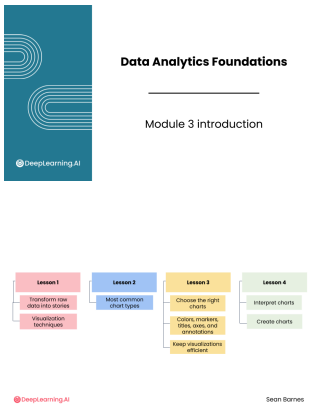
DAG C1M3 scripts

Video title	Isabel	Sean	Slides
L0V1 – Module 3 introduction	✅	✅	✅
L1V1 – What is data storytelling?	✅	✅	✅
L1V2 – The language of data visualizations	✅	✅	✅
L1V3 – Analyzing visualizations	✅	✅	✅
L2V1 – The right chart for the right insight	✅	✅	✅
L2V2 – Demo: Bar & column charts	✅	✅	✅
L2V3 – Demo: Customizing charts	✅	✅	✅
L2V4 – Demo: Scatter plots	✅	✅	✅
L2V5 – Demo: Grouped bar and column charts	✅	✅	✅
L2V6 – Demo: Stacked bar and column charts	✅	✅	✅
L2V7 – Demo: Line charts	✅	✅	✅
L3V1 – Strategies for effective data visualization	✅	✅	✅
L3V2 – Data encoding	✅	✅	✅
L3V3 – Chart elements	✅	✅	✅
L3V4 – Data visualization: the good & the better	✅	✅	✅
L4V1 – Interpreting data visualizations with LLMs	✅	✅	✅
L4V2 – Creating data visualizations with LLMs	✅	✅	✅

Introduction

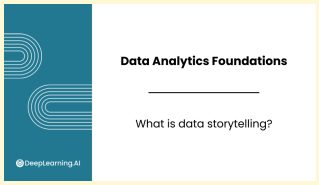
L0V1 – Module 3 introduction

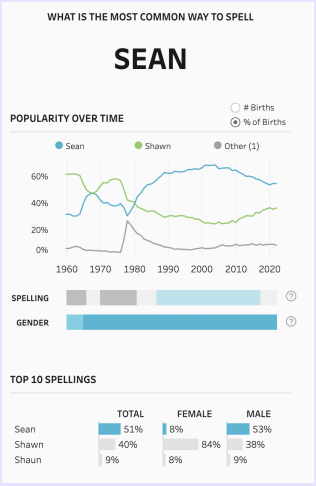
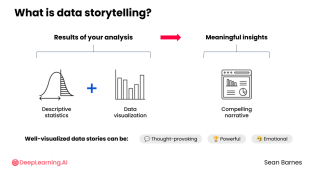
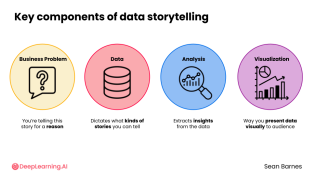
Visual	Script
🎙️ TH	Welcome to Module 3 of Data Analytics Foundations – everything you need to know about data visualization! So far, you've learned what data is, as well as how to process and analyze it using a spreadsheet. Now, you'll get hands on with data visualization, the practice of communicating data with graphics.

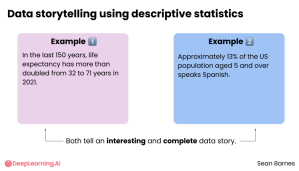
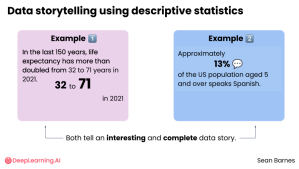
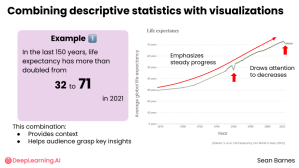
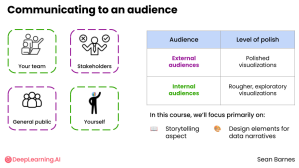
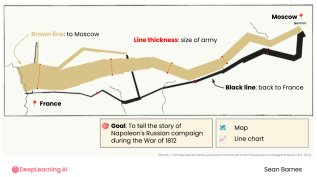
	You'll see how to **[CLICK]** transform raw data into captivating stories that resonate with your audience. Then, you'll explore **[CLICK]** visualization techniques to communicate insights effectively, including the **[CLICK]** most common chart types you'll use as a data analyst. Charts can be misused and misinterpreted. You'll learn how to **[CLICK]** choose the right chart for your data, plus **[CLICK]** how to use colors, markers, titles, axes, and annotations to highlight key insights. Plus, you'll see how to **[CLICK]** keep your visualizations efficient by maximizing the data-ink ratio and minimizing chartjunk... ™ You'll also practice using LLMs to both **[CLICK]** interpret and **[CLICK]** create charts. You'll engage with these tools using a skeptical mindset to save time and avoid mistakes. By the end of this module, you'll be equipped with the core visualization skills you need to drive real-world impact. Let's get started!

Lesson 1 – Data storytelling

L1V1 – What is data storytelling?

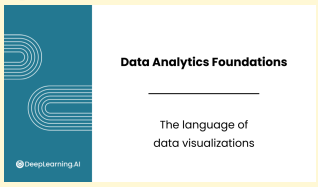
Visual	Script
🎙️ TH – overlay of dry list of numbers 	You're looking at two different presentations of the same data. One **[SEAN HOLDS UP RIGHT HAND]** is a dry list of numbers and statistics,
🎙️ TH – overlay of beautiful data vis e.g. this one	while the other **[SEAN HOLDS UP LEFT HAND]** is a colorful visual narrative that instantly draws you in. Data storytelling is the difference between dry **[RIGHT HAND]** and engaging **[LEFT HAND]**. In this video, we'll explore what data storytelling is, its key components, and why it's such a crucial skill for anyone working with data... starting with an example!
🔗**SCREENCAST** Sean and Shawn 1960 – 2024; USE LASER POINTER FEATURE IN GOOGLE SLIDES? Historical Women in STEM - VOTD	• Let me tell you a story about Sean and Shawn. No, not two Seans. Two spellings of the same name! This story uses the same baby name data you saw in the previous module, and it was created by Kelly Gilbert. • Here's a graph of Sean and Shawn over time, with time on the x axis starting from 1960 to 2024, and the percent of babies born with each name. In green is Shawn with a w, and in blue is the other Sean, the one that looks like bean. • In the late 1960s, Blue Sean began to take off in popularity! It matched Green Shawn for a few years before falling out of favor again, only to

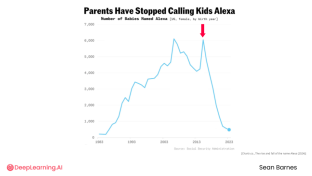
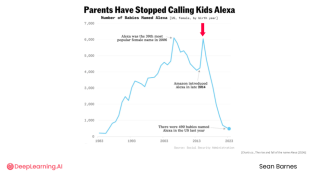
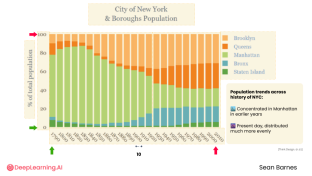
https://public.tableau.com/app/profile/oclaoruid/viz/GenZLeadsLGBTUp/LGBTIdentification https://public.tableau.com/app/profile/takafumi.shukuya/viz/MoM2024W15UKUnemploymentRatesoverDecades/MoM2024_W15 	finally overtake green Sean in the year 1980! Since then, blue Sean has been the dominant spelling, though the gap has been narrowing lately. • If you want to give your child a more unique spelling, go with green Shawn, while blue Sean may be the more popular one. Fun story, right?
	Data storytelling is all about translating the **[CLICK]** results of your analysis into **[CLICK]** meaningful insights. It's the art of **[CLICK]** combining **[CLICK]** descriptive statistics and **[CLICK]** data visualization to **[CLICK]** convey a compelling narrative. As the saying goes, a picture is worth a thousand words. It's just part of being human. **[CLICK]** Well-visualized data stories can be **[CLICK]** thought-provoking, **[CLICK]** powerful, even **[CLICK]** emotional.
	So, what are the key components of data storytelling? • First is the **[CLICK]** **business problem.** Always remember you're **[CLICK]** telling this story for a reason. Focus on your audience and your goal. • Next, there's the **[CLICK]** **data** itself. Because data is your raw material, it **[CLICK]** dictates what kinds of stories you can tell. The name data set you saw a moment ago can compare spellings over time, but not where they were born. • Then, there's the **[CLICK]** **analysis** – the process of **[CLICK]** extracting insights from the data. The descriptive statistics you calculated in the previous module like averages and percentages are fantastic tools, and more complex analyses are certainly valuable too.

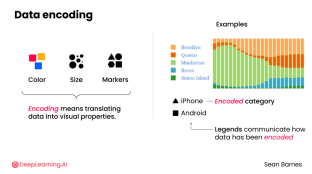
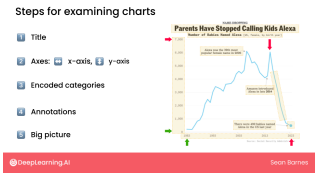
	• Finally, there's the **[CLICK] visualization** – the way you **[CLICK]** present the data visually to your audience.
 • Source • source	You can tell a data story without a visualization, using just descriptive statistics. Here are two examples: • **[CLICK]** In the last 150 years, life expectancy has more than doubled from 32 to 71 years in 2021, reflecting huge advances in areas like nutrition and healthcare worldwide. • **[CLICK]** Approximately 13% of the US population aged 5 and over speaks Spanish, reflecting the deep cultural roots and growing influence of Hispanic and Latinx communities across the nation. **[CLICK]** Both of these descriptive statistics tell an interesting and complete data story.
	You can even use a technique like conditional formatting in a spreadsheet to help visually explain your data. And maybe I can zhuzh these numbers up a little. **[CLICK]** Make 71 larger than 30. **[CLICK]** Add a little speech bubble next to 13%. Already I'm starting to emphasize the key points in a more visual way.
	However, combining these descriptive statistics with a well-crafted visualization can take your data story to the next level. For example, you can combine this story with a **[CLICK]** line graph. Here's one of year on the x axis and average global life expectancy on the y axis. It **[CLICK]** emphasizes steady progress and **[CLICK]** draws attention to decreases. **[CLICK]** This combination **[CLICK]** provides context and **[CLICK]** helps your audience grasp the key insights at a glance.
	Data storytelling is usually about communicating to an audience, whether that's **[CLICK]** your team, **[CLICK]** your stakeholders, or the **[CLICK]** general public. But you can also create them for **[CLICK]** yourself – to quickly spot trends in time series data, or get a rough sense of which revenue streams are the biggest right now. **[CLICK]** More polished visualizations are often needed **[CLICK]** for external audiences, while **[CLICK]** rougher, more exploratory visualizations may suffice for **[CLICK]** internal analytical purposes. **[CLICK]** In this course, we'll focus primarily on **[CLICK]** the storytelling aspect and **[CLICK]** the design elements that make for a compelling data narrative.
 **Must keep this image**	I mentioned emotional data stories earlier. Let's see an example. One of my favorite visualizations is **[CLICK]** Charles Joseph Minard's famous March to Moscow graphic. The text is in French and cursive (!), **[CLICK]** so just focus on the graphics while I walk you through step by step. Minard's **[CLICK]** goal was to tell the story of Napoleon's Russian campaign during the War of 1812. What would take a writer thousands of words, he explained with a single image. ~~Take a look.~~

book/text on one side, **graphic** on the other or replacing it/overlaying it P. 37 of Tufte's book describes it more – https://faculty.salisbury.edu/~jtanderson/teaching/cosc311/fa21/files/tufte.pdf	This visualization is both a **[CLICK]** map and a line chart showing time series data. **[CLICK]** The thickness of the line represents the size of the army as it starts in **[CLICK]** France on the left, marches to **[CLICK]** Moscow on the right, and comes back. **[CLICK]** Brown is the army going towards Moscow. **[CLICK]** Black is the army coming back. ~~**[CLICK]** You can see at the bottom of the graphic, the temperatures on the march back.~~ What do you think, was it a successful campaign? **[pause for thought]** You don't need to read French, or know that Napoleon lost almost 410,000 men to know the campaign was disastrous. As one historian said of this graphic, it "seems to defy the pen of the historian by its brutal elegance."
🎙️ TH	Now you've seen the power of a well-told, well-visualized data story, join me in the next video to learn the language of data visualizations, and how to break them down into their component parts. I'll see you there.

