Contents

- **Unit V: Data Visualization**

- Understanding Data Visualization Principles
- Mapping Data onto Aesthetics
- Visualizing - Distributions, Proportions, Time Series, Trends and Uncertainty;
- Commonly used File Formats and Software.

Reference Book:
Fundamentals of Data Visualization by *Claus O. Wilke, O'Reilly.*

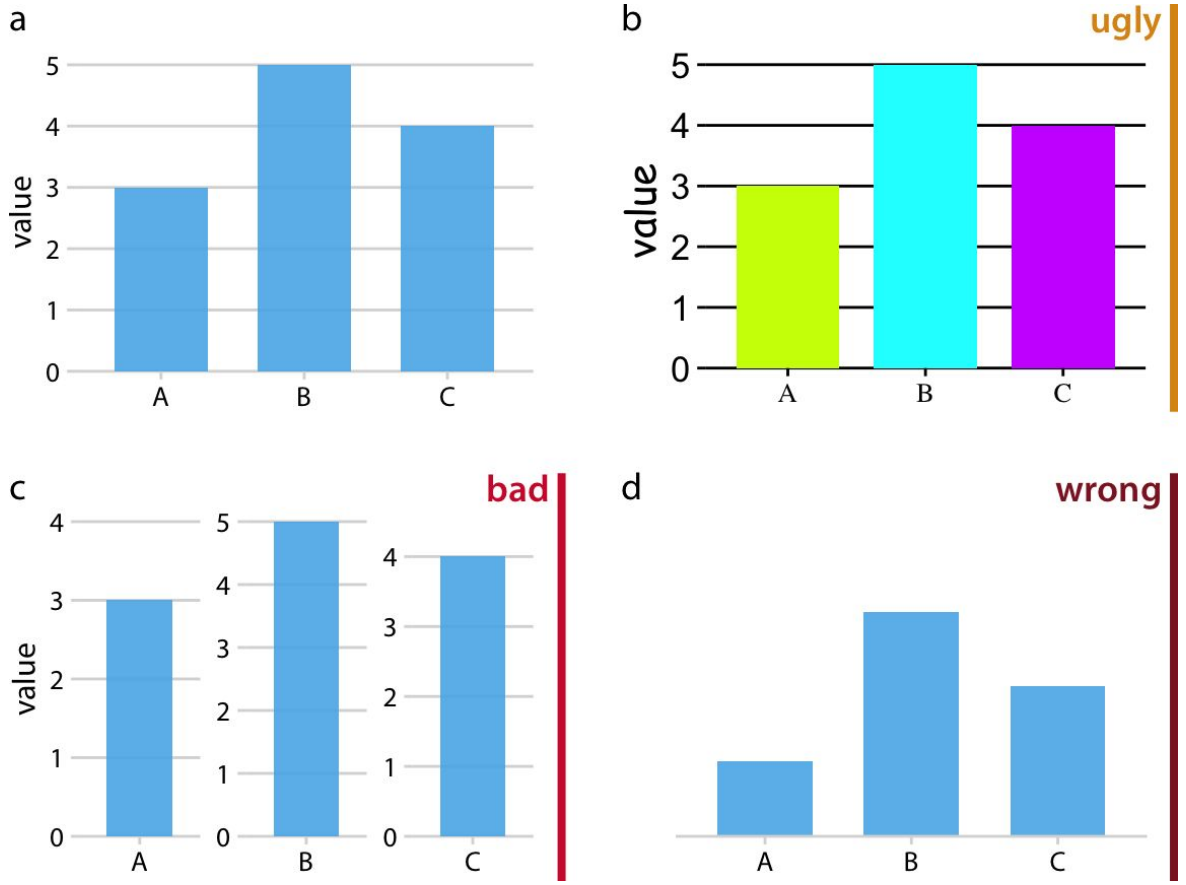
Data Visualization Principles

The five C's of data apply that apply all forms of data.

1. **Clean:-** Clean data means data that has no missing values, no inaccurate data, no out of range data, no typos, etc.
2. **Consistent:-**Consistent data means that the data is the same no matter where it appears in your organization and the definition(s) associated with your data are consistent across your organization**.**
3. **Conformed:-**Conformed data is data that fits within established boundaries. For example, if you have data on students, does it apply to all students, full-time equivalents, daily attendance, or something else?
4. **Current:-** When it comes to data, current can be relative. Stock market data, for example, needs to be real-time if you are using it to trade stocks. On the other hand, if you are doing historical price analysis, then you want data that goes back for months or years.
5. **Comprehensive:-** Comprehensive data is data that spans all the required dimensions of your business case(s); the breadth of your data.

Introduction

- A data visualization first and foremost has to accurately convey the data.
- A data visualization should be aesthetically pleasing.
- Problematic figure are called as “ugly”, “bad”, or “wrong”
  - **ugly**—A figure that has **aesthetic problems** but otherwise is clear and informative.
  - **bad**—A figure that has **problems related to perception**; it may be unclear, confusing, overly complicated, or deceiving.
  - **wrong**—A figure that **has problems related to mathematics**; it is objectively incorrect.



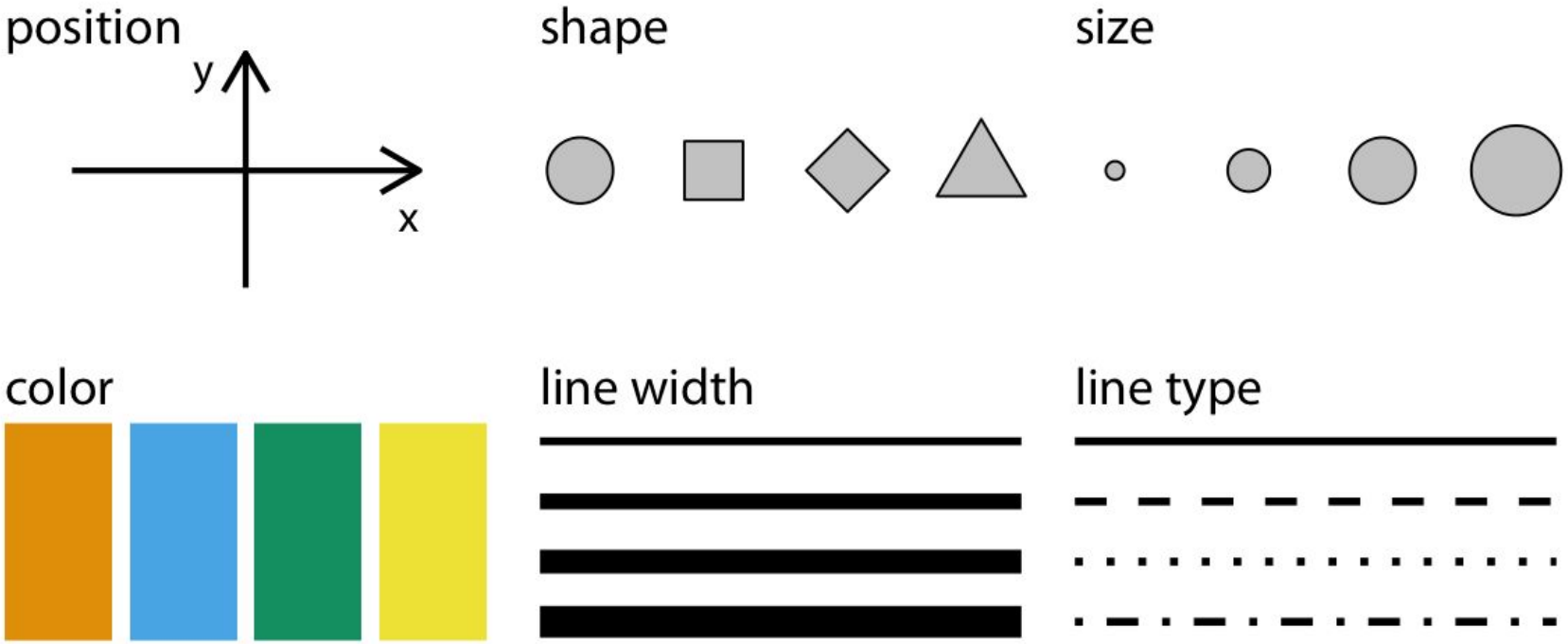
Mapping Data onto Aesthetics

- We take data values and convert them in a systematic and logical way into the visual elements that make up the final graphic.

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

Aesthetics and types of data

- Aesthetics describe every aspect of a given graphical element



- if we want to display **text**, we may have to specify font family, font face, and font size
- if graphical objects overlap, we may have to specify whether they are partially transparent.

Aesthetics and types of data

- **TYPES:**

- Those that can represent continuous data and those that can not.
- **Continuous** data values are values for which arbitrarily fine intermediates exist.
  - For example, time duration is a continuous value.
  - Between any two durations, say 50 seconds and 51 seconds, there are arbitrarily many intermediates, such as 50.5 seconds, 50.51 seconds, 50.50001 seconds, and so on.
- **Discrete**: number of persons in a room is a value. A room can hold 5 persons or 6, but not 5.5.

