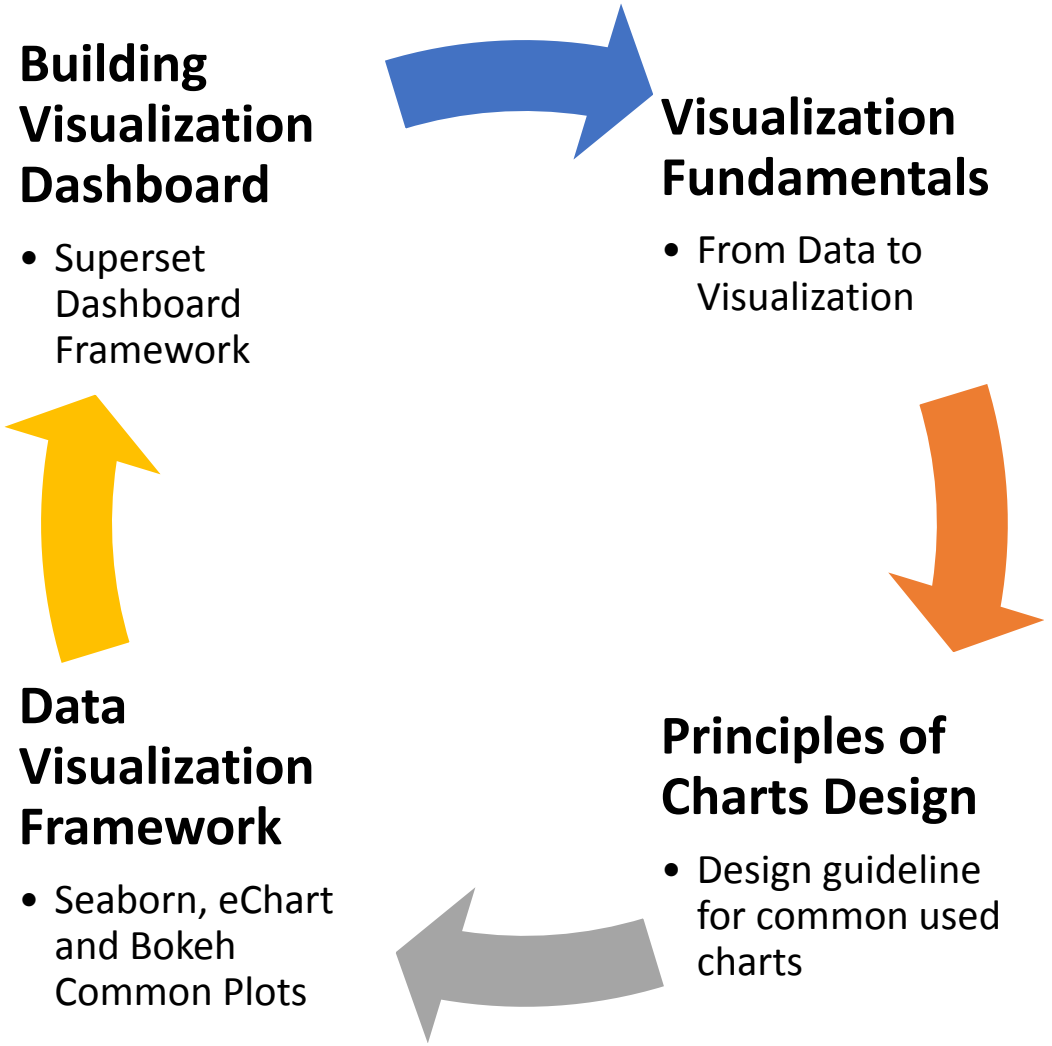
Data Visualization

Risman Adnan, Ph.D

Telkom University

Outline – Data Visualization



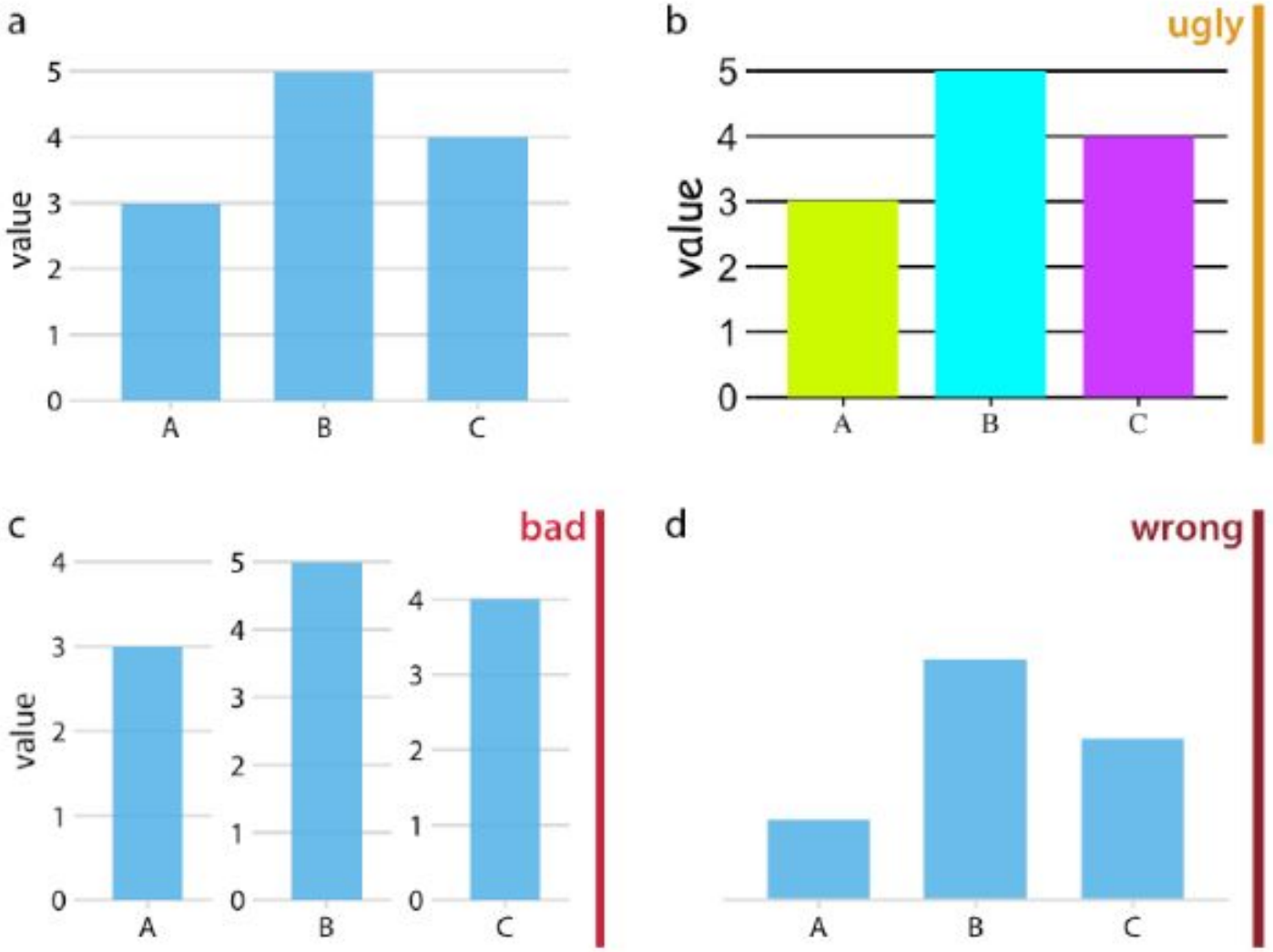
Bad Visualizations are Everywhere

We are Not Good in Visualization!

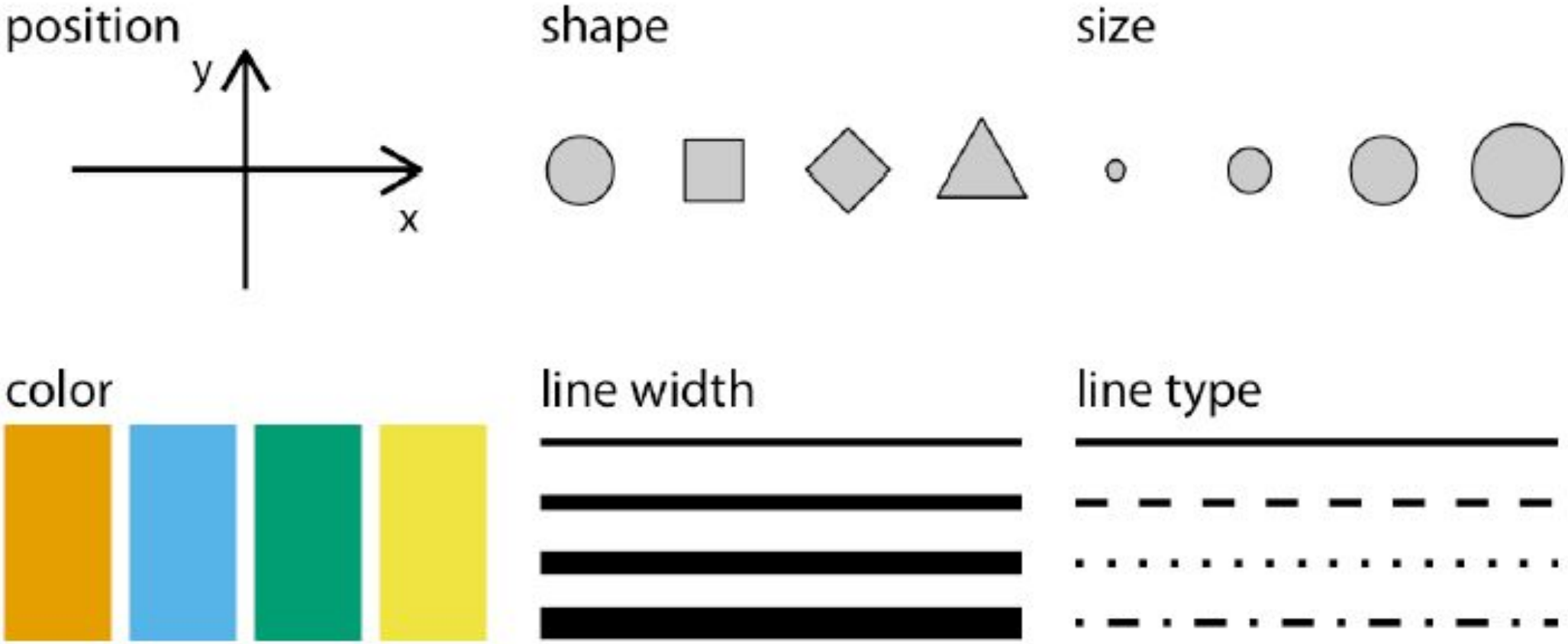
- **Visualization** = science + art
- **Aesthetic** = systematic + logical
- Visualizing = Mapping Data onto Aesthetics!

Key Principles to Avoid:

1. Understand visualization context
2. Choose an appropriate visual display
3. Eliminate clutter
4. Focus attention where you want it
5. Think like a designer
6. Tell a story (the purpose is insights)



Basic Aesthetics



Data Types

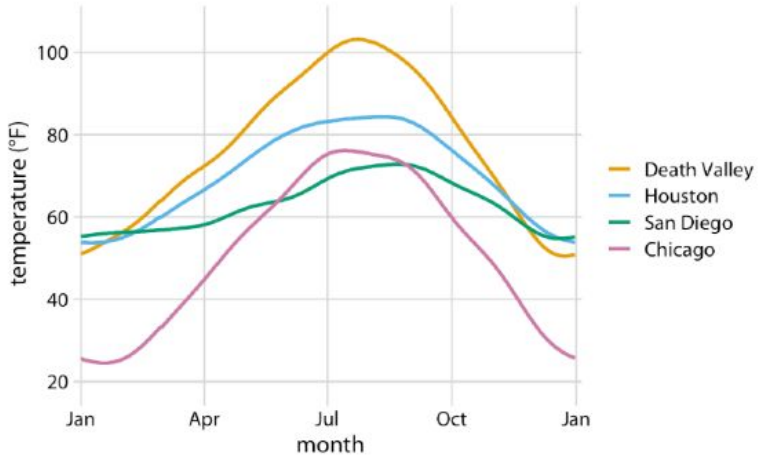
Major Types:

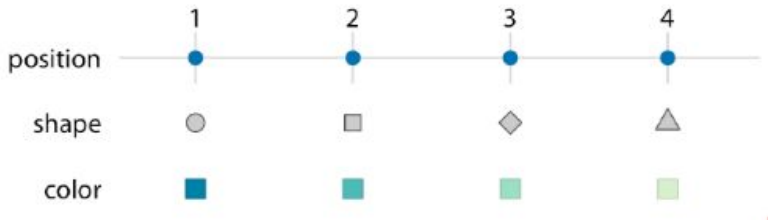
- Numerical
- Categorical
- Date Time
- Text
- Geospatial
- Probability

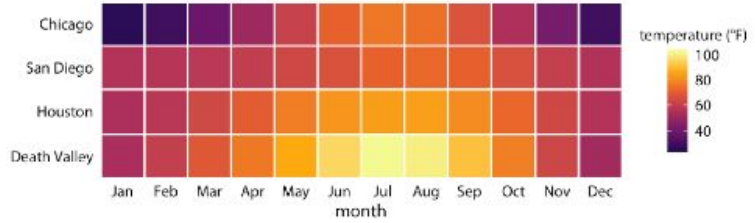
Type of variable	Examples	Appropriate scale	Description
Quantitative/ numerical continuous	1.3, 5.7, 83, 1.5×10^{-2}	Continuous	Arbitrary numerical values. These can be integers, rational numbers, or real numbers.
Quantitative/ numerical discrete	1, 2, 3, 4	Discrete	Numbers in discrete units. These are 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.

Example: US Temperature NOAA Data

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



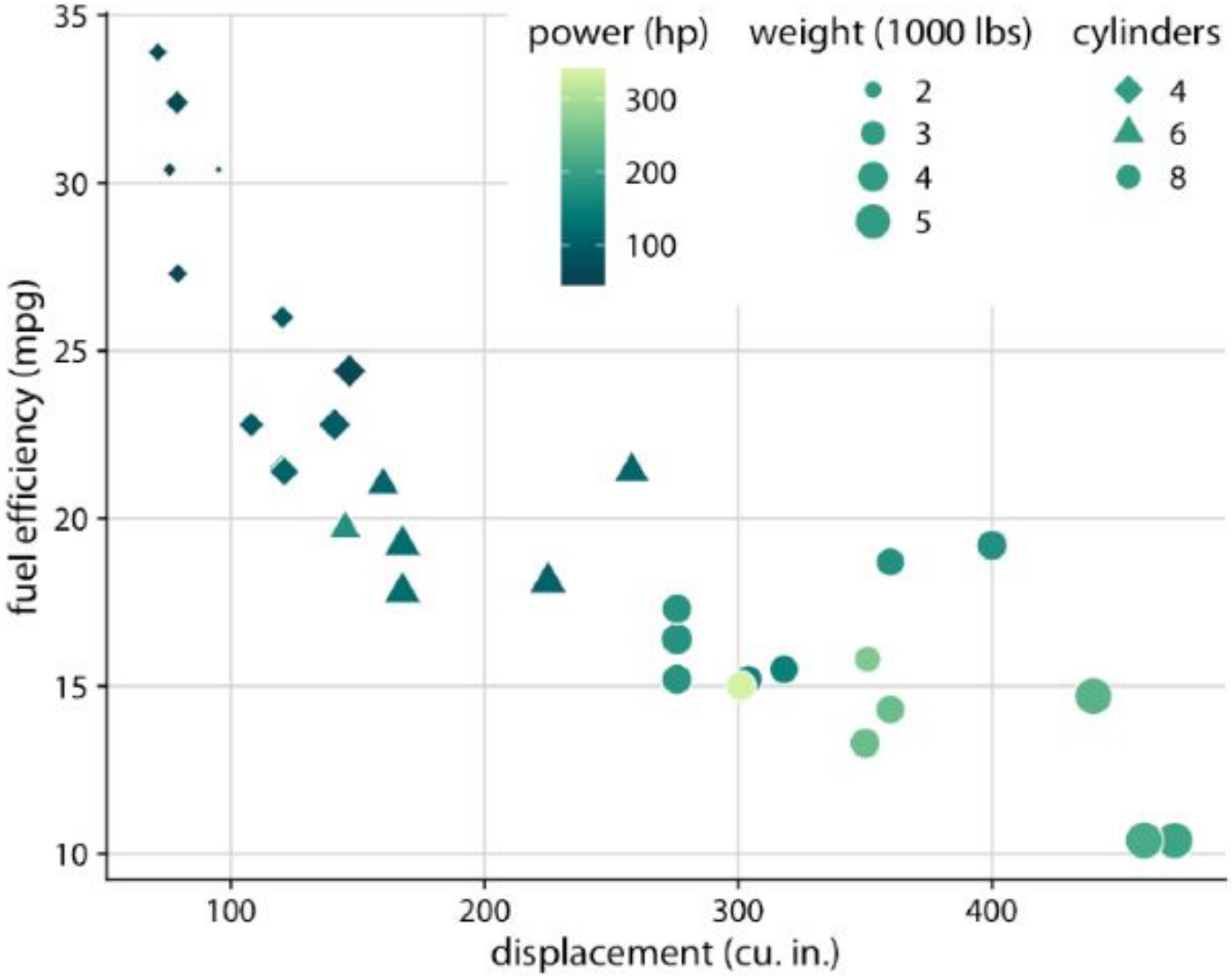




Example: Cars Fuel Efficiency

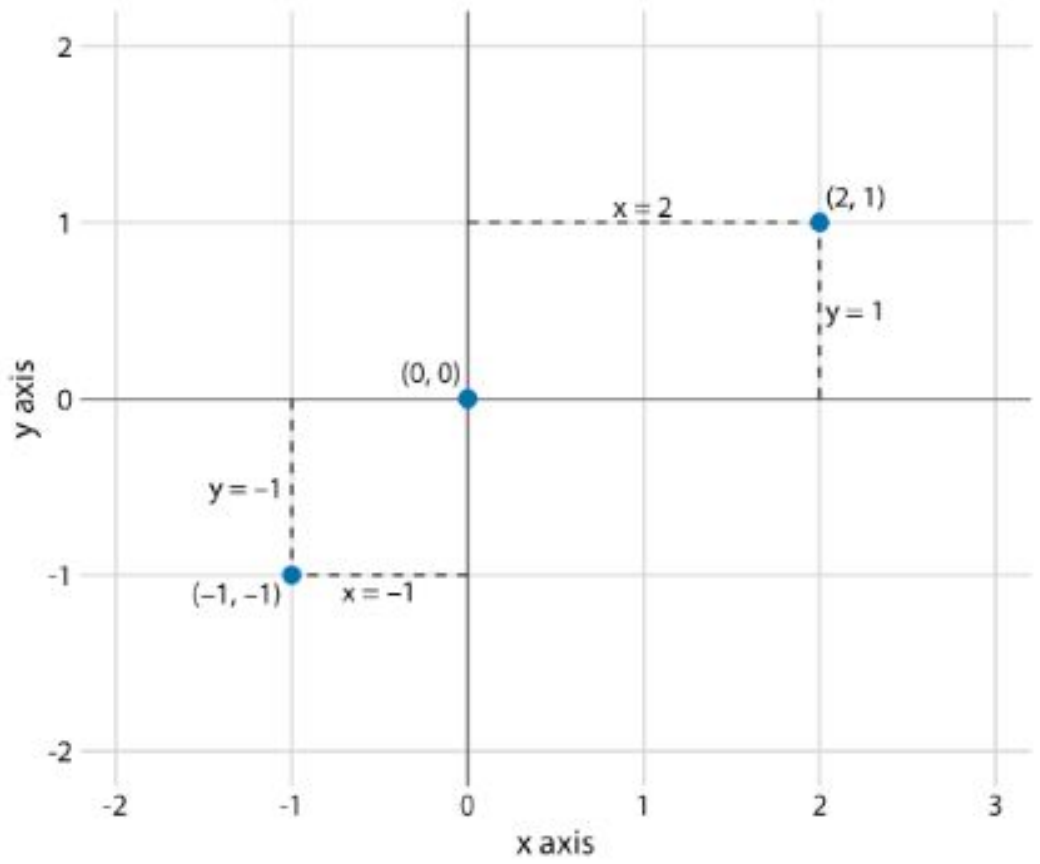
5 scales to describe data:

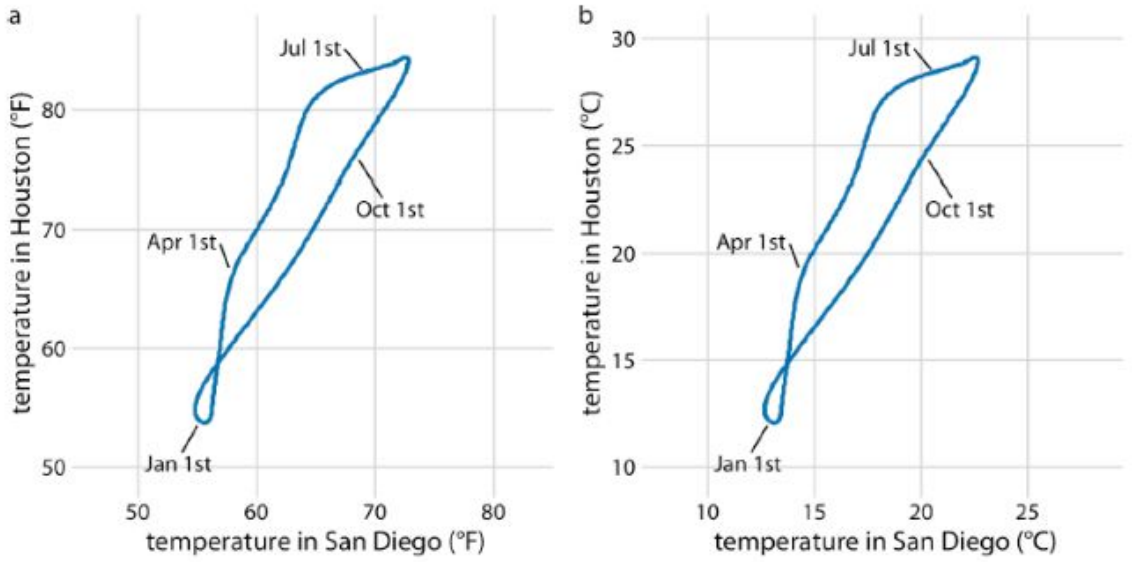
- *x* axis (displacement);
- *y* axis(fuel efficiency);
- Color of data (power)
- Size of data (weight);
- Shape of (# of cylinders).



