Stats 170A/B, Data Visualization

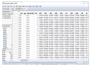
Chen Li[1] and Vladimir Minin[2]

[1]Department of Computer Science
[2]Department of Statistics
Bren School of Information and Computer Sciences
University of California, Irvine

February 3, 2020

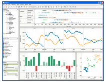
Data science: from data to actions

Raw Data

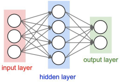


Data Wrangling

Data Management



Exploratory Data Analysis



Predictive Modeling

Consumers

External Business Customers

Internal Business Customers

Government

Scientists

Databases, Algorithms, Software Engineering

Machine Learning, Statistics

Domain knowledge Business knowledge

Why visualize and explore?

- People are good at pattern recognition
  - At spotting clusters, trends, outliers, structure, etc. that computers many miss

- Usually two types of users
  1. The data scientist who wants to explore/analyze/understand
     - For the data scientist, visualization and exploration are part of an iterative process
  2. The person who needs a quick summary to make a decision
     - For the consumer we want to communicate information quickly and clearly
     - e.g., for a medical doctor, for a policy-maker, for a company executive

- For data scientists...its always a good idea to look at your data
  - Helps to understand where the semantics of the data...what the measurements actually mean

What is exploratory data analysis?

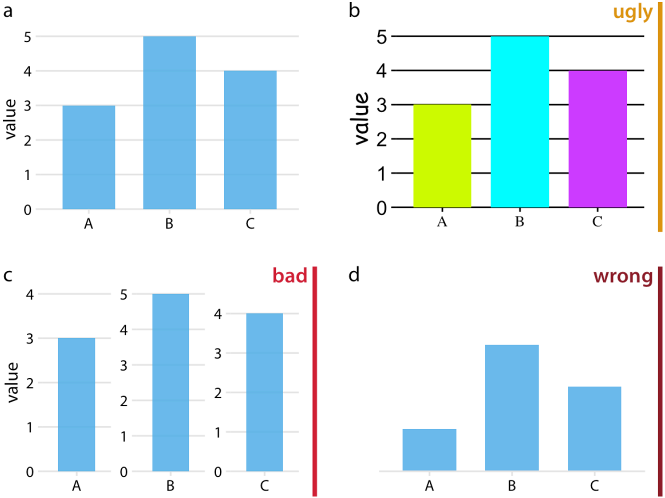
- EDA is broader than just visualization
- EDA = {visualization, clustering, dimension reduction,...}
- For small numbers of variables, EDA = visualization
- For large numbers of variables, we need to be cleverer
  - Clustering, dimension reduction, embedding algorithms
  - These are techniques that essentially reduce high-dimensional data to something we can look at
- Pioneered by John Tukey (statistician at Bell Labs, Princeton) in the 1960's
  - "let the data speak"

Plan for today

Fundamentals of Data Visualization

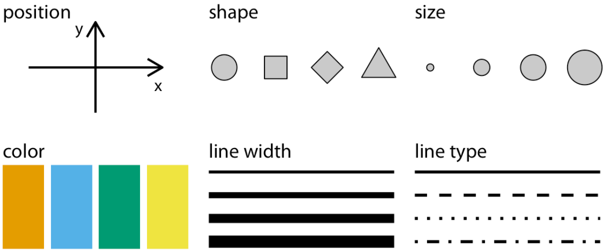
Claus O. Wilke

`https://serialmentor.com/dataviz/`

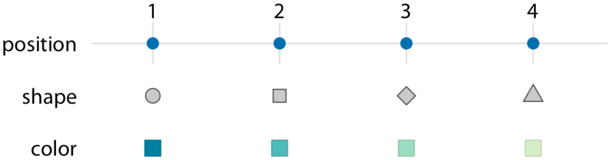


Mapping data onto aesthetics

Types of aesthetics:



Scales map data values onto aesthetics:



Mapping data onto aesthetics — example

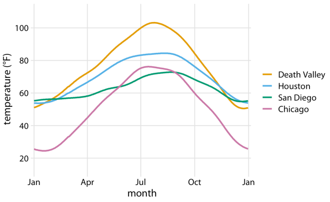
Table 2.2: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.

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

Mapping data onto aesthetics — example

Table 2.2: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.

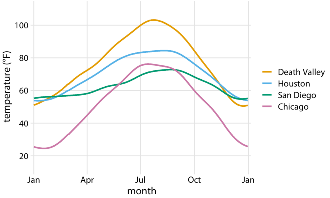
Month	Day	Location	Station ID	Temperature
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319	51.2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

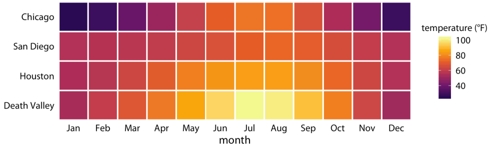


Mapping data onto aesthetics — example

Table 2.2: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.

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

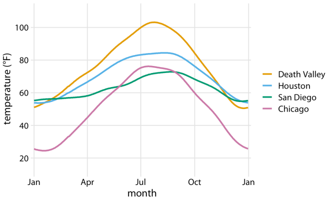


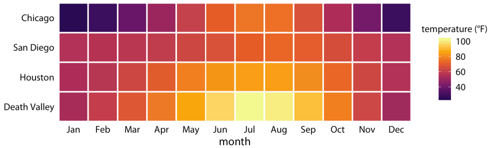


Mapping data onto aesthetics — example

Table 2.2: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.

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





Both plots use three scales in total: two position scales and one color scale

Color as a tool to distinguish

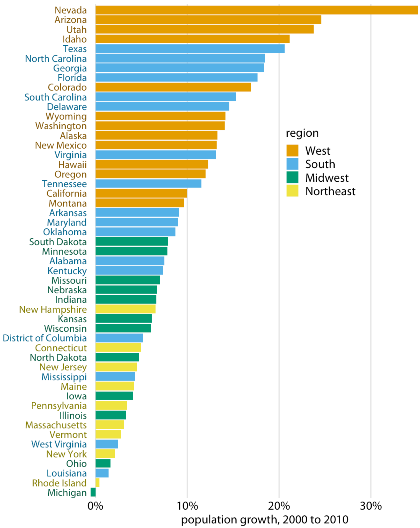


Figure 4.2: Population growth in the U.S. from 2000 to 2010. States in the West and South have seen the largest increases, whereas states in the Midwest and Northeast have seen much smaller increases or even, in the case of Michigan, a decrease. Data source: U.S. Census Bureau

