2月2日_DataScience_数据类型_数据集类型

Tue, 2/2 9:05AM 51:05

SUMMARY KEYWORDS

data, terms, people, numerical data, module, type, calculate, variable, tableau, nominal, categorical, pandas, generally, postcode, week, privacy, table, values, gp, means

00:06

Oh, I will then off that video if I could watch that all day long on screen again. Okay, so to get the ball rolling. Good morning to all of you good ama. Well, and if you have any questions or queries or issues while going through this lecture that just like last week, please do type into chat. I monitoring that as we go. You'll have seen how very diverse This module is in terms of the students from the results of the survey that I posted this morning. That's normal for this module in fact usually but this module is even more diverse in terms of the countries. Although only about two thirds of the students on the module actually responded to the survey. I'm also aware that the zoom recordings of last week are not yet available I put a reply on themes about that the issue I'm facing is that, zoom, even though I've only given two sessions has produced 47 recordings. I haven't yet had the time to sit down and make sure that I download and make available to you, the recordings of the actual sessions that I gave, which is something that in the past has been automated but it is not yet automated. Switch to zoom. Anyway, today. Moving on from around our general data science process by moving on to the topic of data, understanding. As the borrower says there is nothing specific for you to look at in terms of self study for this week. So hopefully that will let you make progress of the getting ahead with a reading particularly of the data science textbook and also in terms of the download materials, we'll have another practical on Tableau this week and next week we'll be setting you up to start working on a farm as a practical analyzing some real data. So where we are in terms of this as a lecture. Let me turn on my right thing. Last week we were dealing with business understanding this week we're going on to data, understanding so this is a general understanding of data. And then next week we will start getting our hands dirty. Looking at the practicalities of preparing data for doing data science. So what I want you to take away from today will hope you're going to learn is the terminology that people use to describe data. And if you end up being slightly confused

by that then

04:59

yes that's nor. That is the real world. But we'll try to simplify things by explaining that there are really only three very basic types of data. Those are the most important to remember, but also to appreciate that data is not always what it appears to be, if you only look at the data source itself that is you need to consider something called semantics. We'll then go on to looking at some of the differences between types of data sources so open and close data, and using that to talk a little bit about preserving privacy privacy is something that you will go into much more detail about when you get into the two weeks of ethics material that starts off with a little bit of terminology. I'm going to think of data, and the sort of four levels of detail starting here on the right. A record a record in a data set or a data table. I will generally be using the term record, but the row or the term instance means exactly the same thing. I one thing you need to get used to is different people will end up using different terms for the very same thing. That is, well, that's the nature of life and as a data scientist, you just need to deal with it. So, record, that's like the rows here in the table. The columns in the table. These correspond directly speaking to the variables in the data set, and other terms that mean the same thing as they use. Sometimes our column or field or attributes. This whole table we say here you can think of this as being a data table, some people would call this a flat files online I call it by the abbreviation CSV stands for comma separated values. This is one of the common types of separase that you can add. And then of course, if you're dealing with large datasets. Then, complex datasets, then you will usually try to put them into a database where the database is in a relational database at least is made up of a number of separate tables so a very simple setup here imagine we have one table, storing information about people. So that's the person table we have here. And we had another day today will storing information about addresses of the houses, or house that you live in, then the relationship we have would be let's say a one to one relationship, whereby any given address could have several people living in it. In other words, you would have an address ID you'd add an address ID into this table here so you could actually link the records in this person table to the address table. If you studied computer science previously that all that will be second nature to you. If you are coming from one of the many many many other academic backgrounds that I know you have of this module we can see the results of that survey. Then some of this may be a little less familiar to you. And you should be gradually covering that in the other. You may be covering that in other computing based modules certainly if you're doing the MSc in data science and analytics, which is data science module is the is the poultry. That's the terminology about data sets, that's the easy bit errors and terminology about data types. The Annoying feature of PowerPoint is the way I have it set up, it changes slide for me in my second monitor but it

doesn't change

09:49

it notes for the slides is rather misleading. Right, okay. these are types. What you're seeing in this table here is a number of quite a large number of terms that can be used to describe different types of data. These are just terms that are drawn from the literature that I work with, don't have these all synonymous with each other. That is, they mean the same thing. Some of them don't mean quite the same thing but has similarities and naturally so I'm all for this. So, this is a how awake you all this morning or this afternoon depending on where you are in the world can, if you can see any of these terms which means the same thing. Can you just type them into chat, that type of pair of values into chat. Alternatively, just unmute yourself and sign up. Well daytime. There, I would say they are similar. So date time is going to be more fine grained than dates. Well, Jason doesn't appear on my list so Jason is that j that j so and

11:18

that is a type of format of data, rather than the data, hides itself so it's how the data is the sort of file structure that that data is presented into courses in a nominal.

