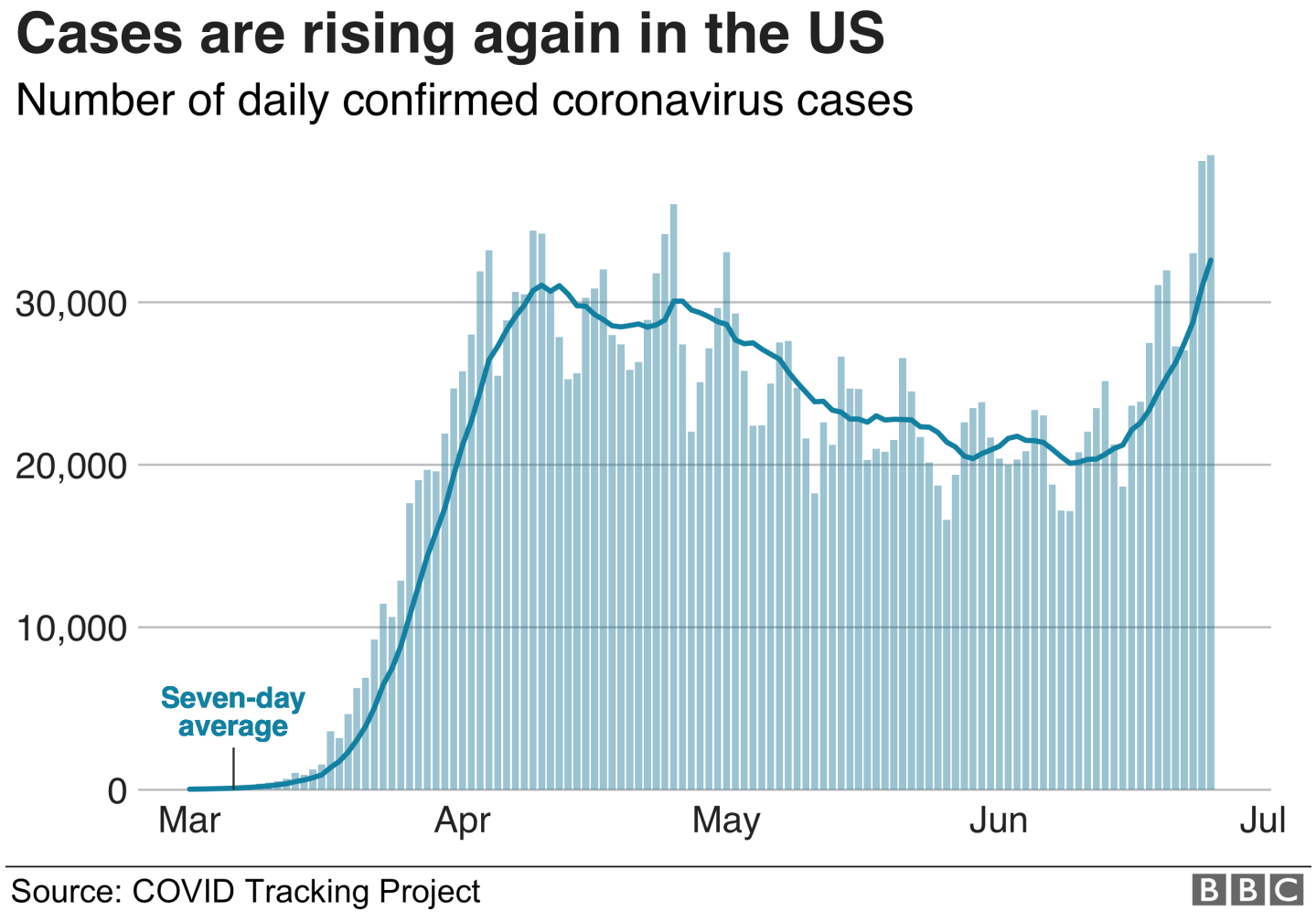
Exploratory Data Analysis & Visualization
Bivariate Data

Ben Winjum

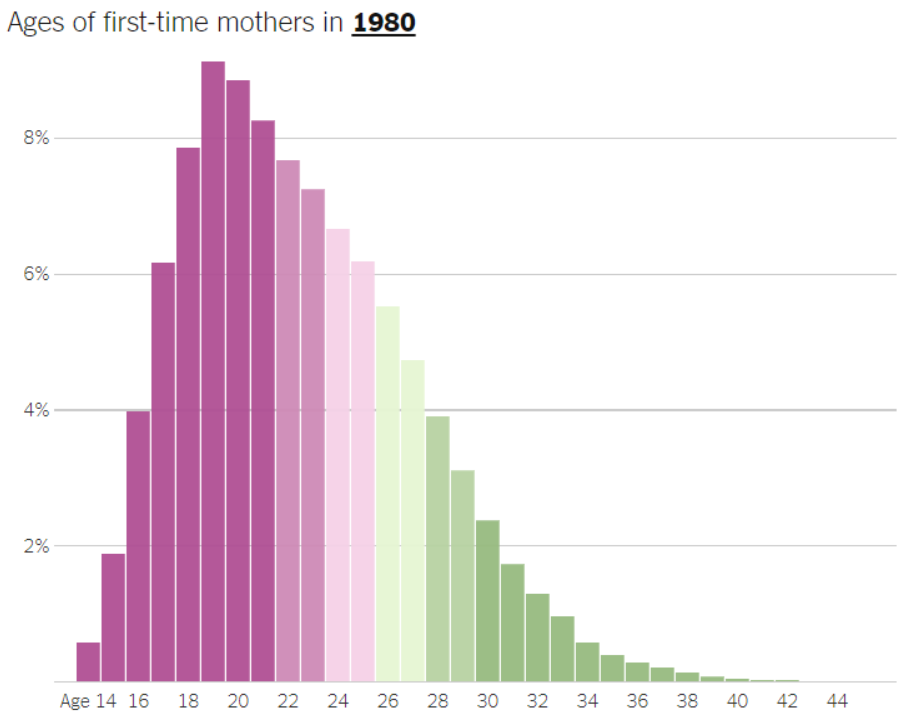
Example Visualizations from Discussions

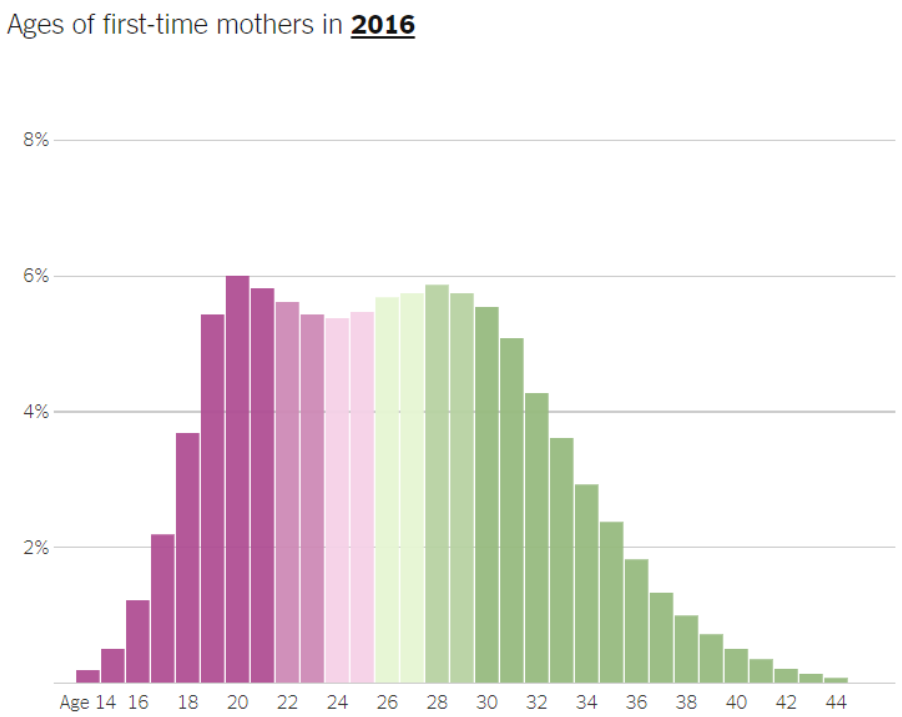
Jing-Wen



Example Visualizations from Discussions

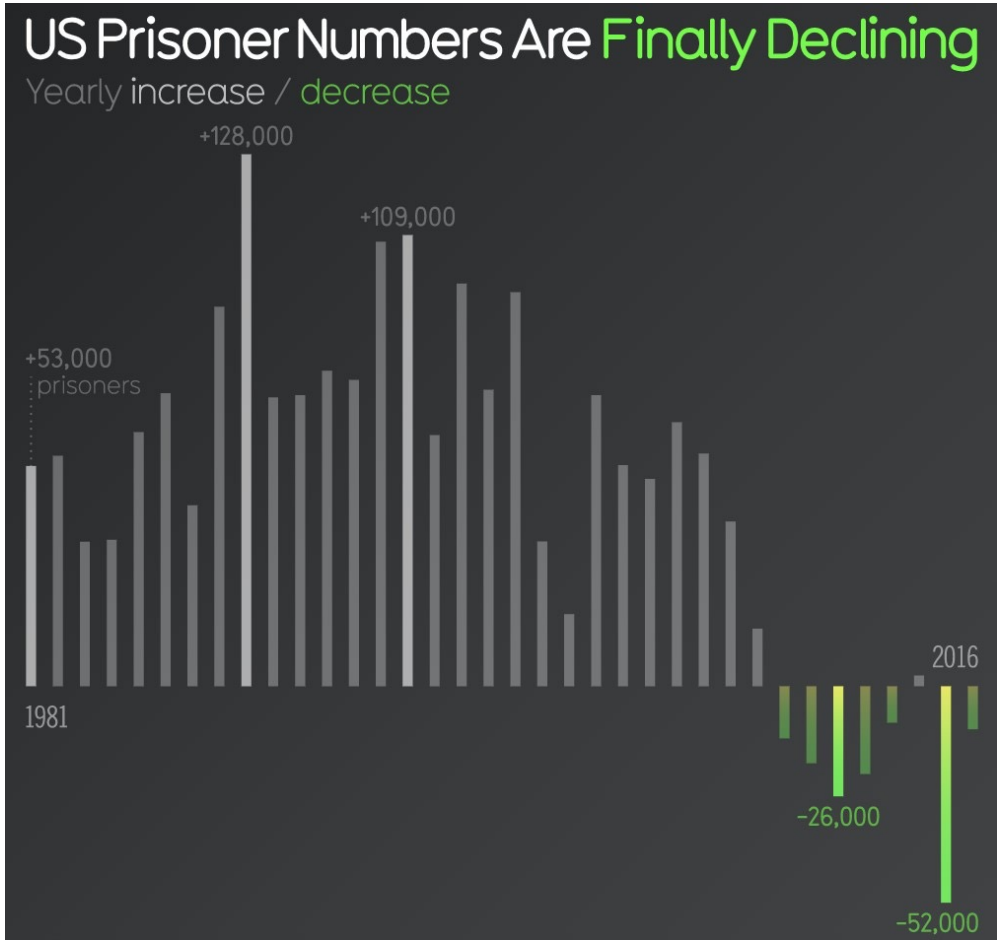
Boonsita



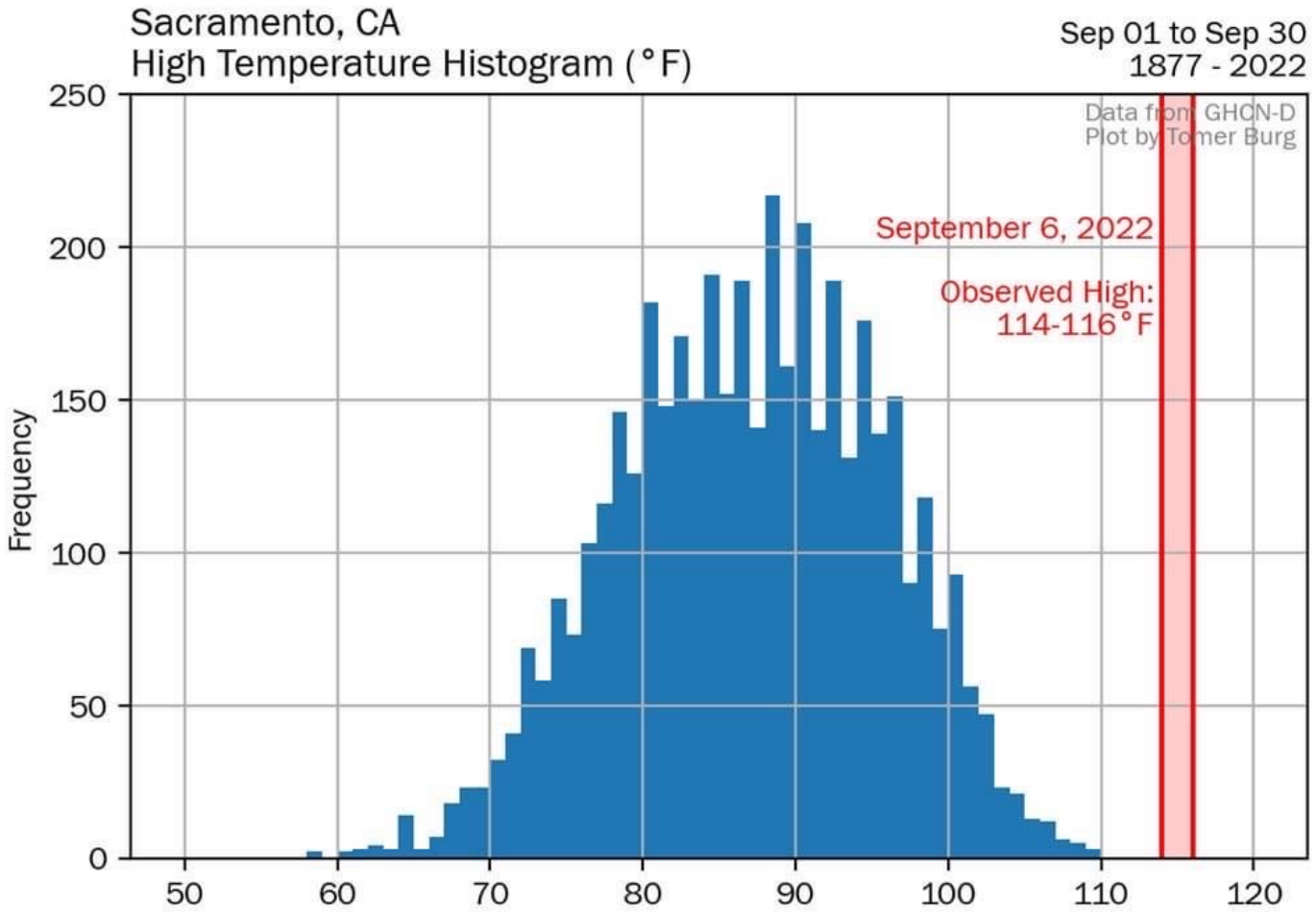


Example Visualizations from Discussions

Charles



Example Visualizations from Discussions



Shanshan

Today: Bivariate Analysis & Viz

Bivariate in the context of statistics

- Descriptive
  - Univariate: single variable
  - **Bivariate: relationships between two variables**
  - Multivariate: relationships between multiple variables

- Inferential
  - Hypothesis testing: testing whether your assumptions are true or not
  - Model fitting: (predictive analytics) use your insights to make decisions or predictions

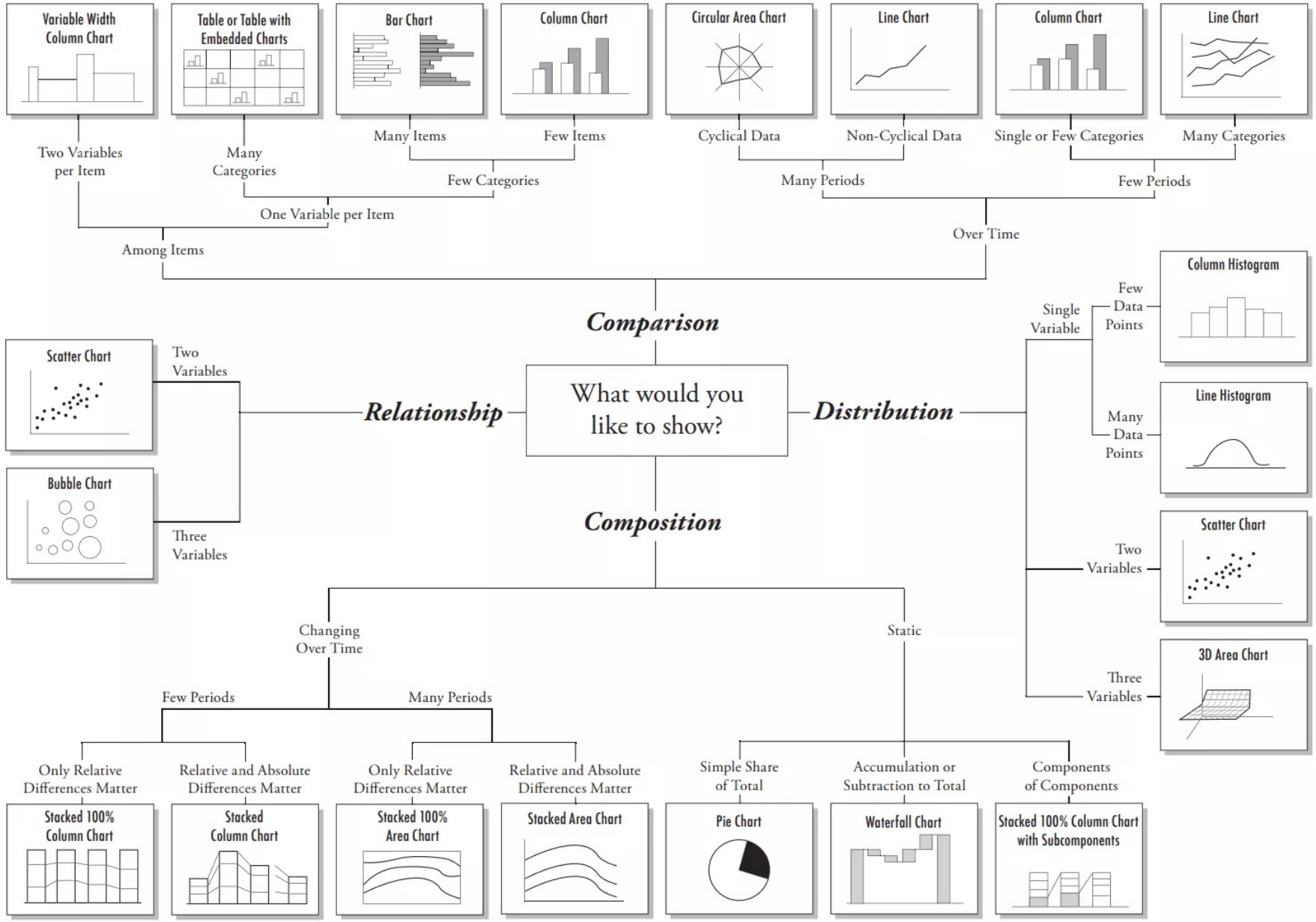
Bivariate in the context of statistics

- Descriptive
  - Univariate: single variable
  - **Bivariate: relationships between two variables**
    - Comparison, dependency, relationship, association
  - Multivariate: relationships between multiple variables
- Inferential
  - Hypothesis testing: testing whether your assumptions are true or not
  - Model fitting: (predictive analytics) use your insights to make decisions or predictions

What can you visualize

- Distribution
  - How is a data variable distributed over a range of values?
- Composition
  - Is a data variable composed of different subgroups?
- Comparison
  - How do trends in different data variables compare?
- Relationship
  - Is a trend in one data variable related to another variable?
- Space and time

Chart Suggestions—A Thought-Starter



Measures of frequency

Histograms & Bar Charts (for Numerical & Categorical)

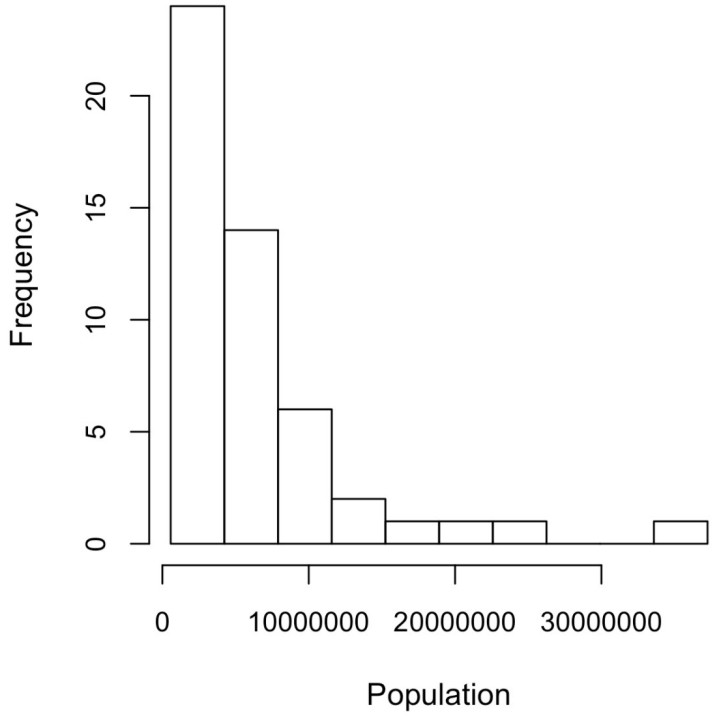


Figure 1-3. Histogram of state populations

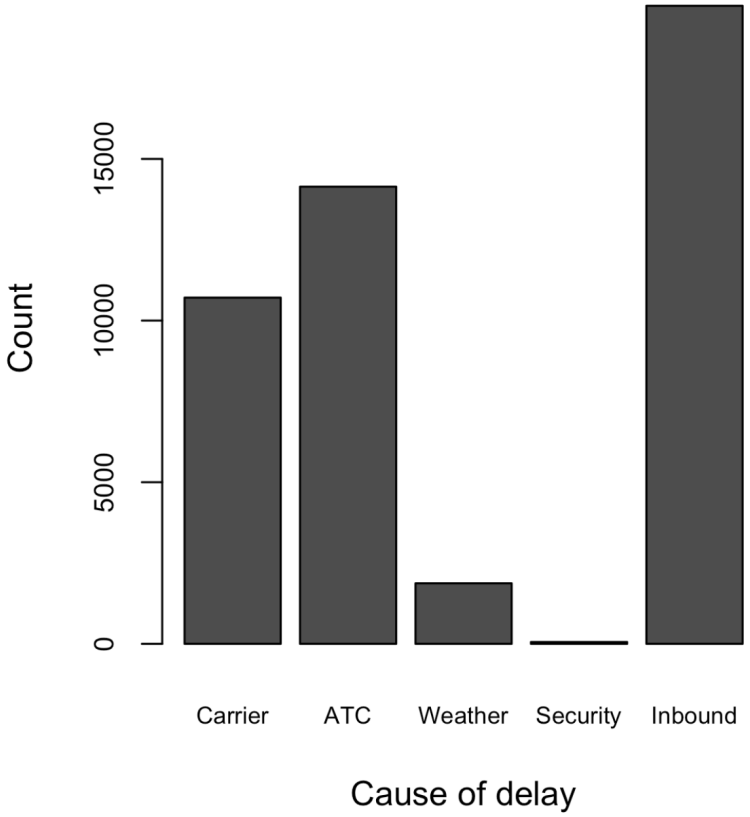


Figure 1-5. Bar chart of airline delays at DFW by cause

*Practical Statistics for Data Scientists, by Peter Bruce, Andrew Bruce, and Peter Gedeck

Exploring the variation of a variable's values

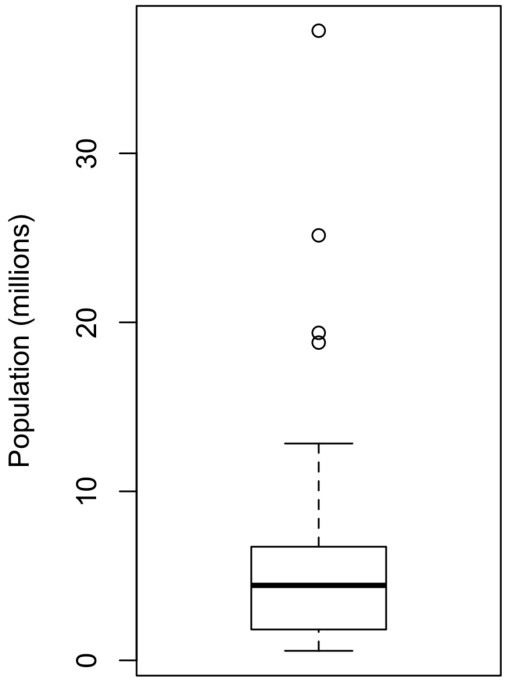


Figure 1-2. Boxplot of state populations

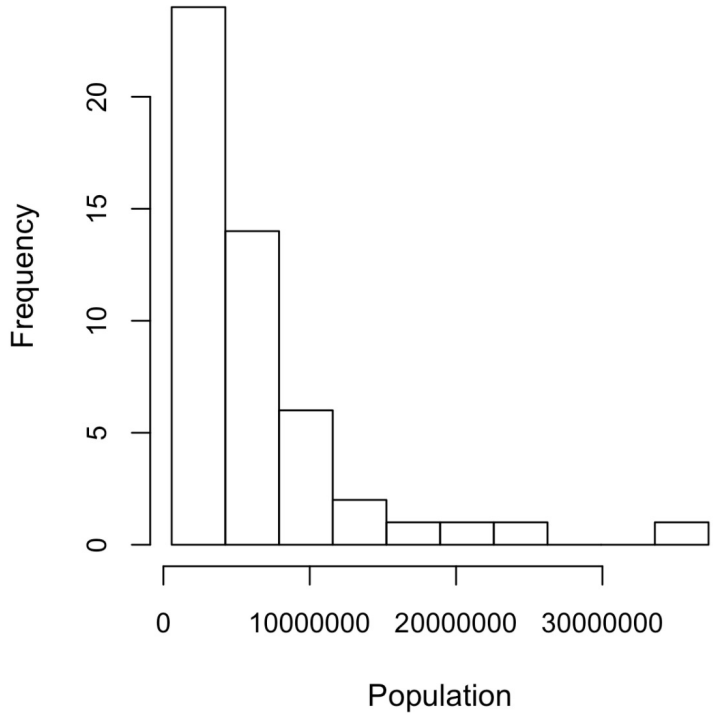


Figure 1-3. Histogram of state populations

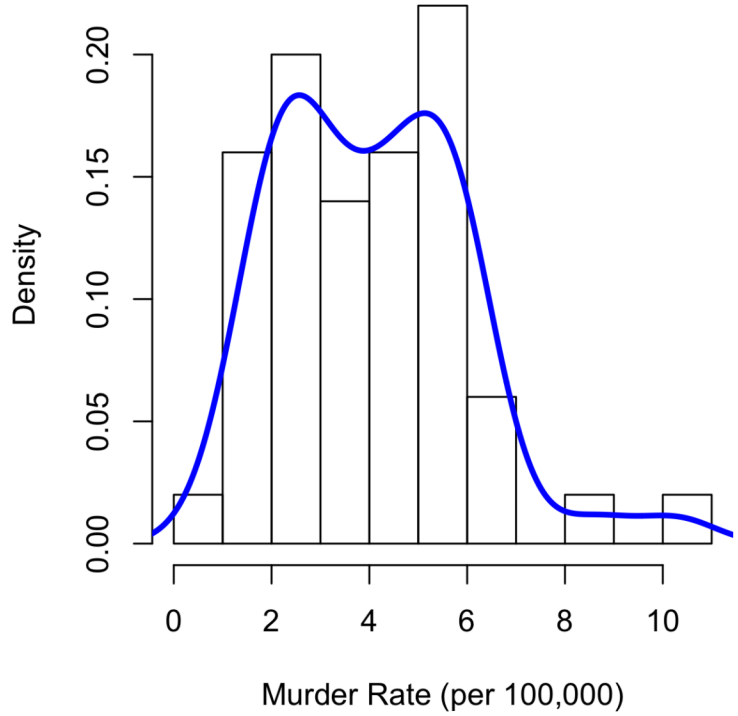


Figure 1-4. Density of state murder rates

*Practical Statistics for Data Scientists, by Peter Bruce, Andrew Bruce, and Peter Gedeck

