Module 1: Casual Analysis

Table of Contents

Lesson 4: Course Introduction

Course Introduction

MUSIC] In the last few years there has been tremendous growth in the volume of data that can be tracked about consumers in parallel. There have also been huge advances in the tools and approaches we can use to understand and analyze big data. Have you ever wondered how companies and researchers make sense of these huge amounts of data? Have you ever wanted to learn how you can use this data to help solve marketing problems? Then you're in the right place. Welcome to applying data analytics in marketing. In this course you will learn four key tools and approaches in marketing that are commonly used by firms and marketing managers in their decision making. These tools are causal analysis, survey analysis, textual analysis and network analysis. Within each tool our discussion will cover two aspects. The first will be an overview of the approach or the tool. The second will be a marketing application in our module on causal analysis. We will focus on experimental and quasi experimental methods and their application to Omnichannel marketing. In survey analysis we will focus on regression models and their application to customer satisfaction data from an airline company in textual analysis.

We will focus on natural language processing and topic modeling using social media data. Finally, in network analysis we will use network structures and influencer brand personality analysis and their application to social media data. As part of this course, you will also get to work with marketing datasets and apply course concepts in are. I will also provide you a deep intuition of when and how these approaches can be used in practice and how to really interpret what you find in them. So welcome again. I look forward to having you in my course. I will also be co teaching this with Joseph Yun, who you will see a lot in module two through module four welcome. [MUSIC]

Instructor Bio: Unnati Narang

Instructor Bio: Unnati Narang

[MUSIC] Hi, my name is Unnati, I'm an assistant professor of marketing and the RC Evans Fellow of data analytics at the Gies College of Business here in Illinois. If you're not familiar with our campus, I'm standing right outside the bowlers courtyard, which is one of our business school buildings in Urbana Champaign. Before I came here I got my PhD in marketing from Texas A&M University. And before I started working on academic research I used to be an entrepreneur. I founded a retail company that was in the business of selling musical instruments and a recording studio for musicians

If you're wondering whether I can play a musical instrument or not? The answer is no. So even though we were a physical space with over 2000 square footage in New Delhi, back in 2011, most of our customers we're still discovering us through digital channels and social media. That's when I got really interested in studying how consumers and firms make decisions in new digital environments. In some of my research now, I have found that mobile apps launched by branded retailers can lead to more purchases not only digitally but also in physical stores of the retailer. Interestingly, I also find that app users are likely to make more product returns in other research. I study digital platforms including Coursera the very platform we're connecting on. I love to teach marketing analytics and all things related to the challenges and opportunities of big data. I'm a big believer in the power of online education, so I'm thrilled to be teaching this course. To thousands of you joining from across the world I look forward to engaging with you in our virtual classroom.

Instructor Bio: Joseph Yun

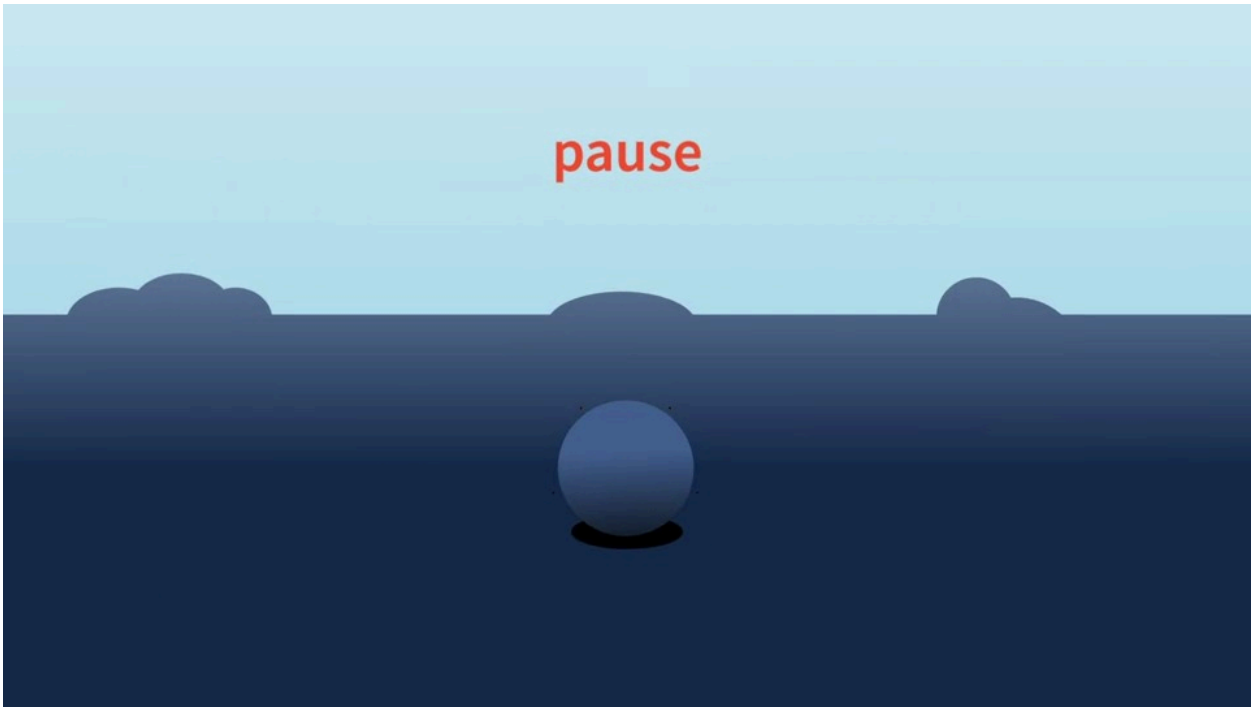
Instructor Bio: Joseph Yun

Hello. My name is Joseph Yun, but you can call me Joe. I am a research assistant professor of accountancy, as well as the director of data science research services here at the Gies College of Business at the University of Illinois, Urbana-Champaign. My educational background is a bachelor's in computer science, a master's in advertising with a focus on social psychology, as well as a doctorate in informatics, bringing together the realms of computer science and advertising psychology, specifically called computational advertising. My research interests are in data science for advertising and marketing, as well as advertising privacy, social data analytics system.

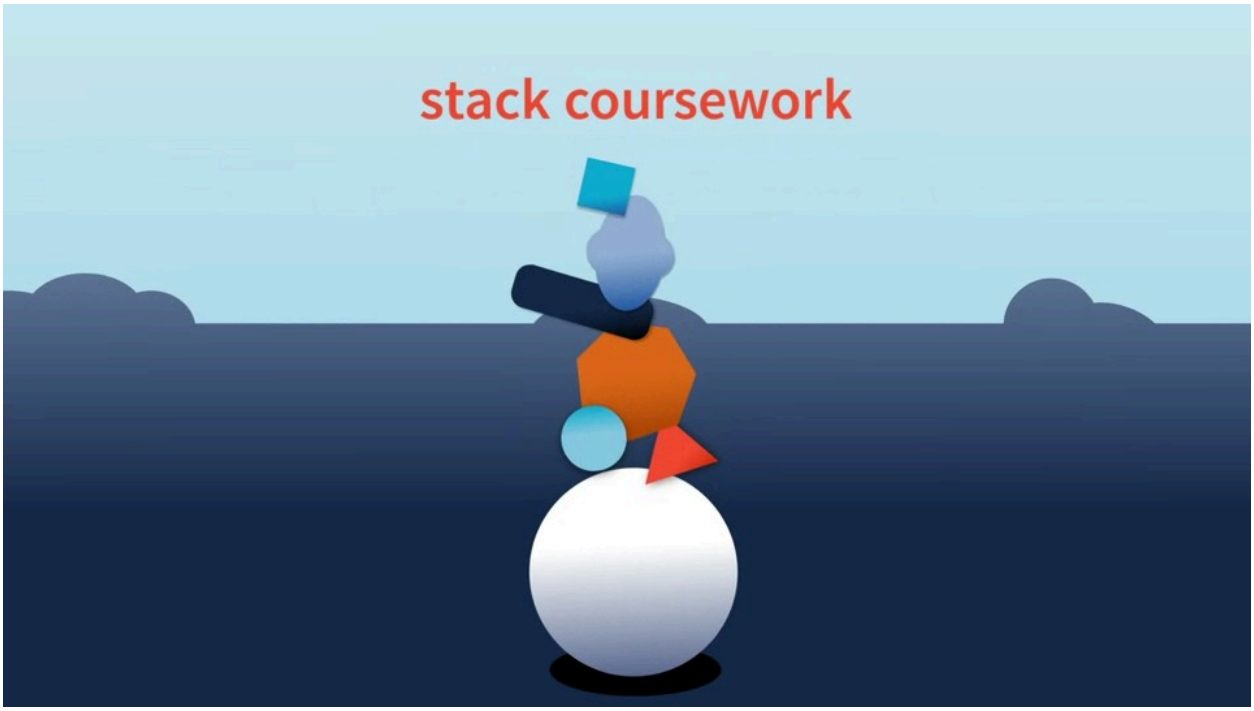
My team builds out a system called the social media macroscope, as well as I am interested in generally this concept of advertising for good. You may be wondering why I am a professor in accountancy, when my background is clearly much more advertising and marketing focused. But basically I was given an unbelievable opportunity, to join, I believe the best accountancy department in the whole wide world, to help them with regards to data science or join many of them that are doing data science, so that we could further data science for the realm of business, and specifically also for this Gies College of Business. I look forward to courses with you, and I just thank you so much for your time.

Learn on Your Terms

Learn on Your Terms



Too often, smart hardworking busy people miss out on education because of traditional linear learning. Learn on your terms.



With stackable online content from Gies College of Business, you can take self-paced classes, earn transcriptable credit, pause, earn a degree. Switch in stack coursework, earn a certificate, or learn however you want. You'll get expert-led education and big or bite-sized increments. Wherever you are in your learning journey, the right time to start this is your time.

Causal Analysis: Module Introduction

Causal Analysis: Module Introduction

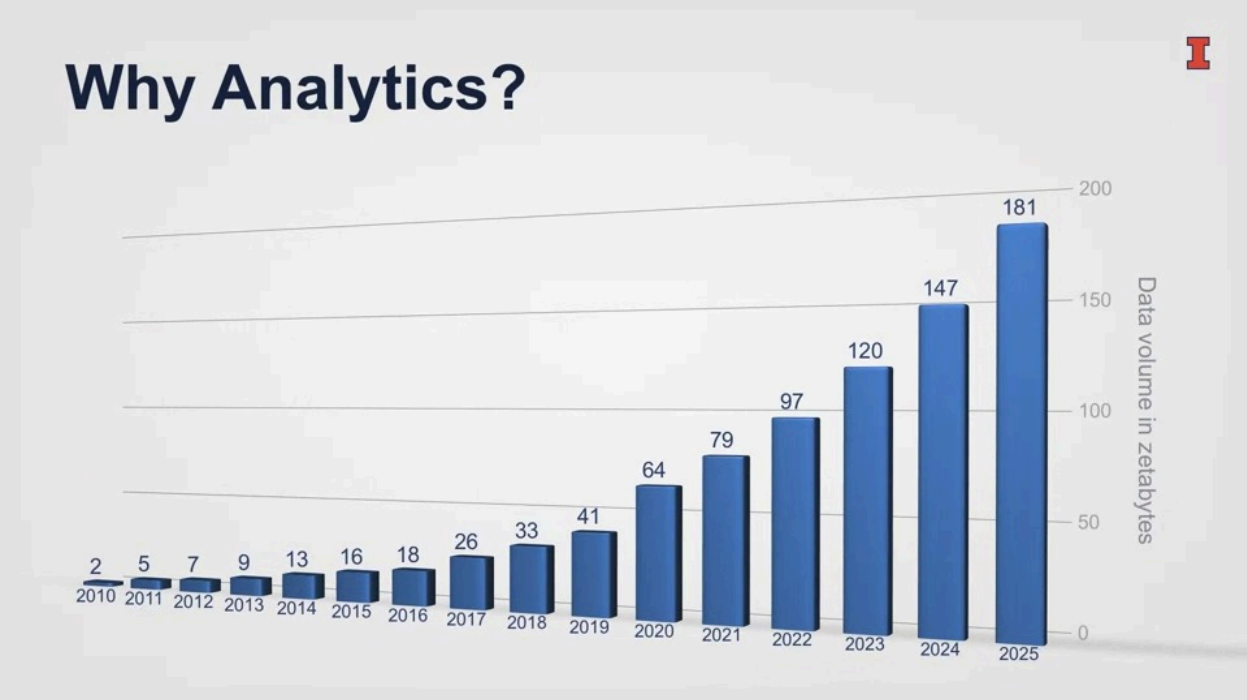
In this module, we will discuss analytics and marketing and dive deeper into causal analysis, an important tool for analytics. We will start with a broad overview of why analytics is important for marketeers.

What are the various types of data, the process of applying analytics, and marketing and the different types of analytics? We will then delve deeper into causal analysis. Within causal analysis, first, we will learn about the conditions necessary for causality, the challenges of getting causal estimates, and randomized and quasi-experimental approaches to causality.

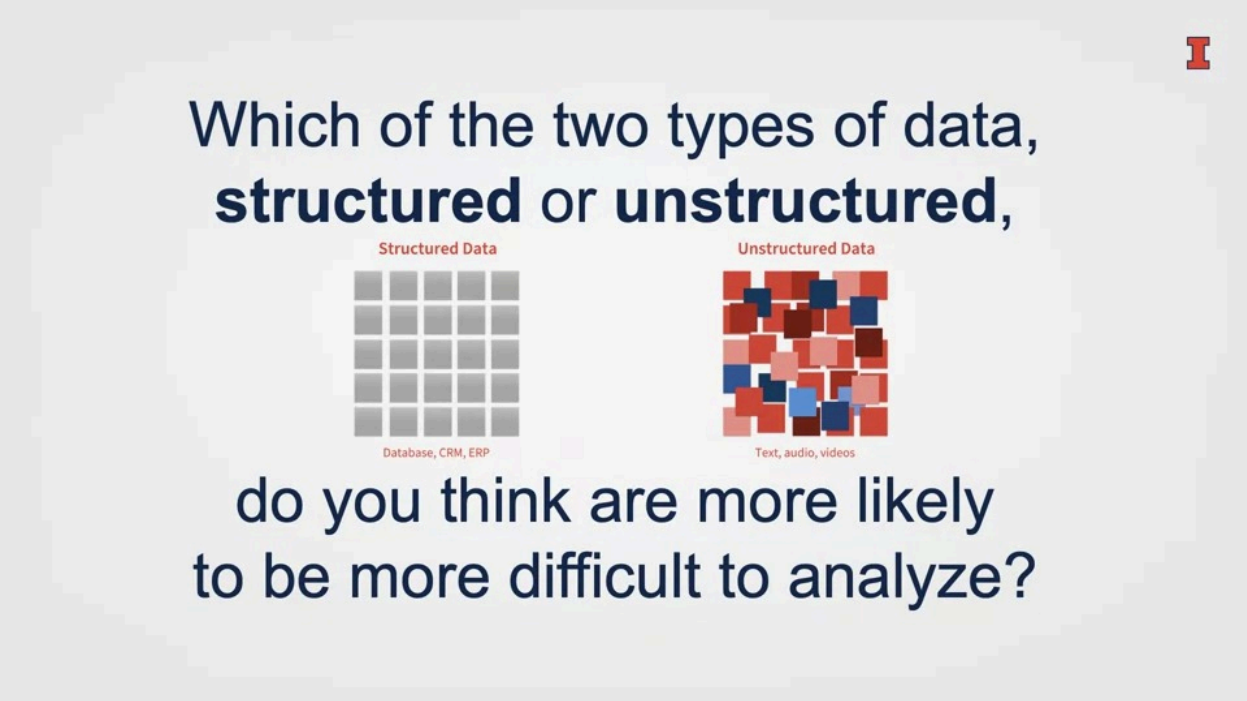
Next, our marketing application in this module will focus on omnichannel marketing and we will discuss the impact of a retailer's launch of a new mobile app. So let's get started. [MUSIC]

Introduction to Marketing Analytics Part 1

Introduction to Marketing Analytics Part 1



The goal of this segment is to give you an overview of analytics and marketing. We will discuss questions such as, why is analytics increasingly important in marketing? How should we think about applying analytics and marketing? And what are the various types of analytics? So why marketing analytics? Well, a compelling reason to learn and apply marketing analytics is the sheer growth in the volume of data.



On this chart, I'm showing you the volume of data created, captured, copied and consumed worldwide in the last decade or so. These data are projected to go from just two Zeta Bytes in 2010, 281 Zeta Bytes in 2025. Each zero Byte is equal to ten to the 21st power number of bytes. It's hard to even visualize, often these increasing volumes of data can contain valuable information and insights to help managers make better decisions.
Play video starting at :1:17 and follow transcript1:17
In addition to the growth in the volume of data, another factor contributing to the growth of analytics is the type of data. There has been a tremendous rise in the amount of unstructured data in the last few years. In fact, 80% of the data generated today tends to be unstructured.
Play video starting at :1:39 and follow transcript1:39
These types of data warrant additional analysis to be used in a meaningful way. So what do we mean by structured and unstructured data? You can think of structured data as data that can be recorded in rows and columns. It typically comprises numbers, dates, strings, for example, the monthly sales of a farm are structured data. On the other hand, unstructured data are data that cannot always be recorded in rows and columns. These can be images, audio, video and textual data. For example, restaurant reviews written by consumers on yelp are unstructured textual data.
Play video starting at :2:24 and follow transcript2:24
So which of the two types of data, structured or unstructured do you think are more likely to be more difficult to analyze?
Play video starting at :2:39 and follow transcript2:39
Unstructured data are typically more difficult to analyze because they need pre processing before any kind of statistical analysis can be performed. In this course, we will talk about text analysis using social media data, now that we've discussed why there is a need for analytics. Let us think about how you can structure a marketing analytics problem. There are various approaches to thinking about analytics. You can think of the analytics pipeline as getting the right data, developing appropriate models, writing up reports of your analysis and then telling a story.
Play video starting at :3:17 and follow transcript3:17

I like to think of it as four different steps. The first step is defining your analytics problem or the question, what is the business problem? What are we trying to learn? The second step is the data. What data do you need to answer your question? What is the source of that data? What is the data generating process? What format and sizes do the data come in?
Play video starting at :3:45 and follow transcript3:45
The third step is the actual analysis. What are the appropriate tools and methods to conduct the analysis? What types of analysis are needed?
Play video starting at :3:56 and follow transcript3:56
The final or the 4th step is insights? What have they learned from the data and analysis about our questions. One key thing to remember about this process is that the steps are not always sequential.

You may learn something from the data that can then help you go back and redefine or refine your initial problem.

Play video starting at :4:20 and follow transcript4:20

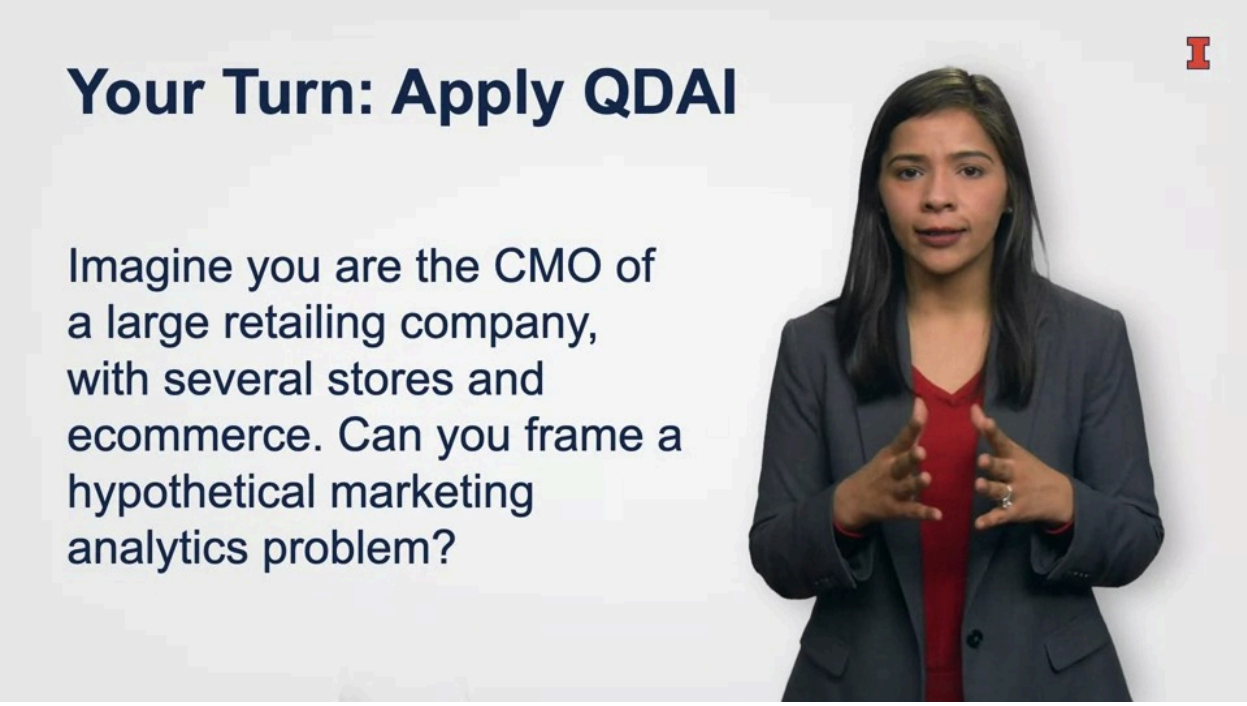
The marketing analytics process can be readily applied to any marketing problem. So why don't we try a quick thought exercise to do this? Imagine you're the CMO at a retail company with several stores and an e commerce website. Feel free to pick a product of your choice for this exercise. How will you frame the marketing question? What data and analysis could you use? What potential insights can come out of this?

Play video starting at :4:50 and follow transcript4:50

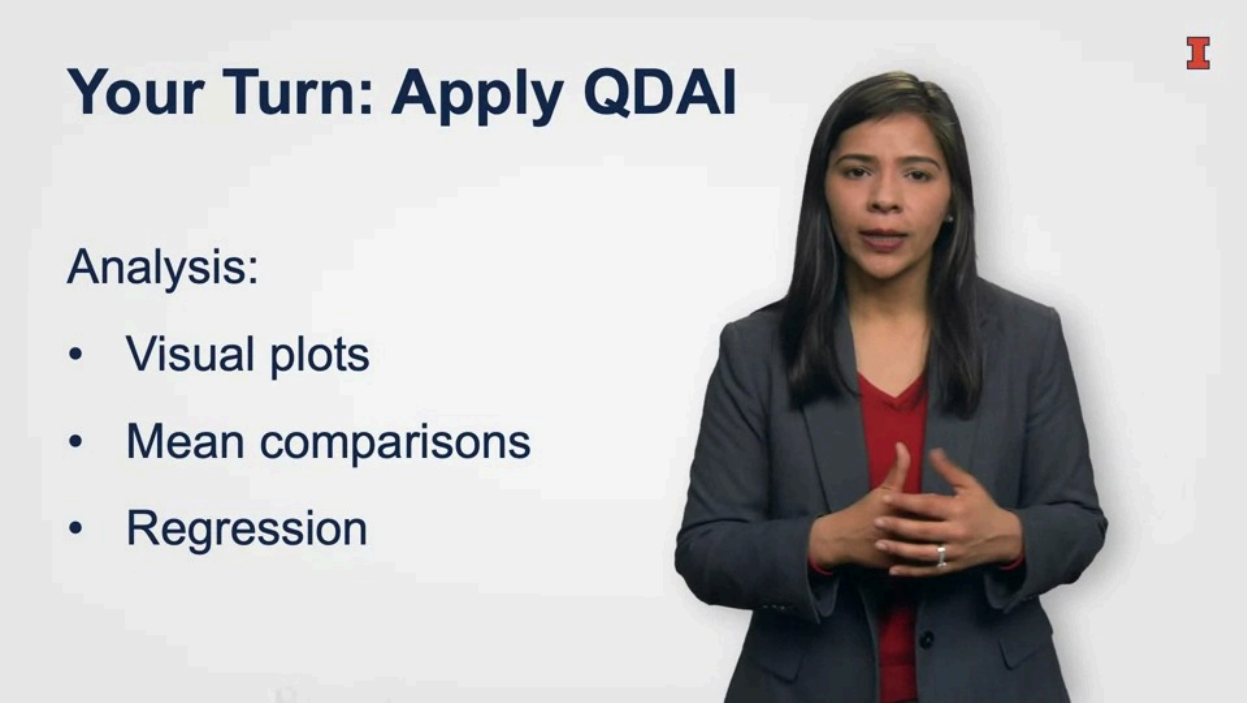
Why don't you think about this? And in the next segment we will examine a hypothetical scenario that I came up with.

Introduction to Marketing Analytics Part 2

Introduction to Marketing Analytics Part 2



Now that you've identified a marketing analytics problem, you've hopefully gained appreciation for how we can look at a wide range of relevant analytics problems in this hypothetical example. Which brands should the retailers stock? Should they open new stores? If so, where should their new stores be? Should they integrate their various channels by offering audit online, pickup in store types of services? One question I came up with is about the launch of a rewards program. The retailer may be interested to know how a new rewards program offered in their app may affect their online and offline sales. What kind of data can help you figure this out? In the ideal world, you could randomize some consumers who get access to the rewards program versus those who do not get access to the rewards program and compare their online and offline purchases. While this kind of randomization is often not possible, companies do conduct hundreds and thousands of such experiments in their digital products. You'd also need data on each consumer's online and offline purchases. Ideally, also whether or not they use the rewards feature. We will discuss the various approaches to analyzing different types of data. But at the very simple level, you could compare the sales from those who were given the rewards feature versus those who were not. You could do this by simply plotting the raw data for the two groups. If the randomization worked, these groups should look alike pre-randomization in their various characteristics and their past spending. To assess the effect of the rewards on sales, you can also plot their online and offline purchases after being given access to the rewards. Finally, you can do statistical tests of mean differences and run a regression analysis.