11:39

So, just to show you where we are here. Nominal qualitative down here, these are identical they're exactly the same thing. And of course categorical also means the same thing. connection and length, Yes, connection here, and link them here. Then, one. Well quantitative measure No, I wouldn't go for that is great and continuous, well they are, they are opposites, based on geographic geographic is a type of spatial data, but also if you think of just position in x y Zed space that is also spatial data but you wouldn't think of that as being geographic knowledge nominal Intex. Thanks over here. Nominal. Generally speaking, yes, text, was the repeat defining categories but of course I might write, you know days of the week, Monday Tuesday Wednesday Thursday Friday as text. Yet, of course, they are referring to data that more generally is a type of date. And just looking at others, that is picked up so discreet and the interval. Now no that's not true. So we had the screen here and interval down here actually interval and ratio. All types of all types of numerical data numerical data is the same as quantum data those two terms are synonymous with each other. So, an example of interval data is temperature in Celsius. That is the difference from zero to five degrees is the same as the difference from five to 10 degrees. Whereas, Celsius is not ratio data, because 10 degrees is not twice as often

five degrees Celsius. However, if you were to express temperature in Kelvin, where absolute zero is zero Kelvin. that can be ratio data referential and hierarchical no that's. Those aren't connected to each other. Couple of the others that are similar to each other so boolean data has a true false on it. Well that's a type of categorical data where there are only two categories. You come up with some good answers to that. And for the purpose here is to introduce you to some of the complexity of data.

15:16

So, Moving on to some of the other terminology. One very very good, a very thorough book, which covers data types is by the Andrey Nkosi I don't expect you to read this as part of the module but there, there is, it is available, both in the university library, although I don't know if it's available online. You could have a look, it's a very very good book cover researchers who have gone through tremendous work in data visualization. And the key thing that they're thinking out there in terms of the concept of referential data characteristic data is referential data refers to the context in which measurements are made, the sort of setting in which measurements are made. And though this depends on the purpose that you're using data for. And in that, that reference can be either in terms of time or space or population where population yes it can involve people or animals or plants but it can also involve this general object, like, you know, sensor one sensor two, and so forth. And then the characteristic, or a characteristic is a measurement that is made about one of these types of reference, continuous and discrete data. This is something that we covered in Thursday's practical last week. It's the thing that distinguishes the, the green from the blue pills in Tableau numerical data are very often, continuous ordinal data and categorical data is always discrete. But of course, integer numerical data is also free. The terms dimension, and measure a dimension. You might think of that as being the same as the variable. And, in fact, some people do use it that way. But other people would make the distinction that a dimension is when you're talking about variables that are really closely related to each other. So for example, the. x x&y coordinates of a position, but not height and weight, height and weight are two separate very separate types of variables. Measure is equivalent with an observation. But of course, it might relate to something that is numerical but also something that is categorical so for example if I want to talk about. Somebody's gender. That's a categorical variable is categorical measure. And then, of course, there are certain companies who then start competing matters a little bit more because they are using the tableau using the dimensions and measures in slightly different ways. There's no answer to this you just have to deal with it. However, what we can do is to simplify things. That is, in terms of really basic data types, there are only three and that is nominal ordinal and imperative is terminology comes from a very influential paper that was published. 24 years ago by john mccain he has to account john mccain is actually one of the most senior people in

Tableau these days. And let me explain them this way, first of all, dealing with, with nominal, or actually I prefer the term categorical here. But you will find all three of these categorical nominal and qualitative. The key thing about this type of data is each value is only identical to or not identical to other body

