

visual narrative

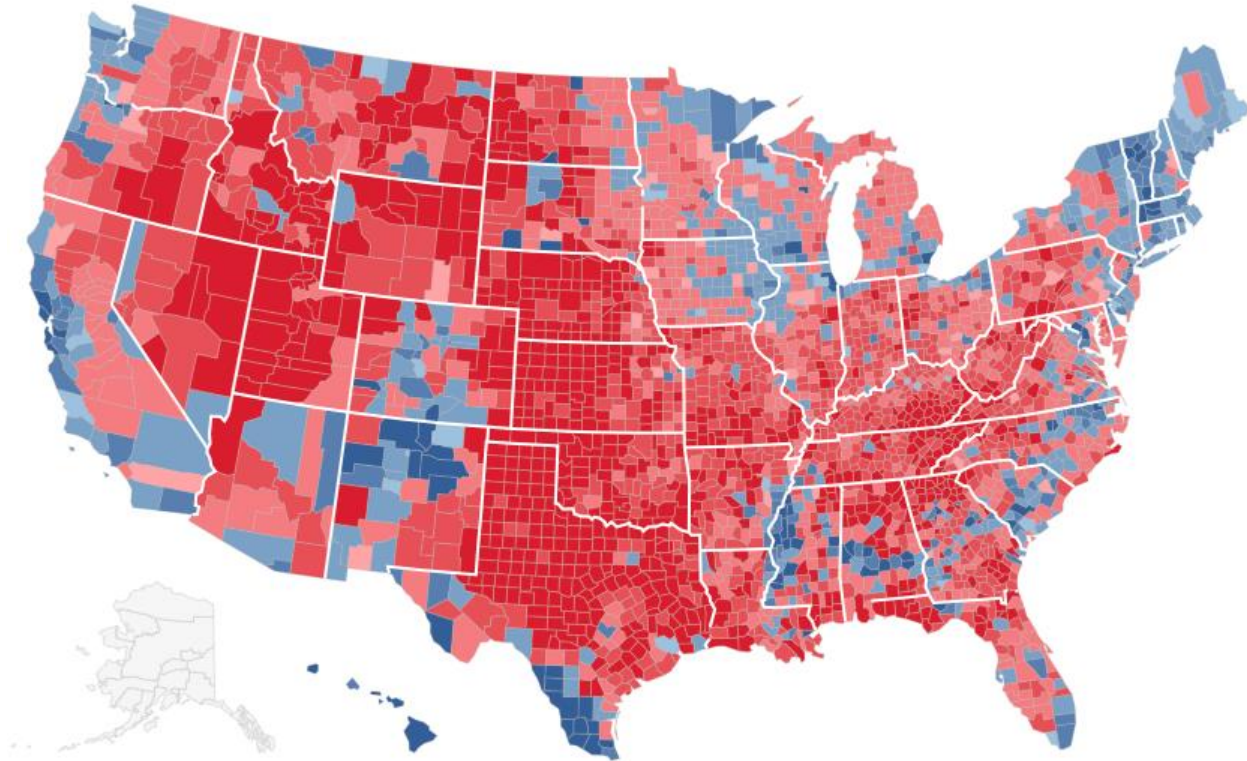
YOUR DATA TELLS A STORY

Does your visual narrative match your actual narrative?

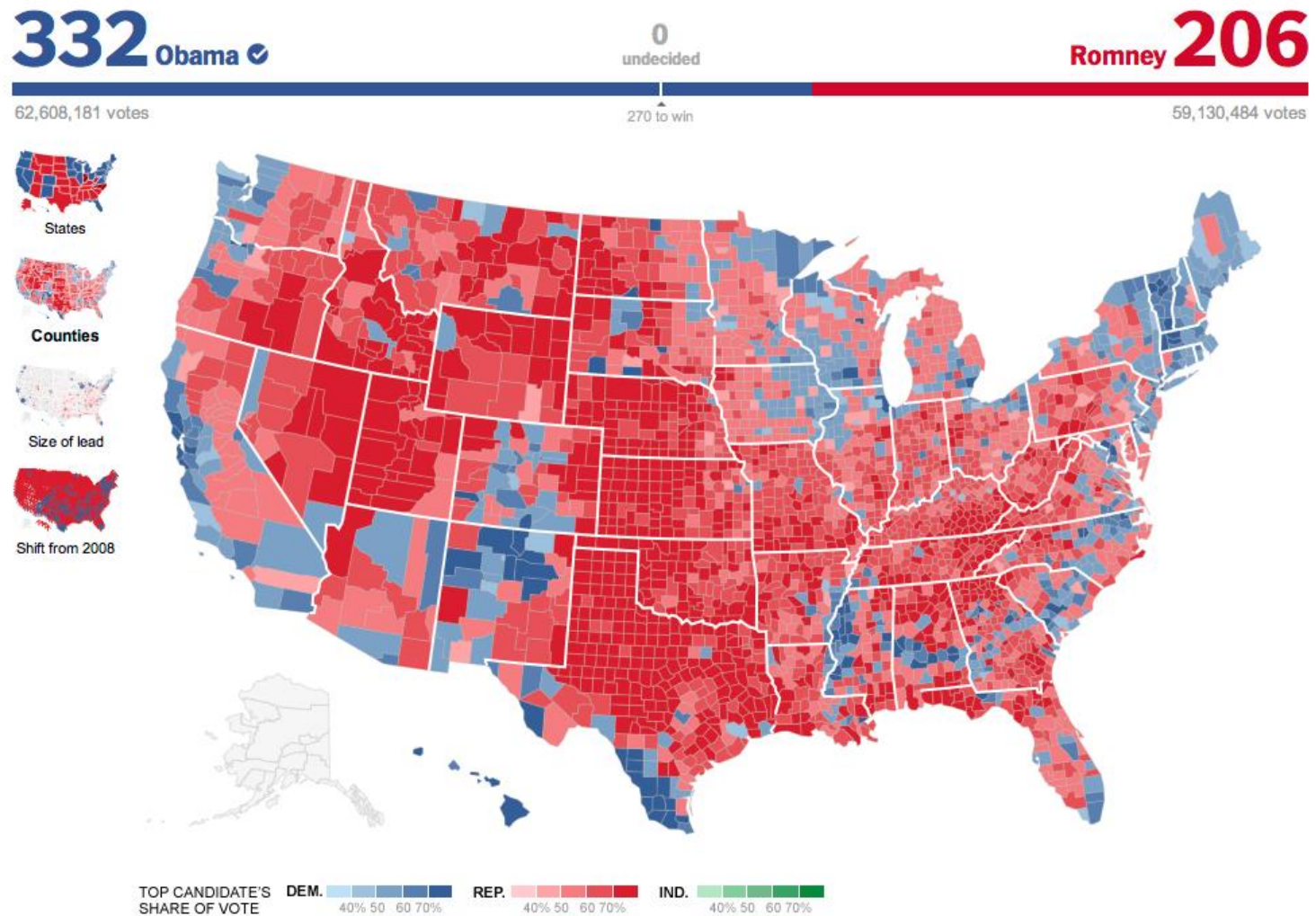
Issue 01: Geographic Units and Proportionality

Intended narrative: one party won the election with more votes.

According to your map who won the election? **RED** is dominant



Land does vote, people vote. PROBLEM: you are visualizing land.

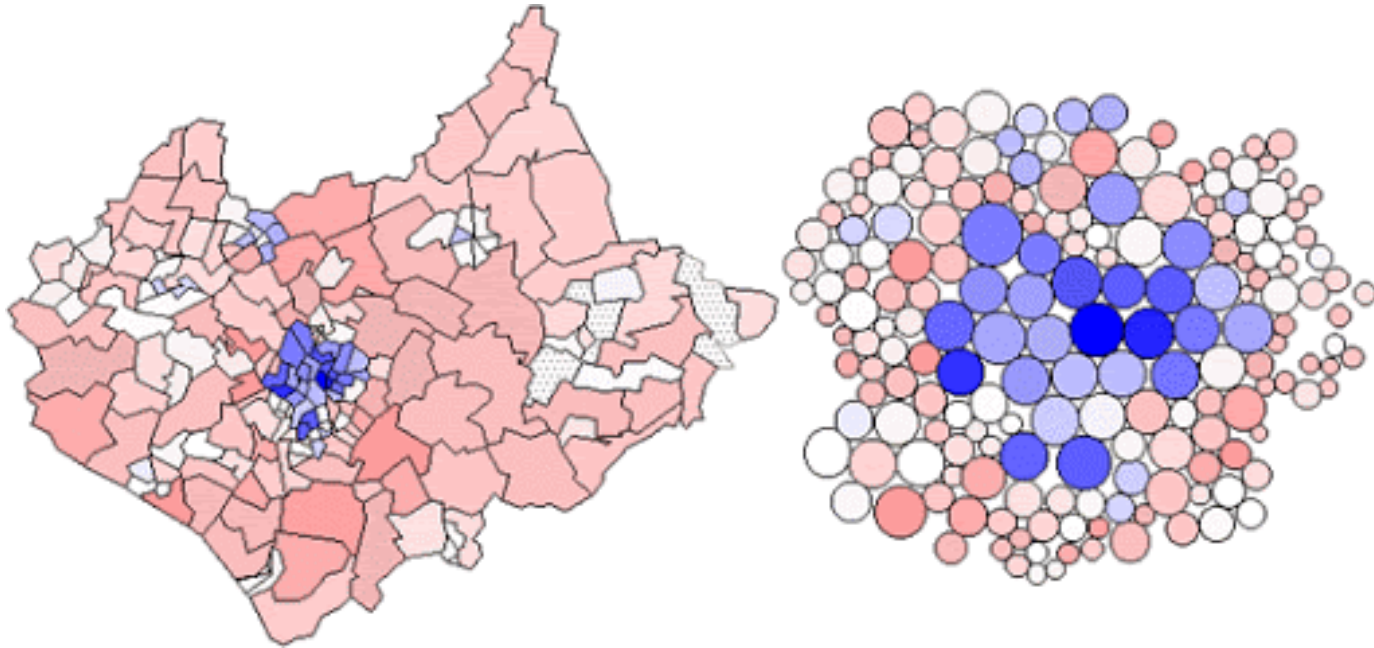


This is what a visualization of people looks like.



The problem with choropleth maps:

When population density varies spatially then simply coloring geographic units like census tracts is extremely misleading. They will over-emphasize people in rural areas and hide people in dense urban areas.



Census tracts are designed to contain approximately 2 to 4 thousand people. Thus the denser the neighborhood the smaller they are. The brain associates size with importance. The choropleth map on the left makes the blue group look like a small minority of all people and thus unimportant.

Dorling Cartograms present a more accurate narrative by **scaling geographies by population to correctly represent group proportionality**.

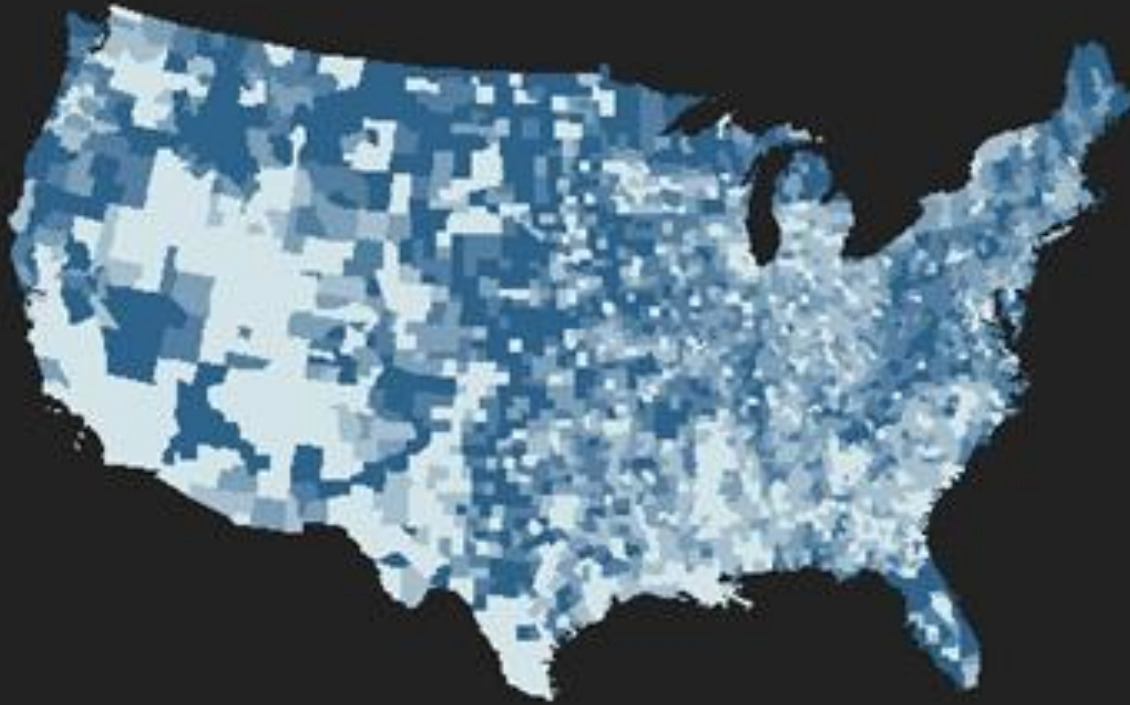
RESOLUTION: If you are presenting data about populations, make sure your data viz tells the story of the people and not the land.

