# Introduction to Data Visualization

## Color Scales
## Visualization Types

**Halil Bisgin, Ph.D.**

# Color Scales

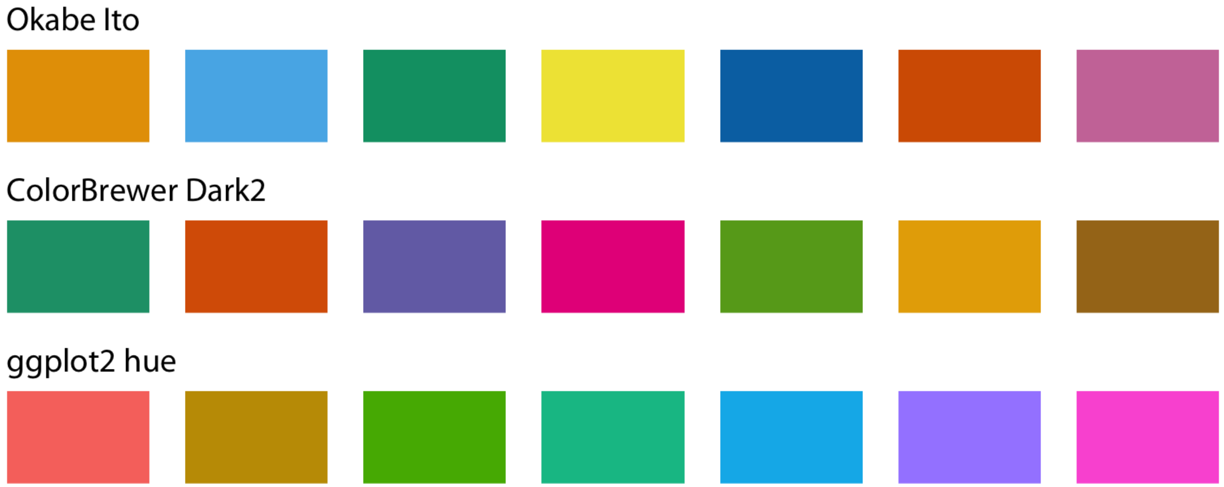
- We can use color
  - *to distinguish groups of data from each other,*
  - *to represent data values, and*
  - *to highlight.*
- The types of colors we use and the way in which we use them are quite different for these three cases.

# Color as a Tool to Distinguish

- We frequently use color as a means to distinguish discrete items or groups that do not have an intrinsic order:
  - *different countries on a map or different manufacturers of a product.*
- Qualitative color scale.
  - *Finite set of specific colors that are chosen to look clearly distinct from each other while also being equivalent to each other.*
  - *No one color should stand out relative to the others.*
  - *The colors should not create the impression of an order.*

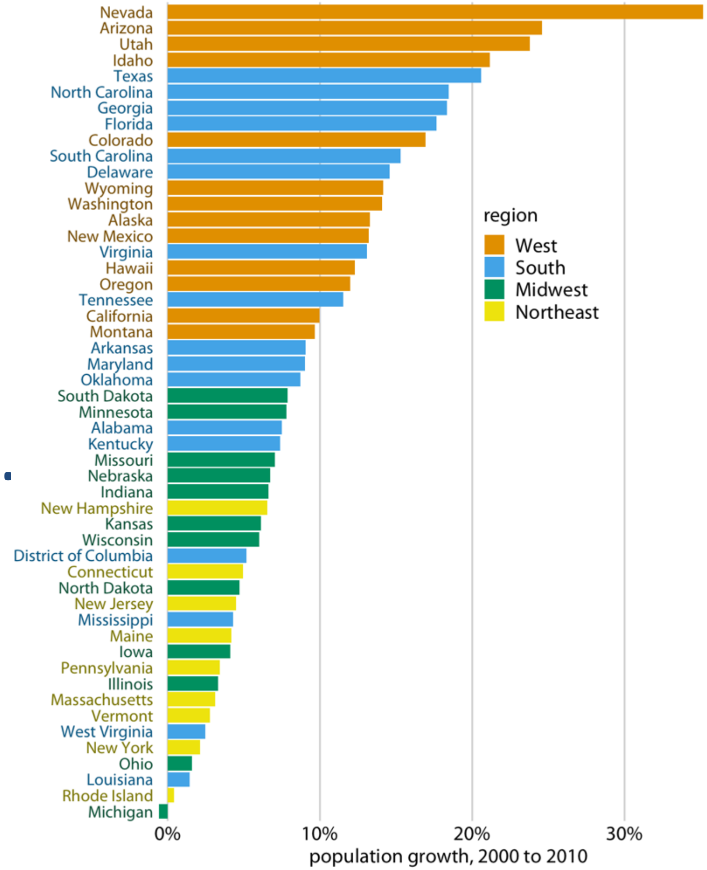
# Qualitative Color Scale-Example

- ColorBrewer and other resources provides such selections.



# Qualitative Color Scale-Example

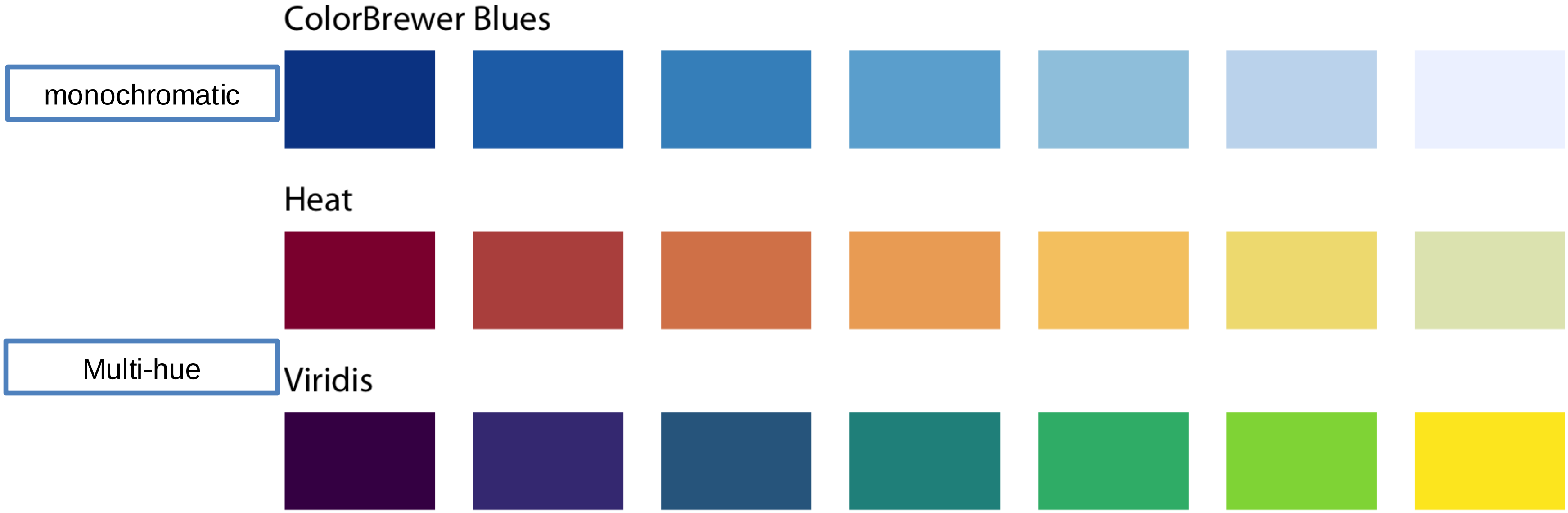
- States arranged based on population growth.
- Colored by geographic region.
- Coloring shows states in the same region have similar growth.



