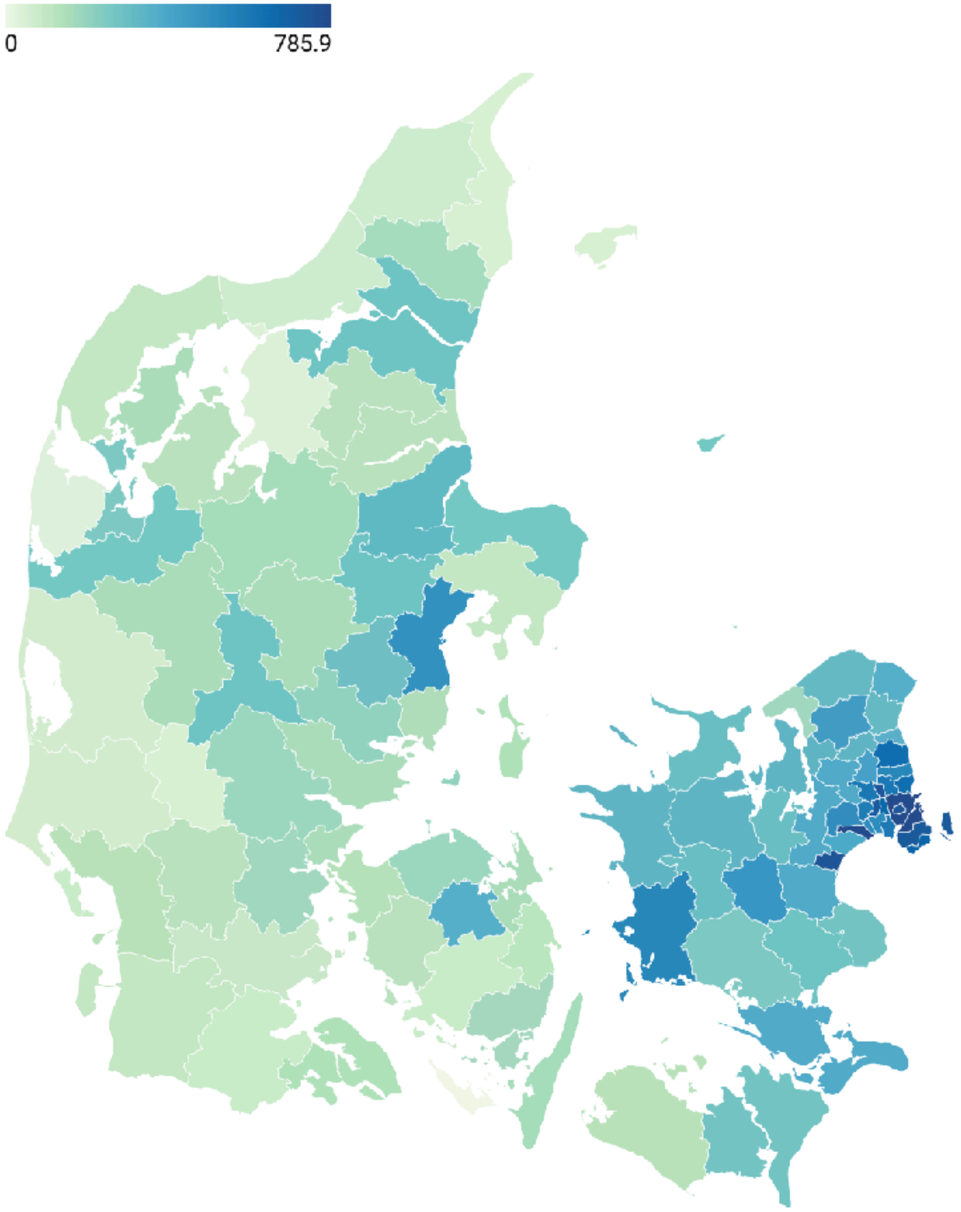
‘Choro’ = region
‘pleth’ = multiple

Combine: Polygons + Attribute values

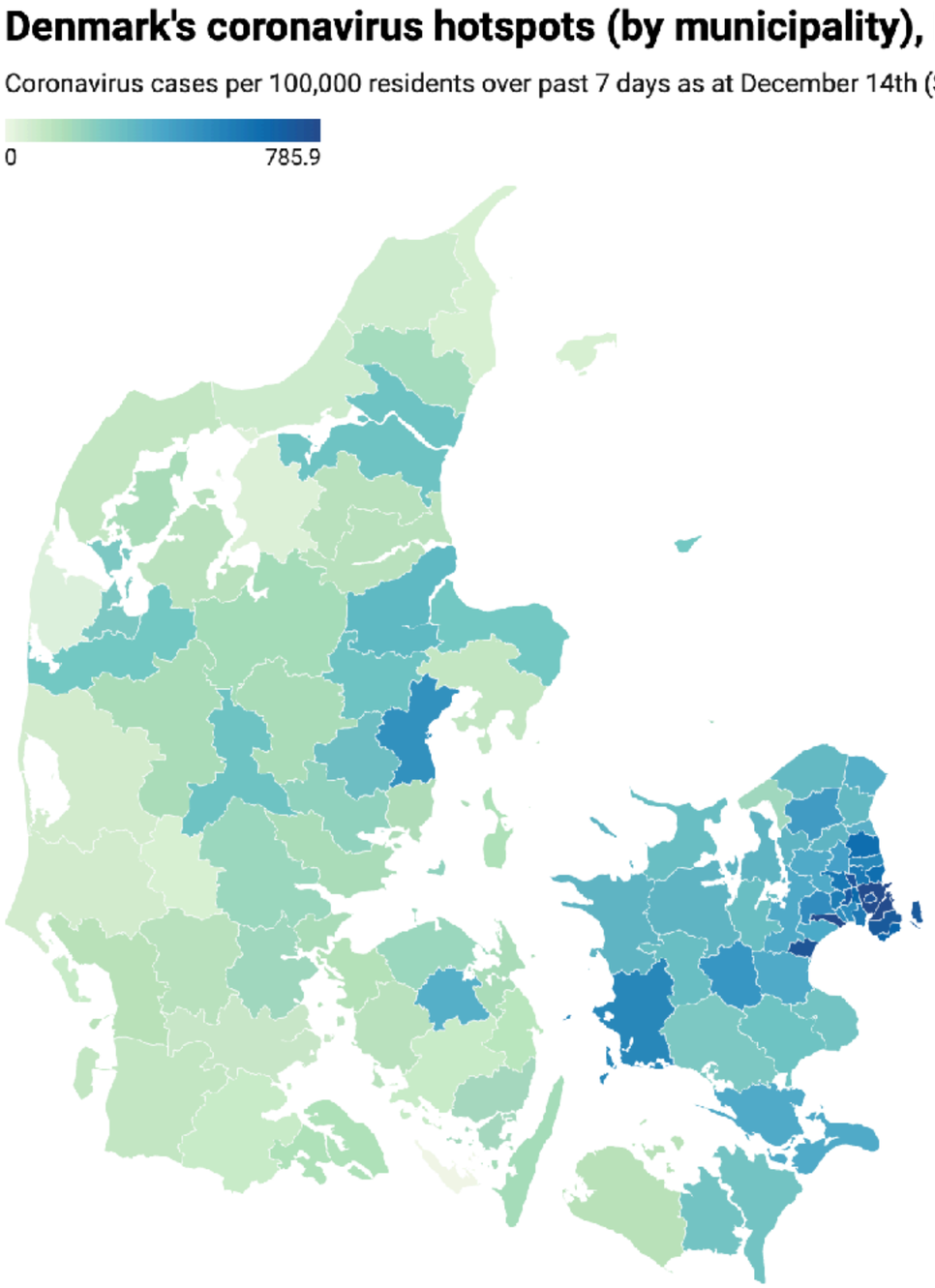
Values are often grouped into classes to reduce complexity

Denmark's coronavirus hotspots (by municipality), December 14th

Coronavirus cases per 100,000 residents over past 7 days as at December 14th (Source: SSI)



https://www.thelocal.dk/20201028/maps-where-are-denmarks-coronavirus-hotspots/

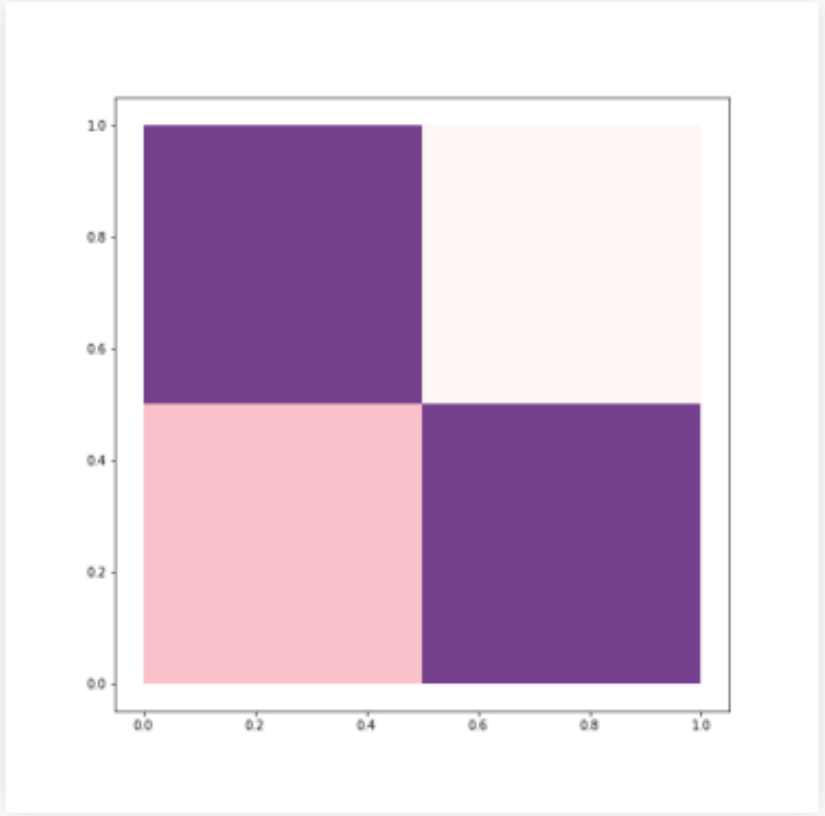


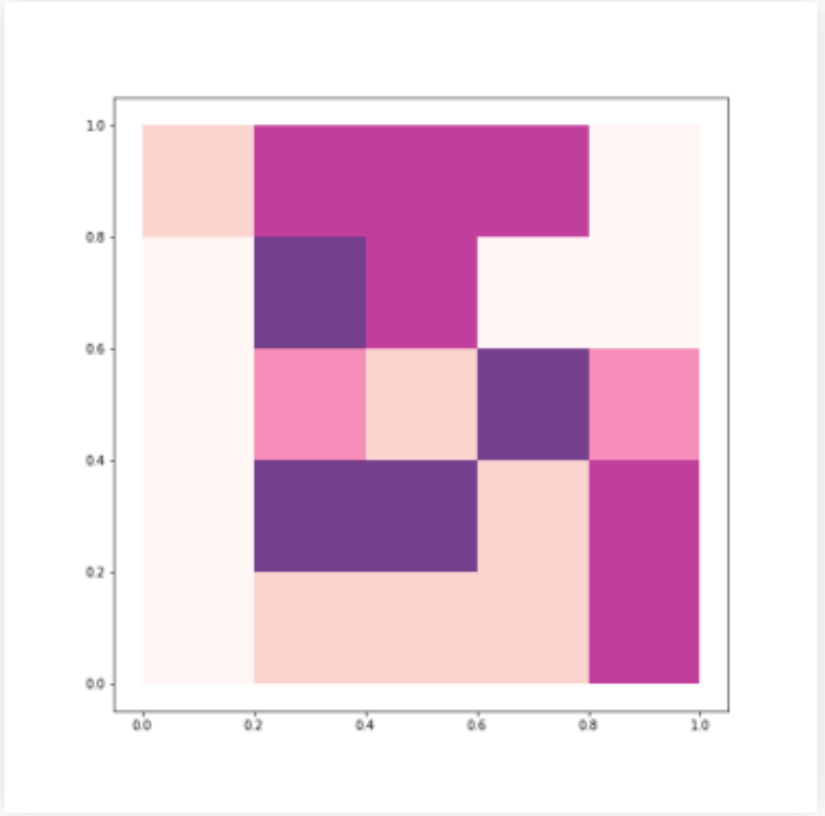
3 choices:

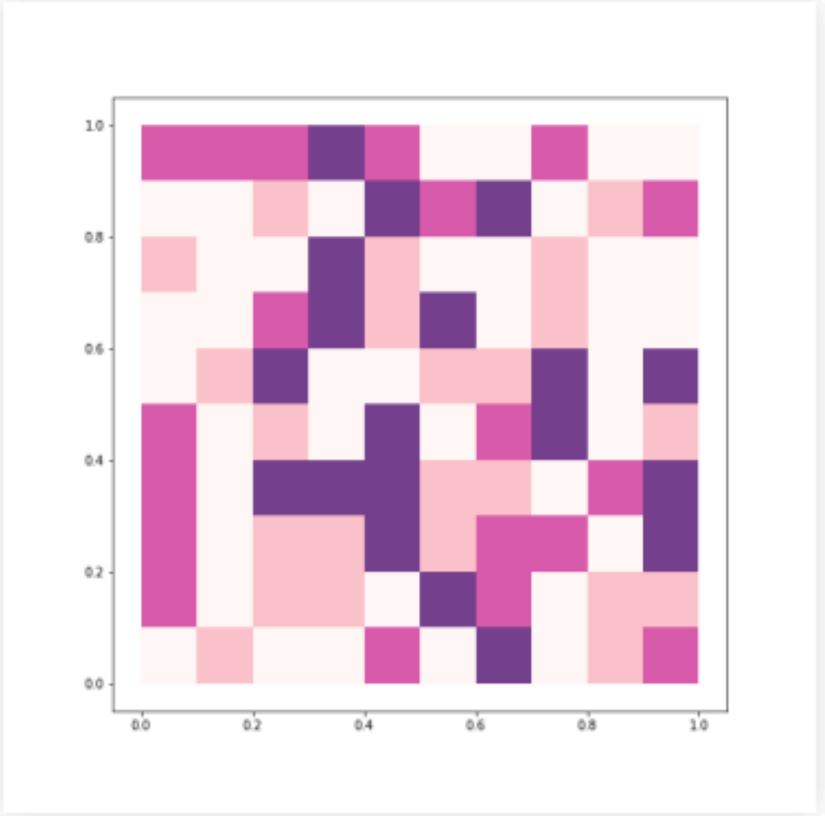
1) spatial units
2) colors
3) classes

https://www.thelocal.dk/20201028/maps-where-are-denmarks-coronavirus-hotspots/

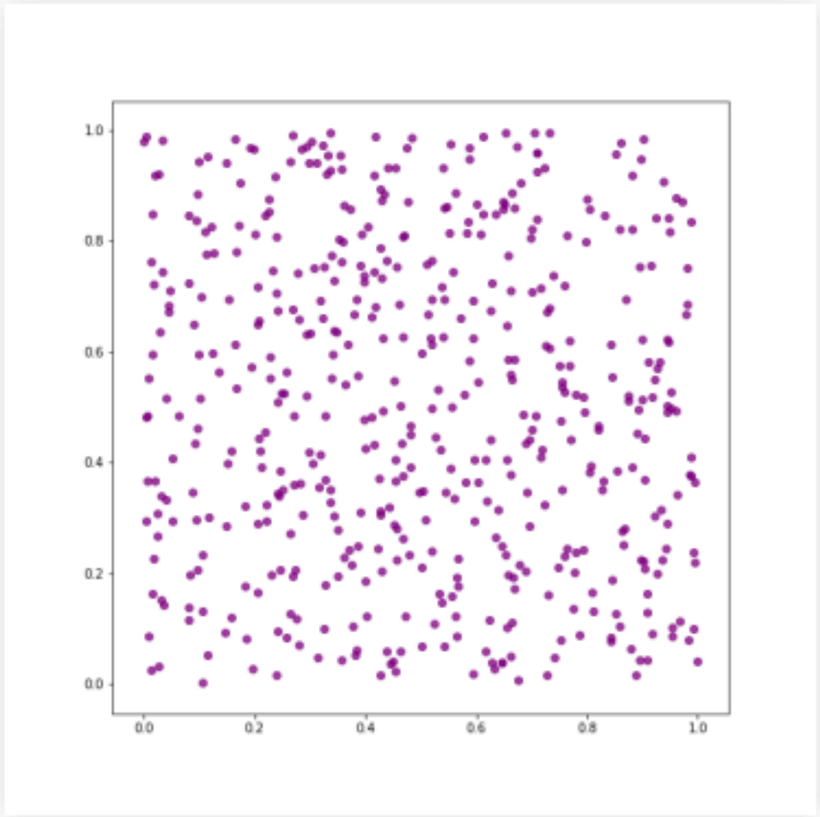
A common source of bias in spatial aggregation: MAUP