Most choropleth maps using Census data are presenting a narrative about people, so make sure your visual narrative is not misleading!

Issue 02: Grouping Data

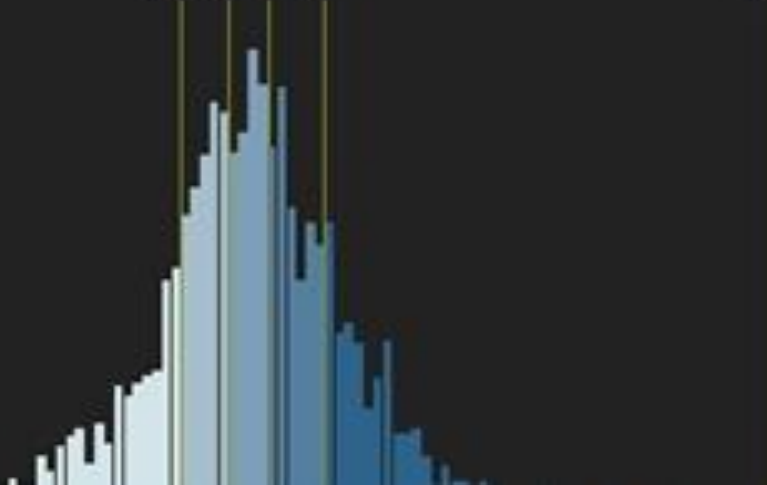
U.S. Census Bureau, 2000
MEDIAN AGE

Classification
QUANTILE



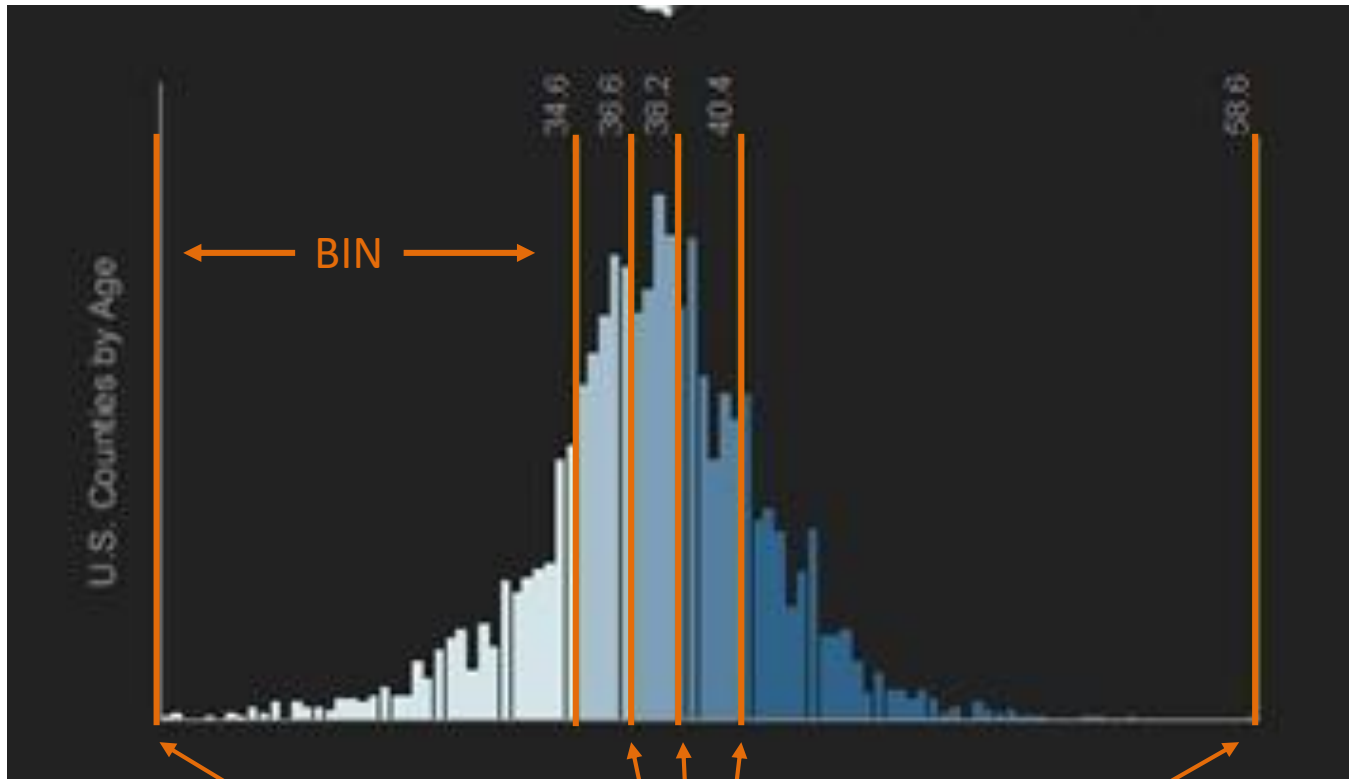
U.S. Counties by Age

34.6 36.6 38.2 40.4 58.6



BINNING DATA

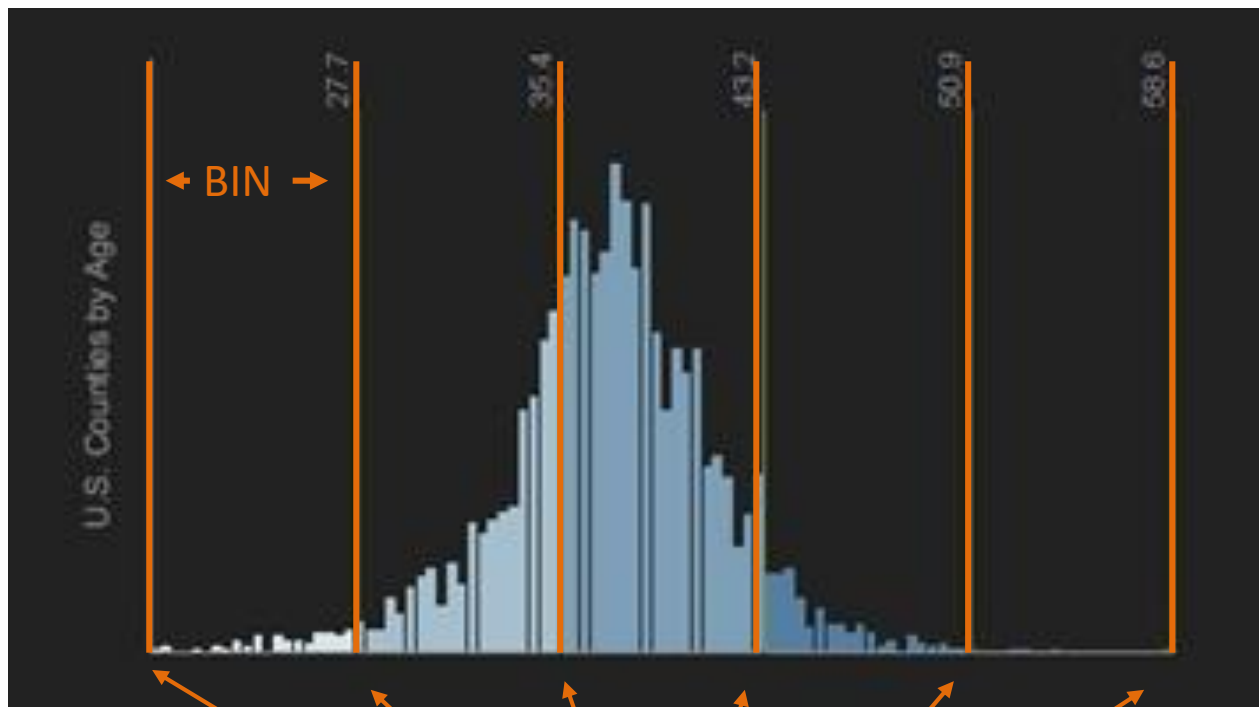
Binning data is the process of converting a continuous variable into a factor by selecting the number of groupings you would like displayed and then identifying corresponding cut points for the bins.



Cut points defined by quantile rule: each bin contains 20% of the data

```
# identify cut points  
bin.vals <- quantile( x, prob=c( 0,0.2,0.4,0.6,0.8,1 ) )  
cut( x, breaks=bin.vals )
```

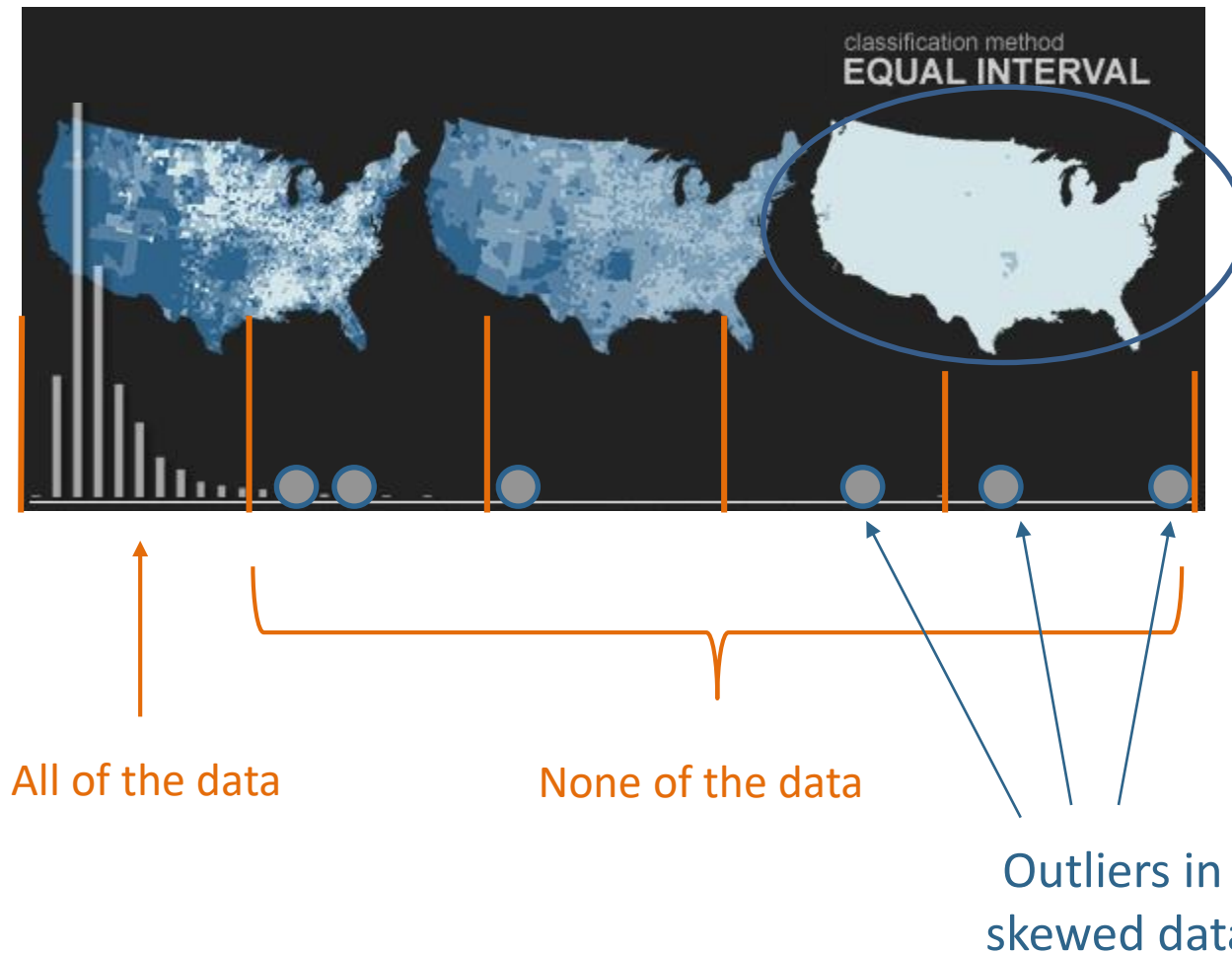
(these are percentiles)



