Introduction to Data Visualization

Visualizing Data: Mapping Data onto Aesthetics
Coordinate Systems and Axes

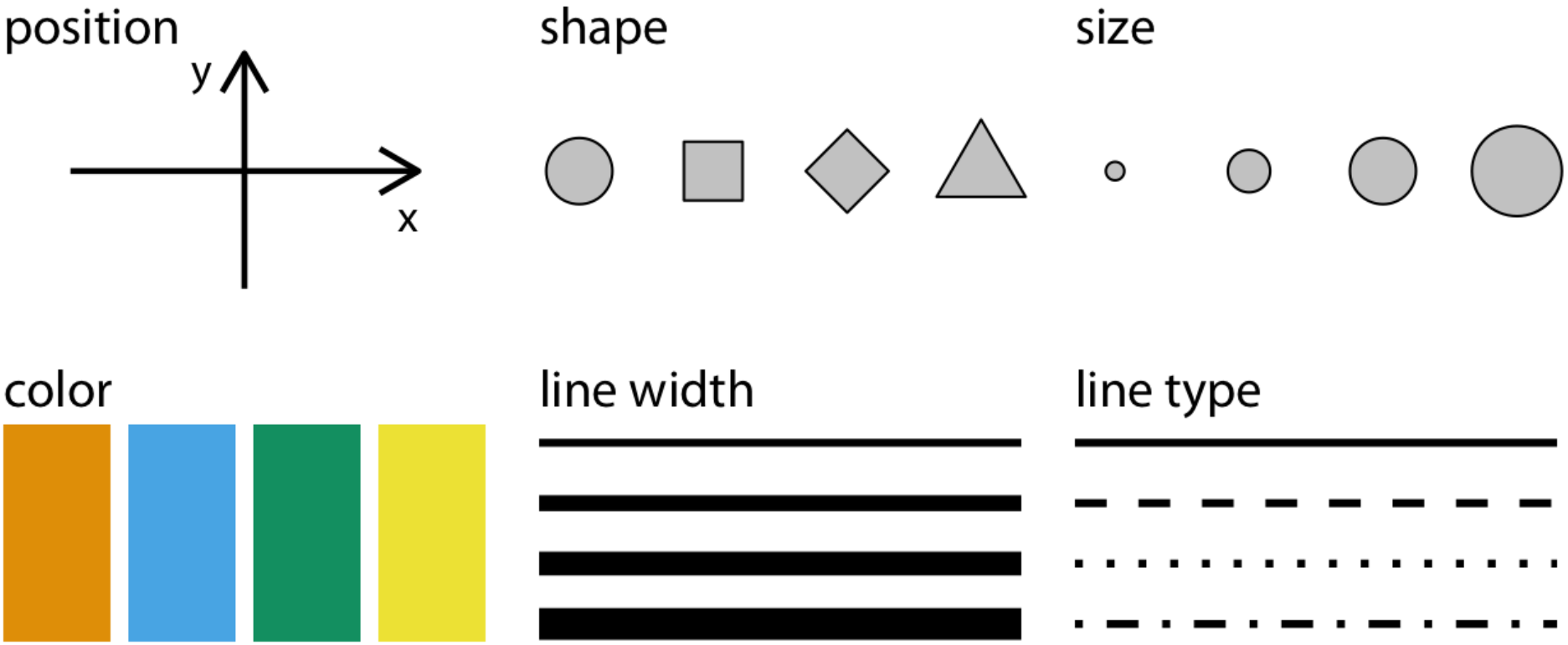
Halil Bisgin, Ph.D.

What is aesthetics?

- Whenever we visualize data, we take data values and convert them in a *systematic and logical* way into the visual elements that make up the final graphic.
- Even though there are many different types of data visualizations, all of them can be described with a common language that captures how data values are turned into colored pixels or ink.
- All data visualizations map *data values into quantifiable features* of the resulting graphic.
- We refer to these features as *aesthetics.*

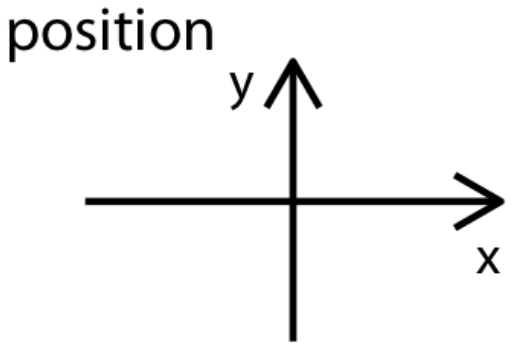
Aesthetics

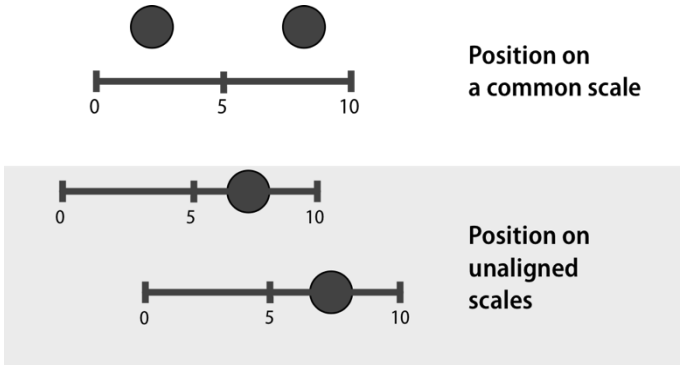
- Aesthetics describe every aspect of a given graphical element.



Components of graphical elements-1

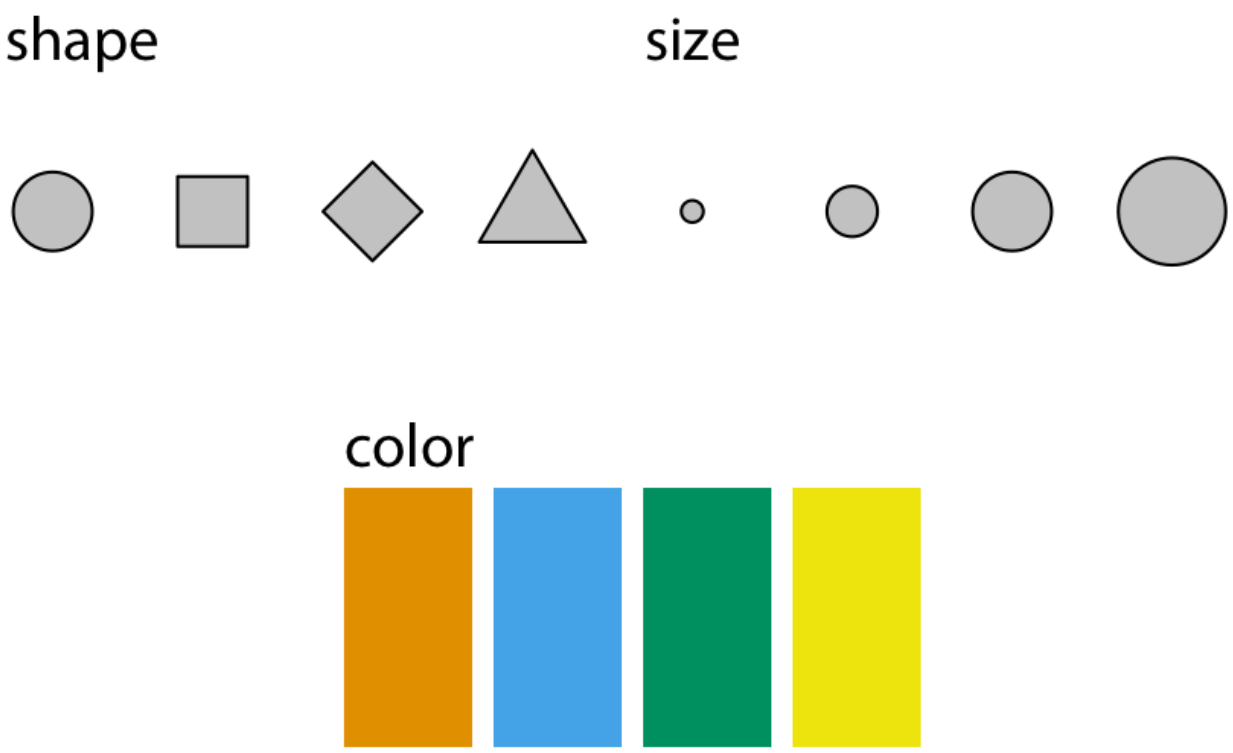
- A critical component of every graphical element is of course its position, which describes where the element is located.
- In standard 2D graphics, we describe positions by an x and y value, but other coordinate systems and one- or three-dimensional visualizations are possible.





