Data Science and Data Visualization

Volodymyr Minin

UCI Department of Statistics
Donald Bren School of Information & Computer Sciences

ISI-BUDS
July 11 2023

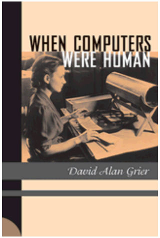
Plan for today

- What is Data Science?
- Data science in the real-world
- Data visualization

Computers and Data

The historical meaning of the term “computer”:

“one who computes” (i.e., a person)



Since the 1700’s, statisticians have been using “computers” to analyze data – so its not a new idea

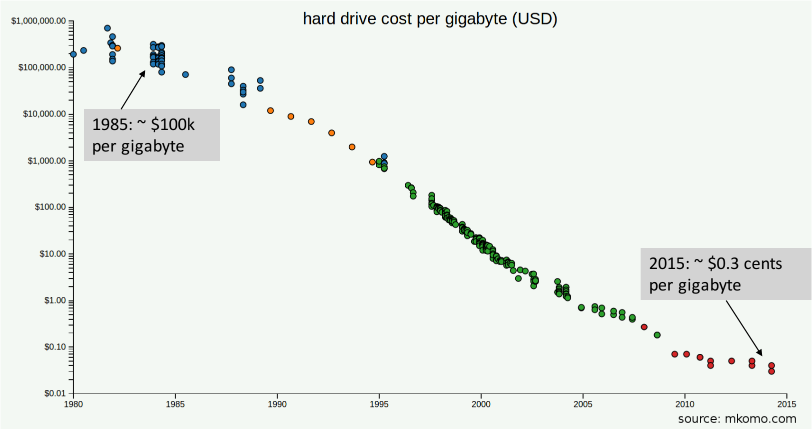
For example, Karl Pearson, one of the founders of statistics, directed a team of “computers” in his lab in London around the early 1900’s

…..but for many years, “computers” could only work on relatively small problems

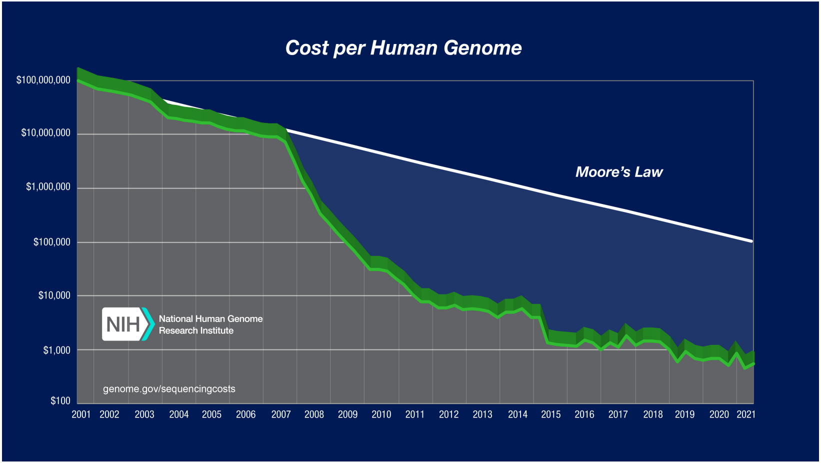
Statistics and Modern Computing

- Post World War II
  - Increasing use of computing to solve algorithmic aspects of statistical analyses
- 1960's
  - Development of statistical computing and exploratory data analysis
- 1980's
  - Computing allowed statisticians to explore more flexible models
  - Increase in use of "non-parametric" techniques and simulation methods
- 1990's
  - Development of "machine learning" — very flexible predictive modeling techniques developed in computer science
- Today
  - Data science = computing + statistics + applications

Data storage became cheaper



Data revolution in Biology



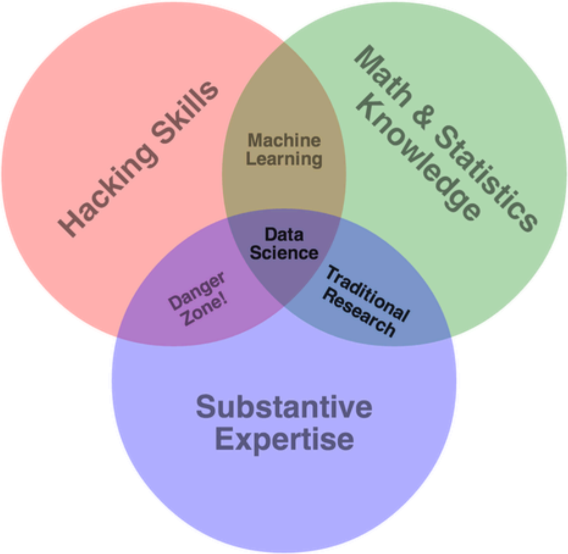
A Paradigm shift in data analysis

- Technological drivers
  - Sensors (cheap and ubiquitous, e.g., GPS on your phone)
  - Data storage (we are all "data owners")
  - Computational power
  - Data analysis methods (statistics and machine learning)
  - Internet and wireless communication (can collect and share data)
- Convergence — tremendous demand for data analysis
  - In business, in sciences, in medicine, in engineering, and more......
- In the past, this demand was met by statistics
  - Does not scale up — there are not nearly enough statisticians
  - Need more tools than just statistics: need databases, algorithms, machine learning,...

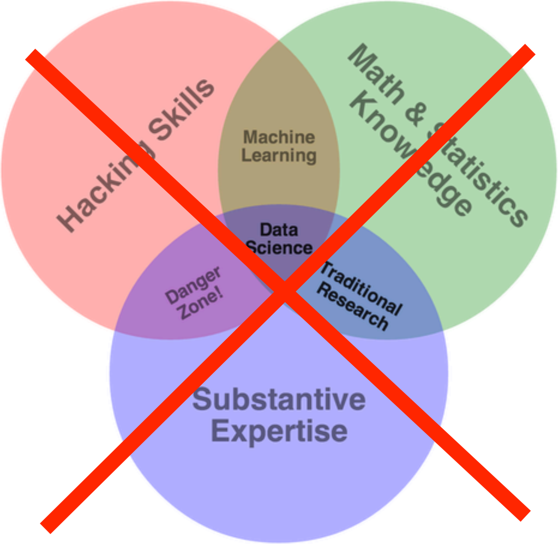
What is Data Science?

- Data science involves the full lifecycle of data: from messy unstructured data to predictions and decisions
- Data science is broader than just databases, statistics, ML, algorithms, but these are all critical components
- Key aspects of data science include
  - Domain knowledge and problem definition
  - Data preparation/organization/management
  - Understanding of uncertainty (statistics)
  - Computing, algorithms, fitting models, machine learning
  - Iterative exploration and experimentation
  - Human judgement and interpretation

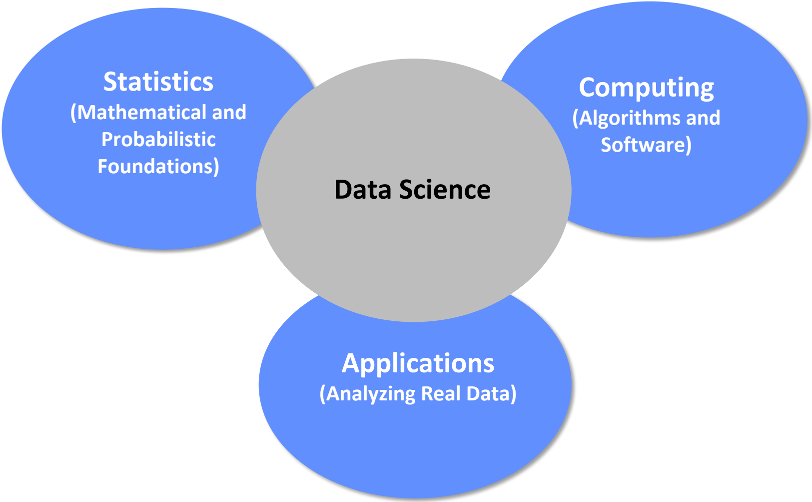
Components of Data Science



Components of Data Science



Components of Data Science

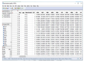


Data pipeline

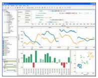
Unstructured Data

Extracted Data

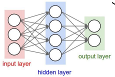
Transformed Data



Data for Modeling

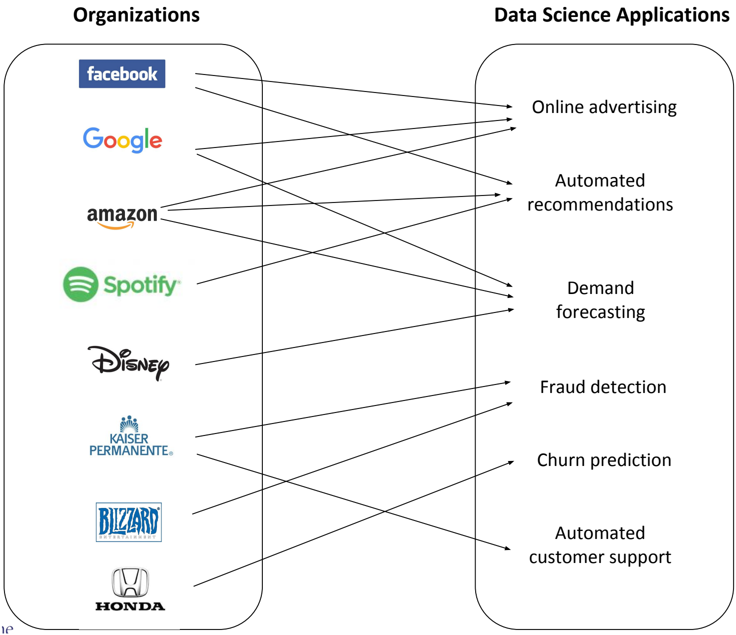


Predictive Model



Predictions/ Decisions

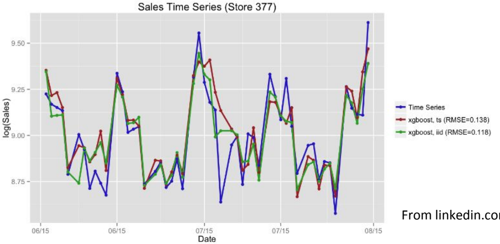
How is Data Science used?



How does Amazon forecast how many items for its warehouses?

