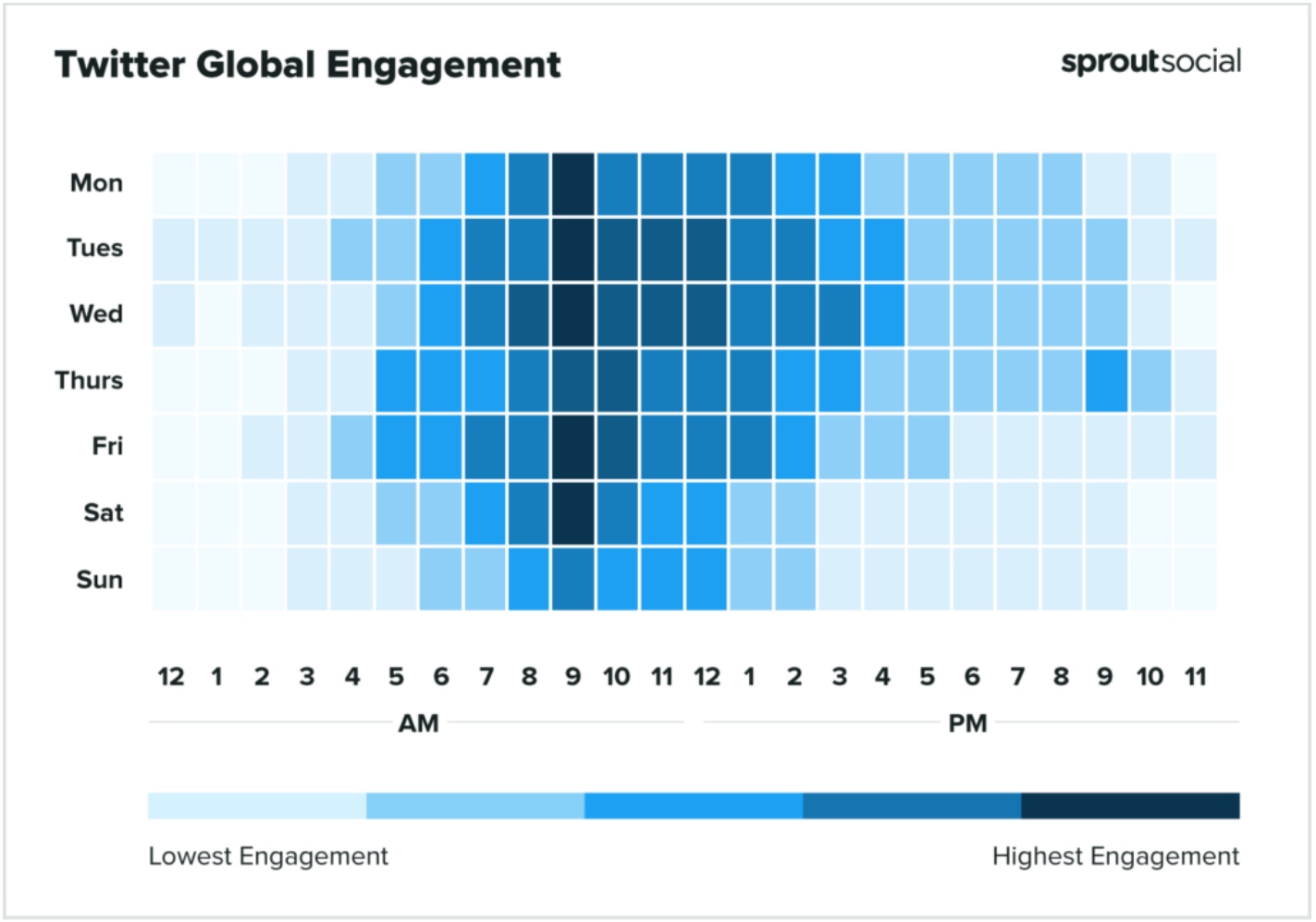
Exploratory Data Analysis & Visualization
Temporal Data

