Set 2: Data Visualization Basics

Skill 2.1: Describe the purpose of data visualizations

Skill 2.1 Concepts

Data visualizations are how we communicate data. We don't read the numbers off a spreadsheet or list every number in a trend to communicate a point – we make a visualization to show them. Data visualizations allow us to,

- Explore and see something new in the data
- Strengthen data-driven arguments
- Share data-driven ideas
- Tell stories

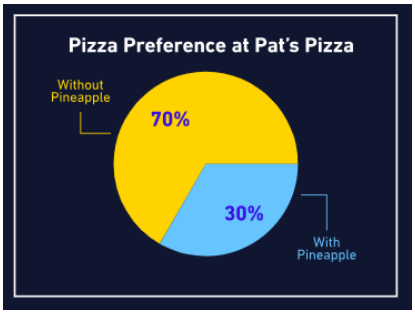
Skill 2.2: Identify an appropriate chart type

Skill 2.2 Concepts

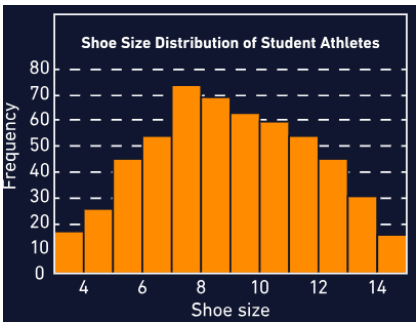
The first step of making a data visualization is choosing a chart type. Chart type isn't our only tool when it comes to visualizing data, but it's an important one for communicating about the relationship we want to show.
In this context, a "relationship" in the data could mean something like…

- "the shop's sales of Gouda were higher in 2021 than any year since 2006"
- "30% of people ordered pizza with pineapple"
- "most people in the sample have a shoe size between 6 and 10.5 "
- "as temperature increases, ice cream sales increase"

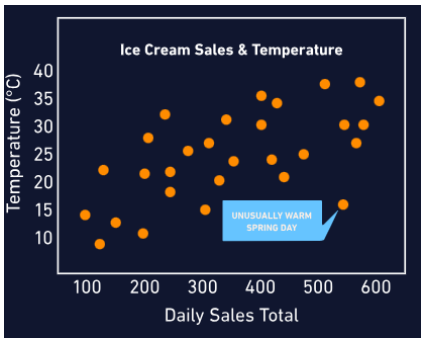
The first example is a change over time – that can be perfect for a line chart or a bar chart.



The second example compares a part to the whole: 30% of people got pizza with pineapple, out of 100% of people who ordered pizza. A pie chart is the classic (sometimes controversial) choice, but newer options include waffle and donut charts. Yum!



The third example is a distribution – the spread of data points in one variable. A histogram is the classic choice for visualizing a distribution.



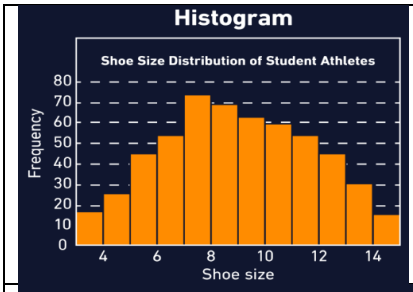
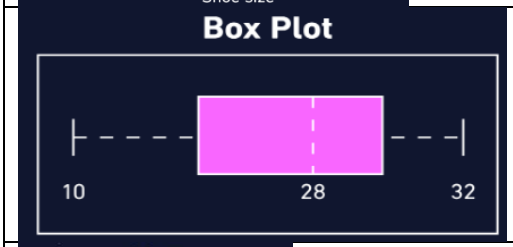
The fourth example is a direct comparison of two variables to help understand a trend. This is perfect for a scatterplot, with or without a trend line.

There's often more than one possible chart we can use for a dataset. But different charts emphasize different questions, arguments, or relationships in the data, and whichever we choose should help translate that data relationship into a visual relationship.

Univariate Charts

One big consideration when choosing a chart type is how many variables we are comparing. Univariate charts help us visualize a change in one variable. Often that means measuring "how much," which can either be a count or a distribution.

