Introduction to Data Visualization

Visualizing Amounts & Distributions
Python (matlibplot, seaborn) examples

Halil Bisgin, Ph.D.

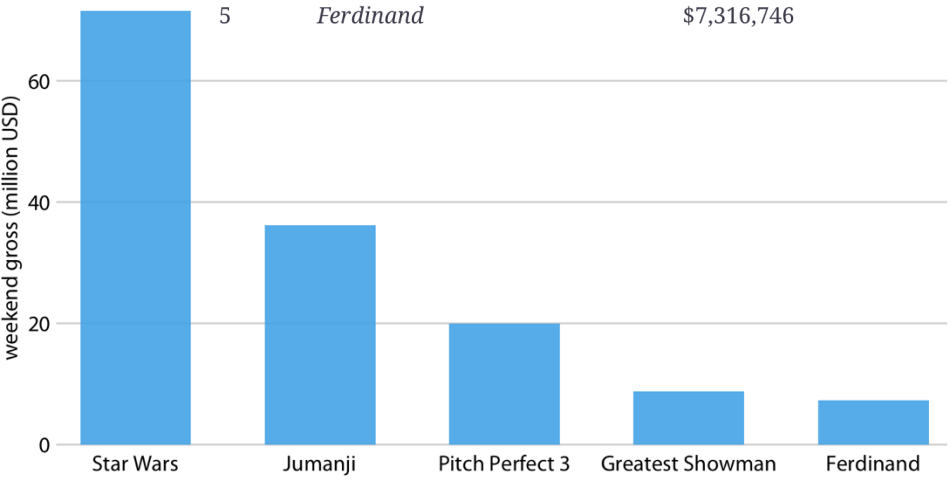
Visualizing Amounts

- In many scenarios, we are interested in the magnitude of some set of numbers.
  - *total sales volume of different brands of cars, the total number of people living in different cities, the age of Olympians performing different sports.*
  - *We have a set of categories (e.g., brands of cars, cities, or sports) and a quantitative value for each category.*
- 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 and its variations.*
  - *Dot plot and the heatmap are the alternatives.*

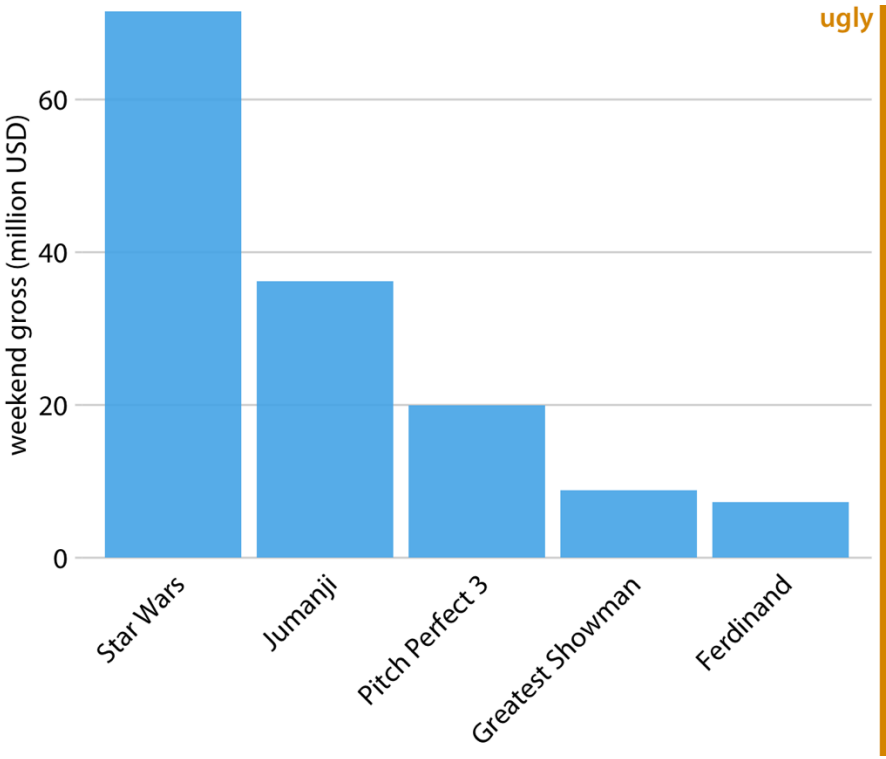
Visualizing Amounts-Bar Plots

- This kind of data is commonly visualized with vertical bars.

Rank	Title	Weekend gross
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746



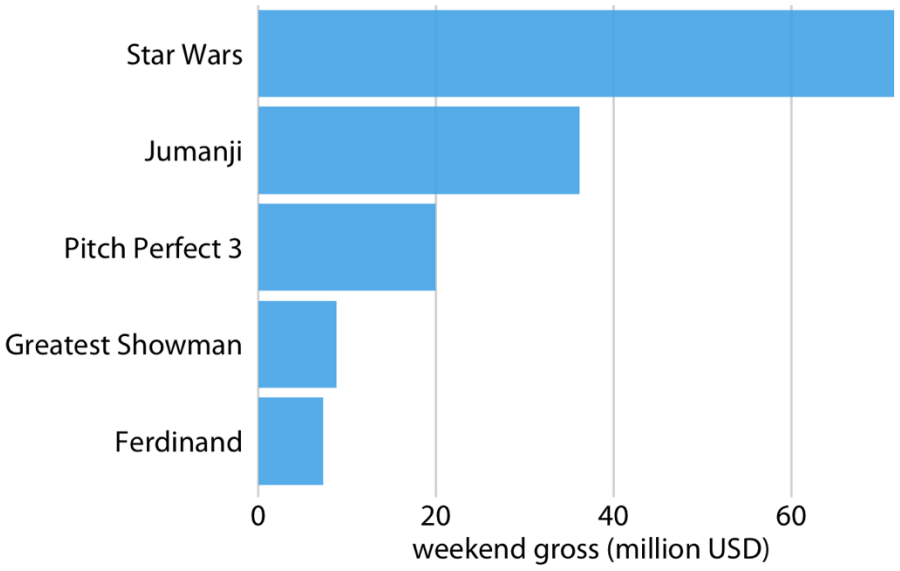
Long labels rotated, but still ugly!

