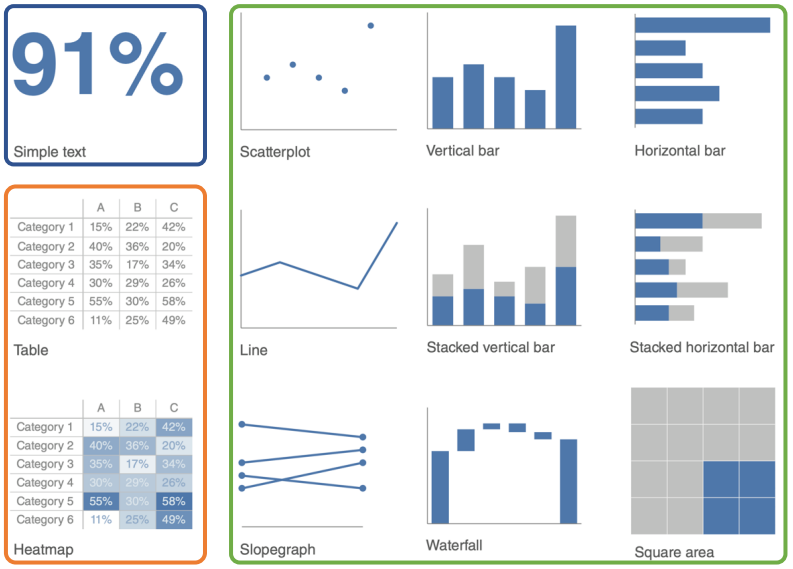
Module 01
Data Literacy
Lecture 2

Wan Fang

Southern University of Science and Technology

Agenda

- Introduction to Data
  - A Simple Example with Infograms on Pandemics
  - A More Advanced Visualization of plastic polution
  - Basic Visualizations of Data
    - Simple Text | Table | Graph
- Dimensional Visualization of Data
  - Dataset of 1D/2D/3D
  - Dataset of higher dimensions



Introduction to Data

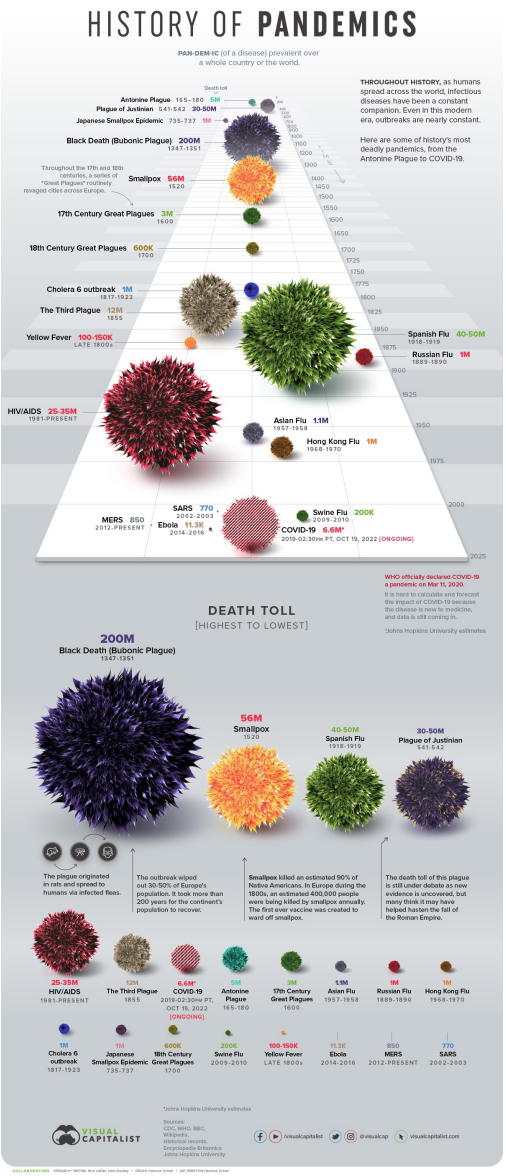
Wan Fang

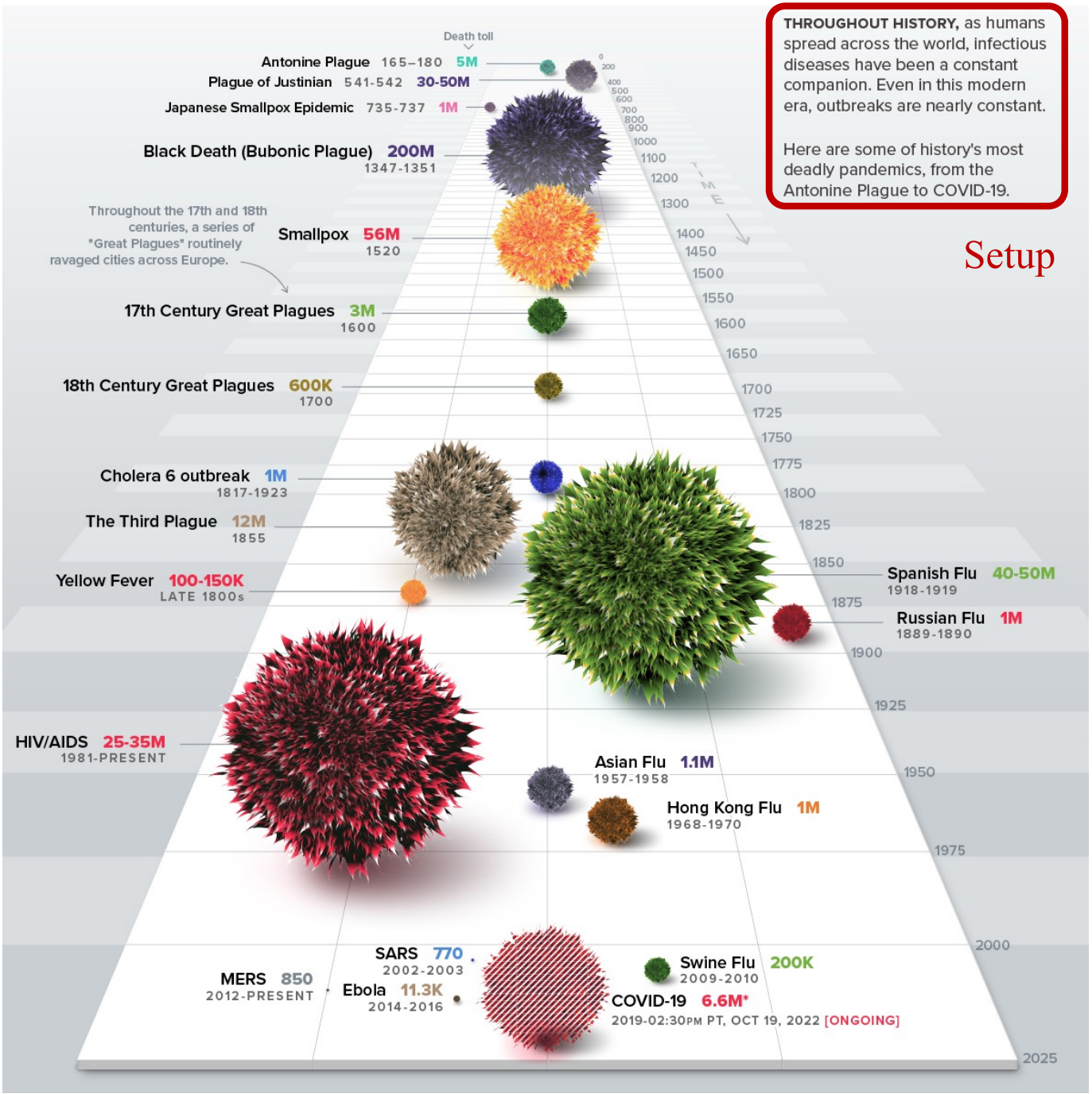
Southern University of Science and Technology

[Adapted from Storytelling with Data by Cole Nussbaumer Knaflic]

