Data Engineering

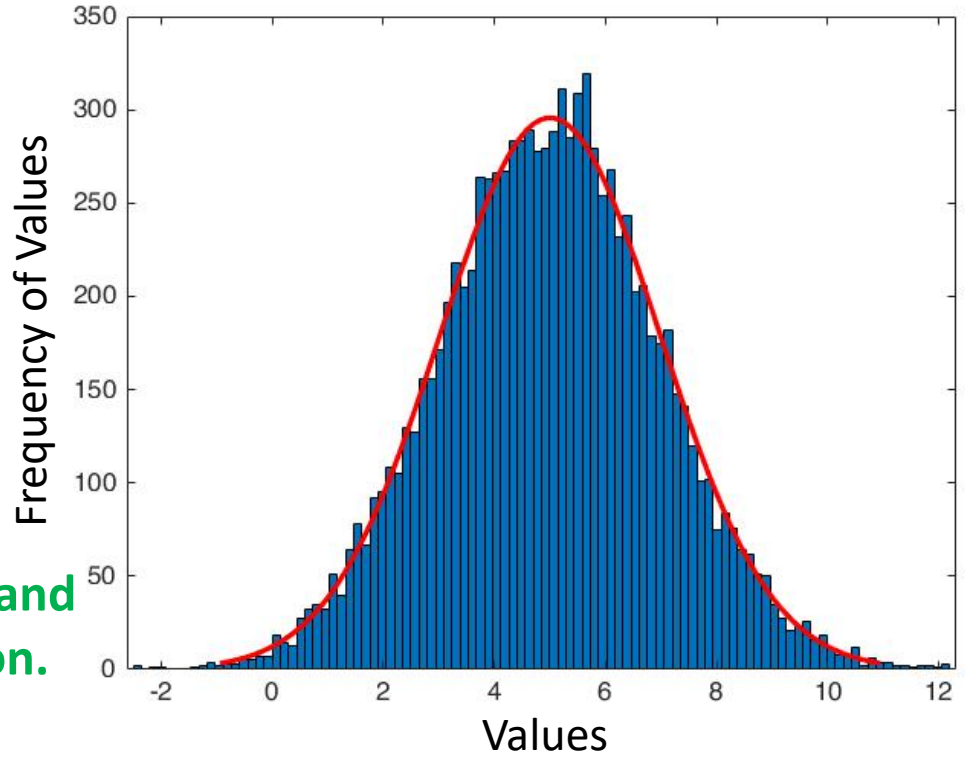
204426

Data Exploration & Data Visualization

Distribution Shape

Mean, Median and Mode

	X_1	X_2	...	X_{10}
$\mathbf{x}_1$				
...				
$\mathbf{x}_n$				



We can slice a feature/variable and describe it as a data distribution.

A distribution in statistics is a function that shows:

- the possible values for a variable (x-axis)
- how often they occur (y-axis).

Distribution Shape

Mean

- A measure of a central or typical value for a probability distribution.
- The sum of all measurements divided by the number of observations in the data set.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Example:

- Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12
- Mean of job performance:

$$\bar{x} = \frac{7+10+11+15+10+10+12+14+16+12}{10} = \frac{117}{10} = 11.7$$

Distribution Shape

Median

- Reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliners.
- The middle value that separates the higher half from the lower half of the data set.
- To compute the middle value, we need to arrange all the numbers from smallest to greatest.
- Then,

$$\tilde{x} = \begin{cases} x_{\frac{(n+1)}{2}}, & if\ n\ is\ odd, \\ \frac{\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)}{2}, & if\ n\ is\ even, \end{cases}$$

Example:

- Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12
- Median of job performance:

n = 10. So, n is even

$$\tilde{x} = \frac{x_5 + x_6}{2} = \frac{11 + 12}{2} = 11.5$$

11.5

7	10	10	10	11	12	12	14	15	16
				x_5	x_6				

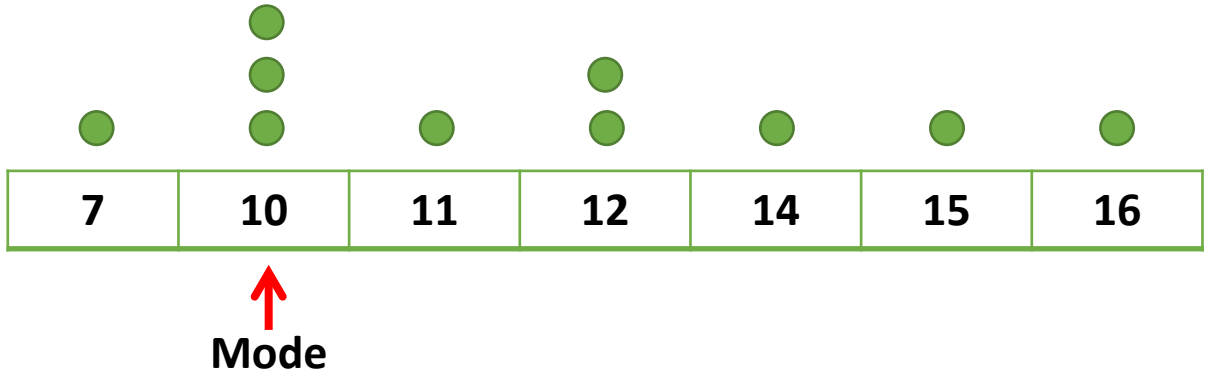
Distribution Shape

Mode

- The most frequent value in the data set.

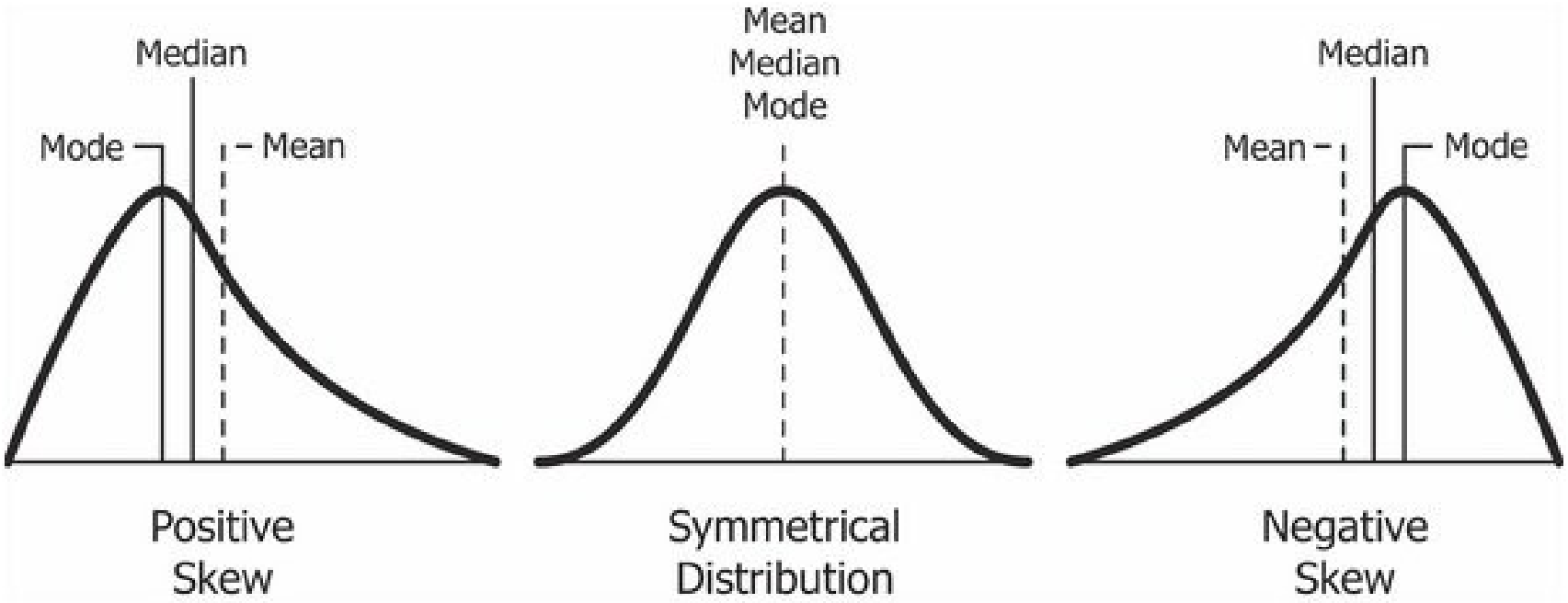
Example:

- Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12
- Mode of job performance:



Distribution Shape

Geometric visualization of the mode, median and mean of an arbitrary probability density function

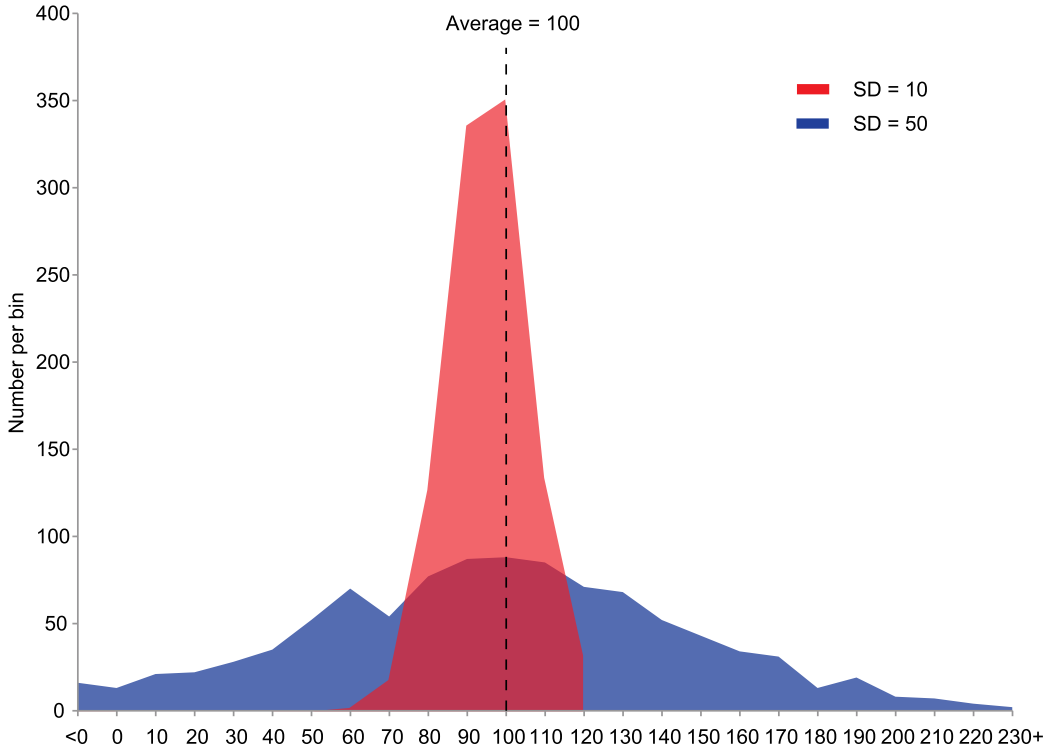


Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Distribution Shape

Standard Deviation (SD, s)

- A measure that is used to quantify the amount of variation or dispersion of a set of data values.
- A low standard deviation indicates that the data points tend to be close to the mean.
- A high standard deviation indicates that the data points are spread out over a wider range of values.



Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Comparison_standard_deviations.svg

Distribution Shape

Standard Deviation (SD, s)

The formula for the sample standard deviation is

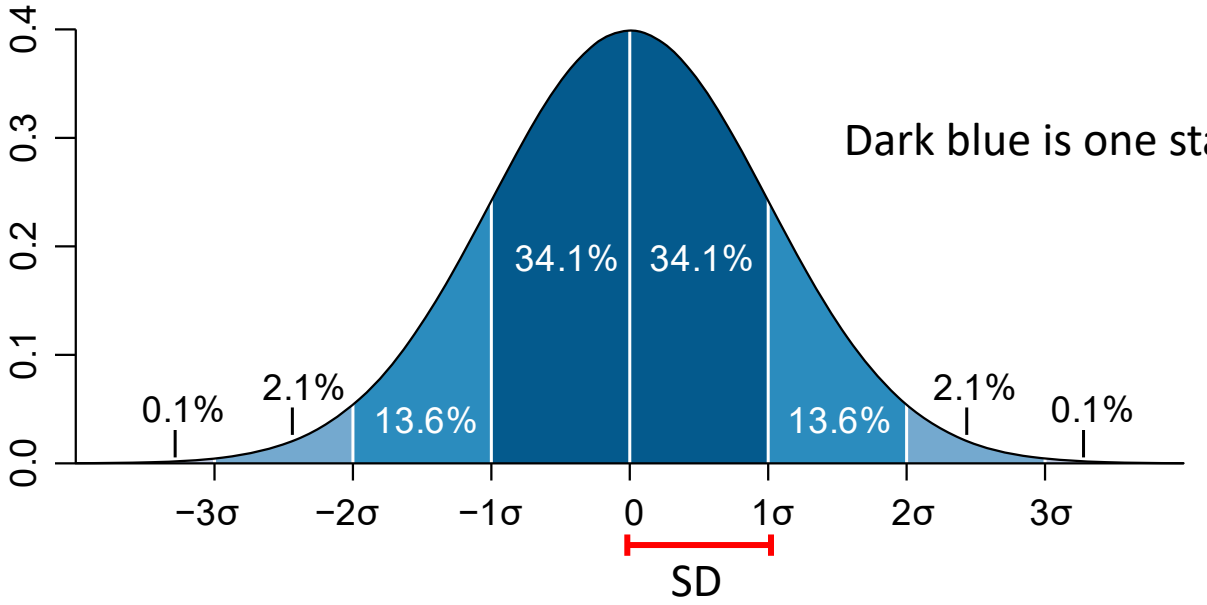
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The formula for the population standard deviation is

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Distribution Shape

Standard Deviation (SD, s)



Dark blue is one standard deviation on either side of the mean.

Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

Distribution Shape

Variance

- How far a set of numbers are spread out from their average value.
- It is the square of the standard deviation
- The formula for the sample variance is

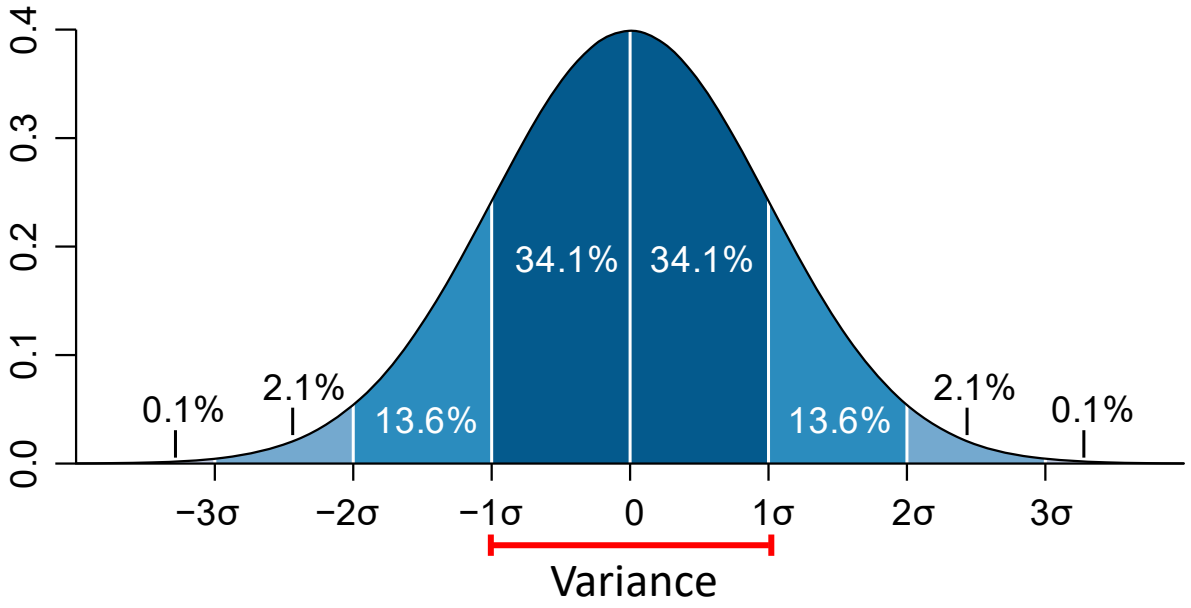
$$s^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}^{2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- The formula for the population variance is

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Distribution Shape

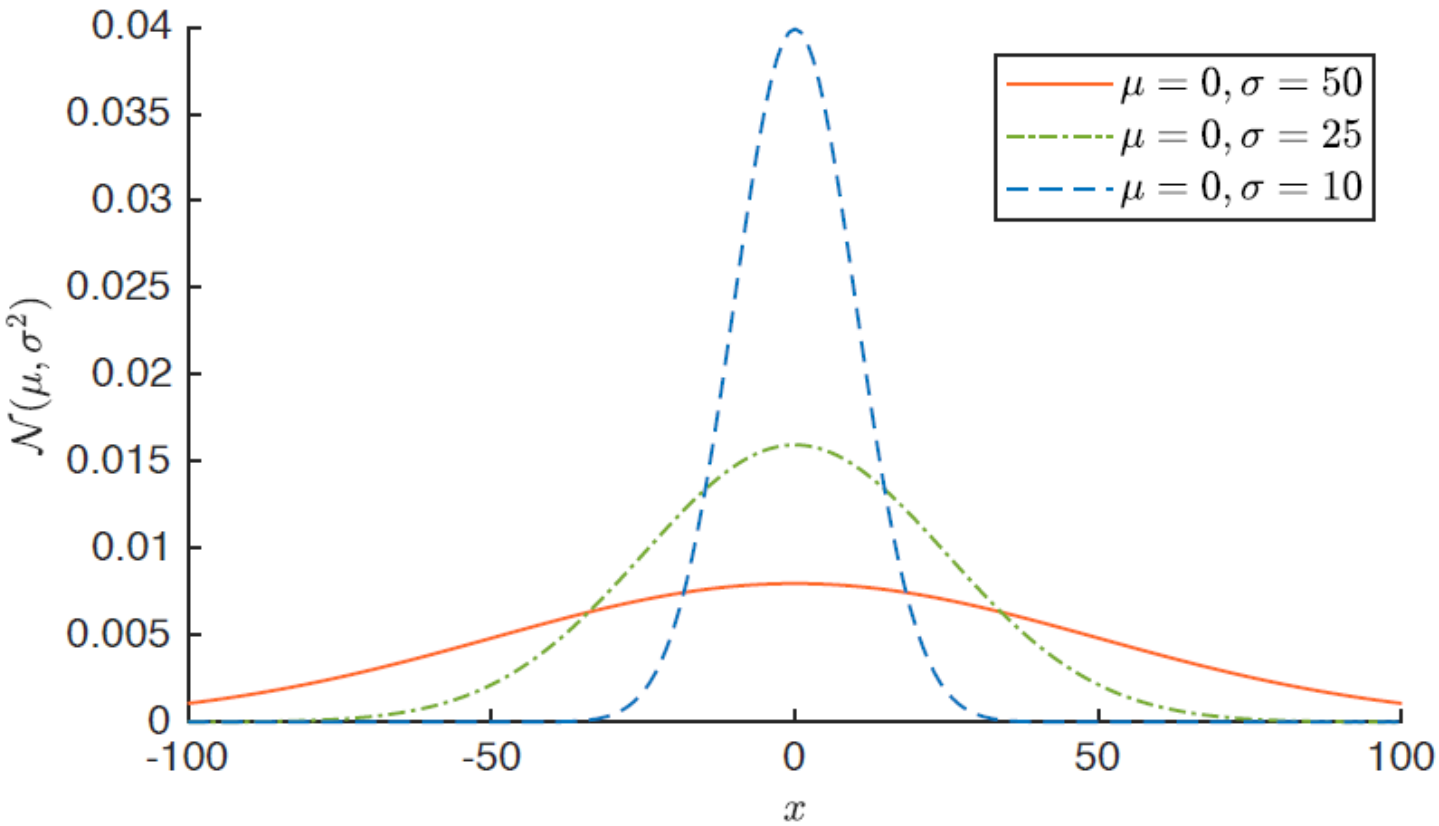
Variance



Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

Distribution Shape

Standard Deviation and Variance

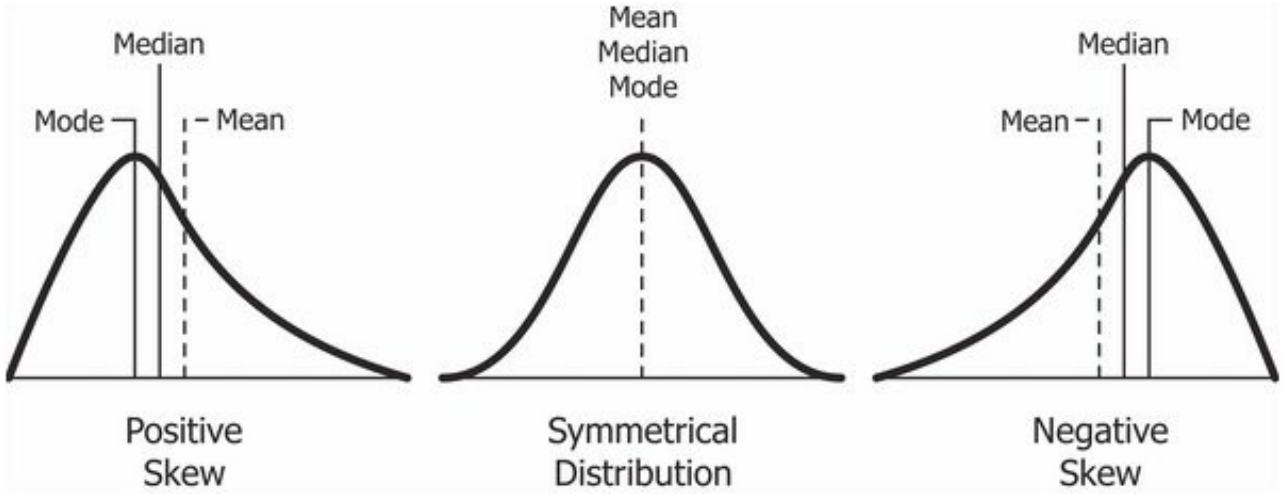


Distribution Shape

Skewness

- Skewness is usually described as a measure of a **dataset's symmetry** – or lack of symmetry.
- The normal distribution has a skewness of 0.
- The skewness can be calculated by

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{2}{3}}}$$



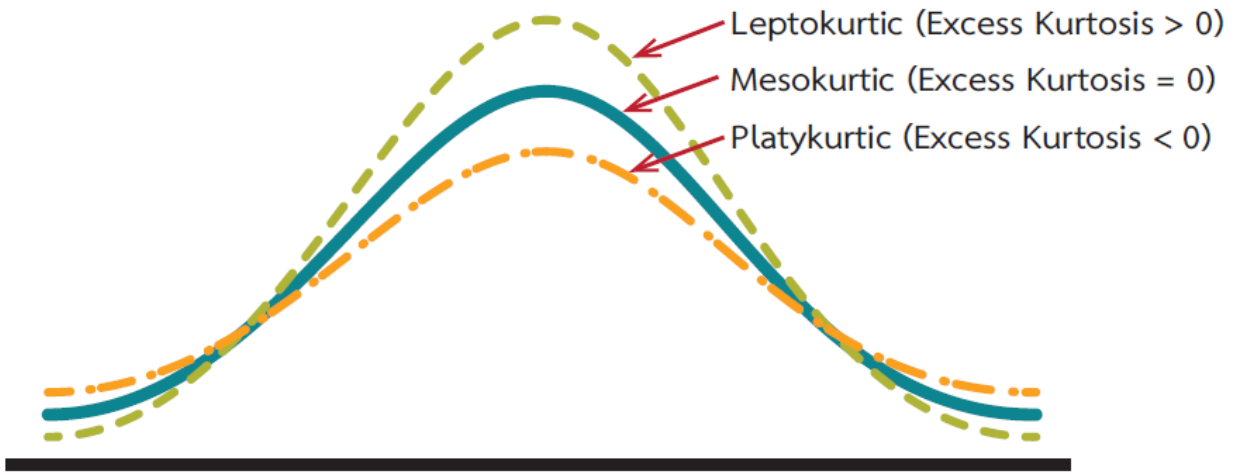
Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Distribution Shape

Kurtosis

- Measures the **tail-heaviness of the distribution**.
- The excess kurtosis for a standard normal distribution is 0.
- The excess kurtosis can be calculated by

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} - 3$$



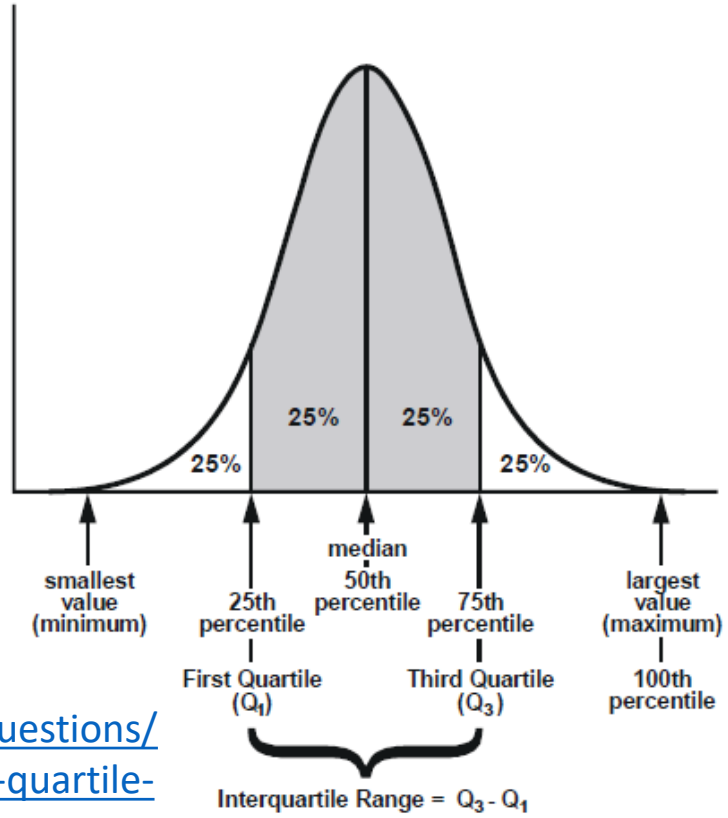
Distribution Shape

Interquartile Range

- The difference between the first and the third sample quartiles.
- This give the range of middle 50% of the data
- It is found from the following

$$IQR = q(0.75) - q(0.25)$$

- **Lower limit**: $LL = q(0.25) - 1.5 \times IQR$
- **Upper limit**: $\mathrm{U}L = q(0.75) + 1.5 \times IQR$
- Observations outside these limits are **potential outliers**.



Source: https://stats.stackexchange.com/questions/470193/representing-quantile-like-quartile-in-form-of-normal-distribution-curve

Distribution Shape

Histogram

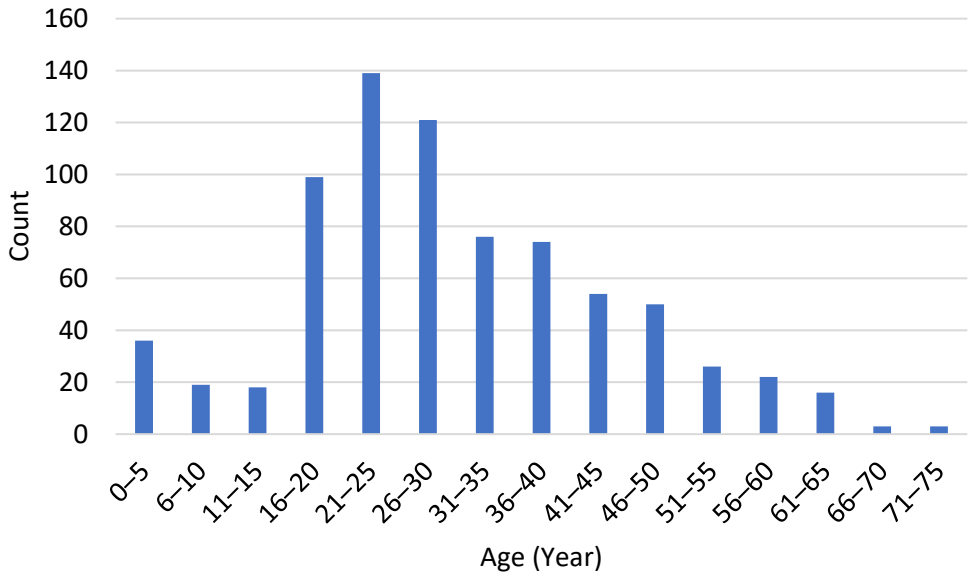
- A histogram takes as input a **numeric variable** only.
- The variable is cut into several bins
- The number of observation per bin is represented by the height of the bar.