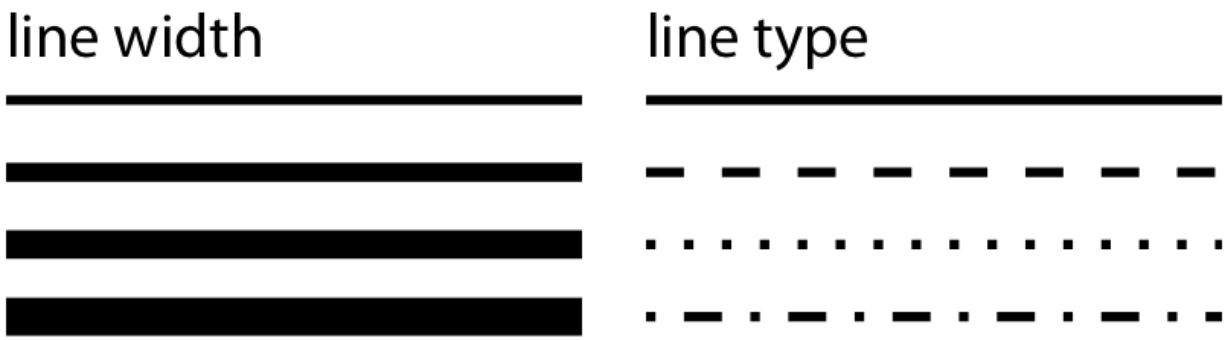
Components of graphical elements-2

- All graphical elements have a shape, a size, and a color.



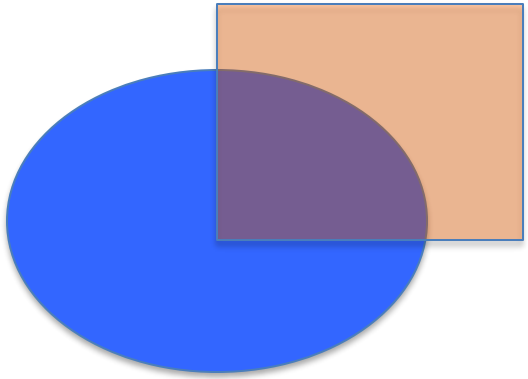
Components of graphical elements-3

- Even if we are preparing a black-and-white drawing, graphical elements need to have a color to be visible: for example, black if the background is white or white if the background is black.
- Finally, to the extent we are using lines to visualize data, these lines may have different widths or dash–dot patterns.



Components of graphical elements-4

- There are many other aesthetics we may encounter in a data visualization.
- For example, if we want to display text, we may have to specify `font family`, **font face**, and font size, and if graphical objects overlap, we may have to specify whether they are partially transparent.



Types of Data

- All aesthetics fall into one of two groups: those that can represent continuous data and those that cannot.
- Continuous data values are values for which arbitrarily fine intermediates exist.
  - *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*
- Position, size, color, and line width can represent continuous data, but shape and line type can usually

Which of the followings can be considered continuous?

A. Temperature
B. Height
C. Weight
D. All of the above
E. None of the above

Beyond numerical data

- Numerical values are only two out of several types.
- Data can come in the form of discrete categories, in the form of dates or times, and as text.
- 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 (dog, cat, fish),*
  - *factors can also be ordered (ordinal), when there is an intrinsic order among the levels of the factor (good, fair, poor).*

Which variable can be ordered and considered ordinal?

A. Color, i.e., red, green, blue, yellow
B. Size of coffee cups, i.e., tall, medium, grande
C. Covid test result, i.e., positive, negative
D. All of the above
E. None of the above

Types of variables

Type of variable	Examples	Appropriate scale	Description
Quantitative/numerical continuous	1.3, 5.7, 83, 1.5×10^{-2}	Continuous	Arbitrary numerical values. These can be integers, rational numbers, or real numbers.

Types of variables

Type of variable	Examples	Appropriate Scale	Description
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.

Types of variables

Type of variable	Examples	Appropriate Scale	Description
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.

Types of variables

Type of variable	Examples	Appropriate Scale	Description
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).

Types of variables

Type of variable	Examples	Appropriate Scale	Description
Text	The quick brown fox jumps over the lazy dog.	None, or discrete	Free-form text. Can be treated as categorical if needed.

Types of variables-Example

- The first few rows of a dataset providing the daily temperature normals (average daily temperatures over a 30-year window) for four US locations.
- This table contains five variables:
  - *Month: ordered factor,*
  - *Day: discrete numerical value,*
  - *Location: unordered factor,*
  - *Station ID: unordered factor,*
  - *Temperature: a continuous numeric*

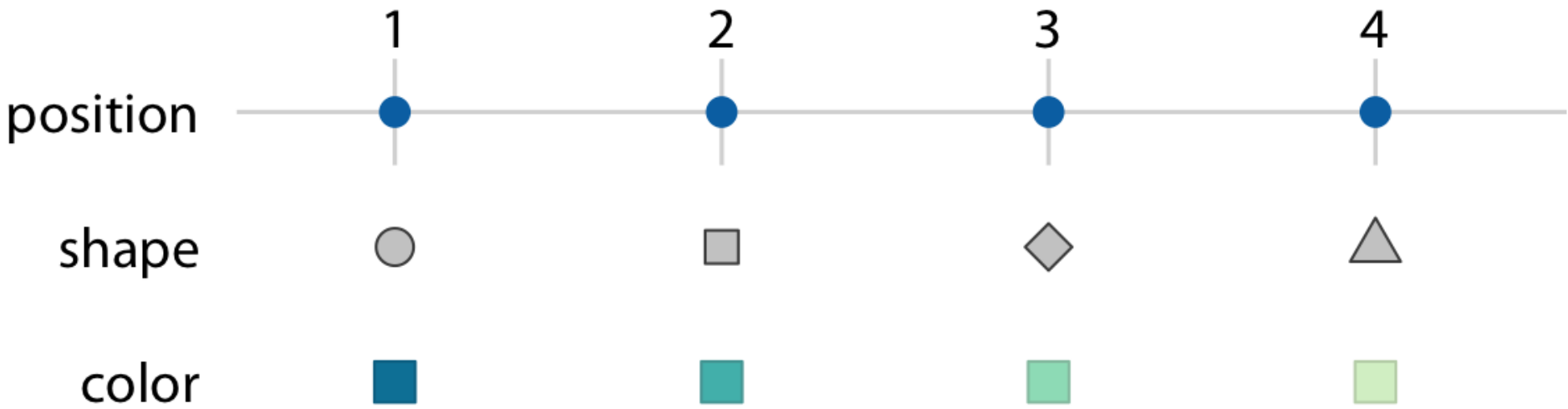
Month	Day	Location	Station ID	Temperature (°F)
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319	51.2

