Data Classification

Data Classification methods

- A ranged thematic map uses shades of color to represent *ranges of data values*.
- How are the ranges determined? All data classification methods revolve around two essential questions:
  - "How many data ranges should there be?" and,
  - “Where does each range begin and end?”
- The answers to these questions are determined in part by the classification method selected

Some common data classification methods are:

- EQUAL-COUNT
- EQUAL RANGES
- NATURAL BREAKS
- STANDARD DEVIATION
- CUSTOM

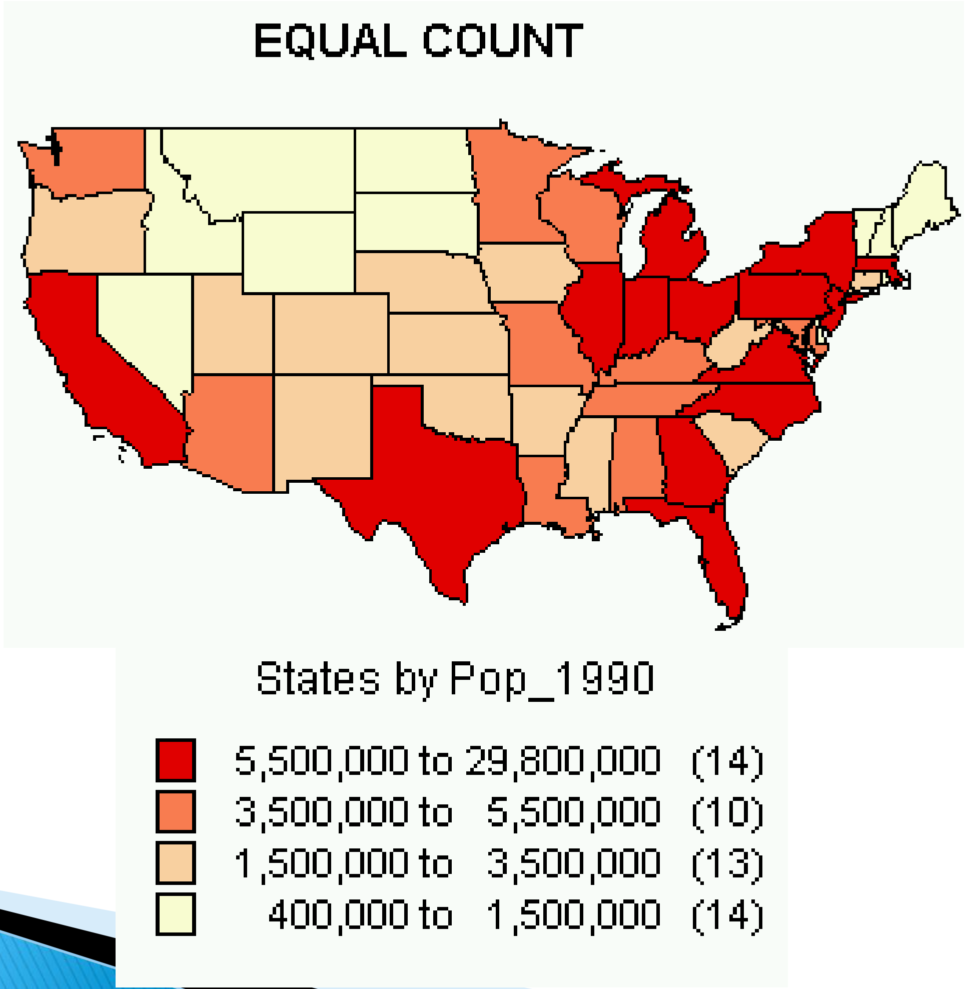
Equal Count

- Equal Count has the same number of records in each range. If you want to group 100 records into 4 ranges using Equal Count, the Equal Count method computes the ranges so that approximately 25 records fall into each range, depending on the rounding factor you set.

$$\text{number of observations per class} = \frac{\text{total ovservations}}{\text{number of classes}}$$

Equal-Count

- classification is to have an equal number of cases in each range.


