DAG C1M4 scripts

Introduction

L0V1 – Module 4 introduction

Visual

Script

🎙️TH

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 and boring [RIGHT HAND] and beautiful 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.

Data storytelling is all about [CLICK] translating the 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?

* [CLICK] First is the business problem. Always remember you're telling this story for a reason. Focus on your audience and your goal.
* [CLICK] Next, there's the data itself. Because data is your raw material, it dictates what kinds of stories you can tell. A rideshare dataset won't help you tell a story about rabbit purchases.
* [CLICK] Then, there's the analysis – the process of extracting insights from the data. The descriptive statistics you calculated in the previous lesson like means and percentages are fantastic tools here, but more complex analyses are certainly valuable too.
* [CLICK] Finally, there's the visualization – the way you present the data visually to your audience.

Doesn't it remind you of the data analytics lifecycle? [SEAN WINKS BEHIND CAMERA]

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 increased by 35 years, from about 30 to 71 in 2021, reflecting huge advances in areas like nutrition and healthcare worldwide. (source)
* [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. (source)

[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. This combination provides context and 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, [CLICK] 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
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

I mentioned emotional data stories earlier. Let's see an example. [CLICK] One of my favorite visualizations is Charles Joseph Minard's famous March to Moscow graphic. [CLICK] Minard's 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.

[CLICK] This visualization is both a 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, [CLICK] marches to Moscow on the right, and [CLICK] 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?

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.

Since data visualization is such a huge part of data storytelling, in this video, you'll see a breakdown of the common components of a visualization. Before you create beautiful ones, you'll need to be able to interpret them skillfully.

Source

Let's start with an example. Let me orient you to this chart.

* [SEAN AD LIBS A BIT] 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 didn't like a particular movie character named Alexa, or that they might accidentally trigger their Alexa by just talking to their child. This chart doesn't attempt to tell that story, but it's 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. Basically, I'll be your guide through this data jungle. Our goal is to understand the main insight of this chart.

* [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. You can see that in the earlier time periods in the graph, from about [CLICK] 1790 through 1920, Manhattan was by far the most populous borough. [CLICK] But around 1920-1930, Brooklyn took over as the most populous borough and that trend has maintained through the turn of the century. [CLICK] It's also interesting that the boroughs appear to be sorted from most to least populous both in terms of the order of the segments in the columns and the order in the legend, based

on where the order landed in 2010. This intentional ordering throughout also enables you to trace the trends within each category over time, in addition to comparing them across categories.
* 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 phone 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. 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.
   1. [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.
   2. [CLICK] Then check your y axis. What is changing from [CLICK] bottom to [CLICK] top? Typically, it should increase.
3. [CLICK] Next, review any encoded categories. Analyze the chart for any indications of color, markers, or size, and read the legend to identify how different categories in your data may be encoded in the chart. 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 in your daily life, whether in the news, at work, or in this course, try to identify these components. It's a great way to practice your data visualization literacy and to start thinking critically about how data is presented visually.

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 – [CLICK] seek to 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?

* 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.

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.
* [CLICK] First the title: AirPods Revenue vs. Top Tech Companies. [CLICK] And the subtitle: as of 2022. I expect to see the revenue generated by airpods alone vs some top tech companies.
* [CLICK] On the x axis, what do you see? These are the top tech companies like Asus, Adobe, Intuit, Spotify, and so on.
* [CLICK] What about the 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. [CLICK] I also notice each bar has a 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.
* [CLICK] Look for encoded categories. What do you see? [CLICK] This chart uses color to highlight airpods compared with the other companies.
* [CLICK] So, big picture, what insight is this chart communicating? [CLICK] It's expecting you to make a comparison between the revenue generated by AirPods and the revenue from these huge companies. [CLICK] AirPods generated 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.com/media?url=https%3A%2F%2Fi.redd.it%2Fuwe4cchpfz5c1.png

* Here's another interesting one, this time a line chart.

* [CLICK] Start with the 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.
* [CLICK] What do you see on the x 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.
* [CLICK] What about the 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.
* [CLICK] Now check the encoded categories. I don't see any markers or 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.
* [CLICK] 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. [CLICK] Meeting online surpassed the previous top method of meeting through friends around 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 contributed positively to 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 methods.

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 scatterplot.
* First, [CLICK] look at the title: Nests above the pivotal temperature produce more female baby green turtles, and [CLICK] the 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 variable, 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 variables.
* [CLICK] On the 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.
* [CLICK] What's on the 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.
* [CLICK] Now the encoded categories. What do you notice? [CLICK] At first the color may not jump out at you, but it's a nice natural encoding with colder temperatures as blues and higher temperatures as pinks. [CLICK] You should also notice the different markers for nests with more male and 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.
* [CLICK] Do you see any annotations? [CLICK] This 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.
* [CLICK] So, big picture, what insights can you get from this chart? [CLICK] Since it's a scatterplot, one technique you can use is to 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 hatchlings 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

Follow me to the next video 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 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: bar or column charts, line charts, scatter plots, and 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] sales per region – here you can see the sales in each continent – or [CLICK] the number of chinchillas per store – in this case, the [bar chart with longer store names] number of chinchillas sold at each of these ten exotic pet store locations.
* 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 bar chart, while chinchillas per store is a column chart. I remember the difference by thinking a column chart has columns aaaand 🤣🤣🤣🤣the other one is a bar chart. Column charts are more common and they're best with fewer than around five categories. Bar charts are 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 here.

