CP321 Data Visualization

- Visualize Distributions

Jiashu (Jessie) Zhao

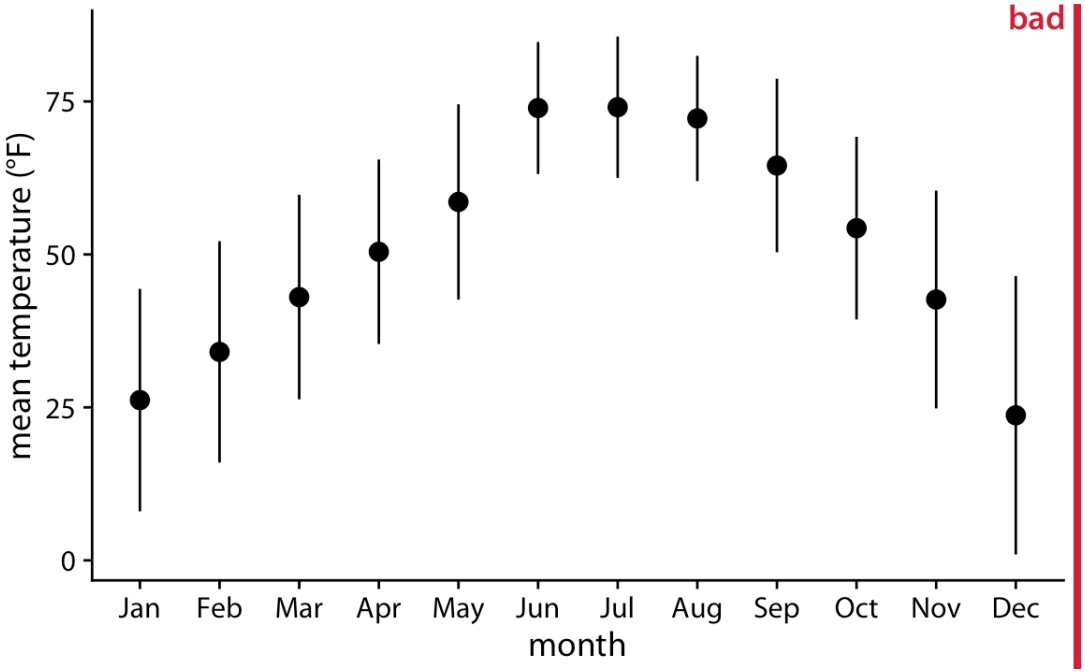
- Visualize multiple distributions at the same time
  - Boxplot
  - Violin plot
  - Strip chart
  - Sina plot
- Design:
  - Reproducibility and repeatability
  - Data exploration VS. data presentation
  - Separation of content and design

Visualize multiple distributions at the same time

- Scenarios in which we want to visualize multiple distributions at the same time.
  - 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

- Example: visualize how *temperature* varies across different *months* while also showing the distribution of observed temperatures within each month.

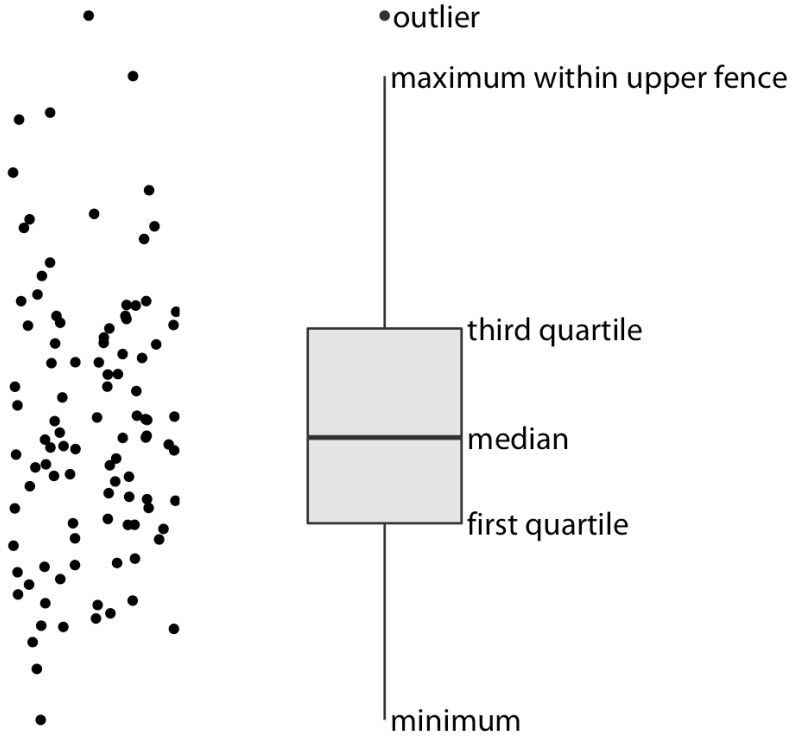
- Why not simply show their mean or median as points, with some indication of the variation by error bars?



Mean daily temperatures in Lincoln, Nebraska in 2016.