	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

Distribution Shape

Histogram bin widths

- **Sturges' rule**

$$k = 1 + \log_2 n$$

  - k is the number of bins.
  - The bin width *h* is obtained by taking the range of the sample data and dividing it into the requisite number of bins.
- **Normal reference rule**

$$h \approx 3.5 \times \sigma \times n^{-\frac{1}{3}}$$

- **Scott's rule**

$$h = 3.5 \times s \times n^{-\frac{1}{3}}$$

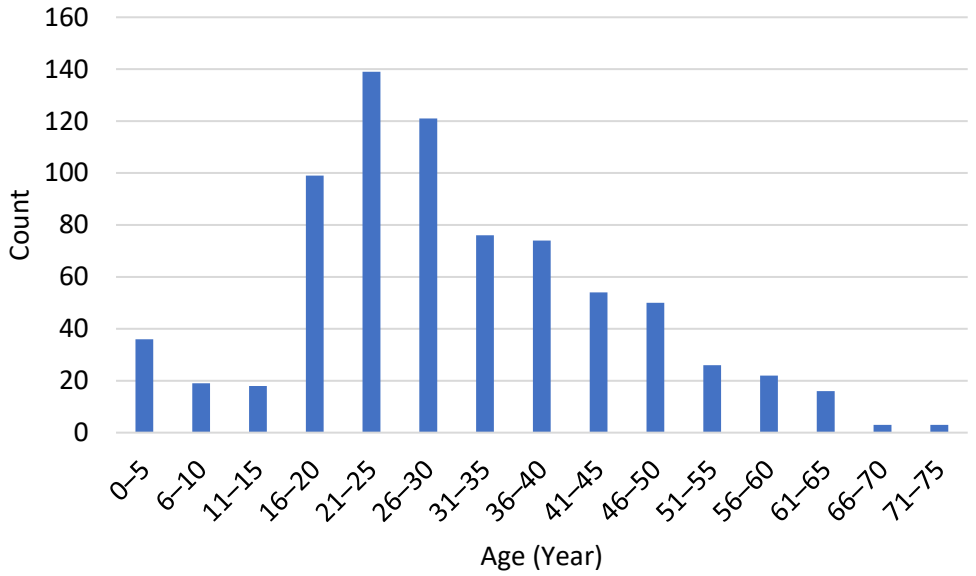
- **Freedman-Diaconis rule**

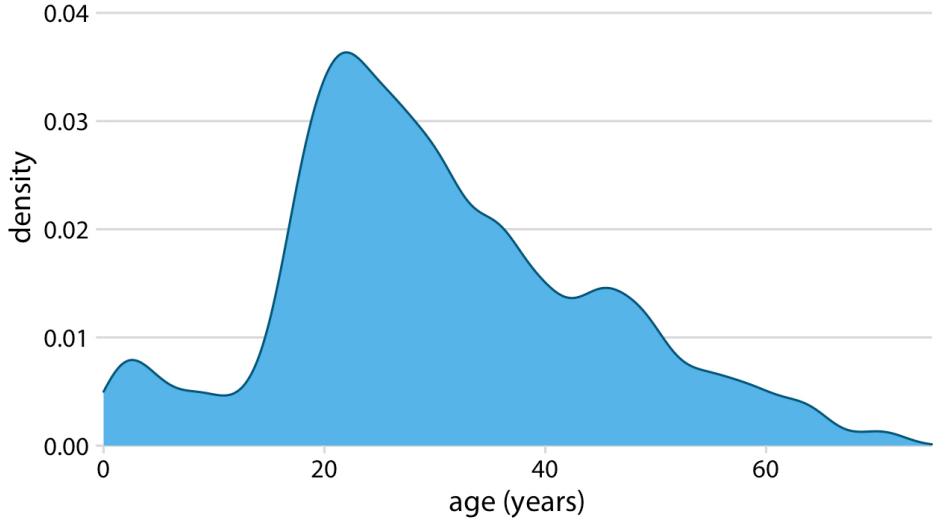
$$h = 2 \times IQR \times n^{-\frac{1}{3}}$$

Distribution Shape

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**.





Density plot

Distribution Shape

Kernel density estimation

- The univariate kernel estimator is given by

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

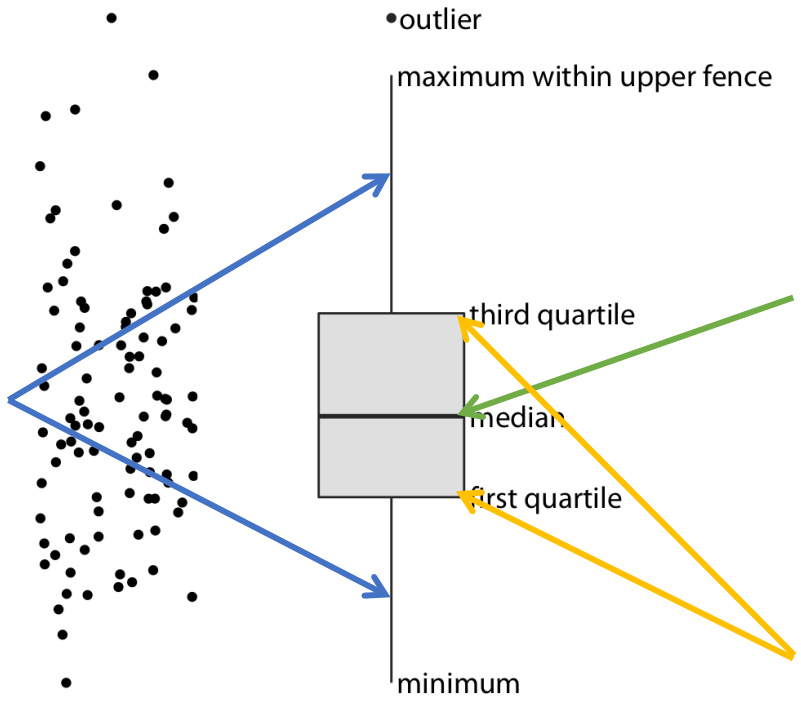
- $K(t)$ is a kernel
- $h > 0$ is a smoothing parameter called the bandwidth, and can be estimated by

$$h = 0.786 \times IQR \times n^{-\frac{1}{5}}$$

Distribution Shape

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

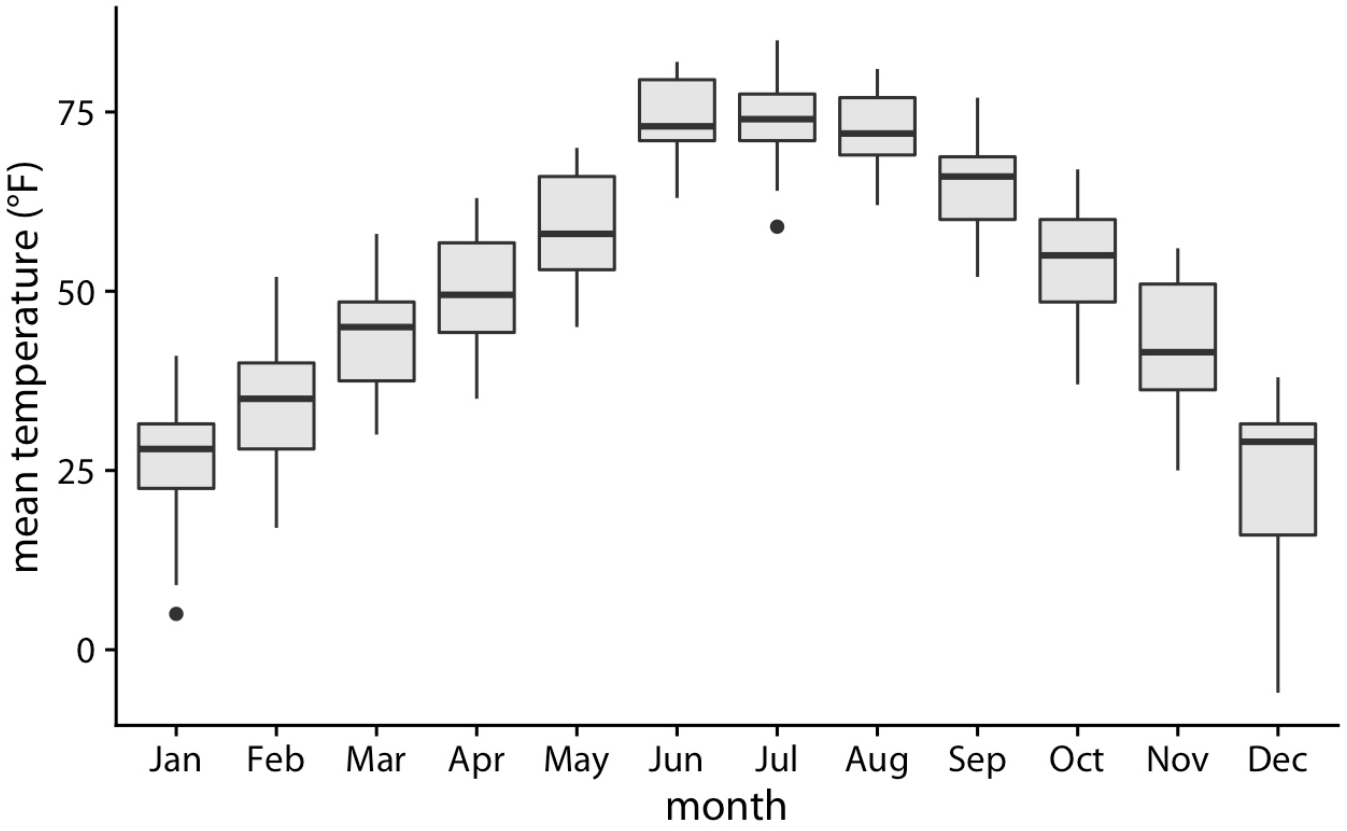
The end of the box shows the third and first quartiles.

Anatomy of a boxplot

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

Distribution Shape

Boxplot

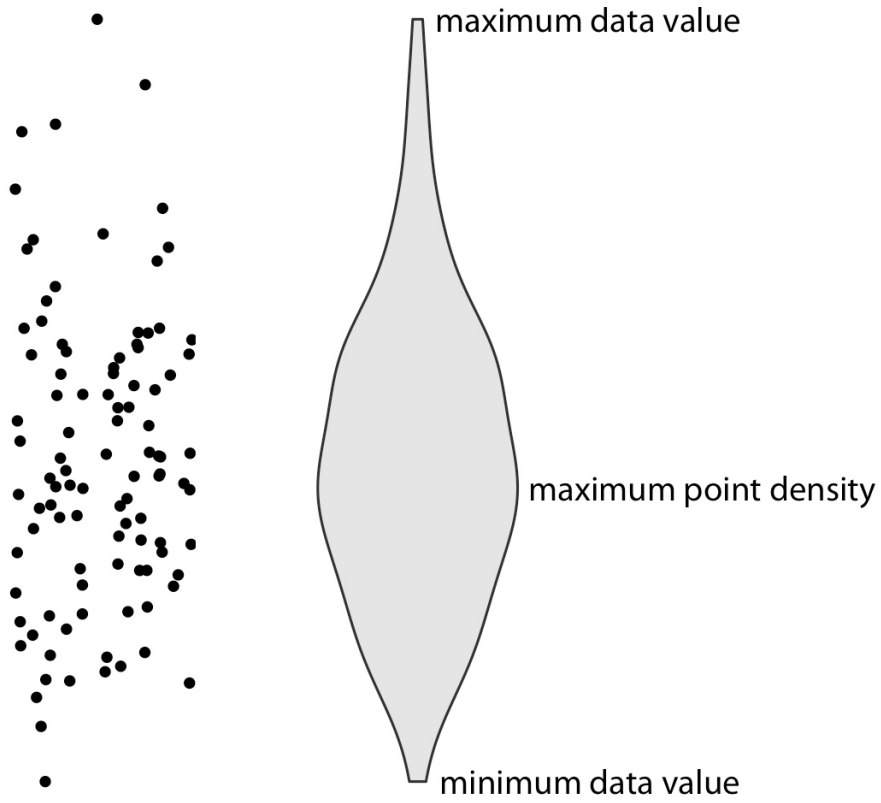


Mean daily temperatures in Lincoln, NE, visualized as boxplots.