Cut points with equal interval rule: each bin is equal size

```
# cut() determines cut points by  
cut( x, breaks=5 )
```

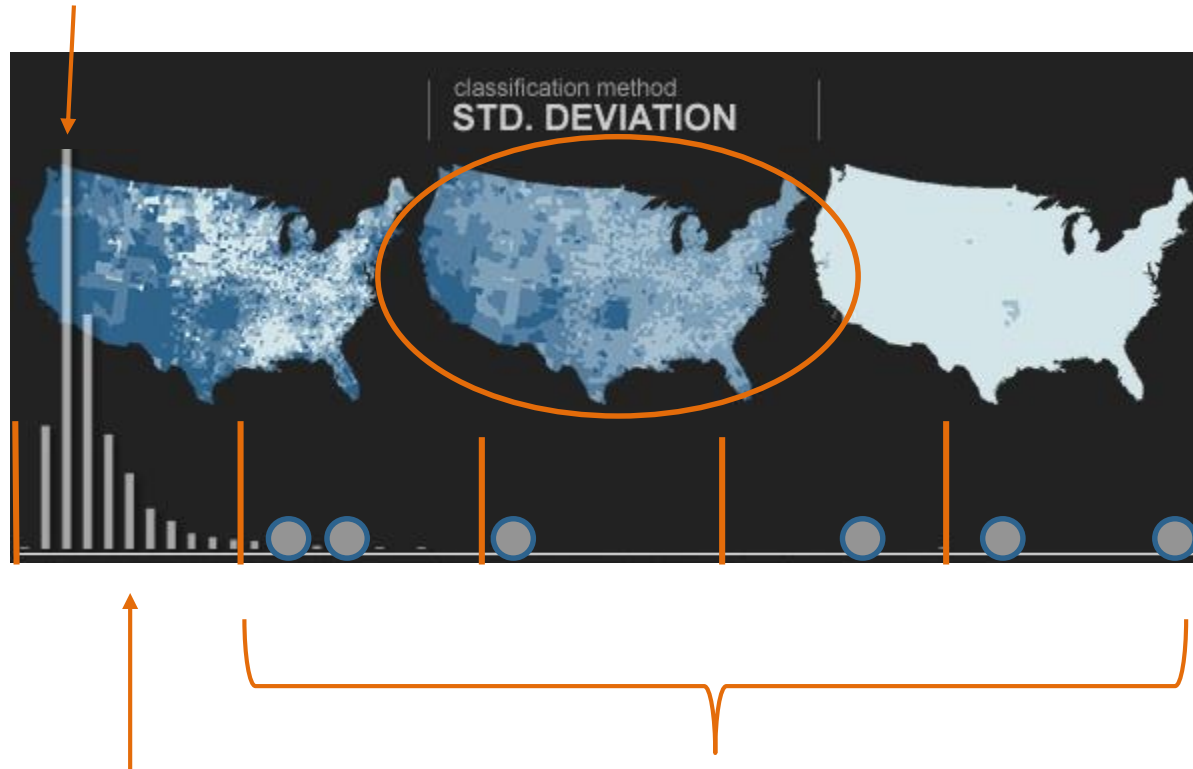
Hiding data with poor binning



No variation on the map – visual narrative is that everyone is the same across all geographies

Hiding data with poor binning

MODE



In skewed data
think about the
“average” case as
the modal case

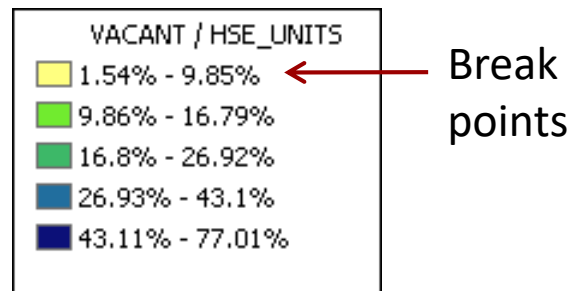
EMPHASIZE
THIS
VARIANCE

BINNING DATA

Process of placing data into groups (classes or bins) that have a similar characteristics or values

Break points

- Breaks the total attribute range up into these intervals
- Keep the number of intervals as small as possible (5-7)
- Use a mathematical progression or formula instead of picking arbitrary values



SELECTING MEANINGFUL BREAK POINTS

Quantiles

- Places the same number of data values in each class
- Will never have empty classes or classes with too few or too many values
- Attractive in that this method produces distinct map patterns
- Analysts use because they provide information about the shape of the distribution.
- Example: 0–25%, 25%–50%, 50%–75%, 75%–100%

SELECTING MEANINGFUL BREAK POINTS

Equal intervals

- Divides a set of attribute values into groups that contain an equal range of values
- Best communicates with continuous set of data
- Easy to accomplish and read
- Not good for clustered data
 - Produces map with many features in one or two classes and some classes with no features

SELECTING MEANINGFUL BREAK POINTS

Increasing interval widths

- Long-tailed distributions
- Data distributions deviate from a bell-shaped curve and most often are skewed to the right with the right tail elongated
- Example: Keep doubling the interval of each category, 0–5, 5–15, 15–35, 35–75 have interval widths of 5, 10, 20, and 40.

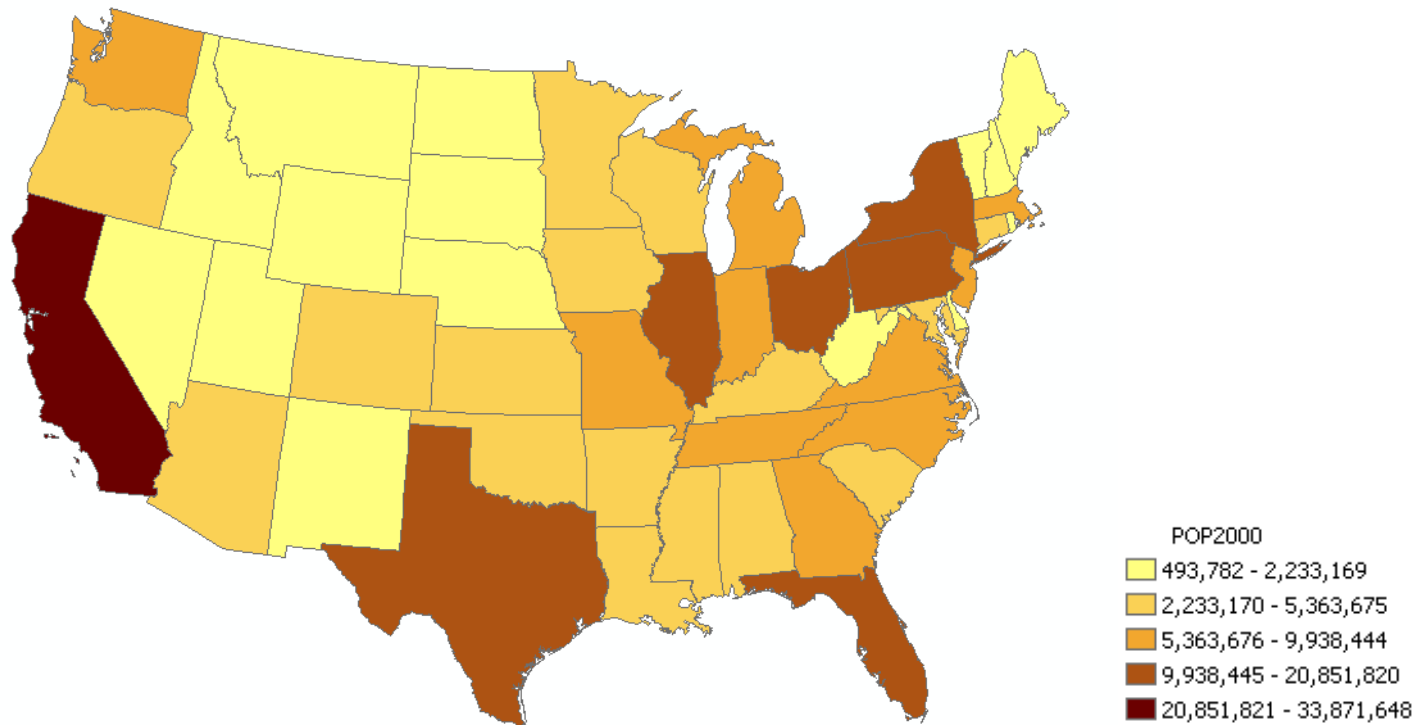
SELECTING MEANINGFUL BREAK POINTS

Exponential scales

- Popular method of increasing intervals
- Use break values that are powers such as 2^n or 3^n
- Generally start out with zero as an additional class if that value appears in your data
- Example: 0, 1–2, 3–4, 5–8, 9–16, and so forth

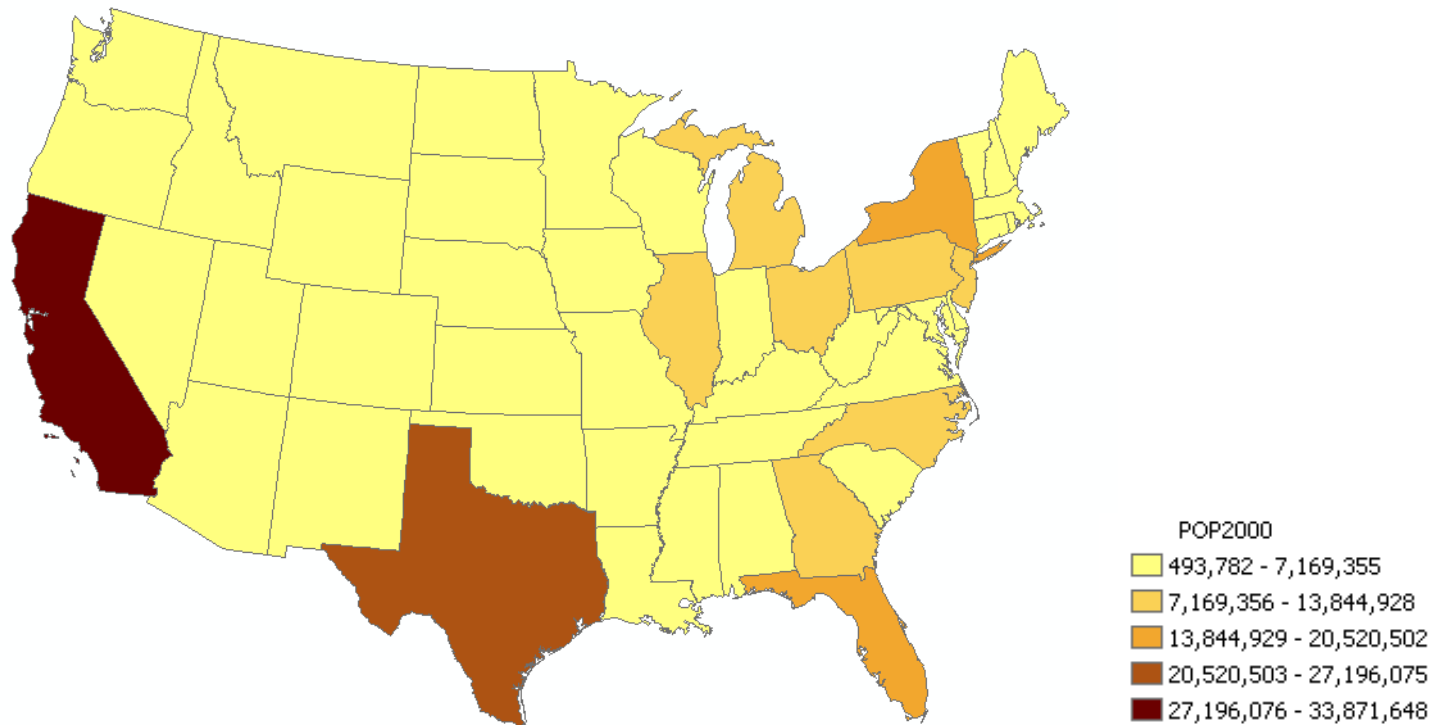
Original map (natural breaks)