L1V2 – The language of data visualizations

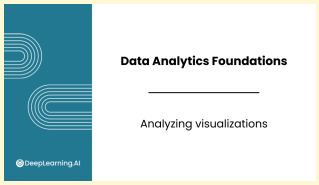
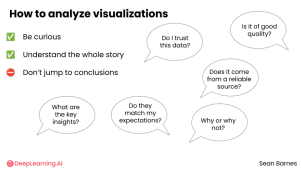
Visual	Script
🎙️ TH 	Have you ever looked at a graph and felt a bit lost? You're not alone! The subreddit for ugly data has over 100,000 members. Not all visualizations are created equal, and some are easier to interpret than others. Let's break down the common components of a visualization so you can interpret them skillfully.
Source 	Let's start with another baby name example. Let me orient you to this chart. • First off, **[CLICK]** the title. "Parents have stopped calling their kids Alexa." That's pretty clear, and it tells me what I should expect to see. • Now, **[CLICK]** the x axis. It doesn't have a title, but since it **[CLICK]** starts at 1983 and **[CLICK]** ends at 2023, I get a clear sense that it's time, and shows the past 40 years with the earliest date on the left, the most recent on the right. • Now **[CLICK]** the Y axis. Again, no title, but it **[CLICK]** starts at 0 and **[CLICK]** ends at 7,000, with **[CLICK]** even increments of 1,000. **[CLICK]** When I look at the chart subtitle, though, I see "number of babies named Alexa, US, female." Okay so per year, this is how many female babies were named Alexa. Sometimes this approach of describing the y-axis label in the subtitle is used to preserve some space on the left-hand side. • Now, I'll look for colors, markers on the line, and a legend. **[CLICK]** The line is blue, but since it's the only one, it doesn't seem like it means anything specific. There's also no legend. **[CLICK]** I see this marker at

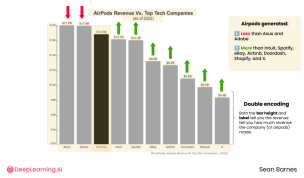
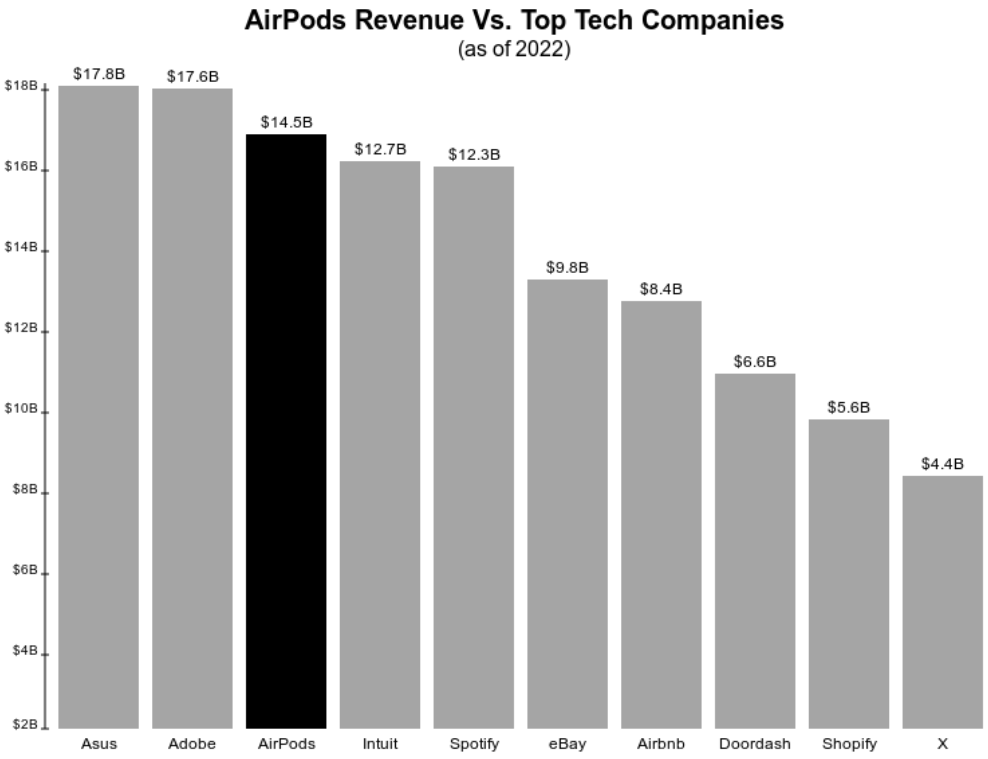
	2023 that indicates the end of the graph. It brings some focus to where the trend sits at the time this chart was made, which helps emphasize the headline. • Now I'll look at annotations. **[CLICK]** I see a peak in the line here indicating Alexa being a very popular name, at almost 6,000 female babies per year. **[CLICK]** Here's another annotation for Amazon introducing Alexa. **[CLICK]** And then all the way in the bottom right, 490 babies were named Alexa in 2023. • Now the insight. It looks like Alexa was pretty popular overall until about 2016. **[CLICK]** Then there's a sharp drop off.
	• Without looking at the title and annotations, that would lead me to wonder, what caused the drop off?
	• But the annotations make it pretty clear: just two years after introducing the Alexa virtual assistant, parents stopped naming their kids Alexa. What do you think was the reason? Maybe people thought they might accidentally trigger their Alexa by calling their child. The chart leaves you to infer the reason, and that mystery is pretty interesting to think about.
Source  Here's another option – https://www.statista.com/chart/17862/apples-annual-revenue-by-operating-segment/ this one has annotations	Now, take a moment to read this chart. What is it trying to communicate? **[PAUSE FOR SEVERAL SECONDS]** You might feel your eyes jumping from one area to the next, and the most important parts might not be standing out to you. Let me walk you through interpreting this graph to understand its main insight. • **[SEAN AD LIBS A BIT]** I notice **[CLICK]** the title says "City of New York & Boroughs Population". So this is a chart about population, and it shows both the populations of New York City, and the five boroughs, which are like districts. • **[CLICK]** Now I'll look at the x axis. It doesn't have a title telling me what this is, but I can tell these are years. So **[CLICK]** starting from the year 1790 all the way on the left, **[CLICK]** then going to the year 2010 on the right. And a quick check tells me **[CLICK]** these are in consistent increments of 10 years, which means it's an even comparison between all these bars going across the x axis • **[CLICK]** Now I'll look at the y axis. **[CLICK]** This one does have an axis label, "% of total population. **[CLICK]** It starts from 0 at the bottom and **[CLICK]** goes to 100 at the top. So that tells me each bar represents 100% of the population of New York City, but broken down by borough. • **[CLICK]** Let's talk about the borough thing. I know you've been thinking about these colors! Color is powerful. • In this case, the color of each column segment represents **[CLICK]** the

	% of the total population that lives in each of the five boroughs. In the earlier time periods, from about **[CLICK]** 1790 through 1920, Manhattan was by far the most populous borough. **[CLICK]** But around 1920-1930, Brooklyn took over and has remained at the top. The boroughs are sorted by population, both in the stack of column segments and in the legend, based on where the order landed in 2010. This consistent ordering allows you to trace the trends within each category over time. • So what's the overall story here? **[CLICK]** For me, this chart summarizes the population trends across the five boroughs throughout the history of the city. **[CLICK]** Again, much of the population was concentrated in the city-center in Manhattan in the earlier years, but we've observed a broadening of where people live over time. **[CLICK]** In the present day, the population is distributed much more evenly across 4 of the boroughs, with Staten Island being a smaller proportion relative to the other boroughs.
	**[CLICK]** Color, **[CLICK]** size, and **[CLICK]** markers are all examples of **[CLICK]** *encoding*, which means translating data into visual properties. **[CLICK]** So light green means Manhattan while blue means the Bronx. **[CLICK]** Or a triangle marker represents iPhone sales while a square marker represents Android sales. Triangle doesn't inherently mean iPhone, **[CLICK]** it's an encoded category – one way the chart can convey meaning visually. **[CLICK]** And we use legends to communicate how the data has been encoded on a particular visualization.
	When you see any chart, take a structured approach to identifying what it is trying to tell you. Here are the steps you just took with me when examining the charts on the previous slides. You should do these for every chart. 1. **[CLICK]** Check the title and subtitle. What is this chart about? Is there a key insight the creator is trying to communicate? 2. **[CLICK]** Next, review the axes. Almost every chart has at least one axis. An axis can have tick marks or grid lines, which mark major steps along that axis, or an axis label which describes the feature plotted on that dimension. a. **[CLICK]** Check your x axis first. What is happening from **[CLICK]** left to **[CLICK]** right? In the case of these two examples, it was time represented in years. Typically the x axis increases from left to right, but don't take that for granted! b. **[CLICK]** Then check your y axis. What is changing from **[CLICK]** bottom to **[CLICK]** top? Like the x axis, it typically increases. 3. **[CLICK]** Next, review any encoded categories. Read the legend to identify the encoded categories, and look for color, marker, or size differences. Make sure you understand what each encoded category means. 4. **[CLICK]** Then look for annotations. Annotations are notes or labels added in the chart to provide context or highlight key points. These

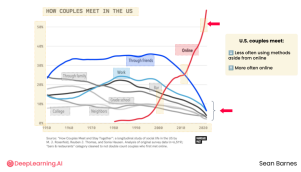
	help draw your attention to what matters most in the chart. 5. **[CLICK]** Once you've reviewed all of these individual components, assess the big picture. What type of insight are you looking for? Should you be making a comparison? Should you be looking for a trend over time? Look for surprising information, big changes, gradual changes. Use annotations and the chart title or subtitles to guide your thinking. Following this principled approach will mean you get the most out of every visualization, even when a single image is conveying an overwhelming amount of information.
TH	In data visualization, the whole is often greater than the sum of the parts – each component that you've just seen serves a specific purpose, and together, they can tell a powerful data story. As you encounter data visualizations, whether in the news, at work, or in this course, try interpreting them using the 5 step process you just learned. It's a great way to practice your data visualization literacy. In the next video, you'll get some more practice extracting insights from data visualizations. See you there!

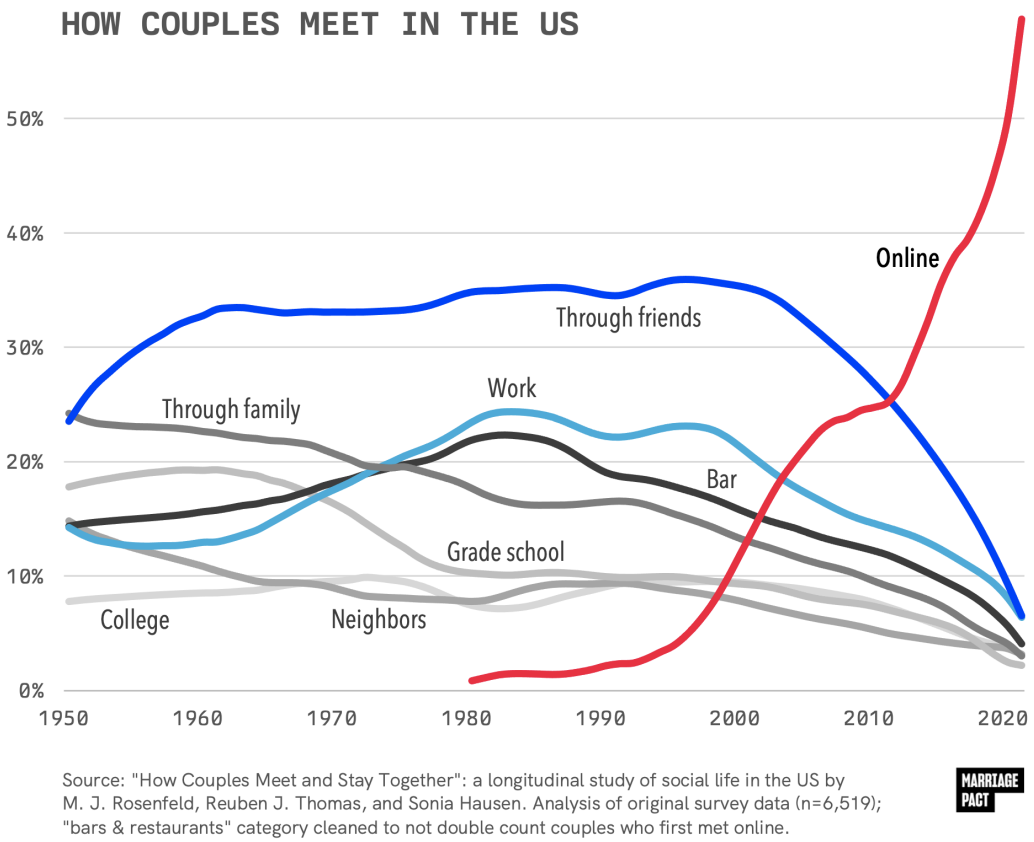
L1V3 – Analyzing visualizations

Visual	Script
TH 	Now that you've seen the core components of data visualizations, you'll put that knowledge into practice by analyzing three visualizations with me.
	Big picture, your main orientation should **[CLICK]** be curiosity – seek to **[CLICK]** understand the whole story and **[CLICK]** don't jump to conclusions. You can think of yourself as interrogating each visualization with questions like: • **[CLICK]** Do I trust this data? **[CLICK]** Is it of good quality and **[CLICK]** does it come from a reliable source? • **[CLICK]** What are the key insights? **[CLICK]** Do they match or not match my expectations, and **[CLICK]** why or why not? Let's get some practice. I have three charts for you to interrogate: a bar chart, a line chart, and a scatterplot. For each chart, I encourage you to pause the video and practice your 5 step process for interpretation. Then, I'll walk you through each chart.

