

# ***Data Visualization***

What, Why, and How?

Henry Novianus Palit

[hnpalit@petra.ac.id](mailto:hnpalit@petra.ac.id)

# Agenda

- ◆ What is data visualization?
- ◆ Why do we visualize data?
- ◆ How do we visualize data?
  - ⊕ Elements of visualization
  - ⊕ Directory of visualization

# Introduction

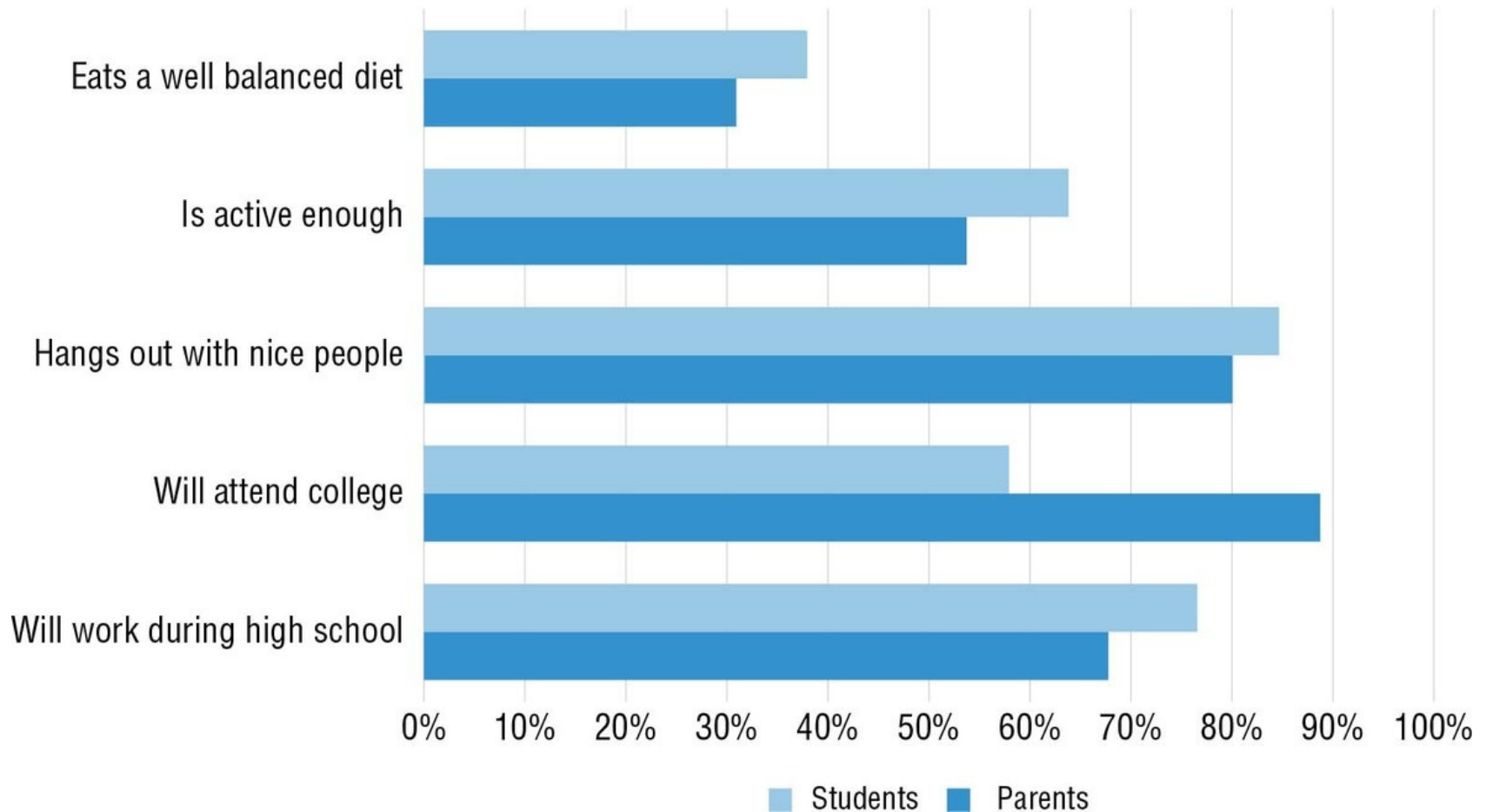
- ◆ Data visualization is part art and part science
  - ⊕ Challenge: get the art right without getting the science wrong, and vice versa
- ◆ Requirements of data visualization:
  - ⊕ Have to accurately convey the data → it must not mislead or distort
  - ⊕ Should be aesthetically pleasing → good visual presentations enhance the message of the visualization
- ◆ Scientists frequently know how to visualize data without being misleading, but may not have a good sense of visual aesthetics and may make visual choices that detract from their desired message

# Why We Visualize (1)

- ◆ The most important question to ask when creating a data visualization: **“What’s your point?”**
- ◆ The primary reason: because we have a point to communicate to the world
  - ⊕ A compelling finding to share
  - ⊕ A big idea revealed in our analysis that we need to say to people
- ◆ Case: **Parent vs. Student Perspectives on Life Plans**

# Why We Visualize (2)

## Parent v. Student Perspectives



# Why We Visualize (2)

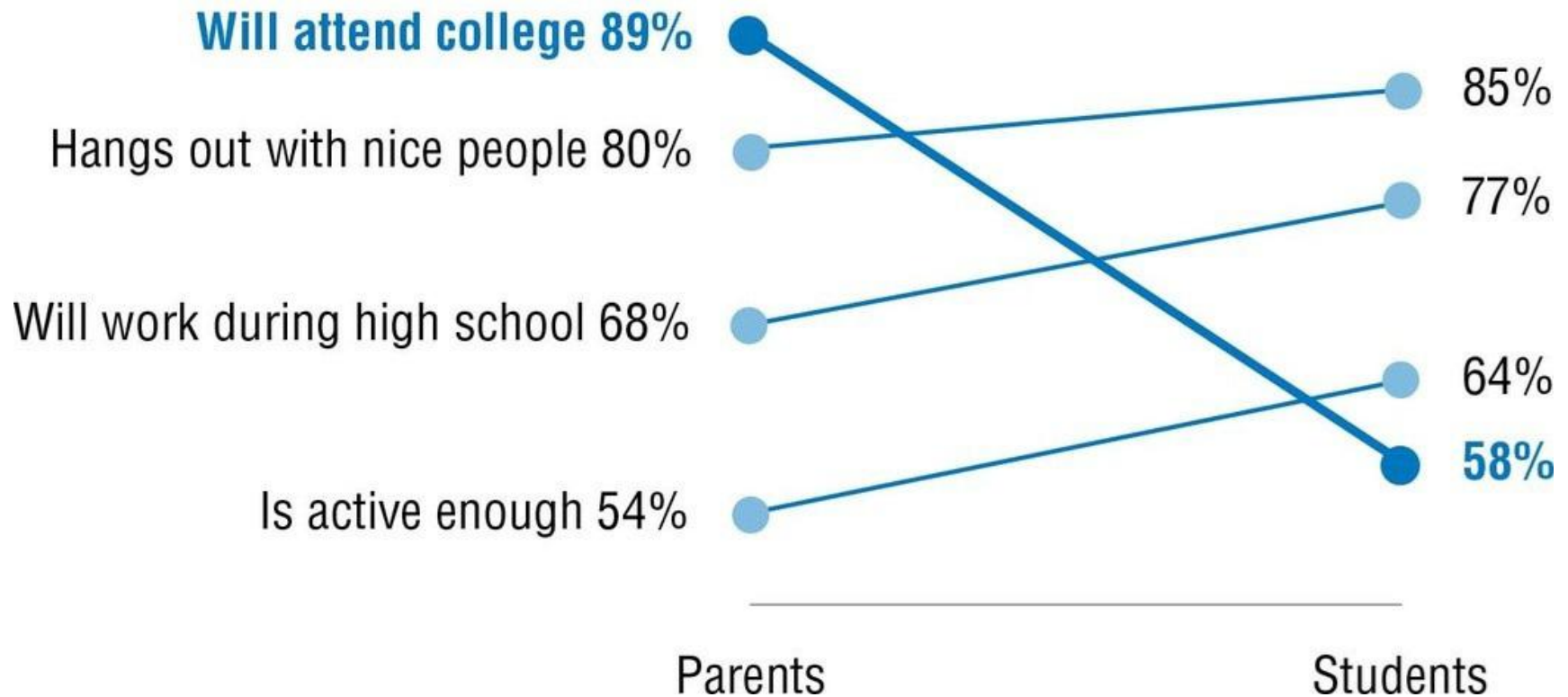
Consultant: “What’s the point of showing this data? Do you want people to compare parents and students?”

