Google Cloud

Data Visualization

Evan Jones

Now, one of the key outputs and deliverables that data analysts create, are those insightful reports you present to your audience.

Data visualization

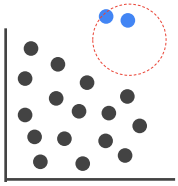
01 Overview of data visualization principles

02 Exploratory vs explanatory analysis approaches

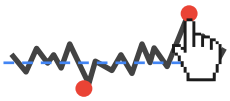
03 Introduction to Google Data Studio

In this data visualization module, we'll cover a little visualization theory and best practices, and then we'll introduce Google Data Studio, as one of the visualization tools in your toolkit for creating those actionable reports.

Use visualization to clearly and concisely present insights



Visualizing a dataset allows you to spot hidden trends



Interacting with a dataset visually is often faster than writing SQL

Deliver powerful insights to your audience through reports

Get scalable performance as your dataset grows with BigQuery-backed visualization tools

Now, visualizing data is another key skill that's part of the data analyst's skill set. And the two major reasons why visualized data is so important; one is probably pretty obvious, building those interactive deliverables and charts and graphs for your audiences, is just a great tool to be able to display those insights. But often the first one that comes to mind for me as a data analyst, is if I don't know those insights off the bat, and I want to explore my data sets, a lot of times the visualization tool if you're a more visual thinker or learner can help uncover those insights a little bit more clearly than just poking around at your data set with SQL. And clearly use case resource before was in the previous course looking at Dataprep. When you're in the transformer and you see those histograms at the top, for the frequency of data values in that particular column, you get a very clear visual picture of a lot of that underlying data. What are the key trends? What are those anomalies that you see just by seeing those bar charts visually? So visualization can do a lot of things.

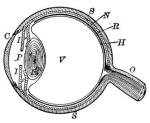
So four quick things that we are going to highlight:

- You can spot those hidden trends.
- You can interact with the dataset, so you can tell a story that has multiple pieces. So, say you start with a time series data, and then you just click into one of the anomalous spikes or troughs in the data and you drill down into those details very visually. So it's a natural tool to let your audiences follow that cohesive story throughout the flow of your explanation.
- And of course, it can be very visually aesthetically pleasing as well, building those dashboards and conveying those insights in a very fast and effective manner.

- And lastly, if a lot of your data is already in BigQuery, putting a visualization tool on top of it like, Google Data Studio will naturally get all the performance benefits of having your data and your queries processed in BigQuery, and then displayed and rendered quickly on the front end in a visualization tool.

Visualization theory: Perception

Stimulus



Eye

Brain

Preattentive processing
Attentive processing

Sensation
Physical process

Perception
Cognitive process

So no lecture on data visualization will be complete without talking a little bit about visualization theory. So as I mentioned before, visualization is both an art and a science. So here we're going to get a little bit of the science of what our brains perceive when you look at a beautiful visualization. In this particular case, we have a stimulus where we have a cat. Your eye Immediately recognizes it and says, "Hey, I've seen that before, that's a cute little kitten." And your brain automatically says, "Hey, I've seen a thousand of these before, I immediately know without thinking that it's a cat."

Now when you actually get into machine learning, it's actually a little bit harder for machines to have that intuition. Intuition is extremely hard to build in for a computer. Whereas as humans, we're built with a lot of that, what we call pre-attentive processing, where we can immediately recognize things.

Now, what does that mean for your data visualizations? It means you can effectively cheat the brain by using common human intuition to not have the brain do a lot of work. As humans we have evolved to not do a lot of mental processing, we want to make those snap judgments very quickly, and then only tap into that really focused thought power when it's required. So let's take a look at an example together. All right.

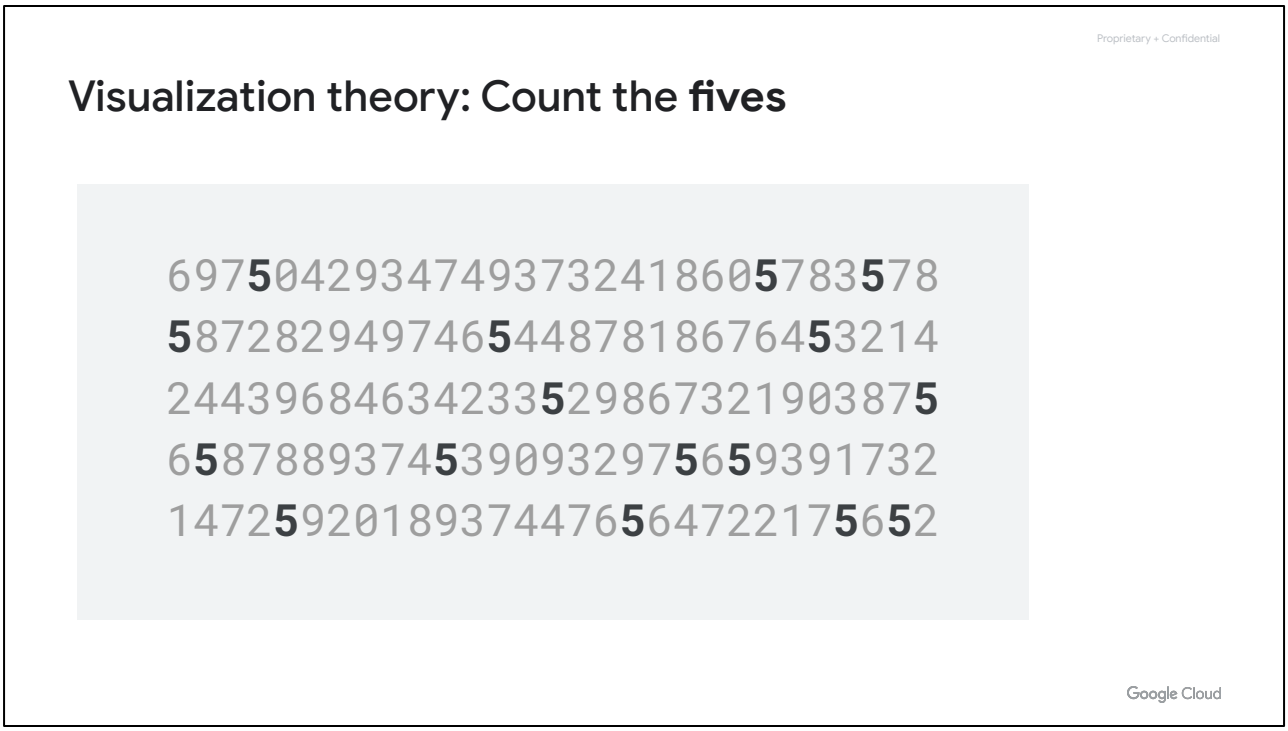
Visualization theory: Count the fives

```
69750429347493732418605783578
58728294974654487818676453214
24439684634233529867321903875
65878893745390932975659391732
14725920189374476564722175652
```