Scales map data values onto aesthetics

- To map data values onto aesthetics, we need to specify which data values correspond to which specific aesthetics values.
  - *If our graphic has an x axis, then we need to specify which data values fall onto particular positions along this axis.*
  - *We may need to specify which data values are represented by particular shapes or colors.*
  - *This mapping is created via scales, which defines a unique mapping between data and aesthetics*
  - *A scale must be one-to-one, such that for each specific data value there is exactly one aesthetics value and vice versa.*
  - *If a scale isn't one-to-one, then the data visualization becomes ambiguous.*

Scales map data values onto aesthetics

- Scales link data values to aesthetics (mapping).
  - *The numbers 1 through 4 have been mapped onto*
  - *a position scale,*
  - *a shape scale, and*
  - *a color scale.*
  - *For each scale, each number corresponds to a unique position,*



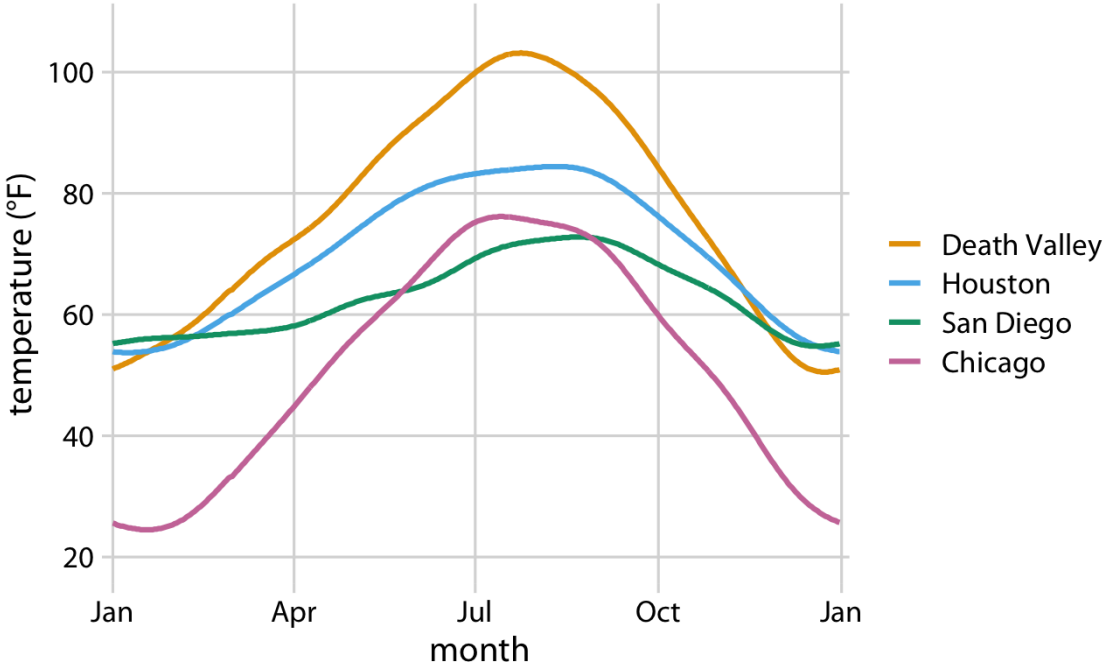
How would you decide on axis …

- if want to explore temperature over time for a given Location ?

Month	Day	Location	Station ID	Temperature (°F)
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

Mapping temperatures-1

- Let's map temperature onto the y axis, day of the year onto the x axis, and location onto color, and visualize these aesthetics with solid lines.
- The result is a standard line plot showing the temperature normals at the four locations as they change during the year.

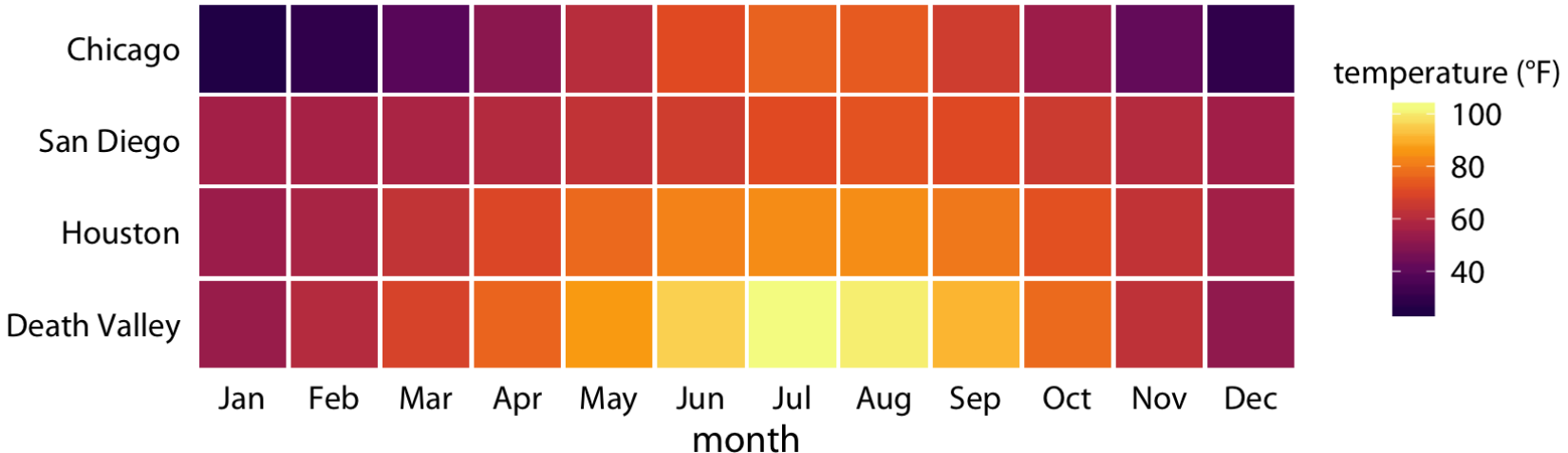


Mapping temperatures-2

- It is up to us which variables we map onto which scales.
- Instead of mapping temperature onto the y axis and location onto color, we can do the opposite.
- Because now the key variable of interest (temperature) is shown as color, we need to show sufficiently large colored areas for the colors to convey useful information.
- We can also choose squares instead of lines, one for each month and location which can be colored by the average temperature normal for each month.