Client : “Actually, no. Our point is that generally **we expect students to report higher than parents on all of these questions, but our data showed that the students’ expectations to go to college were way lower than their parents’**. That set off some alarm bells for us.”

Consultant: “Let’s replace this generic title with your main point. Then, we will swap out a different graph type, maybe something like a **slopegraph** since those are pretty **good at highlighting when one thing is decreasing and the rest are going up.**”

# Why We Visualize (3)

Surprisingly, students have lower expectations to go to college than their parents.



# Why We Visualize (3)

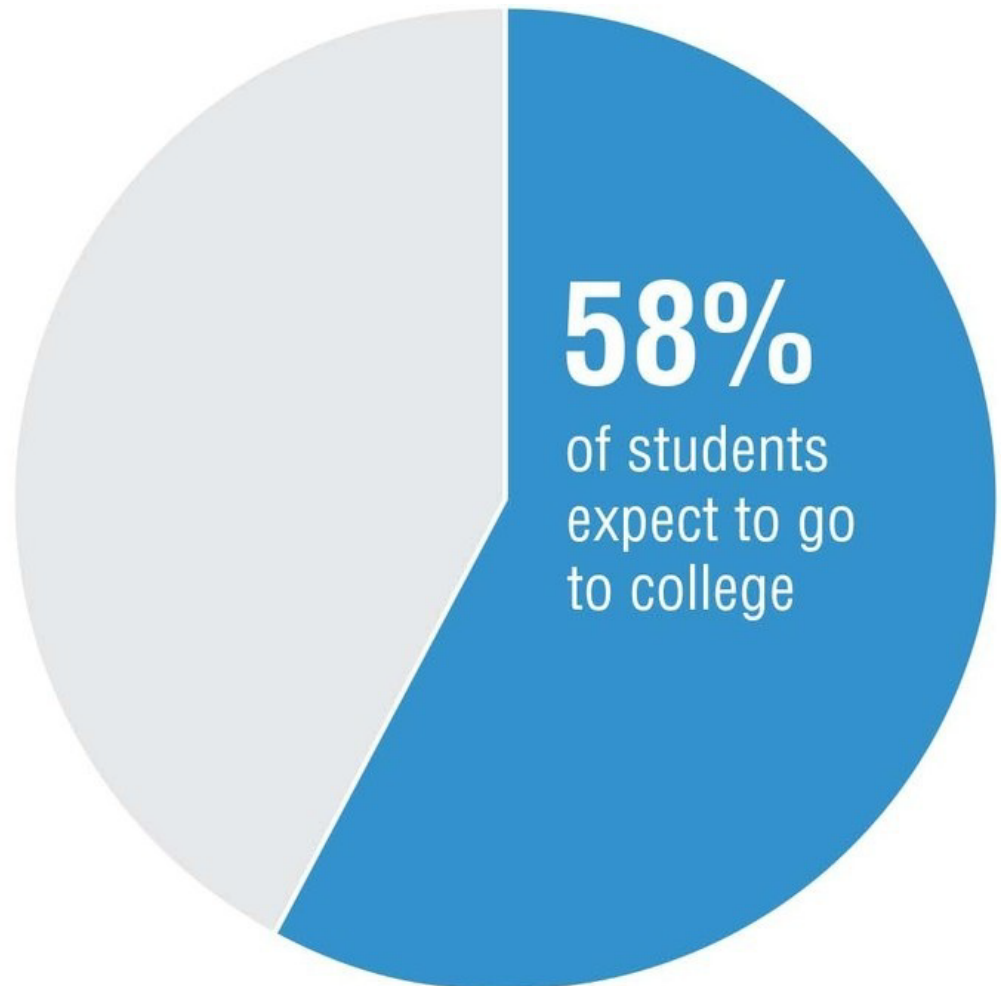
Consultant: “What did you think of that slopegraph?”

Client : “It really does say exactly what we originally thought we needed to show. But, I talked to my colleagues after that and asked them ‘What’s the point?’ We decided that **the real bottom-line point was that so few students have expectations to go to college**. Forget the parents—that’s a secondary issue right now.”

Consultant: “Ah well in that case, you have other options for showing that point. A **single large number** or a **simple pie chart** are two possible ways to help readers **remember one important number**.”

## Why We Visualize (4)

Only  
**58%**  
of students  
expect to go  
to college



# Why We Visualize (5)

- ◆ Figuring out your point sharpens the thinking and the messaging surrounding the data, and in doing so reveals the best way to visualize the data
  - ⊕ If you find you don't have a point, you probably shouldn't bother with graphing the data
  - ⊕ We visualize to communicate a point
- ◆ Visualizing is also to add legitimacy (support) and credibility
  - ⊕ People are persuaded by numbers and stories
  - ⊕ When we can tell stories with numbers, we have a communication powerhouse

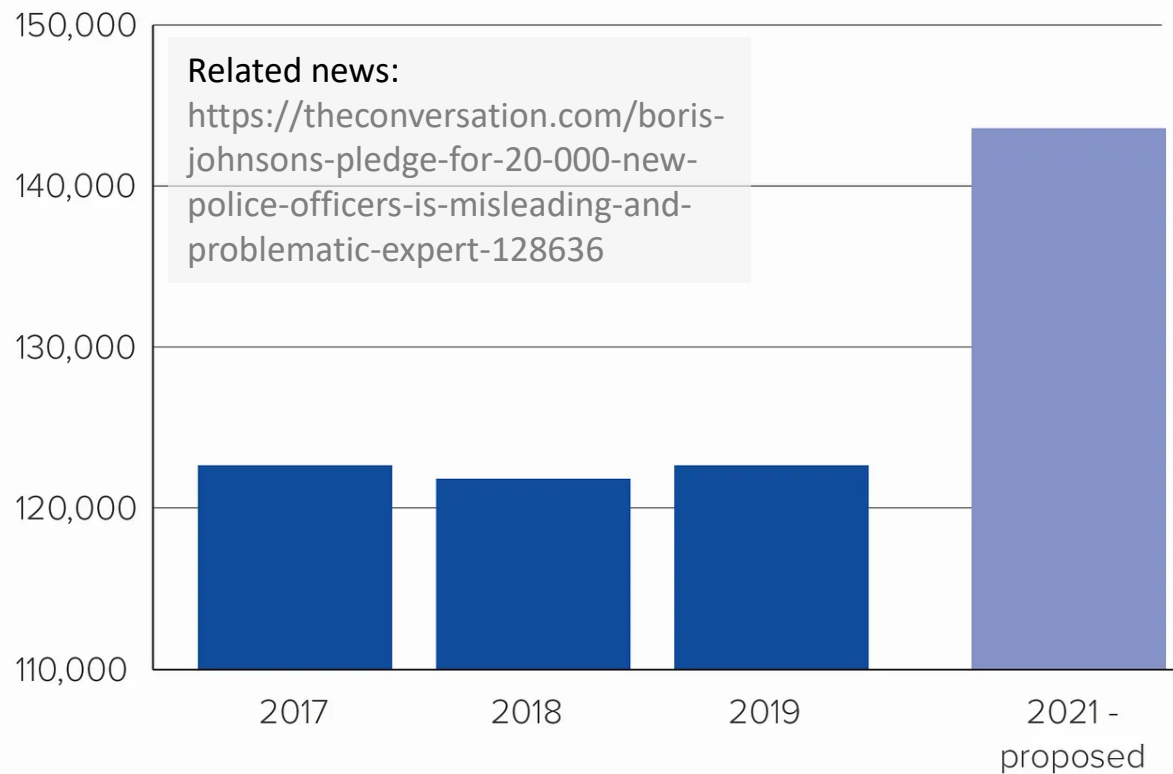
# Why We Visualize (6)

- ◆ Research tells us that data are more persuasive when shown in graphs
- ◆ We are primarily visual beings and most of us, most of the time, are skimming the narrative for things that pop out at us and catch our attention
  - ◆ Data visualization provides the pop
- ◆ Graphs and formulas seem to add credibility to data, even if they don't contain any new insights beyond what already exists in the narrative
- ◆ But, the same tools can be used to *deceive*!