So take a few seconds and count all the fives that are present. This is probably one of the hardest exercises that you can do as part of the specialization.

You're looking at it? All right. Did you get them, or are you still looking? Now, if you counted 16 fives in the 10 seconds that I gave you that's absolutely amazing, must be like a speed reading level. For the majority of us myself included, It's very, very hard to pick out from this noise of numbers here, this very crowded visual what those fives are.

Now, the metal point here is that, this super focused processing that our brains have to do. For majority of us, we we were reading from left to right and serially scanning every single row and counting out all those fives and keeping track of them. Now, there has to be an easier way to do that and naturally you might expect we could do something like this.



Now count the fives.

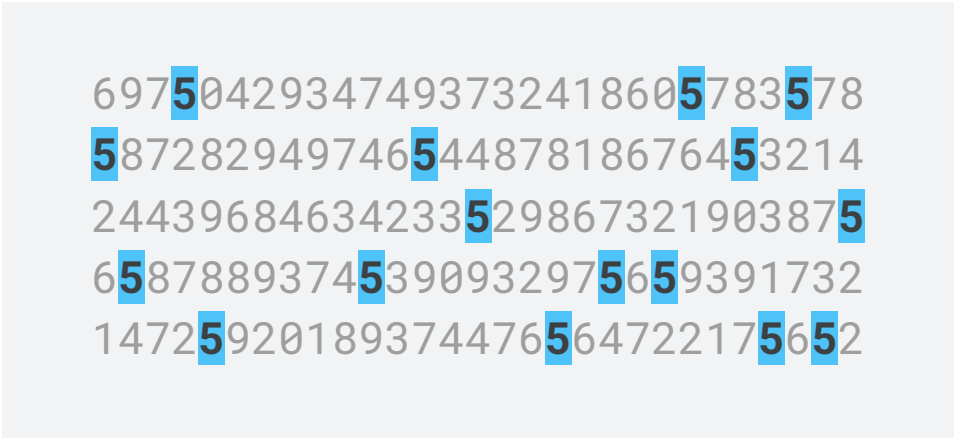
How about this time around, was it much easier? Think about what your brain actually did to count those 16 and why it was much faster for you.

So on the surface you might say, well, you bolded the fives and that made them a lot easier to see. Well, that's absolutely true. Let's think of the theory behind that.

So, in my mind, bolding enabled me to kind of visually cluster those fives into small different chunks and quickly count them out and then pick them. And the two things that stand out from a visualization theory perspective is, when you contrast certain elements, you highlight the focus, and it allows us to treat all the other numbers that aren't fives as background noise that we can safely ignore.

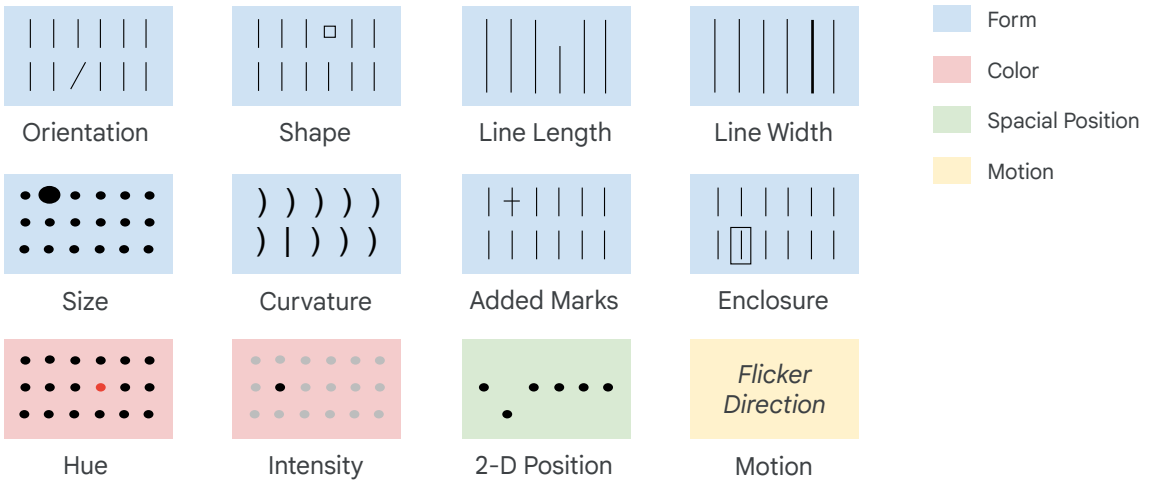
So in essence, it saves your brain's time by applying a prioritization saying, "This is what's important, and this is what you should focus on." And immediately our brains can jump at the task and say, "Bam! Bam! Bam! These are all the elements that I actually care about, let me ignore 80 percent of the rest as visual data.