EQUAL COUNT
States by Pop_1990
🟥 5,500,000 to 29,800,000 (14)
🟧 3,500,000 to 5,500,000 (10)
▨ 1,500,000 to 3,500,000 (13)
□ 400,000 to 1,500,000 (14)

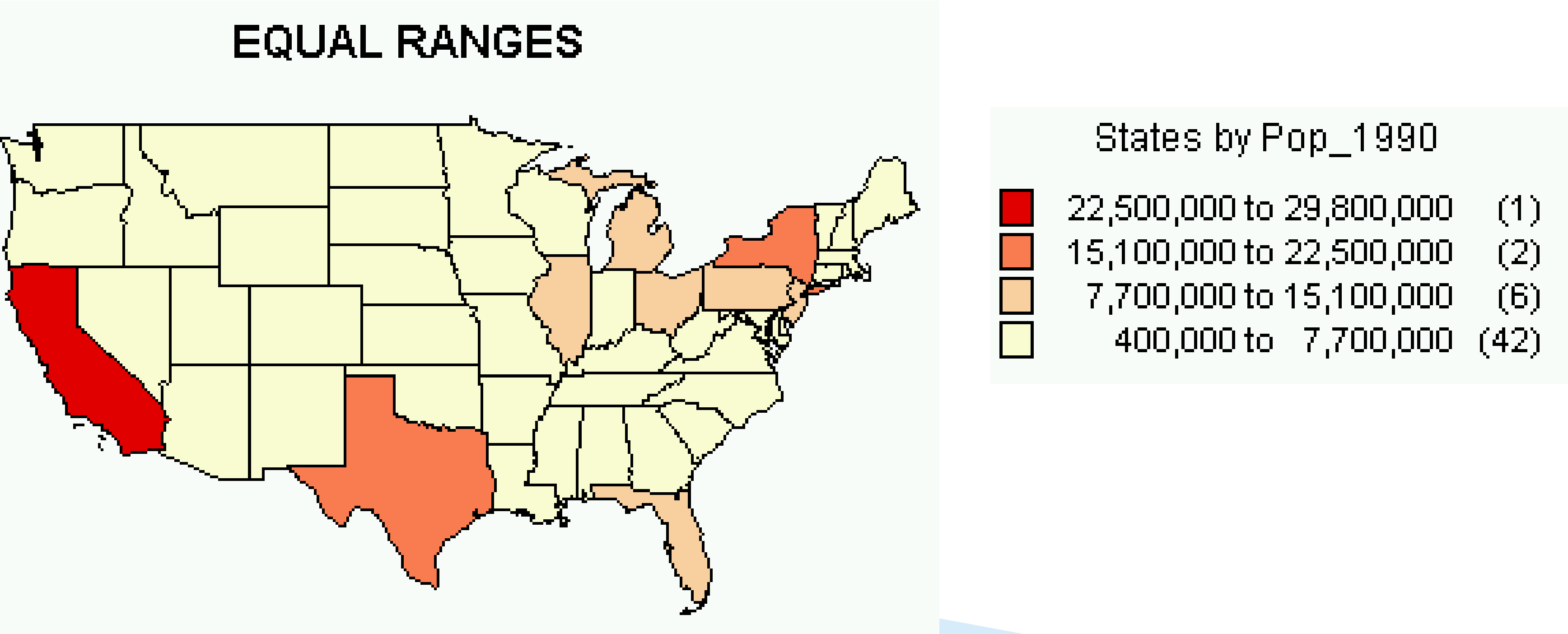
Equal Ranges

- Equal Ranges divides records across ranges of equal size.
- For example, you have a field in your table with data values ranging from 1 to 100. You want to create a thematic map with four equal size ranges. The Equal Ranges method produces ranges 1-25, 25-50, 50-75, and 75-100.
- Keep in mind that the Equal Ranges method may create ranges with no data records, depending on the distribution of your data.

$$\frac{\text{range of data}}{\text{number of classes}} = \frac{\text{(highest value - lowest value)}}{\text{number of classes}}$$

Equal Range

- The object of *Equal Range* classification is to have equal-sized ranges.
- In the example, every range has an interval of about 7 million.
- This method does not always reflect the data very well.
- Notice that the lowest range has 42 cases and the highest range has only 1 case, making for a very "unbalanced" map.


EQUAL RANGES
States by Pop_1990
🟥 22,500,000 to 29,800,000 (1)
🟧 15,100,000 to 22,500,000 (2)
▨ 7,700,000 to 15,100,000 (6)
□ 400,000 to 7,700,000 (42)

Natural Breaks classification

- The object of *Natural Breaks* classification is to create ranges based on clusters or gaps within the data itself. This makes for ranges that reflect the data very well.
- Note however, that the number of cases per range and the size of ranges can vary considerably.