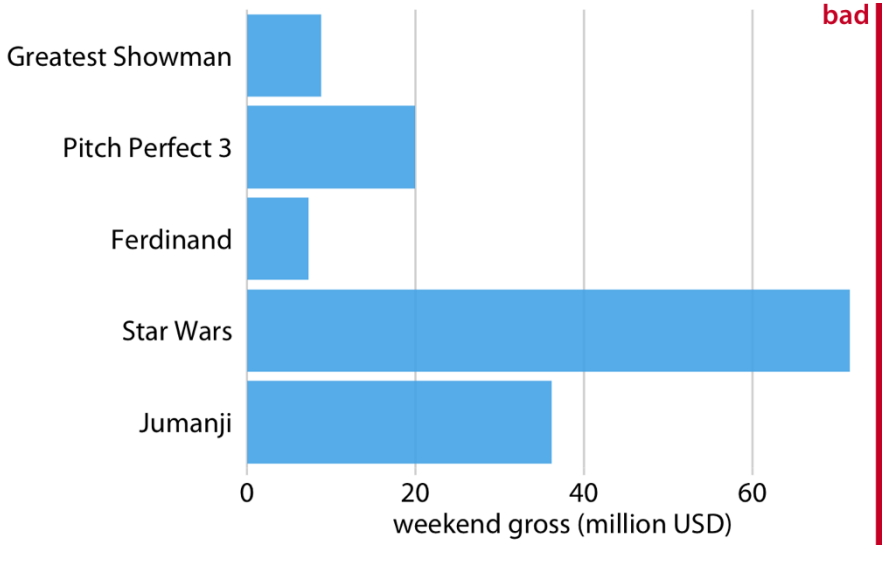


Data: http://www.boxofficemojo.com/

Visualizing Amounts-Bar Plots

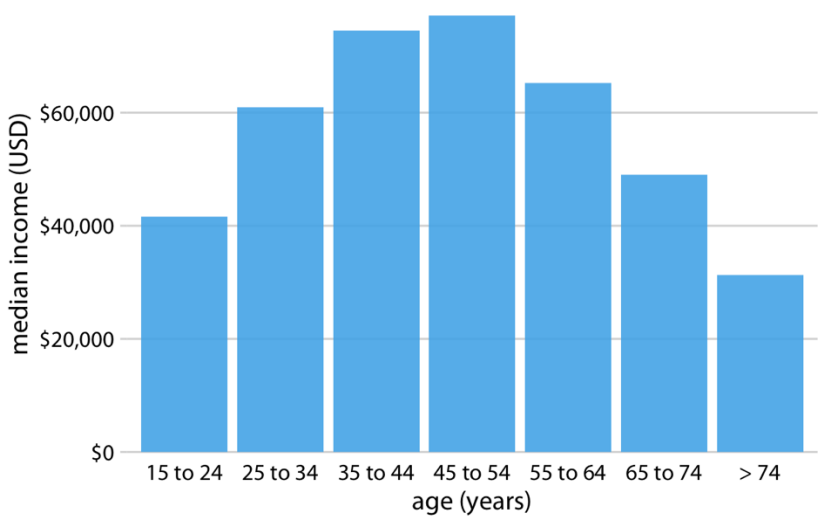
- Better solution: swap x & y axes.
- Be careful about the ordering.

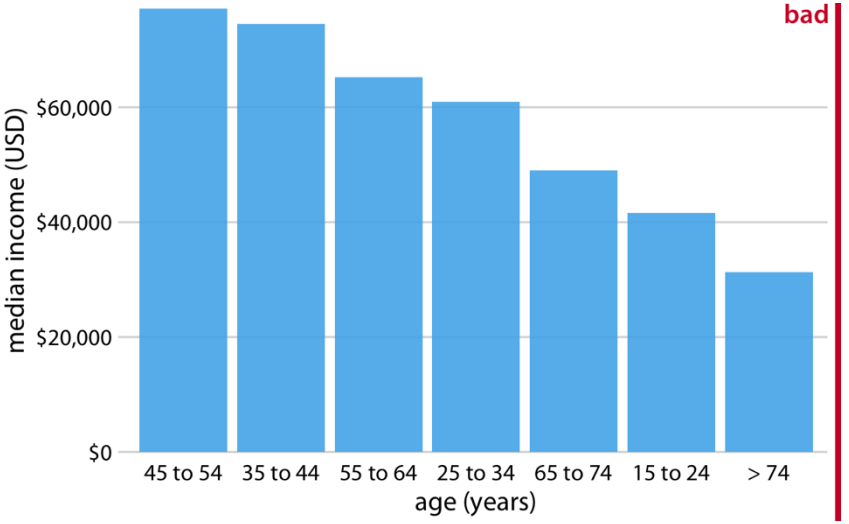




Bar Plots - Ordered vs. Unordered

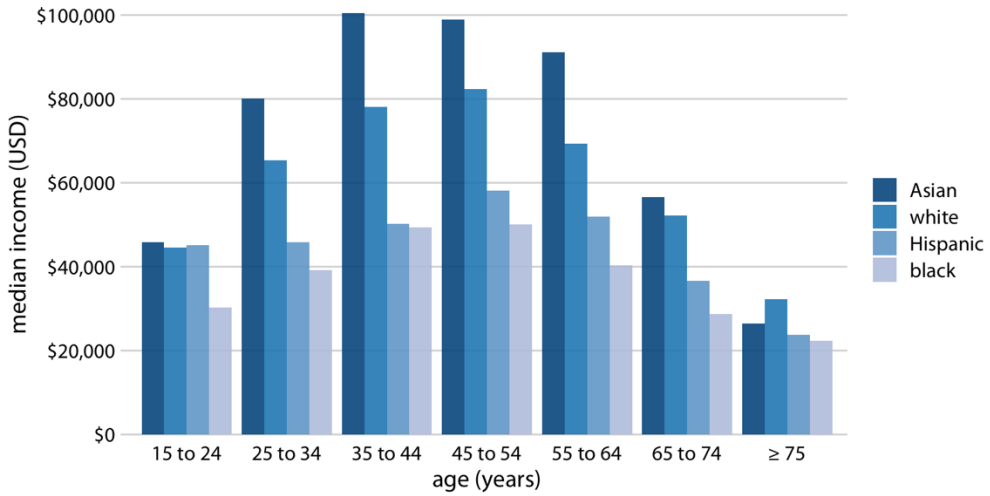
- If ordered, follow the order.
- Otherwise, ascending or descending data values