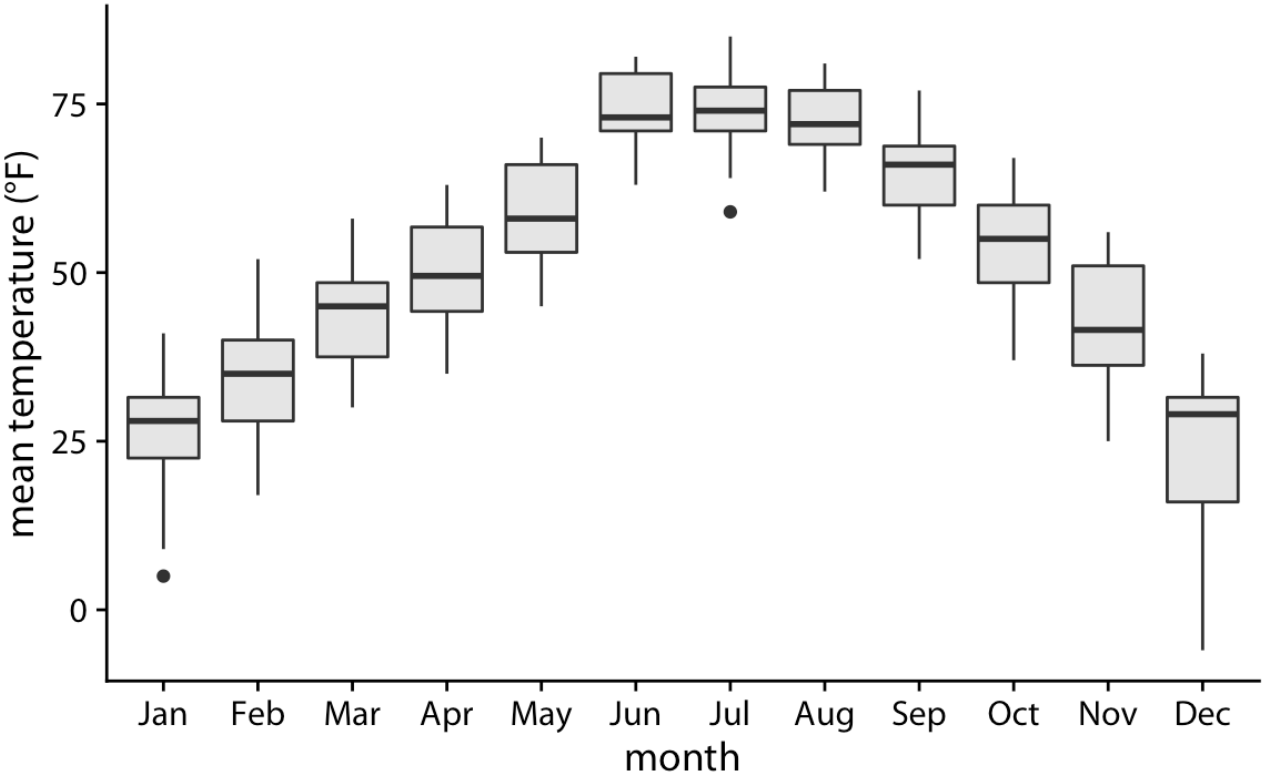
Boxplot

- A boxplot divides the data into quartiles and visualizes them in a standardized manner



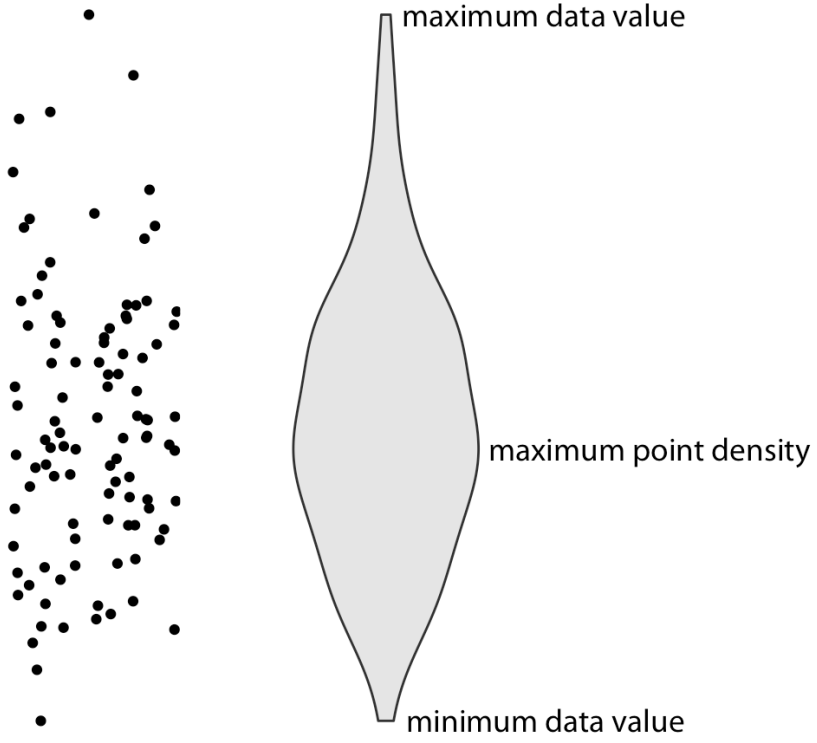
*Fence: 1.5 * inter quantile range away from the box*

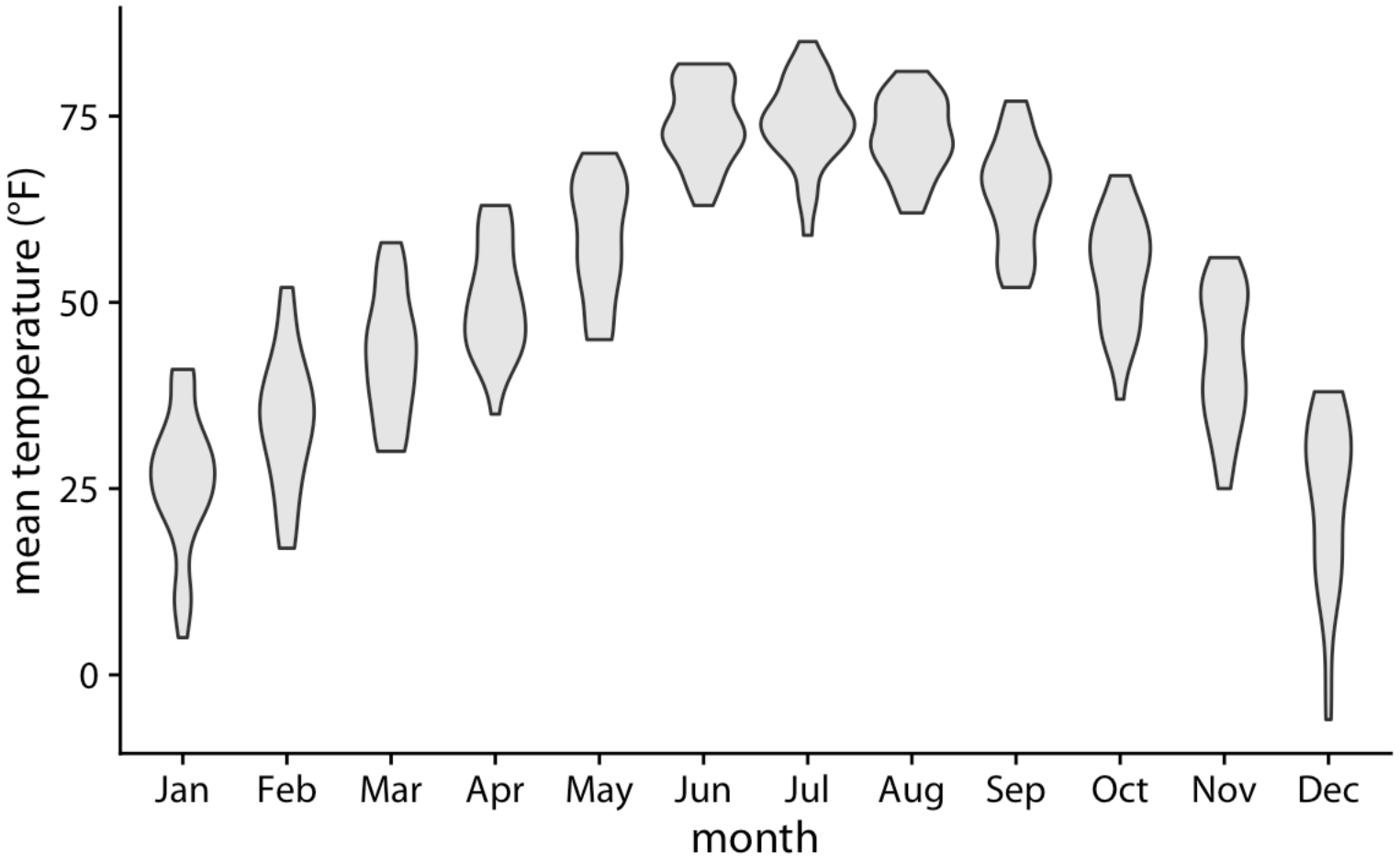
- work well when plotted next to each other to visualize many distributions at once



