Using Consumer Data in Research 3. What skills do we need?

Hi everyone, welcome to this third video on using consumer data and research. I'm Dr. Nick Berman. And I'm going to talk a little bit today about what skills do we need to make best use of consumer data and research? So there, we've covered a little bit on what is consumer data already in the first video? And what sort of examples of research can we do with consumer data. And the second one, so we're moving on, specifically on to the skills in this video.

So remembering that consumer data are data that arise out of everyday transactions for goods and services. When we actually do some of this analysis, there's a kind of couple or two areas that we can focus on. So we can either kind of focus on substantive findings. So this is interesting research questions on this in a specific area. So how do people move? You know, how do bankruptcies impact people, that kind of thing? Or they can be methodological advances. So how do we come up with new processes to to process and analyse some of these very big datasets, some of these very raw data, for example, some of the footfalls stuff, how do we take each individual sensor reading and aggregate it into a way to create a product that we can actually do something useful with either of these is fine, you know, and we've seen lots of examples of where people focusing on one or the other can make an excellent PhD or an excellent research project. If you're trying to do both, that is very, very hard work, you know, so doing both methodological advances and substantive findings is very hard work. Ideally, you want a bit of both. So ideally, you'd be working in a team, and maybe one of you will be focusing on the methodological advances, and one of you will be focusing on the substantive findings. And then you can work together, gain benefits from each other's experience, and both have a great, great project at the end.

So there's lots of different methods that we might need to master to work with consumer data or to know a bit about work with consumer data. So some of these are kind of big data or data science methods. Some of them are GIS, spatial data, methods, some of them have to do with data modelling. And this can be for, you know, either big data or small data, and some of these geographic data science related methods. And some relate to how we work with safeguarded and secure data and aspects related to GDPR. So we'll pick up a little bit on each of those as we go through.

One other really important thing to think about is data triangulation, as well. So when we're working with consumer data, do the findings from your consumer data match your traditional data sources. And so some of the kinds of traditional data analysis tools are really important as well to allow you to do this triangulation. And this is really important, because it will allow you to say are the consumer data representative? Are they a useful source for the topic or question that you're looking at? And it's always worth thinking about who is missing from the data you're working with? You know, are there any groups that are not represented in that data? So the housing data is a great example. So if you remember, we were having we had some sales data and rental data and Zoopla, energy performance certificates and

Lincolnshire registers. And so we've got a kind of database of every adult in the UK and where they've moved. But that's only from 2014 onwards, if someone's not moved since 2014, they won't appear in that list at all, potentially. And I think thinking about it, I'm sure there lots of people, you know, not moved in the last 12 years. And so they're not, not in that data at all, their move. They've not, they're not there, because they've not moved. They might be in the linked registers, depending on how they were in the electoral roll, but they may not be. So they may be missing completely from your analysis. So think about who's missing. And, you know, is that important to the analysis you're doing? And equally also, what's the bias as well, you know, are there specific groups who, even if they are renting won't appear in that data set. And this is where the domain knowledge comes in as well, which we mentioned a bit about last time. So relate your consumer data findings to some traditional data sources as well? Do you get the same results?

And GIS is a key aspect when working with consumer data? You I would be tempted to say this anyway, because I'm a geographer by training. So I'm a big GIS fan. But I think it's really true with consumer data. You know, look, the location element of consumer data is often really important is a frequent way of linking consumer data with other data sets. You know, this may be very explicit by you know, it's an address, or it might be it is a certain dataset, data point, presence or absence in certain locations? Or it might be through a longer role of geo demographic. So do your individuals match particular geo demographic group? And then where might you find those? Yeah. And how important this is varies a lot, depending on your domain, what you're specifically looking at, you know, some areas spatial, is key, you know, migration is a great example, you know, where people are moving to and from is key to understanding migration. Footfall is also quite important. So where is the footfall? How does it vary the space, retail, probably slightly less so where the shops aren't as important whether users are the shops are the customers are as important as the whole aspect of online retail, where location is probably less important for access, it's quite important as well. There's other areas where, you know, location is not key, but still quite important. So for example, hospital admissions, you know, which hospital you go to, depends on where you live. And what treatment you get, probably depends which hospital you go to. So it's not a direct impact, but it's related, and access as well. So how do you actually get to the hospital? You know, do you have the ability to be driven there do you have to go on public transport, how good is transport and so on.

