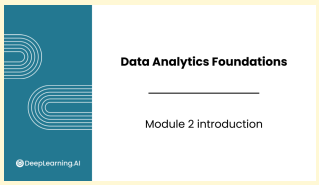
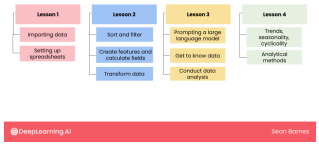
DAG C1M2 scripts

Video title	Isabel	Sean	Slides
L0V1 – Module 2 introduction	✅	✅	✅
L1V1 – Solving problems with data	✅	✅	✅
L1V2 – Spreadsheets for business analytics	✅	✅	✅
L1V3 – Navigating Google Sheets	✅	✅	✅
L1V4 – Importing data	✅	✅	✅
L1V5 – Working with structured data in Google Sheets	✅	✅	✅
L2V1 – Getting to know your data	✅	✅	✅
L2V2 – Summary statistics - MAX, MIN, AVERAGE	✅	✅	✅
L2V3 – Conditional formatting	✅	✅	✅
L2V4 – Summary statistics – COUNTIF	✅	✅	✅
L2V5 – Summary statistics – SUMIF, AVERAGEIF	✅	✅	✅
L2V6 – Summary statistics – COUNTIFS, SUMIFS	✅	✅	✅
L2V7 – Data processing – IF, IFS, RIGHT, LEFT	✅	✅	✅
L2V8 – Where does data come from?	✅	✅	✅
L3V1 – Data exploration with LLMs	✅	✅	✅
L4V1 – Introduction to time series	✅	✅	✅
L4V2 – Real-world time series	✅	✅	✅
L4V3 – Moving averages	✅	✅	✅
L4V4 – Percent change	✅	✅	✅

Introduction

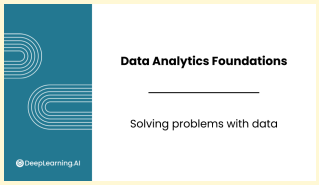
L0V1 – Module 2 introduction

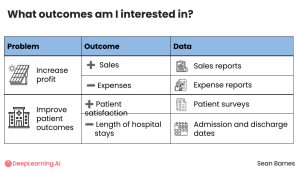
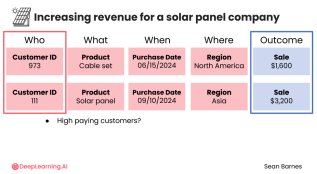
🎙️ TH	Welcome to Module 2 of Data Analytics Foundations! In this module, you'll dive into one of the most powerful and versatile tools in a data analyst's toolkit: spreadsheets. I'm excited to share with you all the incredible things you can

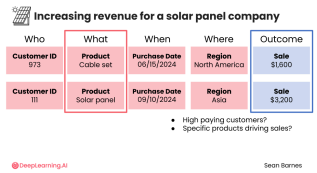
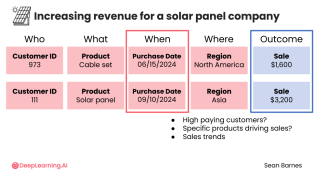
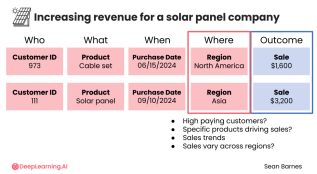
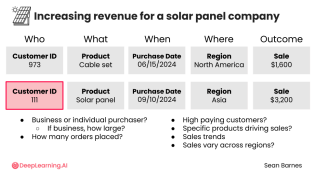
	accomplish with spreadsheets for data analytics.
	We'll start off in **[CLICK]** Lesson 1 by exploring why spreadsheets are such an effective tool for working with structured data. You'll hear how companies like Google and Netflix rely on spreadsheets every day, and you'll get hands-on practice **[CLICK]** importing data into Google Sheets and **[CLICK]** setting up your spreadsheets to enable analysis. In **[CLICK]** Lesson 2, you'll see how to process your data in spreadsheets to extract valuable insights. You'll learn how to **[CLICK]** sort and filter your data, write formulas to **[CLICK]** create new features and calculated fields, and even **[CLICK]** transform your data to make it easier to analyze. We'll work through a real-world example together: understand customer booking behaviors by analyzing hotel reservation data. In **[CLICK]** Lesson 3, you'll practice **[CLICK]** prompting a large language model to really **[CLICK]** get to know your data and **[CLICK]** conduct data analysis. **[CLICK]** Lesson 4 is all about time series data - a specific type of data measured over consistent time intervals. You'll learn to spot the key components of time series, including **[CLICK]** trends, seasonality, and cyclicality. And you'll get lots of practice **[CLICK]** with analytical methods in spreadsheets using a cool real-world dataset of popular US baby names.
TH	So get your spreadsheets ready, it's time to crunch some numbers together! By the end of this module, you'll be well on your way to becoming a spreadsheet power user. Let's jump right in with Lesson 1 and learn how spreadsheets help bring order to the chaos of raw data. See you there!

Lesson 1 – How spreadsheets organize data

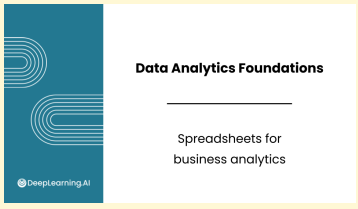
L1V1 – Solving problems with data

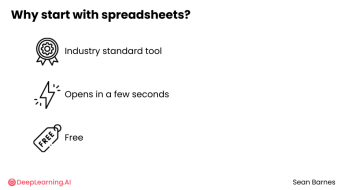
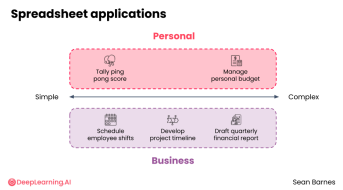
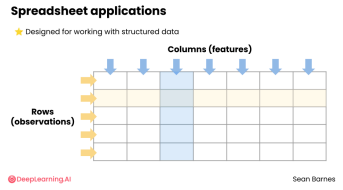
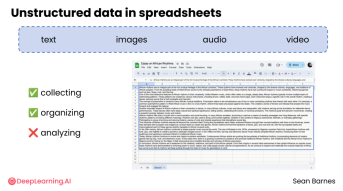
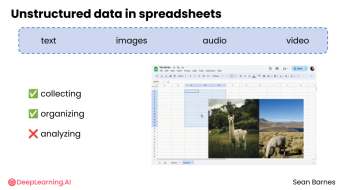
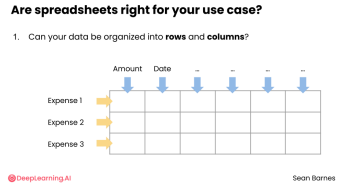
TH 	Data is a fantastic tool for solving problems in a principled way. It can help you solve problems like: how do I spend less? How long will it take to train for a marathon? You might have a hunch, but data will help you understand whether that hunch is likely to be correct. Data can help businesses too, to solve problems like how do we grow memberships faster? Should we create more action games or more platformers? The insights you create as a data analyst will guide objective

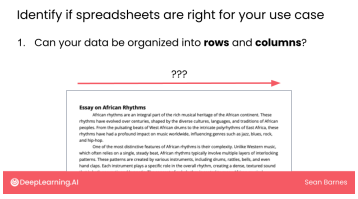
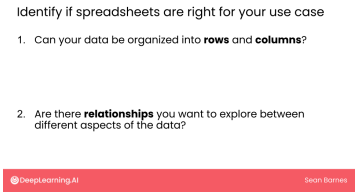
	problem solving that makes a real impact. In order to generate these powerful insights, you'll need to pick the right data to analyze, which can be very challenging. In my experience, there are two key considerations for selecting the right data. Let's take a look.
	Focusing on the problem that inspired your analysis, the first question you should ask yourself is: what outcomes am I interested in? For example, if you are trying to **[CLICK]** increase profit at a solar panel business, then you will want to see a **[CLICK]** positive change in sales, or a **[CLICK]** negative change in expenses. You can pull your **[CLICK]** sales and expense reports for this analysis. Or suppose you're working with a hospital to **[CLICK]** improve patient outcomes ~~after hospital operations~~. In this case, your outcomes of interest might be an **[CLICK]** increase in patient satisfaction and a **[CLICK]** decrease in the length of hospital stays. Starting from these outcomes, you can gather data from **[CLICK]** patient surveys to analyze patient satisfaction, and **[CLICK]** admission and discharge dates can help you figure out whether hospital stays are increasing or decreasing in length.
	Next, identify data that provides context for your outcomes of interest. By "provide context", I mean that this data tells you more information about the outcomes that were observed, such as the 4 Ws: the **[CLICK]** who, **[CLICK]** what, **[CLICK]** when, and **[CLICK]** where. For example, if your **[CLICK]** outcome of interest is sales data, which data points provide context for those sales? Perhaps your sales data is associated with a specific **[CLICK]** customer ID (the who), **[CLICK]** product (the what), **[CLICK]** purchase date (the when), and **[CLICK]** region (the where). All of this information can help contextualize the sales data, so you can compare sales data across different products and regions.
	Let's zoom in on the example of increasing revenue for a solar panel company, and suppose that we only had **[CLICK]** the sales data on the right-hand side. Remember, sales is your outcome of interest. It's great that we have sales, but unfortunately, we don't know a lot more than that. We need *context* about these sales so that we can better understand the driving factors behind the sales. For the solar panel sales, context could look like **[CLICK]** Customer ID 973 purchased a **[CLICK]** cable set on **[CLICK]** June 15 2024 in **[CLICK]** North America. Here is **[CLICK]** another observation to give you a sense of the range of values. **[BRIEF PAUSE FOR LEARNER TO READ]** In this case, you can use the other data points to answer questions like:

	• **[CLICK]** Are there any high-paying customers who should be targeted for additional purchases? Customer ID and sales can help answer this question.
	• Are there specific products that are driving a high proportion of sales? Product and sale will be relevant here.
	• How are sales trending over time? This time, the purchase date combined with sale will answer my question.
	• Does that total amount of sales vary across regions? Now region and sale help with the analysis. It would be impossible to answer any of these questions without this contextual data. It's just as important as the outcome of interest. You can actually dig quite deeply into each of these data points, too.
	For example, you might notice that Customer 111 made the largest purchase. If you want to understand why, you could ask questions like: **[CLICK]** is this customer a business or an individual? If it's a business, **[CLICK]** how large is it? **[CLICK]** How many orders have they placed? Remember that ultimately it's your job as a data analyst to give these data points meaning by connecting them to the broader business goal.
TH	You've now seen how to identify useful data in order to tackle a business problem. What tools can you use to organize and analyze that data once you've identified it? In the next video, you'll learn more about how spreadsheets can be a powerful ally in the world of data analytics.

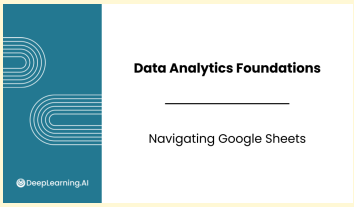
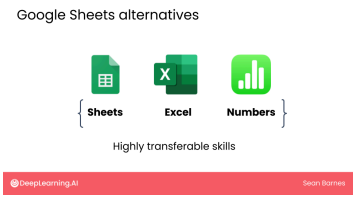
L1V2 – Spreadsheets for business analytics

Slide	Script
TH 	Spreadsheets bring interactivity to structured data. They are an industry standard tool that I have used consistently throughout my career, even as I have learned more complex tools.