violin plots

- Equivalent to the density estimates
- Rotated by 90 degrees and then mirrored

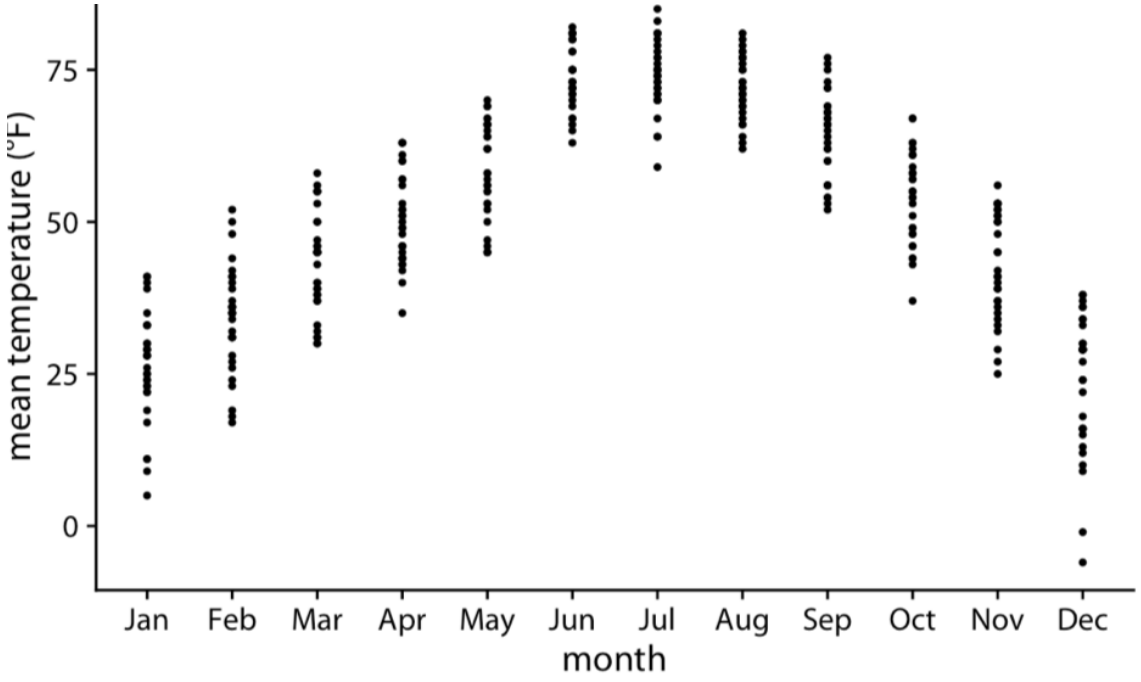



mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

- Before using violins to visualize distributions, verify that you have sufficiently many data points in each group to justify showing the point densities as smooth lines.
- Because violin plots are derived from density estimates, they can generate the appearance that there is data where none exists, or that the data set is very dense when actually it is quite sparse.

