DAG C2M1 scripts

Video title	Isabel	Sean	Slides
L0V1 – Welcome to this course	✅	✅	✅
L0V2 – Generative AI in this course	✅	✅	✅
L0V3 – Module 1 introduction	✅	✅	✅
L1V1 – Populations and sampling	✅	✅	✅
L1V2 – Identifying the population	✅	✅	✅
L1V3 – Probabilistic samples	✅	✅	✅
L1V4 – Non-probabilistic samples	✅	✅	✅
L1V5 – Types of bias	✅	✅	✅
L2V1 – Histograms	✅	✅	✅
L2V2 – Demo: plotting distributions	✅	✅	✅
L2V3 – Central tendency, variability, and skewness	✅	✅	✅
L2V4 – Central tendency: mean and mode	✅	✅	✅
L2V5 – Central tendency: median	✅	✅	✅
L2V6 – Demo: central tendency	✅	✅	✅
L3V1 – Variability: range and interquartile range	✅	✅	✅
L3V2 – Variability: variance and standard deviation	✅	✅	✅
L3V3 – Skewness	✅	✅	✅
L3V4 – Why use these measures?	✅	✅	✅
L3V5 – Demo: variability and skewness	✅	✅	✅
L3V6 – Box plots	✅	✅	✅
L3V7 – Demo: LLMs for spreadsheet formulas & errors	✅	✅	✅
L4V1 – Correlation	✅	✅	✅
L4V2 – Correlation and causation	✅	✅	✅
L4V3 – Demo: correlations & scatterplots in spreadsheets	✅	✅	✅

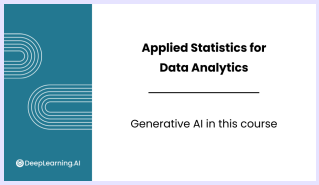
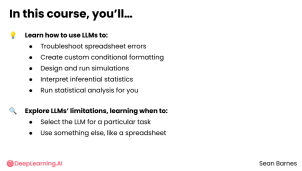
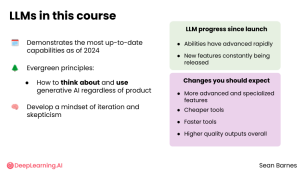
L5V1 – What is segmentation?	✅	✅	✅
L5V2 – Demo: xlookup	✅	✅	✅
L5V3 – Demo: pivot tables	✅	✅	✅

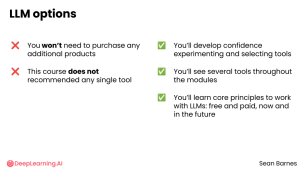
Introduction

L0V1 – Welcome to this course

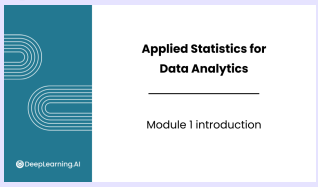
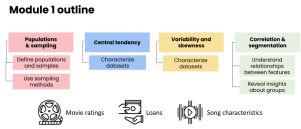
🔗**See DAG C1&C2 scripts for conversations with Andrew**

L0V2 – Generative AI in this course

Visual	Script
🎙️ TH 	One of the key elements of this course is learning to use generative AI, in particular large language models or LLMs like ChatGPT, Claude, Gemini, and so on. Effective use of LLMs will help you streamline your work and stand out.
	In this course, you'll **[CLICK]** learn how to use large language models, or LLMs, to • **[CLICK]** Troubleshoot spreadsheet errors • **[CLICK]** Create custom conditional formatting • **[CLICK]** Design and run simulations • **[CLICK]** Interpret inferential statistics • **[CLICK]** And run statistical analysis for you You'll also **[CLICK]** explore LLMs' key limitations, learning when to **[CLICK]** select the LLM for a particular task, and when to **[CLICK]** use something else, like a spreadsheet. LLMs are ever-evolving, and present challenges for both teaching and learning. I wanted to take a moment to share our team's philosophy about generative AI in this course.
	First, this course **[CLICK]** demonstrates the most up-to-date capabilities as of late 2024, and we expect changes in the coming months and years. This course is designed to teach **[CLICK]** evergreen principles: **[CLICK]** how to think about and use generative AI in your work regardless of which specific product you use. You will **[CLICK]** develop a mindset of iteration and skepticism. **[CLICK]** New models and features are constantly being released, and here are some of the **[CLICK]** changes you should expect in the near future: • First, genAI tools with **[CLICK]** more advanced and specialized features,

	like the ability to use apps for you • Expect **[CLICK]** cheaper tools • And **[CLICK]** faster tools • And **[CLICK]** higher quality outputs overall It's challenging to keep up with rapid progress in this field, but don't worry! In this course, you'll develop the metacognitive skills you need to harness those advancements in your work.
	This course also demonstrates some paid features of LLMs, but you **[CLICK]** won't need to purchase any additional products to complete the assignments. It's important for you to see the available options, including paid options, so that you **[CLICK]** develop confidence experimenting and selecting the best tools in your work as a data analyst. This course **[CLICK]** does not recommend any single tool. You'll **[CLICK]** see several throughout the modules. And remember that the **[CLICK]** core principles you'll learn will prepare you to work with LLMs both free and paid, now and in the future.
TH	You'll encounter your first LLM demo and hands-on lab in Lesson 3 of this module. For now, join me in the next video to see all the exciting topics in this module. I'll see you there!

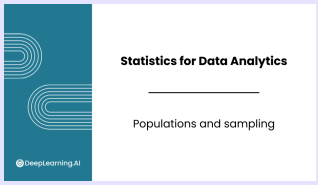
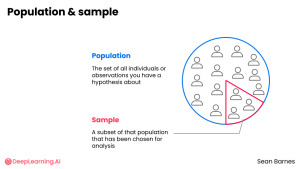
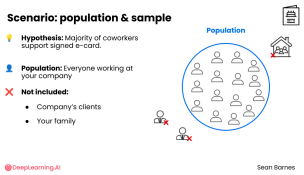
L0V3 – Module 1 introduction

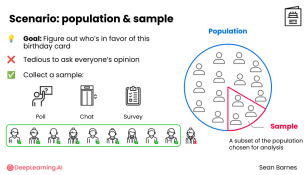
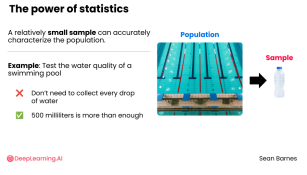
Visual	Script
TH 	Welcome to the first module of Applied Statistics for Data Analytics! Throughout this course, you will learn the fundamental statistical concepts, analyses, and visualizations that serve as the foundation for a career as a data analyst.
	In this module, you'll explore the essential building blocks of statistics that enable rigorous data analysis. You'll learn how to **[CLICK]** define populations, samples, and **[CLICK]** sampling methods; characterize datasets using measures of **[CLICK]** central tendency, **[CLICK]** variability, and skewness; use **[CLICK]** correlation to understand relationships between features; and employ segmentation to **[CLICK]** reveal insights about different groups within your data. You'll apply these concepts to real-world scenarios: analyzing **[CLICK]** movie ratings over time, identifying the most profitable **[CLICK]** loans, and analyzing **[CLICK]** song characteristics to create playlists. You'll also get hands-on

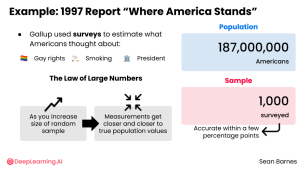
	practice with spreadsheet tools, building on the skills you learned in Data Analytics Foundations to make your analysis more efficient and effective.
	Whether you're new to statistics or looking to refresh your skills, this module will equip you with powerful techniques to extract meaningful insights from your data. By the end of this module, you will feel more confident and capable implementing rigorous statistical analyses in your career as a data analyst! Join me in the next video to get started with populations and sampling. I'll see you in the next video.

Lesson 1

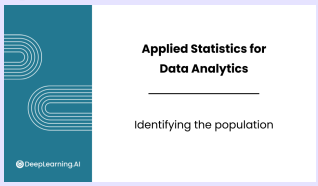
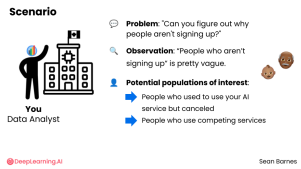
L1V1 – Populations and sampling

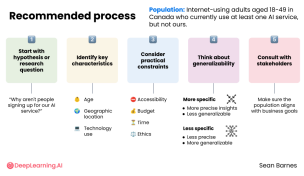
Visual	Script
TH 	Say you're working at a small company of 50 people and you want to start a fun, new birthday tradition: have everyone at the company sign an electronic birthday card! You have a hypothesis that a majority of your coworkers would be on board, but how do you find out if that's true? You could go with the magic 8 ball **[Sean takes out magic 8 ball, then puts it away]** – just kidding. You'd probably ask a couple of coworkers, right? You're trying to figure out if your hypothesis about **everyone** is right by asking just a **few** people. In other words, you're taking a **sample** of a **population**. These two definitions – population and sample – underlie almost every analysis you'll carry out as a data analyst.
	Let's define these concepts more precisely. A **[CLICK]** population is the set of all individuals or observations that you have a hypothesis about, and a **[CLICK]** sample is a subset of that population that has been chosen for analysis.
	In the birthday card example, you hypothesize that the **[CLICK]** majority of your coworkers will support sending a signed e-card. So the **[CLICK]** population is everyone working at your company. That's who this new tradition will affect. Can you think of some people who are **[CLICK]** not included in this population? **[2 second pause]** The **[CLICK]** company's clients or **[CLICK]** your family would be good examples of people who aren't in this population. Their opinion ultimately won't affect the decision.

	So you want to **[CLICK]** figure out who's in favor of this birthday card idea and who's against it. It would be **[CLICK]** tedious to ask everyone their opinion one-by-one in person. Instead, you could **[CLICK]** collect a sample! A **[CLICK]** sample is a subset of the population you've chosen for analysis. There are many ways you could sample this population: **[CLICK]** a meeting poll, **[CLICK]** elevator chat, **[CLICK]** a survey. Each of these samples is just a slice of the population, just a few coworkers. Say you end up asking **[CLICK]** 10 people, and **[CLICK]** 9 of them love the idea. That's a pretty good sign it could work!
	Note that a population doesn't have to refer to people. Here are some **[CLICK]** hypotheses that relate to other types of **[CLICK]** populations • The population might be the **[CLICK]** water in a swimming pool, if your hypothesis is that the **[CLICK]** water quality allows for swimming • Or **[CLICK]** all the iPhones manufactured in a particular facility, if your hypothesis is that **[CLICK]** less than 1% have production defects • Or, **[CLICK]** every bald eagle in the US, if your hypothesis is that their **[CLICK]** nesting behavior varies by region
	This distinction between population and sample matters because **[CLICK]** you won't typically have access to your entire population. **[CLICK]** 50 coworkers, okay, you **[CLICK]** might be able to get a response from all of them. But surveying every Netflix user? There are over **[CLICK]** 200 million of them. Most populations can't or shouldn't be fully captured, and for three key reasons: 1. First, **[CLICK]** knowledge. The population might just be **[CLICK]** unknown. For example, in a political poll, you can survey so-called "likely" voters, but you ultimately don't know who will actually go out and vote. 2. Second, **[CLICK]** practicality. Reaching the entire population could take **[CLICK]** too much time or **[CLICK]** money, and you **[CLICK]** might not be able to access its entirety. For example, you might want to test every iPhone produced in a factory, but that may add an hour to the production of each phone. 3. Finally, **[CLICK]** ethics. If you want to know if a certain fruit is toxic to dogs, well, **[CLICK]** testing that idea could cause harm.
	The good news is, through the power of statistics, even **[CLICK]** a relatively small sample, if collected properly, can accurately characterize the population. For example, say you want to **[CLICK]** test the water quality of a swimming pool. You **[CLICK]** don't need to collect every drop of water. Even for an Olympic size swimming pool, **[CLICK]** 500 milliliters is more than enough – about the **[CLICK]** size of a water bottle. **[brief pause to absorb]**

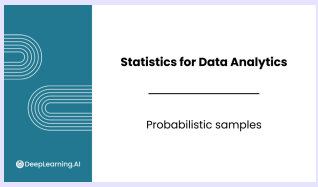
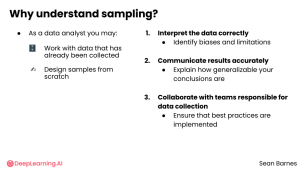
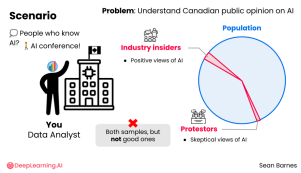
	In 1997, the polling company Gallup released a report called Where America Stands. **[CLICK]** Gallup used surveys to estimate what all 187 million Americans thought about hundreds of issues like **[CLICK]** gay rights, **[CLICK]** smoking, and the **[CLICK]** president. How many people do you think they surveyed to get an accurate answer for all **[CLICK]** 187 million Americans? **[pause for effect]** **[CLICK]** 1,000. That's all! This surprising effectiveness is due to the power of random sampling and statistical theory, particularly a concept known as the **[CLICK]** Law of Large Numbers. This law states that as you **[CLICK]** increase the size of your random sample, your **[CLICK]** measurements from that sample tend to get closer and closer to the true values you would get if you could measure the entire population. When done correctly, a sample of 1,000 people can provide results **[CLICK]** accurate to within a few percentage points of the true values.
TH	Think of the population as the truth: the way the world really is. Of your 50 coworkers, a certain number of them are in favor of sending birthday cards. Your sample is a window into that truth. You ask 10 coworkers and 9 are in favor. It's a glimpse at what everyone might be thinking, but it's not the whole truth. Join me in the next video to see how to identify the population you want to study.

