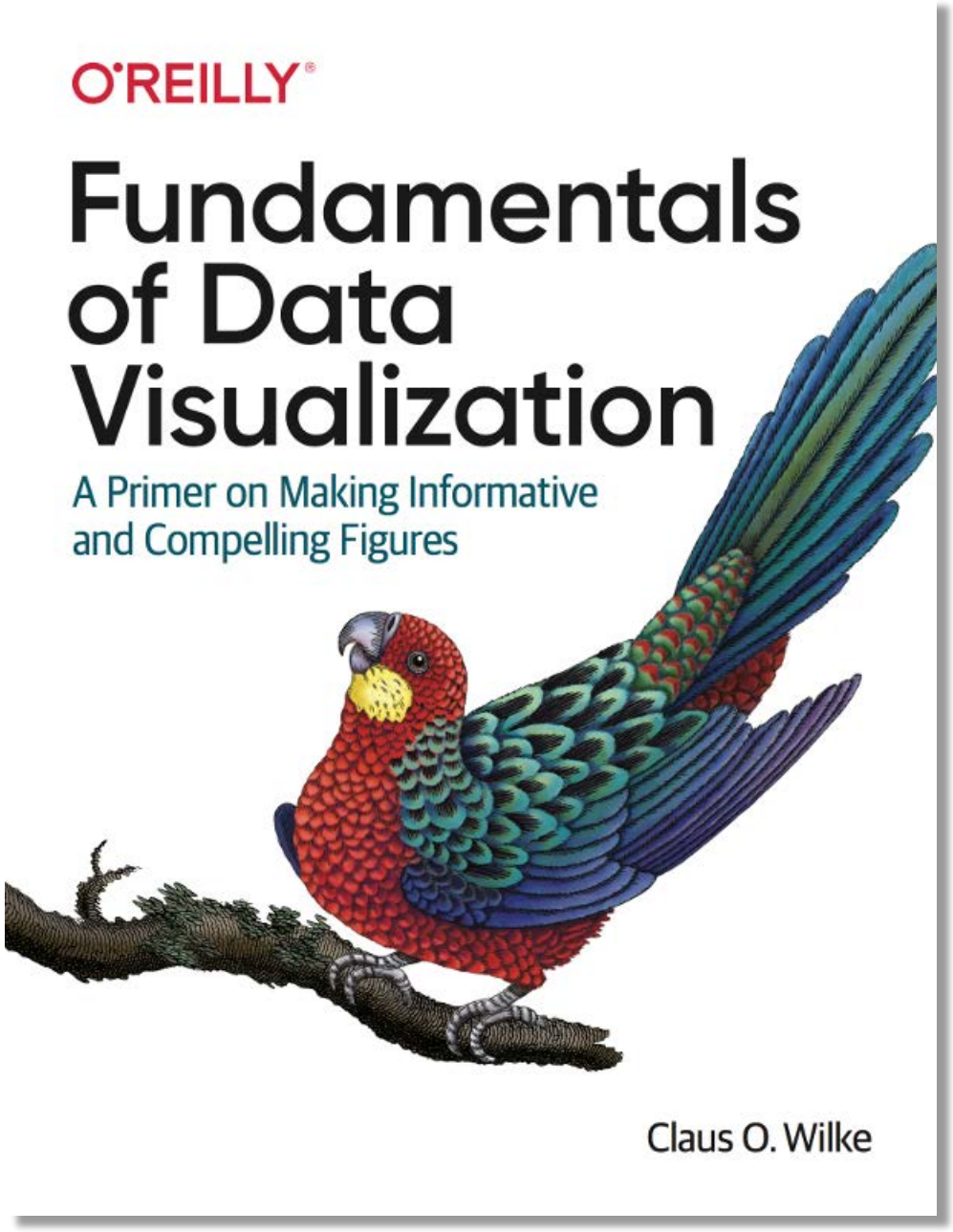
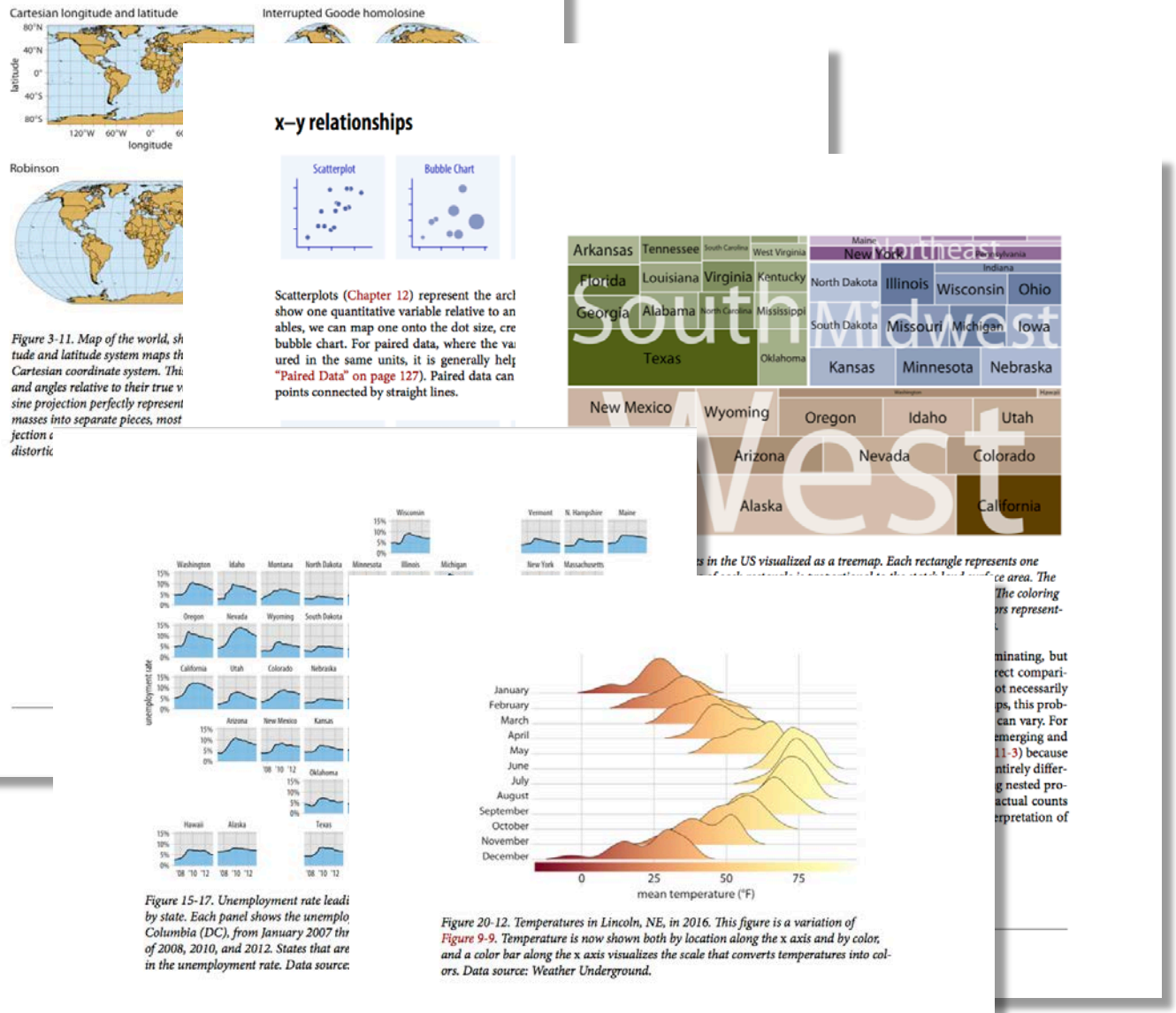
Data Visualization

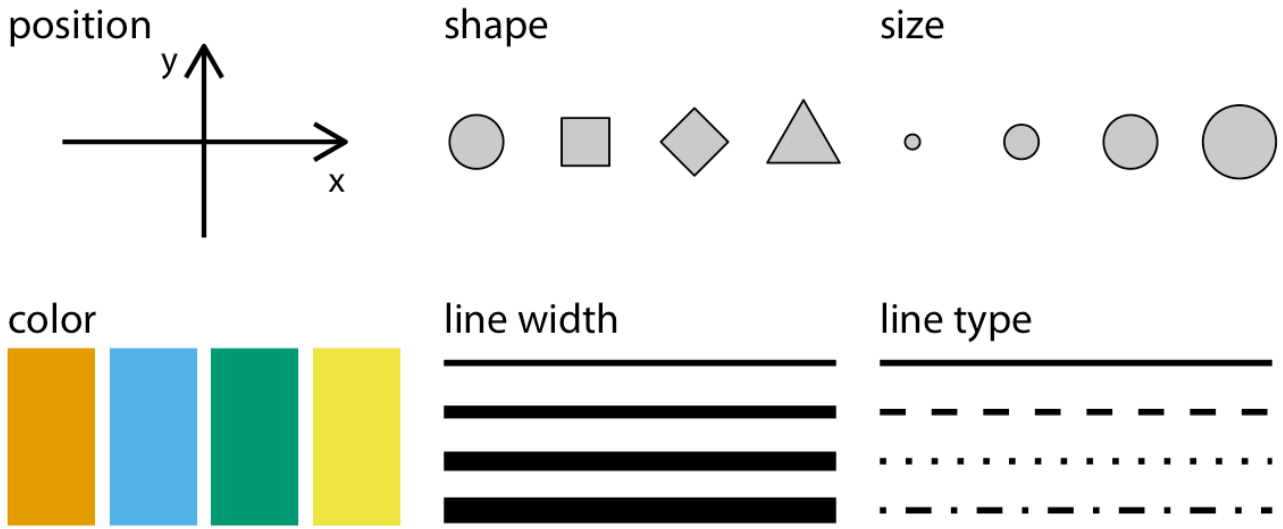




serialmentor.com/dataviz

Visualizing data: Mapping data onto aesthetics

All data visualizations map data values into quantifiable features of the resulting graphic. We refer to these features as *aesthetics.*



Visualizing data: Mapping data onto aesthetics

- **All aesthetics fall into one of two groups: Those that can represent continuous data and those that can not. Continuous data values are values for which arbitrarily fine intermediates exist.** For example, time duration is a continuous value. Between any two durations, say 50 seconds and 51 seconds, there are arbitrarily many intermediates, such as 50.5 seconds, 50.51 seconds, 50.50001 seconds, and so on. By contrast, number of persons in a room is a discrete value. A room can hold 5 persons or 6, but not 5.5. For the example, position, size, color, and line width can represent continuous data, but shape and line type can usually only represent discrete data.
- Next we'll consider the types of data we may want to represent in our visualization. You may **think of data as numbers**, but numerical values are only two out of several types of data we may encounter. In addition to continuous and discrete numerical values, data can come in the form of discrete categories, in the form of dates or times. When data is numerical we also call it ***quantitative*** and when it is categorical we call it ***qualitative***. Variables holding qualitative data are ***factors***, and the different categories are called ***levels***. The levels of a factor are most commonly without order (as in the example of "dog", "cat", "fish"), but factors can also be ordered, when there is an intrinsic order among the levels of the factor (as in the example of "good", "fair", "poor").

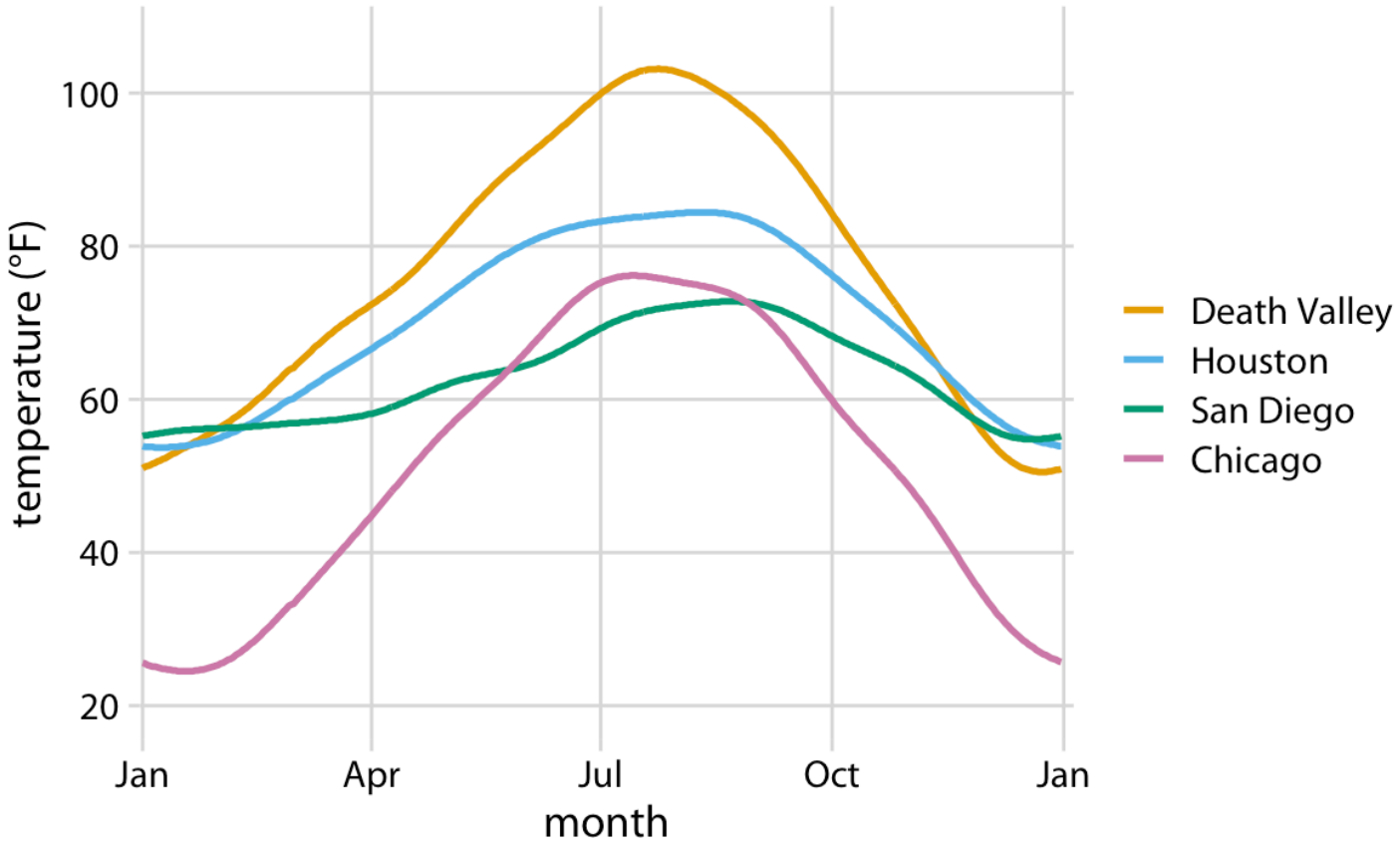
Visualizing data: Mapping data onto aesthetics

Type of variable	Examples	Appropriate scale	Description
quantitative/numerical continuous	1.3, 5.7, 83, $1.5x10^{-2}$	continuous	Arbitrary numerical values. These can be integers, rational numbers, or real numbers.
quantitative/numerical discrete	1, 2, 3, 4	discrete	Numbers in discrete units. These are most commonly but not necessarily integers. For example, the numbers 0.5, 1.0, 1.5 could also be treated as discrete if intermediate values cannot exist in the given dataset.
qualitative/categorical unordered	dog, cat, fish	discrete	Categories without order. These are discrete and unique categories that have no inherent order. These variables are also called *factors*.
qualitative/categorical ordered	good, fair, poor	discrete	Categories with order. These are discrete and unique categories with an order. For example, "fair" always lies between "good" and "poor". These variables are also called *ordered factors*.
date or time	Jan. 5 2018, 8:03am	continuous or discrete	Specific days and/or times. Also generic dates, such as July 4 or Dec. 25 (without year).
text	The quick brown fox jumps over the lazy dog.	none, or discrete	Free-form text. Can be treated as categorical if needed.

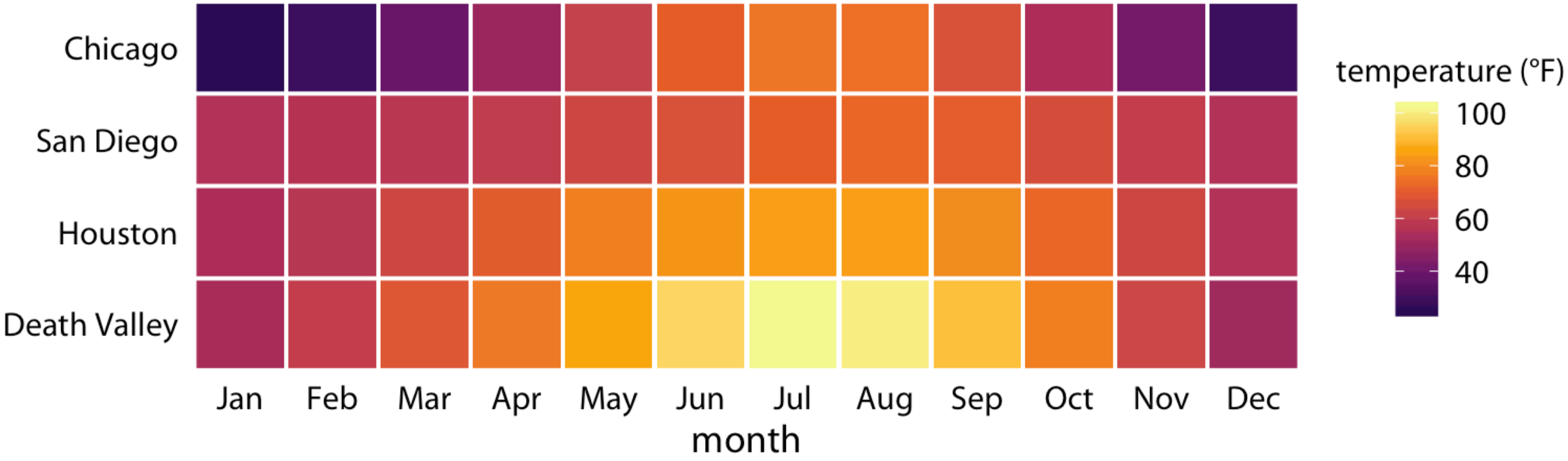
Visualizing data: Mapping data onto aesthetics

Month	Day	Location	Station ID	Temperature
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319	51.2
Jan	3	Chicago	USW00014819	25.3
Jan	3	San Diego	USW00093107	55.3
Jan	3	Death Valley	USC00042319	51.3
Jan	3	Houston	USW00012918	53.8

Visualizing data: Mapping data onto aesthetics

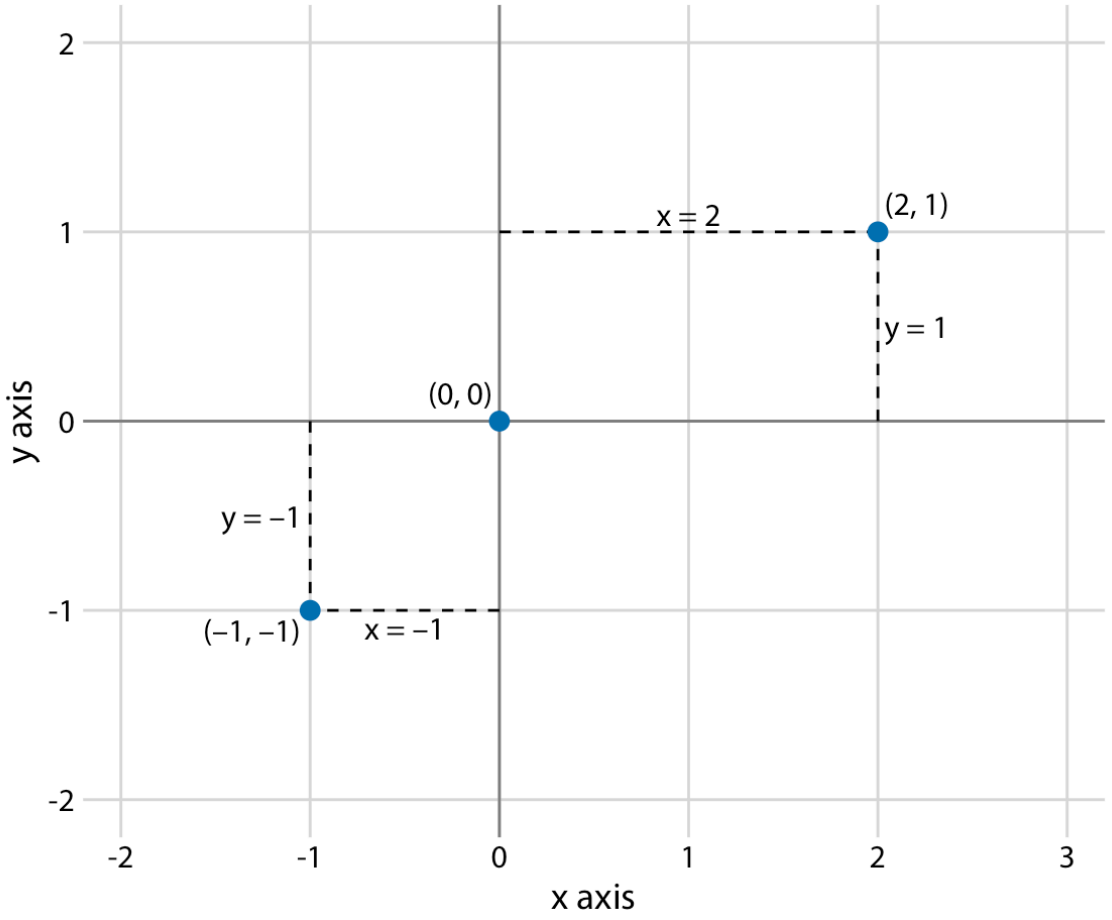


Visualizing data: Mapping data onto aesthetics



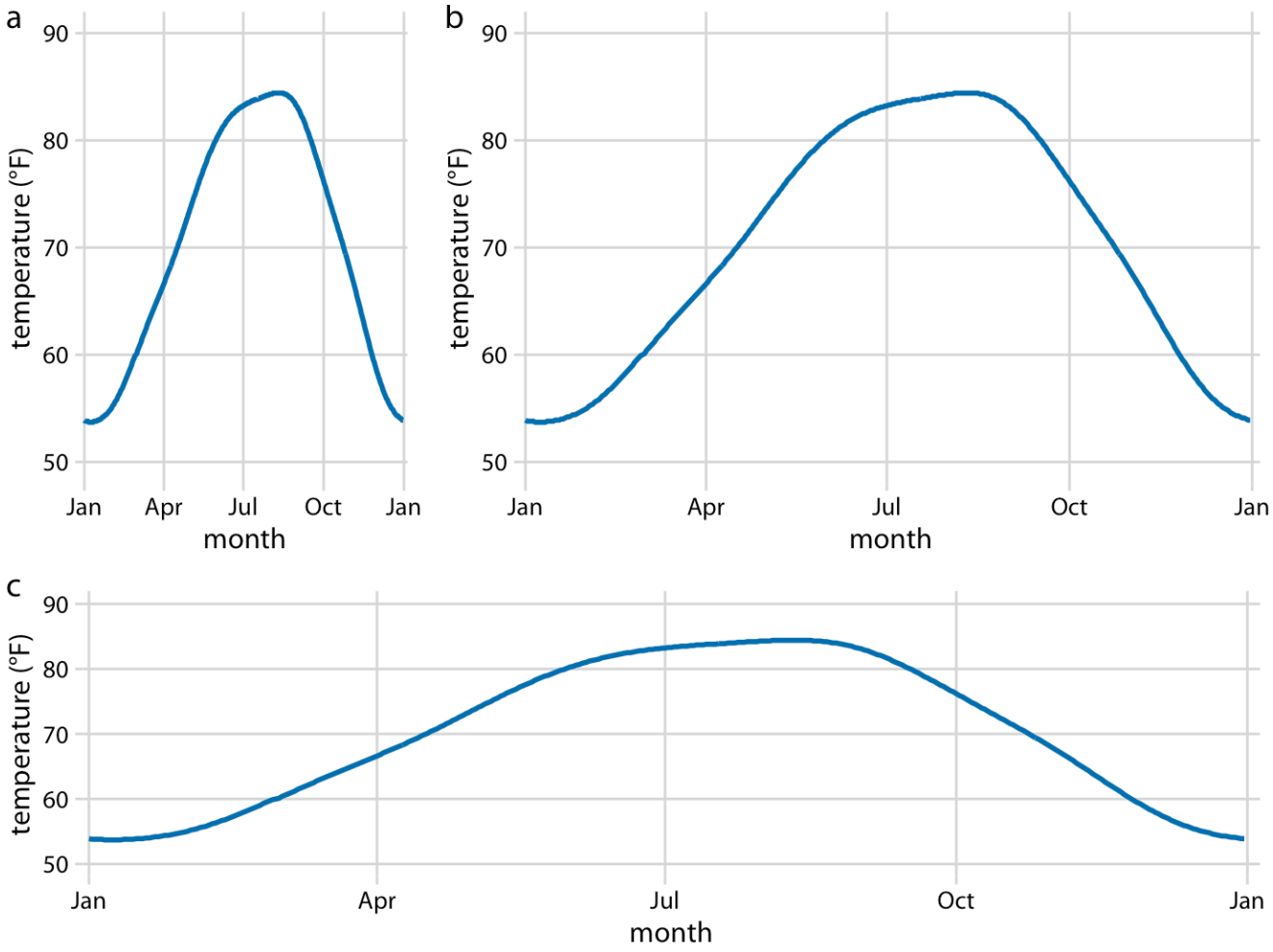
Coordinate systems and axes

The combination of a set of position scales and their relative geometric arrangement is called a *coordinate system.*