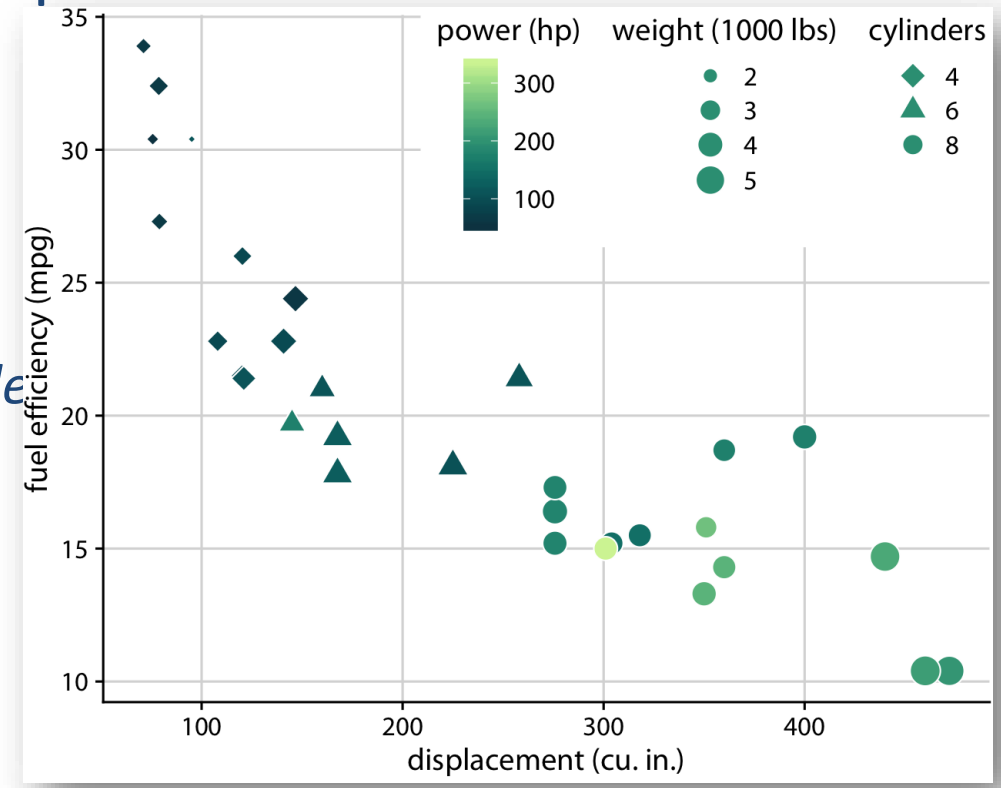
Mapping temperatures-3

- Two position scales (month: x-axis and location: y-axis), but neither is a continuous scale.
- Month is an ordered factor with 12 levels and location is an unordered factor with 4 levels. Position scales are both discrete
- The different levels of the factor at an equal spacing on the axis.
- If the factor is ordered (month), levels follow appropriate order.
- If the factor is unordered (location), then the order is arbitrary.



More scales

- Earlier had two position & one color scales (typical)
- Five separate scales to represent data:
  - *i. x axis (displacement)*
  - *ii. y axis (fuel efficiency)*
  - *iii. color (power)*
  - *iv. size (weight)*
  - *v. shape (number of cylinde*

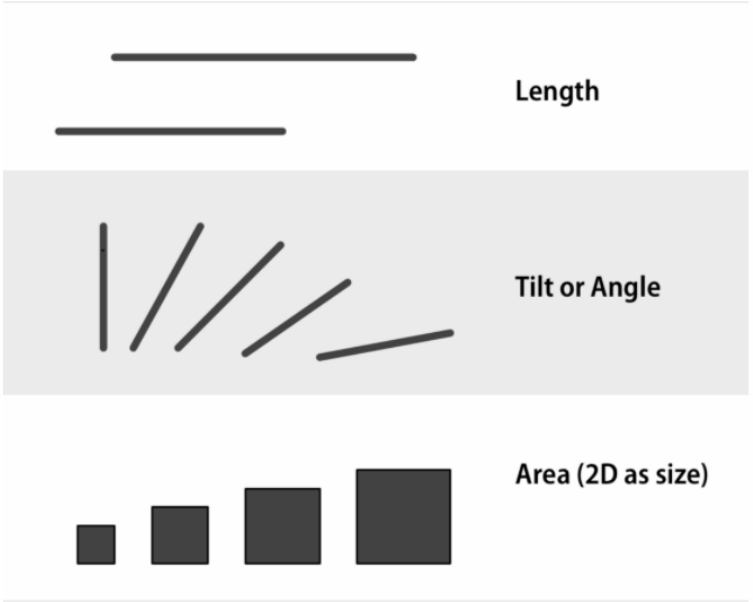


Question

- If our variable is continuous, it would be better to represent it as a series of shapes.
  - *A. True*
  - *B. False*

Channels for representing data

- If we have ordered data, then we should try to encode it as a position on a common scale
- Encoding numbers as lengths (absent a scale) works too, but not as effectively.
- Encoding them as areas will make comparisons less accurate again, and so on.



Channels for representing data

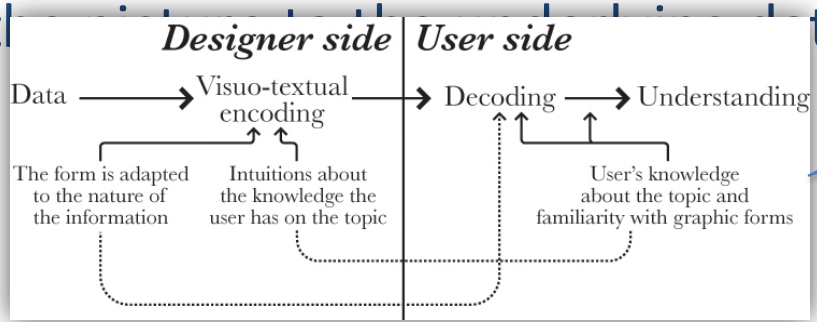
- Perceptual details impact the effectiveness.
- If we have a measure with four categories ordered from lowest to highest, we might correctly decide to represent it using a sequential gradient.
- Wrong sequence of colors is hard to interpret, or actively misleading.
- Bad set of hues for an unordered categorical variable, might result in both unpleasant and misleading.

A note on aesthetics

- Different *channels or mappings* are not in themselves kinds of graphs, but are just the *elements or building blocks* for graphs.
- When we choose how to encode a variable as a position, a length, an area, a shade of grey, or a color, we have made an *important decision that narrows down* what the resulting plot can look like.
- This is not the same as deciding what type of plot it will be, in the sense of choosing whether to make a dot plot or a bar chart, a histogram or a frequency polygon, and so on.

Visual tasks and decoding graphs

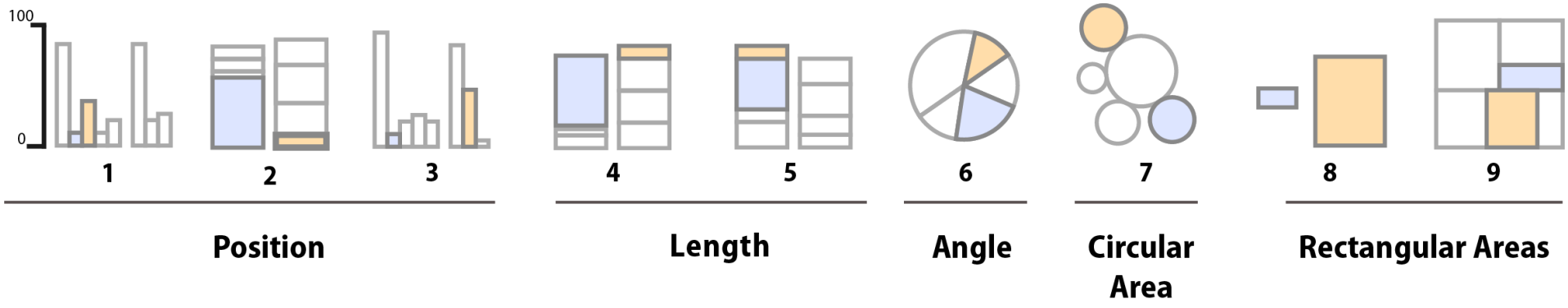
- To understand a graph, the viewer needs to know a lot of general information, such as
  - *what a variable is,*
  - *what an x-y coordinate plane looks like,*
  - *why we might want to compare two variables on one, and*
  - *the convention of putting the supposed cause or "independent" variable on the x-axis.*
- Even if the viewer understands all these things, they must still perform the visual task of interpreting the graph.
- Even well-informed viewers may do worse than we think when connecting t ta



