Fundamentals of Data Visualization

Chapter 5

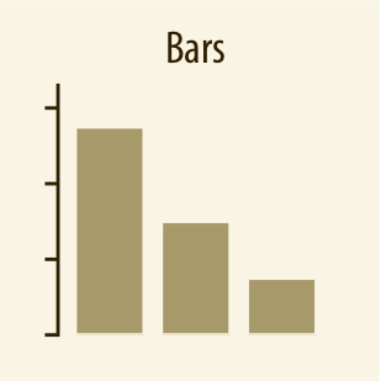
April 25, 2023

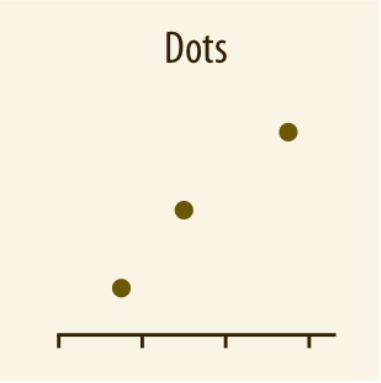
Taxonomy of visualizations

- Amounts
- Distributions
- Proportionas
- x-y relationships
- Geospatial data
- Uncertainty

Visualizing amounts

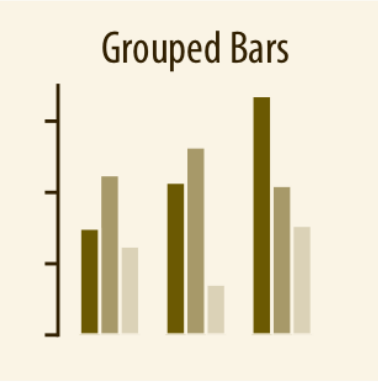
- Most popular are bars or dots

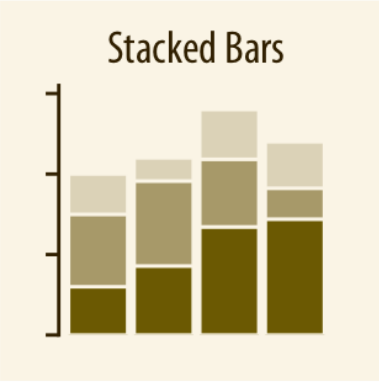




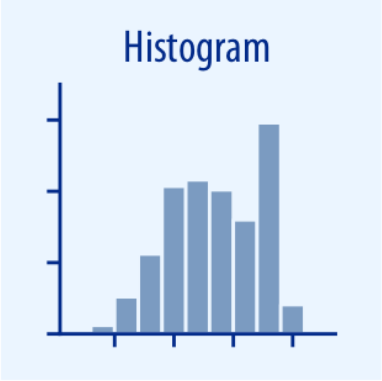
Bars for more sets of categories

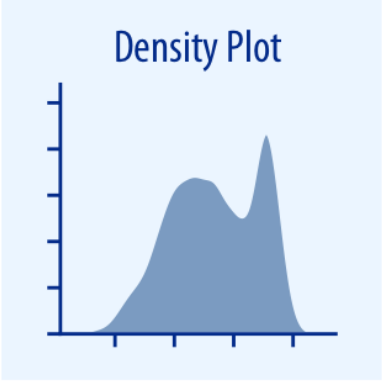
- Grouped bars
- Stacked bars
- Heat map