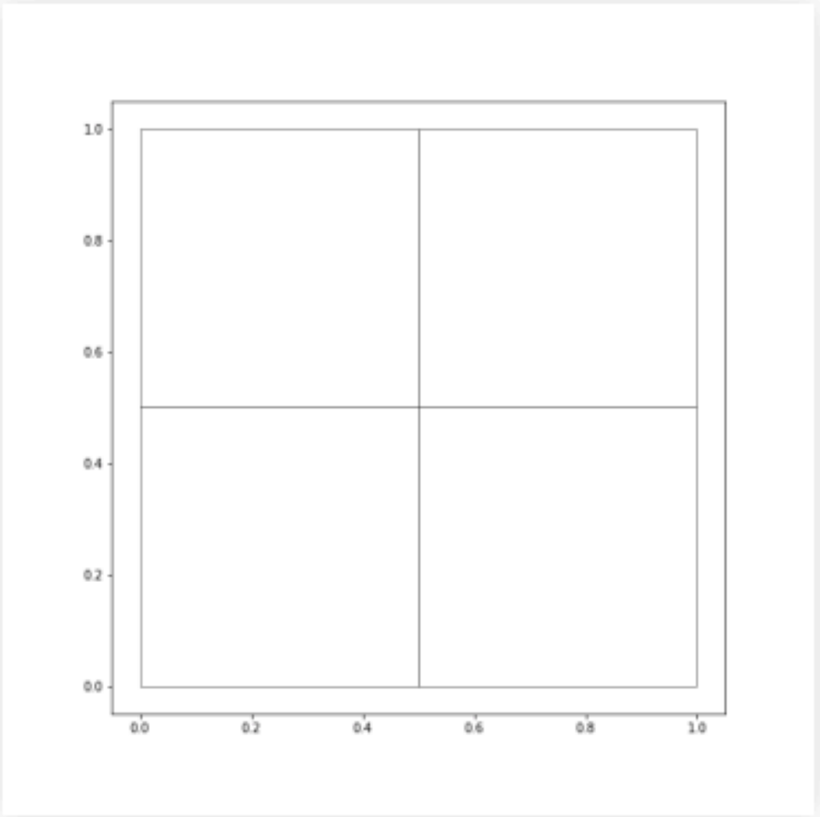


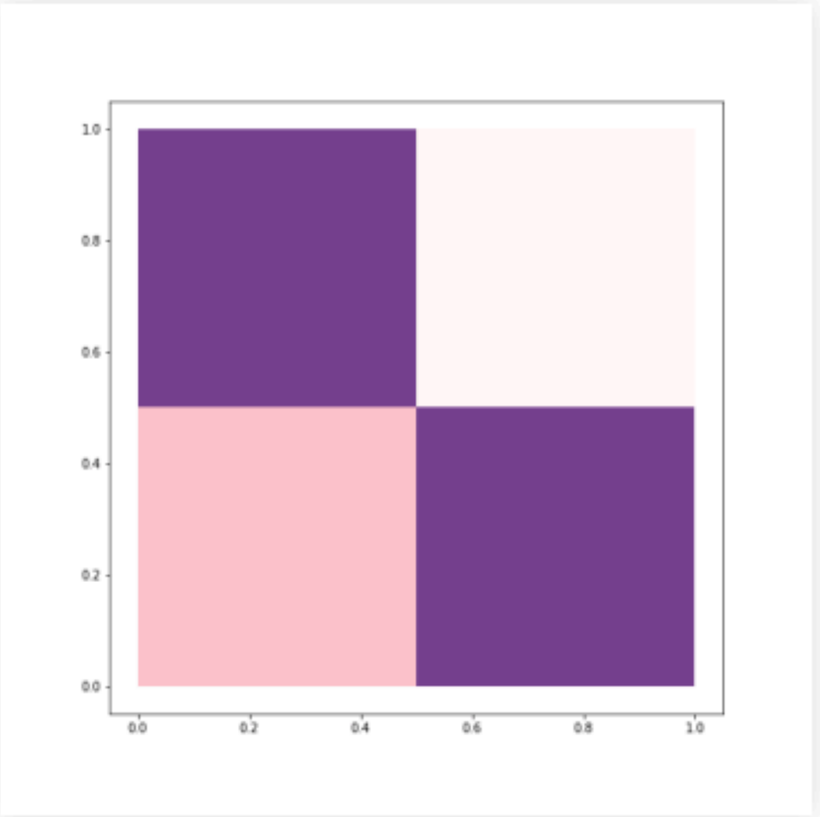




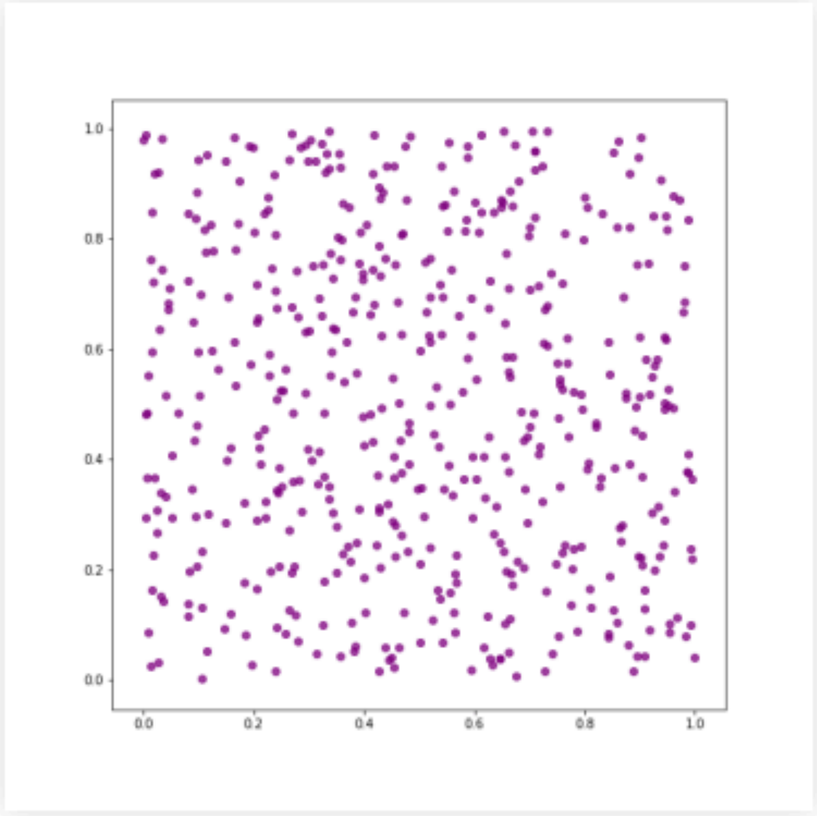
https://gist.github.com/darribas/8b5a7b93d4085223f1c5

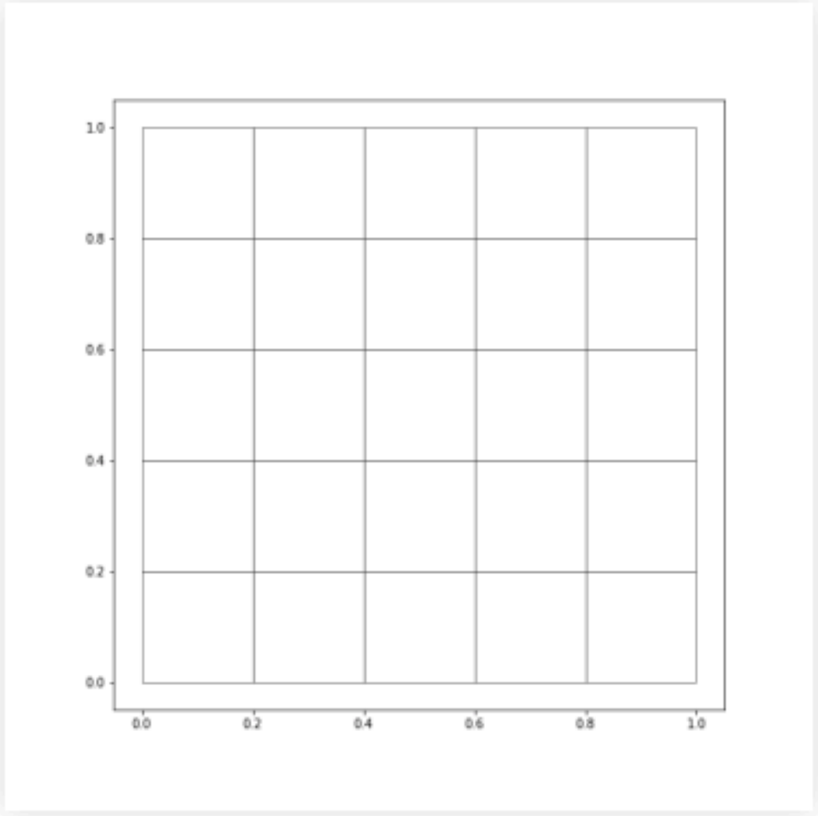


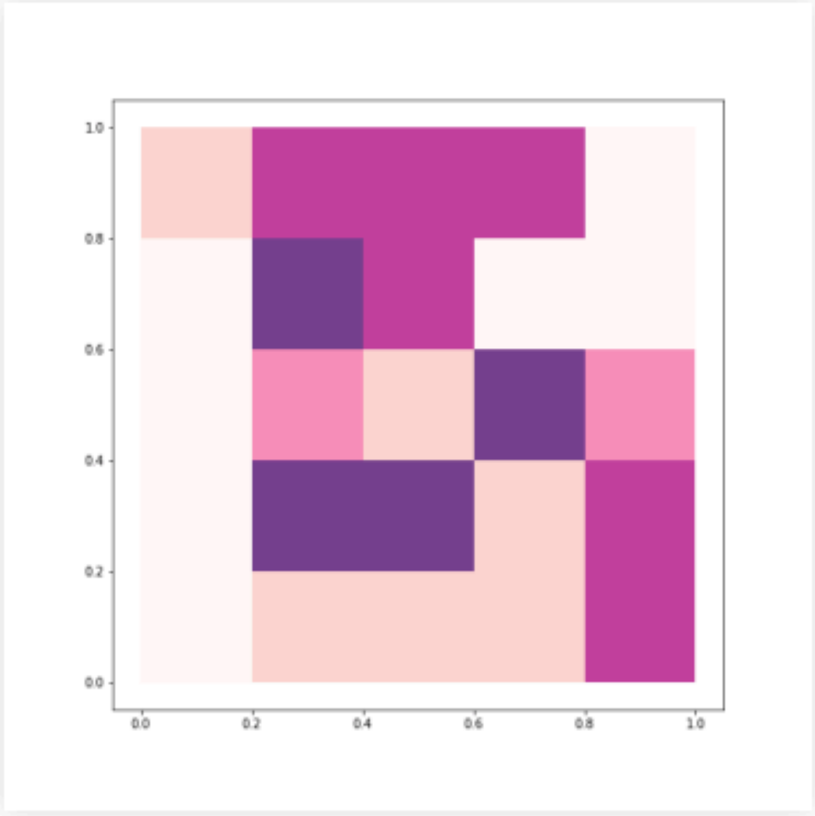




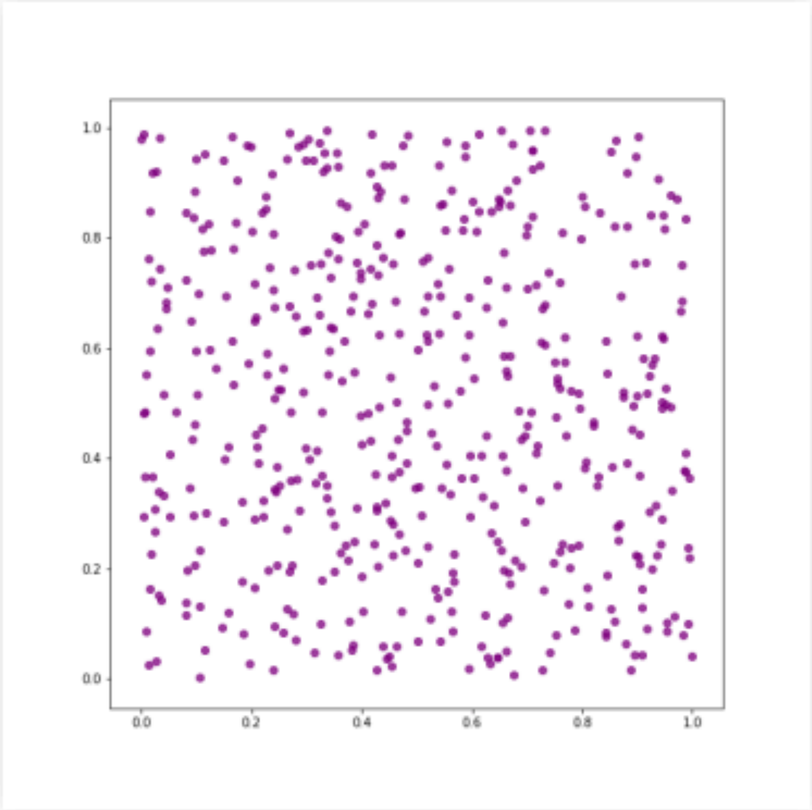
https://gist.github.com/darribas/8b5a7b93d4085223f1c5

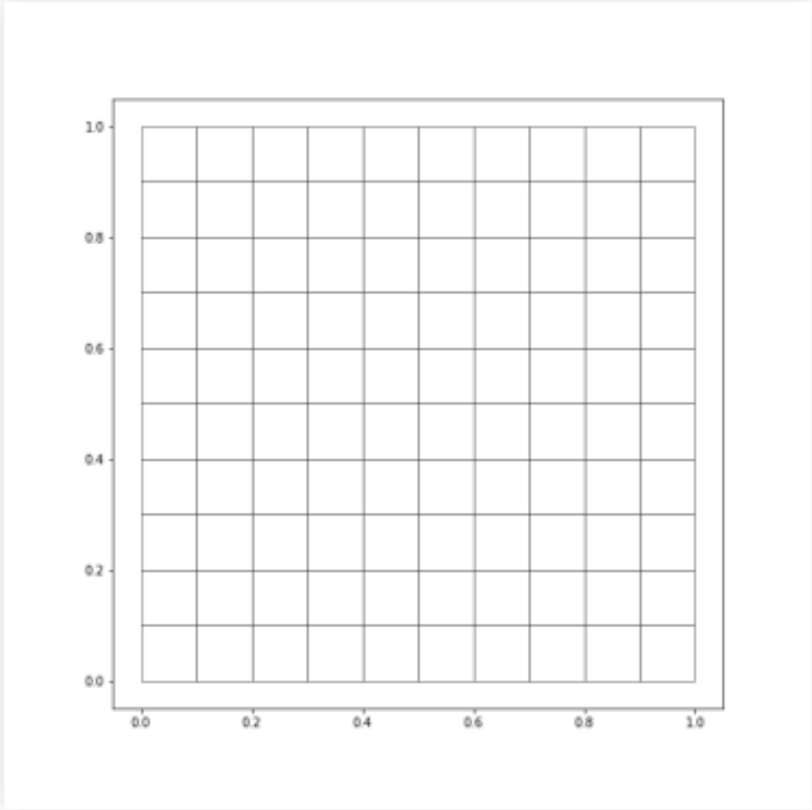


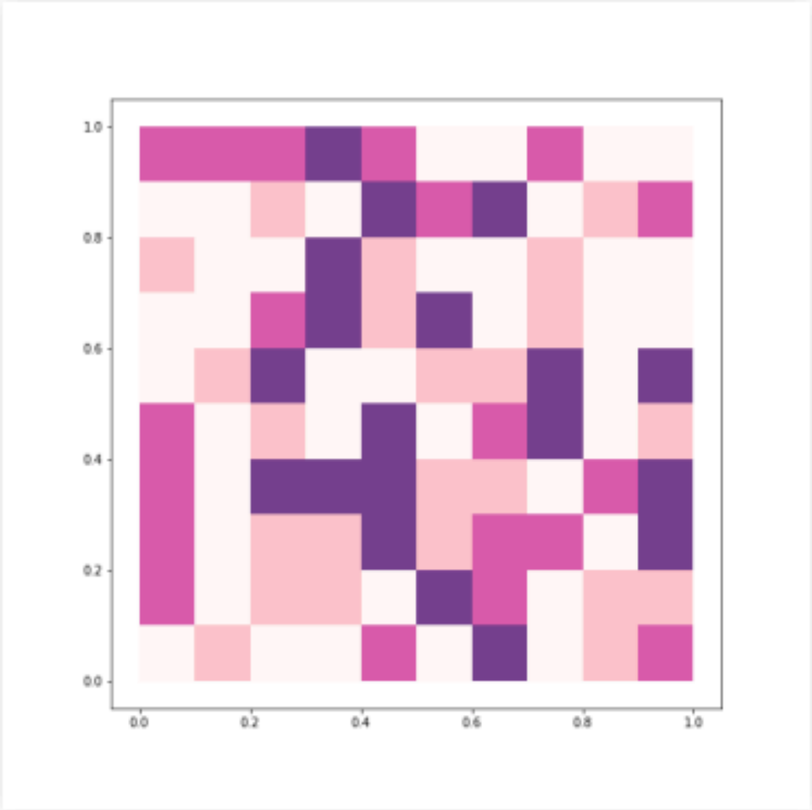




https://gist.github.com/darribas/8b5a7b93d4085223f1c5





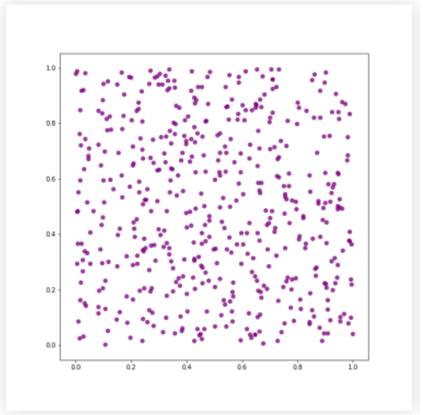


https://gist.github.com/darribas/8b5a7b93d4085223f1c5

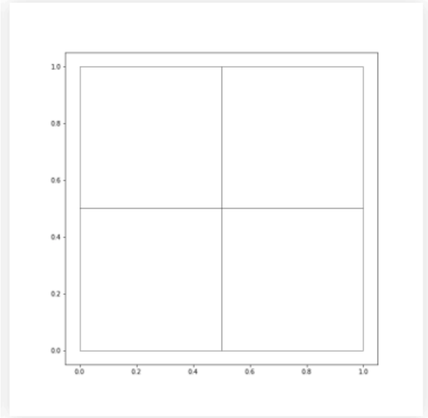
A common source of bias in spatial aggregation: MAUP

