**Storytelling with R – Olga Pierce**

All right. And next up we have Olga Pierce. Hi everybody. I know everyone always says this at data science conferences bigger a good-looking crowd. I am on the data team at ProPublica, which is a journalism non-profit based in New York, were funded by philanthropy. And our mission is to do journalism and the public interest. What I'm going to talk about a little bit today are techniques to get people to consume your work joyfully and not begrudgingly. Kind of things we've learned in the journalism world that might help you communicate your findings as well. So Newton's inferred law of people reading stuff is most people don't want to read or consume things that feel like work. But you can make decisions that make even complex topics and findings less work. Think about sort of where the friction is going to come in as someone is consuming what you have to say. So we step back for a minute and think about actual storytelling. What makes a good story? So here's some things journalists think about when they're working on a narrative. A Story has characters. People seem to sympathise most with people as characters. But a character can also be an entity like a corporation or a system like a school district. A story has protagonists and antagonists. Sometimes it literally is good guys and bad guys. But obviously sometimes that's a ridiculous oversimplification. So more generally, this is the entities whose experiences will serve to advance the story. A Story has a setting making it richer by grounding it in a place. And a story has a beginning, a middle, and an end. This kind of idea of like here's the setup, but wait, there's more. But then the storey is resolved. It sounds tacky, but it's everywhere because it works. The takeaway from that is that the same information structured differently can tell him much more interesting storey. So how does data fit into this? Here's some common ways that data gets used in the journalism world. One is that using data, we've established a pattern and it's telling us which way the story should go. And then what we're doing is looking for characters and settings that illustrate this pattern. Number 2, is that using data we've identified outliers and we want to give them the attention they deserve. And the investigative journalism world where I work usually don't mean attention and a good way are sort of watch word is FFS, which stands for finding fuckers with science. I'm not sure if I'm allowed to say that actually. But whatever, don't ask, don't tell. So also sometimes data contributes insights into the mechanisms by which something happens. We've learned something new about how something works. But I would say if this is the direction your communication is going, specifics will almost always be more readable than generalisations. On extremely rare occasions, the process isn't how you got the data or what you did to it is interesting, but let me reiterate that, that is rare. So another question to ask is, is your work relevant? Our societies assess pool of inequality, corruption, and dysfunction. Does your work highlight someone or something that needs to change? Is someone or something to blame and who has the power to change the situation. So who is the audience that your work could reach that would cause it to have impact. And how does your work intersect with all of this? Another thing that is pretty simple, but very important is to give some thought to units. So can you base your inputs and outputs in tangible things. So number of people versus infections per respirator, our jobs versus full-time equivalents, dollars versus percent change a lot of the time. Your actual analysis won't generate these sort of nominal values, but you can back them out from the analysis that you've done and it's much easier for people to understand. The same goes for methods. Data analysts often don't realise how much the design of an analysis determines what you'll be able to say later. Journalists know that eventually whatever they find will have to be translated into words. One example that's pretty simple as odds versus risk. Most people understand what risk means. I defy you to describe odds to me in a non compound sense. Many have tried, many have failed. So one strategy that we use is to run things both ways. We run something in a very robust way, the way that his data, as data scientists we want to. And then we also run it in a simple, simpler, and more explainable way. So that way you can report a result, but it's easy to understand with a clear methodology. But you deep inside can sleep at night because you know what you're reporting is true. Helping people make sense of what you found. Give people a framework and context to make sense of it all. Often people don't want dashboard type data without a framework for understanding. If you're just vomiting data onto the Internet, people probably won't interact with it the way that you'd like. That doesn't mean you have to be overbearing and ram something down people's throats or non-transparent, like this is what we found. Don't look behind the curtain is not a good strategy, it turns out. So I'm going to give some examples from our newsroom of actual strategies that we've used to communicate findings. So the first project and I want to talk about is called surgeons scorecard. Basically it's an analysis of surgeon performance doing a pretty straightforward procedures. It looks like this when you land on the page. This is what it looks like when I look up my zip code and