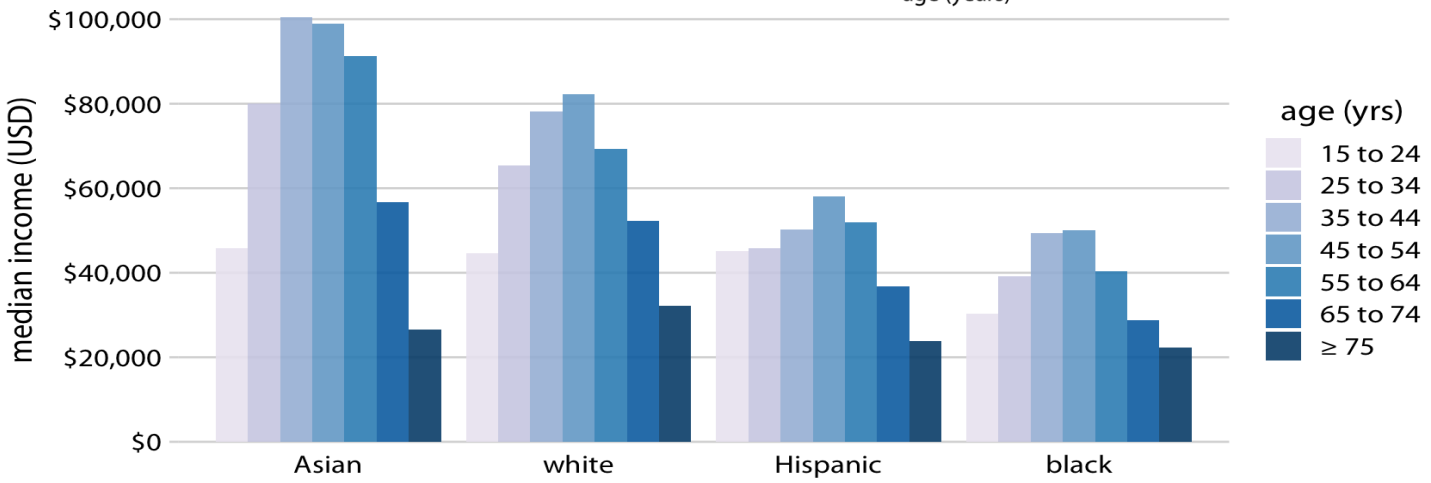




Grouped Bars

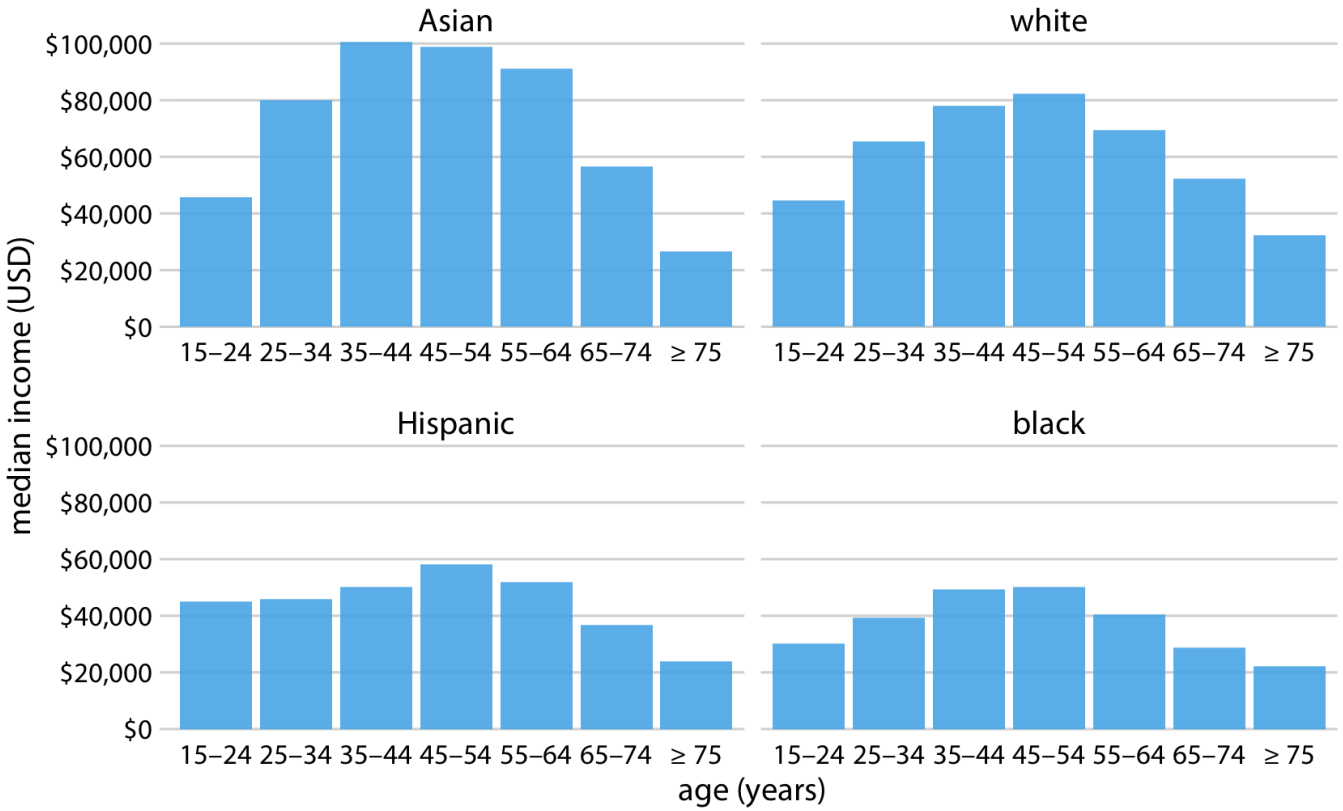
- We are interested in two categorical variables at the same time
- Same info
- Difficult to read





