Using Consumer Data in Research 2. What can we do with Consumer data

Welcome to this second video on using consumer data and research. And in this video, we're going to think a little bit about what can we do with consumer data. So a few examples of what what research has been done with it and why it's such a useful data resource. So consumer data, all comes from everyday transactions for goods and services. And it's a very much an umbrella term. So it covers lots of different sorts of data. And we're going to do a few examples. But there are many, many more out there as well. And fundamentally, it's all about the digital traces that we leave in our everyday lives and some of the things that we can analyse with that.

So data are mostly data about people that consumers and it's often sometimes restricted, because it contains personal information. So we need to bear that in mind when we're working with this data, both in terms of how we actually get hold of the data, which we'll talk a bit about in the third video. And then what analysis can we deal with it once we've got it, which was what we're going to talk a bit about now. It's always secondary data as well. So it's not collected with research in mind. So this limits what we can do with it. And we need to remember that when we're analysing it. And it's also worth mentioning administrative data here as well, which is sometimes termed as consumer data, sometimes considered separately, depending on how you define it. So administrative data, it's kind of similar individual level transaction data, but that collected by the government, so that, you know, at one level, this could be the census. But equally, this is things like HMRC, tax records, DVLA, records, and school census, all that kind of stuff. So very government focused, usually collected with different aims and sources in mind. But sometimes the same approaches and the same tools are really helpful. So we'll mainly focus on consumer data. But some of the tools are relevant for administrative data as well.

One of the first examples we're going to look at is travel data. So this is transaction based travel data. So this is particularly looking at data from bike hire schemes. So when you go and hire a bike, you can put your card in and the bike docking station released the bike, and then you can go off on your journey. And because the way the data is structured, often you need quite a large amount analysis to generate some useful results. So when you get your bike out your bike stand, that's logged, so you know, when you collect the bike, and we stand you left, and then you get some data, each of the bikes have a code. And when you return it to another docking station, it knows where the docking station is, and when you return it. So we've got the time of departure and time of arrival. But nothing else, there's nothing to do with the journey or the route or that sort of thing. We can work out some of these, so we know where it started and finished and the date and time. So you can take a journey, you know when the journey took place where it's from where it's to, and you can take a kind of educated guess about the route, we don't really know much about the route. So you know, to kind of typical output from this would be something like the graph here. So we've got a peak in the morning, Russia, lots people are collecting and returning the bikes. And then I kind of more extended peak for the evening. So evening rush hour plus, you know, early evening travel as well. And this is a great example of where there's

been quite a lot of heavy lifting to get the data into this state. And there's a great example. Todd et al. 2021, talking about the process and how they actually created some output data.

And we can think about footfall data as well. This is another another great example. So this is very interested in how many people there are at one specific point in time. Typically, it's footfall in shopping centres, so where, when the busy times in shopping centres, both times a day, days of week, and times a year as well. And this is all collected by sensors. So these are a few examples of how sensors might be arranged. So we're interested in people kind of going along the footpath, how many there are, it might be we've got a shop or some sort and there's a sensor in there. It might be there's a sensor in like a kiosk on the street somewhere. Or it might be that we've got a sensor, but we have to interpret the data in certain way because there's no outside seating there or a bus stop. So we'll have people passing and people in in a static position outside the sensor. So there's there's quite a bit of interpretation that we need to do this with.

So we had got used to seeing a number of these graphs. So this is from 2017-2018. And we got a nice peak for the Christmas shoppers in December. And then they kind of fall post Christmas where nobody's got any money and nobody does any spending. Yeah, and this is very, very kind of typical output from this footfall graph. And then when Covid hit in February 2020, march 2020, we saw lots of graphs like this, looking at how footfall fell down when we have the national lockdowns and how it's kind of gradually recovered. So here, you know, we're looking at Greater London, we have the kind of precautionary phase, the beginning of March, and then the actual formal lockdown kind of middle towards the end of March. And you can see how the numbers just fell off the cliff here. And the activity was kind of 20-25% of the the normal level. And it gradually increased over time.