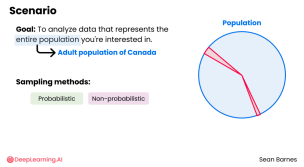
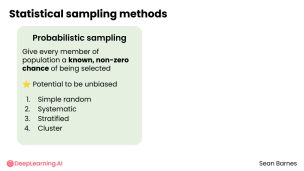
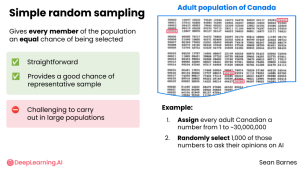
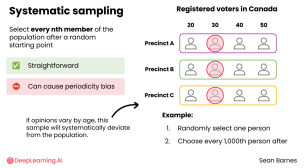
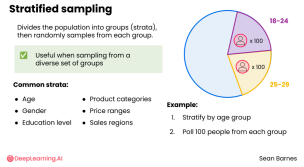
L1V2 – Identifying the population

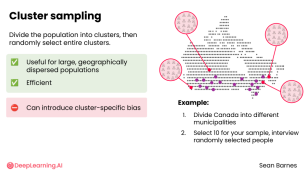
Visual	Script
TH 	Before you collect a sample, you'll need to know what you're trying to take a sample of. That may sound abstract, but the idea is that analyzing different populations can give you different insights.
	Let's explore an example. Suppose you're a **[CLICK]** data analyst at a company developing AI services in Canada. Your boss comes to you with a problem. She says, "We need to increase our user base. **[CLICK]** Can you figure out why people aren't signing up?" Of course you'll say, "Yeah, I can do that!" On looking deeper into the problem, though, you observe that **[CLICK]** "people who aren't signing up" is pretty vague. Everyone who doesn't use your service is a massive group – that's hundreds of millions of people from all over the **[CLICK]** globe, from **[CLICK]** babies to **[CLICK]** grandparents. **[CLICK]** Or maybe you're more interested in **[CLICK]** people who used to use your AI

	service but canceled. Or perhaps you want to focus on **[CLICK]** people who use competing services. Studying each of these populations would lead to very different insights. If you focus on **[CLICK]** former users, you might learn about retention. If you look at **[CLICK]** competitors' users instead, you might better understand what features you're missing.
	Here's a process you can use to nail down your relevant population: 1. **[CLICK]** Start with your hypothesis or research question. In this case, it might be **[CLICK]** "Why aren't people signing up for our AI service?" 2. Then, **[CLICK]** identify the key characteristics that define your population of interest. For an AI service, this might include factors like **[CLICK]** age (are we interested in all ages or just certain demographics?), **[CLICK]** geographic location (are we looking globally or just in certain markets?), and **[CLICK]** technology use (are we only interested in people who already use other AI services?). 3. Next, **[CLICK]** consider practical constraints. While you might ideally want to collect data across the full population, that's probably not feasible. You might need to limit your population based on **[CLICK]** accessibility, **[CLICK]** budget, **[CLICK]** time, or **[CLICK]** ethical constraints. 4. Then, **[CLICK]** think about generalizability. That means how broad do you want your conclusions to be? The **[CLICK]** **[CLICK]** more specific your population, the **[CLICK]** more precise your insights might be, but the **[CLICK]** less generalizable they'll be to other groups. A **[CLICK]** **[CLICK]** less specific population leads to **[CLICK]** less precise but **[CLICK]** more generalizable results. 5. Finally, **[CLICK]** consult with stakeholders. **[CLICK]** Make sure your defined population aligns with the business goals. This step comes last so you can present a concrete idea for review. By following this process, you and your stakeholders might define the population as **[CLICK]** "internet-using adults aged 18-49 in Canada who currently use at least one AI service, but not ours." This clearly defined population can guide everything from your sampling method to how you interpret your results.
TH	Now you've seen how to identify your population. But as you've learned, you're unlikely to have access to the full extent of data. There are many ways to go about sampling your population. Join me in the next video to take a look at the gold standard: probabilistic methods.

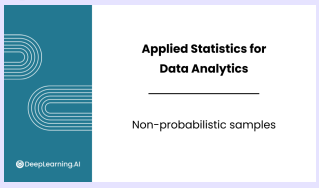
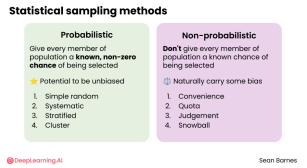
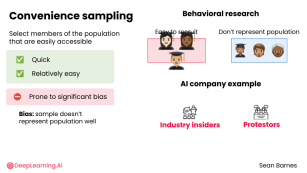
L1V3 – Probabilistic samples

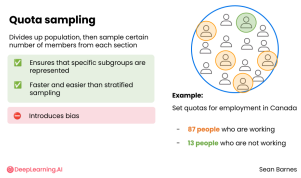
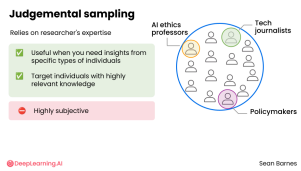
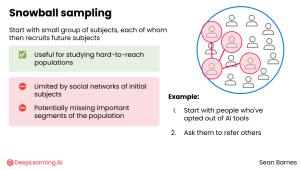
Visual	Script
TH 	Sampling is a critical skill for data analysts. Over the next few videos, you'll explore the tradeoffs involved in sampling as well as the biases it can introduce.
	**[CLICK]** As a data analyst, depending on the size of your team, you may **[CLICK]** work with data that has already been collected, or you may have to **[CLICK]** design a sample from scratch. Regardless, it's crucial for you to be able to understand the principles of sampling for a few key reasons: • First, understanding sampling allows you to **[CLICK]** interpret the data correctly, **[CLICK]** identifying potential biases and limitations in the data • It also allows you to **[CLICK]** communicate your results accurately. In particular, you will be able to **[CLICK]** explain how generalizable your conclusions are, meaning how accurately they reflect the population. • You will also be better able to **[CLICK]** collaborate with teams responsible for data collection, **[CLICK]** ensuring that best practices are implemented. Let's explore sampling through an example.
	Suppose you've been working at the Canadian AI company for a while and your CEO comes to you with a new problem. She wants to **[CLICK]** understand Canadian public opinion on AI to guide the company's roadmap. How would you go about collecting that data? Your first thought might be finding **[CLICK]** people who know what AI is. So you head to an **[CLICK]** AI conference, where you find a lot of enthusiastic **[CLICK]** industry insiders. You interview 100 attendees, and 95 of them express overwhelmingly **[CLICK]** positive views. Things seem promising, but as you're leaving the conference, you notice a group of **[CLICK]** protesters outside. Their signs display slogans about the potential negative impacts of AI. You interview 50 of them, and **[CLICK]** nearly all express skeptical views of AI. You've recorded two very different perspectives, but do either of these groups accurately represent overall public opinion on AI? Probably not. These are **[CLICK]** both samples, but they're not very good ones.

	As a data analyst, your goal is to **[CLICK]** analyze data that represents the entire population you're interested in - in this case, **[CLICK]** the adult population of Canada – **[CLICK]** not just the vocal supporters and opponents. Statistics has a lot to say on how to do that, so let's talk about some **[CLICK]** specific sampling methods, which fall into two categories: **[CLICK]** probabilistic and **[CLICK]** nonprobabilistic. Let's look at the probabilistic ones first.
	Probabilistic sampling methods **[CLICK]** give every member of the population a known, non-zero chance of being selected. These methods have the **[CLICK]** potential to be unbiased, which means that your sample is truly representative of the population. That's what you're aiming for. The four most important probabilistic samples are: • **[CLICK]** Simple random, • **[CLICK]** Systematic, • **[CLICK]** Stratified, and • **[CLICK]** Cluster.
	Simple random sampling gives every member of the population an equal chance of being selected. It's **[CLICK]** straightforward and **[CLICK]** provides a good chance of having a representative sample, but it can be **[CLICK]** challenging to carry out in large populations. • **[CLICK]** In the AI opinion example from earlier, a simple random sample could look like **[CLICK]** assigning every adult Canadian a number from 1 to around 30 million, the total population of Canada, then **[CLICK]** randomly selecting 1,000 of those numbers. You can then call up each person whose number was chosen to ask them their opinion about AI. It sounds tough, right? Not every Canadian has a phone number and it might be expensive to place many phone calls this way.
	In systematic sampling, you select every nth member of the population after a random starting point. For example, you might have a **[CLICK]** list of all registered voters in Canada, **[CLICK]** randomly select one person, and **[CLICK]** then choose every 1,000th person after them on the list. This method is **[CLICK]** straightforward, but **[CLICK]** can cause periodicity bias if there's a pattern in your list.For example, if voters were sorted by **[CLICK]** precinct, **[CLICK]** then by age within that precinct, it would be possible to sample only people with similar ages. **[CLICK]** If opinions on AI vary by age, this sample will systematically deviate from the population.
	Next up: stratified sampling. This method divides the population into groups, or strata, based on shared characteristics, then randomly samples from each group. **[CLICK]** It's useful when you want to make sure to sample from a diverse set of groups. **[CLICK]** Common strata include **[CLICK]** age, **[CLICK]** gender, and **[CLICK]** education level for people, or **[CLICK]** product categories,

	[CLICK] price ranges, and [CLICK] sales regions for products. • [CLICK] In the AI example, you could [CLICK] stratify by age group – [CLICK] 18-24, [CLICK] 25-29, and so on – [CLICK] and poll 100 people from each age group.
	Finally, you have cluster sampling, [CLICK] which is useful for large, geographically dispersed populations. You [CLICK] divide the population into clusters (often geographically), then randomly select entire clusters. [CLICK] It's efficient, but [CLICK] can introduce cluster-specific biases. • In the [CLICK] AI example, you could [CLICK] divide Canada up into different municipalities, [CLICK] then select 10 of them for your sample. You'll send a team to interview randomly selected people within each municipality. You can see how it might be easier to send teams out to just 10 areas, versus randomly sampling every Canadian adult, each of whom could live in the [CLICK] furthest corner of the country.
TH	These four methods form the core of probabilistic sampling, helping to select a sample that's representative of the population you're trying to understand. Now, follow me to the next video to learn about **non**-probabilistic sampling methods.

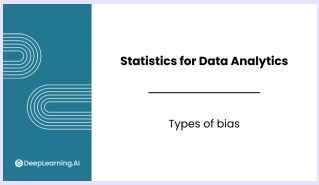
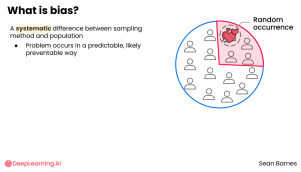
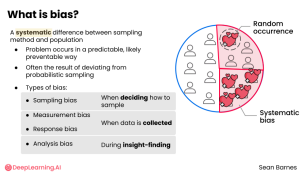
L1V4 – Non-probabilistic samples

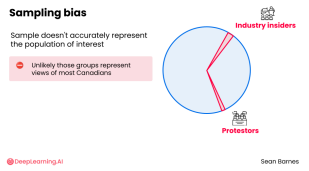
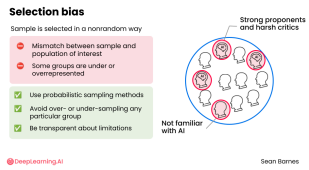
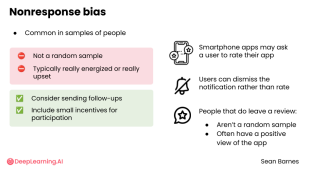
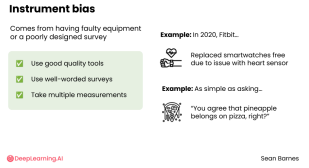
Visual	Script
TH 	When budget, time, or workforce constraints make probabilistic samples infeasible, you can use a non-probabilistic sample. Non-probabilistic samples are generally less rigorous, but often more practical to carry out.
	Nonprobabilistic sampling methods [CLICK] **don't** give every member of the population a known chance of being selected. These methods often arise from practical constraints and [CLICK] naturally carry some bias. The core non probabilistic methods you should know about as a data analyst are [CLICK] convenience, [CLICK] quota, [CLICK] judgment, and [CLICK] snowball samples. Let's get to it.
	In convenience sampling, you [CLICK] select members of the population that are easily accessible. On the upside, it's [CLICK] quick and [CLICK] relatively easy to do. However, it's [CLICK] prone to significant bias. Bias just means your [CLICK] sample doesn't represent your population well. You'll learn more about it in the next video.

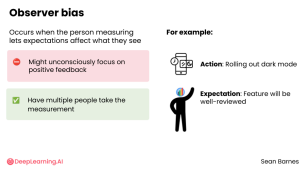
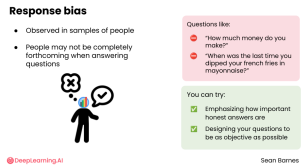
	A commonly cited example of convenience sampling is **[CLICK]** behavioral research. Because many researchers work at universities, they can **[CLICK]** easily recruit students to complete an experiment. But, these students **[CLICK]** don't represent the overall population very well. They're often younger than the average person, for example. • In the **[CLICK]** AI company example from the previous video, surveying just the **[CLICK]** attendees and protesters at one conference is a convenience sample. Just one location, no randomness.
	Quota sampling divides up the population based on certain characteristics, then samples a certain number of members from each section. You can think of it as the nonprobabilistic version of stratified sampling. It is often used to **[CLICK]** ensure that specific subgroups are represented. Like convenience sampling, it's **[CLICK]** faster and easier than stratified sampling, but **[CLICK]** introduces bias. • **[CLICK]** In the AI example, you might set quotas to match the employment demographics of Canada, where the employment rate is 87%. So, you might interview **[CLICK]** 87 people who are employed, and **[CLICK]** 13 who are not. But without random sampling, interviewers may inadvertently skew the results based on who they choose to question.
	Then you have judgmental sampling, which relies on the researcher's expertise to select the sample. **[CLICK]** It can be useful when you need insights from specific types of individuals. The benefit is that **[CLICK]** you can target individuals with highly relevant knowledge. The downside? **[CLICK]** It's highly subjective. • For the **[CLICK]** AI company, using judgmental sampling might mean specifically choosing to interview **[CLICK]** AI ethics professors, **[CLICK]** tech journalists, and **[CLICK]** policymakers, based on the belief that their opinions are particularly valuable.
	Finally, take a look at snowball sampling. You **[CLICK]** start with a small group of subjects, each of whom then recruits future subjects from among their friends, family, and coworkers. Your sample size grows like a rolling snowball. This method is particularly **[CLICK]** useful for studying hard-to-reach populations. For instance, undocumented immigrants may be difficult to source for an interview, but you can start with a few contacts who then refer others from their community. The big drawback is that your sample is **[CLICK]** limited to the social networks of your initial subjects, **[CLICK]** potentially missing important segments of the population. • In the **[CLICK]** AI example, snowball sampling might be used to study

	the opinions of people who've opted out of using AI tools. You could **[CLICK]** start with a few such individuals and **[CLICK]** ask them to refer others who share their habits.
TH	You've heard the term "bias" a few times now. It's one of the biggest issues with sampling. In the next video, you'll explore several common types of bias and how to mitigate them. I'll see you there.

L1V5 – Types of bias

Visual	Script
TH 	**Bias** in sampling happens when your sample doesn't represent your population of interest very well. Similar to how bias can negatively affect interactions between people, bias in sampling can lead to poor decision-making. Let's see how it happens.
	Here's a formal definition for bias in sampling: a systematic difference between the sample and the population that creates an inaccurate depiction of reality. "Systematic" is important here – it means that **[CLICK]** the problem occurs in a predictable, likely preventable way. • For example, if you're **[CLICK]** interviewing randomly selected Canadians about their opinion on AI, and one person happens to respond **[CLICK]** more positively than the average person, that's just a normal **[CLICK]** random occurrence. You expect some variation in opinion.
	• By contrast, if you head to a **[CLICK]** conference filled with AI researchers, you have reason to believe that **[CLICK]** most of these people will view AI more favorably than the average Canadian. That's a **[CLICK]** systematic bias. **[CLICK]** Bias is often the result of deviating from probabilistic sampling methods. Let's take a look at the most important **[CLICK]** types of bias: • **[CLICK]** Sampling bias, • **[CLICK]** Measurement bias, • **[CLICK]** Response bias, and • **[CLICK]** Analysis bias. Sampling bias crops up when a researcher is **[CLICK]** **deciding** how to sample the population. **[CLICK]** Measurement and response bias occur when data is **collected**. **[CLICK]** And analysis bias occurs during **insight-finding**.