Next, take a look at line charts.

* Their purpose is to show trends in a numerical feature over time.
* They're great for any time you want to compare changes over time, such as people bungee jumping each hour, or minute-to-minute stock price.
* Here's a common mistake. I sometimes see time-related data treated as categorical, so using a column chart instead of a line chart. Here's an example with monthly sales over time. There's nothing earth shatteringly wrong here, but as your time series data gets more complex and nuanced, this type of chart will become harder for your audience to interpret correctly.
* Here's the same data as a line chart. The slope of the line between each point emphasizes the rate of change month to month, so you can see HOW sharply sales increased or decreased between each period. Line charts can also show multiple series, or many data points without becoming as overwhelming.
* You may also have seen an area chart, which is a common variation on the line chart. They emphasize the volume of data, not just the trend, especially an accumulation over time. So, total downloads of Beyonce's Lemonade album – that 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 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. You'll see enhancements like color, which can be used to differentiate data points across a categorical variable. Here's an example with virtual meeting length and attention span, but with personal meetings in blue and 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 add quadrants to this plot, I see long meetings with a high attention span, short meetings with a short attention span, and so on. It helps highlight that most meetings are short, with a high attention span.

Finally, let's look at stacked and grouped bar and column charts.

* These are a variation on the vanilla bar and column chart, and their purpose is to compare a numerical feature across multiple categorical features.
* So what are they great for? A stacked chart, which looks like a stack of books, shows a part-to-whole relationship – how much does each product contribute to the total sales in each region, while a grouped chart is better for direct comparison between the categories – which pet store has the most gray chinchillas? In the example with sales by region and product, I can see that in Asia most sales come from Product B, while in Europe, most sales come from Product C. In the Chinchilla graph, I can see that the Elm and 22nd Street location has the most gray chinchillas.

* So these are both absolute charts – right now, you're seeing just sales revenue and the number of chinchillas. A common variation is to have relative charts that show the proportion of each feature combination, rather than the raw number. Here's what that looks like for our data. Relative charts make comparison across groups easier, in particular when the total size of each group is different. So if sales are much smaller in Europe than Asia, but you want to highlight how Product C is selling well there, it might be easier to compare proportions than absolute numbers.
* This is a choice you as the data analyst will need to make – sometimes the magnitude of the impact 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 Product C.

When choosing the right visualization – remember, this is a science, there are right and wrong answers – here are some great questions to ask:
First off, you need to understand your data.
* What types of data are you working with – categorical, numerical, time series?
* How many features are involved?
* And what's the primary outcome of interest?
Then, consider: what's the primary message you want to get across? And who's going to be looking at it? You want to make sure they can easily grasp what you're showing them.
Next, how do your data points relate to each other?
* Are you comparing categories?
* Showing changes over time?
* Displaying the relationship between features?

* Here's a cheat sheet:
  * Time series data often suits line charts
  * Comparisons between categories might use bar or column charts
  * Relationships between two numerical variables could use scatter plots
  * For comparing parts of a whole or multiple categories over time, consider 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
  * 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
  * 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
  * 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
  * The answer is a scatter plot because we want to compare two numerical variables together. However, an important note with scatter plots is not to assume that correlation between two numerical features implies causation. This is just a toy example, but we would not want to conclude there is a direct relationship between these two quantities, unless we performed some more rigorous statistical analysis.

🎙️ 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: Creating Bar & Column Charts in Google Sheets

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 common chart types in Google Sheets.

BUSINESS PROBLEM – INTRODUCE QUESTION

* For this demo video, we are going to work with a sample dataset from Redfin, that summarizes housing sale data at the level of counties and metropolitan areas
* 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

* Let's get started … here I have already imported the Redfin data into a fresh Google Sheet

* Let's explore some of this data
  * They mixed different levels of data (county and metro area)
  * There are different date durations
  *

Some preprocessing –

* Separate the state from the metro area/county

  * Split(Cell, “, “, FALSE - Single delimiter, TRUE - Remove empty spaces)
  * Left(Cell, Number of characters)
* Filter the data so that we can analyze a more consistent subset of data
  * Filter to beginning period (try beginning of Q3) - Home purchases prior to next school year
  * We notice that there still multiple county entries for a given beginning period, why is that?
    * Weird, they include multiple length periods within the data, let’s filter to 12 weeks to capture the full quarter
* Lastly, let’s sort our data in descending order by median home sale price, so that we get the most expensive markets at the top of our data set
* OK, now we’re ready to create the chart

    * Highlight two columns: Region + Median sale price
  * → Cmd/Ctrl + Click
* Insert chart – COLUMN CHART by default
* Summarize the Chart editor
  * If you click off of your chart, don’t fret .. you can just double click your chart to re-open the Chart editor, or select edit chart from the options menu at the top right of the chart
* Oh wow this is unreadable
* Flip to bar chart oh wow that’s so much better

    * Output options
  * Download
  * Copy
  * Put chart in its own tab

    * Analyze the chart
  * “What do you see in the data?”
* Patterns, trends, comparisons

    * Duplicate chart to another tab
* 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. Let’s replace our sales price with the median days on market feature
  * Remember, our data is still sorted by median sale price in descending order, so these are still the top 10 most expensive counties for the selected time period, even though we are now looking at median days on the market
* Now what do you see in the data?

