Blossoms - Correlation and Causation Transcript

Hi, my name is Kamilya Tazhibayeva. I'm doing post-doctoral research in economics at MIT.

I recently came across this eye-catching article headline-- Credit cards can make you gain weight. How bizarre and mortifying. We see these kinds of statements and headlines so often in the news media and in advertising. Like, “Our deodorant has ingredients that will make you irresistible to women”, or ‘Pets reduce the risk of heart disease”.

Bold and assertive, such statements grab our attention by their very incredulity. And yet, the next moment we might already be wondering if maybe there's a way they could be true. How do we make sense of such claims and news articles?

What they usually assert is that when one thing happens, like using a credit card, something else is also likely to happen, like weight gain. This is what's known as a correlation. When you have one, you tend to also have the other. In case of the credit cards, it turns out that when people pay for their groceries with a credit card rather than with cash, they are more likely to buy more of unhealthy junk food items.

But is it just a coincidence, or is there some connection between paying with a card and purchasing more junk food? And if there is a connection, what do you think it's based on? What is causing what? Why don't you break into groups of two-three people and discuss these questions.

I hope you had a fruitful discussion. I'm still not clear myself whether it's just a coincidence or whether there's some causation involved. How do we distinguish the two? That's what today's lesson is about.

First, let's get our definitions straight. Let's start with causation. Causation is when one variable can influence another. One variable can either induce another variable to happen, or make an already existing variable to change.

For example, if you decide to play in the rain without an umbrella, your clothes will get wet. So being rained upon first causes your clothes to become wet, and then changes how wet they get. Since the rain wets the clothes directly, this is an example of direct causation.

As you've probably guessed, causation can also be indirect. Suppose after your clothes got wet in the rain, you got home and put your wet clothes in the laundry. In this case, being rained upon indirectly caused you to do laundry.

Now we have some idea about what causation is. But what about correlation? Correlation measures the extent to which several variables move together.

A positive correlation indicates the extent to which those variables increase or decrease at the same time. A negative correlation indicates the extent to which one variable increases as the other decreases.

For example, when you are a kid, the older you get, the taller you become. This is a case of a positive correlation: when you are still growing, your age and your height both keep increasing.

Note that whether a correlation is positive or negative depends only on whether the variables we're talking about move in the same direction or in opposite directions. It has nothing to do with it being a good or a bad thing.

For example, when you just start playing a new sport, you'll make a lot of mistakes. But the more you practice, the fewer mistakes you'll make. As the amount of time you spend practicing goes up, the number of mistakes you make goes down, so there is a negative correlation between the two. But the fact that you're getting better is a good thing!

A correlation is called "strong" when the variables move together almost in unison. And a correlation is called "weak" when the variables just barely move together.

As you've noticed, there is nothing in the definition of a correlation about causality. So all a correlation tells us is whether the variables we are examining tend to move together or not. It does not tell us why.

So can correlation and causation be connected? Let's look at an example.

100% of people who breathe, die. We know that all living people breathe, and also that all living people eventually die. So, clearly, this is a case of a very strong positive correlation. However, breathing doesn't make people die. So, despite its strengths, this correlation does not imply causation.

On the other hand, 100% of people who don't breathe, die. Take a moment to see for yourself that this is also a case of a very strong positive correlation. But is this correlation based on causation? Well, we know that if a living person stops breathing, they will die. So in this case, we observe a correlation between not breathing and dying because one is indeed directly causing the other.

Do you think any of your own characteristics correlate with one another? Let's find out.

Welcome back. I hope you had fun with this activity.

My guess is that you have found a negative correlation between height and hair length, as boys on average are taller and wear their hair short. But depending on fashion trends or cultural norms, you could have instead found a zero or a positive correlation.

We're now well-equipped to examine the connection between using credit cards and buying more of unhealthy food items. Assuming that the correlation between the two is indeed strongly positive, what does it tell us?

It's reasonable that people pay with cash for small expenses, but use a credit card for larger expenses. So people who shop for a bigger event – say, a large party – are more likely to pay

with a card. But because they're getting more things overall, the number of unhealthy items they get is larger in total, but not in proportion. So we cannot conclude that they would eat more junk food than those who pay in cash.

At the same time, people usually don't like parting with their money. So when they pay in cash, they tend to spend it sparingly. But when you pay with a card, you're not physically giving away your money, even though you are for all practical purposes. As a result, people paying with a card usually spend more freely. So it is possible that people paying with a card would indulge in some junk food that they wouldn't have bought if they were paying in cash.