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

Distribution Shape

Violin Plot



Anatomy of a violin plot

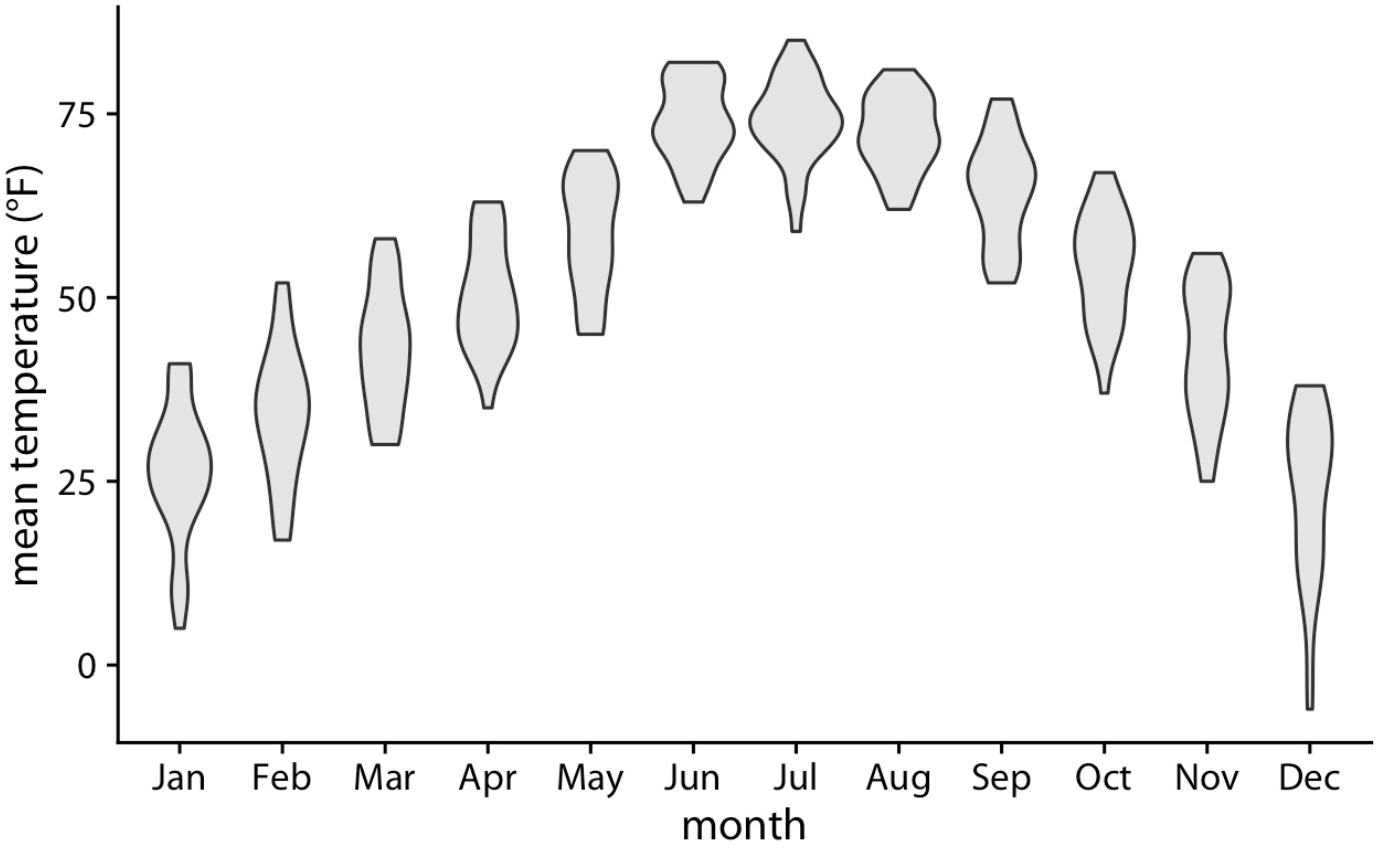
Image Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

- 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.

Distribution Shape

Violin Plot



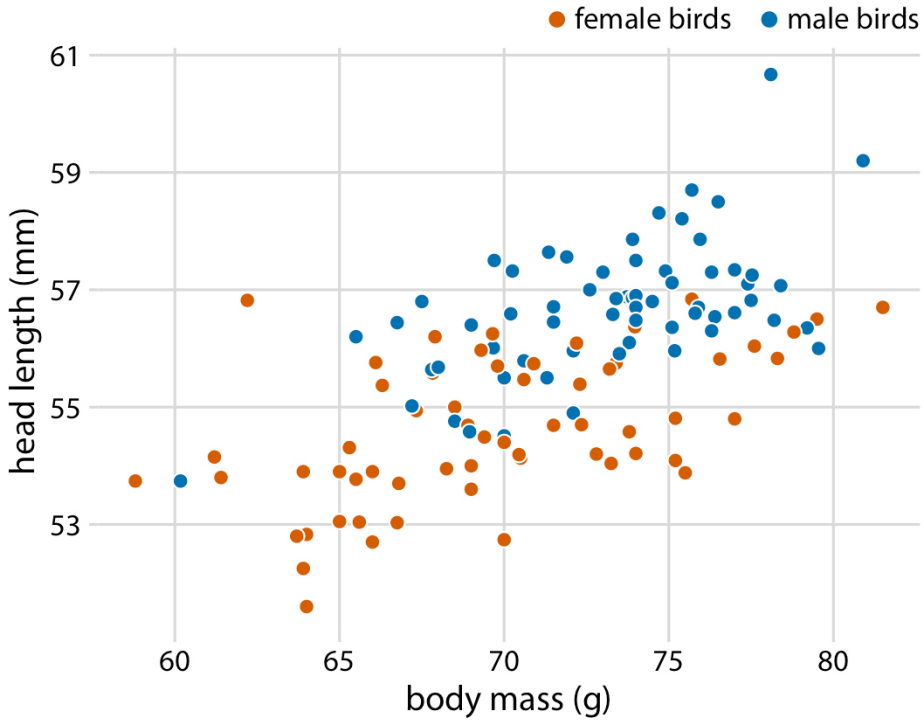
Mean daily temperatures in Lincoln, NE, visualized as violin plot.

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

Multivariate 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.

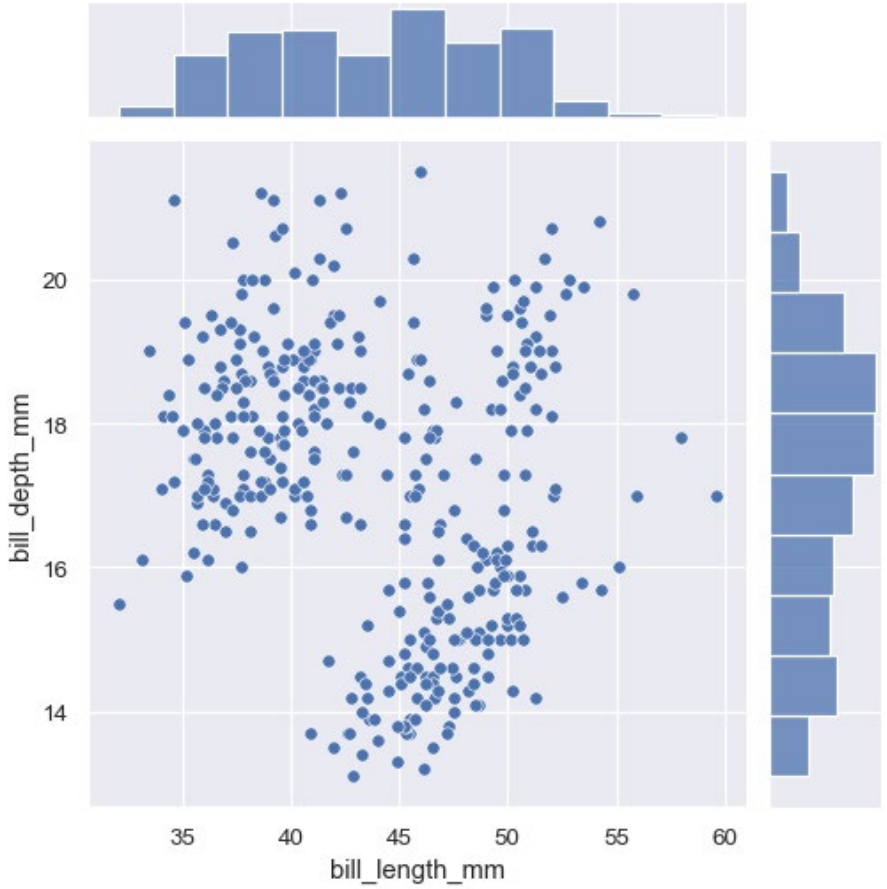


Head length versus body mass for 123 blue jays. The birds' sex is indicated by color. At the same body mass, male birds tend to have longer heads (and specifically, longer bills) than female birds.

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

Multivariate Visualization

Scatterplot with histogram



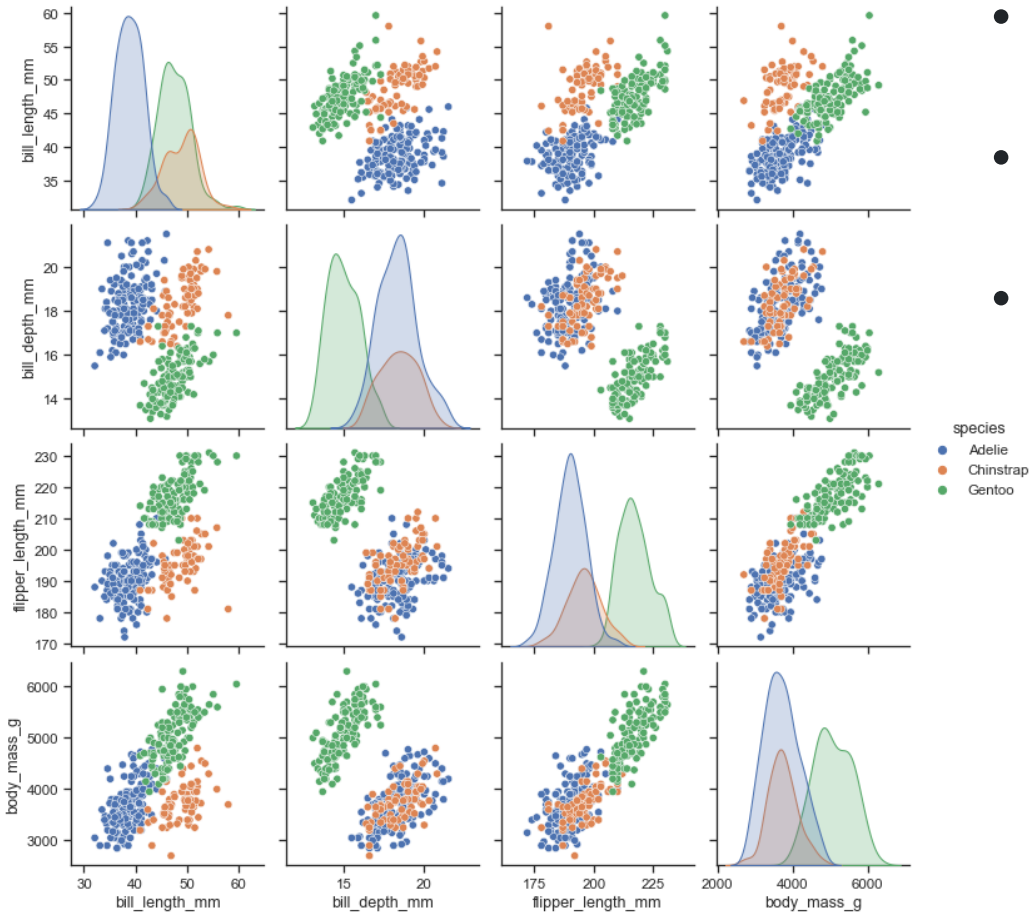
- A 2D scatterplot with univariate histogram along the horizontal and vertical axes.
- The histogram can show useful information about the marginal distributions.

Bill depth versus bill length for penguins data in palmerpenguins dataset .

Source: https://seaborn.pydata.org/tutorial/distributions.html

Multivariate Visualization

Scatterplot Matrices



- Show all possible 2D scatterplots, where the axis of each plot is given by one of the variables
- The scatterplots are arranged in a matrix-like layout for east viewing and comprehension.
- The diagonal boxes can present histograms showing the distribution of each variable.

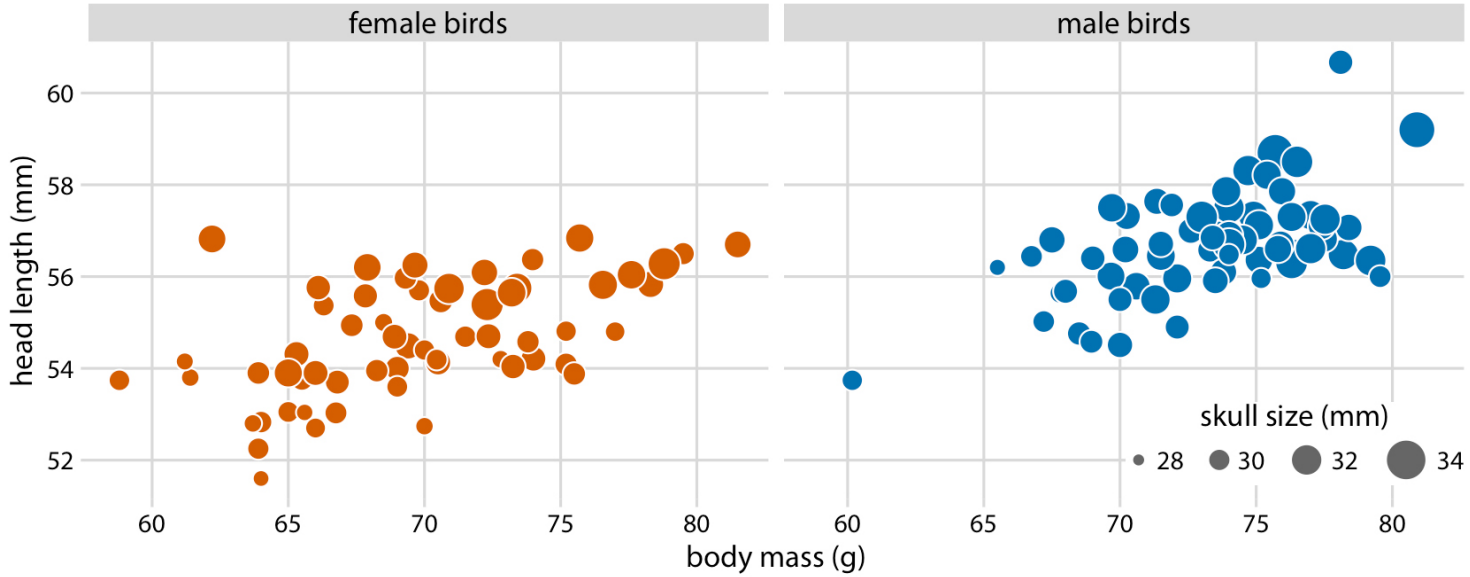
Scatterplot matrix of numerical variables for penguins data in palmerpenguins dataset .

Source: https://seaborn.pydata.org/examples/scatterplot_matrix.html

Multivariate 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.



