Leveraging Geospatial Data in R

Thematic Maps

LIT Learning Event– May 13, 2021
Centre for Special Business Projects, Statistics Canada

Julia Conzon
Geographic Information Scientist
julia.conzon@canada.ca

Outline

- What is GIS
  - Vector vs Raster
  - Coordinate Reference Systems
- Geospatial R packages
- Cartographic Theory
  - Thematic Maps
  - Choropleth Maps
- R Scripts & Tutorial
  - tmap
  - sf + ggplot2

R Mapping Code + Tutorials

tmap

- Code: **https://github.com/Noznoc/r-gis-workshop**
- Tutorial: **https://noznoc.github.io/r-gis-workshop/index.html**

ggplot2 + sf

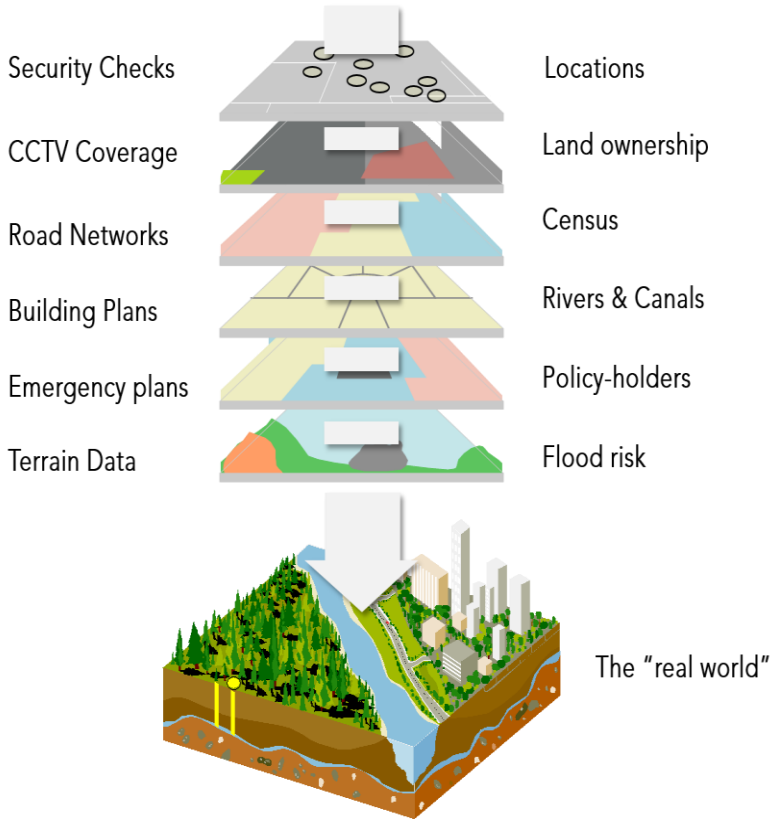
- Code: **https://github.com/Noznoc/bivariate-maps-ggplot2-sf/tree/statcan**
- Tutorial: **https://noznoc.github.io/bivariate-maps-ggplot2-sf/index.html**

What is GISystems?

In general, **Geographic Information Systems leverages geospatial data, statistical/analytical methods and technologies** to interpret the "real world"

These technologies (e.g., software or programming libraries) are used to **collect, store, retrieve, manipulate, explore, analyze and display** geospatial data

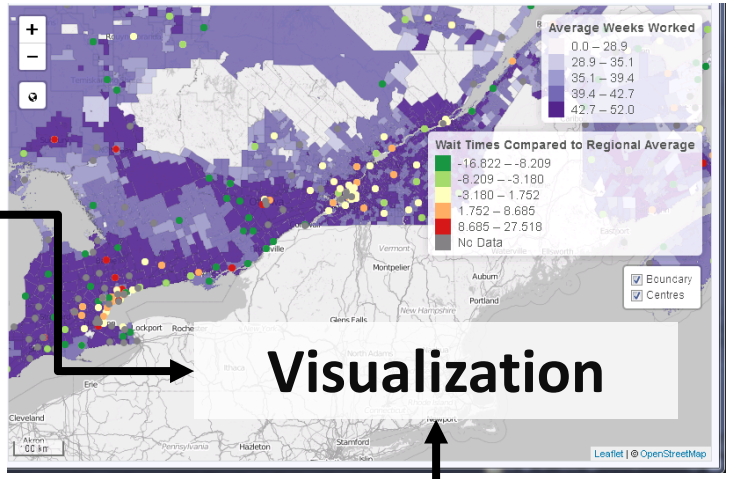
Geospatial technologies, techniques, and data standards through organizations like Open Geospatial Consortium



Source

What is GIScience?

Database ⟷ Server



Geospatial features, layers, ontologies

Processing + Analysis / Statistics

Targeted audience

Data collection (e.g., surveying, crowdsourcing, scraping)

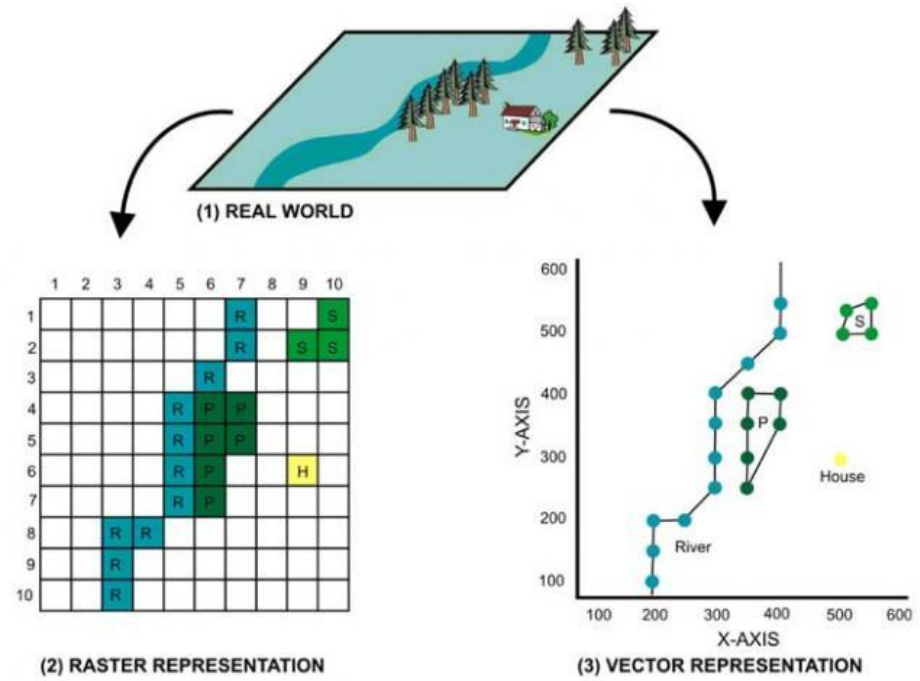
https://github.com/Noznoc/gis-101

Geospatial Features

Geospatial features are digitized geospatial representations

These features can be **discrete** (e.g., buildings and roads) or **continuous** (e.g., pollution, elevation)

Two geospatial data models: **vector** and **raster**

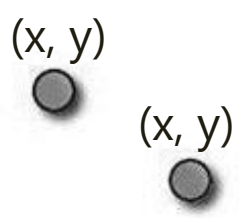


Source

Vector Model Components

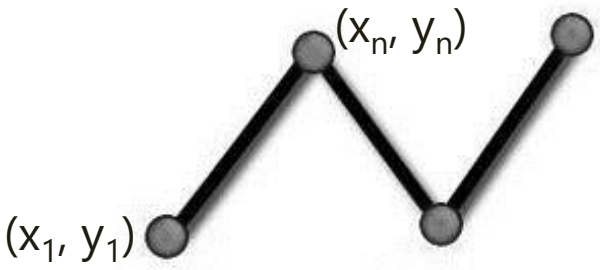
- Geometry: the **shape** of the vector object

Point



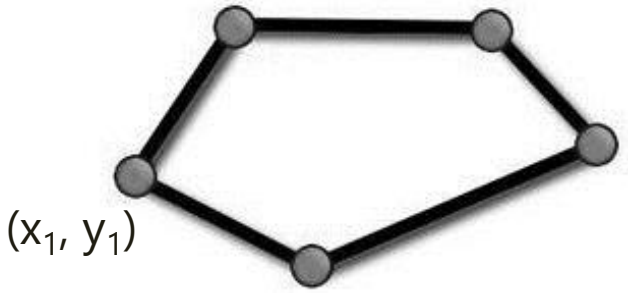
(x,y), (x,y)

Line



$(x_1, y_1, x_2, y_2, \ldots x_{n-1}, y_{n-1}, x_n, y_n)$

Polygon



$(x_1, y_1, x_2, y_2, x_3, y_3, \ldots x_n, y_n, x_1, y_1)$

- **Location**: coordinate system (longitude, latitude)

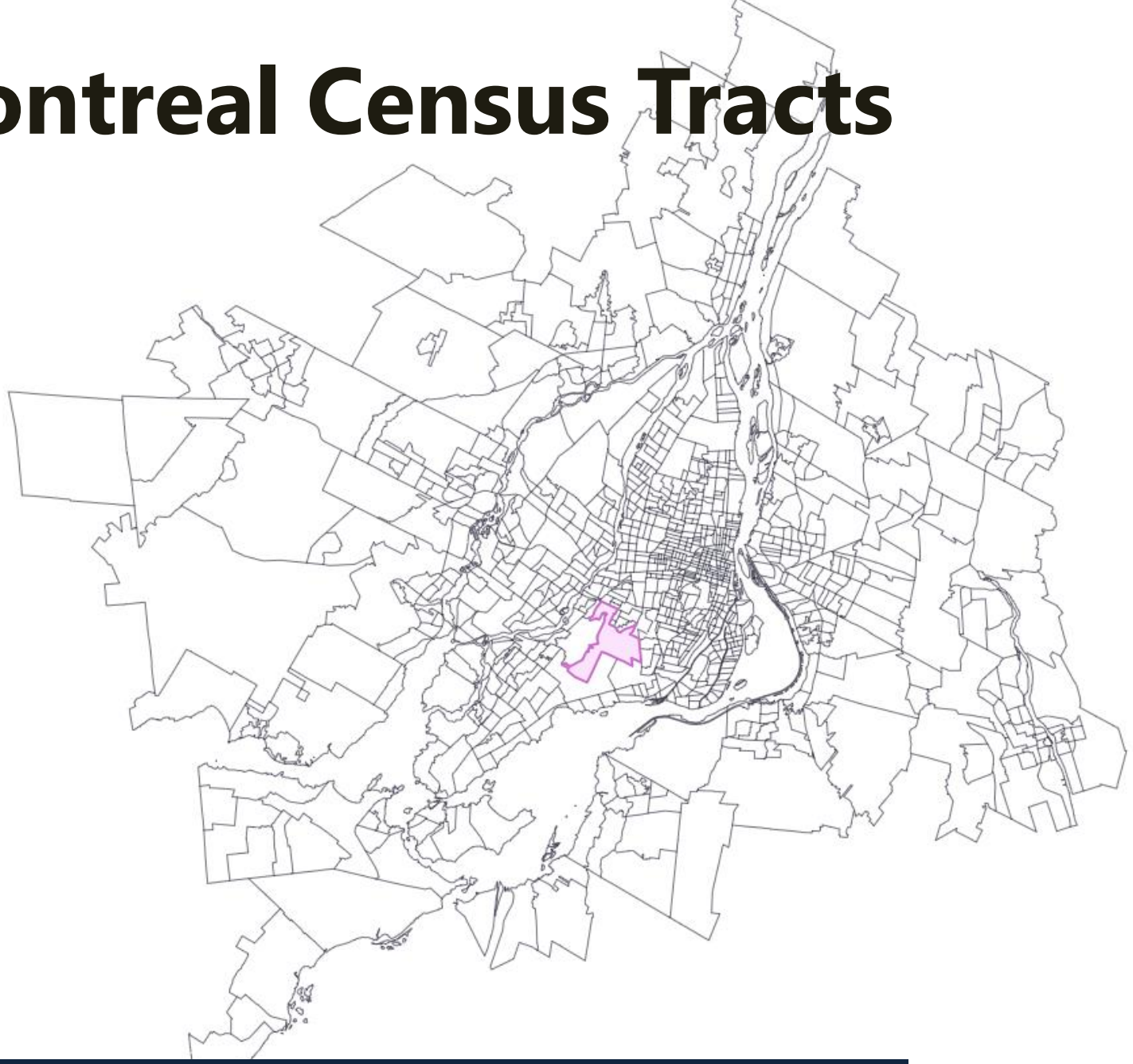
  (x,y) → (4018517.282186, 2008752.042743)

- **Attributes**: variables/fields associated with a feature
- **Style**: color, form, outline, size, pattern, lightness

Source

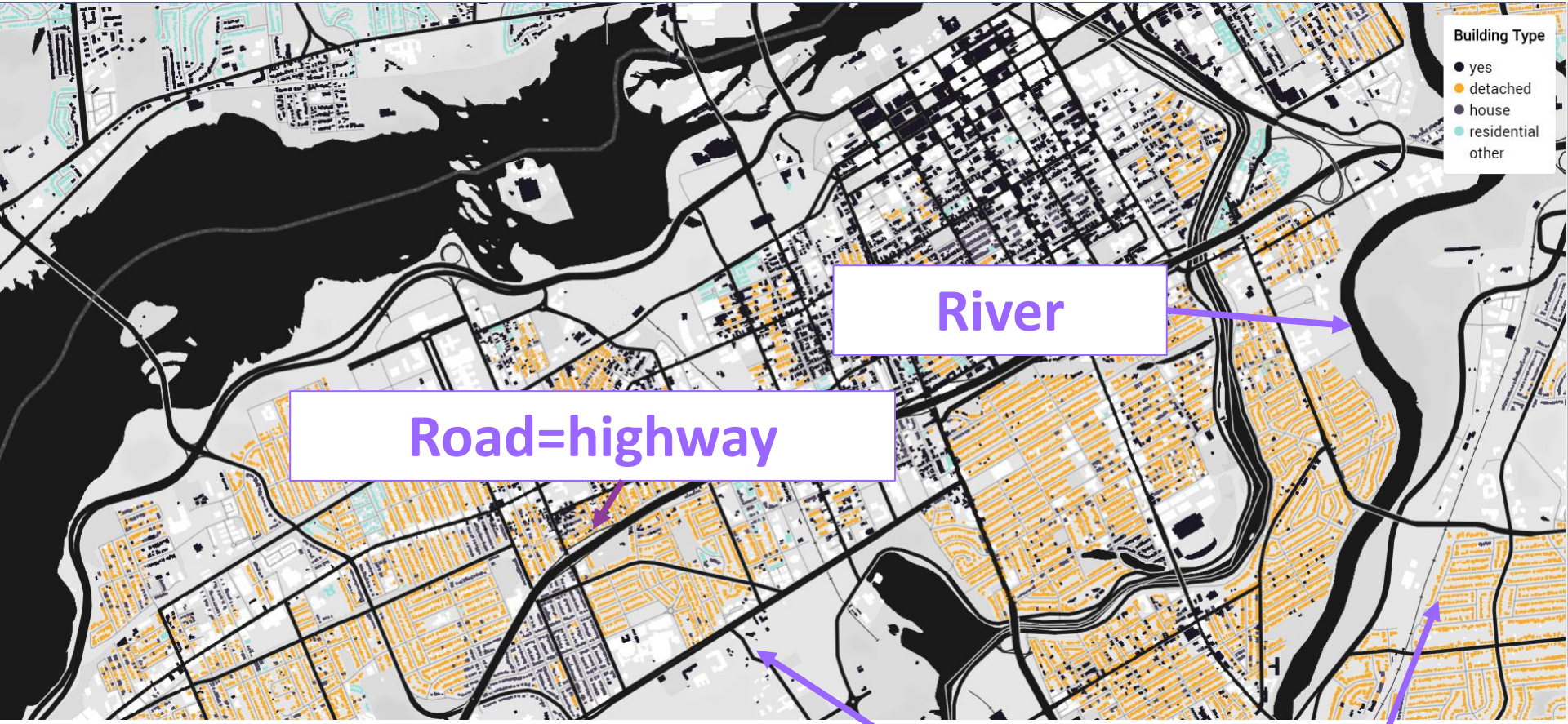
E.g., Montreal Census Tracts

CTUID	462041800
CTNAME	0418.00
PRUID	24
PRNAME	
CMAUID	462
CMAPUID	24462
CMANAME	
CMATYPE	B
Popultn	5907
MedinAg	37.7
LIM-AT	21.8
PpltnDn	4534.8



Geometry and location linked to attributes

E.g., Ottawa/Gatineau Basemap



Style linked to geometry, location, and attribute!

