STRINGFEST ANALYTICS

Five things Excel users should know about Power Query

Contents

0. Getting Started

Thank you for picking up this white paper. Power Query is the most exciting Excel release in 20 years. You are going to love it; your relationship to data and data cleaning will never be the same.

The objective of this white paper is to orient you to the tool and what it can do for you. We'll see how Power Query busts older limitations of Excel in an easy-to-use modern interface.

Please be aware that Power Query is **not fully available on every version of Excel**. You can see all the versions of Excel that feature Power Query here. I am writing this on a Windows PC with the Office 365 desktop version of Excel installed. For best results following along, you'll be on the same.

Companion workbook

One learns best by doing. The more time you have your hands on the keyboard as you learn PQ, the more you'll remember. To that end, **download the companion file for this white paper here.** This is a real-life dataset consisting of housing prices. Learn more about it here.

I will provide a completed version of this workbook at the end of the white paper.

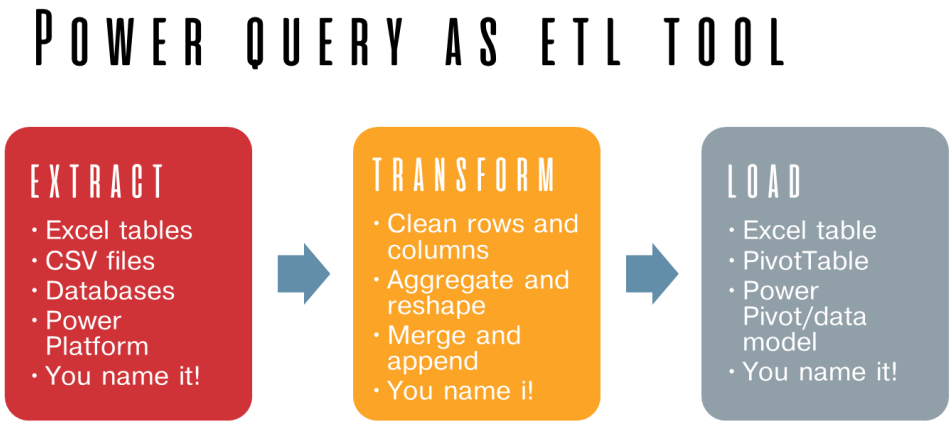
1. It's not just for Excel

While the focus of this white paper is on Excel for Power Query, perhaps the first thing you should know about this tool is **it's not just for Excel.**

You'll also find Power Query in Power BI and Power Automate, among other products in the Microsoft BI stack. Most if not all of the principles you'll read about in this white paper apply equally to these other tools. So you're really getting a bargain here (both from me and Microsoft)! You can learn more about where Power Query is available on the official Power Query page from Microsoft.

At the time of writing, the first sentence on this page to explain Power Query is this: "Power Query is the easiest way to connect, extract, transform and load data from a wide range of sources." Let's drill in on this definition.

2. It's an ETL tool



Let's be honest, there are a lot of terms in tech that sound way more complicated than they really are. Sometimes it seems like these acronyms are

used intentionally to create confusion. But peel back the concept and everything is clear.

ETL, or extract, transform, load is one of those terms. You'll hear database administrators and data engineers fawn and hype over their "ETL pipelines" and "ETL tools." Sounds like something only a certified data geek can do, right?

In fact, Power Query serves to democratize the process of ETL right to an Excel spreadsheet.

a. (connect to and) Extract

The first step of ETL is to "extract" the data from an outside source. To do that, you of course need to connect to the data. Excel Power Query can connect to an outstanding variety of data sources... not just Excel workbooks!