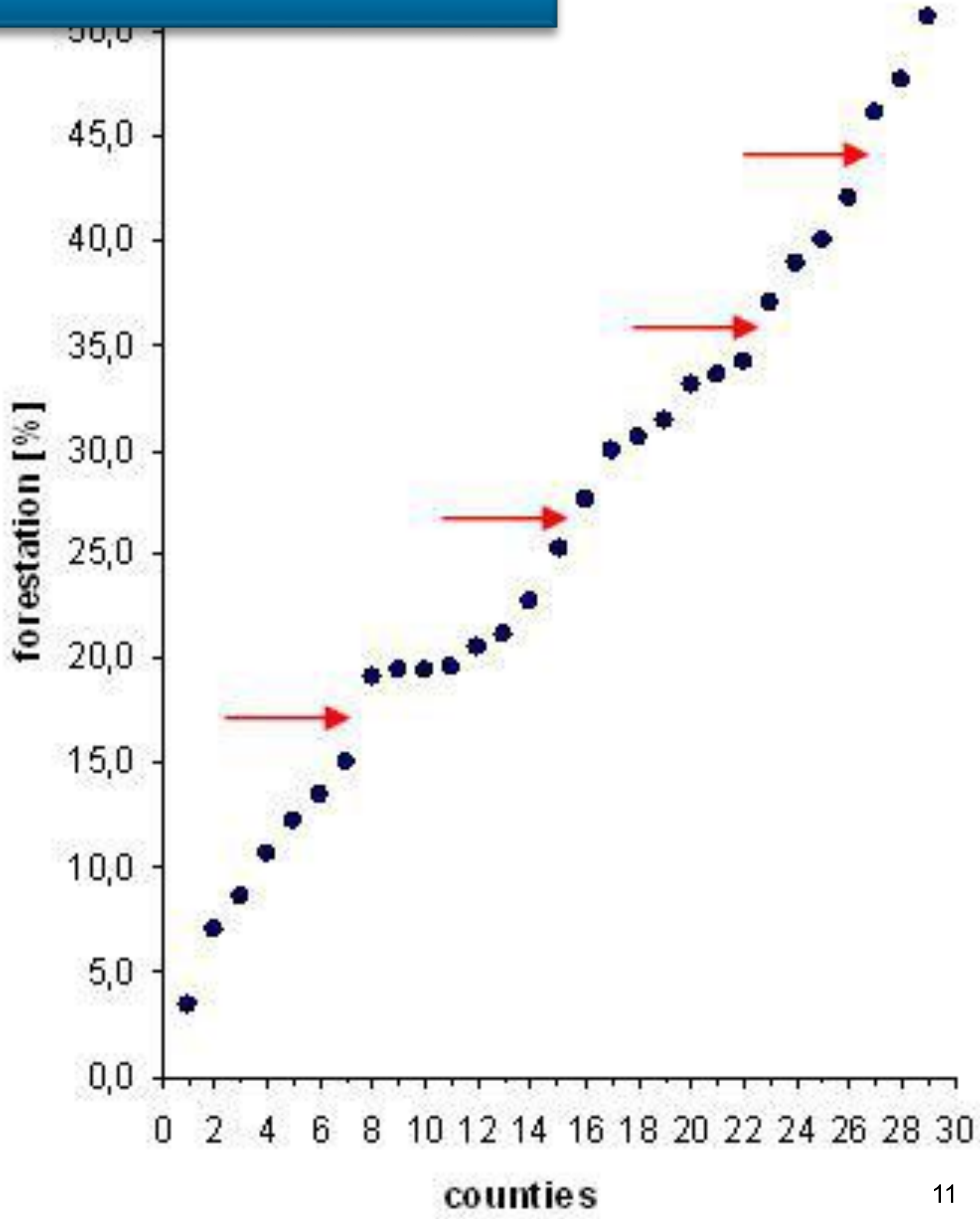
Natural Breaks classification

- Natural Break creates ranges according to an algorithm that uses the average of each range to distribute the data more evenly across the ranges. It distributes the values so that the average of each range is as close as possible to each of the range values in that range. This ensures that the ranges are well-represented by their averages, and that data values within each of the ranges are fairly close together.
- The Natural Break algorithm is based on the procedure described by Jenks and Caspall in their article "Error on Choroplethic Maps: Definition, Measurement, Reduction" from the *Annals of American Geographers*, June 1971.