Though box plots and distributions can be shown within categories as well

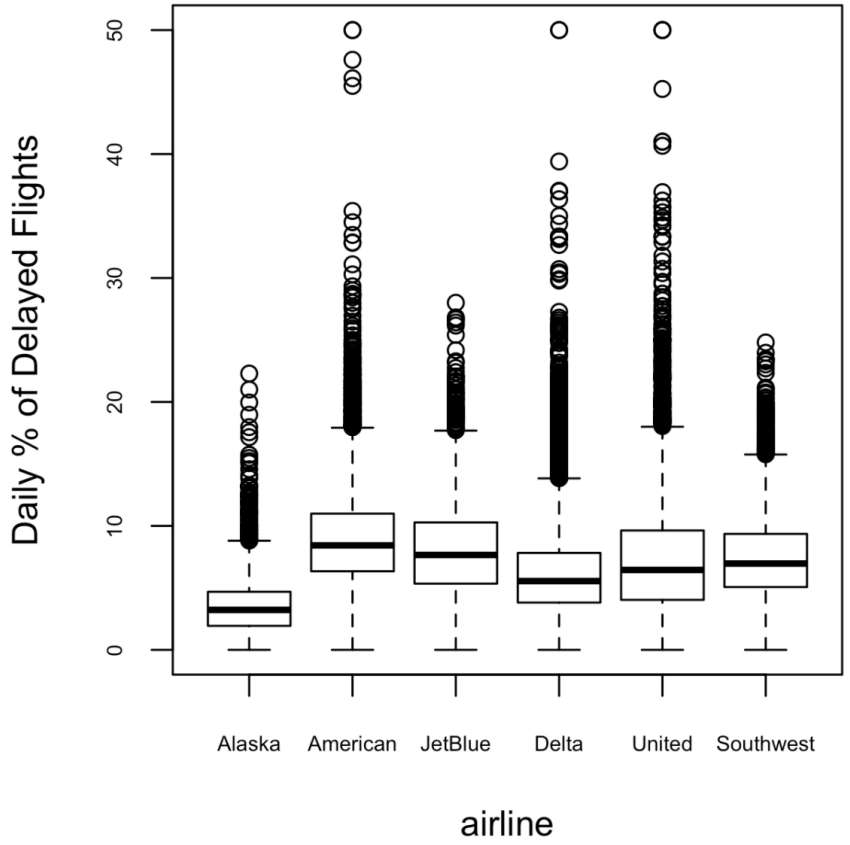


Figure 1-10. Boxplot of percent of airline delays by carrier

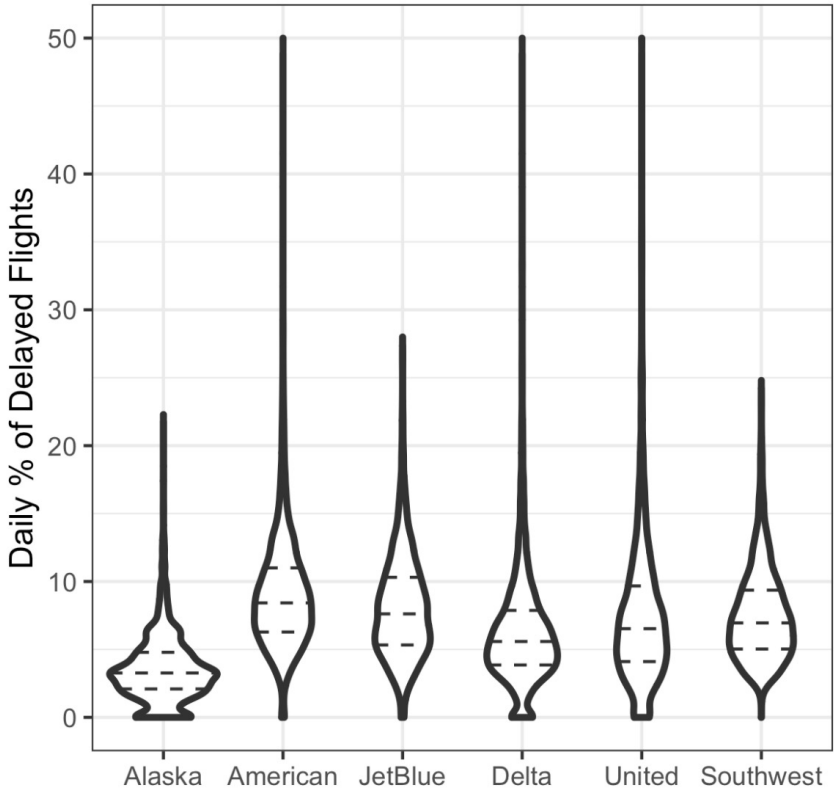


Figure 1-11. Violin plot of percent of airline delays by carrier

*Practical Statistics for Data Scientists, by Peter Bruce, Andrew Bruce, and Peter Gedeck

Exploring the variation between two different variable's values: Scatter Plot

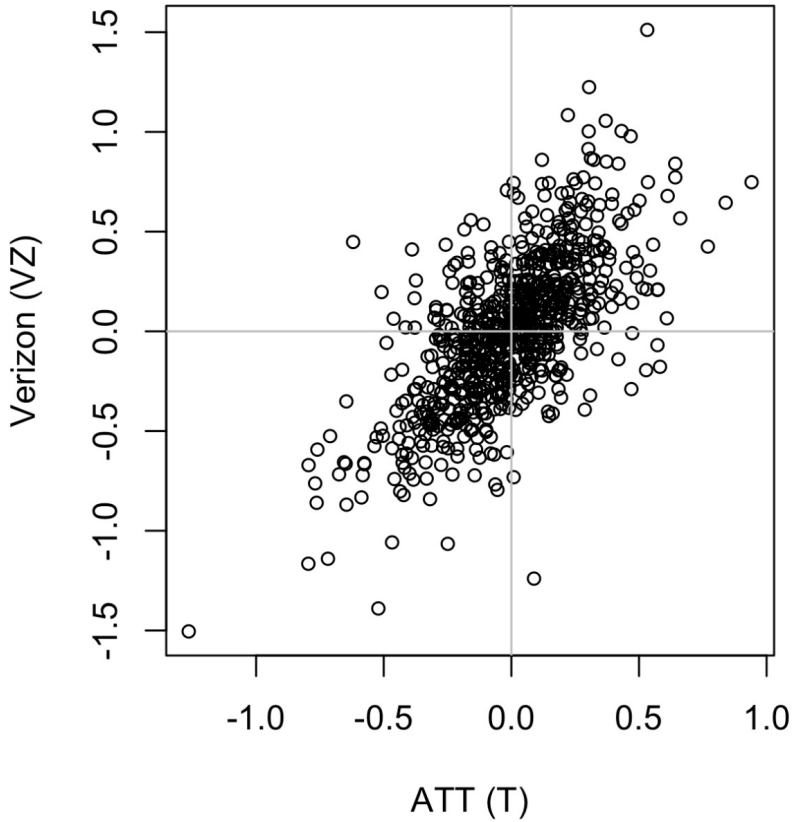


Figure 1-7. Scatterplot of correlation between returns for ATT and Verizon

*Practical Statistics for Data Scientists, by Peter Bruce, Andrew Bruce, and Peter Gedeck

Exploring the variation between two different variable's values: Scatter Plot & Correlation

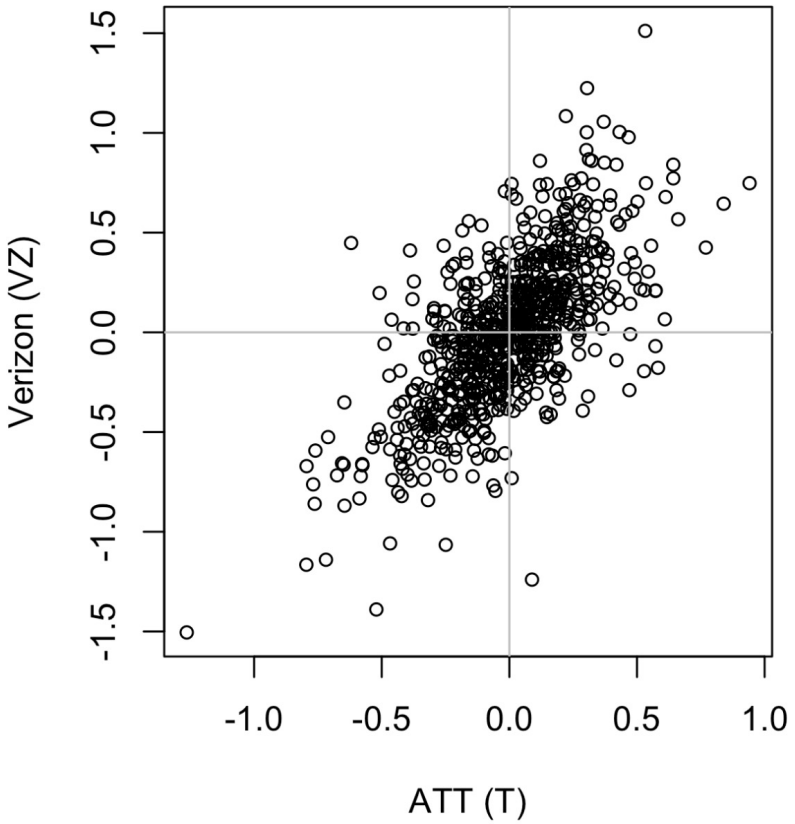


Figure 1-7. Scatterplot of correlation between returns for ATT and Verizon

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}$$

r = +1 : perfectly correlated
r = -1 : perfectly anti-correlated
r = 0 : not correlated

*Practical Statistics for Data Scientists, by Peter Bruce, Andrew Bruce, and Peter Gedeck

Graphical Aside: If there are lots of data points, a scatter plot can get too busy

Hexagonal binning plot example

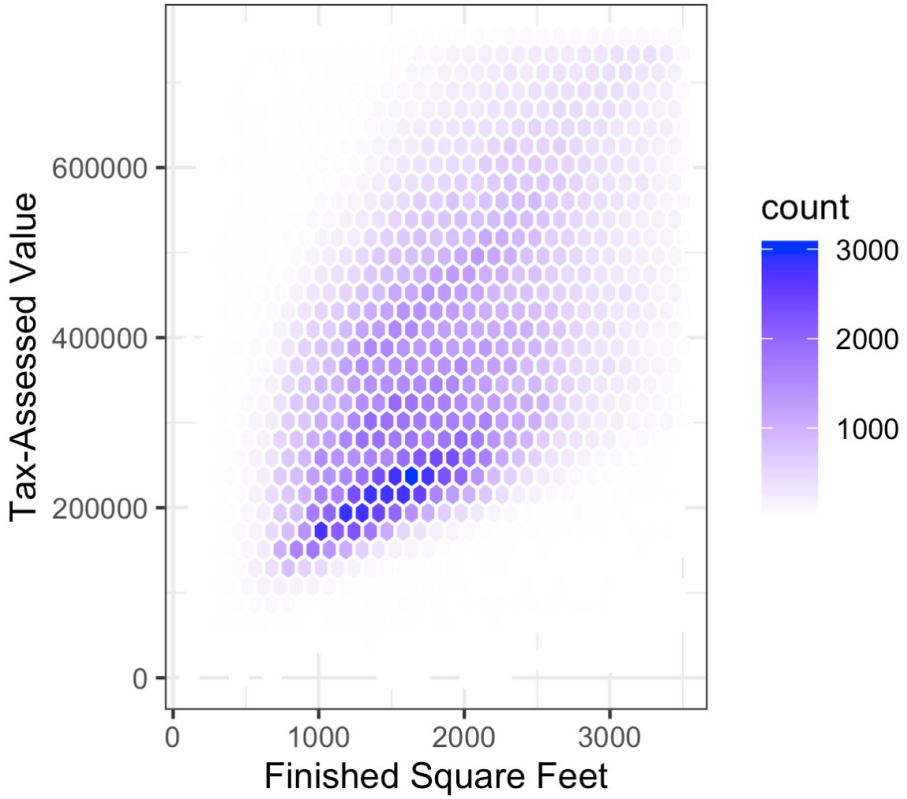


Figure 1-8. Hexagonal binning for tax-assessed value versus finished square feet

Contour plot example

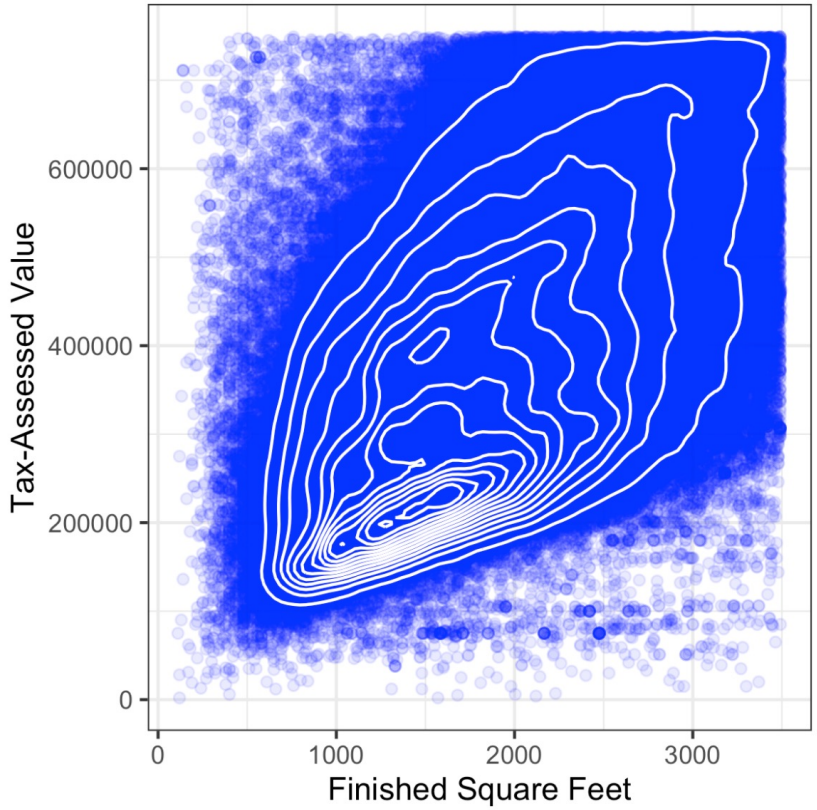
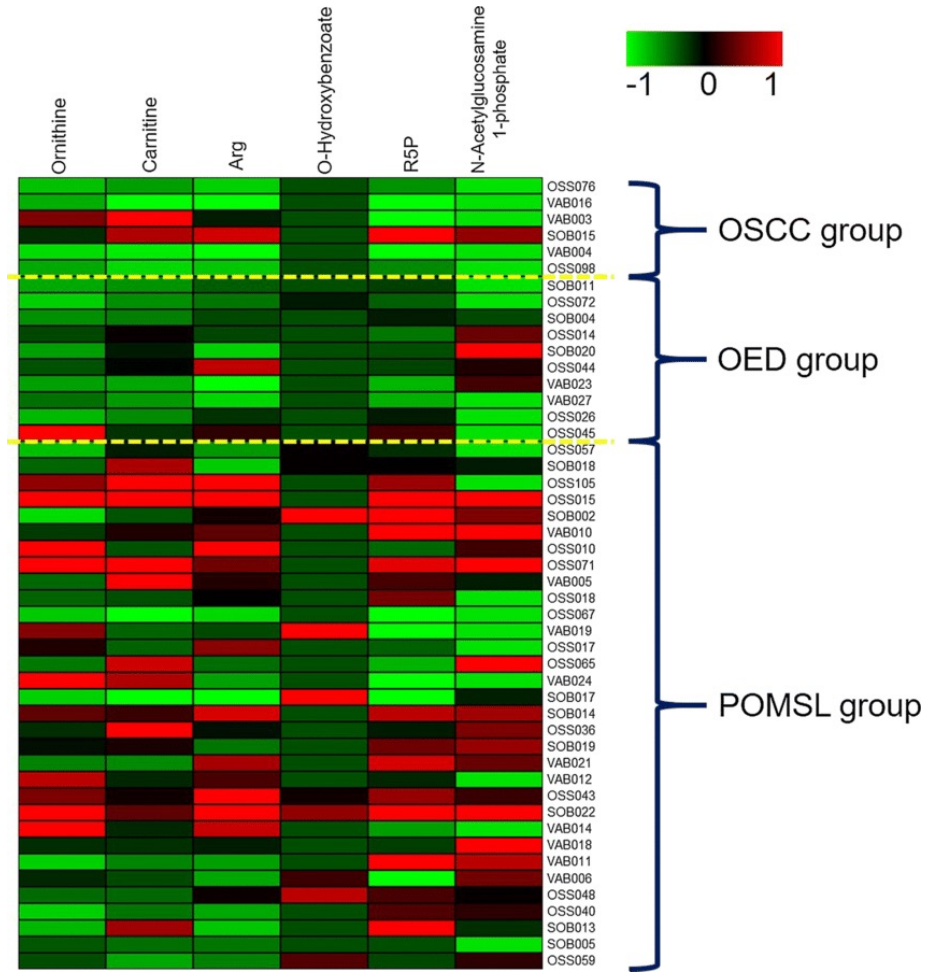


Figure 1-9. Contour plot for tax-assessed value versus finished square feet

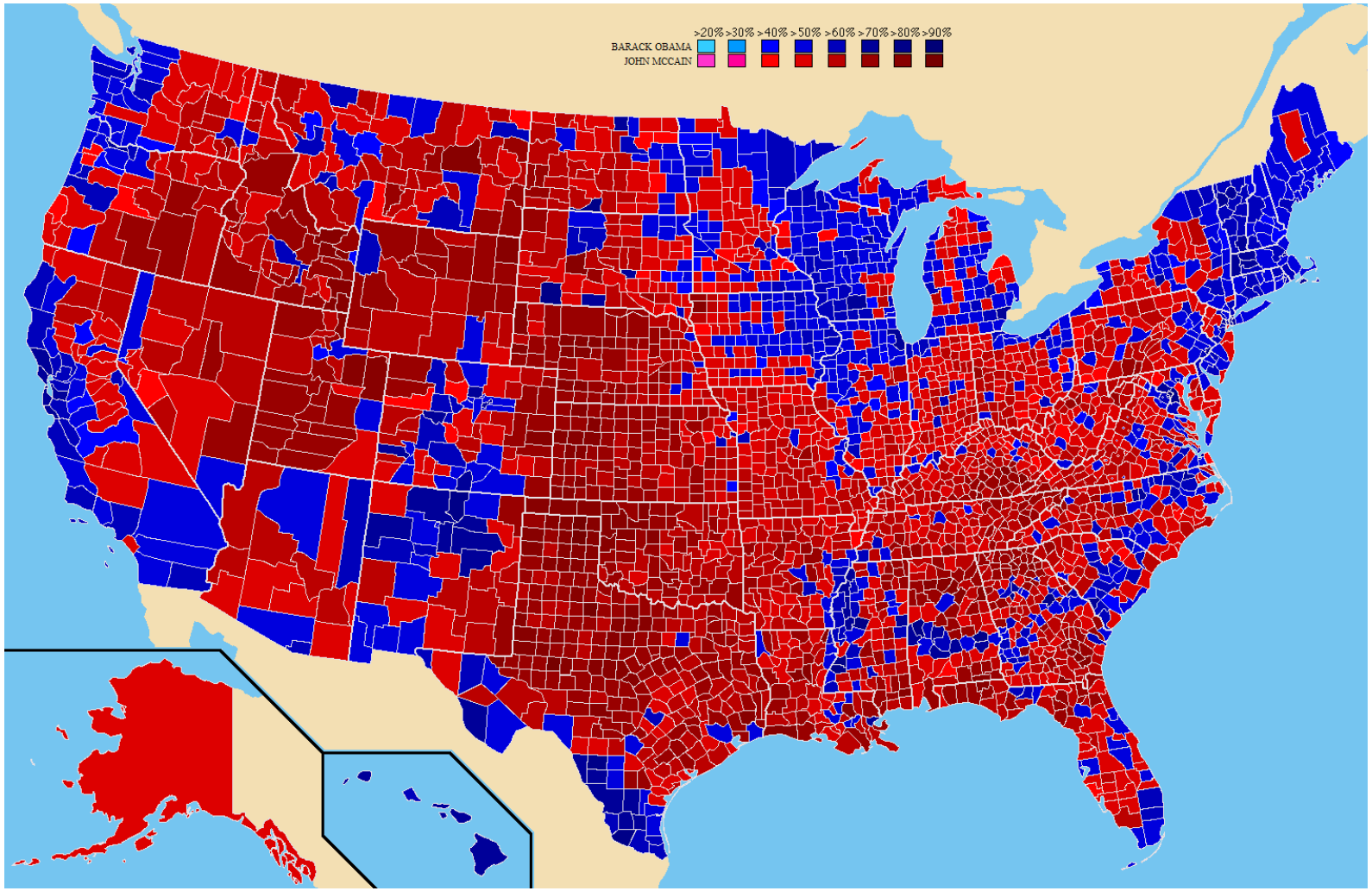
*Practical Statistics for Data Scientists, by Peter Bruce, Andrew Bruce, and Peter Gedeck

Scatter plot for categories??