🎙️ TH

OK, there you have it, we've created our 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 in Google Sheets

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 and make it shine! Customizing your charts is like adding the finishing touches to a masterpiece – it's where you bring your data visualization to life.

Let's explore some of the most common ways to customize charts in Google Sheets:

1. Chart Title: Give your chart a clear and descriptive title that summarizes the main message you want to convey. Think of it as the headline of your data story.
2. Axis Labels: Label your axes clearly and concisely, indicating what each axis represents. This helps your audience understand the context of your data.
3. Legend: If your chart has multiple series of data, use a legend to differentiate them. Make sure the legend is clear and easy to read.
4. Colors: Choose colors that are visually appealing and meaningful. Use contrasting colors to highlight differences between categories or groups. Consider accessibility for colorblind viewers by using patterns or different shades of the same color.
5. Gridlines & Ticks: Gridlines and ticks can help your audience read and interpret your chart more easily. You can adjust the thickness and style of the gridlines or ticks to match your overall design.
6. Data Labels: Add labels to your data points to display specific values. This can be particularly useful in bar charts or pie charts.
7. Trendlines: Add trendlines to highlight patterns or relationships in your data. This can be especially helpful in scatter plots or line charts.

Let's customize the [CHART TYPE] chart we created earlier.

Click on the chart to select it. This will open the Chart Editor on the right-hand side.

In the Chart Editor, navigate to the 'Customize' tab. Here, you'll find a wide range of options for customizing your chart:

* Chart style: Choose from different chart styles and backgrounds.
* Chart & axis titles: Edit the chart title, font, size, and position.
* Series: Customize the colors, line styles, and marker styles for each series in your chart.

* Legend: Change the position and style of the legend.
* Horizontal axis: Adjust the scale, labels, and gridlines.
* Vertical axis: Adjust the scale, labels, and gridlines.
* Gridlines and ticks: Customize the appearance of gridlines and ticks.

Experiment with different settings to see how they affect your chart's appearance. Remember, the goal is to create a chart that is both visually appealing and easy to understand.[a][b]
By customizing your charts, you can:
* Enhance clarity: Make your message more impactful by highlighting key insights.
* Improve readability: Make your chart easier to understand and interpret.
* Create visual appeal: Make your chart more engaging and memorable.

🎙️ TH

So, 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.

L2v4 – Demo: Creating Scatter Plots in Google Sheets

Visual

Script

🎙️ TH

Scatter plots are like constellations in the night sky, revealing hidden relationships between two numerical variables. They're your go-to tool for exploring correlations and spotting trends that might not be obvious at first glance by analyzing the raw data.
In a scatter plot, 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 variables relate to each other.

Let’s explore how you can create a scatter plot in Google Sheets using the Redfin home sales data set. Suppose you wanted to understand the relationship between the size of the typical home sold in a given county, and the median sales price. These are both numerical variables, and thus excellent candidates to visualize via a scatter plot.

To create this chart in Google Sheets, let’s start with our Data tab.
* Let’s select the median pending sqft (column AK) and the median sale price (U) that we analyzed with our bar charts in the previous video
* Then, let’s insert a chart. In this case, you can see that again the chart defaulted to a column chart, which we know is not the appropriate type of chart for this data.

* So let’s adjust the chart type in the setup tab of the chart editor. Ah, there we go!
* Let’s move this chart to its own tab, and then work through some customization.

Now, let's customize our chart.

* First, let’s add a title to the chart, and label the horizontal axis. Since median home sales is in the title, I think we can get away with the vertical, or y-axis label.
* Next, let’s take a look at the options in the Series tab. 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 are not too many data points in this chart, because each one represents a county in CA for this particular time period, we can actually make the markers a little bigger, so I’m going to increase the marker size to 10 points.
* We could also check out the data labels, but you can see that if I enable this option, 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. 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 up, let’s take a look at the horizontal axis. The only change I want to make her is to explicitly set the minimum value to 0, so that we can get a little more perspective on our data. By default, the chart auto-zoomed into the range of median square footage that we observed in the data, but sometimes this can create a false understanding of how the two variables relate to each other. Now, 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.
* OK, let’s move along and experiment with gridlines and ticks. As opposed to our previous examples with bar charts, both axes contain numerical data, and are worth tinkering with. As I mentioned already, data labels are a bit too heavy for this visualization, so we may have to rely more on gridlines or ticks in order to help us estimate the axis values for each data point. Let’s configure some gridlines to serve this purpose.
  * For the horizontal axis, I’m going to add minor gridlines, and place configure an increment of 100 sq ft.
  * For the vertical axis, I’m going to enable minor gridlines as well, and use the same count of 4 to create increments of $100,000.
  * 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, we observe a positive correlation between the median square footage of each pending home sale and the median home sale price. As the median home sizesincreases, 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.
* We may actually consider some of these highest priced counties as outliers, which might be worth investigating further. It seems that there's something really different going on in this group of counties with higher prices that seems to defy the laws of housing physics! Some of the most expensive counties do not seem to have the largest homes, so 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.

L2v5 – Demo: Creating Grouped & Stacked Bar Charts in Google Sheets

Visual

Script

🎙️TH

Grouped and stacked bar and column charts are like the dynamic duos of data visualization. They allow you to compare an outcome of interest across different groups, revealing complex relationships and trends in your data.

Grouped Bar/Column Charts:

Imagine you want to compare the sales figures of different products across multiple regions. A grouped bar chart would display each product as a separate bar within each region, allowing you to see how each product performs in different locations side by side.