U.S. population by state, 2000



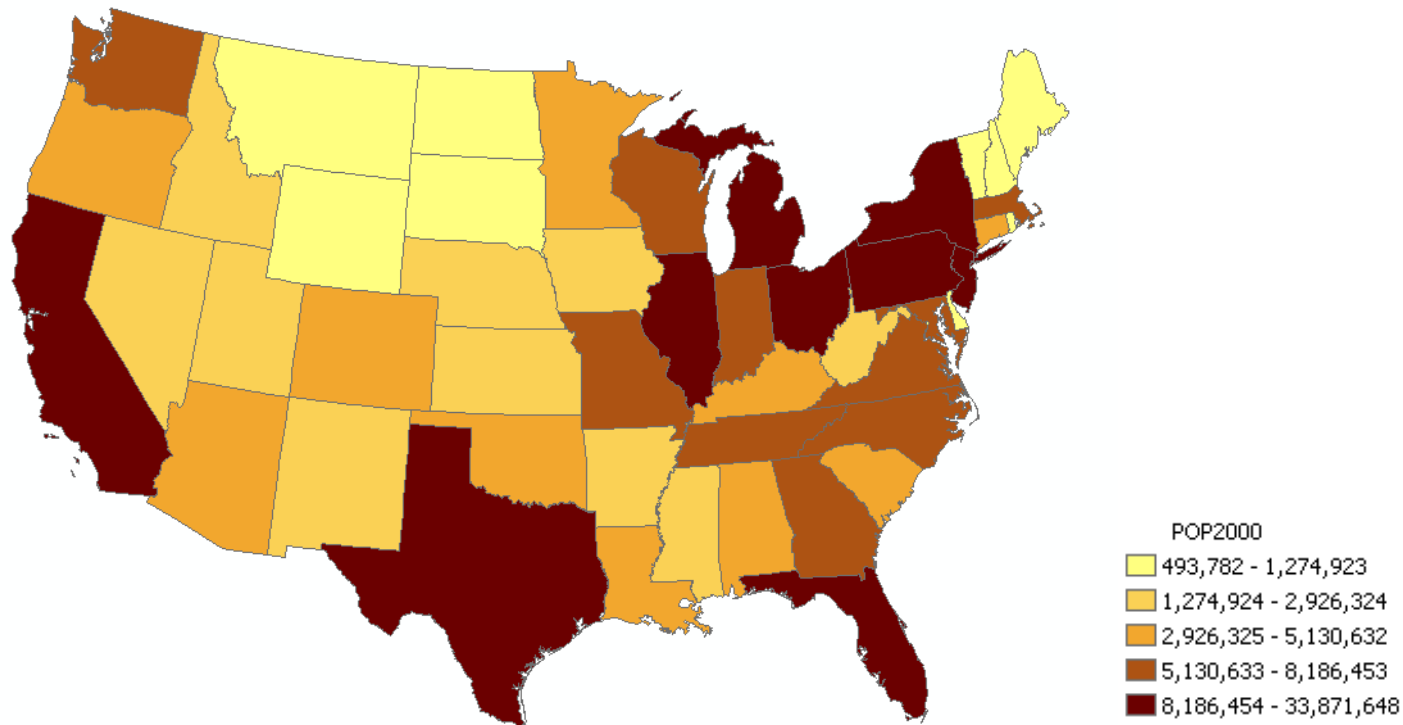
Equal interval scale

Not good because too many values fall into low classes



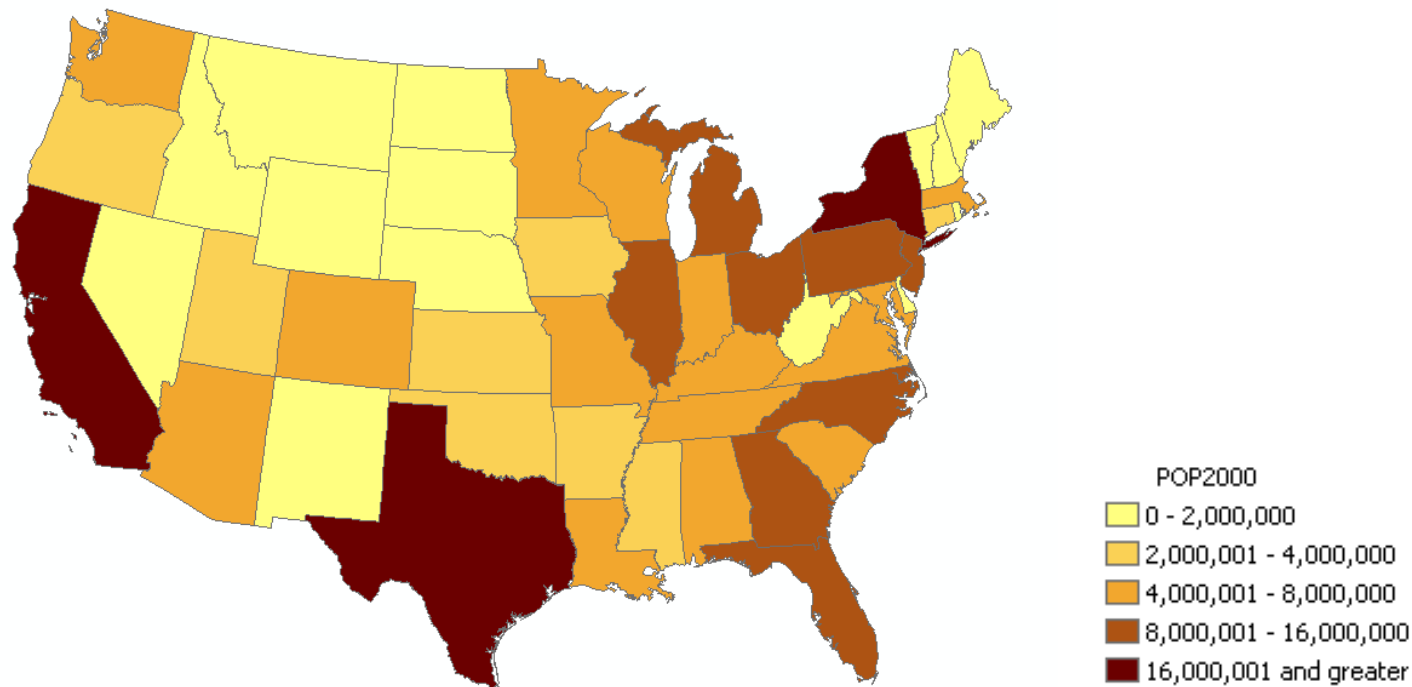
Quantile scale

Shows that an increasing width (geometric) scale is needed



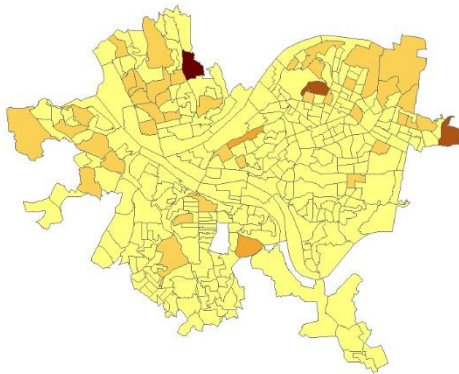
Custom geometric scale

- Experiment with exponential scales with powers of 2 or 3.

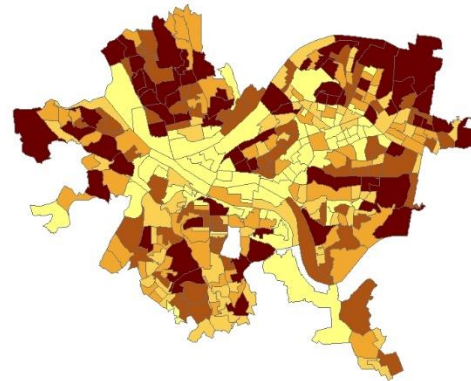


IMPLICATIONS OF BIN SELECTION

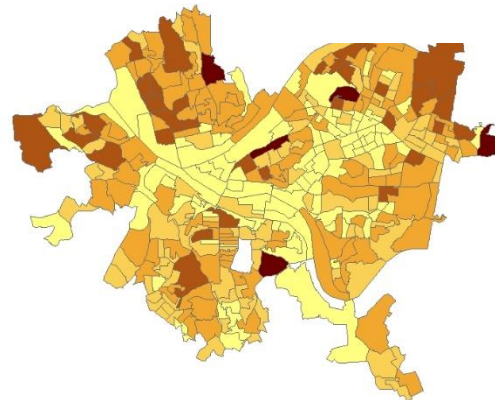
Equal Intervals



Quantiles

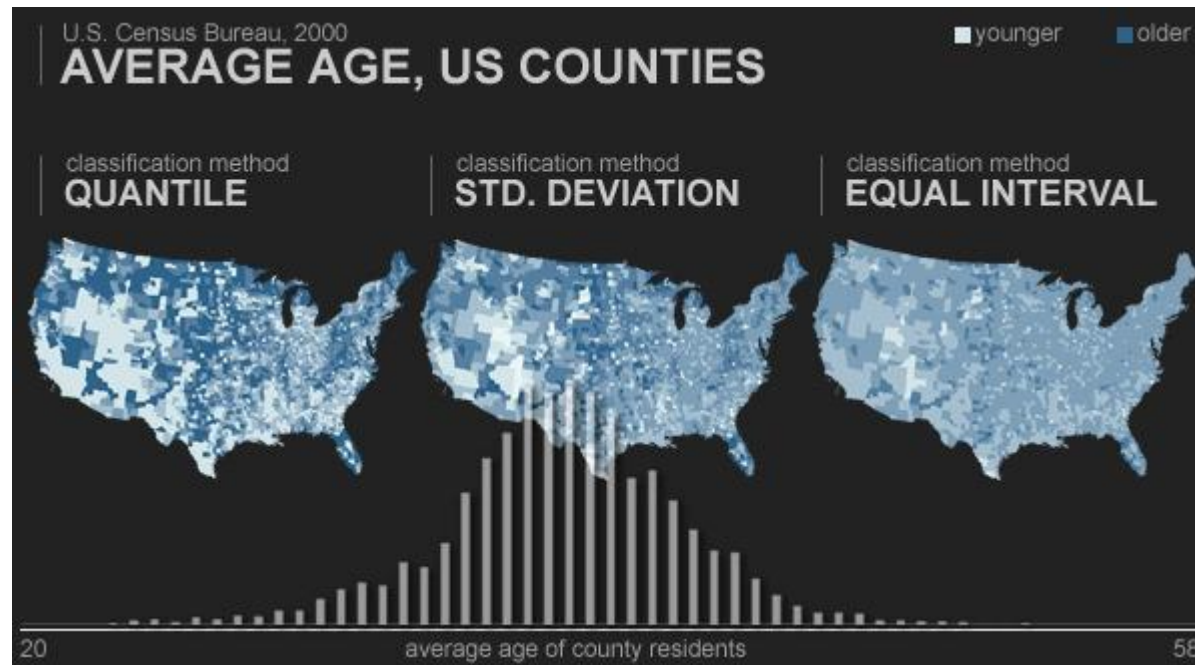


These maps are all made with the same data using different intervals for the break points.