https://www.visualcapitalist.com/history-of-pandemics-deadliest/

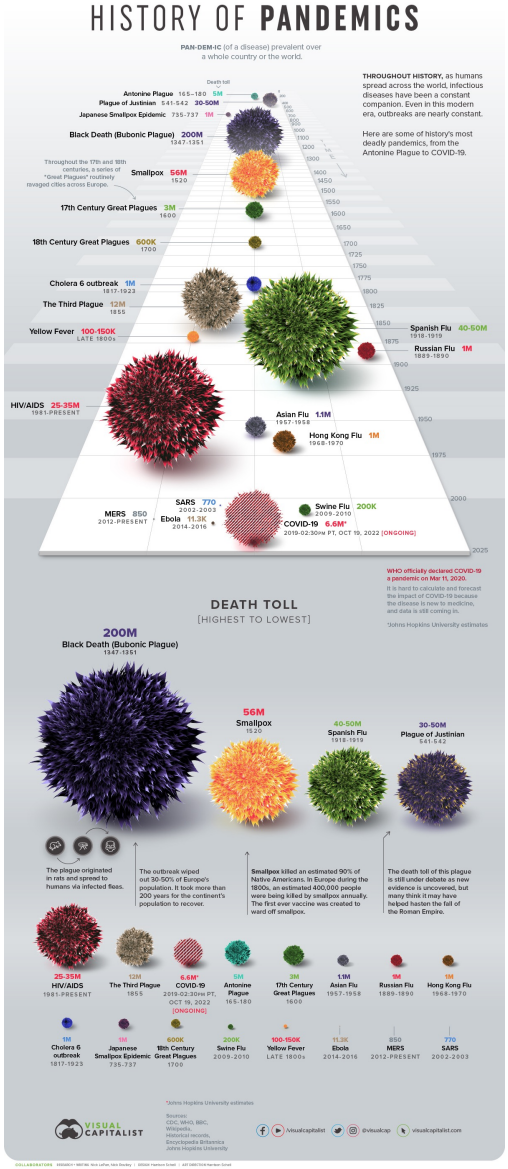
Infograms

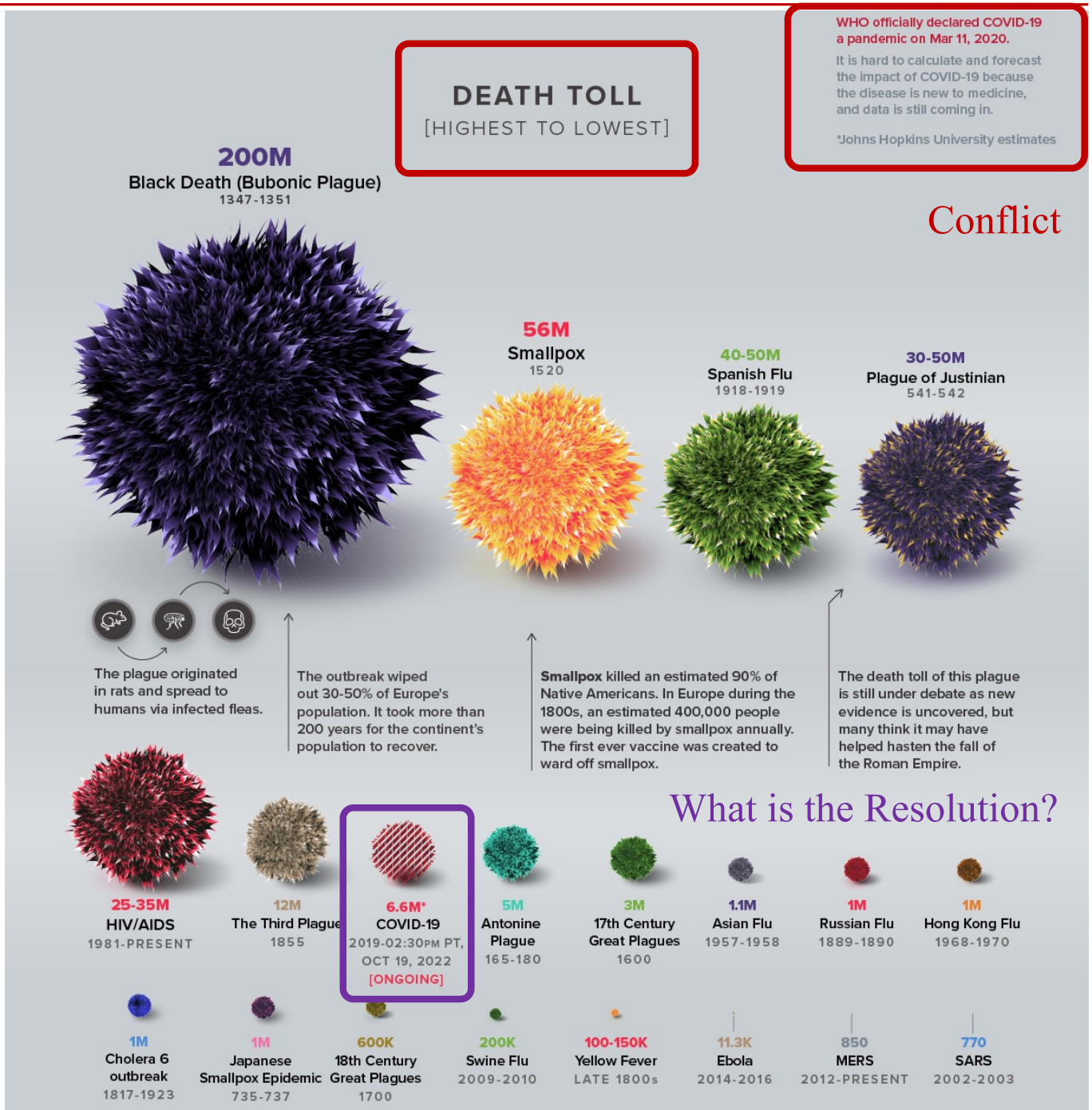




https://www.visualcapitalist.com/history-of-pandemics-deadliest/

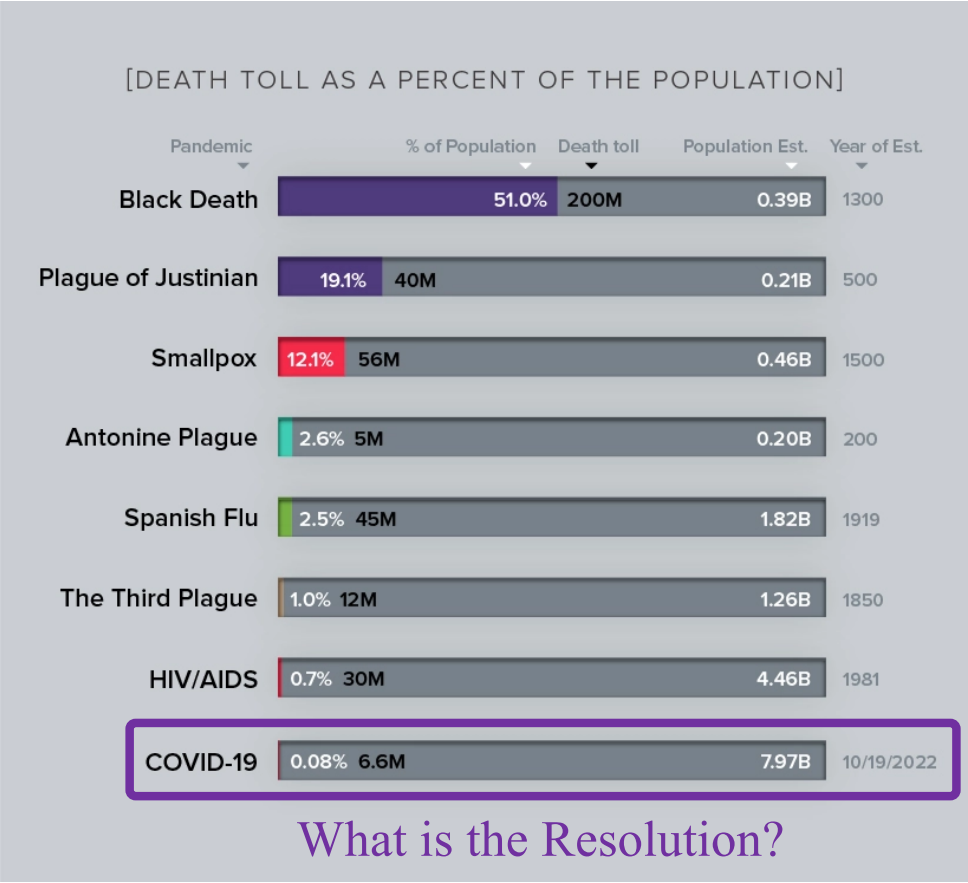
Infograms





Infograms

- Descriptive summary of raw data
- Further insights to illustrate ***impacts behind the story***



What is the Resolution?

Name	Time period	Type / Pre-human host	Death toll
Antonine Plague	165-180	Believed to be either smallpox or measles	5M
Japanese smallpox epidemic	735-737	Variola major virus	1M
Plague of Justinian	541-542	Yersinia pestis bacteria / Rats, fleas	30-50M
Black Death	1347-1351	Yersinia pestis bacteria / Rats, fleas	200M
New World Smallpox Outbreak	1520 – onwards	Variola major virus	56M
Great Plague of London	1665	Yersinia pestis bacteria / Rats, fleas	100,000
Italian plague	1629-1631	Yersinia pestis bacteria / Rats, fleas	1M
Cholera Pandemics 1-6	1817-1923	V. cholerae bacteria	1M+
Third Plague	1885	Yersinia pestis bacteria / Rats, fleas	12M (China and India)
Yellow Fever	Late 1800s	Virus / Mosquitoes	100,000-150,000 (U.S.)
Russian Flu	1889-1890	Believed to be H2N2 (avian origin)	1M
Spanish Flu	1918-1919	H1N1 virus / Pigs	40-50M
Asian Flu	1957-1958	H2N2 virus	1.1M
Hong Kong Flu	1968-1970	H3N2 virus	1M
HIV/AIDS	1981-present	Virus / Chimpanzees	25-35M
Swine Flu	2009-2010	H1N1 virus / Pigs	200,000
SARS	2002-2003	Coronavirus / Bats, Civets	770
Ebola	2014-2016	Ebolavirus / Wild animals	11,000
MERS	2015-Present	Coronavirus / Bats, camels	850
COVID-19	2019-Present	Coronavirus – Unknown (possibly pangolins)	6.6M (Johns Hopkins University estimate as of October 19, 2022)

Note: Many of the death toll numbers listed above are best estimates based on available research. Some, such as the Plague of Justinian and Swine Flu, are subject to debate based on new evidence.

Introduction to Data

https://ourworldindata.org/plastic-pollution#explore-data-on-plastic-pollution

(Some) Basic Visualizations of Data

- Text
- Table
- Graph

- Others

91%

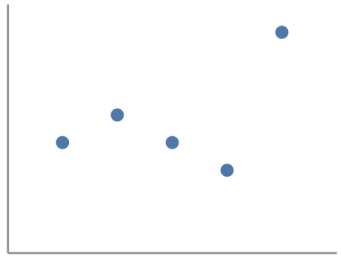
Simple text

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

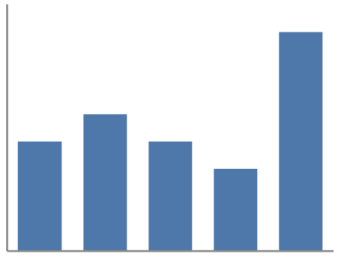
Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

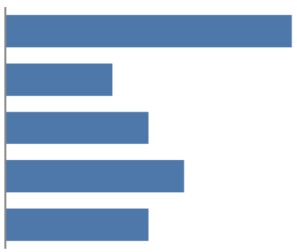
Heatmap



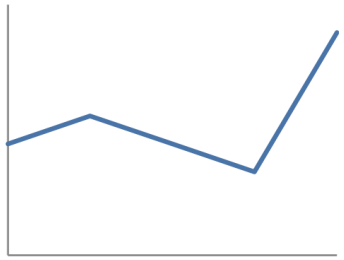
Scatterplot



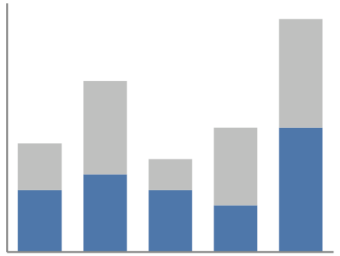
Vertical bar



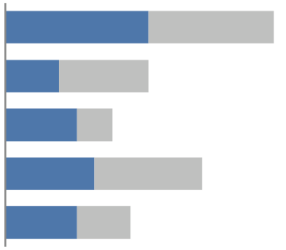
Horizontal bar



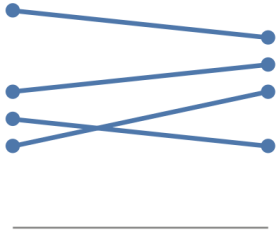
Line



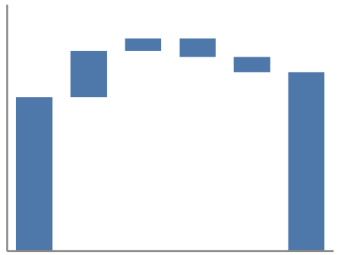
Stacked vertical bar



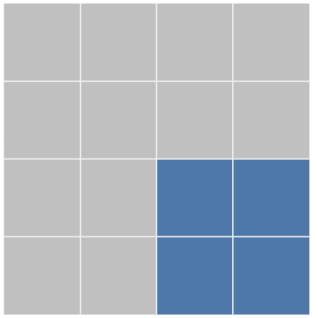
Stacked horizontal bar



Slopegraph



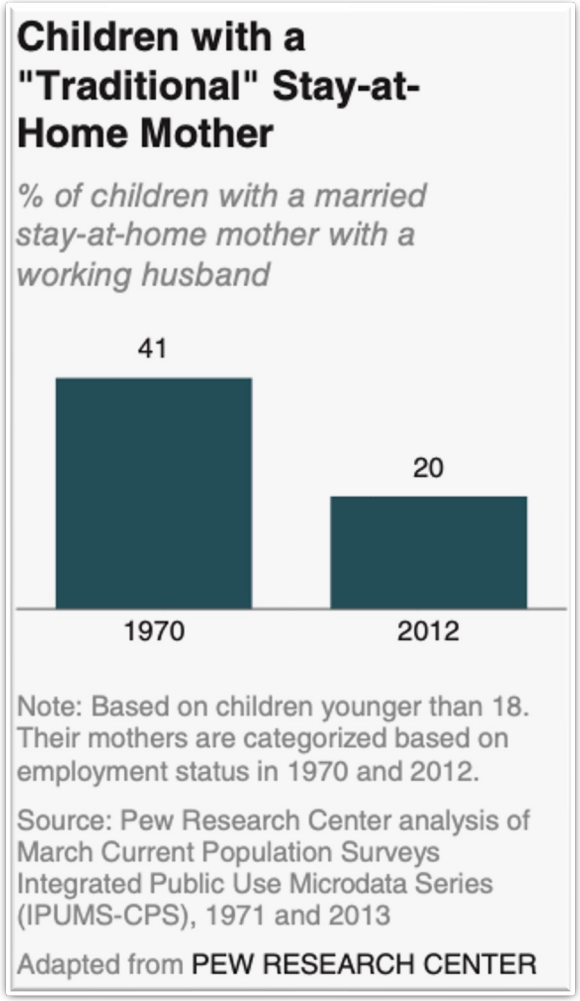
Waterfall



Square area

Simple Text

- When you have just a number or two to share, simple text can be a great way to communicate.



- Think about solely using the number—making it as prominent as possible—and a few supporting words to clearly make your point.
- Beyond potentially being misleading, putting one or only a couple of numbers in a table or graph simply causes the numbers to lose some of their oomph.

The fact that you have some numbers does not mean that you need a graph!

➢ When you have more data that you want to show, generally a table or graph is the way to go.

Tables

- Tables interact with our **verbal system**, which means that we read them.
  - *Reading across rows and down columns or Comparing values*

✓ Tables are great for communicating to a mixed audience whose members will each look for their particular row of interest.

✓ If you need to communicate multiple different units of measure, this is also typically easier with a table than a graph.