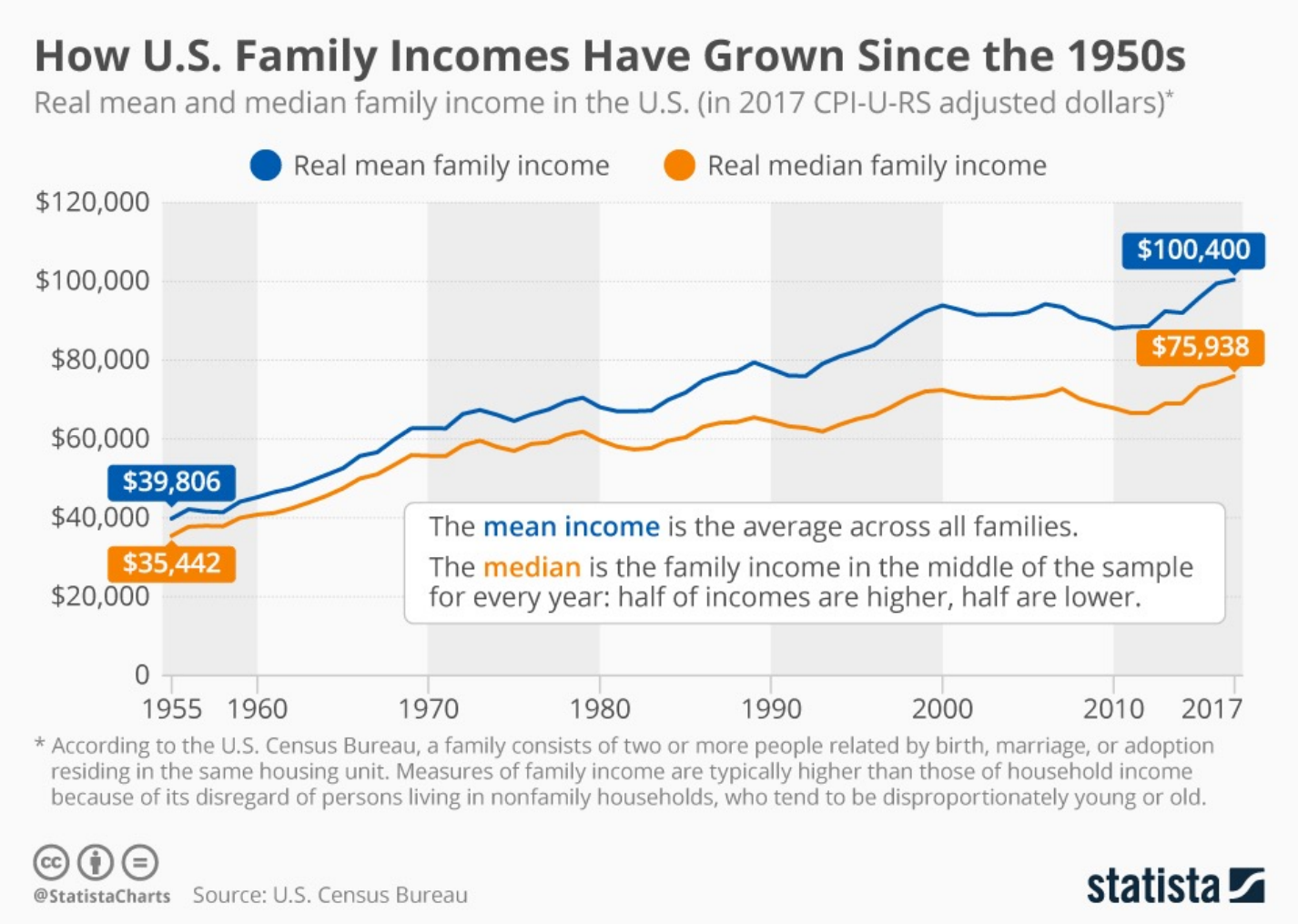
Ben Winjum

Example Visualizations from Discussions



Chi

Example Visualizations from Discussions



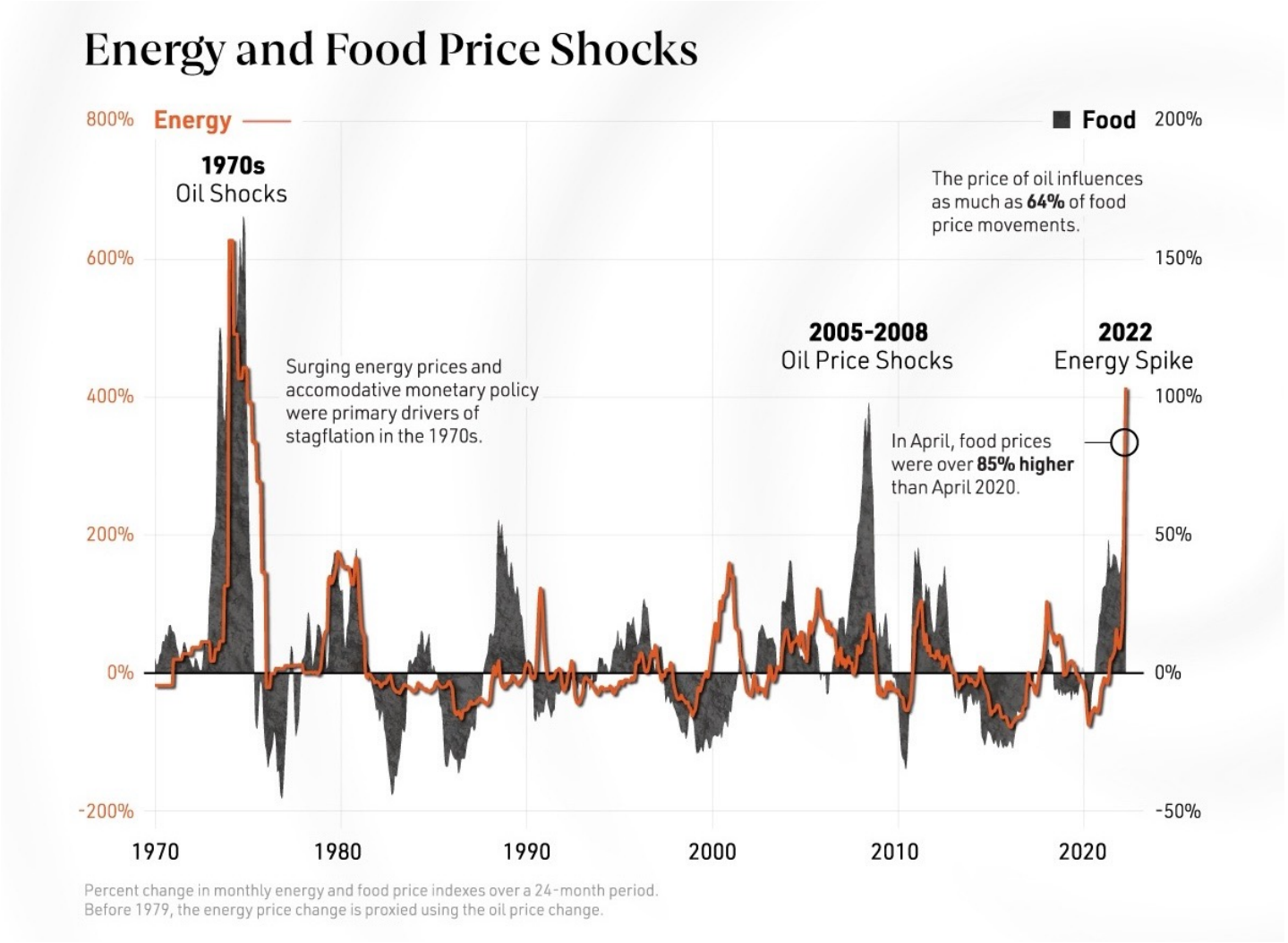
Tim

Example Visualizations from Discussions

Miral

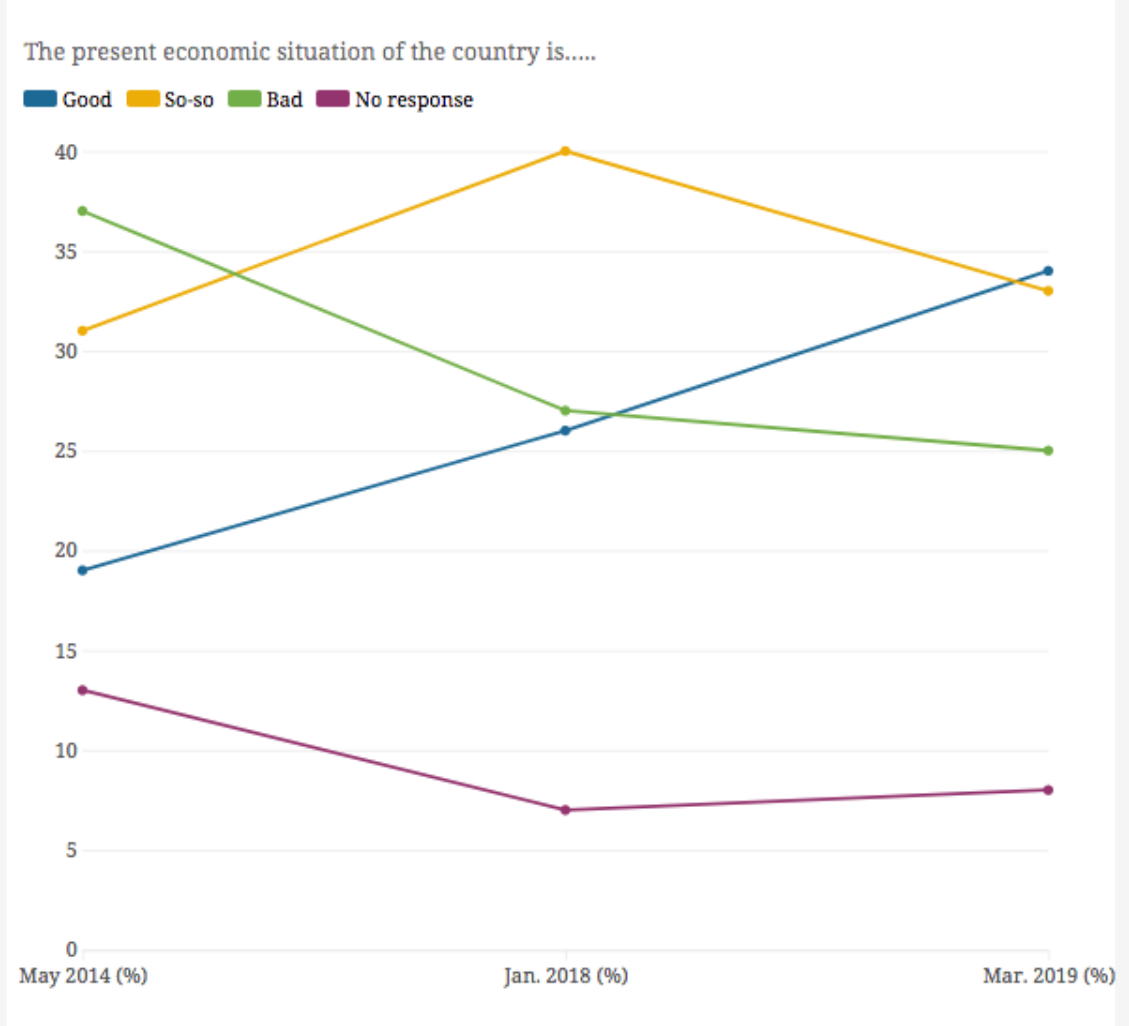
Example Visualizations from Discussions

Boyan



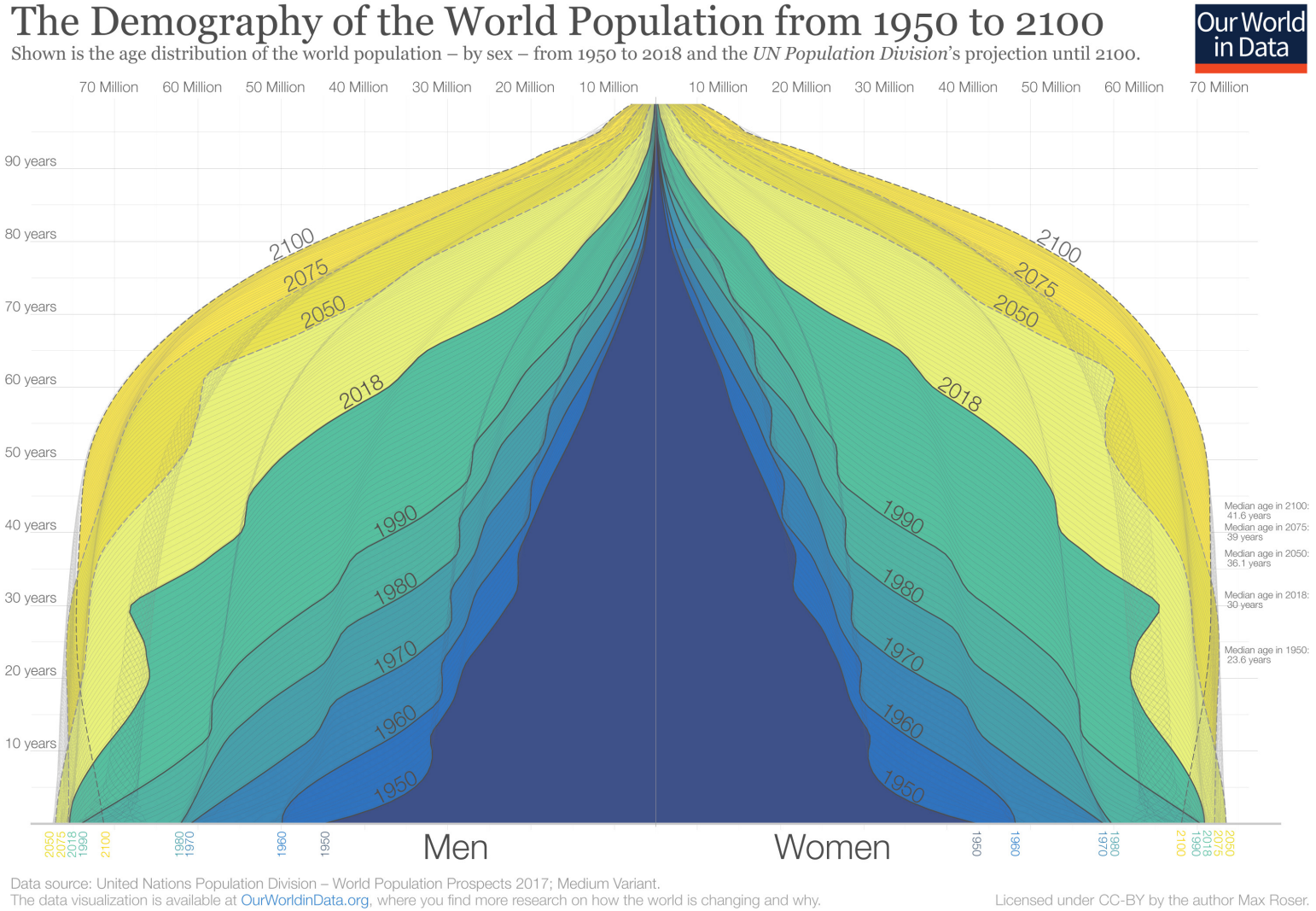
Today: Time Series

Example Time Series from Prior Week's Discussions



Divya

Example Time Series from Prior Week's Discussions

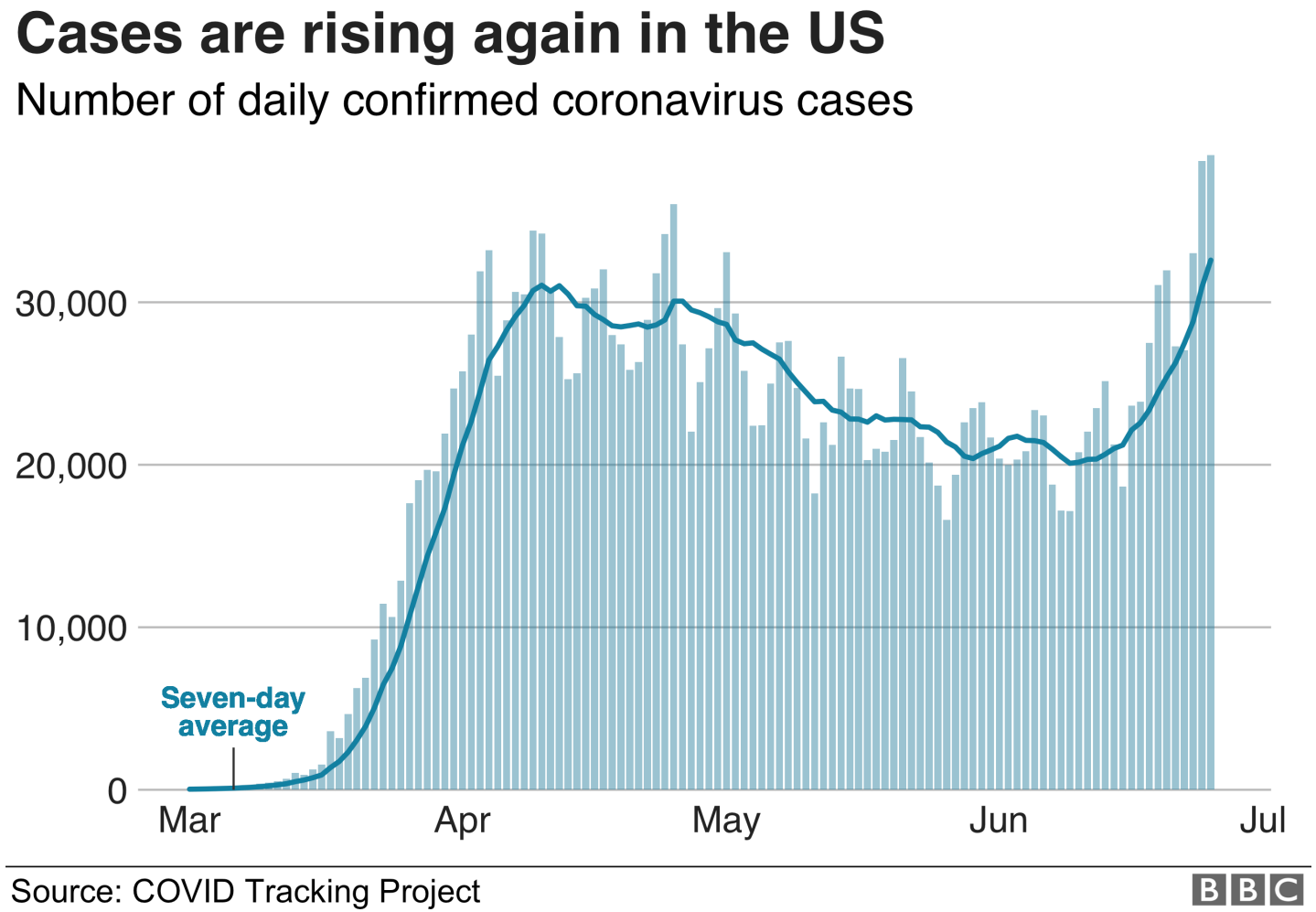


Tristan

https://ourworldindata.org/age-structure

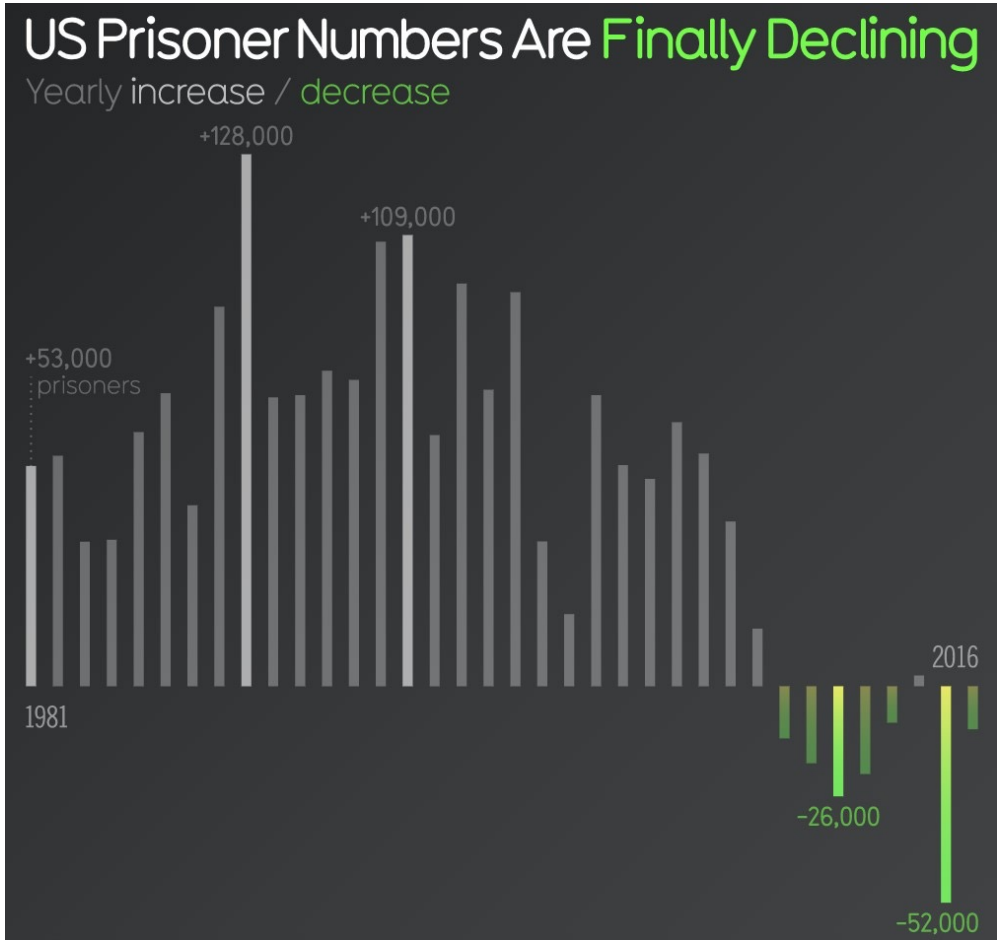
Example Time Series from Prior Week's Discussions

Jing-Wen



