- Next, we will talk about exploring data.

and can show skewness or unusual dispersion.

Boxplots are another type of plot

for showing data distribution.

Line graphs are useful for seeing how values in your data

change over time.

Spikes in the data are also easy to spot.

Scatter plots can show correlation between two variables.

Overall, there are many types of graphs

that visualize data.

They are very useful in helping you

understand the data you have.

Looking ahead, we'll actually spend

the better part of a week just focused

on effective ways to visualize data.

In summary, what you get by exploring data,

is a better understanding of the complexity

of the data you have to work with.

This, in turn, will guide the rest of the process.

After this video, you will be able

to explain the importance of exploring data

and identify methods to perform

preliminary analysis of your data.

After you put together the data that you need

for your application,

you might be tempted to immediately build models

to analyze the data.

We must resist this temptation.

The first step, after getting your data,

is to explore it, as exploring data is a big part

of the two-step data preparation activity.

You want to do some preliminary investigation

in order to gain a better understanding

of the specific characteristics of your data.

In this step, you'll be looking for things like

correlations, general trends, outliers.

And without this step, you will not be able

to use the data effectively.

Correlation graphs can be used to explore the dependencies

between different variables in the data.

General trends show you a simple graph

of how the data is progressing over time.

And outliers show you the data points

that are distant from other data points.

Plotting outliers will help you double-check

for errors in the data due to measurement.

In some cases, outliers that are not errors

might make you find a rare event.

Additionally, summary statistics

provide the numerical values to describe your data.

Summary statistics are quantities

that capture various characteristics of a set of values

with a single number or a small set of numbers.

Some basic summary statistics

that you should compute for your data set

are mean, median, mode, range, and standard deviation.

Mean and median are measures of the location

of specific values.

Mode is the value that occurs most frequently

in your data set.

And range and standard deviation are measures

of spread in your data.

Looking at these measures will give you an idea

of the nature of your data.

They can tell you if there's something wrong

with your data.

For example, if the range of the values for age in your data

includes negative numbers or a number much greater

than hundred, there's something suspicious in the data

that needs to be examined.

Python gives you functions and methods

that we can quickly explore the data

and provide these statistics.

You'll be hearing all about those by the end of this course.

Visualization techniques also provide quick and effective,

overall, a very useful way to look at data

in this preliminary analysis step.

A heat map, for instance, such as the one shown here,

can quickly give you an idea where the hot spots are.

Many different types of graphs can be used.

Histograms show the distribution of the data