The MAUP (Modifiable Areal Unit Problem) is a scale and delineation mismatch between:

Underlying entities $\longleftrightarrow$ Unit of measurement



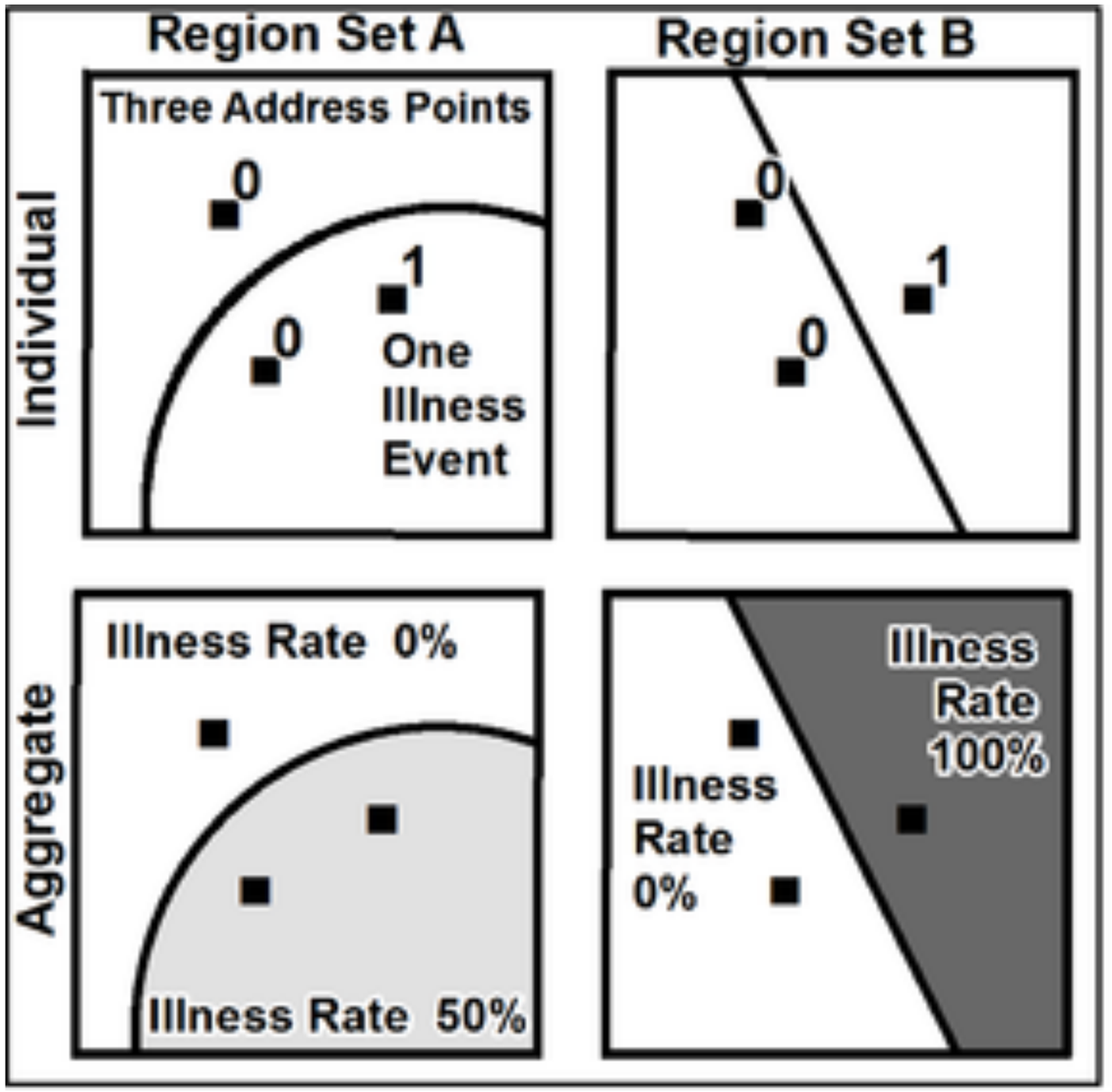
Individuals, shops,..



Districts, regions,…

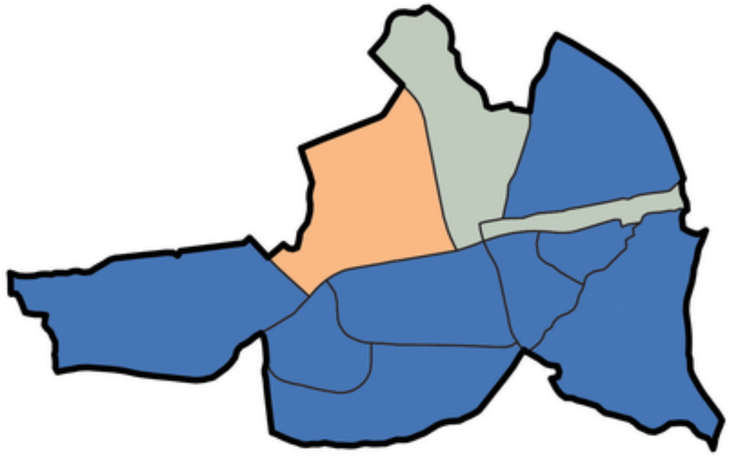
Openshaw, 1984

A common source of bias in spatial aggregation: MAUP

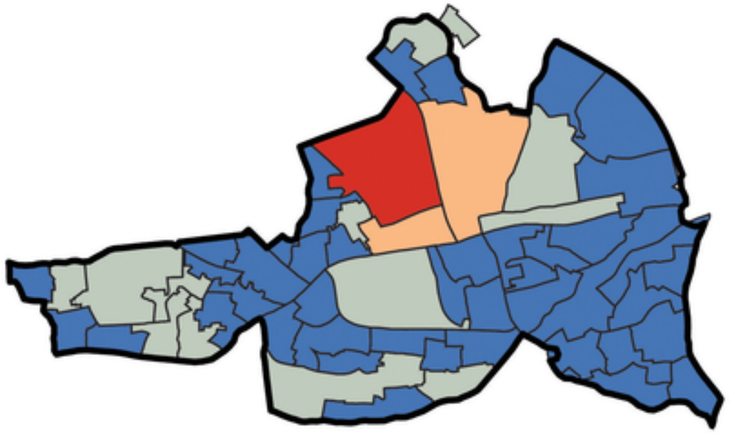


A common source of bias in spatial aggregation: MAUP

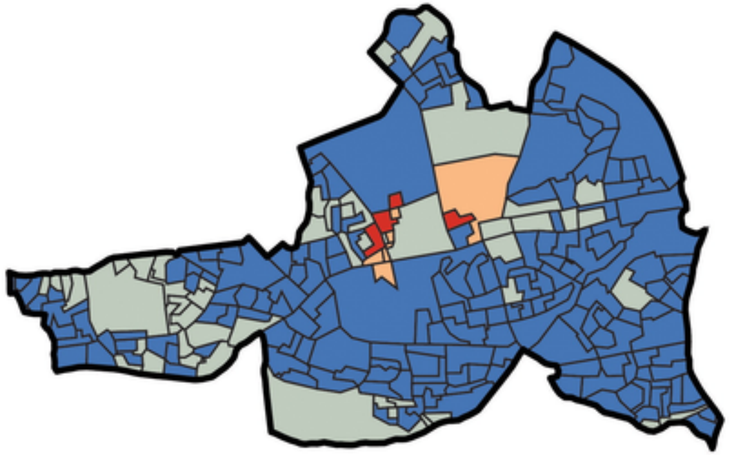
1. Electoral Divisions (EDs), 11 in Study Area



2. Enumerator Area (EAs), 56 in Study Area



3. Small Areas (SAs), 237 in Study Area



Modifiable areal unit problem (MAUP)
Study Area: Tallaght, Dublin
Housing Vacancy Rate, 2011

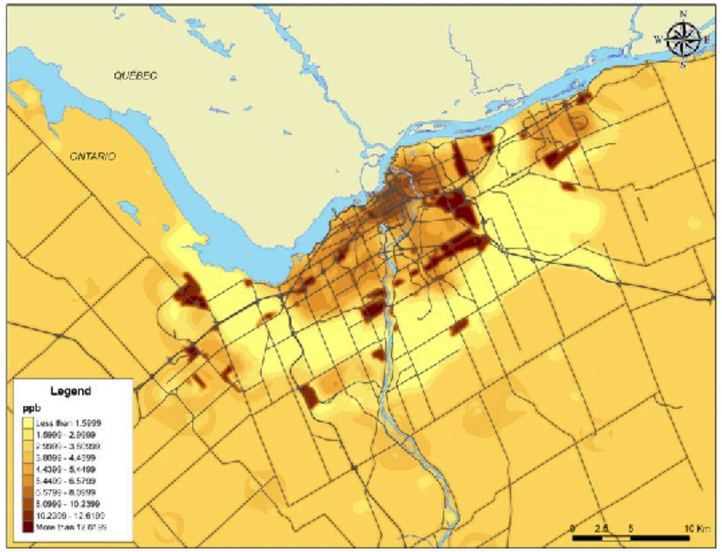
- < 5%
- 5% to < 15%
- 15% to < 30%
- 30% plus

https://www.tandfonline.com/doi/full/10.1080/21681376.2014.983149

A common source of bias in spatial aggregation: MAUP

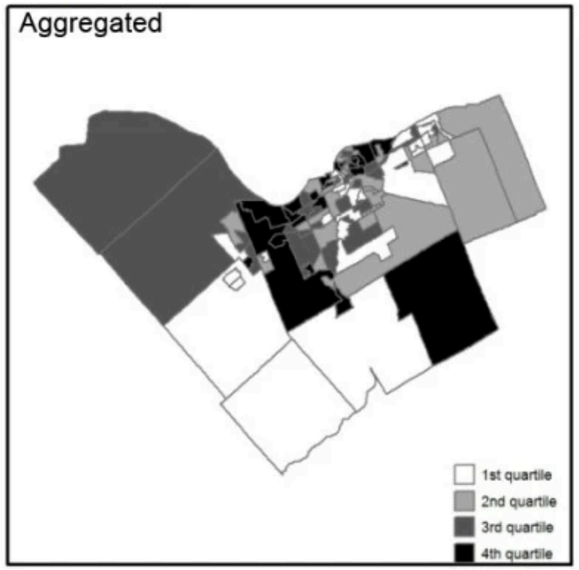
politiken.dk

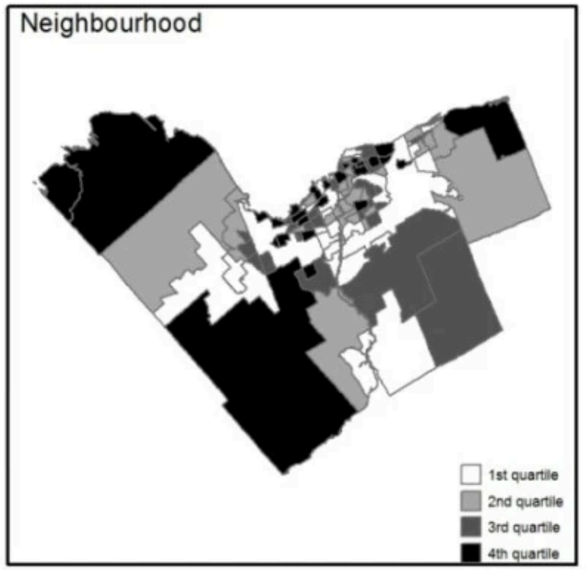
A common source of bias in spatial aggregation: MAUP