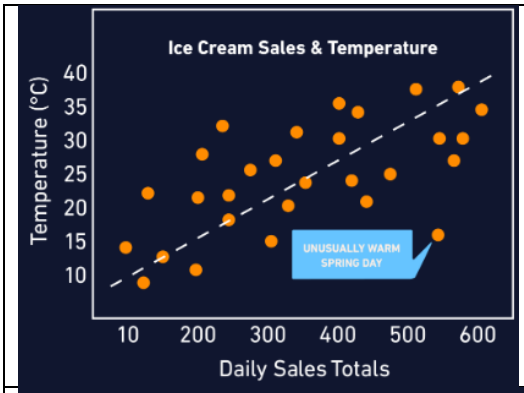
Below are examples,

	Histograms are a great way to show the concept of a normal (or skewed) distribution. We can visualize the answer to questions like… • "how does foot size vary across the population?" • "what is the distribution of pregnancy length across the human population?" • "how is income distributed in my country?"
	A more "math-forward" way to visualize distributions is a box plot or violin plot. These visualizations make percentile and quartile values obvious.
	A univariate map could be used to show location and distance.

Bi- and Multivariate Charts

These charts show the relationships between two or more variables.

Below are some examples,

	Scatterplots translate the relationship between two variables in the data into an easy-to-see spatial relationship. Because we're relying on the idea that each variable increases as we move up the X or Y axis, the scatterplot only makes sense for numeric variables, not categorical.
	A line chart is another common bivariate chart, often measuring a variable changing over time. A stock chart, for example, measures the value of a company over time. A line chart with multiple lines for different variables is a multivariate chart. For an example, check out the line chart that plots both imported and domestic cheese sales.
	A bivariate map shows a basic geographical map plus an additional variable — this example shows roughly where different pasta shapes originated in Italy. We can also map precipitation, altitude or depth, median income, museum locations, or combinations of variables… the list is endless.

Skill 2.2 Exercise 1

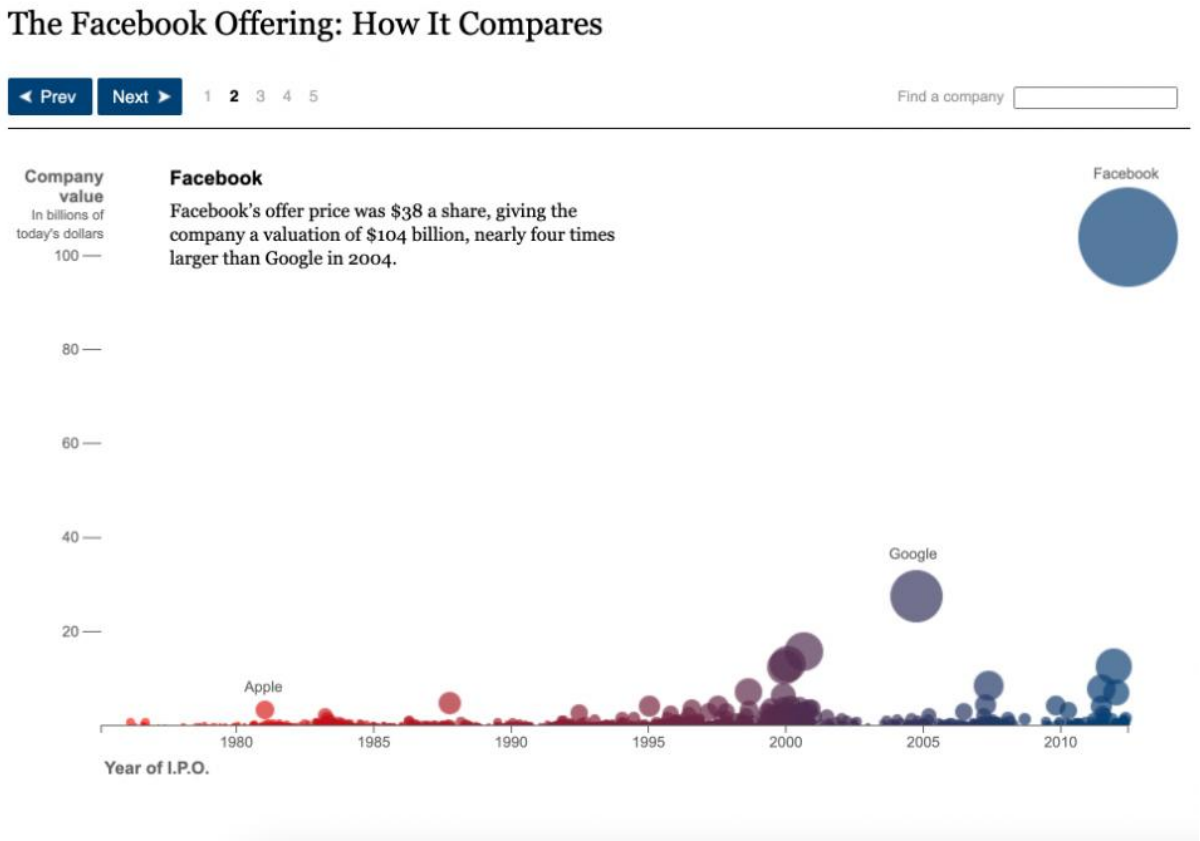
Skill 2.3: Describe the importance of aesthetic properties