Recall

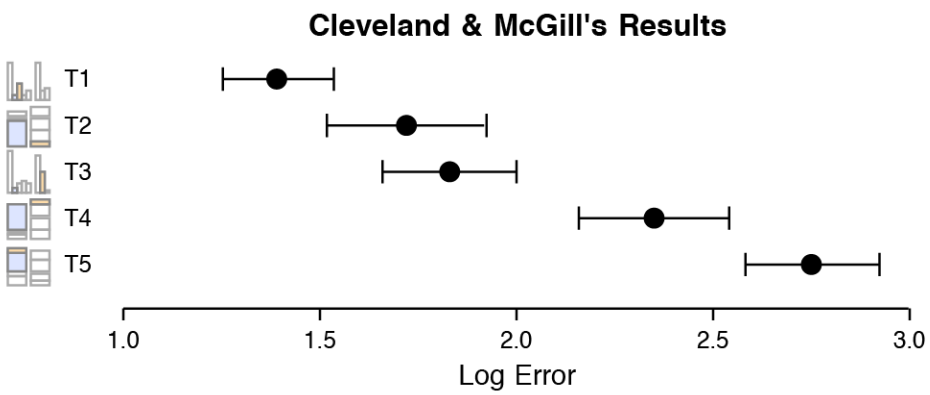
Visual tasks and decoding graphs

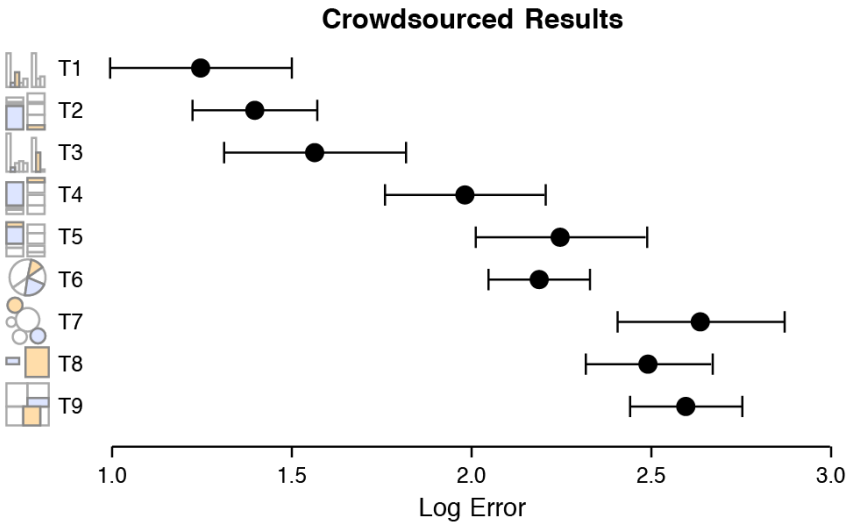
- Participants were asked to make comparisons of highlighted portions of each chart type, and say which was smaller. (by Heer and Bostock, following Cleveland and McGill)



Visual tasks and decoding graphs

- Most often, research subjects were asked to
  - *estimate two values within a chart*
  - *compare values between charts*
- Area comparisons perform even worse than the (justifiably) much-maligned pie chart
- There are better and worse ways when user must estimate and compare.



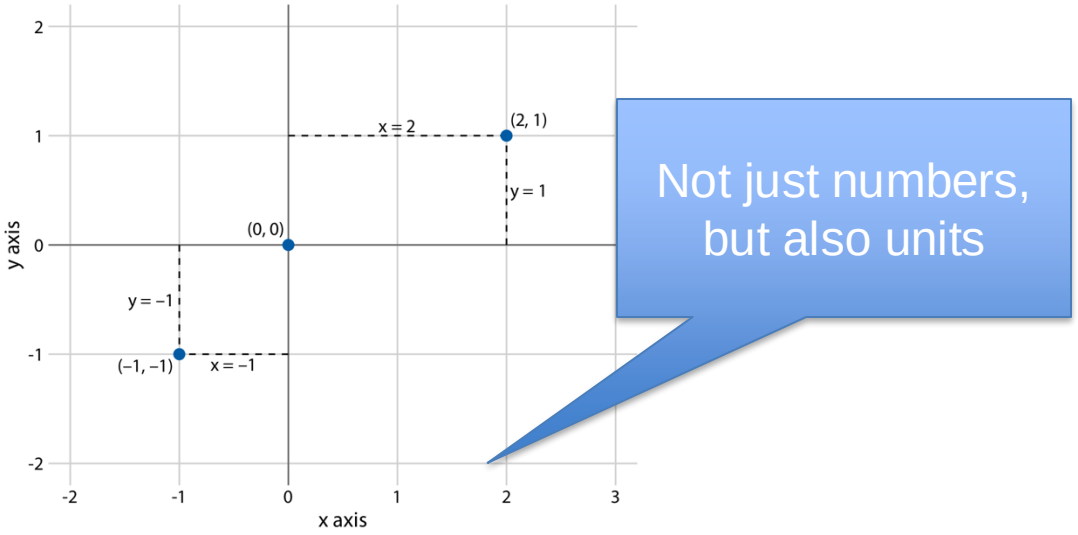


Coordinate Systems and Axes

- To make any sort of data visualization, we need to define position scales, which determine where in a graphic different data values are located.
- We cannot visualize data without placing different data points at different locations, even if we just arrange them next to each other along a line.
- For regular 2D visualizations, two numbers are required to uniquely specify a point, and therefore we need two position scales. These two scales are usually but not necessarily the x and y axes of the plot.

Cartesian Coordinates

- The most widely used coordinate system.
- The x and y axes run orthogonally to each other, data values are placed in an even spacing.
- The two axes are continuous position scales, and they can represent both positive and negative real numbers.

