Data-based Statistical Decision Model

Lecture 4 (Part II) - Data Visualization: Composing/dissecting Data Graphics

Sungkyu Jung

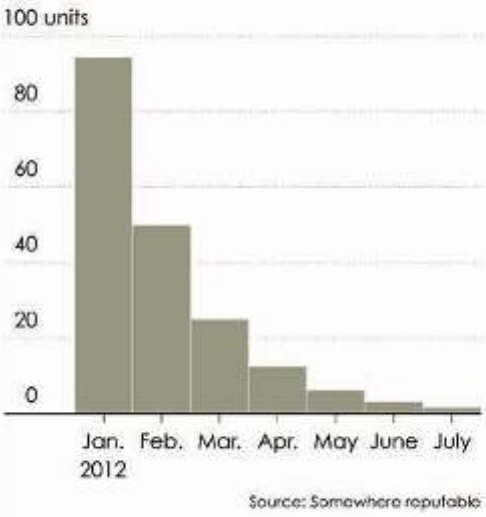
A taxonomy for data graphics

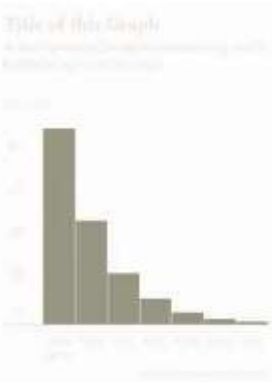
- Nathan Yau (http://flowingdata.com/ (http://flowingdata.com/)) provides a systematic way of thinking about how data graphics convey specific pieces of information, and how they could be improved.
- A complementary grammar of graphics is implemented by Hadley Wickham in the `ggplot2` graphics package in R
- Data graphics can be understood in terms of four basic elements:

1. visual cues
2. coordinate system
3. scale
4. context

Working parts

Several pieces work together to make a graph. Sometimes these are explicitly shown in the visualization and other times they form a visual in the background. They all depend on the data.





Visual Cues

Visualization involves encoding data with shapes, colors, and sizes. Which cues you choose depends on your data and your goals.

Coordinate System

You map data differently with a scatterplot than you do with a pie chart. It's x- and y-coordinates in one and angles with the other; it's cartesian versus polar.

Scale

Increments that make sense can increase readability, as well as shift focus.

Context

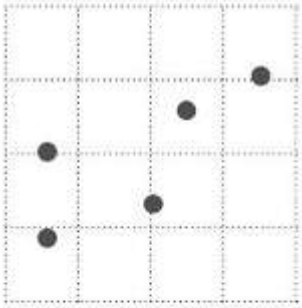
If your audience is unfamiliar with the data, it's your job to clarify what values represent and explain how people should read your visualization.

1. Visual Cues

Visual cues

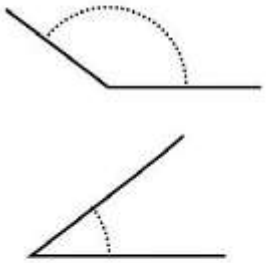
When you visualize data, you encode values to shapes, sizes, and colors.

Position
Where in space the data is

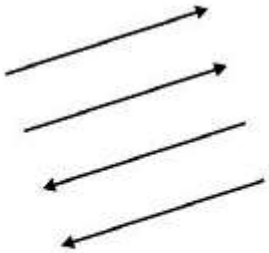


Length
How long the shapes are

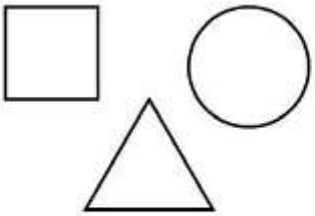
Angle
Rotation between vectors



Direction
Slope of a vector in space

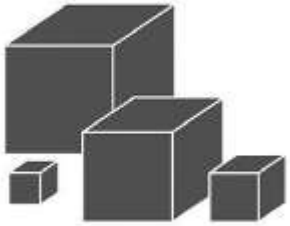


Shapes
Symbols as categories



Area
How much 2-D space

Volume
How much 3-D space



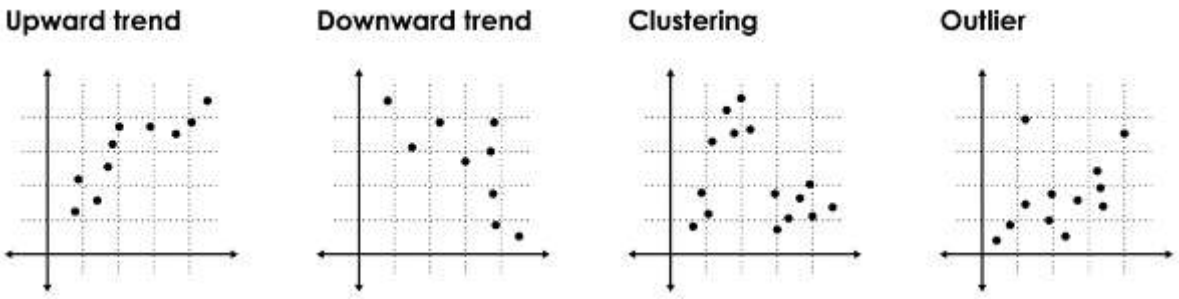
Color saturation
Intensity of a color hue

