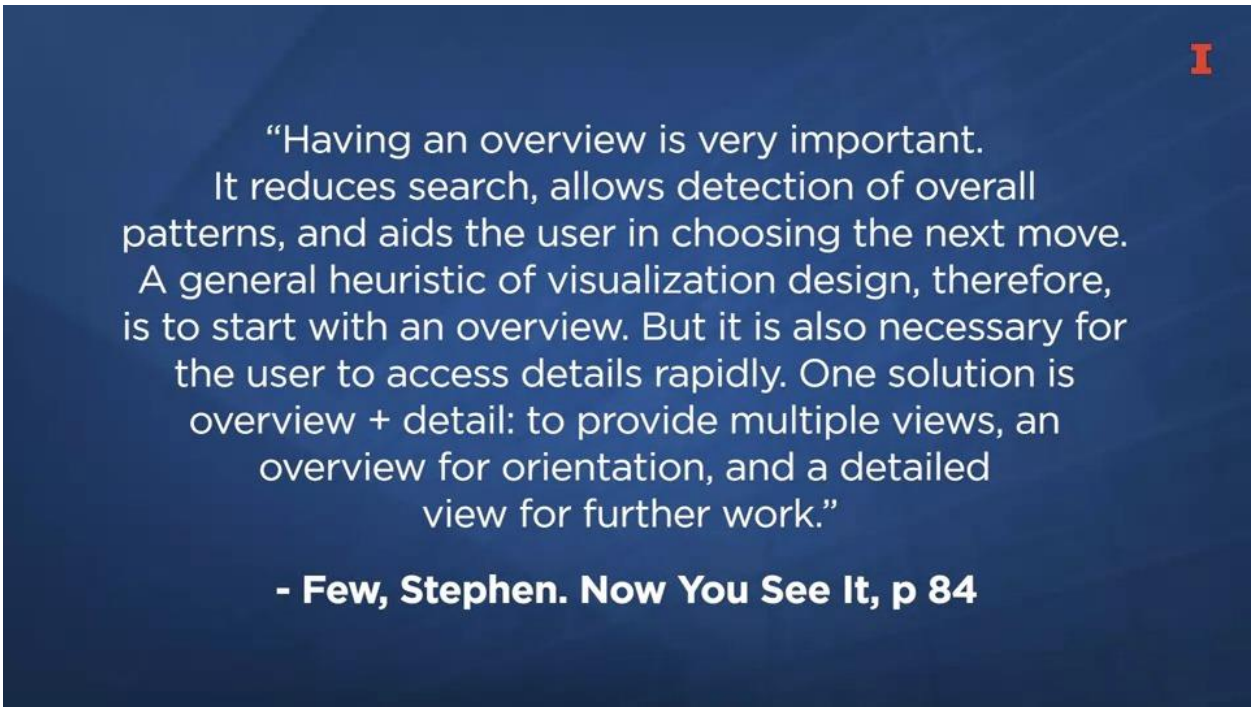
Module 2: ETL and EDA Using PowerBI

Table of Contents

Module 2 Overview

Module 2 Introduction

Stephen Few, a data visualization guru, shares how organizations like the CIA teach new spy recruits to make observations. They're trained to first get an overview of what's going on around them and then when they spot something abnormal, they shift from a broad observational alertness to more focused perspective, in which they analyze the details. In this module, you'll learn how to use power BI to do with data, what spies do when observing their surroundings, get an overview of the data and then narrow in on certain aspects of the data that seem abnormal and analyze them.



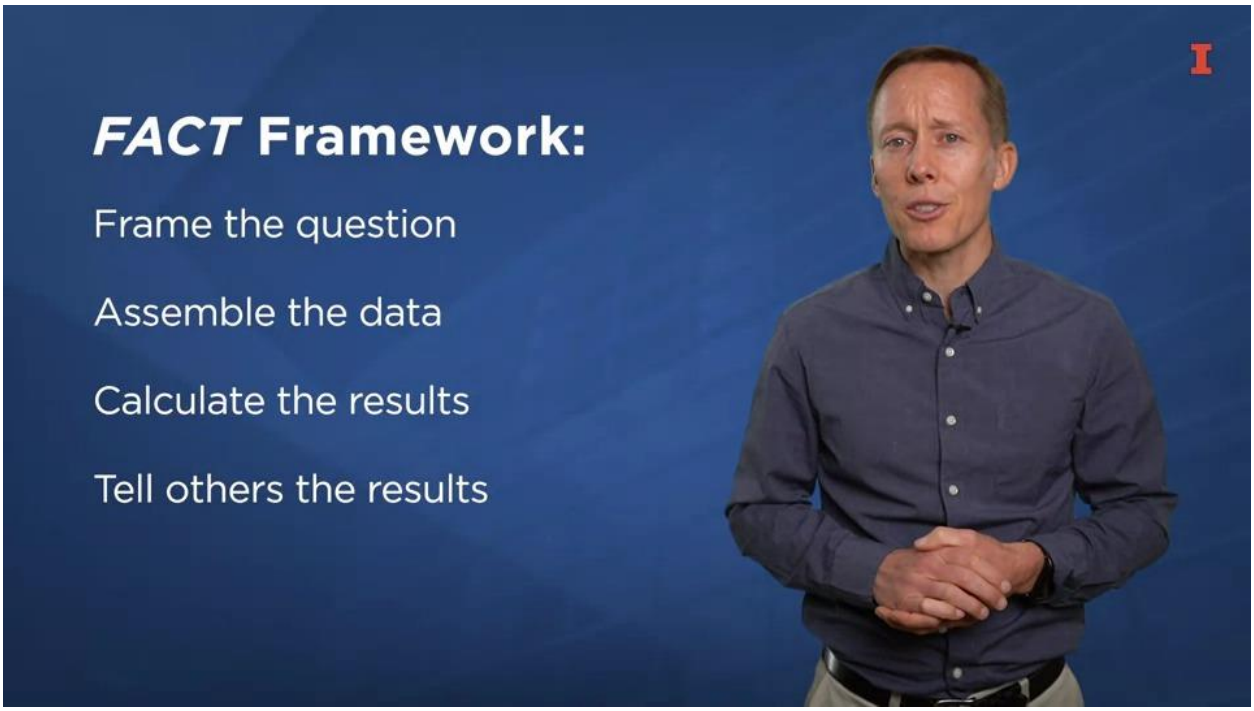
Stephen Few goes on to further quote some data visualization colleagues about why this approach is so useful. Having an overview is very important, it reduces search, allows detection of overall patterns, and aids the user in choosing the next move. A general heuristic of visualization design, therefore, is to start with an overview. But it is also necessary for the user to access details rapidly. One solution is overview plus detail; to provide multiple views, an overview for orientation and a detailed view for further work.

In fact, there's something called Schneiderman's Mantra, which is overview first, zoom and filter, then details-on-demand.

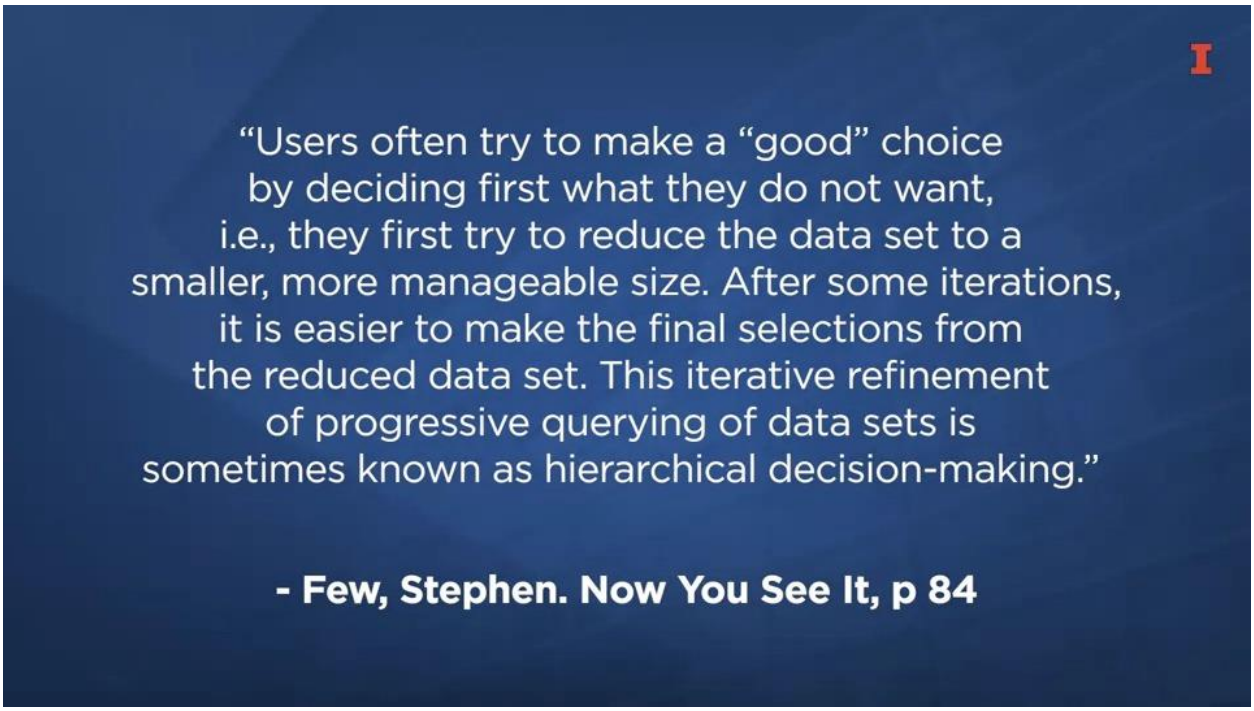
Now Powered BI is a great tool for implementing Schneiderman's Mantra because it is optimized for visualizing data, and so making charts is very quick and intuitive. Power BI is so intuitive it empowers many people, including those who do not have coding skills to make a wide variety of charts and then to use those charts to zoom in and filter the data to get a more focused perspective.

With respect to the Business Analytic Workflow, Power BI excels Exploratory Data Analysis. In some sense, Power BI is like a grown up version of Excel's pivot table and pivot chart functionality, like excel, Power BI is pretty intuitive to use and you don't need to know much about programming to use it.

Also, like Excel, Power BI has a really powerful tool for interacting with the data. The Power Query Editor, you can manipulate the shape of the data from wide to long. You can create new fields that are based on calculations of other fields. You can split or combine text fields. You can also perform Joynes and Lookups.

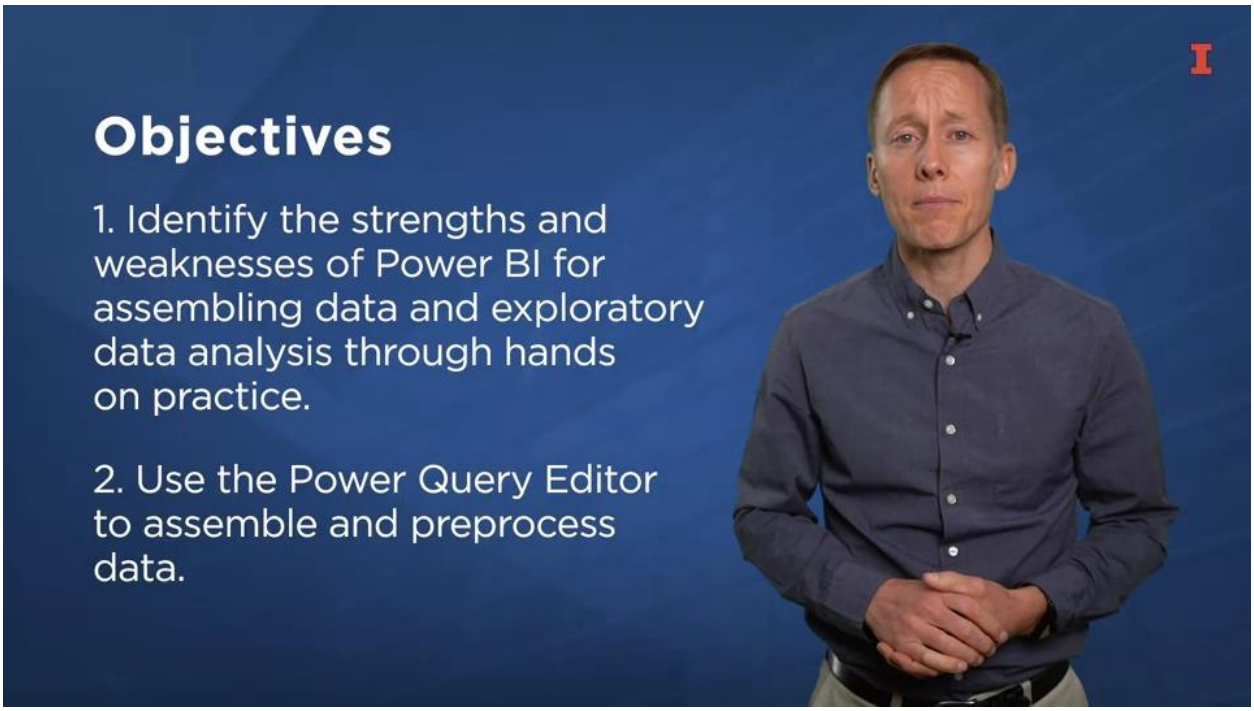


Thus, many of the key assembly processes that can be done in excel, can also be done in the Power Query Editor. The method for doing so is different, because you can't point and click on cells to make changes. However, the benefit is that you can handle much larger amounts of data. While Power BI is a very powerful tool, it is only a tool unless you have first framed a question and then have access to relevant data, which are the first two parts of the fact framework, then you're not as likely to convert data into action even when you do have a well framed question and good data, obtaining actionable insights is not a straightforward process. Quoting his data visualization colleagues, Stephen Few writes,

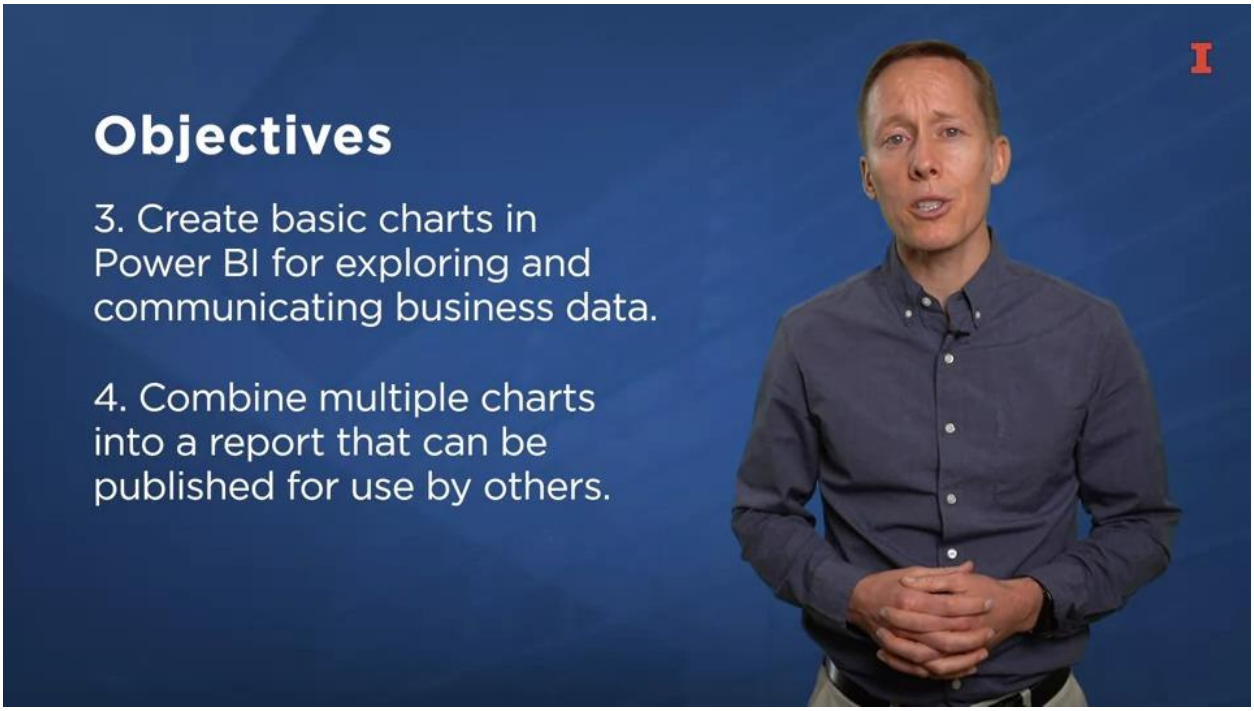


"users often try to make a good choice by deciding first what they do not want, that is, they first try to reduce the data set to a smaller, more manageable size. After some iterations, it is easier to make the final selections from the reduced data set. This iterative refinement of progressive querying of data sets is sometimes known as hierarchical decision making."

Now, in my experience, I found that this quote is accurate. Power BI helps with this hierarchical decision making process by using visualizations to identify and communicate useful relationships in the data. This lessons are not meant to be a comprehensive tutorial on Power BI. There are many excellent tutorials for becoming familiar with all the bells and whistles of Power BI.



The objectives of this module are for you to; one, identify the strengths and weaknesses of Power BI, for assembling data and Exploratory Data Analysis through hands on practice. Two use the Power Query Editor to assemble and preprocess data.



Three, create basic charts in Power BI for exploring and communicating business data, and four, combine multiple charts into a report that can be published for use by others.

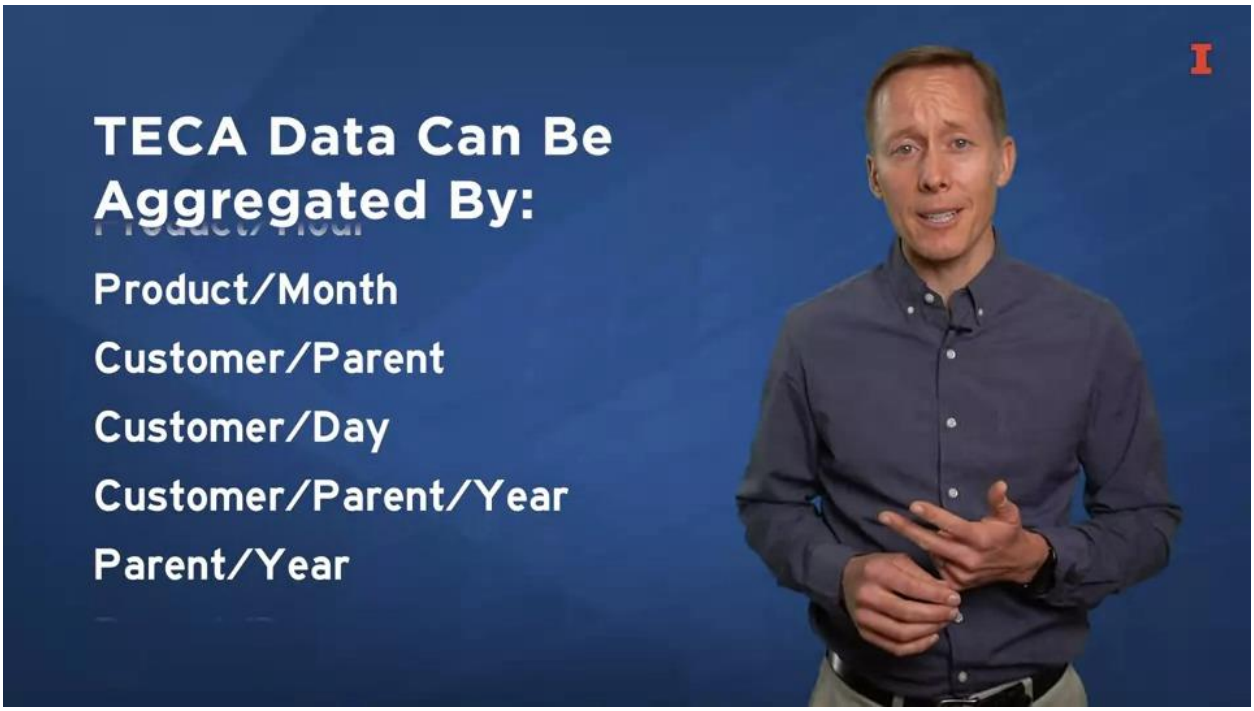
Lesson 2-1: Business Problem

Lesson 2-1.1 Data

In these lessons, we will be using point-of-sale data from Teca, which is the name of a company that owns over 150 convenience stores and gas stations throughout the middle of the United States. These stores sell typical convenience store items; gas, candy, soda, chips, lottery tickets and so on.

This data is known as point-of-sale data because it is generated at the point of the sale, like the cash register or the gas pump. You can probably imagine some of the information that is in this dataset.

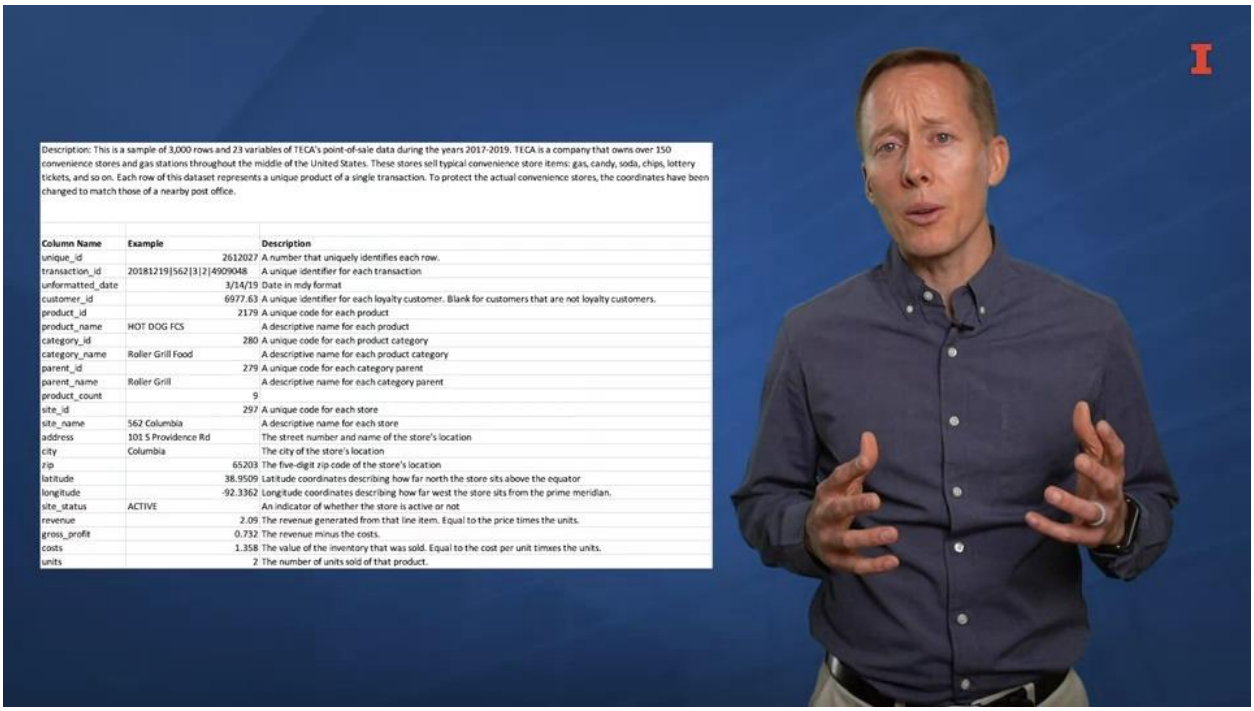
The columns in this dataset contain information about five main categories of the transaction. One, information that identifies when the transaction takes place and the identifier of the transaction. Two, information about the customer. Three, information about the product or products that are sold. Four, information about the store where the transaction took place, and five, revenue and profit data.



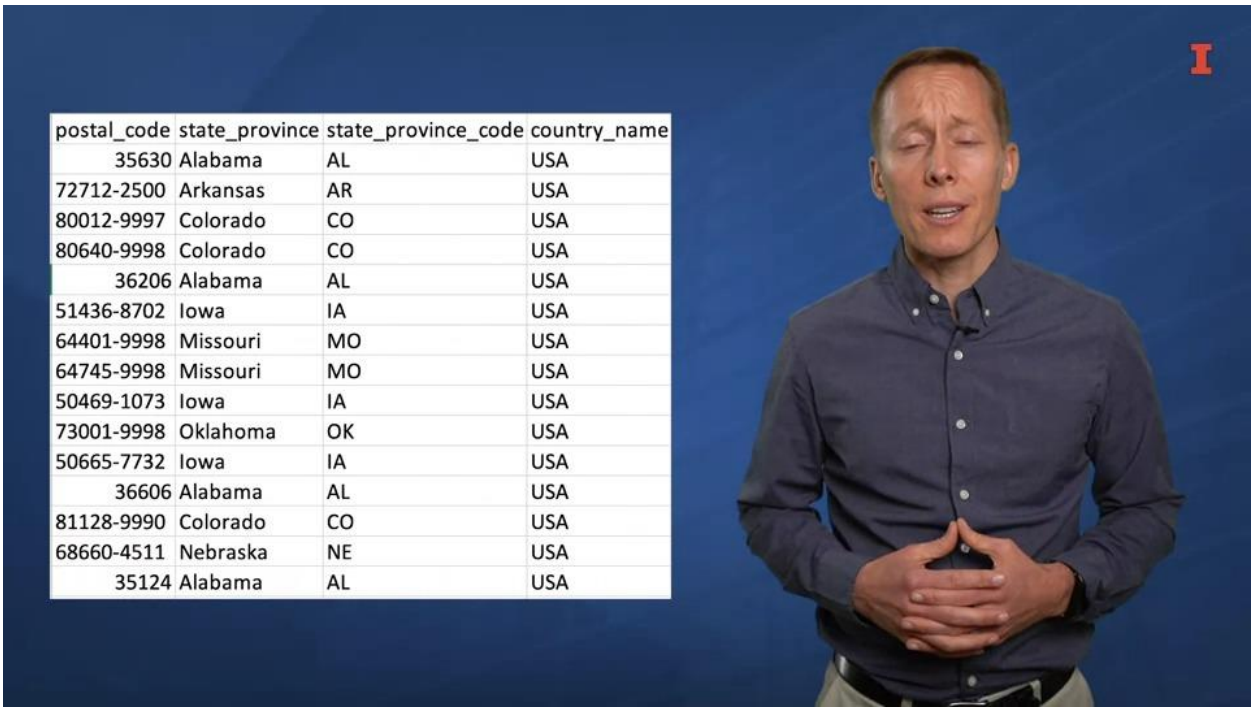
This type of granular data is awesome because we can aggregate it in many different ways, such as by day, customer, product, product category, and even by hour or minute of the day. We will be using a random sample of the data during the years of 2017, 2018, and 2019. Specifically, we will use only 3,000 observations. It would definitely be more informative to look at the full dataset,

but that full data set is so large that we would need a special machine with hundreds of gigabytes of RAM to analyze all of it at once. Now most machines can handle a whole lot more than 3,000 transactions. But since there is bound to be a wide variety of machine computing power among you learners, we wanted to keep it small enough that hopefully, even the weakest machine will be able to process this data.

Another benefit of using such a small data set is that it allows you to drastically reduce wait times as you experiment with different capabilities of the software. Now, even if you had lots of experience with the data analytics software, it would still be good practice to conduct data assembly and exploratory data analytic tasks using a sample of the data. Again, it will reduce the amount of time that you spend waiting for computations to be performed. Once you zero in on the specific data preparation tasks that need to be performed, then you can very easily set those processes to run on the full dataset.



Now there are two other files that go along with this dataset. One is a data dictionary that summarizes the contents of the dataset. Feel free to refer to it to get more specific information of what's in the dataset, especially as you first start working with it. Once you get some experience assembling and exploring the data, you'll learn a lot more about it than what is described in the data dictionary.



The other file contains information about states and zip codes. We will use this file to join the state data to the point-of-sale data using the data analytics software. There's not a data dictionary for this file, but it's pretty self-explanatory, so you should be able to interpret it without much help.

Lesson 2-1.2 Business Problem

In business analytics, sometimes you start with the business problem and then gather data that will help solve the specific problem. In this set of lessons, we are taking the opposite approach and we will explore the data to see what kind of business problems for which it can provide actionable insight.

In this video, we will elaborate on business problems for which we can gain insight by using the TECA data as a reminder that TECA data is point of sale data that comes from over 150 convenience stores in the middle of the United States.

Each row represents a line item of a transaction. The features of each row are related to five areas of the transaction,

one identifying information about the transaction, including the hour and minute of the day when the transaction took place, two identifying information about the customer if they are loyal to a customer or not, three information about the product or products that were purchased by the customer, four information that identifies the store where the transaction took place, and five revenue and profit data.

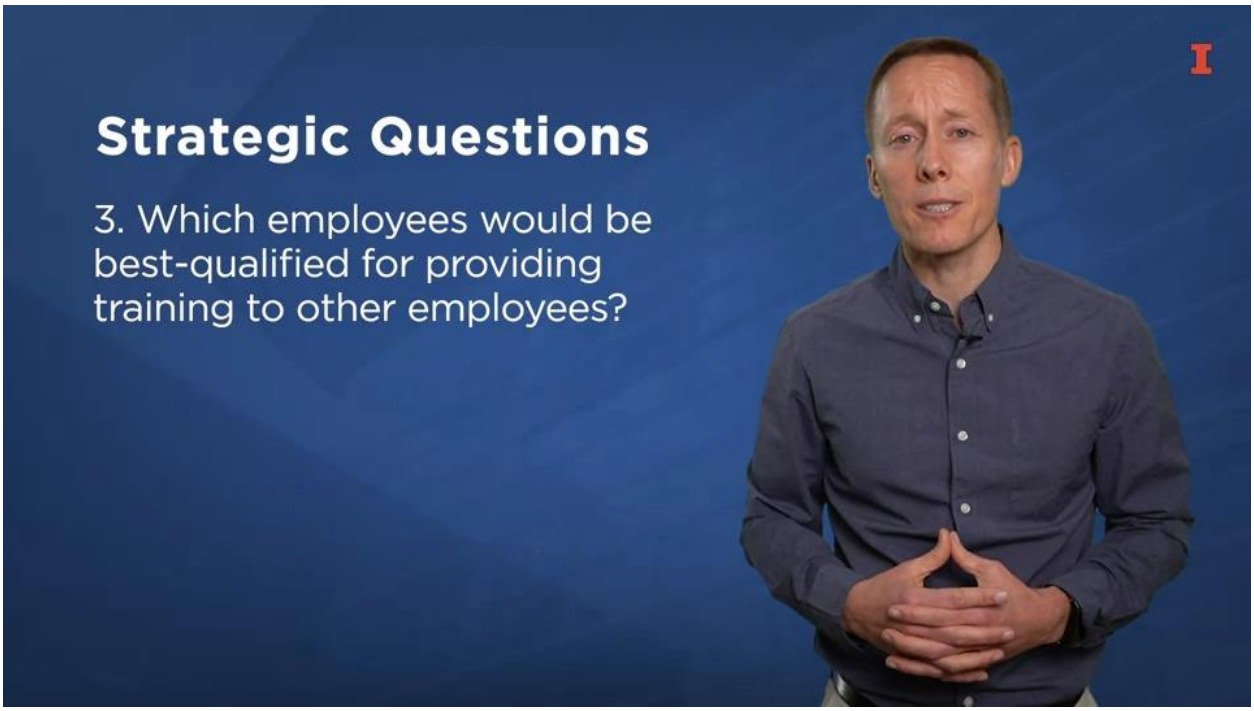
With such granular data, there is a multitude of potential business insights that we could obtain. Remember that we want a specific kind of business insight, actionable insight that will help us achieve some goal. Broadly speaking, those goals could be related to tactical business issues such as increased customer or employee satisfaction, less frequent stock-outs, less frequent stocking of unpopular goods that spoil or become obsolete, reduction of labor costs, and improved employee scheduling.

Specifically, some of these questions are, one, when should we order more of a specific candy bar so that we don't run out of stock? Two what times of the year and day should we schedule extra employees to be working to keep up with customers?

Three what are the high margin products that we can place near the cash register? Four what are complementary products that customers often purchase together so that we can place them next to each other?

Five how can we price our products to obtain the highest profit? Six which customers do we expect to see on a regular basis so that we can offer them a reward?

This data can also be used for strategic decisions, such as introducing new products by entering into a long-term relationship with a supplier, investing in another store, and employee training. Some of these specific questions might be the following. One should we operate and manage a bakery or lease out the space to an external bakery or fast food restaurant? Two if we have the financial means to invest in a new store, where should that store be located?



Three, which employees would be best qualified for providing training to other employees?

This is just a sample of the questions that we could seek to gain insight about from the TECA data. I hope this gets you thinking about additional questions that you could ask and how you would use the data to get insight to those questions. Coming up with questions that you can answer with the data is probably more important than the data itself,

I want to emphasize that the data itself does not provide insight. While it's true that we are going to focus on using business analytic tools to summarize the data and look for patterns in the data, remember that your mind is the most important business analytic tool. Try not to lose sight of the overall business problem as we get into the details of cleaning and exploring the data.

Lesson 2-2: Introduction to Power BI

Lesson 2-2.1 Introduction to Power BI

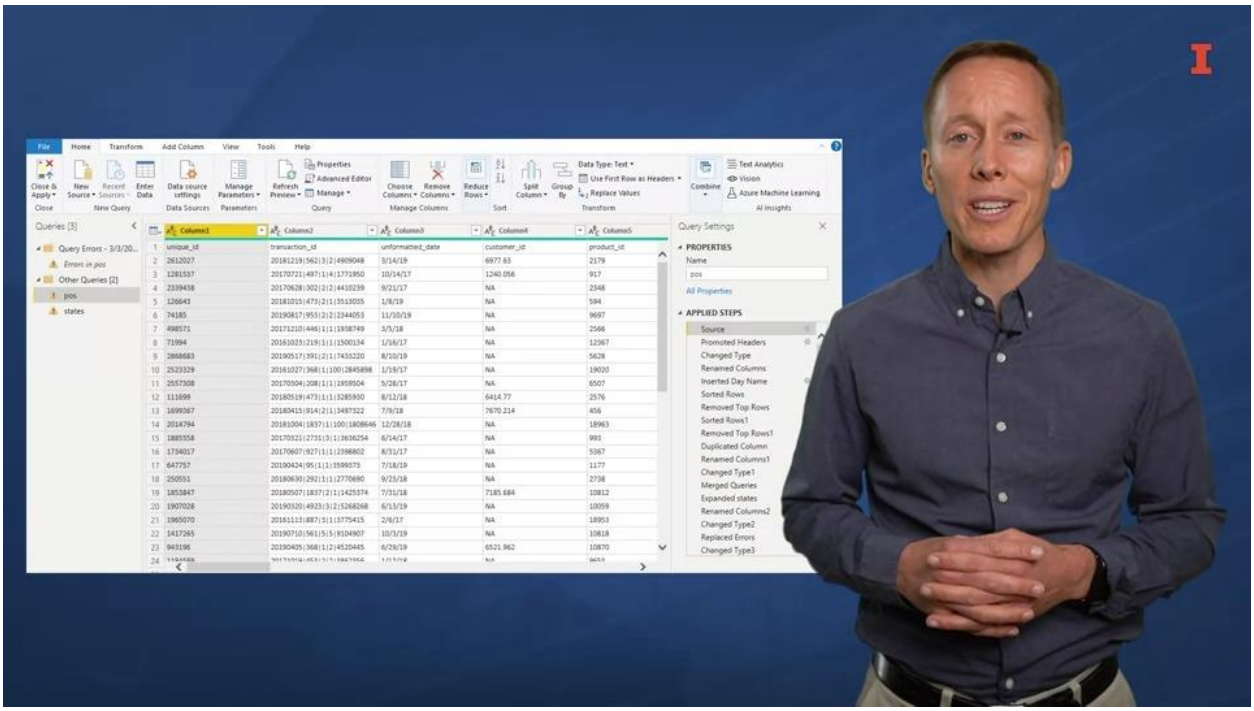
In this lesson, we want to introduce you to a business analytics software tool Power BI, which has a robust set of point and click functions for data exploration and visualization. You may have heard of Power BI because it was developed by the tech giant Microsoft. In recent years, Power BI has become a leader in the business analytics space and there are many reasons why that's the case.

Power BI has a wide variety of stunning data visualizations that can be created very quickly. This makes it a great tool for exploratory data analysis

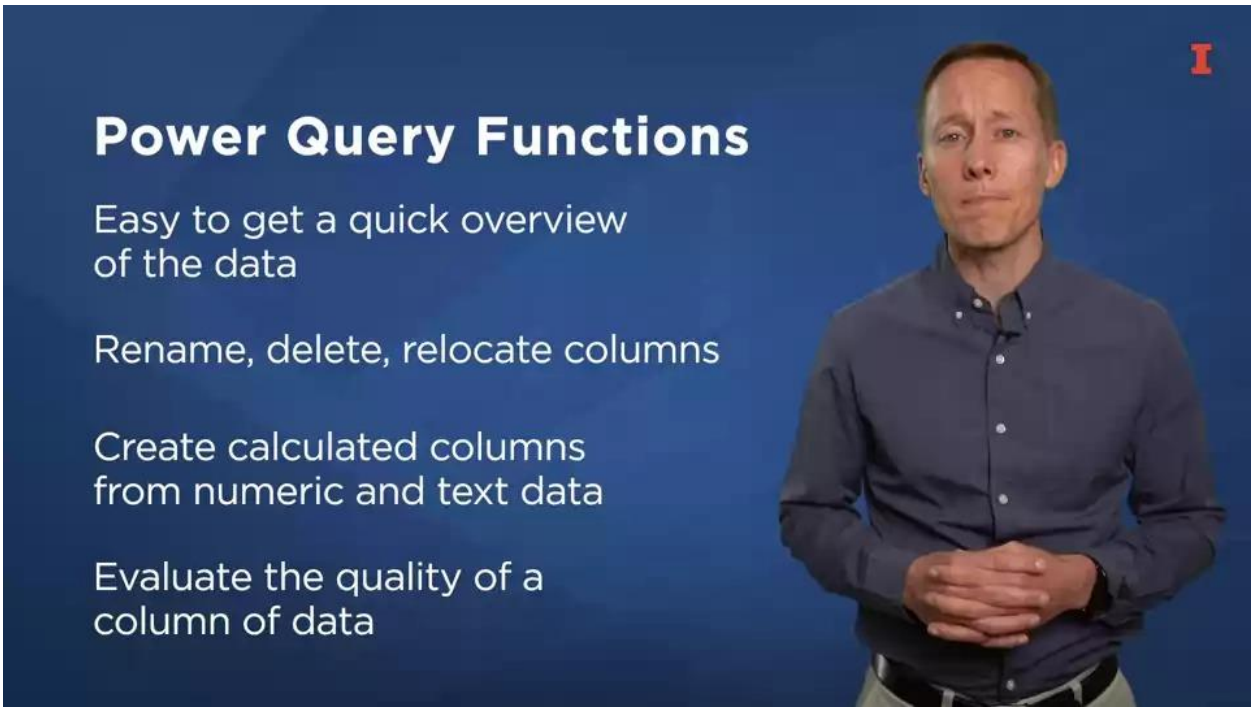
because you can quickly create a variety of multidimensional charts and then filter and drill through the data when you spot a distinctive pattern so that you can zoom in on what is causing that pattern. If you're looking for a chart that isn't included in Power BI, you can easily add in custom visualizations from external sources or create some with R and python.

Power BI is one of the most intuitive business analytic tools that I've used, especially if you're already familiar with other Microsoft products like Word, PowerPoint and especially Excel. The environment of Power BI will be very familiar to those accustomed to other Microsoft products.

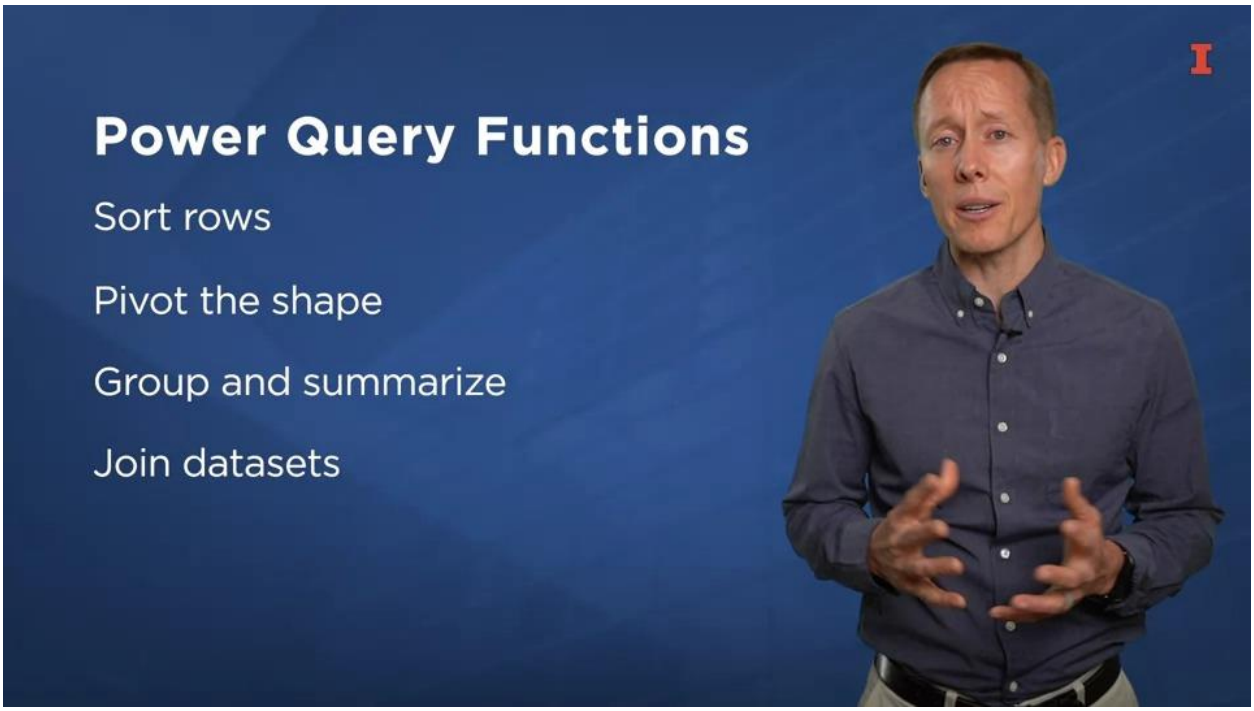
For instance, there's a ribbon along the top with different tabs and familiar icons. New tabs appear when you interact with visualizations and data. You can right click on charts and other elements on the screen to bring up additional options. You can perform common tasks in more than one way, which makes it easy for beginners, but also more efficient for advanced users. Saving and opening files, copying and pasting, formatting, creating shapes, importing data and many other such tasks are all completed in a way that is very similar to how you would do in Excel. Not only is the environment familiar, but many of the common data assembly tasks have also been made available with the click of a button



using the Power Query editor, which is built into Power BI. The data transformation tools available in the Power Query editor are just as important as the data visualization tools. However, it's harder to show data transformation tasks than it is to show a chart. So the visualization tools are probably more memorable to most people than the Power Query editor.



Some of the tasks that you can perform with the Power Query editor include getting a quick overview of the rows and columns, as well as visually exploring the data as you would in Excel, renaming columns, as well as deleting and relocating them, or creating calculated columns based on mathematical combinations of values and numeric columns, or from sub strings of text columns. Quickly evaluating the quality of a column of data to see how many rows have missing values or errors, and then replacing missing values and errors with other values.



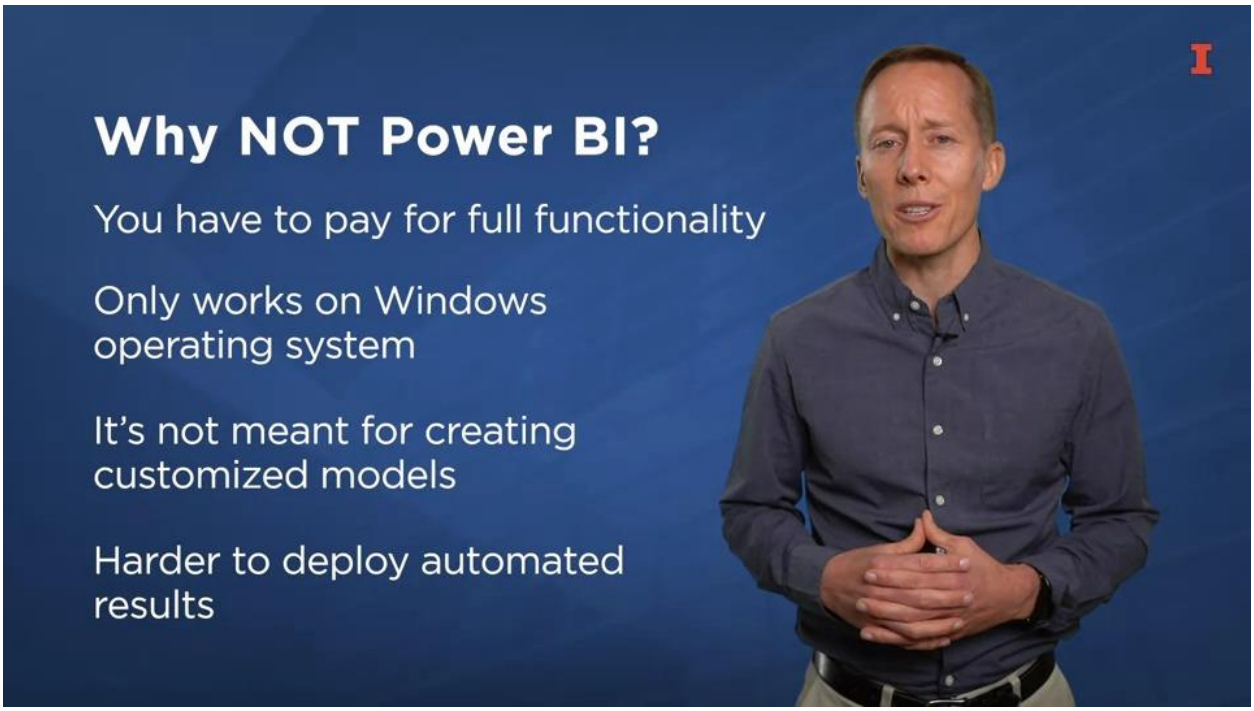
Sorting the rows of data based on values in a column. Pivoting the shape of the data from wide to long or from long to wide. Grouping the data and then calculating summary tables. Joining two data sets together based on a primary key.

There are many other tasks that you can easily perform. We won't have time to go over them, but with a little effort, you should easily be able to find out how to do almost anything you want.

There's one thing that I really like about the Power Query editor that has been missing from Excel, and that is a record of your steps are saved so that you can remember the process for the future, or easily revert back to an earlier version of the data. In my opinion, this has made Excel difficult to use for data analytics because it's hard to replicate a data manipulation process.

While there are a lot of point and click data visualization and transformation tasks that you can perform, there's also an option to use code if there isn't already a button for it. Data analysis expressions or D-A-X can be used for custom data transformations. As I mentioned before, you can use R or Python to create custom visualizations. However, just as a relatively small proportion of users take advantage of VBA in Excel, I suspect that a relatively small proportion of users take advantage of DAX or D-A-X because there's so much you can do without learning to type out code.

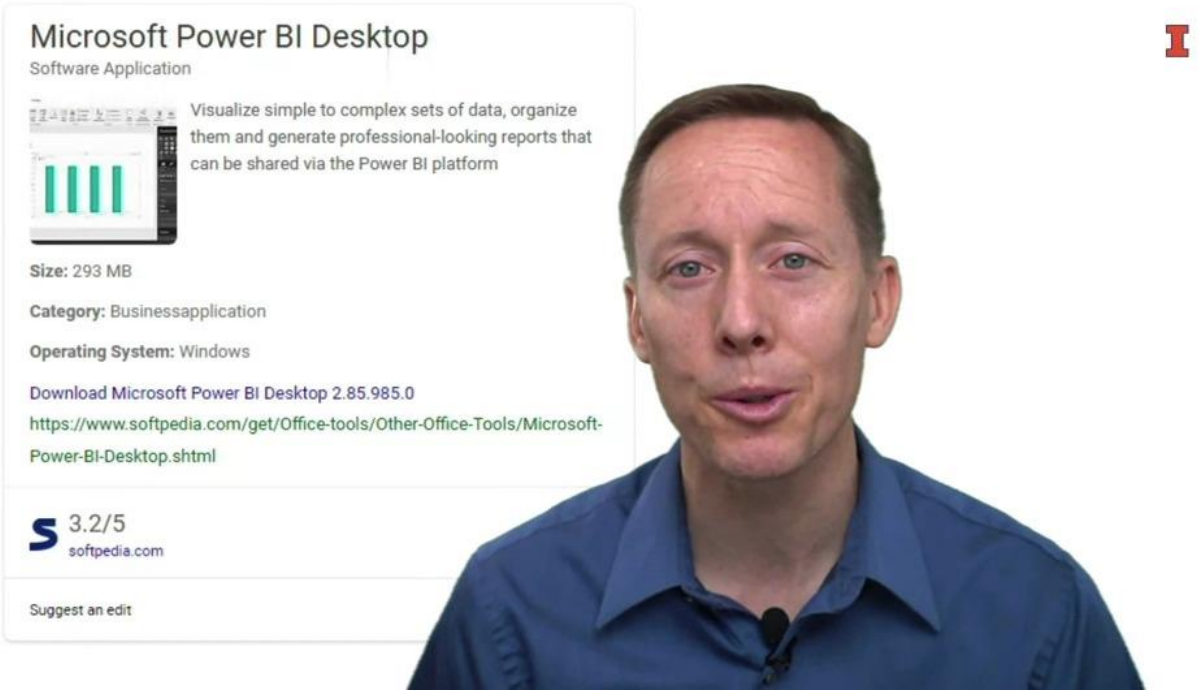
Power BI is also a great way to tell others about data analytic results, but you have to pay for it. If your organization has a subscription, you can publish your reports to a space where others in your organization can access and interact with it. This allows multiple people to analyze a single source of data at the same time, rather than having different versions of the data floating around. Power BI has many other powerful built-in features, such as the ability to send alerts and to apply a limited number of advanced analytic algorithms to make predictions or identify key influencers. There's a lot to like about Power BI.



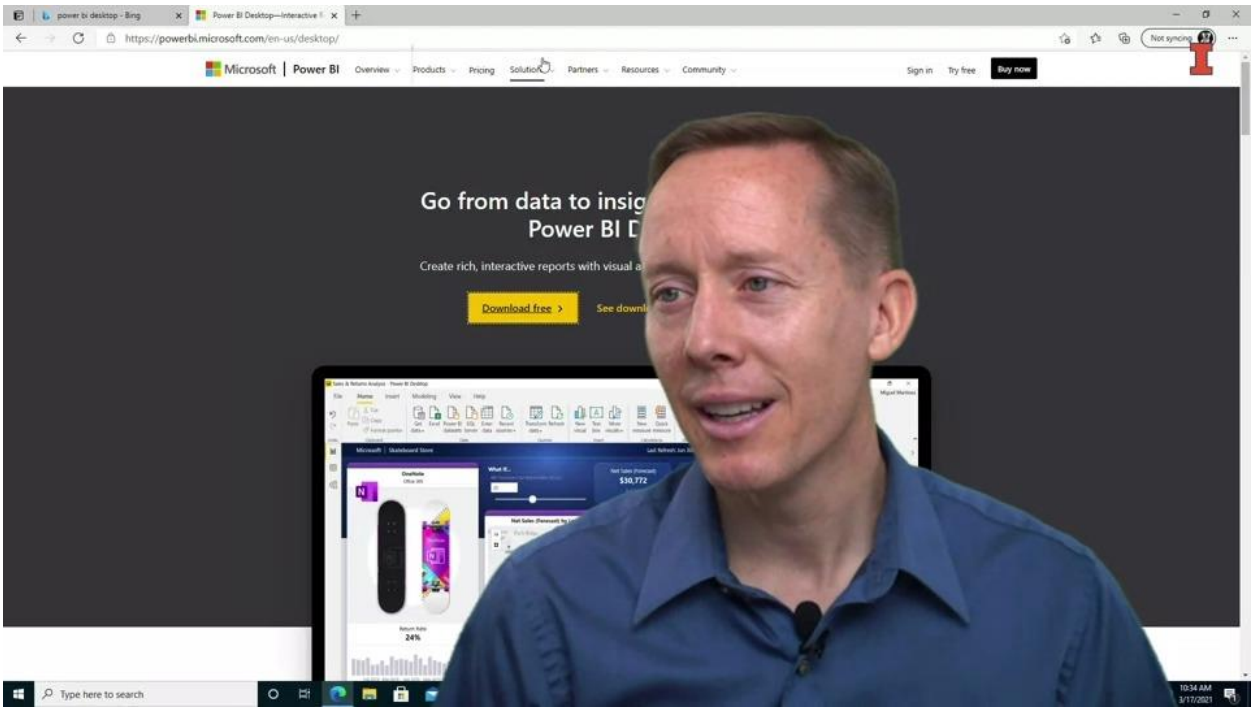
To be fair, I should also mention the weaknesses relative to other commonly used data analytic tools. One weakness relative to open source tools, is that you have to pay for the full functionality of Power BI. While you can do a lot with the free version you have to pay to unlock the most powerful features. Another weakness is that Power BI only works on Windows operating systems. There is a way to use Power BI for those who have Mac or Linux operating systems, but you have to go through a few extra steps, such as partitioning your hard drive or installing some of their software like Virtual Box, which allow you to run the Windows operating system. There's also a cloud based version of Power BI, but it doesn't contain all of the functionality. Even if it did, it's not ideal because you would have to load your data to a cloud based server.

Now, how does Power BI compared to a data analytic language like R or Python? Well, as is often the case with point and click software, there's a trade off in that it's easier to use, but limited in what it can do. For instance, Power BI is not the best tool if you want to create customized models using specific analytic algorithms that have been rigorously tuned and trained. Another limitation relative to a data analytic language is that, I suspect Power BI is limited in its ability to deploy automated results. For example, I suspect that it would be difficult or impossible to set up Power BI to acquire data, send it to Google's image detection model for classifying what is in the picture, and then send the results to some other output platform. Thus, it's probably not the tool of choice for creating customized apps that depend on a lot of sophisticated data analysis. All in all, I would say Power BI is a wonderful tool for allowing a broad range of people in an organization to conduct exploratory data analysis because it's very intuitive to use. But it's not the tool of choice for those who want to get deep into data analytics or who want to deploy automated data analytics solutions.

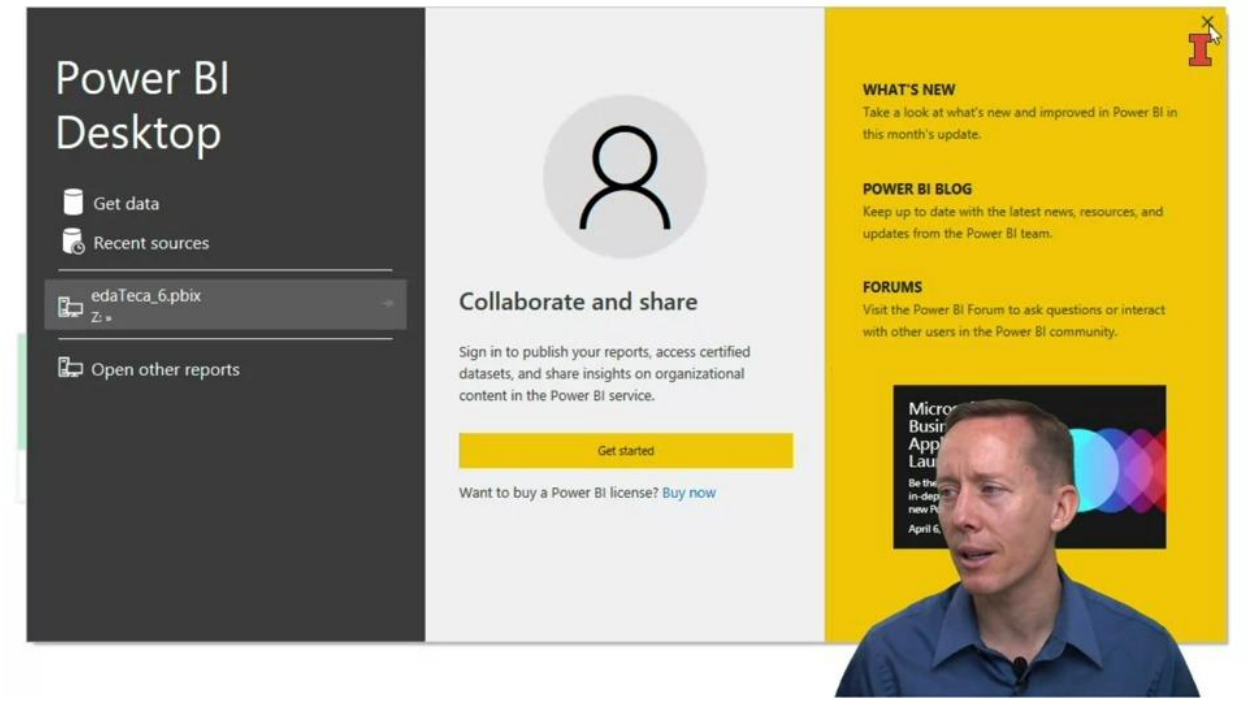
[Lesson 2-2.2 Installing Power BI]



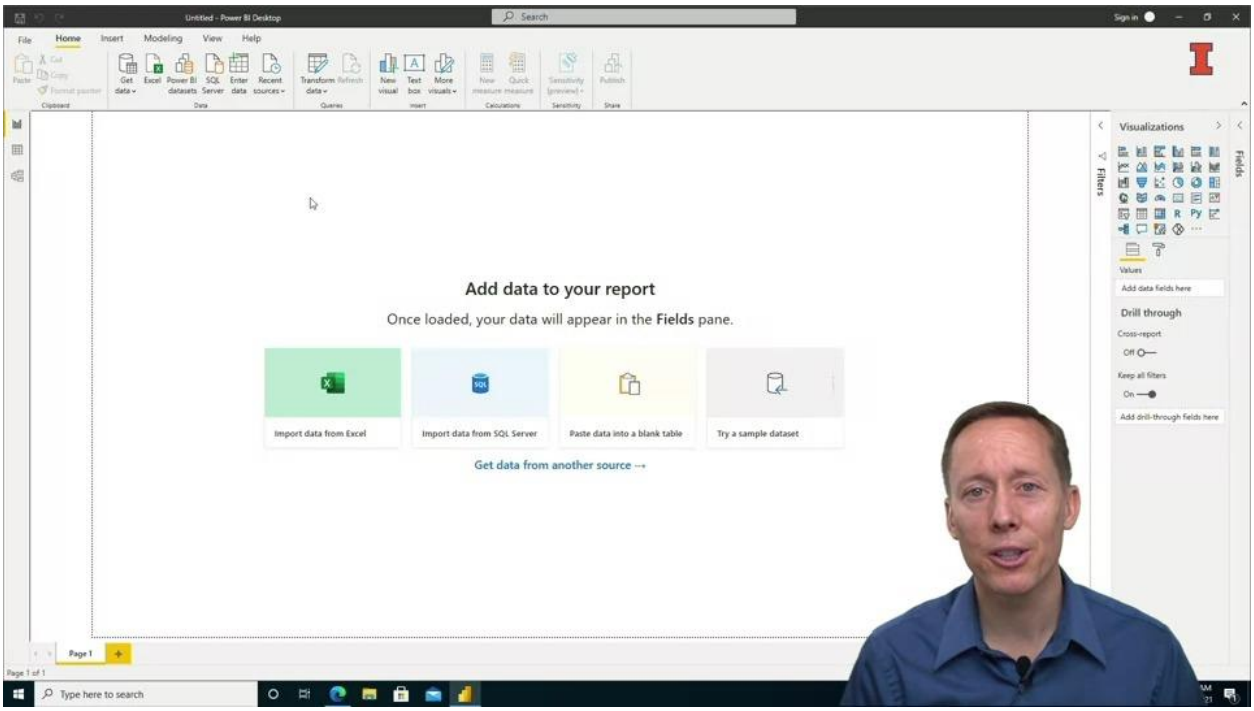
In this video, we want to help you get started using Power BI I on your own machine and then give you a high level overview of the Power BI I environment. The first thing to recognize is that Power BI I only works on Windows operating systems. So if you're using a machine that has a Mac operating system or a Linux operating system, then you'll have to install some other software that allows you to emulate a Windows operating system. And that's what I'm doing.