Scatter plot for categories: Heat map
(here for salivary metabolomic profiles)

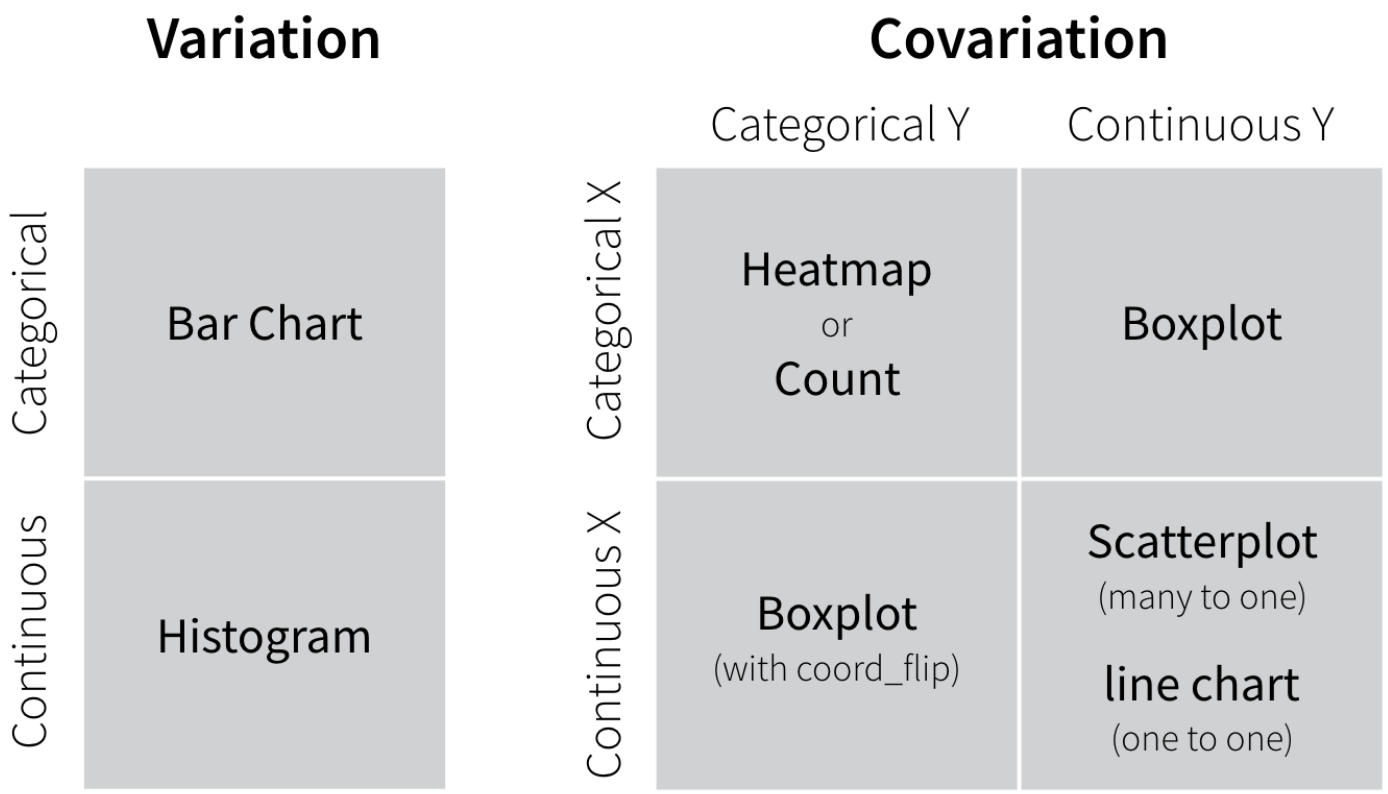


Heat “Map”: here categorical vs numerical



https://visage.co/data-visualization-101-heat-maps/

Measures of variation



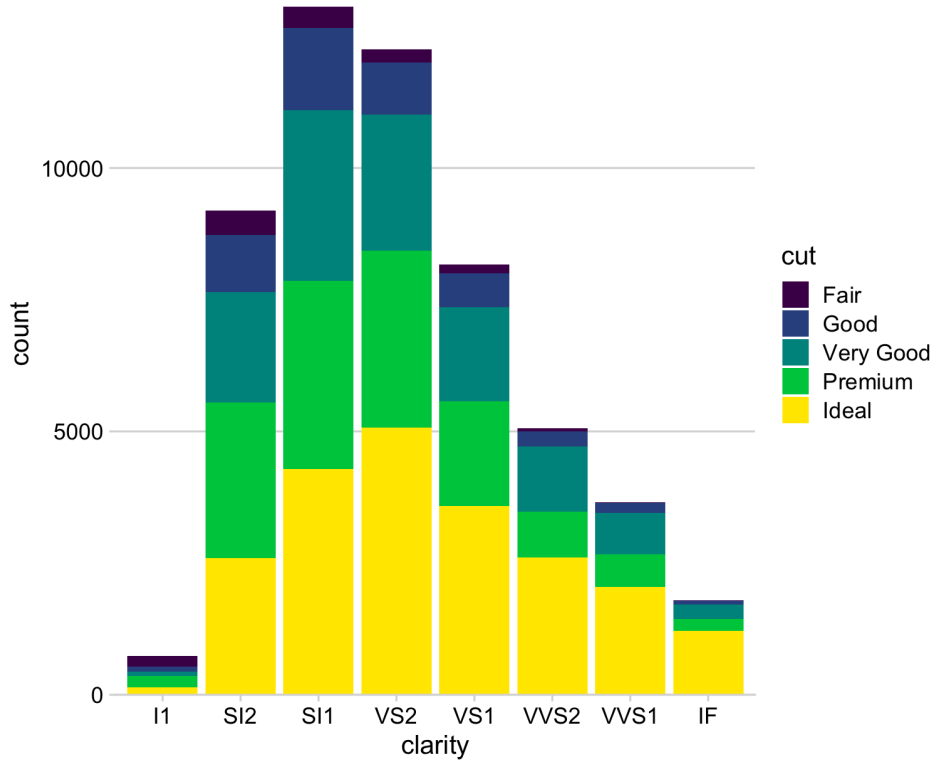
*Practical Statistics for Data Scientists, by Peter Bruce, Andrew Bruce, and Peter Gedeck

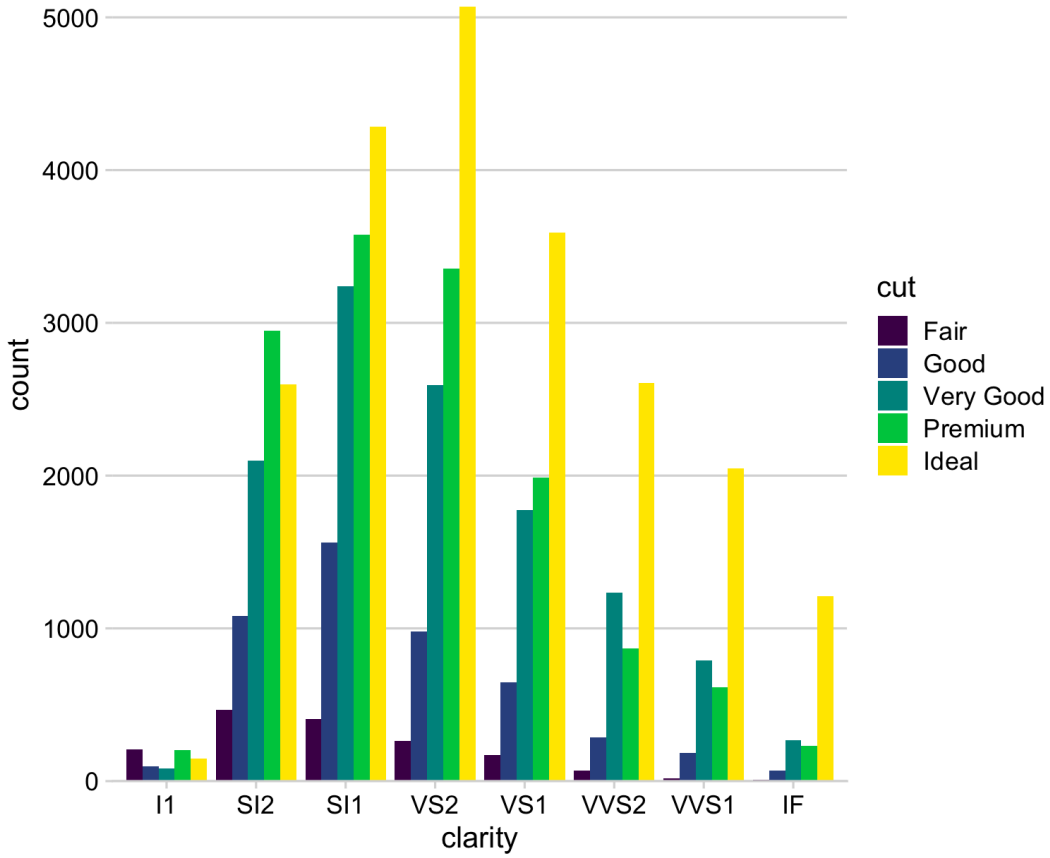
Measures of variation

Let’s extend last week’s Tableau examples

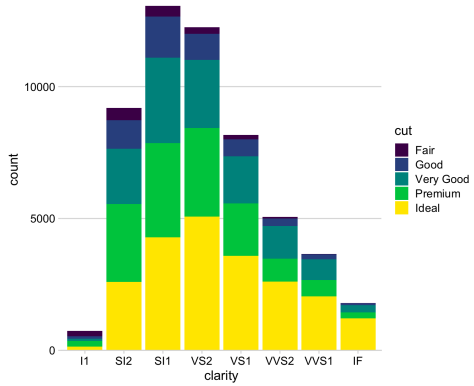
*Practical Statistics for Data Scientists, by Peter Bruce, Andrew Bruce, and Peter Gedeck

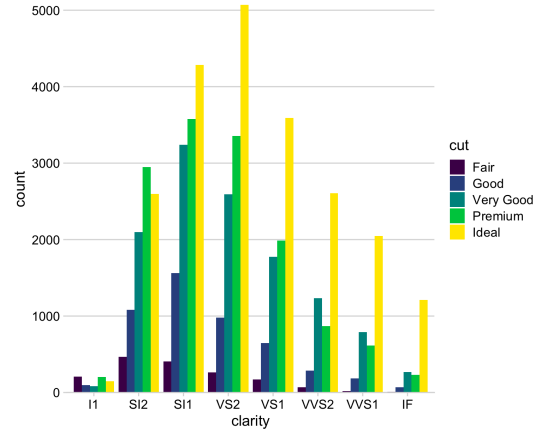
Bar plots of two categorical variables

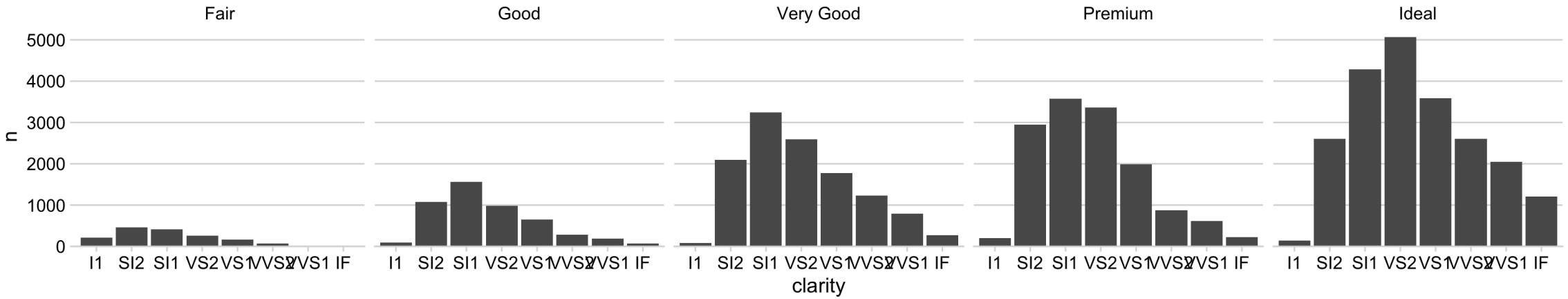




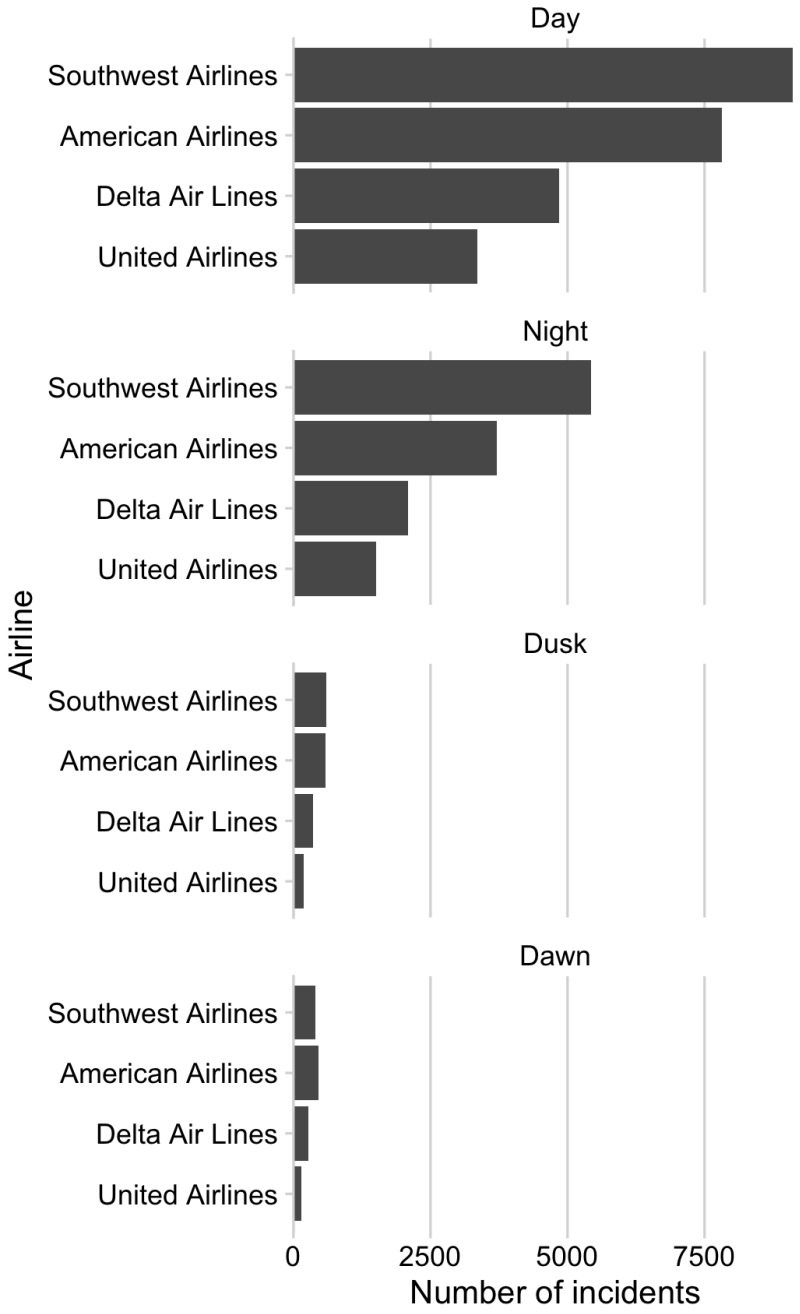
Bar plots of two categorical variables
-> splitting into facets





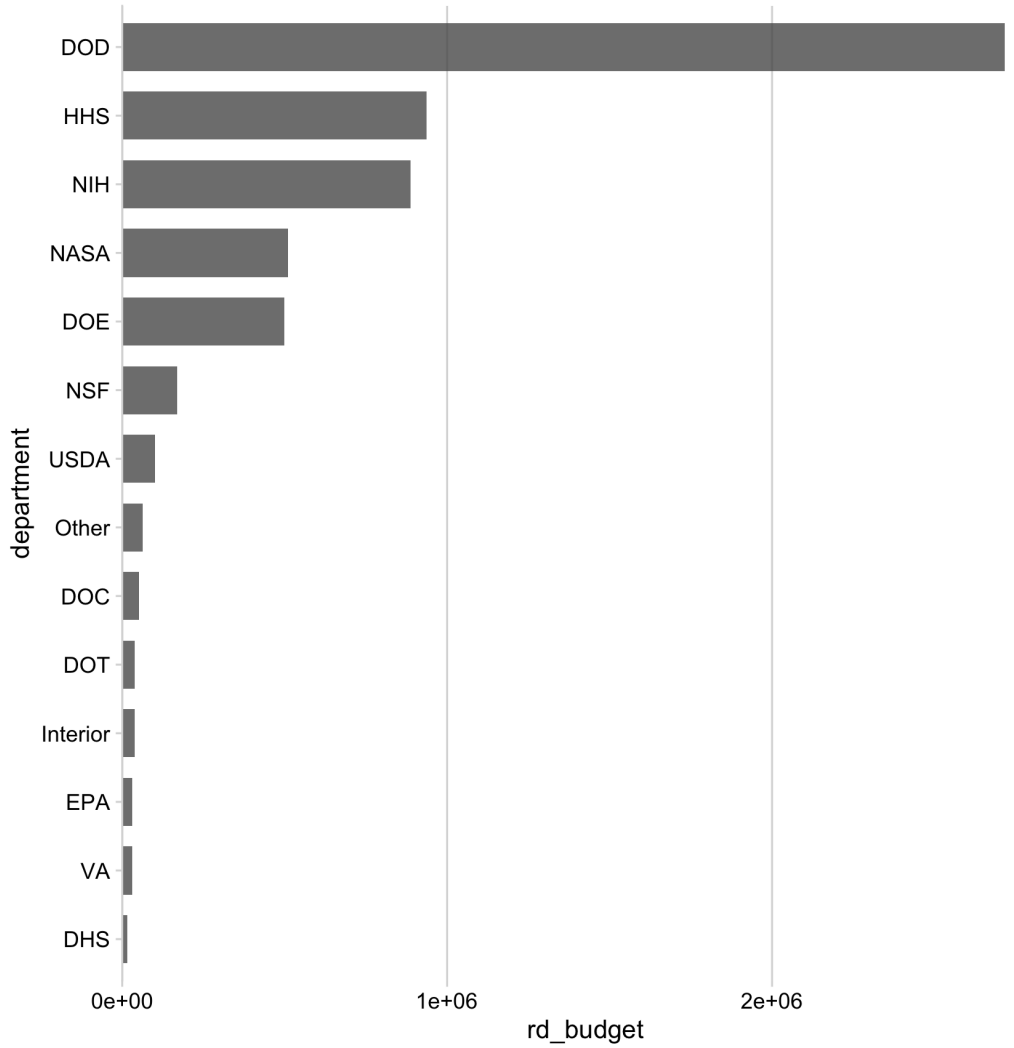


Variations

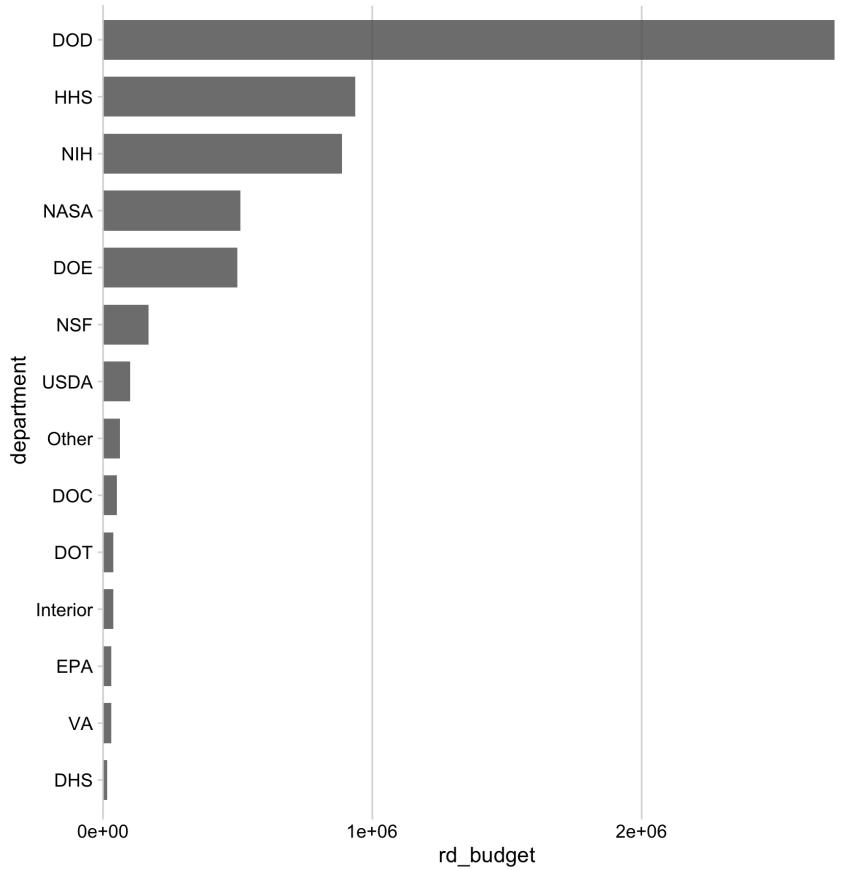


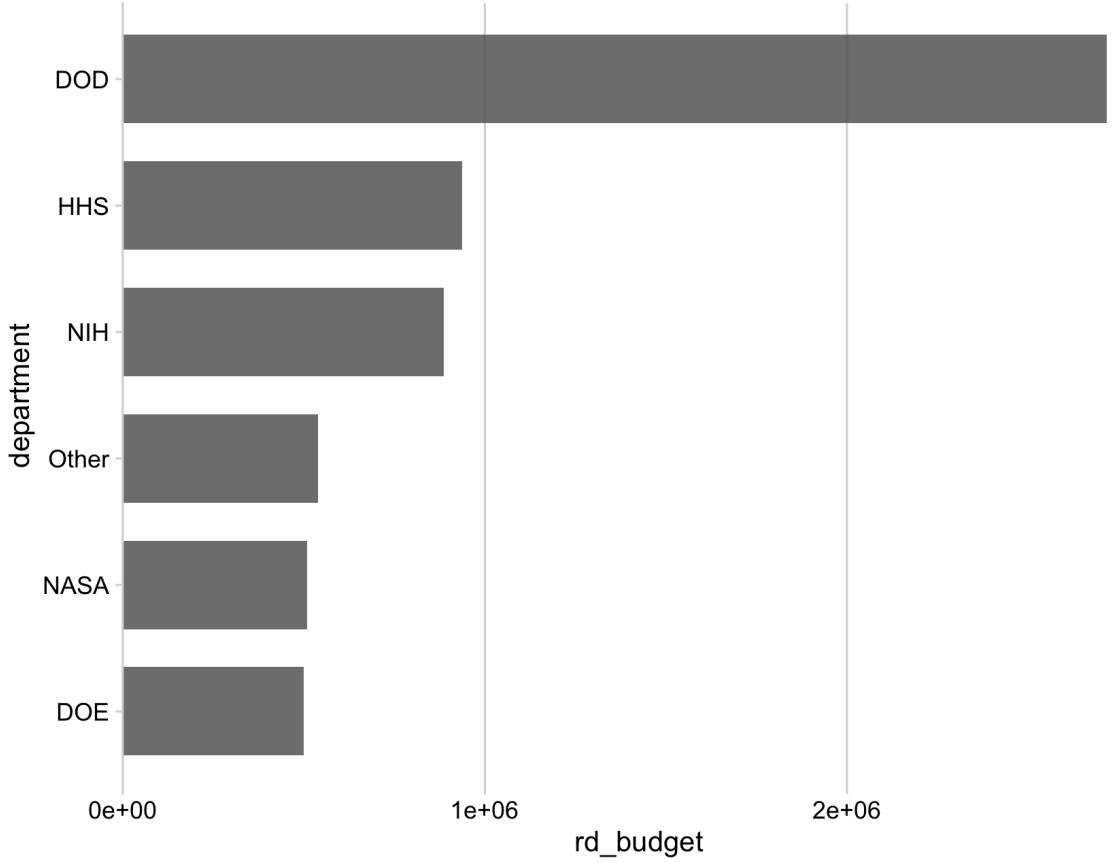
(not really a bivariate plot)

What if there are too many levels?
(is this univariate?)

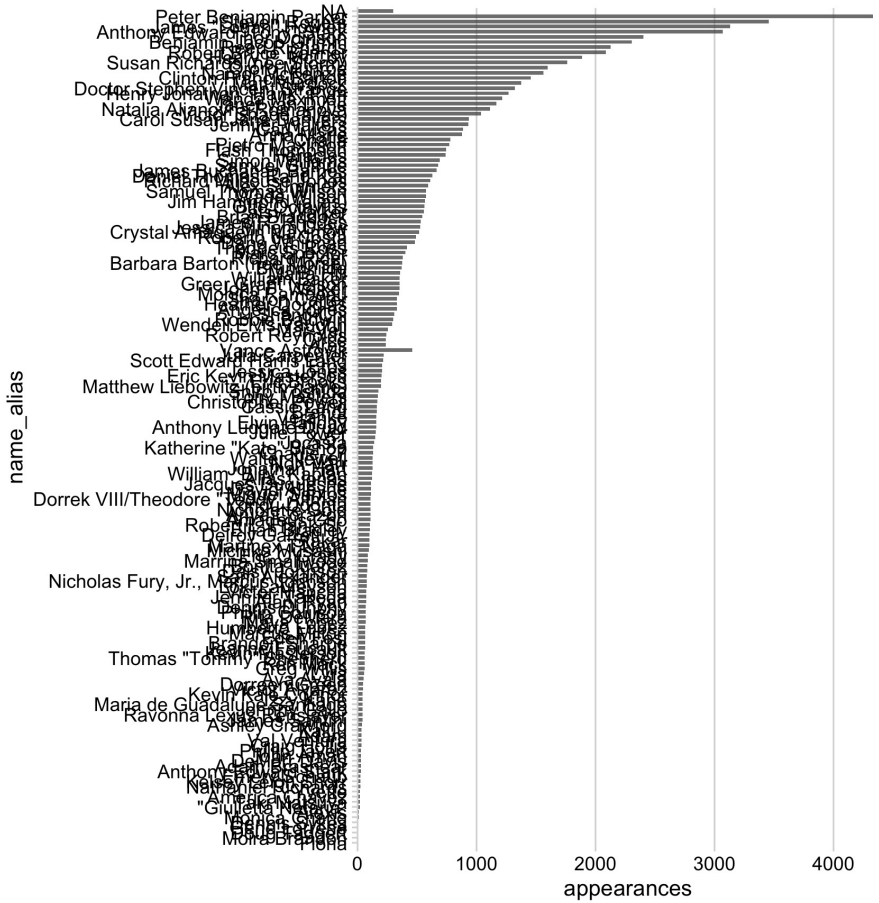


What if there are too many levels?
-> include extraneous into an “Other” category

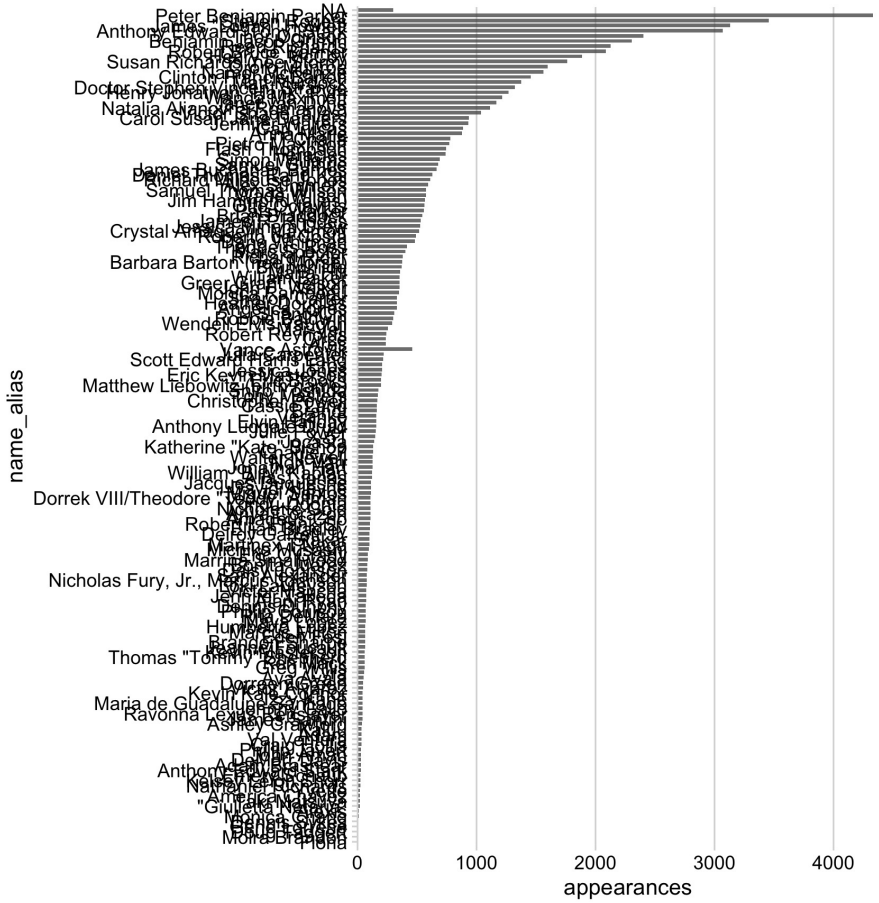


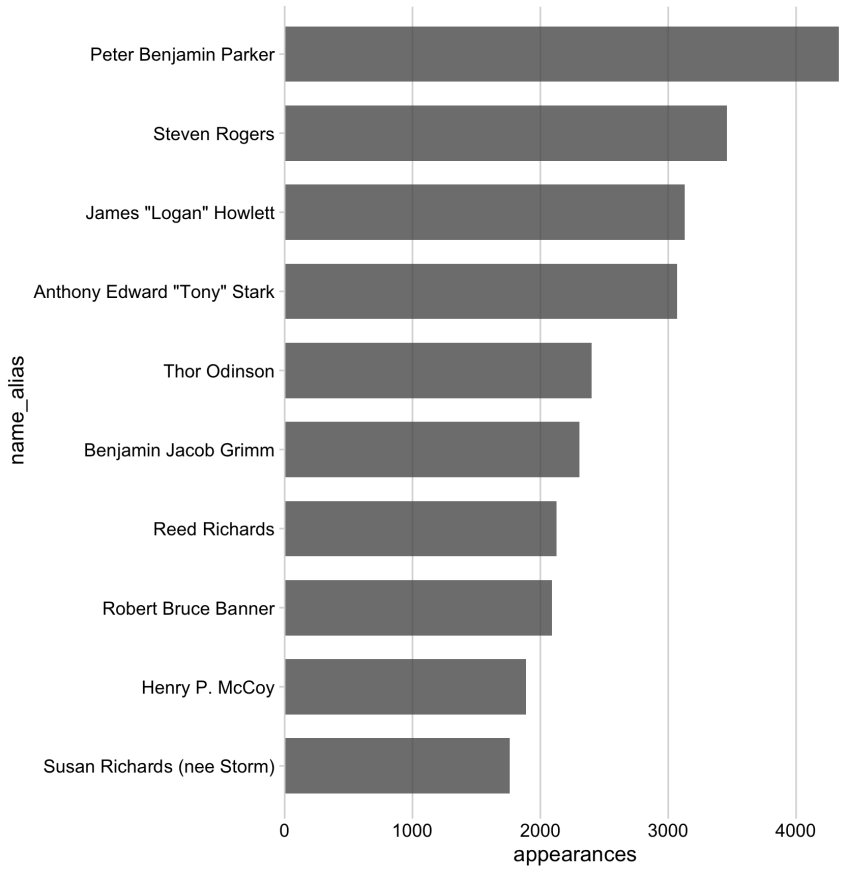


What if there are TOO TOO many levels?