Source

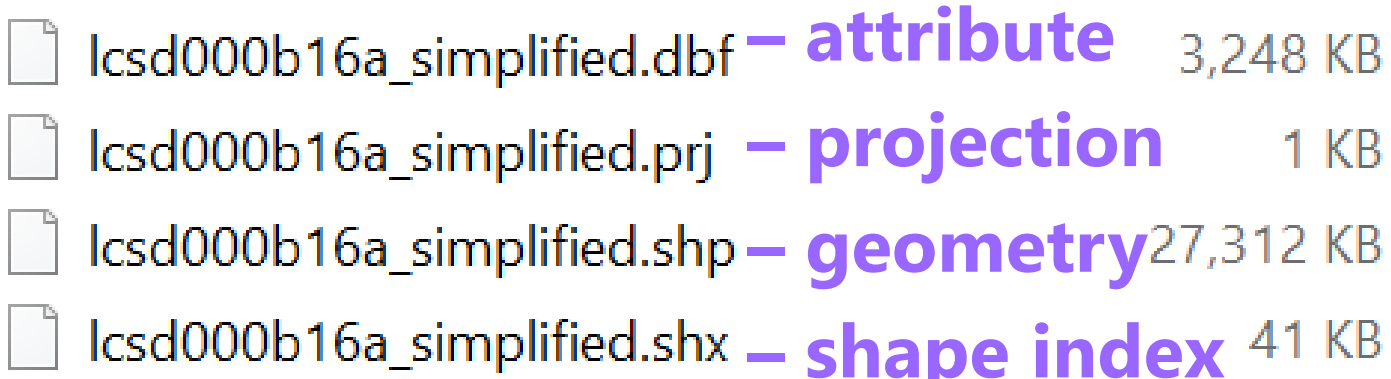
Vector Formats + Standards

Open Geospatial Consortium (OGC) Compliant

- Well-Known Text (WKT)
- GeoPackages
- Shapefiles
- Keyhole Markup Language (KML)
- GeoJSON

https://www.opengeospatial.org/docs/is

Shapefiles (Esri)



GeoJSON (open)

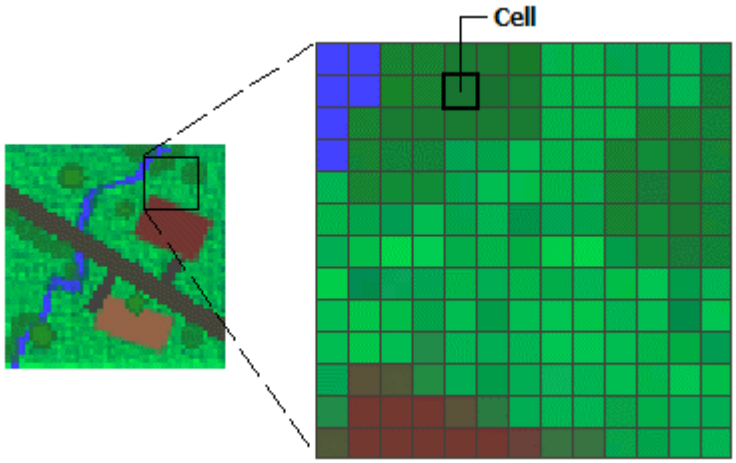
Represents simple geographic features (points, lines, strings, polygons) and non-spatial attributes

Popular for web maps

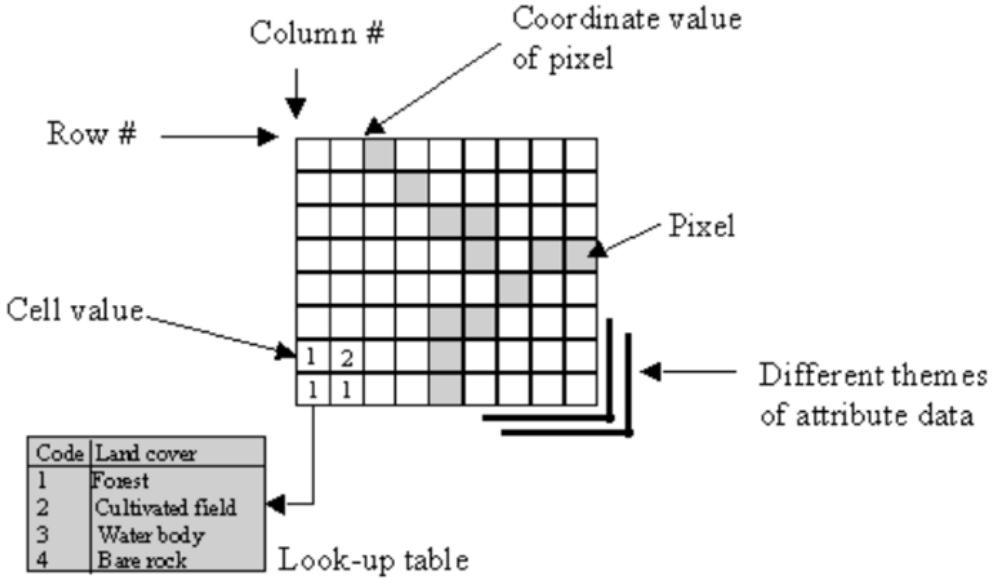

```
{
  "type": "FeatureCollection",
  "features": [
    {
      "type": "Feature",
      "properties": {},
      "geometry": {
        "type": "Polygon",
        "coordinates": [
          [
            [
              -123.10558319091798,
              49.264668080750134
            ],
            [
              -123.07296752929688,
              49.26534019822459
            ],
            [
              -123.08876037597656,
              49.2727328862137
            ],
            [
              -123.10558319091798,
              49.264668080750134
            ]
          ]
        ]
      }
    }
  ]
}
```

Raster Model Components

- Rasters: digital aerial photography, satellite imagery, scanned maps, digital images (e.g., street-level images)
- Matrix of equal sized cells/pixels represented as a grid of rows and columns
- Each cell/pixel has an attribute value



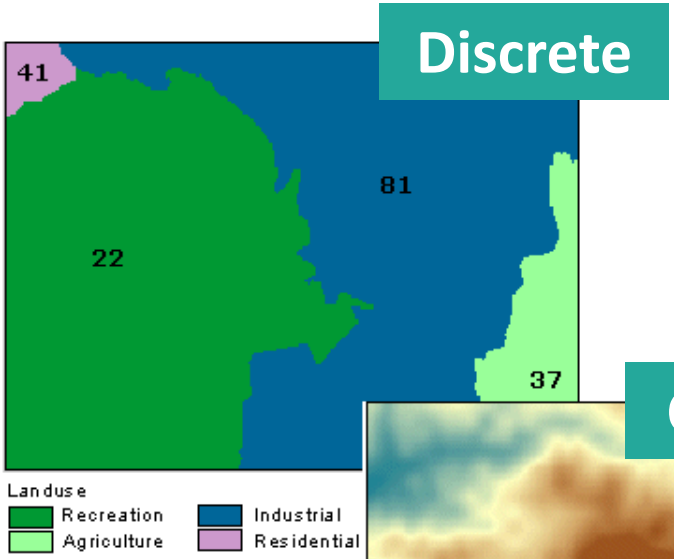
Source

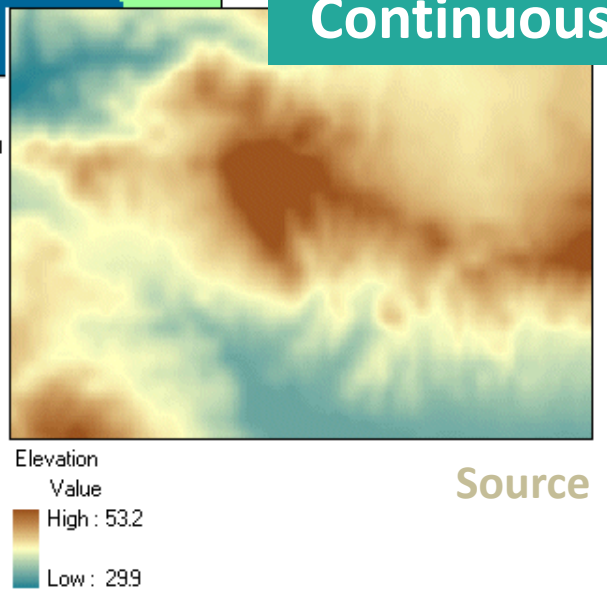


A.K. Yeung 1998-10-10u51-15

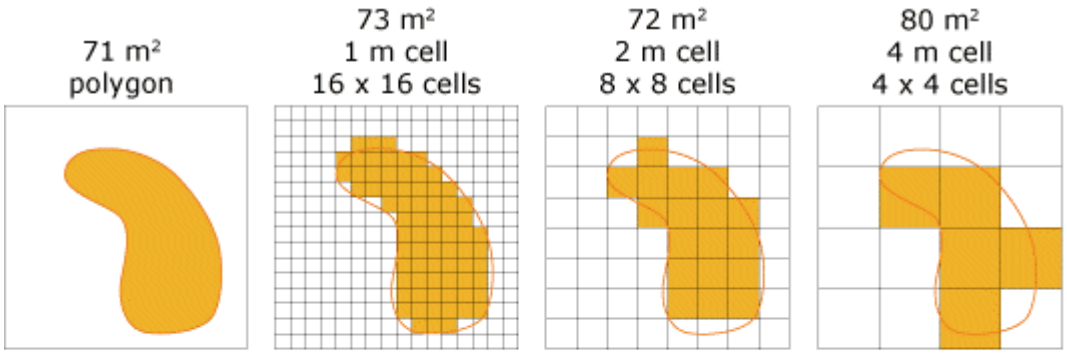
Source

Raster Data





Source



- Smaller cell size
- Higher resolution
- Higher feature spatial accuracy
- Slower display
- Slower processing
- Larger file size

- Larger cell size
- Lower resolution
- Lower feature spatial accuracy
- Faster display
- Faster processing
- Smaller file size

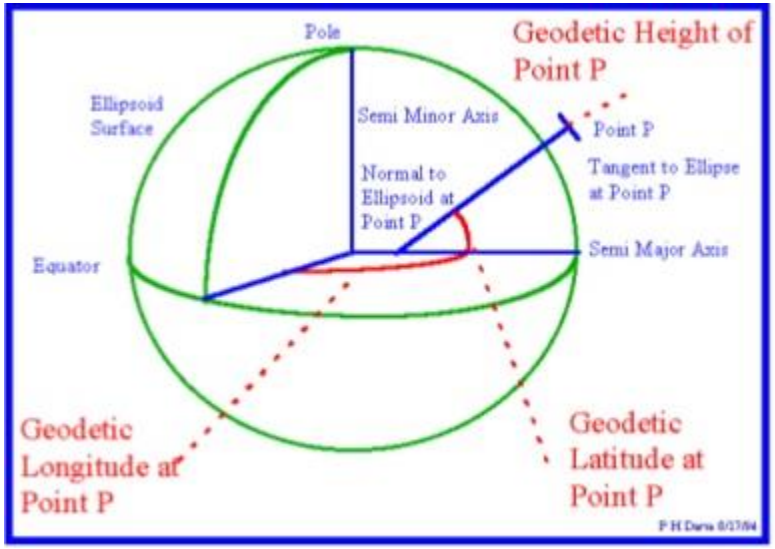
Source

Coordinate Systems

A coordinate reference system (**CRS**), is a reference system for representing the location of a vector feature or a raster. There are two types of CRS:

Geographic coordinate systems

- Takes earth as a spheroidal surface
- Defined by datum, an estimate of earth's surfaced based on an ellipsoid
- Longitude (x) and latitude (y) as reference lines
- Uses lat and long and angular measurement to define a position



Projected coordinate systems

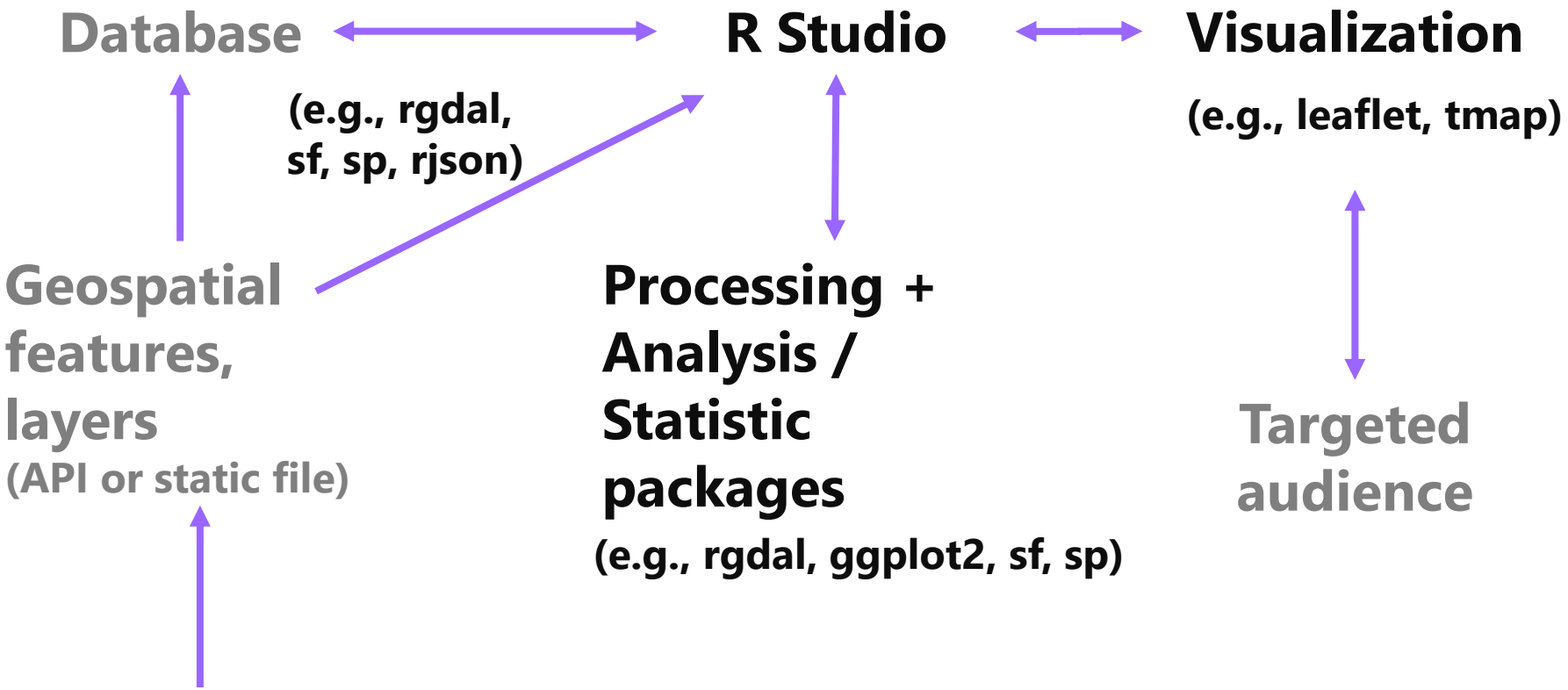
- Takes earth as a planar/flat surface

https://www.slideshare.net/yourmohsin/projections-and-coordinate-system

How can R Support GIS?