Tables in live presentations

Using a table in a live presentation is rarely a good idea. As your audience reads it, you lose their ears and attention to make your point verbally. When you find yourself using a table in a presentation or report, ask yourself: what is the point you are trying to make? Odds are that there will be a better way to pull out and visualize the piece or pieces of interest. In the event that you feel you're losing too much by doing this, consider whether including the full table in the appendix and a link or reference to it will meet your audience's needs.

Heavy borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Light borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Minimal borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

You want the design to fade into the background, letting the data take center stage.

Note how the data stands out more than the structural components of the table in the second and third iterations (light borders, minimal borders).

Heatmap

- A way to visualize data in tabular format, where in place of (or in addition to) the numbers, you leverage colored cells that convey the relative magnitude of the numbers.

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

LOW-HIGH

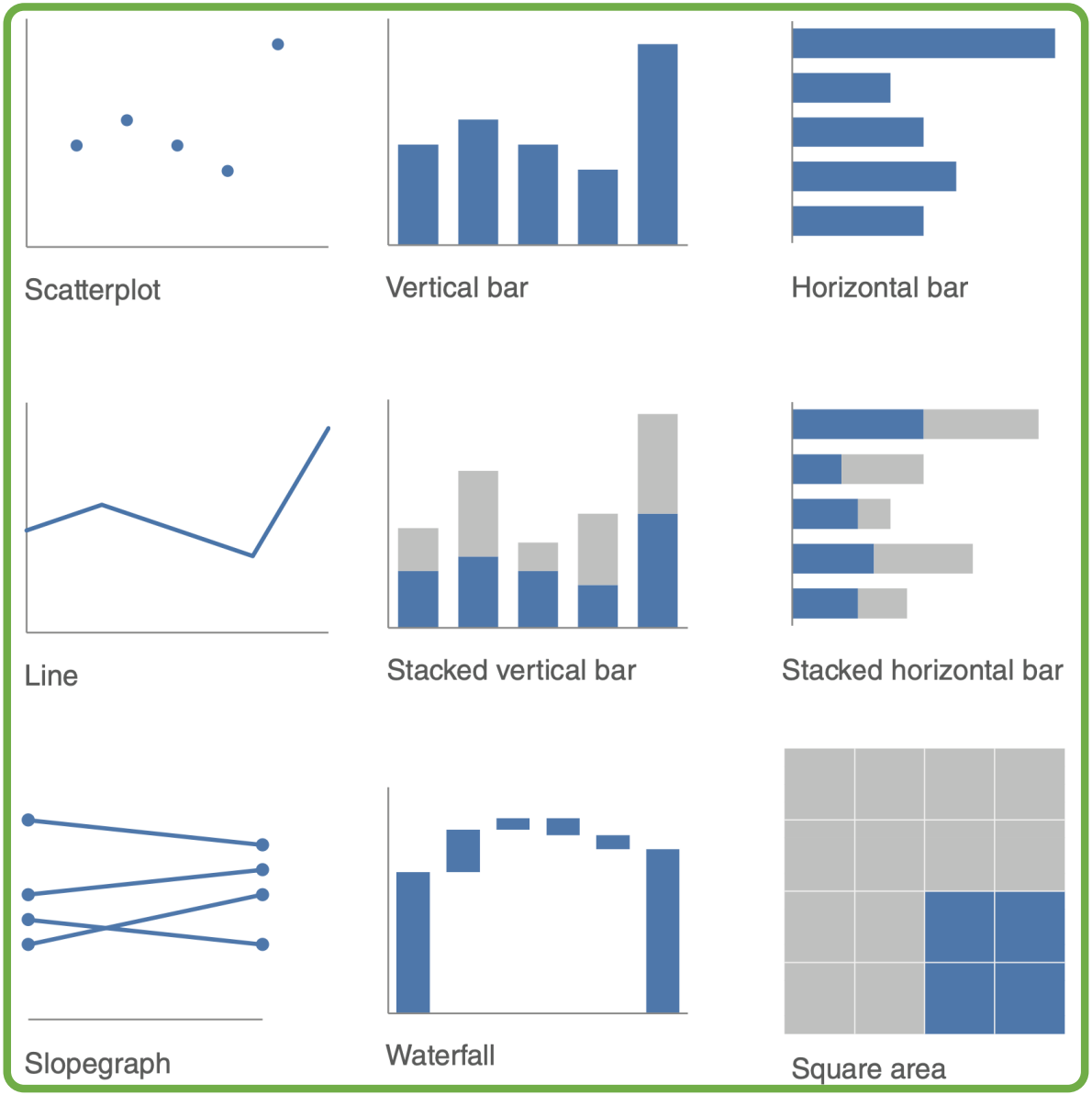
	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

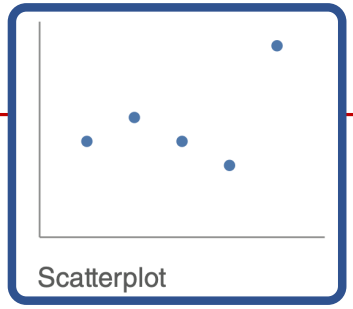
- Graphing applications (like Excel) typically have conditional formatting functionality built in that allows you to apply similar formatting shown here.
  - Be sure to include a legend to help the reader interpret the data (in this case, the LOW-HIGH subtitle on the heatmap with color corresponding to the conditional formatting color serves this purpose).

Graphs

- Graphs interact with our **visual system**, which is **faster** at processing information.
  - This means that a well-designed graph will typically get the information across more quickly than a well-designed table.
  - There are a plenty of graph types out there. The good news is that a handful of them will meet most of your everyday needs.

- Common types of graphs:
  - points, lines, bars, and area

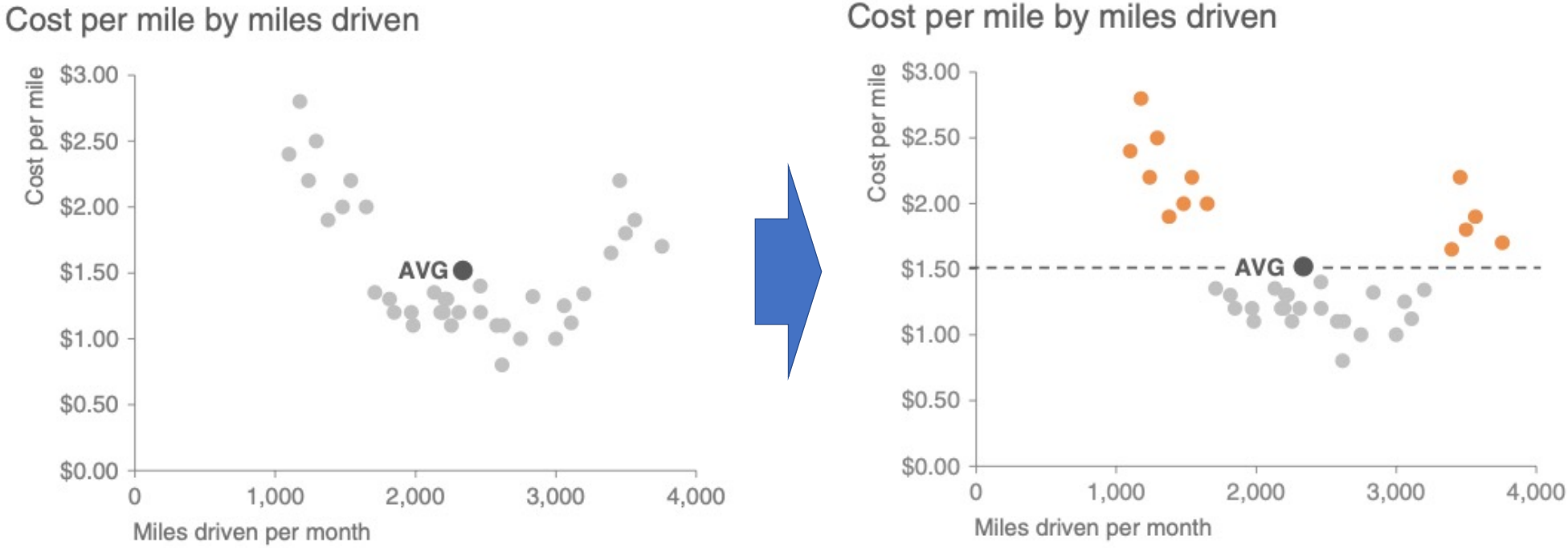




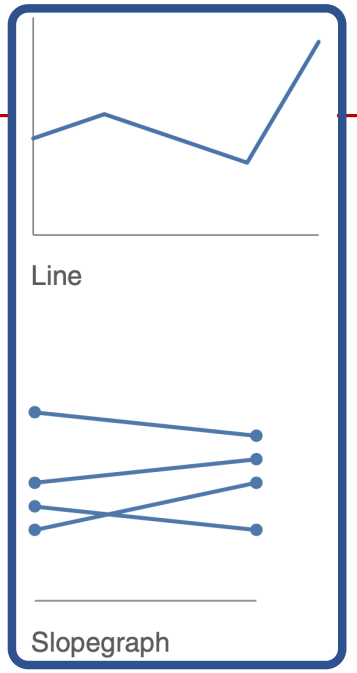
Points: Scatterplot

- Can be useful for showing the relationship between two things, because they allow you to encode data simultaneously on a horizontal *x*-axis and vertical *y*-axis to see whether and what relationship exists.
  - They tend to be more frequently used in scientific fields (and perhaps, because of this, are sometimes viewed as complicated to understand by those less familiar with them).
  - Though infrequent, there are use cases for scatterplots in the business world as well.

- If we want to focus primarily on those cases where cost per mile is above average, a slightly modified scatterplot designed to draw our eye there more quickly might look something like those on the right.



Lines



- Line graphs are most commonly used to plot continuous data.
  - Because the points are physically connected via the line, it implies a connection between the points that may not make sense for categorical data (a set of data that is sorted or divided into different categories).
  - Often, our continuous data is in some unit of time: days, months, quarters, or years.

➢The standard line graph

➢The slopegraph

Example

*Sample Raw Data for **the Slopgraph plot above***

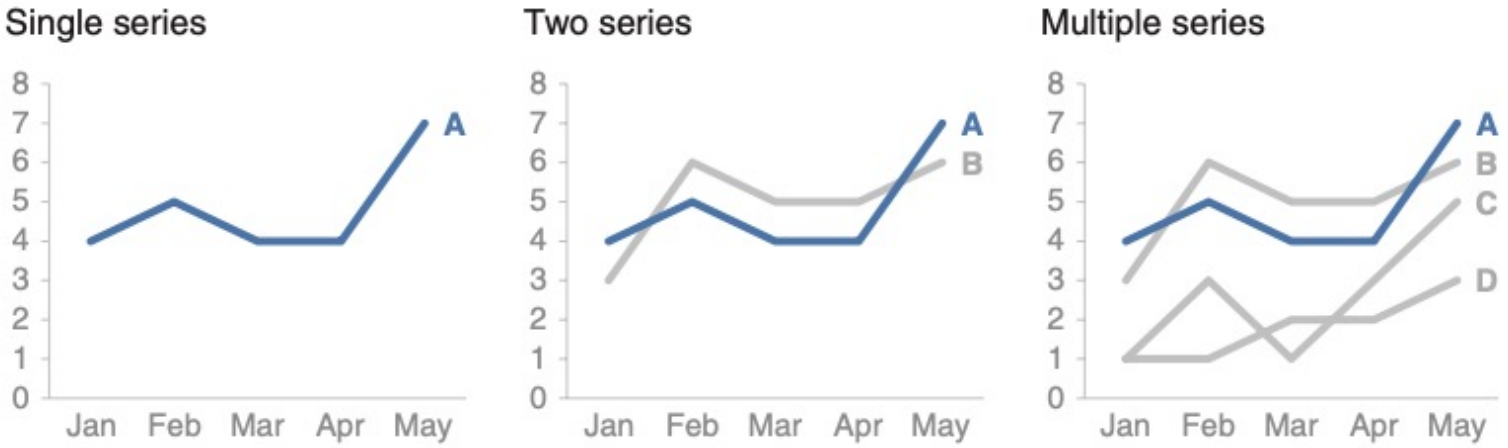
Time	Metric A	Metric B	Metric C	Metric D
Timestamp 1	AA	BB	CC	DD
Timestamp 2	AA	BB	CC	DD
Timestamp 3	AA	BB	CC	DD
Timestamp 4	AA	BB	CC	DD

*Sample Raw Data for **the Line plot above***

Horizontal in Time
Vertical in Metrics

Lines: Line graph

- The line graph can show a single series of data, two series of data, or multiple series

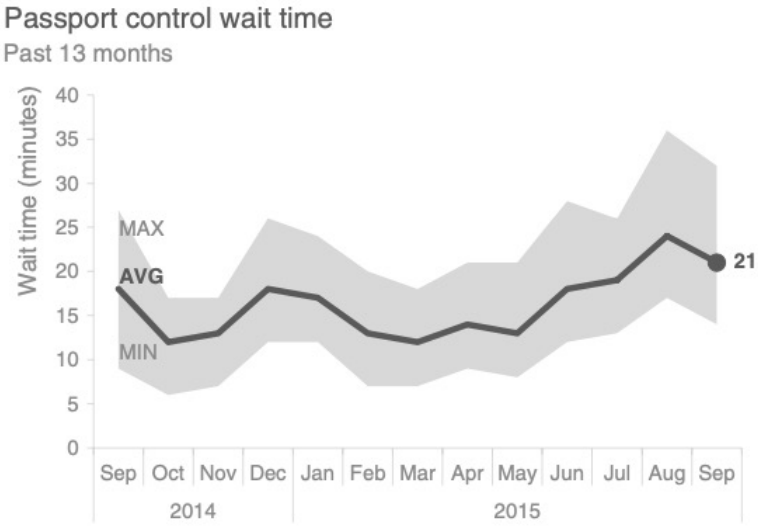


Note that when you're graphing time on the horizontal x-axis of a line graph, the data plotted must be in consistent intervals.