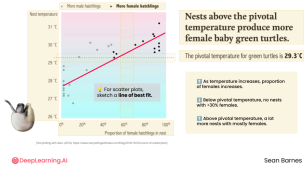
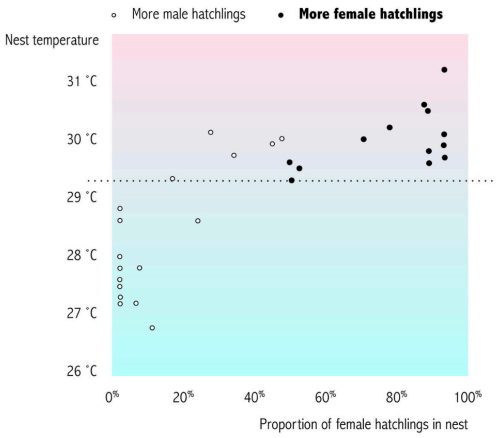
Must keep this image Fixed image – https://i.imgur.com/Ora6W5J.png Original image: https://www.reddit.com/media?url=https%3A%2F%2Fi.redd.it%2Fyfd30i4rbvjb1.png clean it up - Do a bar chart vs a column chart (no angled labels) - Emphasize airpods bar - Gray out or de-emphasize the other ones - Remove the airpods - Optionally embed labels in bar Fixed image 	 • Let's start with this bar chart. Pause for a moment and give it a shot. **[pause for thought]** • First the **[CLICK]** title: AirPods Revenue vs. Top Tech Companies. And the **[CLICK]** subtitle: as of 2022. I expect to see the revenue generated by airpods alone vs some top tech companies. • On the **[CLICK]** x axis, what do you see? These are some top tech companies like Asus, Adobe, Intuit, Spotify, and so on. • What about the **[CLICK]** y axis? There's no axis title, but based on the chart title and axis labels you can infer that this is revenue as of 2022 specifically, and in billions of dollars. I also notice each bar has a **[CLICK]** label with the revenue, which makes it easier to compare revenue between them. This strategy is called **[CLICK]** double encoding – both the bar height and this label tell you how much revenue the company (or airpods) made. • Look for encoded categories. What do you see? This chart uses color to **[CLICK]** highlight airpods compared with the other companies. • So, big picture, what insight is this chart communicating? It's expecting you to make a comparison between the revenue generated by AirPods and the revenue from these huge companies. **[CLICK]** AirPods generated **[CLICK]** slightly less revenue than Asus and Adobe as a whole, **[CLICK]** but more than Intuit, Spotify, and all the other companies in this chart. That's a pretty surprising insight! Because it's surprising, you may want to check either the data or your assumptions.
Must keep this image https://www.reddit.co	• Here's another interesting one, this time a line chart. Pause for a moment and see if you can figure out the story. **[pause for thought]**

m/media?url=https%3A%2F%2Fi.redd.it%2Fuwe4cchpfz5c1.png





- Start with the **[CLICK]** title: “How couples meet in the US.” What does that tell you about the data represented in this chart? It’s only people in the US. It doesn’t say anything more about how a “couple” is defined, so you may need to learn more in order to understand what the data is telling you. For instance, I happen to know that this data only represents heterosexual couples.
- What do you see on the x **[CLICK]** axis? It’s unlabeled, but I see 1950 to 2020, so you can safely assume that this represents the year of the data. It’s also a line chart, so that typically means the data will be measured over time.
- What about the **[CLICK]** y axis? The labels go from 0% to 50%, and the y axis is unlabeled. ~~It would be helpful to have a label, but~~ based on the title you can assume this means “percent of couples.” Whether that’s all couples in the US, just new couples, couples of a certain kind isn’t clear.
- Now check the encoded categories. I don’t see any markers or a legend, but each line has a different color. **[CLICK]** Online is red and it really stands out, **[CLICK]** friends and **[CLICK]** work are different shades of blue, and **[CLICK]** the other lines are different shades of gray. It’s not clear whether the blue categories are more relevant than the gray ones, but they are the next most popular ~~after the online category.~~
- Next, you should check for annotations, but I don’t see any here.
- So what’s the big picture? **[CLICK]** This chart is encouraging you to compare the change in couples meeting online with couples meeting in

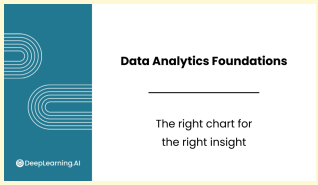
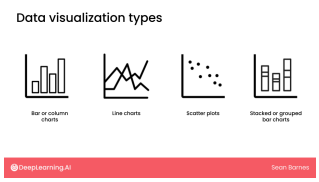
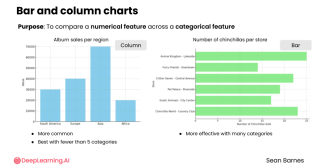
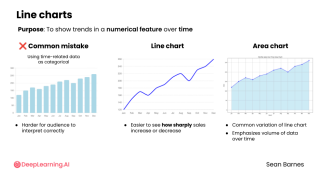
	all the other ways displayed here. **[CLICK]** Each of the categories aside from online has a sharp drop off while **[CLICK]** meeting online shows a sharp increase. **[CLICK]** Starting around 2000, social media began to proliferate, starting with the likes of MySpace, Friendster, and Facebook. Remember MySpace and Friendster? **[chuckle]** Meeting online surpassed the previous top method of meeting through friends around **[CLICK]** 2012 or so – the year Tinder was released – and **[CLICK]** in 2020 over half of couples met online. Smart phones also became increasingly more prevalent through the last decade or so of this time period, and likely also boosted the online category. • You can find some other insights in this chart as well, like how around 10% of US couples met their partner in college consistently from 1950 to about 2000, or the consistent decline of couples meeting through family, grade school, or neighbors. But the key insight here is about meeting online compared with in-person.
Must keep this image https://www.storytellingwithdata.com/blog/2018/10/23/scores-of-scatterplots about halfway down – the temperature should be on the x axis in my opinion  	  • Last one, here's an interesting scatter plot about … science! This one is a little more challenging! Pause the video for a moment and see if you can figure out the story. • First, look at the **[CLICK]** title: Nests above the pivotal temperature produce more female baby green turtles. Ugh, turtles are so cute. And the **[CLICK]** subtitle: The pivotal temperature for green turtles is 29.3 degrees Celsius. ~~After reading these, I'm expecting a chart showing the relationship between temperature and the number of female baby green turtles. Temperature is a continuous numerical feature, and the number of baby green turtles is a discrete count, so~~ the scatter plot makes sense because we are trying to visualize the relationship between two numerical features. • On the **[CLICK]** x axis, what do you see? It's a percent from 0 to 100 showing the percent of female hatchlings found in the nest. Since each hatching is likely to produce different numbers of babies, it makes sense to focus more on the percentage than that absolute count. • What's on the **[CLICK]** y axis? It's labeled nest temperature and it's in

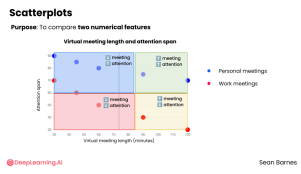
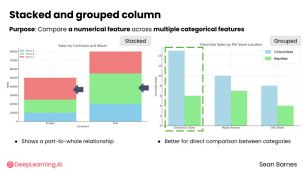
	degrees celsius. If you're not familiar with celsius, it might be helpful to convert the temperature to degrees Fahrenheit: 26 degrees Celsius is around 78 degrees Fahrenheit, and 31 degrees Celsius is around 87 degrees Fahrenheit. • Now the encoded categories. What do you notice? At first the color may not jump out at you, but it's a nice **[CLICK]** natural encoding with colder temperatures as blues and higher temperatures as pinks. You should also notice the **[CLICK]** different markers for nests with more male hatchlings vs more female hatchlings, based on the legend at the top: an empty circle for hatchings with more males, and a filled circle for hatchings with more females. ~~So at 50% on the x-axis, we see a clear separation of the markers. This is another subtle example of double encoding.~~ • Do you see any annotations? Annotations don't have to be text! This **[CLICK]** dotted line shows the "pivotal temperature" for green turtles, which is around 29.3 degrees C. You may not know what that means, but the annotation here indicates you should look for changes around that line. • So, big picture, what insights can you get from this chart? Since it's a scatterplot, one technique you can use is to **[CLICK]** sketch what a line through the middle of all these points could look like (this is called the line of best fit). **[CLICK]** As the temperature increases, so going up the y axis, the proportion of majority female hatchings increases. The effect is pretty dramatic. • Now examine the pivotal temperature line. **[CLICK]** Below the pivotal temperature, there are no nests with greater than 30% female hatchlings. **[CLICK]** Above it, there are a lot more nests with mostly females. • This gets me curious about the scientific mechanism behind this finding!
🎙️ TH	Great work analyzing those charts! In this lesson, you've seen the power of data storytelling and the role that data visualization plays in crafting a compelling data story. You've also practiced a structured process for analyzing data visualizations. Once you're done with the practice assessment for this lesson, follow me to the next one where you'll learn how to create beautiful visualizations in Google Sheets. See you there!

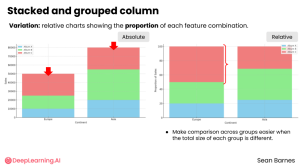
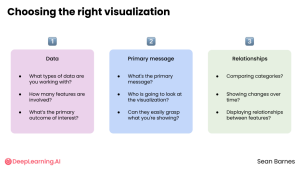
Lesson 2 – Creating charts

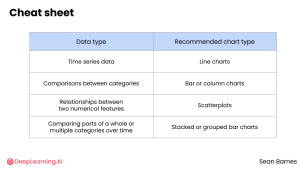
L2V1 – The right chart for the right insight

Visual	**Script**

TH 	The other day I saw my friend dip his french fries in mayonnaise. Right in front of me! You're probably wrinkling your nose right now and thinking, wow, that is just SO WRONG. It should've been aioli. I mean, really. By the end of this lesson you'll see some data visualizations the same way as I see my mayonnaise-dipping friend: that is just SO WRONG. Because data visualization is both an art and a science. As a science, there are right and wrong answers. Specifically, it's crucial to choose the right type of visualization for the insight you're trying to communicate. Let's see how to do that.
	There are hundreds of data visualization types, but we'll focus on the core four: **[CLICK]** bar or column charts, **[CLICK]** line charts, **[CLICK]** scatter plots, and **[CLICK]** stacked or grouped bar charts. These four types cover a majority of the insights you'll want to communicate. If I had to throw out a number, I'd say you can effectively communicate close to 80% of insights with just these four types. You don't have to get fancy!
	First, bar and column charts. • Their **purpose** is to compare a numerical feature across a categorical one. • For example, they're **great for** visualizing data like **[column chart]** **[CLICK]** album sales per region – here you can see continent on the x axis and sales on the y axis – or **[CLICK]** the number of chinchillas per store – in this case, the **[bar chart with longer store names]** number of chinchillas sold on the x axis at each of these six exotic pet store locations on the y axis. ~~Hey look, one of them is grandma's!~~ • ~~I've been saying "bar AND column" because there's a genuine difference. It IS helpful to be specific with your terminology. The sales per region chart is a **[CLICK]** column chart, while chinchillas per store is a **[CLICK]** bar chart.~~ **[CLICK]** Column charts are more common and **[CLICK]** they're best with fewer than around five categories. Bar charts are **[CLICK]** more effective with many categories. You can see how annoying it would be to fit all the text with the pet store names in such a small space at the bottom of this column chart. That's the main difference. I remember the difference by thinking: a column chart has columns aaaand the other one is a bar chart.
	Next, take a look at line charts. • Their **[CLICK]** **purpose** is to show trends in a numerical feature over time, ~~such as people bungee jumping each hour, or minute-to-minute stock price.~~ • **[CLICK]** Here's a line chart of monthly sales over time, with month on the x axis and album sales on the y axis. ~~Say it's sales of Beyonce's Lemonade. Great album.~~ The slope of the line between each point

	emphasizes the rate of change month to month, so you can see **[CLICK]** HOW sharply sales increased or decreased between each period. Line charts can also show multiple series, or many data points without becoming as overwhelming. ● ~~Here's a **[CLICK]** common mistake. I sometimes see **[CLICK]** time-related data treated as categorical, so using a column chart instead of a line chart. **[CLICK]** Here's the same monthly album sales over time, this time in a column chart. It's harder to calculate rates of increase and decrease, plus it breaks our 5 or fewer categories rule for column charts. There's nothing earth shatteringly wrong here, but as your time series data gets more nuanced, this type of chart will become **[CLICK]** harder for your audience to interpret correctly.~~ ● You may also have seen **[CLICK]** an area chart, **[CLICK]** which is a common **variation** on the line chart. It **[CLICK]** emphasizes the volume of data, not just the trend, especially an accumulation over time. ~~So, total downloads of Beyonce's Lemonade album.~~ The rate of sales goes up and down, but the total number only ever gets bigger – you can't un-sell albums.
	And now we come to my personal favorite: Scatterplots. They're just underrated, okay? ● Their **[CLICK]** **purpose** is to compare two numerical features ● They're **great for** exploring relationships between those features. ~~So insights like "age goes up, so wisdom goes up" or "the longer a virtual meeting, the lower my attention span" – maybe you feel the same.~~ ● Scatter plots have a lot of flexibility. **[CLICK]** Here's an example with virtual meeting length on the x axis and my attention span on the y axis, with **[CLICK]** personal meetings in blue and **[CLICK]** work meetings in red. ~~So yeah, attention span suffers more at work.~~ Another common enhancement is to add quadrants, emphasizing high/low combinations. If I **[CLICK]** add quadrants to this plot, I see long meetings with **[CLICK]** a high attention span, short meetings with a short attention span, and **[CLICK]** so on. It helps highlight that most meetings are short, with a high attention span.
	Finally, let's look at stacked and grouped column charts. You can also have bar charts with this variation, which are very similar. ● These are a variation on the standard column chart, and their **[CLICK]** **purpose** is to compare a numerical feature across multiple categorical features. ● So what are they great for? **[CLICK]** A **stacked chart**, which looks like a stack of books, shows a part-to-whole relationship. In this case, it