## Blue wave

Conservatives will put 20,000 extra police officers on the street by 2021

Total number of police officers in England and Wales

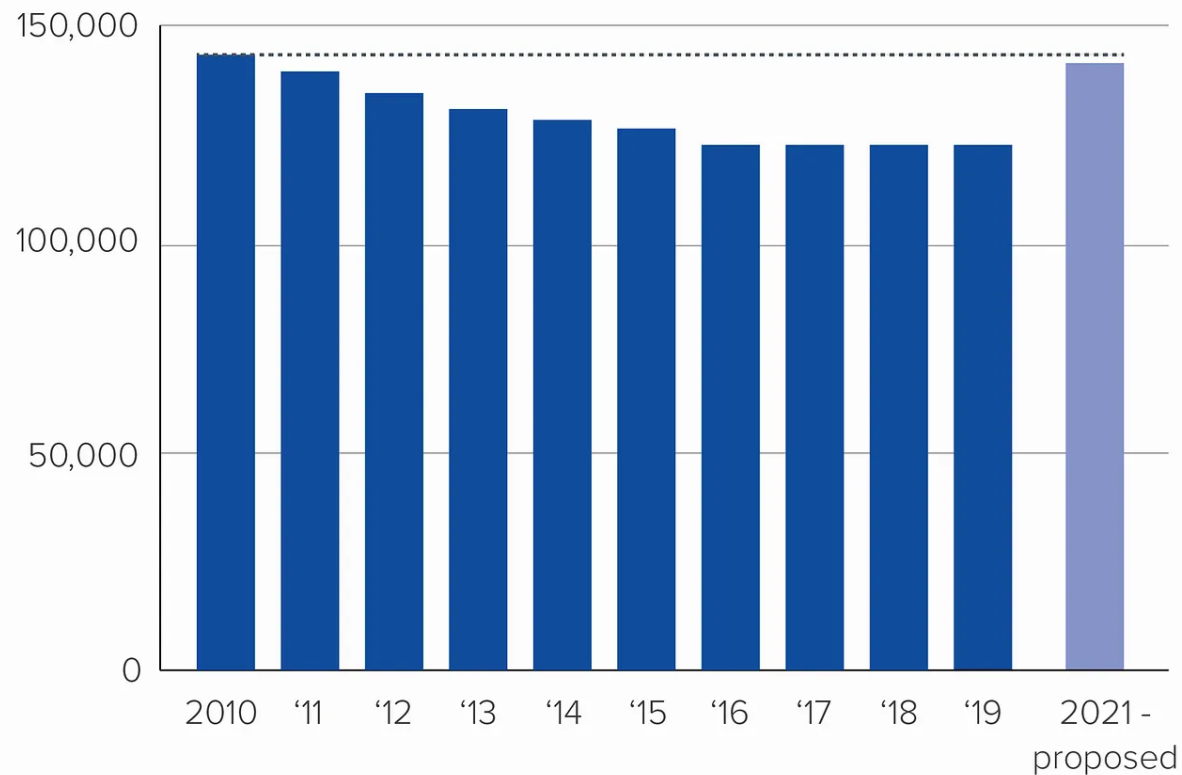


Source: UK Home Office, 2019

## Blue murder

Conservatives' proposed increase of 20,000 police officers will not make up for numbers cut during austerity (2010-19)

Total number of police officers in England and Wales



Source: UK Home Office, 2019

# When Visualization is Harmful *(1)*

- ◆ At best, data visualization errors are unintentional mistakes that lead to misinformation
- ◆ At worst, they are purposeful manipulations designed to influence the story a graph can tell
- ◆ Elements like the scale of the axis or the size and shape of the graph can distort data and produce interpretation errors
- ◆ Changing aspects of the graph can lead to deception, whether intentional or benign

# When Visualization is Harmful (2)

- ◆ There are justifiable reasons for truncating the y-axis on a line graph; alteration to support decision making can be warranted
- ◆ In general, distortion is real, common, and harmful

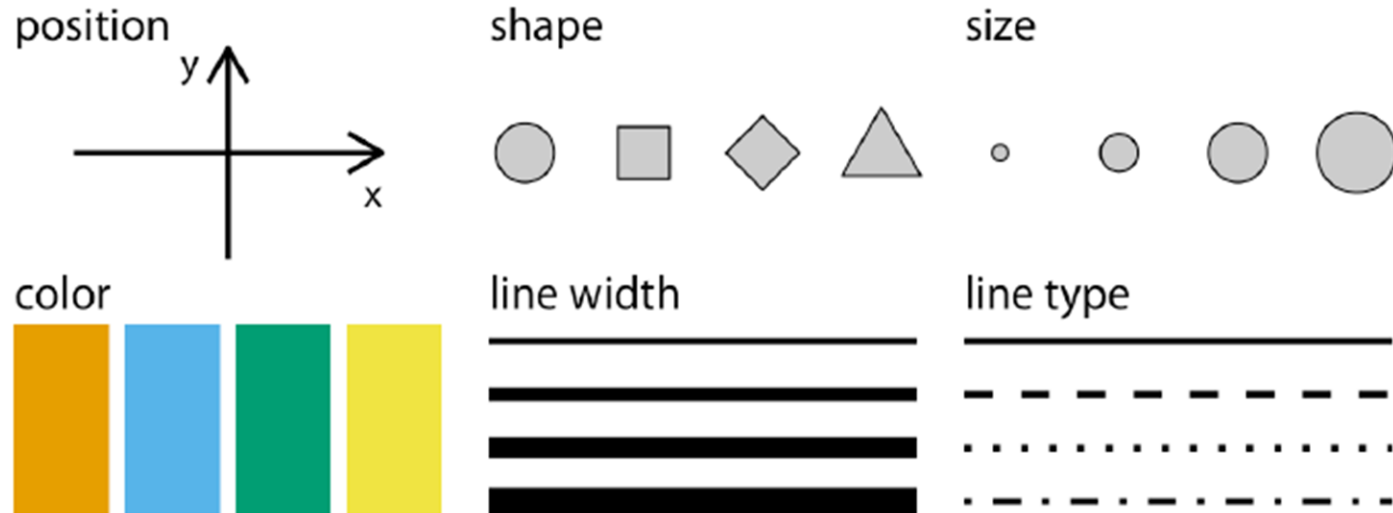
# ELEMENTS OF VISUALIZATION

# Visualizing Data

- ◆ When 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 (e.g., a scatterplot, a pie chart, a heatmap, etc.), all these visualizations can be described with a common language that captures how data values are turned into blobs of ink on paper or colored pixels on a screen
- ◆ All data visualizations map data values into quantifiable features of the resulting graphic; we refer to these features as ***aesthetics***

# Aesthetics (1)

- ◆ Aesthetics describe every aspect of a given graphical element



- ◆ Other aesthetics:
  - ⊕ Text – font family, font face, and font size
  - ⊕ Overlapped graphical objects – transparency

# Aesthetics (2)

- ◆ All aesthetics fall into one of two groups:
  - ⊕ Continuous data values are values for which arbitrarily fine intermediates exist
    - Time duration: between any two durations (e.g., 50 seconds and 51 seconds) there are arbitrarily many intermediates (e.g., 50.5 seconds, 50.51 seconds, 50.50001 seconds, and so on)
    - Position, size, color, and line width can represent continuous data
  - ⊕ Discrete data values
    - Persons: a room can hold 5 persons or 6, but not 5.5
    - Shape and line type can usually only represent discrete data

# Types of Data (1)

- ◆ In addition to continuous and discrete numerical values, 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 (e.g., dog, cat, fish)
  - ⊕ Factors can also be ordered when there is an intrinsic order among the levels of the factor (e.g., good, fair, poor)

# Types of Data (2)

Type of variable	Examples	Appropriate scale	Description
Quantitative/ numerical continuous	1.3, 5.7, 83, $1.5 \times 10^{-2}$	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 <i>factors</i> .
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 <i>ordered factors</i> .
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.