Be consistent in the time points you plot

Showing average within a range in a line graph

In some cases, the line in your line graph may represent a summary statistic, like the average, or the point estimate of a forecast. If you also want to give a sense of the range (or confidence level, depending on the situation), you can do that directly on the graph by also visualizing this range. For example, the graph in Figure 2.9 shows the minimum, average, and maximum wait times at passport control for an airport over a 13-month period.



Lines: Slopegraph

- Slopegraphs can be useful when you have two time periods or points of comparison and want to quickly show relative increases and decreases or differences across various categories between the two data points
- Example:
  - Imagine that you are analyzing and communicating data from a recent employee feedback survey.
  - To show the relative change in survey categories from 2014 to 2015, the slopegraph might look something like the one on the right.

Slopegraphs pack in a lot of information

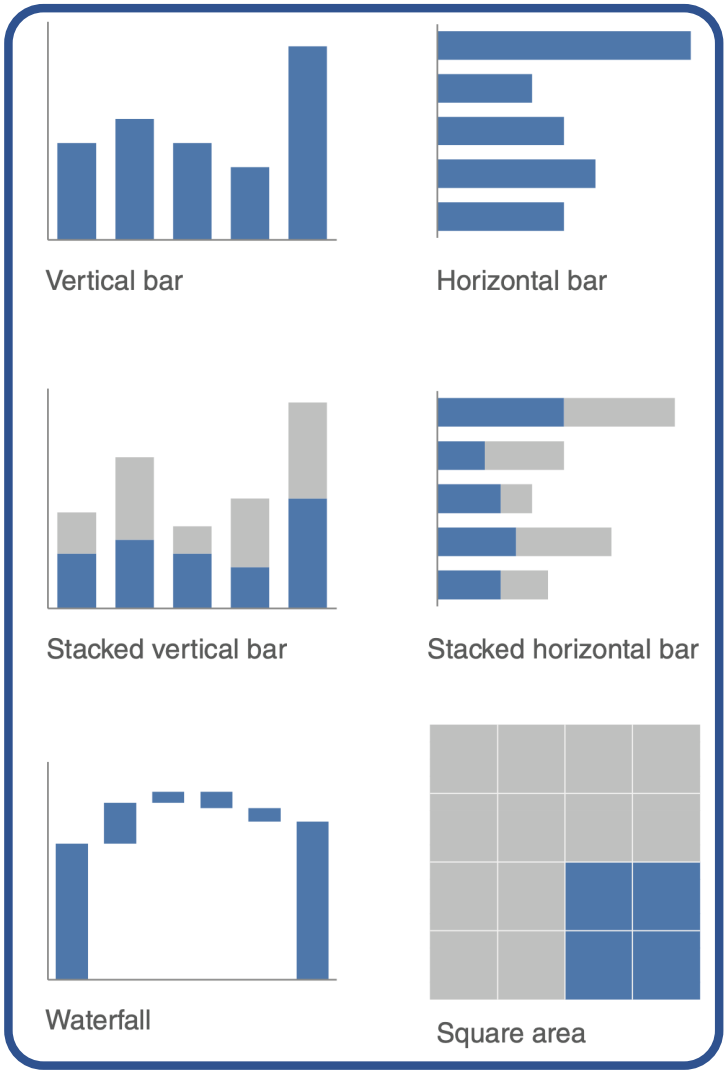
- The lines that connect them give you the visual increase or decrease in rate of change (via the slope or direction) without ever having to explain that's what they are doing, or what exactly a "rate of change" is—rather, it's intuitive.

Slopegraph template

Slopegraphs can take a bit of patience to set up because they often aren't one of the standard graphs included in graphing applications. An Excel template with an example slopegraph and instructions for customized use can be downloaded here: storytellingwithdata.com/slopegraph-template.

Bars

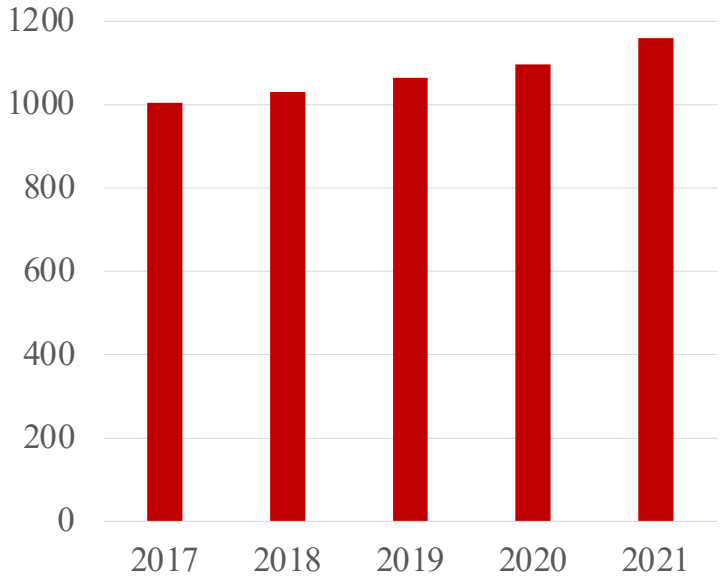
- Bar charts are easy for our eyes to read.
  - Our eyes compare the end points of the bars, so it is easy to see quickly which category is the biggest, which is the smallest, and also the incremental difference between categories.
  - Note that, because of how our eyes compare the relative end points of the bars, it is important that bar charts always have a zero baseline (where the *x*-axis crosses the *y*-axis at zero), otherwise you get a false visual comparison.



Introduction to Data

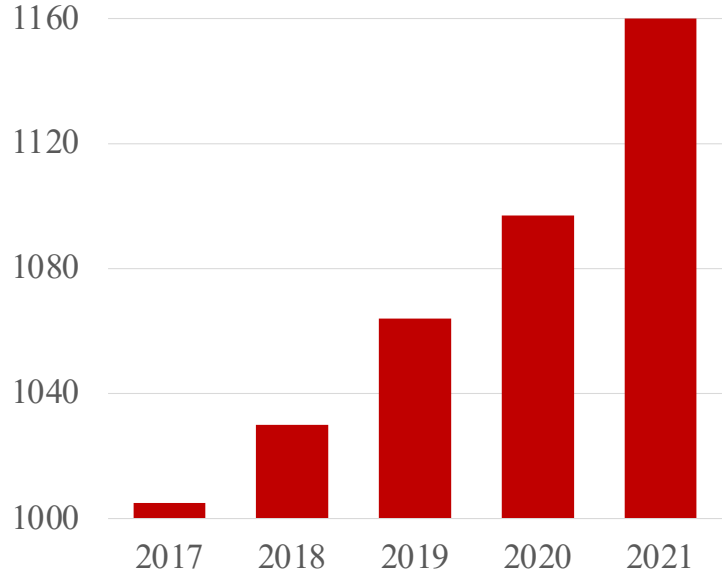
腾讯新闻：https://new.qq.com/rain/a/20220719A0AGQ700

知乎：https://zhuanlan.zhihu.com/p/391503737



With the same data, are they telling the same story?

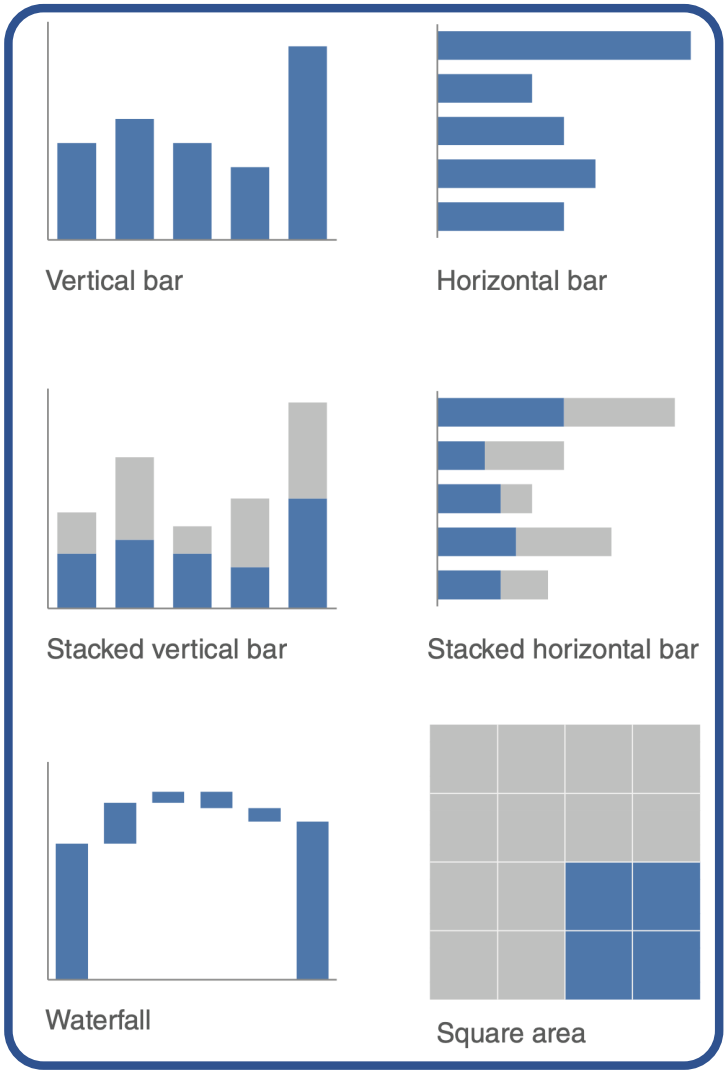
VS.



Bars

- Bar charts are easy for our eyes to read.
  - Our eyes compare the end points of the bars, so it is easy to see quickly which category is the biggest, which is the smallest, and also the incremental difference between categories.
  - Note that, because of how our eyes compare the relative end points of the bars, it is important that bar charts always have a zero baseline (where the *x*-axis crosses the *y*-axis at zero), otherwise you get a false visual comparison.