# Color to Represent Data Values

- Color can also be used to represent quantitative data values, such as income, temperature, or speed.
- Sequential color scale.
  - *To represent larger or smaller values and their distances.*
  - *Needs to be perceived to vary uniformly across its entire range.*
  - *Can be based on a single hue or on multiple hues (e.g., from dark red to light yellow).*
  - *Multi-hue scales tend to follow color gradients that can be seen in the natural world, such as dark red, green, or blue to light yellow, or dark purple to light green.*
  - *The reverse (e.g., dark yellow to light blue) looks unnatural and doesn’t make a useful sequential scale.*

# Sequential Color Scale-Example



ColorBrewer Blues

monochromatic

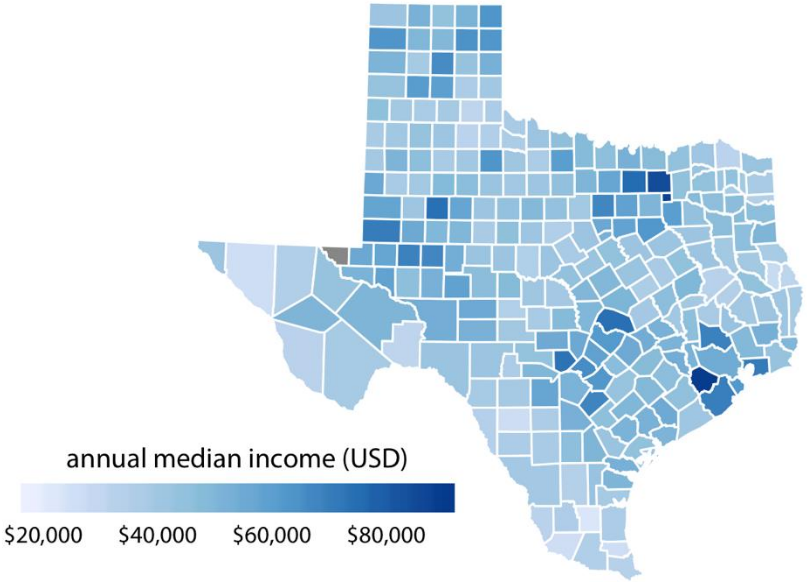
Heat

Multi-hue

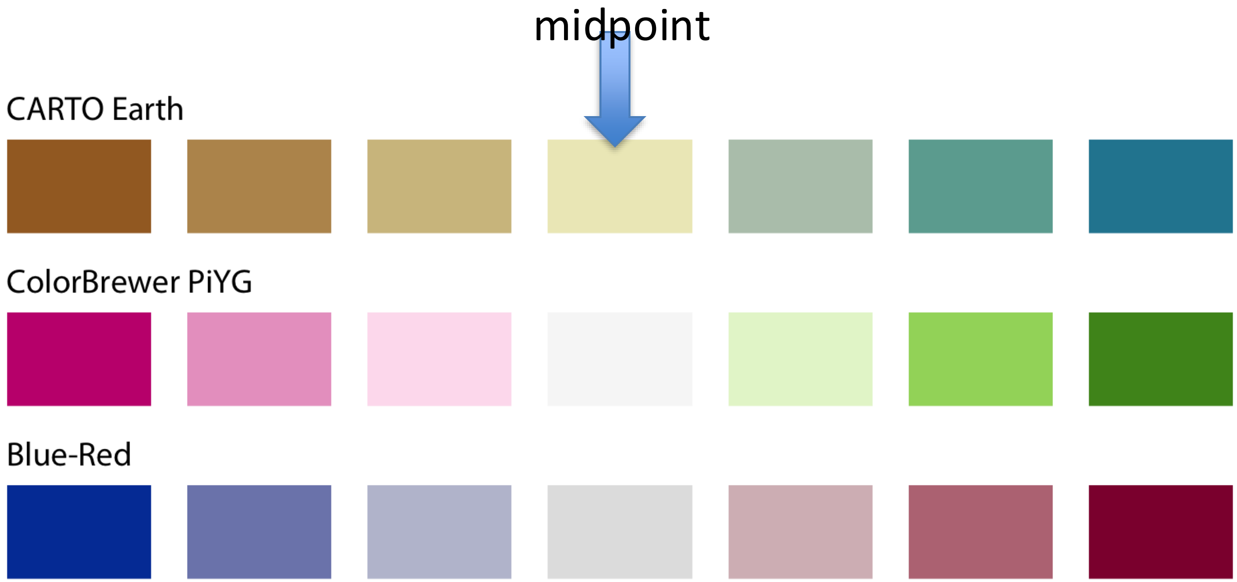
Viridis

# Sequential Color Scale-Example

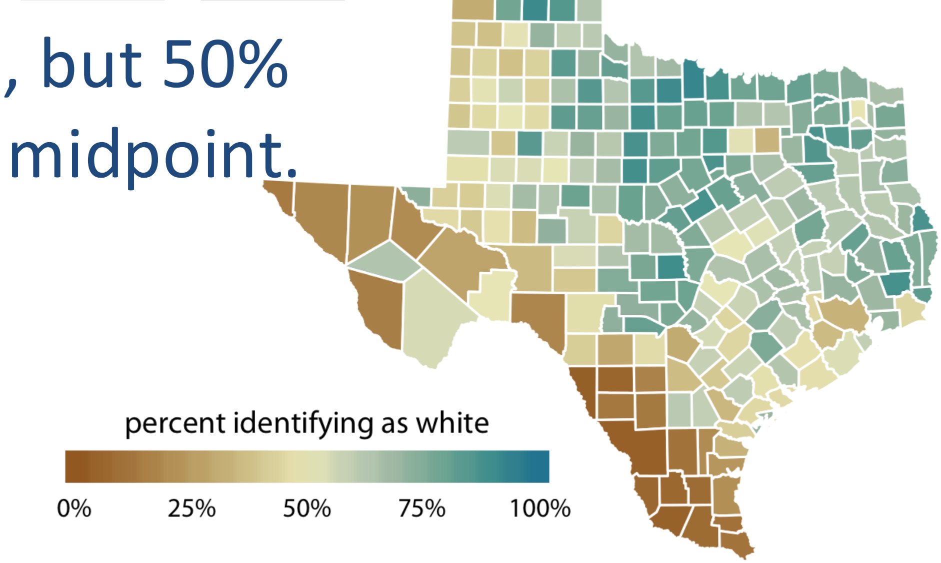