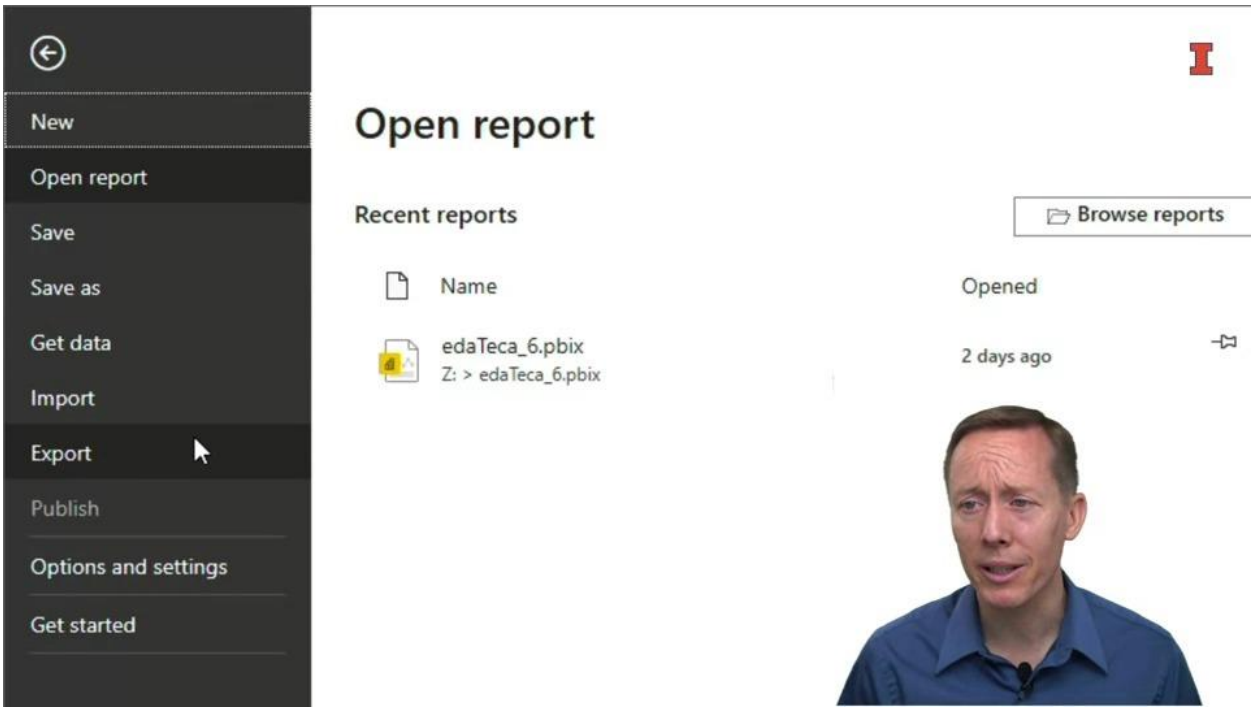
This demonstration is based on a Mac and so I have virtual box running, which allows me to emulate a Windows operating system. All right, once you get that going, go ahead and open a browser and just do a search for Power BI desktop, you don't want the online version it doesn't have as many features. One of the top links should be Power BI desktop and if you click on that it'll take you to a website where you can download Power BI I desktop for free. And if you click on that it will do everything for you. At least it did for me and you can install it and then launch it and there you go. So using power bi desktop is very easy to install on your own machine. There may be some differences for some of you, but I can't anticipate what all of those would be. Now is Power bi desktop really free? Yes, it is. But if you want the full functionality, especially to use it with your organization, you will have to pay for additional features.



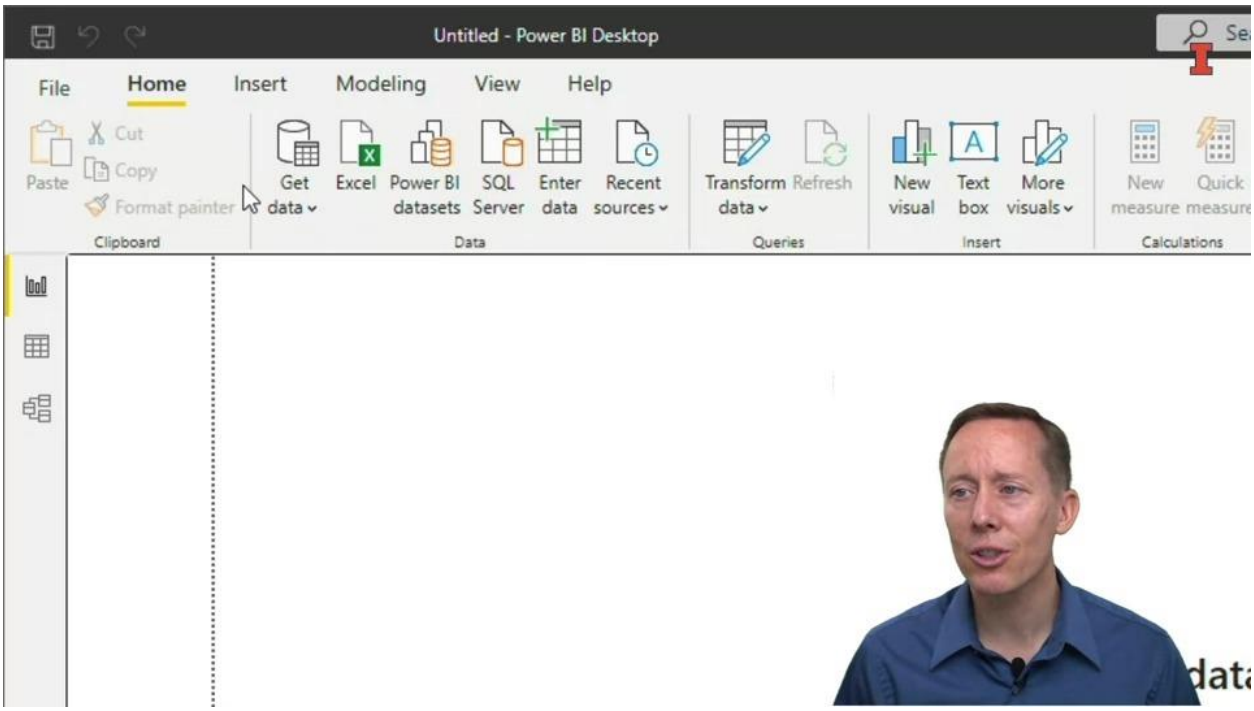
Once you've installed Power bi desktop, you can go ahead and launch it. All right, once Power BI I starts up, you'll get this welcome window that intends to help you get started with the most recent project that you're working on and gives you other information about updates. I'm going to go ahead and close that and what I want to do here is just explain to you the overall environment so that you get a feel for how it works.



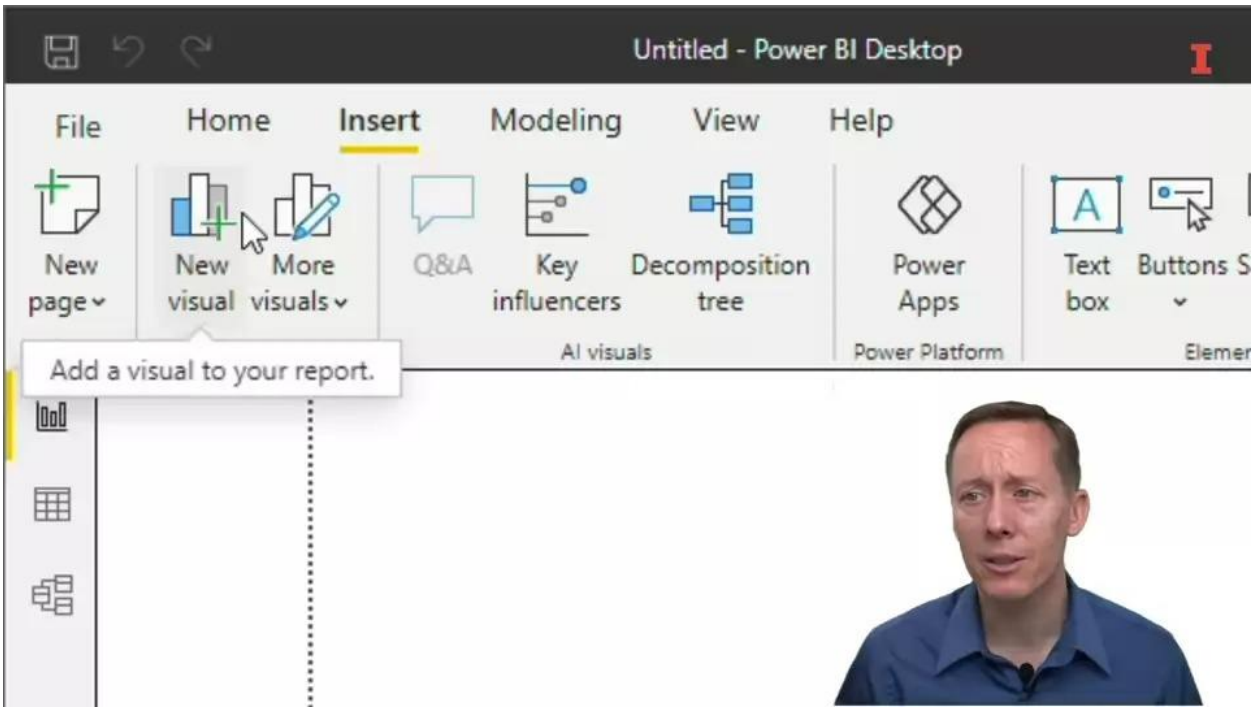
So the first thing to recognize is that this environment is very similar to an environment such as Excel. This main area here is the canvas where you will build visualizations and interact with the data and we'll get to that in a little bit.



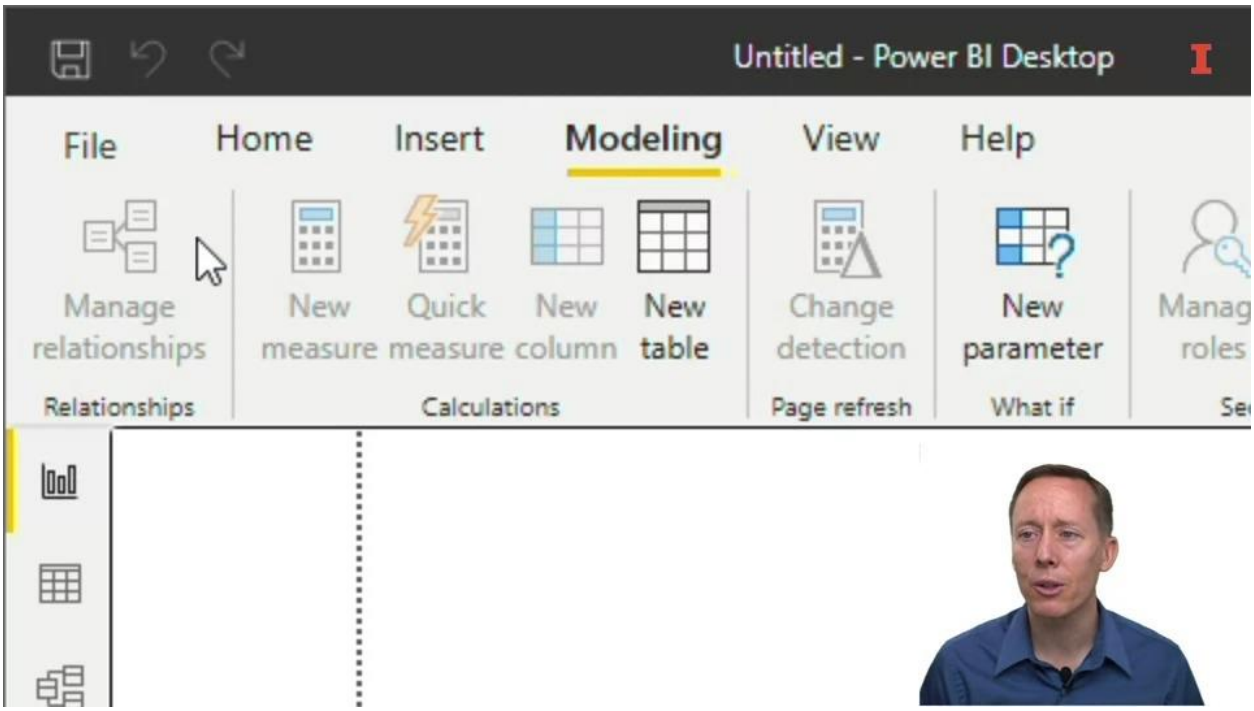
Let's talk about what everything on the side and top are all about. So let's start with this top menu up here. You've got the file menu which if you're a user of other Microsoft products, this will be very familiar where you can open a new report, open an existing report, save your progress, save as, get data, import, export and so on. So you can see here is a report that I've recently been working on.



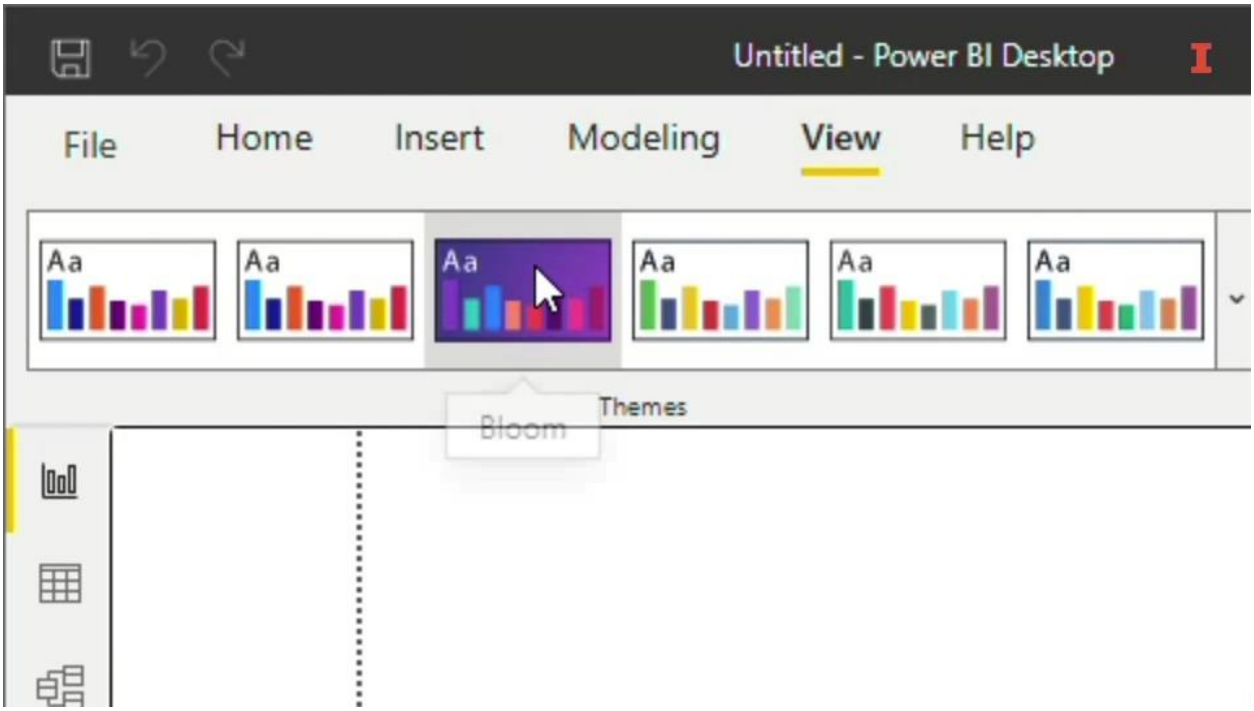
Then you've got the home menu which is kind of the default menu. This has different sections in here for commonly used features in Power BI. So this first section is the clipboard and this again if you're familiar with Microsoft products, this is very familiar. You've got paste, cut, copy and a format painter here. And then you've got a data section. And this section is very useful for accessing data. So you can get data from a CSV file or an Excel file from Power BI data set or a SQL server, so this allows you to connect to a remote SQL server. You could enter data manually and then just recent sources. Then you've got this query section and we will be using this a lot more in other lessons. This allows you to manipulate the data. Then you've got an insert section for inserting some new visualizations and text boxes, some of this you'll be able to use in more detail on this insert tab here. You can also create some quick calculations, some new measures or columns in the data and then a couple of other things, you've got the share here if you want to publish a report online for others to access. And this is one of the features that you have to pay for if you want to allow your whole team to access it.



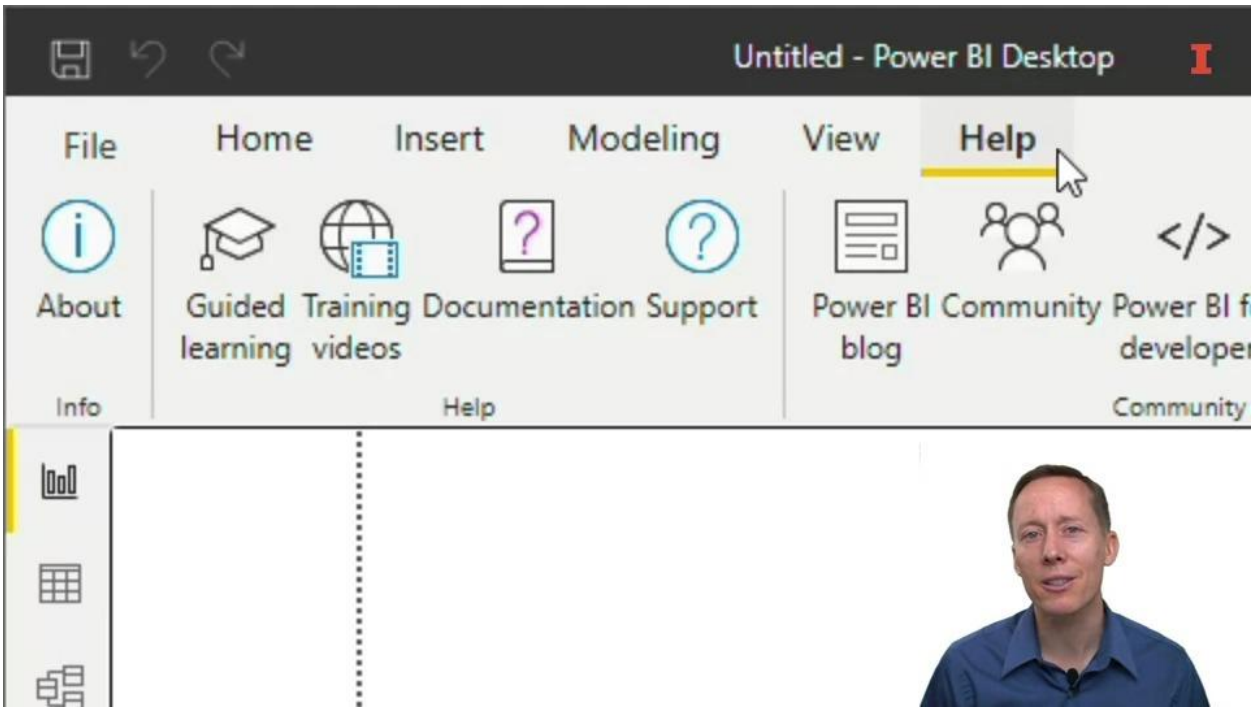
All right then we've got the insert tab at the top and this allows you to insert a new page of your report down here at the bottom or insert new visualizations and a lot of these things you can do in other areas as well. So you can go to the sidebar on the right here and insert visualizations that way. So some of this is redundant, just depends on how you want to use it. And then you have this AI visual section and this is actually really cool, we're not going to go into this. But key influencers allows you to evaluate what is influencing categorical data, whereas the decomposition tree I find is more useful for continuous or numeric data. Then you've got a power app section where you can actually create applications for your business that are based on power BI. And finally you've got this Element section where you can add in a text box or buttons, shapes, images. Again, this should be very similar to what you may have seen in Excel.



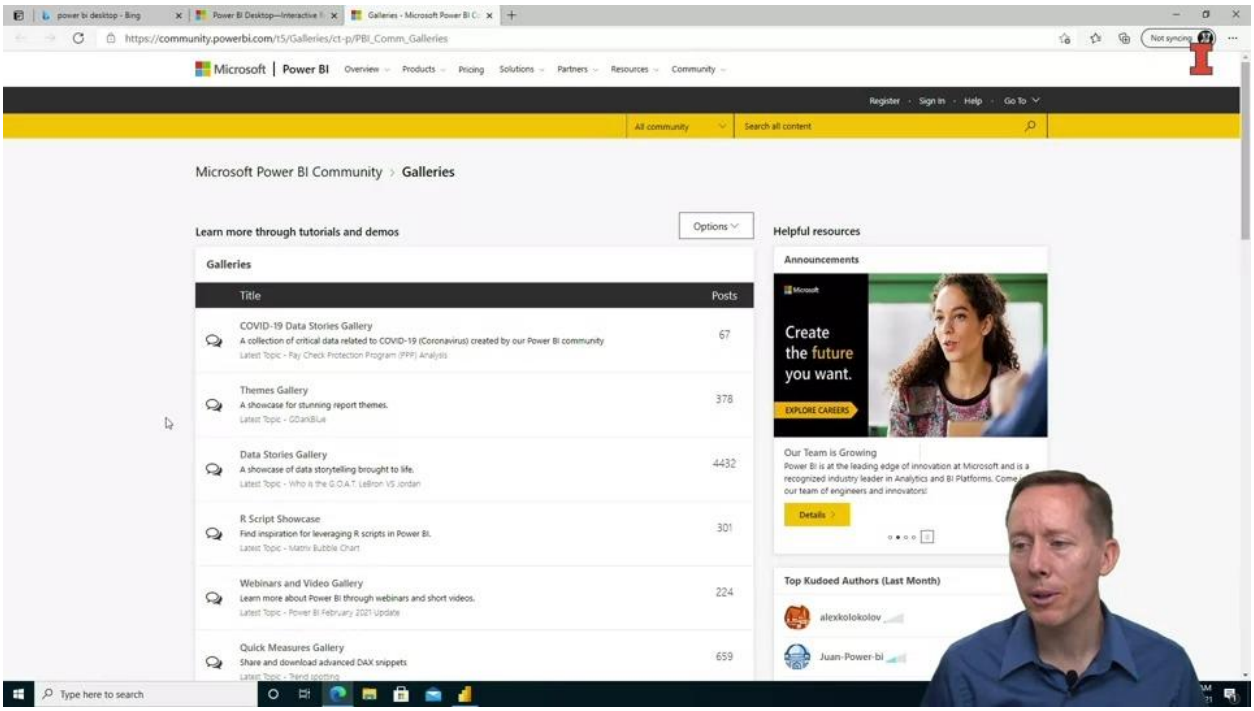
All right, then you've got the Modeling section at the top, which we're not going to use a whole lot, but you can manage relationships here, which you can also access in another area. You can again create calculations or new columns in your data and a variety of other things.



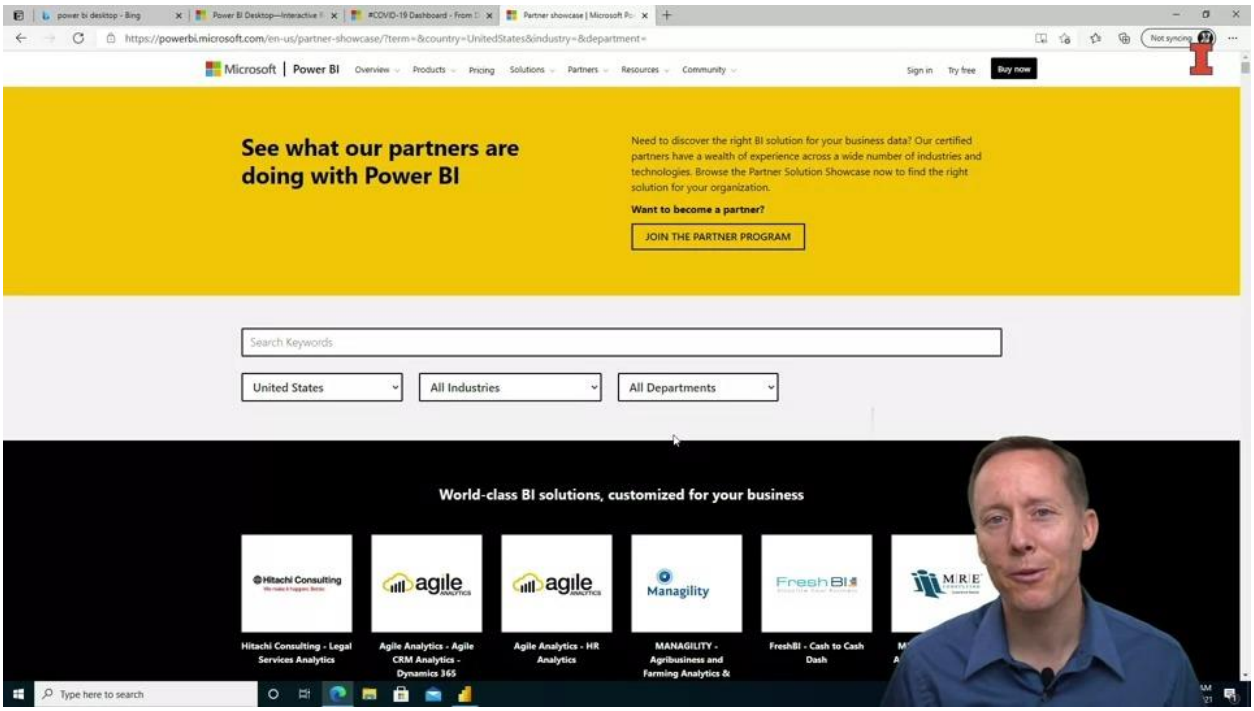
Then we've got the View tab here which allows you to change the aesthetics of what your reports looks like, so the Color scheme, the Layout.



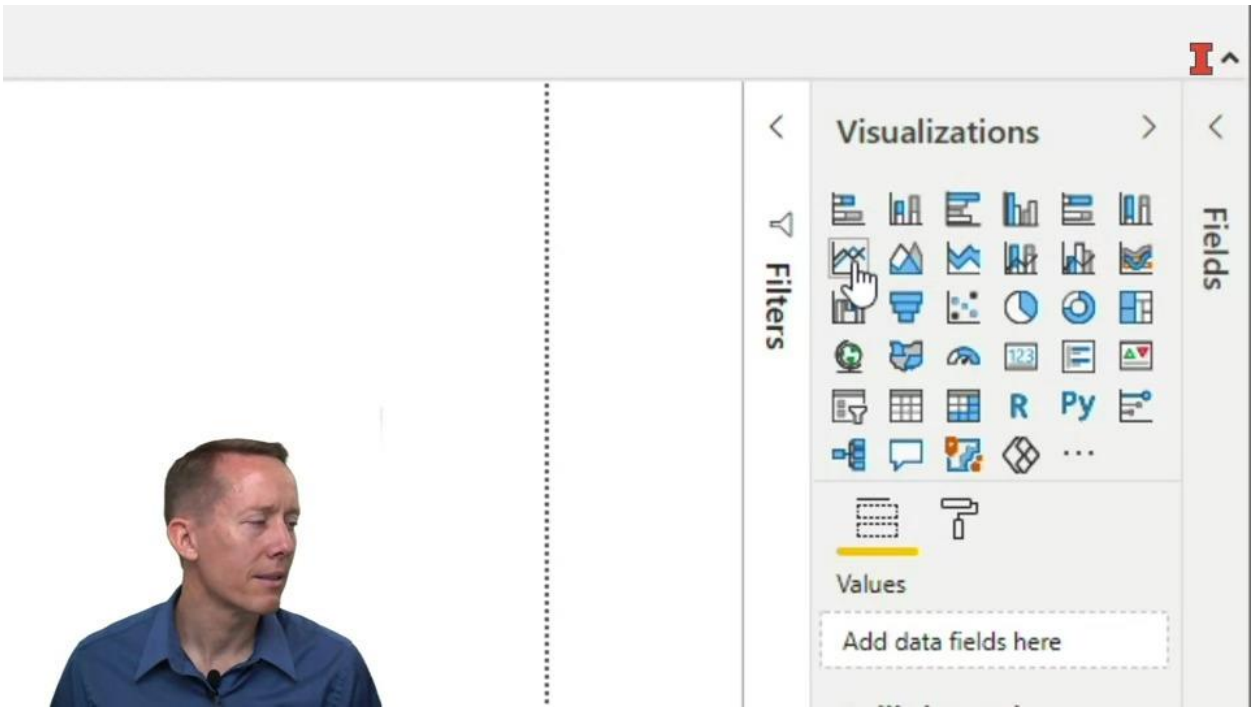
And then one thing that I think is very helpful in any software tool is learning how to use the Help features. So I would take some time exploring what is in the help section here. If you know how to use this, you'll be able to extend your application much beyond what we'll talk about in these lessons. So you can see that Microsoft has done a great job of providing Help. You got Guided learning, Training videos, Documentation, Support, where you can chat in. There's a Power BI blog, there is a large Community of people that use Power BI. And then for developers you can Submit ideas, External tools that you can import and then what I think is super useful are these Examples. So let me click on this and just show you a couple of things. You can use a Sample data set if you want, that is included with Power BI. You can also look at Sample reports and then the thing that I really want to focus on in this lesson here is this community galleries and partner showcase.



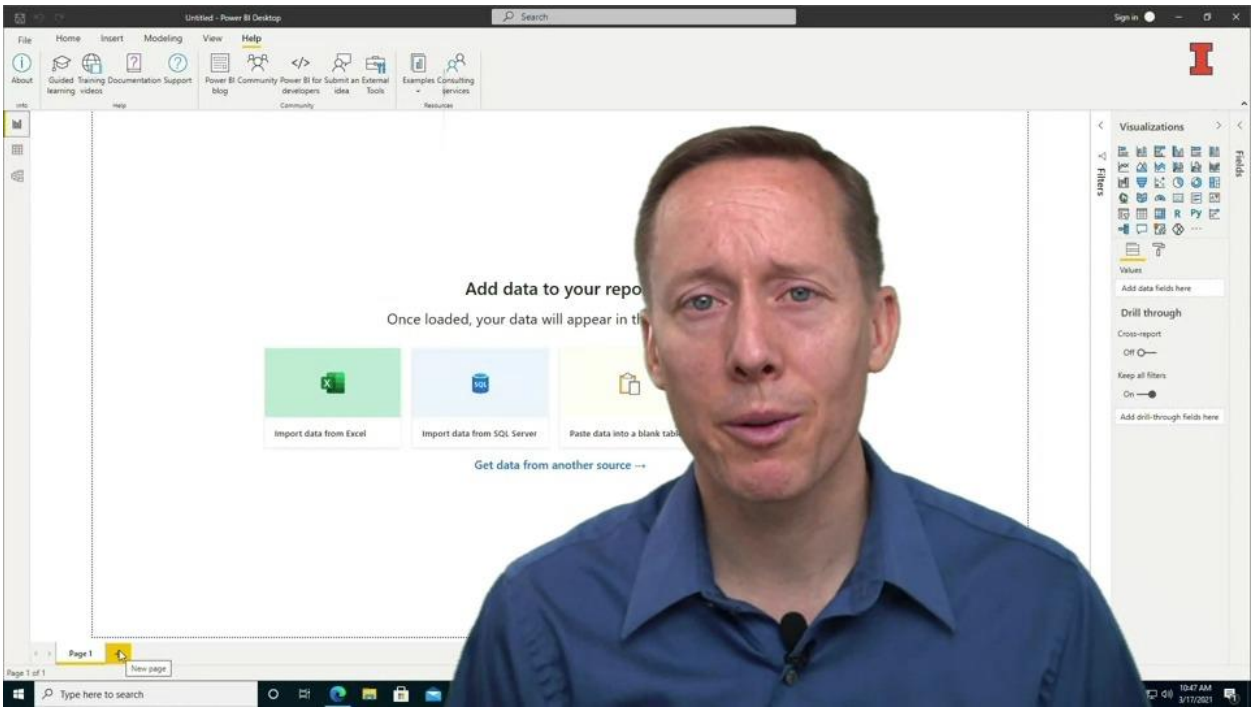
So if I click on community galleries, this will bring up a website with a variety of reports that people have made for a specific theme. So let me click on this top one here, COVID-19 Data Stories Gallery, and this will be a collection of reports, that people have made using the COVID-19 data. I'll go ahead and just click on this first one here and you can see that this report comes up that I can now interact with and explore all online. I don't have to have it on my desktop per se. But I think it's really helpful for getting an idea of what power BI is capable of doing. And can do all these hover overs that interact with the charts and then it gives you an explanation of how to use this dashboard down below.



The other section in this example that I want to show you is the Partner showcase. So if you click on Partner showcase it will take you to a website where you can explorer reports made from businesses that use power BI. So let's go ahead and look at the Explorer Solutions and then from here I can explore Power BI reports from a variety of specific industries. So if you click on this all industries here, you can see that there are many different industries I could choose form. I'll go ahead and choose education and then I can even go into different departments and look at how they're using Power BI. Let me just go ahead and select accounting, and here are some companies that are using Power BI. And I'll click on this Hitachi Consulting and this gives me an overview of the report that they use. I can even click on View Report and interact with their report and read more about it.

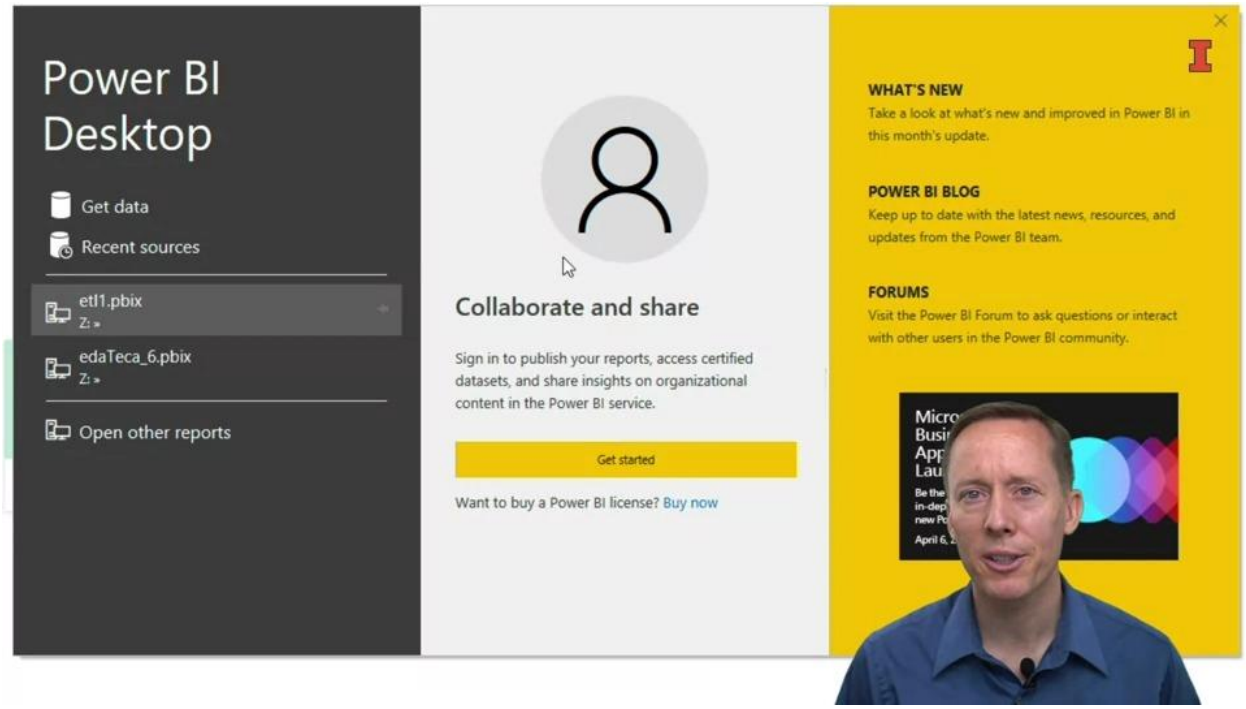


All right, let's go ahead and go back to Power BI and there are just a few other elements here that I want to illustrate to you. So on the right side bar we've got the main controls where we'll spend a lot of time for creating visualizations. So there are all sorts of different charts that you can create, Bar charts, Funnel charts, Line charts and many others. Then on the left sidebar we have menu items that allow us to quickly toggle between a Report view, a Data view, or the Data Relationship view. And then at the bottom you can see that you can add different pages either to the Data, Modeling, Relationships, or to the Reports.

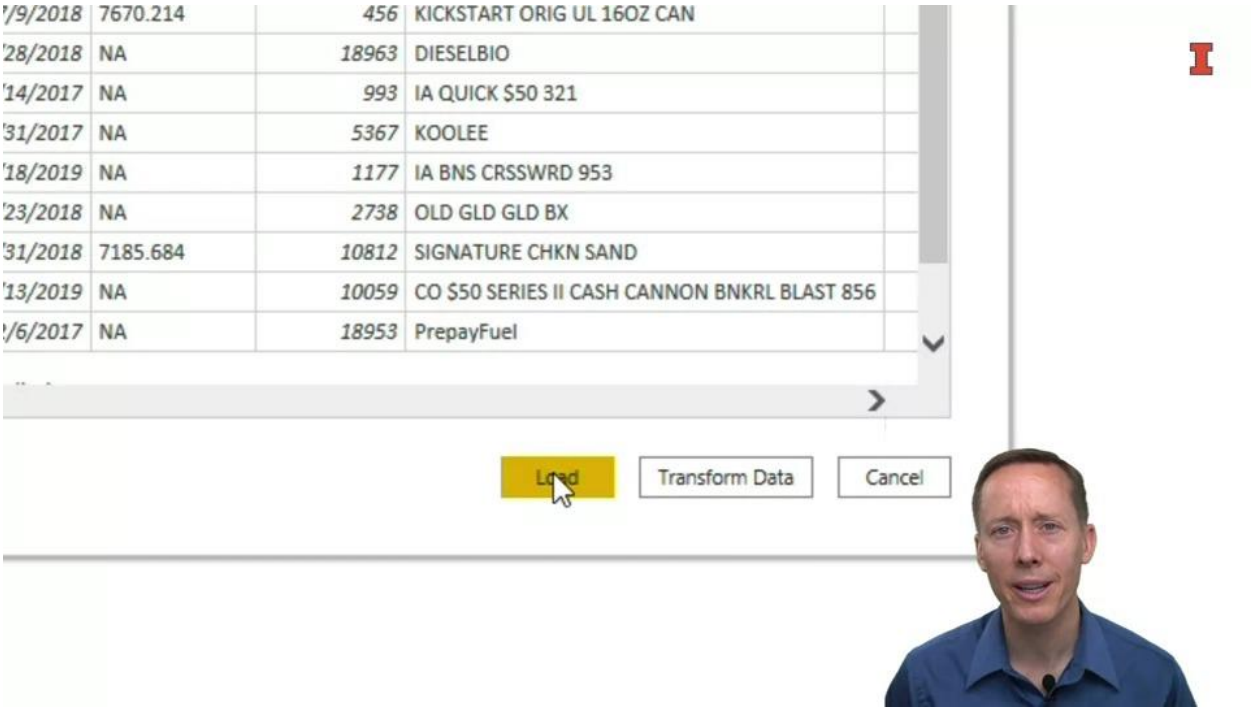


So there's a quick overview of the Power BI environment of course, as we get into working with data, all of this will become much more meaningful to you.

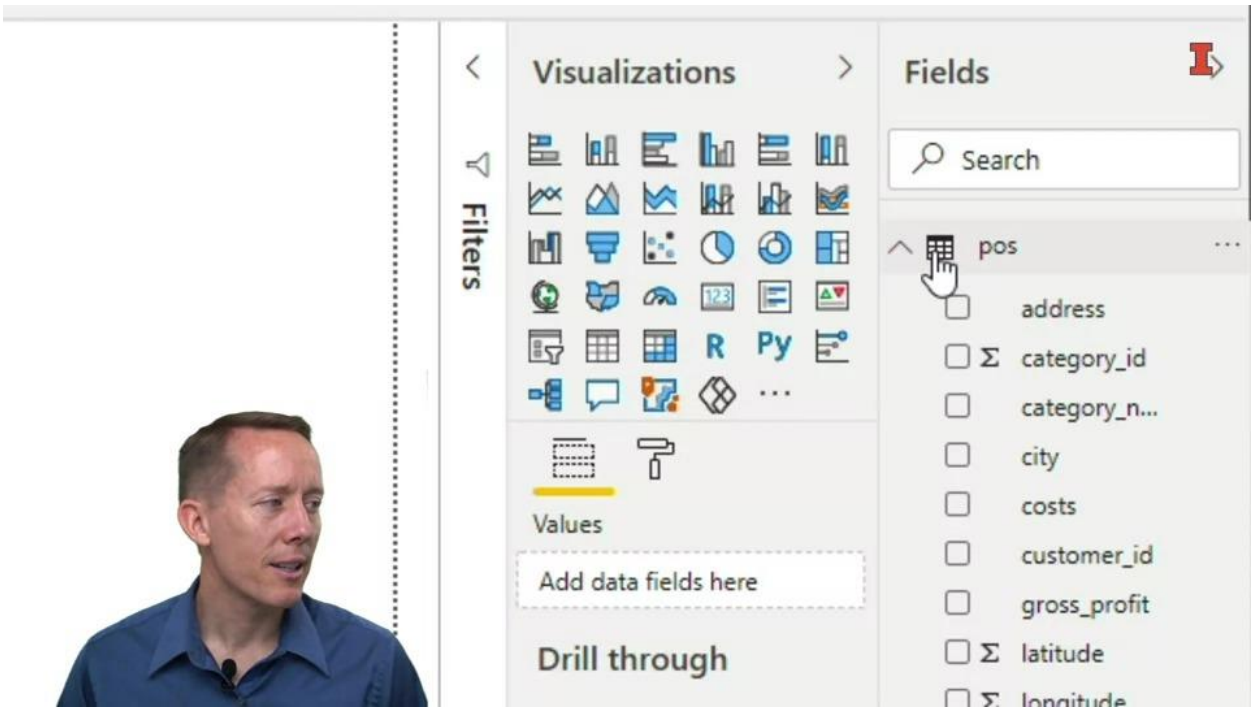
Lesson 2-2.3 ETL 1: Examine Data with Power Query Editor



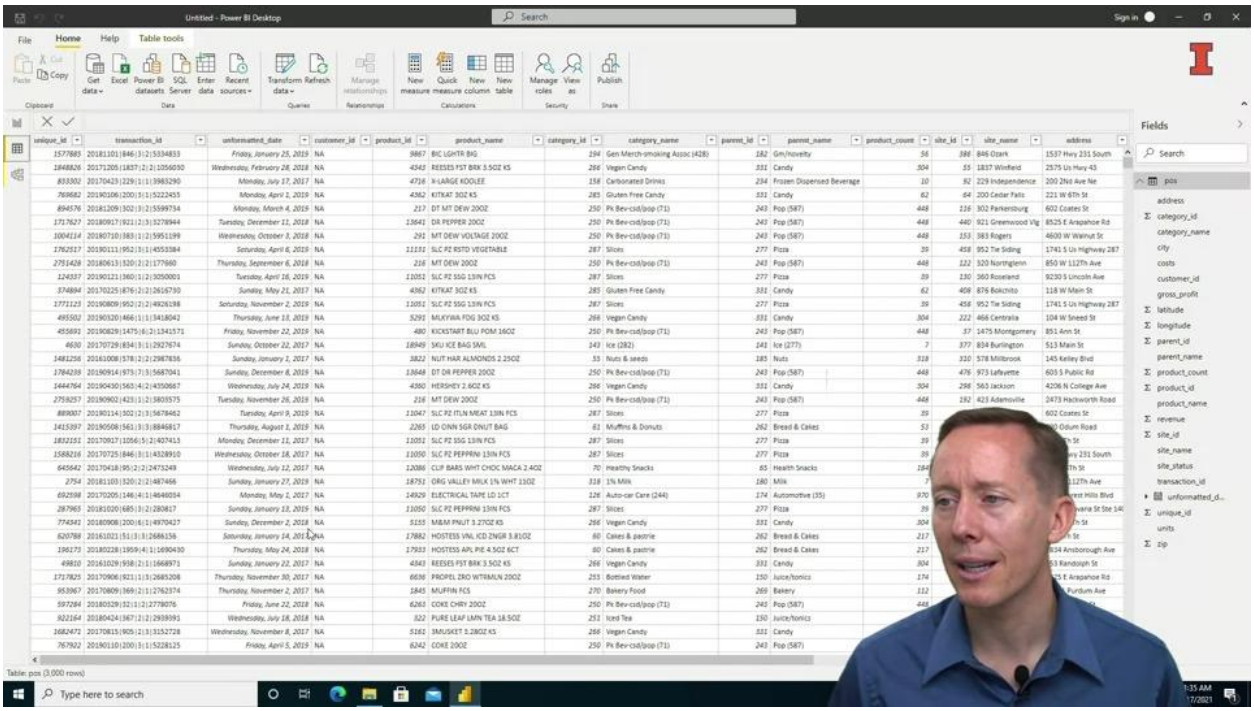
In this video, we want to show you how you can connect Power BI to a data source and then open that day to explore it and start making some transformations with it. So the first thing I need to do is start up Power BI so that I don't have to continue going to this menu and searching for it this way I'm going to pin it to my task bar and then just click on it from there. All right, and once Power BI has started up, I can access some things very quickly from here. But I want to show you how to do this without this. So I won't use that.



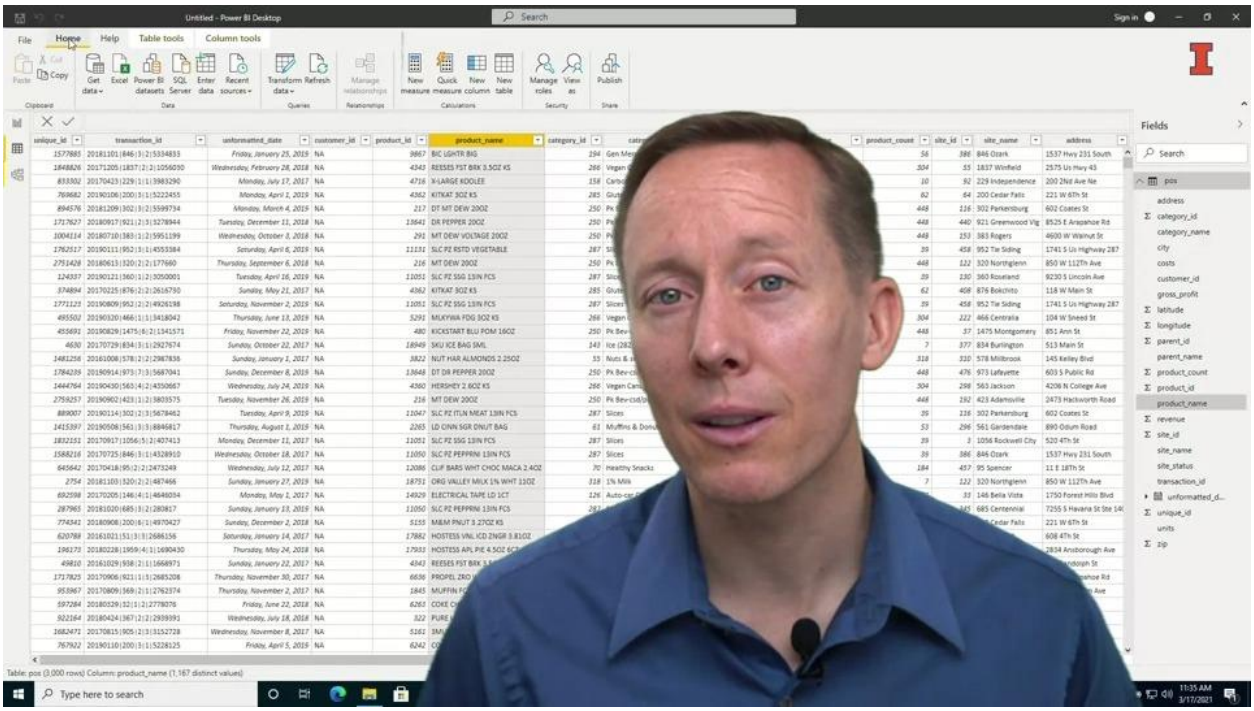
And so what I want to do is get some data. So I'm going to go to the home tab and click get data. Now, if I knew I wanted to connect to an Excel file, I could click on that Excel file. What type of file are we going to use? We're going to use a file, I'll show you where it is on this machine. I go to this tecaData folder, we're going to connect to this tecaPosSmall.csv file. So if you're following along, you want to download that data to your machine and then you'll point Power BI I to the folder where it's located. And so anyway, that's the file that I'm looking for. And since it's a csv file, I want to click on this get data icon. So this wizard pops up and I can see the different options for data sources that I can connect to. And there are a whole bunch of different data sources which is pretty cool. We're doing a very basic one which is text/CSV. So I'll click on that and then connect. And then I will navigate to where this file is located and there it is. TecaPosSmall.csv, click open. And I get this preview of the data and I can see the type of data that it's assuming this is and so we're assuming it's Western European data that's being eliminated with commas. And the data type detection is based on 200 rows, I could change that if I wanted to. We could also go directly to transforming the data from here. But I want to click load so that I can give you a quick overview of where the data is in the general Power BI environment.



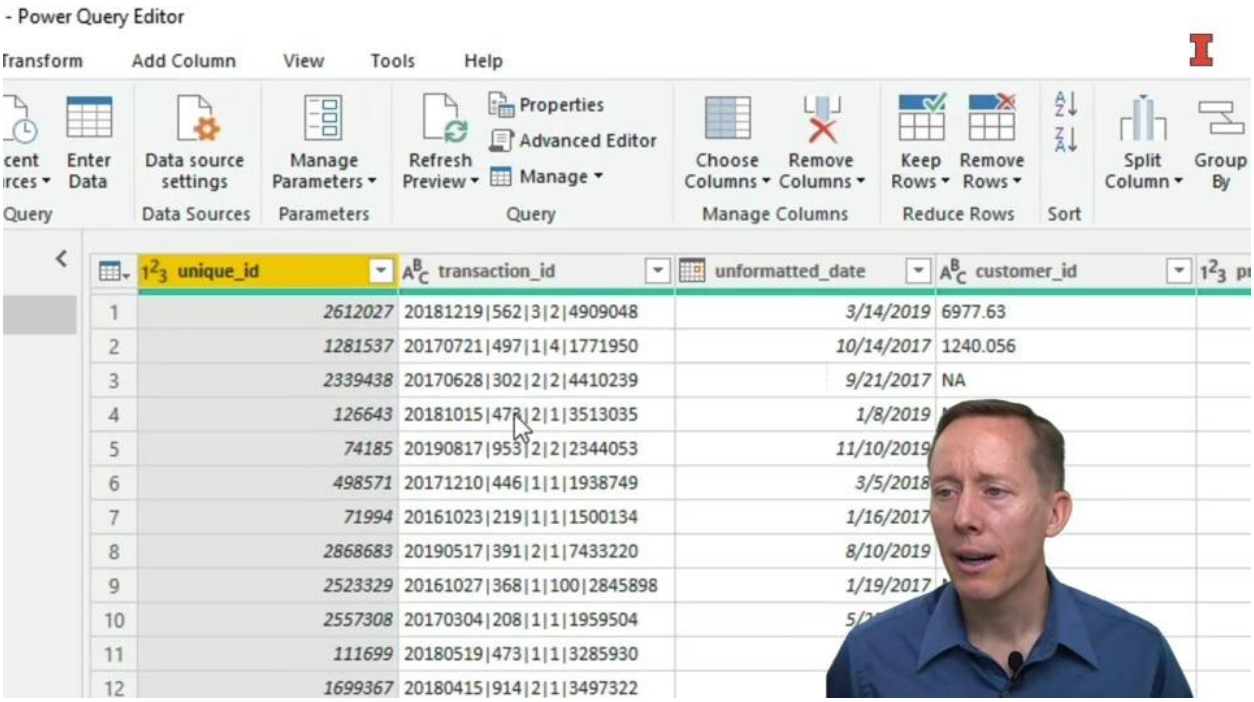
And then the first thing you'll notice is that on the far right hand side, you'll see that data set and a list of each column in that data set. So you can see that I've got the address, category_id and so forth. And you'll see that address doesn't have an icon next to it. Category_id has this summation icon next to it, which means it's a numeric field. So I've got several numeric fields. So, if it doesn't have an icon it means it's a text field or a character string. And I've got this field down here that has a calendar icon next to it, meaning it's a date field. Now from here, I could actually click on these three dots and do a lot of other things with this data. But the main thing I want to do from here is rename this by double clicking on it from tecaPosSmall to something simpler, just like pos, for point of sale data and click enter. And it does some updates. Now I could connect to another data source and if I did that, I would see that other data source listed here as well.



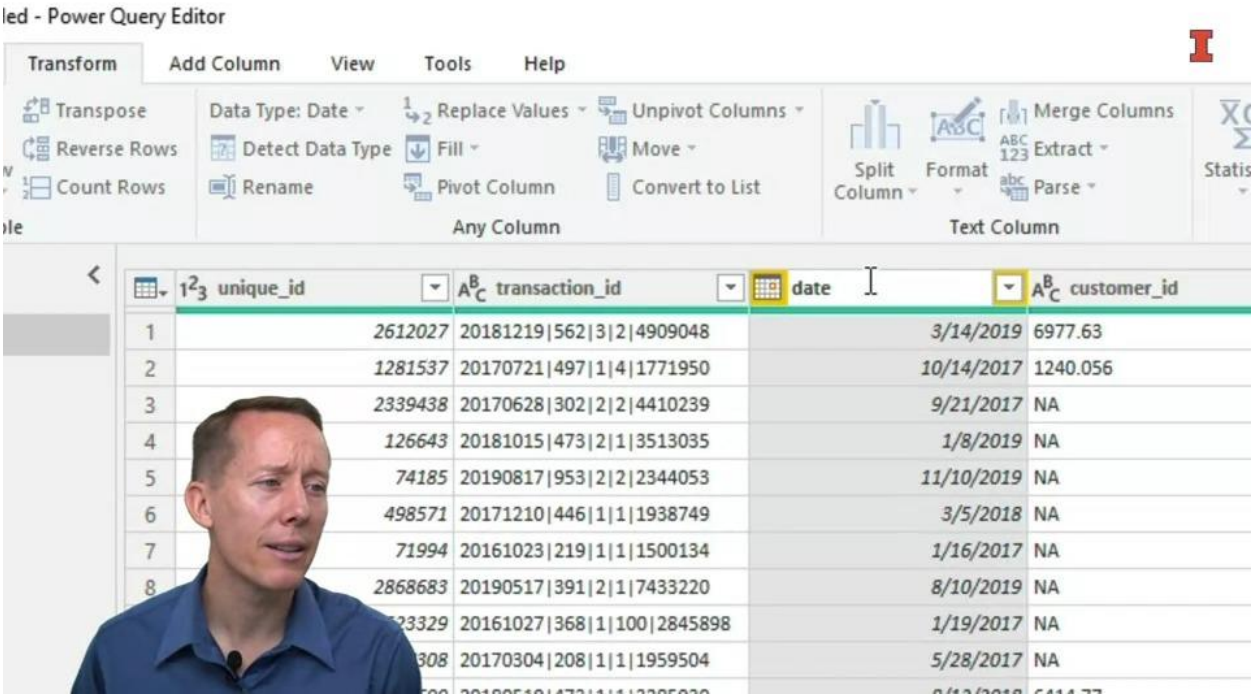
Now for a quick exploration of the data, I can go over to the left sidebar and click on this grid icon which is the data tab. And from here I can explore the data as if it were a worksheet in Excel. So it's got that very familiar feel to it. I can even click on this down arrow next to a column name and see the number of different unique values in that field. And I can do some filtering and sorting from here.



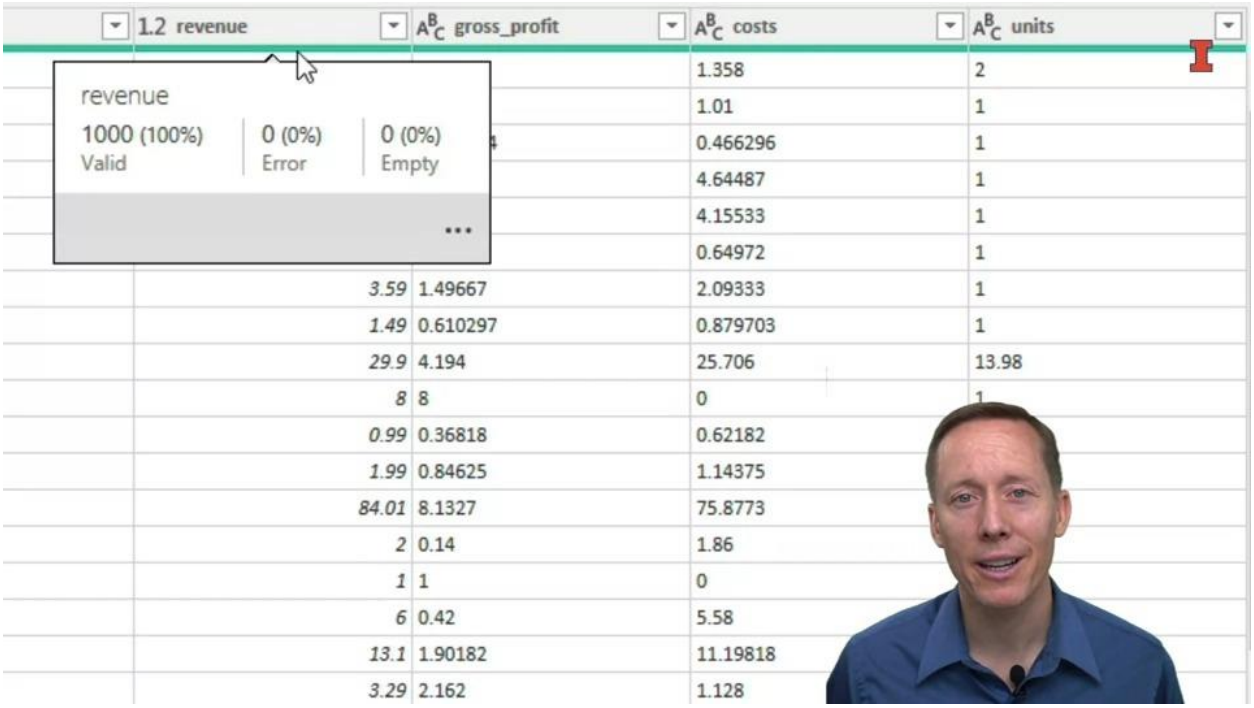
So in my mind this is more of a quick exploration or quick access to some basic tools for manipulating the data. What we want to do is really get into a setting where we can do a lot of transformations with the data.