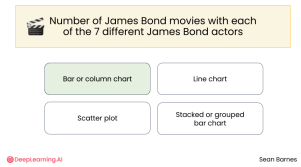
	shows continent on the x axis and sales on the y axis, with different albums shown in the three different colors. This chart answers the question: how much does each album contribute to the total sales in each region? I can see that in Asia, most sales come from **[CLICK]** Album B, while in Europe, most sales come from **[CLICK]** Album C. • A **[CLICK]** grouped chart is better for direct comparison between the categories. Here's a grouped column chart with pet store location on the x axis and chinchilla sales on the y axis. Which pet store sells the most chinchillas? It looks like **[CLICK]** Grandma's! • ~~So these are both absolute charts – right now, you're seeing just album sales and chinchilla sales.~~
	• A common **variation** is to have relative charts that show the proportion of each feature combination, rather than the raw number. **[CLICK]** Here's what that looks like for the album sales data. The x axis is the same, but now the y axis shows the proportion of sales rather than the number of sales. Relative charts **[CLICK]** make comparison across groups easier, in particular when the total size of each group is different. In this case, **[CLICK]** sales are much smaller in Europe than Asia, but you can highlight how **[CLICK]** Album C is selling well there using this type of chart. • This is a choice you as the data analyst will need to make – sometimes the magnitude is most important – you want to highlight that Asia has many more sales than Europe, but sometimes you want to highlight that Europe buys more of Album C proportionally.
	~~– here are some great questions to ask:~~ ~~First off, you need to **[CLICK]** understand your data.~~ • ~~**[CLICK]** What types of data are you working with – categorical, numerical, time series?~~ • ~~**[CLICK]** How many features are involved?~~ • ~~**[CLICK]** And what's the primary outcome of interest?~~ ~~Then, consider: what's the **[CLICK]** primary message you want to get across? And **[CLICK]** who's going to be looking at it? You want to make sure they can **[CLICK]** easily grasp what you're showing them.~~ ~~Next, how do your data points **[CLICK]** relate to each other?~~ • ~~**[CLICK]** Are you comparing categories?~~ • ~~**[CLICK]** Showing changes over time?~~ • ~~**[CLICK]** Displaying the relationship between features?~~



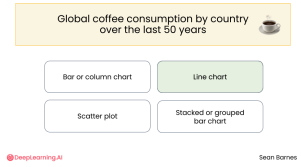
When choosing the right visualization – remember, this is a science, there are right and wrong answers

- Here's a cheat sheet:
  - **[CLICK]** Time series data often suits **[CLICK]** line charts
  - **[CLICK]** Comparisons between categories might **[CLICK]** use bar or column charts
  - **[CLICK]** Relationships between two numerical features could use **[CLICK]** scatter plots
  - **[CLICK]** For comparing parts of a whole or multiple categories over time, consider **[CLICK]** stacked or grouped bar charts

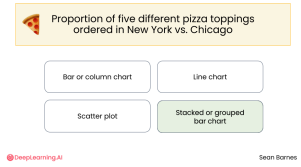
Let's do a lightning round. I'll give you an insight and you take a moment to think of the right chart for it.



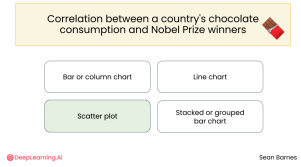
- Number of James Bond movies with each of the 7 different James Bond actors **[3 second pause for learner to consider]**
  - **[CLICK]** The answer is bar or column, and I would have a slight preference for bar here so that you could easily fit each actor's name as a label on the axis



- Global coffee consumption by country over the last 50 years **[3 second pause for learner to consider]**
  - **[CLICK]** The answer is a line chart, because we are comparing consumption over time, and we can incorporate each country with a distinctly colored line



- Proportion of five different pizza toppings ordered in New York vs. Chicago **[3 second pause for learner to consider]**
  - **[CLICK]** The answer is a stacked bar chart, because we want to analyze the relative proportion of each topping across the two locations



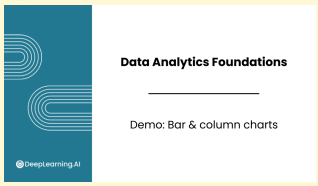
- Correlation between a country's chocolate consumption and Nobel Prize winners **[3 second pause for learner to consider]**
  - **[CLICK]** The answer is a scatter plot because we want to compare two numerical features together.

🎙️ TH

Great work! So now you'll start to have french-fries-in-mayonnaise moments with data visualizations. And hopefully a lot of "that's so right" when you see

	your own! In the remaining videos in this lesson, you'll see how to create each of these foundational chart types in Google Sheets. I'll show you all the tips and tricks I know. See you there!

L2V2 – Demo: Bar & column charts

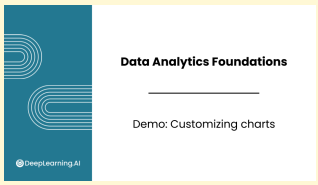
Visual	Script
🎙️TH 	Bar and column charts are one of the most common data visualizations. They're versatile, easy to understand, and can be used to compare different categories or groups. Let's explore how to create these charts in Google Sheets. We are going to work with a dataset from Redfin that summarizes housing sales for different counties in California. Suppose we wanted to analyze how the median home sale price varies across different regions in the state of California, in order to identify the most expensive housing markets.

🔗**DEMO SPREADSHEET**	• 💬Each observation represents the home sales in a given time period in a given county. For example Row 2 shows home sales in Stanislaus county from February 27 2023 to May 21 2024, which is 12 weeks.
	  ○ ▶️ Highlight Row 2
	• 💬Each feature in this data set is some characteristic of home sales in that period. Some of these are sales-y and might be hard to understand.
	  ○ ▶️Highlight D
	  ○ ▶️Highlight J
	  ○ ▶️Highlight L
	  ○ ▶️Highlight N
	  ○ ▶️Highlight W
	• 💬let's do a bit of preprocessing to better understand this data
	  ○ ▶️Freeze 1 row
	• 💬Okay, let's reduce this data, you can't have a column chart with 60 columns. Remember also the original goal was to identify counties with the most expensive homes
	  ○ ▶️Filter beginning period → Q3
	• Next step
	  ○ ▶️D:D → conditional formatting color scale
	  ○ ▶️W:W → conditional formatting diverging color scale at 0, lower is better
	  ○ ▶️Format L and N as plain numbers SELECT BOTH WITH CTRL
	• 💬And a little reconnaissance
	  ○ ▶️County → count summary
	  ○ ▶️J:J → show max, min, mean using corner summary

	○ ▶️N:N →show max, min, mean using corner summary
	• • 💬These are home prices prior to the new school year. Now you can sort so the most expensive markets are at the top ○ ▶️N:N → Sort Z to A
	• 💬Let's add our first chart ○ ▶️Highlight top C1:11 and N1:11 → insert chart (cmd click) • 💬Column chart by default • 💬Here's the chart editor. You have a lot of options! If you click off your chart, just double click on it again. • 💬This is a bit unreadable, let's flip to a bar chart ○ ▶️Change chart type → bar • 💬Using this kebab menu, you can download the chart, which is quite useful if you need to add it to a report, or move it to its own sheet. ○ ▶️Move to own sheet ○ ▶️Rename "Median home sales" • 💬Let's analyze this chart. ○ ❓What do you see in this data? ○ Patterns, trends, comparisons • 💬Here's another cool visualization you can make – median days on market. Let's modify the chart to show how a different numerical feature behaves across these same top 10 counties in terms of median home sale price. ○ ▶️Duplicate chart to another tab ○ ▶️Replace series with column V – median days on market • 💬These are still the top 10 most expensive counties for the selected time period, starting with San Mateo, even though we are now looking at median days on the market ○ ❓What do you see? ○ Explain more ...
🎙️TH	OK, there you have it, you've created your first data visualization in Google Sheets! Let's take it over to the next video to learn how you can customize your chart to tell a stronger data story.

L2V3 – Demo: Customizing charts

Visual	**Script**
🎙️TH	Alright, you've created a chart in Google Sheets, but it's looking a little... plain. Don't worry, it's time to unleash your creativity! You can bring your data visualizations to life with customizations.

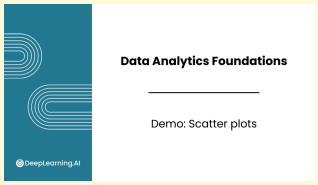


🔗[DEMO SPREADSHEET]

- 💬Let's customize the median days on market bar chart from the previous video.
  - ▶️Open chart editor → customize
- 💬Here, you'll find a wide range of options for customizing your chart.
  - ▶️Change background color to gray, change back
  - ▶️Change to 3D → change back
  - ▶️Title text “Median home sales in CA Q3 2023”
  - ▶️Horizontal axis → “Median home sales” → Bold → 16pt
  - ▶️Vertical axis title → “Region” → remove it
- 💬Let’s take a look at the series tab
  - ▶️Fill color → orange, pink → go back to blue
  - ▶️Add data labels → 14pt
- 💬This helps us not have to rely on interpreting data from the gridlines alone, it’s a great option for bar & column charts.
  - ❓What if you wanted to highlight median home price in San Francisco, since that’s the county your company is working in?
  - ▶️Format data point → SF → orange
- 💬Suppose you actually wanted to de-emphasize your other counties. Rather than having blue and orange, you can change these other bars to gray.
  - ▶️Series → gray
- 💬You don’t need a legend to distinguish SF from the other counties, that’s already clear from the axis
- 💬For the horizontal axis, this is the median home sales price. Overall looks good, we can increase the font size.
  - ▶️Font size → 14pt
  - ▶️Number format → currency
- 💬Vertical axis contains our counties, I think these could stand out a bit more
  - ▶️Bold → 14pt
- 💬Lastly, the gridlines and ticks. These are most relevant for numerical data, so not our vertical axis. Gridlines refer to the lines that span the entire chart, while ticks sit on the axis itself.
  - ❓Is it really important to read across the entire chart or just to visualize on your axis?
- 💬Major axis is coarser grained intervals, in this case the increment of $500k, could also implement minor gridlines and space according to count or step. I like 4 because it segments each section into $100k increments, or you can do step with $100k.
  - ▶️ Gridlines → minor → count 4
- 💬I don’t think we need ticks, given our data labels. I think major grid

	lines are sufficient. If you like the minor gridlines, you could consider taking off the axis labels • 💬Overall, I think this final design is pretty clean. ○ ▶️Toggle back and forth between old & new design
🎙️ TH	Don't be afraid to experiment and have fun with it! With a little creativity and attention to detail, you can transform your charts into powerful tools for communicating your data story. Follow me to the next video to explore creating a scatter plot that shows the relationship between home size and sale price.

L2v4 – Demo: Scatter plots

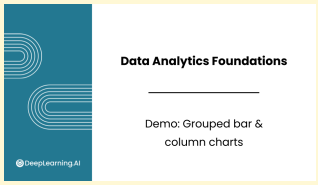
Visual	Script
🎙️ TH 	Scatter plots reveal relationships between two numerical features. Each point represents a pair of values: one on the x-axis and one on the y-axis. The position of the point shows how the two features relate to each other. Suppose you wanted to understand the relationship between home size and the median sales price. These are both numerical features, and thus excellent candidates to visualize via a scatter plot.