Types of variables encountered in data visualization scenarios

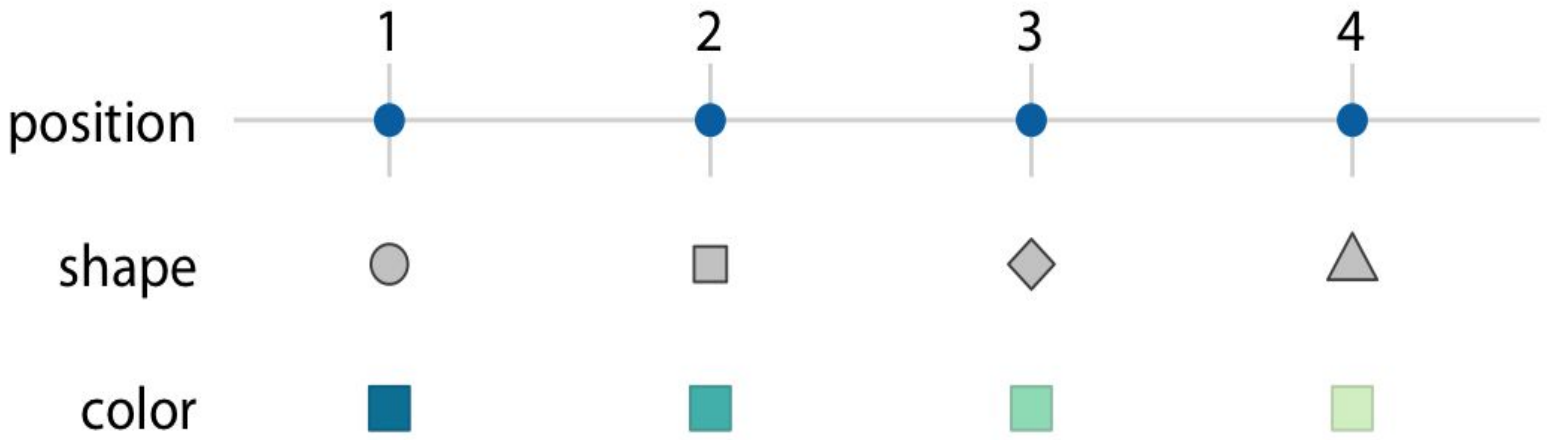
Type of variable	Examples	Appropriate scale	Description
quantitative/numerical continuous	1.3, 5.7, 83, $1.5x10^{-2}$	continuous	Arbitrary numerical values. These can be integers, rational numbers, or real numbers.
quantitative/numerical discrete	1, 2, 3, 4	discrete	Numbers in discrete units. These are most commonly but not necessarily integers. For example, the numbers 0.5, 1.0, 1.5 could also be treated as discrete if intermediate values cannot exist in the given dataset.
qualitative/categorical unordered	dog, cat, fish	discrete	Categories without order. These are discrete and unique categories that have no inherent order. These variables are also called *factors*.

Types of variables encountered in data visualization scenarios

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.

Scales map data values onto aesthetics

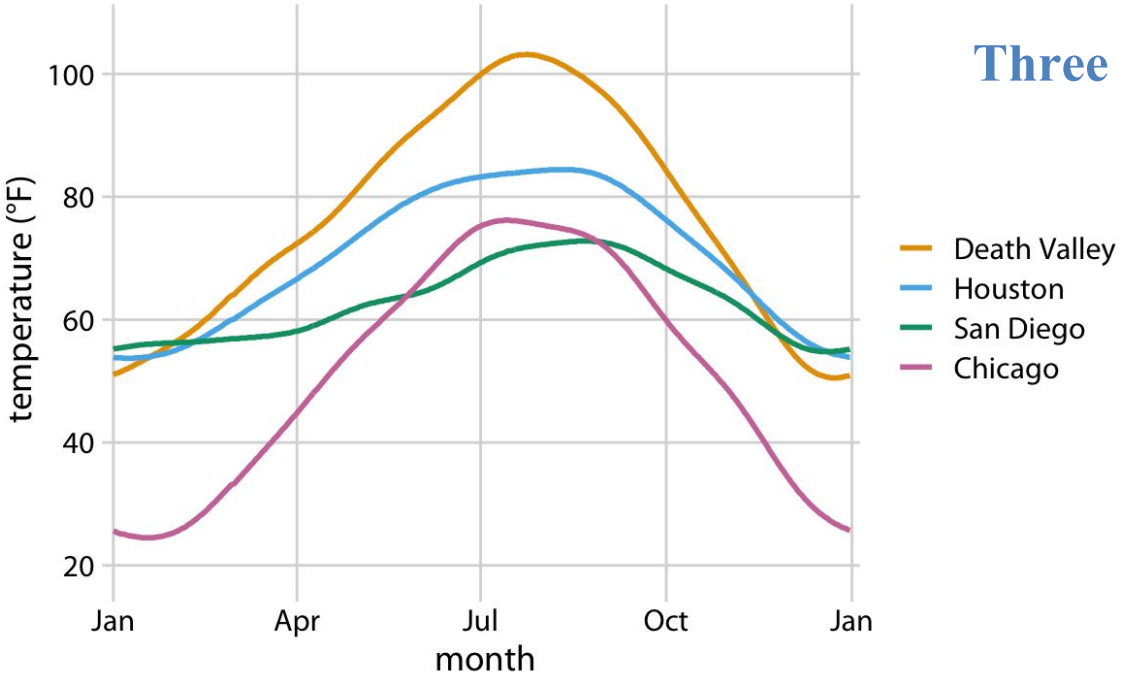
- To map data values onto aesthetics, we need to specify **which data values correspond to which specific aesthetics values**.
- For example, if our graphic has an *x* axis, then we need to specify which data values fall onto particular positions along this axis.
- Similarly, we may need to specify which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