- We can draw a map (choropleths) of the geographic regions and color them by the data values.
- What if you want to visualize the deviation of data values in one of two directions relative to a neutral midpoint?
  - –*diverging color scale*



# Sequential Color Scale-Diverging



- Values are always positive, but 50% can be assumed to be the midpoint.

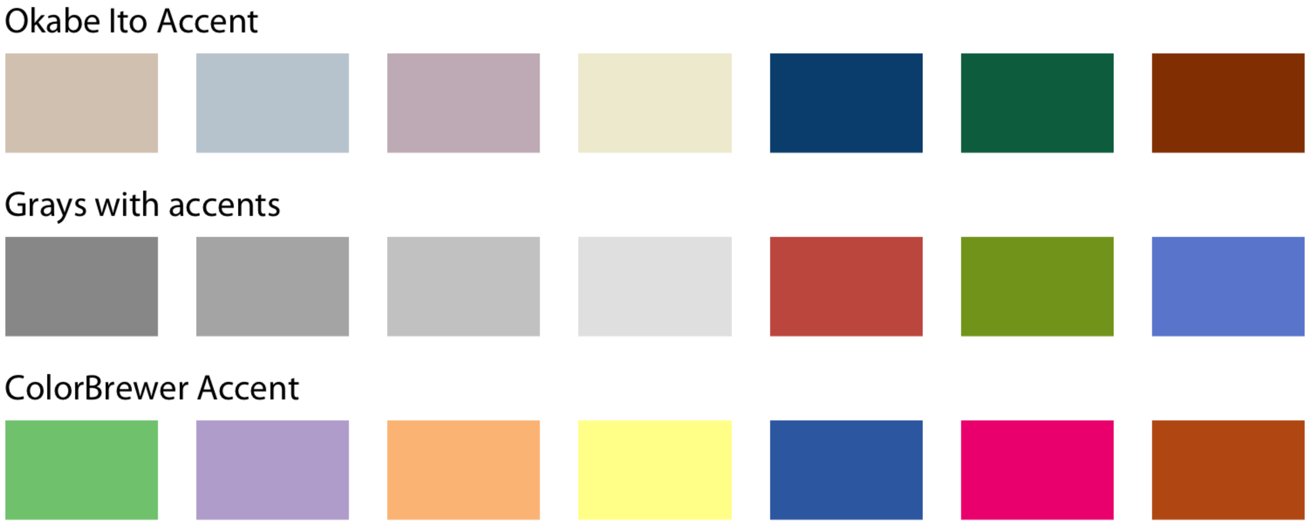


# Color as a Tool to Highlight

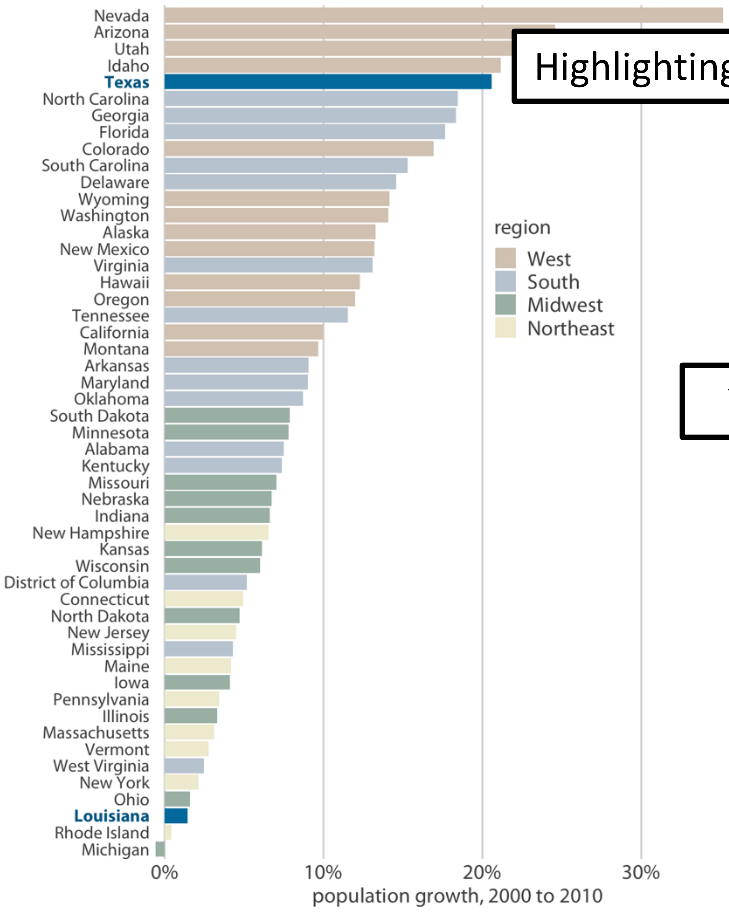
- Effective tool to highlight specific elements.
- There may be specific categories/values in the dataset that carry key information about the story we want to tell.
- An easy way to achieve this emphasis is to color these figure elements in a color or set of colors that vividly stand out against the rest of the figure.
- This effect can be achieved with accent color scales, which contain both a set of subdued colors and a matching set of stronger, darker, and/or more saturated colors