Grab color scales at `http://colorbrewer2.org`

Color as a tool to highlight

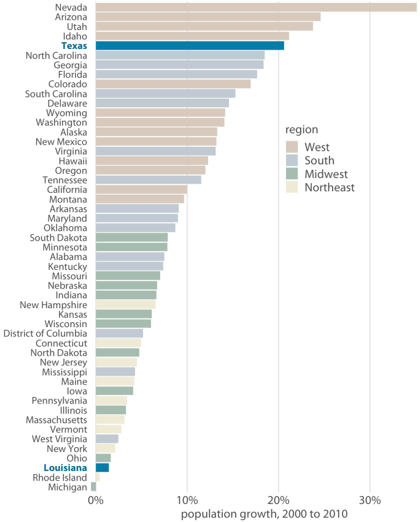


Figure 4.8: From 2000 to 2010, the two neighboring southern states Texas and Louisiana have experienced among the highest and lowest population growth across the U.S. Data source: U.S. Census Bureau

Grab color scales at `http://colorbrewer2.org`

Color to represent data values

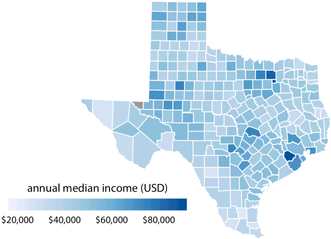


Figure 4.4: Median annual income in Texas counties. The highest median incomes are seen in major Texas metropolitan areas, in particular near Houston and Dallas. No median income estimate is available for Loving County in West Texas and therefore that county is shown in gray. Data source: 2015 Five-Year American Community Survey

Sequential color scale

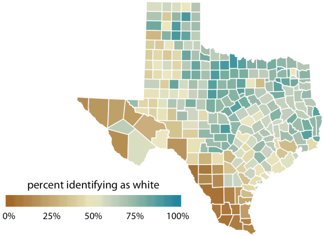


Figure 4.6: Percentage of people identifying as white in Texas counties. Whites are in the majority in North and East Texas but not in South or West Texas. Data source: 2010 Decennial U.S. Census

Divergent color scale

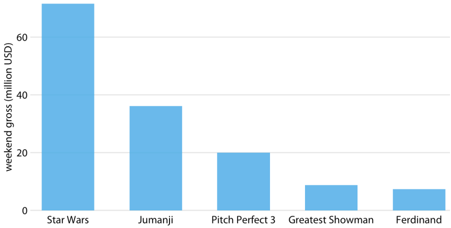
Okabe, M., and K. Ito. 2008. "Color Universal Design (CUD): How to Make Figures and Presentations That Are Friendly to Colorblind People." `http://jfly.iam.u-tokyo.ac.jp/color/`.

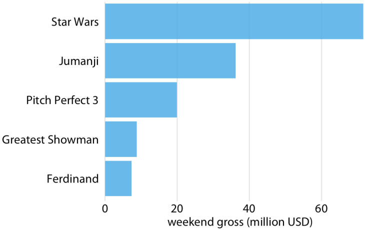
Visualizing amounts

Visualizing amounts — example 1

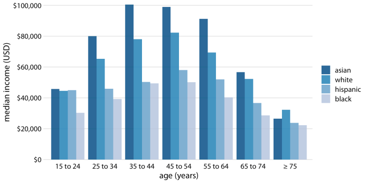
Table 6.1: Highest grossing movies for the weekend of December 22-24, 2017. Data source: Box Office Mojo (http://www.boxofficemojo.com/). Used with permission

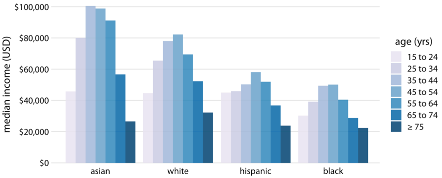
Rank	Title	Weekend gross
1	Star Wars: The Last Jedi	$71,565,498
2	Jumanji: Welcome to the Jungle	$36,169,328
3	Pitch Perfect 3	$19,928,525
4	The Greatest Showman	$8,805,843
5	Ferdinand	$7,316,746