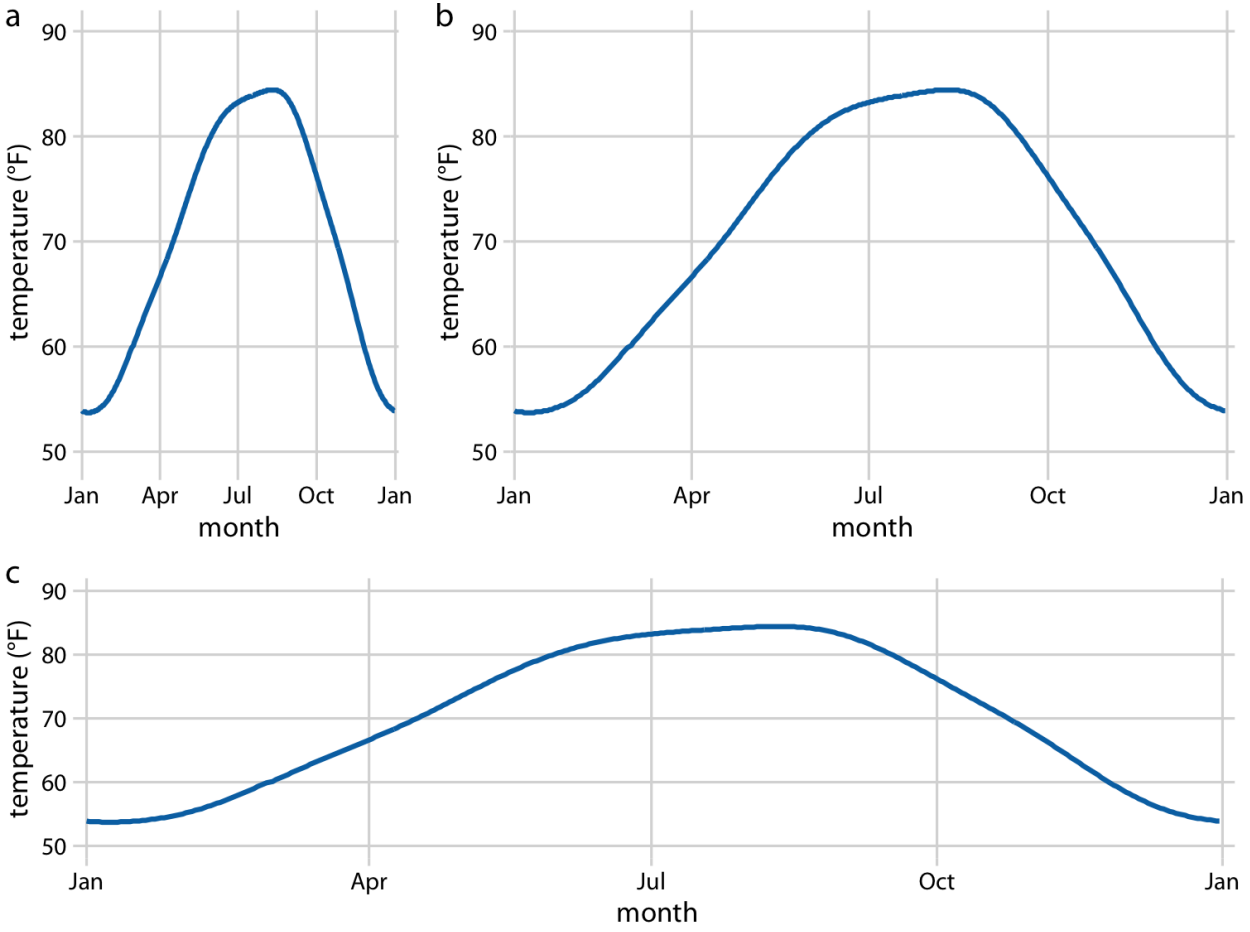


Cartesian Coordinates

- Can have two axes representing two different units.
- If in different units, we can stretch or compress one relative to the other to maintain a valid visualization.
- Which version is preferable may depend.
- A tall and narrow figure emphasizes change along the y axis and a short and wide figure does the opposite.
- Ideally, we want to choose an aspect ratio that ensures that any important differences in position are noticeable.

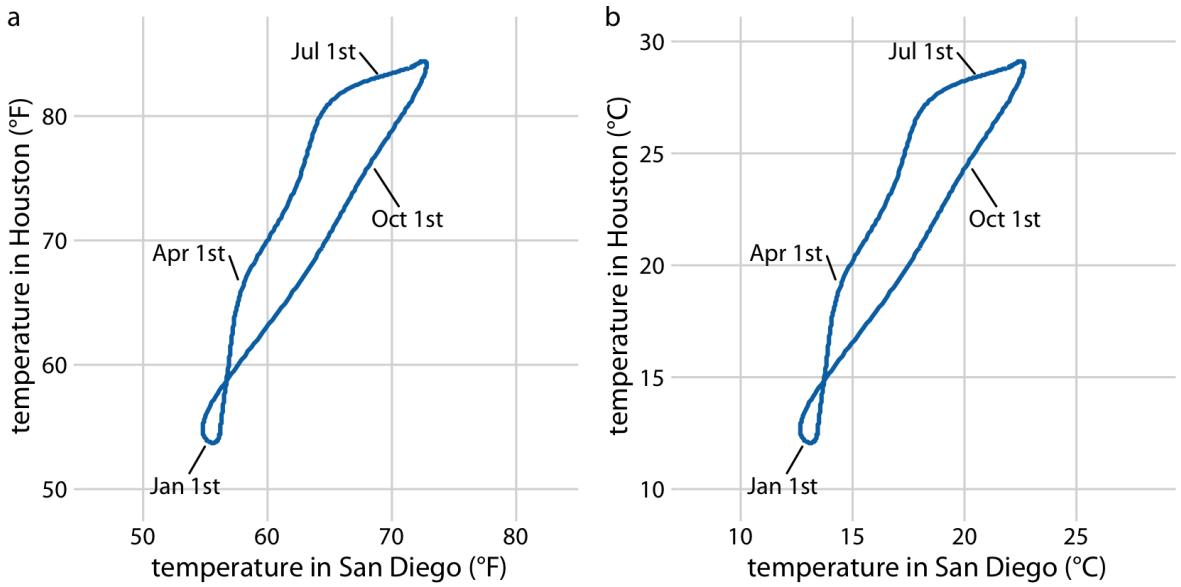
Cartesian Coordinates

- All three parts are valid. (Data source: NOAA)



Cartesian Coordinates

- If the x and y axes are measured in the same units, then the grid spacings for the two axes should be equal.
  - *We can plot the temperature in Houston vs. in San Diego.*
  - *We need to make sure that the grid lines form perfect squares.*

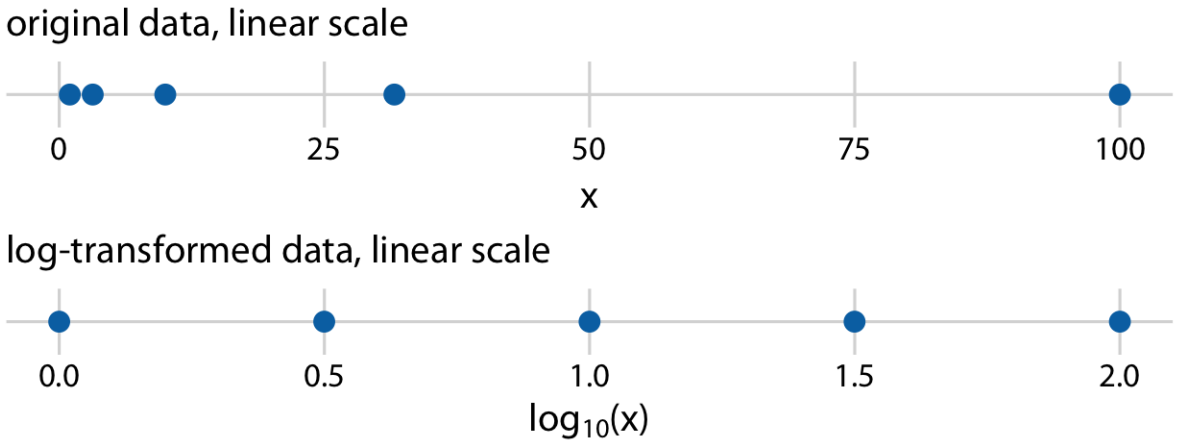


Nonlinear Axes

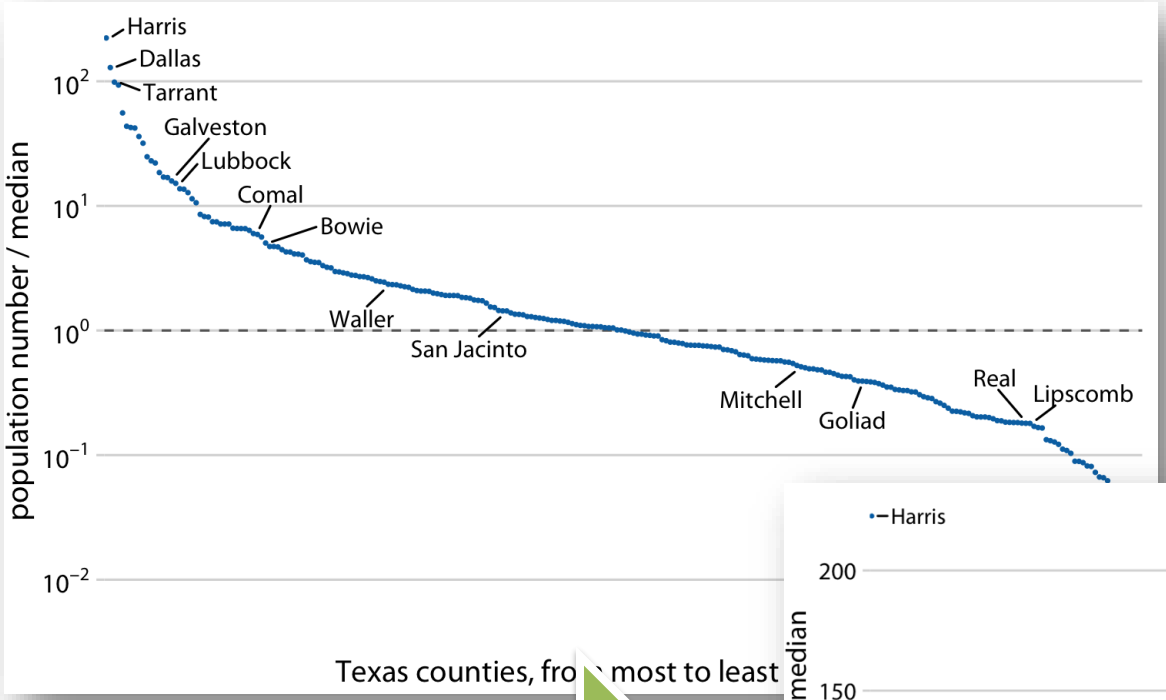
- There are scenarios where nonlinear scales are preferred.
- In a nonlinear scale, even spacing in data units corresponds to uneven spacing in the visualization, or conversely even spacing in the visualization corresponds to uneven spacing in data units.
- The most commonly used nonlinear scale is the *logarithmic* (*log) scale.*
- Need to log-transform the data values while exponentiating the numbers that are shown along the axis grid lines.

