* One thing you’ll see that’s interesting about Power Query is that is has brought a lot of best practices in IT into Excel, one of those being the “extract, transform and load” methodology. So, I want take a brief pulse check here, is this a term you have heard of before? And honestly if you don’t, that’s why you are here.
* [Observation about the results] Now I’m going to be very scientific here and go to Wikipedia to explain how extract, transform, load or ETL methodologies work:
  + A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions
* So, you can see ETL achieves quite a few purposes, and as the name implies it works in three steps:
  + First, you want to extract the information from some outside source, this can be anything from a database to a web page.
  + Then, you want to transform it. This could be cleaning the data, filtering it, checking for assumptions, you get the idea.
  + Finally, loading the data, this places it in a format that is accessible for end users to work with.
* You will see that Power Query essentially works in this three-step process, so some of the tasks that you might use it for include:
  + Data retrieval
  + Data cleaning
  + Data combining
  + Data sharing/reporting
* As I mentioned, ETL is all over the IT landscape and is considered a best practice when you are working with large amounts of disparate data that need to be “fit for general office consumption.” It’s not just an Excel thing, and in fact this is a pretty disruptive tool to have in Excel, and it really busts a lot of the myths about what Excel can and can’t be which I would like to cover right now.
* So, let’s bust some myths in Excel.
* First, that Excel is not reproducible.
  + Now, the idea of reproducibility is that, the process of getting to some output is fully transparent and it’s possible to walk through each step of the process and get the same result at the end. A lot of typical Excel reporting is *not* fully reproducible, for example consider a report that depends on you deleting columns A through C to work. Now, if I was going into that final report to understand how it was built, there would really be no way for me to know that, right? On the other hand, a fully reproducible report would show me that, and ideally I’d just click a button and it would do it for me.
  + In data work, reproducibility is often achieved by powering an analysis with code, so that you can indeed run a script and watch the results emerge one line at a time. Now, the nice thing about Power Query is that it *is* a code-driven process, so it is reproducible. So does this mean you need to learn how to code in this class? Not necessarily right now. For everyday data cleaning processes, we can use menu-driven options, that will produce code in a Microsoft computer language called M, and now our data-cleaning steps are fully reproducible, in Excel, with no code needed. So that is really cool and a huge myth busted.
* Another myth, that Excel can really only hold structured, or tabular data.
  + This is the type of data that in fairness we are more used to working with in classic Excel, data that comes in a nice rectangular shape, usually coming from an Excel workbook or csv file. If the data were in a so-called “unstructured” format, such as song lyrics or a website structure, good luck with using Excel for that.
  + Come over to Power Query and there is quite an expanded range of possibilities for where data can come from, I will list some of them here:
    - Access
    - Text files
    - SQL
    - XML
    - Web
    - SharePoint
    - Hadoop
    - oData
  + So this is a combination of quite a few data sources, and to make it even better, with Power Query we are able to take data from these disparate sources, combine and clean the results, and this whole process is pretty fast.
* And that’s the last myth I want to bust here, is that Excel can’t handle large datasets. Now, with “classical Excel” this was probably true, that Excel just couldn’t handle some of those bigger datasets that a lot of data analysts are working with these days, but that’s just not true when you are using PowerQuery, and here is an example of that, a blog post about analyzing 50 million records using Excel. 50 million is pretty significant for most anyone, and Excel could handle it, so dismissing Excel out of hand on account of performance is just not as true anymore with Power Query.
* Now you are hyped up Power Query, I think you are going to love it and it will be hard for you to remember the days when you didn’t have it. So let’s think about what we would do without Power Query.
  + Here I have a dataset, wholesale\_customers.xlsx. Let me take a look at this dataset.
    - Now what I would like to do is turn this into a PivotTable, so that I can easily get the total sales by channel and region. So, I am going to ask, what is making that difficult to do with this data in its current state? What would I need to do to fix it? [Ask questions]
    - That’s right, so because we have sales being tallied across several columns here, it’s difficult to summarize this data. Now you’ll see in Power Pivot, I can correct this problem pretty easily.
    - But before getting too into the “how’s” of data cleanup, what I would like to do is really focus on the “why’s.” Why exactly is this data so hard to analyze? What does this tell us about how data works? Let’s cover that in this second module.