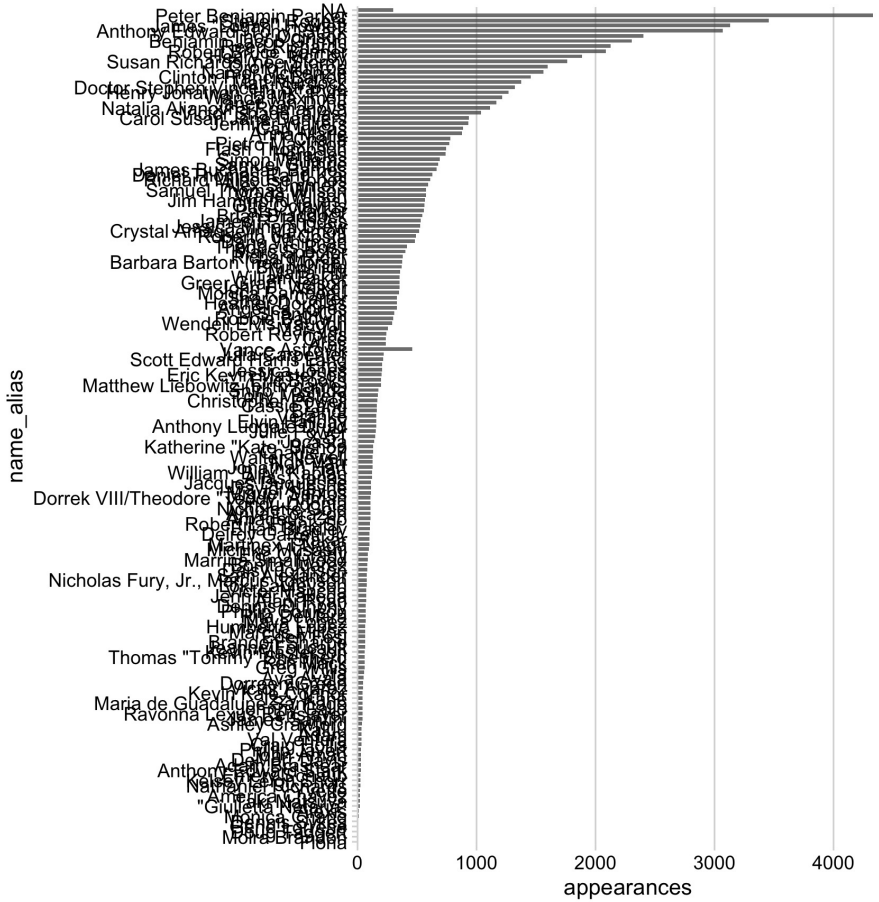


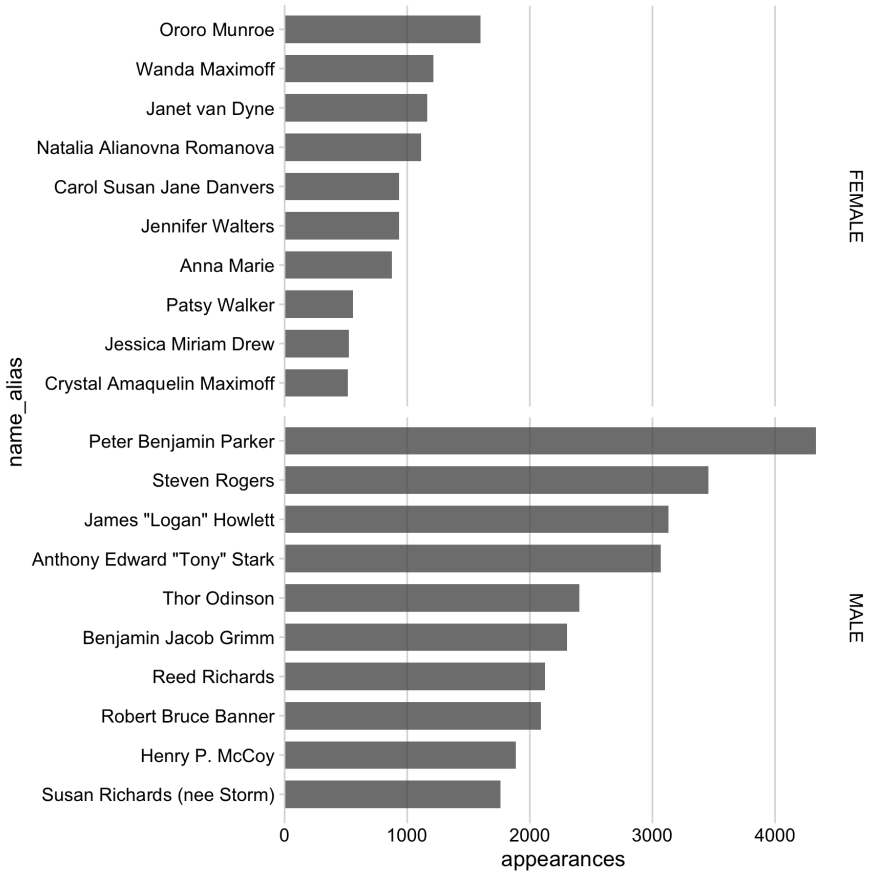
What if there are TOO TOO many levels?
-> keep the top N



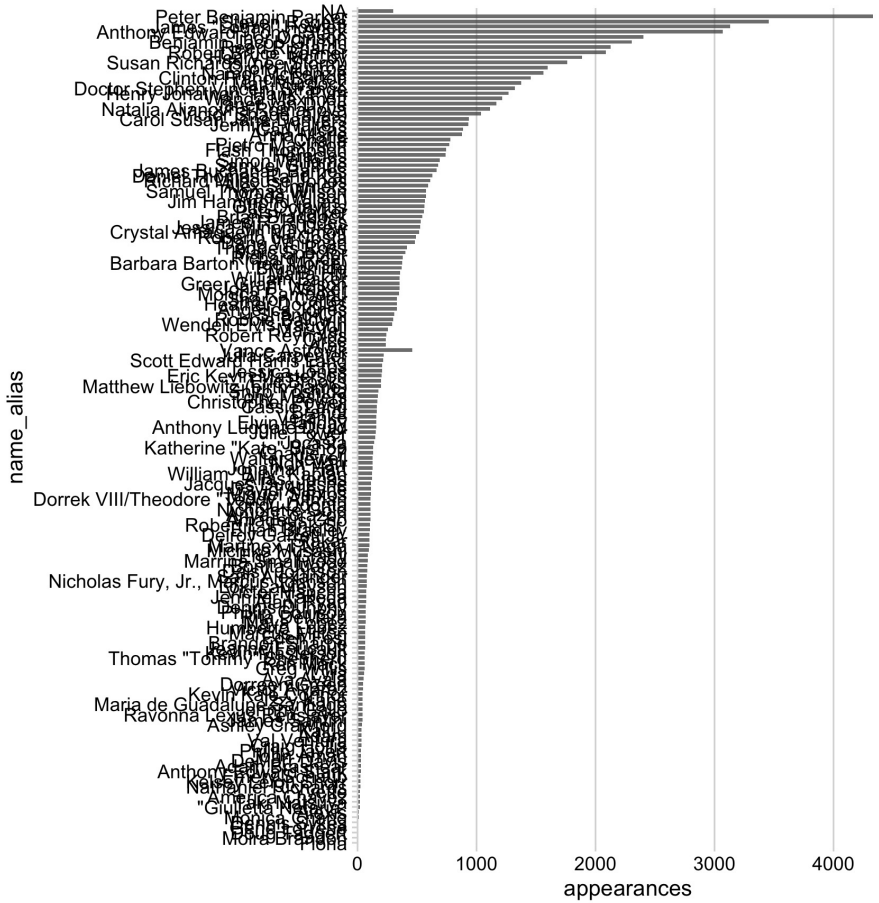


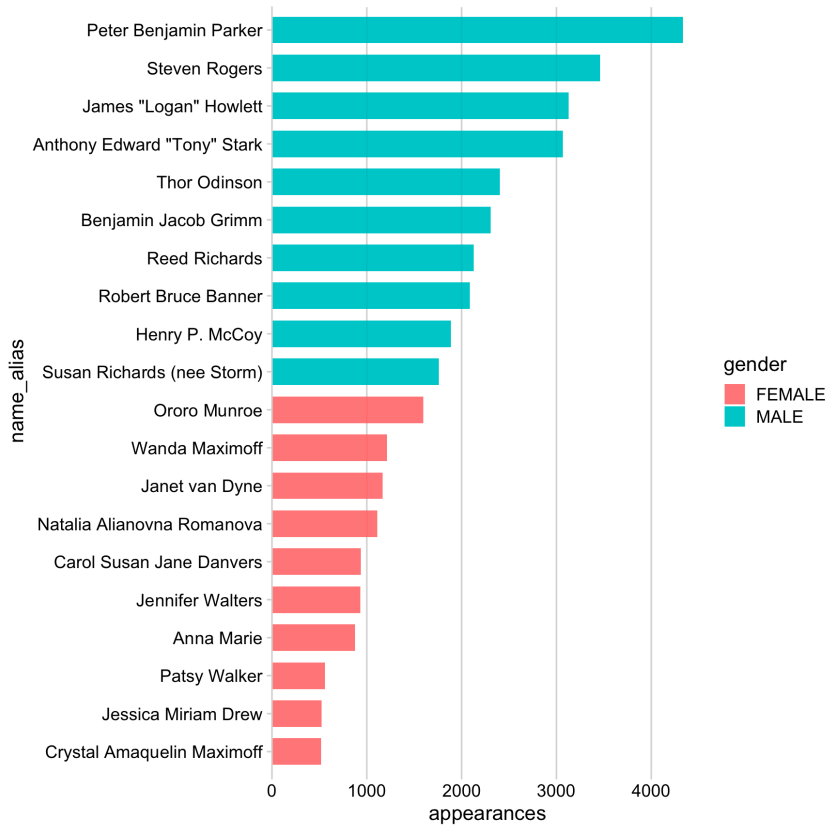
What if there are TOO TOO many levels?
-> can take advantage of faceting
(and is this bivariate?)



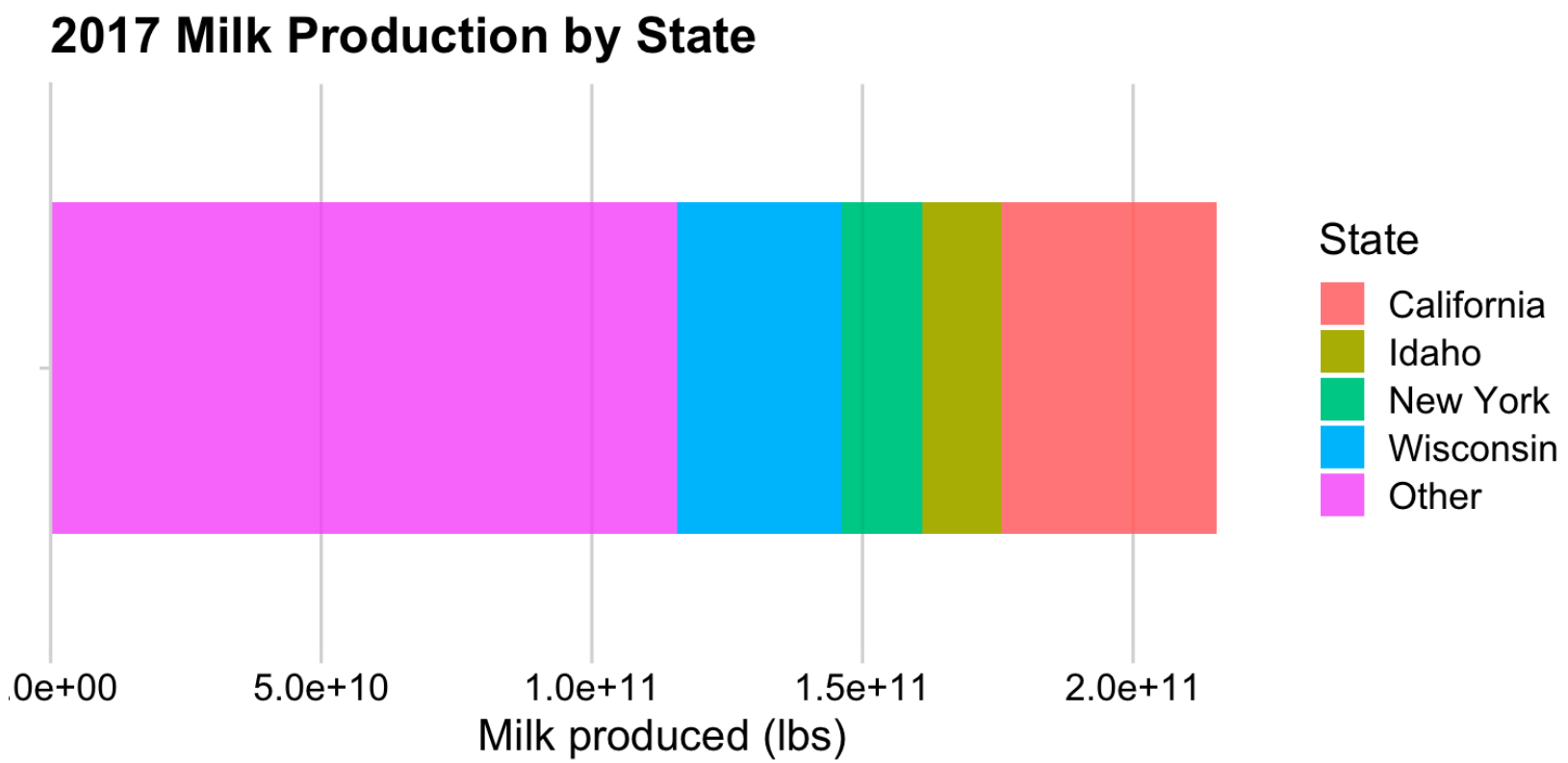


What if there are TOO TOO many levels?
-> can take advantage of color
(and is this bivariate?)



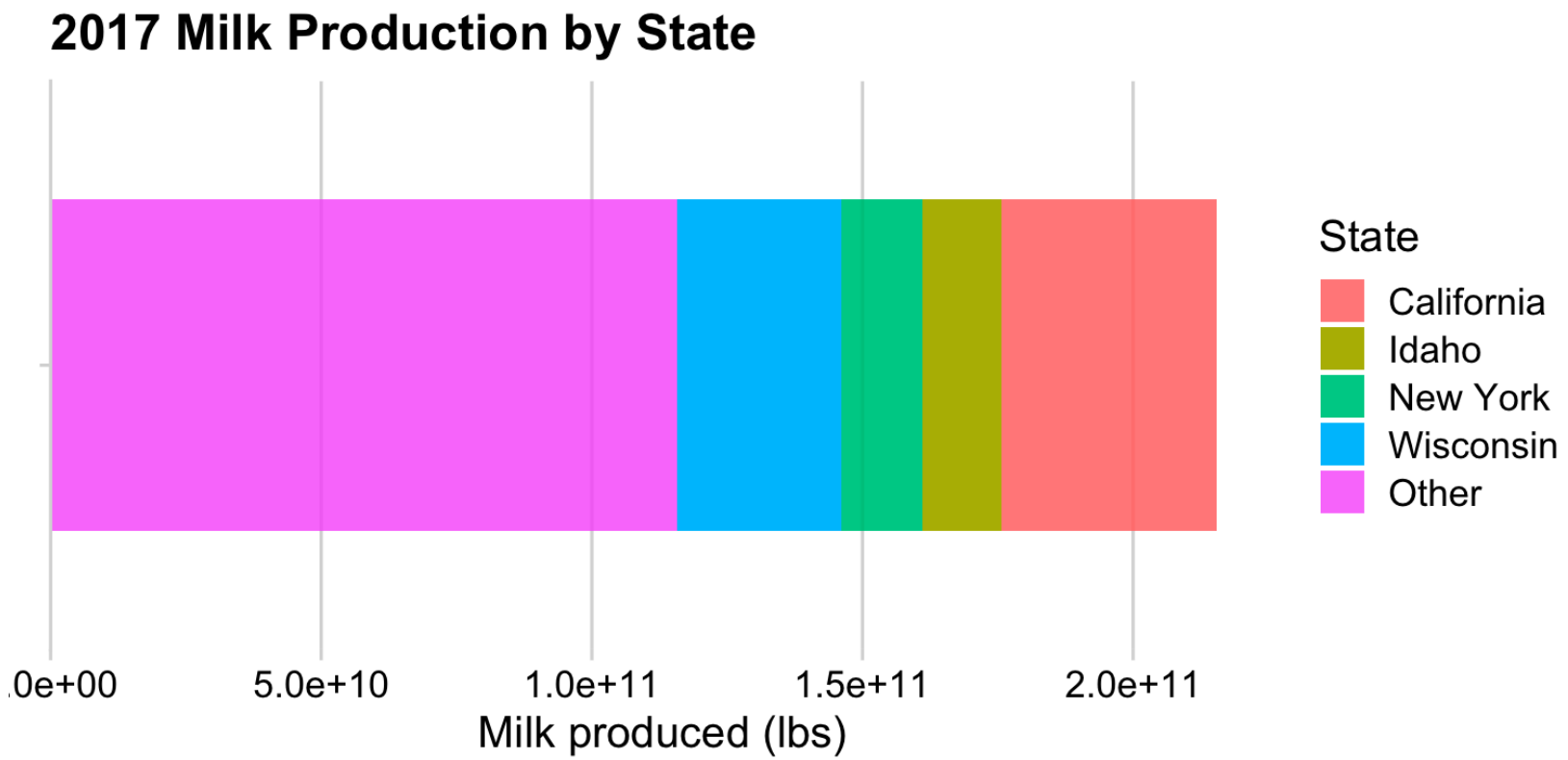


Stacked bars



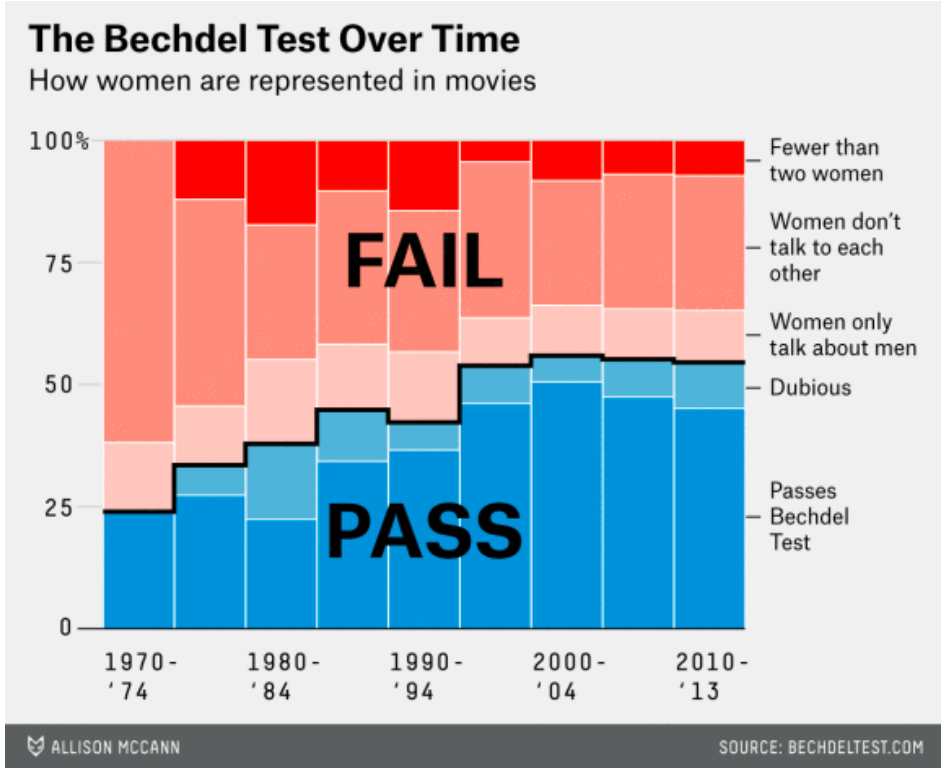
Stacked bars

- Not good for more than a couple categories



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

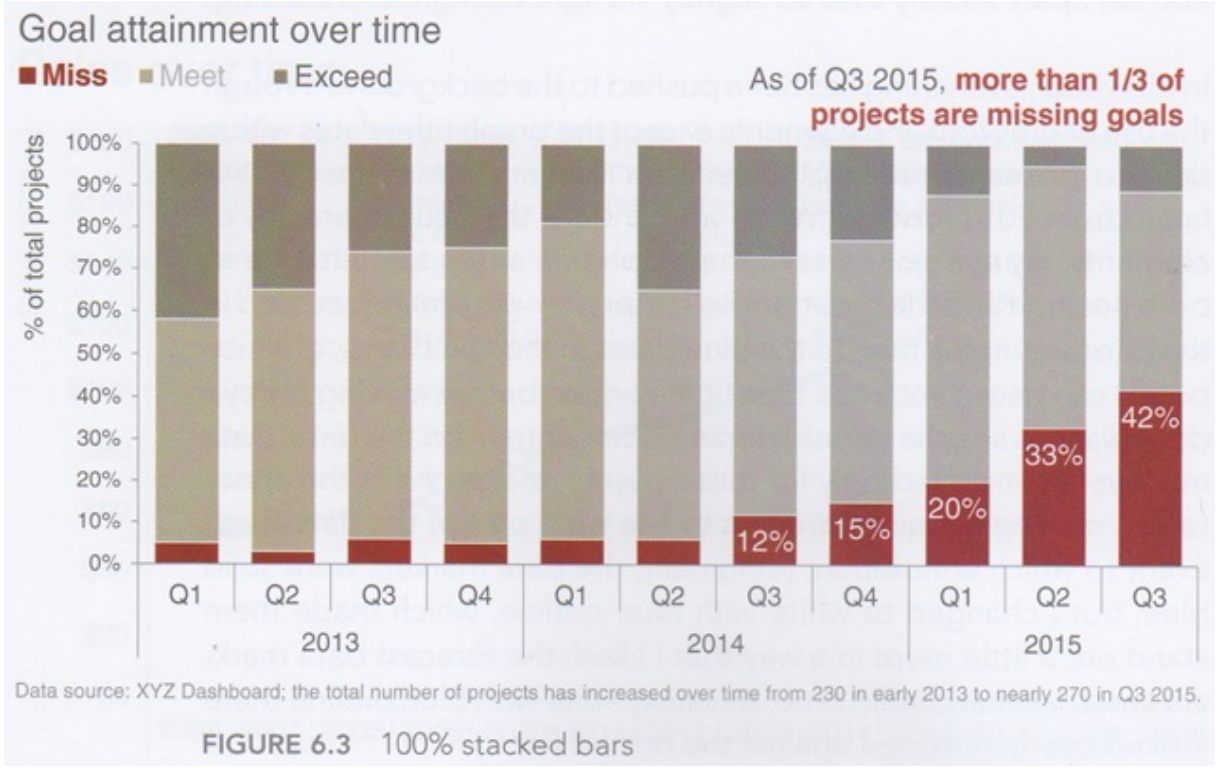
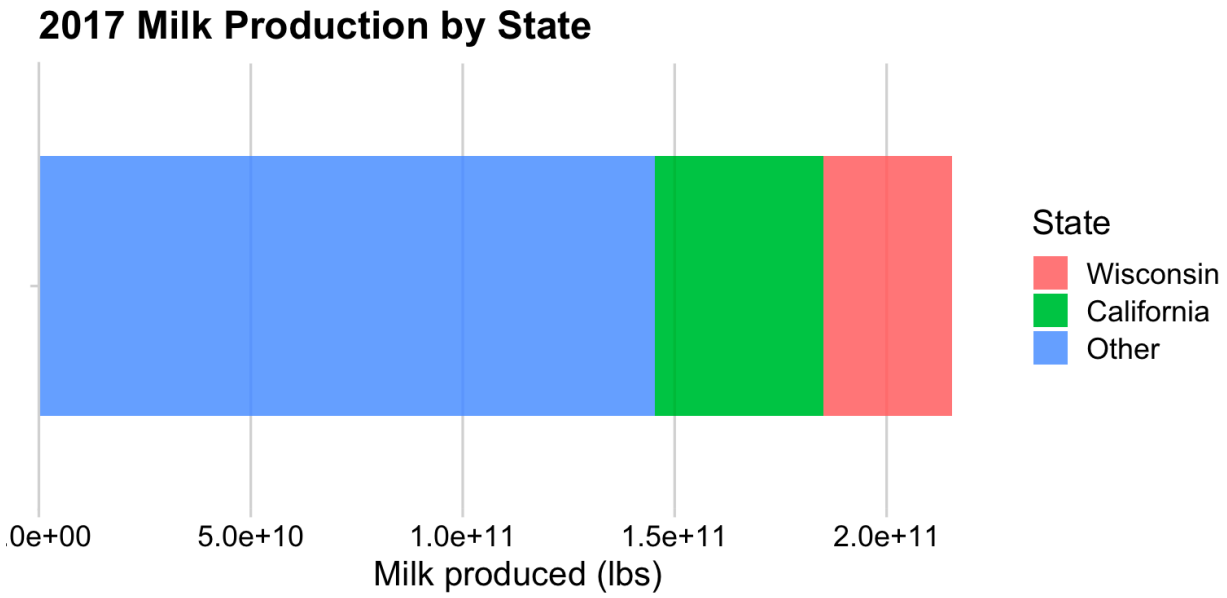