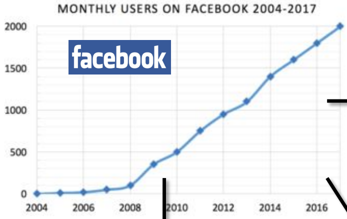
From dailymail.co.uk

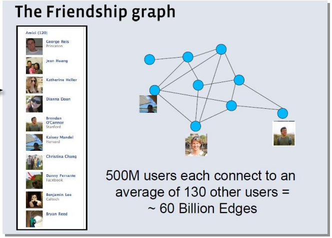
From www.formaspace.com

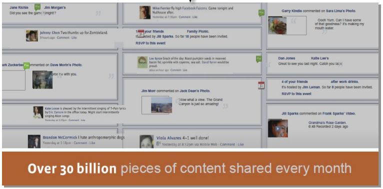


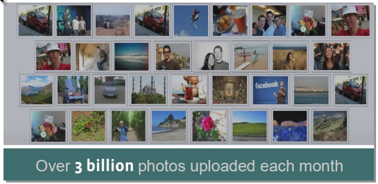
From linkedin.com

How does Facebook predict what content to show you?









Graphics from Lars Backstrom, ESWC 2011

How do companies decide what ads to show you?


U.S. INTERNATIONAL 中文网

?

The New York Times

?

Tuesday, March 4, 2014 | Today's Paper | Personalize Your Weather

WORLD U.S. NEW YORK BUSINESS OPINION SPORTS SCIENCE ARTS FASHION & STYLE VIDEO All Sections

?

TURMOIL IN UKRAINE

Putin, Flashing Disdain, Defends Action in Crimea

By STEVEN LEE MYERS 59 minutes ago

President Vladimir V. Putin's first public remarks on the political upheaval in Ukraine were aimed at both international and domestic audiences, defending Russia from the fury of global criticism and rallying support at home.

NEWS ANALYSIS

No Easy Way Out of Ukraine Crisis

By PETER BAKER 54 minutes ago

White House officials are weighing their options, knowing that reversing the occupation of Crimea would be difficult, if not impossible, in the short run.

Uriel Sinai for The New York Times

Ukrainian riot police officers stood guard at an anti-Russian rally in Donetsk on Tuesday.

Crimea's Pro-Russian Leader Says Region Is Secure

By DAVID M. HERSZENHORN 8:21 PM ET

The prime minister of the autonomous region offered the assurance on Tuesday even as armed standoffs continued.

RELATED COVERAGE

- **Kerry Takes Offer of Aid to Ukraine** 33 minutes ago
- **Cyberattacks Rise as Crisis Spills to Internet** 6:47 PM ET
- VIDEO: **Confrontation in Crimea**

The Opinion Pages

OP-ED CONTRIBUTOR

Has Privacy Become a Luxury Good?

By JULIA ANGWIN

It takes a lot of money and time to avoid hackers and data miners.

- **Editorial: Frustration With Afghanistan**
- **Brooks: Putin Can't Stop**
- **Cohen: Russia's Crimean Crime**

DRAFT

My Character to Kill

By ALEX BERENSON

I'm not sure I can say goodbye to a man who has defined my creative life for so long — and who will pay the mortgage for at least one more contract.

- **Op-Docs: 'Chinese, on the Inside'**

MARKETS » At 10:03 PM ET

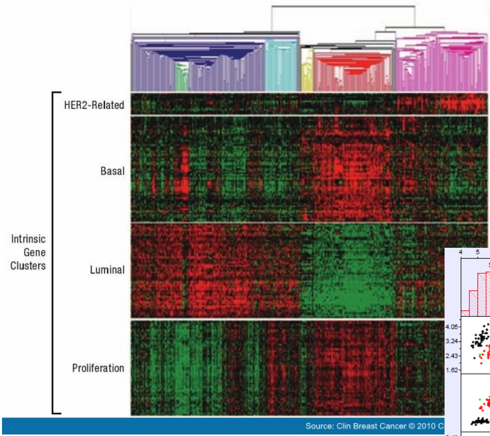
JAPAN		CHINA
Nikkei	HangSeng	Shanghai
14,942.78	22,690.46	2,059.39
+221.30	+32.83	−12.09
+1.50%	+0.14%	−0.58%

Data delayed at least 15 minutes

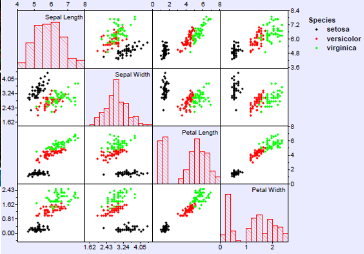
Get Quotes | My Portfolios »

?

How can we make personalized recommendations in medicine?



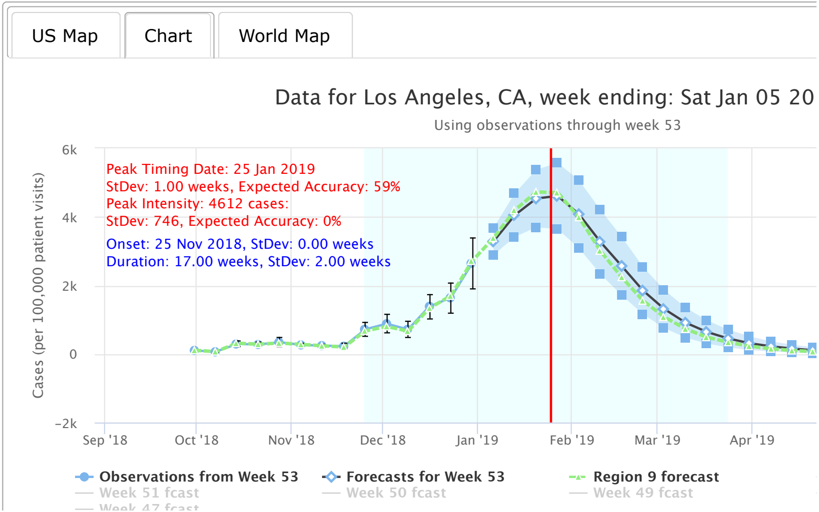
Data Matrix:
Rows = genes
Columns = patients



From www.originlab.com

How do public health workers predict infectious disease outbreaks?

Influenza Observations and Forecast

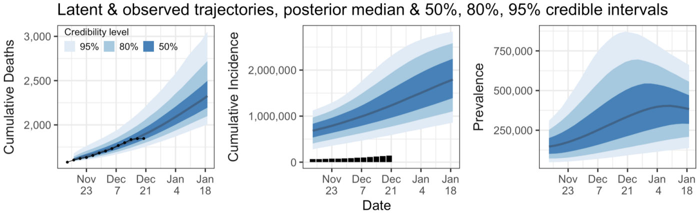


https://cpid.iri.columbia.edu

Orange County, CA COVID-19 Situation Report, December 28, 2020

Report period: Nov 15 - Dec 20 (we don't use the most recent data due to reporting delays)

The goal of this report is to inform interested parties about dynamics of SARS-CoV-2 spread in Orange County, CA and to predict epidemic trajectories. Methodological details are provided below and in the accompanying manuscript. We are also contributing to COVID Trends by UC Irvine project that provides data visualizations of California County trends across time and space.



https://www.stat.uci.edu/oc_covid_model/

Questions?

Data visualization: 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

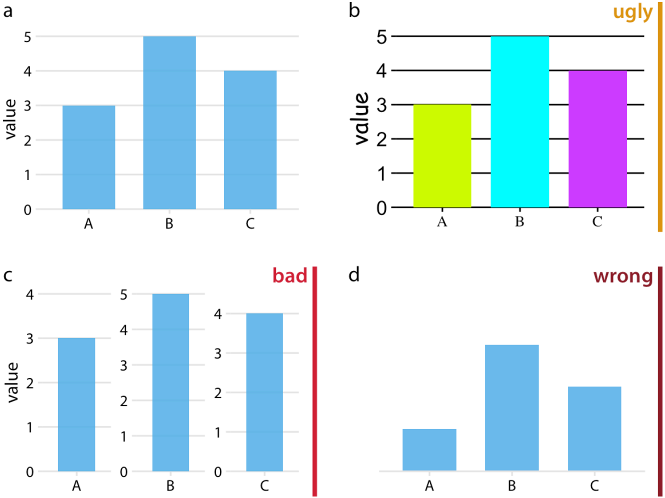
- Pioneered by John Tukey (statistician at Bell Labs, Princeton) in the 1960's
  - "let the data speak"

Recommended reading

Fundamentals of Data Visualization

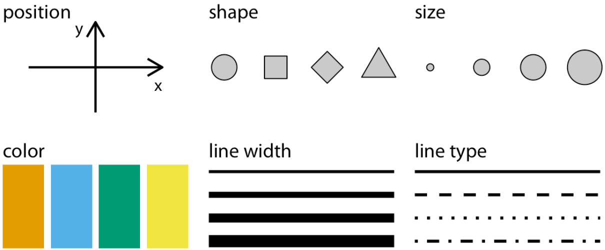
Claus O. Wilke

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

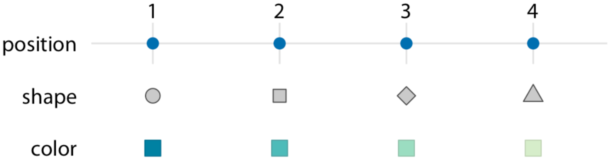


Mapping data onto aesthetics

Types of aesthetics:



Scales map data values onto aesthetics:



Mapping data onto aesthetics — example

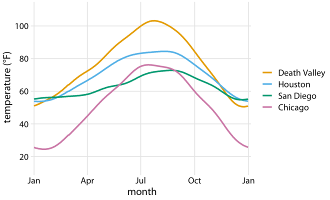
Table 2.2: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.

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.