Color hue
Usually referred to as color

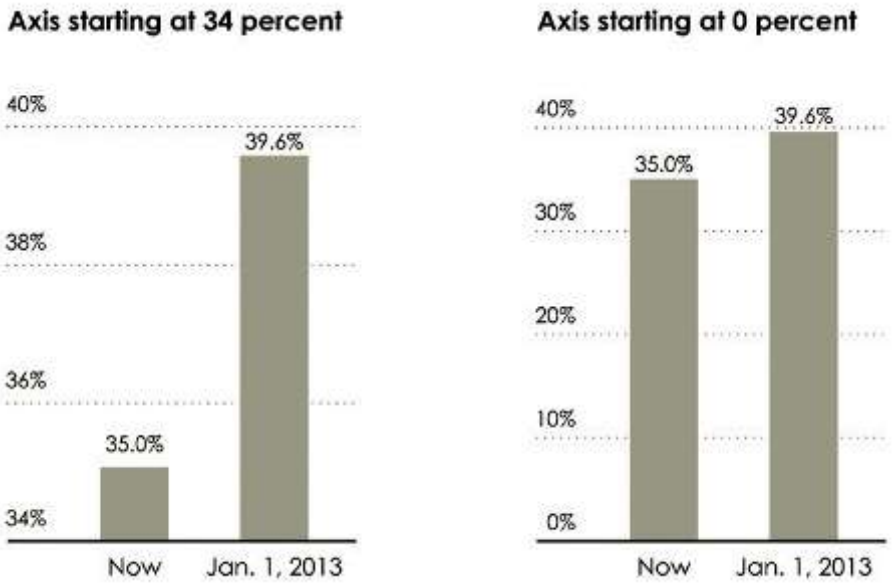
FIGURE 3-3 *Visual cues*

Visual Cues

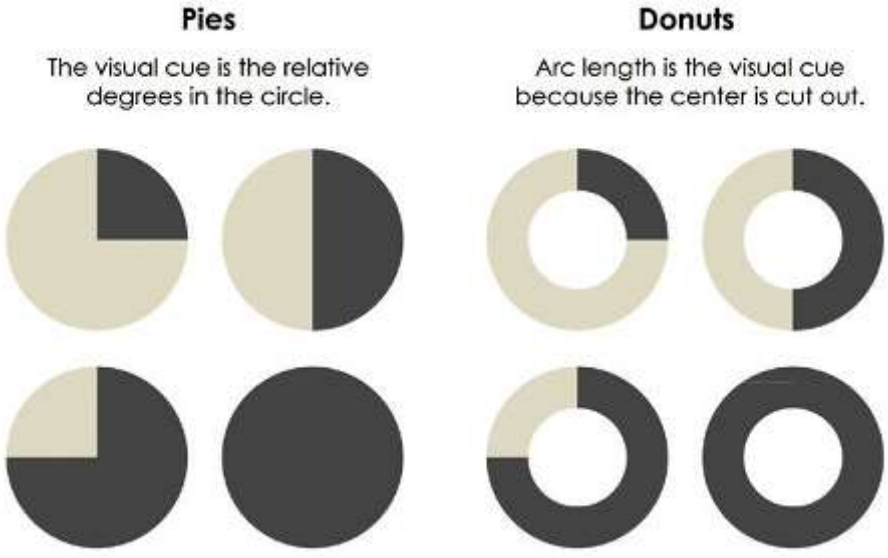
1. Position (numerical) where in relation to other things?



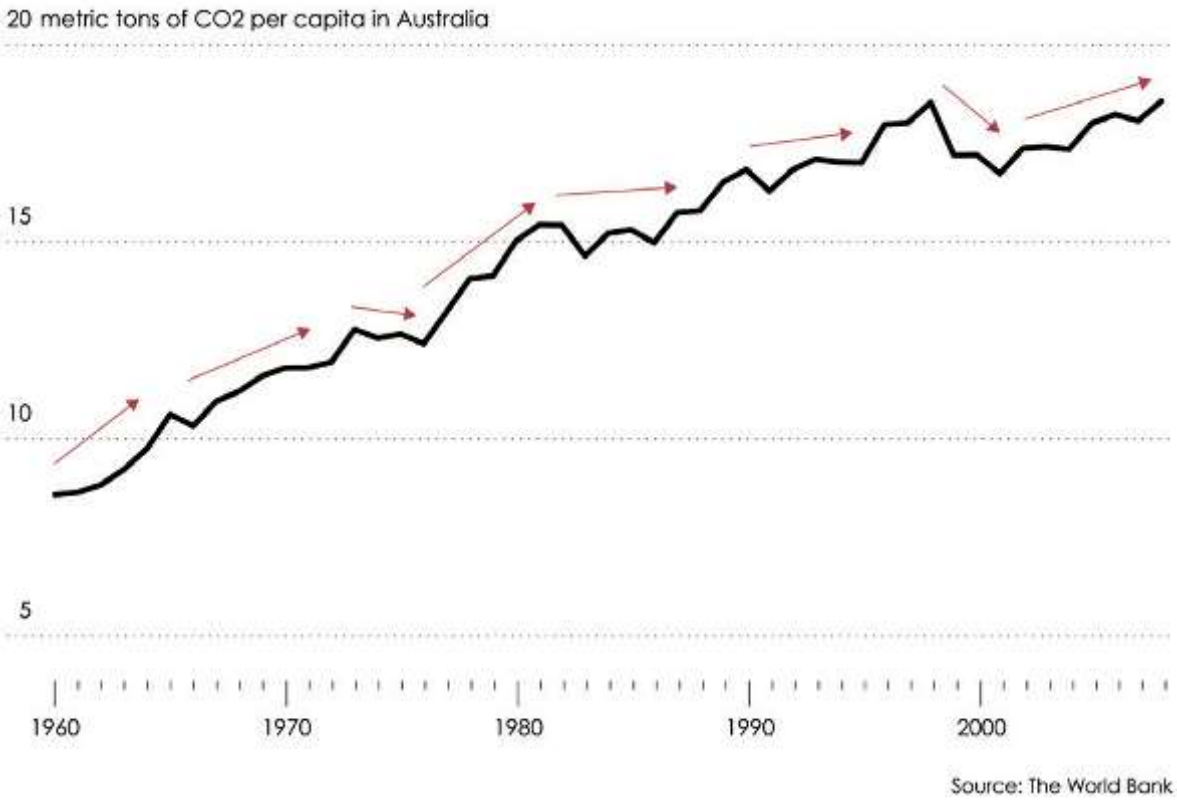
2. Length (numerical) how big (in one dimension)?



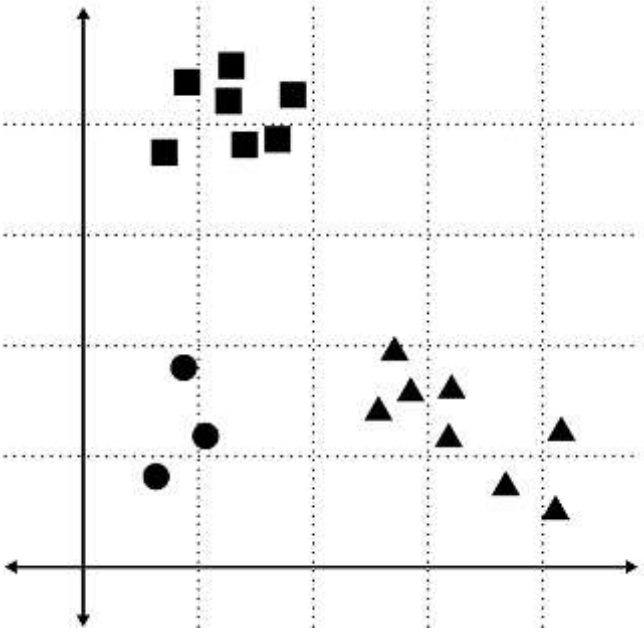
3. Angle (numerical) how wide? parallel to something else?