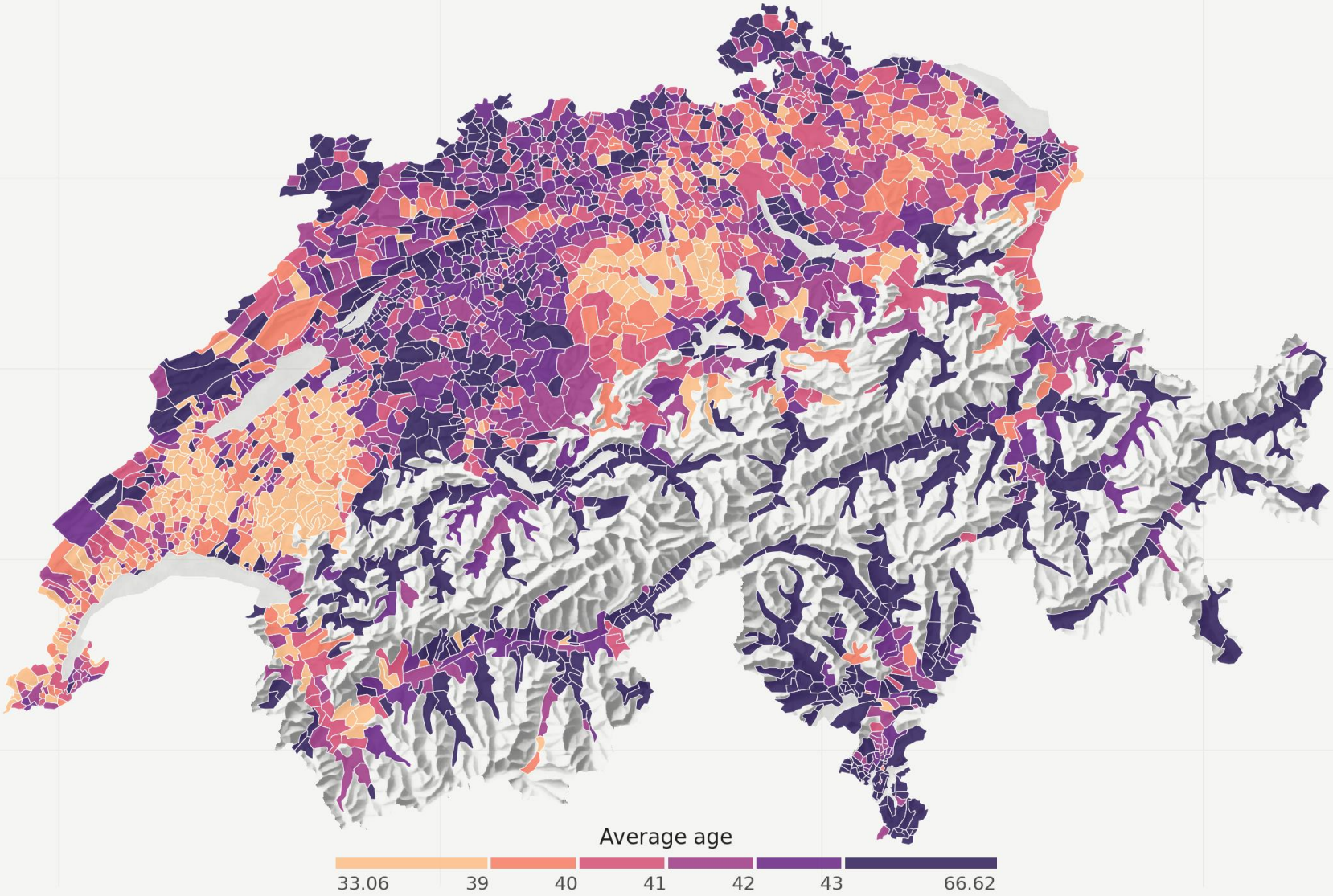
- Large library of spatial analytical and statistical packages for both **raster** and **vector data**
- Several packages to support both **static** and **interactive maps**
- R package bindings of popular open source libraries, like Leaflet and GDAL
- Reads and writes various spatial data formats: GeoJSON, shapefiles, GeoPackages, KML, spatial database connections (PostgreSQL)
- Allows non-spatial data to be assigned to spatial data

R Geospatial Architecture



Switzerland's regional demographics

Average age in Swiss municipalities, 2015

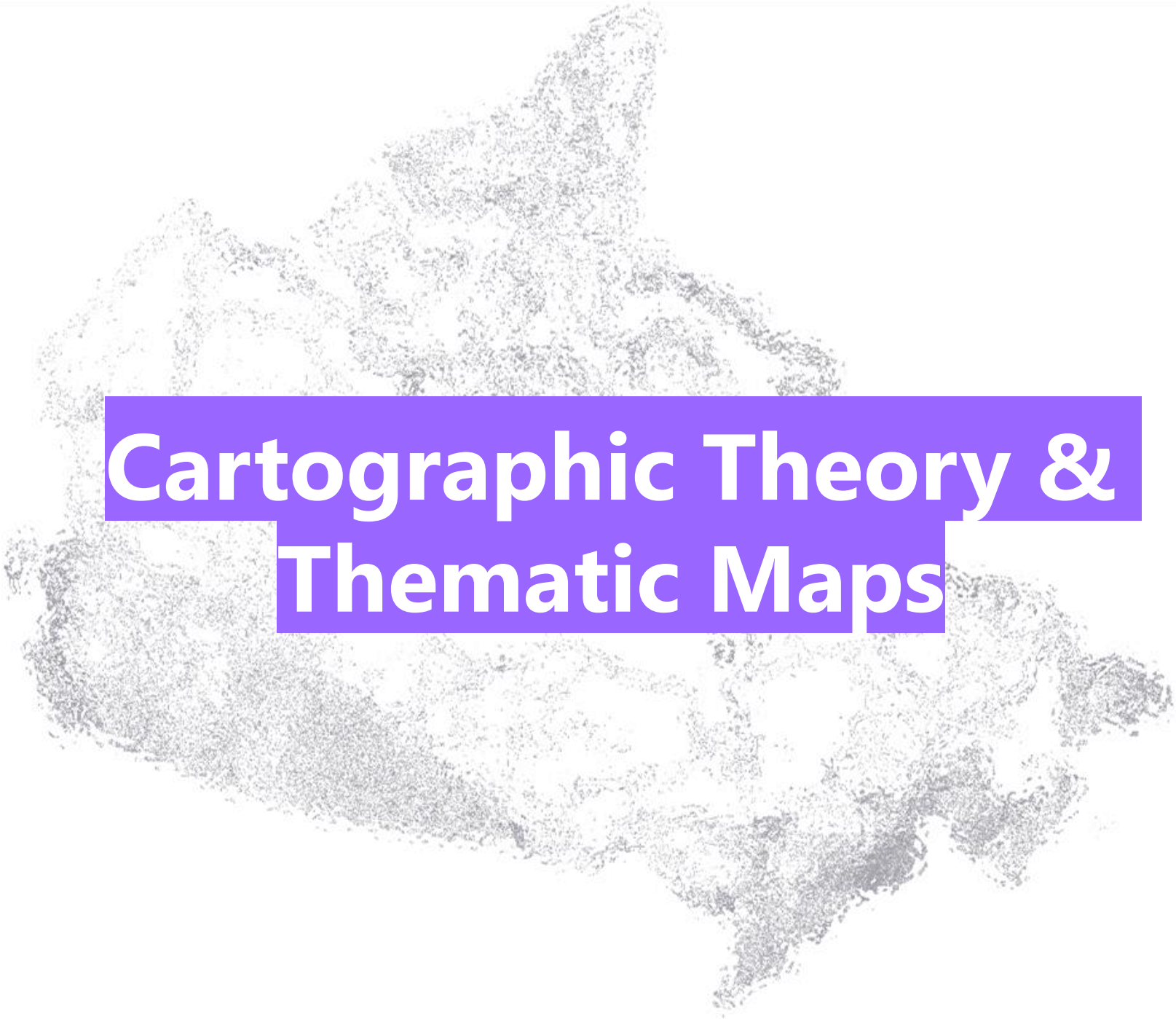


https://timogrossenbacher.ch/2016/12/beautiful-thematic-maps-with-ggplot2-only/

R Geospatial Packages

There's lots: https://cran.r-project.org/web/views/Spatial.html

- ~~**sp**: classes and methods for spatial data; stores data as slots~~
- **sf**: classes and methods for spatial data; stores data as simple features (well-known text geometry) within a data frame
- **rgdal**: geospatial data abstract library (GDAL) binding to read/write geospatial data (e.g., shapefiles, GeoPackages)
- **rgeos**: geometry engine (open source) binding to manipulate and query geometric data
- **ggplot2**: create static map plots
- **ggmap**: extension of ggplot2 for additional map features
- **maptools**: mapping functionality
- **tmap**: thematic static or interactive maps
- **leaflet**: JavaScript library binding for interactive mapping



Cartographic Theory & Thematic Maps

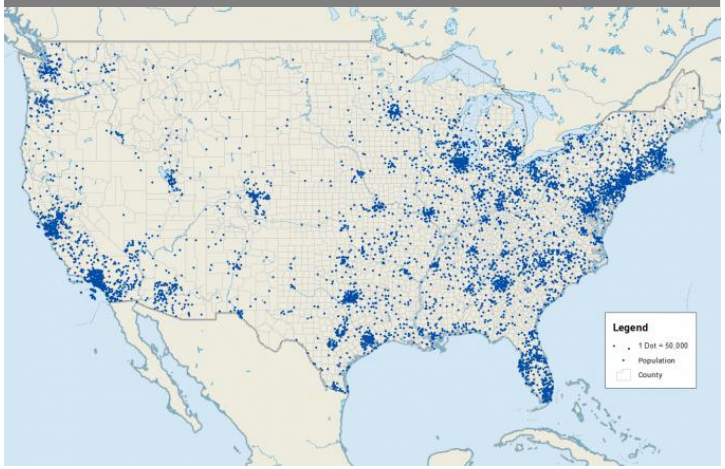
Thematic Maps

- Maps that show **spatial distributions** related to single or multiple themes
- Linked to a geographic **space** (any scale) to interpret the **place**
- Data can be qualitative or quantitative
- Useful in **exploratory spatial data analysis**
  - Discover **patterns** and **relationships** between variables
- Used to help form **hypotheses**

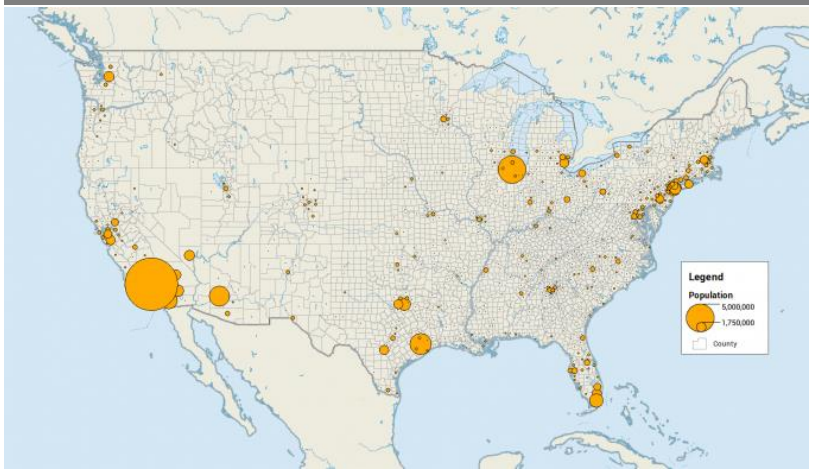
https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/other-autre/theme/def-eng.htm