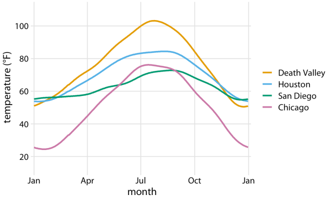
Month	Day	Location	Station ID	Temperature
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319	51.2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

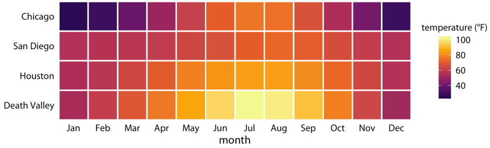


Mapping data onto aesthetics — example

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

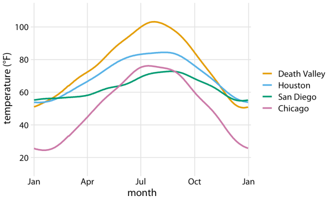


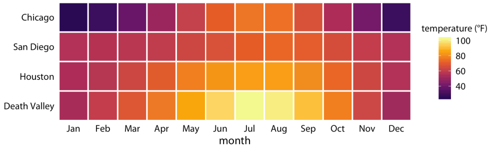


Mapping data onto aesthetics — example

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

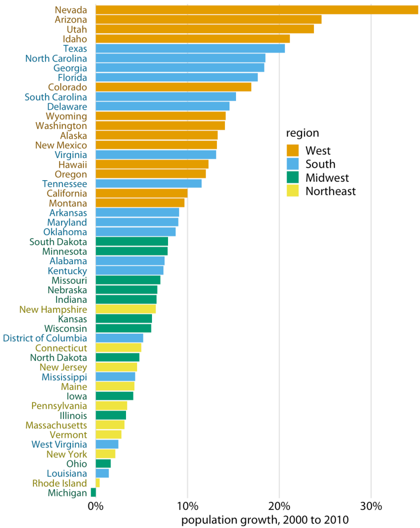


Figure 4.2: Population growth in the U.S. from 2000 to 2010. States in the West and South have seen the largest increases, whereas states in the Midwest and Northeast have seen much smaller increases or even, in the case of Michigan, a decrease. Data source: U.S. Census Bureau

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

Color as a tool to highlight

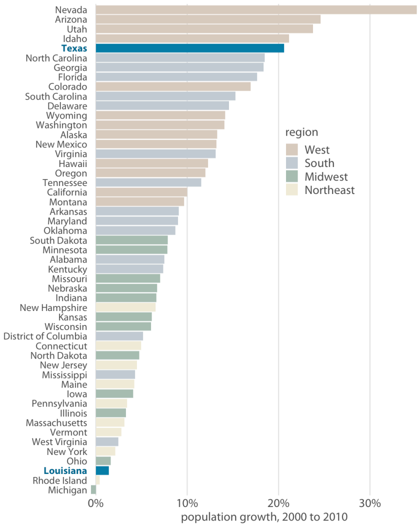


Figure 4.8: From 2000 to 2010, the two neighboring southern states Texas and Louisiana have experienced among the highest and lowest population growth across the U.S. Data source: U.S. Census Bureau

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

Color to represent data values

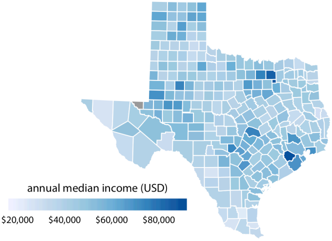


Figure 4.4: Median annual income in Texas counties. The highest median incomes are seen in major Texas metropolitan areas, in particular near Houston and Dallas. No median income estimate is available for Loving County in West Texas and therefore that county is shown in gray. Data source: 2015 Five-Year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

Sequential color scale

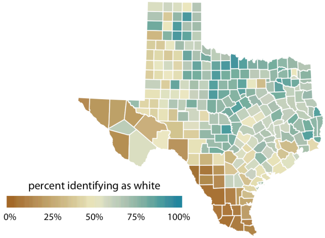


Figure 4.6: Percentage of people identifying as white in Texas counties. Whites are in the majority in North and East Texas but not in South or West Texas. Data source: 2010 Decennial U.S. Census

Divergent color scale

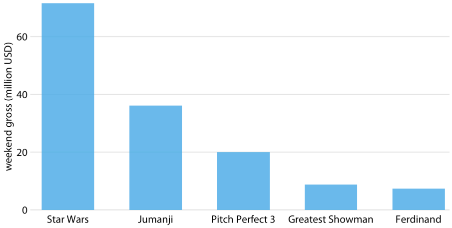
Okabe, M., and K. Ito. 2008. "Color Universal Design (CUD): How to Make Figures and Presentations That Are Friendly to Colorblind People." `http://jfly.iam.u-tokyo.ac.jp/color/`.

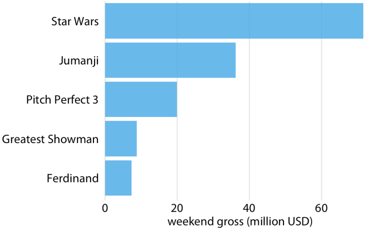
Visualizing amounts

Visualizing amounts — example 1

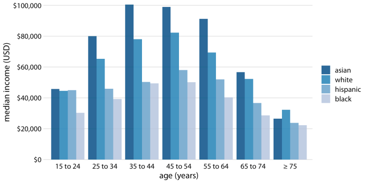
Table 6.1: Highest grossing movies for the weekend of December 22-24, 2017. Data source: Box Office Mojo (http://www.boxofficemojo.com/). Used with permission

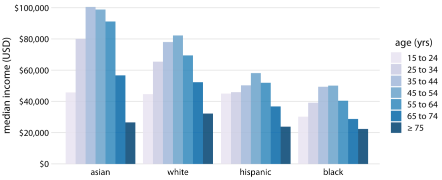
Rank	Title	Weekend gross
1	Star Wars: The Last Jedi	$71,565,498
2	Jumanji: Welcome to the Jungle	$36,169,328
3	Pitch Perfect 3	$19,928,525
4	The Greatest Showman	$8,805,843
5	Ferdinand	$7,316,746



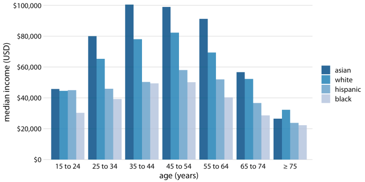


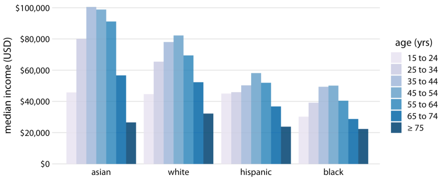
Visualizing amounts — example 2

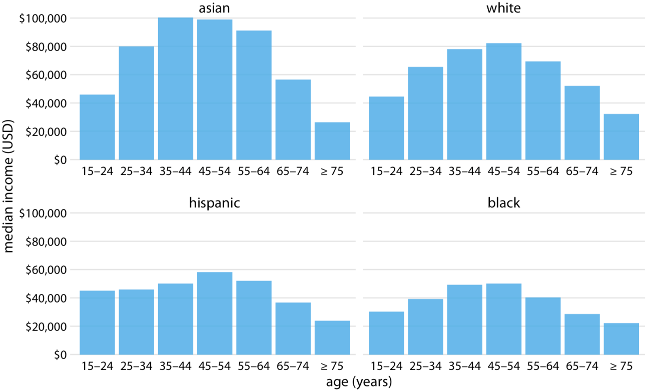




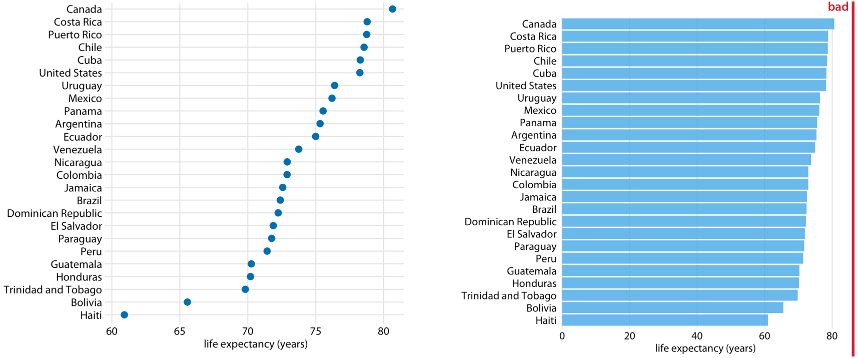
Visualizing amounts — example 2





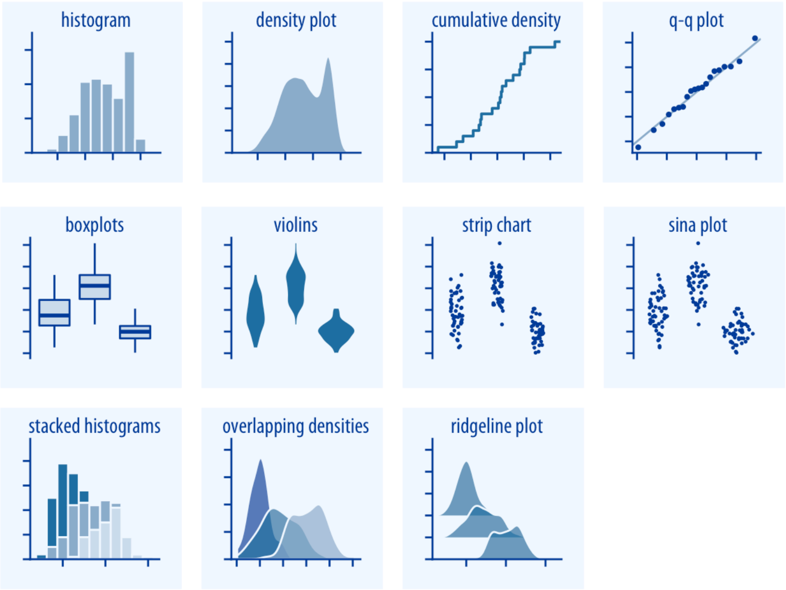


Visualizing amounts — example 3



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

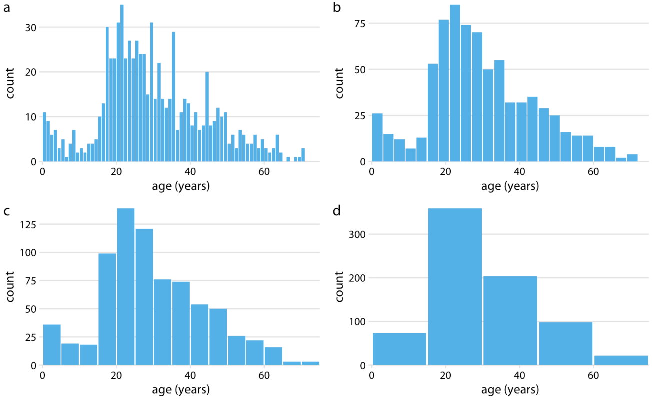
Visualizing distributions



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

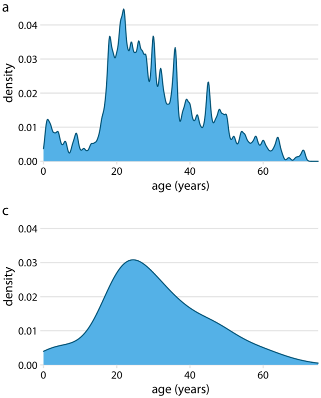


When making a histogram, always explore multiple bin widths

Visualizing distributions — examples

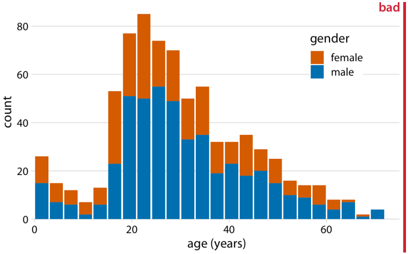
Table 7.1: Numbers of passenger with known age on the Titanic.