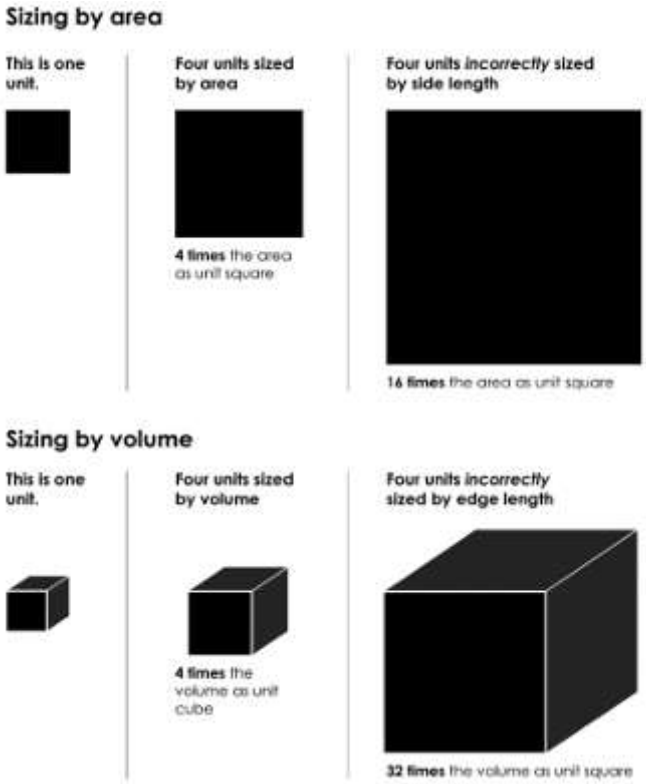
4. Direction (numerical) at what slope? In a time series, going up or down?



5. Shape (categorical) belonging to which group?



6. Area (numerical) how big (in two dimensions)?
7. Volume (numerical) how big (in three dimensions)?



8. Shade and color (color saturation and color hue) to what extent? how severely? Beware of red/green color blindness

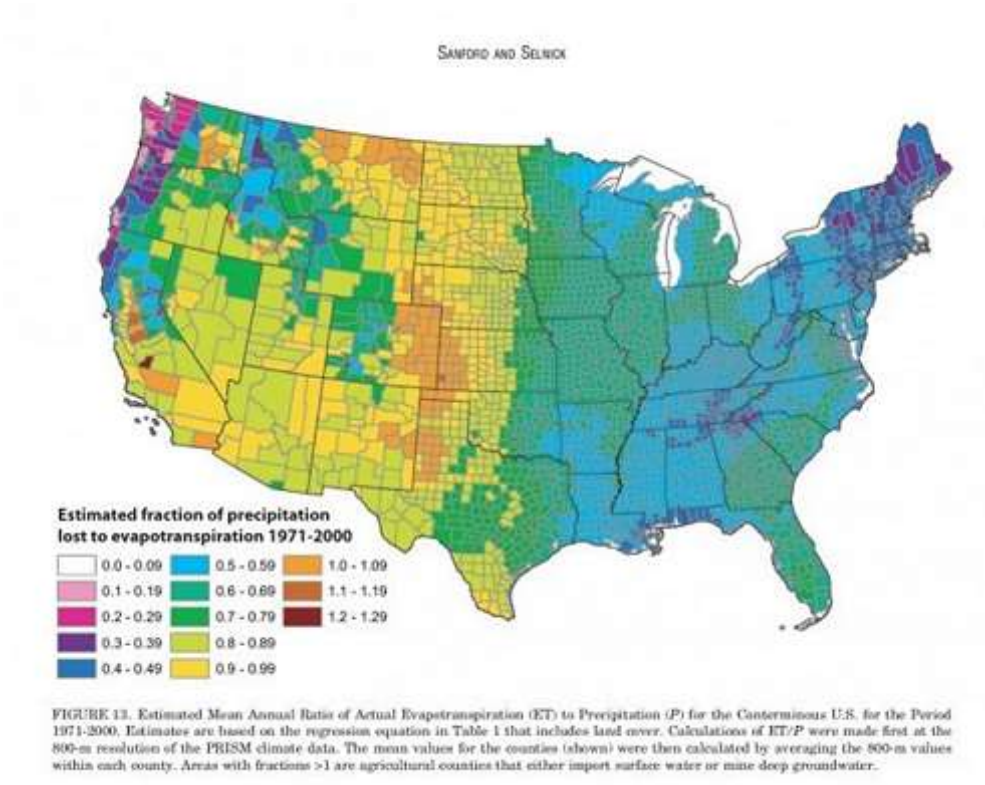
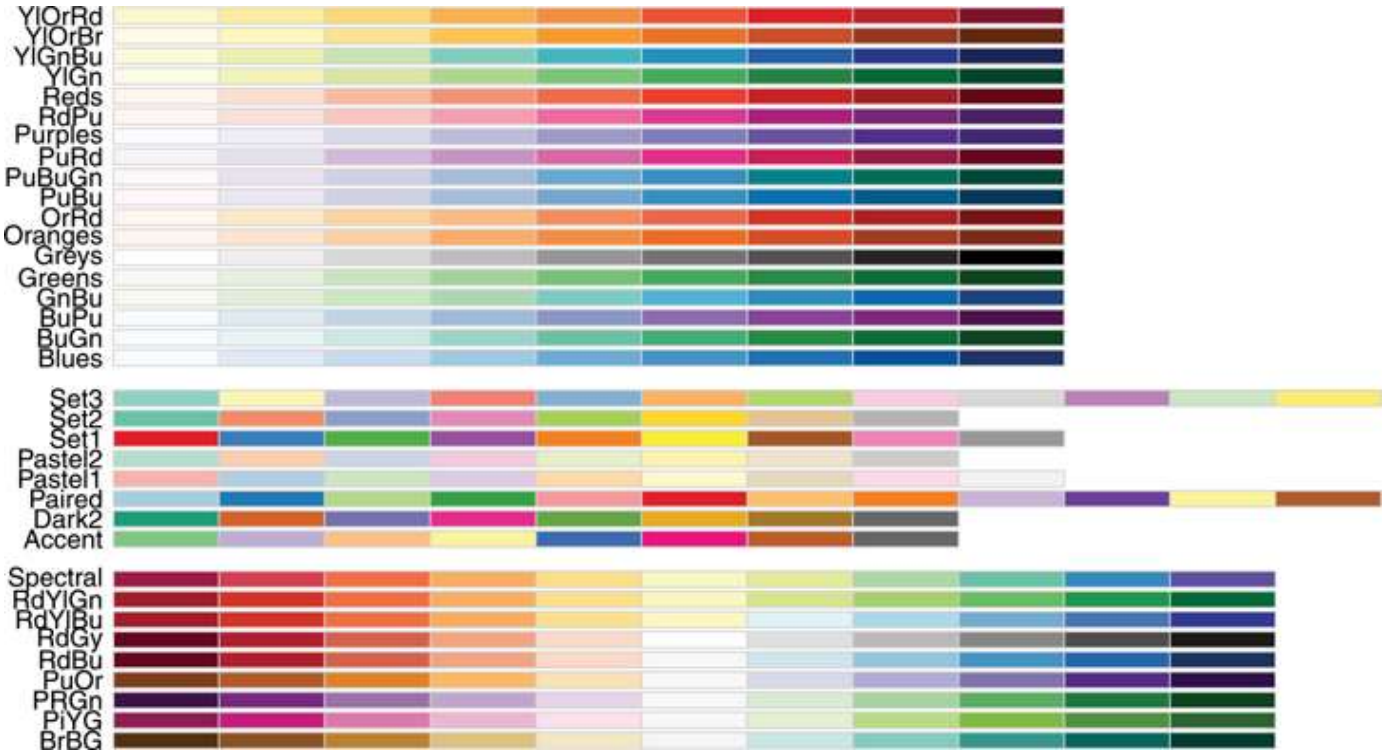