Stacked Bar/Column Charts:

Now, suppose you want to see how the sales of different products contribute to the total sales in each region. A stacked bar chart would stack the bars for each product within each region, with the height of each stack representing the total sales for that region.

Number of babies grouped by gender and name

* Generate a dataset of a sum of all the babies across

Let's create a grouped bar chart using our [YOUR DATA SET NAME] data.

First, select the relevant columns of data. For this example, we'll use [COLUMN NAMES FOR CATEGORIES] for the x-axis, [COLUMN NAMES FOR SUBCATEGORIES] to group by, and [COLUMN NAME FOR VALUES] for the y-axis.
Next, click on the 'Insert' menu and select 'Chart.' This will open the Chart Editor.
In the Chart Editor, select 'Bar chart' or 'Column chart' from the 'Chart type' menu. Under the "Setup" tab, select "Switch rows / columns" to switch between grouped and stacked.

Now, let's customize our chart. You can add a title, label your axes, change the colors, and adjust other settings to make your chart visually appealing and informative.
For this example, let's add the title "[YOUR CHART TITLE]" and label the axes "[YOUR X-AXIS LABEL]" and "[YOUR Y-AXIS LABEL]."

Now, let's analyze our chart.
What do you see in the data?
[DESCRIBE THE PATTERNS, TRENDS, OR COMPARISONS YOU OBSERVE IN THE CHART]
For example, you might say something like:
* "We can see that [SUBCATEGORY] consistently outperforms [SUBCATEGORY] across all [CATEGORIES]."
* "In [CATEGORY], [SUBCATEGORY] makes up the largest proportion of the total [VALUE]."
* "The [VALUE] for [SUBCATEGORY] has increased/decreased over time in [CATEGORY]."

🎙️TH
Remember, grouped and stacked bar/column charts are powerful tools for showcasing complex relationships between multiple variables. Use them wisely to reveal hidden patterns and tell compelling stories with your data.

L2v6 – Demo: Creating Line Charts[c][d][e][f] in Google Sheets
Visual
Script
🎙️TH

The last technique for analyzing time series data is data visualization. As you have seen in the last couple of videos, a significant part of time series data analysis involves visualizing the data and evaluating the different components. 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.

Line charts are like the superheroes of analyzing time series data. They are defined with the unit of time, for example minutes, hours, or days, on the [CLICK] x-axis, and the measured outcome plotted on the [CLICK] y-axis. They [5 EVENLY SPACED CLICKS DURING THIS SENTENCE] connect each data point with a line, showing how the measurements are changing over time and that they all are connected together in a series. You can compare multiple time series on the same chart by using different colors, such as hourly temperature forecasts for different cities or daily stock market prices for different companies.

SC1

Going back to our spreadsheet on baby names, let's perform one more analytical task before we create the line chart for this data.

Suppose I want to calculate a 10-year moving average, which would represent the average number of female babies named Ruby over a 10-year period. First, I will add a new column to the right of the data and format it similarly to the other columns. I will also apply the same whole number format as I did for the result of my average calculation. Next, I will need to use the AVERAGE function, but instead of selecting the entire column, I will select the 10 previous values and execute the formula.

Remember, we can only start our simple moving average calculation once we have accumulated a full window size worth of observations in the data. So I will start with 1889, which is the first year that has 10 historical observations up to and including that year. You will see that the spreadsheet returns the moving average for each year, once I autofill the formula down the rest of the column. You can validate the formula is working properly by highlighting any 10 consecutive cells and making sure the formula result to the right of the last cell you highlighted matches up with the average that you can see in the lower-right corner of your spreadsheet.

SC2

Now, I am going to create a line chart of the female Ruby data. I am going to select the Count and Year columns, and insert a chart via the Insert menu. In the Chart Editor that appears on the right, I will select Line chart from the Chart type menu, and then configure the axes. Remember, a time series chart has the unit of time on the x-axis and the measured outcome on the y-axis, so I am going to make those changes here. I am also going to add a 2nd series of data, which will be the 10-year moving average of name popularity that I created in the previous step. You will notice that there is a legend created that will help me differentiate between the two time series of data.

Now, I have generated a basic line chart, but I should not stop here. There are a lot of additional features that I could add to the chart. These features are available on the Customize tab, where I can configure the colors on the chart, chart & axis titles, the legend, and more. In this case, I will add a chart title (Ruby Name Popularity - Assigned Female at Birth) and format the text (BOLD and Black). The default settings are sufficient for the rest of the chart.

Now, let's analyze the chart, starting with the original count data plotted in blue. What do you see in the time series data for Ruby? I will summarize my observations based on the components of time series data:
* For the trend, I see things follow the same observations that I have already identified. There was an initial rise in popularity of the name for female babies 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 resurgence in the popularity of the name, but it has not yet reached the peak of popularity that it achieved 100 years ago.
* On seasonality, I do not see a trend at the yearly level that this data was captured. However, that does not necessarily mean that there is no seasonality in this data. Suppose this data was collected monthly instead of yearly, and there was a spike each year in the name's popularity in July, which corresponds to the month associated with the ruby gemstone. If there were such a pattern, then I would say that there is monthly seasonality in the data.
* We can, however, see a clear cyclical pattern in the data, given the initial rise and decline 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, or perhaps the name's popularity will follow a more consistent trend. We'll just have to see...
* Lastly, there is some noise in the data, but overall, I would say that the noise in this data is small relative to the trend and cyclical components, which are very clear. So while there are some minor fluctuations throughout, they do not distract us from observing the overall patterns over time.

