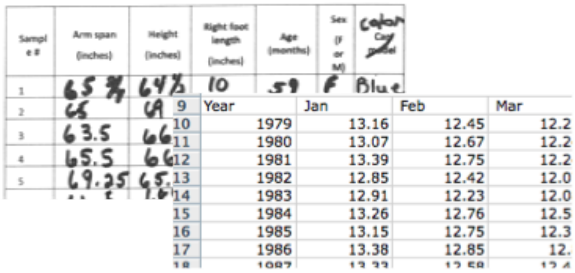
Thinking your way through data analysis

So you have a table of data. What now?

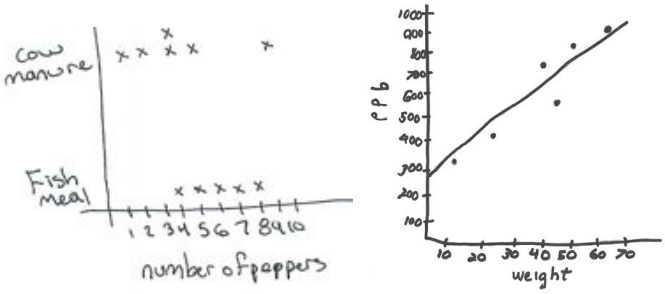


	Sampl e #	Arm span (inches)	Height (inches)	Right foot length (inches)	Age (months)	Sex (F or M)	color Curl tongue
1		65 ¾	64 ½	10	59	F	Blue
2		65	69	9	Year	Jan	Feb	Mar
3		63.5	66	10	1979	13.16	12.45	12.2
4		65.5	66	11	1980	13.07	12.67	12.2
5		69.25	65.	12	1981	13.39	12.75	12.2
				13	1982	12.85	12.42	12.0
				14	1983	12.91	12.23	12.0
				15	1984	13.26	12.76	12.5
				16	1985	13.15	12.75	12.3
				17	1986	13.38	12.85	12.

Whether you have data that you collected yourself, or data accessed from an online archive, your first task is to turn an array of numbers into a visually meaningful form of evidence – usually a graph.

A graph (or graphs) displays the data as evidence to support a claim or an answer to the question that you are investigating.

1. If the dataset is very small (a dozen points or fewer), and not in a digital spreadsheet, sketch a quick graph by hand. This can give you a quick sense of what you have. Making a dot can quickly show how similar or different two groups are or how two factors are related. It does not have to be fancy!



2. To make a graph using a computer, enter the data into a spreadsheet program, such as Excel or Google Sheets. Then upload the spreadsheet into a "drag and drop" type of data visualization platform such as Tuva, CODAP, Tinkerplots, or Fathom. These platforms allows you to simply drag and drop different "attributes" (variables) to different axes, select different graph types, and (or) easily filter the data to look at only subsets if desired.

How to begin? A scientist might approach a new dataset in a couple of ways.

Explore and play with data

Sometimes a scientist has no clear idea of what can be learned from a new dataset, so she or he might begin to freely explore the different attributes it contains. She might put various attributes on different axes and try out different types of graphs. She may even filter out certain categories, or look at only a sub-range of the data to focus on a particular subset with no clear question in mind, just to see what is there. *Does anything interesting or puzzling pop up during the exploration?*

Exploration is a great way to become familiar with a new dataset and poke around with what it might be able to tell you. Free exploration can even help you develop a question or pose a claim.

Dragging and dropping attributes and using the filter in Tuva make it easy to freely and quickly explore data.

But at some point, exploration begins to feel fruitless or even frustrating. At this point a more deliberate approach can help you move towards finding a rewarding story in the dataset.

Transform data into evidence deliberately, for a specific purpose

The challenge of data analysis involves making a series of purposeful decisions, such as *What kind of graph should I make?*

Analyzing data requires thinking about what you might want to know from the data. The following questions can help you begin, or return to a productive track if you feel derailed:

--> What data do I have to work with?

Review the dataset attributes. Which are categorical and which are numeric? What groups are included in the categorical attributes? For example, if Location is an attribute, click on Location to see which locations are included in the dataset. Click on a numeric attribute to see the range of values. If it is a time attribute (such as hour, date, or year), what time span does the dataset cover? What is the source of the dataset? (Is the source reputable?)

Example: The table of data below includes mean temperatures (monthly mean, high, and low) and precipitation for 12 locations (1 African, 1 Southeast Asian, and 10 US cities). The dataset comes from the National Climate Data Center, supported by NOAA. It is reputable.

