0:13
In any data set, there are several different types of values or fields that you will find.
0:20
You can broadly bucket them into categorical, numerical, and Composite. And each has its
0:26
own characteristics. Categorical allows you to do relatively fewer computations. You can list them,
0:33
maybe you can sort them, but by and large, these are pieces of information that are there that
0:38
you have to infer by themselves. For example, if you have colors like red, blue, and green,
0:44
other than knowing that these are the colors, there is not that much you can do with them.
0:48
In comparison with Numerical, where there are a series of operations
0:52
that you can perform. These are numbers that could be real numbers, that could be integers.
0:57
But you can add them. You can take ratios; you can multiply two columns and perform several
1:03
derived operations out of these attributes. The third is Composite, where you have

1:10
even more operations that you can perform because Composite typically comprises multiple elements.
1:16
For example, composites may contain an array that has a list of numerical values
1:21
or a list of numerical and a list of categorical values. Let us look at the different types of
1:26
categorical values that are. Boolean is one example. Things that can be true or false, yes,
1:32
or a no, are just one of two values indicating something that exists or does not exist. Or
1:38
you could have unordered categories, unordered categories are categories that are distinct, and
1:44
you cannot sequence them in any way. For example, colors red, blue, and green, it is not like red
1:48

is bigger than blue is bigger than green. Or cities, it is not like London is bigger than
1:55
Paris, or so on. However, there are other ways in which you can order them. For example, you can
2:00
order cities by population. You can order cities by area. In these cases, we are not really using
2:06
the categorical value that is the name of the city itself to order. We are using another attribute,
2:12
which is their population or their area, you could also sort them alphabetically, but then again, you
2:17
are not treating them as individual categorical values. You are treating them as a sequence of
2:21
letters and then ordering based on the letters and alphabets, ABCD. They are, in fact, ordered.
2:28
And that brings us to ordered categorical things like low, medium, and high; the letters A,
2:33
B, and C, are things that you can place in a sequence. But there are some sequences that are
2:38
slightly different from just ordered. These are called Cyclical sequences, and an example would be
2:46
Monday, Tuesday, Wednesday, and Thursday, the days of the week, and it is not like Monday is
2:51
bigger than Tuesday is bigger than Wednesday, but you can order them in the sense that Monday comes
2:56
before Tuesday comes before Wednesday, and so on. And it is cyclical in that at the end, it repeats,
3:02
similarly, for January, February, March, etcetera.
3:05
These two can be considered under ordered, but you could treat them as a different sub-segment
3:10
of ordered called cyclical. And then, there is the unstructured categorical. These are similar to
3:17
unordered, except that we just recognize that it is not a specific set of values that you can have.
3:23

These are not enumerable. You could have anything literally. So, unlike the list of colors,
3:28
which is a finite and known, defined list of colors, or the list of months, which is
3:33
a defined list of 12 months, the text could contain anything, binary could contain anything,
3:38
it could be an attachment, it could be an image it could be a video, literally any possible item.
3:45
And each of these is distinct. Keep in mind that it is possible to get
3:49
derived values from categorical variables. Like for example, you can look up London's population
3:55
and create a new column. You can look up the length of a string that is ordered,
4:01
and use that as an attribute, you can look up the first letter. Or, if you have arbitrary text,
4:07
you can, for example, process the text to see if it has a positive or negative sentiment.
4:12
If it is an image, you can extract the names of people whose faces are in that image.
4:17
These are ways of extracting additional columns from categorical, but the categorical columns
4:22
in themselves have relatively few operations that you can perform on them directly. Numerical values
4:28
are what you are probably most familiar with. They could be integers, which could be negative
4:32
or positive like minus 2, minus 1, 0 1 2, or a subset of integers like whole numbers, which
4:38
start with 0 1 2, and so on, or just natural numbers which start with 1 2 3, and so on.
4:46
Or you could have fractions or decimals, which can be expressed as real numbers. And that can contain
4:53
1.5, 2.7, 3.1, or 3.5, or transcendental numbers like e and pi. Composite values combine multiple
5:02
values in a specific structure. You are probably familiar with some of them, like

dates or times.
5:08
A date is a day, a month, and a year combined into one structure, or time in hours, minutes,
5:17
and seconds combined into one structure. You can also have spatial structures. For example,
5:24
a point on the map can be characterized based on its latitude and longitude. Or you could have a
5:31
shape. It could be a line from a certain X and Y coordinate to another X and Y coordinate.
5:37
You could have an arc that passes through two points with a certain curvature. Or you could
5:43
have a ring, a polygon that goes from point A to point B to point C to point D, and so on. These
5:48
are examples of spatial structures that are used, for example, in shapefiles to create maps.
5:55
You could have structured data within a value. For example, a single field can comprise an
6:02
entire XML or JSON document, which internally has its own structure. And that structure could be
6:09
based on a specific schema that is well defined or could be completely arbitrary. Examples of
6:15
such composite structures are values that you will find when you parse tweets and get
6:21
JSON objects and arrays of JSON objects, each of which has its own set of fields inside them.
6:28
This apart, there are several specialized composite structures, like IP addresses. These can
6:33
be if it is an ipv4, it comprises of four integers from 0 to 255, or ipv6 is six such integers from
6:41
0 to 255. Currencies are just numbers but specialized with a specific prefix, let us
6:48
say the US dollar or the British pound. Composite values primarily are collections of values, and
6:56

they are very varied, but they, therefore, support even more operations than numerical types. And the
7:02
reason is that you can extract more information you have more fields to play around with. So,
7:07
it is a superset of everything that is categorical or numerical, plus a whole lot more.