Now, let's talk about the moving average. As I just mentioned, this data is not very noisy, so the overall patterns in the time series are clear to the trained observer. But since I created it, let's talk about it 🙂
The first thing you may observe is that the moving average follows a similar trajectory as the original time series data. In addition, you may notice that the moving average is in fact a little smoother, and it also is shifted a bit to the right, indicating a bit of a lag in time. The size of all of these differences are directly related to the window size of the data. The window size effectively determines how reactive your moving window will be to the original data. The smaller the window size, the more similar the moving average will be and the smaller the time lag, and the inverse is true for larger window sizes.

I want to make a couple of final points about these observations:
1. First, the observation about the spike in name popularity in 1963 is not very apparent in this chart, which emphasizes the fact that the analysis performed in the previous video can complement the insights observable in the line chart here.
2. Second, the observations I have shared are specific to this particular subset of data. If you repeat this analysis for a different name and gender, you are likely to observe very different behavior! The analytical

approach will be similar, and there may even be some common insights, but you have to do the work in order to find out. Do not make assumptions that the observations will be the same.

🎙️ TH

OK, that's all for our exploration of time series data analysis in spreadsheets. The techniques that we've covered in this Lesson should help you explore a variety of time series data and answer many different types of questions. You can explore some of these possibilities in the ungraded lab with the U.S. baby names data set.

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%27s_stairs.svg.png and can we emphasize each possible option visually as Sean mentions them? Must keep this image

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] It could be either one. Two reasonable people can come to two completely different conclusions about this same image.

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.

So… how to avoid that? Two things. First, follow a structured and iterative process for creating your visualizations. And second, follow the principles for effective visualization design. Let's see how.

Visualization design process (Initial design → build → evaluate → share → finalize)

The process of creating an effective visualization typically follows these steps: initial design, build, evaluate, share, and finalize.

* [CLICK] First you sketch out an initial design. Often, there are several ways to convey the same information, but one way is likely best.

* [CLICK] Then build the initial draft. As you work, consider how your visualization will be consumed. Will you have a chance to explain it verbally, or will it need to stand alone with just a caption? The latter will need more annotation.
* [CLICK] Next, evaluate whether your visualization effectively conveys the key insights – I'll share some strategies in a moment.
* [CLICK] Then, share your visualization to get colleagues' initial reactions. See whether it's clearly understood.
* [CLICK] Lastly, finalize the visualization. Incorporate insights from the previous steps. Depending on the importance of the visualization and the stakes involved, you might iterate through these steps multiple times.

When evaluating your own visualizations, focus on three key principles: clarity, context, and efficiency. Let's break these down.

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. [CLICK] First, choose the appropriate chart type for your data, which you already know how to do!
2. [CLICK] Second, avoid unnecessary complexity. Use simple, clean designs and avoid clutter. More on this in a moment.
3. [CLICK] Third, use clear labels and titles so your audience understands what they're looking at, plus annotations to highlight key insights.
4. [CLICK] Next, make sure you're consistent with your color schemes, fonts, and scales.

5. [CLICK] Lastly, share your work with others! Your manager, peers, and trusted stakeholders can provide valuable feedback. Remember, you're not the one who will be interpreting it in the end, so getting outside perspectives is crucial.

[CLICK] Here's a visualization of monthly sales. What do you think? Is it clear what's happening? [CLICK] First of all, since this is a pie chart, it's hard to compare the size of the 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. [CLICK] I don't get a clear sense of the trend in sales either.

https://i.imgur.com/2OSUQnJ.png

Now here's another chart of the same data. 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?

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] Data-ink includes [CLICK] bars, [CLICK] markers, [CLICK] the line in a 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. It can be pictures, extra text, too many colors, or something else. Be mindful of any chartjunk that you may be adding to your visualization just to make it more interesting, let the data speak!

Must keep this image
As the airpods chart – https://i.redd.it/yfd30i4rbvjb1.png

Nice airpods chart – https://i.imgur.com/EKQDNqC.png

Here's an example of a chart you saw earlier – about Airpods revenue – with the original chartjunk intact. Ask yourself, [CLICK] what does this picture of airpods in the corner really do? [CLICK] It dilutes the data, and [CLICK] distracts the audience's attention from the core insight about just how much money Airpods make on their own.

[CLICK] You may remember the same data from Lesson 1, presented in this next chart. [CLICK] Extraneous elements – like the airpods picture and logos – have been subtracted out, while [CLICK] the airpods revenue has been highlighted with color.

Context is about grounding your audience's understanding. Consider what background knowledge your audience has about this data. Context can look like
* [CLICK] Creating a clear narrative structure – storytelling!

* [CLICK] Providing relevant background information
* [CLICK] Comparing the insight with familiar concepts
* [CLICK] Defining any jargon
* [CLICK] Explaining 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

A well-designed visualization that tells a clear story will have lasting impact. You know you've done a good job when others start using your visualization in their own work.

Remember that your goal is to create a visualization where everyone who looks at it comes away with the same, correct interpretation. That's the power of effective data visualization.

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 – Creating effective data visualizations: data encoding

Visual

Script

🎙️TH

You know the strategy for effective data visualization design – clarity, efficiency, and context – but how do you actually create a chart that follows them?