Coordinate System Axes

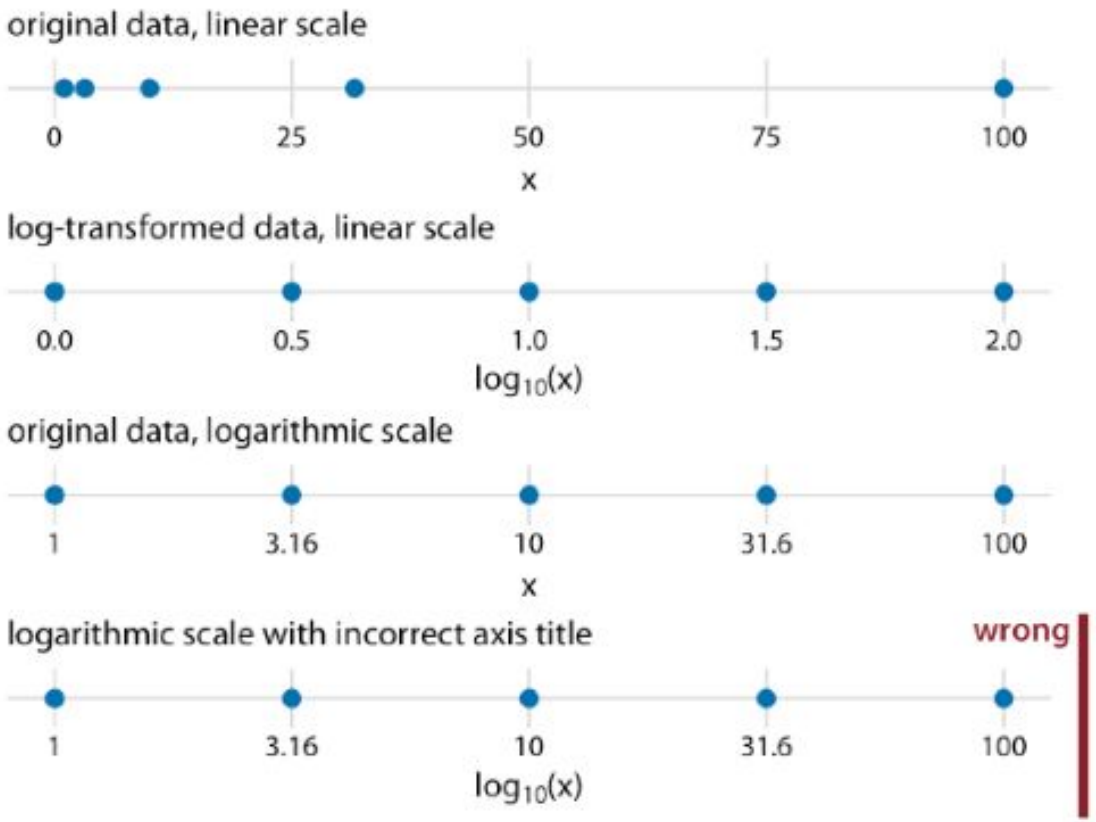
Cartesian Coordinates

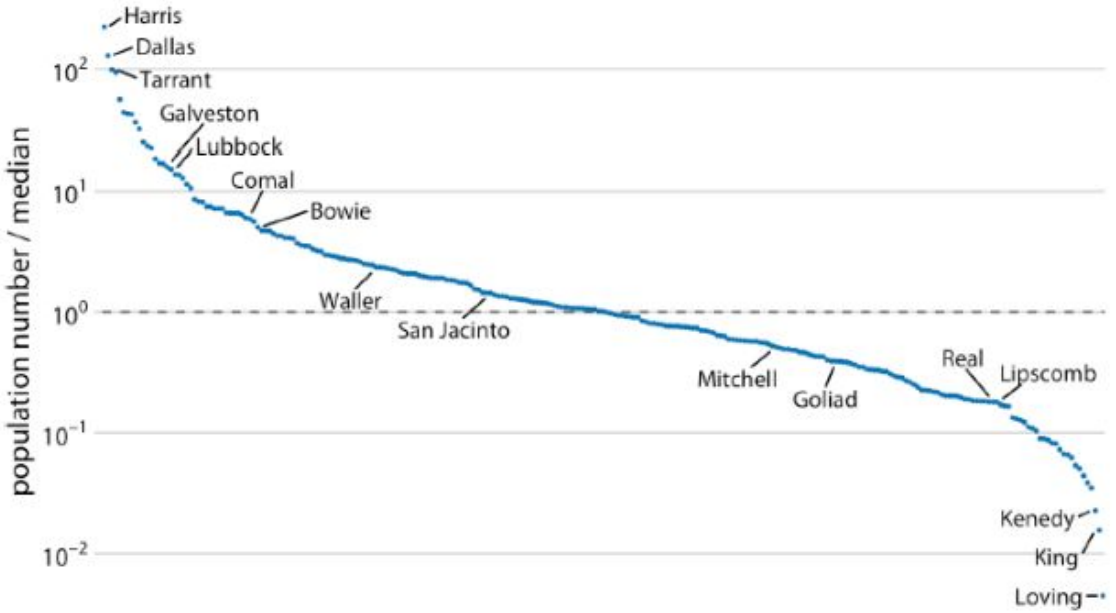




Coordinate System Axes

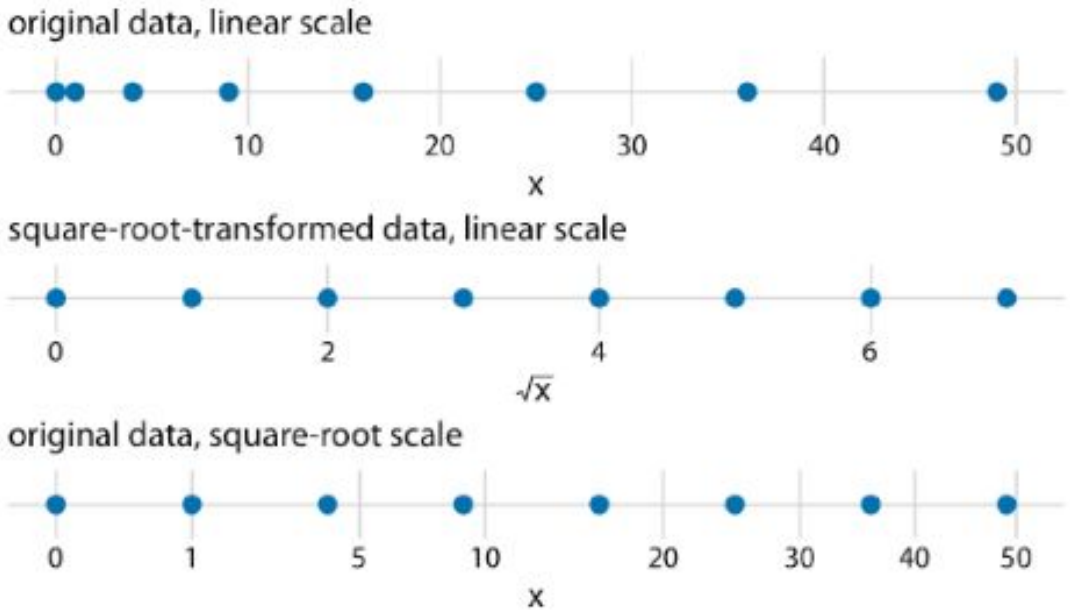
Non-Linear Axes

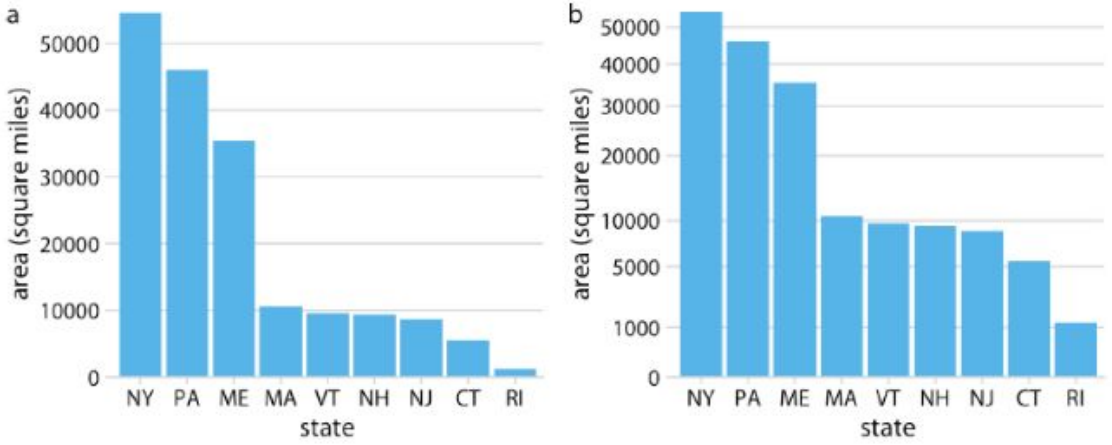




Coordinate System Axes

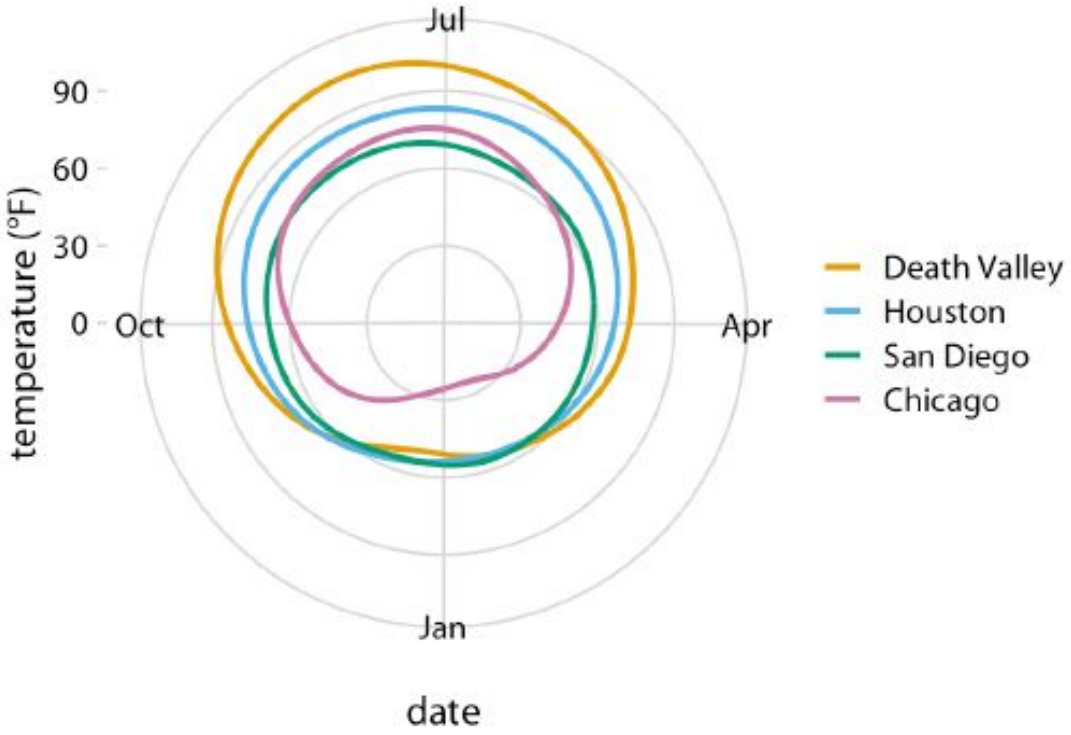
Square Root Axes

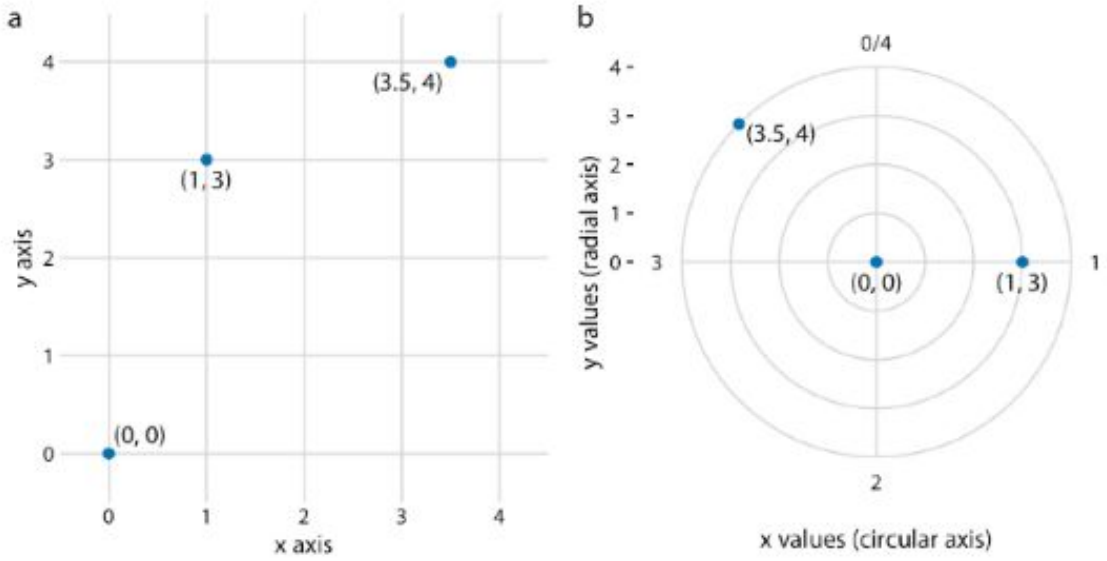




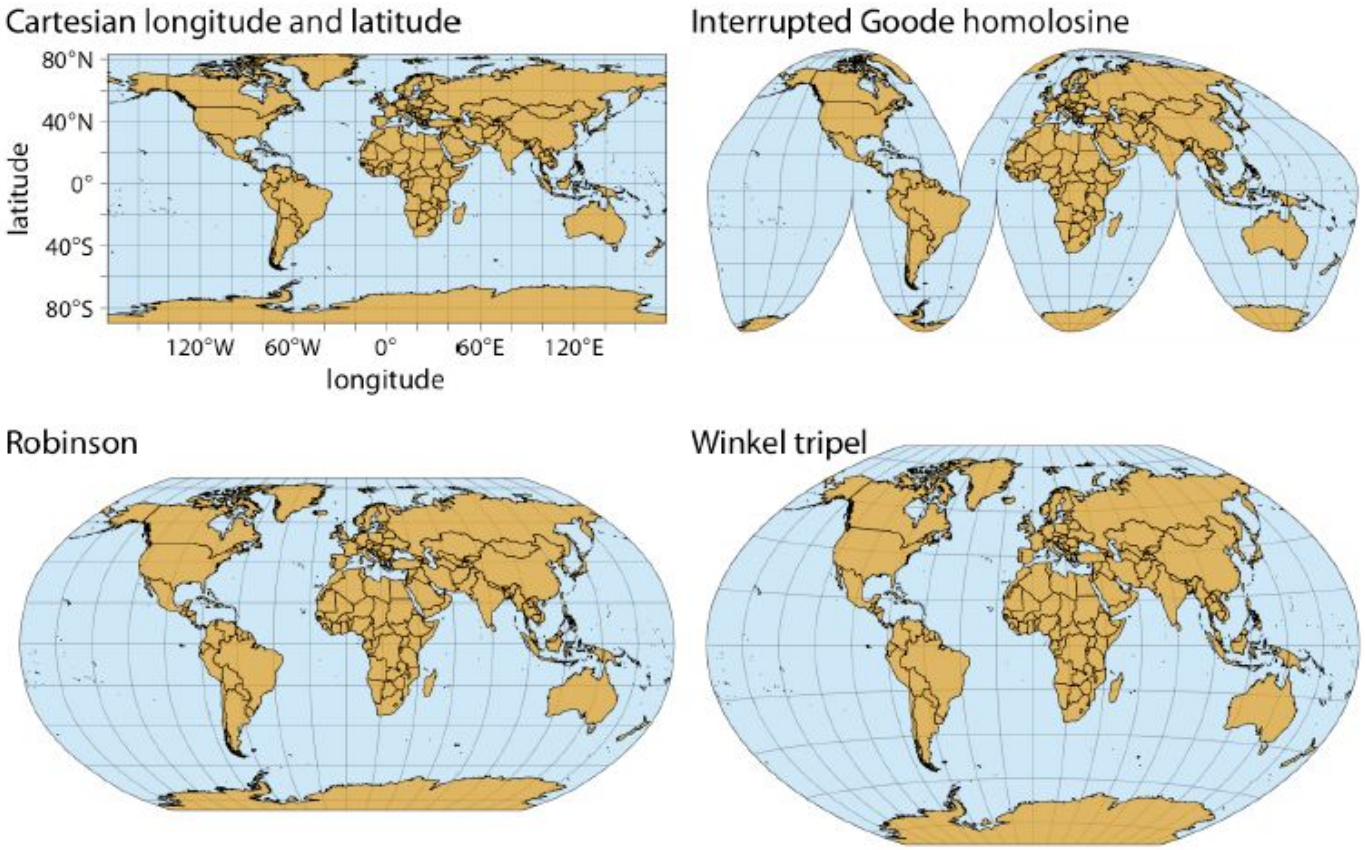
Coordinate System Axes

Circular Axes



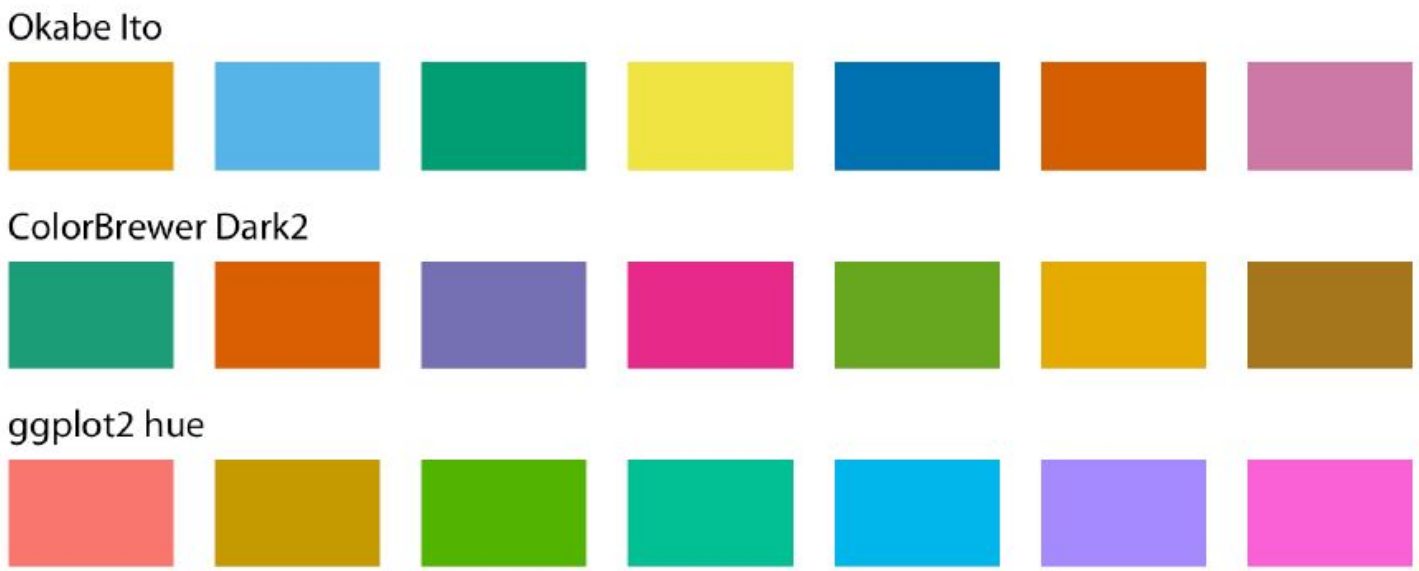


Example: Map of the World



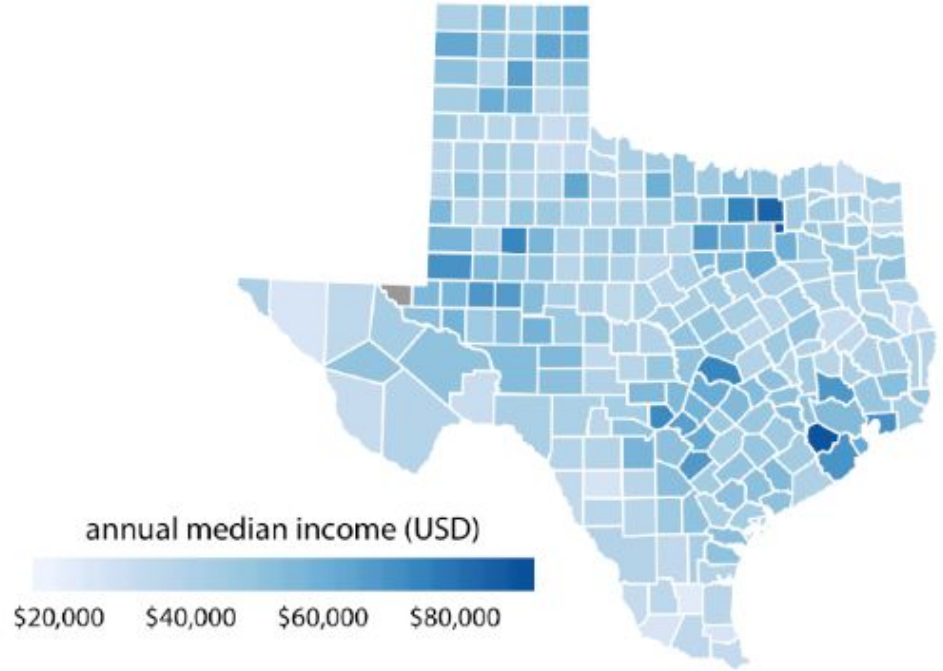
Color Scales

- Colors to distinguish groups of data from each other, to represent data values, and to highlight.
- In this case, we use standard qualitative color scale. Many of them available in community.