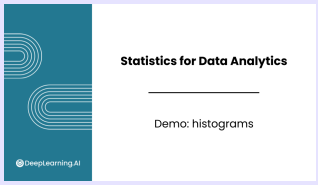
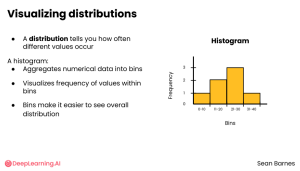
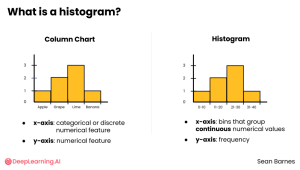
	Let's start with sampling bias. It's quite common, and it happens when a **[CLICK]** sample doesn't accurately represent the population of interest. This is the kind of bias that a researcher would run into if they interviewed **[CLICK]** people at an AI conference plus the **[CLICK]** protestors outside. **[CLICK]** It's unlikely those two extreme groups adequately represent the views of most Canadians.
	A common form of sampling bias is called selection bias. It happens when a **[CLICK]** sample is selected in a nonrandom way, causing a **[CLICK] [CLICK]** mismatch between the sample and the population of interest. As a result, **[CLICK]** some groups in the population are under or overrepresented. **[CLICK]** In the AI conference example, **[CLICK]** strong proponents and harsh critics are overrepresented, while people who **[CLICK]** aren't very familiar with AI are underrepresented. To avoid selection bias, **[CLICK]** use probabilistic sampling methods. **[CLICK]** Avoid over- or under-sampling any particular group. You should also **[CLICK]** be transparent about the limitations your sample may have. If you know your sample introduced selection bias, explain how that bias affects your conclusions.
	Another common type of sampling bias is nonresponse bias. It's **[CLICK]** common in samples of people. For example, **[CLICK]** smartphone apps may ask a user to rate their app after using it for a while. But **[CLICK]** users can dismiss the notification rather than actually rating the app. By consequence, the **[CLICK]** people that do leave a review **[CLICK]** aren't a random sample. They **[CLICK]** often have a positive view of the app, since they're willing to accept the prompt and invest time in a review. To counteract nonresponse bias, **[CLICK]** consider sending follow-ups or **[CLICK]** even including small incentives for participation.
	Next up: two key types of measurement bias. The first one, instrument bias, **[CLICK]** comes from having faulty equipment or a poorly designed survey. **[CLICK]** For example, in 2020, Fitbit **[CLICK]** replaced some of its smartwatches free of charge due to an issue with their heart rate sensor. But instrument bias can also be **[CLICK]** as simple as asking, **[CLICK]** "You agree that pineapple belongs on pizza, right?", which is a leading question. Make sure you're **[CLICK]** **[CLICK]** using good quality tools – **[CLICK]** and well-worded surveys – **[CLICK]** and take multiple measurements if you can.

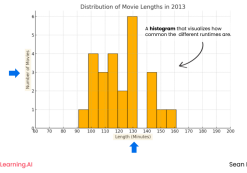
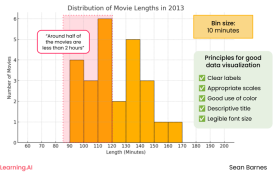
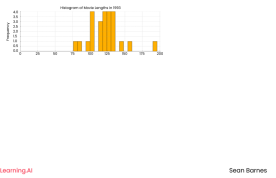
	Another type of measurement bias is observer bias. It **[CLICK]** occurs when the person doing the measuring lets their own expectations affect what they see. **[CLICK]** For example, say your company is **[CLICK]** rolling out dark mode for one of its apps. If you as the data analyst **[CLICK]** expect this feature to be well-reviewed, then when you're interviewing a sample of users, you **[CLICK]** might unconsciously focus on their positive feedback. Try **[CLICK]** having multiple people take the measurement if possible.
	Another type of bias is response bias, which is often **[CLICK]** observed in samples of people. Essentially, **[CLICK]** people may not be completely forthcoming when answering questions. So **[CLICK]** questions like **[CLICK]** "How much money do you make?" or **[CLICK]** "When was the last time you dipped your french fries in mayonnaise?" might elicit untruthful responses, even on an anonymous survey. Mitigating this type of bias isn't easy. **[CLICK]** You can try **[CLICK]** emphasizing how important honest answers are, or **[CLICK]** designing your questions to be as objective as possible.
	Finally, let's consider analysis bias. You've sampled your population and collected your data, but the potential for bias isn't over! If you're not careful, **[CLICK]** the way you perform your analysis can lead to results that don't reflect the truth. The biggest pitfall is **[CLICK]** confirmation bias – looking for evidence for something you already believe, rather than looking at all the evidence objectively. For example, a **[CLICK]** product manager might **[CLICK]** hope that a new feature is improving user retention. They may be under some pressure to show that the feature was worth developing. They might **[CLICK]** analyze 10 metrics, but choose to **[CLICK]** focus on the two that show the most positive results. It's okay to have a hypothesis or a goal, of course, but **[CLICK]** stay open-minded and **[CLICK]** let the data tell its own story, even if it's not the one you were hoping to hear.
TH	Bias is often unavoidable, and most samples will contain some degree of it. However, by following best practices, in most cases you can mitigate its effects. That takes you to the end of this lesson! Great work so far. You've learned the core concepts of populations and sampling, which underlie all of statistics. After the practice assessment for this lesson, join me in the next module to learn how to characterize samples through measures of central tendency,

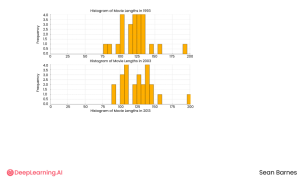
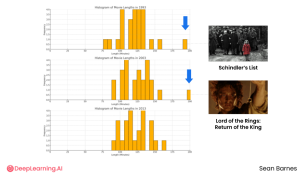
	variability, and skewness. I'll see you in the next lesson.

Lesson 2 – Central tendency

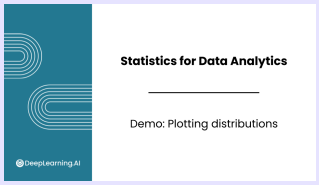
L2V1 – Histograms

Visual	Script
TH 	Once you've collected a sample, it's often useful to start your analysis by visualizing its distribution.
	**[CLICK]** A distribution tells you how often different values occur in a population or in your sample. Descriptive statistics is all about characterizing the distribution of your data. One useful visualization for distributions is **[CLICK]** the histogram. A histogram **[CLICK]** aggregates numerical data into bins and **[CLICK]** visualizes the frequency of the values within those bins. The purpose of the bins is to make it **[CLICK]** easier to see the overall distribution by showing how common ranges of values are.
	Histograms are a special version of a **[CLICK]** column chart. In a column chart, you are trying to compare a numerical feature across different categories, so you display the **[CLICK]** categorical (or discrete numerical feature) on the x axis and **[CLICK]** the numerical feature on the y axis. In a histogram, you're interested in visualizing the frequency of different values in a continuous numerical feature. On the x-axis, it plots **[CLICK]** bins that group the **continuous** numerical values into categories, with the **[CLICK]** frequency on the y axis. Frequency means the number of observations in the data that fall into a particular bin. You may also see the proportion of observations plotted on the y axis, rather than the raw frequency.
TH	Let's work with some real data together. Do you feel like movies have been getting longer lately? **[pause]** Maybe it's just my attention span, but I've heard this hypothesis several times lately. Say you wanted to figure out how long movies typically are, and if they've been getting

	longer over time. You sample the 25 most popular movies from 2013. What can you do to describe this distribution of movie durations?
🖥️SC 🔗**Spreadsheet** 🔗**Source data**	So here's the data of movie durations of the top 25 most popular movies in the year 2013. If you'd like to explore the data yourself, you can find this spreadsheet in the downloads tab accompanying this video. The columns in the data set are the name of the movie, the year it was released – all 2013 – the rating out of 10 – this is from the International Movie Database or IMDB, the number of IMDB ratings, and the duration of the movie. My personal fav here was Iron Man 3, but maybe you liked The Great Gatsby. You learned a moment ago that a distribution represents how often values in the sample data occur. You have a sample of movies here. So what **values** should you examine to answer the question? **[pause for learner to think]** Those would be the durations of the movies!
 https://i.imgur.com/Y15Hrj5.png	Here's a **[CLICK]** histogram that visualizes how common the different movie durations are. On the x axis is the **[CLICK]** duration of the movie in minutes, and **[CLICK]** the y axis shows the number of movies with that duration. Note that duration is a continuous numerical feature, which has been grouped into **[CLICK]** 7-minute bins, giving this histogram **[CLICK]** 10 bins. Any movie with a length between 91 and 98 minutes is represented by **[CLICK]** this column.
 https://i.imgur.com/0mqXBsD.png	Here's the same data but this time in histogram with **[CLICK]** bins of 10 minutes, which are a bit easier to interpret, though the overall picture of the data is similar. These bin sizes make it easier to say something like **[CLICK]** "around half of the movies are less than 2 hours." Note that too few bins can overly simplify the data, and too many bins can make it difficult to identify any overall patterns. Aside from bin size, keep in mind your **[CLICK]** principles for good data visualization: **[CLICK]** clear labels, **[CLICK]** appropriate scales, **[CLICK]** good use of color, **[CLICK]** descriptive title. And make sure your **[CLICK]** font size is legible for your audience.
 	Multiple histograms can also be plotted next to each other to compare distributions. Here are the distributions of movies in 1993 at the top,

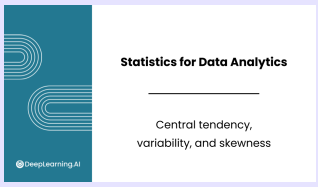
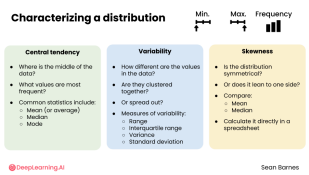
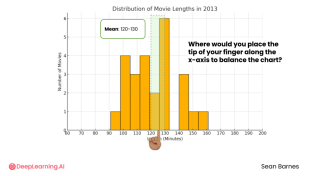
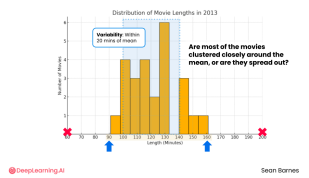
	then 2003,
 https://i.imgur.com/zCcXnuz.png	then 2013 on the bottom. These all use the same bin size, to make comparison easier. What can you say about the change in durations over time? **[pause for learner to reflect]** It's a bit hard to say. Perhaps there was a small shift to the right? They look pretty consistent overall. **[CLICK]** Check out those 200-minute movies though. **[CLICK]** Can you guess what they are? **[brief pause]** in 1993, **[CLICK]** that was Schindler's List; and in 2003, it was **[CLICK]** The Lord of the Rings: Return of the King.
🎙️ TH	Histograms help visualize the distribution of a continuous numerical feature. Follow me to the next video to see how to create them in Google Sheets.

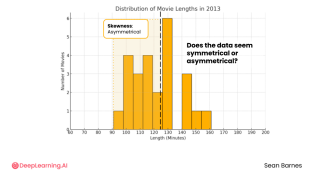
L2V2 – Demo: Plotting distributions

Visual	Script
🎙️ TH 	Now that you've learned how to interpret histograms, let's take a look at how to plot distributions, using both histograms and column charts.
🖥️ SC 🔗 **Start** 🔗 **Solution** 🔗 **Source**	The demos in this module use a subset of the Lending Tree Loan Dataset. The data set includes thousands of loans made through the lending tree platform, which allows individuals to lend to other individuals. • Each row represents a particular loan, and each column includes features about the person who took out the loan, like their job title **[A]** and annual income **[F]**, as well as information about the loan like the amount **[K]**. Remember, if you'd like to follow along with the demo, you can find this spreadsheet and the solution in the downloads tab. Imagine you are thinking of becoming a lender on this platform. Before committing to anything, you want to better understand the risks involved. You can perform a statistical analysis on the existing loans to try to identify the level of risk of each loan, and which factors seem to affect risk.
🖥️ SC	As a potential lender, you might be interested in the paid interest feature **[P]**. This feature essentially represents the amount of money the lender has made

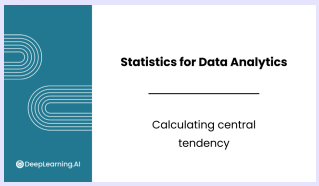
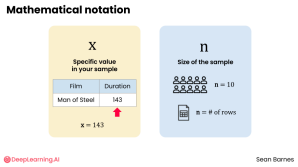
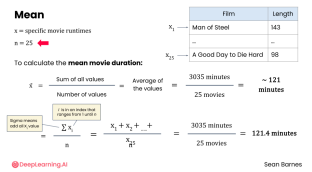
	so far on the loan. **Histogram** ● Select column P ● Insert Chart → should be a histogram by default, or change it to one ● What if you want to customize the bar width? ● Customize → Histogram → Bucket size → 50 ● Move to own sheet “paid_interest_chart” ● Title → Distribution of Paid Interest ● Comment on chart → looks like most loans have paid out around $50 to $600 dollars, with some loans paying a lot more than that For discrete features you need to use a column chart rather than a histogram. You already learned this visualization in Data Analytics Foundations, but let’s see how a column chart can show you the overall distribution of your data. One categorical feature present in this dataset is the loan **grade**. The grade is the quality score given to each loan based on the borrower's credit history and some other factors. It’s useful to you as a potential lender to better understand the distribution of good and bad loans on the platform **Column chart** ● Highlight the column N2:N ● Insert Chart - column chart ● Notice that the grades on the x axis are not in order, making it hard to visualize the distribution. ● You can sort the data to avoid this issue ○ Apply filter → sort A to Z ● Move to own sheet “grade_chart” ● Add title “Distribution of Loan Grades” ● Notice the Aggregate box. This is on by default, and what it does is count the frequency of each grade in the dataset, rather than trying to display each value individually ● Great! Now you can see many loans are in the higher grades, A, B, and C. That may motivate you more as an investor, but it’s worth investigating these terms more. Is a C significantly worse than an A? It’s worth looking into
🎙️ TH	Great work on histograms! Join me in the next video to learn how to complement this visualization approach by calculating descriptive statistics.