So if we go to the home tab, click on transform data icon, this will bring up the Power Query Editor. Now the Power Query Editor is Microsoft's powerful tool for transforming data. And it brings up this window on top of Power BI and it's got a familiar feel to it, just like all Microsoft products. We've got this hierarchical relationship between the functions in here where we start with this menu icon along the top and then for each menu icon in the ribbon here we've got a grouping of different functions. And then on the left sidebar we've got a list of the different data sets. I've only read in one data set but you can see it's pos, point of sale. And then in the main area here we've got the data that we can explore. And then on the right side we've got the name of the data set. And what I think is super cool is this applied steps area? This will keep track of the steps that we make for transforming the data so that we have a log of what has been done. And we can undo any transformations that we have made to the data. All right, and then along the bottom we've got an overview of the structure of the data set and we also have that in the Power BI environment as well. We can see that I've got 3000 rows and anyway, it tells us some of the same information in the Power Query Editor. Where I've got 23 columns and over 999 rows of data. And then the profiling that we see in here is just based on the top 1000 rows and I'll come back to that in a little bit.

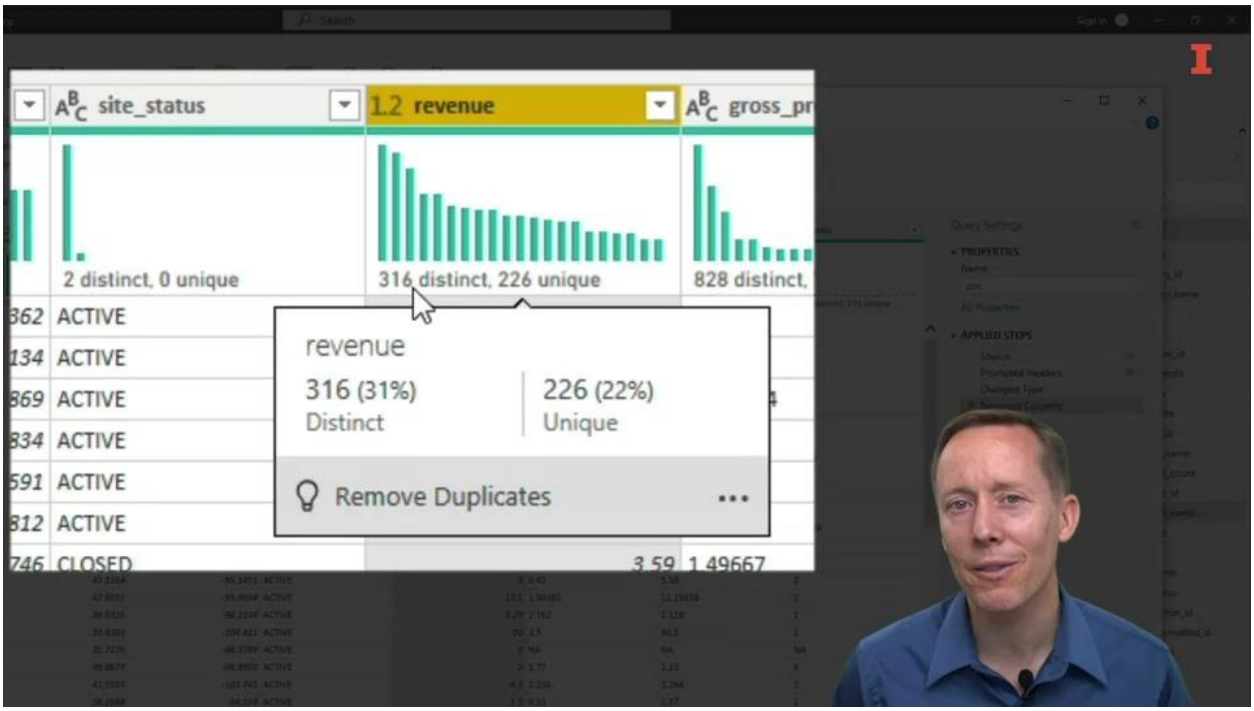


All right, so let's start looking at the data in here. We can see the different columns here I've got unique_id, I've got this icon next to it. And this means it is a numeric field and I can get more information about that if I click on the transform tab and I can see that the data type is actually a whole number. And if I wanted to change this to a different data type, I could simply click on that menu button and select one of these data types that make sense. Then I've got transaction_id. This data type, you can see it's ABC. That means it's a text or character string data type, unformatted date has a calendar icon. So that means it's a date data type and so on and so forth. Now let's say that I don't like the name of a particular column of data. I can simply double click on that and type in a new name and it will update.



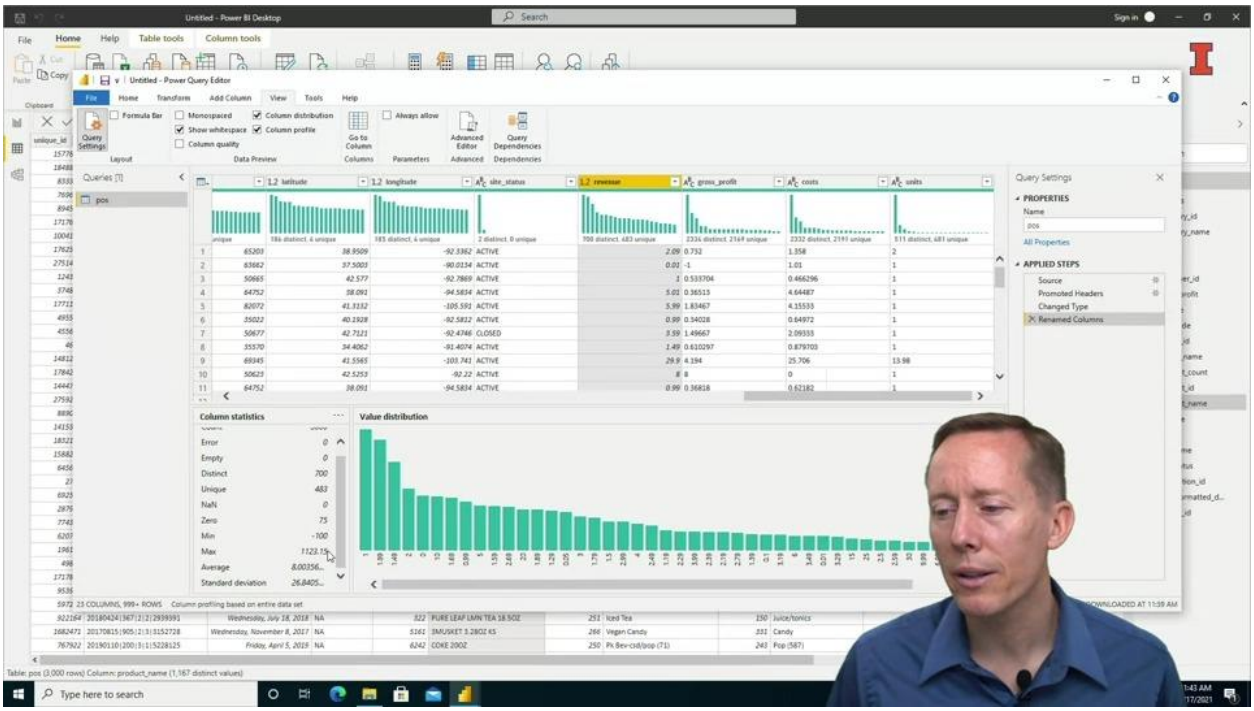
1.2 revenue	gross_profit	costs	units
		1.358	2
		1.01	1
		0.466296	1
		4.64487	1
		4.15533	1
		0.64972	1
3.59	1.49667	2.09333	1
1.49	0.610297	0.879703	1
29.9	4.194	25.706	13.98
8	8	0	1
0.99	0.36818	0.62182	
1.99	0.84625	1.14375	
84.01	8.1327	75.8773	
2	0.14	1.86	
1	1	0	
6	0.42	5.58	
13.1	1.90182	11.19818	
3.29	2.162	1.128	

I'll now scroll over to the numeric columns that I am really interested in which are revenue, gross_profit, costs and units. So below each column name, you'll see this green bar here and that tells us some information about the quality of that column of data and I'll come back to that in a minute.

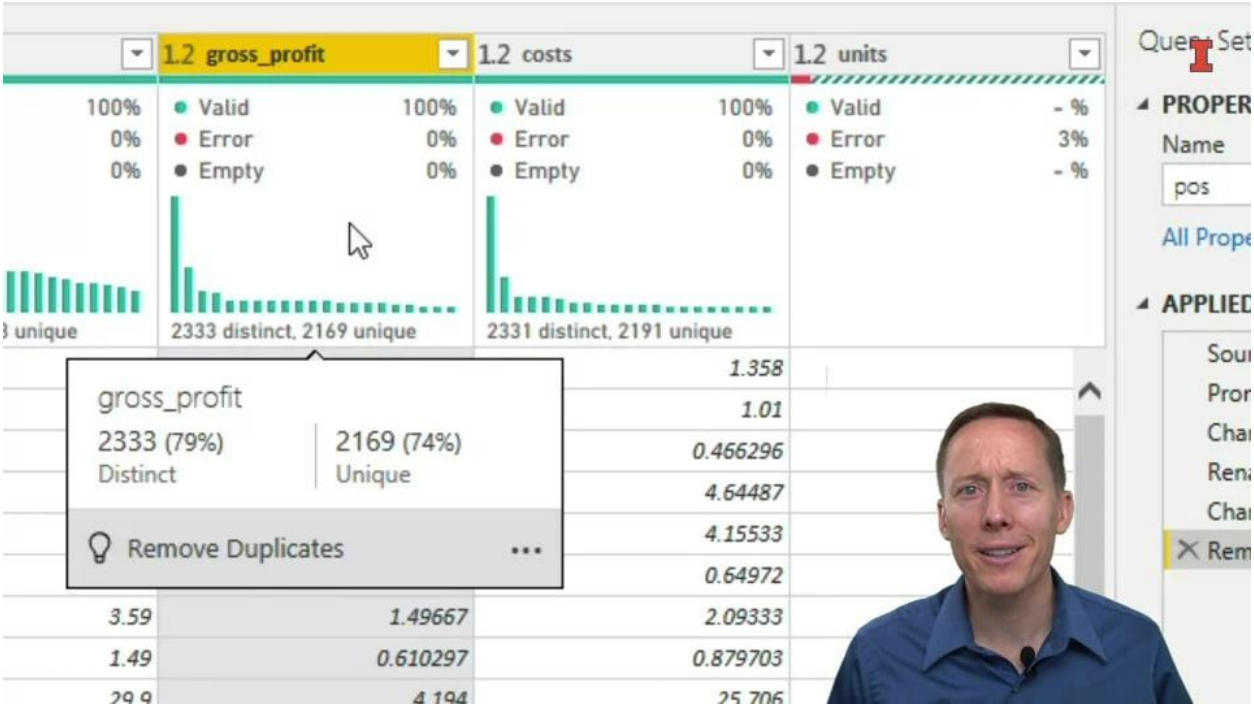


Let me just click on the revenue column here and let's go ahead and go to the view tab and let's click on column distribution. When we do that, you can see we get this really quick overview of the number of distinct values in a column and how frequently they occur. So it doesn't give us a bar for every unique value, but it does give us a bar for the top 20 or so unique values and it tells us that there are 316 distinct values and 226 are unique. What that means is that 226 values show up only one time.

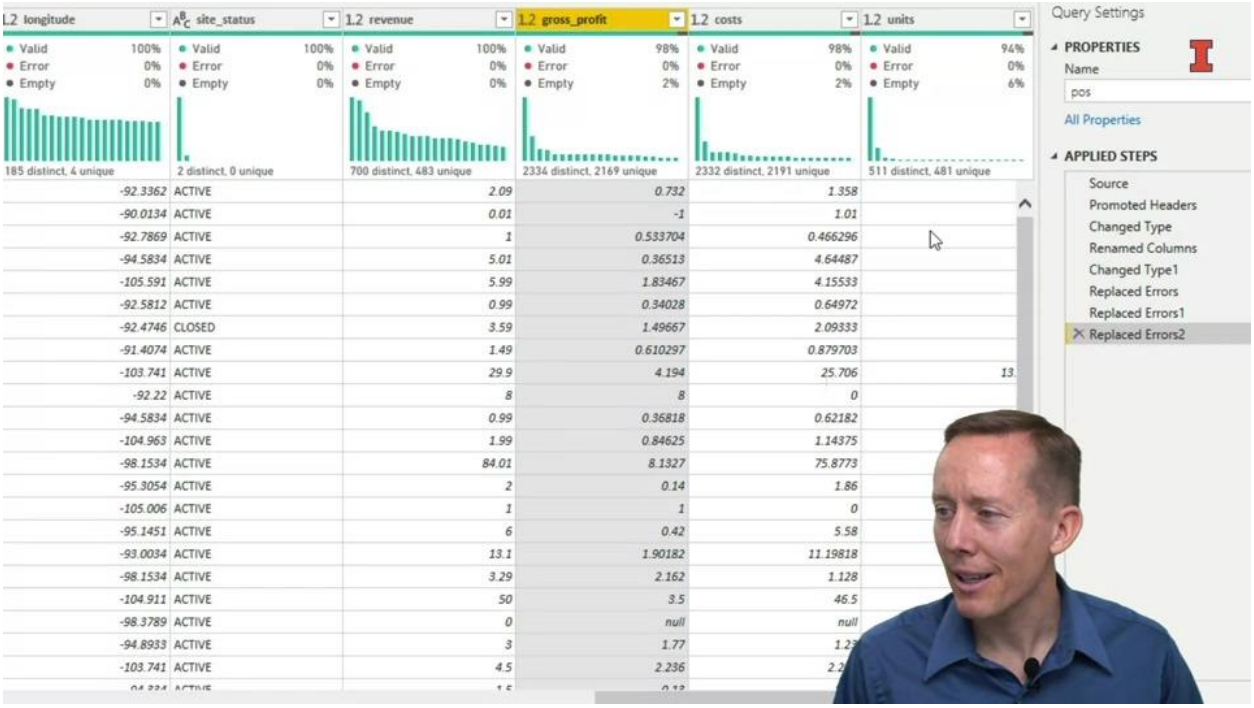
Now let's get a little bit more information and click on the column profile box here. And now down here in this bottom section, I can see more specifically what these values are. So I can see that a dollar and 99 shows up five percent approximately for these 1000 rows of data and how often these other values show up. This is convenient store data. So you would expect a lot of these numbers, these smaller numbers to show up associated with snacks and drinks and so forth. I can also see the column statistics. So I can see the number of rows of data that this is based on. So this is just based on 1000 rows, the number of error entries, empty values, distinct and unique again. NaN, so not a number of values, such as if you divide something by zero. Zero values, so zero shows up 32 times. The min is negative 20 the max is 1123.15, the average and standard deviation. Now this may be sufficient, but if you really want to get an idea for these numeric columns, what the min and max is. And we don't want to just base it off of a sample of 1000 rows.



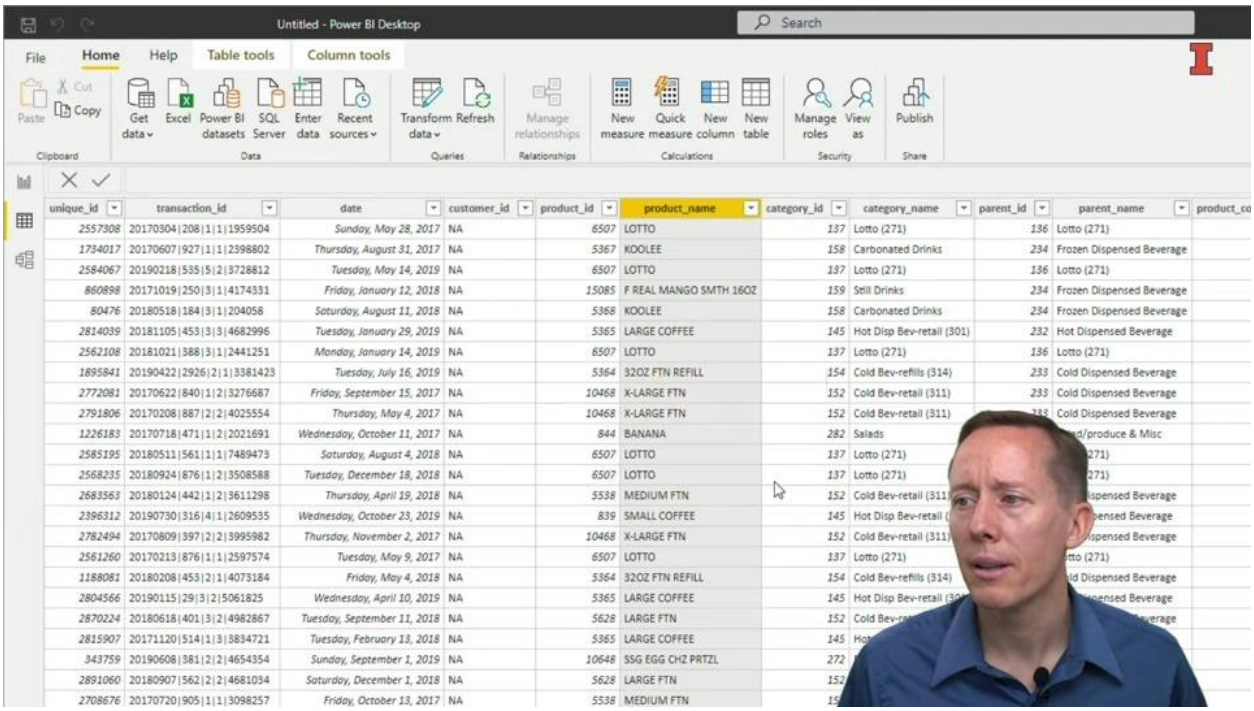
We can go ahead and change this column profiling to be based on the entire data set. And the trade off is that this will now give us more information, but it takes a little bit longer. In this instance here, I only have 3000 versus 1000 rows, it's not going to make much of a difference. If you're dealing with a million rows, then that would be a significant trade off in terms of time that you'd have to deal with. Anyway now I can see that I've actually got 700 distinct values, 483 are unique and the min is negative 100, the max is 1123.15. Alright, from here, let's go ahead and explore the quality of the state of a little more.



And we'll go up to the column quality box and check that and now along the top for each column, you can see the number and with the percentage of valid entries, error entries and empty values in here. Now as I look at revenue, it looks great. What about gross_profit? Gross_profit looks great. However, it is a text data type. So I want to change that. So if I click on gross_profit and go to transform, I can select a data type here. And I want this to be a decimal number. And when I do that it now says that there are 71 errors in here, or two percent are errors. And I could actually change the data type if I just click on that icon there and go to decimal number for costs, units, I will also change this to a decimal number. And now you can see that for all of these. I've got some values which are valid and others which are errors in here. And this bar along the top tries to indicate, send a quick signal about the quality of the colon. Now, what if I want to change? Maybe I just want to remove the errors, I can click on this remove errors button. And when I do that, it will remove all the rows that have errors in this gross_profit column. What if I decide that was a rush decision, I didn't actually want to remove those rows. I want to replace them with something else.



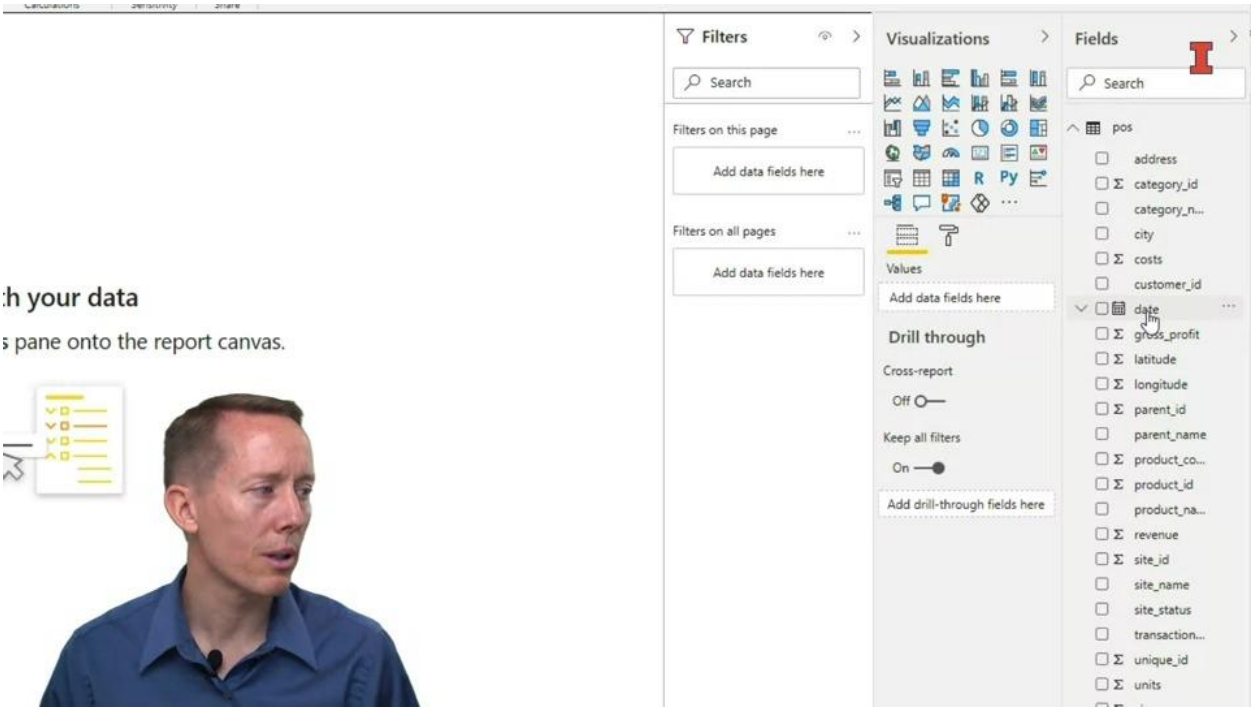
Well I could go over here to these applied steps and remember it's keeping track of everything I do. So I can see that I've renamed columns, I've changed the type and I've removed errors. So I'm going to just unclear or click that undo button and now I've got these errors that are back in the gross_profit column. And if I decided hey, I want to do something else with these missing values, these error values, I want to replace the errors with something else. I can do that, click on replace errors and I want to replace them with null. Now null means it's missing. And you'll see down here that it just replaced that error with a null. So I could go ahead and clean up these other columns as well. I'll go ahead and do that for these remaining columns.



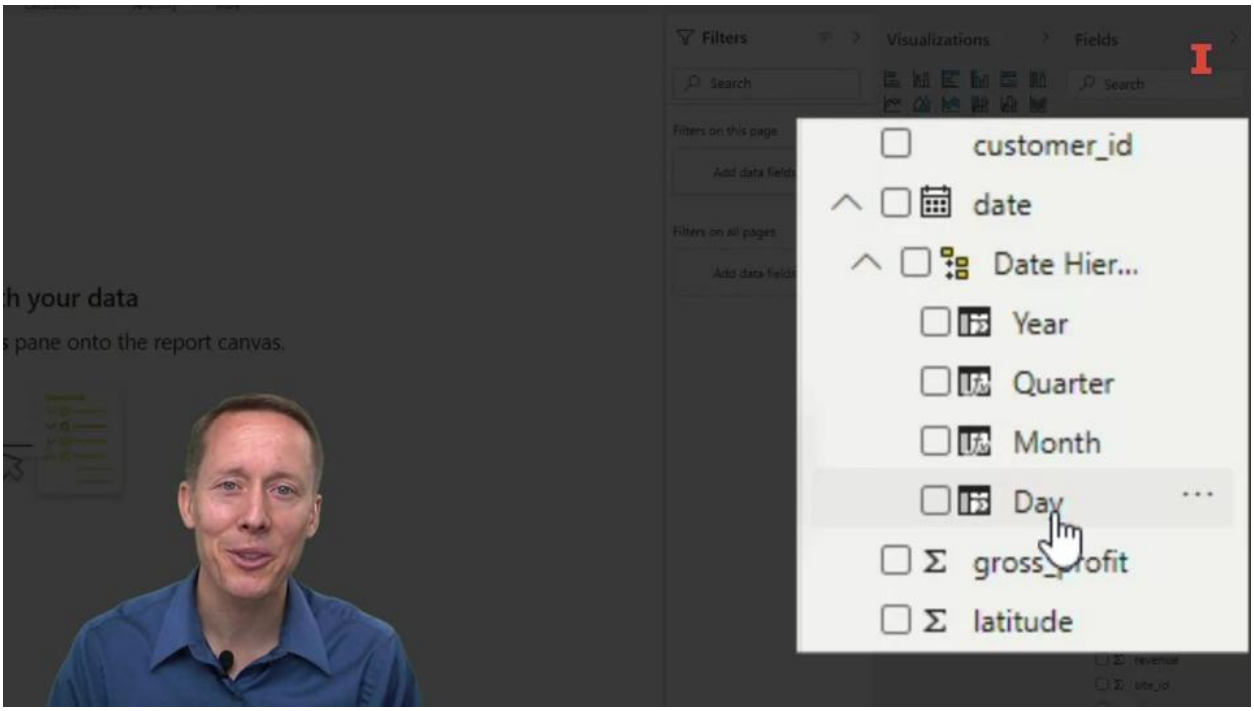
unique_id	transaction_id	date	customer_id	product_id	product_name	category_id	category_name	parent_id	parent_name
2557308	20170304|208|1|1|1959504	Sunday, May 28, 2017	NA	6507	LOTTO	137	Lotto (271)	136	Lotto (271)
1734017	20170607|927|1|1|2398802	Thursday, August 31, 2017	NA	5367	KOOLEE	158	Carbonated Drinks	234	Frozen Dispensed Beverage
2584067	20190218|535|5|2|3728812	Tuesday, May 14, 2019	NA	6507	LOTTO	137	Lotto (271)	136	Lotto (271)
860898	20171019|250|3|1|4174331	Friday, January 12, 2018	NA	15085	F REAL MANGO SMTH 16OZ	159	Still Drinks	234	Frozen Dispensed Beverage
80476	20180518|184|3|1|204058	Saturday, August 11, 2018	NA	5368	KOOLEE	158	Carbonated Drinks	234	Frozen Dispensed Beverage
2814039	20181105|453|3|3|4682996	Tuesday, January 29, 2019	NA	5365	LARGE COFFEE	145	Hot Disp Bev-retail (301)	232	Hot Dispensed Beverage
2562108	20181021|388|3|1|2441251	Monday, January 14, 2019	NA	6507	LOTTO	137	Lotto (271)	136	Lotto (271)
1895841	20190422|2926|2|1|3381423	Tuesday, July 16, 2019	NA	5364	32OZ FTN REFILL	154	Cold Bev-refills (314)	233	Cold Dispensed Beverage
2772081	20170622|840|1|2|3276687	Friday, September 15, 2017	NA	10468	X-LARGE FTN	152	Cold Bev-retail (311)	233	Cold Dispensed Beverage
2791806	20170208|887|2|2|4025554	Thursday, May 4, 2017	NA	10468	X-LARGE FTN	152	Cold Bev-retail (311)	233	Cold Dispensed Beverage
1226183	20170718|471|1|2|2021691	Wednesday, October 11, 2017	NA	844	BANANA	282	Salads		d/produce & Misc
2585195	20180511|561|1|1|7489473	Saturday, August 4, 2018	NA	6507	LOTTO	137	Lotto (271)		271)
2568235	20180924|876|1|2|3508588	Tuesday, December 18, 2018	NA	6507	LOTTO	137	Lotto (271)		271)
2683563	20180124|442|1|2|3611298	Thursday, April 19, 2018	NA	5538	MEDIUM FTN	152	Cold Bev-retail (311)		ispensed Beverage
2396312	20190730|316|4|1|2609535	Wednesday, October 23, 2019	NA	839	SMALL COFFEE	145	Hot Disp Bev-retail (		pensed Beverage
2782494	20170809|397|2|2|3995982	Thursday, November 2, 2017	NA	10468	X-LARGE FTN	152	Cold Bev-retail (311)		ispensed Beverage
2561260	20170213|876|1|1|2597574	Tuesday, May 9, 2017	NA	6507	LOTTO	137	Lotto (271)		tto (271)
1188081	20180208|453|2|1|4073184	Friday, May 4, 2018	NA	5364	32OZ FTN REFILL	154	Cold Bev-refills (314)		ld Dispensed Beverage
2804566	20190115|29|3|2|5061825	Wednesday, April 10, 2019	NA	5365	LARGE COFFEE	145	Hot Disp Bev-retail (3		ispensed Beverage
2870224	20180618|401|3|2|4982867	Tuesday, September 11, 2018	NA	5628	LARGE FTN	152	Cold Bev-		verage
2815907	20171120|514|1|3|3834721	Tuesday, February 13, 2018	NA	5365	LARGE COFFEE	145	Hot		
343759	20190608|381|2|2|4654354	Sunday, September 1, 2019	NA	10648	SSG EGG CHZ PRTZL	272			
2891060	20180907|562|2|2|4681034	Saturday, December 1, 2018	NA	5628	LARGE FTN	152			
2708676	20170720|905|1|1|3098257	Friday, October 13, 2017	NA	5538	MEDIUM FTN	15			

All right, now that I've cleaned up those numeric columns, I feel pretty good about where things are at and so what if I want to close my project and apply these changes? I can't simply click on this X box here. What I need to do is go to the home tab here and click close and apply and it will apply those changes to the data and then close out of that and then it will update any visualizations that I may have that are based on that data. And at this point if I am done with my session of Power BI I can save this. As a specific file and I'll save it as etl1 and click save. And I will now have access to this when I want to use Power BI. Next time I can quickly open up and it will preserve these changes. And to close out Power BI I'll click on this X icon at the top. That's hopefully a quick introduction to how you can start using Power BI to making transformations to your dataset.

Lesson 2-2.4 ET2: Dates and Calculated Columns

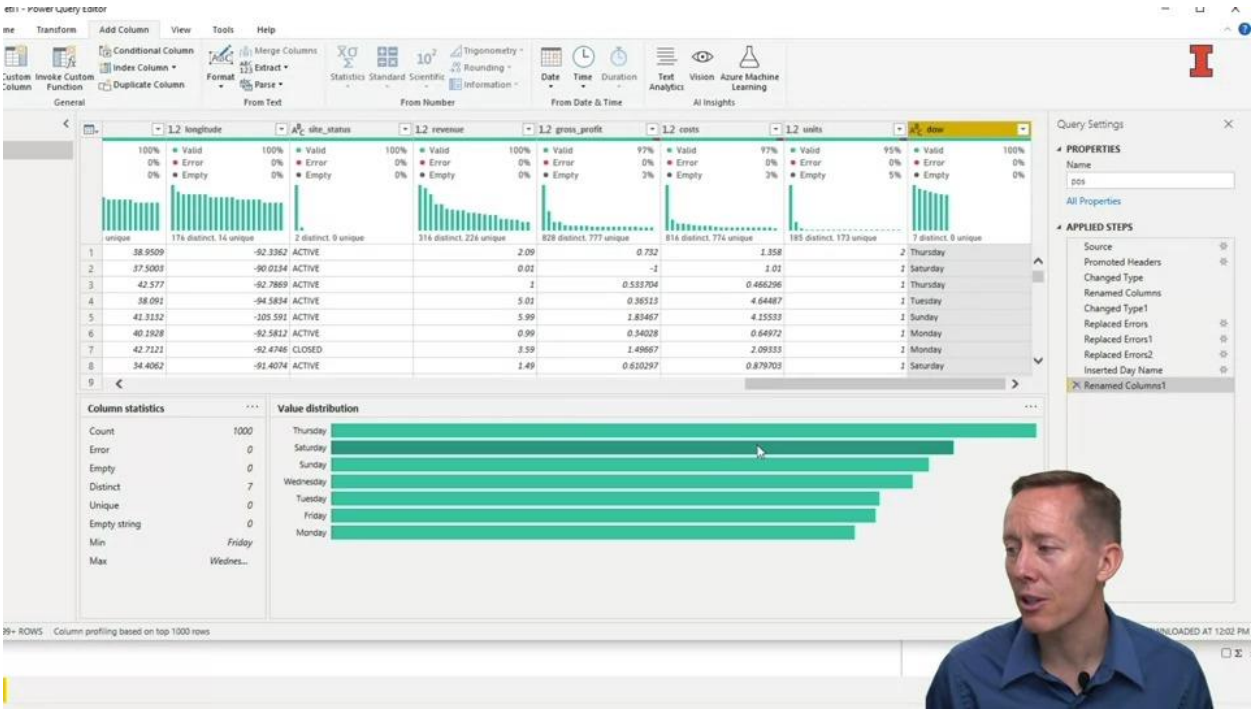


In this video, we will focus on dealing with dates and calculated columns in Power BI. Let's go ahead and start up power BI. At this point, let's go ahead and just go with the suggestion to open this recent Power BI file, PBIX file, ETL1. Double click on that, and it will load the data and any other reporting that I have done with this file. Now that the data is loaded, let's just go to the right sidebar and explore this date column here.

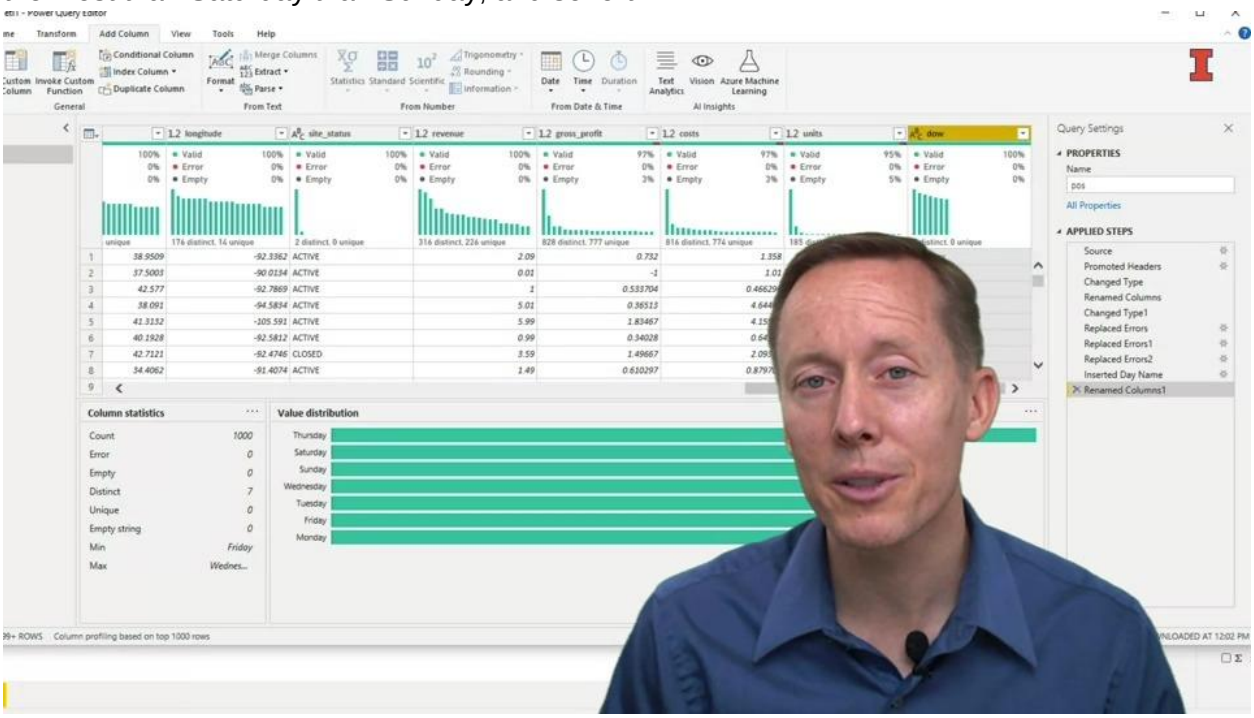


You'll notice that it has this drop-down arrow next to it. Now, Power BI is really awesome at dealing with dates. It does a lot of things automatically. If I click on that down-arrow, I see this date hierarchy column, and click on that down-arrow and you can see that I've got a year, quarter, month and day that I could use for creating visualizations with my data. Now this is really cool because it's not a new column that has been created, and that's important because it doesn't make our data any bigger by adding new columns. It just is a quick measure in here that it can use as if it were a new column, and then Power BI will make those transformations. If I wanted to create a chart that shows maybe revenue by year, I can do that, or by quarter or by month or by day. Now, this day is not day of the week, it's day of the month. Numbers one through 31.

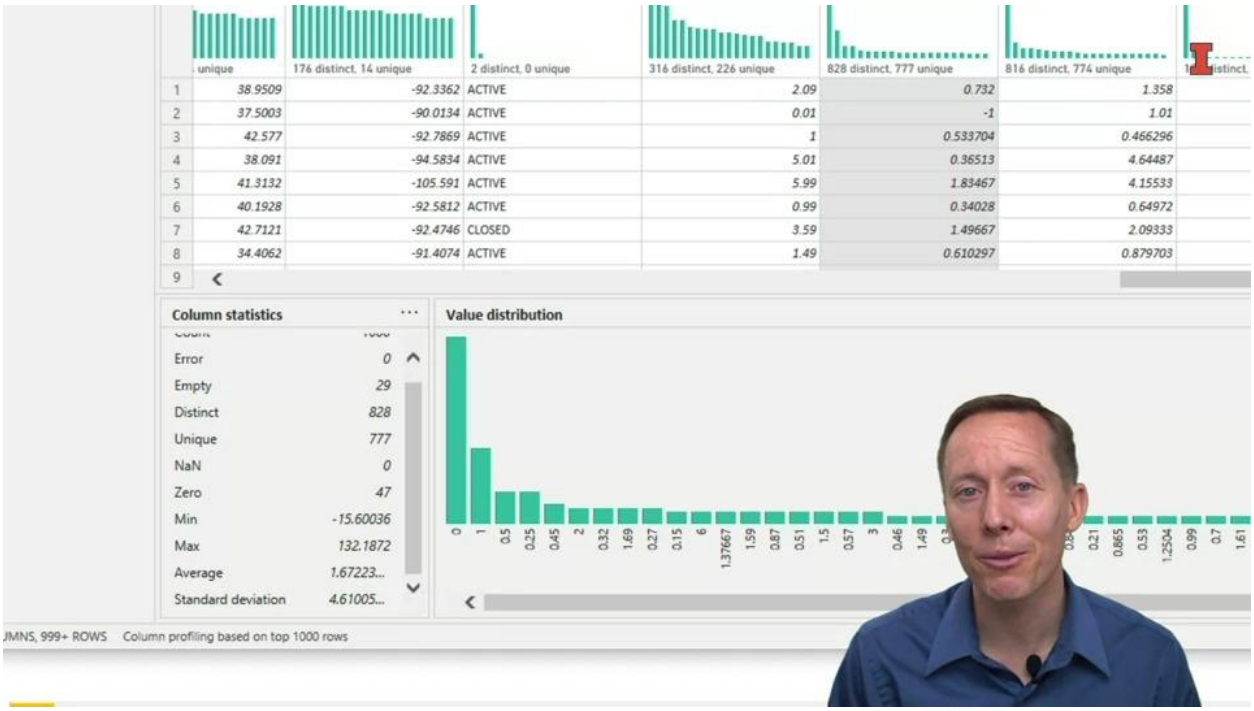
What if I did want to create another column such as day of the week? I can definitely do that. The way I'll do it is I'll illustrate that by going into the power query editor, by clicking on this transform data icon. Now I can click on the date column here, and if I go to the add column menu item at the top, I can go to the column from example's button. I just want to do it for this column from selection, and now if I double click on this new column here, you can see it's pointing to the date column here, it gives me a list of possible new columns that I might want to create from this date column here. I've got the date in this format 3/14/2019. The month, the year format, the number of days that have passed from a specific date, the day of the month, day of the week, Thursday, and a bunch of other things, maybe the first day of the month, first day of the quarter, just the months name, a lot of different options that I can choose from here. Now I want day of the week, so I'm going to click Thursday, double click that and then click "Okay". You can see I've got this new column day and name, and I've got the day of the week for every observation. I'm going to simplify this and change the state name to dow, make it a little bit simpler.



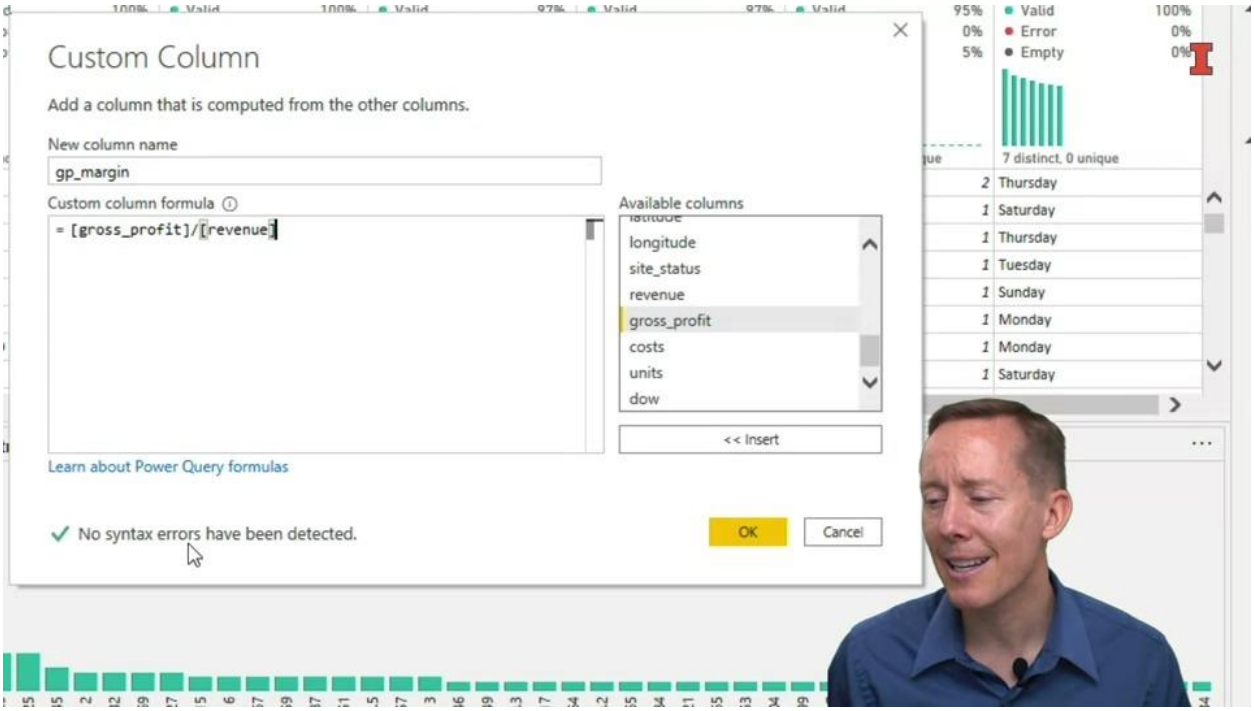
You notice that when I click on this here or when I just look at these quick summary information here at the top, I can see that there are seven distinct values, zeros are unique, meaning that all of them show up more than once. Down here at the bottom, I can see that Thursday shows up the most than Saturday than Sunday, and so forth.



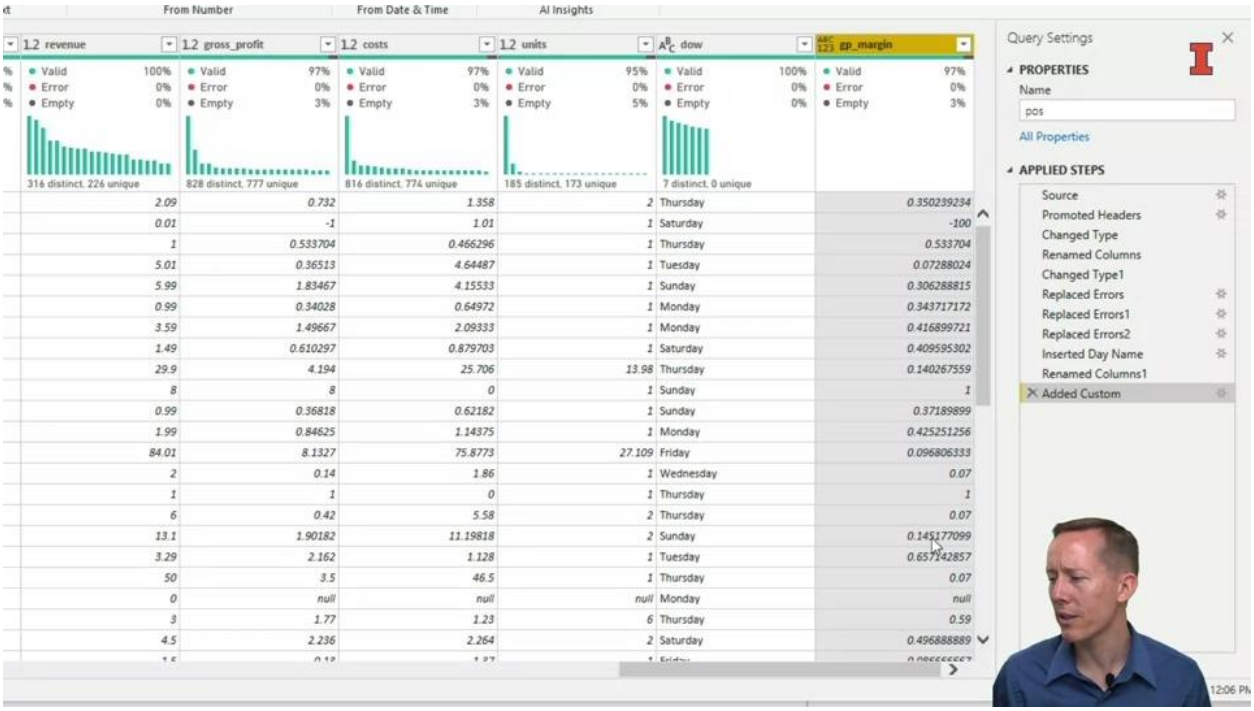
That's pretty awesome. Power query is very efficient at dealing with dates. Now let's illustrate how to create a calculated column based on two or more columns in the data set.



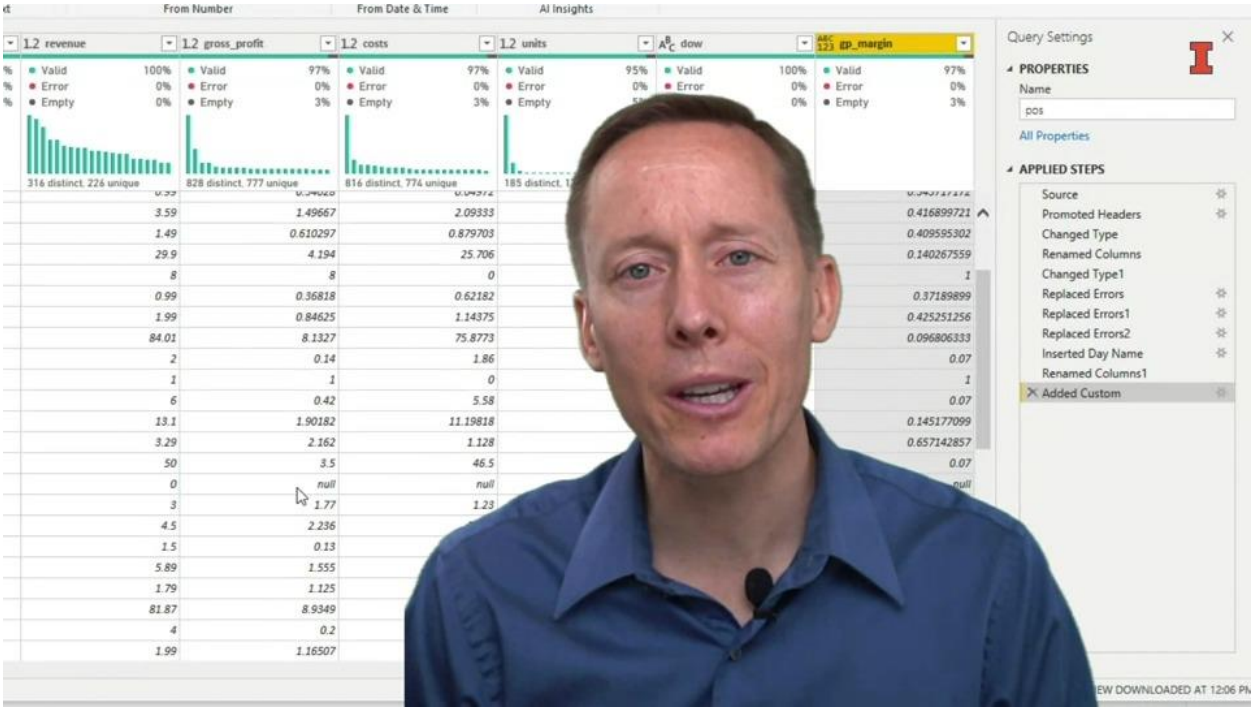
If I look at the different numeric values, a couple of that are relevant here would be revenue and gross profit. As I explore the data, I can see that gross profit is a number that ranges from negative 15-132. This tells me it's not a percentage number. What if I want to create a gross margin or at gross profit as a percent of revenue? I can easily do that.



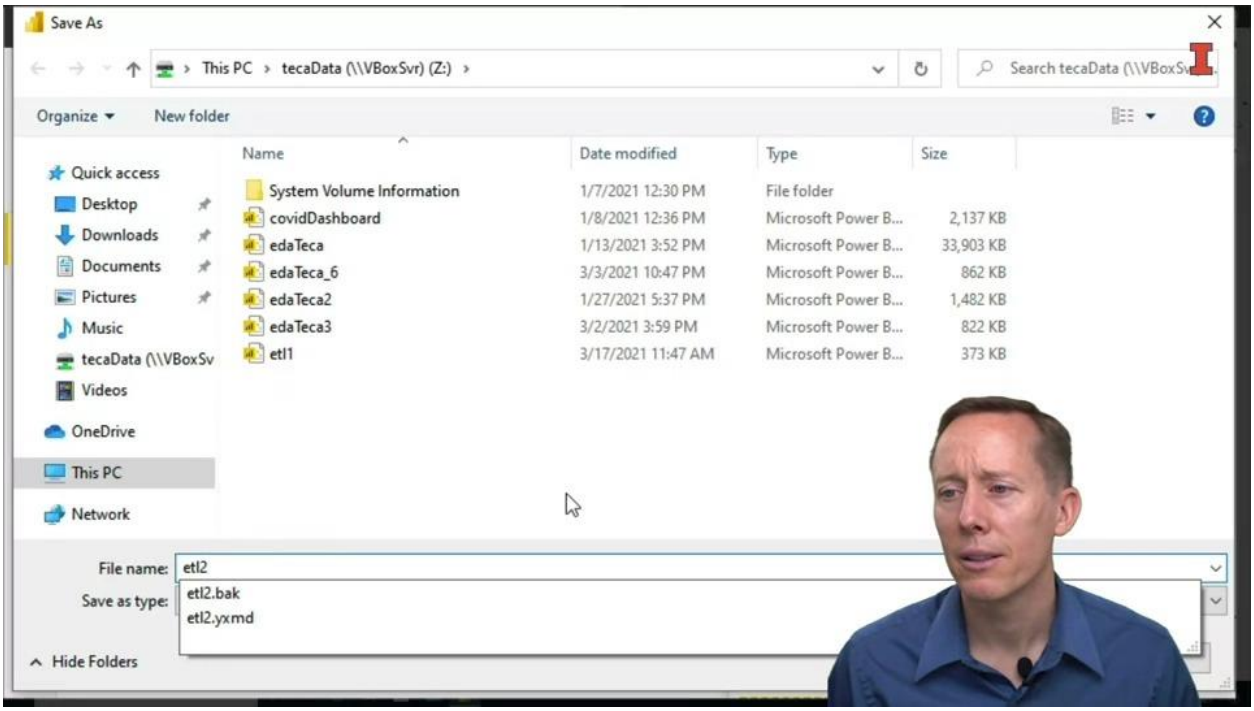
The way I would do that, is I would go in this add column tab, or go to this custom column here and click on that, and it will bring up a wizard that will allow me to type in a formula for creating a new column. Now, first thing I'll do is I'll give this a new column name and I'll label it gp_margin. Then from here, I can go down into this custom column formula box and start typing things. If I just type a letter d, it'll bring up a suggestion of possible functions that I might want to use on a column of data. Now, I don't want to use this function here, but I think that's important to recognize that it has this auto complete feature built into it. What I want to do, is I want to calculate gross profit as a percent of revenue. I'm actually going to go down to the available columns in here, and you can see they're ordered in terms of how they show up in the data set, not alphabetical. I'll double click gross profit, and the way it indicates a column in this formula is it simply put square brackets around it. If I want to take gross profit and divide it by revenue, I can click on the forward slash arrow. At this point you see this red squiggly line and this red line over here, which is indicating that there's an error with this formula. If I were to play this now, it actually won't let me, if I were to do that, it had just result in an error. I could navigate to revenue here. But let me just show you, if I know the column name, I can simply click the open square bracket and I'll lists all the columns in here. If I start typing revenue, it will narrow it down to the column that I'd want. Anyway, there's revenue, gross profit divided by revenue, there's no longer a syntax error. I can go ahead and click "Okay".



There I've got this new column, gp_margin, and I can see that there are 97 percent valid entries, no errors and three percent that are empty. If I want to see the empty values, I could scroll down and look at them and I can see that I've got no here. That's one of the empty values. Why is it null? I might want to understand why. Well, it's because the gross profit value is null here, and so null divide by zero is going to just give us null.



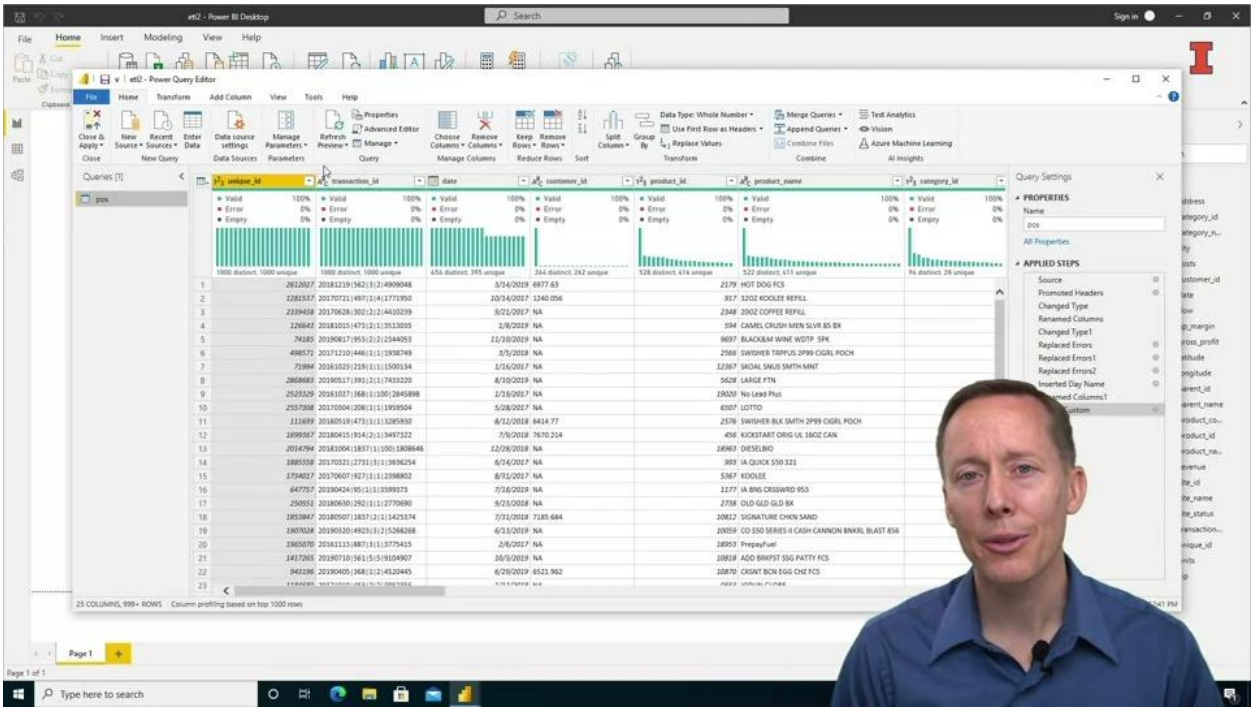
There you go. That's how you can very easily deal with date columns and create calculated columns in power query editor.



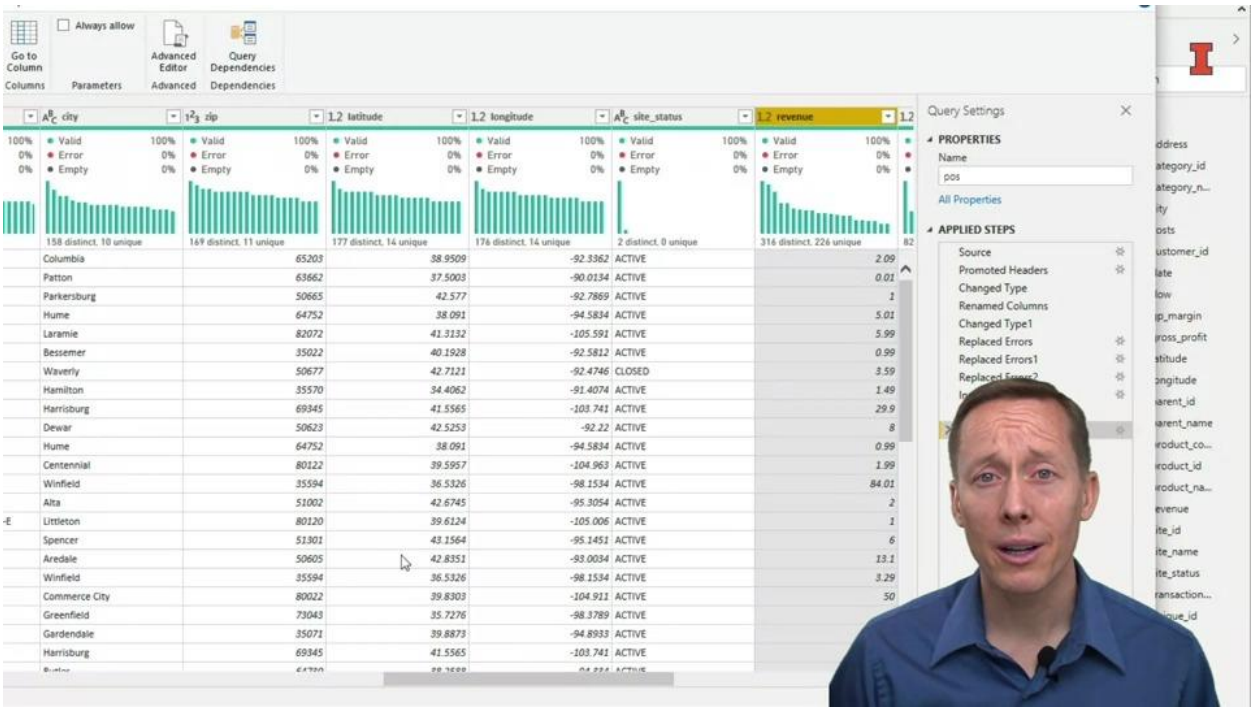
I'll go ahead and save this. I'll close and apply this. Once those changes have been applied, I can go over and verify that they're here. If I look at the different columns, I've got now the dow, for day of week and I also have gp_margin here that I just created, which I can now use to create visualizations. Now that I've done that, I will save this file. I could just click save, but I want to save it as a new file, and so I will save it as etl2, to separate what we did from a prior lesson. Click "Save" and then exit out of Power BI.

There's an example of how you can deal with date columns and Power BI, as well as create calculated columns.

