Introduction to Data Science

Last Update: 4 October 2020

Chapter 6
Data Visualization

Papangkorn Inkeaw, PhD

Department of Computer Science, Faculty of Science

Chiang Mai University

Outline

Data Visualization

1. Data Visualization
2. Visualization for Numerical Data
3. Visualization for Categorical Data
4. Visualization for Time Series Data
5. Map and Network

Data Visualization

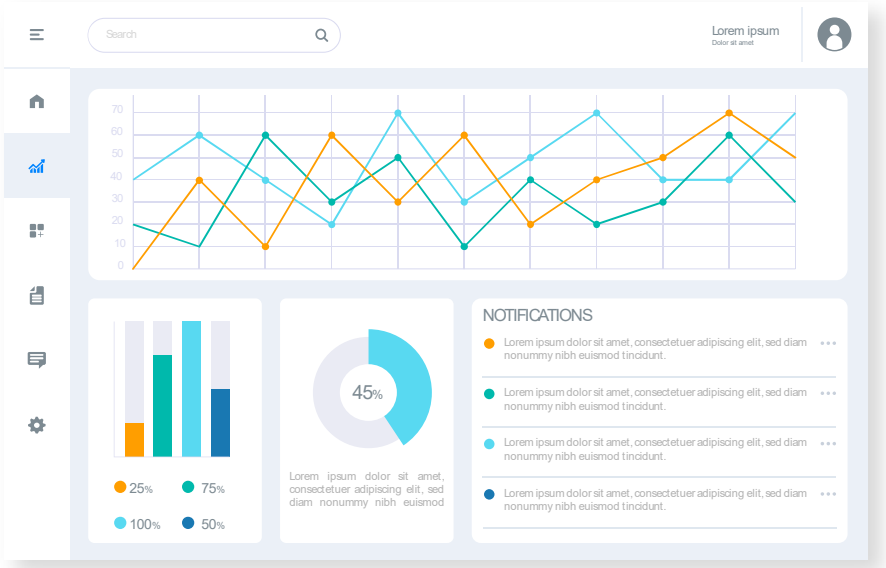
Data Visualization is:

- the graphic representation of data
- part art and part science
- a branch of descriptive statistics.

	Date	Rainfall	Mean temperature	Humidity
$\mathbf{x}_1$	1/1/1990	0	18	56
$\mathbf{x}_2$	2/1/1990	0	17.5	62
…				
$\mathbf{x}_n$	31/10/2019	10	29.5	72

Large Data Table

=



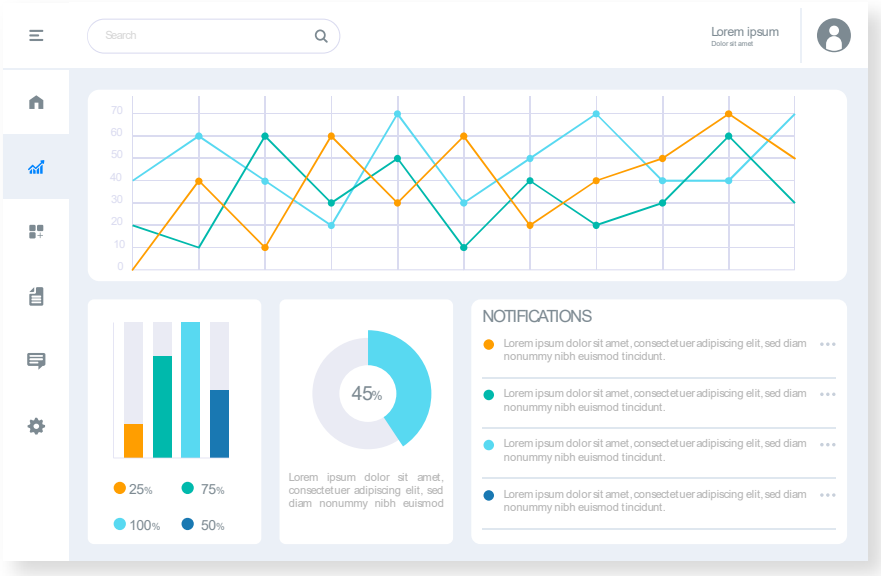
Graphs representing the data

Data Visualization

	Date	Rainfall	Mean temperature	Humidity
$\mathbf{x}_1$	1/1/1990	0	18	56
$\mathbf{x}_2$	2/1/1990	0	17.5	62
...				
$\mathbf{x}_n$	31/10/2019	10	29.5	72

Large Data Table

=



Graphs representing the data

- The challenge is to get **the art right without getting the science wrong**, and vice versa.
- A data visualization first and foremost has to accurately convey the data. It must not mislead or distort.

Data Visualization

Data visualization plays two key roles:

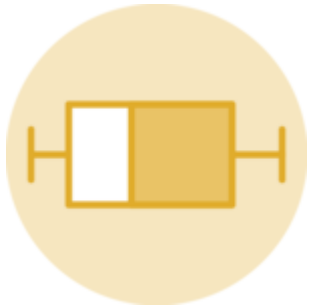
1. Communicating results clearly to a general audience.
2. Organizing a view of data that suggests a new hypothesis or a next step in a project.

Visualization for Numerical Data

Data Visualization

Histogram

Density



Boxplot

Violin plot

Ridgeline plot

Scatterplot

Bubble plot

Correlogram

Heatmap

Etc.

Source: https://www.data-to-viz.com/

Visualization for Numerical Data

Data Visualization

Histogram

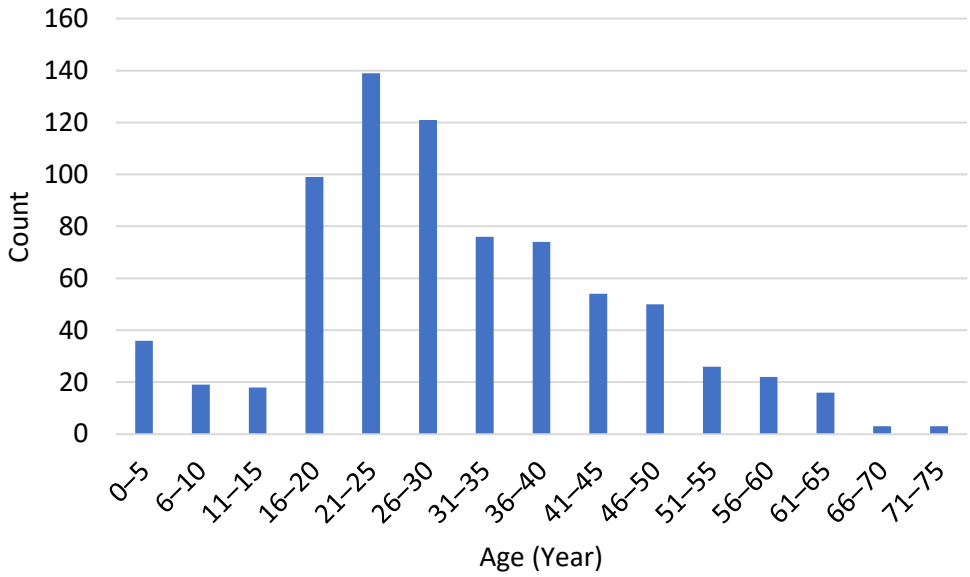
- A histogram takes as input a **numeric variable** only.
- The variable is cut into several bins
- The number of observation per bin is represented by the height of the bar.

	Age	…	…	…
$\mathbf{x}_1$	35			
$\mathbf{x}_2$	25			
…	…			
$\mathbf{x}_n$	62			

$Age \in \{0,75\}$

Data

Bin of Age	Count
0-5	36
6-10	19
11-15	18
…	…
61-65	16
66-70	3
71-75	3



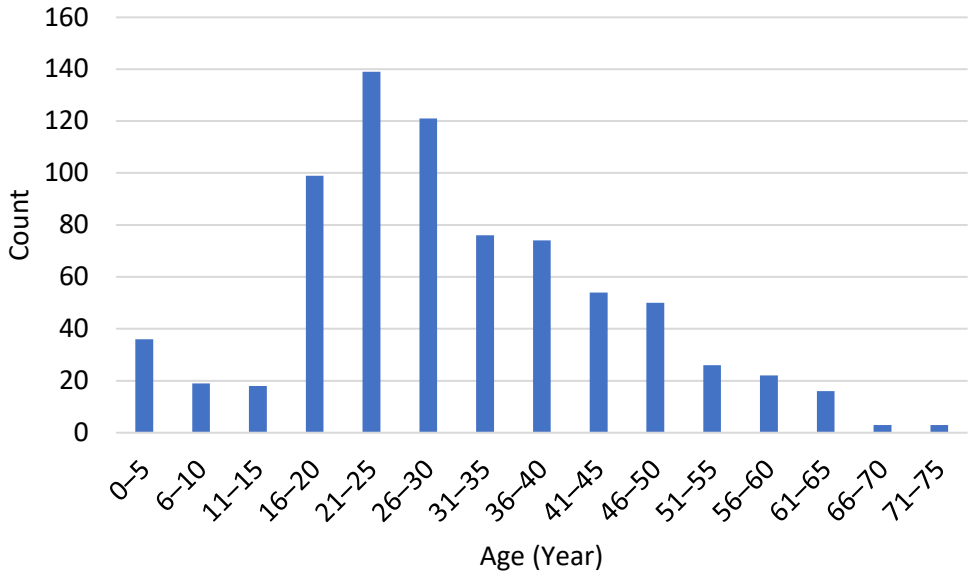
Histogram

Visualization for Numerical Data

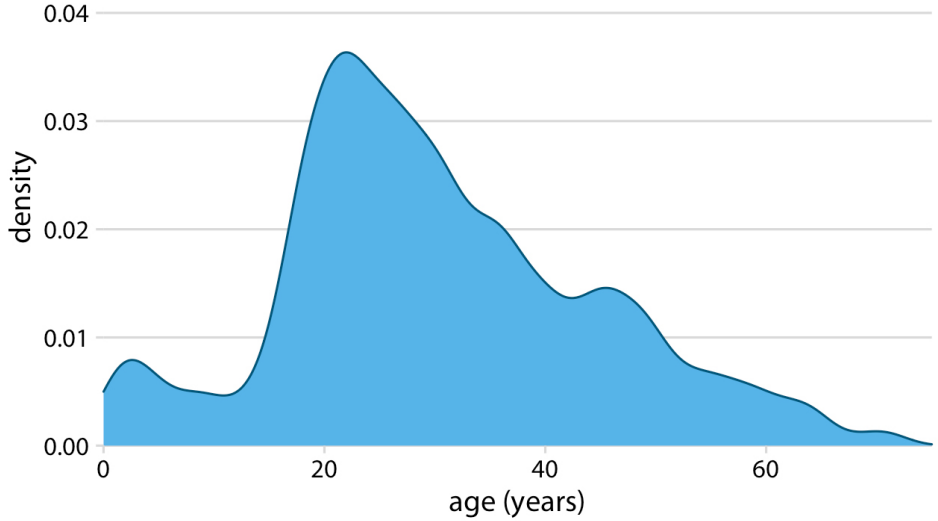
Data Visualization

Density

- Visualize the underlying probability distribution of the data by drawing an appropriate **continuous curve**.
- This curve needs to be estimated from the data using **kernel density estimation**.