Visualization theory: Count the fives



We can take this yet a step further and continuously add more, what we call visual encoding on these particular elements that we want to add focus. So here we introduce an element of color where you can further highlight the elements that you want folks to focus on, and it gives it greater attention to what you're calling out as those key points.

Visualization theory: Preattentive attributes



Now, there's a variety of methods that you can use to "cheat" the brain, and then really tap into that fast processing time, that instant decision making that we as humans have developed over the years.

Now, you can mess around with things like, the orientation, the shape and the skew, the length of a certain attribute, the size of the mark on the page. Maybe some elements are curved and some aren't. You're adding things like a box around it, changing the intensity or the hue, moving around in the positioning or even adding things like a motion element to it as well.

So, all of these will help the brain focus in on what's important. And then leveraging these and potentially in combination and in concert with each other, will really help make your visualization stand out and ultimately convey that message very quickly, without your audience having to stare at your screen and really focus hard.

You choose: Effective or ineffective (or wrong) visuals

Okay, now we're going to play a little bit of a game. So, it's the good and the bad and the ugly when it comes to data visualization. There's many different ways to visualize the same data as you're going to see when we jump into our Google Data Studio exercise as many different chart options are available to you.

So, here's an example and I want you to think which visual presents this data set better. And the data set is Game of Thrones characters by gender. So, take a look.

Most of you would probably say the one on the left.

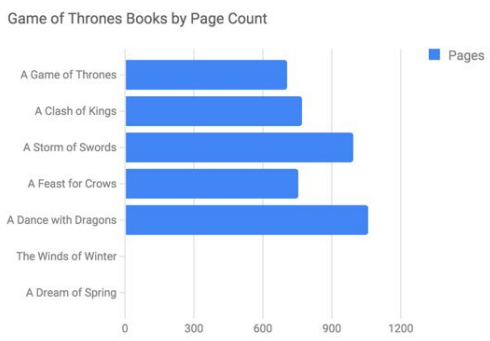
So, the one on the right, using something like a time series for this discrete analysis, generally is not the best thing that we want to do. I would actually never do that. And on the left, you just see a much clearer picture of the division between the count of male characters and count of female characters. I think of what are some additional ways you could add other than encoding attributes that could be beneficial here.

So, one of the things that you could say that the graph in the right actually does give you is the actual numeric count of those characters. Whereas, the graph on the left does not have any kind of labeling or any other type of measure documentation that we can see what the actual values are.

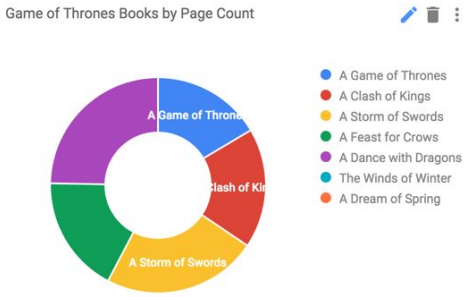
So, both of these graphs could definitely be improved but I would lean towards the one on the left and then add in something like a percentage or other data label that you conclude. Again, this is a continuation of something like the count-to-fives

exercise where we added that blue coloring. Adding additional encoding measures like data labels, a color, a good use of white space can continuously help make your message clear. But at the same time, you don't want to overload the audiences' eye with too much going on on the screen. Let's try another one.