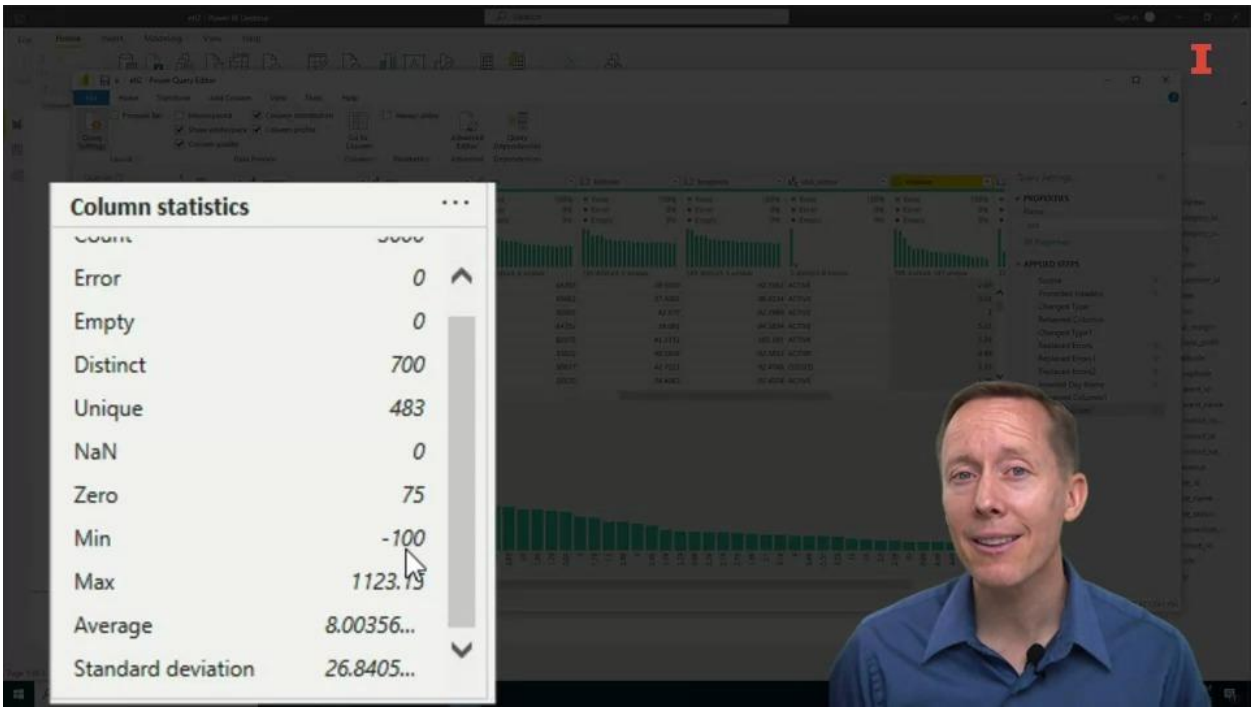
Lesson 2-2.5 ETL 3: Checking for and Eliminating Outliers



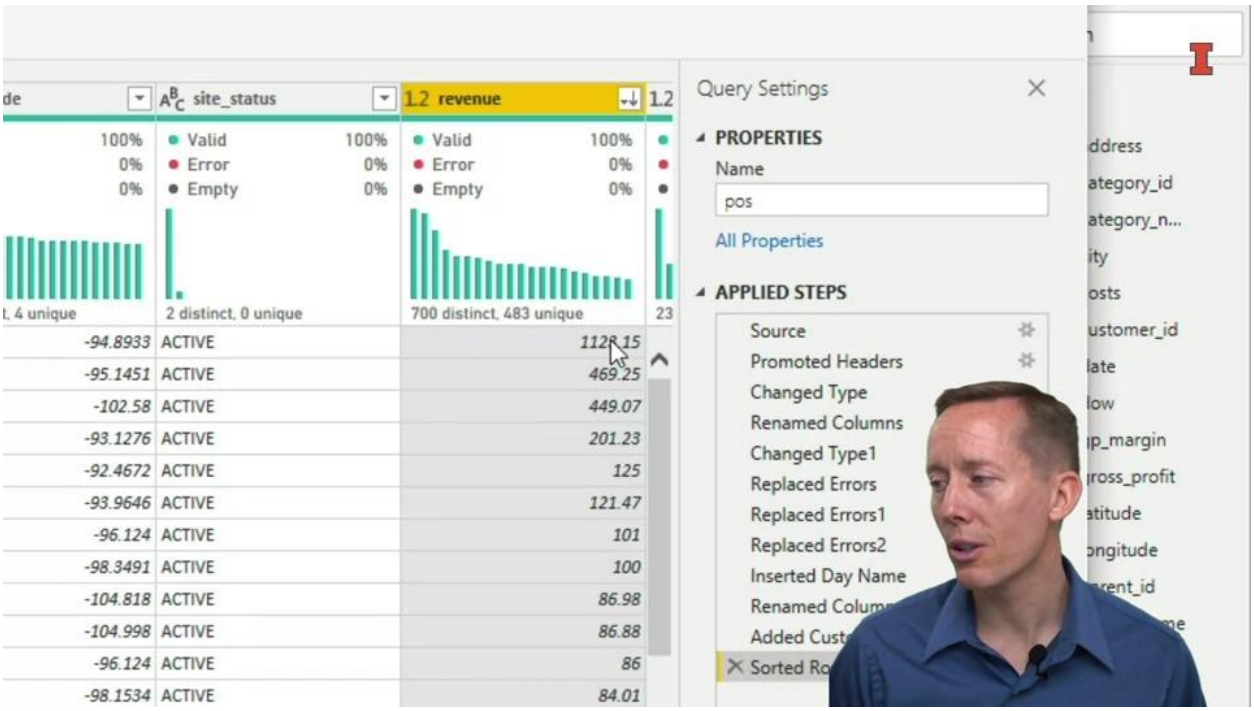
In this video, we want to show you how to check for outliers in both numeric and categorical columns and then remove outliers or deal with them in some other way. The first thing I'll do is once I open up Power BI as I will select the report that I've been working on, etl2, and once it has loaded the data and any visualizations that I may have I will click on the transform data icon, and now we are in the power query editor. Now, I want to focus first on looking at the revenue column because that is a column that we really want to understand more.



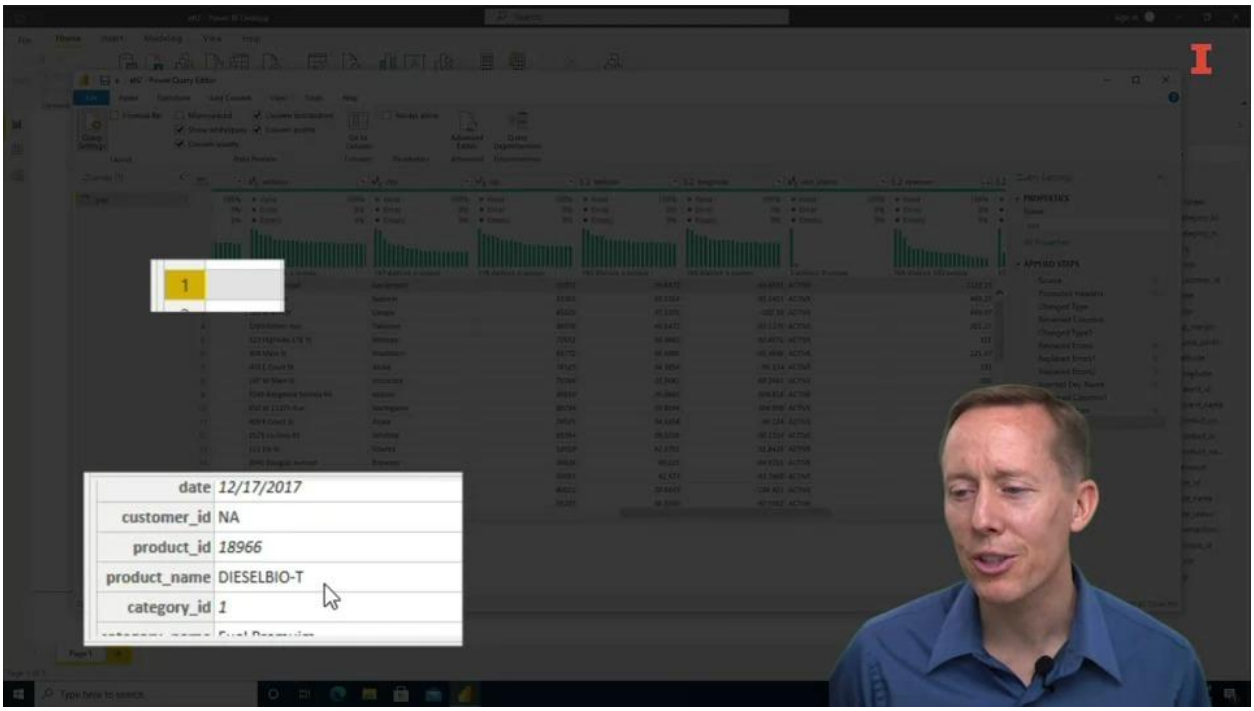
So to find that column, I know that it's on the far right, but I'll show you another thing that you can do. If you've got a wide data set and you don't know where the column is located, if you go into the View tab, you can click on go to column and then you can either scroll through the different columns or just start typing in its name and it will pop up. You can click on it, click "Okay", and it will move over, scroll to where that column is located. Now, let's explore whether or not there are outliers in the revenue column.



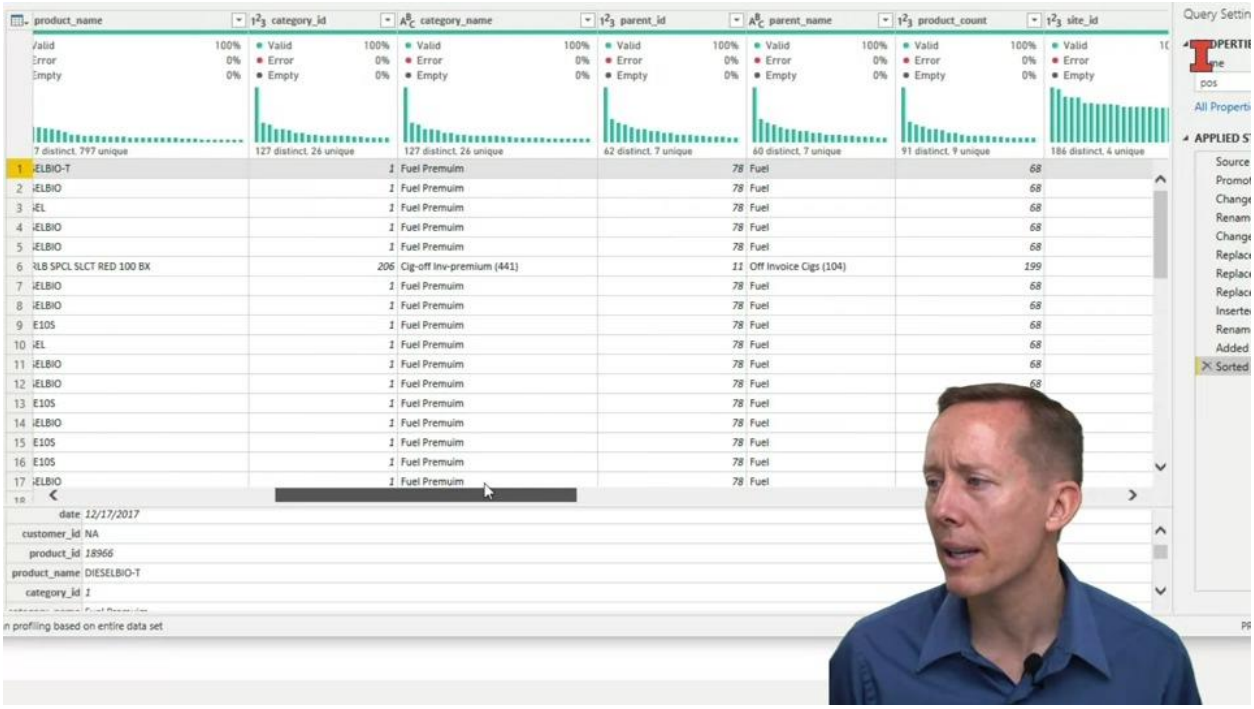
So I'll click on that column header and make sure that I have the column profile selected, and also make sure that I'm profiling this on the entire data set. Now I can look at the Min, max, average, and standard deviation to get an idea of whether or not there are outliers. So let's first focus on the average. The average is eight, the standard deviation is about 26, and if this were a normally distributed column of data then any outliers are typically either above three standard deviations above the mean, or three standard deviations below the mean. So if I take 26 and multiply it by 3, that's about 78 and then add that to the average of 8, that means anything that's above about 86 would be considered an outlier, and if I take that 78 and subtract it from 8, that means anything below a negative 70 would be considered an outlier as well. So now I can compare those two numbers, negative 70 and 86 to the mean and the max and see that indeed there are observations that extend beyond those points.



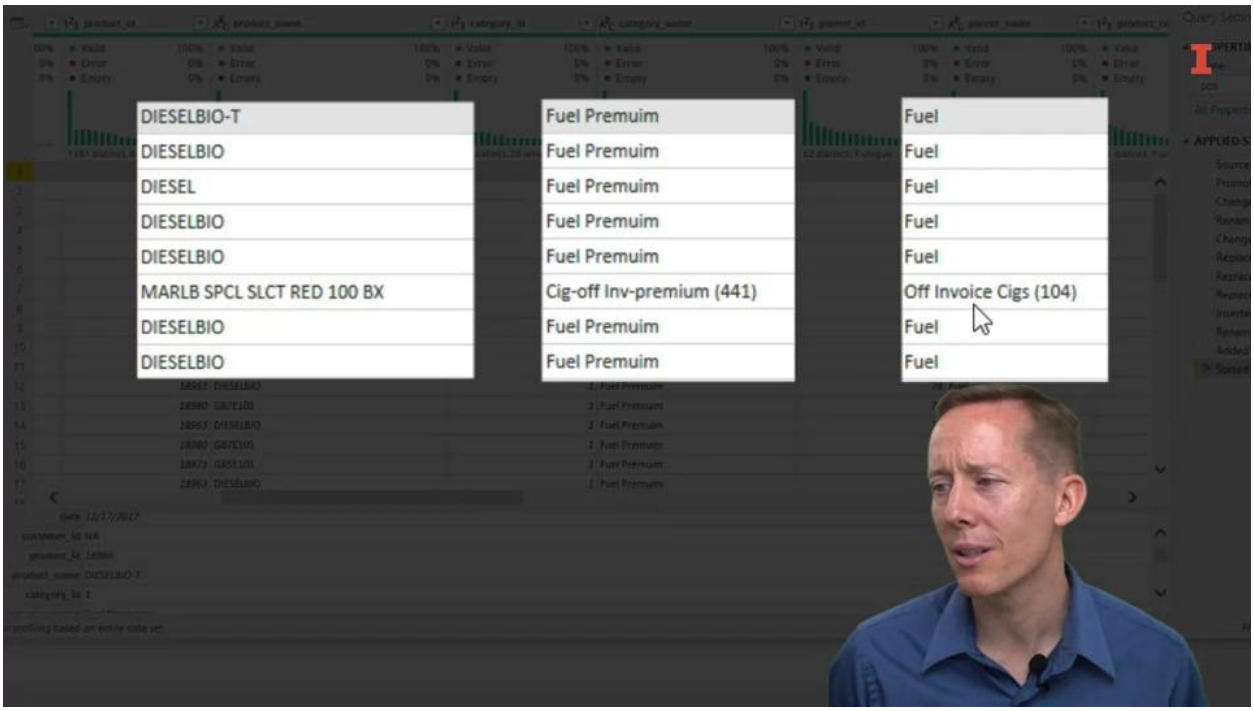
So let's first identify those observations and see if we can figure out what's going on. So I can go to the revenue column, click on the drop down arrow next to the column name and let's sort descending first. Now I can see that the data has been sorted, the rows have been sorted based on the value of revenue with the highest value at the top, and you can visually explore and verify that these are the highest values.



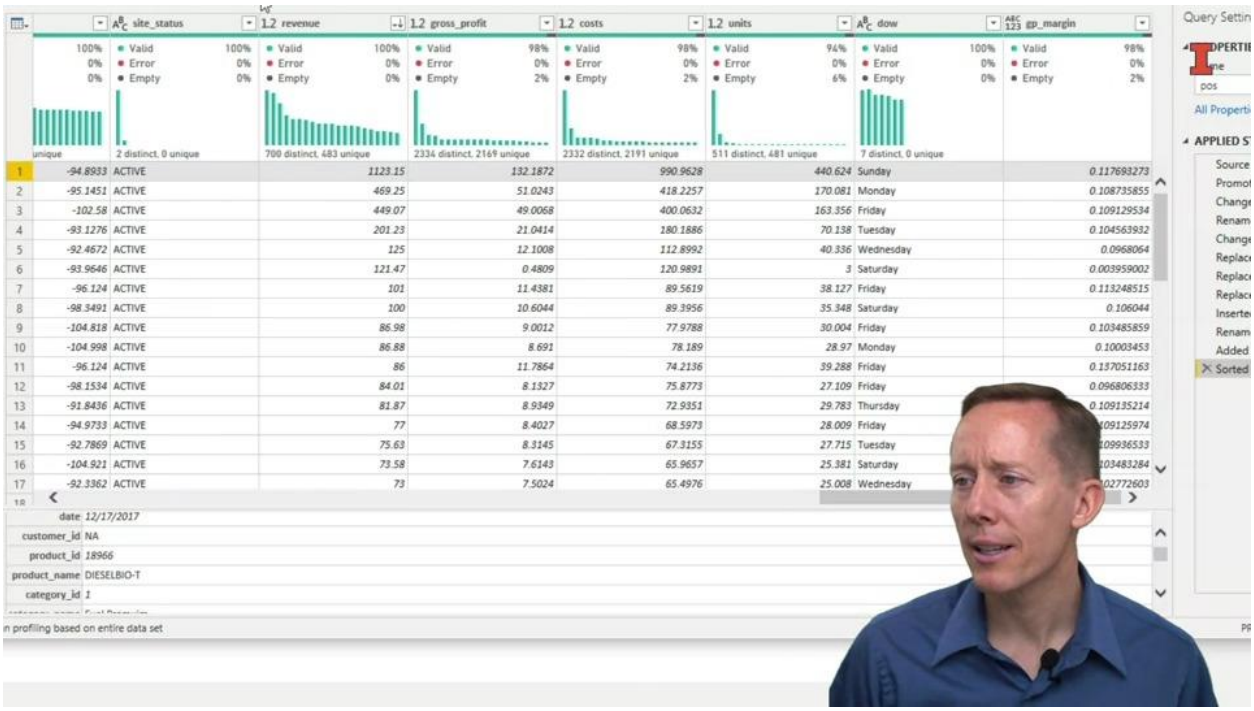
Now, if I click on a column label here, I can get information about the values in the other columns for this observation and I can see that for this observation, it came from diesel bio, so it was a fuel observation.



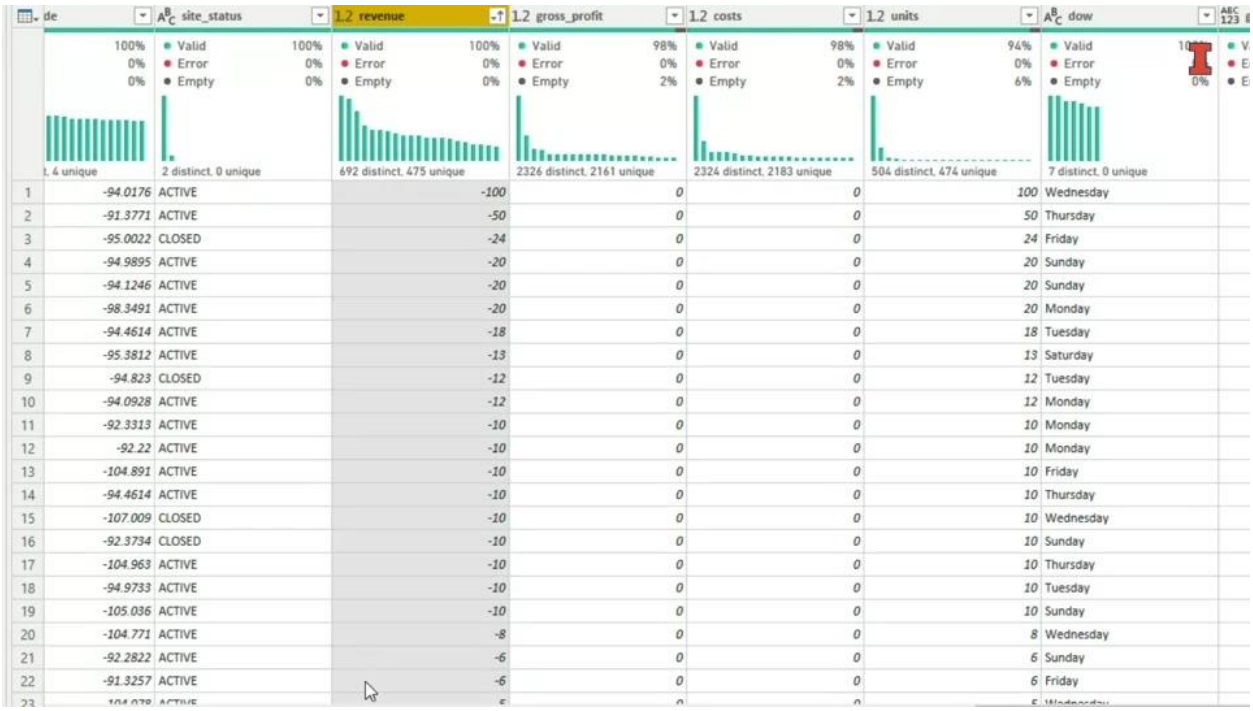
Anyway, so that gives me a way to quickly see what this observation is about. If I want to look at all of these high observations, it looks like about the first eight are above 86, we'll say 87. Then I may want to figure out why, what's going on with these? I could just scroll over until I get to the parent name and category name



and product name and basically real quickly see that these are all related to fuel except for one, which is a large purchase of cigarettes, apparently.



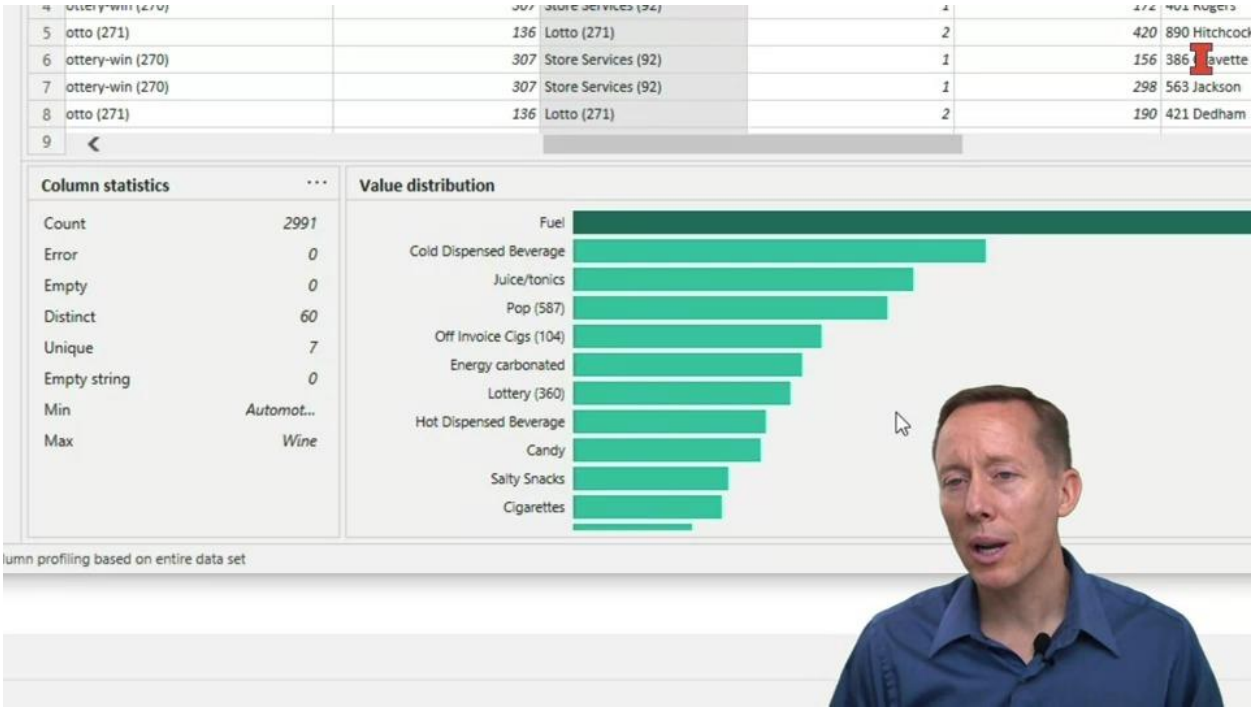
So at this point, I may think to myself, okay, I want to get rid of these. They're not entered in correctly. If they were, then I'd want to go deal with that, but if they're not, then I may not want them though, because they're not representative of most of the other transactions. So one way that I can get rid of these observations is I can go back to the revenue column, and since we've already got it sorted in descending order of revenue, I can go down to this and see the unique values and I can just manually say, "Hey, I don't want these top eight values in here, anything that's above 86 and click Okay," and that will remove those rows from this data set from any additional analysis that I do with this data set. Now, there's another way that I can deal with these and to show you that I'm going to undo that way of filtering. I'll go to the Home tab. In here, I can go to the Reduce Rows section of the ribbon and click on this "Remove Rows" button. I want to remove the top X number of rows. All I have to do is say, hey, since I know these top eight rows are the ones that I don't want, I can click eight there and click "OK". Now this only works because I first sorted the data. If I want to remove the bottom rows of data, I could scroll down to the bottom or I could resort it again, but sort in ascending order. Notice, again, that in this Applied Steps area, it's keeping track of all these different transformations I do with the data. I sorted the data. Now I can see I've got this negative 100 value here. There's really just one observation that's below negative 70.



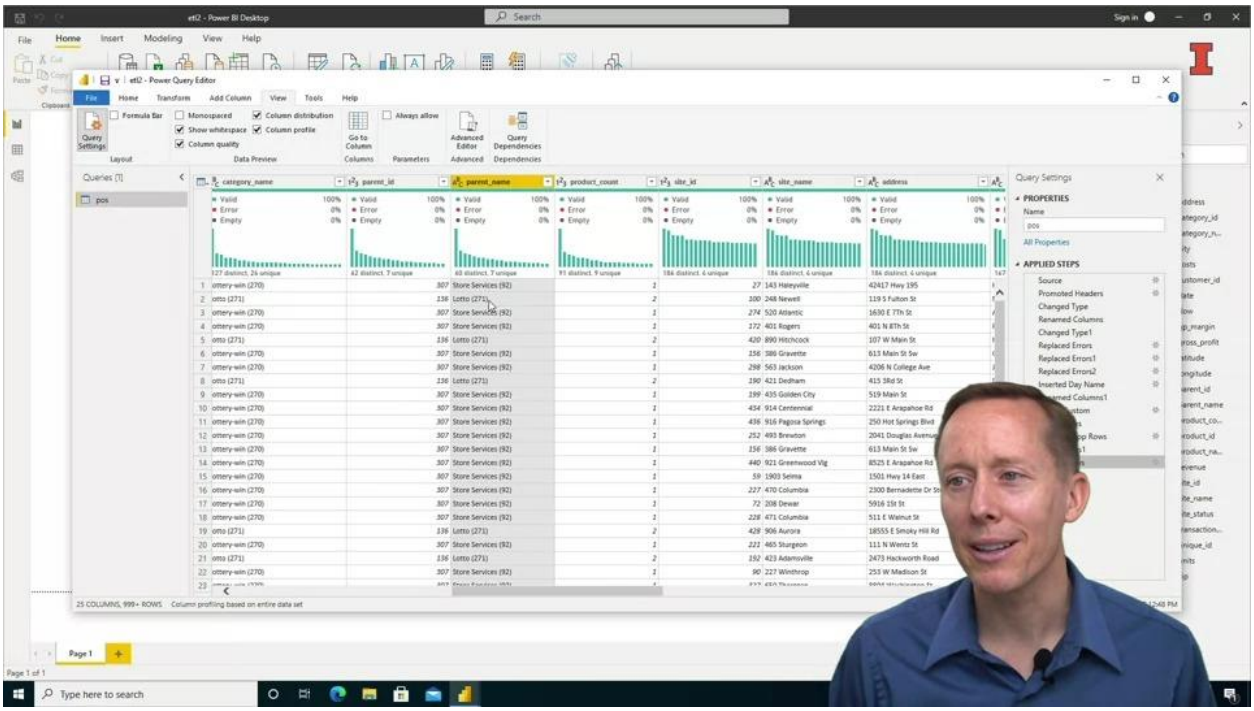
	de	site_status	revenue	gross_profit	costs	units	dow
	100% Valid, 0% Error, 0% Empty	100% Valid, 0% Error, 0% Empty	100% Valid, 0% Error, 0% Empty	98% Valid, 0% Error, 2% Empty	98% Valid, 0% Error, 2% Empty	94% Valid, 0% Error, 6% Empty	Valid, Error, Empty
	4 unique	2 distinct, 0 unique	692 distinct, 475 unique	2326 distinct, 2161 unique	2324 distinct, 2183 unique	504 distinct, 474 unique	7 distinct, 0 unique
1	-94.0176	ACTIVE	-100	0	0	100	Wednesday
2	-91.3771	ACTIVE	-50	0	0	50	Thursday
3	-95.0022	CLOSED	-24	0	0	24	Friday
4	-94.9895	ACTIVE	-20	0	0	20	Sunday
5	-94.1246	ACTIVE	-20	0	0	20	Sunday
6	-98.3491	ACTIVE	-20	0	0	20	Monday
7	-94.4614	ACTIVE	-18	0	0	18	Tuesday
8	-95.3812	ACTIVE	-13	0	0	13	Saturday
9	-94.823	CLOSED	-12	0	0	12	Tuesday
10	-94.0928	ACTIVE	-12	0	0	12	Monday
11	-92.3313	ACTIVE	-10	0	0	10	Monday
12	-92.22	ACTIVE	-10	0	0	10	Monday
13	-104.891	ACTIVE	-10	0	0	10	Friday
14	-94.4614	ACTIVE	-10	0	0	10	Thursday
15	-107.009	CLOSED	-10	0	0	10	Wednesday
16	-92.3734	CLOSED	-10	0	0	10	Sunday
17	-104.963	ACTIVE	-10	0	0	10	Thursday
18	-94.9733	ACTIVE	-10	0	0	10	Tuesday
19	-105.036	ACTIVE	-10	0	0	10	Sunday
20	-104.771	ACTIVE	-8	0	0	8	Wednesday
21	-92.2822	ACTIVE	-6	0	0	6	Sunday
22	-91.3257	ACTIVE	-6	0	0	6	Friday

I could explore this and see what it's related to. In this way it's store services. Again, let's assume that this is entered correctly, but it's not representative of what we normally do, so we want to get rid of it. At this point, I might just want to click on this and uncheck the negative 100 and click "OK".

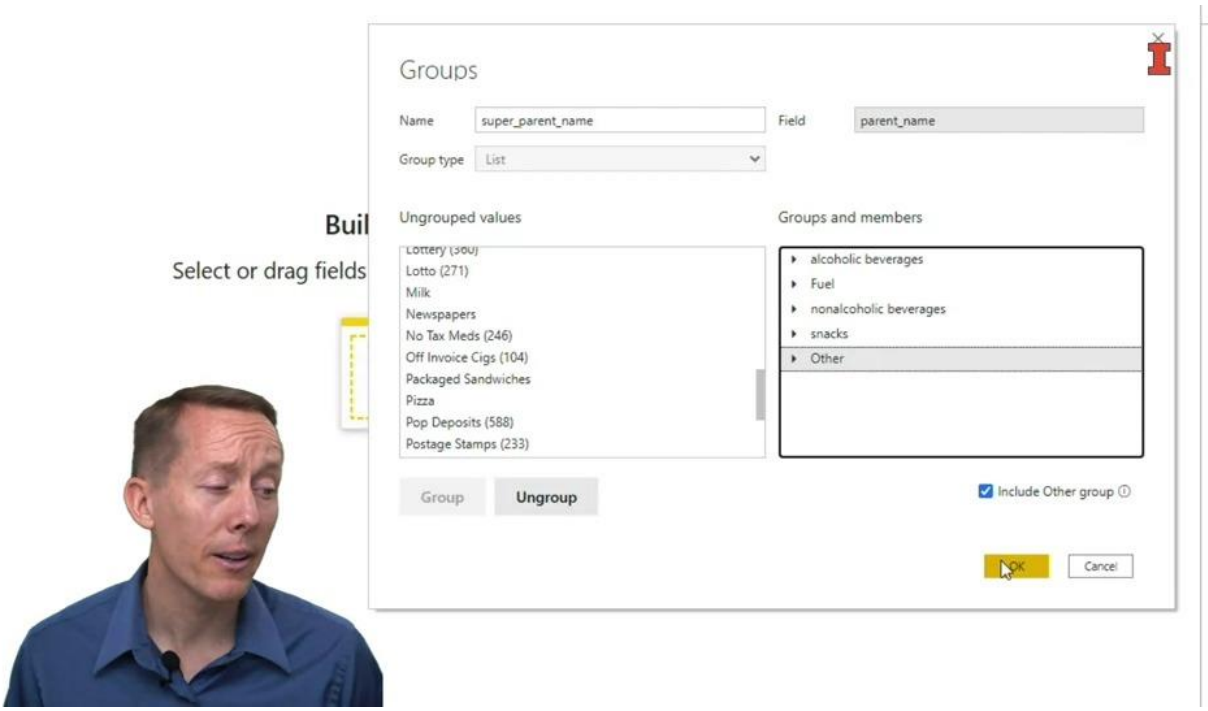
Now I have dealt with my outliers for the numeric columns. Now, this is just one way of dealing with them. Oftentimes, in Power BI, we don't want to remove those rows here, but we may want to remove them using slicers or filters after we create observations. But anyway, you can remove them from the data in the Power Query editor. That's how you do with numeric columns of data. Let's say we want to look at categorical columns of data.



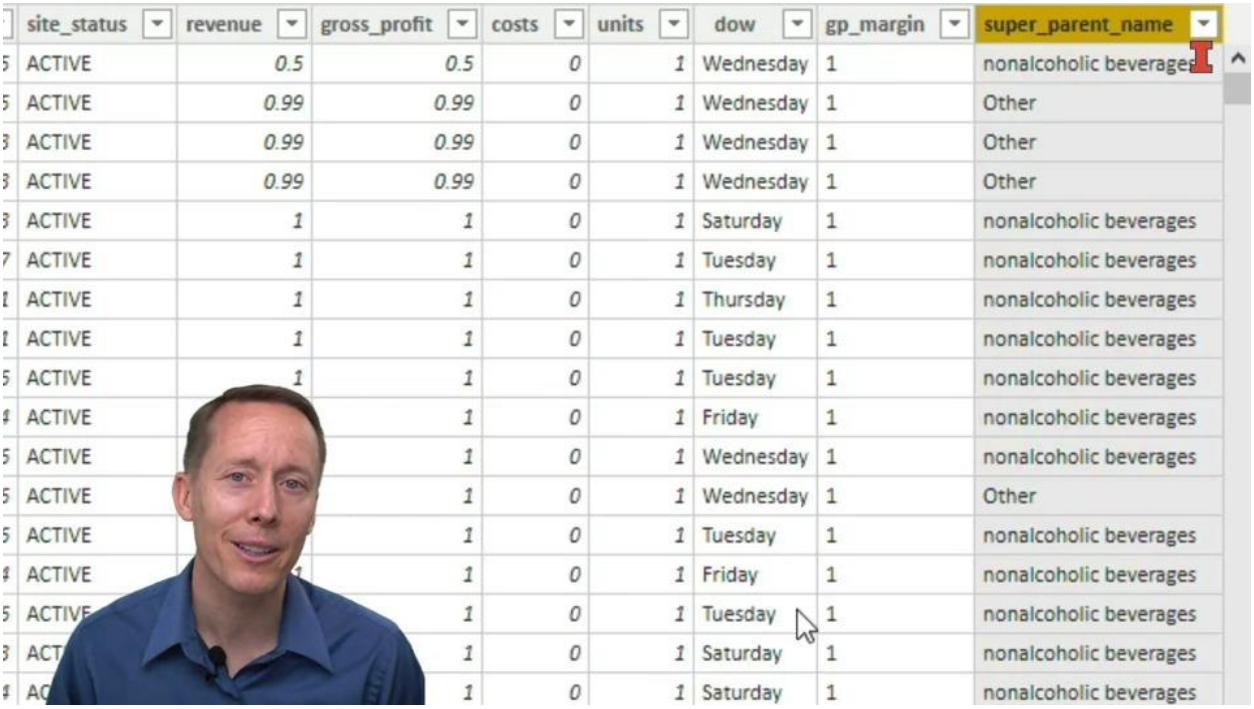
Let's go to the parent name column and let's explore this column a little bit. We can see that in the parent name column, I've got 60 distinct values and seven of those are unique. If I click on that column header and I have the column profile checked in the view menu item, I can get an idea of what those values are in the 60 columns and how often they occur. I can very quickly see that fuel shows up most often, cold dispense beverages, then juice/tonics, and then pop. Hey, three of these top four are beverages. That might stand out to me, so I keep that in mind. Anyway, I can go down through here. If I think about visualizations and how I want to explore things, having 60 different values, seven of which only occur once, that could make it problematic for color coding things. There'd be too many colors. The technical name for this is high cardinality. High cardinality means there's a lot of different values in a set. Oftentimes, it just makes it too cumbersome and hides some of what's going on. I want to reduce the cardinality somehow.



One thing I could do is click on the parent name or hover over it and go to "Remove Duplicates". If I do that and ask myself, does that work? Well, actually, what happened is now I say, I've just got 60 rows. It removed any observation for which the parent name occurred a second time. That's not what I wanted to do. Fortunately, we don't have to worry about, how do I undo that? I can just go over to the Applied Steps and I click on that "Remove Duplicates" button. Now they're all back there. I've got all the data. I want to show you a cool way for dealing with high cardinality in Power BI. First, I need to exit the Power Query editor, so I'll close and apply these changes. Great.

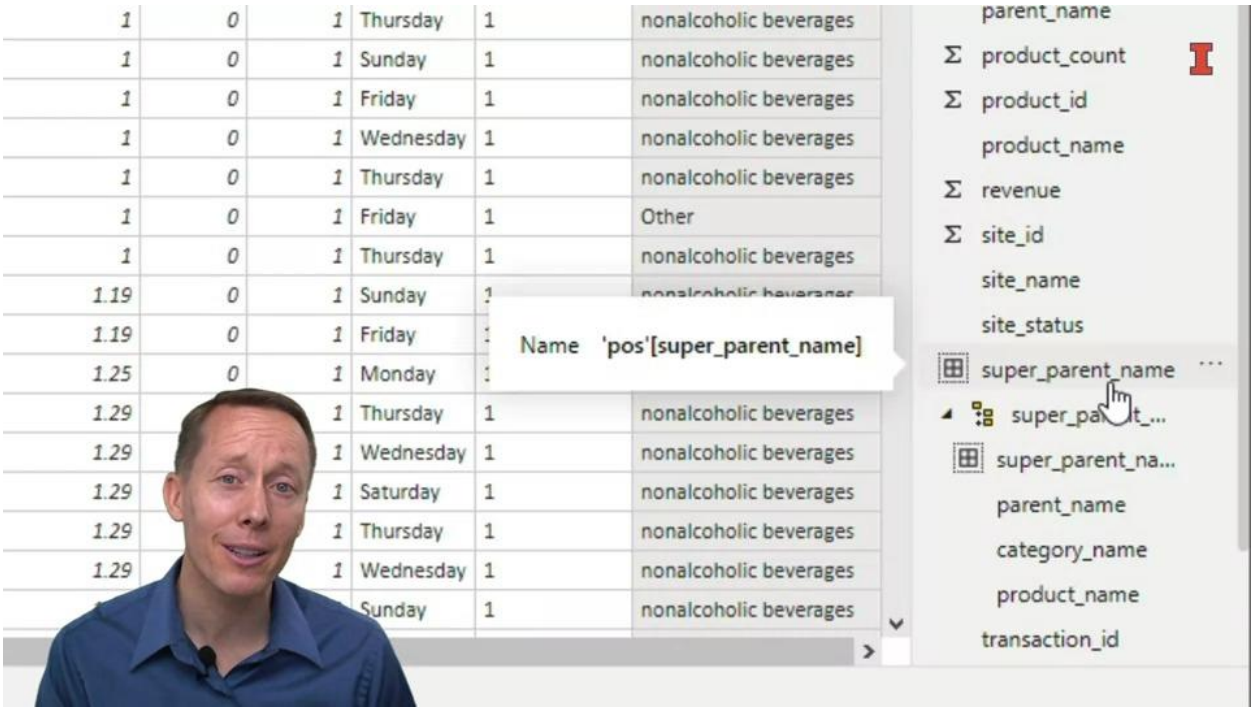


Now, let's go ahead and go over to the fields section and let's go to the parent name and click on the three dots next to it. Let's go to New group. What we will do is we will manually create groups for all of these different parent names. I'm actually going to create a new column. This column I'll call super parent name. Now what I can do is group together, perhaps any parent name that is a beverage. As I start exploring these parent names, now they're listed in alphabetical order, I might, real quickly see here I've got beer. I'll click on that and I can see that this group comes up. I could create a group, but I want to add more to that. Let's go down and identify any other perhaps alcoholic beverages in here. I've already done this, so I know that there's a liquor in here. I'm holding Control Down and clicking on "Liquor" and then there's wine. I've got those three parent names selected. I'll click "Group", and now I've got this new group and the members associated with it and I want to rename this to alcoholic beverages. That's one group. Now, I also may want to create just a beverages group. Let's go down and identify all the beverages that are nonalcoholic and put those into a group. Now I've got this nonalcoholic beverages group here that's made up of these items. Now let's go ahead and create another group that is made up of snacks. I created a new snacks group made up of these parent name items and I can collapse that. I can say I've got these three groups so far, alcoholic beverages, nonalcoholic beverages, snacks. I also want to create a group that's made up just of fuel. I'm doing that because I know fuel is a high-volume item. Now, all these other things may not be as important to us. They may make up a smaller amount of revenue, and so I don't necessarily want to leave those out or leave them separate. I can real quickly, click this "Include Other Groups" and everything else falls into this other category. Then I'll click "Okay".



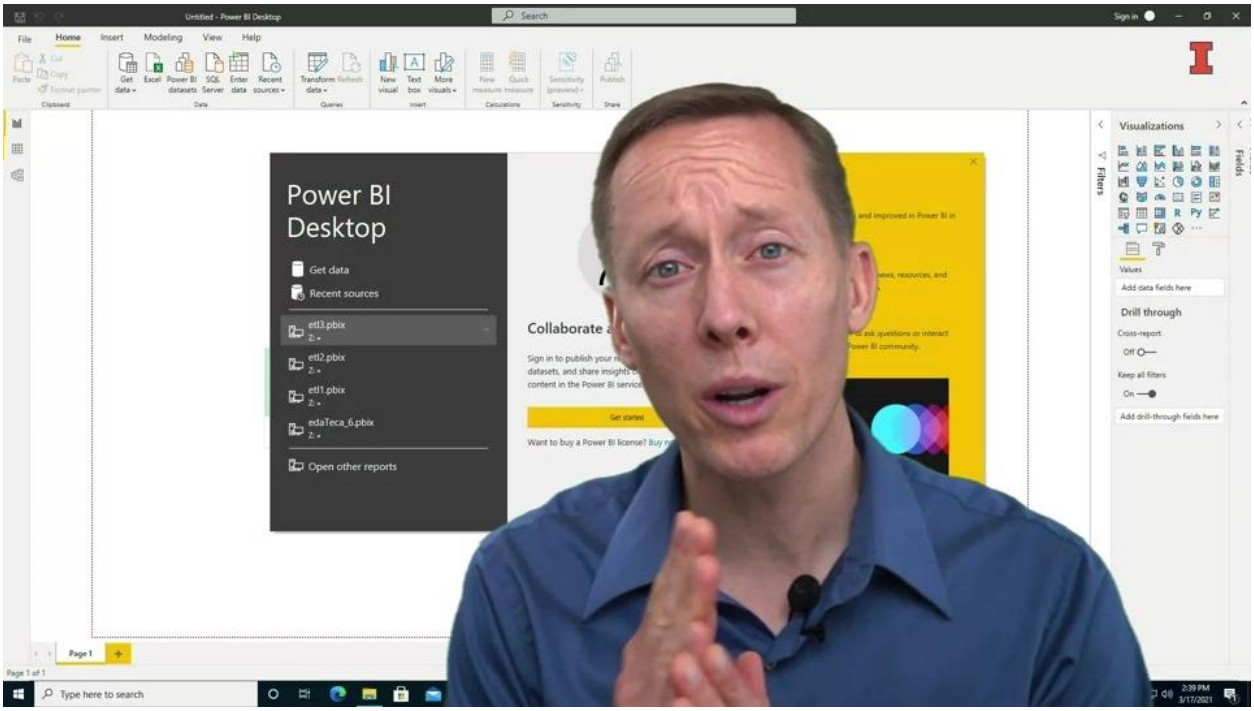
site_status	revenue	gross_profit	costs	units	dow	gp_margin	super_parent_name
ACTIVE	0.5	0.5	0	1	Wednesday	1	nonalcoholic beverages
ACTIVE	0.99	0.99	0	1	Wednesday	1	Other
ACTIVE	0.99	0.99	0	1	Wednesday	1	Other
ACTIVE	0.99	0.99	0	1	Wednesday	1	Other
ACTIVE	1	1	0	1	Saturday	1	nonalcoholic beverages
ACTIVE	1	1	0	1	Tuesday	1	nonalcoholic beverages
ACTIVE	1	1	0	1	Thursday	1	nonalcoholic beverages
ACTIVE	1	1	0	1	Tuesday	1	nonalcoholic beverages
ACTIVE	1	1	0	1	Tuesday	1	nonalcoholic beverages
ACTIVE	1	1	0	1	Friday	1	nonalcoholic beverages
ACTIVE	1	1	0	1	Wednesday	1	nonalcoholic beverages
ACTIVE	1	1	0	1	Wednesday	1	Other
ACTIVE	1	1	0	1	Tuesday	1	nonalcoholic beverages
ACTIVE	1	1	0	1	Friday	1	nonalcoholic beverages
ACTIVE	1	1	0	1	Tuesday	1	nonalcoholic beverages
ACT	1	1	0	1	Saturday	1	nonalcoholic beverages
AC	1	1	0	1	Saturday	1	nonalcoholic beverages

Now I can say I've got this super parent name column over here and if I click on the data and explore it quickly, I can see that this column is added to the end, and if I click on the Drop Down arrow, I can see that I've only got these five different categories in here; alcoholic beverages, fuel, nonalcoholic beverages, other and snacks. That's one way of dealing with high cardinality columns. Now, here's another cool thing you can do with Power BI.

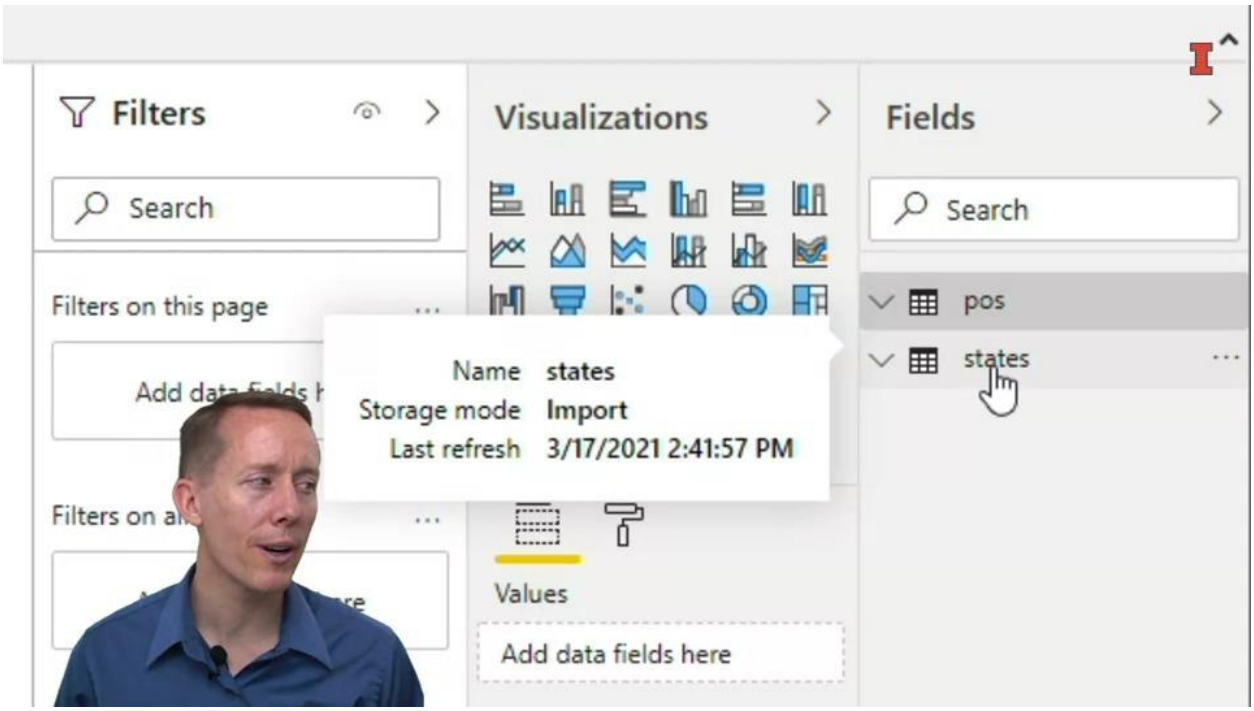


Remember how with the date column, it created automatically this date hierarchy so we could scroll down through the data and start with year then go down to quarter and then month and date, we can do that with any column of data or columns of data. This will be very helpful as you create visualizations that you may want to drill down into. Let me show you how to do that. I can go to super parent name, click on the three dots there and click on New hierarchy. Now I've got this super parent name hierarchy created and I can add different columns to that hierarchy, so I can drill down. After the super parent name, comes the parent name. I'll go up to the parent name column, click on the three dots next to it, and click "Add to Hierarchy", super parent name hierarchy. If there were more than one hierarchy, I can choose which hierarchy I'd want to add it to. Now, under the super parent name hierarchy, I've got parent name and after parent name, I would want to add category name. I go up to category name, click on the three dots, and add that to the super parent name hierarchy. There is category name, and then I can go to product name and add that to the hierarchy as well. Creating this hierarchy is a bonus, it's something that'll be useful as we create visualizations. Sometimes you want to create these hierarchies after exploring the data a little bit, as well as dealing with outliers. Once you start exploring the data visually, you'll see they're outliers. We'll come back to dealing with outliers later. But this is one way to deal with outliers in both numeric and categorical columns.

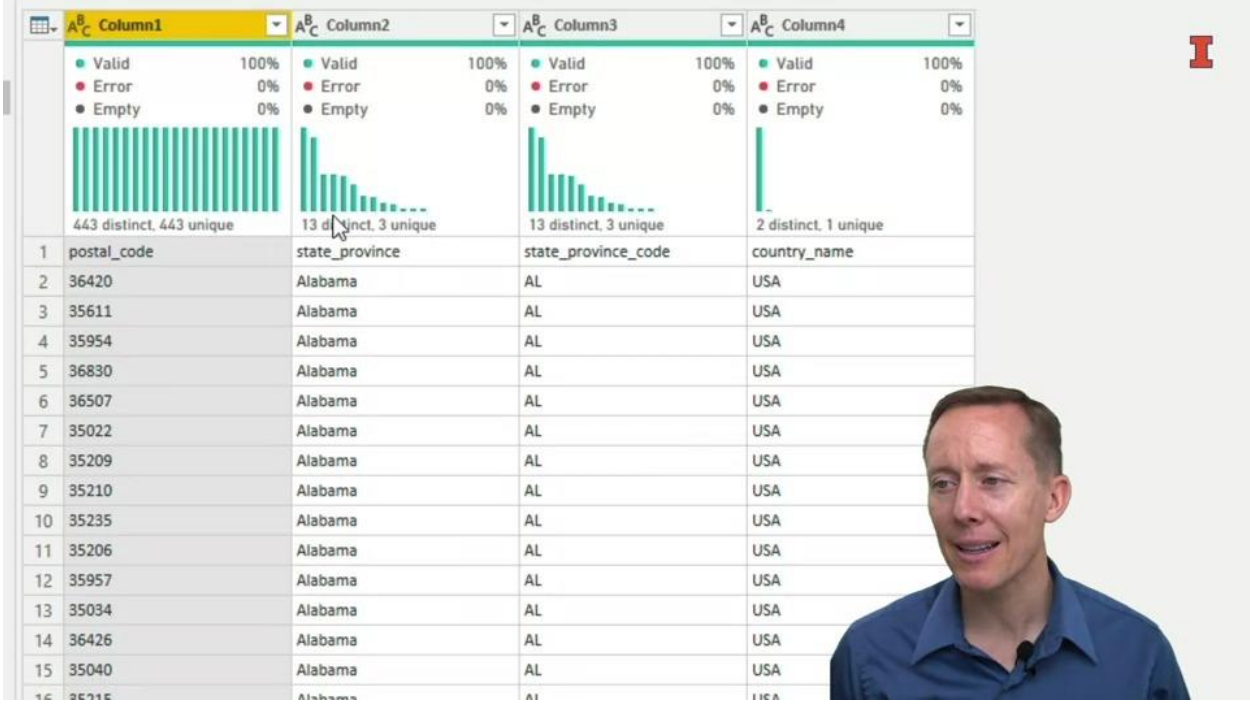
Lesson 2-2.6 ETL 4: Data Models and Joins



In this video, we will focus on joining data together in Power BI. This is important because oftentimes you may have one table or a data set that you're focused on, but you want to bring in information from other data sets. We'll need to do that by matching values on one or more columns. We will focus on how to do that in Power BI, and along the way, we will also illustrate how to create a simple chart and how to do some data wrangling as well.

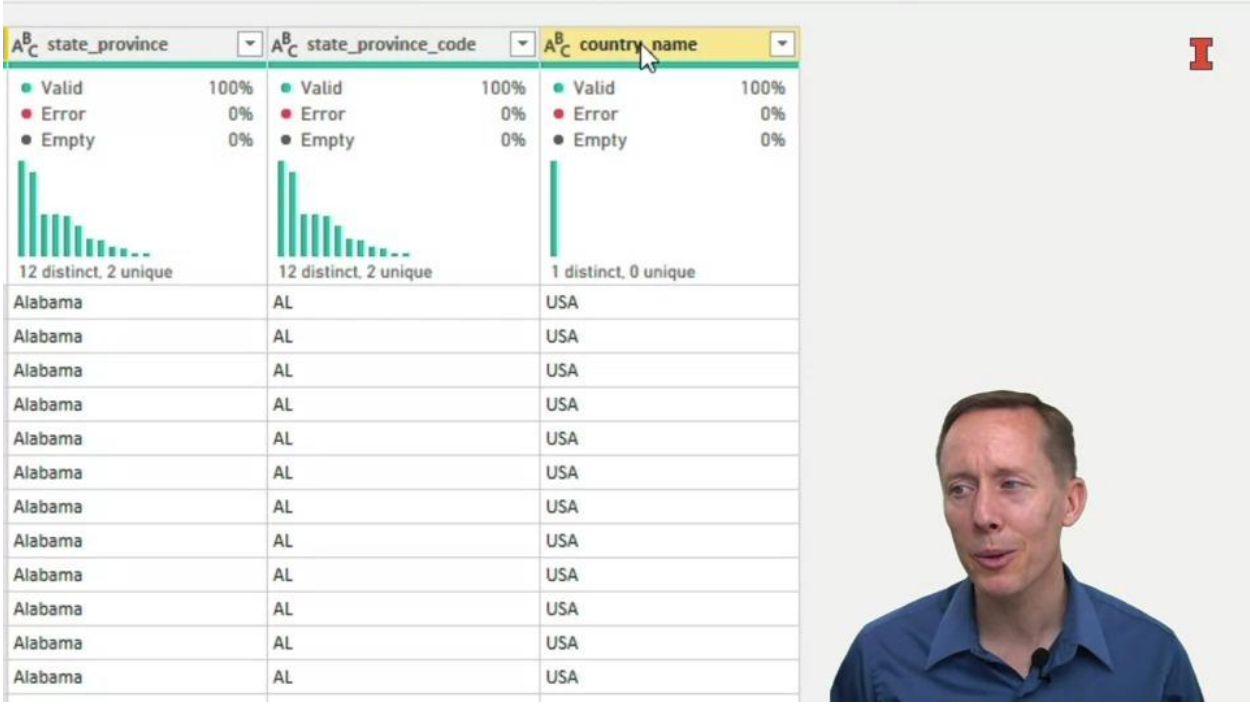


The first thing that we should do is open up Power BI and I will open this project that we've been working etl3. Now we've got this project loaded and we can see in the field section that we've got one data set, pos, or the point of sale data, and within this data set, we have a column called Zip, for zip code, also known as postal code. We can see it's a numeric value which will be important a little bit later on. Now, what if I want to aggregate data and visualize data based on state or province? I don't currently have that information in here. But if I had a list of zip code and which state they belong to, I could bring in that state information and then do that type of aggregation. Fortunately, I have that data and that data is located in the same folder as this data, it doesn't have to be, but it's called states.csv. Let's first read in that data by going to the home tab and clicking on "Get Data." Since it's a.csv file, we'll click "Text/CSV" and Connect. Then indicate we want the states.csv file and here is a preview of the data. I'll go ahead and load it. Now if I look over in the field section, you can see I've got pos and states. Now, I like the name states, so I'll just leave it at that. But let's go into the Power Query Editor to transform and further analyze what this data set looks like.