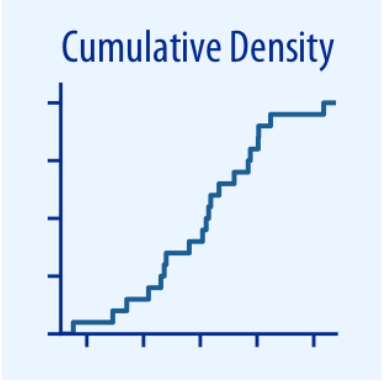




Distributions

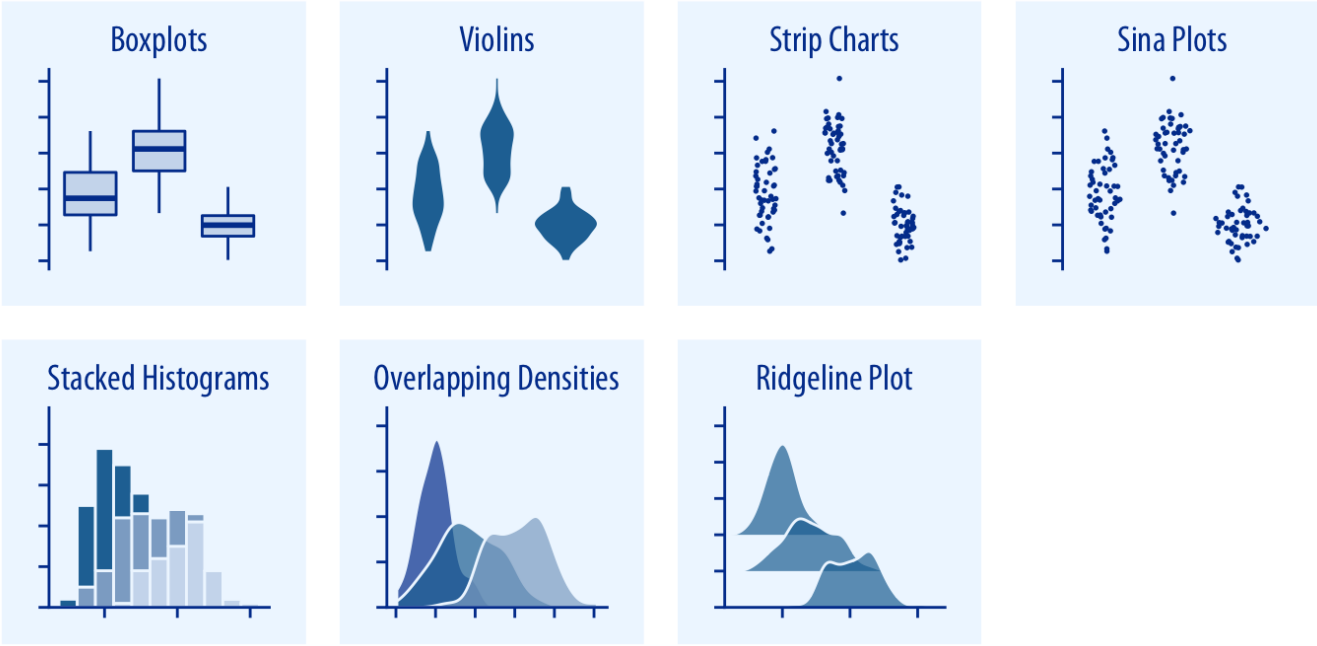






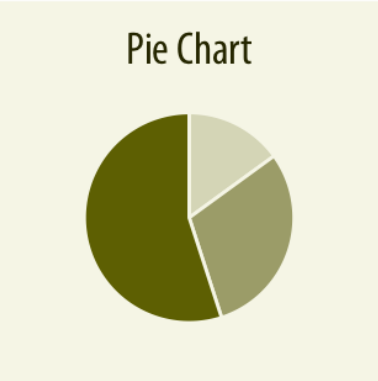
- Histograms and density plots are intuitive visualizations
- Both require arbitrary parameter choices and can be misleading
- Cumulative density and Q-Q plots always represent the data faithfully
- CD and Q-Q can be difficult to interpret

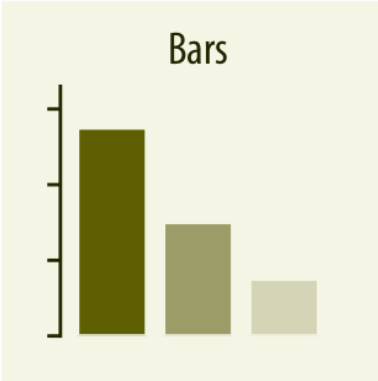
Multiple distributions

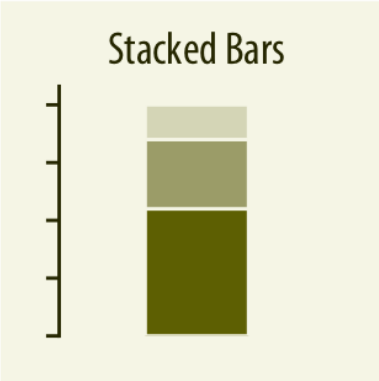


- Boxplots, violins, strip charts, and sina plots are useful for many distributions at once.
- Good at showing overall shifts among distributions.
- Stacked histograms and overlapping densities allow a more in-depth comparison of a smaller number of distributions.
- Stacked histograms are difficult to interpret and best avoided.