Skill 2.3 Concepts

Aesthetic properties are the attributes we use to communicate data visually:

- Position
- Size
- Shape
- Color / pattern

Consider the visualization below published by The New York Times in 2012. It shows different tech IPOs (initial public offering, or when a formerly-private company becomes available as a publicly traded stock).



Let's walk through what this visualization is showing…

- Starting with position: from left to right (the x-axis), the graph measures time. From bottom to top (the y-axis) the graph measures company value in billions of dollars.
- The size of each circle tracks its company value. Companies with a larger IPO amount get a bigger circle.
- Color corresponds to time. Earlier corresponds to red, and later to blue. The middle portion, around 1995, is purple. This visually separates the three decades into three general zones.

Let's take a closer look at this graph. There's a connection here between size and y-position (how high or low a circle is): they actually tell us the same information twice!

This is an example of *information redundancy*, or encoding the same information in different visual properties. We already know that Facebook has the largest company value because it's the highest circle on the chart. Its large size gives us another way to visually compare it to the other data points.

Info redundancy is also helpful for prioritizing values. There are lots and lots of smaller companies on this graph – if every circle were the size of Google's circle, the bottom part of the graph would be an unreadable ball pit. Or, if all the circles were the size of the smallest ones, the chart would lose some of its emphasis on Facebook's large IPO value. *Information redundancy helps key data points to stand out.*

Skill 2.3 Exercise 1

Skill 2.4: Explain the importance of universal design as it applies to data visualizations

Skill 2.4 Concepts

Universal design for data visualizations means creating charts and graphs that are accessible and understandable to the widest possible audience, regardless of their abilities or background.

Universal design should be considered when it comes to,

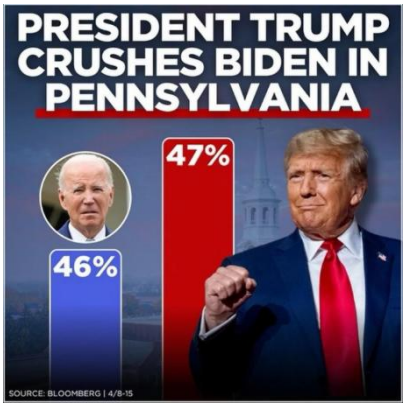
- Readability: keep the reading level to a high school level whenever possible
- Prior knowledge: define unfamiliar terms and avoid unnecessary jargon
- Information overload: introduce new information with intentional pacing and organization
- Visual impairments: keep in the mind the following,
    - Colorblind-friendly color palettes
    - Large enough font size
    - Readable, web-accessible font type
    - Alt text on data visualization images online

Skill 2.4 Exercise

Skill 2.5: Recognize the elements of good data visualizations

Skill 2.5 Concepts

While the best graphs really do teach us something new, and help us understand a deeper truth using data, graphs can also be misleading, both intentionally or unintentionally. We shouldn't ever assume that a data visualization shows us the truth, the whole truth, and nothing but the truth.



Here we will consider the elements of misleading and confusing graphs so that we can avoid making them ourselves.