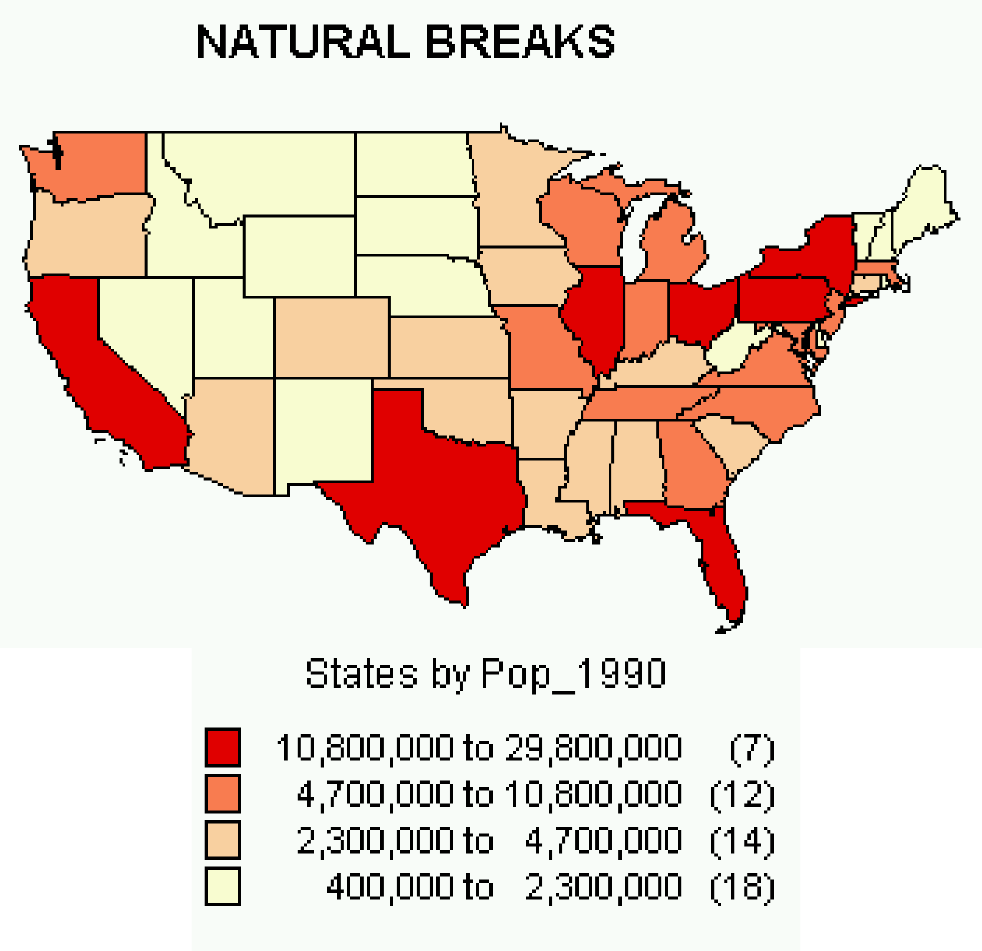
Jenks algorithm

- Choose the number of classes
  - Compute ADAM (the sum of absolute deviations about
- the mean for the entire data set) Σ (xi–ma)2 where ma is
- array mean
- For each iteration {
  - Determine class boundaries
  - Compute ADCM (the sum of absolute deviations about class mean) : $\Sigma\ \Sigma$ (xi–mc)2 where $mc$is class mean
  - ADF (Goodness of absolute deviation fit) = 1 – ADCM/ADAM
- Repeat the iteration until GADF cannot be maximized further

Natural Breaks classification



Natural Breaks classification


NATURAL BREAKS
States by Pop_1990
🟥 10,800,000 to 29,800,000 (7)
🟧 4,700,000 to 10,800,000 (12)
▨ 2,300,000 to 4,700,000 (14)
□ 400,000 to 2,300,000 (18)