Histogram



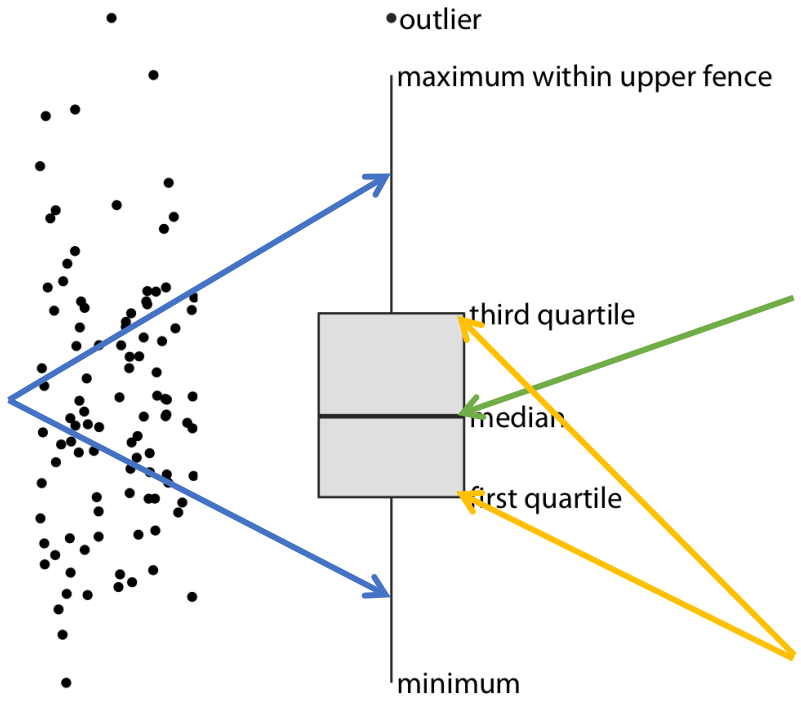
Density plot

Visualization for Numerical Data

Data Visualization

Boxplot

Boxplot gives a nice summary of one or several distributions.



Extreme lines show the highest and lowest value excluding outliers

Line that divides the box into 2 parts represents the median of the data

The end of the box shows the third and first quartiles.

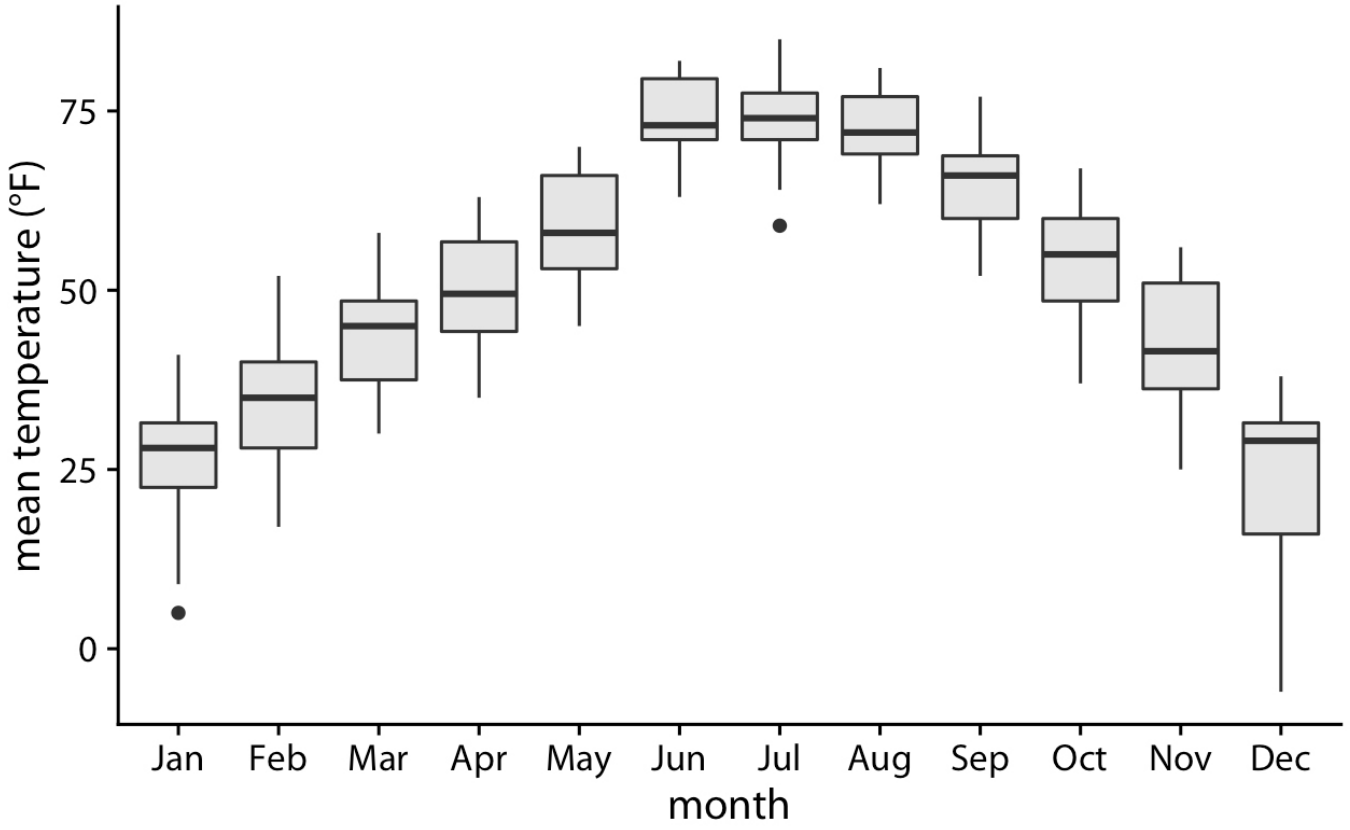
Anatomy of a boxplot

Image Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Numerical Data

Data Visualization

Boxplot



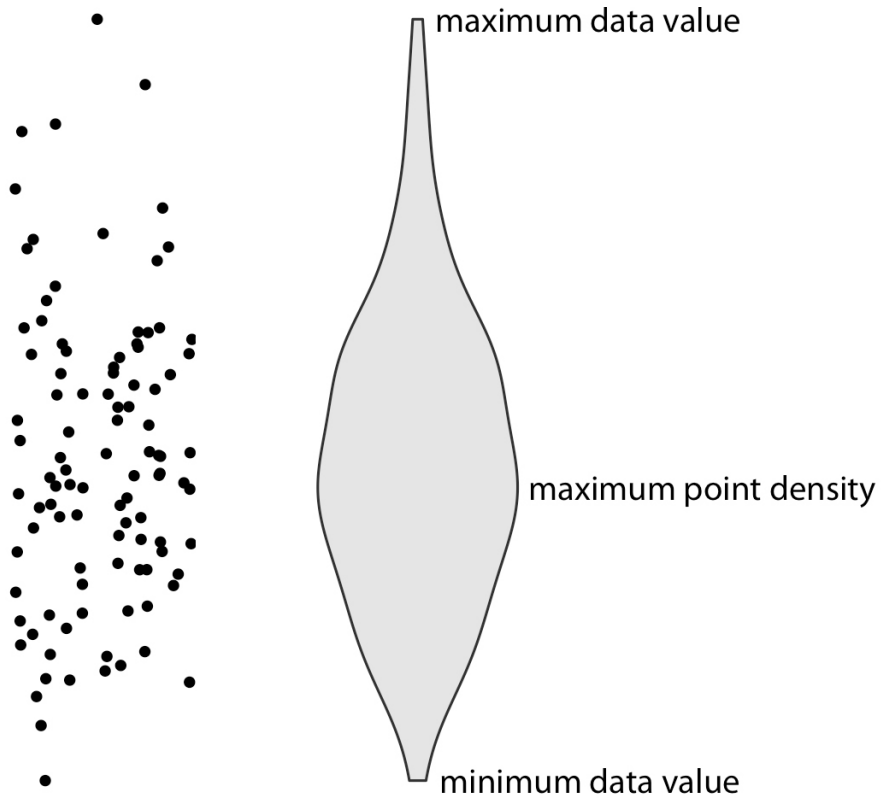
Mean daily temperatures in Lincoln, NE, visualized as boxplots.