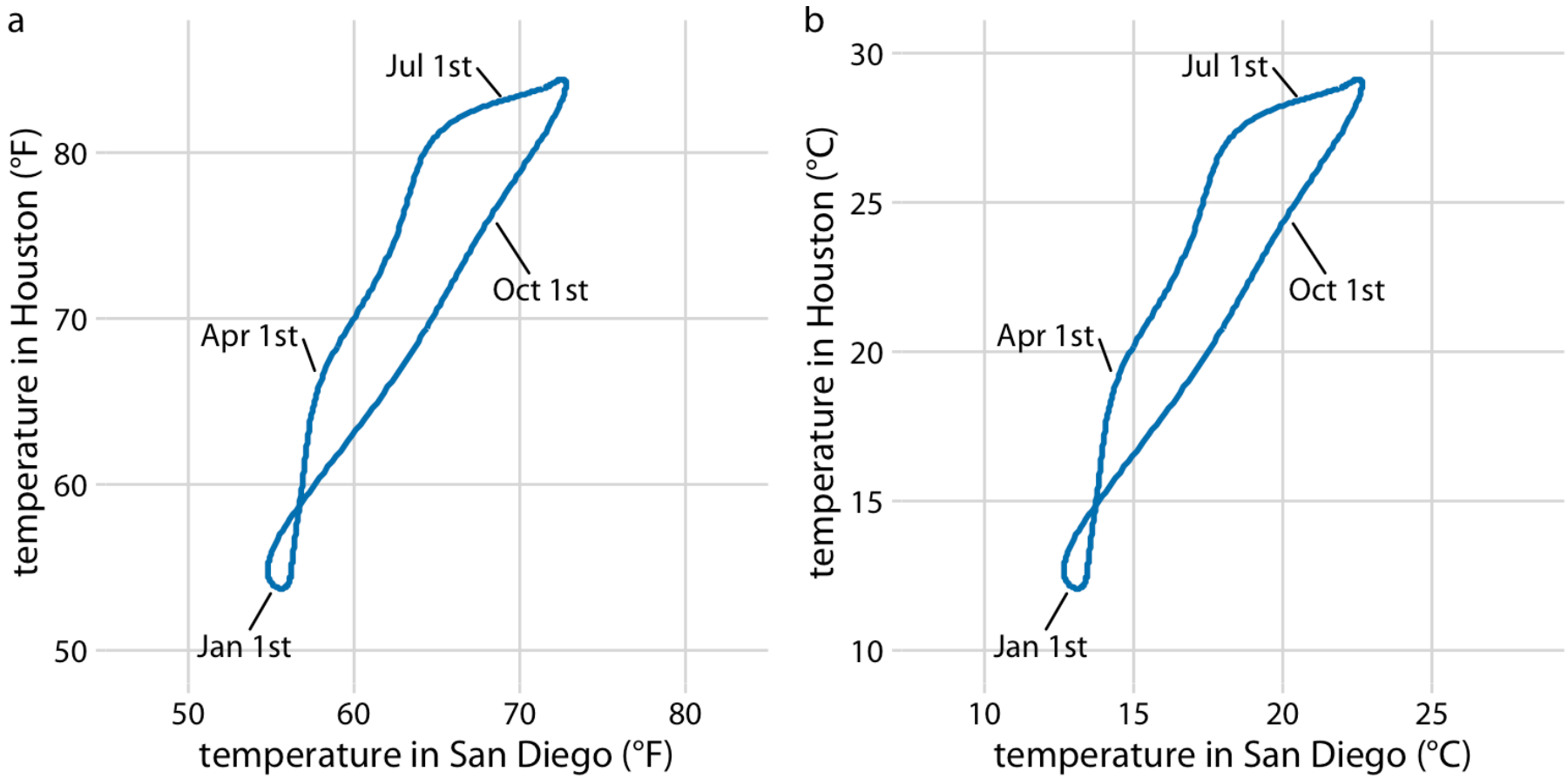
Coordinate systems and axes

Good practice: represent axes with different units with different grid sizes

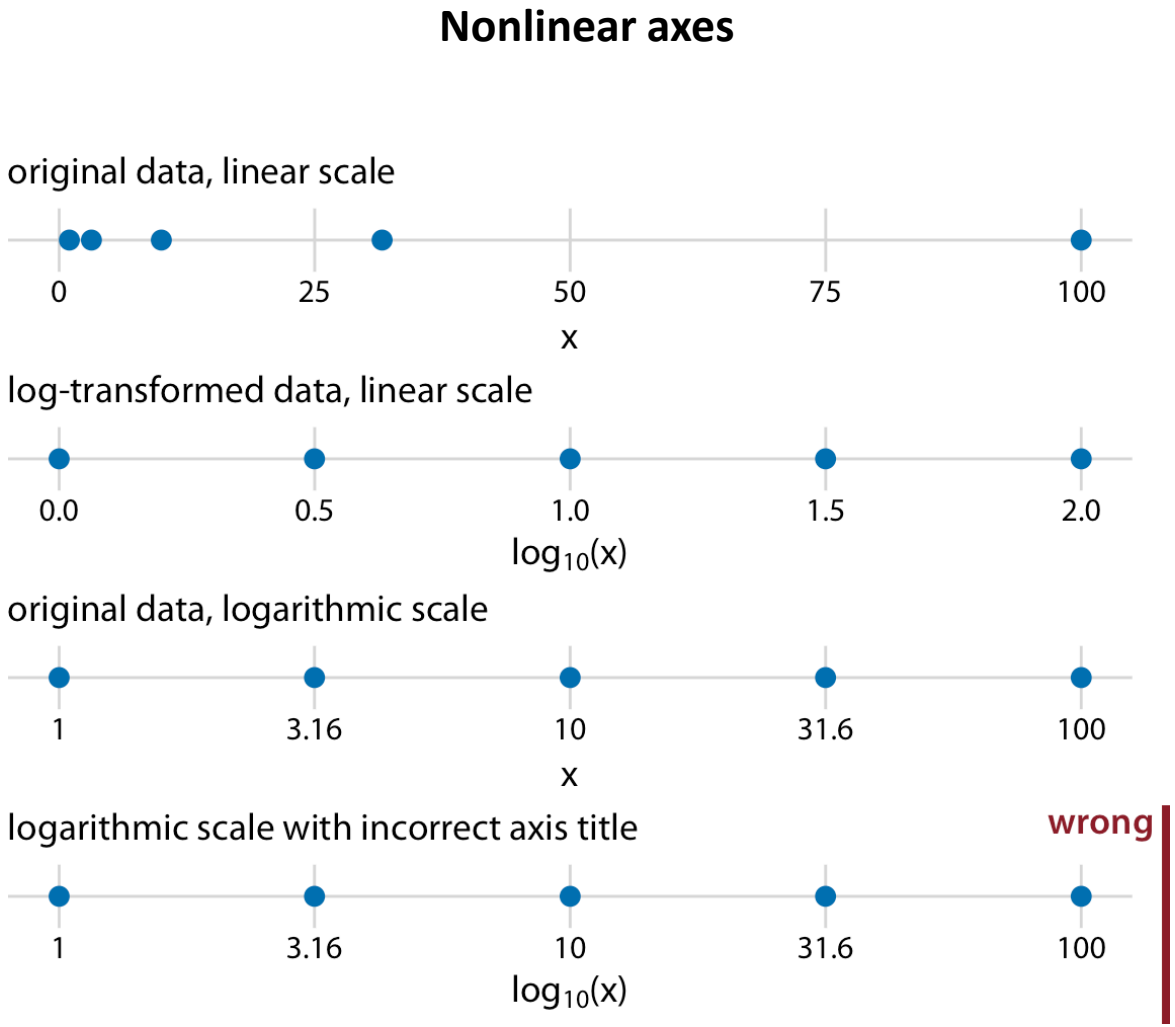


Coordinate systems and axes

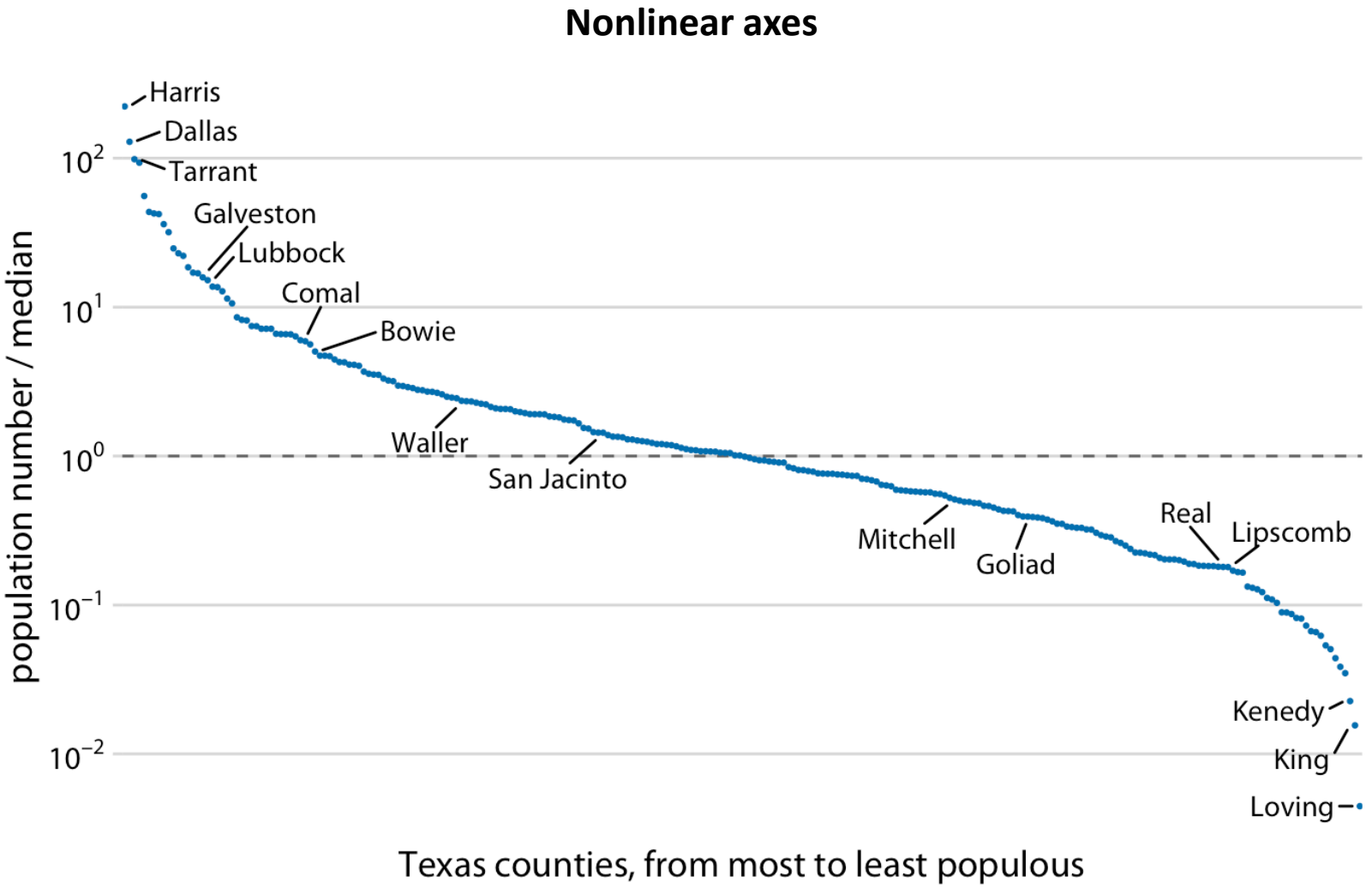
Good practice: represent axes with the same units with the same grid sizes



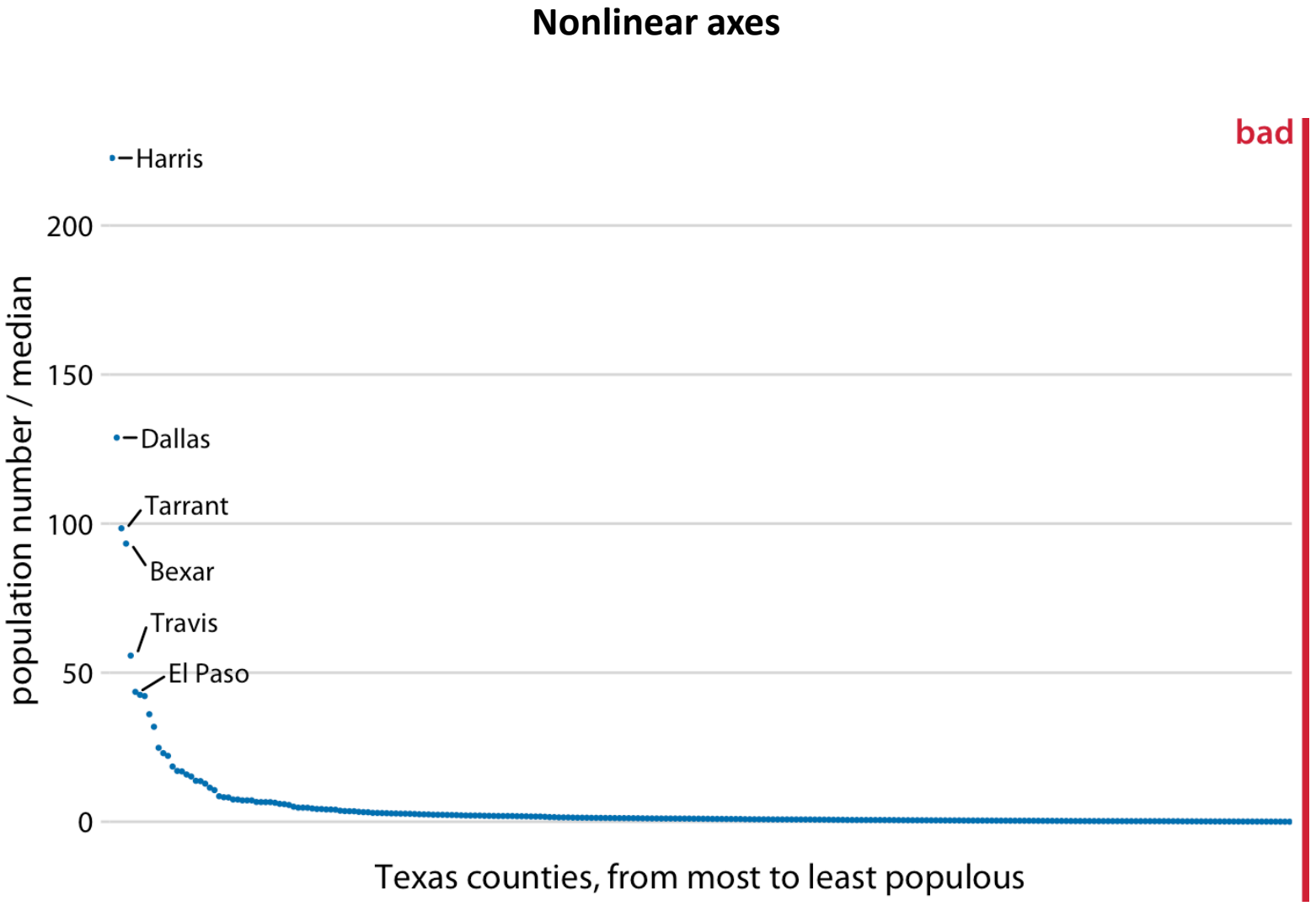
Coordinate systems and axes



Coordinate systems and axes

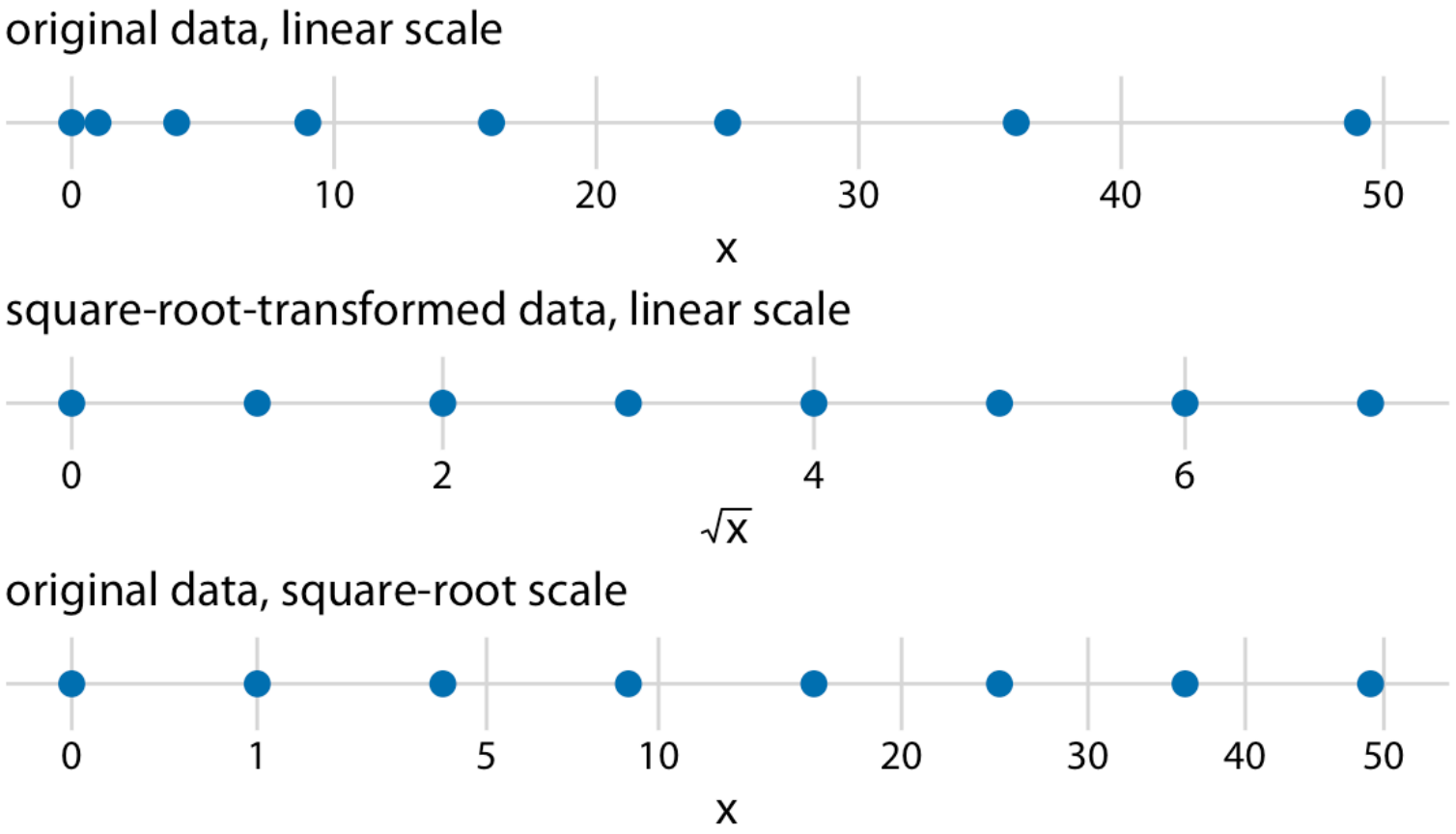


Coordinate systems and axes



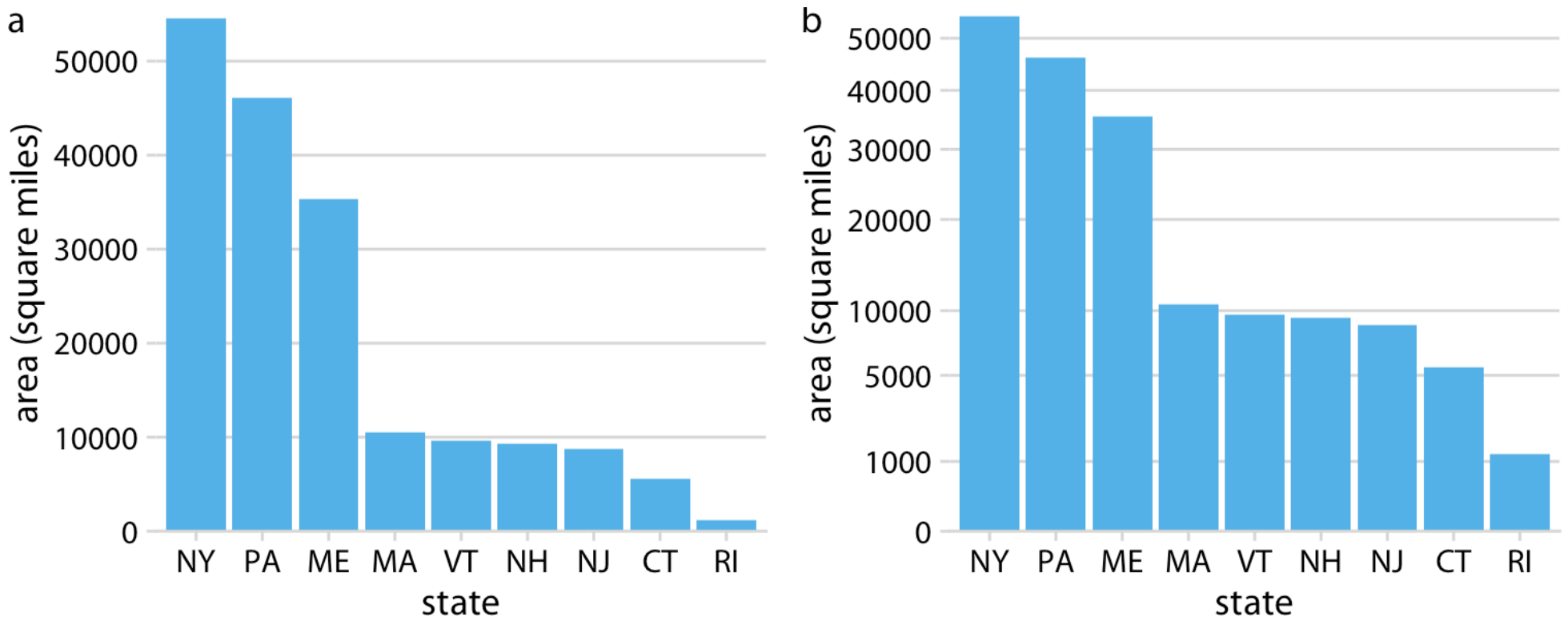
Coordinate systems and axes

Nonlinear axes



Coordinate systems and axes

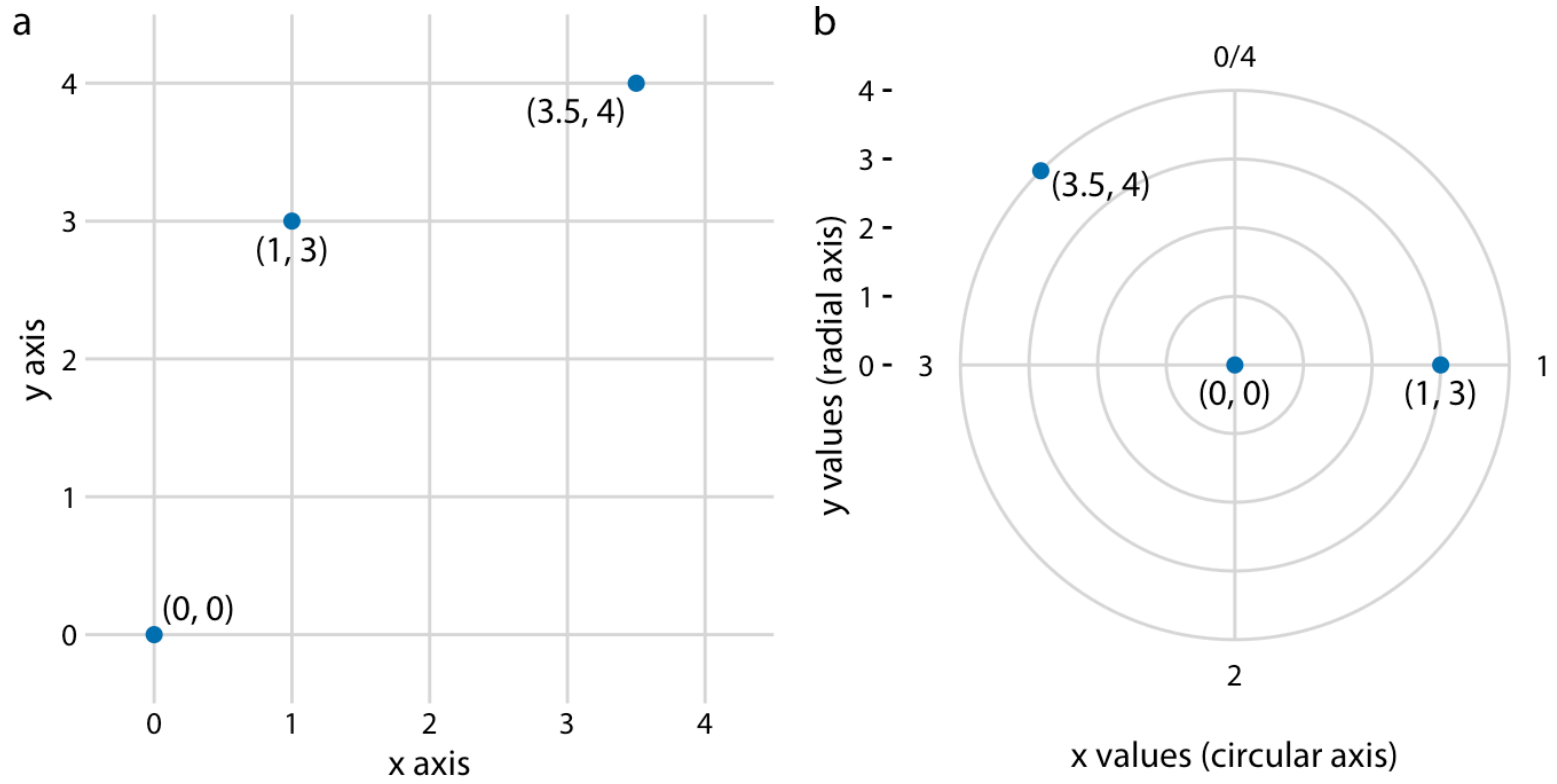
Nonlinear axes



Areas of Northeastern U.S. states. (a) Areas shown on a linear scale. (b) Areas shown on a square-root scale.