# Accent Color Scales

- Accent color scales can be derived in several different ways:
  - *Take an existing color scale and lighten and/or partially desaturate some colors while darkening others (top).*
  - *Take gray values and pair them with colors (middle).*
  - *Use an existing accent color scale (bottom).*

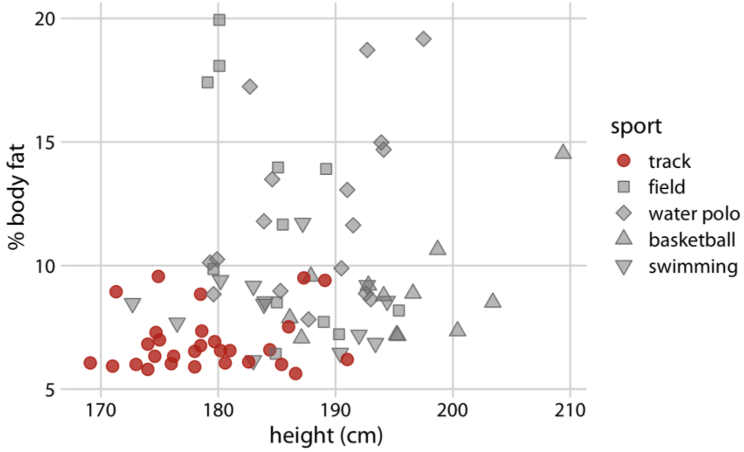


# Accent Color Scales-Example



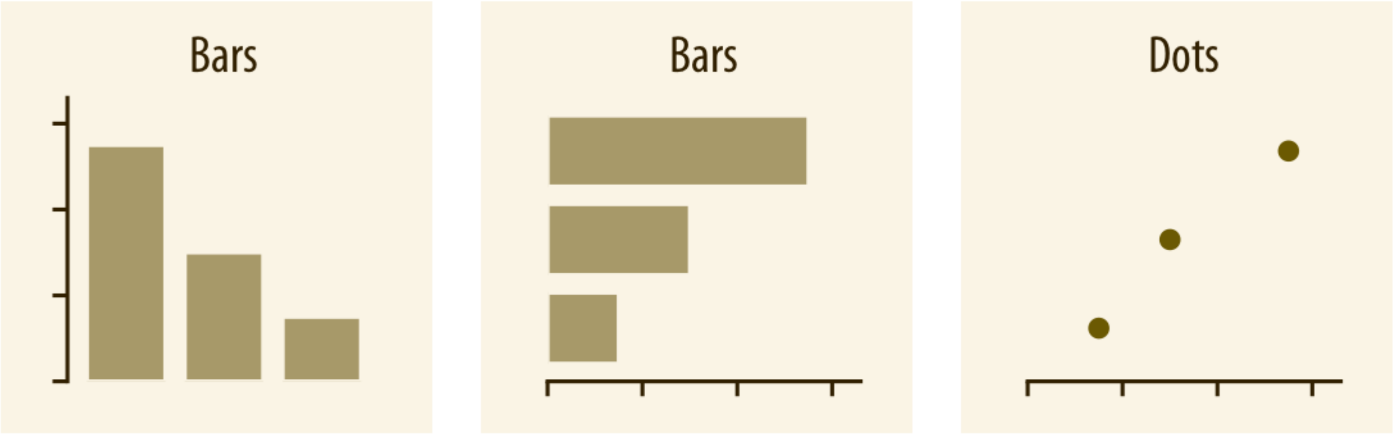
Highlighting two neighboring states. Fifth fastest vs. third slowest

You can also remove color not to cause any color competition



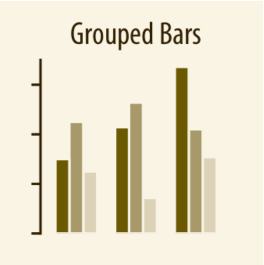
# Visualization Types-Amounts

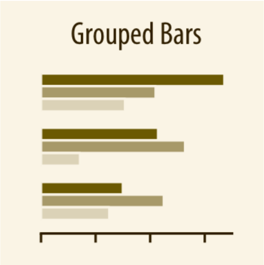
- The most common approach to visualizing amounts (i.e., numerical values shown for some set of categories) is using bars (vertically or horizontally).
- We can also place dots at the location where the corresponding bar would end

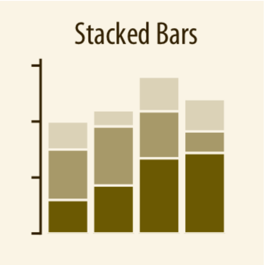


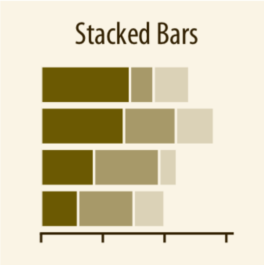
# Visualization Types-Amounts

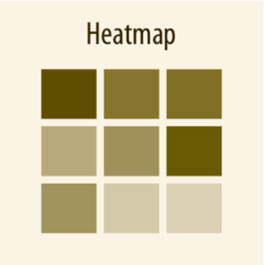
- For two or more sets of categories for which we want to show amounts, we can group or stack the bars.
- We can also map the categories onto the x and y axes and show amounts by color, via a heatmap.