Some of these include:

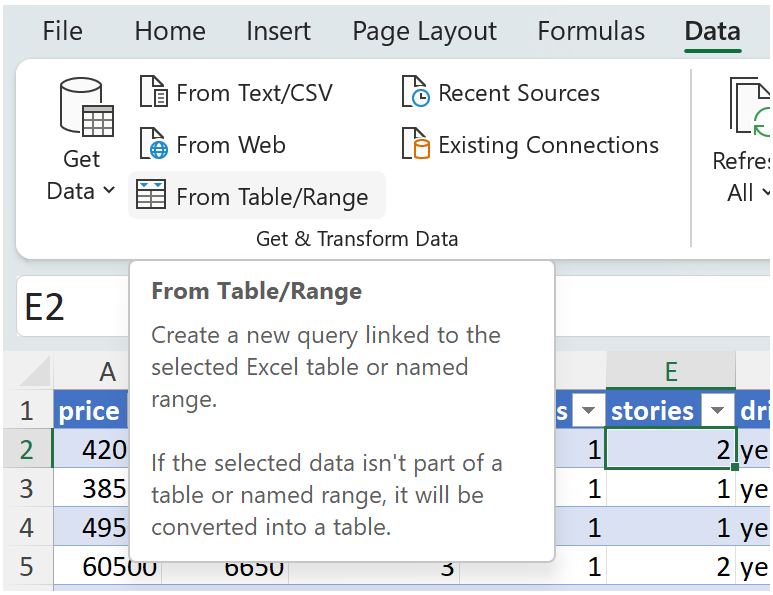
- Text and CSV files
- Databases
- SharePoint
- XML, HTML, Web data

For right now, let's just stick with a plain old Excel workbook. If you haven't downloaded the starter file yet, here's another chance. Download it now and open.

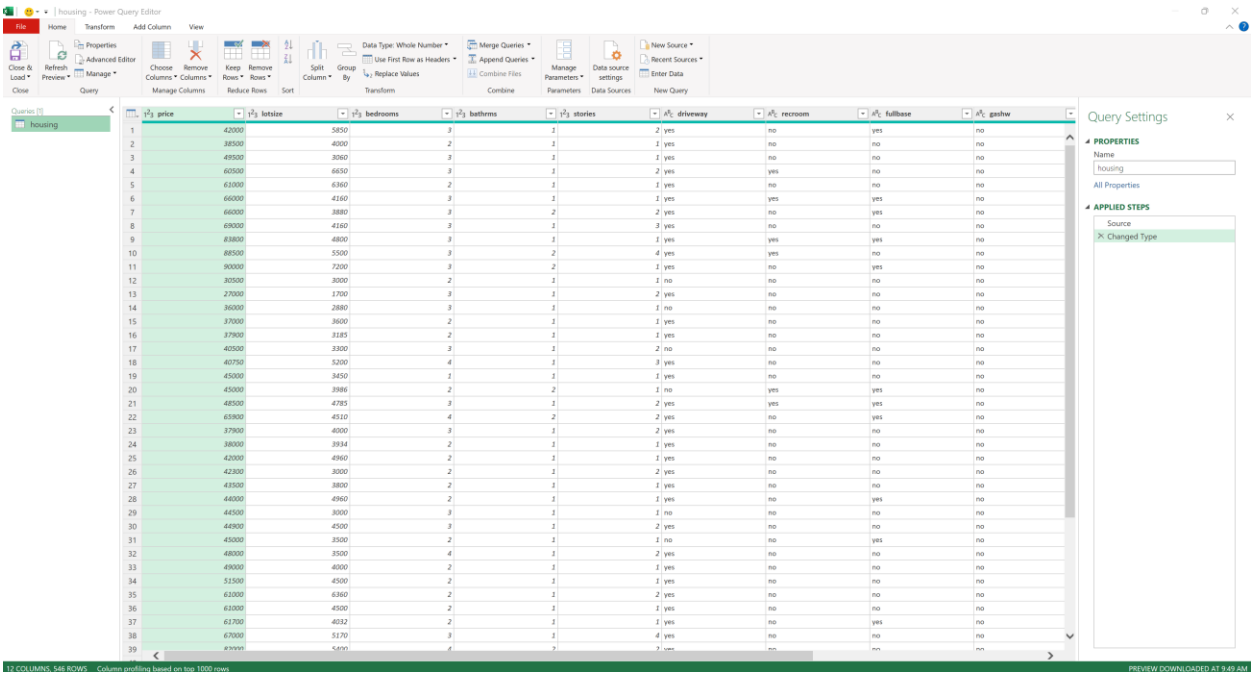
Our first order of business is to *extract* the data from the `housing` table. Now it seems a little strange to be "connecting to and extracting" data that we already have, right in this workbook. It's arguably overwrought/

What makes this such a powerful idea is that *Power Query is forcing us to keep the raw data intact*. We must take an *extract* of it (There's the E). So let's connect to this data, *then* transform it (There's the T!).