# Types of Data (3)

◆ Example: daily temperature normals (average over a 30-year window) for four weather stations

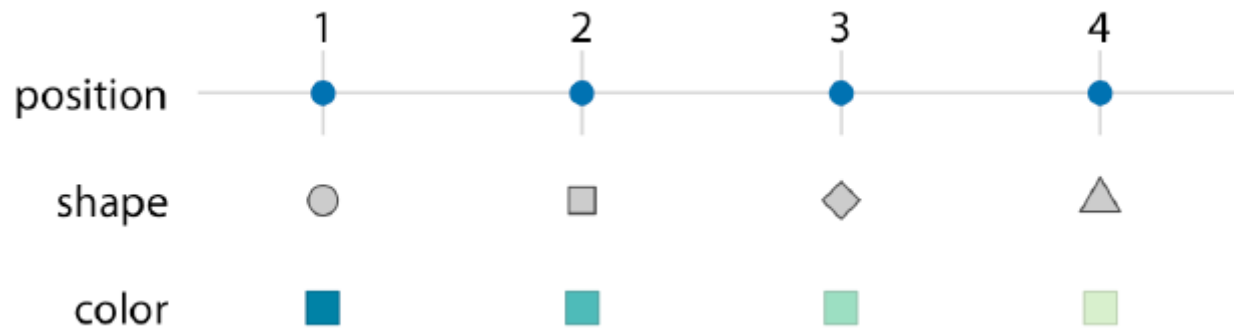
Month	Day	Location	Station ID	Temperature (°F)
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319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

- ⊕ Month: ordered factor
- ⊕ Day: discrete numerical value
- ⊕ Location: unordered factor
- ⊕ Station ID: unordered factor
- ⊕ Temperature: continuous numerical value

# Scales (1)

- ◆ 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
  - ⊕ Similarly, we may need to specify which data values are represented by particular shapes or colors
- ◆ This mapping between data values and aesthetics values is created via **scales**
- ◆ A scale must be one-to-one, i.e., for each specific data value there is exactly one aesthetics value and vice versa; or else, the data viz becomes ambiguous

# Scales (2)



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, shape, or color, and vice versa

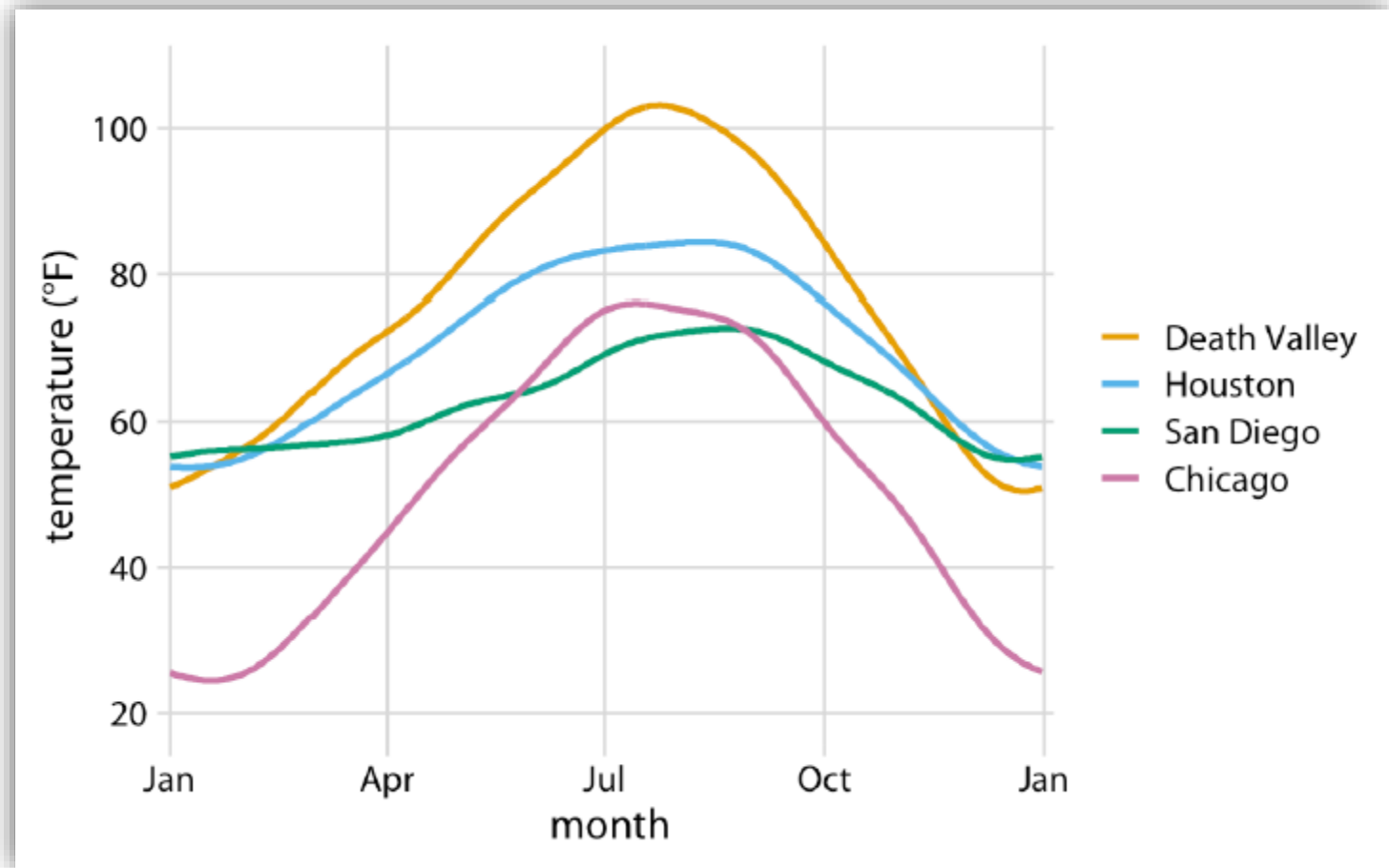
# Scales (3)

◆ Let's put things into practice

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

We map temperature onto the y axis, day of the year onto the x axis, location onto color, and visualize these aesthetics with solid lines

# Scales (3)



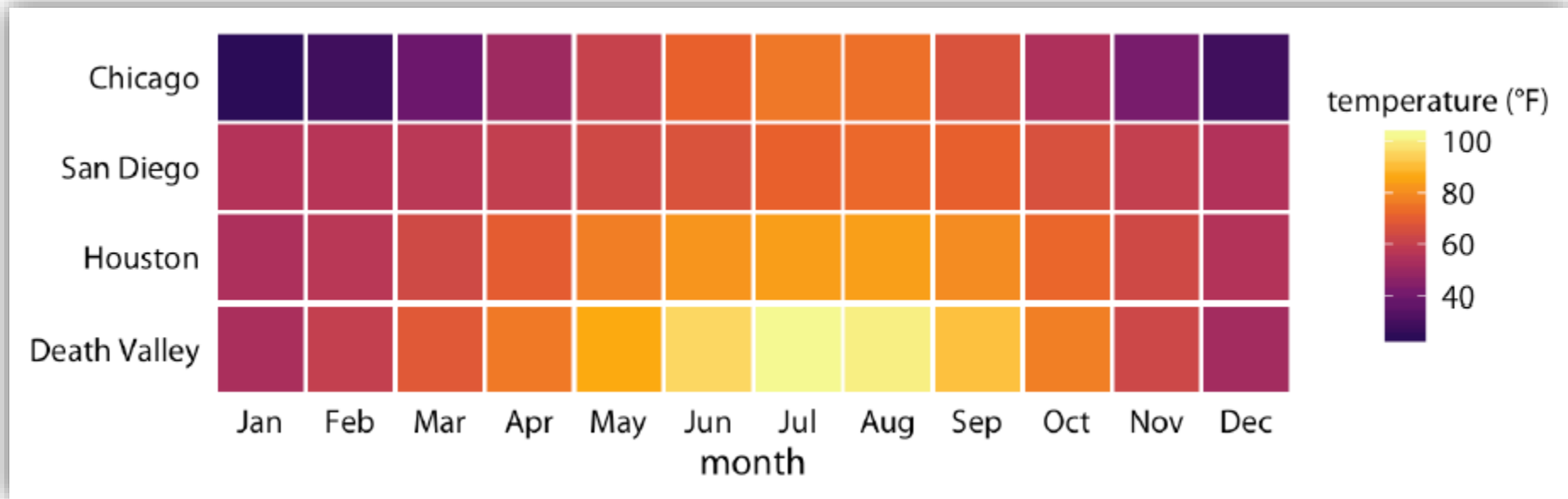
# Scales (4)

◆ Let's put things into practice

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

We map location onto the y axis, day of the year onto the x axis, temperature onto color, and visualize these aesthetics with squares (one for each month and location, colored by the average temperature normal for each month)