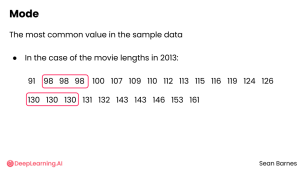
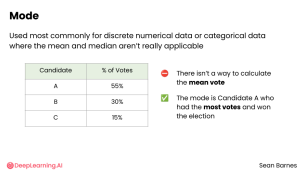
L2V3 – Central tendency, variability, and skewness

Visual	Script
TH 	Once you've visualized the data in your sample, you'll want to describe it. But how? What are the most useful statistics for explaining what your sample is like?
	You can characterize a distribution through measures of **[CLICK]** central tendency, **[CLICK]** variability, and **[CLICK]** skewness. Here are the questions each type of descriptive statistic answers: • For central tendency: **[CLICK]** where is the middle of the data? **[CLICK]** What values are most frequent? Common statistics include **[CLICK]** mean (or average), **[CLICK]** median, and **[CLICK]** mode. • For variability: **[CLICK]** how different are the values in the data? **[CLICK]** Are they clustered together, or spread out? There are many **[CLICK]** measures of variability, including **[CLICK]** range, **[CLICK]** interquartile range, **[CLICK]** variance, and **[CLICK]** standard deviation. • For skewness: **[CLICK]** is the distribution symmetrical, or does it lean to one side? Skewness characterizes the direction and extent of any asymmetry in the data. You can **[CLICK]** compare the **[CLICK]** mean and **[CLICK]** median to approximate skewness, or **[CLICK]** calculate it directly in a spreadsheet. You've learned many other useful descriptive statistics in Data Analytics Foundations, so don't forget about them! These include **[CLICK]** min, **[CLICK]** max, and **[CLICK]** frequency.
	Let's return to the movie durations data. In an earlier video, you saw this histogram of the durations. For now, let's explore the central tendency, variability, and skewness of this data by eye, then you'll see how to calculate these measures. • For the mean, imagine you are trying to balance the distribution on the tip of your finger. You're finding the center of mass of the data. **[CLICK]** Where would you place the tip of your finger along the x-axis to balance the chart? **[pause for learner to think]** The center of mass looks to be **[CLICK]** somewhere around 120 or 130 minutes.
	• What about variability? Look at how **[CLICK]** spread out the movie durations are on the x-axis. **[CLICK]** Are most of the movies clustered closely around the mean, or are they spread out? **[pause for learner to think]** If the mean is around 120, most of the durations are **[CLICK]** within 20 minutes to the left and right of the mean. That's ⅙ of the average duration, so they're relatively close together. There aren't any movies in this sample that are **[CLICK]** 60 minutes or 200 minutes long.

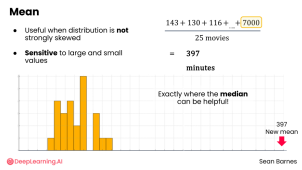
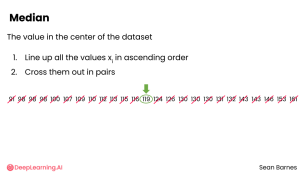
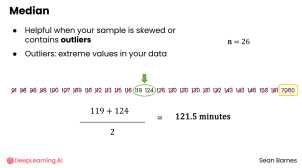
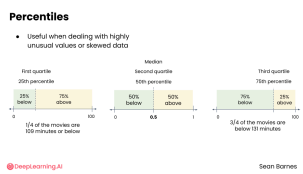
	• How about skewness? **[CLICK]** Does this data seem symmetrical or asymmetrical? **[pause for learner to think]** If you **[CLICK]**draw a line down the center here at the mean, you can see some asymmetry. There seem to be **[CLICK]** more movies on the left side, with shorter durations. However, the skewness isn't extreme.
TH	Now that you have an intuition about what these different statistics calculate, follow me to the next video to see how to calculate central tendency.

L2V4 – Central tendency: mean and mode

Visual	Script
TH 	You have a few different tools to measure central tendency: mean, median, and mode. When should you use each one? Let's look at how to calculate and interpret these measures, starting with mean and mode.
	First, you'll need a little bit of notation. Statisticians use mathematical notation to represent how the same calculation applies to **all** circumstances. It's common to use **[CLICK]** lowercase x to represent a **[CLICK]** specific value you found in your sample, like a **[CLICK]** specific movie duration. You'll see x used in many upcoming formulas. You'll also see **[CLICK]** lowercase n used to represent the **[CLICK]** size of the sample. So for example if you interview **[CLICK]** 10 people, **[CLICK]** n equals 10. In a sample dataset that you might have in a **[CLICK]** spreadsheet, **[CLICK]** n is the number of rows or observations. With these two terms now defined, we can now build up the formulas used to calculate these measures.
	Let's return to the movie duration example. You may already be familiar with how to calculate the mean, but you'll see it formalized using mathematical notation here. For this example, **[CLICK]** x represents specific movie durations. Note that the number of possible values for x is quite large, since movies can have many different durations, even if you round to the nearest minute. The sample size of the data, **[CLICK]** n, is 25 since you're working with the top 25 most popular movies from 2013. Let's walk through how to **[CLICK]** calculate the mean movie duration of your

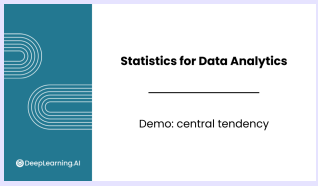
	sample. The sample mean is written as **[CLICK]** x-bar. • The mean is calculated by **[CLICK]** summing up all of the values in the sample and **[CLICK]** dividing by the **[CLICK]** number of values, or the sample size. This calculation is **[CLICK]** an average of the values, which is why you may hear mean and average sometimes used interchangeably. • Intuitively what's happening is that you're distributing the total sum of all values equally among the number of values in the sample. So in total all of these movies are **[CLICK]** 3,035 minutes, and **[CLICK]** there are 25 of them. If these movies were all the same duration, each one would be about **[CLICK]** 121 minutes. • Using the notation you saw earlier, the mean would be the **[CLICK]** sum of all of the values x_i , **[CLICK]** where *i* is in an index that ranges from 1 (the first data point in the sample) all the way until n (the last data point in the sample). So **[CLICK]** x_1 would correspond to the duration of Man of Steel, which was 143 minutes, and **[CLICK]** x_{25} would correspond to the duration of A Good Day to Die Hard, which was 98 minutes. Remember, **[CLICK]** n is the sample size of the data, in this case 25. Don't be scared of this big Greek letter Sigma, it just **[CLICK]** means to add all of these x_i values up. And lastly, **[CLICK]** divide the sum by the sample size of **[CLICK]** *n*. • Here's what that looks like all written out. You have **[CLICK]** $x_1 + x_2$, all the way to x_{25}, since you had 25 movies in your sample. The sum of all of these individual movie durations gives you the **[CLICK]** 3,035 total minutes that you've already seen. Then, you'll **[CLICK]** divide by the sample size of **[CLICK]** 25 to get the final result of **[CLICK]** 121.4 minutes. **[pause for absorption]**
	Another common measure of centrality is the mode: the most common value in the sample data. **[CLICK]** In the case of the movie durations in 2013, you actually have two modes: **[CLICK]** 3 movies were 98 minutes long, and 3 movies were 130 minutes long.
	Mode is much more commonly **[CLICK]** used for discrete numerical data or categorical data where the **[CLICK]** mean and median aren't really applicable. Voting results are a great example. If there are **[CLICK]** three candidates, A, B, and C, with 55%, 30%, and 15% of the votes, respectively, **[CLICK]** there isn't a way to calculate the mean vote. What would you say is the "center of mass" of this data? **[pause for thought]**In this case, it would be **[CLICK]** candidate A, because they earned the most votes in the election. So, the **[CLICK]** mode being Candidate A is quite useful for describing this data.
TH	Now you've seen how to use mean and mode appropriately, but there's one more useful measure of central tendency. Follow me to the next video to calculate the median.

L2V5 – Central tendency: median

Visual	Script
🎙️ TH	The median has a more complicated formal definition compared with mean and mode. For now, let's get an intuition of how it works.
	While the mean is especially **[CLICK]** useful when your distribution is not strongly skewed, it's quite **[CLICK]** sensitive to very large and very small values. Say the last movie in your **[CLICK]** 2013 dataset was actually the longest movie ever made, and it was **[CLICK]** 7000 minutes long! This very rare case would change the mean to **[CLICK]** 397 minutes. That's not the most useful way to think about the "middle" of the data. Looking at the **[CLICK]** histogram again, none of the observations are anywhere **[CLICK]** near that mark. This scenario is **[CLICK]** exactly when the **median** can be helpful!
	Median is calculated by selecting **[CLICK]** the value in the center of the dataset. One way to calculate it is to **[CLICK]** line up all the values x_i from your sample in ascending order. Then, **[CLICK]** cross them out in pairs. **[CLICK]** One on the left, **[CLICK]** one on the right. **[CLICK]** One on the left, **[CLICK]** one on the right, and **[CLICK]** eventually **[CLICK]** you'll **[CLICK]** come to the **[CLICK]** middle. Since there are 25 values in the movie duration dataset, there's one value left and that's **[CLICK]** 119. Of course computers are able to efficiently find the middle value, so you won't be crossing these out by hand.
	The median is **[CLICK]** helpful when your sample is skewed or contains outliers. **[CLICK]** Outliers are extreme values in your data. Imagine adding the longest movie ever – the one that's **[CLICK]** 7000 minutes back into the dataset. Now you have **[CLICK]** 26 movies, so n is even. **[CLICK]** Repeat the process to calculate the median, and you can see that 7000 minutes gets thrown out right away. **[CLICK]** In the end, there are two remaining values: **[CLICK]** 119 and 124. To find the median, take the average of these two numbers by **[CLICK]** adding 119 and 124 together, **[CLICK]** then dividing the result by 2. This gives you a median of **[CLICK]** 121.5. That's pretty close to the original median of 119. Compare that 2.5 minute difference to the 276-minute difference between the original and new **means**. The median represents the center of this modified sample data much better.
	The median is closely tied to the notion of percentiles. Percentiles are a powerful way to understand how values in your dataset are distributed. **[CLICK]** The median you just calculated is actually the 50th percentile - it's the value that's right in the middle, **[CLICK]** with 50% of the data below it and **[CLICK]** 50% above it. Some applications, like Google Sheets, have you select a number between **[CLICK]** 0 and 1 to calculate a percentile. So in that case,

	[CLICK] 0.5 would be the 50th percentile. But you're not limited to just the 50th percentile, the median. You can calculate any percentile from 0 (the minimum) to 100 (the maximum). For example, the **[CLICK]** 25th percentile, also known as the first quartile, is the value below which **[CLICK]** 25% of your data falls, in this case 109 minutes. So a **[CLICK]** quarter of the movies in this sample were 109 minutes or below. **[CLICK]** The second quartile is the median of 119, as you just saw. 50% above, 50% below. **[CLICK]** The 75th percentile, or third quartile, is the value below which 75% of your data falls, in this case 131. **[CLICK]** So three quarters of the movies are below 131 minutes. Percentiles are particularly **[CLICK]** useful when dealing with highly unusual values or skewed data. Remember the 7000-minute movie? It might dramatically affect the mean, but it has minimal impact on most percentiles. The 90th percentile, for instance, tells us about the longer movies in the sample without being swayed by this extreme value.
🎙️ TH	Now you've seen how to select and calculate an appropriate measure of central tendency, one of the most crucial descriptive statistics you can use to describe a sample. Follow me to the next video to calculate these measures in a spreadsheet.

L2V6 – Demo: central tendency

Visual	**Script**
🎙️ TH 	Let's see how you can use a spreadsheet to calculate central tendency in order to answer a business question. If you need a refresher on spreadsheet functions, I encourage you to revisit the Data Analytics Foundations course, where you can learn all the functions you'll see in this demo.
🖥️ SC 🔗 **Start here** 🔗 **Solution**	Recall that you previously created a histogram for the **paid interest** feature, essentially how much profit had accrued on the loan. Let's complement the histogram with some descriptive statistics, to help you get an idea of how much money you could be making from the interest on loans!

And don't forget, if you'd like to follow along with the demo, you can find this spreadsheet and the solution in the downloads tab.

- Mean:
  - B2 → =AVERAGE(Data!Q:Q)
- Median:
  - B3 → =MEDIAN(Data!Q:Q)
- Percentile:
  - B4 → =PERCENTILE(Data!Q:Q,0.1).
    - Here you put the percentile you want, a number between 0 and 1.
    - If you want a different percentile you change the last value,
  - B5 → =PERCENTILE(Data!Q:Q,0.5)
    - Same values as the MEDIAN
  - B6 → =PERCENTILE(Data!Q:Q,0.99)
- Mode:
  - B7 → =MODE(Data!Q:Q)
  - This isn't really helpful, as your most common value may not be in the center

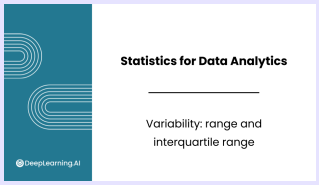
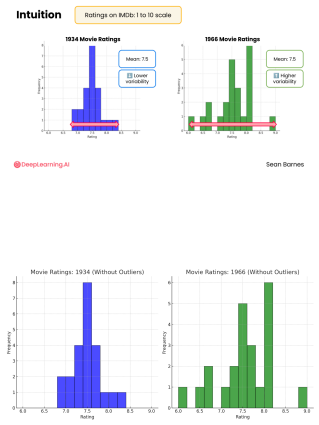
However, say you're interested in finding the center of mass of homeownership, to better understand the kind of collateral that is available to those who are taking out loans.

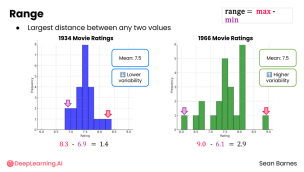
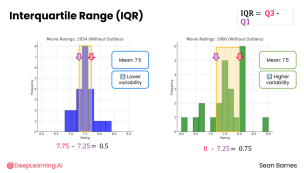
- Look at column D → It turns out you can't really find the median of this feature, what is the average of these?
- The mode function only uses numbers, so you can use homeownership numeric which has 1 for rent, 2 for mortgage and 3 for own.
  - Note that if you take the mean of this, you will get a number, but how would you interpret that? If the average is 1.5 for example, that doesn't correspond to a meaningful category. This operation also implies an order to the data that may not be accurate
- So, now take the MODE of this categorical column which will give you the center of mass
- Mode:
  - C7 → =MODE(Data!E:E)

	○ So mortgage is the most common! Helps you get a better picture of where the people taking out loans are in life and the kinds of collateral they might have.
TH	By calculating central tendency for the paid interest, you're starting to understand the center of mass of the sample data, which gives you a sense of roughly what payout you can expect on average. Once you've completed the practice assessment and the practice lab for this lesson, join me in the next one to learn more about variability and skewness in your data.

