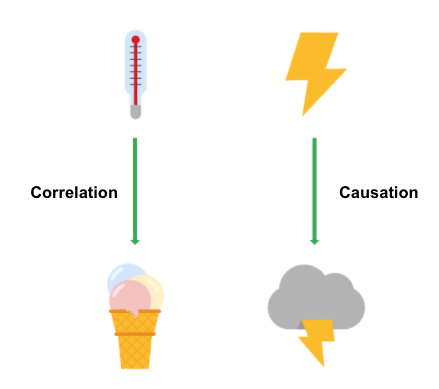
* **Correlation** in statistics is the measure of the degree to which two variables move in relationship to each other. An example of correlation is the idea that “As the temperature goes up, ice cream sales also go up.” It is important to remember that correlation doesn’t mean that one event causes another. But, it does indicate that they have a pattern with or a relationship to each other. If one variable goes up and the other variable also goes up, it is a positive correlation. If one variable goes up and the other variable goes down, it is a negative or inverse correlation. If one variable goes up and the other variable stays about the same, there is no correlation.
* **Causation** refers to the idea that an event leads to a specific outcome. For example, when lightning strikes, we hear the thunder (sound wave) caused by the air heating and cooling from the lightning strike. Lightning causes thunder.



## **Why is differentiating between correlation and causation important?**

When you make conclusions from data analysis, you need to make sure that you don’t assume a causal relationship between elements of your data when there is only a correlation. When your data shows that outdoor temperature and ice cream consumption both go up at the same time, it might be tempting to conclude that hot weather **causes** people to eat ice cream. But, a closer examination of the data would reveal that every change in temperature doesn’t lead to a change in ice cream purchases. In addition, there might have been a sale on ice cream at the same time that the data was collected, which might not have been considered in your analysis.

Knowing the difference between correlation and causation is important when you make conclusions from your data since the stakes could be high. The next two examples illustrate the high stakes to health and human services.

### **Cause of disease**

For example, pellagra is a disease with symptoms of dizziness, sores, vomiting, and diarrhea. In the early 1900s, people thought that the disease was caused by unsanitary living conditions. Most people who got pellagra also lived in unsanitary environments. But, a closer examination of the data showed that pellagra was the result of a lack of niacin (Vitamin B3). Unsanitary conditions were related to pellagra because most people who couldn’t afford to purchase niacin-rich foods also couldn’t afford to live in more sanitary conditions. But, dirty living conditions turned out to be a correlation only.

### **Distribution of aid**

Here is another example. Suppose you are working for a government agency that provides food stamps. You noticed from the agency’s Google Analytics that people who qualify for food stamps are browsing the official website, but they are leaving the site without signing up for benefits. You think that the people visiting the site are leaving because they aren’t finding the information they need to sign up for food stamps. Google Analytics can help you find clues (correlations), like the same people coming back many times or how quickly people leave the page. One of those correlations might lead you to the actual cause, but you will need to collect additional data, like in a survey, to know exactly why people coming to the site aren’t signing up for food stamps. Only then can you figure out how to increase the sign-up rate.