And there's other areas where, you know, location is probably not as important by any means. So for example, energy use, you know, how much energy you use depends a lot on what sort of house you live in, as well as what you what you do, what type of house you live in, might be influenced by where you live. And your income or deprivation might also be influenced by where you live. But they're kind of one or two steps removed from the actual data itself. And there's lots of training out there on GIS and spatial data. You know, there's lots of free software stuff online, so you can use QGIS, or R. And there's lots of stuff out there. So there's plenty of opportunities to find out more about what GIS is. And sometimes these GIS skills need to be combined with the big data skills, databases, and other keywords, or geographic data science as well.

On the note of big data, and data science, sometimes these are quite closely related, sometimes less. So these are really useful tools for working with big data. Yeah, anything more than a million rows was our fairly arbitrary definition. And then there's some very, there's a lot of acronyms surrounding this and

some tools. So high performance computing is a key area. MapReduce is a kind of approach to doing data analysis. And there's various cloud processing resources available as well. This is not a big data, or data science presentation. So I'm not gonna go into those. But there's lots of resources available if you need to use those. Two of the key things are use of databases. And that's really, really useful when you're working with large datasets, even if it's not as big as big data, and scripting or coding. So you know, this is becoming much more popular in GIS and across the board in other areas. So r in Python are the two big ones there. There are others available. And we often kind of wrap this up as data science, this area. So if you start thinking about data science, this is the kind of thing we're talking about.

And the other aspect that, you know, might easily be missed from something like this, but it's quite important is modelling so data modelling and statistical modelling. So, you know, there's kind of two, two elements to this. So firstly, kind of more classical statistics to how do we make sense of this large dataset. So we can start off with some descriptive statistics. And we might start to look at kind of things like correlation. And we might do some more space, more advanced stuff, so like, principal component analysis, and we might go down the spatial route as well. So we can look at spatial autocorrelation, spatial clustering, spatial regression, all sorts of stuff there. Geographically weighted regression was another one as well. So how we kind of apply the stats varies very much depending on what the data are and what you're interested in. I'd also say that modelling is quite important as well. So you know, particularly with some of the the larger data sets, they often might have missing data. And modelling can be used to fill in some of the these gaps in the raw data. How effective it is really depends on what data you've got, how good a model it is, and what you're trying to predict, as well. So Van Dijk paper is a very good overview of how the modelling in a number of different ways to fill in some of the gaps in the linked consumer register.

And we also, you know, you've probably come across this already geographic data science. And this is a kind of mix of GIS and data science, and a little bit of modelling thrown in as well. And it's combining all of these existing methods, but also focusing on how we go about creating new methods and new tools and new new analysis. There's loads more out on geographic data science, key bit, I would say, is scripting. So you know, this is ever present in geographic data science. So I'd really recommend everybody at least have a go at learning R or Python. And in most of the kind of novel applications working with consumer data, you often combine one or more of the above. So the the kind of work and effort needed is not to be underestimated. And that's why we made this distinction about methodological advances and substantive research at the beginning, you know, it's easier to focus on one of those rather than try and do both well. And equally, there are lots of new methods being developed all the time. So do keep your eyes open and see what the literature contains, see what's been presented to Conferences for new methods that might be really, really useful for the work you're doing.

The other aspect is often seen as a kind of slightly non technical side, but it's still really important in terms of data governance, for working with consumer data. The data we're working with is often individual level data and could contain personal and or confidential information. You know, loyalty card data is a great example of this, you know, how do you get the individual level records without compromising people's security? The how it applies is really variable depending on what data you're

working with. And so it's really variable. But there's a few key concepts that are useful for this. And GDPR provides a kind of structure for this. But ultimately, it's remembering that about the individuals who provided that data originally. We need to manage this data responsibly. And that's a specific request in the GDPR rules is to manage and secure this data responsibly. And we also have the concept of data minimization. So this is where we think about, okay, what what data do we actually need for our analysis, what's the minimum we need to do, and then we don't have use any of the extra data, we just we can get rid of the data that we don't need. Not storing data is the best way of keeping it secure. So minimising exactly what you need is a great technique. And training is key for this, you know, there's the safe research training that UK Data Service offer. And this is a kind of key underpinning to how we work with this data. And that that train is training is compulsory for most work with individual level data.