And we might have seen a few more detailed graphs. So splitting out that sort of fall by location. So it might be we're looking at, you know, park or residential. So, you know, Park usage, you know, loads beginning but then came up, remember, of course, the time of year,, so we're getting into June, July. And then various stats about, you know, driving and supermarkets, which kind of went down a bit. And some workplaces and city centres and retail, which fell much more significantly, and have been generally slower to recover, depending on what kind of specific area you're looking at.

And one really interesting thing about this is how we actually collect the footfall data. And the technology behind it is constantly evolving. So originally, at the beginning, we were using WiFi sensors and MAC addresses to account and identify phones and one of the tools we use to remove some of those ws, if the MAC address is staying the same, these people that are always close by the sensor. So these could be people waiting for the bus or people sat at a restaurant table, where if people walking past it would look very different in the dataset. But you know, there was an update with IOS, so the iPhone operating system, and this change to dynamic MAC addresses. So people's MAC addresses change. So it's very hard to use the same process. Android phones weren't affected by this so that that data process still works. But Apple phones wernt so we couldnt really use the data. So with that processing, all of the iPhone data is now not very useful. But iPhone data and Android phones, I would say are not kind of equally representative. Different groups typically buy one or the other depending on a whole range of things. So we need to bear that in mind when you're doing that analysis. Now the technology has moved on. So now we've got a series of apps that collect GPS data. And these are

aggregated together. So they've got location in a very different way. But some of the output is broadly comparable. And we can use it in similar ways. But when we're using data or different sources, we need to bear in mind these differences.

Another great example is housing, looking at housing and migration and where people move to. So Zoopla data is a great resource for this. So it will tell us quite a lot about how sales house rentals and some of the work done at UCL with linked consumer registers combines this information with people's names and addresses. So you can see where people have moved to and from so what, you know, are they moving to bigger houses? Are they didn't smaller houses? How far are they moving? Particularly interesting for moving around in London? You know, are people moving out of London to get a bigger house? And how far out are they going? And some of the work combining this with the electoral roll data and energy performance certificates have allowed all of these kinds of datasets to be tied together. And that allows us to look at housing and migration as a whole. So kind of in the round, if you like and these are some interesting comparisons of the Zoopla data. So on the y axis and then the moves in the private rented sector from census data. So on the x axis and looking at different years, you know the correlation there is pretty good. So the it'd be fair to the data in the Zoopla your data is representative of our whole population data in the census.

We also look at hospital admissions rates. So hospital episode statistics are a great data set looking at different disease prevalences, who's been admitted to hospital. And then we've done CRC have done some processing, looking at grouping this by ethnic groups, and a whole range of other characteristics. So ethnicity is something that's not captured particularly well in a hospital episode statistics, you know, but there's some ethnicity data there. But you know, good, nearly 10% is missing from the hospital episode statistics. And that missing percentages, probably not equally spread across the different ethnic groups. Yeah, it's hard to say exactly what's missing, because we don't know what's missing. But overall, it isn't representative by any means. So we, if we combine that with the ethnicity estimator tool, this allows us to model ethnicity. And we can reduce that down to, you know, less than half percent of missing, which would be missing data. And so this is one of these datasets, where there's quite a lot of analysis to get it into a usable form. And, you know, this data set is now available on the CDRC site, for for application and use and further analysis. And it's a great input into a lot of large scale health data analysis, and this key to a lot of the work that's done.

Another nice example is Smart Energy meter readings. So with the rollout of smart metres, you can get usage information every half an hour. And one approach that taking this data is we can analyse it to to kind of get typical user profiles to help understand people's energy uses and how it varies. And this can be used to target energy initiatives, energy efficiency, provide insights into the temporal nature of energy usage. Yeah, previously, we could do this at a national scale. And so there's some useful information there. But this allows us to pick out different groups look at it at a regional scale, as well as as well as an individual scale. And it's a it's a great input, the temporal element is really helpful. You know, it's an interesting analysis done by a colleague looking at, you know, how feasible would it be to use wind farms to generate power for every UK Home? And that's a data story there. And so that talks about, you know, how much energy can you generate? But also, when is it used? And then how do you store energy, and there's some of this energy storage challenges that the sector is facing. And all of

this is even more important with a very recent increase in gas prices and the energy cap as well. So having this extra data allows us to do some much more interesting and detailed analyses.