Example Data

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

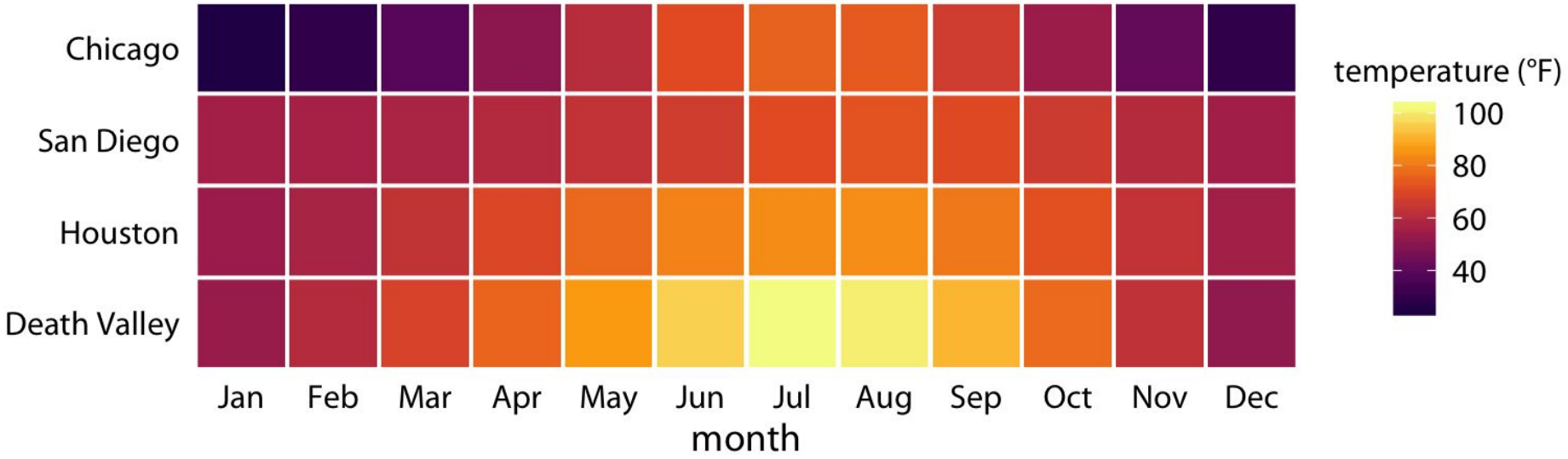
Example Data

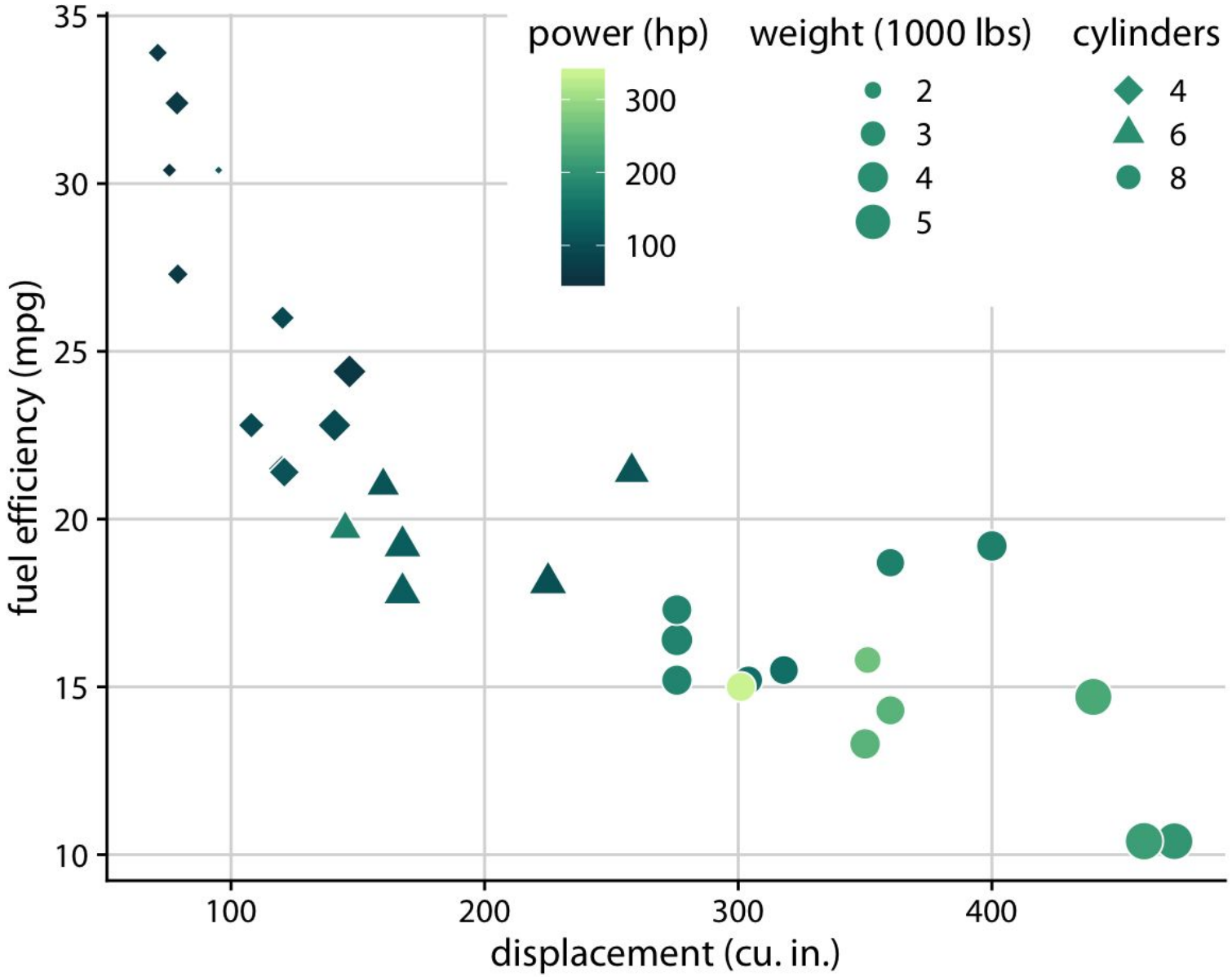


Three scales in total, two position scales and one color scale

two position scales (month along the *x* axis and location along the *y* axis) but neither is a continuous scale

Month is an ordered factor with 12 levels and location is an unordered factor with four levels. Therefore, the two position scales are both discrete. For discrete position scales, we generally place the different levels of the factor at an equal spacing along the axis.





five scales:
two position scales
one color scale
one size scale, and
one shape scale,

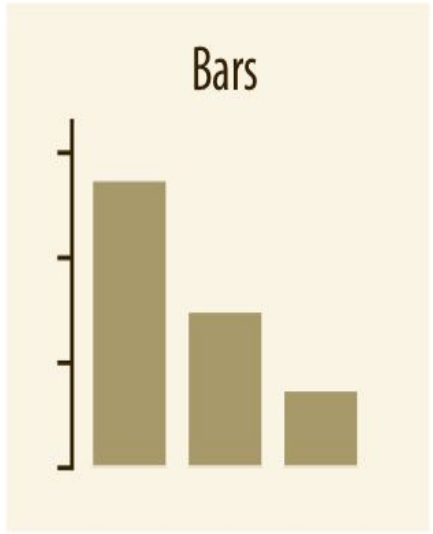
and all scales represent a different variable from the dataset.

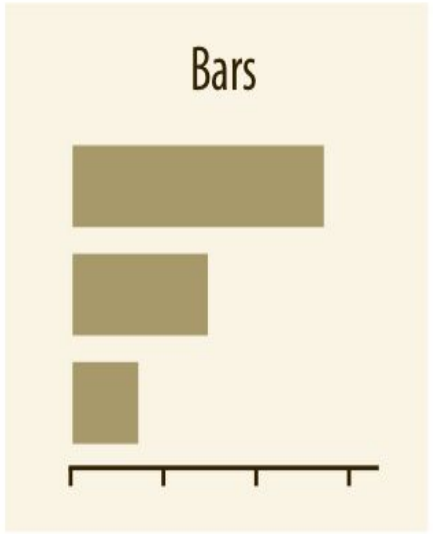
Visualizing Amounts

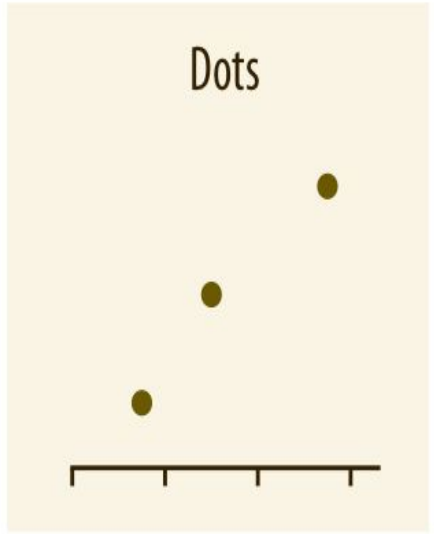
- For example, we might want to visualize the *total sales volume of different brands of cars,* or the *total number of people living in different cities*, or the *age of olympians performing different sports.*
- In all these cases, we have a set of categories (e.g., brands of cars, cities, or sports) and a quantitative value for each category.
- There are visualizing amounts, because the main emphasis in these visualizations will be on the magnitude of the quantitative values.

- The standard visualization in this scenario is the **bar plot**, which comes in several variations, including simple bars as well as grouped and stacked bars. Alternatives to the bar plot are the **dot plot** and the **heatmap.**

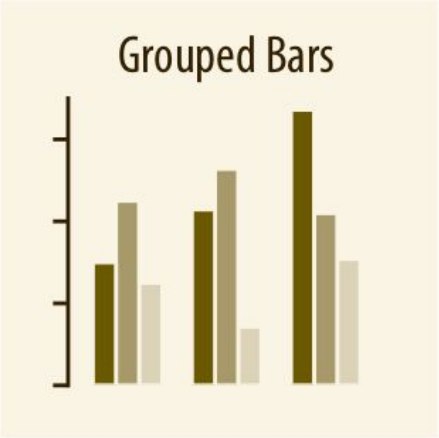
Visualizing Amounts

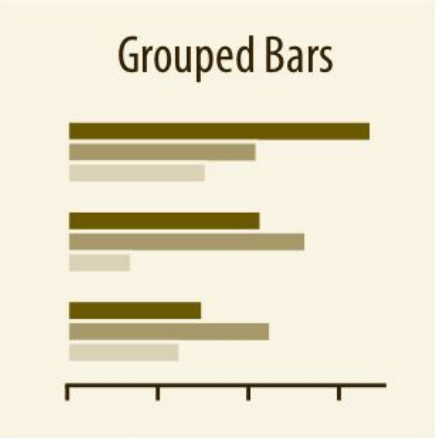


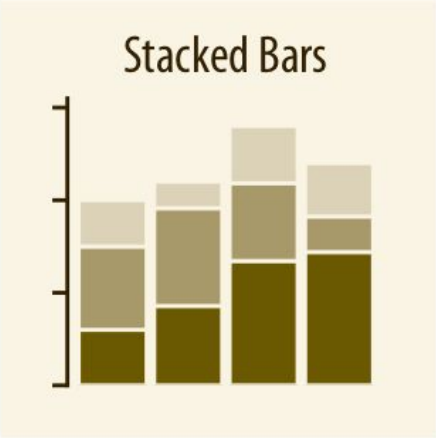


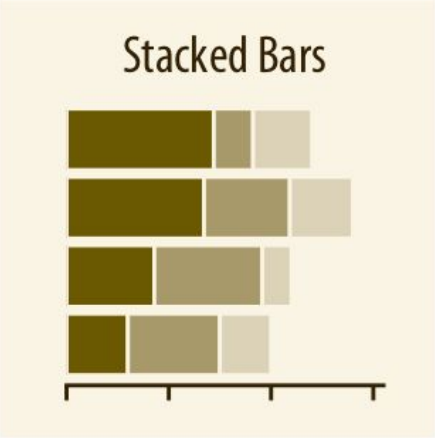


Visualizing Amounts









Visualizing Distributions

- Understand how a particular variable is distributed in a dataset.
- Example:
  - There were approximately 1300 passengers on the Titanic (not counting crew), and we have reported ages for 756 of them. We might want to know how many passengers of what ages there were on the Titanic, i.e., how many children, young adults, middle-aged people, seniors, and so on. We call the relative proportions of different ages among the passengers the *age distribution* of the passengers.

Example: **Visualizing a single distribution**

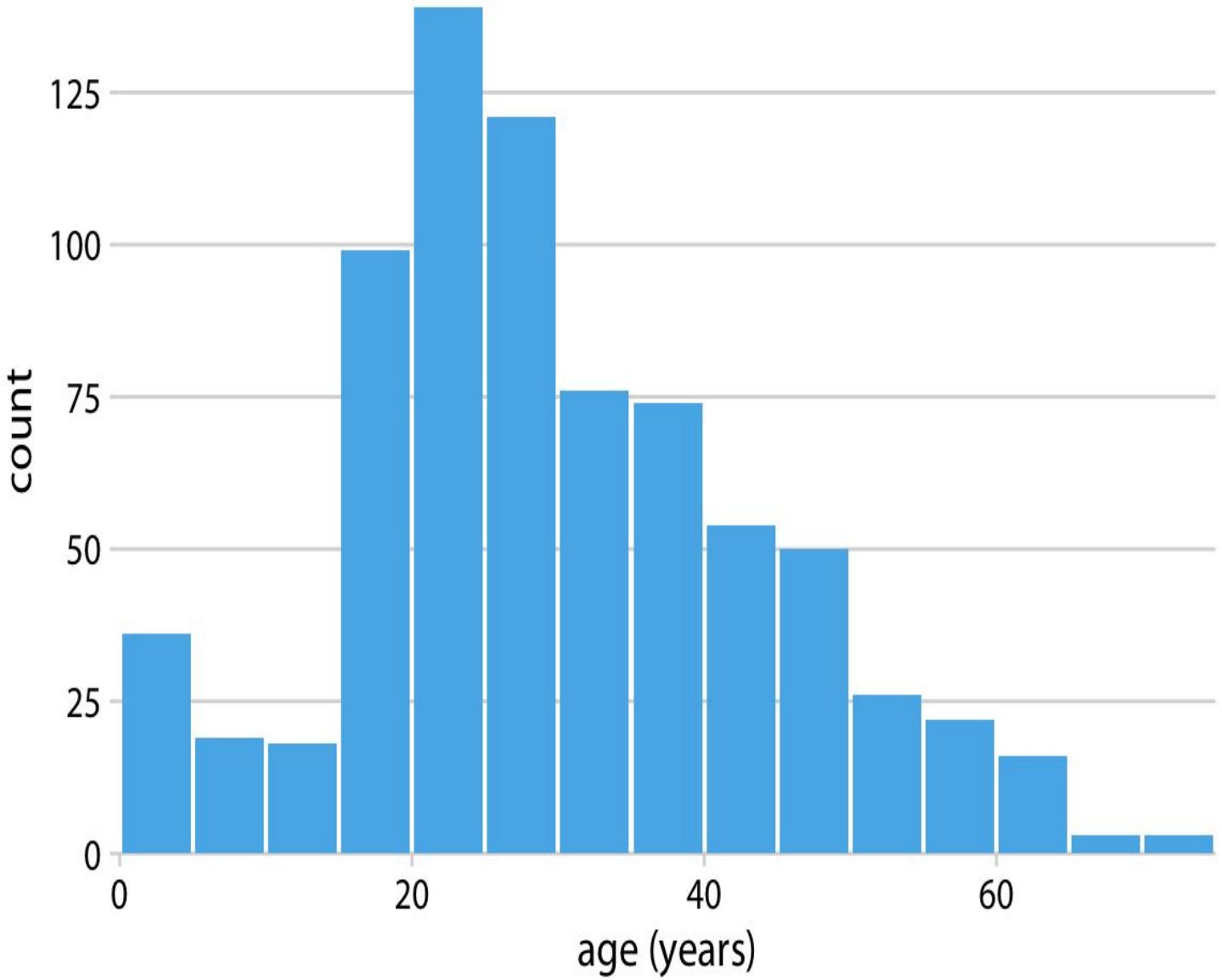
Numbers of passenger with known age on the Titanic.