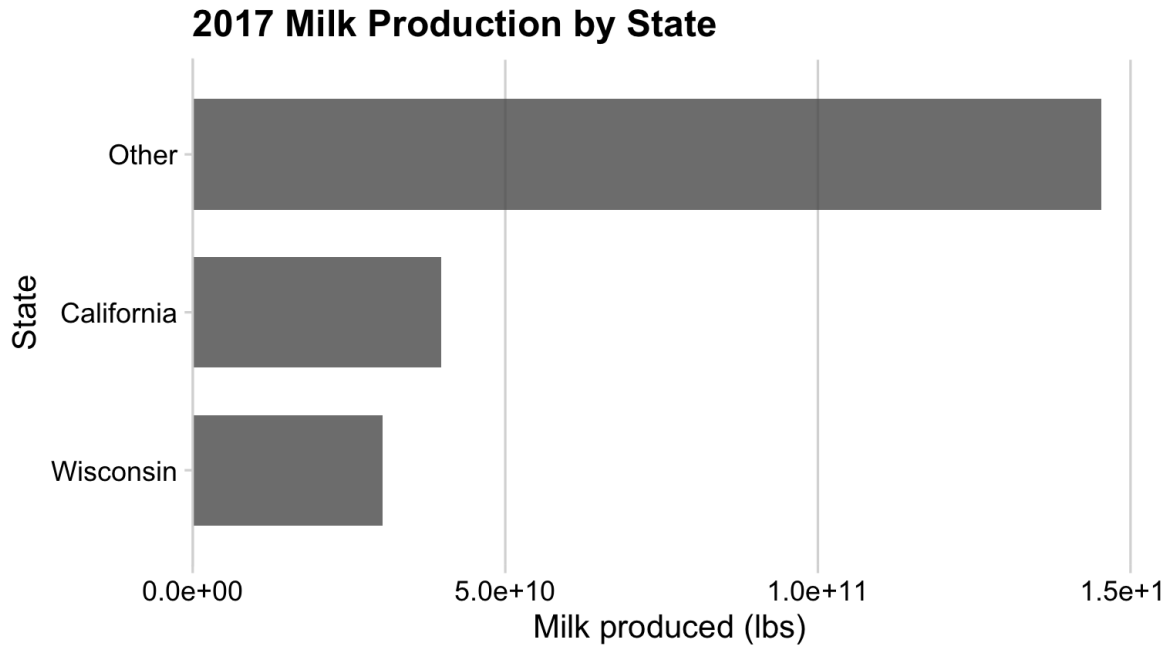
FIGURE 6.3 100% stacked bars

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

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

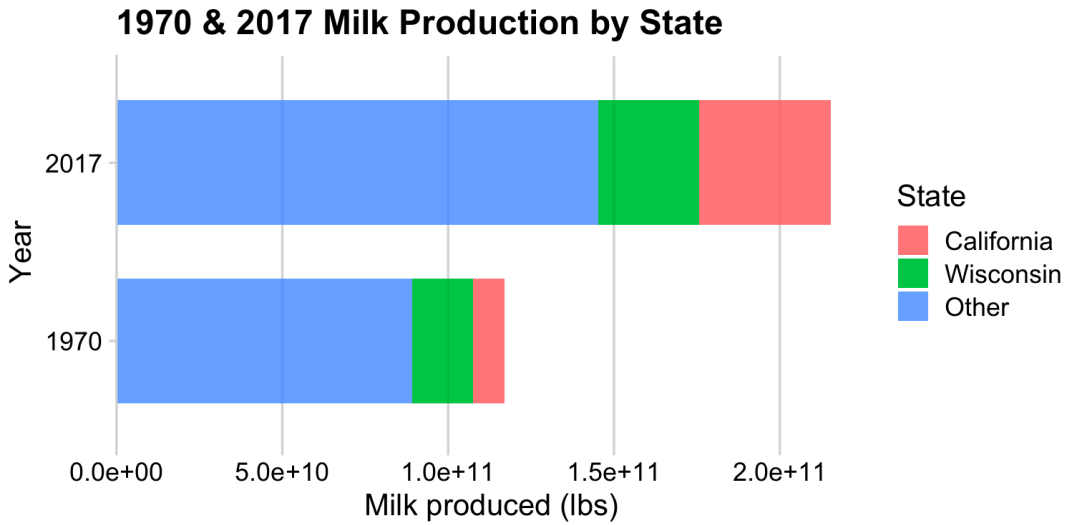
- Stacked : better for showing single part-to-whole comparison



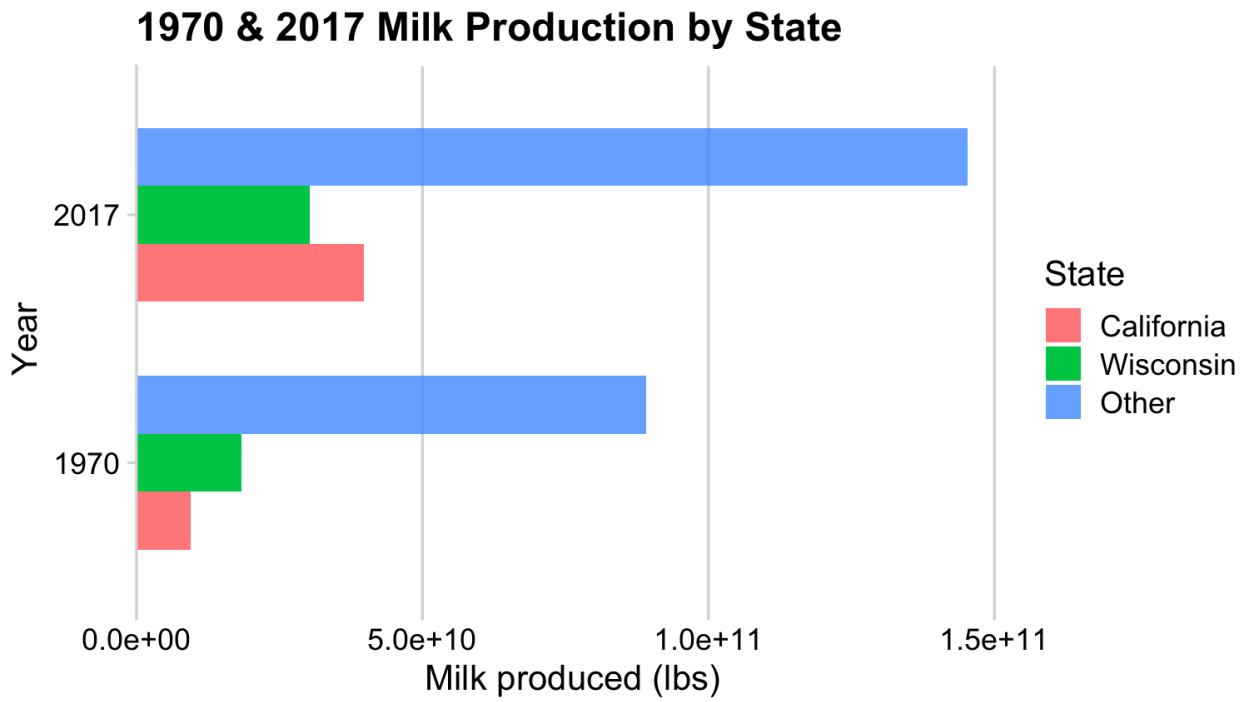


Dodged bars

- Dodged : better for comparing individual components

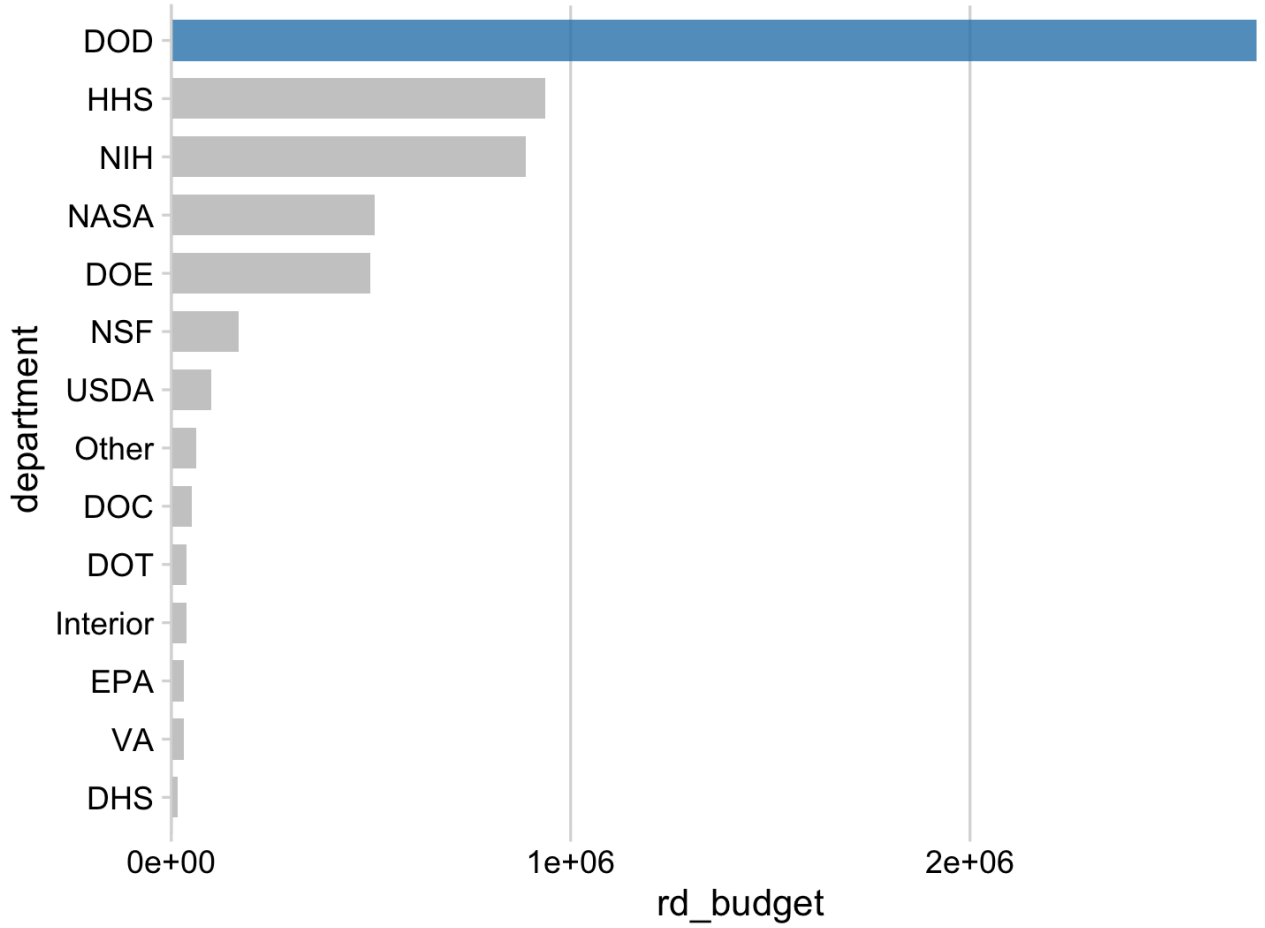


Better for **comparing total**



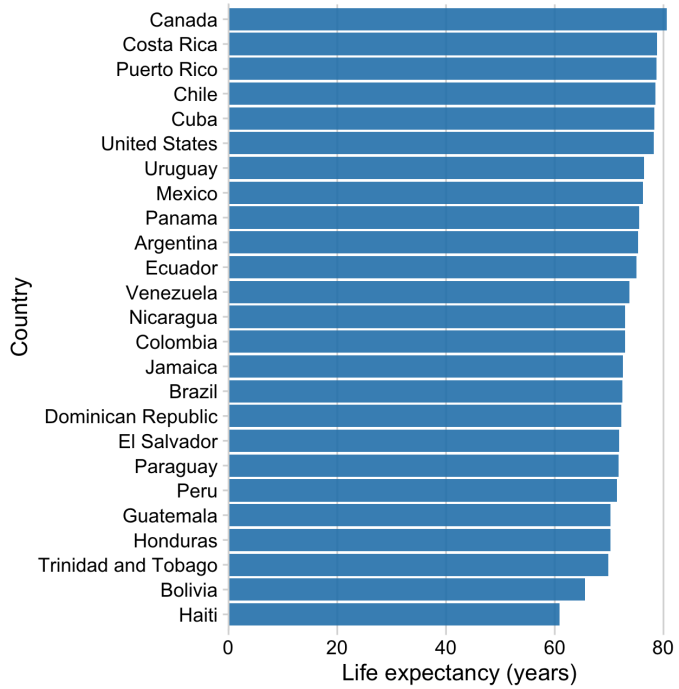
Better for **comparing parts**

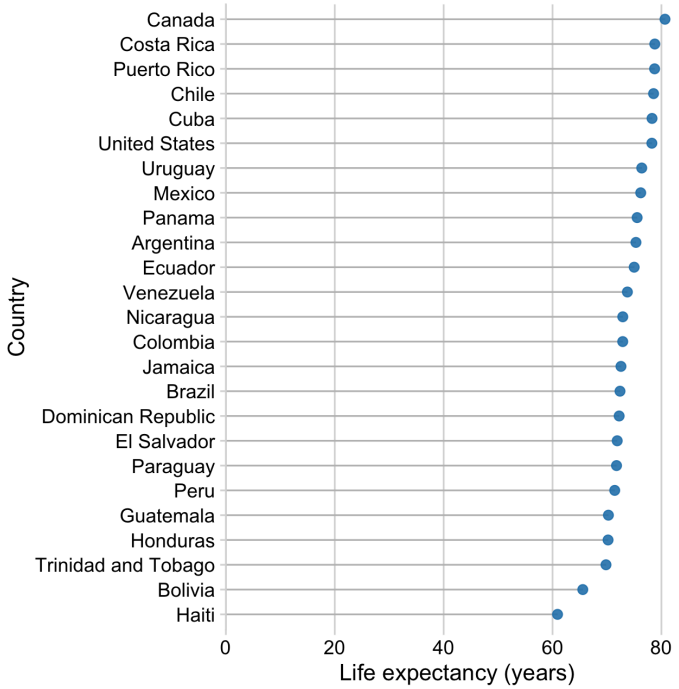
Color can be used to great advantage with bar plots



Lollipops

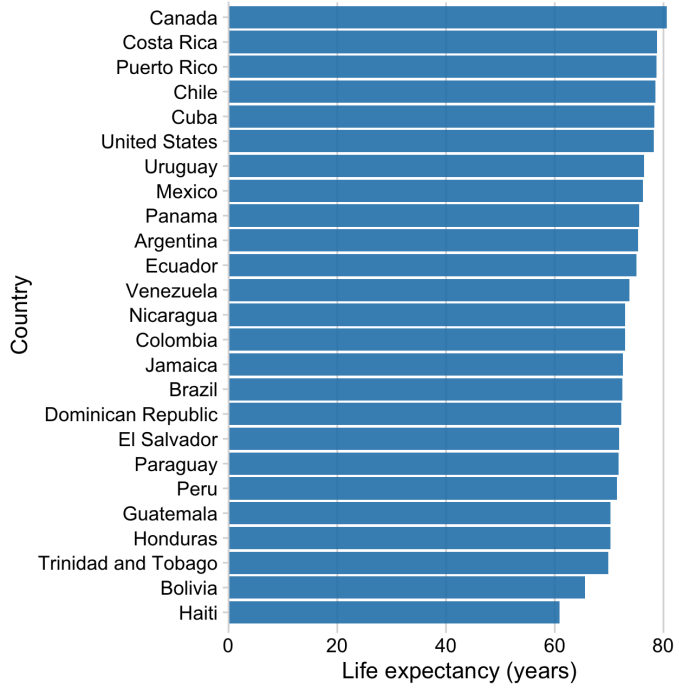
- Good when bars are overwhelming

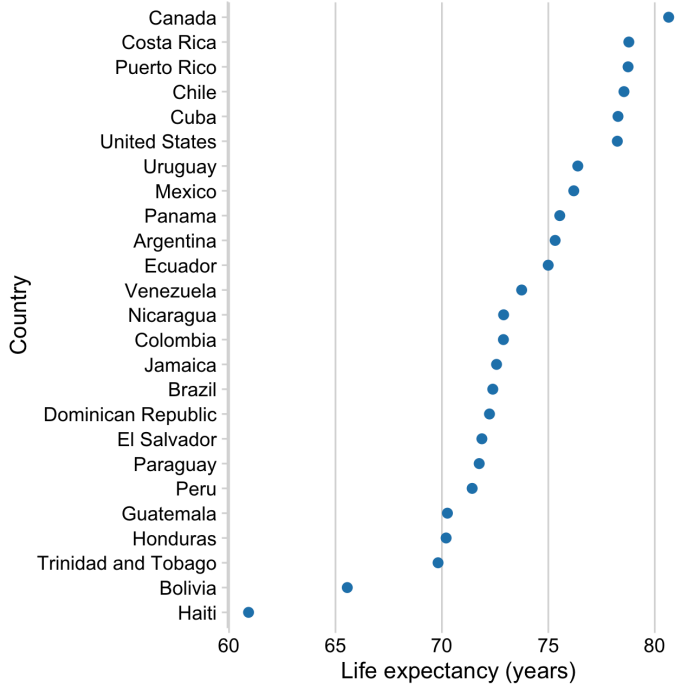




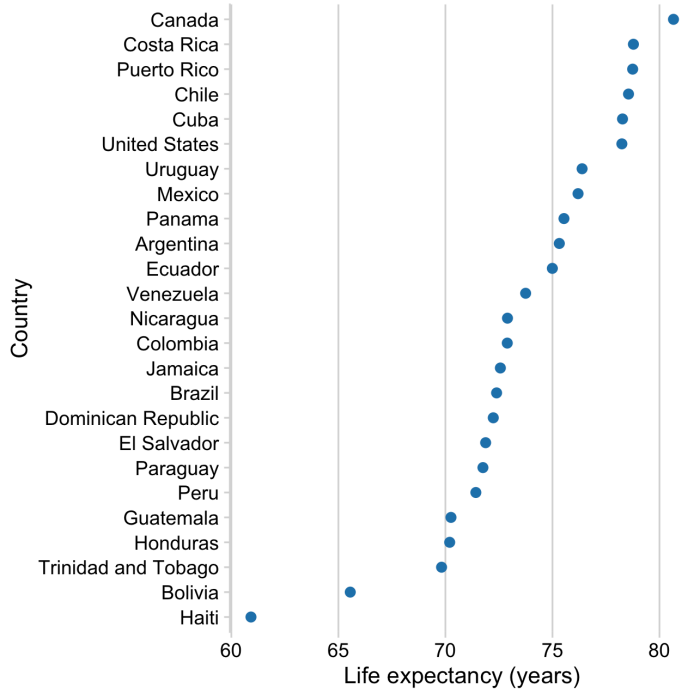
Dot Plots

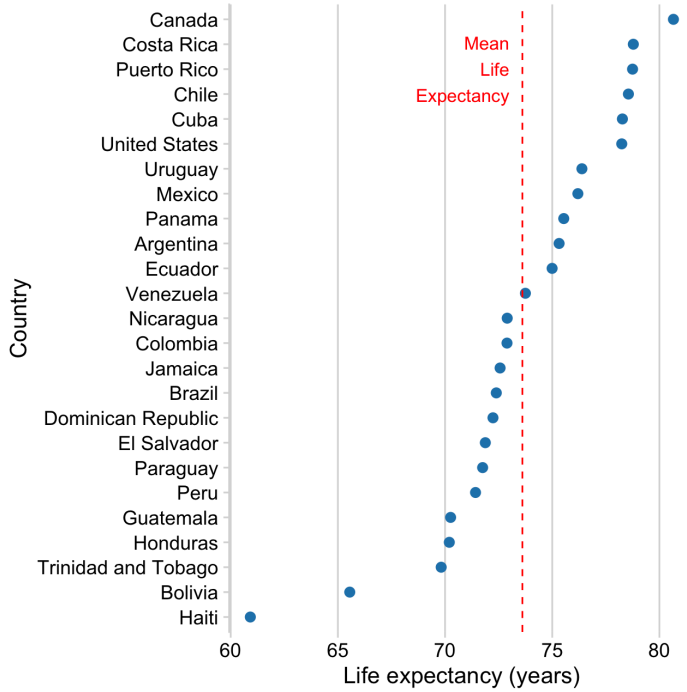
- When using dots, don't have to set min to 0



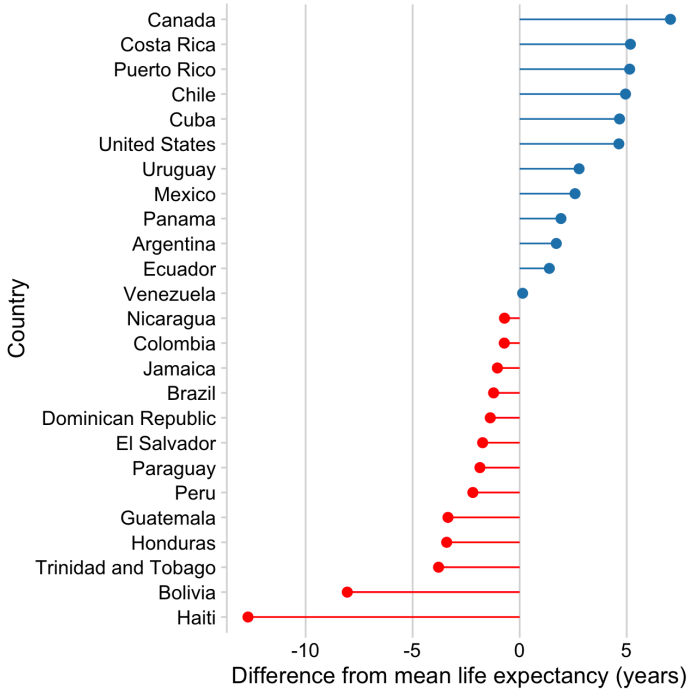


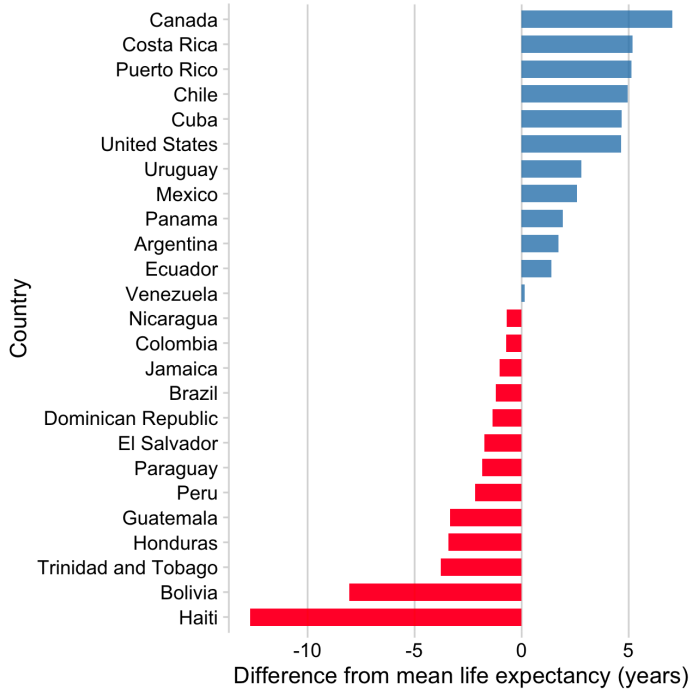
Reference Lines for Context





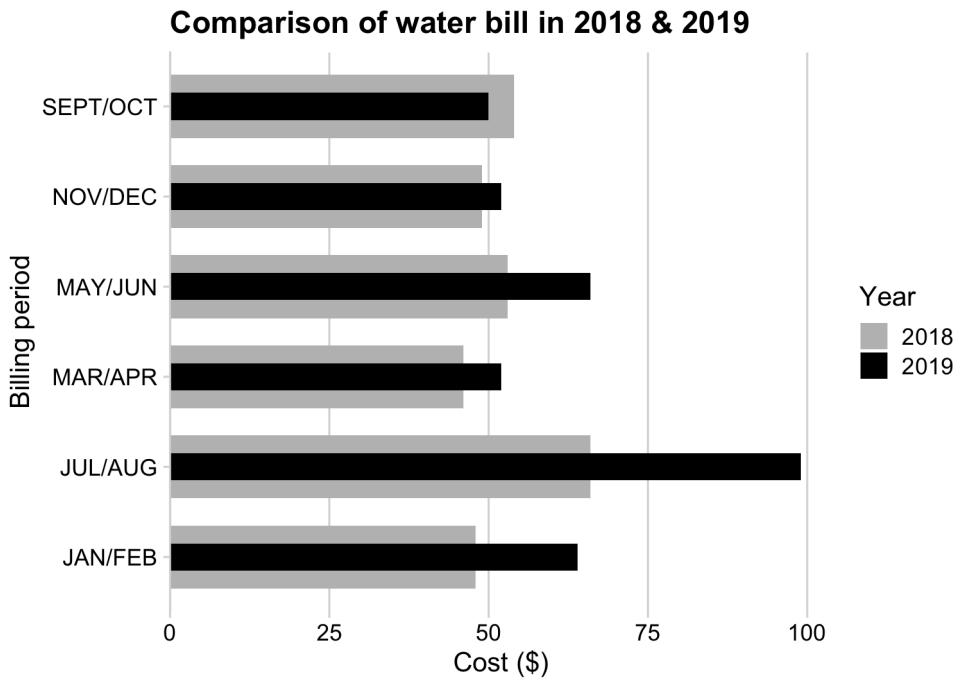
“reference line” can also be highlighted by plotting the difference relative to the line