Nonlinear Axes

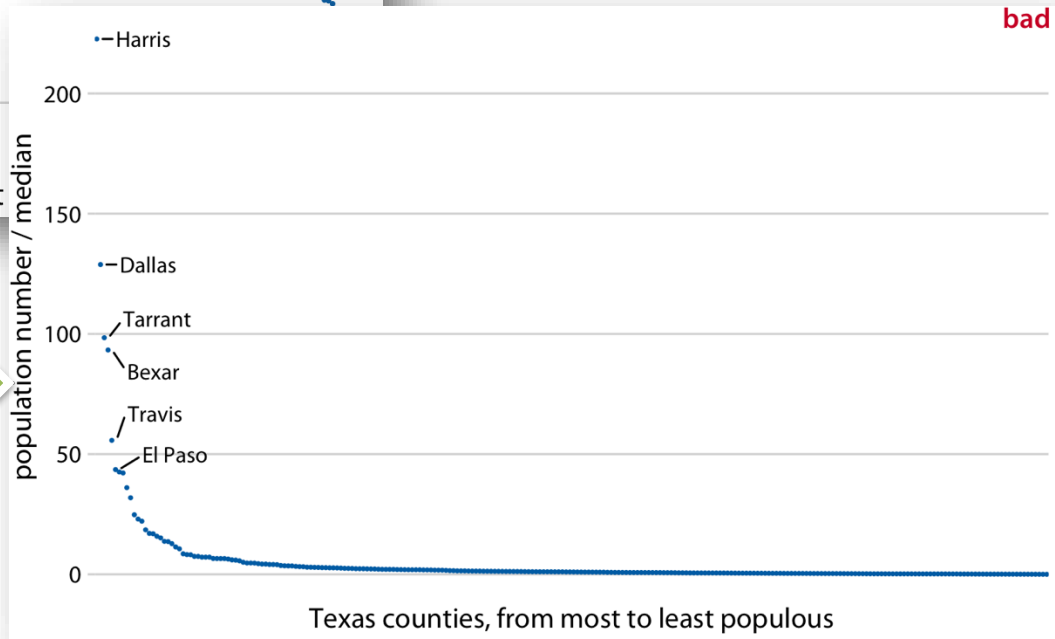
- The numbers 1, 3.16, 10, 31.6, and 100 placed on linear and log scales.
- The labeling for a logarithmic scale is preferable



Ratio data w/ log and linear scale

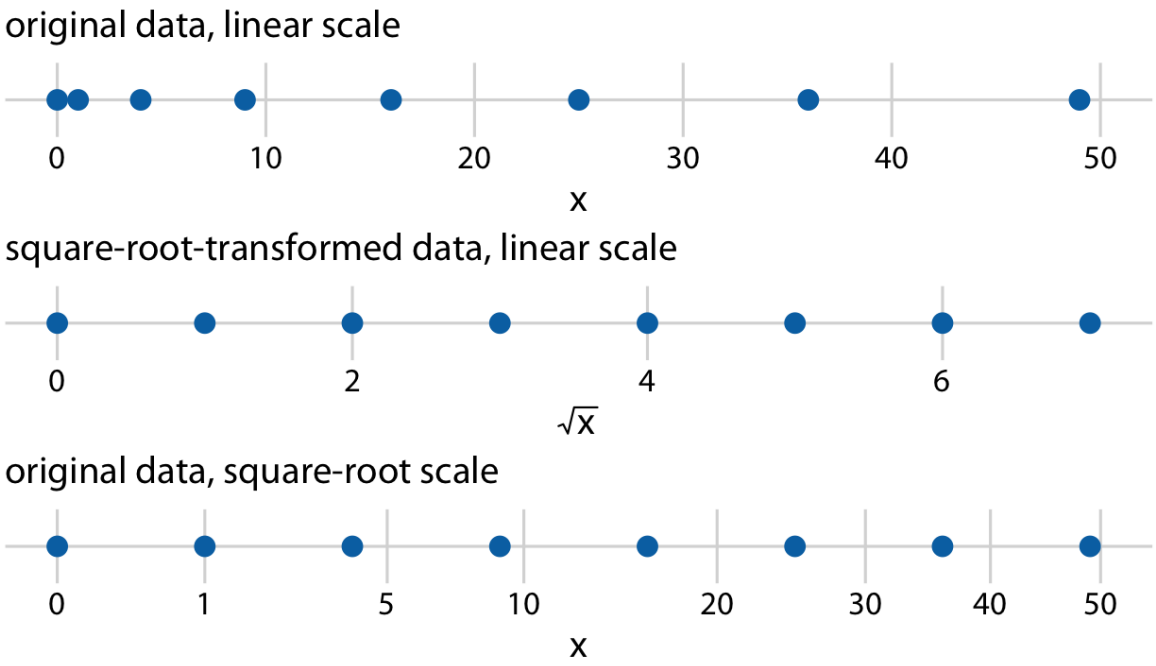


obscures the differences!