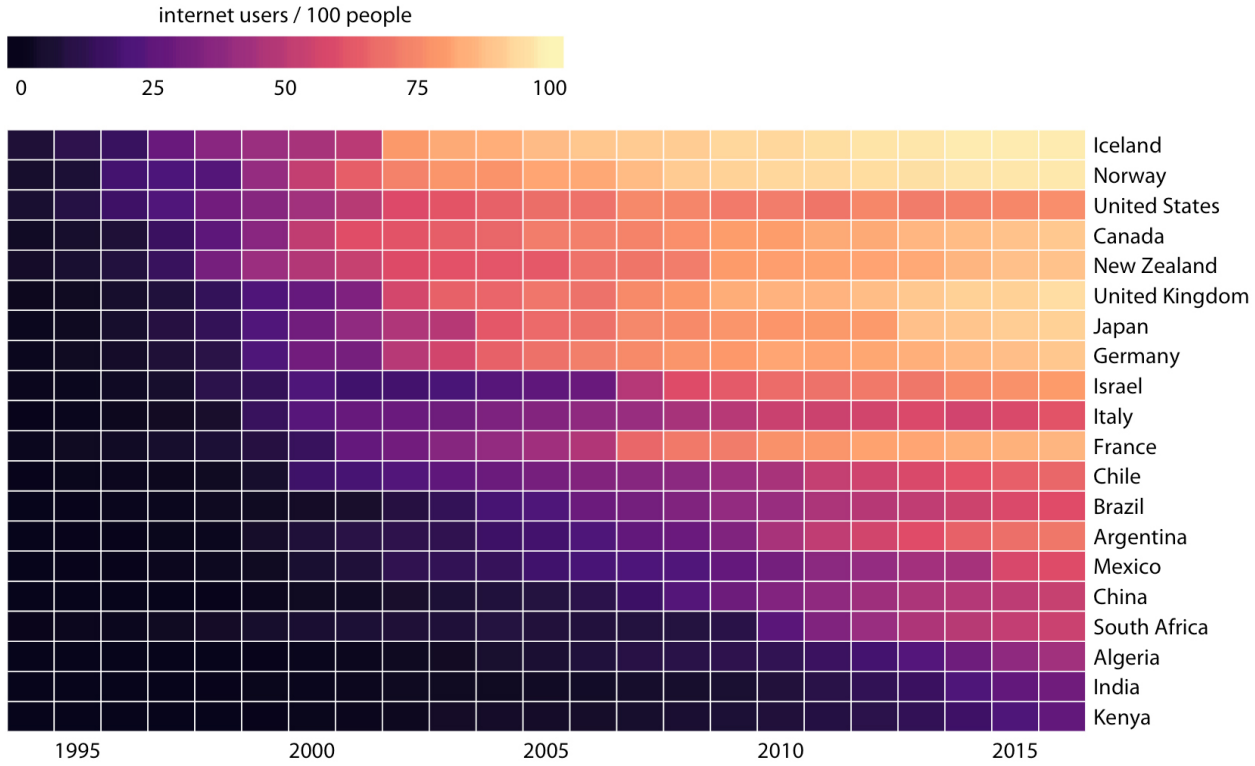
Head length versus body mass for 123 blue jays. The birds' sex is indicated by color and the birds' skull size by symbol size. Head length measurements include the length of the bill while skull size measurements do not. Head length and skull size tend to be correlated, but there are some birds with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

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

Multivariate 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.



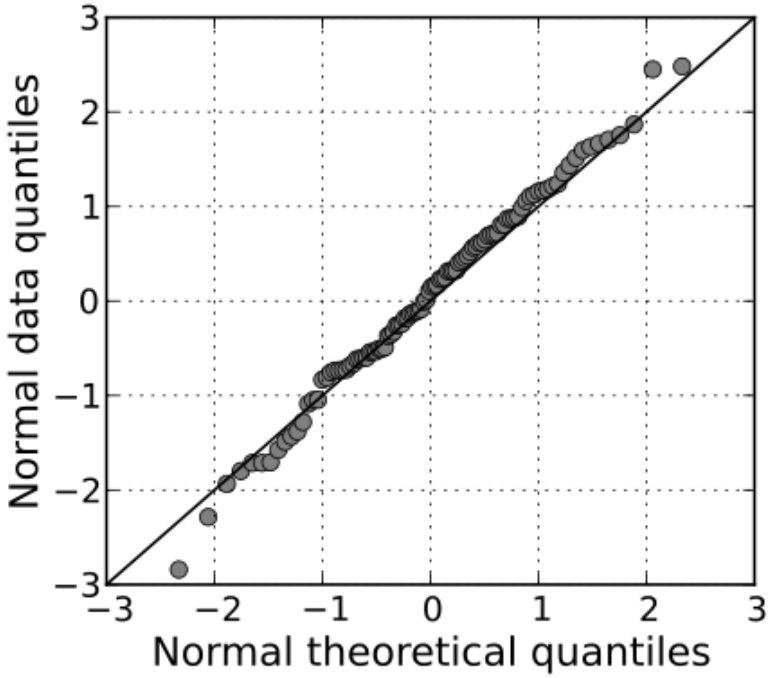
Internet adoption over time, for select countries. Countries were ordered by the year in which their internet usage first exceeded 20%.

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

Determining Data Distribution

Quantile-Quantile Plot (Q-Q Plot)

- A scatterplot created by plotting two sets of quantiles against one another.
- If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



Source:
https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot#/media/File:Normal_normal_qq.svg

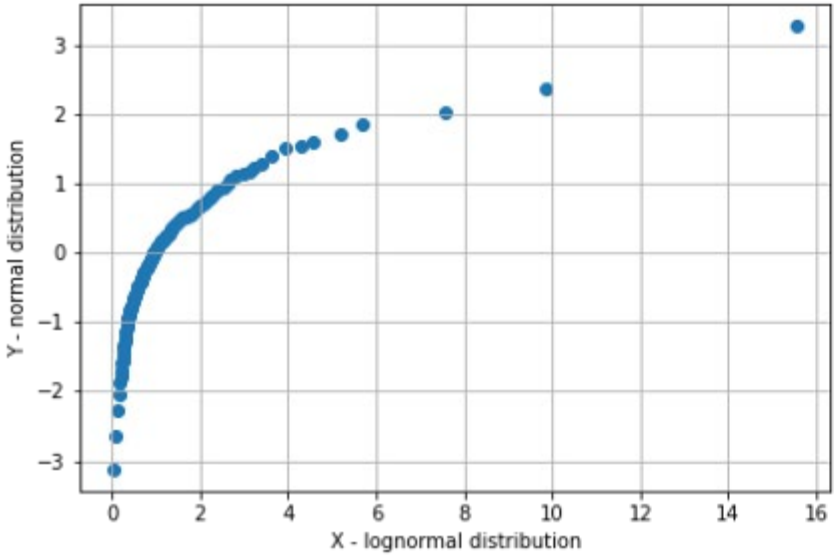
Determining Data Distribution

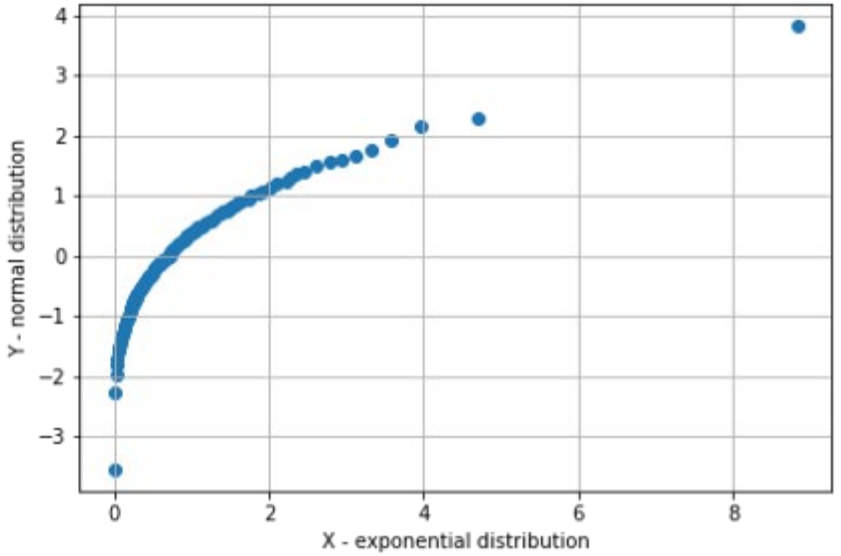
How to Make a Q-Q Plot

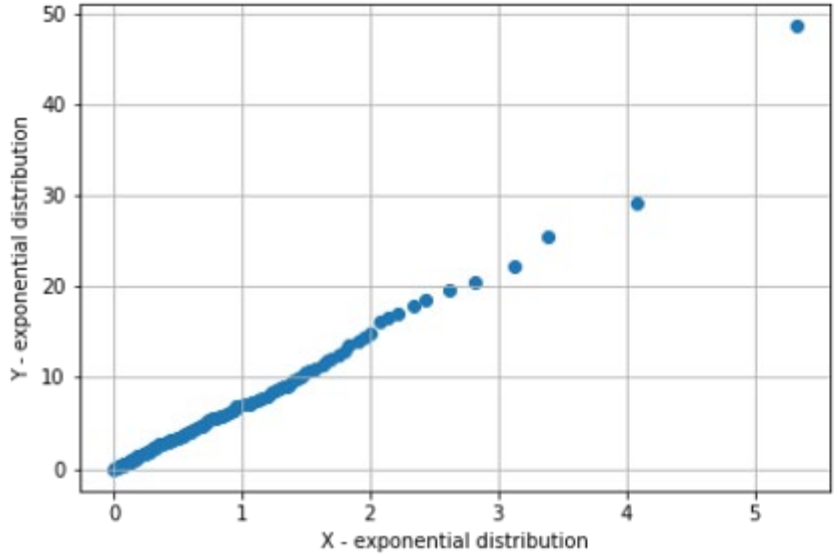
Step 1: Given an unknown random variable.
Step 2: Find each integral percentile value or 100 z-values.
Step 3: Generate a known random distribution and follow steps 1–2 for this distribution too.
Step 4: Plotting Q-Q plot

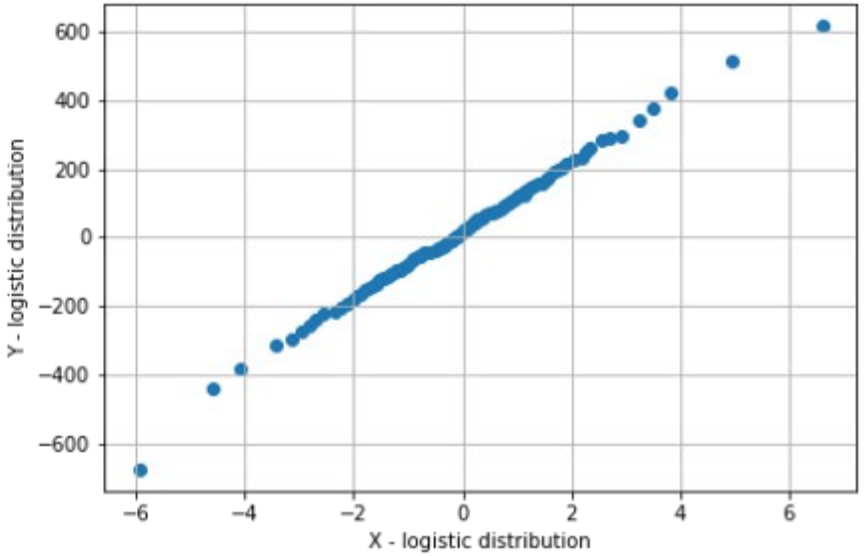
See the following YOUTUBE video for example:
https://www.youtube.com/watch?v=okjYjClSjOg

Determining Data Distribution







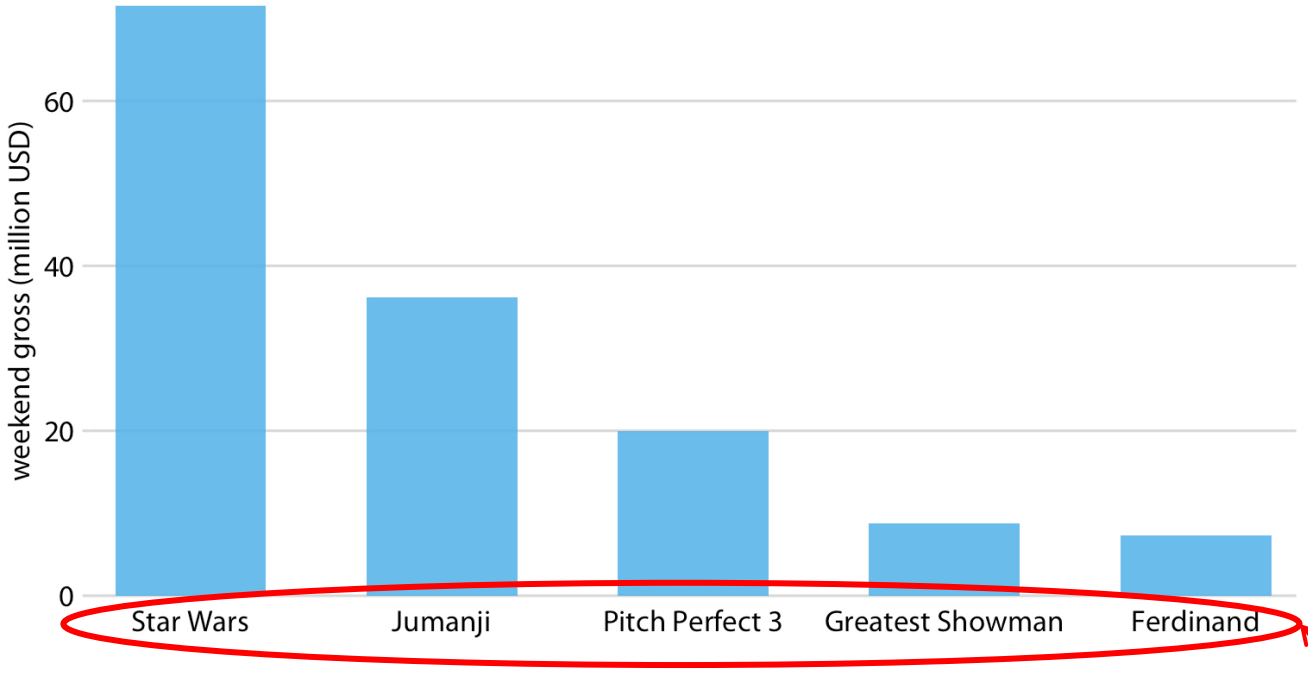


Source: https://towardsdatascience.com/how-to-verify-the-distribution-of-data-using-q-q-plots-acdb7ca2d576