Proportions

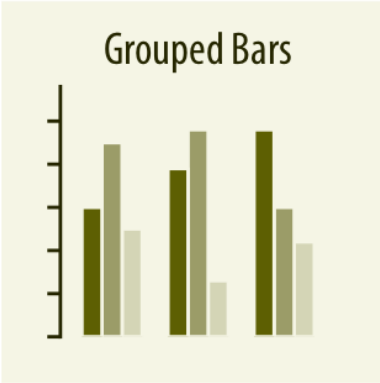


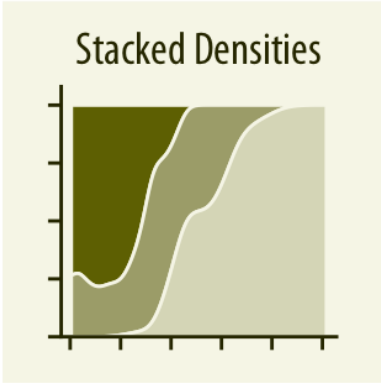




- Pie charts emphasize that the parts add up to a whole.
- Pie charts are difficult to compare individual sizes.
- Stacked bars look awkward for a single set of proportions
- Can be useful when comparing multiple sets of proportions

Multiple Proportions



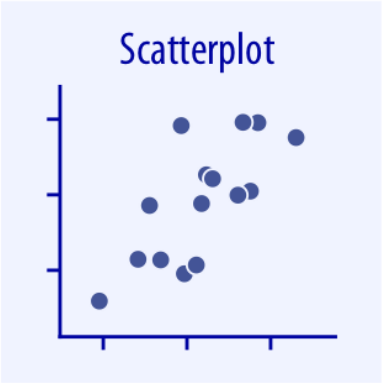


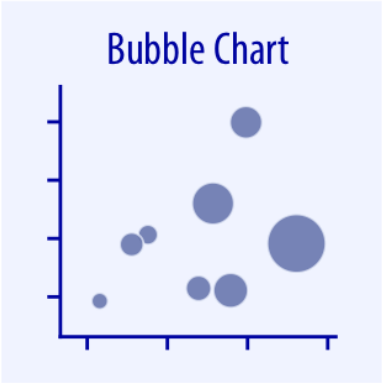
- Pie charts tend to be space-inefficient and often obscure relationships.
- Grouped bars work well as long as the number of conditions compared is moderate
- Stacked bars can work for large numbers of conditions
- Stacked densities are appropriate when the proportions change along a continuous variable.

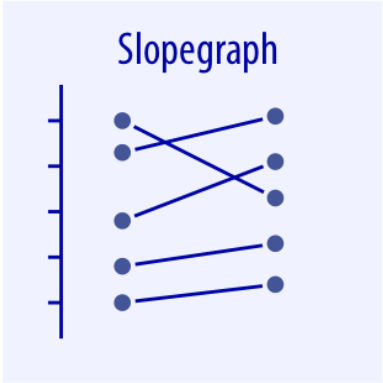
Proportions with multiple grouping variables

- Mosaic plots, treemaps, and parallel sets can visualize proportions according to multiple grouping variables.
- Mosaic plots assume every level of one grouping variable can be combined with every level of anothre grouping variable.
- Treemaps do not make that assumption.
- Parallel set work better when there are more than two grouping variables.

x-y relationships

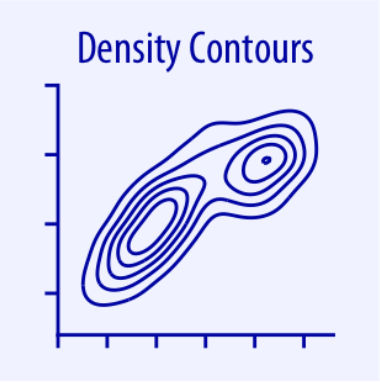


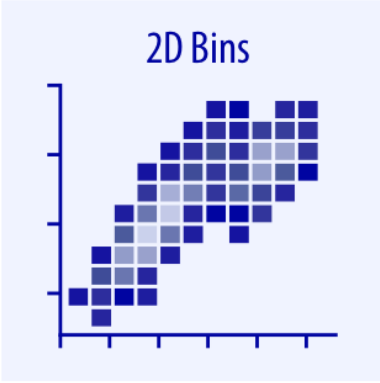


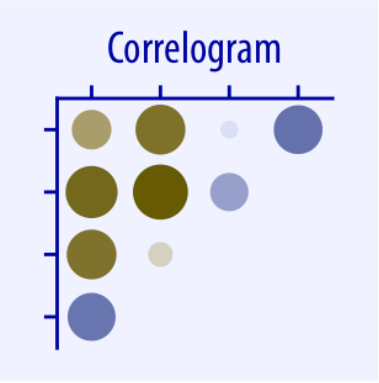


- Scatterplots are the archetypical visualization when we want to show one quantitative variable relative to another
- If we have three quantitative variables, we can map one onto the dot size, creating the bubble chart
- For paired data, where the variables along the x and the y axes are measured in the same units, it is generally helpful to add a line indicating $x = y$
- Paired data can also be shown as a slope graph of paired points connected by straight lines