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/P were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

Note: Colors can represent both quantitative and categorical variables, using the following

- **Sequential** The ordering of the data has only one direction.
- **Diverging** The ordering of the data has two directions.
- **Qualitative** There is no ordering of the data



Which visual cues are more effective?

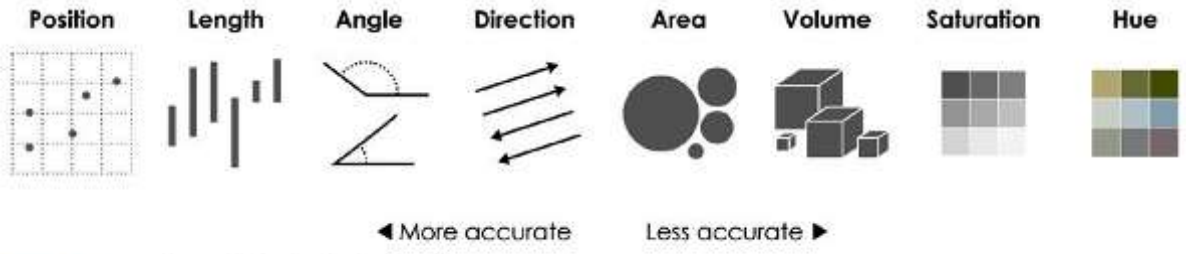
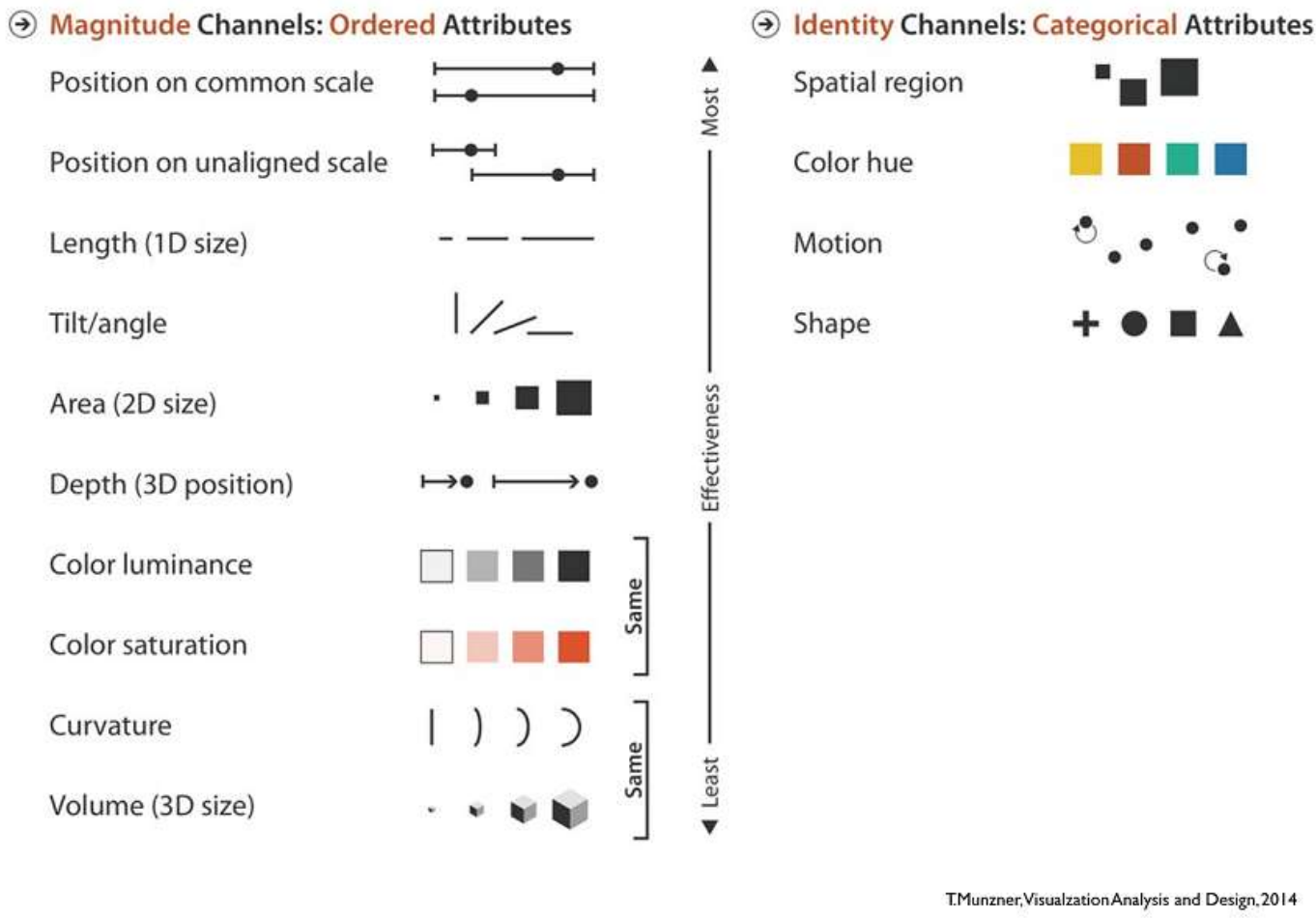


FIGURE 3-12 *Visual cues ranked by Cleveland and McGill*

Which visual cues are more effective? (2)



2. Coordinate systems

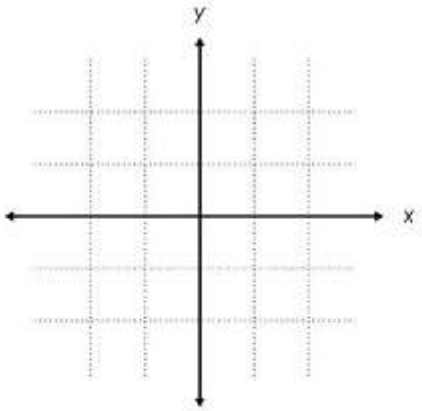
How are the data points organized? While any number of coordinate systems are possible, three are most common:

Coordinate systems

There are a variety of them, from cylindrical to spherical, but these three will cover most of your bases.