🔗DEMO SPREADSHEET	• 💬Starting in the data tab, I'll select median pending sqft and median sale price, then insert a chart. A scatter plot can more clearly display a large number of points, so selecting all the counties is okay!
	  ○ ▶️AD:AD, N:N → insert chart
	• 💬The chart defaulted to a column chart, which is not appropriate
	  ○ ▶️Open chart editor
	  ○ ▶️Change to scatter plot
	  ○ ▶️Move to tab
	  ○ ▶️Rename tab → "Median sqft vs median price"
	• 💬Now we can start our customizations!
	  ○ ▶️ Title → "Median home size vs median sales price in CA – Q3 2023"
	• 💬 Since median home sales is in the title, I think we can get away with the vertical, or y-axis label. The units are missing from home size, though.
	  ○ ▶️ Horizontal axis title → Median square footage → bold → 14pt
	• 💬 Remember each of these dots represents a county, not an individual house
	• 💬 Let's consider the markers. One thing to consider with scatter plots is the size of the markers. If you have a lot of data, you may want to use a smaller size, or reduce the fill opacity, which adds some transparency to the markers. These techniques help ensure you can see all of the individual data points, vs. having some (or many) overlap with each other. Since there aren't too many data points, you can make them

	larger. ○ ▶️ Marker size → 10pts • 💬 Data labels worked well in our column chart, what about here? ○ ▶️ Check data labels • 💬The chart gets very busy and difficult to read each label, so let's leave that off for now. • 💬 We could, however, look at adding a trendline, which is a useful tool in scatter plots that helps visualize the linear trend through the data. Let's see what that looks like. ○ ▶️ Check trendline → 4px → 60% • 💬We can see that there's a positive slope to this trendline, which suggests that as the median square footage increases for the county, the median sales price also increases on average. • 💬 Next, the horizontal axis. The chart zoomed into the observed bounds of our data. What does it look like if I set the minimum value to zero? ○ ▶️ horizontal axis → min → 0 • 💬 In reality, we are not likely to observe any median house sizes of, say, 200 square feet, but it helps to know how the trend behaves all the way from zero. • 💬 Let's customize the gridlines and ticks. Both axes contain numerical data, so these are worth tinkering with. Since data labels are too heavy, your audience may have to rely more on gridlines to estimate the axis values for each data point. ○ ▶️ Horizontal axis → Minor gridlines → 100 sqft increment ○ ▶️ Vertical axis → minor gridlines → count of 4 ($100k) • 💬 Now, we have a grid that enables us to estimate each data point fairly well, and the gridlines are not so heavy that they distract us from the data itself. I'm pretty happy with this.
	Now, let's analyze our chart. *What do you see in the data?* For example, you might say something like: • Based on the trendline, you observe a **positive correlation** between the median square footage and the median home sale price. As the median home size increases, so does the median sale price. However, this trendline doesn't run perfectly through the data. It actually seems to dissect the data into two groups: One lower priced group that follows a fairly consistent trend, and one higher priced group that doesn't appear to follow a trend at all. • Some of these highest price counties may be considered **outliers**, and are worth investigating further. Something seems different with the higher priced counties that seems to defy the laws of housing physics! Since some of the most expensive counties do not seem to have the largest homes, perhaps there are some other drivers of these high prices at play.

🎙️ TH	Remember, scatter plots are powerful tools for uncovering hidden relationships in your data. Use them to explore correlations, identify outliers, and generate hypotheses for further investigation. Follow me to the next video for another new chart type: grouped bar and column charts.

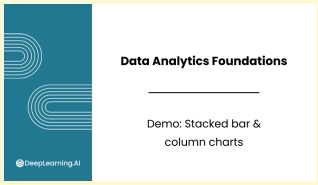
L2v5 – Demo: Grouped bar and column charts

Visual	Script
🎙️ TH 	Grouped column charts allow you to compare an outcome of interest across different groups, revealing complex relationships and trends in your data. Suppose you wanted to visualize the most popular male and female baby names across all years, using the baby names data set you've seen a few times so far. You would need to visualize the number of babies by both the rank and gender. Two features at once! That makes this data set a great candidate for a grouped bar chart.
🔗 **DEMO SPREADSHEET**	Let's start by taking a look at the data needed to create this chart. Here's a summary of the top 10 male and female names of all time based on the data. This data sums up the total baby count for each name and gender across all years, sorted by rank. It takes some work to assemble this data, and in Course 2 you'll learn more about pivot tables which make short work of this task! For now we'll start here. You'll need data in this format to create grouped bar charts in Google Sheets: a common set of categories, here it's Rank, and then distinct sets of counts for each gender.

	• 💬 To create this chart in Google Sheets, highlight the prepared data and insert a chart.
	  ○ ▶️ A1:E11 → insert chart
	• 💬 It looks like there is a grouped column chart created, but there is some work to do in order to configure this chart properly
	• 💬 Since there are 10 male and 10 female baby names, I am going to switch the chart type to a bar chart
	  ○ ▶️ chart type → bar chart
	• 💬 Next, I will assign the Rank column to the y-axis, and you'll see that the 1-10 labels appear as expected
	  ○ ▶️ y axis → rank
	• 💬 Next, there is some work to clean up the data series that were automatically populated in the chart. In the end, I would like to have one series for the male baby counts, and one series for the female baby counts.
	  ○ ▶️ Series → remove rank
	  ○ ▶️ Series → male count → add label male name

	○ ▶️ Series → female count → add label female name ○ ▶️ Use row 1 as headers • 💬 OK, that's a pretty good start. Let's customize! ○ ▶️ Move to sheet
	• 💬First, I'm going to add a title: ○ ▶️ Title → "Most Popular Baby Names in the U.S." → bold → center ○ ▶️ Subtitle → "1880 - 2022" → bold → center • 💬 Next, let's soften these colors in the series menu, they're a bit hard on the eyes ○ ▶️ Female count → dark green ○ ▶️ Male count → dark blue • 💬 I notice that the x-axis values are a bit hard to read with so many zeros, so let's add commas ○ ▶️ Horizontal axis → Number format → custom → other custom formats → #,##0 • 💬 Lastly, it's pretty challenging to estimate the baby counts with just the major gridlines, so I'm going to add some minor gridlines as well. ○ ▶️ Major gridlines → count 7 ○ ▶️ Minor gridlines → count 1 • 💬 That will achieve a 500,000 count increment. So now it's pretty clear that Elizabeth through Barbara are hovering around the 1.5M mark.
	OK, I think that's pretty good! What insights do you see? • For me, what stands out the most is that the Top 10 male baby names are significantly more popular than their female counterparts. In fact, the Top 10 male baby names are more popular than all of the Top 10 female baby names except for Mary. This trend seems to suggest that the male baby names are more concentrated, whereas the female names are more diversified.
🎙️ TH	Interesting! Great work on that Grouped chart. Follow me to the next video to see how to create a stacked chart!

L2v6 – Demo: Stacked bar and column charts

🎙️ TH 	Now, suppose we want to analyze trends over time in the proportion of male vs female babies for a unisex name. For that, you can create a stacked bar chart. Let's take a look at Riley, which is a fairly common unisex name with an interesting trend.
	Let's take a look at the data. You can tell there was some prep work here. This table shows the count of male and female babies for each year in the data set.

	Many of the earlier years do not have any female Riley babies.

🔗 **DEMO SPREADSHEET**	• 💬Now, let's create the chart
	  ◦ ▶️ Select A:D → insert chart
	  ◦ ▶️ Chart type → column
	• Even though we have many years in the data set, a column chart makes more sense because we typically want to have time on the x-axis. OK, let's clean this up
	  ◦ ▶️ Stacking → standard
	  ◦ ▶️ X axis → year
	  ◦ ▶️ Series → F, M
	  ◦ ▶️ Move to separate sheet

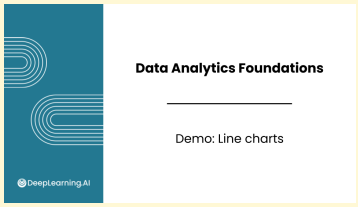
	• 💬 Let's customize!
	  ◦ ▶️ Title → "Riley Gender Distribution" → bold → center
	  ◦ ▶️ Subtitle → "1880 – 2022" → bold → center
	  ◦ ▶️ Series → F → dark green
	  ◦ ▶️ Series → M → dark blue
	• 💬I'm just going to make a few additional subtle changes to finalize things.
	  ◦ ▶️ Legend → position → inside
	• 💬That expands our chart space. Then I'll lighten up the gridlines a bit since they are not too important for the story.
	  ◦ ▶️ Gridline color → light gray

	Now, let's analyze our chart. What do you see? • For me, the trend is pretty interesting, in that the name Riley was fairly uncommon from 1880 all the way until the 1980s when it began to rise in popularity. It was also used exclusively for boys prior to this rise in popularity, and the began a steady increase in usage for female babies as well. After the turn of the century, Riley actually became a more popular name for female babies than males. It has remained that way ever since.

	Suppose I was less concerned about the **total count** of Riley babies, and I wanted to focus more on the **distribution** of male vs. female Riley babies. To answer this question more directly, I could use a 100% stacked column chart instead.
	
	• 💬Start by duplicating this chart, and then make a few changes:
	  ◦ ▶️ Setup → stacking → 100%
	• 💬 Wow, this looks pretty different, and the y-axis labels changed from counts to %ages.
	• 💬The blue is a little aggressive, so I'm going to lighten up a bit, and add a little bit of transparency. And I'll move the legend while I'm at it.
	  ◦ ▶️ Customize → Series → M → fill opacity 50%
	  ◦ ▶️ Legend → top
	• 💬 If I'm only focused on which is the dominant gender, I could actually

	reduce the gridlines to just include the midpoint, and make it a little darker so that it's easy to see which gender is the majority. ○ ▶️ Major gridlines → vertical axis → count 3 → dark gray • 💬It looks like it was the year 2003 when the female Riley babies began to outnumber their male counterparts.
🎙️ TH	Stacked bar and column charts are powerful tools for showcasing complex relationships between multiple features. Join me in the next video to learn about your final visualization type: line charts! Line charts are great for time series data. I'll see you there.

L2V7 – Demo: Line charts

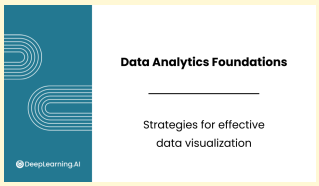
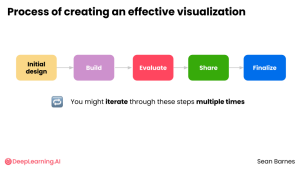
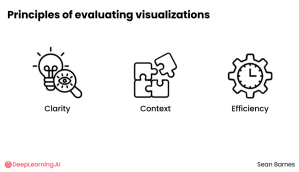
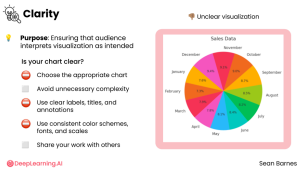
Visual	Script
🎙️ TH 	The primary chart type for visualizing time series data is called a line chart. Though they may seem simple, this type of chart is delivering business value daily.
🔗 **DEMO SPREADSHEET**	• 💬Does this look familiar? This is where we left off with Ruby names in Module 2. Does it feel like a long time ago?! • 💬Let's create a chart with both the count and moving average. ○ ▶️ C:C to E:E → insert chart ○ ▶️ Chart type → line ○ ▶️X axis → year ○ ▶️Series → remove year • 💬 You saw previously that these two series overlap, since the moving average is smoothed out over 10 years. • 💬Let's customize this chart! ○ ▶️Move to own sheet ○ ▶️Title → "Ruby Name Popularity - Assigned Female at Birth" → bold → center ○ ▶️Subtitle → "1880 - 2022" → bold → center ○ ▶️Horizontal axis title → Year → bold → 14pt ○ ▶️Vertical axis title → Baby count → bold → 14pt ~~• 💬The moving average is obscuring the data somewhat, so you can give it some transparency or make it lighter~~ ~~○ ▶️Series → Count → dark green~~ ~~○ ▶️Series → Moving average → lime green~~ • 💬Let's add some gridlines so the audience can more easily spot particular years. ○ ▶️Major gridlines → step → 25

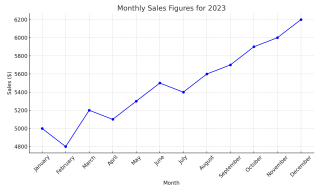
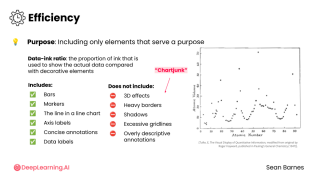
	○ ▶️Minor gridlines → step → 4
Continue in previous sheet	Now, let's analyze the chart, starting with the original count data plotted in dark green. What do you see in the time series data for Ruby? • The trend shows an initial rise in popularity in the late 1800s and early 1900s, peaking in the early 1920s and then proceeding to decline consistently for the next 50 years, until reaching a minimum around 1975. Then, in more recent years, there appears to have been a small resurgence. • You may remember that there was a sharp increase in Rubies in 1963. That isn't apparent in this chart. • This chart doesn't show yearly seasonality, but it's still possible Ruby has some seasonality. For example, if this data was collected monthly, there could possibly be a spike each year in July, which is associated with the ruby gemstone. If there were such a pattern, that would be monthly seasonality. • We can, however, see a clear cyclical pattern in the data, given the initial rise and fall of popularity, followed by the recent resurgence. This cycle would have been impossible to predict, but is clear from the historical data. Perhaps there will be another cycle in the future! We'll just have to see... • Overall, while there is some noise in the data, it doesn't distract from observing the overall patterns. The moving average follows the data quite closely.
🎙️TH	And that's a wrap on creating charts in Google Sheets! I hope you're feeling inspired to develop your own beautiful charts. You have a lot of capabilities now. Coming up, I hope you'll try out the practice lab where you will visualize the hotel reservations dataset you worked with in the previous module. Once you've finished the practice lab and assessment, follow me to the next lesson to learn about best practices in data visualization. See you then!