The conclusion? The relationship between using credit cards and eating more junk food is far from certain. But at the same time, there is some evidence that we should take note of and see that our own behavior is not affected in the same way without us realizing it.

We already saw that correlation is not always based on causation, either direct or indirect. So what are other possible reasons behind the correlation? Well, often there are additional factors that influence both variables between which we see a correlation.

For example, it is known that for adults, the smaller their palm, the more likely they are to eventually reach very old age. That is, the size of a person's palm is negatively correlated with their lifespan.

It's an interesting fact, but what could be causing it? It's hard to see at first. Obviously having larger hands doesn't directly cause people to die younger. Here's a trick. When trying to find what lies behind a correlation, it's important to look not only for causation, but also for associations. Let's try it.

Which adults have smaller palms? Well, women usually have smaller hands than men. We also know that women tend to live longer than men. So women have smaller hands and have, on average, longer life spans than men. And as a result, we observe a negative correlation between the size of adults' palms and their longevity.

So just like that, by looking for associations, we have uncovered what this confusing correlation is based on. Such additional factors behind a correlation are called omitted – or lurking – variables. They are called so because they are omitted from the original correlation statement, and are lurking in the background.

Here is another interesting real-life example. One study has found that the amount of ice cream sold by street vendors is positively correlated with the number of serious crimes committed in New York City during 1980s.

What do you think of this finding? Is this pure coincidence, or can these two factors – something so wonderful as ice cream and so horrid as crime rates – be related? What are possible omitted variables that are lurking in the background?

Why don't you break into pairs and discuss these questions. Remember to look not only for causation, either direct or indirect, but also for associations.

Guys, while you were discussing the study, I ran into my friend Stefano.

Hi guys.

Stefano, have you heard that the more ice cream is sold in New York, the more street crime happens there?

Wow. That sounds crazy. Is that just a coincidence, or can you find a reasonable explanation?

I don't know. Let's first examine the possibility of causation. Can more crime cause people to buy more ice cream?

So maybe people get stressed out by all the crime, and they eat more ice cream to feel better? So crime causes the stress, and stress causes the ice cream sales.

You know, ice cream does tend to cheer me up. But this story seems a little far-fetched.

So let's think about this for a second. Can causality go in the other direction? That is, higher ice cream sales cause higher crime rates?

You know, ice cream does have a lot of sugar, which tends to make people a bit more hyper.

Yes, but this doesn't seem plausible either. I mean, you'd really have to eat a lot of ice cream to lose control. Even kids don't do that.

Yeah, ice cream does tend to fill me up. So perhaps we'd better look for omitted variables. So let's think of associations. What does ice cream associate with?

Summer! And hot weather!

This sounds promising. Does the amount of crime also associate with hot summer weather?

Well, when it's cold people staying indoors a lot more, so there are fewer people on the streets. And it also gets dark a lot earlier in the winter, so people are less likely to go out in the evening.

Exactly. And during summer, the days are long and the weather is hot. People go on vacation, they spend weekends at outdoor markets and festivals, and long evenings in cafes.

That's true. So houses stay empty longer during summers, making break-ins more likely. And with more people out and about, as well as with large crowded events – such as fairs and festivals – street crime is also more likely.

It seems that both ice cream sales and crime rates plausibly associate with hot summer weather. So it's likely that summer is the underlying lurking variable causing the positive correlation between ice cream sales and serious crime rates.

Well, I'm really relieved that ice cream doesn't cause crime! Because, let's face it – I love ice cream!

Me too! Notice, though, how we managed to come up with stories of causality between ice cream sales and crime rates, either saying that people ate more ice cream to comfort themselves, or that ice cream made people more hyper. But these stories didn't seem very likely.

And although it wasn't straightforward to work out the association that lead us to thinking of summer as the underlying lurking variable, the reasoning behind it was much more sound.

You know, omitted variables are a very common explanation for correlations. After all, the world is very complex, and there exist multiple connections between different events and things. So remember to look for associations to uncover the less obvious, but more plausible reasons for a correlation.

But so is there always a reasonable explanation for a correlation?

That is a very important question. Let's find out during our next activity.

Consider these two graphs. Each traces a pair of variables over 11 years. We see that each pair is strongly positively correlated. Divide into pairs and decide which one of you gets which graph. Then take turns explaining to each other what you think is going on in the graph you've chosen.

Can there be some reasonable explanation for this correlation, or is it pure coincidence? Explain to your partner the reasoning behind your thinking, and see if you can convince them, or if they can poke holes in your argument.

I hope you had fun discussing these graphs. We talked about them with Stefano and decided that they both are pure coincidence.