Here's the basic process: 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 for creating effective data visualizations.

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. The next video will focus on chart elements

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 easy to read, labeled, and intuitive. [CLICK] Strongly consider including zero for numerical features.
* [CLICK] Scale your axes appropriately. Avoid exaggerating or compressing the data, which can 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 [CLICK] most powerful tools for creating clarity and context. For instance, [CLICK] you can use color to highlight key insights, like your company's performance compared to competitors.[CLICK] Or you can use color to 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. [CLICK] About 4.5% of people worldwide have some form of colorblindness, mostly men. [CLICK] When possible, combine color with another element like markers or labels. Additional clarity helps everyone.

Can use these graphs or similar looking ones with same subject

3 dimensions – https://i.imgur.com/r3guFdx.png

3 plots with 2 dimensions – https://i.imgur.com/ly4JbYg.png

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 next to each other.

Here's an example. [CLICK] Say you're trying to create a chart to show the number of birds you observe each day in your backyard based on the temperature. There are three species of birds you typically track. So your data has three dimensions: [CLICK] temperature, [CLICK] number of birds observed, and [CLICK] species.

[CLICK] Let's see all those dimensions plotted on the same chart: temperature on the x axis, number of birds on the y, and the bird species in different colors.

It's clear these three species have different habits based on the temperature – [CLICK] Bird 1 prefers moderate temperatures, [CLICK] Bird 2 prefers higher temperatures, and [CLICK] Bird 3 prefers the cold. [CLICK] I find myself to be most similar to Bird 2. [CLICK] But there are a lot of observations on this graph and it's hard to tell where exactly they overlap. [CLICK] In particular, observations with less than 5 birds are muddled.

One option to improve clarity is to separate this data into three scatterplots each with a single species. The patterns for the individual species become clearer, while still allowing your audience to compare habits across species.

3 dimensions in color – (source)
3 dimensions with markers(source)

The remaining elements you'll see in this video should be used judiciously, starting with markers.

Markers are a data encoding element typically used in scatterplots to add a third dimension to the data.

[CLICK] You saw the bird scatterplot a moment ago with color differentiating the three series. [CLICK] Here's that same data but using markers rather than color to differentiate. 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. [CLICK] Consider 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. [CLICK] They 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 GDP 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] [CLICK] They're here, the two largest bubbles! [CLICK]

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!

Earlier you saw how using both color and labels or markers can help individuals with colorblindness interpret your chart. This strategy of using multiple visual cues to reinforce the same information is called double encoding. It can greatly enhance clarity.

Remember, efficiency is key. Don't overdo it with visual elements. Each addition should serve a clear purpose in enhancing understanding.

🎙️ TH

Now that you're familiar with using data encoding and double encoding in your charts, join me in the next video to explore how additional chart elements can make your insights even clearer. See you there!

L3v3 – Creating effective data visualizations: chart elements

Visual

Script

🎙️ TH

Let's talk chart elements. These are the parts of the chart that don't directly represent the data.

https://upload.wikimedia.org/wikipedia/commons/c/c6/Internet_host_count_1988-2012_log_scale.png – logarithmic graph

Take a look at this graph of internet hosts, which means the number of devices connected to the internet. It's clear that internet usage skyrocketed between the start and end years. [CLICK] But the 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] A logarithmic scale changes the distances between values on the y-axis. [CLICK] On a linear scale, values are evenly spaced. Think about plotting 10, 100, and 1000. On a linear scale, the second distance is ten times larger than the first. [CLICK] A logarithmic scale transforms these values, making 10, 100, and 1000 evenly spaced.

The logarithmic scale [CLICK] spreads out smaller values and [CLICK] compresses larger ones, [CLICK] making patterns across the entire range more visible.

Linear graph – https://i.imgur.com/9SDetRB.png

spread out clustered data points

Here's the same graph of internet hosts, but with a logarithmic scale this time. [CLICK] It's much clearer that there was substantial growth between 1981 and 1997.

[CLICK] Consider a log scale when [CLICK] you want to cover a large range of data, [CLICK] emphasize proportional changes over absolute values, or [CLICK] spread out clustered data points so they can be seen better.

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/

While I am a big fan of log scale, you have to be careful about how well the audience can interpret what they're seeing. People's brains don't naturally think logarithmically. Consider whether the value to your story is worth the added complexity.

Here's an example of two graphs of COVID19 deaths over time. [CLICK] The graph on the left uses a linear scale, [CLICK] while the graph on the right uses a logarithmic one. Researchers studied regular people's interpretations of these graphs. Which do you think looks more scary? [CLICK] They found that people who were asked to interpret the graph on the right had significantly lower comprehension when asked to compare weeks and predict the number of deaths in the following week.

[CLICK] So, a good rule of thumb: if your data is bunched up towards one end 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 scale, consider including zero. In most cases, including zero helps communicate the magnitude of your data, while excluding it can mislead your audience into thinking data is much smaller or larger than it is.

Leaving out zero is a common tactic in misleading graphs. Here's a graph of pretzel sales in 1967 it looks like Salty Serpentine is really beating out the Twist and Shout brand. But look – the scale starts at 950. It exaggerates the differences, since the green bars seem twice or three times higher than the purple ones.

Here’s the same data with zero included in the scale, where it looks like these two brands are pretty competitive.