## **Key takeaways**

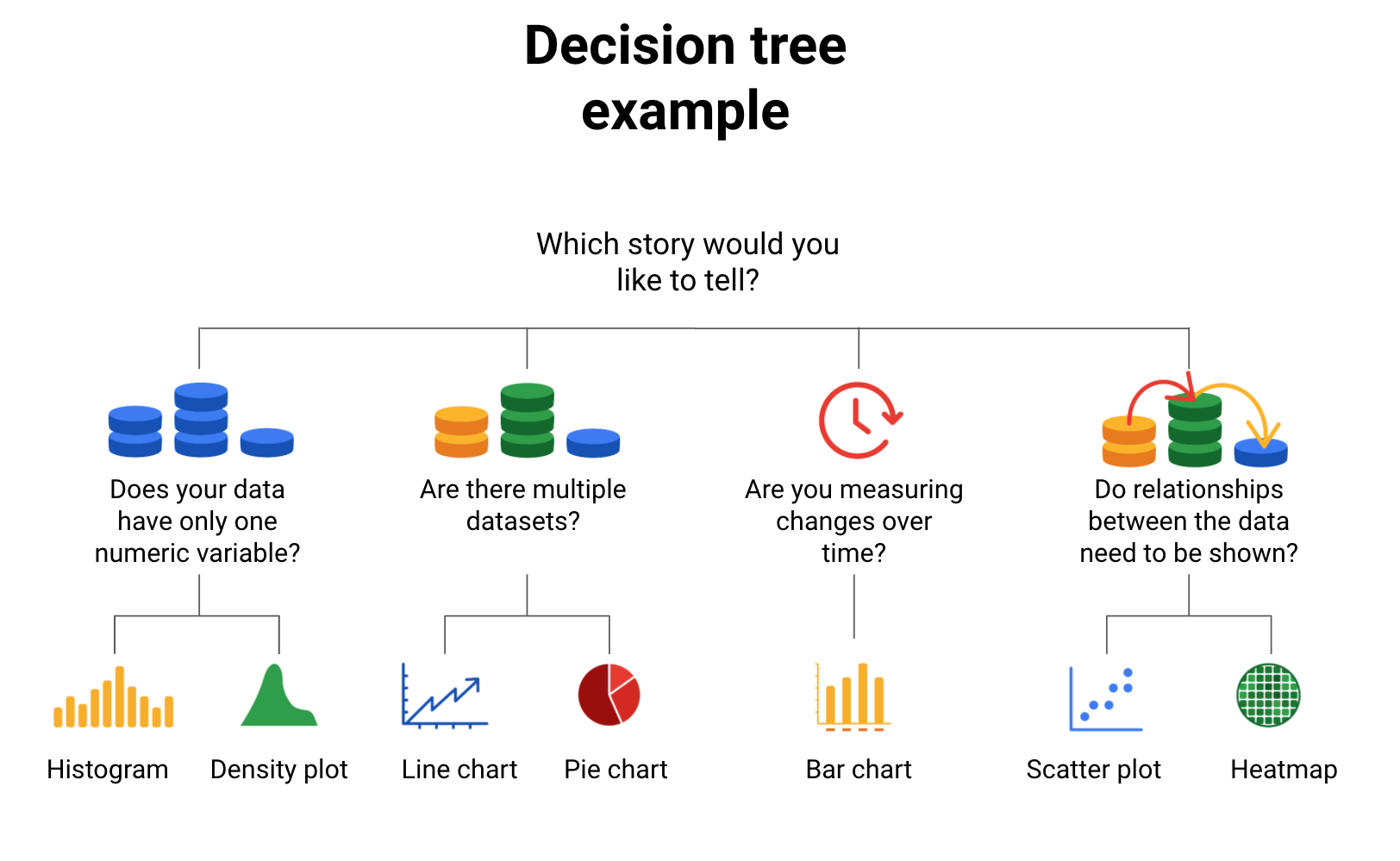
In your data analysis, remember to:

* Critically analyze any correlations that you find
* Examine the data’s context to determine if a causation makes sense (and can be supported by all of the data)
* Understand the limitations of the tools that you use for analysis

# Data grows on decision trees

With so many visualization options out there for you to choose from, how do you decide what is the best way to represent your data?

A **decision tree** is a decision-making tool that allows you, the data analyst, to make decisions based on key questions that you can ask yourself. Each question in the visualization decision tree will help you make a decision about critical features for your visualization. Below is an example of a basic decision tree to guide you towards making a data-driven decision about which visualization is the best way to tell your story. Please note that there are many different types of decision trees that vary in complexity, and can provide more in-depth decisions.