You choose: Effective or ineffective visuals



OR

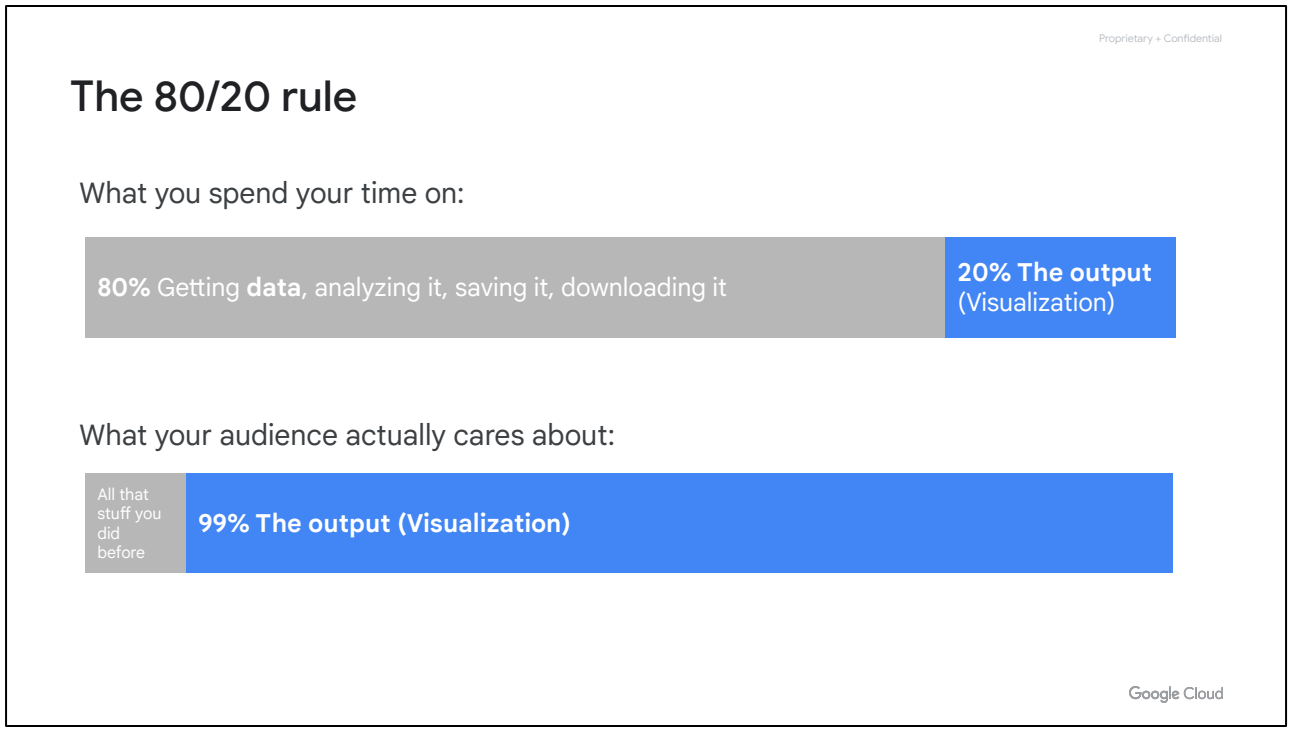


Okay, now we have the same doughnut chart, but it's going to be on the right. And then we have a horizontal bar chart on the left. And then this is the dataset of Game of Thrones books by page count. Take a look.

Okay, so let's take a look at the one on the right. So, this doughnut chart has labels this time and a little bit hard to read because you have that white on a potentially white background there. And then you have a lot of these different books here, some of which are not showing like that teal color, The Winds of Winter, so because you can't actually display potentially negative values or zero values on a doughnut chart. That doesn't exist in that space.

Whereas, you contrast that to the bar chart on the left, you can see the last two books: The Winds of Winter, and A Dream of Spring. Why would those have zero values for the page count? And you probably have to be a Game of Thrones fan to know this, but those books have not yet been written as of late 2017 when this was recorded.

And at the end of day, you have to think of what is the ultimate message that you want to convey. If it's comparing page count between the books, maybe something like a bar graph with an additional label saying how many pages are in the actual book.



Now, this is one of my favorite slides when it comes to talking about visualization theory. We tell you the benefits of cleaning up your datasets, writing beautiful queries and ultimately, finding those insights. In large and part, if we had to do a pie chart on how you actually spend your time, maybe about 80 percent of that pie chart is going to be filled with doing data analysis and getting to that point of those insights, and maybe 20 percent of your time is actually going to be spent on building up those beautiful visualizations and sharing them with your peers.

But, as we all know, a beautiful picture or visualization of those insights is large in part what the audience is actually going to care about. How you're delivering that message is just as important if not more important than the actual methods that you did to actually get those insights in the first place. Because at the end of the day, even though we both write beautiful SQL statements, unless you're presenting it to a peer who also wants to review the code quality that you have, a lot of times the executives or the other members in your organization are just going to see those visuals. So, put a lot of good conscious thought in how those visuals are being created and how that message is being conveyed.

Visualization core concept: Dimensions and measures

	Description	Examples
1. **Dimensions**	Independent variable Categorical information	• Name • Location • Part number # • Job title
2. **Measures**	Dependent variable Any field containing quantitative information	• Revenue • Salary • Expenses • Count of errors

One of the core concepts that we're going to dive into, and this is directly related to any visualization tool that you play around with, is called a dimensions versus measures. So, you can think of the data fields and your datasets as ingredients that we can potentially cook with to build and bake these beautiful visualizations.

So, a dimension is a field that you can classify, that is an independent variable. So normally, I like to think of that as your qualitative or categorical information. So, if you had your IRS 990, this could be like the name of the charity, their address, anything that's not necessarily mathematical, …

… whereas your measure, is going to be dependent. So, that's going to be generally quantitative information. So, things like revenue metrics, expenses, salaries counts, sums. And that doesn't mean that one particular field can't serve as a potential dimension and measure, for example if you had the employer identification number, what would that be? Would that be a dimension or measure? If you said dimension, you're absolutely correct. Just because it's an integer, or a number, doesn't necessarily mean it's a measure, because that's a qualitative information. So, that would actually classify itself as a dimension. But if you had the count of all those EIN numbers, you could then treat that dimension as a measure, as that count would then be a measure.

Class question

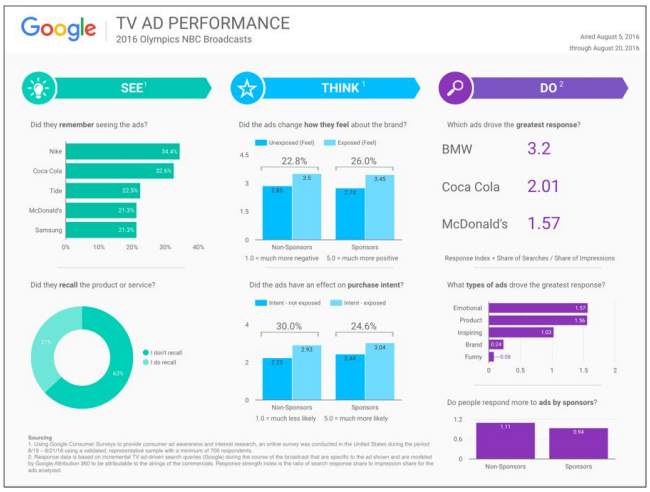
Which of the below are measures?

