NUS National University of Singapore

GET1030

Computers and the humanities

Lecture 3

Visualizing data

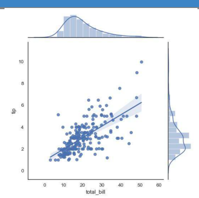
Dr Miguel Escobar Varela

Learning Objectives

1. To describe the main types of visualizations used today
2. To identify different approaches to data visualization
3. To identify the potential for bias in visualizations
4. To offer critical perspectives on data visualization

Lecture 3

Visualizing data



Part 1: Most common scientific data visualizations today

To describe the main types of visualizations used today

What is a visualization?

Representing numerical and categorical data with graphical elements (color, shape, position, size)

What is it useful for?

- To give an overview of the data
- As a first step for further research

Charts in this session

Boxplots
Barplots
Lineplots
Scatterplots
Histograms
KDE plots
Violinplots
Joint plots

Boxplot

Index	Actors
0	3
1	4
2	7
3	8
4	9
5	9
6	9
7	$\boxed{10}$
8	11
9	11
10	11
11	12
12	13
13	16
14	17

Description of a univariate distribution (one variable)
For this toy example: number of actors required for a theatre play

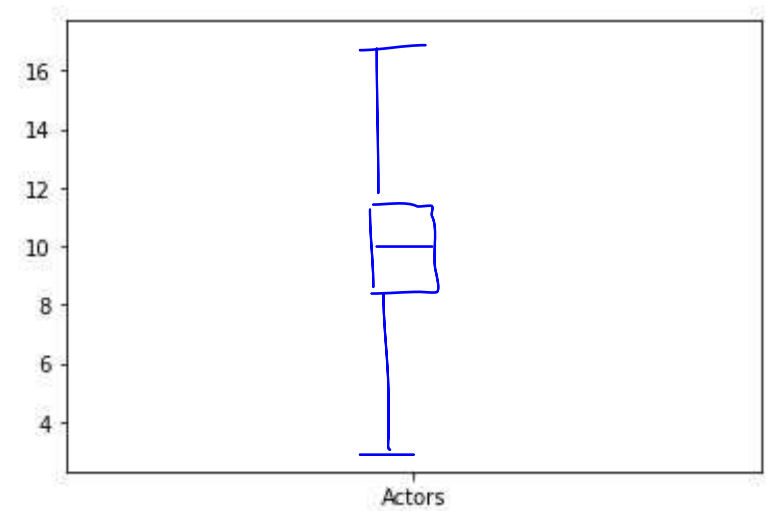
1st quartile =

Median =

3rd quartile =

Minimum =

Maximum =



Quartile
include median

Boxplot (Outliers)

Index	Actors
0	3
1	4
2	7
3	8
4	9
5	9
6	9
7	10
8	11
9	11
10	11
11	12
12	13
13	16
14	17

Calculating outliers using Tukey's rule

IQR = 11.5 − 8.5
= 3

1st quartile = 8.5

Median = 10

3rd quartile = 11.5

Minimum = 3

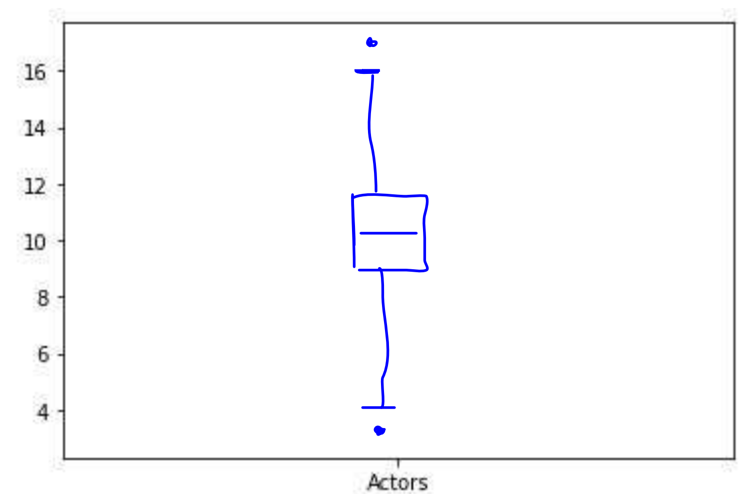
Maximum = 17

Upper bound
= 11.5 + 4.5
= 16

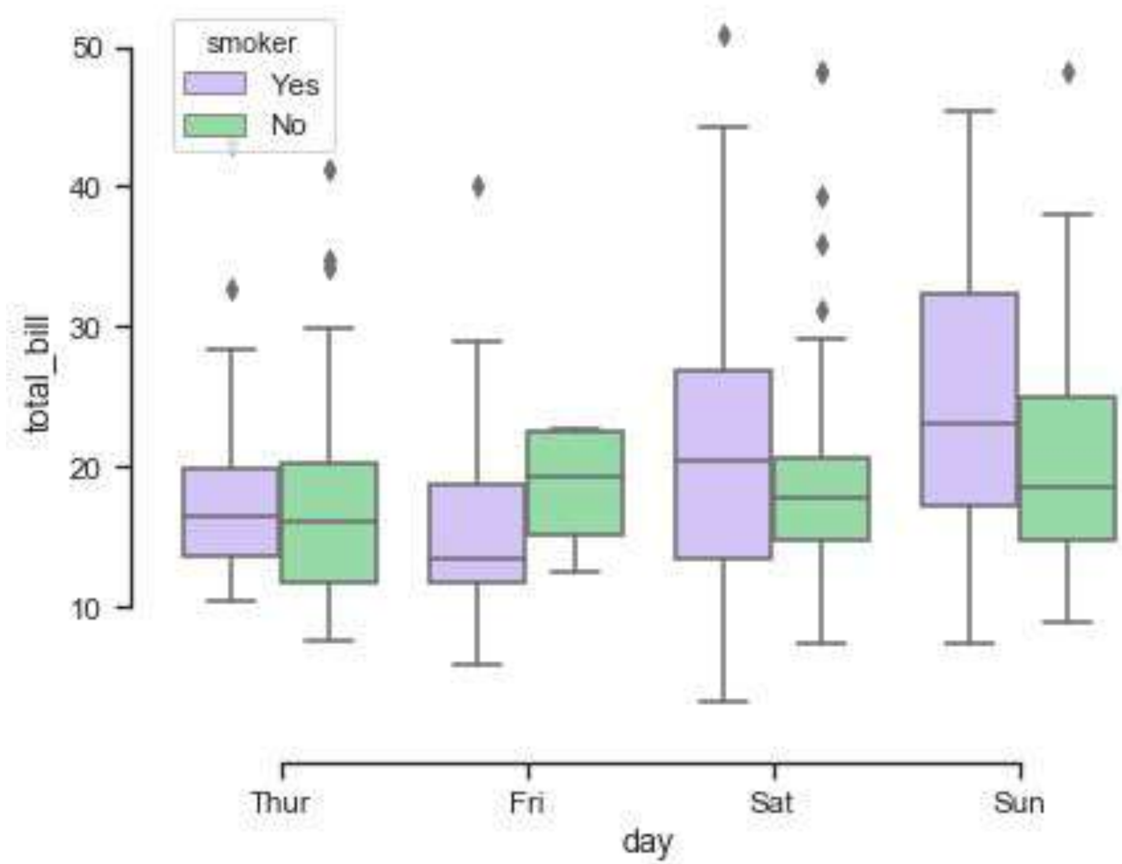
Lower
= 8.5 − 4.5
= 4

1.5 x IQR = 4.5

IQR = 75th percentile - 25th percentile



Categories in boxplots



How many variables are represented in this graph?

no.	total	Day	Smoker
0	17	Thurs	Y
1	17	T	Y
2			
⋮			

Standard deviation in bar charts

Measure of the dispersion of the data
Square root of the **variance**
Variance is the average of the squared differences from the mean

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Index	Actors
0	2
1	3
2	4
3	4
4	4
5	5
6	5
7	5
8	6
9	7

Mean = 10

Variance = $\sigma^2 = \frac{(10-3)^2 + (10-4)^2 + \ldots}{10}$

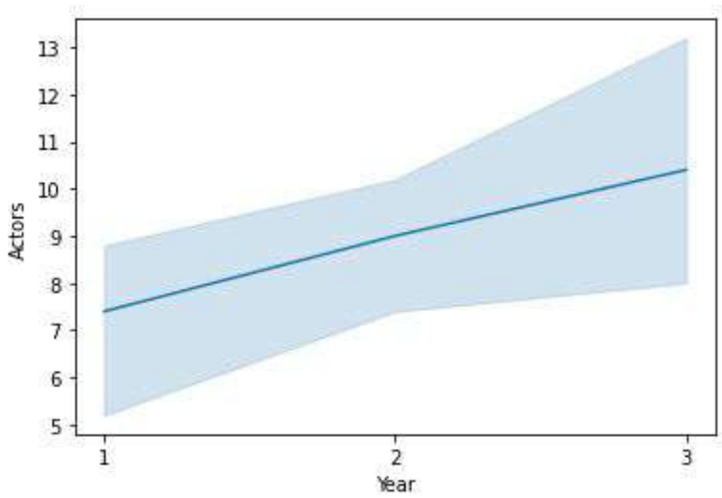
$= 13.46$

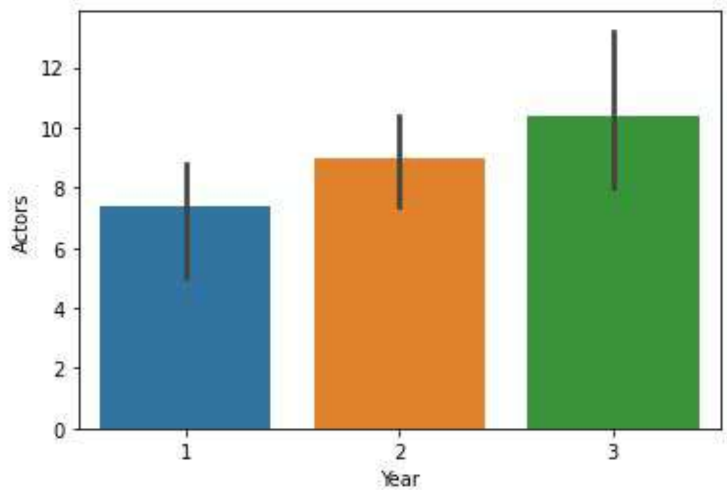
Standard Deviation = 3.66

Standard deviation in lineplots

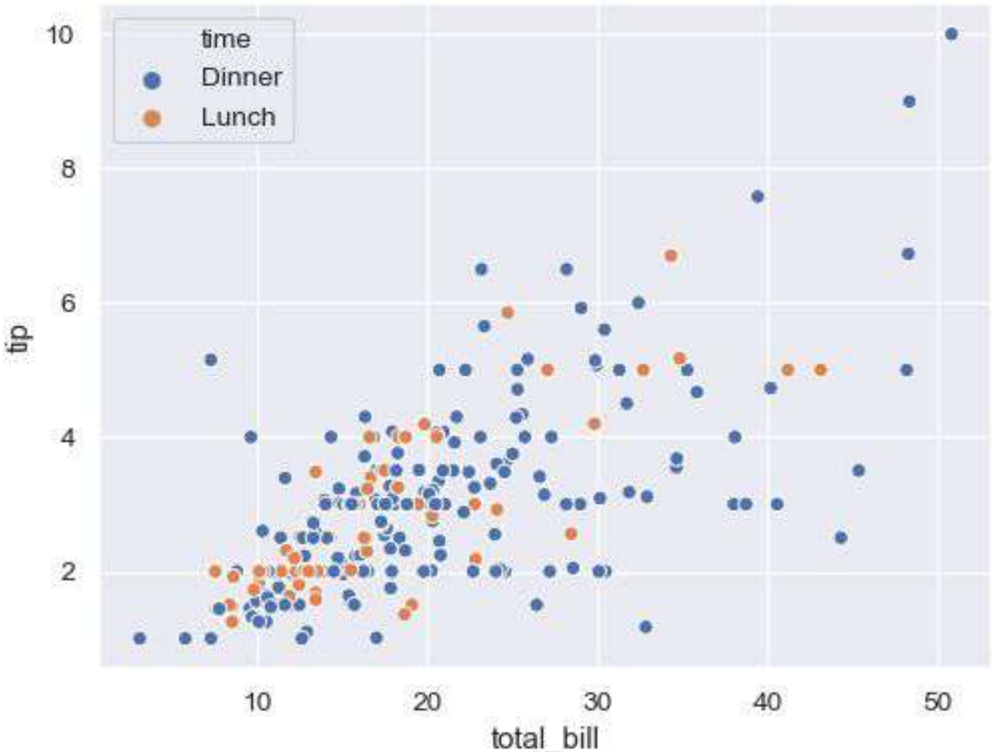
Consider this example of actors in plays over time

	Actors	Year
0	8	1
1	9	1
2	9	1
3	8	1
4	3	1
5	11	2
6	6	2
7	10	2
8	9	2
9	9	2
10	8	3
11	8	3
12	8	3
13	15	3
14	13	3





Scatterplot

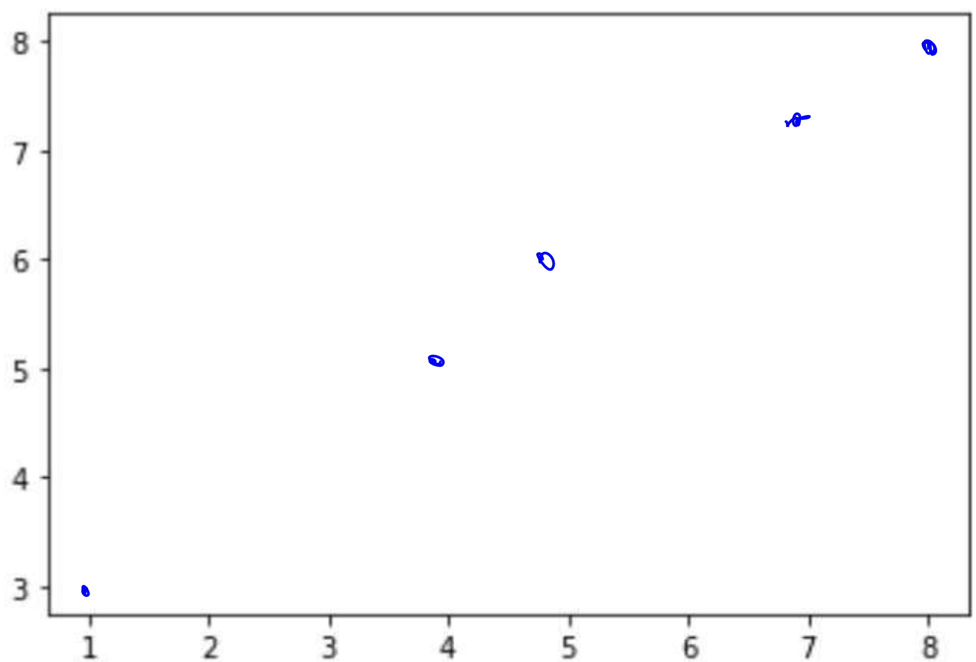


Shows the numerical relationship between two variables. Color can be used to indicate categorical variables.