Age range	Count
0–5	36
6–10	19
11–15	18
16–20	99
21–25	139
26–30	121

Age range	Count
31–35	76
36–40	74
41–45	54
46–50	50
51–55	26
56–60	22

Age range	Count
61–65	16
66–70	3
71–75	3

Plot a histrogram

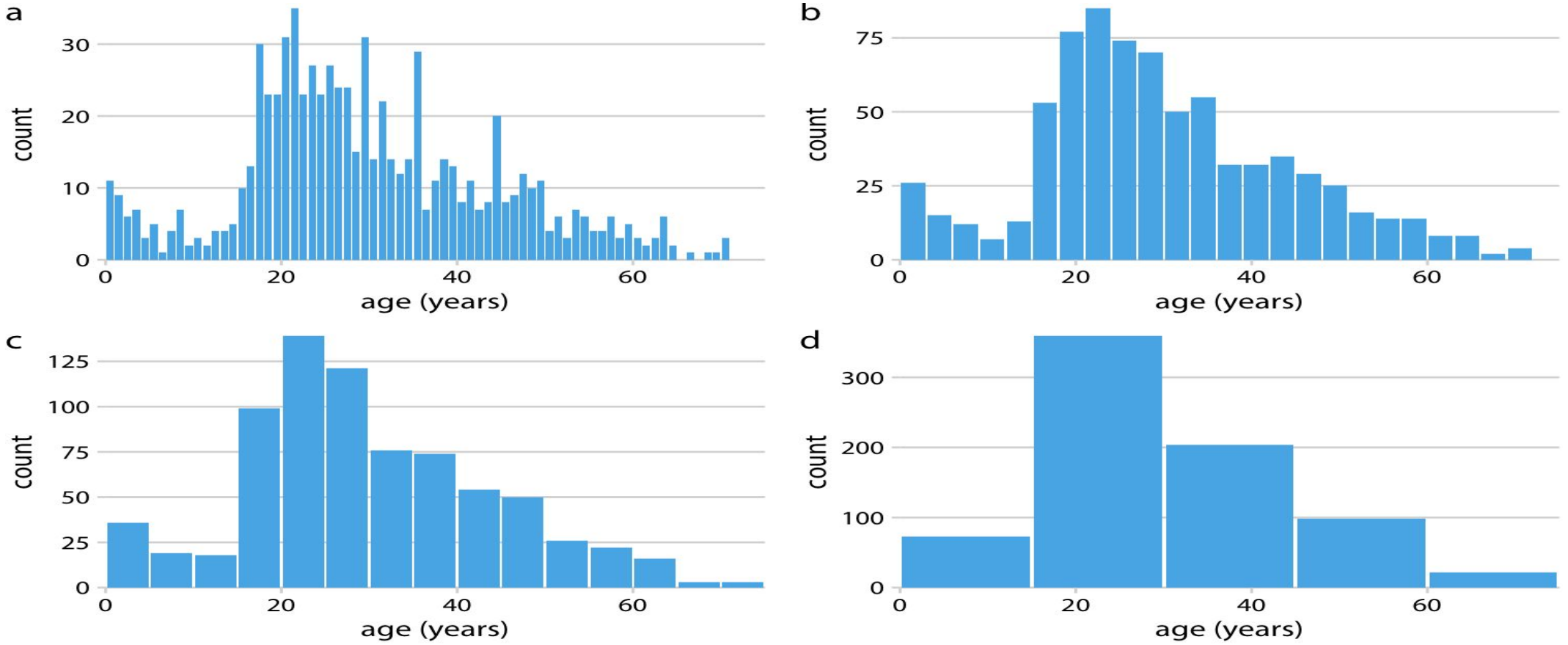


histograms are generated by binning the data, their exact visual appearance depends on the choice of the bin width

In general, if the bin width is too small, then the histogram becomes overly peaky and visually busy and the main trends in the data may be obscured.
On the other hand, if the bin width is too large, then smaller features in the distribution of the data, such as the dip around age 10, may disappear.

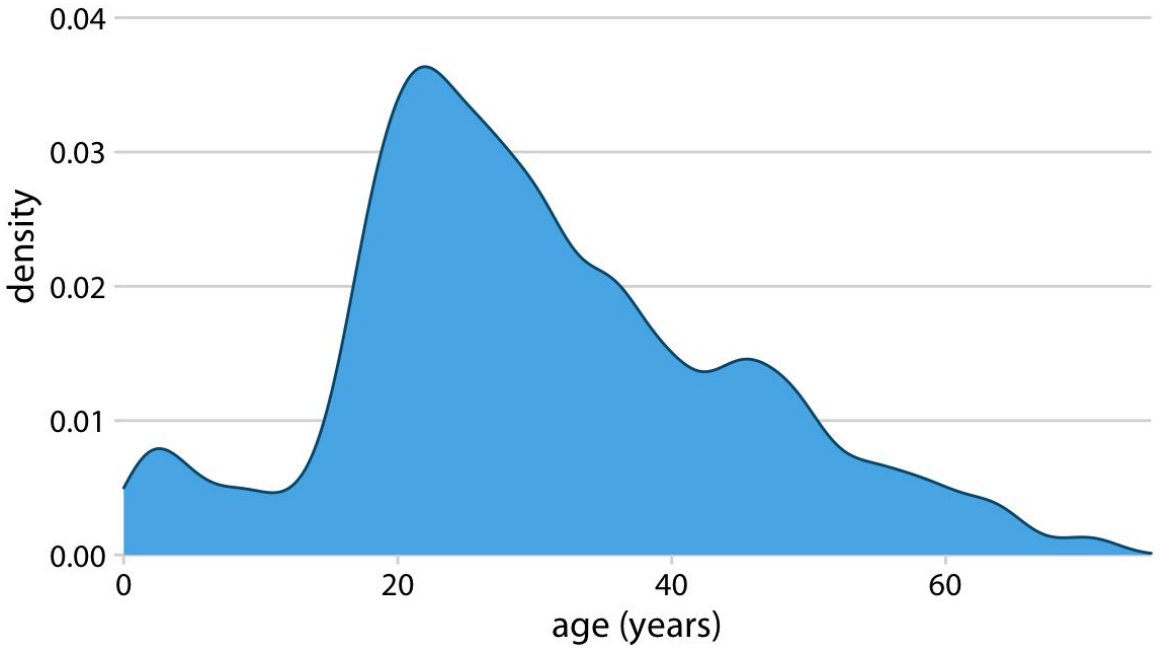
Example: **Visualizing a single distribution**

For the age distribution of Titanic passengers, we can see that a bin width of one year is too small and a bin width of twenty years is too large, whereas bin widths between three to five years work fine



Density Plot

- In a density plot, we attempt to visualize the underlying probability distribution of the data by drawing an appropriate continuous curve
- This curve needs to be estimated from the data, and the most commonly used method for this estimation procedure is called *kernel density estimation*. In kernel density estimation, we draw a continuous curve (the kernel) with a small width (controlled by a parameter called *bandwidth*) at the location of each data point, and then we add up all these curves to obtain the final density estimate. The most widely used kernel is a Gaussian kernel (i.e., a Gaussian bell curve), but there are many other choices.