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Numerical Data

Data Visualization

Violin Plot



Anatomy of a violin plot

Image Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

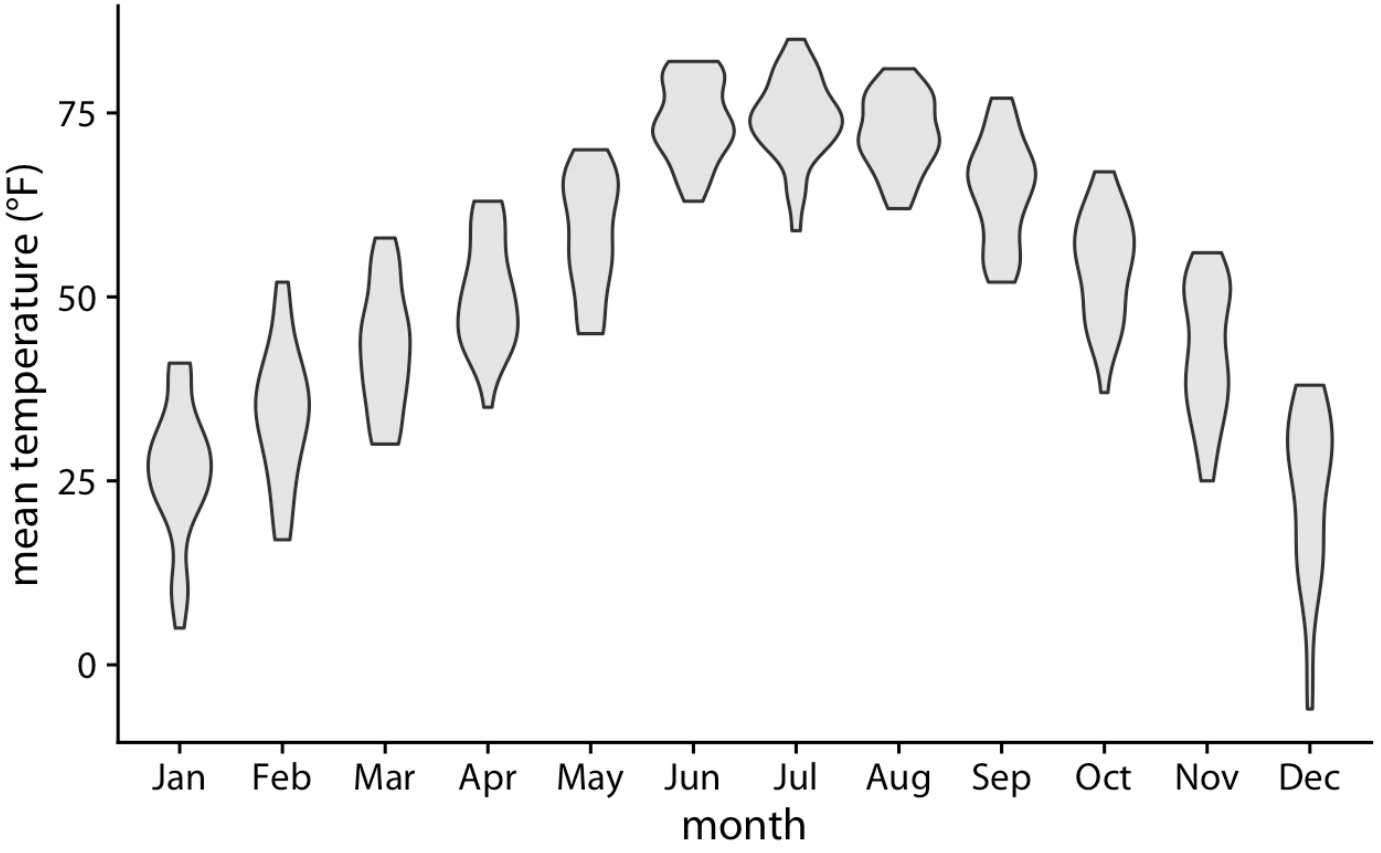
- Visualize the distribution of a numeric variable for one or several groups.
- It is really close from a **boxplot** but allows a deeper understanding of the distribution.

- Violins are particularly adapted when
  - the amount of data is huge
  - showing individual observations gets impossible.

Visualization for Numerical Data

Data Visualization

Violin Plot



Mean daily temperatures in Lincoln, NE, visualized as violin plot.

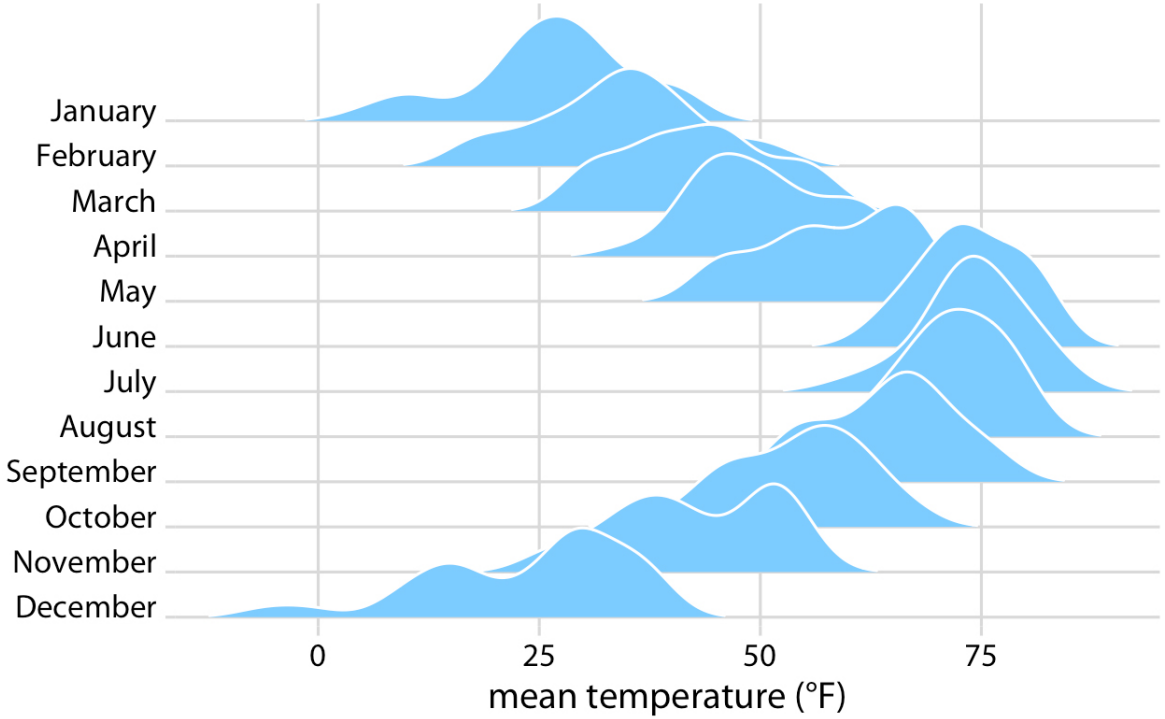
Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Numerical Data

Data Visualization

Ridgeline Plot

- The distribution of a numerical value for several groups.
- Distribution can be represented using histograms or density plots, all aligned to the same horizontal scale and presented with a slight overlap.



Temperatures in Lincoln, NE, in 2016, visualized as a ridgeline plot. For each month, we show the distribution of daily mean temperatures measured in Fahrenheit.

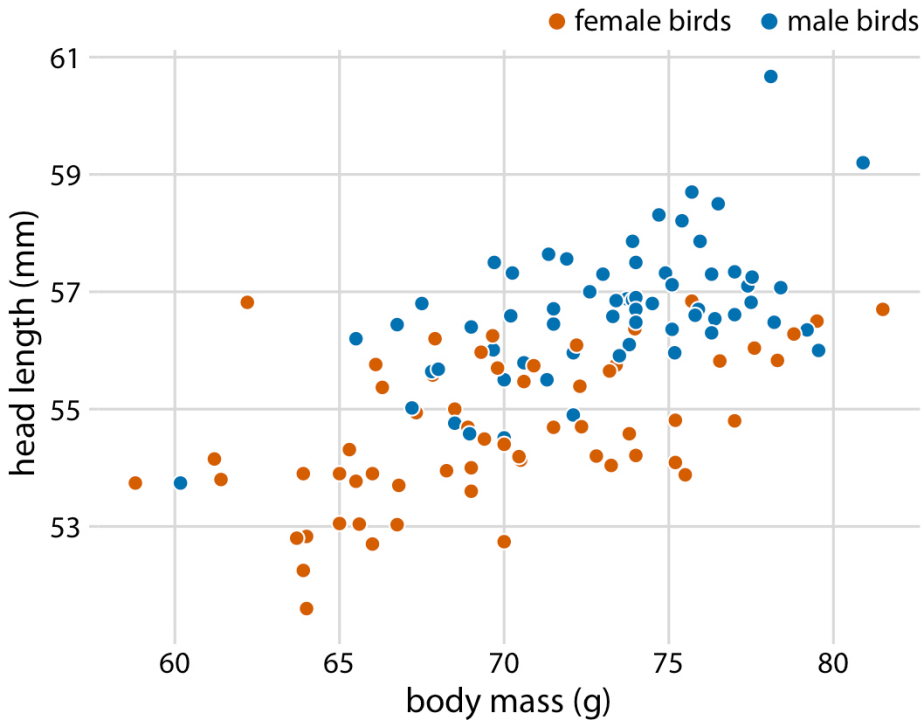
Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Numerical Data

Data Visualization

Scatterplot

- Displays the relationship between 2 numeric variables.
- For each data point, the value of its first variable is represented on the X axis, the second on the Y axis.



Head length versus body mass for 123 blue jays. The birds' sex is indicated by color. At the same body mass, male birds tend to have longer heads (and specifically, longer bills) than female birds.

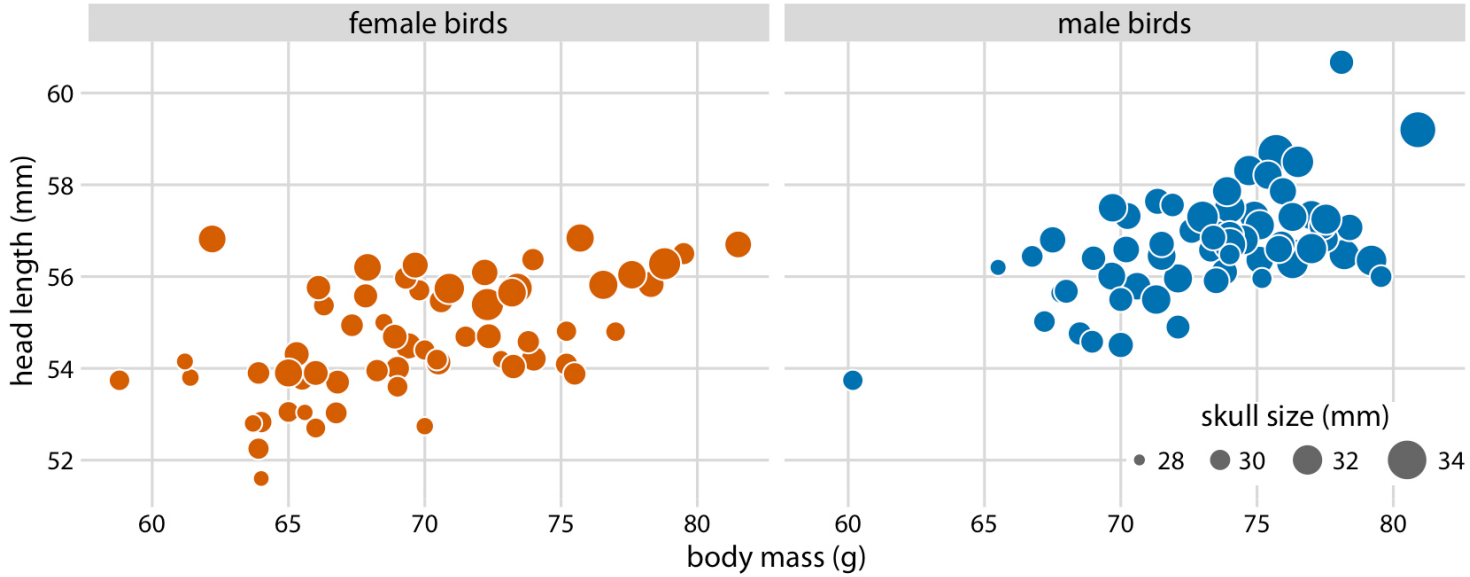
Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Numerical Data

Data Visualization

Bubble plot

- A scatterplot where a third dimension is added: the value of an additional numeric variable is represented through the size of the dots.