Example Time Series from Prior Week's Discussions

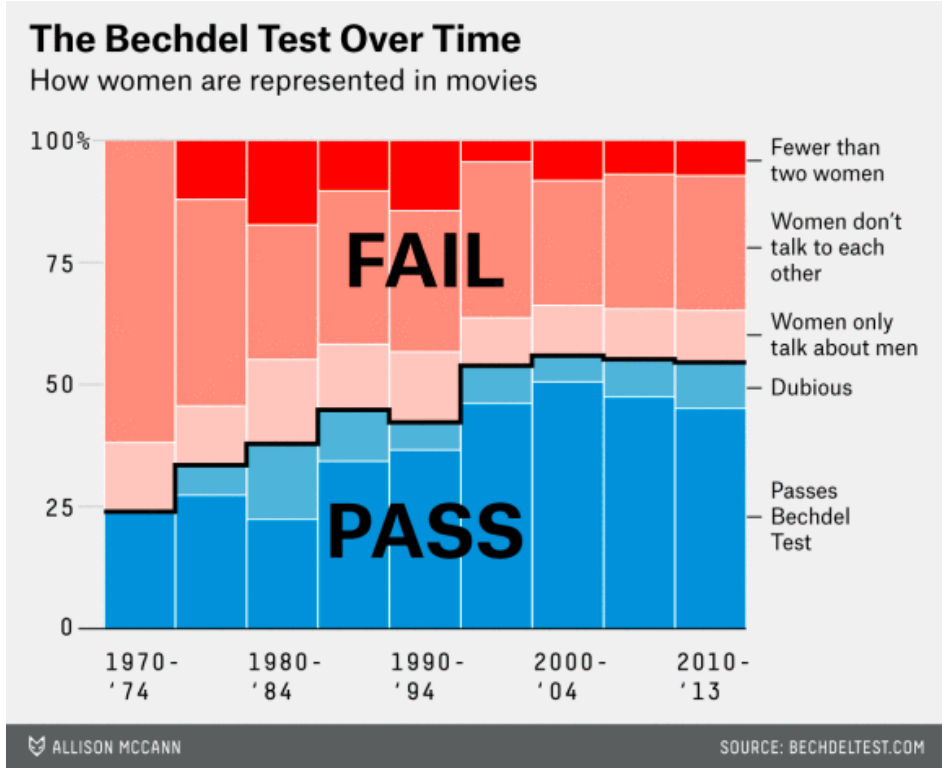
Charles



Prior plots of other lectures showing Time Series

Stacked bars

- Not good for more than a couple categories
- Here 2-3 groups



https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

Stacked bars

- Effective here for portions over time

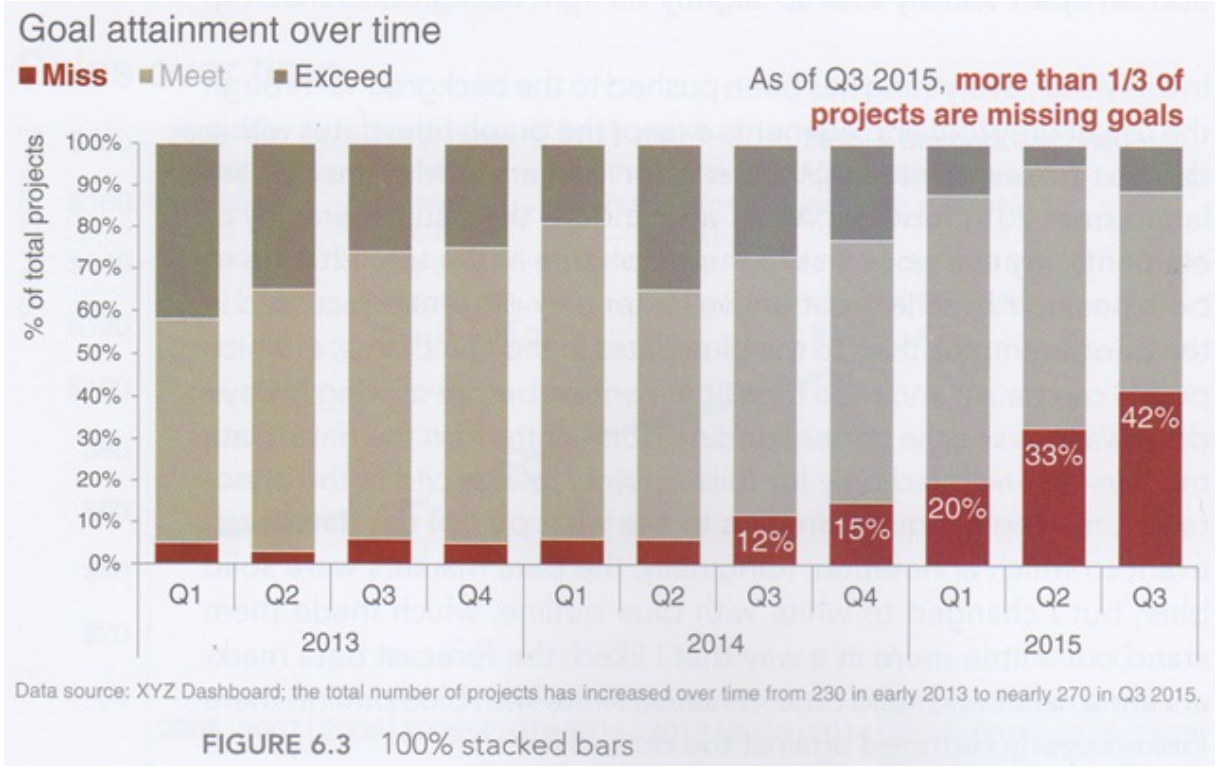
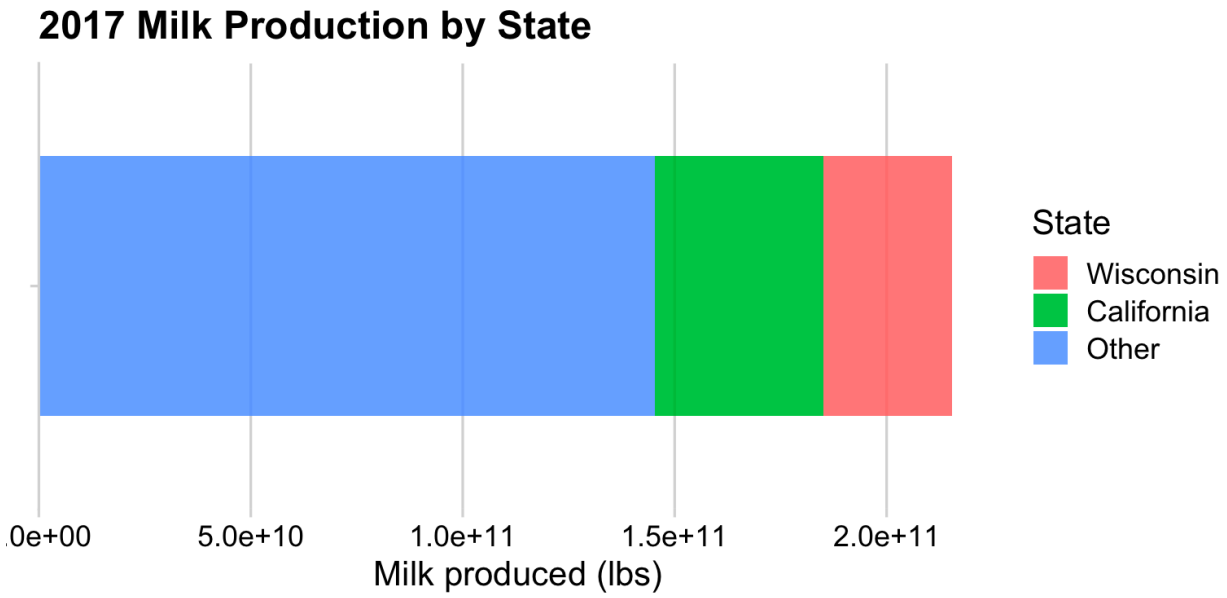


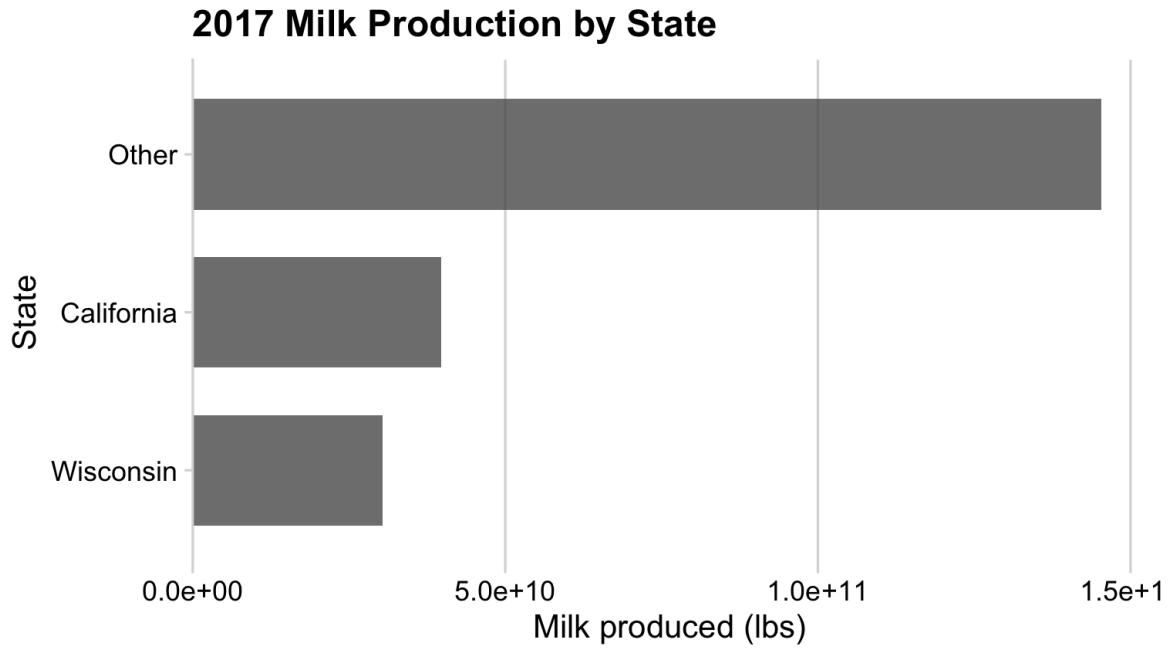
FIGURE 6.3 100% stacked bars

https://www.perceptualedge.com/blog/?p=2239

Dodged bars
(the bars are getting out of each other's way)

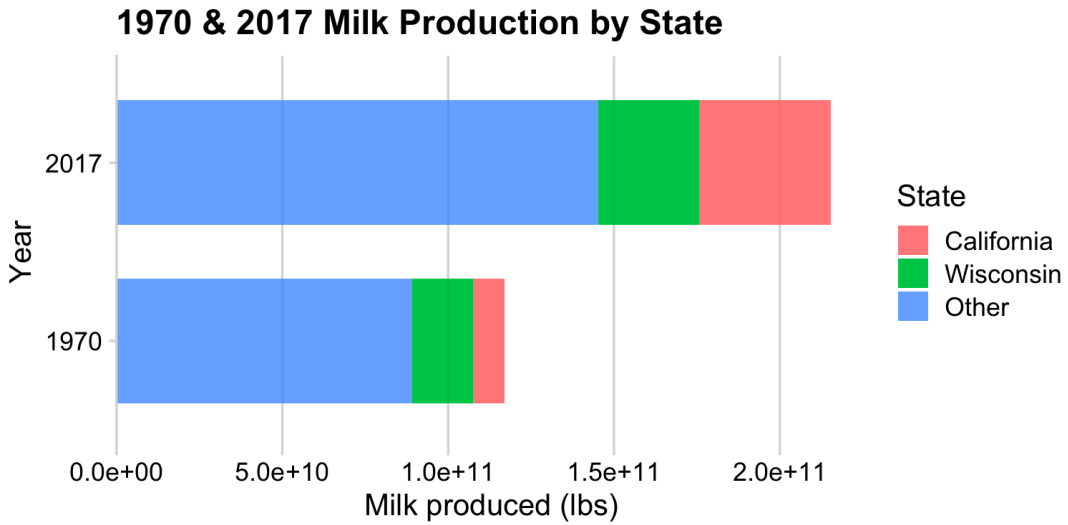
- Stacked : better for showing single part-to-whole comparison