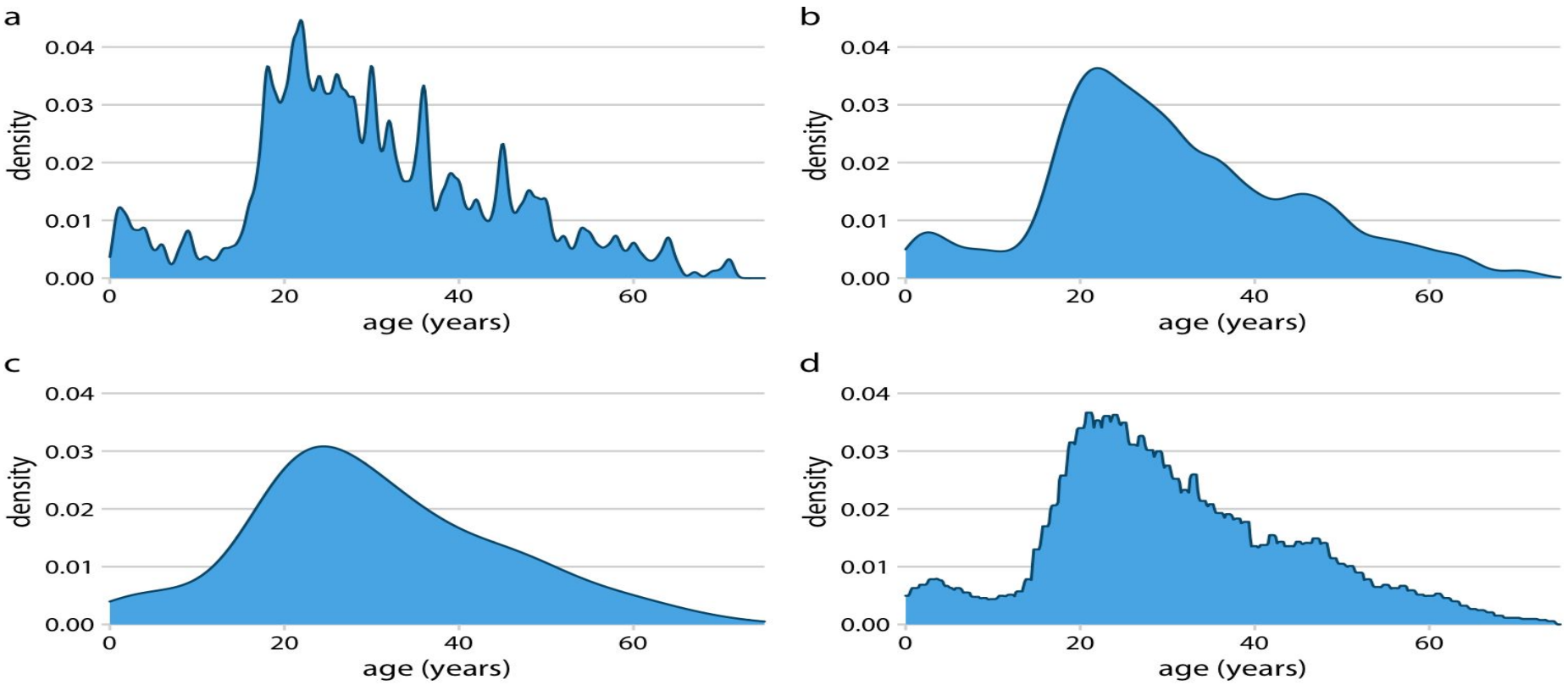


Density Plot

- Just as is the case with histograms, the exact visual appearance of a density plot depends on the kernel and bandwidth choices
- The bandwidth parameter behaves similarly to the bin width in histograms.
- If the bandwidth is too small, then the density estimate can become overly peaky and visually busy and the main trends in the data may be obscured.
- On the other hand, if the bandwidth is too large, then smaller features in the distribution of the data may disappear.
- In addition, the choice of the kernel affects the shape of the density curve.
- For example, a Gaussian kernel will have a tendency to produce density estimates that look Gaussian-like, with smooth features and tails.
- By contrast, a rectangular kernel can generate the appearance of steps in the density curve

- In general, the more data points there are in the data set, the less the choice of the kernel matters. Therefore, density plots tend to be quite reliable and informative for large data sets but can be misleading for data sets of only a few points.

Density Plot

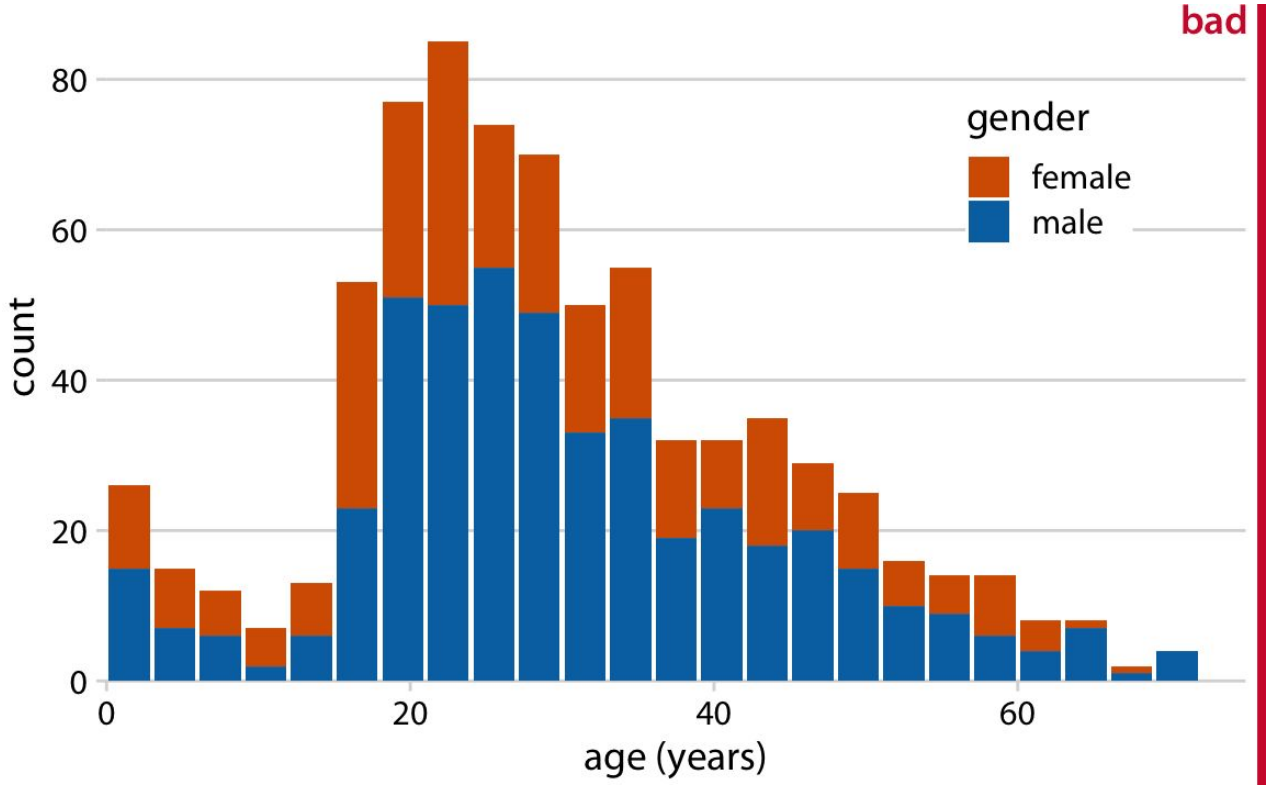


- Density curves are usually scaled such that the area under the curve equals one.

To visualize several distributions at once, kernel density plots will generally work better than histograms.

Visualizing multiple distributions at the same time

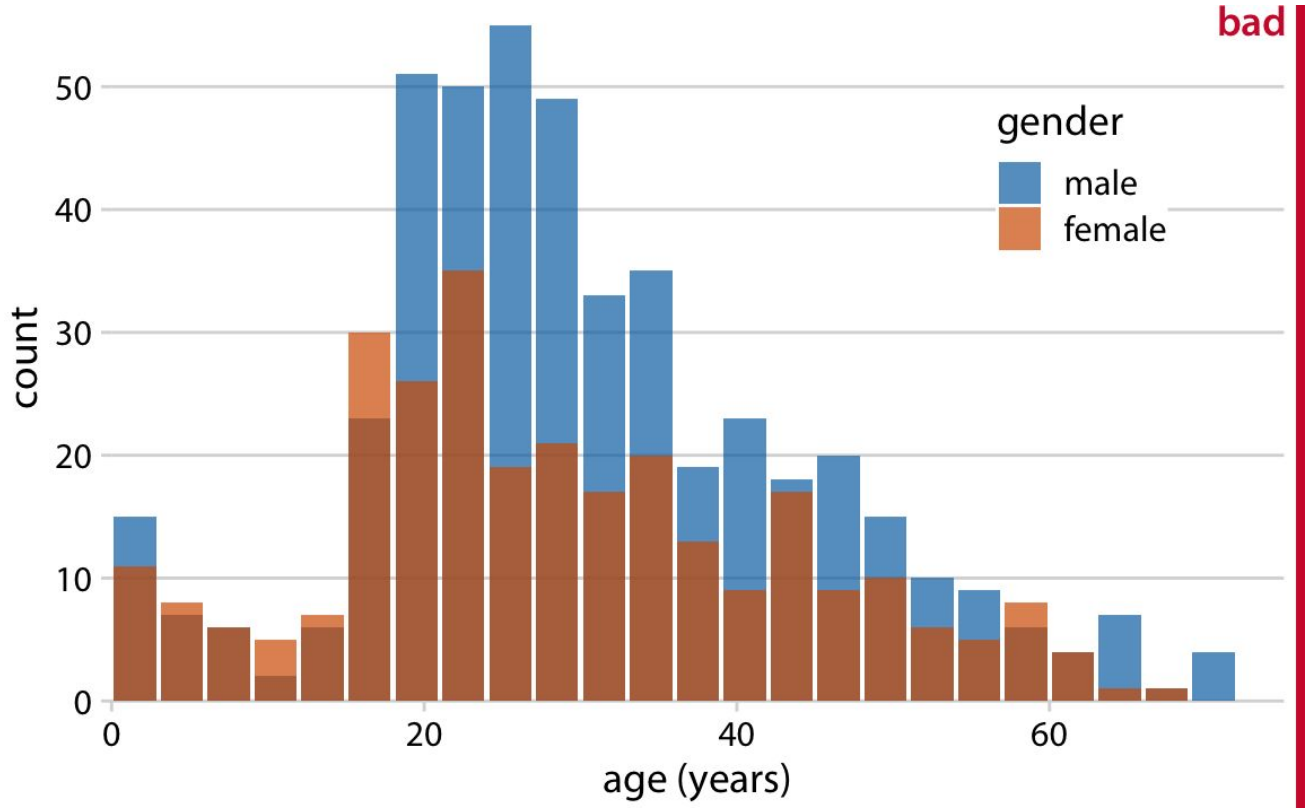
- how the ages of Titanic passengers are distributed between men and women. Were men and women passengers generally of the same age, or was there an age difference between the genders? One commonly employed visualization strategy in this case is a stacked histogram



it is never entirely clear where exactly the bars begin. Do they start where the color changes or are they meant to start at zero? In other words, are there about 25 females of age 18–20 or are there almost 80?