As well as the GDPR. And the information governance it's worth thinking about the infrastructure as well. So if you're working with individual level data, we often might use a secure research infrastructure, a trusted research environment. And this is a way of managing the secure data. So you have different levels of secure and safeguarded data, depending exactly what the data is. And there's a kind of similar model that most of the groups use. So safeguarded data is when there's a there's a process to apply for the data. So you've got to say what you're going to do with it, and it has to be reviewed, and so on. And then usually you get the data, you have it for a set period of time, you can do your analysis, sometimes your outputs might need to be, your publication might need to be checked, sometimes they might not. The alternative approach is you get secure data, and you have to use that in a secure environment. So this can be a physical lab. So it can be a computer lab where you physically have to travel to and use the data. And there's no external Internet access in that lab. So you just do your work in there. More common now is a virtual secure lab. So you can use your own computer but you connect through a virtual environment to access the data and do your analysis within this secure infrastructure in this environment. And then to actually take any results out or to extract anything from the map. Yeah, there's an output in process or a release process where someone else checks the data to make sure you're not breaching the confidentiality or releasing any identifiable data. There's lots more information on the the specific processes depending on where you get your data from. The key bit is it is it takes time, you know, both the application process upfront and the output checks at the end, you need to build this into your plans, you can't just click your fingers and immediately get the data there's a process to go through. So do bear that in mind.

The other really, really useful skill, which is really important is the aspect of reproducible research. So this is the concept that good science is reproducible. So if someone does a bit of analysis, and publishes a paper on it, the idea is that someone else with a similar level of knowledge could come along, pick up the paper, and then replicate the analysis. So a lot of journals now ask people to post data and code with your submission. And if you're doing using scripting to do your analysis, this is a really helpful because you can include the script. And this brings into the concept of version control as well, which is not something we've mentioned here. But it's not only that, it's kind of a really useful tool for working with reproducible research. At first glance, it might, you might think that, you know, your journal submission requests to submit the data and the code might directly conflict with submitting secure data to a journal. And you know, that is true, you secure data, you are only allowed to use it, you can't submit that to a journal. But there are a number of processes for managing this. So you can still

document the methods and the code using much the same process. And that can be shared. And then you can reference exactly what dataset you're using. So the usual approach for this would be with a data DOI to say, Okay, I'm using this particular dataset, and possibly even a specific version of the data as well. And depending exactly what you're working with, it might be appropriate to provide some sample data. So the person reproducing research can take the sample data, run the code, and it can give an output. Whether that's appropriate or not, it depends on the situation depends on what data you're using, there's no right or wrong. It's not black and white, but it depends on what you're doing. And it's always really good to publish your output data, or your your analysis ready data products, if you're, if you're developing a new method, and you create a nice dataset as a result of it, if you can publish that as an output, because that can be really, really useful for future researchers. So do bear that in mind when you're doing work.

So we've covered a few of the kind of key skills, there's quite quite a lot there. But we can kind of group them together. You might need some big data or some data science skills, you probably need some GIS geographical information science skills. And you may also need some data modelling skills, and possibly some geographic data science skills. And or a combination of all of those four, because the combination is where you can make some real progress and generate some really interesting results by combining these in a novel way. It's also really important to be aware of some of the secure data requirements and the information governance requirements and GDPR. To ensure that you do this in a way that is secure and meets the expectations of the GDPR and the safe research training. And the trust that we're doing the correct thing to these data is key to the whole system. Because if our use of the data is abused, people will likely not provide data for future research. So it's in our own best interest to follow the codes of practice and do the right thing with data to ensure we still have access to data in the future.

So thanks very much for listening to all of these videos, it's been great to share some of the bits of knowledge of consumer data with you. So we've talked a bit about what consumer data is. And we've talked a bit about what can we do with consumer data with quite a range of examples. And then we talked a bit about skills, working with consumer data as well. Please do check out the references are all at the end of each of the presentations and the reading lists on the NCRM portal. And I wish you the best of luck. Thank you very much