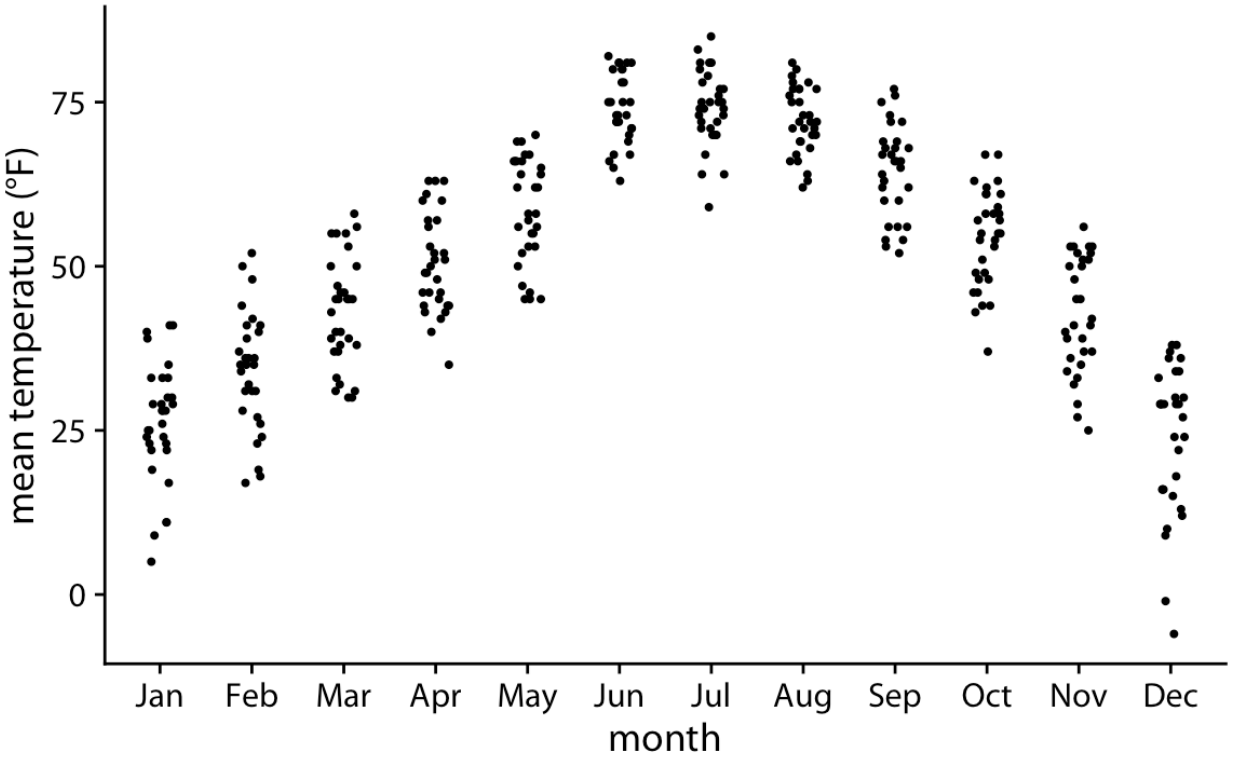
Strip chart

- Plot all the individual data points of the response variable directly



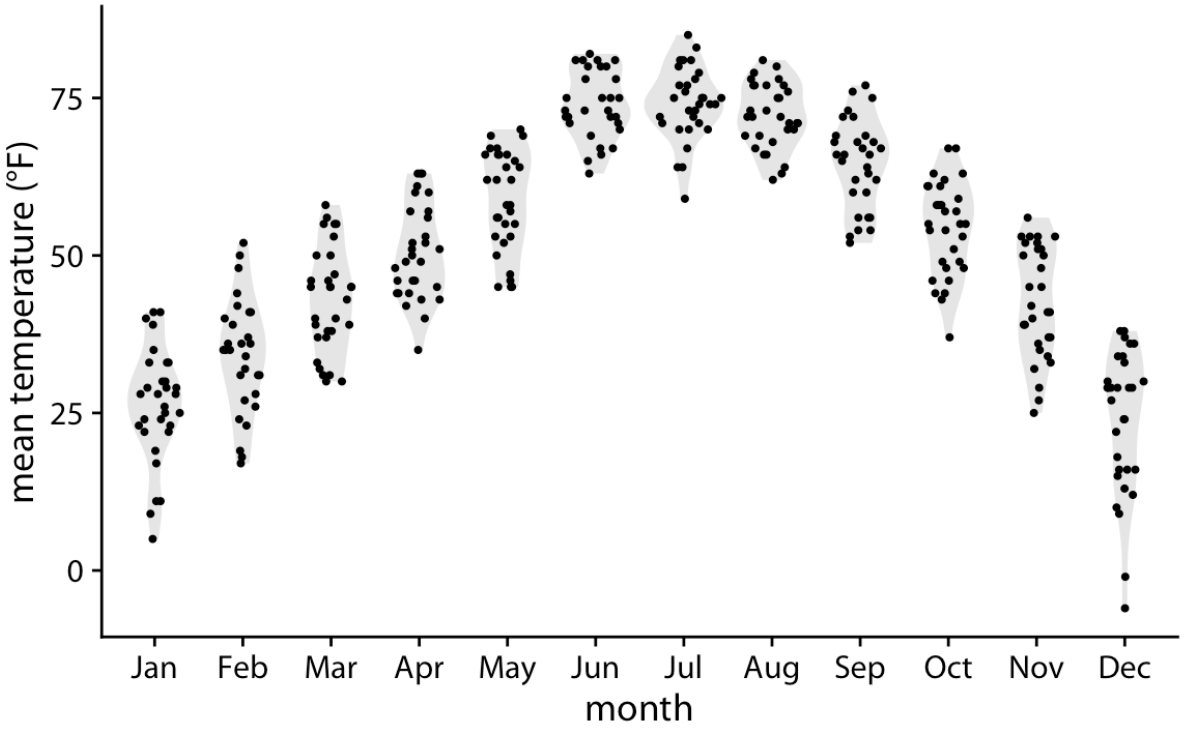
Note: don’t plot too many points on top of each other

- *Jittering:* A simple solution to overplotting is to spread out the points somewhat along the *x* axis, by adding some random noise in the *x* dimension