- *Bar charts must have a **zero** baseline*

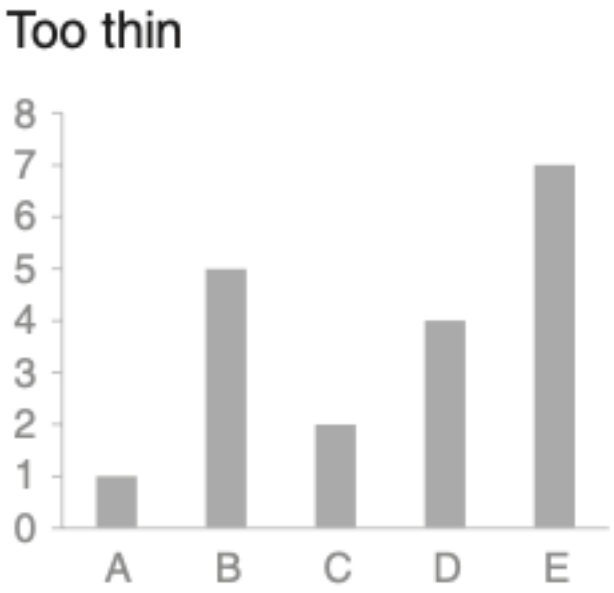


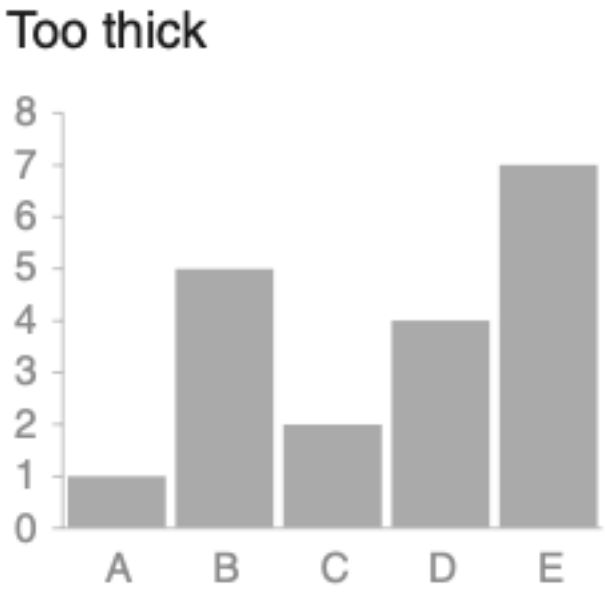
Ethical Concerns

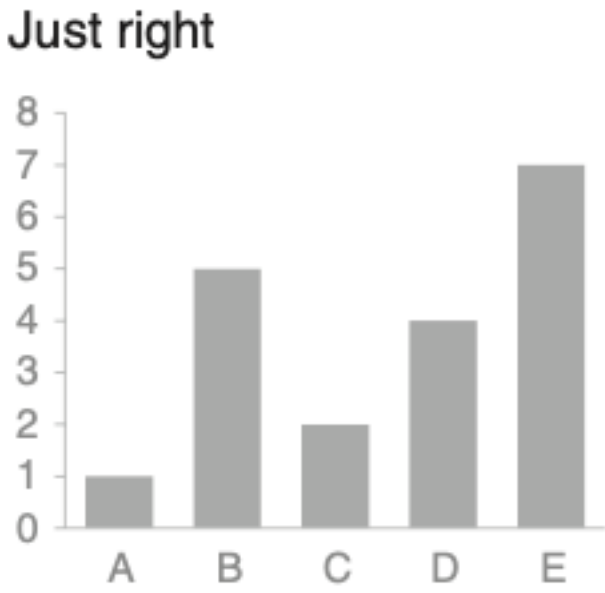
Ethics and data visualization

But what if changing the scale on a bar chart or otherwise manipulating the data better reinforces the point you want to make? Misleading in this manner by inaccurately visualizing data is not OK. Beyond ethical concerns, it is risky territory. All it takes is one discerning audience member to notice the issue (for example, the *y*-axis of a bar chart beginning at something other than zero) and your entire argument will be thrown out the window, along with your credibility.

- While we're considering lengths of bars, let's also spend a moment on the width of bars.
- There's no hard-and-fast rule here, but in general the bars should be wider than the white space between the bars.
- You don't want the bars to be so wide, however, that your audience wants to compare areas instead of lengths.
- Consider the following "Goldilocks" of bar charts: *too thin*, *too thick*, and *just right*.

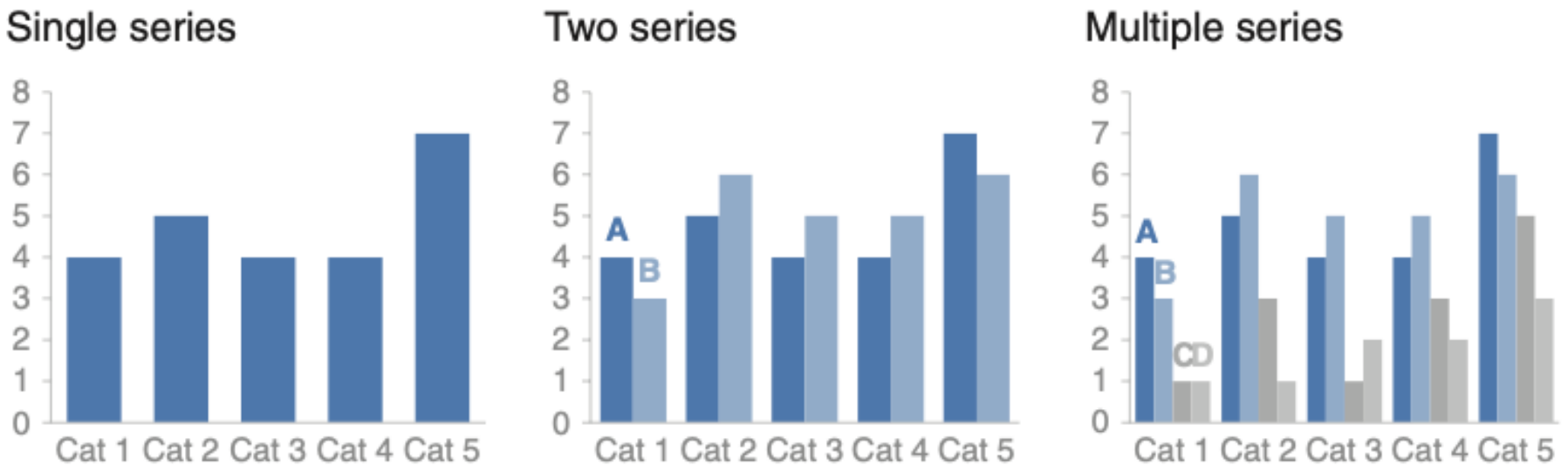






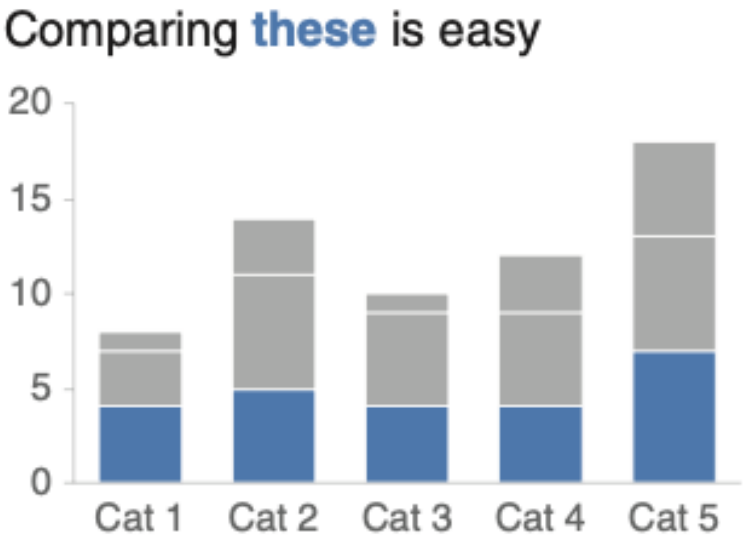
Vertical bar chart

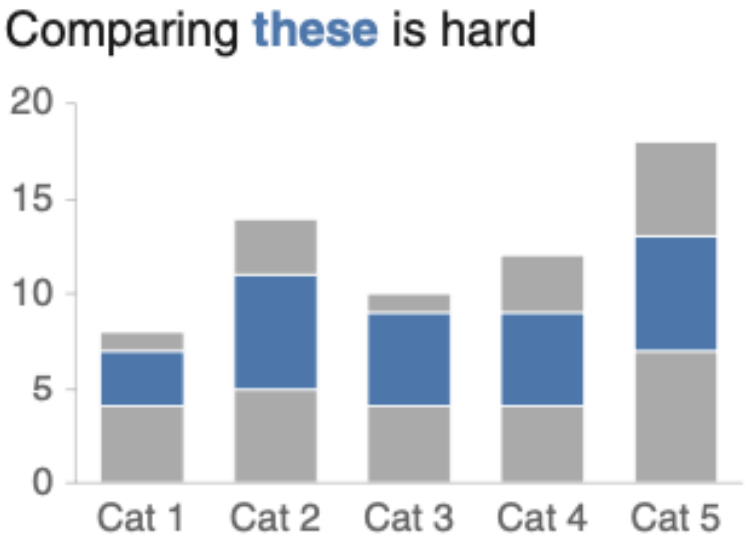
- Like line graphs, vertical bar charts can be single series, two series, or multiple series.
  - Note that as you add more series of data, it becomes more difficult to focus on one at a time and pull out insight, so use multiple series bar charts with caution.
- Be aware also that there is visual grouping that happens as a result of the spacing in bar charts having more than one data series. This makes the relative order of the categorization important.
  - Consider what you want your audience to be able to compare and structure your categorization hierarchy to make that as easy as possible.



Stacked vertical bar chart

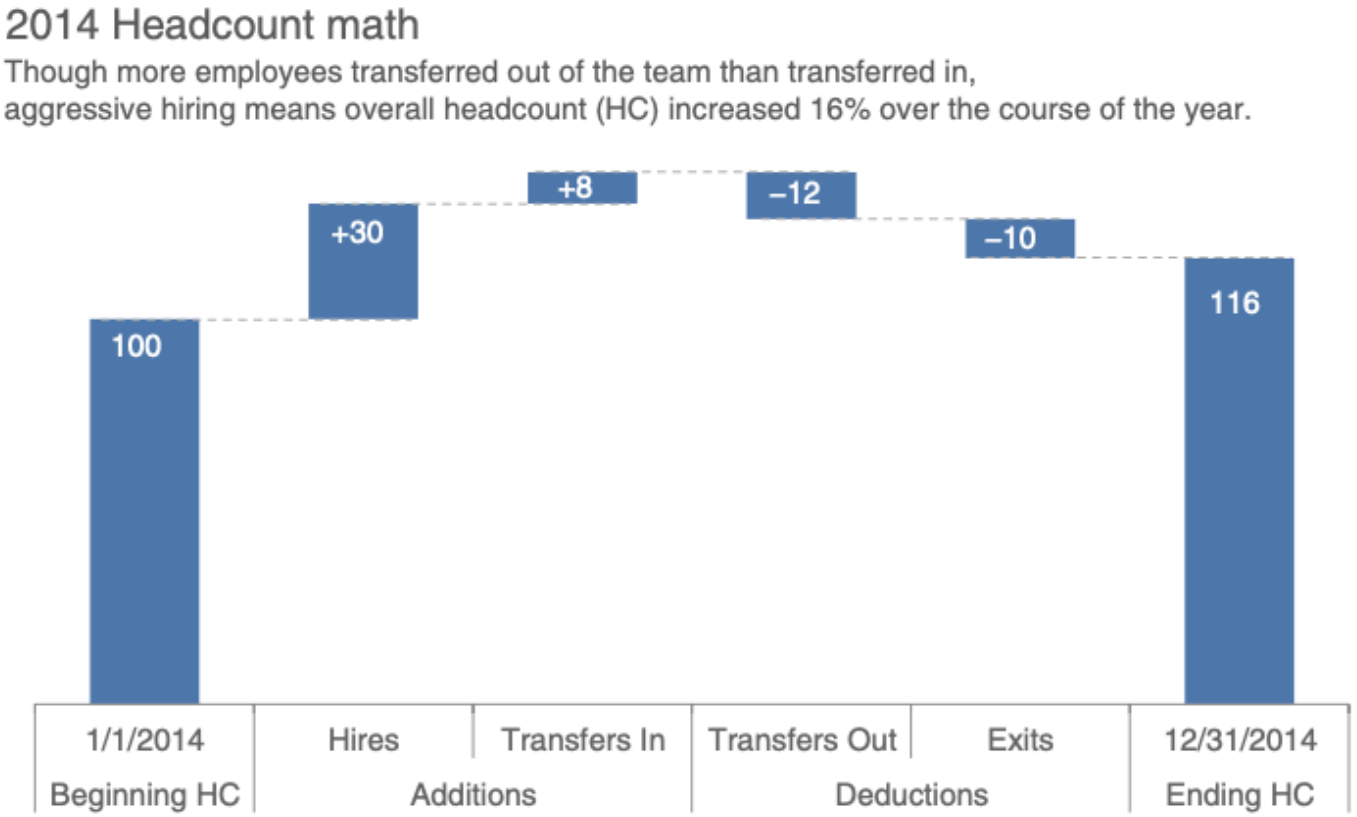
- Use cases for stacked vertical bar charts are more limited.
  - Meant to allow comparing totals across categories and also see the subcomponent pieces within a given category.
  - Can quickly become visually overwhelming, however—especially given the varied default color schemes in most graphing applications (more to come on that).
- Hard to compare the subcomponents across the various categories once you get beyond the bottom series (the one directly next to the x-axis) because you no longer have a consistent baseline to use to compare.
  - This makes it a harder comparison for our eyes to make.