1. Phone number
2. Employee ID
3. Age
4. Date of birth
5. Tenure at work (in years)
6. Job title

Remember, measures are usually quantitative fields

Which of the below are measures? Some measures again are those quantitative fields. Which ones can you perform math on? Well, let's take it from top to bottom. Number one, phone Number, no, that is definitely dimension. You cannot perform math in a phone number. I mean you could, definitely shouldn't. Employee ID, no, not technically, but again, if you're doing things like counts of employees, you can apply functions on top of your dimensions to treat them as measures. Age, absolutely. You could do average age, you could do mean, max. Date of birth, I wouldn't really consider that as a measure. Tenure at Work (in years), definitely, that's something that would make sense to do math on, and Job Title, I would not consider that as a measure, because again that's qualitative information.

Reports transform data into information



Tell a clear story with your data

Share and collaborate on reports with others

Let's talk a little bit more about one of the chief deliverables of what you're actually going to be producing which is lo and behold, the report.

And the report is a canvas for you to clearly tell whatever message or story about the insights that you've gathered, and tell it in a clear, and effective, and well laid out manner, so that your audience gets it very quickly.

And again, think back to the count of fives example, you want them to immediately look maybe within the first 10,15 seconds, and hone in on those really key insights that they the audience care about.

And of course, you can share and collaborate these reports with your peers as you're developing them. And this is an area to process. Just because you create a beautiful visualization, and you launch it out into the world, doesn't necessarily mean that that's never going to change in the future, with new data or new feedback from your audience members.

Data visualization

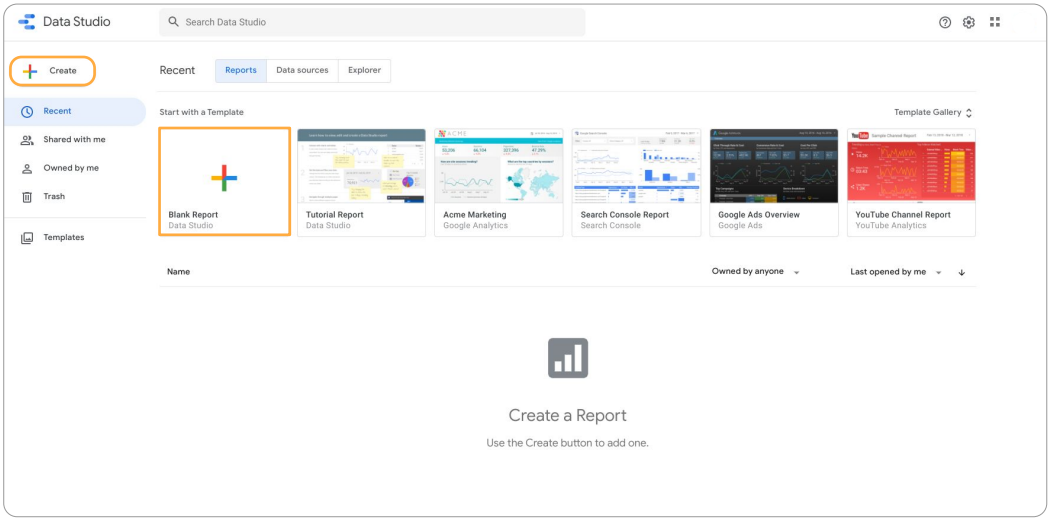
01 Overview of data visualization principles

02 Exploratory vs explanatory analysis approaches

03 Introduction to Google Data Studio

Now it's time to take a look at the Google Data Studio UI,
and see one of the ways that you can visualize all these insights that we've been talking about.

Create new reports in the Data Studio UI



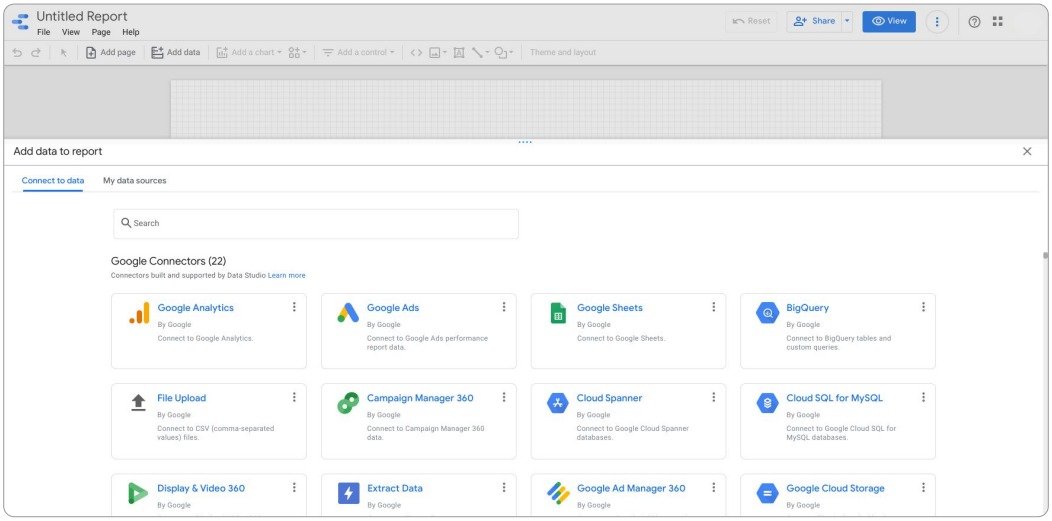
What you are looking at here is the Data Studio home screen.

Here is where you can see the reports you created previously, you can create new reports, or copy existing reports and use them as a starting source.

There are two ways to create a new report from scratch:

- Select **Blank Report** in the templates panel in the middle of the screen.
- Or click the **Create** button in the navigation pane on the left of the screen.