Some Thematic Map Types

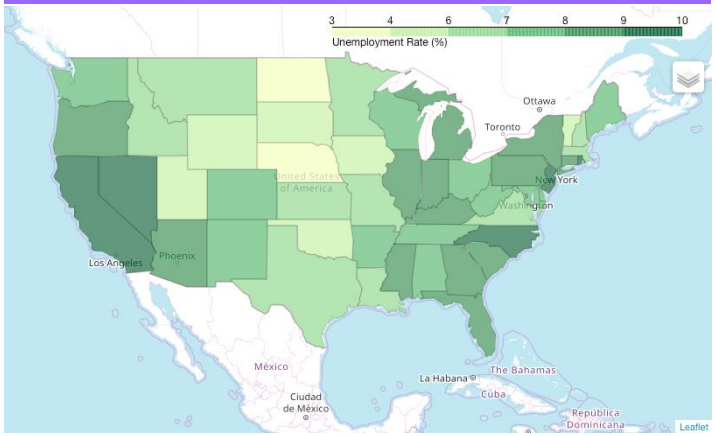
Dot Density



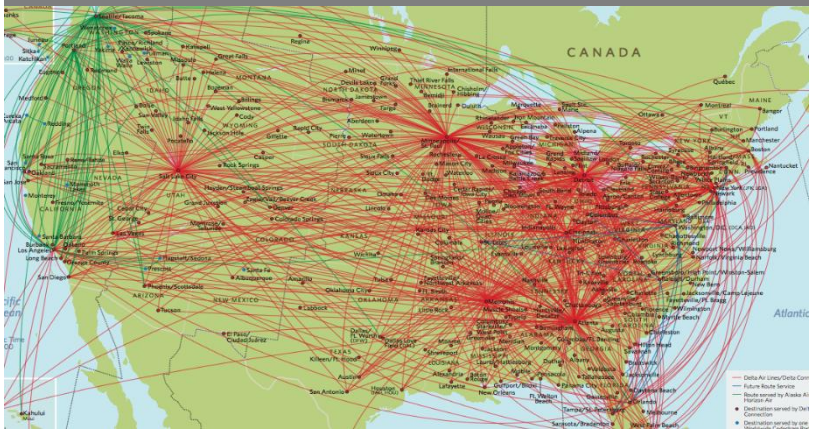
Proportional Symbol



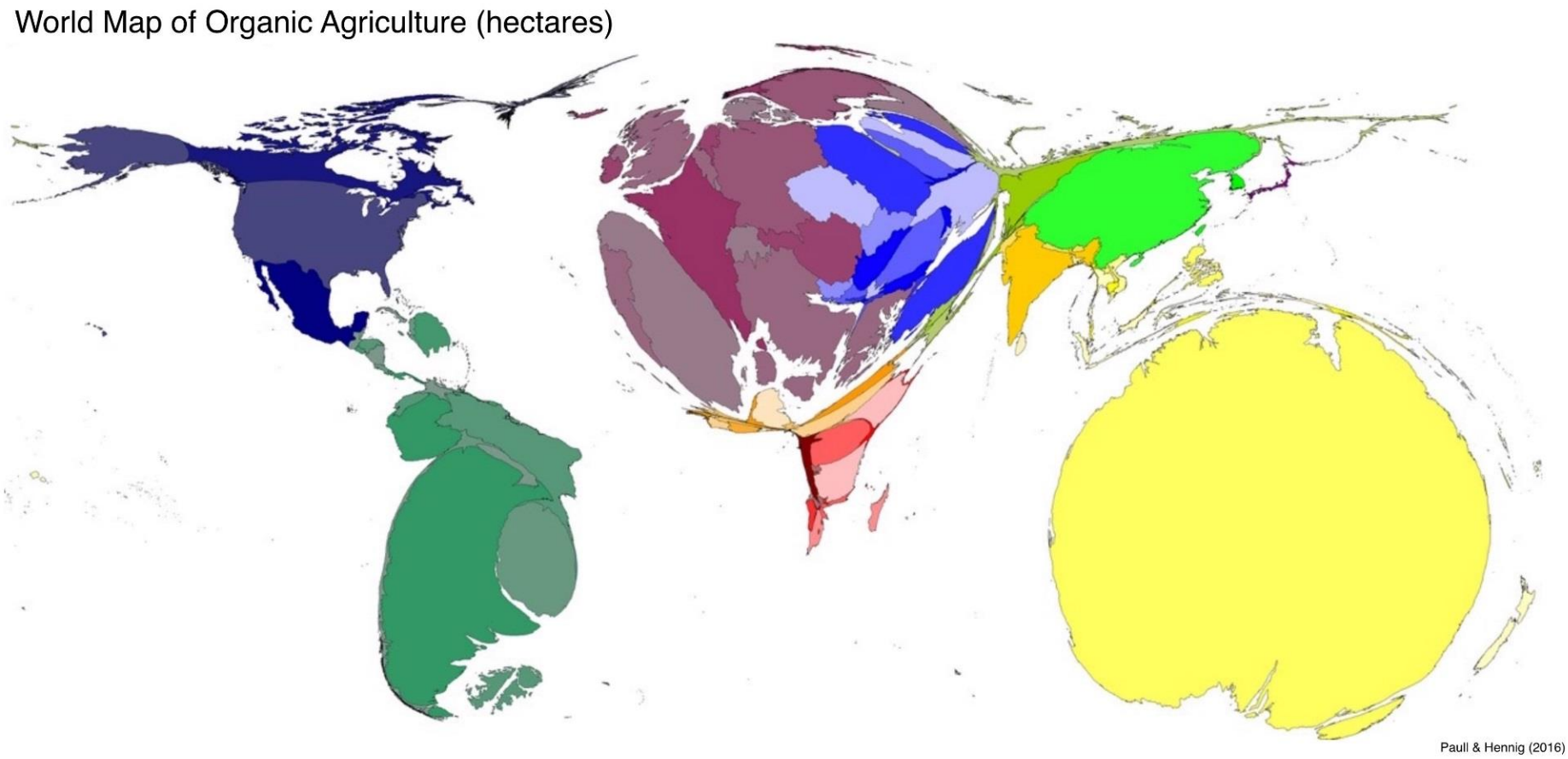
Choropleth



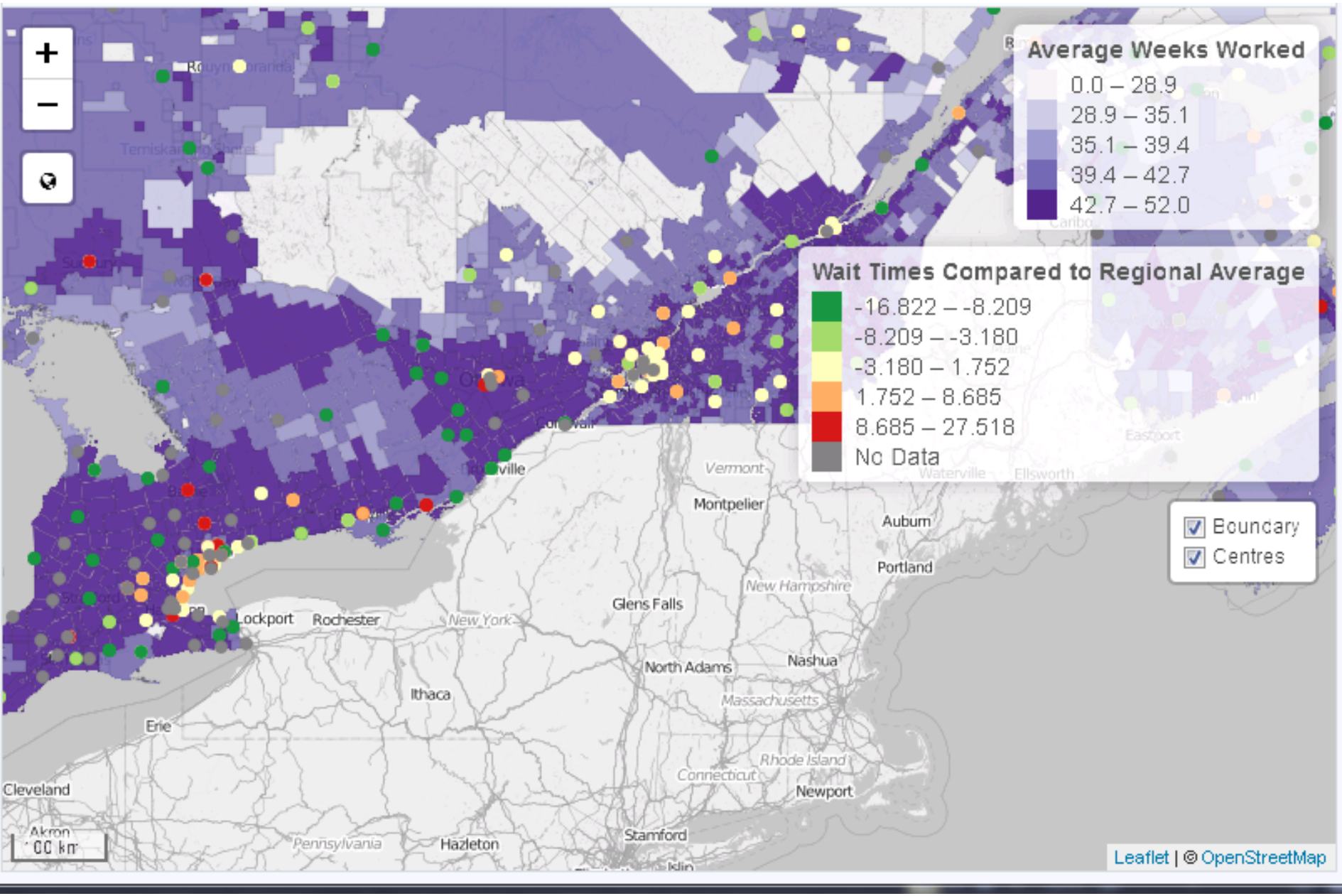
Flows

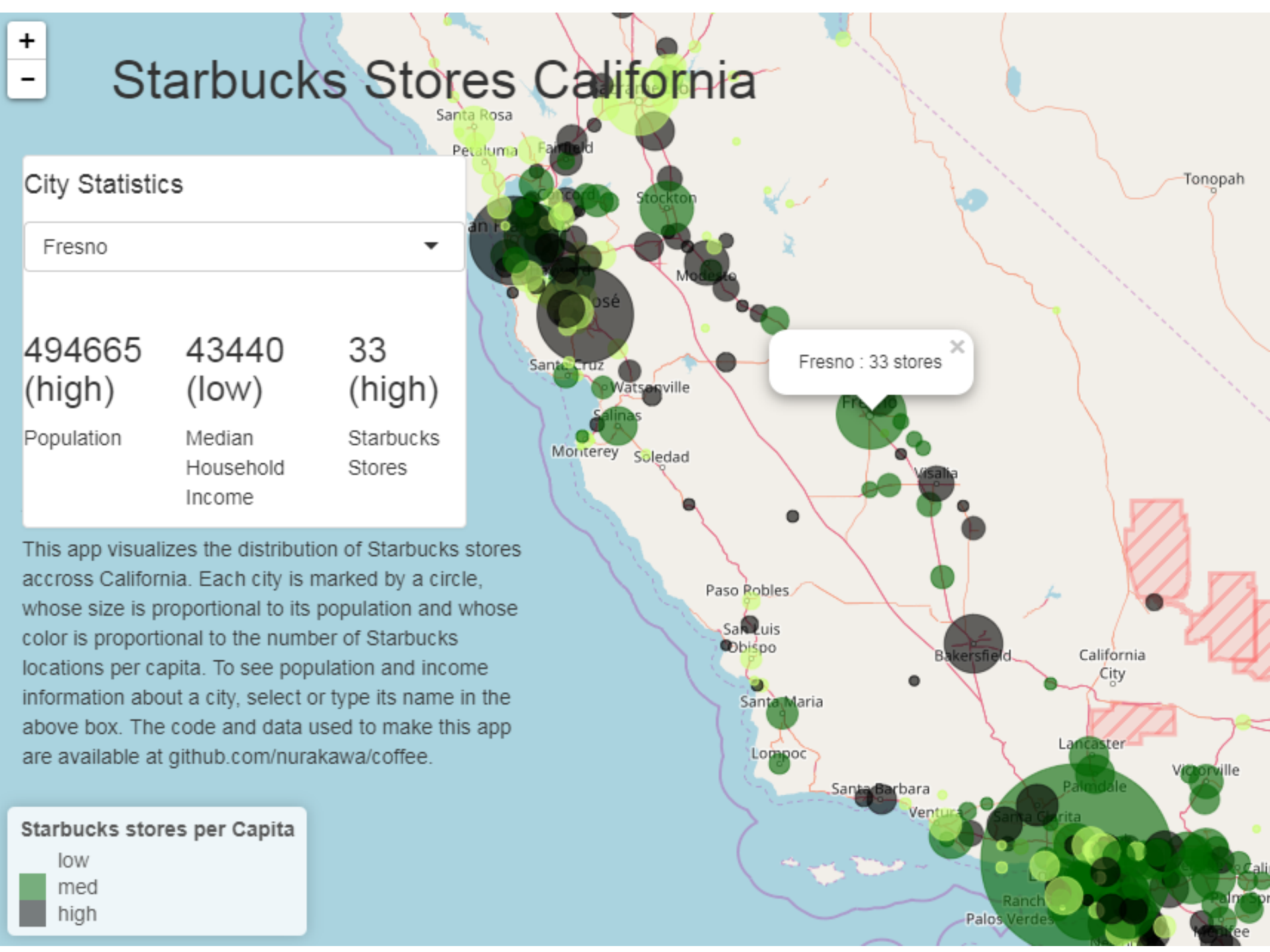


Cartograms



https://www.loxodrome.io/post/effective-cartograms/


Average Weeks Worked
0.0 – 28.9
28.9 – 35.1
35.1 – 39.4
39.4 – 42.7
42.7 – 52.0
Wait Times Compared to Regional Average
-16.822 – -8.209
-8.209 – -3.180
-3.180 – 1.752
1.752 – 8.685
8.685 – 27.518
No Data
Boundary
Centres
Vermont
Montpelier
Auburn
Portland
New Hampshire
Glens Falls
Lockport
Rochester
New York
North Adams
Nashua
Massachusetts
Ithaca
Erie
Rhode Island
Connecticut
Newport
Cleveland
Akron
Pennsylvania
Hazleton
Stamford
Waterville
Ellsworth
Eastport
Leaflet | © OpenStreetMap

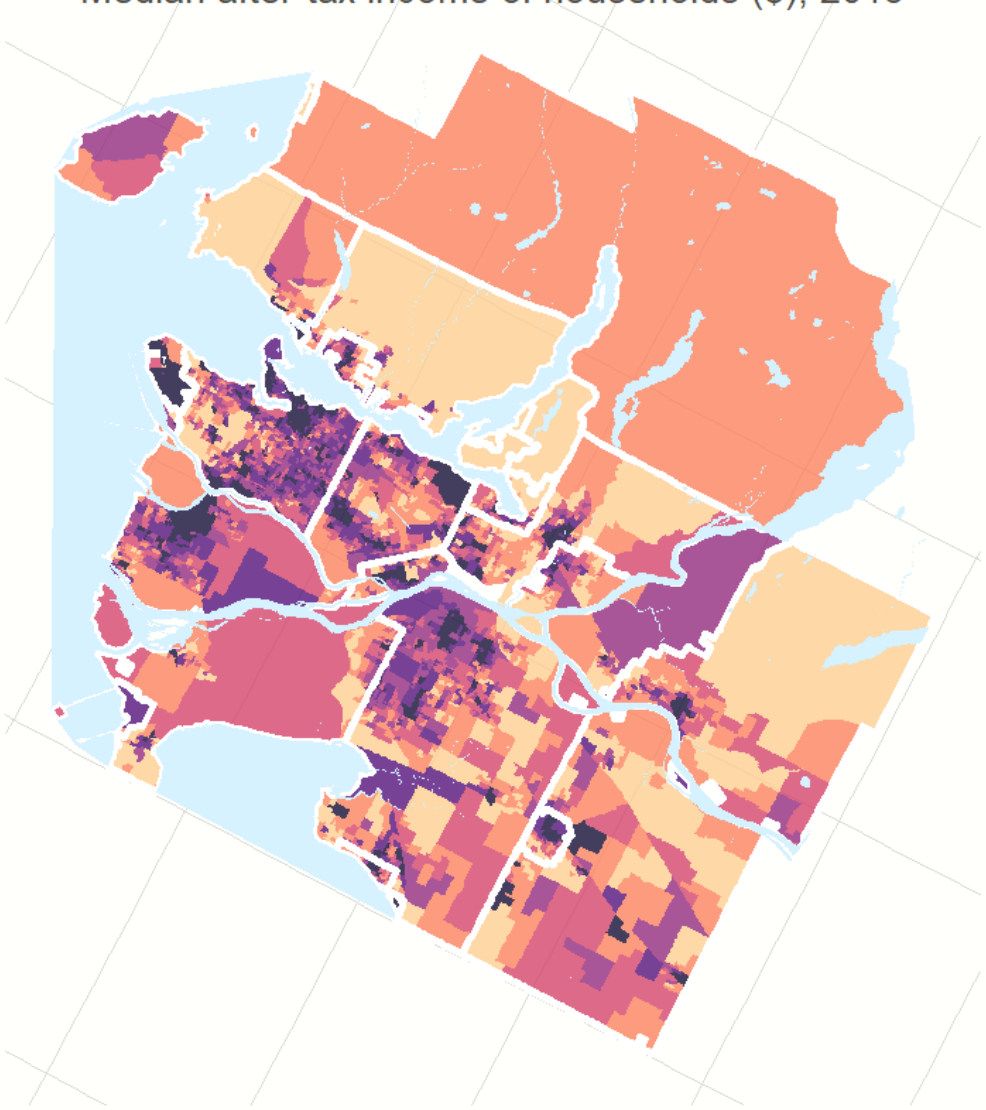

Starbucks Stores California
City Statistics
Fresno
494665 (high)
Population
43440 (low)
Median Household Income
33 (high)
Starbucks Stores
This app visualizes the distribution of Starbucks stores accross California. Each city is marked by a circle, whose size is proportional to its population and whose color is proportional to the number of Starbucks locations per capita. To see population and income information about a city, select or type its name in the above box. The code and data used to make this app are available at github.com/nurakawa/coffee.
Starbucks stores per Capita
low
med
high
Fresno : 33 stores
Santa Rosa
Petaluma
Fairfield
Stockton
Modesto
Tonopah
Santa Cruz
Watsonville
Salinas
Monterey
Soledad
Visalia
Paso Robles
San Luis Obispo
Bakersfield
California City
Santa Maria
Lompoc
Lancaster
Victorville
Santa Barbara
Ventura
Palmdale
Palos Verdes

Choropleth Maps

- **Maps that show the spatial distribution of data linked to a specific areal unit** (e.g., province/territory, census tract)
- The data is either **classified** or **unclassified** and *should* be **normalized** to avoid population maps
- Display data values through **saturation/lightness** (quantitative) or **hues** (qualitative)
- Can organize, simplify, and *generalize* large information and complex spatial patterns into informative visualizations

Income Distribution in Greater Vancouver

Median after-tax income of households ($), 2015



Median after-tax income ($)

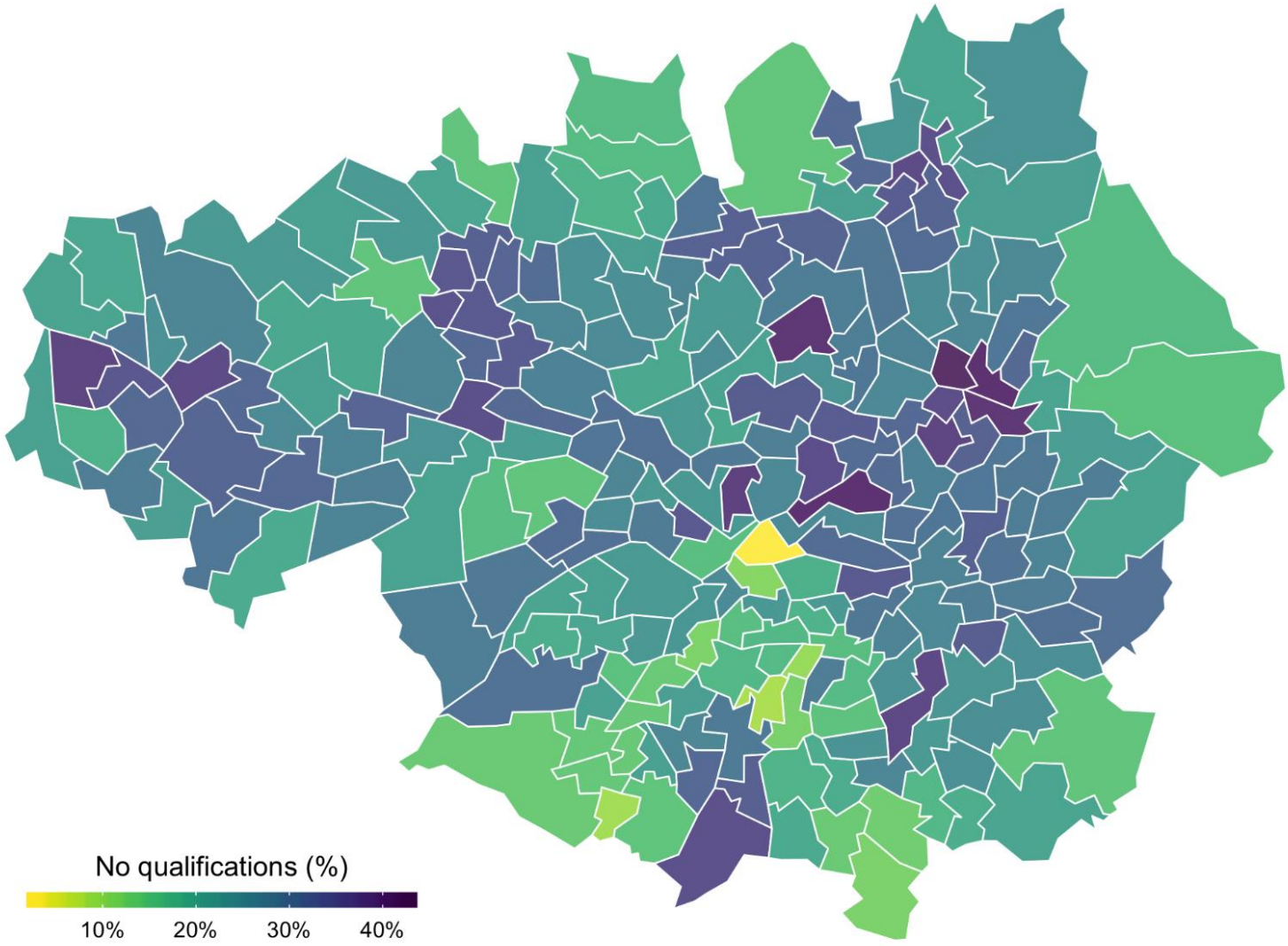
- NA
- 91k - 157k
- 79k - 91k
- 71k - 79k
- 62k - 71k
- 50k - 62k
- 12k - 50k

Bins

Authors: Julia Conzon (@Noznoc)
Geometries: Dissemination Area; Data: StatCan Census, 2016, & Proximity Measures, 2020