	Not only is the spreadsheet an **[CLICK]** industry standard tool used regularly at Google, Netflix, and more, but you can open up a spreadsheet right now – in **[CLICK]** just a few seconds and for **[CLICK]** free. They have a broad range of use cases, whether your goal is to analyze your household finances or calculate year-over-year revenue growth for a company.
	Spreadsheet applications range from the very **[CLICK]** simple to the fairly **[CLICK]** complex and for both personal and business use cases. For example, on the **[CLICK]** personal side, you can utilize spreadsheets to **[CLICK]** tally scores in ping pong or **[CLICK]** manage your personal budget. On the **[CLICK]** business side, spreadsheets can be used to **[CLICK]** schedule employee shifts, **[CLICK]** develop a project timeline, or **[CLICK]** draft a quarterly financial report. The use cases are nearly endless, as long as you have a mechanism to collect the data in a structured fashion and store it.
	Spreadsheets are **[CLICK]** designed primarily for working with structured data, which as you learned in Module 1 means data that can be organized into **[CLICK]** rows representing the observations of your data and **[CLICK]** columns that represent various features. **[BRIEF PAUSE BEFORE MOVING ON]**
	When it comes to unstructured data in spreadsheets, like **[CLICK]** text, **[CLICK]** images, **[CLICK]** audio, and **[CLICK]** video, spreadsheets can be used for **[CLICK]** collecting and **[CLICK]** organizing it, but their capabilities for **[CLICK]** analyzing this type of data are limited. ~~They are just not designed to handle it in a meaningful way.~~ Imagine trying to **[CLICK]** write an essay in this interface **[PAUSE TO WATCH GIF FOR LIKE 2 SECONDS]**
	Or organize your photos. **[PAUSE TO WATCH GIF FOR LIKE 2 SECONDS]** **[STOP GIF FROM PLAYING]** Where would you even start? It would probably make your task a lot harder. So if you identify a need to work with unstructured data, you may need to rely on a computer programming language such as Python or generative AI tools, both of which you should explore as you expand your data analytics toolkit.
	Here are two questions you can ask yourself to identify if spreadsheets are right for your use case. First, **[CLICK]** can your data be organized into **[CLICK]** rows and **[CLICK]** columns? ~~That is, are you working with structured data?~~ This organization is fundamental to spreadsheets. A budget can be organized into **[CLICK]** one row for each expense, and **[CLICK]** columns for features like the amount, transaction date, and so on.

	Meanwhile, unstructured data like a **[CLICK]** essay can't easily be organized in the same way. What are the **[CLICK]** columns of an essay? It just doesn't work. **[BRIEF PAUSE TO UNDERSTAND VISUAL]**
	**[CLICK] [CLICK]** Second, are there relationships you want to explore between different aspects of the data? Spreadsheets can effectively calculate these relationships, like **[CLICK]** organizing all the expenses in your budget by category or analyzing purchases to **[CLICK]** find the month where you spent the most. If the answers to both questions are an enthusiastic yes, then spreadsheets will be a fantastic tool for solving the problem you're trying to tackle.
🎙️ TH	Now that you've seen how powerful spreadsheets can be, I hope you'll join me in the next video to get hands on with a renovation project in Google Sheets.

L1V3 – Navigating Google Sheets

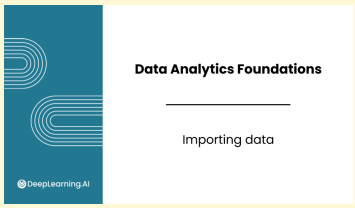
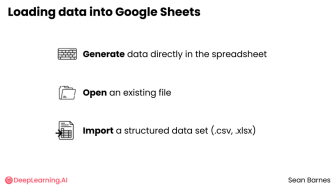
Slide	Script
🎙️ TH 	One of the most common spreadsheet applications is Google Sheets. It's both accessible and useful. Best of all, it's available for free to individuals, and you can share your spreadsheet to collaborate with friends and teammates.
	While Google Sheets is widely used, you have other options, such as **[CLICK]** Microsoft Excel, **[CLICK]** Apple's Numbers, and more. The skills that you will develop in this course using Google Sheets are **[CLICK]** highly transferable to each of these other tools ~~should you find yourself working with them.~~ Let's take a tour of Google Sheets to see how it works.

🔗 **DEMO SPREADSHEET**
🔗 **RENOVATION NOTES**
	• 💬 You can create a new sheet by going to sheets.new, as long as you're signed into your Google account. You can see the instructions in the following reading item for more
	  ○ ▶️ Go to sheets.new
	  ○ ▶️ Name file "Home renovation project"
	• 💬 You have all these options up here on the ribbon, you'll learn these as you go!
	• 💬 My friends have been helping me with my home renovation. I

	have some data here and some questions about it. ○ ▶️ Show renovation notes document overlaid on spreadsheet ○ ❓Just looking at this data, did I go over budget? ○ ❓What item was the most over or under? ○ ❓What was the first transaction? ○ ❓How much did Joi spend? • 💬 Each one of these transactions I've been taking notes on has several features. This could fit into rows and columns! ○ ▶️ A10 → Copy/paste in one section ○ ▶️ Turn the data into a table • 💬 Each of these intersections between a row and column is a "cell." Notice that I can click a cell and I can click into a cell. ○ ▶️ Show how to → select multiple cells ○ ▶️ Show how to → select multiple rows/columns ○ Navigating with the arrow keys • 💬 Let's make it pretty. Don't focus too much on what I'm doing, you'll learn all this in the coming videos, I just want to show you some cool capabilities ○ ▶️ Header → Bold ○ ▶️ Header → Add background color ○ ▶️ Header → Add bottom border • 💬 I'll add all the data ○ Copy all rows from solutions sheet • 💬 Now I can answer, did I go over budget? ○ ▶️ Summary for budget & cost in bottom right ■ ▶️ Navigating with the arrow keys • Shift plus arrow keys selects multiple cells ○ ▶️ Show =sum() for both in a total row ○ ▶️ In A14, "total" ○ ▶️ Bold total row ○ ▶️ Add top border to total row • 💬 Ah, I just remembered another transaction for the Whale soap holder ○ ▶️ Add in a new row, show how column sums update • 💬 Now I want to answer, what was the first transaction? ○ ▶️ Sort by date • 💬 And what did Joi buy? ○ ▶️ Filter just to Joi, show what she bought ○ ▶️ Select all in filter again • 💬 And what were the labor costs? ○ ▶️ Filter to just Labor ○ ▶️ Select all in filter again • 💬 Now I'm wondering, what was the most over budget purchase? ○ ▶️ New column =C2-D2 ○ ▶️ Fill handle drag

	○ ▶️ Sort by this column to show most over/under purchase ● 💬 I've answered all my questions! What a powerful interface 😀
🎙️ TH	Now you're in a fantastic position to work with any spreadsheet in Google Sheets. Join me in the next video to see how to import data. I'll see you there.

L1V4 – Importing data

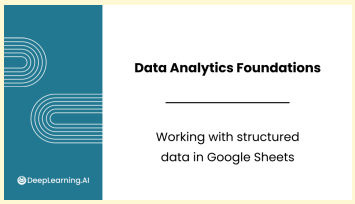
🎙️ TH 	There are several common ways to load data into Google Sheets. It depends on your use case. Let's see how.
	● First, you can **[CLICK]** generate data directly in the spreadsheet. Essentially typing in your data. It's pretty common for small scale personal applications and you just saw that approach in the previous video's home renovation budget. ● You can also **[CLICK]** open an existing file. You'll use this approach when you've already been working with some data in Google Sheets and you just want to pick up where you left off. ● Most often, you are going to access data using the third approach, which is **[CLICK]** importing a structured data set. Most data starts out in a csv or xlsx file rather than a Google Sheet. These are two types of spreadsheet files that work with most software. It's helpful to see this in action, so let me show you how it's done.
🔗 **SCREENCAST**	Quickly, if I wanted to re-open my budget from the previous video here, I can go to docs.google.com/spreadsheets and either select it – it's my most recent as I was just working on it! – or search for it.
🔗 **SCREENCAST**	● 💬 Now most of the time you'll be working with more complex data sets ● 💬 For example, if you're studying booking patterns for a hotel, you might look for a dataset that's already been collected, like this one. You'll work with this data throughout this lesson and the next, it's quite interesting. ○ ▶️ Show abstract ● 💬 This article has two datasets related to hotel demand ● 💬 31 variables and 40,060 observations. Wow, I don't want to be typing that by hand ● 💬 This is real hotel data, and it's been anonymized. ○ 💬 By the way typos here are likely from the translation, as

	this data is from Portugal • 💬 This data is actually publicly available ◦ ▶️ Scroll down to “Transparency document. Supplementary material”

🔗 **DEMO SPREADSHEET**	• 💬 I want to work with this data in Google Sheets. First I’ll go to sheets.new. I have downloaded it from the previous site, you can as well if you’d like to follow along
	  ◦ ▶️ Go to sheets.new
	• 💬 It’s large, so we’ve created a smaller version that is a subset of these reservations and columns to make it easier to work with. Automatic detection of the data types typically works
	  ◦ ▶️ File → Import → Upload → select dataset
	  ◦ ▶️ Replace current sheet → Detect automatically
	• 💬 Now the data has appeared! I’ll review it.
	  ◦ ▶️ Rename sheet → Hotel reservations
	  ◦ ▶️ header → background color
	  ◦ ▶️ header → bottom border
	  ◦ ▶️ add filter
	• 💬 Looks great! If it didn’t, you could try correcting the original file or your columns here.
	• 💬 There are a lot of rows. One trick is to freeze the top row
	  ◦ ▶️ Format → freeze → 1 row
	  ◦ ▶️ select column a → count is 36k observations
	• 💬 I want to share with my collaborators on the data team. This is a major plus for Google Sheets
	  ◦ ▶️ Share → add someone via email **[blur email(s)]**
	• 💬 If you wanted your analysis to be public, you could put share to anyone with link, can copy link to message it to someone
	  ◦ ▶️ Copy link
	• 💬 Omg I closed my file!! Oops
	  ◦ ▶️ Paste in link, go
	• 💬 The good news is, it’s automatically saved! You can also see other versions. If you make a mistake you can go back to the previous one.
	  ◦ ▶️ Go to revision history **[blur revision history]**
	  ◦ ▶️ select previous version
	  ◦ ▶️ Go back to revision history
	  ◦ ▶️ Select most up to date version again **[blur revision history]**
	
	Now you’re ready to start analysis!

🎙️ TH	Now that you’ve seen how to import data, you can work with any data set on the internet. 😀 Join me in the next video to work with powerful sorting, filtering, and analysis techniques. I’ll see you there.

L1V5 – Working with Structured Data in Google Sheets

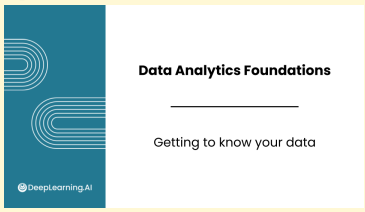
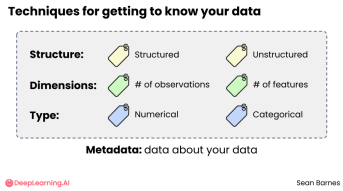
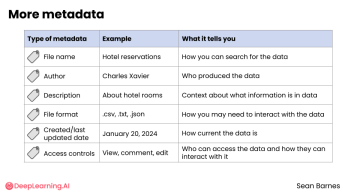
🎙️ TH  	Now that you've imported this massive real world data set, let's investigate it a bit further. You've been tasked with examining hotel bookings to understand which bookings are the most profitable, when they occur, who's making them, and anything else you can find out. When you're exploring a dataset your first goals should be to understand it and make sure it's formatted correctly. Let's do that!

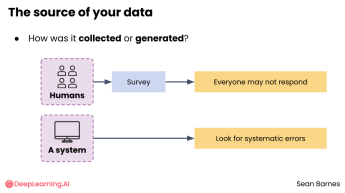
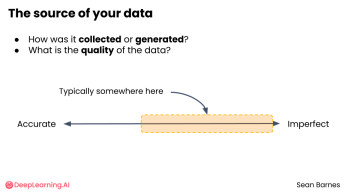
🔗 **DEMO SPREADSHEET**	• 💬 First, let's review the structure. One observation per row. Each observation is a booking. "Cell" is the intersection of row and column.
	• 💬 There's a lot of info here. Let's review some important features – arrival year day month, children, adults, required car parking space, lead time, average price per room, booking status **[the point is so learners don't have to read all the text every time they look at this]**
	• 💬 Mix of numerical & categorical features. Notice something like required car parking space seems categorical
	  ▪ In G36278 → =unique(G2:G36276)
	• 💬 Say you want the date on the left.
	  ○ ▶️ Select J shift L → move between A and B
	• 💬 As easy as that!
	• 💬 Notice these years are all mixed up. Say I want to order the rows by date. Unlike reordering columns, organizing the rows of your data is not typically a manual exercise, since you are likely to have many more rows than columns.
	  ○ ▶️ Sort alphabetical → a to z
	  ○ ❓ How do you think we could sort by month?
	  ○ ▶️ Select all data → Sort → Sort range → ✅ has header → arrival year A to Z → arrival month A to Z → arrival date A to Z
	• 💬 This is a tad inconvenient, isn't it?
	  ○ ❓Why do you think it might be stored this way?
	• 💬 Maybe you want to conduct analysis across months, like examining summer months. But a date column could be helpful 😀
	• 💬 Say you're only interested in repeat bookings. You can filter your data.
	  ○ ▶️ Column N → filter → 1
	  ○ ▶️ Select N:N → count summary (931)
	  ▪ Ctrl click to remove the top cell
	• 💬 Only 931 out of 36,000. I notice a lot of these are corporate customers!
	• 💬 While I'm here, what do you notice about average price per room? **[pause for thought]** It should be in Euros, based on the dataset.