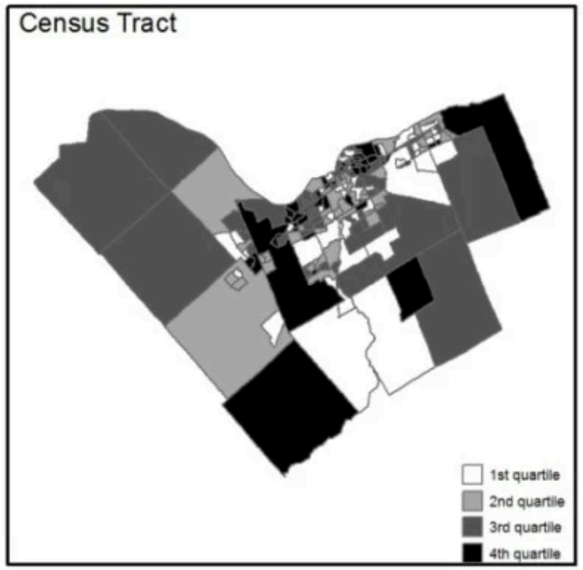


Pollution

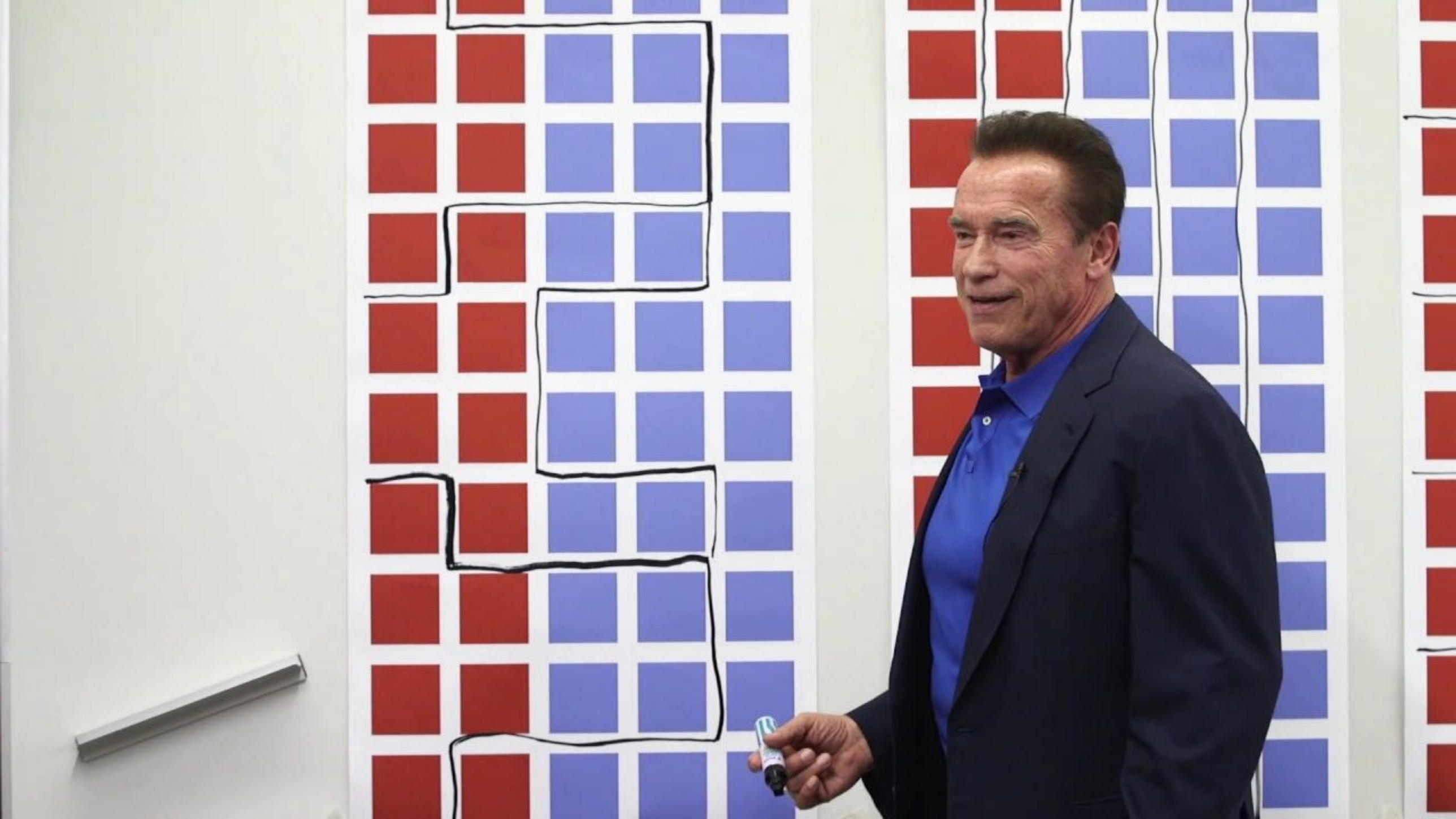
Morbidity



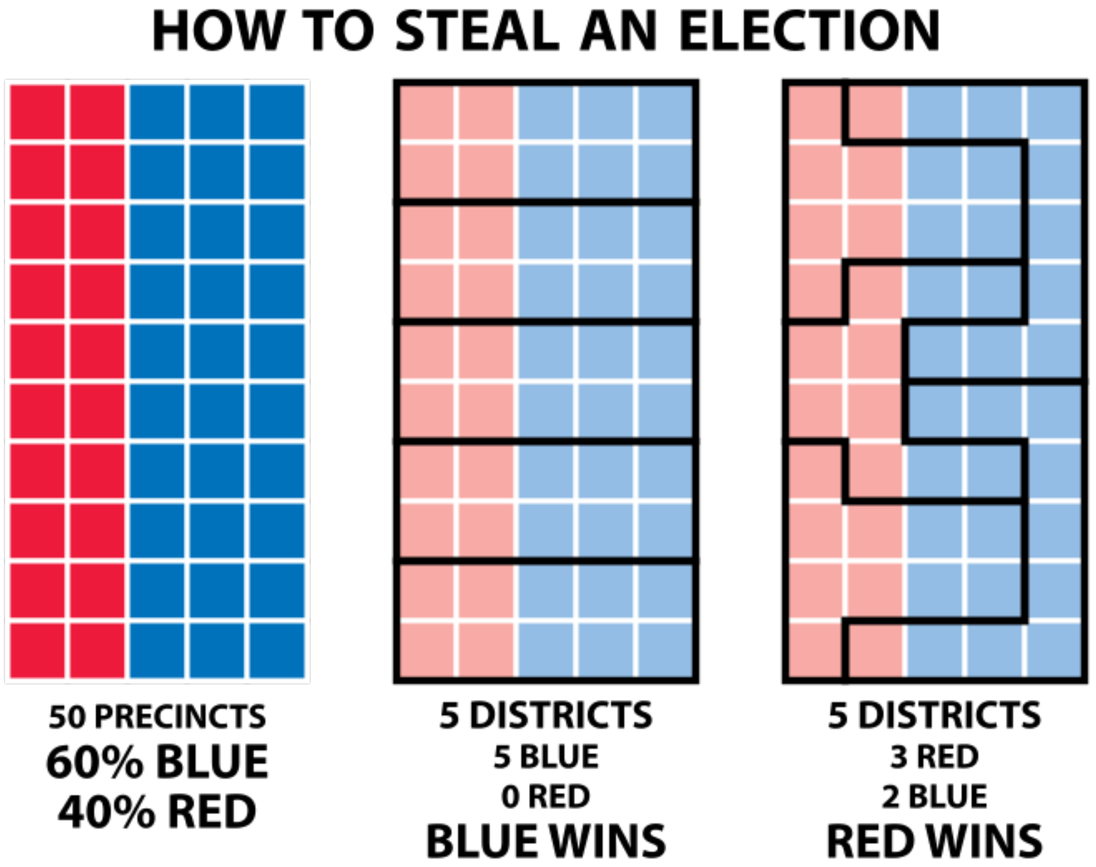


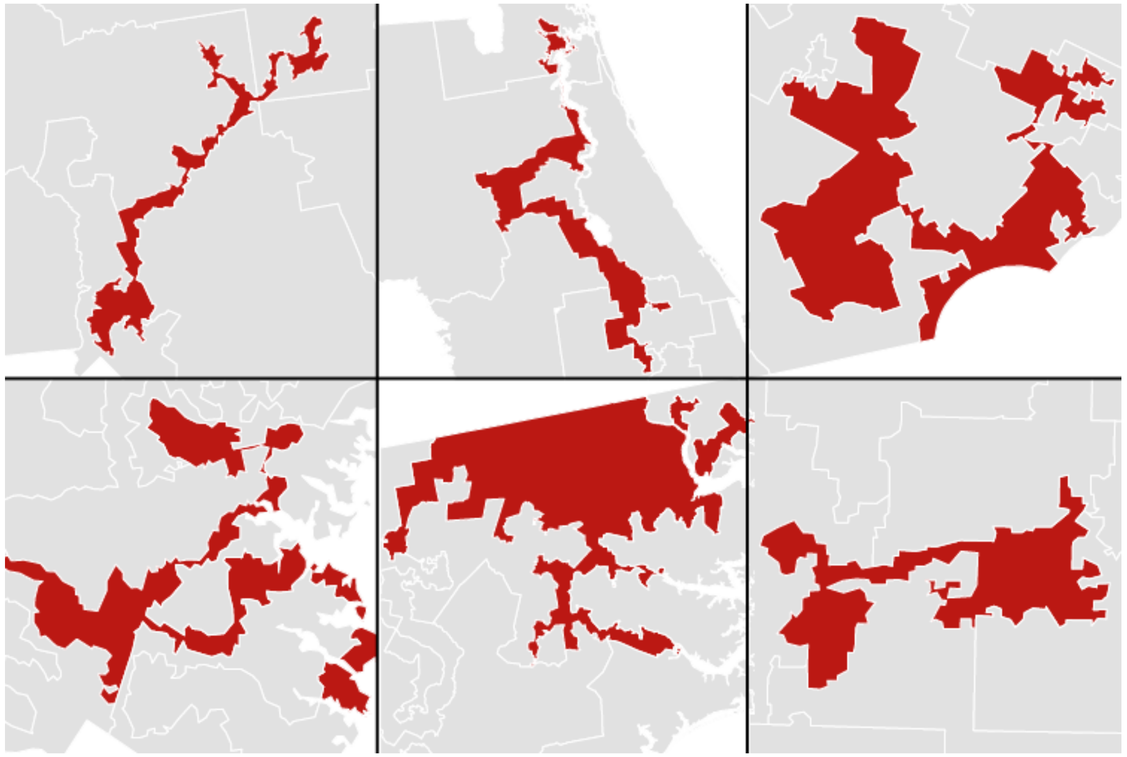


https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-10-58



The MAUP is abused for Gerrymandering





Can You Gerrymander Your Party to Power?

By Ella Koeze, Denise Lu and Charlie Smart

Gerrymandering is the intentional distortion of political districts to give one party an advantage, and it has been criticized for disenfranchising many voters and fueling deeper polarization.

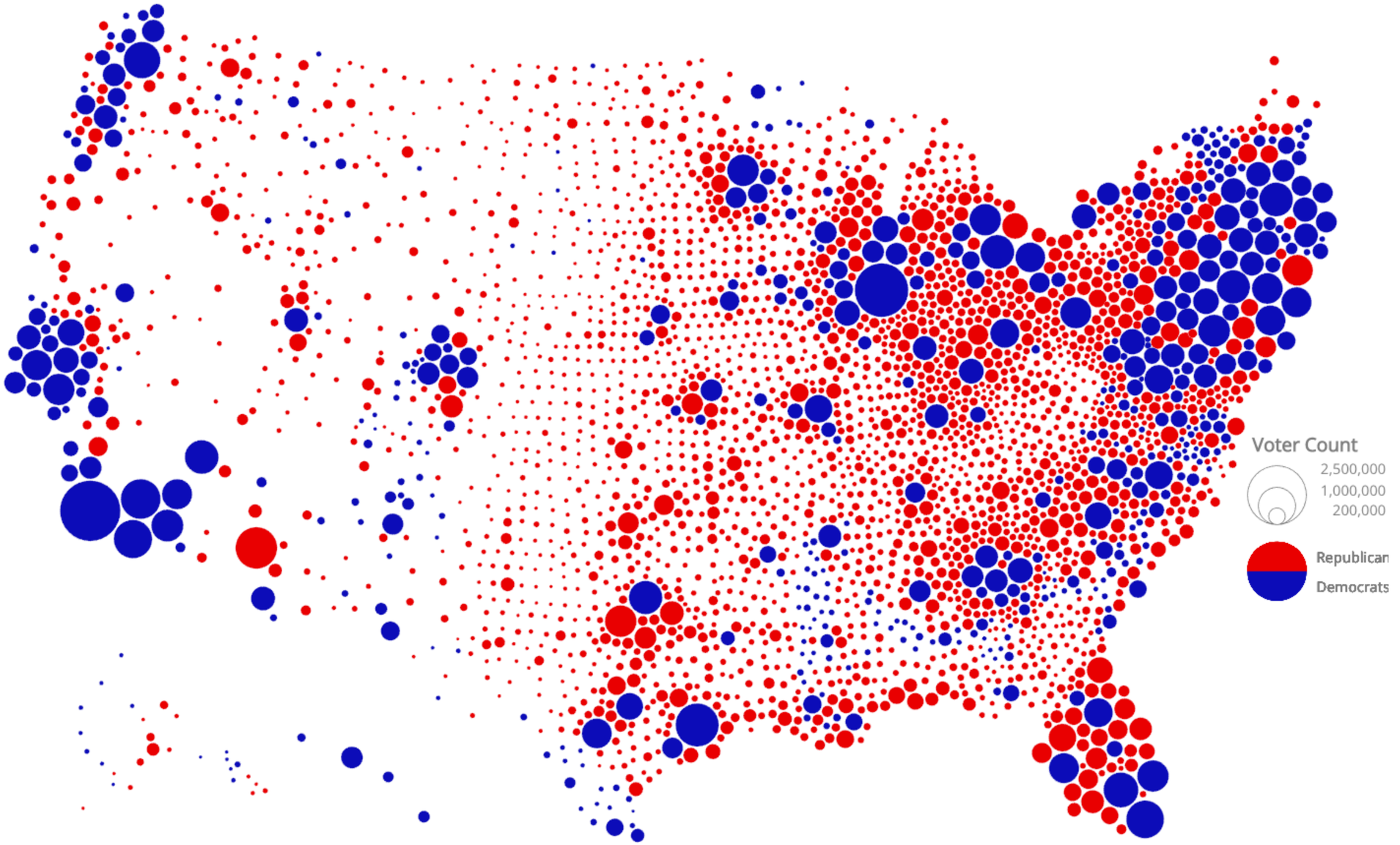
To help you understand it better, we created an imaginary state called **Hexapolis**, where your only mission is to gerrymander your party to power.