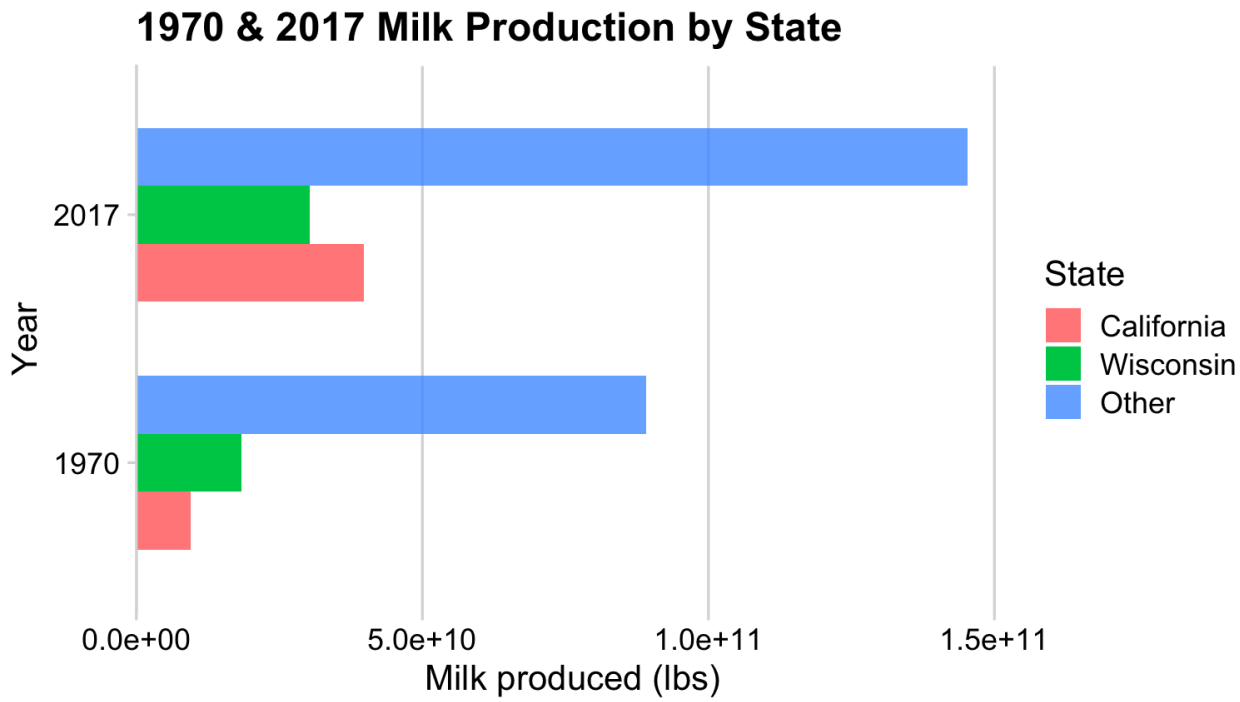


Dodged bars

- Dodged : better for comparing individual components

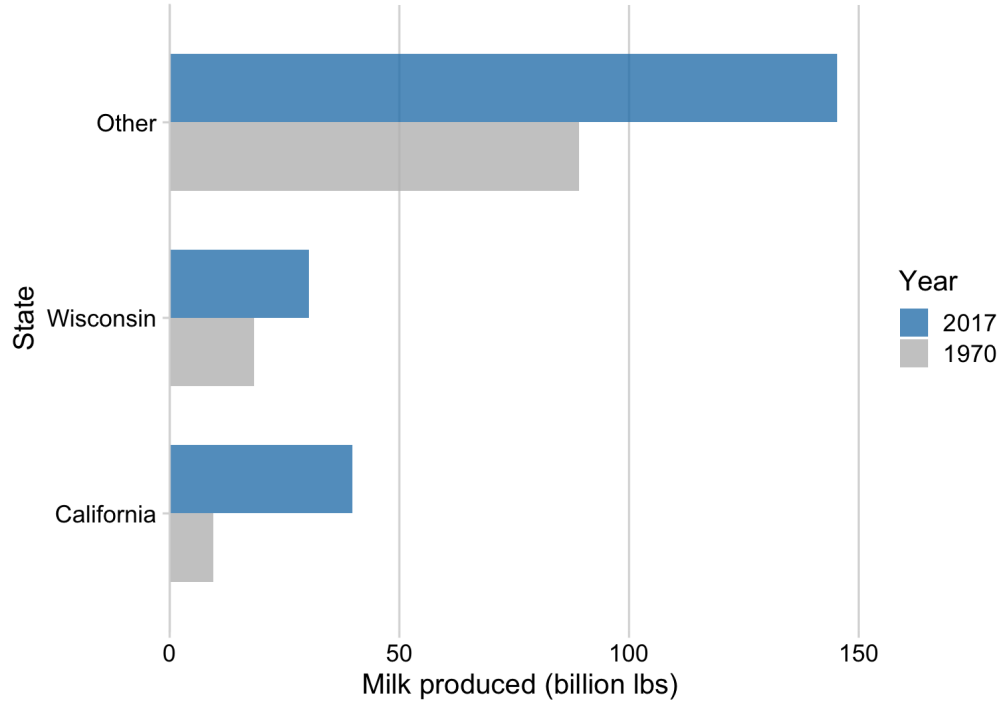


Better for **comparing total**

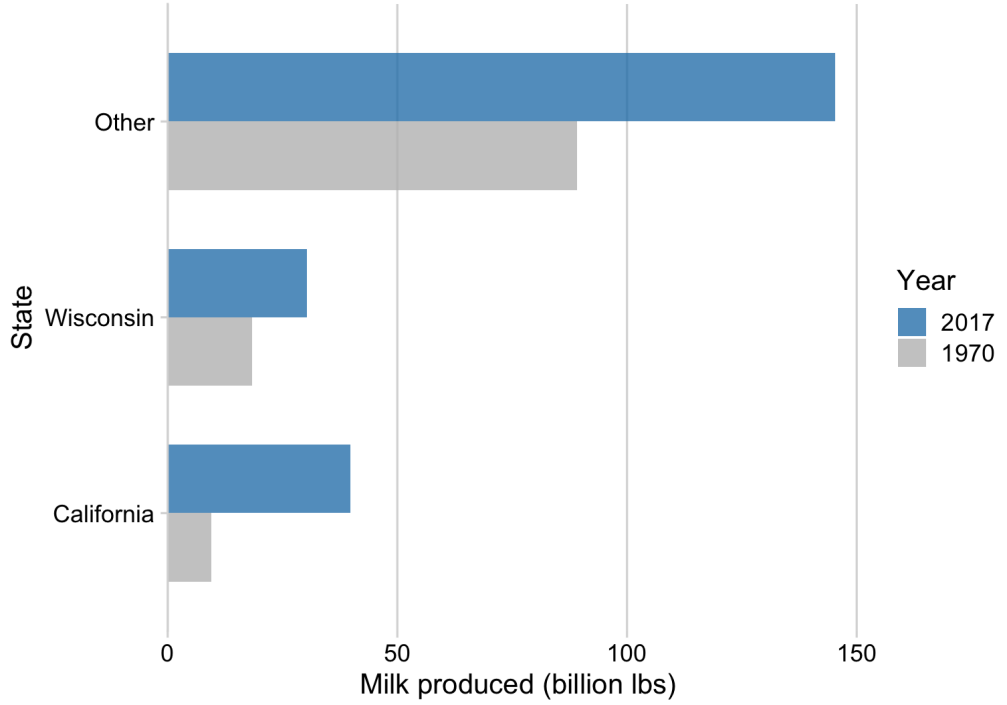


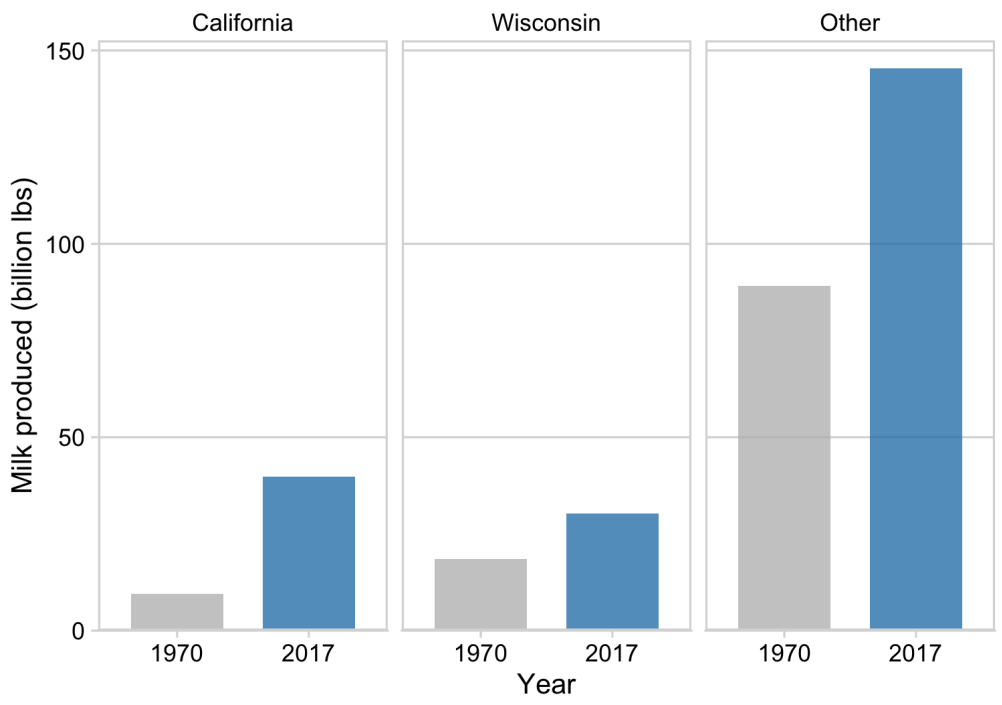
Better for **comparing parts**

Dodged bars: useful for comparing two things

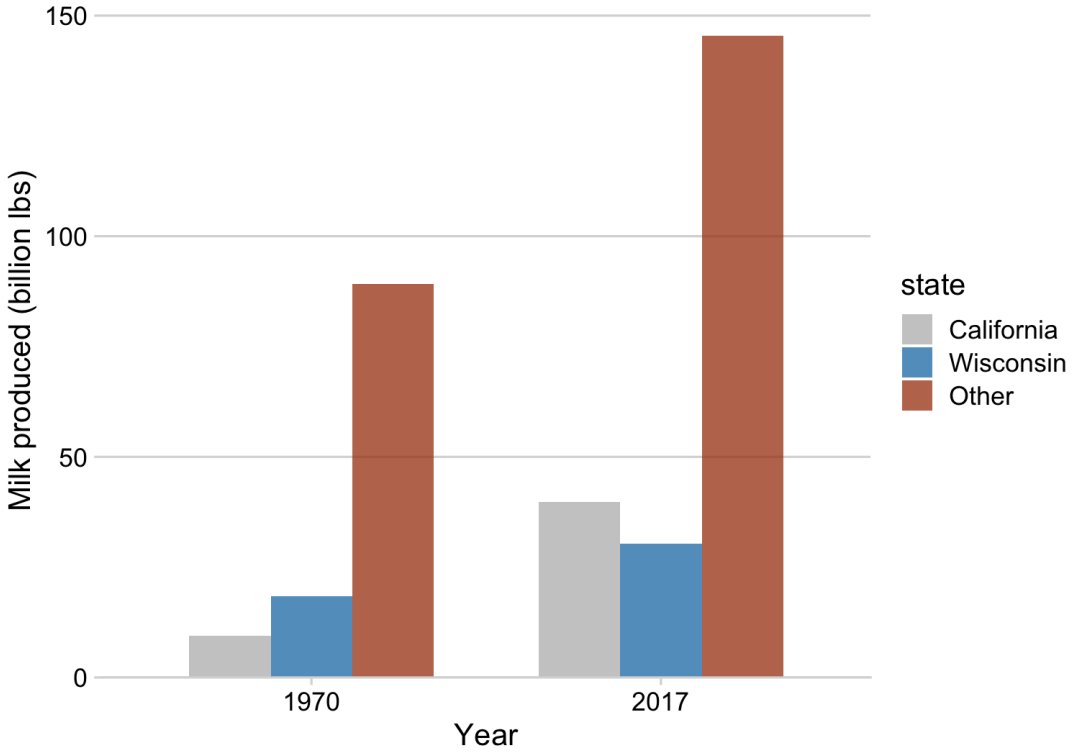


Dodged bars: facets can be helpful

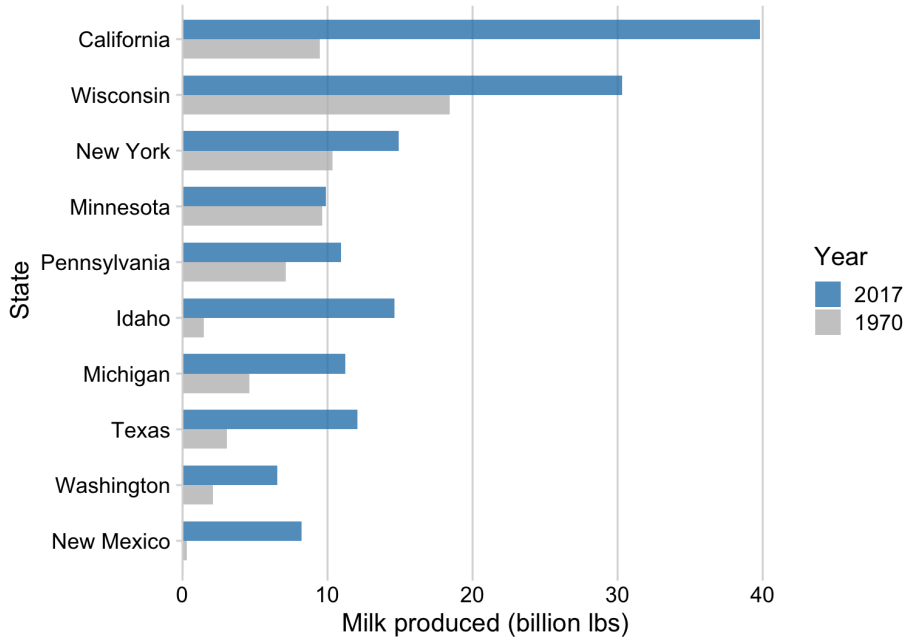




Dodged bars: more than two starts to get confusing



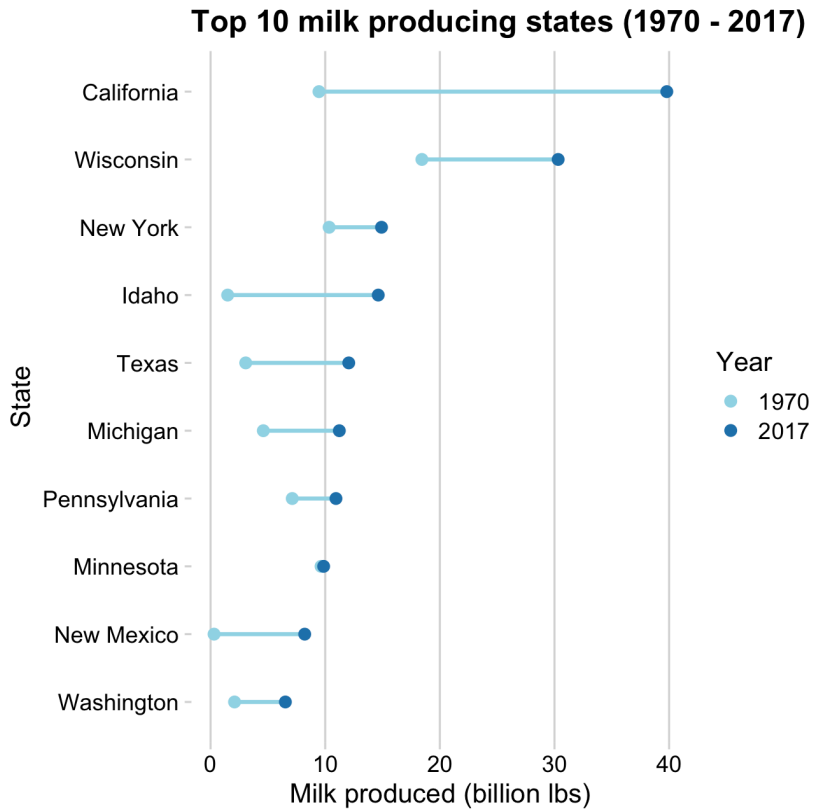
2 years, but more than 2 items per group



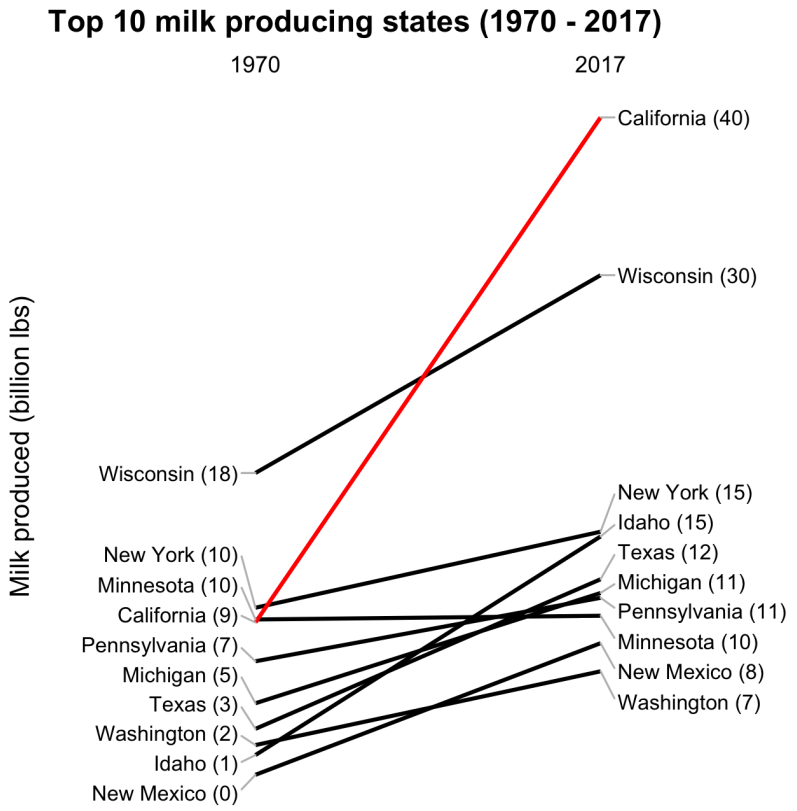
2 years, but more than 2 categories

Comparisons across more than two categories

- Dumbbell: good for comparing change in magnitudes
- Slope: good for comparing change in ranking, and how one is different than another