Visualizing Categorical Data

Bar Plot

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

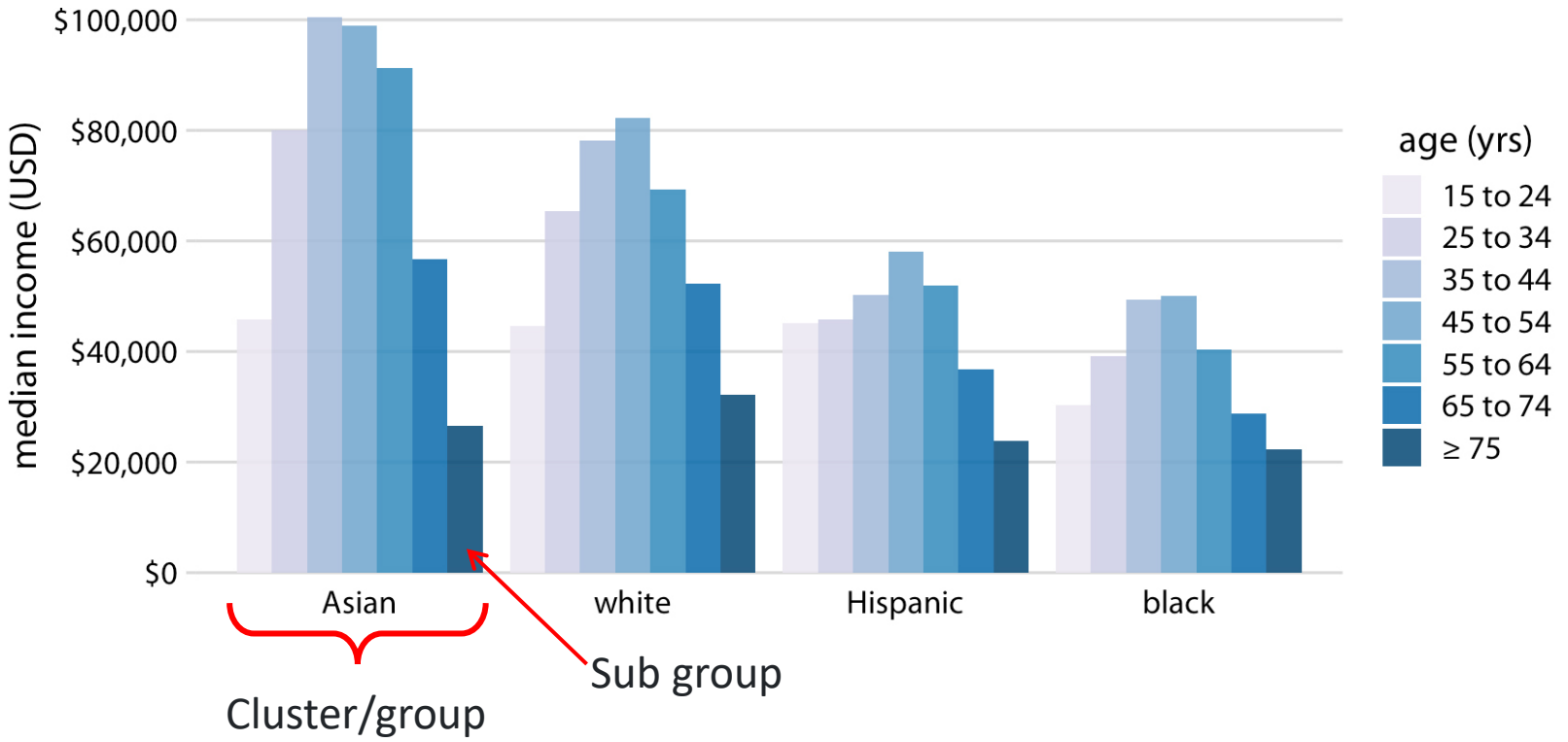


Highest-grossing movies for the weekend of December 22–24, 2017, displayed as a bar plot.

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

Visualizing Categorical Data

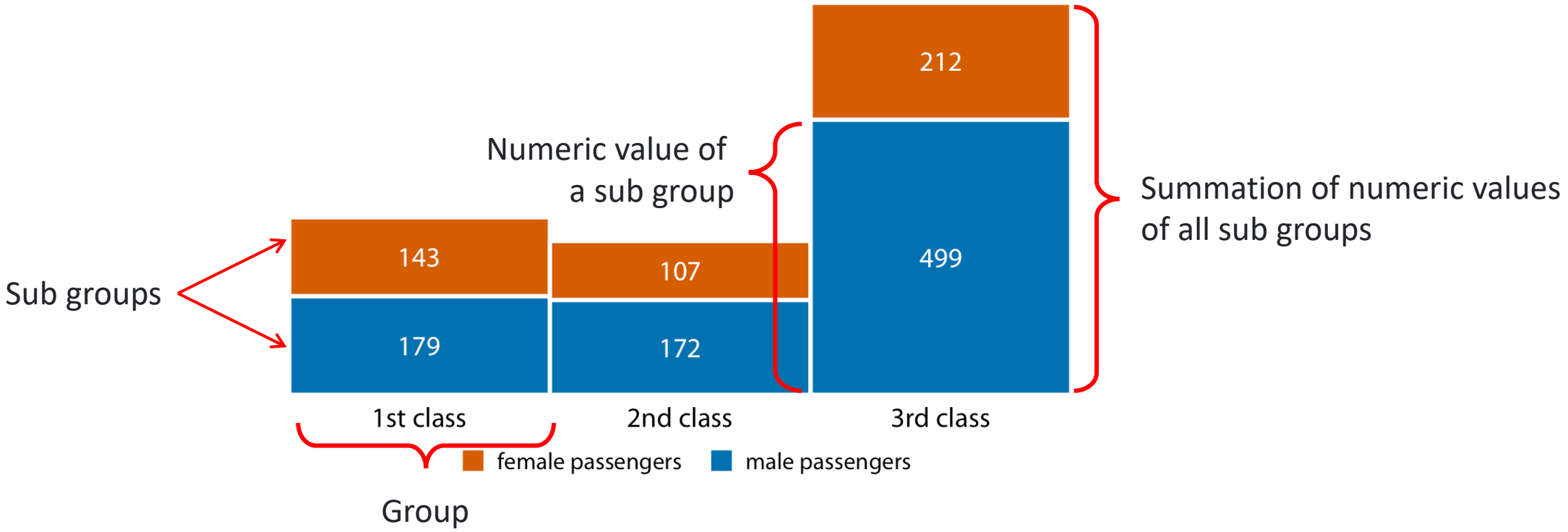
Clustered Bar Plot



Visualizing Categorical Data

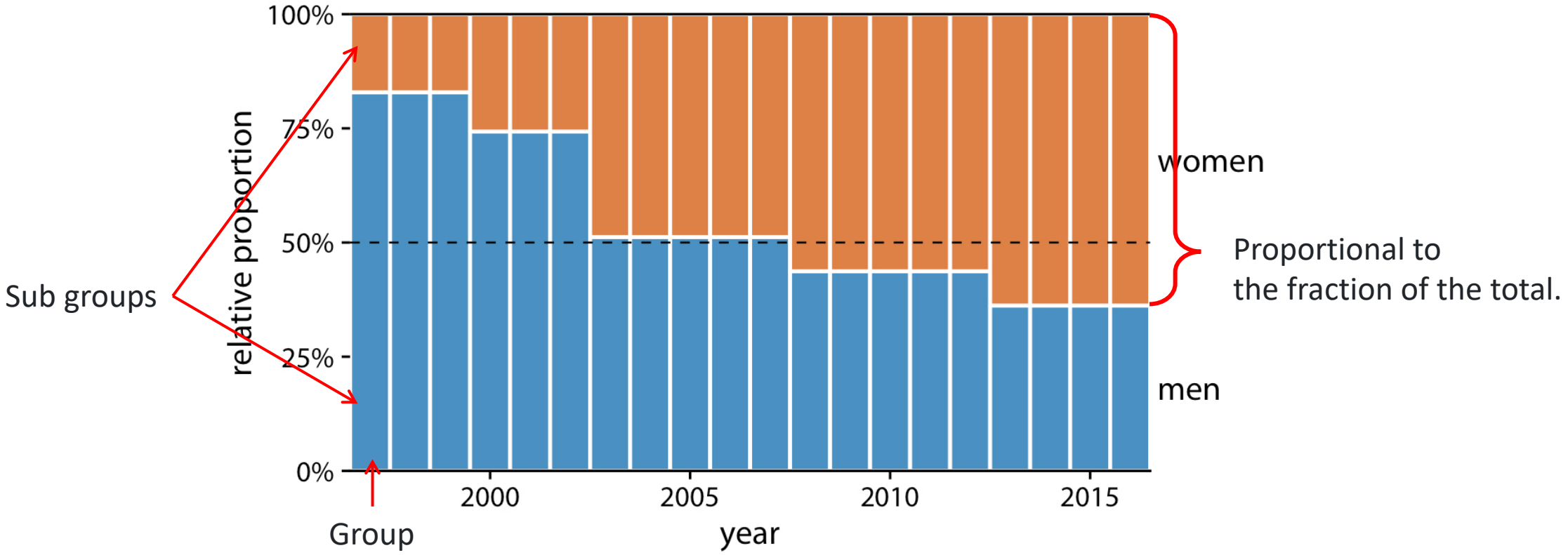
Different Types of Bar Plots

Stack Bar Plot



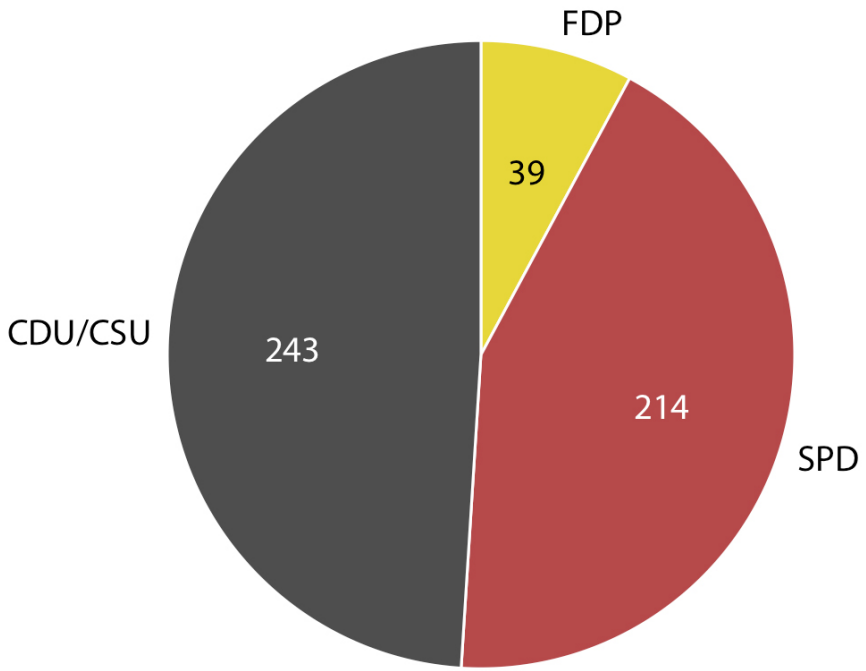
Visualizing Categorical Data

100% Stack Bar Plot



Visualizing Categorical Data

- 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.

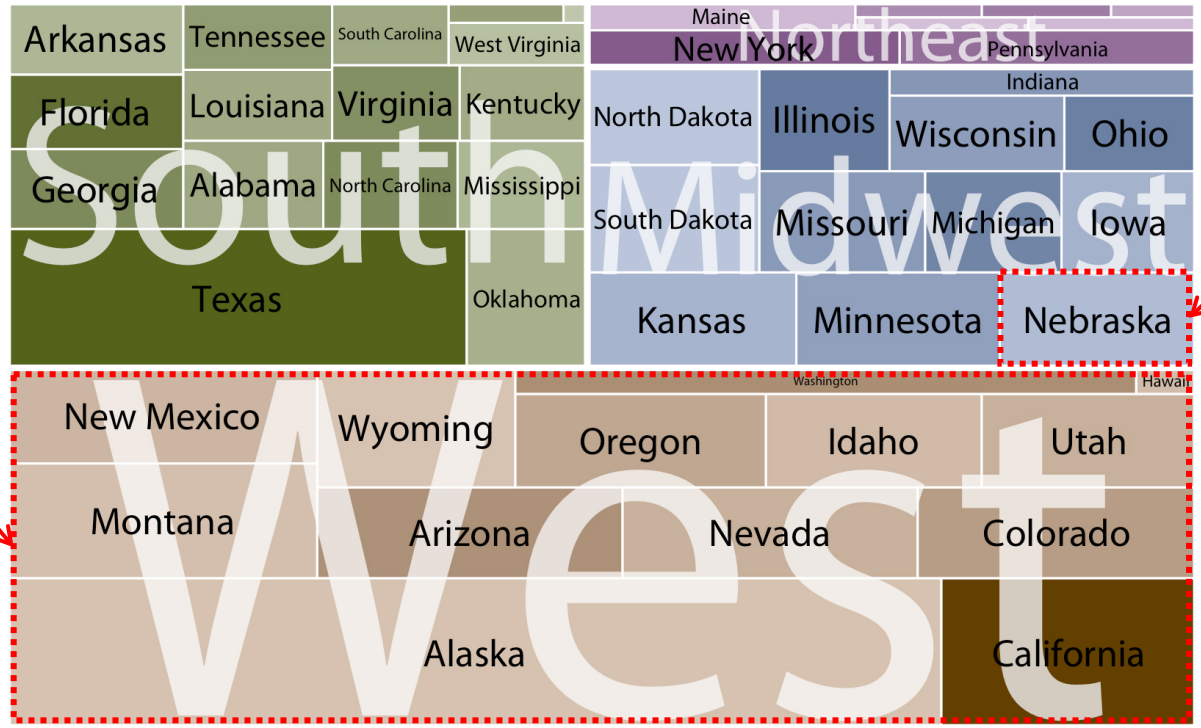
Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Visualizing Categorical Data

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.