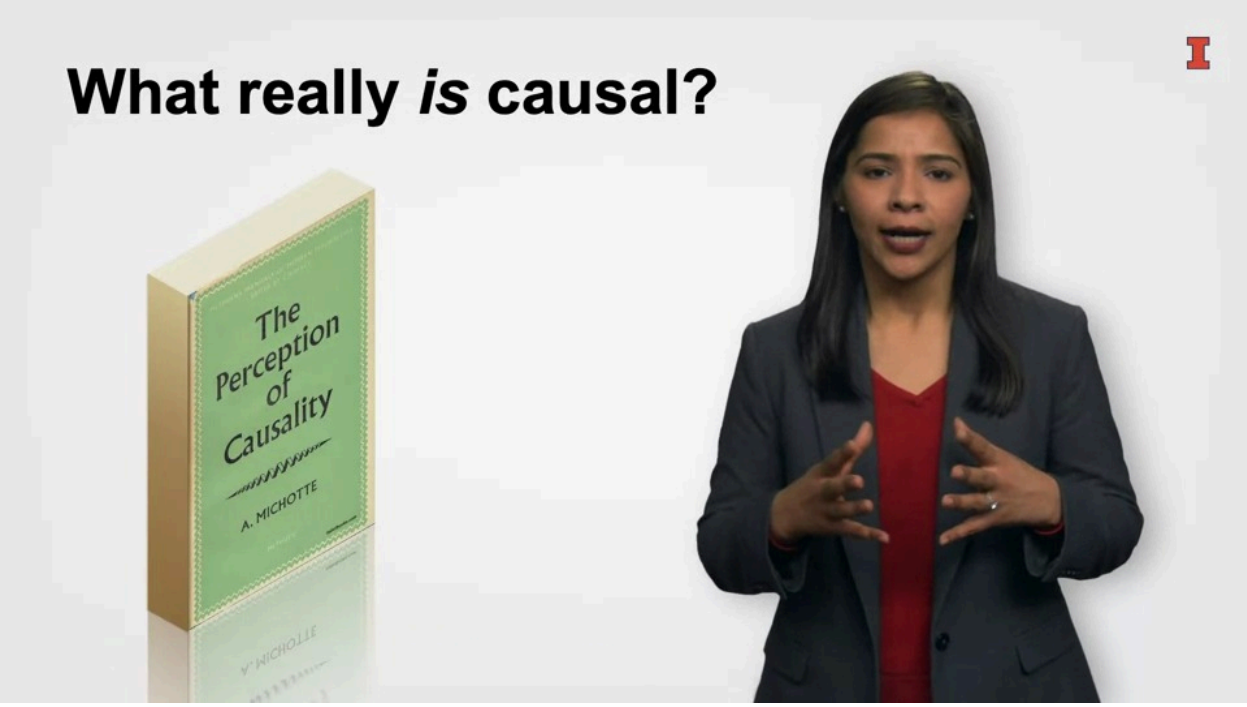


Based on the data and analysis, what are some insights you might discover about your question? Remember in this case, our question was, what will be the effect of a rewards program on sales and stores and on websites? We might find, for example, that those assigned to the rewards feature spent more online or in the app, since it allows them to discover this rewards feature but that their spending in stores remains unchanged. Or if the rewards are redeemable in stores we may also see some increases in stores spending with even a potential decrease in online spending. The final topic for this segment on the introduction to marketing analytics is the types of analytics. Descriptive analytics is about describing what happened. It simply documents the current state. For example, what is the current conversion rate from online traffic to sales? Predictive analytics is about forecasting what will happen in the future. It projects into the future based on some past patterns or trends. For example, based on a user's past purchases, is he or she likely to buy next month? Much of the machine learning analysis tends to be predictive. Finally, causal analysis is about understanding the cause and effect of one variable on another.

For example, will introducing a new product affect the sales of existing products of a firm? Think about the marketing analytics problem we framed in the last segment. What type of problem is it? Is it descriptive, predictive, or causal? Recall that we were interested to find out the effects of a rewards program on sales in different channels of a retailer. As such, this is a causal problem. While descriptive and predictive analytics can provide useful insights to managers, causal analysis helps identify the chain of causality and truly understand the impact of various actions, interventions, or external policy changes on outcomes of interest. Causal inference is a powerful tool for marketers. In the next segment, we will more formally discuss the notion of causality. We will learn to apply it for both experimental and observational settings.

Causal Analysis: Overview

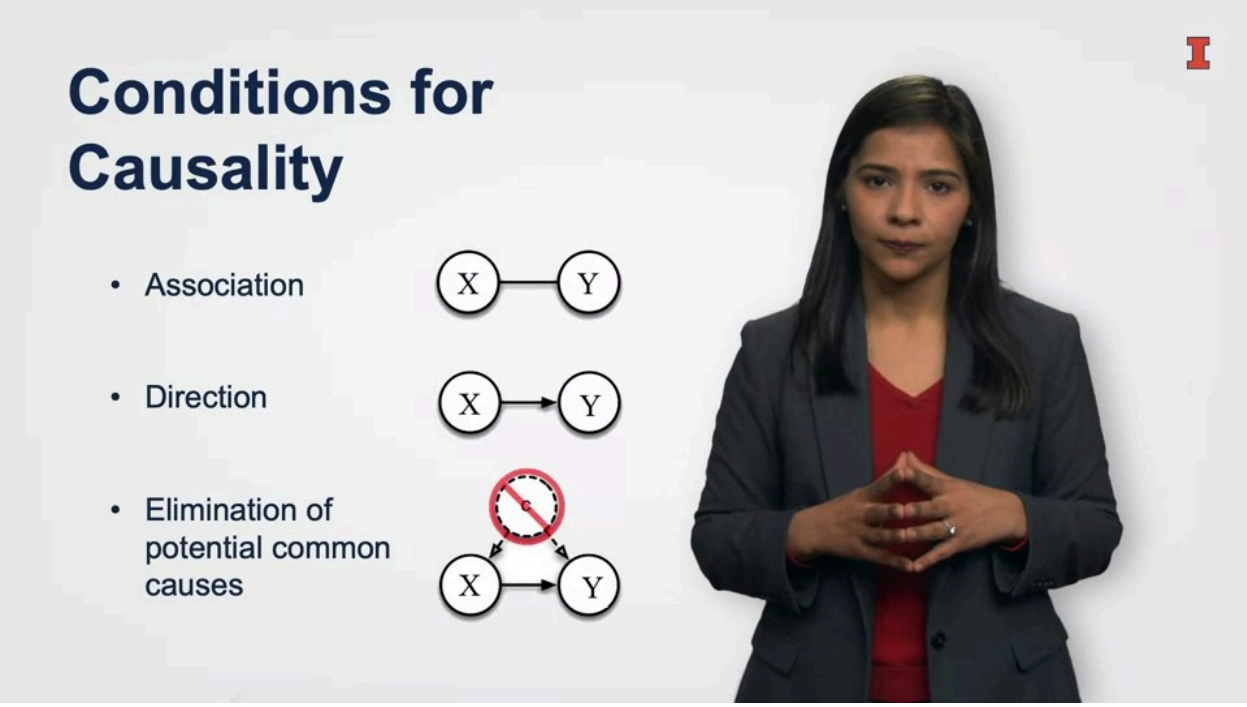
Causal Analysis: Overview



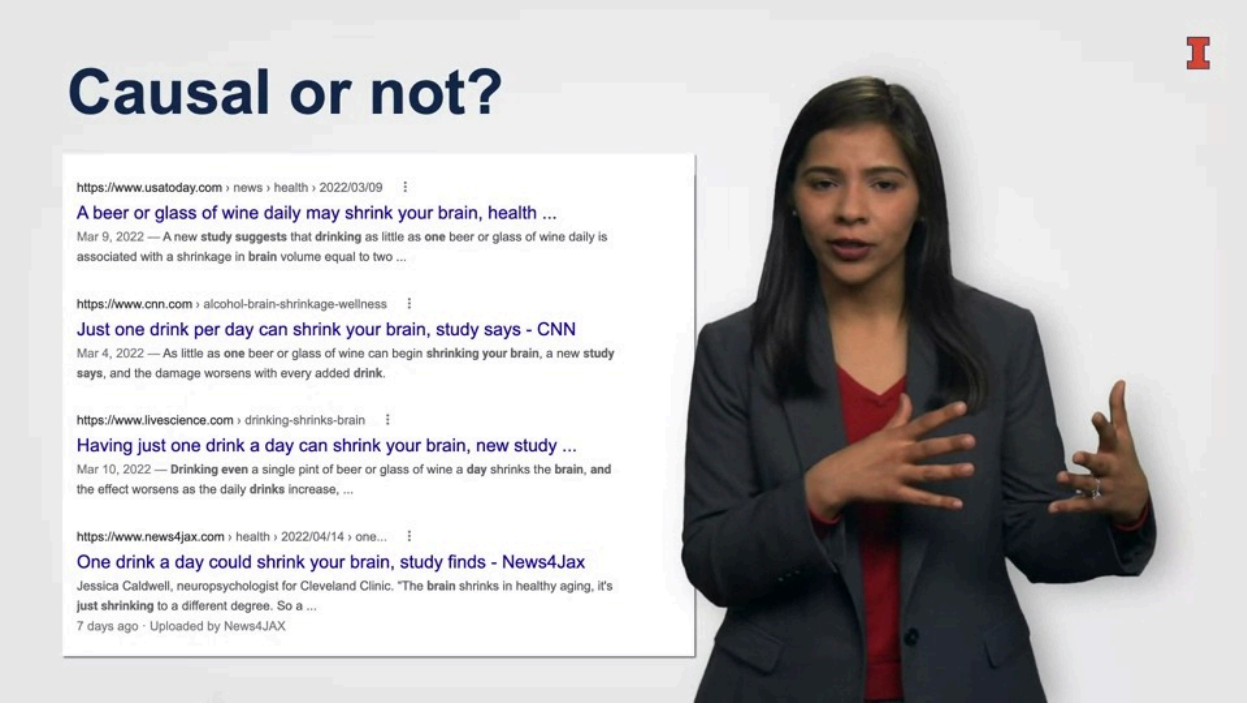
In this segment, we will explore the notion of causality. Causal analysis, as the name suggests, is the field of using experiments and observational data to help establish cause and effect. Causal analysis can be a powerful tool for marketers who are often concerned about causal questions. An e-commerce company may be interested to know the effect of their online ad campaigns on traffic and conversions on their website. A social media platform may want to estimate the effects of a new privacy regulation on its online advertising revenues. Similarly, a brick-and-mortar store may want to see if sending targeted mobile messages to consumers who are physically located near the store will get them to come in and spend more. Cause and effect is how we make sense of the world. In the 1940s, psychologist Albert Michotte theorized we see causality just as directly as we see color.

Way, way before this, in the 1700s, philosopher David Hume, however, believed that we can never truly recover causal relationships with certainty. Causality is simply a creation of our mind to explain repeated and constant conjunction between events that we call cause and that we call effect. Then, what really is causal and what isn't? Fortunately, we do have some answers now. This is because social scientists have been thinking about causal inference and ways to estimate causal effects for decades. Remarkable progress has been made and causal inference in just the last few years across multiple disciplines, most notably, in statistics and economics. In the 1920s, statistician Jerzy Neyman made some of the earliest and most noteworthy contributions to the field of causal inference. In his 1923 article, he discussed randomized experiments and proposed ways for testing hypotheses by repeated sampling. This discussion has formed the fundamental tenet for progress and causal inference. Donald Rubin and William Cochran, amongst others, developed this model into a more general framework for causal inference. In the 1970s, the idea of potential outcomes made a comeback. Since then, it has been used as a framework to understand causality and also to help get to causal inference in the absence of randomization. Many statisticians and economists developed quasi-experimental methods.

Paul Rosenbaum, Donald Rubin, introduced propensity score matching, a method that allows comparing similar treated and untreated units by matching them on observed data. Alberto Abadie and his co-authors proposed synthetic control-based methods. In parallel, computer scientists and Turing Award winner, Judea Pearl, introduced graph-based methods for causal inference that he had adapted for his work on artificial intelligence. In computer science, the idea of transfer learning and multi-arm bandits also made sequential experimentation more popular. Most recently, two books, Causal Inference by Scott Cunningham, and The Effect by Nick Huntington-Klein, have added to the rich discussion on causality, the research design for getting causal estimates and it's estimation in modern-day statistical software like R, which we will be using extensively in this course. Many of these developments and causal inference not only allow us to better design and analyze experiments, but to also recover causal relationships in the absence of randomization. At the forefront of some of these quasi-experimental work, our contemporary economists such as Joshua Angrist, Jorn-Steffen Pischke, David Card and Guido Imbens among others. In 2021, Angrist, Card and Imbens were awarded the Nobel Prize for their contribution to causality. They introduced the notion of natural experiments to identify causal effects by using chance events or policy shocks that leads some groups of people to be treated differently than others without any experimental manipulation. While randomization is considered the gold standard of causality, and we will discuss why this is, what's powerful is that in some cases, even without running an experiment, you may get close to causal estimates. Sounds magical, doesn't it? Well, takes a lot more than magic. Now that I've provided a 20,000 feet view of the causality literature, let's begin by taking a step back. Let us formalize when we can say there is a causal relationship between two variables, X and Y. We will draw from these advances to apply tools of causal analysis to marketing problems. For causal relationship to exist between an independent variable X, and a dependent variable y, three conditions must exist. First, two variables should be correlated. Second, X must have happened before Y.



Finally, the correlation between X and Y should not be driven by a third variable that causes the two under consideration. This is a simple textbook idea of causality. In real life, outside of a controlled lab setting, it is very difficult to achieve. Yet we're surrounded by statements that could suggest causality or be interpreted as causal. Here's one example. Drinking milk can make you win a Nobel Prize. Can you think of any reasons that could make this claim problematic? Let me give you a hint. Perhaps there is a third variable that's driving both milk consumption and winning the Nobel Prize. As it turns out, a lot of milk producing countries also tend to have very good education systems. Here's another fun one. A single drink can shrink your brain.

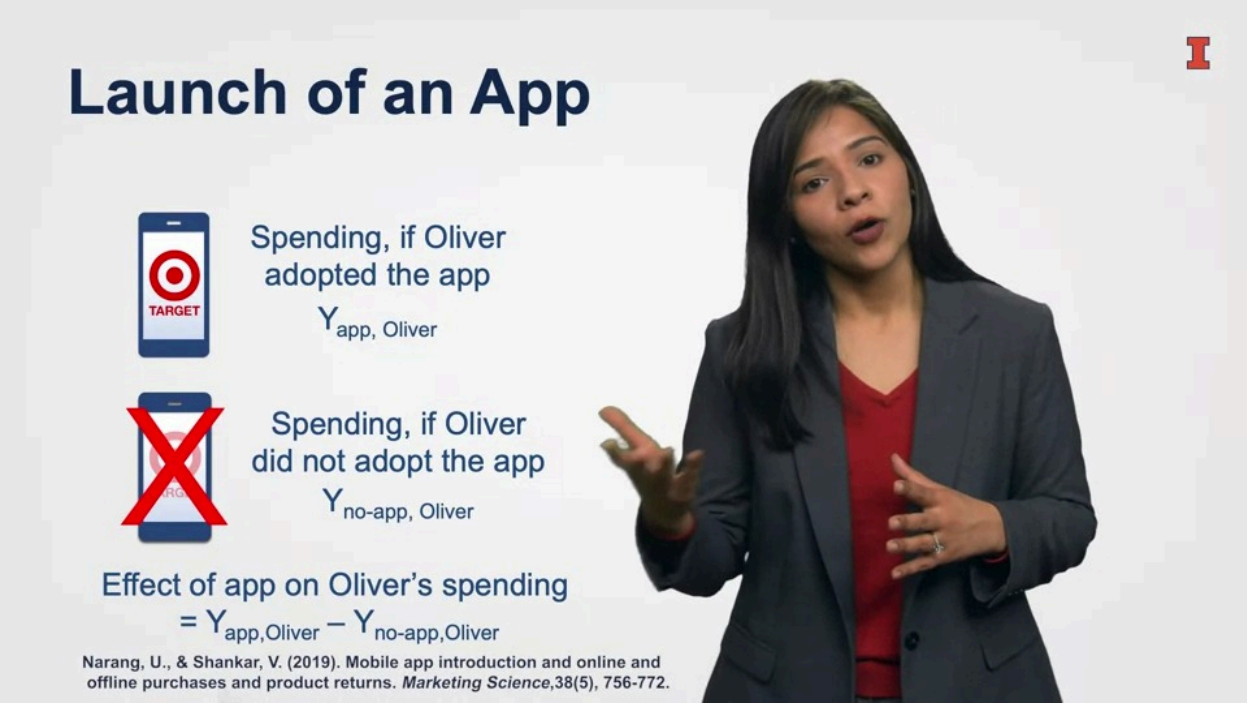


Does this mean having one drink each day makes your brains smaller? The researchers in this study found that those who drank the equivalent of just one daily glass of wine had slightly smaller brains on average than non-drinkers using MRI scans. Does it sound believable? Well, the challenge with a study like this is that brain scans were taken at a given snapshot in time so we cannot rule out pre-existing differences between the drinkers and the non-drinkers. Perhaps people with a smaller brains tended to drink more of a priority. Participants also self-reported their drinking, so due to these limitations, causal inference is not straightforward, and even if there was a causal relationship, the direction of causality is not clear.

Causal Analysis: Motivating Example and Key Concepts

Causal Analysis: Motivating Example and Key Concepts

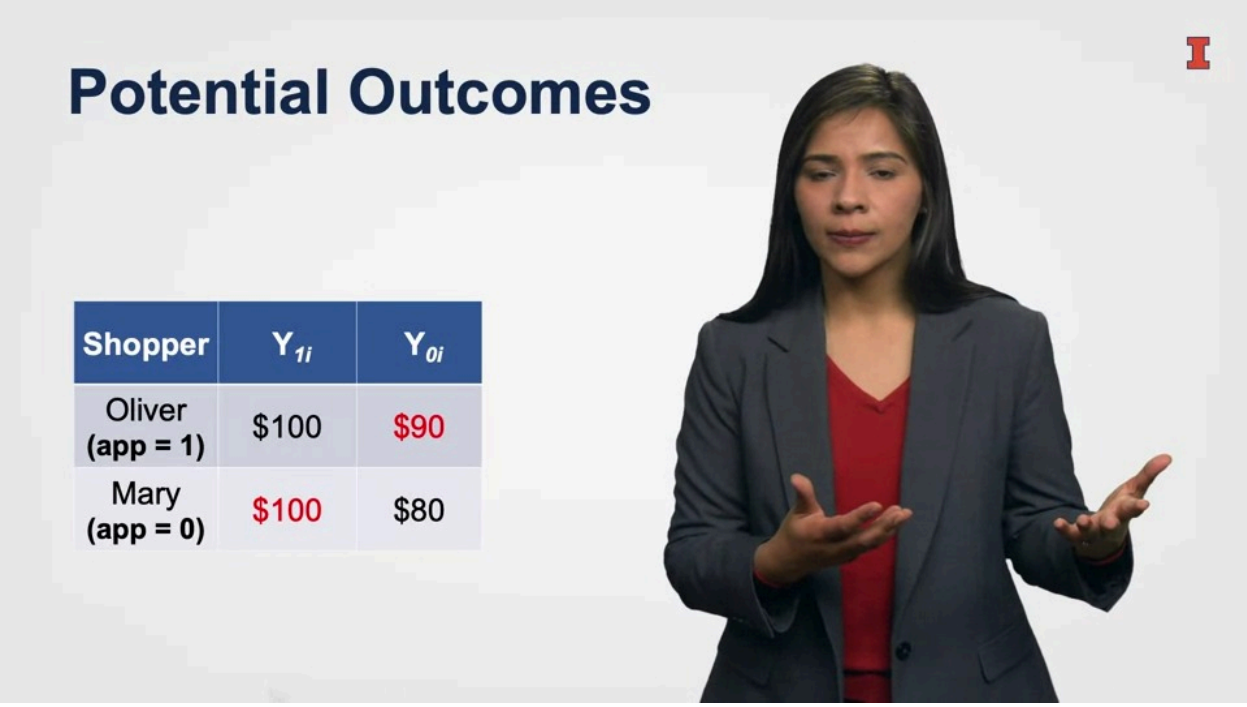
[MUSIC] What makes it difficult to identify causal relationships, in this segment, we will take an example from retailing to think about this question. Imagine that a retailer launches a new mobile app and wants to understand the effect of the app on its customers shopping in other channels. Now we know that mobile usage is growing rapidly. Americans today spent more than four hours each day on their mobile devices. Most of the time in mobile devices is spent in mobile apps, retail apps are among the fastest growing app categories. So in this new world of omnichannel retailing where consumers are always online and are surrounded by communication from brands in multiple channels including mobile apps, websites and physical stores.



How would the launch of a new mobile app affect consumers in a retailer's online and offline channels?

Play video starting at :1:10 and follow transcript1:10

Think of a multi channel retailer like Macy's or targets when they launch an app, it may allow shoppers to better find their nearest stores through the app or search for specific products and even deals that they can redeem in stores or online. So perhaps the app will facilitate more spending in other channels. At the same time, the app could also act as a substitute and reduce spending in other channels, assuming the retailer is able to track individual purchases in these various channels as well as consumers app use. How can we start to think about identifying the causal effect of the app? Remember, not every customer uses this new app, either a customer is an epidural doctor or they aren't. To get the causal effect of the app in the ideal world, we would want to compare the outcomes for an app adopter with what the outcomes would have been had this individual not adopted the app. Suppose as a consumer named Oliver, we can denote Oliver's spending if you adopted the app as Y app Oliver and Oliver spending if he did not adopt the app as why no app Oliver. If we observe the outcome under both these scenarios we would obtain the causal effect as the difference between the two. Unfortunately, we only observe one of the two outcomes in reality at any given point in time



Why is that? Well if Oliver adopts the app, we would only observe by app Oliver as his spending and we would never observe what his spending would have been had he not adopted the app. This is the potential outcome which was never realized. Let me introduce you to some more general notation. So you would notice that I'm using Y to denote spending which is the primary outcome of interest in this case, Yi is the spending of individual I depending on whether we are denoting outcomes under treatment or not. We can specify Yi as either Y1i or Y0i. In our case app adoption is the treatment. So the outcomes with app for individual, I will be Y1i and the outcomes without the app for individual i will be Y0i.

Play video starting at :3:39 and follow transcript3:39

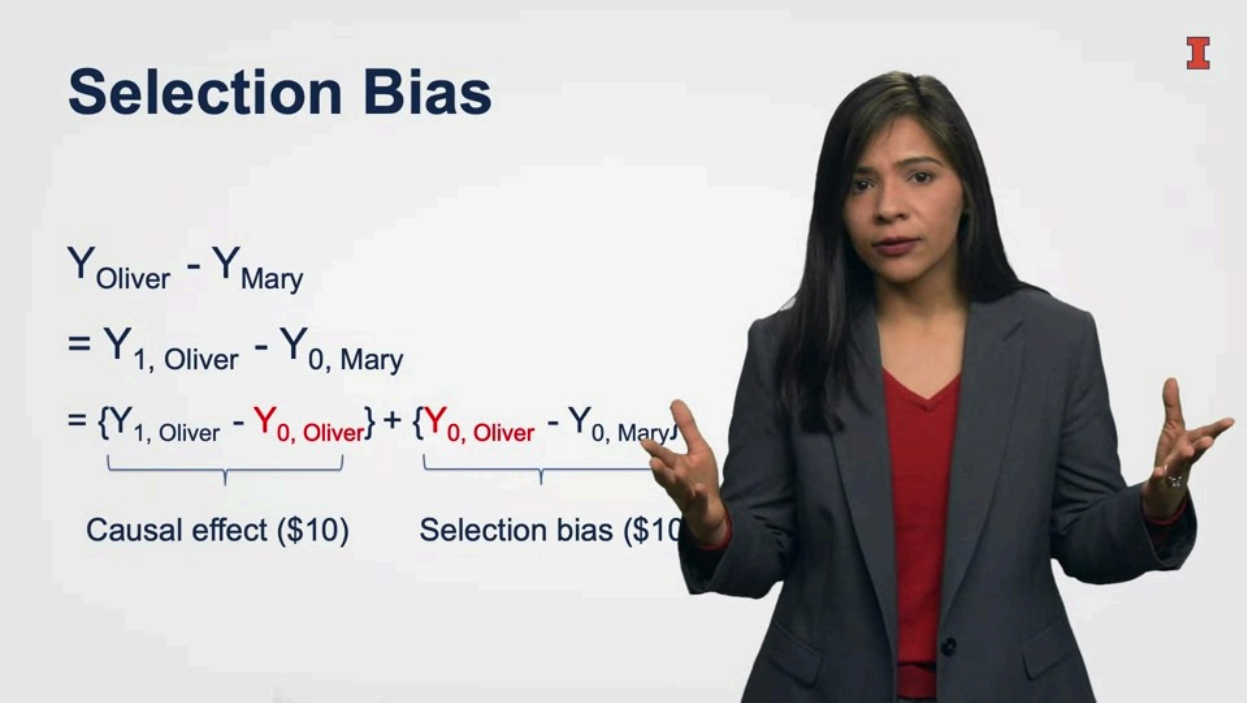
The causal effect can then be recovered as the outcomes under treatment equals 1 minus the outcomes under treatment equal 0.

Play video starting at :3:51 and follow transcript3:51

The fundamental problem of causality and why it is so difficult to get causal effects is that we observe only one of these for a given individual depending on their treatment status. If Oliver adopts the app, we would only observed by one eye for Oliver and never observed what would have happened had he not adopted the app. One way to think about this problem is through the length of the potential outcomes framework. This framework comes from Donald Rubens Classic paper in 1974 and is often the starting point of thinking about causal effects. Under the potential outcomes framework each shopper can have two potential outcomes based on their treatment status. If Oliver adopts the app, the potential outcome under treatment would be Y1i. Assume that this is $100 of spending for the sake of this example. If Oliver does not adopt the app, the potential outcome under no treatment would be Y0i. Assume this is $90, so the treatment effect should be the difference between the two which gives us $10 as our effective interest. In reality, remember we only observe one of the two. For example, if Oliver adopts the app, you would only see $100 as the outcome.

Play video starting at :5:18 and follow transcript5:18

Now imagine there's another shopper named Mary, again, we can have two potential outcomes for $100 if she adopts the app $80 if she does not adopt the app. In reality, suppose Mary does not adopt the app, we would only observe her outcome as $80. So assuming Oliver adopts the app and Mary does not. We would observe their actual outcomes as $100 and 80 dollars



We would not observe the potential outcomes highlighted in this table, based on what we observe in reality, what if we compared the treatment outcomes for Oliver who is an epic doctor with the untreated outcomes for Mary who is a non adopter? Would this give us the causal effects of the app?
Play video starting at :6:8 and follow transcript6:08
In other words, could we simply compare app adopters with non adopters and conclude that any differences in their shopping behavior are due to the app? What is wrong with such a comparison? Well, it's easy to see why this comparison should not be taken at face value. People who adopt the app are probably more inclined to make purchases with the retailer to begin with. So this comparison will likely overestimate the effects, recall that in our example, the true treatment effect was $10. This comparison is overstating the treatment effect to be twice as big. Let's try to understand why this happens.
Play video starting at :6:54 and follow transcript6:54
Let's use a simple math exercise to reveal a powerful insight. The difference of the outcomes we observe for Oliver and Mary or the treated outcome for Oliver and the untreated outcome for Mary, can be rewritten as a sum of two sub parts. The first part is a true treatment effect we desire which is the difference between Y1 Oliver and Y0 Oliver. And the second part is the difference between the untreated outcomes for Oliver and Mary. The second term is the source of bias and it is often termed as the selection bias, in our example because app adopters are more likely than non adopters to adopt the app, those who adopt the app likely have more spending. This makes the selection effect positive. In some cases, it can be negative and it can even flip the sign of the true treatment effect.