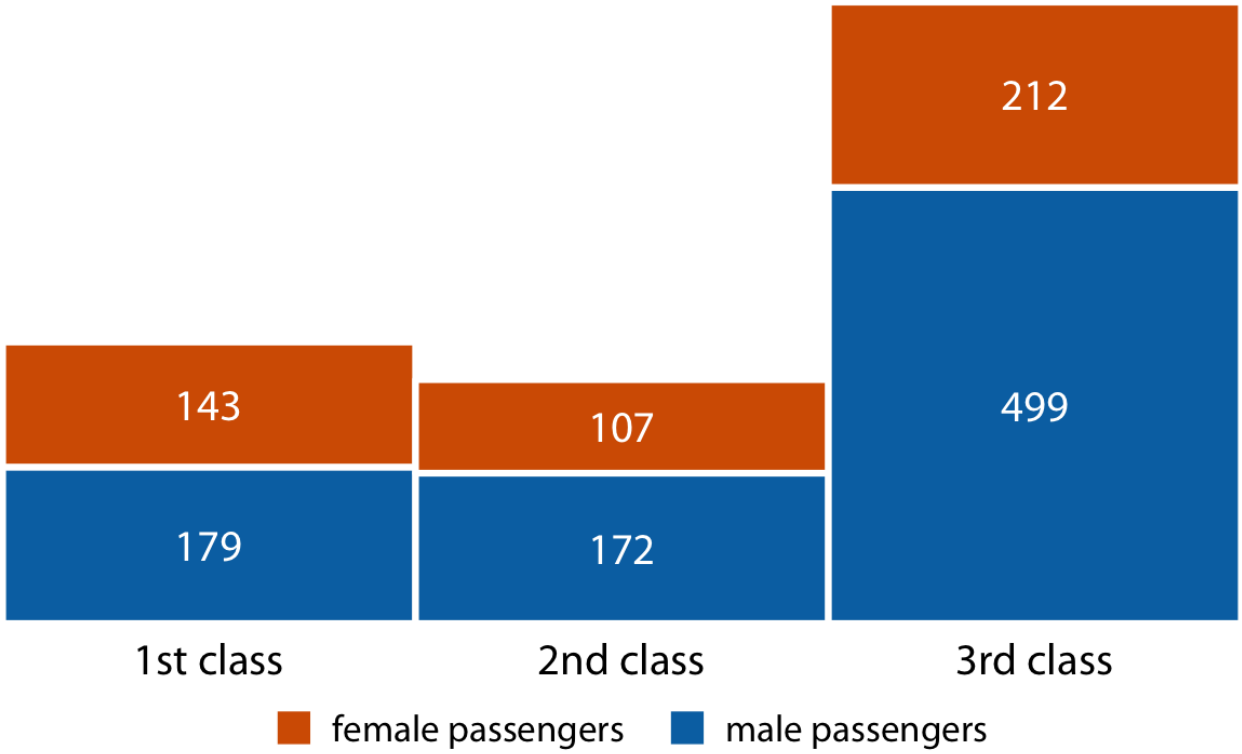
Alternative to Grouped Bars

- Maybe preferable over earlier ones.



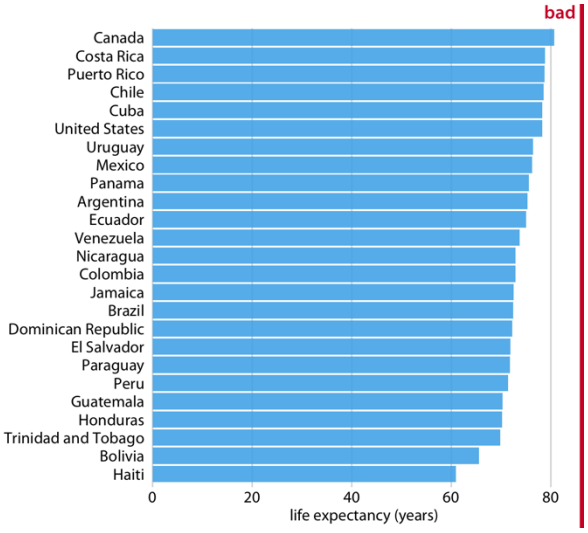
Stacked bars

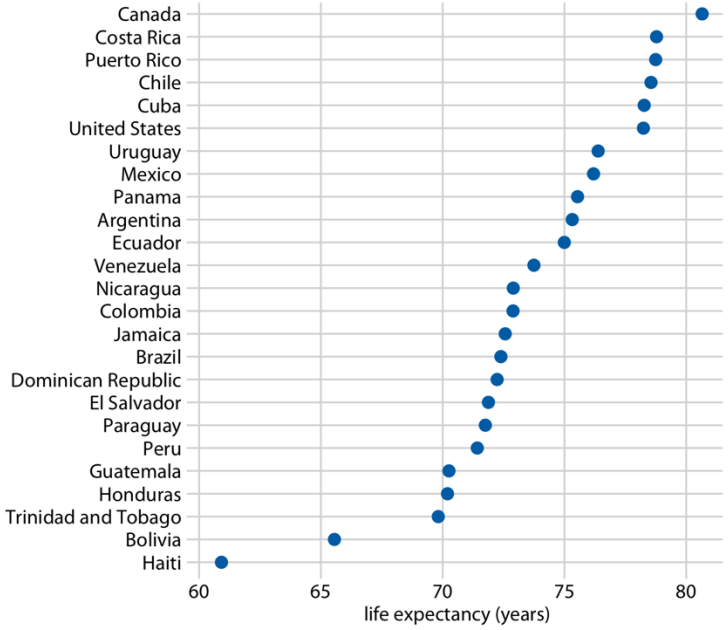
- Sometimes preferable to stack bars on top of each other.
- Useful when the sum is meaningful.

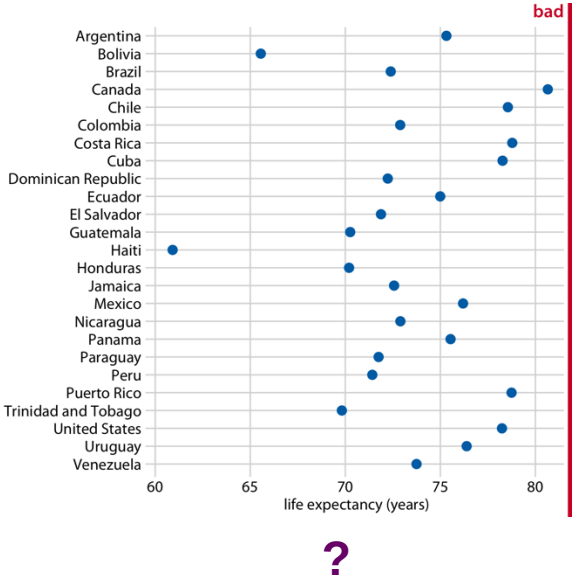


Dot Plots

- Bars should start from zero for a proportional presentation of the amount.
- For some other data, bars are impractical and may obscure key features