Sina plot

- A hybrid between a violin plot and jittered points

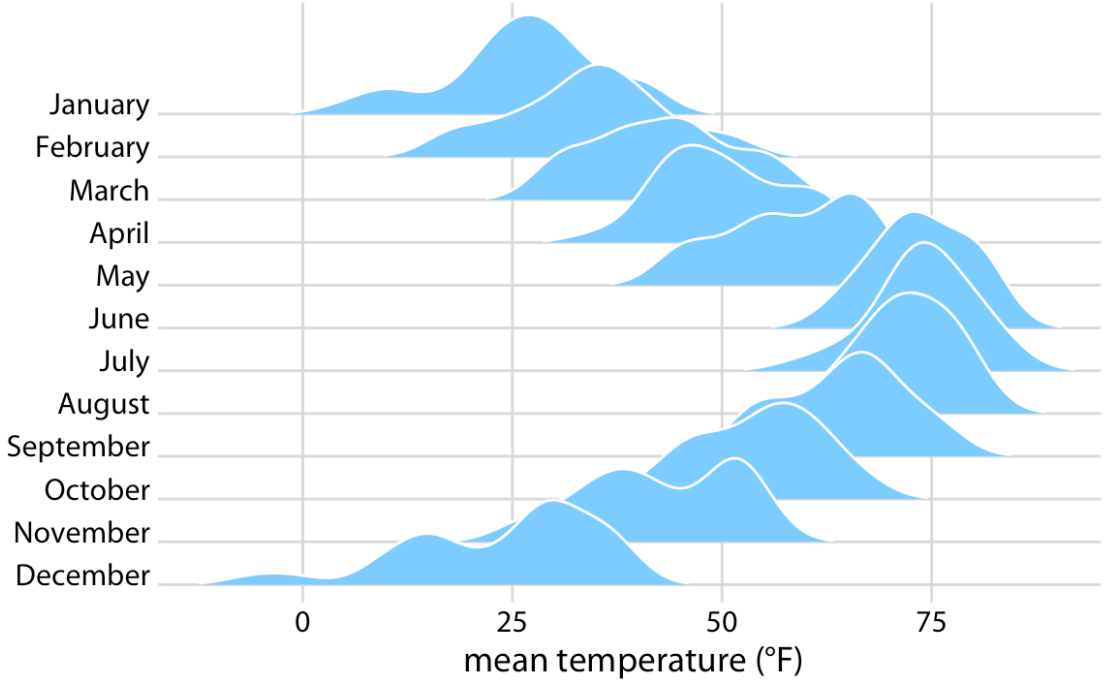


Visualizing multiple distributions along the horizontal axis

- Ridgeline plot
  - Standard: use density estimation
  - use histograms

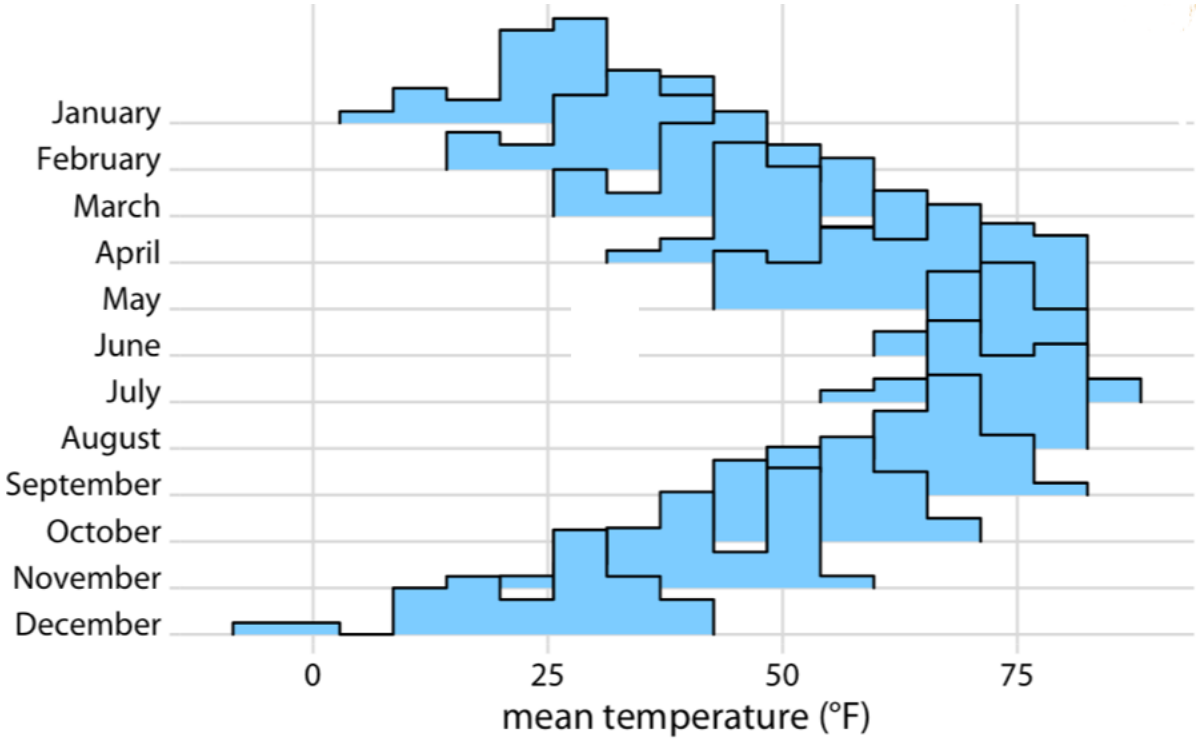
Ridgeline plot - Visualizing multiple distributions along the horizontal axis

- Closely related to the violin plot
- Density estimates are shown alongside the grouping variable.