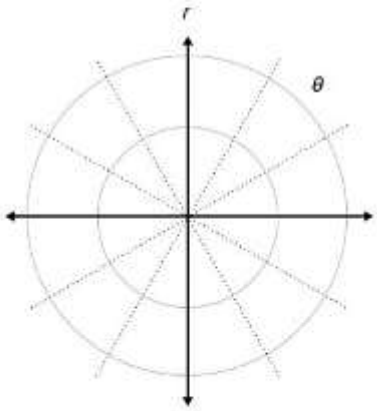
Cartesian

If you've ever made a graph, the x- and y-coordinate system will look familiar to you.



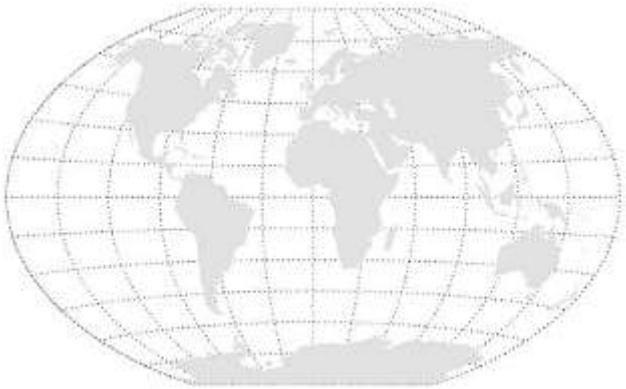
Polar

Pie charts use this system. Coordinates are placed based on radius *r* and angle *θ*.



Geographic

Latitude and longitude are used to identify locations in the world. Because the planet is round, there are multiple projections to display geographic data in two dimensions. This one is the Winkel tripel.



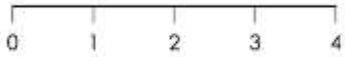
An appropriate choice for a coordinate system is critical in representing one's data accurately, since, for example, displaying spatial data like airline routes on a flat Cartesian plane can lead to gross distortions of reality

3. Scale

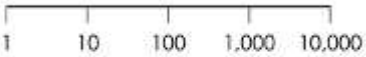
Scales translate values into visual cues. The choice of scale is often crucial. The central question is how does distance in the data graphic translate into meaningful differences in quantity? Each coordinate axis can have its own scale, for which we have three different choices:

1. **Numeric** A numeric quantity is most commonly set on a linear, logarithmic, or percentage scale.
2. **Categorical** A categorical variable may have no ordering (e.g., Democrat, Republican, or Independent), or it may be ordinal (e.g., never, former, or current smoker).
3. **Time** Time is a numeric quantity that has some special properties. First, because of the calendar, it can be demarcated by a series of different units (e.g., year, month, day, etc.). Second, it can be considered periodically as a wrap-around scale.