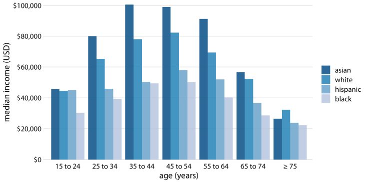


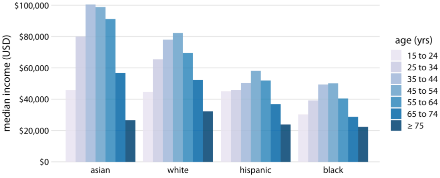
Visualizing amounts — example 2

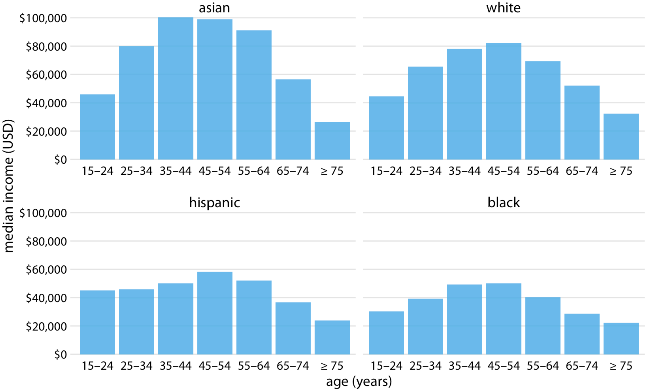




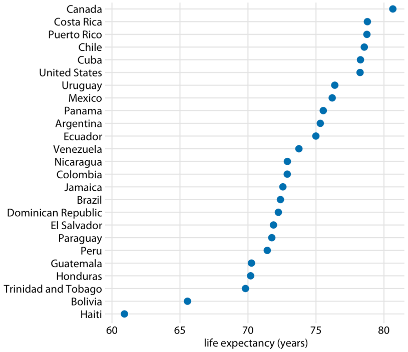
Visualizing amounts — example 2

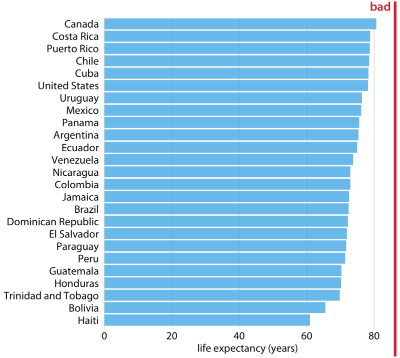






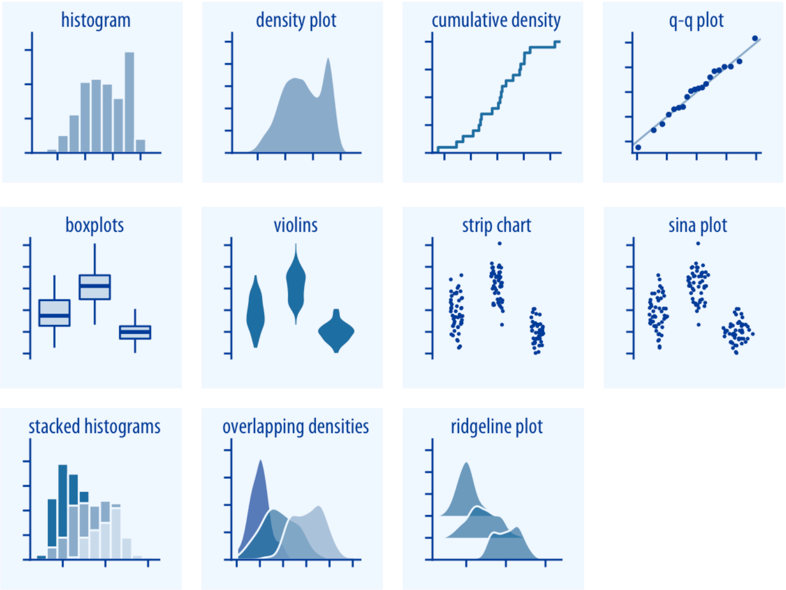
Visualizing amounts — example 3





This dataset is not suitable for being visualized with bars. The bars are too long and they draw attention away from the key feature of the data, the differences in life expectancy among the different countries. Data source: Gapminder project

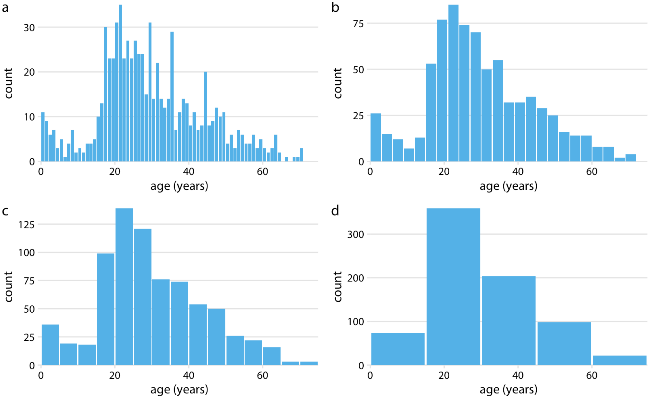
Visualizing distributions



Visualizing distributions — examples

Table 7.1: 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
31–35	76
36–40	74
41–45	54
46–50	50
51–55	26
56–60	22
61–65	16
66–70	3
71–75	3

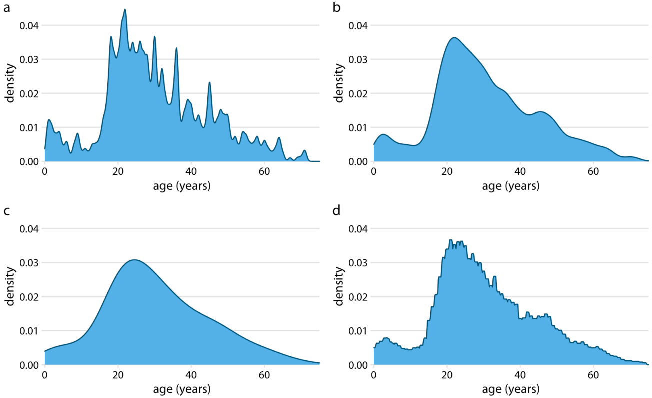


When making a histogram, always explore multiple bin widths

Visualizing distributions — examples

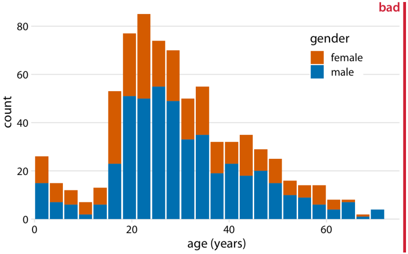
Table 7.1: Numbers of passenger with known age on the Titanic.

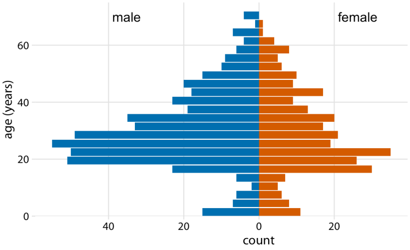
Age range	Count	Age range	Count	Age range	Count
0–5	36	31–35	76	61–65	16
6–10	19	36–40	74	66–70	3
11–15	18	41–45	54	71–75	3
16–20	99	46–50	50		
21–25	139	51–55	26		
26–30	121	56–60	22		

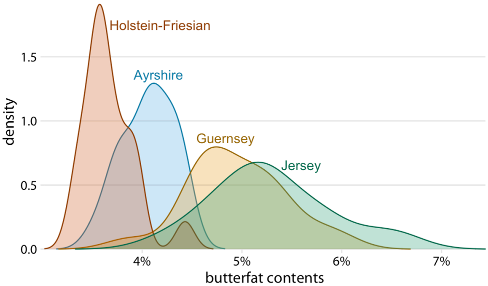


Verify that density doesn’t predict the existence of nonsensical data

Visualizing multiple distributions

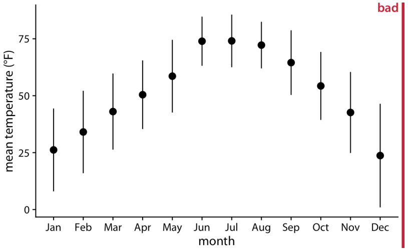




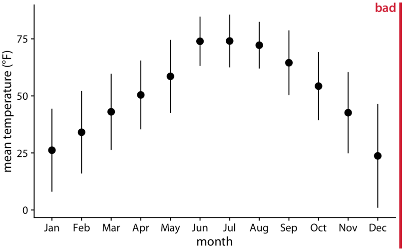


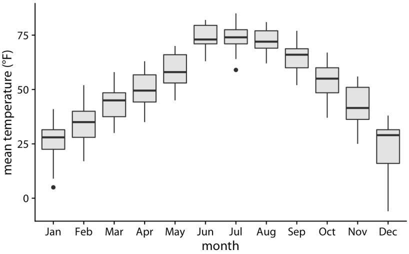
To visualize several distributions at once, kernel density plots will generally work better than histograms.