Axes

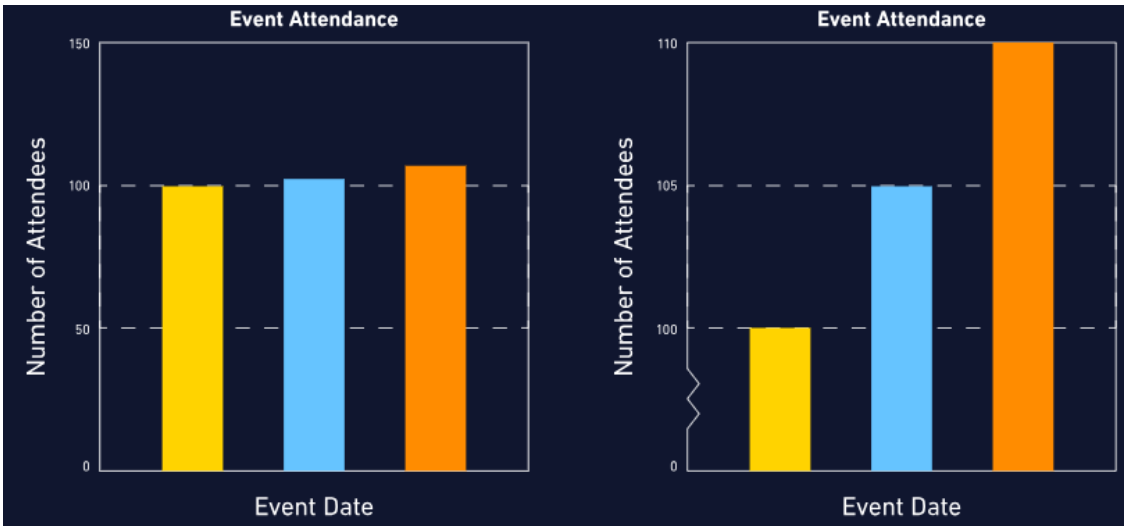
Axes and scaling are like the page layout and spacing of a paper book: they're not the most exciting parts, but they do present plenty of opportunities to make it harder to read.

Let's start with axes – the x-axis (horizontal, left-right) and y-axis (vertical, top-bottom). A common misleading aspect of an axis is a **break**. A break starts the count at a number that's not zero, or jumps ahead – this can distort the amount of

difference between data points by removing context, and make small differences in data seem bigger.

Here's an illustration of that idea: it would be almost impossible to tell at a glance if there were 100, 105, or 110 people standing in a room – but you'd be able to easily tell the difference between 0, 5, and 10 people standing in a room. Using a break on an axis can have the same effect, amplifying the **change** rather than the **context** because it alters the proportions in the visualization.

Check out the graphs below to the right to see what this looks like in practice.



So what to do? If you're looking at a graph, take a second to check where the axis starts. If there's a break, factor that in as you think about what the numbers mean.

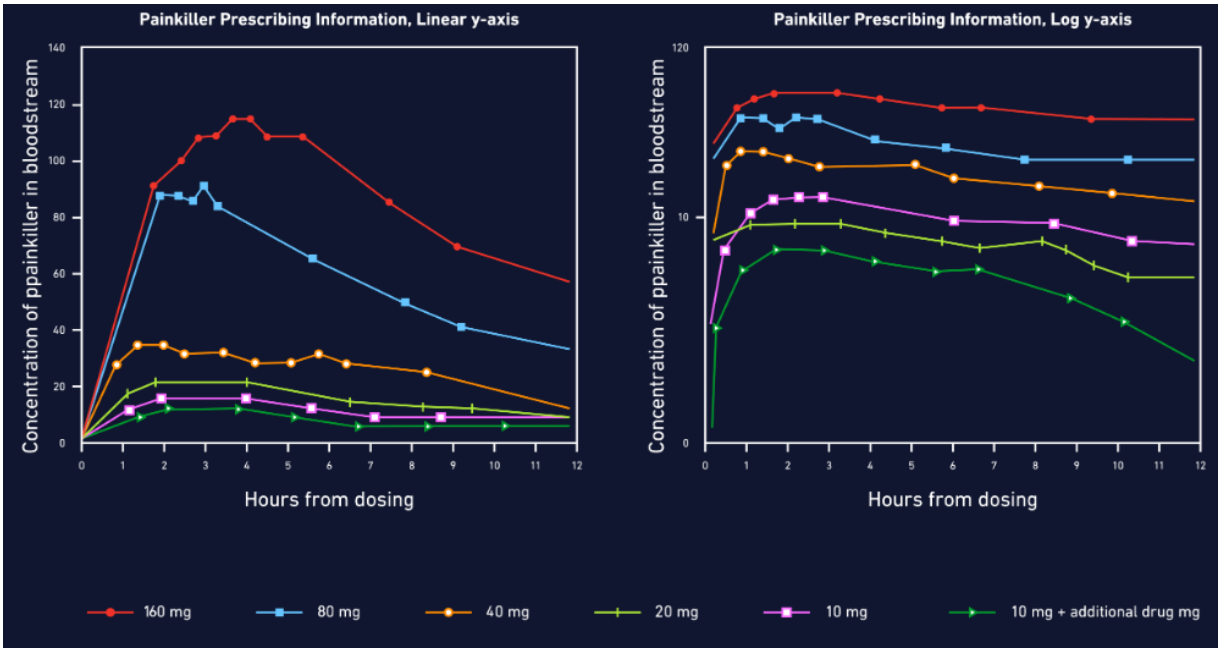
If you're making the graph, instead of using a big break…