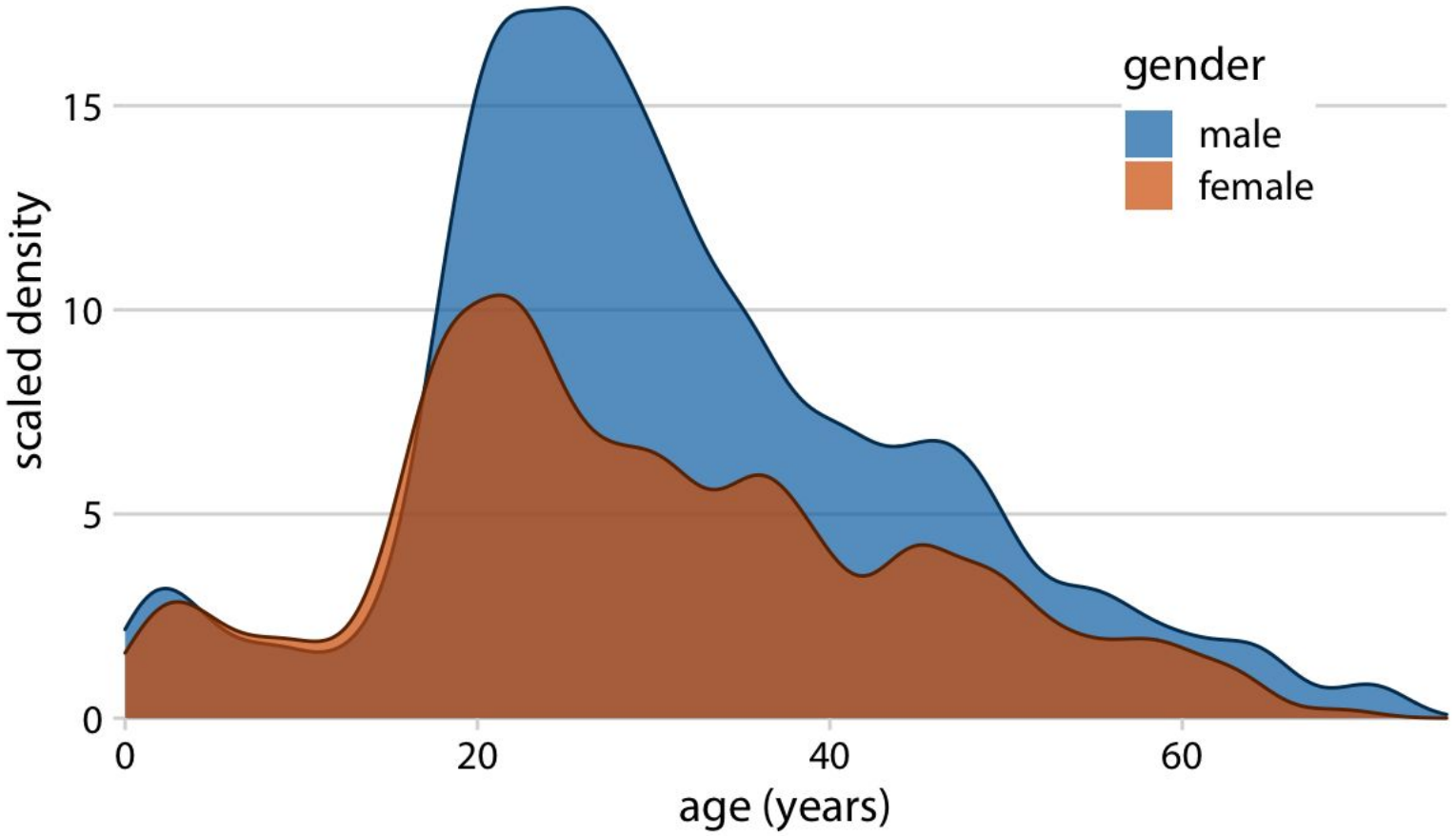
the bar heights for the female counts cannot be directly compared to each other, because the bars all start at a different height.

having all bars start at zero and making the bars partially transparent

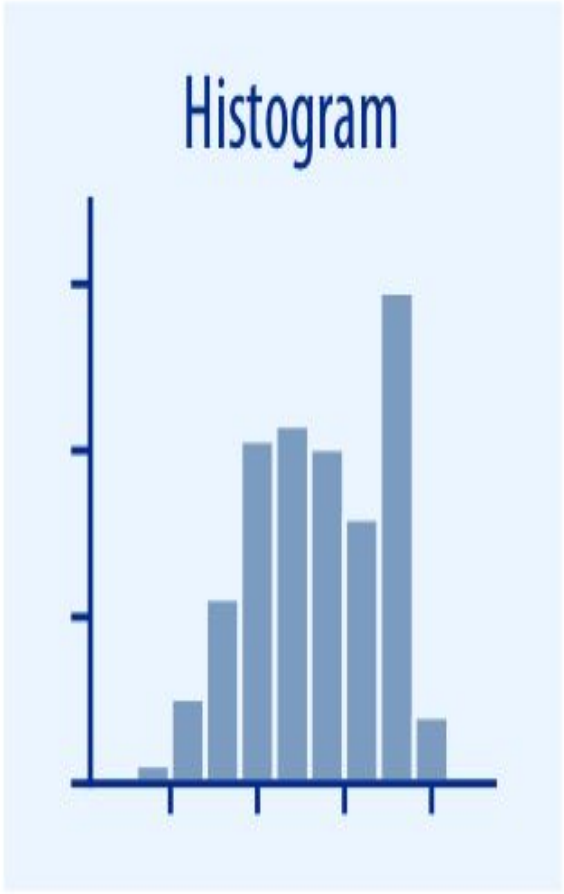


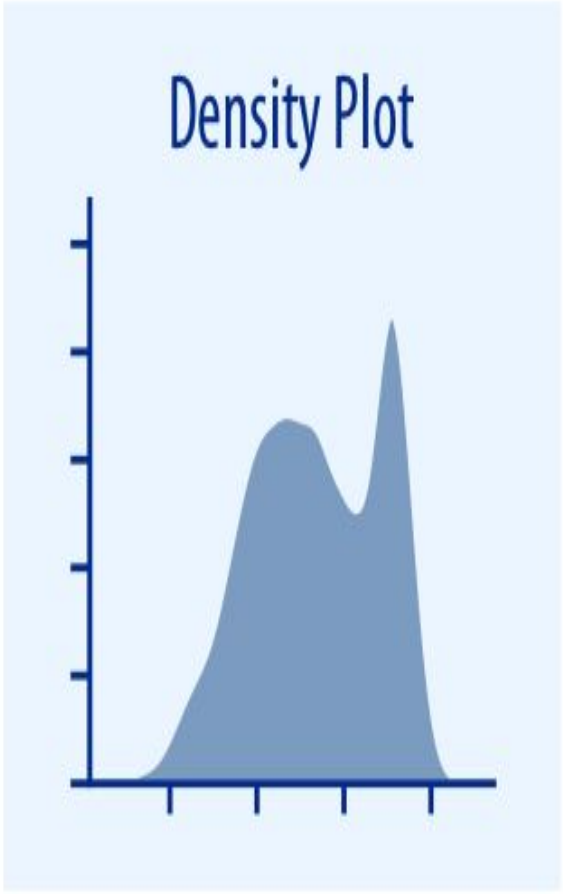
Now it appears that there are actually three different groups, not just two, and we're still not entirely sure where each bar starts and ends. Overlapping histograms don't work well because a semi-transparent bar drawn on top of another tends to not look like a semi-transparent bar but instead like a bar drawn in a different color.