	○ ▶️ Format → number → custom currency → Euro • 💬 I don't see any other special formats, but you'll see a lot more tools in the next lesson!
🎙️ TH	Great work developing your spreadsheets skills. It's fun right?? You have a lot of capabilities already. Once you've finished the practice assessment for this lesson, join me in the next one to learn more about where data comes from, how to get to know it, and how to write formulas in spreadsheets to enable meaningful analyses.

Lesson 2 - Spreadsheet fundamentals

L2V1 – Getting to know your data

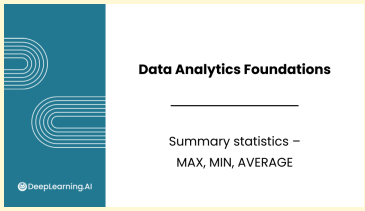
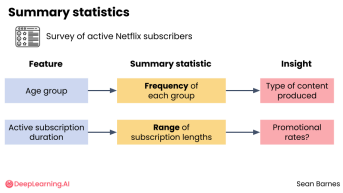
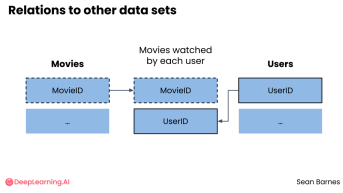
Visual	Script
🎙️ TH 	Before you can perform impactful analysis, you must get to know your data. The first time you open up a data set, it's kind of like meeting someone new. Datasets have history and personality! What are some questions you might want to know about a new acquaintance? Maybe how old they are, where they're from, what kind of work they're in.
	You've already seen some strategies for getting to know your data: determining whether it's **[CLICK]** structured or unstructured, counting the **[CLICK]** number of observations and features, and **[CLICK]** differentiating between numerical and categorical features. This type of information about your data is called **[CLICK]** *metadata*, or data about your data, which is a very 'meta' concept 🙂.
	Let's look at some more metadata that you'll commonly encounter, and I'll show you the **[CLICK]** type of metadata, an example, and what that metadata can tell you. • First you have the **[CLICK]** file name of the data set, such as **[CLICK]** "Hotel Reservations", which tells you fundamentally **[CLICK]** how you can search for or find the data • The original **[CLICK] [CLICK]** author of the data, which tells you **[CLICK]** who produced the data, so you can follow up with questions • A **[CLICK] [CLICK]** description of the data, which **[CLICK]** provides

	context about what information is contained in the data • The **[CLICK]** file format (**[CLICK]** .csv, .txt, .json, etc.), which informs **[CLICK]** how you may need to interact with the data • **[CLICK] [CLICK]** When the file was created, or last updated, which tells you **[CLICK]** how ~~old or~~ current the data is • **[CLICK] [CLICK]** Access controls, which tells you **[CLICK]** who can access the data and how they can interact with it
	You'll want to understand the *source* of your data, or its origin story: • **[CLICK]** How was it collected or generated? Was it generated by **[CLICK]** humans or was it generated by **[CLICK]** a software system? If you know your data was collected by **[CLICK]** survey, you'll want to consider the fact that **[CLICK]** everyone you sent the survey may not actually respond. If you know it was collected via software, you may want to **[CLICK]** look for systematic errors.
	• What is the general quality of the data? Is it **[CLICK]** accurate or are there **[CLICK]** imperfections? Typically it's **[CLICK]** the latter, and understanding the source may help you identify likely issues. Let's ask some of these questions to explore the origin story and personality of the Hotel Reservations data set you met in the previous lesson.

🔗 **SCREENCAST**
🔗 **SCREENCAST - MAP**
	• 💬 Here's the journal article with the source of this data. There's a lot of text here, let me walk you through it.
	• 💬 How old is this data?
	  ○ ▶️ Highlight volume date at top of page
	• 💬 Looks like it was published in February 2019, so you shouldn't expect bookings newer than that.
	• 💬 Let's read on. [go through abstract, highlight interesting things – location, type of hotel, observations, identity]
	  ○ ▶️ Highlight as you read
	• 💬 Where does this data come from?
	  ○ ▶️ Scroll down to specifications table
	• 💬 For data source location, I see that both hotels are in Portugal, H1 at the resort of Algarve and H2 in Lisbon
	  ○ ▶️ Open map, show Faro region and Lisbon
	• 💬 They're far from each other and may have different characteristics.
	• 💬 How was this data collected? [read rest of specifications, explain your thoughts]
	  ○ ▶️ Highlight "how data was acquired"
	• 💬 Extraction from the PMS is a signal that this data is reliable. The records may be generated automatically with minimal human error. It also signals that this data is owned by the hotels themselves.

	Automatic extraction also often leads to large datasets like this: 36,000 observations spanning two years, so over 50 bookings per day.
TH	Great work analyzing the source of your data! Follow me to the next video to see how to explore some key summary information.

L2V2 – Summary statistics – MAX, MIN, AVERAGE

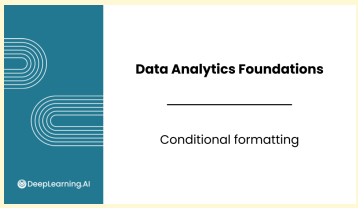
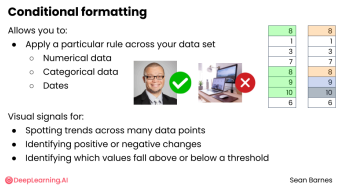
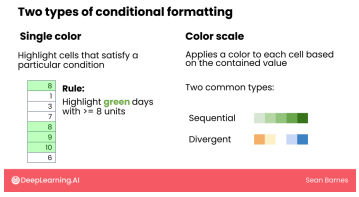
Visual	Script
TH  	You know where your data is from and how it was collected, but do you know what's in it? It's fun to explore, looking for trends and unusual values that can help you answer your business question.
 	Once you know your data's source, you should calculate a few summary statistics to better understand the features. Consider a **[CLICK]** survey of active Netflix subscribers 🙂 You may have a feature for **[CLICK]** age group like 18-24, 25-34, and so on. The **[CLICK]** frequency of each group forms the subscriber age distribution. Does your subscriber base lean younger or older, and how does that information factor into the **[CLICK]** type of content you produce? Your data set might also include **[CLICK]** active subscription duration. What is the **[CLICK]** range of these durations? Perhaps the minimum subscription length is 3 months because you offered a **[CLICK]** promotional rate for new signups. And maybe the longest one is only 2 years because that goes back to when you first launched the service. You'll also want to understand the typical duration, so you can figure out how to lengthen it. You should also understand **[CLICK]** relationships between features. For example, consider the relationship between **[CLICK]** age group and active subscription duration. **[CLICK]** How does duration vary across different age groups? Why?
 	Your data set may also relate to other data sets, often connected by one or more common features. For example, a dataset of movies **[CLICK]** may relate to a dataset of **[CLICK]** users, linked based on **[CLICK]** movies watched by **[CLICK]** each user. Let's take a look at some of the features in the hotel reservations dataset

	to get to know it better.

🔗 **SCREENCAST**	• 💬 Let's explore some of these features, starting with number of adults. I'll create a new sheet to store the values
	  ○ ▶️ Copy top 2 rows from solution sheet
	  ○ ▶️ Delete formulas
	  ○ ▶️ In B2 → =min('🚀 Start here - Data'!E:E)
	  ○ ▶️ In C2 → =max('🚀 Start here - Data'!E:E)
	  ○ ▶️ In D2 → =average('🚀 Start here - Data'!E:E)
	  ○ ▶️ Reduce average to 2 decimals
	• 💬 Minimum number of adults is 0? That's odd. How can a booking have zero people? Well, there are children too! How many children only bookings?
	  ○ ▶️ Sort E → A to Z
	  ○ ▶️ Highlight top 139 rows → bottom right summary → count
	• 💬 139 out of these 36,000 bookings were just for children, only with 2 or 3 of them. Could it be a mistake?
	• 💬 So let's look at children then. I'll copy over the same formulas this time for column F.
	  ○ ▶️ Copy row 3 from solution sheet
	• 💬 Min of zero, okay makes sense, but who had 10 children in a room?? The average is quite low, though. Most bookings have none.
	  ○ ▶️ Sort F → Z to A
	• 💬 Okay only 1 has 10, but 2 have 9, and then it goes down to 3
	• 💬 Last up, check out lead time. This is the number of days before the check in time that the reservation was booked. I'll copy over the same formulas for column L
	  ○ ▶️ Copy row 4 from solution sheet
	• 💬 The max is quite a bit higher than the mean. Surprised at the last minute bookings! Let's investigate bookings over 400 days
	  ○ ▶️ Filter lead time → clear all → 418, 433, 443
	• 💬 This looks odd, doesn't it? Bookings above 400 days are rare, yet there are only 3 values represented. Perhaps it was a promotion? It might be worth looking into further.
	• 💬 It's valuable to take a look at features this way! I encourage you to check out more of them. As you advance your skills, you'll work with programming languages that can give this type of summary quickly.

🎙️ TH	So you have a good sense of what's going on in the hotel reservations dataset. How can you analyze what's happening here? Join me in the next few videos to see some cool techniques for analyzing your data, starting with conditional formatting.

L2V3 – Conditional formatting

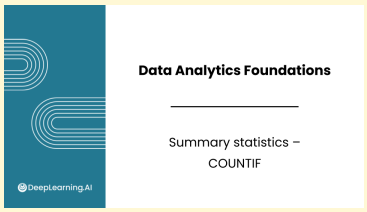
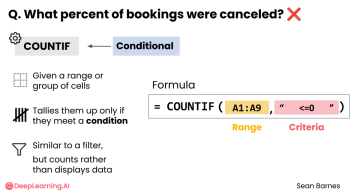
Slide	Script
TH 	One powerful visual tool for quickly understanding your data is conditional formatting. It can help you discover meaningful insights more easily than just looking at raw data. Let's see how it works.
	Conditional formatting in a spreadsheet looks like this **[CLICK]** or this **[CLICK]** and **[CLICK]** allows you to efficiently **[CLICK]** apply a particular rule across your data set, including for **[CLICK]** numerical data, **[CLICK]** categorical data, and **[CLICK]** even dates. Its main benefit is that it provides a visual layer on top of your data, so that you don't have to mentally evaluate how each data point compares with the rule. As a **[CLICK]** human and not a **[CLICK]** computer, I appreciate the **[CLICK]** visual signals that conditional formatting provides, because it makes it easier to identify patterns in data, such as: • **[CLICK]** Spotting trends and patterns across many data points • **[CLICK]** Identifying positive/negative changes • Identifying outliers, or • **[CLICK]** Identifying which specific values fall above or below a threshold There are many types of rules that you can apply to your data!
	The two main types of conditional formatting are **[CLICK]** single color, which highlights yes/no conditions, and **[CLICK]** color scale, which shows a range of values in a range of colors. • Apply **single color** formatting when you want to **[CLICK]** highlight cells that satisfy a particular condition. For example, you could **[CLICK]** highlight the days with 8 or more solar panels sold, which you may classify as a great day. Applying conditional formatting would enable you to easily identify which days were great. Single color formatting also allows you to choose font styles. • Your other option is a **color scale**, which **[CLICK]** applies one of several colors to each cell based on its value. You can't apply other styles like bold and italics, since there's no clear way to scale those across many values. There are **[CLICK]** 2 common types of color scales, including **[CLICK]** **sequential** and **[CLICK]** **divergent**. Sequential uses increasingly darker shades of the same color, while divergent uses different colors on each side of a central value. Let's take a look at each of these in a spreadsheet, so you can see where

	each one is useful.