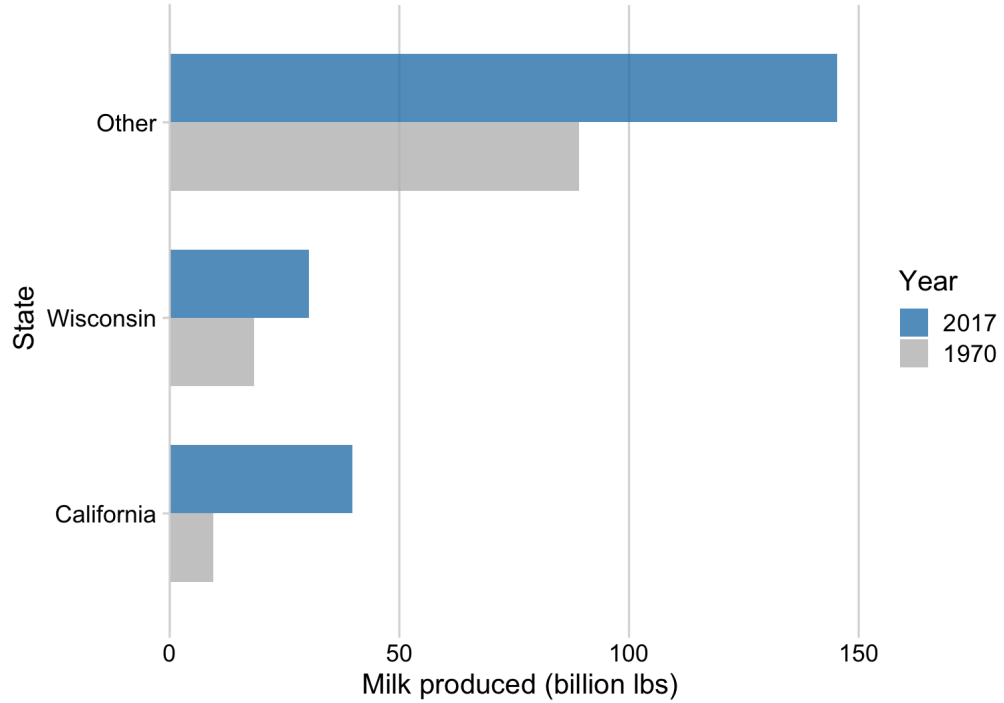


Overlapping bars

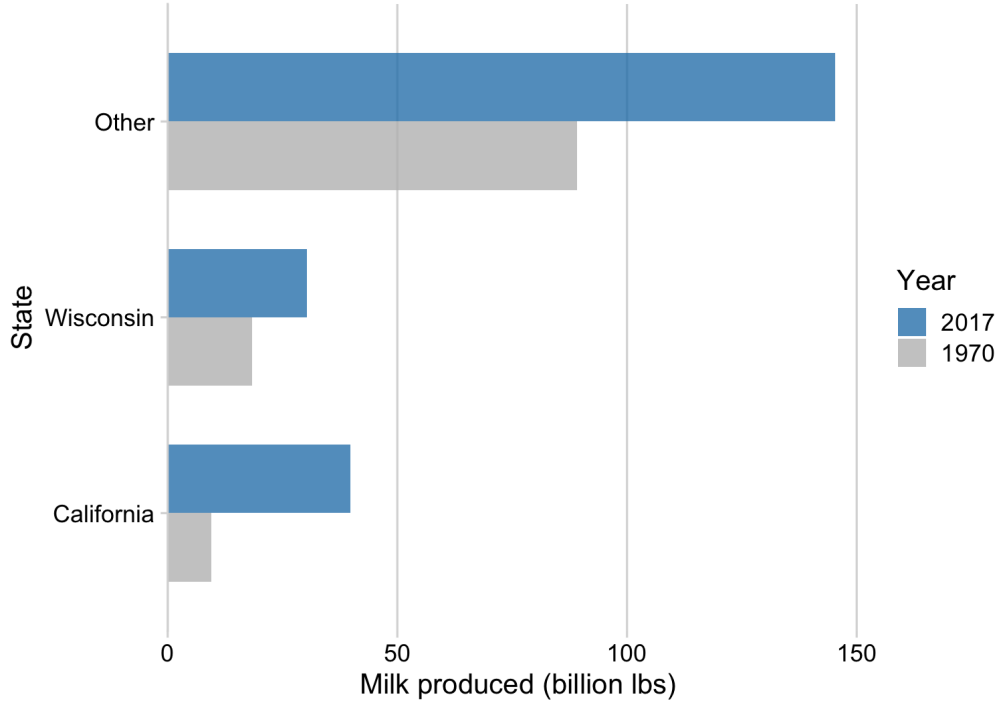
- Useful for comparing references

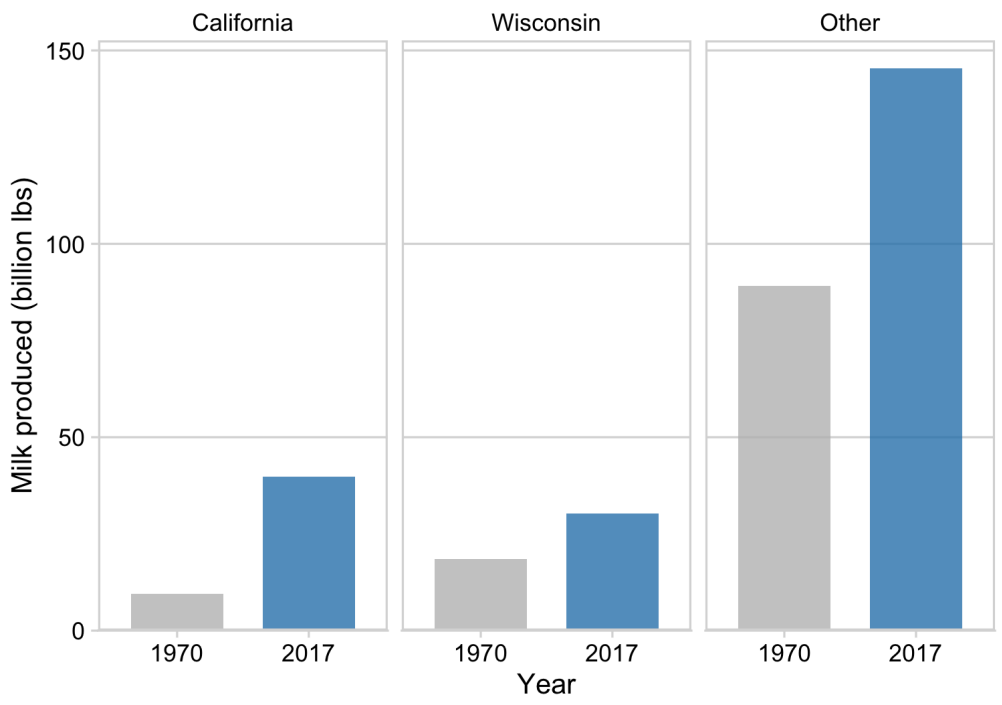


Dodged bars: useful for comparing two things

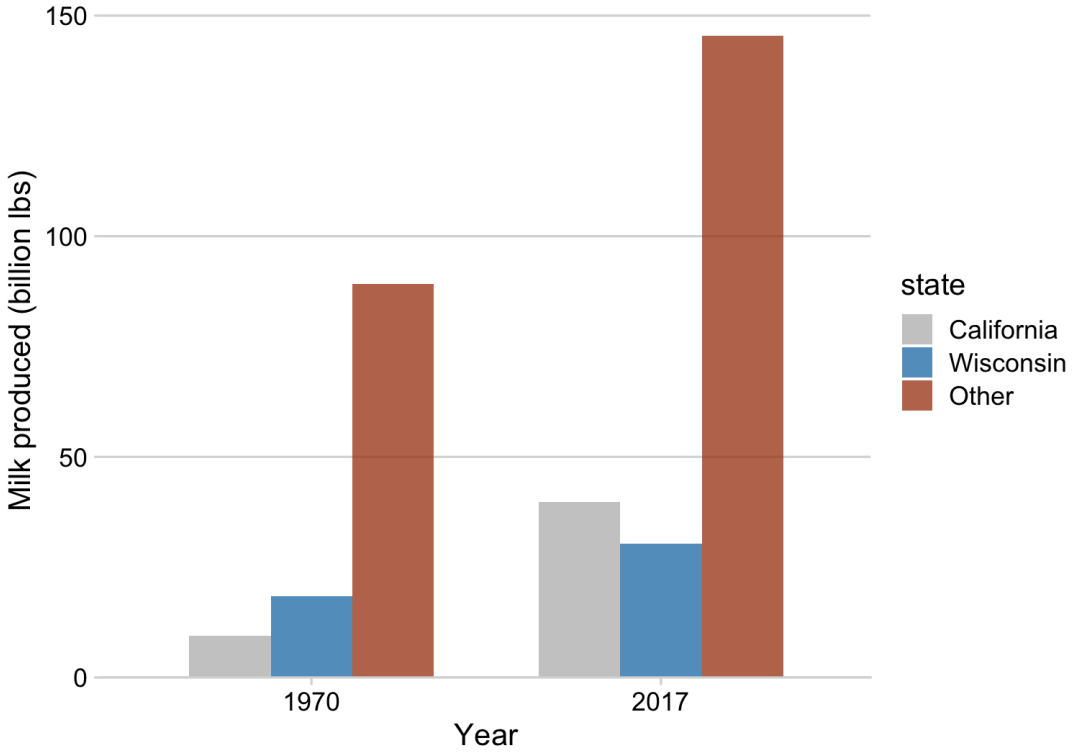


Dodged bars: facets can be helpful

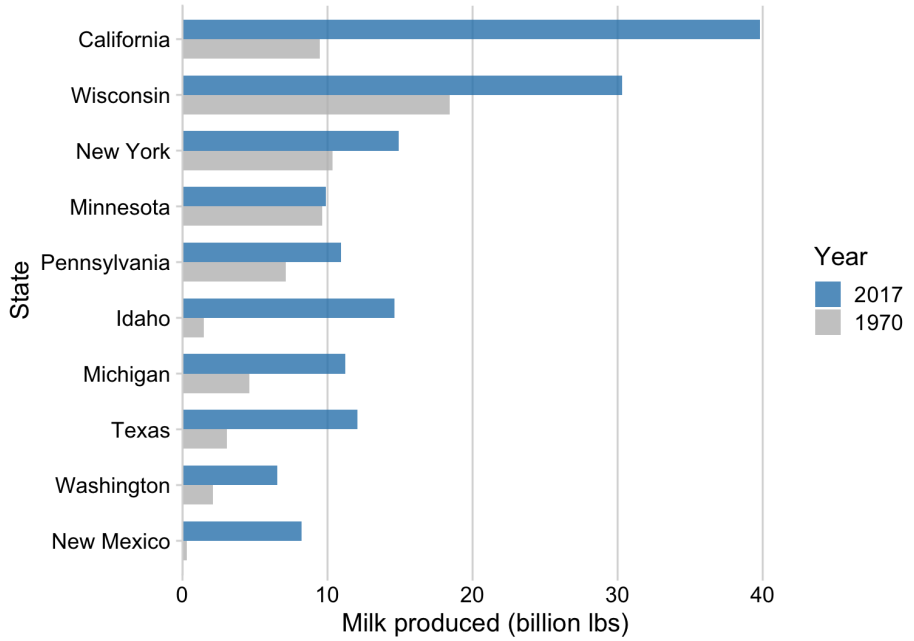




Dodged bars: more than two starts to get confusing



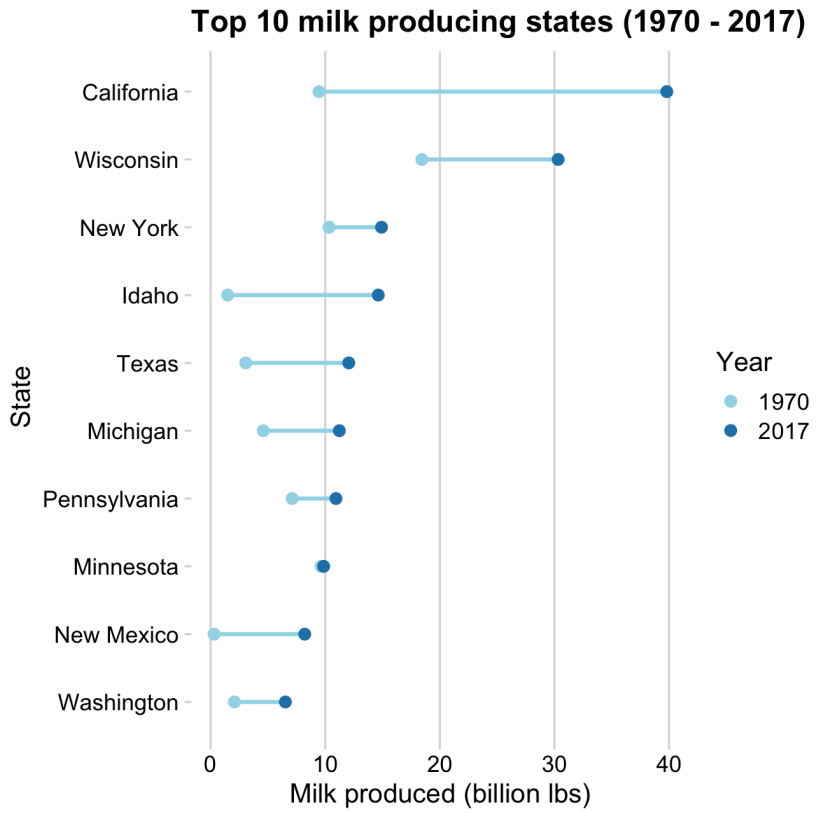
2 years, but more than 2 items per group