Coordinate systems with curved axes

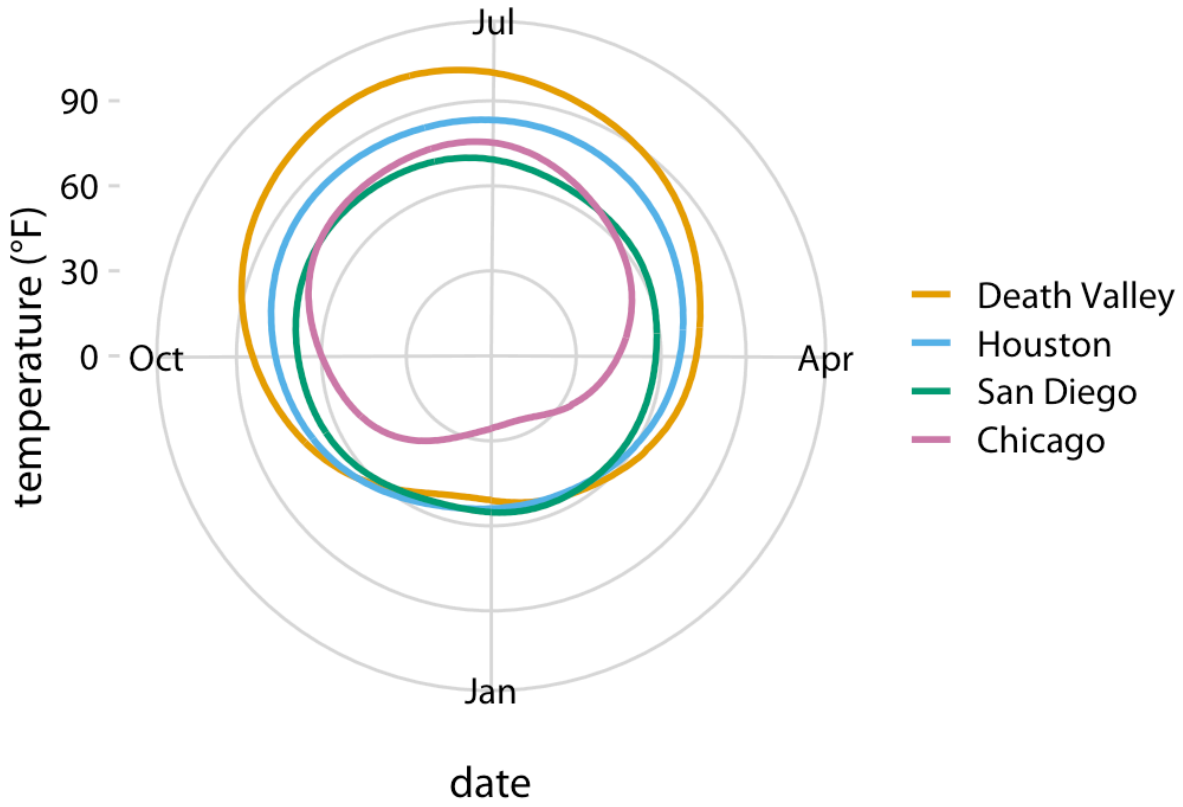
Polar coordinates can be useful for data of a periodic nature, such that data values at one end of the scale can be logically joined to data values at the other end.



Relationship between Cartesian and polar coordinates. (a) Three data points shown in a Cartesian coordinate system. (b) The same three data points shown in a polar coordinate system. We have taken the *x* coordinates from part (a) and used them as angular coordinates and the *y* coordinates from part (a) and used them as radial coordinates. The circular axis runs from 0 to 4 in this example, and therefore $x = 0$ and $x = 4$ are the same locations in this coordinate system.

Coordinate systems with curved axes

Polar coordinates can be useful for data of a periodic nature, such that data values at one end of the scale can be logically joined to data values at the other end.

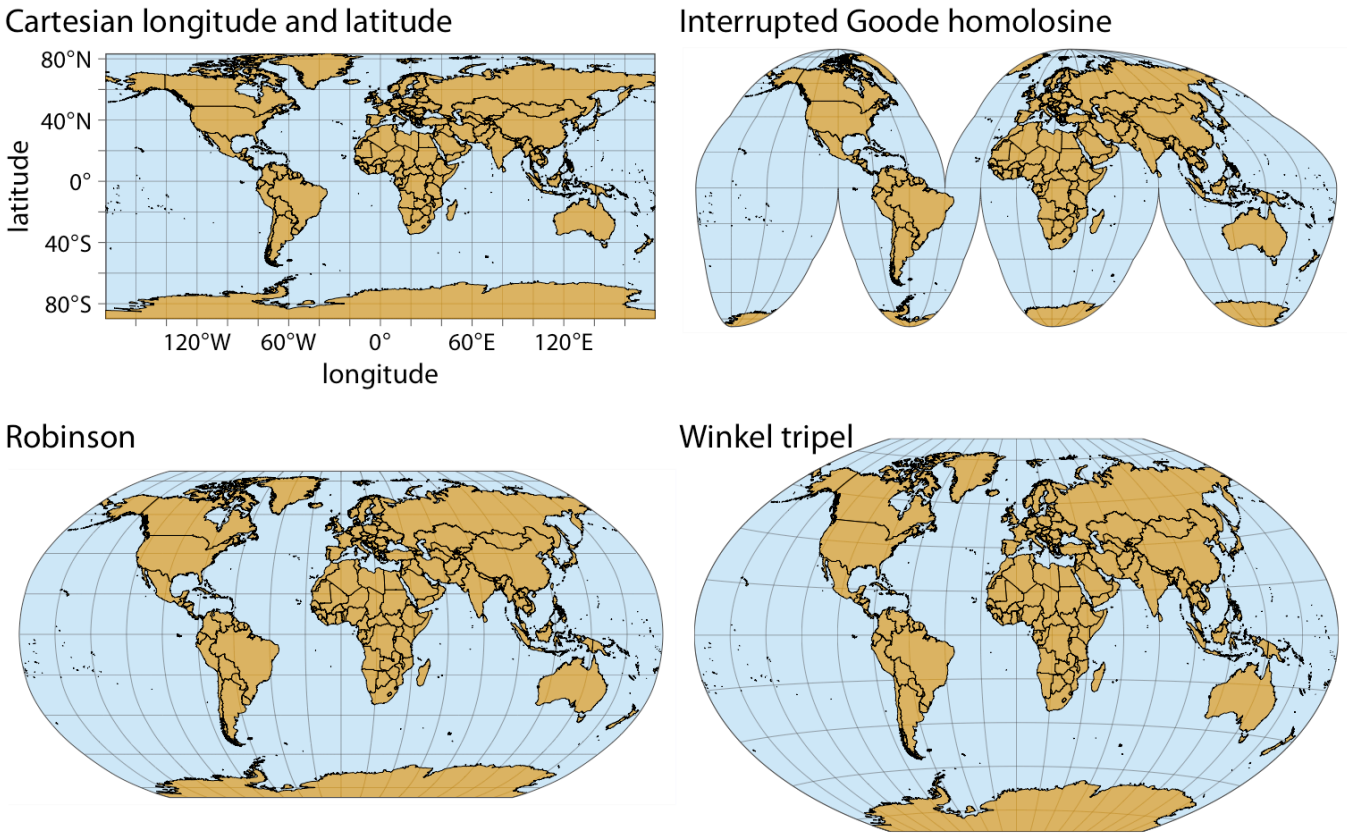


Daily temperature normals for four selected locations in the U.S., shown in polar coordinates. The radial distance from the center point indicates the daily temperature in Fahrenheit, and the days of the year are arranged counter-clockwise starting with Jan. 1st at the 6:00 position.

Coordinate systems with curved axes

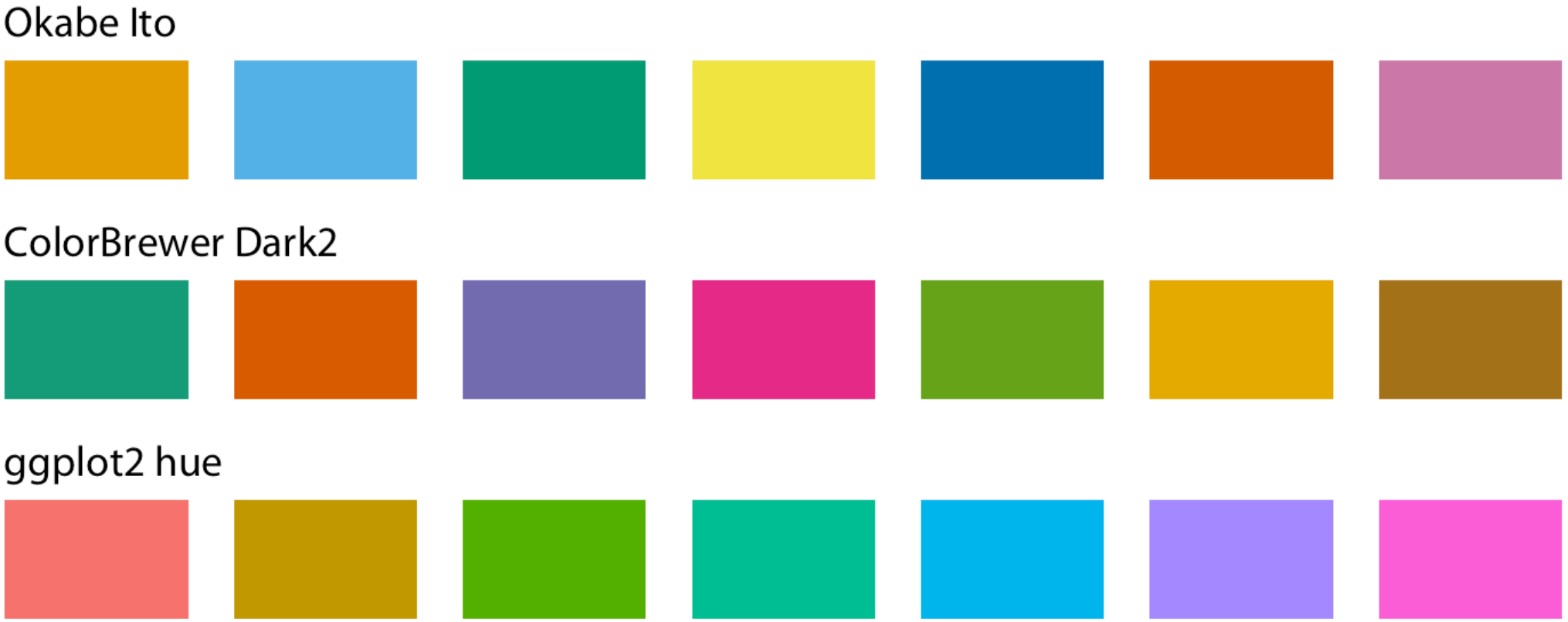
A second setting in which we encounter curved axes is in the context of geospatial data, i.e., maps. Locations on the globe are specified by their longitude and latitude.

Map of the world, shown in four different projections. The Cartesian longitude and latitude system maps the longitude and latitude of each location onto a regular Cartesian coordinate system. This mapping causes substantial distortions in both areas and angles relative to their true values on the 3D globe. The interrupted Goode homolosine projection perfectly represents true surface areas, at the cost of dividing some land masses into separate pieces, most notably Greenland and Antarctica. The Robinson projection and the Winkel tripel projection both strike a balance between angular and area distortions, and they are commonly used for maps of the entire globe.



Color scales

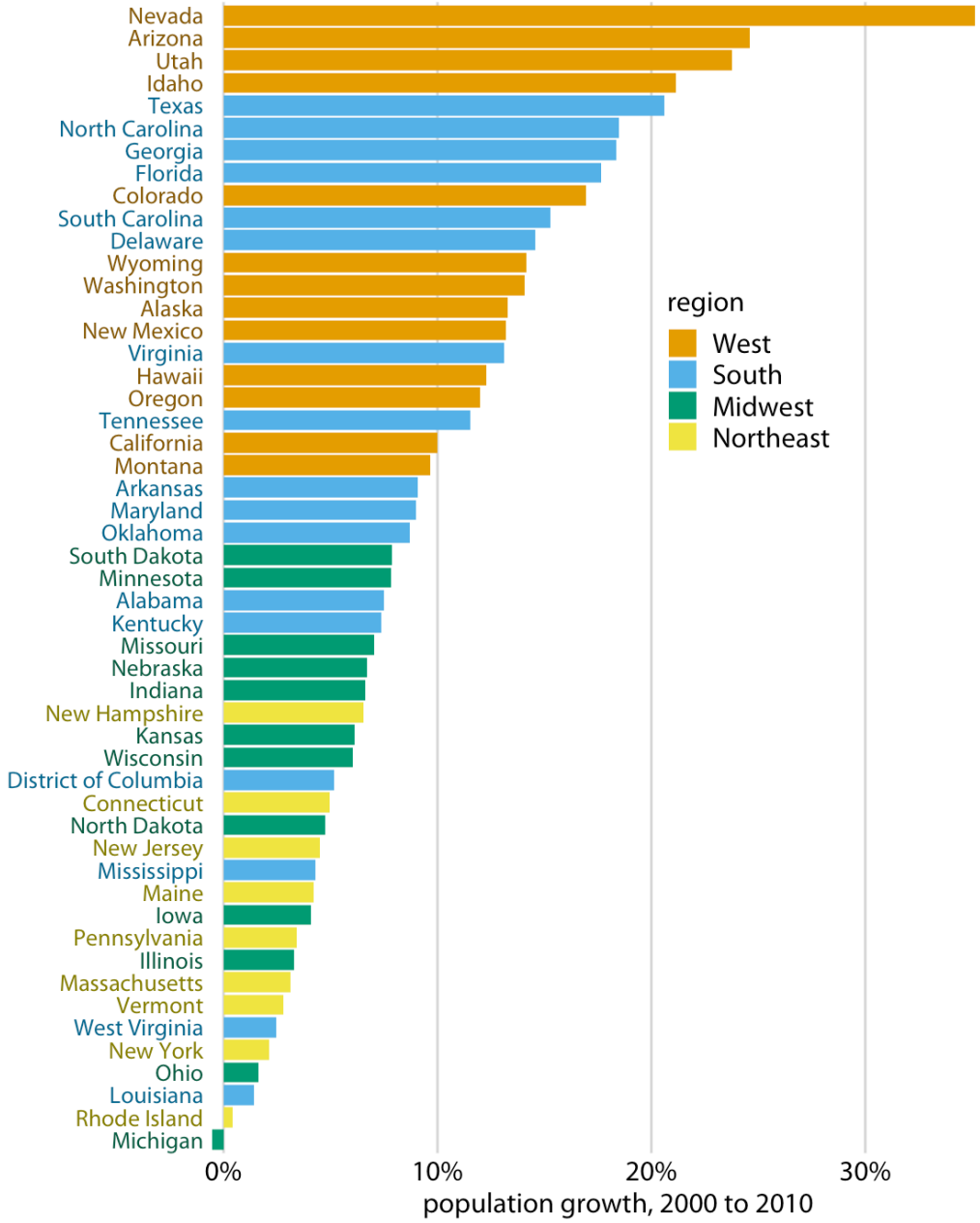
qualitative color scale: a means to distinguish discrete items or groups that do not have an **intrinsic order**, such as different countries on a map or different manufacturers of a certain product.



Color scales

qualitative color scale: a means to distinguish discrete items or groups that do not have an intrinsic order, such as different countries on a map or different manufacturers of a certain product.

Population growth in the U.S. from 2000 to 2010. States in the West and South have seen the largest increases, whereas states in the Midwest and Northeast have seen much smaller increases or even, in the case of Michigan, a decrease. Data source: U.S. Census Bureau



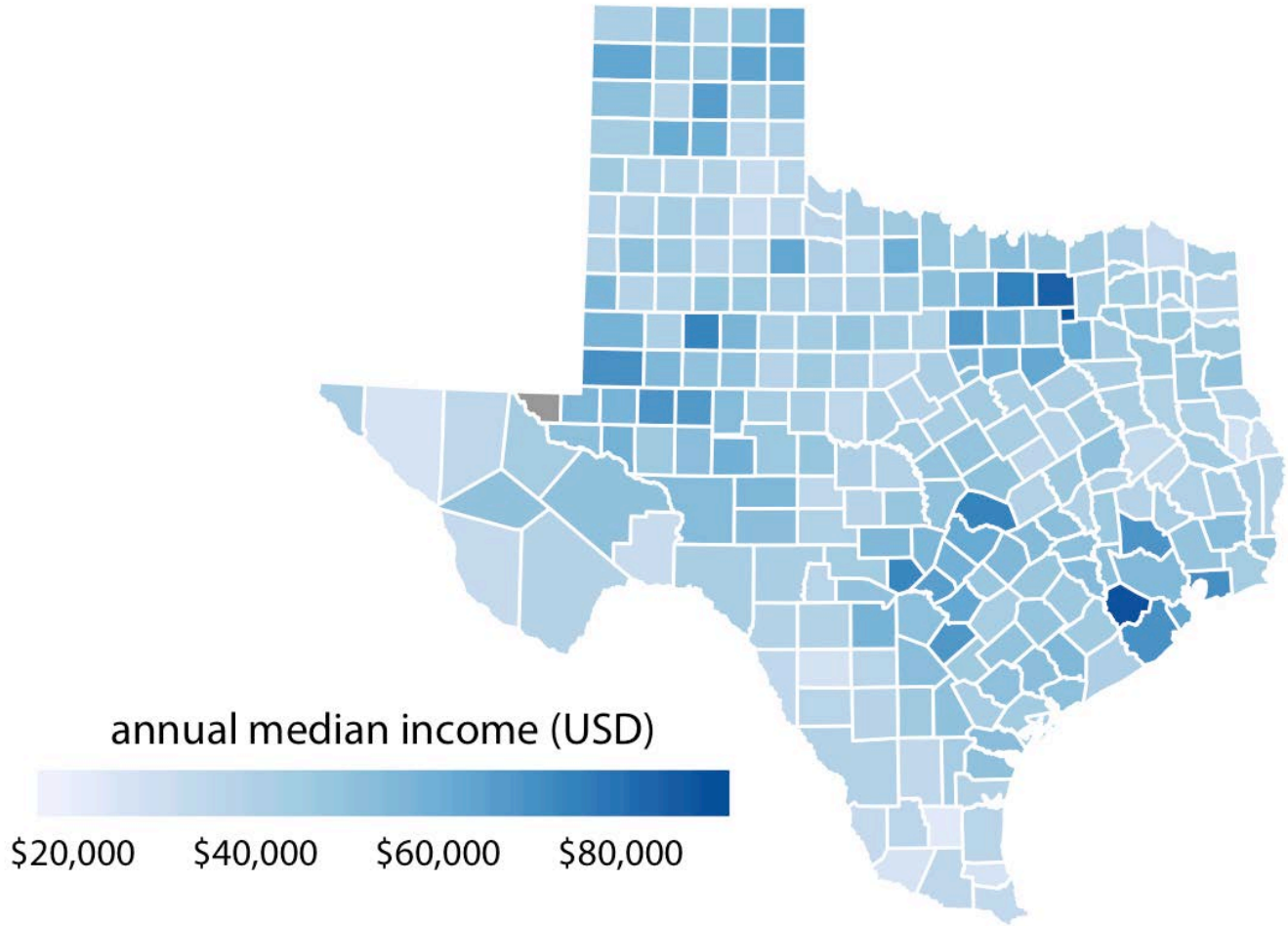
Color scales

Color to represent data values: Color can also be used to represent data values, such as income, temperature, or speed. In this case, we use a *sequential* color scale. Such a scale contains a sequence of colors that clearly indicate (i) which values are larger or smaller than which other ones and (ii) how distant two specific values are from each other. The second point implies that the color scale needs to be perceived to vary uniformly across its entire range.

Example sequential color scales. The ColorBrewer Blues scale is a monochromatic scale that varies from dark to light blue. The Heat and Viridis scales are multi-hue scales that vary from dark red to light yellow and from dark blue via green to light yellow, respectively.