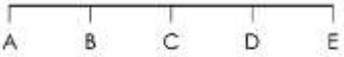
Linear
Values are evenly spaced



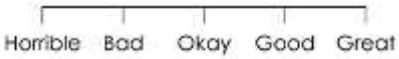
Logarithmic
Focus on percent change



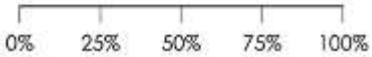
Categorical
Discrete placement in bins



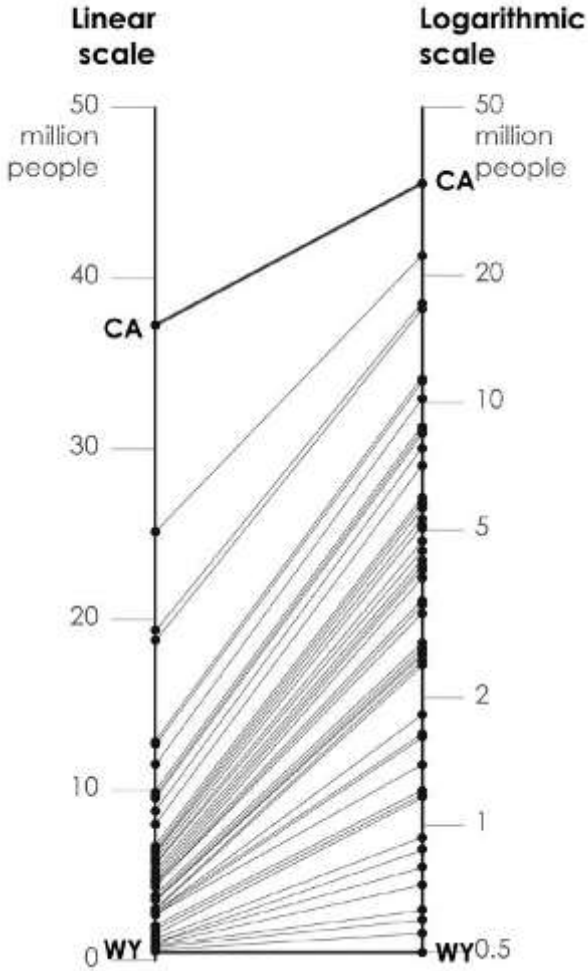
Ordinal
Categories where order matters

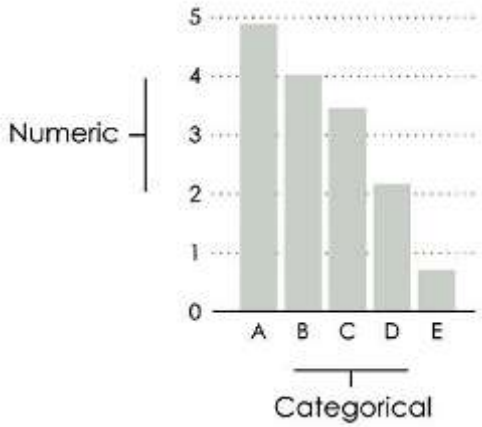


Percent
Representing parts of a whole

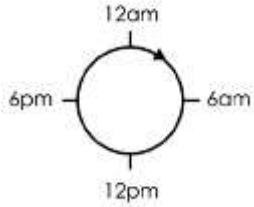


Time
Units of months, days, or hours

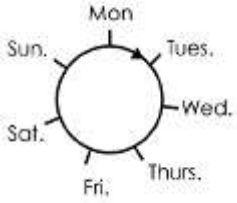




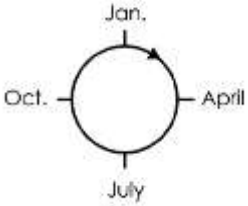
Time of day



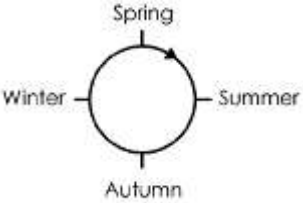
Day of the week



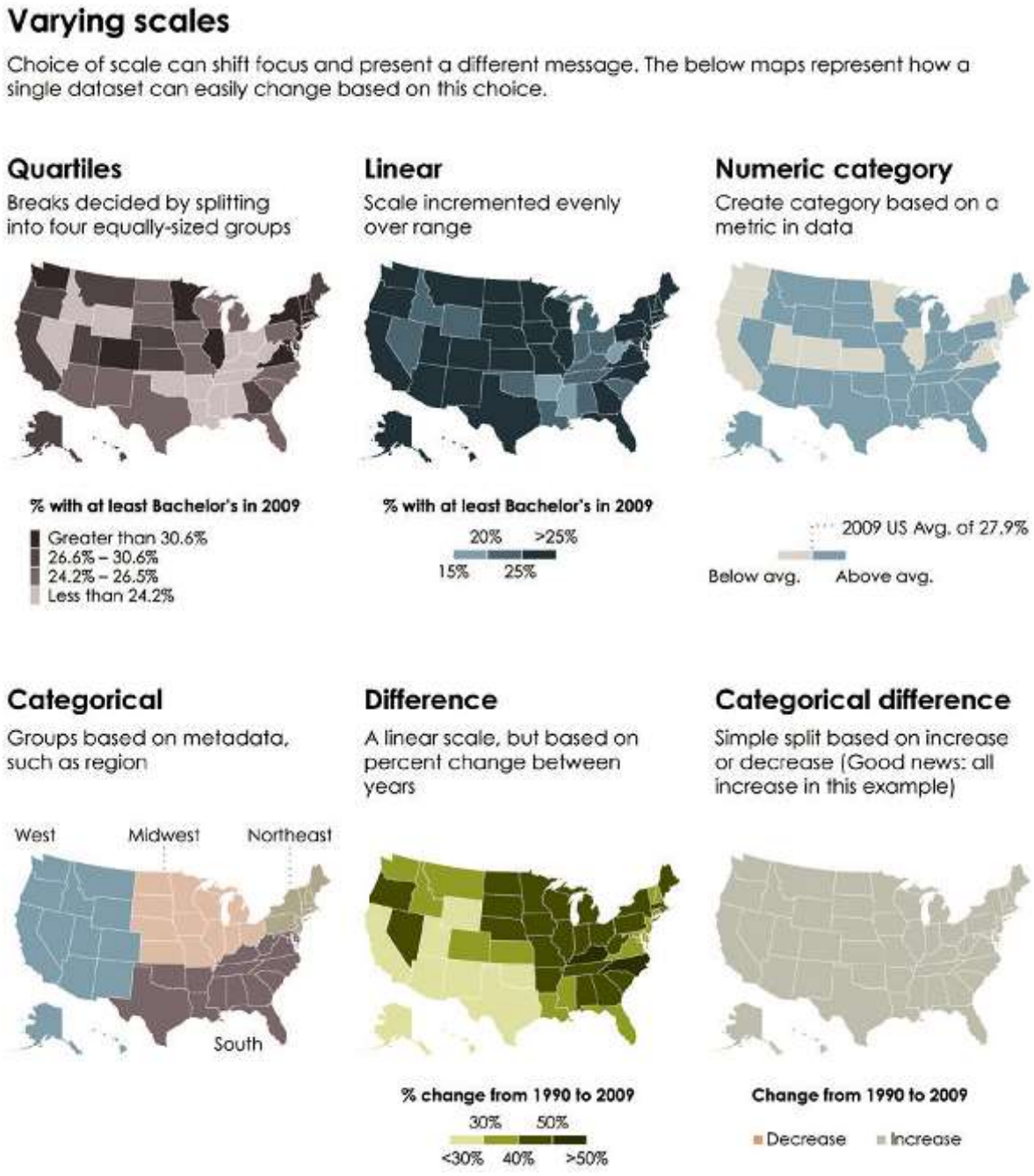
Month in the year



Seasons



Use data transformation (mutation) to choose the most effective scale



4. Context

The purpose of data graphics is to help the viewer make meaningful comparisons. Context can be added to data graphics in the form of

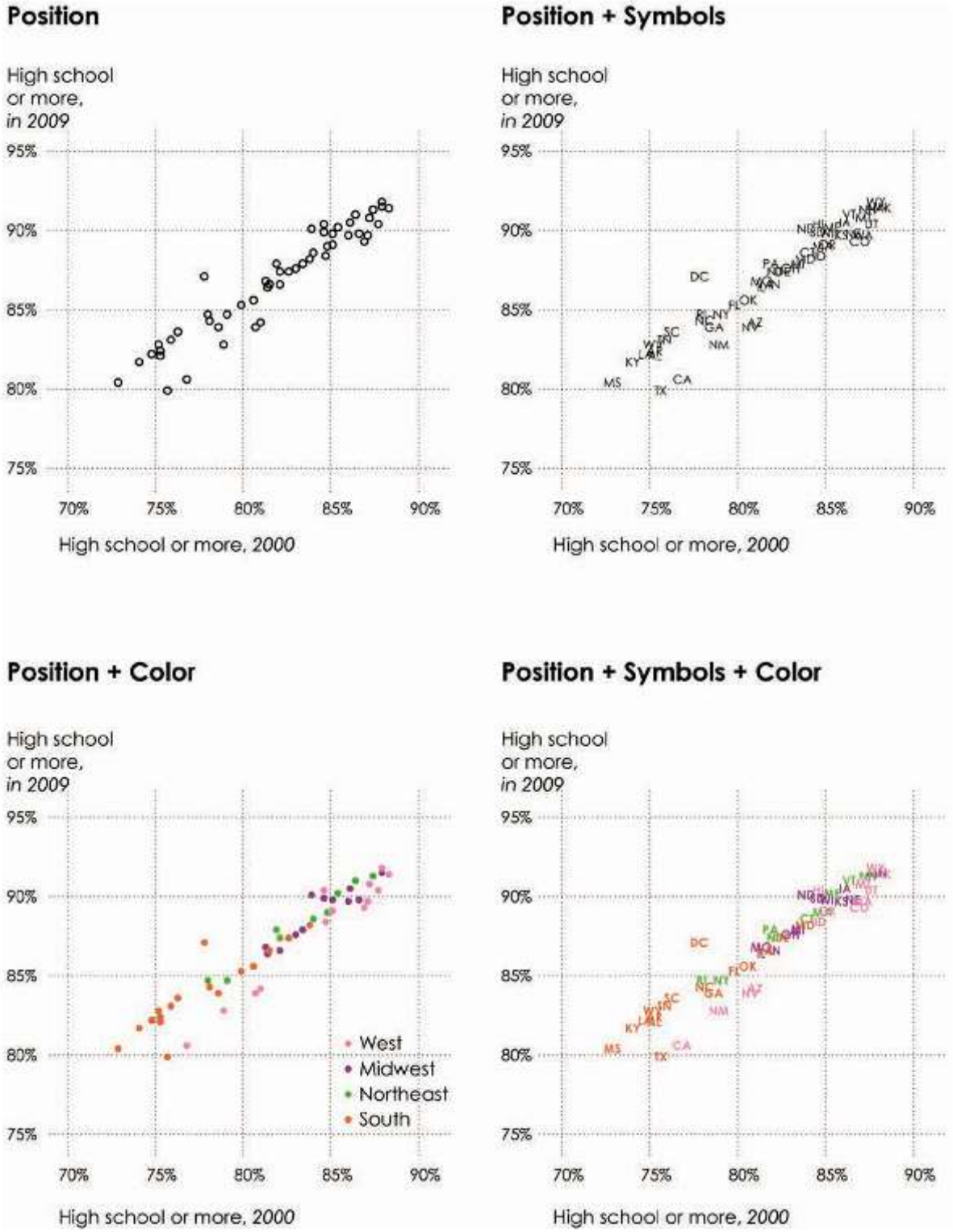
- titles or subtitles
- axis labels
- reference points or lines