Geometric

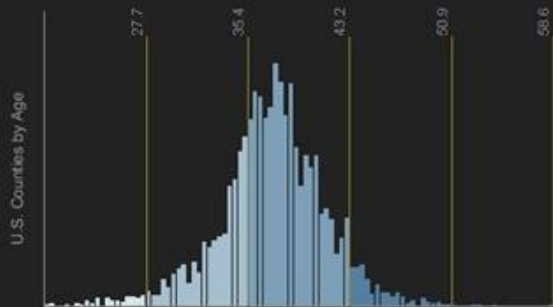
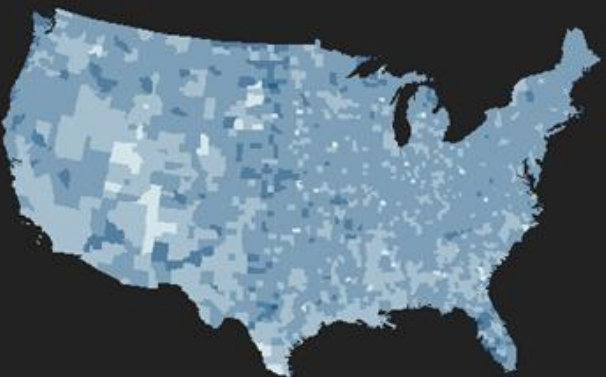
BREAK POINTS FOR NORMAL DISTRIBUTIONS



<http://uxblog.idvsolutions.com/2010/03/crazy-world-of-range-breaks.html>

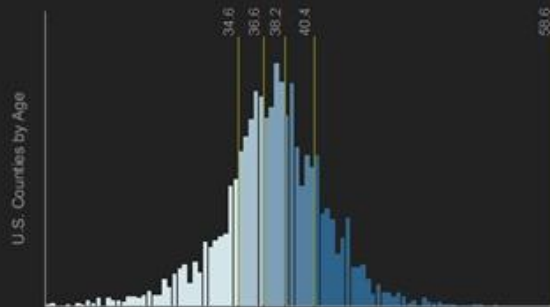
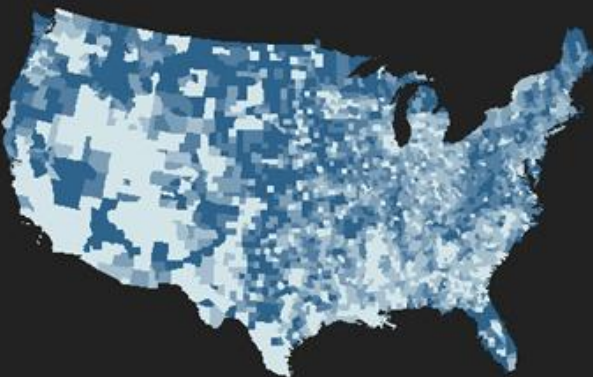
U.S. Census Bureau, 2000
MEDIAN AGE

classification
EQUAL INTERVAL



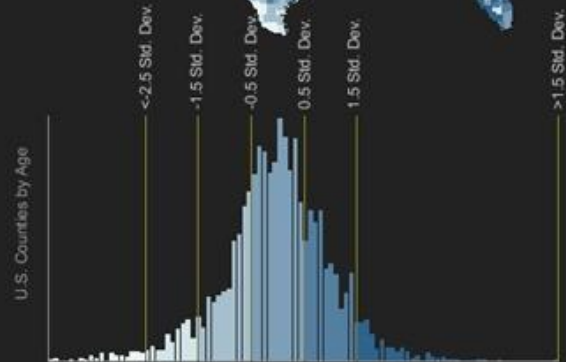
U.S. Census Bureau, 2000
MEDIAN AGE

classification
QUANTILE

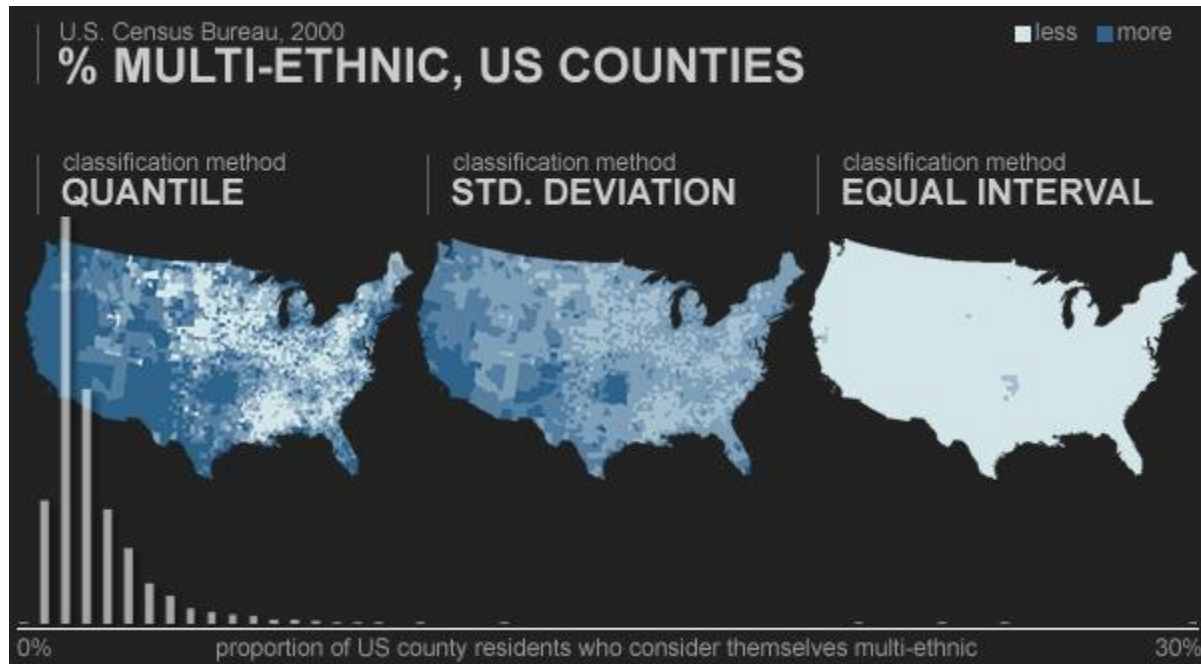


U.S. Census Bureau, 2000
MEDIAN AGE

classification
STD. DEVIATION



BREAK POINTS FOR SKEWED DISTRIBUTIONS



U.S. Census Bureau, 2000
% MULTI-ETHNIC

classification
EQUAL INTERVAL

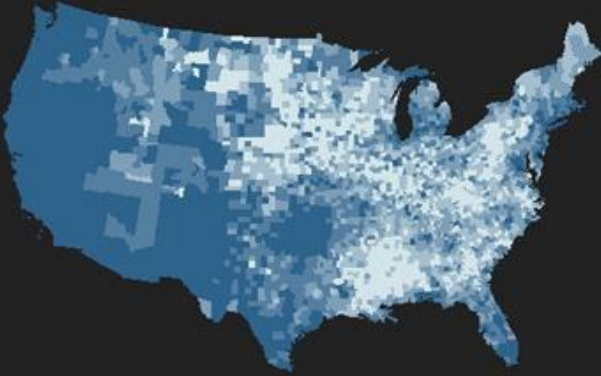


U.S. Counties by % Multi-Ethnic

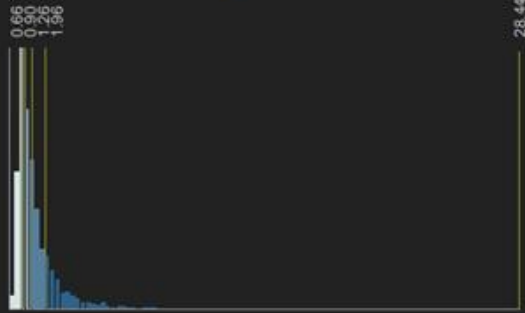


U.S. Census Bureau, 2000
% MULTI-ETHNIC

classification
QUANTILE

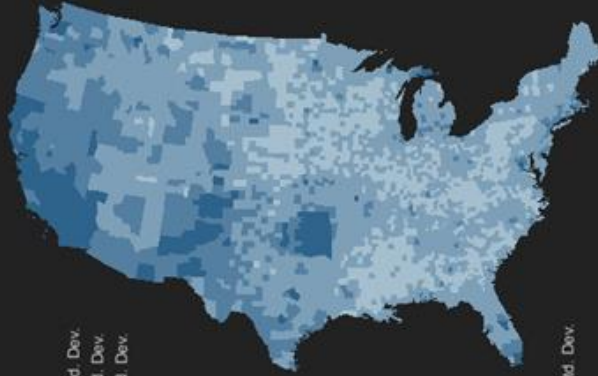


U.S. Counties by % Multi-Ethnic



U.S. Census Bureau, 2000
% MULTI-ETHNIC

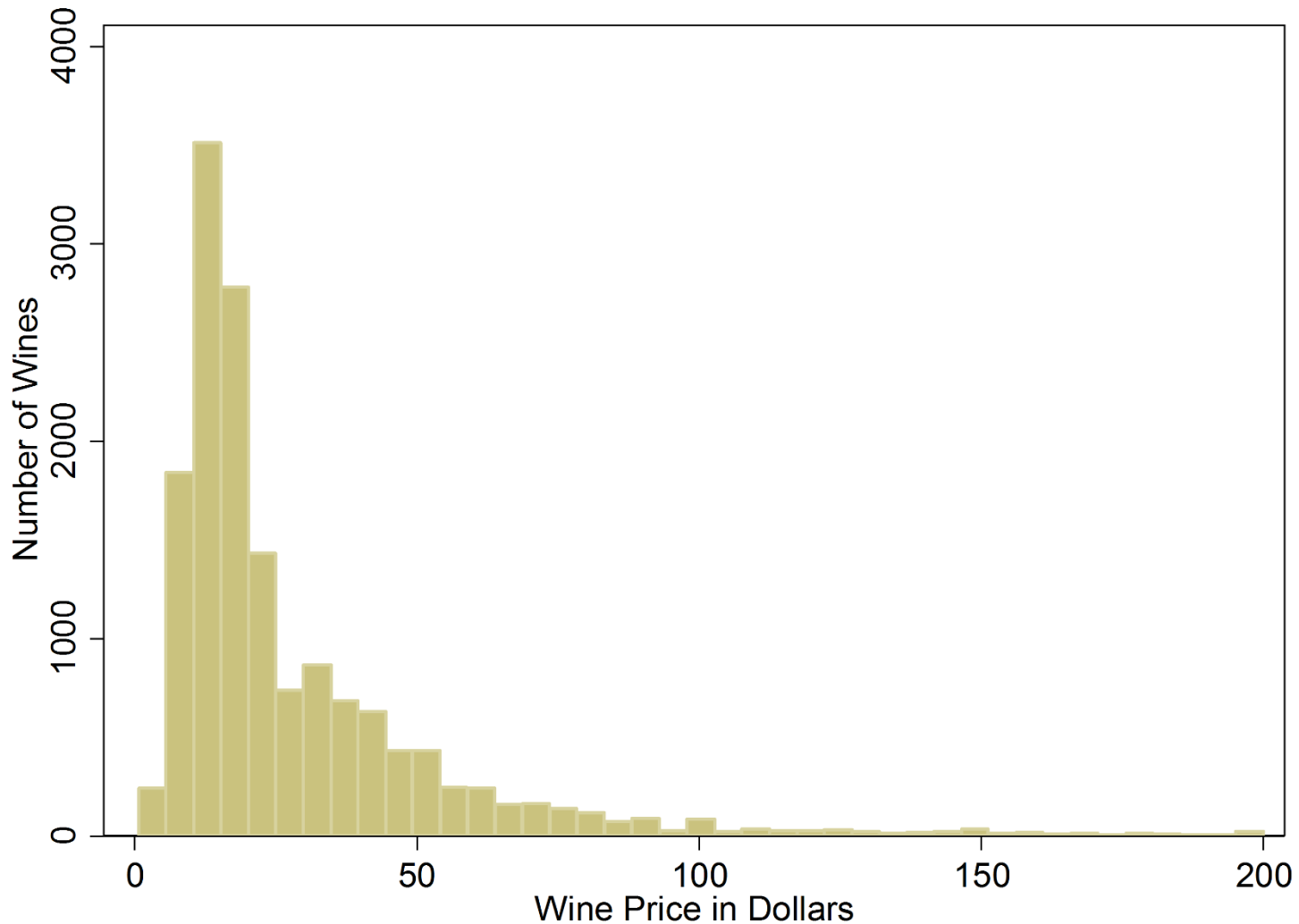
classification
STD. DEVIATION



U.S. Counties by % Multi-Ethnic



HIGHLY-SKEWED DISTRIBUTIONS:



HIGHLY-SKEWED DISTRIBUTIONS:

Think about skewed distributions like you think about classes of wine.

The first level of quality is a bottle for \$3, the next level is a bottle at \$6, the next level is at \$10-12, the next level at \$25, the next level at \$50, then \$100.

To move up one class, you basically double the price.