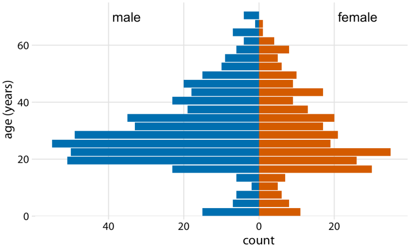
Age range	Count	Age range	Count	Age range	Count
0–5	36	31–35	76	61–65	16
6–10	19	36–40	74	66–70	3
11–15	18	41–45	54	71–75	3
16–20	99	46–50	50		
21–25	139	51–55	26		
26–30	121	56–60	22		

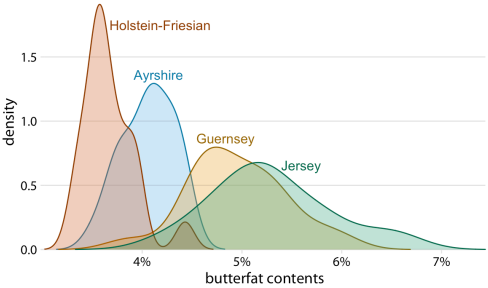


Verify that density doesn't predict the existence of nonsensical data

Visualizing multiple distributions

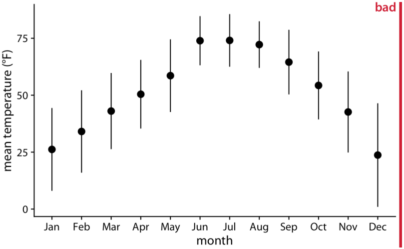




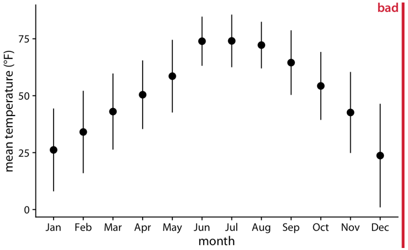


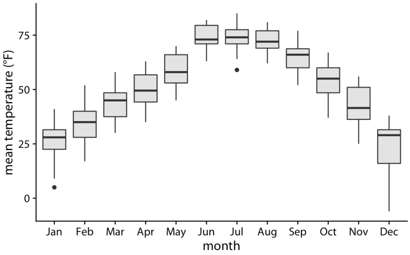
To visualize several distributions at once, kernel density plots will generally work better than histograms.

Visualizing many distributions

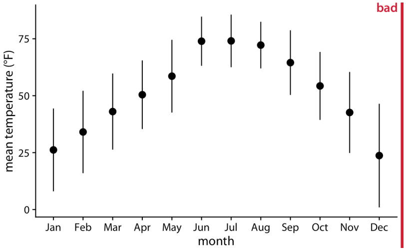


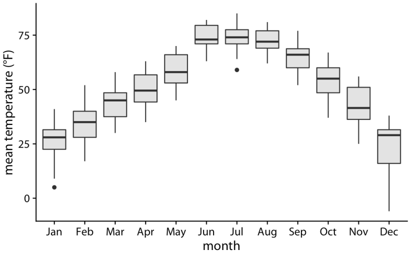
Visualizing many distributions

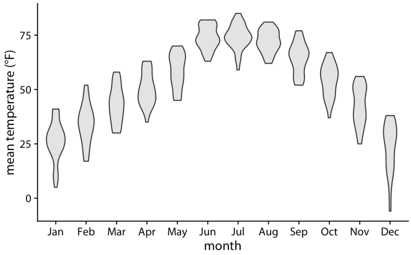




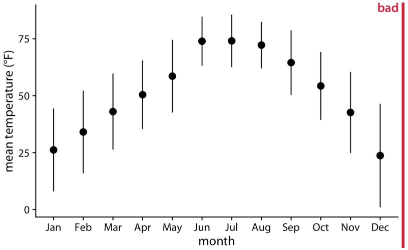
Visualizing many distributions

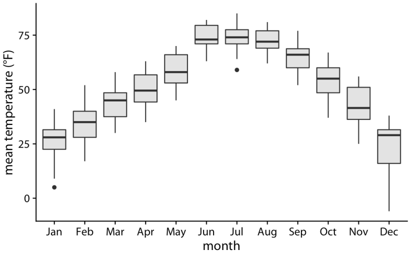


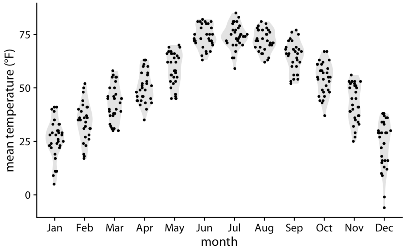




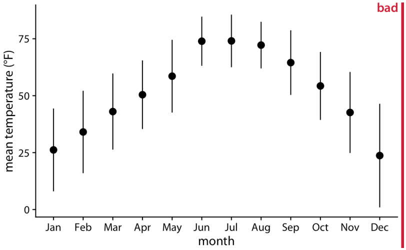
Visualizing many distributions

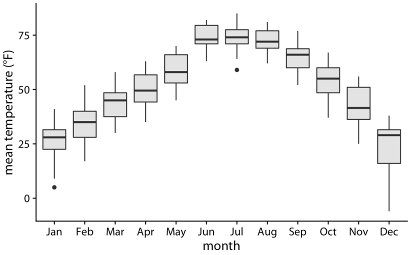


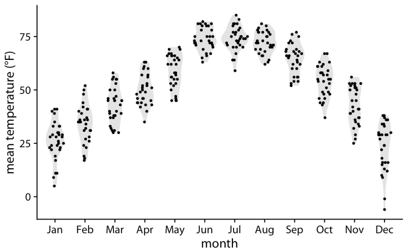


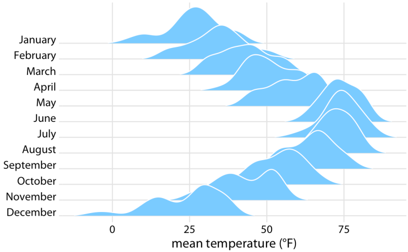


Visualizing many distributions

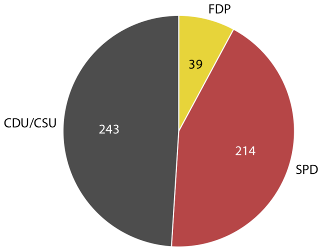




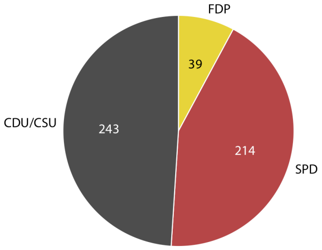


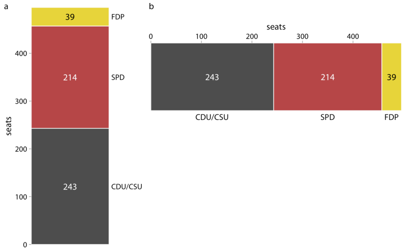


Visualizing proportions

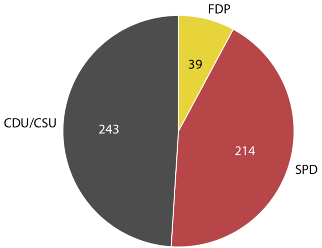


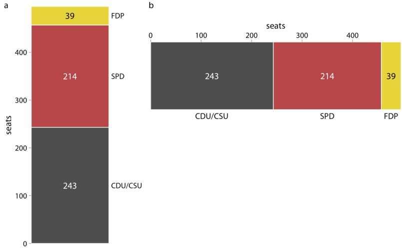
Visualizing proportions

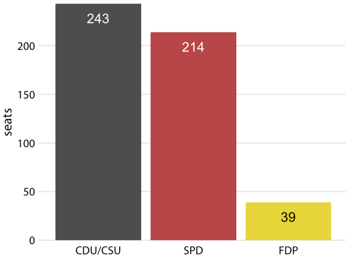




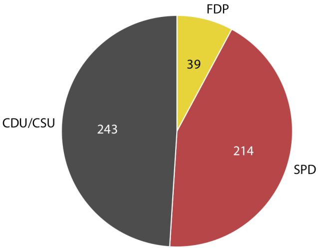
Visualizing proportions

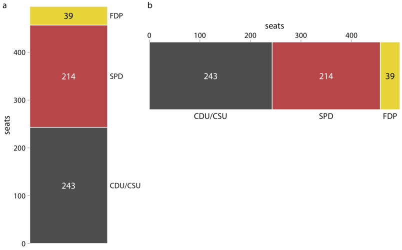






Visualizing proportions





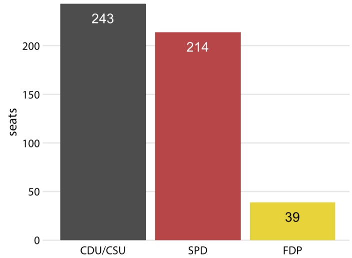


Table 10.1: Pros and cons of common approaches to visualizing proportions: pie charts, stacked bars, and side-by-side bars.

	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

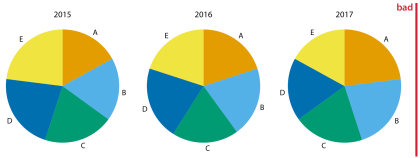


Figure 10.4: Market share of five hypothetical companies, A–E, for the years 2015–2017, visualized as pie charts. This visualization has two major problems: 1. A comparison of relative market share within years is nearly impossible. 2. Changes in market share across years are difficult to see.

When side-by-side bars win

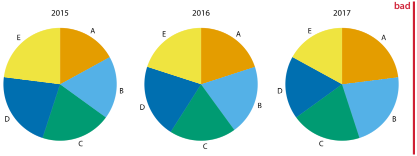
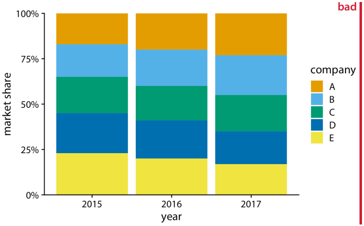


Figure 10.4: Market share of five hypothetical companies, A–E, for the years 2015–2017, visualized as pie charts. This visualization has two major problems: 1. A comparison of relative market share within years is nearly impossible. 2. Changes in market share across years are difficult to see.