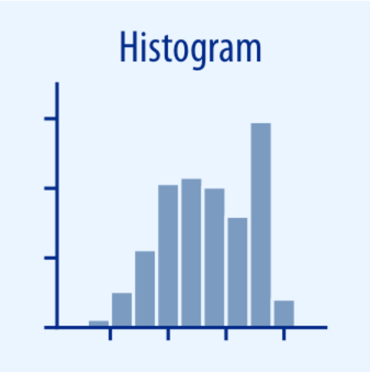


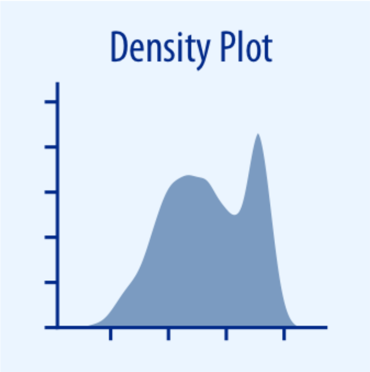


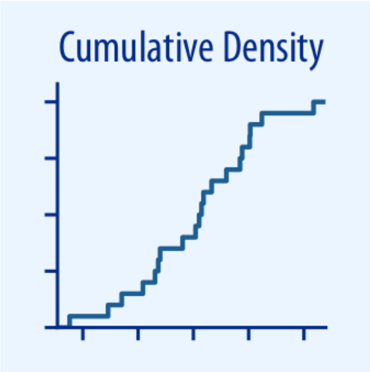


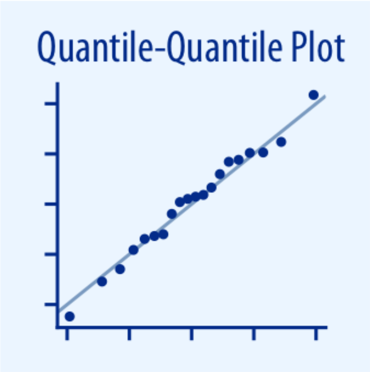
# Visualization Types-Distributions

- Histograms and density plots provide the most intuitive visualizations of a distribution, but both require arbitrary parameter choices and can be misleading.
- Cumulative densities and quantile-quantile (q-q) plots always represent the data faithfully but can be more difficult to interpret.


Histogram

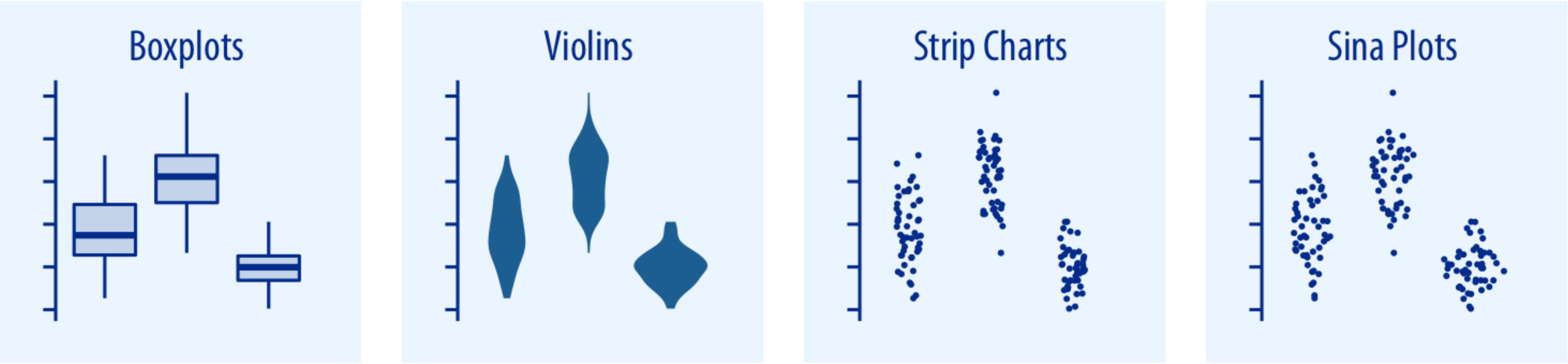

Density Plot


Cumulative Density


Quantile-Quantile Plot

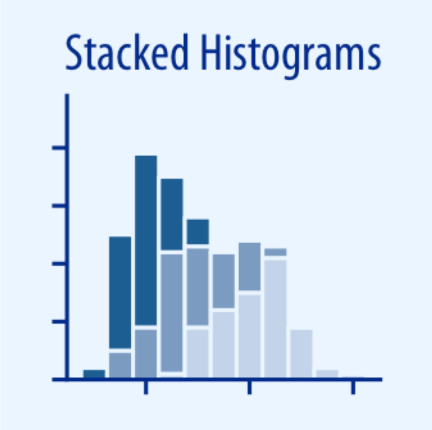
# Visualization Types-Distributions

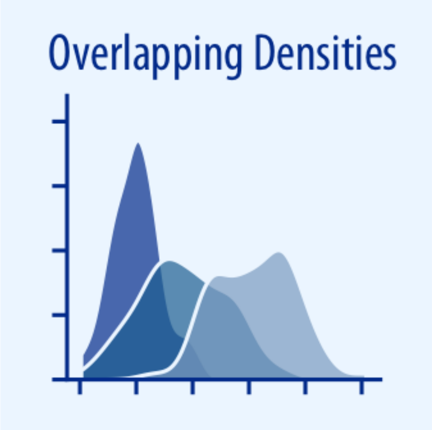
- Boxplots, violin plots, strip charts, and sina plots are useful to visualize many distributions at once and/or if we are primarily interested in overall shifts.

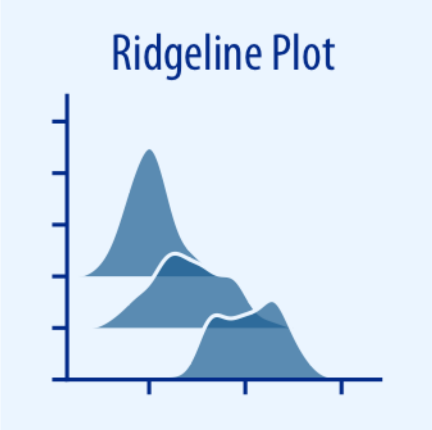


# Visualization Types-Distributions

- Stacked histograms and overlapping densities allow a more in-depth comparison of a smaller number of distributions.
  - *Can be difficult to interpret and are best avoided.*
- Ridgeline plots can be a useful alternative to violin plots and are often useful when visualizing many distributions or changes in distributions over time