Head length versus body mass for 123 blue jays. The birds' sex is indicated by color and the birds' skull size by symbol size. Head length measurements include the length of the bill while skull size measurements do not. Head length and skull size tend to be correlated, but there are some birds with unusually long or short bills given their skull size.

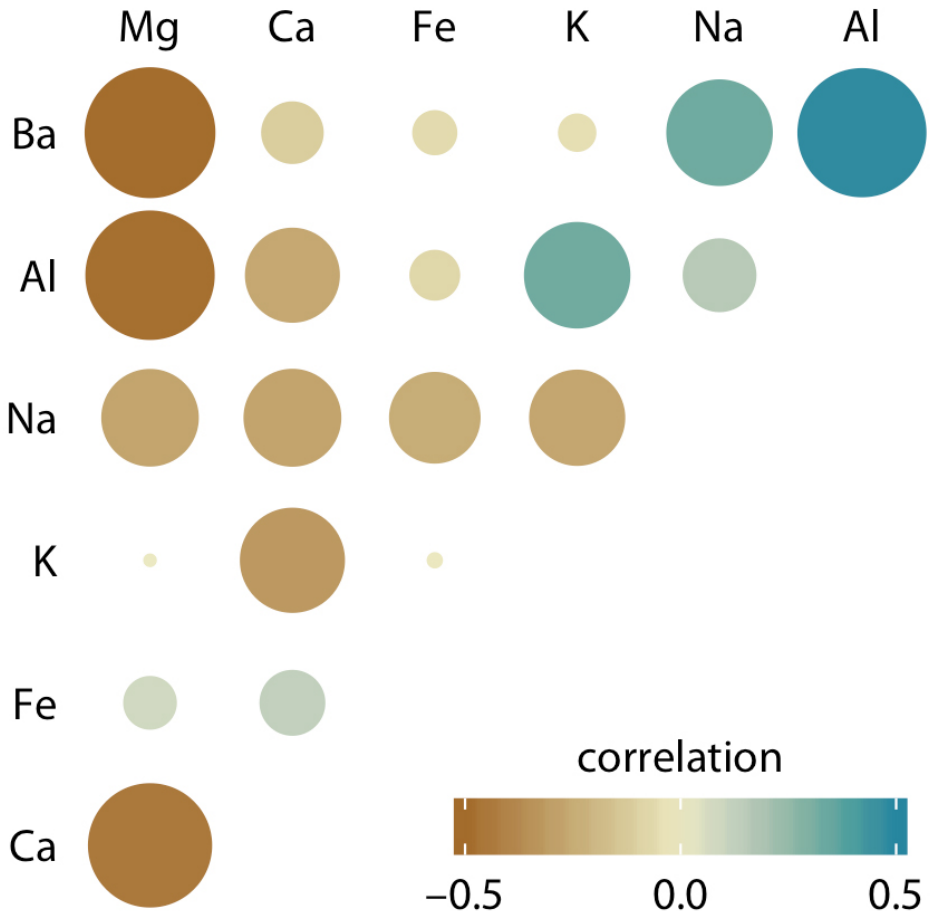
Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Numerical Data

Data Visualization

Correlogram

- **Correlation matrix** allows to analyze the relationship between each pair of numeric variables of a matrix.
- The correlation between each pair of variable is visualized through a scatterplot, or a symbol that represents the correlation.



Correlations in mineral content for forensic glass samples. The magnitude of each correlation is encoded in the size of the colored circles and color scale.

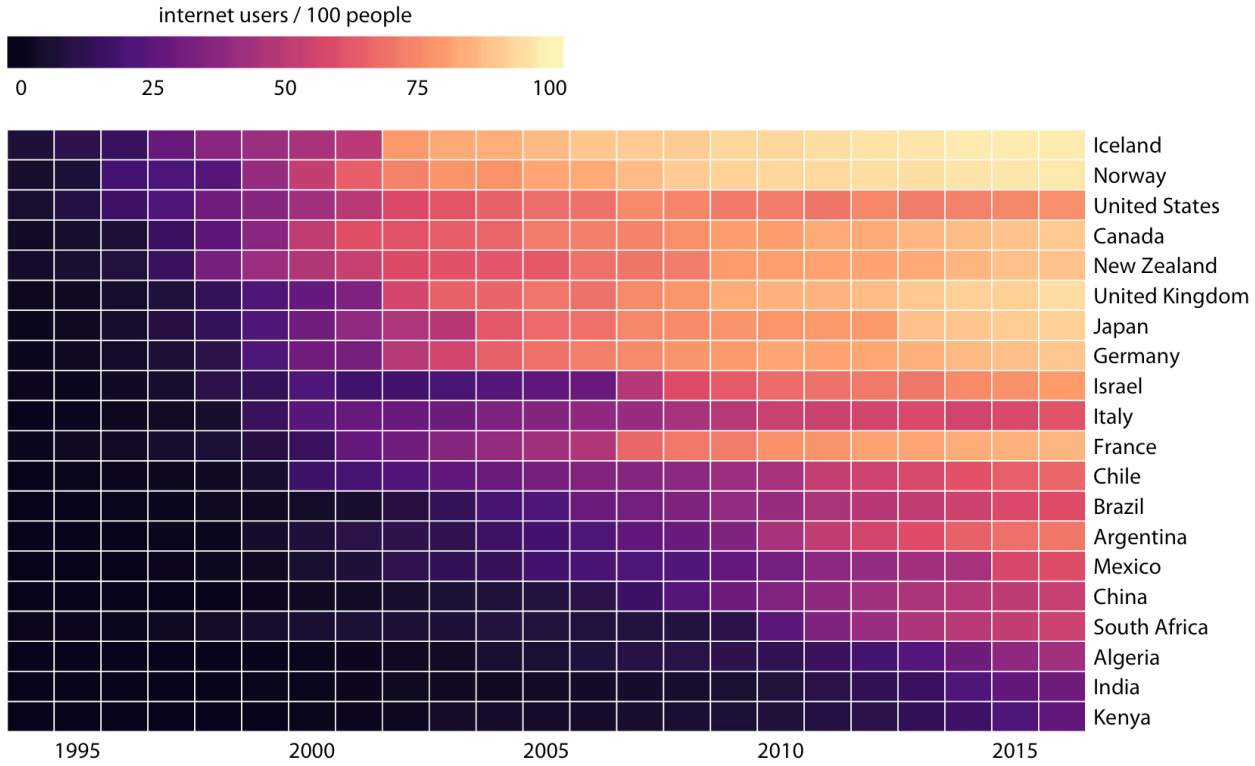
Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Numerical Data

Data Visualization

Heatmap

- Representation of data where the individual values contained in a matrix are represented as colors.
- It is really useful to display a general view of numerical data, not to extract specific data point.



Internet adoption over time, for select countries. Countries were ordered by the year in which their internet usage first exceeded 20%.

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Categorical Data

Data Visualization

Bar plot

Pie

Tree map

Parallel sets plot

Etc.

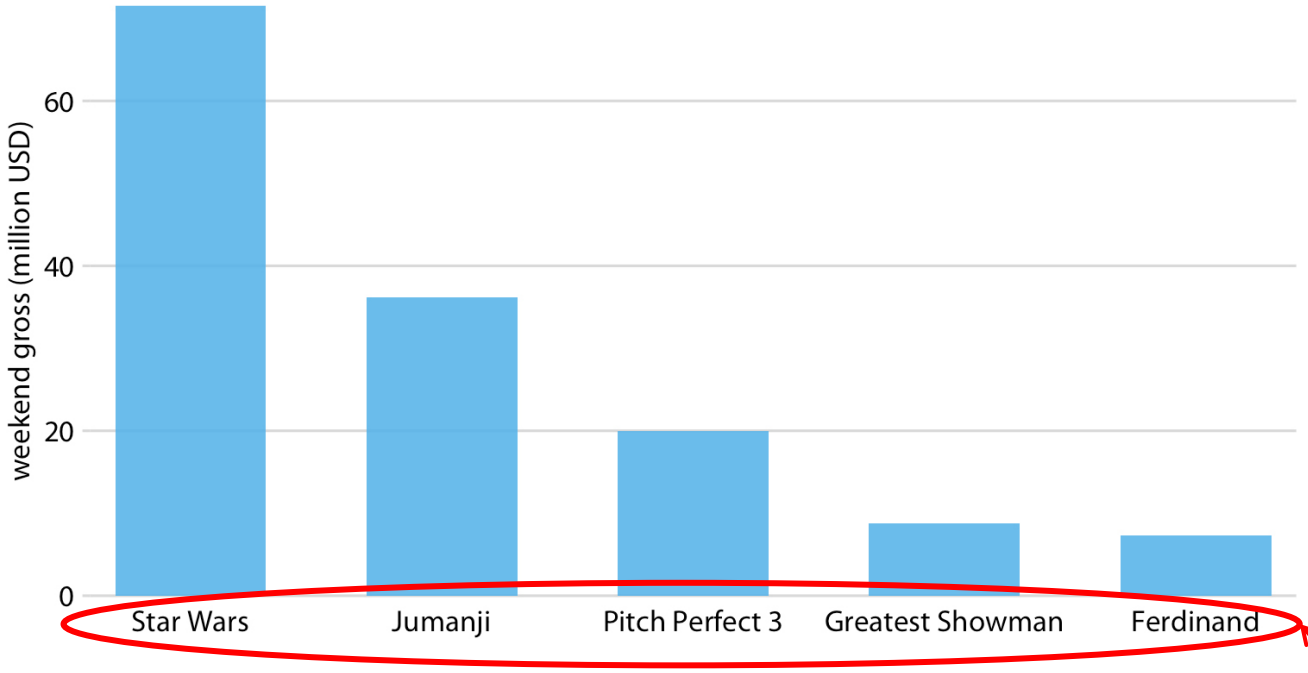
Source: https://www.data-to-viz.com/

Visualization for Categorical Data

Data Visualization

Bar Plot

- Each entity of the categoric variable is represented as a bar.
- The size of the bar represents its numeric value.