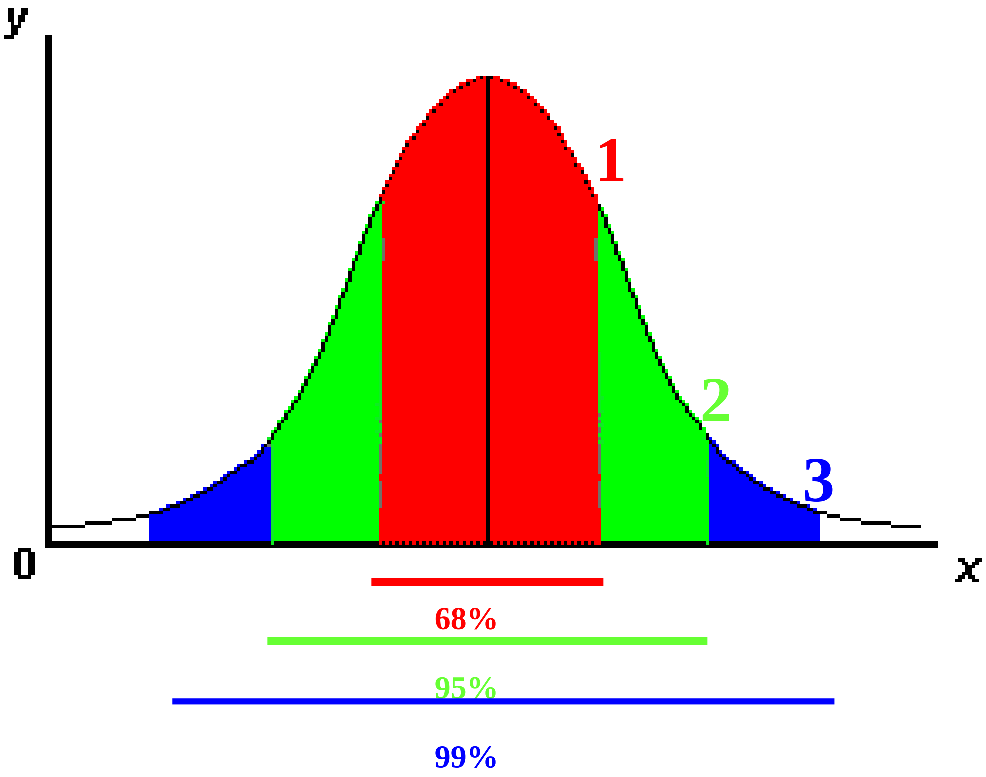
Standard Deviation classification

- When you create ranges using Standard Deviation, the middle range breaks at the mean of your values, and the ranges above and below the middle range are one standard deviation above or below the mean.

Standard Deviation example

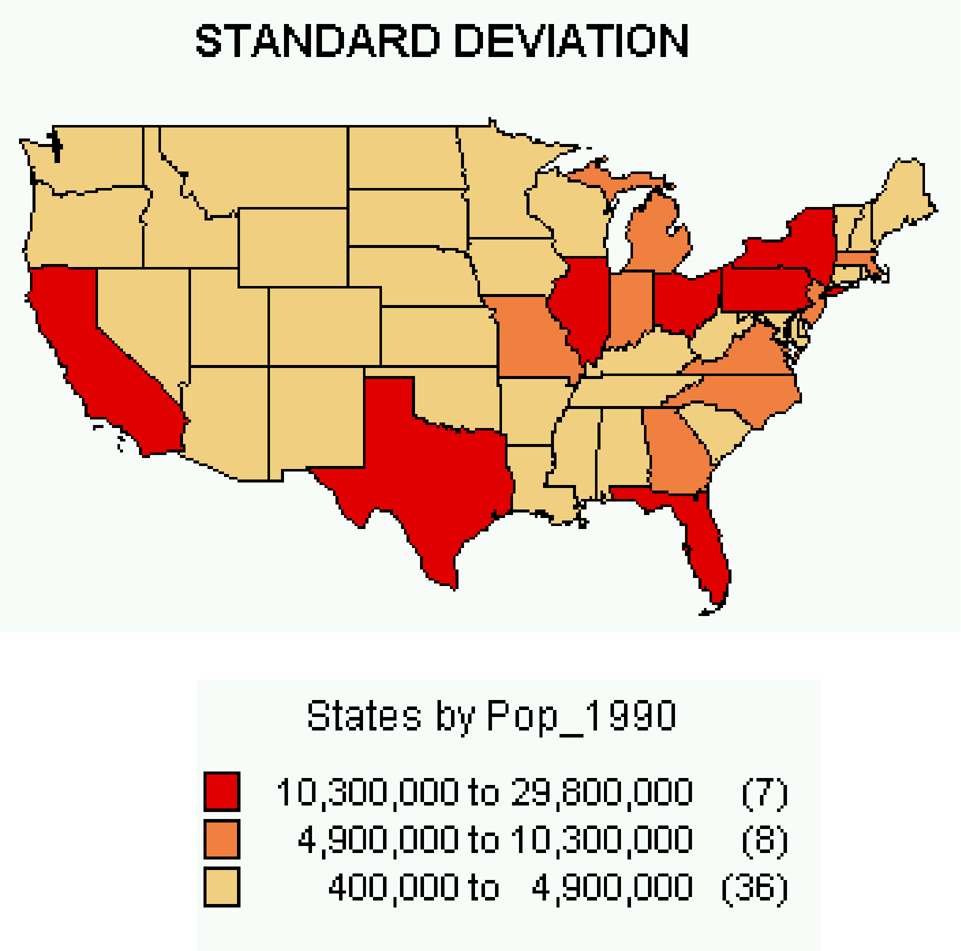
- if the average value is about 4.9 million and the standard deviation (SD) is about 4.5 million.
- In effect, the map shows above and below average values.
- This type of classification is popular for highlighting extremes of data.