If you want a scale that translates wine prices into class, it would look something like:

\$0 - \$4	Cheap wine
\$4 - \$8	Low quality
\$8 - \$20	Medium quality
\$20 - \$80	High quality
\$80 +	Excellent quality

RULES OF THUMB FOR SELECTING BINS:

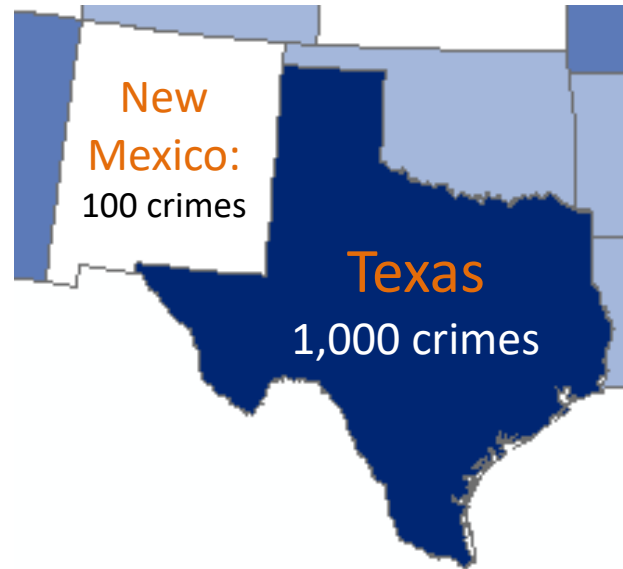
1. **First Rule:** use common sense! What do your groups represent, and are they meaningful? Are you misleading your audience with unreasonable breaks?
2. Binning by quantiles is typically a safe way to create breaks to show low, medium, and high values.
3. If a lot of your data is bunched together (for example, half of your values are close to zero), quantiles will not be meaningful because it will imply differences that do not exist.
4. If your distribution is skewed, consider increasing-interval or exponential scales.

For example, define the first group as 0-2, second as 2-6, third as 6-14, next as 14-30 (your interval size doubles each break).

Issue 03:

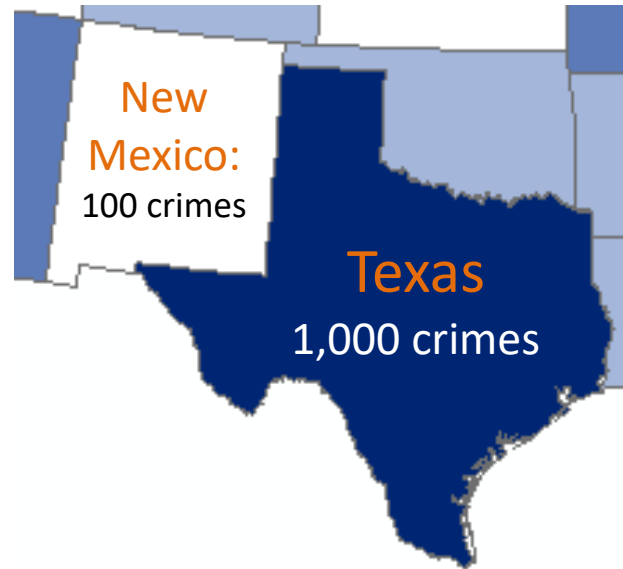
COUNTS of things
versus
RATES of things

NORMALIZING DATA



I want to compare crime rates in Texas and New Mexico.

If they are 10 times higher in Texas, does that mean its 10 times more violent / crime-ridden?



NEW MEXICO

Population: 2 million

Crime: 5 per 100k

TEXAS

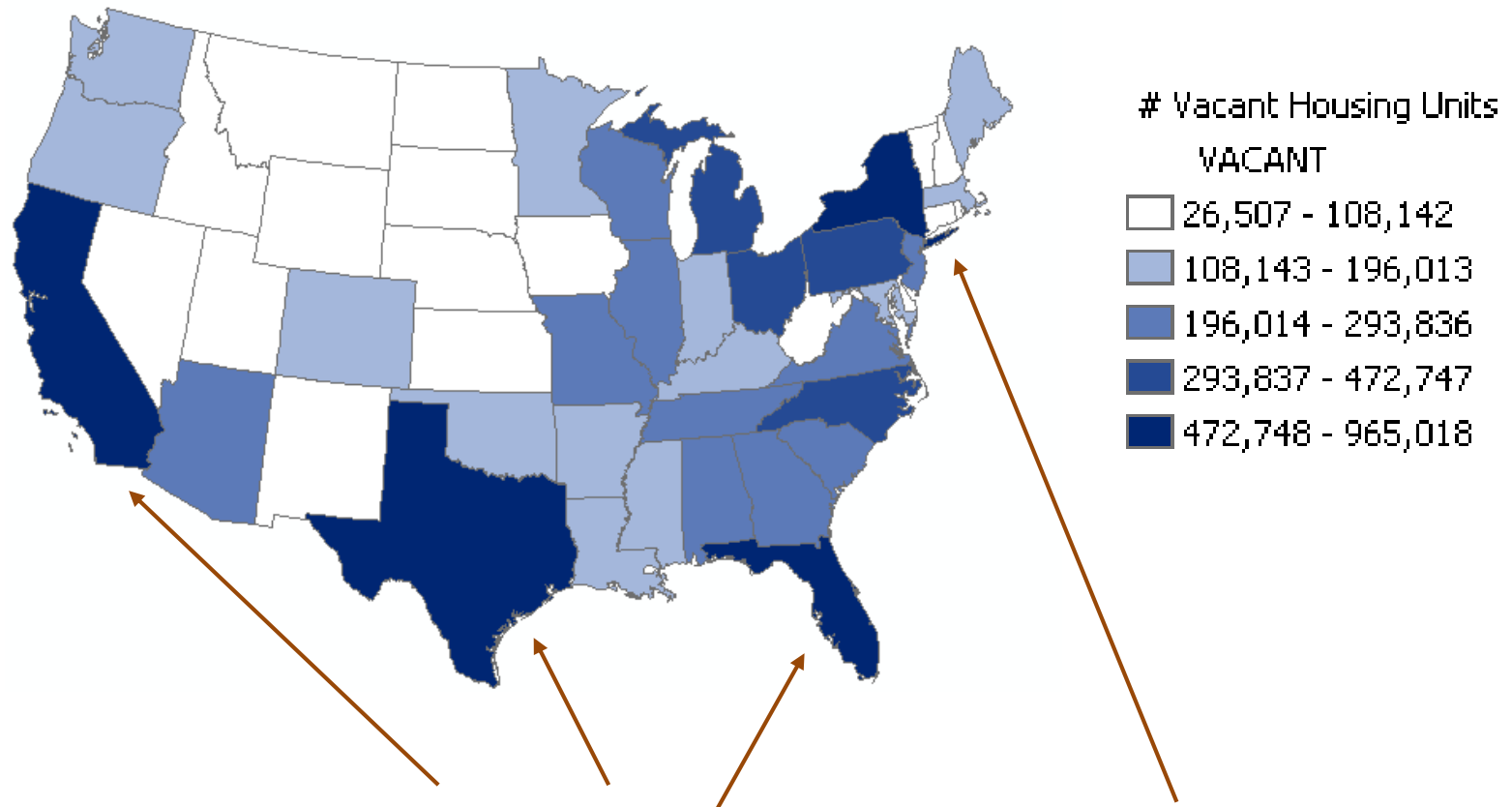
Population: 29 million

Crime: 3.5 per 100k

NORMALIZED TO POPULATION SIZE: CRIME RATE
(BETTER APPLES TO APPLES COMPARISON)

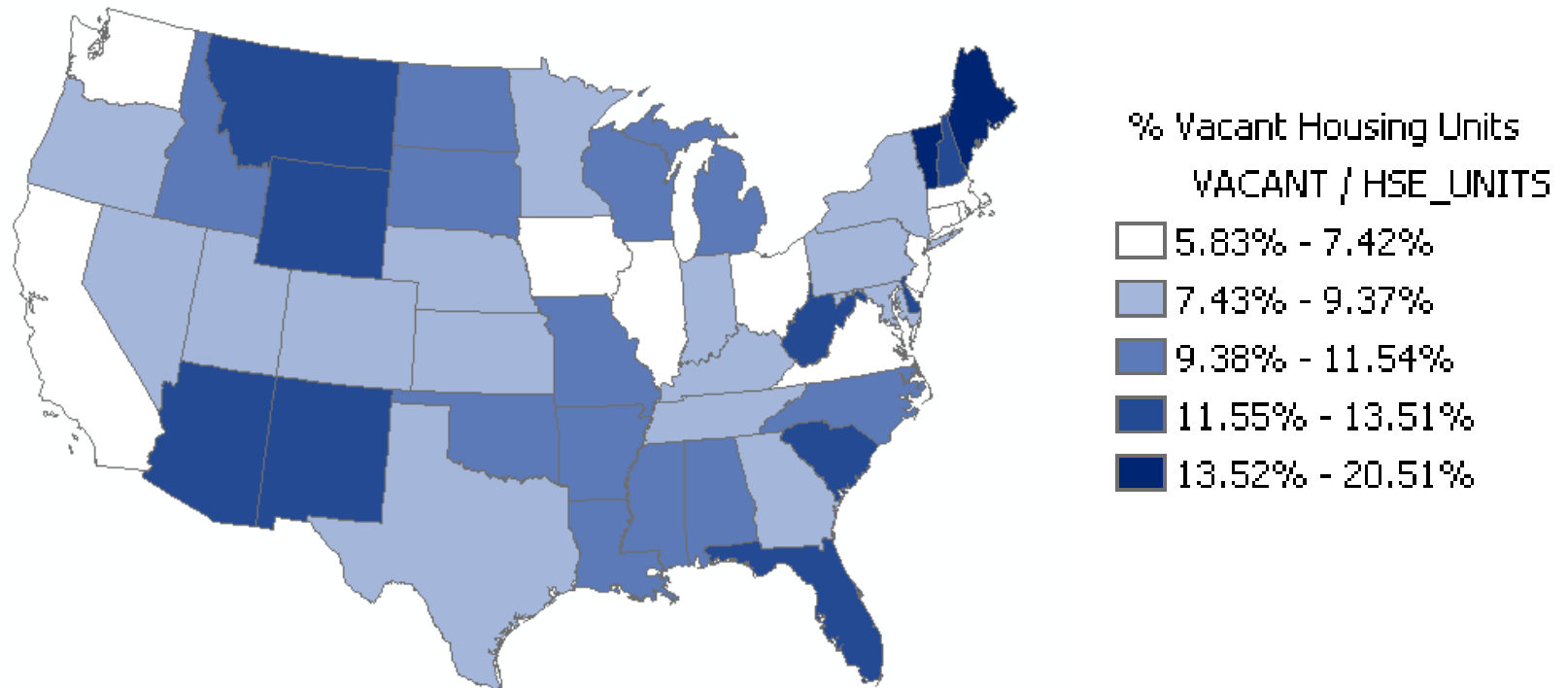
RAW DATA

NUMBER of vacant housing units by state, 2000



NORMALIZED DATA

Percentage vacant housing units by state, 2000



Now we can see RATES OF VACANCY independent of population size

NORMALIZING DATA

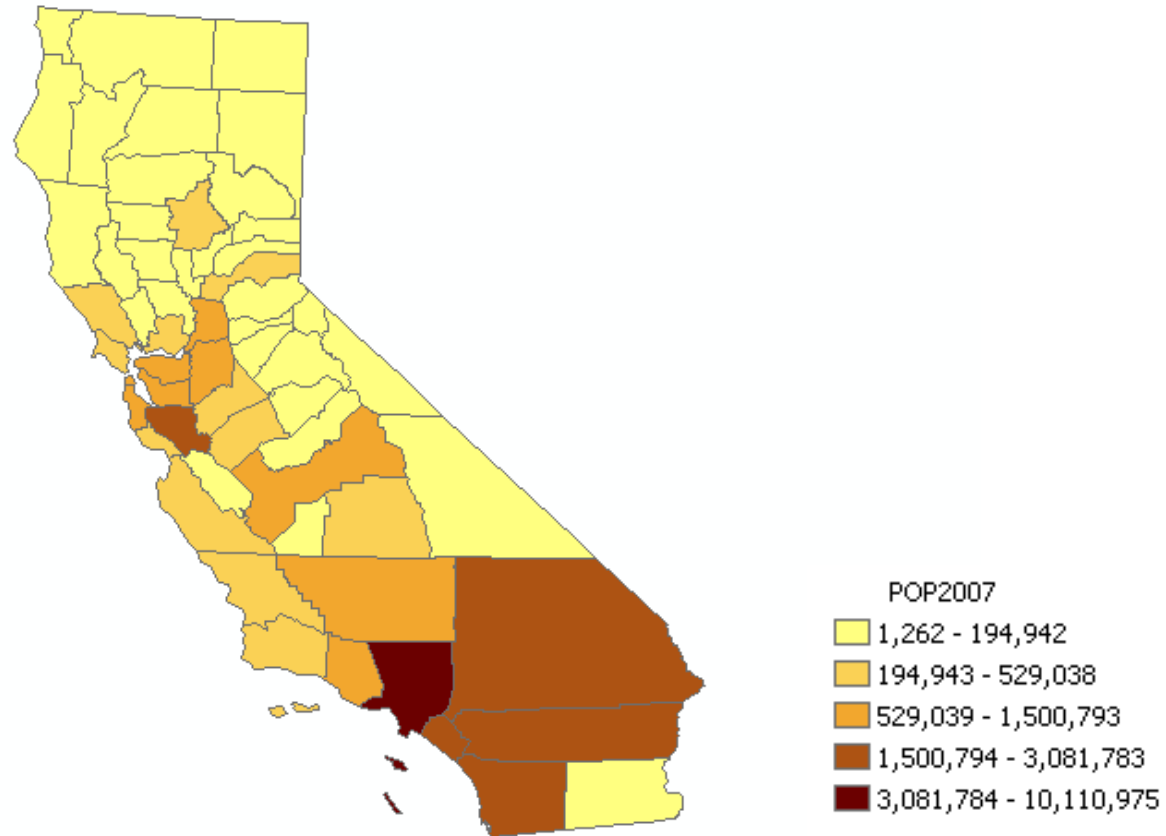
Divides one numeric attribute by another in order to minimize differences in values based on the size of areas or number of features in each area