To do so, click anywhere inside the housing table, then go to Data > Get Data > From Table/Range:



The Power Query editor will now launch in a new window:



We'll cover the basics of the Power Query editor in this white paper. For a more detailed tour, check out this post.

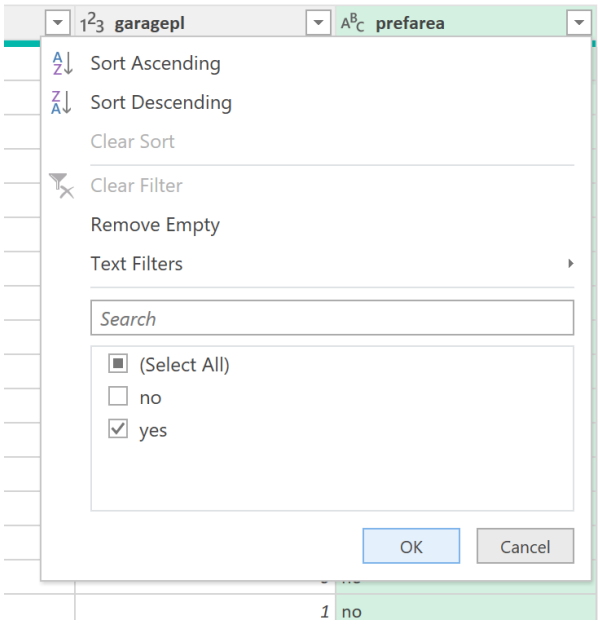
Congratulations on getting the ETL started! An E for effort... and extract! Let's move to the T.

2. Transform

The next step is to transform the data. This could be a whole lot of things, but it's really all the steps you must take to make this data usable, such as:

- Sorting or filtering rows
- Adding, dropping, renaming or calculating columns
- Grouping by and aggregating categories
- Merging multiple data sources

Let's make a very simple transformation to this data: we will filter it to keep only the records where `prefarea` are set to `yes`. Scroll the data to the right if you can't see this column on your screen. Then click the dropdown menu on the column header and click off the `no` selections.

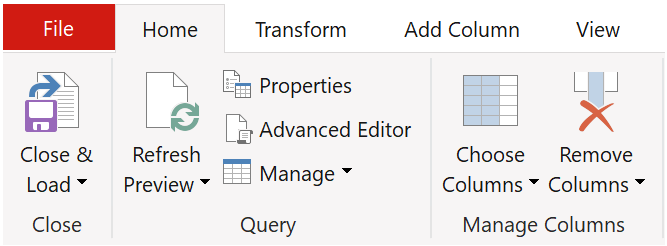


Click OK. And the data transformation begins!

There's of course a *lot* more you can do with data transformation in Power Query, and most of your time is going to be in this step of ETL. But let's move to that last leg, or L for "load" (say that 10 times fast!).

3. Load

Finally, click Home on the ribbon of the Power Query editor, then Close & Load:



You will see the "transformed extract" of your `housing` data loaded into Excel:

	A	B	C	D	E	F	G	H	I	J	K	L
1	price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase	gashw	airco	garagepl	prefarea
2	106000	7440	3	2	1	yes	yes	yes	no	yes	0	yes
3	155000	7500	3	3	1	yes	no	yes	no	yes	2	yes
4	141000	8100	4	1	2	yes	yes	yes	no	yes	2	yes
5	75000	10269	3	1	1	yes	no	no	no	no	1	yes
6	78000	6100	3	1	3	yes	yes	no	no	yes	0	yes
7	95000	6420	3	2	3	yes	no	no	no	yes	0	yes
8	97000	12090	4	2	2	yes	no	no	no	no	2	yes
9	107000	6600	3	1	4	yes	no	no	no	yes	3	yes
10	130000	6600	4	2	2	yes	yes	yes	no	yes	1	yes

Congratulations for completing an entire ETL job!