Residents with no qualifications in Greater Manchester, 2011

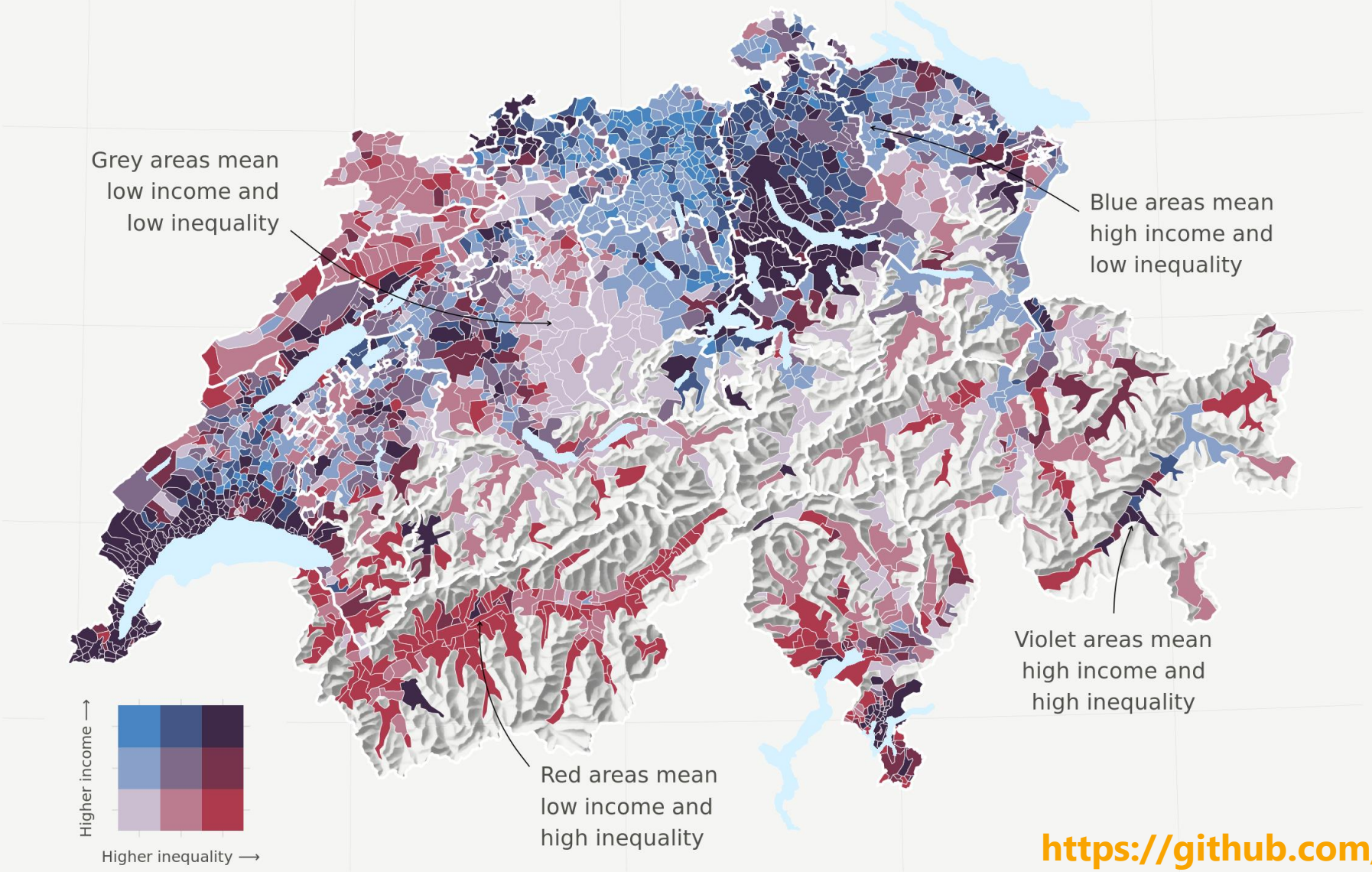
Source: Table QS502EW, Census 2011



https://medium.com/@traffordDataLab/lets-make-a-map-in-r-7bd1d9366098

Switzerland's regional income (in-)equality

Average yearly income and income (in-)equality in Swiss municipalities, 2015



https://github.com/grssnbchr/bivariate-maps-ggplot2-sf

Choropleth Development & Design

1. Select a **study area (areal unit)**, data (e.g., indicators), and **theme** of interest
2. Classify the values
    - Assess **distribution (histogram)**
    - If data is quantitative:
        - Determine whether to use classes, and if so the **number of classes**
        - Choose the **classification method**
3. Select **colour scheme**
4. Plot map (static or interactive)

Areal Units

Data is usually aggregated to a **Statistic Canada boundary files**:

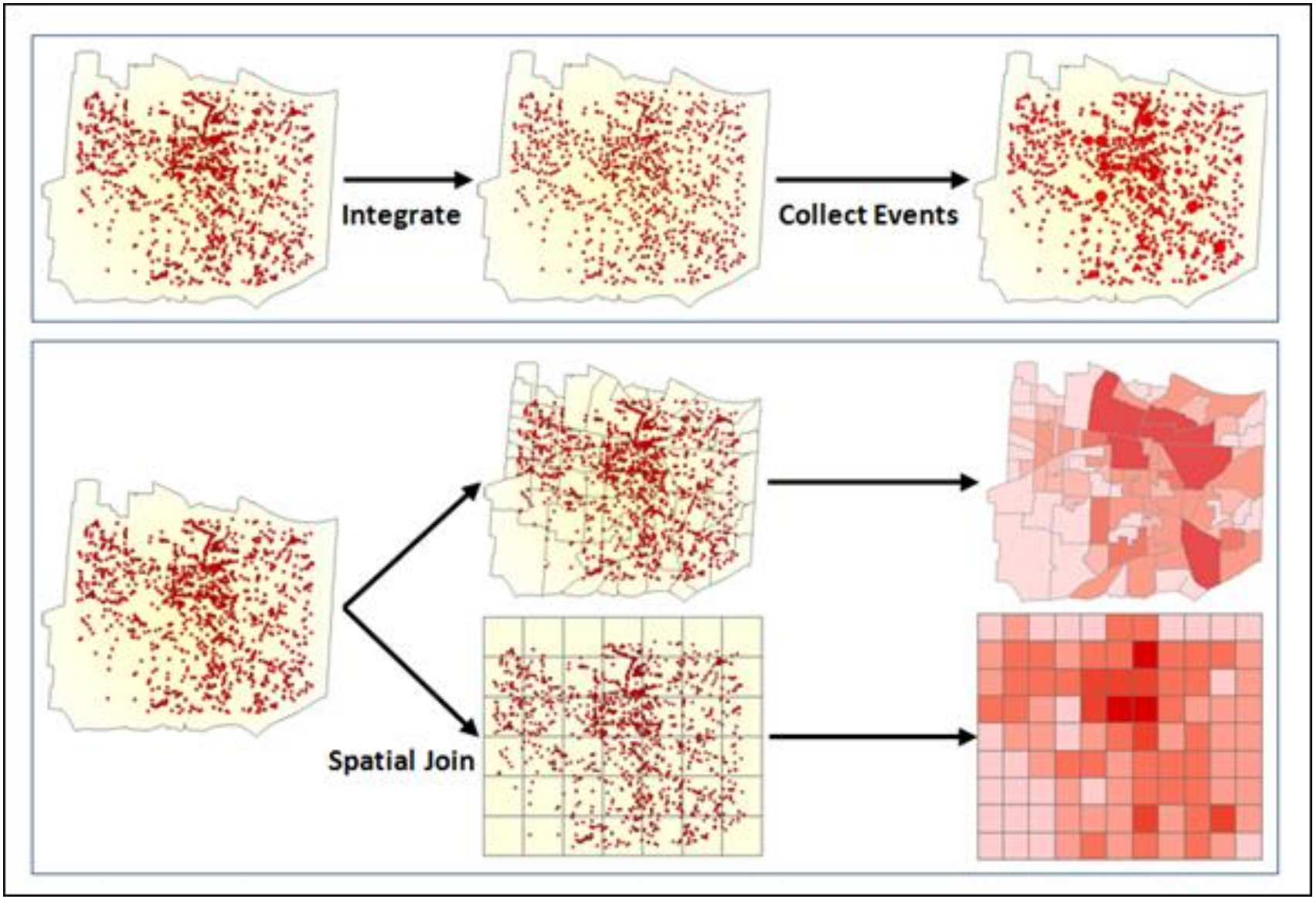
Provinces/territories; Federal electoral districts; Economic regions (ER); Census divisions (CDs); Aggregate dissemination areas (ADAs); Census subdivisions (CSDs); Census metropolitan areas and agglomerations (CMAs and CAs); Census tracts (CTs); Dissemination areas (DAs) and blocks (DBs)

Forward Sortation Areas (FSA) and Postal Codes, owned by Canada Post

Building footprints

https://www150.statcan.gc.ca/n1/pub/92-196-x/92-196-x2016001-eng.htm

Aggregate Points to an Areal Unit



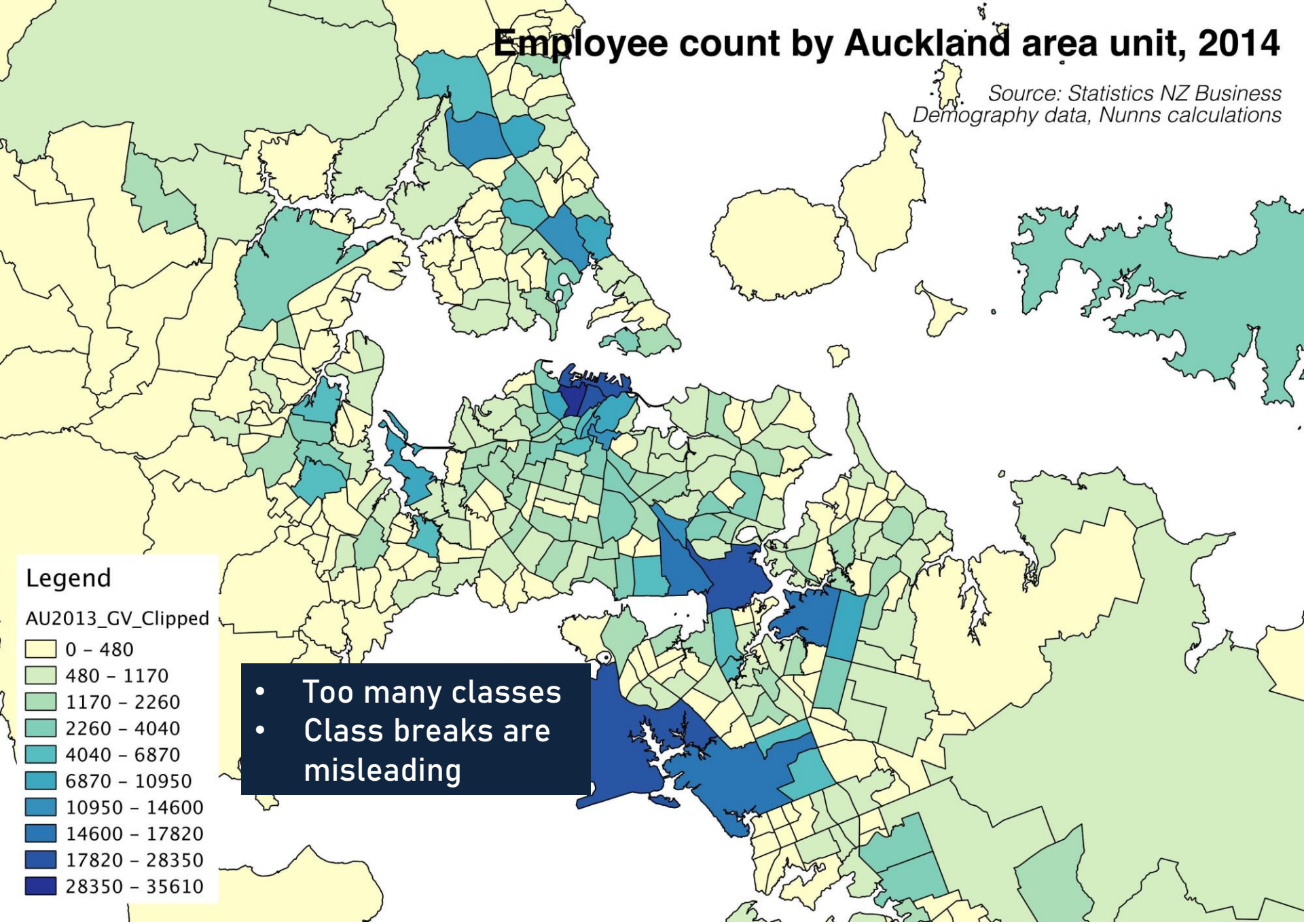
To Classify or Not

Depends on your data range, theme/message, and audience

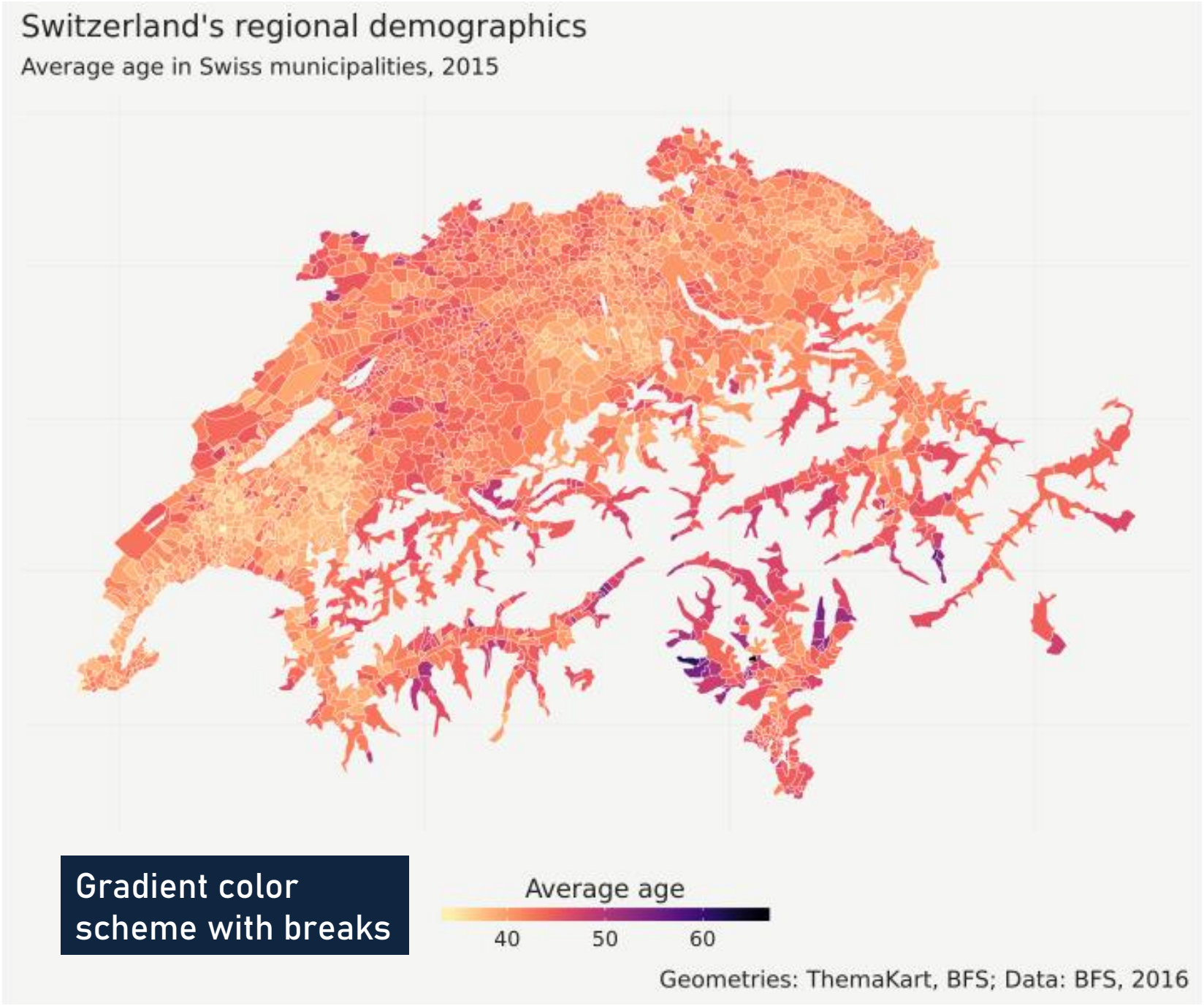
5-7 breaks/classes are optimal (Brewer and Pickle 2002)

If your variable's value range is large, then consider unclassified (graduated color scheme)