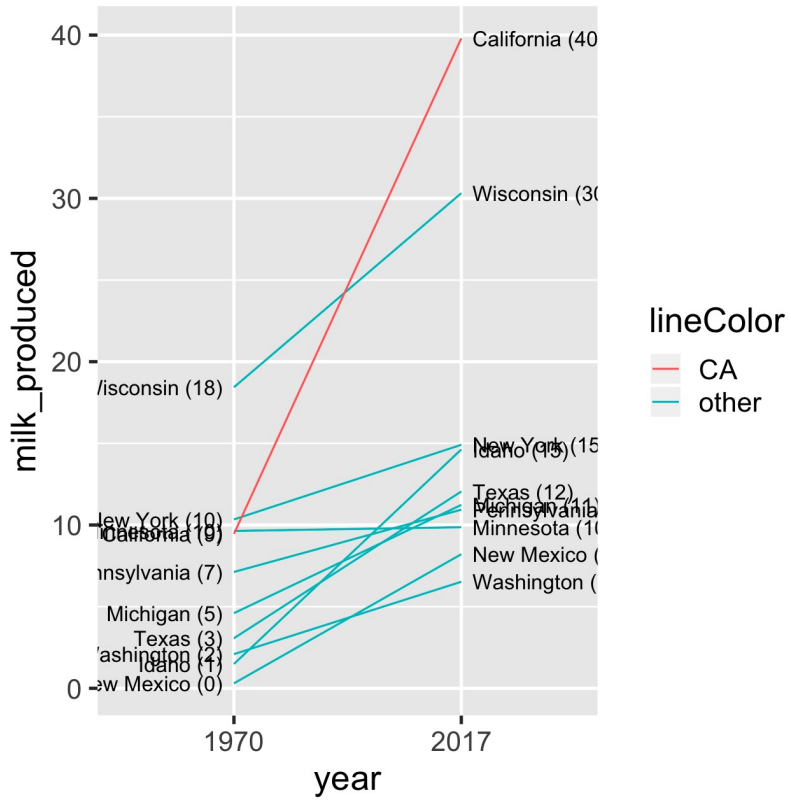
Dumbbell chart



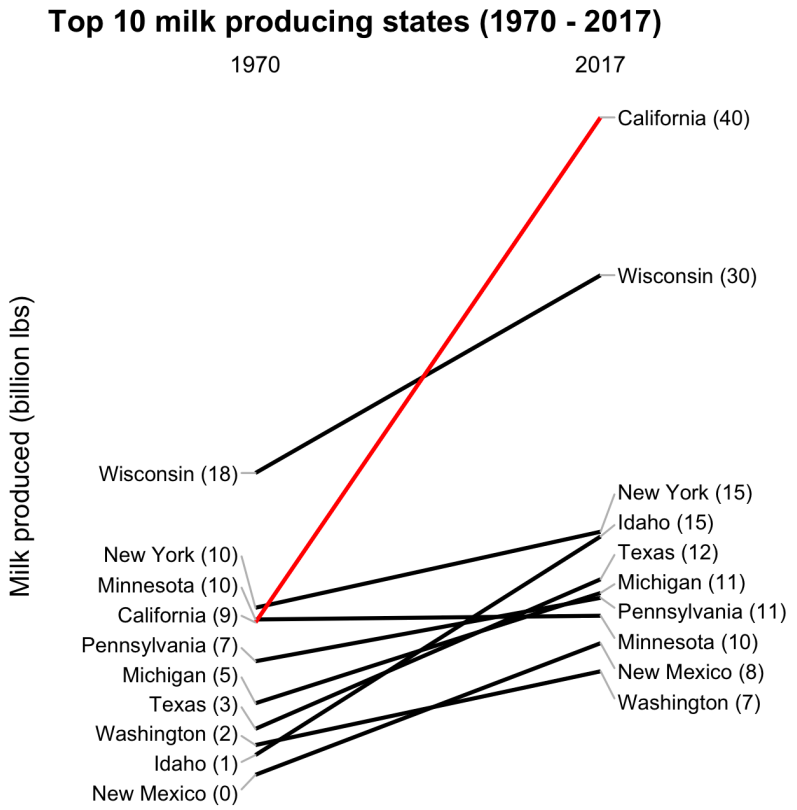
Slope chart

Comparisons across more than two categories

- Remember to label clearly

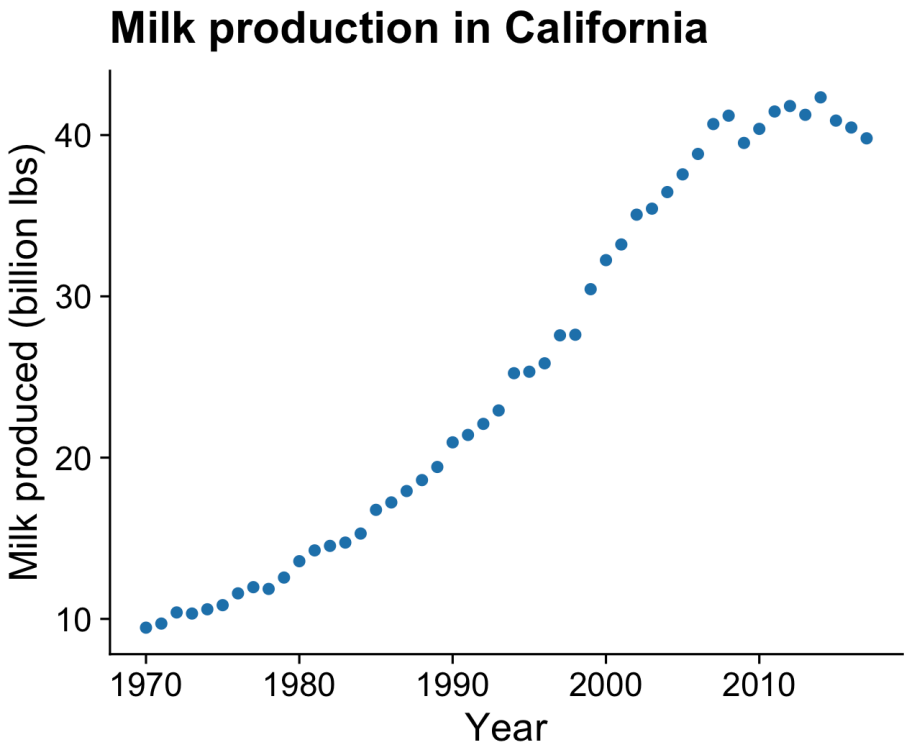


Dumbbell chart

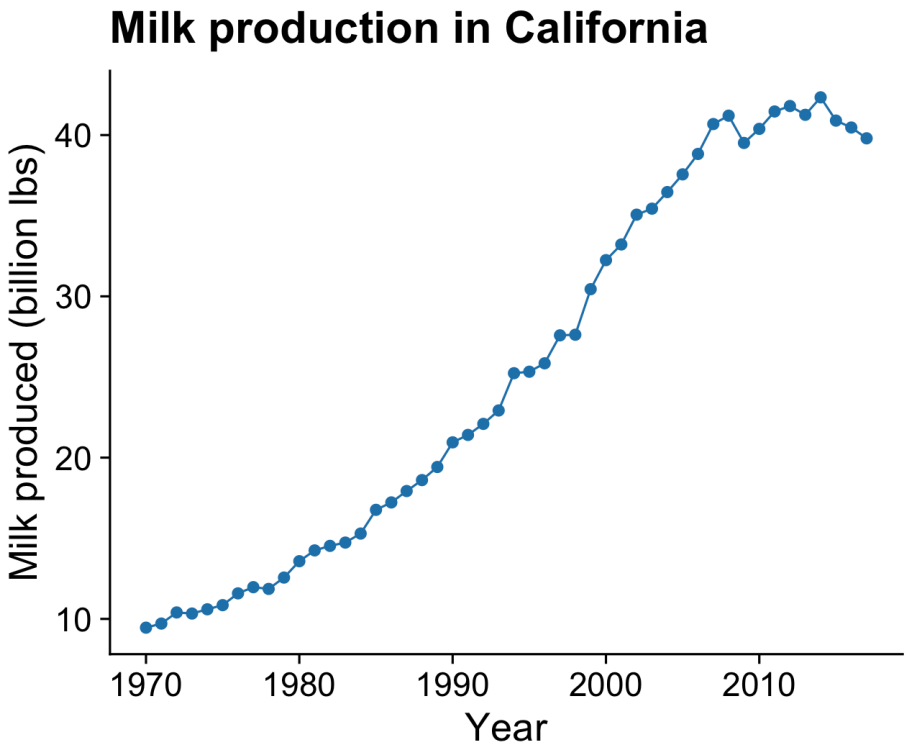


Slope chart

Trends

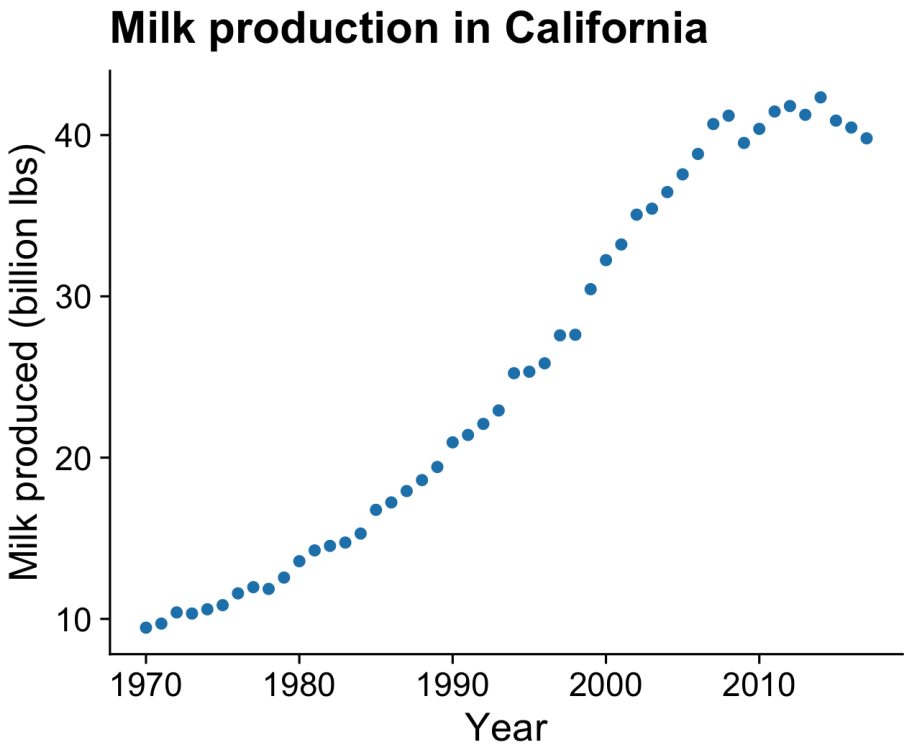


Points

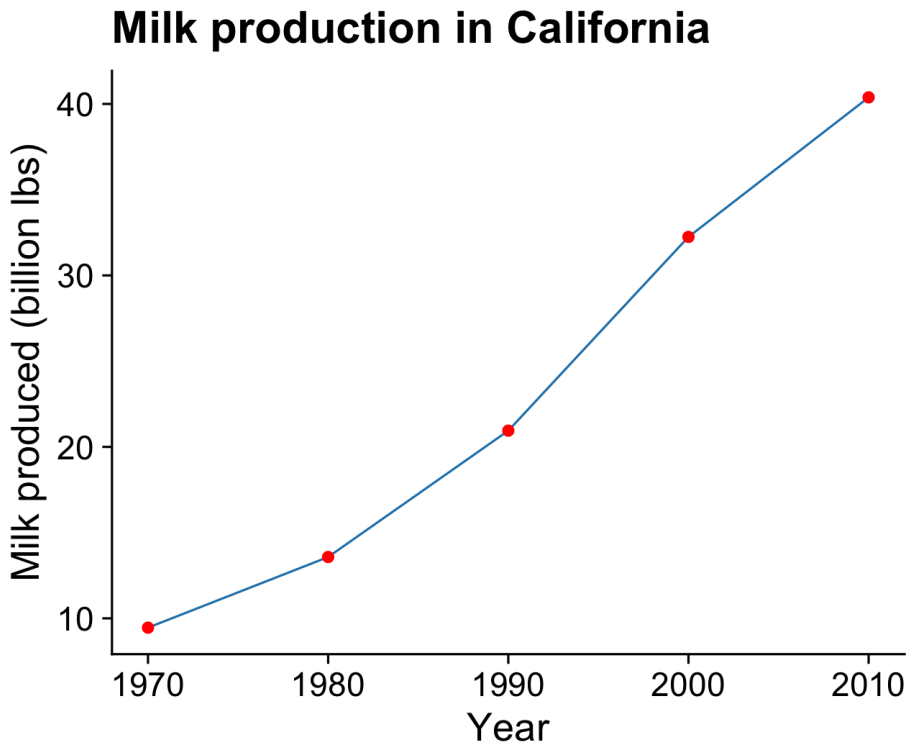


Points + Line: helps emphasize the overall trend

Trends

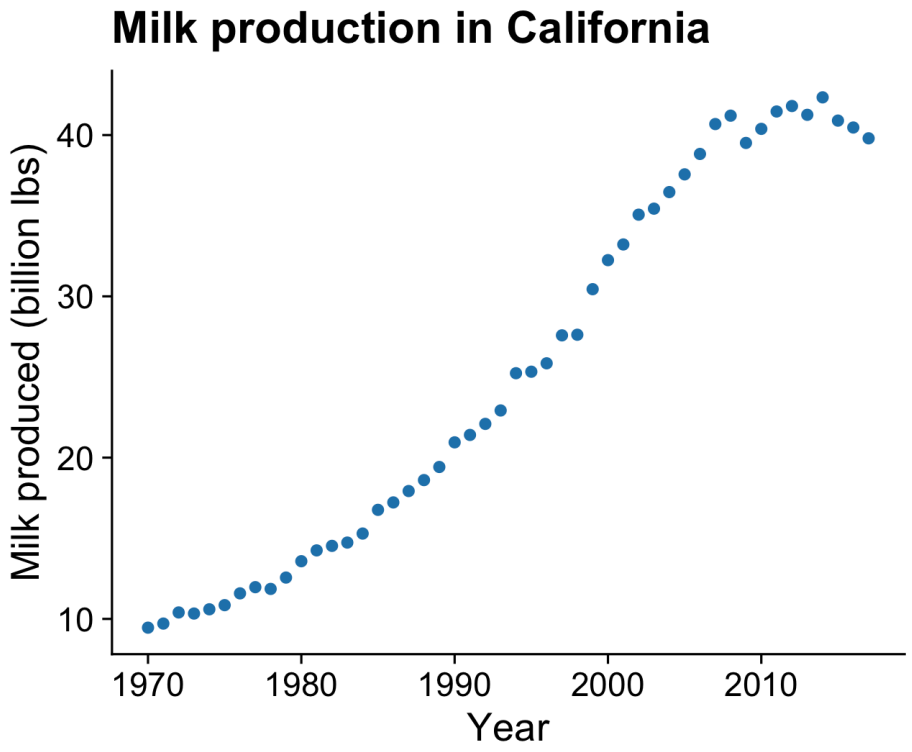


Points

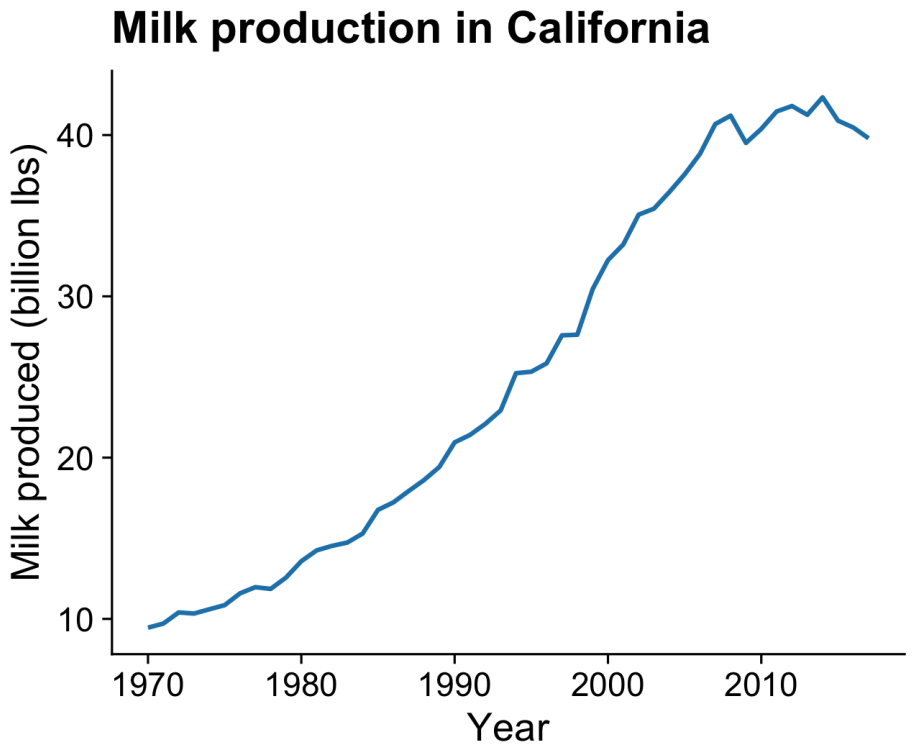


Points + Line: Note that for sparse data, a line can potentially be misleading

Trends

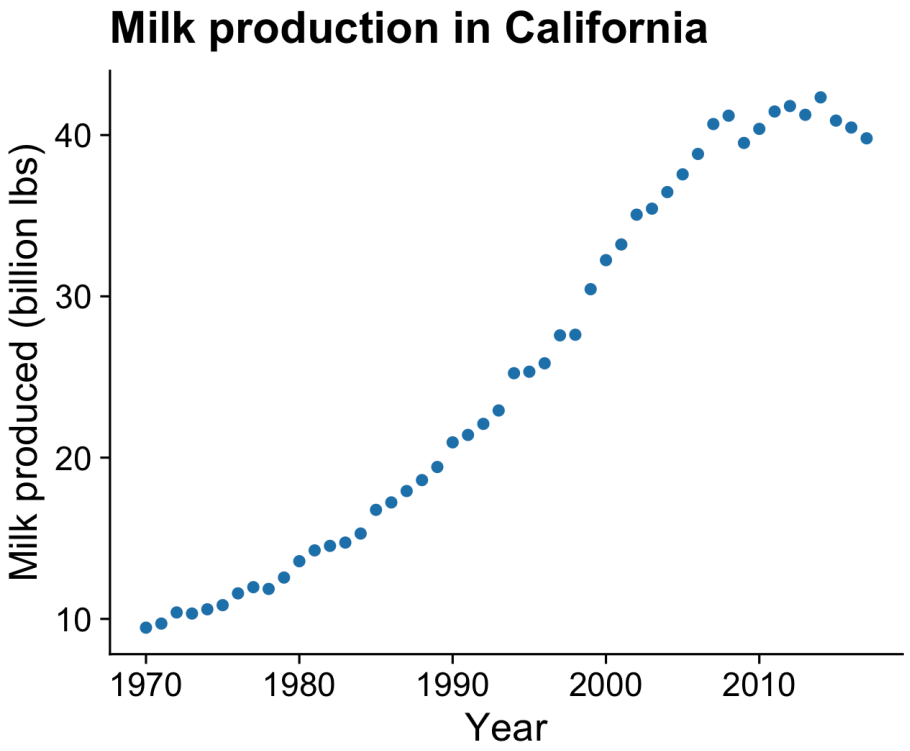


Points

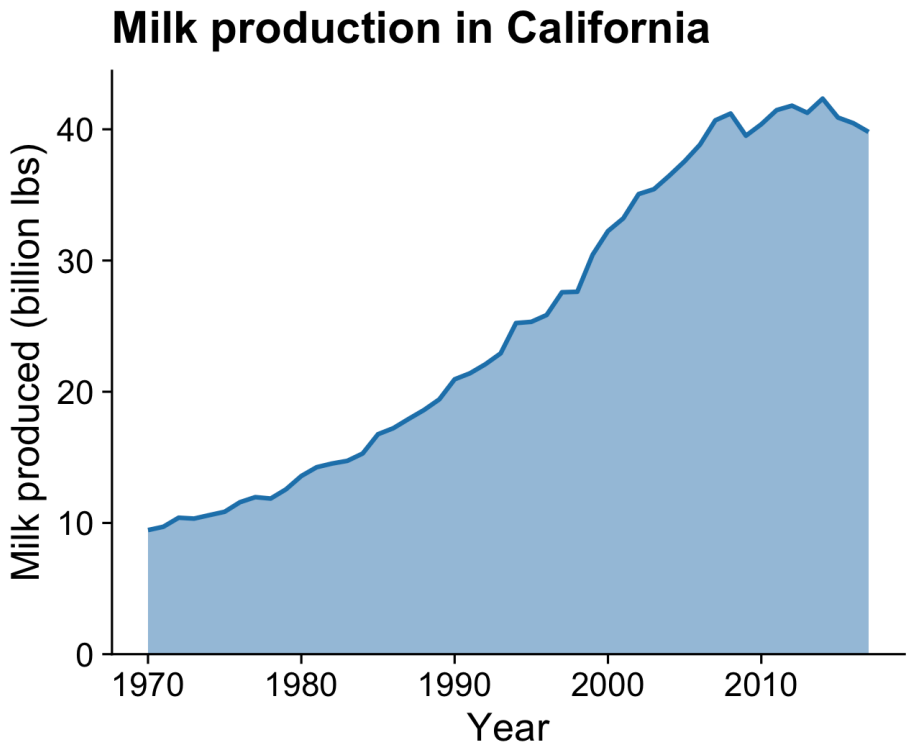


Line: omitting points emphasizes the overall trend

Trends



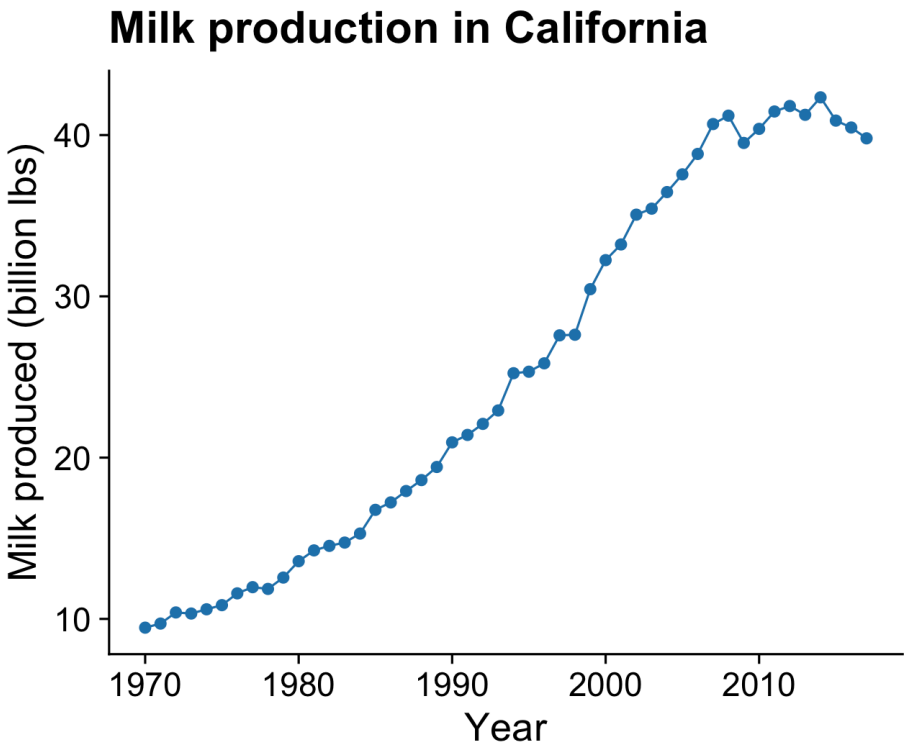
Points



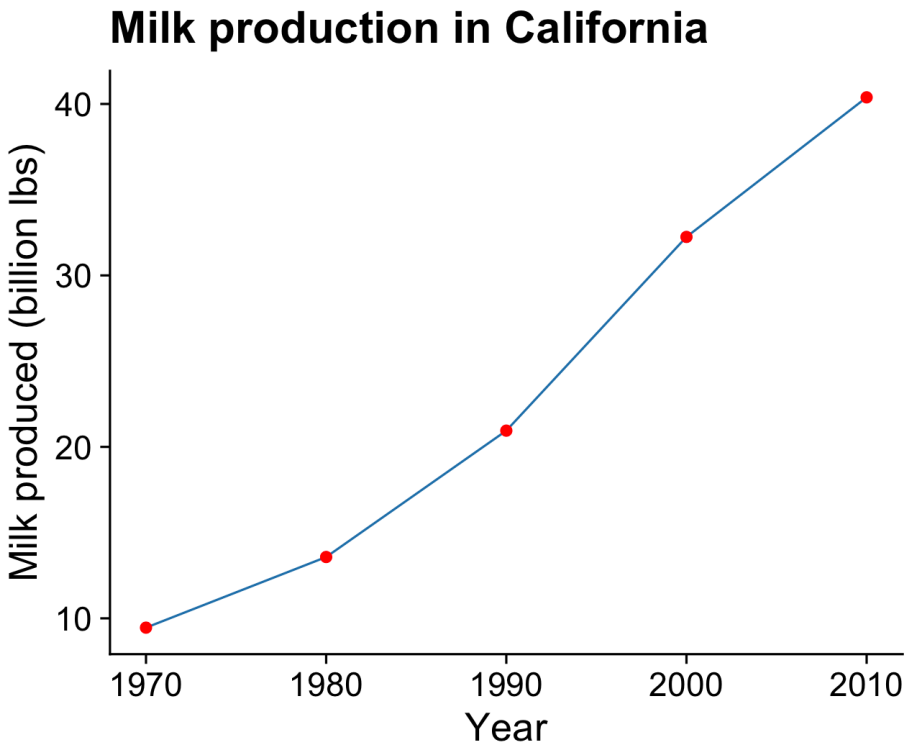
Line + Area: further emphasizes the overall trend

but y-axis must start at 0

Trends

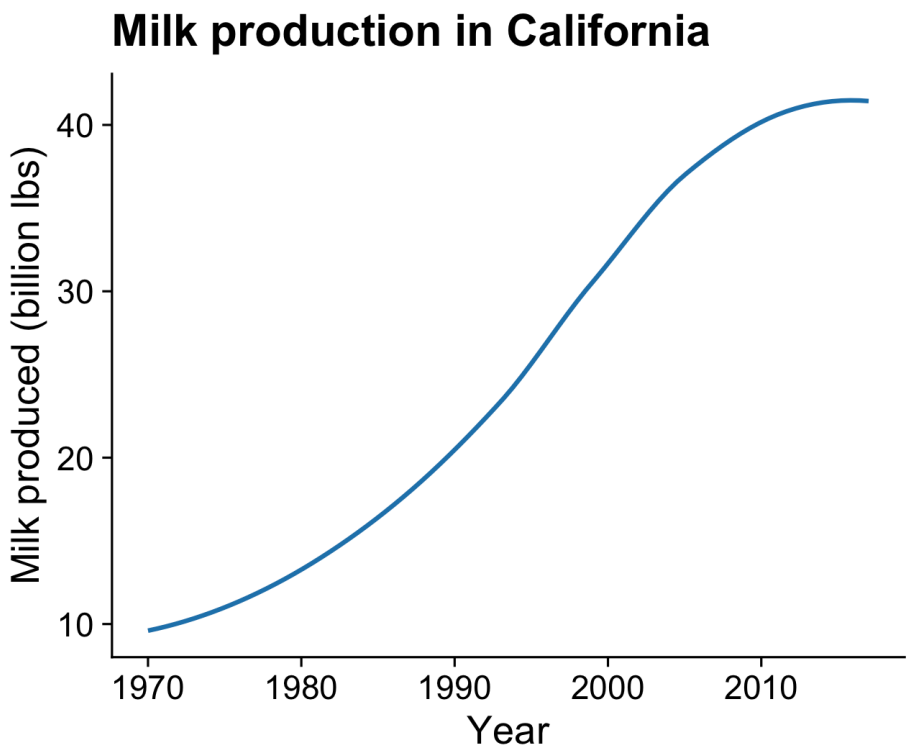


Points + Line: helps emphasize the overall trend

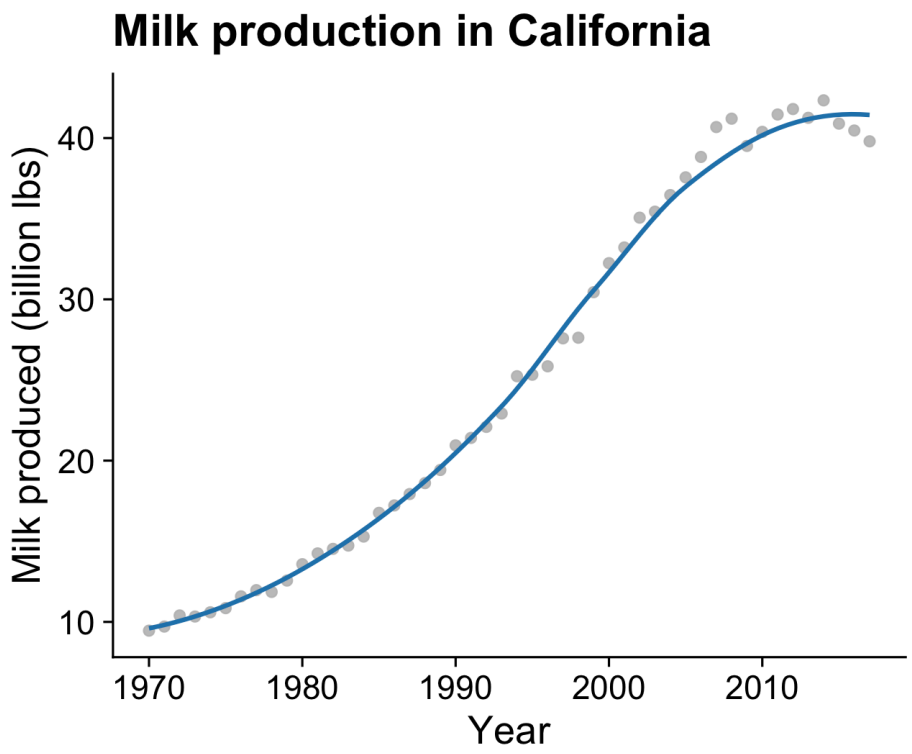


Points + Line: Note that for sparse data, a line can potentially be misleading

Trends



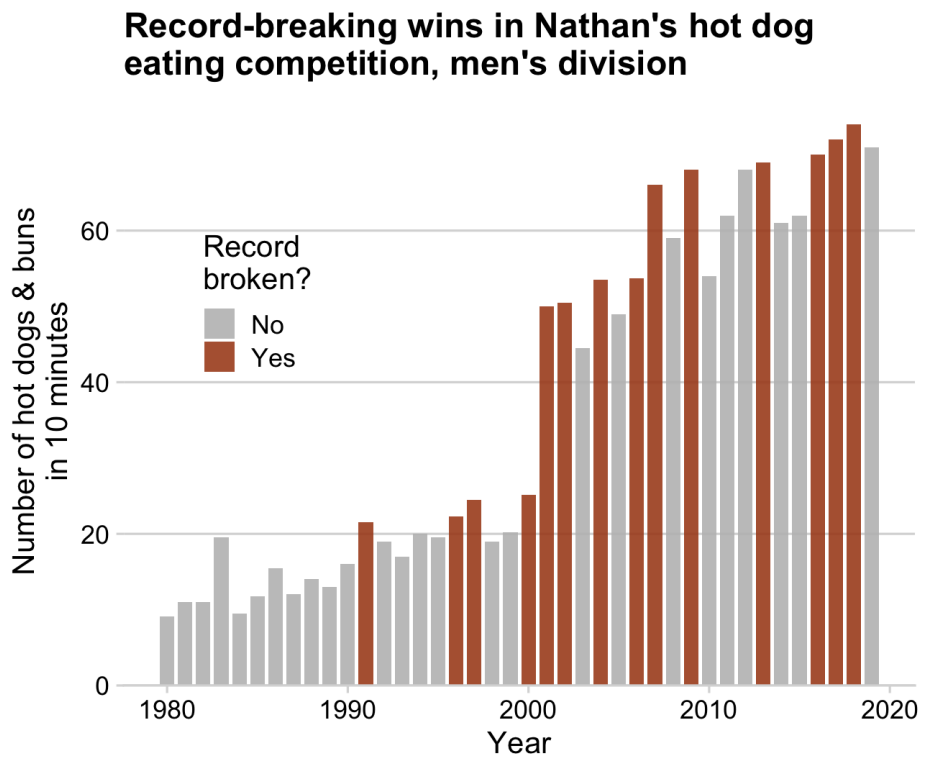
Smoothed line: shows modeled representation of the trend