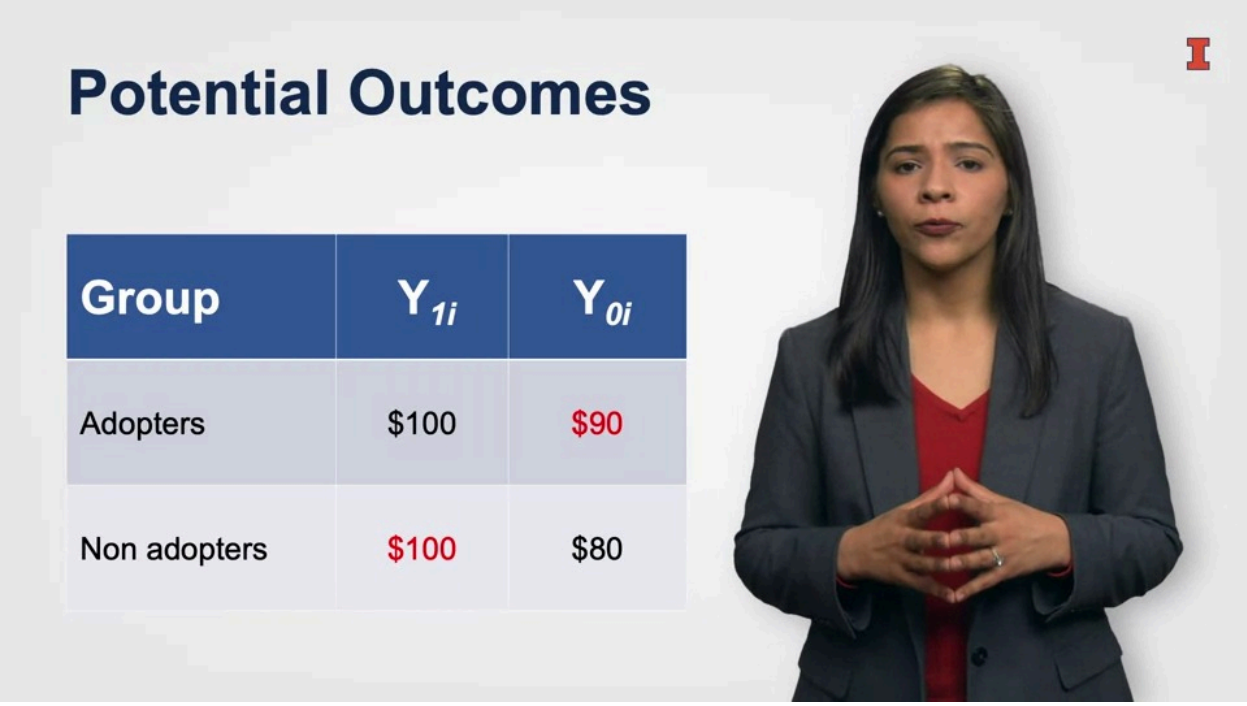
.

Through this example we learned about the fundamental problem of causality which is that all the potential outcomes are never observed. In reality the outcomes we do observe are based on the treatment status and this treatment status may be driven by selection bias. In the next segment, we will learn about various approaches to mitigate selection bias and get to causal estimates. We will discuss the power of randomization and what to do in the absence of a randomized data generating process.

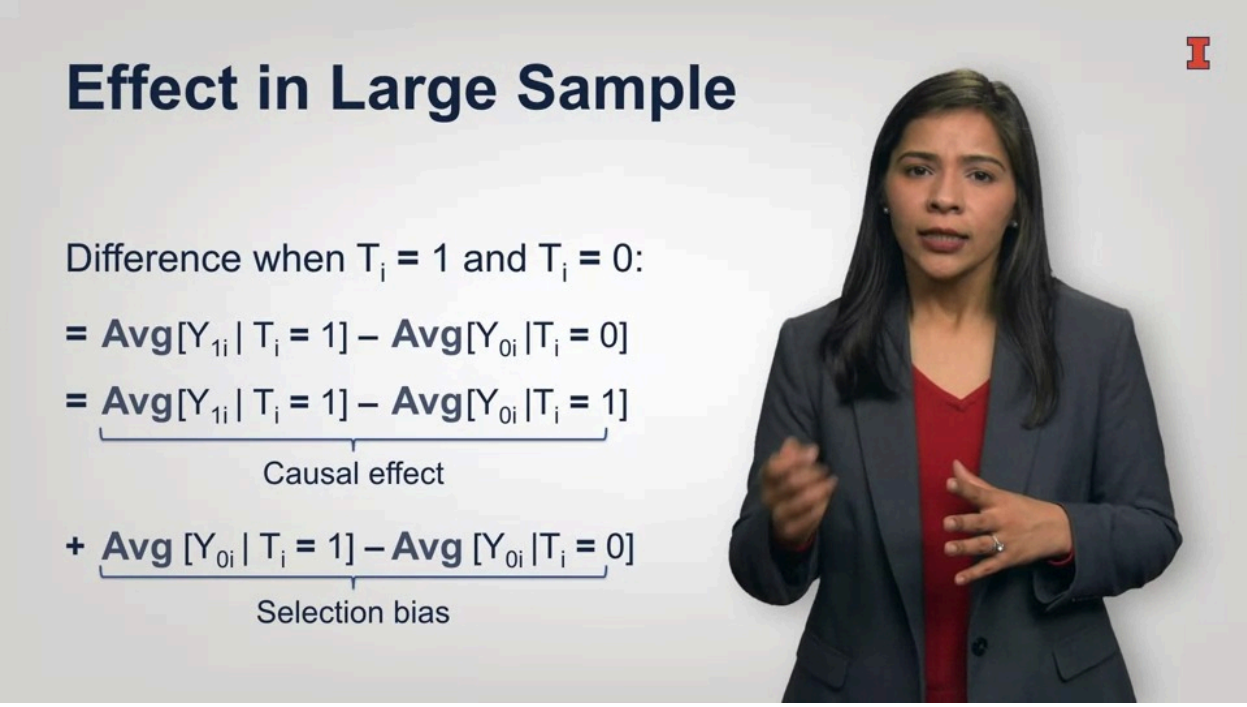
Causal Analysis: Randomized Experiments

Causal Analysis: Randomized Experiments

[MUSIC] So far we've discussed why comparing the observed outcomes for the treated and the untreated individuals can give us biased causal effects. We've seen this in the simple case of comparing one individual's outcomes with another. Can we solve this problem by taking larger samples? If we used data on large groups of consumers that are treated and not treated, does the selection bias go away? Turns out that it doesn't. We need something more than large samples. In this segment, I will introduce the concept of randomization.



Let's assume that the hypothetical outcomes I showed you were average outcomes for a group of adopters and a group of non adopters instead just for Oliver and Mary. We observed the treated outcome Y1i for app adaptors and we observe the untreated outcome Y0i for non adopters. We can extend our simple math for one individual to groups of individuals by averaging over individuals in each group. By taking averages over the treated and untreated groups of individuals, we can show that the selection bias still exists if we simply compared the observed outcomes for the treated and the observed outcomes for the untreated groups. As we noticed before in the two individual example, this naive comparison of observed outcomes, even if group averages, does not give us the causal effect were looking for. This comparison gives us an estimate that includes the selection bias term.



The selection term is the potential untreated outcome had the treated not been treated minus the observed outcome for the untreated. In case of groups, you can think of the selection term as how much the outcomes of the treated and untreated groups differ on average in the absence of treatment. In our app example, this is the same as asking would the app adopters and non adopters have differed in their spending even in the absence of the app.

Play video starting at :2:28 and follow transcript2:28

If this election term is non 0, then that means that even if there was no app adoption, these groups would have differed in their outcomes. So we cannot truly attribute this difference to the app.

Play video starting at :2:43 and follow transcript2:43

Now that I've provided some intuition for the selection term, as a quick aside, I want to point out that these averages can also be written as conditional expectations of outcome Y given treatment equals one or zero. You can think of this expectation as the value the outcomes would take on average over a large number of occurrences given that a certain set of conditions is known to occur. In this case, we have two conditions treatment and no treatment, so we take conditional expectations over these.

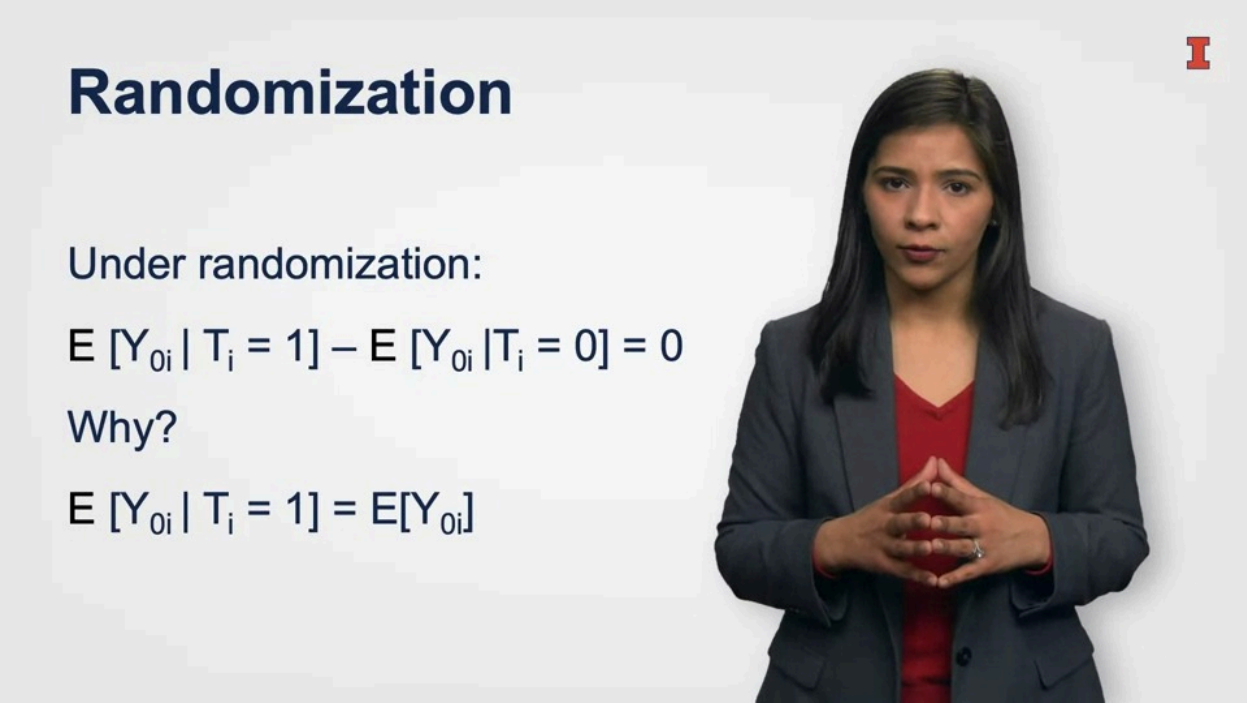
Play video starting at :3:22 and follow transcript3:22

So how can we mitigate the selection bias? Many of you may be familiar with the idea of running experiments, either in your jobs or at some point in your education, you may even have run experiments.

Play video starting at :3:36 and follow transcript3:36

Companies today run thousands of real time AB tests. Let's discuss how randomization helps us get to causality. In experimental design, randomization is the idea of randomly assigning participants to the treatment conditions. In our app example, if we could randomly give some customers access to the app and not to some others, we would randomize app assignment to them. When treatment is assigned randomly and samples are sufficiently large, selection bias disappears. This is a powerful idea. To understand why randomization works in large samples.

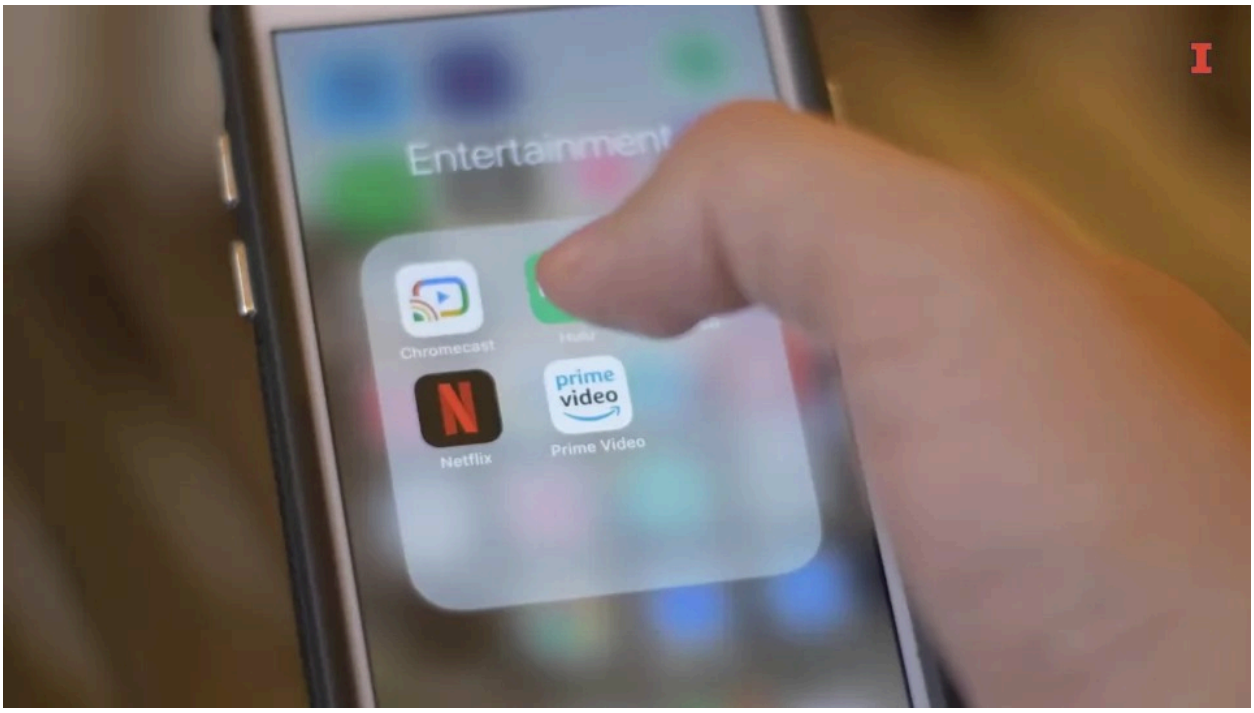
Imagine that you tossed a coin. We know that a coin toss should give us heads with 50% chance and tails with 50% chance. So let me flip a coin twice and see if this textbook principle works.
Play video starting at :4:33 and follow transcript4:33
Here we go.
Play video starting at :4:37 and follow transcript4:37
I get heads on my first toss, so I should be getting tails now, right? Well let's see.
Play video starting at :4:47 and follow transcript4:47
It's heads again. So what happened? Did something go wrong? Is there something wrong with this coin? Not really. The principle of randomization would work over many, many, many coin tosses. If I kept going over a large number of coin tosses, I will start to observe that the average likelihood of getting heads will be close to 50%.
Play video starting at :5:13 and follow transcript5:13
So we understand now why we need large samples to make randomization work. But how does randomization solve selection bias? Let's look at the map for a quick insight. We know that the selection bias comes from the second term which is the difference between the untreated outcomes when treatment is one and when treatment is zero.



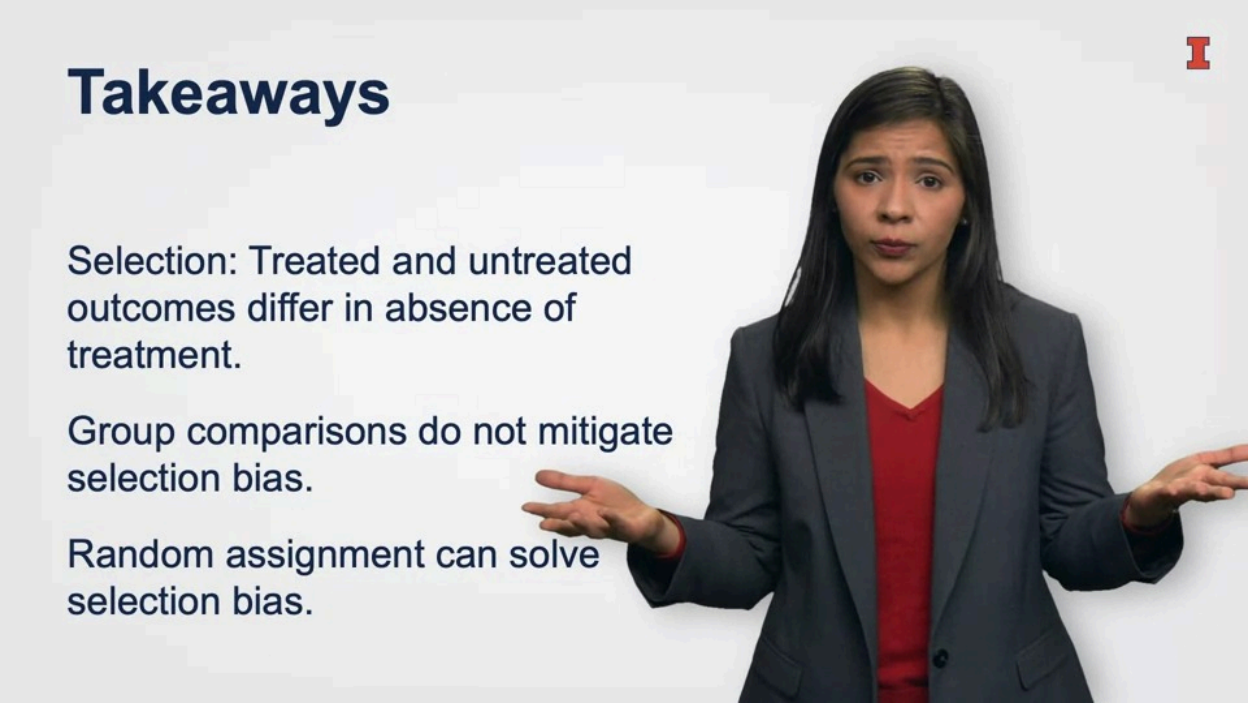
If we were to randomize over many, many individuals under both these conditional expectations, we would get the unconditional expectation of the untreated outcome. This is free of which conditional group you're assigned to. Since the two terms are equivalent under random assignment, the selection bias is reduced to zero.

Play video starting at :5:56 and follow transcript5:56

Now you may be wondering, this all sounds cool we believe you, but how often do companies actually do this kind of randomization in the real world to make decisions? What if the cost to randomize is too high? Or what if you simply cannot randomize or what if you simply don't want to randomize? If the retailer in our example created really, really good app, they'd probably want to roll it out to everyone, right, and not keep it for a subset of their customers. Randomized trials are not yet as common in social sciences like marketing as they are in medicine where they have been used extensively for clinical trials and for other biological experiments. But, they are becoming much more popular with the rise of the internet. Internet companies like Amazon and Netflix run thousands of randomized trials. Netflix for example, tested the top 10 shows feature to help members discover popular shows.



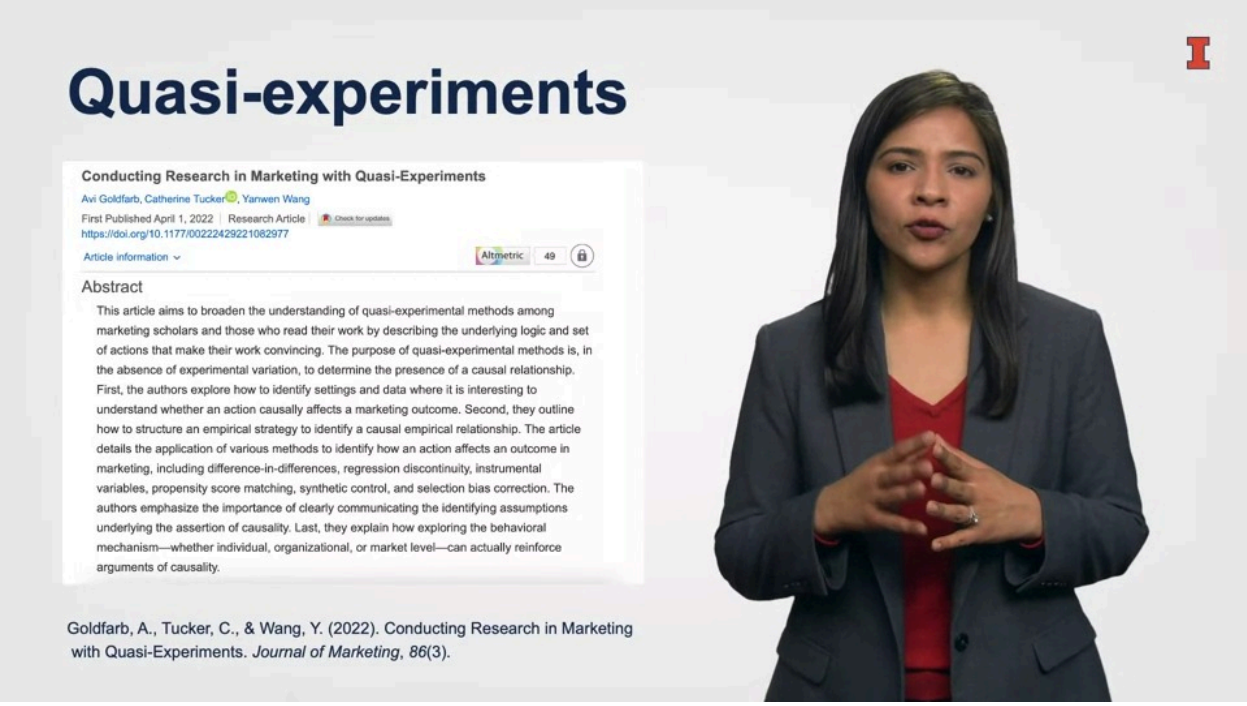
They also tested the random show generator to help choose a show for viewers if they can't decide what to watch. With the digital revolution, the cost of conducting large number of randomized trials real time on the internet is much lower. So we have seen an increasing number of field experiments in areas like online advertising, search and e-commerce. However, there are many important settings in which randomization is not possible. There are many big questions which often cannot be answered with randomization. How do you randomize locations where the retailer opens a physical store if you wanted to understand the effect of a new store opening? In this segment we discussed that comparing observed outcomes for treated and untreated groups in the absence of randomization does not give us causal estimates due to the presence of selection bias. Selection bias does not go away by simply using larger samples. Random assignment is a powerful tool that can eliminate selection bias.



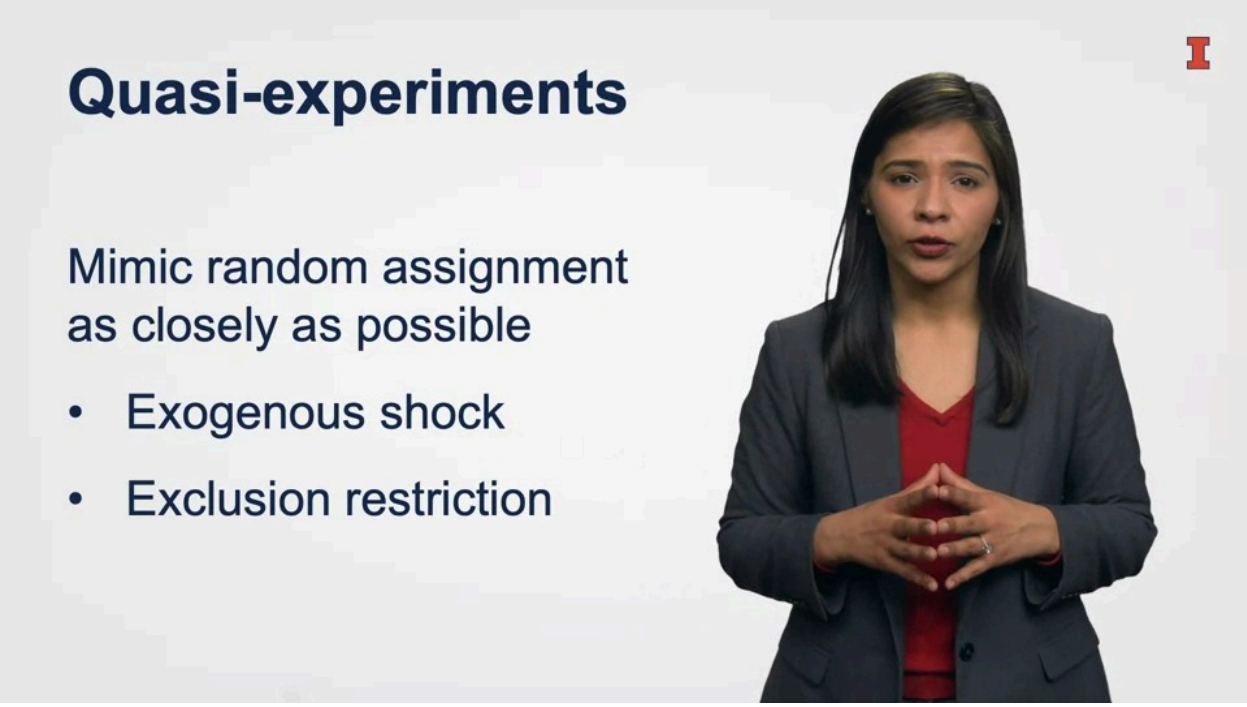
However, randomization is not always achievable. In the next segment we will discuss observational studies in which there is no manipulation of the treatment, but we may find some quasi randomness that can help us get to causality
Play video starting at :8:28 and follow transcript8:28
[MUSIC]

Causal Analysis: Observational Approaches

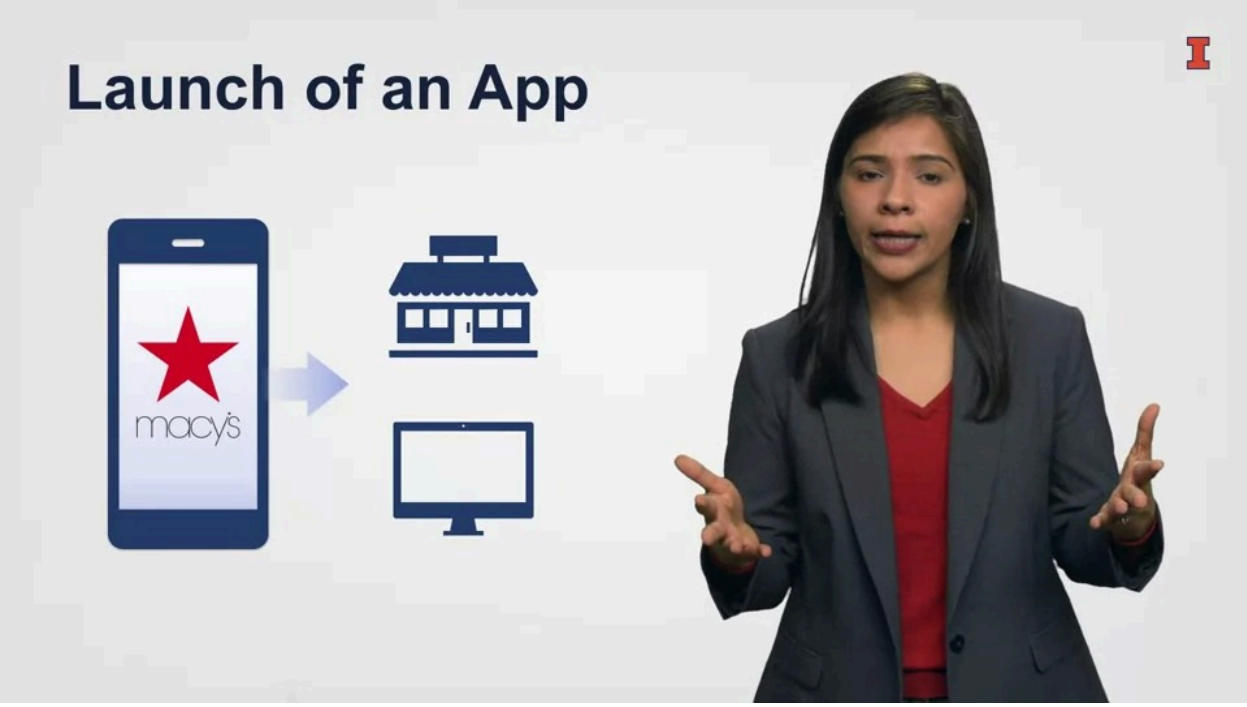
Causal Analysis: Observational Approaches