Highest-grossing movies for the weekend of December 22–24, 2017, displayed as a bar plot.

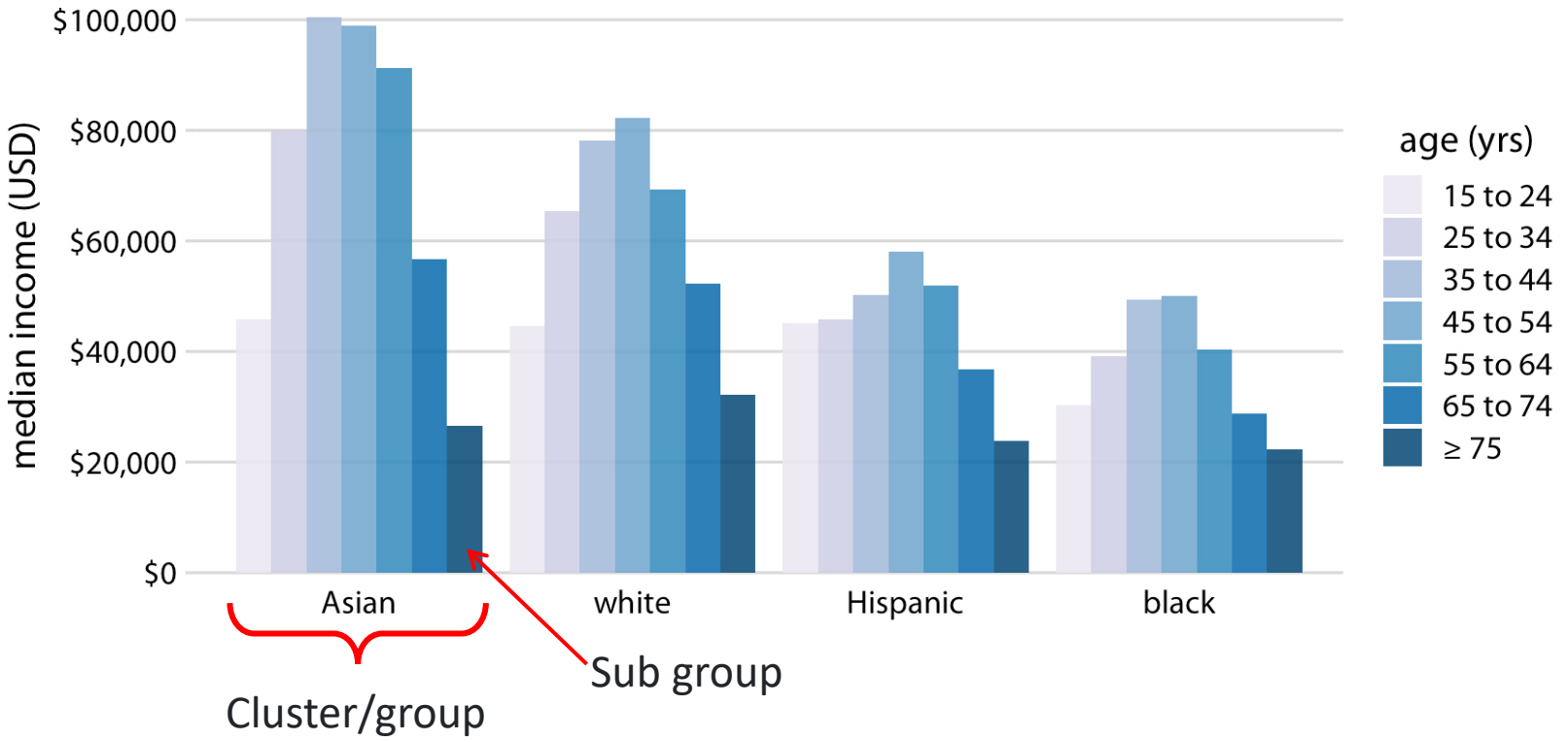
Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Categorical Data

Data Visualization

Different Types of Bar Plots

Clustered Bar Plot

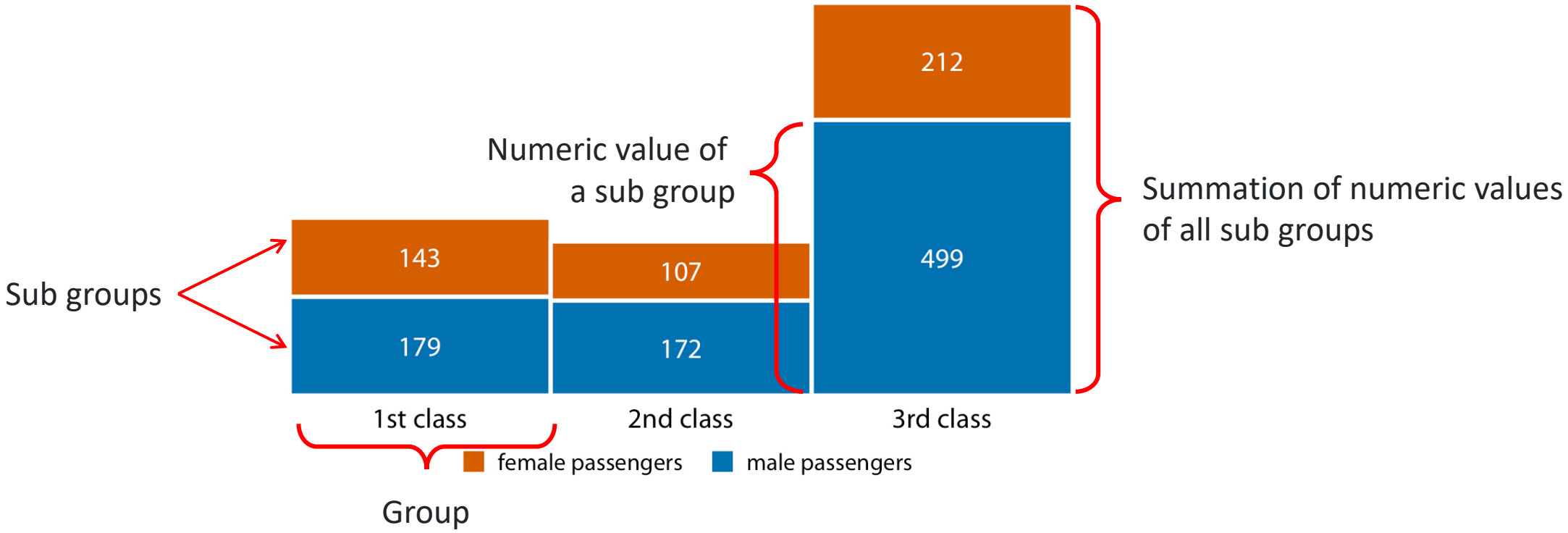


Visualization for Categorical Data

Data Visualization

Different Types of Bar Plots

Stack Bar Plot

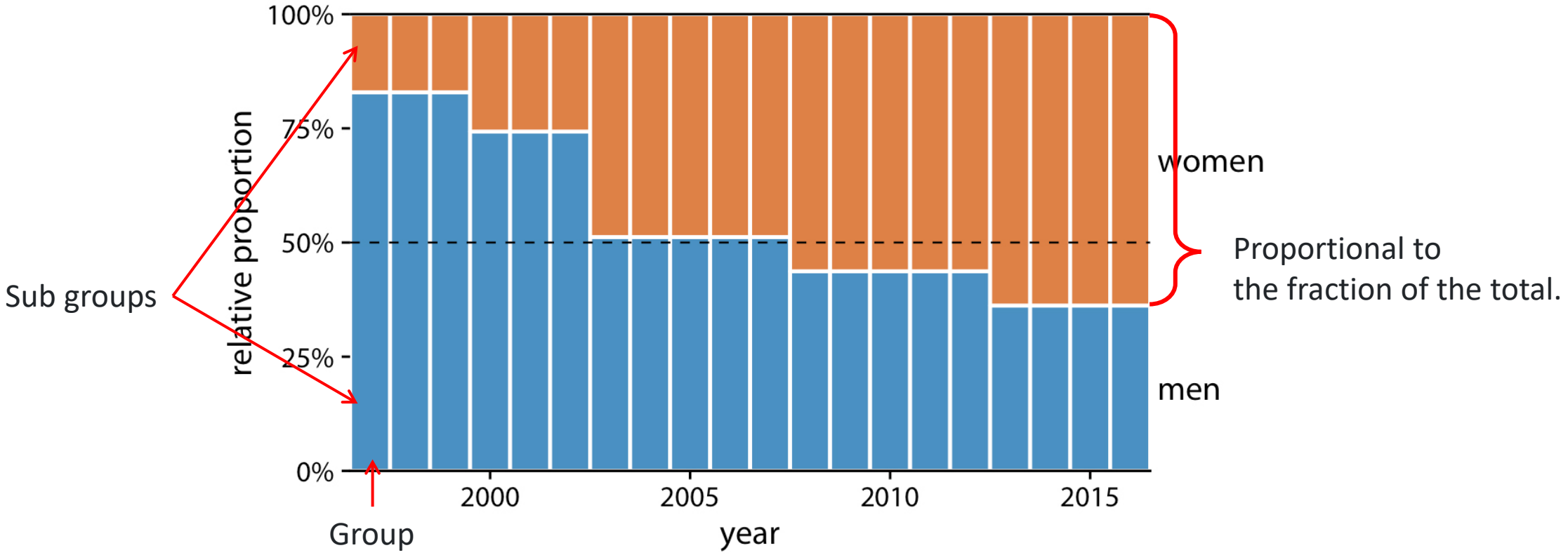


Visualization for Categorical Data

Data Visualization

Different Types of Bar Plots

100% Stack Bar Plot

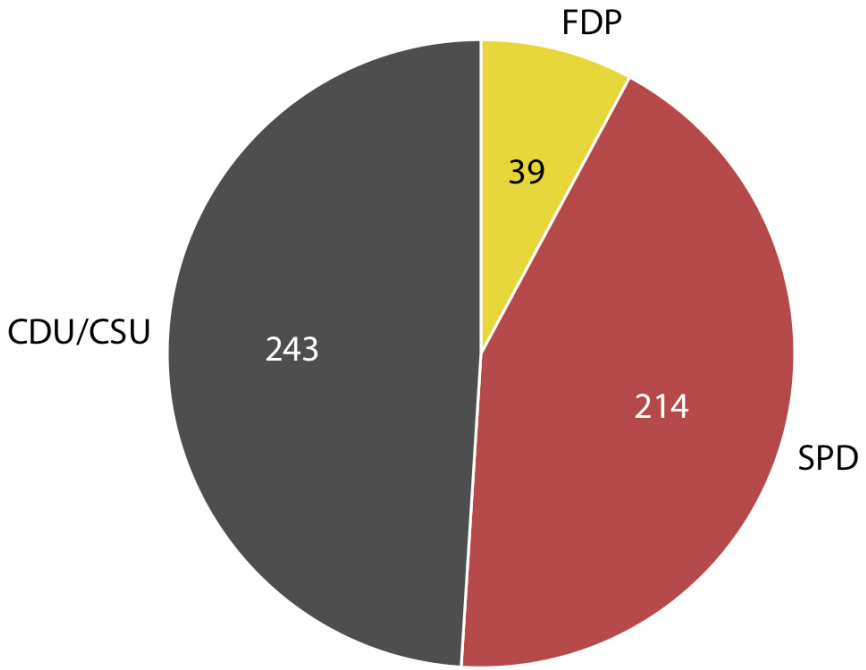


Visualization for Categorical Data

Data Visualization

Pie

- A circle divided into sectors that each represent a proportion of the whole.
- It is often used to show proportion, where the sum of the sectors equal 100%.