Scatterplot

Shows the relationship between two variables.

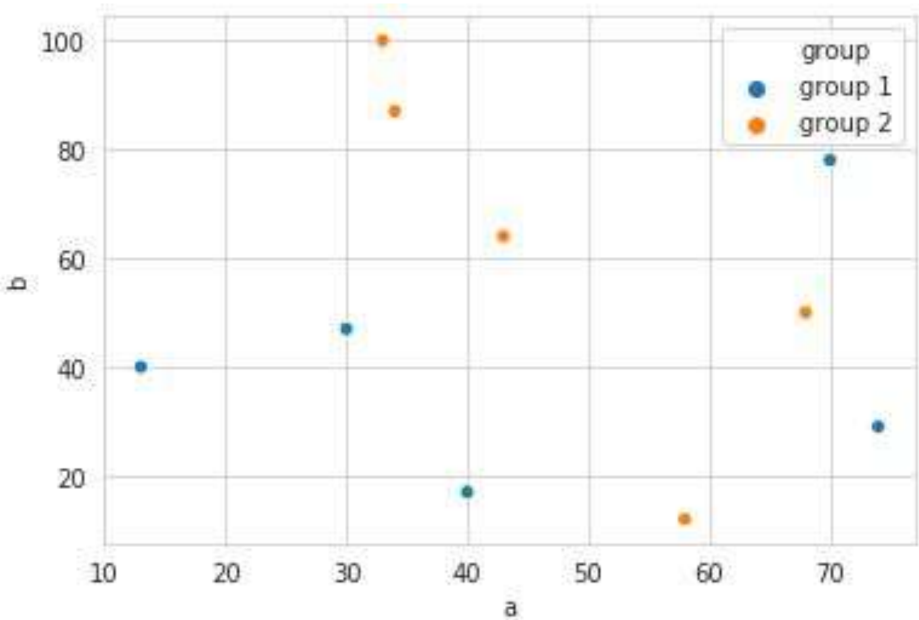
a	b
1	3
4	5
5	6
7	7
8	8



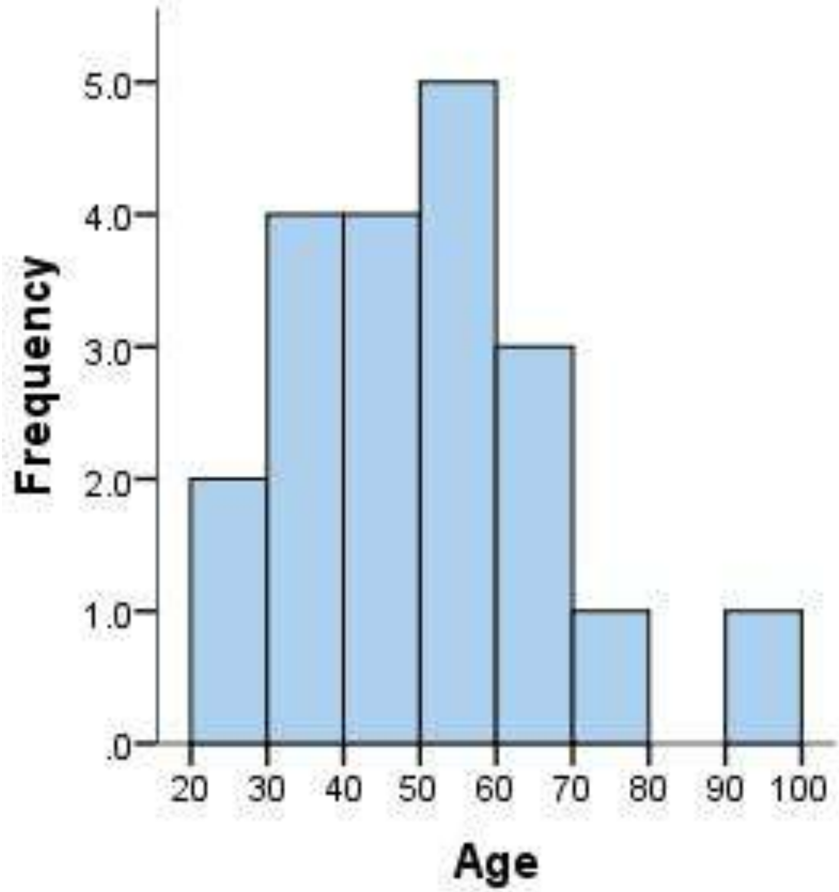
Scatterplot

Shows the numerical relationship between two variables. Color can be use categorical variables.

a	b	group
40	17	group 1
30	47	group 1
74	29	group 1
70	78	group 1
13	40	group 1
43	64	group 2
68	50	group 2
58	12	group 2
33	100	group 2
34	87	group 2



Histogram

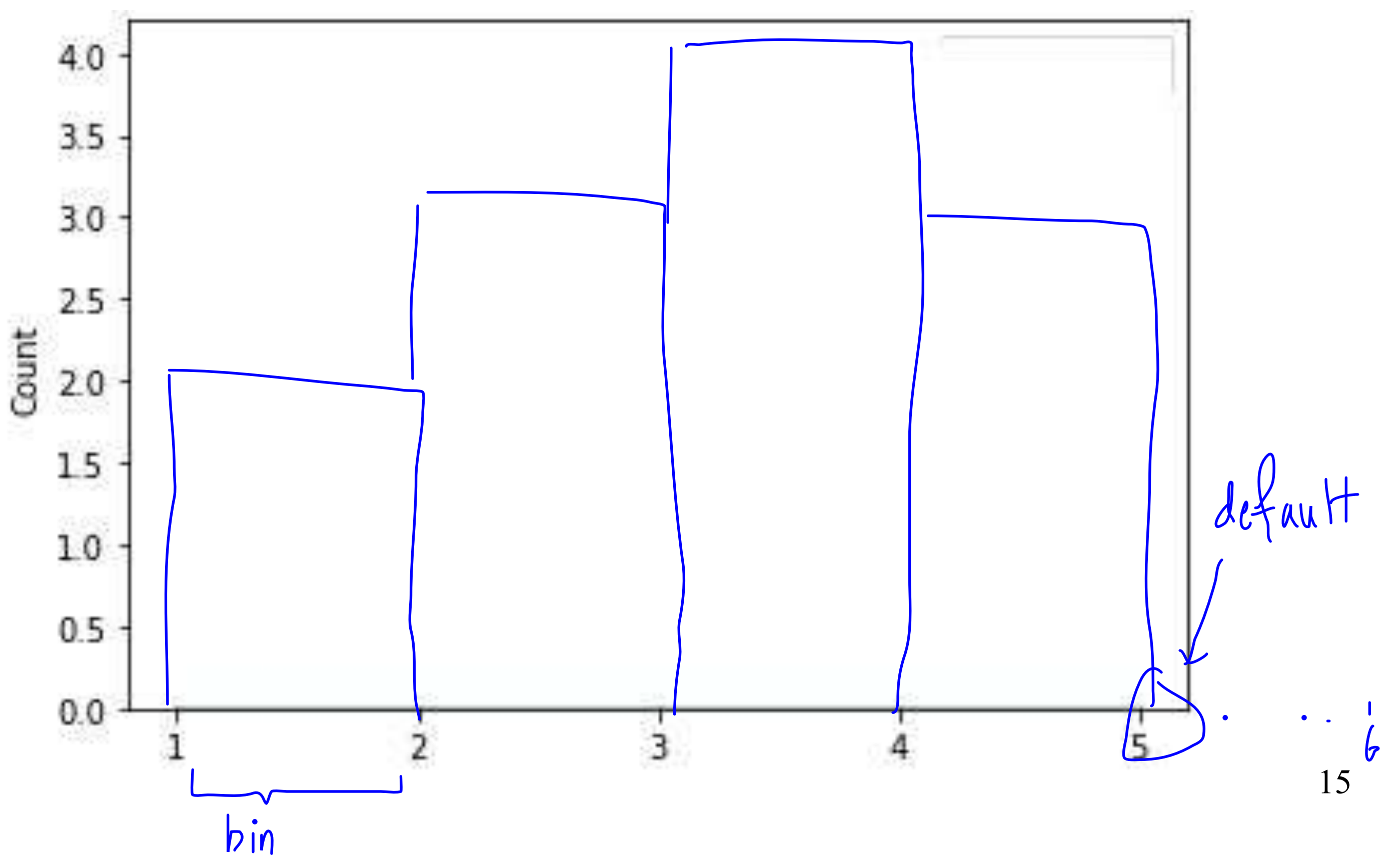


Each bar groups numbers into ranges. Taller bars show that more data falls in that range. It displays the shape and spread of the data.

Histogram

Consider this example of stars given to films in a group of reviews.

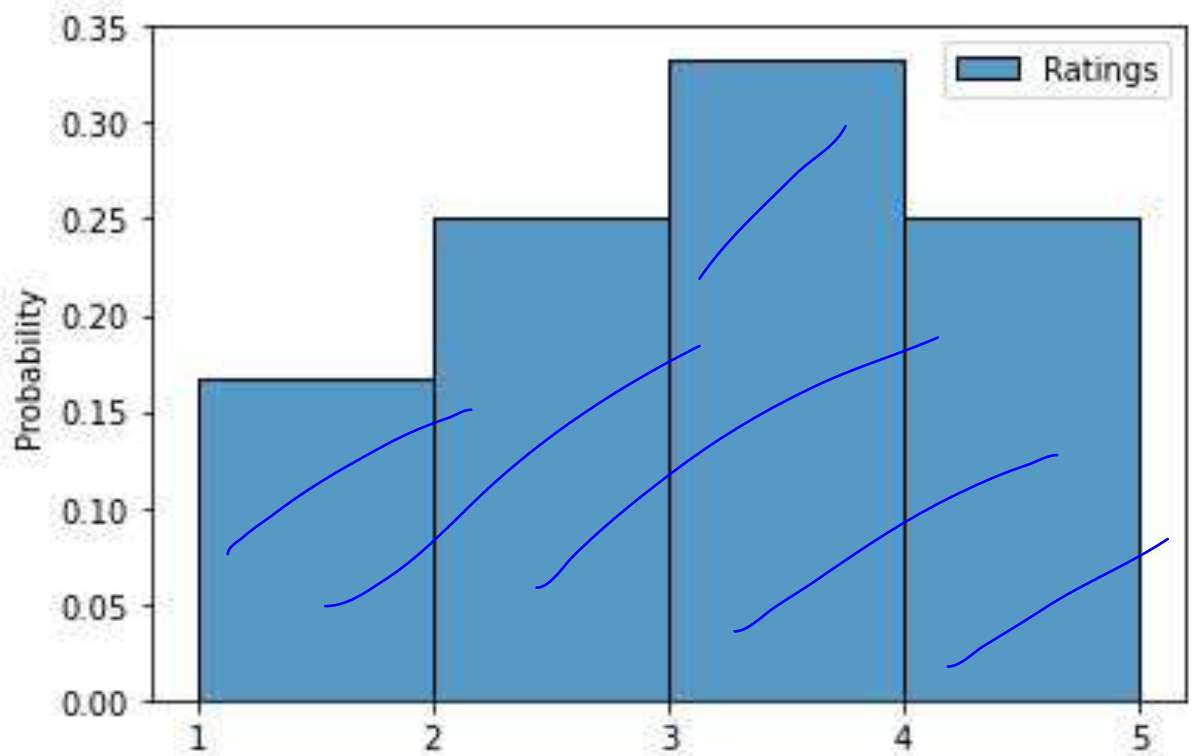
Index	Ratings
0	1
1	1
2	2
3	2
4	2
5	3
6	3
7	3
8	3
9	4
10	4
11	5



Histogram (relative frequency)

Consider this example of stars given to films in a group of reviews.

Index	Ratings
0	1
1	1
2	2
3	2
4	2
5	3
6	3
7	3
8	3
9	4
10	4
11	5

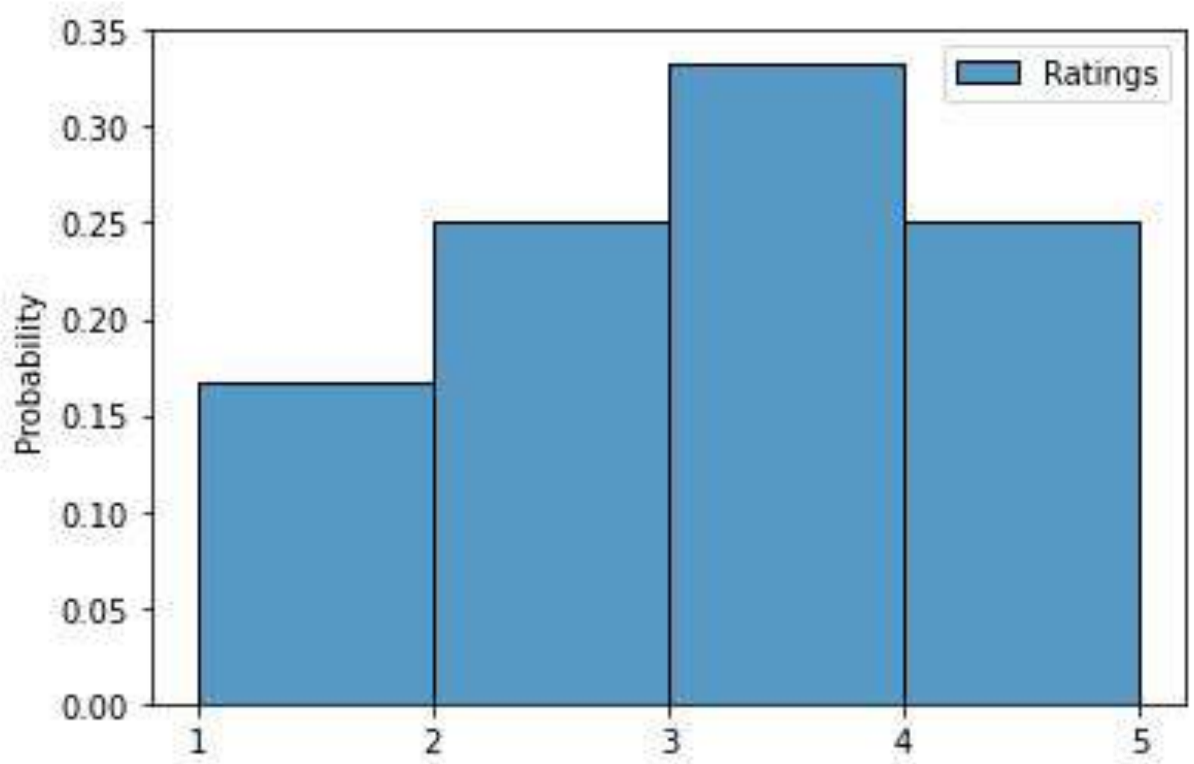


$= 1$

Histogram (relative frequency)

Consider this example of stars given to films in a group of reviews.