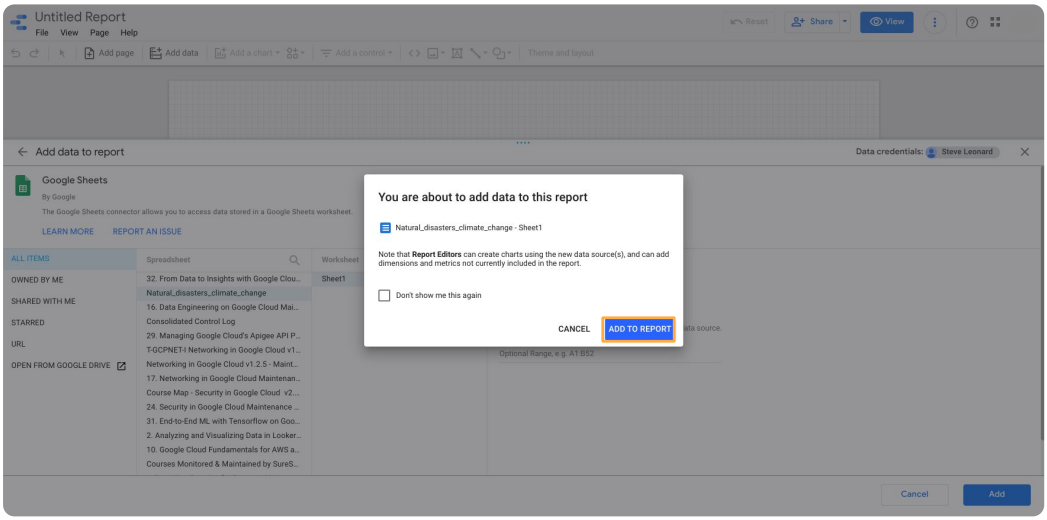
Connect to multiple different types of data sources



Note that you can have any or all of these data sources in a single Data Studio report.

In addition to the Google Connectors, there is an increasing list of Partner Connectors to choose from as well.

Add the data source to your report



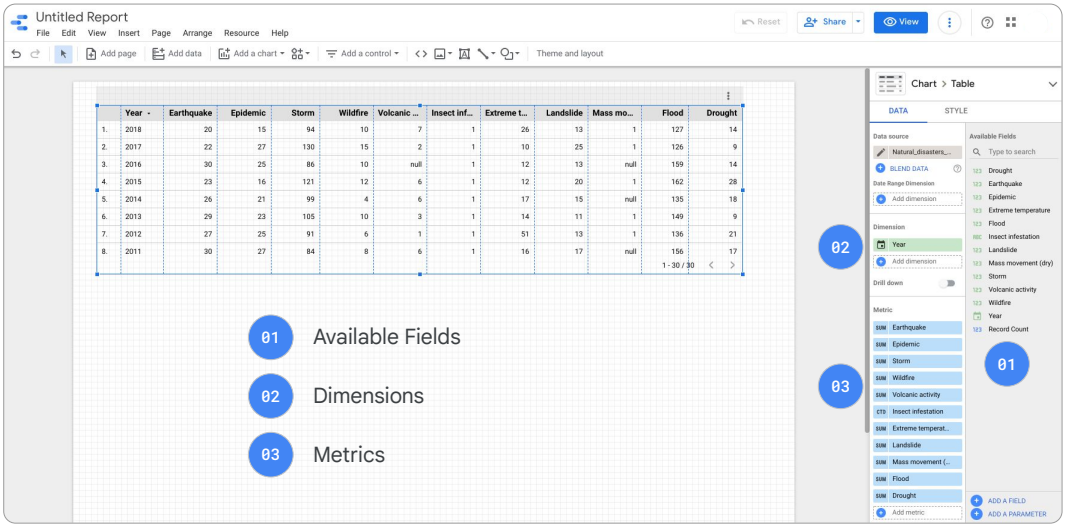
Since Data Studio reports can be shared, you should be aware of the ramifications of adding a data source.

When you add a data source to a report, other people who can view the report can potentially see all the data in that data source.
And anyone who can edit the report can use all the fields from any added data sources to create new charts with them.

We'll talk about how to control access to data and sharing later on in this course.

Click Add to report.

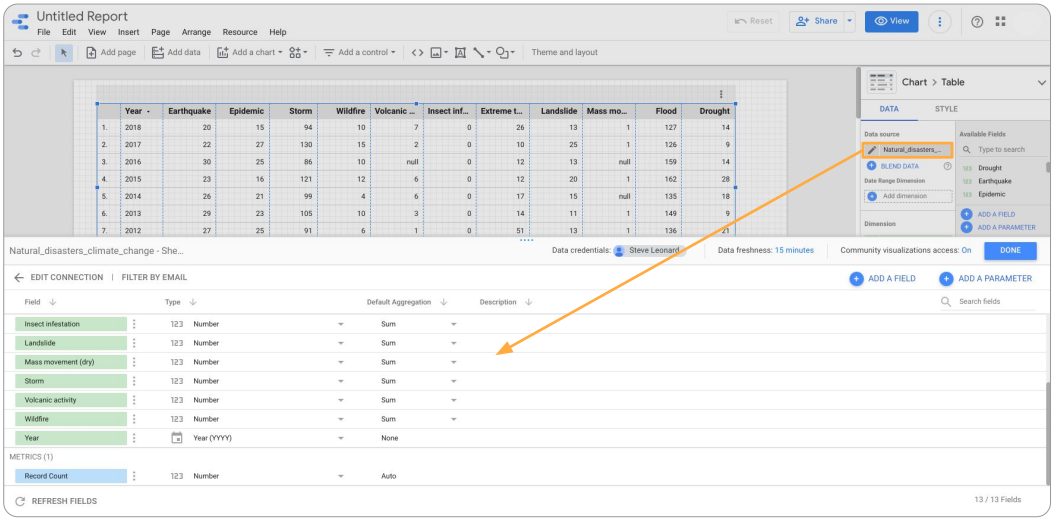
Select your data fields to build your visualizations



Having selected a dataset, you can specify what elements of the dataset you wish to visualize.