Waterfall chart

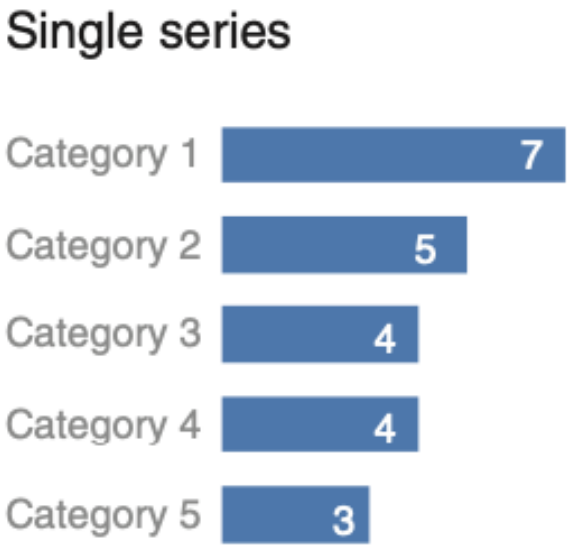
- The waterfall chart can be used to *pull apart the pieces of a stacked bar chart to focus on one at a time*, or *to show a starting point, increases and decreases, and the resulting ending point*.

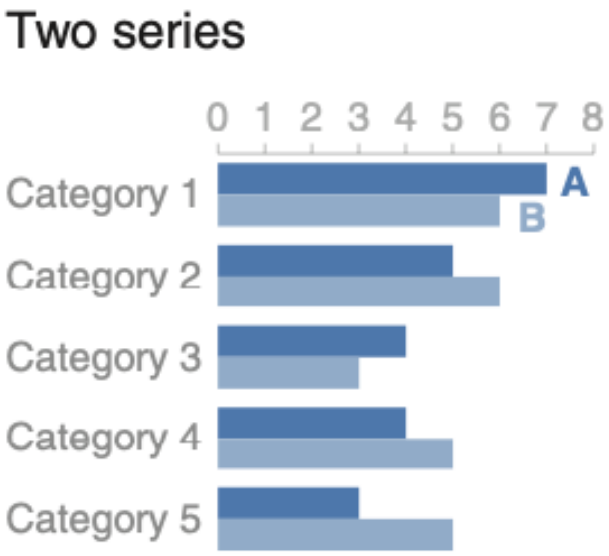


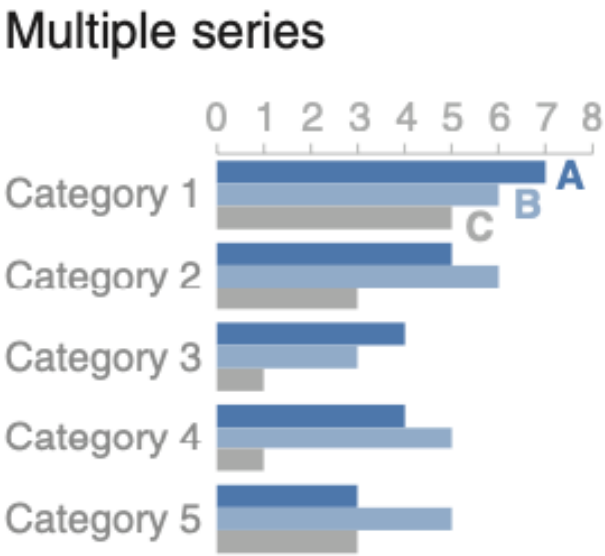
Imagine that you are an HR business partner and want to understand and communicate how employee headcount has changed over the past year for the client group you support.

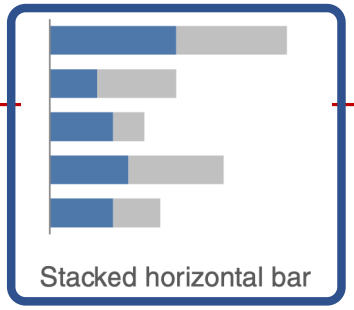
Horizontal Version

- Extremely easy to read
  - The single go-to graph for categorical data, which flips the vertical version on its side
  - **Especially useful if your category names are long**, as the text is written from left to right, as most audiences read, making your graph legible for your audience.
- Also, because of the way we typically process information the structure of the horizontal bar chart is such that **our eyes hit the category names before the actual data**.
  - starting at top left and making z's with our eyes across the screen or page
- This means by the time we get to the data, we already know what it represents.



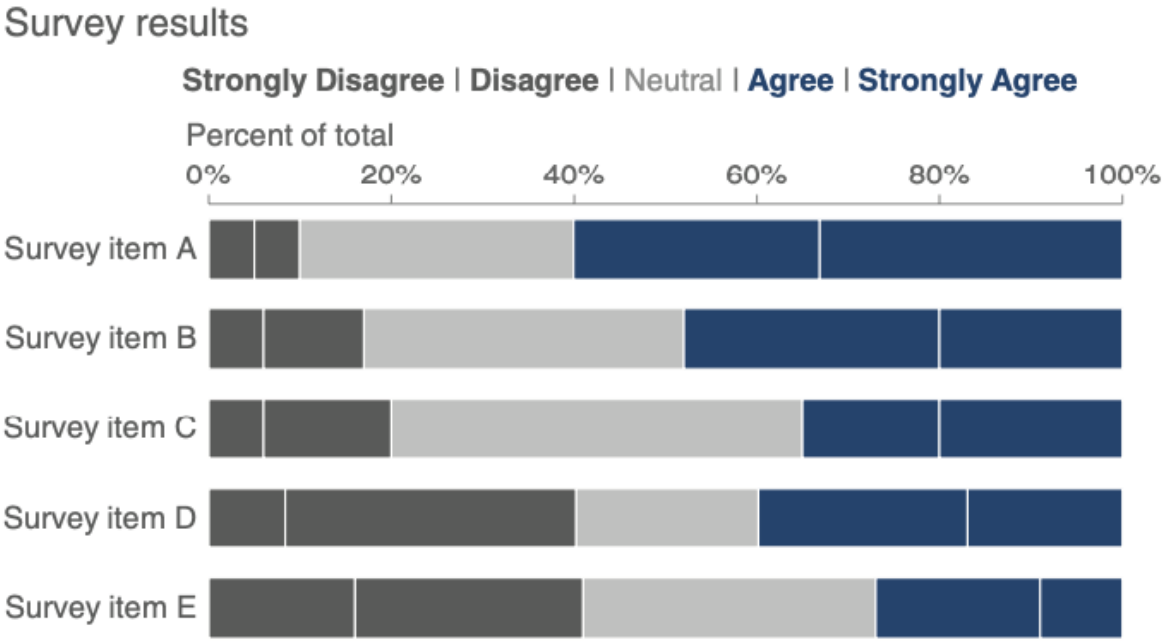






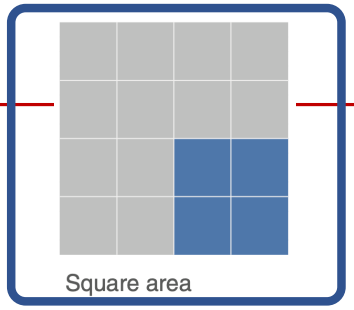
Stacked horizontal bar chart

- To show the totals across different categories but also give a sense of the subcomponent pieces
  - Can be structured to show either absolute values or sum to 100%
- Work well for visualizing portions of a whole on a scale from negative to positive
  - because you get a consistent baseline on both the far left and the far right, allowing for easy comparison of the left-most pieces as well as the right-most pieces.



For example, this approach can work well for **visualizing survey data collected along a Likert scale**

- A scale commonly used in surveys that typically ranges from Strongly Disagree to Strongly Agree



Area

- Avoid area graphs in general
  - Humans' eyes don't do a great job of attributing quantitative value to two-dimensional space,
  - which can render area graphs harder to read than some of the other types of visual displays we've discussed.

For this reason, avoid them with one exception—when you need to visualize numbers of vastly different magnitudes.

- The second dimension you get using a square for this (which has both height and width, compared to a bar that has only height or width) allows this to be done in a more compact way than possible with a single dimension

A Short Summary

- What do you need your audience to know?
  - In many cases, there **isn't a single correct visual display**; rather, often there are different types of visuals that could meet a given need.
  - The most important is to have that need clearly articulated. Then choose a visual display that will enable you to make this clear.

- Whatever will be easiest for your audience to read?
  - There is an easy way to test this, which is to create your visual and show it to a friend or colleague.
    - Have them articulate the following as they process the information: where they focus, what they see, what observations they make, what questions they have.
    - This will help you assess whether your visual is hitting the mark, or in the case where it isn't, help you know where to concentrate your changes.