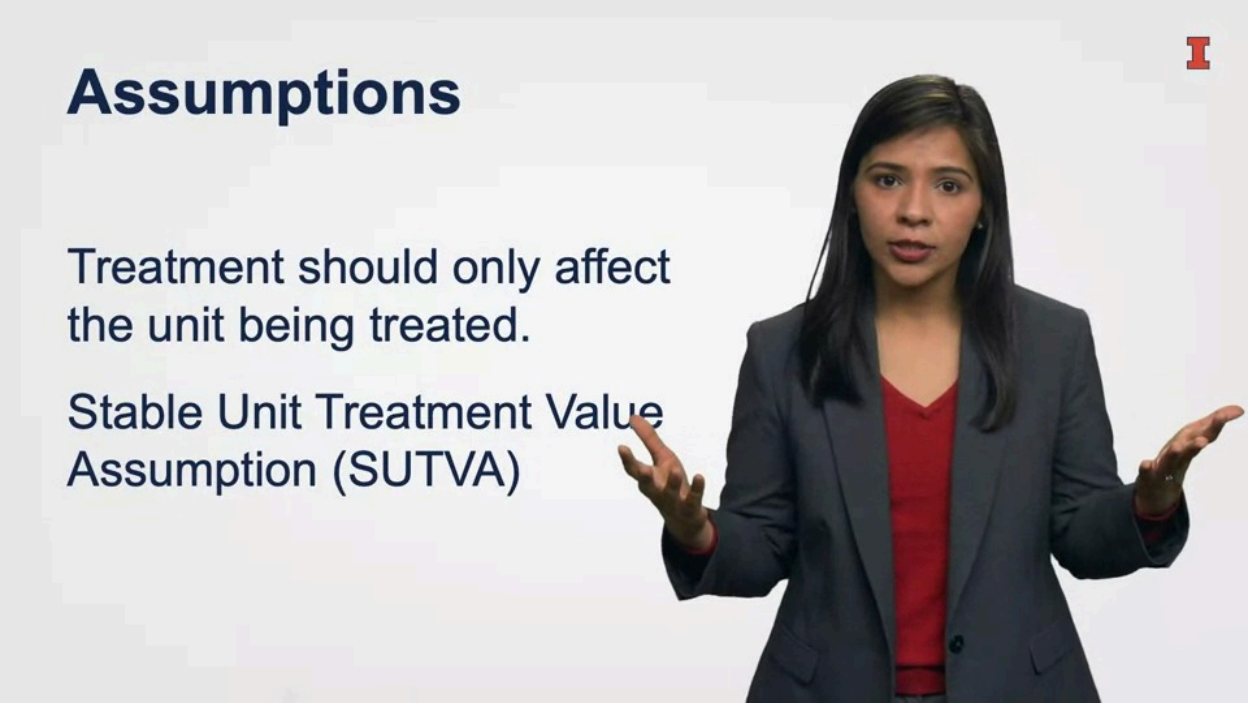
Even though random assignment is considered the goal standard and causal inference, it is sometimes not feasible or appropriate in practice. In these situations, it may be possible to arrive at causal or close to causal estimates using quasi-experiments. Quasi-experiments mimic random assignment as closely as possible using observational data. These data, unlike experimental data, are not generated by manipulating the treatment of interest. Instead, they occur naturally in the world around us and require clever and clear identification. To discuss quasi-experiments, I will draw from a paper from the Journal of Marketing and the various books and references in this paper. How can we identify quasi-experiments? Such settings exist when there is exogenous variation that leads one group to take up the treatment and not others without affecting their outcomes directly other than through the treatment. Exogenous variation can come from different sources. For example, it can come from events like changes across geographical boundaries, contract changes, shifts in foreign policy, individual level life changes, and regulatory changes. For an event to be truly exogenous, it should meet the exclusion restriction criteria. Under this restriction, the exogenous event should not affect the outcome in any way except through the treatment. We often need the event to induce treatment but not affect the outcomes directly.



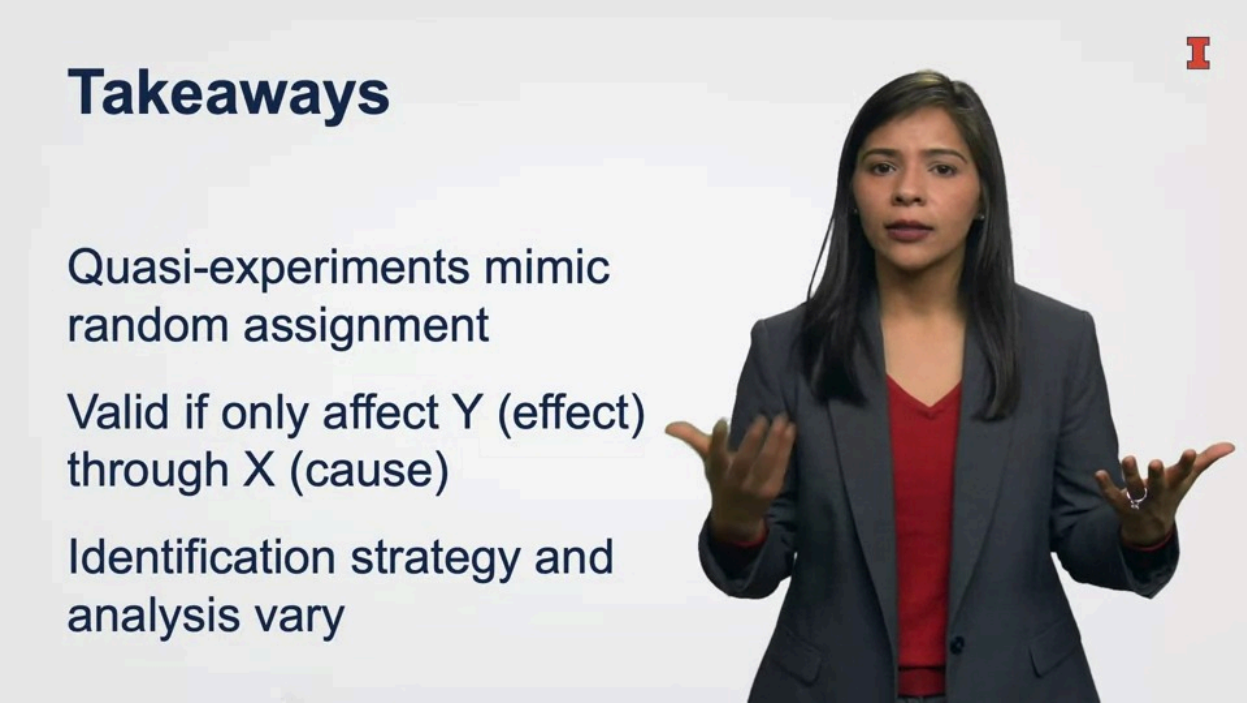
This is not trivial. Finally, if the source of variation cannot be argued to be exogenous, we can still make some valid comparisons by comparing treatment groups with only those control units that are similar to the treated group on average. A class of matching methods such as propensity score matching and synthetic control allow us to do this. To understand these concepts, let's continue with our app adoption example. A retailer launches an app but does not want to randomize the app availability, so they release it to all the customers at once. Some customers adopt the app and some others don't. We know that customers who adopt the app likely prefer the retailer and will buy more anyway. We cannot do a naive comparison of app adopters with non-adopters. In our app example, a quasi-experimental approach would mean finding something that gets people to adopt the app more but not affect their spending directly.



In the paper, which was part of my dissertation research, I showed that a potential source of such variation is the number of cell towers. Due to better connectivity, higher number of cell towers in a zip code may actually induce app adoption, but may not necessarily drive more spending, particularly if more purchases take place in physical stores. In another paper, the author study a social media type platform for windsurfers. They wanted to understand if there was any rule of content creation on forming friendships on social media for windsurfing. They cleverly used windspeed as a way to provide exogenous variation for content creation. Windspeed should not directly affect friendships except through more content creation on this social media platform. The key in any quasi-experimental work is a careful evaluation of the assumptions that go into identifying the causal effect of interest. This is tricky because there is really no formal test of whether a variable meets the exclusion restriction criteria or not. We have to make some untestable causal assumptions and look at the empirical support, and the institutional setting to verify if these assumptions have some face validity. For this reason, most quasi-experimental work using observational studies will warrant a fair amount of caution in making causal claims. In the windsurfing example, for instance, it is possible that windy days with high windspeed could directly affect friendship formation on the platform because windier days attract more people to the surf locations and these people likely become online friends later on. How can we check this from the data? One evidence that inspires confidence is that ties of friendships that form online don't reflect geography. It's unlikely that these people met on surf locations before becoming friends online. In the app launch example, it is possible that cell tower scheme and response to increasing demand and urbanization, so the direction of causality may be less clear.



We tested this in the data and found that most cell towers were already in place before our data period, and also there was little correlation between population economic growth and other indicators of economic development with cell towers during our data period. Finally, we also matched individual shoppers on their observed characteristics like demographics, pass spending to make sure we were comparing similar app adopters and non-adopters. Matching methods like propensity score matching allow us to do this by creating a single propensity score based on many different underlying observed characteristics. Ultimately, through these various checks, we showed that the launch of an app results in almost 36 percent more spending across the retailers' channels over a period of three years. Surprisingly, we also discovered higher product returns from app adopters. This result was pretty useful for the retailer to try to reduce returns from app adopters. Another assumption for the causal estimates to be valid is that the treatment should only affect the unit being treated and not the control units. This is also called the SUTVA or Stable Unit Treatment Value Assumption. Under SUTVA, treatment of unit I affects only the outcomes of unit I.

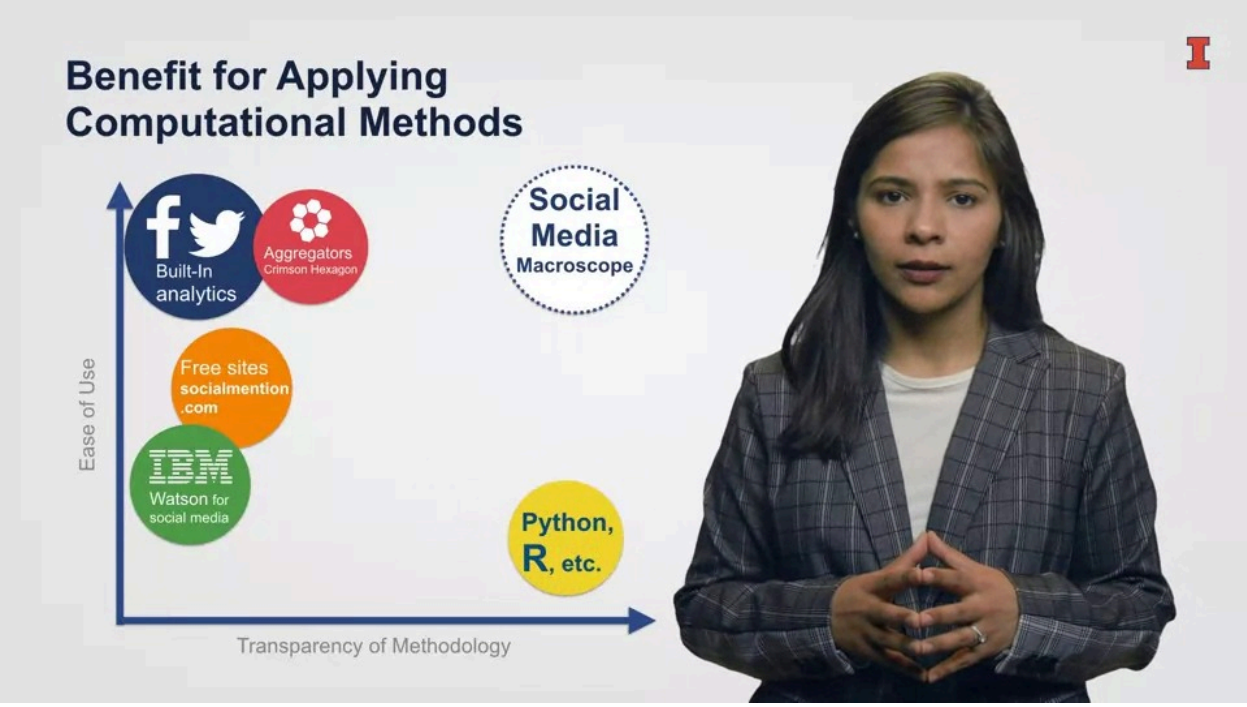


Again, this is often difficult to defend if there are spillovers or introductions between treatment and control groups. Overall, quasi-experiments can be a useful tool for learning about causal effects in the absence of randomization, however, they do warrant a thorough and careful discussion of these assumptions. In this segment, we discussed quasi-experiments, we talked about when is an event considered to be exogenous, what is an exogenous shock, and what does it mean for the exogenous events to only affect outcomes through the treatment. In other words, what is an exclusion restriction and why do we need it? If you're worried that the variation is not exogenous, how can we still make valid comparisons between the treated and the control groups by matching them on some characteristics? Hopefully, this gives you a broad overview of causality in the absence of randomization. Ultimately, observation methods may not work as well as randomized trials but they can still help us address really important questions for which experiments are often infeasible.

Introduction to the Social Media Macroscope (SMM)

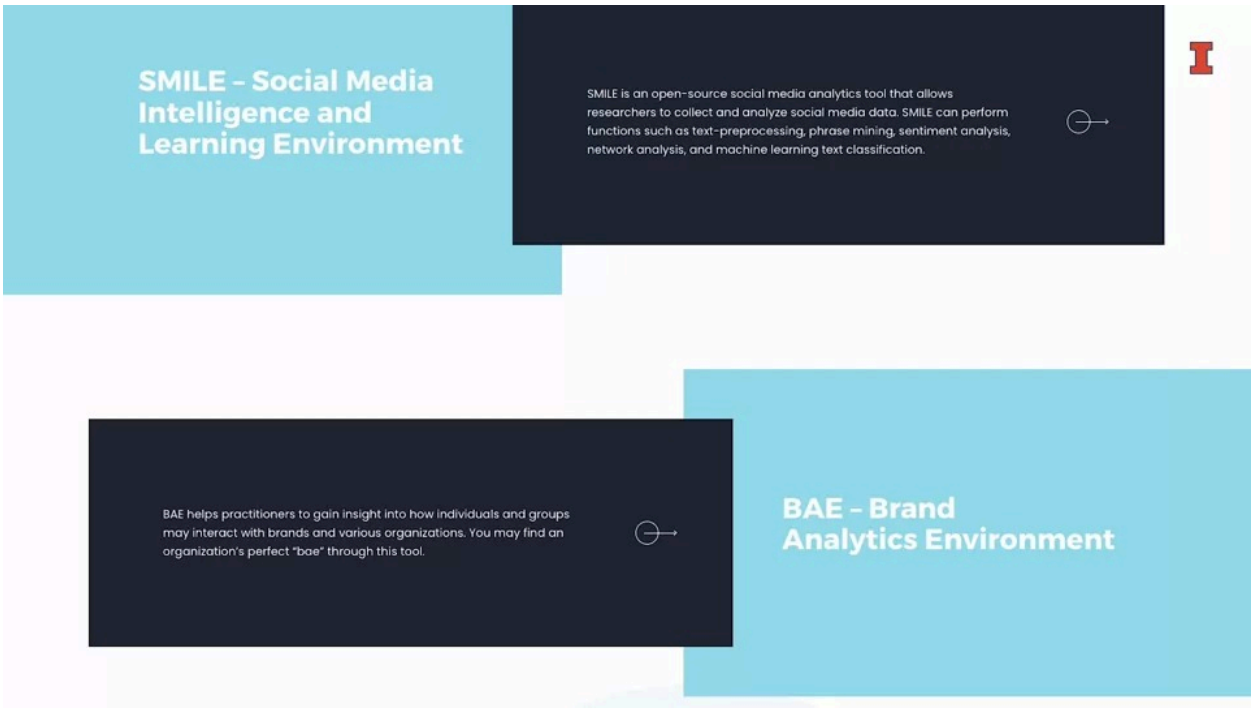
Introduction to the Social Media Macroscope (SMM)

[MUSIC] In this video, I will introduce you to a new tool called the social media microscope or the SMM. This tool was created by Joseph Yun and his team here at the University of Illinois at Urbana, Champaign. The idea behind the social media microscope is to assist those who don't necessarily have programming backgrounds to be able to conduct various data science analyses on social data. More formally, the SMM is a science gateway with the goal of making social media data analytics and visualization accessible to researchers and students of all levels of expertise.



We will be using this tool throughout our course to demonstrate various analysis. Here is our attempt at creating one picture that explains the SMM. The world creates social data and the microscope is in between a telescope and a microscope that allows people to do research. A little blurb here says an open social analytics science gateway for research. The word open is very important. This project is not only open source but everything in the SMM, all the methods and algorithms that we use, none of them are black box. They're all explained and they all have papers behind them. Now I want to explain this via a graph as to why the microscope is so unique. This is a two dimensional graph that explains how the microscope benefits individuals who are focused on applying computational methods to research questions using social data but who may not necessarily want to deal with programming or who may not even have a programming background. The two critical criteria that are important to people in this realm are ease of use and transparency of methodology. So what do I mean by that? Let's look at some of the tools that are mentioned here.

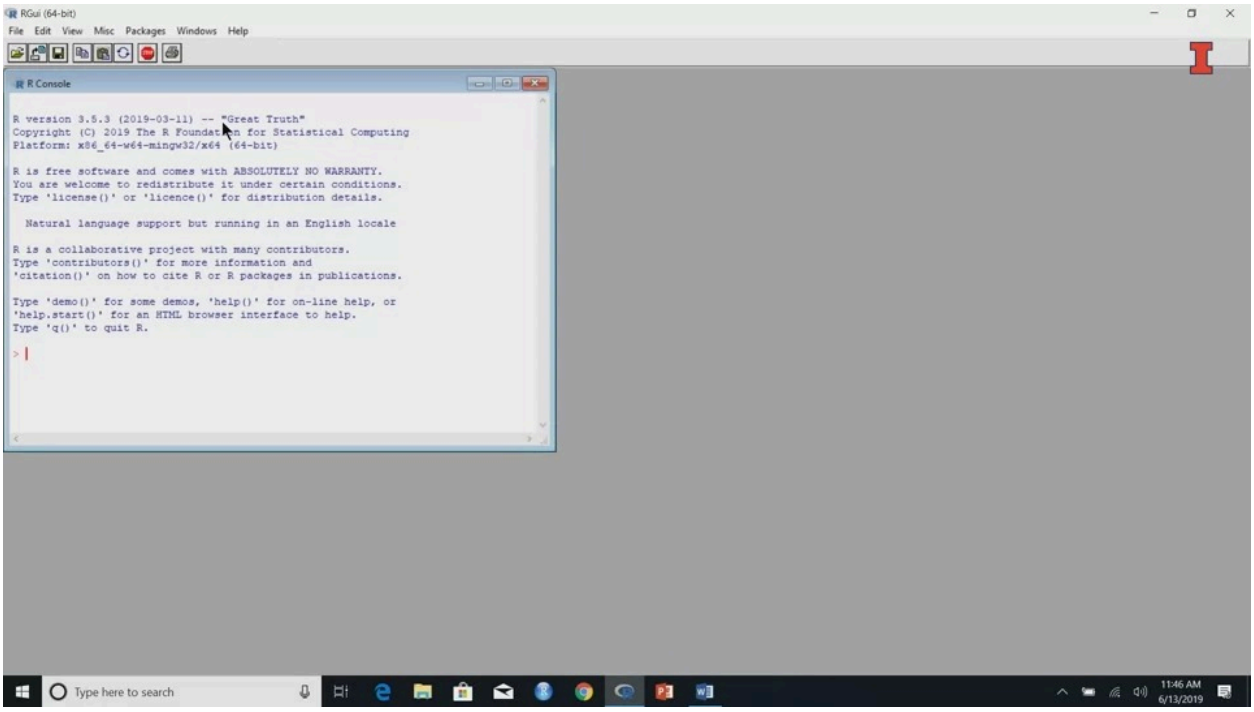
Facebook and twitter have their in built analytics and they're super easy to use. Anyone can just fire up twitter analytics and just start clicking on buttons and there you go. You can use it. But the problem is it's not that transparent in its methodology. So if twitter uses some sort of sentiment analysis, they're not going to explain to you how that sentiment analysis works. Which is really important for data science on the other end of the spectrum. You could also fire up your python terminal or your art studio and you can start to use various packages there and do sentiment analysis or any other type of analysis. And that's very transparent in its methodology but it's not necessarily straightforward and easy to use for everyone. So the S. Mm attempts to be in the top right quadrant of, easy to use as well as transparent in its methodology. While the initial goal for S mm was to help those with non coding backgrounds, they're also computer scientists that participate in this project. You may ask well why don't they need the S mm. Since they are natively inherently programmers and that's what they're already comfortable with. The microscope is also beneficial for those with computational backgrounds because it gives them a breadth of social media data access as we get more and more researchers that are working within the platform.



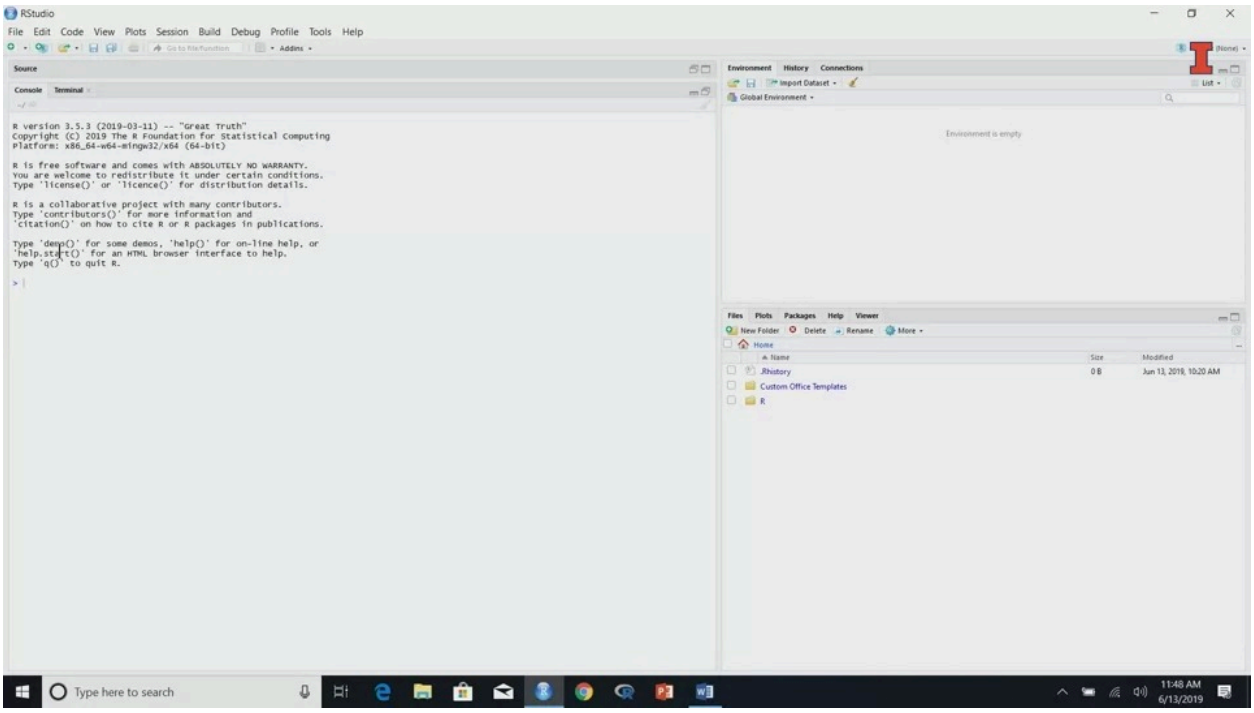
They may start putting out some of their data sets if they're allowed to into the microscope. This enables other computational researchers to have access to more data as well as the ability to test in house customized models. There's many times where in the course we will just use the microscope to do certain methods because it's easier and faster. Even though I do know how to program it myself, you can find more information about the social media microscope on this website, www. Social media microscope dot com as well as look at the running instance of this open source software at social media microscope dot org. What I mean by a running instance is that we've actually installed the software on cloud computing and then presented it as a web portal. Now, one last note with regard to the microscope is that it might sound like the microscope itself is a tool to analyze social data if you're thinking that you're kind of right, but kind of off because I guess I haven't explained it well enough. Let me better explain using a metaphor here. If you think about the app store on your phone, the app store on your phone is an app. Absolutely. But it's not necessarily an app where you run things. It's an app where you go and look for other apps. So the app store is like a container for all the actual apps that you can use. Similarly, the social media microscope or the S mm is like a container that contains the actual apps or as we call them, tools to analyze social data and it is an open source platform. What we're trying to do is encourage other developers to create their own open source social data analytics tools to contribute to quote and quote our app store our microscope. The goal eventually is to have over time more and more tools that are being hosted by the S. Mm. Throughout this course, we will give you a live demonstration of various analysis using S. Mm. We will also be using our in our studio which is an open source software you can download and install on your own systems in the next few videos. We will go through a basic tutorial of our to make sure you're all set up and ready to dive deeper into applying data analytics in marketing. Yeah.

Primer on R and Rstudio

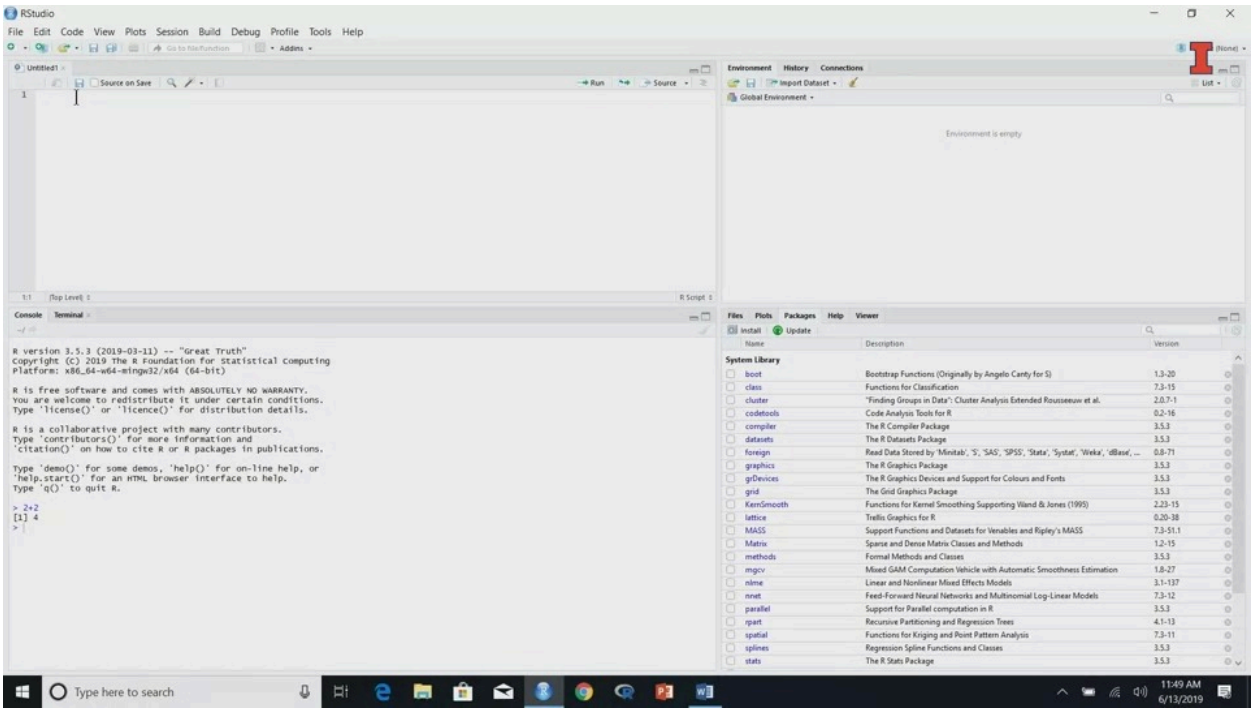
Primer on R and Rstudio



By now you have hopefully installed R and RStudio. You'll see two icons that look like this, R and RStudio. First I'd like you to follow along and click on the R icon. This is probably the only time you'll ever click on this. But I want to show you what it is. This is the original version of native R. Here you can see some introductory messages. The version of R you are running. The programmers like to give it a great name for each version.