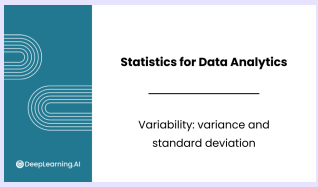
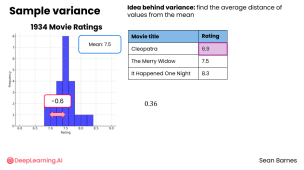
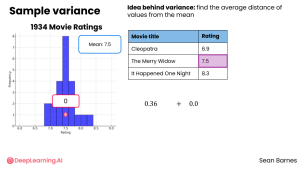
Lesson 3 – Variability and skewness

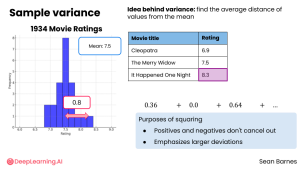
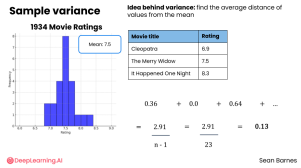
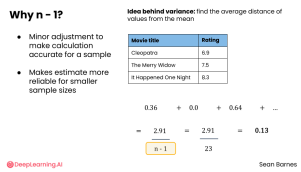
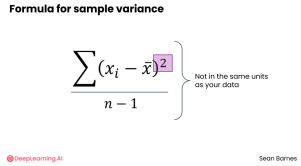
L3V1 – Variability: range and interquartile range

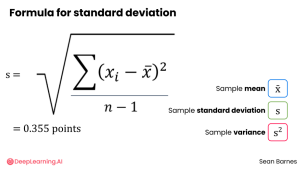
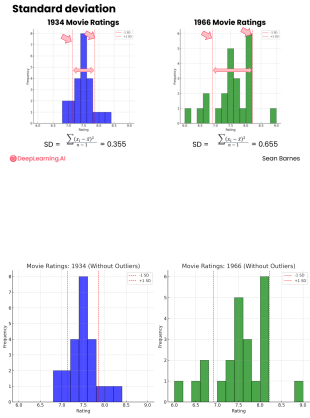
Visual	Script
TH  🔗Source	In the 2020 Tokyo Olympics, elite swimmer Margaret Mac Neil won the women's 100 meter butterfly swim event by 0.05 seconds. **[brief pause for effect]** In fact, all 8 swimmers that competed in the finals achieved times within 1 and a half seconds of each other. That's a very, very small difference. Contrast that with your local high school swim tryouts. You might see the fastest swimmer finish in just 70 seconds, while others take 200 seconds or more. The range is much broader because the skill levels are more varied. This measure of spread across the data is known as *variability*: a measurement of how tightly or loosely data points cluster around the mean.
 https://i.imgur.com/dLzfL5v.png	Let's start with some intuition. You saw earlier that the movie dataset also includes **[CLICK]** ratings, so you can compare how movies have been rated on the International Movie Database or IMDb, which is on a 1 to 10 scale. Let's zoom in on two years, **[CLICK]** 1934 and **[CLICK]** 1966. Take a look at **[CLICK]** these two histograms. What can you tell about the central tendency of these two years? **[pause for thought]** It turns out that on average, movies in 1934 and 1966 both have the **[CLICK]** same mean rating of 7.5. A quick note, by the way, that one outlier has been removed from each year, so there are only 24 movies here. Even though these two years have the same mean score, the distributions look quite different. What major difference can you spot? **[pause for thought]** Well, the distribution in 1934 looks **[CLICK]** tightly clustered around the mean

	of 7.5, while the distribution for 1966 is **[CLICK]** much more spread out. The values are more different from each other. In other words, 1966 has **[CLICK]** higher variability, while 1934 has **[CLICK]** lower variability. Let's look at the different tools you have available for calculating variability.
	First up: range. Range is simply the **[CLICK]** max minus the min. It answers the question, what is the **[CLICK]** largest distance between any two values? It's a simple yet useful measure of variability. In 1934, the highest rated movie was an **[CLICK]** 8.3, with the lowest a **[CLICK]** 6.9, making the range of 8.3 minus 6.9 a difference of 1.4. Contrast that year with 1966, which had a max rating of **[CLICK]** 9.0 and a minimum rating of **[CLICK]** 6.1, making the range **[CLICK]** 2.9, more than twice as large. Calculating the range is a quick way to see how spread out the ratings are, and whether a year had consistent or inconsistently rated movies. In this case, it looks like 1934 had more consistent ratings compared with 1966.
	A similar measure of variability you can use in conjunction with the median is the interquartile range, or IQR. It's calculated by **[CLICK]** subtracting the first quartile from the third quartile. Remember the first quartile defines the lowest 25% of the data and the third quartile defines the upper 75% of the data. That difference means that the IQR contains the middle 50% of the data. For 1934, the **[CLICK]** first quartile is 7.25, the median is 7.5, and the **[CLICK]** third quartile is 7.75, making the **[CLICK]** IQR 0.5. For 1966, the **[CLICK]** first quartile is also 7.25, the median is 7.6, so slightly higher than 1934, and the **[CLICK]** third quartile is 8, making the **[CLICK]** IQR 0.75, or 50% wider. So you can see IQR follows a similar pattern to the range: higher for 1966 compared with 1934, reflecting the greater variability in the data.
🎙️ TH	You've seen two common measures of variability: range and interquartile range. In the next video, you will learn about variance and standard deviation, two more useful measures of variability.

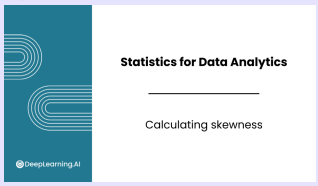
L3V2 – Variability: variance and standard deviation

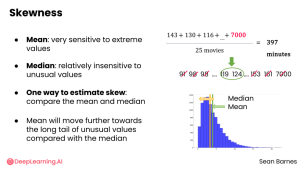
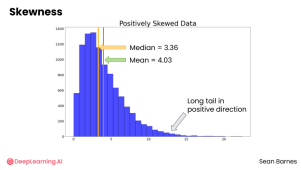
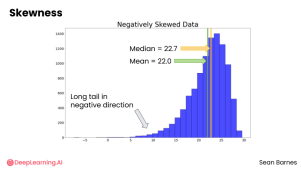
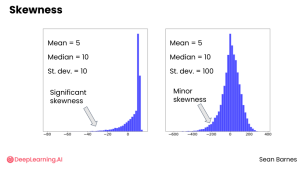
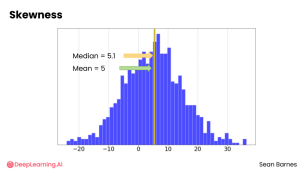
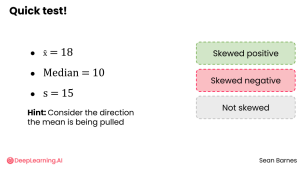
Visual	Script
TH 	Now let's explore variance. Variance is a more complex calculation than range and IQR, and is the foundation of many other calculations in statistics, including the standard deviation.
 See cell G73 for values – https://docs.google.com/spreadsheets/d/1CVrznZIt1ieQB2Obr4f9RZFymMjp3pIf33jPKaFcEuk/edit?gid=1892502851#gid=1892502851	Remember that the whole point of calculating variability is to quantify how spread out the values are. You can step through the calculation for variance using the **[CLICK]** movie ratings from 1934. The **[CLICK]** idea behind variance is roughly to find the average distance of the individual values from the mean. The process for calculating it involves looking at **[CLICK]** each value in your sample, in this case each movie rating, finding the **[CLICK]** difference between that value and the mean, **[CLICK]** squaring that difference, and then
	repeat **[CLICK]** that **[CLICK]** for all **[CLICK]** the values. Then, you **[CLICK]** sum all of these squared differences and **[CLICK]** divide by one less than your sample size. We'll discuss the "one less" point in more detail shortly. The squared differences are also sometimes called **[CLICK]** "squared deviations", with deviation just meaning how different each value is from the mean.
	In the case of the 1934 movies, you can start with Cleopatra, the lowest rated movie. Start with the movie's rating, **[CLICK]** 6.9. Subtract the mean of 7.5 from it giving you **[CLICK]** negative 0.6. Then square that, **[CLICK]** 0.36. Great, pause there.
	Now you need to do the same for each other movie. So for example, The Merry Widow with a rating of **[CLICK]** 7.5, subtract the mean from that and you get **[CLICK]** 0, squared is still **[CLICK]** 0.

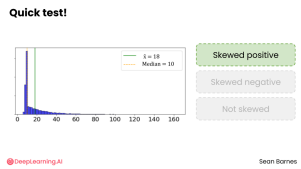
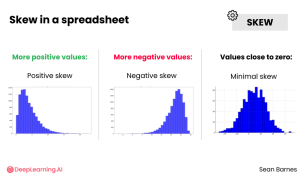
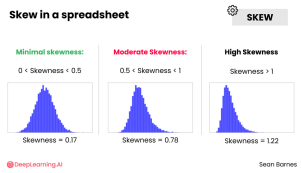
	How about **[mysterious voice]** It Happened One Night, which received a rating of **[CLICK]** 8.3? Subtract the mean from 8.3 and you get **[CLICK]** 0.8, squared is **[CLICK]** 0.64. And you'll do this for **[CLICK]** each of the values. ~~Here's how it looks.~~ Notice how each of these values is either 0 or positive. That's one **[CLICK]** purpose of the squaring operation: to make sure that when you add these values all up, you **[CLICK]** don't have positive and negative values canceling each other out. Another effect of squaring is that it **[CLICK]** emphasizes larger deviations. So observations that are far away from the mean contribute much more significantly to the overall variance!
	Your final step is to divide the **[CLICK]** total of these squared differences by the **[CLICK]** sample size n - 1, which in this case is **[CLICK]** 23. The result you get is **[CLICK]** 0.13.
	A quick note here, you might be wondering why you don't divide by just *n*, and calculate the average sum of squared deviations. That's great intuition! However, n - 1 is **[CLICK]** a minor adjustment that helps make your calculation more accurate when working with a sample rather than the entire population. By using n - 1, you're applying a small correction that **[CLICK]** makes your estimate more reliable, especially for smaller sample sizes. Don't worry too much about the exact reasons for this adjustment right now; it's a bit of an advanced concept.
	Compare the variance you just calculated with the **[CLICK]** variance for 1966, which is 0.42, almost three times higher. So this calculation is confirming your intuition that the movie ratings in 1966 were much more spread out.
	Now that you've calculated it through, let's formalize the definition for sample variance. Each value in your sample is **[CLICK]** x sub i, and you're subtracting the **[CLICK]** sample mean, x-bar from that. **[CLICK]** Square this up. Do you remember the notation for summing up all these individual values? **[pause for learner to think]** That would be **[CLICK]** sigma, the big E-looking symbol. And finally, **[CLICK]** divide by n minus 1. Now one issue with interpreting variance is that, because you **[CLICK]** squared the differences between the mean and each value, the variance is **[CLICK]** not in the same units as your data. Right now, your variance is in the units points-squared. Awkward.

	However, if you just **[CLICK]** take the square root, you'll get back to having a **[CLICK]** calculation in points, in this case 0.355. This square root of the variance is called the **[CLICK]** standard deviation. **[brief pause]** The sample standard deviation is written as **[CLICK]** s. So you already saw **[CLICK]** x-bar for the sample mean, now you have **[CLICK]** s for the standard deviation. Sample variance is written **[CLICK]** s-squared, since it's the standard deviation squared.
 https://i.imgur.com/hVciiKO.png	Let's compare with the **[CLICK]** standard deviation for movie ratings from 1966, which is 0.655. We have the same interpretation as before, that 1966 has a higher standard deviation. Now you can **[CLICK]** visualize these statistics on the histograms from before. The dashed red lines each indicate the mean minus one standard deviation on the left and the mean plus one standard deviation on the right. The difference on the left is **[CLICK]** much more narrow, while the difference on the right is **[CLICK]** wider. It looks like the values in the 1966 histogram are almost pulling those dashed lines outward, which mathematically, is exactly what happens! In practice, you'll see both standard deviation and variance, since variance is used to derive multiple other calculations and standard deviation is commonly used to characterize your distributions.
🎙️ TH	Great work on calculating these measures of variability! Join me in the next video to better understand the third key measure of your distribution: skewness. I'll see you there.

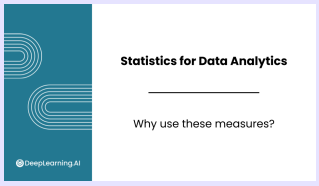
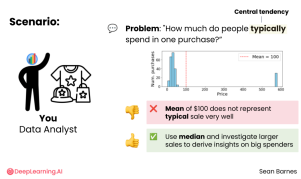
L3V3 – Skewness

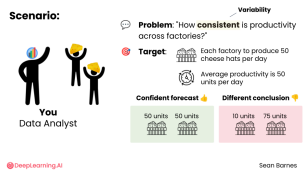
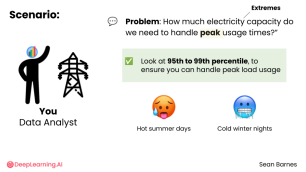
Visual	Script
🎙️ TH  Peek formula onto screen in whichever direction Sean sideyes: $b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$	Unlike mean, median, variance, and standard deviation … skewness isn't regularly calculated manually. It has a rather complicated formula … **[formula peeks dramatically onto the screen, Sean sideyes it]** No! Go away! It's quite a useful concept, though. Let me show you a quick shorthand for estimating it.