19:55

language males, male or female, going on to the numerical data that we have here

20:03

with some of the other. The key thing about numerical data what defines data as being American is that you can do arithmetic on it. For example, you could calculate the difference between two numerical values, you could, you can calculate the mean of the vertical batteries, whereas an ordinal data. You can't do arithmetic on them it's not sensible for example to calculate the mean. But it does follow less than and greater than relationships. So an example here would be the classification of a degree so you're probably all of you are studying a master's degree at the moment. Whether you end up getting a, an ordinary BB in your master's degree or a distinction in your master's degree. The distinction means that you've got a greater number of marks than the ordinary. But you can't calculate the mean of the math from, from those two degrees. It is also worth noting over here that for the purposes of this module in order to keep things simple, date and time data that is a type of numerical data. Of course it is rather special because it contains all sorts of different levels of detail and complexity, but at the end of the day you can do arithmetic on dates, times. And the same goes in terms of spatial data in terms of what it is, again, it's a very special type of data that you can do arithmetic on it if you got a series of of locations, for example, you could calculate the central average of the location. So this hopefully will send the simplify things a little bit from the very large number of different types of terms around the data, we looked at a few slides ago. And I simplify things. Now I'm going to make things a bit more complex. You've got an example dataset. And it's take a minute or two to look at this data. Column variables in this data table and ask yourself, which data type. Is it is it categorical ordinal, or numerical. Give yourself a minute to do that.

22:59

And then I'll ask some questions about that. This variable height. What type is it out of this. These three is five for C for categorical and an overall or n for numerical depending on what you think I did. There was a lot of anxiety variable you're absolutely right that.

Yeah, ethnicity, again the same thing, obviously you know or. Again, we're saying off the stage absolutely right. Gender by the CEO or an N. I'm saying lots of CS

24:36

fibers. Yes, I think

24:43

what you should notice is that although the name of the variable is gender, the values are into development. So if you read this data into pandas. For example, this column is variable will be given the data type of integer, because the values are all integers. So that illustrates the difference between a datatype as defined by the values of a, the values of data, versus the datatype as defined by the semantics of data. So when you read thought about the name of the variable gender you thought well that of course refers to male, female. Therefore, these values here. One probably refers to one probably being male and another one probably means female. You don't know which is which because you don't have to documentation. But once you were doing there is making an inference about the semantics of data in order to come up with a data type and of course you would be correct.

26:01

What about Heidi. Again five Percy and L or M. Welcome, I was coming up with the CSE. So, again, the values here are all numbered. You realize correctly that an ID and different identifier, you can't average it meaning, calculate the average of an ID.

26:42

But actually, there's nothing here to say that an ID, for example this ID here in romantic is greater than this ID here, so actually the ID is an example of categorical data. It's just that it's stored as a number, partly so it could be auto generated, one by one, but also that it will then make comparisons, very very quick. However, you need to watch out when you're using numbers for IDs, in some software because often like pandas because of some of the underlying software, it depends on cannot handle missing data for integers. So it would actually, if you had a missing ID in here, pandas would absolutely convert this variable into floating point numbers, not integers is the limitation for pandas. Data birth, what type is a gun fight. No, no, no. Mari is mostly hands. A few hours, if you go back if I

go back to the previous, look, go back to the previous slide. Note that what I'm saying is that, for the purposes of this module. All daytime data, it is a type of numerical data is a special type of numerical data but it is a type of numerical data and daytime. If I had asked you to classify in terms of one of the three dives the answer would be numerical. What about month of last visit. I couldn't say for him that. This is a nice example there's lots of you know you're coming up with every, every time under the sun. The LDC. Now, If you again if he was to read this into something like pandas it would say oh well that's tax that's a categorical variable. But of course, semantically june july Bay you know these are months. You can do arithmetic on love. They are part of the day. They are the central part of the day, semantically. These are numerical data. That is the distinction there between was defined by the values or the data diabetes as defined by the values or what the data type is as defined by the member. This is why it's very, very important to read the manual rtfm is an acronym that stands for read the manual, I'll leave you to fill in the missing letter with the word. Because unless you have access to documentation that about data, you can often find it impossible to correctly interpret the data, we go back to that previous table. We can make an inference that is one and two referred to male referred to male or female, but unless we have documentation, we don't know with with as. When you go when you go into the practicals in two weeks time in fact you're very welcome to look at it now if you look at it before if you were to go to data Ville north. This is our data repository for datasets from the north of England, and one of the datasets there off street parking fines I'll show you that.