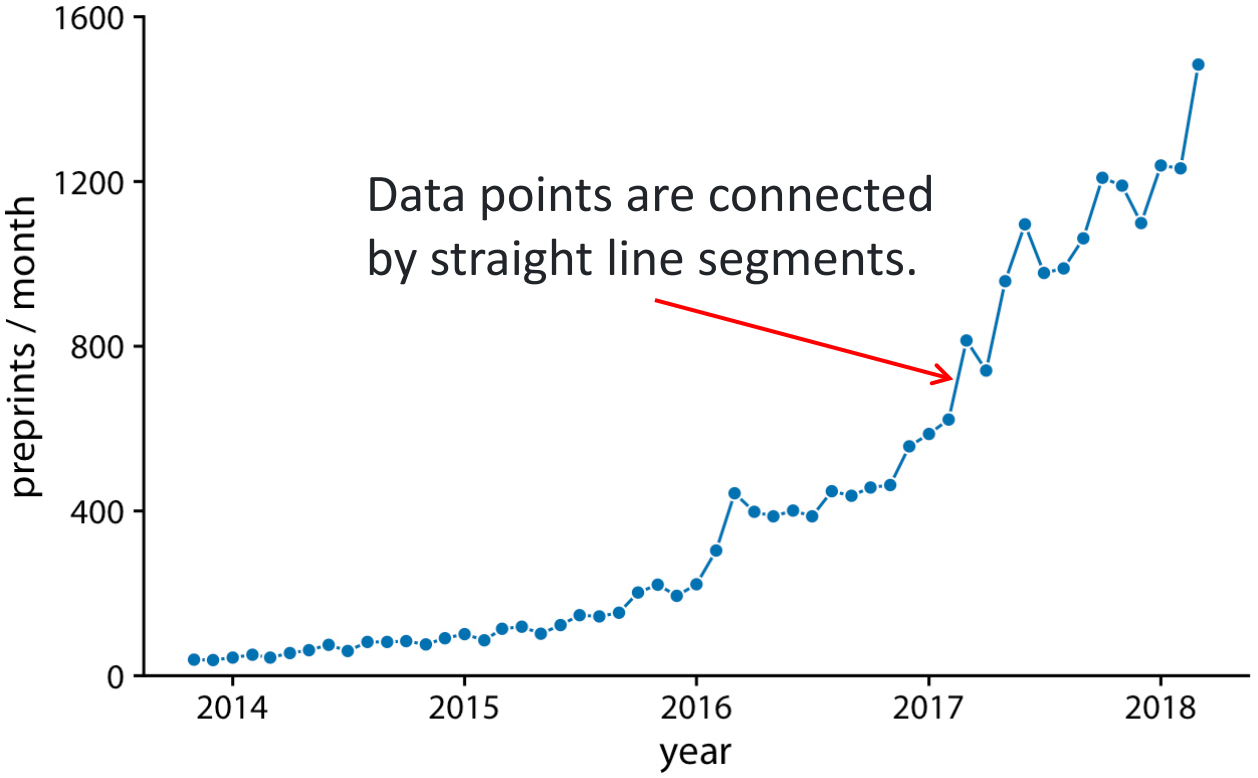
States in the US visualized as a tree map. Each rectangle represents one state, and the area of each rectangle is proportional to the state's land surface area. The states are grouped into four regions, West, Northeast, Midwest, and South. The coloring is proportional to the number of inhabitants for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

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

Visualizing Time Series Data

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.



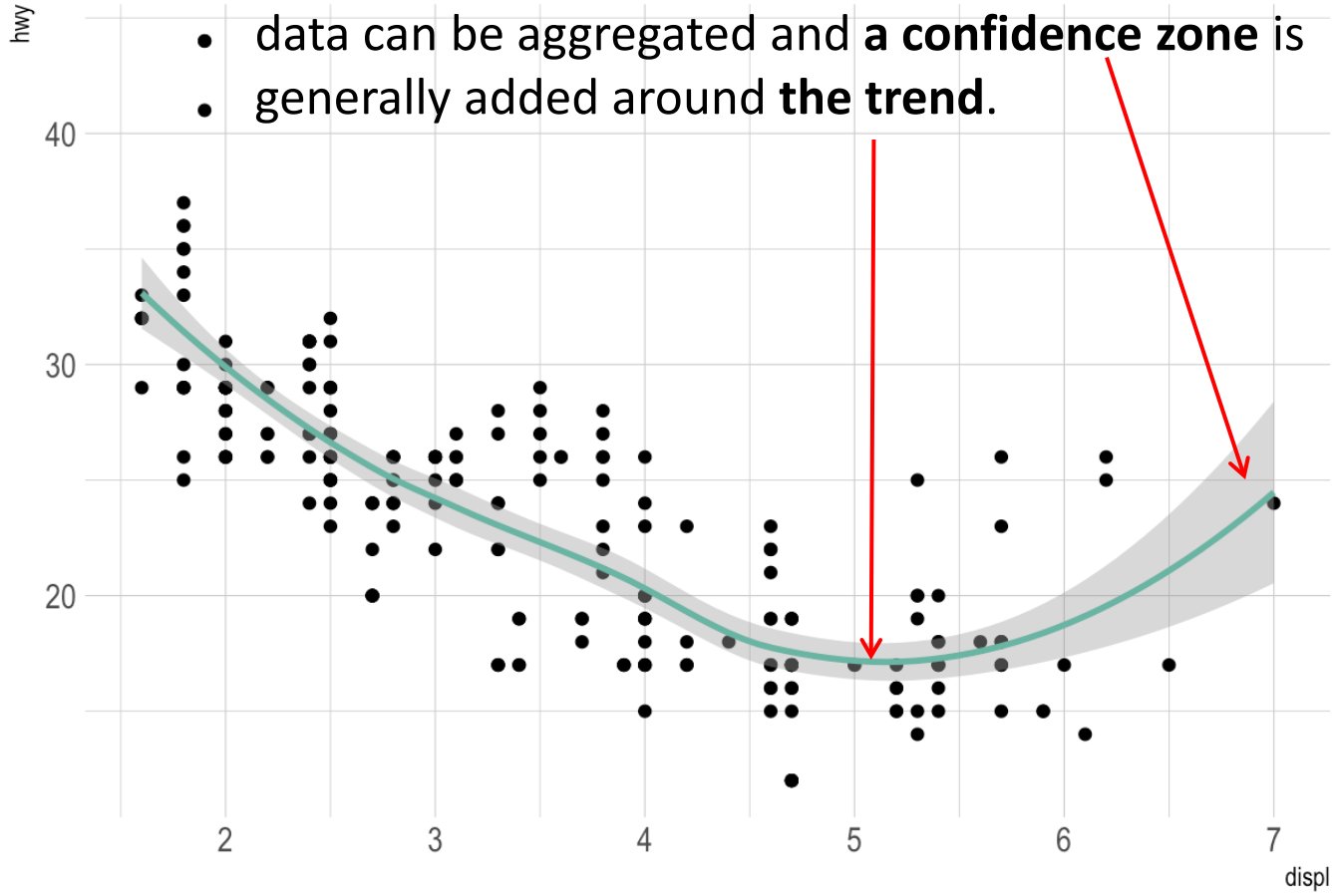
Monthly submissions to the preprint server bioRxiv, shown as dots connected by lines. The lines do not represent data and are only meant as a guide to the eye. By connecting the individual dots with lines, we emphasize that there is an order between the dots: each dot has exactly one neighbor that comes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

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

Visualizing Time Series Data

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**.

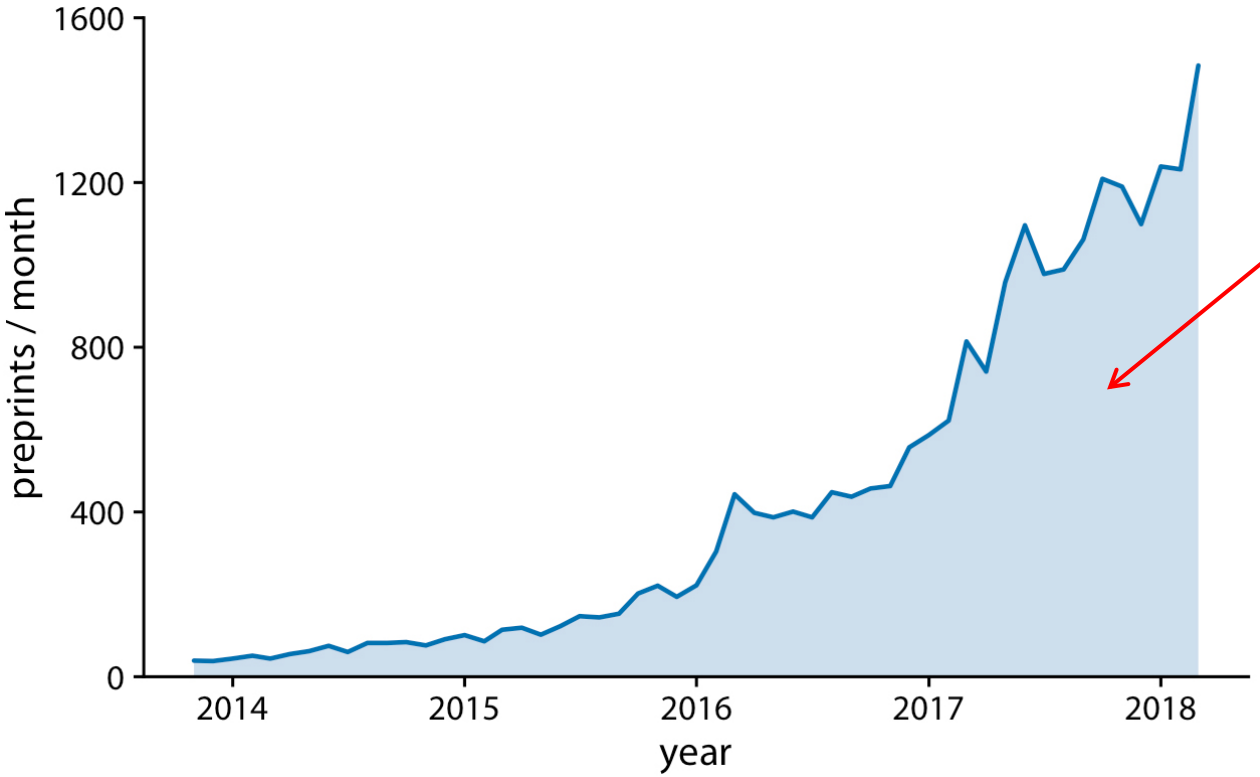


Source: https://www.data-to-viz.com/

Visualizing Time Series Data

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.

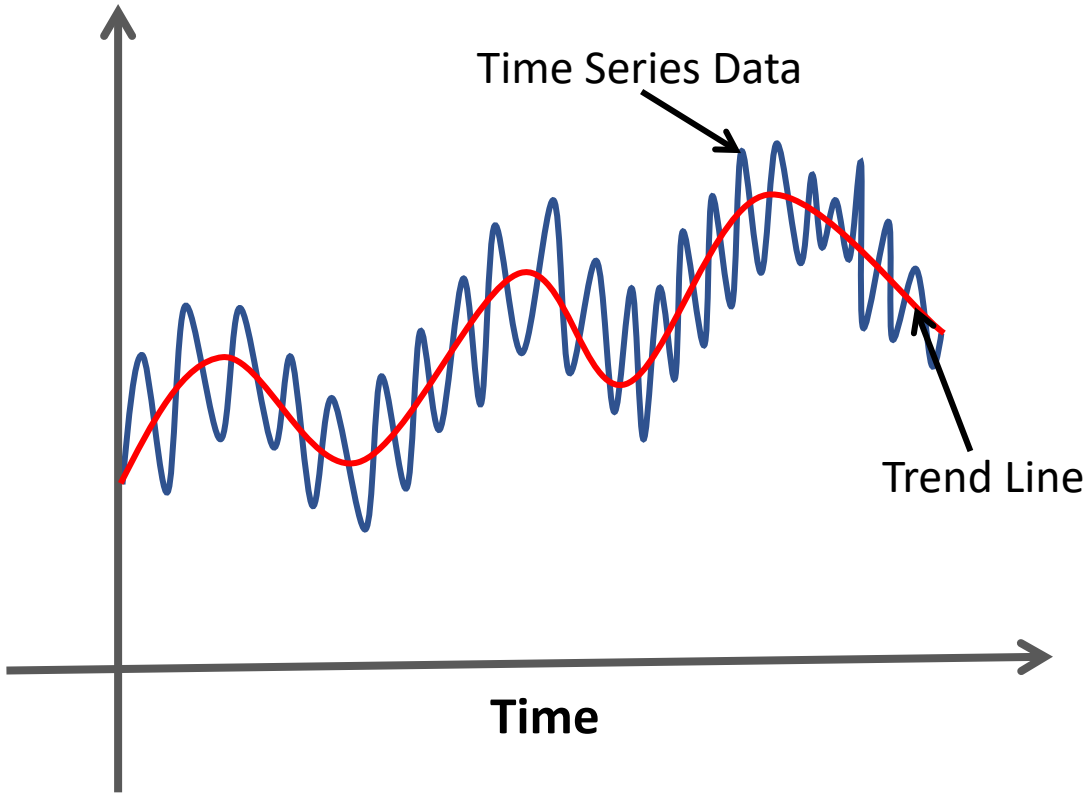


The area between the x axis and the line is filled.

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.

Smoothing Time Series Data

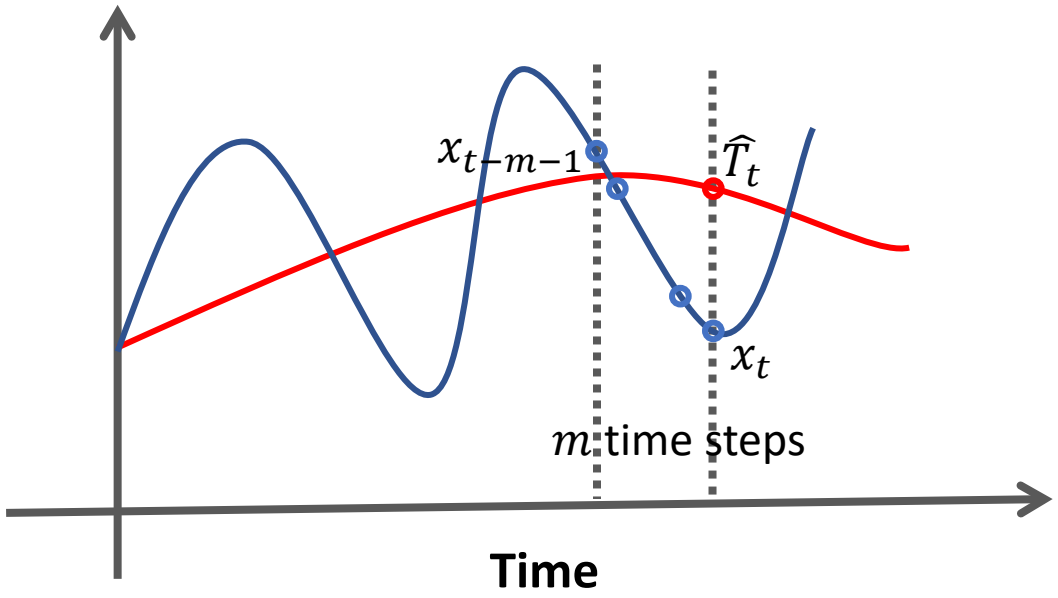


- Smooth out short-term fluctuations
- Highlight longer-term trends or cycles.