Case	Location	Month	Mean Temp (F)	High Temp (F)	Low Temp (F)	Total Precipitation (In)
73	Singapore	Jan	78.4	85.8	73.6	7.72
74	Singapore	Feb	79.5	87.8	74.3	6.01
75	Singapore	Mar	80.2	88.5	75	6.67
76	Singapore	Apr	81	89.1	75.7	5.5
77	Singapore	May	81.5	88.9	76.3	6.16

(Fig. 1)

--> What do I want to find out?

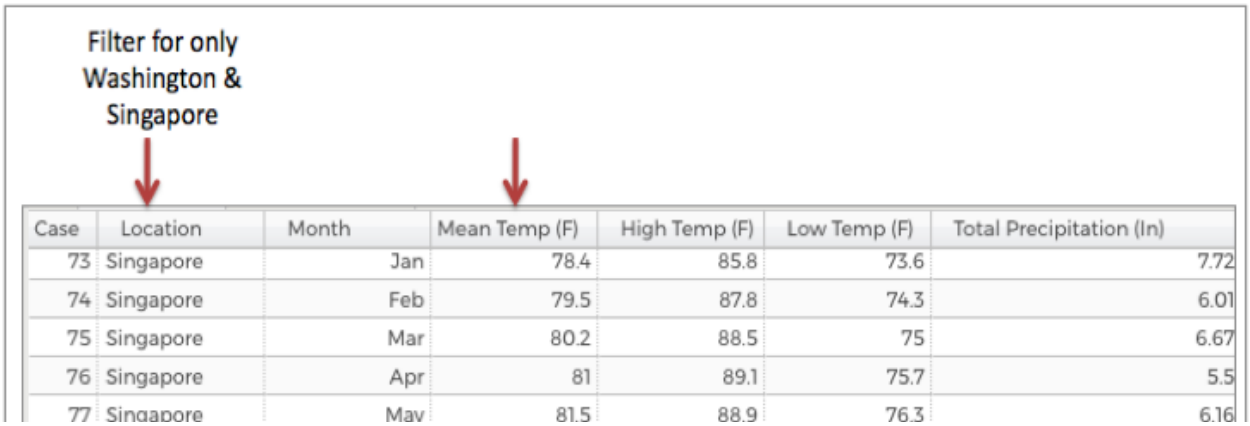
Based on the attributes and groups that are available in the dataset, what questions could the dataset help answer? What would you like to find out? Write down a question, claim, or hunch you would like to investigate. Having a clear question provides a basis for making decisions in your analysis.

Example: How do average temperatures in Washington DC and Singapore change through the year?

--> Which attributes (or subsets of data) do I need to answer the question?

Identify which of the attributes you'll need to answer the question. Think about whether or not some of the attributes can be filtered to focus on a subset that relates directly to the question (such as a particular location, or a certain period of time).

Example: Attributes needed for the question above are Location and Mean Temperature. Location can be filtered to show data for just Singapore and Washington.



Case	Location	Month	Mean Temp (F)	High Temp (F)	Low Temp (F)	Total Precipitation (In)
73	Singapore	Jan	78.4	85.8	73.6	7.72
74	Singapore	Feb	79.5	87.8	74.3	6.01
75	Singapore	Mar	80.2	88.5	75	6.67
76	Singapore	Apr	81	89.1	75.7	5.5
77	Singapore	May	81.5	88.9	76.3	6.16

(Fig. 2)