Visualizing many distributions

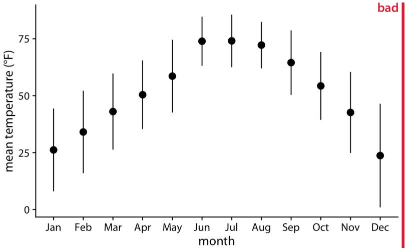


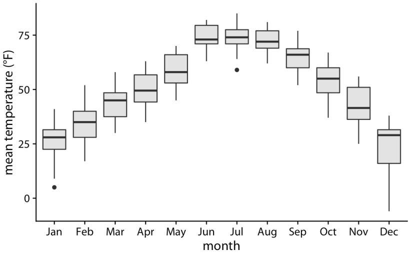
Visualizing many distributions

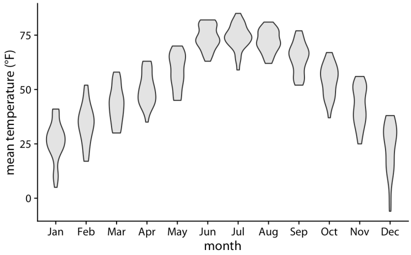




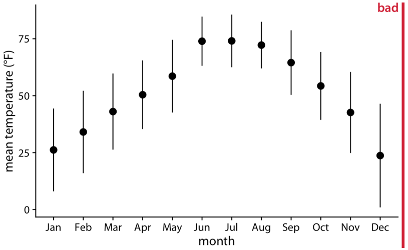
Visualizing many distributions

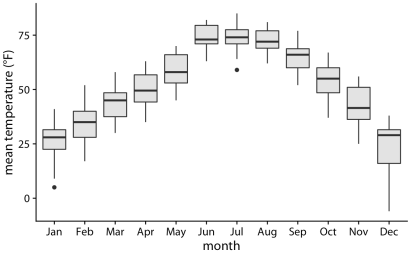


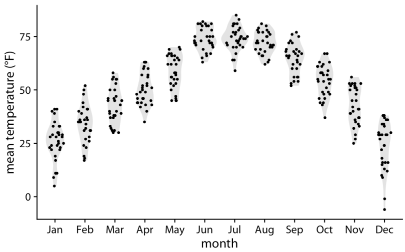




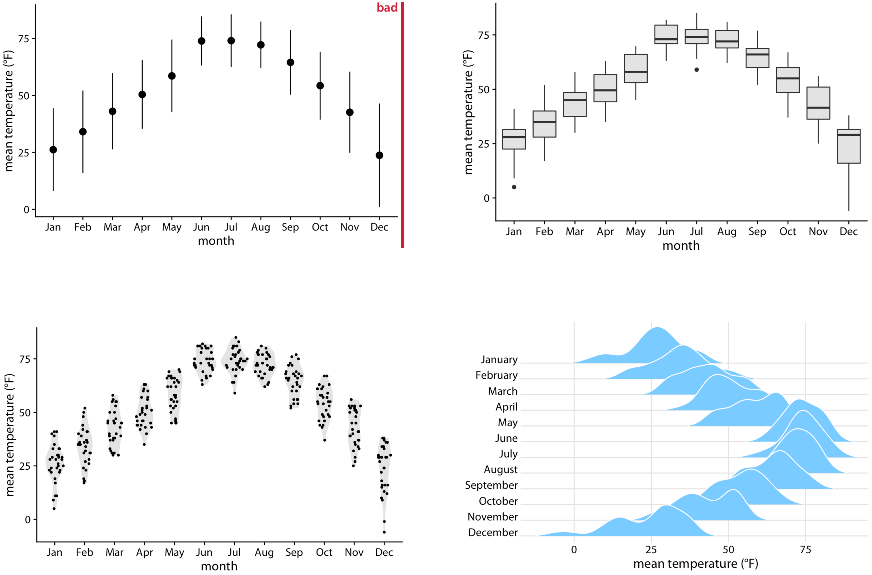
Visualizing many distributions



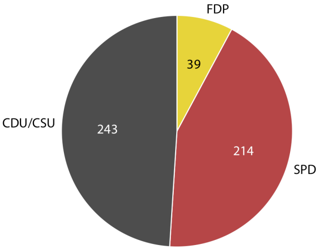




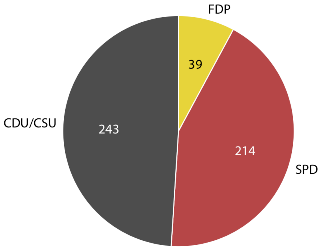
Visualizing many distributions

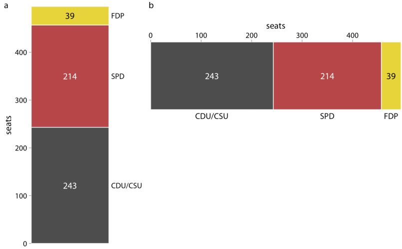


Visualizing proportions

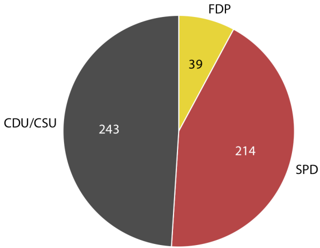


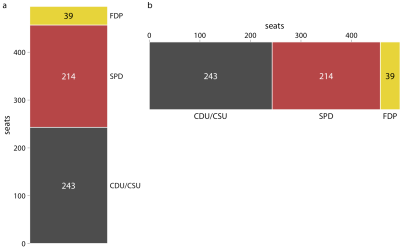
Visualizing proportions

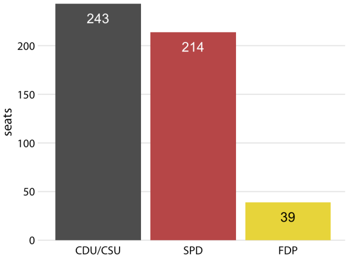




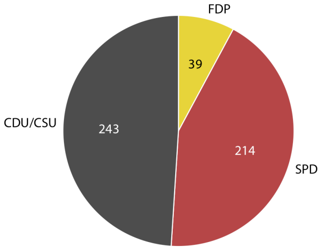
Visualizing proportions

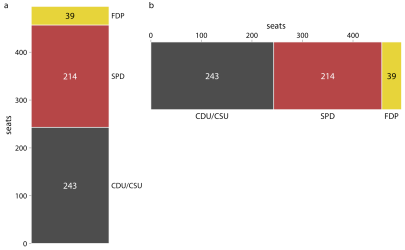






Visualizing proportions





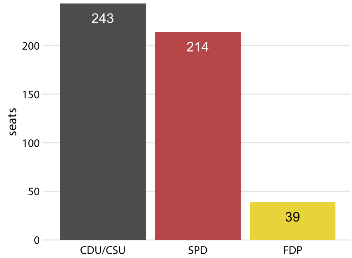


Table 10.1: Pros and cons of common approaches to visualizing proportions: pie charts, stacked bars, and side-by-side bars.