Party composition of the eighth German Bundestag, 1976–1980, visualized as a pie chart. This visualization highlights that the ruling coalition of SPD and FDP had a small majority over the opposition CDU/CSU.

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

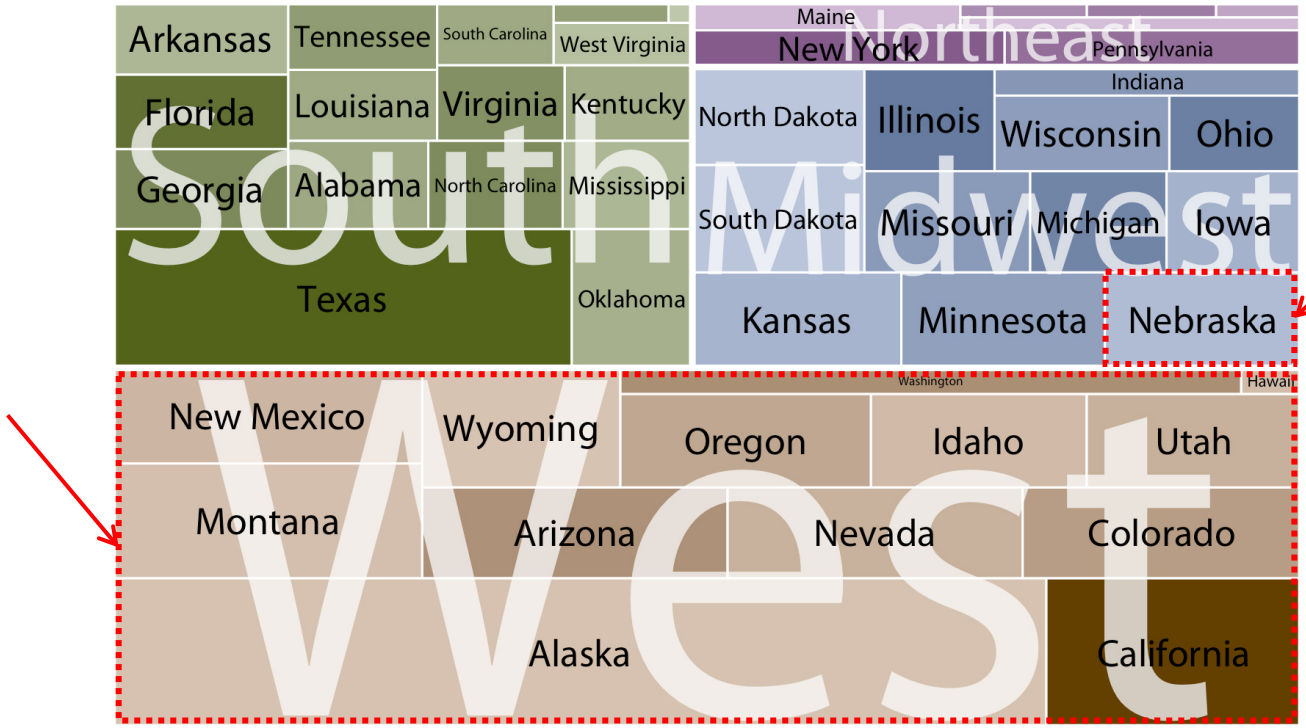
Visualization for Categorical Data

Data Visualization

Tree map

Displays hierarchical data as a set of nested rectangles.

Use color schemes to represent several dimensions: groups, subgroups etc.



Each group is represented by a rectangle, which area is proportional to its value.

States in the US visualized as a tree map. Each rectangle represents one state, and the area of each rectangle is proportional to the state's land surface area. The states are grouped into four regions, West, Northeast, Midwest, and South. The coloring is proportional to the number of inhabitants for each state, with darker colors representing larger numbers of inhabitants.

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Categorical Data

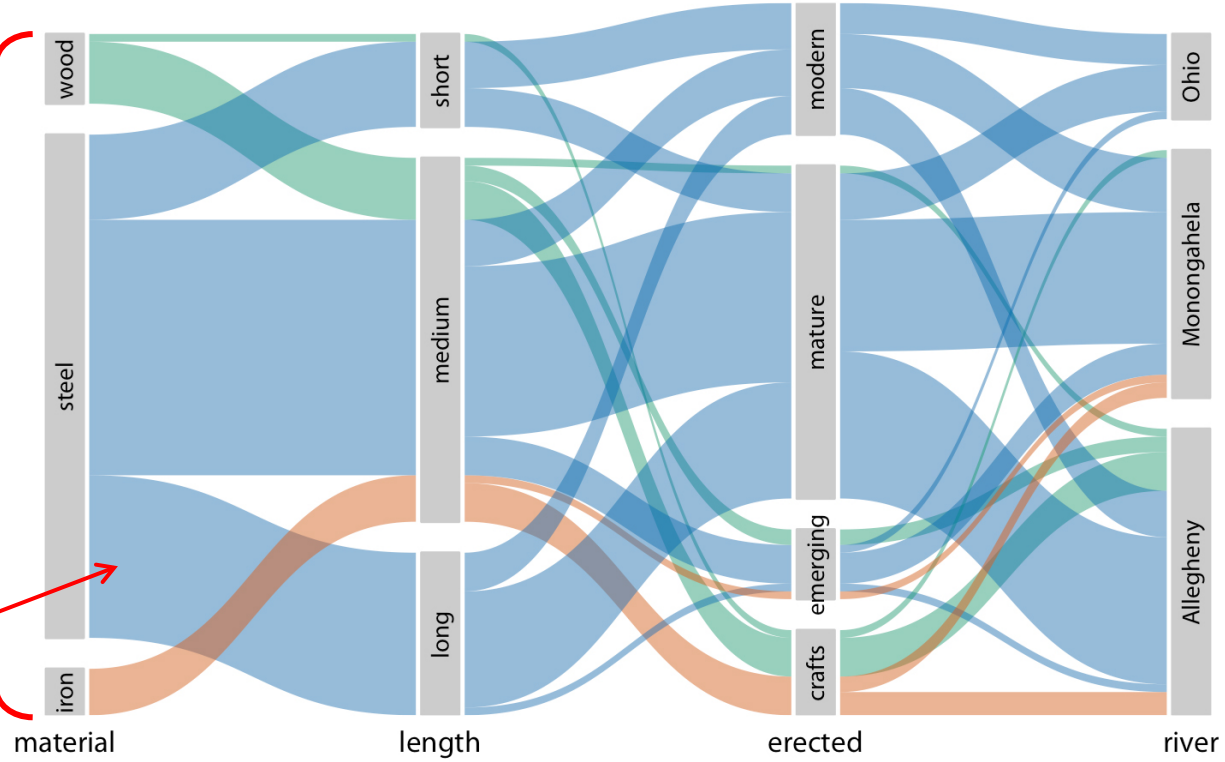
Data Visualization

Parallel sets plot

- Show how the total dataset breaks down by each individual categorical variable,
- Draw shaded bands that show how the subgroups relate to each other

All possible values of a categorical variable

shaded bands show how the subgroups relate to each other



Categorical Variables

Breakdown of bridges in Pittsburgh by construction material, length, era of construction, and the river they span, shown as a parallel sets plot. The coloring of the bands highlights the construction material of the different bridges.

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

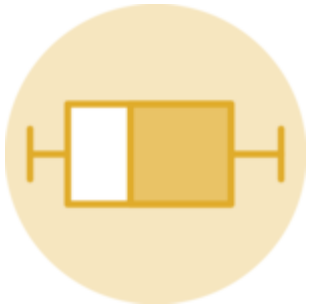
Visualization for Time Series Data

Data Visualization

Line chart

Area chart

We can use the following techniques to visualize time series data by representing time on X or Y axis.