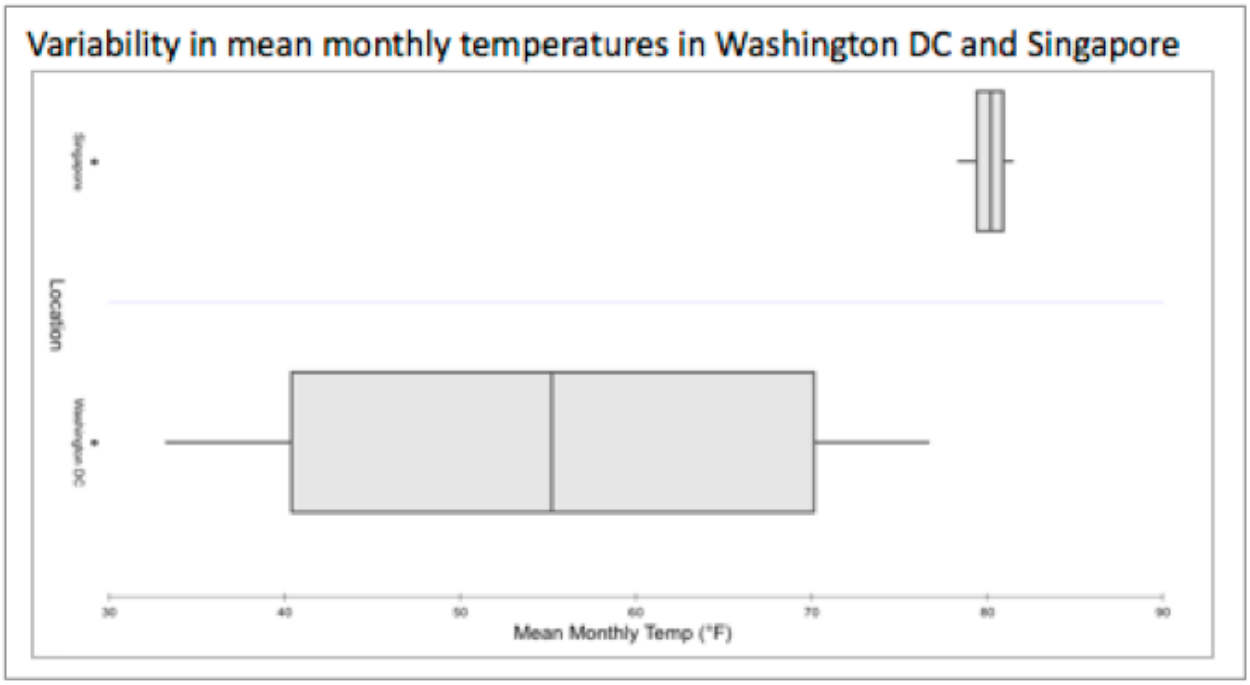
--> What kind of graph should I use?

Determine what kind of question you are asking.

- Is your question about variability or consistency within a group? (*How fast do roller coasters go?*) H
- How similar or different are two or more groups? (*Do wooden and steel roller coasters have similar or different speeds?*)
- What is the a relationship between two numeric attributes? *(What is the relationship between roller coaster heights and speeds?)*
- How does something change through time? *(How have roller coaster heights changed since the first roller coaster was built?)*
- What is the composition of a group? (What proportion of roller coasters are wood and what proportion are steel?)
- How is something geographically distributed? (Where are roller coasters located?)

Then use the Graph Choice Chart to help you decide what kind of graph to use based on the kind of question you are asking.

Example: To compare two cities in mean monthly temperatures, a dot or box plot works well (Fig. 3). If your question is: *How do mean monthly temperatures at each city change through the year?*, you could use a line graph to show how temperatures changed through time. (Fig. 4).



(Fig. 3)