-Does your data have only one numeric variable? Histogram or Density plot

-Are there multiple data sets? Line chart or pie chart

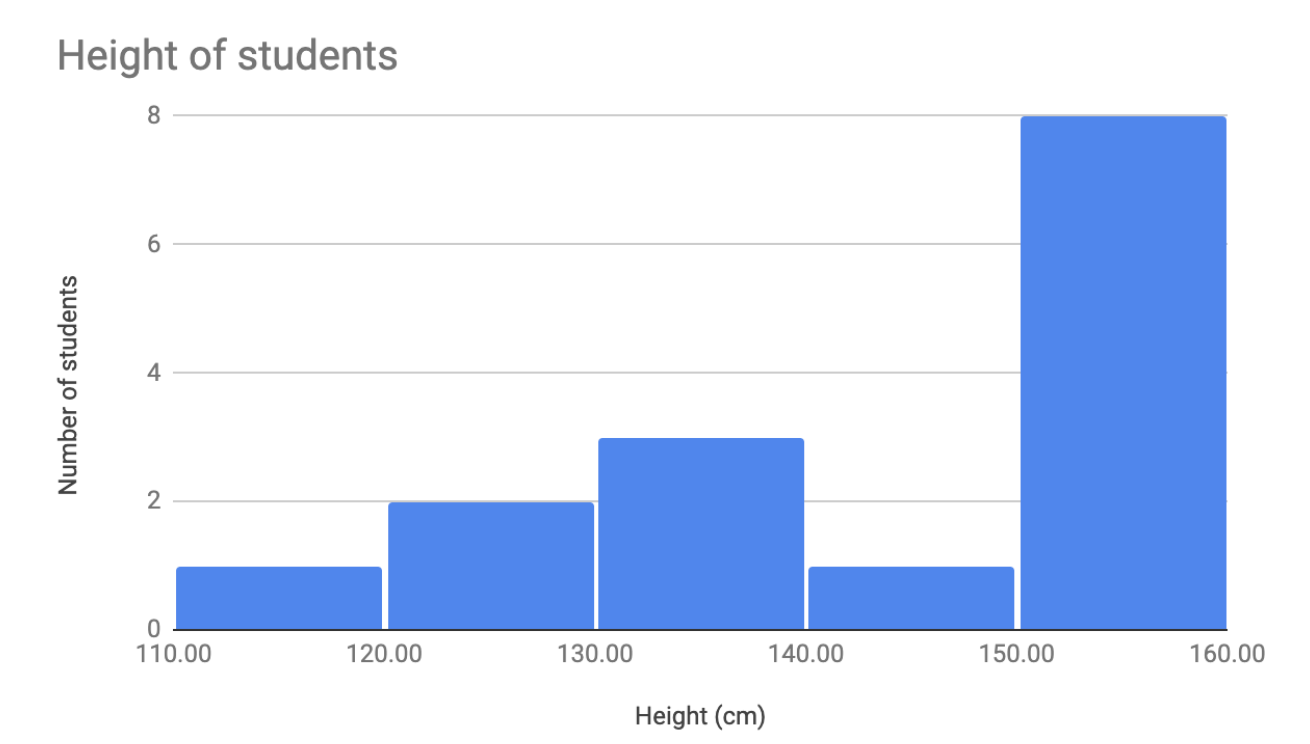
-Are you measuring changes over time? Bar chart