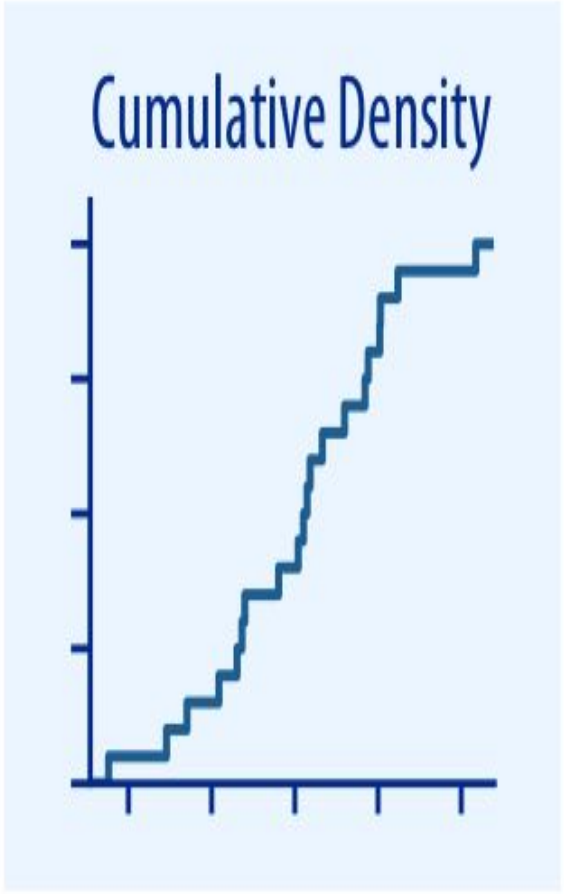
Overlapping density plots

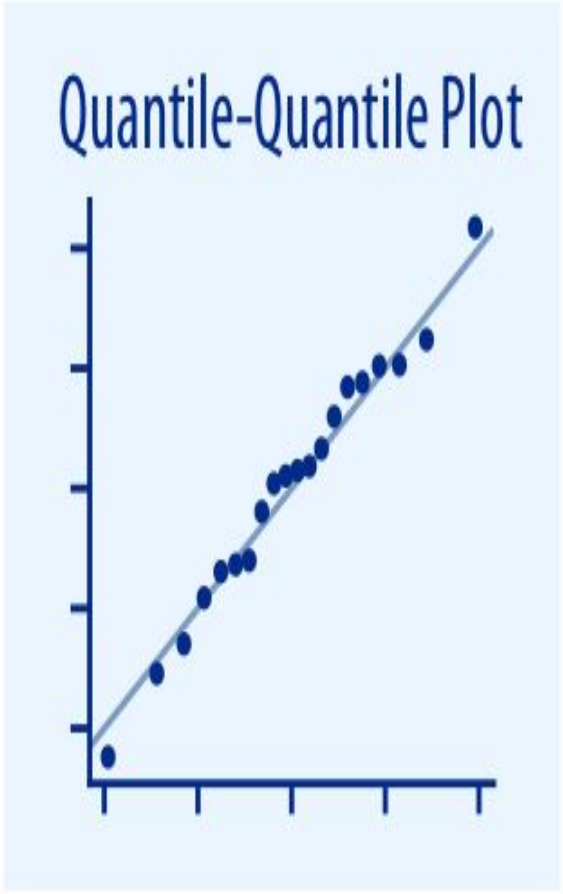


Visualizing Distributions

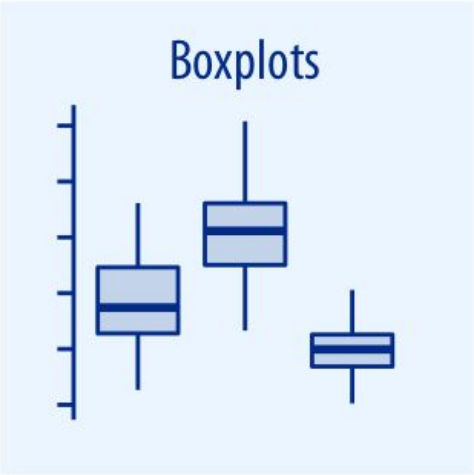


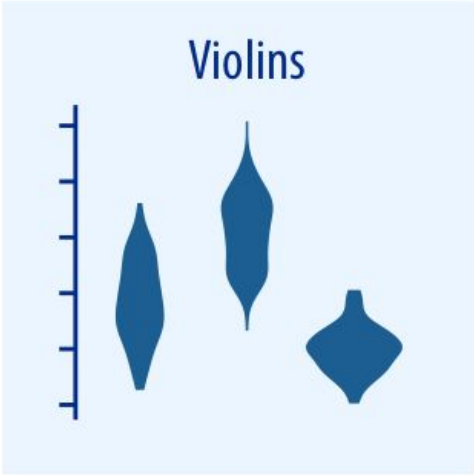


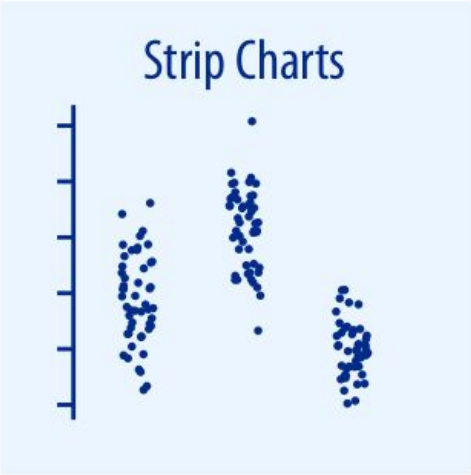


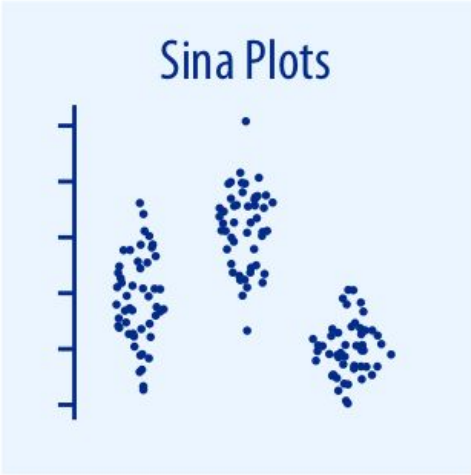


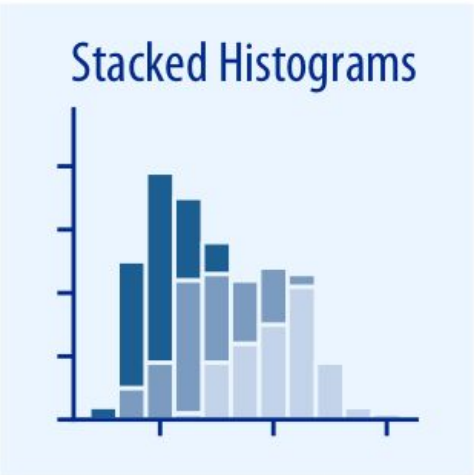
Visualizing Distributions

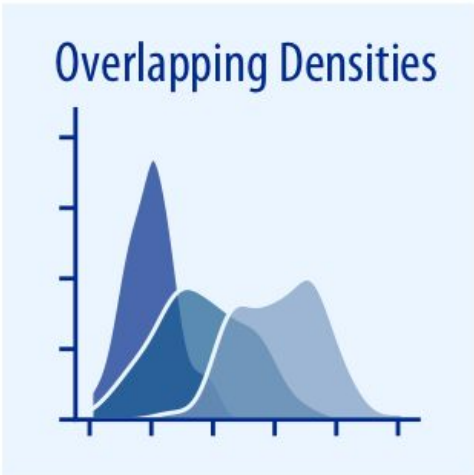


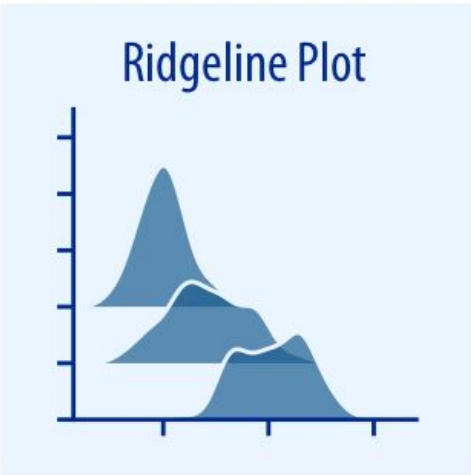












Visualizing Distributions

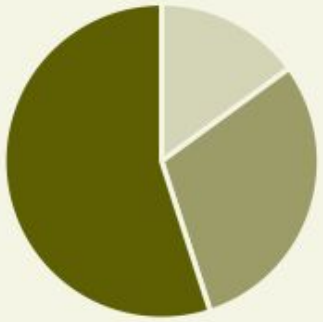
- Boxplots, violins, strip charts, and sina plots are useful when we want to visualize many distributions at once and/or if we are primarily interested **in overall shifts among the distributions**.

- Stacked histograms and overlapping densities allow a more in-depth comparison of a smaller number of distributions, though stacked histograms can be difficult to interpret and are best avoided.

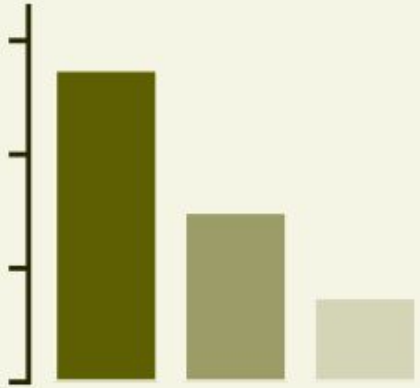
- Ridgeline plots can be a useful alternative to violin plots and are often useful when visualizing very large numbers of distributions or changes in distributions over time