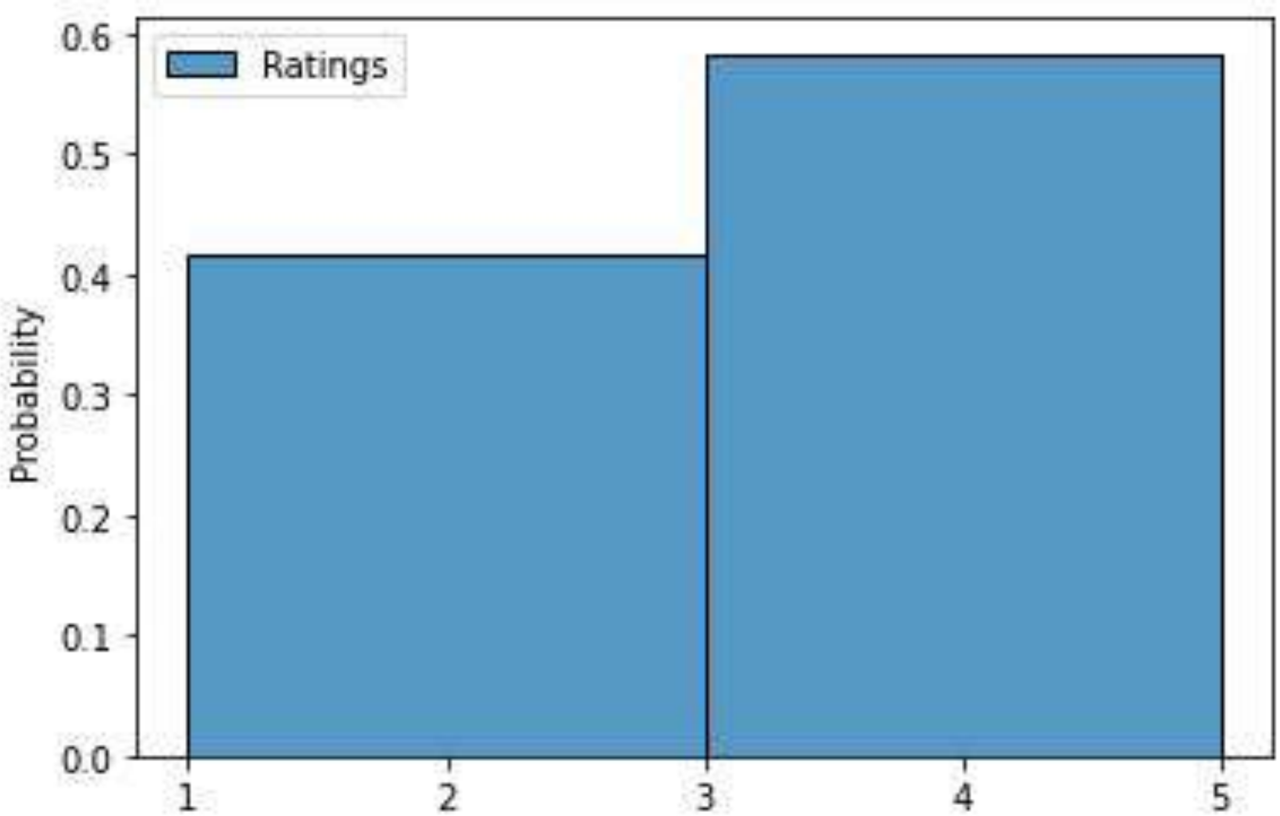
Index	Ratings
0	1
1	1
2	2
3	2
4	2
5	3
6	3
7	3
8	3
9	4
10	4
11	5



Histogram (bin size)

Consider this example of stars given to films in a group of reviews.

Index	Ratings
0	1
1	1
2	2
3	2
4	2
5	3
6	3
7	3
8	3
9	4
10	4
11	5



Histogram

Exploring Histograms, an essay by Aran Lunzer and Amelia McNamara

Bin-breaks: Why these bins?

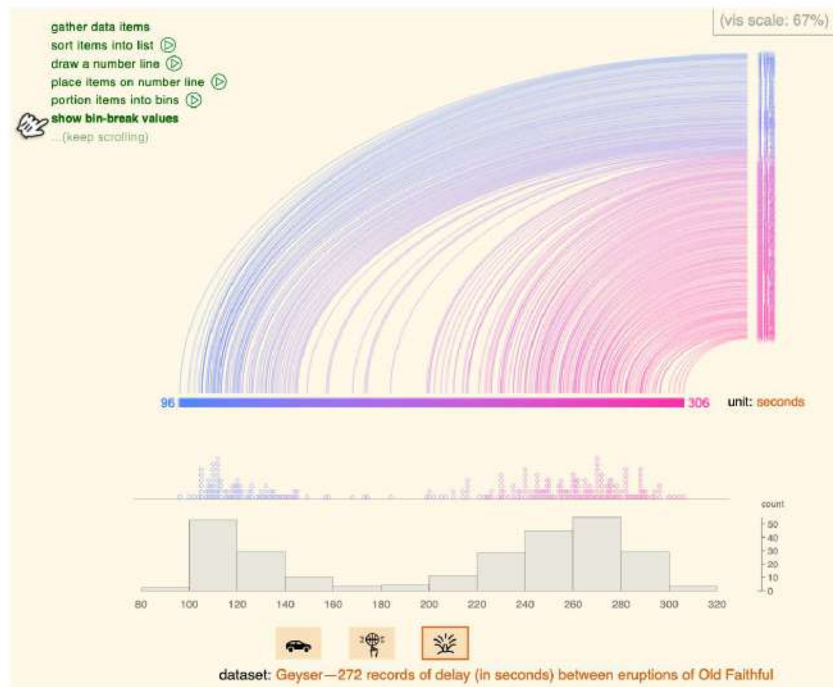
For a start, you probably noticed that the histograms shown for our sample datasets have different numbers of bins. This is because we used Sturges' formula, a common method for estimating the number of bins for a histogram, given the size of a dataset.

Given a suggested number of bins, how did we then decide the precise values for the bin boundaries (the so-called "breaks")? Again we used a common method: look for nearby round numbers. This is why the breaks for "MPG" are all multiples of 5, and those for "NBA" are multiples of 2.

For those two datasets, the bins turn out to cover the range of the item values rather tidily. But look at the first and last bins for "Geyser". Their placement relative to the value range looks a little arbitrary, right? That's because it is.

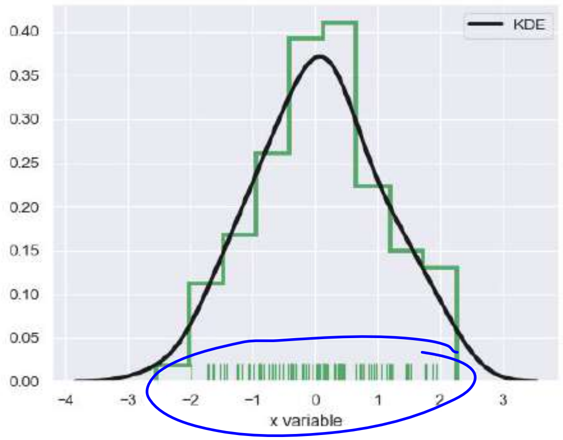
The fact is that there are few hard-and-fast rules for drawing a histogram. Instead of Sturges' formula, we could have chosen the number of bins using Scott's choice or the Freedman-Diaconis choice, among many other methods. And there's certainly no rule saying that bin-break values have to be rounded to the nearest multiple of 2 or 5.

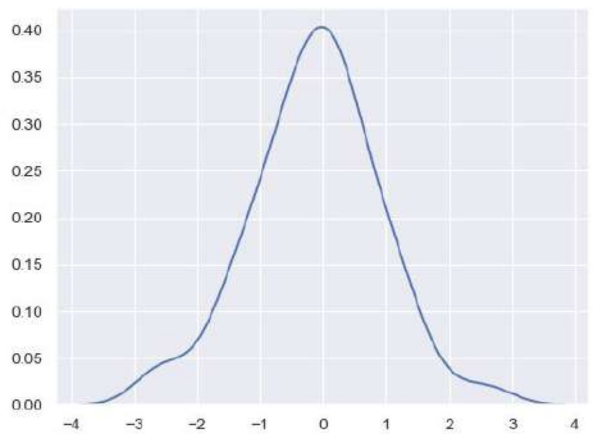
What's important is whether a given histogram is a **representative summary** of its underlying dataset. One way to judge this is to try varying the positions of the breaks, and see what impact that has on the summary that the histogram conveys.



https://tinlizzie.org/histograms/

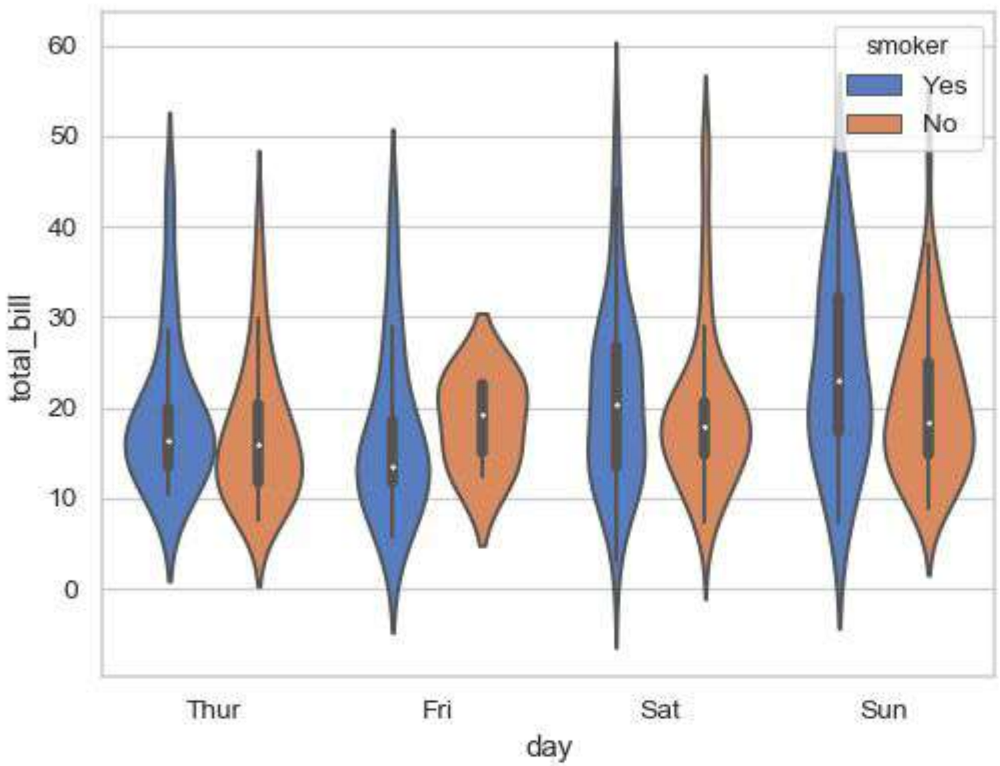
Kernel Density Estimation (KDE) plots





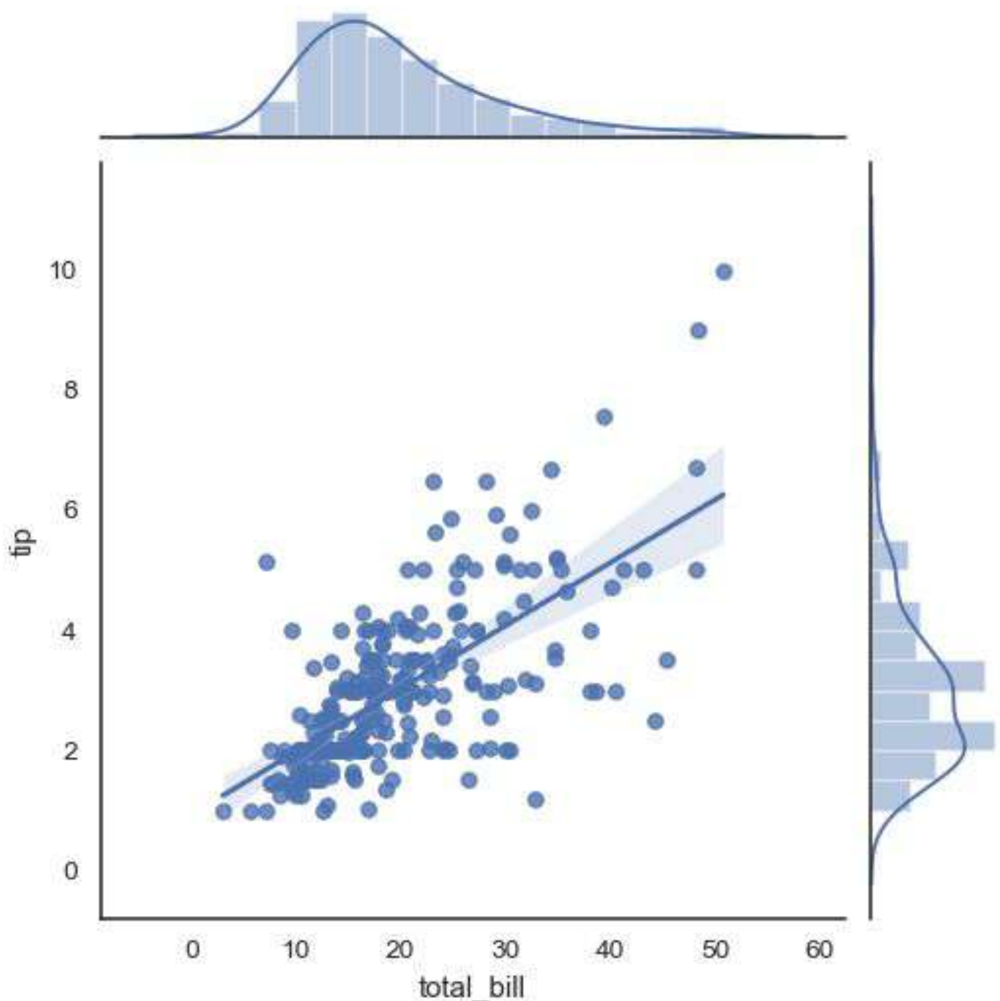
Closely related to histograms. They show a smoothed representation of the data distribution.