	Pie chart	Stacked bars	Side-by-side bars
Clearly visualizes the data as proportions of a whole	✔	✔	✖
Allows easy visual comparison of the relative proportions	✖	✖	✔
Visually emphasizes simple fractions, such as 1/2, 1/3, 1/4	✔	✖	✖
Looks visually appealing even for very small datasets	✔	✖	✔
Works well when the whole is broken into many pieces	✖	✖	✔
Works well for the visualization of many sets of proportions or time series of proportions	✖	✔	✖

When side-by-side bars win

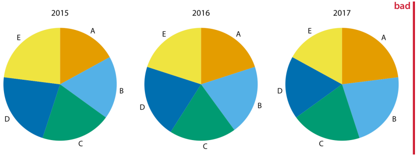


Figure 10.4: Market share of five hypothetical companies, A–E, for the years 2015–2017, visualized as pie charts. This visualization has two major problems: 1. A comparison of relative market share within years is nearly impossible. 2. Changes in market share across years are difficult to see.

When side-by-side bars win

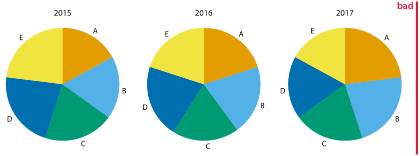
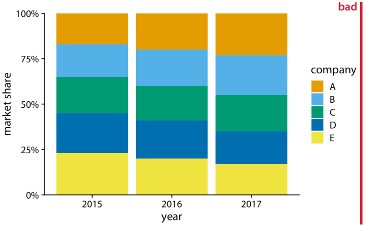


Figure 10.4: Market share of five hypothetical companies, A–E, for the years 2015–2017, visualized as pie charts. This visualization has two major problems: 1. A comparison of relative market share within years is nearly impossible. 2. Changes in market share across years are difficult to see.



When side-by-side bars win

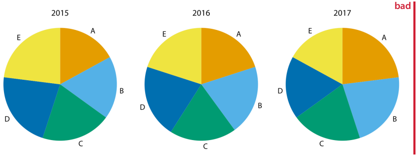
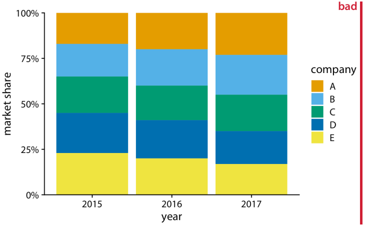


Figure 10.4: Market share of five hypothetical companies, A–E, for the years 2015–2017, visualized as pie charts. This visualization has two major problems: 1. A comparison of relative market share within years is nearly impossible. 2. Changes in market share across years are difficult to see.



When side-by-side bars win

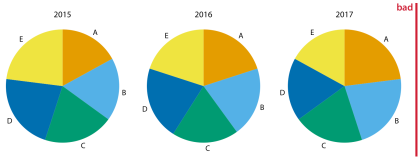
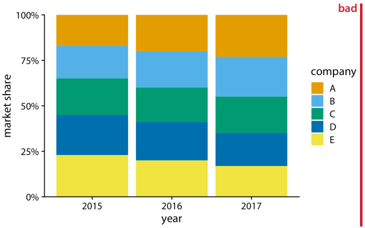
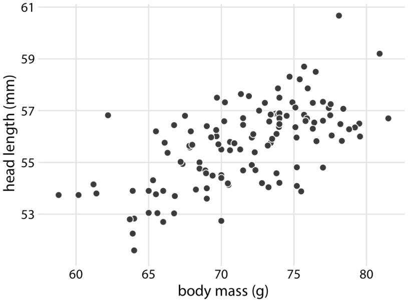


Figure 10.4: Market share of five hypothetical companies, A–E, for the years 2015–2017, visualized as pie charts. This visualization has two major problems: 1. A comparison of relative market share within years is nearly impossible. 2. Changes in market share across years are difficult to see.

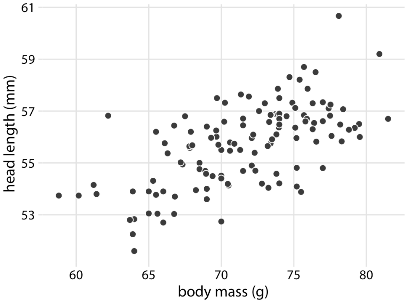


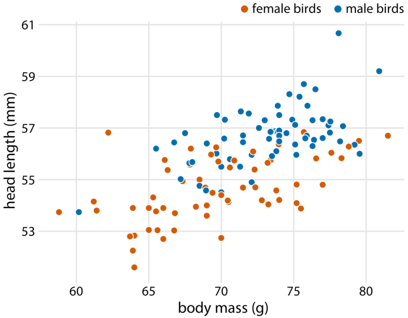
Humans are not good at computing integrals in their heads, so comparing lengths is much easier than comparing areas.