Large numbers of points

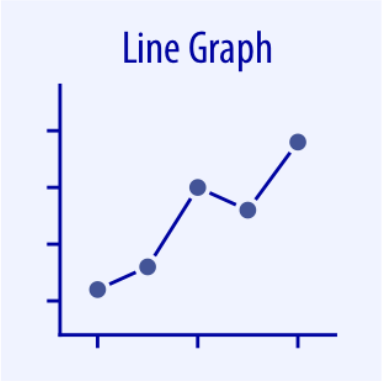


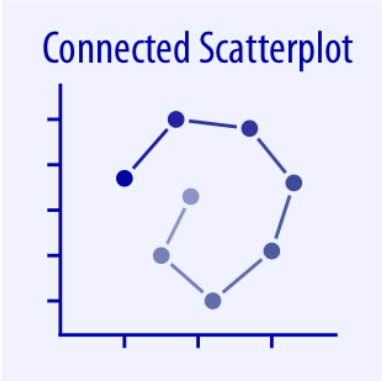


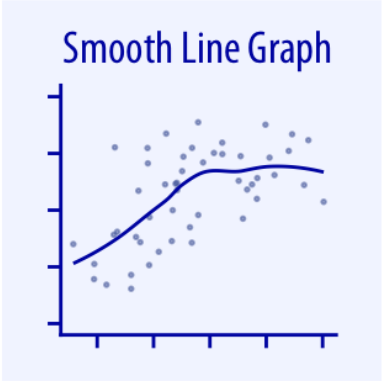


- For large numbers of points, regular scatterplots can become uninformative due to overplotting.
- Contour lines, 2D bins, or hex bins may provide an alternative
- For more than two quantities, we may want to plot correlation coefficients in a correlogram

Lines

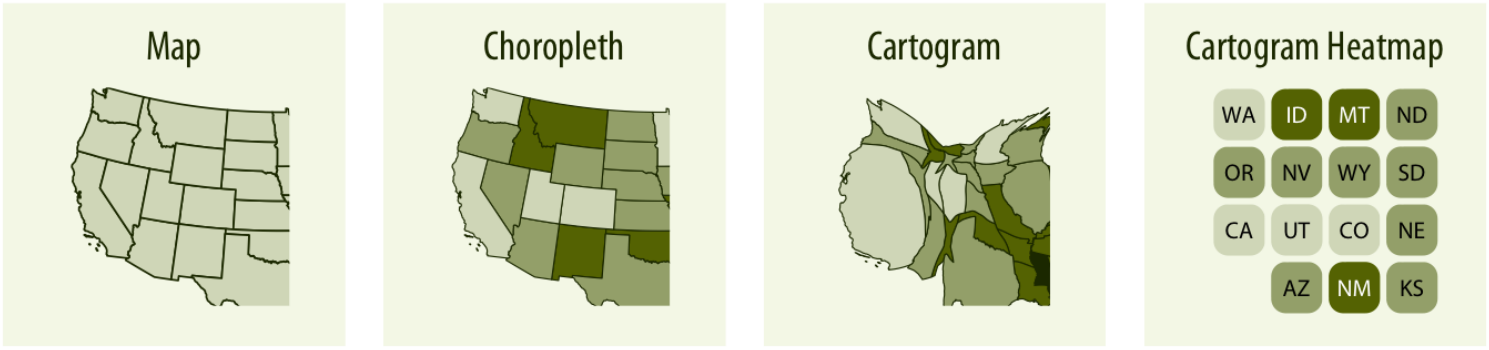






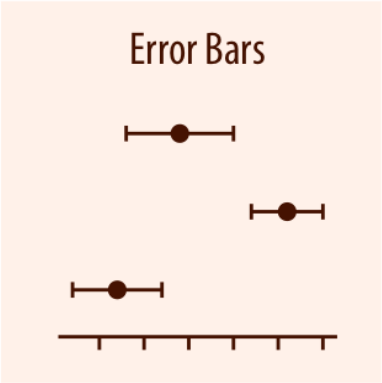
- When the x axis represents time or a strictly increasing quantity such as a treatment dose, we commonly draw line graphs
- If we have a temporal sequence of two response variables, we can draw a connected scatterplot
- We can use smooth lines to represent trends in a larger dataset