-Do relationships between the data need to be shown? Scatter plot or heatmap

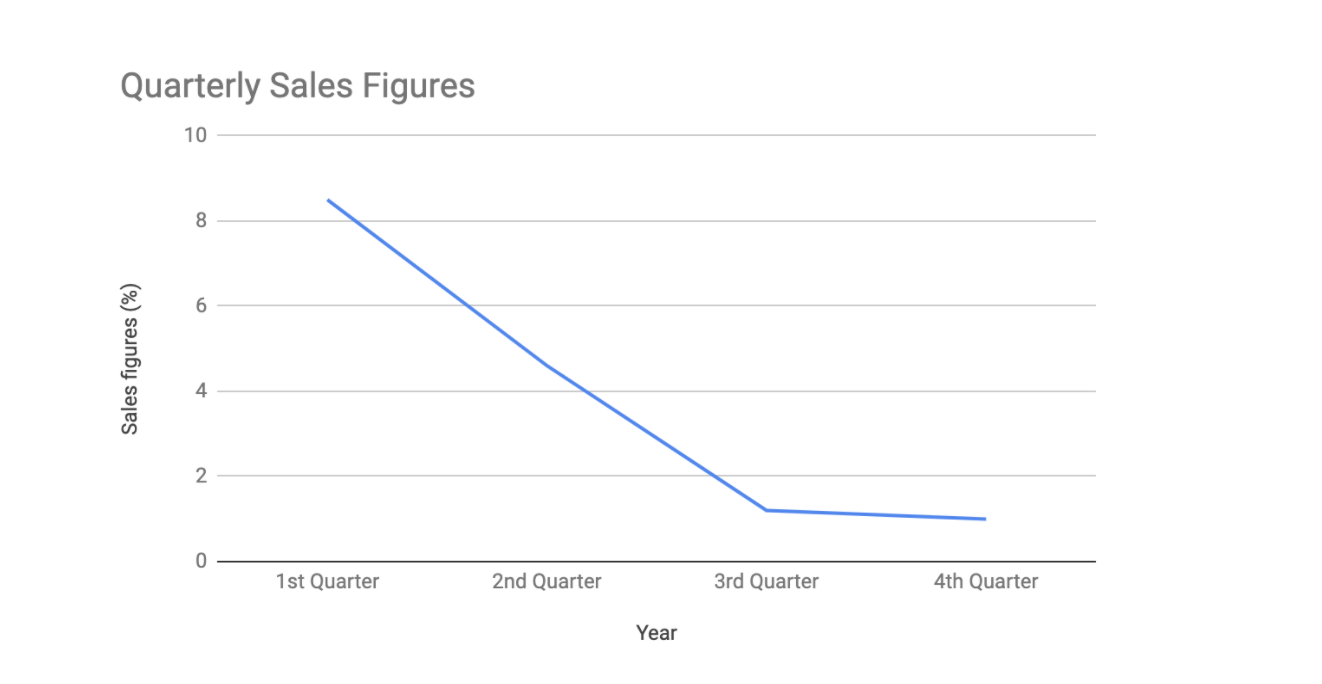
## **Begin with your story**

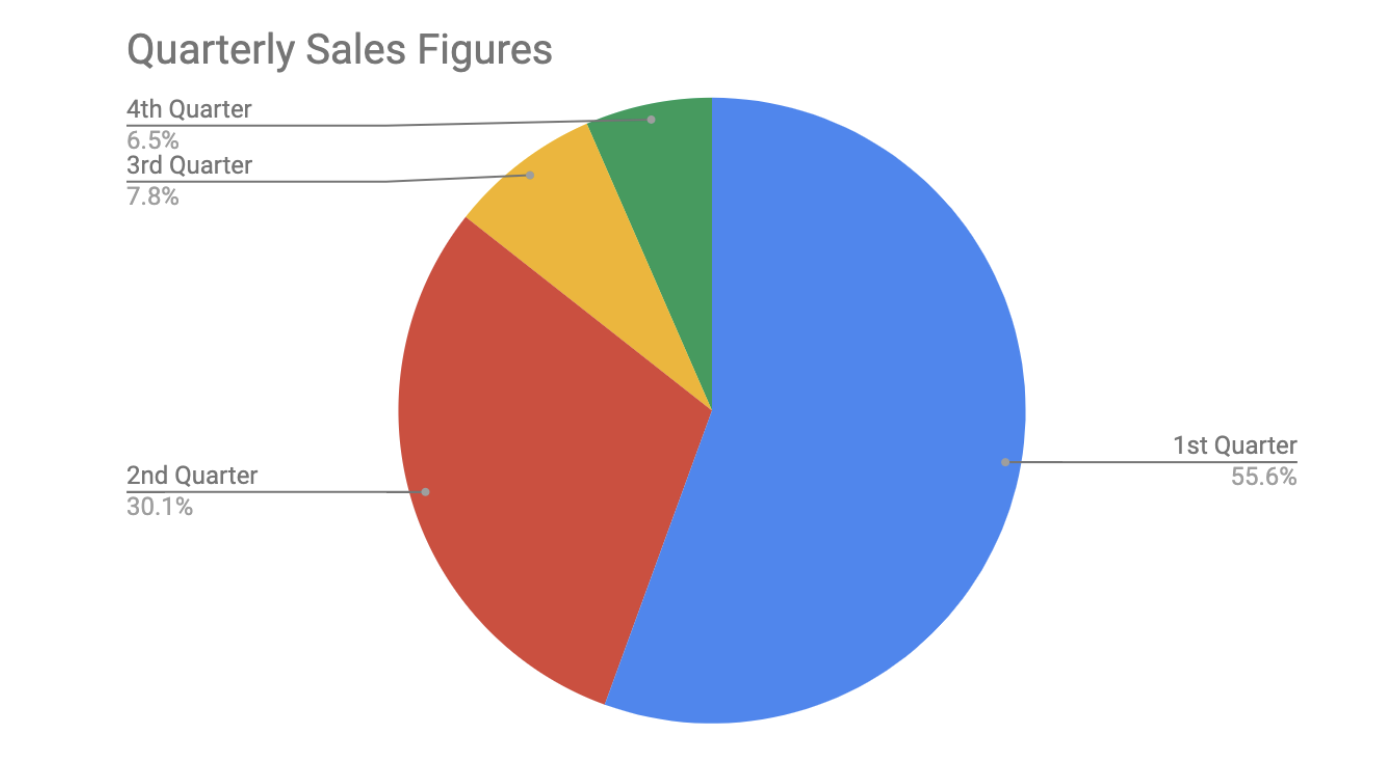
Start off by evaluating the type of data you have and go through a series of questions to determine the best visual source:

* **Does your data have only one numeric variable?** If you have data that has one, continuous, numerical variable, then a histogram or density plot are the best methods of plotting your categorical data. Depending on your type of data, a bar chart can even be appropriate in this case. For example, if you have data pertaining to the height of a group of students, you will want to use a histogram to visualize how many students there are in each height range:



* **Are there multiple datasets?** For cases dealing with more than one set of data, consider a line or pie chart for accurate representation of your data. A line chart will connect multiple data sets over a single, continuous line, showing how numbers have changed over time. A pie chart is good for dividing a whole into multiple categories or parts. An example of this is when you are measuring quarterly sales figures of your company. Below are examples of this data plotted on both a line and pie chart.