Visualizing x-y relationships

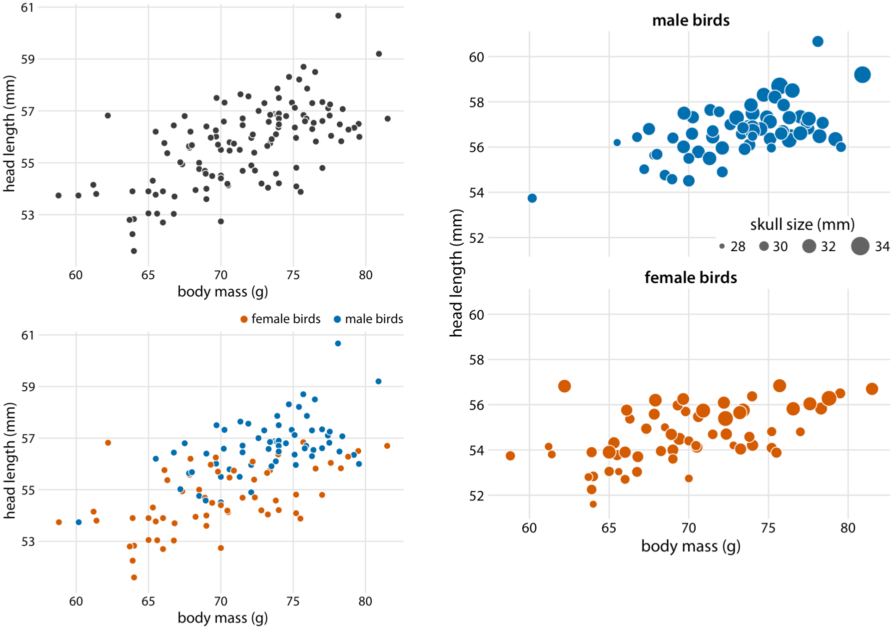


Visualizing x-y relationships

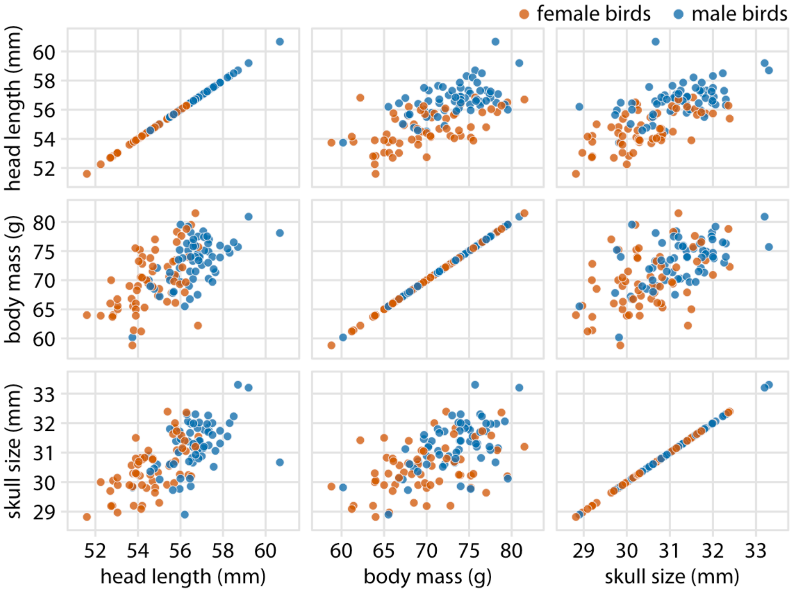




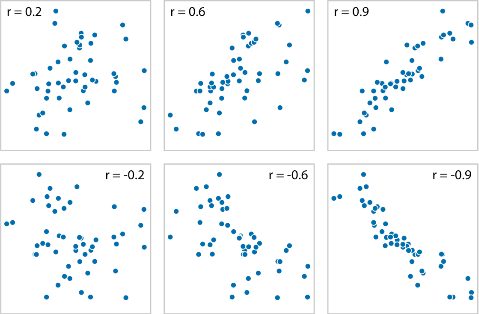
Visualizing x-y relationships



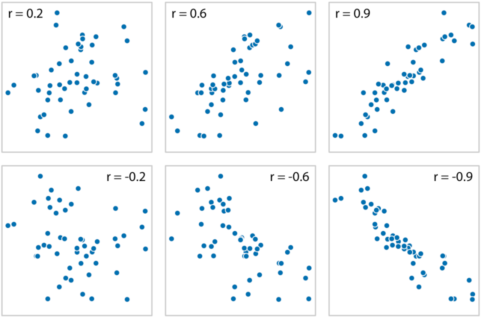
Scatter matrix plot

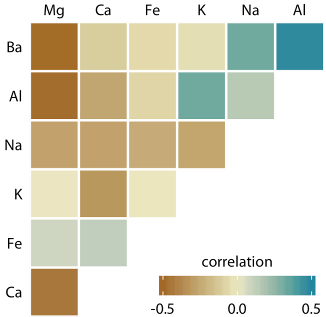


Correlograms

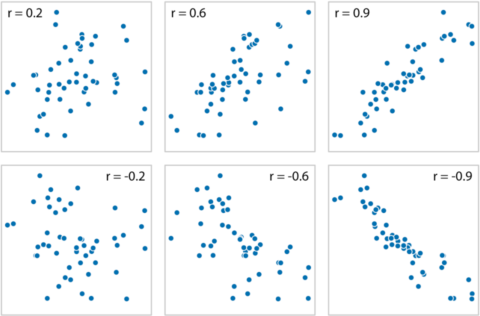


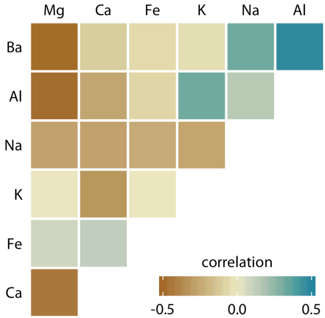
Correlograms





Correlograms

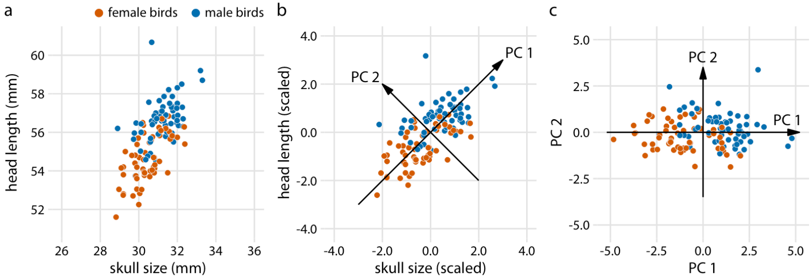




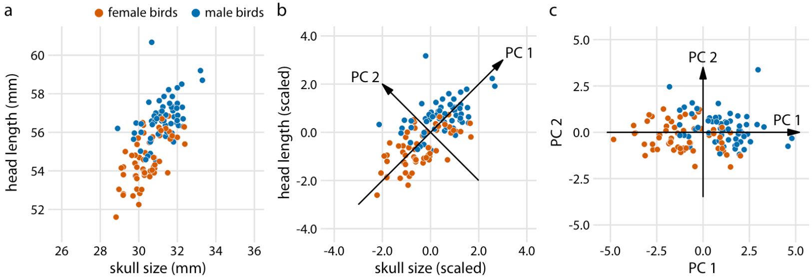
Non-Linear Dependence

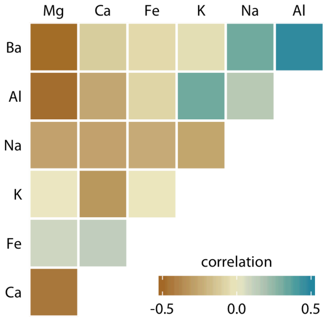
Lack of linear correlation does not imply lack of dependence