- Keep enough context to view differences **in proportion** to a meaningful amount, OR
- Make two graphs, one without a break and one "zoomed in", OR
- Choose a visualization type that shows the change, rather than the raw numbers

Scaling

Scaling refers to the distances between numbers on an axis. Almost all graphs use a **linear scale**, where the numbers count up by a consistent interval – tenths of a centimeter or millions of dollars, if it's the same interval, it's a linear scale.

The other scaling option is a **logarithmic scale**, a.k.a. log scale. The log scale is common for showing exponential growth that won't fit on the page with a linear scale, but it's almost never a good choice for a general audience. Unless people use log scales regularly, they tend to have trouble interpreting them correctly.
Check out the graphs below to see how the pharmaceutical company Purdue infamously used this misinterpretation to their advantage in the early 2000s. The linear scale shows how the concentration of a painkiller drug spikes sharply in the bloodstream at higher doses – the log scale makes it look like all doses behave pretty similarly. (These are reproductions of the original graphs, but we can definitely see how differently they represent the same numbers.)



In general, just like it's always worth checking for a break, it's always worth checking how a graph is scaled.
Last thing about axes and scaling: **generally, we measure time horizontally**, putting that variable on the x-axis. For the vast majority of circumstances, this makes the most sense and helps readers to intuit what the graph measures.

Color scales

Color is often the first thing we register when looking at data visualizations. There are three types of color scales, used for the three major types of relationships we can visualize with color.

Sequential scales are colors in a sequence – often, this is the same hue with more and more white added to or taken away from the color. Sequential scales are used to show a variable increasing or decreasing in intensity or amount, like income, depth, or percent of population that owns a chinchilla.

Divergent scales are anchored by colors from opposite sides of the color wheel, a.k.a. complementary colors. A divergent scale is used to visualize data where the middle is a baseline, and either side represents a contrasting change. For example, divergent scales do a good job of showing a positive/negative swing in voting or polling, temperatures above and below freezing, or gains and losses over time.

Categorical scales use a variety of colors to differentiate categories without assigning a rank or order to them. In other words, "purple" doesn't necessarily mean more than "green" – the two are just different colors. Categorical scales are for

categorical data, like types of vegetables in a supermarket, or different treatments tested in a controlled study, or organizational blocks on a calendar.

Color associations

Once we've picked the right color scale, there are still a few more considerations to be made to reduce confusion.