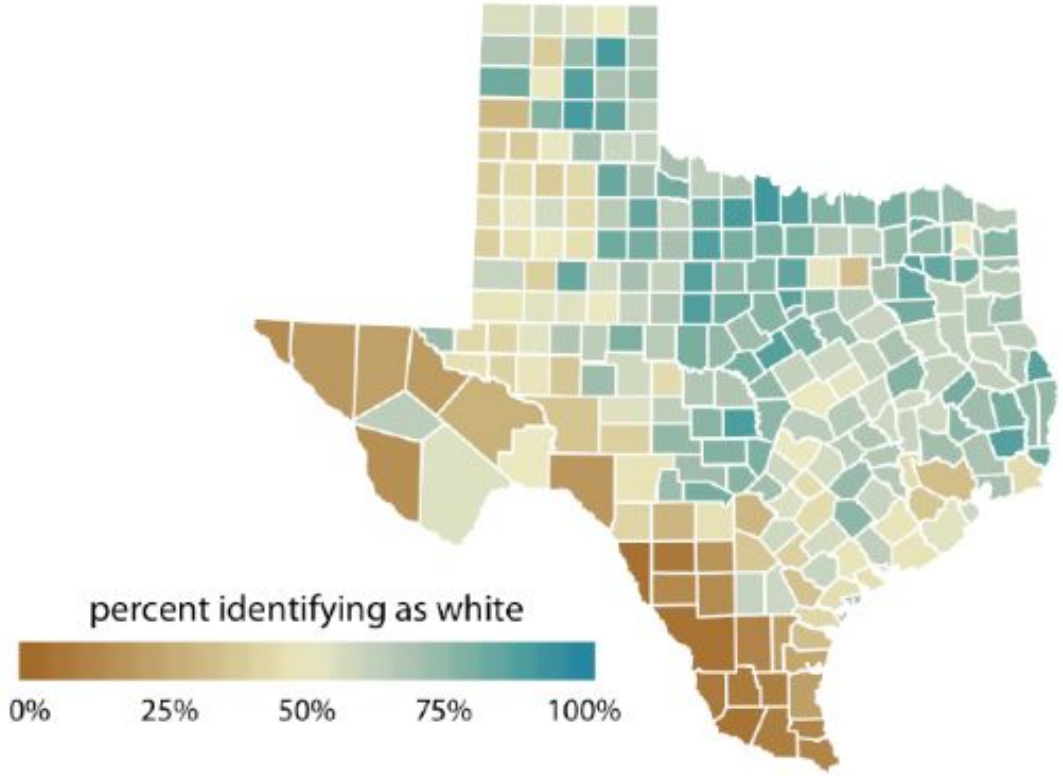
Color Scales

Represent Data Values



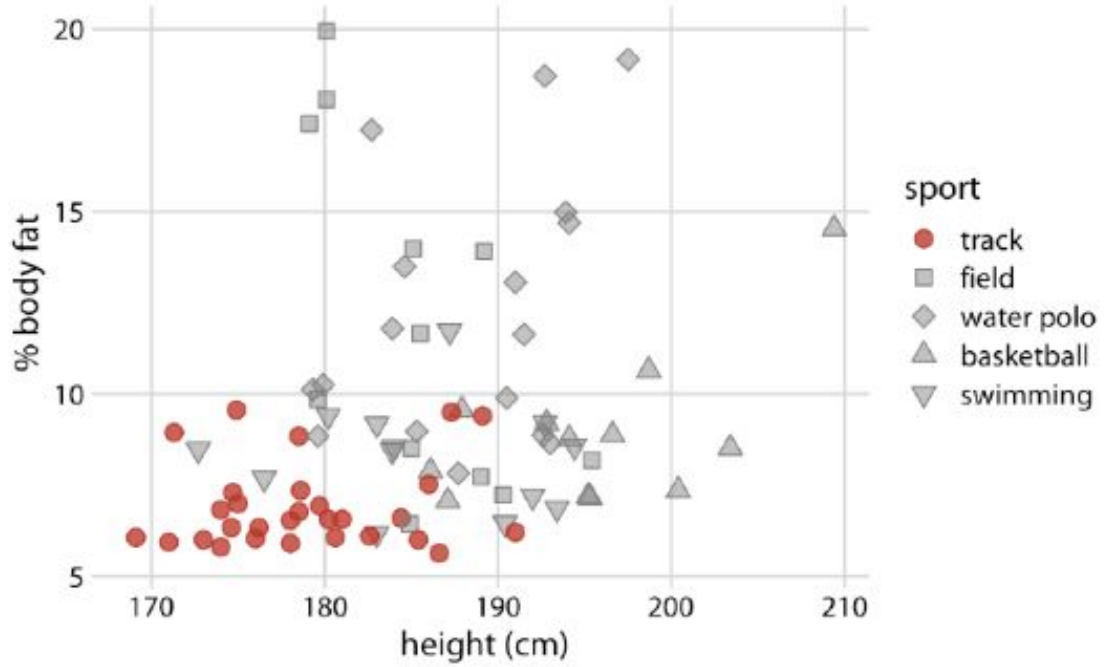
Color Scales

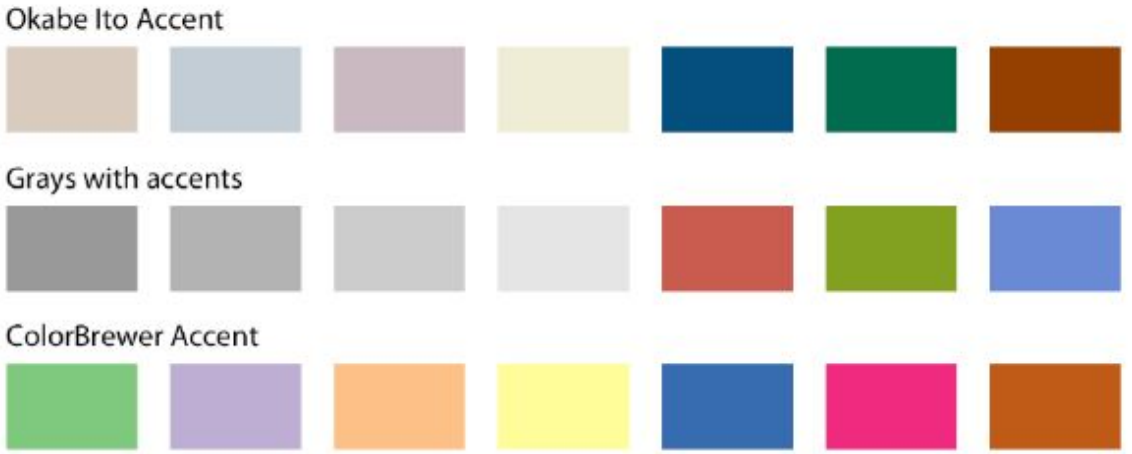
Diverging Data Values



Color Scales

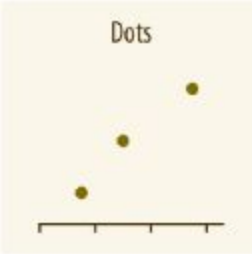
Color to Highlight



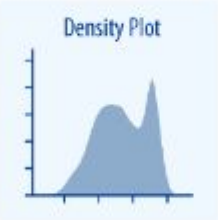


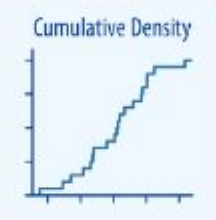
Directory of Visualizations

Amount

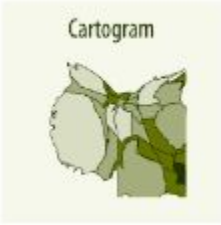


Distribution

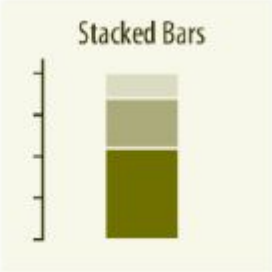




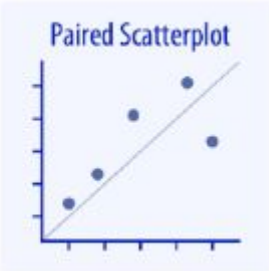
Geospatial

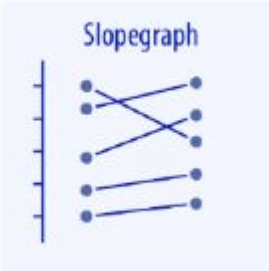


Proportion

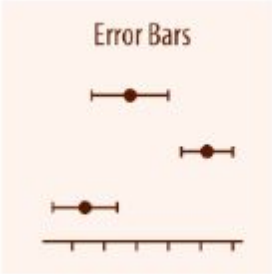


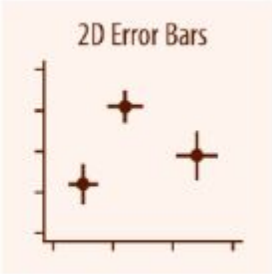
X-Y Relationship



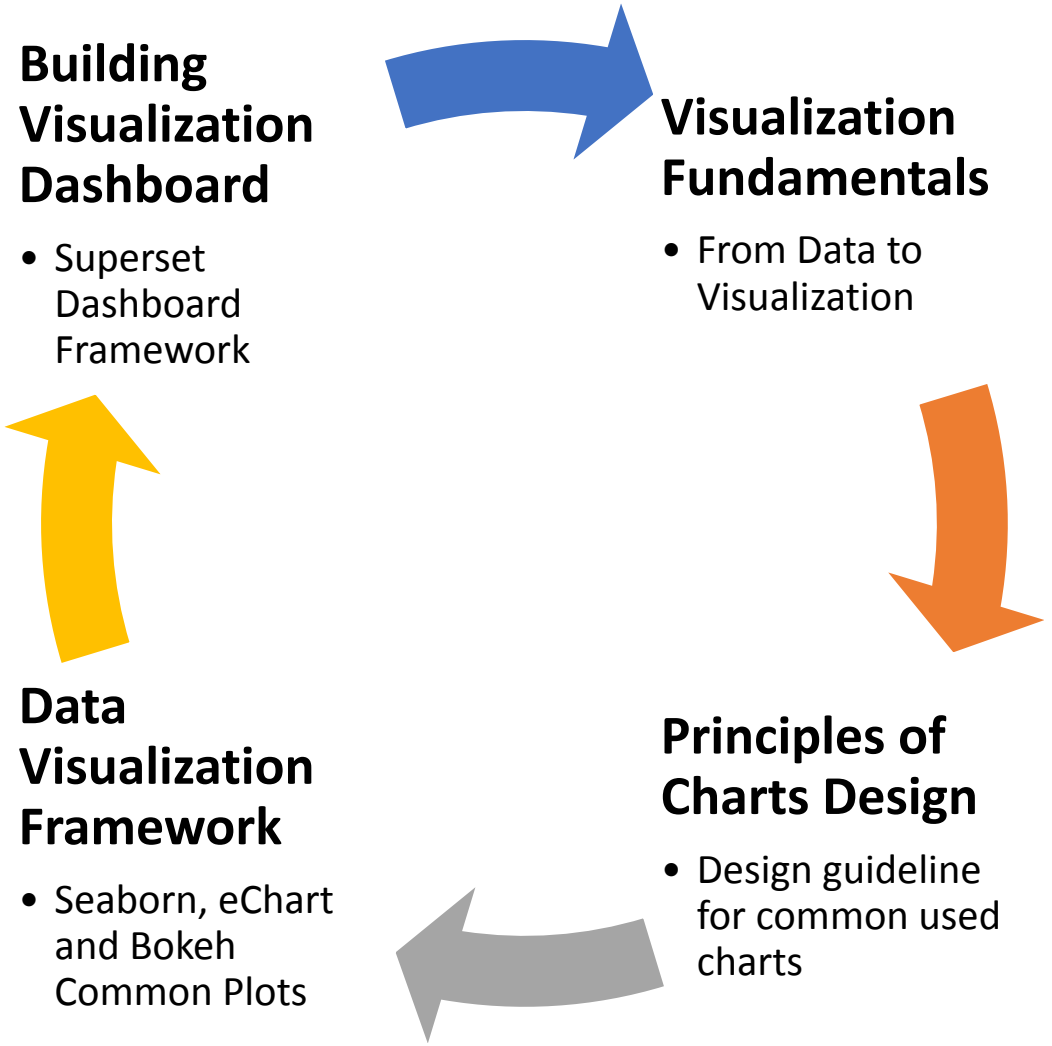


Uncertainty



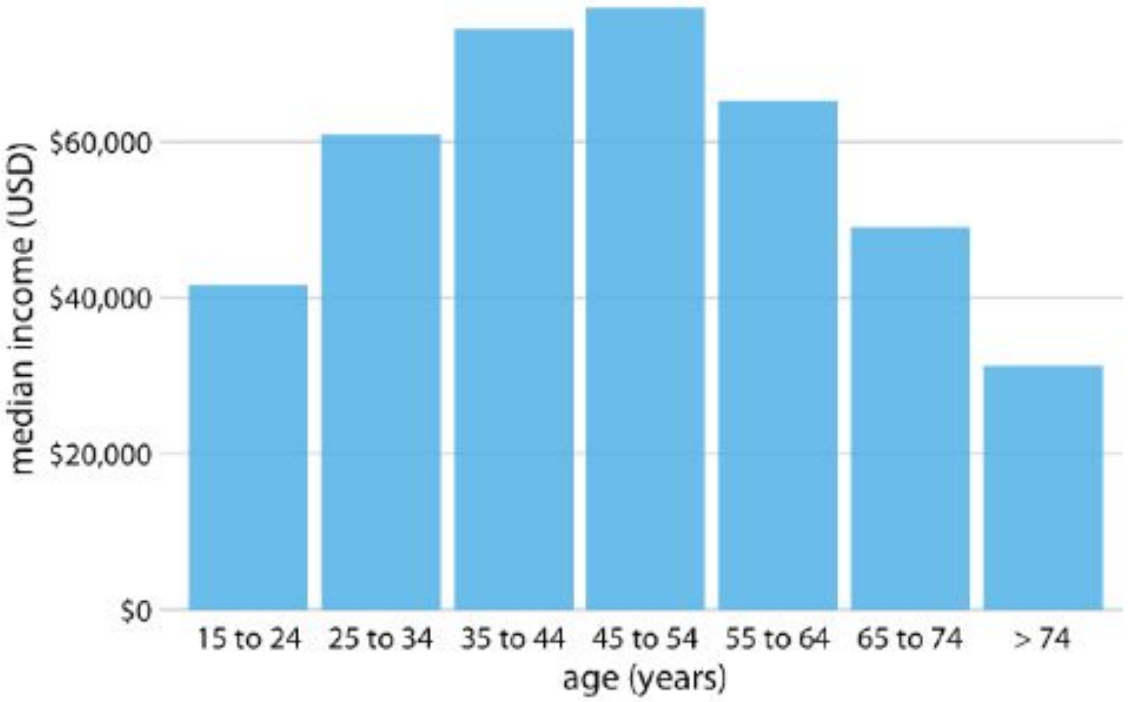


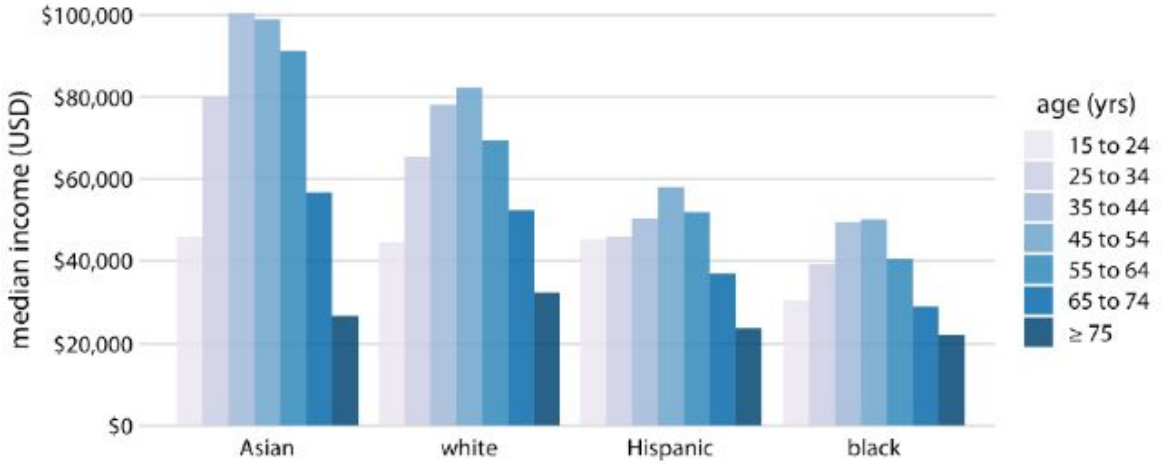
Outline – Data Visualization



Visualizing Amounts

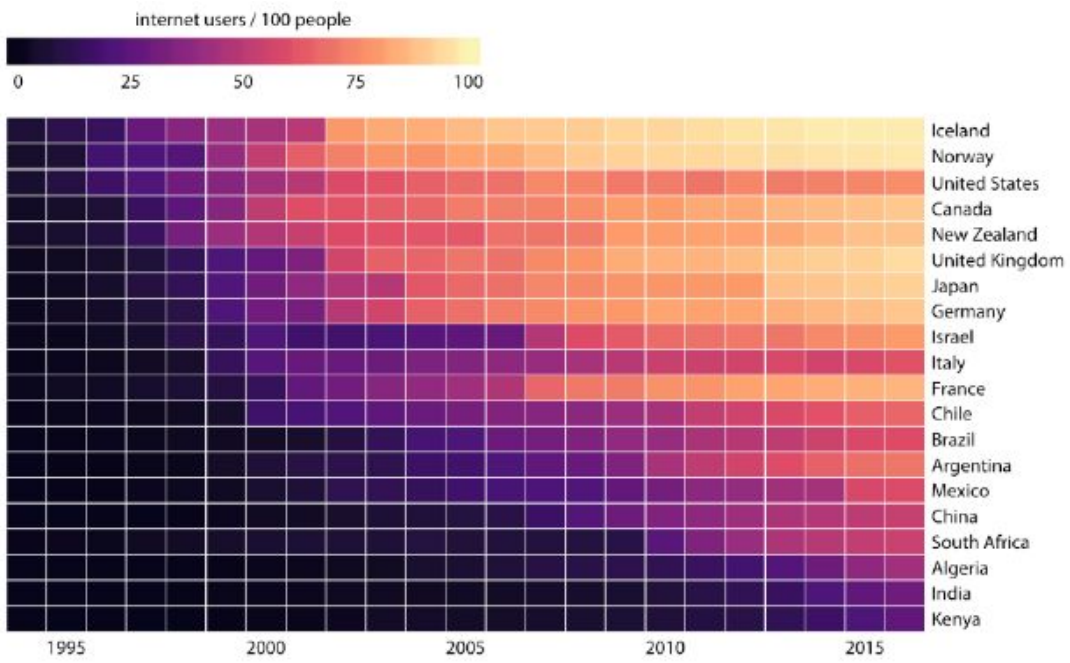
- We are interested in the magnitude of some set of numbers.
- Samples of visualization: **Bar and Stacked Bars**

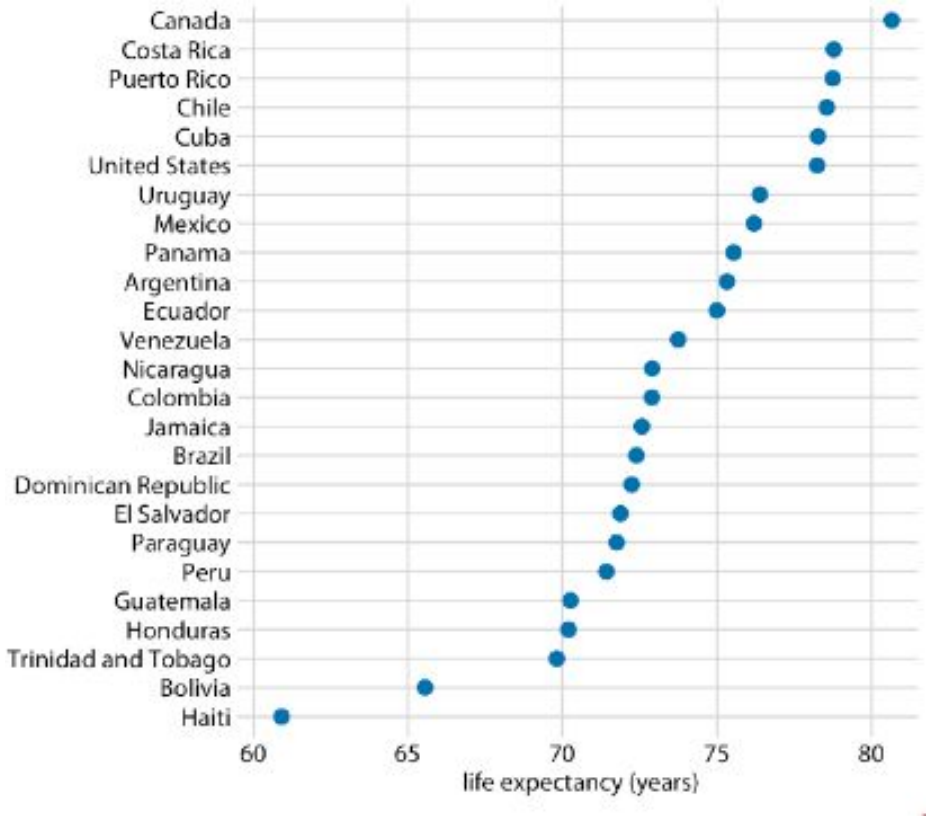




Visualizing Amounts

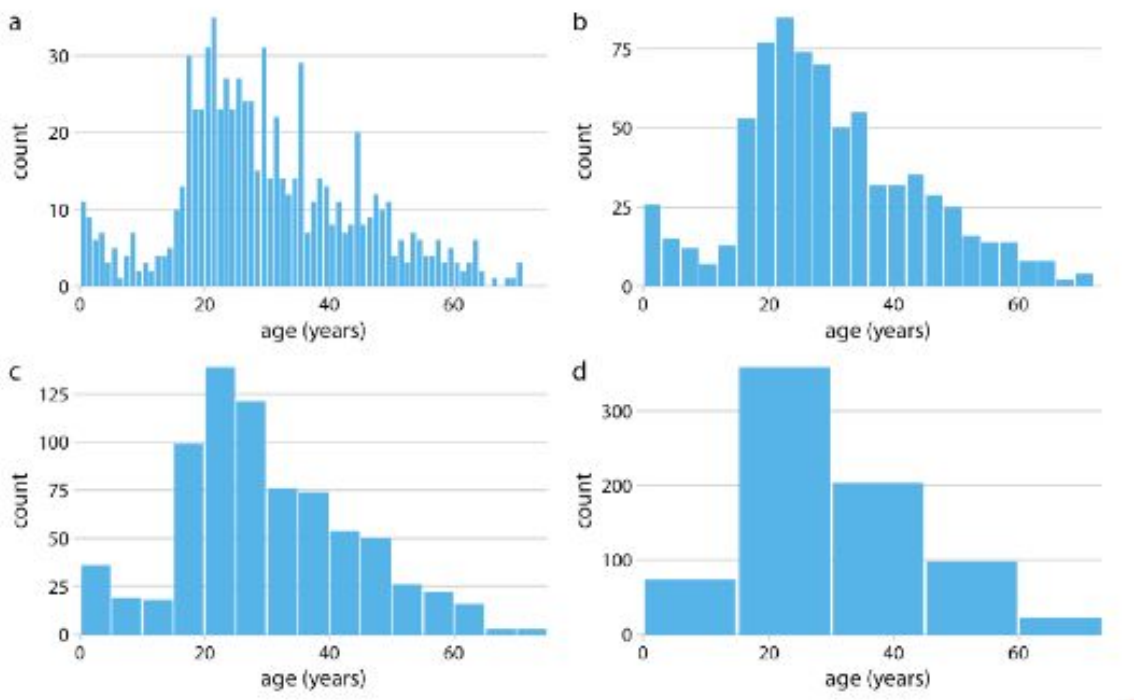
Samples of visualization: **Dot plot and Heatmap**

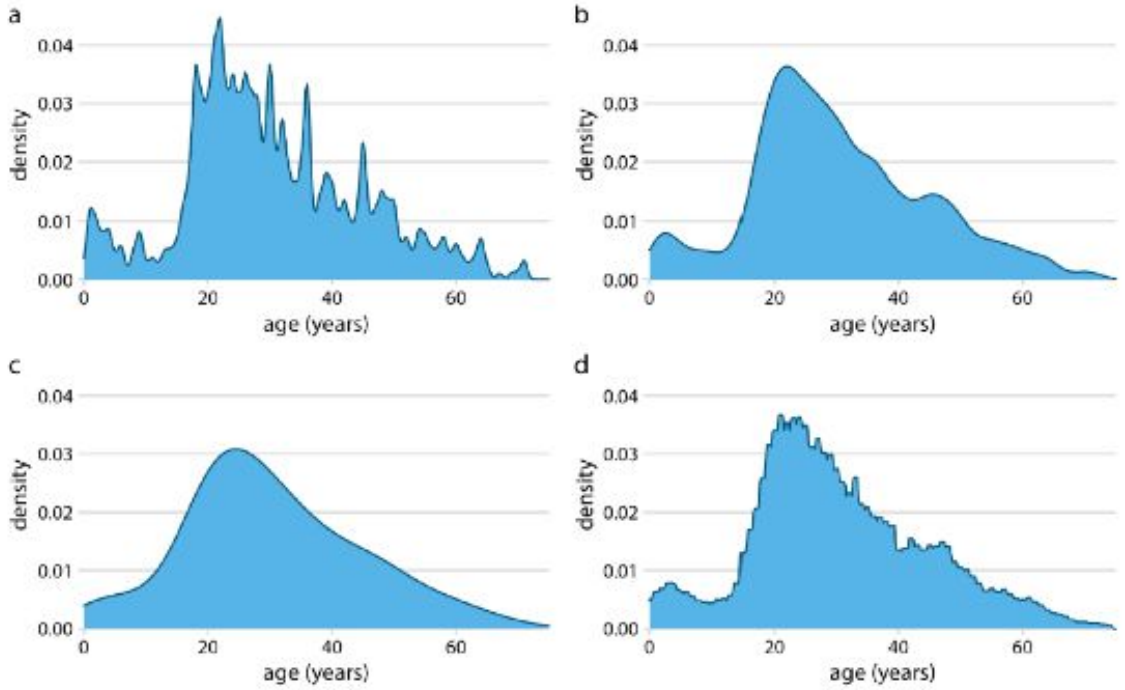




Visualizing Distributions

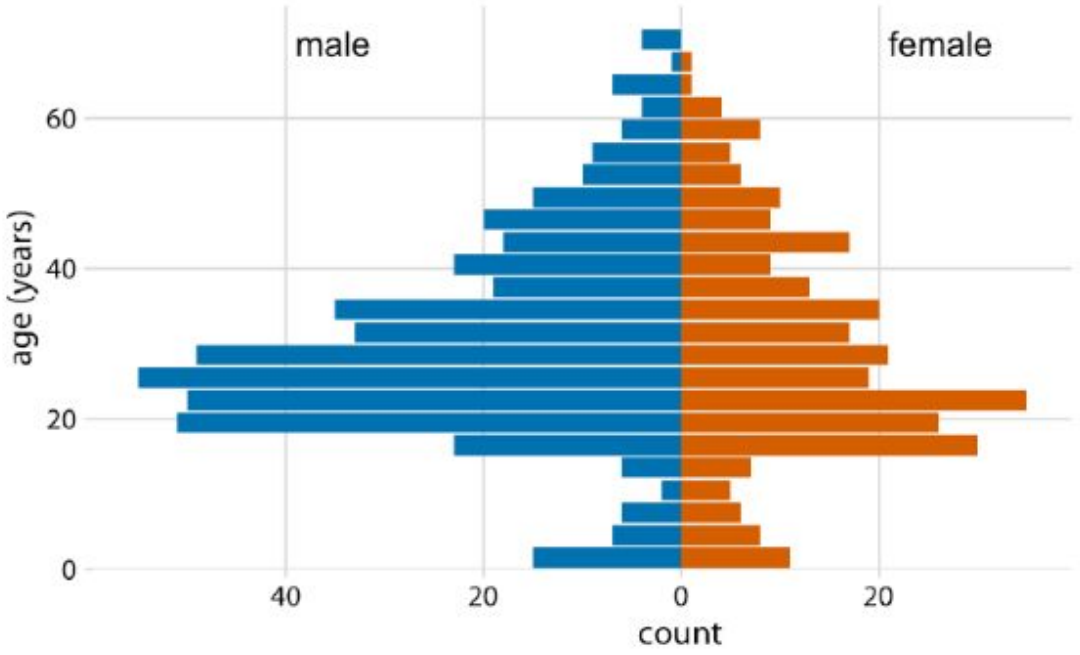
- Understand how a particular variable is distributed in a dataset.
- Sample of visualization: **Histograms and Density Plots**

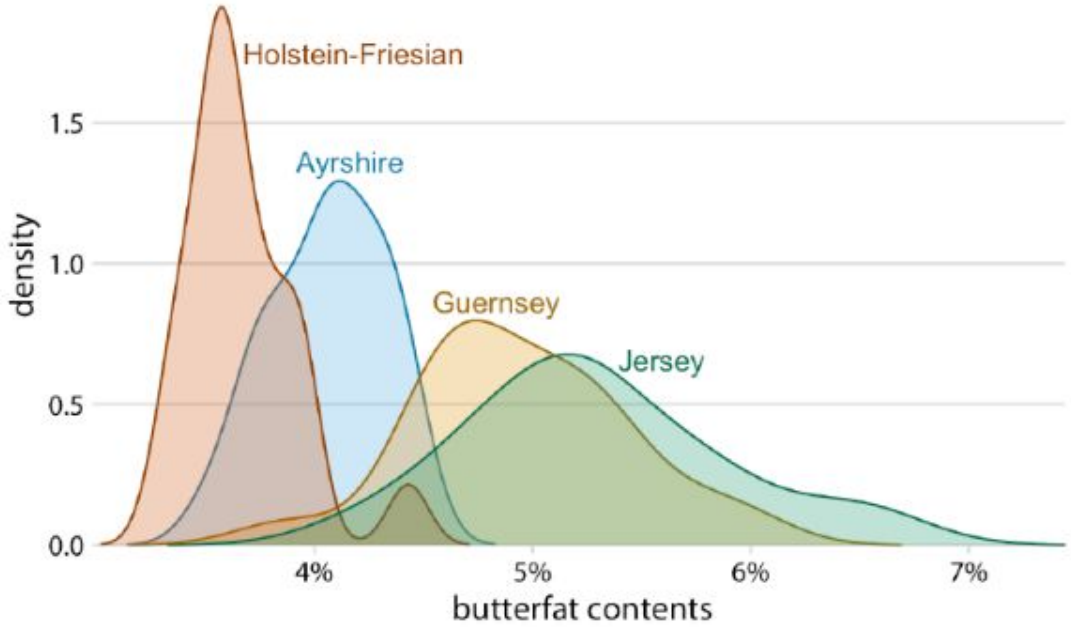




Visualizing Distributions

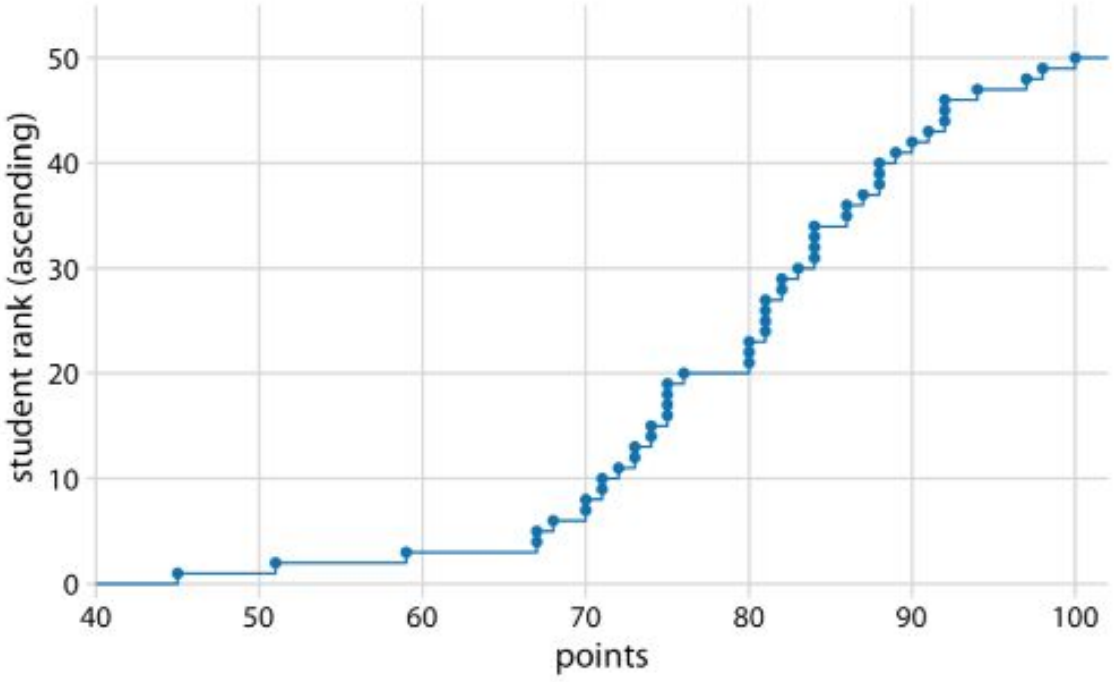
Sample of visualization: **Multiple Distributions**

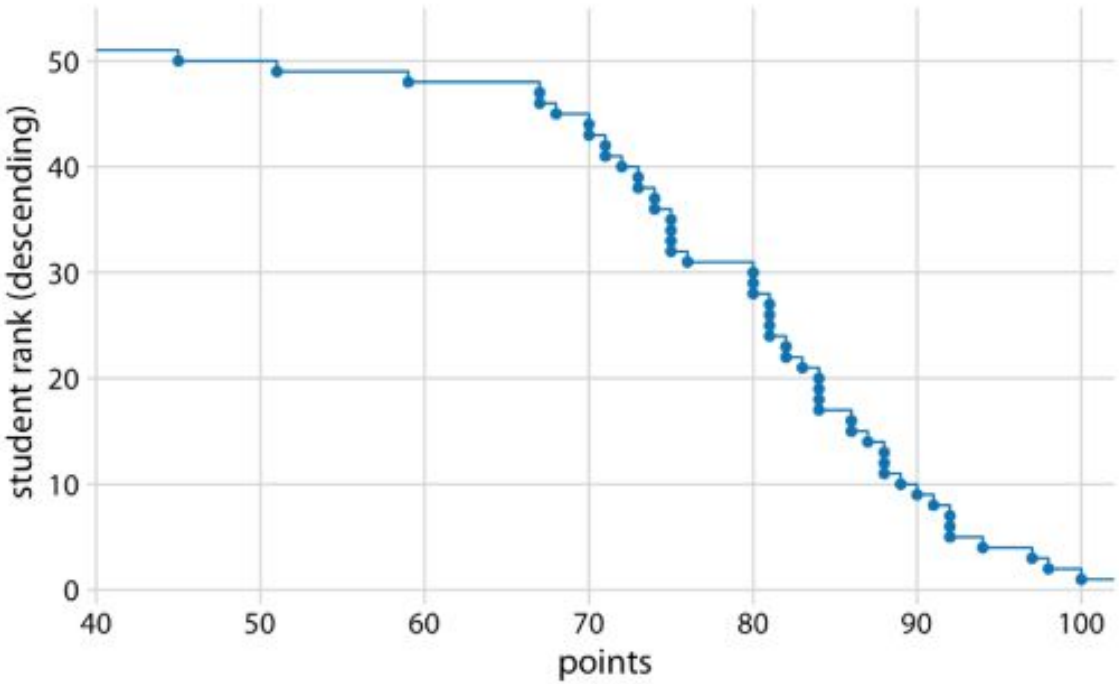




Visualizing Distributions

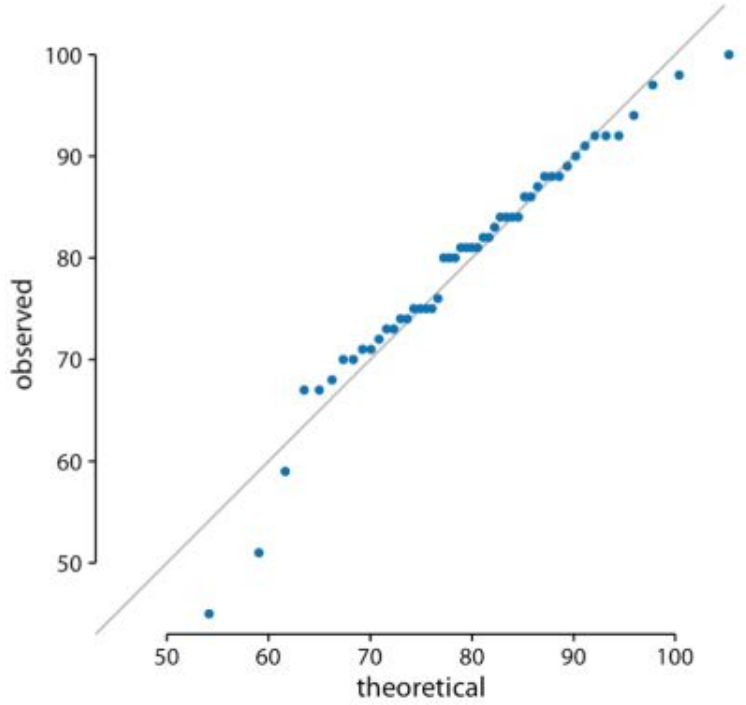
- **Empirical cumulative distribution functions (ECDFs)**, no arbitrary parameter choices, and they show all of the data at once.

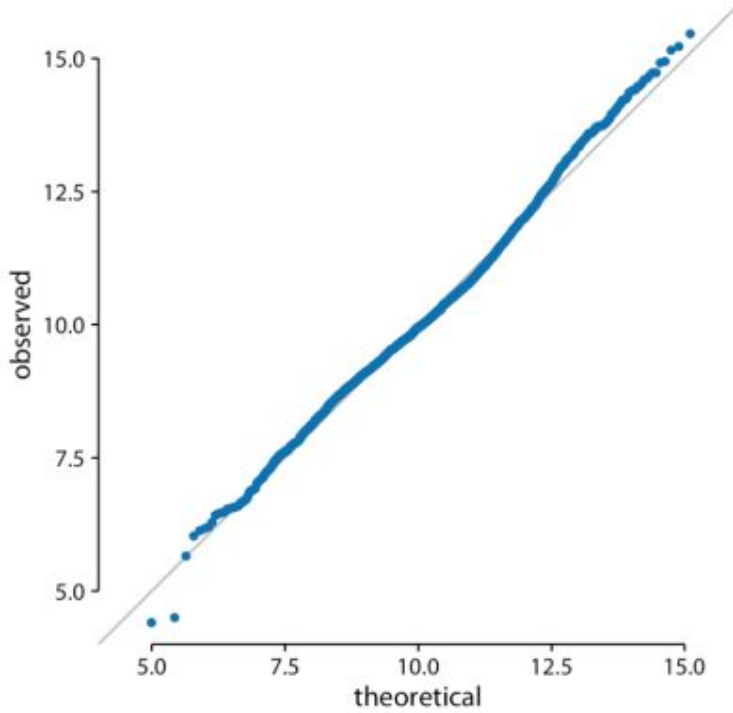




Visualizing Distributions

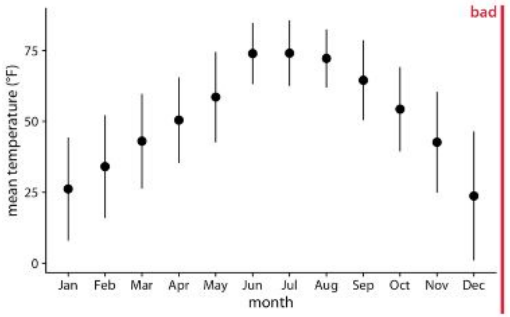
- **Quantile-quantile (q-q) plots** are useful when we want to determine to what extent the observed data points do or do not follow a given distribution.

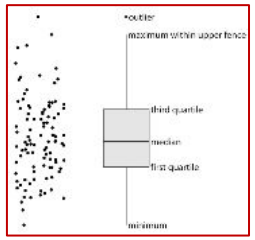


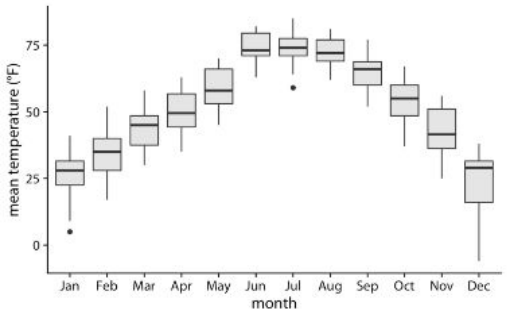


Visualizing Many Distributions at Once

- Think in terms of the **response variable** and one or more **grouping variables**. The *response variable* is the variable whose distributions we want to show. The *grouping variables* define subsets of the data with distinct distributions of the response variable. Sample: **Boxplot**.