 Show histogram of positive skew (mean > median)	Earlier you saw how the mean is **[CLICK]** very sensitive to extreme values. That's because the mean distributes the total of all the values over the sample size. You also saw that the median is **[CLICK]** relatively insensitive to these types of unusual values. So, **[CLICK]** one way to estimate skew is to compare the mean and median. When the values are asymmetrical, the **[CLICK]** mean will move much further towards the long tail of unusual values compared with the median.
	Here's what that looks like in a histogram: • If the **[CLICK]** mean is greater than the **[CLICK]** median, the data is skewed towards the right. This is also called positive skew. You can remember this because this **[CLICK]** long tail of values goes up in the positive direction.
	Meanwhile here's a different distribution. Here, the **[CLICK]** mean is less than the **[CLICK]** median, indicating a negative skew. You can see how the mean was pulled more strongly towards the left, towards the **[CLICK]** tail that goes out negatively from the center of the data.
	The larger this difference is, the more skewed your data is, because the mean will continue to be pulled further and further away from the median. The magnitude of this difference should be put in context with your standard deviation. • For example, say you have a **[CLICK]** mean of 5, a **[CLICK]** median of 10, and a **[CLICK]** standard deviation of 10. The large difference between the mean and median compared with the standard deviation suggests **[CLICK]** significant skewness. **[brief pause for learner to read graph]** • Now take the **[CLICK]** same mean and median, but the standard deviation is **[CLICK]** 100. The **[CLICK]** difference is not as impactful, but skewness is still present. **[brief pause for reading]**
	What if your data isn't skewed? Here's a **[CLICK]** histogram of data that is not skewed. In this case, the mean and median will be **[CLICK]** roughly equal, since the mean hasn't been influenced significantly by asymmetry in the data.
	Quick test, say I have a **[CLICK]** mean of 18, a median of 10, and a standard deviation of 15. Would you say this data is **[CLICK]** skewed positive, **[CLICK]** skewed negative, or **[CLICK]** not skewed? **[CLICK]** Hint: consider the direction that the mean is being pulled. **[second pause for learner to think]**

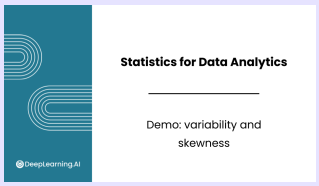
	This data is positively skewed! Since the mean is larger than the median, that difference suggests there is a long tail on the right or positive side of the distribution. The difference of 8 compared with the standard deviation of 15 suggests a considerable degree of positive skewness.
	When calculating skew in a spreadsheet, you'll use the **[CLICK]** SKEW() function, which returns a number, so you'll need to know how to interpret it. **[CLICK]** More positive values indicate more **[CLICK]** positive skew, more **[CLICK]** negative values indicate **[CLICK]** negative skew, and **[CLICK]** values close to zero indicate **[CLICK]** minimal skewness.
	Absolute values of skewness **[CLICK]** less than 0.5 indicate minimal skewness. Skewness between **[CLICK]** 0.5 and 1 is moderately skewed and skewness **[CLICK]** greater than 1 is highly skewed.
TH	Now you're familiar with how to calculate and understand the key measures of central tendency, variability, and skewness. But how can these be used in data analytics? Follow me to the next video to better understand how to select each measure based on the business question you're trying to answer.

L3V4 – Why use these measures?

Visual	Script
TH 	Why are these three measures – of central tendency, variability, and skewness – so important?
	Depending on your business question, you may care **more** about **different aspects** of your distribution. Sometimes you care most about the **[CLICK]** average person, sometimes about the **[CLICK]** extremes, sometimes the **[CLICK]** variability. Let's look at a few examples of business questions and how descriptive statistics can help you answer them.
	A personalized merch company might wonder, **[CLICK]** how much do people **typically** spend in one purchase? In this case, **[CLICK]** you're interested in central tendency. Basically, where's the middle? You have to be mindful of how you characterize the typical purchase amount. The average could be misleading if you have a skewed distribution. **[CLICK]** For example, if your mean purchase is $100, but you have lots of small purchases and a few very large ones, that **[CLICK]** mean of $100 doesn't represent your typical sale very

	well. In this case, you might want to **[CLICK]** use the median, and maybe investigate those large sales further to see if you can derive insights into your big spenders.
	Now say you're working for a **[CLICK]** cheese hat manufacturer and they want you to figure out, **[CLICK]** how **consistent** is productivity across factories? In this case, **[CLICK]** variability is quite valuable. Say your **[CLICK]** target is for **[CLICK]** each factory to produce 50 cheese hats per day, and the **[CLICK]** average productivity is 50 units per day. It looks good on the surface, but a mean value with no variance means something very different than a mean value with a lot of variance. If **[CLICK]** each factory produces almost exactly 50 cheese hats on average per day, you can be pretty **[CLICK]** confident in your production forecasts. But if a factory produces **[CLICK]** 10 units in one day and **[CLICK]** 75 on another, you'll come to a **[CLICK]** different conclusion. It's possible to reach up to 75 cheese hats per day, and you may want to implement the practices used on these high productivity days company-wide.
	Lastly, consider an electricity company that may investigate: **[CLICK]** how much electricity capacity do we need to handle **peak** usage times? For this question, you're not just interested in the average electricity usage, **[CLICK]** but also the extremes. You'd want to look at the highest usage times, perhaps the **[CLICK]** 95th or 99th percentile, to ensure you can handle peak loads without blackouts. This analysis is especially important for utilities planning for **[CLICK]** hot summer days or **[CLICK]** cold winter nights when usage spikes.
🎙️ TH	Cool! You've seen how to match up these different measures with the different insights they best support. Follow me to the next video to see conduct these analyses in a spreadsheet environment.

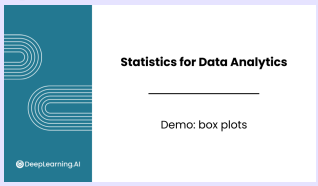
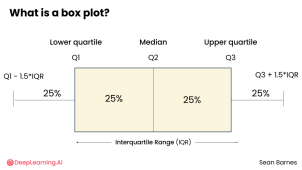
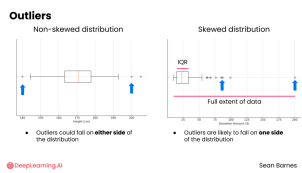
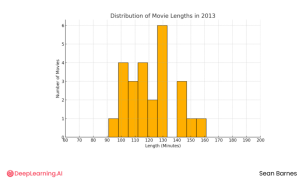
L3V5 – Demo: variability and skewness

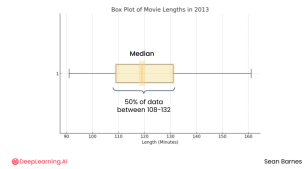
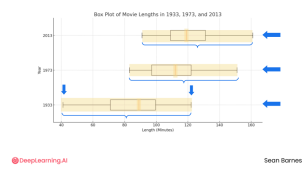
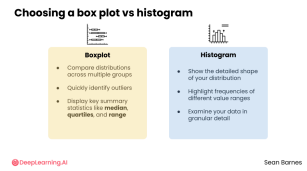
Visual	Script
🎙️ TH 	Now let's see how you can use a spreadsheet to calculate measures of variability and skewness in the Lending Tree data. These measures complement your previous exploration of the central tendency of the sample data.
🖥️ SC 🔗 Start 🔗 Solution	Let's take another look at the paid interest feature. You know that the mean value is around $617 and the median is around $456. But that doesn't tell you much about the variability in the data. So, while you may be interested in

	seeing the average payout, you're probably also interested in understanding the variability too – how common is it to make less than the average or more than average? Is the amount of interest paid spread out over hundreds of dollars or is it tightly clustered around the mean? Remember that if you'd like to follow along with the demo, you can find this spreadsheet and the solution in the downloads tab. ● You can start with the range, which tells you the difference between the maximum and minimum payments. This will help you understand the difference between the largest and smallest possible payouts ● Range ○ B9 → =MAX(Data!Q:Q) ○ B10 → =MIN(Data!Q:Q) ○ B11 → =B9-B10 ● That tells you the maximum possible distance between values but you have more tools available to you. Variance accounts for all the data points, not just these two extremes ● Variance ○ B13 → =VAR(Data!Q:Q) ○ Reduce decimal points to 2 ○ This calculation is useful, but it's not in dollars, it's in dollars squared which is weird. Let's convert to standard deviation to have it be more interpretable ● SD ○ B14 → =STDEV(Data!Q:Q) ○ Show that STDEV = SQRT(VAR) ■ B14 → =SQRT(B13) ○ Make standard deviation into dollars ● Skewness ○ Based on the histogram you saw earlier, what would you say is the skew of this feature? ○ You can also look at the mean, median and mode. Do you think that the distribution is symmetrical or skewed? **[pause for thought]** ■ The mean>median, which indicates Positive Skewness ○ B16 → =SKEW(Data!Q:Q) ○ Reduce to 2 decimal points ○ Skewness is positive sign as expected
🎙️ TH	Even just by calculating central tendency, variability, and skewness, you learn a

	lot about the potential interest payout. In the next video, you'll wrap up this discussion of distributions by interpreting another common visualization: the box plot.

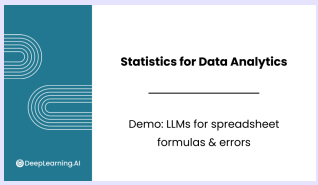
L3V6 – Box plots

Visual	Script
TH 	Another great visualization tool for distributions is the box plot, also sometimes called a box and whiskers plot by cat enthusiasts **[Sean puts on cat ears]** They're a bit less common than histograms, but quite useful. You should know how to interpret them.
	A box plot visualizes the quartiles of the data, **[CLICK]** including the median, which you saw previously is the second quartile. **[CLICK]** The first and **[CLICK]** third quartiles form the box, and the whiskers typically extend to the **[CLICK]** smallest and **[CLICK]** largest values that are within 1.5 * the IQR from the first and third quartiles. **[CLICK]** The box covers the length of the interquartile range.
	Box plots also help visualize outliers. Outliers are any values that lie beyond the whiskers, and are often represented as individual markers. In **[CLICK]** non-skewed distributions, **[CLICK]** outliers could fall on either side of the center of the distribution. For example, there are outliers on either side of the distribution for human height – **[CLICK]** people who are unusually short, and **[CLICK]** unusually tall. For **[CLICK]** skewed distributions, however, most of the **[CLICK]** outliers are likely to fall on one side of the distribution. This is because most of the values in skewed distributions are highly concentrated in one region of the distribution, which causes the **[CLICK]** interquartile range to be small relative to the **[CLICK]** full extent of the data. For example, in this distribution of charity donations, most donations are 10 to 20 dollars, but there are a **[CLICK]** few larger donations. So, most outliers are likely to be large donations, not really small ones.
	As a quick reminder, you saw this histogram of movie durations in the year 2013 in an earlier lesson. You guessed that the mean movie duration was somewhere between 120 and 130. Now that you're familiar with IQR, you can construct a box plot for this same data to compare the two visualizations.

 https://i.imgur.com/JF2b8uw.png	Here's a box plot of the same movie durations data. It's easier to tell that the **[CLICK]** median is a bit below 120 minutes and that **[CLICK]** 50% of the data is roughly between 108 and 132 minutes. Box plots are especially good for comparing multiple distributions, since you're directly comparing the medians, quartiles, and potential outliers, as well as the variability, whereas a histogram does not enable those comparisons as directly, even when you control for bin size and axis scale.
	Let's see how box plots enable better comparison for the movie duration data. **[CLICK]** Here are box plots for movie durations from **[CLICK]** 1933 on the bottom, **[CLICK]** 1973 in the middle, and **[CLICK]** 2013 on the top **[CLICK]** What can you tell about the distributions? **[pause for learner to think]** The median value for 2013 as you have seen is around **[CLICK]** 120, while for 1973 it's close to **[CLICK]** 115. That's not a massive difference, and the **[CLICK]** variability in those two years is similar, considering both the range and IQR. 1933 is a bit of a different story. Its median duration is around **[CLICK]** 90 minutes – what a time! And it seems movies regularly were as **[CLICK]** short as 40 minutes! The range only barely reaches **[CLICK]** 120, which was the median length in 2013. It seems between 1933 and 2013 there was a noticeable shift in movie durations, but that in the latter half of this period, the change had pretty much run its course.
	So when should you choose a **[CLICK]** box plot versus a **[CLICK]** histogram? Histograms are a great choice when you want to **[CLICK]** show the detailed shape of your distribution, **[CLICK]** highlight frequencies of different value ranges in your data, or **[CLICK]** examine your data in granular detail. Box plots on the other hand are ideal when you want to **[CLICK]** compare distributions across multiple groups, **[CLICK]** quickly identify outliers, or **[CLICK]** display key summary statistics like the median, quartiles, and range at a glance.
🎙️ TH	Histograms and box plots are both powerful tools for data visualization. Before you complete the upcoming ungraded lab, I hope you'll join me in the next video to see how to use LLMs to help with spreadsheet errors and formulas. I know it will be so helpful to you. I'll see you there!

L3V7 – Demo: LLMs for spreadsheet errors & formulas

Visual	Script
🎙️ TH	You've encountered a lot of complex formulas and spreadsheet tasks so far. A

	large language model can serve as a helpful thought partner for writing formulas as well as troubleshooting any errors you may encounter.
🔗Start 🔗Solution	All right, so you've seen this data set before. This is the data set of the top 25 movies in each year, their rating on IMDB, the number of ratings they have, and their duration. Let's say you're working with this data in a spreadsheet and you've set up a few calculations that you want to do. Let's say you want to see how many movies are over two hours. You know you want to use COUNTIF, so... count the number of movies if it's greater than 120 minutes. So, you select your E column, and you say greater than 120. But you get this error. Frustrating, because it just says formula parse error. It really doesn't tell you that much. Let's see if we can figure out the problem with an LLM. ***PROMPT:*** *Hey, I got a formula parse error, can you help me fix it? Here is my formula: =COUNTIF(E:E,>120).* *Here is the first few rows of my data:* I'm just going to demonstrate ChatGPT 4-oh here. I'm going to start a new conversation and I'm going to use this prompt right here: “Hey, I got a formula parse error, can you help me fix it? Here is my formula.” I'm just pasting the formula here. And add a few rows of the data. ***PROMPT:*** *Hey, I got a formula parse error, can you help me fix it? Here is my formula: =COUNTIF(E:E,>120).* *Here is the first few rows of my data:* *Film Year Rating Number_Ratings Duration_Minutes* *M 1931 8.5 71033 10* *City Lights 1931 8.6 65709 87* *Frankenstein 1931 8 37746 70*

	So I want to give the LLM as much context as possible. Maybe it's not relevant, but maybe it can help me figure out what the problem is. Okay, it looks like the issue with your formula is related to the syntax. In Google Sheets, you need to enclose a criterion in quotation marks. Okay, so, I forgot about my quotation marks. Now, let's say I'm not 100 percent sure that this is correct. Remember, you should have a mindset of skepticism. So let's just ask, ***PROMPT:*** *Are you sure the greater than sign needs to be within the quotes?* And just double check. It says yes, I'm sure. just by asking again, we're just giving it a chance to rethink. So, let's go ahead and copy this function right here, go back to our sheet, and paste it in. And there we have 638 movies that are greater than two hours. Okay. Now, say I want to see the percent of movies that are greater than two hours. If I want to count the total number of movies here, I can scroll down to the bottom and see 2,051 (minus 1 for the header row). That's kind of tedious. Maybe the number of movies will change if I add more years. So in this case, I'm just going to ask, ***PROMPT:*** *How can I count the number of rows in my dataset with a formula? Give me some options.* Okay, so it gives me a couple of options here. **[READ AND COMMENT ON SUGGESTIONS - FIND AN INCORRECT ONE AND DO IT]** countA for the A column, rows for the A column. Here's another one using countIf, but this is only counting rows where the value in column A is greater than 120, so not the most useful suggestion, that isn't what I asked. And counting rows with multiple criteria, okay, but I don't care about these criteria. So let's go ahead and borrow the first option that it gave us. And I'll put under total movies, COUNTA for this column. Well, 2052, okay, that's clearly not correct. It looks like this is counting the header row. Let's go back to the model and just say,

PROMPT:
I tried the first option, but it isn't working for me. It counts the header row as well.