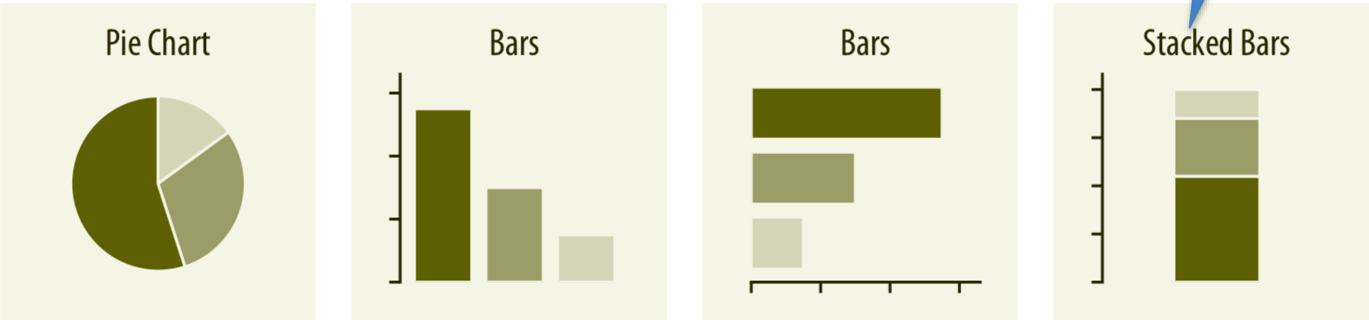




# Visualization Types-Proportions

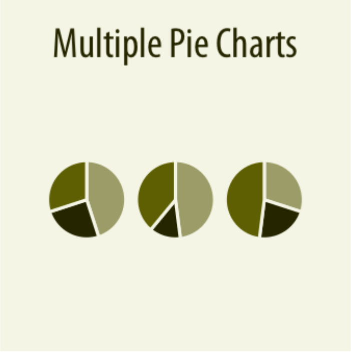
- Proportions can be visualized as pie charts (!), side-by-side bars, or stacked bars.
- Pie charts emphasize that the individual p to a whole and highlight simple fractions.
  - *The individual pieces are more easily compared i bars.*
- Stacked bars look awkward for a single set o proportions, but can be useful when comparing mu

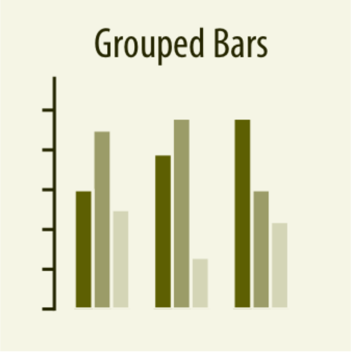
What is a good alternative for stacked bars when there's one set of proportions?

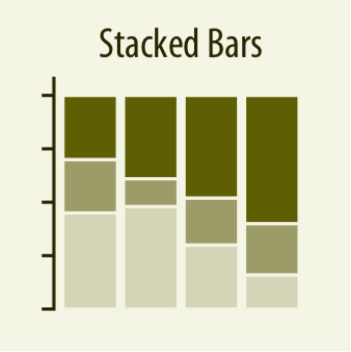


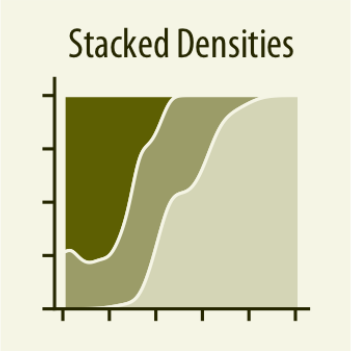
# Visualization Types-Proportions

- Pie charts tend to be space-inefficient and often obscure relationships when visualizing multiple sets.
- Grouped bars work well as long as the number of conditions compared is moderate, and stacked bars can work for large numbers of conditions.
- Stacked densities are appropriate when the proportions change along a continuous variable.



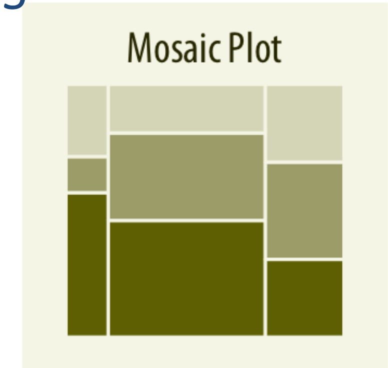


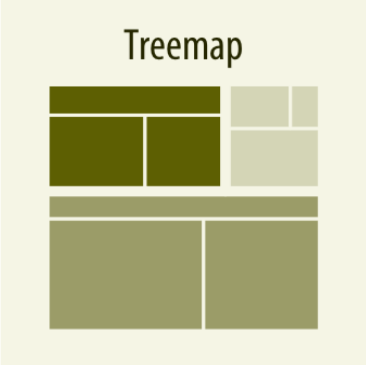


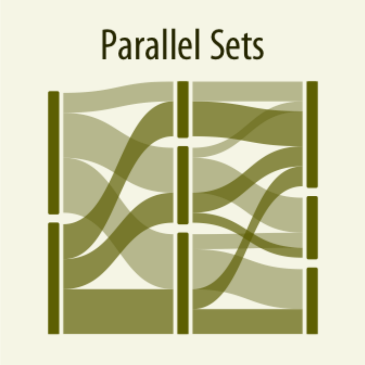


# Visualization Types-Proportions

- Proportions can be based on multiple grouping variables. (survivors from Titanic: gender vs. class)
  - *Mosaic plots: every level of one grouping variable can be combined with every level of another grouping variable.*
  - *Treemaps: do not make such an assumptions and work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.*
  - *Parallel sets: work better when there are more than two grouping variables.*

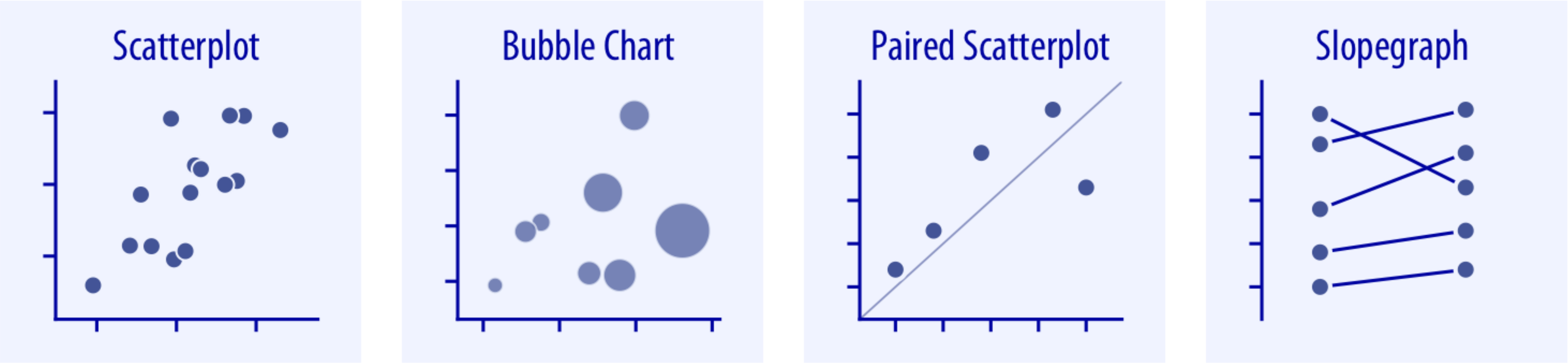

Mosaic Plot


Treemap


Parallel Sets

# Visualization Types-*x-y relationships*

- Scatterplots represent the archetypical visualization to show x-y relationships.
- Bubble chart: a variant of scatterplot that maps one variable to the dot size.
- For the same units, it is helpful to add a diagonal line.
- Paired data can also be shown as a slope graph.