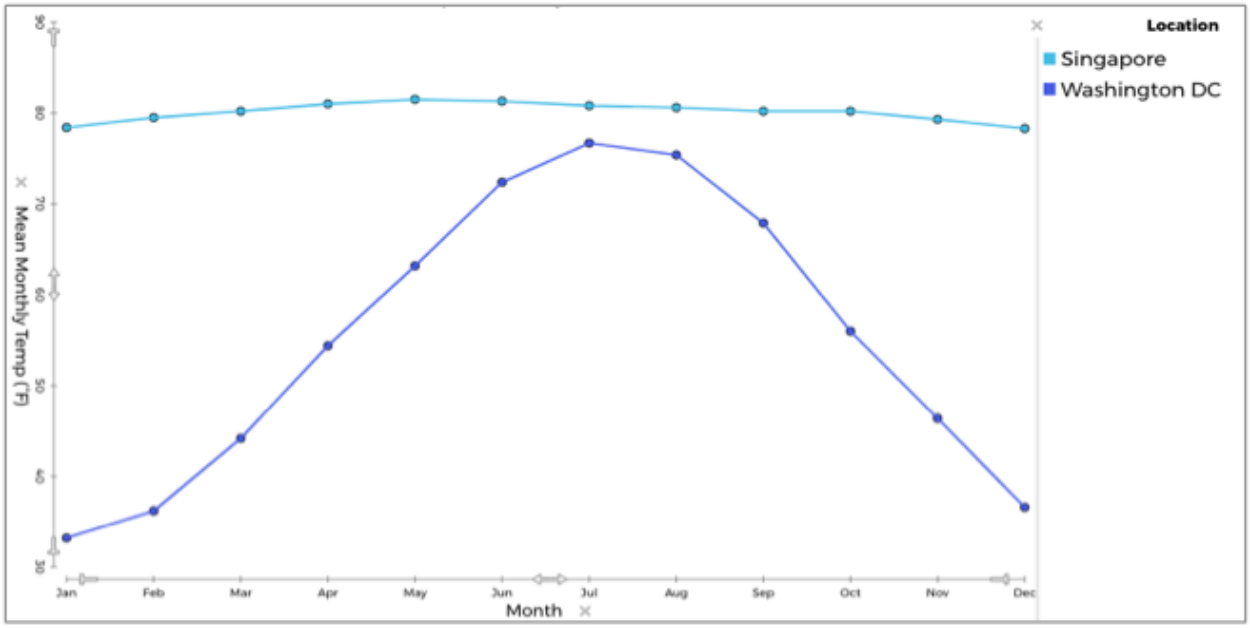


Fig. 4

--> What patterns or relationships do I see in the graph?

Is something steadily increasing? Are changes gradual or abrupt? Does it rise and fall periodically? (Line graph). Is one group more wide-ranging than the other? (Box plot, dot plot, or histogram). Is there a linear or oval shape in where the points lie, or do they look random? (Scatter plot). Is one sector much larger than all the rest? (Pie or stacked bar graph).

For example, In Washington, mean monthly temperatures are around 32° in January, then rise steadily until July they are 75°, then fall steadily back down to around 35° in December. (Fig. 3)

--> What do the patterns or relationships suggest about the question?

Describe the patterns or relationships you see. Be specific. Include quantities in your description.

For example*, instead of "Singapore is warmer", a more compelling description would be "Singapore's climate is warmer and more consistent through the year than Washington's is. The median temperature is 55° in Washington DC, and it's around 80°F in Singapore. Mean monthly temperatures in Washington vary by 45° (from around 33°F in January to almost 78°F in July), and in Singapore they vary only 3 degrees (between 78°F and 81°F) throughout the year."*

--> What does the evidence mean *in terms of a general scientific process or a wider context?*
Connect your evidence to a fundamental scientific concept that is possibly at work, or to implications in a wider context. If you found that the pulse rate generally increases after exercise in 12 different trials, what biological response could be at work? How likely would it happen for other people who you didn't measure?

For example, if a graph shows winter temperatures are rising overall, but are highly variable from year to year, connections to make could be:

"The rising trend line is consistent with the claim by scientists that Earth's climate is warming."

"CO2 in the atmosphere traps heat, and therefore could make winters warmer in the long term, but not necessarily from one year to the next."

"The year to year variability means that one year winter might be warm and the next year could be colder, which is a signal of weather. But in the long term, the trend is warming, and that is climate change."

--> What strengths or limitations does the evidence have?

Here is your chance to talk about how confident you are in your claim. Do the data come from a reliable source? Are there any gaps in the data? Is the number of samples large enough to fairly represent reality? Does the degree of variability among cases cast any doubt on the strength of the claim?

Example: The data come from the National Climate Data Center, based on 30-year averages for each month. There are no obvious known gaps in the data. The trend line has a steep positive slope. I am confident in my interpretation.

--> What new questions are raised by the data?

When a scientist analyzes one aspect of a dataset, new questions about other aspects of the data often arise. What new questions arise for you? Don't be afraid to wonder and follow your nose with a new question.

Example: How do the two cities compare in precipitation? How to the cities compare with a third city at a different latitude?

_____

Summary:

You have a dataset, but you aren't sure what to do with it.
Play with it for a bit, see what attributes and categories it includes, then think about what you can learn from it -- what you would like to find out.

--> What data do I have to work with?

--> What do I want to find out?

- Something about variability or consistency within a group? (*How fast do roller coasters go?*)
- How similar or different are two or more groups? (*Do wooden and steel roller coasters have similar or different speeds?*)
- What is the relationship between two numeric attributes? *(What is the relationship between heights and speeds of roller coasters?)*
- How does something change through time? *(How have roller coaster heights changed since the first roller coaster was built?)*
- What is the composition of a group? (What proportion of roller coasters are wood and what proportion are steel?)
- How is something geographically distributed? (Where are roller coasters located? -- Possible if you have latitude and longitude data).

--> Which attributes (or subsets of data) do I need to answer the question?

--> What kind of graph should I use *(based on my kind of question)?*

--> What patterns or relationships do I see in the graph?

--> What do the patterns or relationships suggest about the question?

--> What does the evidence mean *in terms of a general scientific process or a wider context?*

--> What strengths or limitations does the evidence have?

--> What new questions are raised by the data?