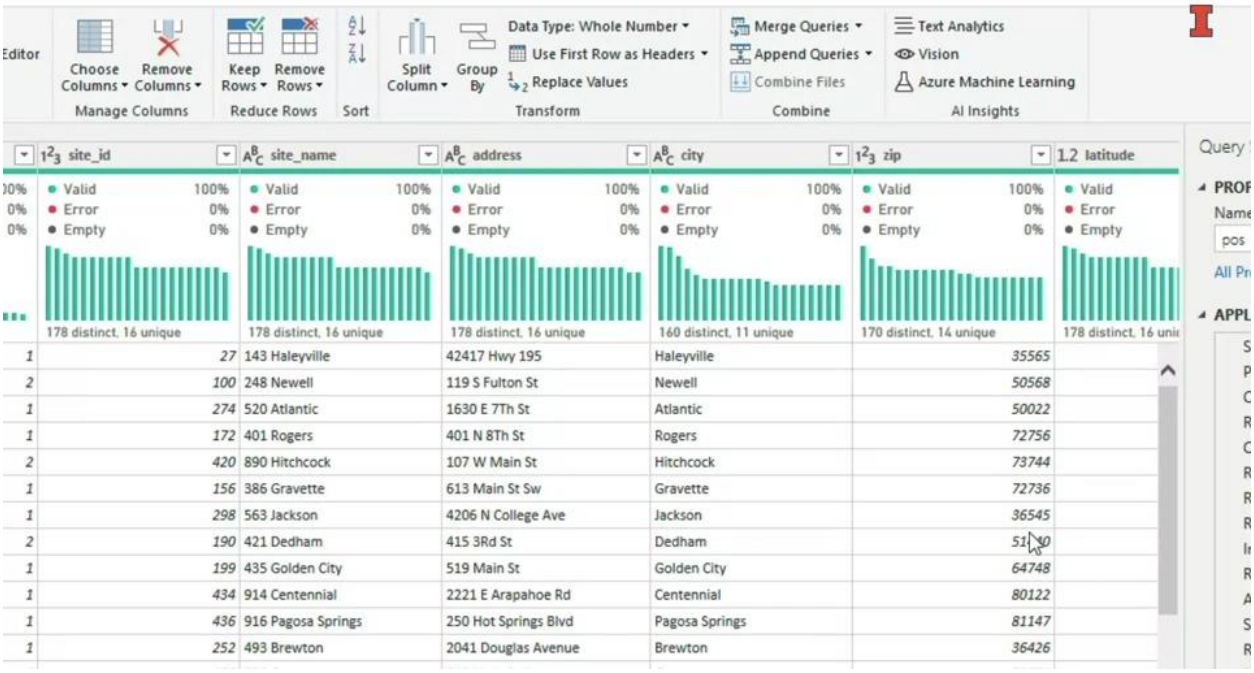
	Column1	Column2	Column3	Column4
	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
	443 distinct, 443 unique	13 distinct, 3 unique	13 distinct, 3 unique	2 distinct, 1 unique
1	postal_code	state_province	state_province_code	country_name
2	36420	Alabama	AL	USA
3	35611	Alabama	AL	USA
4	35954	Alabama	AL	USA
5	36830	Alabama	AL	USA
6	36507	Alabama	AL	USA
7	35022	Alabama	AL	USA
8	35209	Alabama	AL	USA
9	35210	Alabama	AL	USA
10	35235	Alabama	AL	USA
11	35206	Alabama	AL	USA
12	35957	Alabama	AL	USA
13	35034	Alabama	AL	USA
14	36426	Alabama	AL	USA
15	35040	Alabama	AL	USA

Here on the left-hand side, we can see we've got pos and the states, I'll click on "States," and the first thing that I think we should do is just look at the structure of this data set. We can see that it has four columns and 443 rows. Now, if you browse this data, you can see that the first row appears to be the column header, but Power BI did not know that it should be the column header. So we want to somehow get this up to the column name up here. Fortunately, there's a really easy way to do that because this happens quite a bit.

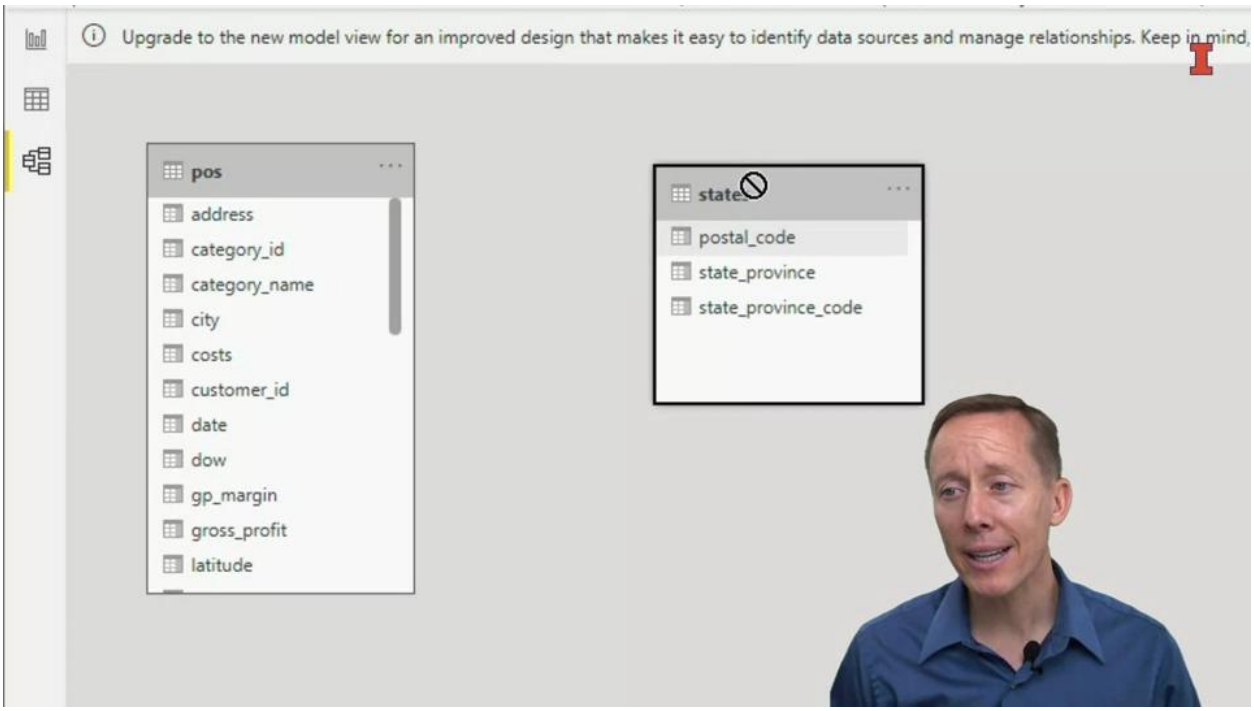


state_province	state_province_code	country_name
Valid 100%	Valid 100%	Valid 100%
Error 0%	Error 0%	Error 0%
Empty 0%	Empty 0%	Empty 0%
12 distinct, 2 unique	12 distinct, 2 unique	1 distinct, 0 unique
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA
Alabama	AL	USA

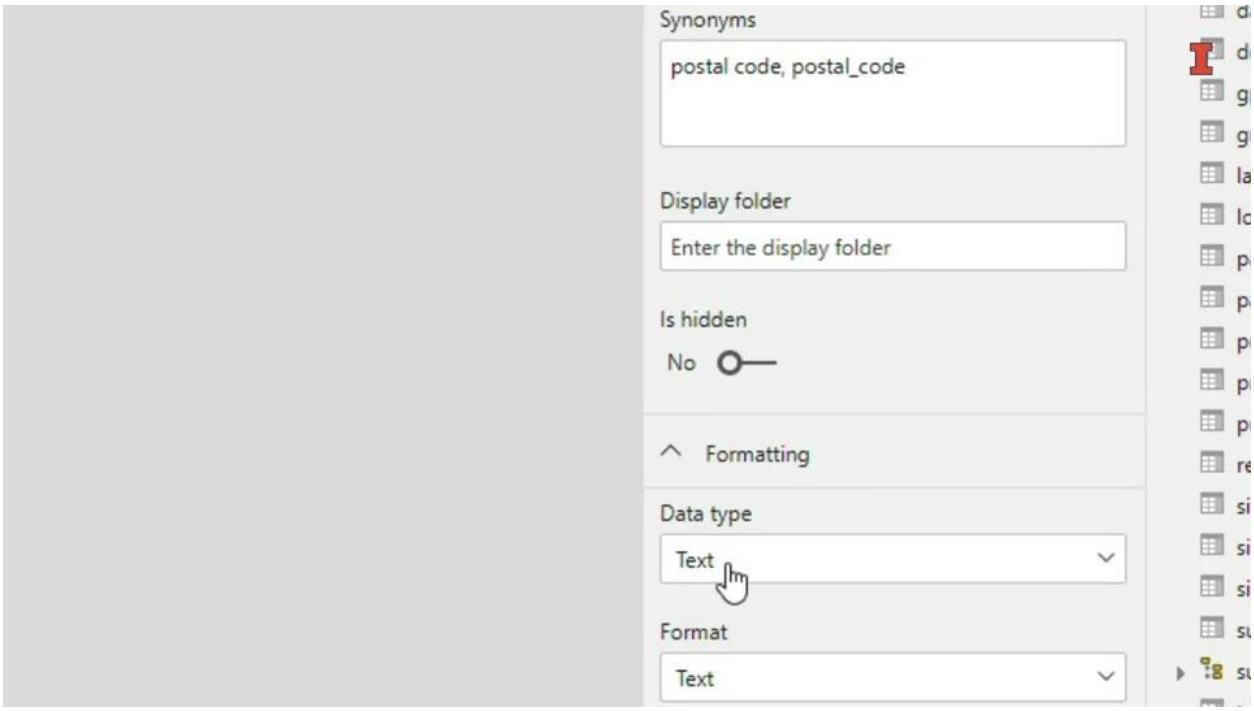
In the home tab, we can go over here and click "Use First Row as Headers." Now we can see that there are only 442 rows instead of 443, and these column names are more appropriate. Let's start looking at the values in the columns, inspect the quality, and so forth. For the postal code, we can see that there are 442 distinct values and 442 unique values. This means that each postal code only shows up one time. If we look at the state_province column, we can see there are 12 distinct values and two are unique. So two of the states only show up once. We can click on the "state_province_code" or hover over that and see that it's very similar to a state or province, it's just that these are two letter abbreviations. Then the country_name has only one distinct value and zero unique, meaning everything is USA. Well, this column probably, really isn't that important. Let's go ahead and get rid of it.



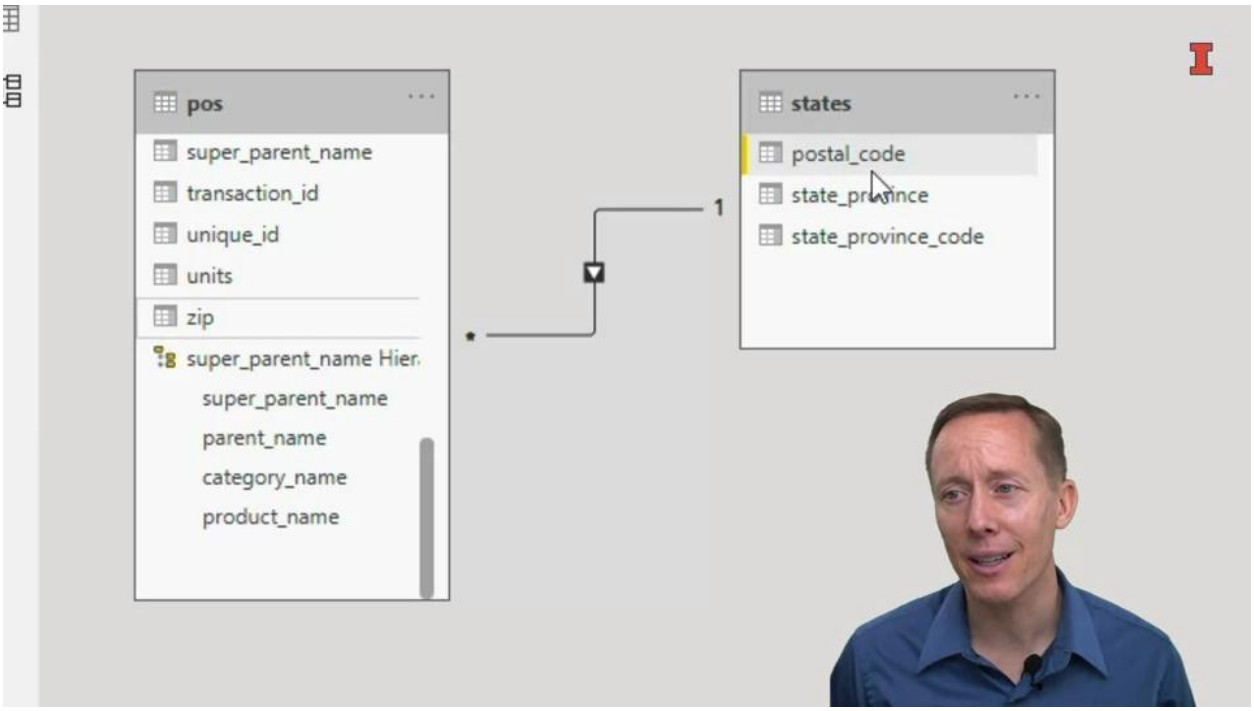
We'll click on that column header and then in the home tab here, click on "Remove Columns," and we got rid of that column. As always, if we make a mistake or want to revert back to a previous step, we can go over to applied steps and remove that. But I'm going to go ahead and leave that here. The idea here is that we want to take these postal codes here and maybe we just browse them in. As we go down, look at these, you can see that some of these observations have a nine-digit zip code. We've got hyphenate the first five digits, so hyphen in the last four digits. What we want to do is take these state values and link it to each transaction here based on the zip code. If we scroll over here a little bit, we can see we've got the zip code in here and there are the unique values. These are only five-digit zip codes. At some point, we're going to have to get rid of those last four digits.



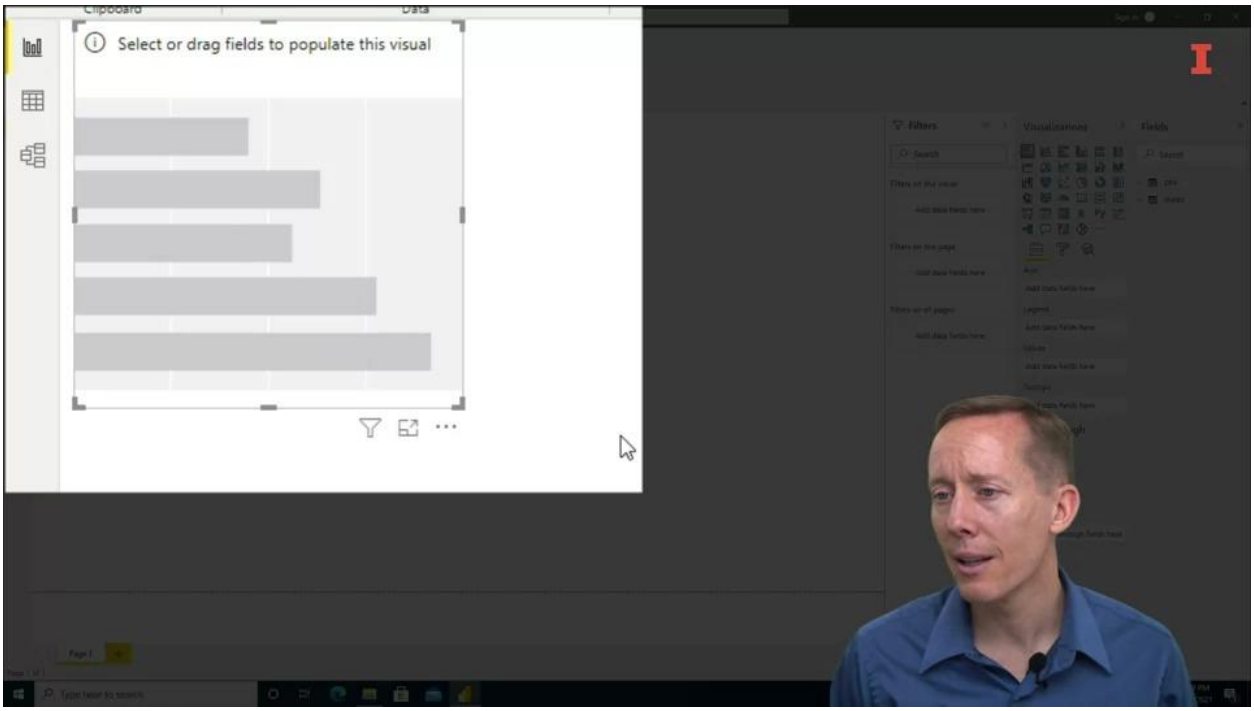
But once we've done that with the data, let's just go ahead and close and apply these changes that we made to the states' data set. Now that we've made those changes, let's talk about how to join data using the data modeling tool. I'll click on this Data Model, on the left sidebar. This gives me a rectangle for each data set that I have. Now, if there are any data sets that have the same column name and the same type, it will automatically recognize that these can be joined together.



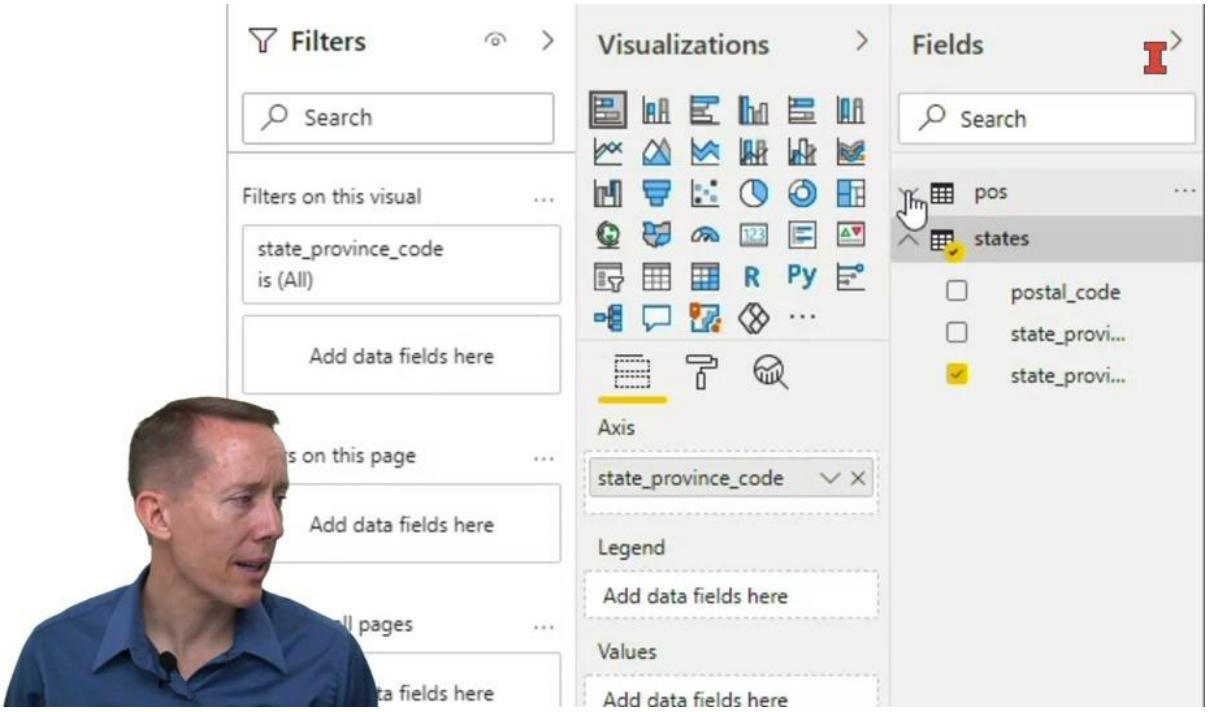
That's not the case here, because in the POS data set, we've got the zip column. If I click on that, I can see that the data type is a whole number. In the states column I've got the postal code, and while it's really the same thing, this is a text value here. Now, that doesn't really make a difference in this data model here.



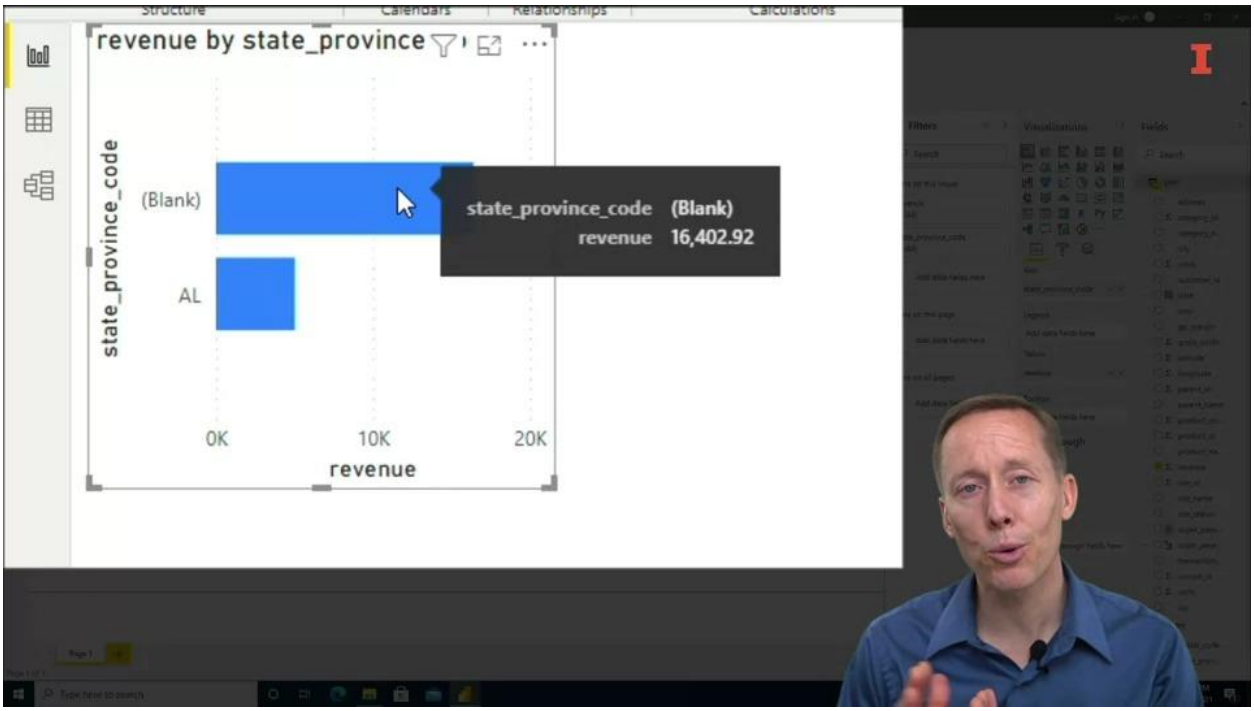
Let's just go ahead and pretend we don't know there's any difference or that there are the last four digits in the states data set. What I can do is I can explicitly define a relationship by clicking on a column name and dragging it over to a column name on another table. I can click on this arrow here and see that these tables are joined together using the zip column and the postal code column. I can also see a star on the side of this arrow that's closest to the POS table and a 1 on the side of this arrow that's closest to the states table. What that means is that there's a 1 to many relationship. In other words, each transaction in each row in POS matches up with only one observation in the states data set, but the observations in the states data set based on postal code could match up with many observations in the POS data set.



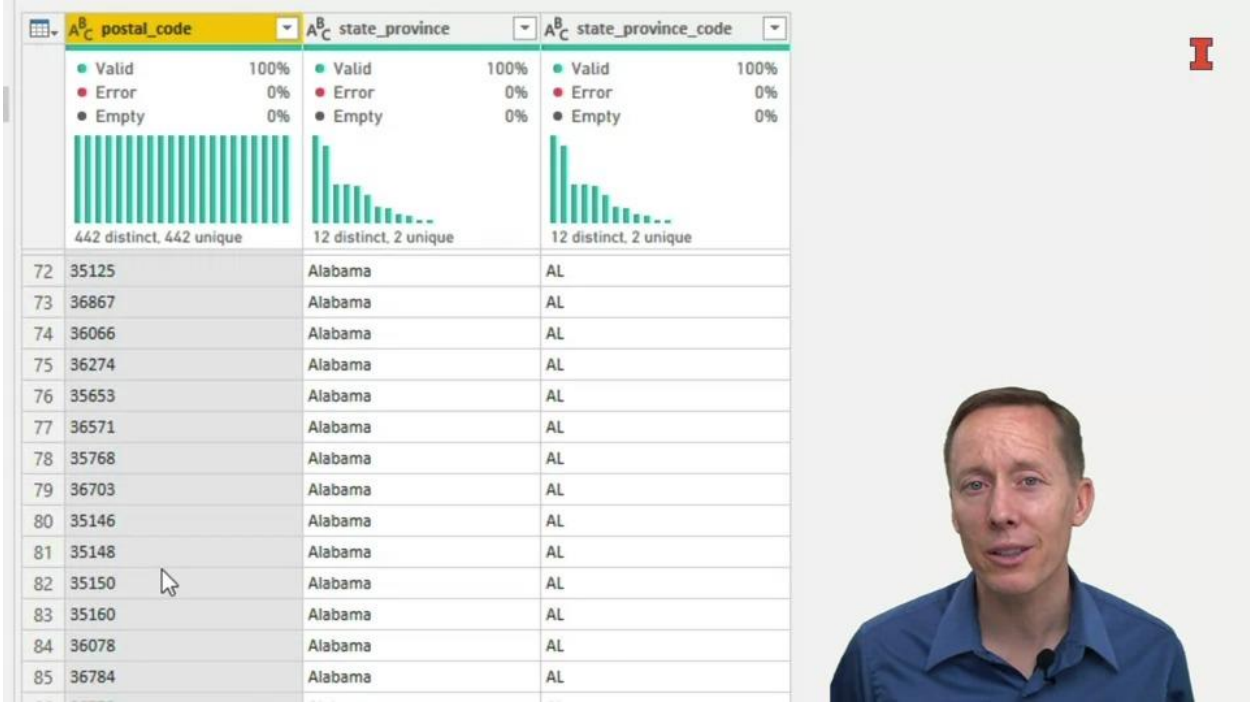
Right now, since we've done that, let's go ahead and just try creating a basic chart. Now, I realize you may not know how to create charts yet in Power BI, but whether you do or you don't it's really easy. Let's just go ahead and create a bar chart here. I will click on that stacked bar chart icon. You can see it creates this folder for the chart here,



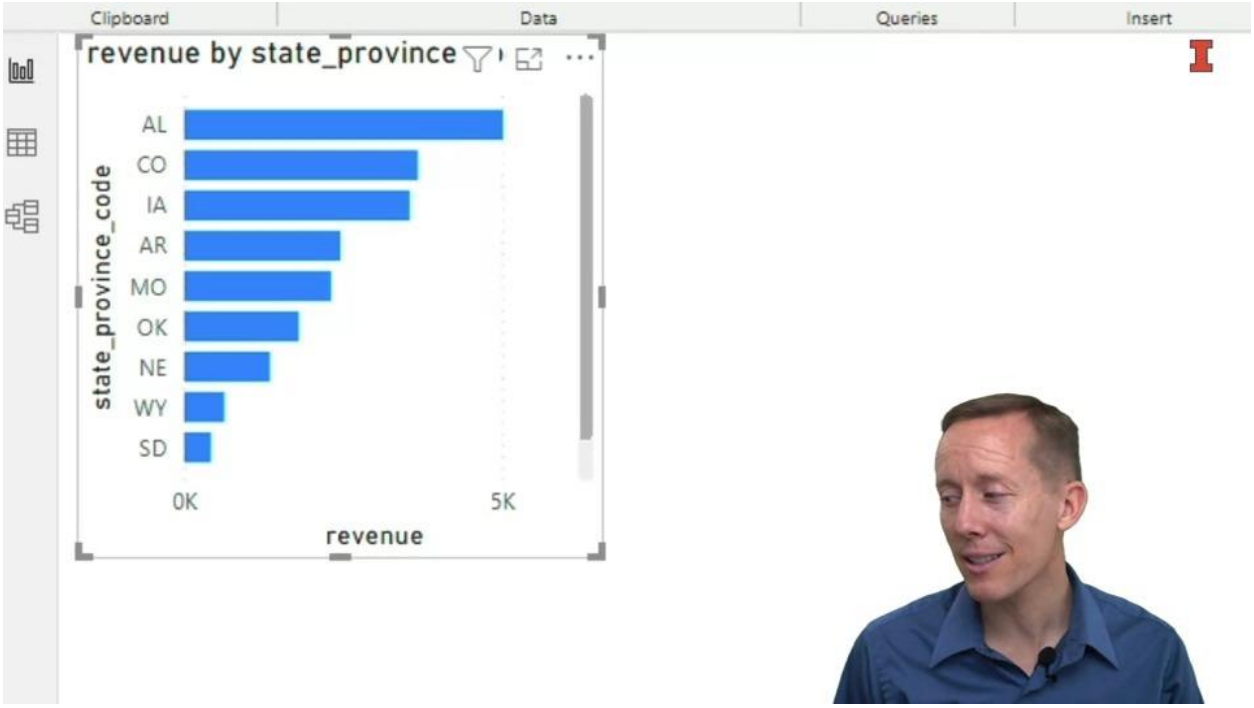
and then if I go into the states data set in here, I will drag state or province to the axis, and then in the POS data, I will drag revenue to the values.



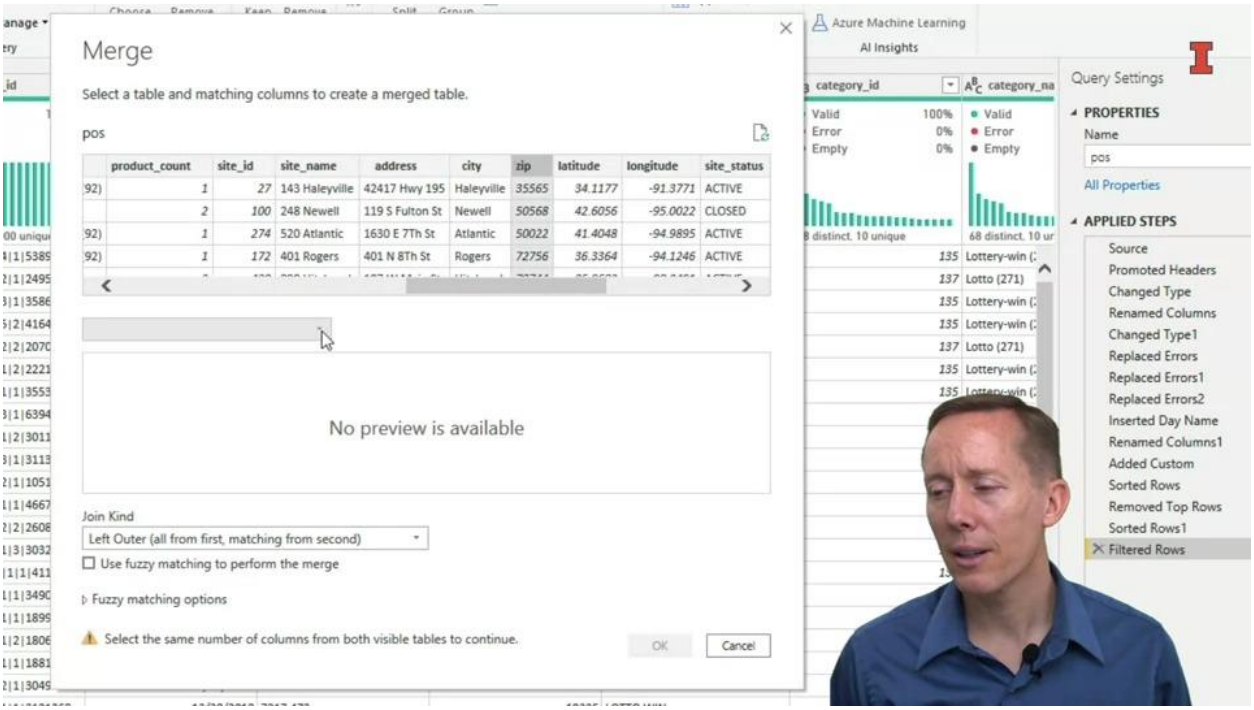
I'm not going to focus on what that means at this point here, but it's pretty easy to see that we created a bar for each state and we added up the revenue. Now you can see that I've got a lot of blanks and one AL for Alabama. I presume there's more than just one state in the this data set. What's going on here? What are all these blanks? Remember how some of the postal codes in the states data set have the four additional digits for the postal code, so we need to go in and clean that up.



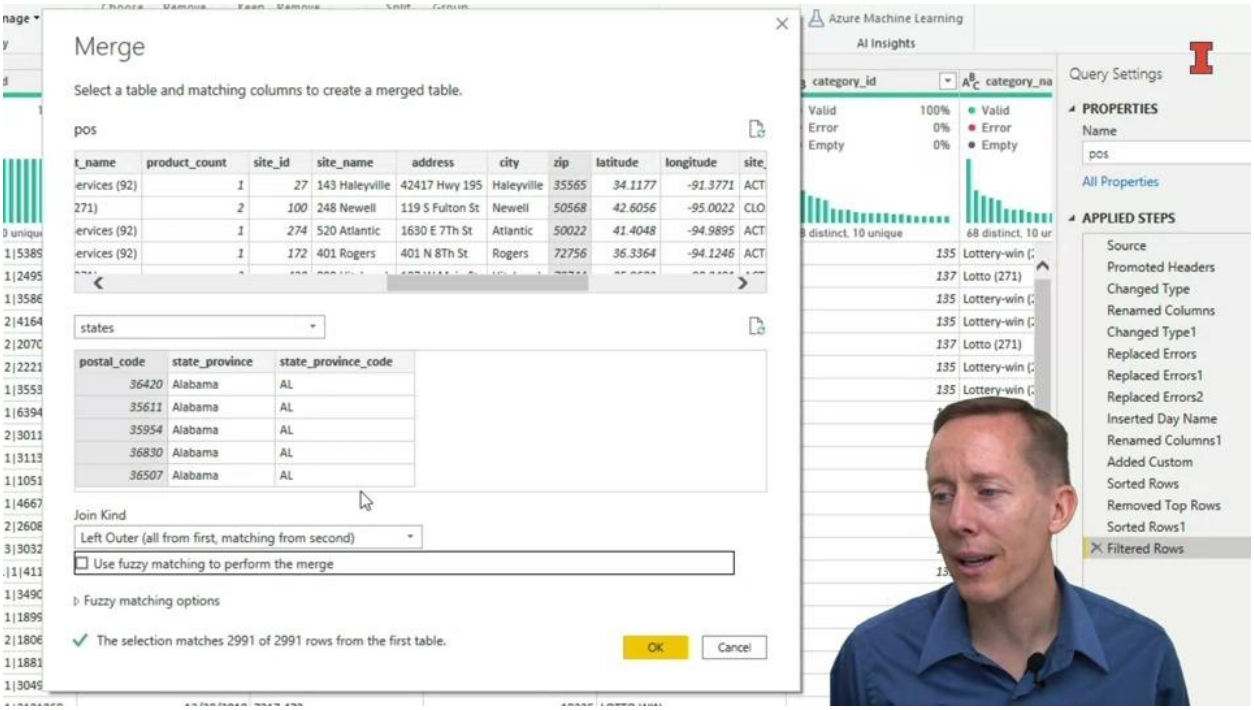
Let's go ahead and go to the home button and click on the transform data. Now that we're in the power query editor, we want to go into the states data set and recall that this data set has some observations that have the additional four digits in there. It would be awesome if we could just somehow extract only the information before that hyphen. There is a way to do that. Make sure we've highlighted that column. We'll go to transform, and then we'll go over to the extract area here for text columns, and we're going to select text before delimiter and our delimiter is the hyphen. Click "Hyphen" click "Okay" and now as I go down and look at these different postal codes, you can see that there are no longer the last four digits in that hyphen.



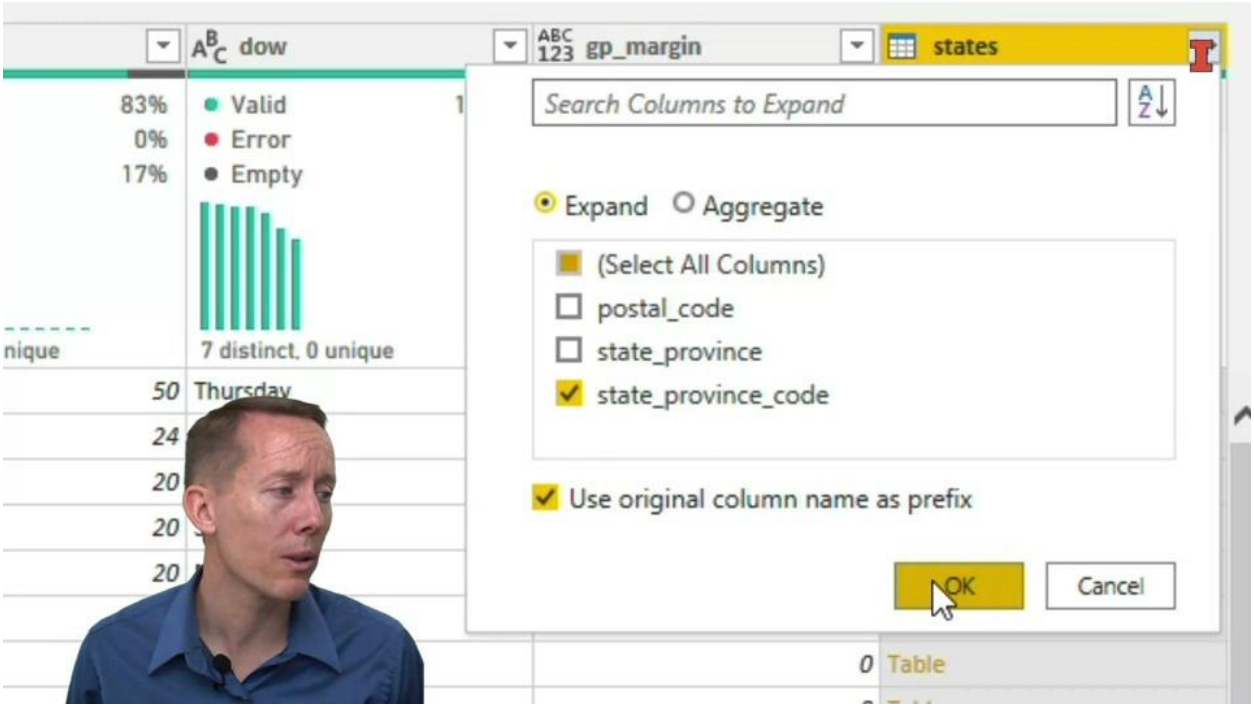
There are a lot of other ways that you can wrangle text data or character strings in here. You can split it, format it and so forth, anyway that's just one example. But now that we've done that let's go ahead and close and apply this, and notice how this chart will be updated here once that has finished processing. Now, we've got more than just Alabama in here and there are no blank states in there. That data model is really cool because even though states doesn't show up in one data set, we can use it from another data set.



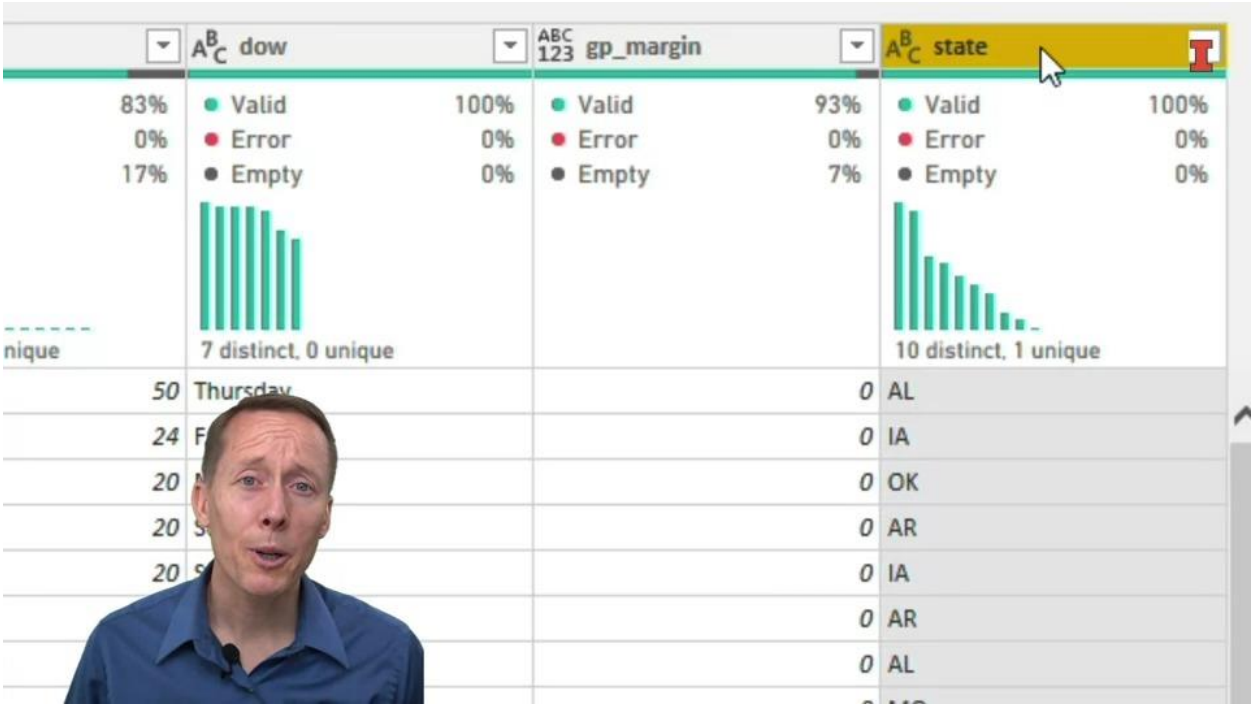
Sometimes, though, we do want to create a new column for this POS data set. Let's talk about how to do that. I'll click on the "Transform data" and in the home tab here, I can go over to the combine section and just click on "Merge queries" and now what I have is a representation of the POS data set. I could change this data set if I wanted to, but I have this data set. I'm going to scroll over to the zip column, click on that. Down here, I indicate the other data set I want to connect it to and that will be states, and I want to link these together based on the postal code and the zip code. Now we get this prompt to indicate a privacy may be an issue here; we can ignore it for this data set. Now, once we do that, it's not allowing us to link these data sets together. Well, the reason is because it says select columns of the same type.



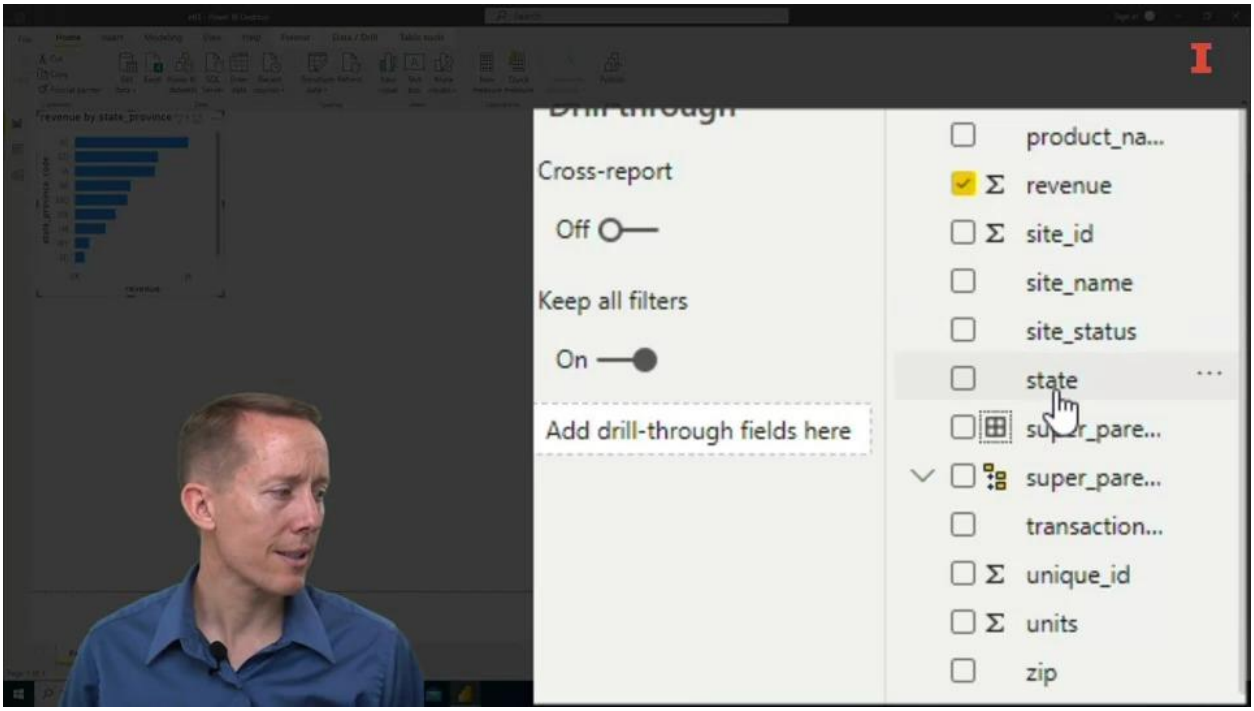
Let's go ahead and make that translation. We can see that the zip column and the point of sale data is a numeric value, whereas in states it is a text or character string value. We need to do convert one to the other. Let's just go ahead and convert postal code to a whole number and then we'll go back over to the point of sale data, click on Merge Queries, and now I can scroll over to zip, select that, select the States, and now select Postal code. This time it says it will match up, it will allow us to move forward. But before we do that, let's just look at a couple of other options here. We can do different types of joins here; left, right, full, inner, and then anti-joins as well. You don't have to necessarily start with POS, you could start with states and then link it to the POS data and then use a right join. In this case, we're going to stick with that left join and click Okay.



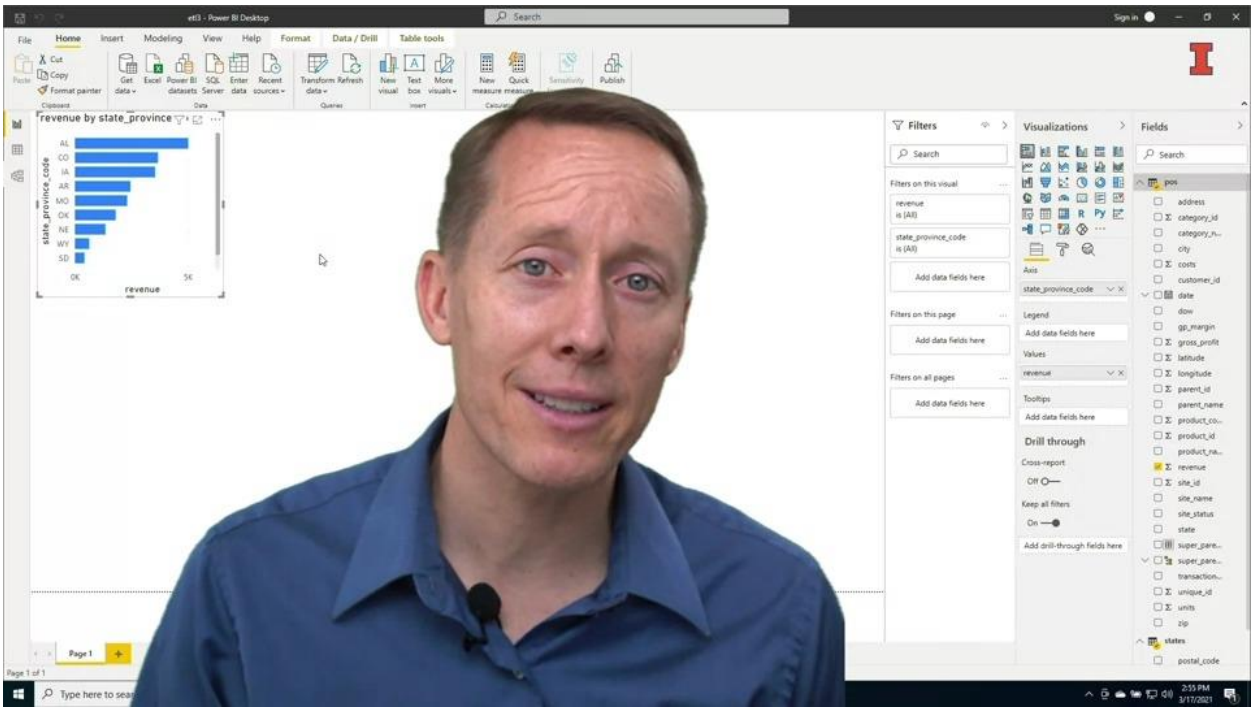
Now you can see in this POS data we've got this new column that states and it says we've got a table in every row here, so let's click on that arrow here. This allows us to indicate which columns we want to keep. Well, I really only want to keep the state province code. I don't want to keep the postal code again. We've already got that in the zip, and I don't want the state province; that's a longer name of the code. I'll just make sure I'm only selecting the state province code, click Okay.



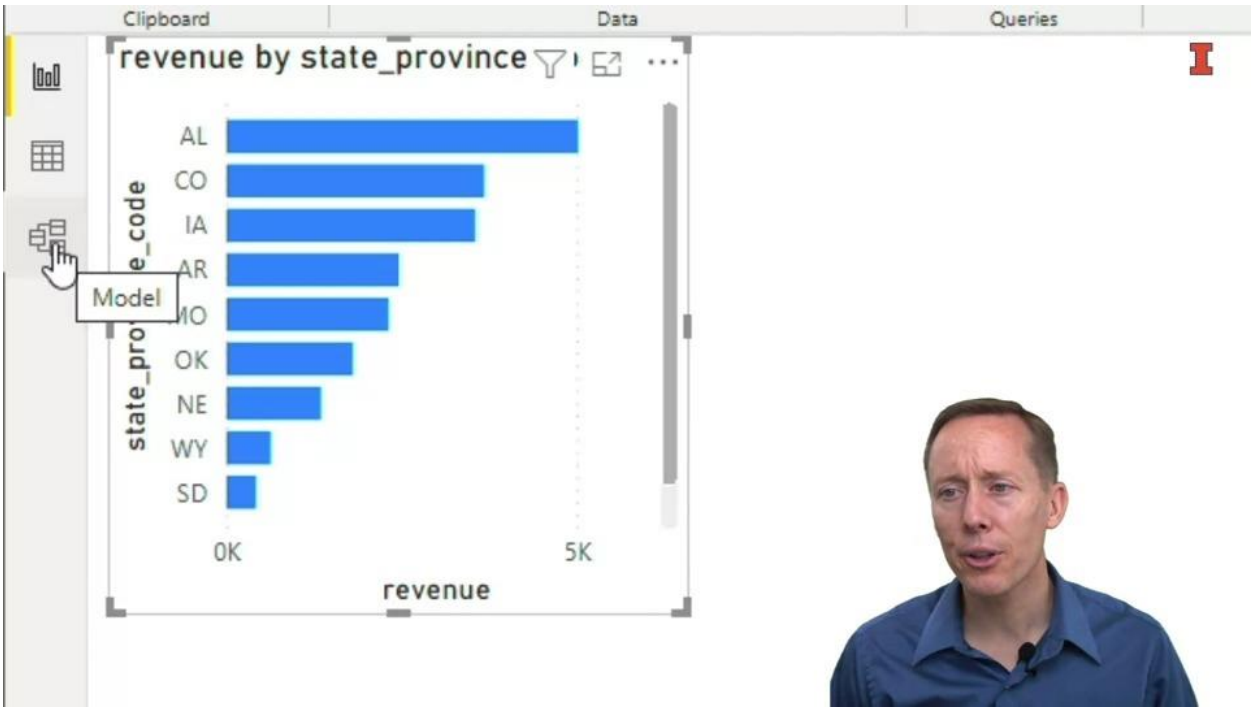
Now I've got that two digit abbreviation for each state in here, and it's a long name so we double-click on that and rename it just to state. All right, perfect. We have now created a new column in the POS data set. I can close and apply this.



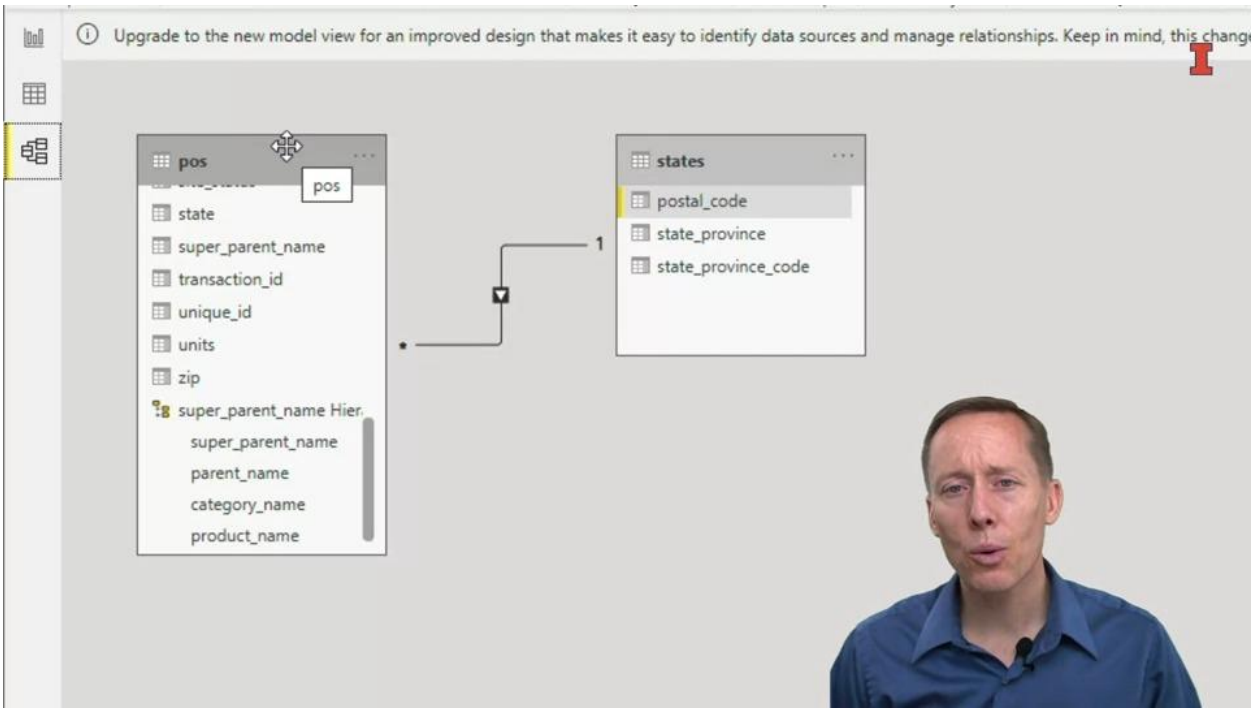
Now you can go over to the POS data and see that state is in here so we don't necessarily have to join it together using a data model.



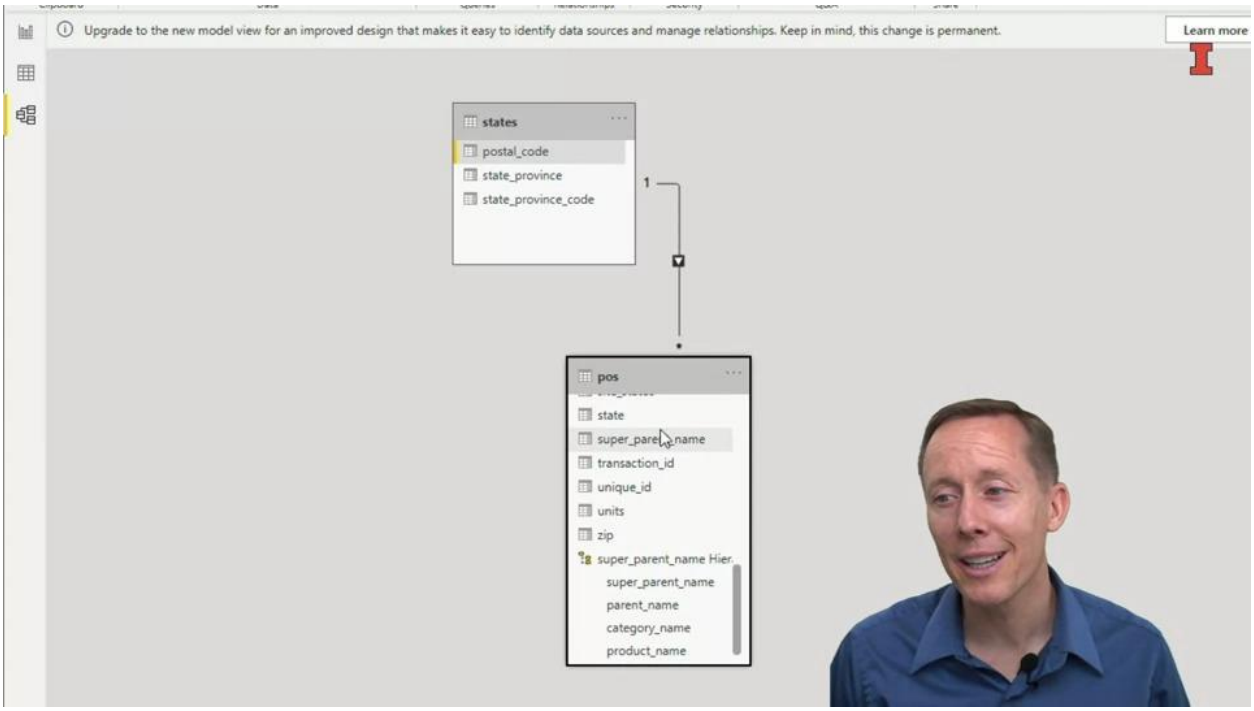
At this point you might be wondering, why are there two different ways to join data? Well, there are advantages and disadvantages to both approaches. As I've already mentioned, adding a new column to a data set can increase the storage size of that data set and all of this file that we're working on and that's not a good idea especially if we're dealing with large amounts of data.



The benefit of the data model is that we don't have to create more data to store. On the other hand,



if you are using a data model to join data and you've got four or five different tables in here, you may have this long hierarchical relationship and that can take a long time for queries to run and also just be overwhelming. What's often done is a combination approach, and it's called a Star Schema in which you've got one main data set

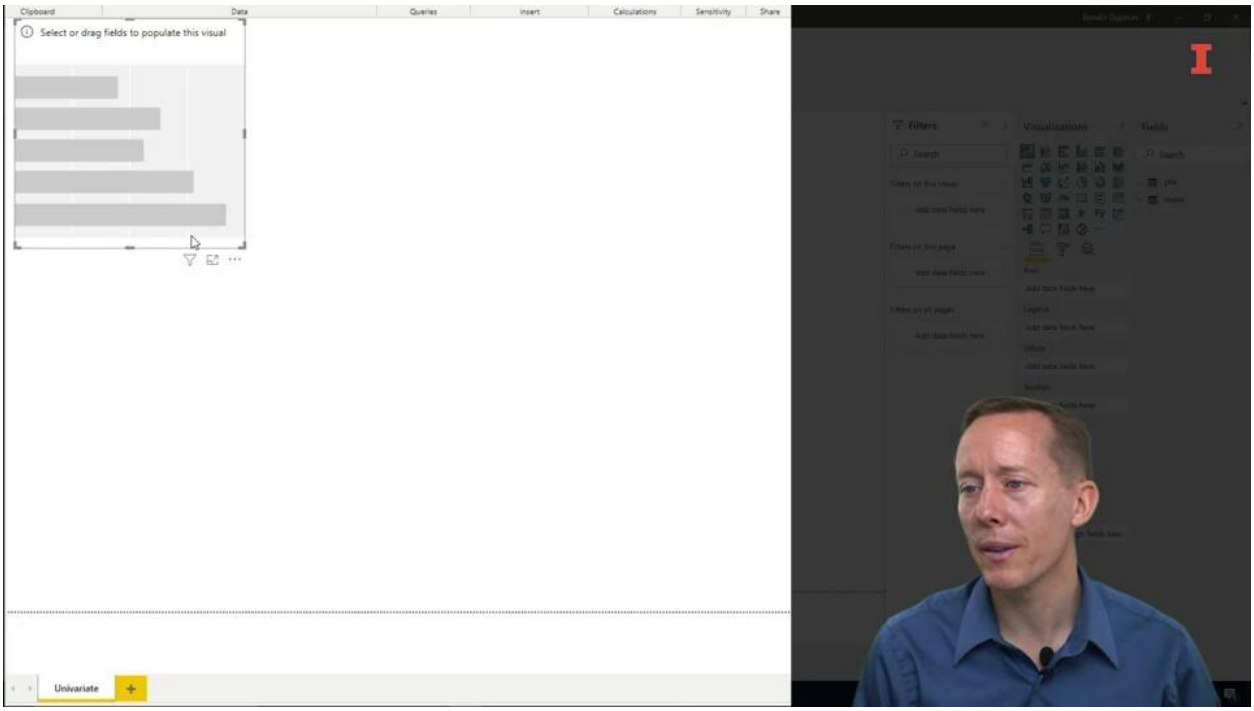


and any other data sets that you want to join together to it, you'll just make sure that they join directly to that key data set. In that way it won't have to query a lot of different hierarchical relationships in here. If you need to add in a column of data from, so you can match one table with another that is not this key table then you'll add in a new column into this key table in the middle. Anyway, that's a star schema. There is an introduction for joining data sets together in Power BI. It's really pretty easy. The benefit is that by bringing data together from different sources, it can really open up the insights that you can gain.