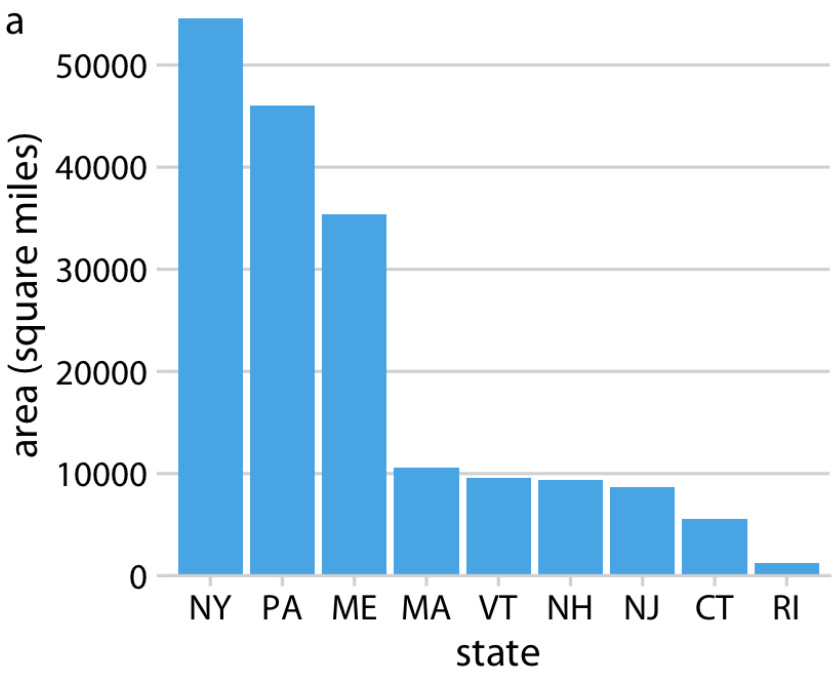
Square-root scale

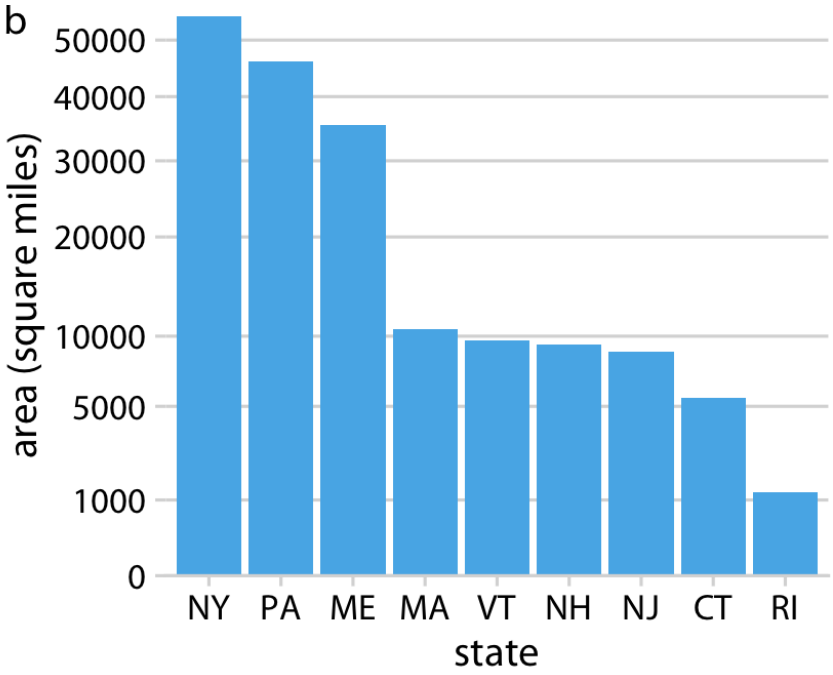
- A square-root scale compresses larger numbers into a smaller range, but unlike a log scale, it allows for the presence of 0.
- There may some flaws, but appropriate applications exist.



Square-root scale

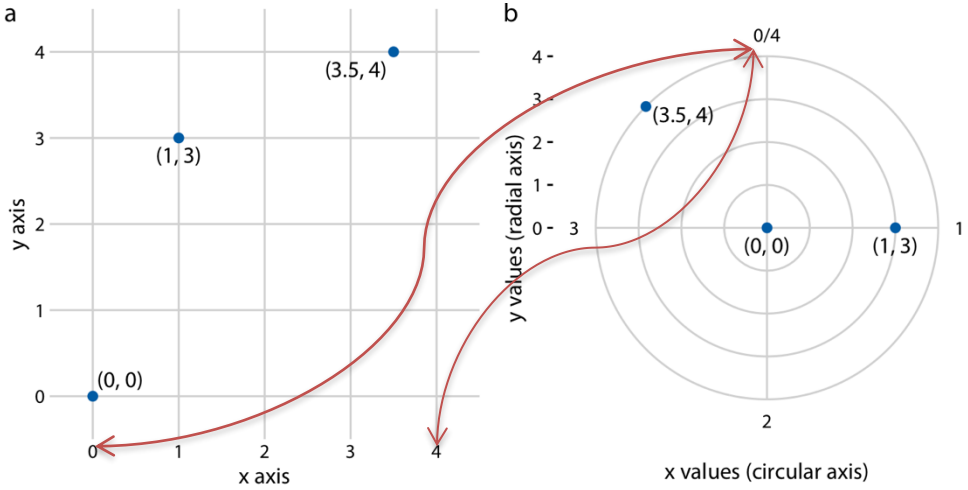
- The relative time it will take to drive across each state is more accurately represented by the figure on the square-root scale than the figure on the linear scale





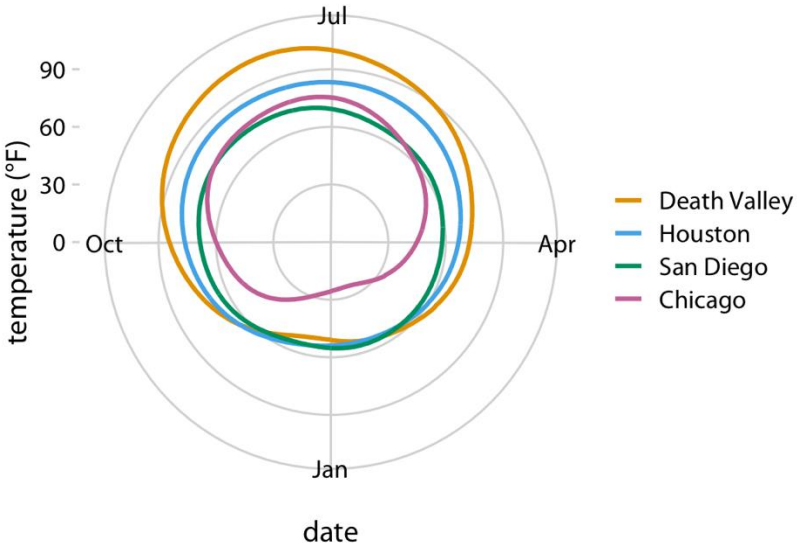
Coordinate Systems with Curved Axes

- In the polar coordinate system, we specify positions via an angle and a radial distance from the origin, and therefore the angle axis is circular.
- 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.
  - *Dec 31 is the last day & one day before the first day.*



Coordinate Systems with Curved Axes

- December 31st is the last day of the year, but it is also one day before the first day of the year.
- The radial distance from the center point indicates the daily temperature in Fahrenheit, and the days of the year are arranged counterclockwise starting with Jan. 1st at the 6:00 position

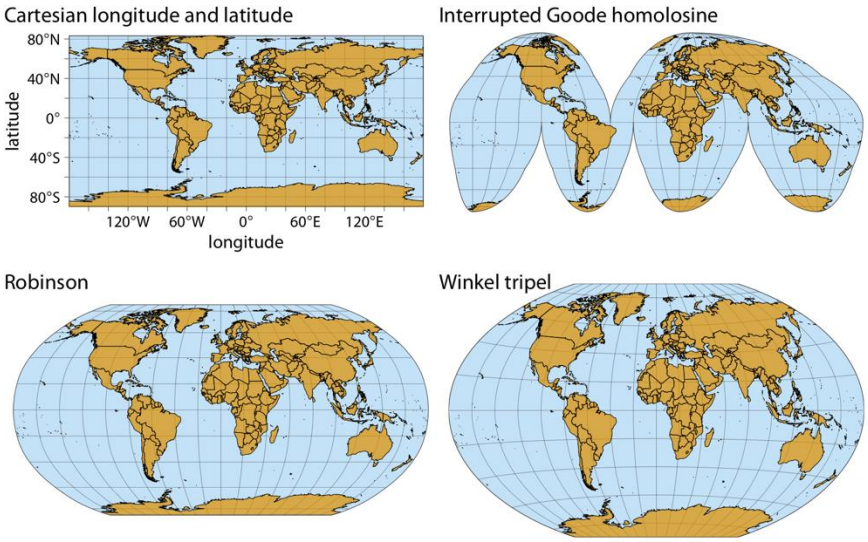


Coordinate Systems with Curved Axes

- We also encounter curved axes is in geospatial data.
- The earth is a sphere and drawing latitude and longitude as Cartesian axes is misleading.
- Nonlinear projections are used to minimize artifacts and strike different balances between conserving areas relative to the true shape lines on the globe

Coordinate Systems with Curved Axes

- Interrupted Goode homolosine projection perfectly represents true surface areas, at the cost of dividing some land masses into separate pieces.
- The Robinson and Winkel tripel projections both strike a balance between angular and area distortions, and they are commonly used for maps of the entire globe.



In-class Assignment

Please see the mtcars data with the descriptions below:

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	Engine (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears

1. Which two variables would you visualize to make a comparison? Why?
2. Choose 4 variables and state which visualization components you would use for them? Line, color, shape?