Color scales

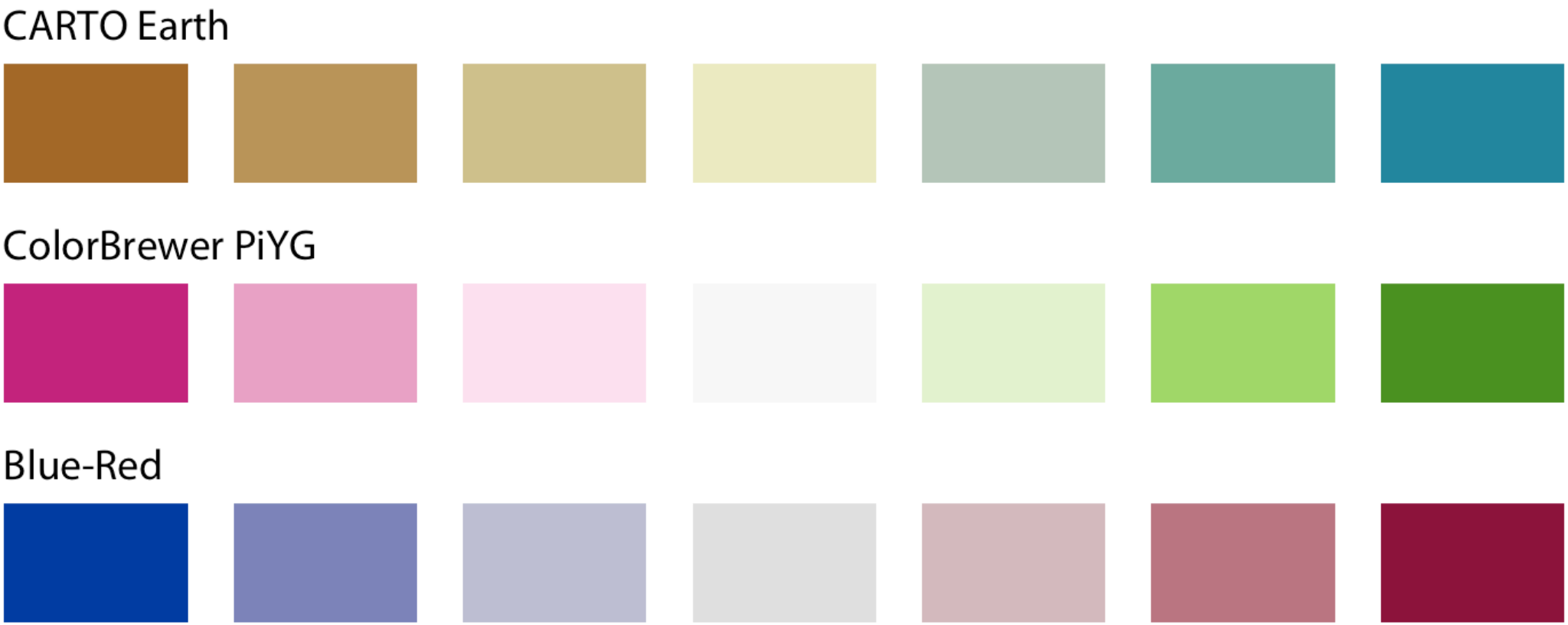
Color to represent data values: Color can also be used to represent data values, such as income, temperature, or speed. In this case, we use a *sequential* color scale. Such a scale contains a sequence of colors that clearly indicate (i) which values are larger or smaller than which other ones and (ii) how distant two specific values are from each other. The second point implies that the color scale needs to be perceived to vary uniformly across its entire range.



Median annual income in Texas counties. The highest median incomes are seen in major Texas metropolitan areas, in particular near Houston and Dallas. No median income estimate is available for Loving County in West Texas and therefore that county is shown in gray. Data source: 2015 Five-Year American Community Survey

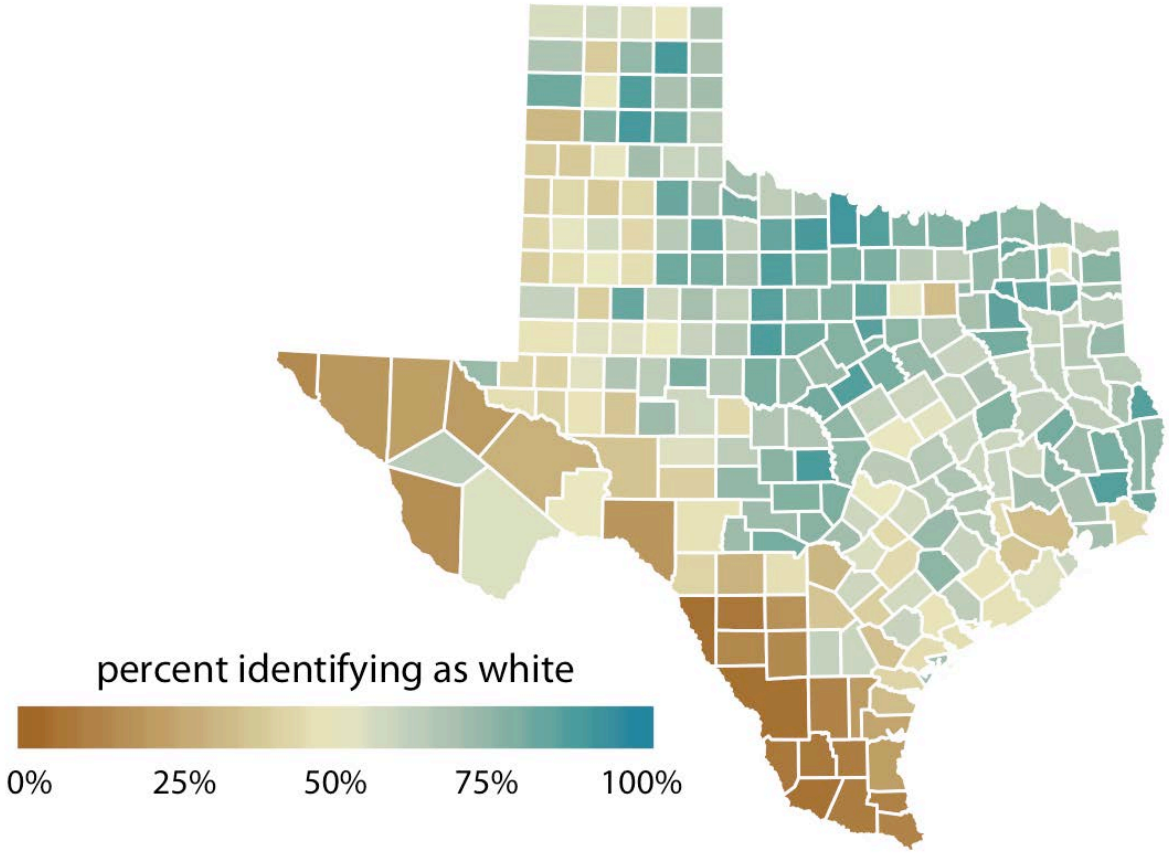
Color scales

In some cases, we need to visualize the deviation of data values in one of two directions relative to a neutral midpoint. One straightforward example is a dataset containing both positive and negative numbers. We may want to show those with different colors, so that it is immediately obvious whether a value is positive or negative as well as how far in either direction it deviates from zero. The appropriate color scale in this situation is a ***diverging* color scale**.



Color scales

In some cases, we need to visualize the deviation of data values in one of two directions relative to a neutral midpoint. One straightforward example is a dataset containing both positive and negative numbers. We may want to show those with different colors, so that it is immediately obvious whether a value is positive or negative as well as how far in either direction it deviates from zero. The appropriate color scale in this situation is a ***diverging* color scale**.



Percentage of people identifying as white in Texas counties. Whites are in the majority in North and East Texas but not in South or West Texas.

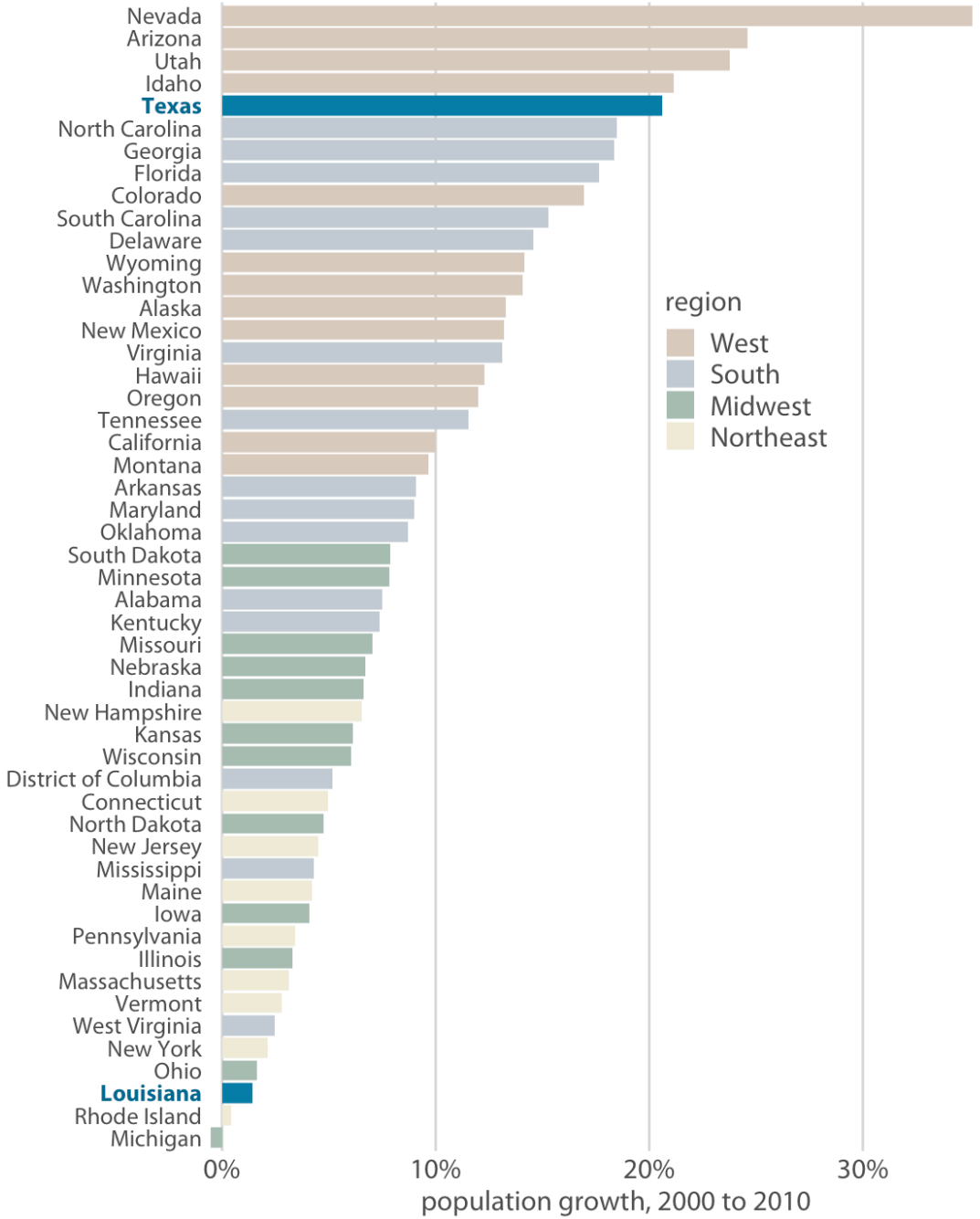
Color scales

Color as a tool to highlight: Color can also be an effective tool to highlight specific elements in the data. There may be specific categories or values in the dataset that carry key information about the story we want to tell, and we can strengthen the story by emphasizing the relevant figure elements to the reader. 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 are color scales that contain both **a set of subdued colors and a matching set of stronger, darker, and/or more saturated colors**.

Color scales

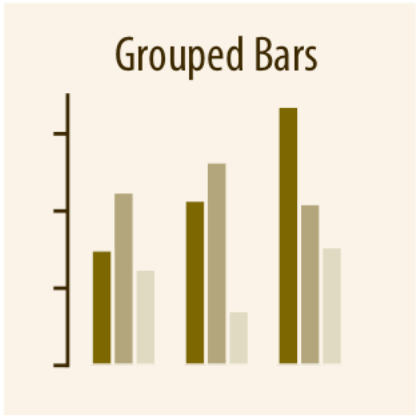
Color as a tool to highlight: Color can also be an effective tool to highlight specific elements in the data. There may be specific categories or values in the dataset that carry key information about the story we want to tell, and we can strengthen the story by emphasizing the relevant figure elements to the reader. 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 are color scales that contain both a set of subdued colors and a matching set of stronger, darker, and/or more saturated colors

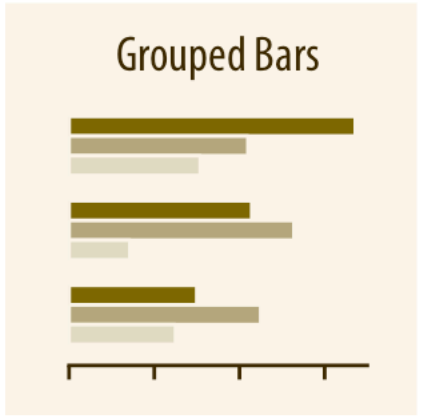
From 2000 to 2010, the two neighboring southern states Texas and Louisiana have experienced among the highest and lowest population growth across the U.S.

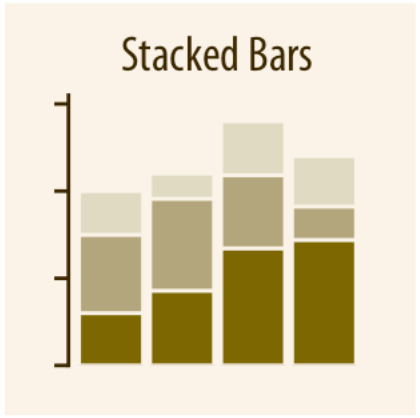


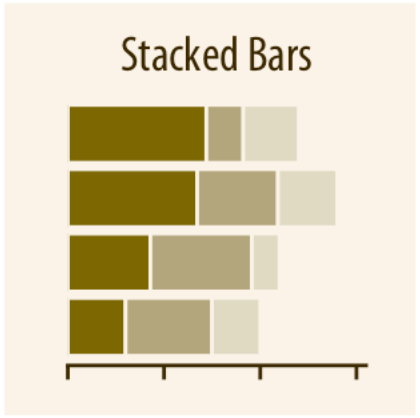
Directory of visualizations: Amounts

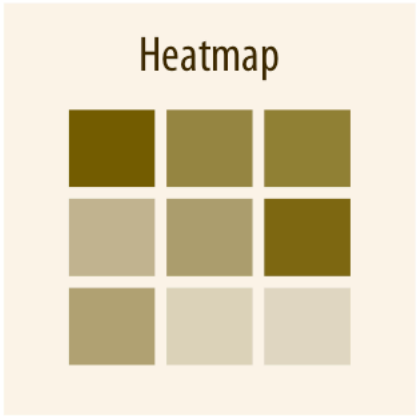
The most common approach to visualizing amounts (i.e., numerical values shown for some set of categories) is using bars, either vertically or horizontally arranged

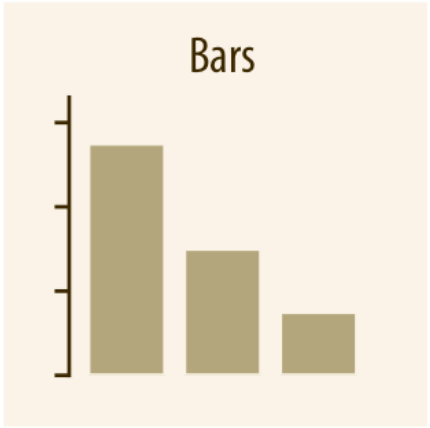


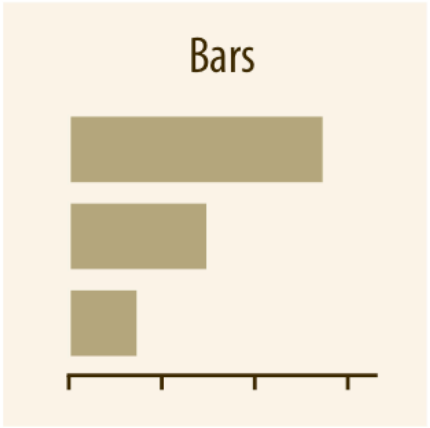


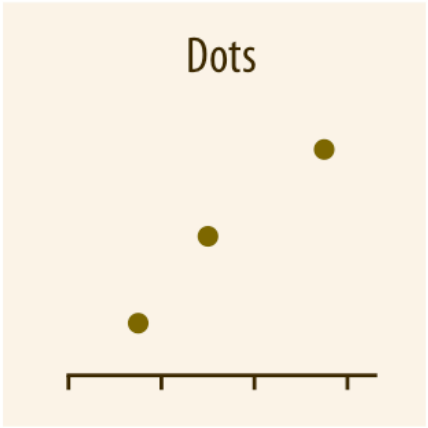






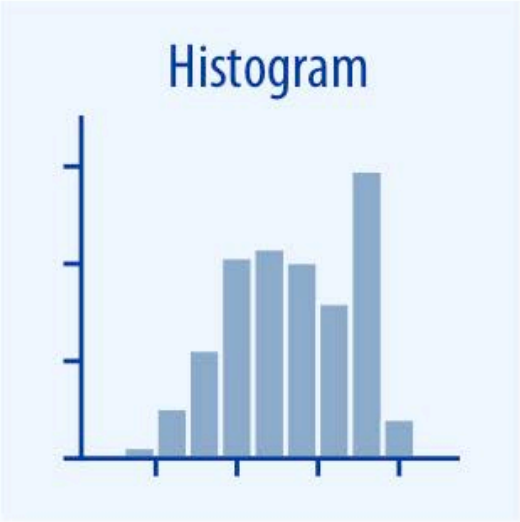


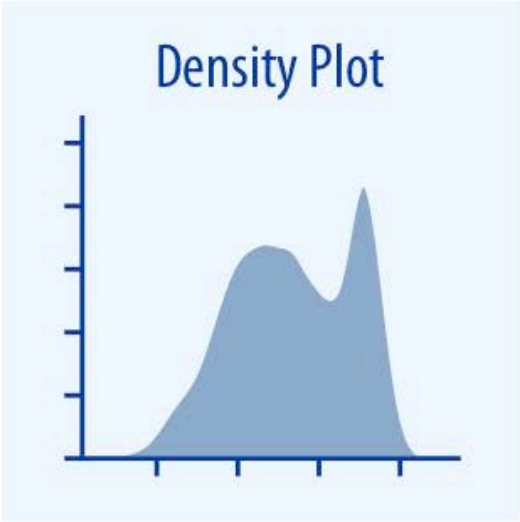


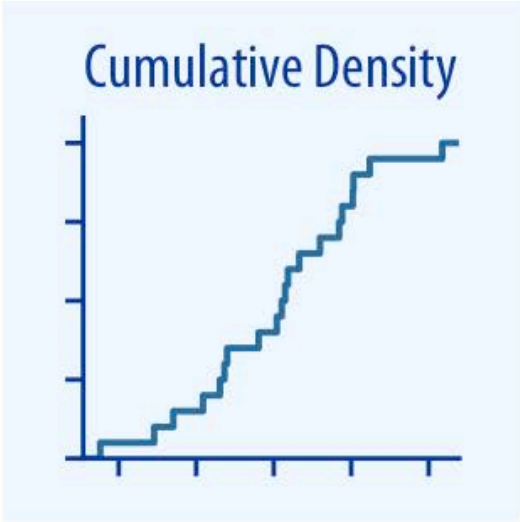


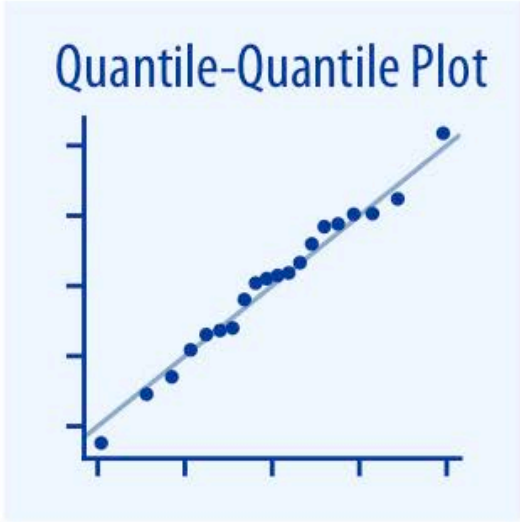
Directory of visualizations: Distributions

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.



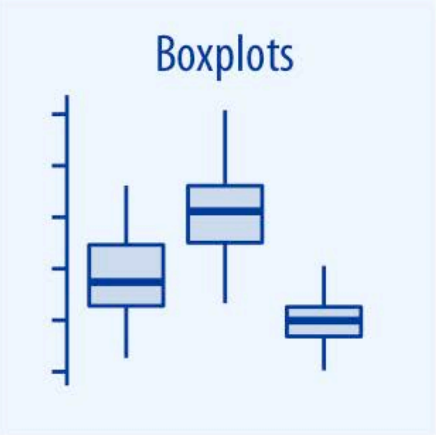


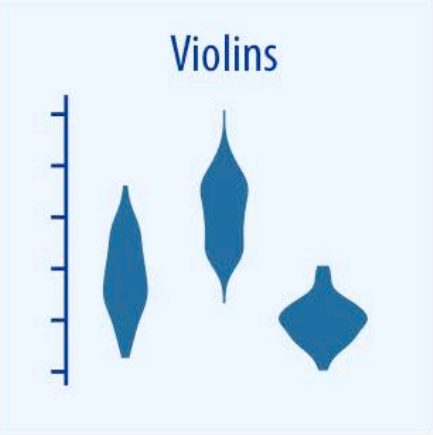


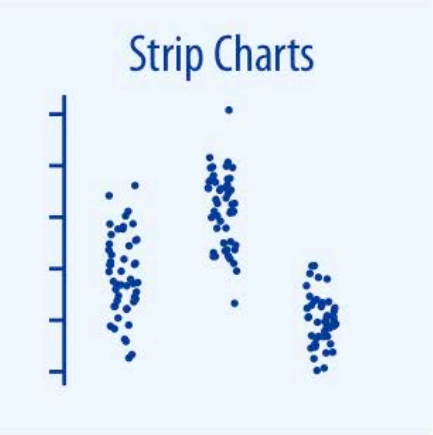


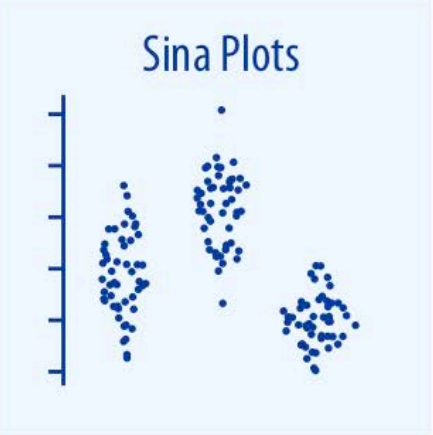
Directory of visualizations: Distributions

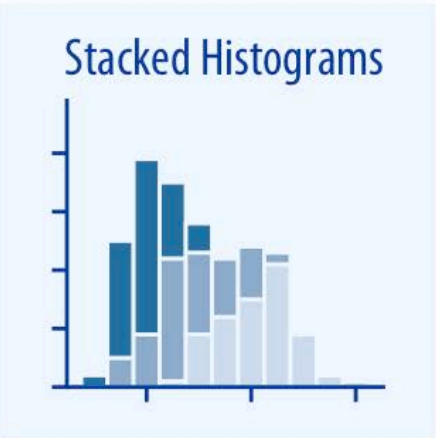
Histograms and density plots provide the most intuitive visualizations of a distribution, but both require arbitrary parameter choices and can be misleading. Cumulative densities and quantile-quantile (q-q) plots always represent the data faithfully but can be more difficult to interpret.