Yes, it's hard to find any plausible causality behind these correlations, or any likely omitted variables.

Yeah, even if you try really hard and line up a chain of, like, dozens and dozens of omitted variables, connecting one into another, to link, say, Nicolas Cage movies and drowning in pools, all these “connections” would be pretty far-fetched.

Absolutely. Well, this was fun! But I’ve got to get back to work.

OK. Well, it was great to see you!

Take care!

When a correlation is based purely on coincidence, it's called spurious correlation, which means “false”, “fake”, or “illegitimate”.

Here is my favorite example of a spurious correlation that I saw online. This is a picture of a beautiful mountain range. Since we've been talking about crime rates in New York, let's plot the murder rate in New York state. Is this possible? Can this unknown mountain range be somehow connected to the number of murders in New York state?

Well, I highly doubt it. To me, it looks like a pure coincidence, although a very good one. This is a great reminder that no matter how strong a correlation is, we shouldn't be fooled into thinking that it implies causation, or that there is some reasonable explanation behind it. In many cases, a perfect correlation is just a coincidence.

So when you come across a correlation, say you read about it online or hear about it on TV, don't jump right away to a conclusion that it implies causation. Instead, think about other possible explanations that could be behind that correlation, including likely associations, and the lurking variables these associations point to. This will give you a more complete picture of what is going on. Then you can think how plausible each potential explanation is before making up your mind what this correlation actually tells you.

So far, we've covered direct causation, indirect causation, omitted variables, and spurious correlations.

Next, let's talk about reverse causality. Reverse causality is when the variables in question are indeed connected, but not in a way that you would at first expect.

Here is a simple example of reverse causality. The kid is complaining that he wishes the pilot didn't turn on the seat belt sign so often, because every time he does, it gets bumpy. The way this correlation is stated, it seems to imply that turning on the seat belt sign causes the bumpiness.

I'm sure you can see where the kid got it backwards. The causation is in fact going the other way. Because of the air turbulence, the pilot turns on the seat belt sign.

Let's practice a bit more in identifying different types of causation. Here are two examples that may or may not contain causality. Divide into pairs and decide which one of you gets which cartoon. Then take turns explaining to each other what you think is going on in the cartoon that you've chosen.

Does the statement in the cartoon contain any causality? If so, what type of causality is it? Which direction is it going? Explain to your partner the reasoning behind your thinking, and see if you can convince them, or if they can poke holes in your argument.

I hope you had fun with the cartoons. As you've probably figured out, they're both good examples of reverse causality. For example, it is tempting at first to think that this person isn't going to try doing something because they're terrible at it, whatever "it" is.

But when you stop and think about it, it seems more reasonable that the causation is going the other way. They are terrible at it because they've never, in earnest, tried doing it, and so never got a chance to get better at it. After all, practice makes perfect!

By now, we know better than to jump to conclusions about causality when we see a correlation. That doesn't mean, though, that all correlations should be mistrusted.

As we've seen earlier, in many cases there's an underlying reason for a correlation. However, that reason is often well hidden. Sometimes examining a correlation helps us uncover this hidden root of the cause. That is exactly what physician John Snow did to uncover the way cholera is transmitted.

In the 19th century, there were a number of large cholera outbreaks in London, killing hundreds of people in a matter of days. At the time, no one knew how cholera spread. A lot of physicians thought that it spread through the air. However, during those outbreaks, only a part of people in the same neighborhood, who breathed the same air, would contract the disease. There was no clear pattern as to who got sick and who didn't. It was a puzzle.

Dr. Snow went to examine the Soho district during its cholera epidemic. At that time, people had to pump all their water from public water pumps, and generally they would use the closest pump. During his visit, Dr. Snow realized that most of the people who got sick lived very close to one specific pump. So he mapped the neighborhood, noting on the map the number of sick people in each building, and then added to the map the locations of all water pumps.

A clear pattern emerged. All incidences of disease were concentrated around one water pump. The map showed that those living closer to any other pump were much less likely to get cholera. Dr. Snow also did some statistical data analysis to confirm this result.

Based on this evidence, he came to the conclusion that cholera spread not through air, but through water, which is indeed the case. At the time, no one knew about germs or how they spread. So Dr. Snow had no way to understand the mechanism of how cholera could possibly spread through water. Despite this limitation, the clear correlation in the data helped him reach the right conclusion. He convinced the city officials to close the suspected pump. After it was shut down, the rate of infection dropped significantly.

And this is the whole point of correlations – they guide us in promising directions to help us understand various problems, both simple and complex. Chances are, we all, at one point or another, learned something from a correlation. Let's try to recall some of those useful experiences.