Scatterplot

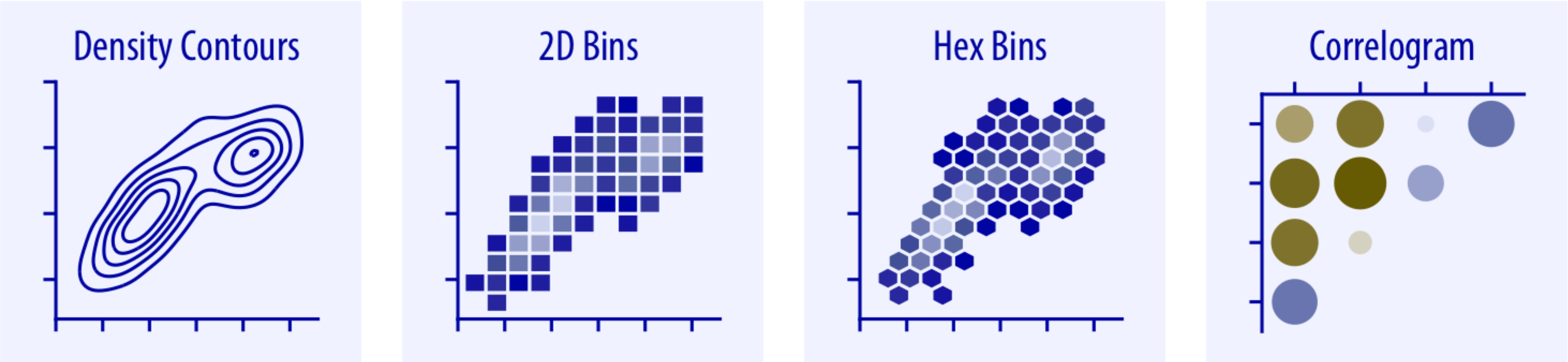
Bubble Chart

Paired Scatterplot

Slopegraph

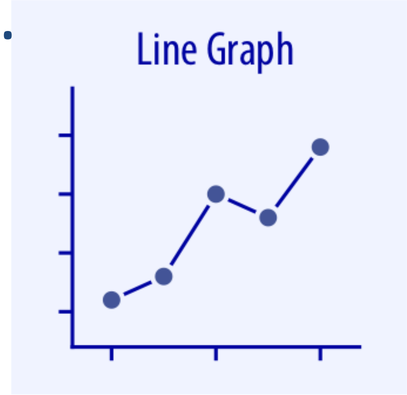
# Visualization Types-*x-y relationships*

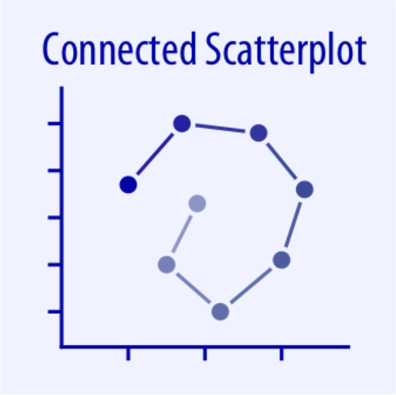
- For large numbers of points, regular scatterplots can become uninformative due to overplotting. Alternatives:
  - *contour lines, 2D bins, or hex bins.*
- When there are more than two quantities, correlation coefficients can be visualized in the form of a correlogram instead of the underlying raw data.

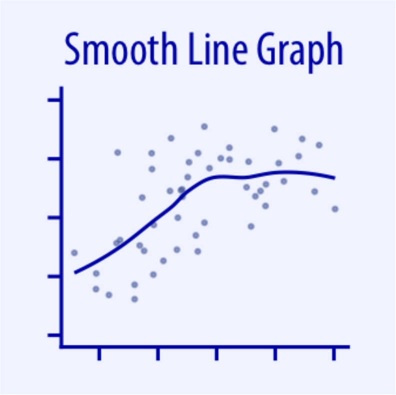


# Visualization Types-*x-y relationships*

- Line graphs: when *x* represents time or a strictly increasing quantity such as a treatment dose.
- Connected scatterplot: when a temporal sequence of two response variables.
  - *1) plot the two response variables in a scatterplot, 2) connect dots corresponding to adjacent time points.*
- We can use smooth lines to represent trends in a larger dataset.