Boxplot

Violin plot

Ridgeline plot

Bar plot

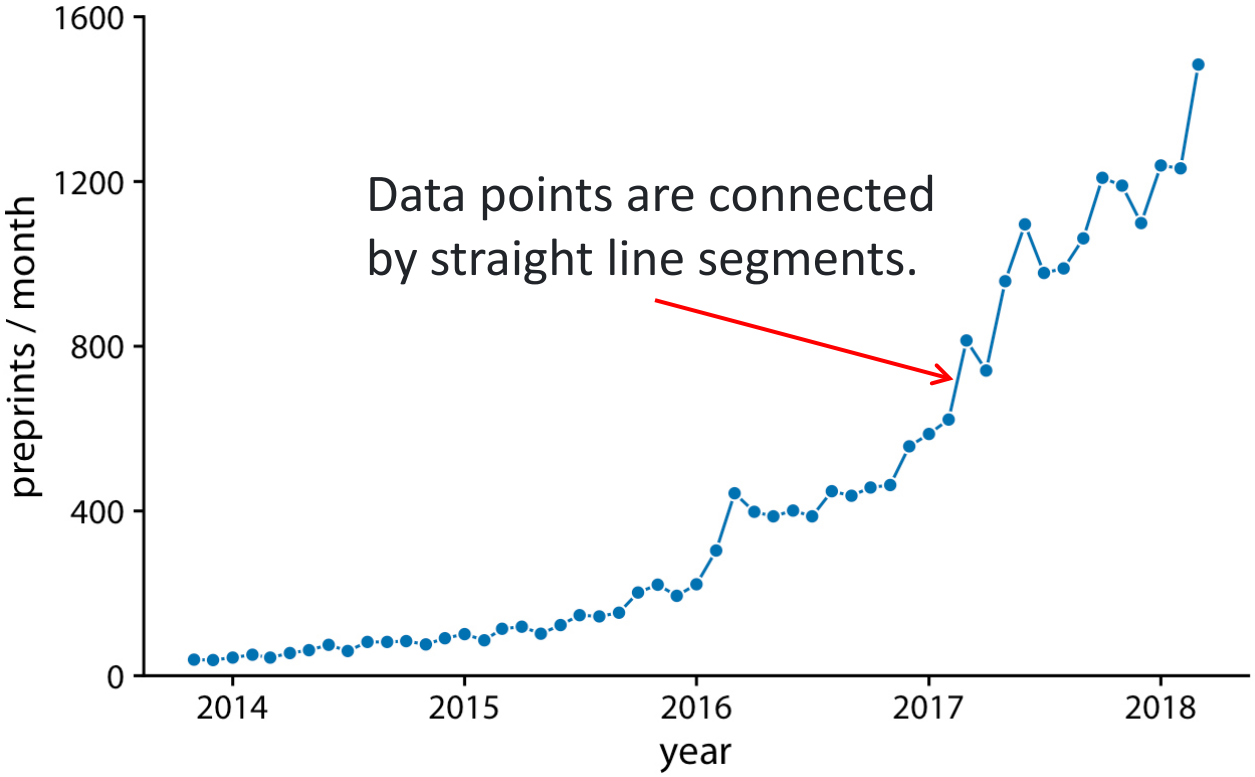
Source: https://www.data-to-viz.com/

Visualization for Time Series Data

Data Visualization

Line chart

- Display the evolution of one or several **numeric variables**.
- The measurement points are ordered.
- A line chart is often used to visualize a trend in data over intervals of time.



Monthly submissions to the preprint server bioRxiv, shown as dots connected by lines. The lines do not represent data and are only meant as a guide to the eye. By connecting the individual dots with lines, we emphasize that there is an order between the dots: each dot has exactly one neighbor that comes before it and one that comes after.

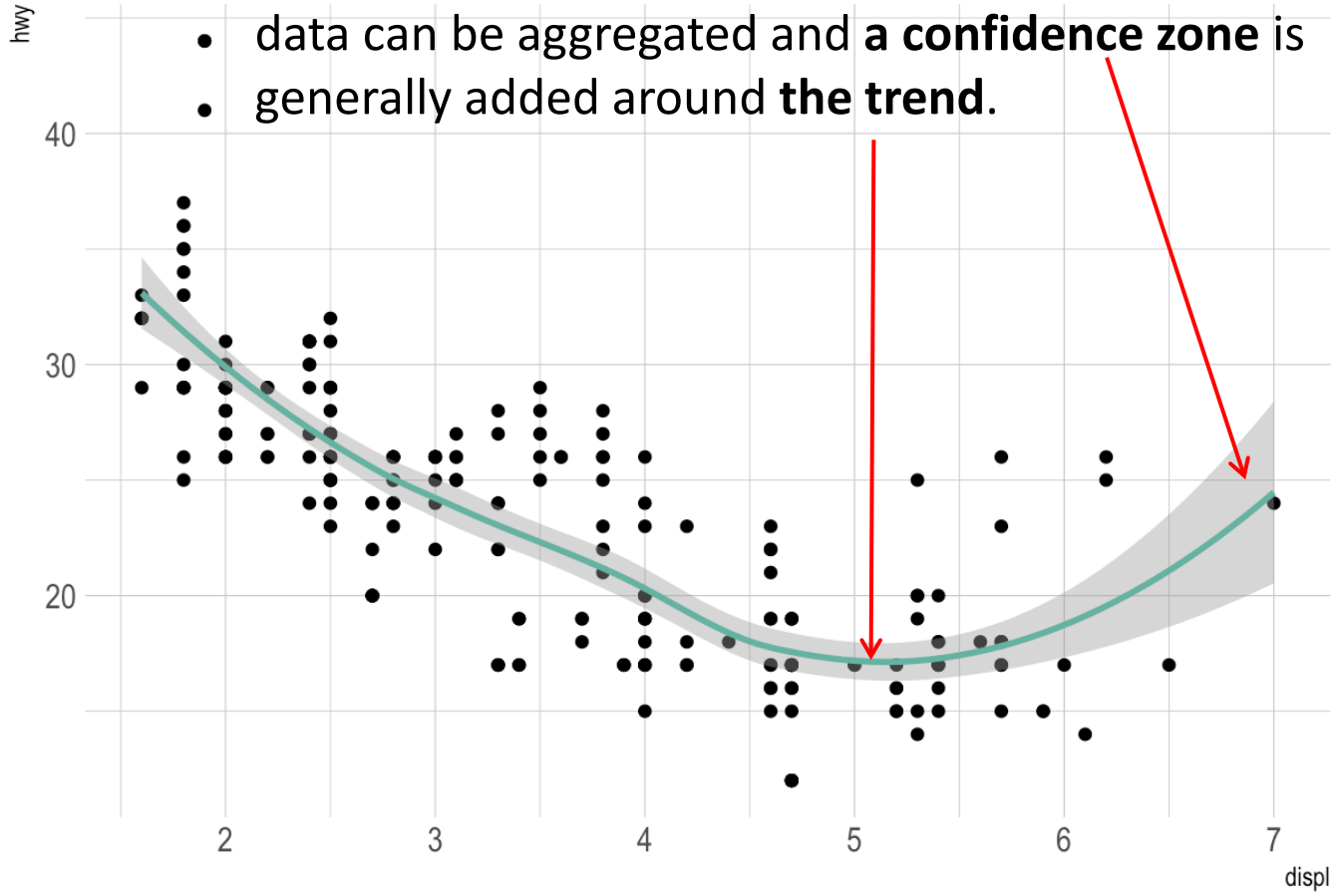
Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualization for Time Series Data

Data Visualization

Line chart

- When several values are available for each time stamp, data can be aggregated and **a confidence zone** is generally added around **the trend**.



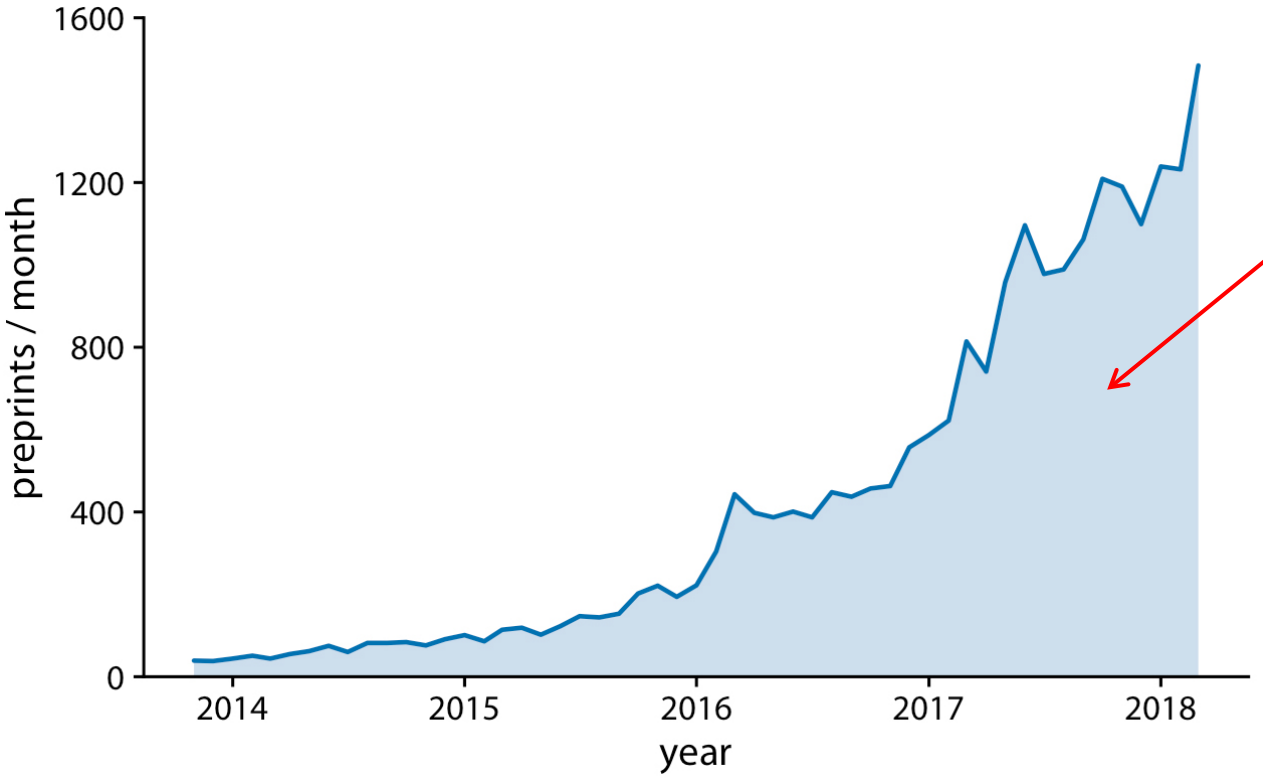
Source: https://www.data-to-viz.com/

Visualization for Time Series Data

Data Visualization

Area chart

Similar to a line chart, except that the area between the x axis and the line is filled in with color or shading.



The area between the x axis and the line is filled.

Monthly submissions to the preprint server bioRxiv, shown as a line graph with filled area underneath. By filling the area under the curve, we put even more emphasis on the overarching temporal trend than if we just draw a line.

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Map and Network

Data Visualization

Choropleth Map

Bubble Map

Network Diagram

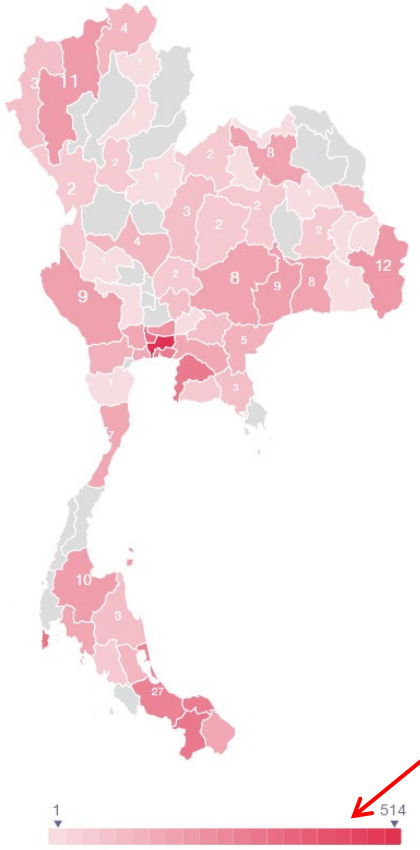
Source: https://www.data-to-viz.com/

Map and Network

Data Visualization

Choropleth Map

- Displays divided geographical areas that are colored in relation to a numeric variable.