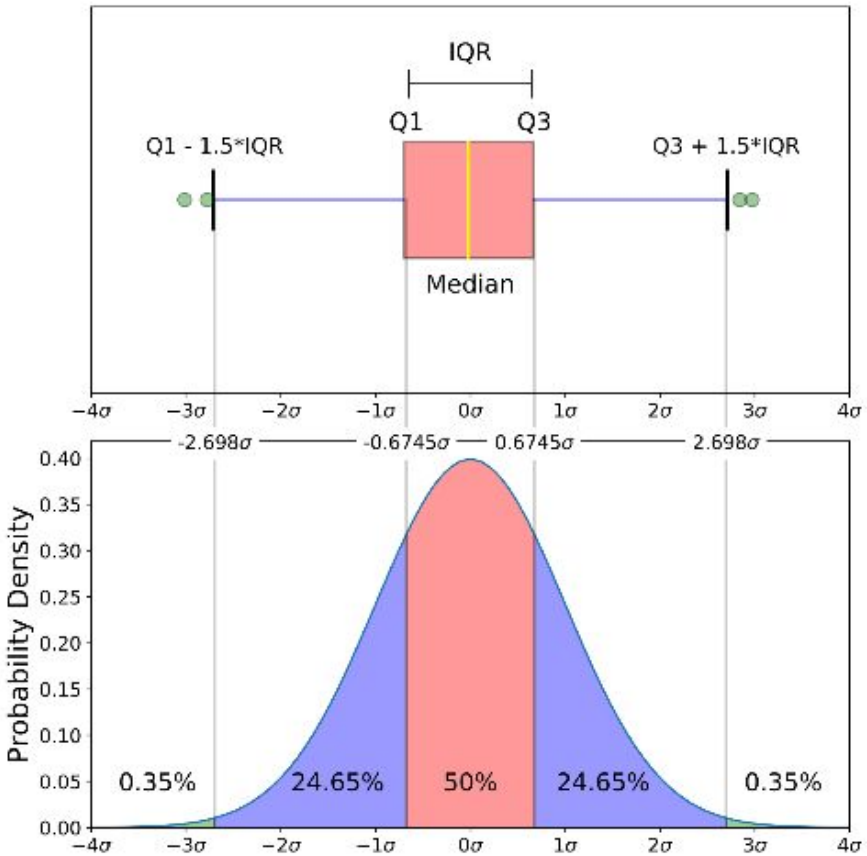


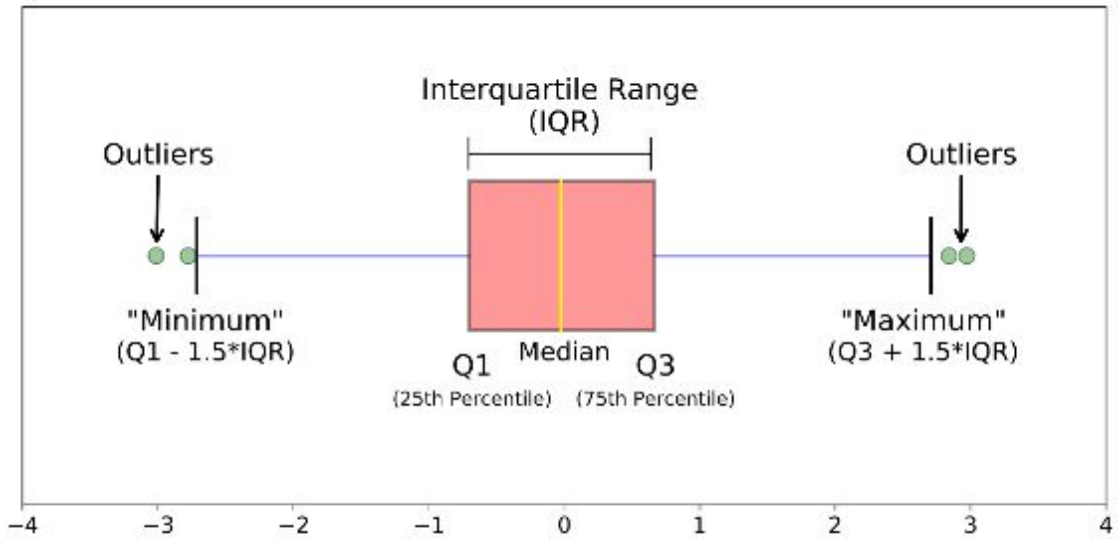




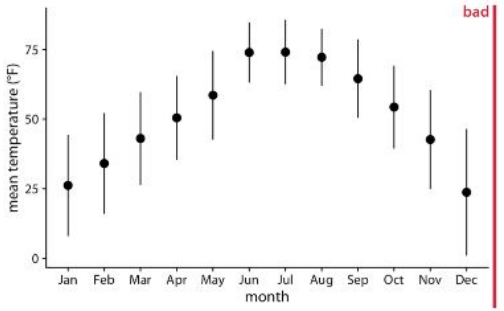
Boxplot Definition

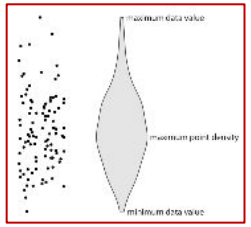
Show distributions by depicting groups of numerical data through quartiles.

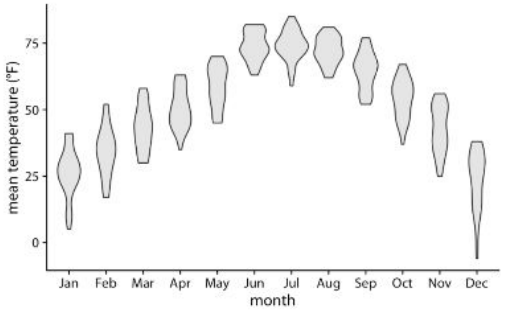




Visualizing Many Distributions at Once



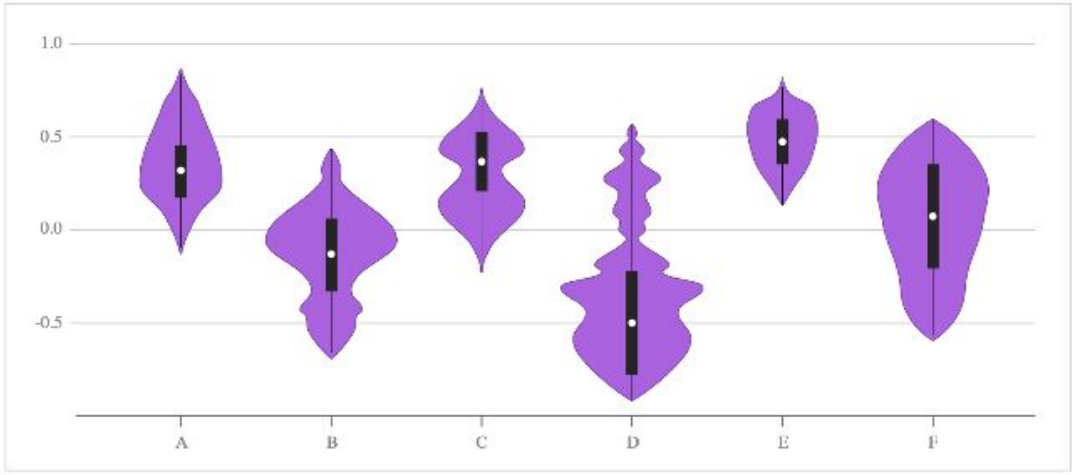


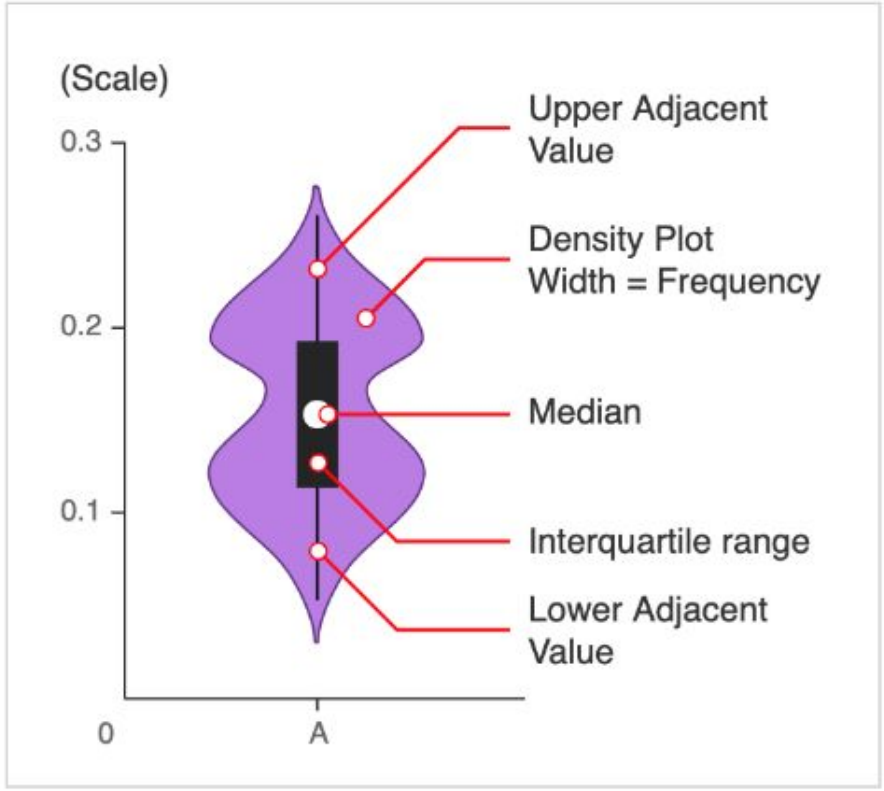


- Violins can be used whenever one would other-wise use a boxplot.
- Violin plots will accurately represent bimodal data whereas a boxplot will not.

Violin Plot Definition

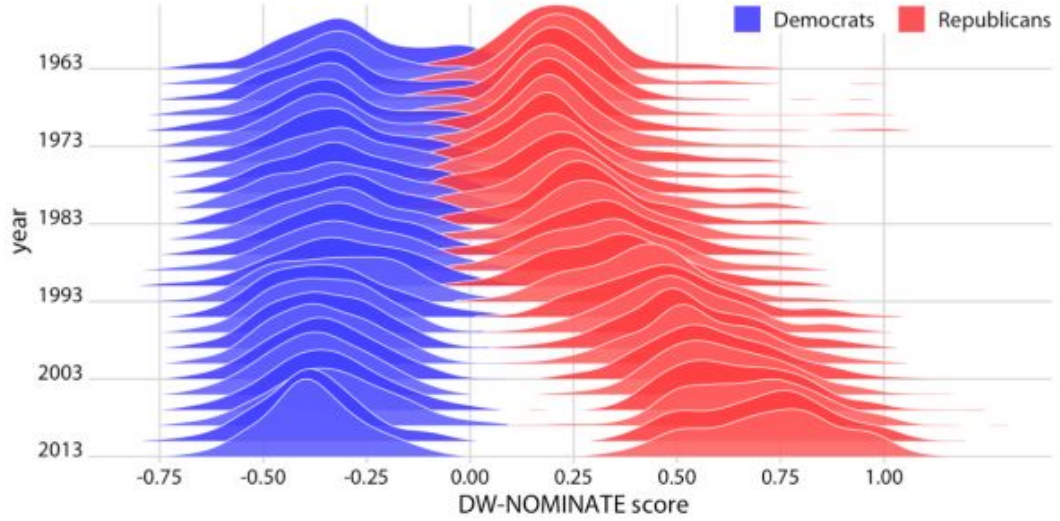
Visualize the distribution and its probability density = boxplot + density plot.

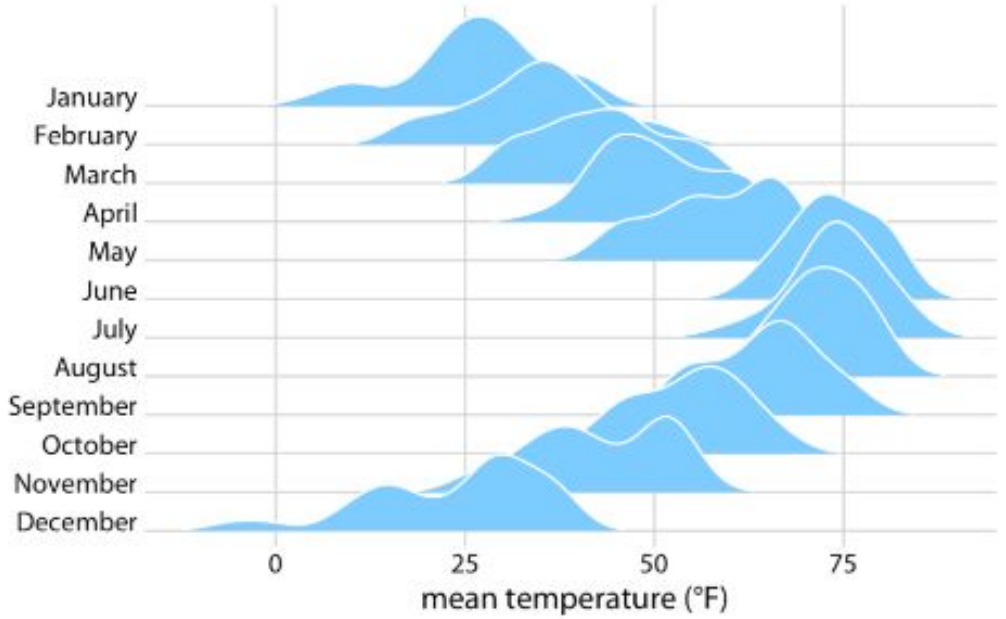




Ridgeline Plot Definition

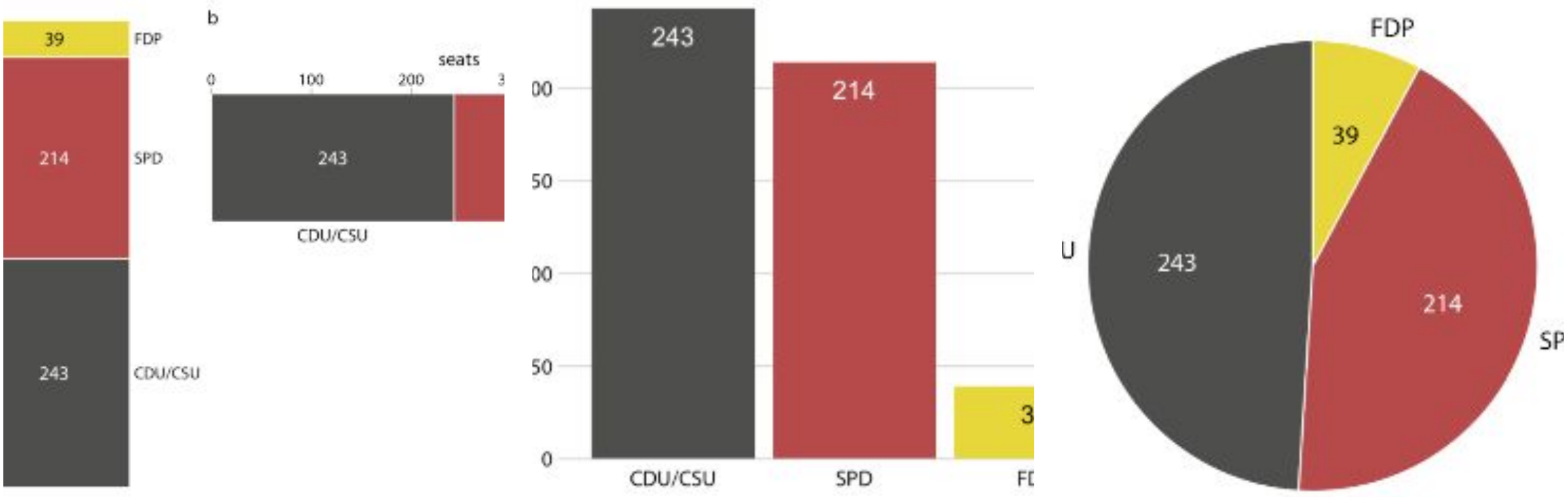
It looks like mountain ridgeline ☺️. Use it to show trends of distribution over time.





Visualizing Proportions

- Looks simple, but visualizing proportions can be challenging, in particular when the whole is broken into many different pieces or when we want to see changes in proportions over time or across conditions.



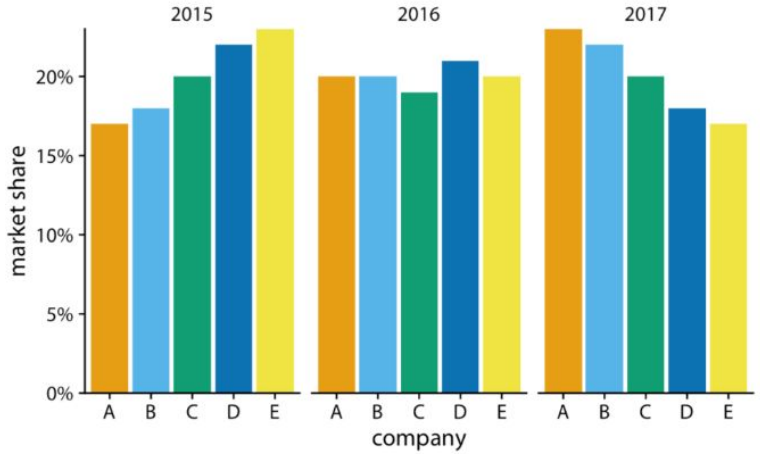
Visualizing Proportions

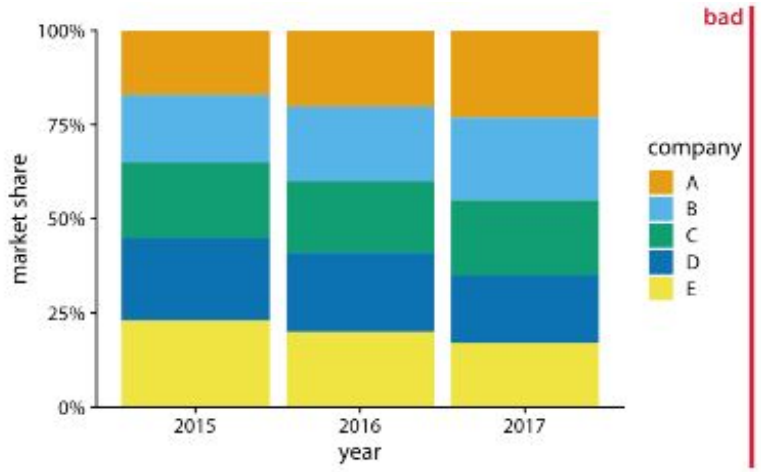
There is no single ideal visualization that always works.

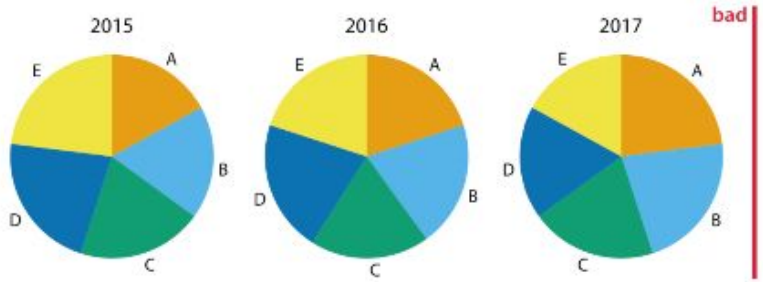
	Pie chart	Stacked bars	Side-by-side bars
Clearly visualizes the data as proportions of a whole	✓	✓	✗
Allows easy visual comparison of the relative proportions	✗	✗	✓
Visually emphasizes simple fractions, such as 1/2, 1/3, 1/4	✓	✗	✗
Looks visually appealing even for very small datasets	✓	✗	✓
Works well when the whole is broken into many pieces	✗	✗	✓
Works well for the visualization of many sets of proportions or time series of proportions	✗	✓	✗

Visualizing Proportions

Side-by-side bars case.

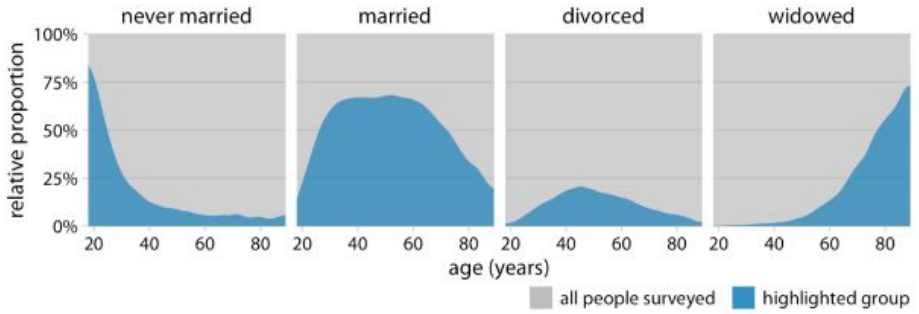


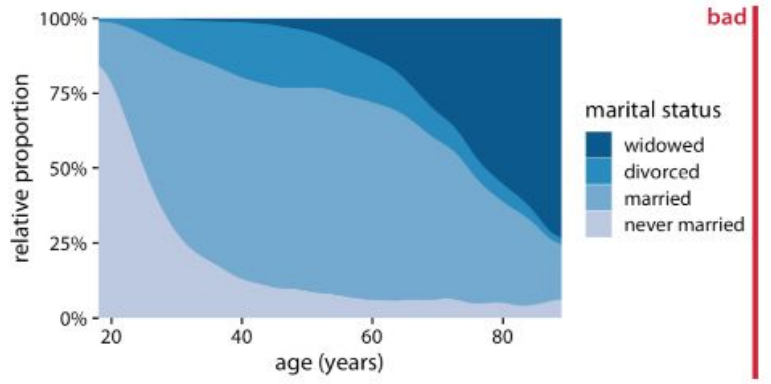


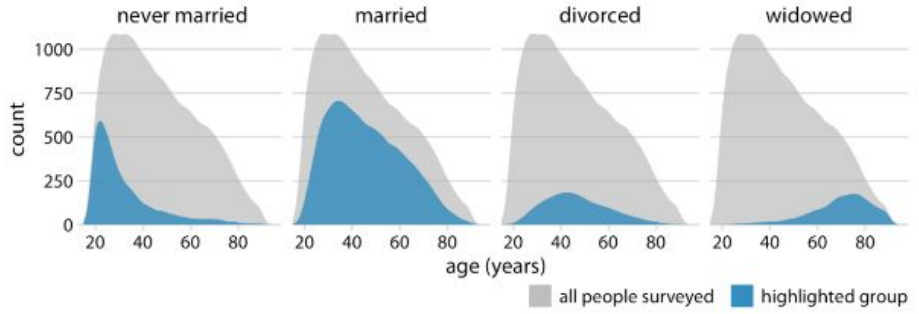


Visualizing Proportions

Case of Stacked Densities

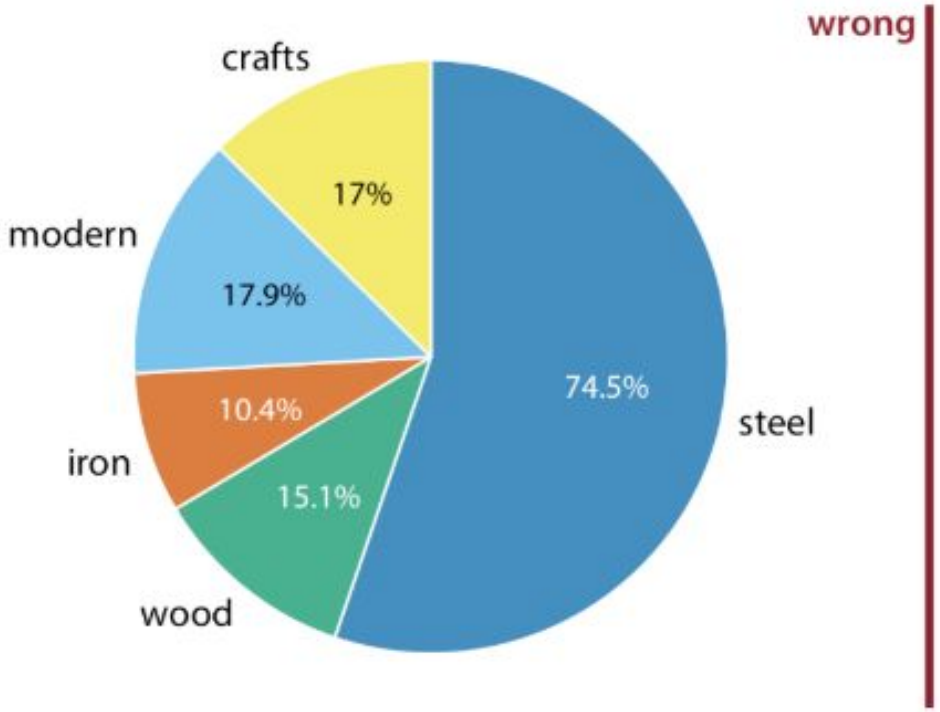


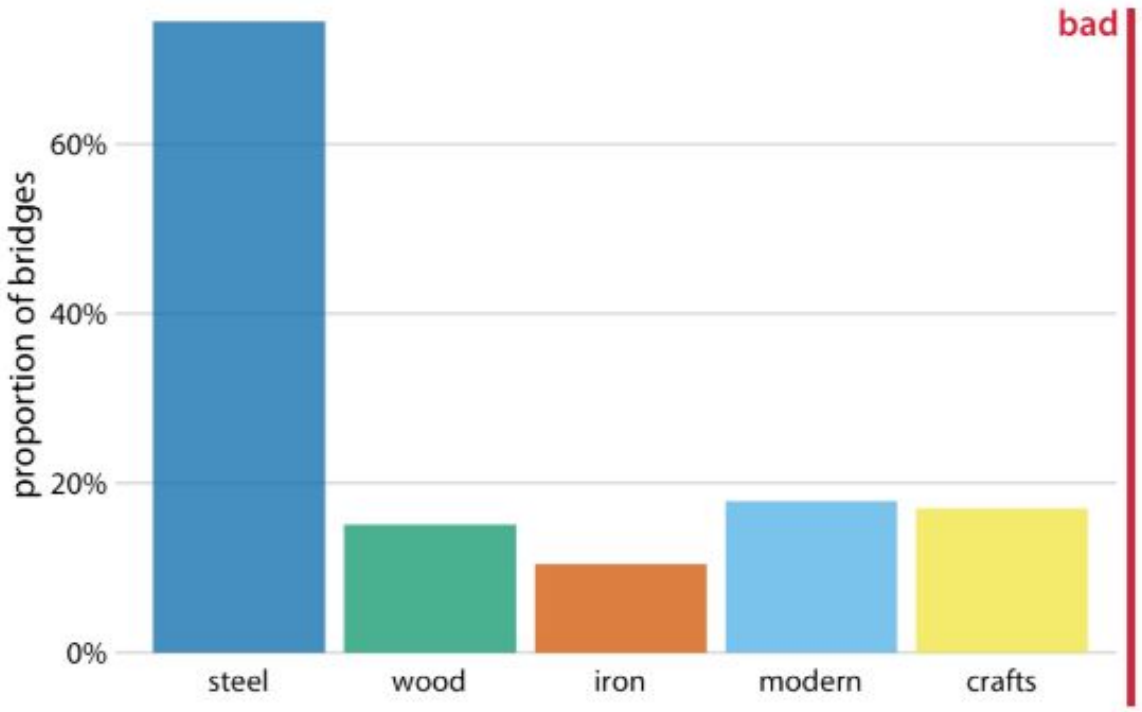




Visualizing Nested Proportions

- Nested proportions, because each additional categorical variable that we add creates a finer subdivision of the data nested within the previous proportions.