Brooklyn, please don't stalk me. So what you can see right there, a lot of new surgeon and some are good, some are bad. Clicky, clicky, clack, clack. If I'm looking for a surgeon for grandma, this is useful, But we're journalists and we want to actually communicate a finding, right? So our strategy for this was to find a hospital where there were surgeons that were both very good overall on our data and surgeons that really did not look so great in our data and then go to that hospital and just find out what was going on. As you can see, there's the surgeon at the bottom is naming is Constantine Toumbis, and he really stood out in our data is not being great. Then when we went to the actual hospital, what we found is that the hospital was basically bankrupt. And the way it works in the hospital world is that surgeons recruit patients and then do their surgeries at that hospital. So this particular surgeon was bringing millions of dollars a year into this hospital. We found out again from talking to people familiar with the situation, that the hospital knew that the surgeon was a problem and was sort of deliberately overlooking the risk to patients for financial reasons. And we kept looking at him and you'll never believe what we found next, I've actually decoloured this photo because it's pretty shocking. So it turns out that while in medical school this surgeon actually slashed another med students throat. And then years later when he was applying for licensure in Florida, he didn't disclose that this had happened. Interestingly, the hospital actually defended him and he's probably doing surgery like on Tuesday. So the second case study I want to highlight is a project we did called the colour of debt. We were looking at the issue of garnishment, Which is basically when a creditor takes an individual to court and gets a judgement that allows them to take up to 25 percent of their after-tax income directly from their paycheck. So the person who's the subject of the judgement just has money taken away. And the question we wanted to answer was, Who are the people who are subject to garnishment? And here we go. So what we did was we got hundreds of thousands of court records, a challenge that we faced was that the race of the person the judgement was against was actually not in the data. But we did have their zip code. And good for us, bad for America. There are places where the racial distribution, the racial distribution of the racial makeup of zip codes is essentially perfectly bimodal, right? These are places where there's extreme segregation. And you can guess pretty accurately the race of someone by looking at their zip code. So when we look at the judgments per 100 people and the median income and then coloured these dots by race. You can see that here and basically Newark, New Jersey, see this pattern. This is St. Louis County. So we see this pattern as well. And you can see that. And if you can see most. So these blue dots are African-American neighbourhoods essentially, and the white dots are essentially white neighbourhoods. And then this is Cook County, which is basically Chicago. And again, you can both see that for the most part, neighbourhoods are either dark blue or nearly light blue. And that the way that they experience wage garnishment appears to be very different even if you look at the median income. So the challenge was, if we go back to these charts, right, you've got a lot of tricky things going on. You've got two rates, you've got a median which people don't always fully understand. So how are we going to actually communicate this? So here's the solution that we came up with. I'm gonna see if I can. We made an interactive media can make bigger that looks like this. We have it for each of the three cities that we highlighted. So from this. Let's get right. You can see a couple of things. One, that wage garnishment is pretty geographically concentrated. And also that there are places where the number of judgments per 100 people is 25% So essentially, there's just piles of money being taken out of these neighbourhoods due to wage garnishment. If you are a household with an income of $20 thousand per year and 25% of that gets taken out to pay creditors, That's catastrophic. So the way the interactive works is you can sort of scroll through these neighbourhoods. Some of you may remember Ferguson. And you get a little a baseball card that gives you the median household income for that zip code. The racial breakdown of population, and the number of judgments per 100 residents. And then you can also, if you hover, you get for Google Street View of the neighbourhood itself as well. But we didn't stop there. We actually sent a reporter out to go to a neighbourhood and say, what does it look like when you have essentially one in three people are experiencing wage garnishment. And we actually found a town where the mare and seven out of eight city council members were actually all having their wages garnished, but they didn't know it because it's something people don't talk about with each other. So I'm going to keep trying to zoom through. Sorry. If I sound like an auctioneer. Say the third project I wanted to talk about briefly was a project called machine bias, where we were looking at scores that sort of guess the risk of recidivism that are used in our legal system in some jurisdictions. So essentially they plug in your answers to some questions and then say this person is likely to recidivi or sort of medium likely are not very likely. But a lot of the questions