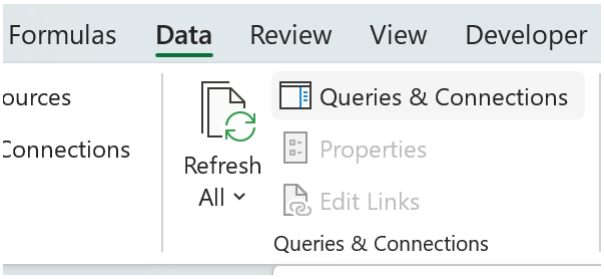
- You extracted the raw data from an Excel table
- You transformed the data using the Power Query editor
- You loaded the results back to Excel

This is of course a quite diminutive ETL. But regardless of complexity, Power Query lets you build repeatable data cleaning processes using a variety of sources. **Very** helpful, if I do say myself.

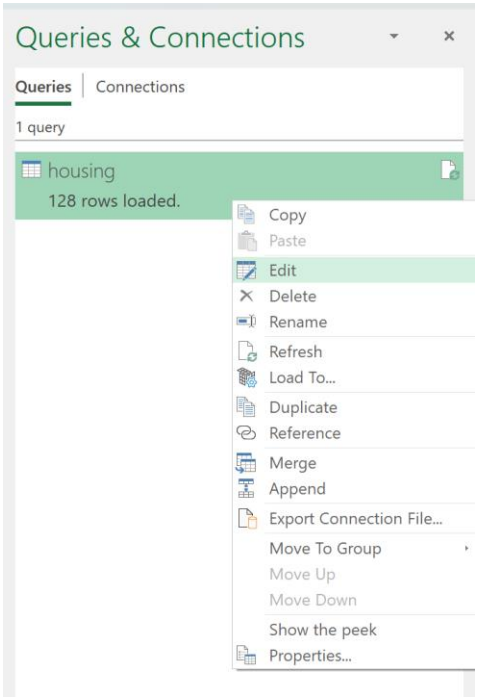
3. It generates code

Guess what... you just produced code in Power Query! Don't believe me?

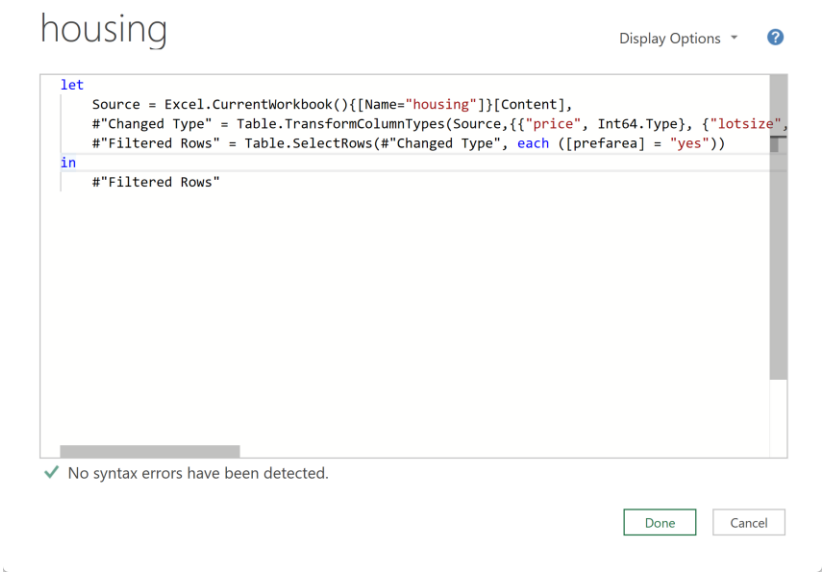
Let's go back to the Power Query Editor to find out. To do this, go to the home ribbon in Excel, then Data > Queries & Connections:



If it wasn't there already, a Queries & Connections menu will appear to the right of your data. Currently we just have the one, `housing`. Go ahead and right-click on this in the menu, then select Edit.