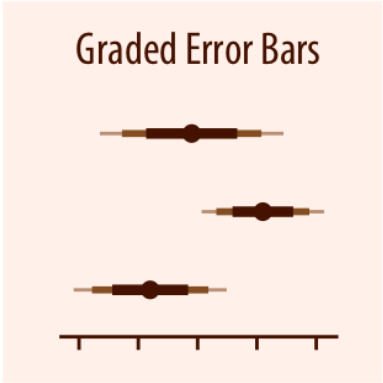
Geospatial data



- The primary mode of showing geospatial data is in the form of a map
- A map takes coordinates on the globe and projects them onto a flat surface
- Data values in different regions are shown by different colors.
- Such a map is called a choropleth
- In some cases, it may be helpful to distort the different regions according to some other quantity (e.g., population number) or simplify each region into a square.
- Such visualizations are called cartograms.

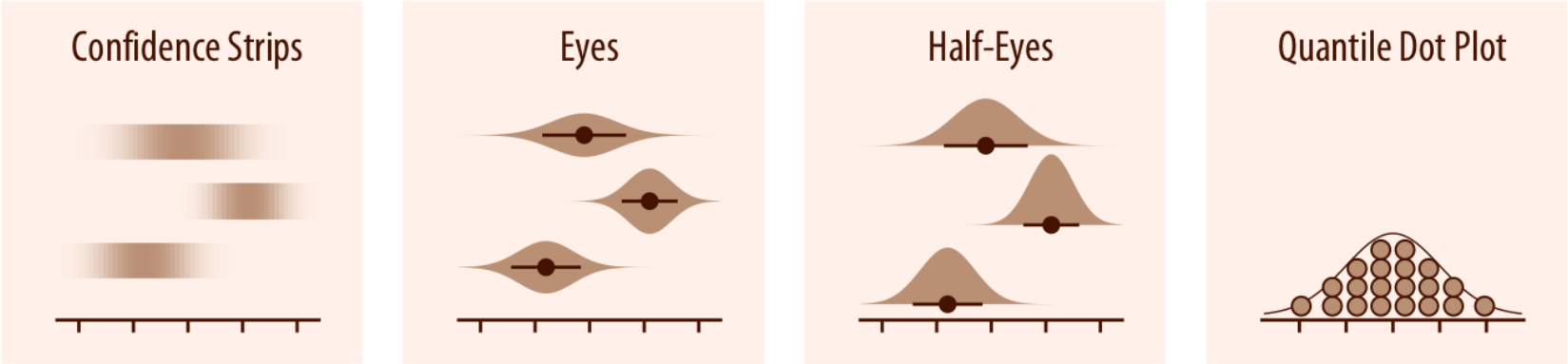
Uncertainty





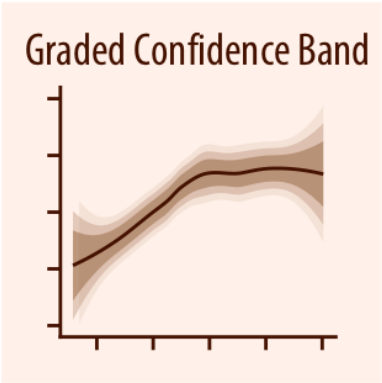
- Error bars are meant to indicate the range of likely values for some estimate or measurement
- E.g. mean and standard deviation, or median and inter-quartile range
- Graded error bars show multiple ranges at the same time, where each range corresponds to a different degree of confidence.
- They are in effect multiple error bars with different line thicknesses plotted on top of each other.

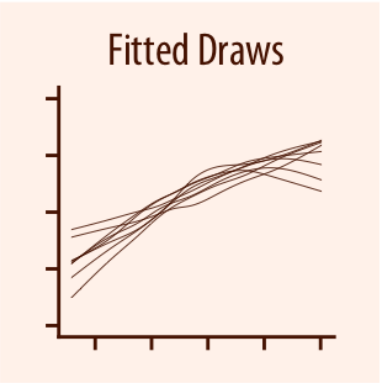
More detail than error bars



- We can visualize the actual confidence or posterior distributions
- Confidence strips provide a clear visual sense of uncertainty but are difficult to read accurately.
- Eyes and half-eyes combine error bars with approaches to visualize distributions (violins and ridgelines, respectively)
- By showing the distribution in discrete units, the quantile dot plot is not as precise but can be easier to read than the continuous distribution shown by a violin or ridgeline plot.

Confidence bands





- For smooth line graphs, the equivalent of an error bar is a confidence band