For multivariate data

Challenging to condense multivariate information into a two-dimensional image. Use

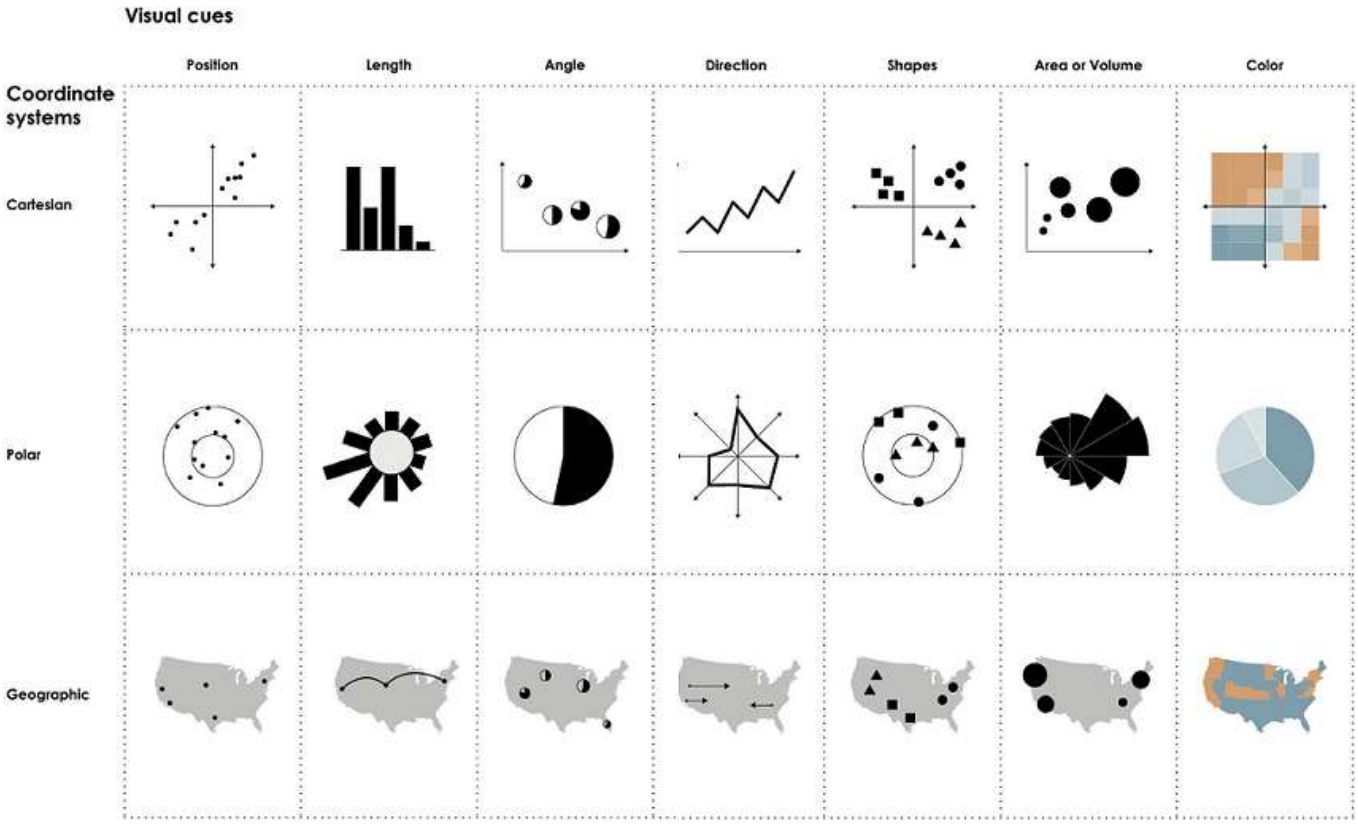
- Small multiples Also known as *facets*

- Layers
- Animation



(We will revisit facets and layers while learning *A Layered Grammar of Graphics*, implemented in `ggplot2`)

Putting it all together

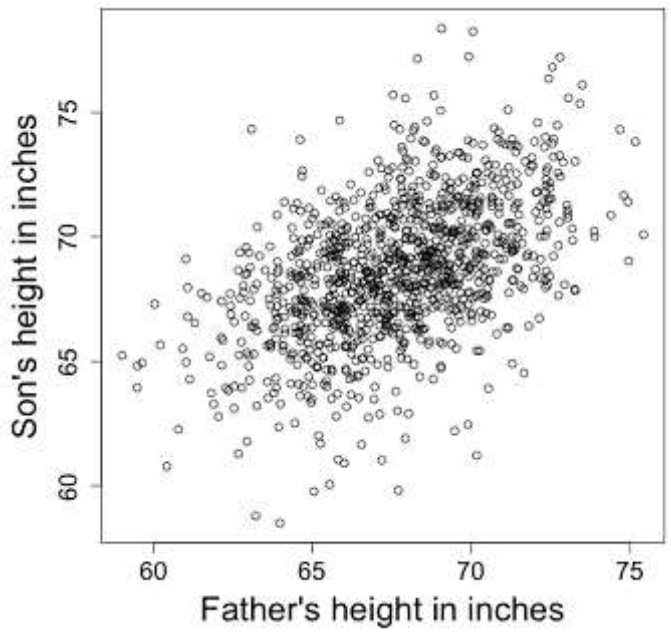


Exercises

For each of data graphics, answer the following:

1. Which variables are used, and what are the types of variables?
2. Which visual cue is used?
3. On which coordinate system, and on which scale?
4. How context is provided?

Exercise 1.