Violinplot



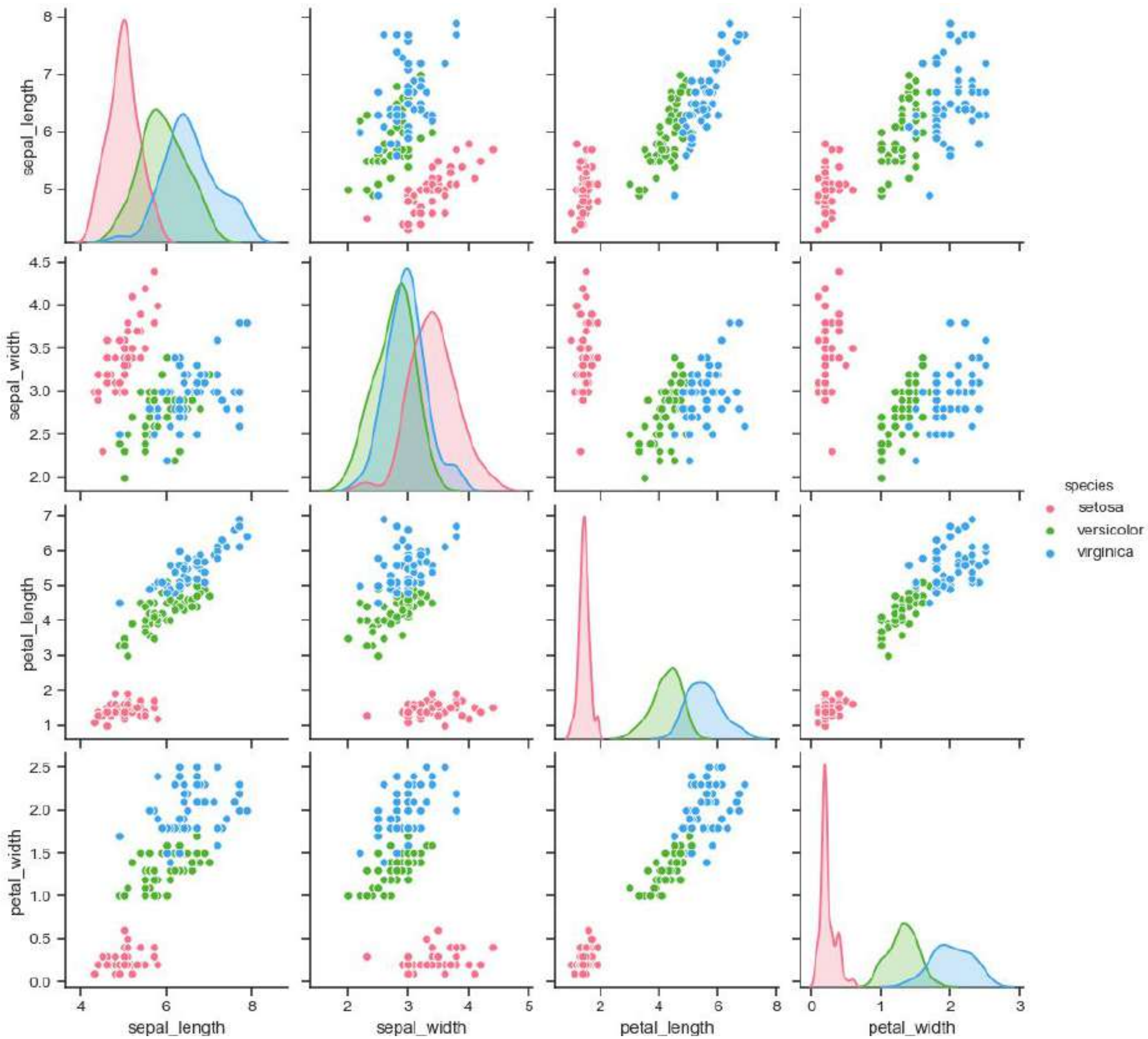
Another visual representation of the 5-number summary but also represents the distribution of values, in a way similar to a KDE.

Jointplots



Combined scatterplot with regression line, confidence interval, histogram and KDE.

Pairplots



Scatterplots and KDE of multiple variables for the same samples.

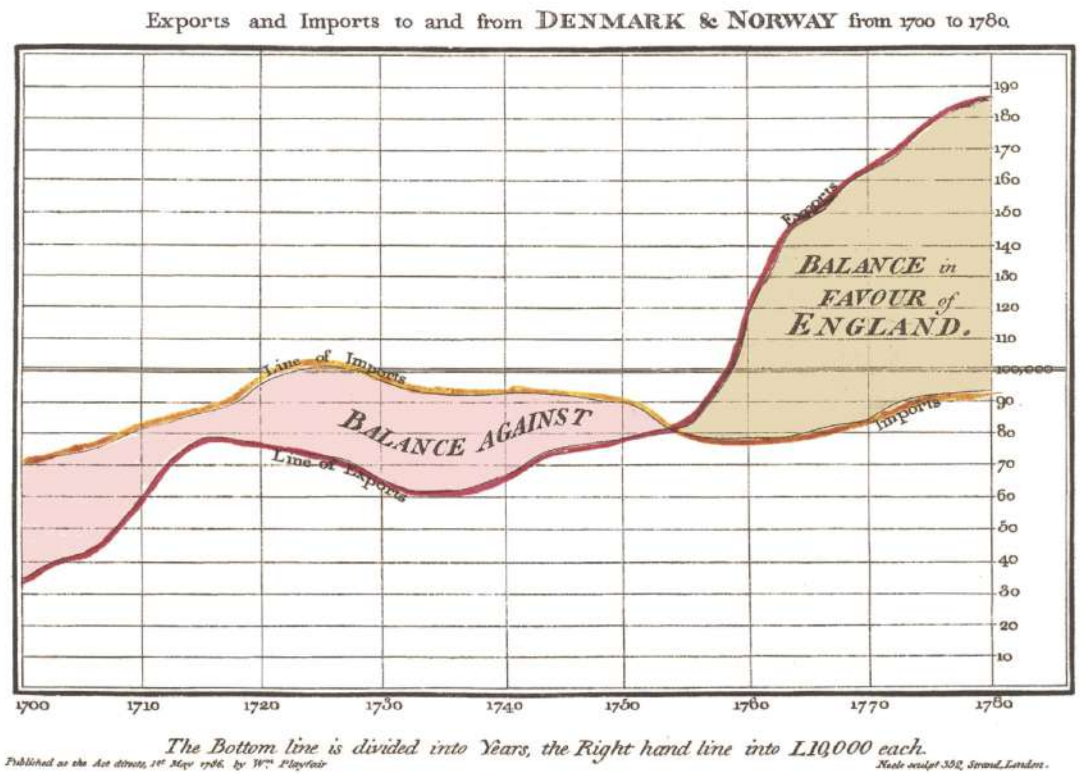
Lecture 3

Visualizing data

Part 2: Different approaches to Data Visualization

GET1030

William Playfair (1759-1823)

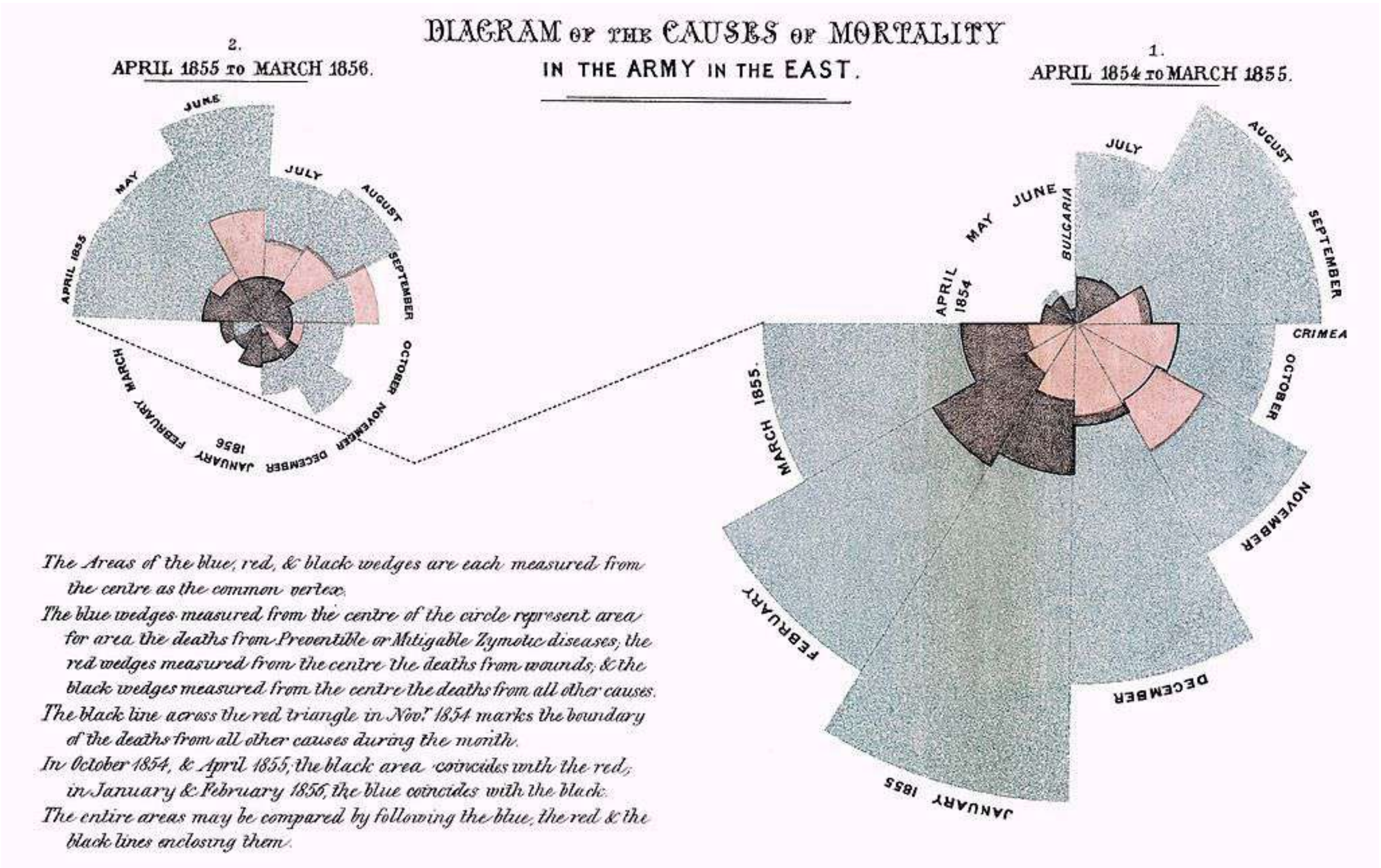


From *The Commercial and Political Atlas; Representing, by Means of Stained Copper-Plate Charts, the Exports, Imports, and General Trade of England, at a Single View* (1785)

John Snow (1813-1858)

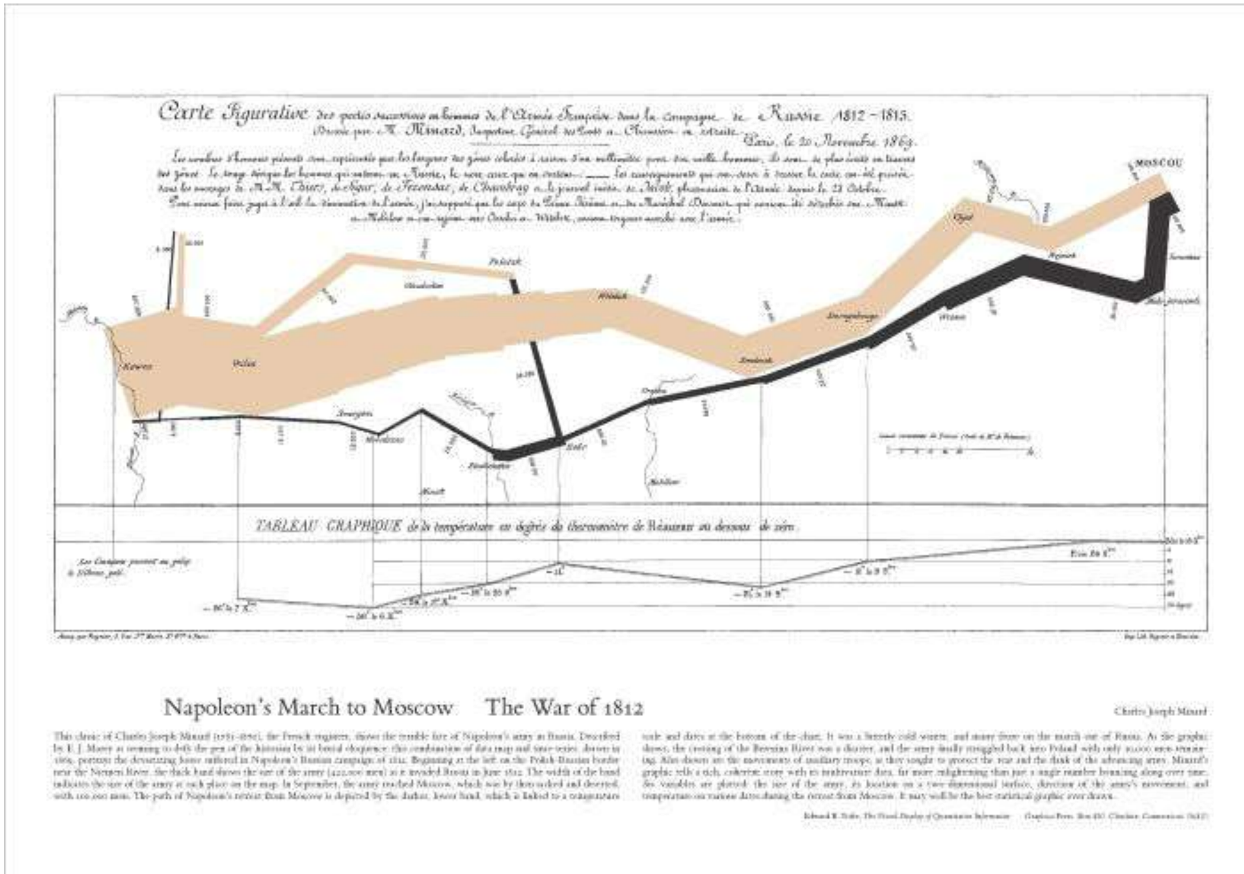
John Snow’s map of cholera outbreaks and wells (1854)
Digital version by Robin Wilson (2013)
http://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/