https://tinyurl.com/a72r229w

Related bias: Area is not population

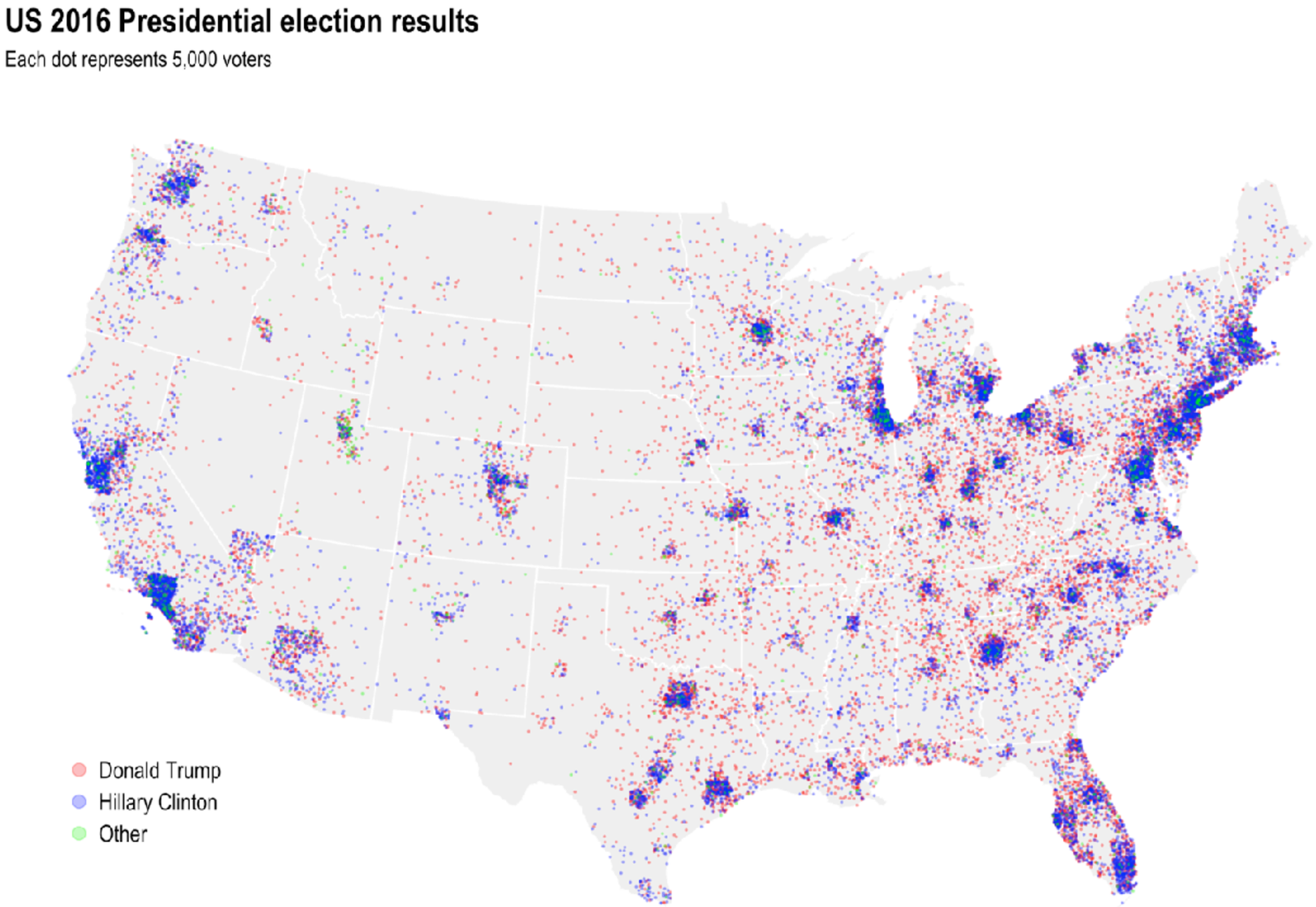
http://try-to-impeach-this.jetpack.ai/

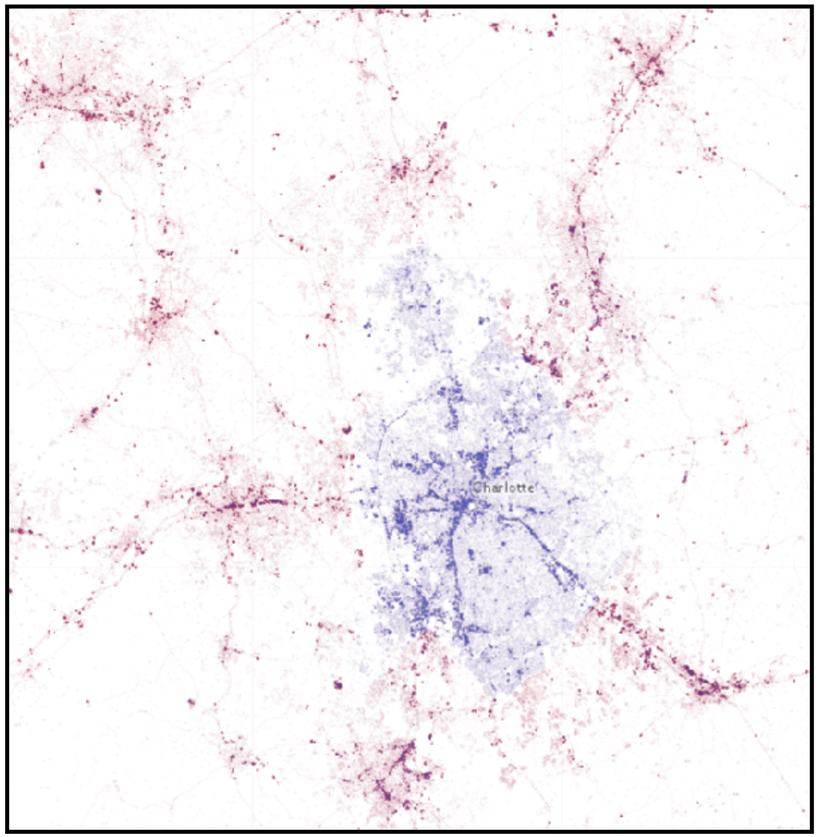
Related bias: Area is not population



http://try-to-impeach-this.jetpack.ai/

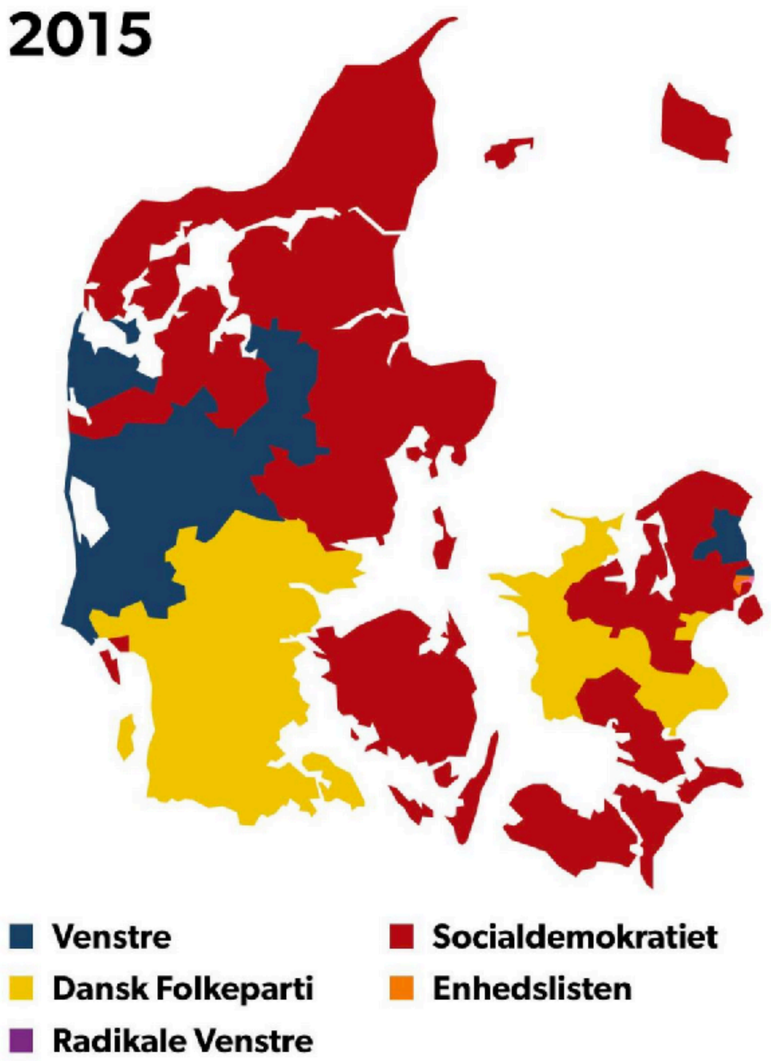
A dot density map can reduce the bias

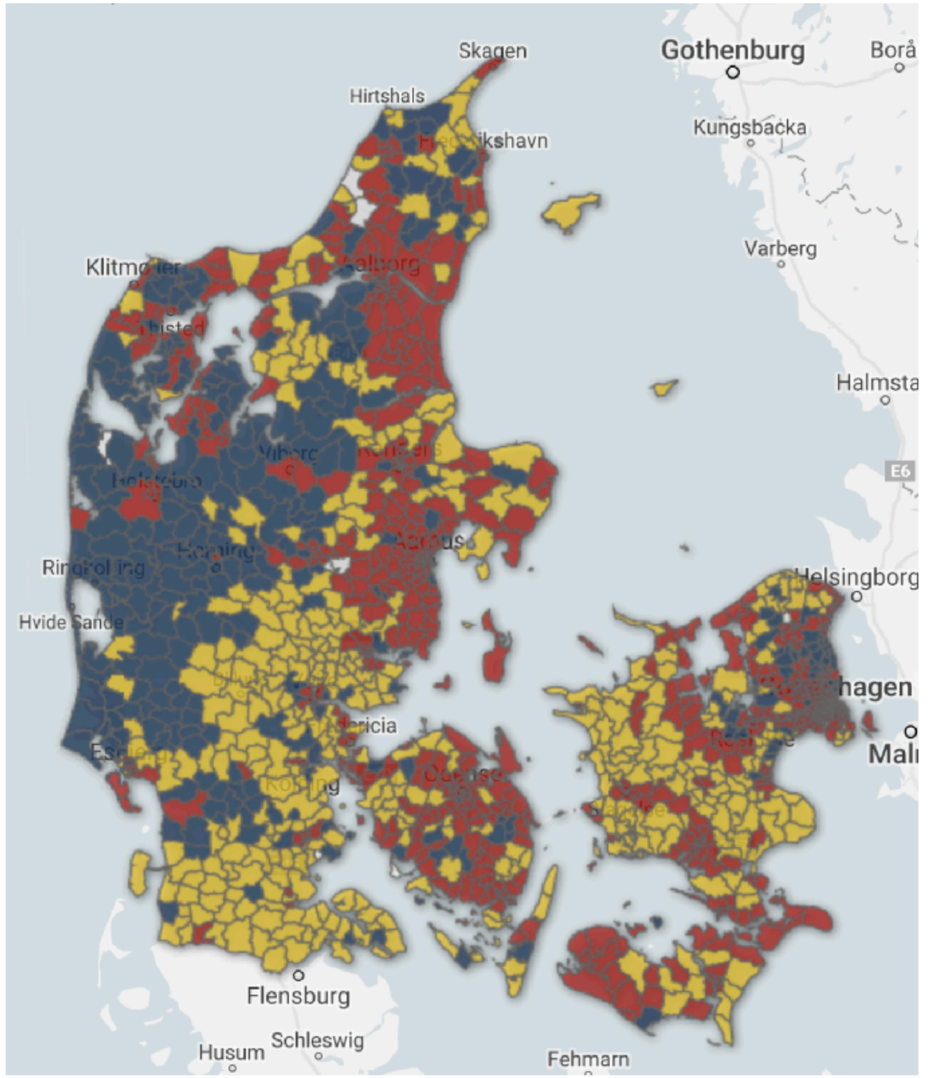




https://www.andybeger.com/2018/05/11/dot-density-map-of-the-2016-election/
https://www.maproomblog.com/2018/04/kenneth-fields-dot-density-election-map-redux/

Poorly aggregated data exaggerate regional differences





Always mind the MAUP when exploring aggregated data

Classification: An intellectual process that groups similar phenomena to gain relative simplicity in communication and user interpretation. Also known as **binning**.

Classification = Transforming continuous into categorical

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

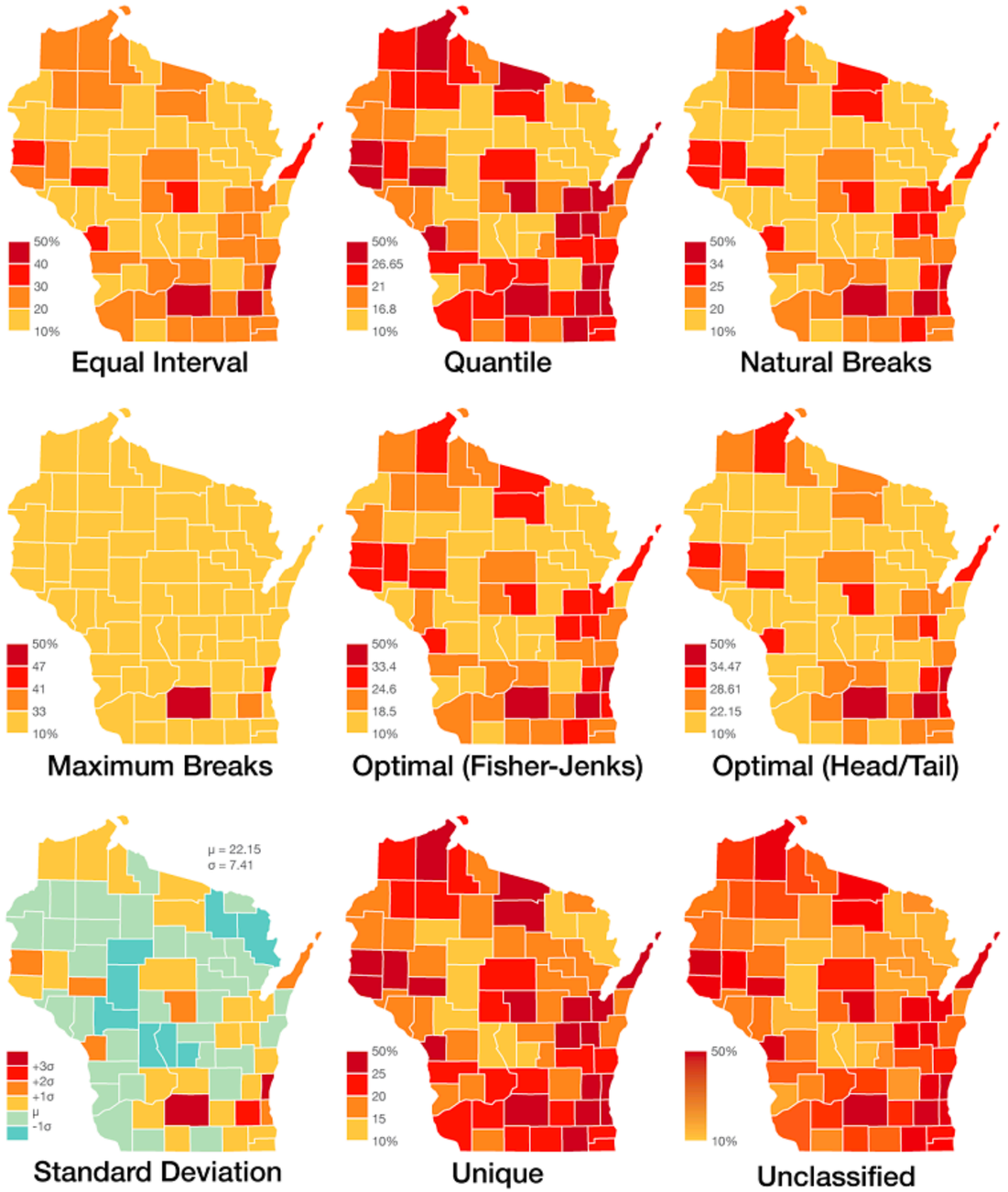
reject
accept
accept

How many categories should there be? If 2: Binarization
How should the values be mapped? What should be the thresholds?

Data can be classified very differently

Same data, different split points

Percentage of residents over 25 with a Bachelor's degree



https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

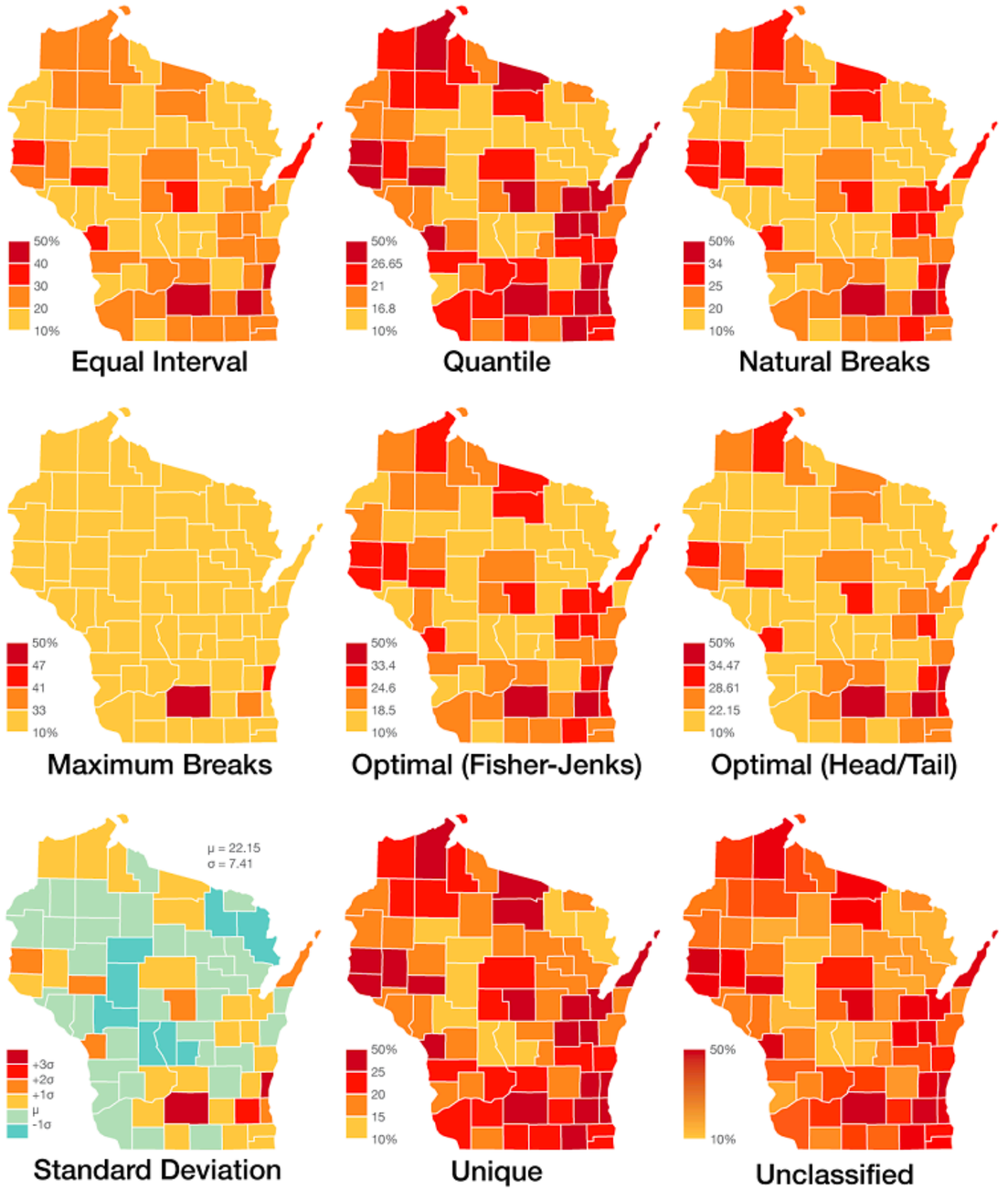
Data can be classified very differently

How you do this depends on your:

- Audience
- Goal (highlight outliers, show distribution, deceive ..)

A data set does not have a “perfect" choropleth map

Percentage of residents over 25 with a Bachelor's degree



https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Formally, you must select thresholds to classify your data

The classification problem is to define class boundaries such that

$$c_j < y_i \leq c_{j+1} \; \forall y_i \in C_j$$

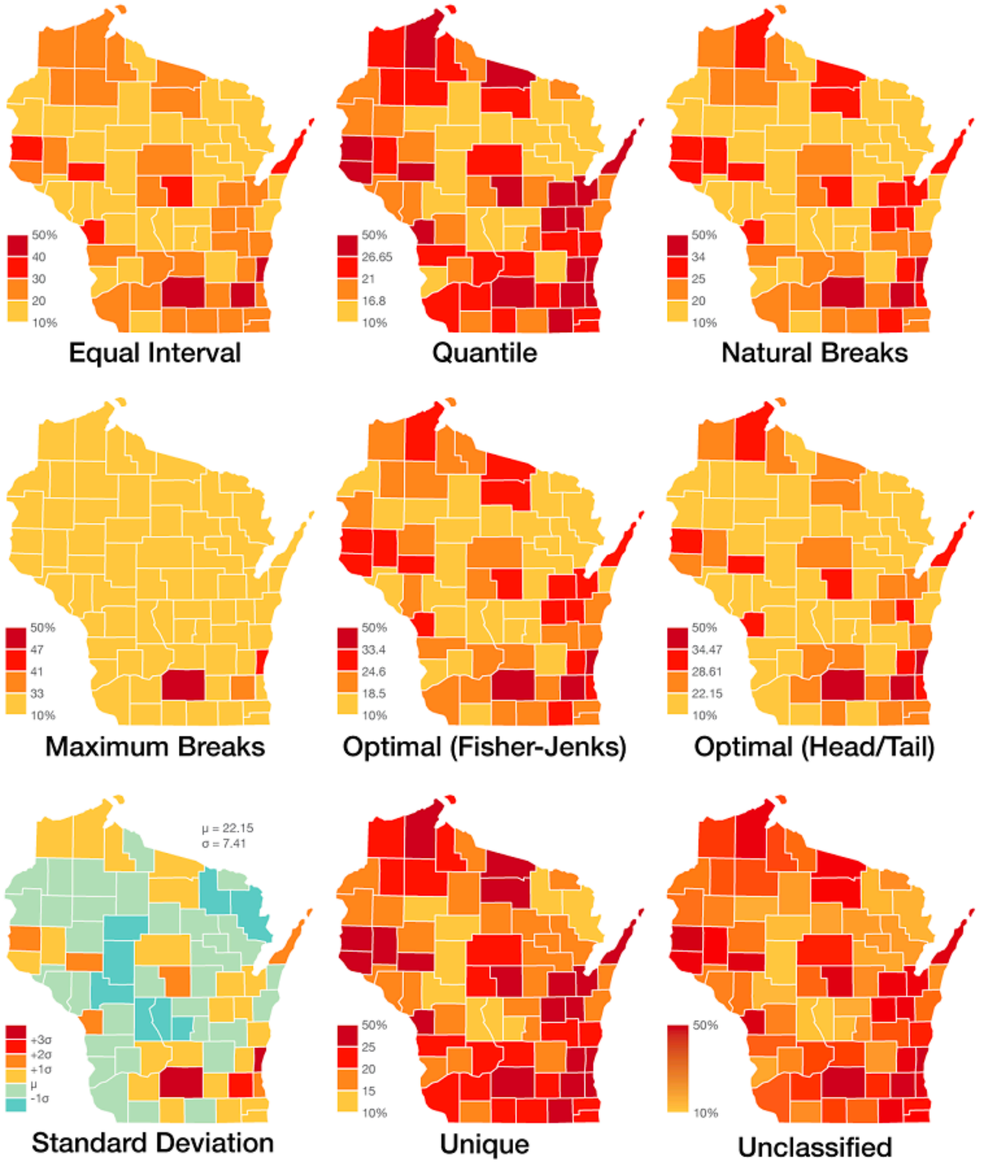
where y_i is the value of the attribute for spatial location i,
j is a class index,
c_j represents the lower bound of interval j.

The choice of class boundaries defines the classification scheme

mapclassify:

```
  'EqualInterval',
  'FisherJenks',
…
  'HeadTailBreaks',
  'JenksCaspall',
…
  'MaximumBreaks',
  'NaturalBreaks',
  'Percentiles',
…
  'Quantiles',
  'StdMean',
  'UserDefined',
```

Percentage of residents over 25 with a Bachelor's

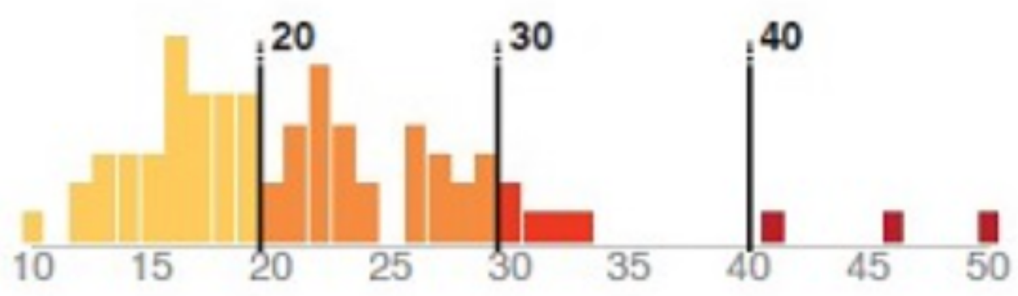


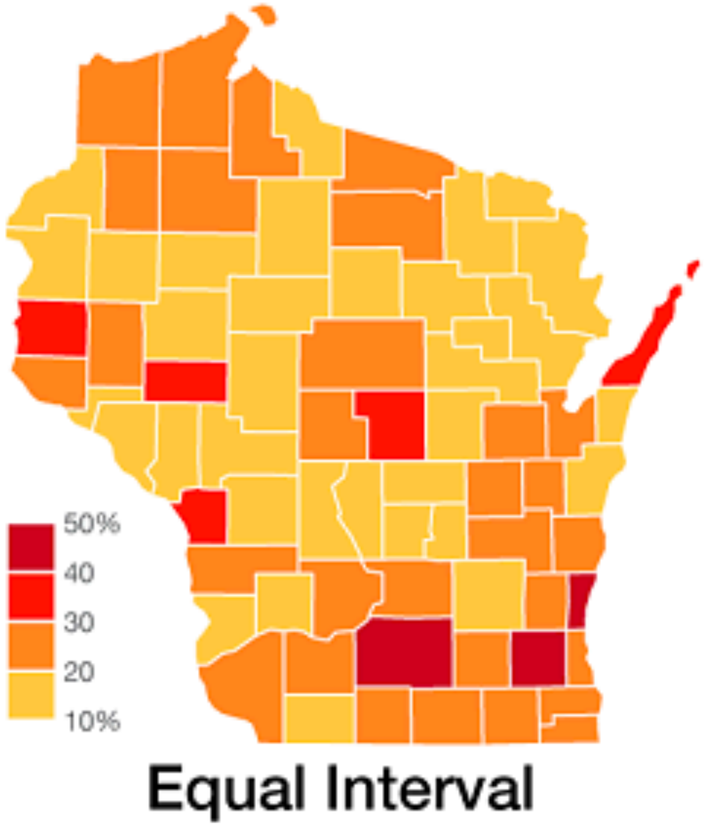
https://pypi.org/project/mapclassify/

https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Equal Interval

For uniformly distributed data with familiar data ranges.