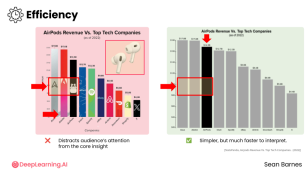
Lesson 3 – Best practices in data visualization

L3V1 – Strategies for effective data visualization

Visual	Script
🎙️TH – overlay of https://upload.wikimedia.org/wikipedia/commons/thumb/4/4c/Schroeder%27s_stairs.svg/580px-Schroeder%27	Look at this image for a moment. What do you see? **[pause]** Is it a staircase leading **[emphasis]** downwards from left to right, or the same staircase but **[emphasis]** upside down? **[eyebrows]** Hm. It could be either one. Two reasonable people can come to two completely different conclusions about this same image.

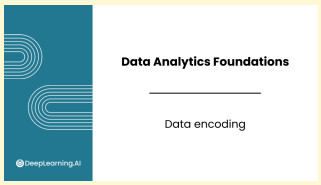
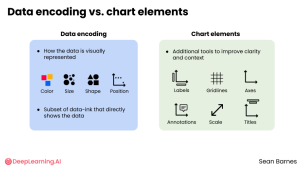
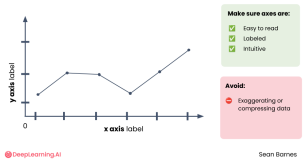
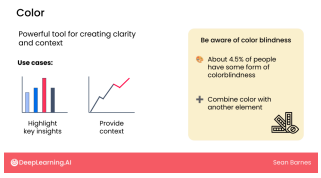
s_stairs.svg.png and can we emphasize each possible option visually as Sean mentions them? 	If you're not careful, your data visualizations can end up like this optical illusion – also called Schroeder stairs. You'll show them to your stakeholders and each one will come away with a different insight. How do you avoid this confusion? Two things. Follow a structured process for creating your visualizations. And implement principles of effective visualization design. First, I'll talk about the process.
Visualization design process (Initial design → build → evaluate → share → finalize) 	• Start by sketching out an **[CLICK]** initial design. While there are often several ways to convey the same information, one is likely best. • Then **[CLICK]** build the initial draft. As you work, consider how your visualization will be consumed. Can you give a guided tour or will it need to stand on its own? • Next, **[CLICK]** evaluate whether your visualization effectively conveys the key insights – I'll share some strategies in a moment. • Then, **[CLICK]** share your visualization to get colleagues' initial reactions. See whether it's clearly understood. • Lastly, **[CLICK]** finalize the visualization. Incorporate insights from the previous steps. Depending on the importance of the visualization and the stakes involved, **[CLICK]** you might iterate through these steps multiple times.
	When evaluating your own visualizations, focus on three key principles: **[CLICK]** clarity, **[CLICK]** context, and **[CLICK]** efficiency. Let's break these down. **[1 second pause]**
 Unclear chart (same thing here): https://i.imgur.com/HLsU9Hp.png	Clarity is about ensuring that your audience interprets your visualization in the way you intend. Your goal should be to have as many people as possible, especially your most important stakeholders, come away with the same insights. **[CLICK]** How do you know if your chart is clear? 1. First, **[CLICK]** choose the appropriate chart type for your data, which you already know how to do! 2. Second, **[CLICK]** avoid unnecessary complexity. Use simple, clean designs and avoid clutter. 3. Third, **[CLICK]** use clear labels and titles so your audience understands what they're looking at, plus annotations to highlight key insights. 4. Next, make sure you're **[CLICK]** consistent with your color schemes, fonts, and scales. Triple check that your chart is legible. Font size that's too small is a common mistake, since it can be challenging to fit so much information in one graphic. Larger font sizes are necessary for presentations, while you may be able to get away with a smaller size for

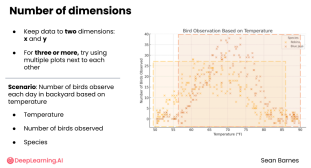
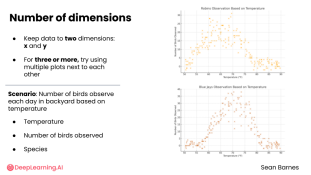
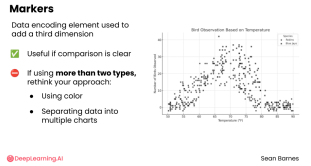
	individual viewing. 5. Lastly, **[CLICK]** share your work with others! Your manager, peers, and trusted stakeholders can provide valuable feedback. Getting a fresh perspective is crucial. **[CLICK]** Here's a visualization of monthly sales. What do you think? Is it clear what's happening? **[brief pause for digestion]** First of all, the **[CLICK]** chart type isn't right. In a pie chart, it's hard to compare the relative sizes of many slices, and it's not suited for time series data. **[CLICK]** There are so many colors and **[CLICK]** it's missing key information like the year and the actual sales figures. I don't get a clear sense of the trend in sales either. What else can I say, but **[CLICK]** "that's so wrong"!
  https://i.imgur.com/2OSUQnJ.png	Now here's another chart of the same data, this time a line chart with time on the x axis and sales on the y axis. At first glance, I can tell that sales trended consistently upward throughout the time period, which was the year 2023. The axes are clearly labeled, and I can estimate the sales figures from the y-axis.
I also like pp. 102 - 103 from Tufte subtracting the non data ink out – https://faculty.salisbury.edu/~jtanderson/teaching/cosc311/fa21/files/tufte.pdf 	Efficiency means including only the elements that serve a purpose. Ask yourself: is there anything in your chart that doesn't contribute to the story you're trying to tell? Statistician Edward Tufte, one of my inspirations, first coined the term **[CLICK]** **data-ink ratio**. It's the proportion of ink – or pixels – that is used to show the actual data compared with decorative elements. Above all else, focus on the data. **[CLICK]** Here's an example of a chart with a very high data-ink ratio, which shows the atomic number of an element on the x axis and its volume on the y axis. It displays only the most crucial information for interpretation. **[CLICK]** Data-ink includes **[CLICK]** bars, **[CLICK]** markers, **[CLICK]** the line in a line chart, **[CLICK]** axis labels, **[CLICK]** concise annotations, and **[CLICK]** data labels. **[CLICK]** On the other hand, **[CLICK]** 3d effects, **[CLICK]** heavy borders, **[CLICK]** shadows, **[CLICK]** excessive gridlines, **[CLICK]** overly descriptive annotations and so on are non data-ink. Tufte also used the term **[CLICK]** **chartjunk** to characterize distracting decorations that do not enhance the audience's understanding. Decorations, extra text, too many colors... If you find yourself adding stuff just to make it

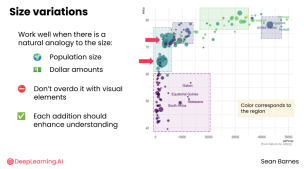
	your chart more interesting, stop! Let the data speak!
 **Must keep this image** As the airpods chart – https://i.redd.it/yfd30j4rbvjb1.png Nice airpods chart – https://i.imgur.com/EKQDNqC.png	Earlier, you saw a similar chart of AirPods revenue. This one is actually the original, with the chartjunk intact. Ask yourself, **[CLICK]** what does this picture of airpods in the corner really do? **[CLICK]** Why do the columns have shadows? Primarily, it **[CLICK]** distracts the audience's attention from the core insight about just how much money Airpods make. You may remember the same data from **[CLICK]** this chart in an earlier lesson. **[CLICK]** Extraneous elements – like the airpods picture and logos – have been subtracted out, while **[CLICK]** the airpods revenue has been highlighted with color. **[CLICK]** It may be simpler, but it's much faster to interpret.
	Context is about grounding your audience's understanding. Consider what background knowledge your audience has about this data. You can provide context in a few ways. You'll want to • **[CLICK]** Create a clear narrative structure – storytelling! • **[CLICK]** Provide relevant background information • **[CLICK]** Compare the insight with familiar concepts • **[CLICK]** Define jargon • **[CLICK]** Explain the significance of the data Depending on who you're presenting to, you may need to adjust your context. **[CLICK]** Err on the side of including more context. You'll rarely hear complaints about providing too much information, but a lack of it can lead to misinterpretation.
🎙️TH – https://cdn.mos.cms.futurecdn.net/rQkQZ6pDZbEHz23rxckWPm-320-80.jpg	A well-designed visualization that tells a clear story will have lasting impact. Remember that your goal is to create a visualization where everyone who looks at it comes away with the same, correct interpretation. I'll leave you with one last image. Do you see an old woman, or a young one? Join me in the next video to learn more about data visualization design.

L3v2 – Data encoding

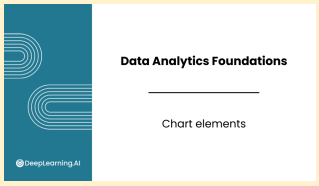
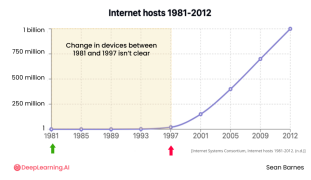
Visual	Script
🎙️TH	You know the playbook for effective data visualization design – clarity, efficiency, and context – but how do you actually create a chart that follows

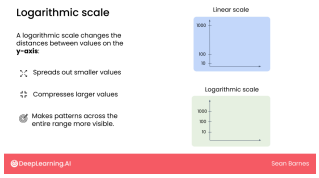
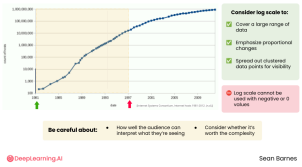
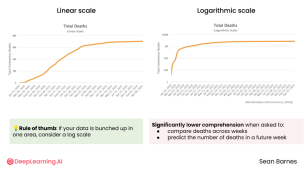
	them? The way I typically think about it is to first make sure your axes, including scale and labels, are impeccable. Then think really hard about how you can use color. And then think really really hard before you add anything else. Let's talk more about this hierarchy.
	First, there's a distinction between data encoding elements and chart elements. Data encoding means **[CLICK]** how the data is visually represented using elements like **[CLICK]** color, **[CLICK]** size, **[CLICK]** shape, and **[CLICK]** position. You can think of data encoding as **[CLICK]** the subset of data-ink that directly shows the data – not the labels, gridlines, axes and so on. Data encoding forms the backbone of your visualization. Chart elements includes everything else: **[CLICK]** labels and **[CLICK]** gridlines and **[CLICK]** axes, plus **[CLICK]** annotations, **[CLICK]** scale adjustments, and **[CLICK]** titles. **[CLICK]** These are additional tools you can use to improve clarity and context, but they should be used judiciously to maintain efficiency. In this video, I'll focus on data encoding, with chart elements in the next one.
Building up a single data visualization e.g. a multiple line chart 	Let's return to that basic process from earlier. • Start with your **[CLICK]** x and **[CLICK]** y axes. **[CLICK]** Make sure they are **[CLICK]** easy to read, **[CLICK]** labeled, and **[CLICK]** intuitive. **[CLICK]** Consider including zero for numerical features. • **[CLICK]** Scale your axes appropriately.**[CLICK]** Avoid **[CLICK]** exaggerating or compressing the data, which can **[CLICK]** distort the message. • Labels should be clear and concise. They're especially helpful when it's difficult to read exact values from gridlines.
	Color is one of your most powerful tools for creating clarity and context. For instance, you can use color to **[CLICK]** highlight key insights, like your company's performance compared to competitors. Or you can use color to **[CLICK]** provide context, like graying out historical data to focus attention on the current year's performance. **[CLICK]** Be aware that some of your audience may have difficulty differentiating colors. About **[CLICK]** 4.5% of people worldwide have some form of colorblindness, typically red/green colorblindness. When possible, **[CLICK]** use double encoding by combining color with another element like unique markers. **[CLICK]** Additional clarity helps everyone.
Can use these graphs or similar looking ones with same subject	But be mindful of how many dimensions you're asking your audience to interpret simultaneously. In general, **[CLICK]** keeping your data to two dimensions – x and y – helps your audience interpret the right insight. **[CLICK]** If you truly need to show three or more dimensions, try having multiple plots

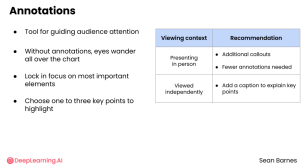
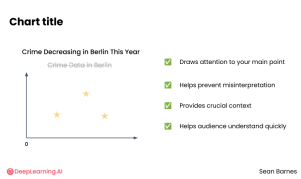
3 dimensions – https://i.imgur.com/r3guFdx.png 3 plots with 2 dimensions – https://i.imgur.com/ly4JbYg.png 	next to each other. Here's an example. Say you're trying to create a chart to show the **[CLICK]** number of birds you observe each day in your backyard based on the temperature. There are two species of birds you typically track: robins and blue jays. So your data has three dimensions: **[CLICK]** temperature, **[CLICK]** number of birds observed, and **[CLICK]** species. Let's see all those dimensions plotted on the **[CLICK]** same chart: temperature on the x axis, number of birds on the y, and the bird species in different colors. These two species seem to prefer different temperatures – **[CLICK]** robins prefer moderate temperatures, **[CLICK]** while blue jays prefer higher temperatures. I find myself to be most similar to robins. But there are a lot of observations on this graph and it's hard to tell the two apart. ~~In particular, observations with less than 5 birds are muddled.~~ One option to improve clarity is to
	separate this data into two scatter plots, each with a single species. The patterns for the individual species become clearer, while still allowing your audience to compare habits across species. The remaining elements you'll see in this video should be used judiciously. They're often more difficult to interpret. Let's start with markers.
3 dimensions in color – (source) 3 dimensions with markers(source) https://i.imgur.com/bON6WVQ.png https://i.imgur.com/RRaFi3D.png and https://i.imgur.com/Cw5N4OH.png  https://i.imgur.com/KUDGgU0.png	Markers are a **[CLICK]** data encoding element typically used in scatterplots to add a third dimension to the data. You saw the bird scatterplot a moment ago with color differentiating the two series. **[CLICK]** Here's that same data, this time using markers rather than color for differentiation. Do you think this graph is easier or harder to interpret? **[pause for learner to think]** I'd say harder. Since the markers are so small, the different shapes don't jump out at me. **[CLICK]** Markers can be useful if the comparison is clear. **[CLICK]** But if you find yourself using more than two types, it might be time to rethink your approach. Consider **[CLICK]** using color instead, or **[CLICK]** separating your data into multiple charts.