I hope you had a productive trip down the memory lane.

Let's review what we've learned today. We now know that two variables are correlated if they move together. Positive correlation means that these variables move in the same direction, and negative correlation means they move in opposite directions. And we should remember that, for

correlations, the adjectives positive and negative have nothing to do with whether what we see is a good thing or a bad thing.

We learned that, in many cases, a correlation is based on causation, which can be either direct or indirect. Very often, while there's no causality between the correlated variables, there is a reasonable explanation for the correlation in the form of additional factors that are lurking in the background.

We've also talked about cases where the correlation we see is pure coincidence, and has no causality behind it.

Finally, we talked about cases when causation goes in the opposite direction from what we would expect at first.

So what is the takeaway from all this information? Well, the first thing to remember is that correlation does not imply causation. A correlation also does not imply an existence of a reasonable explanation. It could be just pure coincidence.

At the same time, keep in mind that correlation is a good place to start at when looking for an explanation. And, when looking for a reasonable explanation, don't forget to think about associations.

So that's pretty much what a correlation is. It is a useful indicator that signals to us that we could potentially learn something new from examining it.

But to make good use of a correlation, we have to do our part. We should examine the involved variables, think of possible directions of causation, and of potential omitted variables, and consider the possibility that the correlation is spurious. After that, we can decide if any of the explanations will have thought of are plausible enough.

So correlation is a promising starting point for finding the answer. But it's not the end of the process. It's not the answer itself.

Here is a good quote on this. "Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing: “Look over there."

Thank you for spending time with us today. I hope you’ve learned something new from this lesson, and I certainly enjoyed doing it. To help you remember the main point, let me leave you with this comic.

Hello, and thank you for being interested in this lesson. It introduces the students to the concepts of correlation and causation. This lesson will guide the students in identifying and examining various possible reasons behind a correlation. And it will help them to understand the usefulness as well as limitations of the information provided by a correlation.

The ability to make well-founded decisions and draw accurate conclusions makes a huge difference in a person's life. From professional activities, to social and personal areas of life, it enables us to continuously make positive changes and improve our performance.

Understanding the distinction between correlation and causation is an important component of this ability. It enables the person to develop their own take on the situation and to think independently, rather than having to take someone else's viewpoint on faith. It prepares us to analyze and handle independently various possible circumstances. After all, one cannot simply memorize the correct course of action for every possible situation in life.

In this lesson, we will cover the cases of omitted variables behind the correlation, of spurious correlation, and of reverse causality. Throughout it, students will work actively through a variety of cases and examples. This lesson is suitable for all high school students, but it might be especially valuable for older students.

The lesson starts by introducing the students to the concept of a correlation. The goal of the first activity is to get the students, with the help of an example, to start thinking critically about connections between a correlation and a causation.

It might be helpful to encourage the discussion by asking pointed questions, such as: “When are you more likely to use a credit card?” or “When do people generally buy junk food?”

The second activity takes this one step further. It uses a hands-on exercise to illustrate the overwhelming presence of all kinds of correlations in our everyday lives. The goal is to demonstrate to the students that a correlation is not some abstract concept, but rather something common and relatable. The purpose of the accompanying discussion is to help the students start questioning the value of the information that is provided by a correlation.

In activity 3, students practice identifying omitted variables behind a correlation. Here, again, it might be helpful to lead the discussion in the right direction by asking the class questions that help them start thinking critically about realistic connections between different events. It is important to keep a balance between encouraging the discussion on the one hand, and at the same time not revealing the answers, so that the students get to practice their analytical skills.

Segment 4 discusses in detail the case that the students have worked on during the preceding activity, activity number 3. So depending on time availability, and on how easy or confusing the students have found activity 3, it might be fruitful to start activity 4 by first having a short class-wide discussion on what the students have thought about the explanation provided in segment 4, whether or not they agree with the proposed conclusions, and why.

In activity 4, students practice distinguishing between a correlation based on causation and a spurious correlation. Making this a paired activity, so that each student gets to explain their reasoning, encourages them to develop and structure their arguments. This activity should require little guidance, although some pairs of students might need a little nudging in the form of guiding questions to help them start the discussion.

Activity 5 has a similar structure. It takes the students one step further – here, they practice identifying the correct reason behind a correlation among various alternatives they've already learned about in this lesson.

The final activity helps students relate the lesson's concepts to their own experience. It might be helpful to give the class a short, relatable example from everyday life.

I hope you and your class will enjoy this lesson. Thank you so much for watching!