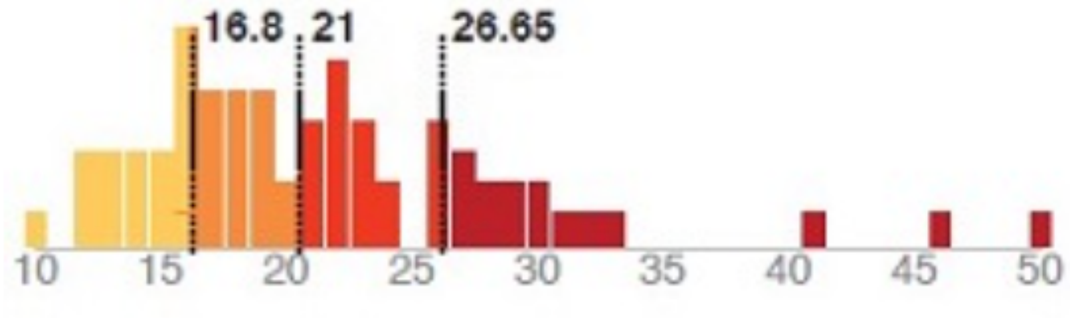


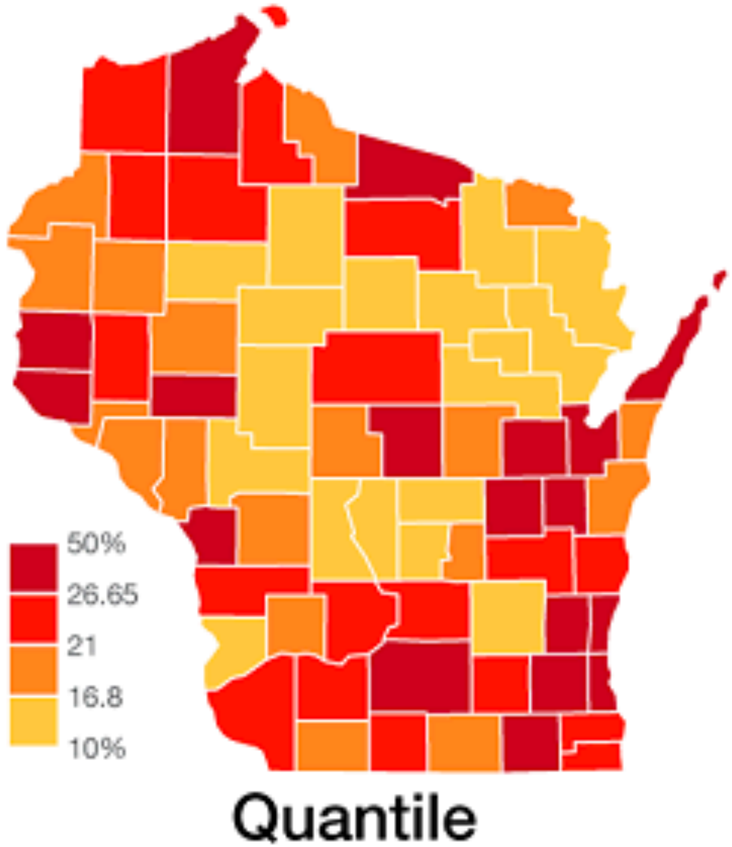


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Quantiles (Equal Count)