Purpose: to help improve understanding of the time series

Smoothing Time Series Data

Simple Moving Average

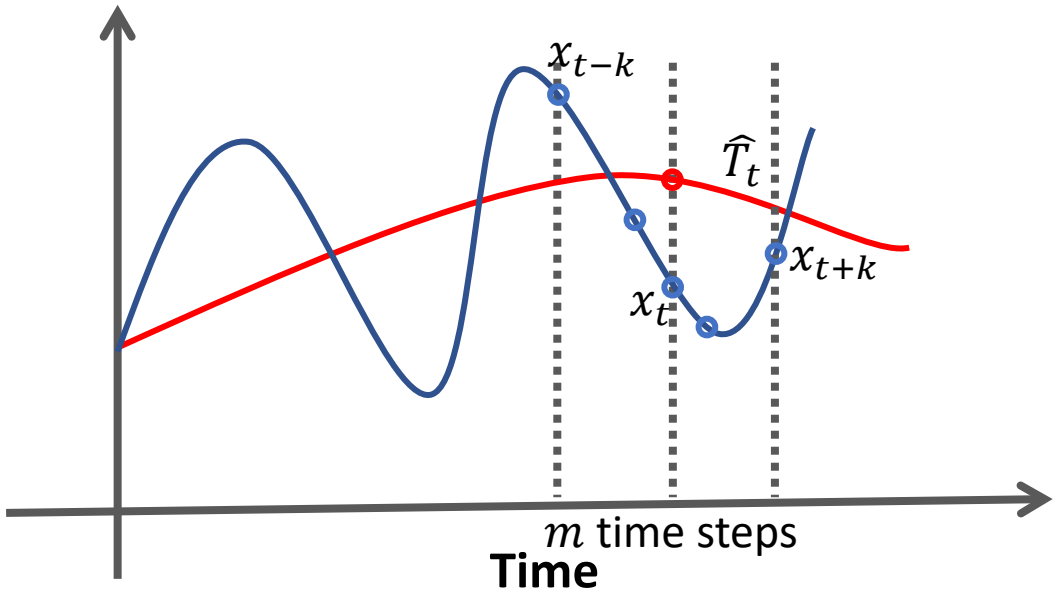


Financial Applications: the unweighted mean of the previous n data.

$$\widehat{T}_t = \frac{1}{m}\sum_{i=0}^{m-1} x_{t-i}$$

Smoothing Time Series Data

Simple Moving Average



Science and Engineering: the mean is taken from an equal number of data on either side of a central value.

$$\widehat{T}_t = \frac{1}{m} \sum_{i=-k}^{k} x_{t+k}$$

where $m = 2k + 1$

Smoothing Time Series Data

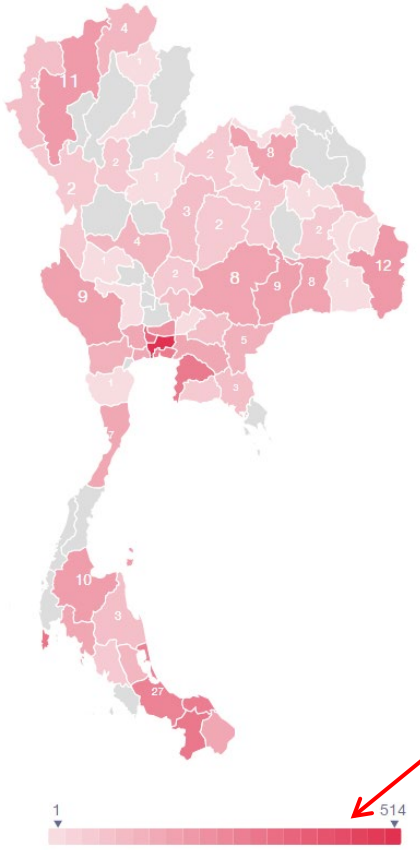
Simple Moving Average

Example: Different moving averages applied to the residential electricity sales data.
Source: https://otexts.com/fpp2/moving-averages.html

Map and Network

Choropleth Map

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



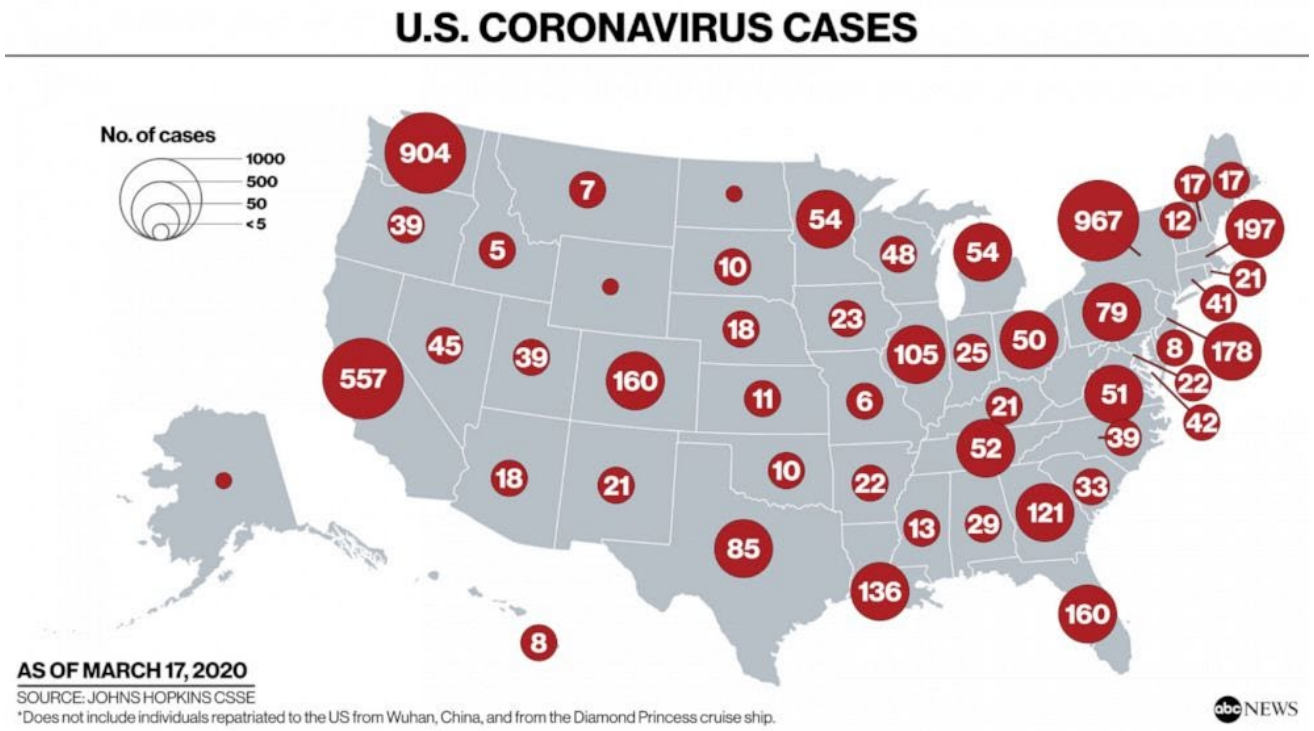
The number of COVID-19 patients in every Thailand provinces. The data is officially reported on March 30, 2020.
Source: https://covid19.workpointnews.com/ (Retrieved on 30/3/2020)

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

Map and Network

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.



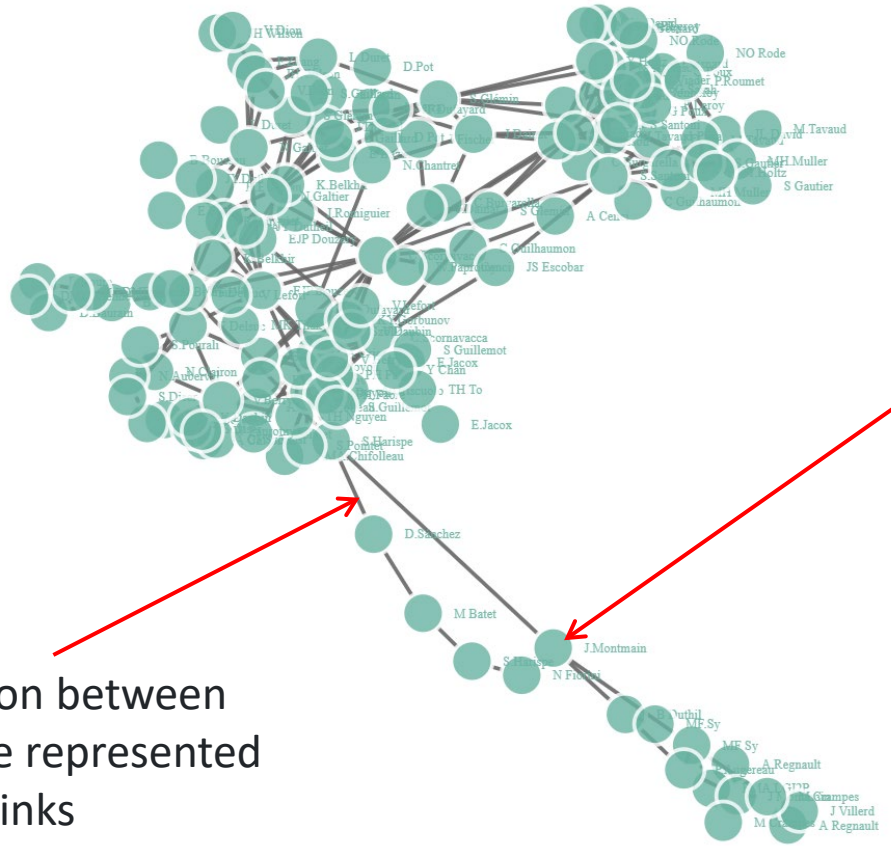
The number of COVID-19 cases in the U.S. reported on March 17, 2020.

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

Network Diagram

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



Entity is represented by a Node

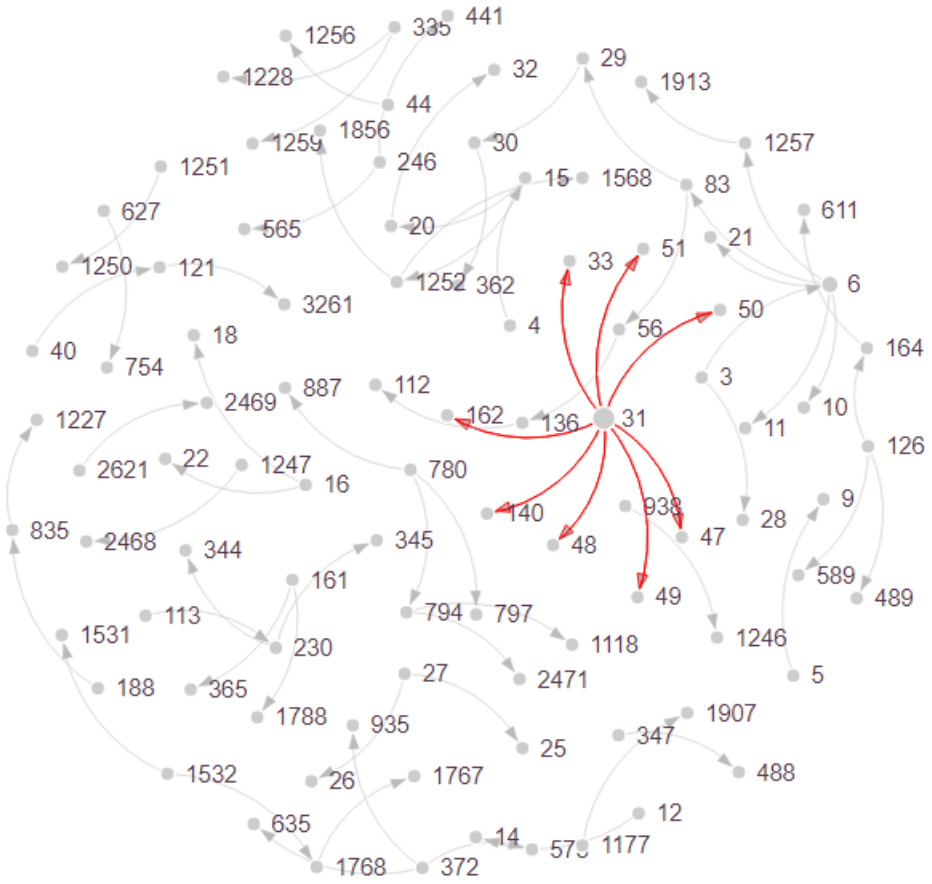
Connection between nodes are represented through links

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

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.

References

- Joanes, D. N. and C. A. Gill (1998). “Comparing measures of sample skewness and kurtosis.” In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1, pp. 183–189.
- Raymond H. Meyers, Ronald E. Walpole and, Sharon L. Meyers, and Keying E. Ye (2012). *Probability & Statistics for Engineers & Scientists*. 9th. USA: Prentice Hall.
- Westfall, Peter H. (2014). “Kurtosis as Peakedness” In: *The American Statistician* 68.3, pp. 191–195.
- Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O’Reilly Media, Inc.
- Martinez, Mendy L., Martinez, Angel R. and Solka, Jeffrey L. (2017). Exploratory Data Analysis with MATLAB. USA: CRC Press.
- Stephanie Glen. "Q Q Plots: Simple Definition & Example" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/q-q-plots/