And just like that, you're back in the Power Query editor. This time, go to the View tab on the ribbon, then Advanced Editor. You will see an editor like this:



You can increase the font of this editor by pressing `Ctrl + Shift + Enter` on your keyboard.

Where did this code come from? The Power Query Editor is a *syntax generator*. That means that code, or syntax, is created each time you make a change in the Editor. The first step of the query is the first line: where did this data come from? That's the extract. After that, Power Query detects the data type of each column: does this column contain text? Numbers? Dates? We won't talk much about data types in Power Query; you can learn more here.

But after that? Those steps are all you, my friend. You are coding... but not. If you've used SPSS or the VBA Macro Recorder before, this type of behavior may be familiar.

But what *kind* of syntax is this? VBA? Python? Something else? In fact, this is the M language developed specifically by Microsoft for Power Query. To learn more about M, check out this article from Microsoft.

The vast majority of your work in Power Query can be done without ever interacting with the generated M code. But as you progress in your Power Query journey, you may find yourself coming here for certain tasks.

Let's make an edit here ourselves: change the filter so that `prefarea` must now equal `no`:

You should see at the bottom of the editor a checkmark and message that "No syntax errors have been detected." Great! Go ahead and click Done. You'll see the data in the Power Query editor change.

From here, you can Close & Load the query again. Rinse... or should I say transform... and repeat!

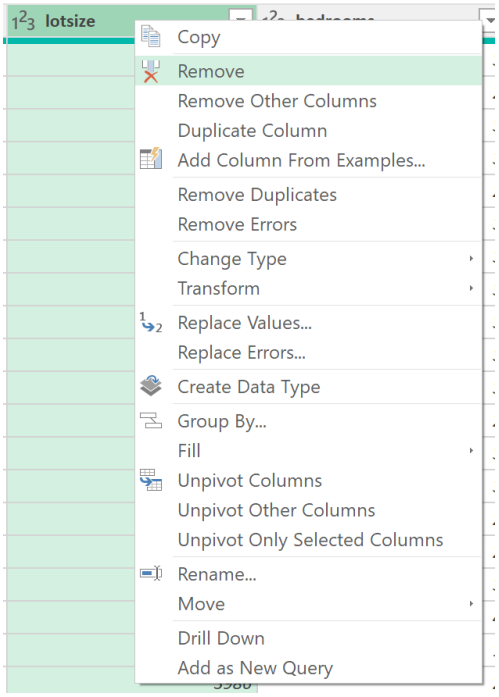
4. It's fully reproducible

Picture this: You're working in an unfamiliar Excel file and delete a column. Everything in it breaks. Or you are trying to audit through some process and just can't figure out where the numbers came from.