For evenly distributed data and ordinal data

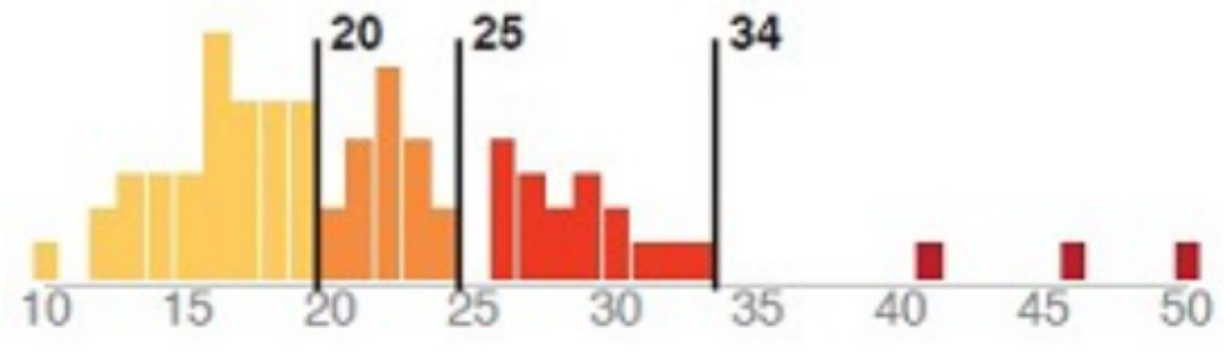


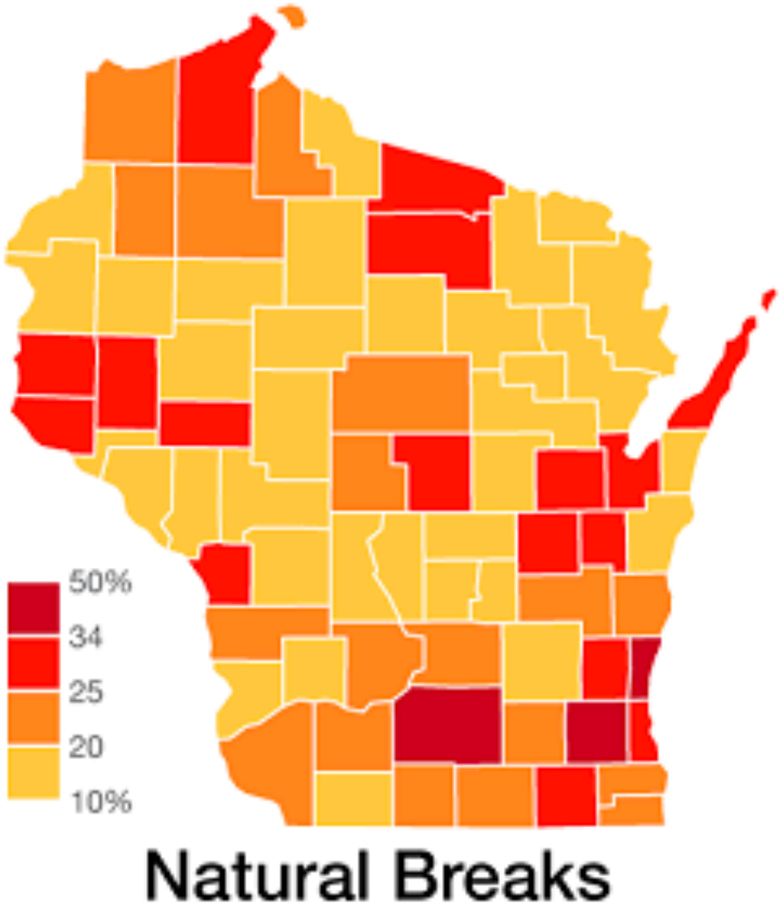


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Natural Breaks

For clustered data

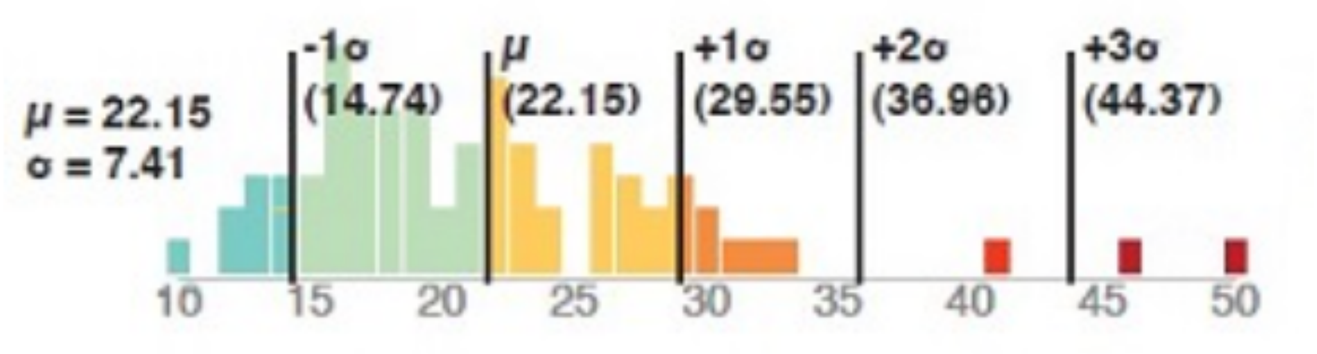


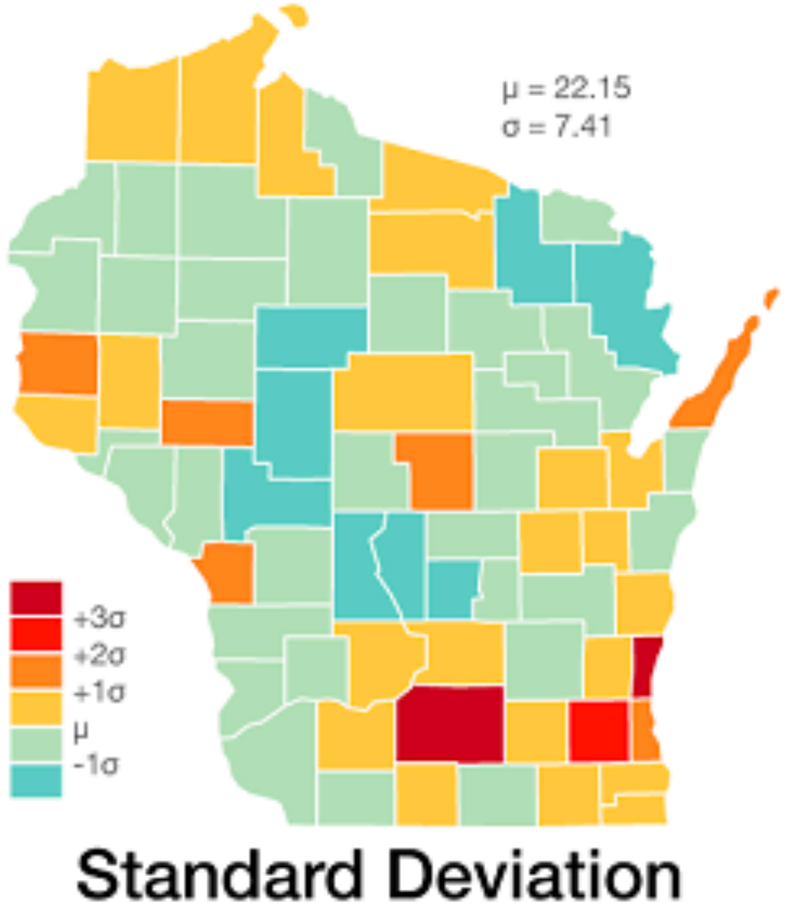


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Mean-Standard Deviation

For normally distributed data

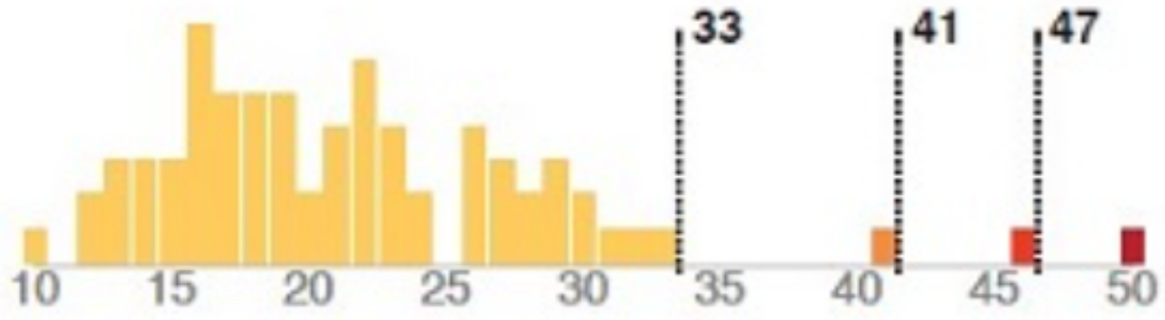


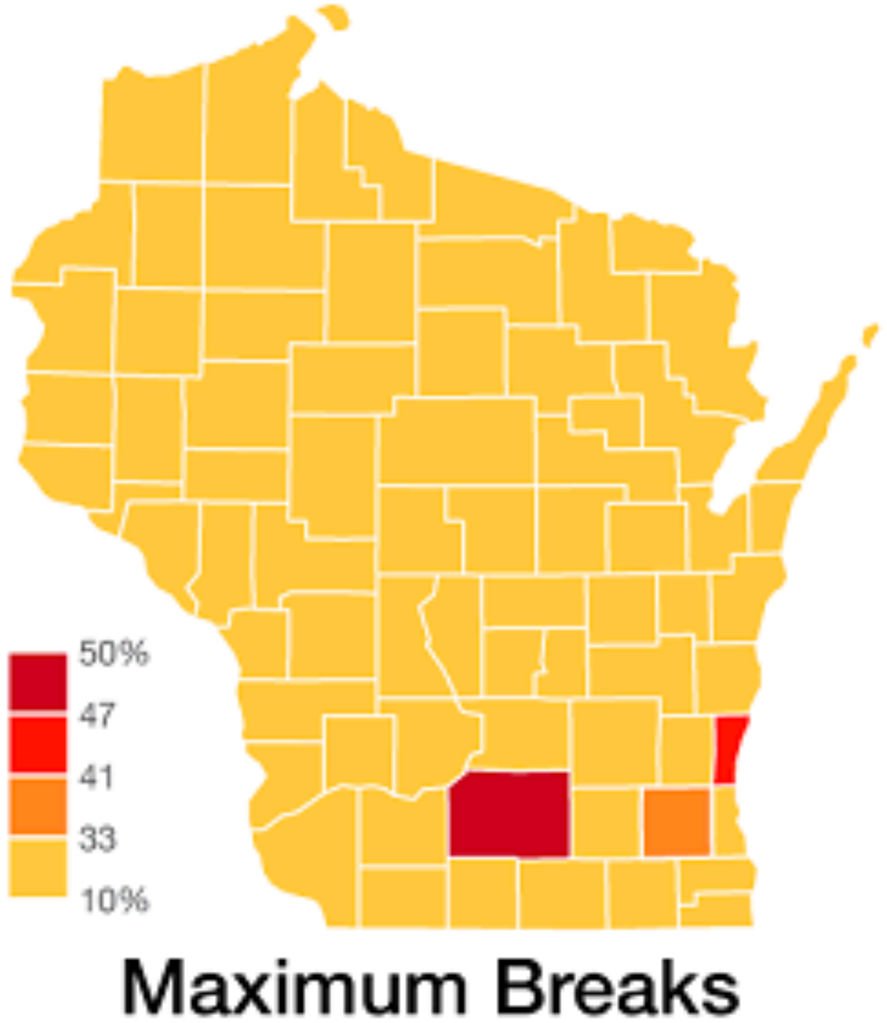


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Maximum Breaks

For piecewise and clustered data

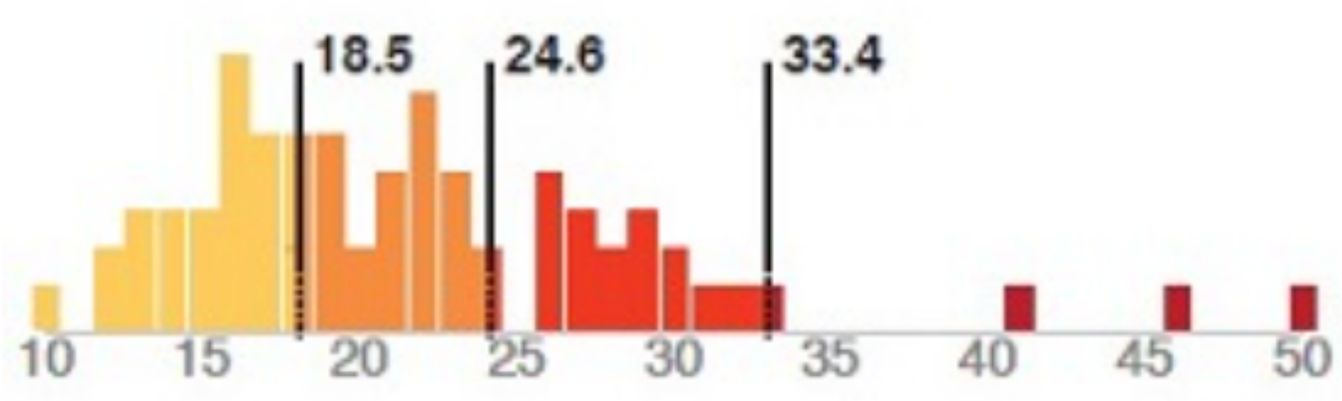


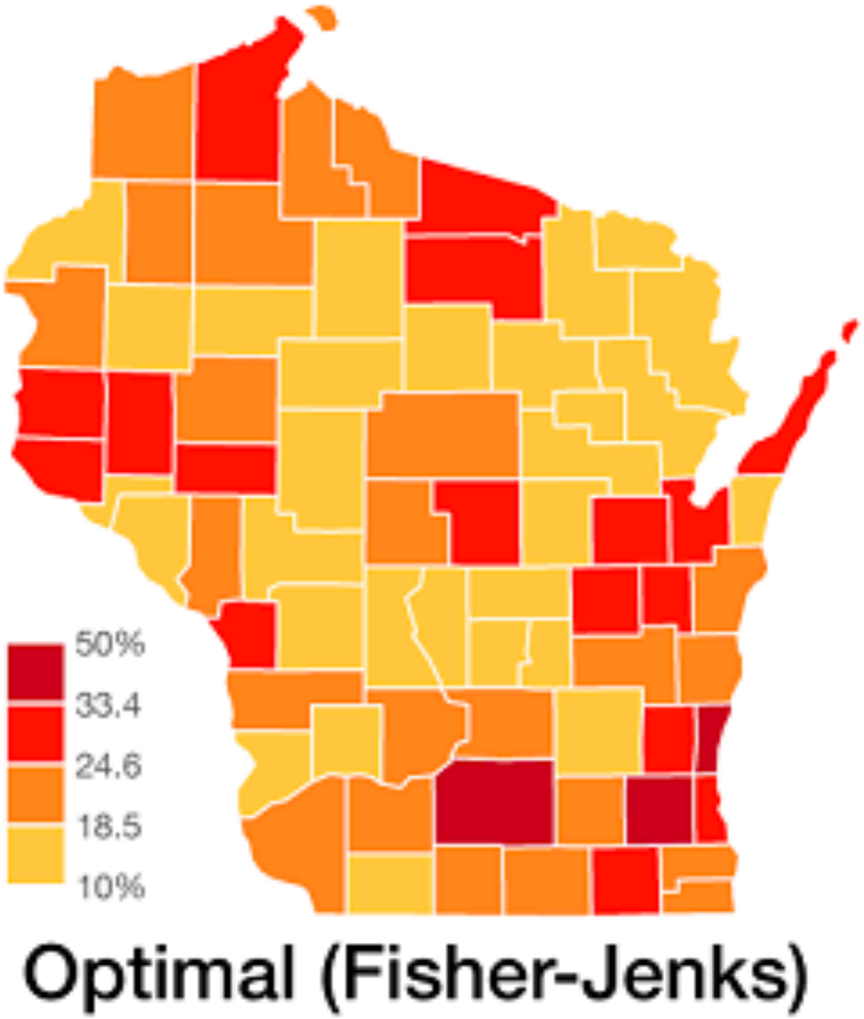


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Jenks-Caspall & Fisher-Jenks

For clustered and skewed data

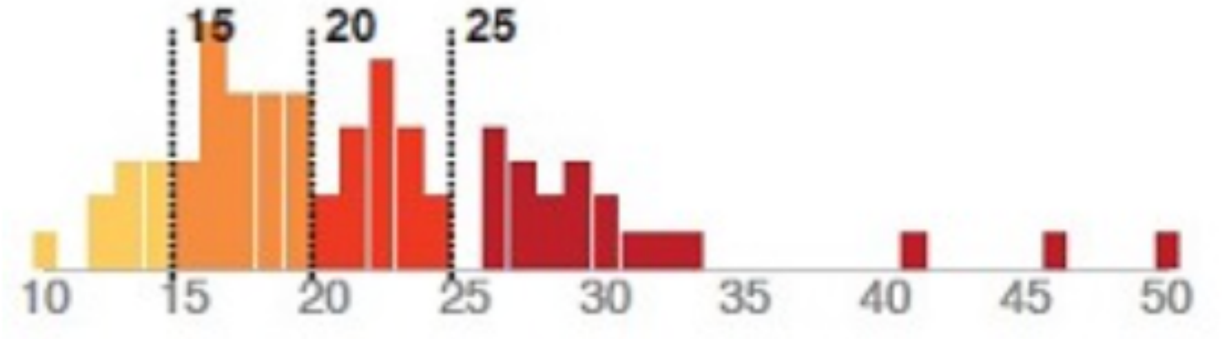


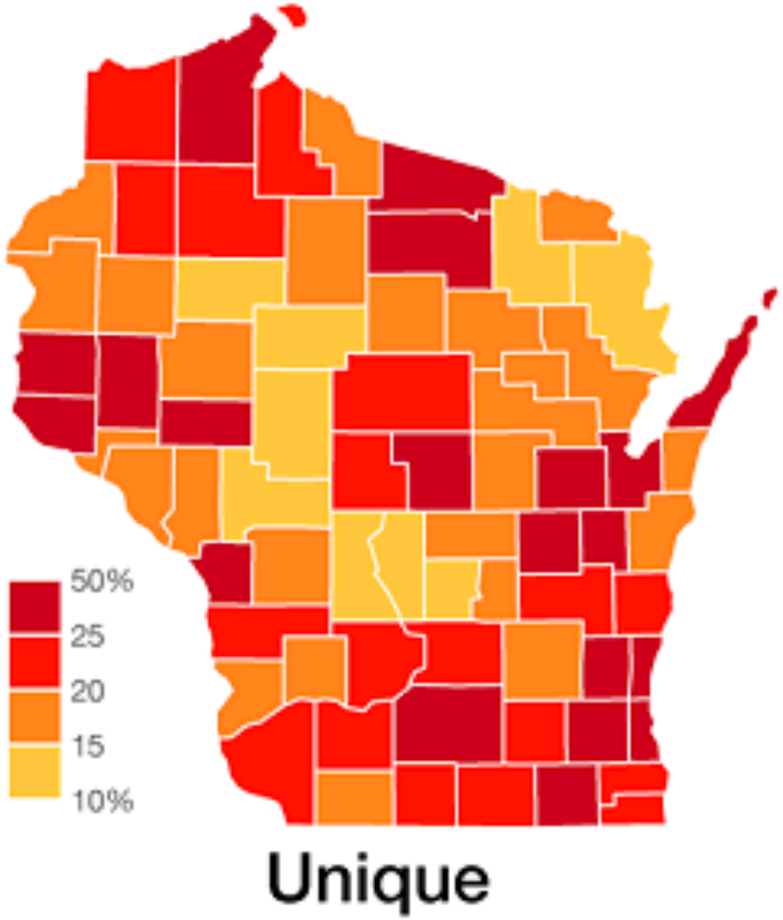


https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

Unique/Manual

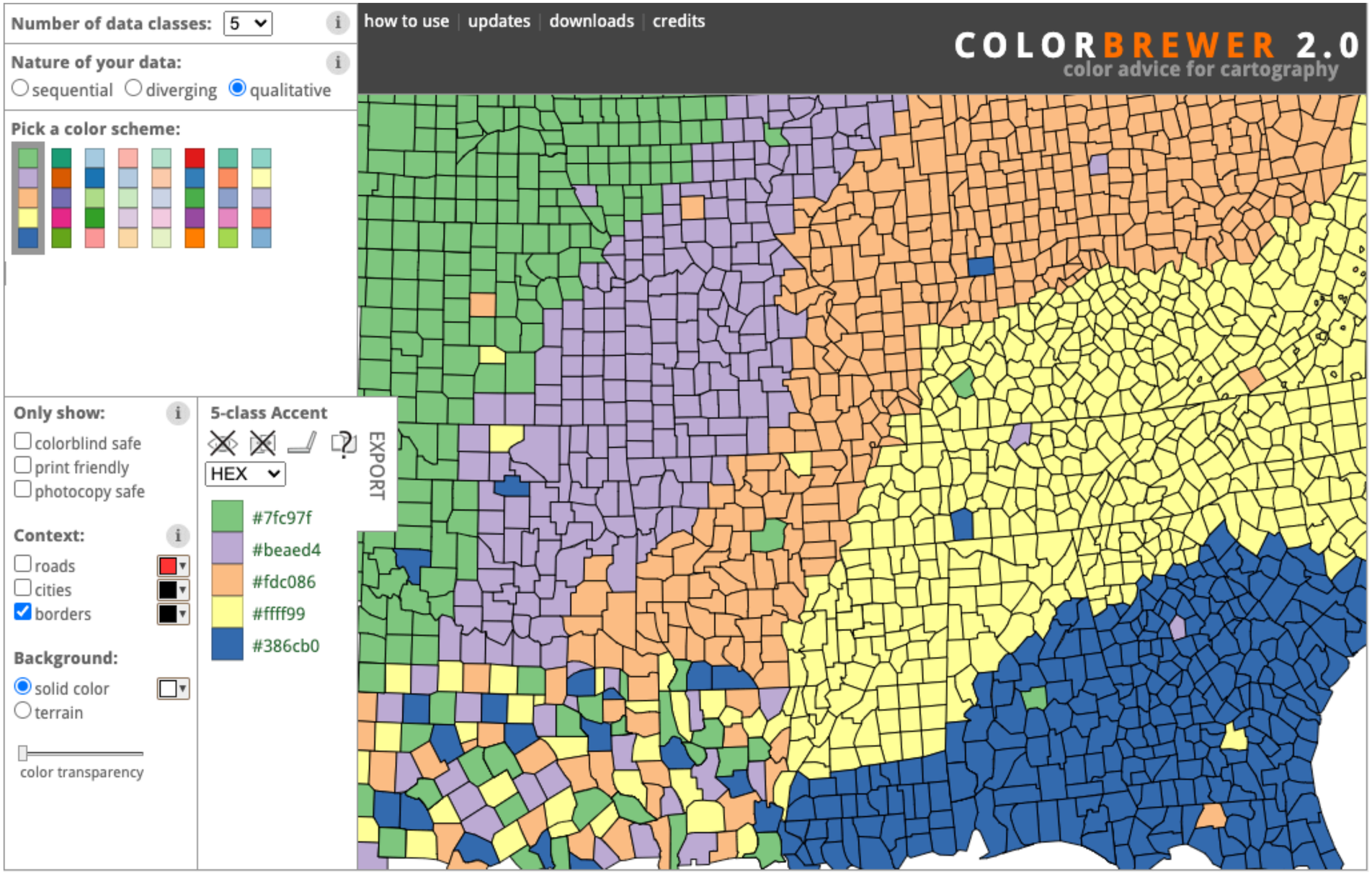
For data where key numbers are important





https://gistbok.ucgis.org/bok-topics/statistical-mapping-enumeration-normalization-classification

For choice of colors use tools like color brewer

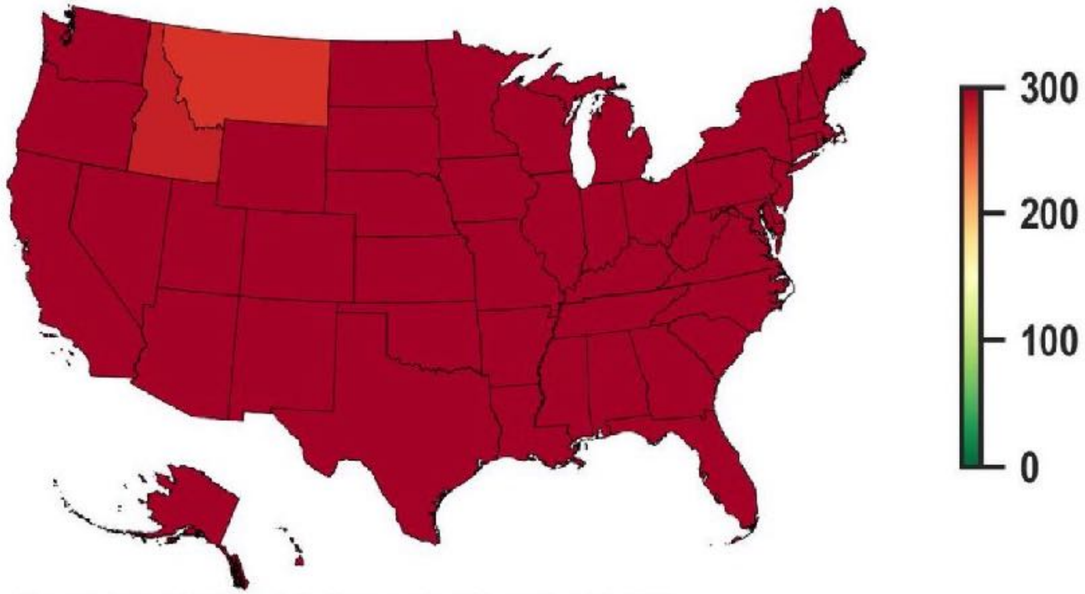


https://colorbrewer2.org/

Example: Wrong use of colors and scales

How the year started

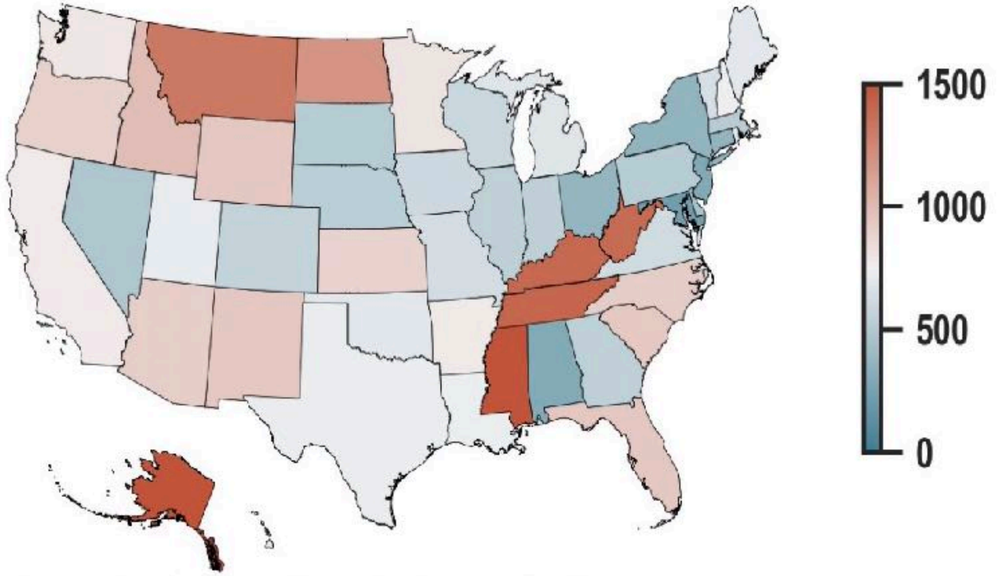
New COVID-19 cases per million population



Source: Johns Hopkins, J. P. Morgan. As of December 31, 2021

How it is going

New COVID-19 cases per million population



Source: Johns Hopkins, J. P. Morgan. As of February 08, 2022

https://twitter.com/carlquintanilla/status/1491824980648247296?s=20

Choropleth maps with Geopandas

```
gdf.plot(
    column=“col_name”,
    categorical=True,
    legend=True,
    cmap='tab20b'
);
```

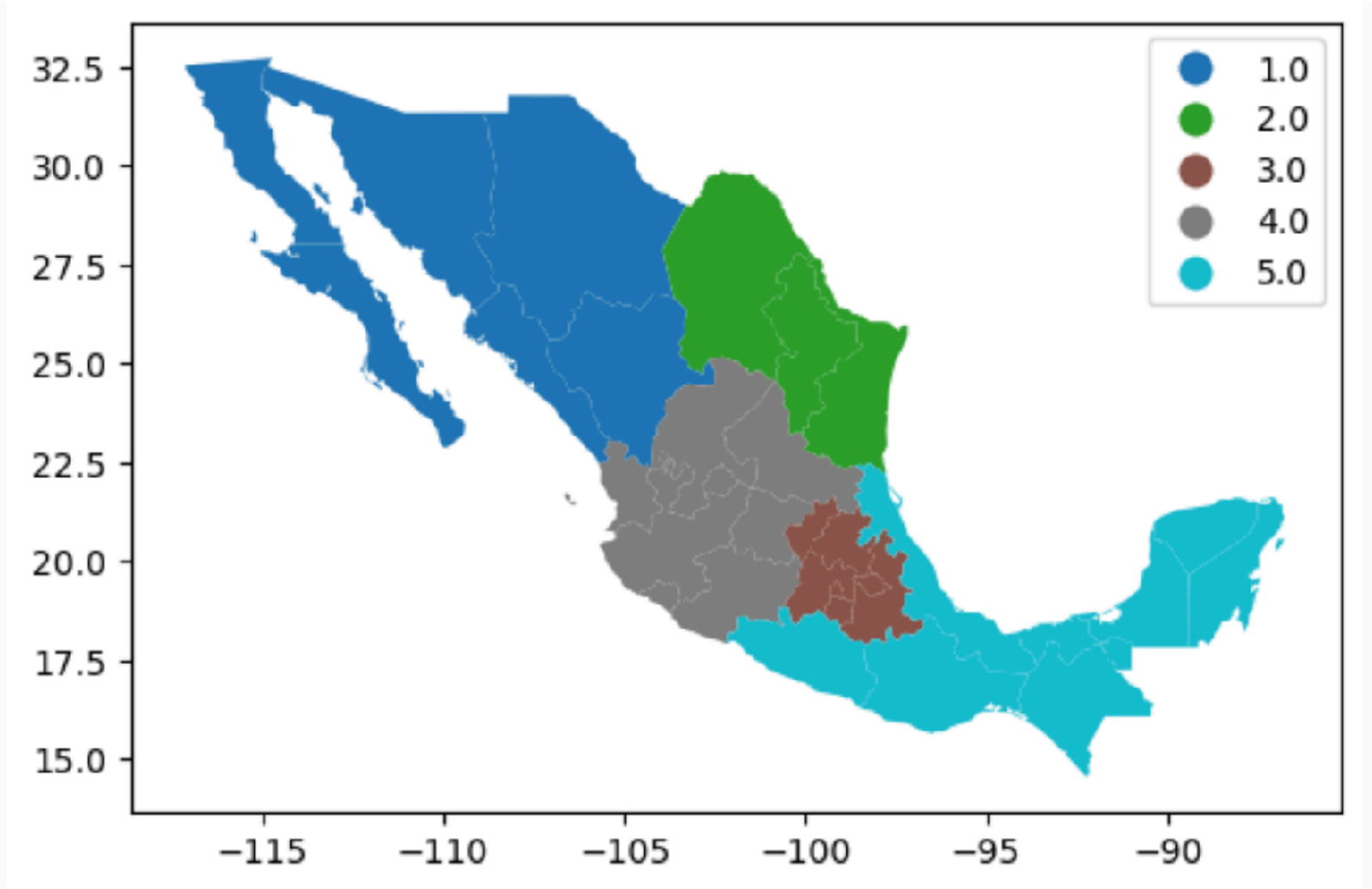
column with variable to plot

specify categorical if non-sequential data

show legend

choose color map (or use default)