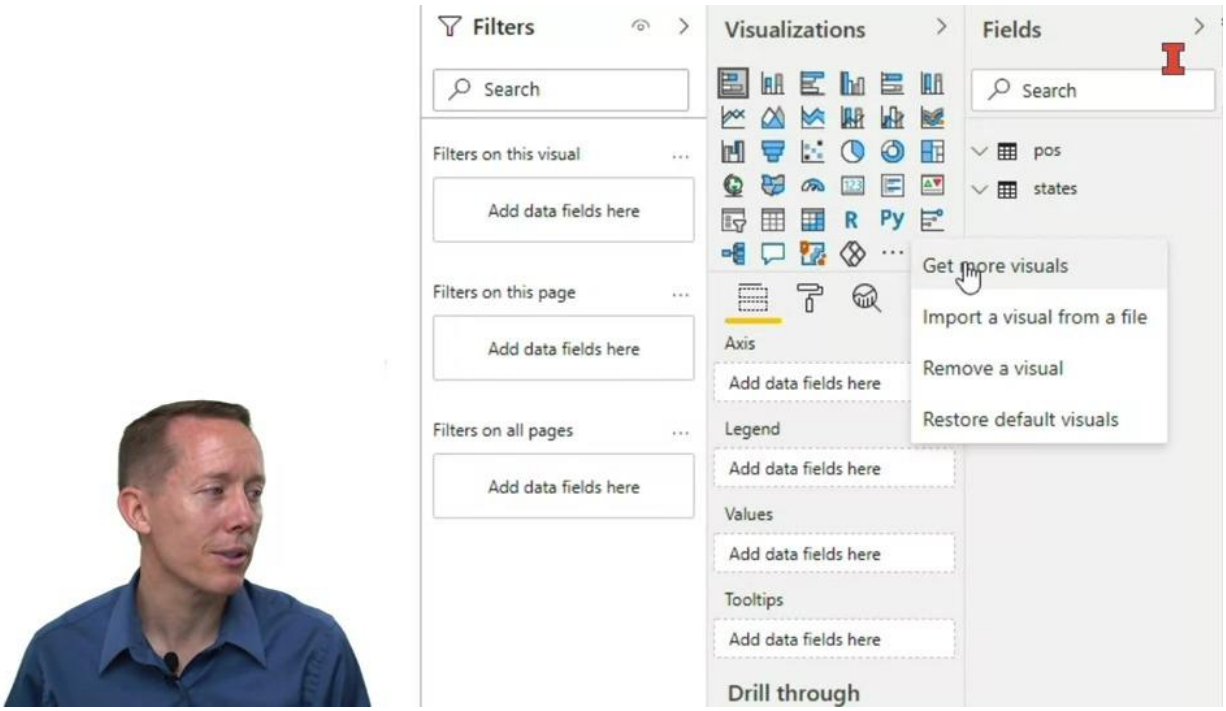
Lesson 2-3: EDA with Power BI

Lesson 2-3.1 EDA 1: Univariate Plots for Numeric Data: Histograms and Boxplots

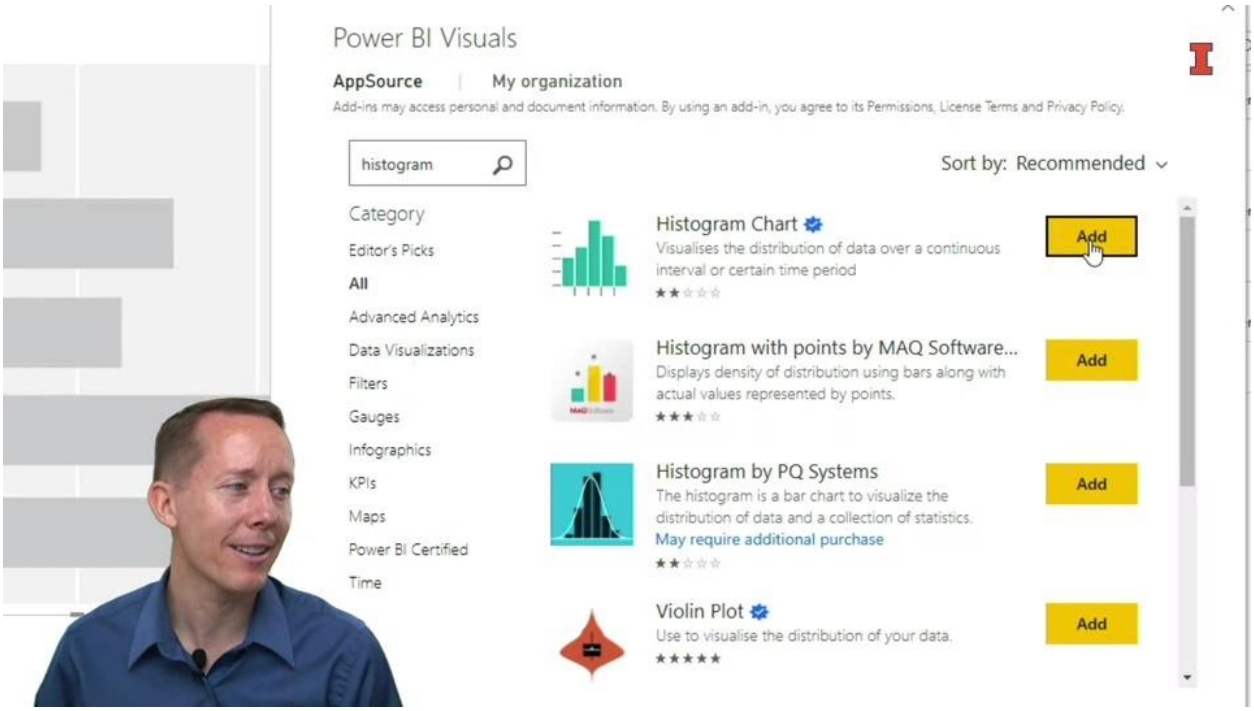
In this video, we focus on creating visualizations with Power BI. Specifically, we will demonstrate how to create visualizations for exploring one column of data at a time, which is known as univariate data analysis. In this video we'll focus on creating histograms and box plots, which are used for columns that have a numeric and data type.



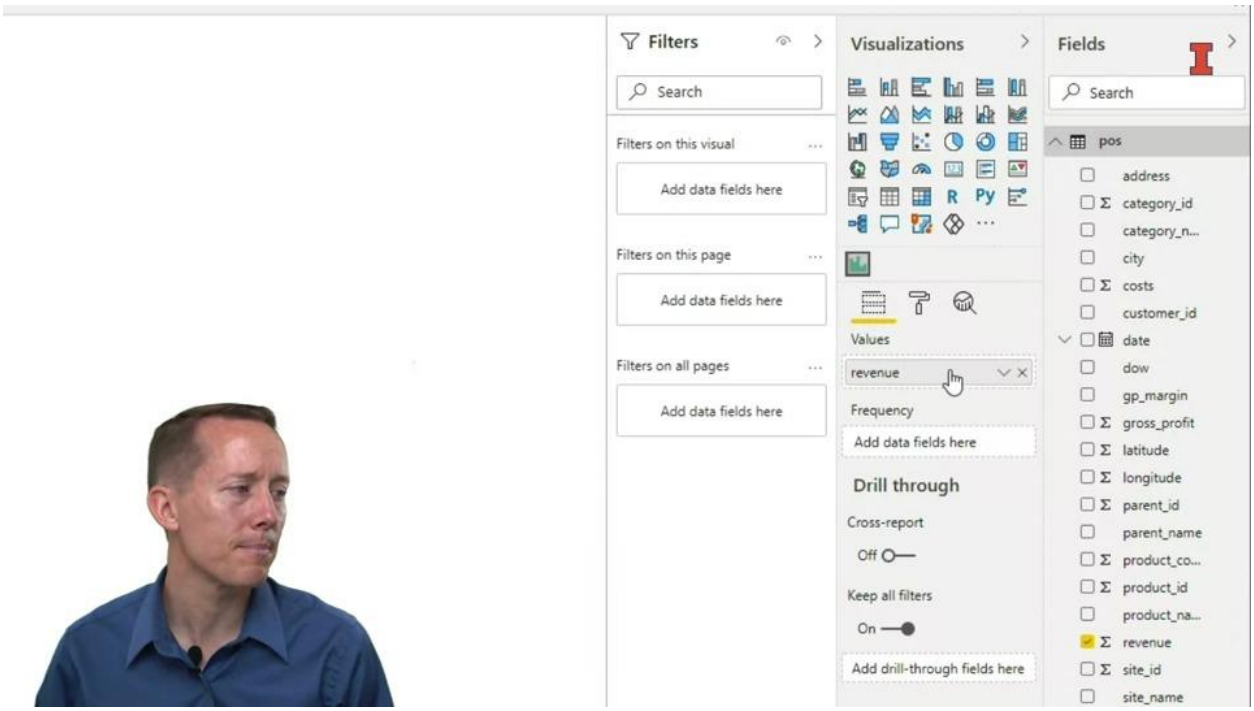
The first thing that we should do is open up Power BI and I'm going to open up a project that I've been working on which is called etl4. In this project, you can see I've got two data sets, I've got POS for point of sale data and states data. Now we're going to create visualizations in this main area here in the Canvas. But before we do that, let's go ahead and just name this tab, Univariate, so that we can remember as we add more tabs, what is on each tab. Creating visualizations is super simple in Power BI. We've got a variety of default visualizations that come built into Power BI. This first one here is a stacked bar chart, you can see we've got line charts, we've got pie charts, doughnut charts, tree maps and many others. Let's just go ahead and illustrate how to create a basic chart. You simply click on it and you'll see that a placeholder shows up on the Canvas, which we can resize, and that's often a good thing to do when you're creating the chart



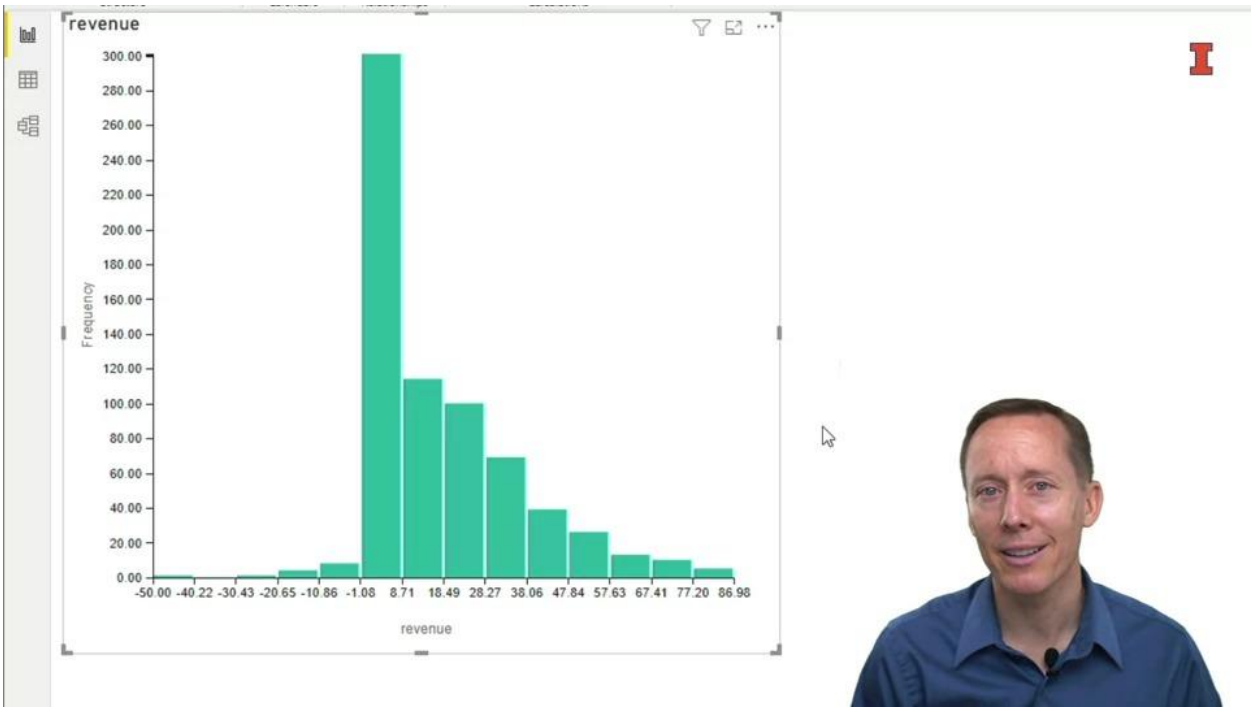
and then what we will do is we will add fields to the access legend values tool tips and any other placeholders that show up in this area below the visualization. Now we want to create a histogram, so as you look through here, you'll notice that there isn't a histogram that comes built in to Power BI. Fortunately, we can add in more charts by clicking on these three dots here that says get more visuals and then select get more visuals.



Now, in order to get more visuals using this approach, you'll have to sign in, and if your organization has a Power BI account, you can sign in with those credentials. I'm signed in with my organization's credentials. Now that I'm in here, I can go ahead and just search for a specific type of chart. I'll type in histogram, there are a variety of histogram charts that I can choose from. I'll just select this first one and click add and it tells me that this visual was successfully imported into this report so I'll click "Okay".

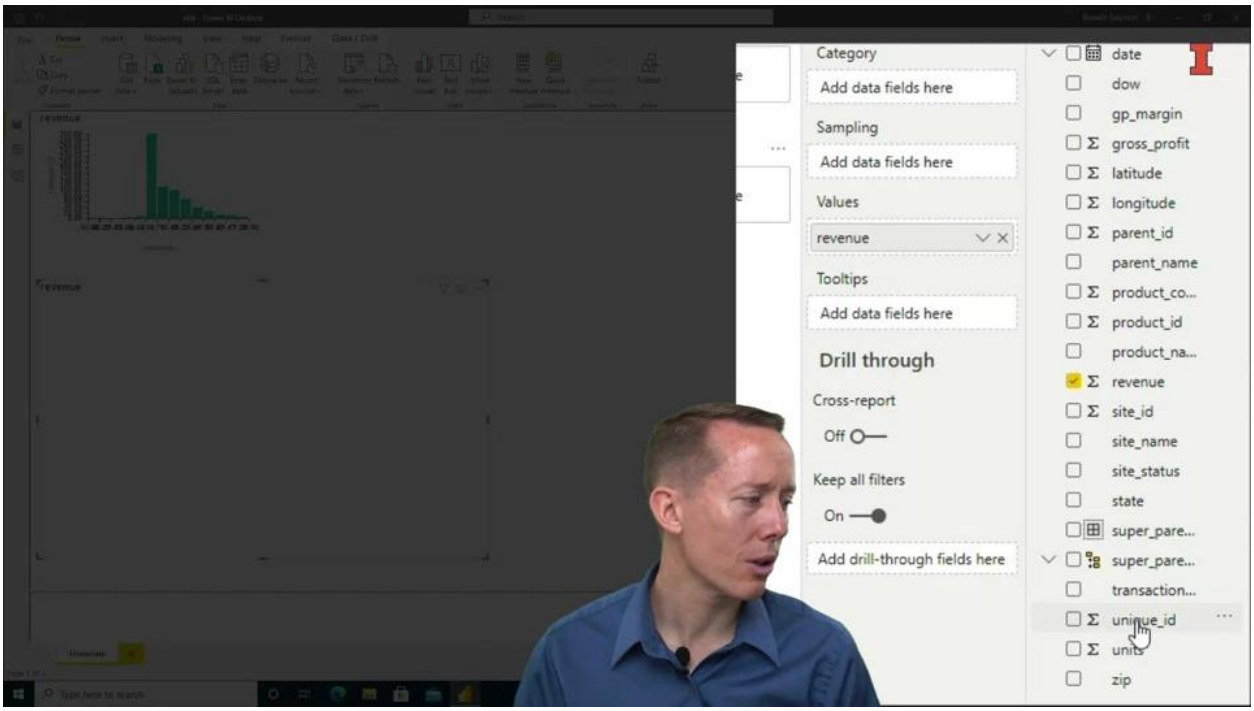


Now you can see I've got the imported visualizations down here below. Let's go ahead and just change the type of this chart by simply clicking on "Histogram", and you can see that updates to a histogram type chart or a different template here and now what I need to do is add column names to the values and perhaps frequency box as well. I'm interested in exploring revenue, so let's go ahead and open this POS data and go down to revenue and drag that into the Values box

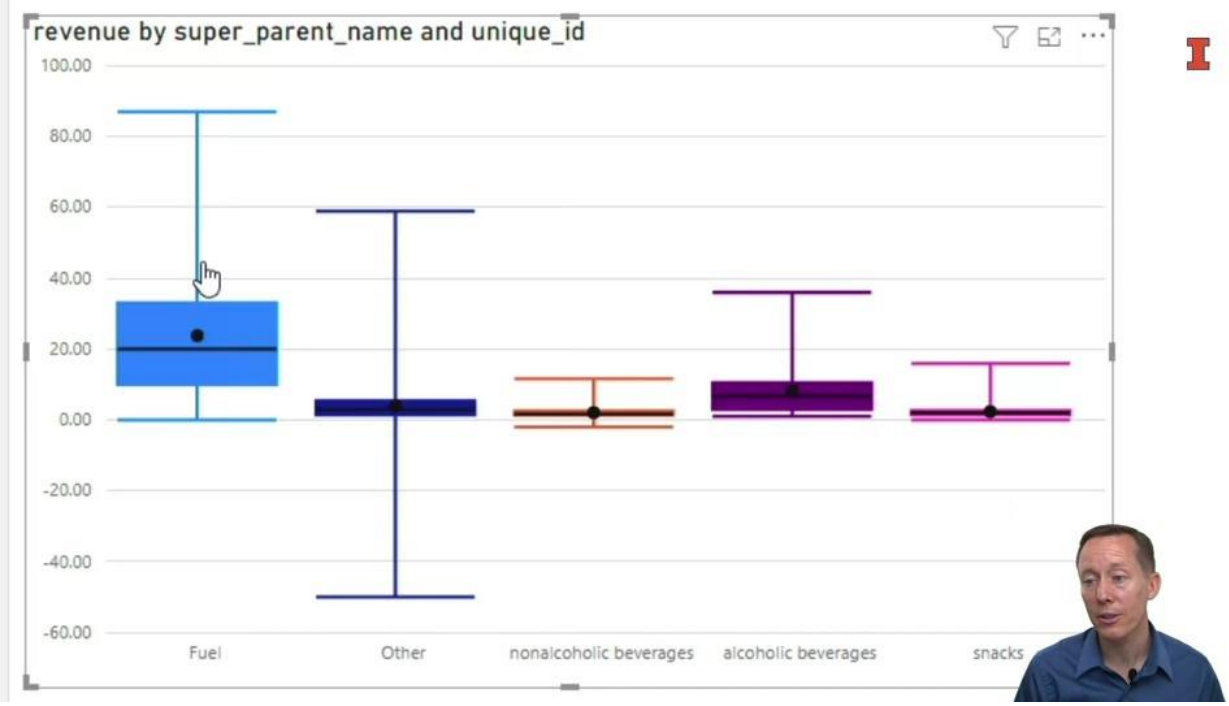


and there we have a histogram. Let's take a minute to just identify what a histogram is doing. A histogram takes a continuous variable or a column of data that is numeric and it divides it into equally sized bins and in this case here, the default looks like it's about 12 different bins or so. Let's just hover over this top one, it stands out the most. You can see that it says frequency is 301, meaning there are 301 of approximately 3,000 observations that fall into this range of negative one dollar and eight cents to eight dollars and 71 cents. That's not for a transaction necessarily, it could be but if a transaction has more than one line item, then it will just be for that one line item. You can see that the range here is negative dollar eight cents to eight dollars and 71 cents. The next being right above that only has 114 observations and that ranges from eight dollars 71 cents to $18.49. It gives you a real quick overview not only of where most of the observations fall, but also how that is distributed. You can see there are a lot more positive values relative to negative values. That's a histogram. Another great chart for exploring numeric data is a Box and Whisker Plot.

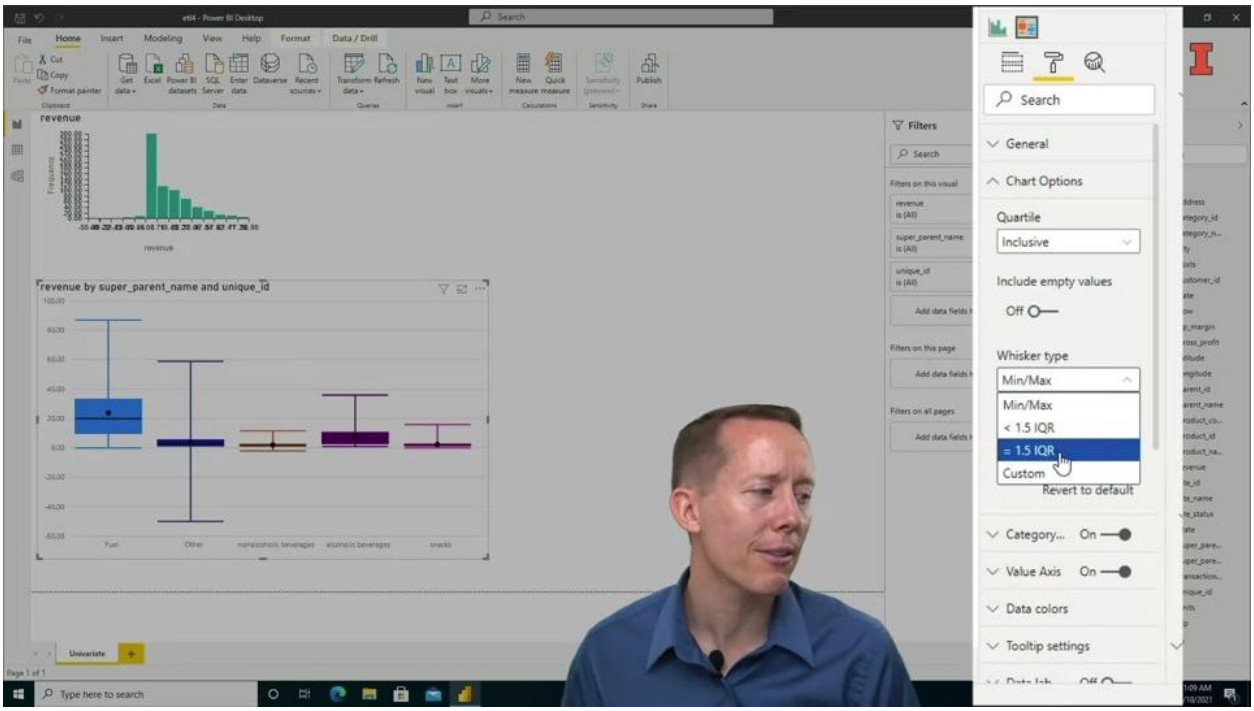
You'll notice that there's not a Box and Whisker Plot built into Power BI by default, so let's go ahead and add one in. I will search for just Box and sure enough, I get at least a couple of different Box and Whisker Plots here. I'm going to go with this one that has a checkmark by it. Click "Add" and we can see that this was successfully imported. We'll click "Okay", and I'll go ahead and make sure I'm not clicking on the histogram click off of that and actually, I might want to even resize this now that this looks how I want it to look, I'll just shrink it down. Now I will add in this Box and Whisker Chart and let's go ahead and make this bigger



and we can see that this chart has four different potential variables that I could add in here. Let's start with the values. I want to focus on revenue, so I'll drag revenue into there and nothing shows up and that's because Power BI is really good at aggregating data and if we click on this down arrow next to revenue, we can see that what it's doing is it's summing up all of the revenue observations into one data point so that will not allow us to create a Box and Whisker Plot. We need more than one observation. If we wanted to, we could average or minimum, maximum, you can do other aggregations, but sometimes there will be an option that just says don't aggregate and if it had it, then I would select that. In this case here though, let's look at the sampling. Let's try using sampling and I suspect that what we need to do is identify that there is a real label for each of these observations and we want to identify that we're using a row label as a unique observation. Let's go and take unique ID, drag that to sampling.

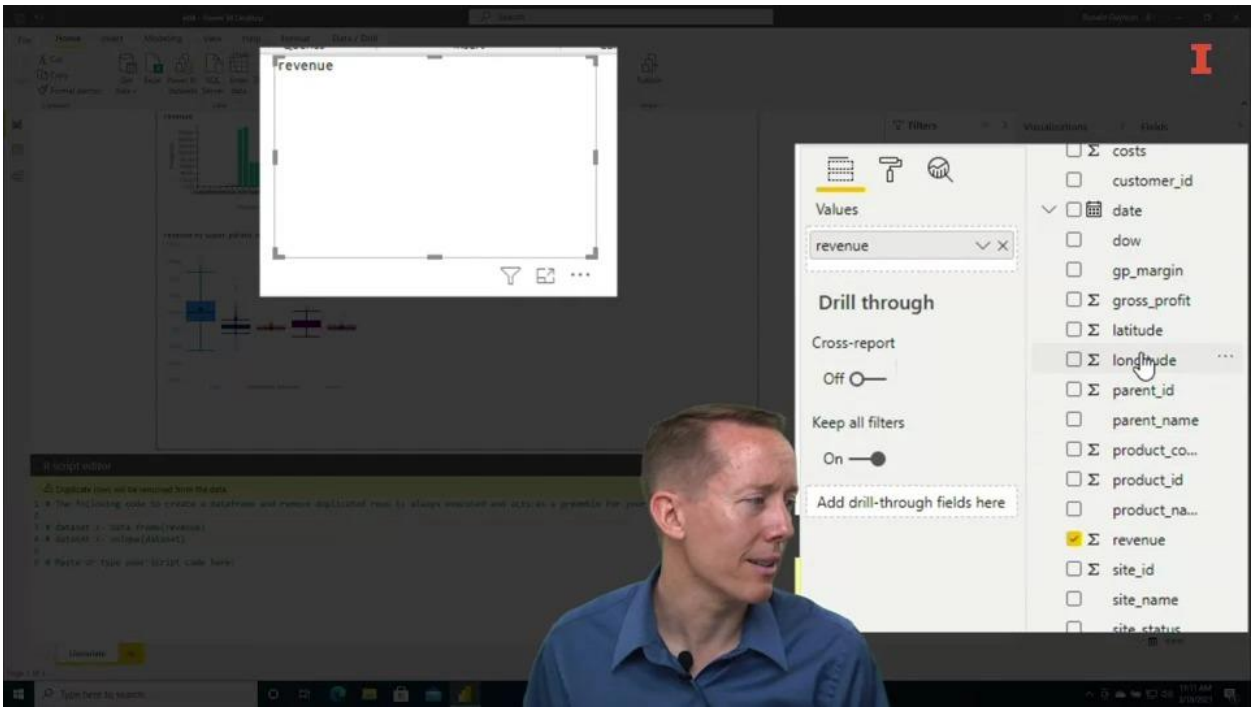


Still nothing shows up, and we can see that by the icon here, it looks like this is intended to create a box and whisker plot for different categories. Let's add something to the category box there. I'm going to select super parent name, drag that into there. All right, there we are. This now is a box and whisker plot for each of the five different super parent name categories that we created, fuel, other non-alcoholic beverages, alcoholic beverages and snacks. It's pretty cool, we can see the median by the line there, and then the dot is the average, then the whiskers extend out to looks like the extreme observations. Now, in all of these plots, we can customize them.

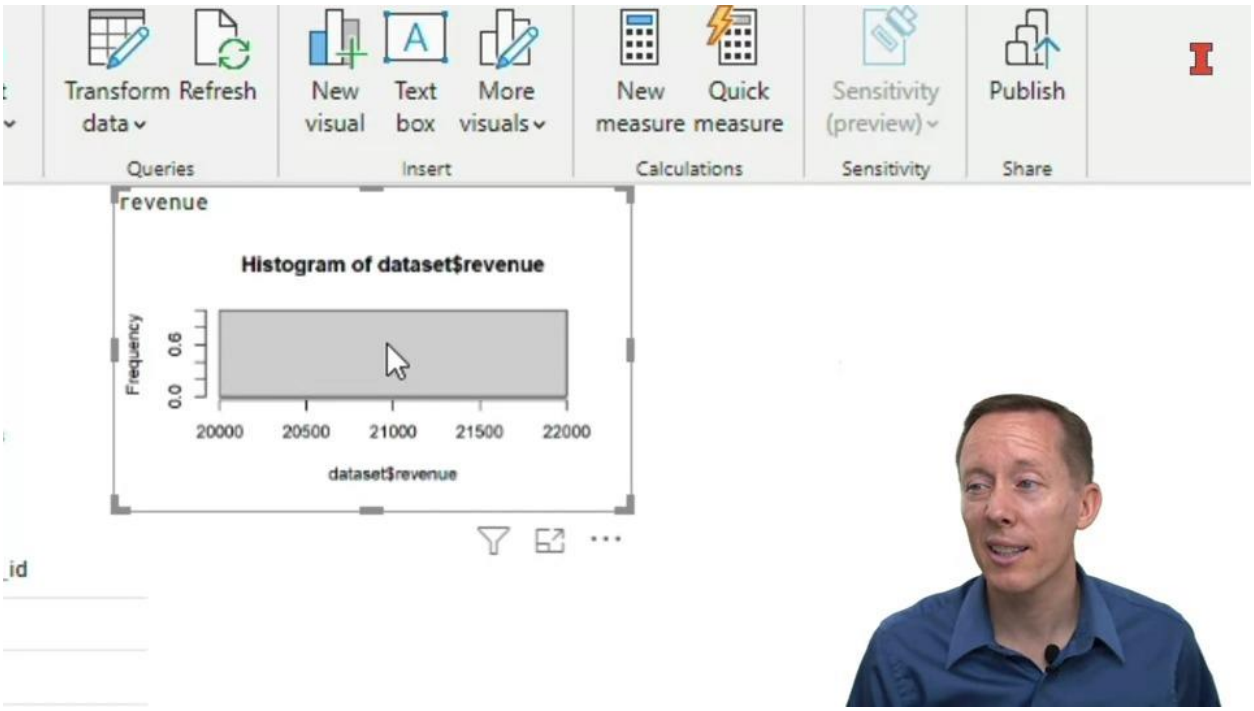


If you go over to this ruler paintbrush, this is how you format plots. With that box and whisker plot selected, I will click format. You can see all the different ways that I can format this, I can add a title, I can change the background, the aspect ratio, data labels, so on and so forth. I'm just going to go into chart options here and I want to maybe identify that these whisker types, I don't want to go all the way to the min the max, I want to do what is often done, which is equal to one and a half times the interquartile range.

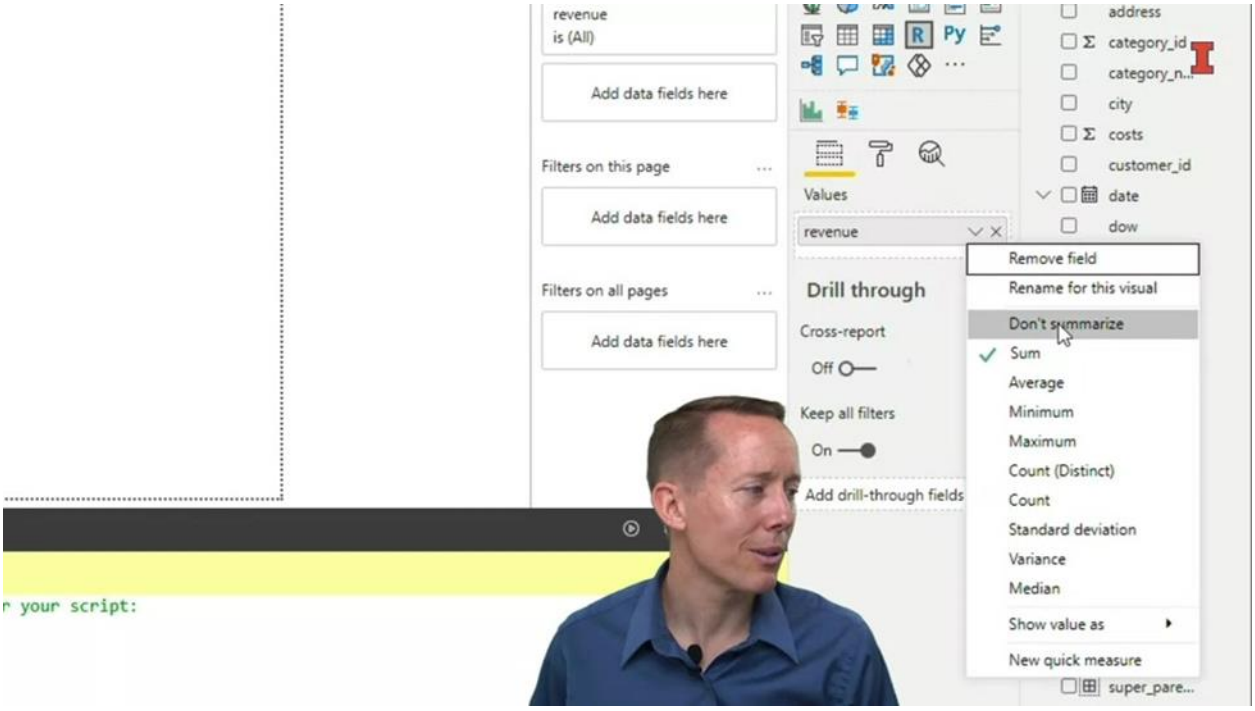
I'll select that, you can see this changed a little bit. These whiskers now, the interquartile range is the distance between the top of the box and the bottom of the box, and that's the twenty third percentile and the seventy five percentile or third quarter in first quarter. We're taking one and a half times that distance and making these whiskers go out that far. This just shows, that's why it's symmetric now, it's going out there. It just shows us that there are observations in that range.



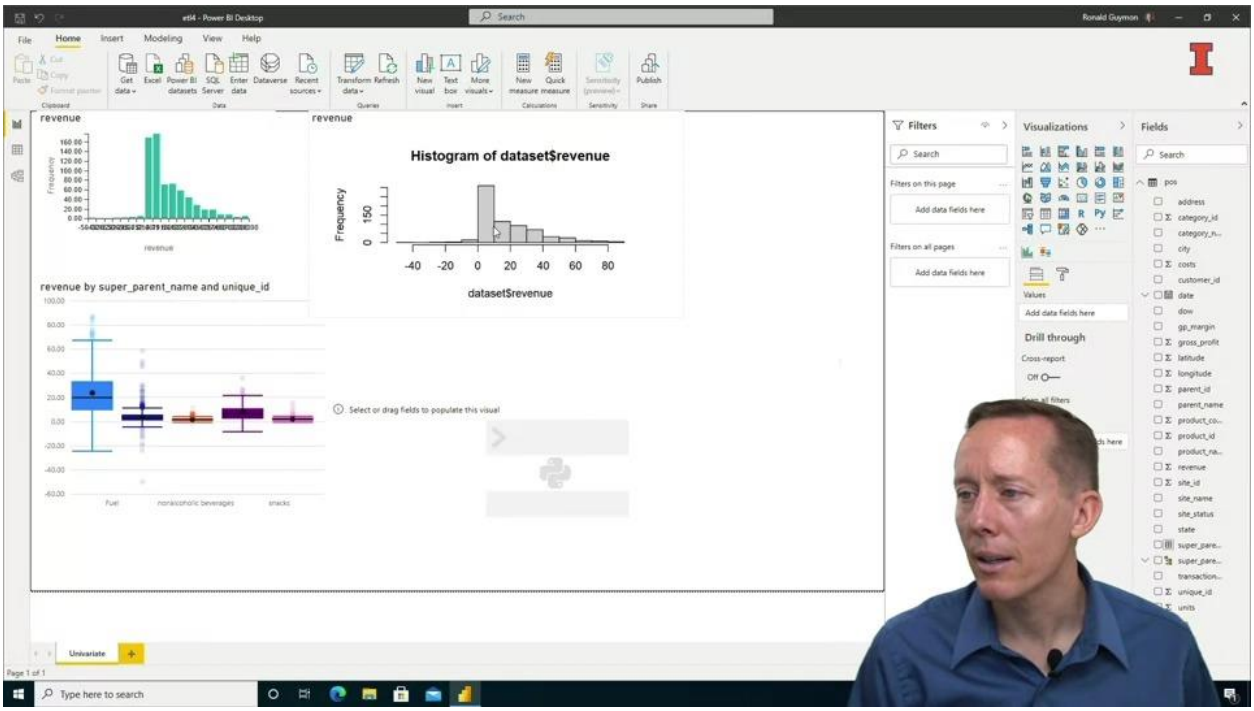
Now, we can also go in here and say, "Hey, I want to toggle on the outliers." When I do that, you can see that now I have the stubble. You can see that there are observations that go beyond that range and that allows us to get a sense of how the data is distributed in each of these columns here. We could also customize the histogram if we wanted to by indicating. We're going to general hair, we can say, "Hey, beans instead of maybe 12 or so, I want 20 beans." I type in 20 and hit enter and you can see that it added more beans in there. Creating and editing charts, formatting charts and power BI is so simple, it's awesome. Let me show you one other type of chart that we can add in here. Let me just resize this, make it a little bit smaller. Make room for another chart, and this other type of chart is in our chart. You can see you can add in our scripts or Python scripts let me just click r. You'll get this prompt to enable it, because it could damage your computer, someone is doing something nefarious, so I'm going to go ahead and click enable. You can see it creates a placeholder here and there's just one item for this type of chart so far, and that is a value. Now, I could actually add in more than one value, but I'm just going to start with revenue. If I wanted, I could add in longitude or some other column as well.



When I do that, it populates this r script editor and it's basically saying, "Hey, we're going to take this revenue, create a new data set and it will have revenue as a column in there." Now, I can click in there and I can type our code to create any type of plot I want. In this case, I'm going to type, I'll just do hist for a histogram and I will say data set and the column is revenue, and click run. You can see it doesn't look very pretty there,



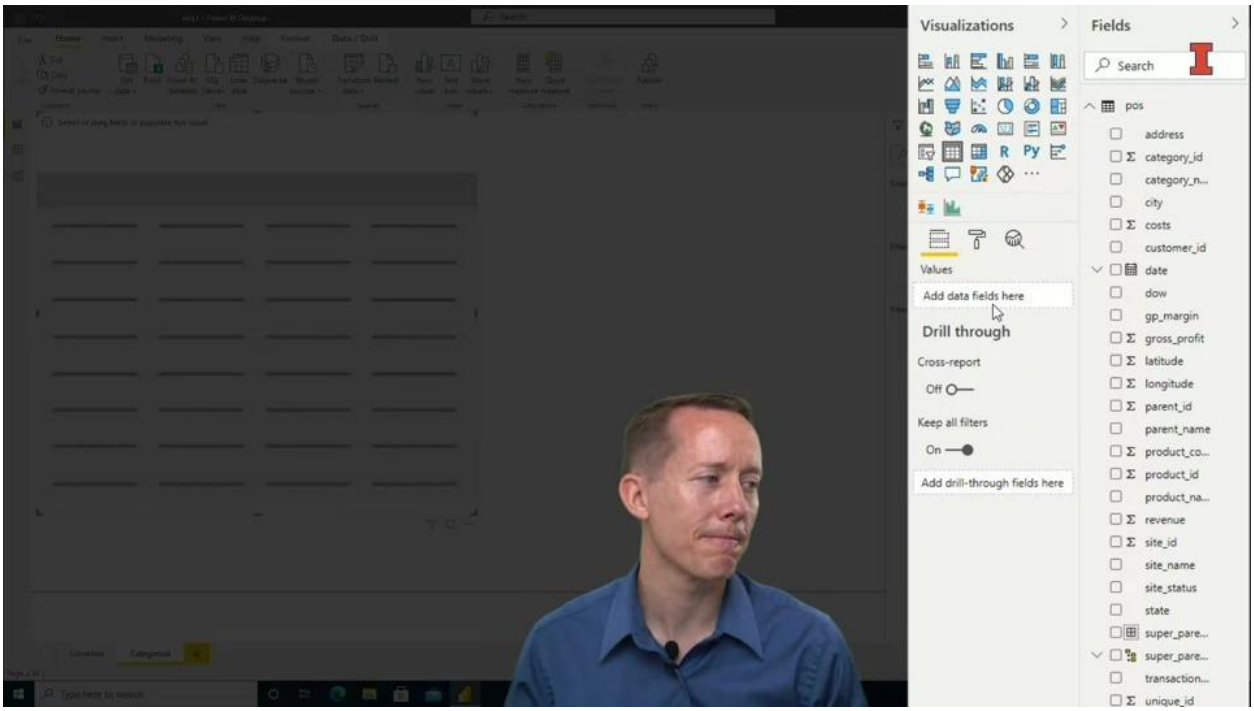
and let's go ahead and explore why by clicking on this down arrow next to revenue. Sure enough, it is summarizing it all into one data point, one observation, so let's click on the don't summarize option.



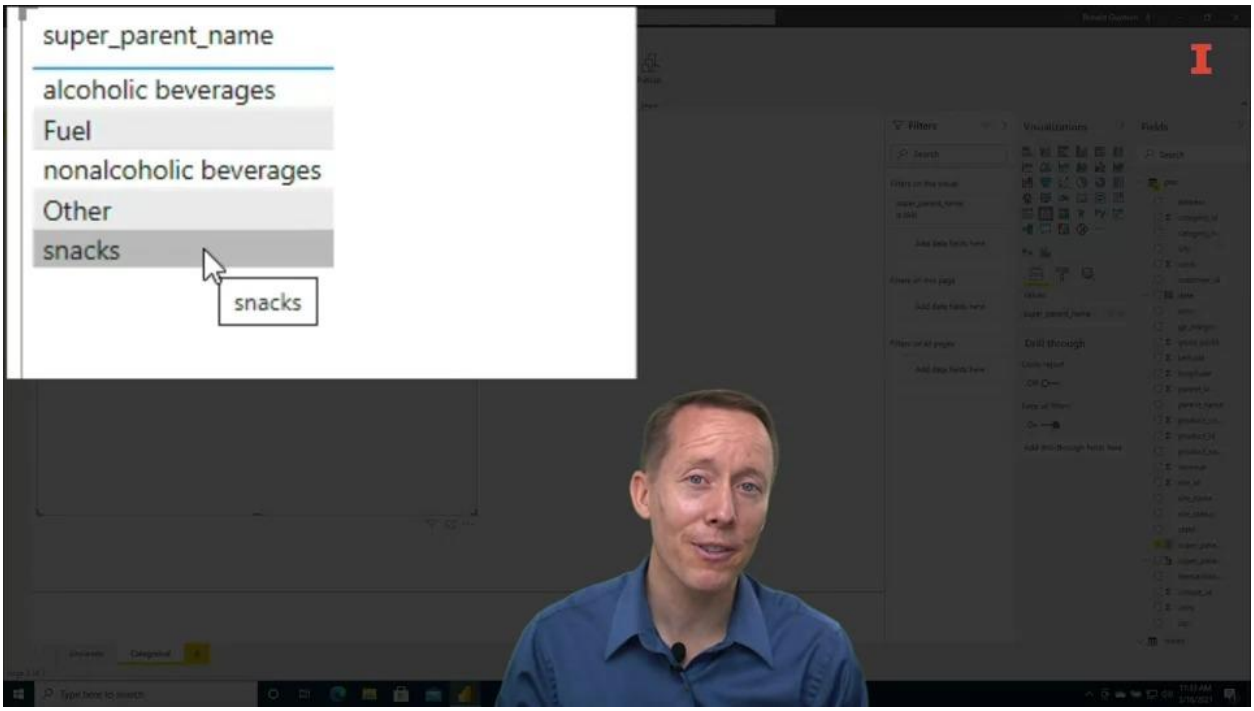
There we go, this looks like a much better histogram that's somewhat similar to this other histogram that was added. The point is, if there is a type of plot that you really like using from our Python, you can add that in. Now, one thing that you didn't see is that before adding in or being able to use our scripts in Python scripts, you need to make sure that you have these loaded on your machine. If you try to add in something for Python here, it will tell you, "Hey, first make sure it's loaded on your machine." If it is, and you probably won't get that prompt, but if it isn't, then you'll need to add it in and there will be help that will allow you to point you into how to add that software. You can see that creating visualizations is very simple. We looked at some uni-variate visualizations, you can see it's very easy to resize them. We can also move them around on the canvas here and display them in a variety of different ways. We can get rid of them by clicking on it, clicking on these three dots and clicking remove or simply hitting the backspace or delete key. It's so intuitive, creating charts in power BI. Those are introduction to creating charts as well as how to create uni-variate charts.

Lesson 2-3.2 EDA 2 - Univariate, Bivariate, and Multivariate Plots with Categorical

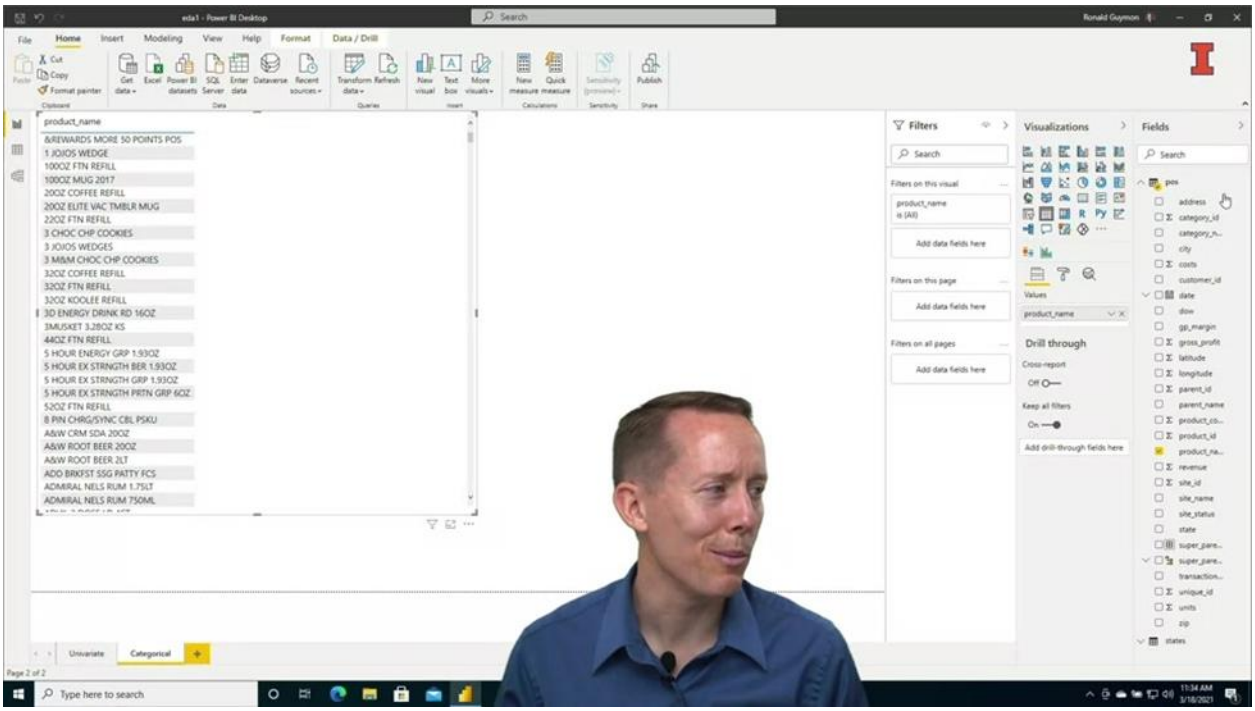
In this video, we'll talk about exploratory data analysis that focuses on categorical columns of data. When exploring categorical data, it's often useful to identify the cardinality or the number of unique values. We can do that in the Power Query Editor, but we can also do that with visualizations. In this lesson, we will first talk about data visualization tools for univariate data analysis with categorical columns. But we will also show you bivariate data analysis in which you analyze two columns of data together.



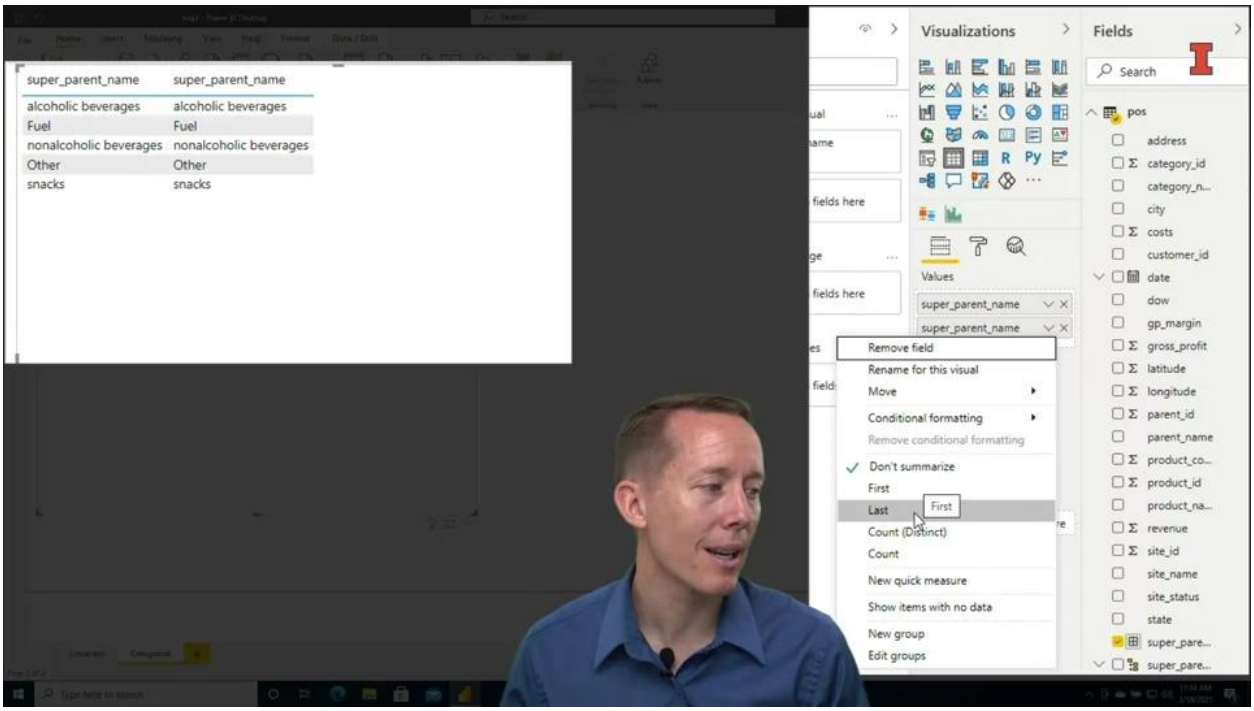
I'm going to open up Power BI here and open up this project I've been working on eda1. When I open it, I have to identify that I want to enable script visuals, meaning I have some visualizations from are in here. I'll click that. Here is the tab that has the univariate analysis on it. Let's go ahead and create a new tab at the bottom. Click that "Plus Button" and we'll name this categorical. Let's start by analyzing the super parent name in the point of sale data, so we've got the super parent name here. We need to decide how we want to analyze it. What I want to do is start with the table. Click on table. This adds a table visualization to my canvas here and I'll just drag super parent name to the values.



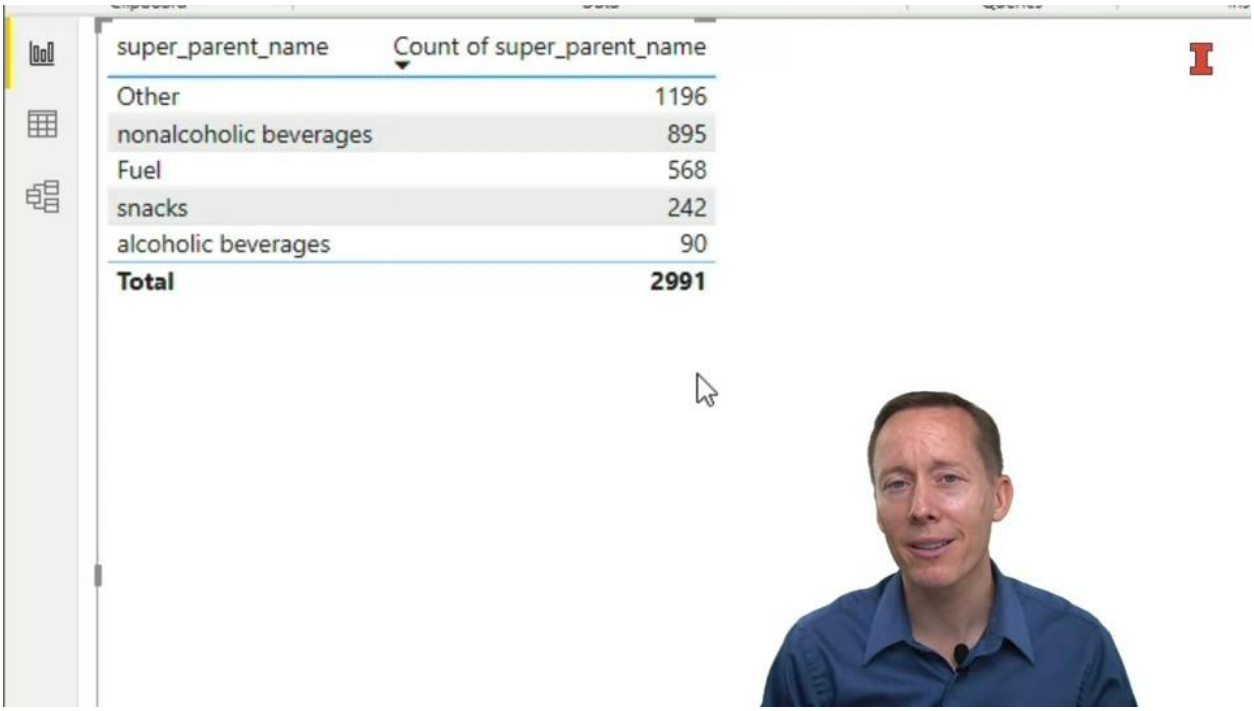
This is a categorical field because it's not numeric data. You can see that we've got five different super parent names, alcoholic beverages, fuel, non-alcoholic beverages, other and snacks. That's the first thing, is just looking at the number of different values in a column of data. This is relatively low cardinality because it only has five unique values.



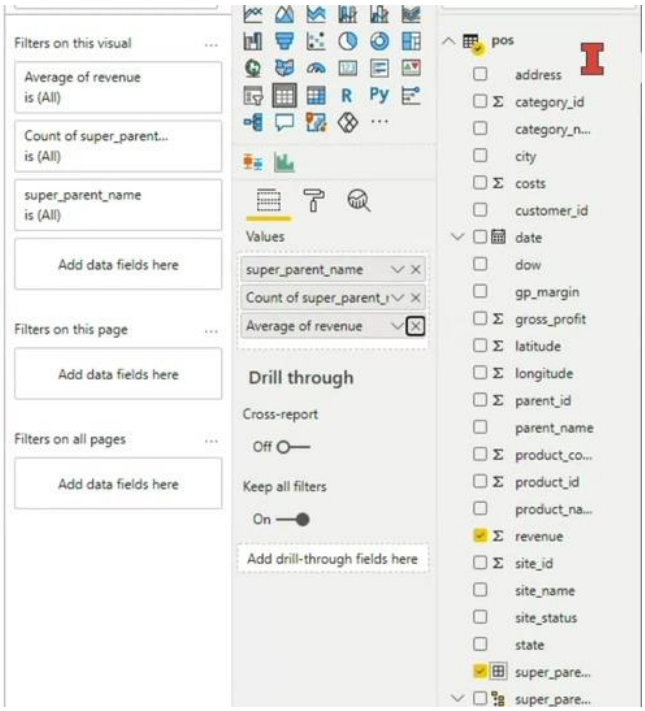
If we were to compare this to product name, let's take out super parent name, compare it to product name. You can see by the scroll bar here just scrolling down, we've got a lot of different values here. That can be overwhelming.



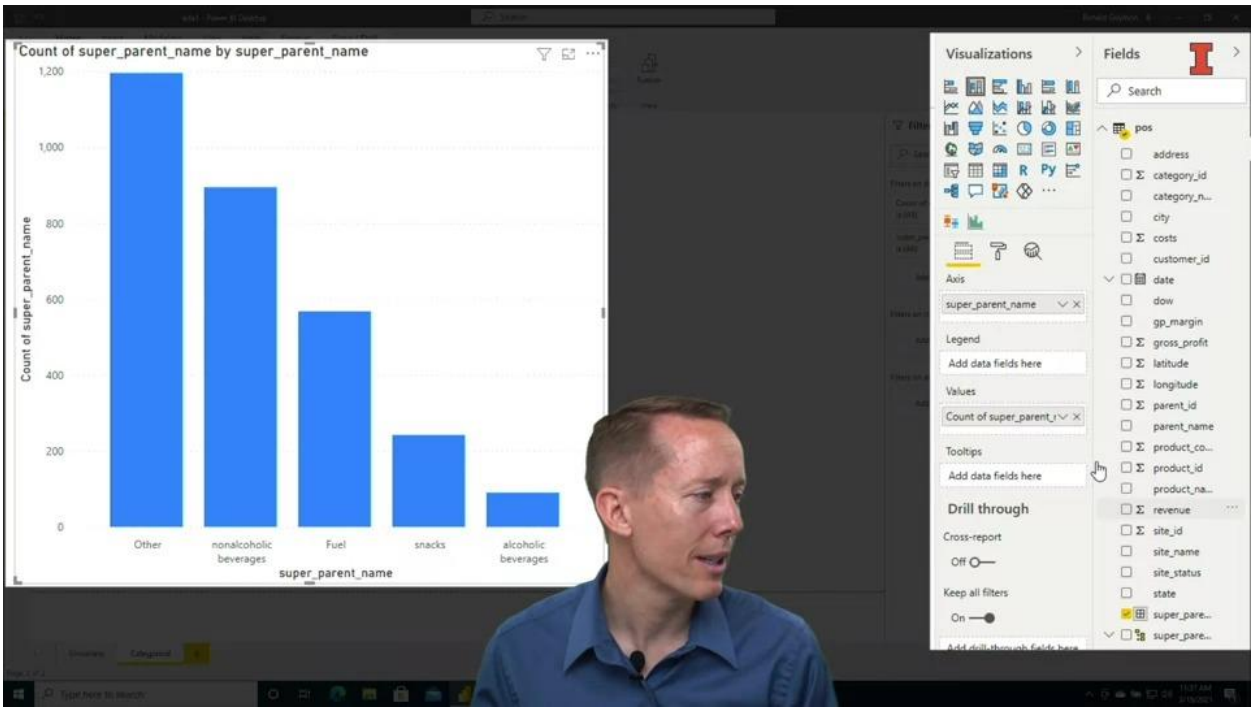
Let's start with a low cardinality category, which is super parent name. Once we have that, we can start by just looking at the count of observations for each of those values. The way we can do that is by simply dragging super parent name again to the values box, and we can see now it's just two columns with the same thing. But if you click the drop-down arrow next to the second one, we can indicate that different ways of summarizing this data and what we want is count. Click on "Count".