🔗 **SCREENCAST**	Let's see how each of these color scales could work in the Hotel reservations data set
	• 💬Identify the most valuable bookings w/ conditional formatting. One idea: does booking have children? You saw that bookings with children are relatively rare (average 0.11) Two conditions: above 0 children or not → single color scale.
	  ○ ▶️F:F → Apply single color
	  ○ ▶️Blue
	  ○ ▶️Bold
	• 💬Spot unusual values. Bookings with children are relatively rare
	• 💬Another idea is to show the range of lead times, days booked in advance. That may help you identify at a glance unusual values. Choose a scale that emphasizes the higher values
	  ○ ▶️L:L → Apply color scale
	  ○ ▶️Green
	• 💬Interesting! Now for example if I filter by corporate, I can see there is less notice (lighter colors)
	  ○ ▶️ M → Filter → clear all → corporate
	  ○ ▶️ remove filter
	• 💬Next, you think of looking at the average price per room. Say your breakeven point is 45 Euros, so any lower than that and you're losing money, and the lower the price is below that you lose more money. Above that is profit, and higher profit is better. You can use a diverging color scale.
	  ○ ▶️Apply a color scale
	  ○ ▶️Q:Q
	  ○ Midpoint: number → 45
	• 💬I could use these options with red and green, but those may be hard to see for folks who are colorblind. I'll make a custom one
	  ○ ▶️Low values – orange
	  ○ ▶️High values – blue
	• 💬Most of these rooms are profitable! Lighter values are near the breakeven point, while darker green ones show more profit. I notice one low booking which was online with one adult for one night. And some very low value bookings of 0 and 1 that all appear complementary.

🎙️ TH	Great work applying conditional formatting to your data set! Conditional formatting is quite powerful for exploring your data and communicating insights. Now that you've seen how to apply it to real-world data, follow me to the next video to see how you can extend these insights to summarize data in a spreadsheet.

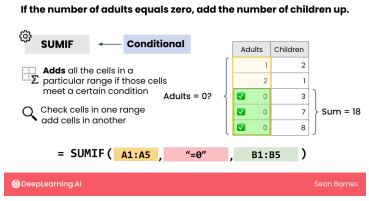
L2V4 – Summary statistics – COUNTIF

TH  	When a data set has a lot of features to analyze, like the hotel reservations data set, you might wonder where to even begin? One strategy is to segment your data to try to understand the different underlying groups. Let's see how, starting with the outcome variable, booking status.
 	Here's a question, what percent of bookings were canceled? It might be difficult to tally up using your current tools. You can use the **[CLICK]** COUNTIF function to help answer this question. **[CLICK]** Given a range of cells, or a group of cells, COUNTIF **[CLICK]** tallies them up only if they meet a certain condition. COUNTIF is **[CLICK]** similar to a filter, which only displays data that meets a certain condition, **[CLICK]** except it counts up that data rather than displaying it. So COUNTIF is a type of **[CLICK]** **conditional** – take some action only if a certain condition is met. Your **[CLICK]** formula will look like this. • First, **[CLICK]** an equals sign. Formulas have to start with an equals sign; if you don't include it, whatever you type here will typically be treated as plain text. • Then **[CLICK]** COUNTIF, the name of the function, **[CLICK]** open parentheses. COUNTIF has two inputs, **[CLICK]** range and **[CLICK]** criteria. • **[CLICK]** Select the range, • And then in quotes add the criteria. Example criteria could be **[CLICK]** "Hot pocket", if you want to count the number of cells in the range that contain simply "Hot pocket". Or, if you're checking numbers you could have **[CLICK]** ">100" or **[CLICK]** "<=0". Let's see that function in action.

🔗**DEMO SPREADSHEET**	• 💬It's good practice to separate summary statistics from your data, let's build on the previous demo's statistics. And our idea is to tally up the number of canceled and not canceled reservations. I'll stay organized with this table
	  ○ ▶️Copy over A6:C8 from solution – summary statistics
	• 💬 First, count the number of canceled reservations. First copy the "Canceled" so I know exactly how it's spelled, good trick.
	  ○ ▶️ B7 → =COUNTIF(S:S, "Canceled")
	  ■ ▶️ Show help menu for COUNTIF
	• 💬 This hotel was dealing with 12,000 cancellations in just two years. That's almost 17 cancellations a day. How about not

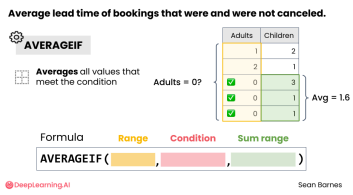
	canceled? ○ ▶️ Copy Sheet1:S2 ○ ▶️ B7 → =COUNTIF(S:S, “Not_Canceled”) • 💬 Twice as many! ○ ❓ If I sum up these two values, what should I get? ○ ▶️ B9 → =SUM(B7:B8) • 💬 The total bookings! • 💬 A note about COUNTIF and formulas in general, case doesn’t matter but characters do. “not_canceled” works, but “not canceled” gives 0 ○ ▶️ B7 → =COUNTIF(S:S, “not_canceled”) ○ ▶️ B7 → =COUNTIF(S:S, “Not Canceled”) ○ ▶️ B7 → =COUNTIF(S:S, “Not_Canceled”) • 💬 It’s a little hard for me to gauge what that means. I’d like to see the percent of all bookings that are cancellations. I don’t know about you, but I’m not going around dividing 11,000 by 36,000 all the time in my head. ○ ▶️ C7 → = B7/B9 • 💬 This is a proportion out of 1, you can multiply by 100, but an easier way is to format as percentage ○ ▶️ C7:9 → percentage ○ ❓ Can you guess the formula for C8? ○ ▶️ C8 → = B8/B9 ○ ❓ And what should these sum to? ○ ▶️ C9 → =SUM(C7:C8)
🎙️ TH	Cool summary! There are a lot of cancellations, perhaps more than you would’ve expected. I mean, I do cancel stuff a lot 👀 In any case, follow me to the next video to sum up the number of children who stayed with and without an adult!

L2V5 – Summary statistics – SUMIF, AVERAGEIF

🎙️ TH	Say you want to look at the relationship between children and adults staying in the hotel. You want to sum up the number of children who stayed on their own without an adult. Let’s see a powerful way to do that.
	In other words, what you want to do is, **[CLICK]** “If the number of adults equals zero, add the number of children up.” For this purpose, you can use **[CLICK]** SUMIF, another useful function. SUMIF **[CLICK]** adds up all the cells in a particular range if those cells meet a certain condition. You can also **[CLICK]** check cells in one range and **[CLICK]** add cells in another. Like COUNTIF, SUMIF is also a **[CLICK]** **conditional** – sum these values only if a certain condition is met.

	[CLICK] Here's how it works: • First, an **[CLICK]** equals sign. • Then **[CLICK]** SUMIF, the name of the function,**[CLICK]** open parentheses. SUMIF has three inputs, **[CLICK]** range, **[CLICK]** condition, and **[CLICK]** sumrange. • **[CLICK]** Select the range you want to check. This is what comes right after your "if", so in this case **[CLICK]** "number of adults" • And then in quotes add the criteria. **[CLICK]** You want "=0". • And then **[CLICK]** the column you want to add up is **[CLICK]** children. Let's see that in action.

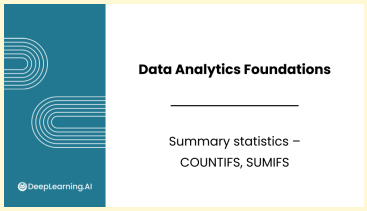
Continue from previous spreadsheet	• 💬 Let's sum the number of children who stayed with and without an adult.
	  ○ ▶️ Copy over A11:C14, remove formulas
	  ○ ▶️ B12 → =sumif('🚀 Start here'!E:E,"=0",'🚀 Start here'!F:F)
	• 💬 About 282 children stayed on their own. How does that compare with accompanied children? Sum if adults is greater than 0
	  ○ ▶️ B13 → =sumif('🚀 Start here'!E:E,">0",'🚀 Start here'!F:F)
	  ○ ▶️ B14 → =sum(B12:B13)
	• 💬 What percent is 282 children?
	  ○ ▶️ C12 → =B12/B14; fill out C13 and C14
	  ○ ▶️ Format as percent
	• 💬 Interesting! So now you've segmented your data about children based on whether they were with an adult or not.

	Say you wanted to take the **[CLICK]** **average** of one column rather than the sum based on a particular condition. For example, you might want to investigate the **[CLICK]** average lead time of bookings that were and were not canceled. For that purpose, you can use SUMIF's cousin **[CLICK]** AVERAGEIF. AVERAGEIF has the exact same **[CLICK]** inputs as SUMIF, except it will take the **[CLICK]** average of all the values that meet the condition rather than adding them all up. Let's get right into it.

Continue with same demo sheet	• 💬 Maybe lead time can give us a hint about what kind of person cancels.
	  ○ ❓What's your prediction? Do you think people who cancel tend to book farther in advance or closer to the date?
	• 💬 Here's how you can calculate that.
	  ○ ▶️ Copy over A16:B19 from solution, remove formulas
	  ○ ▶️ B17 → =averageif('🚀 Start here'!S:S,"Canceled",'🚀 Start

	here'!L:L) ○ ▶️ Decrease decimals • 💬 So canceled bookings on average are booked 139 days in advance. How does that compare with ones that are not canceled? ○ ▶️ B18 → =averageif('🚀 Start here'!S:S,"Not_Canceled",'🚀 Start here'!L:L) • 💬 About 80 days less! An average of 59 days. That's a big difference. What's the overall average? This will weight more heavily towards not canceled bookings, since they make up ⅔ of bookings ○ ▶️ B19 → =average('🚀 Start here'!L:L) What do you think could explain this difference? Here's one potential interpretation: • Canceled bookings were reserved on average over 4 months in advance, which suggests to me that these were planned vacations where plans must have changed • On the other hand, bookings that were not canceled were only booked around 2 months out. At that point, plans were probably more stable. • Regardless, lead time seems to be related to cancellation in some way, although we can't know exactly how without further analysis.
🎙️ TH	Great work on that segmentation task! Follow me to the next video to learn a similar technique that counts and sums based on multiple conditions.