Florence Nightingale (1820-1910)



‘Coxcom’ (polar area graph) visualization of the cause of death in the army (1850s Crimea War)
*preventable causes, wounds, accidents

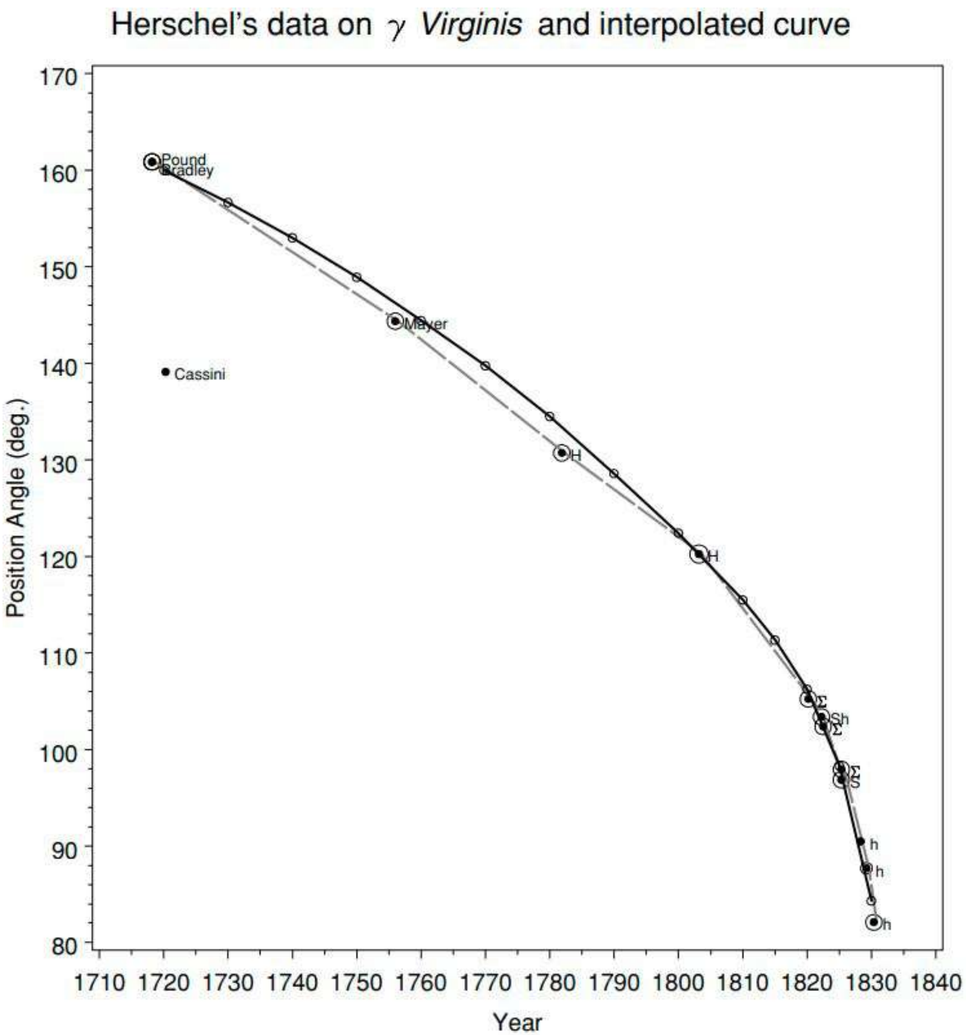
Charles Joseph Minard (1781-1870)



Napoleon’s March to Moscow (published 1869)
Beginning at the Polish-Russian border, the thick band shows the size of the army at each position. The path of Napoleon's retreat from Moscow is depicted by the dark lower band, which is tied to temperature and time scales.

GET1030

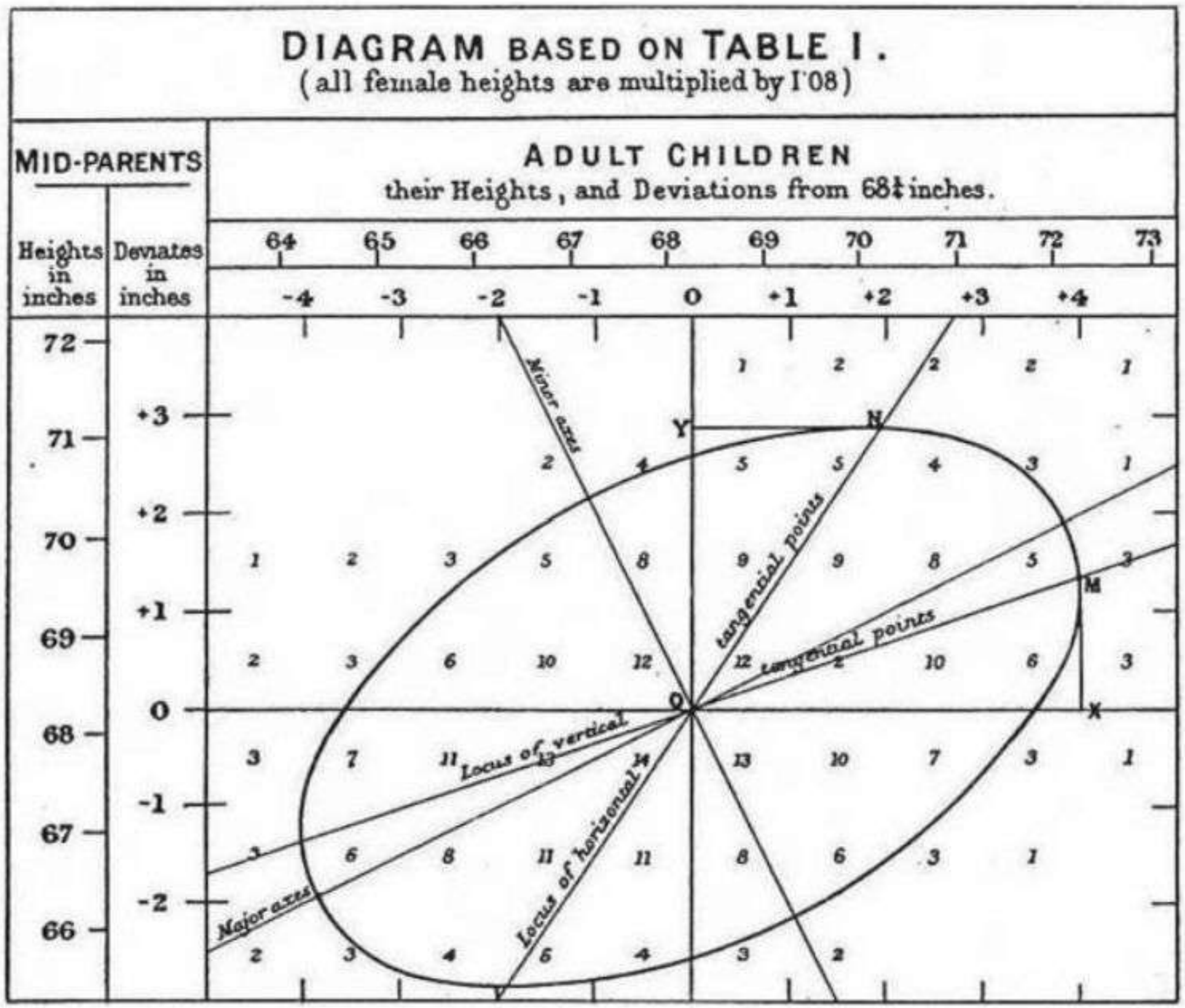
The scatterplot



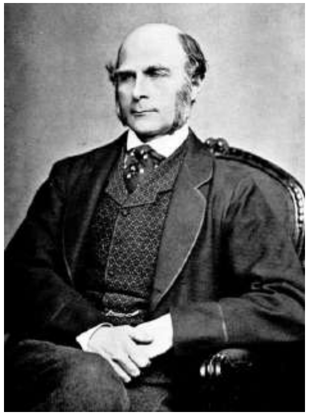
*This is a reconstruction based on data from Herschell's 1833 paper, "On the Investigation of the Orbits of Revolving Double Stars" by Friendly and Denis (2005).

Sir John Frederick William Herschel (1972-1871)

The scatterplot



Francis Galton's (1986) smoothed correlation diagram for the data on heights of parents and children, showing one ellipse of equal frequency.

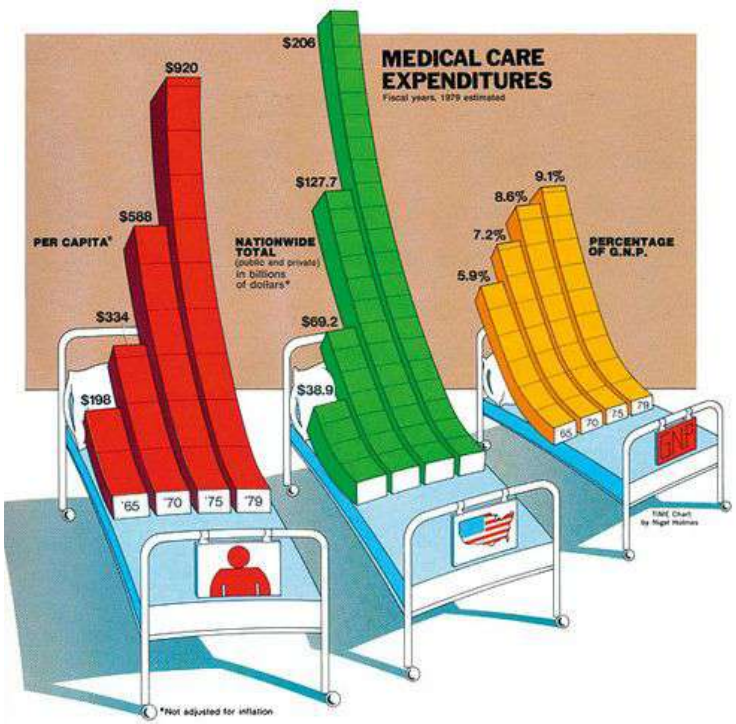


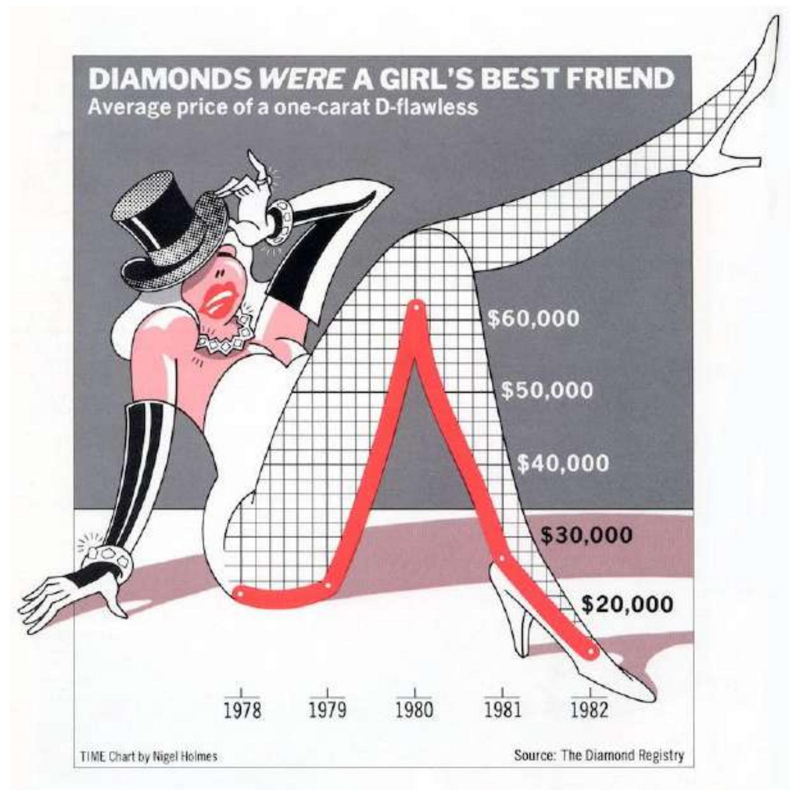
Francis Galton (1822-1911)

The ‘Infoviz’ Approach

Statisticians often disagree with a type of visualization aesthetic common in the news, which was most influential developed by Nigel Holmes, while working for TIME magazine in the 1970s.