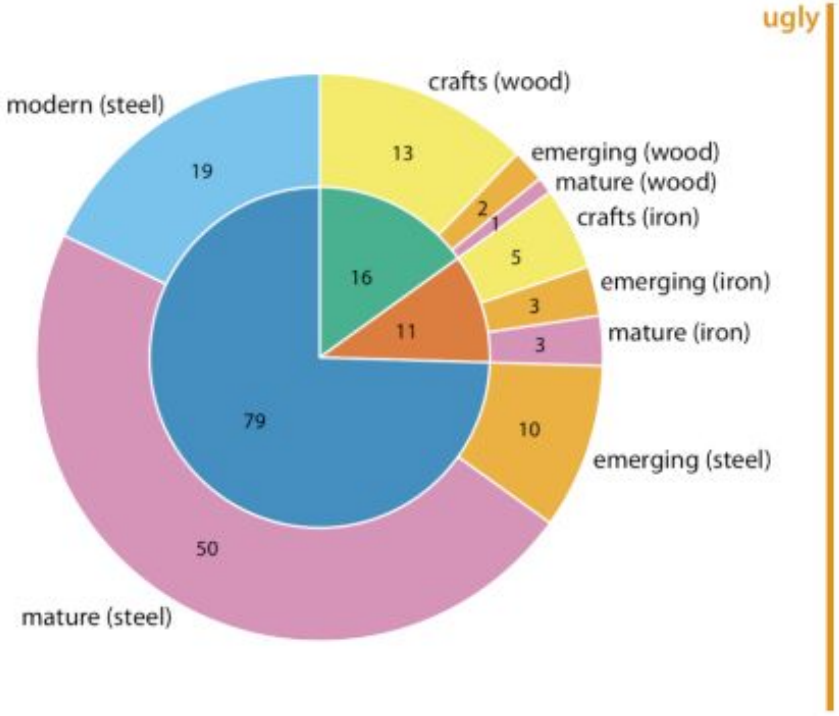


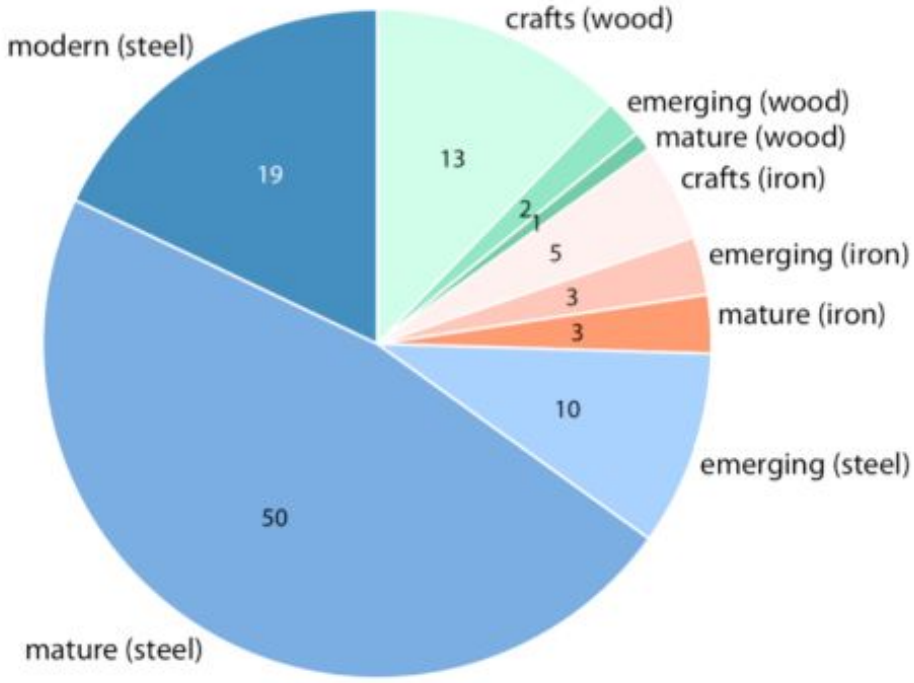


Visualizing Nested Proportions

- We can use nested map, but it is not intuitive to understand.

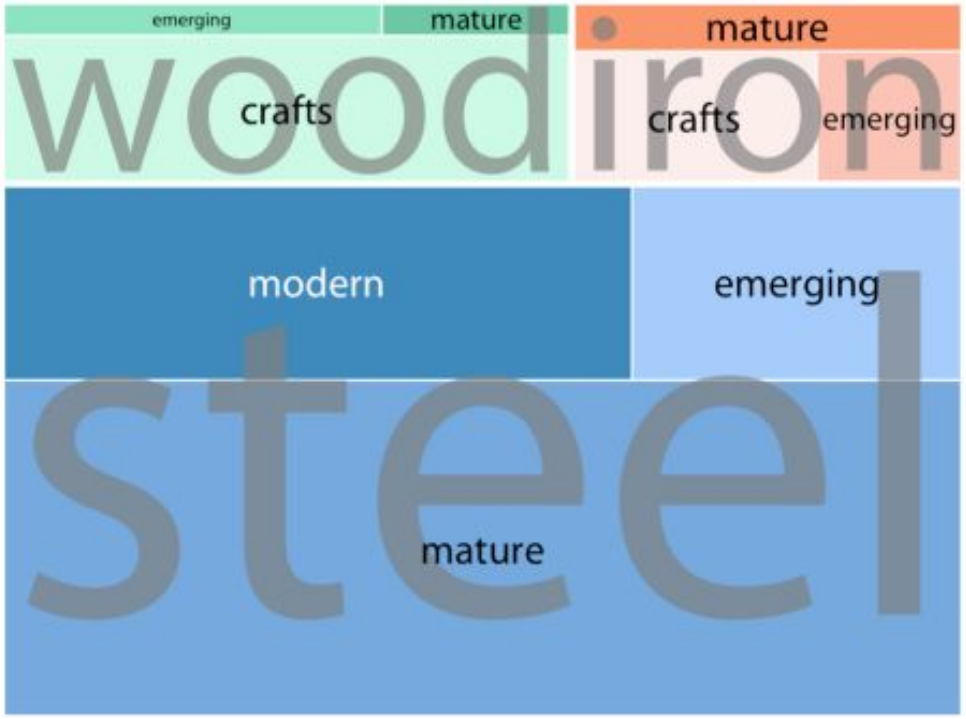
ugly

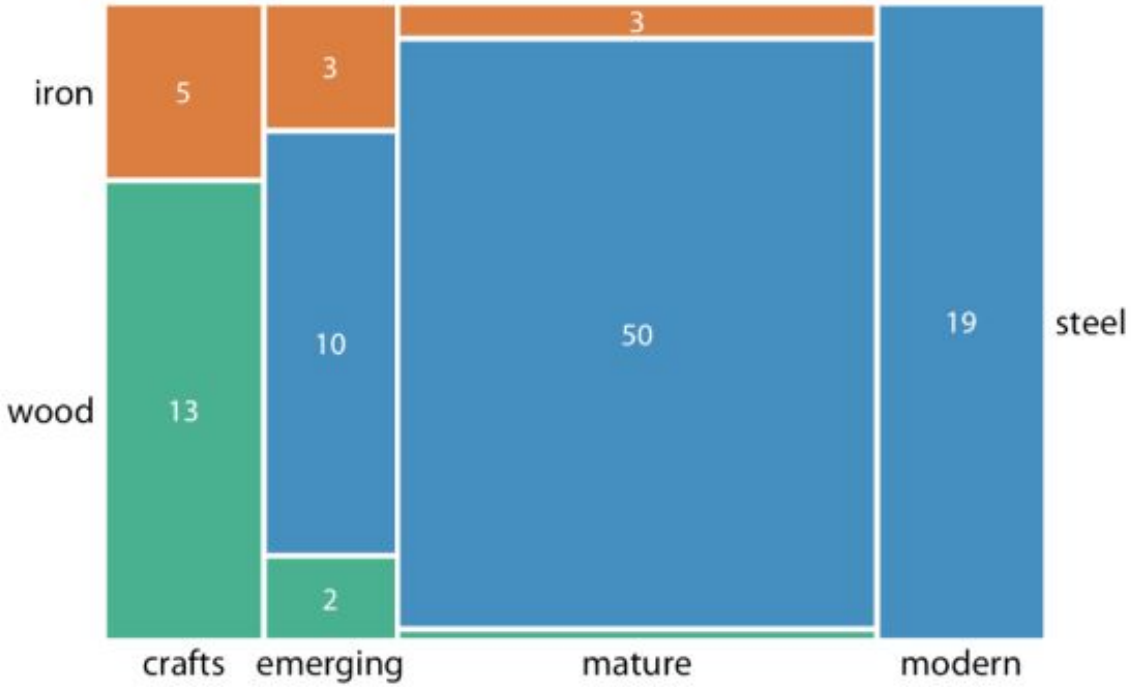




Visualizing Nested Proportions

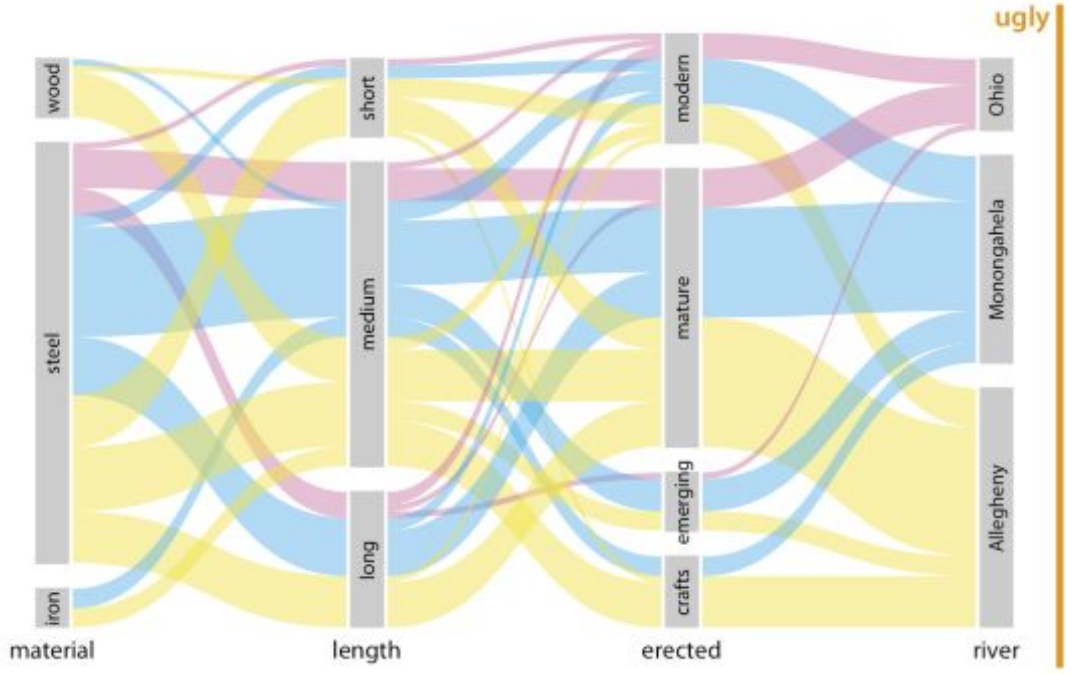
- There are several suitable approaches to visualize such nested proportions, including **mosaic plots** and **treemaps** below.

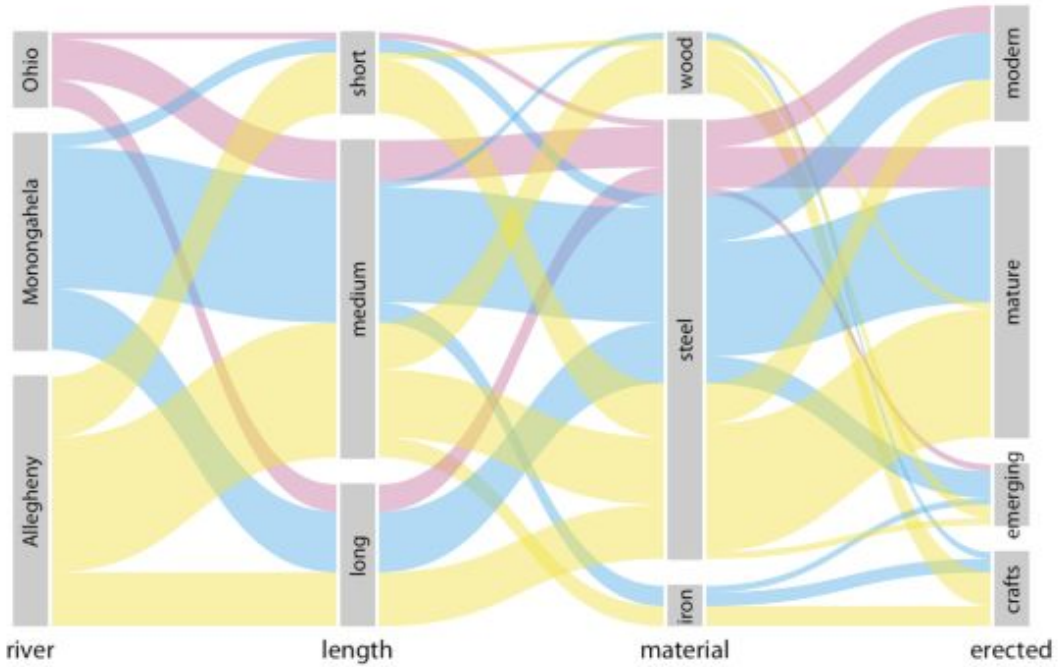




Visualizing Nested Proportions

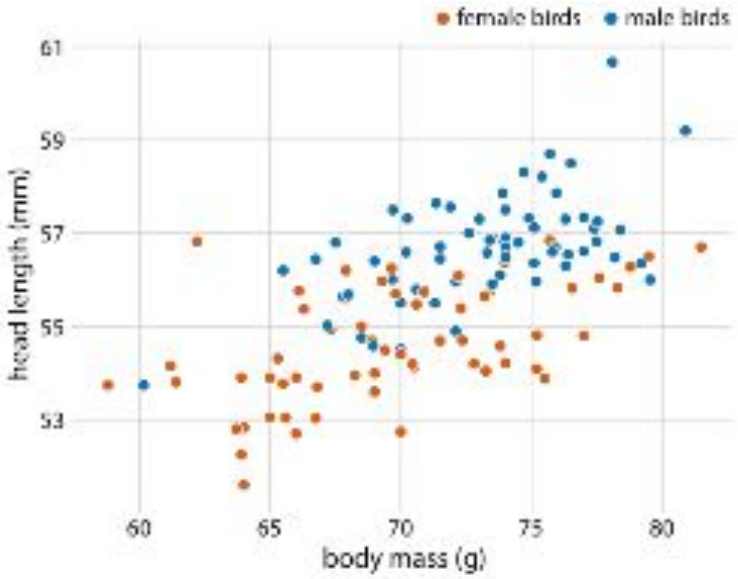
Mosaic plots, treemaps, and pie charts are OK, but choice is **parallel sets plot**.

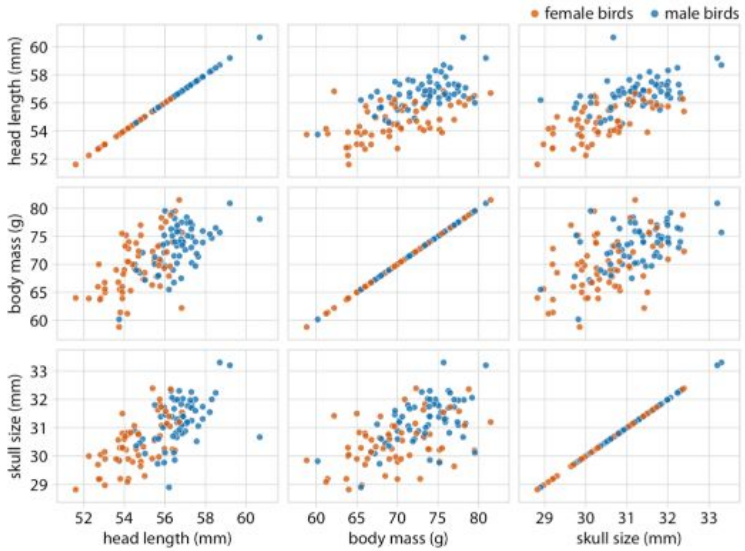


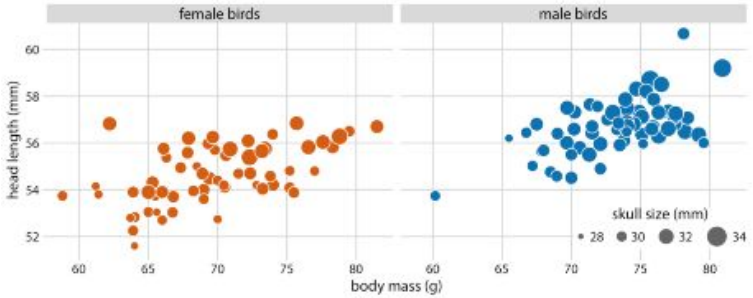


Visualizing Associations

- Many datasets contain two or more quantitative variables, and we may be interested in how these variables relate to each other. For example, **Scatter Plot**.

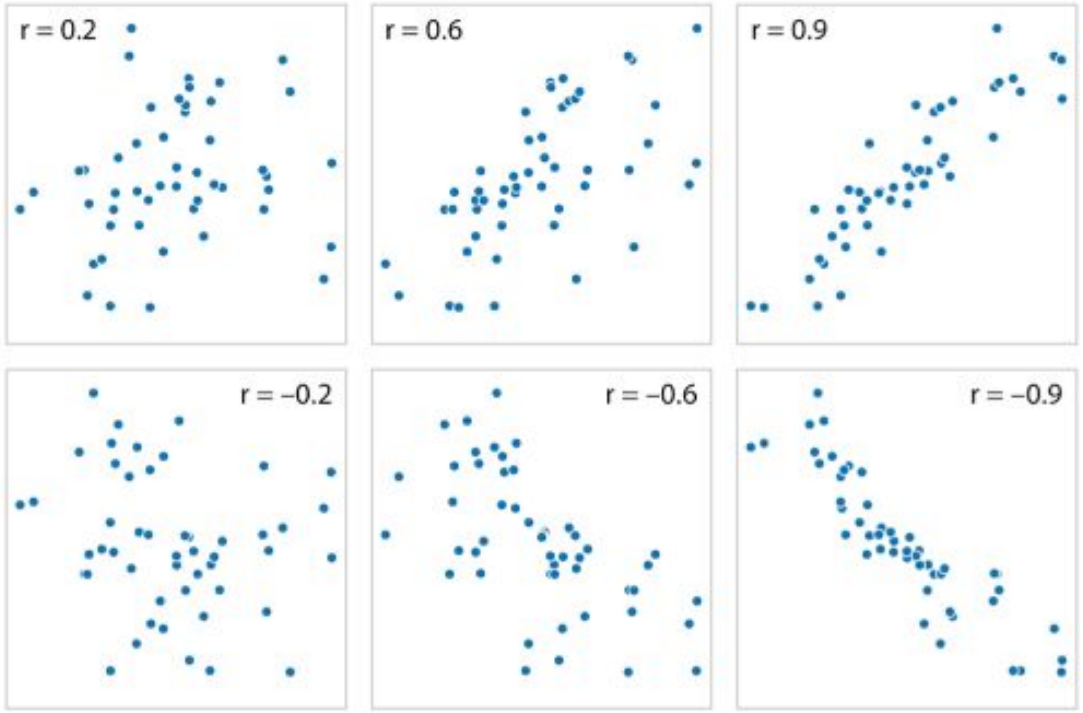






Visualizing Associations

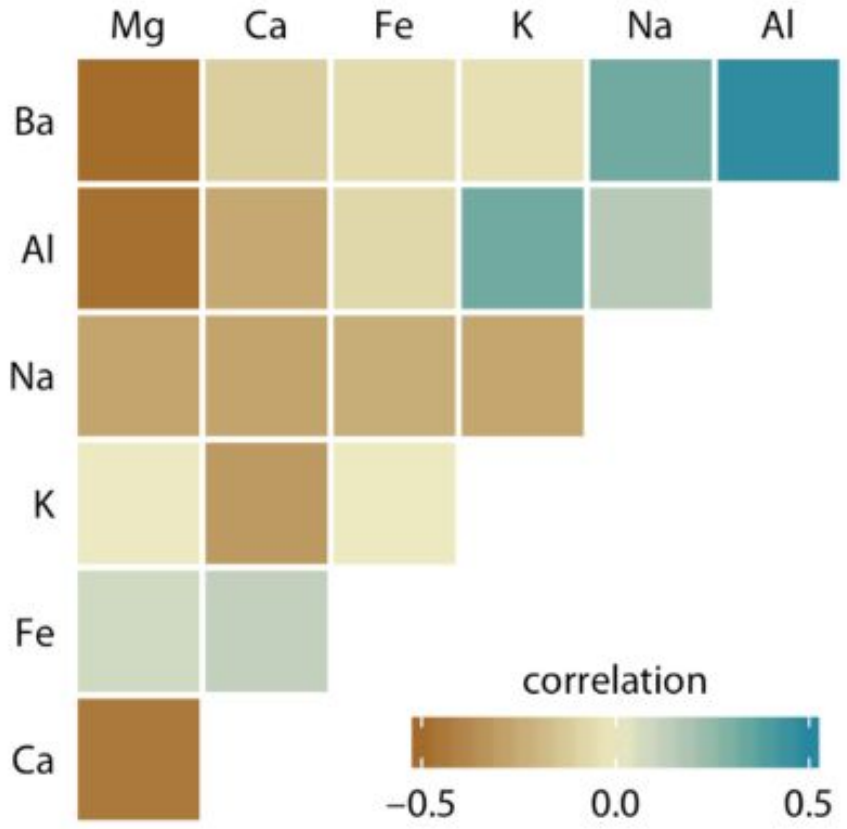
- When we have more than three to four quantitative variables, all-against-all scatter- plot matrices quickly become unwieldy. In this case, it is more useful to quantify the amount of association between pairs of variables and visualize these quantities rather than the raw data. One common way to do this is to calculate *correlation coefficients*.

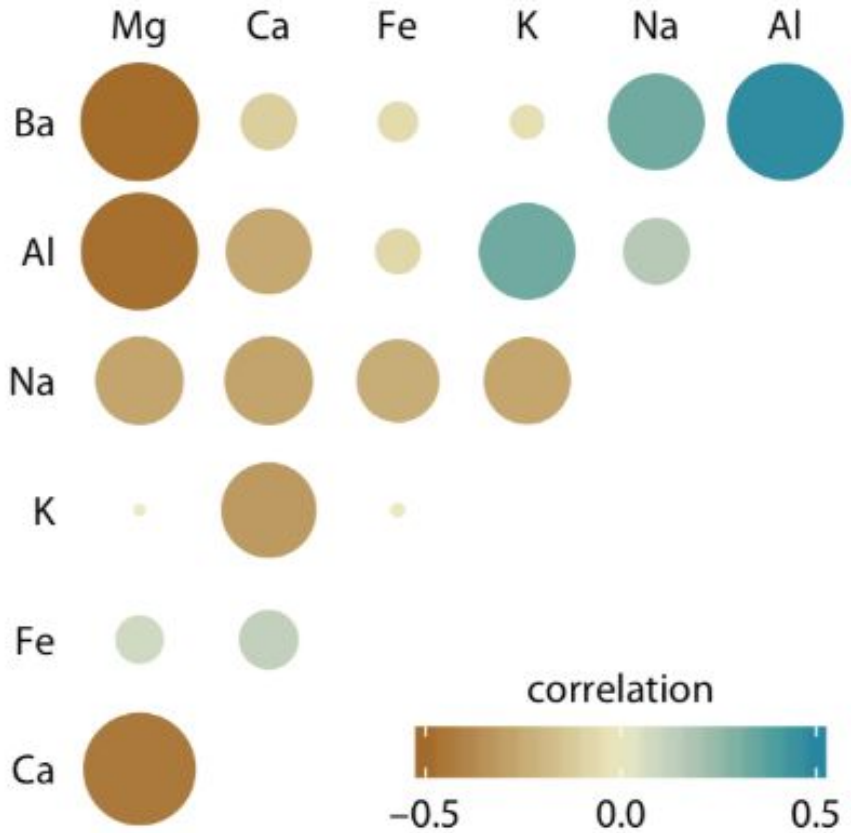


$$r = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i (x_i - \bar{x})^2} \sqrt{\Sigma_i (y_i - \bar{y})^2}}$$

Visualizing Associations

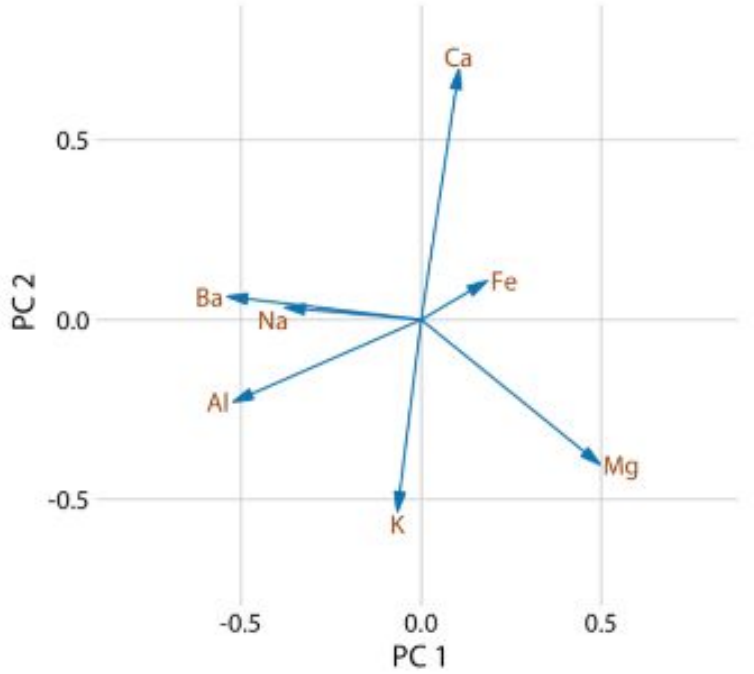
Visualizations of correlation coefficients are called **correlograms**.

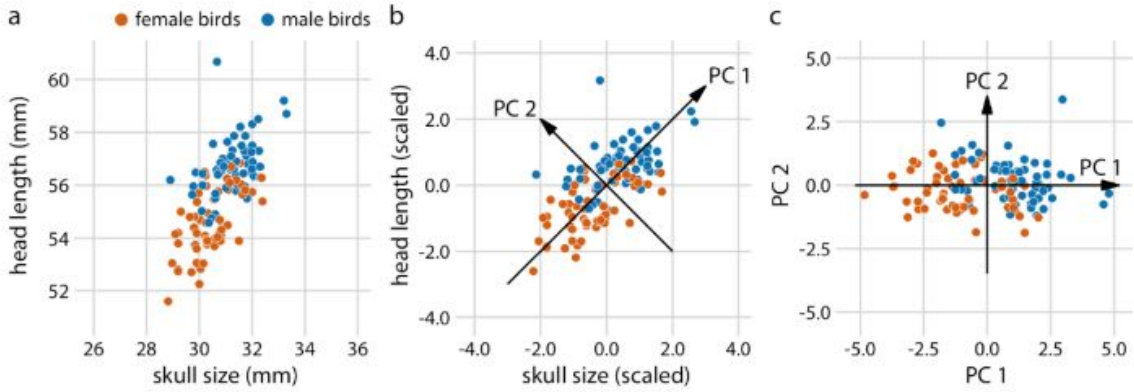




Visualizing Associations

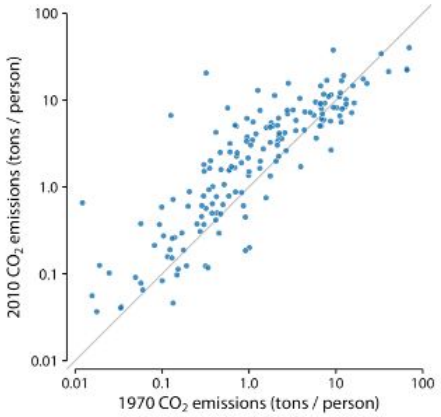
- Dimension reduction relies on the key insight that most high-dimensional datasets consist of multiple correlated variables that convey overlapping information. Widely used technique is PCA - *Principal Components Analysis.*

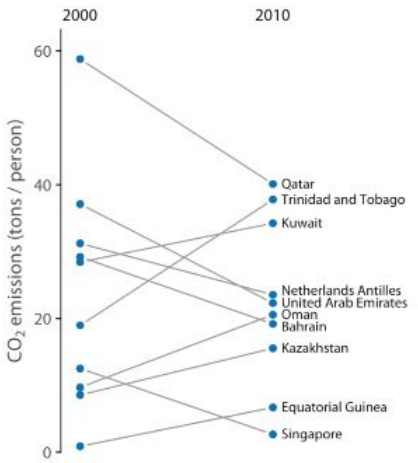


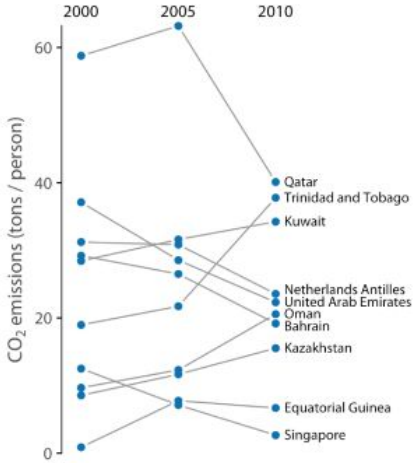


Visualizing Associations

- A special case of multivariate quantitative data is *paired data*: data where there are two or more measurements of the same quantity under slightly different conditions.