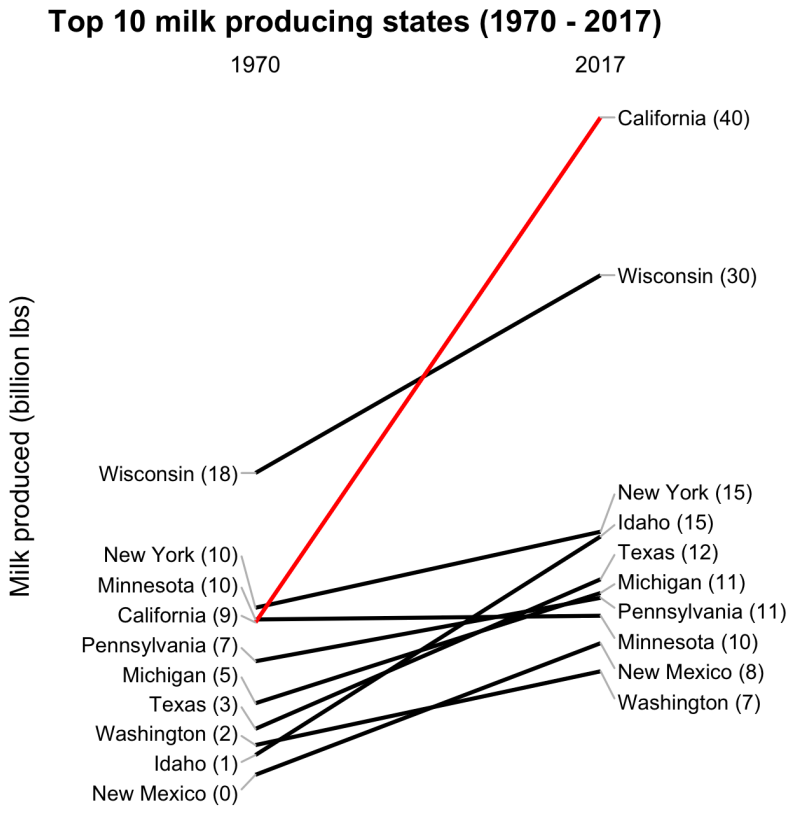
2 years, but more than 2 categories

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



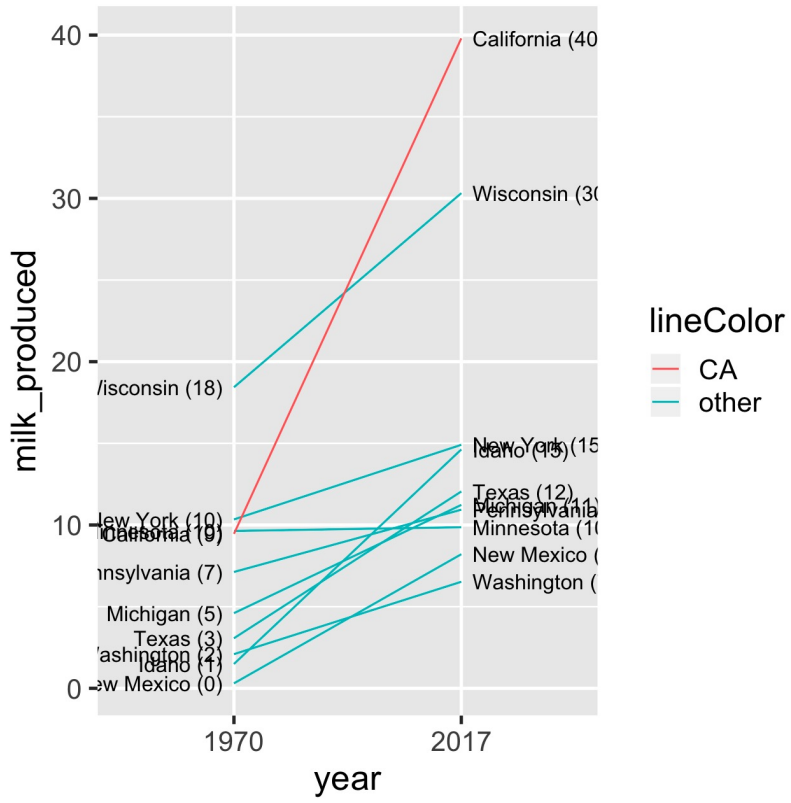
Dumbbell chart



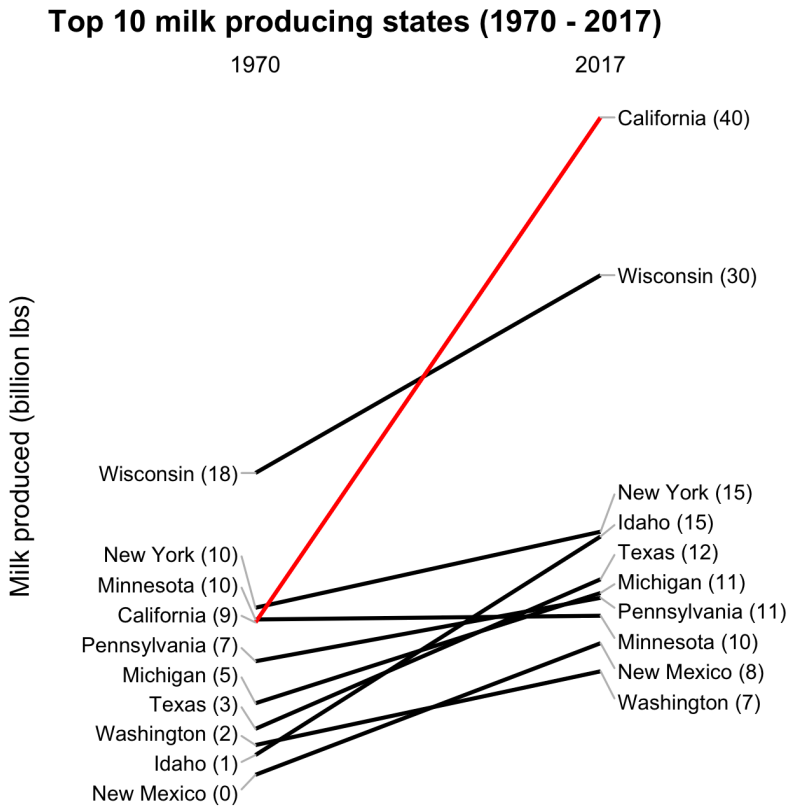
Slope chart

Comparisons across more than two categories

- Remember to label clearly

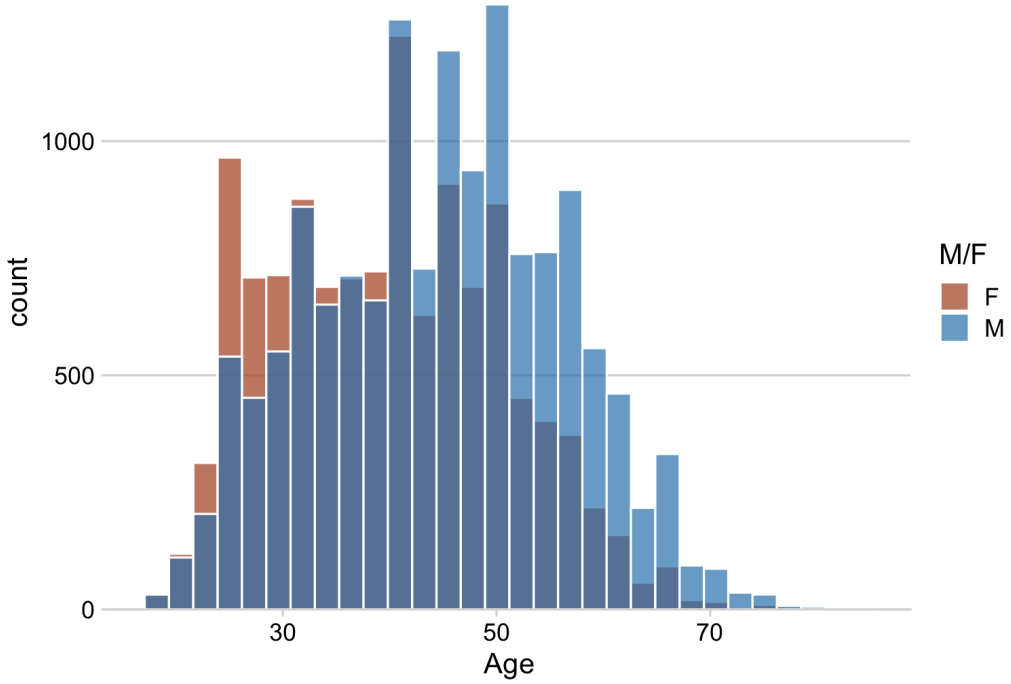


Dumbbell chart

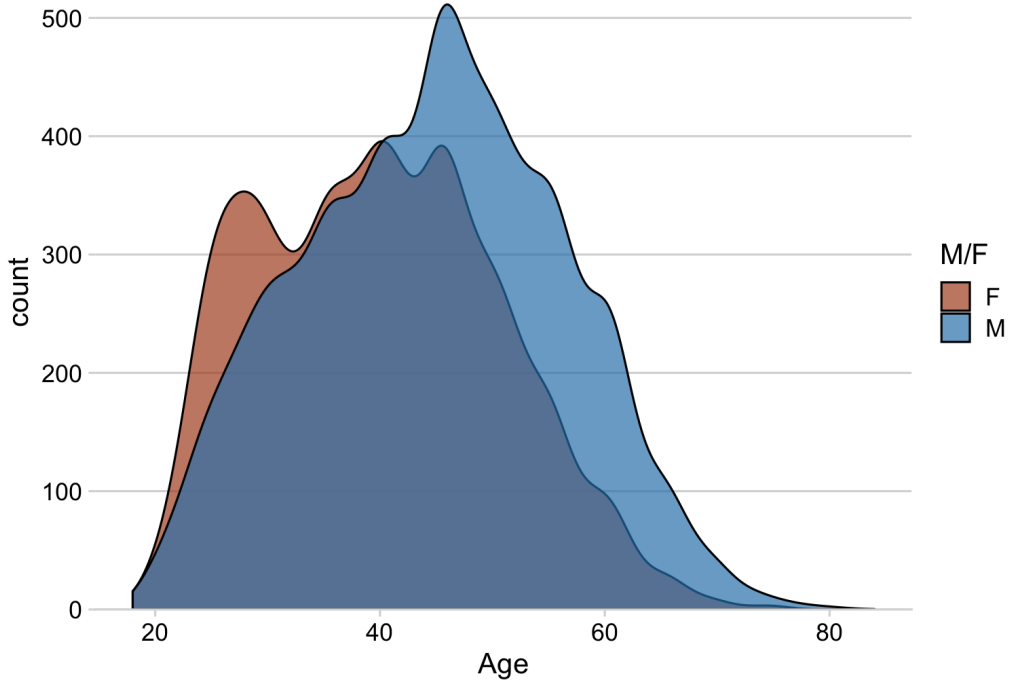


Slope chart

Continuous : now not bar but histogram



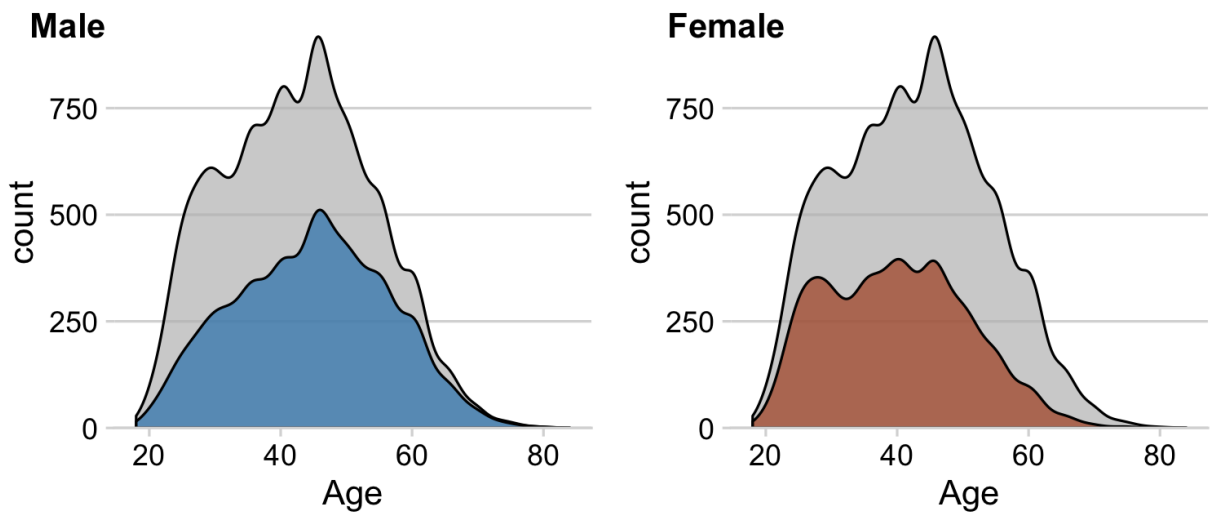
bad



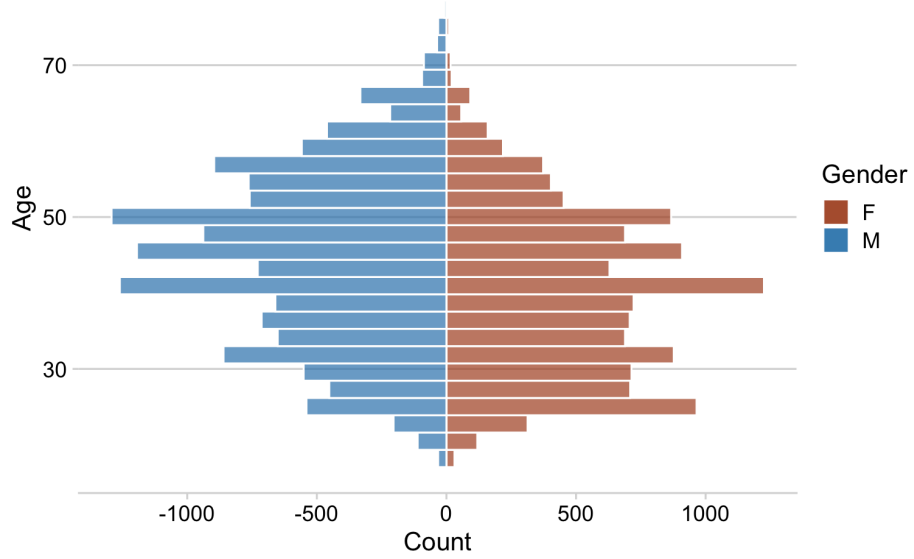
Somewhat better

Comparing histograms

- Overlapping or diverging can be okay for a small number of categories

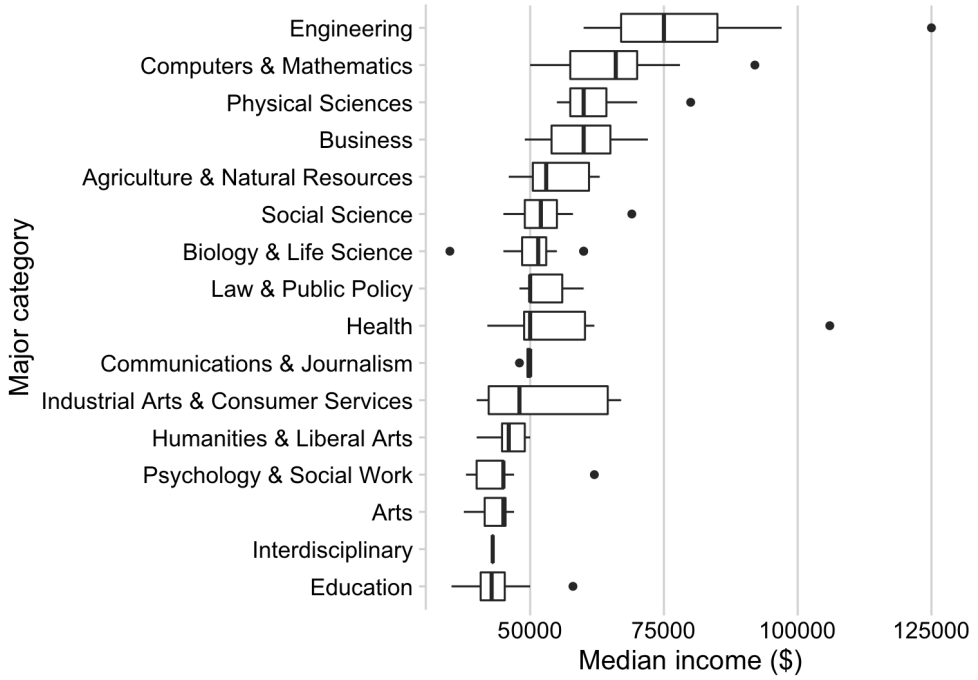


Density facets

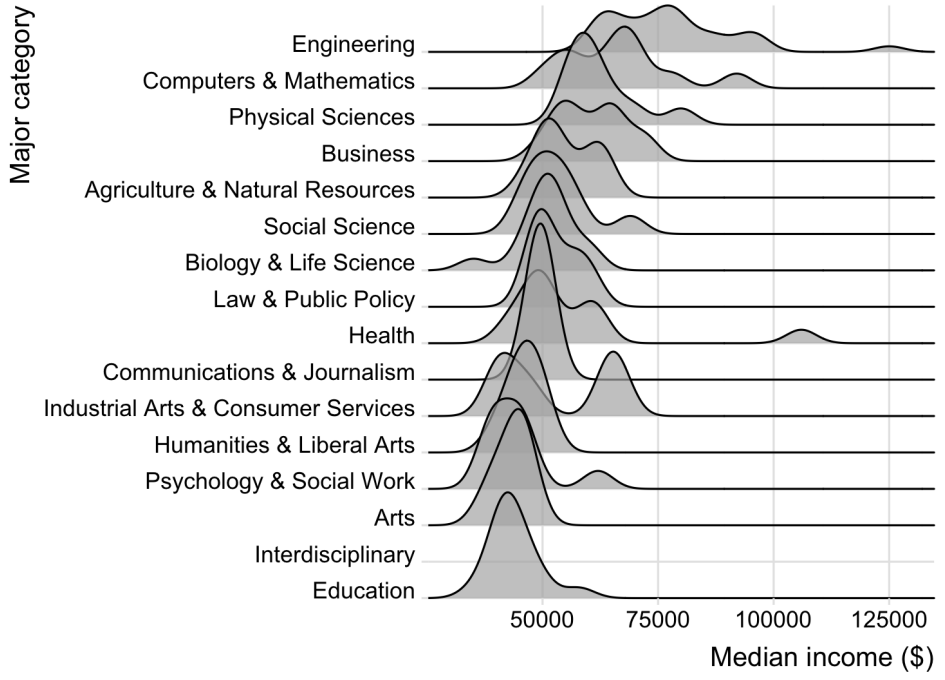


Diverging histograms

Box and ridge plots are good for large numbers of categories

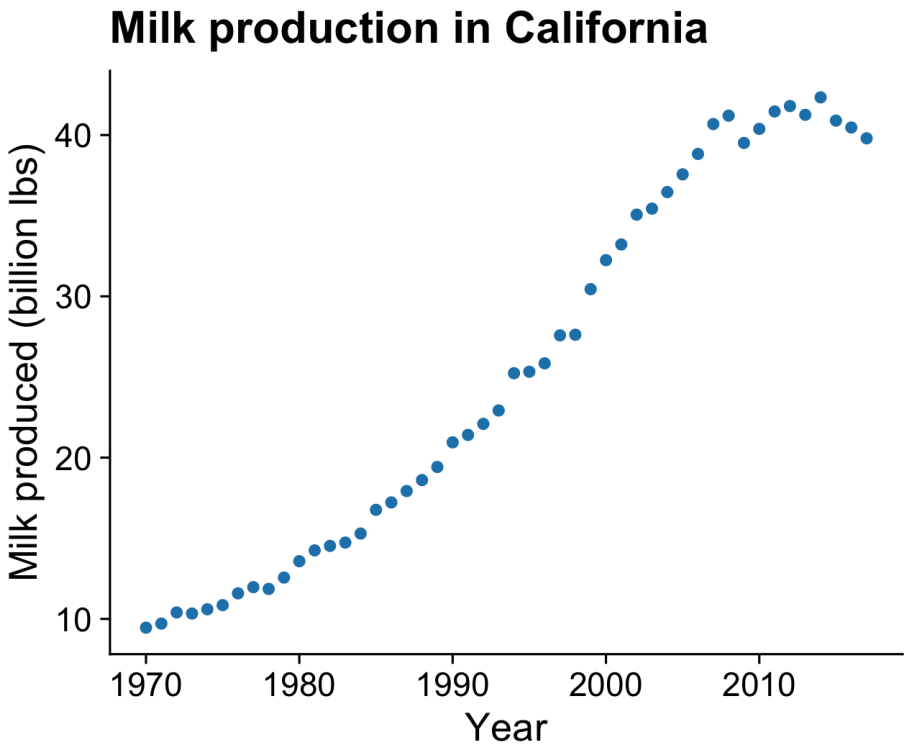


Box plot

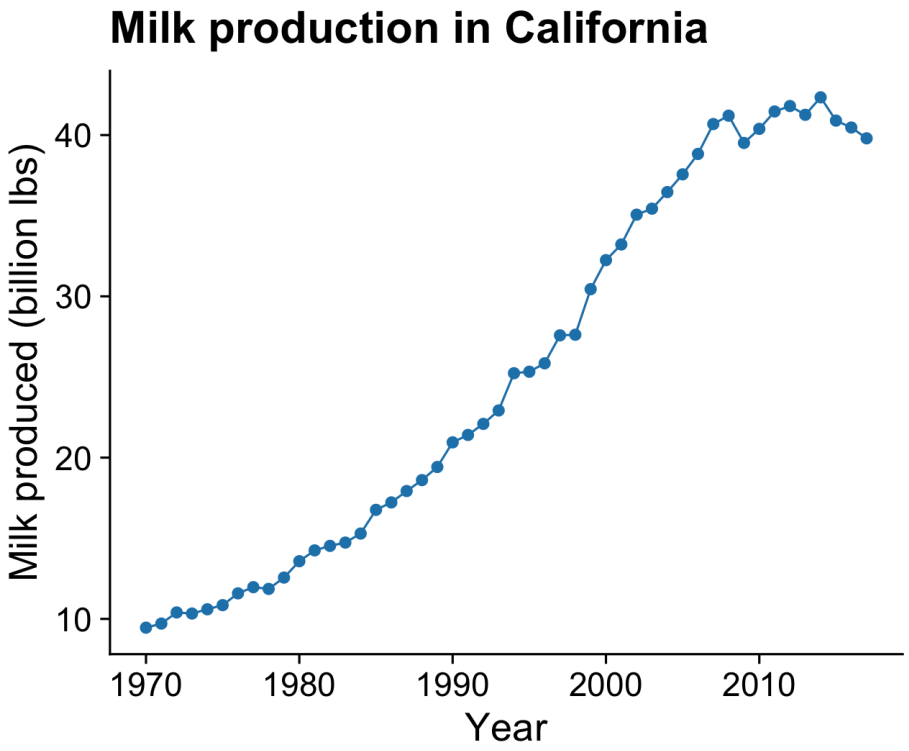


Ridge plot

Numerical vs Numerical
Trends

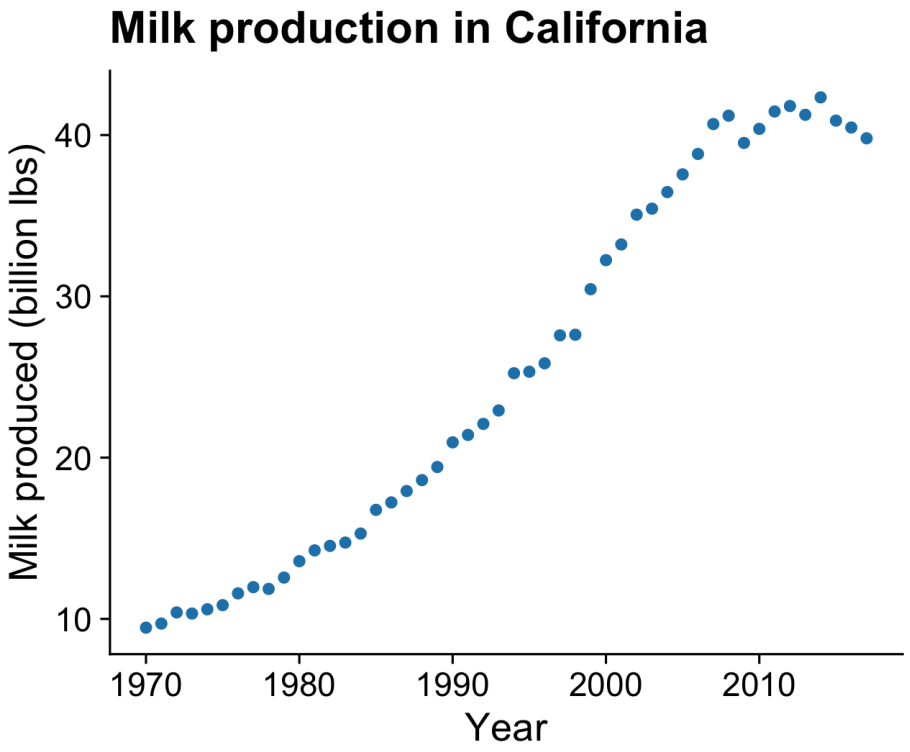


Points

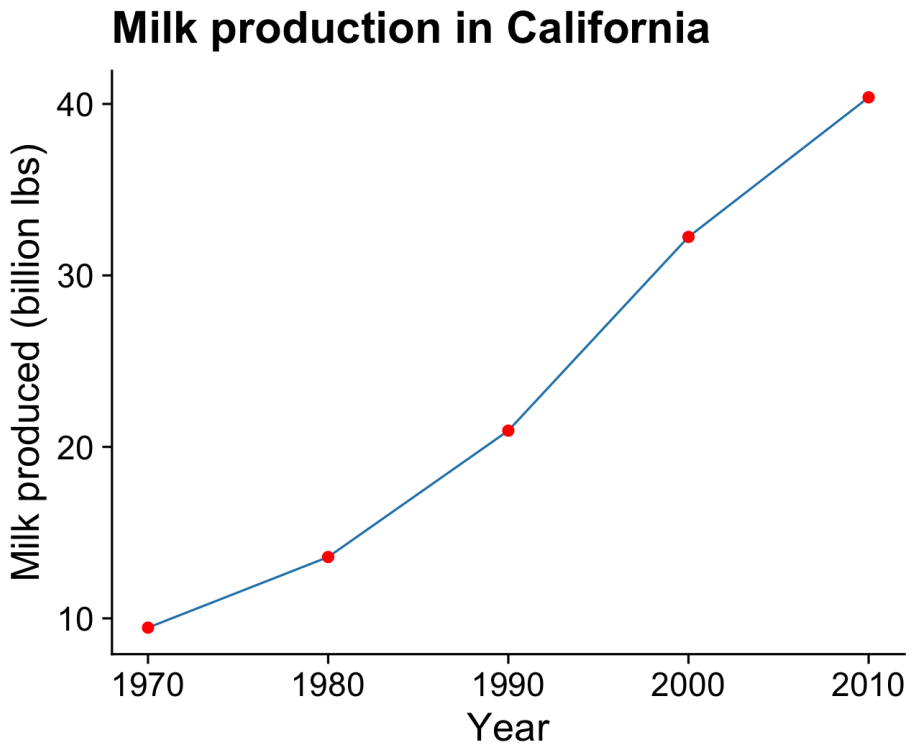


Points + Line: helps emphasize the overall trend

Trends

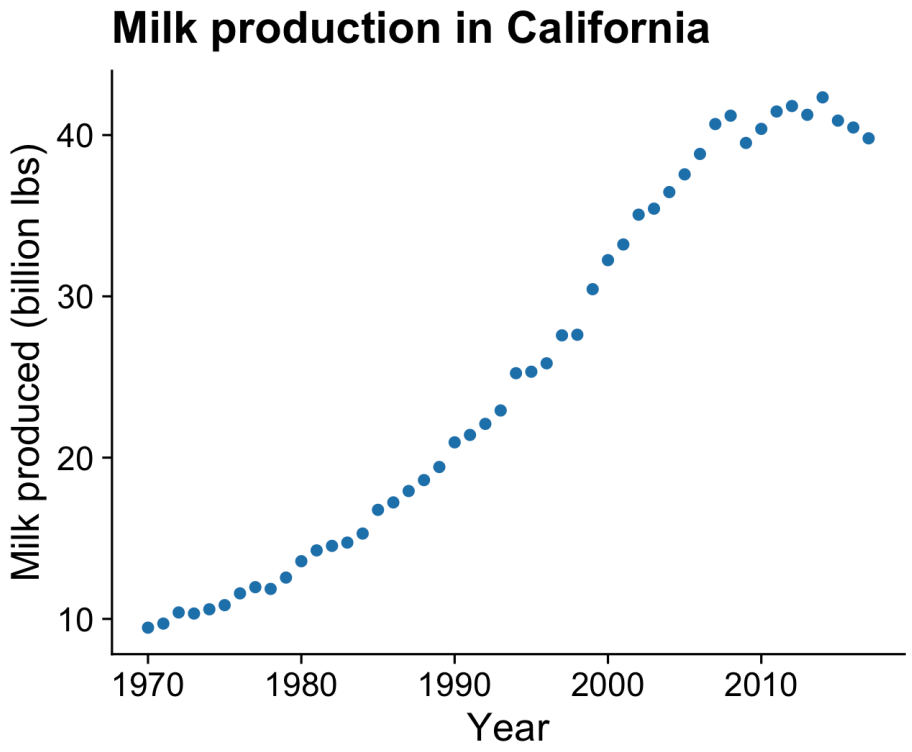


Points

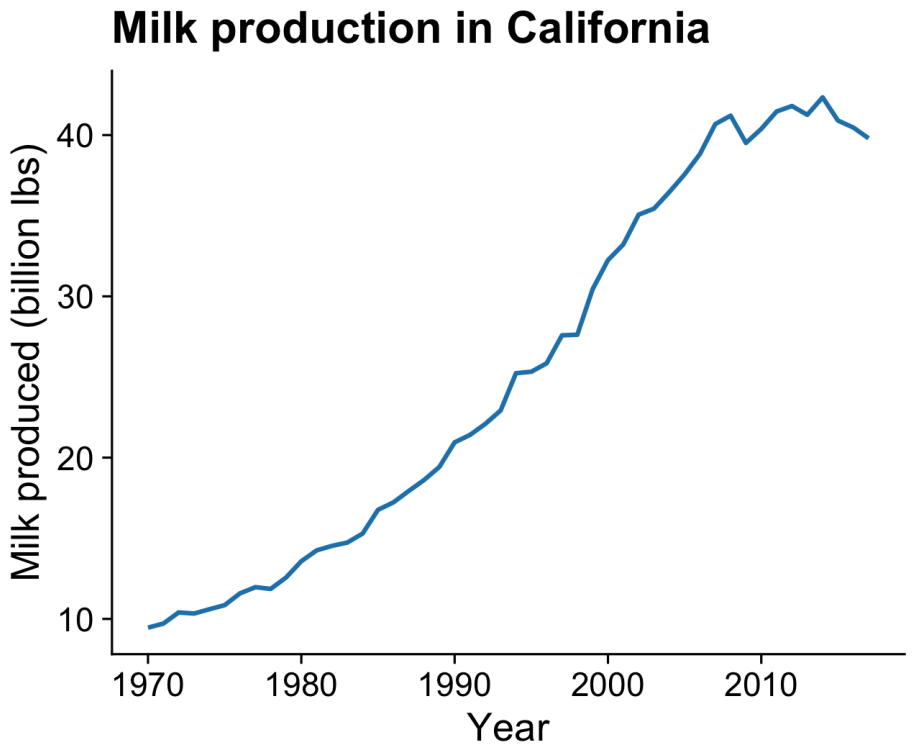


Points + Line: Note that for sparse data, a line can potentially be misleading

Trends

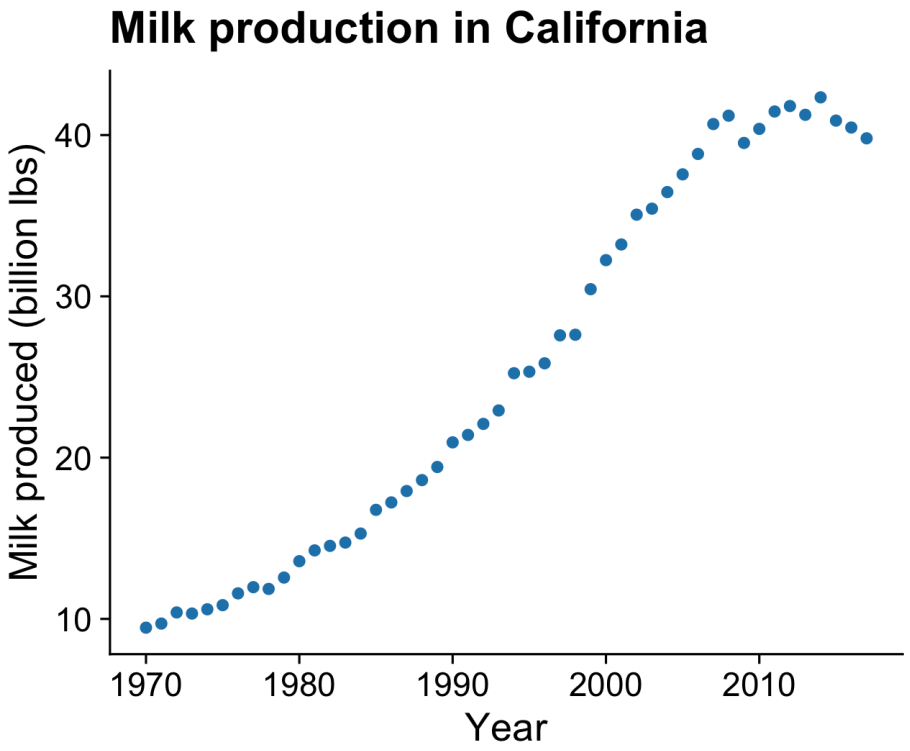


Points

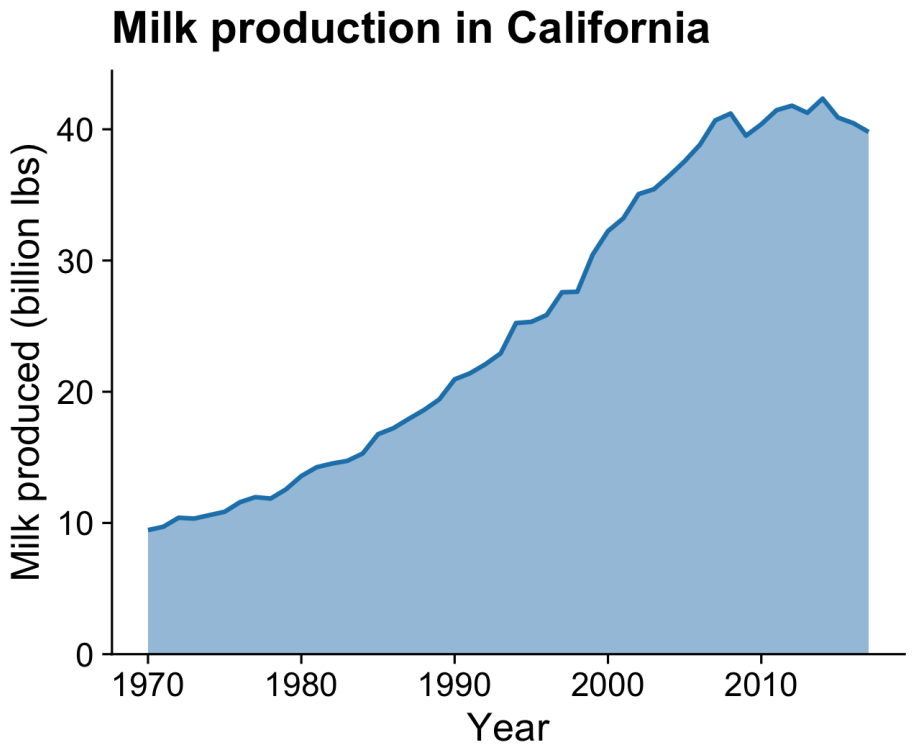


Line: omitting points emphasizes the overall trend

Trends



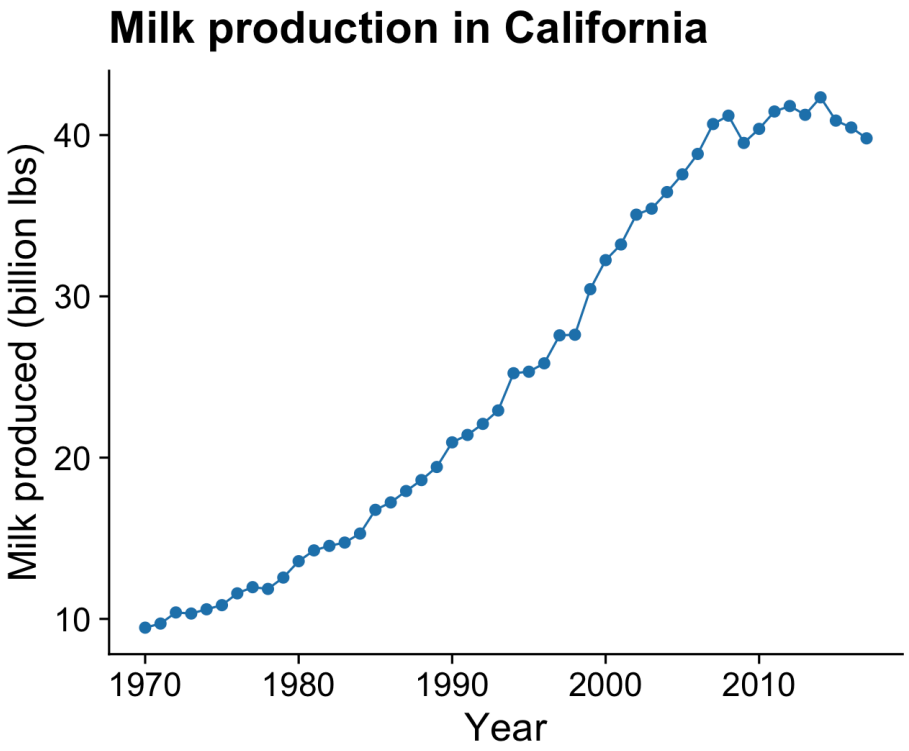
Points



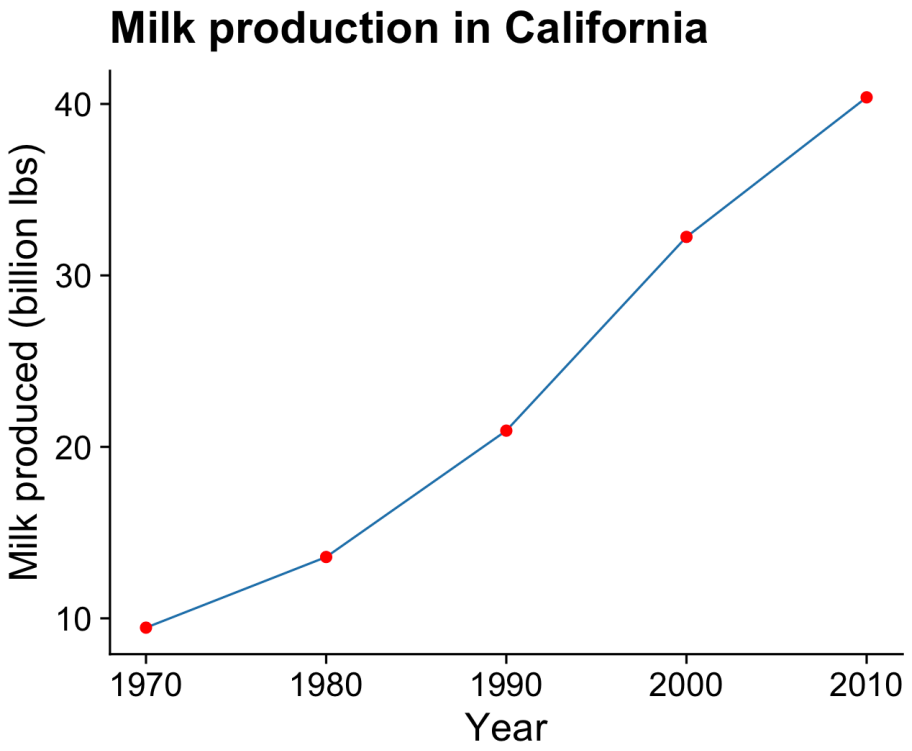
Line + Area: further emphasizes the overall trend

but y-axis must start at 0

Trends

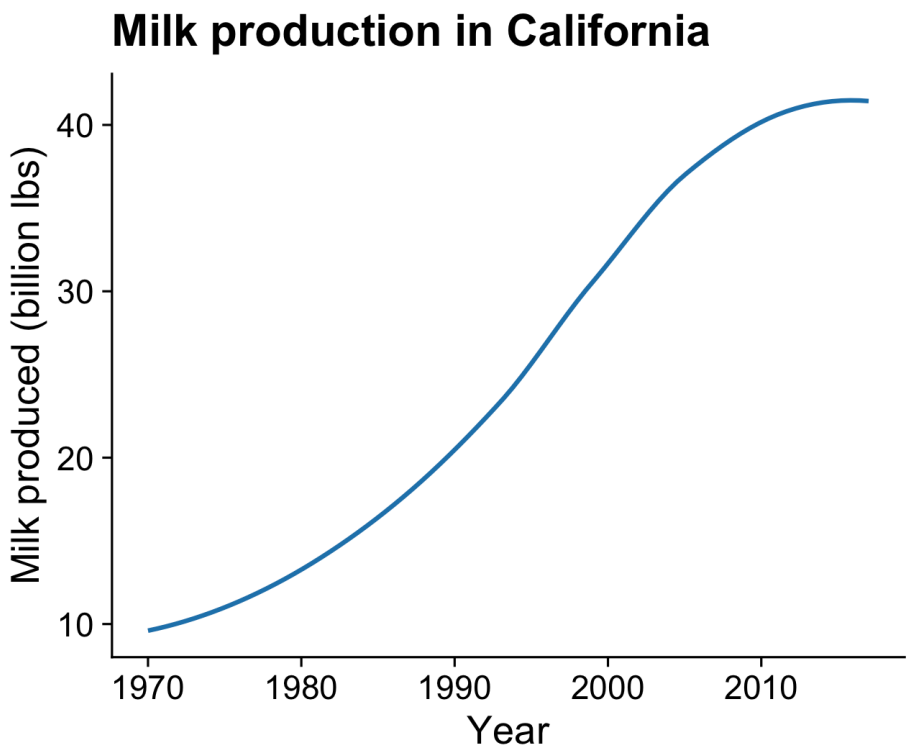


Points + Line: helps emphasize the overall trend

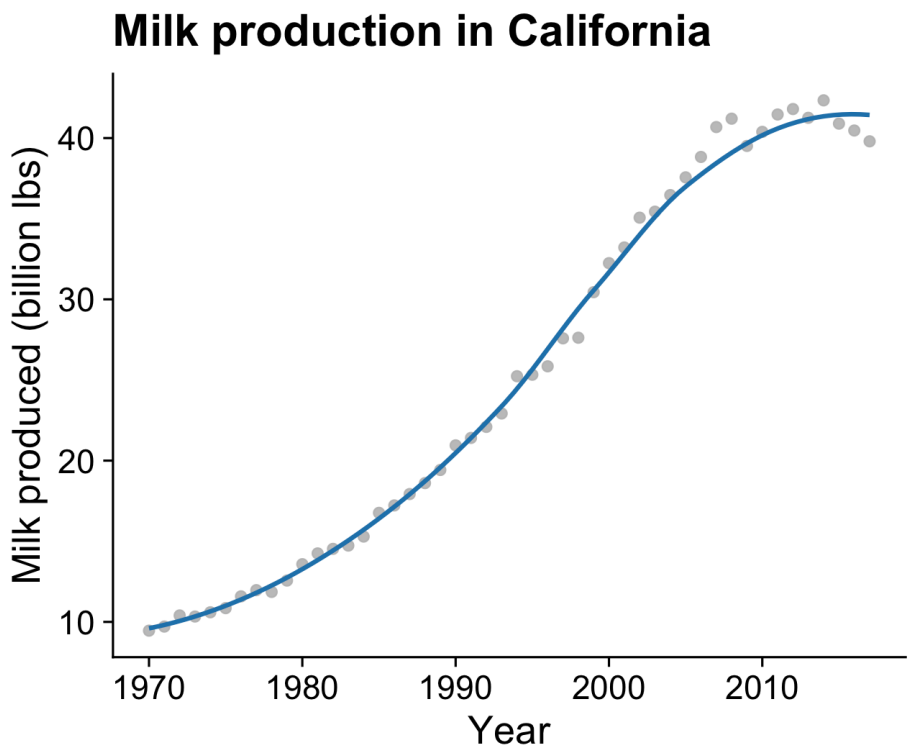


Points + Line: Note that for sparse data, a line can potentially be misleading