When side-by-side bars win

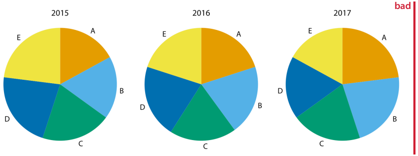
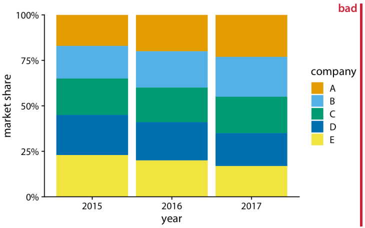


Figure 10.4: Market share of five hypothetical companies, A–E, for the years 2015–2017, visualized as pie charts. This visualization has two major problems: 1. A comparison of relative market share within years is nearly impossible. 2. Changes in market share across years are difficult to see.



When side-by-side bars win

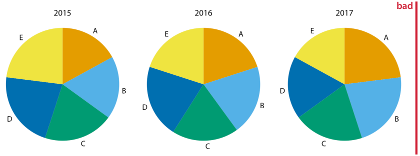
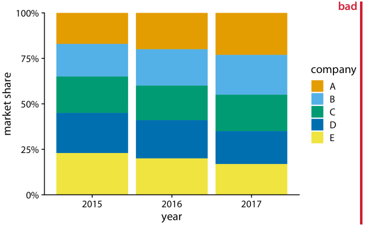
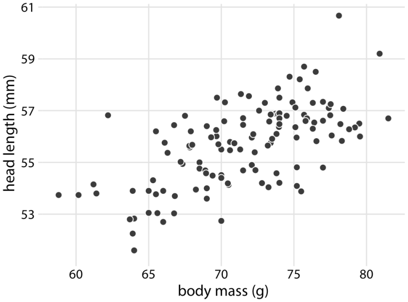


Figure 10.4: Market share of five hypothetical companies, A–E, for the years 2015–2017, visualized as pie charts. This visualization has two major problems: 1. A comparison of relative market share within years is nearly impossible. 2. Changes in market share across years are difficult to see.

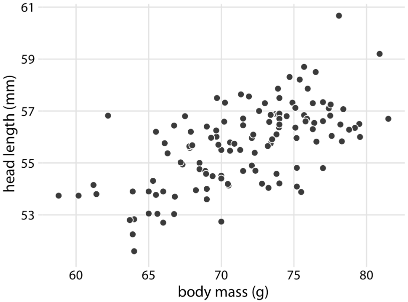


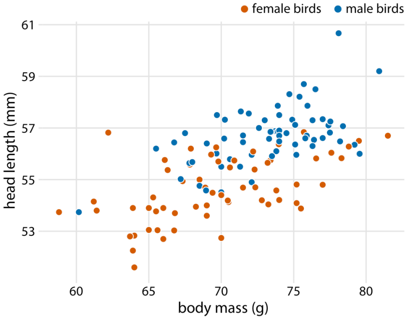
Humans are not good at computing integrals in their heads, so comparing lengths is much easier than comparing areas.