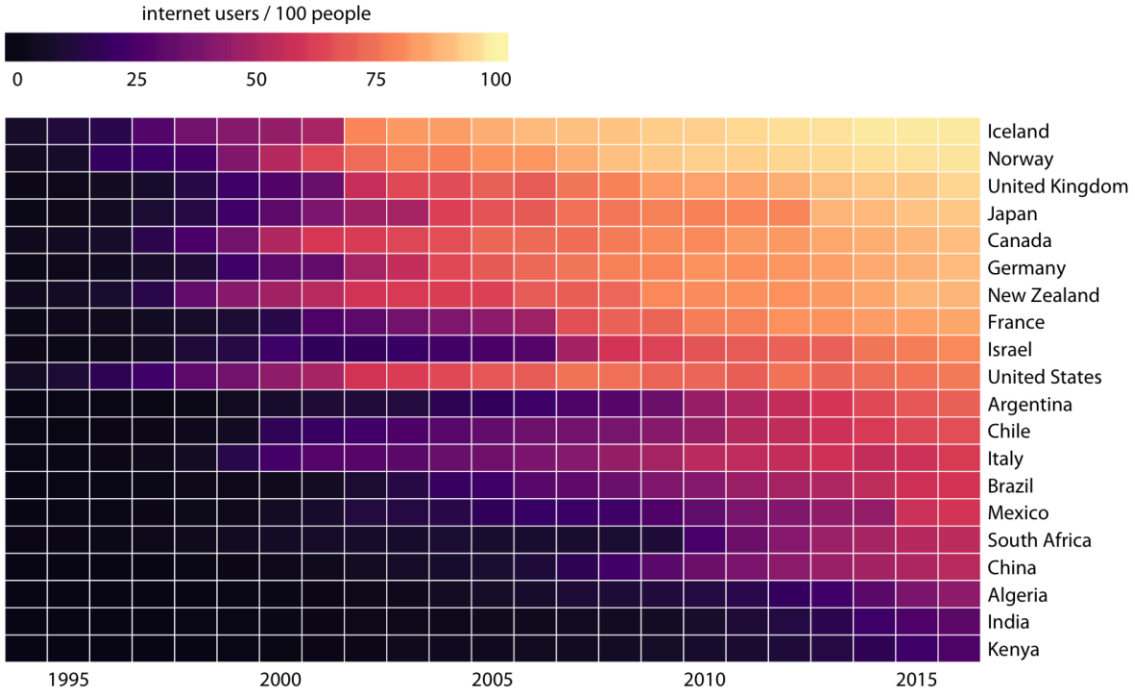




?

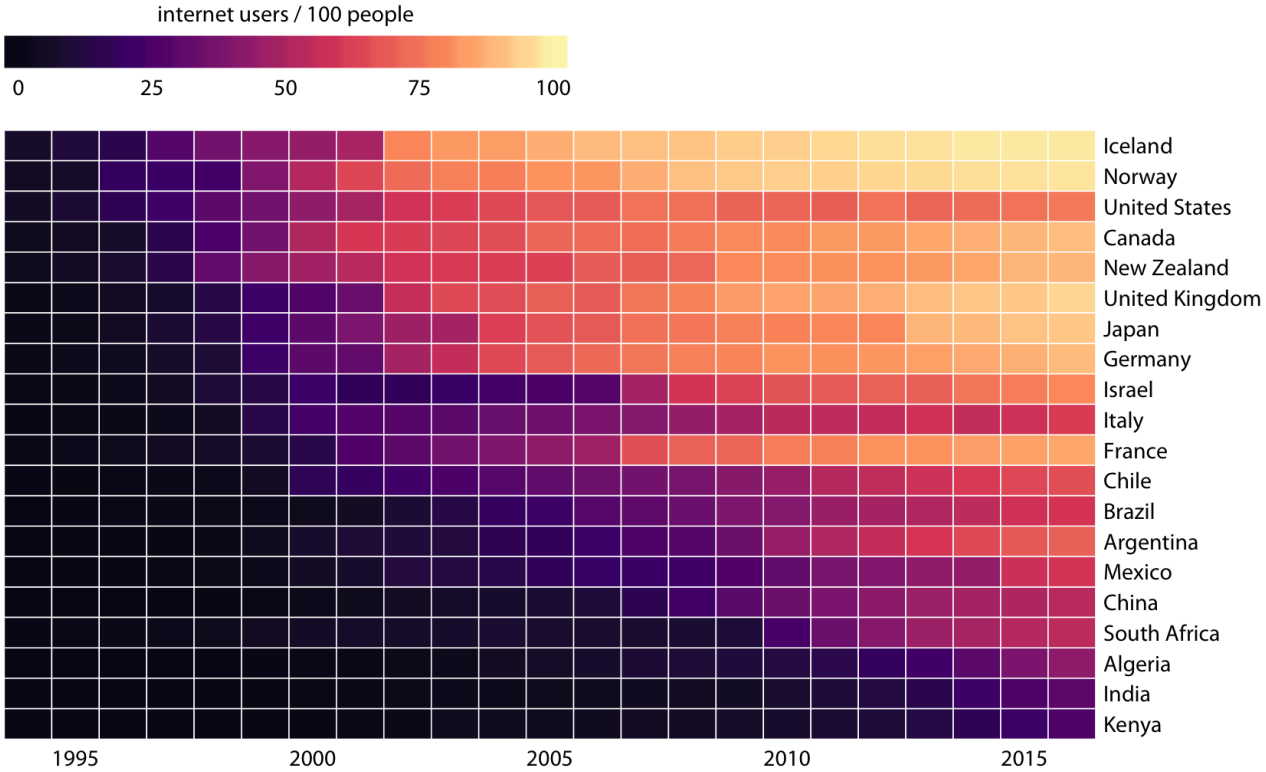
Heatmaps

- As an alternative to mapping data values onto positions via bars or dots, we can map data values onto colors.
- We can't infer exact values, but it helps us see the trend.
- BIG picture



Heatmaps

- Ordering can make a difference again.
- Countries are ordered by the year in which internet usage first rose to above 20%.



Visualizing Distributions: Histograms and Density Plots

- 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.