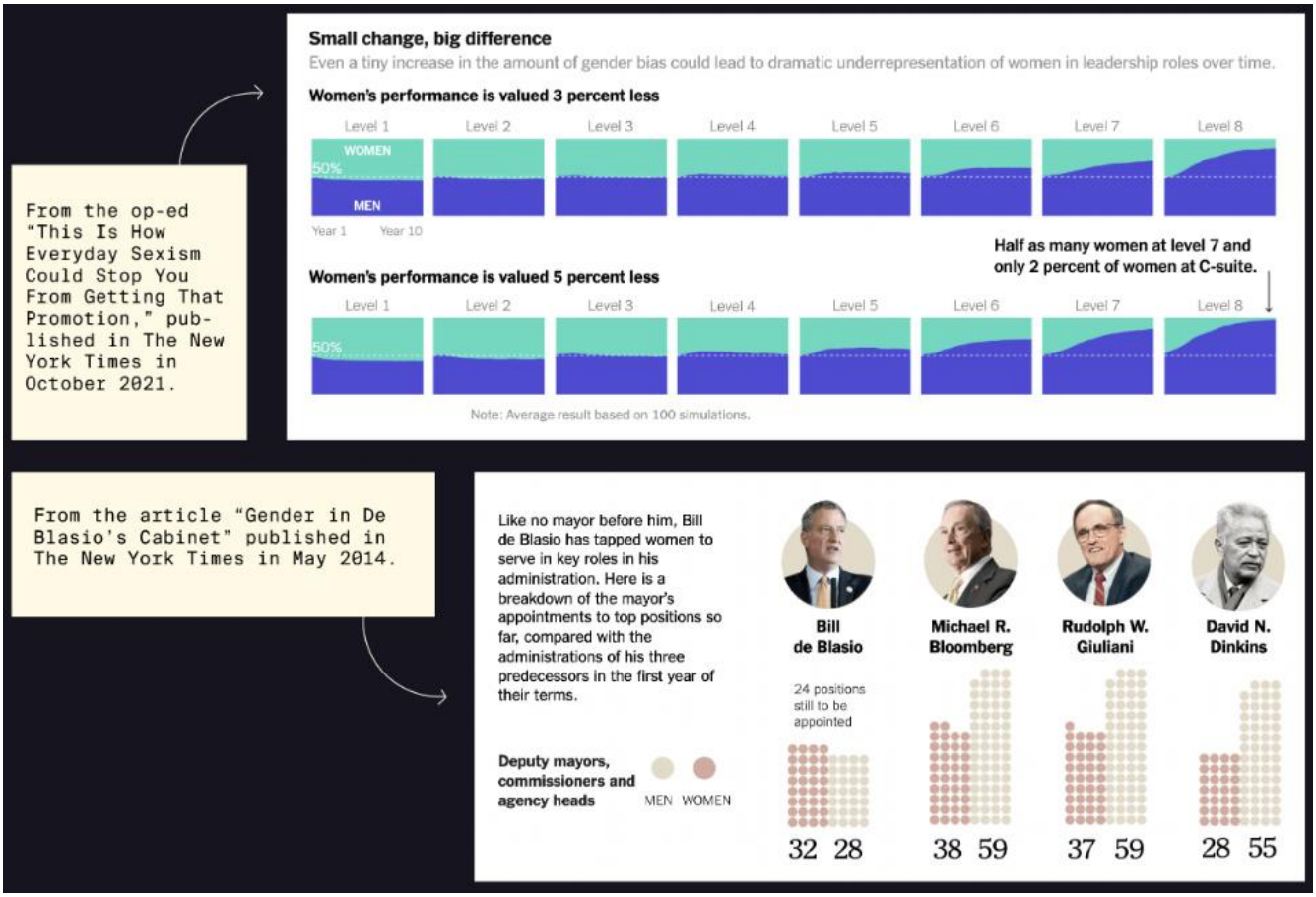
First up, we tend to view darker colors as "more" and lighter colors as "less." For example, if we're visualizing which US states have the most pet ferrets, California – with the most pet ferrets of any state – should be the darkest state on the map. When this scale is reversed, people will tend to just read the graph wrong rather than reading the legend carefully.

We also come to data visualizations with pre-existing associations for certain colors. These can be culturally specific (red means bad vs. red means lucky), or influenced by the norms for a particular field (red means negative financial balance).

Sometimes it's good to stick with what's recognizable: it would be confusing for US voters if a major newspaper decided to visualize Democrats in red and Republicans in blue, since these colors are overwhelmingly associated with the opposite party.

But in other cases, switching up colors that have existing cultural associations can reduce harmful stereotyping. Using pink for women and blue for men reinforces an outdated, binary view of women as soft and passive, and men as strong and unemotional. This design choice will not only turn off some viewers, it may also distract on a graph where gender is a relevant variable but not the whole focus.

It would be confusing to just reverse this stereotypical color palette, but there are plenty of good alternatives – check out two examples from The New York Times below. The important thing is to be consistent with the alternative palette



Labels and titles

A good title is one of the best and fastest tools for making a more understandable visualization. Lots of confusion can be saved with a descriptive title.

If the graph doesn't have a good title (or even a title!), viewers have to do more legwork to first figure out what each axis measures and then what the data points show.
The title can be a question that visualization answers, like, "Who speaks more in Disney movies, male or female characters?"

Titles can also be a statement of what the visualization shows, like "Comparing denim inseam lengths through the decades" or "Millennials really do spend more on rent than on avocados" or "The effect of hunger on mood level."

Like a good title, annotations on a graph also help the viewer to understand what's going on. Annotations are perfect for calling out points of interest, explaining outliers, or including background information that a viewer won't necessarily know from just looking at the graph.

Check out this *Live Births* graph from FlowingData to see how much value the annotations add. They…

- add detail to the highest and lowest points on the graph
- explain what the 0% baseline means
- provide a caveat for the 2021 data
- reinforce in words that the percents on the y-axis show "more births" and "fewer births"

Just a few lines of thoughtful annotation here and there give the audience so much more ability to interpret the graph at a deeper level!