**1. A plot tells a thousand words**

Hi, I'm Richie. Welcome to Data Visualization for Everyone!

**2. What you'll learn**

In this course, you'll learn how to choose an appropriate plot to answer common questions about different data types. You'll also learn how to interpret common plot types like histograms, box plots, scatter plots, line plots, and bar plots. Finally, you'll learn about best practices for drawing plots, and how to avoid pitfalls. Throughout this course you will see lots datasets from many different fields. I want you to get a sense that data is everywhere, and that drawing plots is a great way to make sense of it.

**3. Three ways of getting insights**

There are three main ways of getting insight from a dataset. Firstly, you can calculate summary statistics. That includes measures of quantity like mean or median, and measures of variation like standard deviation. Secondly, you can run statistical models like linear regression and logistic regression to model relationships between variables. Thirdly, you can visualize data by drawing plots like a scatter plot or a histogram. These three ways are often used together and have different benefits. In this course, we'll focus in on the third way.

**4. The Datasaurus Dozen**

The Datasaurus dozen is a collection of 13 datasets, with names like away and bullseye. Each dataset has two variables: the x and the y coordinates. "Variable" is just statistics jargon for a column of data.

1. 1 Matejka, J., & Fitzmaurice, G. (2017) https://www.autodeskresearch.com/publications/samestats

**5. Mean of x for each dataset**

If you calculate the mean of the x values in each dataset, you can see that it's more or less the same value.

**6. Mean of x and y for each dataset**

It's the same situation for the means of the y coordinates. The value is the same in each dataset.

**7. Standard deviations for each dataset**

Similarly, we can look at the variation of the x and y values by calculating the standard deviation for each dataset. Variation describes how spread out values are. Each dataset has the same standard deviation for x and y.

**8. plotting dino**

Here is a scatter plot of each dataset, and even a quick glance shows what the calculations failed to. That is, every dataset is completely different. Until you physically look at the datasets, it's hard to tell that you have lines and circles and a star and a dinosaur. The datasets are artificial, but I hope this example has convinced you of the importance of plotting your datasets.

**9. Continuous and categorical variables**

Before diving deeper into plotting, it's important to acknowledge that there are different types of data. Choosing a type of plot depends on whether your variables are continuous or categorical. Continuous variables are numbers, such as heights, or temperatures or revenues. You can do arithmetic on continuous variables, like adding two temperatures together.

**10. Continuous and categorical variables**

Categorical variables are things that can be classified, and are usually written as text. They include eye color, which takes the categories blue, brown and a few others. Other examples include country and industry, which have a longer list of categories.

**11. Continuous and categorical variables**

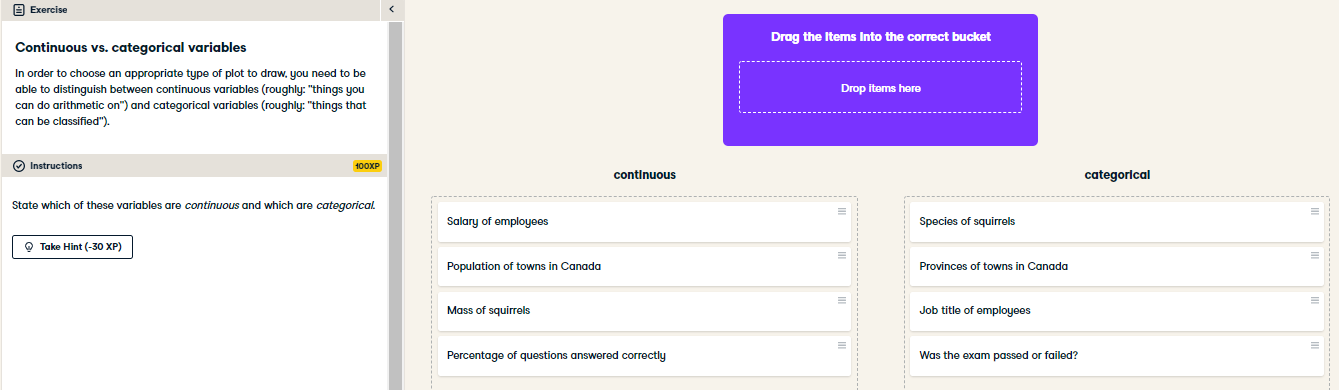
Finally, some things can either be continuous or categorical. Age is a number, so by default it's a continuous variable. However, many surveys use age groups like 25 to 30. Those age groups are categories. Similarly, time is naturally a continuous variable, but if you have to produce a report on how prices change each month, you might want to think of the month of the year as a categorical variable. In this case, you have the freedom to treat the time as either continuous or categorical. It depends on the question you're trying to answer!

**12. Let's practice!**

It's time for your first set of exercises!



Perfect price plot probing! Often, looking at a plot is a quicker way to answer questions than poring through data.



Classy classification! The distinction between continuous variables (numbers) and categorical variables (text) is important when deciding what kind of plot to draw.