This version has to be called the Great Truth. A little bit about the licensing and some simple commands for demonstrations and help options. Here you see that little greater than sign, this is where you would type in a command. Let's type in something real quick. 2 plus 2 and you get an answer of 4. You enter a command and a response will come back. I'm going to close this now this is the last time I want you to open up native R because we'll be using RStudio, Save workspace image. No, I don't really want to save that. Let's click on RStudio and open that up. The first time you open it, let me maximize my window. You should see on your screen something that looks like this. You should see four quadrants. If you see three, you can click on this window icon here. Now you see four quadrants here, 1, 2, 3, 4. I can adjust the size of these paints by grabbing in the middle, I can go left and right. That's how you would customize your environment. I'm going to maximize the window on the bottom left. Here you can see the same introductory message. Before we open up our native R and we got this introductory message, and now it's embedded into RStudio in this quadrant called the Console. There it is. Here you can see the arrow, the greater than sign



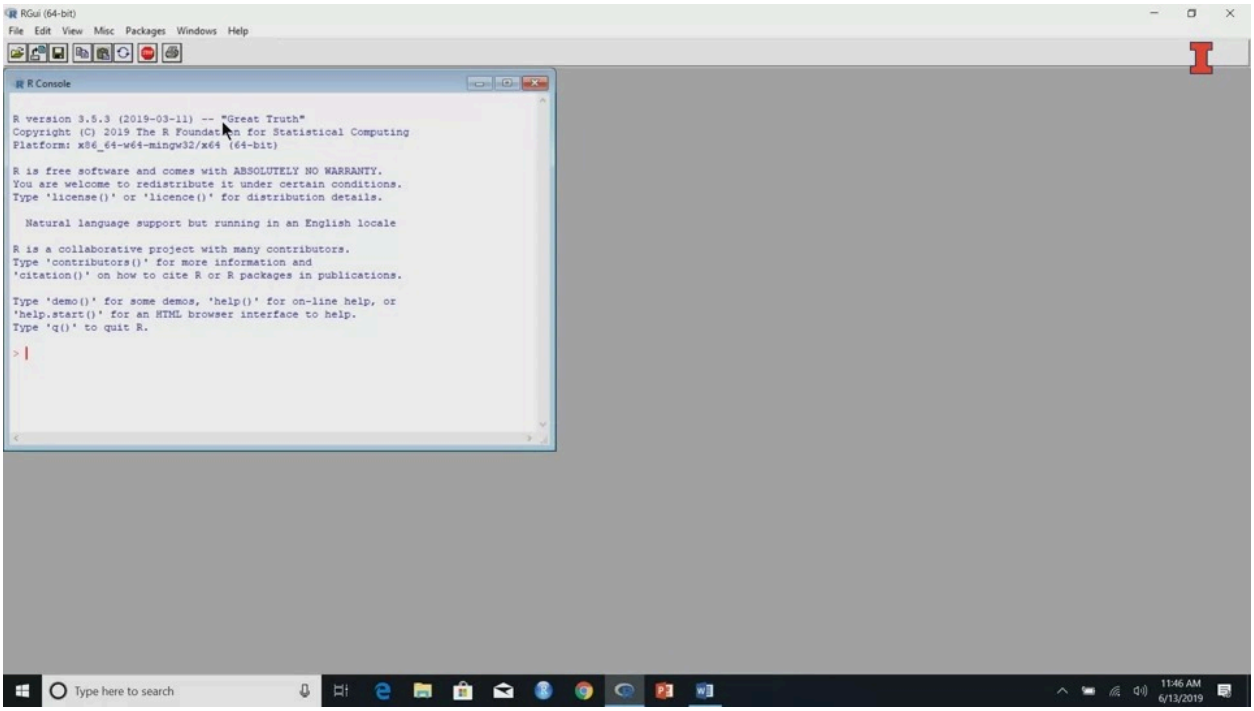
Again, I can enter a command 2 plus 2, and I get a response of 4. Let's go to the other quadrants real quick.

Play video starting at :2:47 and follow transcript2:47

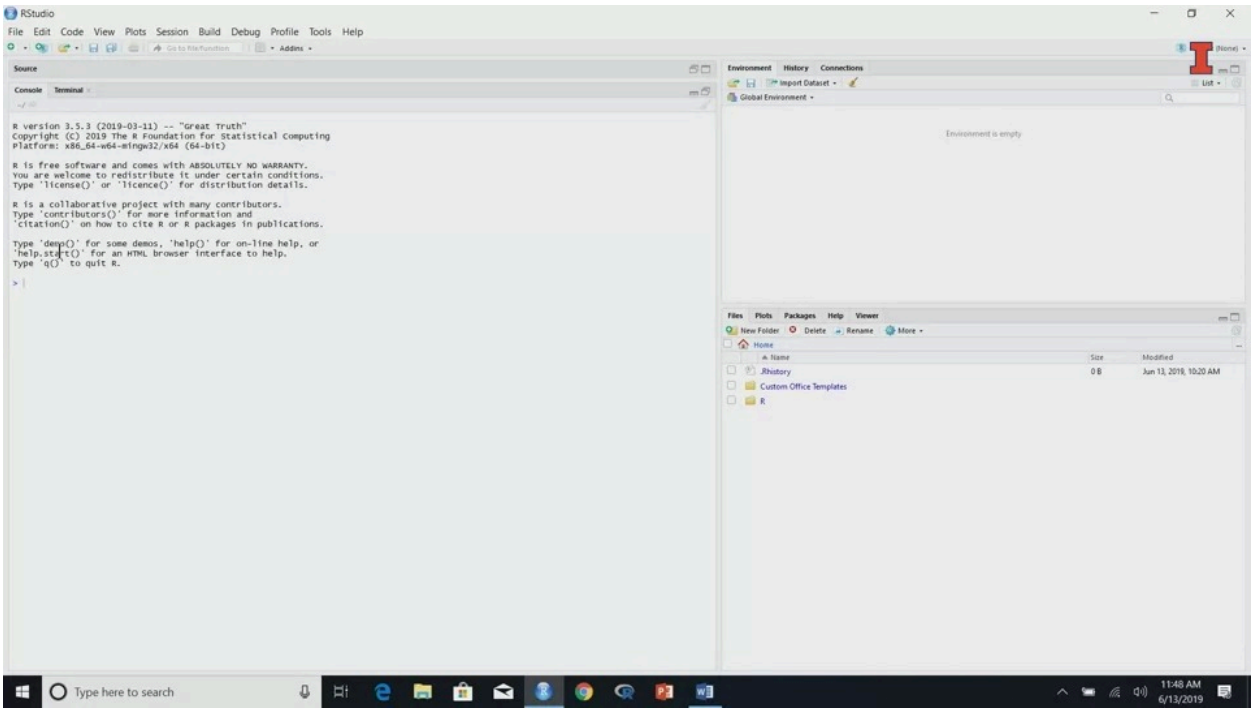
In the top right is your Environments, this is where your variables will show up. We'll talk about that in a minute. That's the primary one that I want you to use right now. Don't worry so much about History and Connections. History just has a history of commands that you've entered. Down below, here's files. This is a window where you can navigate through your file directory to find our files and supporting files that you might need data files. Here if we have any graphics, the plots will show up here. There is help. If you need help with a command, we can look here. This does connect to the Internet, so your Wi-Fi needs to be turned on. Here are the various packages or libraries that are available to you and we'll talk about that later. On the top right is an Editor. It's untitled now.

Tour of R and RStudio

Tour of R and RStudio



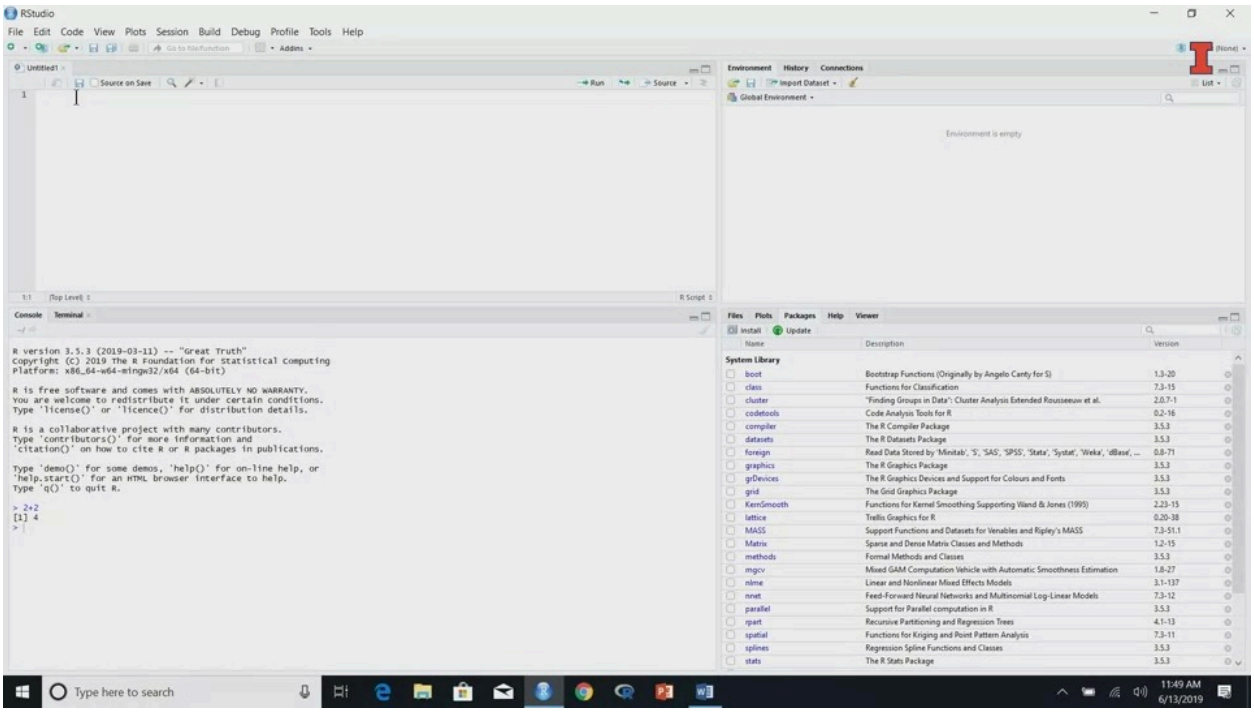
By now you have hopefully installed R and RStudio. You'll see two icons that look like this, R and RStudio. First I'd like you to follow along and click on the R icon. This is probably the only time you'll ever click on this. But I want to show you what it is. This is the original version of native R. Here you can see some introductory messages. The version of R you are running. The programmers like to give it a great name for each version. This version has to be called the Great Truth. A little bit about the licensing and some simple commands for demonstrations and help options. Here you see that little greater than sign, this is where you would type in a command. Let's type in something real quick. 2 plus 2 and you get an answer of 4. You enter a command and a response will come back. I'm going to close this now this is the last time I want you to open up native R because we'll be using RStudio, Save workspace image. No, I don't really want to save that. Let's click on RStudio and open that up. The first time you open it, let me maximize my window. You should see on your screen something that looks like this. You should see four quadrants. If you see three, you can click on this window icon here. Now you see four quadrants here, 1, 2, 3, 4. I can adjust the size of these paints by grabbing in the middle, I can go left and right. That's how you would customize your environment. I'm going to maximize the window on the bottom left. Here you can see the same introductory message. Before we open up our native R and we got this introductory message, and now it's embedded into RStudio in this quadrant called the Console.



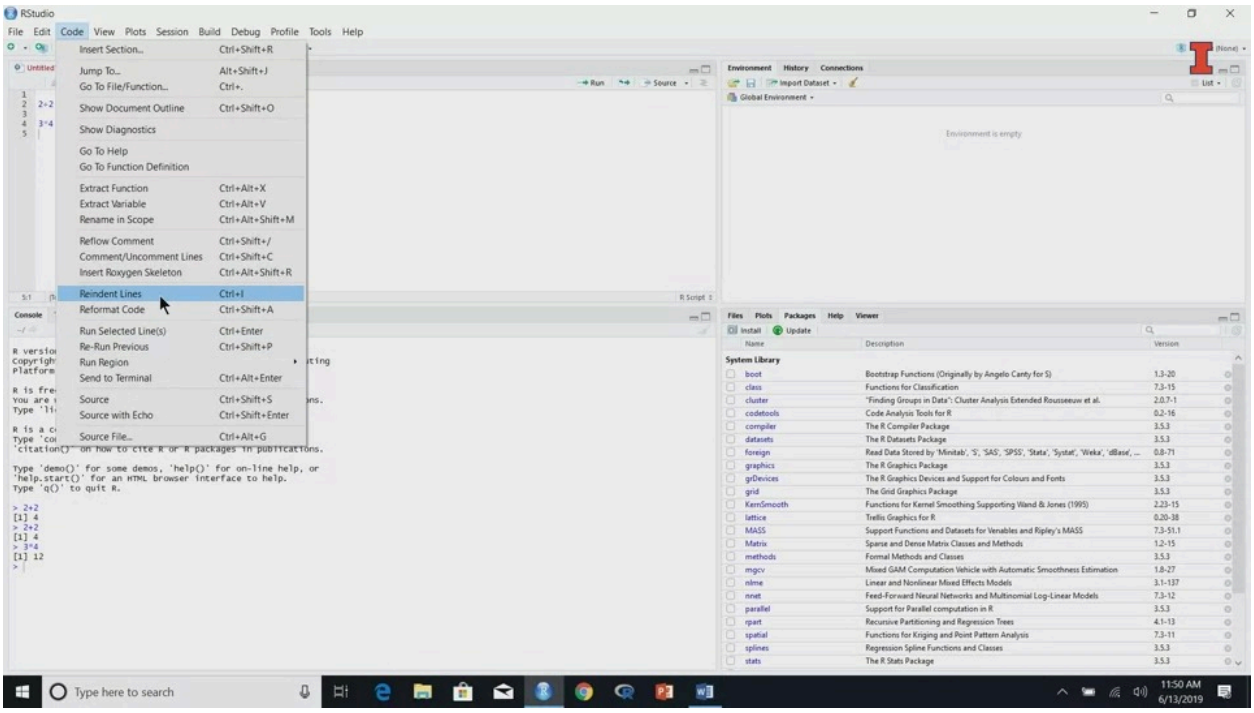
There it is. Here you can see the arrow, the greater than sign. Again, I can enter a command 2 plus 2, and I get a response of 4. Let's go to the other quadrants real quick.

Play video starting at :2:47 and follow transcript2:47

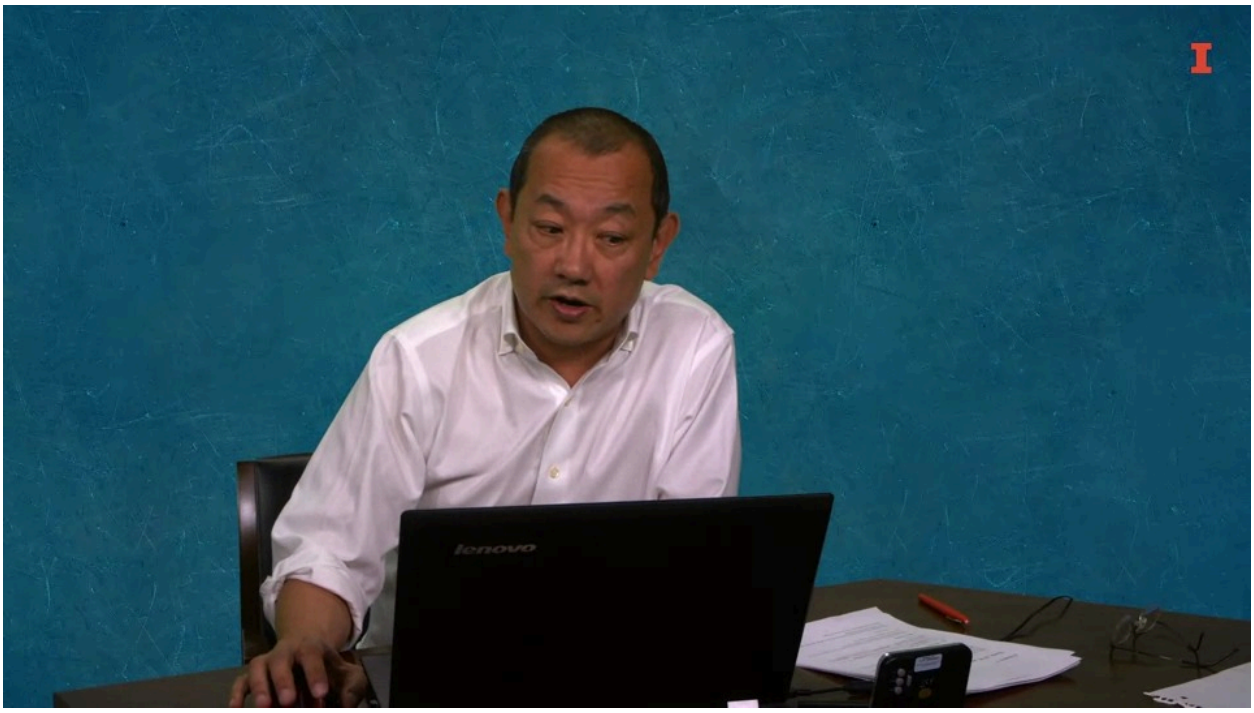
In the top right is your Environments, this is where your variables will show up. We'll talk about that in a minute. That's the primary one that I want you to use right now. Don't worry so much about History and Connections. History just has a history of commands that you've entered. Down below, here's files. This is a window where you can navigate through your file directory to find our files and supporting files that you might need data files. Here if we have any graphics, the plots will show up here. There is help. If you need help with a command, we can look here. This does connect to the Internet, so your Wi-Fi needs to be turned on. Here are the various packages or libraries that are available to you and we'll talk about that later. On the top right is an Editor.



It's untitled now. This is where you can write programs for your assignments. Here I can say we did something basic like 2 plus 2. Then I might want to do 3 times 4. That's a program. Notice nothing's executing down below. We can put the cursor there on line 2. Do you see that? If I hit Run, it'll run the current line. It's essentially cutting and pasting that line down below into the Console and hitting Enter for me. There it is, and there's 2 plus 2 equals 4. This cursor up here moves down to 3 times 4. I can hit Run. There it is 3 times 4, and the answer is 12. There are some more commands that you can use, they're under the Code menu. I'm just going to show you where they are right now, but don't worry about them. But you can Run the Selected Line. You can highlight some code and Run Selected.



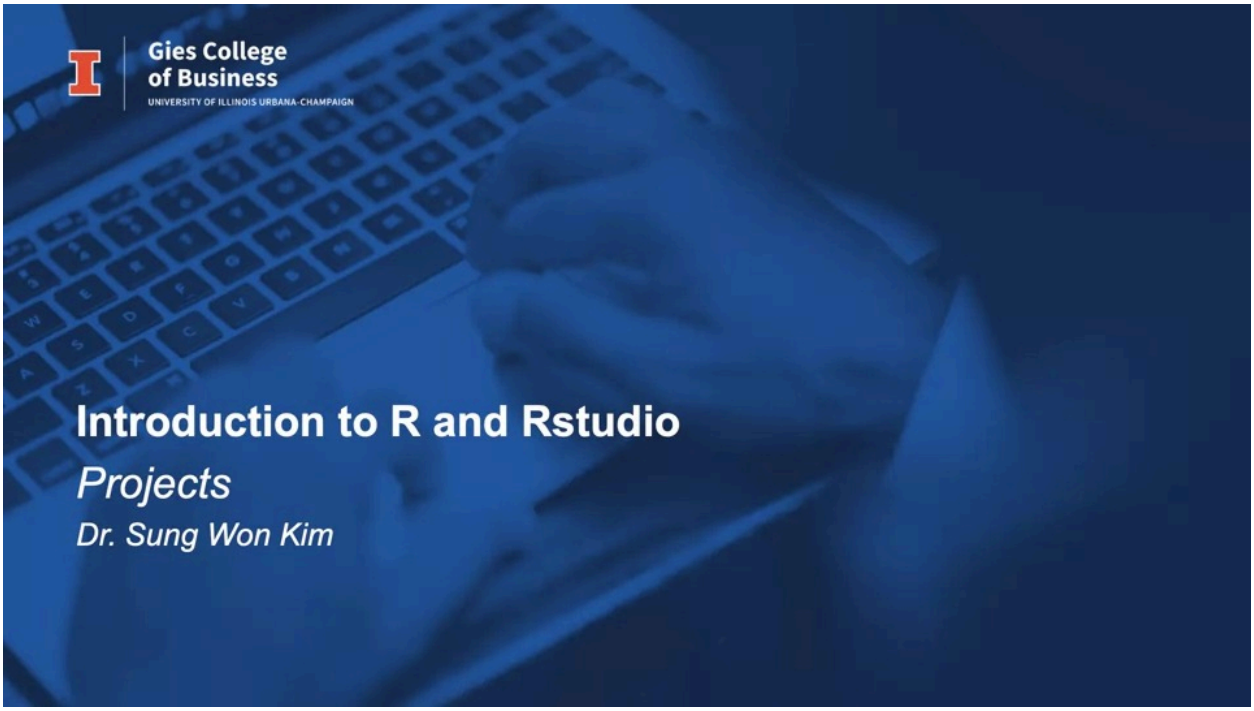
You can Run Previous, etc.
Play video starting at :4:54 and follow transcript4:54
Finally you can save your program. All this work up here. You might have a lot of code that you'll want to save into an R file. The extension is.R and it's like saving a Word document. But before we do that, let me talk about Projects. Let's create File, New File. Here you can see we can create a new R script, which is generally what you'll be doing for this class.



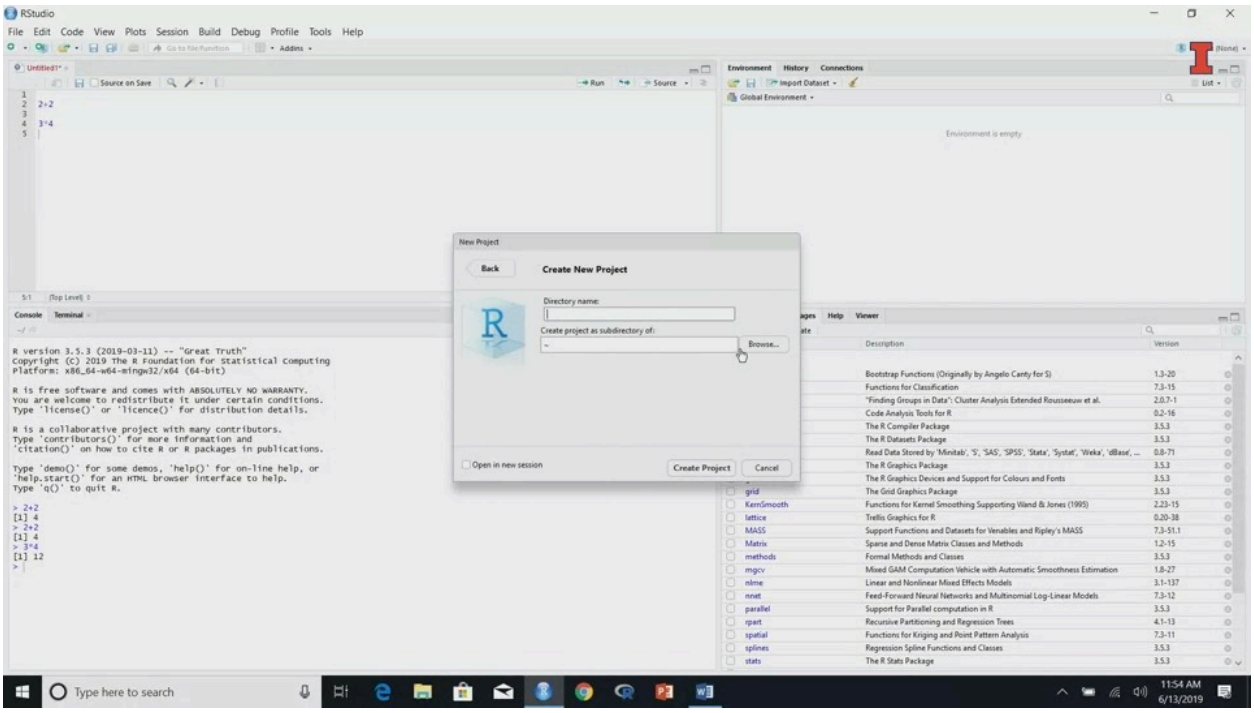
I've discussed the four quadrants, the Editing quadrant in the top-left, the Console, the Environment, and the bottom right, which has various utilities. Next we'll talk about Projects.