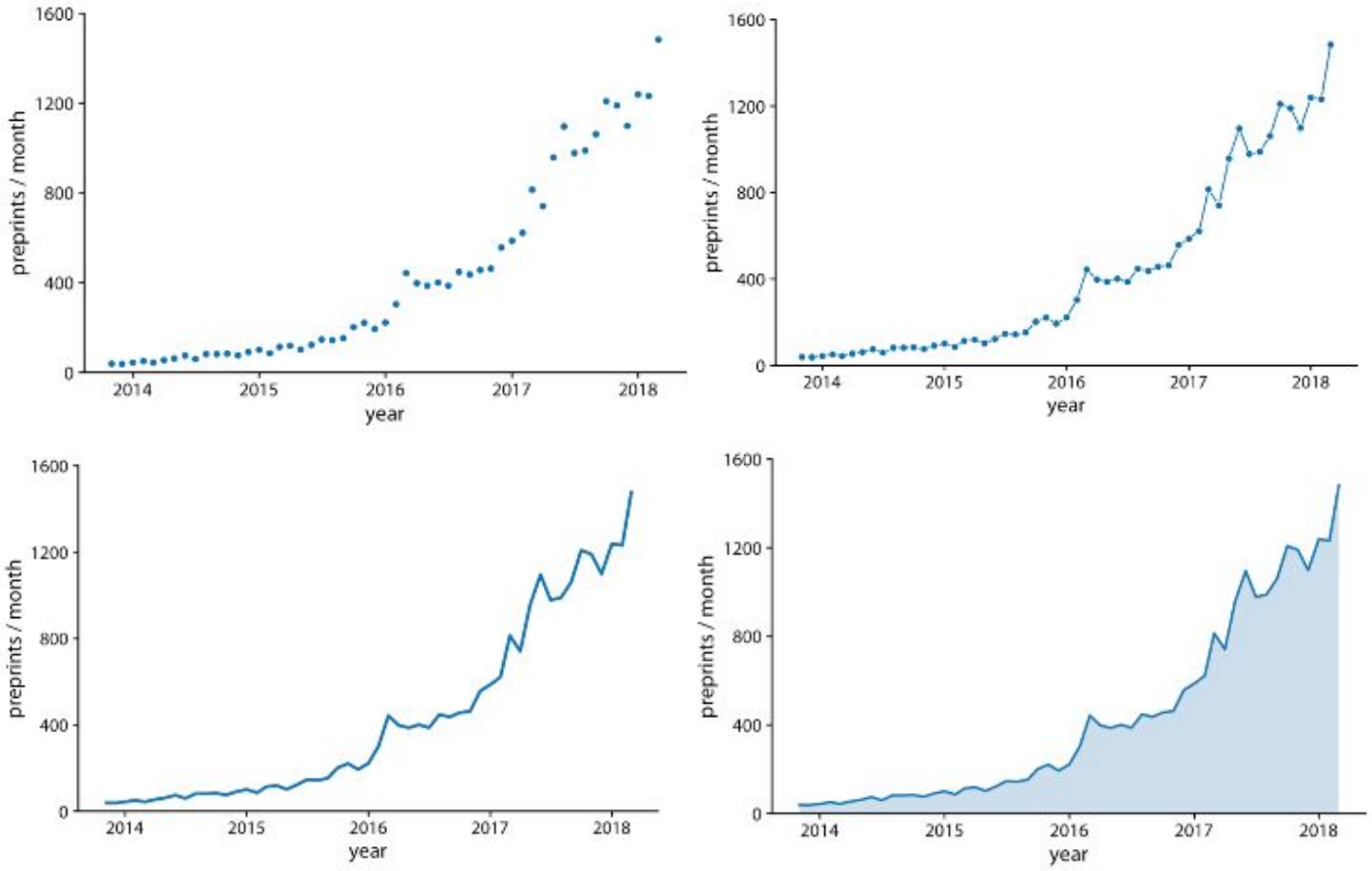






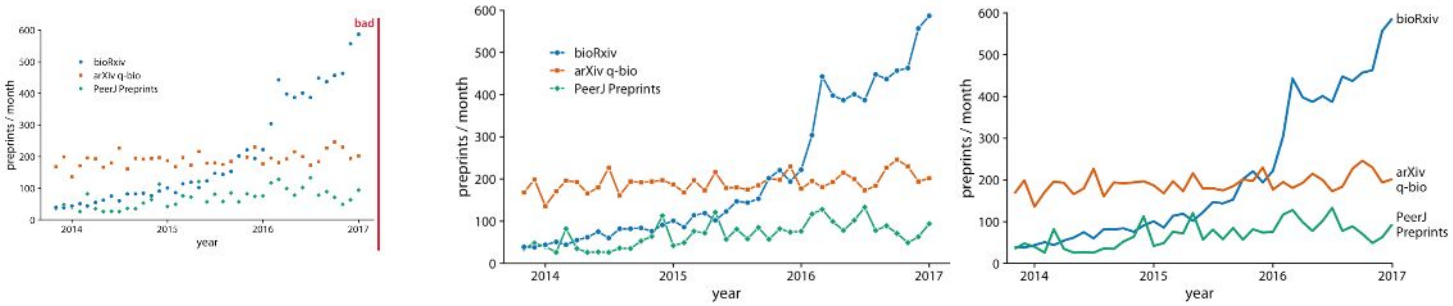
Visualizing Time Series

- One of the two variables can be thought of as time which imposes additional structure on the data.
- We can arrange the points in order of increasing time and define a predecessor and successor for each data point.



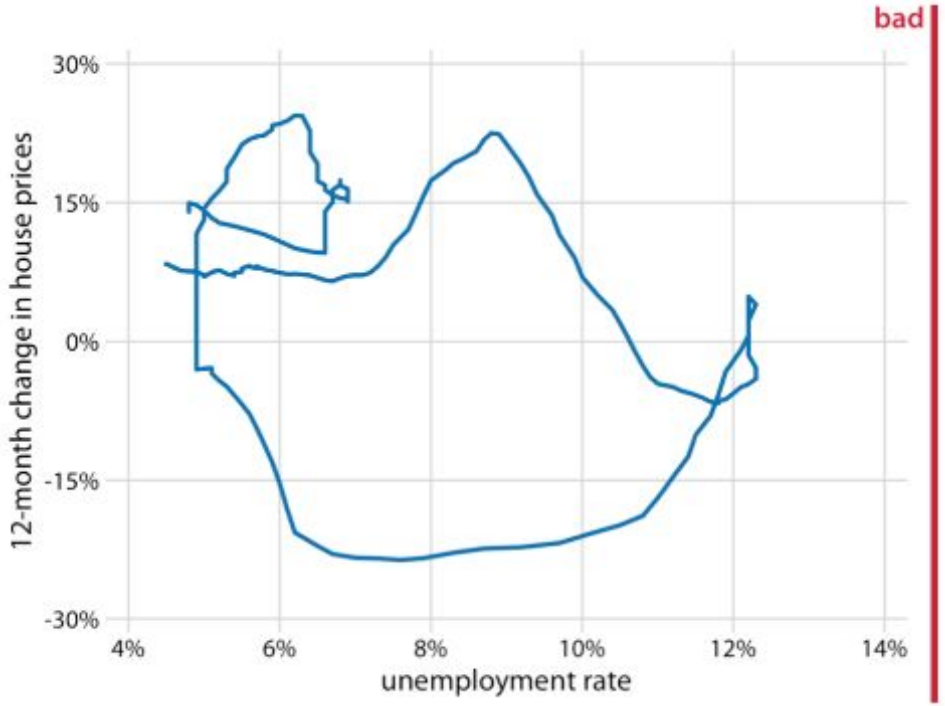
Visualizing Time Series

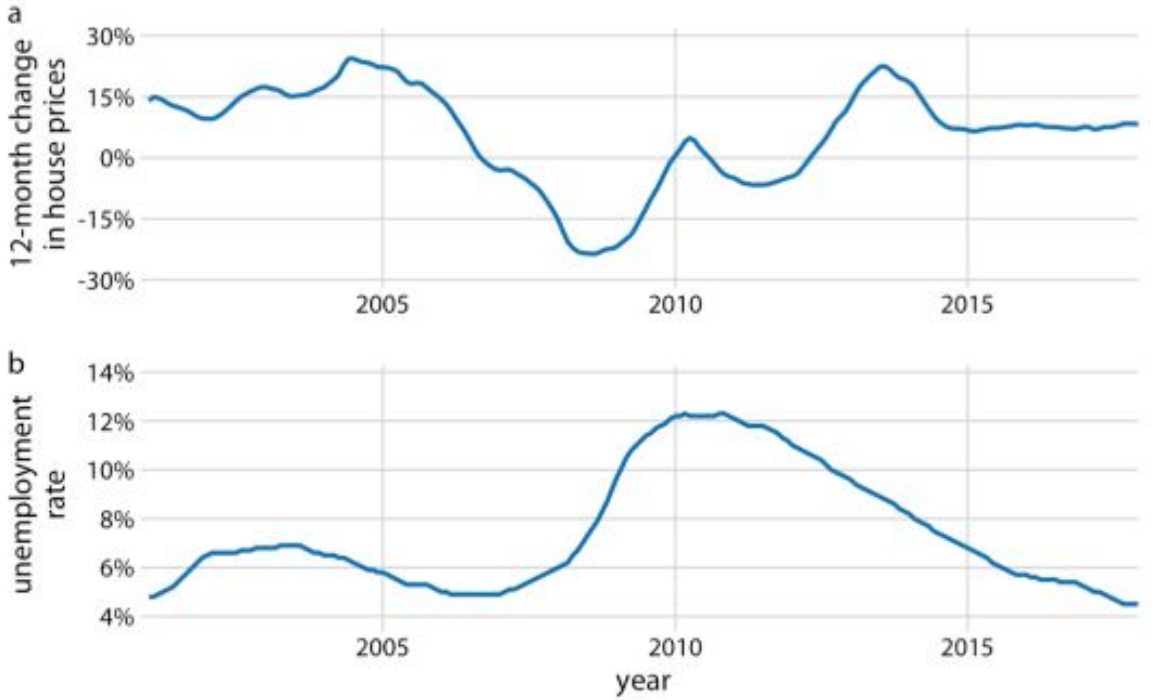
- We often have **multiple time curves** that we want to show at once. In this case, we have to be more careful in how we plot the data, because the figure can become con- fusing or difficult to read.



Visualizing Time Series

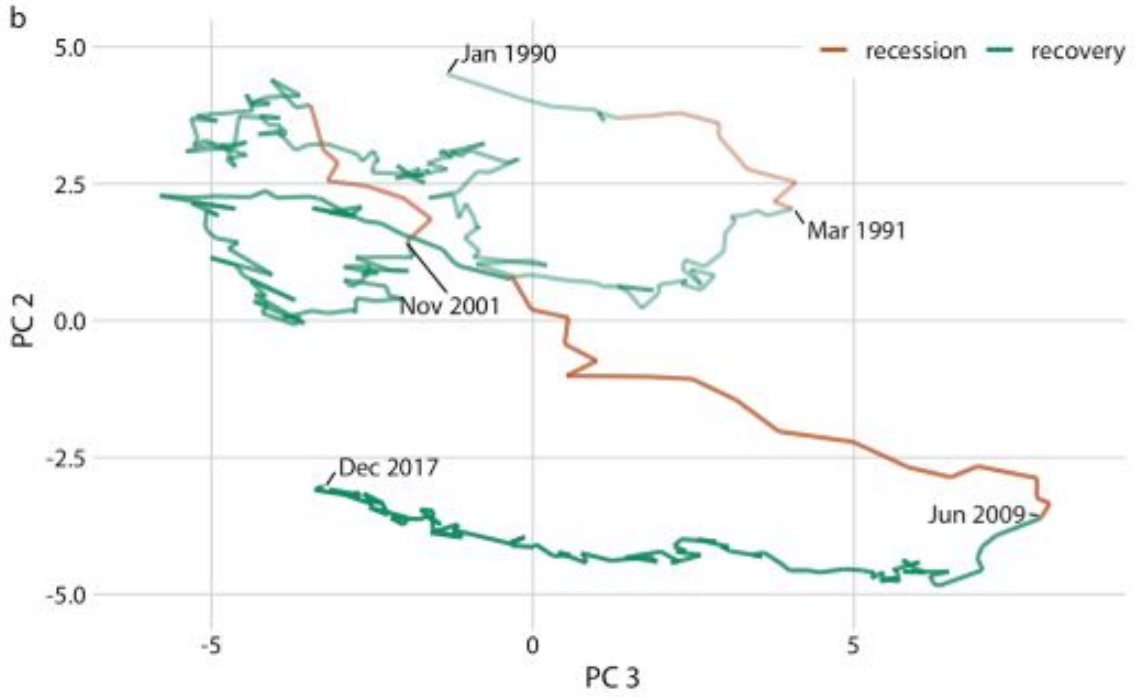
- It is not unusual, to have more than one response variable. Such situations arise commonly in macroeconomics.

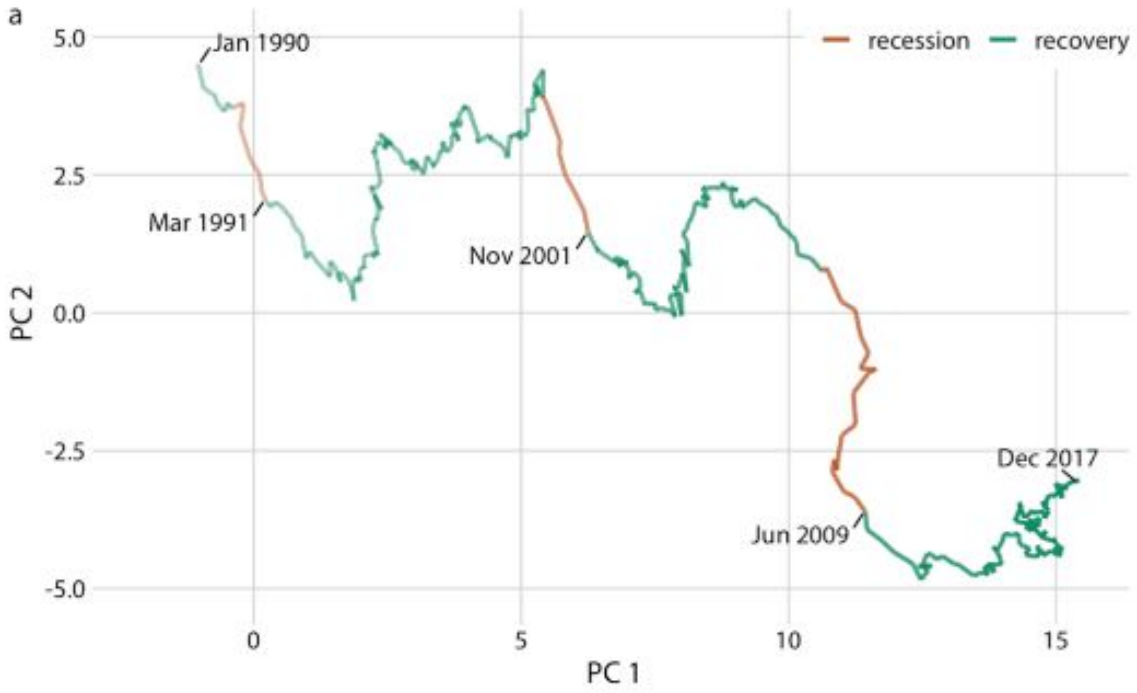




Visualizing Time Series

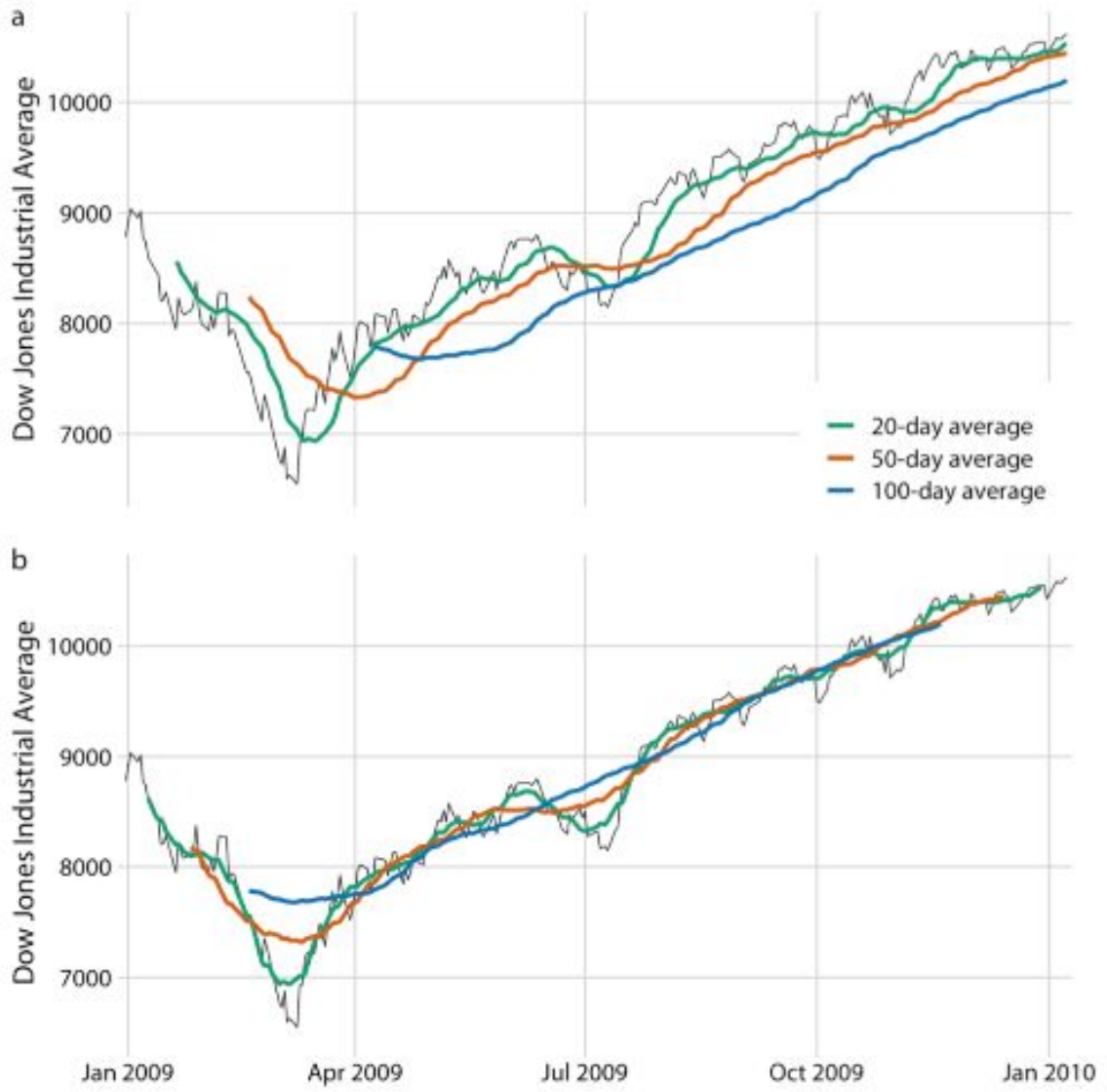
- The trick is to apply dimension reduction first then draw a connected scatterplot in the dimension-reduced space.





Visualizing Trends

- There are two fundamental approaches to determining a trend: **smoothing** (and **defined functional trend**.
- The **moving average** is the most simplistic approach to smoothing, and it has some obvious limitations.
- Statisticians developed numerous approaches to smoothing that alleviate the downsides of moving averages. One widely used method is **Locally Estimated Scatterplot Smoothing** (LOESS)

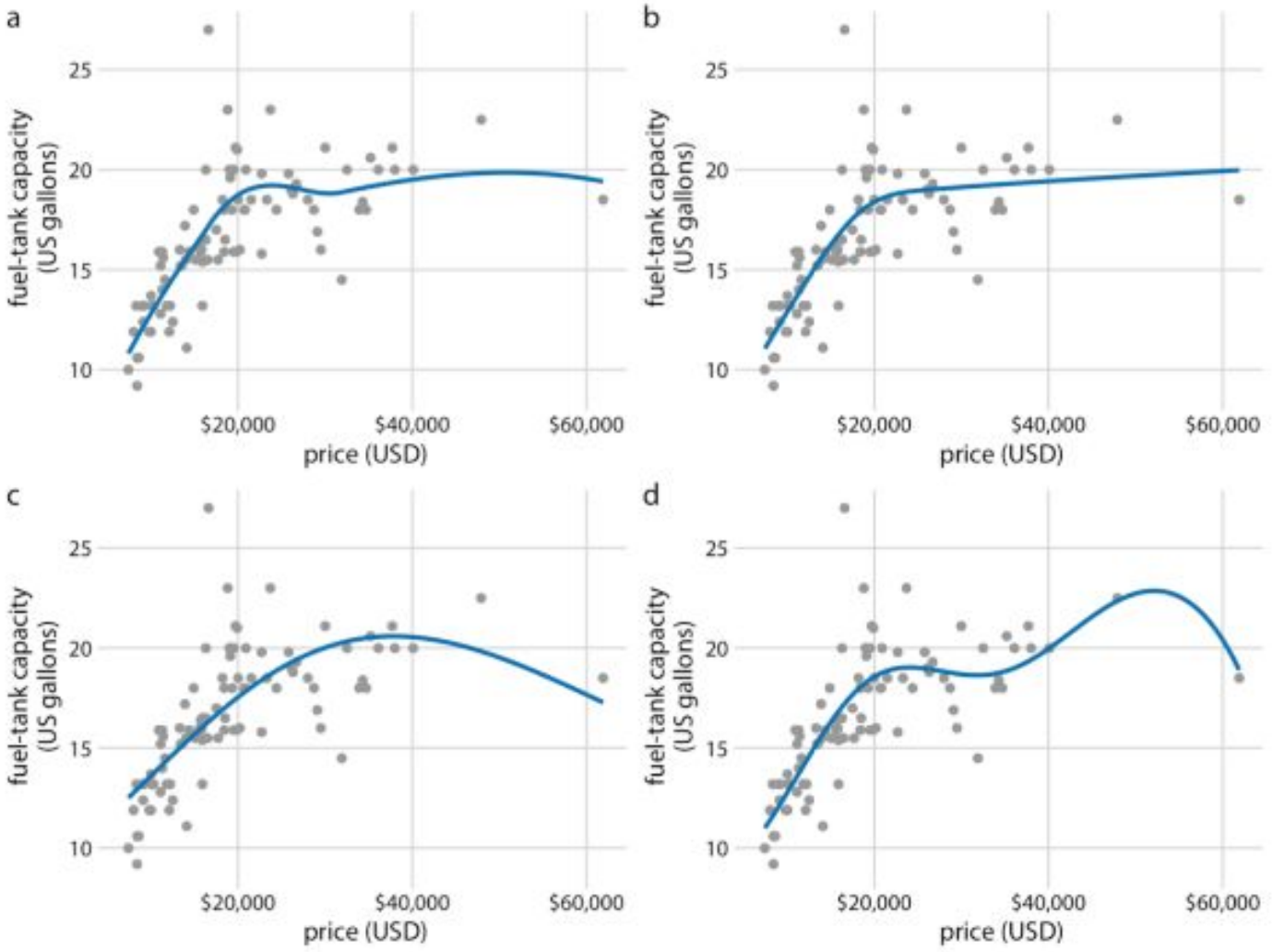


Visualizing Trends

- **LOESS** fits low-degree polynomials to subsets of the data. The points in the center of each subset are weighted more heavily than points at the boundaries, and this weighting scheme yields a much smoother result.
- **LOESS** is a popular smoothing approach because it tends to produce results that look right, however, it requires the fitting of many separate regression models.

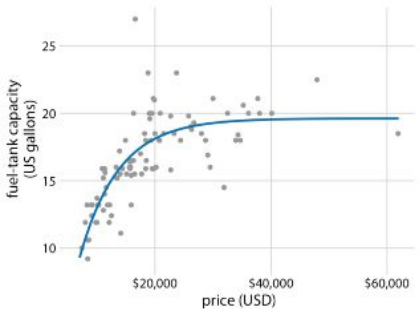
Visualizing Trends

- As a faster alternative is spline models. A **spline** is a piecewise polynomial function that is highly flexible yet always looks smooth.
- There is a bewildering array of different types of splines, including cubic splines, B-splines, thin-plate splines, Gaussian process splines, and many others, and which one to pick may not be obvious.
- **Most data visualization software will provide smoothing features.**

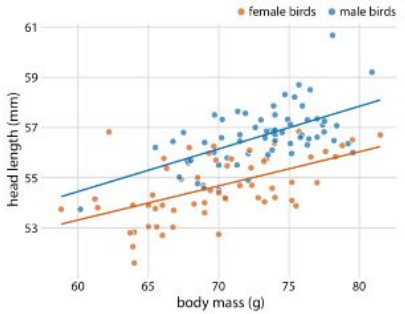


Visualizing Trends

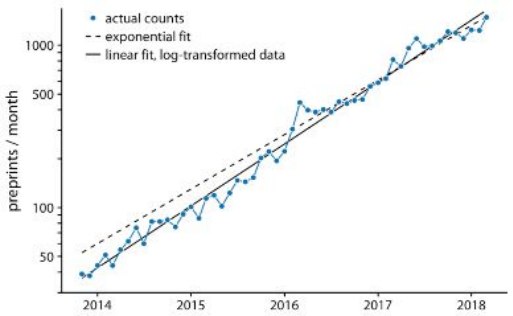
- These smoothers also do not provide parameter estimates that have a meaningful interpretation. Therefore, whenever possible, it is preferable to fit a curve with a **specific functional form** that is appropriate for the data and that uses parameters with clear meaning.