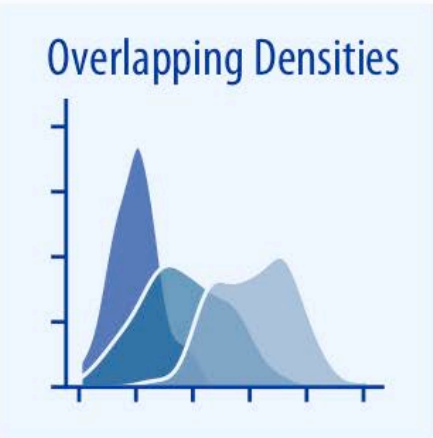


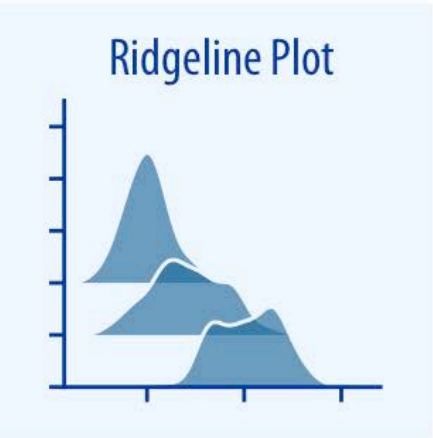






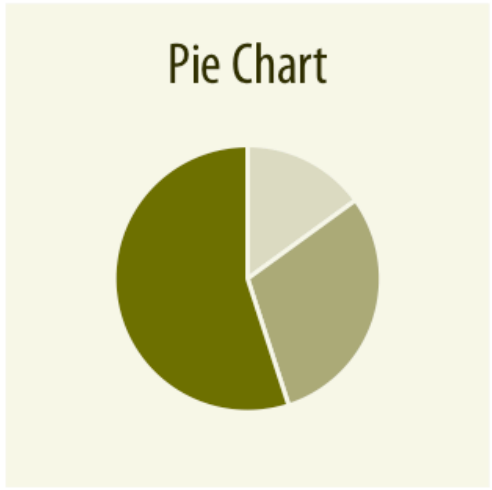


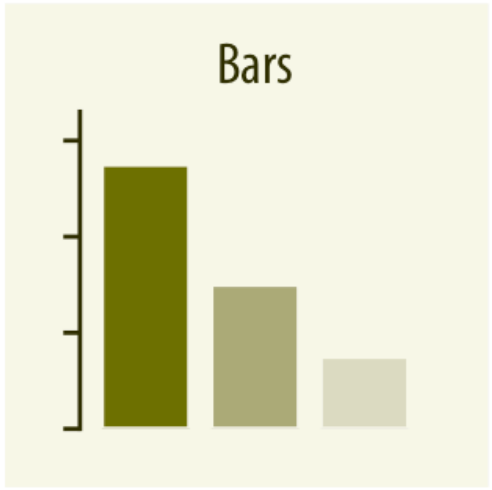


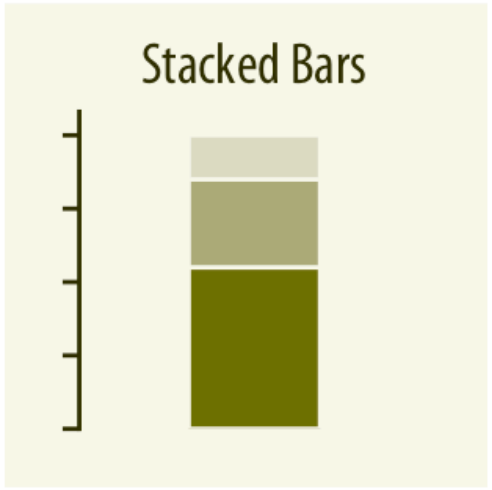


Directory of visualizations: Proportions

Proportions can be visualized as pie charts, side-by-side bars, or stacked bars and as in the case for amounts, bars can be arranged either vertically or horizontally. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars. Stacked bars look awkward for a single set of proportions, but can be useful when comparing multiple sets of proportions

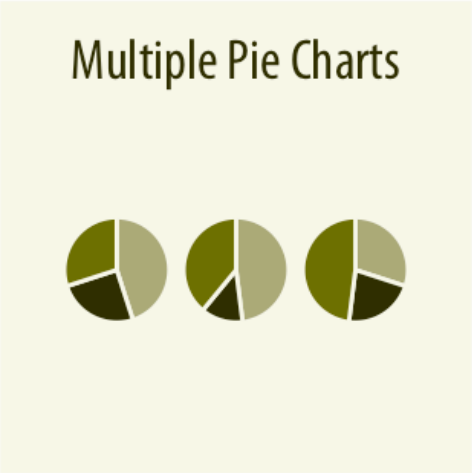


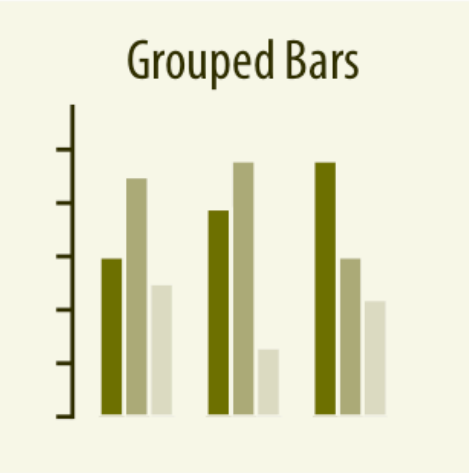


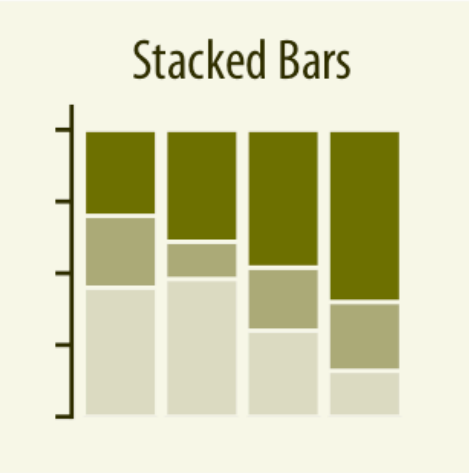


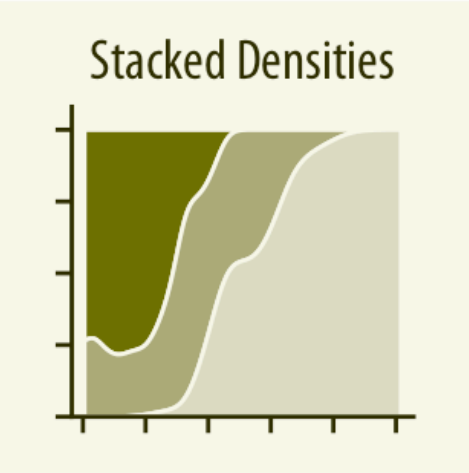
Directory of visualizations: Proportions

Proportions can be visualized as pie charts, side-by-side bars, or stacked bars and as in the case for amounts, bars can be arranged either vertically or horizontally. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars. Stacked bars look awkward for a single set of proportions, but can be useful when comparing multiple sets of proportions







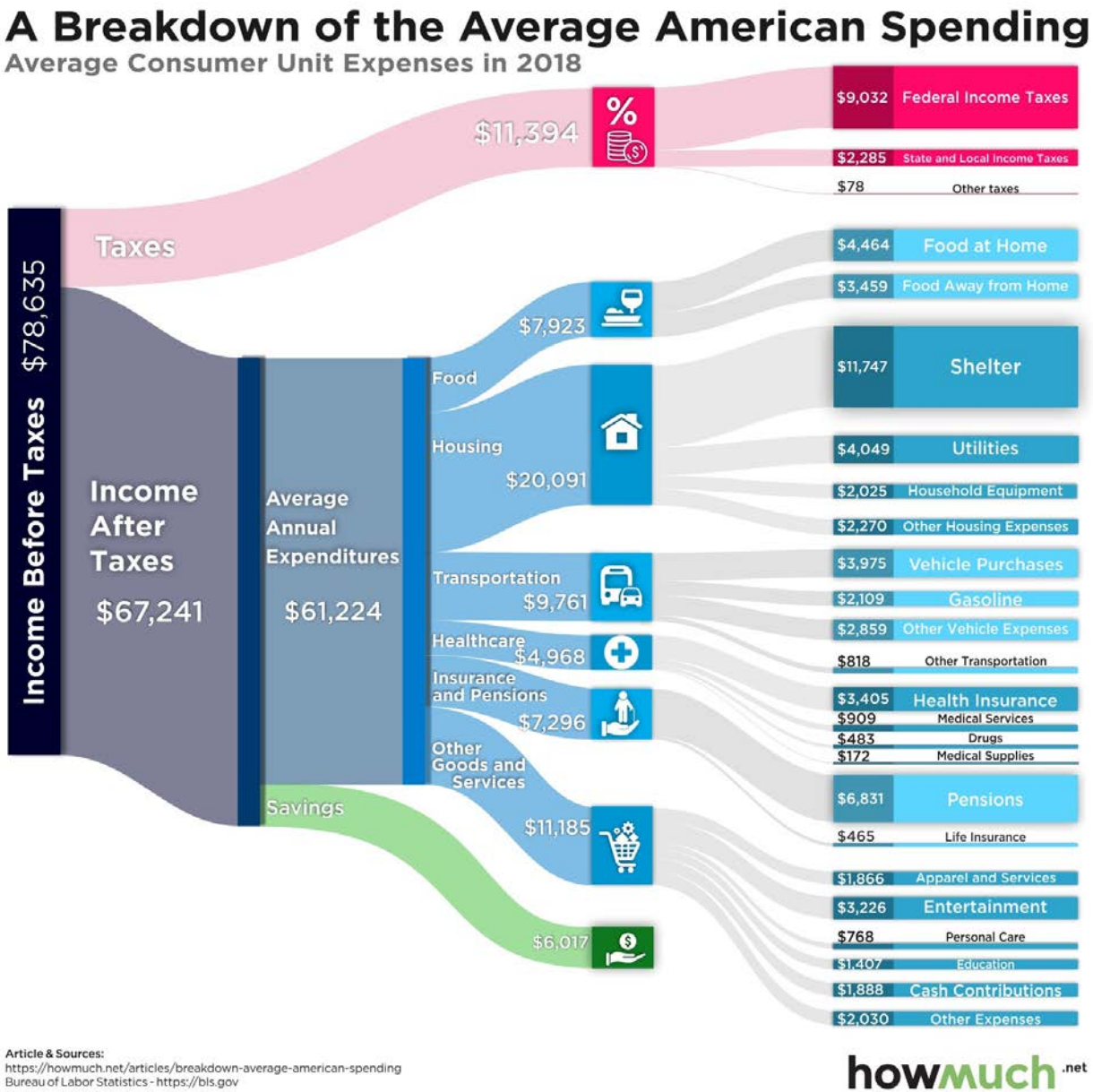


Directory of visualizations: Proportions

Proportions can be visualized as pie charts, side-by-side bars, or stacked bars and as in the case for amounts, bars can be arranged either vertically or horizontally. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars. Stacked bars look awkward for a single set of proportions, but can be useful when comparing multiple sets of proportions

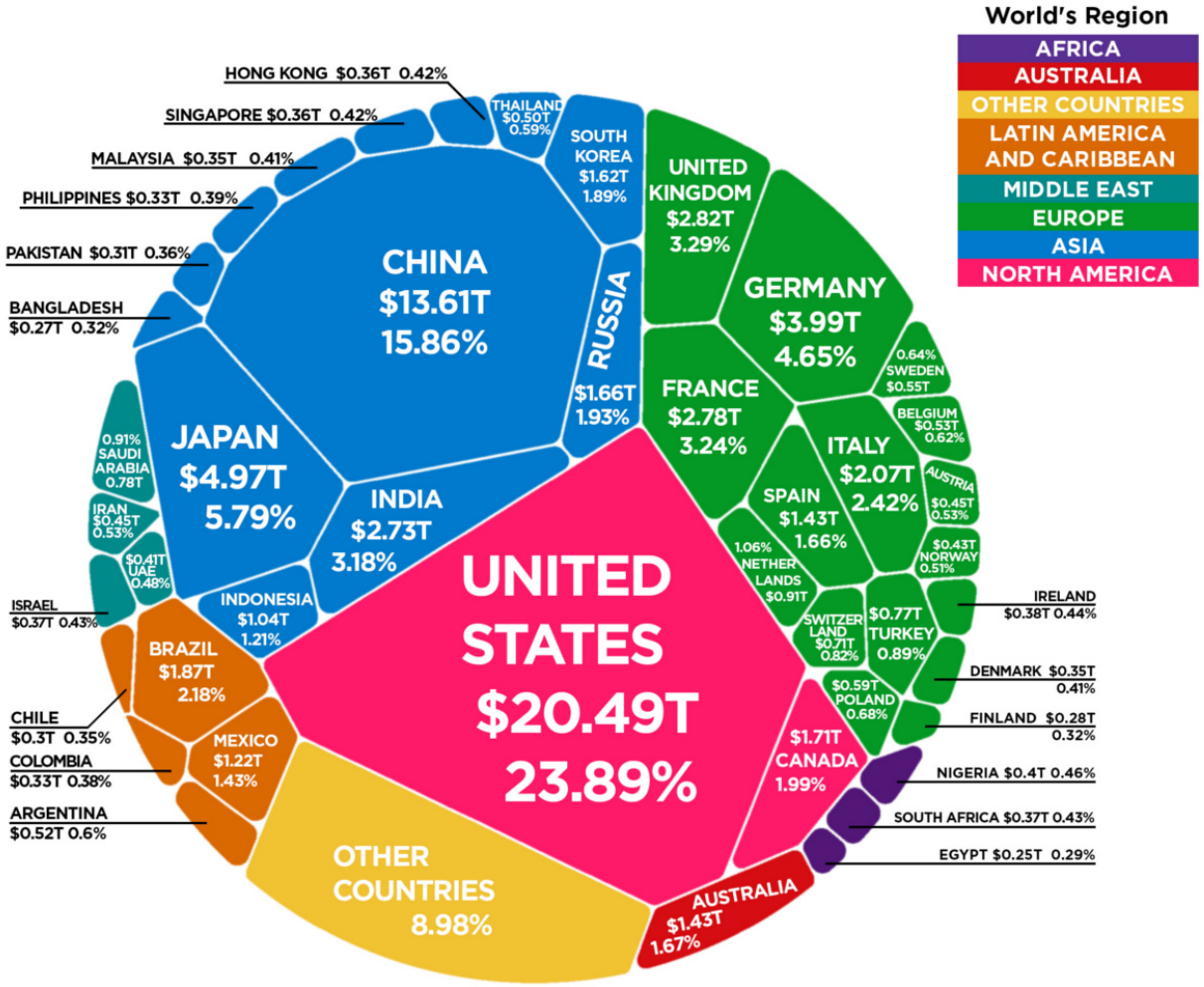
Directory of visualizations: Proportions

Proportions can be visualized as pie charts, side-by-side bars, or stacked bars and as in the case for amounts, bars can be arranged either vertically or horizontally. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars. Stacked bars look awkward for a single set of proportions, but can be useful when comparing multiple sets of proportions



Directory of visualizations: Proportions

Proportions can be visualized as pie charts, side-by-side bars, or stacked bars and as in the case for amounts, bars can be arranged either vertically or horizontally. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars. Stacked bars look awkward for a single set of proportions, but can be useful when comparing multiple sets of proportions

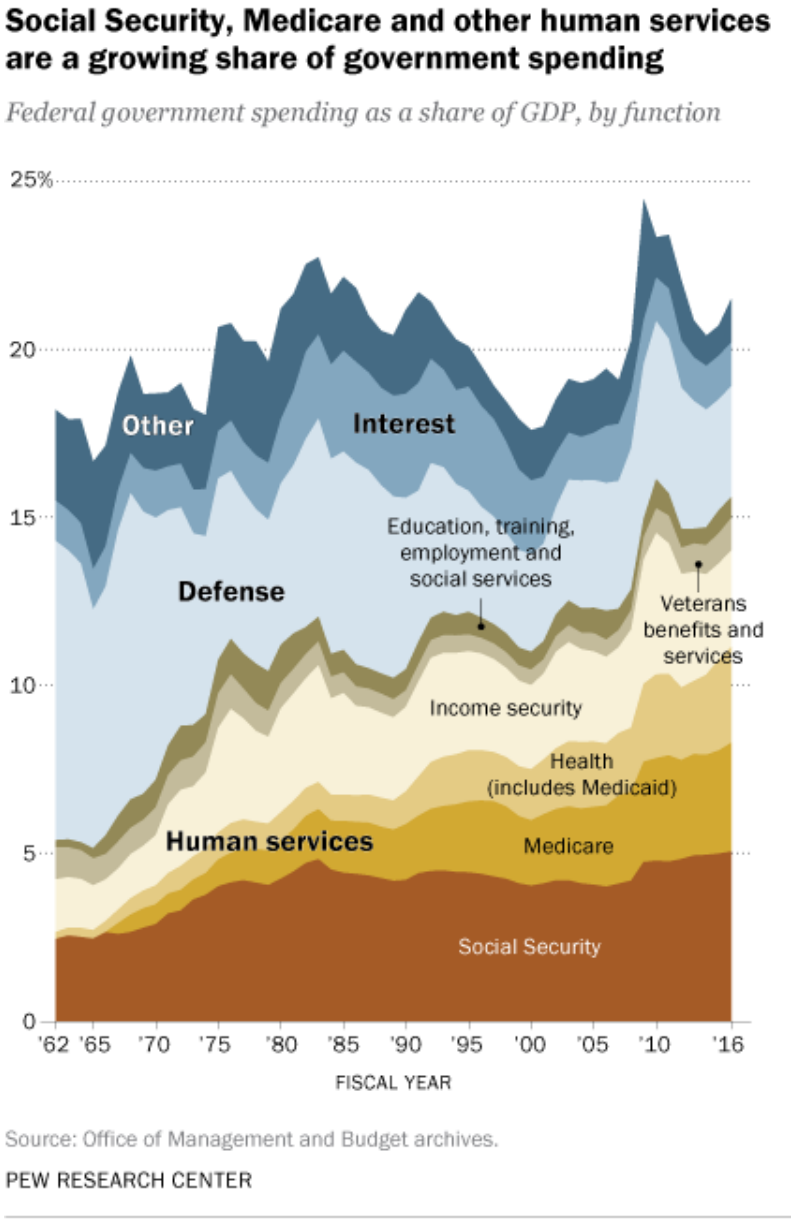


Article & Sources:
https://howmuch.net/articles/the-world-economy-2018
https://databank.worldbank.org

howmuch.net

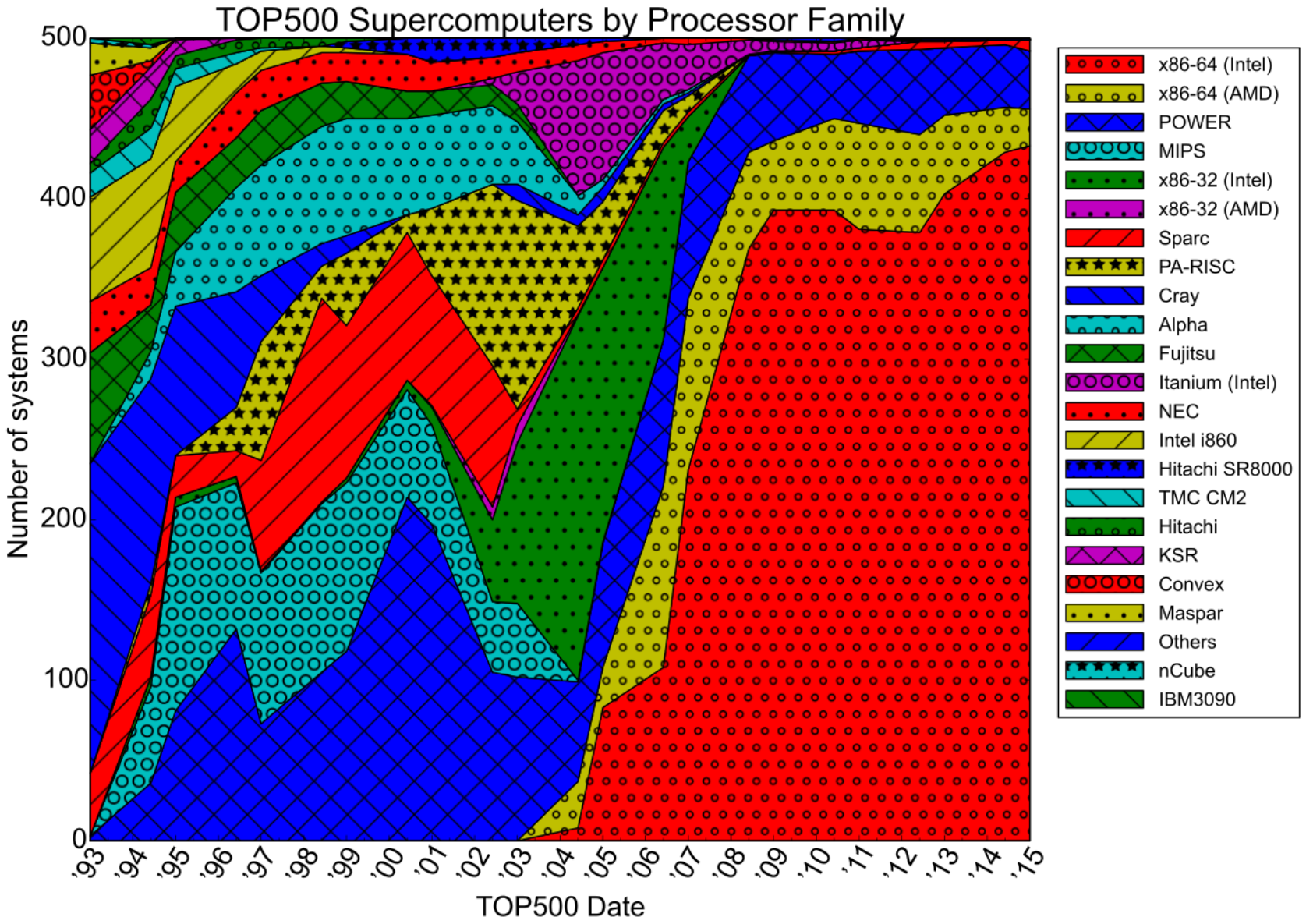
Directory of visualizations: Proportions

Proportions can be visualized as pie charts, side-by-side bars, or stacked bars and as in the case for amounts, bars can be arranged either vertically or horizontally. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars. Stacked bars look awkward for a single set of proportions, but can be useful when comparing multiple sets of proportions



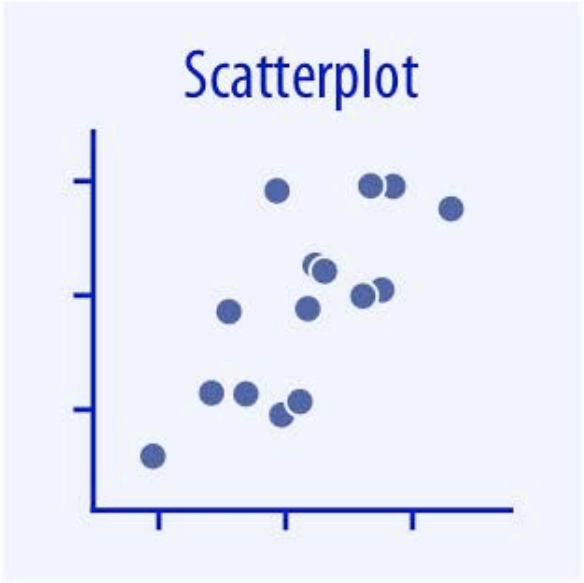
Directory of visualizations: Proportions

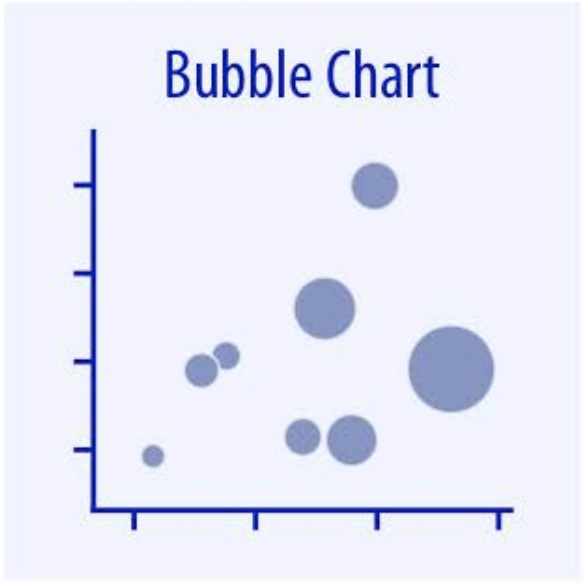
Proportions can be visualized as pie charts, side-by-side bars, or stacked bars and as in the case for amounts, bars can be arranged either vertically or horizontally. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces are more easily compared in side-by-side bars. Stacked bars look awkward for a single set of proportions, but can be useful when comparing multiple sets of proportions