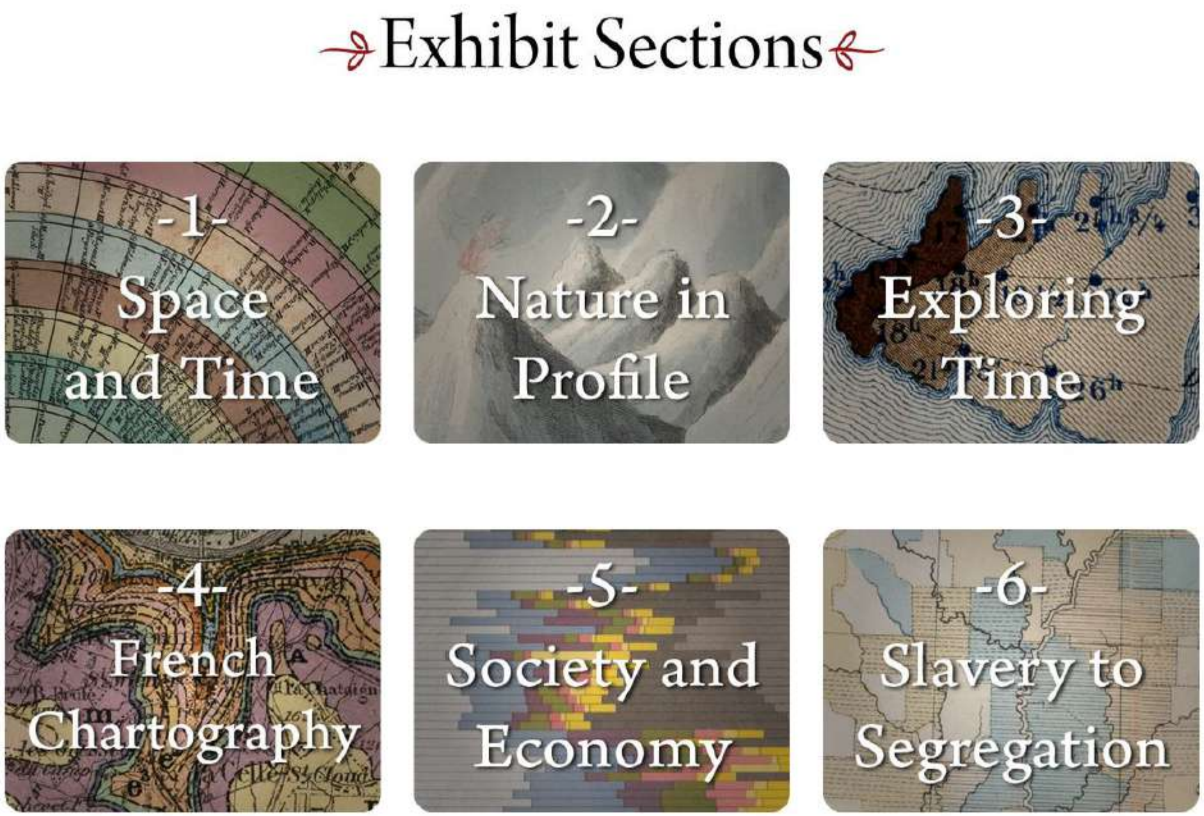
*as we’ll see later, this is the kind of graph that Tufte would classify as ‘chartjunk’





GET1030

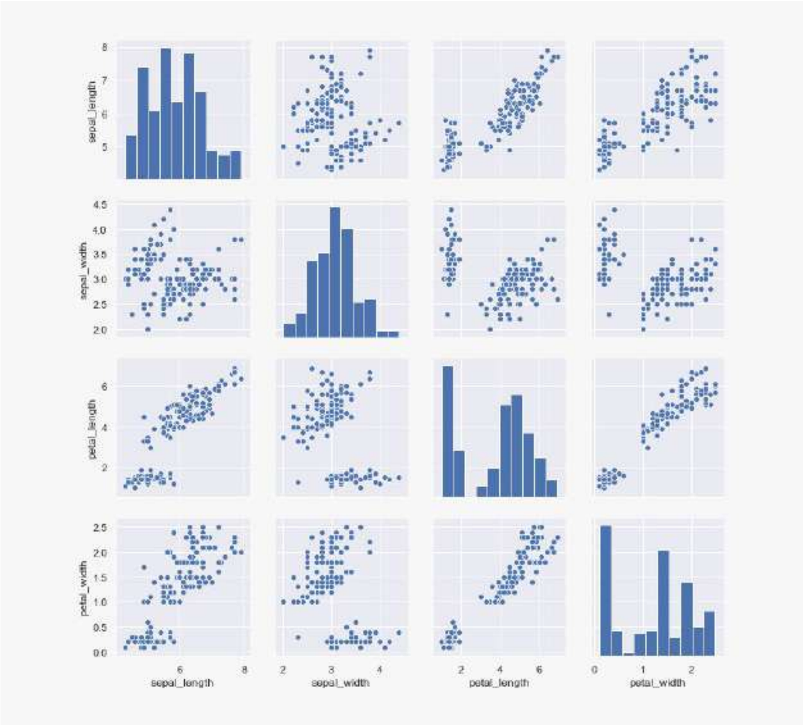
Additional resources



https://exhibits.stanford.edu/dataviz

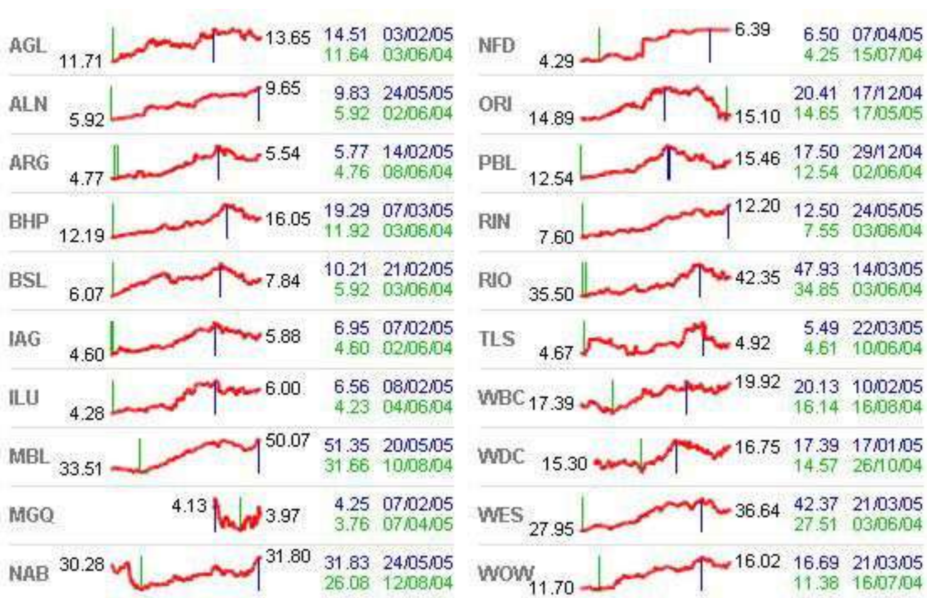
Exploratory data analysis (EDA)

Theorized by John Tukey (1915-2000).
Visualizations are often central to EDA.
Often includes statistical information (error bars, confidence intervals, standard deviation)
Many different visualizations with shared axes.



Modern scientific approach to dataviz

Edward Tufte:
Against “chartjunk”
Popularized sparklines.
Famous for several concepts:
lie factor, the data-ink ratio (against decoration), small multiples and the data density of a graphic.

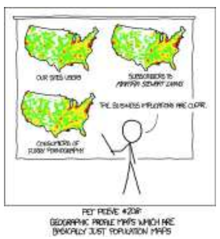


What's the purpose?

To grab attention?
To show trends?
To help scientists analyze data?

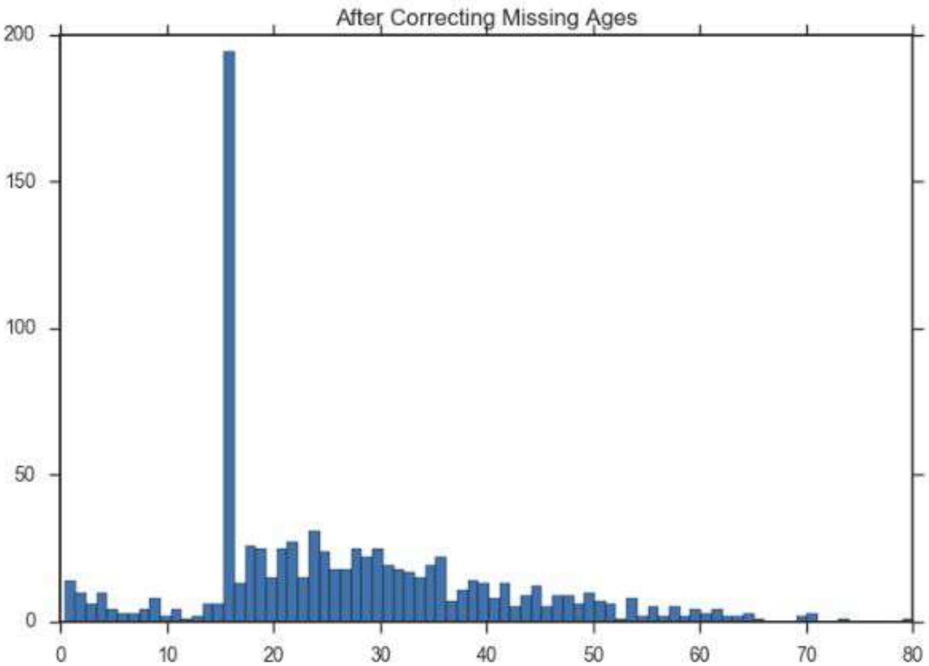
Lecture 3

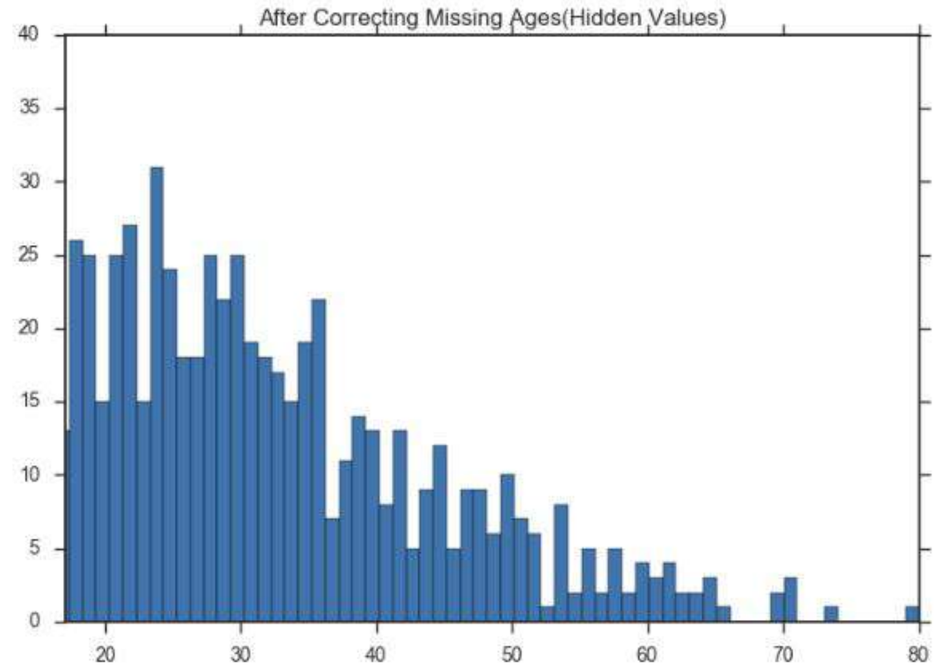
Visualizing data



Part 3: Sources of bias in data visualization

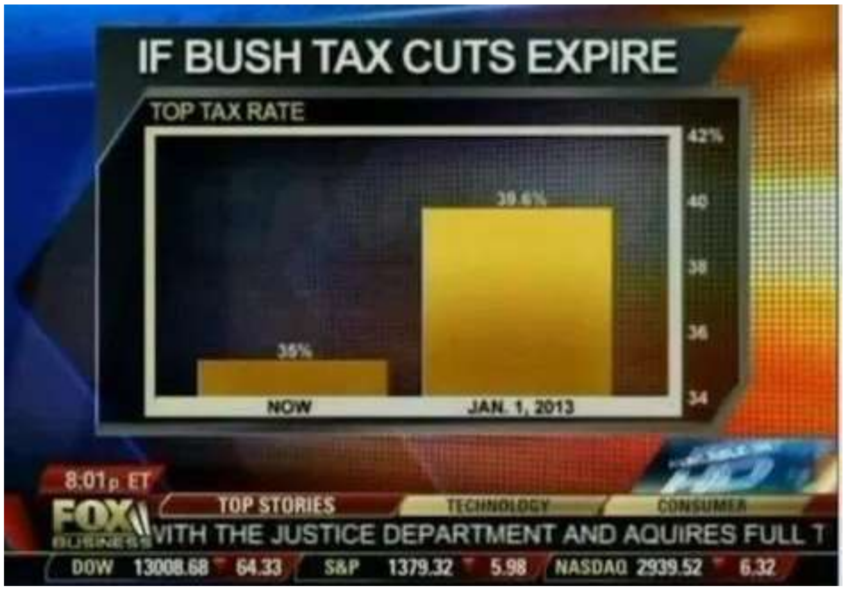
Axis Cropping

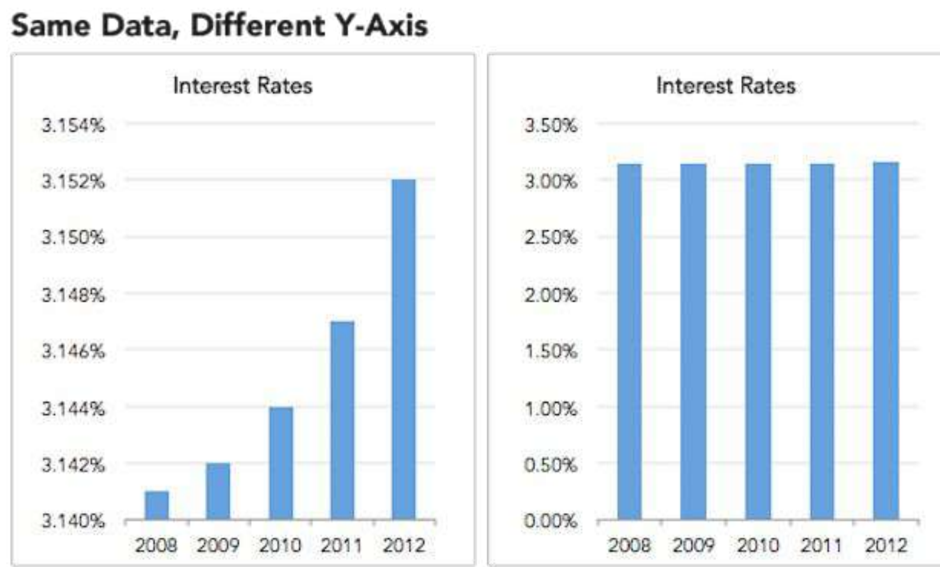




Kendall Fortney, "5 Ways Data Visualizations can Lie", *Towards data science,* https://towardsdatascience.com/5-ways-data-visualizations-can-lie-46e54f41de37