Trends



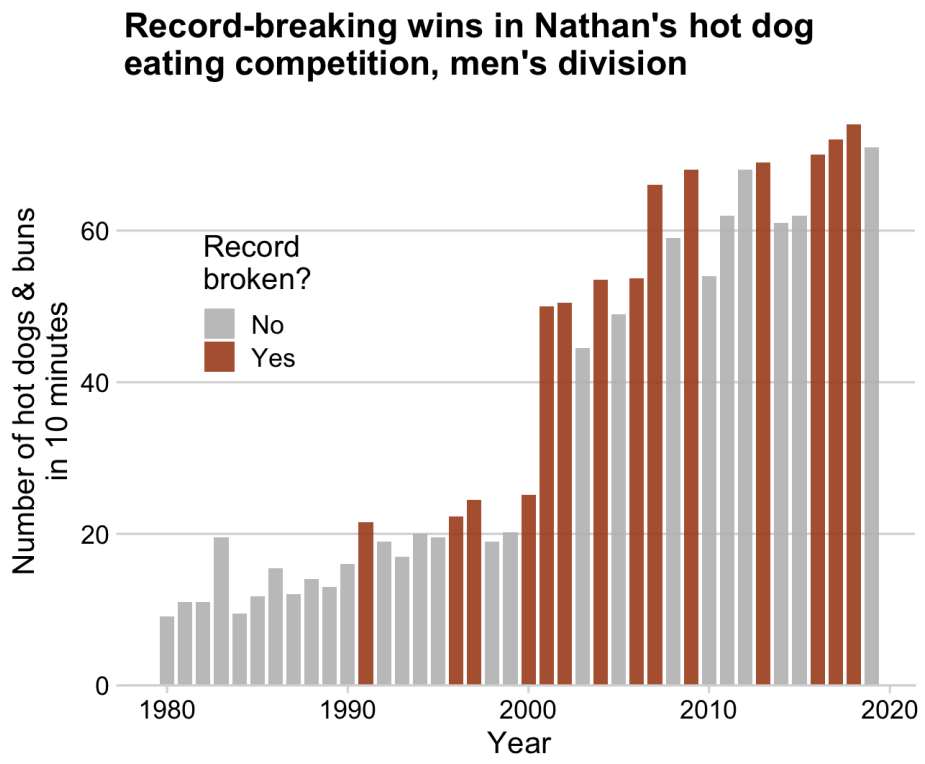
Smoothed line: shows modeled representation of the trend

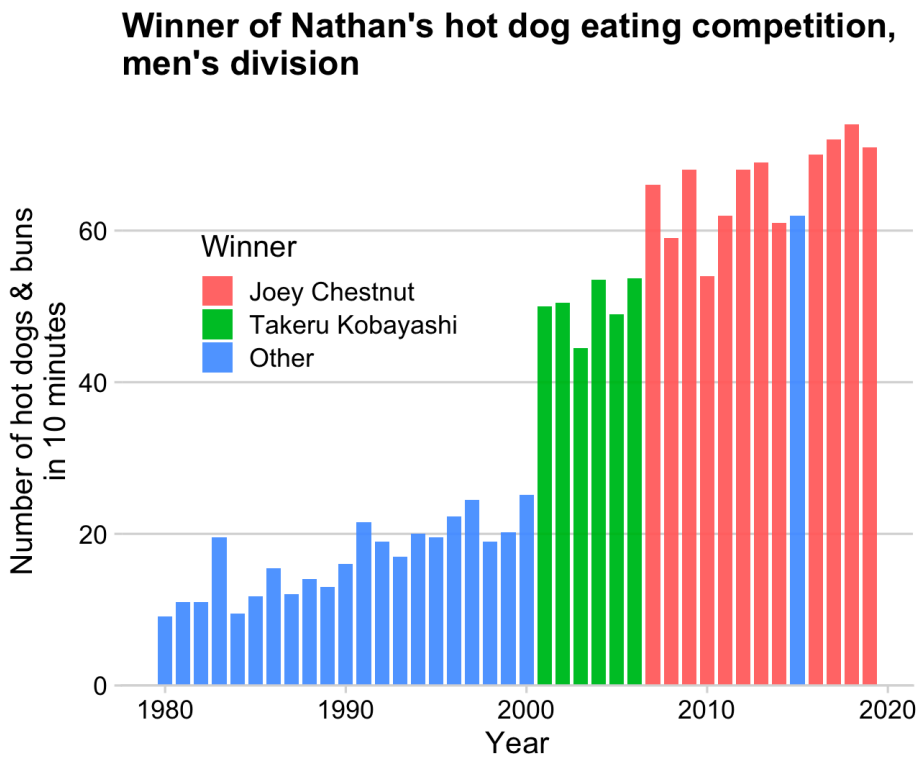


Smoothed line + points: helps show whether outliers are driving the trend

Trends

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





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

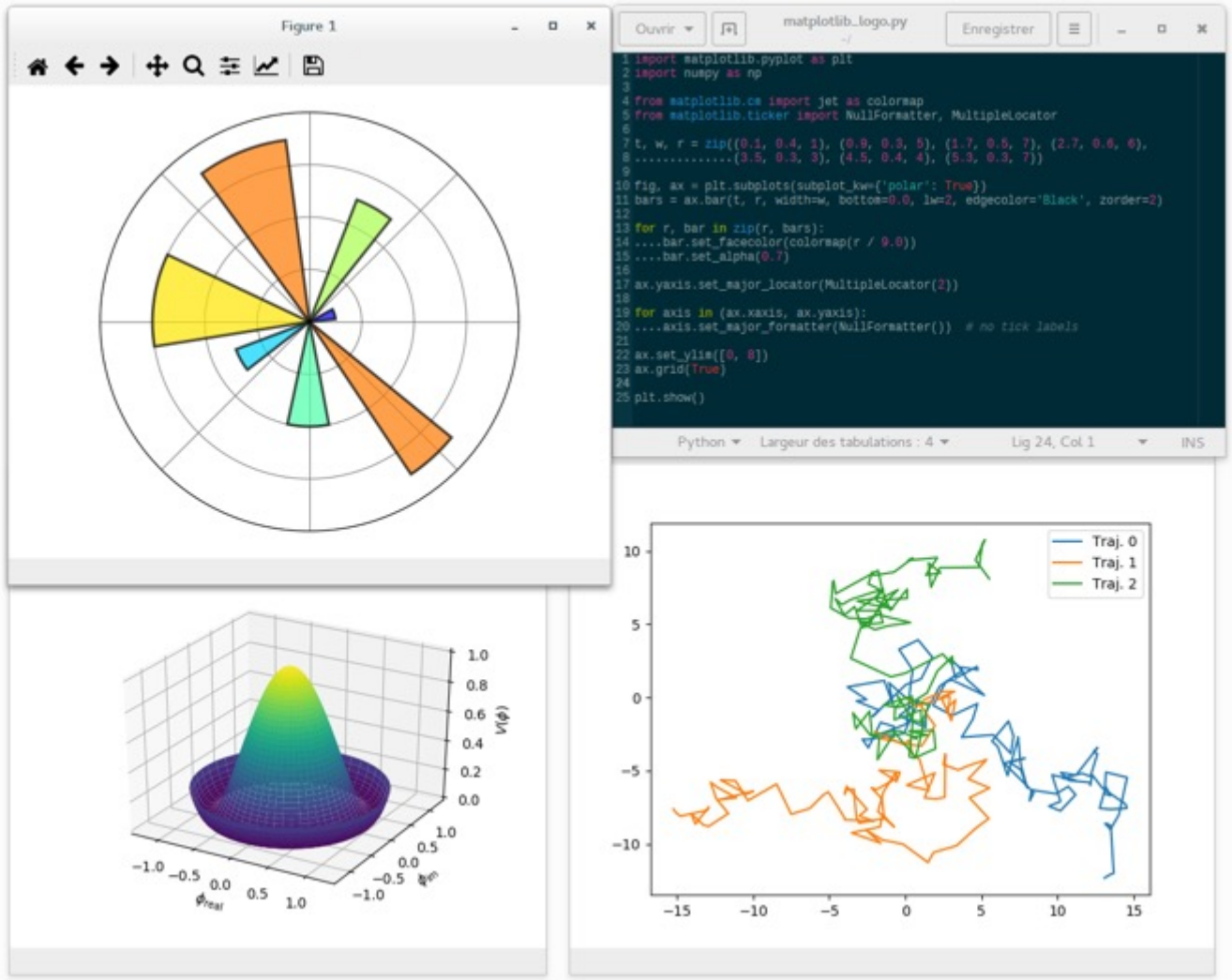
Smoothed line + points: helps show whether outliers are driving the trend

Python plotting

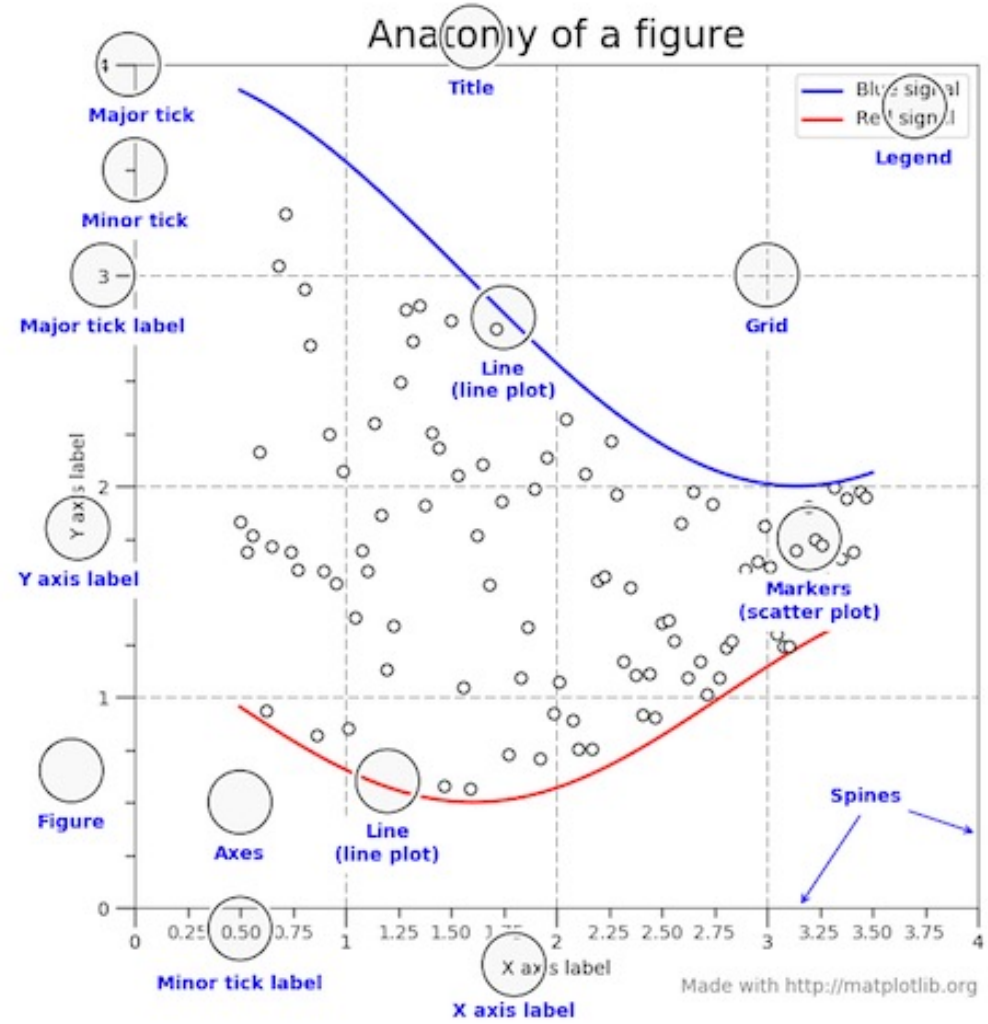
- For the remainder of the time, I want to get us up to speed with some Python plotting basics

Matplotlib

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



https://matplotlib.org/



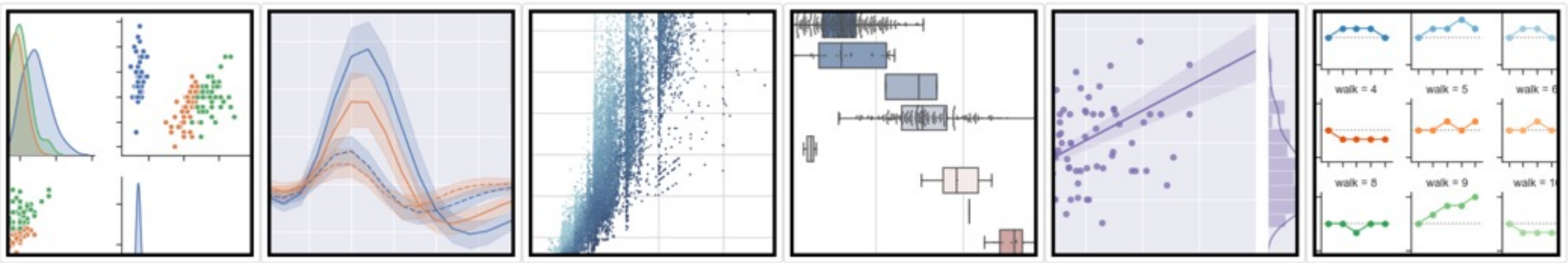
Matplotlib

- Built on the NumPy and SciPy frameworks and initially made to enable interactive Matlab-like plotting via gnuplot from iPython
- Gained early traction with support from Space Telescope Science Institute and JPL
- Easily one of the go-to libraries for academic publishing needs
  - Create publication-ready graphics in a range of formats
  - Powerful options to customize all aspects of a figure
- Matplotlib underlies the plotting capabilities of Pandas and Seaborn

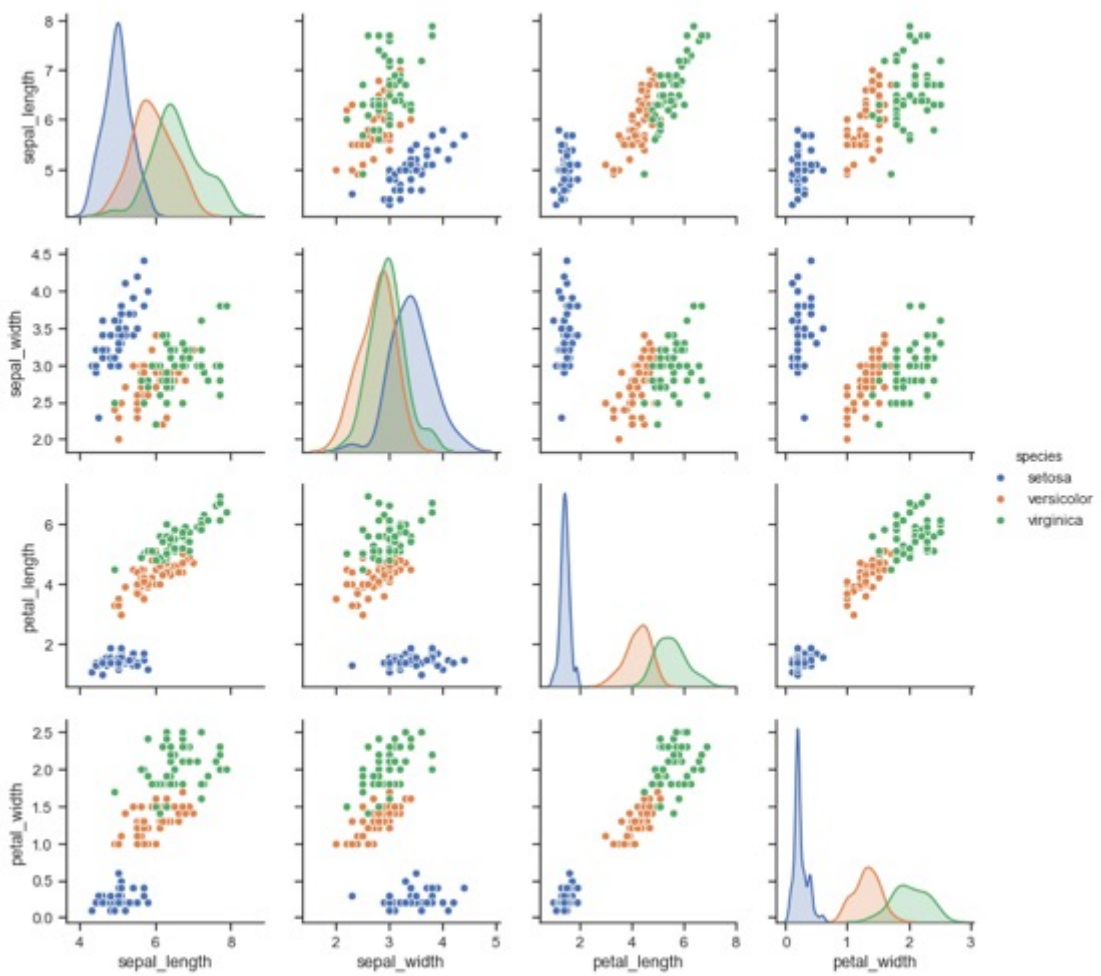
https://matplotlib.org/

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/

- Matplotlib is organized in a hierarchy
- At the top: matplotlib.pyplot
  - This is a module that provides high-level functions to add elements to the current axes in the current figure
- Lower levels can be accessed by figure and axes objects
  - “Figure”: an object that keeps track of child “Axes” objects (and other things like titles, legends, and the canvas)
  - “Axes”: an object that can be thought of as the plot
  - “Axis” is a different object than “Axes”