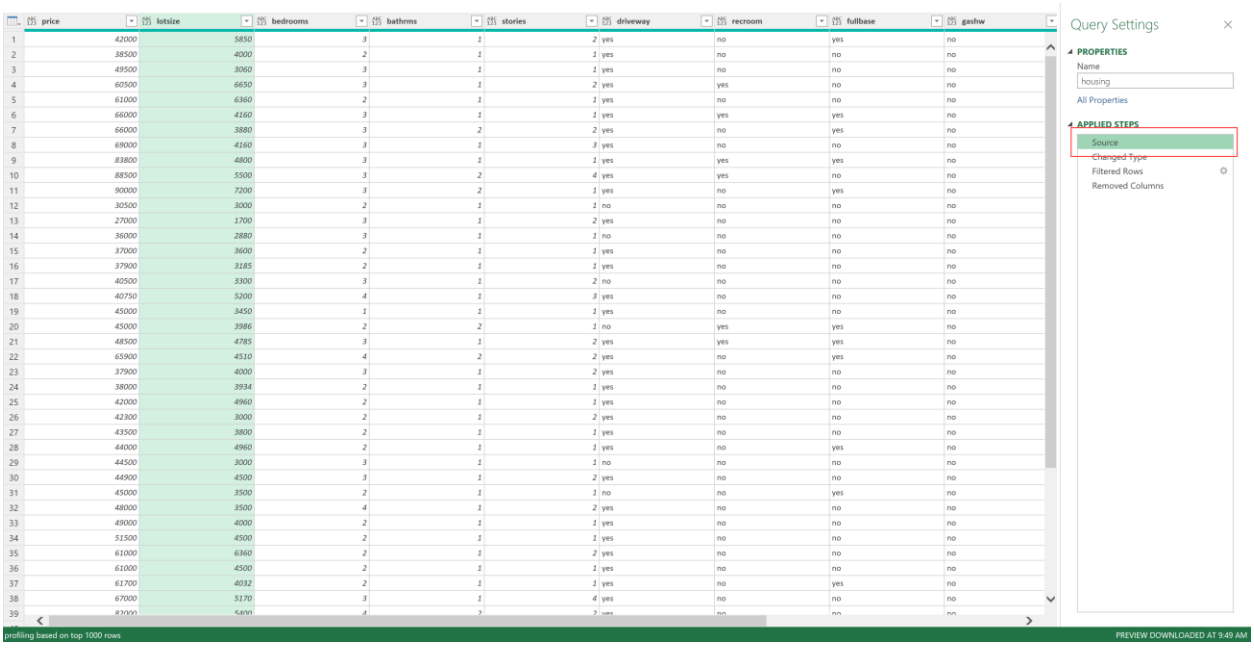
Every process in Excel and data analysis should be like a recipe. You are given a starting set of inputs and clear procedures to follow. At the end, everyone ends up with the same thing. This concept is known as reproducibility, and it's not always been Excel's strongest suit.

Contrast this to Power Query, where every step of the process is clearly recorded as code... and displayed in a more accessible Applied Steps menu. To demonstrate this, go back to your `housing` query. We are going to delete the `lotsize` column.

You can do this by right-clicking on the column header and selecting Remove:



In traditional Excel deleted columns could prove catastrophic. In Power Query it's not a big deal, because you always have a copy of the source data to go back to. And to view that source data, head over to the Applied Steps menu toward the right of your dataset, then click Source:



Guess what... the `lotsize` column is back. Not only that, but if you scroll over to `prefarea` you'll see that it's no longer filtered. This is a picture-perfect view of the original data source.

The Applied Steps menu allows you to step through each part of the data cleaning process to take another look at the data. It lists in plain English what was done to the data, when.

This is your recipe for reproducibility! Perhaps the most potent argument against Excel has been defeated here.

5. It's an Excel myth buster

Excel is a legendary business application. But sometimes, legend passes to myth. And not all myths are true, at least not anymore. At least not with Power Query. Consider this tool the ultimate Excel myth buster:

Myth 1: "Excel is not reproducible"

We already covered this myth. You'll hear often that Excel is not reproducible. People will point out that there's not "recipe to follow" when rows and columns are manually filtered and deleted. That's all true in basic Excel, but not in Power Query. You know better now. Worth the price of this white paper, I'd say.

Myth 1: "Excel can't handle over one million rows"

OK, now *this* one may be my favorite myth to bust. The Excel-averse often make it their trump card: we're in the age of "big data," right? Excel can't even go past ~1 million rows... that's peanuts these days!

Talk about a half-truth... it *is* still true that an Excel workbook *itself* can only contain 1,048,576 rows. But this *does not* mean that you can't analyze more than a million rows in an Excel workbook!

How can this be? You guessed it — Power Query. Again, with this tool we can *extract* data from a variety of sources ranging from CSV files to databases and more. It's true that we can't *load* more than a million rows, but we are free to aggregate and summarize the data in Power Query, then load those results to Excel. And who wants to skim > 1 million rows in a worksheet, anyway?