Must use this graph 	Size variations, often seen in bubble charts, also add a third dimension to your visualization. They **[CLICK]** work well when there's a natural analogy to size, like **[CLICK]** population size or **[CLICK]** dollar amounts. **[CLICK]** Here's an example of a bubble chart with population size determining the bubbles. It plots countries by wealth on the x axis and life expectancy on the y axis. Annotations help the audience spot some of the most interesting points. Can you spot China and India, the world's two most populous countries? **[pause for learner to look]** They're **[CLICK]** here, the two largest bubbles! By the way, color adds a fourth dimension to this chart. Can you guess? **[pause]** **[CLICK]** Color corresponds to the region, with **[CLICK]** purple for Africa, **[CLICK]** light blue for Asia, **[CLICK]** green for Europe, and **[CLICK]** dark blue for the Americas. That's a lot of data to interpret at once! Remember, efficiency is key. **[CLICK]** Don't overdo it with visual elements. **[CLICK]** Each addition should serve a clear purpose in enhancing understanding.
TH	Now that you're familiar with using data encoding in your charts, join me in the next video to explore how chart elements can make your insights even clearer. See you there!

L3v3 – Chart elements

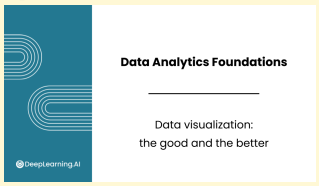
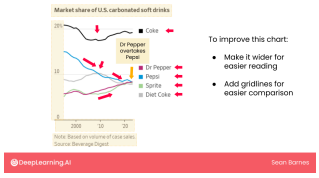
Visual	Script
TH 	Let's talk chart elements. These are the parts of the chart that don't directly represent the data.
 Linear graph – https://i.imgur.com/9SDetRB.png spread out clustered data points	Take a look at this graph of internet hosts, which means the number of devices connected to the internet. The x axis shows time, while the y axis shows the number of internet hosts. It's clear that internet usage skyrocketed between the start and end years. But the **[CLICK]** change in devices between 1981 and 1997 isn't as clear. How could you better show this data?

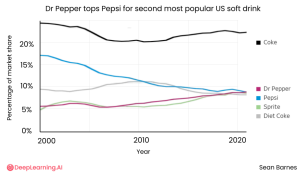
	One technique is to use a logarithmic scale instead of a linear one. **[CLICK]** On a linear scale, values are evenly spaced. If you plot 10, 100, and 1000, the second distance is ten times larger than the first. **[CLICK]** A logarithmic scale transforms these values, making 10, 100, and 1000 evenly spaced. **[CLICK]** It changes the distances between values on the y-axis, **[CLICK]** spreading out smaller values and **[CLICK]** compressing larger ones, **[CLICK]** making patterns across the lower range more visible. You can create a log scale by transforming the values in your data using a base 10 logarithm. Don't worry about the math – software will do this step for you.
https://upload.wikimedia.org/wikipedia/commons/c/c6/Internet_host_count_1988-2012_log_scale.png – logarithmic graph 	Here's the same data set of internet hosts, this time with a logarithmic scale on the x axis. Notice how the distance between **[CLICK]** 1 and 10 thousand, which is 9 thousand, is the same length as the distance between **[CLICK]** 10 and 100 thousand, which is 90 thousand. It's much clearer in this version of the chart that there was **[CLICK]** substantial growth between 1981 and 1997. **[CLICK]** Consider a log scale when you want to **[CLICK]** cover a large range of data, **[CLICK]** emphasize proportional changes over absolute values, or **[CLICK]** spread out clustered data points for visibility. One limitation of the log scale is that it **[CLICK]** cannot be used with negative or 0 values. Again, don't worry too much about the math, but the log of negative numbers and 0 is undefined, which you just can't graph. While I am a big fan of log scale, **[CLICK]** you have to be careful about **[CLICK]** how well the audience can interpret what they're seeing. People's brains don't naturally think logarithmically. **[CLICK]** Consider whether the storytelling value is worth the added complexity.
Use covid line graphs from here: https://blogs.lse.ac.uk/covid19/2020/05/19/the-public-doesnt-understand-logarithmic-graphs-often-used-to-portray-covid-19/ 	Here's an example of two graphs of COVID19 deaths over time. Both of them show time on the x axis and total deaths on the y axis. However, the graph on the left uses a **[CLICK]** linear scale on the y axis, while the graph on the right uses a **[CLICK]** logarithmic one. Researchers studied regular people's interpretations of these graphs. Which do you think is easier to interpret? **[brief pause for thought]** They found that people who were given the graph on the right scored **[CLICK]** **[CLICK]** significantly lower on comprehension when asked to **[CLICK]** compare deaths across weeks and **[CLICK]** predict the number of deaths in a future week. One issue is that the log scale can't display values before around March 1st, when covid cases were 0. That's because the log of 0 is undefined. It's tough for the casual viewer to understand what the abrupt start of the line means. **[CLICK]** So, a good rule of thumb: if your data is bunched up in one area on a linear scale, it might be time to consider a log scale or adjust your axis limits.

	Just be mindful of how your audience will interpret the scale.
 Misleading graph – https://i.imgur.com/MsoXkmP.png Graph with zero – https://i.imgur.com/SoLVcRN.png	Finally, with regard to axis scale, consider including zero. **[CLICK]** This tactic helps communicate the magnitude of your data, in particular when absolute value matters more than relative value. Leaving out zero is a common tactic in misleading graphs. That said, **[CLICK]** excluding zero by truncating one of your axes can be helpful if you need to emphasize small differences. You can think of it like zooming in on your range of interest. **[CLICK]** Here's a graph of pretzel sales in 1967, with time on the x axis and sales on the y axis. It looks like Golden Loops is really beating out the Twist and Shout brand. **[CLICK]** But look – the scale starts at 950. It emphasizes the differences, since the ~~green~~ bars seem twice or three times higher than the purple ones. **[CLICK]** Here's the same data with zero included in the scale, showing the absolute values. It looks like these two brands are pretty competitive. Either of these graphs could be appropriate depending on the circumstance. The graph on the left may be helpful in a Twist and Shout brainstorm about taking over Golden Loops, since it **[CLICK]** zooms in on the differences between the two brands. The graph on the right might be more helpful for a supermarket looking to **[CLICK]** assess the most popular brand.
	Next up, annotations. These are a fantastic **[CLICK]** tool for guiding your audience's attention. **[CLICK]** Without annotations, people's eyes might wander all over your chart. With well-placed annotations, you can **[CLICK]** lock in their focus on the most important elements. Remember, efficiency is key. Don't overwhelm your chart with annotations. **[CLICK]** Choose just one to three key points to highlight. **[CLICK]** You should also consider how your audience will encounter the chart. If you're **[CLICK]** presenting in person, you can use a laser pointer for **[CLICK]** additional callouts, **[CLICK]** so fewer annotations may be needed. If your chart will be **[CLICK]** viewed independently, such as in a report, **[CLICK]** consider adding a caption to explain key points, since you won't be there to tell the narrative.
	Once you've chosen your scale and added annotations, you'll want to pick a great chart title. Your title isn't limited to just describing what the chart shows. Use it as an opportunity to highlight your key insight. Instead of **[CLICK]** "Crime Data in Berlin", consider something like **[CLICK]** "Crime Decreasing in Berlin This Year". This immediately **[CLICK]** draws attention to your main point and **[CLICK]** helps prevent misinterpretation. Titles can also **[CLICK]** provide crucial context, like the time period your data covers, which **[CLICK]** helps your audience quickly understand what they're looking at.

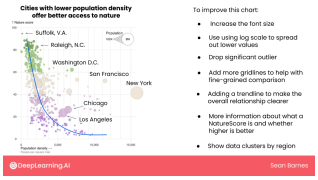
🎙️ TH	At this point, you've seen the core techniques for building beautiful and functional data visualizations. Join me in the next video to see how to improve some cool charts! You won't look at a graphic in the news the same way again, trust me. See you there!

L3V4 – Data visualizations: the good & the better

Visual	Script
🎙️ TH 	Well… **[Sean puts on sunglasses]** it's time to adopt a critical eye and explore some data visualizations I've found in the wild. For each chart, I encourage you to pause the video and think about the story it's trying to tell and what improvements you could make to help the story stand out. Then you can hear my side of things. Let's get into it.
https://i0.wp.com/flowingdata.com/wp-content/uploads/2024/06/Dr-Pepper-ties-with-Pepsi.png?w=1384&quality=80&ssl=1 	Here's our first graph. Pause the video for a moment and see how you might interpret and improve it. • So right off, we're looking at a graph showing the **[CLICK]** market share of U.S. carbonated soft drinks. Pretty straightforward **title** there. • On the **[CLICK]** **x-axis**, we've got time running from 2000 to right after 2020. That's a two-decade span. • The **[CLICK]** **y-axis** shows the percent market share of each drink. • Because this is time series data, a line chart seems to be the **right visualization**. So far so good. • Checking for **encoded categories**, the different soft drink brands are color-coded, which is nice. That technique makes it easy to track each one without having to squint at labels. We've got **[CLICK]** Coke in black, **[CLICK]** Dr Pepper in magenta,**[CLICK]** Pepsi in blue, **[CLICK]** Sprite in green, and **[CLICK]** Diet Coke in gray. The colors also more or less match the brand colors associated with each drink, which is a nice touch. • There aren't many **annotations**, but **[CLICK]** the note at the bottom about volume of case sales provides context on what we're actually measuring here. • Looking at the **big picture**, you can see some interesting trends. Coke's been dominating the whole time, **[CLICK]** but it had a dip in the middle there. Meanwhile, Pepsi's been on a **[CLICK]** steady decline. **[CLICK]** Dr Pepper and Sprite have been slowly climbing up, while **[CLICK]** Diet Coke saw a bump but then **[CLICK]** started falling again. One key insight – which I can barely make out – is that **[CLICK]** Dr Pepper

	appears to be overtaking Pepsi, especially when you consider the trends of each soda.
	• To improve this chart, you could **[CLICK]** make it wider for easier reading. It looks like the chart may have been designed for a constrained space, but a wider one will show the change over time more clearly. • **[CLICK]** Adding some gridlines wouldn't hurt, so the audience can more easily compare the sales across the different brands. Just having 10 and 20% makes analysis pretty coarse. • The axis labels could also be more consistent. On the x axis, 2010 and 2020 can be **[CLICK]** lengthened for consistency, and the percent sign could be added to the y axis labels. **[CLICK]** Axis titles could improve readability. • Finally, an **[CLICK]** improved title like “Dr Pepper tops Pepsi for second most popular US soft drink” would help the audience focus on the main insight.
🔗**SCREENCAST** – we can cut the middle categories – purchasing to caring for non household members to save space	Alright, let's check out this chart, which is published on the website for the US Bureau of Labor Statistics. • First up, we've got a pretty clear **title** here: 'Average hours per day spent in selected activities by age, 2023 annual averages'. Straightforward enough. It doesn’t tell us a key insight, but rather sets the context for the chart. • Looking at the **x-axis**, we're dealing with hours per day, running from 0 to 12. • On the **y-axis**, we've got a list of different activities. No label needed here since it's self-explanatory. • Now, **encoded categories**. Color is used to represent age group,. 15 to 19 year olds are dark blue, 35 to 44 year olds in light blue, and 75+ in green. The distinct colors make comparison easy. ○ However, even distinct colors likely wouldn’t make **[select all boxes]** comparing across all 8 categories easy, especially since these colors don’t have a logical correspondence. There's no particular reason why 35 should be blue and not red. The bars also get quite small. Allowing the audience to select only the age groups relevant to them makes the graph easier to interpret. • A horizontal bar **chart makes sense** here given all these categories we're comparing across. A column chart wouldn’t have enough space for each label. Since the chart is grouped, that helps the audience compare within an activity, to interpret insights like “15-19 year olds