Next up, annotations. These are a fantastic tool for emphasizing your story and guiding your audience's attention. Without annotations, people's eyes might wander all over your chart. With well-placed annotations, you can lock in their focus on the most important elements.

Remember, efficiency is key. Don't overwhelm your chart with annotations. Choose just one to three key points to highlight. You should also consider how your audience will encounter the chart. If you're presenting in person, you can use a laser pointer for additional callouts, so fewer annotations may be needed. If your chart will be viewed independently, such as in a report, 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 "Crime Data in Los Angeles", consider something like "Crime Increasing in Los Angeles This Year". This immediately draws attention to your main point and helps prevent misinterpretation. Titles can also provide crucial context, like the time period your data covers. This 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 data visualizations! You won’t look at a graphic in the news the same way again, trust me. See you there!

L3V4 – Data visualization examples: 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.

* So right off, we're looking at a graph showing the market share of U.S. carbonated soft drinks. Pretty straightforward title there.
* On the x-axis, we've got time running from 2000 to somewhere around 2020. It's giving us a good two-decade span to see how things have changed.
* The 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 Coke in black, Pepsi in blue, Dr Pepper in that dark pink, Sprite in light green, and 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 the note at the bottom telling us it's based on volume of case sales is helpful. Gives us context on what we're actually measuring here.
* Looking at the big picture, we can see some interesting trends. Coke's been dominating the whole time, but it had a dip in the middle there. Pepsi's been on a steady decline. Dr Pepper and Sprite have been slowly climbing up, while Diet Coke had a bump but then started falling again. One key insight – which I can barely make out – is that Dr Pepper appears to be overtaking Pepsi, especially when you consider the trends of each soda.
* To improve this chart, you could 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.
* Adding some gridlines wouldn't hurt, so the audience can more easily compare the sales across the different brands.

https://www.bls.gov/charts/american-time-use/activity-by-age.htm

Alright, let's check out this graph.

* 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 set 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.
* 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.
* Now, encoded categories. Each age group has distinct colors, which helps the audience compare the values easily. However, even distinct colors likely wouldn’t make comparing across 8 categories easy, especially since these colors don’t have a logical correspondence. There’s no reason why 35 should be light blue. However, being able to select only three age groups, as I did here, helps make the graph easy to interpret.

* 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 we wanted to improve this chart, we might consider grouping 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. Pretty fascinating stuff when you dig into it.

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.

* So, the title:"Access to nature where you live". It's relatively clear but doesn’t share the key insight.
* On the x-axis, we're looking at population density, measured in people per square mile. It runs from 0 to 30,000, giving us a wide range to work with.
* The y-axis shows us the NatureScore, which goes from 0 to 100. I'm assuming that higher scores mean better access to nature.
* For the chart type, we're dealing with a scatterplot here. It's a solid choice when you want to show the relationship between two numerical variables, 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 NatureScore using a diverging color scale. Greener means higher score, purple means lower, and 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 works is as intuitive on the lower scale.
* The size of the dots is another encoded variable - representing population. Bigger cities get bigger dots, which makes sense. 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 bigger cities and some outliers like Union City, N.J. and Suffolk, Va. It's not cluttered, but gives us some reference points, especially for more populous cities where the audience may live or be able to picture those cities.
* Looking at the big picture, we can see a pretty clear relationship here. 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.
* One key insight is how some cities buck the trend. 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, you might consider adding a trendline to make the overall relationship even clearer. 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 "Cities with lower population density offer better access to nature"? Also, 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 – there are some beautiful and fascinating charts out there. Don’t just stop once you’ve displayed the data. Think about how you can make every one of 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 and for reflecting on your personal progress.

Join me in the next lesson to learn how to leverage LLMs to assist you with data visualization. I’ll see you there!

Lesson 4 – Data visualization with LLMs
L4V1 – Interpreting data visualizations with LLMs
Visual

Script

🎙️ TH

Link to initial cut & screencast for timestamping

So earlier in this course, I kind of lied to you. I told you that Claude 3.5 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 multiple modalities of inputs, so 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.

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. So it's a pretty simple prompt.

So 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. So I'm going to click on this image so that we can see a little better, and we saw that population density increases to [00:02:00] the right, and you can kind of see that 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. Let's see also 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. So 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 [00:03:00] 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. So 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 for the points.[00:04:00]

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 really 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.
* [00:05:00] And the gridlines are quite prominent. Okay, I actually think that the gridlines help organize the data really well.

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. We saw in the previous

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 [00:06:00] 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.

Between these two use cases, it seems like our generative AI tools like Claude and ChatGPT are going to be more useful for interpreting data visualizations and maybe less useful, but still a good second pair of eyes, on our own data visualizations to spot possible areas for improvement. So don't hesitate to take a screenshot, throw it into Claude, throw it into ChatGPT, see what kind of insights they can help you out with and if they can help you improve your visualization, even if it's just a tiny increase.