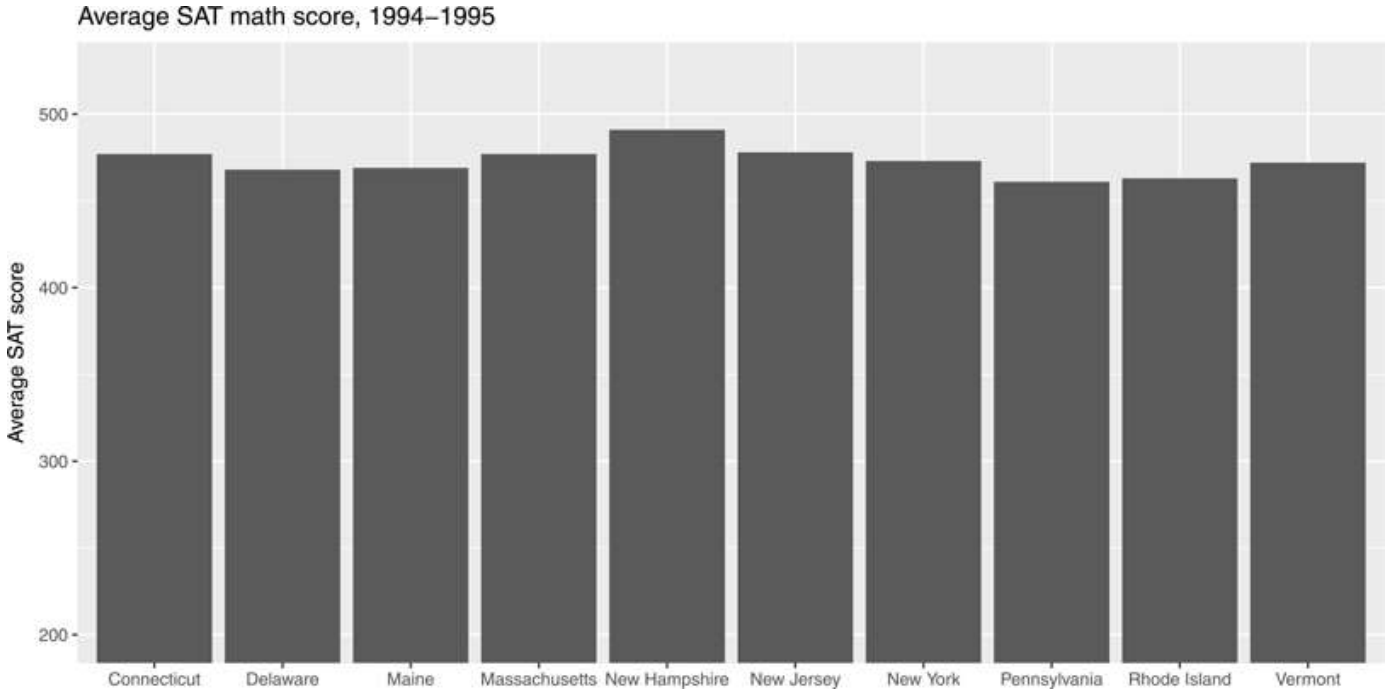


1. *Two quantitative variables* (Son's height and father's height) are used

2. using the visual cue of *position*,
3. in the *Cartesian* plane with *linear* scales
4. Context is provided by the axis lables (to show the positive association).

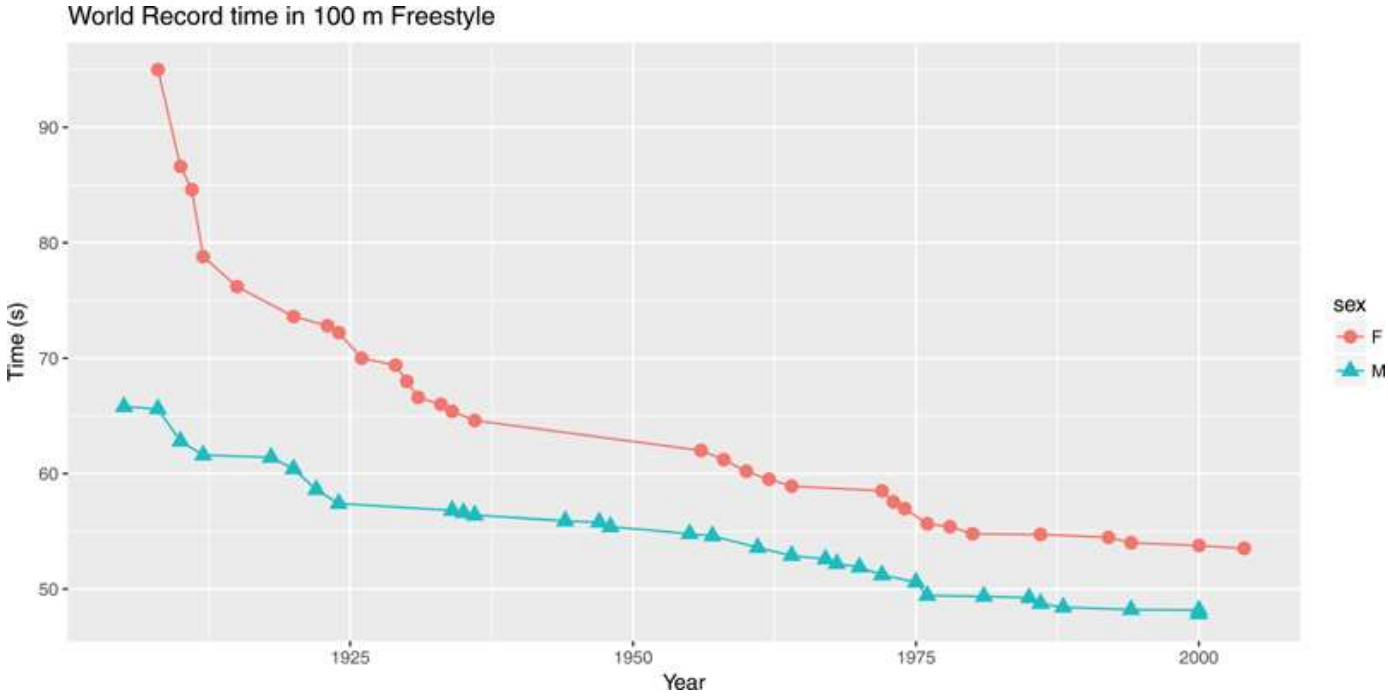
Exercise 2.

The bar graph displays the average score on the math portion of the 1994–1995 SAT (with possible scores ranging from 200 to 800) among states for whom at least two-thirds of the students took the SAT.



Exercise 3.

A time series shows the progression of the world record times in the 100-meter freestyle swimming event for men and women. The time series plot displays the times as a function of the year in which the new record was set.



Exercise 4.

A choropleth map showing the population of Massachusetts by the 2010 Census tracts