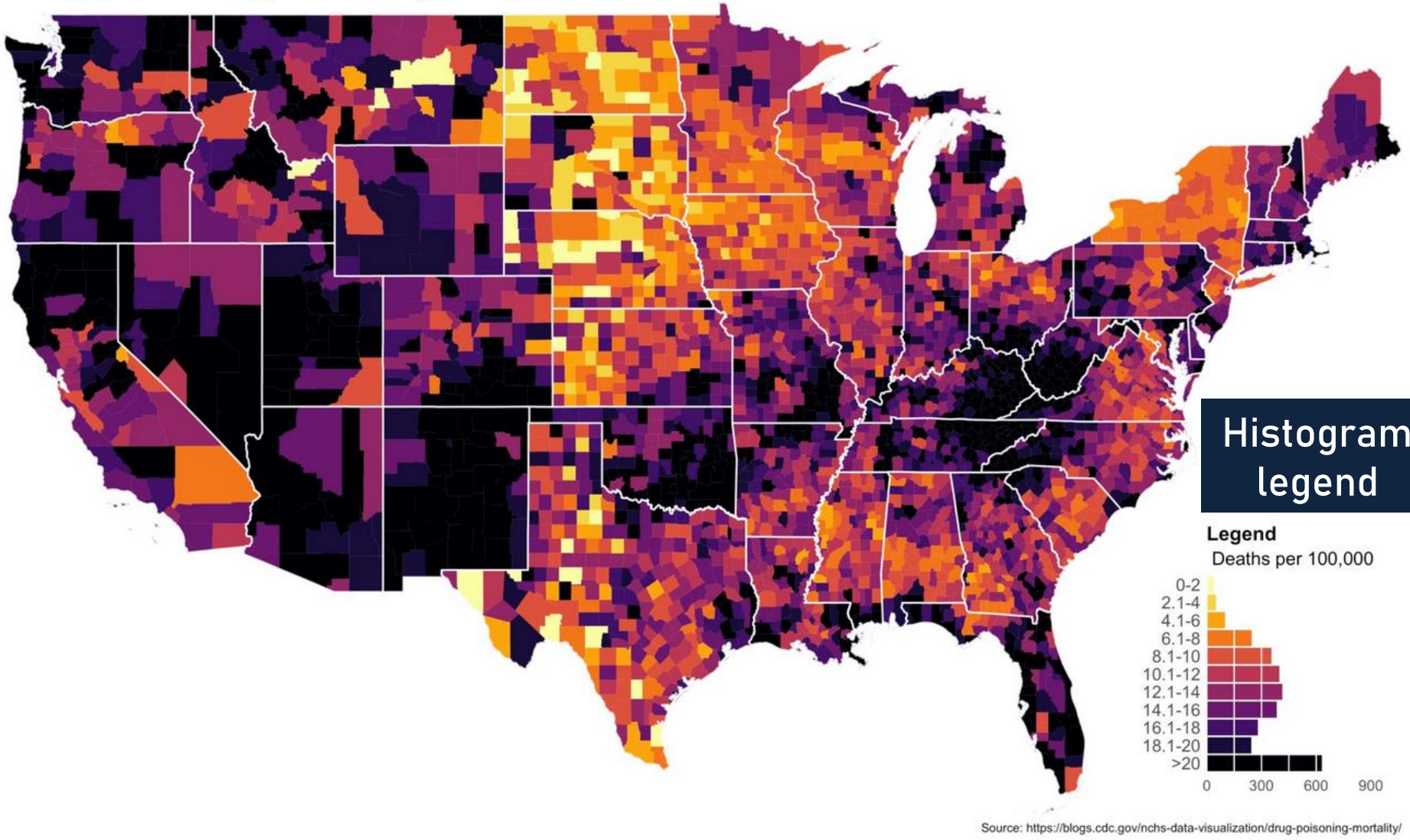
How are the data distributed throughout the range? And what, if any, class breaks might have a particular meaning to the map user? – Mark Monmonier (p. 163)



https://www.greaterauckland.org.nz/2015/01/13/the-geometry-of-urban-labour-markets/



https://blog.revolutionanalytics.com/2016/12/swiss-map.html



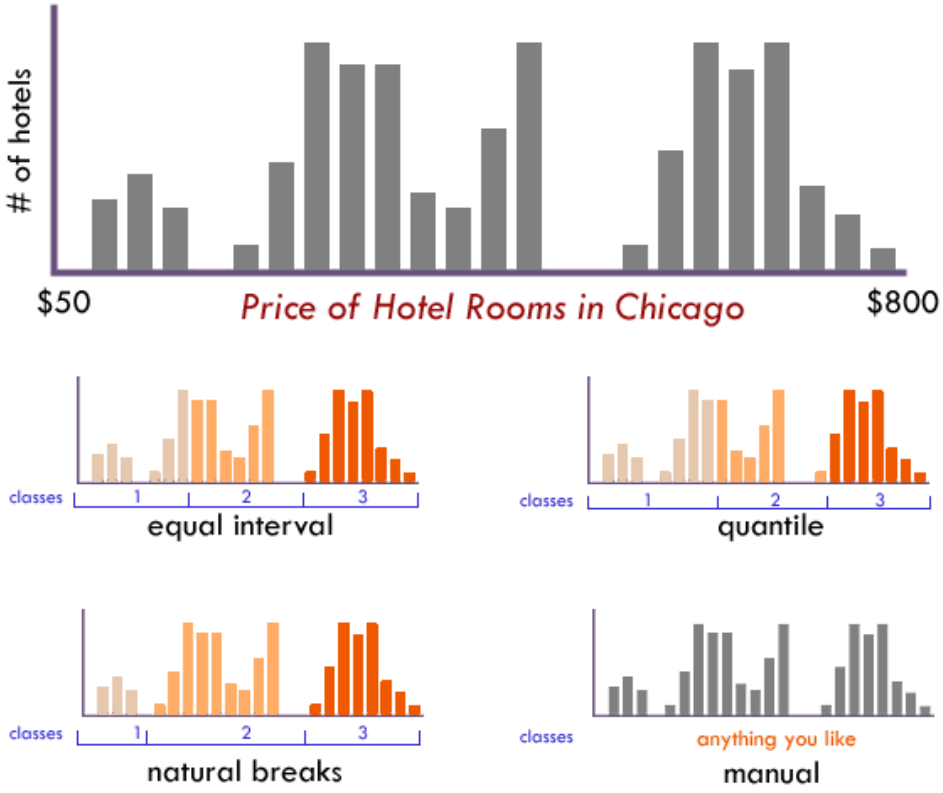
https://mathewkiang.com/2017/01/16/using-histogram-legend-choropleths/

Classification

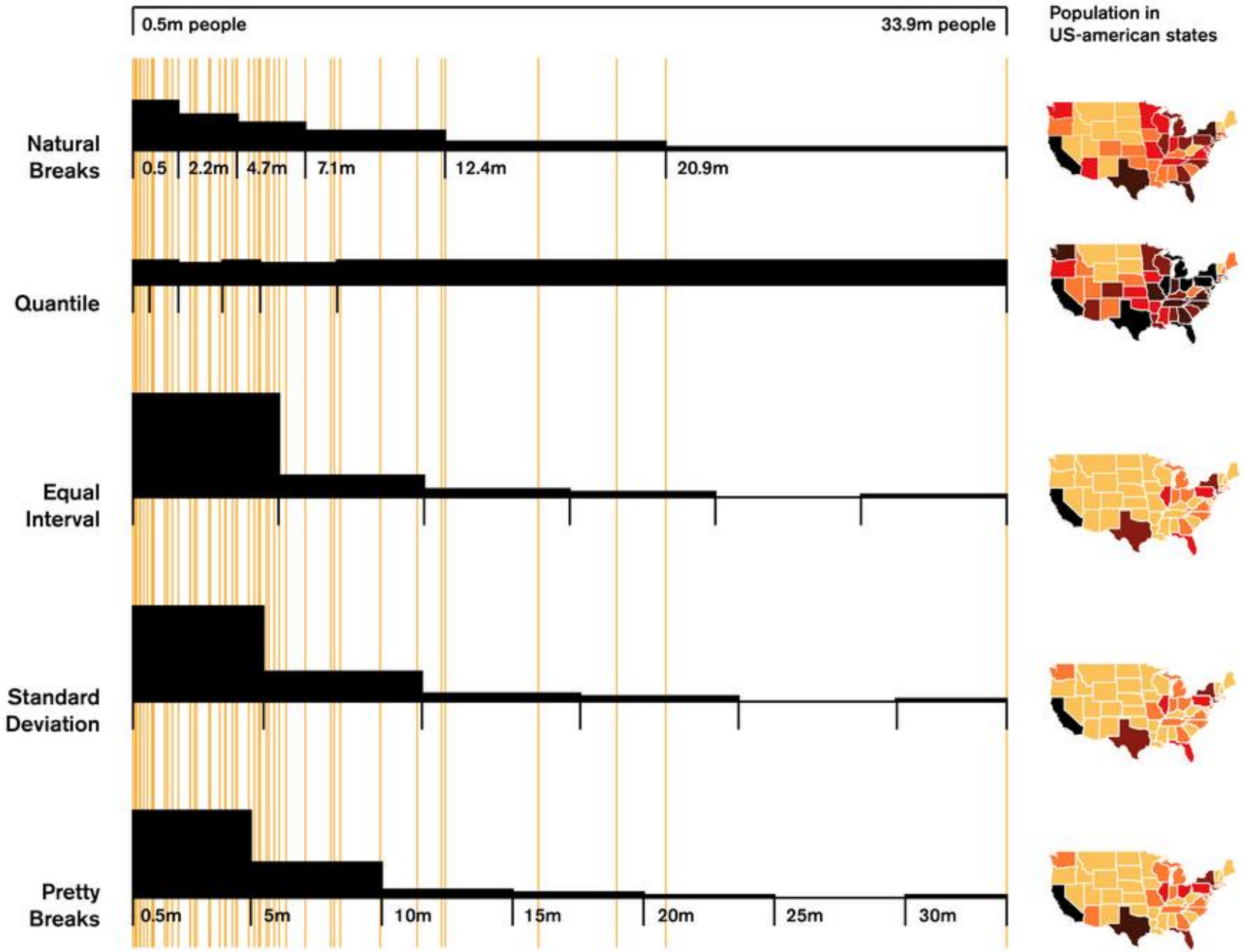
Classes/Bins group data values while **stops/breaks** are the values that separate the classes

Classification methods organize similar values into classes

- Quantile
- Standard deviation
- Equal interval
- Jenks (Natural Breaks)
- Manual
- Pretty breaks



https://www.axismaps.com/guide/data/data-classification/


0.5m people
33.9m people
Population in US-american states
Natural Breaks
0.5
2.2m
4.7m
7.1m
12.4m
20.9m
Quantile
Equal Interval
Standard Deviation
Pretty Breaks
0.5m
5m
10m
15m
20m
25m
30m

Classification Methods in R

With **tmap** R package to create the choropleth maps. This package depends on the **classIntervals** package for the classification methods:

https://www.rdocumentation.org/packages/classInt/versions/0.3-3/topics/classIntervals

- Manual = "fixed"
- Quantile = "quantile"
- Pretty breaks = "pretty"
- Jenks = "jenks"
- Standard deviation = "sd"
- Equal intervals based on range = "equal"

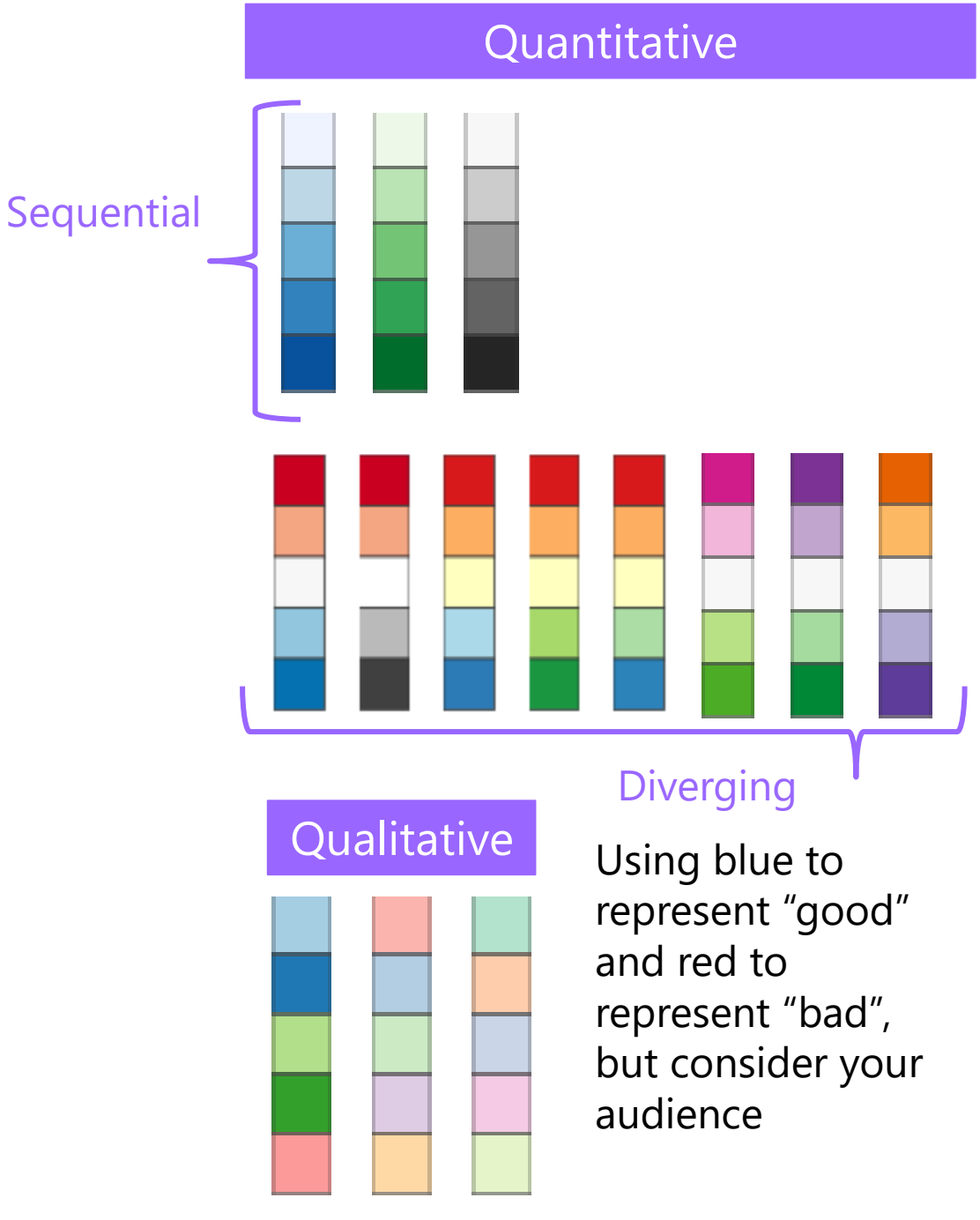
Not discussed, but are offered in package:

- "k means"
- "fisher"
- Hierarchical clustering = "hclust"
- Bagged clustering = "bclust"

Colours

Is your data qualitative or quantitative? Sequential or diverging?

What message are you trying to convey?



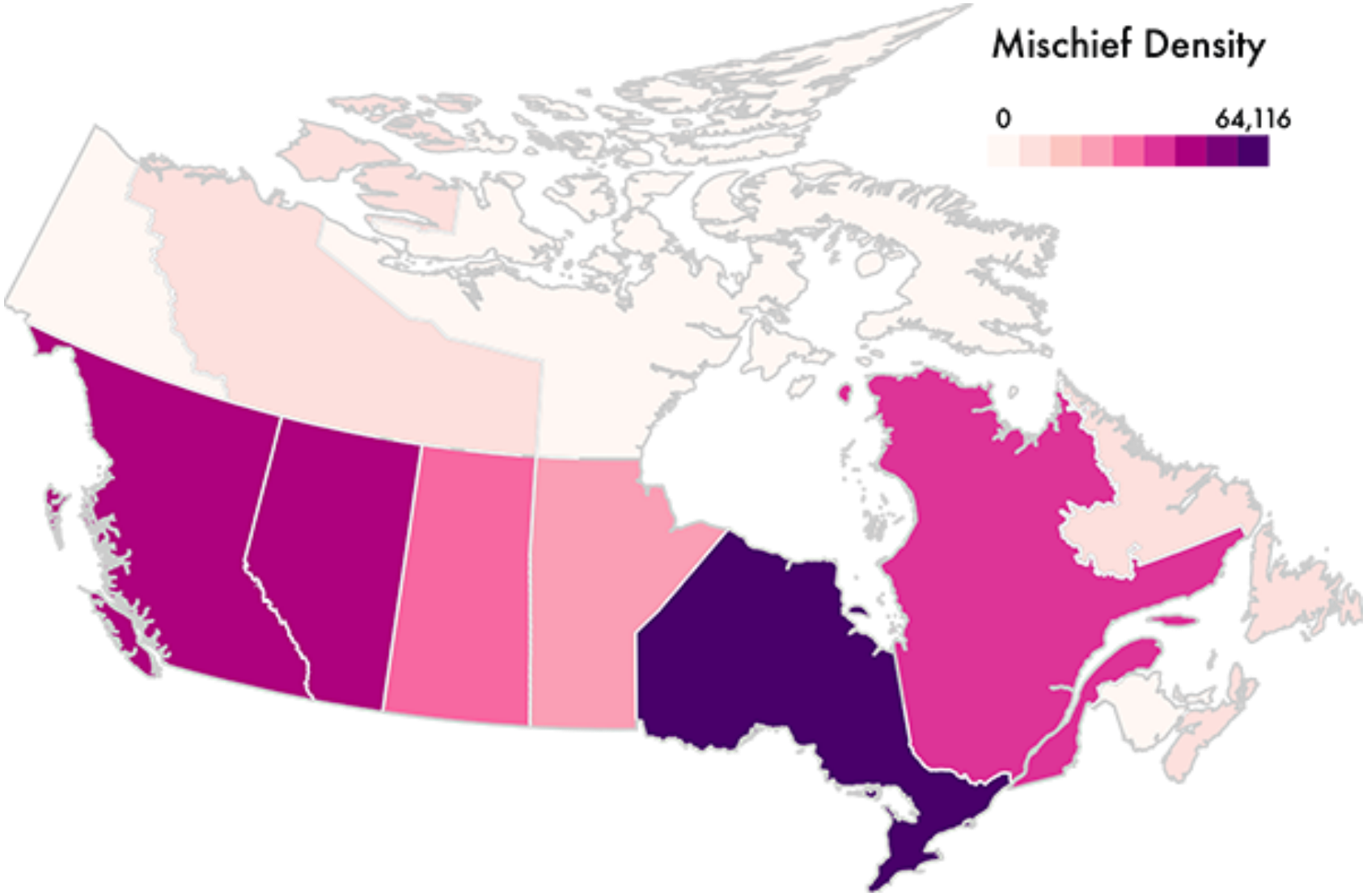
Using blue to represent “good” and red to represent “bad”, but consider your audience

http://colorbrewer2.org/

Choropleth Reflections + Limitations

- Default classification on software
- Modifiable Areal Unit Problem (MAUP): boundaries can influence the presentation of the data, different boundary aggregates will yield different results
- Comparing choropleth maps / correlation choropleth maps
- Always question the map!
  - Data sources
  - Classification
  - What is the map trying to show or hide?