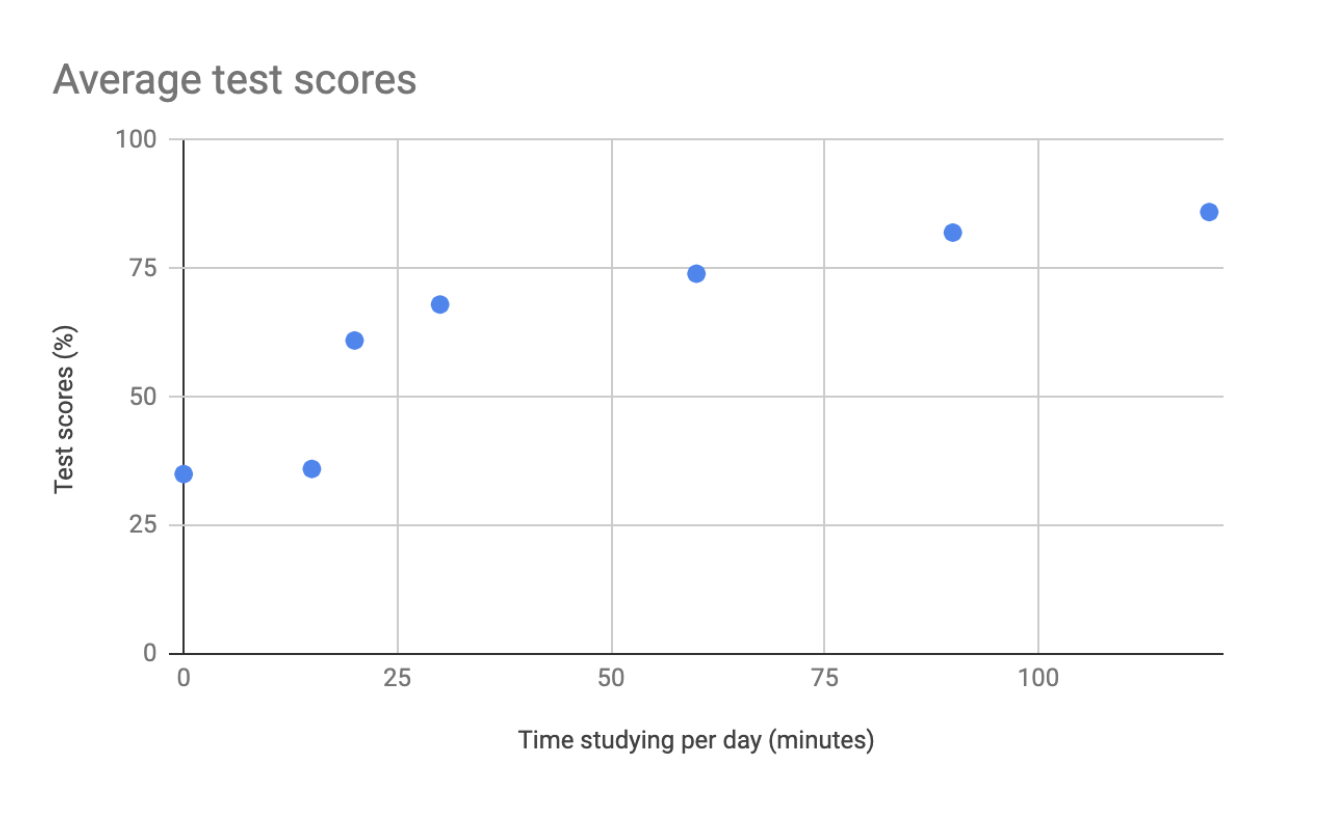




* **Are you measuring changes over time?** A line chart is usually adequate for plotting trends over time. However, when the changes are larger, a bar chart is the better option. If, for example, you are measuring the number of visitors to NYC over the past 6 months, the data would look like this:



* **Do relationships between the data need to be shown?** When you have two variables for one set of data, it is important to point out how one affects the other. Variables that pair well together are best plotted on a scatterplot. However, if there are too many data points, the relationship between variables can be obscured so a heat map can be a better representation in that case. If you are measuring the population of people across all 50 states in the United States, your data points would consist of millions so you would use a heat map. If you are simply trying to show the relationship between the number of hours spent studying and its effects on grades, your data would look like this:



## **Additional resources**

The decision tree example used in this reading is one of many. There are multiple decision trees out there with varying levels of details that you can use to help guide your visual decisions. If you want more in-depth insight into more visual options, explore the following resources:

* [From data to visualization](https://www.data-to-viz.com/): This is an excellent analysis of a larger decision tree. With this comprehensive selection, you can search based on the kind of data you have or click on each graphic example for a definition and proper usage.
* [Selecting the best chart](https://www.youtube.com/watch?v=C07k0euBpr8): This two-part YouTube video can help take the guesswork out of data chart selection. Depending on the type of data you are aiming to illustrate, you will be guided through when to use, when to avoid, and several examples of best practices. [Part 2](https://www.youtube.com/watch?v=qGaIB-bRn-A) of this video provides even more examples of different charts, ensuring that there is a chart for every type of data out there.