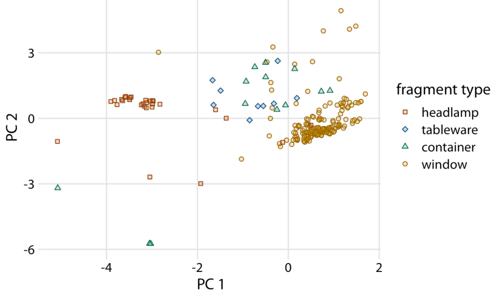
Dimension reduction



Dimension reduction

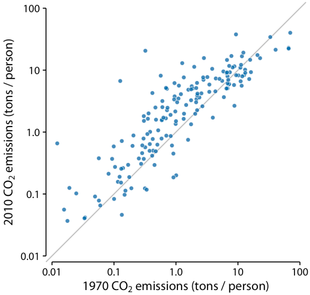






Paired data

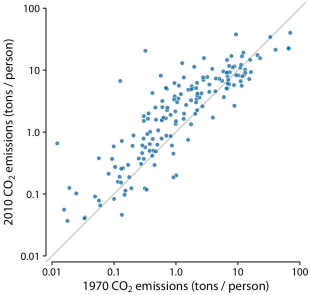
Scatterplots and slopegraphs are two main choices for plotting paired data.

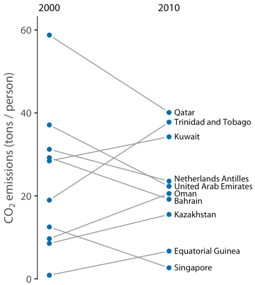


The last plot shows that slopegraph can accomodate short time series.

Paired data

Scatterplots and slopegraphs are two main choices for plotting paired data.

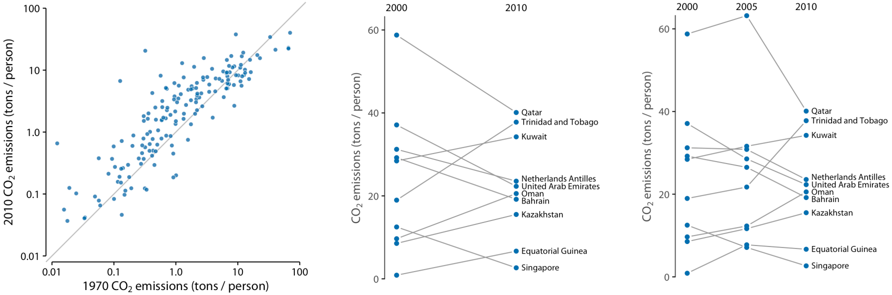




The last plot shows that slopegraph can accomodate short time series.

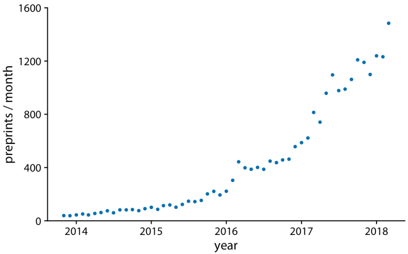
Paired data

Scatterplots and slopegraphs are two main choices for plotting paired data.

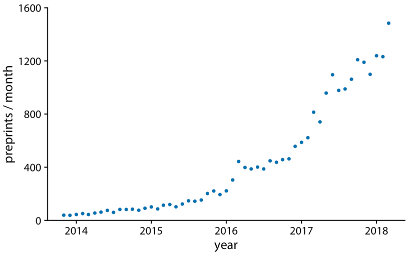


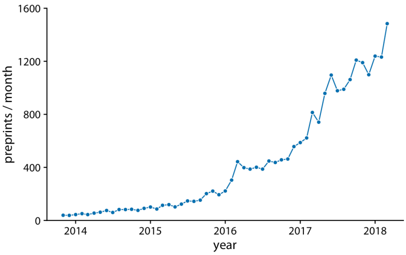
The last plot shows that slopegraph can accomodate short time series.

Visualizing time series — univariate

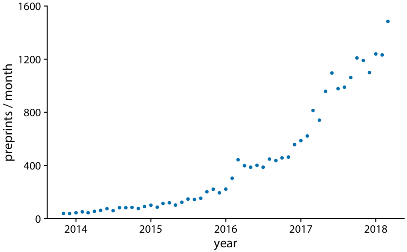


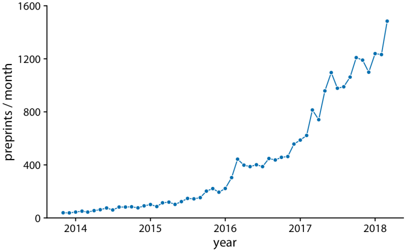
Visualizing time series — univariate

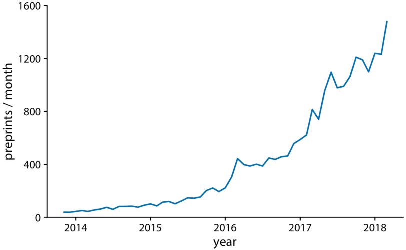




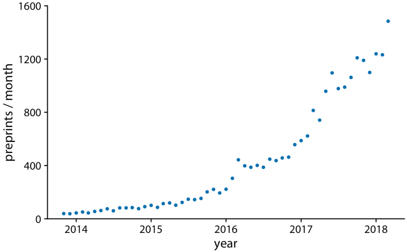
Visualizing time series — univariate

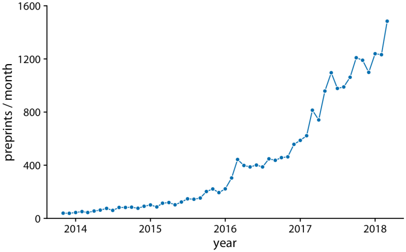


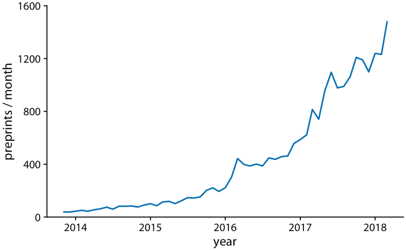




Visualizing time series — univariate

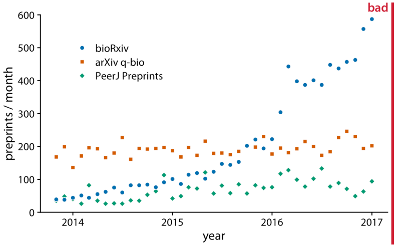




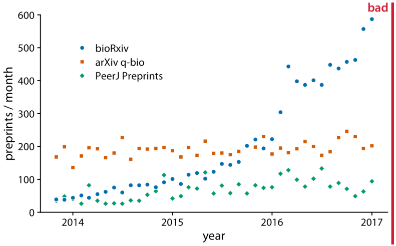


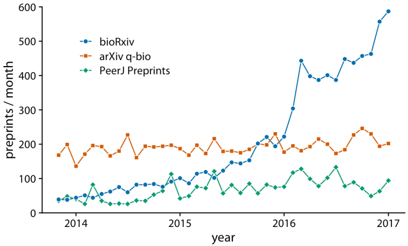
For dense time series, connect the dots and omit them.