Loyalty card data is also really, really helpful as well. So this is some work using a high street retailer. So one chain that that we've got some data for. And we'll come to some of the benefits in a minute, but we have to remember the data is not representative. You know, this is an analysis looking at this particular retailer and splitting out by geo demographic group, you know, and some groups are really underrepresented in the census, in the loyalty card data. So you know, rural tenants here, they're much less likely to have one of these loyalty, cards and the kind of average person and equally some are overrepresented. So in semi detached suburbia, they're much more likely to have one of these ability cards than average. So we got to bear this in mind when working with this data. So not to say we can't do anything with this data. But remember, who we're actually targeting here.

And equally it's not geographically representative, as well. So you know, a map of the UK looking at what proportion of adults are in this old scheme, you can see it varies quite a lot, you know, in the inner kind of urban centre. And then, you know, suburbia can be quite high in anything 20 to 30%. But in some areas, it can be very, very low, you know, two and a half percent or less. So, it's not representative of the population. It's not geographically representative, either.

But, having said all of that, there's still lots of lots of really interesting and useful analysis that you can do with this data. So, you know, there's a great study by some colleagues at Imperial looking at cancer patients, so patients who are diagnosed with ovarian cancer, but also what were their spending habits before they were diagnosed. And they did a very small pilot study looking at 11 people and before the these patients were diagnosed with ovarian cancer, there was a marked change an increase in the over the counter medicine they bought for pain and indigestion compared to the control group who didn't experience this change. So there's potential to use loyalty card data to identify particular conditions and, you know, possibly even target early screening to help with this. They're doing a longer study with a much larger group now to see if it can be replicated. And this is kind of in process. So we'll see what happens with that. But it shows some of the kind of potential that that can be done.

County court judgments are also another really interesting data set. So these are a key measure of financial health, including bankruptcies. And a couple of people have done some work on this, looking at links between health and financial well being so you know, of people subject to personal bankruptcies, who's impacted the most are the financial or demographic patterns in this. And there's lots of extra work that could be done with this. So potentially developing some economic health indicators, looking at different research aspects and the impact of politics on debt on indebtedness, and there's a financial stress tracker there, the registry trust have been developing, looking at, you know, how, how different areas and different groups suffering, more financial stress, depending on time of year, and the various aspects that are going on, you know, and relate to the, you know, the gas prices and cost of living and all the rest of it is even more important. So there's lots of things that can be done that and there's lots of potential work out there as well.

So, you know, there are some limitations as well, you know, there's lots of benefits, but there's limitations as well, you know, some of the raw consumer data can be quite hard to analyse, so

particularly, for example, footfall data, bike data, you know, the, the actual raw data needs quite a lot of work to generate some useful results. And sometimes these datasets have already been, had some work done on them. So there's these things called analysis ready data. So this is data that's been pre processed already. And it's ready for you to do some some work with. Also, we need to think a little bit about the representativeness of the data and represent the results fairly, that we're doing, you know, what have we found, but what are the limitations as well as the advantages? And there's a there's a very good presentation by Danny Arribas-Bel, r links at the bottom there. And this was a keynote presentation for the Spatial Data Science Conference and explaining and looking into some of these issues about what we can do with consumer data. Also, what are some of the limitations as well? How does it fit into the kind of wider spatial data science.So, we we've looked at various different datasets here we've got a few different examples and some of the other potential sources as well as theres lots and lots of other potential sources out there.

So for the final video in the series, we're going to have a look at some of the what skills do we need to work with consumer data? What are the kind of key skills and techniques and how can you go about learning these if you need to? Thank you very much. I'll see you in the final video.