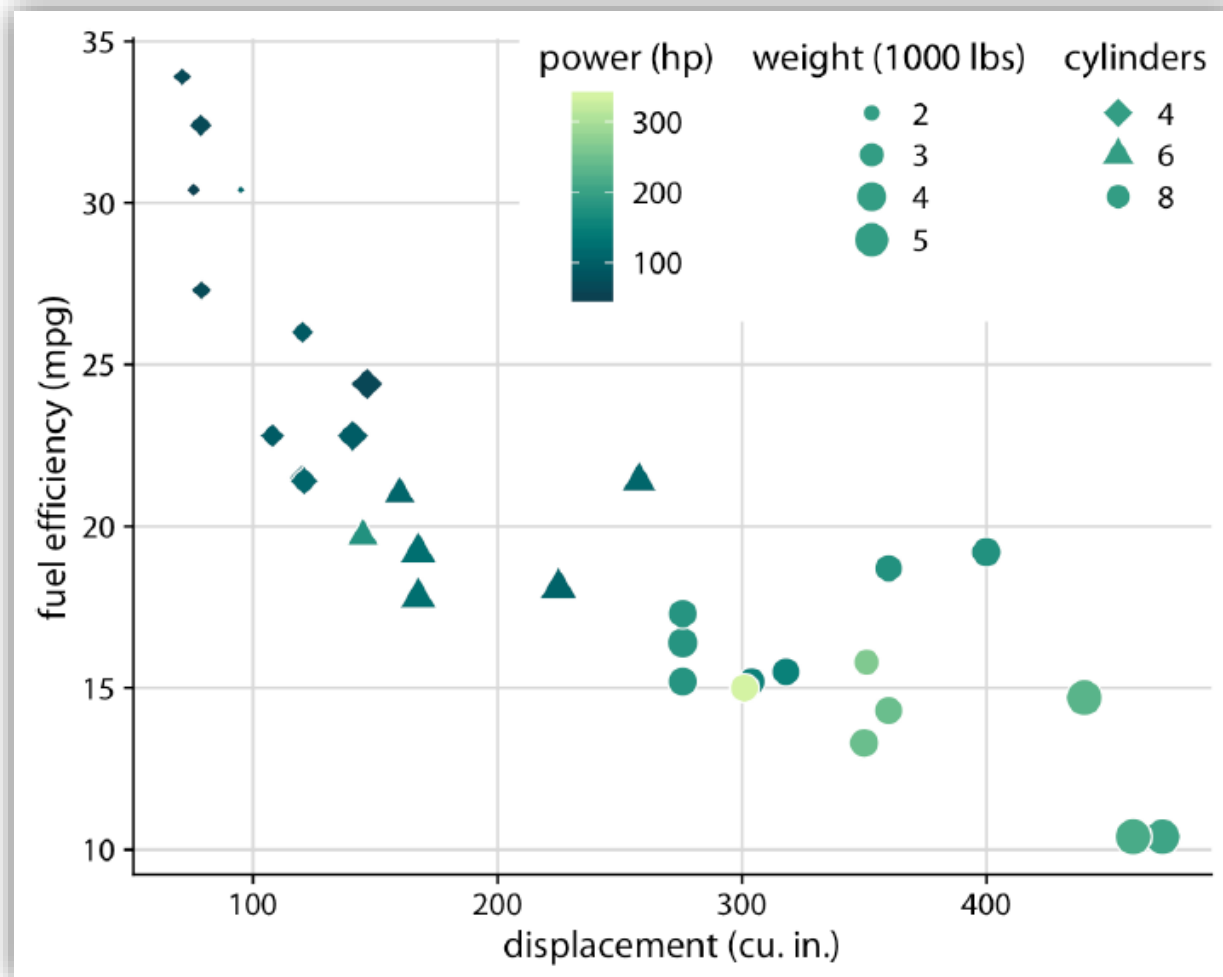
# Scales (4)



# Scales (5)

- ◆ In the last figure, the two position scales are both discrete (month and location)
  - ⊕ For discrete position scales, we place the different levels of the factor at an equal spacing along the axis
  - ⊕ If the factor is ordered (as is the case for month), then the levels need to be placed in the appropriate order
  - ⊕ If the factor is unordered (as is the case for location), then the order is arbitrary
- ☞ Note: In the figure, the locations were ordered from overall coldest (Chicago) to overall hottest (Death Valley) to generate a pleasant staggering of colors

# Scales (6)



Fuel efficiency versus displacement, for 32 cars (1973–74 models).

This figure uses five separate scales to represent data:

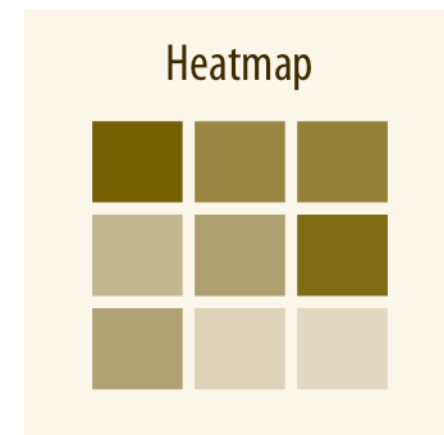
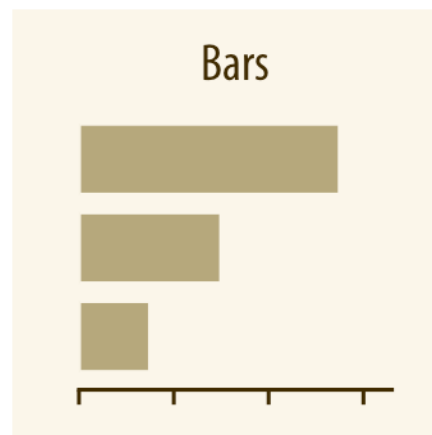
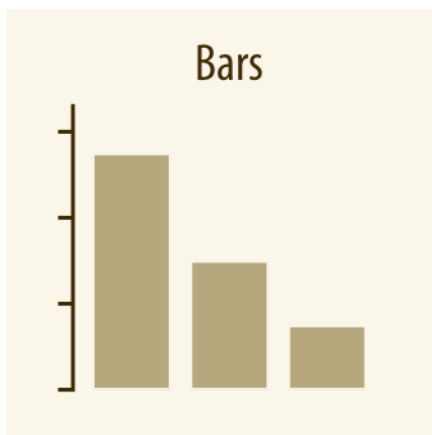
- (i) the x axis (displacement);
- (ii) the y axis (fuel efficiency);
- (iii) the color of the data points (power);
- (iv) the size of the data points (weight);
- (v) the shape of the data points (number of cylinders).

Four variables displayed are numerical continuous. The number of cylinders can be considered to be either numerical discrete or qualitative ordered.

# DIRECTORY OF VISUALIZATION

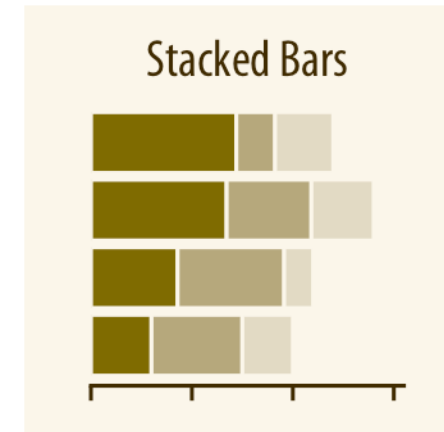
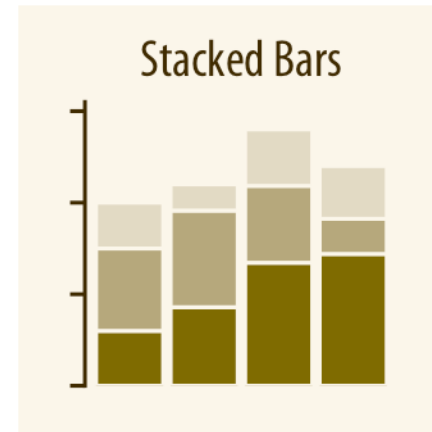
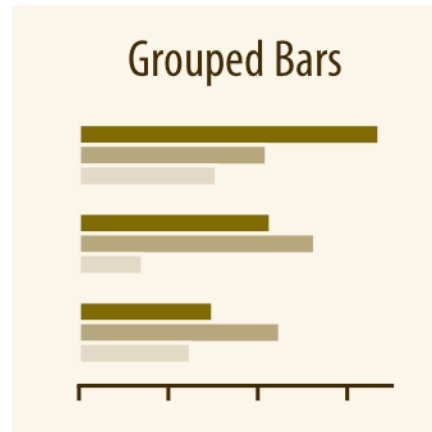
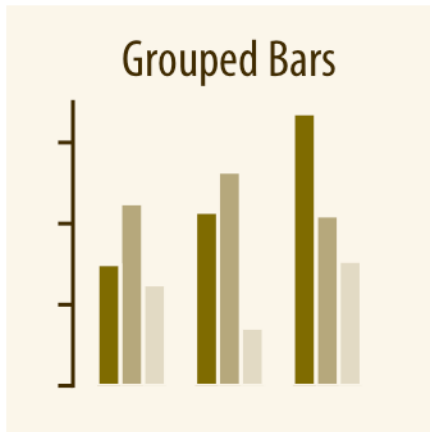
# Amounts (1)