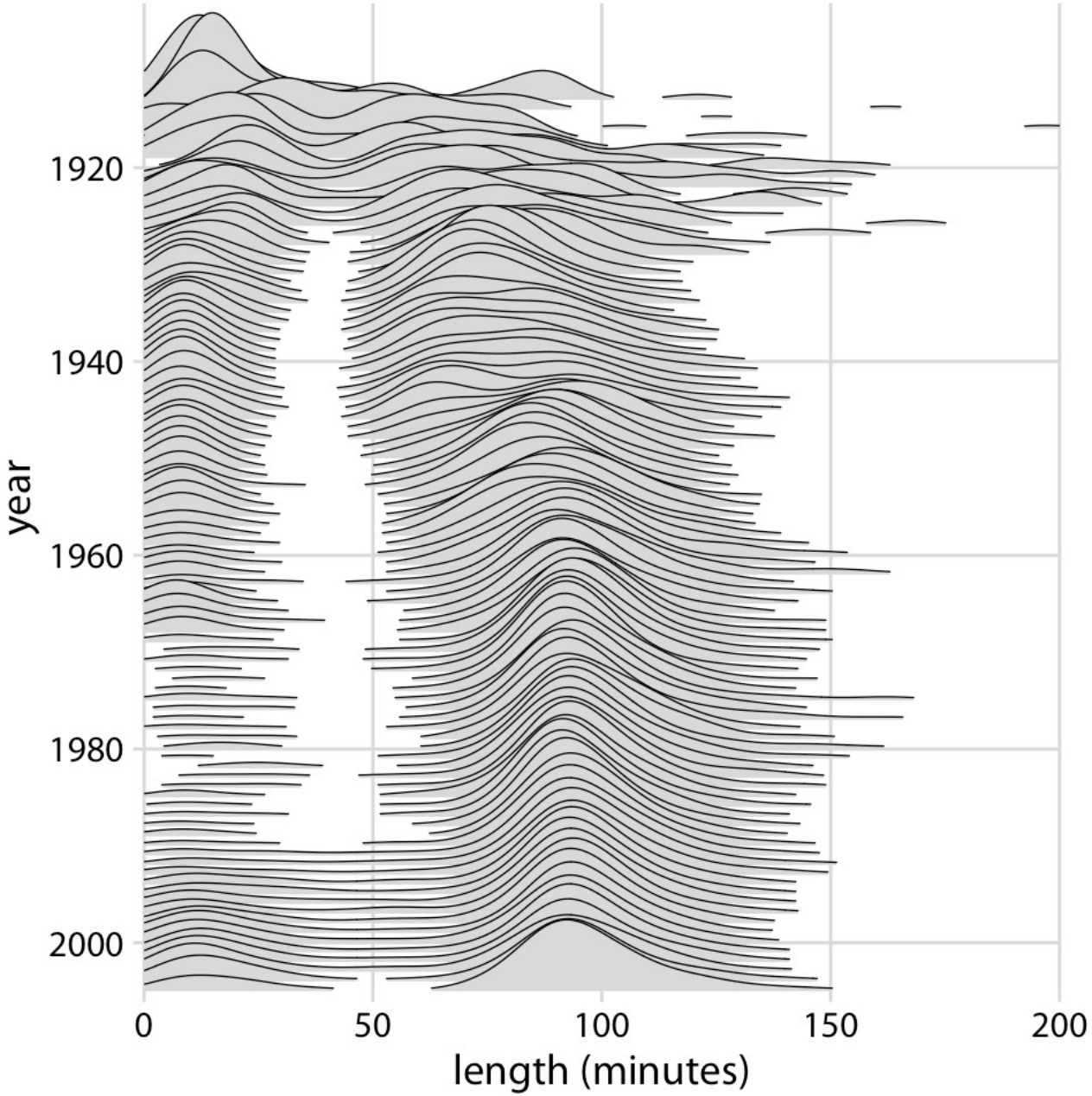


Another Example in python: https://seaborn.pydata.org/examples/kde_ridgeplot.html

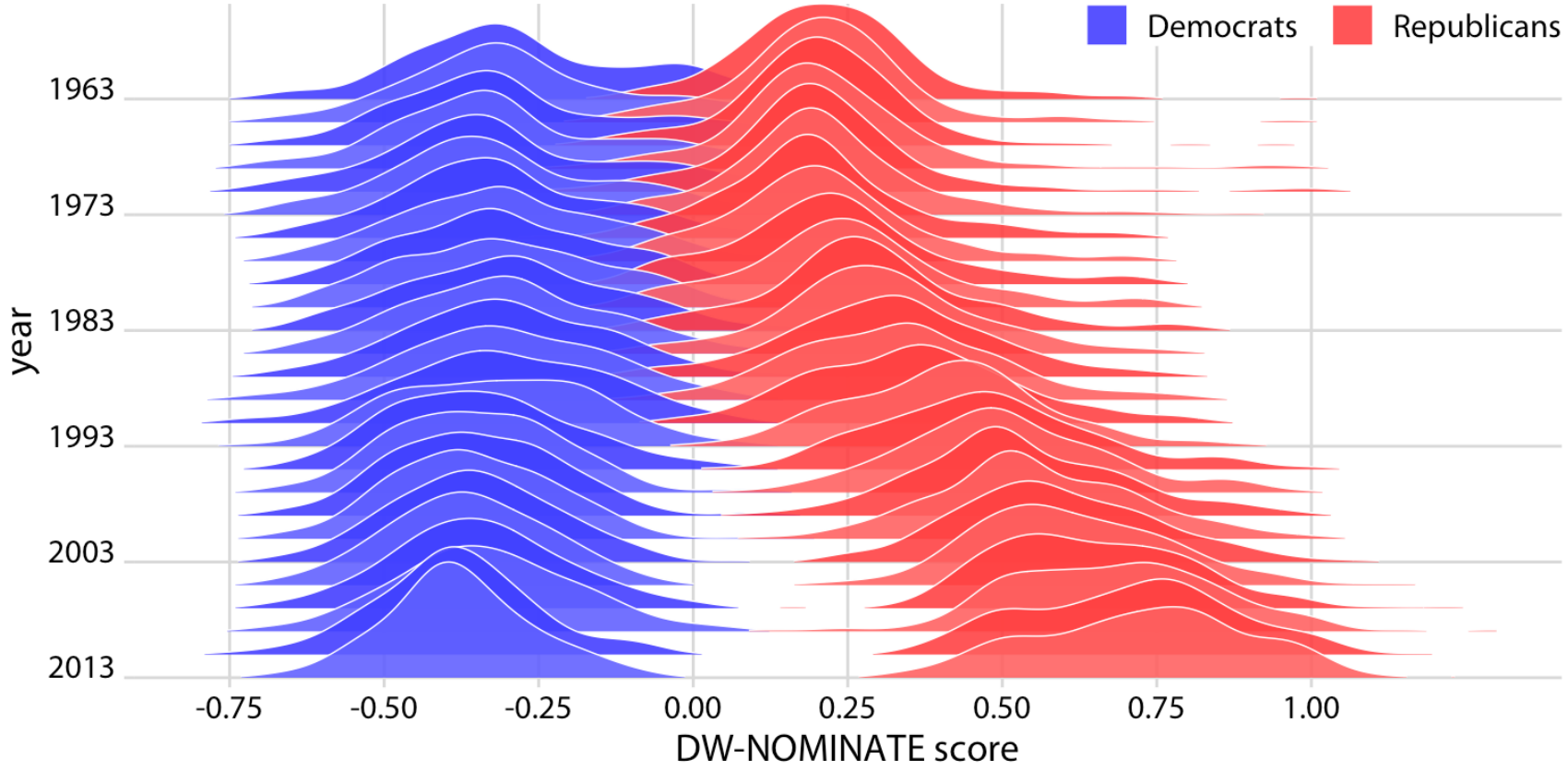
- Ridgeline plot use histograms



- Note: Easy to be overlapped


year
1920
1940
1960
1980
2000
0
50
100
150
200
length (minutes)

- Compare two trends over time



Voting patterns in the U.S.

- Visualize multiple distributions at the same time
  - Boxplot
  - Violin plot
  - Strip chart
  - Sina plot
- Design:
  - Reproducibility and repeatability
  - Data exploration VS. data presentation
  - Separation of content and design