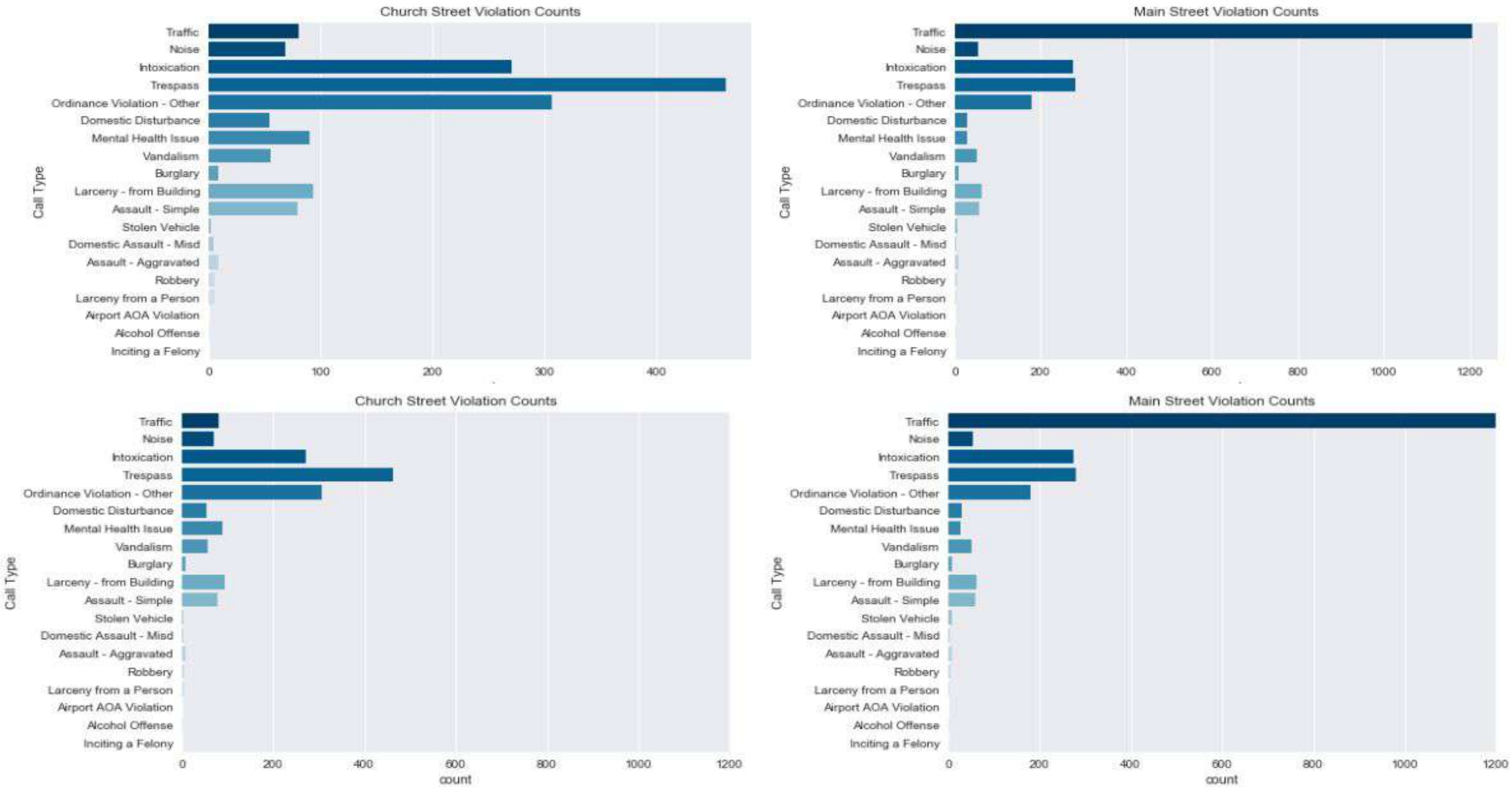
Axis Scalling





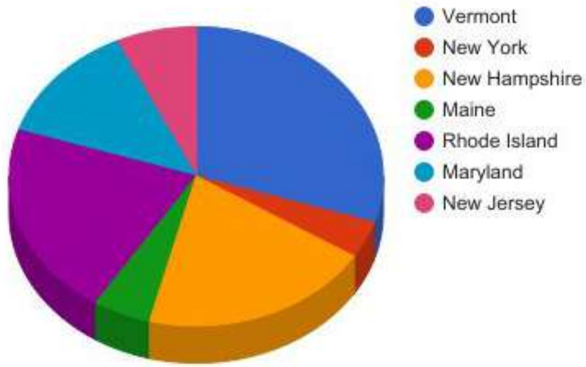
Ravi Parikh, ‘How to lie with data visualization’, *Gizmodo,* 2014, https://gizmodo.com/how-to-lie-with-data-visualization-1563576606

Axis Scaling

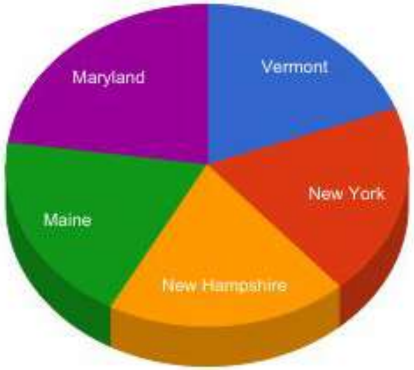


Kendall Fortney, “5 Ways Data Visualizations can Lie”, *Towards data science,* https://towardsdatascience.com/5-ways-data-visualizations-can-lie-46e54f41de37

The problem of pie charts



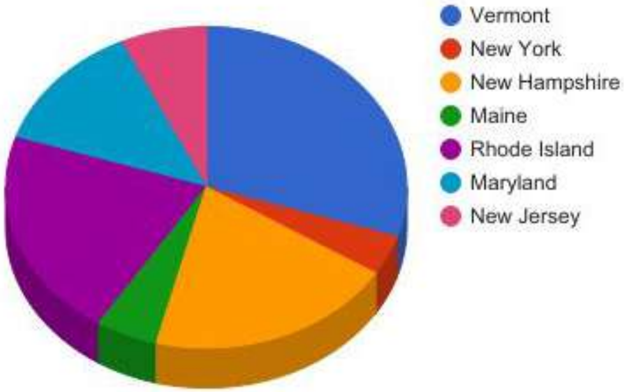
The green slice is actually equal to a quarter of the yellow one, and pink is a third of the value of the purple slice



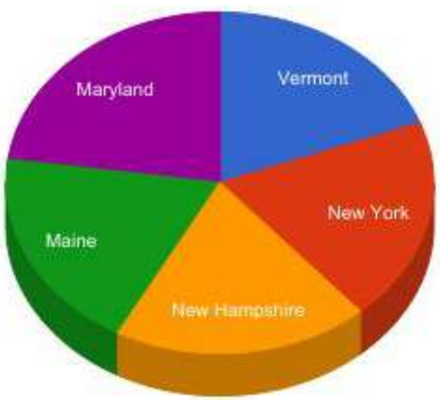
Maryland is bigger than the others (by 3%). The 3D effect visually adds more volume to NH, tricking your eyes. Without labels telling the percentages there would be little to no chance of accurately guessing it.

Kendall Fortney, “5 Ways Data Visualizations can Lie”, *Towards data science,* https://towardsdatascience.com/5-ways-data-visualizations-can-lie-46e54f41de37

The problem of pie charts



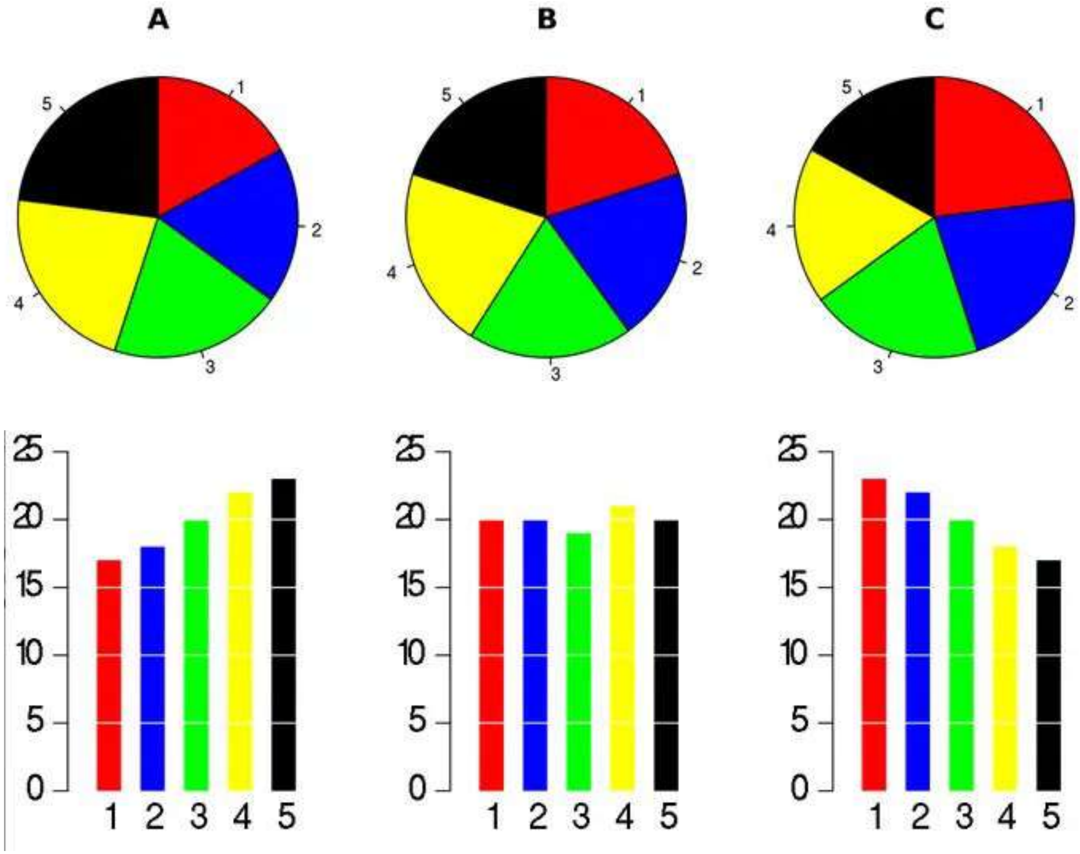
The green slice is actually equal to a quarter of the yellow one, and pink is a third of the value of the purple slice



Which is bigger? If you choose New Hampshire you would be wrong, it was actually Maryland. The 3D affect visually adds more volume to that slice, tricking your eyes. Without labels telling the percentages there would be little to no chance of accurately guessing it.

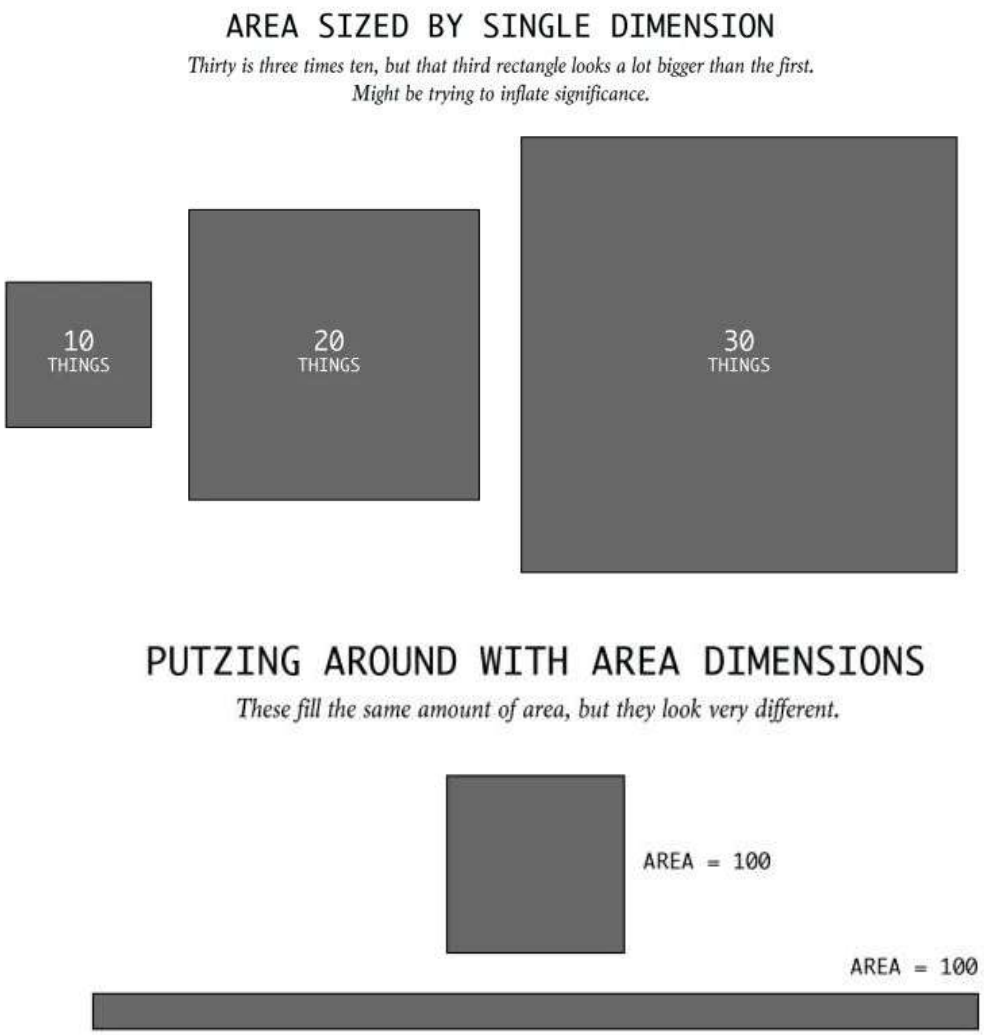
People tend to underestimate the size of acute angles (<90°) and overestimate the size of obtuse ones (>90°) (Nundy et al, 2000, text at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC25873/)

The problem of pie charts



Walt Hickey, “The Worst Chart In The World”, *Business Insider,* 2013, https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