- ◆ 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
- ◆ Instead of using bars, we can also place **dots** at the location where the corresponding bar would end



# Amounts (2)

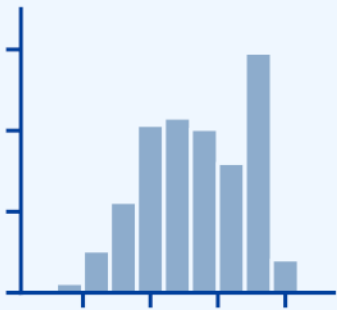
- ◆ If there are two or more sets of categories, we can map the categories onto the x and y axes and show amounts by color, via a **heatmap**
- ◆ Alternatively, we can also **group or stack the bars**



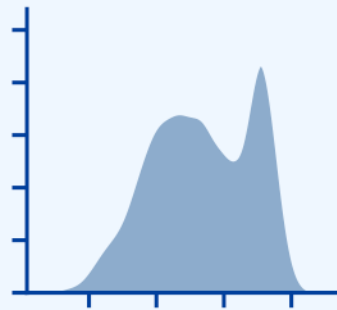
# Distributions (1)

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

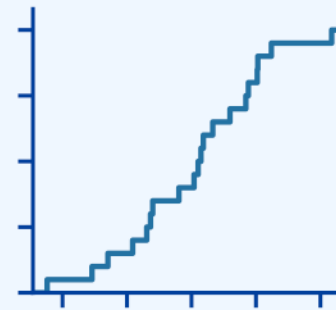
Histogram



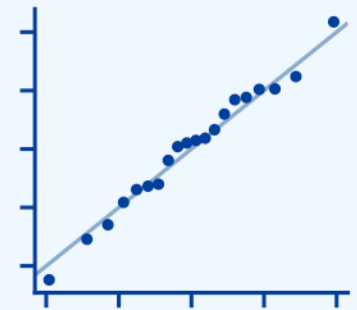
Density Plot



Cumulative Density



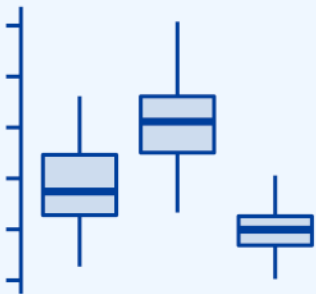
Quantile-Quantile Plot



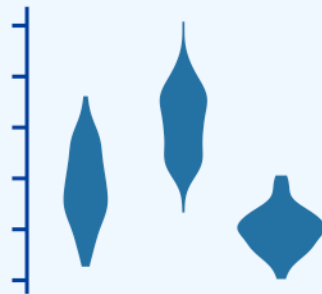
# Distributions (2)

- ◆ **Boxplots, violin plots, strip charts, and sina plots** are useful when we want to visualize many distributions at once and/or if we primarily interested in overall shifts among the distributions

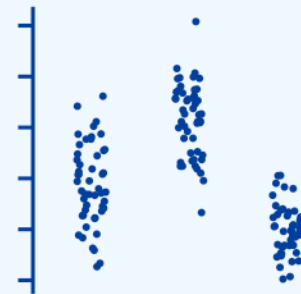
Boxplots



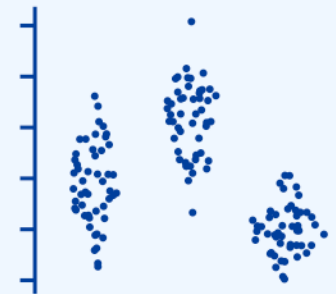
Violins



Strip Charts

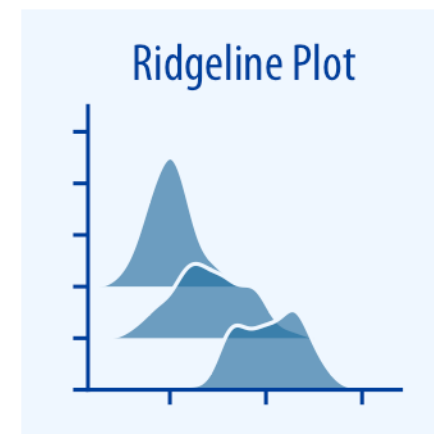
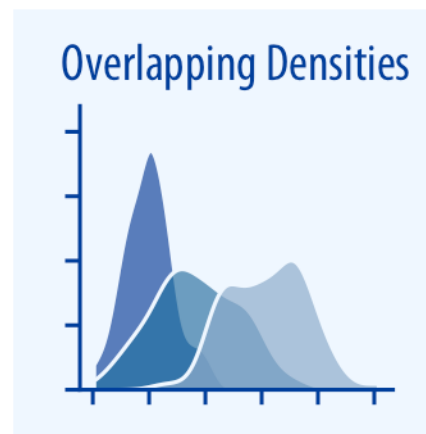
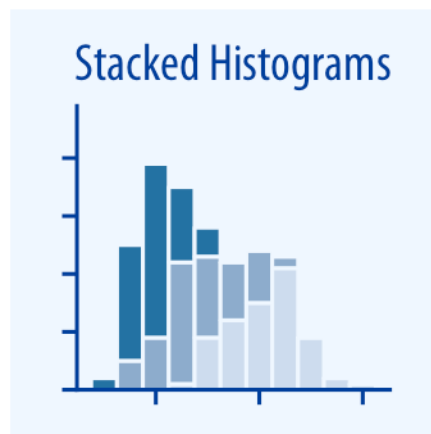


Sina Plots



# Distributions (3)

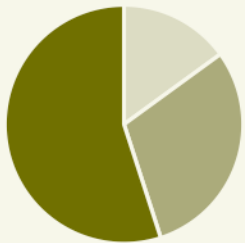
- ❖ **Stacked histograms** and **overlapping densities** allow a more in-depth comparison of a smaller no. of distributions, though stacked histograms can be difficult to interpret and are best avoided
- ❖ **Ridgeline plots** can be a useful alternative to violin plots and are useful when visualizing very large no. of distributions or changes in distributions over time



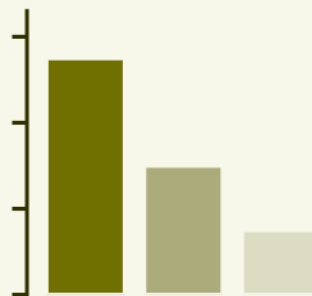
# Proportions (1)

- ◆ Proportions can be visualized as **pie charts, side-by-side bars,**  
**or stacked bars**
  - ⊕ Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions
  - ⊕ 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

Pie Chart



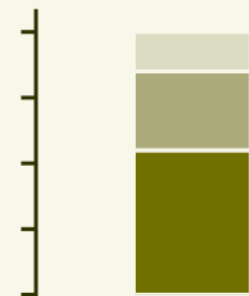
Bars



Bars



Stacked Bars



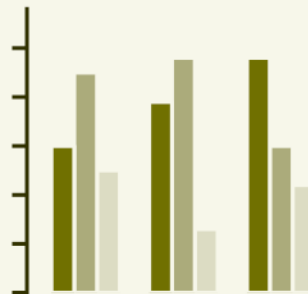
# Proportions (2)

- ◆ When visualizing multiple sets of proportions or changes in proportions across conditions, **multiple pie charts** tend to be space-inefficient and often obscure relationships
- ◆ **Grouped bars** work well as long as the no. of conditions compared is moderate, and **stacked bars** can work for large no. of conditions
- ◆ **Stacked densities** are appropriate when proportions change along a continuous variable