Projects

Projects



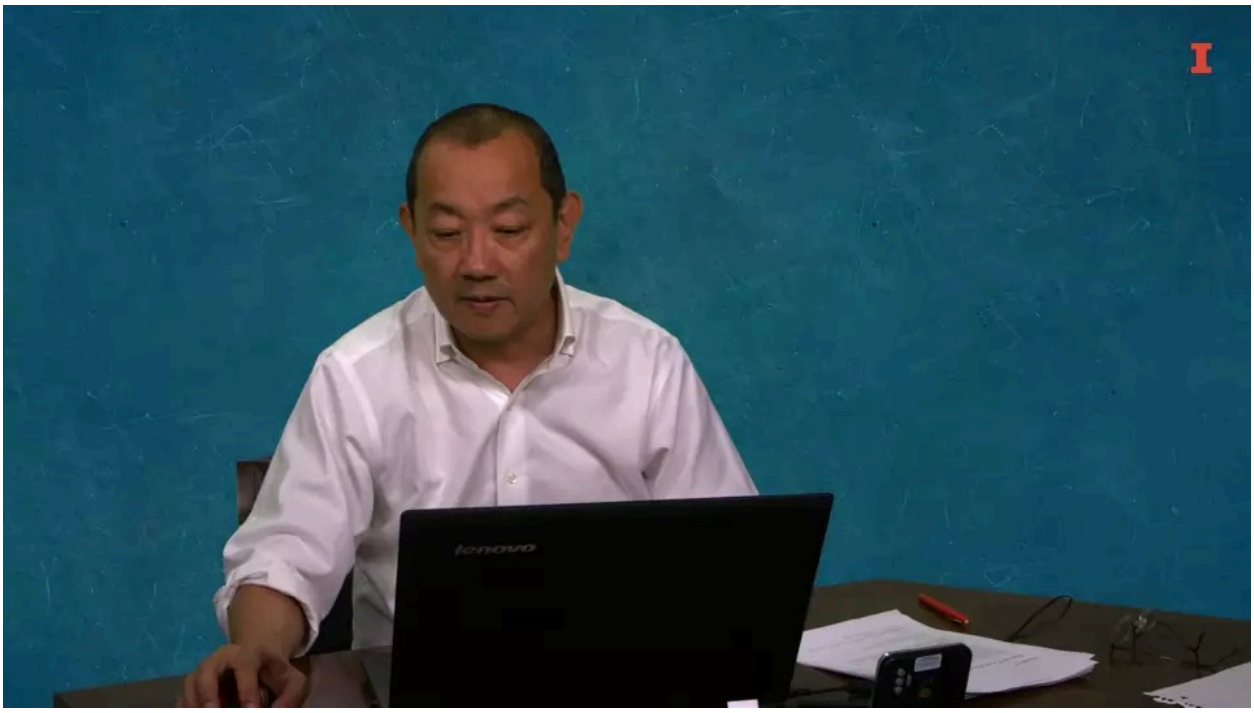
The first thing you want to do is to create a project. A project is essentially a folder that allows you to organize your files. Let's do that. File, New Project. You can use an existing folder on your computer or create a new folder. I'm going to create a new directory, a new project. Maybe I went too fast. New Directory. The first line is New Project.



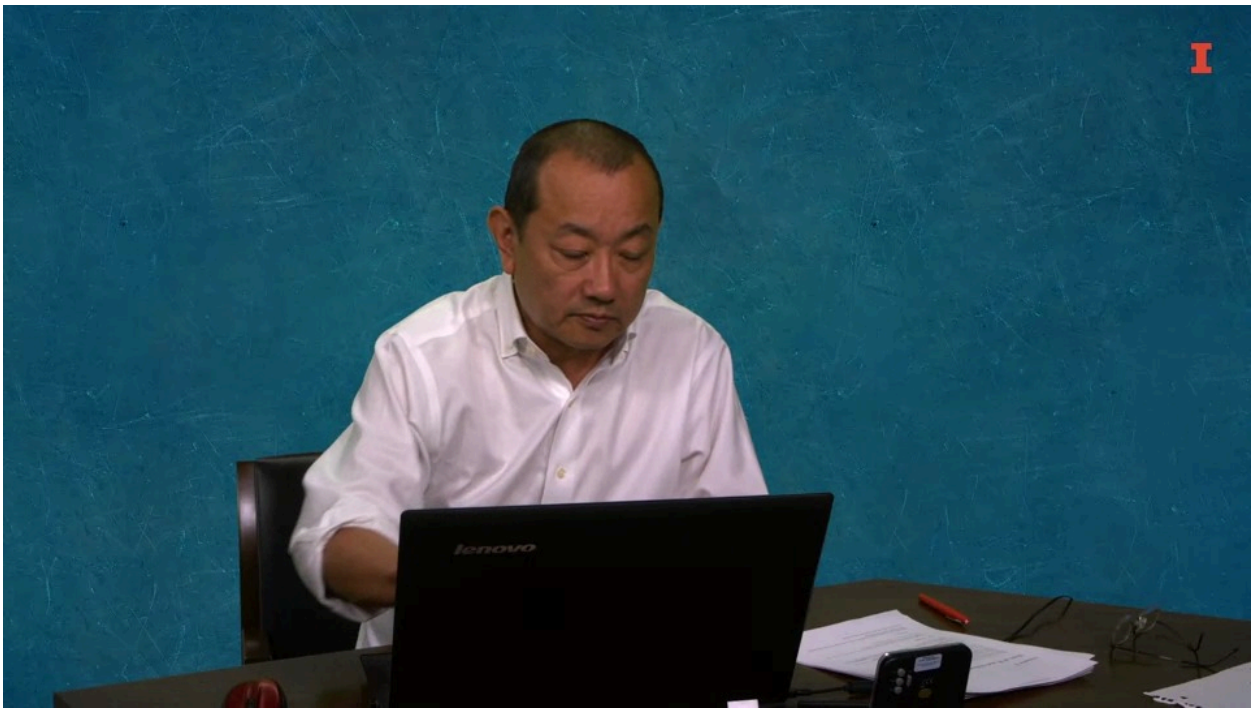
I'm going to search for a folder where I want to put it. Here, I happen to have on my desktop something for this Coursera class, so I'm going to put it there. I'm going to create a name. Let's call this Project 1, and then Create Project.

Play video starting at :1:5 and follow transcript1:05

You have a project. All this really does is create a folder on your computer. On the bottom right, if you click on the Files tab right here, you can see that it's a project, and RStudio will automatically put this file in here with a.PRPRJ file, and that helps R just manage your environment. Also notice in RStudio in the top right, here I am, there's a little drop-down here and it says Project 1. This is your current work environment or current workspace.



If you had multiple projects, you could switch projects, and that'll help you switch environments. That includes data files you might have in there, or R programs that you might have in there, or any supporting documents that you might have in there such as markdown. Now, say you wanted to create a program. Before we created a simple program, two plus two, three times four. I know, not very exciting. I want to save this file. I can use File Save As or I can use this little Save button, Control S on the menu. I'm going to click that. It's going to give me a file name and I'll just call this Program 1. Save. Now notice it automatically puts the. R extension, which lets your operating system millets in R program. Also notice in the folder in the bottom right, you can see here we have program1.R file. Then like before, we can execute it, and it shows up on the bottom right. One last quick note before I wrap up this session. Down below you see the console

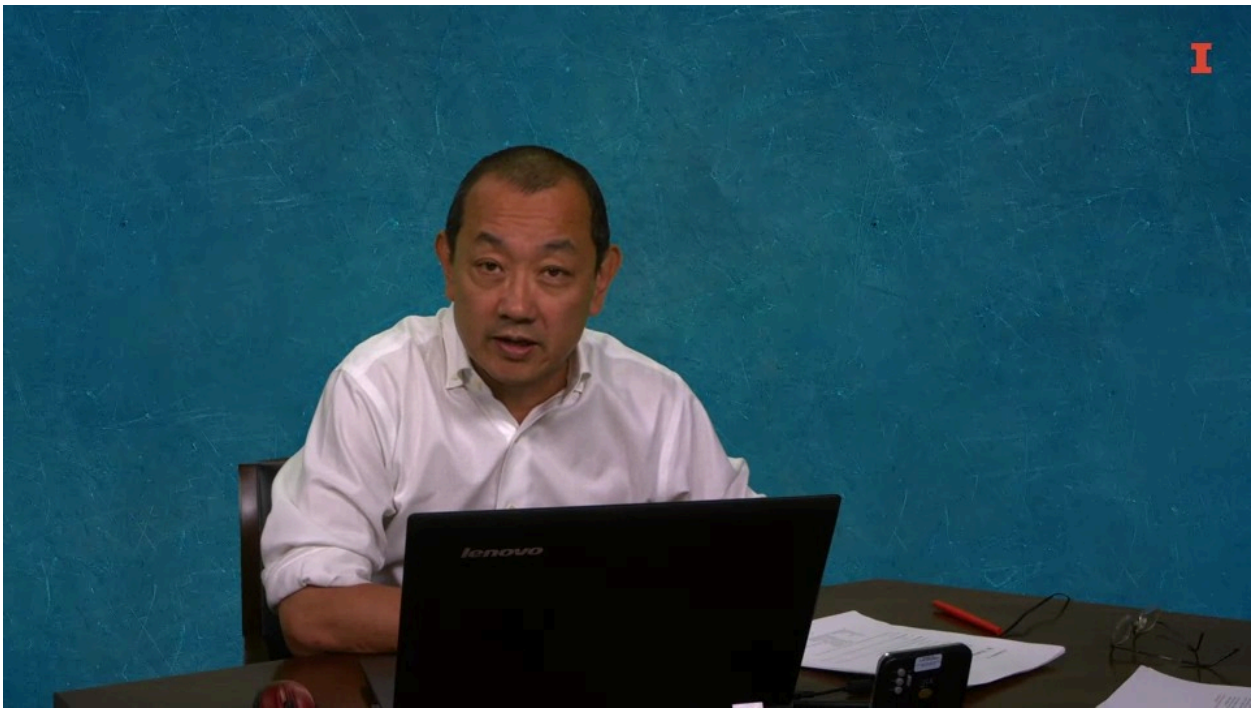


If it gets a little busier, just want to clean it out. To basically erase the console, you use the command Control L, and it'll clear the console. Note that when we get to variables, it does not remove any variables or any other data structures. That wraps up projects.

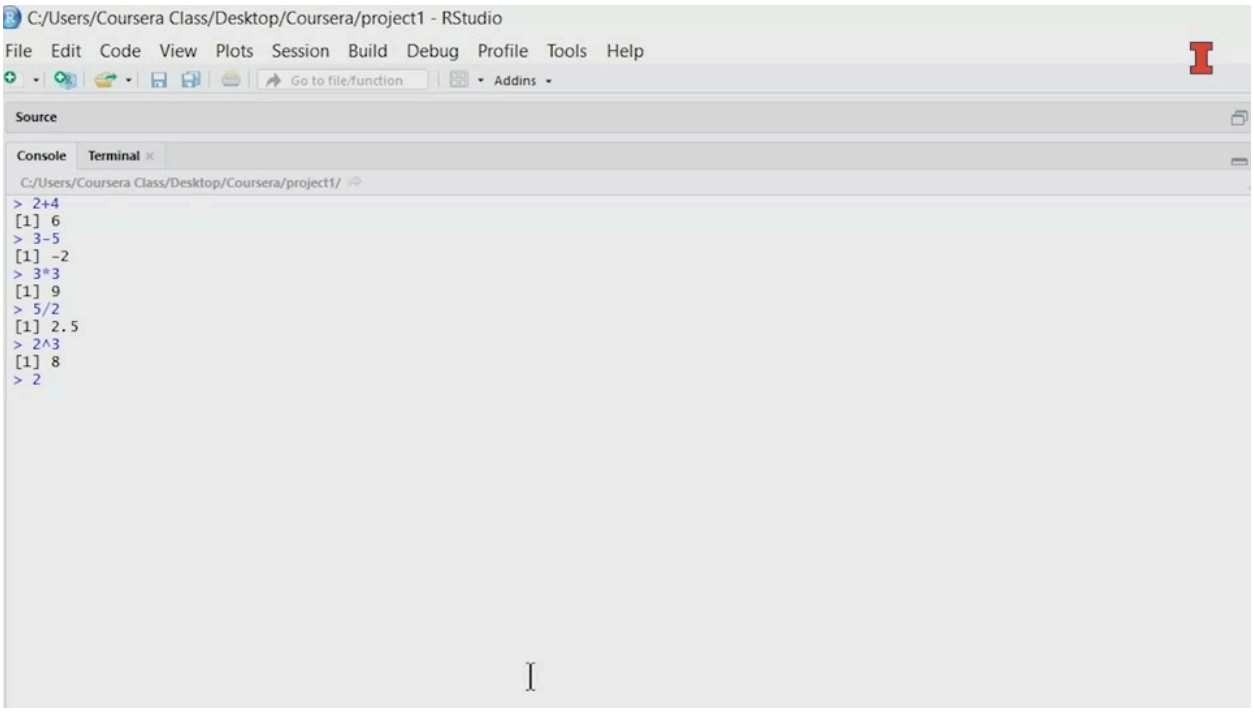
Math Function

Math Function

[MUSIC] Okay, so in this lesson we will start to look at some elementary R commands. I know some of you taking this class may not have any programming experience and some of you might have extensive programming experience. But for those of you who do not have any experience in programming in any language, please don't worry. In this class, the level of difficulty will be about the same as writing a function in Microsoft Excel.



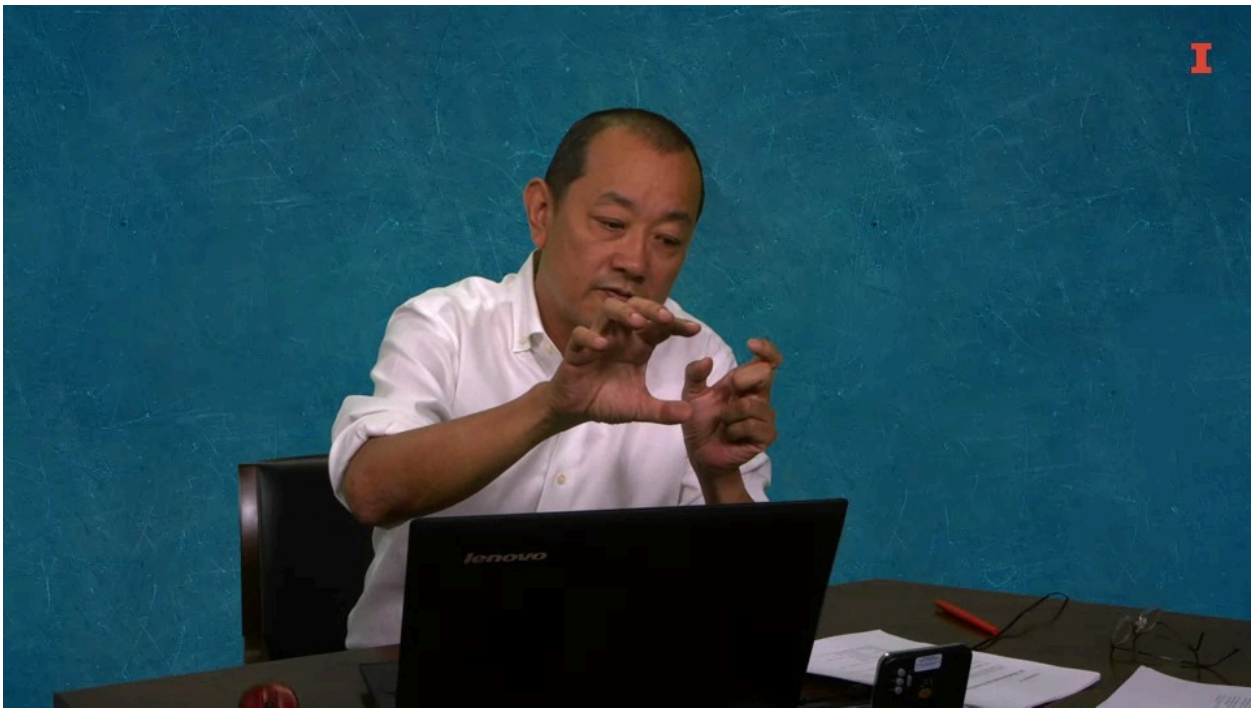
And we'll be taking baby steps along the way, and feel free to pause the video at any point in time in case you want to look over your notes. So let's begin. And the first thing I'd like to talk about is using R with mathematical functions almost like a calculator. So for this lesson, I'm going to focus mostly in the console area. So I'm going to maximize the screen.
Play video starting at :1:10 and follow transcript1:10
There we go. And I think I've showed you in previous videos some basic commands, like 2+4, so that's addition and we're rolling. Subtraction is with the -, so they can use 3 -5,oops, and that's -2. We can obviously do multiplication, 3*3 is 9. So far so good. Division is with a /, So 5/2 is 2.5. If you want to do an exponent for example, 2 to the third power, 2*2*2, that would be 8.



And another way to write the same thing is to use 2 and then instead of one asterisk use double asterisk, 2 to the third power and that is also 8. So those are your basic arithmetic functions. But you might want to go beyond that and use more complicated functions that you might have learned in the past. So we can take the absolute value of a number,right? Which essentially strips away the sign. So the command there is abs(-7). Say, oops I typed a 5 hit enter and the absolute value of -5 is 5.

Play video starting at :2:35 and follow transcript2:35

So one thing to note about functions and this goes for all of R, is that there's a function name in this case it's abs. There's an open parentheses.

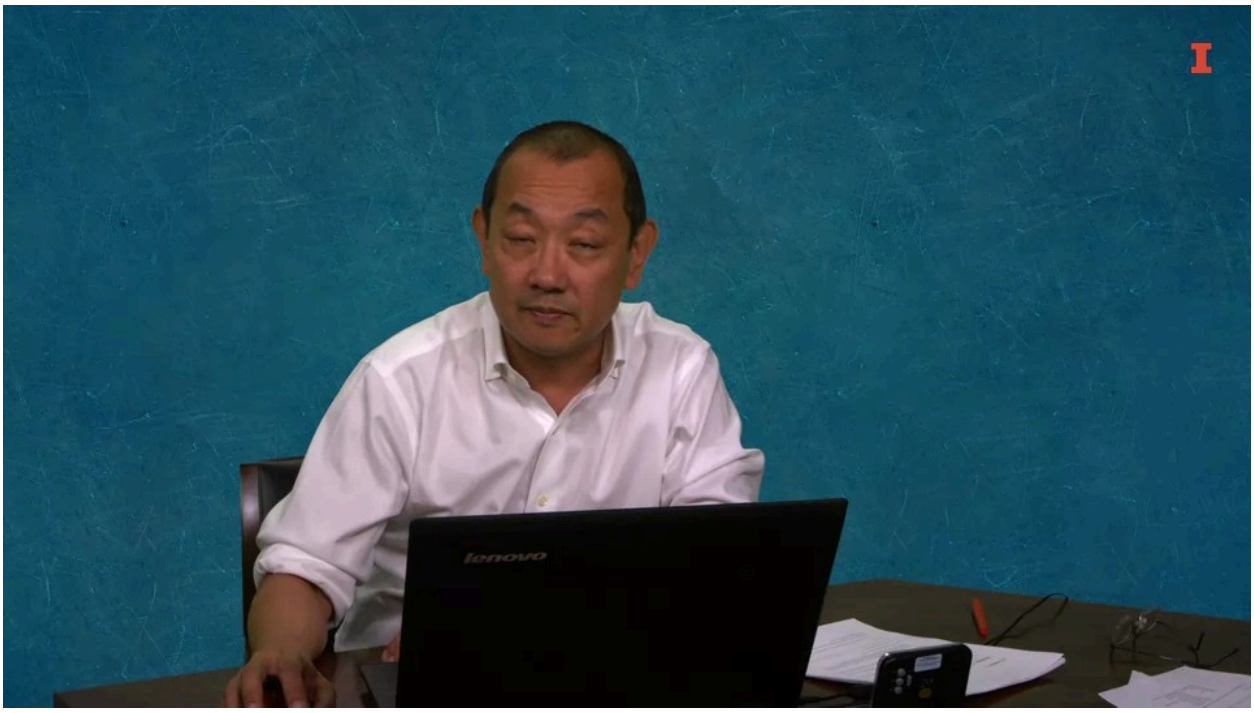


And then inside you'll put in your arguments, the variables of interest and then a close parentheses and that's the general form of a function. Other functions might have other arguments that you can put in there, but these are usually one number at a time. [COUGH] Okay so we have absolute value of 5. We can take the log of 5, we can take log(5, base=3, if you want to do it that way. Exponent exp (3). Square root, square root of 4 is 2. Factorial(9) is 362,000 plus. And then your traditional trig functions sign of 0, sine of pi there you have it. So those are your basic functions. I encourage you to look at the Archie sheet and practice some of these functions on your own. If you don't know what these functions mean, because you've never been exposed to something like, for example, factorial. Don't worry about it. Your knowledge will drive what you need to know in terms of the functions. So perhaps someday you will come across the need to use the factorial function. Maybe you're doing some combinatorics and when you're studying that, then this will become useful. And that wraps up arithmetic functions in R.

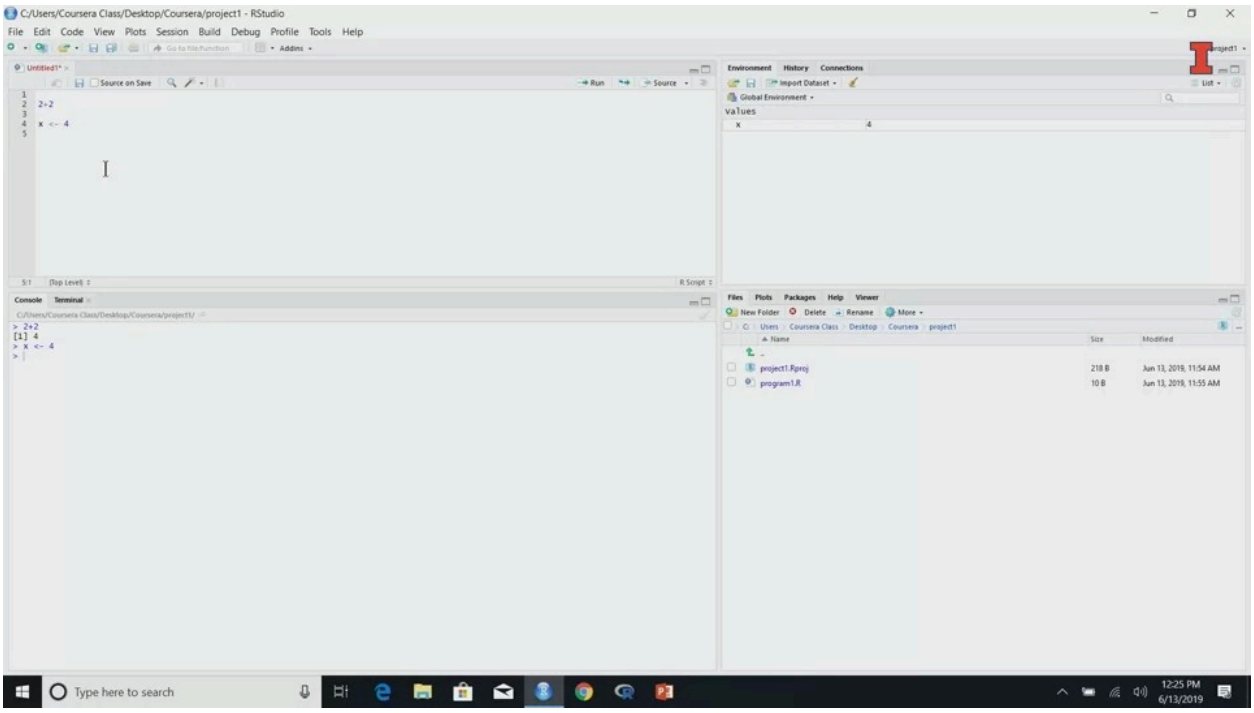
Scalar Variables

Scalar Variables

In the past video, we described how to use mathematical functions in r. But notice in this example here I have two plus two, if I run this code I get the answer for but the answer is not stored anywhere. So in this lecture we're going to talk about variables in our basically how to store information and the way to do that is really simple. First you create a variable name, I'm going to call it x. Not a very creative name but it'll do, then you use this thing that looks like an arrow. So it's a less than sign and a hyphen and then I'm going to put a value in there of four.
Play video starting at ::58 and follow transcript0:58
It hasn't run yet. So let me run that code down below. And as I mentioned in a previous video, when I hit that run button, it's like cutting and pasting that command down into the console and hitting enter and so on the console you see x is assigned a value of four.



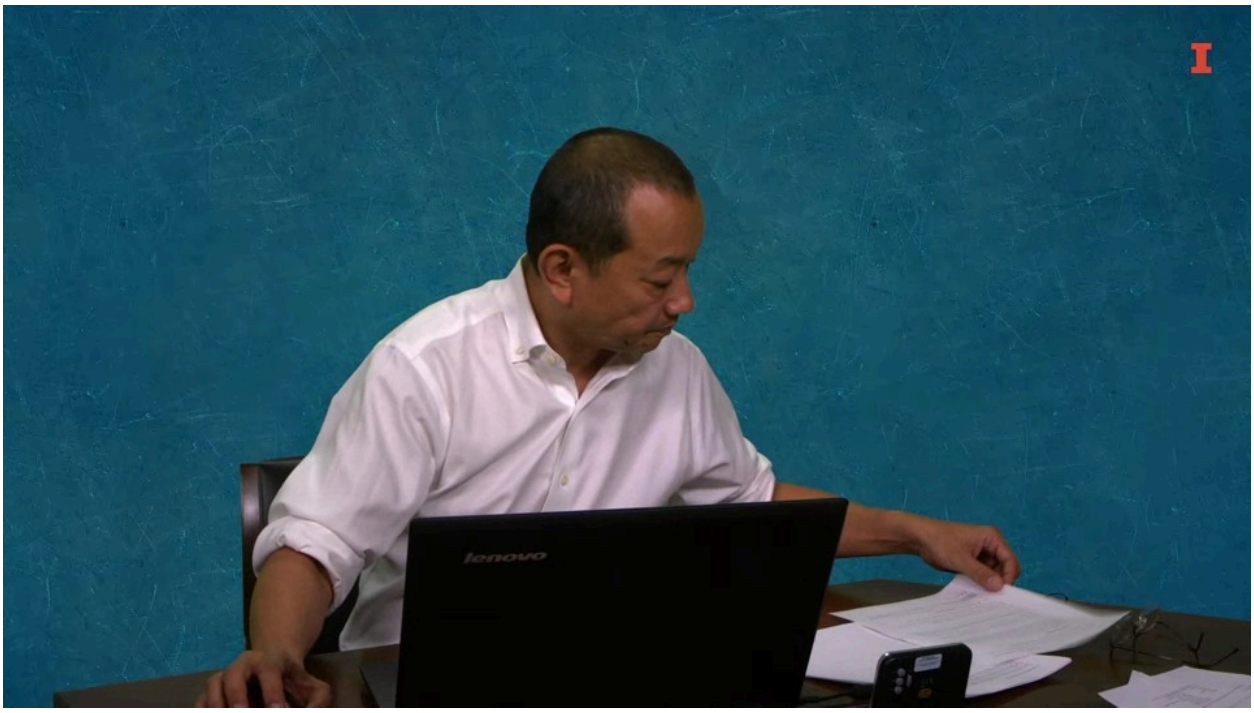
Now what does that mean? Creating a variable in r is creating, it's like creating space on the computer. So imagine you have a shoebox and that's your space on the computer. And then you label that shoebox with the stuff that you're going to put on it. In this case I named it x. It could be things like height. So if you have a team roster, this is the list of all the heights of all the players, it could be the weights, it could be birthdates, it could be whatever you want. So you want to label the shoebox and put like kind of data within that shoe box. Here, I just put the value of four. Also notice in the top right of our studio here, you can see in the global environment types or click on that if it's not active there's your variable x



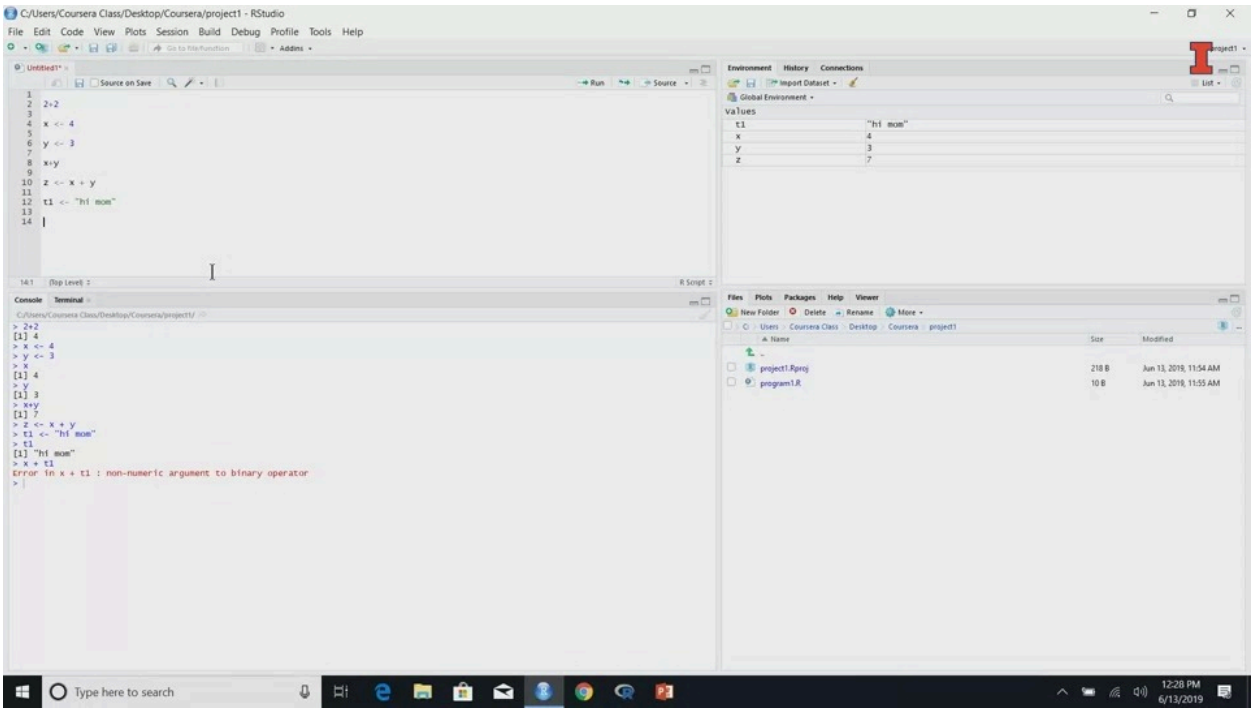
And here's the contents of that shoe box, its our, okay? Now let's create a second variable y. A value of three, okay? And there, now you can see I've executed that code here and you can see it in the global environment up here. I've created a little space for it. And y is equal to a value of three. Another way you can see the value of what's inside a variable.