Visualizing x-y relationships

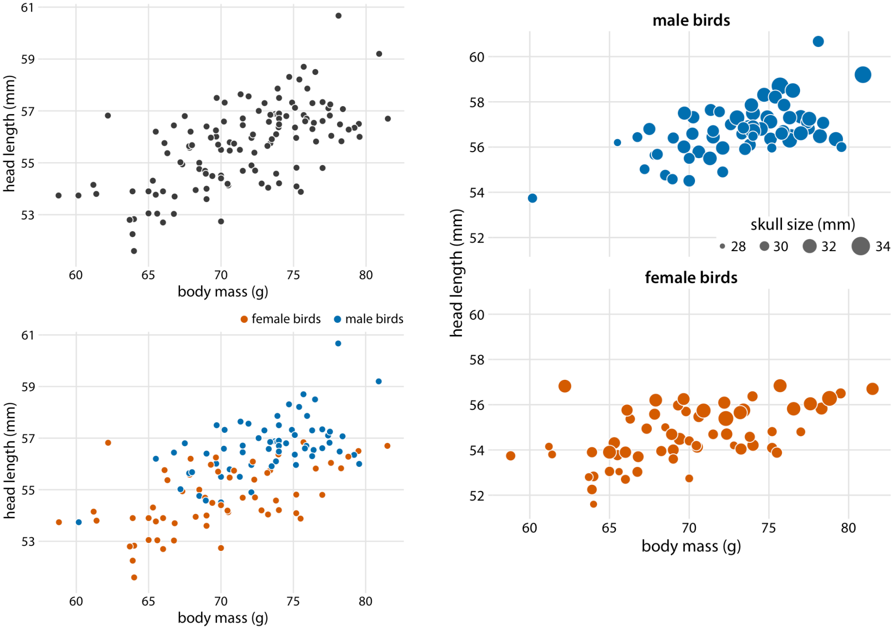


Visualizing x-y relationships

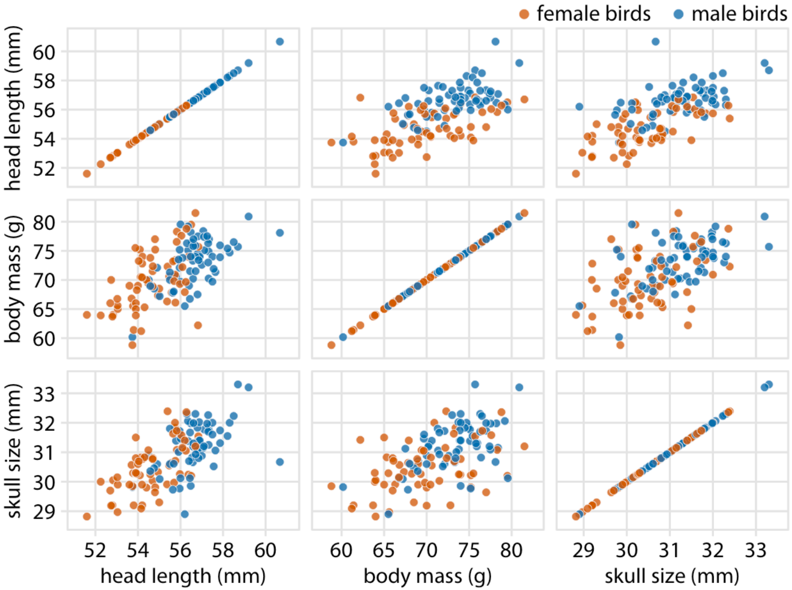




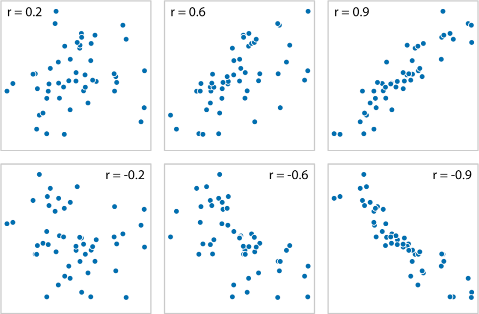
Visualizing x-y relationships



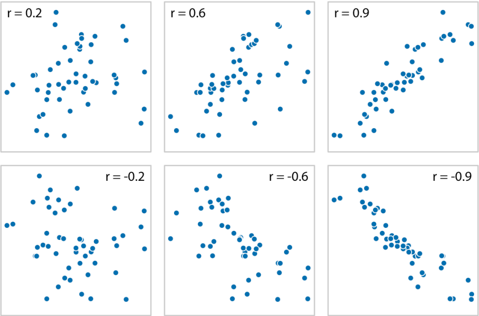
Scatter matrix plot

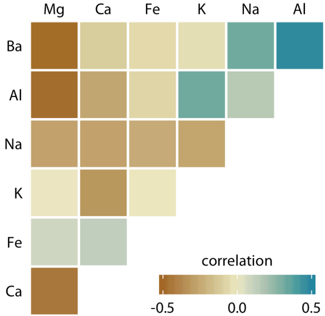


Correlograms

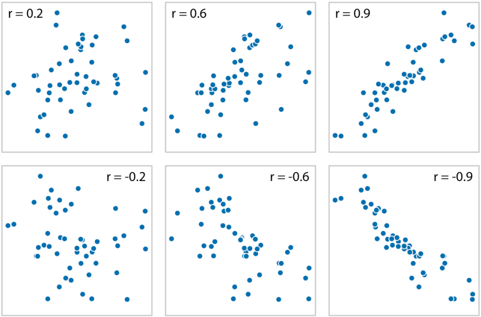


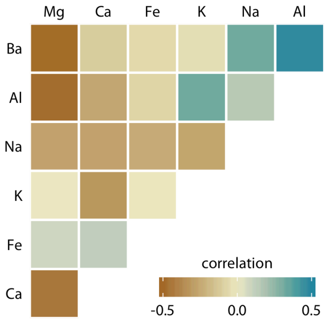
Correlograms





Correlograms

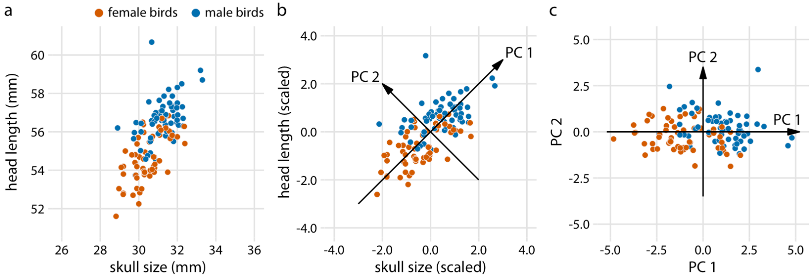




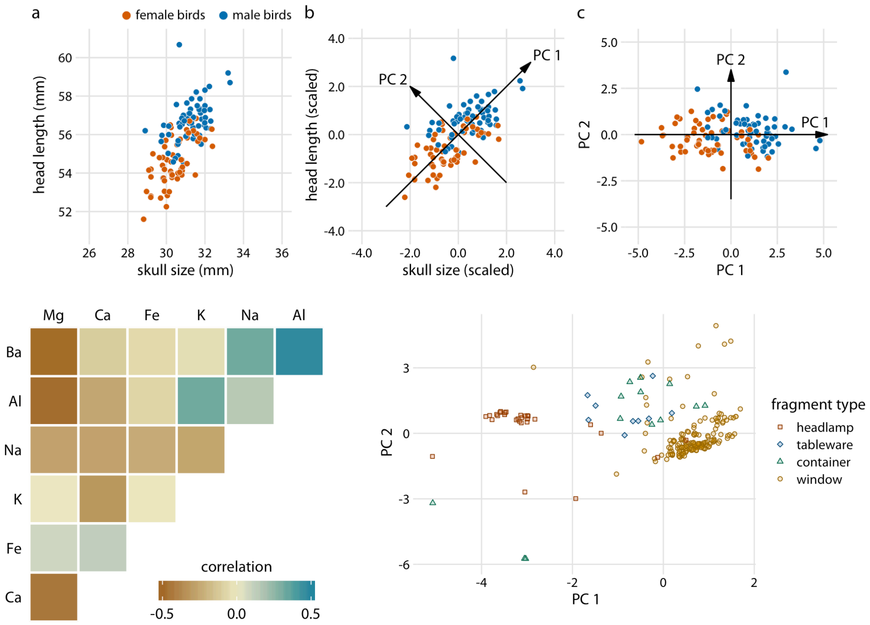
Non-Linear Dependence

Lack of linear correlation does not imply lack of dependence

Dimension reduction

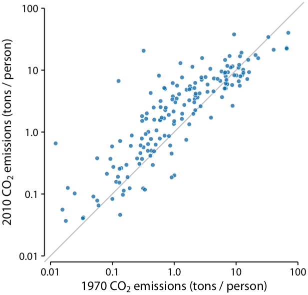


Dimension reduction



Paired data

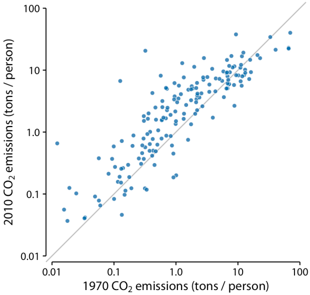
Scatterplots and slopegraphs are two main choices for plotting paired data.

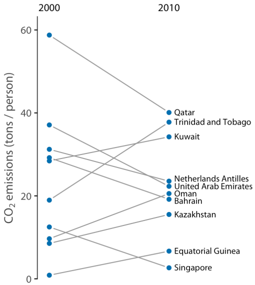


The last plot shows that slopegraph can accomodate short time series.

Paired data

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

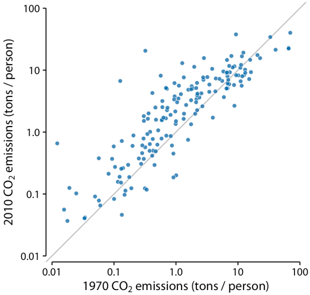


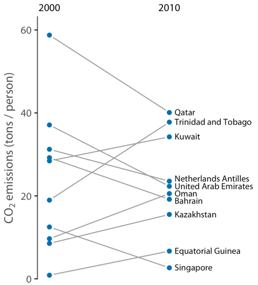


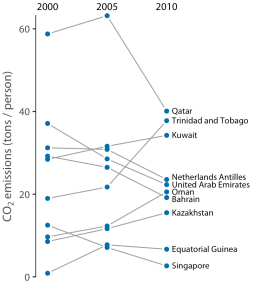
The last plot shows that slopegraph can accomodate short time series.

Paired data

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

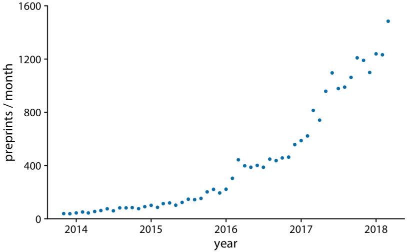




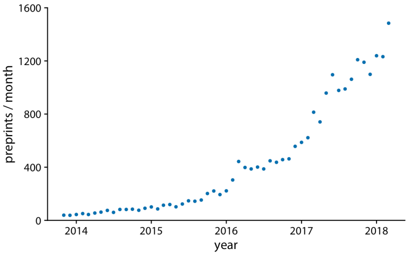


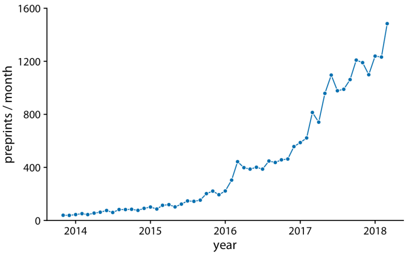
The last plot shows that slopegraph can accomodate short time series.

Visualizing time series — univariate

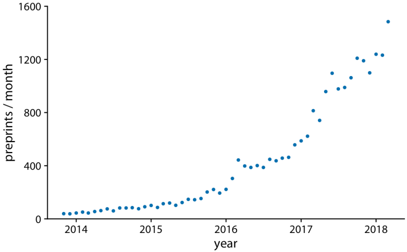


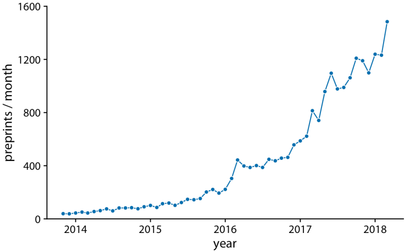
Visualizing time series — univariate

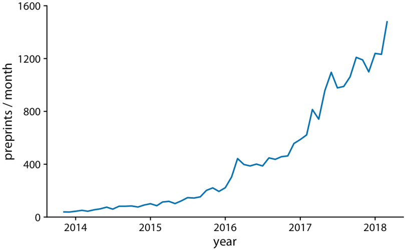




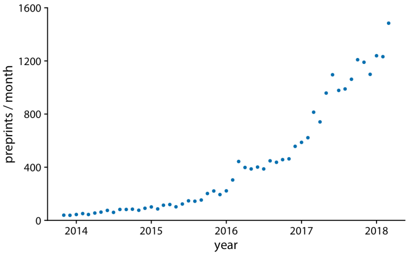
Visualizing time series — univariate

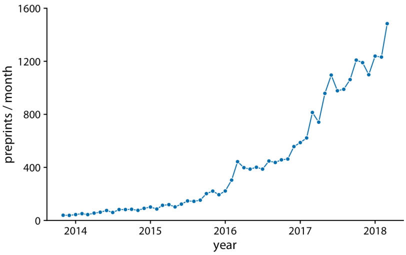


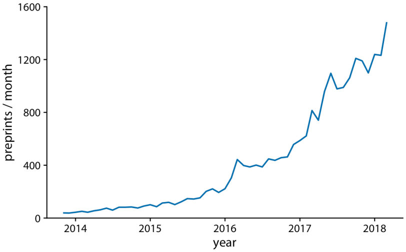




Visualizing time series — univariate

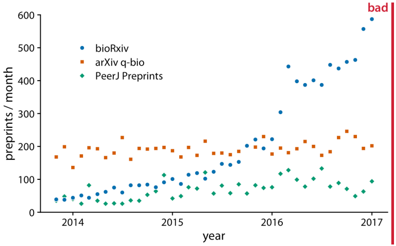




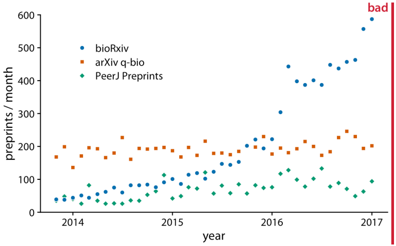


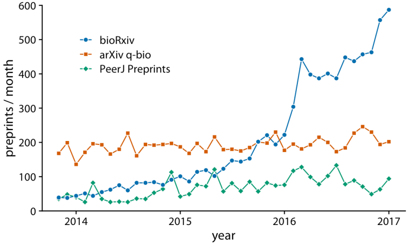
For dense time series, connect the dots and omit them.

Visualizing time series — multivariate

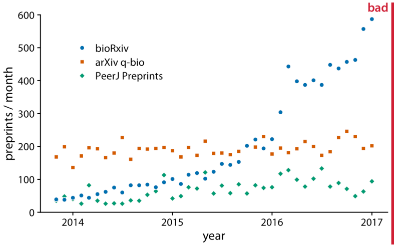


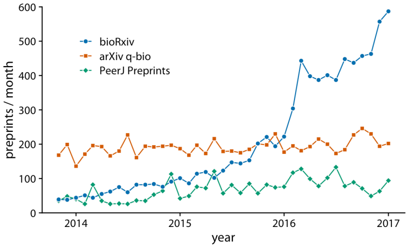
Visualizing time series — multivariate

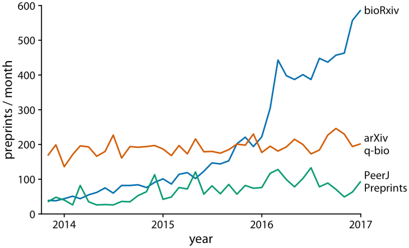




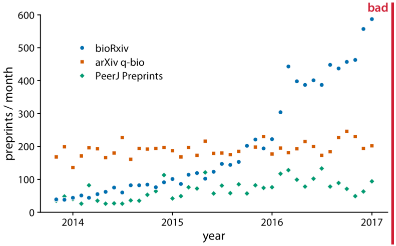
Visualizing time series — multivariate

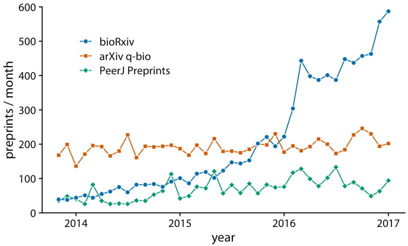


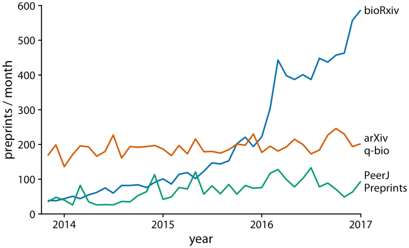




Visualizing time series — multivariate



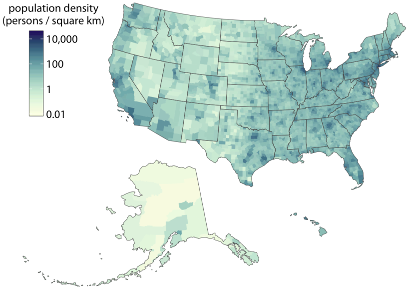




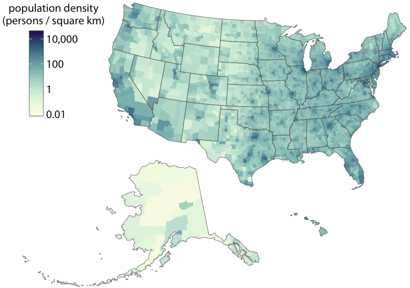
Consider replacing legends with direct labeling.