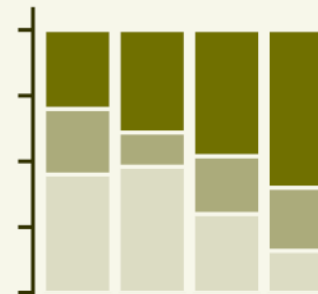
Multiple Pie Charts



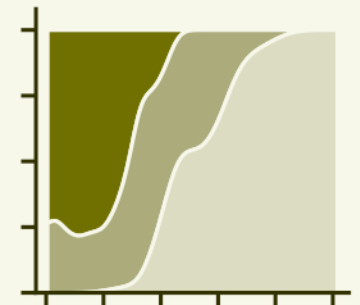
Grouped Bars



Stacked Bars

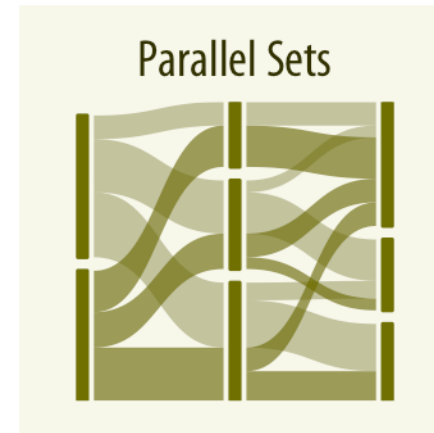
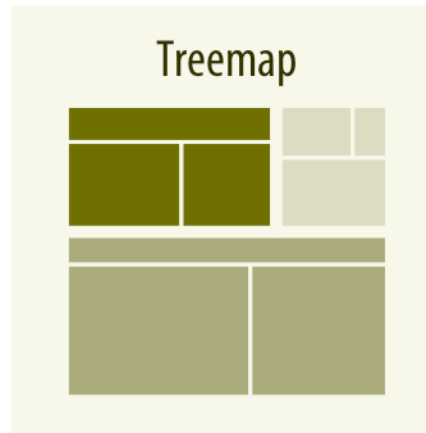
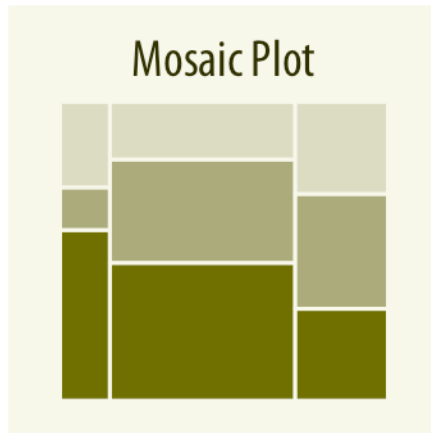


Stacked Densities



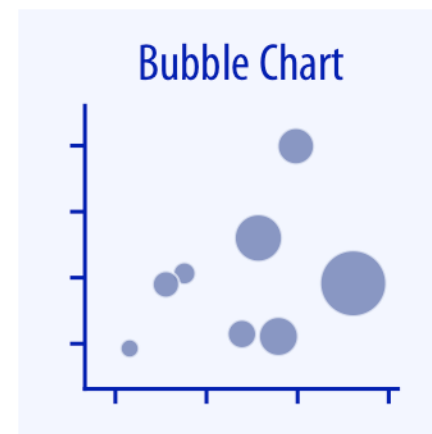
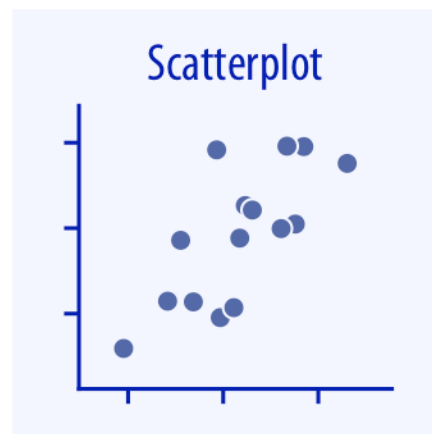
# Proportions (3)

- ◆ When proportions are specified according to multiple grouping variables, **mosaic plots**, **treemaps**, or **parallel sets** are useful visualization approaches
  - ⊕ Mosaic plots assume that every level of one grouping variable can be combined with every level of another grouping variable
  - ⊕ Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another
  - ⊕ Parallel sets work better when there are > 2 grouping variables



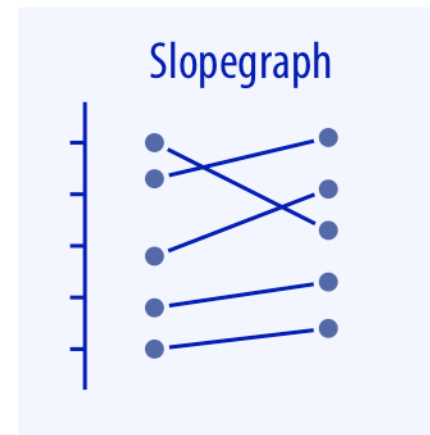
# x-y Relationships (1)

- ◆ **Scatterplots** represent the archetypical visualization when we want to show one quantitative variable relative to another
- ◆ If we have three quantitative variables, we can map one on the dot size, creating a variant of the scatterplots called a **bubble chart**



## x–y Relationships (2)

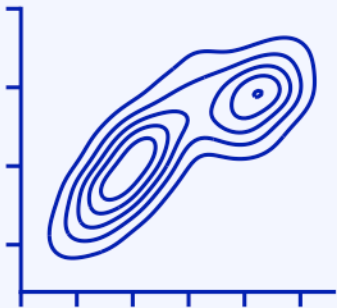
- ◆ For **paired scatterplots**, where the variables along x and y axes are measured in the same units, it is generally helpful to add a line indication  $x = y$
- ◆ Paired data can also be shown as a **slopegraph** of paired points connected by straight lines



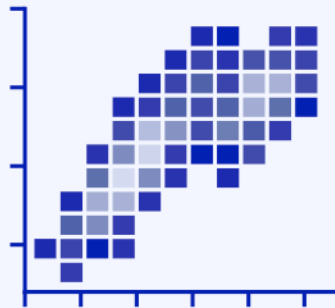
## x-y Relationships (3)

- ◆ For large no. of points, regular scatterplots can become uninformative due to overplotting
- ◆ In this case, contour lines, 2D bins, or hex bins may provide an alternative
- ◆ To visualize > 2 quantities, we may choose to plot correlation coefficients in the form of a correlogram

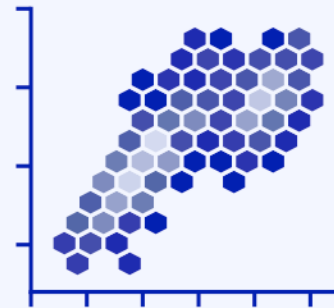
Density Contours



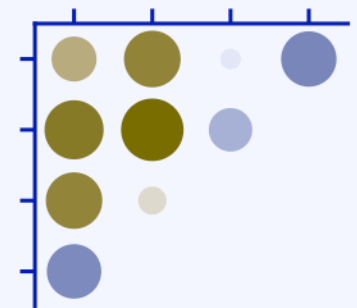
2D Bins



Hex Bins

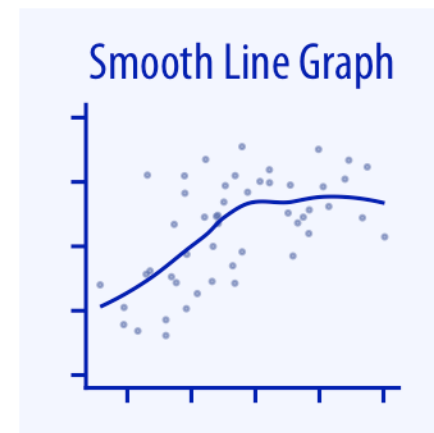
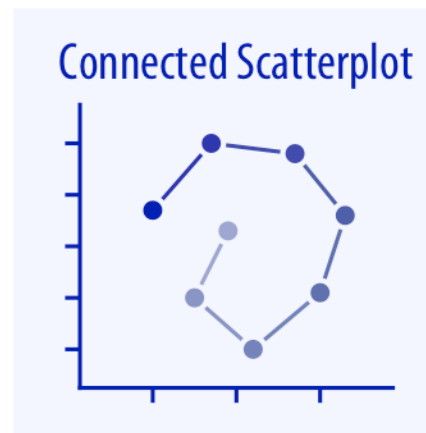
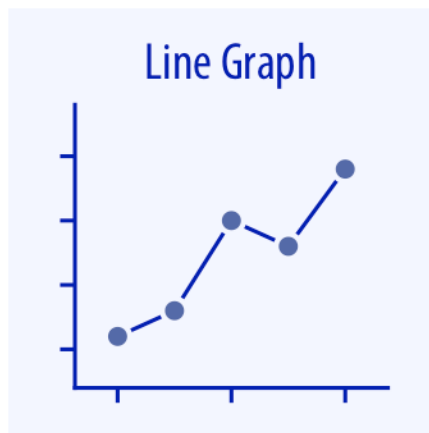