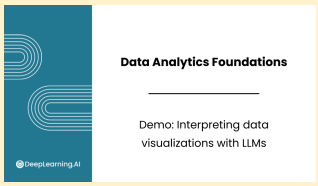
	spend significantly more time on educational activities than either of the other two groups". A stacked chart could also be suitable, if you wanted to better compare how the composition of time spent changes across age groups. • The light gridlines are just right - they give the audience a sense of scale without cluttering things up. No need for super precise measurements here. • I don't see any **annotations**. Honestly, you don't really need them. The data speaks for itself here. • **Big picture**, there's a lot to unpack here. Personal care and sleep is taking up the most time across all age groups, no surprise there. But check out how work time peaks in the middle age groups and then drops off. And leisure time? That's on the up and up as people get older. Education time is pretty much owned by the younger crowd, as you'd expect. • If you wanted to improve this chart, you might consider making the x axis label clearer, maybe "Average hours spent per day". It might be nice to be able to sort these categories by age group as well. • You could also group by age instead of activity. That could give us a clearer picture of how a typical day looks for each age group. • Overall, this chart isn't shouting one clear message at us. It's more like a buffet of insights about how we spend our time throughout our lives.
https://flowingdata.com/2024/04/12/access-to-nature-where-you-live/ 	We'll wrap up with a bubble chart, which is a variant of a scatter plot you've seen before. I know the font size is tough to read – I'll address that in a moment. • **[CLICK]** So, the **title**: "Access to nature where you live". It's relatively clear but doesn't share the key insight. • **[CLICK]** On the **x-axis**, you're looking at population density, measured in people per square mile. It runs from 0 to 30,000, showing a wide range. • **[CLICK]** The **y-axis** shows us the NatureScore, which goes from 0 to 100. I'm assuming that higher scores mean better access to nature. • The **chart type**, a scatterplot, is a solid choice when you want to show the relationship between two numerical features, in this case population density and the NatureScore. • Now, let's talk about those **encoded categories**. This chart has double encoding going on with color, which helps define categories of **[CLICK]** NatureScore using a diverging color scale. **[CLICK]** Greener means higher score, **[CLICK]** purple means lower, and **[CLICK]** tan is in the middle. Green for a higher nature score is a nice touch, since most

	people will associate green with nature. I'm not sure purple is as intuitive on the lower scale. • The **size** of the dots is another encoded feature - representing population. Bigger cities get bigger dots, which makes sense. **[CLICK]** The legend in the top right is helpful, showing us the scale for population size. It gives context to the dot sizes we're seeing. • In terms of **annotations**, they've done a good job here. We've got labels for just a few of the **[CLICK]** bigger cities and some outliers like **[CLICK]** Union City, N.J. and **[CLICK]** Suffolk, Va., which seems like a nice place to live! It's not cluttered, but gives us some reference points, especially for more populous cities the audience may be familiar with. • Looking at the **big picture**, you can see a relationship. **[CLICK]** As population density increases (moving right on the chart), the NatureScore tends to decrease (moving down). But there's some variability - it's not a perfect correlation. • Some cities buck the trend. **[CLICK]** Washington D.C., for example, has a relatively high NatureScore despite its population density. On the flip side, Union City, N.J. is way out there with super high density and very low NatureScore. Unfortunate.
	• If you wanted to improve this scatter plot, first **[CLICK]** increase the font size. You'll also notice how the data is all **[CLICK]** clustered together on the left hand side of the chart, with nearly half of the x axis dedicated to just two cities. Can you think of one way to solve that? **[pause for thought]** You could use a **[CLICK]** log scale! Log scales help spread out lower values.
	• Another option is just to drop Union City, NJ, since it's such a significant outlier. • **[CLICK]** More gridlines would help with fine-grained comparison. • You could also consider **[CLICK]** adding a trendline to make the overall relationship even clearer. **[CLICK]** More information about what a NatureScore is and whether higher is better could be helpful. Right now, the audience has to assume. The title could use some work, too. Maybe something like **[CLICK]** "Cities with lower population density offer better access to nature"? Also, **[CLICK]** it might be interesting to see how this data clusters by region - are West Coast cities different from East Coast ones, for example?

	• Overall, this chart does a great job of presenting complex data in a visually appealing and intuitive way. The relationship between population density and access to nature is clear, but there's enough nuance to keep it interesting. It's the kind of chart you could spend a while exploring and still find new insights.
🎙️ TH	Great work analyzing those charts with me. I want you to feel inspired! Make your visualizations memorable. And one professional tip for you: have a personal hall of fame of your best data vis work. It will come in handy in the job search. That takes us to the end of this lesson. Once you've completed the practice assessment, join me in the next lesson to learn how to leverage LLMs for data visualization. I'll see you there!

Lesson 4 – Data visualization with LLMs

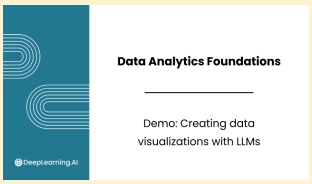
L4V1 – Interpreting data visualizations with LLMs

Visual	Script
🎙️ TH 	So earlier in this course, I didn't quite tell you the whole truth. I told you that Claude 3.5 Sonnet and ChatGPT 4o are both LLMs, large language models, but they're actually more than that. They're large multimodal models. Large multimodal models can interpret different types of inputs, typically text and images. This is really cool for data visualization because they can help you interpret charts as well as create them, which you'll see in the next video. These models do make mistakes, and you will have to check them, but they can help your process a lot. Let's see what that looks like.
Link to initial cut & screencast for timestamping 🔗**SCREENCAST** - 🔗nature score Walk me through this chart and explain the key insights it contains.	Earlier in this module, you saw this visualization of population density and nature score. Let's ask Claude to interpret this chart for us. So in this case, I'm going to screenshot the chart and drag it here to add the file to the chat. And I'm going to ask the model to walk me through this chart and explain the key insights. It's a pretty simple prompt. It says the chart visualizes the relationship between population density, access to nature, and population size for various cities and regions in the US. So it breaks down the three different dimensions that the chart contains, which is awesome. First, it goes through the axes. So on the x axis, we have population density.

	On the y axis, nature score, which it correctly points out is measured from zero to 100. And each bubble represents a city or region, with the size of the bubble indicating the size of the population. It also walks through the key insights. So first it points out this inverse relationship. There's a trend showing that as population density increases, the nature score decreases. I'm going to click on this image so that we can see a little better. We see that as the population density increases to [00:02:00] the right, the general trend is that the nature score decreases. So that seems pretty accurate. There are also some outliers as we saw. So Washington DC has a relatively high nature score. And then it also points out that the really, really large cities tend to have lower nature scores. So you just saw Claude 3.5 Sonnet's capabilities. Keep in mind that at the time that you're viewing this video, there are likely more advanced models out than this one, as model capabilities are changing very quickly.
🔗**SCREENCAST** - 🔗https://i.imgur.com/r3guFdx.png You are an expert data analyst with a passion for data visualization. I am trying to visualize the habits of three different bird species in my backyard. Here is a chart I created to show how many times I spotted each species in a day based on the temperature. Explain each distinct area for improvement in this chart with a specific suggestion for how to fix the issue. Be concise and don't waste my time.	Let's see how you can use a large multimodal model to critique your own charts. Let's take a look at ChatGPT 4o's capabilities, and in this case, we're going to use this visualization of the number of birds observed versus the temperature. In this case, there are actually three bird types on the graph. I can just right click and copy this image, and paste that image in there. And I'll use a little bit more complex of a prompt. I'll start off by giving the model a role: you're an expert data analyst with a passion for data visualization. Then a task: explain each distinct area for improvement in this chart with a specific suggestion for how to fix the issue. And then I'm going to give it additional instructions. Be concise and don't waste my time. So don't be afraid to tell the model exactly what you want it to do. So let's see the output. Because of the way I formatted my prompt – asking for an issue and a suggestion – it's going to give me those pretty clearly. The first issue is that the colors for bird 1 and bird 2 are very similar and can be hard to differentiate, so it asks me to choose distinct colors. We identified this issue before, that it was really hard to tell these colors apart from each other. Now arguably all three of these colors are too similar, so it's interesting that it only pointed out that bird one and bird two are too similar. There are also many overlapping points, so we can use transparency or alpha to resolve the issue or jitter them slightly to reduce the overlap. This is an interesting suggestion. You may not have the capability to be able to do this in Google Sheets, but this might be a completely new idea you hadn't yet

	considered. Previously, you resolved this issue by having three separate plots next to each other. The axis labels are clear but could be more descriptive,okay, so here's a mistake, right? It just suggested the same axis labels, which is not that helpful, but it's affirmation that you did that correctly. **[chuckle]** • It also pointed out that the legend overlaps with the data points. That's possible, but it doesn't really look like it's overlapping. • The title is general and could provide more context. That's definitely true. • Data point size. Reduce the size of the data points. So do I think that's an issue? Probably not. These data points seem like they're pretty well sized. Maybe as dots instead of Xs, they might be a little more visible, but that wasn't its exact suggestion. • Lastly, the gridlines are quite prominent. Okay, I actually think that the gridlines help organize the data really well.
🔗**SCREENCAST** You are an expert data analyst with a passion for data visualization. I am trying to visualize the habits of three different bird species in my backyard. Here is a chart I created to show how many times I spotted each species in a day based on the temperature. Explain each distinct area for improvement in this chart with a specific suggestion for how to fix the issue. Be concise and don't waste my time.	Let's see that same prompt with Claude. It's great to test different models and see what's working best, especially when it comes to really advanced capabilities like interpreting a very complex data visualization. Different models may perform better. Okay, Claude says that the data points overlap, obscuring the patterns, and it suggests transparency or a 2D density plot. Interesting suggestion! This may also be something you hadn't considered. It also suggests that you should use more contrasting colors. And it noted that the y axis scale had a large range and masked lower value patterns. So this is a really interesting observation. You saw in an earlier video that there were a lot of observations between zero and five birds. A log scale seems a little excessive here, but it's something you could consider if you really wanted to highlight the differences in that lower range. Again, it commented on the legend placement, so it's picking up on the fact that the legend is over the chart area. It felt the gridlines were too prominent and distracted from the data. We already identified that we disagreed with that, and the title lacks specificity.
🎙️TH	Between **these** two use cases, it seems like LLMs are more useful for interpreting data visualizations and maybe less useful – though still a good second pair of eyes – for critiquing them. Don't hesitate to take a screenshot and throw it into Claude or ChatGPT to see what kind of insights they can help you with.

L4V2 – Creating data visualizations with LLMs

Visual	Script
🎙️ TH  **Link to initial cut & screencast for timestamping**	Let's see how a generative AI tool like Claude or ChatGPT can help **create** data visualizations.
🔗**SCREENCAST** Here is some data about hotel reservations:	Let's start with Claude. I'm going to grab my data set and add it to the conversation. And right away you can see that this file is much too large for Claude to handle. I have 36,000 rows and Claude is telling me that this is too large. That's a sign that I need to interact with this tool in a different way. Instead, I'll use ChatGPT.
🔗 **SCREENCAST** Help me visualize the total number of bookings in each of the 12 months. Please adjust the chart so it is a bar chart rather than a column chart, and so it displays the month name rather than number on the y axis. Please create a 100% stacked column chart showing the canceled and not canceled reservations based on market segment. I want to be able to easily compare cancellation rates across market segment.	So let's go over to ChatGPT. I'm going to upload my entire dataset. All 36,000 rows. And ChatGPT is completely fine with that. And I'm just going to say, Help me visualize the total number of bookings in each of the 12 months. It gives me this nice visualization, which is very, very similar to the one that we just created. The scale is a bit different on the y axis and it has more gridlines, which helps with readability. Say I want the month name rather than the number to be displayed on the x axis. And I want this to be a bar chart rather than a column chart. Right now month is on the x axis, but I'm treating these more as categories rather than a series over time. Let's say I wanted to visualize cancellation rate across market segments. Here's one way to do that. Looks good! Gut check: you can check against your spreadsheet. Remember that about ⅔ of reservations aren't canceled, so that seems to check out here. Let's try two more visualizations for fun. **[next two prompts with commentary]** Let's ask one more question. How can I visualize the different values for average price per room for canceled vs not canceled bookings? It suggests a box or violin plot. You might not have seen these before! Let's just try a box plot for fun. You'll learn more about box plots in the following course. The range of values is between 0 and around 550. This central line is the median or

Can you plot average booking value by month? Can you create a bar chart to visualize lead time by repeated guests and booking status? How can I visualize the different values for average price per room for canceled vs not canceled bookings? Let's do a box plot.	the middle value. For canceled bookings it looks like the median is a tad higher than for not canceled ones, maybe by about 10 Euros.
🎙️ TH	You've seen that LLMs can be a really helpful part of your workflow to both enhance your own visualizations as well as create new visualizations for you. This strategy is great for quickly exploring data. Great work! You're almost to the end of this module. Next up, you'll complete a practice lab to test your skills working with LLMs for data visualization. When you're done, you'll take the graded assessment and graded lab. The lab explores market research for a bike sharing service. I know you'll have fun creating tons of cool visualizations to support your insights. Once you're done, follow me to the next and final module of this course, which is all about your role in the data analytics lifecycle. See you there!