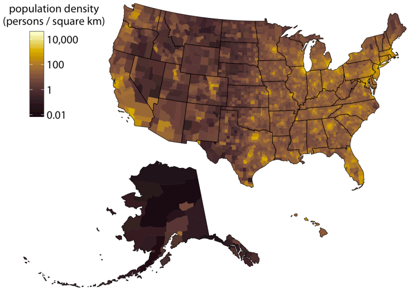
Make sure it is easy to compare objects of interest

Visualizing geospatial data

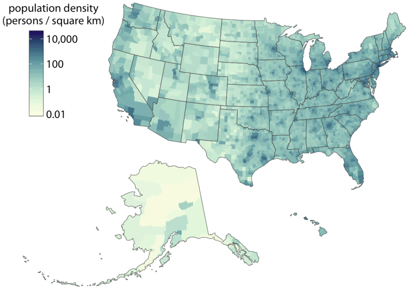


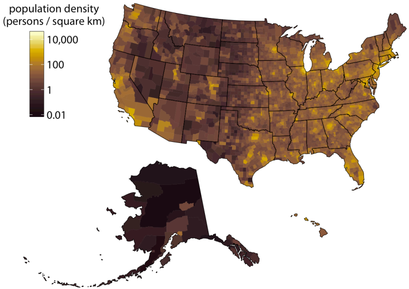
Visualizing geospatial data

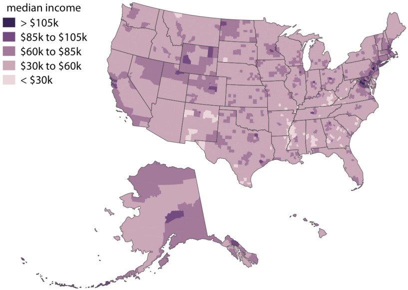




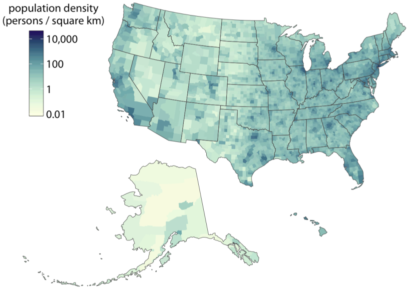
Visualizing geospatial data

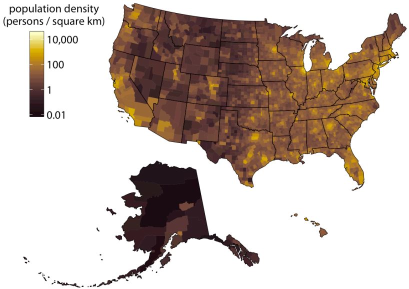


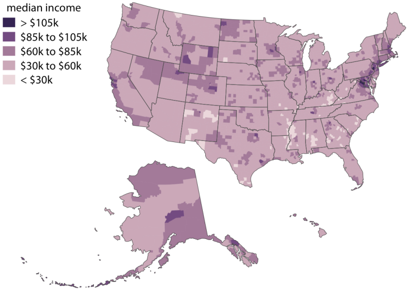


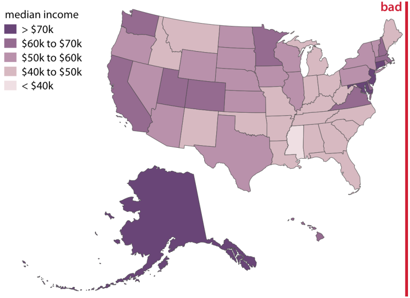


Visualizing geospatial data

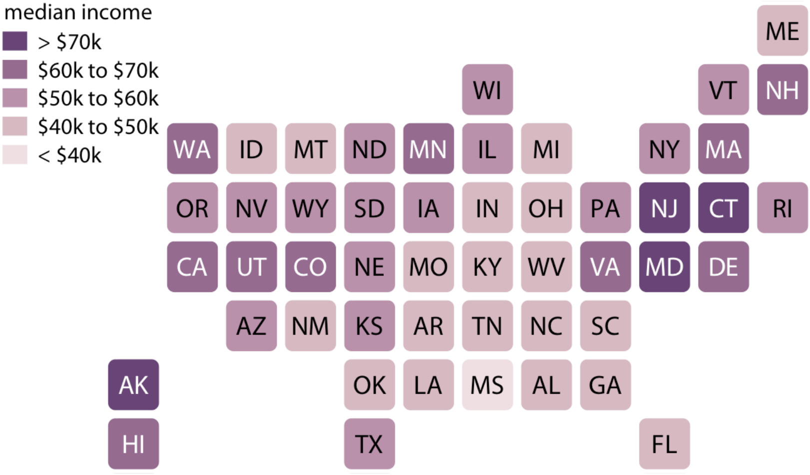




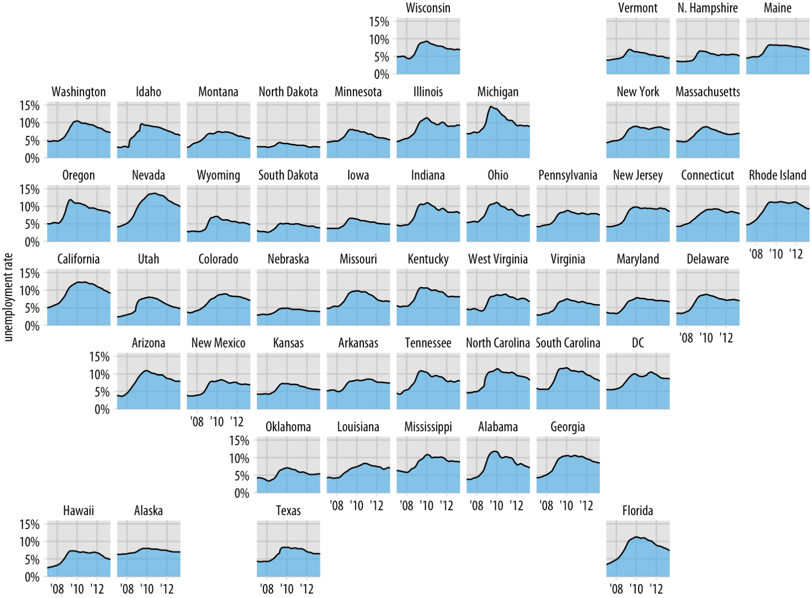




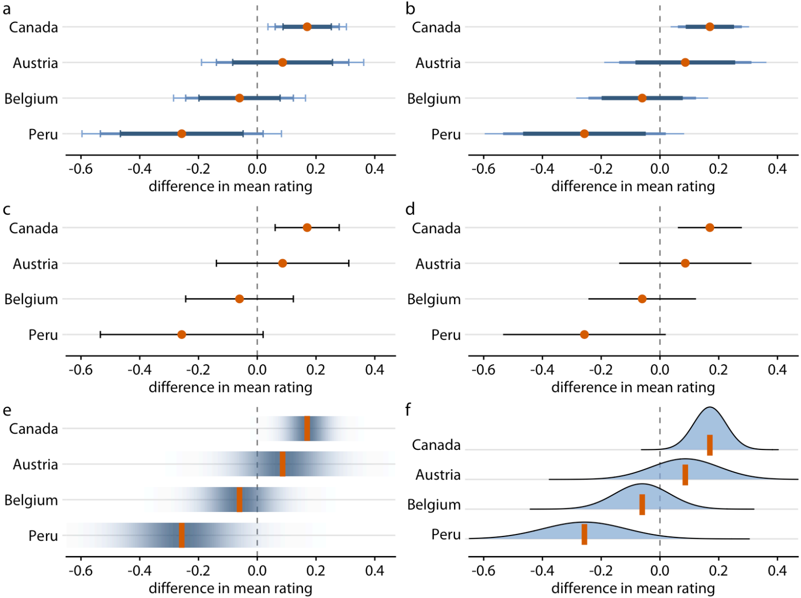
Visualizing geospatial data without maps



Visualizing geospatial data without maps

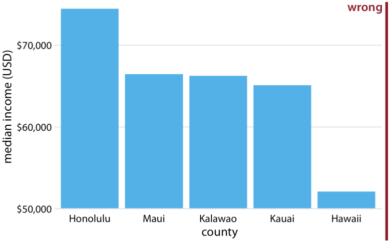


Visualizing the uncertainty of point estimates



The principle of proportional ink

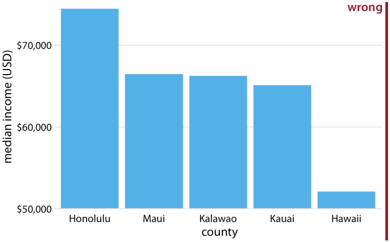
The principle of proportional ink: The sizes of shaded areas in a visualization need to be proportional to the data values they represent.

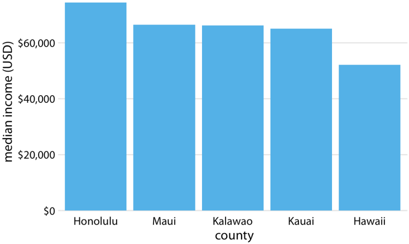


Bars on a linear scale must always start at 0.

The principle of proportional ink

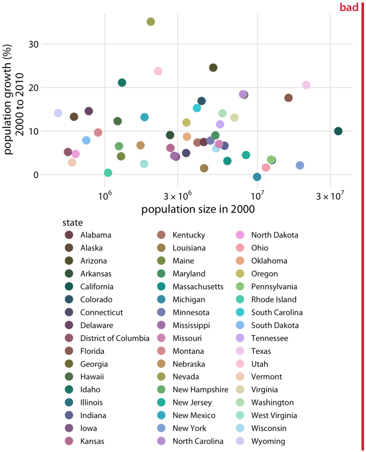
The principle of proportional ink: The sizes of shaded areas in a visualization need to be proportional to the data values they represent.