31:48

Have you gotten that web page you would see this, you would go to this site which is giving you information on parking buying but people are paid leads and you'll see this day that goes by and through the present day, mostly with data in a CSV format that we've got one file here that's an Excel spreadsheet, surprisingly, going back all the way to play that thing.

32:22

So the documentation is very important, but often you'll find documentation is not complete, and sometimes you'll have to do some additional research to explain things that are not explained in that documentation. Okay, let's go on to something about data sources and what I want to talk about here are like a distinction between different types of data source, starting with publicly available data. And with There are two main types open data, and other publicly available data. So first of all open data. This is data. Generally speaking, it will be aggregated but not always, but steps will be made to the

privacy, where necessary and other data generally is provided with no restriction as to how you can use it, which means you are free to make money from it. You want to use open data to develop applications that are useful to people that you then sell to those people. And in fact, that's one of the motivations behind government, making data freely available. The UK Open Data license is an example there. And one example of open data, which is available in the UK is the, the distance that children travel to their school. So here we've got three schools in the general area of Leeds, and this. If you live within this distance of the school. This is where they put the files, then you would generally expect to automatically get entrance to that school. If you live further away from that from the school, then it is much less likely that you will that your children will be able to go to that school and that of course, then influences where people choose to buy housing aggregated data because it's not saying anything about the individual children, going to the school, or an example where each and every record is provided. So it's not aggregated actually, is that carparking bind data that I showed you on data mineral bills now. There's a unique identifier here every record. But there's no information there on that fine was given. So that's our privacy. There's lots of other ultimately available data publicly available data, because we're the UK. That means you're free to look at it, but you'll not necessarily free to make use of it in a commercial manner as studying as a standard there's no problem there. And so what you need to do is to check the licensing conditions before you try to commercially exploit data, otherwise you may not be thought. And it is your responsibility to protect that other types of data so you can think of all these other types of data as being data where access to them is restricted when sharing data. This is where a sort of collective effort is made to gather the data, and then under certain conditions that data is shared with the parties to use for specific purposes. For example, fine grained data about environmental measurements. By be shared with an organization, under a license for a specific purpose such as for helping to teach the geography students in a university. A university may get that license for free. Or it may have to pay for that license. If they break the terms of that license and start using it for a different purpose, then they can end up in court. The same goes for a lot of health data in the UK. So, some of the research I use makes use of very, very large like 20 million record data sets of treatments and so forth that individual patients have had in Boston when it is all anonymized

37:21

that data, you have to pay for evil anonymized for research. And, one has to keep that under very strict security rules. Even though the data is anonymized, and with for example the NHS data. It can only be kept on systems that are accreditation known as nhfb of what this is good for is that it promoting reuse of data that is actually expensive to obtain. So these hospital episode statistics are used for a vast amount of research, benefits,

public health, and diseases and so forth. In the UK, some of the data that I'm using of this type, actually, at this very moment, relates to our patients are being treated in the UK. During the current pandemic in order to provide the one that helps doctors understand and how to which treatments are most appropriate for which conditions. Very often this data is record level, it is in the case of these examples of almost always it will be a lot of miles so you can't directly identify individual people and again. You'll learn about some of the issues there in the ethics lesson. The last type of data close data. Generally speaking, this is data which is identifiable or commercially confidential. So in terms of your health record your health record where you are identified, of course that is being shared it is shared between organizations so for example your GP works for a different organization than the lead hospital. Legally, they are separate organizations. Of course you need to be able to identify a person, treating them for a disease to come to the right treatment. But in order to do that, there has to be all kinds of ethical and danger agreements that are signed it is not trivial to do that. The same goes for commercial information. This is mostly confidential and this is why you encounter in terms of agreements. You know, they do, if you want to use a certain on line private a certain app, you have to agree to let the developers of that. Share your data with other organizations. Pay they're doing that to force you to agree to something that will make them money, or it might sometimes it's the case that actually certain aspects of the management of their app is done by a third audio