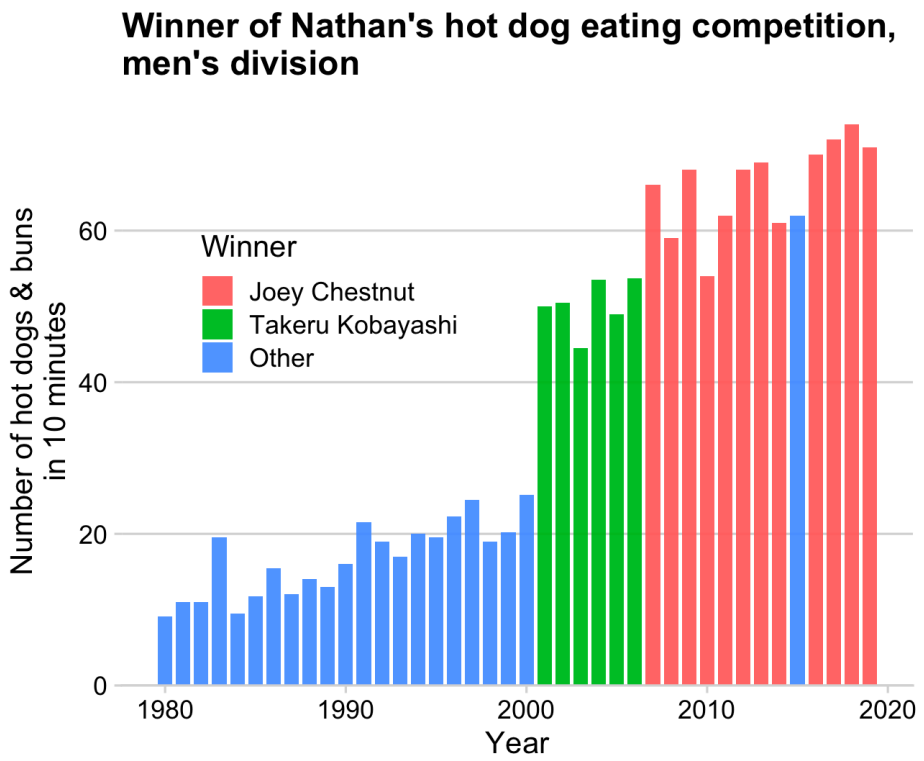


Smoothed line + points: helps show whether outliers are driving the trend

Trends

- Bars: useful to emphasize data points rather than slope between them

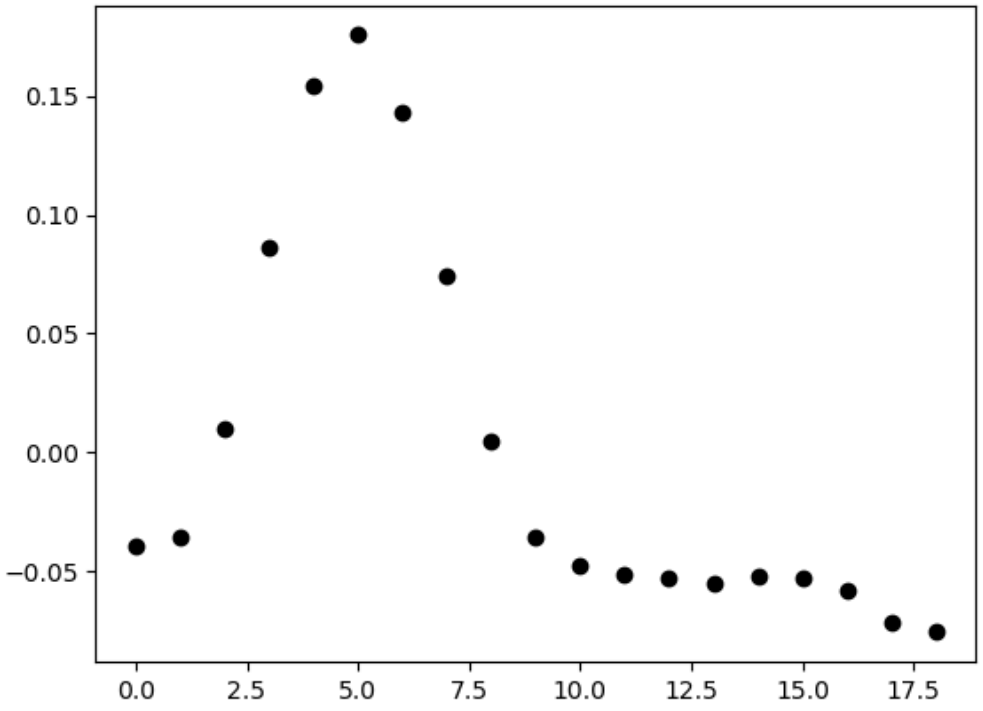




Review using FMRI data from last week

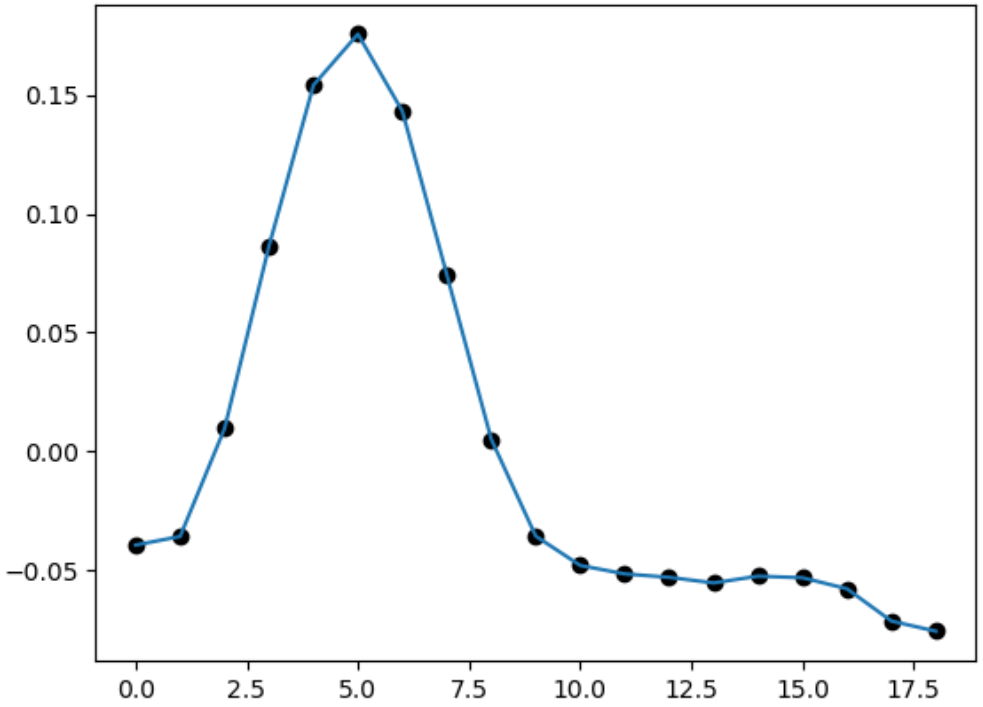
Review using FMRI data from last week

- Scatter plot for time series -> points are evenly spaced



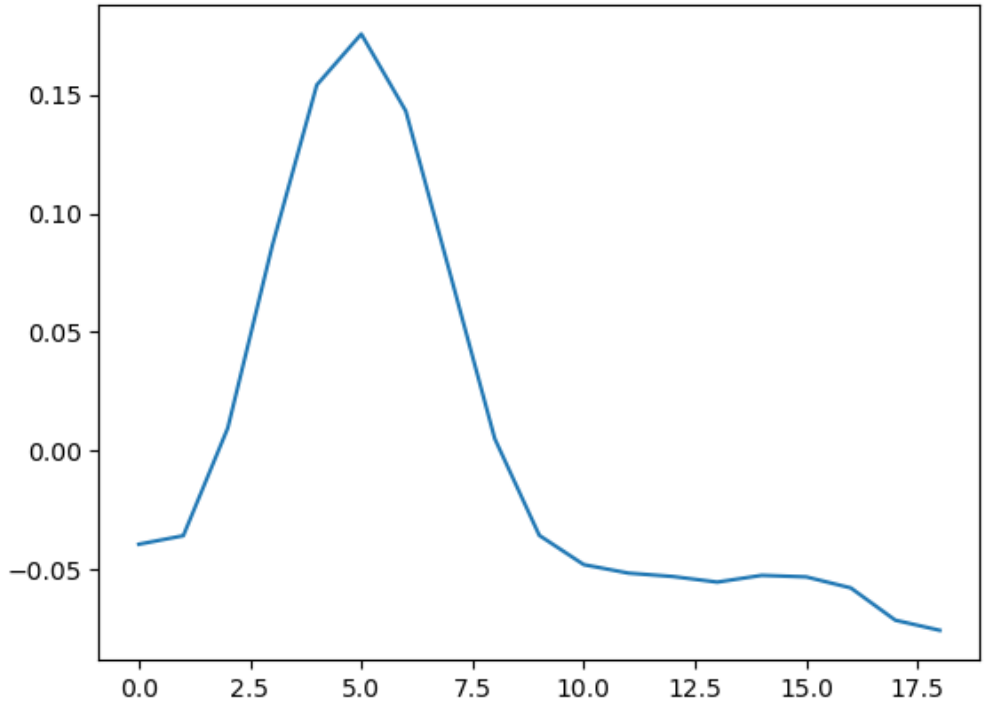
Review using FMRI data from last week

- Line plot for time series -> emphasizes the trend
- But inherently assumes that a linear connection between points is ok



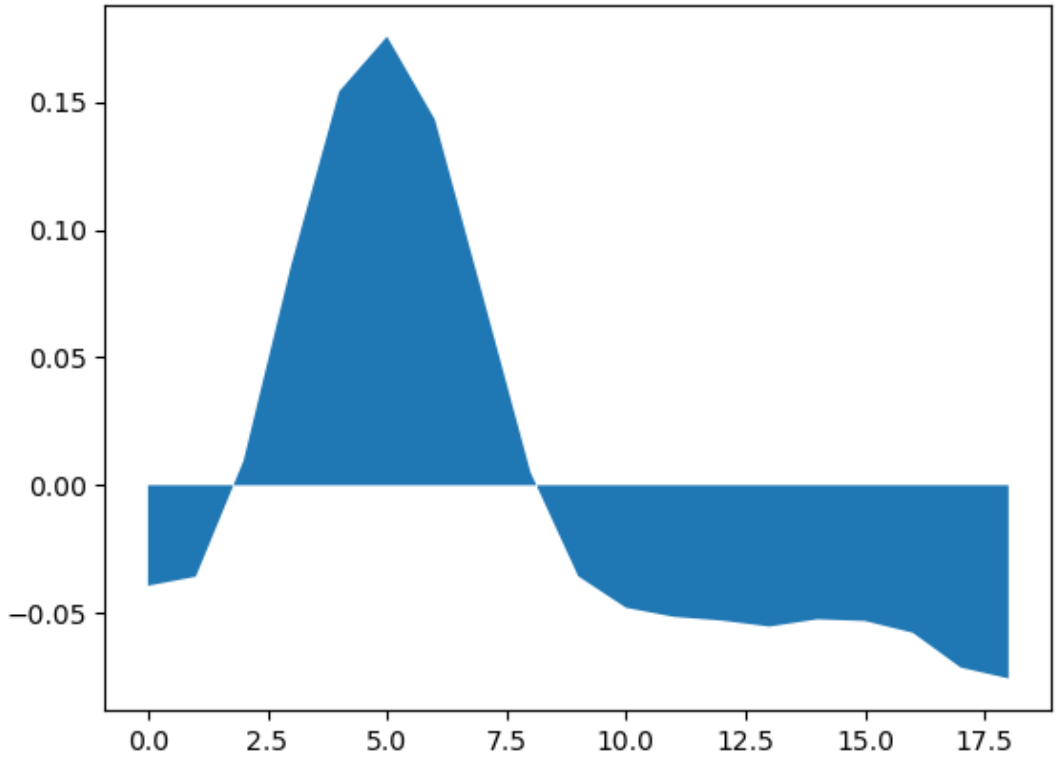
Review using FMRI data from last week

- Removing the points is commonly done when the emphasis is on the trend and less on the actual data points
- Questionable for points spaced far apart, but good for closely spaced points
- It also helps to remove the clutter



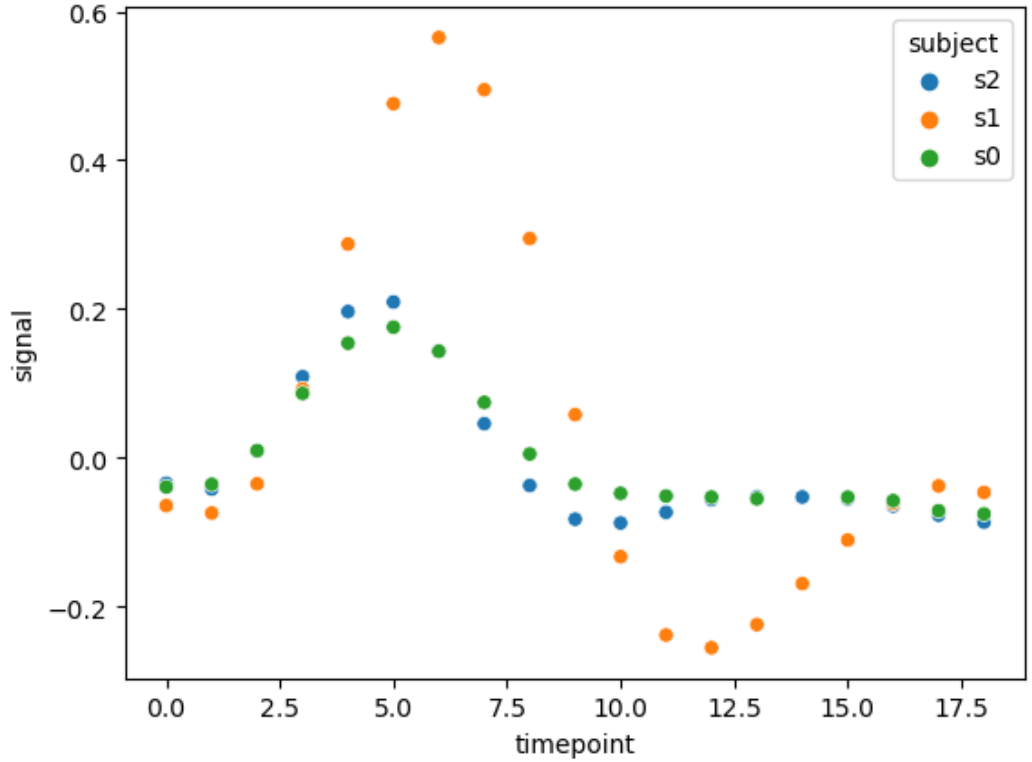
Review using FMRI data from last week

- Area charts are good to emphasize difference relative to reference point



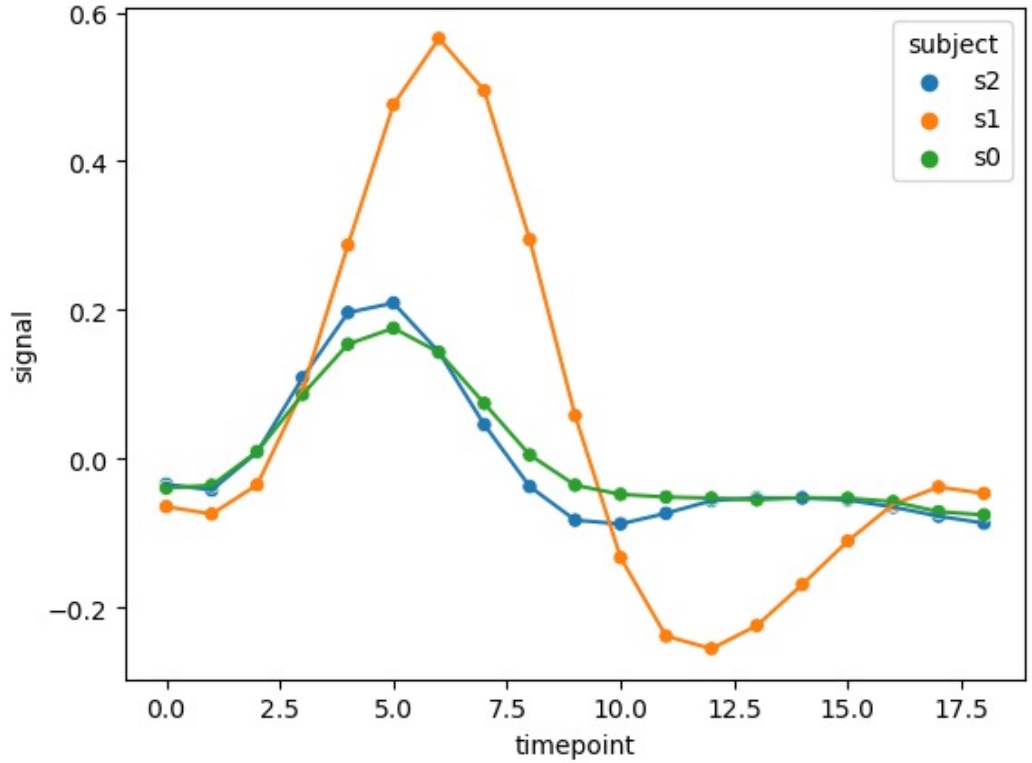
Review using FMRI data from last week

- Multiple time series: scatter plots are difficult to disentangle



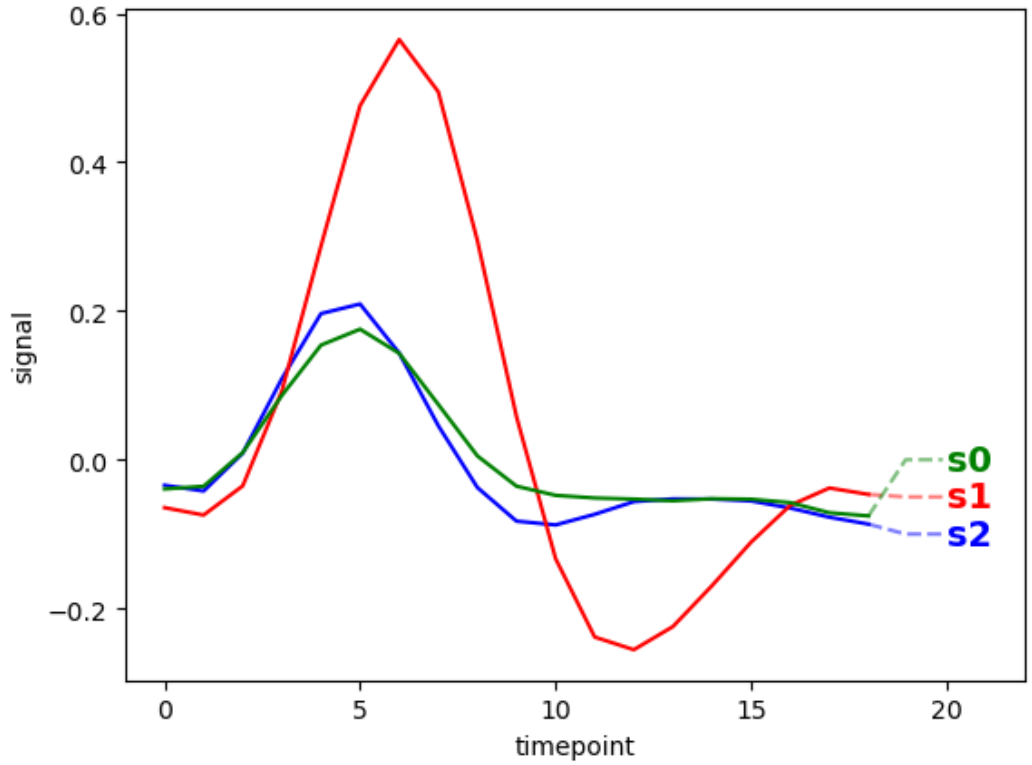
Review using FMRI data from last week

- Multiple time series: scatter plots are difficult to disentangle
- Lines visually assist us in following the trends across multiple category values



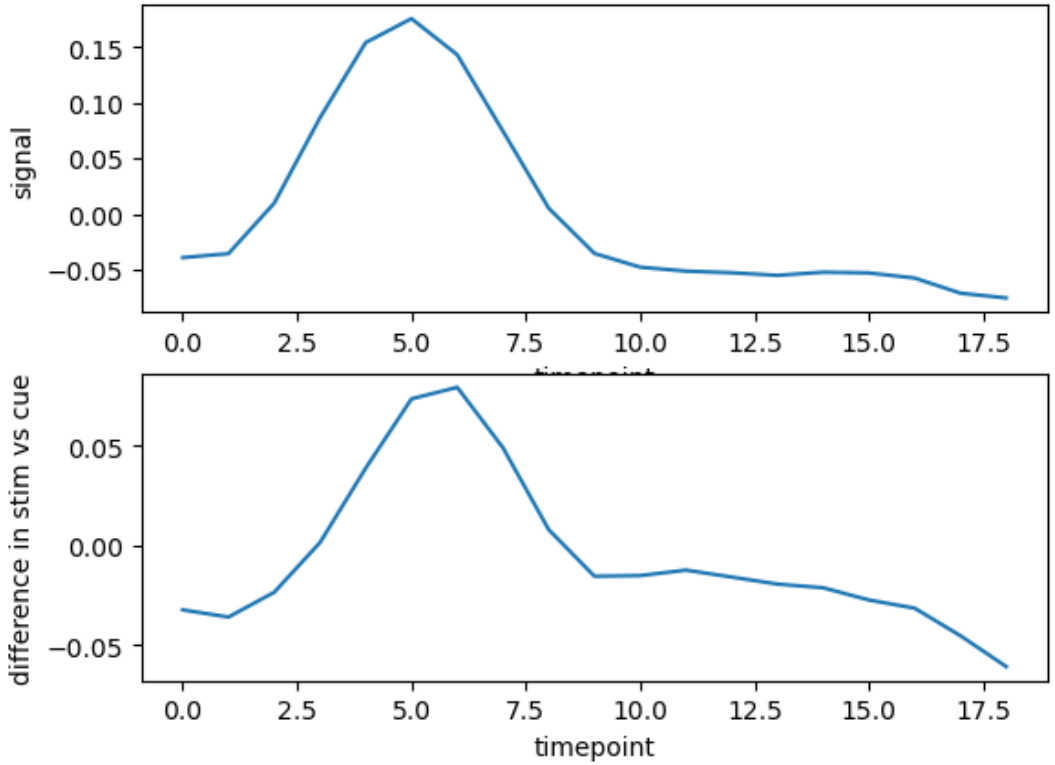
Review using FMRI data from last week

- Multiple time series: scatter plots are difficult to disentangle
- Lines visually assist us in following the trends across multiple category values
- We can also reduce the cognitive load even more by removing the legend and points



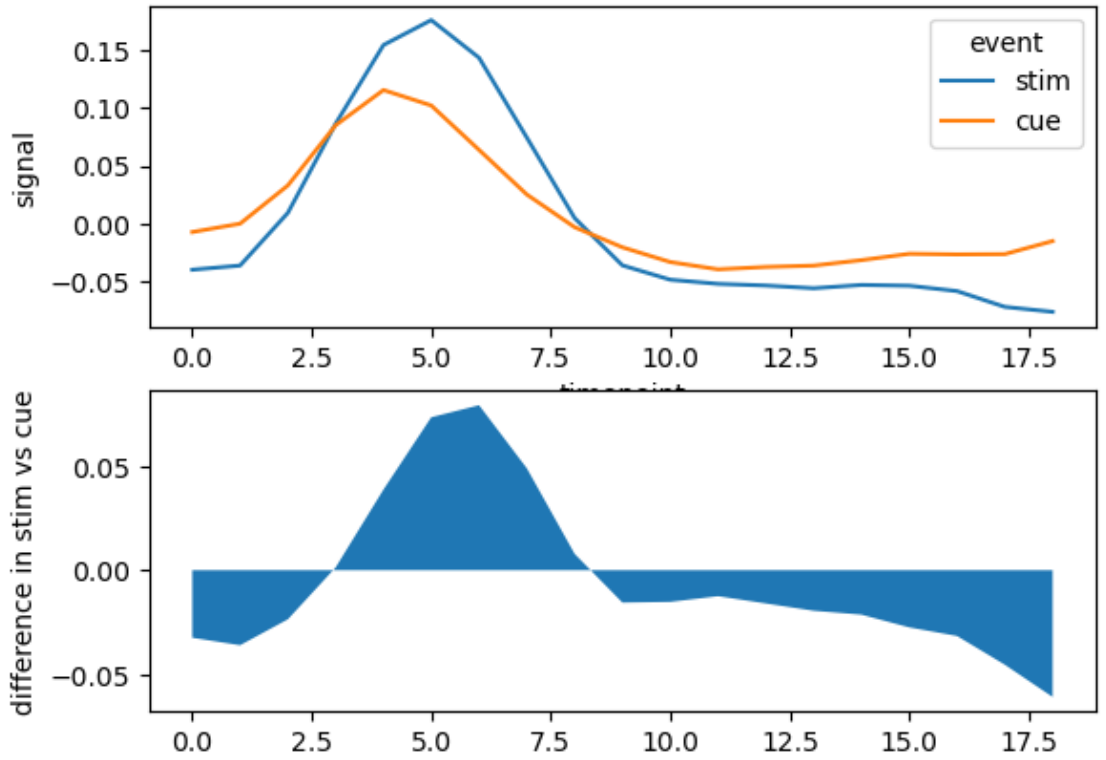
Review using FMRI data from last week

- Time series with two or more response variables
- Plot on top of each other to see what happens at similar times