Using Standard Deviation to Classify Values

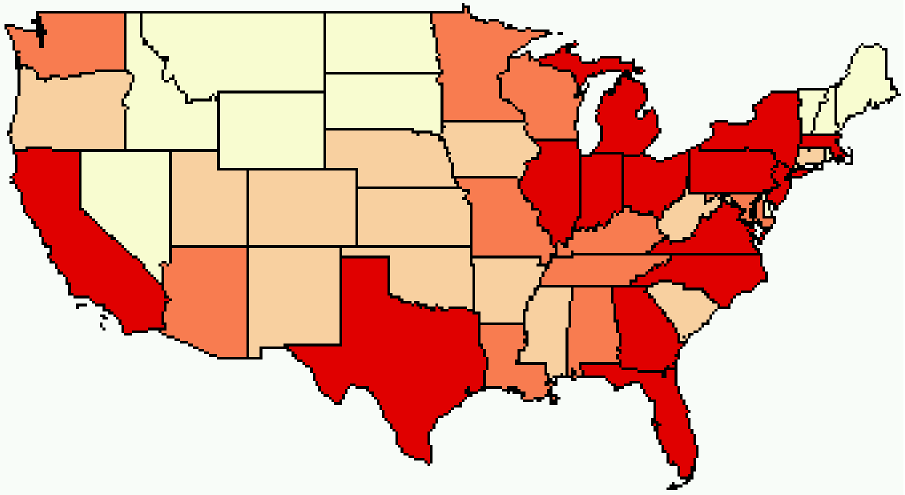


Standard Deviation classification

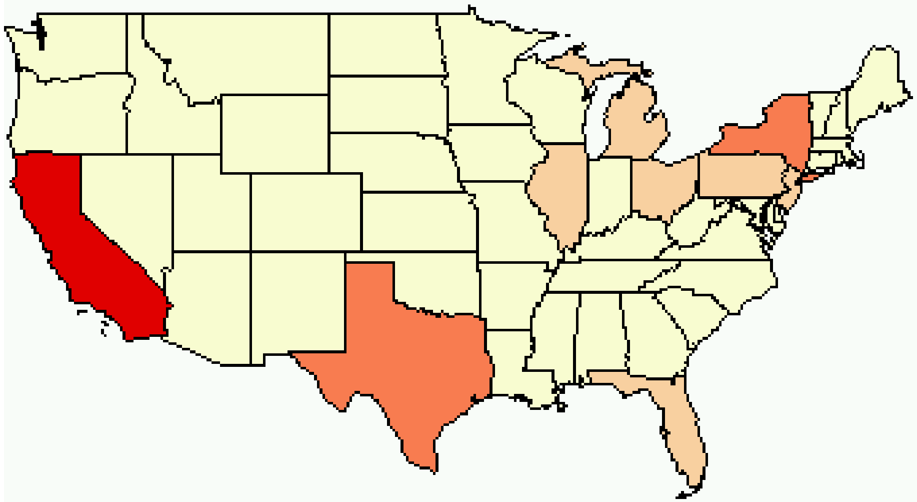
- Three ranges cover all the data values
- range 1 = 1 SD below the average
- range 2 = 1 SD above the average
- range 3 = 2 SD above the average).