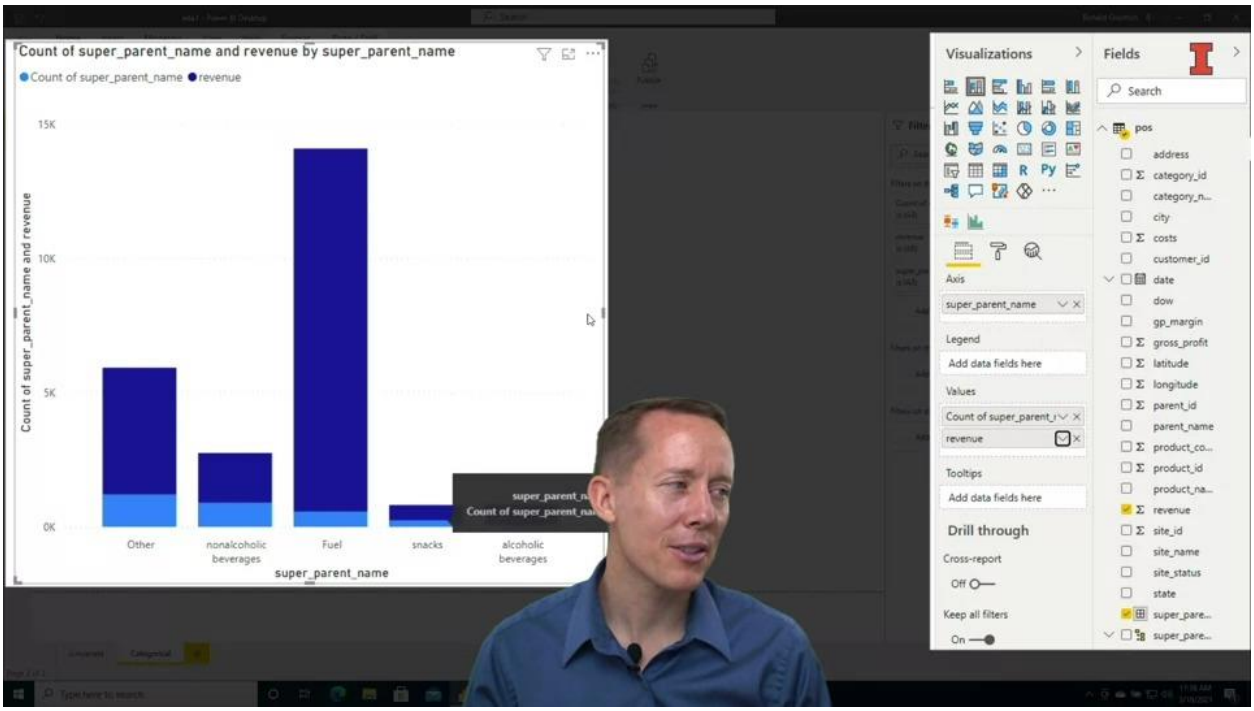
Now this table updates so that we can see that we have 90 observations that are alcoholic beverages, 568 that are fuel and so on, and we can see the total there. Basically in this, If you remember, each line of data is a line item that is purchased in a transaction. We've got 2991 items. We started with 3000 then we got rid of some outliers. You can see this is by default, ordered alphabetically. We could order it, if you it over the column name and click on the "Arrow", we can switch the ordering so that it's in a descending alphabetical order, or we can order it by the count of super parent name, which might be more useful in this case here. We can see the other makes up most of the observations. Then non-alcoholic beverages and fuel and so forth. This, I would say, is a univariate way to analyze categorical data.



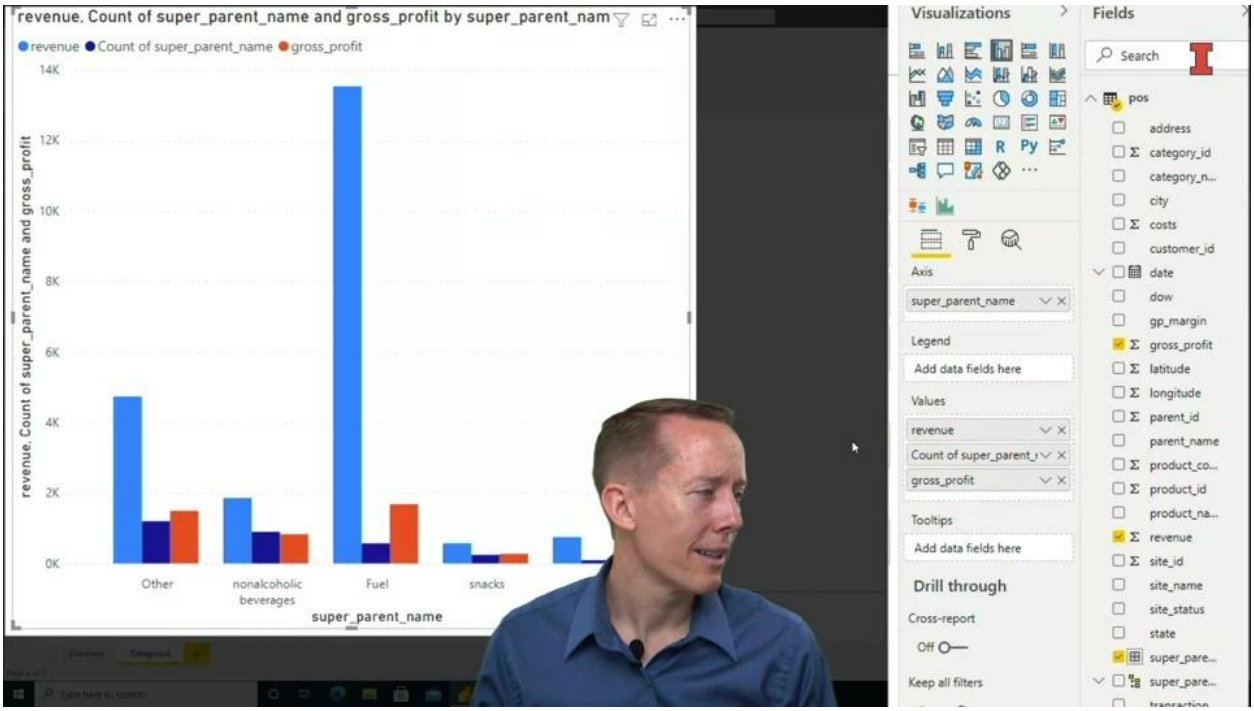
If we wanted to make this a bivariate analysis, we could drag in some other column and let's drag in revenue, since that's something that we're interested in. Now you can see that we've got a column that summarizes the revenue. You can see if you click on revenue, if you want to look at the average, we could use average instead, aggregate it to see on average what are people spending for each of these different super parent names. That's pretty cool. I'm going to go ahead and just get out of the average. Click that out of the values and start with this.



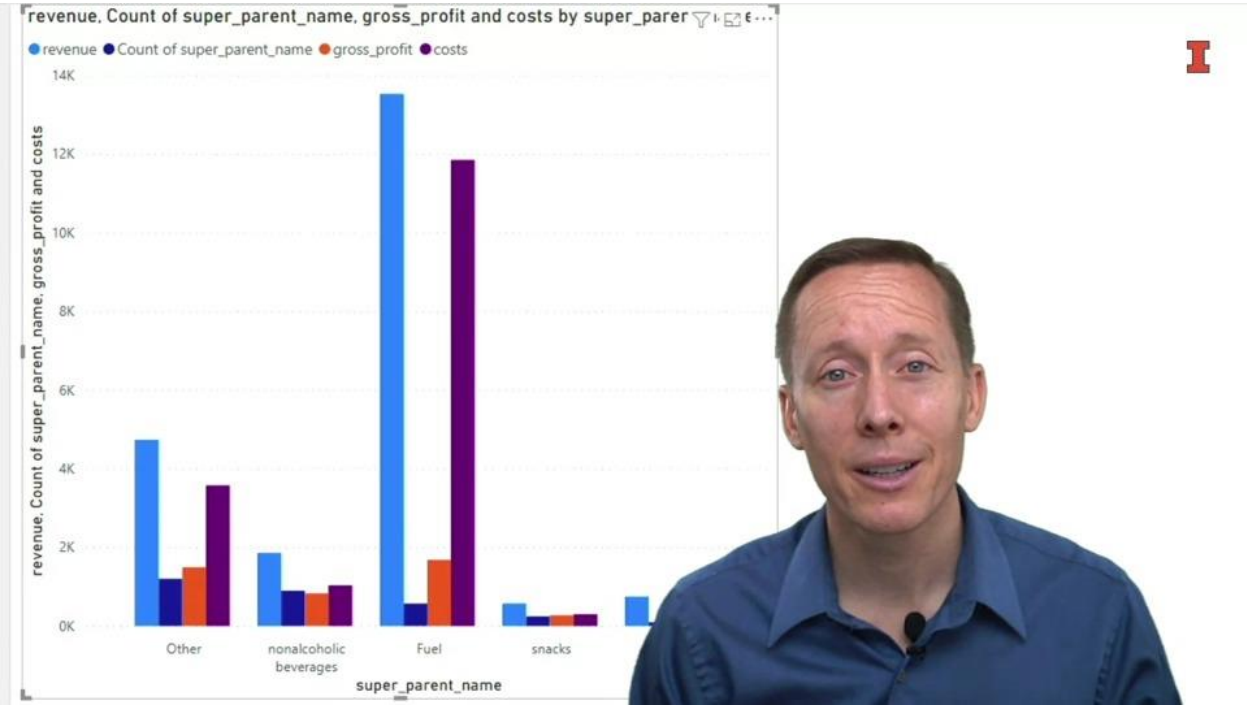
Now let's convert this table into a bar chart. The reason is because when we look at relationships between numeric values, they're much easier to see using visualizations. So I'm going to just click on this stacked bar chart. You can see it's horizontal, and now it shows us very quickly that other is about a fifth more than non-alcoholic beverages. We can convert this to a stacked or a vertical content very easily. You can see it's so simple to rotate among different types of charts here. What if I do want to look at revenue and count together? I can very easily do that by dragging revenue to this values box here.



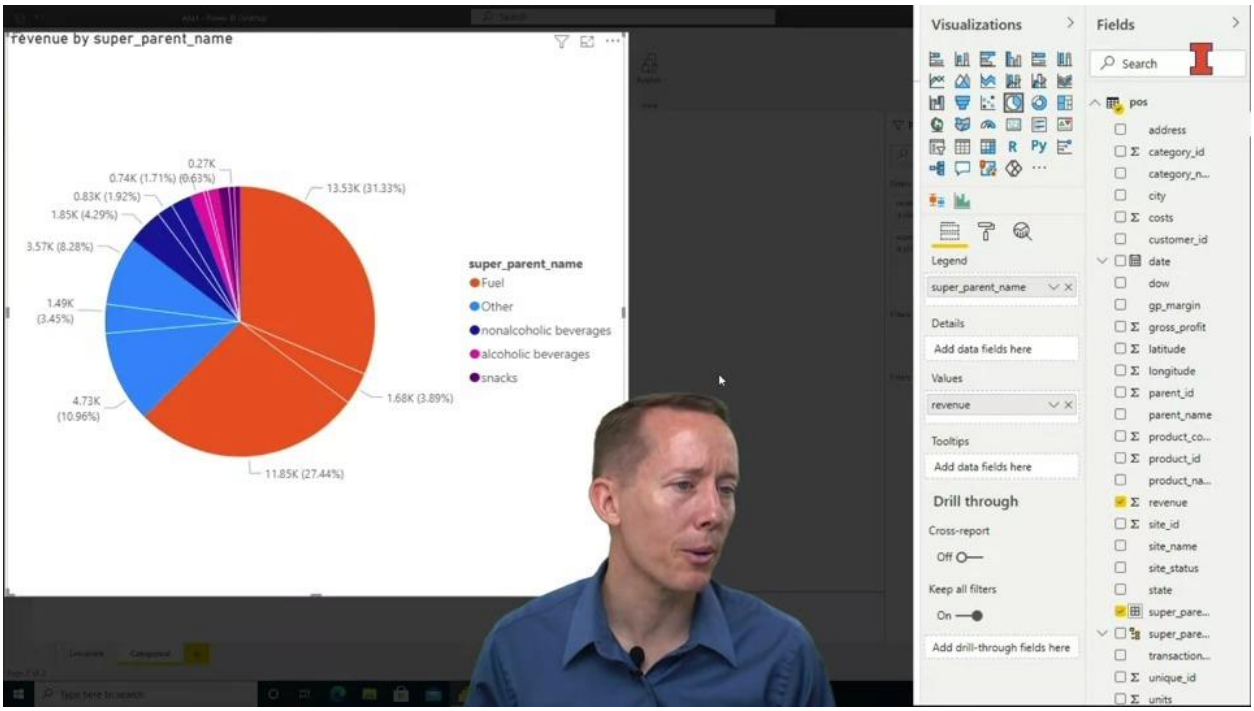
Now I've got a stacked bar chart where the light blue is the count and the dark blue is the sum of the revenue. Again, we could change these to aggregate it in different ways if we wanted to look at average amount spent on revenue. You can see the stacked bar charts can be helpful. I don't know that in this case, stacking count and revenue on top of each other is useful. Let's say, hey, we want to look at a grouped bar chart, clustered bar chart.



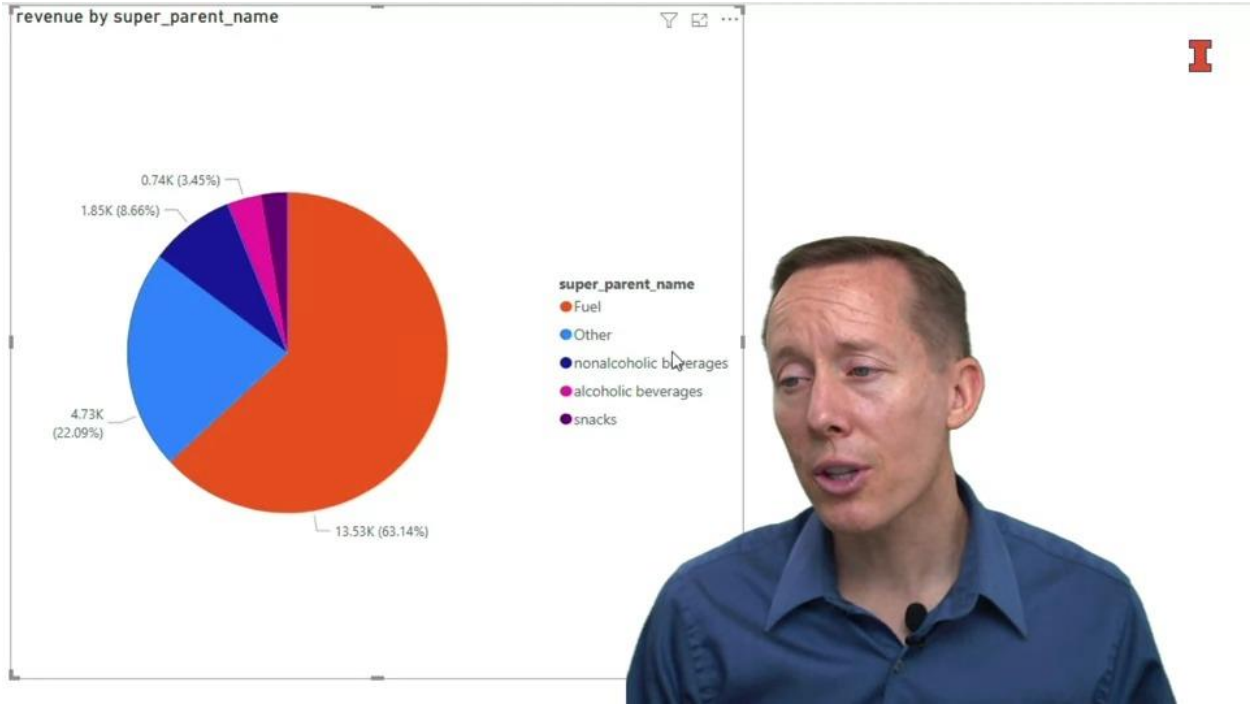
I can navigate that way to cluster bar chart and very quickly convert it to a clustered bar chart. If I want to change the order so that I have revenue before the count, I can just drag the revenue pill up above the count pill and that changes the order there. I can obviously add in a lot of other columns in here, so I can look at gross profit, I could look at the costs.



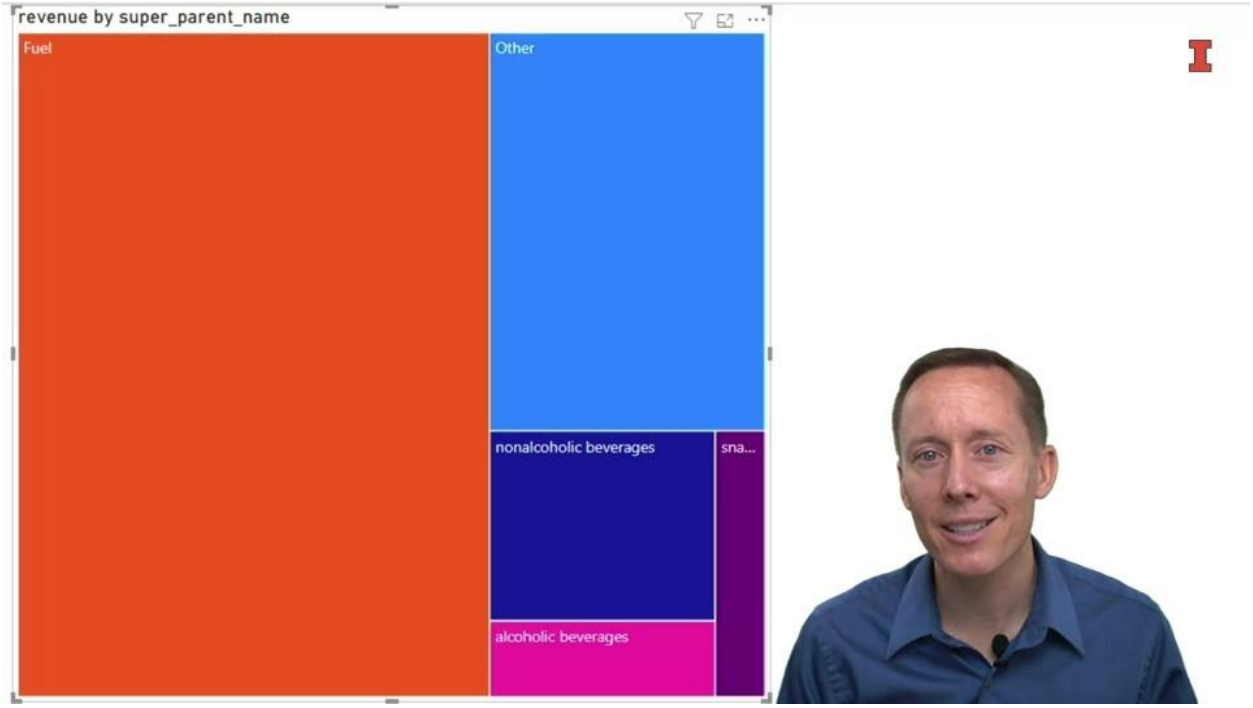
Pretty cool how we can see all of that data. I encourage you to think about the use of these visualizations. They can very easily become overwhelming and may communicate more than what you intend to communicate. The hope is that there are some intraocular traumatic impact. There's a conclusion that hits you right between the eyes, oftentimes we want to communicate that. Although if you want something for people to explore and really take time to look at, maybe you do want something that's busy like this. So I encourage you to play around with the different types of charts and the different formats for these charts and to think about how quickly each chart is communicating what you hope it's communicating.



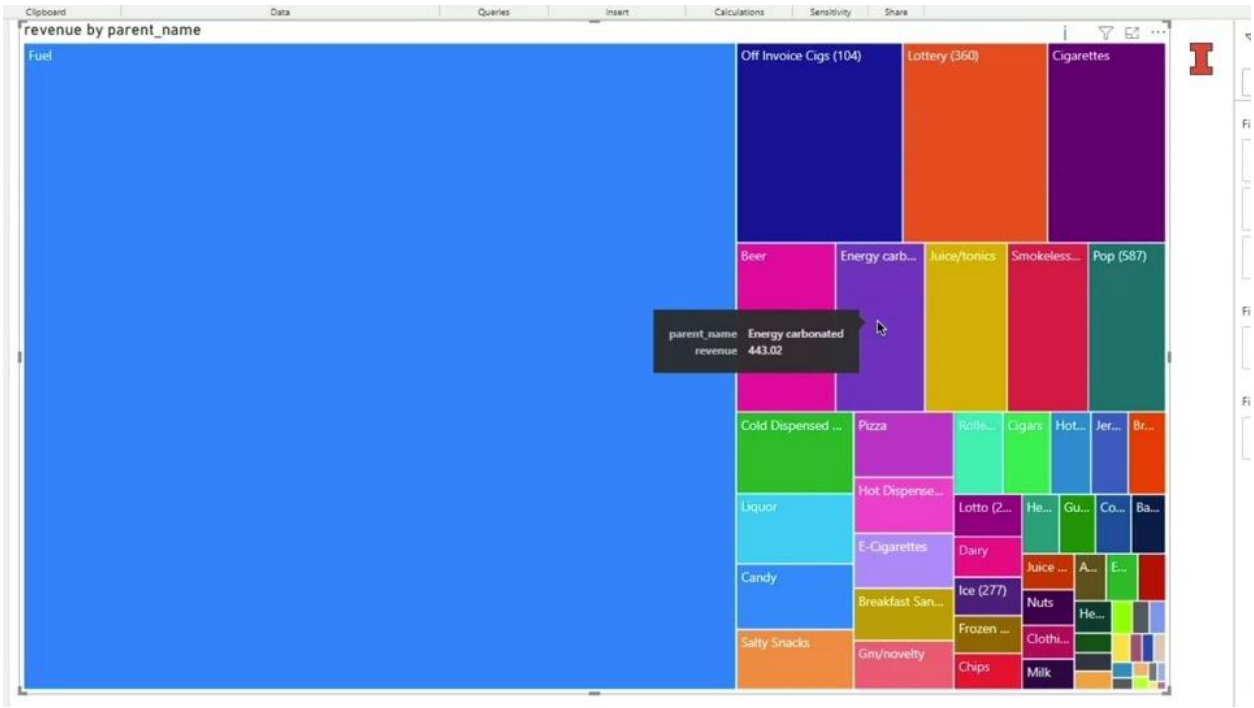
For exploratory data analysis, it may be useful to rotate between different types of charts. Bar charts are helpful to see relative amounts, a pie chart is often useful to show percentages. This pie chart is very messy. Pie charts aren't usually intended to show more than one different type of value. Let's narrow it down to just revenue. We'll get rid of these other ones here.



This is helpful. This allows us to see that fuel makes up 63.14 percent of the revenue or $13.530. Pie charts are very helpful, my rule of thumb is six or fewer slices or categories in a pie chart can be useful, especially when you're looking at the percent of the total. The downside with pie charts is that they don't display lots of pie slices very well. Another downside is that it's typically not ideal to have the legend separate from the slice of the pie. We would like real close so your eyes don't have to keep moving back and forth.

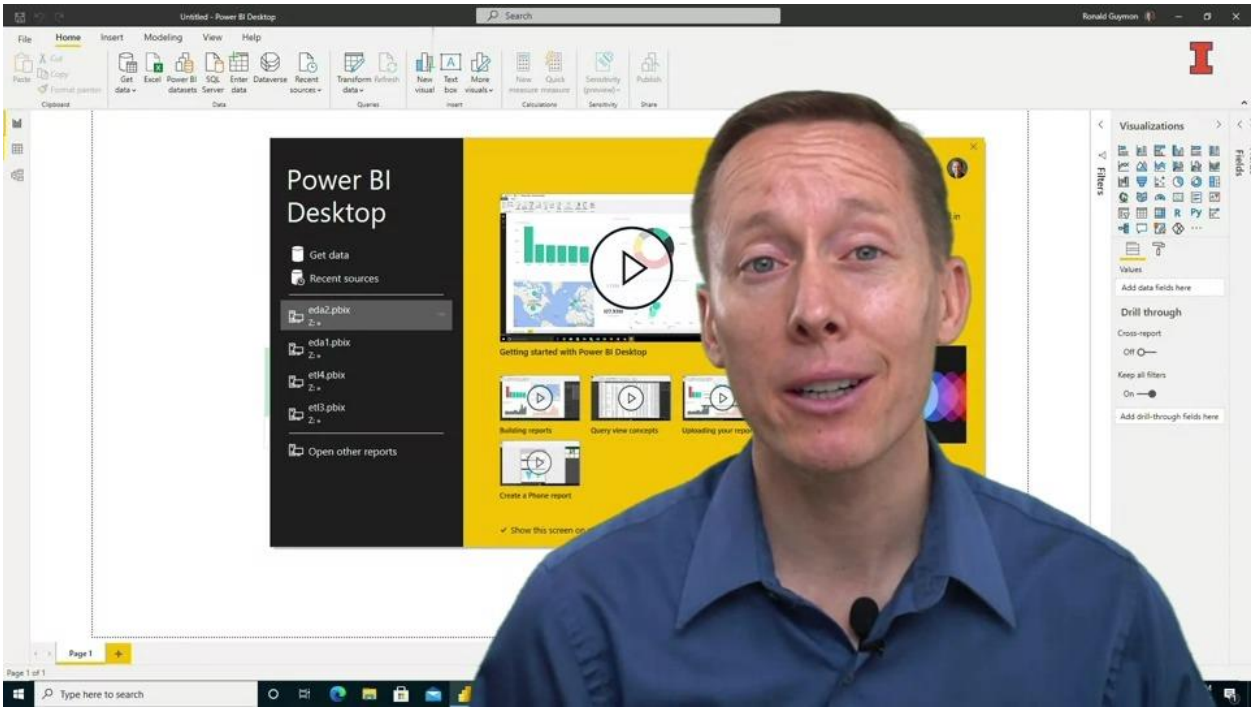


Another type of chart that you may be aware of is called a tree map. A tree map is like a pie chart, but in rectangular form. The thing I like about tree maps is that often the pieces of the tree map have the label on them. They are also, I think, much more helpful for displaying categorical data that have more than six different values in them.

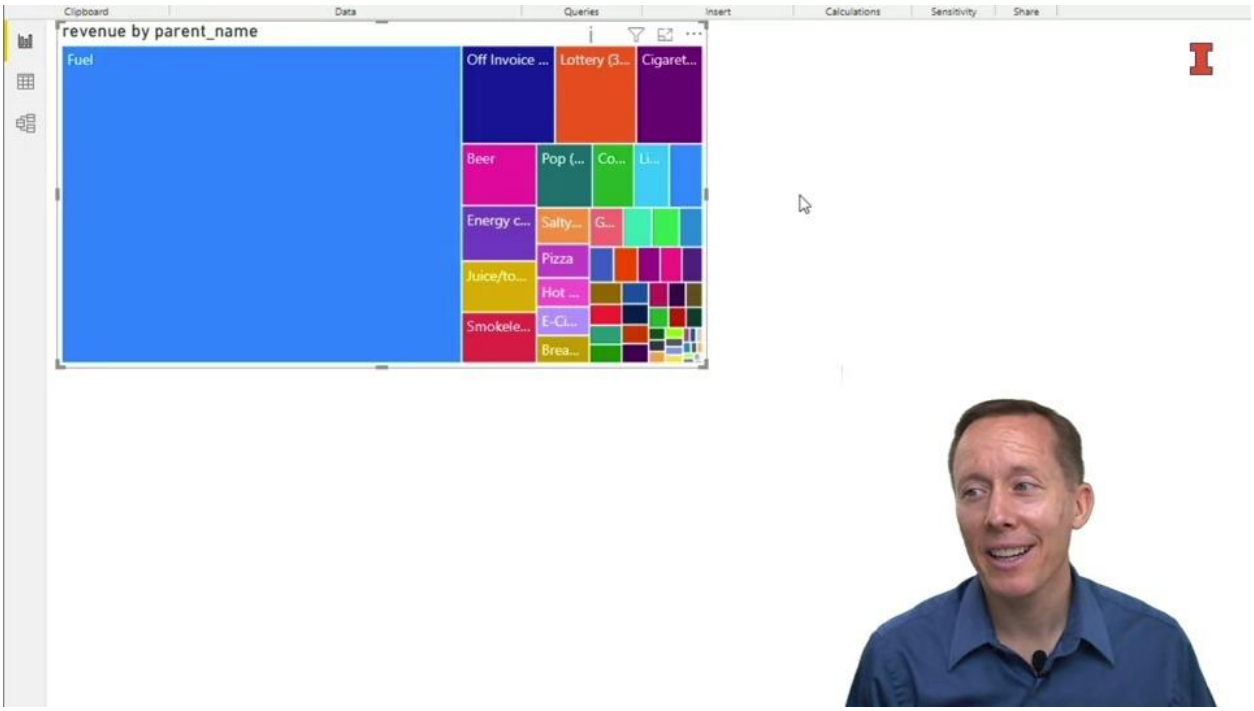


If we wanted to look at maybe instead of super parent name, we want to look at parent name, which has 60 different unique values in there. You can see how quick it is to update a chart and we can make this bigger. Now we can see that fuel makes up about the majority, but it's easy to see some of these other parent names that are populating the total revenue. That's how you can do some exploratory data analysis with categorical data in Power BI.

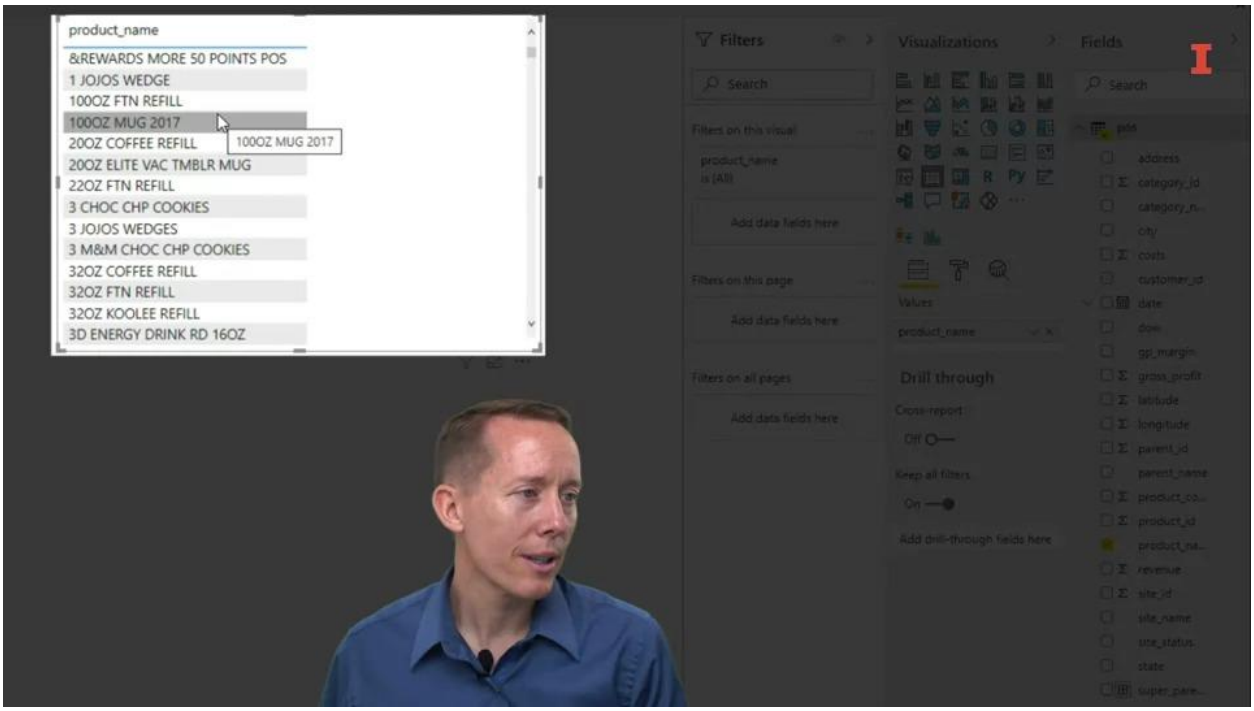
Lesson 2-3.3 EDA 3: Filters, Slicers, and Drill Through



In this video, we want to show you some tools for implementing Schneiderman's mantra, which is overview first, Zoom filter, then details on demand. Specifically, we'll show you how easy it is to use charts and tables as filters and also how to use the filter pane slicers and finally, how to use the drill-down features, in Power Bi for hierarchical data.

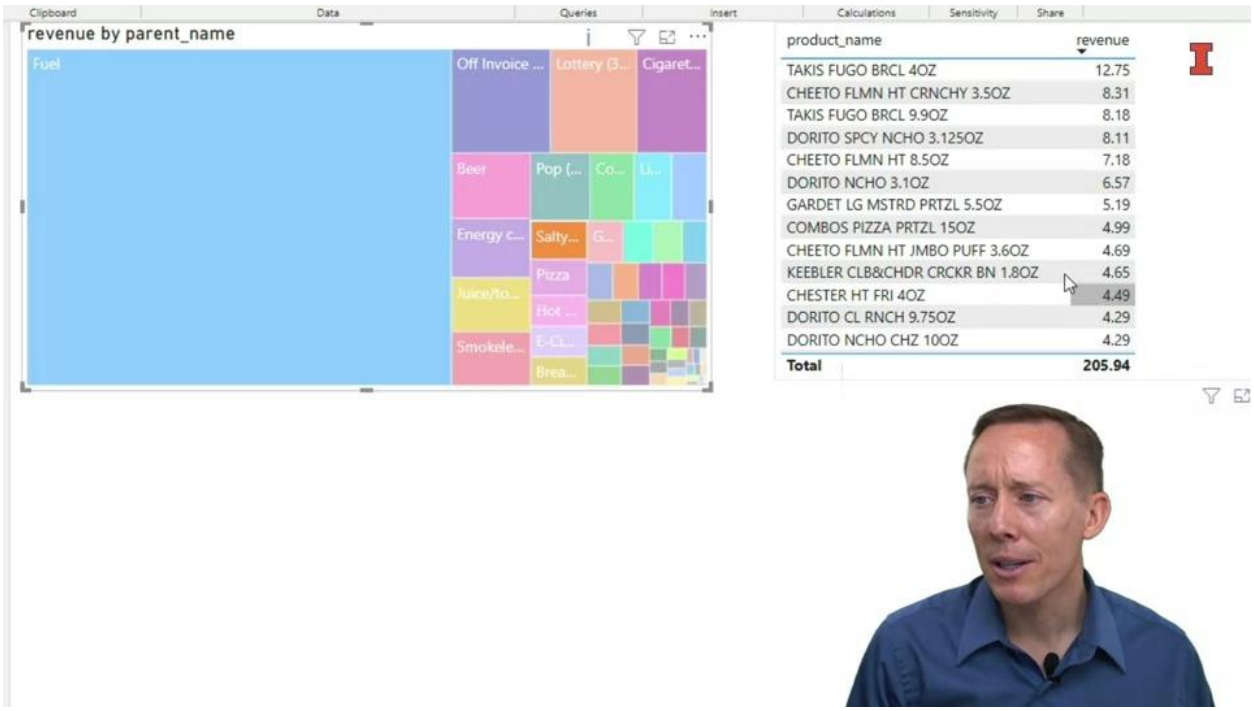


The first thing I will do to start with is load this project I've been working on EDA2, and I will go ahead and click on this and able script visuals. You can see in this report, I've got a univariate tab and a categorical tab and the univariate tab has histograms and a box and whisker plot, the categorical tab right now just has a tree map. First of all, I'm going to resize this tree map, and then I will add in a table,

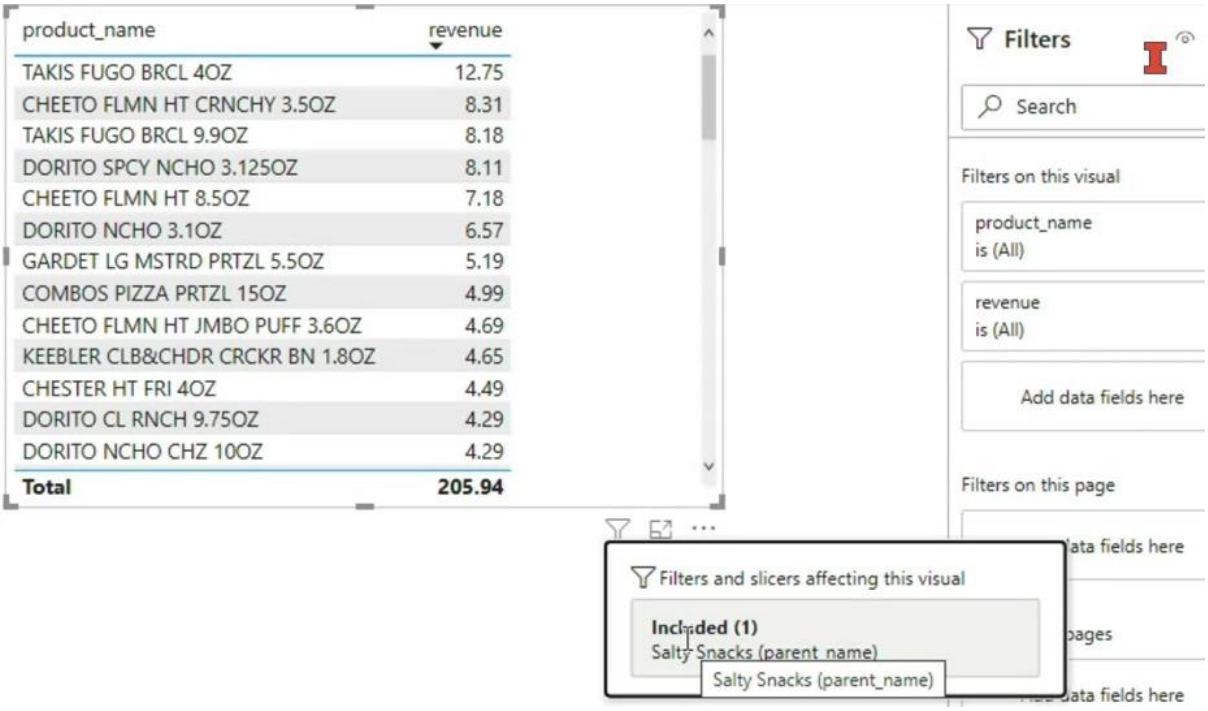


so I'm going to just click on this table icon in the visualizations paint, and I'm going to resize this and move it up next to the tree map and in this table, values filled, I'm going to add in product name. This gives me a list of all the different products that are sold or that are in the data center,

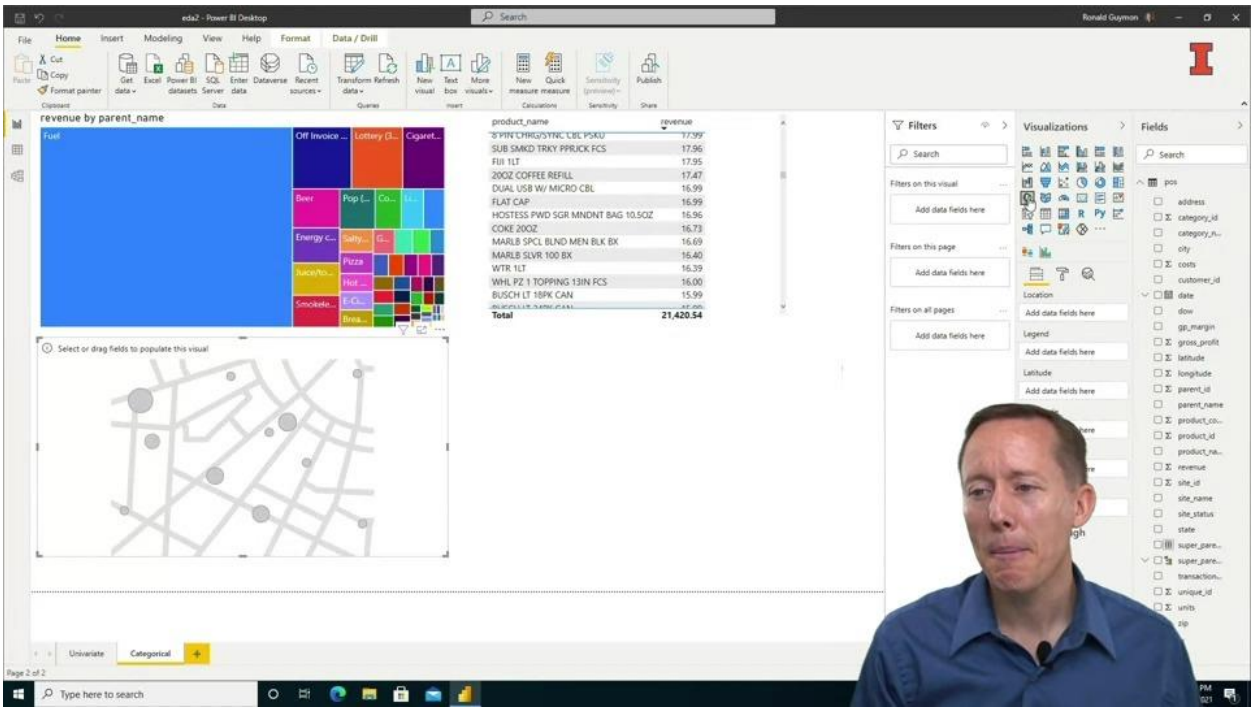
and I want to see what the total revenue is for each product, so I will also drag revenue, to the value's filled. You can see maybe it doesn't appear to be summarized, but it is, if you click on the download extra revenue, you can see that it's summing up the total amount of revenue for each of these product names, and it might make more sense to sort this in terms of, the highest revenue items to the lowest revenue items. The first thing I want to show you is that inherent in power BI plots is the ability to filter.



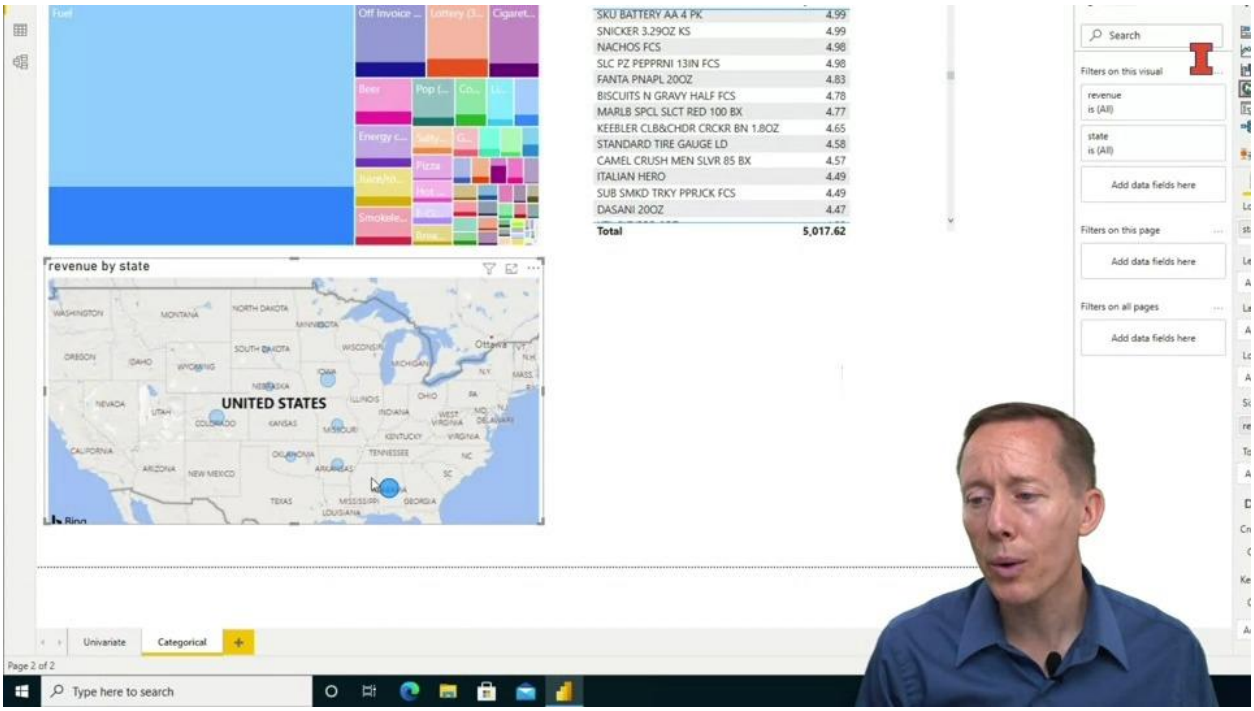
If I click on one square in this tree map, let's say salty snacks. Notice how this table updates and includes the product names only for the salty snacks, and I can see the total is 2,005 dollars 94 cents in salty snacks. If I'm ever curious as to what is filtering this table,



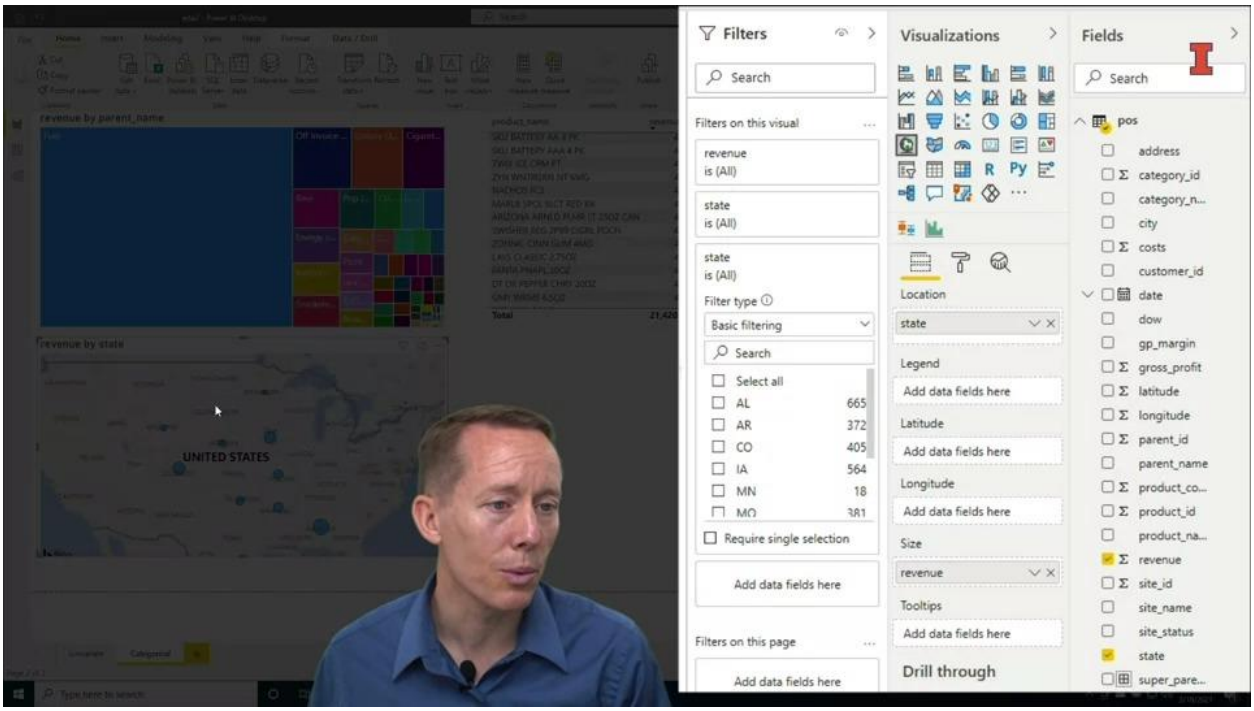
I can hover on this, click on the table and then hover over it and I see the funnel button, and it will tell me what filters are currently applied. I'll go ahead and click salty snacks again and click on the funnel and I can see that salty snacks is being used to filter this table.



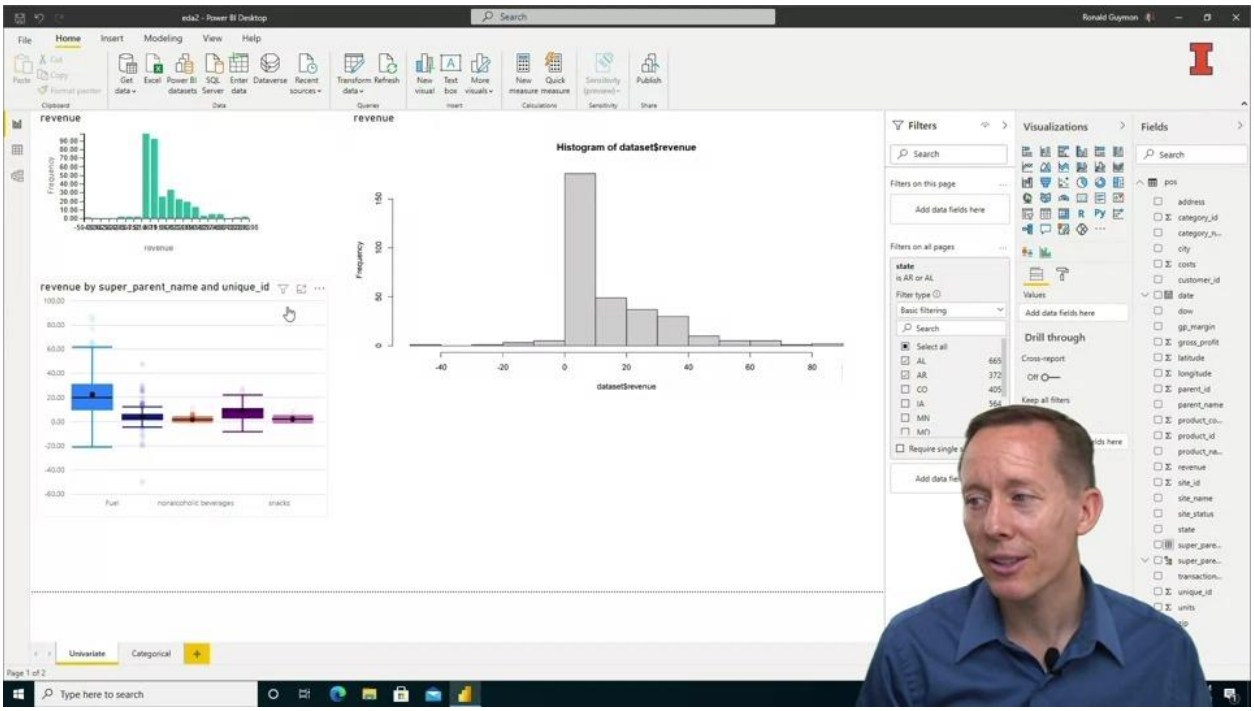
Now, I can also use the table to filter the tree map as well. I'll turn off the filter on the tree map by just clicking on the square that had been clicked on. Let's go ahead, and I will click on, coffee refill and this is a very small segment of the hot drinks item there, so it highlights it for me and I actually like it. Sometimes it's hard to find if it's a small item, but I like that it leaves the context of everything else there rather than just show me one rectangle. I really like how it leaves the context in there, so anyway, filtering is really very easy and powered BI, but there are some other ways to filter that. I'd like to illustrate for you. To do that, let's go ahead and turn off this filter and let's add in a new chart and let's add in a map.



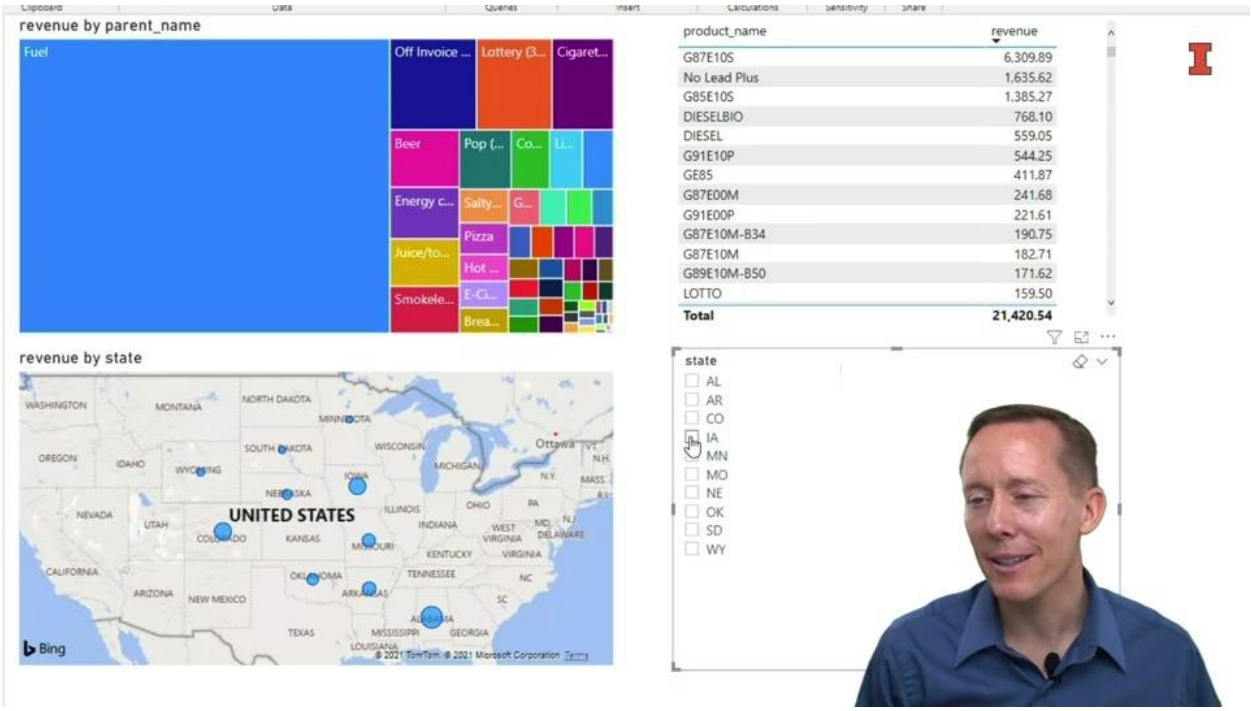
Now, power bi is really good at recognizing locations by name. You can also add in latitude and longitude coordinates, but let's go ahead and just use state, to add into the location box for this map and see what happens. Beautiful. You can see that we've got a dot on each of these states, so it recognizes the states and where it should go on a map. Now, I can resize these dots. Let's say I want to see which state sells the most. I can drag revenue to the size field. Now, I can very easily see that it looks like Alabama, [inaudible] Colorado has the most sales, and now I could use this as a filter for the other plots as well, if I want to see the total sales just for Alabama, also I click on that, the tree map and the table update to show me the sales just in Alabama. Again, I really like how the tree map includes the context of everything else. I can see how much of the total sells fuel cells Alabama makes up out of all the sales and all these other apparent names as well. Now, what if I want to filter the visualizations, not just on this tab, but also on the univariate tab as well?



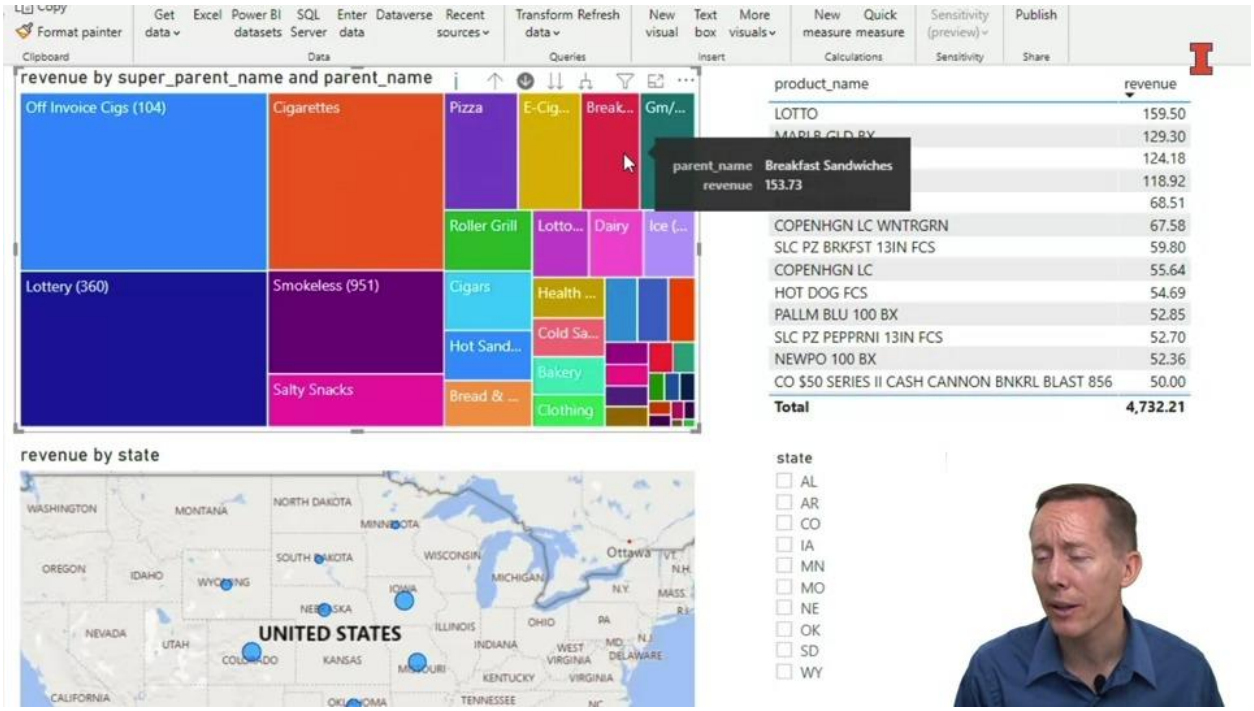
One way I can do that, is by using this filters bar over here and if I drag state into this first box here, you'll see this will allow me to filter only on a particular visual, which is the map here.



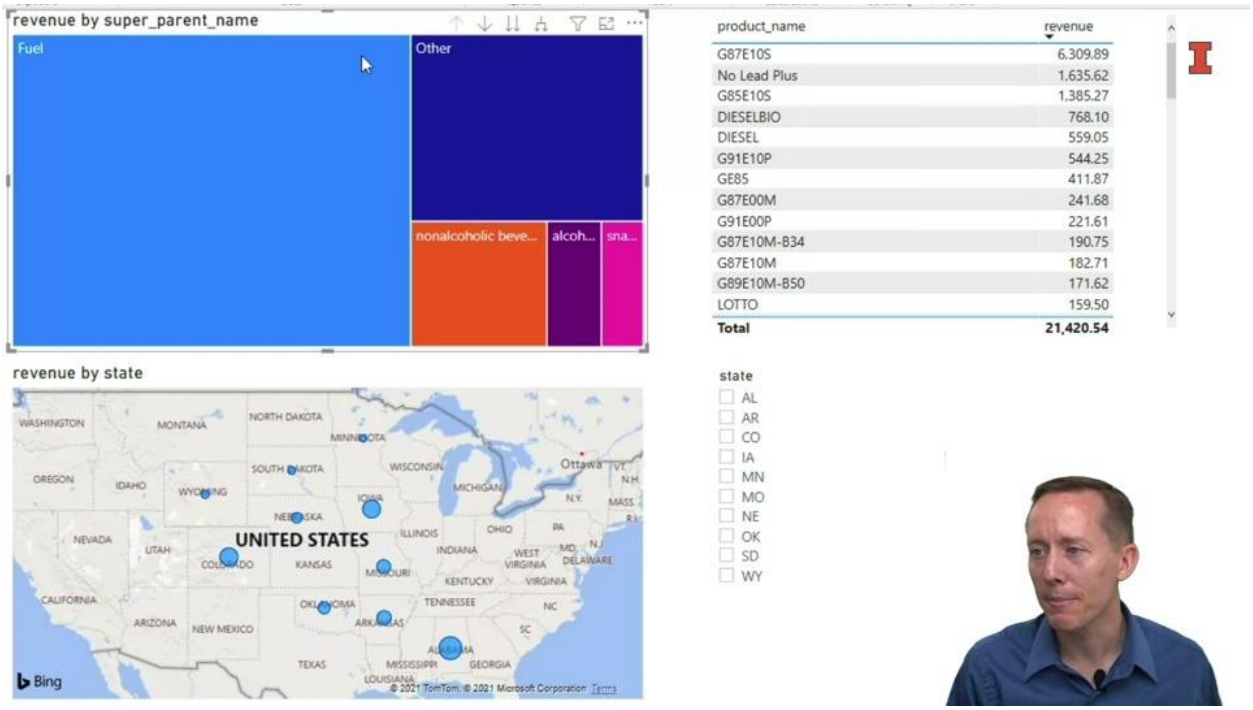
I can just click on Ar for Arkansas and you can see it zooms in on Arkansas and filters it just to that. I'll remove that filter and then drag state to filters on this page and I can just click on Arkansas again, now that not only filtered the map, but the tree map and the table as well. Finally, I could. You state to filter the observations on all the pages. If I click on Arkansas here and go over to the Univariate tab, you probably can't tell but these have been filtered. Let me click on another state here AL for Alabama. Notice how these charts also are updated. That's another way that you can explore the data using filters.



