5.1.0 Biases

Hi, today we're going to be talking about biases when analysing data sets and when dealing with big data.

So, it's a well-known idiom that you can prove anything you want to, if you use the right questions, and you approach it in the right way.

So, bias comes in many forms, but any form of bias, whether it's conscious or unconscious, whether its intended or unintended, will affect the outcomes of all analysis. So, just using the scientific method, and we've got a simplified version here, as a process for carrying out analysis.

So first we observe, then, once we've observed, we're going to come up with a hypothesis, we're going to create a test to prove our hypothesis, we'll then take a look at what came out of the test, we'll draw some conclusions. And eventually, we'll publish a result of our experiment. This is well known and it's well documented as the scientific method.

But what if somebody started their analysis the other way around? What if we started with the result? And we worked backwards from there. So, what if we already knew the result that we wanted to happen and we only looked for conclusions and we only looked for outcomes of tests that supported our result, then once we find that, we then look for a hypothesis that would lead to those tests having been carried out. Once we have that hypothesis that we've now created, or found, we can then look for, what would we observe, or what would we have observed that would have led us to that hypothesis. So, in effect, we can go and find the right and relevant proof that we need and push it through our process to create our own reality.

This happens, a huge amount when you look at Big Data Analysis. It happens when somebody comes to you and says, “I have some funding and I want you to do this research”. Unconsciously, you're going to want to do research and you're going to want to analyse the data that they're asking you to analyse in a way that is going to make the person that has funded you happy. This is subconscious bias and it could well be that you end up looking for results that will make them happy, as opposed to following the method.

Now, it also works the other way around.

If we went into our analysis, with no hypothesis at all, no expectation of what we might find no idea of even what we're looking for, then the likelihood is and the percentage shot here is people tend to jump on the first pattern or trend that they see whether that then bears out through the rest of the analysis or not, you'll latch on to it. And there's a tendency to kind of grab hold of that first thing that you've seen and then the rest of the analysis is geared around making it true.

So, for argument's sake, very simply, if we look at the cluster of information, the cluster of data points on the left-hand side, we can sit there and say, Okay, our attention was

immediately drawn to the large, teal coloured data points. So, we then go through and say, because they were the largest data points, they therefore make up most of the data. And we're now going to conclude that the data is mostly teal coloured. Now, that may or may not be correct, because there's more of the pink ones. In fact, there's even more of the small purple ones. And for those of you with keen eyes, each of those circles is surrounded by a white edge. So, if we were to add all of those things up, what would then be our outcome?

There's a rule that Dr Occam came up with and it's applied to philosophy as opposed to data analytics although, in my opinion, it applies equally to both and this is “the simplest explanation is usually the right one”. So, as Dr Occam puts it, suppose there exists two explanations for the same occurrences. The one that requires the smallest number of assumptions is usually correct.

So, another way to put this is, if you're standing outside, and you hear hoofbeats coming down the road, it's likely to be a horse, not a zebra.

There are other biases that exist.

There's the bigger is better theory and the bigger is better bias or the bigness bias. This is a belief that because we are analysing large data sets, and because we are analysing big data, we are going to find more relevant and more meaningful results than an analysis of a smaller data set word. This is not true, but we must be aware that this bias exists, we must be aware that people believe it, because it may taint our results.

We've also got the works on mine bias, also referred to as overfitting. This is where you create a hypothesis, you test it, you get the results you want from it. And it works every single time and it's replicable on the data you hold using the testing conditions you have. But as soon as you try to apply it to anything outside of your own controlled environment, it no longer fits an insistence that well, it works on mine, so yours must be broken and insistence that it's correct, regardless of what's the extended evidence would show you is called overfitting or trying to overfit your results.

There's something called a blending bias and easiest way to understand this would be looking at collection methods. If I came back with a voting census, or polling, as some might call it, so I sent out a whole bunch of people and I employed a whole bunch of different techniques like social media questions and email campaigns, people knocking on people's doors and asking them questions, people telephoning people and asking them and asked all of these people in all of these different ways, about their opinions on the current political climates, and who they might be voting for in the next elections. When I try and bring all of those datasets together into a single data set, each one of them will have its own bias, it could be because of the way that somebody asked the question. It could be because of the type of people that answered the question. You know, typically, when you look at it, and you know, excuse any stereotypes here, but typically, when you look at it, you are likely to get younger people filling out social media questionnaires than you are older people, you're likely to get more engagement from older people on the phone, or even when you knock on

their door than you will from younger people. It's just a thing. It's a truism and it leads to blending bias when you try and bring those data that all of those data points together.

This is one of my favourites, it's the what if bias or the statistical method bias. If you give a statistician enough data points and the freedom to ask any question they want, they will be able to prove anything. Because they can simply say, what if we did this? What if this was slightly different? What if the person that answered the question felt this way at the time? What if this is an anomaly, and they can include or discard values that don't fit their hypothesis very, very easily by creating a scenario in which that shouldn't that that value or that data point should never have been included to begin with.

And finally, here we've got the unknown or the ambiguity bias. Now, this is where there are certain points of systems there are certain points of the world that we simply Don't know what impact that might have had. If your data has been collected through, for simplicity sake, a third-party system, you may not understand, or you may not have access to all of the information for how that system processed the data on its way through. Which means you might have an unknown bias within your data. That is just ambiguous. You don't know what it was, you don't know how it got there, and you don't know how it's going to affect the results. But again, it's important to be aware of it because that might affect what kind of weighting we put on those results, or how we normalise them into other records.

So this concludes our session on biases.

Next we're going to be talking about sub setting and subsets of data