Visualizing Proportions

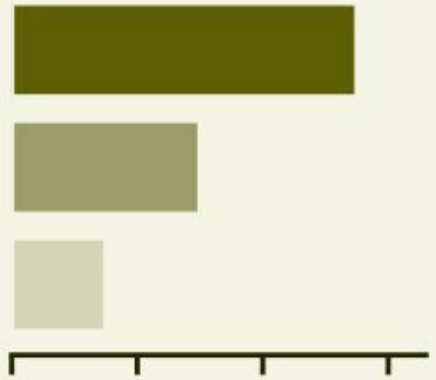
Pie Chart



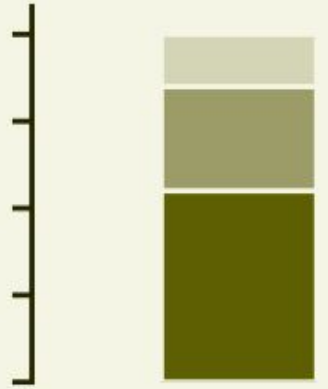
Bars



Bars

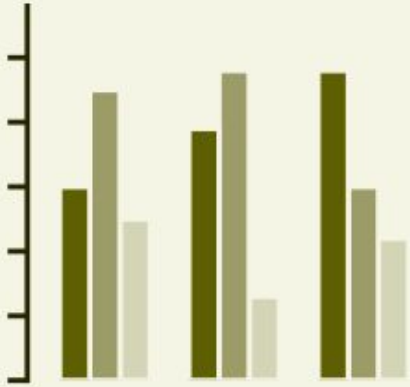


Stacked Bars



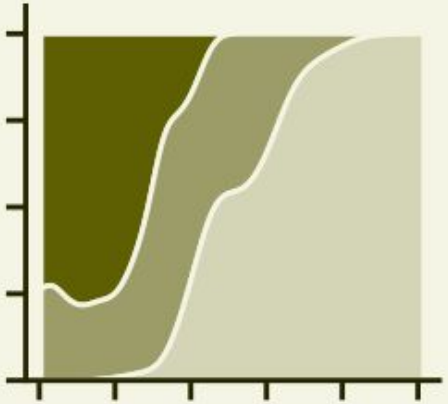
Multiple Pie Charts

Grouped Bars



Stacked Bars

Stacked Densities



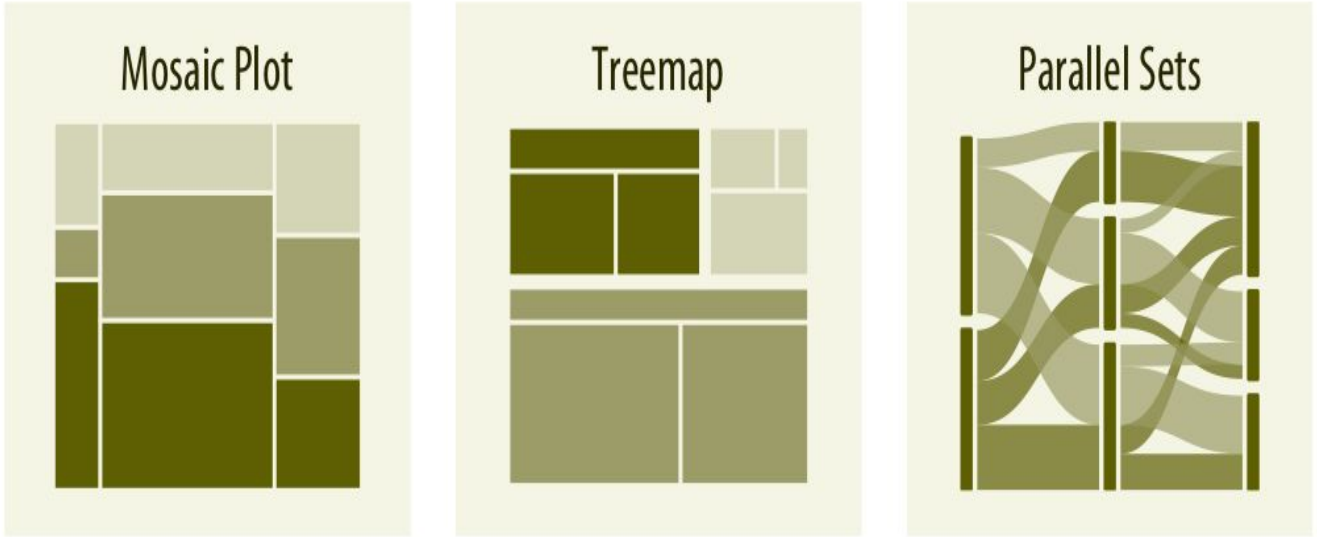
Visualizing Proportions

When visualizing multiple sets of proportions or changes in proportions across conditions, pie charts tend to be space-inefficient and often obscure relationships.

Grouped bars work well as long as the number of conditions compared is moderate, and stacked bars can work for large numbers of conditions.

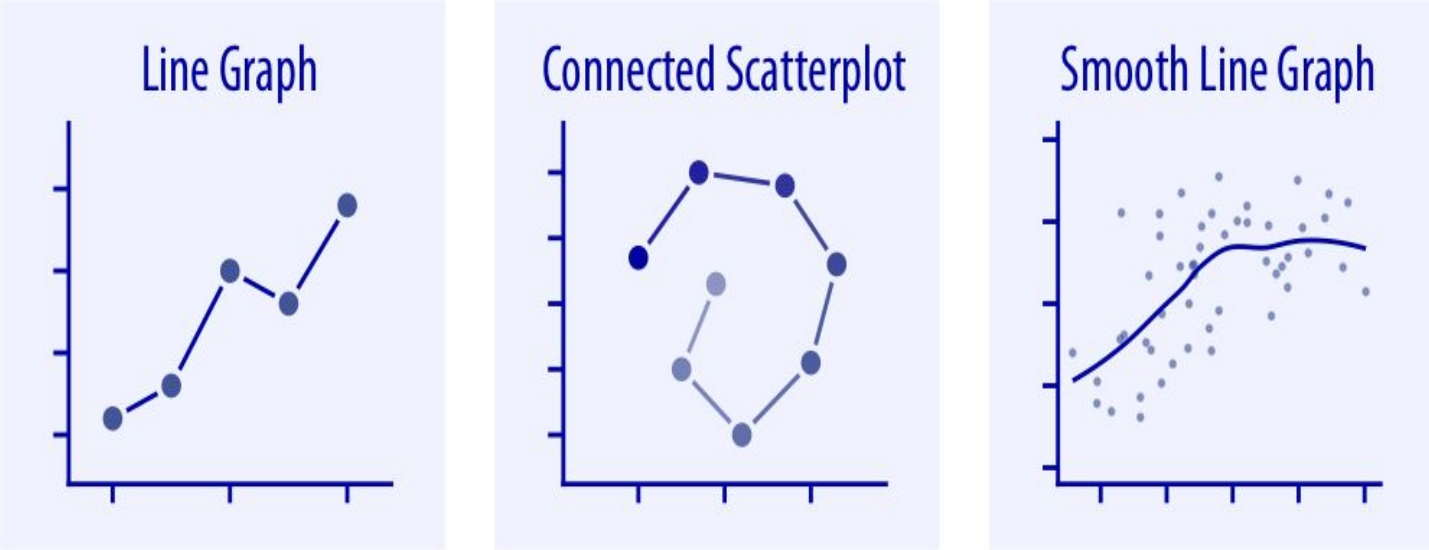
Stacked densities are appropriate when the proportions change along a continuous variable.

Visualizing Proportions



When proportions are specified according to multiple grouping variables, then 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, whereas treemaps do not make such an assumption. Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another. Parallel sets work better than either mosaic plots or treemaps when there are more than two grouping variables.

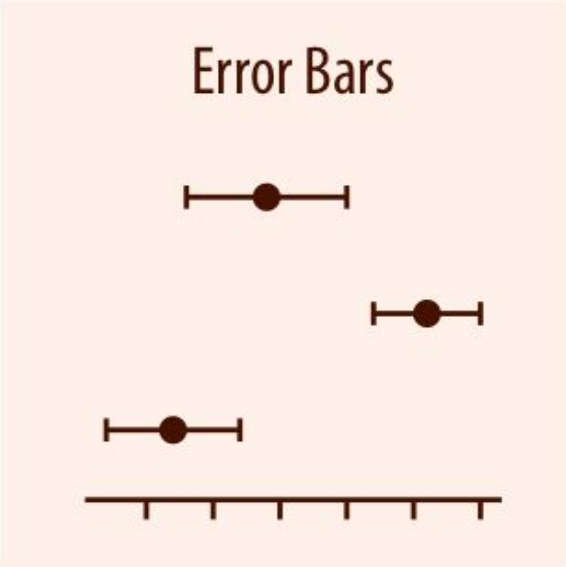
Visualizing Time Series/ Trends

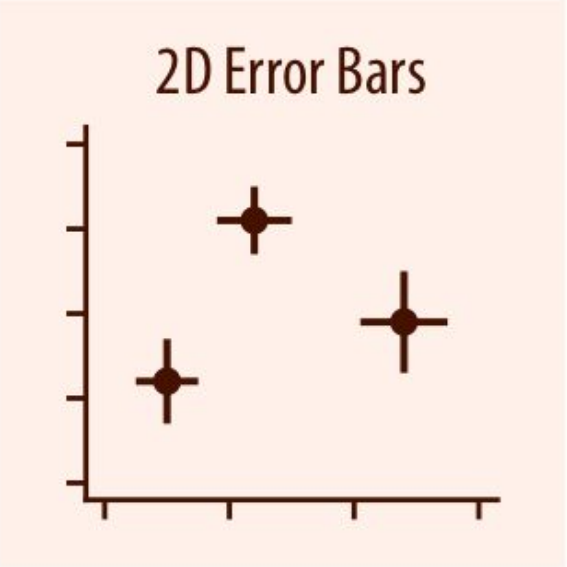


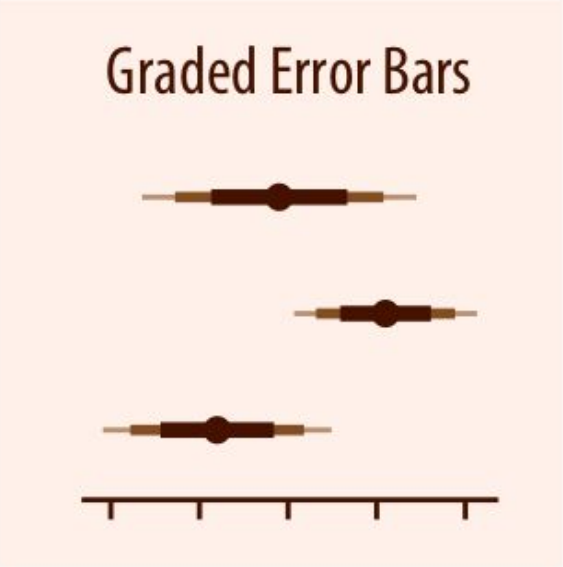
- When the *x* axis represents time or a strictly increasing quantity such as a treatment dose, we commonly draw line graphs.

- 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. We can use smooth lines to represent trends in a larger dataset.

Visualizing Uncertainty







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