Okay, so now it suggests **[READ AND COMMENT ON SUGGESTIONS]**

This is interesting because I happen to know that the count function is also a good option. So you can just do equals COUNT, this counts the number of numeric values in a data set, and you can just select any numeric column: year, rating, anything like that, and that will give you 2051, which we already checked manually is correct. But these other options work as well. So we could do COUNTA for column A minus one, that works as well. To get percent, we can just divide these two results. And that'll give us 31.11%. Okay, interesting. So about a third of movies are greater than two hours.

Let's say that you want to see – where do these 2 hour movies fall in the distribution of all movies? So first we'll calculate the median of duration. It's about 109 minutes.

So it seems like the 2 hour movies will be above the median. Possibly above the 75th percentile. Let's check. So we'll do percentile, and we'll select the duration column, and I want to know the 75th percentile. But I get an error. Okay. “Error. Function percentile parameter 2 value 75 is out of range”

Okay, it tells me that it's out of range, but it doesn't tell me what the range is. I can head to ChatGPT and ask

PROMPT:
What is causing a parameter 2 value 75 is out of range error. =percentile(E:E, 75)?

what is causing the range error. And just give it the function here. So the error occurs because the percentile function expects a percentile to be between 0 and 1, and it gives us the corrected formula. Okay, now I remember. Okay, that's helpful. Alright, so 120 is just below the 75th percentile. And we can go ahead and fill in the same for 25th percentile if we're curious.

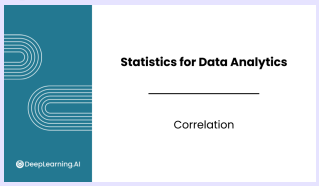
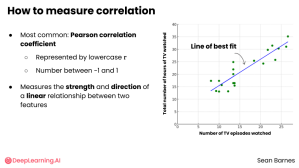
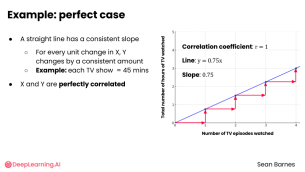
Okay, one cool trick I want to show you is using LLMs for conditional formatting. Conditional formatting is actually really, really flexible, So, when you go to Format and Conditional Formatting, under Format Rules, one of the options is “Custom Formula Is...”.

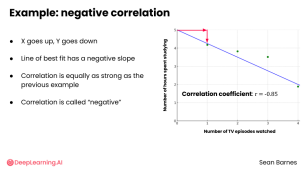
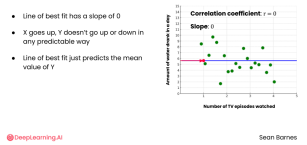
	And so that gives you a lot of flexibility, but it is kind of a pain to remember how to format custom formulas. Let's say I want to highlight each row from column A to column E if and only if the rating of that movie is in the 75th percentile or higher. So let's just go ahead and calculate the 75th percentile. Okay, so 75th percentile rating is 7.9. So let's go over to chat GPT Here is my prompt: ***PROMPT:*** *How can I use conditional formatting to highlight the entire row from column A to column E green, if the rating is in the 75th percentile or higher. My 75th percentile is in cell H6.* So, this is important because now I'm giving the LLM the ability to reference that cell in the formula rather than having it write a formula and then fix that formula later myself. So let's see what it says. Okay, so calculate 75th percentile, check. Select the range of cells, for example, select A2, okay. Open conditional formatting. Set the rule using custom formula is. Okay, Select the range. Make sure not to select the header row, because that can cause some issues. And then change to “custom formula is...”. And paste in that value that the LLM just gave you. Okay, so right away I see that it's applied. Let's double check. It should only be the rows with a rating greater than or equal to 7.9 that are highlighted. So this 7.9 is highlighted. 8.0 is highlighted here. It looks like it's working really well. So this conditional formatting helps me see at a glance which of these movies are in the 75th percentile or higher based on their duration. So, it's pretty cool. LLMs can help you discover stuff that you didn't even know existed. Great work!
🎙️ TH	LLMs can help you discover spreadsheet functionality and enjoy working with them more. You’ve learned a ton so far in this module. Coming up, you’ll take the practice

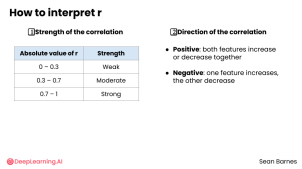
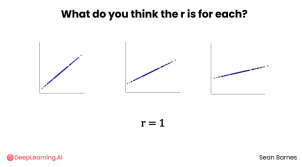
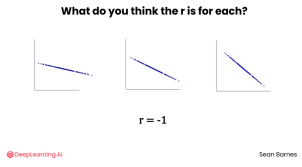
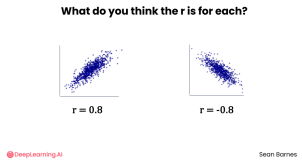
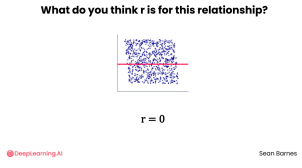
	assessment for this lesson, as well as complete two practice labs. In the first, you'll practice calculating and interpreting variability and skew using the dataset of Spotify songs from the previous lesson. In the second, you'll work with a large language model to explore new spreadsheet functionality and troubleshoot errors. Once you're done, join me in the next lesson to explore correlation.

Lesson 4 – Correlation

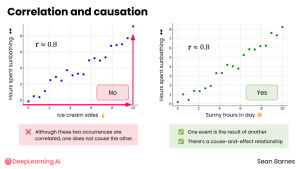
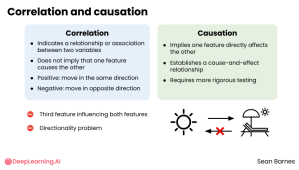
L4V1 – Correlation

Visual	Script
TH 	Correlation is a way to quantify the relationship between two numerical features. You've seen scatter plots that visually represent how two features are related. Now you'll formalize that relationship.
	There are multiple ways of measuring correlation, but **[CLICK]** the most common is the Pearson correlation coefficient, which is **[CLICK]** represented by lowercase r. r is always a **[CLICK]** number between -1 and 1. Let's start with the intuition behind this measure. Say you have a **[CLICK]** scatterplot of two numerical features. On the x axis you have **[CLICK]** number of tv episodes watched, and on the y axis the **[CLICK]** total number of hours of tv watched. You can think of r as measuring how well you can fit a **[CLICK]** straight line through these points, called a **[CLICK]** line of best fit. More technically, r **[CLICK]** measures the strength and direction of a linear relationship between two features.
	Let's take the perfect case. Say you have these points for your scatter plot. **[CLICK]** **[CLICK]** A straight line has a consistent slope, meaning **[CLICK]** for every one-unit change in X, Y changes by a consistent amount. That would make sense for a tv show where **[CLICK]** every episode is the exact same length, say 45 minutes. If you know how many episodes of tv someone watched, you also know exactly how long they've been watching tv for. In this example, you have a line at **[CLICK]** y = 0.75x, meaning that each episode is three quarters of an hour. So this line has a **[CLICK]** slope of

	positive 0.75, and **[CLICK]** for every change of 1 in X, **[CLICK]** Y changes by 0.75. **[next 6 clicks before the next line of script] [CLICK] [CLICK] [CLICK] [CLICK] [CLICK] [CLICK]** In this case, the correlation coefficient **[CLICK]** r equals positive 1. You can say that **[CLICK]** X and Y are perfectly correlated because changes in X produce exactly predictable changes in Y.
	Now imagine your data is less perfect, with this set of points. It's not as easy to fit a line through the middle, but it is **[CLICK]** definitely possible to get a good approximation. **[CLICK]** Each point is either a little above or a little below the line, but as **[CLICK]** X goes up, **[CLICK]** Y also goes up, and it's relatively predictable how much it will go up. You can tell that the more tv episodes you watch, the more time you've watched tv, but the **[CLICK]** episodes aren't all the same length. The correlation coefficient might be close to positive **CLICK]** 0.85. In this scatter plot, the line of best fit is the **[CLICK]** same as the previous one, but the correlation isn't as strong.
	Here's another example where as **[CLICK]** X goes up, Y goes down. In this case, X is the **[CLICK]** number of tv episodes watched, and y is the **[CLICK]** number of hours spent studying. So the more tv a person watches, the less time they might have to study. In this case, **[CLICK]** the line of best fit has a negative slope. However, **[CLICK]** the correlation is equally as strong as the previous example. The correlation coefficient would be around negative 0.85. Remember, you should think of correlation as essentially about how well you can fit a straight line through the data, regardless of whether the relationship is positive or negative. So, when **[CLICK]** X goes up, **[CLICK]** Y goes down, and it goes down a roughly predictable amount. This correlation is as strong as the previous example, **[CLICK]** it's just called "negative", like the slope of the line of best fit.
	Here's what a scatter plot would look like with no correlation, which means the **[CLICK]** correlation coefficient is equal to zero. This graph represents the **[CLICK]** number of tv episodes watched and the **[CLICK]** amount of water each person drank in a day. You can drink water whether you watch tv or not, so there doesn't appear to be any relationship. **[CLICK]** The line of best fit has a slope of zero, because **[CLICK]** as X goes up, **[CLICK]** Y doesn't go up or down in any predictable way. So, **[CLICK]** the line of best fit just predicts the mean value of Y no matter the X value – that's the best possible guess, and it's not very good.

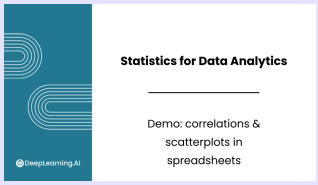
	Calculating r by hand can get complicated, but computers nowadays make it much easier. Let's examine how to interpret r, which tells you two important things about the relationship between two features: • First, **[CLICK]** the strength of the correlation. Here's a general guide for interpreting r based on its **[CLICK]** absolute value, or distance from zero. ○ **[CLICK]** Positive or negative values between 0 and 0.3 indicate a weak correlation. ○ **[CLICK]** Between 0.3 and 0.7, positive or negative suggests a moderate correlation. ○ **[CLICK]** Between 0.7 and 1 positive or negative indicates a strong correlation. • Second, r tells you **[CLICK]** the direction of the correlation. A **[CLICK]** positive r means both features tend to increase or decrease together. A **[CLICK]** negative r means that as one feature increases, the other tends to decrease.
	Here are three scatterplots. What do you think the r is for each? Here's a hint: see how they all look like a straight line? **[pause for learner's answer]** **[CLICK]** These each have an r of 1. You can perfectly fit a line through this data, and as one feature goes up, so does the other. The slope of the line is positive, no matter how big or small, and the relationship is highly predictable.
	What about these three scatter plots, what r do you think they have? **[pause for learner's answer]** **[CLICK]** These each have an r of -1. You can fit a line perfectly through them, however as X goes up, Y actually goes down, making the correlation negative.
	Here are two scatter plots with correlations of 0.8 and -0.8. Can you tell which is which? **[pause for learner's answer]** **[CLICK]** The one on the left is positive 0.8, **[CLICK]** while the one on the right is -0.8.
	Okay last one, what's your guess for the correlation of this scatter plot? **[pause for learner's answer]** **[CLICK]** It's 0! There's no discernable relationship here, and the **[CLICK]** straight line here would just be flat.
🎙️ TH	Great work studying correlations! It's exciting to calculate a number that maps onto your intuition about the data. Join me in the next video to learn more about the difference between correlation and causation.

L4V2 – Correlation and causation

Visual	Script
TH	Correlation is easy to misinterpret because it's often confused with a related concept – causation.
	Correlation and causation might sound similar, but they refer to two different aspects of the relationship between two features. Take a look at **[click]** this plot of **[click]** ice cream sales on the x axis and the **[click]** number of hours spent sunbathing on the y axis. You can see that as ice cream sales go **[click]** up, **[click]** so do hours spent sunbathing. The r value might be around **[click]** positive 0.8. Based on that information, can you conclude that buying ice cream causes people to sunbathe more? **[pause for thought] [click]** No. **[click]** Although these two occurrences are correlated, one does not cause the other. Have a look at **[click]** this very similar plot of number of **[click]** sunny hours in the day and number of **[click]** hours spent sunbathing. You can see that the plot is quite similar to the one you saw a moment ago. The r value might again be around **[click]** positive 0.8. But is there causation? **[pause for thought] [click]** Yes there is, because the availability of sunlight directly affects people's ability to sunbathe. You would be less likely to sunbathe on a cloudy day. Causation means that **[click]** one event is the result of another; **[click]** there's a cause-and-effect relationship.
	To sum up the difference, **[click]** correlation indicates **[click]** a relationship or association between two features, but **[click]** does not imply that one feature causes the other. It **[click]** is positive when the features move in the same direction or **[click]** negative when the features move in opposite directions. **[click]** Causation, on the other hand, implies that **[click]** one feature directly affects the other and it **[click]** establishes a cause-and-effect relationship. You can't establish causation using a scatterplot or the Pearson correlation coefficient r. **[click]** It requires more rigorous testing than correlation. To establish whether there is a causal relationship between the features, you should note that • **[click]** There might be a third feature influencing both features of interest, giving the illusion that they directly affect each other. In the

	case of the ice cream and sunbathing example, while these two correlate, they are both caused by sunnier days, a third feature. • Another potential pitfall is **[click]** the directionality problem. Even if a causal relationship exists, it might be challenging to determine which feature is the cause and which is the effect without experimental evidence. It's clear that **[click]** sunnier days **[click]** cause more **[click]** sunbathing and **[click]** not vice versa. But how about if you're interested in the relationship between loneliness and social media usage? Do lonely people use more social media, or does social media make people feel more lonely?
🎙️ TH	While correlation can be established using scatterplots and the Pearson correlation coefficient, causation requires a higher level of rigor. Be mindful of how you interpret correlation. Follow me to the next video to see how to run correlation analysis in a spreadsheet!

L4V3 – Demo: correlation in spreadsheets

Visual	Script
🎙️ TH 	Let's put correlation into practice on the lending tree Loan Dataset. You can choose any pair of numerical features to see how they correlate!
🔗Start 🔗Solution 🔗Dataset	As a reminder, this dataset includes thousands of loans made through the lending tree platform, which allows individuals to lend to other individuals. • Each row represents a particular loan, and each column includes features about the person who took out the loan, like their job title **[A]** and annual income **[F]**, as well as information about the loan like the amount **[K]**. And don't forget, if you'd like to follow along with the demo, you can find this spreadsheet and the solution in the downloads tab. So far, you've taken a look at the distribution of the paid_interest feature, and examined its central tendency, variability, and skewness. This is an interesting feature from a lender's point of view, but it's even more useful with more

context. You're also interested in which factors contribute to higher interest and loan amounts.