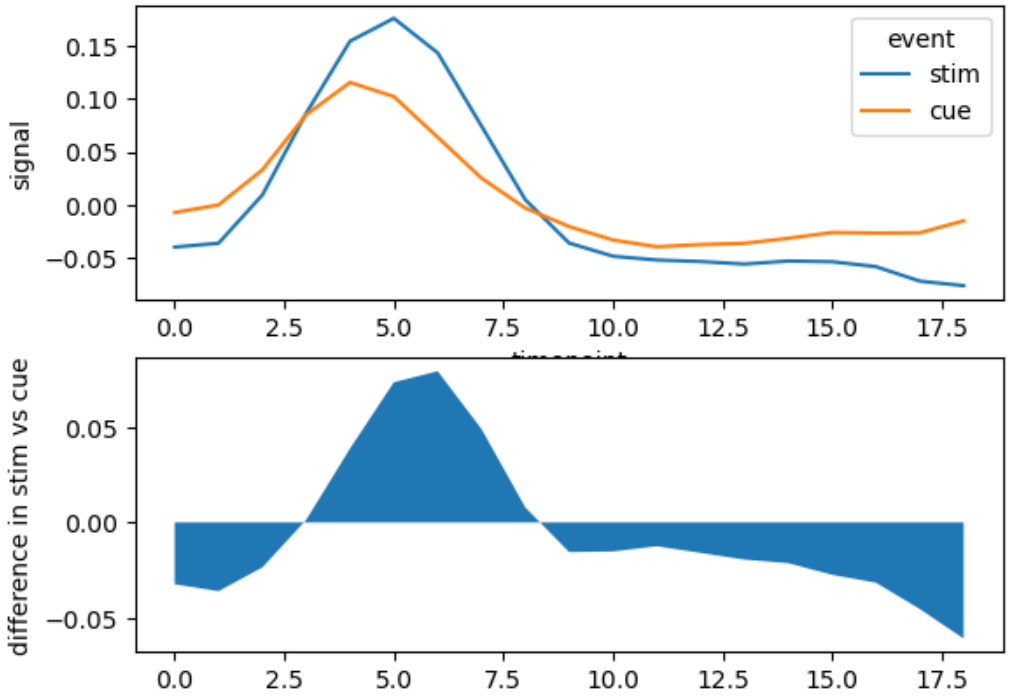
Review using FMRI data from last week

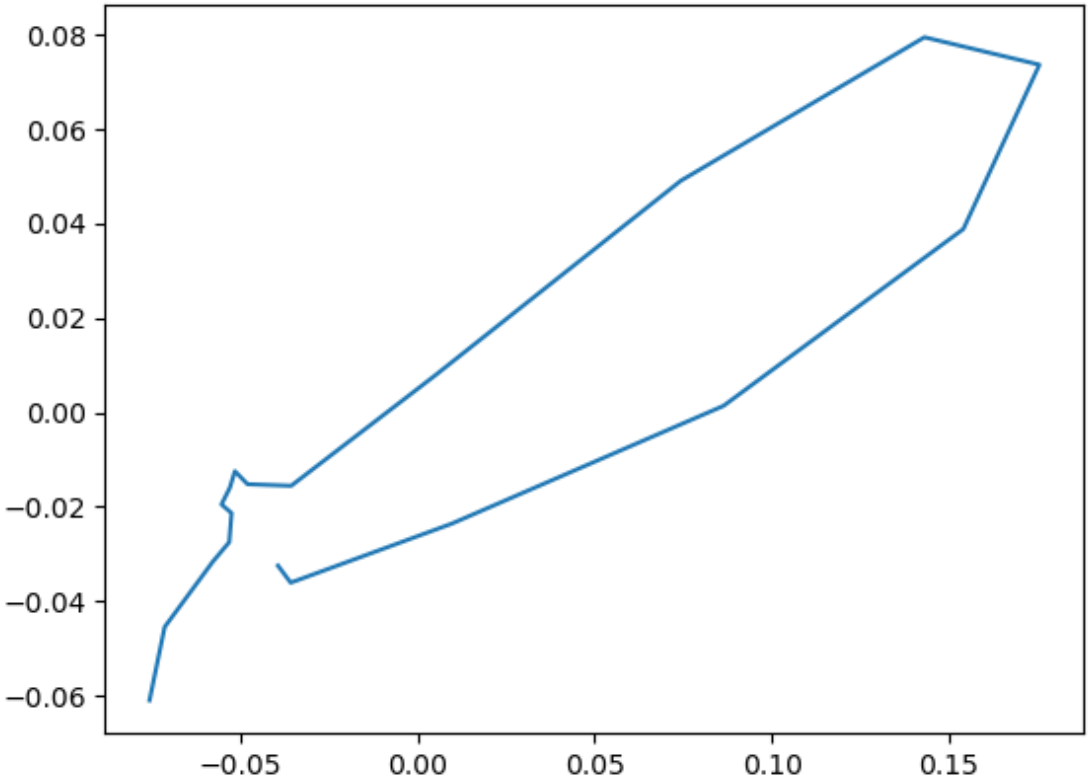
- Time series with two or more response variables
- Plot on top of each other to see what happens at similar times



Review using FMRI data from last week

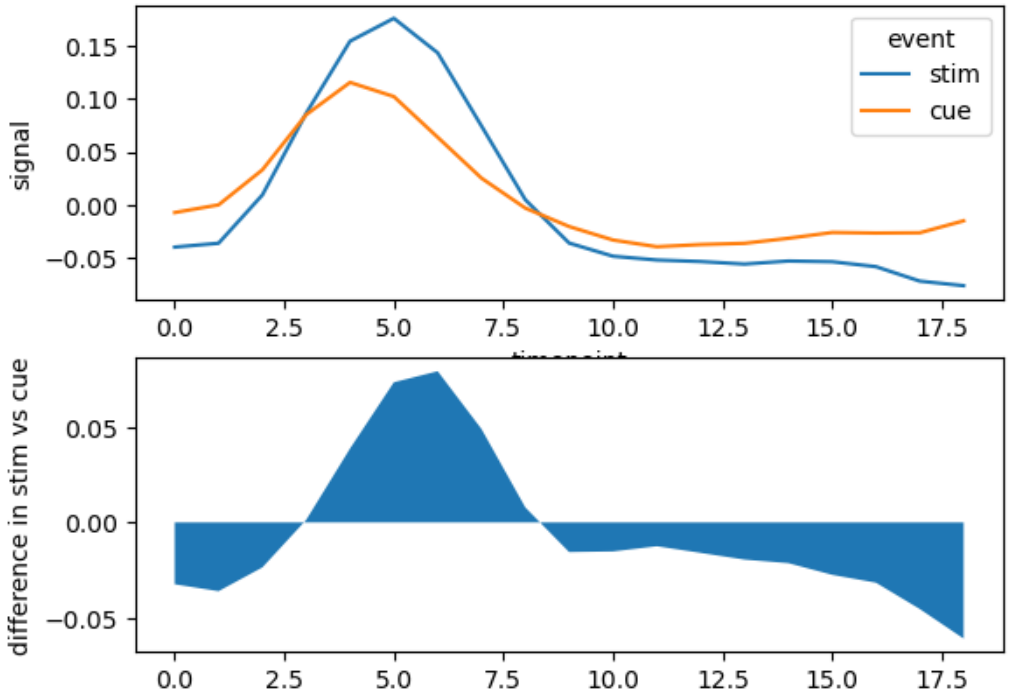
- Time series with two or more response variables
- Plot on top of each other to see what happens at similar times
- Or in “phasespace” plot -> but important to show direction and scale too

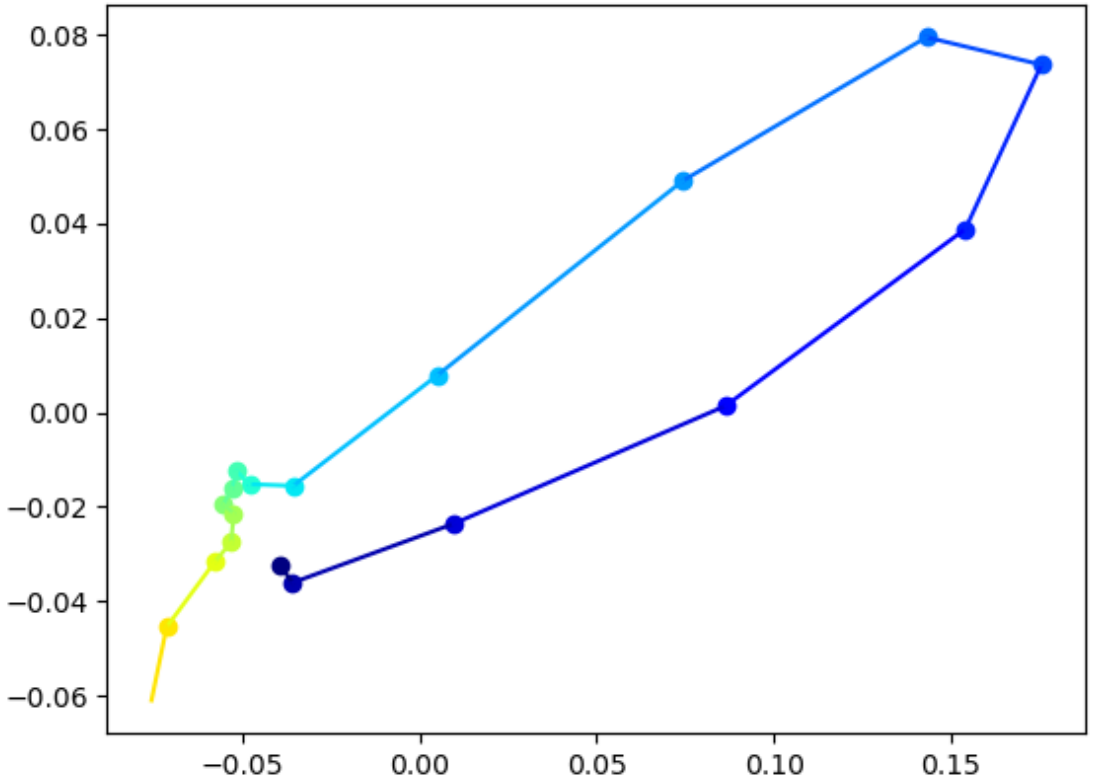




Review using FMRI data from last week

- Time series with two or more response variables
- Plot on top of each other to see what happens at similar times
- Or in “phasespace” plot -> but important to show direction and scale too





Review using FMRI data from last week

Visualizing trends

- Often it’s more important to visualize the trend than the specific details of every point
- Trend lines can help viewers identify the key features
  - Draw on top of points, or instead of points
- Approaches:
  - Smooth the data, say by moving average
  - Fit a curve

Smoothing

- Window average
- Some ambiguity as to where in the window the average is plotted
  - Financial analysts -> endpoint
  - Statisticians -> middle point
- Window length sets scale over which fluctuations are visible
- Limitations
  - Smoothed curve is shorter
  - Not necessarily smooth -> can be affected by new data points at the edge

Alternatives to Smoothing

- LOESS
  - locally estimated scatterplot smoothing
  - Fits low-degree polynomials to subsets of the data
  - Points in the center are weighted more heavily than at the edges
  - Slow for large data -> requires fitting many separate polynomials
- Splines
  - Piecewise polynomial that looks smooth
  - Computationally efficient
  - Many different types: cubic splines, B-splines, thin-plate splines, …
    - Hard to know which to choose

Alternatives to Smoothing

- Smoothers don’t provide parameter estimates with a meaningful interpretation
- Whenever possible, it’s preferable to fit points with a curve that has a specific functional form
  - Gives parameters with interpretable meaning

Code time

Visualizing trends

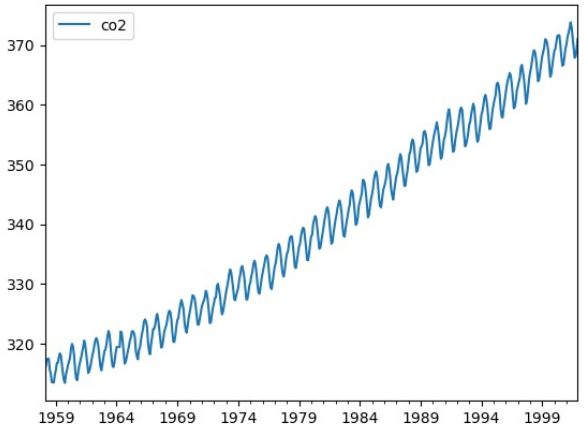
- Often it’s more important to visualize the trend than the specific details of every point
- Trend lines can help viewers identify the key features
  - Draw on top of points, or instead of points
- Approaches:
  - Smooth the data, say by moving average
  - Fit a curve

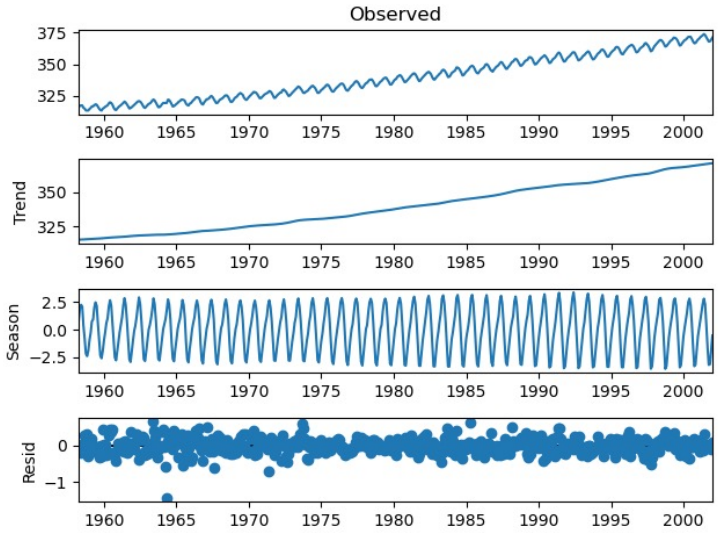
Visualizing trends

- Often it’s more important to visualize the trend than the specific details of every point
- Trend lines can help viewers identify the key features
  - Draw on top of points, or instead of points
- Approaches:
  - Smooth the data, say by moving average
  - Fit a curve
- When identifying a trend, it may also be useful to look at deviations from the trend or to split the trend into separate components

De-trending

- Time series may have several underlying trends
  - Long-term trend
  - Seasonal trend (sales of Christmas trees)
  - Daily trend (high and low temperature)
- There are a variety of ways to decompose signals in order to look for various trends occurring together

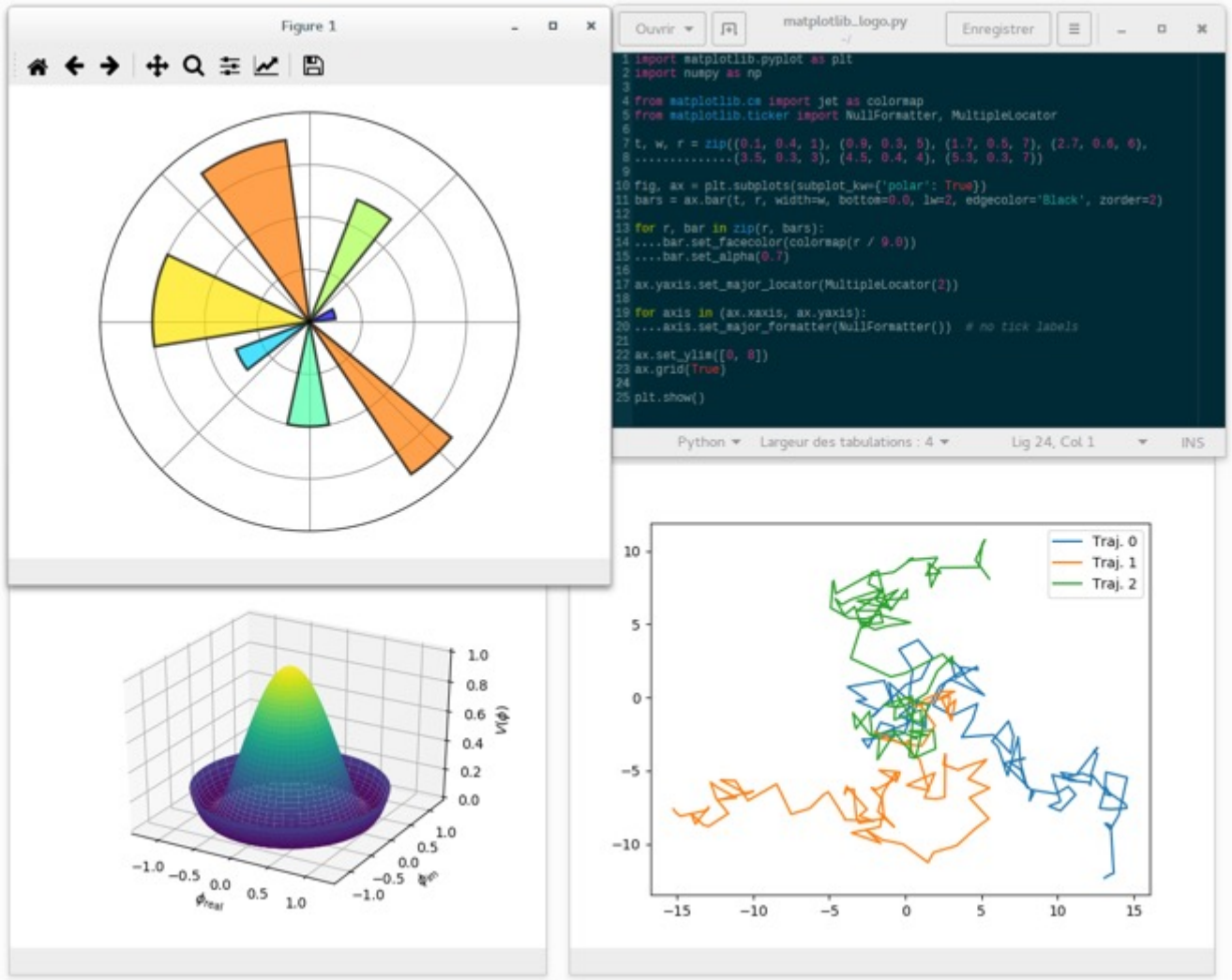




Code time

Visualization in Python: a standard library is matplotlib

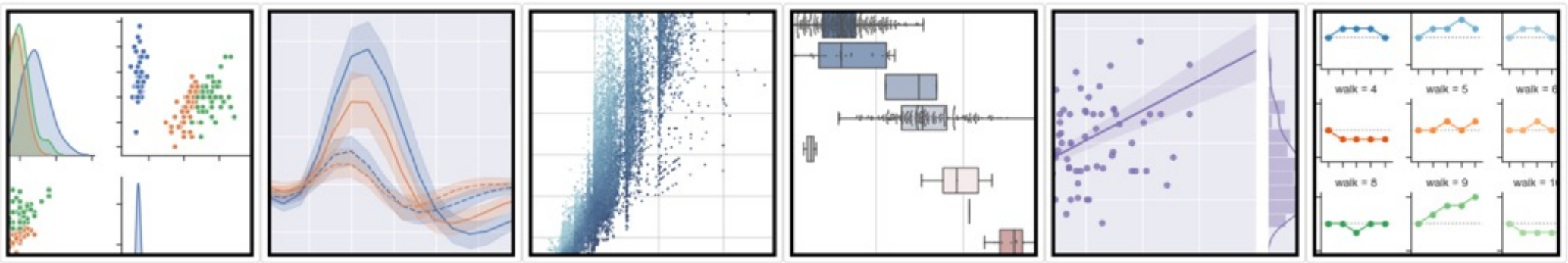
"matplotlib tries to make easy things easy and hard things possible."