$y = A - B\,exp(-mx)$



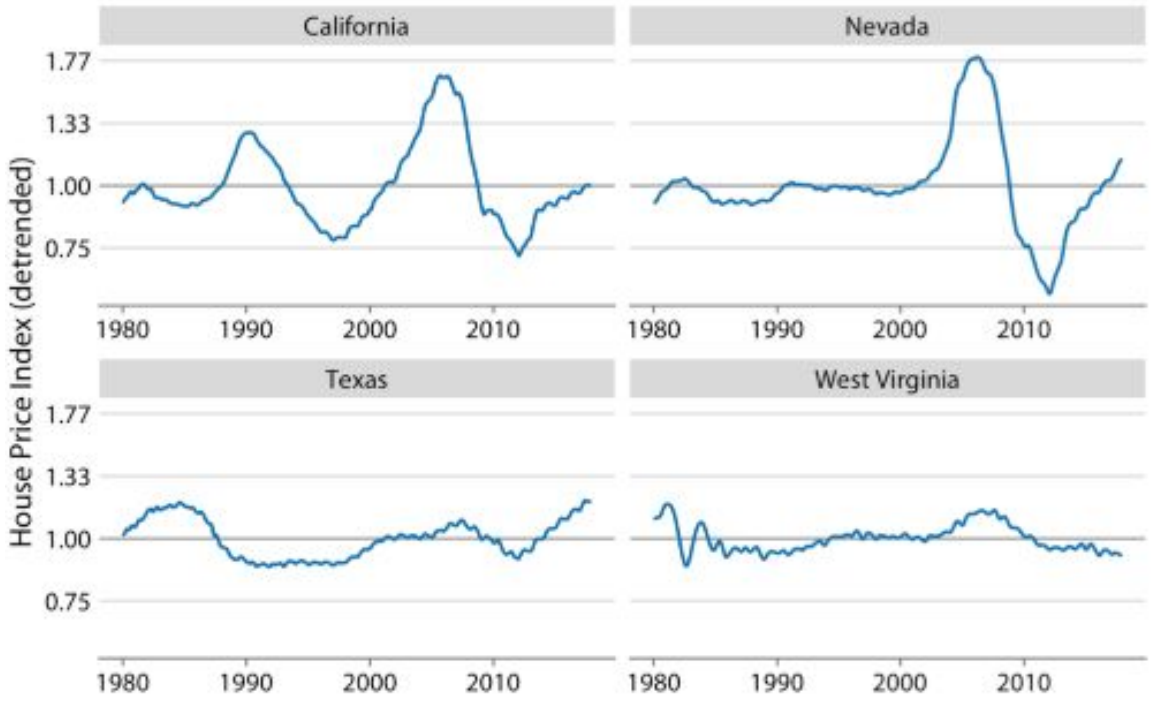
$y = A + mx.$

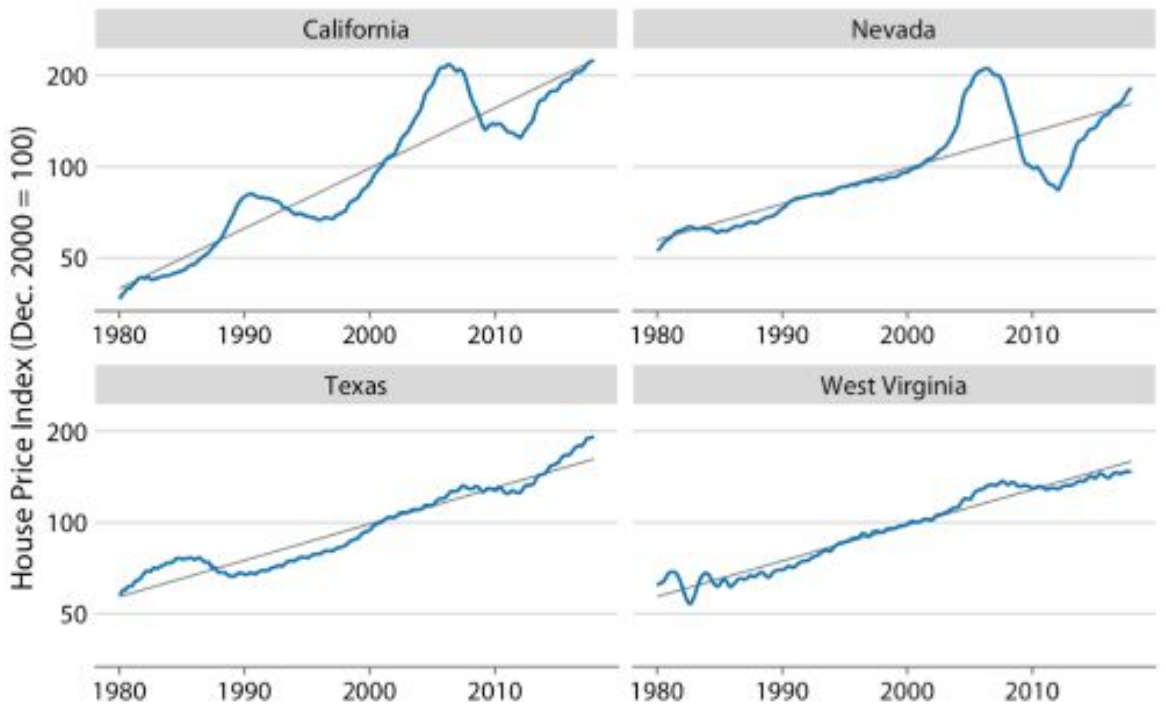


$y = 43\,exp[0.88(x - 2014)].$

Visualizing Trends

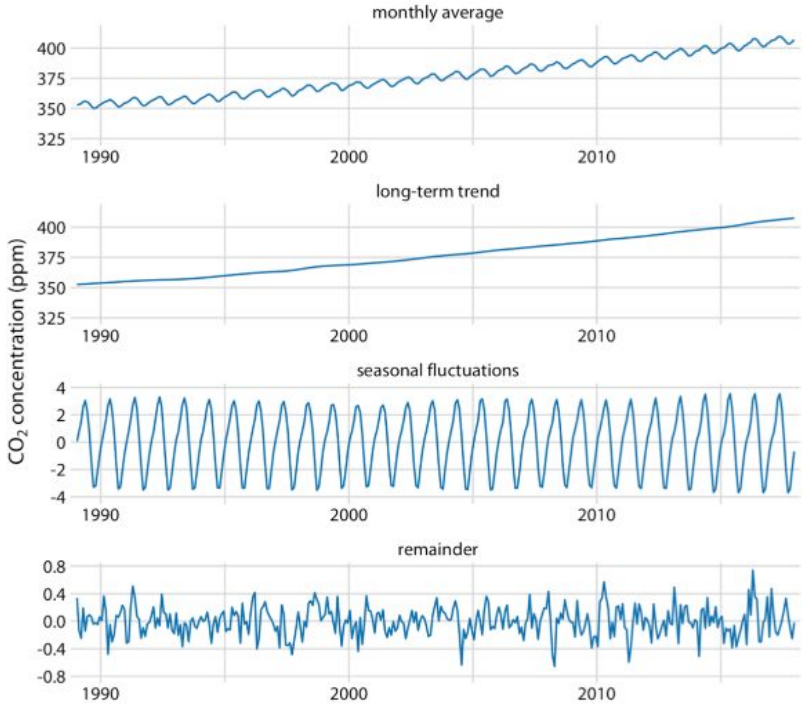
- For any time series with a prominent long-term trend, it may be useful to remove this trend to specifically highlight any notable deviations. This technique is called **detrending**.

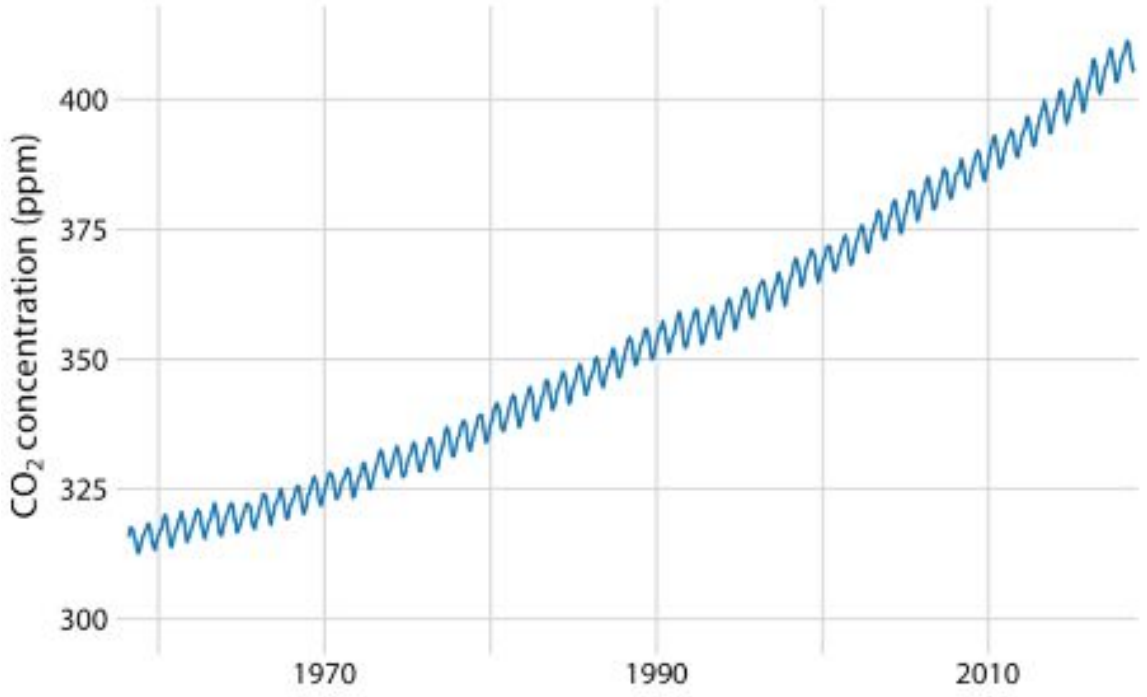




Visualizing Trends

- What we want to see in time series: **long term trend**, random **noise**, **external events**, **cyclical variations**. We called this **Time Series Decomposition**.

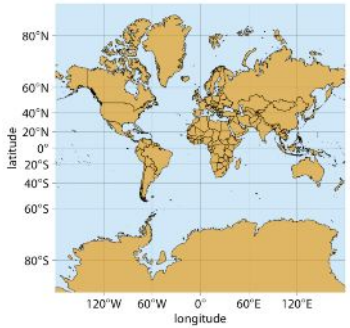




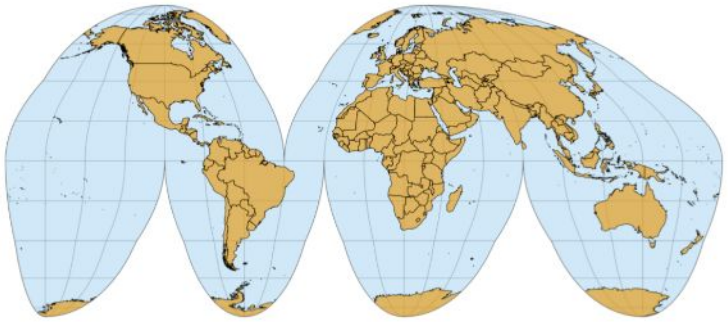
Visualizing Geospatial Data

- Maps tend to be intuitive to readers, but they can be challenging to design. A common mapping techniques, the ***choropleth map*** *and* ***cartograms***.

Orthographic projection



Mercator projection

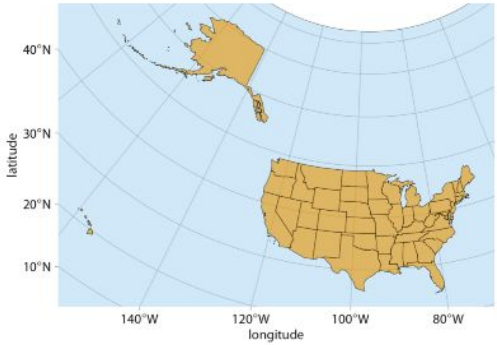


Interrupted Goode homolosine projection

Visualizing Geospatial Data

- Shape or area distortions due to map projections are particularly prominent when we're attempting to make a map of the whole world, but they can cause trouble even at the scale of individual continents or countries.

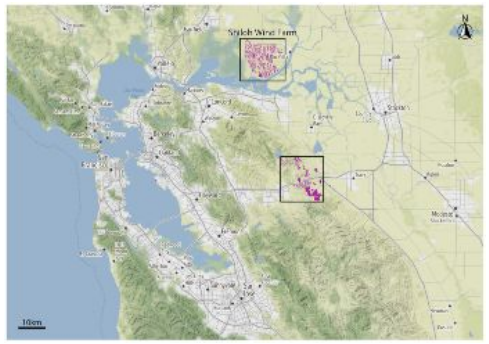
Orthographic projection



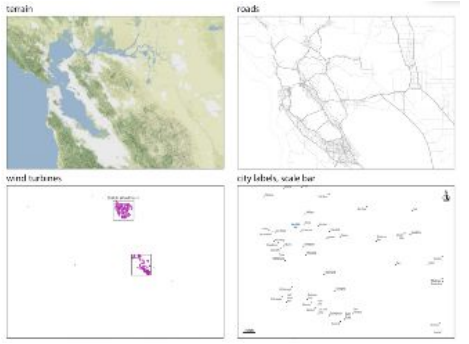
Albers projection

Visualizing Geospatial Data

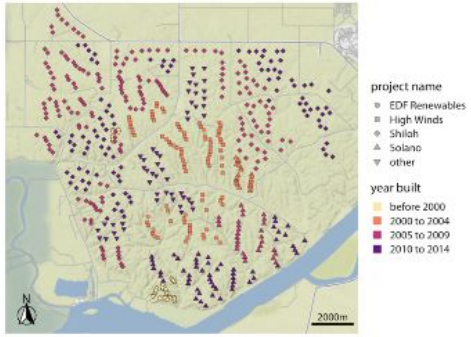
- To visualize geospatial data in the proper context, we usually create maps consisting of multiple layers showing different types of information.



Google map with 4 layers



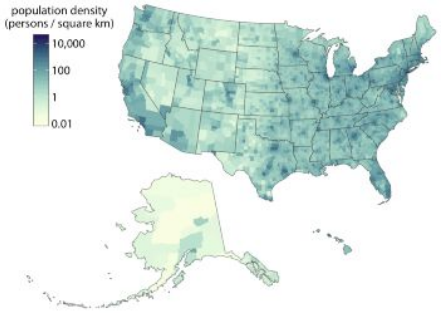
Individual layers

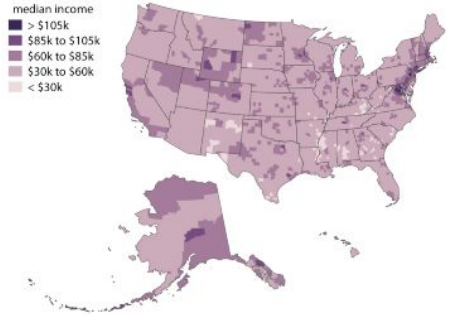


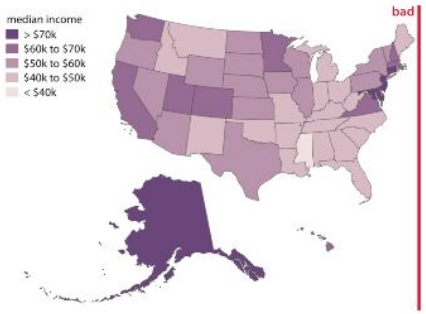
Detail map

Visualizing Geospatial Data

- We frequently want to show how some quantity varies across locations. We can do so by coloring individual regions in a map, we called it ***choropleth maps***.

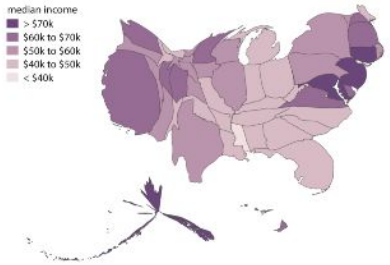


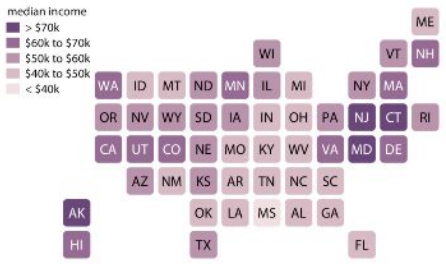


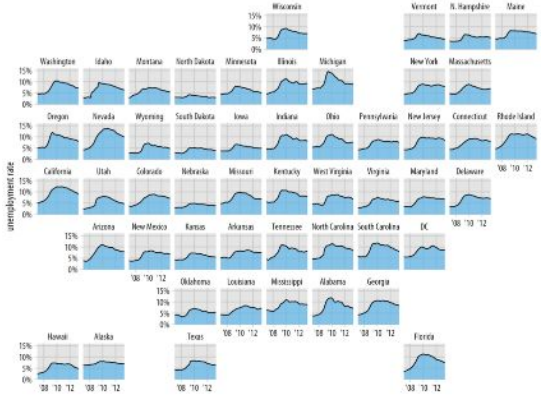


Visualizing Geospatial Data

- Not every map-like visualization has to be geographically accurate to be useful. What if we deformed the states so their size was proportional to their number of inhabitants? Such a modified map is called a ***cartogram.***

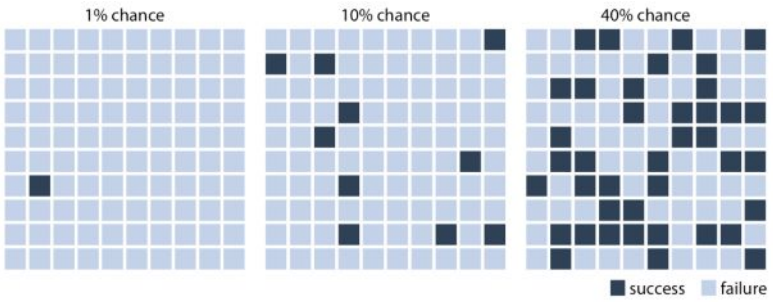




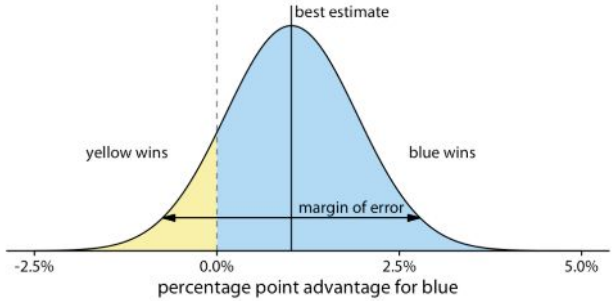


Visualizing Uncertainty

- One of the most challenging aspects of data visualization is the visualization of uncertainty like probability data. Two commonly used approaches to indicate uncertainty are **error bars** and **confidence bands**.



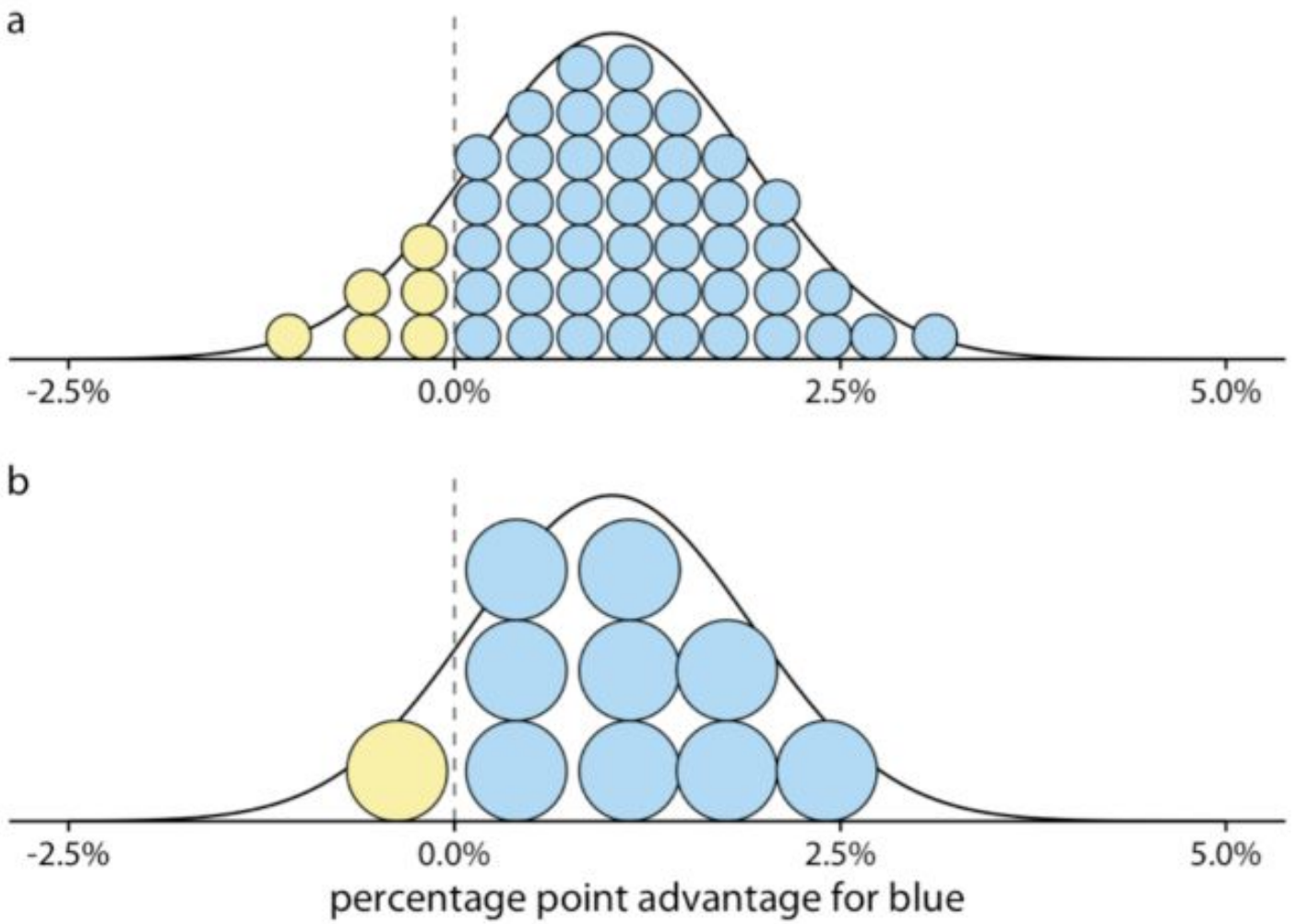
Visualizing probability as frequency



Visualizing probability as distribution

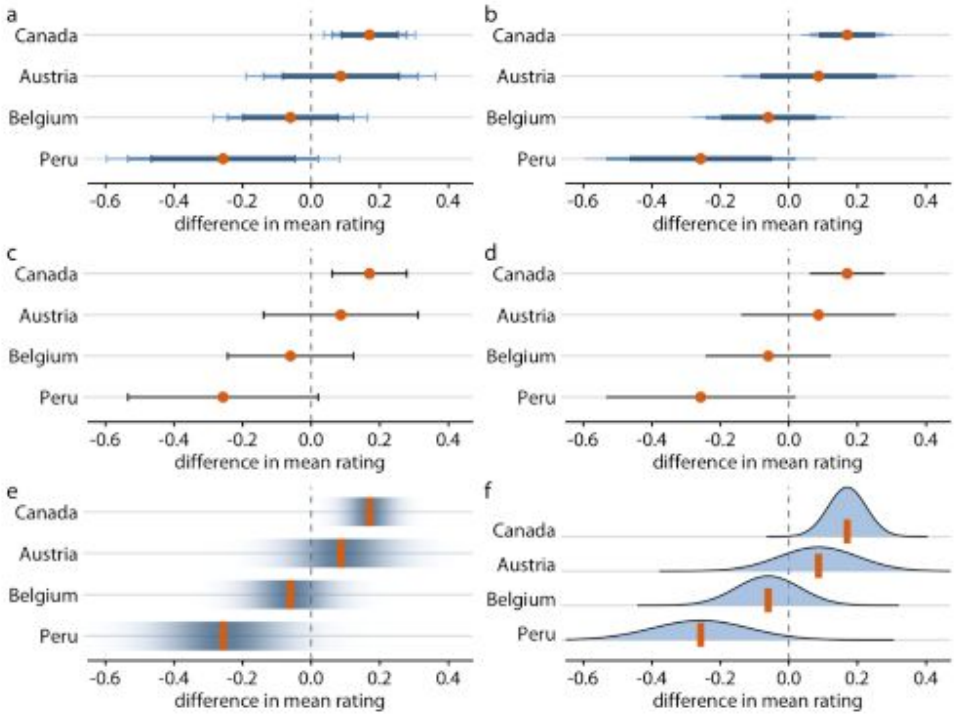
Visualizing Uncertainty

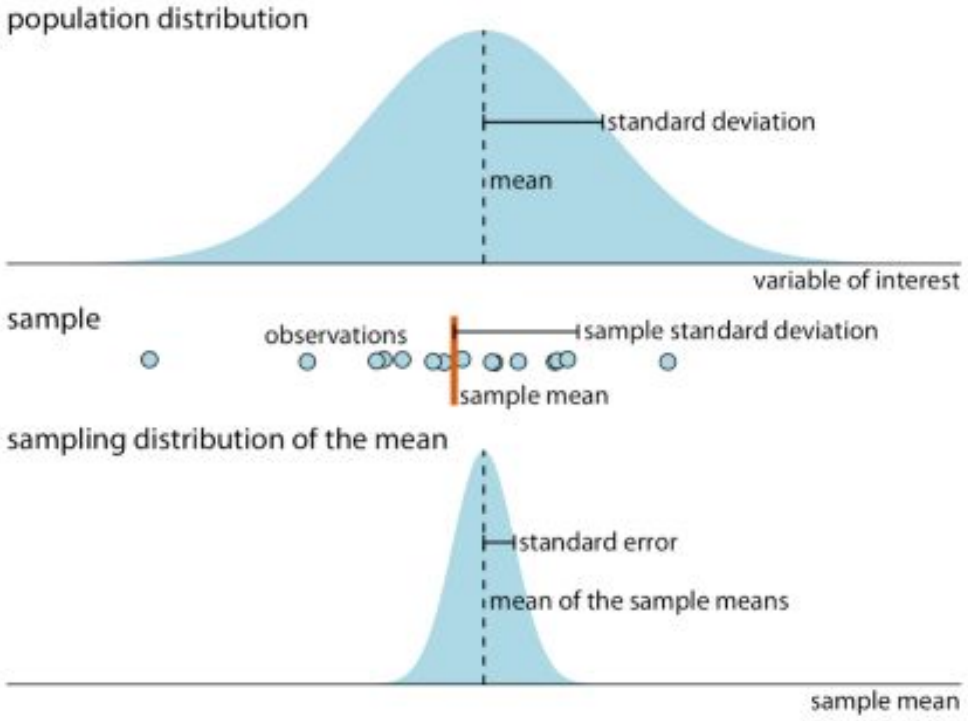
- We can combine the discrete outcome nature frequency with a continuous distribution using drawing a ***quantile dot plot***.
- In the quantile dot plot, we subdivide the total area under the curve into evenly sized units and draw each unit as a circle.
- As a general principle, quantile dot plots should use a small to moderate number of dots.



Visualizing Uncertainty

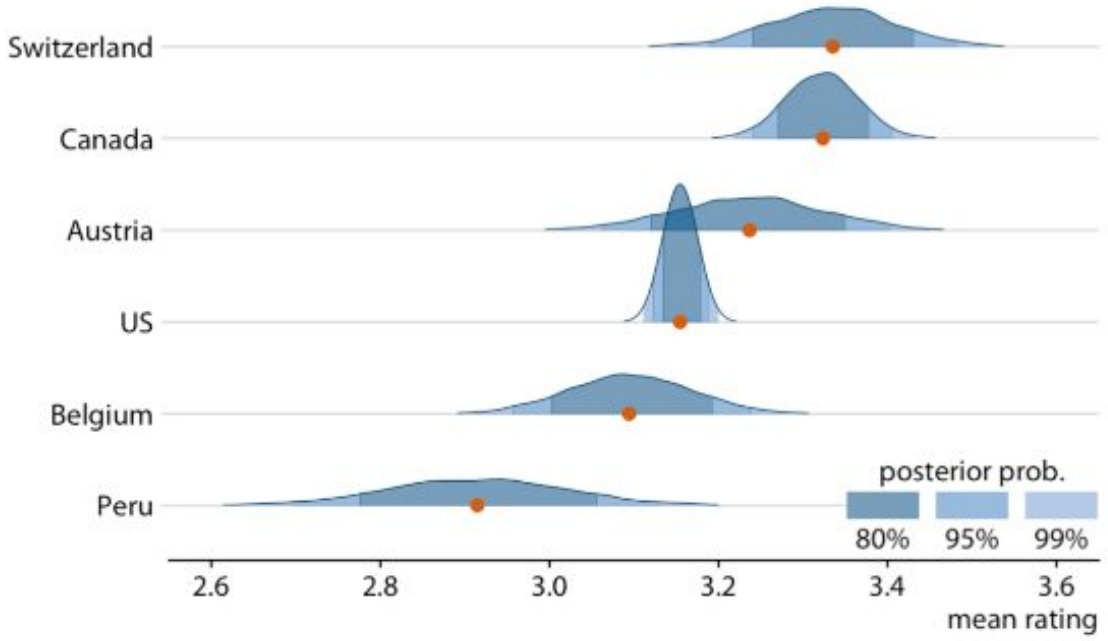
- **The concept of statistical sampling.** The variable of interest that we are studying has some true distribution in the population, with a true population mean and standard deviation.

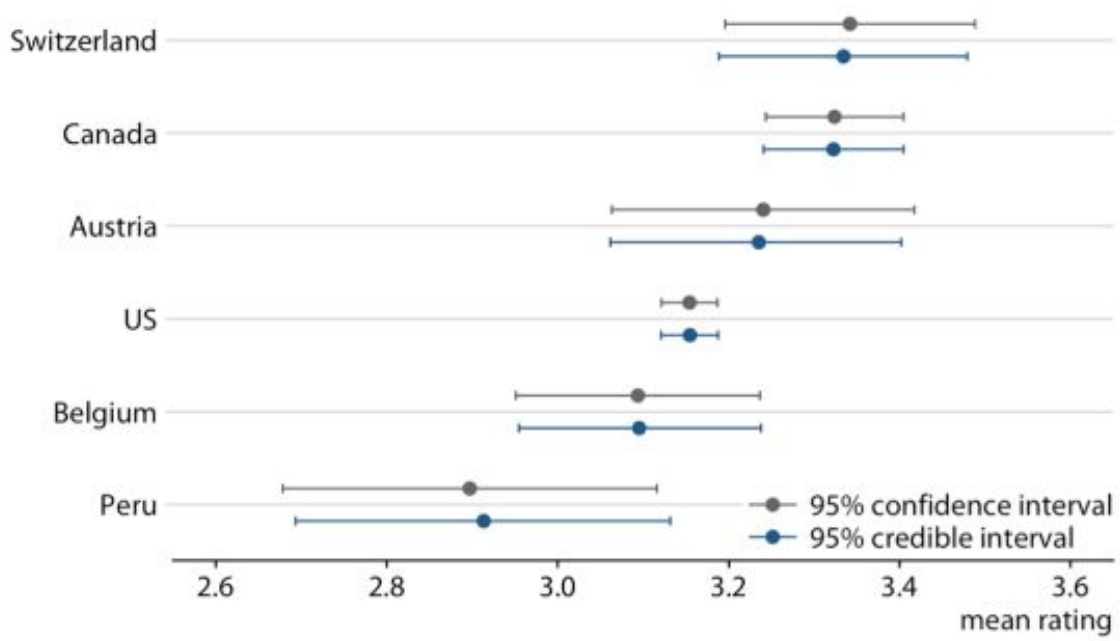




Visualizing Uncertainty

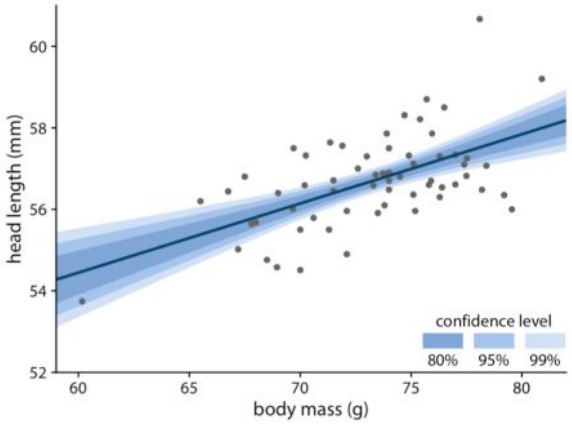
- There are two types of probability interpretations, **frequency** and **bayesian**. They use different styles to visualize probabilities.