The number of COVID-19 patients in every Thailand provinces. The data is officially reported on March 30, 2020.
Source: https://covid19.workpointnews.com/ (Retrieved on 30/3/2020)

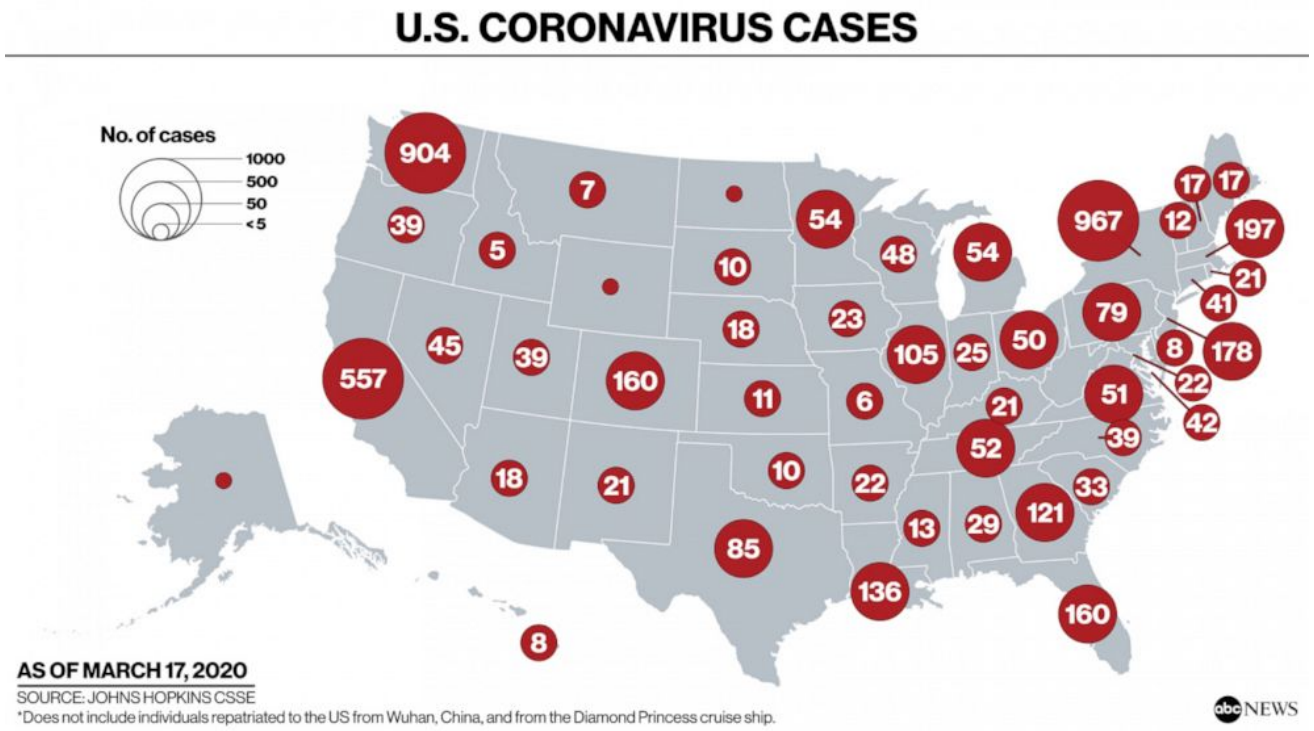
Choropleth map uses different color tones to represent the numerical values.

Map and Network

Data Visualization

Bubble Map

- Similar to choropleth map but uses circles of different size to represent a numeric value on a territory.
- It is possible to display a bubble per geographic coordinate, or a bubble per region.



The number of COVID-19 cases in the U.S. reported on March 17, 2020.

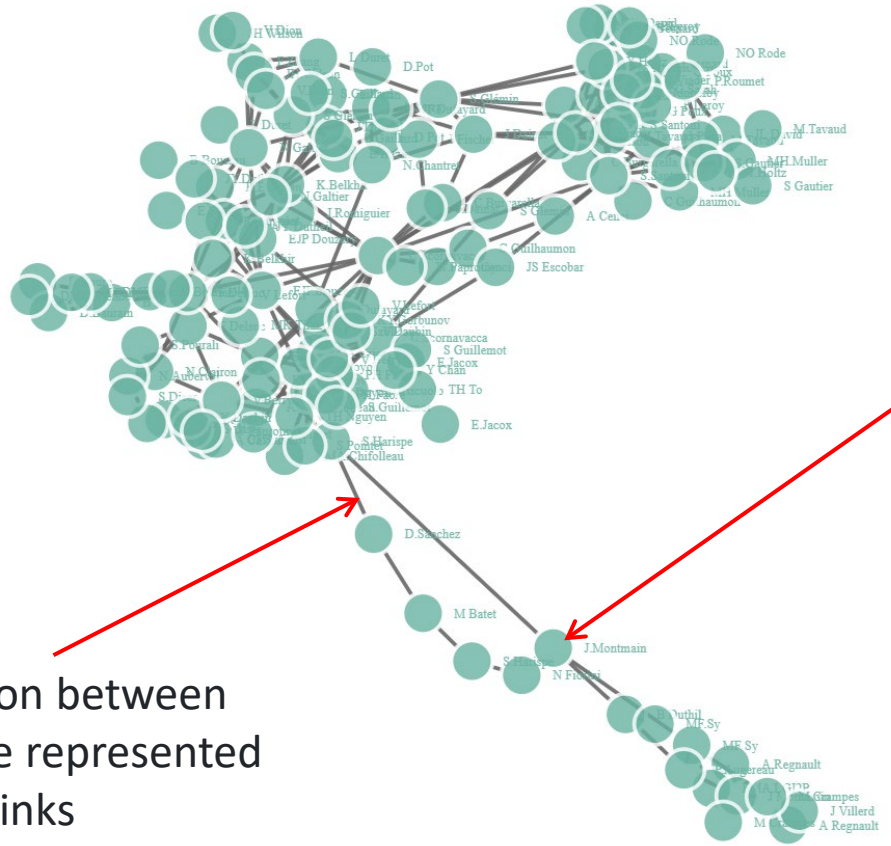
Source: http://www.kake.com/story/41905619/coronavirus-map-tracking-the-spread-in-the-us-and-around-the-world (Retrieved on 30/3/2020)

Map and Network

Data Visualization

Network Diagram

Show interconnections between a set of entities (data point).



Entity is represented by a Node

Connection between nodes are represented through links

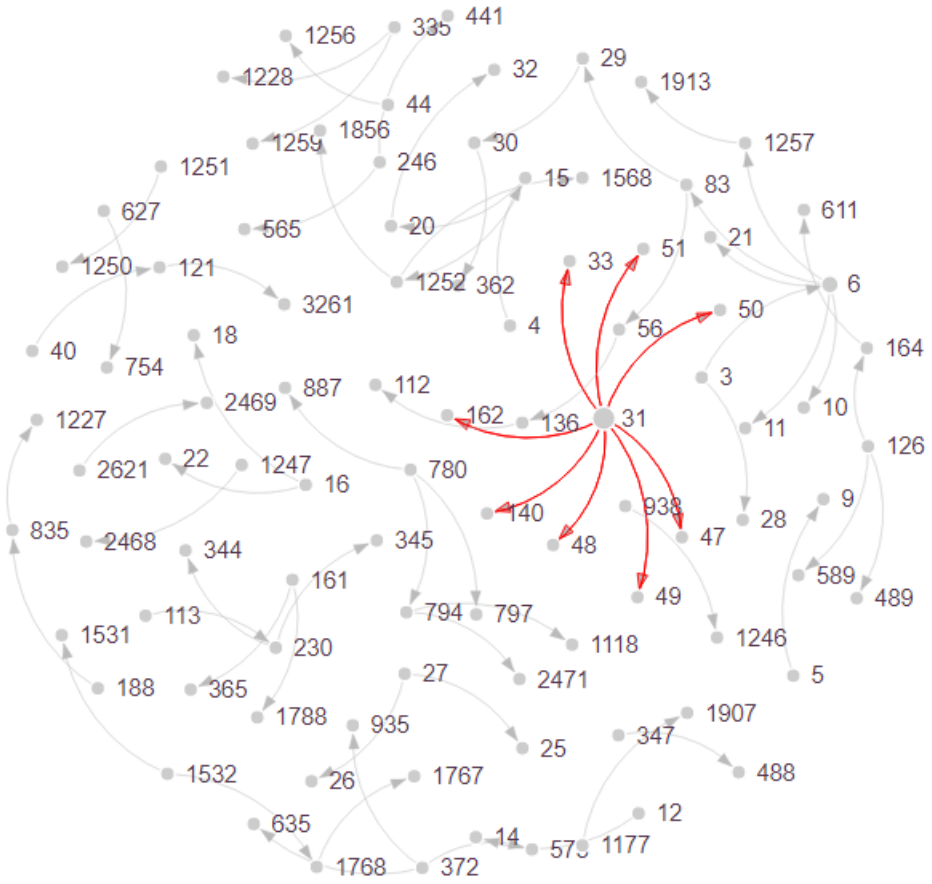
The co-authors network of Vincent Ranwez, a researcher who's my previous supervisor. Basically, people having published at least one research paper with him are represented by a node. If two people have been listed on the same publication at least once, they are connected by a link.

Source: https://www.data-to-viz.com/graph/network.html

Map and Network

Data Visualization

Network Diagram



The network of COVID-19 patients in Korea. Each node represents a patient with patient's ID. By representing by a link, the patient on tail of arrow was infected by the patient on head of arrow.

Further Study

Book:

- Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Website:

- https://towardsdatascience.com/10-viz-every-ds-should-know-4e4118f26fc3
- https://medium.com/@nutdnuy/1-%E0%B8%9A%E0%B8%97%E0%B8%99%E0%B8%B3-867a02e07a74
- https://www.data-to-viz.com/