Now, another way to filter the data is by using slicers. If you've ever used Excel's pivot tables and added slicers to those, it's very much the same way. I'll go ahead and remove this filter from the filter bar and I'll go back to the Categorical tab, make sure I haven't selected any of the visualizations and I'll click on this slicer icon here. Now, I can drag state into the field and now I can essentially have that what was in the Filter bar on the right here within the chart itself. This is helpful if we want to publish a report for others to use and that filter bar for me is more like a design mode, whereas this is kind of production where this will allow other users who are not able to change the layout of this report to filter the data. Anyway, we can filter in those three different ways, using the visualizations and tables, using the filter bar, and using slicers. The last thing I want to show you in this lesson is how you can use the drill-down functionality and Power BI. This is really exciting. I think this is super cool. To illustrate this, I'm going to play with the treemap

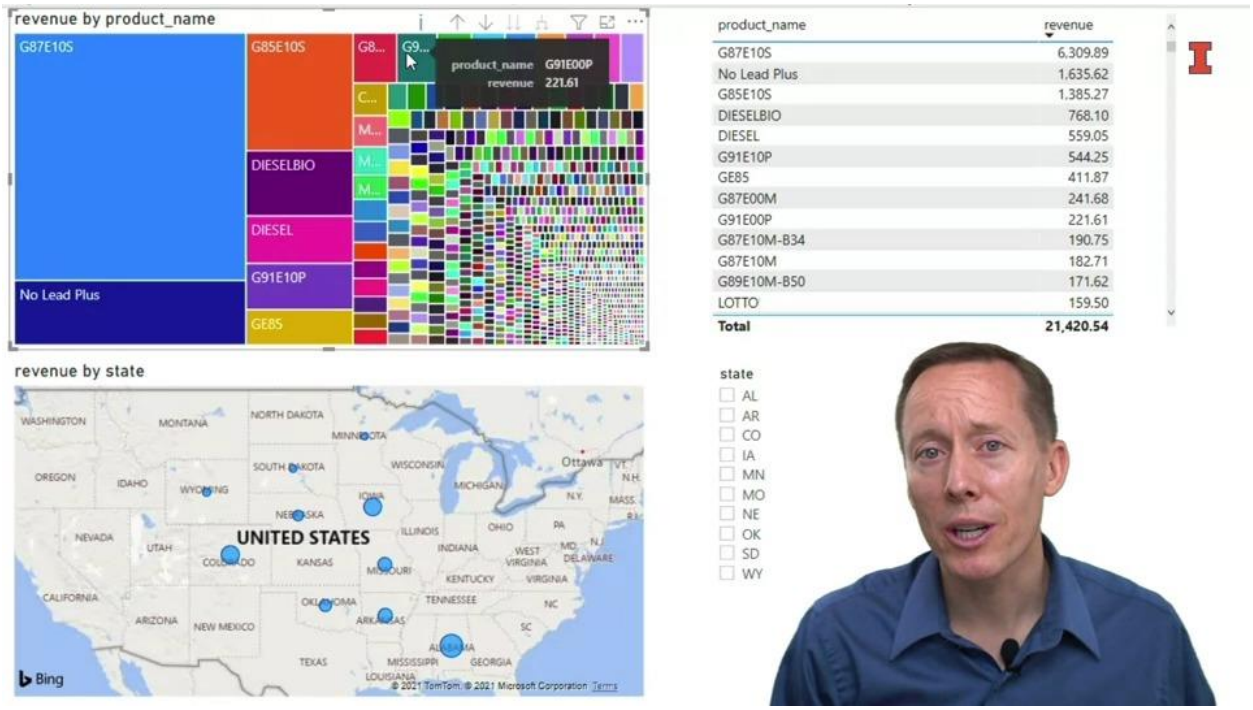


and I'm going to remove the parent name and replace it with this super parent name hierarchy. Now, if you remember the super parent name hierarchy has not just superpower name, but then below that, the parent name, then the category name, and then the product name. This is a hierarchical relationship in the way Teka organizes the categories of products that it sells. Now that I've done that, notice that in this treemap I've got some additional icons up here. Let me turn off the filters. If I hover over these, this allows me to click, to turn on, to "Drill down". Let's start with this. If I click on that now, if I click on one of these super parent names, let's try Other. I know it has a lot. Now, you can see that it's showing me what makes up the total amount in the Other category. I can see the largest amount is off-invoice cigs and then lottery, cigarette smokeless, so on and so forth. Now, I can drill down even further. Let's say I want to go to breakfast sandwiches and when I click on that also note what happens to the dots on the map.



All right, so those are changing as well and we can see the amount of breakfast food that is now purchased in each of the states, relatively speaking, as well as in this table over here. Now, let's click on breakfast food one more time to see what makes up breakfast food. We can see the different breakfast food items and they're laid out in this table very well. Now that we're in this lowest level, I may want to use this as a filter, not as a drill down. I can turn off that drill down mode and now I can click on, for instance, this pretzel bun and now everything will filter to show me just the pretzel buttons. I can see those are sold mostly in Iowa and Alabama, total of $18.05, a little bit in South Dakota as well. Then I can undrill if I click on this up arrow button and go back up to the top layer or I guess that's called drilling up, not on drilling. I can turn off that drill mode and turn off all the filters.

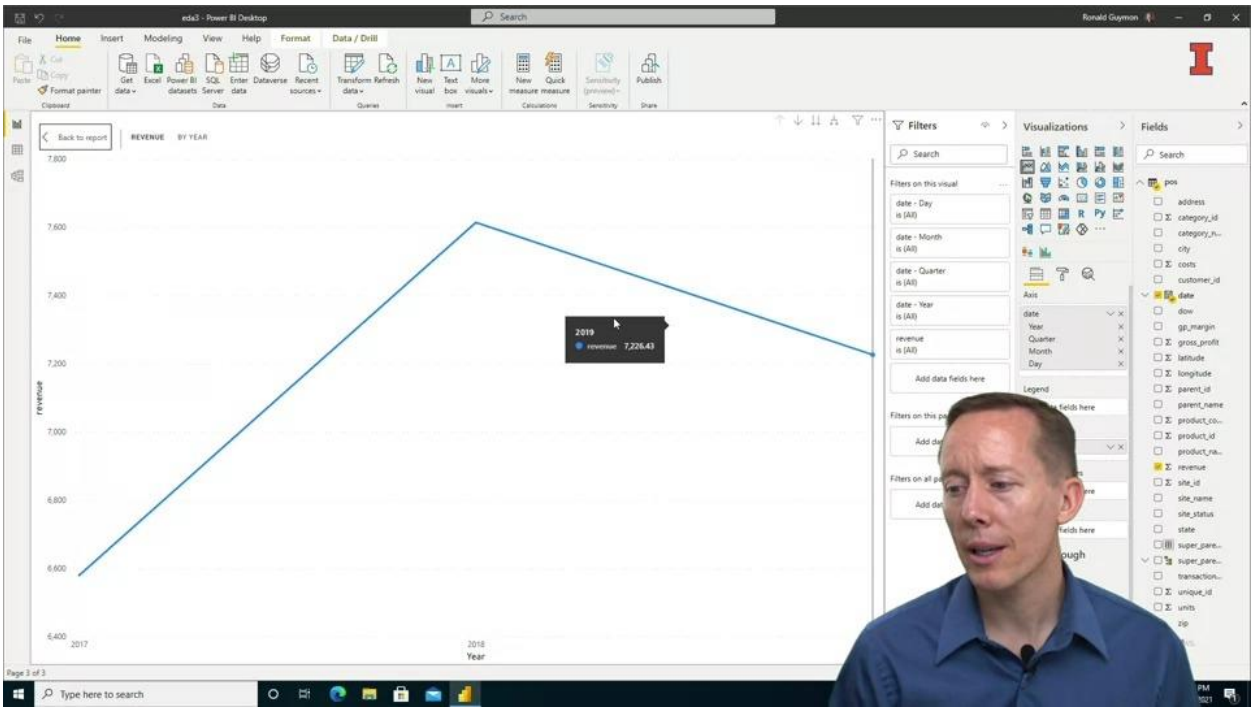
Now, another way that you can drill down to other levels is by clicking on this double arrows, and watch what happens here if I click on that. Watch what happens to fuel and other. They then now are broken down into the various components still on the same treemap. So fuel, its product name or the parent name category is just fuel. But these other super parent names are made up of more than one parent name, so those got broken down into more categories. I can keep going down to the next level and you can see that now this is getting overwhelming in here. I've got all these different product names basically. Now I've got all the product names that I'm at and it's just overwhelming. Now, if you do have something like this and you really want to explore it more, you can click on this focus mode here and that will make it so that it fills up the whole screen and allows you to see a little bit more detail. This is still way too much, though, to really learn much from other than maybe these ones that make up the majority. Then I can click on "Back to report" to go back to the report.



These are demonstration of how to use the drill-down feature and Power BI. I think it's super awesome. In conclusion, I'll leave you with the challenge, see if you can replicate this state's map so that you can drill down from state, down to zip code and then down to city.

Lesson 2-3.4 EDA 4: Multivariate Plots: Scatter Plots and Line Plots

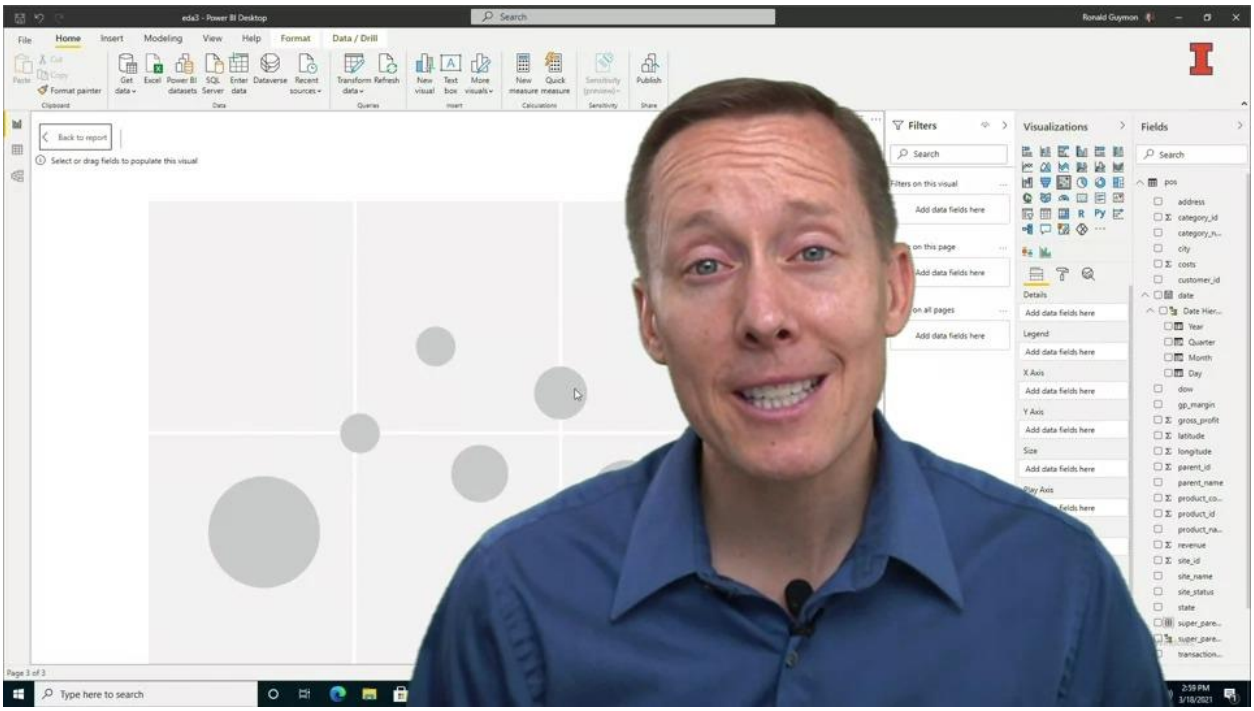
In this lesson, we are going to focus on multivariate plots, in contrast to univariate plots to look at one column of data and bivariate plots to look at two columns of data. Multivariate plots look at three or more columns of data. And these types of plots can include a mixture of categorical and numeric data. Now the focus will be on scatter plots and line plots.



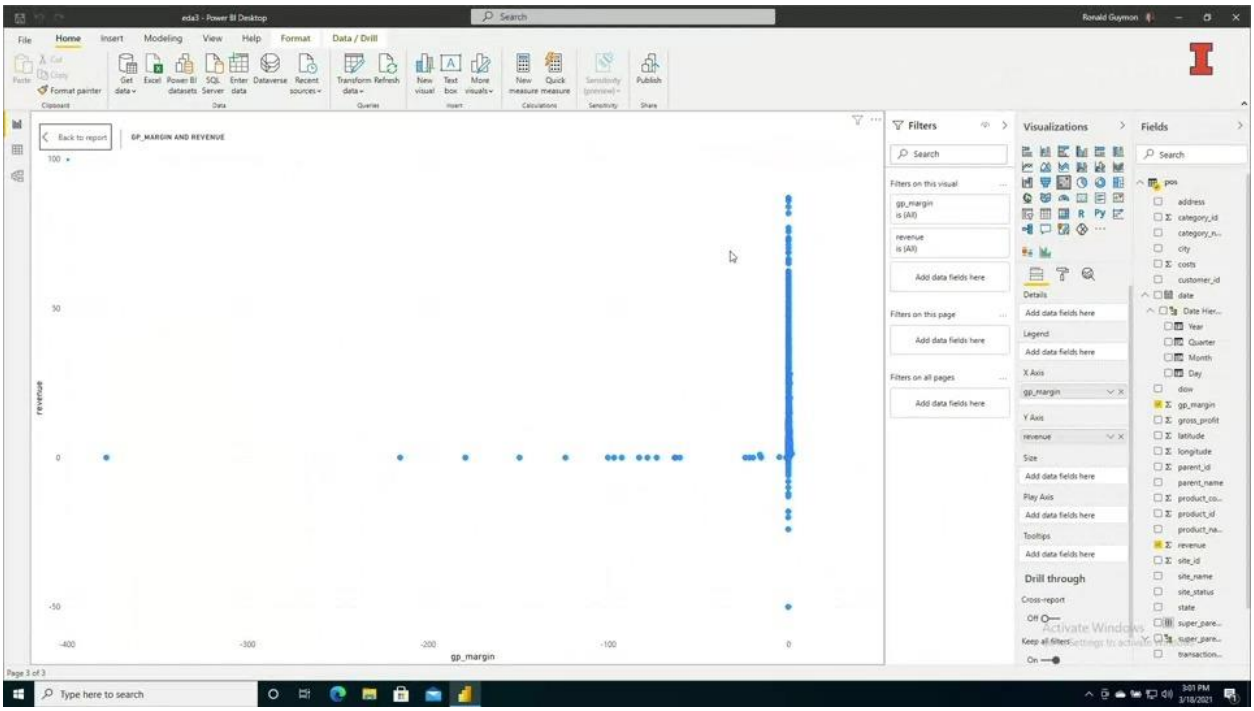
So first let's open up our Power BI and I'm going to open up this project I've been working on ETA 3. And on this report you can see that we have a tab for univariate plots and then a tab for categorical plots, I'm going to create a new tab for multivariate plots. And in this tab I am going to create a line chart. Now line charts are really helpful for connecting points that are related to each other in terms of time. So I'm going to go ahead and drag onto the axis here, the date and notice how it includes the hierarchies of the date, we'll come back to that. And then in the value section I'm going to drag revenue and I'm going to focus on this plot, I'll blow it up so we can see it better. Now, notice what it does is it's taking the date at the highest level of the hierarchy year and adding up the values so that we've got a total revenue for each year. And if I look at revenue, I can see that it's summing it up there. So that's pretty awesome, how quickly you can create a line plot in Power BI. What if I want to look at total revenue in terms of quarter?

Well, there's a couple of ways I can do that, I can go up here and since there's a hierarchical relationship, I can just drill down to the next level in the hierarchy. And you can see that it's grouping it by a quarter, now, it's not by quarter and years, just by quarter. So all quarter one overall three years has being grouped together and I can continue to drill down to see month and then to see day, which is really day of the month. All right, I'll go back up to the top, I'll drill up about two year again. And if I want to, I could expand all down one level in the hierarchy and when I do this, it will break it up into quarter and year and then month by year and finally day and year. So that's using the drill down feature, but you may not want the drill down feature.

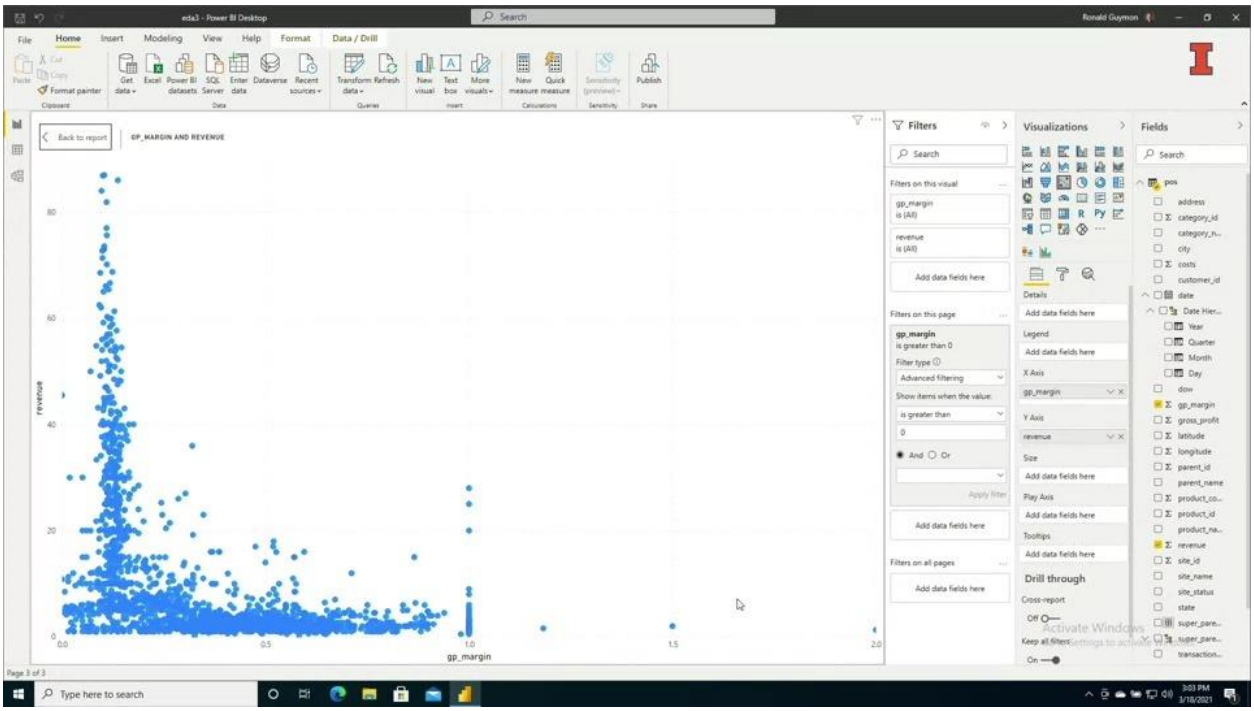
So another thing that you can do is just go over to the date here and select the down arrow and say, hey I don't want the date hierarchy, I just want the date. And if I do that, now you can see that we've got an individual data point for each day in this data set. All right, so this is two columns of data, date and revenue. I can add a third column of data by adding in year. So if I go to the date hierarchy and drag year to the legend. Now I can see all the daily sales and it's colored by date. So that's allowing me to investigate three different columns of data at once and this may not be as useful here. All depends on what you're trying to do, again if I want to maybe see year over year monthly sales, that may be more useful. I could go back over to the date and go to date hierarchy and I can drill down and say I don't want it a year or a quarter, but I wanted that month and now I've got a separate line for each year and a dot for each month. So pretty awesome how we can create a multivariate line plot in Power BI, so quickly and intuitively.



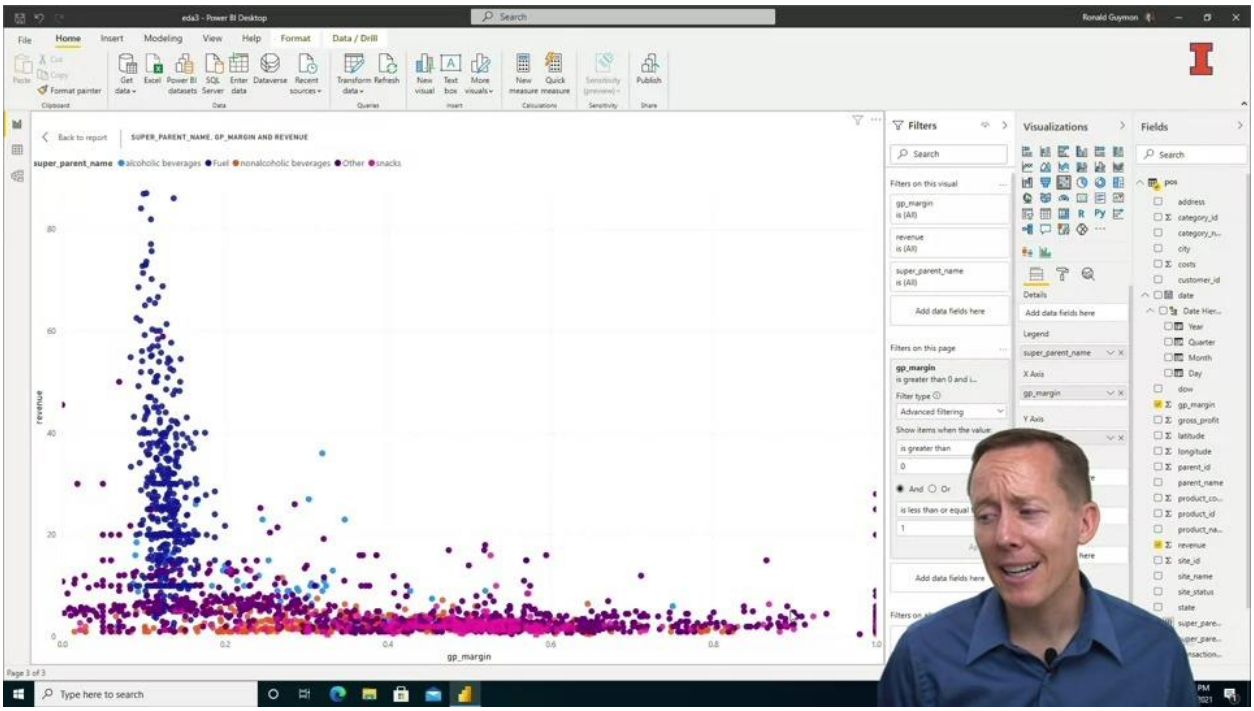
All right, let's go back to the report here and create a scatter plot. So I will go ahead and click on the scatter chart icon and then I will focus on this. And let's look at the relationship between our gross profit margin and revenue, and the rationale for this is we might want to investigate if units or line items that we sell that have a high gross profit margin also have a high revenue. That would be ideal, right? If our highest gross profit margin, if we have 99% gross profit margin also brings in a ton of revenue. So let's go ahead and investigate that.



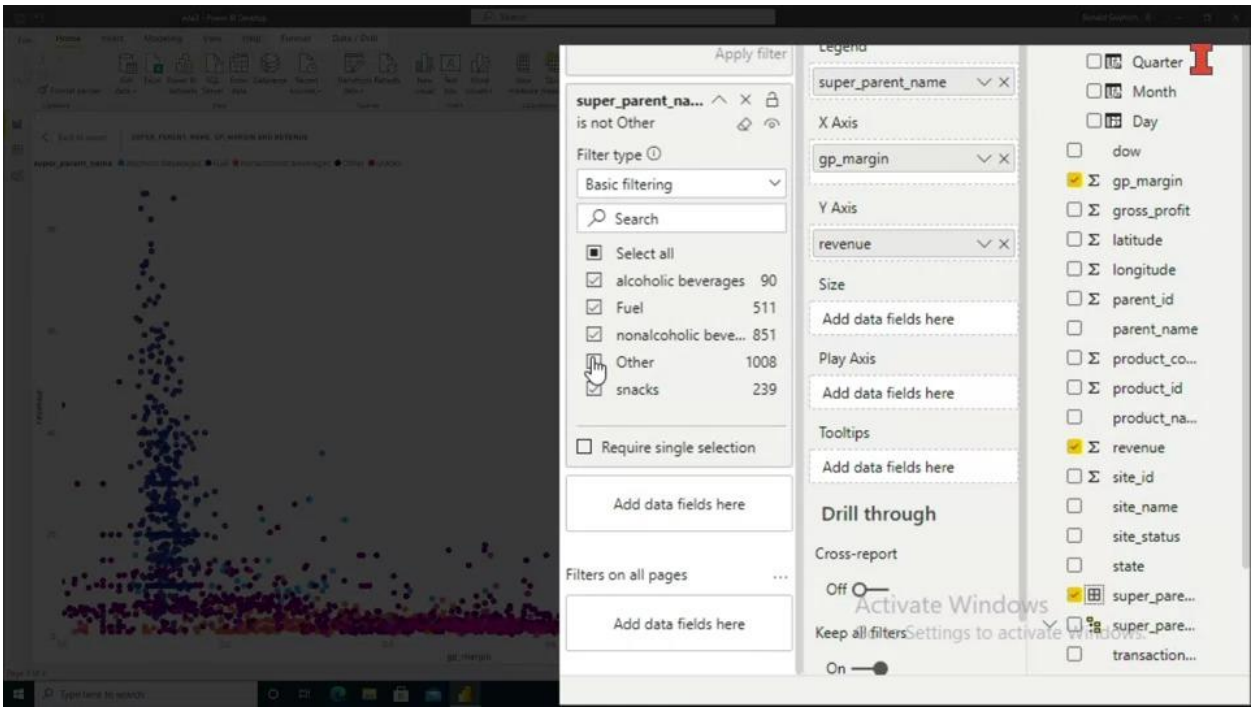
So we will move gp margin to the x axis and on the y axis we will put revenue. Now at this point there's not a whole lot of information here that's because there's some aggregation going on. So we need to click on the down arrows and let's say don't summarize and it's saying it's not playing well, well why is that the case? So if we look at gross profit margin, gp margin, we can see that it is a text data type right now. So we need to go in and change that. So we'll go to the home tab, go to transform data. This brings up the power query editor and we'll scroll over to gross profit margin and we'll click on this data type icon and change it to a decimal number. All right, now we can see that there are 93% valid entries and 7% that are empty, perfect. So this is what we want to go ahead and close and apply this. All right, now we've got a scatter plot that is working.



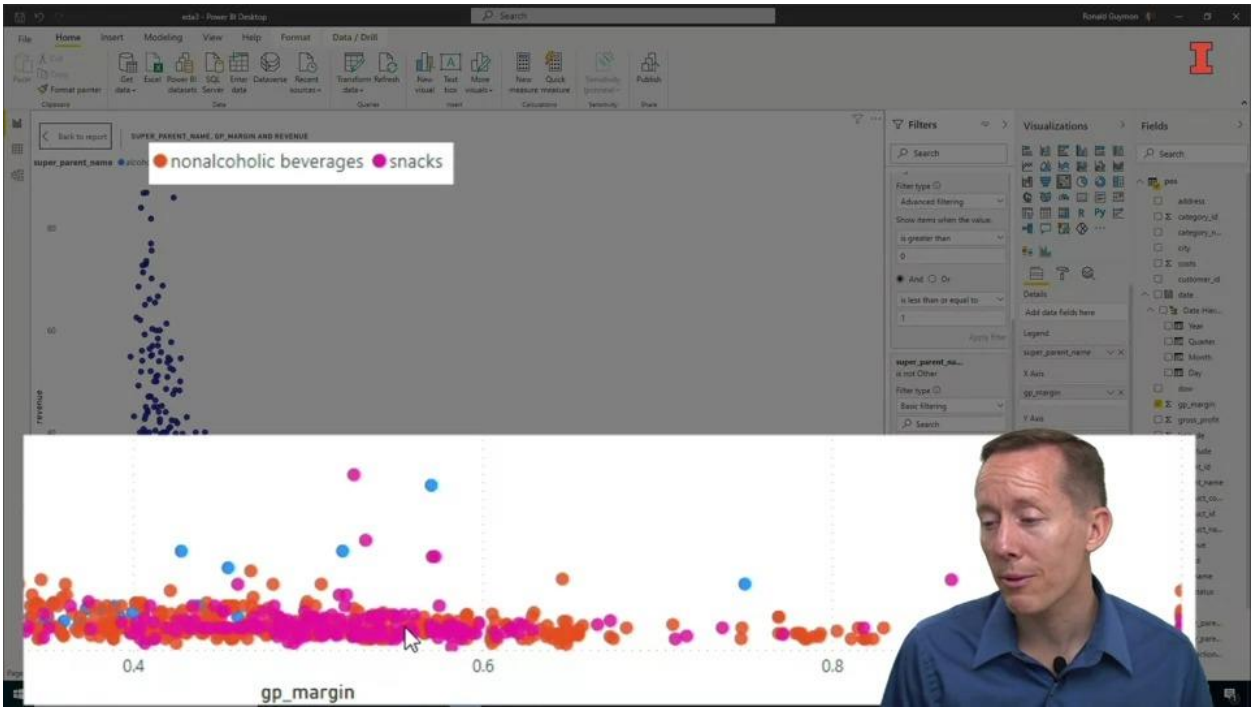
And let's go ahead and make sure that gross profit margin is not being aggregated and it's not so perfect. But this isn't real informative and so let's try to explore what's going on here. As we look at this, we can see that on the x axis there are definitely some outliers on the low end, so negative 400 up to zero. And we may want to remove those outliers, I suspect that most of the observations will be between zero and one. And so we could go back into the power of query editor and remove outliers that way but there's often a better way to do that in Power BI and that's by using filters or slicers. Let's go ahead and just use the filter here. And let's drag gross profit margin onto the filters on this page, and notice that we can use filters with numeric data as well. So we want to say we are using gross profit margin and we want to say it's not less than but greater than zero. And let's go ahead and apply that filter. All right, so this is looking much better. We're not being overwhelmed by the extreme negative outliers, but we do have some positive outliers that we may not want to include either.



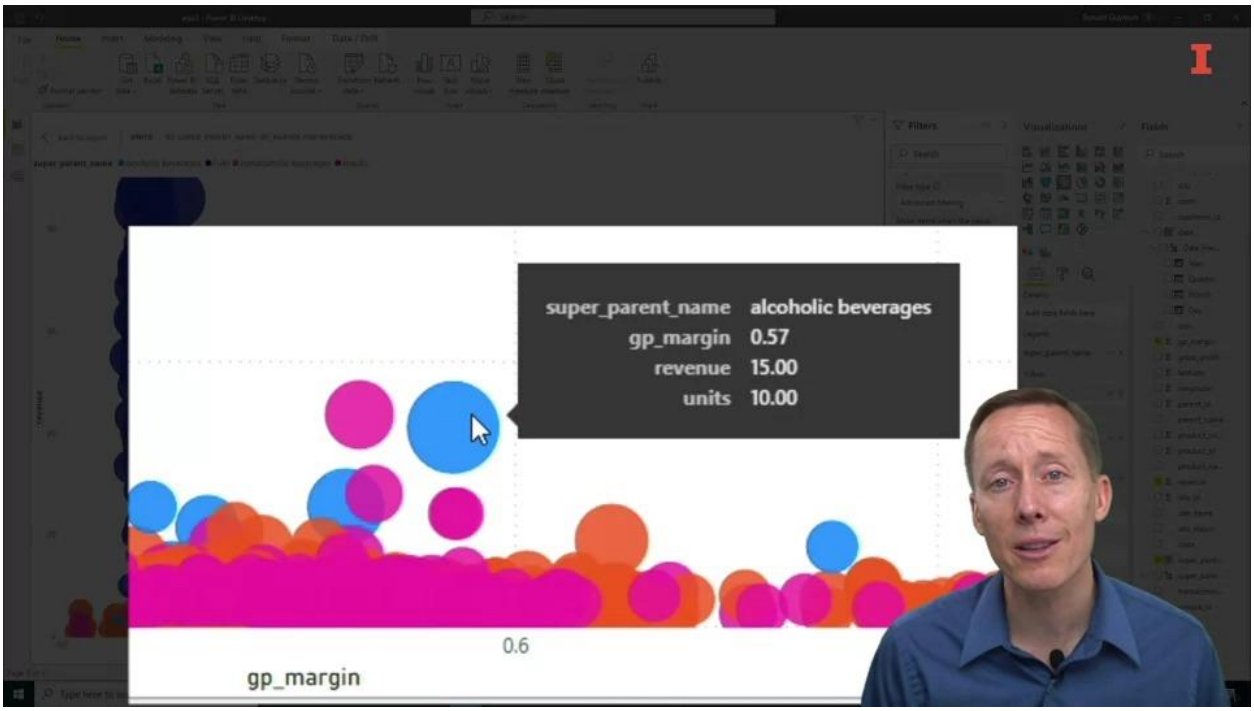
So we can go ahead and go back to our filter and say we also want to filter observations such that gross profit margin is less than 1 and actually might say less than or equal to 1, then apply the filter. Now this lets us see a lot more of the nuance in the data. And at this point we're looking at two different columns of data, gross profit margin and revenue. And the next thing we might want to ask ourselves is does this relationship, which it's not real strong. It looks like we've got a bunch of observations that are kind of low profit margin, but high revenue. And then a bunch of observations that are high profit margin, low revenue. Well, we might want to ask ourselves, is that due to a particular type of product? So let's go ahead and color these data points using the super parent name. So drag super parent into the legend box. All right, now this is pretty cool, we can see that all these points here that are blue are the fuel points and so those are the ones that have the low gross profit margin but high revenue typically. Then we've got some over here that are this dark purple and these are the other category here. So this, if you may recall, this is a bunch of different types of products that are in that other category. Now this may be kind of overwhelming, we're not interested in seeing the other in there.



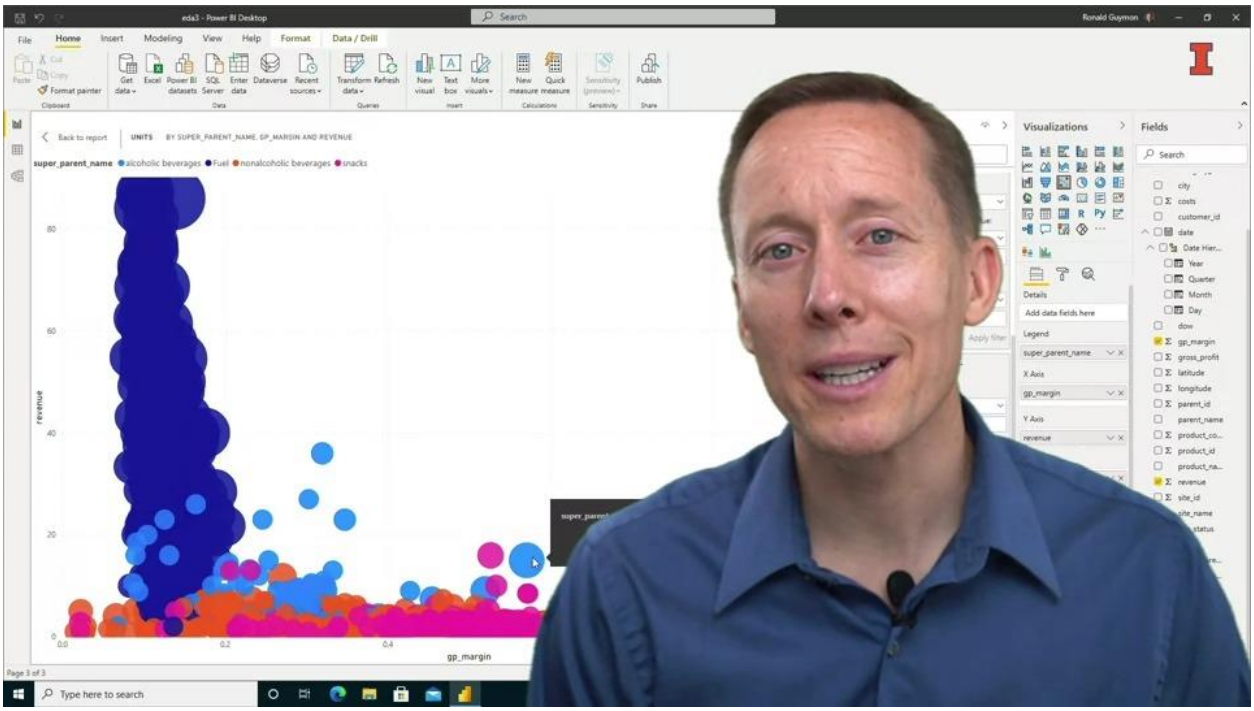
So we could filter out these other observations by also dragging super parent name into this filter And now we have a new filter box and we'll start with selecting all and then we will just uncheck the other.



So now we're not being overwhelmed by all the other observations and it's a little easier to see the gross profit margin and revenue relationship between our four main categories. And we can see that alcoholic beverages, these light blue points often have a higher revenue and a higher gross profit margin, whereas the orange non alcoholic beverages and pink snacks often have a really high gross profit margin, but pretty low revenue.



Now you might ask yourselves as you think about the data. Well, maybe some of these observations have high revenue because there are a lot of units that people are buying at once. Well, we can actually investigate that and include that in here as a fourth variable. So let's go ahead and drag units into this size box here. All right, now that's kind of overwhelming, it doesn't show a whole lot here. That's because remember fuel, the units are the gallons purchase and that's often in the teens or 20's, whereas these snacks are often in the 1's or 2's. But we can see here that this observation for an alcoholic beverage has a really high gross profit margin and high revenue is because they purchase ten units at once.

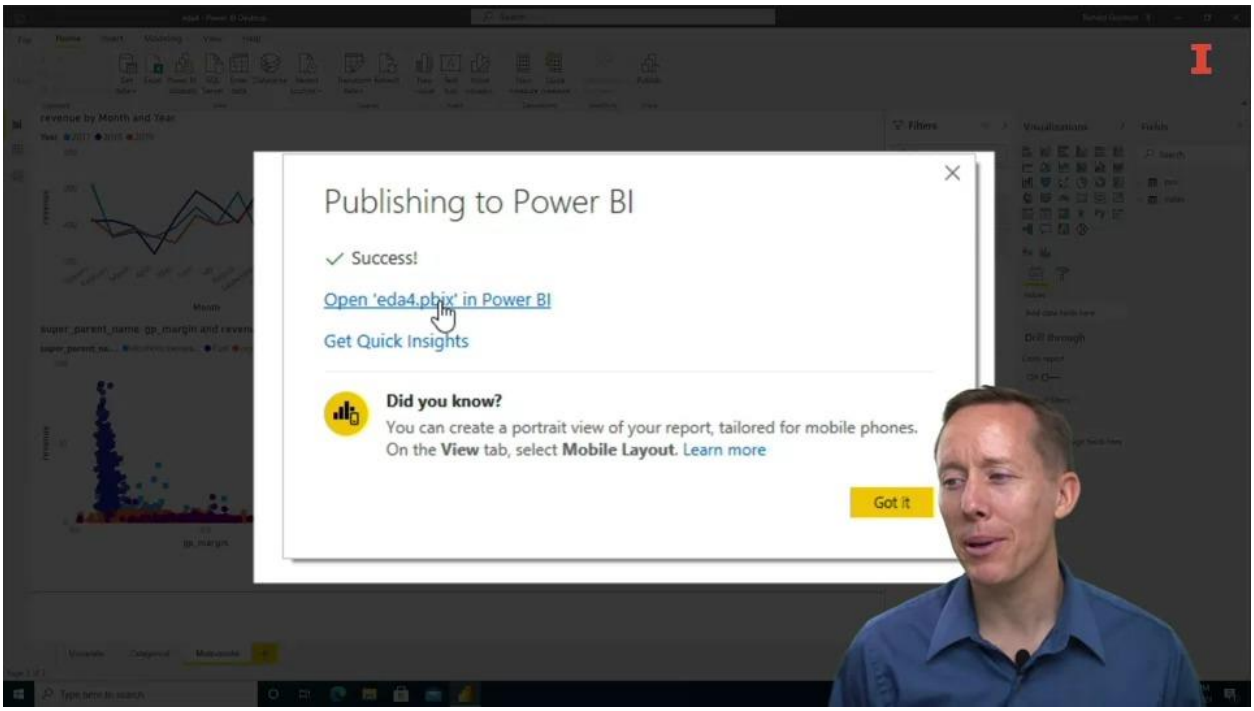


So that helps kind of explain things as well. So real quickly, you can see how we can create a multivariate scatter plot in a multivariate line chart and it's easy to include three or four different columns of data at once. Now, word of warning is to be careful in using too many multivariate plots on a report like this.

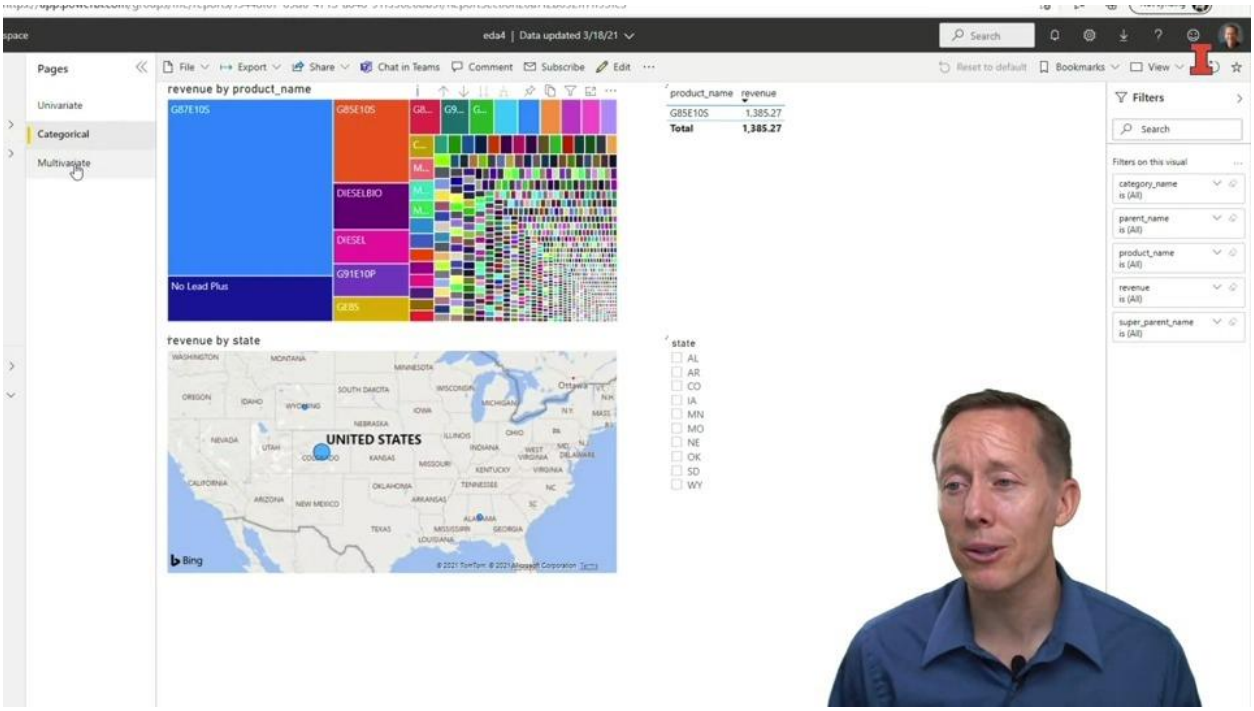
When you do that and you have all these plots that have lots of different variables on there it can get overwhelming very quickly. So I would recommend to maybe just put one multivariate plot on a tab in a report, maybe two, but just consider how much information processing will need to occur by the people interpreting this. And if your purpose is to just explore the data, maybe it does make sense to put all these plots on there. But on the other hand, you could replace one multivariate plot with several bivariate and univariate plots and then filter through and look at the relationships that way.

Lesson 2-3.5 EDA 5: Publishing a Power BI Report

In this video, we want to show you how you can really take advantage of the Microsoft environment by publishing your Power BI report online so that everyone in your organization can access it.

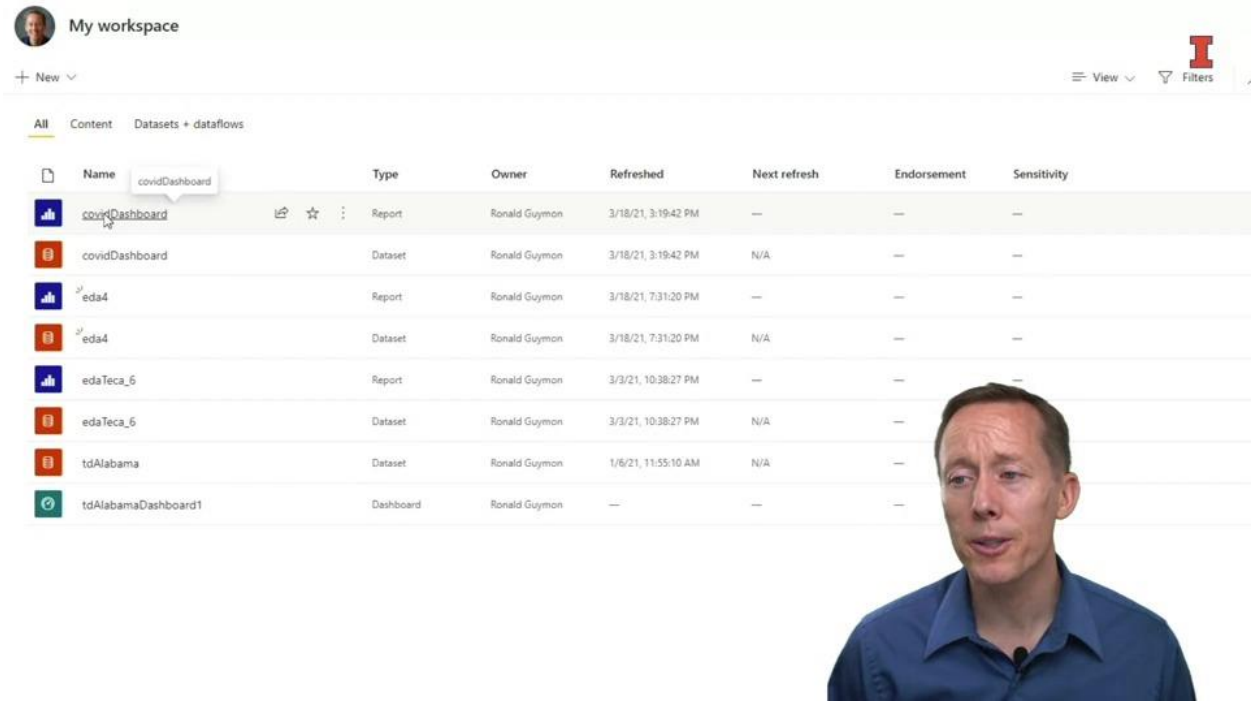


First of all, we'll go ahead and open Power BI and open a report that we've been working on. All right, now that we have this report open, we just make sure that we go to the home tab and then over to the public icon, we'll save any changes that had not been saved, and then if you have not logged in you'll be asked to log in, make sure you're logged in to your organization account, and then you can choose if you want to publish this to just your workspace to different work-spaces. I'll just publish it to my workspace and click select. All right, it has published it, let's go see what that means. We will open this in Power BI online. All right.

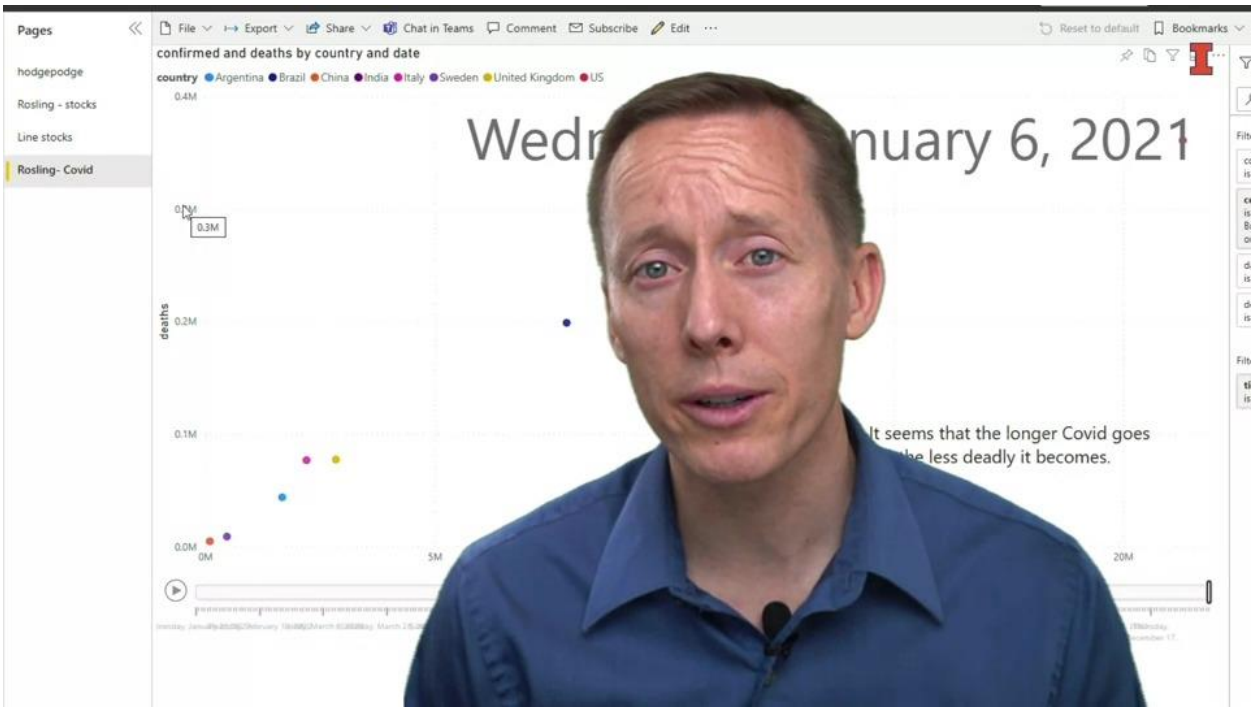


Here's my report in my personal workspace online, and if you explore it a little bit, you'll see that you can access the different tabs of this report, you will notice that our visuals are a Power BI Pro feature, so that's one thing that you have to pay for. That's a limitation, but it's still pretty awesome that for free, I guess your organization needs an account. You can publish these reports and you could share them to an area where others could access them, and they could apply filters, they could click on parts of a treemap or a map to apply filters that way.

They can't design and edit the layout of the report, but they can apply filters and explore it. This is a really powerful feature, I think, of Power BI, this is just an exploratory data analysis report. But even then, if you're working with lots of people, what a great way to allow them to access your data. Now, let me show you something else about this environment here.



If I go to the far left sidebar, click on this home tab, I can see favorite's and frequents, I can see data stories from Power BI community, there's some other tutorials. I'll just click on my workspace. In here, I can see other reports that I have created like this COVID dashboard, as well as the data that goes along with it.



If I want to view another dashboard, I can click on that, and now I can access this other report here and explore it. This is enhanced Rosling type visualization where you can look at the evolution of COVID over time. This is an example of posting an exploratory data analysis report online, but certainly, you could use this to post a finished overall report to highlight findings in your analysis.

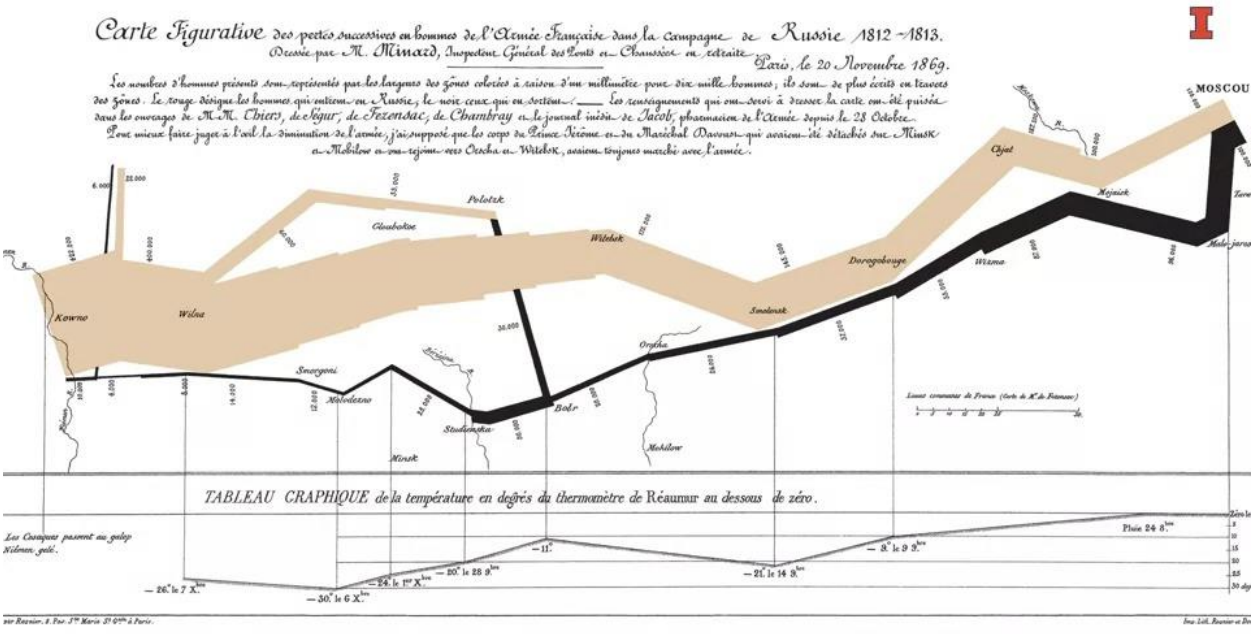
Module 2 Review

Module 2 Conclusion

In the summer of 1812, the French leader Napoleon created an army of 680,000 soldiers and began an invasion of Russia. At that point, it was the largest assembly of soldiers ever known. Napoleon marched with his troops to Moscow, Russia, and expected the Russian army to surrender. However, they never did.

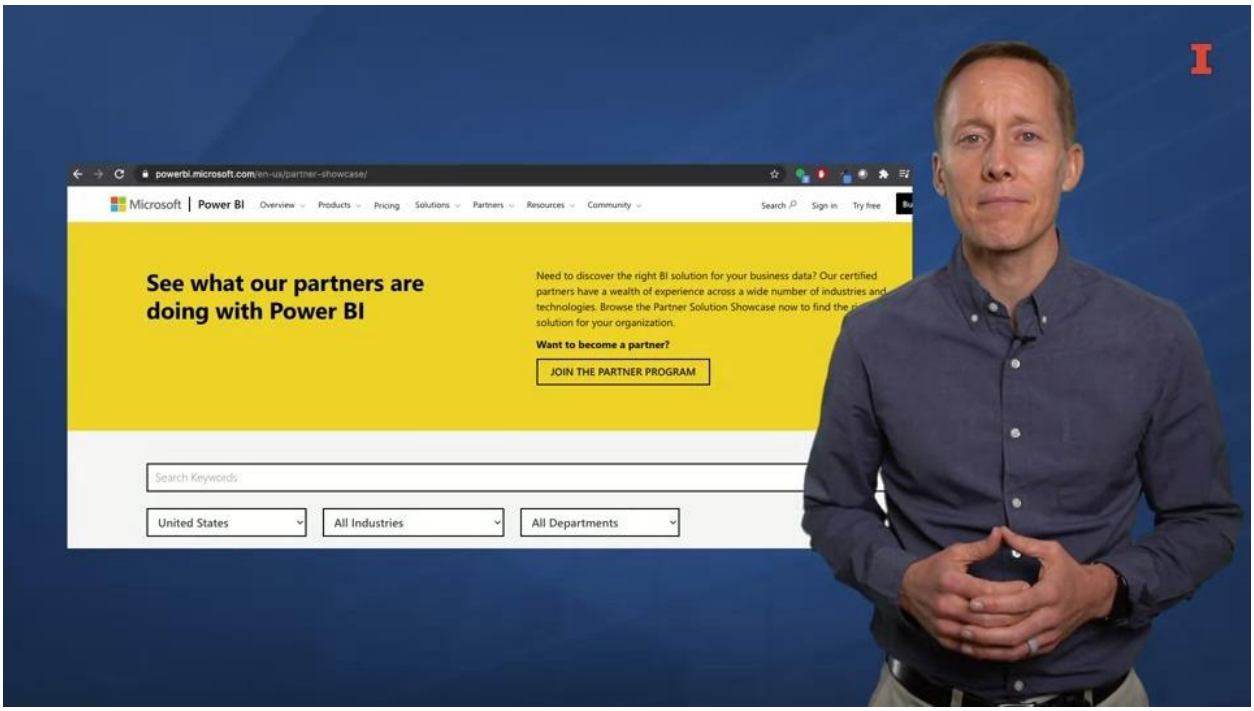
Scorched earth tactics in which the armies would burn everything and leave nothing for the opposing army to subsist upon left Napoleon's army lacking supplies. The distance and terrain made it hard for the French to deliver adequate provisions to its army. Eventually, Napoleon retreated and headed back to France.

By that time, the temperature had started to drop. As a result of the cold, inadequate provisions, and repeated attacks from Russian peasants, Napoleon returned to Paris with only 27,000 soldiers. I tell you that story because one of the greatest data visualizations of all time is attributed to Charles Minard's depiction of that ill-fated campaign.



Minard used basic elements to succinctly communicate at least five dimensions on the chart. One, bars communicate the location of the paths upon which the troops traveled. Two bars also communicate the distance that the troops traveled. Three, color communicates the direction that the troops traveled. Four, bar width communicates the number of troops in the army, and five, a separate line chart on the bottom is used to communicate the temperature during the French army's return from Moscow.

This is a great example of using data visualizations to tell a story. It's more impressive when you consider that Charles Minard was an engineer who died in 1870. Can you imagine how long it must have taken him to create that chart by hand? Can you imagine how happy he would have been if he had a tool like Power BI? How do you think he would have used the filtering and zooming tools to improve his visual story? I hope at this point you have a better idea of how power BI fits in with the business analytic workflow. It's a powerful tool for assembling and exploring data, as well as for communicating results with others. There are many wonderful modern examples of effective data visualizations that can be created with Power BI.



Microsoft has a partner showcase website where you can explore examples from many different industries and departments. I encourage you to consider these examples inspiration for how to use data visualization tools that you've learned about not only to explore data but also to effectively communicate a story. As you can see by now in these lessons, we've only scratched the surface of what is possible with Power BI.