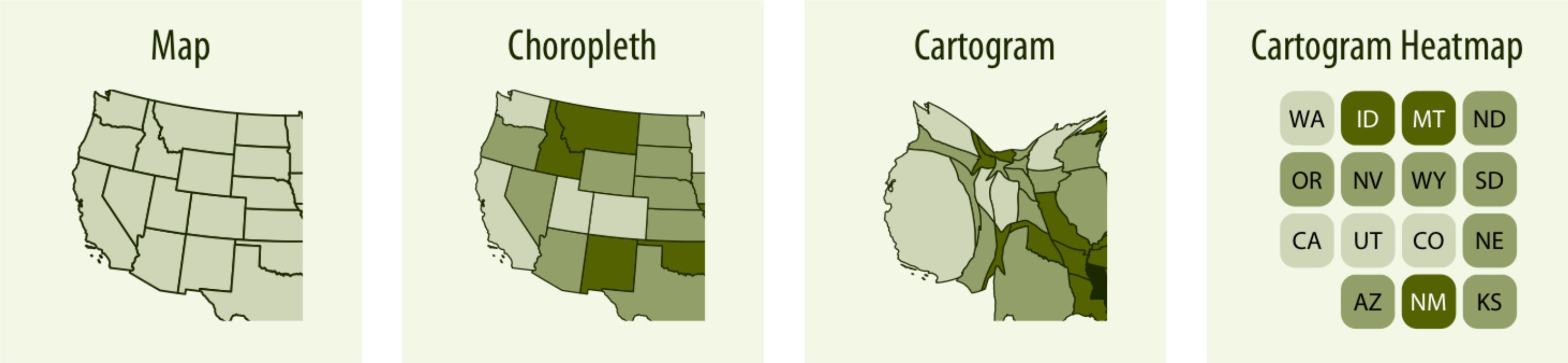






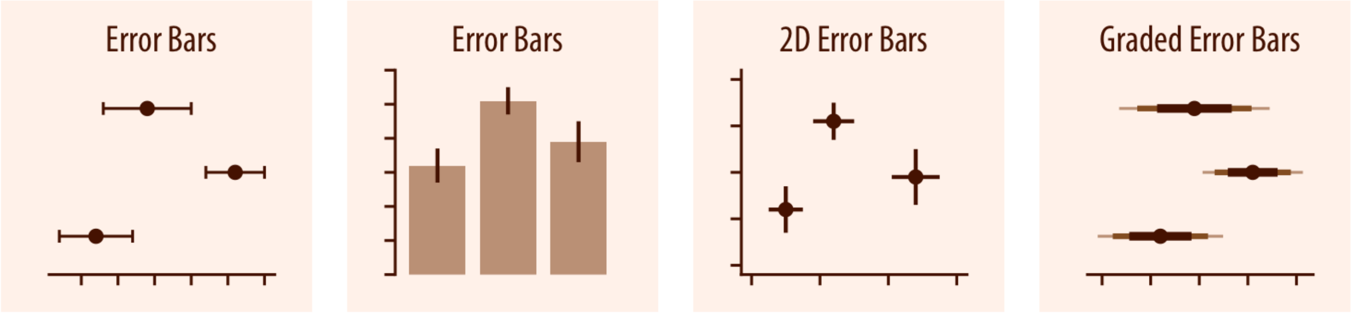
# Visualization Types-Geospatial Data

- Geospatial data visualized in the form of a map.
- A map takes coordinates on the globe and projects them onto a flat surface (2D representation).
  - *We can show data values in different regions by coloring those regions in the map according to the data (choropleth).*
  - *It may also be helpful to distort the different regions according to some other quantity (e.g., population number) or simplify each region into a square (cartograms).*



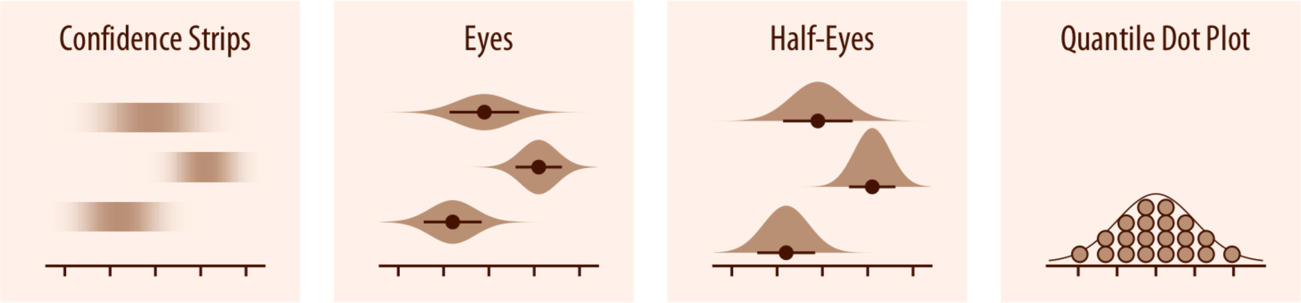
# Visualization Types-Uncertainty

- Error bars are meant to indicate the range of likely values for some estimate or measurement.
  - *Horizontally and/or vertically from some reference point representing the estimate or measurement.*
  - *Reference points can be shown in various ways (e.g., dots or bars).*
  - *Graded error bars show multiple ranges at the same time, where each range corresponds to a different degree of confidence.*
  -



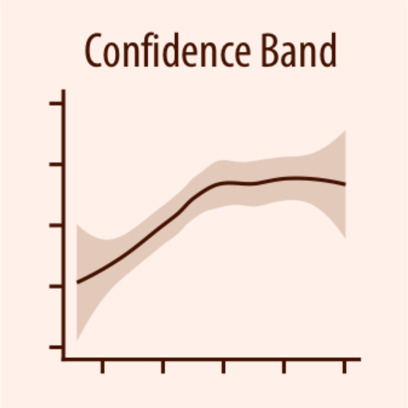
# Visualization Types-Uncertainty

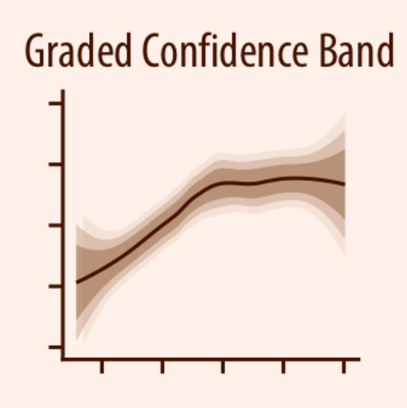
- We can visualize the actual confidence or posterior distributions for more detailed visualization.
  - –*Confidence strips: a visual sense of uncertainty but difficult to read.*
  - –*Eyes and half-eyes: combine error bars with approaches to visualize distributions (violins and ridgelines, respectively), and thus show both precise ranges for some confidence levels and the overall uncertainty distribution.*
  - –*A quantile dot plot: alternative visualization of an uncertainty distribution. Shows the distribution in discrete units. Not precise but easier to*

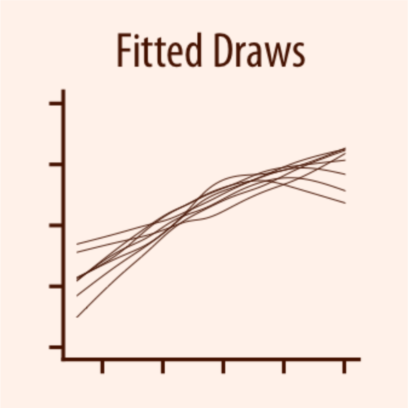


# Visualization Types-Uncertainty

- For smooth line graphs, the equivalent of an error bar is a confidence band.
  - *Shows a range of values the line might pass through at a given confidence level.*
  - *Like with error bars, we can draw graded confidence bands that show multiple confidence levels at once.*
  - *We can also show individual fitted draws in lieu of or in addition to the confidence bands.*


Confidence Band


Graded Confidence Band


Fitted Draws