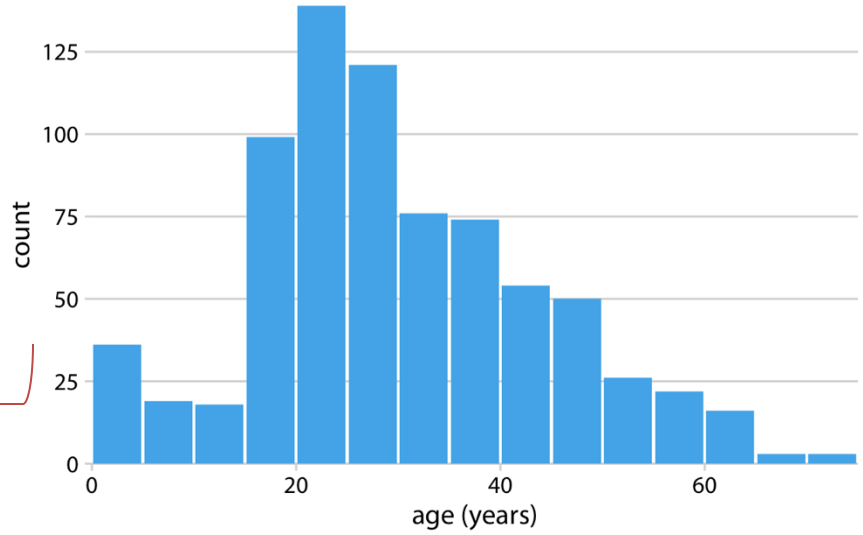
Visualizing a Single Distribution

- Counts for age *bin*s in Titanic …

Age range	Count	Age range	Count	Age range	Count
0–5	36	31–35	76	61–65	16
6–10	19	36–40	74	66–70	3
11–15	18	41–45	54	71–75	3
16–20	99	46–50	50		
21–25	139	51–55	26		
26–30	121	56–60	22		

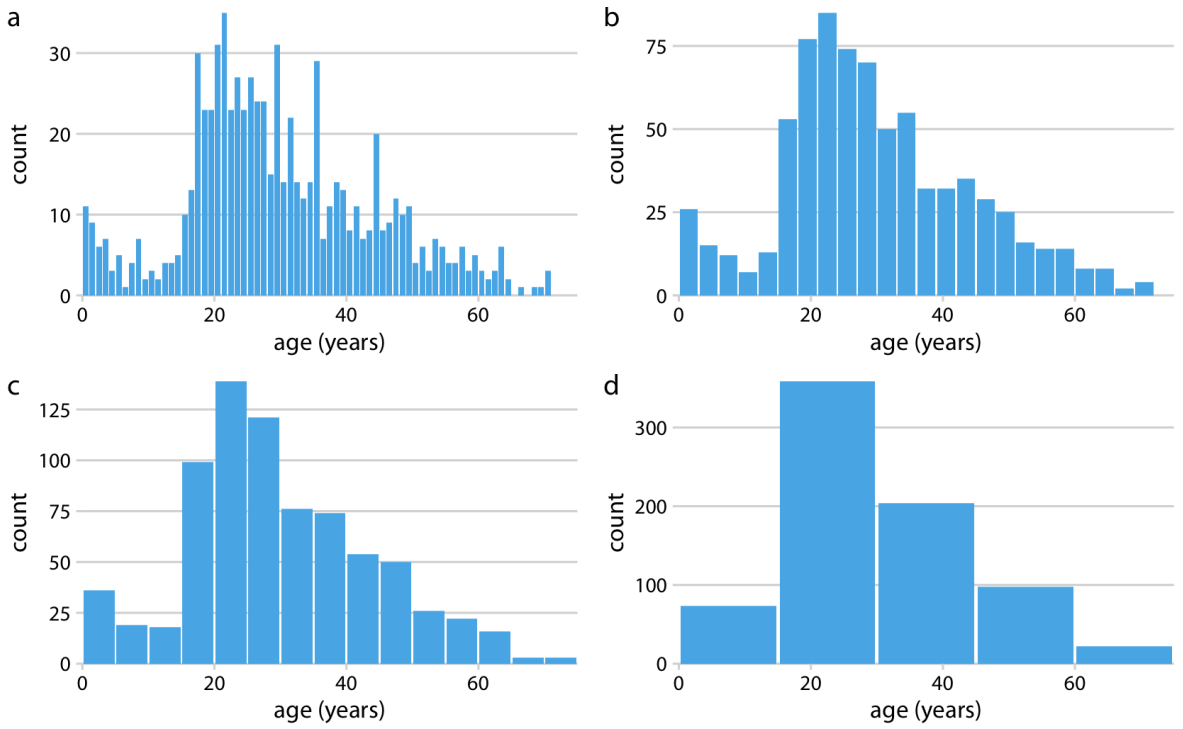
15 bins

… and the *histogram*



Histograms

- Softwares have default bin size (# of bins) which can be changed.
- When making a histogram, always explore multiple bin widths.