Play video starting at :2:41 and follow transcript2:41

I'm in the console here now and I can just type that variable name, x. And it will respond with four in this case or y and it responds with three. So I can see the inside of the box that way, I can see it in the environment. And now I can do some commands as we noted before. So let's just do x plus y. There we go. We run that and x plus y in this case is seven. Which makes sense because x contained a value of four and y contained the value of three. But notice that the sum of x and y is not stored anywhere, so I can create a third variable, let's call it z. Use the assignment operator which is the less sand sign hyphen and x plus y. Let's run that line and you can see now in the top right in the environment section of our studio, we have created the variable z and at the same time put in the value of x plus y.



So that's how you create a scalar variable. One thing to note is that the variables can be of any type. So if you want to put text in there, you have to put it in quotes. So let's put, I'll just call this variable t1, for text example one. And I can write hi mom. And there it is and you can see in the environment there's hi mom. If I go down here and type T1, it'll show hi mom. But this is a character string. So what would you expect if I did something like x plus t1.
Play video starting at :4:41 and follow transcript4:41
Have you tried that? Don't be afraid to eat dinner but you can see that it's an error because it's trying to add a text string. String is a bunch of characters with a numeric. One final note about variable names. There are a number of ways you can write variable names. There are almost no limitations on what you can call a variable, but in terms of style and readability, you want to pick a style that's consistent throughout your programs. And one style I like to use is called lower camel case

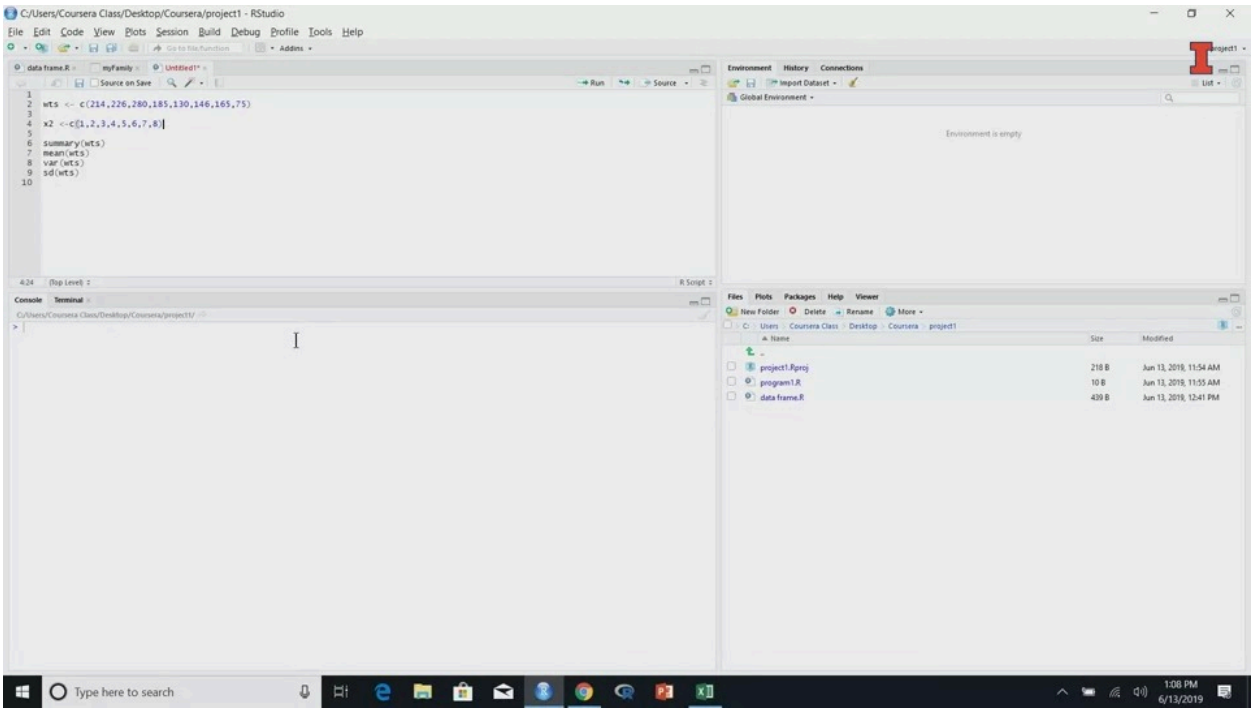


So what does that mean? Say you have a column of data or variable that is something like shoe size, camel case would look like this. Shoe size. So notice I've capitalized S here and so you can string along a phrase and then capitalize each word. Some people like to capitalize the first letter as well, so it looks more like capitalist for shoe size, something like this, it doesn't matter. Just pick a style and and stick with it for readability. Also note if you go into the corporate environment, they might have their own data variable naming standards. So that's something you should think about and try to stay consistent with the corporate environment, and that wraps up scalar variables.

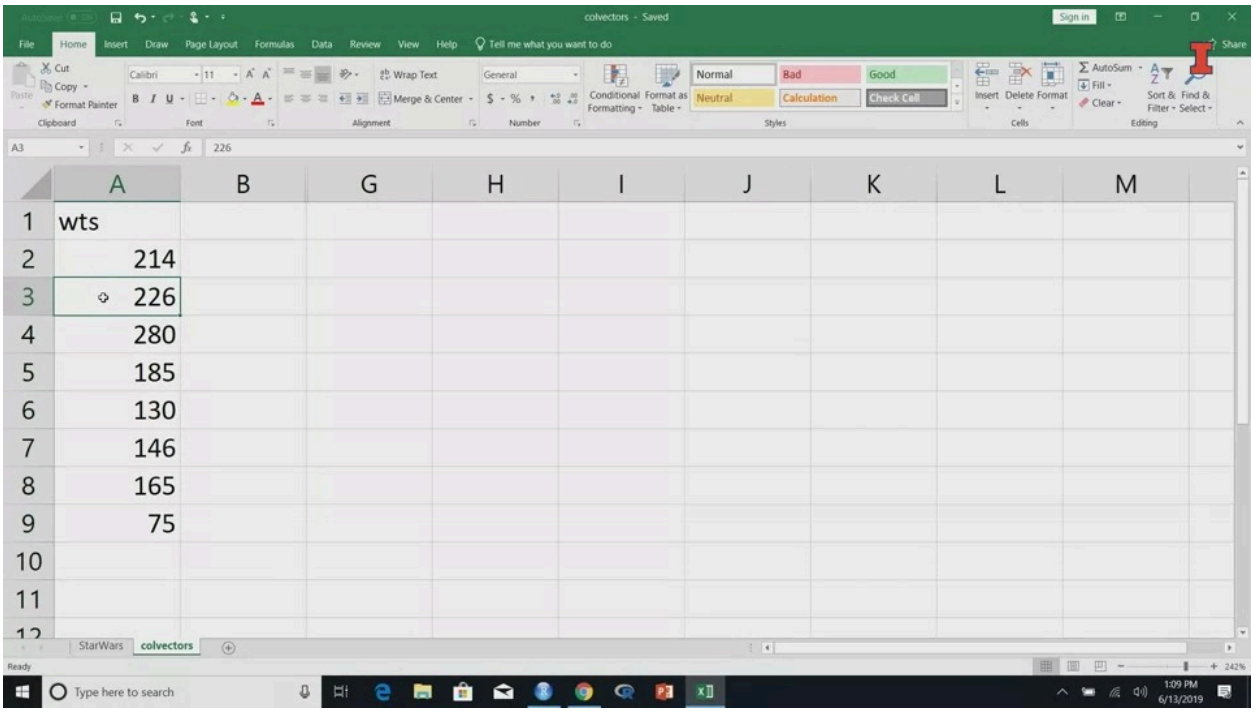
Column Vectors

Column Vectors

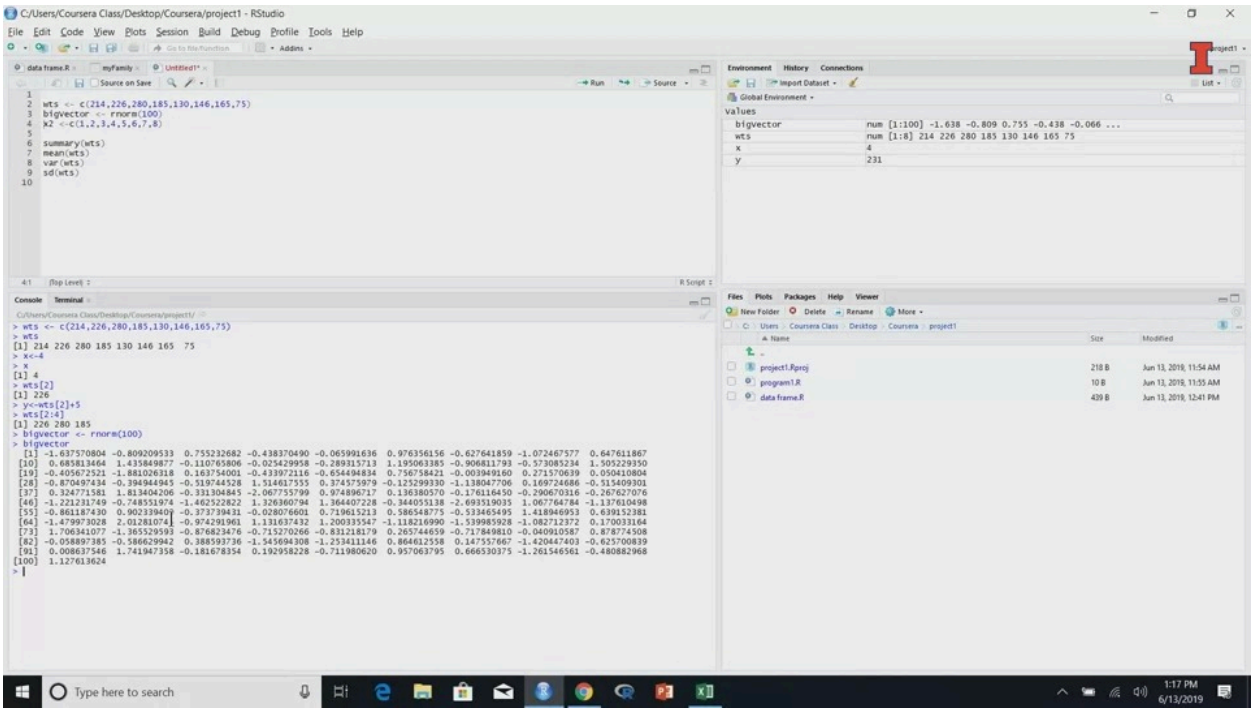
In this lecture, I'm going to talk about column vectors in R. Column vectors another basic datatype in R that'll help us store data and be able to manipulate data. The concept of a column vector is not too hard, I'm sure you've seen before. Here in Microsoft Excel, I have a column of weights. I call the column wts for weights, and I have some weights of objects ranging from 75 at the bottom to a maximum of 280 pounds or kilograms or whatever they may be. Here's some weights. We might want to create something similar in R. It's on this line 2 here. The column vector name is wts. You still have that less than sign hyphen for the assignment operator.



To create a column vector, you use C open parentheses, and there's the whole bunch of numbers there, the data 214, 226, etc. down 75 close paren. That's how you create a column vector. Let's do that. I execute the code, it shows up in my global environment. There it is. I can also type in wts and there it is. It's listed out 214, 226, 280, etc. One thing to note, I can finally talk about this square bracket one item here. If I created just a scalar variable, x equals 4, and I show x, I have that square bracket 1. In fact, a scalar variable is a column vector with only one element in it. This is telling you which element it is in the column vector.



. Here I am in Excel and 214 is the first element in the column vector, 226 is the second element in the column vector, and so forth. Seventy five is the last element or the eighth element in the column vector. If we go over to RStudio, we can see 214, 226, etc. If I wanted to just get the second element, the 226 pounds, wts square bracket 2, I can access that. There it is. I can even put it into another variable, y, y is wts square bracket 2 plus 5 and you can see y is now value of 231. If I wanted to access more than just that second cell, if I wanted to get maybe the second through fourth cell, it would be wts square bracket 2, colon 4 and there they are. This square bracket one on the response will tell you this is the first element. If it runs over, you will see the next number. Here, I've accessed the second through fourth variables of the weights and I get 226, 280, 185, which we can verify here 214, 226, 280, and 185. So that works. This square bracket one is the address of your column vector. Here this is the first element, this is your second element, this is your third element, etc. If it wraps around, you'll see the address of the bigger vector. Let's create a big vector, and I'm going to call it big vector. I'm going to use a random number generator. Let's create 100 random numbers.



If I want to look at big vector now, you can see that here, the square bracket 1, this is the first element, 1 minus 1.67 etc. That's the first element. Then we have here a square bracket 10. This item here is the 10th element, and so forth until we get the 100th element. That's what this addressing here is. The next thing I want to talk about are some basic commands that we can use on a vector. I have them here in lines 6 through 10. The first is, we can get a summary of the weights. Here are your descriptive statistics. The minimum value, the maximum value, the mean, the median, and the quartile ranges. If I wanted just the mean, I could do that with the mean command. There it is, and the variance and the standard deviation as well.

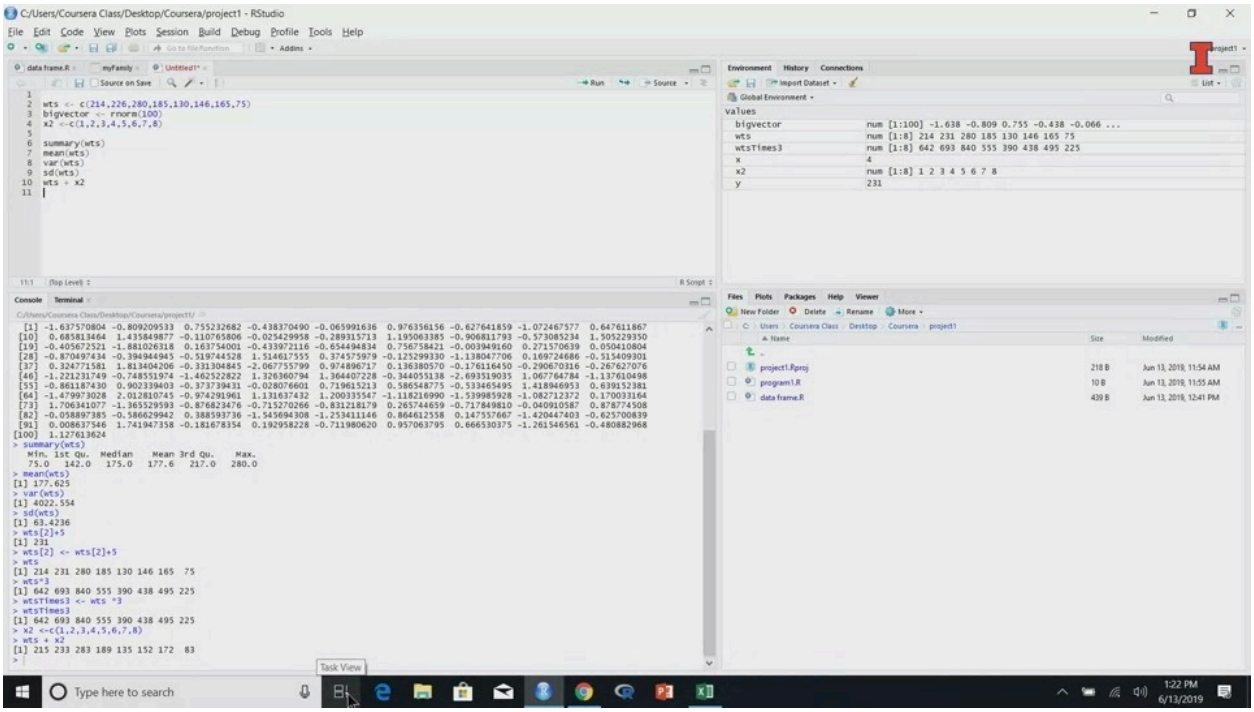
Play video starting at :5:33 and follow transcript5:33

We might want to manipulate some of the data within a vector. Up above I showed you y is equal to weights, the second element plus 5, wts second element plus 5. But note that's not being stored anywhere. I could store it in another variable y which I did up above, or I could store it back into that same location by saying wts 2 is assignment operator, wts 2 plus 5. Keep an eye on wts 2, which is 226. I'm going to add five to it and essentially I might take it out of that box add five and then put it back into that variable. Now, you can see it. Now, if I look at wts you can see that the second element has changed. We can also manipulate all the values within a column vector, Unhide. For example, Sam wanted to multiply this column. This is my original column vector. Note I've changed that second one added five to it. Say I wanted to multiply by three and these would be the values. That's very simple to do. We just go wts star for multiplication 3 and there they are. Again, I can put that back into that same column vector or I could put it into a new column vector. Let's do that. Let's call it wts 2, or let's say times 3 is wts times 3. There you have it. We can inspect the value of it , and there it is.

Play video starting at :7:36 and follow transcript7:36

We can also manipulate two column vectors. Here's a column vector here that I've created, and I've just created a sequence of numbers one through eight. I've named it x2, and I'm going to add the two together so that the corresponding entries add up. Here 214 plus 1 is 215, 226 plus 2 is 228, etc. Note that I did change the value of that second element. I'm going to create a column

vector x2 here it is, in line four. Let me run that. X2 is a column vector 1, 2, 3, 4, 5, 6, 7, 8. Now, I can do something like wts plus x2. Run that. There it is. Now, the corresponding entries have added up.

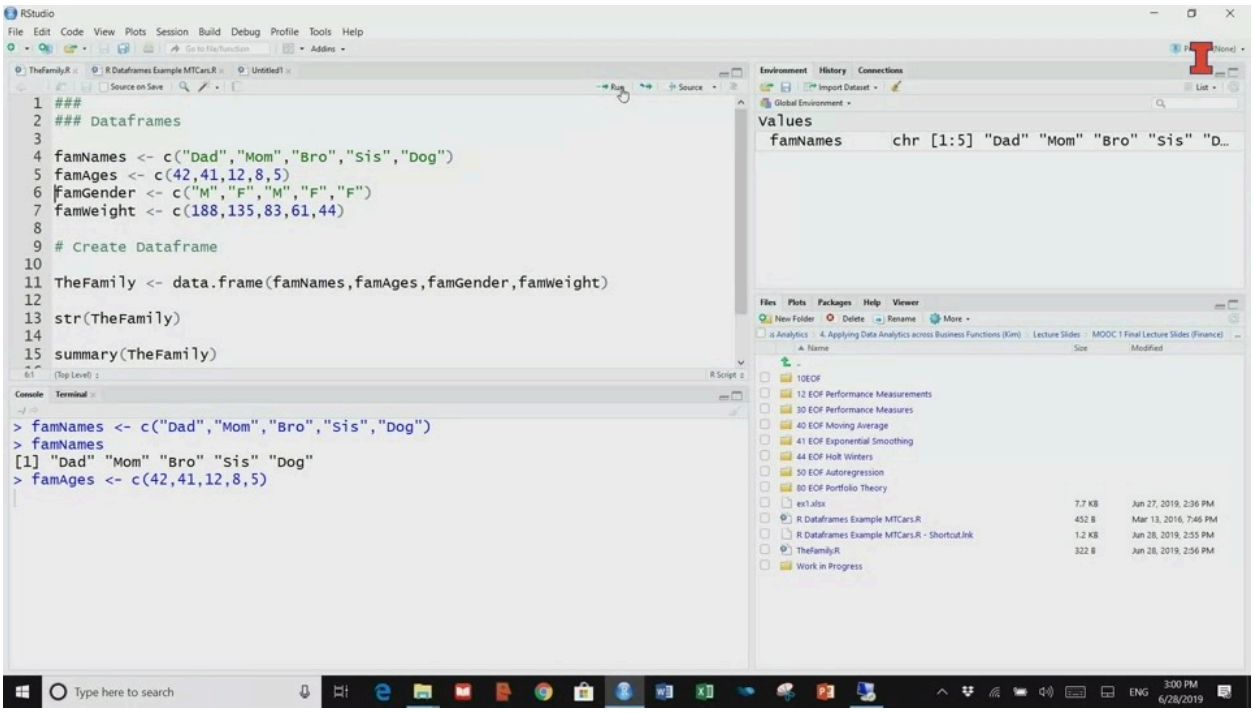


That's how you manipulate column vectors. You can also use multiplication and division. If you're unsure of how these arithmetic operators will behave, just create a small subset like I did and test it out for yourself. One thing to note is that you want to make sure that your column vectors are the same length. R does have special handling techniques for column vectors of different techniques which are beyond the scope of this video in this class. If you're interested, you can look up the documentation, but for now, and in most cases in real life, column vector should be about the same length. That wraps up column vectors.

Data Frame

Data Frame

In this video, I'd like to talk about dataframes, another data structure type in are. And it's nothing more than a table that you might see in an Excel spreadsheet or something like that but let's build one. The first line here and this code is a bunch of family names. It's a column vector and the names are creatively dad, mom bro, sis and dog. So let's create that vector. And if we take a look at it, we can see there's the column vector.



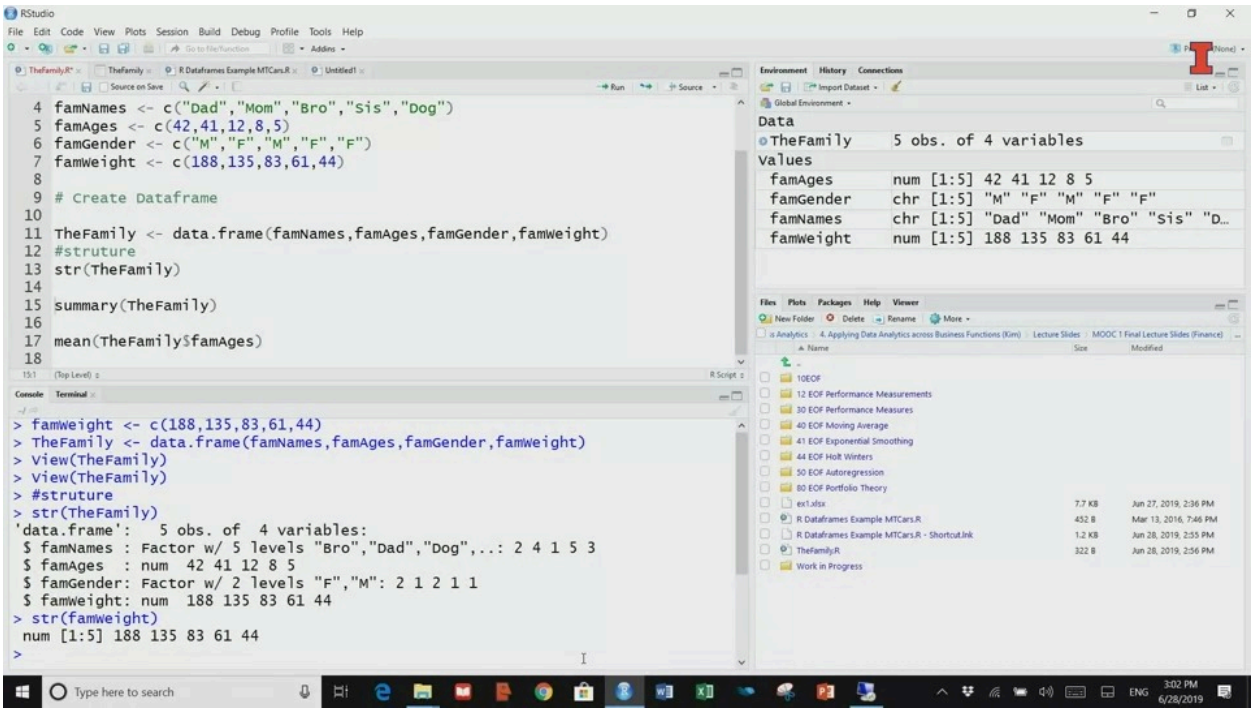
Similarly, we have the creative column vector of their ages and we can take a look. There they are and their genders and their family weights. So now I have these four columns. One is their names, one other ages, one is a column for their genders and their weights, right? So this is some data that you might typically find. I'm going to create a dataframe. And to create a dataframe, you need this sort of dataframe open print and then the names of each of the columns that you're going to put in there. And then I want to name the dataframe the family. So let's run that line of code, there it is. And now we can actually take a look at it. And you can see there it looks like almost like an Excel spreadsheet. The column names are at the top and there they are. Here the names, there's their corresponding ages, their genders and their weights.