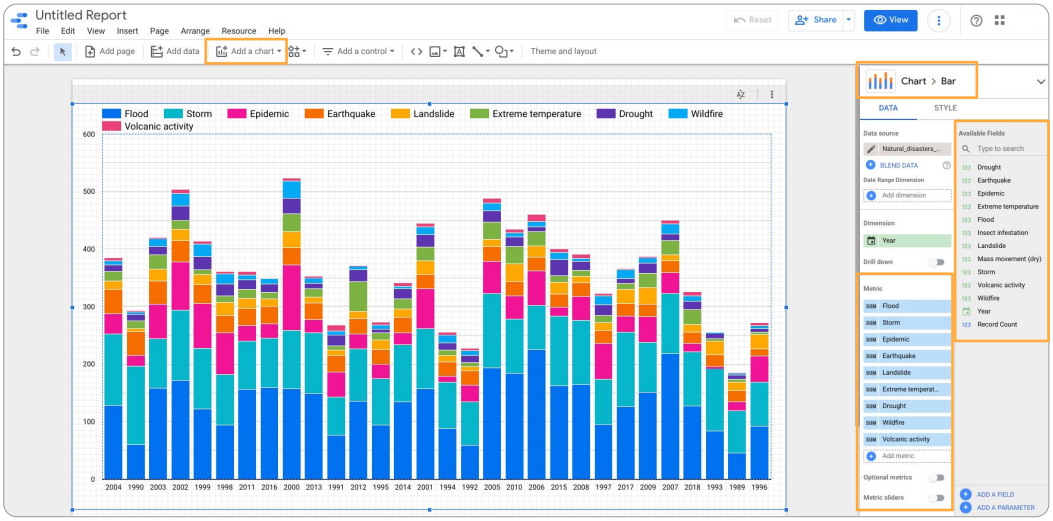
This includes selecting the **Dimensions** and **Metrics** that you want to use from the **Available Fields** of your dataset.

Edit your data source fields, if necessary



The **Edit data source picker** in front of the **Data source** name can be selected to edit the dataset fields.

Create charts to visualize data relationships



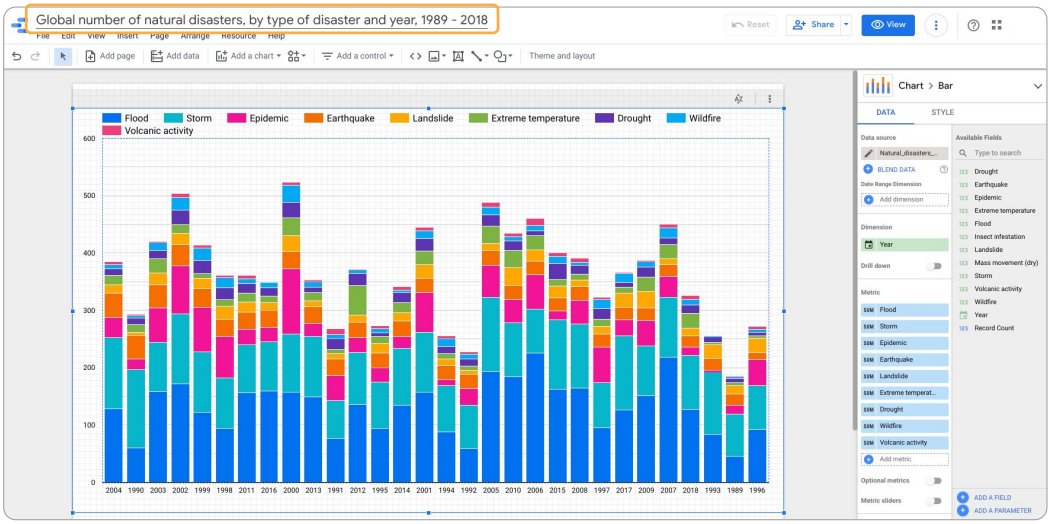
Easily change your **data table** view to a **chart** by clicking **Chart** in the properties panel and selecting a chart type from the options provided.
You can edit the style of your chart, or even change your chart type selection.
You can also revert back to a **data table** view.

You can also add separate charts by selecting **Add a chart** from the toolbar.
You can resize the components on the canvas to arrange data tables and different chart types as required.

In the same way that you defined **Dimensions** and **Metrics** earlier,
you can do the same for your chart by adding selections from the **Available Fields** list.

Here's a useful tip: The sequence of the fields under **Metric** will determine the order in which the data is displayed in the chart.
Use the drag feature to easily change the sequence of the fields.