L2V6 - Summary statistics – COUNTIFS, SUMIFS

🎙️ TH 	In the previous videos, you used conditional formulas including COUNTIF, SUMIF, and AVERAGEIF. These only check one condition. What if you want to calculate based on multiple conditions, say comparing counts based on cancellation status and market segment. Let's take a look.
Continue with the same sheet as before.	• 💬 First, set up your data to address this question. ○ ▶️ Copy A21:E26, remove formulas INCLUDING =UNIQUE • 💬 For market segments, I'm not sure at a glance how many there are. I'll use the unique formula to get all the unique values for my rows. Make sure not to include the column header. Also add total. ○ ▶️ A22 → =UNIQUE(M2:M) ○ ▶️ A27 → "Total" • 💬 So now in this first cell, I want to count the number of canceled cells if the market segment is also "Offline". ○ ▶️ Show tooltip, explain briefly ○ ▶️ B22 → =countifs(Data!T:T,"Canceled",Data!M:M,A22)

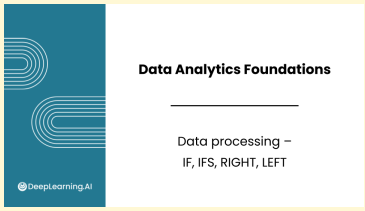
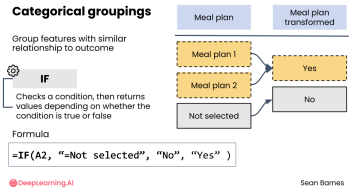
	• 💬Good work, so 3153 are canceled and offline. Notice that A22 is a relative cell reference. If you use the fill handle, you can calculate all the counts for each market segment.
	  ○ ▶️ Fill handle from B22 to B26
	• 💬 Great! [share some insights]
	• 💬 Now sum those. What should those add up to? The total number of bookings.
	• 💬 Let's do the percent
	  ○ ▶️ C22 → =B22/B27
	• 💬 Can we use the fill handle here?
	  ○ ▶️ Try fill handle
	• 💬 Ah, no. That's because B27 here is also a relative reference. Notice I get this suggestion. You can use the dollar sign to prevent a row or column reference from changing. In this case, I want to stop the 27 from changing
	  ○ ▶️ C22 → =B22/B$27
	• 💬 Now if I drag this down, the first cell reference changes, but the second one doesn't
	  ○ ▶️ click into B25 as an example
	  ○ ▶️ Format as percent
	• 💬 I'll do the same quickly for not canceled, you can try this on your own but I just want to do a quick comparison.
	  ○ ▶️ Copy over D22:E27 from solution

Continue with same sheet as before	SUMIFs
	• I've added a column here called total value, this is the average price per room times the number of nights. Let's sum the total value for these segments using SUMIF.
	  ○ ▶️ Show tooltip
	• 💬 It looks similar, but it has another parameter at the beginning, the sum range. Choose the column you want to sum over. Then the criteria are the same as for the countifs. Now I can use both the booking status and market segment as cell references. Have to use dollar sign for booking status
	  ○ ▶️ B30 → =sumifs(Data!R:R,Data!T:T,B$29,Data!M:M,A30)
	  ○ ▶️ Drag down
	  ○ ▶️ Add sum in total cell
	• 💬 That's a lot of potential revenue! More than 4 million Euros. How about for not canceled bookings?
	  ○ ▶️ C30 → =sumifs(Data!R:R,Data!T:T,C$29,Data!M:M,A30)
	
	Now you can do all sorts of analyses and business predictions and planning given these numbers!

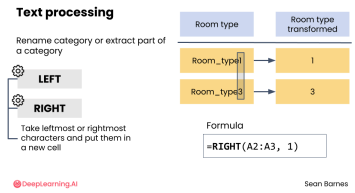
🎙️ TH	You saw that COUNTIFS and SUMIFS give you a ton of flexibility. In the practice lab for this lesson, you will calculate the revenue per booking, not

	just total revenue. Remember: if you forget about which parameters the function uses, you can always check the help menu. You're almost done with this lesson. Follow me to the next video to learn about data processing techniques in spreadsheets.

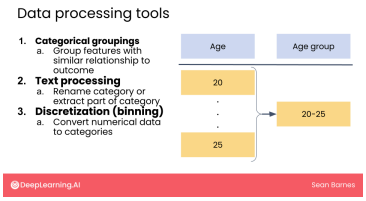
L2V7 – Data processing – IF, IFS, RIGHT, LEFT

Visual	Script
🎙️ TH 	So far, we've been working with the hotel reservations data set as you found it. But there's no rule that you have to use data as it comes. You can process the data as much as needed to solve your problem, as long as the changes are valid. Let's see how to apply common data processing techniques.
	Let's start with **[CLICK]** categorical groupings. - Combining multiple categories into one. Why? You can **[CLICK]** group features with a similar relationship with your outcome. Or maybe there's too low of frequency across some categories and you just want to group them into 'Other'. - In the hotel dataset, for example, you can collapse the **[CLICK]** "meal plan" feature into just **[CLICK]** two categories: yes and no. That may help make your analysis cleaner if there is no difference in cancelation rate between meal plans 1 and 2. The **[CLICK]** IF function is one powerful tool for categorical groupings. It's another conditional function like COUNTIF. • It **[CLICK]** checks a condition and then returns a different value depending on whether the condition is true or false. Here's an example. **[CLICK]** IF **[CLICK]** this cell **[CLICK]** is "Not selected", we just want to return **[CLICK]** "No", and otherwise we want to return **[CLICK]** "Yes". Let's do some data processing on the hotel reservations data set to prep it for more analysis in the upcoming practice lab.

🔗 DEMO SPREADSHEET	• 💬 I mentioned how we may want to group the meal plan categories together into just Yes or No. In fact, let's take that idea further and just make it 0 for no meal plan and 1 for any type of meal plan.
	  ○ ▶️ Create new column, has_meal_plan
	  ○ ▶️ in J2 → =if(I2="Not Selected",0,1)

	• 💬 Notice instead of having the cell and condition as separate inputs, IF takes a logical expression. • 💬 Now at a glance I can tell much more easily that most people got a meal plan of some kind. I can add conditional formatting more easily too. ○ ▶️ J:J → conditional formatting → single color → is greater than 0 → blue
	Your next tool for data processing is Text processing - **[CLICK]** Renaming a category or extracting a part of it so the text is easier to read. • For text processing, you can use the **[CLICK]** LEFT and **[CLICK]** RIGHT functions. These functions **[CLICK]** take the leftmost or rightmost characters and put them in a new cell. • For example, say you want to just extract the number of the **[CLICK]** Room type. You can use the **[CLICK]** RIGHT function to extract the **[CLICK]** single rightmost character of the original feature, which makes the data more readable. Let's try that.

Continue with previous cell	• 💬 Make a new column for hygiene purposes. Never overwrite your original data, it causes major headaches.
	  ○ ▶️ New column M → "room_type_numerical"
	  ○ ▶️ M2 → =RIGHT(1)
	• 💬 1 is the default but for consistency you can put 1 character here
	• 💬 Nice! The result is so much easier on the eyes.

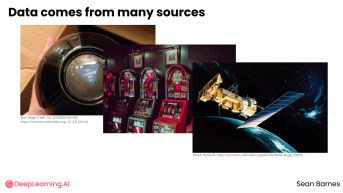
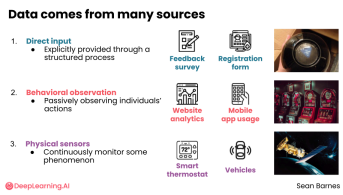
	For numerical data, you'll often use it as-is, but **[CLICK]** grouping into categories is useful, a process known as binning. It's useful if the direct relationship between a numerical feature and your outcome is unclear. • A common example of binning is the use of **[CLICK]** age groups. Oftentimes, there is not much of a difference between, say, income earnings or health outcomes of a 22 year old vs. a 24 year old. You can simplify your analysis by grouping people into **[CLICK]** age groups and reanalyzing. This strategy can help you discover new insights. For this technique, it's quite useful to employ the **[CLICK]** IFS function. Can you guess what it does? **[pause for thought]** **[CLICK]** It checks multiple conditions! IFS also uses the same "logical expression" idea you saw with IF. **[CLICK]** You'll put the whole **[CLICK]** condition together as the first input, then the **[CLICK]** value you want displayed if the cell meets that condition. **[CLICK]** Then repeat! **[pause to absorb]**

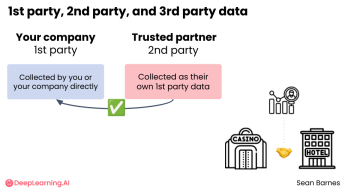
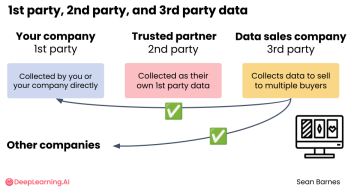
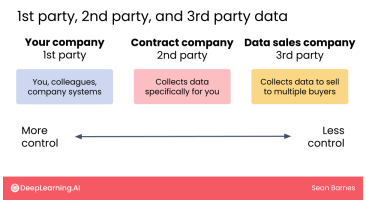
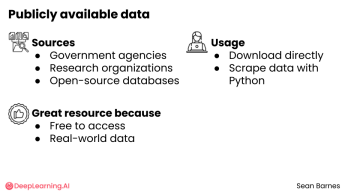
	Let's see that in action on the hotel reservations data set.

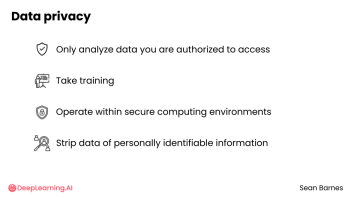
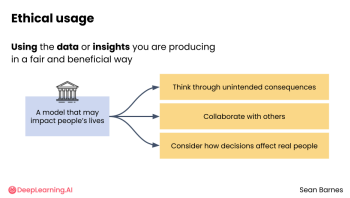
Continue with previous screencast spreadsheet	• 💬 Let's go through an example where we bin lead time into bins of <50 days, 50-100 days, and >100 days. I'll create a new column
	  ○ ▶️ New column O → "lead_time_binned"
	• 💬 Now what I want to do is check this cell, if it's less than 50, I'll put <50, then I can check if it's less than 100, and so on.
	  ○ ▶️ O2 → =ifs(N2<50,"<50",N2<100,"50-100",N2>=100,">100")
	  ○ ▶️ Apply to column
	• 💬 Cool! One thing this allows me to easily do is filter by these conditions. Before it would have been harder.
	  ○ ▶️ Filter O → <50
	  ○ ▶️ Show Filter for N

🎙️ TH	Great work processing that data! These techniques will make further analysis in the practice lab much easier. Follow me to the next video to zoom out a bit and talk about where all this data is coming from and how that affects your analysis.

L2V8 – Where does data come from?

Slide	Script
🎙️ TH	We've been working with this data for awhile – numbers, dates, categorical data. Let's take a step back for a moment. I'd like to take a few minutes to talk to you about where data comes from. Yes, we're going to have "the talk".
	As you learned in the previous module, data can come from almost anywhere. A **[CLICK]** customer leaving a review about the magic 8 ball they just purchased. A **[CLICK]** casino that tracks spending down to each square meter. The hundreds of **[CLICK]** weather satellites currently in orbit. Each of these data sources is distinct. Let's see how to describe those differences.
	First, data can be collected via **[CLICK]** direct input, which means that the data is **[CLICK]** explicitly provided through a structured process such as a **[CLICK]** customer feedback survey or a **[CLICK]** registration form at a doctor's office. Your **[CLICK]** magic 8 ball review data falls into this category. Second, data could be collected via **[CLICK]** behavioral observation, which means a system that gathers data by **[CLICK]** passively observing individuals' actions. This type of data includes **[CLICK]** website analytics, **[CLICK]** mobile app usage, or social media engagement. **[CLICK]** Casino monitoring also falls into this category.

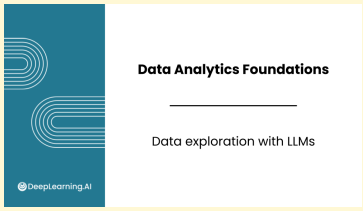
	Third, data could be collected via **[CLICK]** physical sensors that **[CLICK]** continuously monitor some phenomenon. **[CLICK]** Smart thermostats that measure temperature, **[CLICK]** vehicles that track driving patterns, **[CLICK]** or environmental sensors like **[CLICK]** satellites all fall into this category. Even when you know *how* a certain data set was generated, ~~that is, whether it is direct input, behavioral observation, or sensor data,~~ there is still more information you should know about where it comes from.
	For example, who collected the data? **[CLICK]** 1st party data is **[CLICK]** owned by you or your company directly. For example, **[CLICK]** a casino installs its own cameras throughout the gaming area to monitor patrons.
	2nd party data is collected by **[CLICK]** another company as their own 1st party data. You typically **[CLICK]** acquire this data from a trusted partner. **[CLICK]** A casino might **[CLICK]** partner with a neighboring hotel to share **[CLICK]** customer data to gain insights about big spenders.
	A 3rd party may collect the data for the general purpose of selling the data to **[CLICK]** multiple buyers. A casino could buy a large data set of people who have **[CLICK]** visited online gambling websites for a new marketing campaign. If you had to guess, which type of data would you have the most control over? **[pause for thought]**
	You have more control over 1st and 2nd party data and can usually make sure it will serve your specific purpose. For 3rd party data, you may have to work with a low number of observations, irrelevant features, or systematic inaccuracies.
	A lot of data is also publicly available, such as **[CLICK] [CLICK] [CLICK]** data published by government agencies, **[CLICK]** research organizations, and **[CLICK]** open-source databases. This data is often intended to enable research that benefits society at large. Publicly available data is a **[CLICK] [CLICK]** great resource for you as a data analyst because it's typically **[CLICK]** free to access and it's often **[CLICK]** real-world data. Some of this data is **[CLICK] [CLICK] [CLICK]** directly downloadable from public-facing websites, and in other cases, you may be able to **[CLICK]** scrape data using a programming language such as Python.