Visualizing time series — multivariate, the same y-axis

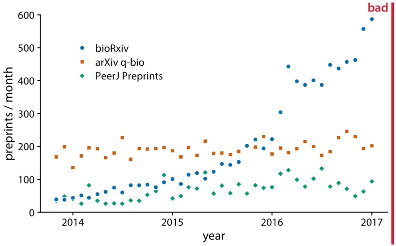


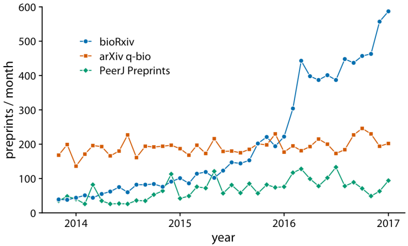
Visualizing time series — multivariate, the same y-axis

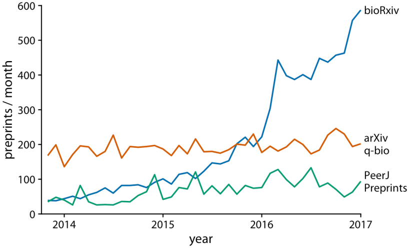




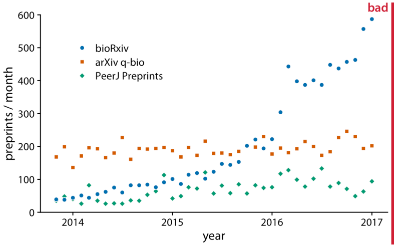
Visualizing time series — multivariate, the same y-axis

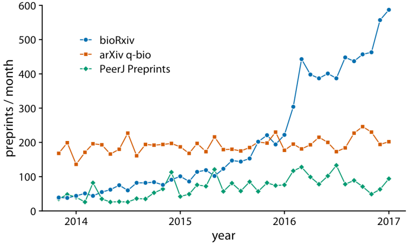


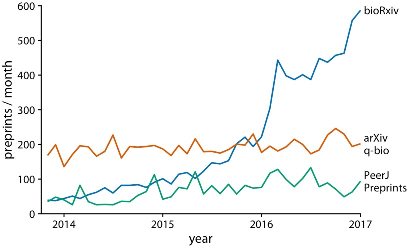




Visualizing time series — multivariate, the same y-axis



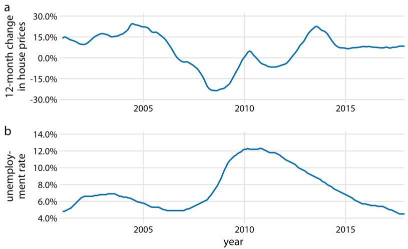




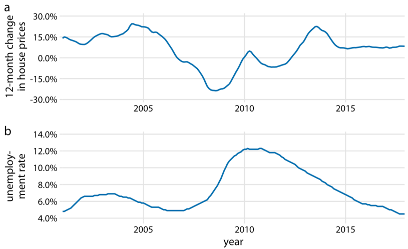
Consider replacing legends with direct labeling.

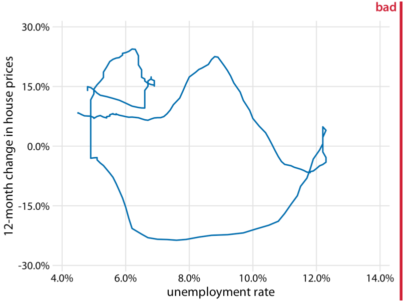
Make sure it is easy to compare objects of interest

Visualizing time series — more than one y-axis

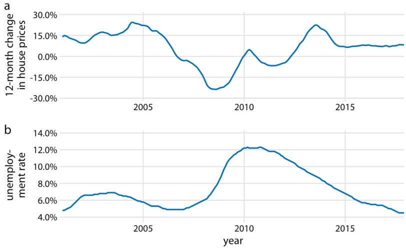


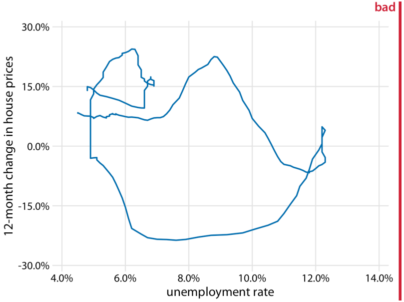
Visualizing time series — more than one y-axis

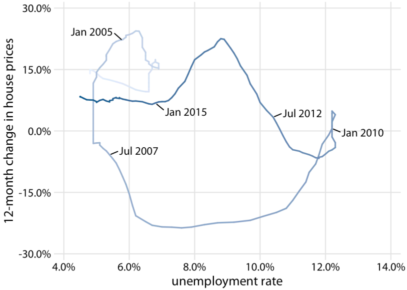




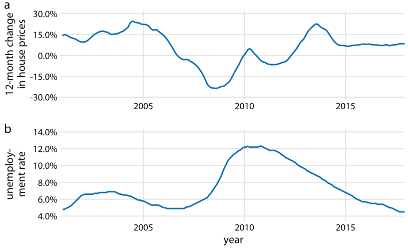
Visualizing time series — more than one y-axis

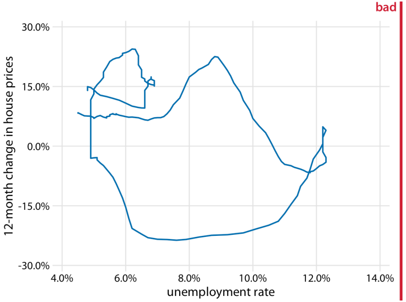






Visualizing time series — more than one y-axis





Connected scatter plots are great, but don't forget to indicate both the direction and the temporal scale of the data. .

When you have more than two y-axes, use dimension reduction techniques to map $\mathbb{R}^n$ onto $\mathbb{R}^2$.