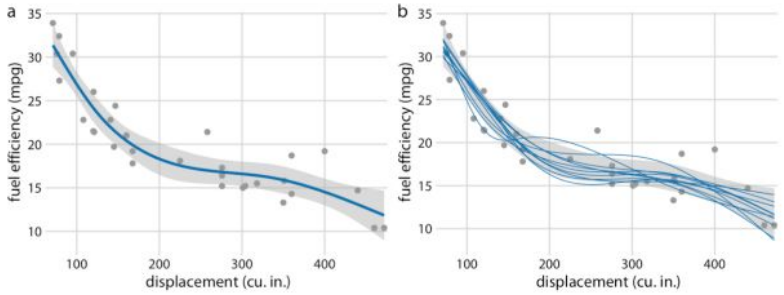




Visualizing Uncertainty

- In case of polynomial curve fitting, we can use **confidence band** to show uncertainty. It can be combined with curve fit.

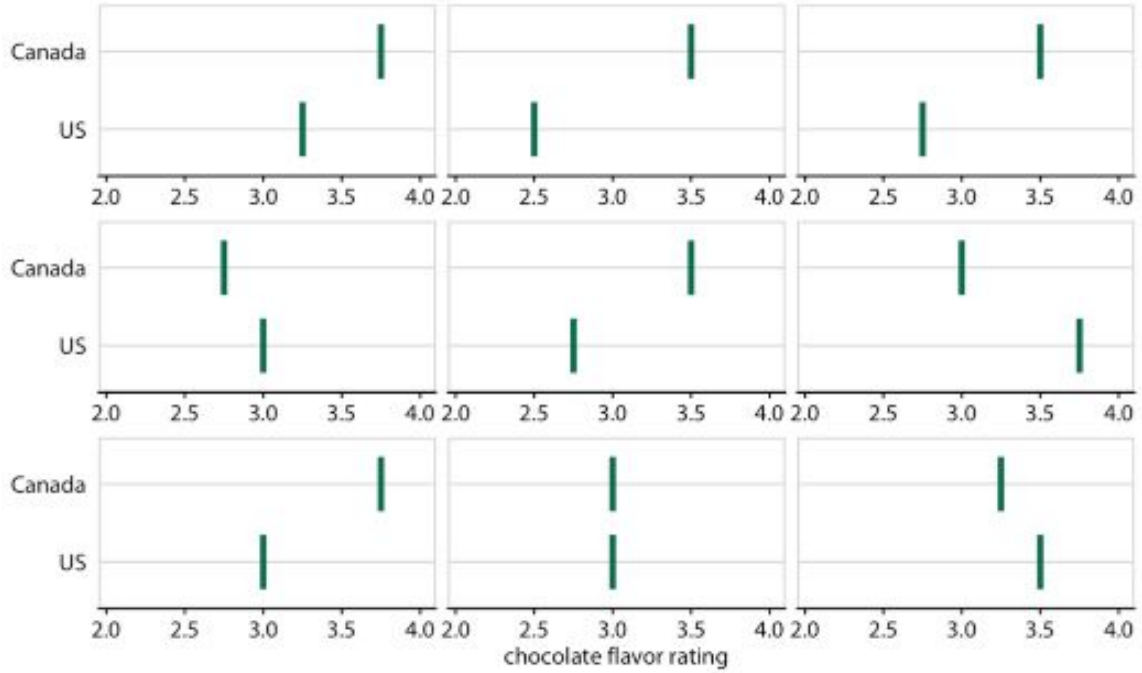


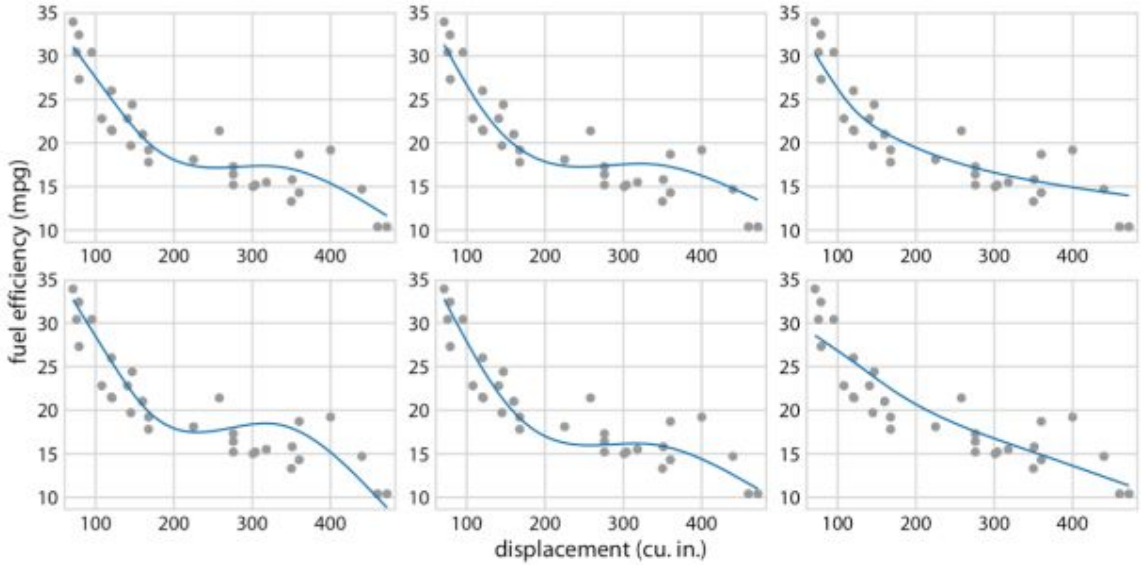


Best fit spline and confident band.

Visualizing Uncertainty

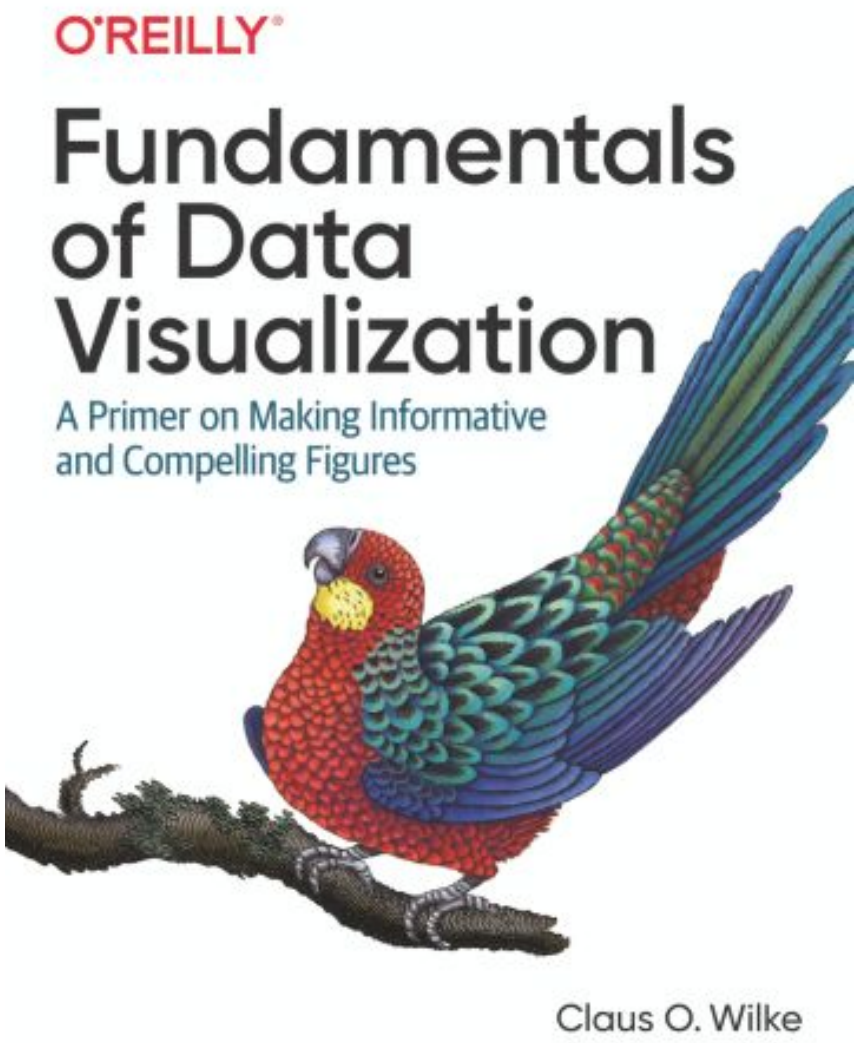
- Audiences may interpret uncertainty visualization as a deterministic feature. We can avoid this problem by visualizing uncertainty through animation, by cycling through a number of different but equally likely plots, this called a ***hypothetical outcome plot*** (HOP).



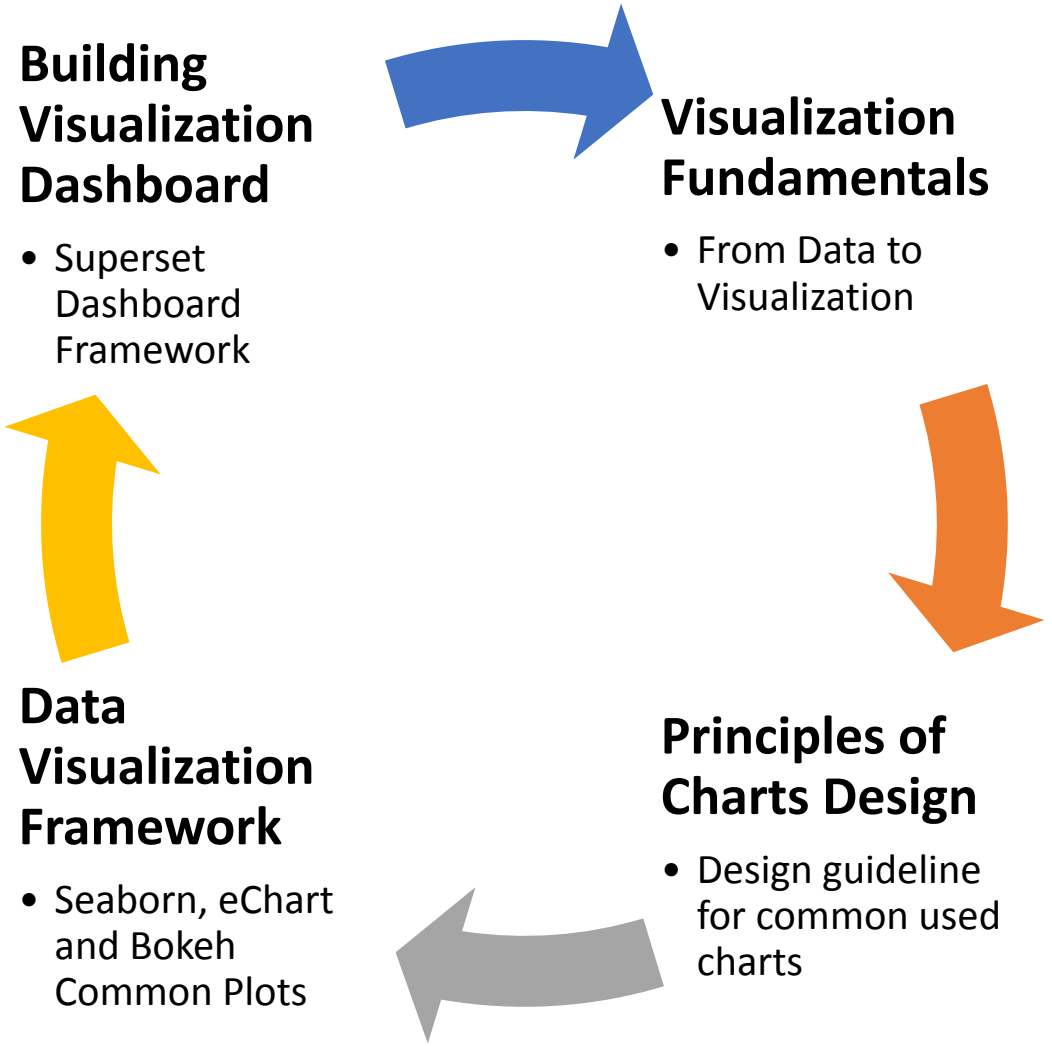


Some Key Principles

1. Proportional Ink
2. Handling Overlapping Points
3. Common Pitfalls of Color Use
4. Redundant Coding
5. Multipanel Figures
6. Title, Captions and Tables
7. Balance the Data and the Context
8. Use Larger Axis Labels
9. Avoid Line Drawing
10. Don’t go 3D



Outline – Data Visualization



Seaborn Framework

- Data visualization library based on matplotlib, closely integrated with Pandas.
- Here is some of the functionality that seaborn offers:
  - Dataset-oriented API for examining relationships between multiple variables
  - Support categorical variables to show observations or aggregate statistics
  - Options for visualizing univariate or bivariate distributions
  - Automatic estimation and plotting of linear regression models
  - Convenient views onto the overall structure of complex datasets
  - High-level abstractions for structuring multi-plot grids
  - Concise control over matplotlib figure styling with several built-in themes
  - Tools for choosing color palettes that faithfully reveal patterns in your data

How to Learn Seaborn Framework

Plotting functions

- Visualizing statistical relationships
  - Relating variables with scatter plots
  - Emphasizing continuity with line plots
  - Showing multiple relationships with facets
- Plotting with categorical data
  - Categorical scatterplots
  - Distributions of observations within categories
  - Statistical estimation within categories
  - Plotting "wide-form" data
  - Showing multiple relationships with facets
- Visualizing the distribution of a dataset
  - Plotting univariate distributions
  - Plotting bivariate distributions
  - Visualizing pairwise relationships in a dataset
- Visualizing linear relationships
  - Functions to draw linear regression models
  - Fitting different kinds of models
  - Conditioning on other variables
  - Controlling the size and shape of the plot
  - Plotting a regression in other contexts

Multi-plot grids

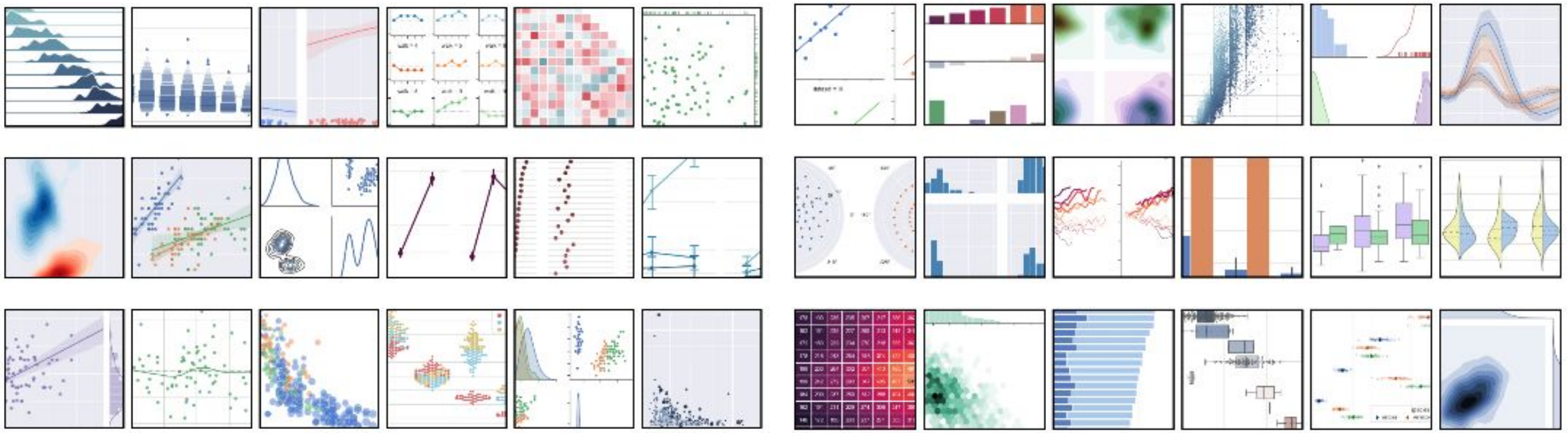
- Building structured multi-plot grids
  - Conditional small multiples
  - Using custom functions
  - Plotting pairwise data relationships

Plot aesthetics

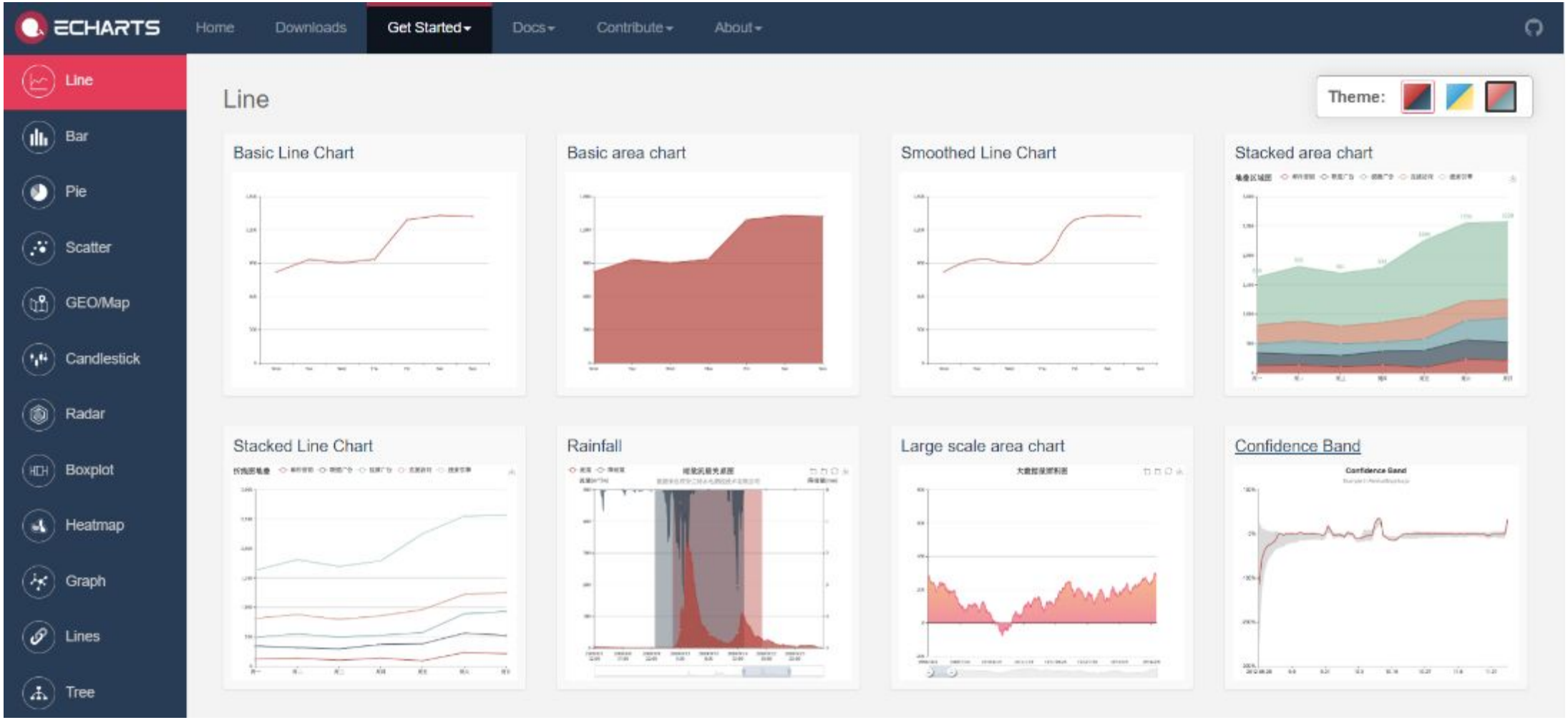
- Controlling figure aesthetics
  - Seaborn figure styles
  - Removing axes spines
  - Temporarily setting figure style
  - Overriding elements of the seaborn styles
  - Scaling plot elements
- Choosing color palettes
  - Building color palettes
  - Qualitative color palettes
  - Sequential color palettes
  - Diverging color palettes
  - Setting the default color palette

Common Examples

- Please see gallery: https://seaborn.pydata.org/examples/index.html



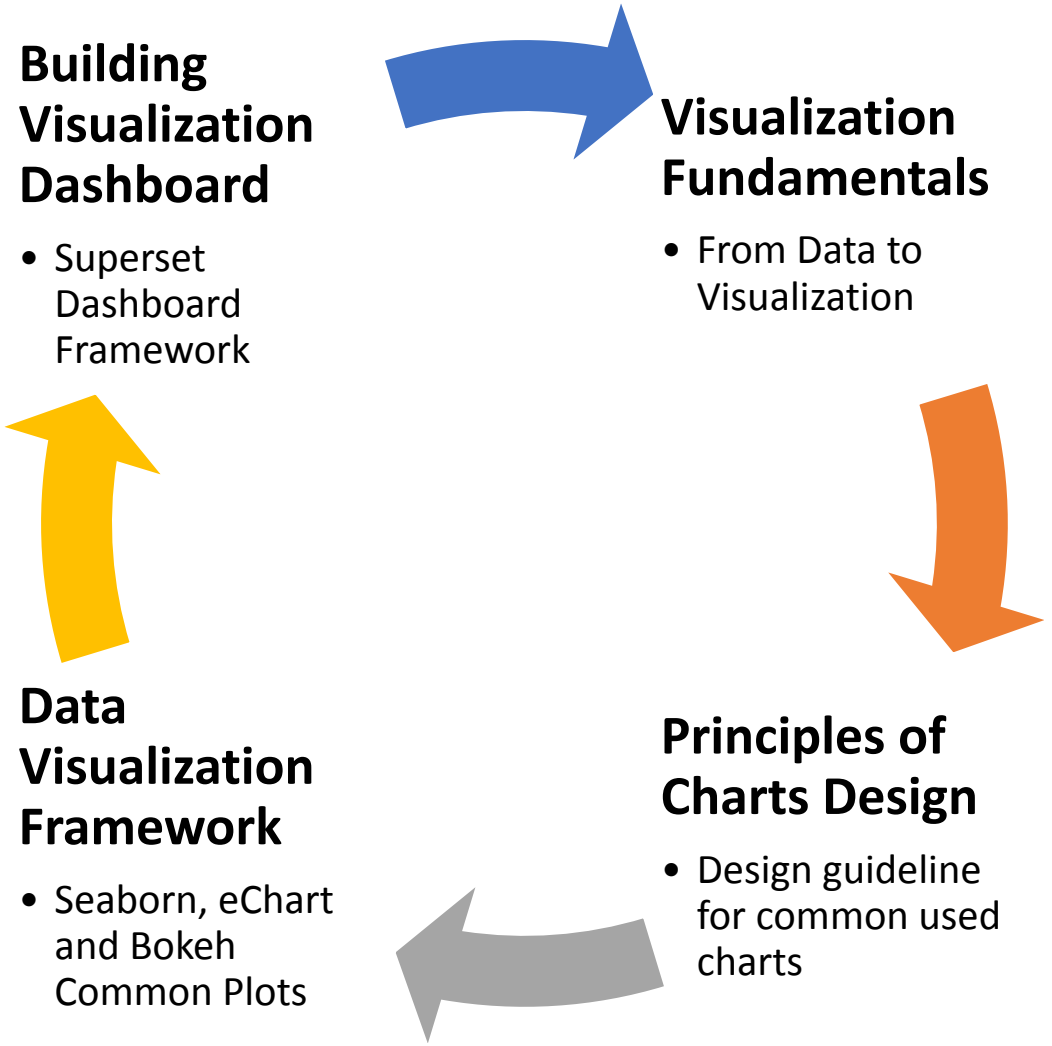
eChart Framework



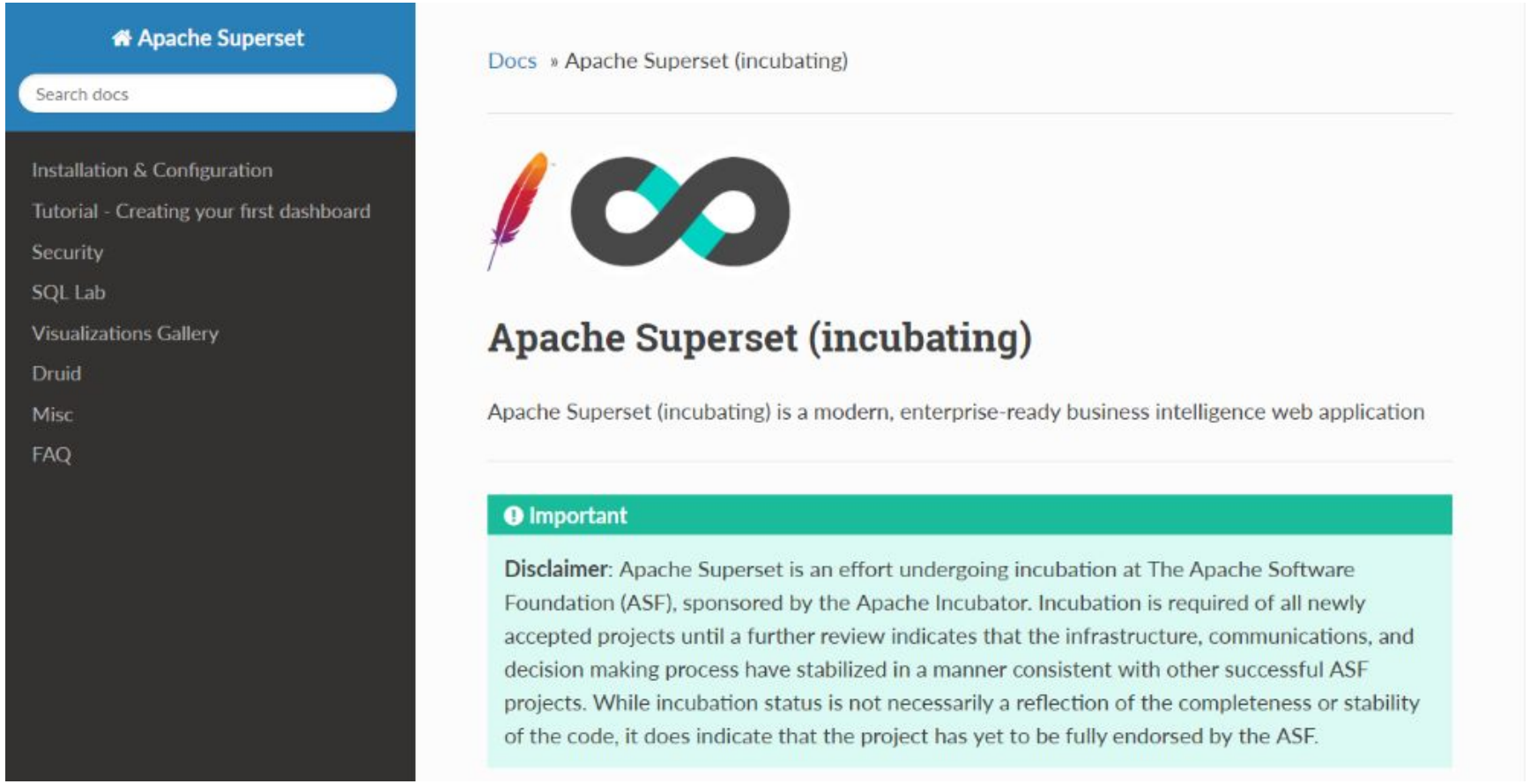
Bokeh Framework

- Bokeh is an interactive visualization library, its goal is to provide elegant, concise construction of versatile graphics, and to extend this capability with high-performance interactivity over very large or streaming datasets.

Outline – Data Visualization



Introduction to Superset



Please work on your homework

THANKS