Play video starting at :2:6 and follow transcript2:06

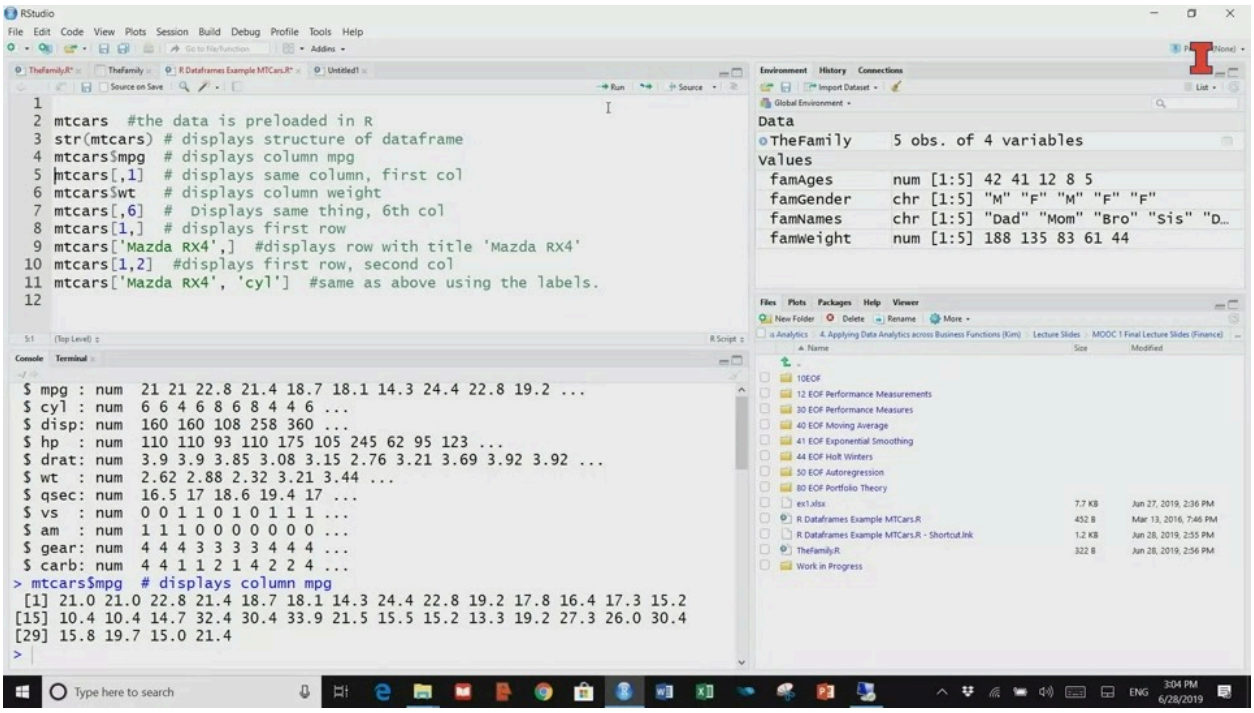
If you come across in someone else's code or some sort of data structure you're unfamiliar with, you can use this str command which stands for structure. I can even put that in the comments here, structure.

Play video starting at :2:26 and follow transcript2:26

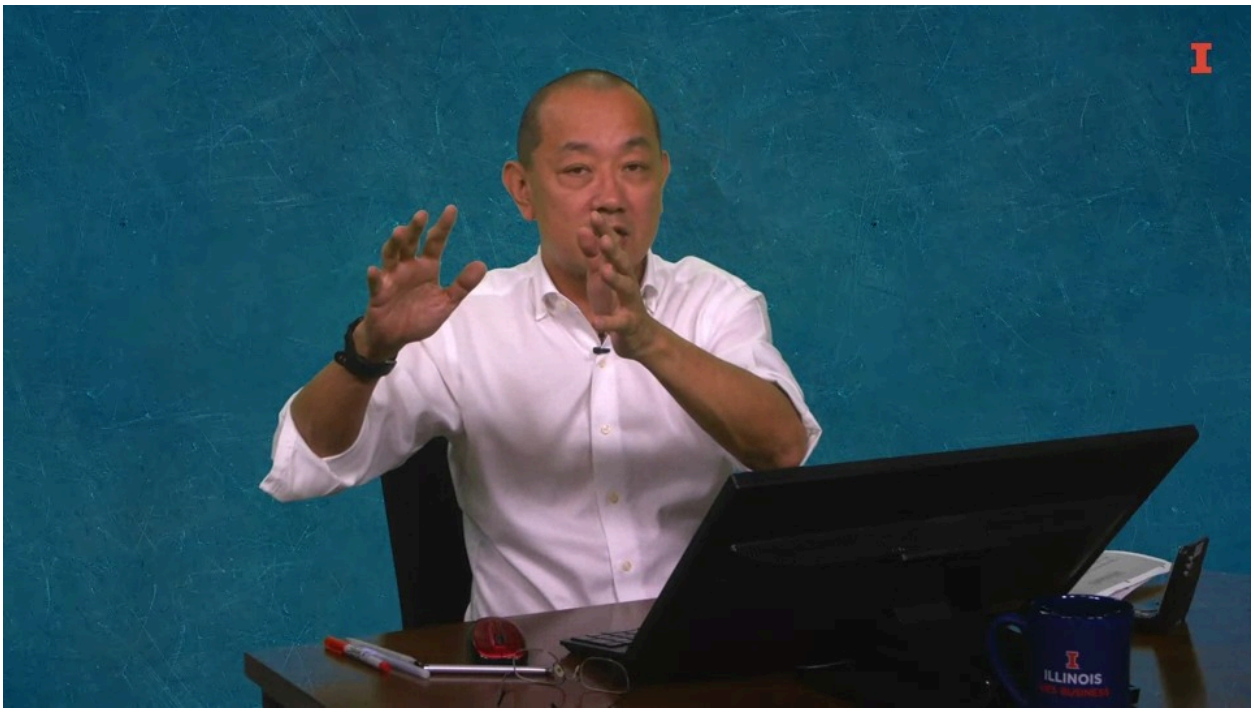
And that will tell you what kind of data you have. So, here we did structure of the family. And you can see that there are five observations of four variables and it sort of lists them out. We can also get, let me go below here, the structure of maybe the weights, just that column vector, f-a-m-w-e-i-g-h-t-s.



And you can see here it's a numerical column vector ranging from 1 to 5 elements and there are the actual values. So, structure command tells you a little bit about the data that you're working with. If you wanted to access just one of the columns in the dataframe, you can use this str sign name notation. So, in the structure command, you see the name is the family, and then you see data family names, family ages, etc. And if you want to just reference that, you can do something like the family, dollar. And let's get the ages, and there they are. And there's the column vector. And now we can do things with that calculate the mean. I've already had the command up here, what does the mean or what is the average age in the family? And there you have it. One thing you need to be able to do when you have a table is to be able to access a row or a column or an individual cell. And there are different ways of addressing each of those components. Here is a preloaded data set in not, it's called mt cars. Mt stands for motor trend magazine, mt cars. And here's a dataframe. It has the model of the car, mpg, cylinder displacement, etc. We might want to look at the structure. Is this a dataframe? And there you go, it's a data frame. It has 32 observations of 11 variables in this case. If we just wanted to get the mpg column, like I showed you in the family example, that would be dollar mpg and there they are.



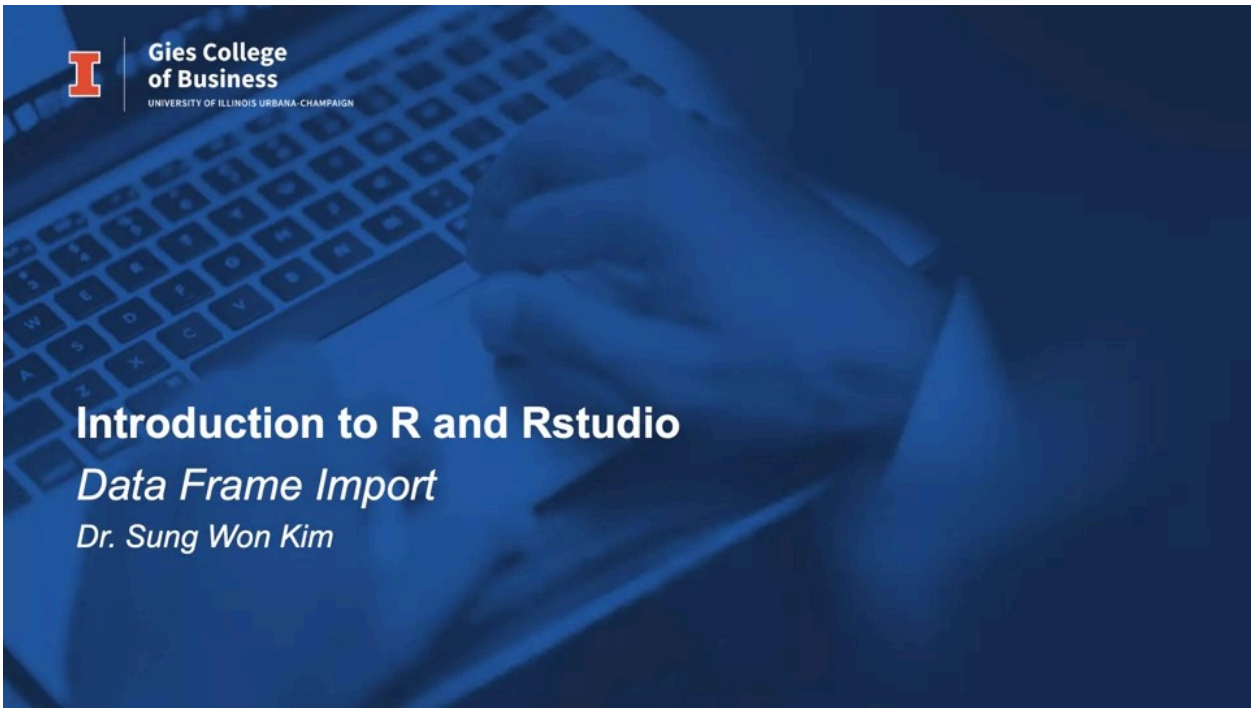
If I wanted to get just the first column, I use this notation square bracket something, something. So, let's look at line 10. This is row 1 column 2. So if I run that line of code, I get the value of 6 which is row one column 2, its that value there. If I wanted just the first column, I put nothing in that first element before the comma and just say the first column and there you go. That happens to be the same as the column for mpg, all right? That's the first column. If I wanted to get the column weights, there are their hopes, mt cars dollar weight, there they are. If I wanted the sixth column, I can do that, if I wanted the first row, right? So now I'm switching square bracket row one, give me all the columns in that first row. There they are. In this case this the state of frame does have some labels. So we can do it by the label of the row, and there you have it. That basically will help you get around the data sets. Generally speaking for this class, we will be using just give me the column. So, these first set of commands on what you really need to know even the name of the column. So data frame name and column name.



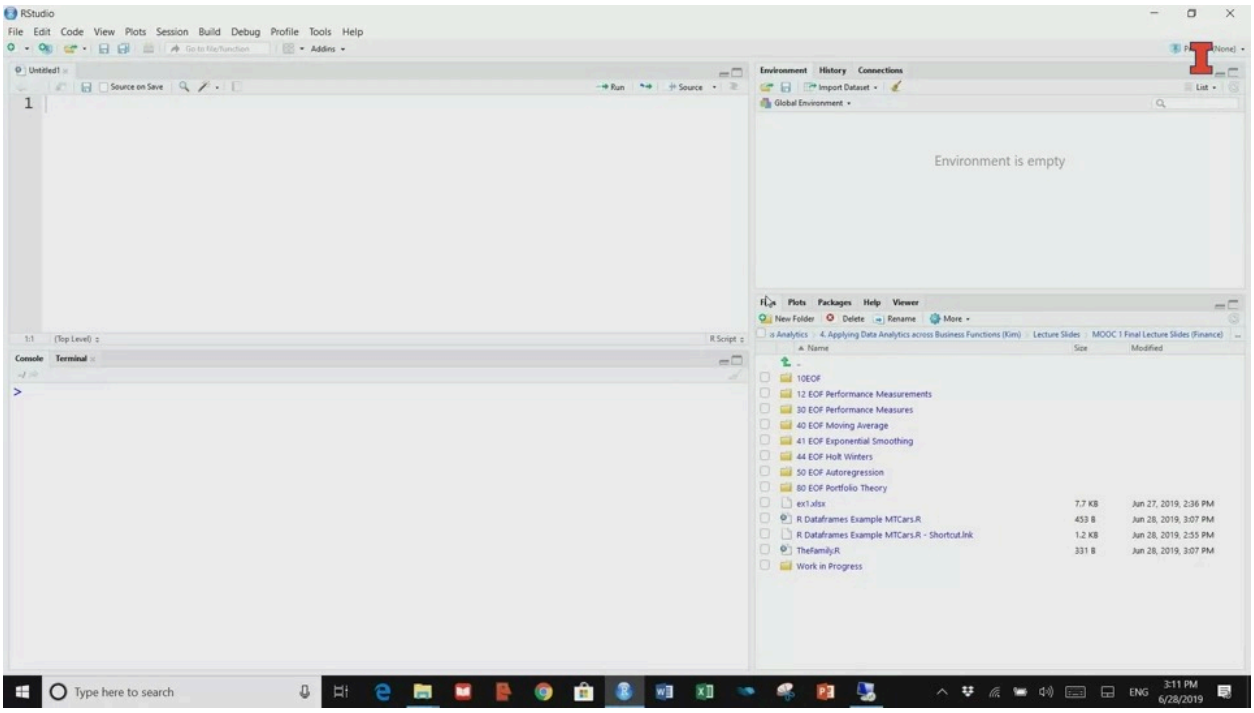
If you're used to Excel, that would be your Excel file name and then your column name and then just get picking out the individual columns. And that sort of wraps it up for dataframes.

Data Frame Import

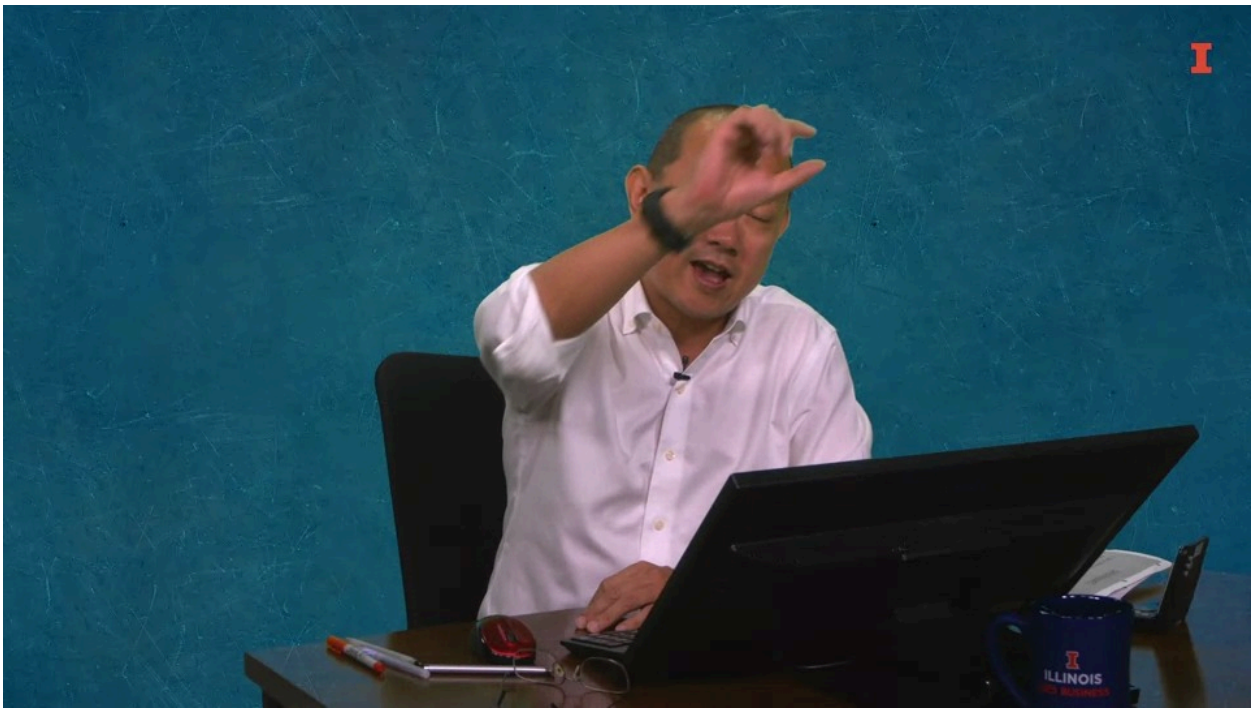
Data Frame Import



[MUSIC] We talked about data frames as being one of the workhorses, one of the primary data types that we'll use in r and it's essentially a table of data. And in the video about data frames, I showed you how to create one from scratch within the r environment. But in reality in the in the corporate environment, if you're usually given a data set in some sort of table, usually an Excel table or a comma separated value table and then you want to import it into r meaning, you want to bring that data into the r environment. And this is how you do it. One is to use this import command up here and as you can see there are some choices you can import from text files, from Excel's files and also data structures that come from other statistical packages like SPSS, SAAs data.



The second method is to go to this tab in the bottom right hand corner, under files, you click this ellipsis to navigate to the directory that you want to be in. So here's a bunch of directories. I've already navigated to the directory and there's an Excel file that I'm going to show by way of example, I can click on that, import data set and now I can see a preview of the data set and I can import it. One of the options is first row as names. Now in this data set, I didn't put names in the first column. So I'm going to unclip that. And now I can see my full data table. I have three rows, three columns. There they are.

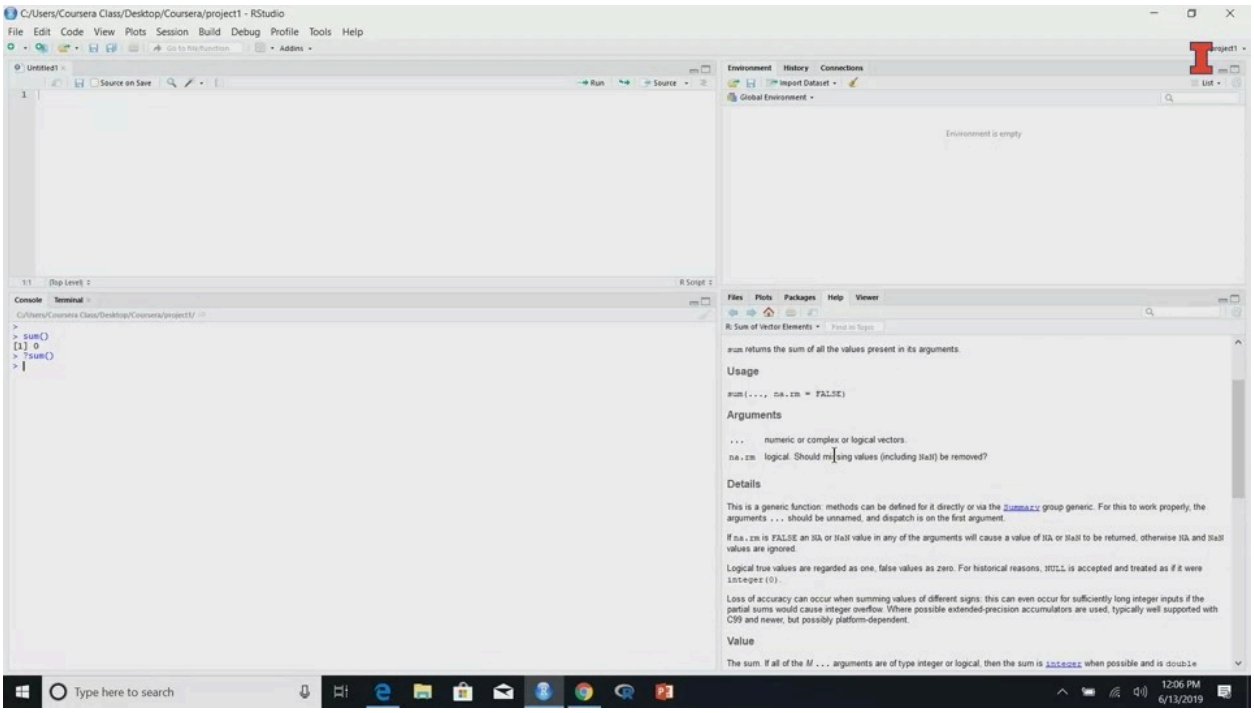


If I had used the first row as names of column headings, like name, age, height, weight, something like that, then if you click on that box here, r will know to import that first line of the table. So I can import that. There it is. I have an example of the table, and now if I do a structure command of EX1, I can see there they are and now I can access different elements within that table. So just by way of example, say I want to get the second column and there they are.
Play video starting at :2:47 and follow transcript2:47
And that wraps up how you import a data set.

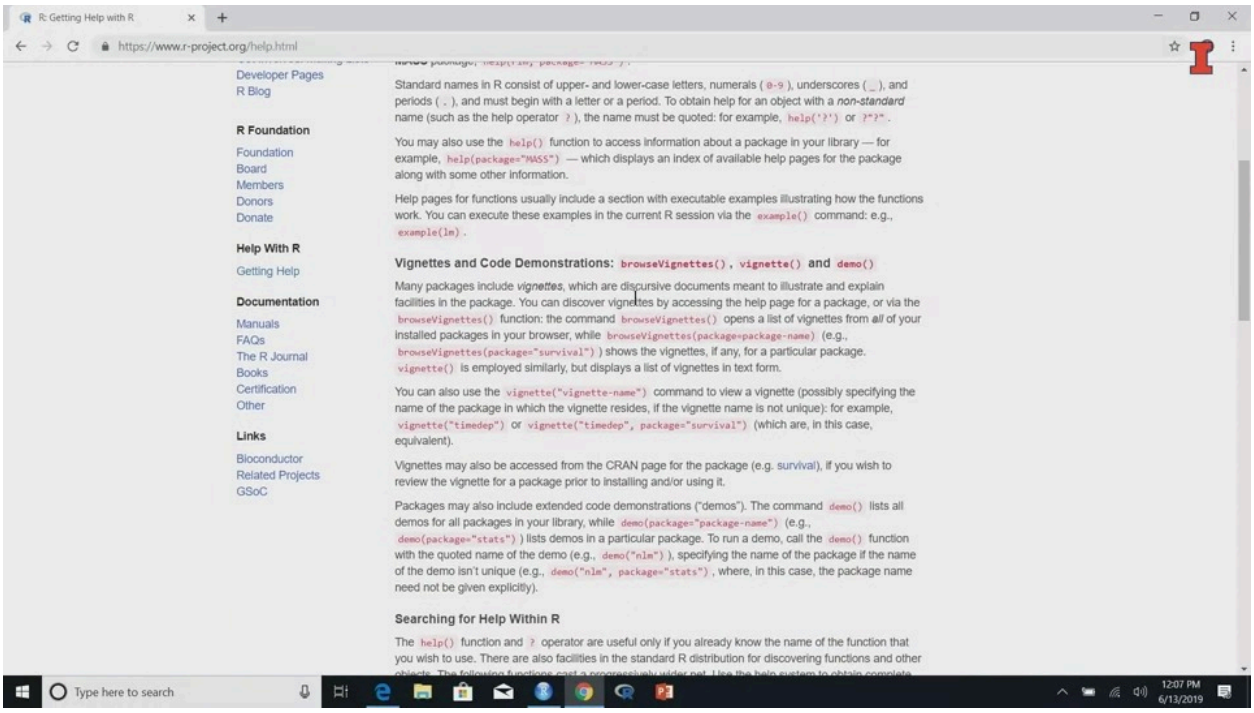
Help and Cheat Sheets

Help and Cheat Sheets

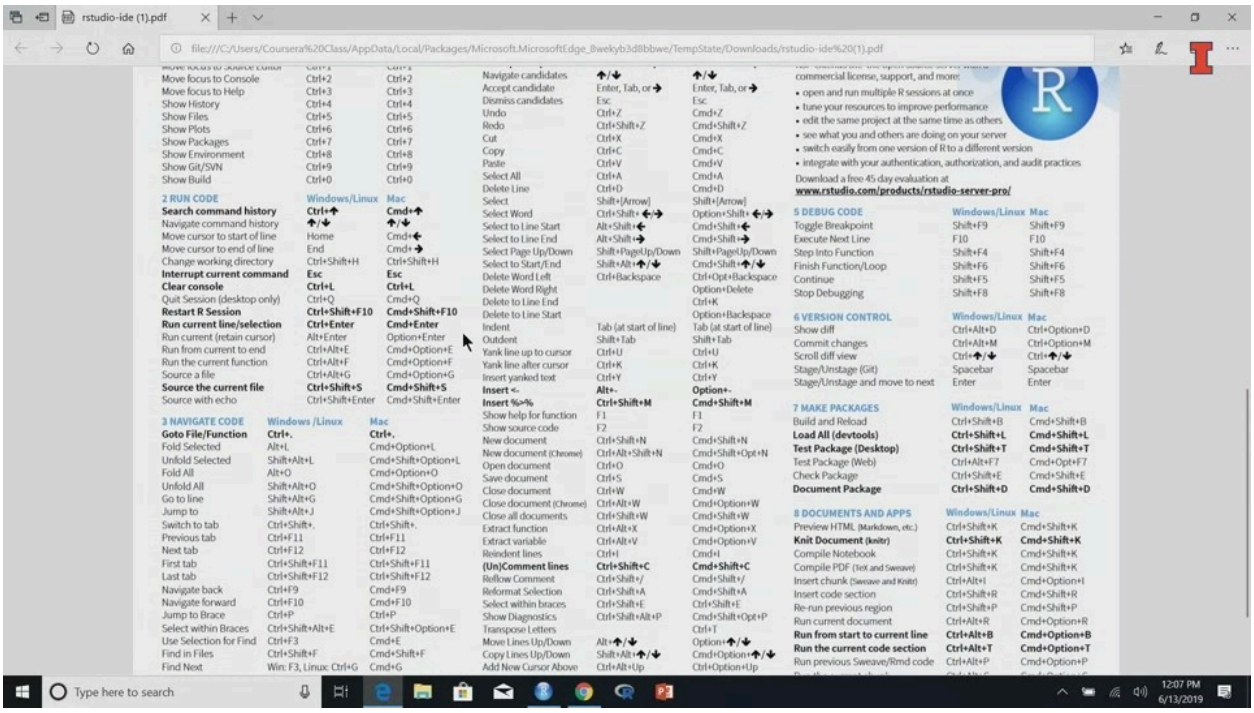
[MUSIC] So you might find yourself having trouble within our command and you need to get some help. There are a number of ways to do this. You can for example you might be having trouble with a function. Let's use the sum function which adds up a bunch of numbers but you're really unsure of how to use it. So let's let's try to get some help.



One is to use a question mark if you know the exact function name that you're using and notice here on the bottom right there is a some documentation on how to use this function. If you use question mark, question mark, sum it will do is a search of those terms sum and we'll do a broader search. The other place you can look here is that the rproject.org/ help.html.



And this gives you a list of commands that you could use to search for help searching within our our documentation or asking for help. Here's some a link to stack overflow. Another great resource. Google is another great resource. Also within the our studio environment, if you click on your help cheat sheets, there's a number of different cheat sheets that have been created by the folks at our studio. And let me click on this first one, Rstudio IDE cheat sheet and it is this pdf which has a list of all the commands and R that you can use.



In this tutorial I'm showing you the basics, but as you go along, you might want to try out these different commands. And then even tick mark than when you have mastered and learn them. Some of these are pretty straight forward, like the things you might see in Microsoft Word cut and paste, for example. And then there are some others that are more specific to the R studio environment, for example, fold all, unfold, all etc, etc. And that is how you can get some help in our R studio.