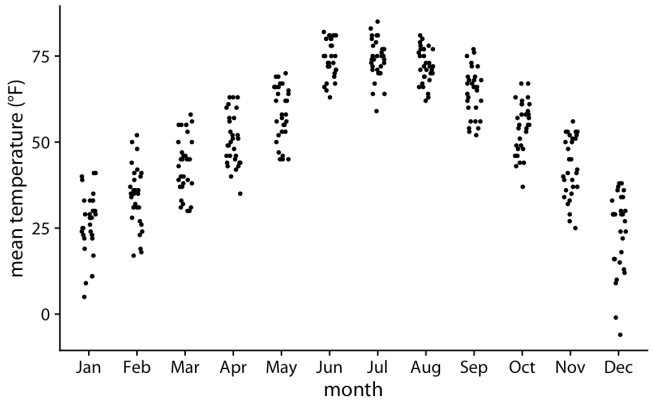
Reproducibility and repeatability

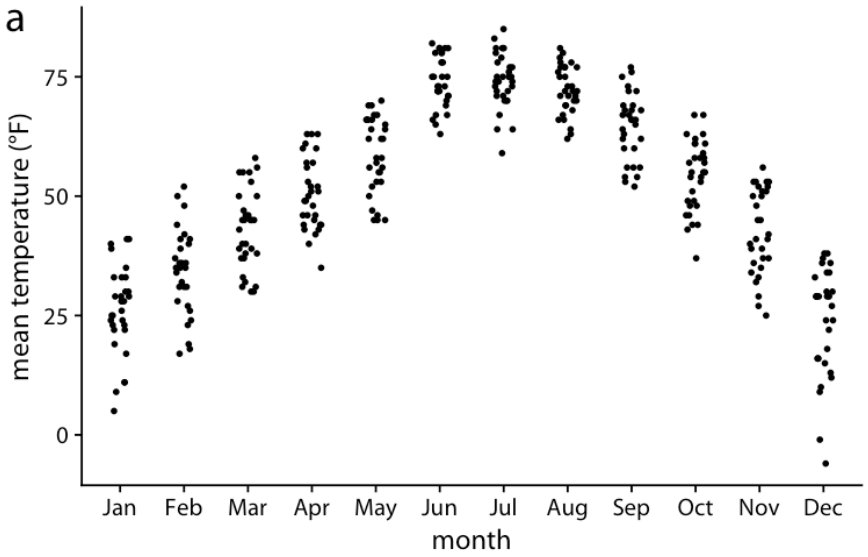
- In scientific experiments, we refer to work as repeatable if very **similar or identical** measurements can be obtained by the same person repeating the exact same measurement procedure on the same equipment.
- Work is reproducible if the overarching scientific finding of the work will remain **unchanged** if a different research group performs the same type of study.

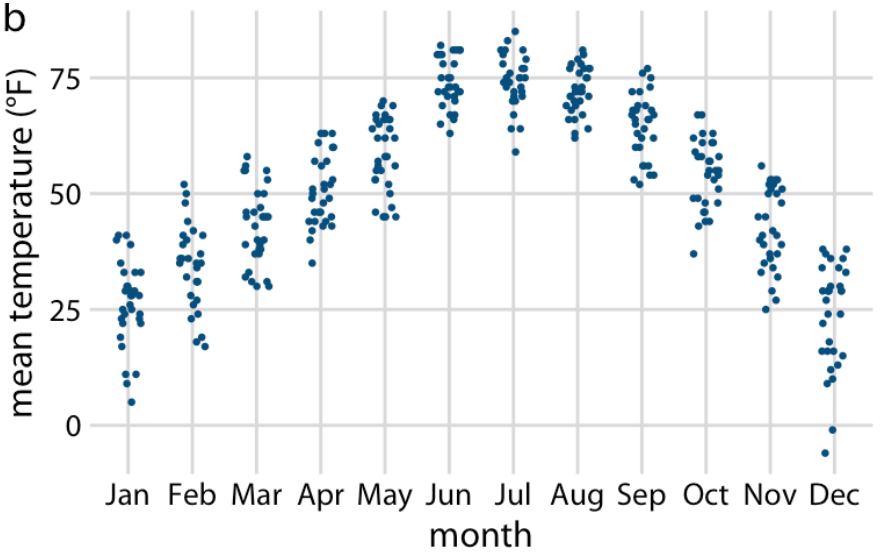
Reproducibility and repeatability - in data visualization

- A visualization is repeatable, if it is possible to recreate the exact same visual appearance, down to the last pixel, from the raw data.

- A visualization is reproducible, if the plotted data are available and any data transformations that may have been applied are exactly specified.


mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month


a
mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month


b
mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

Data exploration versus data presentation

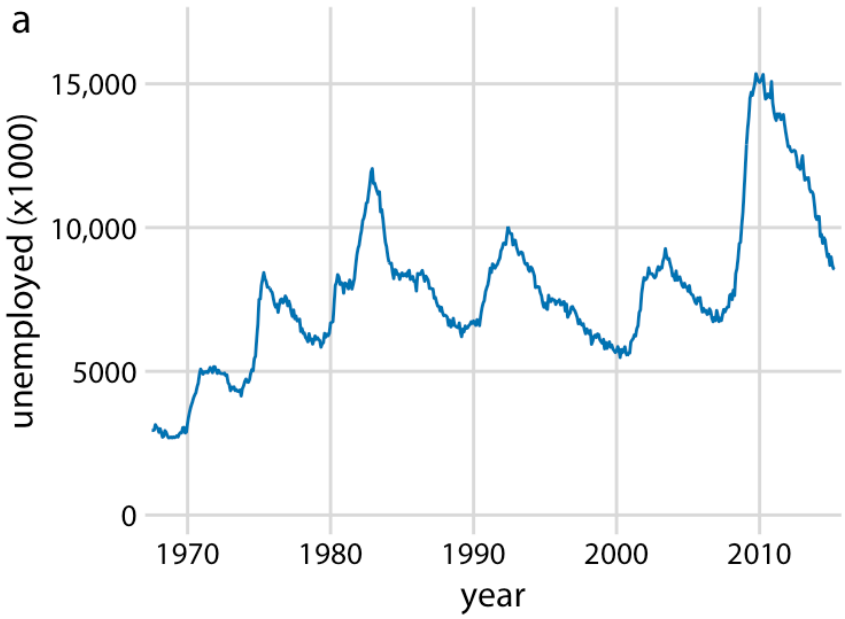
- Two distinct phases of data visualization with different requirements: data exploration and data presentation
- Data exploration: Whenever you start working with a new dataset, you need to look at it from different angles and try various ways of visualizing it, just to develop an understanding of the dataset's key features.
- Data presentation: The key objective in this phase is to prepare a high-quality figure