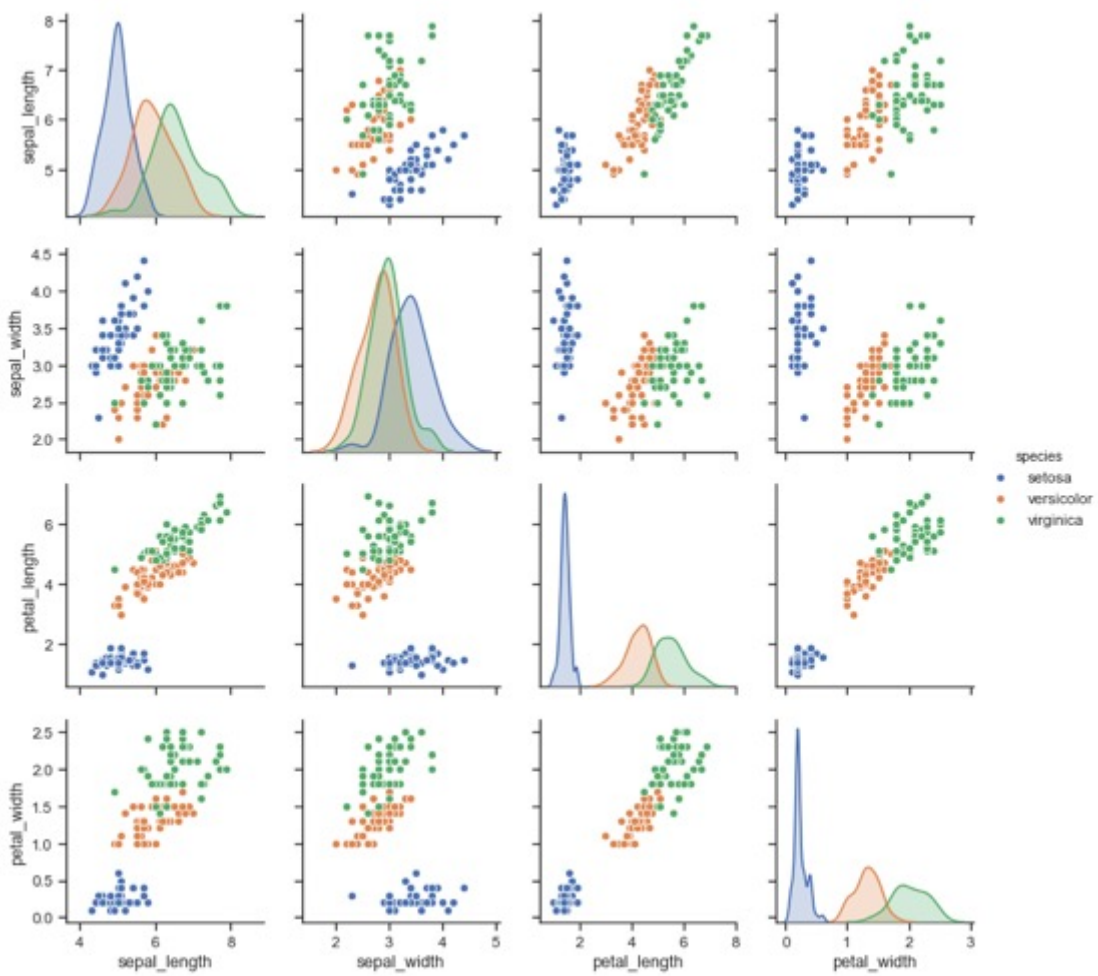
https://matplotlib.org/

seaborn

If Matplotlib “tries to make easy things easy and hard things possible,”
Seaborn tries to make a well-defined set of hard things easy too.



https://seaborn.pydata.org/



seaborn

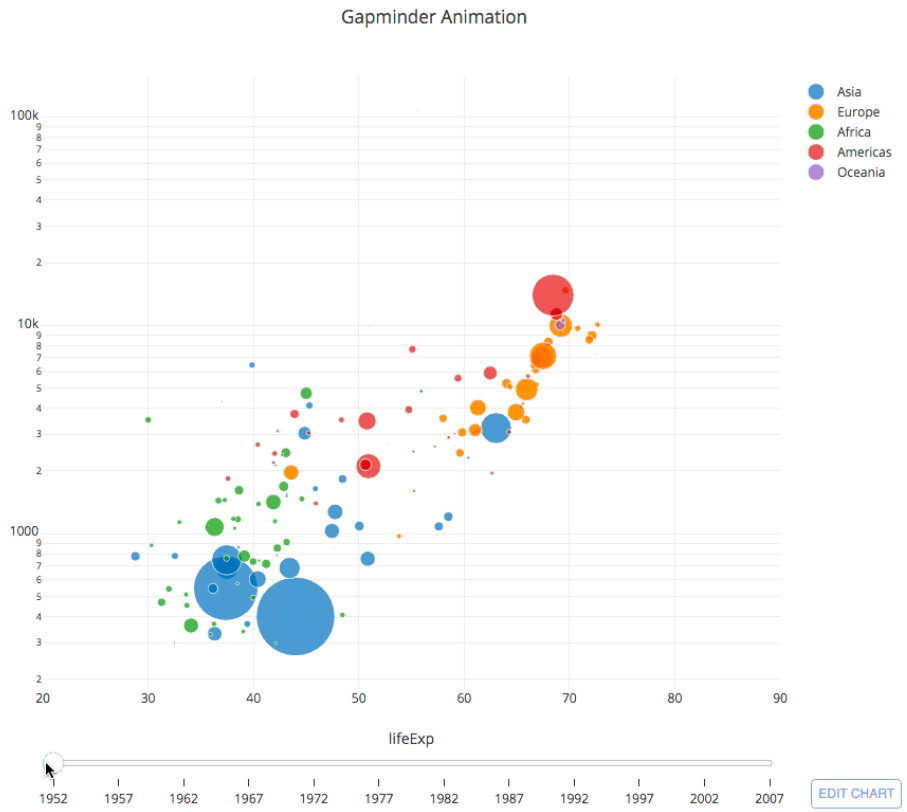
- Built on top of matplotlib and closely integrated with pandas data structures
- Used for making statistical graphics and using visualization to quickly and easily explore and understand data
- The style settings can also affect matplotlib plots, even if you don't make them with seaborn

https://seaborn.pydata.org/

plotly

The plotly Python library (plotly.py) is an interactive, open-source, and browser-based graphing library

https://plot.ly/python/



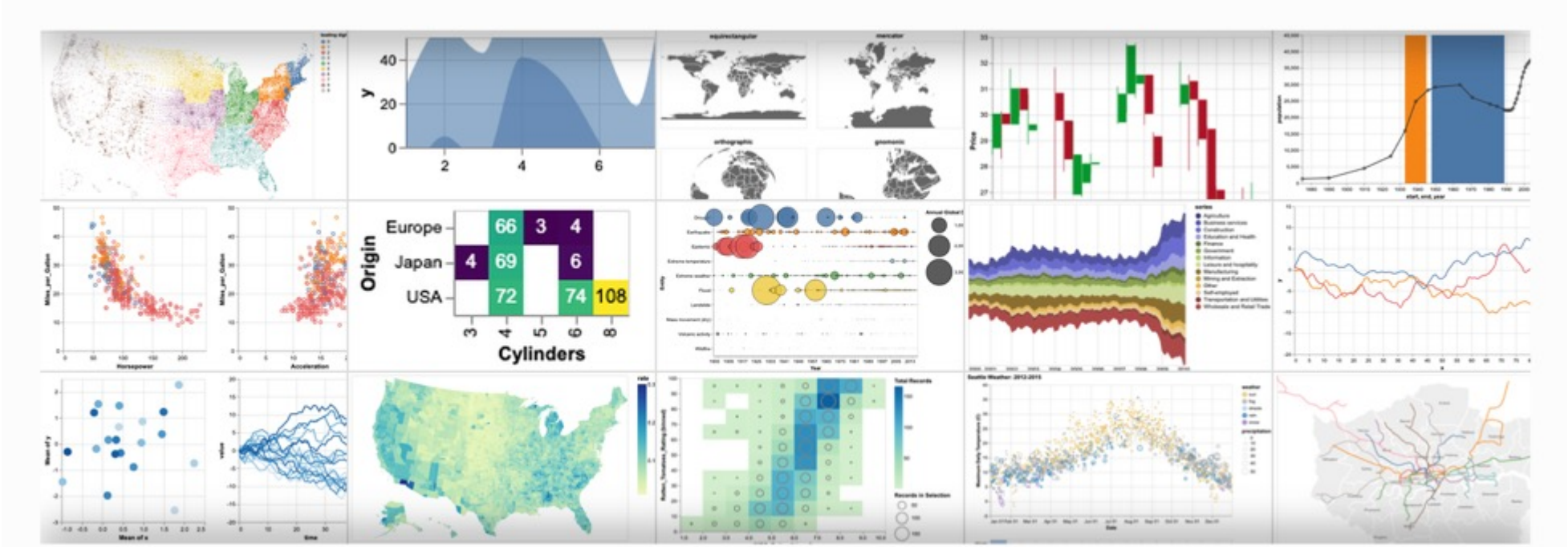
plotly

- An open-source product of Plotly, Inc., that is built on top of Javascript (plotly.js).
- Enables Python users to create beautiful interactive web-based visualizations that can be displayed in Jupyter notebooks, saved to standalone HTML files, or served as part of pure Python-built web applications using Dash.
- Also has a version for R, as well as other web visualization products

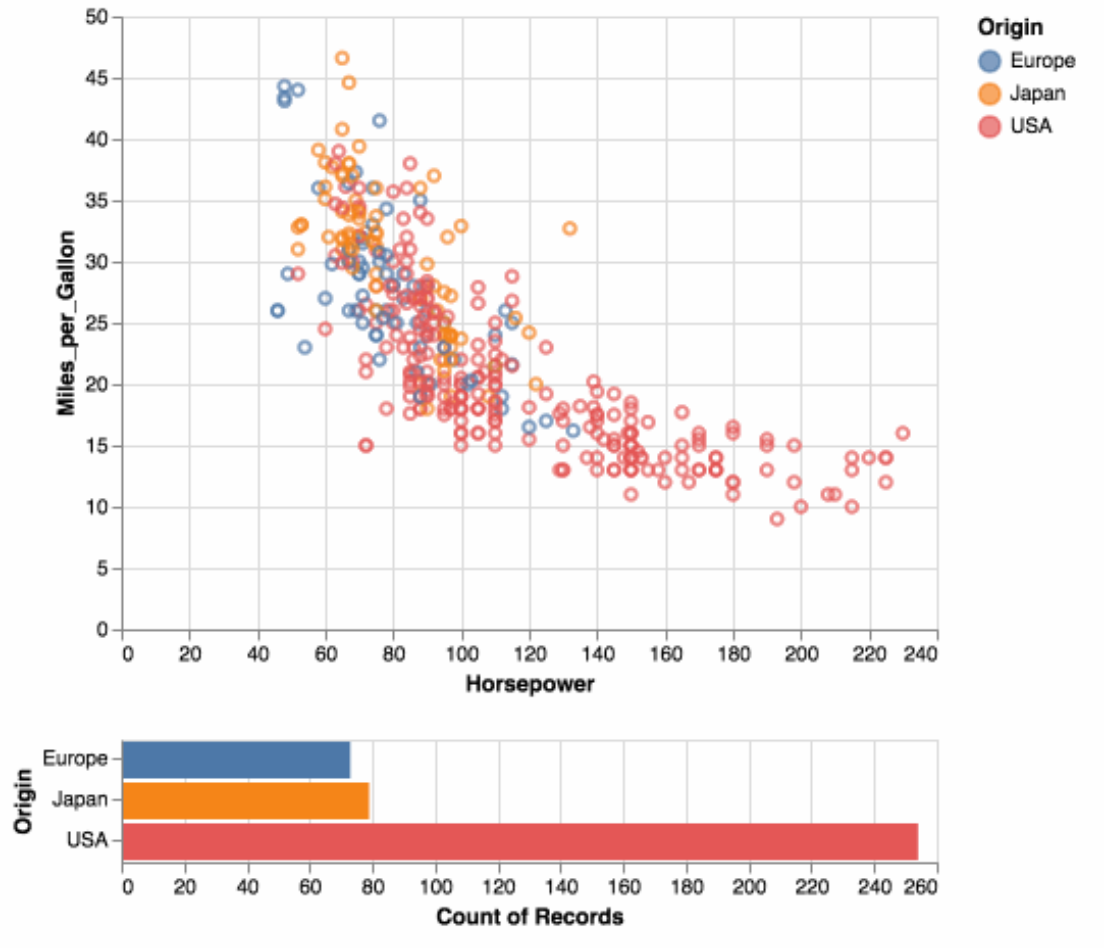
https://plot.ly/python/

Altair

Altair is a declarative statistical visualization library for Python,
based on Vega and Vega-Lite (high-level grammar of interactive graphics)



https://altair-viz.github.io/



Altair

- Based on Vega and Vega-Lite (high-level grammar of interactive graphics)
  - Vega-Lite provides a concise JSON syntax for rapidly generating visualizations to support analysis
  - Its specifications describe visualizations as mappings from data to properties of graphical marks
- Aims for elegant simplicity so focus can be on understanding data

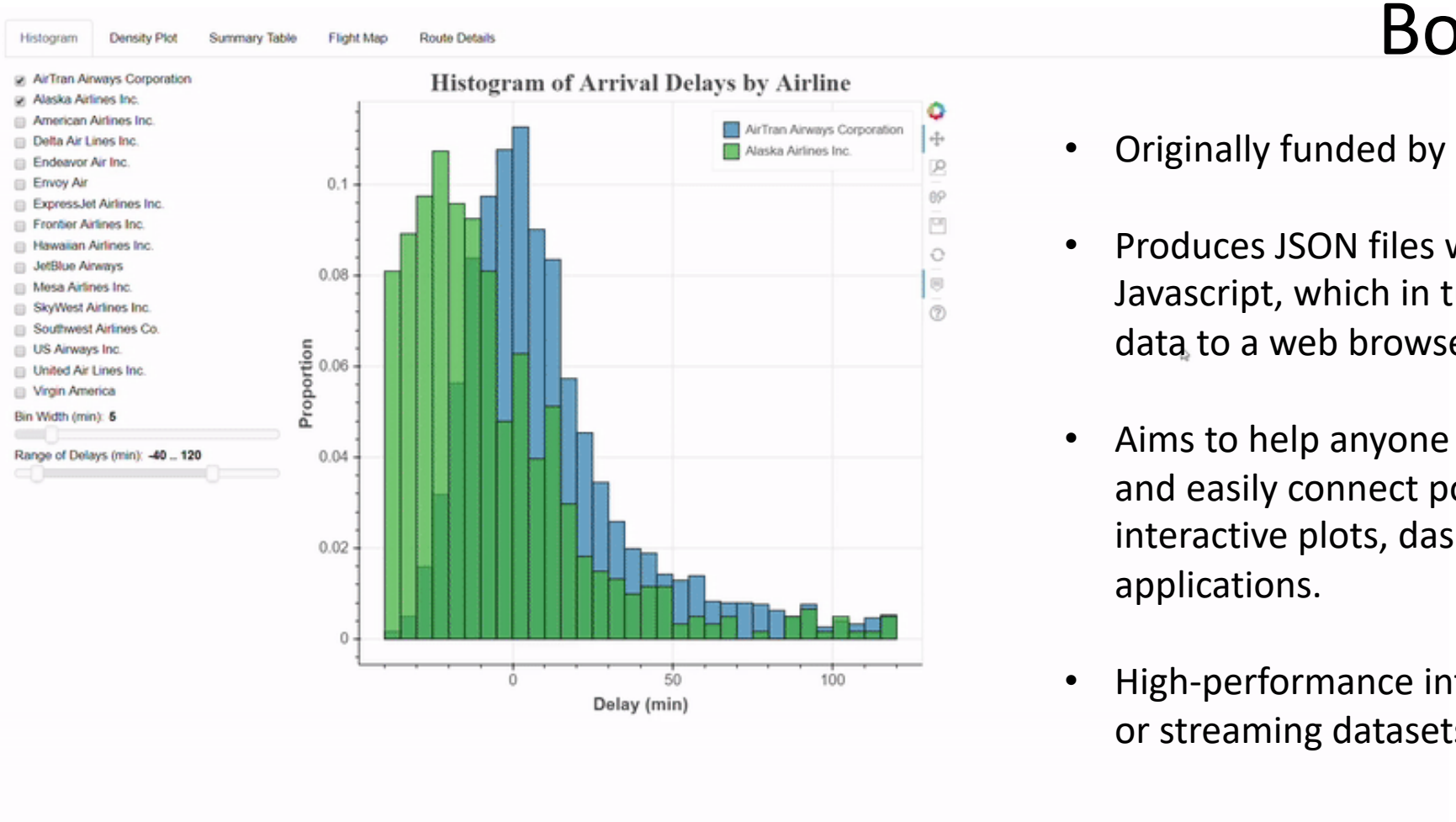
https://altair-viz.github.io/

Bokeh

Bokeh creates shareable, interactive data applications for modern browsers …
all without having to delve into JavaScript or “web tech”.

https://docs.bokeh.org/en/latest/index.html

Bokeh



- Originally funded by DARPA
- Produces JSON files which work as input for Javascript, which in turn are used to present data to a web browser
- Aims to help anyone who would like to quickly and easily connect powerful PyData tools to interactive plots, dashboards, and data applications.
- High-performance interactivity over very large or streaming datasets

https://docs.bokeh.org/en/latest/index.html
gif obtained from https://towardsdatascience.com/data-visualization-with-bokeh-in-python-part-iii-a-complete-dashboard-dc6a86aa6e23

Code time