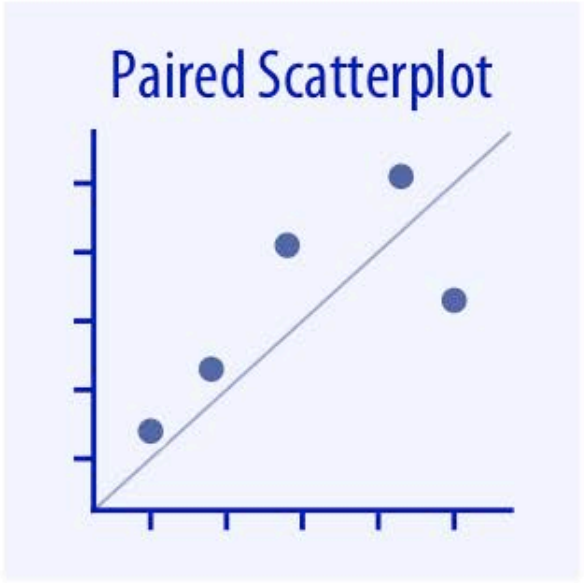


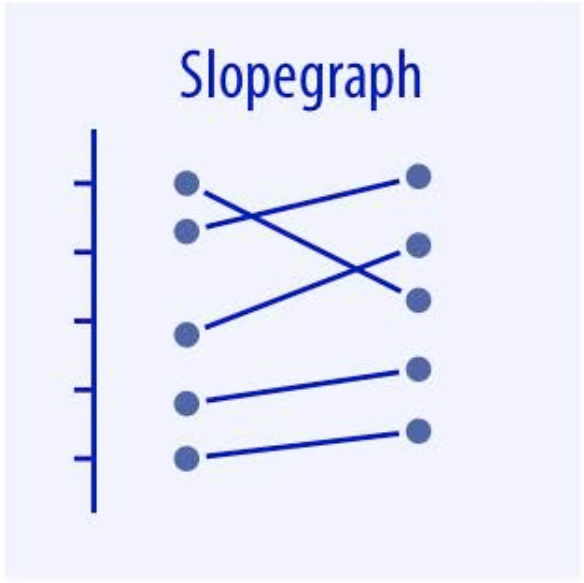
Directory of visualizations: *x*–*y* relationships

Scatterplots represent the archetypical visualization when we want to show one quantitative variable relative to another



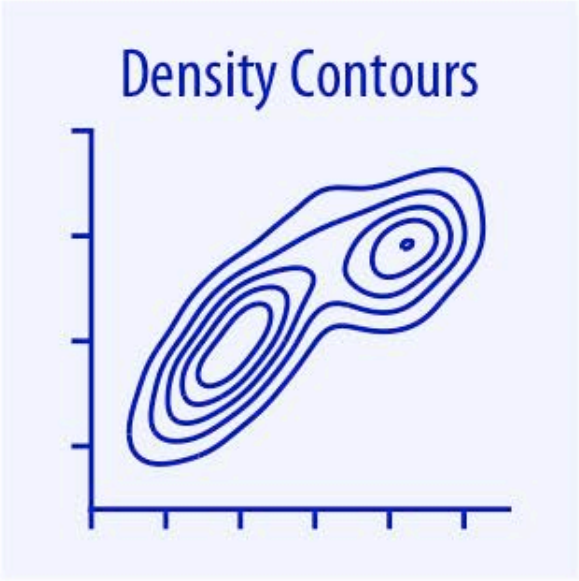


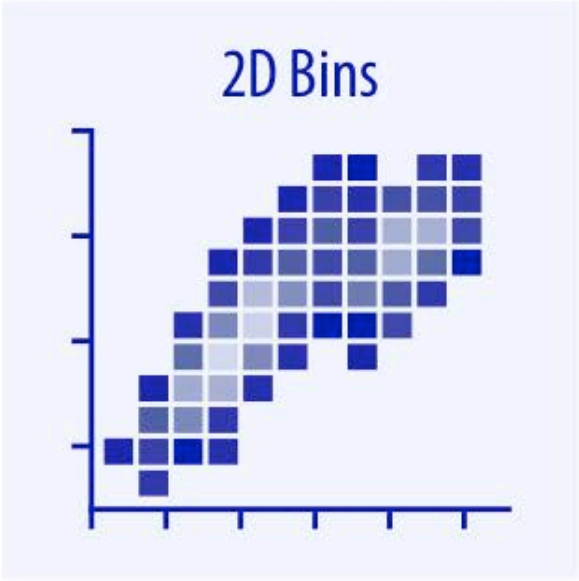


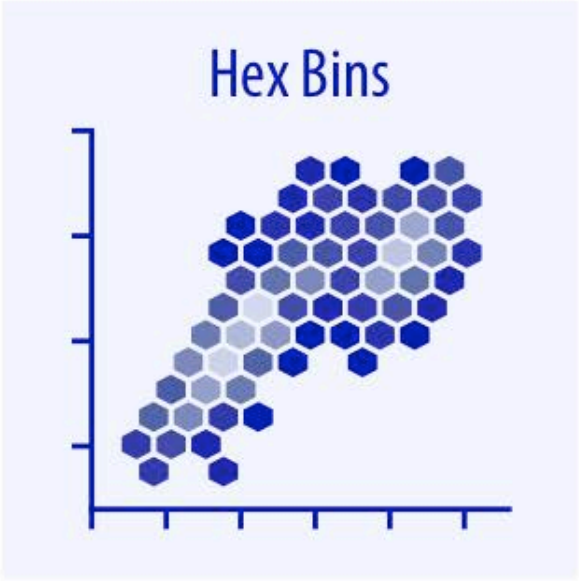


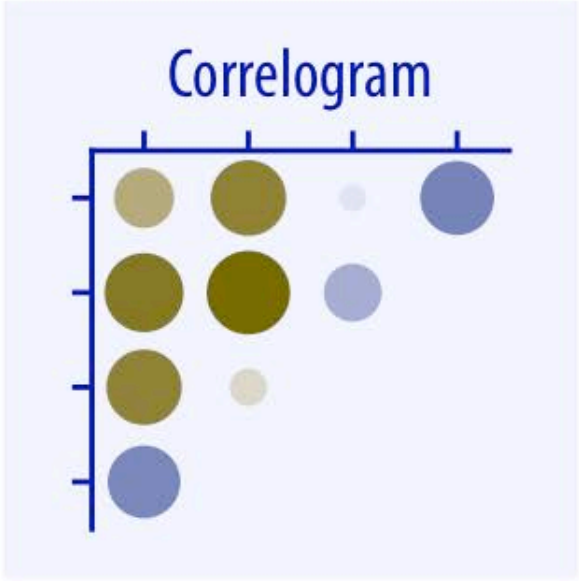
Directory of visualizations: *x*–*y* relationships

For large numbers of points, regular scatterplots can become uninformative due to overplotting. In this case, contour lines, 2D bins, or hex bins may provide an alternative



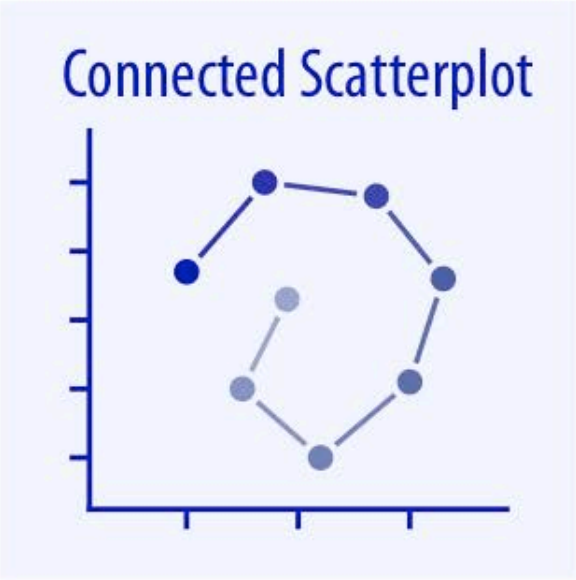


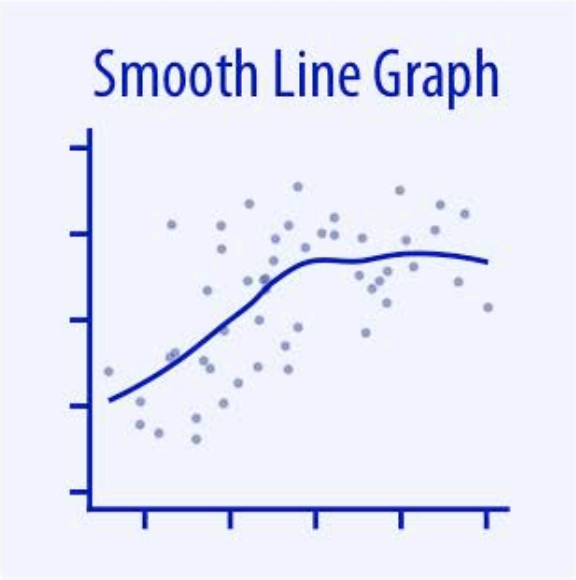




Directory of visualizations: *x*–*y* relationships

When the *x* axis represents time or a strictly increasing quantity such as a treatment dose, we commonly draw line graphs.





Directory of visualizations: 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, such that shapes and distances on the globe are approximately represented by shapes and distances in the 2D representation.

Map

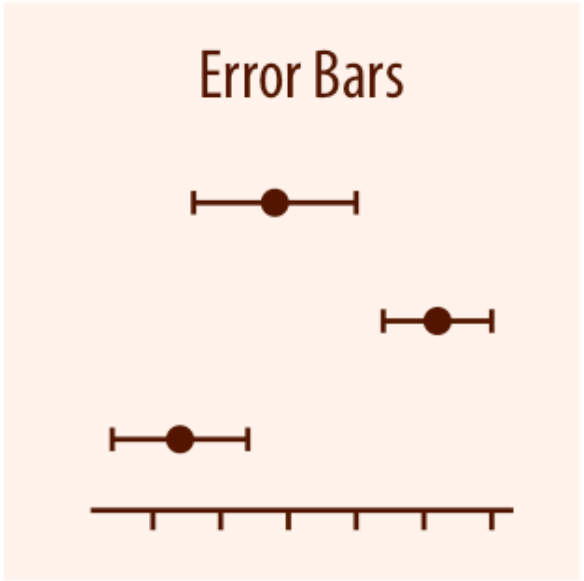
Choropleth

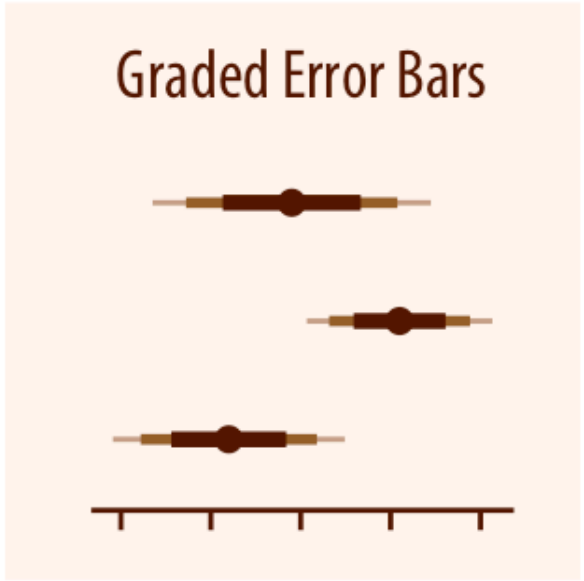
Cartogram

Cartogram Heatmap

Directory of visualizations: Uncertainty

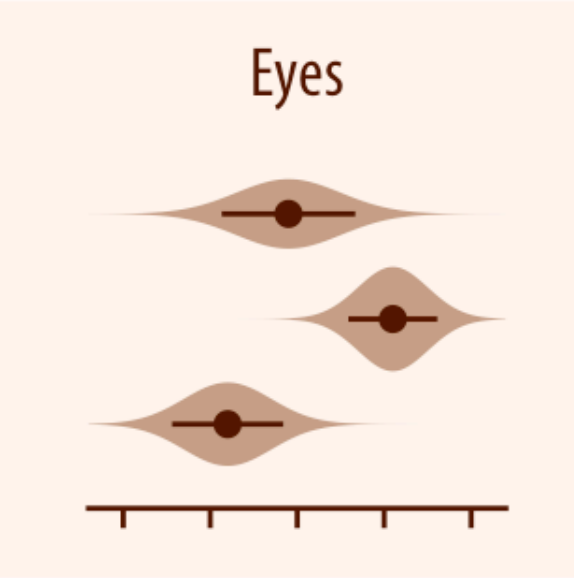
Error bars are meant to indicate the range of likely values for some estimate or measurement. They extend horizontally and/or vertically from some reference point representing the estimate or measurement.

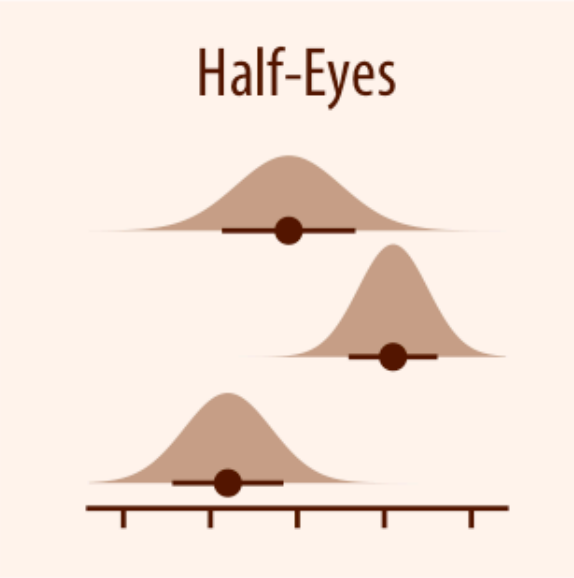


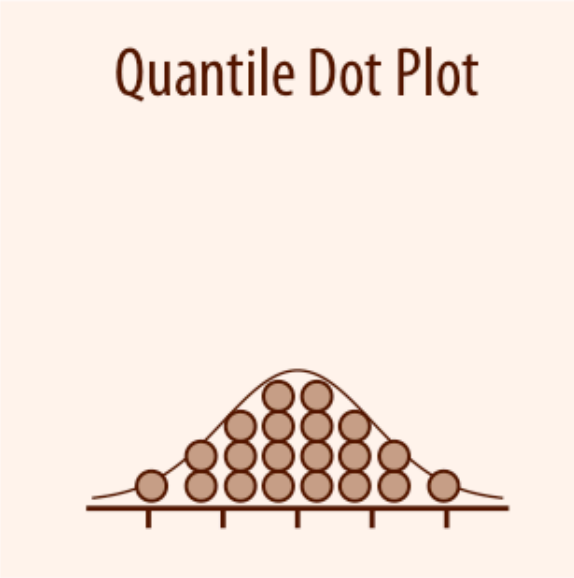


Directory of visualizations: Uncertainty

Error bars are meant to indicate the range of likely values for some estimate or measurement. They extend horizontally and/or vertically from some reference point representing the estimate or measurement.



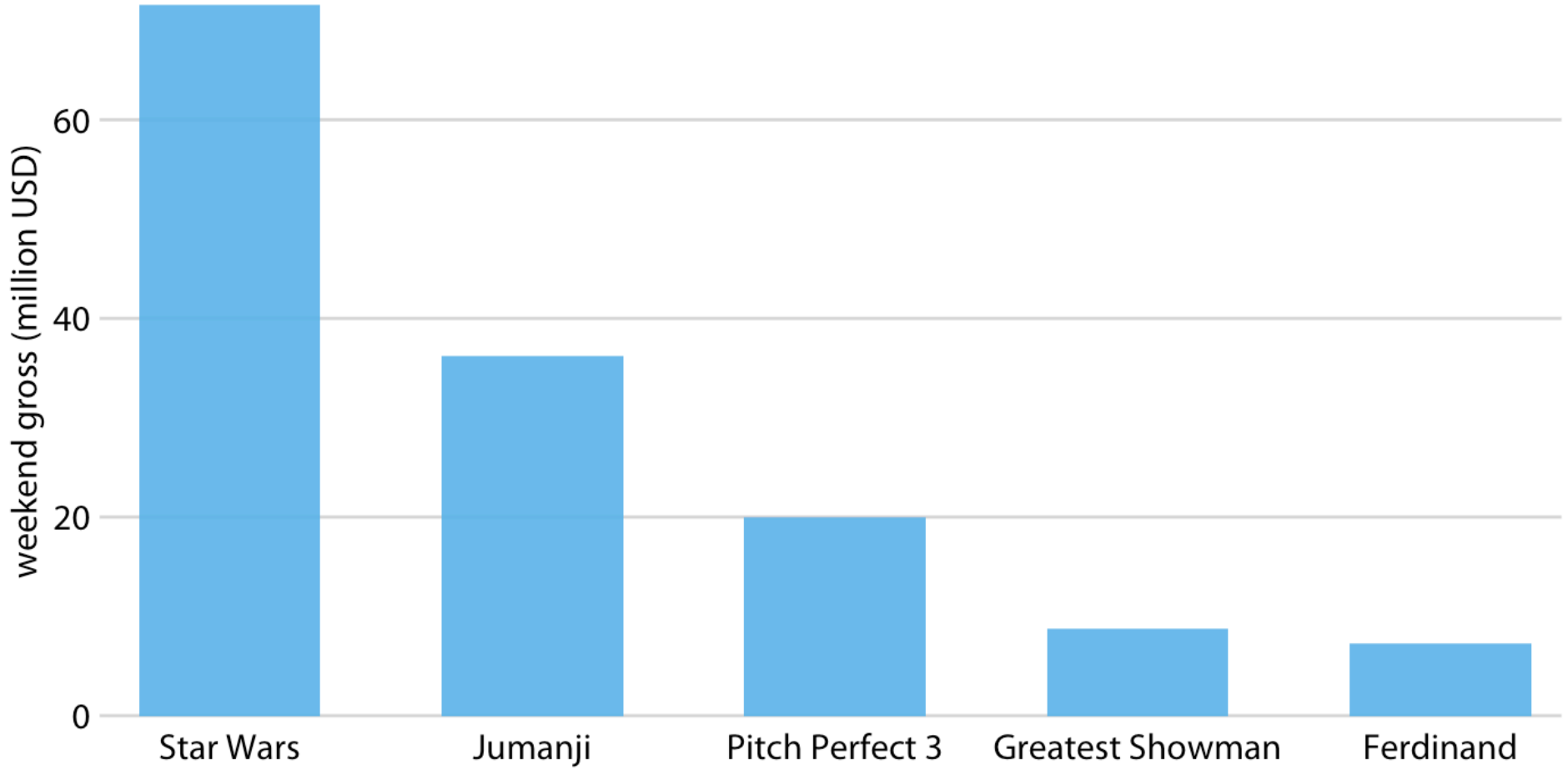




Visualizing amounts

Bar plots

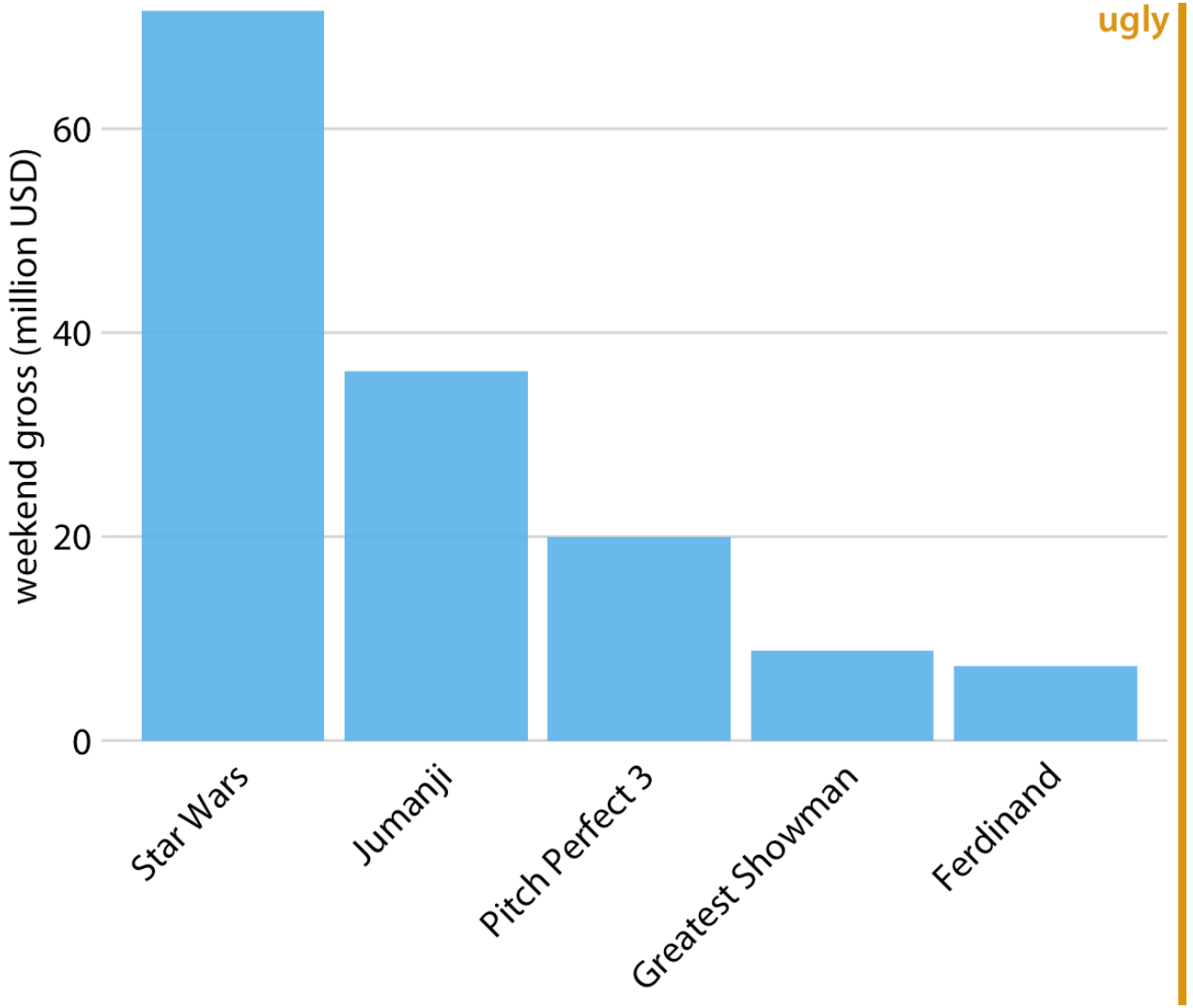
Highest grossing movies for the weekend of December 22-24, 2017, displayed as a bar plot. Data source: Box Office Mojo