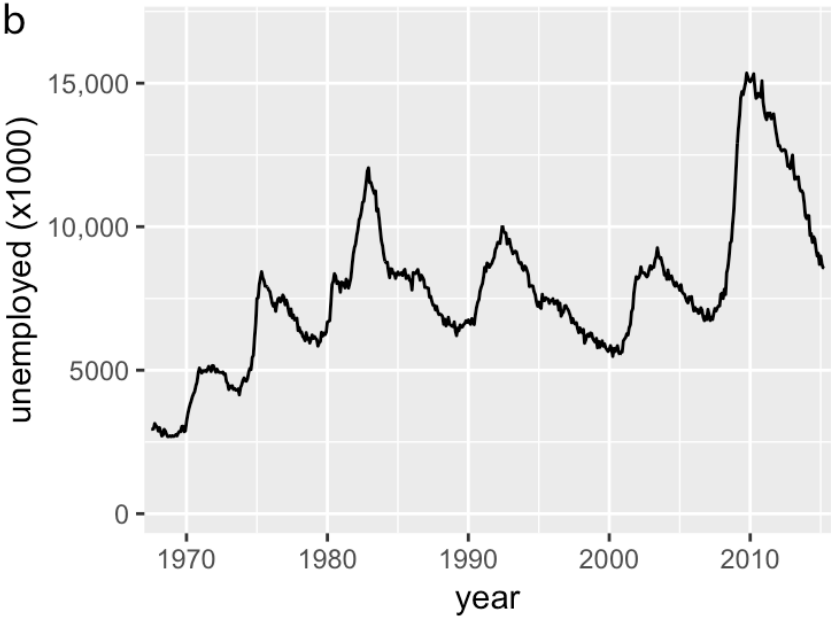
- Data exploration
  - determined how exactly we want to visualize our data, what data transformations we want to make, and what type of plot to use
  - speed and efficiency are of the essence.
  - whether the figures you make look appealing is secondary
  - what is critical is how easy it is for you to change how the data are shown
  - provide a wide range of different visualization options within a single coherent framework (not with many programmatic figure generation tools!)

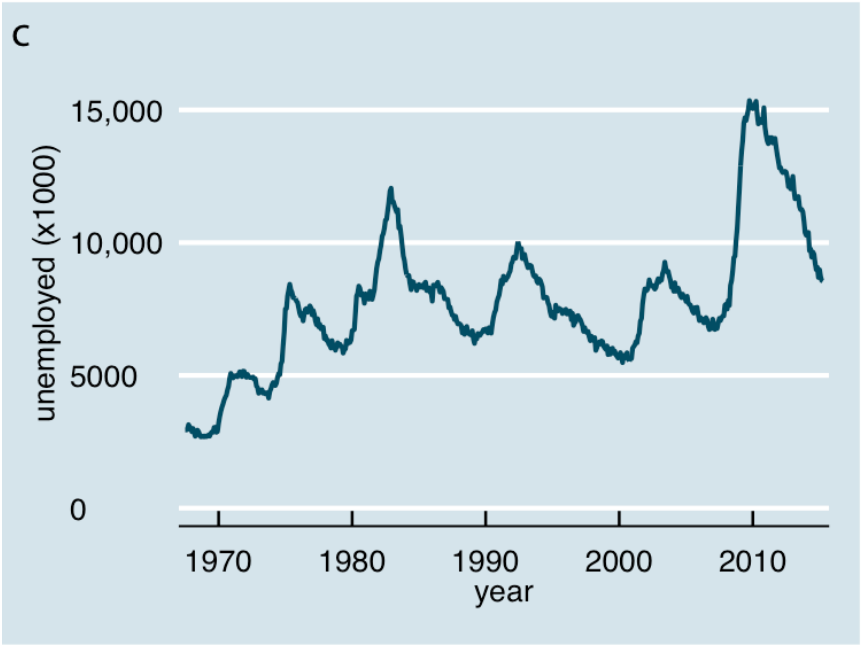
- data presentation
  - prepare a high-quality figure
  - can finalize the figure using same software platform we used for initial exploration
  - can switch platform to one that provides us finer control over the final product
  - can produce a draft figure with a visualization software and then manually post-process with an image manipulation or illustration program
  - can manually redraw the entire figure from scratch
  - reproducing and repeating

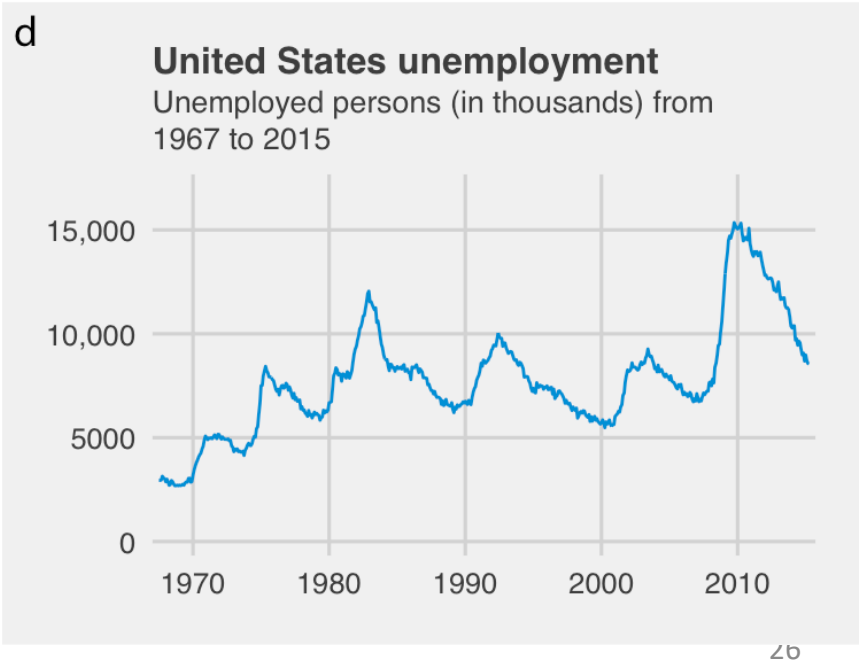
Separation of content and design

- Content: the specific data set, the data transformations applied (if any), the specific mappings from data onto aesthetics, the scales, the axis ranges, and the type of plot (scatter plot, line plot, bar plot, boxplot, etc.).
- Design, describes features such as the foreground and background colors, font specifications (e.g. font size, face, and family), symbol shapes and sizes, the placement of legends, axis ticks, axis titles, and plot titles, and whether or not the figure has a background grid.


a
15,000
10,000
5000
0
unemployed (x1000)
1970
1980
1990
2000
2010
year


b
15,000
10,000
5000
0
unemployed (x1000)
1970
1980
1990
2000
2010
year


c
15,000
10,000
5000
0
unemployed (x1000)
1970
1980
1990
2000
2010
year


d
United States unemployment
Unemployed persons (in thousands) from
1967 to 2015
15,000
10,000
5000
0
1970
1980
1990
2000
2010