STANDARD DEVIATION
States by Pop_1990
🟥 10,300,000 to 29,800,000 (7)
🟧 4,900,000 to 10,300,000 (8)
🟨 400,000 to 4,900,000 (36)

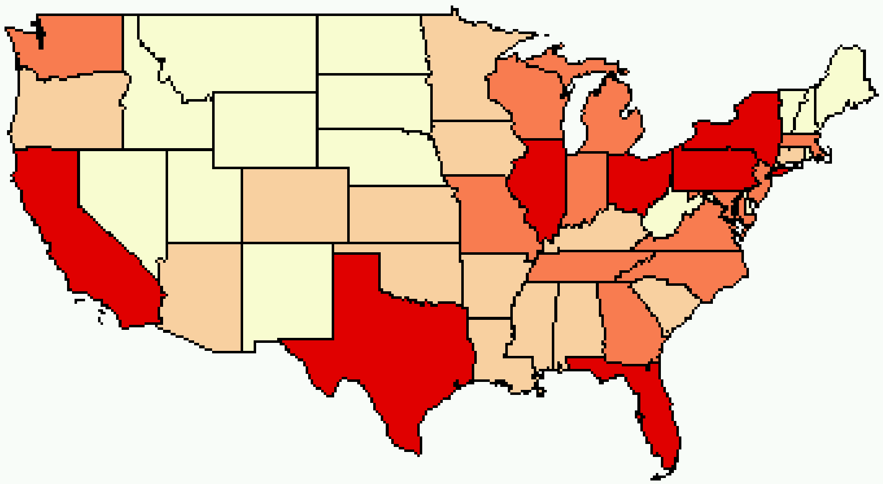
EQUAL COUNT



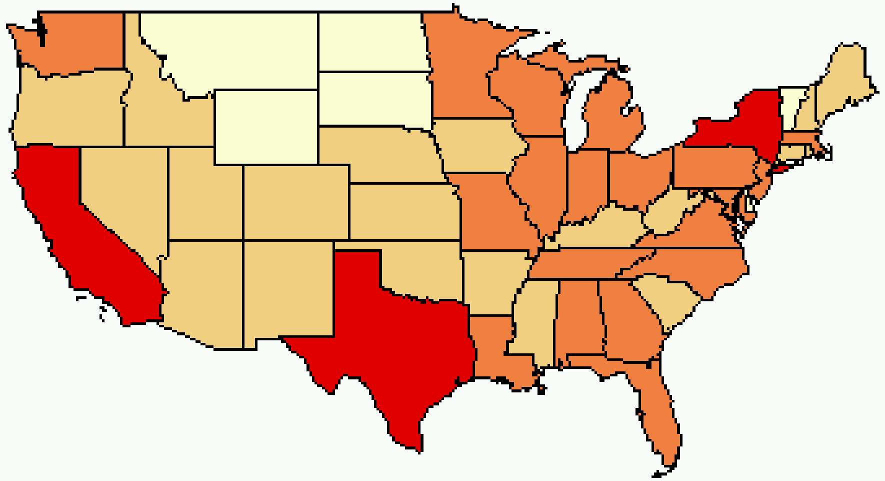
EQUAL RANGES



NATURAL BREAKS

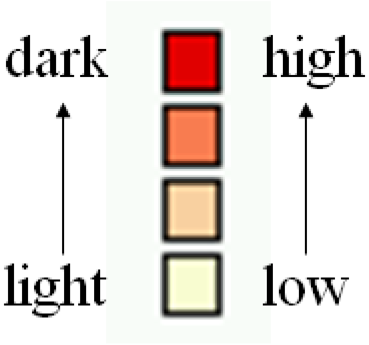


CUSTOM



General Rules:

- Notice that all of the examples shown illustrate a convention in thematic mapping – progressively darker shades represent progressively higher value ranges.
- On the question of how many ranges to use – research has shown that between 4 and 6 ranges is the most visually effective.
- Clearly the appearance of a thematic map can vary greatly depending on the classification method used



Evaluating classification methods

Classification methods	Works best when…	But there is a pitfall in that…
Equal intervals	Rectangular distribution Want to compare?	Rectangular distribution is rare
Quantiles	Interested in flattened pattern of skewed distribution?	Hide the fact data is skewed
Standard deviations	Normal distribution Interested in how typical/untypical?	What if map users don’t understand mean and standard deviation
Natural breaks	Any distribution Flexible, intuitive	Breaks not always obvious Not good for comparison