Visualizing amounts

Bar plots

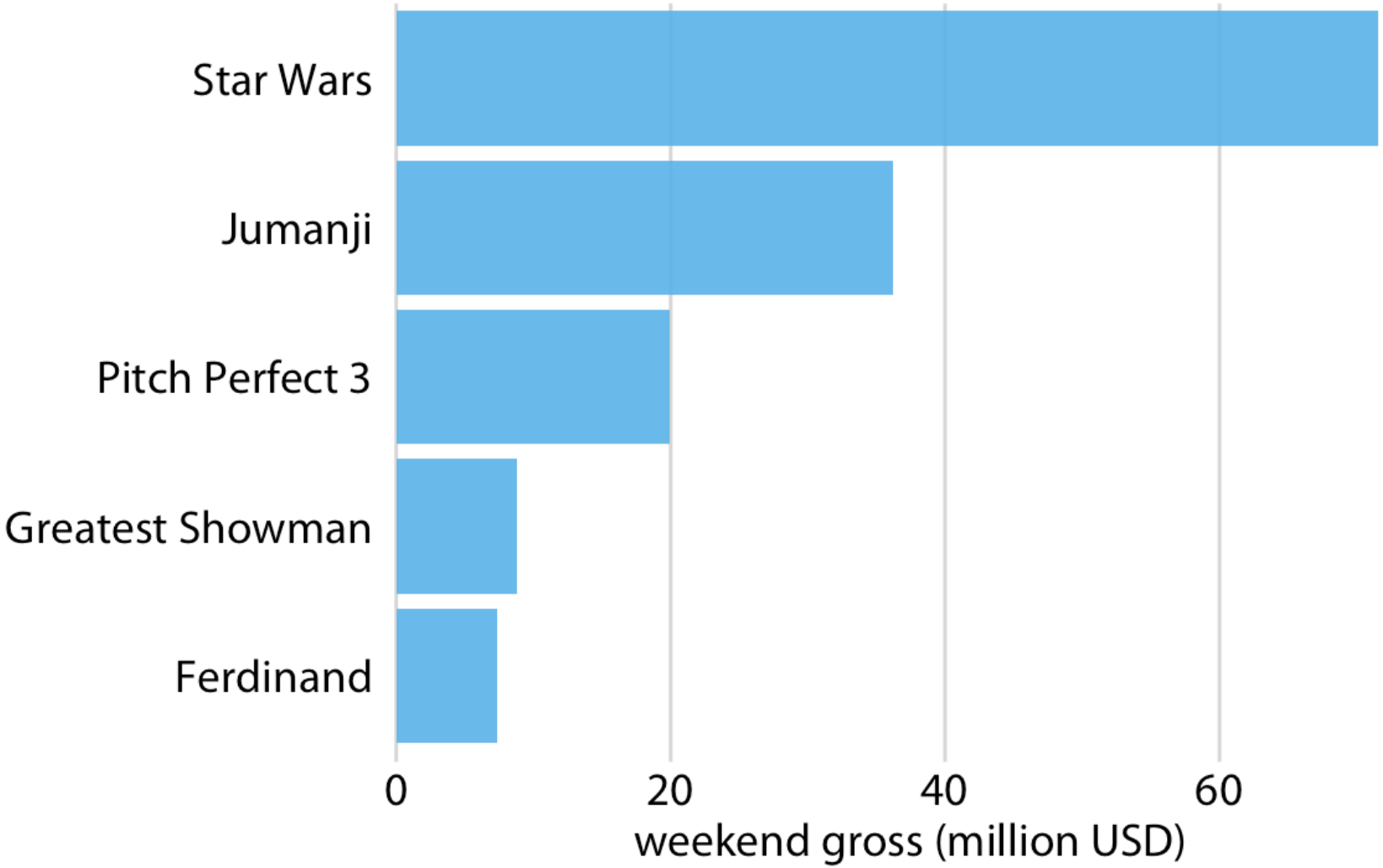
Highest grossing movies for the weekend of December 22-24, 2017, displayed as a bar plot. Data source: Box Office Mojo



Visualizing amounts

Bar plots

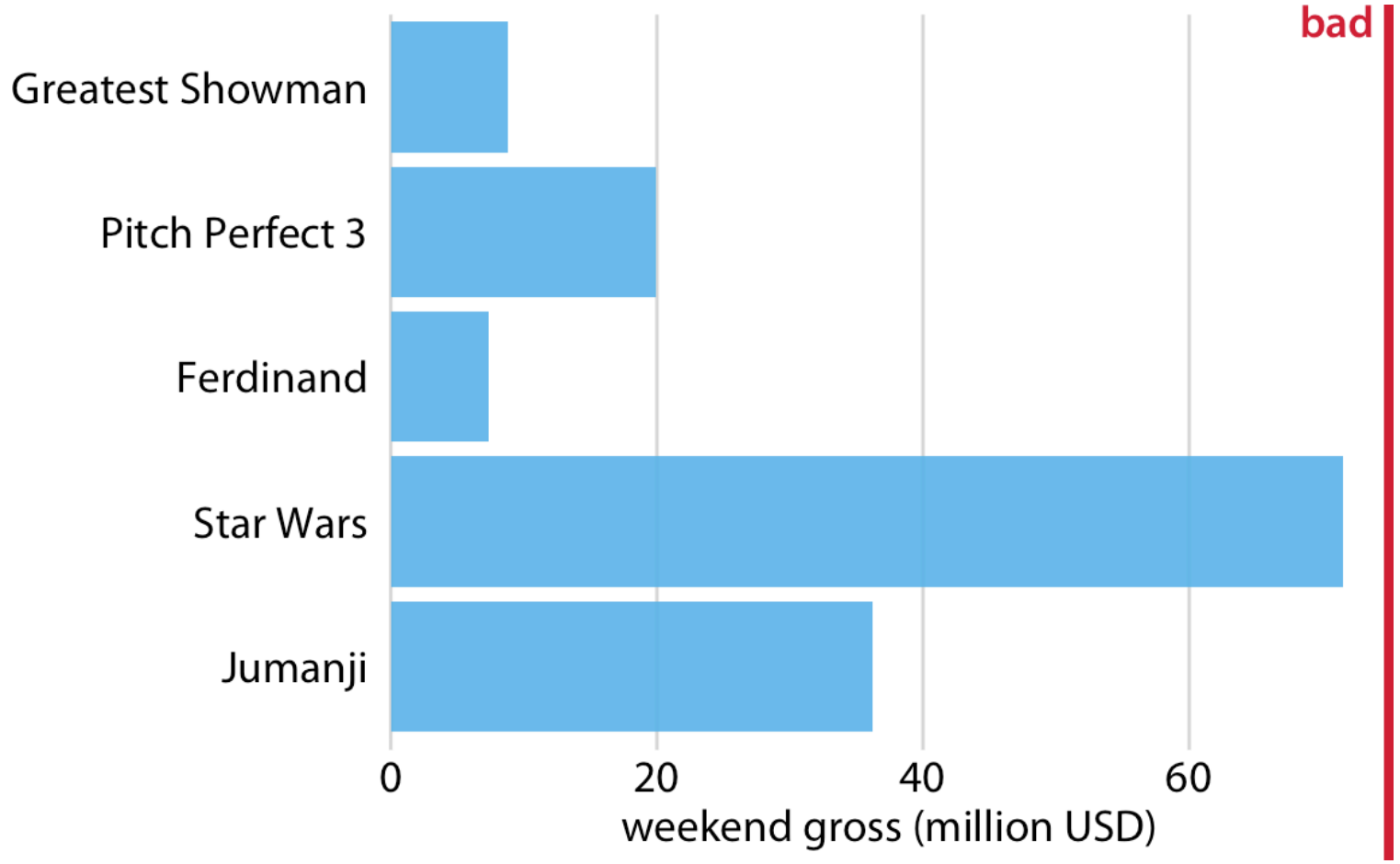
Highest grossing movies for the weekend of December 22-24, 2017, displayed as a bar plot. Data source: Box Office Mojo



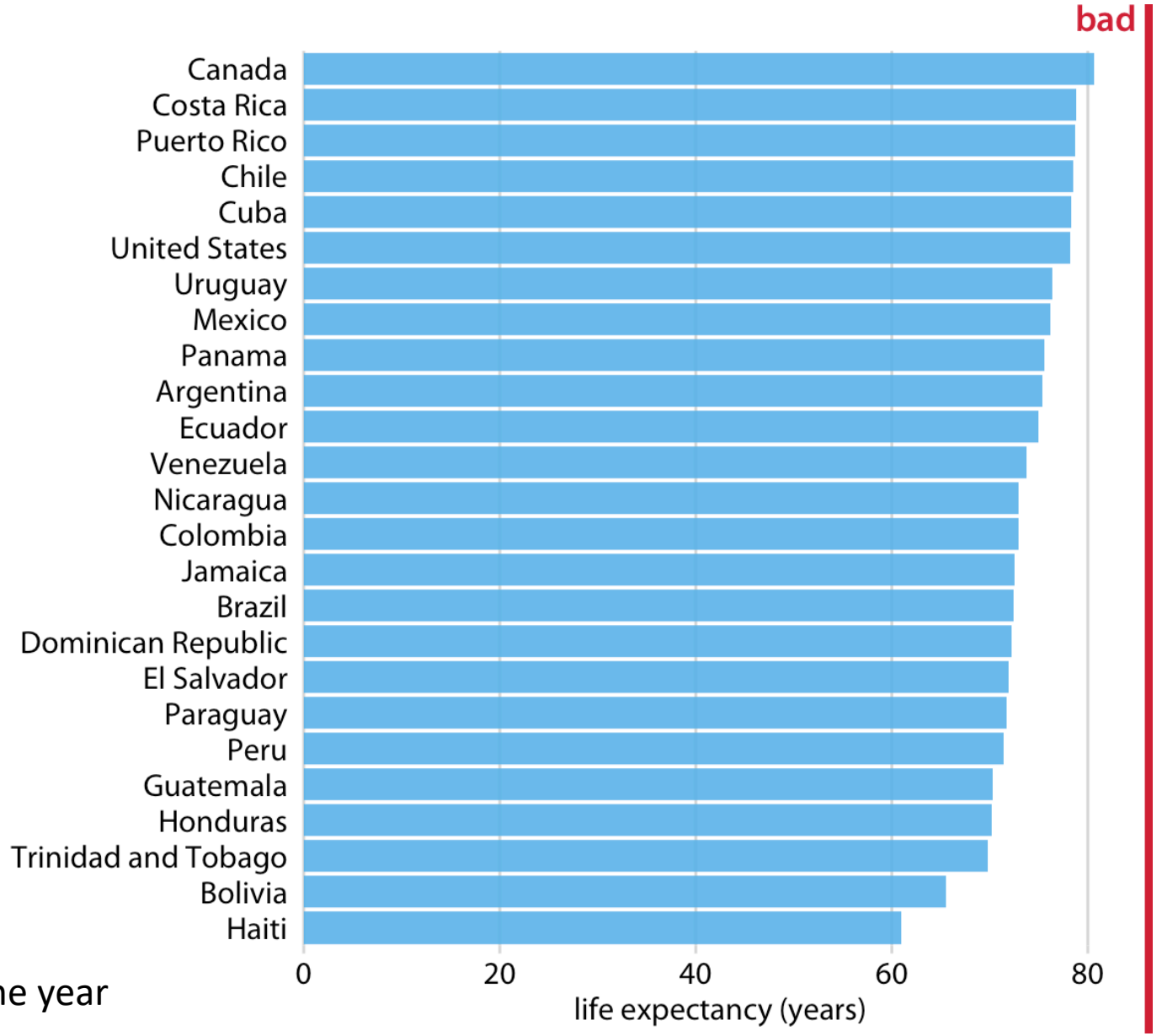
Visualizing amounts

Bar plots

Highest grossing movies for the weekend of December 22-24, 2017, displayed as a bar plot. Data source: Box Office Mojo

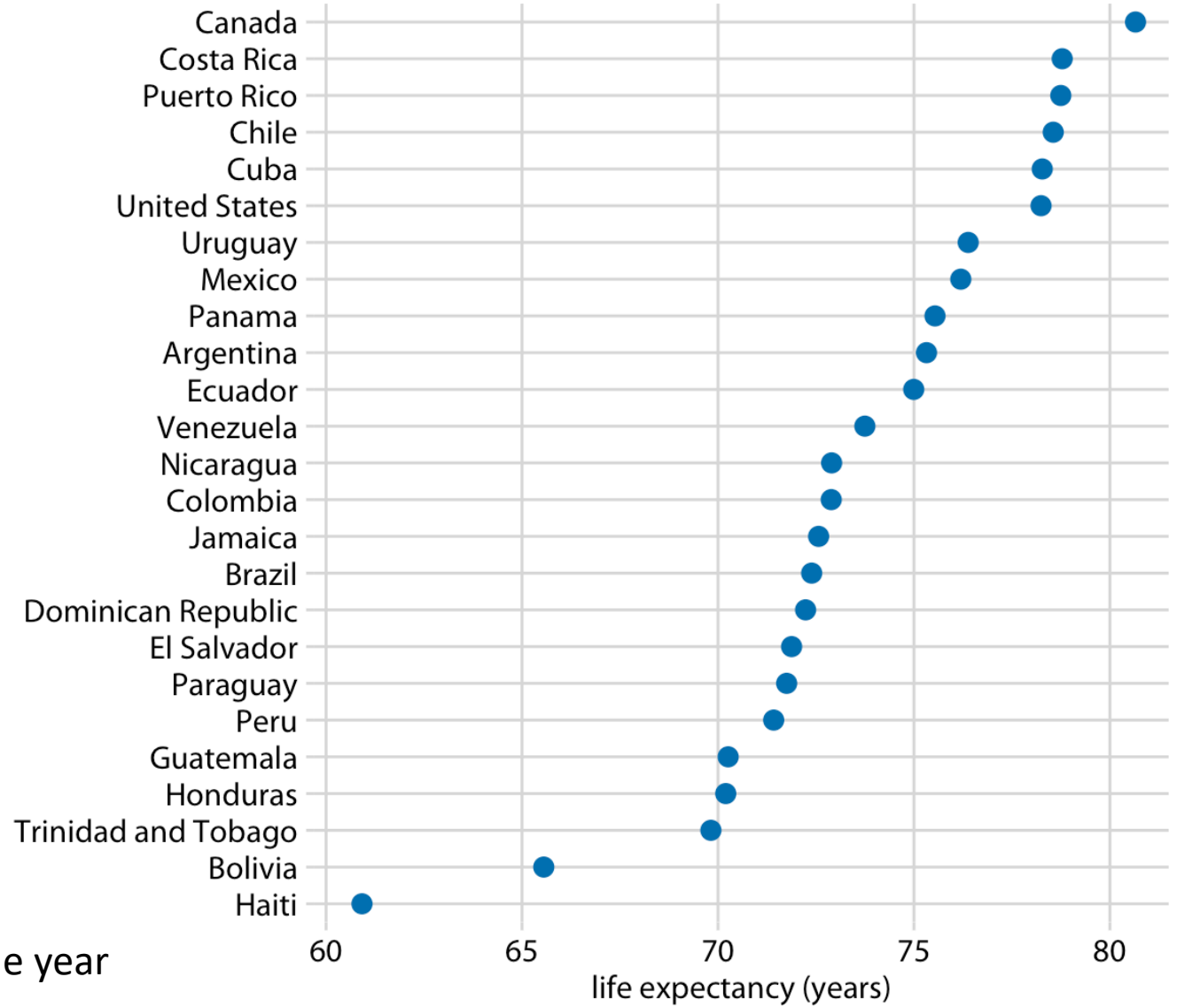


Visualizing amounts



Life expectancies of countries in the Americas, for the year 2007. Data source: Gapminder project

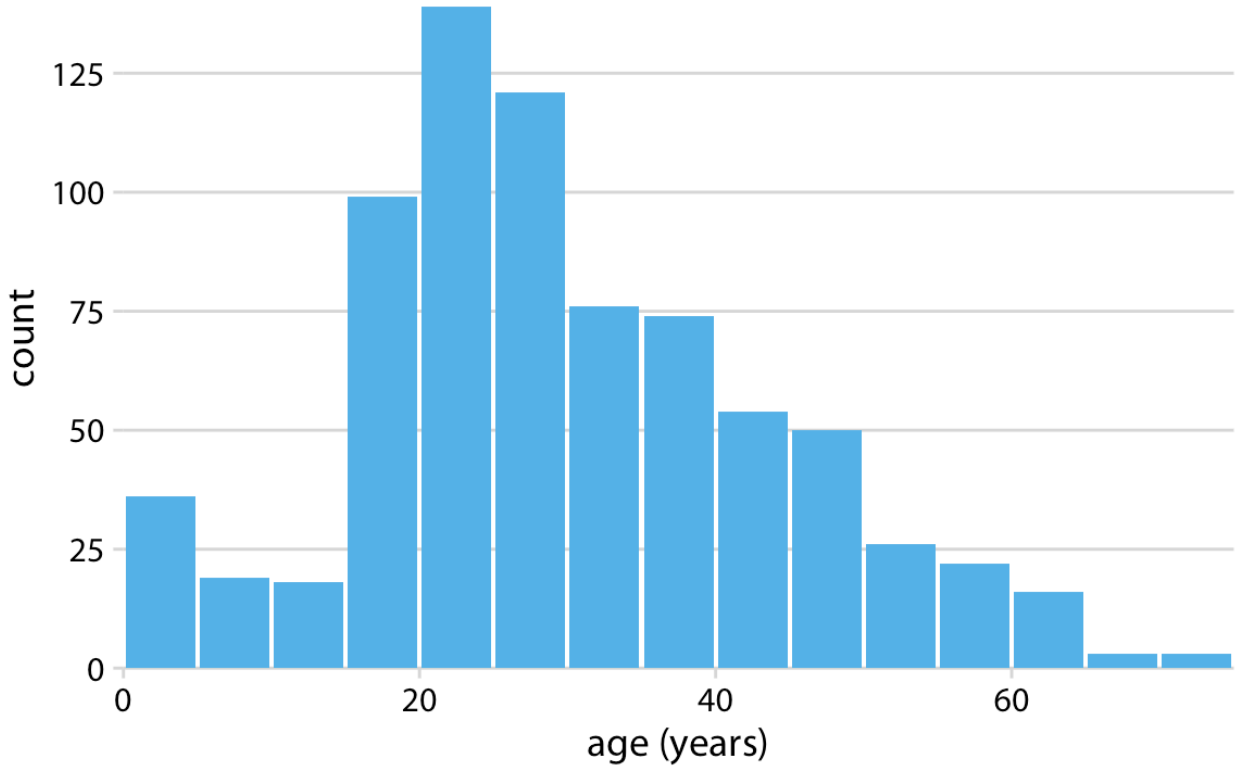
Visualizing amounts



Life expectancies of countries in the Americas, for the year 2007. Data source: Gapminder project

Visualizing amounts

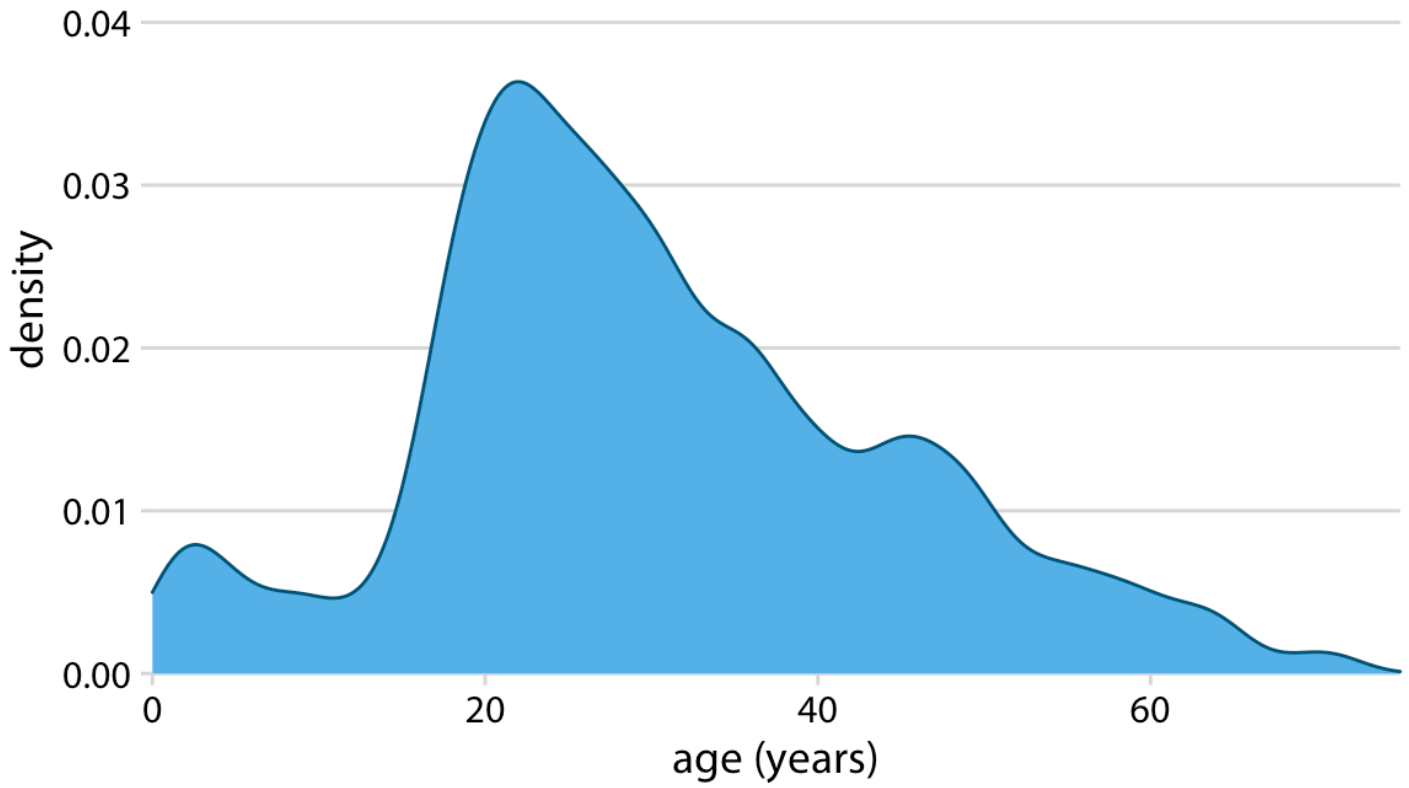
Visualizing distributions: Histograms and density plots



Histogram of the ages of Titanic passengers.

Visualizing amounts

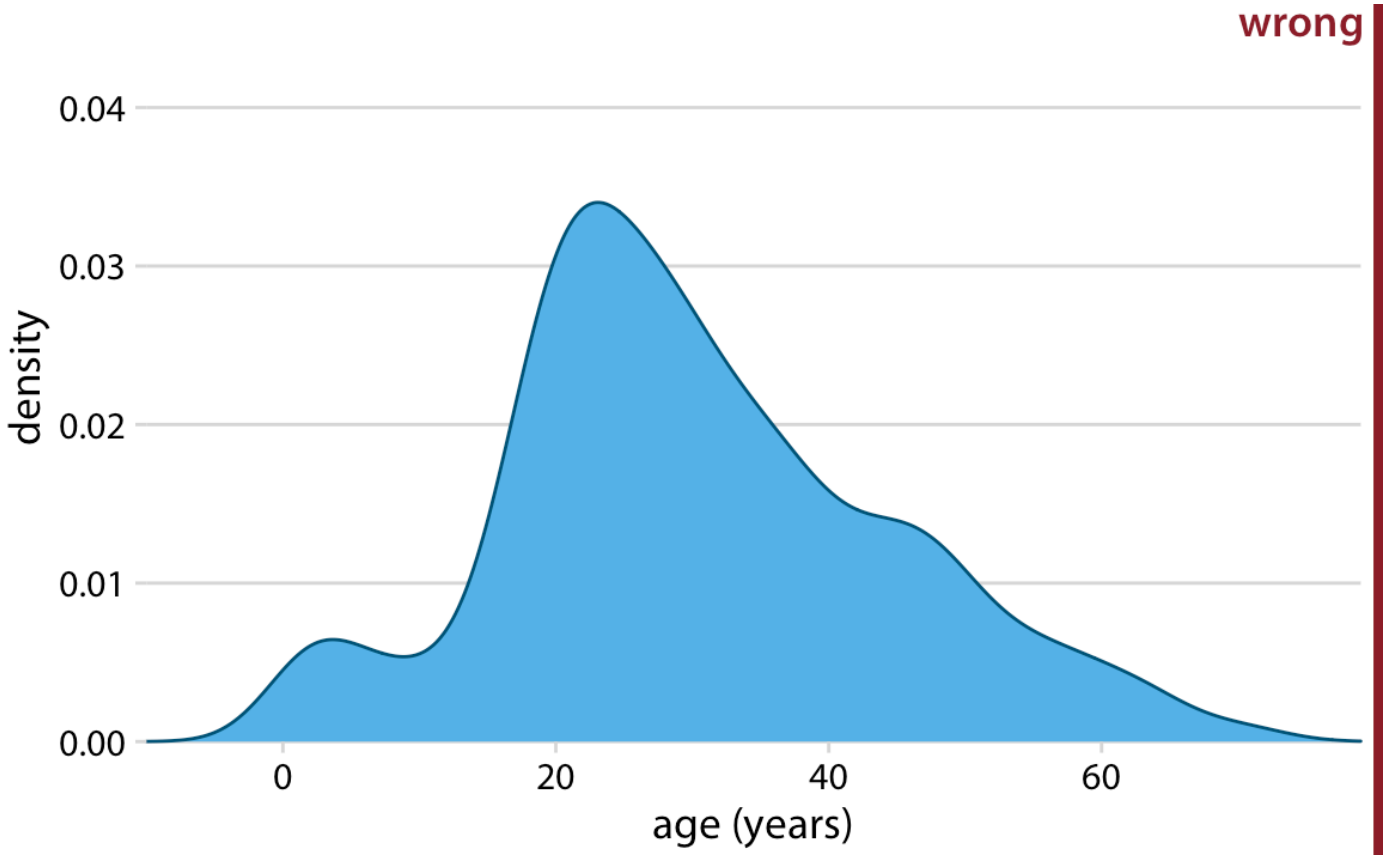
Visualizing distributions: Histograms and density plots



Kernel density estimate of the age distribution of passengers on the Titanic. The height of the curve is scaled such that the area under the curve equals one. The density estimate was performed with a Gaussian kernel and a bandwidth of 2.