pretty quickly just become a proxy for race. There are things like, did you get in trouble in school as a child? Do your parents own a home? So we wanted to do was look and see basically what is the output of the algorithms versus the actual eventual outcome. So the way we decided to model it was actually using survival curves. So in this case, it's not people actually dying, but it's people recidivating at some point. And the first thing you can see here is that for violent offences, the model is actually not especially predictive at all. Which is problematic if people's futures are being determined based on the output of an algorithm. The second thing we looked at was what is the difference and false positive rates. So what you can see here, it's a true table or a matrix of confusion. Depending on what discipline you ask. The positive predictive value for black and white defendants is pretty similar. But the false positive rate is much higher for black defendants, which means when we go all minority reports and try to estimate the likelihood that someone will be violent in the future. We're much more likely to guess wrongly that an African-American individual will do that than someone who is white. So the way we tackle that problem, obviously if we publish this, it might not sort of grip people the way that we would want. So we actually found some individuals in the data. So here's an example. Vernon Prater was arrested for shoplifting at a Home Depot. He was given a low-risk rating by the algorithm. Brisha Borden, stole a bicycle from a yard and rode around on it, and then actually returned it because it was a bicycle for a child and not very useful to her. And she was rated as high risk. So it turns out that are low-risk, I actually had committed to armed robberies and one attempted armed robbery. And that after he was assessed by this algorithm, he went on to you commit one grand theft offence and Brisha Borden had some juvenile misdemeanours and then subsequently did not commit any future offences or was not arrested for them. So this is a way to take that matrix of confusion which could be seen as very dry and hard to understand and put it in human terms. And that's the end of the show. If anyone has questions. It took longer than this when I practise. So what adrenaline I'll do Horrea. So how do you find that you keep the power of the data? Because when you, when you show that example that was on that last slide, wasn't the pops to my mind is Okay. Well, that's just 22 anecdotes, but you have a lot of data that led you to find those anecdotes, how you make sure that you're reporting keeps that power. Yeah. I mean, so behind this, right, there's a simple fact which is that the false positive rate for African-Americans is, I don't remember the exact like 20 percent higher rate. And that's definitely included in the storey that you would publish. We didn't publish this graphic alone, right? It was part of a probably too long chunk of text. So distilling it down, she sort of a simple fact that people can grasp, I think one way to make sure your data kind of comes along. One thing we found that sadly is not especially effective is to say things like we analyse 0.5 million rows. Like, you know, again, that sort of friction of where somebody you guys might glaze over and they stop reading. But we do also have all the code that was used to generate. This is on GitHub, all the findings or in a notebook. So we try to, for an audience that wants more, make that available, but also have a version that's sort of accessible for people who don't want to kinda to you on the bristle of life. Well, there any consequences for the company developing the algorithm for the Machine Bias, one that they face and the penalty, the police or stop using them well, so, so what's interesting, right, is that the word bias in criminal justice has a very specific meaning rate, which is, is the algorithm wrong at the same rate for both races? And it's like, Yeah, it is. So in terms of how kind of like fare is defined in that discipline, the algorithm is fair. Our argument was more, okay, so technically it's fair, but is it just right if we're making a guess about somebody's future violent behaviour based on an algorithm rate. Is it fair for it to be wrong in one direction for one race and in different direction for another is the company called North points, formerly Hitler gods now. But also they changed their name. They made claims that they were sort of fine tuning their algorithm was started the whole fight that's ongoing. So this is a shameless plug for ProPublica. This is unsolicited one, please go to donate the ProPublica.org and donate to provoke will go because they do great work. So you talked about how in some cases the storey is really in the outlier. So how do you go about deciding what to investigate and knowing where, where there's a storey before you actually get there. So in general, an interesting storey, right? Like the idea of news isn't something serve runs counter to expectation, I guess I would say. And often that's what an outlier divs. Sometimes it's so sometimes an outlier is just someone who looks especially bad. Frankly, in the investigative world, sometimes it's actually someone who looks weirdly good. And then because we have dark souls, were always like is that person cheating in some way? So I mean, I think there's kind of something inherently interesting about outliers. If the question you're asking is an interesting question to begin. Thanks so much everyone for attending.