Correlogram



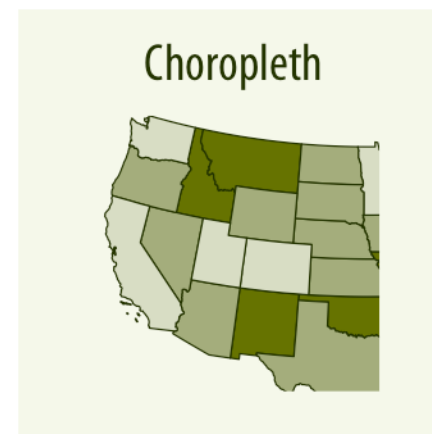
# x-y Relationships (4)

- ◆ When the x axis represents time or a strictly increasing quantity, we commonly draw line graphs
- ◆ If we have a temporal sequence of two response variables, we can draw a connected scatterplot, where we first plot the two response variables in a scatterplot and then connect dots corresponding to adjacent time points
- ◆ Smooth lines can represent trends in a large dataset



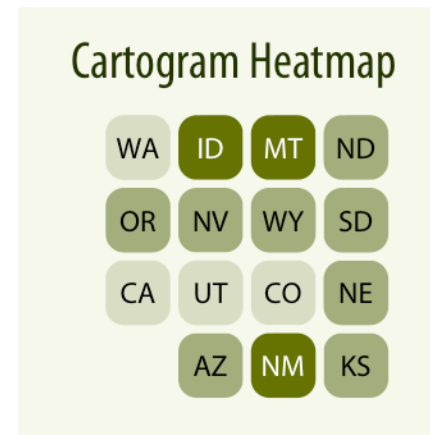
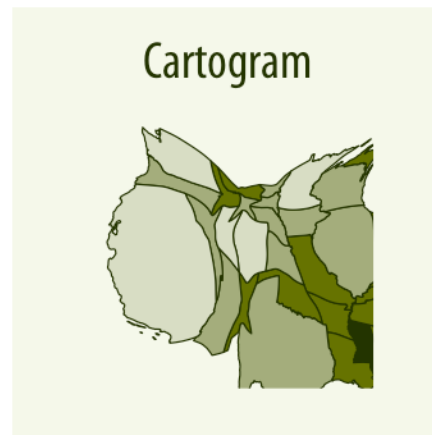
# Geospatial Data (1)

- ◆ The primary mode of showing geospatial data is in the form of a **map**, taking coordinates on the globe and projecting them onto a flat surface
- ◆ We can show data values in different regions by coloring those regions in the map according to the data; such a map is called a **choropleth**



# Geospatial Data (2)

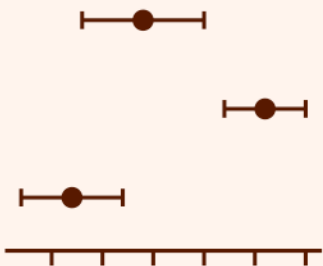
- ◆ In some cases, it may be useful to distort the different regions according to some other quantity (e.g., population number) or simplify each region into a square; such visualizations are called **cartograms**



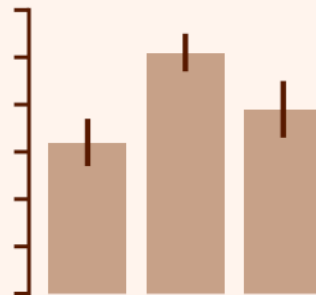
# Uncertainty (1)

- ◆ **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 (shown by dots or bars) representing the estimate or measurement
- ◆ **Graded error bars** show multiple ranges at the same time, 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

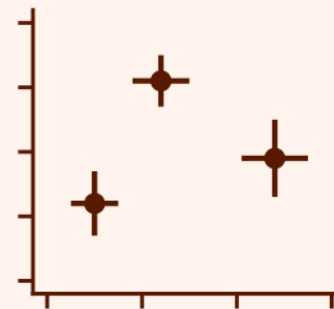
Error Bars



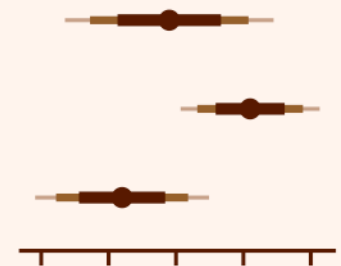
Error Bars



2D Error Bars



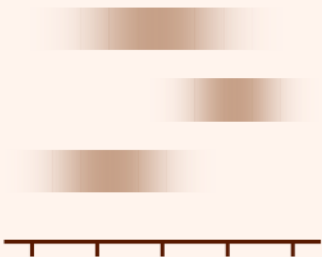
Graded Error Bars



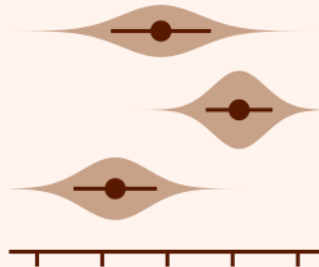
# Uncertainty (2)

- ◆ To achieve a more detailed visualization, we can visualize the actual confidence or posterior distributions
  - ⊕ **Confidence strips** provide a 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)
  - ⊕ A **quantile dot plot** can serve as an alternative visualization; because the distribution is in discrete units, the quantile dot plot is not as precise but can be easier to read than the continuous distribution

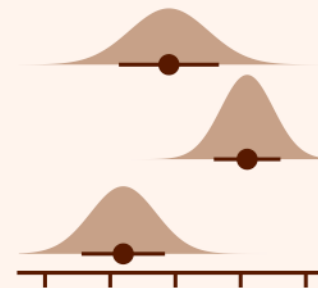
Confidence Strips



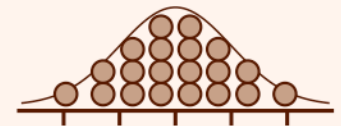
Eyes



Half-Eyes

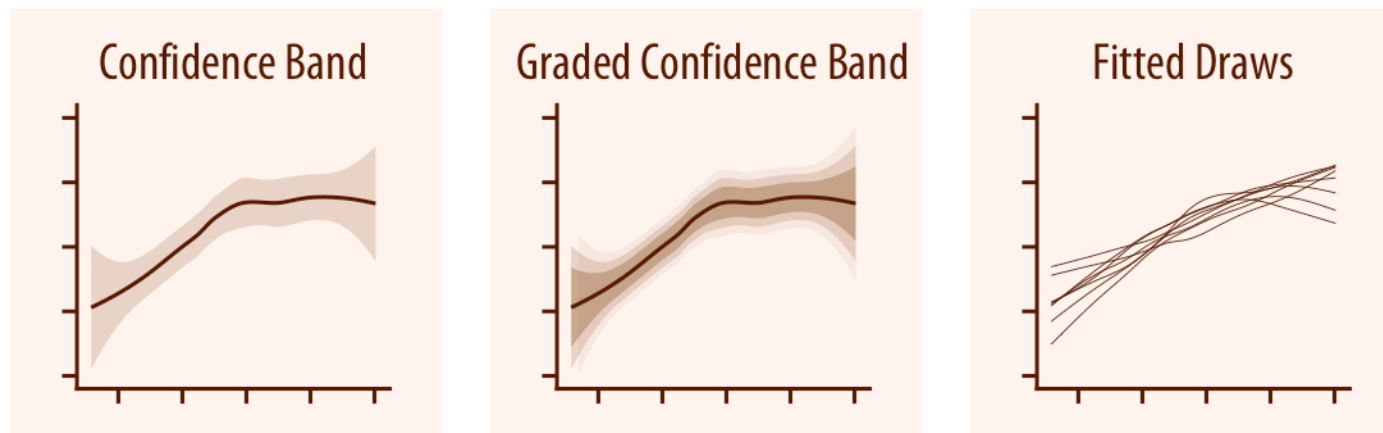


Quantile Dot Plot



# Uncertainty (3)

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



# References

- ◆ Stephanie D.H. Evergreen (2017), “Effective Data Visualization: The Right Chart for the Right Data”, SAGE Publications.
- ◆ Claus O. Wilke (2019), “Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures”, O’Reilly Media.