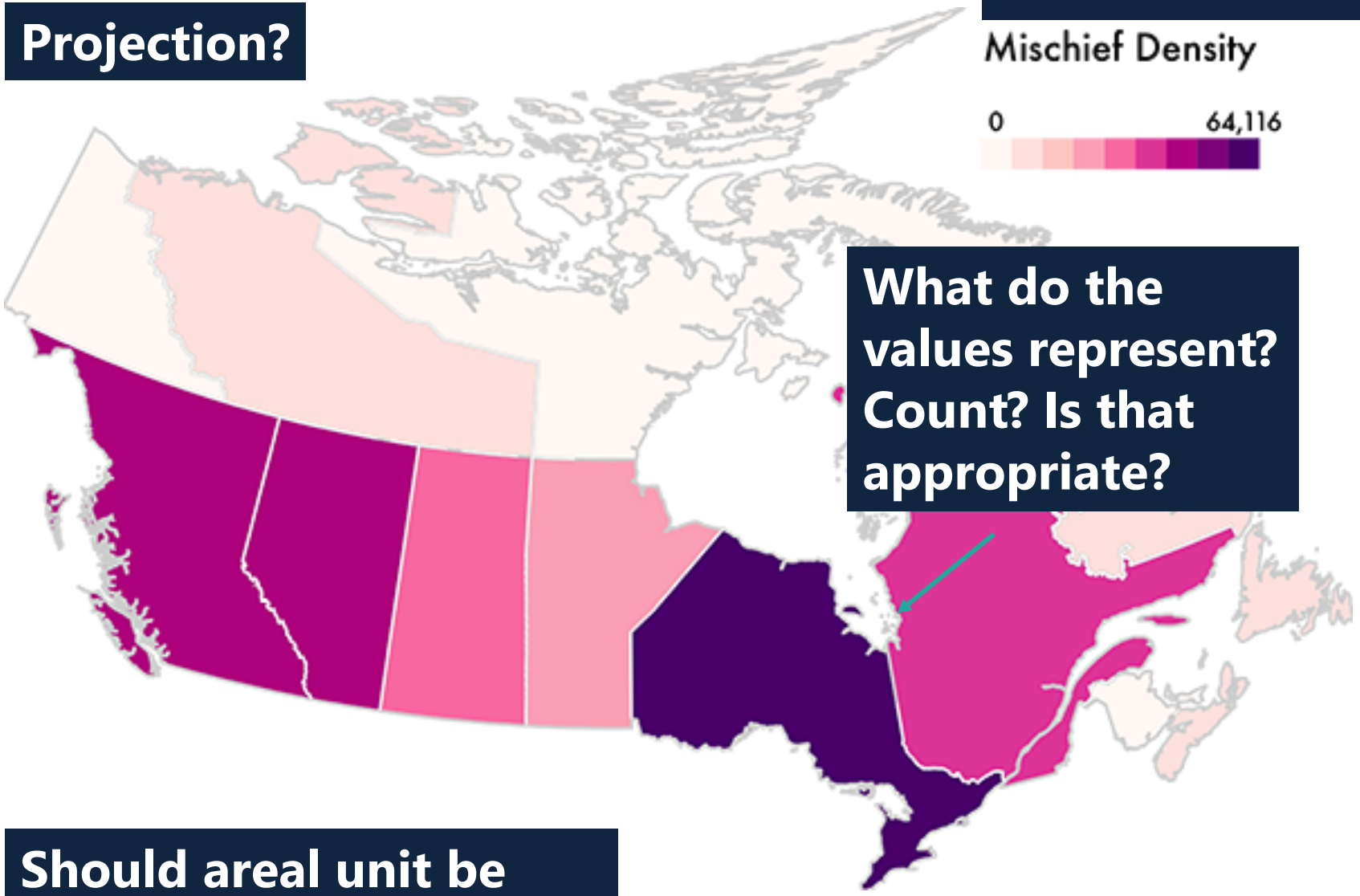
Question the map!



Title?

Projection?

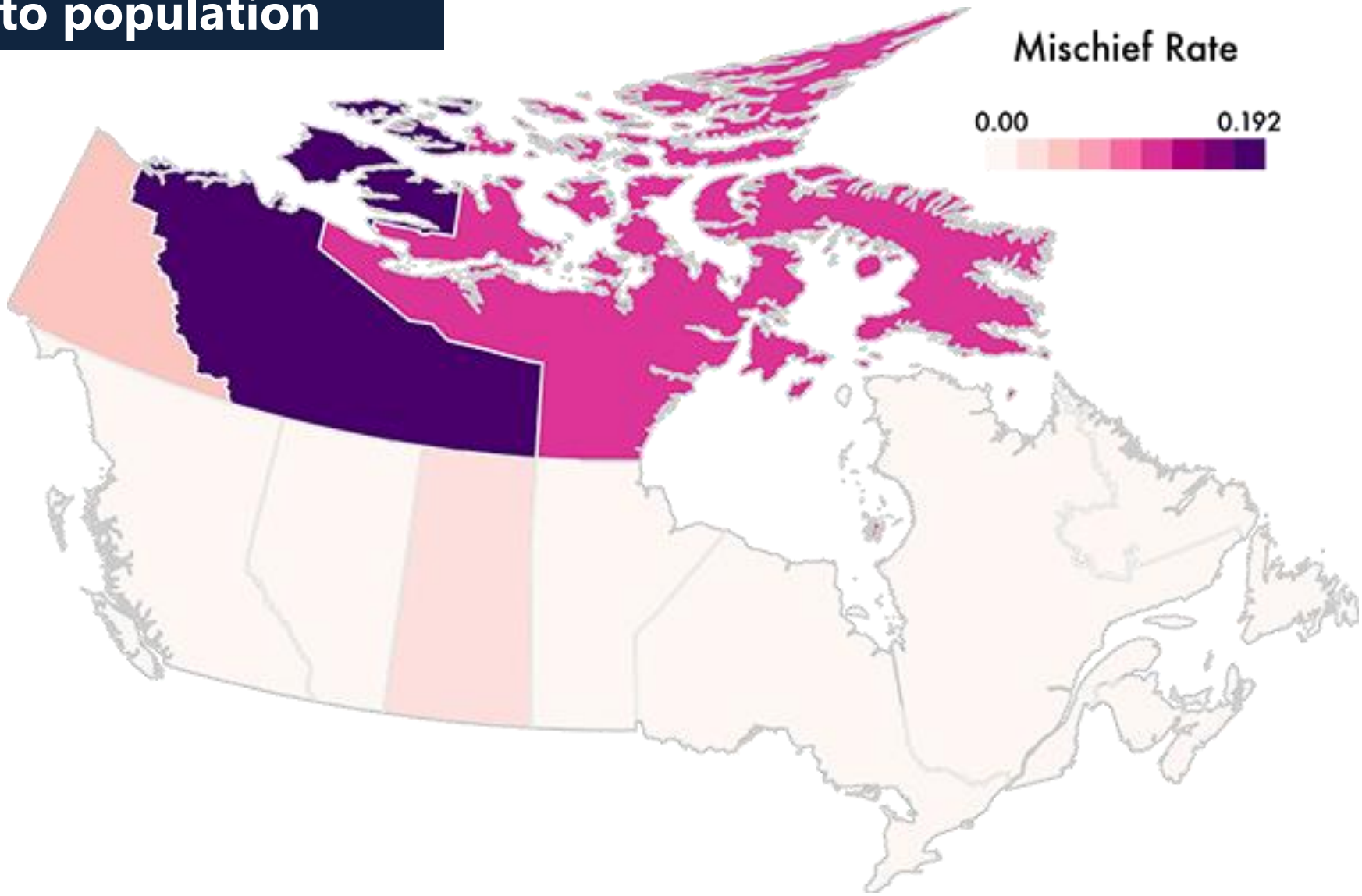
How is mischief defined?



What do the values represent? Count? Is that appropriate?

Should areal unit be more granular?

Now proportional to population



https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865

R Mapping Code + Tutorials

tmap

- Code: **https://github.com/Noznoc/r-gis-workshop**
- Tutorial: **https://noznoc.github.io/r-gis-workshop/index.html**

ggplot2 + sf

- Code: **https://github.com/Noznoc/bivariate-maps-ggplot2-sf/tree/statcan**
- Tutorial: **https://noznoc.github.io/bivariate-maps-ggplot2-sf/index.html**

R + Choropleth Resources

- https://www.r-graph-gallery.com/chloropleth-map/R choropleths
- https://r4ds.had.co.nz/workflow-basics.html
- https://geocompr.robinlovelace.net/adv-map.html
- https://www12.statcan.gc.ca/census-recensement/2011/geo/map-carte/ref/thematicmaps-cartesthematiques-index-eng.cfm
- https://timogrossenbacher.ch/2019/04/bivariate-maps-with-ggplot2-and-sf/

Citations

Brewer, Cynthia A. and Linda Pickle. 2002. “Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series.” Annals of the Association of American Geographers, 92(4), 662-681. **http://www.personal.psu.edu/cab38/Brewer_Annals.pdf**

De Smith, Michael J., Michael F Goodchild, Paul A Longley. *Geospatial Analysis: A Comprehensive Guide to Principles Techniques and Software* Tools. 2018. **www.spatialanalysisonline.com/HTML**

Jenks, G. F., Caspall, F. C., 1971. "Error on choroplethic maps: definition, measurement, reduction". Annals, Association of American Geographers, 61 (2), 217-244

Monmonier, Mark. *How to Lie with Maps*. University of Chicago Press. 2018.

Appendix

What is R?

Open source interpreted language and environment for data handling, storage, analysis and graphics

Users leverage existing **packages** within the Comprehensive R Archive Network (**CRAN**) and can contribute to, or develop their own, packages

The **R environment** supports the entire data science/analysis workflow, from data collection to visualization/communication

https://cran.r-project.org/

What is RStudio?

Most popular Integrated Development Environment (IDE) for R

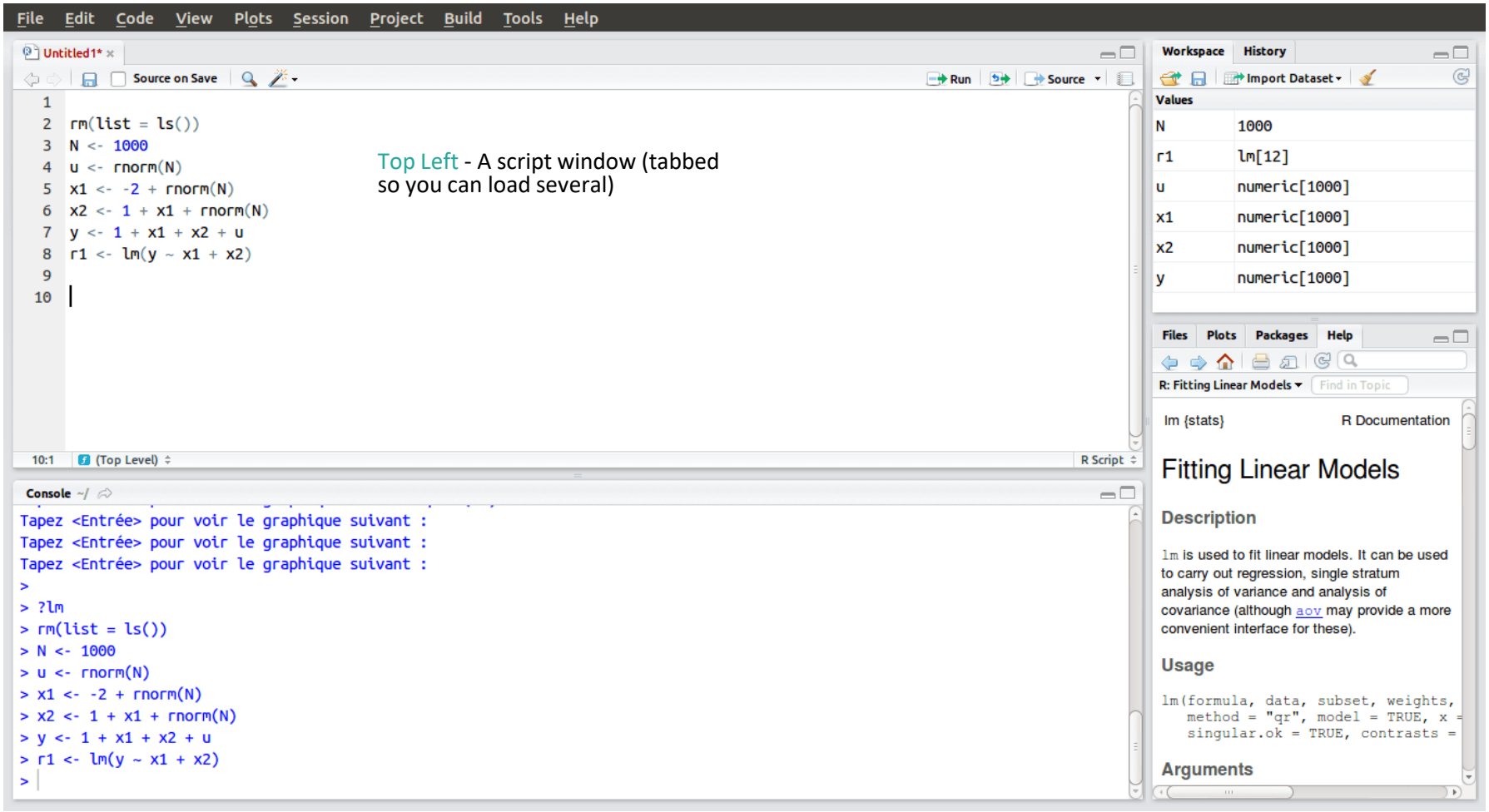
Available in two editions: RStudio Desktop and R Studio Server

Runs on Windows, Mac and Linux

Newest version (v1.2.1335) have support for Python and D3

https://www.rstudio.com/

RStudio User Interface



Well-Known Text (open)

Text markup language that stores vector geometry

Well-Known Binary (WKB) is used to transform this information in databases

Type	Example (WKT)
Point	POINT (30 10)
LineString	LINESTRING (30 10, 10 30, 40 40)
Polygon	POLYGON ((30 10, 40 40, 20 40, 10 20, 30 10))
	POLYGON ((35 10, 45 45, 15 40, 10 20, 35 10), (20 30, 35 35, 30 20, 20 30))

R simple features (sf)

https://www.r-spatial.org/r/2016/02/15/simple-features-for-r.html

KML (Google)

```
<Placemark>
  <name>Extruded placemark</name>
  <visibility>0</visibility>
  <description>Tethered to the ground by a customizable
    &amp;quot;tail&amp;quot;</description>
  <LookAt>
    <longitude>-122.084578742152 5</longitude>
    <latitude>37.42215078737763</latitude>
    <altitude>0</altitude>
    <heading>-148.4126684946234</heading>
    <tilt>40.55750733918048</tilt>
    <range>365.2646606980322</range>
  </LookAt>
  <styleUrl>#globeIcon</styleUrl>
  <Point>
    <extrude>1</extrude>
    <altitudeMode>relativeToGround</altitudeMode>
    <coordinates>-122.0857667006183,37.42156927867553,50</coordinates>
  </Point>
</Placemark>
```

https://developers.google.com/kml/documentation/kml_tut

GeoPackages (open)

Fairly new format for geospatial data (2014)

Stores geospatial information in a SQLite database:

- Vector features
- Imagery/Raster tilesets
- Attribute information

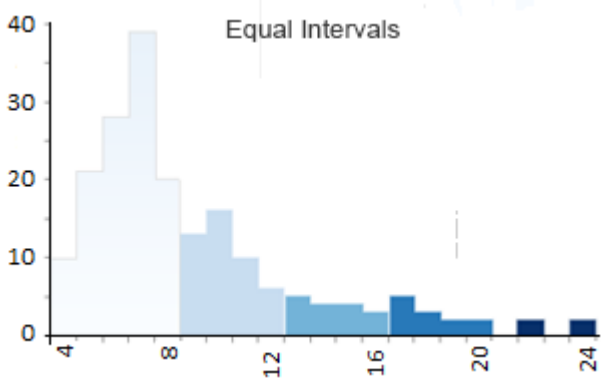
The GeoPackage Encoding Standard defines the schema

https://www.geopackage.org/

Classification Methods

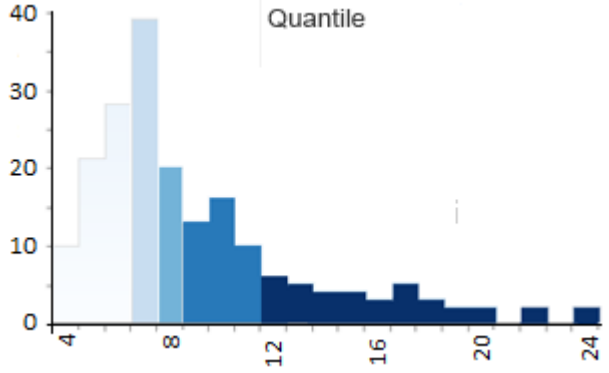
Equal interval based on range: groups data values into equal intervals

- (maxValue – minValue) / class #
- This can cause unequal amounts of values across the classes



Quantile: groups data into classes of same quantity (4 = quartile, 5 = quintile)