	As you advance your skills as a data analyst, you'll learn these more sophisticated methods for acquiring data programmatically.
TH	Lastly, let's talk about ethical usage. These are important aspects of your work as a data analyst. You'll do more than just "crunching numbers." Oftentimes, you'll act as a steward of truth and an advocate for the people that exist in your data.
	You should only **[CLICK]** analyze data that you are authorized to access. Data is often legally protected, such as financial ~~data,~~ or personal health information ~~in the US, which is protected by the Health Insurance Portability and Accountability Act~~. You may have to **[CLICK]** take training to access sensitive data, or **[CLICK]** operate within secure computing environments. In some cases, it may be necessary to **[CLICK]** strip data of personally identifiable information, such as names, addresses, or social security numbers. The hotel reservations data set was one example of a dataset where personally identifiable information was removed.
	Ethical usage means that **[CLICK]** you use the data or insights that you are producing in a fair and beneficial way. For example, are you training a **[CLICK]** model that may significantly impact people's lives, such as in **[CLICK]** criminal justice decisions? ~~In which case,~~ how can you make sure the model is fair, and not perpetuating historical discrimination? **[CLICK]** Think through potential unintended consequences of your work, which may require you to **[CLICK]** collaborate with others. You must **[CLICK]** consider how the business decisions derived from your insights will affect real people.
TH	OK, that concludes "the talk" about where data comes from, and also the videos for this lesson! I encourage you to look for sources of data in your everyday life, even in the smallest and strangest places. In the practice lab for this lesson, you'll explore the hotel reservations dataset yourself to find some cool insights and practice what you've learned. Once you've completed the practice lab and assessment, I hope you'll join me in the next lesson to learn how to explore data using LLMs. I'll see you there.

Lesson 3: Data exploration with LLMs

L3V1 – Data exploration with LLMs

Visual	Script

🎙️TH  	In this lesson and accompanying lab, you'll use an LLM to explore a data set. You'll learn prompting techniques and practice a skeptical mindset using the hotel reservations data set.

🔗 **SCREENCAST**
 	
How might a dataset of hotel bookings be generated?
 	
Summarize this information about the dataset and tell me the observations and features:
• 🔗**Copy description from website**
	• 💬Now let's ask some questions about a hotel reservations data set. Here we're just brainstorming with the LLM, since it doesn't have the ability to understand the entire data set.
	  • 🤖How would a dataset of hotel bookings be generated?
	  ○ ▶️Respond to LLM
	• 💬Now let's give it some more information to work with.
	  • 🤖Summarize this information about the dataset and tell me the observations and features:
	  • ▶️Copy and paste in abstract & specifications table
	• 💬I'll ask the LLM to summarize this information about the dataset and tell me the observations and features. Note that I can combine multiple questions. This info can help inform our analysis.
	  • ▶️Respond to LLM
	• 💬This type of response can help us understand what kinds of business questions we can answer with this data set.

🔗 **SCREENCAST**
 	
What is this dataset about?
 	
What range of dates are these bookings for?
 	
Are there any missing values?
	Let's try the next step up: uploading our data set so the LLM can see more of the data for itself.
	
	You can do so in Coursera by clicking this file button, then selecting the dataset. Note that you can only use data sets pre-loaded into the lab, not your own.
	• 💬To upload a dataset, click on the paperclip and select your dataset.
	• 💬Note that the LLM has to read the entire data set every time you ask a question, so the larger the dataset, the slower your conversation will be. I'll use an abridged dataset of **200 rows**.
	• 💬Remember that the LLM can't perform precise calculations, like calculating a mean or or percent. LLMs are good at reading and writing, not math. So I'll play to its strengths while asking specific questions about the data. Let's first ask, what is this dataset about?
	  ○ 🤖What is this dataset about?
	  ○ ▶️Respond to what LLM says
	• 💬Next I can ask, what range of dates are these bookings for?
	  ○ 🤖What range of dates are these bookings for?
	  ○ And it will respond that these bookings are for dates in 2017 and 2018, and show how it got that information.
	• 💬Finally, I'll ask, are there any missing values? Again this is a

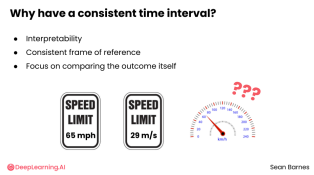
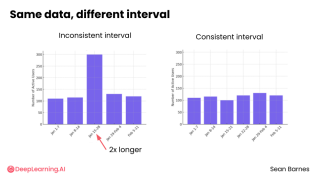
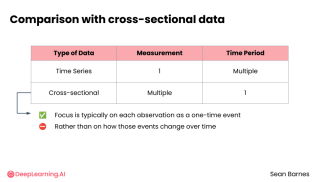
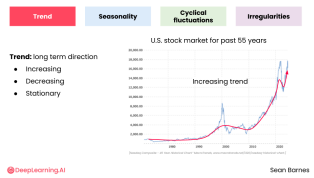
	question specific to this dataset and the LLM will use the data file to answer the question. ○ 🤖Are there any missing values? ○ ▶️Respond to LLM
🔗 **SCREENCAST** What are the number of observations and features in this data set? Are these observations in chronological order? What is the range for the number of children feature. Please visualize the range of number of children. What percent of people book at least one month in advance? Create a graph.	• 💬ChatGPT is an LLM that can run code. What's great about this feature is that it compensates for some of the LLM's shortcomings and can actually compute calculations that are accurate. For this example, I'll use ChatGPT Pro, which has the Advanced Data Analytics feature that will write and run code for you. I'll upload the entire dataset as well rather than a subset, since ChatGPT Pro can handle large files, as it doesn't directly read them as part of the prompt each time. ○ ▶️Upload dataset to ChatGPT • 💬Note that I'm only uploading this anonymized, publicly available dataset. You shouldn't upload proprietary data, as you can't control what OpenAI does with this information. • 💬First, I'll ask ChatGPT again about the number of observations and features in our data set. ○ 🤖What are the number of observations and features in this data set? • 💬This "analyzing" section writes code to answer the question. After it finishes its response, you can click on this little code symbol to see what code it ran. • 💬We have 36,275 observations and 19 features, great! • 💬Now let's get into some more sophisticated analysis that will require actual calculations. First: are these observations in chronological order? ○ 🤖Are these observations in chronological order? • 💬And the LLM explains they are not in chronological order. • Okay, now I can ask for the range of the number of children feature. ○ 🤖What is the range for the number of children feature? • 💬It may go through a process here to find the right column. Wow! It looks like someone arrived with 10 children. So how common is that? Let's get a quick visual for that using the prompt "Please visualize the range of number of children". ○ 🤖Please visualize the range of number of children. • 💬Now that's pretty interesting, nearly all of the bookings were made with 0 children. • 💬Final question, what percent of people book at least one month in advance? And I'll ask it to show me a graph too. ○ 🤖What percent of people book at least one month in advance? Create a graph. • 💬Interesting, about 66.5% of people book one month in advance. Only about a third of these bookings are more spur of the moment. It looks like people plan ahead more than I thought!

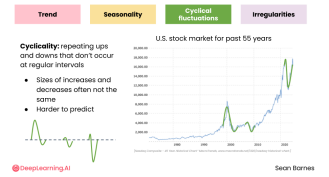
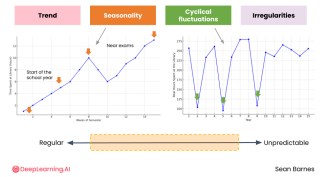
	I encourage you to explore more questions using LLMs that can write code, if you have access to one. The LLM isn't guessing at its math, it's actually making accurate calculations. Even though this feature is part of the Pro subscription, it may make financial sense for you as a data analyst for quick exploratory data analysis.
🎙️ TH	Now you've seen three different ways to use LLMs for data analytics. In the following practice lab, you'll continue to develop your prompting skills. I hope you continue to the lab and check out how you can partner with LLMs in your data analytics work. Once you're done, join me in the next lesson to learn all about time series data.

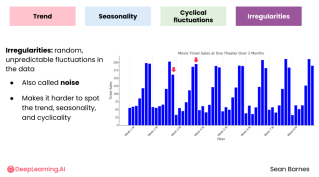
Lesson 4 - Time series data

L4V1 – Introduction to time series

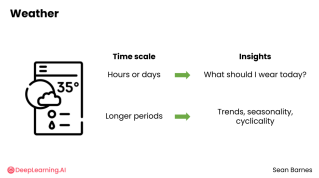
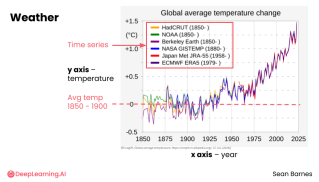
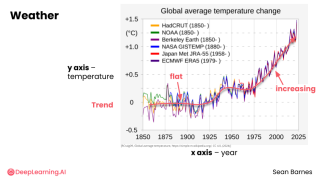
Visual	Script
🎙️ TH	Understanding climate change, tracking animal populations, even predicting earthquakes... All of these tasks rely on analyzing data over time. Time series data is a fundamentally different type of data. Let's take a look.
Introduction to time series • Measuring one or more outcomes over fixed periods of time ◦ Minutes ◦ Hours ◦ Days • Goal: Understand how the outcomes vary over time ◦ Identify increasing or decreasing trends ◦ Discover unusual events ◦ Forecast future outcomes DeepLearning.AI Sean Barnes	Time series data is generated by **[CLICK]** measuring one or more outcomes over fixed time periods such as **[CLICK]** minutes, **[CLICK]** hours, or **[CLICK]** days. The goal with time series analysis is to **[CLICK]** understand how these outcomes change over time. You can use it to **[CLICK]** identify increasing or decreasing trends, **[CLICK]** discover unusual events, or **[CLICK]** *forecast* future outcomes.
Common use cases for time series Measurement | Time interval | Purpose Rainfall in the Amazon | Annually | Global warming Active users | Weekly | User growth Sales | Monthly, quarterly | Revenue forecasting Stock market prices | By minute, hourly | Day trading Stock market prices | Monthly, quarterly, yearly | Long-term investing DeepLearning.AI Sean Barnes	Many industries use time series data, capturing the same **[CLICK]** measurement over a consistent time interval. • An environmental group might measure **[CLICK]** rainfall in the Amazon annually to better understand global warming trends • Tech companies often measure the **[CLICK]** number of active users each week to see how many people regularly use the product, which can help them forecast user growth. • Most companies have some type of **[CLICK]** sales. As a data analyst, you'll often measure sales monthly or quarterly to help understand trends over time. • **[CLICK]** Stock market prices are often analyzed across multiple time intervals. **[CLICK]** Day traders may be interested in how a company's stock varies per minute or per hour, whereas **[CLICK]** long-term investors are probably more interested in intervals of months, quarters, or years.