Multiple dimensions



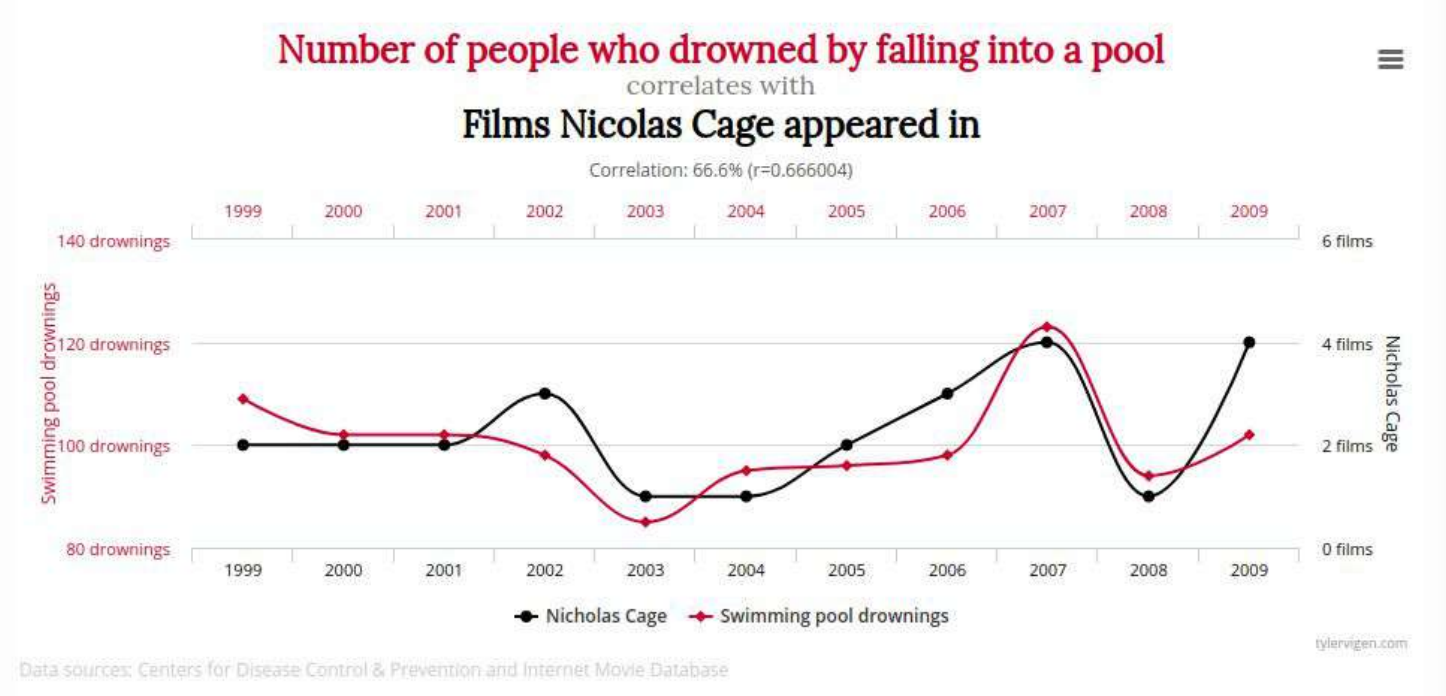
Nathan Yau, How to Spot Visualization Lies, *Flowingdata,* 2017, https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Values not normalized

Chiqui Esteban, ‘A Quick Guide to Spotting Graphics That Lie, *National Geographic,* May 2015, https://www.nationalgeographic.com/news/2015/06/150619-data-points-five-ways-to-lie-with-charts/

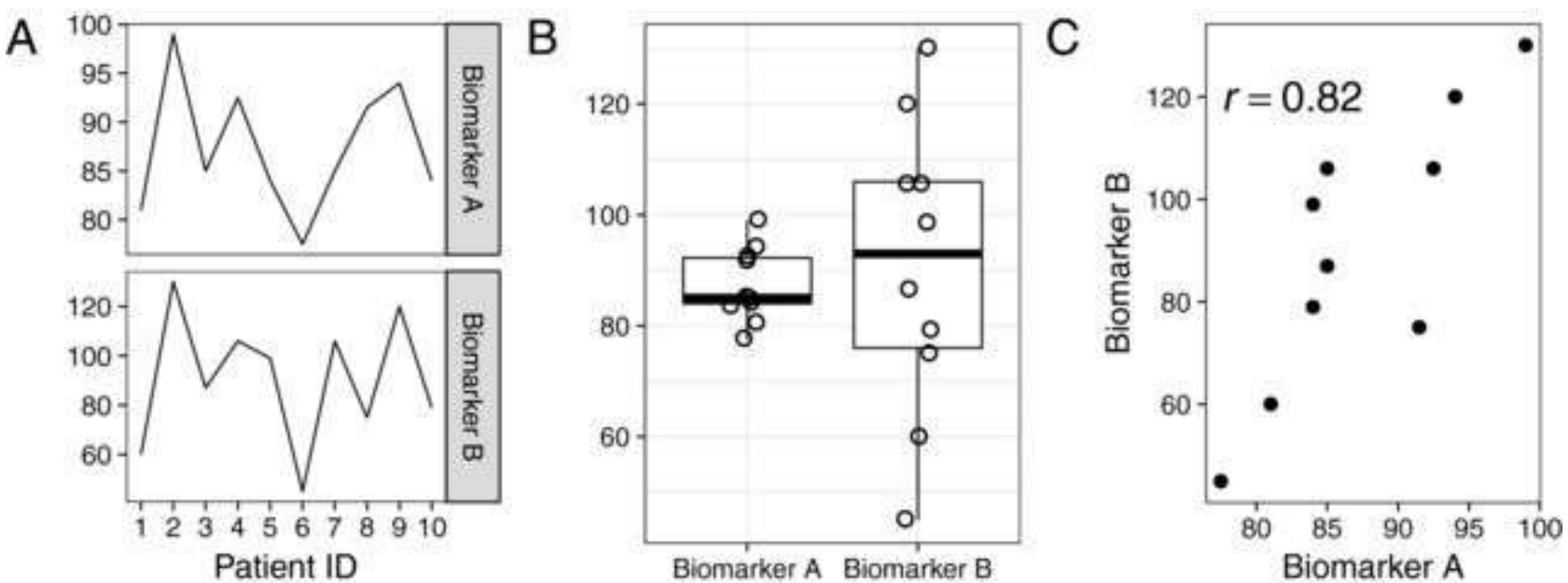
Spurious correlations

Linecharts tend to suggest correlation



http://www.tylervigen.com/spurious-correlations

Even good graphics tell different stories



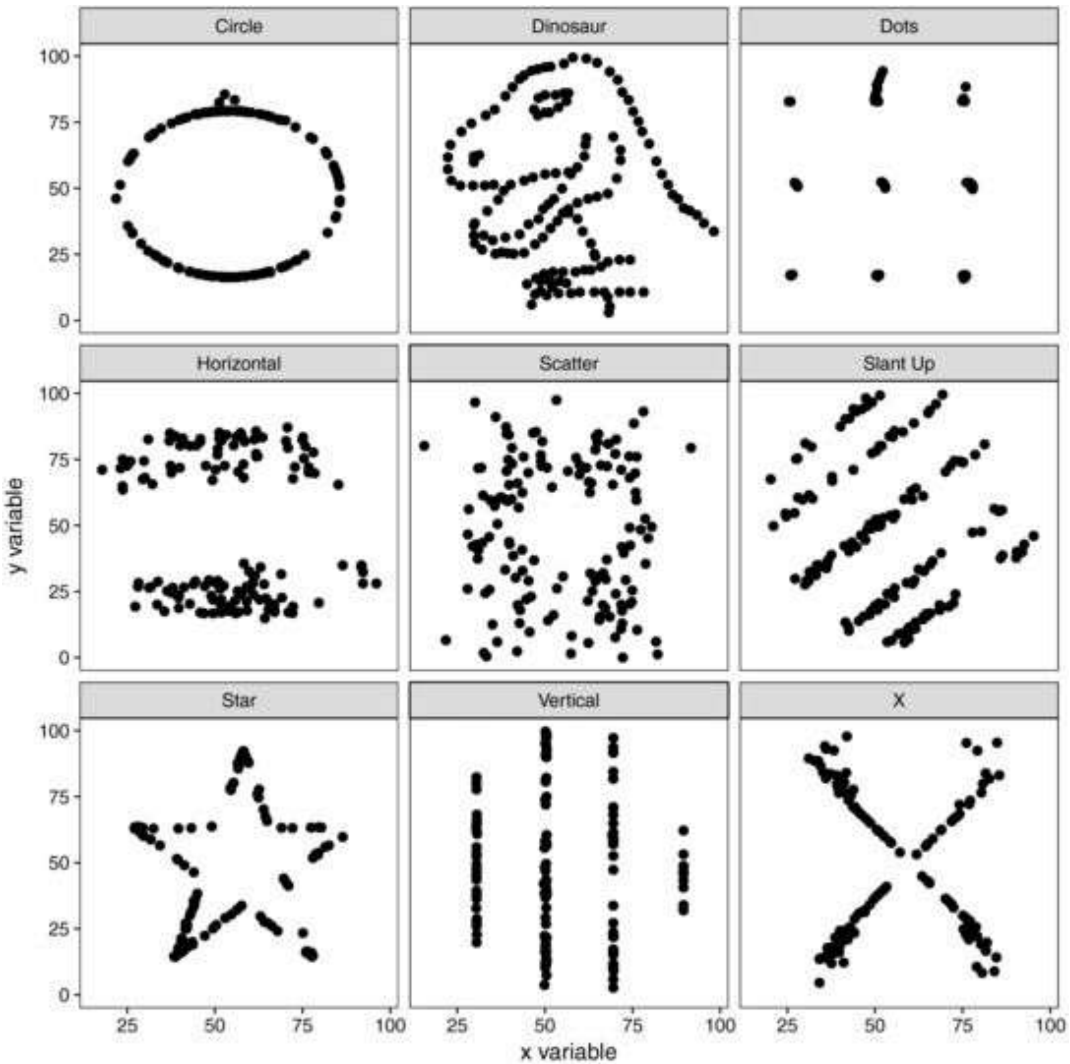
line plot here shows that something is changing over time

box and whisker plot show distribution which we cannot see from line

can see correlation which the others cannot see

Cabanski, C., Gilbert, H., & Mosesova, S. (2018). *Can Graphics Tell Lies? A Tutorial on How To Visualize Your Data*. Clinical and translational science, 11(4), 371–377. doi:10.1111/cts.12554

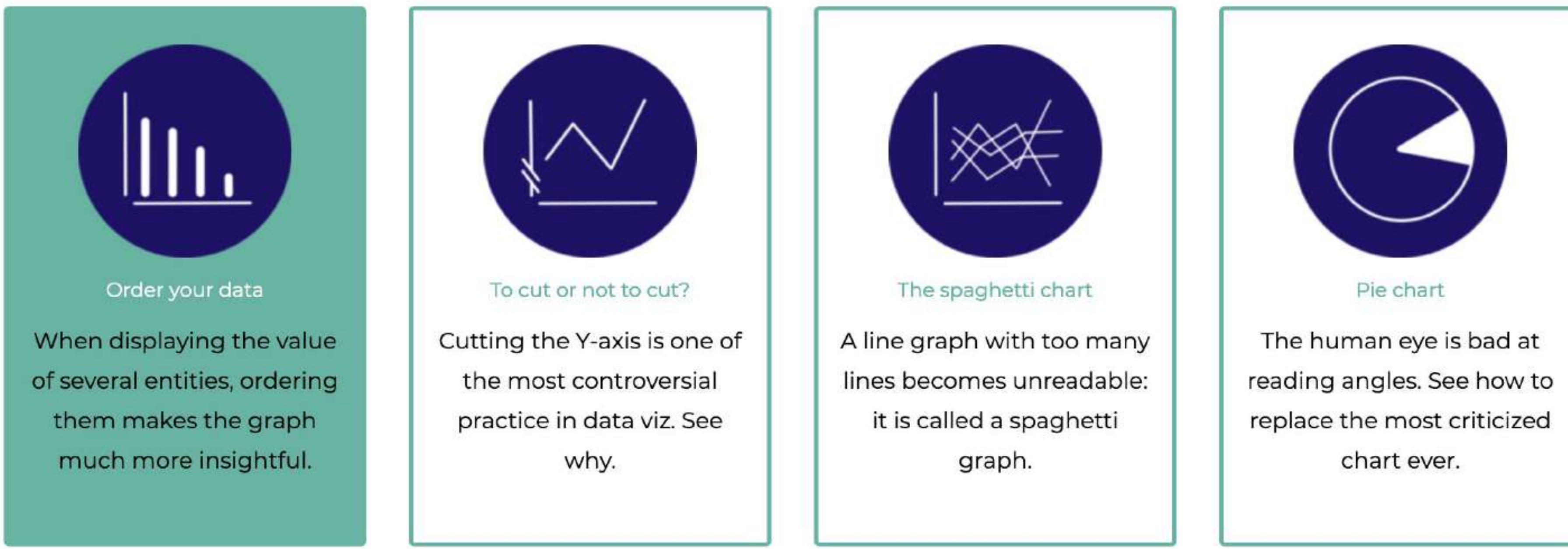
Raw data, not just summary statistics



Nine data sets with equivalent summary statistics. Each data set has the same x mean (54.26), y mean (47.83), x SD (16.76), y SD (26.93), and Pearson correlation coefficient (−0.06). The nine distinct patterns show the importance of plotting the raw data rather than only displaying summary statistics or models.

Cabanski, C., Gilbert, H., & Mosesova, S. (2018). *Can Graphics Tell Lies? A Tutorial on How To Visualize Your Data*. Clinical and translational science, 11(4), 371–377. doi:10.1111/cts.12554

List of caveats



https://www.data-to-viz.com/caveats.html

References

Drucker, Johanna. 2011. “Humanities Approaches to Graphical Display.” *Digital Humanities Quarterly* 005 (1).

Hepworth, Katherine, and Christopher Church. 2019. “Racism in the Machine: Visualization Ethics in Digital Humanities Projects.” *Digital Humanities Quarterly* 012 (4).

Gray, Jonathan, Liliana Bounegru, Stefania Milan, and Paolo Ciuccarelli. 2016. “Ways of Seeing Data: Toward a Critical Literacy for Data Visualizations as Research Objects and Research Devices.” In *Innovative Methods in Media and Communication Research*, edited by Sebastian Kubitschko and Anne Kaun, 227–51. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40700-5_12.

Thorp, Jer. 2017. “You Say Data, I Say System.” Hacker Noon. July 13, 2017. https://hackernoon.com/you-say-data-i-say-system-54e84aa7a421.

Friendly, Michael and Daniel Denis. “The early origins and development of the scatterplot.” *Journal of the History of the Behavioral Sciences*, 41(2), 103–130.

Nundy, Surajit et al. 2000. “Why are angles misperceived?”. *Proc Natl Acad Sci* 97(10): 5592–5597.