Examples:

- Dividing the number of vacant housing units by the total number of housing units yields the percentage of vacant units
- Dividing the population by area of the feature yields a population density

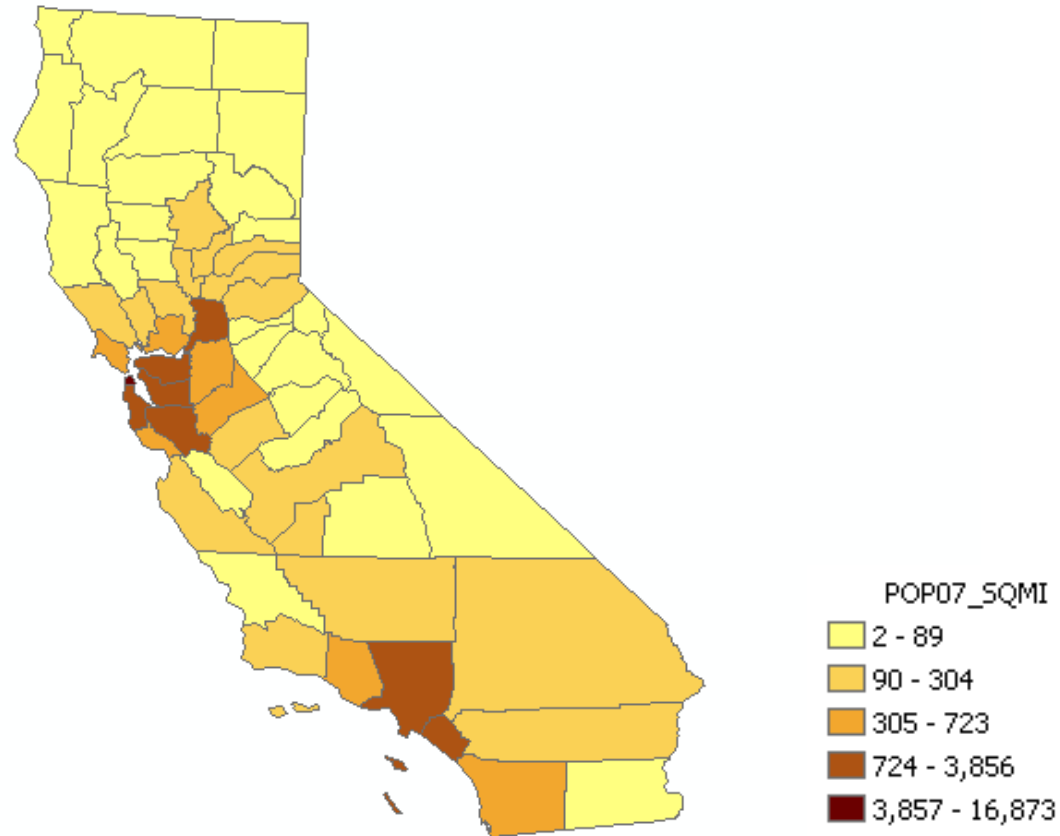
NON-NORMALIZED DATA

California POPULATION BY COUNTY



NORMALIZED BY GEOGRAPHY

California POPULATION DENSITY



Issue 04:

Low to High Scale

versus

Average / Not Average


SELECTING COLORS

If you have naturally categorical data you select colors to maximize contrast between categories so membership is clear.

If you have numeric data you need to decide:

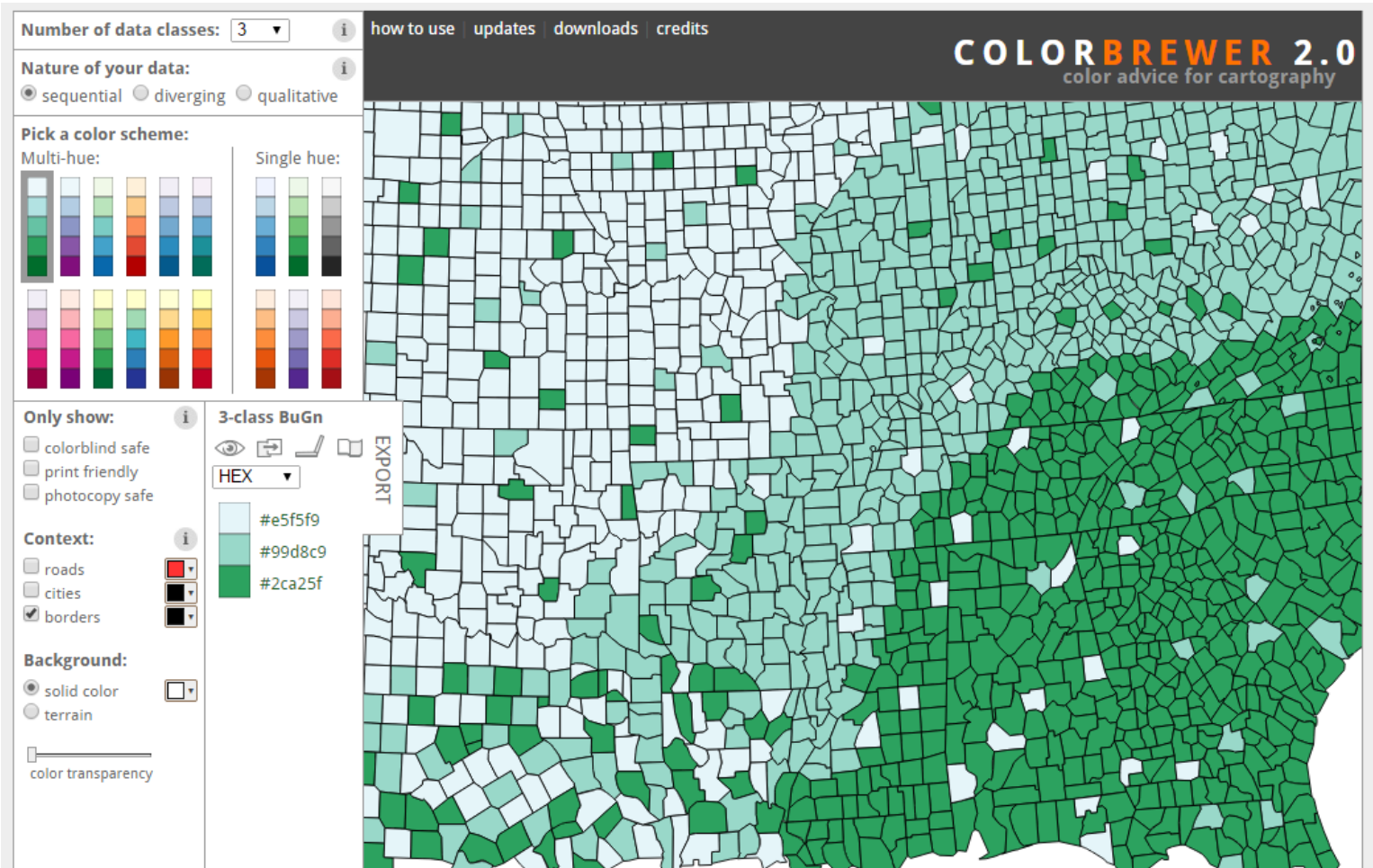
Is my narrative about the low or high performers? Select a **SEQUENTIAL** (light to dark) scale.

Or is my narrative about the non-average cases (both low and high groups)? Select a **DIVERGENT** scale (dark-light-dark).

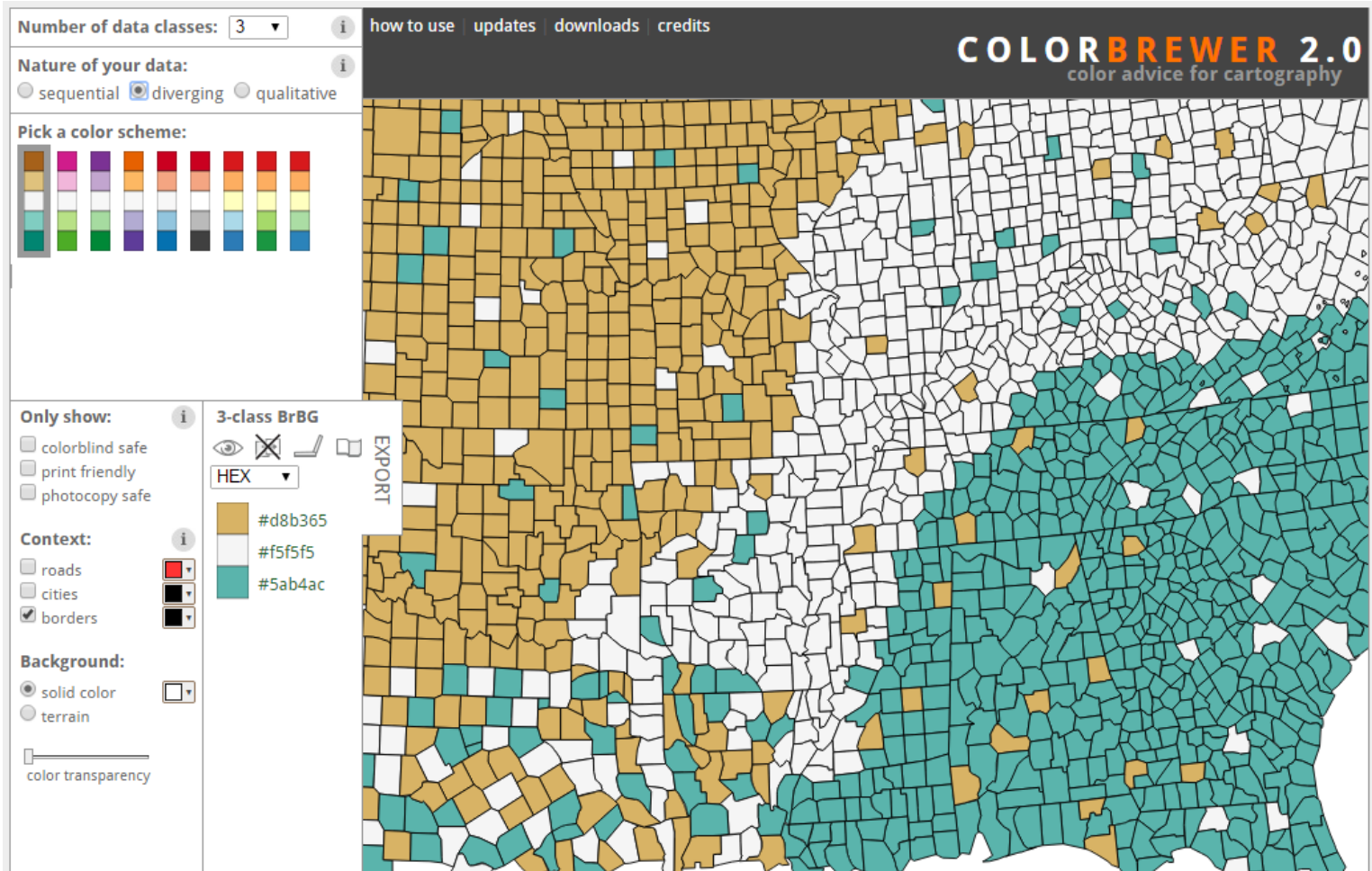


Comparison
group is
always light

SEQUENTIAL SCALE



DIVERGENT SCALE

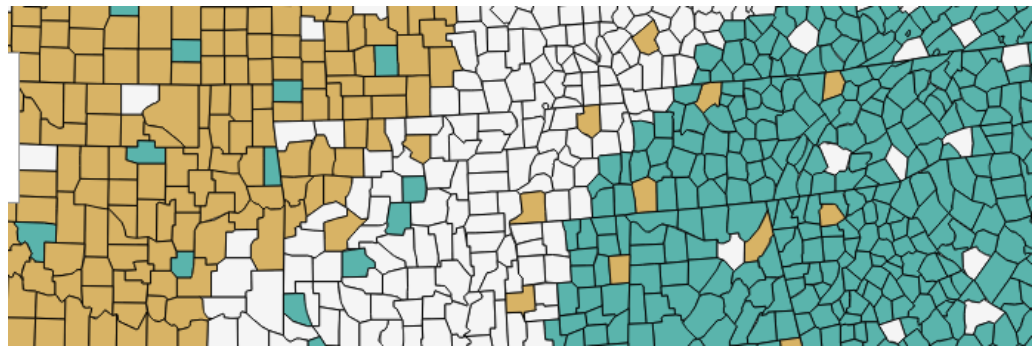


RULES OF THUMB:

1. If you are highlighting a class of individuals, like those in poverty, a single color might be sufficient (FE red for in poverty, gray for not).
2. When you are highlighting performance along a single dimension, use a sequential scale with white or gray at the bottom of your color scale and a dark color at the top.
3. If your comparison is relative to the average put white or gray in the middle to represent “average.” Consider red for low, blue or green for high (depending on the context if high is good, low is bad).
4. Five to seven categories is typical.

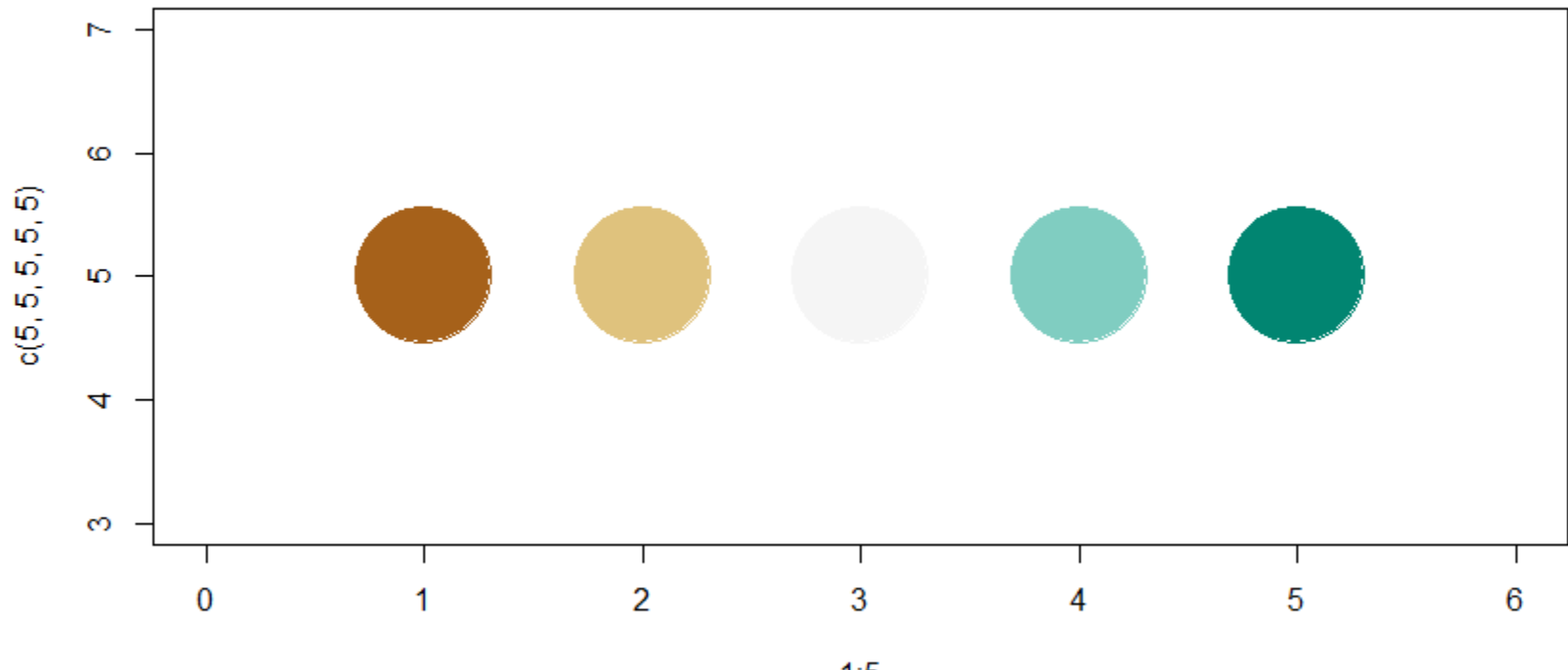
COLOR FUNCTIONS

<http://colorbrewer2.org/>



MANUALLY:

```
color.vals <- c("#a6611a", "#dfc27d", "#f5f5f5", "#80cdc1", "#018571" )  
plot( 1:5, c(5,5,5,5,5), col=color.vals, pch=19, cex=10, xlim=c(0,6) )
```



```
library( RColorBrewer )
```

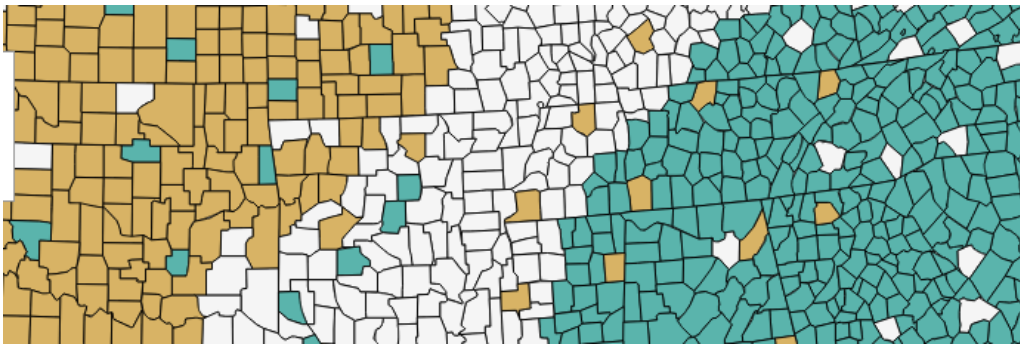
```
display.brewer.pal( 7, "BrBG" )
```

```
display.brewer.pal( 5, "BrBG" )
```

```
display.brewer.pal( 7, "BuGn" )
```

```
brewer.pal( 5, "BrBG" ) # identical to previous slide
```

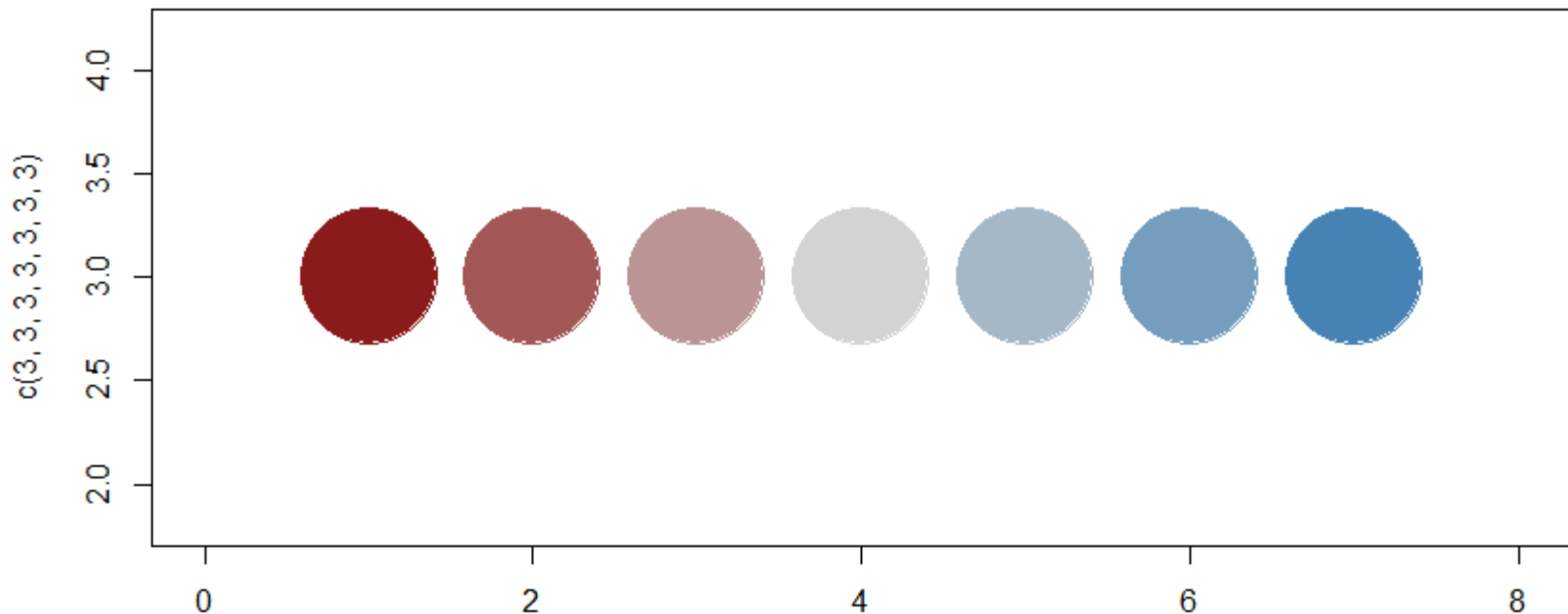
```
color.vals <- brewer.pal( 5, "BrBG" )
```



CREATE YOUR OWN COLOR RAMP:

Set center and end colors

```
color.function <- colorRampPalette( c("firebrick4","light gray","steel blue" ) )  
color.function(5) # number of classes you desire  
col.vals <- color.function(7)  
plot( 1:7, c(3,3,3,3,3,3,3), pch=19, cex=10, col=col.vals, xlim=c(0,8) )
```



IDENTIFY BREAK POINTS FOR QUANTILES:

```
norm.vec <- rnorm(10000) + 10
```

```
# Five groups, 20% of the data in each
```

```
quantile( norm.vec, probs=c(0, 0.20, 0.40, 0.60, 0.80, 1 ), na.rm=T )
```

```
# Seven groups
```

```
quantile( norm.vec, probs=seq( from=0, to=1, by=1/7 ), na.rm=T )
```

```
bin.values <- quantile( norm.vec, probs=seq( from=0, to=1, by=1/7 ) )
```

```
cut( norm.vec, breaks=bin.values ) # assigns data to bins
```

```
col.vals <- color.function(7)
```

```
cut( norm.vec, breaks=bin.values, labels=col.vals ) # adds color labels
```

Issue 05:

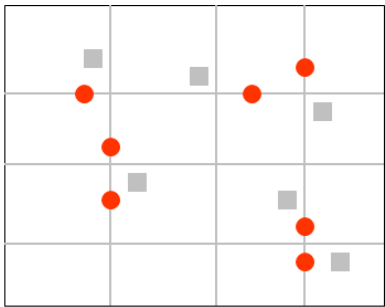
Don't hide your subject

GRAPHICAL HIERARCHY

GROUND AND FIGURE

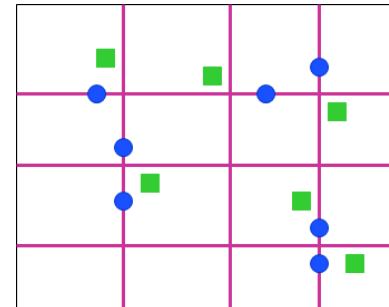
- Assign bright colors (red, orange, yellow, green, blue) to important graphic elements
- Features are known as **figure**, or the **subject** of the map.
- Assign drab colors to the graphic elements that provide orientation or context, especially shades of gray
- Features known as **ground**, or the **context** for the subject.

GOOD



Circles in figure, squares and lines in ground

BAD



All features in figure