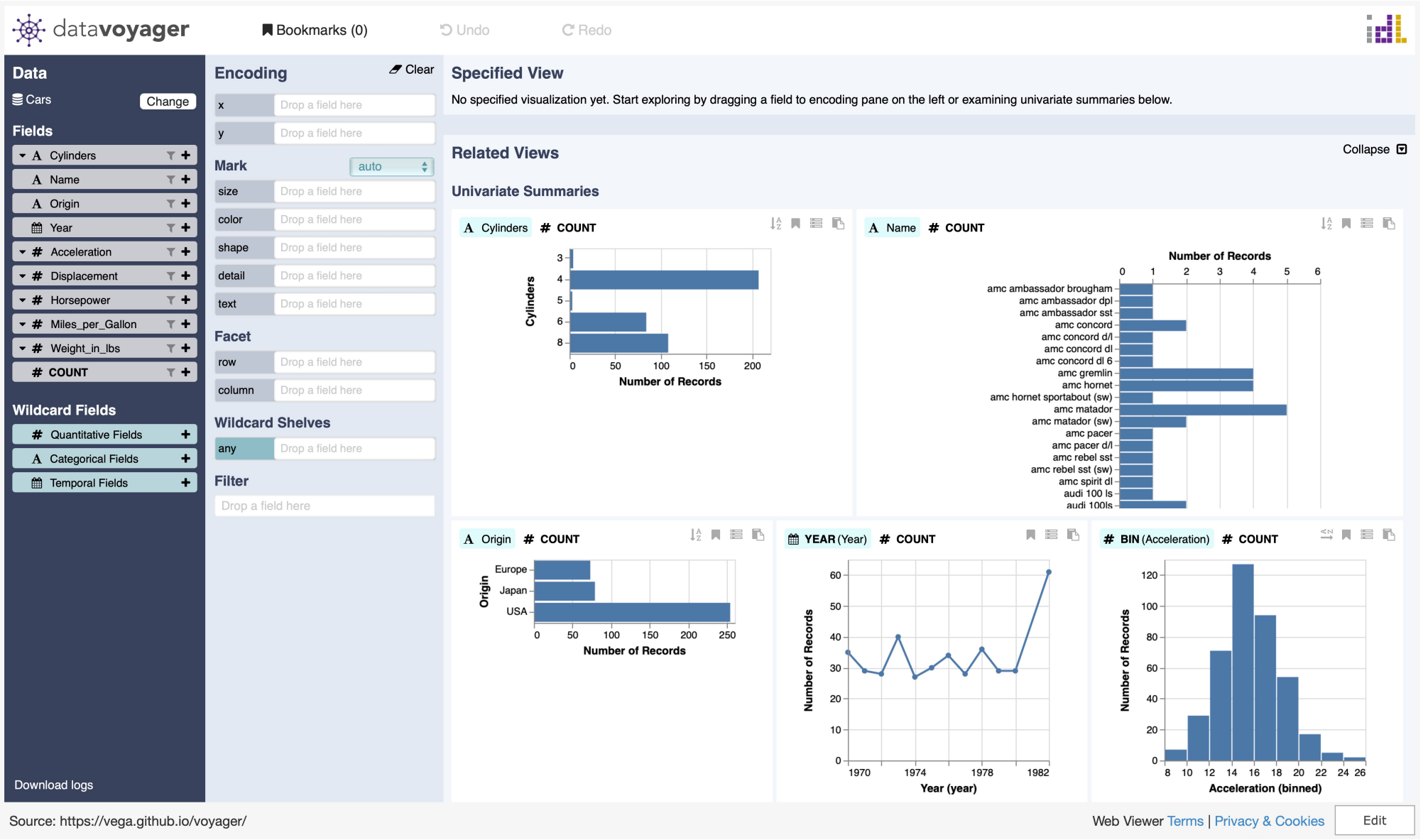
Dimensional Visualization of Data

Wan Fang

Southern University of Science and Technology

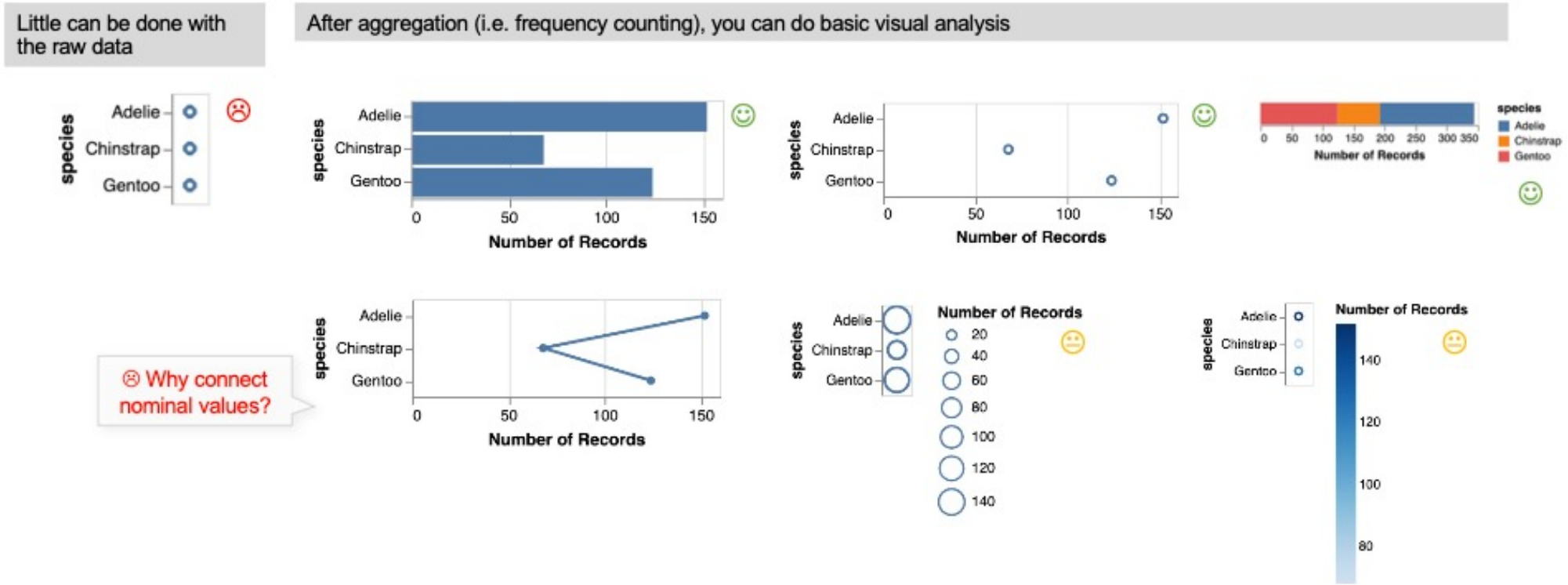
[Adapted from Data Analytics for Designers by Tak Yeon Lee]

datavoyager



1D Nominal

- When you are interested in a single column containing nominal values (i.e., only frequency counting is allowed)
  - E.g., **species** column of the penguin dataset

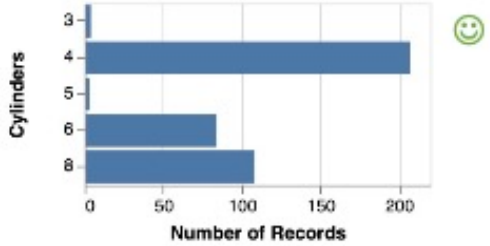


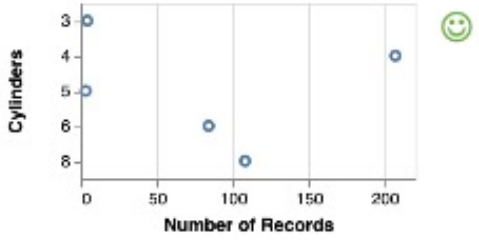
1D Ordinal

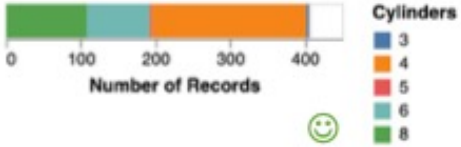
- When you are interested in a single column containing ordinal values (i.e., counting and ranking are allowed)
  - E.g., **# of cylinders** column of the car dataset

Little can be done with the raw data

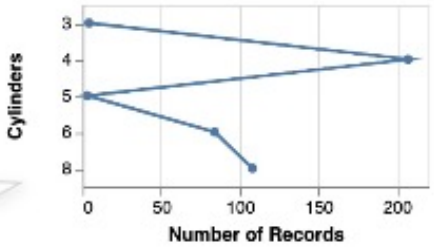
After aggregation (i.e. frequency counting), you can do basic visual analysis

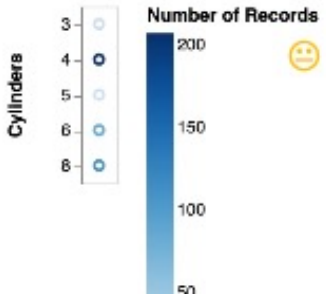






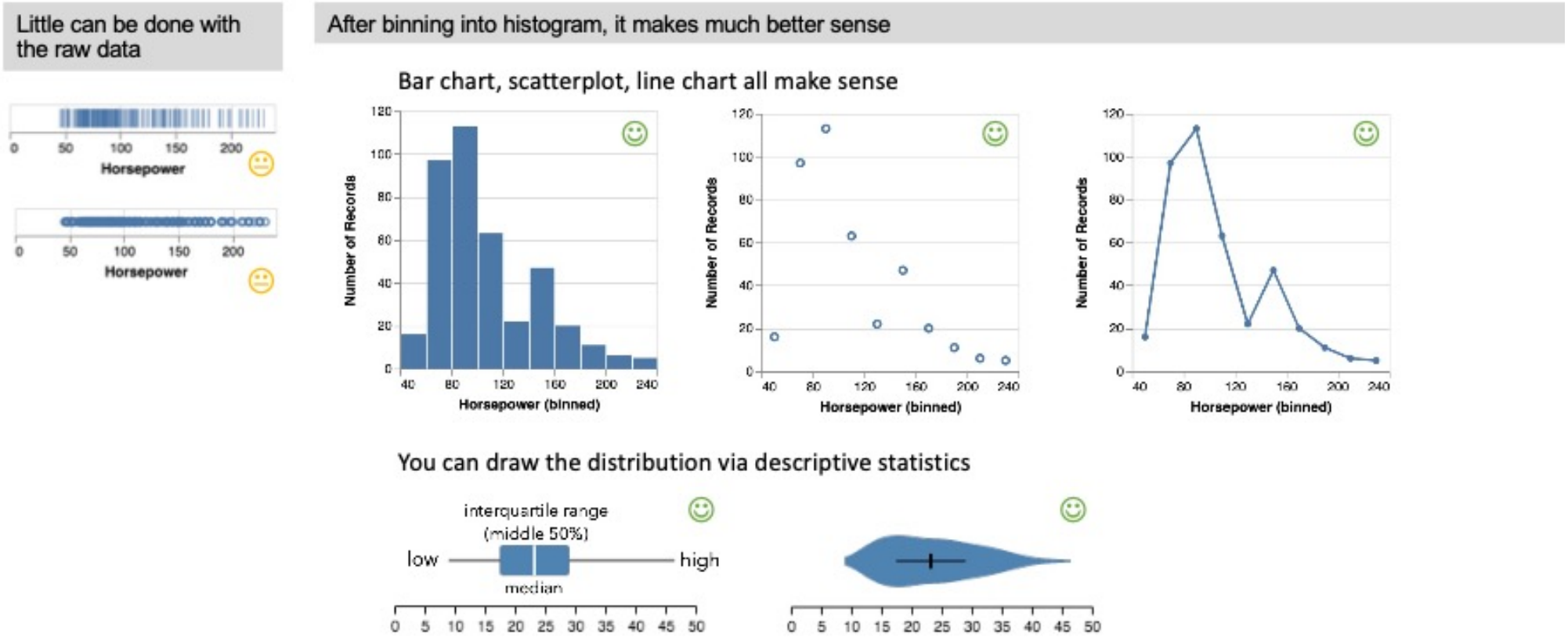
☹ Line chart makes sense as # cylinders have ordinal relationship





1D Quantitative

- When you are interested in a single column containing quantitative (interval or ratio) values (i.e., numerical operations are allowed).
  - E.g., **horsepower** column of the car dataset



Summary of 1D charts

- **Aggregation** is the key to draw meaningful charts from 1D
  - Frequency counting for nominals and ordinals
  - Binning (to get histogram) or Descriptive Statistics (to get distribution) for quantitative values

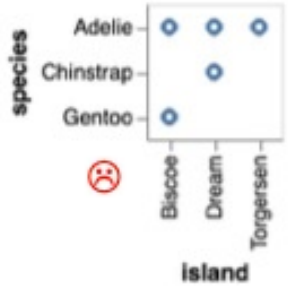
- EDA (Exploratory Data Analytics) begins with 1D charts
  - Suitable for finding outliers or incomplete values
  - Suitable for knowing distribution (mean, median, min, max)

- Once you found an interesting column(s), quickly move on to 2D
  - If 1D is not interesting, adding another column in 2D is unlikely to be interesting
  - Trial-and-errors of finding an interesting pair of columns is the core activity of EDA

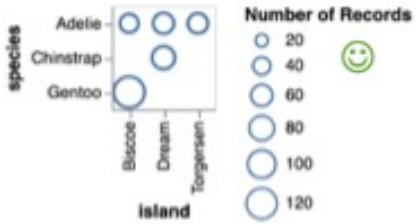
2D Nominal x Nominal

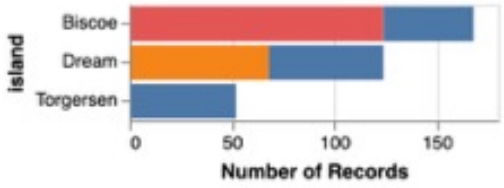
- If you are interested in how two nominal columns are correlated
  - E.g., **species** and **island** columns of the penguin dataset

Little can be done with the raw data

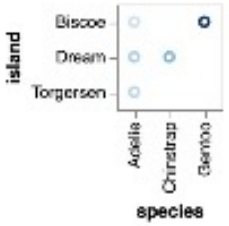


After aggregation (i.e. frequency counting), you can do basic visual analysis





☺ Good for accurate comparison of frequencies

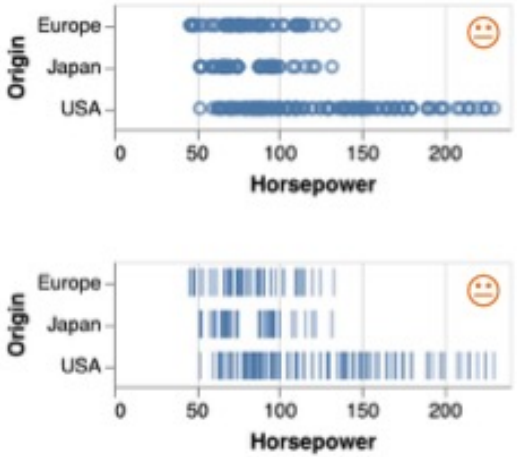


☺ Good for checking existence and rough comparison of frequency with small space

2D Nominal x Quantitative

- If you are interested in how one nominal and one quantitative columns
  - E.g., **origin** and **horsepower** columns of the car dataset