Let's consider the correlation between **paid_interest** and **installment**. Remember the installment is the monthly payment the applicant should be making.

- How do you expect these features to be correlated? Positive or negative? Strong, moderate or weak? **[pause for thought]**
- Create the scatter plot.
    - Select the columns → Insert Chart
    - Chart Setup → Chart type Scatter
    - Move to new sheet
    - Add x label
        - Customize → Chart and axis title → Horizontal axis title → installment
    - Add y label
        - Customize → Chart and axis title → Vertical axis title → paid_interest
    - Add the trendline: Customize → Series → Select trendline at the bottom. Make it thicker and change the color so it stands out more.
    - What kind of correlation do you see in the scatterplot? **[pause for thought]**
        - Trendline has positive slope
        - For small values of paid_interest, you get small values of installment, and you have a lot of these points
        - For larger values of paid_interest you get bigger values of installment, but with a lot more dispersion, but less density of points.
        - It suggests positive correlation, probably between moderate and strong.
- Now find the actual correlation between features
    - Pick a cell → =CORREL(→ Select paid_interest → CTRL → installment →ENTER
        - Mention you can swap the columns and it is the same, because correlation is "symmetric".
        - You get a correlation of 0.69, which is right on the edge between moderate and strong correlation. Remember moderate is between 0.3 and 0.7, and strong is above 0.7

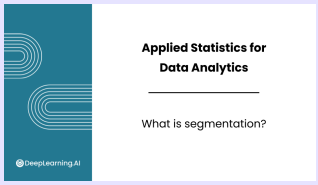
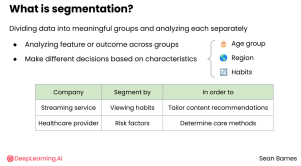
	Let's see some other examples! Let's take a look at how the **paid_interest** correlates with the **annual_income**. ● How do you expect these features to be correlated? Positive or negative? Strong, moderate or weak? **[pause for thought]** ● Create scatter plot with trendline (same as before) ○ Select the columns → create plot → Scatter plot ○ Add axis labels ○ Add trendline (fix color, opacity and width) ● What kind of correlation do you see in the scatterplot? **[pause for thought]** ● Interpret: ○ There is a higher density of points with low paid_interest and low annual_income, but the points are pretty dispersed ○ You also get some points pretty far away, that will affect correlation ○ The points very loosely fit the trendline, meaning that there is likely not a strong linear relation between the features. ○ This suggests a weak positive correlation ● Get the actual correlation ● This gives you a correlation of 0.206 which is below 0.3, so it's weak but positive. The more people make in a year, the more interest they pay, but there are a lot of other factors that explain the variation in interest paid. It's difficult to predict the exact value of interest paid based on income. Let's look now at two new features: **annual_income** and **debt_to_income.** Debt to income is the ratio between how much money you owe on loans and how much money you make in a year. A lower debt to income ratio means a person has proportionally less debt and is more likely to be able to pay off a new loan. This correlation helps answers the question what's the relationship between how much money these borrowers make & the relative amount of debt they have? ● How do you expect these features to be correlated? Positive or negative? Strong, moderate or weak? **[pause for thought]** ● Create scatter plot ○ Select the columns → create plot → Scatter plot ○ Add axis labels ○ Add trendline (fix color, opacity and width) ● What kind of correlation do you see in the scatterplot? **[pause for thought]**

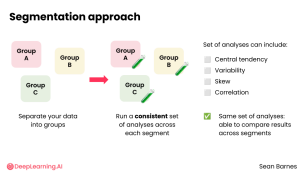
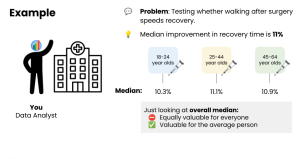
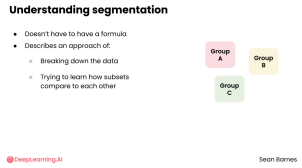
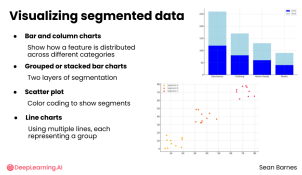
	• Interpret:
	  ○ Trendline has negative slope
	  ○ You have a lot of points where you get high values of annual_income paired with low debt_to_income, and vice versa, this suggests negative correlation
	  ○ The scatter plot shows a pretty non linear relationship between features. This will impact correlation value
	• Find the actual correlation
	  ○ -0.177
	  ○ This is a weak negative correlation. As income goes up, people taking out loans tend to have a lower debt to income ratio.

TH	Great work using scatter plots spot the sign and magnitude of a correlation, as well as interpreting the output of the CORREL function. That takes you to the end of this lesson. Next up, you'll complete the practice assessment as well as the practice lab for this lesson. In the practice lab, you'll explore the correlations between different features of music in order to build an even better playlist.

Lesson 5 – Segmentation

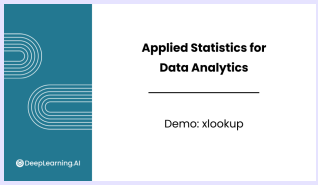
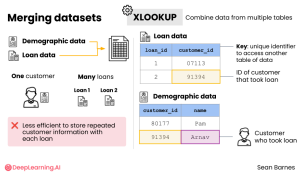
L5V1 – What is segmentation?

Visual	Script
TH 	Segmentation is a powerful technique that allows you to develop insights for subsets of your data. You saw several examples of segmentation in Data Analytics Foundations, but now let's formalize this concept.
	At its core, segmentation is about **[CLICK]** dividing your data into meaningful groups and analyzing each group separately. Generally, it's useful when you're interested in **[CLICK]** analyzing a particular feature or outcome across different groups. It's also valuable when you might want to **[CLICK]** make different decisions based on group characteristics. Common segmentations include **[CLICK]** age group, **[CLICK]** geographical region, or **[CLICK]** habits. For instance, **[CLICK]** a streaming service might segment its users by **[CLICK]** viewing habits to **[CLICK]** tailor content recommendations. Or **[CLICK]** a

	healthcare provider might segment patients by **[CLICK]** risk factors to **[CLICK]** determine appropriate interventions.
	The approach is straightforward: **[CLICK]** you separate your data into these groups and **[CLICK]** then run a consistent set of analyses across each segment. **[CLICK]** These analyses often **[CLICK]** include descriptive statistics, like measures of **[CLICK]** central tendency, **[CLICK]** variability, or **[CLICK]** skew, and **[CLICK]** correlation analyses. Using the same set of analyses is important because you want to be **[CLICK]** able to compare results across your segments.
	Let's look more closely at the healthcare example. Say you're **[CLICK]** testing whether walking after surgery speeds recovery. You know the **[CLICK]** median improvement in recovery time is 11%. But you might want to compare the effect across age groups. Is walking after surgery more effective for the young, the old? So, **[CLICK]** you segment your data by age group and **[CLICK]** measure the median recovery time for each group. Let's say you find that for 18-24 year olds, the median improvement in recovery time was **[CLICK]** 10.3%, for 25-44 year olds, it was **[CLICK]** 11.1%, and for 45-64 year olds, it was **[CLICK]** 10.9%. So walking after surgery is beneficial no matter the age! If you just looked at the overall median, you **[CLICK]** wouldn't be able to say for certain whether it was equally valuable for everyone, just that **[CLICK]** it's valuable for the average person. Segmentation is crucial in the medical field. Drugs and other interventions often affect men and women differently and require different dosages for children.
	There's no magical "segmentation method." **[CLICK]** It doesn't have a formula like the Pearson Correlation Coefficient does. **[CLICK]** It describes an approach of **[CLICK]** breaking down the data and **[CLICK]** trying to learn about how different subsets compare to each other.
	For visualizing segmented data, you have several options: • **[CLICK]** Bar and column charts are **[CLICK]** great for showing how a single feature is distributed across different categories. • You can also further segment your data using **[CLICK]** grouped or stacked bar charts. So, **[CLICK]** two layers of segmentation! • **[CLICK]** Scatter plots can be **[CLICK]** enhanced with color coding to show different segments within the data. You can also show several smaller scatter plots in a grid, one for each segment. • **[CLICK]** Line charts can be segmented by **[CLICK]** using multiple lines,

	each representing a different group. As always, when creating these visualizations, don't forget the principles of good design that you've already learned. Clarity, efficiency, and context are still crucial when displaying segmented data.
TH	Great work with segmentation. Join me in the next video to see a powerful spreadsheet technique for working with data from multiple files, which is common in segmentation tasks.

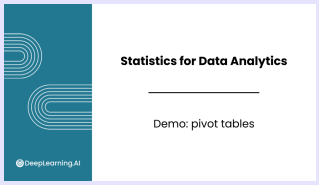
L5V2 – Demo: xlookup

Visual	Script
TH 	Often, customer data and product data are stored separately. They may be collected and analyzed in different ways. However, if you want to segment product usage patterns based on customer features, you'll need to merge these datasets. It's a common task in data analytics.
	In the case of the Lending tree Loan dataset you've seen throughout this module, the **[CLICK]** customer demographic data, like their income, and the **[CLICK]** loan data, like the interest paid, were likely originally stored in separate files. **[CLICK]** One customer may have **[CLICK]** many loans, so it's **[CLICK]** less efficient to store this repeated customer information together with each loan. Here's what typically happens. The loans dataset has a feature called **[CLICK]** customer_id, and its value for each loan is the **[CLICK]** id of the customer that took it out. So, if you take that id and **[CLICK]** search the dataset of customers, you'll come back with **[CLICK]** one unique result: the **[CLICK]** customer who took out the loan. The main advantage of this approach is that you can store the customer and loan data separately, while still having access to both. The feature customer id is called a **[CLICK]** key, a unique identifier that allows you to access another table of data. Now, in order to perform segmentation analysis, you'll need to **[CLICK]** merge these two datasets. That way you can answer questions like how does income correlate with interest paid? Spreadsheets have a powerful function called **[CLICK]** XLOOKUP that allows you to **[CLICK]** combine data from multiple tables, in this case to create detailed profiles of customers with the most profitable loans. Let's see how it works.

🔗Start
🔗Solution
	• Check out the data – demographics in one sheet, loan info in the other
	  ○ Use this sheet. A **customer_id** column was added
	  ○ I randomly sorted the customer card info so it wasn't just like, copy over the rows
	• Which one helps us match?
	• Point out **customer_id** is in both,
	• Do it the manual way first
	  ○ Copy first **customer_id**, ctrl f, look up in second sheet
	  ○ **Grade** is A
	• We need to do this 3,000 times, let's do it programmatically
	• Start with one column, grade
	  ○ Add the column name in demographic sheet
	  ○ Use =XLOOKUP(), look at help menu
	  ■ Search key = loan id
	  ■ lookup_range= where you're looking (loan_id in second sheet)
	  ■ Result_range = which column you want returned (grade in the second sheet)
	• Can manually drag handle down, let's use arrayformula though
	  ○ =ARRAYFORMULA(XLOOKUP(A2:A,applicants_credit_history!A:A, applicants_credit_history!M:M))
	• Let's do one more column
	  ○ =ARRAYFORMULA(XLOOKUP($A2:$A,applicants_credit_history!$A:$A, applicants_credit_history!J:J))
	  ○ You can see that the result_range changed

🎙️ TH	Great work! Follow me to the next and final video of this module, where you'll explore how to apply descriptive statistics to segmented data using pivot tables.

L5V3 – Demo: pivot tables

Visual	Script
🎙️ TH 	In addition to charts, you'll typically want to examine the descriptive statistics for your segmented data. Spreadsheets power users employ pivot tables for this task. Let's see how they work.
🔗Start here 🔗Solution	Different kinds of loans have different risk levels for the lender, which reflects the likelihood that the borrower will not be able to pay the loan. and the interest rate reflects those risk levels. Higher risk loans have higher interest rates, so these loans can be more profitable as long as the borrower is able to

repay. You're interested in investigating the differences in interest rates to understand your potential loans better.

Start segmenting by **grade**. Remember that grade is the classification system that involves assigning a quality score to a loan, where grade A represents the lowest risk and G is the riskiest.

Remember to check the downloads tab for this spreadsheet and the solution.

- Select the entire dataset
- Go to **Insert** → **Pivot table**.
- A pop up menu appears, select insert to New sheet. Alternatively you could use an existing sheet for this, and select the cell where you want the pivot table to be in.
- On the right you will get the **Pivot table editor**, here is where you set up the pivot table.
    - Rows: Add the **grade** feature (this is the one you use for segmentation
    - **Values**: Add the feature you want to aggregate
        - Click on **Add** under **Values**.
        - Select **interest_rate** .
        - Now choose the summarization method from **SUM** to other options like **AVERAGE**. This should show you the mean interest rate for each card category
    - Observations: As expected, grade A loans pay the smallest interest rate, and it gets consistently higher for riskier grades
- If you get out of the pivot table and need to edit, don't worry, just click on the pencil button on the lower left corner, and you get all the options again.
- **Columns**: If you want to pivot over two features, this is the place to add it, but let's skip this for now.
- Add more statistics
    - Now add another statistic, like standard deviation. Values → Add → Select **interest_rate** → change the summarization method to STDEV
    - Observations: what happened with the standard deviation of grade G? **[pause for thought]** It probably has just one observation, so the variability is 0!
- You can also add more features, it's the same step.

	○ Values → Add → Select **total_credit_utilized** → change the summarization method to AVERAGE ○ You can keep adding as many features and statistics as you like! Another cool thing is that you can create categories within a feature to do the segmentation. Let's see how ● Select the data table ● Insert → Pivot table ● Select Insert to Existing Sheet → Choose a cell below the previous pivot table (Say A9) ● For Rows select **emp_length** ● For Values select **paid_interest** and summarize by **AVERAGE**. ● Observe that there seems to be some difference between the **emp_length = 1,2,3** segments and the rest. Let's segment with this criteria ○ Right Click anywhere in the **emp_length** column in the pivot table ○ Select Create pivot group Rule ○ Minimum Value: 4(so it groups 1, 2 and 3) ○ Interval Value: 7 (so you cover all values above 3 up to 10, which is the maximum value) ● You can add MODE, VAR, STDEV or whichever statistics you like!
🎙️ TH	Excellent work with segmentation. You learned how to assemble two datasets into one, visualize the segments in your data, and calculate descriptive statistics for each one using pivot tables. Coming up, you'll take the practice assessment for this lesson. You'll also complete the two graded items for this module, including the graded assessment and the graded lab. In the lab, you'll use all the skills you learned in this module to help the Portuguese National Park Service prevent forest fires. After you've completed these items, you'll move on to the next module: Probability and Simulation. Once you've taken a sample of the population and described the distribution of that sample, you can apply rules of probability and statistics to estimate features of the entire population. I'll see you in the next module to learn more.