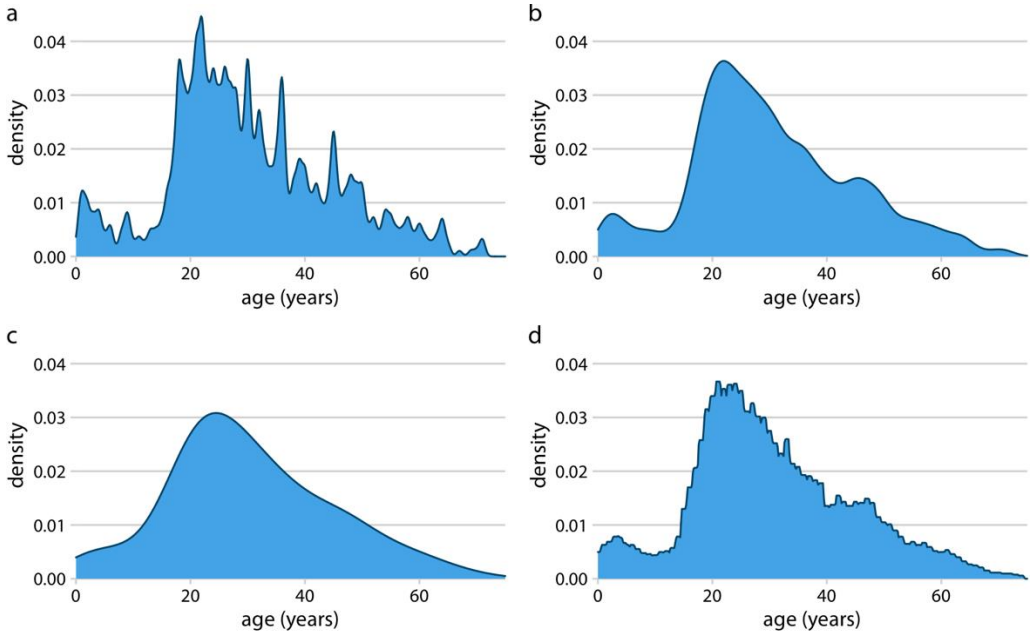


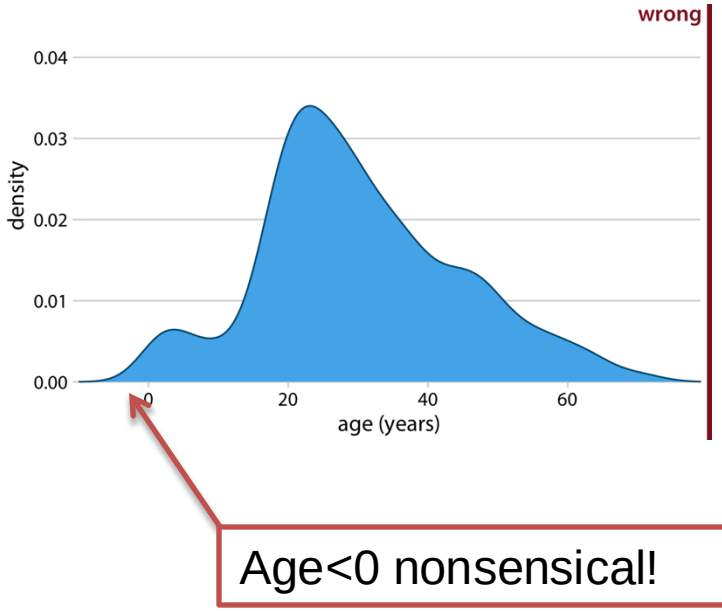
Density Plot

- We try to visualize the underlying distribution by drawing an appropriate continuous curve.
- Needs to be estimated from the data, and the most commonly used method for this estimation procedure is called kernel density estimation.
  - *Draws a continuous curve (the kernel) with a small width (controlled by a parameter called bandwidth) at the location of each data point.*
  - *Adds 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

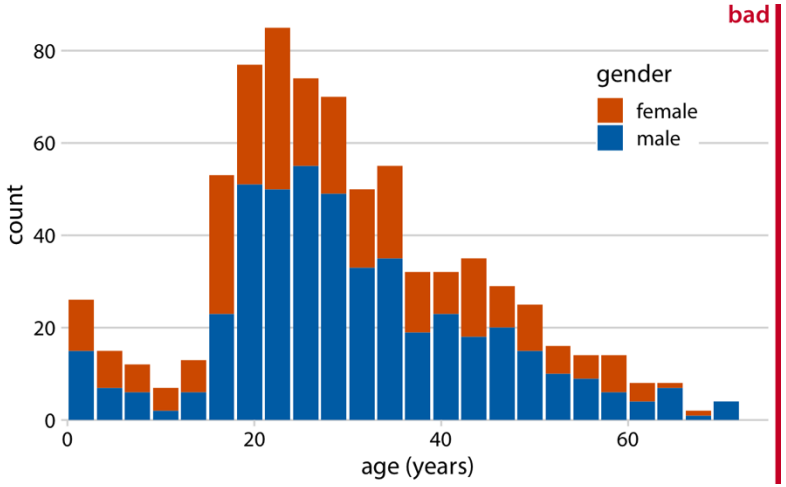
- The bandwidth parameter ~ bin width in histograms.
- Small → peaky and visually busy
- Large → smaller features may disappear
- The kernel affects the shape of the density curve



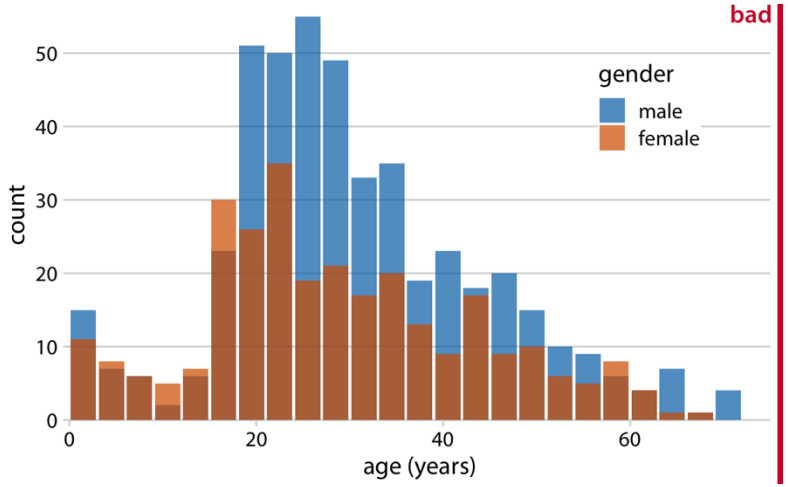


Visualizing Multiple Distributions

- What if more than one distribution simultaneously?
  - *How are the ages of Titanic passengers distributed between men and women?*
  - *Were male and female passengers generally of the same age, or was there an age difference between the genders?*



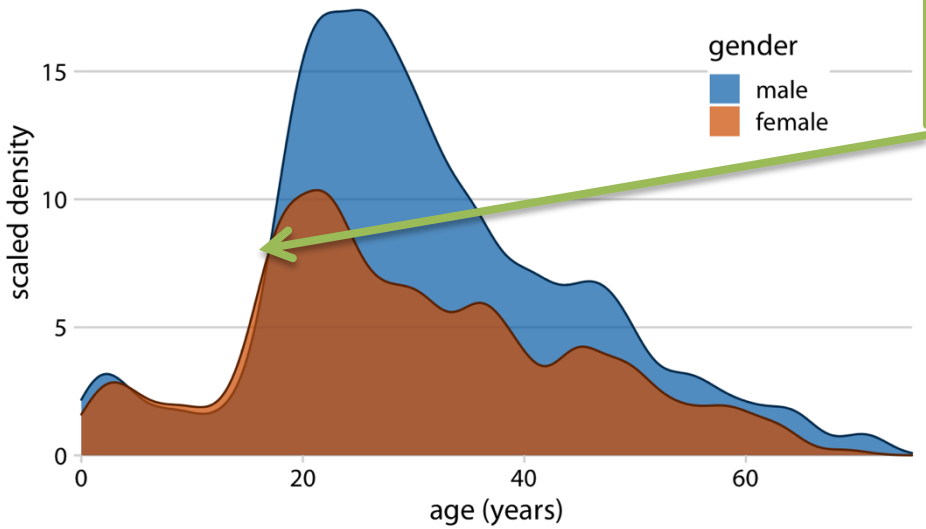
Where does it start?
What are the counts?