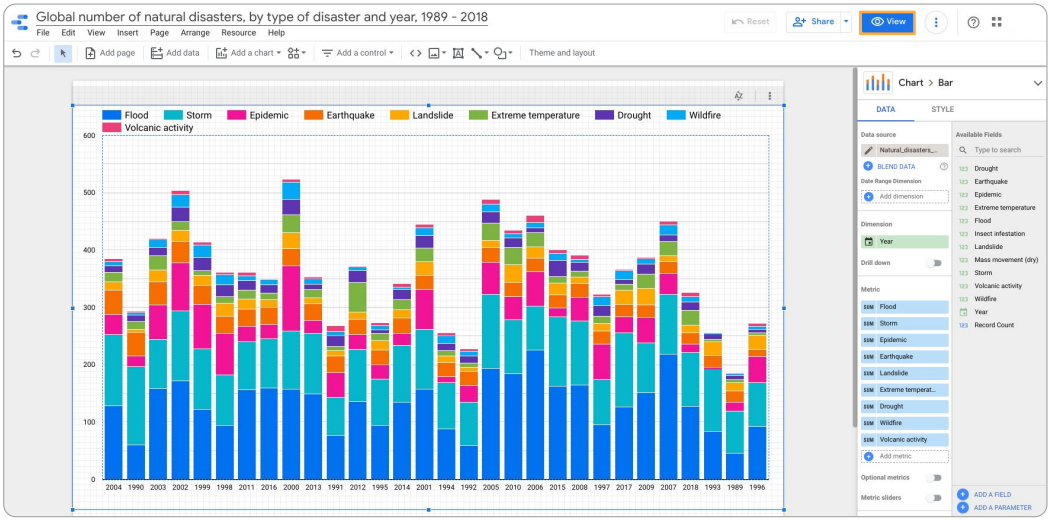
Add a descriptive name to your report



Give your report a name.

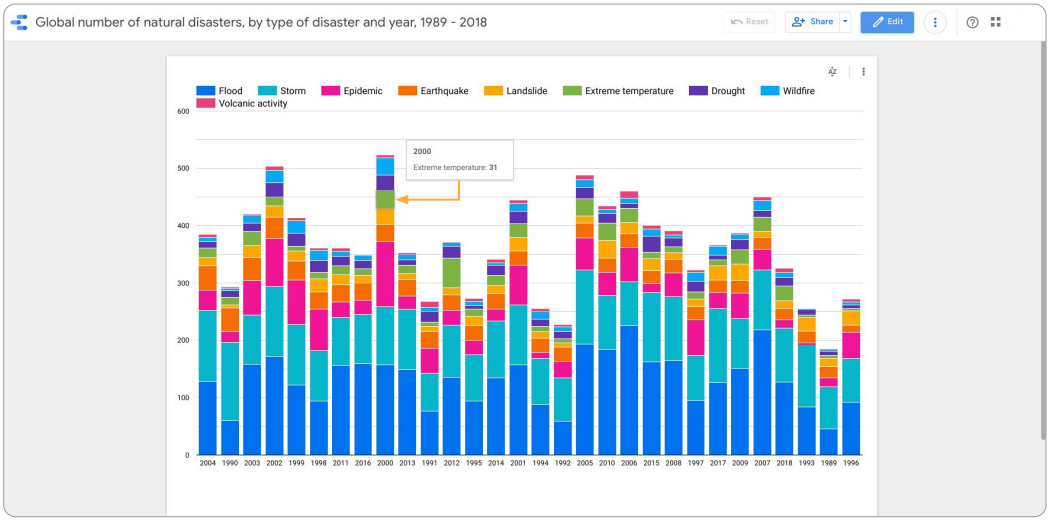
Since Data Studio is based on Google Drive you can have duplicate file names.

View the end-user version of the report



Click the **View** toggle button to view the end-user version of the report.

View your report as an end-user

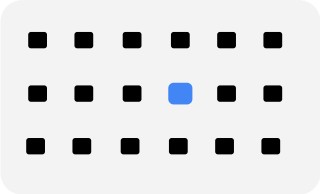


And here is your report. Notice it looks very similar to when you were editing it, but as a viewer, you can't modify the report.

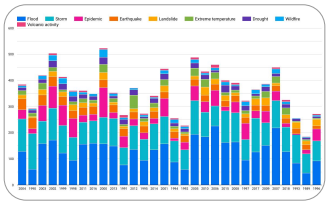
When a viewer mouses over the chart, they are able to view live data.
In this example, the viewer is able to see that in the year 2000, there were 31 natural disasters related to extreme temperature.

Note that users cannot edit your reports unless you give them permission.

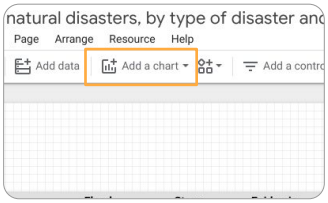
Summary: Explore and present your insights visually



Guide the eye of your user with preattentive attributes



Use the right visual to convey the right message



Create new report charts inside Google Data Studio

As we wrap up this module, we see that visualizing data is both an art, and a science. We've barely scratched the surface of visualization theory, and we discussed concepts like …

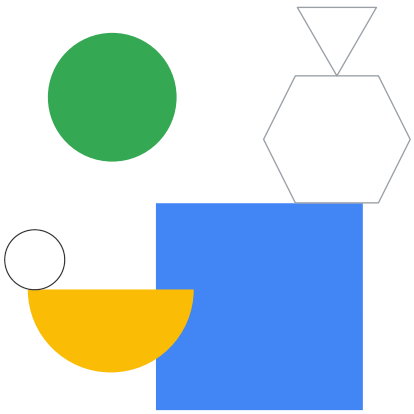
… pre-attentive versus post-attentive processing for that quick eye-to-brain understanding.

Along the way, we saw some bad ways to visualize data, as well as a few practices to do it right.

Lastly, we looked at Google Data Studio, which is the visualization platform we'll be exploring in more depth in our next lab.

Lab Intro

Explore and Create Reports with Data Studio



Now, it's time for us to explore an ecommerce dataset that has millions of Google Analytics records for the Google Merchandise Store loaded into BigQuery.

In this lab, you'll pair your knowledge of the ecommerce dataset with what you just learned about visualizing dimensions and measures. You'll create a new report and add visuals and interactive filters for your reporting users.