For a more serious demonstration on how to blow past the alleged million-row Excel limit, check out this post from Master Data Analysis on analyzing **fifty million rows** using Excel Power Query. And for something more humorous, check out my meme-video on the topic…

Myth 2: "Excel has no true `null`"

If you've ever used relational databases you're familiar with the concept of a dedicated keyword for any observation that is missing: it's `null`. When something is `null`, *it does not equal zero*! If we *knew* what the value should be, we would just fill in a zero.

It's very easy to confuse these values, especially in basic Excel which has no `null` keyword. Fortunately, Power Query makes up for this shortcoming with the conventional `null` keyword. For more on how to understand missing values and to see them done right in Power Query, check out this post.

Myth 3: "Excel only works with structured data"

Now this one I can be a little more sympathetic to, because in general Excel really isn't the best tool for working with unstructured data. That said, it's worth pointing out that Power Query does make it a little more possible and should expand our horizons about what Excel can do.

If you're not familiar with the idea of unstructured or its opposite, structured data, think of how spreadsheets are laid out — in rows and columns. Generally there are some patterns and rules you can use to determine what type of data goes in those rows and columns. This is structured data in a nutshell. Excel thrives in this territory.

Then there is the messier data like text, images, audio and so forth. Not real simple to cram this into a spreadsheet, right? This is unstructured data.

Again, I agree that Excel shouldn't be the go-to solution when working with unstructured data. That said, Power Query does make it possible to do these things. For example, we can Power Query to perform basic sentiment analysis or parse so-called "semi-structured" JSON data.

What other myths will you find?

I hope these Power Query possibilities are getting you to "think outside the cell…"

As you progress with this tool, think about what other Excel myths you're able to bust! Feel free to drop me a line with how it's going. Which brings me to the end of our white paper...

Conclusion and next steps

Congrats for taking your first steps using Excel Power Query! I hope you see the vast possibilities here: fully reproducible workflows, from a variety of sources, right in Excel, using a low-code interface with easy access to the source code. What's not to love?

A few closing resources follow:

Companion solution file

First of all, here is the worked-through download file to accompany this white paper. Were you able to follow all the twists and turns along with me? If you weren't, you know where to look to see every step applied... (Hint... it's the Applied Steps menu!)

First Steps in Power Query course

If you're excited for more, I invite you to take my First Steps in Power Query for Excel course:

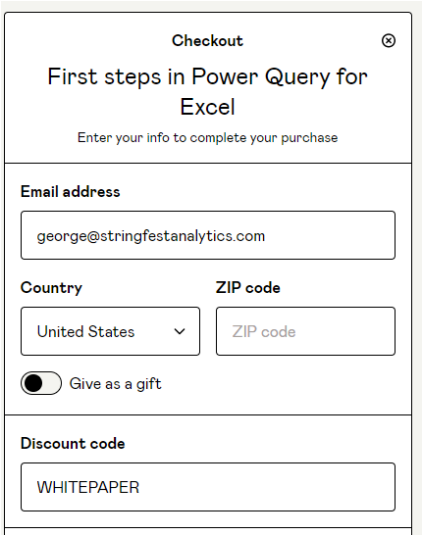
This course contains nearly 3 hours of video instruction on the basics of Power Query for Excel, along with in-depth demo notes, drills, datasets and more. Learn the fundamentals of Power Query in under half a day, then use these techniques day after day.

Promocode for course

Thank you for taking the time to read this white paper! I can tell you are serious about this stuff. As a special thank-you, I'm offering **25% off the course** with promocode WHITEPAPER.

To redeem, click under "Buy this" on the course listing and enter under "Discount code."

Looking forward to seeing you in class!



Thank you

Thanks for picking up this white paper as a guide for adding Power Query to your Excel toolkit.

I invite you to continue reading my newsletter and perusing my blog for more analytics content. You're also welcome to get in touch if I can help your organization with the services listed here.

I offer a full-day Excel Power Query workshop to corporate teams up to 20 people in-person or 35 online.

One last thing: please give me a follow on Twitter and LinkedIn. Your support is appreciated.