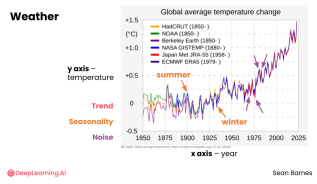
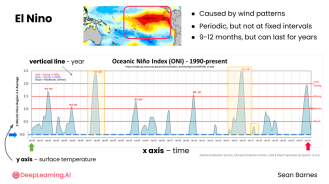
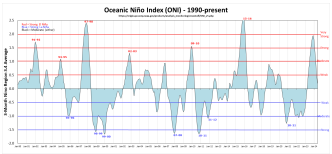
	Time series data requires you to have a consistent time interval ~~because of the [CLICK] interpretability provided by a [CLICK] consistent frame of reference. That way, you can [CLICK] focus on comparing the outcome itself, rather than calculating different lengths of time. Imagine seeing a speed limit sign that says [CLICK] 65 mph one day, and [CLICK] 29 m/s the next, meanwhile your [CLICK] speedometer is measured in km/hr. It would be chaos! Your focus would be on converting between all the units rather than obeying the law.~~
	Here's an example of interpreting inconsistent time intervals. I'll show you two graphs of time on the x axis and number of active users on the y axis • In the chart on the **[CLICK]** left, the middle period is **[CLICK]** twice as long as all of the others, which may lead you to believe there was a spike in user activity. **[3 SECOND PAUSE FOR LEARNERS TO READ CHART]** • However, when the intervals are **[CLICK]** consistent, like the chart on the **[CLICK]** right, the overall trend does not include such a spike. **[3 SECOND PAUSE FOR LEARNERS TO READ CHART]**
	You saw a moment ago that **[CLICK]** time series data is the measurement of **[CLICK]** one outcome over **[CLICK]** multiple time periods. If you invert that definition, you get data that is collected over a **[CLICK]** single time period, but across **[CLICK]** different measurements. This type of data is called **[CLICK]** *cross-sectional data*. Cross-sectional data can also be collected over time, but the **[CLICK] [CLICK]** focus is typically on each observation as a one-time event **[CLICK]** rather than on how those events change over time. In the hotel reservations dataset, each booking is treated as a one-time event. You're not attempting to track individual bookings over time.
 https://www.macrotrends.net/1320/nasdaq-historical-chart	Time series data is often analyzed based on four components: **[CLICK]** trend, **[CLICK]** seasonality, **[CLICK]** cyclical fluctuations, **[CLICK]** and irregularities. Check out **[CLICK]** this graph of the US stock market over the past 55 years, with time on the x axis and the size of the market on the y axis. A higher value is better! What can you say about the size of the market over time? There are ups and downs, but in general, what direction is it going? **[pause for thought]** **[CLICK]** This graph has an increasing trend. The trend is the **[CLICK] [CLICK]** long-term direction of the data. On the whole, is it going up, down, or staying the same? Trends can be: • **[CLICK]** Increasing: The values tend to go up over the observed period. • **[CLICK]** Decreasing: Values generally go down ~~as time progresses.~~ • **[CLICK]** Stationary: No consistent long-term increase or decrease. This

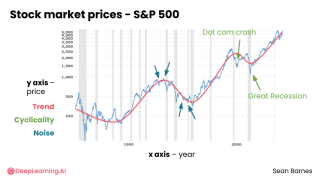
	is also called having no trend.
	Take a look at this graph of movie ticket sales at a local theater over two months. The x axis shows time and the y axis shows the number of tickets sold, with each bar representing one day. What repeating pattern do you notice? **[pause for thought]** The movie ticket sales are **[CLICK]** higher on the weekend compared with weekdays. **[CLICK] [CLICK]** This is seasonality – a repeating, predictable pattern that occurs at regular intervals. Seasonality can occur **[CLICK]** daily, **[CLICK]** weekly, **[CLICK]** monthly, or **[CLICK]** yearly; it **[CLICK]** doesn't have to be across weather seasons. **[CLICK]** It can even appear at multiple time intervals. For example, movie ticket sales often increase during summers or holidays, with that pattern repeating year over year
	Let's return to this graph of the stock market you saw a moment ago. Can you spot any patterns of increases and decreases that seem to repeat at irregular intervals? **[pause for thought]** You're identifying stock market bubbles and crashes, like this **[CLICK]** dot com bubble, **[CLICK]** the financial crisis of 2008, and **[CLICK]** the pandemic. These are called **[CLICK] [CLICK]** cyclical fluctuations. The stock market undergoes repeating ups and downs that don't occur at regular intervals like weekends or seasons. **[CLICK]** The sizes of increases and decreases are often not the same. These irregularities make cyclical patterns **[CLICK]** harder to predict compared with seasonality. It's tough to know when the next stock market bubble will happen.
	Here's a more relatable example to help you remember the difference between seasonality and cyclicality. **[CLICK]** Think about the time you spend studying at the campus library. **[CLICK]** At the start of the school year, you might be studying a bit less, since classes have just started. **[CLICK]** Then, near your exams, you'll spend a lot of time studying at the library. Are these seasonal or cyclical patterns? **[pause for thought]** These are **[CLICK]** seasonal patterns, since they happen at regular, predictable intervals. Each semester, you would expect the same pattern. **[CLICK]** Now consider an event like construction at the library, which happens every 2-3 years but not on a fixed schedule. **[CLICK]** The construction may make it difficult for you to study there due to the noise and dust. Is this event seasonal or cyclical? **[pause for thought] [CLICK]** It's cyclical, since construction does periodically happen at the library, but not at regular intervals. Real world data often doesn't fit into clear cut categories of seasonality and cyclicality. On a **[CLICK]** spectrum of predictability, from highly regular to completely unpredictable, many events fall **[CLICK]** somewhere in the middle.

	Let's return for a moment to the graph of movie ticket sales. Why is the **[CLICK]** Sunday value for Week 2 much lower than the other weeks? Why is Week 3 the only week where **[CLICK]** Sunday has the most sales? These are **[CLICK]** **[CLICK]** random, unpredictable fluctuations in the data, so they are called irregularities or **[CLICK]** noise. Think of noise as the data version of static in the background of a phone call – random sounds that make it harder to hear the person on the other end. Similarly, noise in time series data **[CLICK]** makes it harder to spot the trend, seasonality, and cyclicality.
TH	Great work identifying the different time series components in movie ticket sales and the stock market! As a data analyst, you will often analyze the trends, seasonality, and cyclicality of time series data. Join me in the next video to see some specific examples of real-world time series data you might work with.

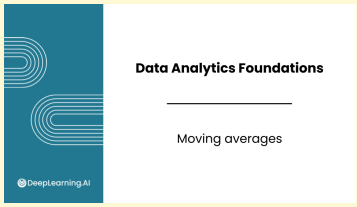
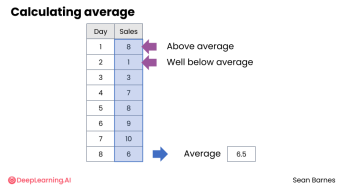
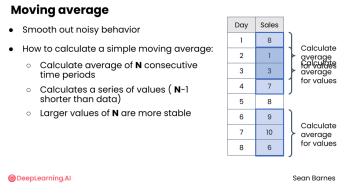
L4V2 – Real-world time series

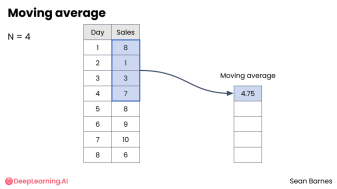
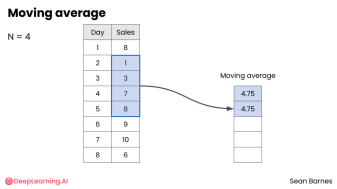
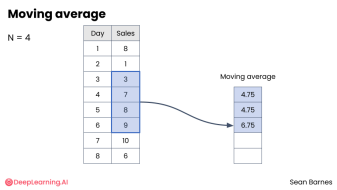
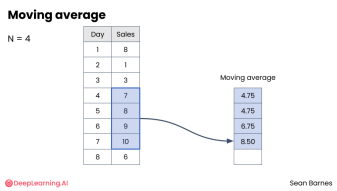
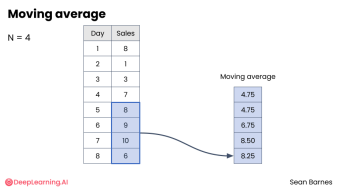
Slide	Script
TH 	Time series data can exhibit all or none of the components you saw in the previous video. How do they combine in real-world situations? Let's see how you can analyze graphs of time series data to identify these components visually.
	When you open the weather app on your phone, you're often looking at a temperature forecast for the next **[CLICK]** few hours or days. That can help you answer questions like **[CLICK]** "what should I wear today?" However, when you look at **[CLICK]** longer time periods, you can analyze the long-term **[CLICK]** trends, seasonality, and cyclicality of temperatures, rainfall and other weather measurements.
	Consider **[CLICK]** this graph of average global temperature since 1850. The **[CLICK]** y-axis values represent the temperature, in degrees Celsius, relative to the **[CLICK]** average temperature between 1850 - 1900, which is used as a reference for the pre-industrial age. The **[CLICK]** multiple time series plotted on this chart, the different colored lines, are measurements from different types of temperature sensors.
	You can see that the **[CLICK]** trend is **[CLICK]** flat from 1850 - 1925, but then starts to **[CLICK]** increase fairly consistently on average, perhaps with a bit of a pause between 1940-1975.

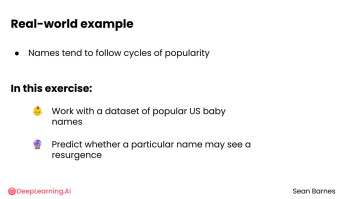
	Throughout the time series, you can also see a combination of **[CLICK]** seasonality patterns that move up and down with the weather seasons, and also some **[CLICK]** noise that makes the patterns look not quite perfect. It's difficult to determine what cyclical patterns are present using this chart, since there is a lot of seasonality that could be obscuring them, and weather patterns are often local.
  Just the top half if possible?	Here's a weather-related example of cyclicality: El Nino. El Nino refers to **[CLICK]** warming of the Pacific ocean's surface **[CLICK]** caused by particular wind patterns. It happens **[CLICK]** periodically, but not at fixed intervals. It often lasts for **[CLICK]** 9-12 months, but can last for years. Here's **[CLICK]** a graph of El Nino, and this shows on the **[CLICK]** x axis time, from **[CLICK]** january 1990 to **[CLICK]** january 2024 **[CLICK]**, with each **[CLICK]** vertical gray line representing a year, and on the **[CLICK]** y axis is a measurement related to the ocean's surface temperature. Sustained values above **[CLICK]** this 0.0 baseline are El Nino years. **[CLICK]** The 97 – 98 El Nino was very strong and lasted about a year, while the **[CLICK]** 15-16 El Nino was even stronger and lasted almost a year and a half. Meanwhile there are many smaller examples, such as this one **[CLICK]** from around February to July 2017 that would only be classified as weak. El Nino is considered cyclical because it does happen periodically, but the strength and duration is difficult to predict. You know El Nino will happen again, but it's hard to say exactly when, for how long, and how strong it will be.
	Let's move on from weather to a more sophisticated version of the stock market graph you saw in the previous video. This data is often analyzed to make investment decisions. In this case, you are looking at a line chart of the S&P 500, which is a combined measure of stock prices across the 500 largest U.S. companies. Take a look at the trend, which can help an investor decide whether to buy, hold, or sell a particular stock. What kind of trend do you see? **[pause for thought]** Generally, it's increasing. However, over the short term, these trends are nearly impossible to predict. Most investors aren't looking to keep a stock for 120 years. Imagine if you were to zoom in to different time periods on this chart. The trend could be **[CLICK]** increasing, **[2 SECOND PAUSE]** **[CLICK]** decreasing, **[2 SECOND PAUSE]** or **[CLICK]** flat **[2 SECOND PAUSE]** depending on where you start and stop.

	But when you zoom out, the trend is clearly increasing over time, and even some of the largest recessions in the last 30 years, such as the **[CLICK]** "dot com" crash in 2000, the **[CLICK]** Great Recession in 2008, or even the Covid-19 pandemic seem relatively insignificant, despite the impact they had on the world at the time. These are **[CLICK]** cyclical patterns that correspond to broader economic conditions. The other half of these cycles are periods of economic growth, such as the 2010s. Stock market prices also demonstrate a lot of **[CLICK]** noise. Prices are impacted by many factors, not all of which are perfectly understood. **[2 SECOND PAUSE TO ABSORB INFO]**
TH	Now that you've seen real-world time series examples, join me in the next video to learn about some concepts we can use to work with time series data.