Raw data is already useful



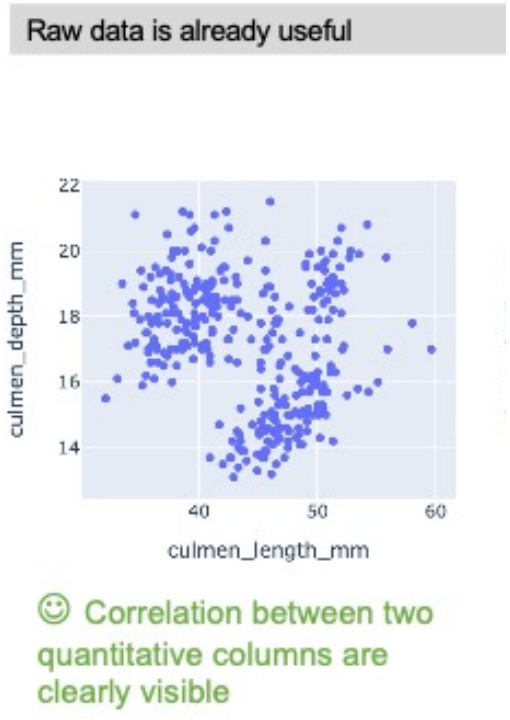
A typical use case of scatterplot. However, we can do better by applying descriptive stats on horsepower

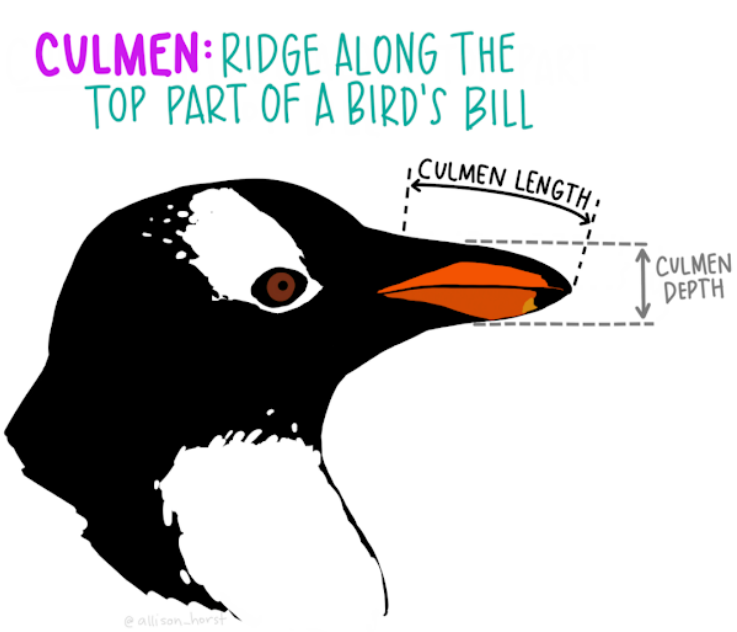
With statistical summary on quantitative values, we can draw box, violin plots

☺ We can compare distributions!

2D Quantitative x Quantitative

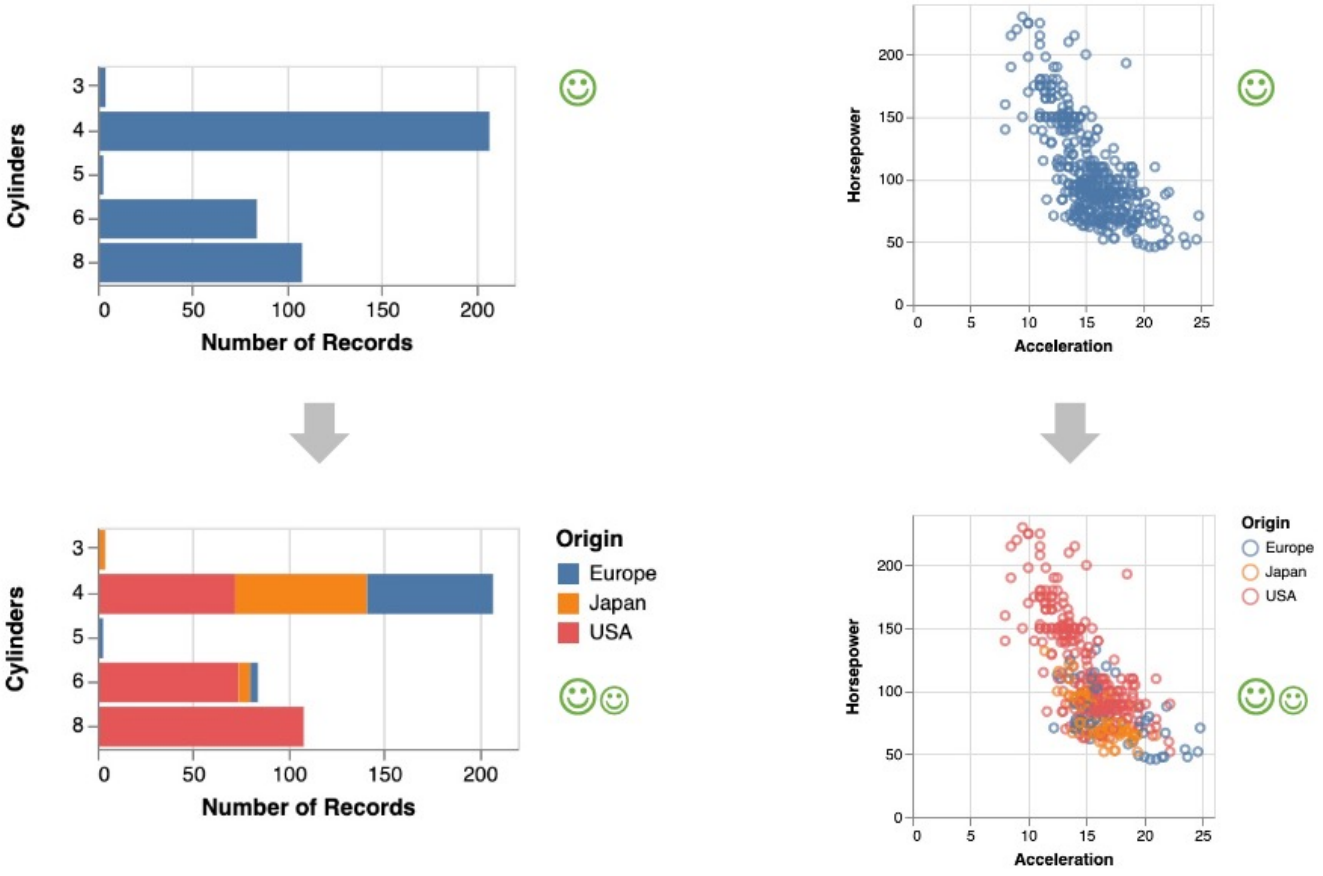
- If you are interested in how two quantitative columns
  - E.g., **culmen length** and **culmen depth** columns of the penguin dataset





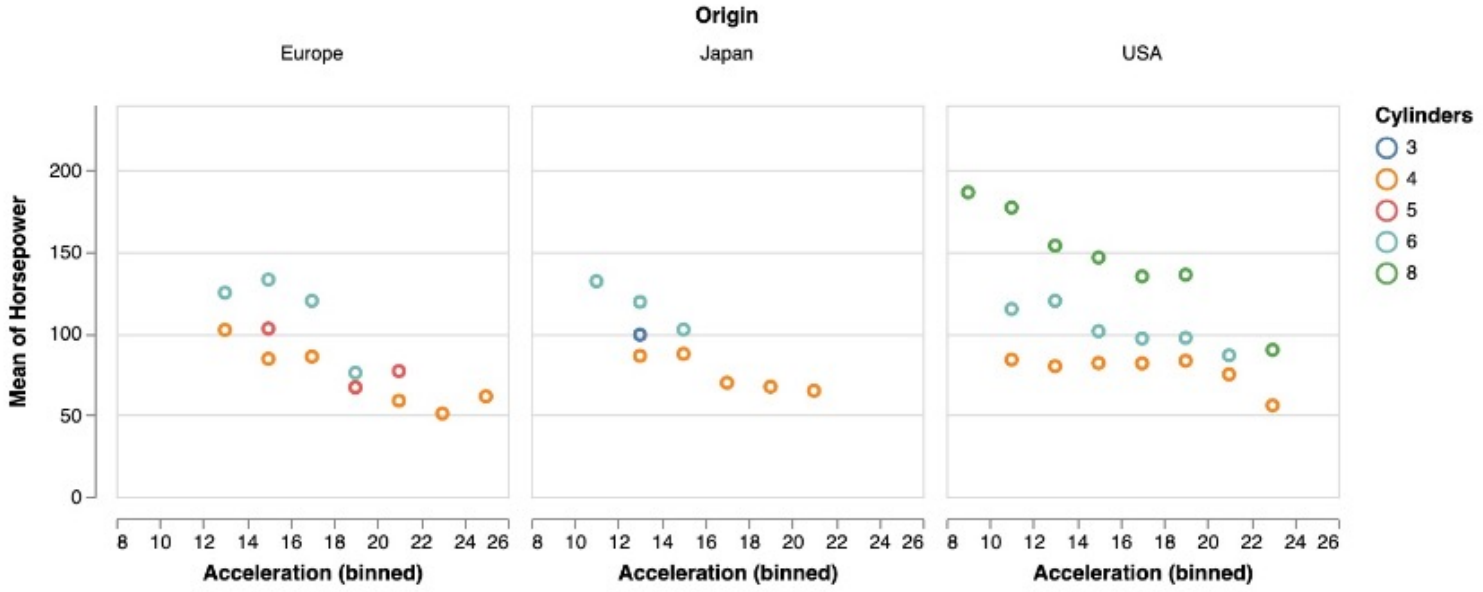
3D ANY

- Each visualization can accommodate 1-2 extra columns with color or size encodings. Why not explore higher-dimensions?



Higher Dimension

- Single charts usually cannot accommodate larger than 5 dimensions.
  - However, we can use **composite charts**.
  - For example, we have used scatterplot matrix in the previous tutorial.

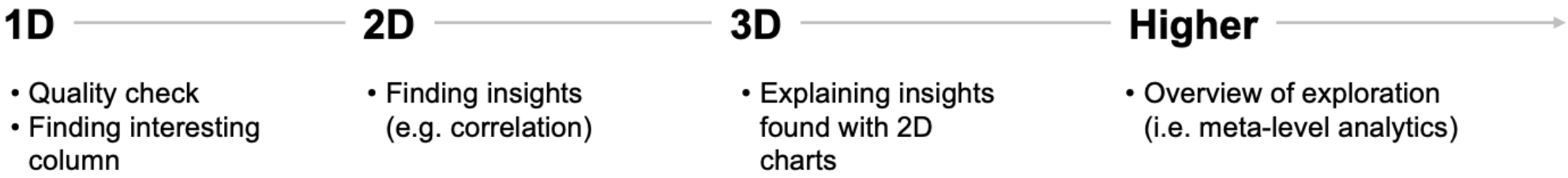


☺☺☺ Using subplots we can add another field

EDA Progression in general

- Why did we learn 1D, 2D, 3D, and higher? It seems that higher dimensions are better?

- 1. Data exploration usually starts with 1D for...
  - Checking data **quality** of each column
  - Finding **interesting** column for further exploration
- 2. # combinations grows very quickly for higher dimensions
  - E.g., If a dataset has 10 columns, there are 1000 combinations for 3D charts.
  - Thus, we need to narrow down columns to explore through 1D and 2D



Thank you~

Wan Fang
Southern University of Science and Technology