MITOCW | MIT15_071S17_Session_7.2.05_300k

In this video, we'll discuss the meaning of data visualization, and why it's often useful to visualize your data to discover hidden trends and properties.

Data visualization is defined as a mapping of data properties to visual properties.

Data properties are usually numerical or categorical, like the mean of a variable, the maximum value of a variable, or the number of observations with a certain property.

Visual properties can be (x,y) coordinates to plot points on a graph, colors to assign labels, sizes, shapes, heights, etc.

Both types of properties are used to better understand the data, but in different ways.

To motivate the need for data visualization, let's look at a famous example called Anscombe's Quartet.

Each of these tables corresponds to a different data set.

We have four data sets, each with two variables, x and y.

Just looking at the tables of data, it's hard to notice anything special about it.

It turns out that the mean and variance of the x variable is the same for all four data sets, the mean and variance of the y variable is the same for all four data sets, and the correlation between x and y, as well as the regression equation to predict y from x, is the exact same for all four data sets.

So just by looking at data properties, we might conclude that these data sets are very similar.

But if we plot the four data sets, they're very different.

These plots show the four data sets, with the x variable on the x-axis, and the y variable on the y-axis.

Visually, these data sets look very different.

But without visualizing them, we might not have noticed this.

This is one example of why visualizing data can be very important.

We'll use the ggplot2 package in R to create data visualizations.

This package was created by Hadley Wickham, who described ggplot as "a plotting system for R based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts.

It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics." So what do we gain from using ggplot over just making plots using the basic R functions, or what's referred to as base R?

Well, in base R, each mapping of data properties to visual properties is its own special case.

When we create a scatter plot, or a box plot, or a histogram, we have to use a completely different function.

Additionally, the graphics are composed of simple elements, like points or lines.

It's challenging to create any sophisticated visualizations.

It's also difficult to add elements to existing plots.

But in ggplot, the mapping of data properties to visual properties is done by just adding layers to the plot.

This makes it much easier to create sophisticated plots and to add to existing plots.

So what is the grammar of graphics that ggplot is based on?

All ggplot graphics consist of three elements.

The first is data, in a data frame.

The second is an aesthetic mapping, which describes how variables in the data frame are mapped to graphical attributes.

This is where we'll define which variables are on the x- and y-axes, whether or not points should be colored or shaped by certain attributes, etc.

The third element is which geometric objects we want to determine how the data values are rendered graphically.

This is where we indicate if the plot should have points, lines, bars, boxes, etc.

In the next video, we'll load the WHO data into R and create some data visualizations using ggplot.