L4V3 – Moving averages

Visual	Script
TH 	Some interesting analyses are unique to time series. One of these, called moving averages, allows you to smooth out potentially noisy data. It's especially useful for data collected over small intervals, or when you have a lot of data. Let's take a look.
	Let's refer back to our solar panel sales exercise from Lesson 1. Suppose you have a **[CLICK]** time series for the number of units sold each day. In this example, you would calculate the **[CLICK]** average of 8 time series values to be **[CLICK]** 6.5 units, and now you have a reference to know how each individual value compares to this reference. For example, on Day 1, you sold 8 units, which was **[CLICK]** above average. But on Day 2, you only sold 1 unit, which was **[CLICK]** well below average. **[1 SECOND PAUSE]**
	As you saw a moment ago, time series data can be noisy, making it difficult to clearly identify the data's overall behavior. Moving averages **[CLICK]** smooth out this noisy behavior, and can bring clarity to your analysis. Let's take a look at how to **[CLICK]** calculate a simple moving average. Simple moving averages **[CLICK]** calculate the average outcome across several consecutive time periods. The number of timer periods is represented as *N*. Imagine placing **[CLICK]** a window over your data, *N* units tall, then **[CLICK]** calculating the average of the values inside the window. You can then

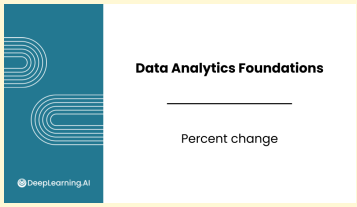
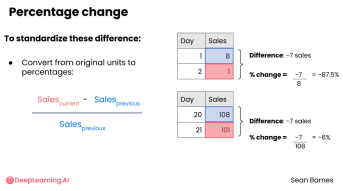
	[CLICK] slide this window along our data – one period at a time – **[CLICK]** until you get all the way to the end. Each window pools the total solar panels and redistributes them across N days as if you sold the same number each day. So Instead of calculating a single number to summarize the data, a moving average **[CLICK]** calculates a series of values. This series is *N* -1 units shorter than the size of our data. You can't calculate a simple moving average with fewer than *N* data points in the window. **[CLICK]** Larger values of *N* tend to be more stable over time, whereas smaller values of *N* tend to be more noisy.
	Let's refer back to our solar panel sales exercise from Lesson 1. Suppose you have a time series for the number of units sold each day. Here's how to calculate a simple moving average for this data. Let's pick **[CLICK]** N=4, a window size of 4. Start by **[CLICK]** placing the window over the first N values and calculate the average within the window. So in this case, the numbers 8, 1, 3, and 7 appear within the window, which average out to **[CLICK]** 4.75.
   	Then slide the window one place down. Now it contains the values 1, 3, 7, and 8, which also **[CLICK]** average out to 4.75. And so on, **[CLICK]** until the **[CLICK]** window reaches the **[CLICK]** end of the time series. Notice that the length of the simple moving average series on the right is 3 periods shorter than the original time series. You can calculate this length as the number of periods, 8 in this case, minus the window size (or N) minus 1.

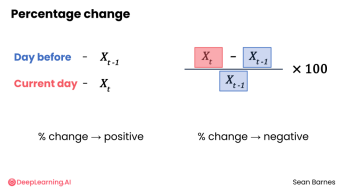
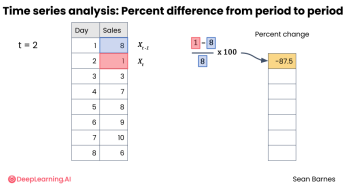
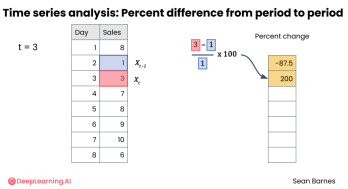
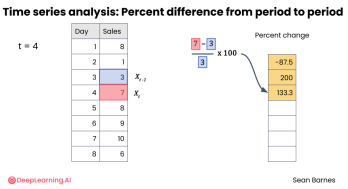
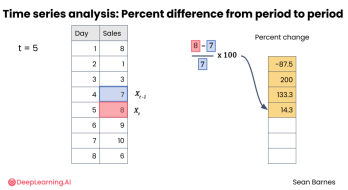
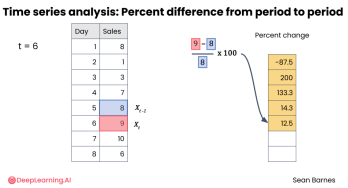
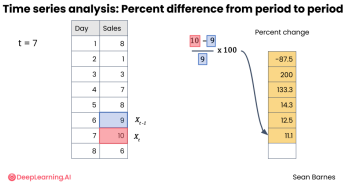
	Let's take a look at how to apply moving averages to a real world data set. Just like fashion trends, **[CLICK]** names tend to follow cycles of popularity. Older names often come back into vogue! **[CLICK]** In this exercise, you'll **[CLICK]** see how to work with a dataset of popular US baby names to **[CLICK]** predict whether a particular name may see a resurgence in the coming years.
🔗 **SCREENCAST** – **Baby Names from Social Security Card Applications**	First, let's have a quick look at the data source. • 💬 It's from the SSA, that's reliable. • 💬 Last updated 2022, so relatively recently • 💬 Any biases? Well, missing any undocumented immigrants [others you can think of]

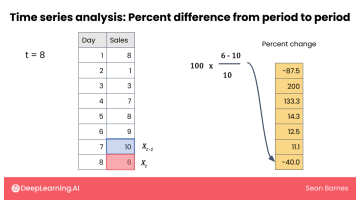
🔗 **DEMO SPREADSHEET**	Let's take it over to the data.
	• 💬 Review features; not many in the data – interested in count; unique identifier is name itself; count is outcome of interest.
	• 💬 How many observations are there?
	  ○ ▶️ Select A:A, bottom right summary → count
	• 💬 Notice these are already sorted by count and year, giving us the most popular names in each year. You can mess around with these on your own 🙂
	• 💬 That's a lot of names. Let's look at one! My dad's mom, Ruby.
	  ○ ▶️ Filter → Ruby; Filter → F
	  ○ ▶️ Copy to new sheet
	• 💬 How is the data changing over time? Hard to see at a glance. I'll add conditional formatting.
	  ○ ▶️ D:D → conditional formatting → color scale
	• 💬 Allows me to see at a glance what's happening; goes up in the 20s and more recently. These values vary a lot.
	• 💬 Let's visualize this. You'll learn how to do all this in the next module.
	  ○ ▶️ Insert → chart
	  ○ ▶️ Copy chart from solution sheet, remove moving average series
	• 💬 Interesting trend! [describe some more] My grandmother was born around the 1920s. This name is experiencing a resurgence in popularity. Hard to see the trend in recent years, though.
	• 💬Now let's see what's happening numerically, starting with average
	  ○ ▶️ G2 → =Average(D:D)
	  ○ ▶️ Decrease decimals to whole number
	• 💬 Average is 2560. It's an exclusive club! How helpful is this one number? Kinda general. Let's calculate a moving average. Moving average can be calculated from anywhere in the period. I have it in the center so you'll be able to directly compare visually.
	•

	○ ▶️ E11 → =average(D2:D10) ○ ▶️ Drag all the way down • 💬 Hard to tell if it's smoother. Let's visualize that. ○ ▶️ Add moving average series to line chart • 💬 It's much smoother! Especially in recent years, the upward trend is clear, but it seems to be leveling out [point out specific peaks, anything else interesting]
🎙️ TH	Great work analyzing the moving average of Ruby names! You can see how moving averages cut through the noise in data to reveal the overall trend. Follow me to the next video to learn another powerful tool for time-series analysis: percent change. I'll see you there!

L4V4 – Percent changes

Slide	Script
🎙️ TH 	If you want to identify when a time series is changing steadily or suddenly, you can calculate the percent change from period to period. Let's see how to do that, then apply this technique to the baby names dataset.
	Percent change can be more consistently interpreted compared with raw differences. **[CLICK]** Take a look at your solar panel sales again. If they changed from **[CLICK]** 8 one day to **[CLICK]** 1 the next, how big of a change is that? It is clear that **[CLICK]** the difference is -7 sales, but as numbers get bigger, it is easy to lose your sense of what constitutes a big or small change. For example **[CLICK]** you might have sold **[CLICK]** 108 panels one day and **[CLICK]** 101 the next. **[CLICK]** The difference is still -7, but it is much less significant compared with overall sales. **[CLICK]** To standardize these differences, you can **[CLICK]** convert from your original units to percentages. First calculate the difference as you just did, **[CLICK]** subtracting the previous day's sales from the current day, then **[CLICK]** divide by the number of sales the previous day. What you're essentially doing is answering the question, compared with the previous day's sales, how big of a difference is the current day's sales? In these two cases you ge t**[CLICK]** -7 over 8 and **[CLICK]** -7 over 108. That gives you a proportion, so multiply by 100 to get a percent. You get **[CLICK]** -87.5% and the second is **[CLICK]** -6%. At a glance, it's easier to tell which one is a bigger deal.

	Formally, you can assign variables **[CLICK]** X(t-1) to the day before and **[CLICK]** X(t) to the current day, where X is the number of units sold and the t is the day you are referring to. The final equation is **[CLICK]** X(t) - X(t-1) **[CLICK]** / X(t-1) **[CLICK]** * 100. This is exactly what you just calculated for your solar panel sales, but generalized to any time series. Notice that percent changes can be negative or positive. **[CLICK]** If your values are going up, percent changes will be positive. **[CLICK]** If they're trending downward, percent changes will be negative.
	Let's work through the solar panel sales example. You already calculated the % change for the **[CLICK]** 2nd value in this time series relative to the **[CLICK]** 1st. Subtract the **[CLICK]** previous day's sales from the **[CLICK]** current day's sales, and divide by the **[CLICK]** previous day's sales. Times 100 gives you a **[CLICK]** -87.5% change, which represents a pretty large drop in sales.
    	Slide the window one time period down, and **[CLICK]** repeat the calculation, now with the value of **[CLICK]** 1 representing the previous day's sales. The % change is (positive) 200%. Your sales increased by 200% from Day 2 to Day 3, great work 🙂 Repeat **[0.5s CLICKS THROUGH EACH SLIDE WHILE SEAN IS TALKING]** this process until you get to the end of the series, and there you have it. As you can see, the sales data is fairly volatile. It changes significantly day over day, which is common for smaller numbers. Some days you sell a lot, and some days you don't.

🔗 **DEMO SPREADSHEET**	Let's return to our Ruby baby name data! You already calculated the moving average, now you can see how the percent change can help you identify sudden spikes and drop offs.
	• 💬 Here's where we left off [reorient to spreadsheet]
	• 💬 I want to see the percent change. Will help identify decreasing and increasing periods, sharp changes
	  ○ ▶️ Insert column right of E
	  ○ ▶️ Title "Percent change"
	  ○ ▶️ In F3 → =(D3-D2)/D2
	  ○ ▶️ Drag all the way down
	  ○ ▶️ Format as percent, decrease decimals
	• 💬 Interesting! Overall an increase until 1899, which I'll admit I didn't spot.
	• 💬Notice there's nothing in F2
	• 💬Let's make these changes even easier to spot.
	  ○ ❓What kind of color scale do you think would be good for conditional formatting?
	  ○ ▶️ F:F → conditional formatting → color scale → custom → orange to blue → midpoint 0
	• 💬 Cool! Darker colors are bigger changes, orange is negative, blue is positive. Notice that darker counts lead to negative percent changes.
	• 💬 Here's something interesting, a 20% increase in 1963
	  ○ ❓If you're familiar with US history, do you know what was happening in 1963?
	• 💬 Let's investigate!

Google "ruby name" Ruby (given name) - Wikipedia	**[Google "Ruby given name"]** You can try to do some research on the internet, for example you can look at this wikipedia page **[click on the wikipedia link]** about the name Ruby. Looking through the famous people here, the first one who was alive in 1963 is Ruby Bridges. **[click on Ruby Bridges]**, who was the first African American child to attend a desegregated school in Louisiana. She likely inspired many parents to name their children after her. Her story even inspired the 1998 Disney movie of the same name. So, based on its recent resurgence, it's possible the name may get even more popular. Ruby is one name on an upward trend!
🎙️ TH	That's what data analytics is all about – spotting interesting trends,

	investigating, and coming to a useful conclusion that can drive business decisions. Great work exploring that data. Coming up, you'll put your skills to the test in the module quiz and graded lab, which is all about investigating video game sales and ratings. You'll get hands on with both cross-sectional and time series data. It will be a blast. Once you're done, follow me to the next module to explore data visualization! It's one of my favorite topics and I know you'll enjoy developing beautiful ways of displaying data. I'll see you there.