L4V2 – Creating data visualizations with LLMs
Visual
Script
🎙️ TH

Link to initial cut & screencast for timestamping
[00:00:00] Let's see how a generative AI tool like Claude or ChatGPT can help create data visualizations. Here I have a data set of online course engagement with a few different features, including the user ID, the course category,
the time the user spent on the course, some other features, and then all [00:00:20] the way on the right, this is the outcome we're interested in: whether or not the learner completed the course. So each row here is one observation of one person taking one course
This data set is publicly available and there doesn't appear to be any identifying information, so I'm [00:00:40] not too concerned about uploading this to a large language model.
For example, I can just remove the user ID column to further anonymize this, since I don't really need that information. But if you have any kind of sensitive information, you shouldn't be uploading it to a generative AI tool.
Let's start with Claude. I'm going to grab my data set and add it [00:01:00] to the conversation. And right away you can see that this file is much too large for Claude to handle. So I have 9,000 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, let's just take the first five rows, copy those, [00:01:20] and and I'm going to tell Claude here is some data about online course

completions. If I paste this in directly, it pastes in awkwardly with all the data squished together. If you use command shift v or control shift v, that helps with formatting.
Let's say that I want to visualize [00:01:40] the course completion based on the category of the course. I'll just go ahead and ask Claude what would be a good visualization for that. Now it's going to try and create a bar chart, but you already know that that's not going to be very helpful because it only has a couple of observations here.
But it suggests using a bar chart. So [00:02:00] let's go ahead and create a bar chart of course category and course completion. You can insert a chart like so.
Okay, so this chart looks pretty strange. What I was hoping to see was just each category on the x axis and then on the y axis the number of completed courses. [00:02:20] So let's go back to Claude and ask it: I made a chart, it looks weird, how can I aggregate each row into just one column for each course type? Okay, and it's going to suggest creating a pivot table to aggregate the data. For rows, course category. For values, completion rate. And then to average the completion rate. I notice it suggests completion rate, but I actually want the course completion feature. I'm going to delete this [00:02:40] chart, and let's go ahead and do that quickly.
For rows, we want course category. For values, we want course completion. And the sum is helpful, but we may want to go ahead and do average,
So that we can compare those values no matter how many observations we had for each one. And then let's go ahead and [00:03:00] create a chart for that.
So we can see that overall the course completion rate is pretty similar. Health seems to be the lowest; whether or not that's a meaningful difference is yet to be determined. And maybe arts and programming are completed a bit more often.
Let's say instead of doing it myself, I actually want the computer to create a visualization for [00:03:20] me. I'm going to use a different tool: ChatGPT, in this case 4o, because, as you saw, at the moment Claude can run its own code. However, it cannot process the entire dataset.
So let's go over to ChatGPT. I'm going to upload my entire dataset. And ChatGPT is completely fine with that. And I'm just going to [00:03:40] say, Help me visualize course completion by course category.
It gives me this nice visualization, which is very, very similar to the one that we just created, but it looks like it ordered them based on the number of completions from most to fewest. Okay, so now I can ask it to modify this. Let's change that to percent completed. Remember, [00:04:00] if I click on analyzing, I can see all the code that it's running. I don't necessarily need to see that, but it can be helpful. We see that the percentages are really, really similar to the previous graph.
That makes sense because we have a very similar number of observations in each of these categories.
Let's say I wanted to look [00:04:20] at the relationship between quiz score and completion rate. So, did people who have higher quiz scores tend to complete more of the course? Let's take a look. So, it is telling me that a scatter plot is a good choice because this is a relationship between two continuous numerical variables. All right, this is just a [00:04:40] mass of data and it looks like there's absolutely no relationship. Does that seem strange to you? Seems a little strange to me. That leads me to believe maybe this is artificially generated data. Hard to say.
Let's try two more visualizations here. Let's create a grouped bar chart of the time spent on the course, as well as [00:05:00] the course category, by device type. So, in this case, 0 is desktop and 1 is mobile. So the

question is, do people taking different types of courses spend different amounts of time on those courses, and does that change based on the device type they're using? If I had to guess, I would say people on mobile [00:05:20] probably spend less time.

So let's take a look. Okay, it's a little challenging to view these colors, if I click this little button right here, it will change to a chart with a white background that's not interactive. Looks like the average time spent is very, very similar across all these categories. Again, that's a little odd. You might expect the behavior of desktop [00:05:40] and mobile to be distinct.

Let's ask one more question. I've already taken a look at this data, and I happen to know that whether or not someone completed each course is related to the average grade they received on all the quizzes. So I'm going to ask ChatGPT, how can I visualize that, since I have a continuous numerical variable and a [00:06:00] categorical variable. And it's gonna suggest a box plot.

You may not have seen this type of plot before. The range of values is between 50 and 100, but it looks like for people who did not complete the course, their median quiz score, which is this red line, appears to be a little bit below 70, whereas for people who did complete the course, their [00:06:20] median appears to be a little bit above 80. So that could potentially be something of interest.

And it gives me a violin plot as well; this is an alternate visualization that's similar to a boxplot. If you're not 100 percent sure what's going on here, the model will attempt to explain what's happening.

So generative AI [00:06:40] tools can be a really helpful part of your workflow when you're trying to visualize data. They can both help you enhance your own visualizations as well as create new visualizations for you. Maybe this isn't the kind of quality I would put in a report, but this is great for exploring the data.

Great work! You're almost to the end of this module, and to the end of this course. I hope you continue on to the lab and capstone project for this course, which tests your data analytics skills on a really cool data set of customer churn for a telecom company. I know you'll enjoy testing your analysis and visualization skills to help the company understand the factors that lead a person to quit being a customer!

You've made the first crucial step in your data analytics journey and I'm excited to see what you accomplish next. Once you've completed the lab and capstone exercise, I hope you'll join me in the next course to expand your statistics and spreadsheets skills for deeper and more meaningful analysis. Thanks for sticking with me and I'll see you there.