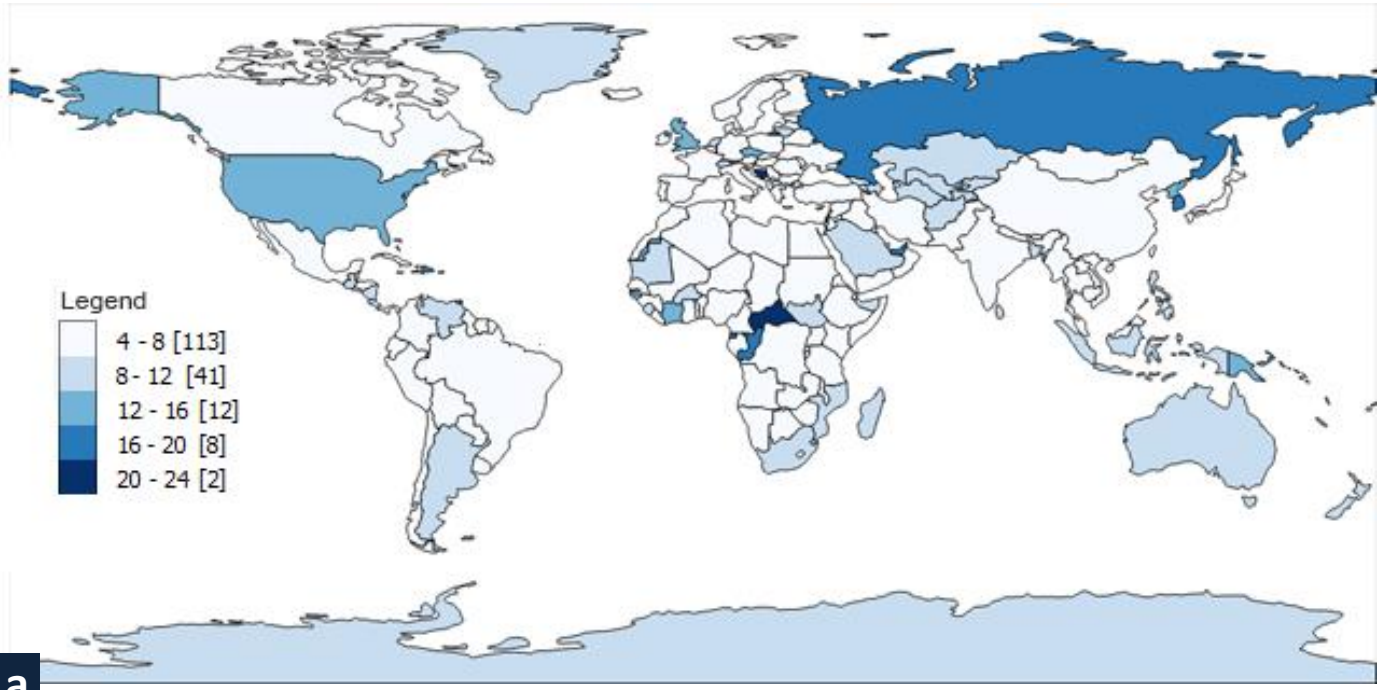
- (total # of features / class #) to get the average, then the method tries to count the quantity in each group closest to the average
- Can be easier to interpret, but the method can cause very different values being classified into same groups



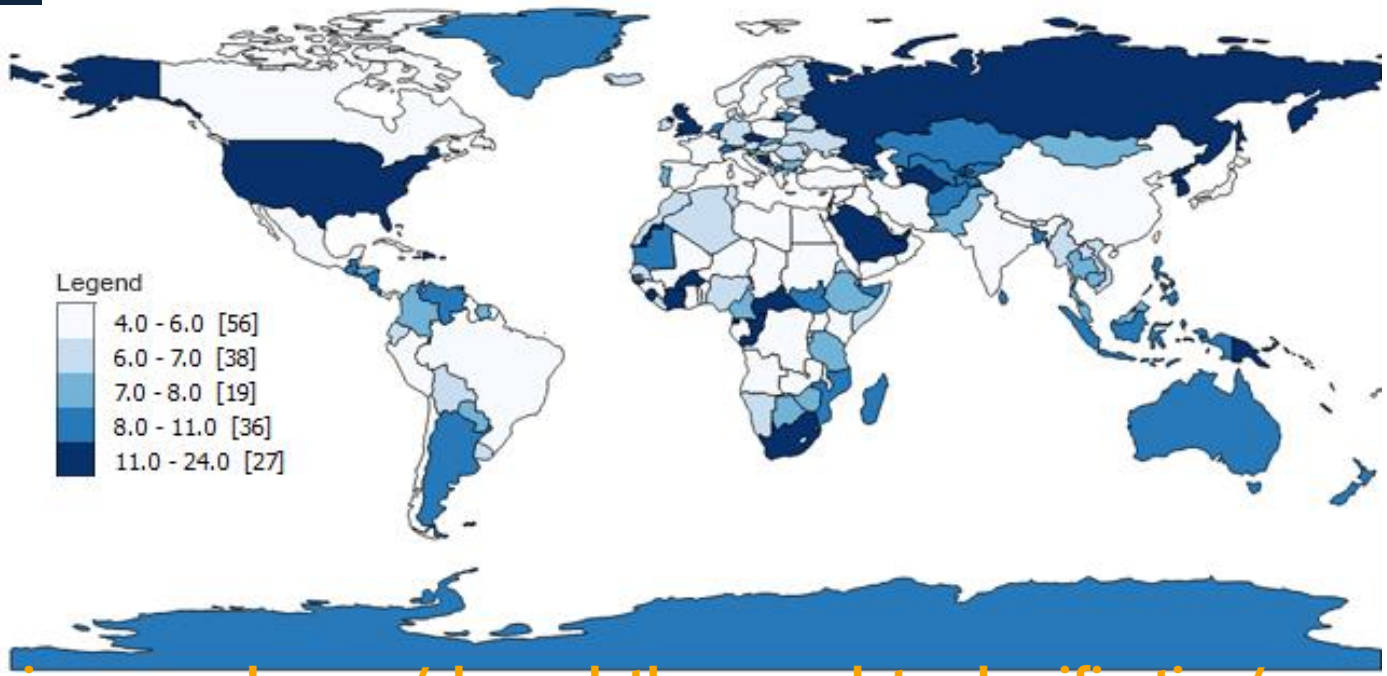
(De Smith et al. 2018)

Equal Interval

(values represent number of letters in a country's names)



Quantile

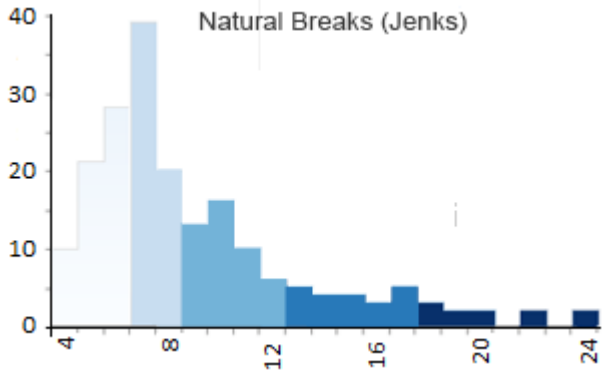


https://gisgeography.com/choropleth-maps-data-classification/

Classification Methods (continued)

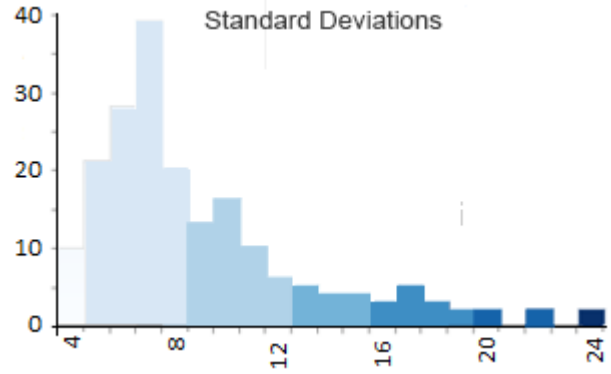
Jenks (Natural Breaks): groups data by reducing the amount of variance within classes and maximize variance between classes

- Method randomly selects breaks, then through an iterative process, minimizes each class's average deviation from the class mean
- Can highlight outliers in own classes
- Cannot be used to compare maps because of how the method's initial randomization step



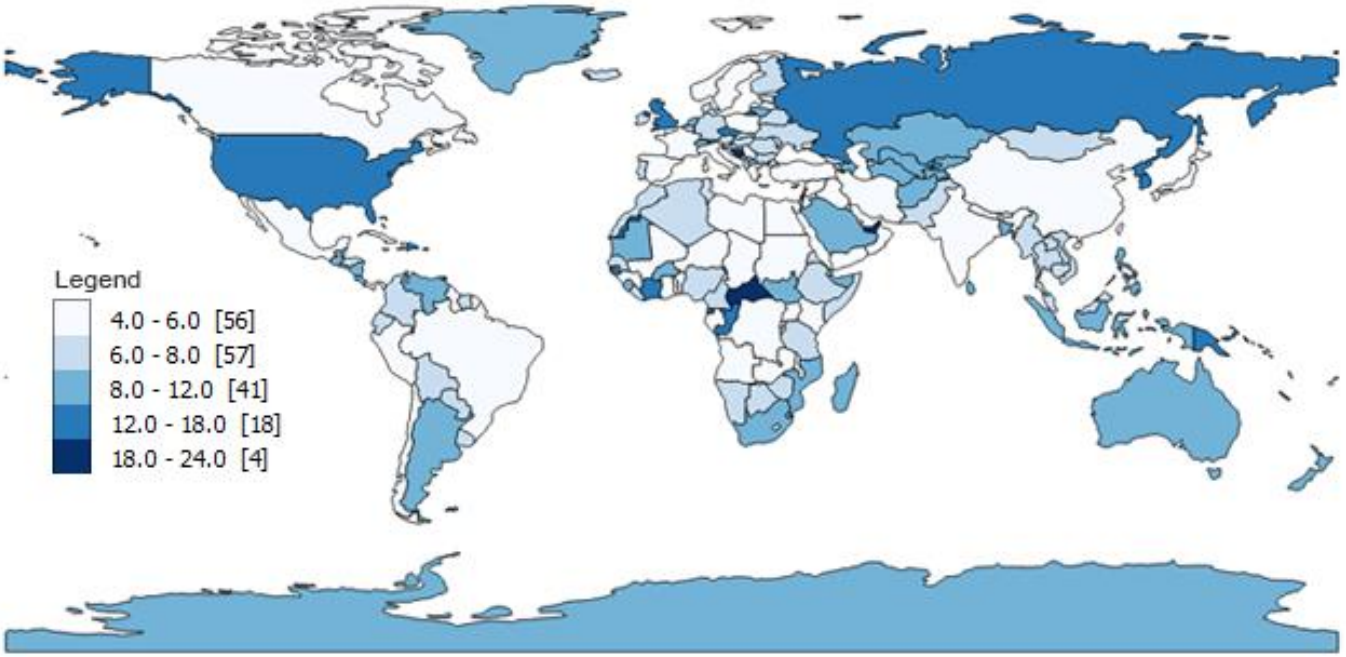
Standard Deviation: groups data by adding/subtracting the standard deviation from the mean

- Best for data that is normally distributed

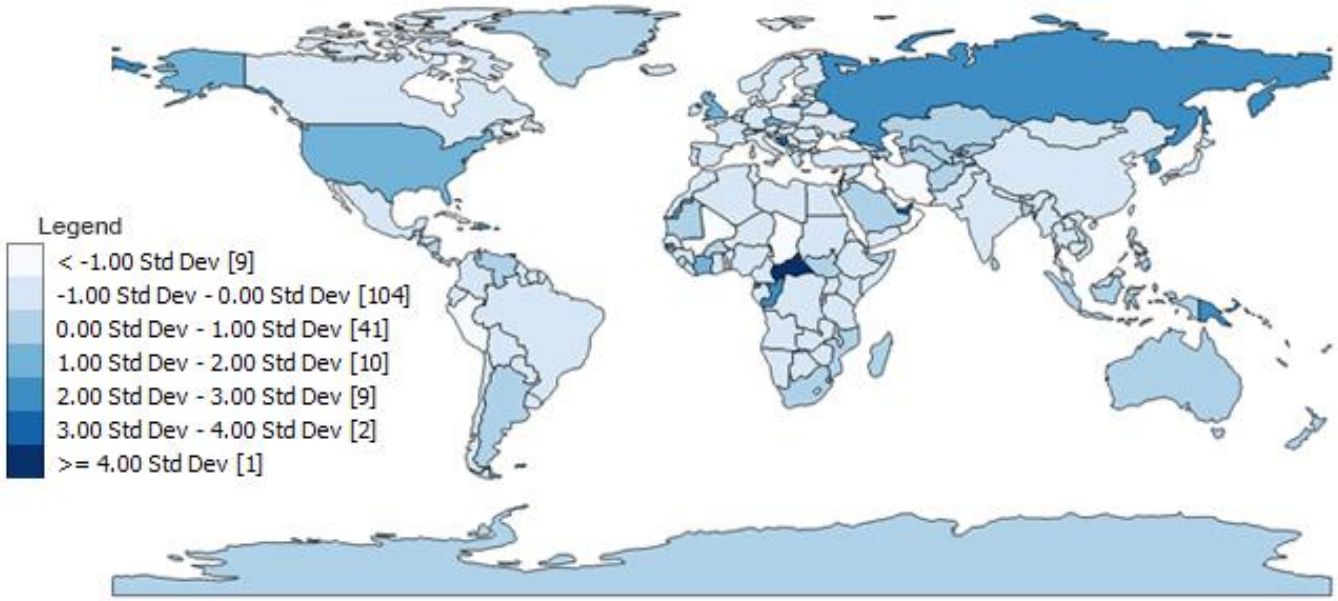


(De Smith et al. 2018; Jenks 1971)

Jenks



Standard Deviation

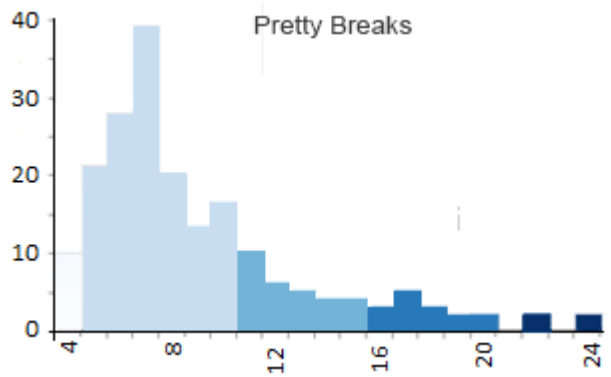


https://gisgeography.com/choropleth-maps-data-classification/

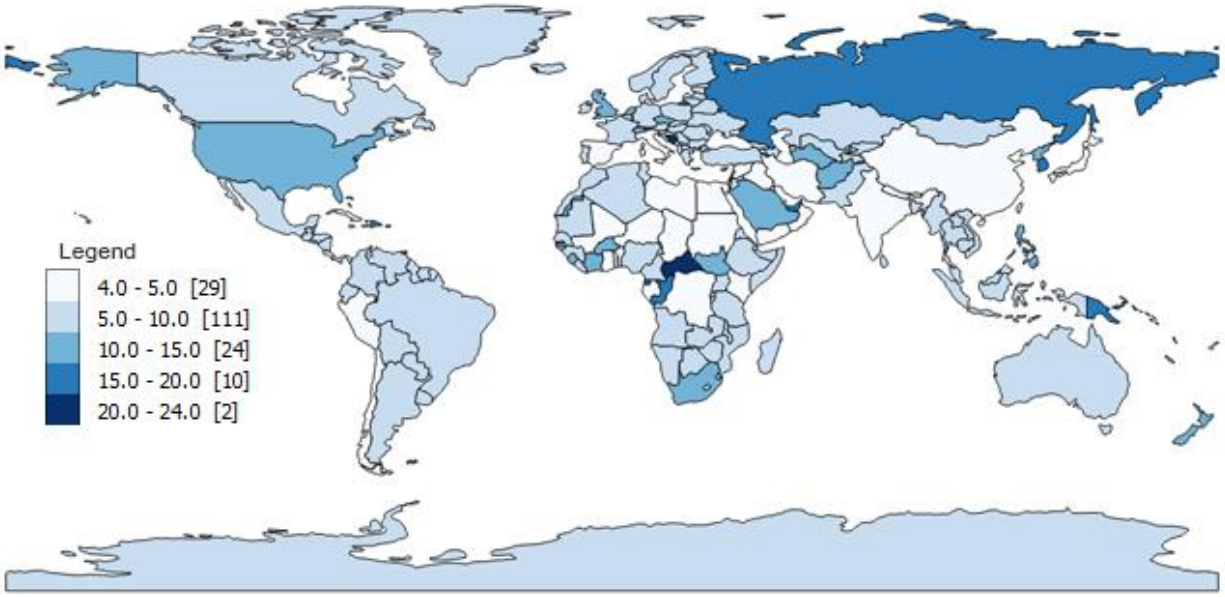
Classification Methods (continued)

Pretty breaks: groups data values into classes that allow the breaks to be rounded values

- Instead of a break being 499.231, it would be 500



Manual: manually select the breaks



(De Smith et al. 2018)