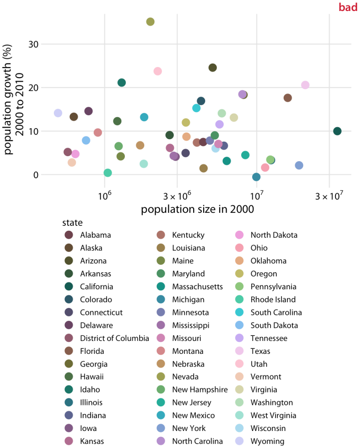


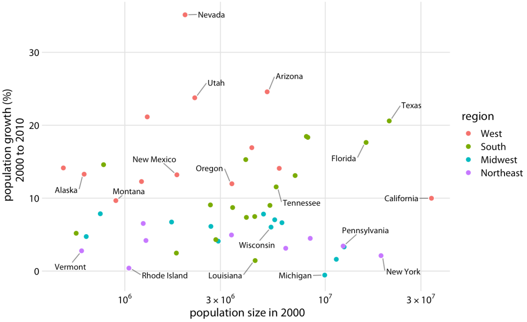
Bars on a linear scale must always start at 0.

Common pitfalls of color use

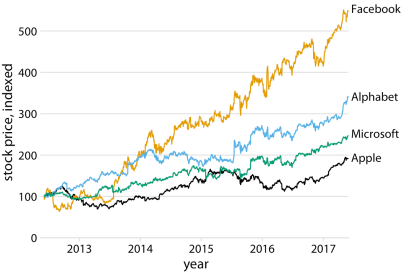


Common pitfalls of color use

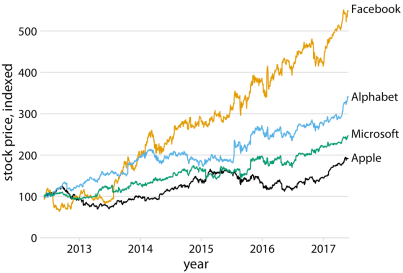


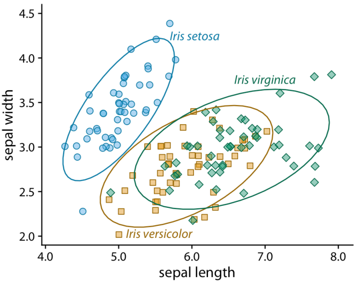


Figures without legends

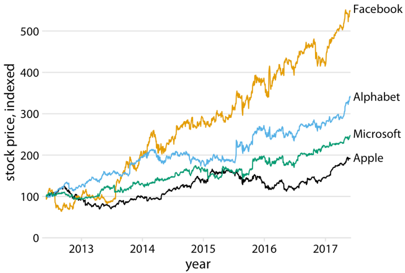


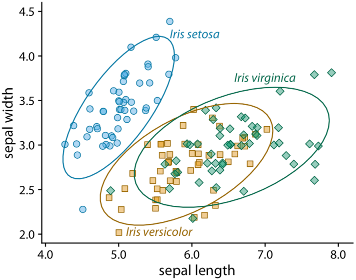
Figures without legends

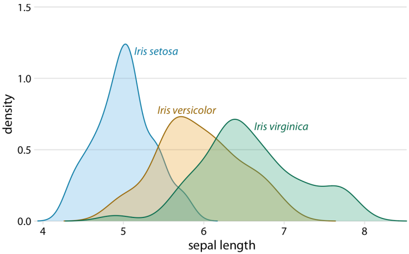




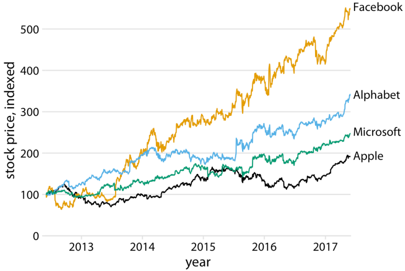
Figures without legends

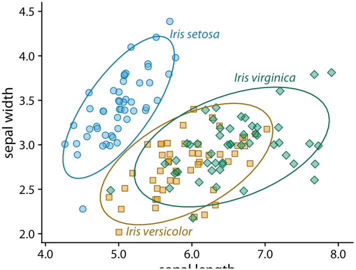


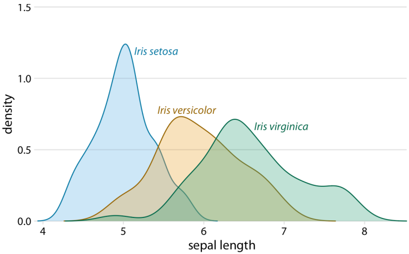


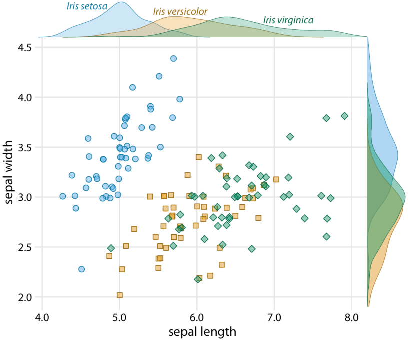


Figures without legends

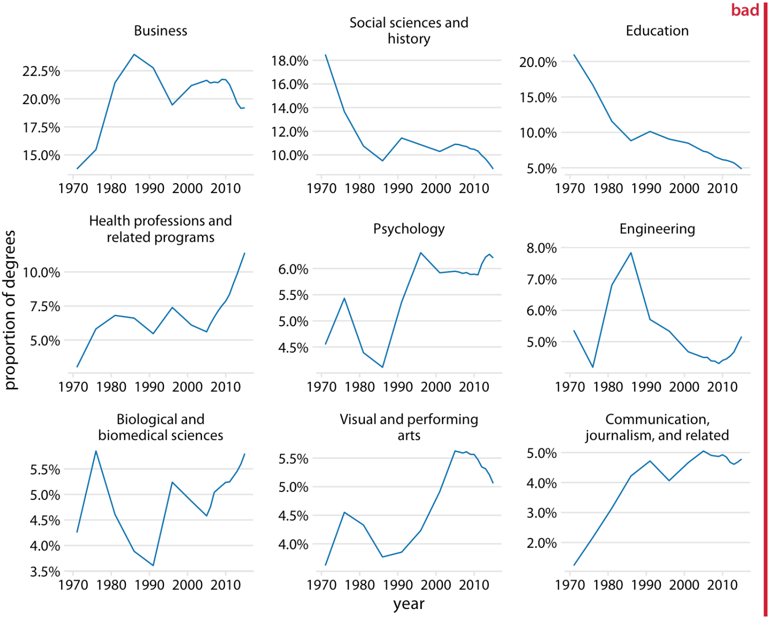




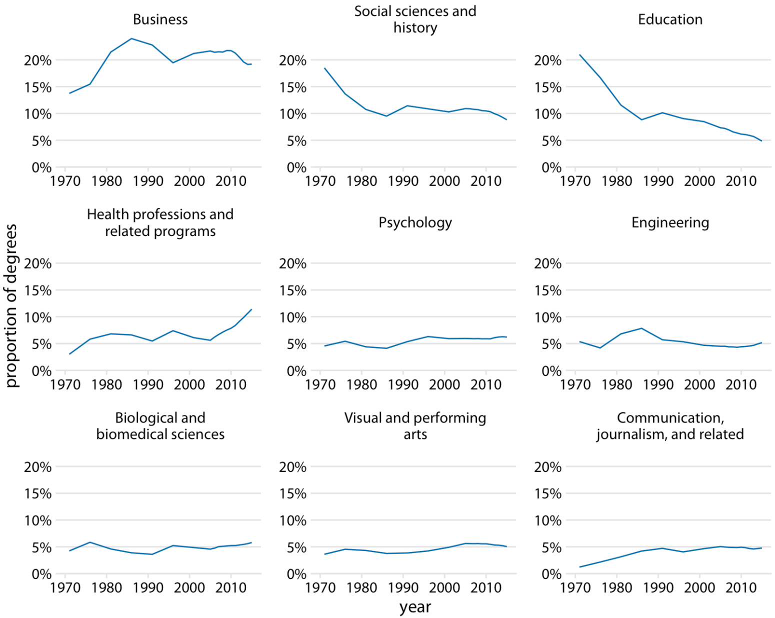




Multi-panel figures



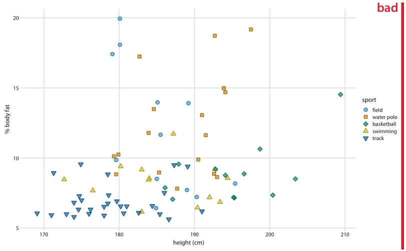
Multi-panel figures

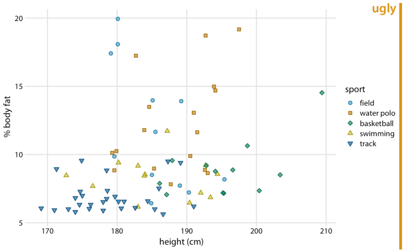


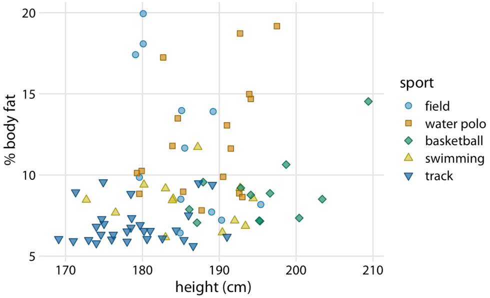
Titles and captions

- Always label your axes!
- Captions of figures and tables should be self-explanatory.

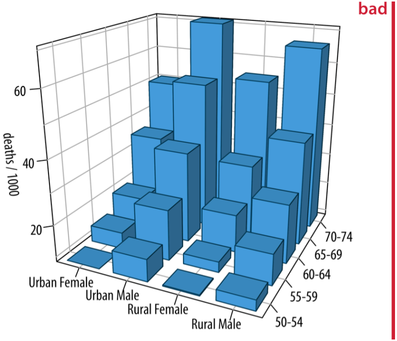
Your axis labels are too small



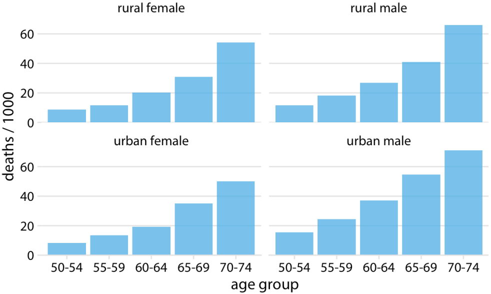




Don’t go 3D



bad
60
40
20
deaths / 1000
Urban Female
Urban Male
Rural Female
Rural Male
70-74
65-69
60-64
55-59
50-54



rural female
rural male
urban female
urban male
60
40
20
0
deaths / 1000
50-54
55-59
60-64
65-69
70-74
age group