40:38

finish off this lecture by looking at ways of preserving privacy. And if we go back to the children going to school to work out the data here for the schools and the average distance away from schools about the distance away from schools and pupils live in order to be pretty one guaranteed access to that school. These values there are actually calculated from the fine grained data about each pupil, which means taking the postcode which identifies who and then generally to within a few 10s of meters, where somebody lives. And the post ends the position of the entrance of the schools from that you can calculate the straight line distance but each people in the school. And then you can average it, or this guy's calculate the maximum. And what you find is is open data is just the maximum. So, if we had the full listing here of the school the name of the pupil, and then from that school that will be a part of close data. If you took away the name of the pupil then perhaps you would make that available as shared data. And when you want to aggregate them together like we have in this example here. We can now make it openly available because it's not breaking anybody's purchasing. Otherwise, that you can protect privacy when you're creating data sources. One is to reduce the precision, though, in terms of age, rather than providing information about people's date of birth, you provide information about their age, the year of birth, or more often in terms of the, The age band.

An example that comes from some of the health data that publicly available. This is property information in the UK about the number of patients registered each GP is unique code for the practice here and as you can use that code to look up the name and location of that code. And then you will find. For example, for males, a whole series of variables here in this case running in a new band so this is basically less than a zero, a one year, so forth all the way up to 94, and then grouping together. All of the patients who are age 95, or greater. This tells you the number of patients, and therefore the age profile of patients GP. And the reason of course that this last column is grouping 95 plus, is that there are not so many people have that kind of a in the UK. And in the case of yeah this is real data. Are these GP practices have no male patients in that age band, whereas this one has nine male bass in a band. The numbers in here get very small which means that it would become possible to link this data to other data sources and have an absolutely a high debt defy the individual people who are a little brave. So that's how you begin to break privacy. The minute I think in this way, grouping them in this way. and generally reducing the precision, protect privacy. You can group data spatially so in the UK we every place every household Every business has a postcode. The university postcode is this one lm two nine JT. And this postcode is made up of four different items. These first two are what's known as the house code. And these and these two are known as the in code. So,

45:28

if you just use the, the, the area. ls stands for leads that of course is giving you a very approximate spatial location. ls two is giving you a finer grained location, right down to. ls two nine JT will actually that will move into one location the University of Leeds. And so if you remember, in the tabular practicals that we were doing last week in the GP surgeries dataset. Each GP surgery, there was both the full postcode of the surgery, but also a separate column, giving you the aos code.

46:14

And it was the alcove. That was then linked to the different counties in. In another data table. And so finally, other things to consider. In terms of understanding your data. Think about the data provider, because if you're going to do some data as data governance you want your data to be a good quality. And

46:51

one of the places we've got there and think about the reputation of your data provider you know do they curate their data, they put a lot of work into cleaning their data, or is it

very messy data, which is going to leave you with, with a lot of work done. Even if your provider has a good reputation there still may be lots of issues with data quality, and some of those are unavoidable. And then secondly, think about making sure that you use data in a lawful manner. Some of that relates to licensing and a lot of that, particularly when you're dealing with data about people relates to ethics and information governance, which is something that you will spend two weeks on later on. That's where we get to for the day. I'm just going to go back to the nerva. On the learning results is, first of all, if you want to look at the results of the survey you sent in to anonymize it just has the dump of your responses. Then you'll see the results down here. And on Thursday on the practicals. Hopefully, you've already dealt with the Getting Started I'll spend a little bit of a bit of time on the calculation. Part of that tutorial, I'll go over that in the beginning, because we didn't have time to do that last week, but then focus on these challenges and in particular, 38 of them but the most important ones for this module, are challenges for with the focusing on aspects of data quality. So it will

48:59

be that what I've done with these challenges is show you what you need to produce, but not how to produce them in the tutorial I showed you exactly what to do.

49:12

But in the challenges, you've got to work it out yourself. But do use Google do put postings on teams get started on it now. Help each other there it's a team effort. And in doing so, you will

49:25

a learn about some of the issues of real data, and this is using the parking fine data.

49:39

So you learn about the issues the real data, you'll learn about how to think about tackling those issues with data science software, and you'll also learn will improve your knowledge on a particular piece of software like tableau. Of course, you might also choose to do these challenges in a different way, which is using R or matplotlib or pandas or whatever. I'll leave things that are all hanging around for a couple of minutes in case anybody has any questions you want to post on the chat side and I will see you all next week. Sorry, see

you all on Thursday.