Choropleth map with categorical data



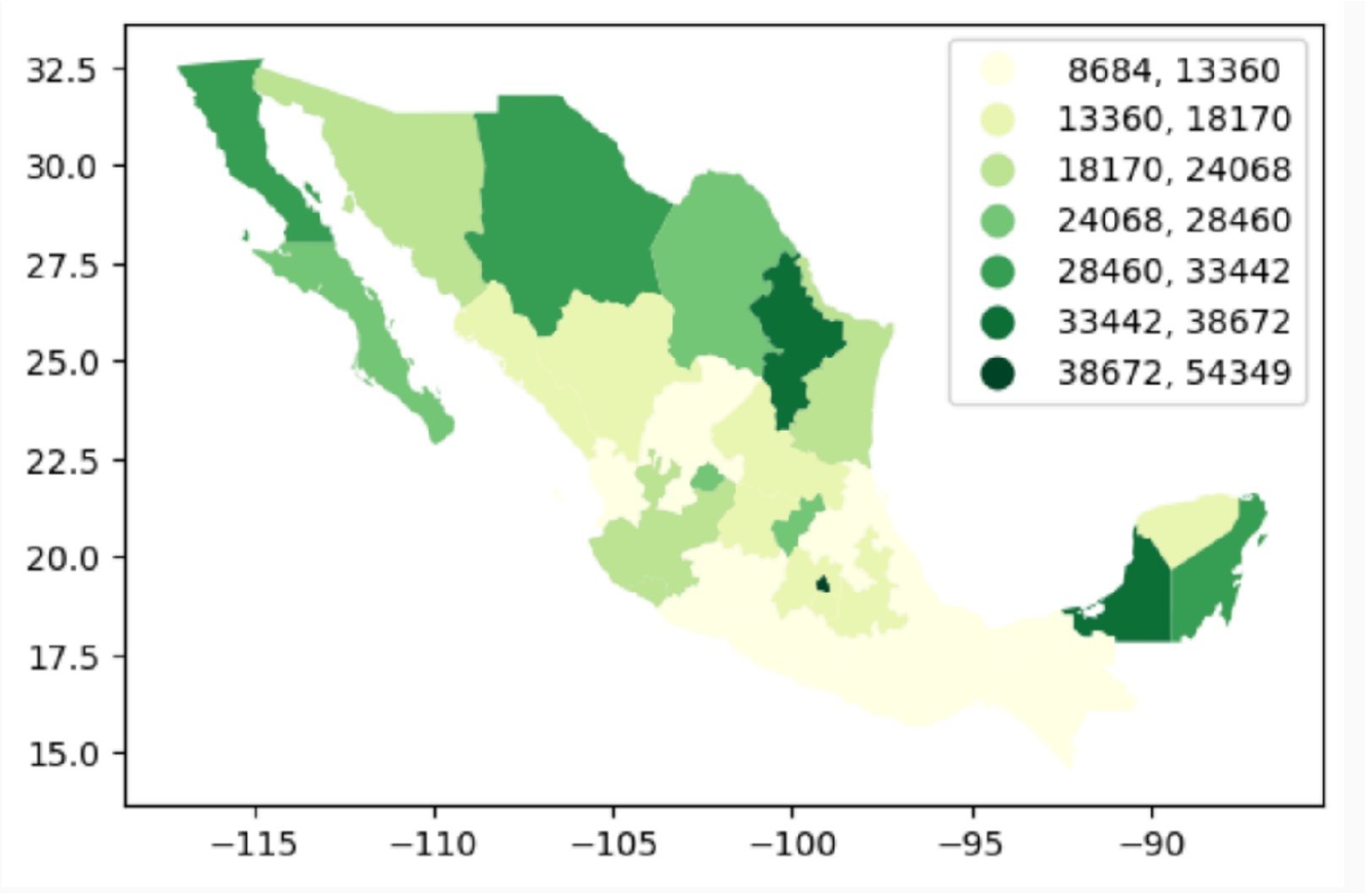
Geopandas + mapclassify

```
gdf.plot(
    column=“col_name",
    scheme="fisher_jenks",
    k=7,
    cmap="YlGn",
    legend=True,
);
```

choose classification method

specify number of classes

Choropleth map with sequential (ratio/interval)



PySal: Python Spatial Analysis Library

Core library for geospatial analysis

https://pysal.org/

mapclassify

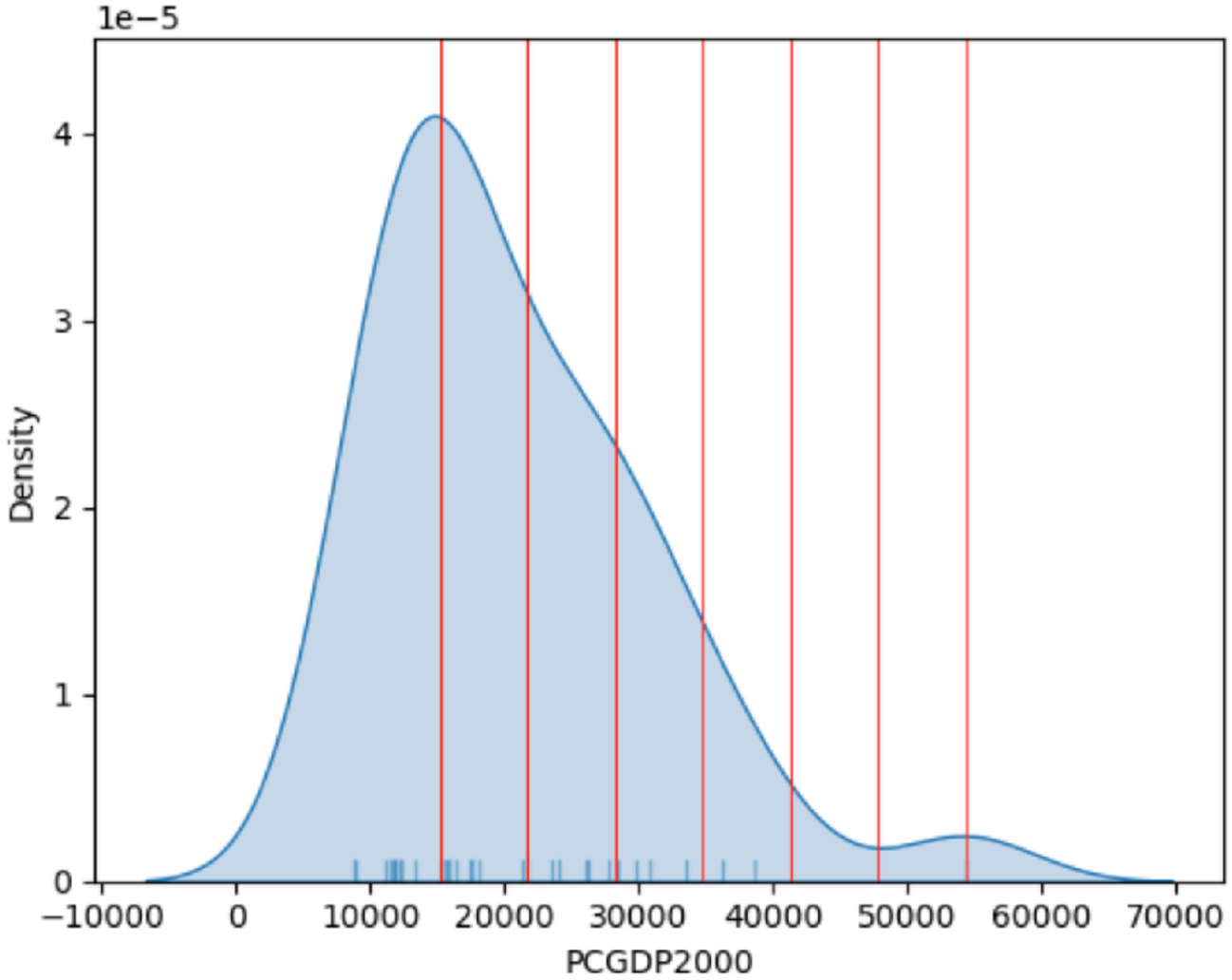
```
from pysal.viz import mapclassify

eq = mapclassify.EqualInterval(gdf[“col_name”], k=7)
```

```
      Interval          Count
----------------------------
[ 8684.00, 15207.57] |    10
(15207.57, 21731.14] |    10
(21731.14, 28254.71] |     5
(28254.71, 34778.29] |     4
(34778.29, 41301.86] |     2
(41301.86, 47825.43] |     0
(47825.43, 54349.00] |     1
```

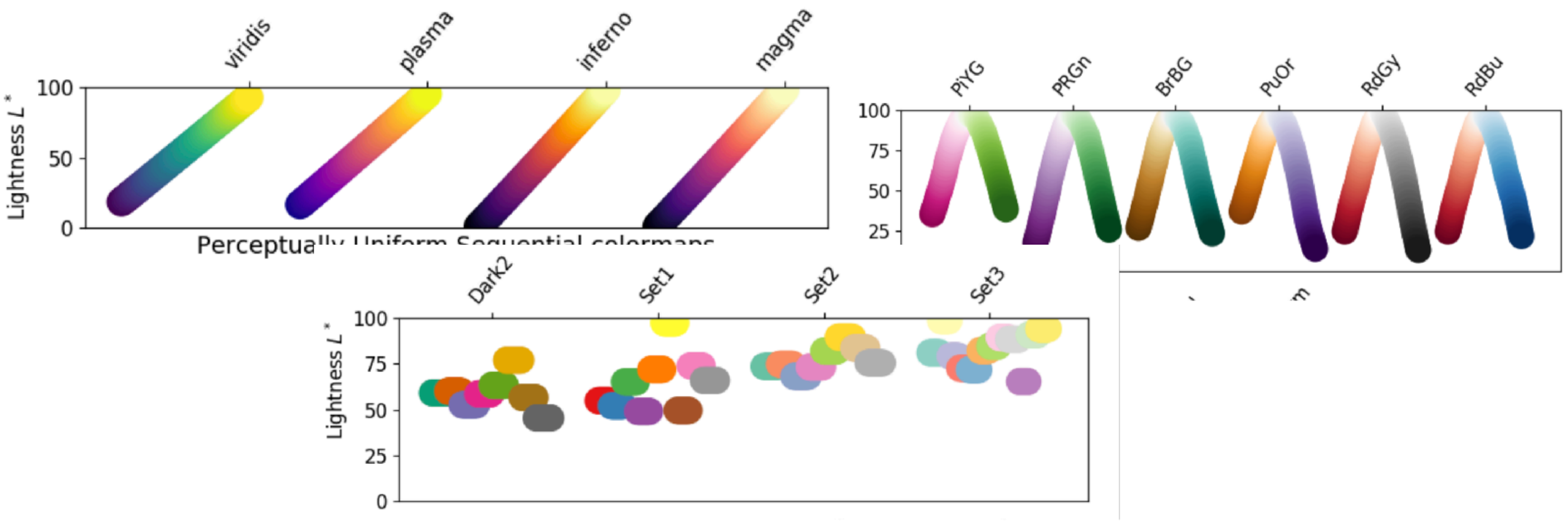
seaborn: statistical data visualization

Use KDE plots to explore how your data are distributed in the classes



https://seaborn.pydata.org/

Colormaps




```
import matplotlib.pyplot as plt

plt.colormaps()
```

https://matplotlib.org/2.0.2/users/colormaps.html

Jupyter

Sources and further materials for today's class

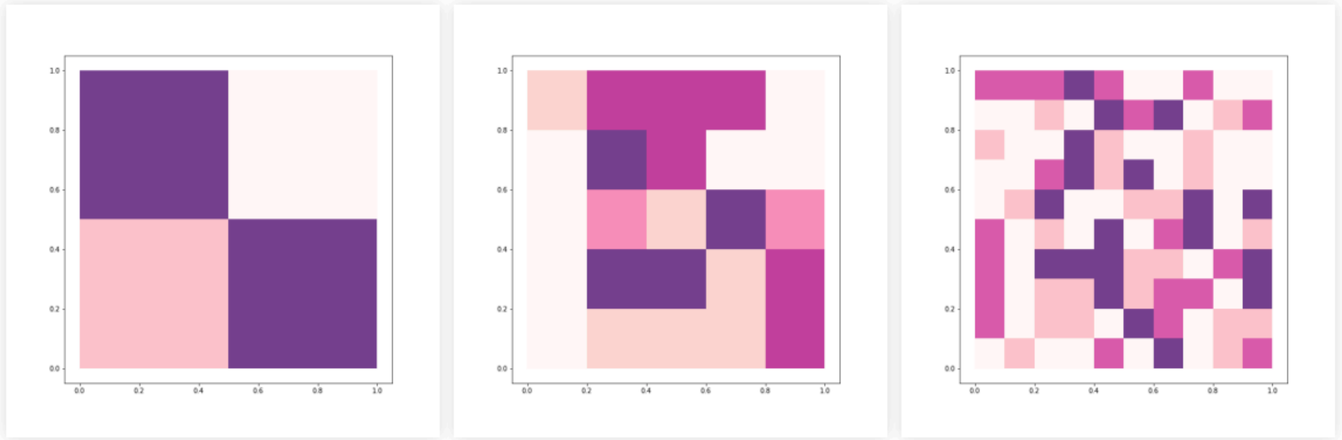
Geographic Data Science with Python

https://geographicdata.science/book/notebooks/05_choropleth.html

https://darribas.org/gds_course/content/bD/concepts_D.html

Take home messages for today

The choice of scale (MAUP) and classification scheme allows to create *arbitrary* interpretations



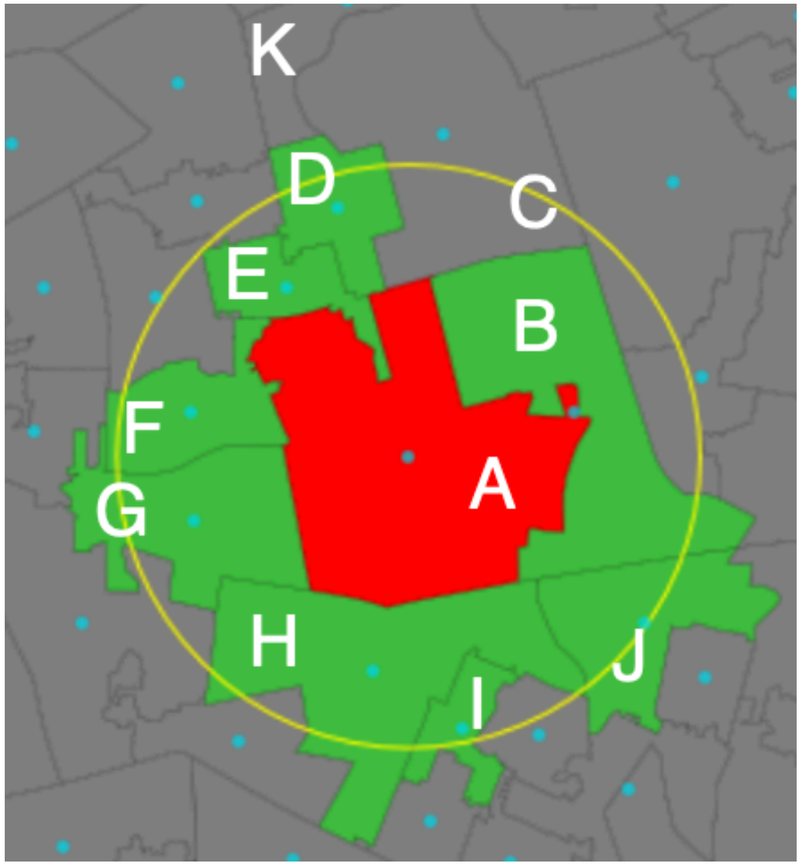
Aggregated data is simpler,
but can be very biased…

…but abstraction & generalizations are necessary

…In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. **In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless,** and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography

Jorge Luis Borges, Collected Fictions, 1946

Next week: Spatial weights



$$W = \begin{pmatrix} 0 & w_{12} & \dots & w_{1N} \\ w_{21} & \ddots & w_{ij} & \vdots \\ \vdots & w_{ji} & 0 & \vdots \\ w_{N1} & \dots & \dots & 0 \end{pmatrix}$$