Semitransparent, but a third color?
Still ambiguous

Visualizing Multiple Distributions

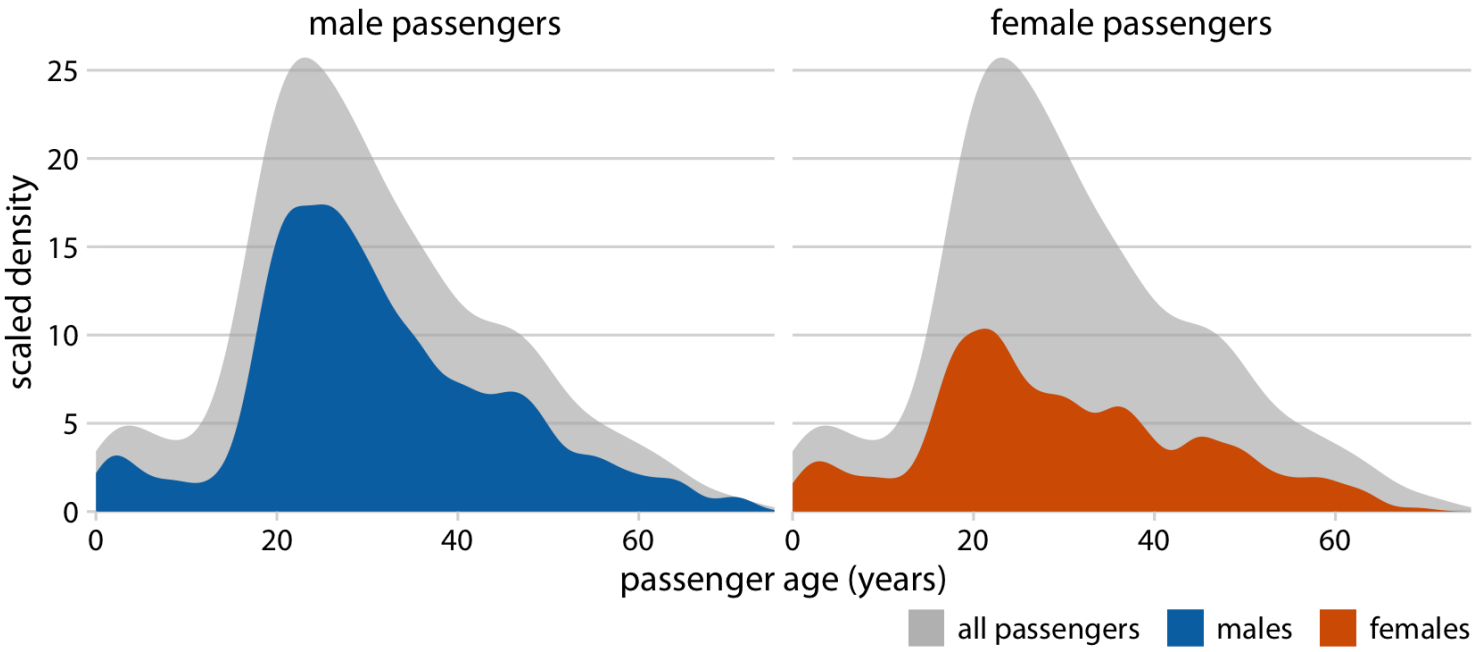
- *Overlapping density plots* don't typically have the problem that overlapping histograms have, because the continuous density lines help the eye keep the distributions separate



Shows identical until age 17.
Not really ideal in this case, but OK.

Visualizing Multiple Distributions

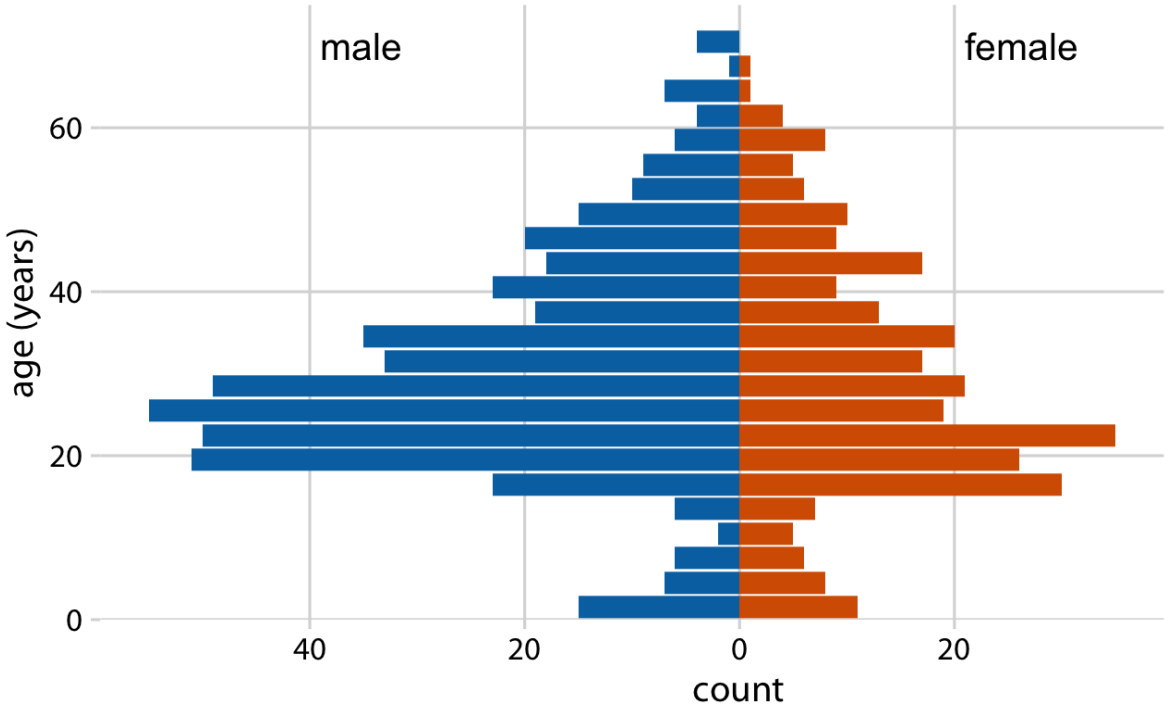
- Proportional to the whole population.
- This visualization shows intuitively and clearly that there were many fewer women than men in the 20-to-50-year age range on the Titanic.



Visualizing Multiple Distributions

- When to visualize exactly two distributions,
  - *we can also make two separate histograms,*
  - *rotate them by 90 degrees, and*
  - *have the bars in the opposite direction of the other.*

- Age pyramid→→



Visualizing Multiple Distributions

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