Visualizing amounts

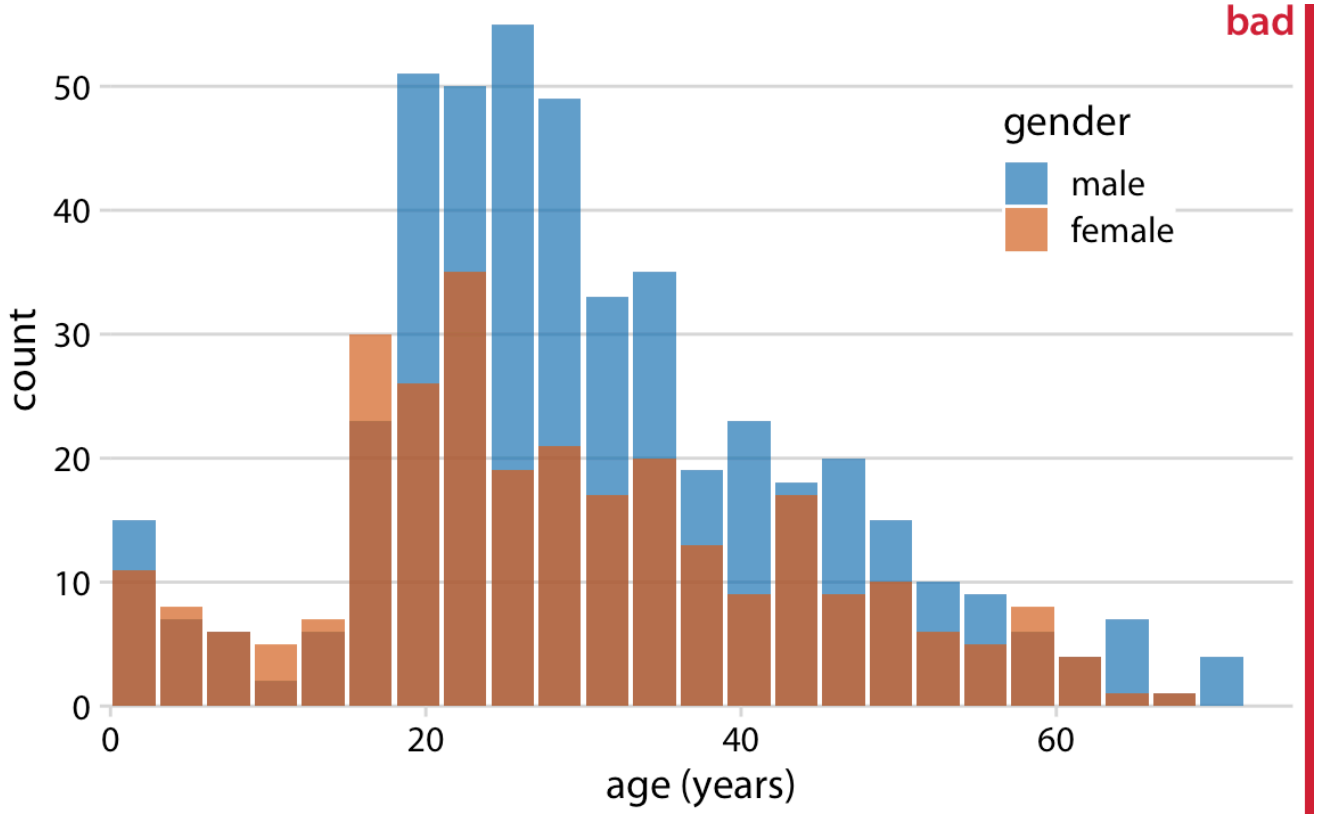
Visualizing distributions: Histograms and density plots



Kernel density estimate of the age distribution of passengers on the Titanic. The height of the curve is scaled such that the area under the curve equals one. The density estimate was performed with a Gaussian kernel and a bandwidth of 2.

Visualizing amounts

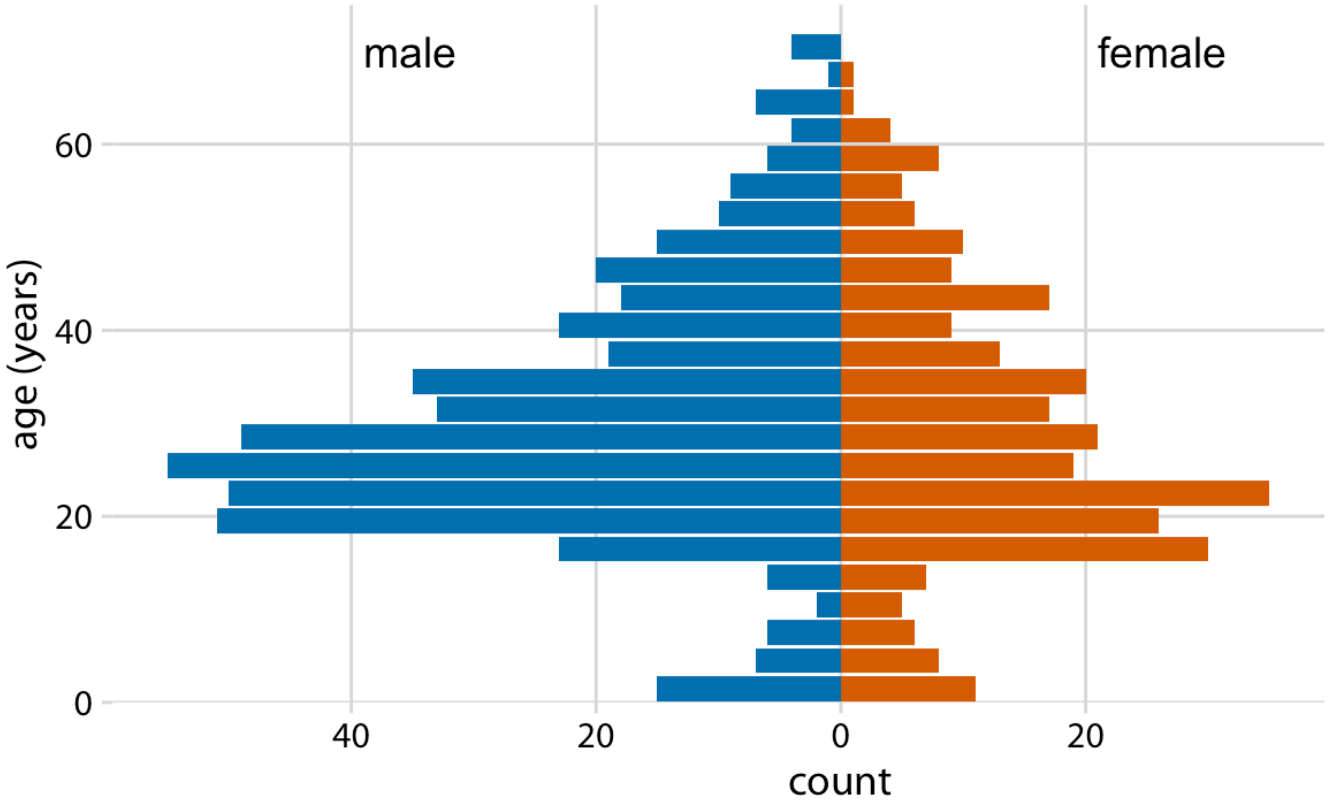
Visualizing distributions: Histograms and density plots



Age distributions of male and female Titanic passengers, shown as two overlapping histograms. This figure has been labeled as "bad" because there is no clear visual indication that all blue bars start at a count of 0.

Visualizing amounts

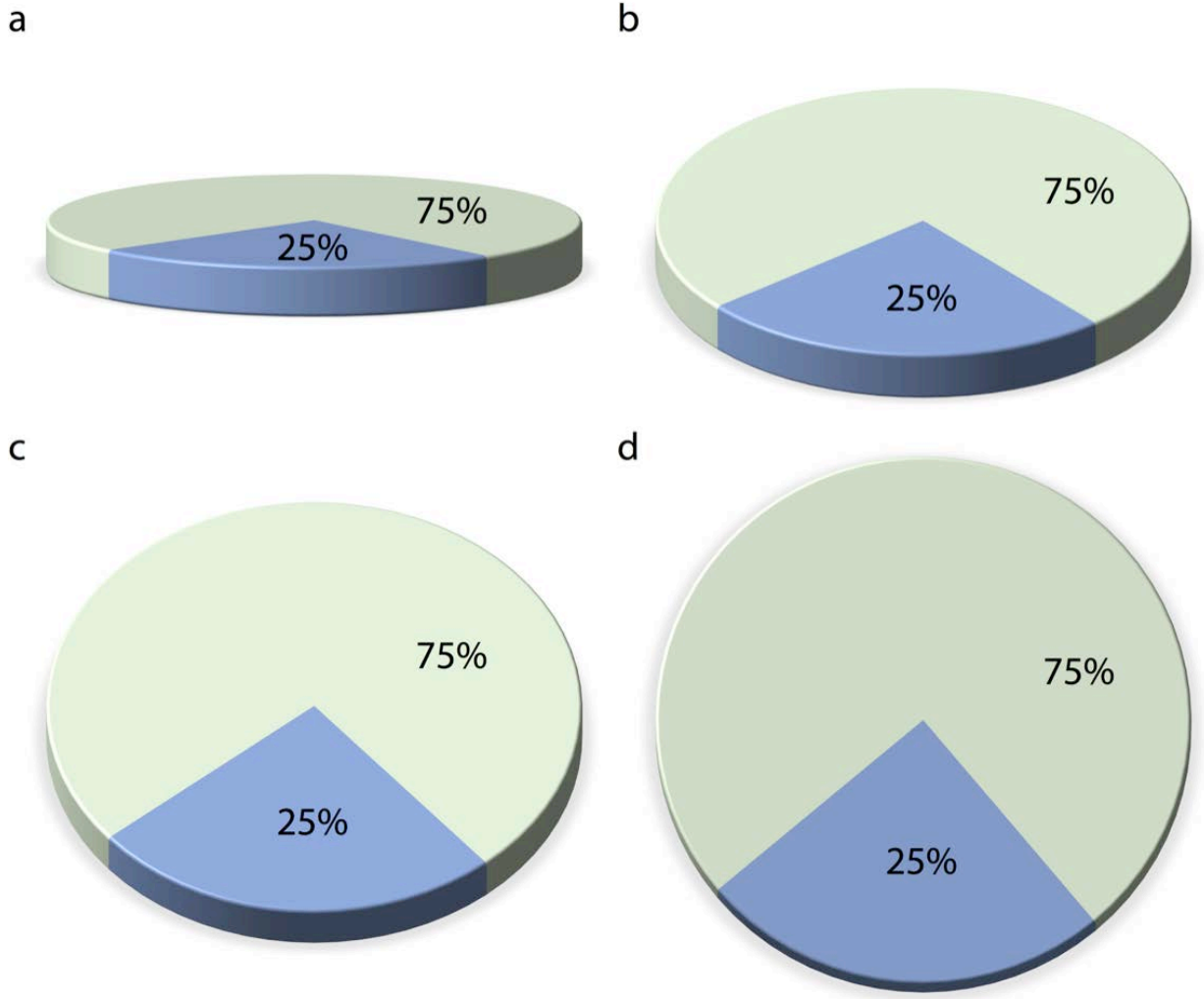
Visualizing distributions: Histograms and density plots



The age distributions of male and female Titanic passengers visualized as an age pyramid.

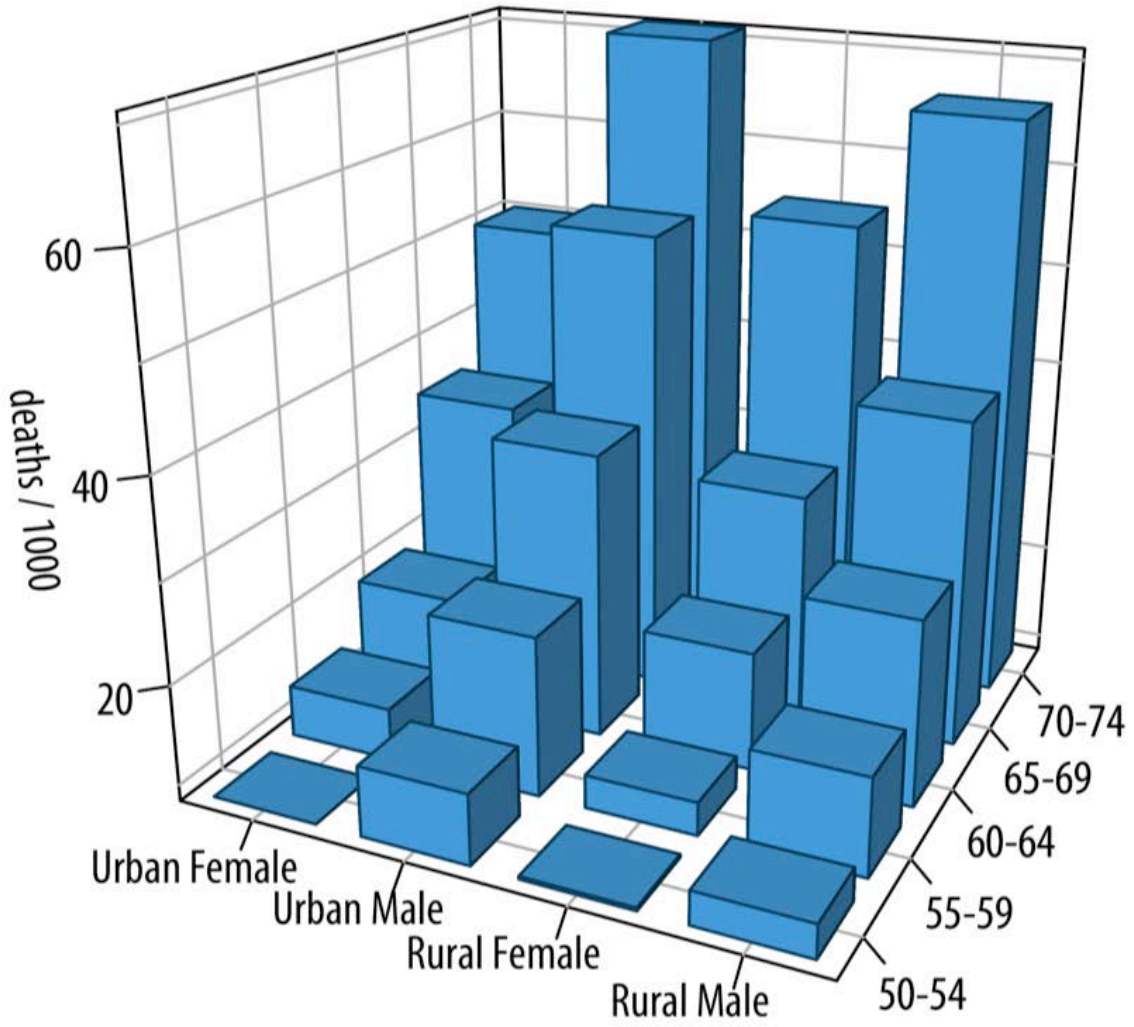
Don’t go 3D

The same 3D pie chart shown from four different angles. Rotating a pie into the third dimension makes pie slices in the front appear larger than they really are and pie slices in the back appear smaller. Here, in parts (a), (b), and (c), the blue slice corresponding to 25% of the data visually occupies more than 25% of the area representing the pie. Only part (d) is an accurate representation of the data.



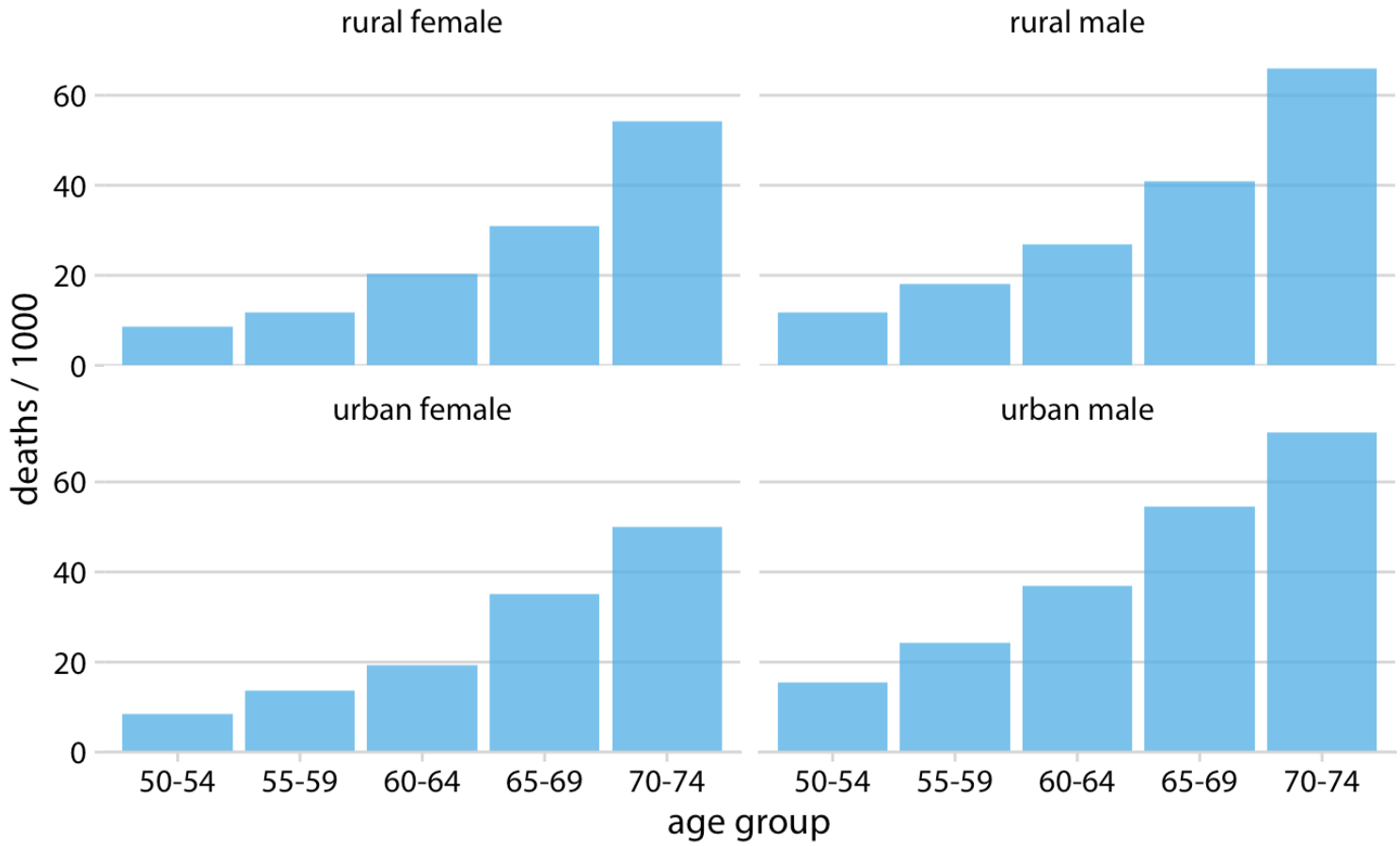
Don’t go 3D

Mortality rates in Virginia in 1940, visualized as a 3D bar plot. Mortality rates are shown for four groups of people (urban and rural females and males) and five age categories (50–54, 55–59, 60–64, 65–69, 70–74), and they are reported in units of deaths per 1000 persons. This figure is labeled as “bad” because the 3D perspective makes the plot difficult to read.

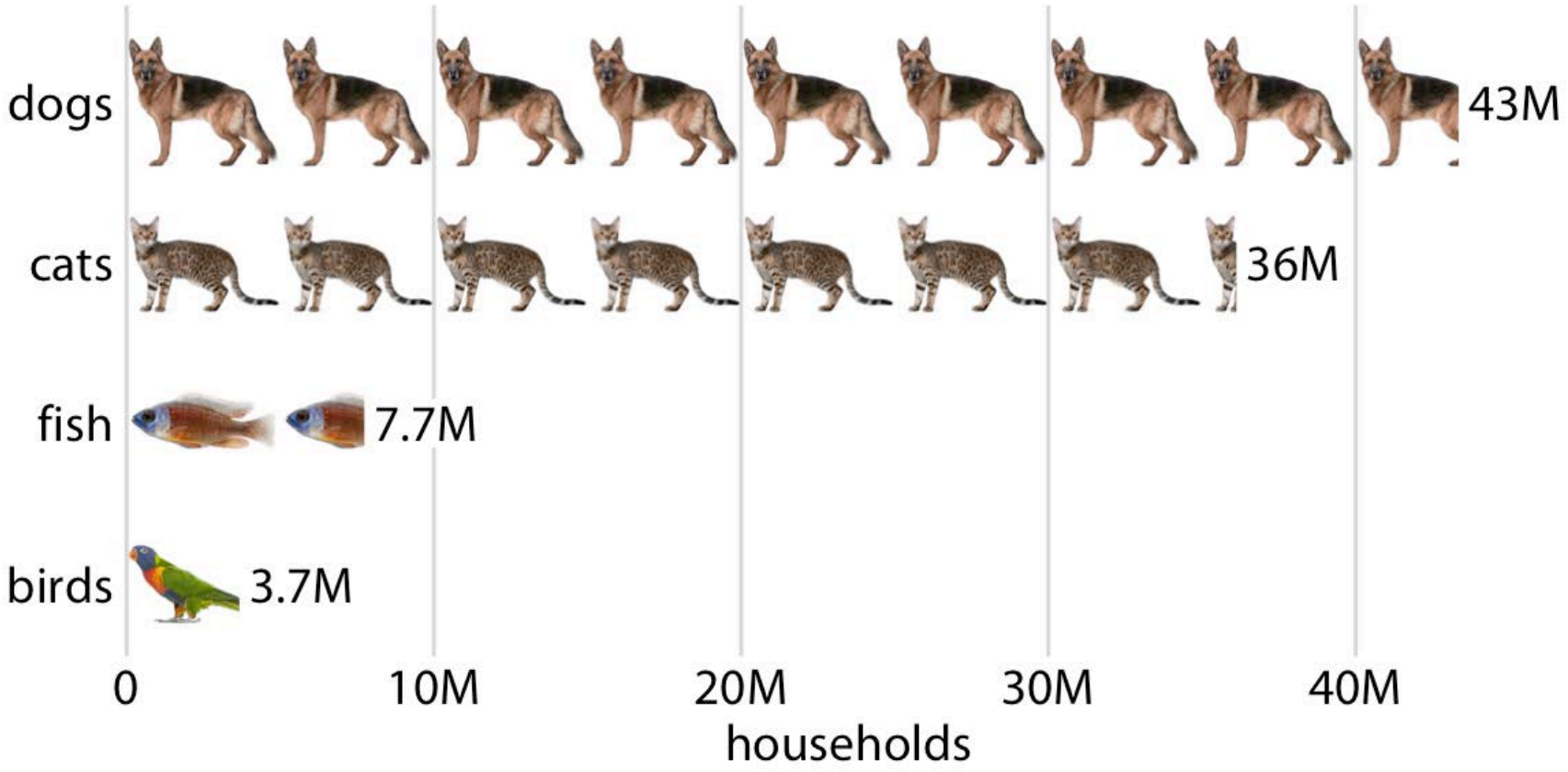


Don’t go 3D

Mortality rates in Virginia in 1940, visualized as a Trellis plot. Mortality rates are shown for four groups of people (urban and rural females and males) and five age categories (50–54, 55–59, 60–64, 65–69, 70–74), and they are reported in units of deaths per 1000 persons.



Visualizing amounts



Number of households having one or more of the most popular pets, shown as an isotype graph. Each complete animal represents 5 million households who have that kind of pet. Data source: 2012 U.S. Pet Ownership & Demographics Sourcebook, American Veterinary Medical Association