High level stuff

* We will just focus on tables of data
* Find the list of the page’s queries and connections by going to Data tab, Queries & Connections. d

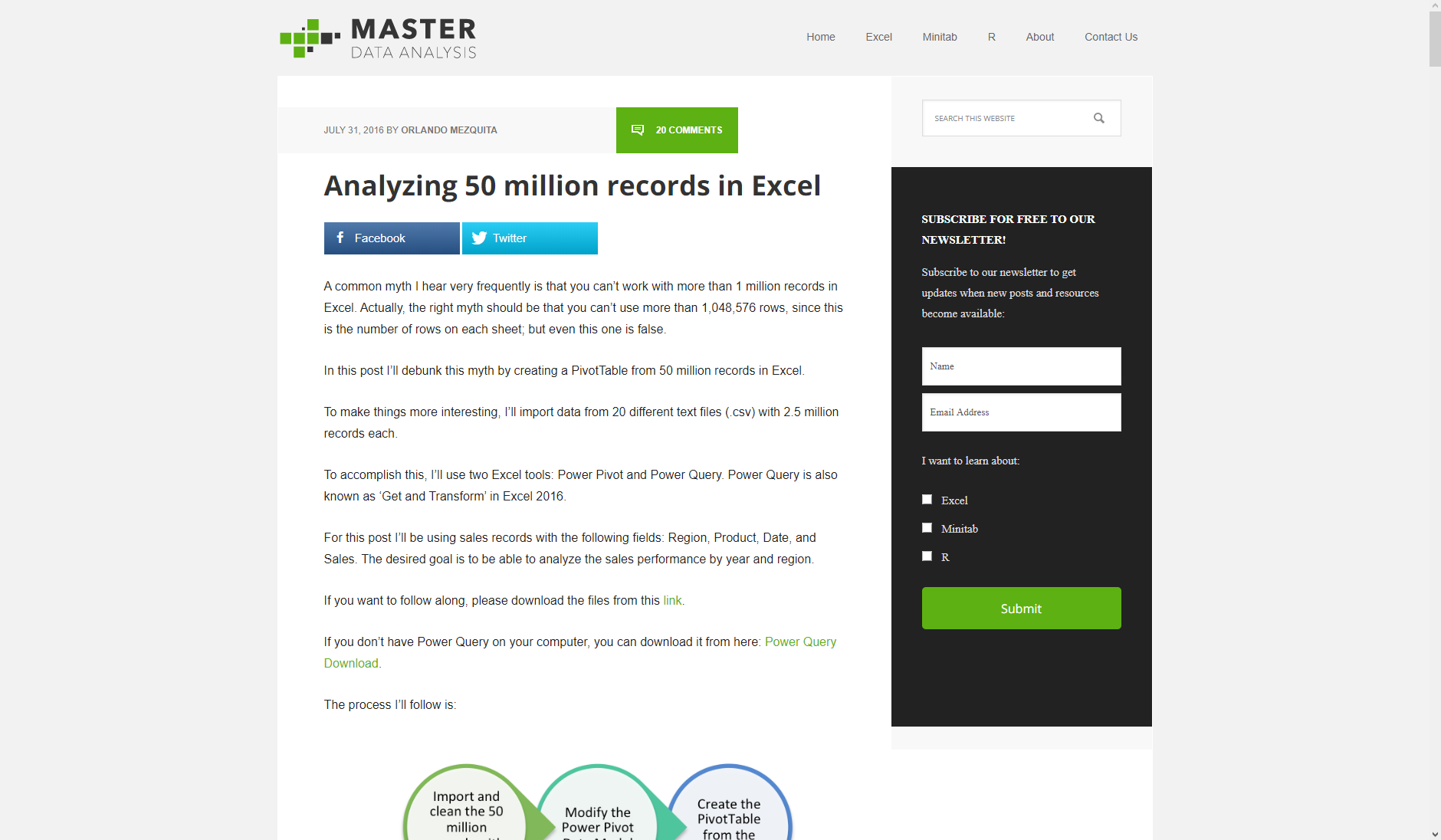
1. **Power Query as Excel’s ETL tool** (25 minutes)

* Presentation: Why ETL for business analytics?
  + Overview of what is meant by “extract, transform and load” processes, with applications in business analytics and reporting
* Presentation: What would we do without Power Query?
  + Exercise: Assess how to clean a messy dataset using known tools. Explain what makes that dataset messy.
    - Wholesale customers dataset
  + Tour of the alternatives to Power Query: complex formulas, VBA, SQL, and more
  + Overview of Microsoft’s “Power Platform” as applied to Excel: Power Pivot, Power Query, M, DAX.
* Q&A

1. Power Query vs Power Pivot vs Power View
   1. Power Query is the ETL process of this.
2. How do I get Power Query?
3. What will I use Power Query for?

Power Query & Excel Fact-Checking.

* “It’s not reproducible”
  + Fully powered by a script
  + You can generate syntax with your menu choices
    - (If you’ve ever used SPSS it’s like that)
  + You can also write your own code (we aren’t)
* “It’s slow/can’t handle large datasets”
  + <https://www.masterdataanalysis.com/ms-excel/analyzing-50-million-records-excel/>



* “It can only handle tabular data” (Some options include)
  + Access
  + Text
  + SQL
  + XML
  + Web
  + Azure
  + SharePoint
  + Hadoop
  + oData

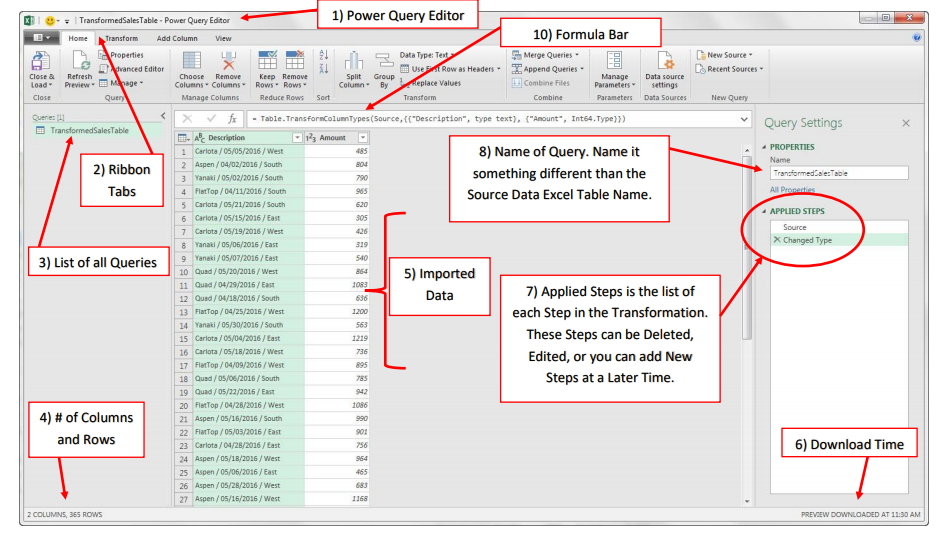
What do we do with Power Query?

* Data retrieval
* Cleanse
* Combine
* Share

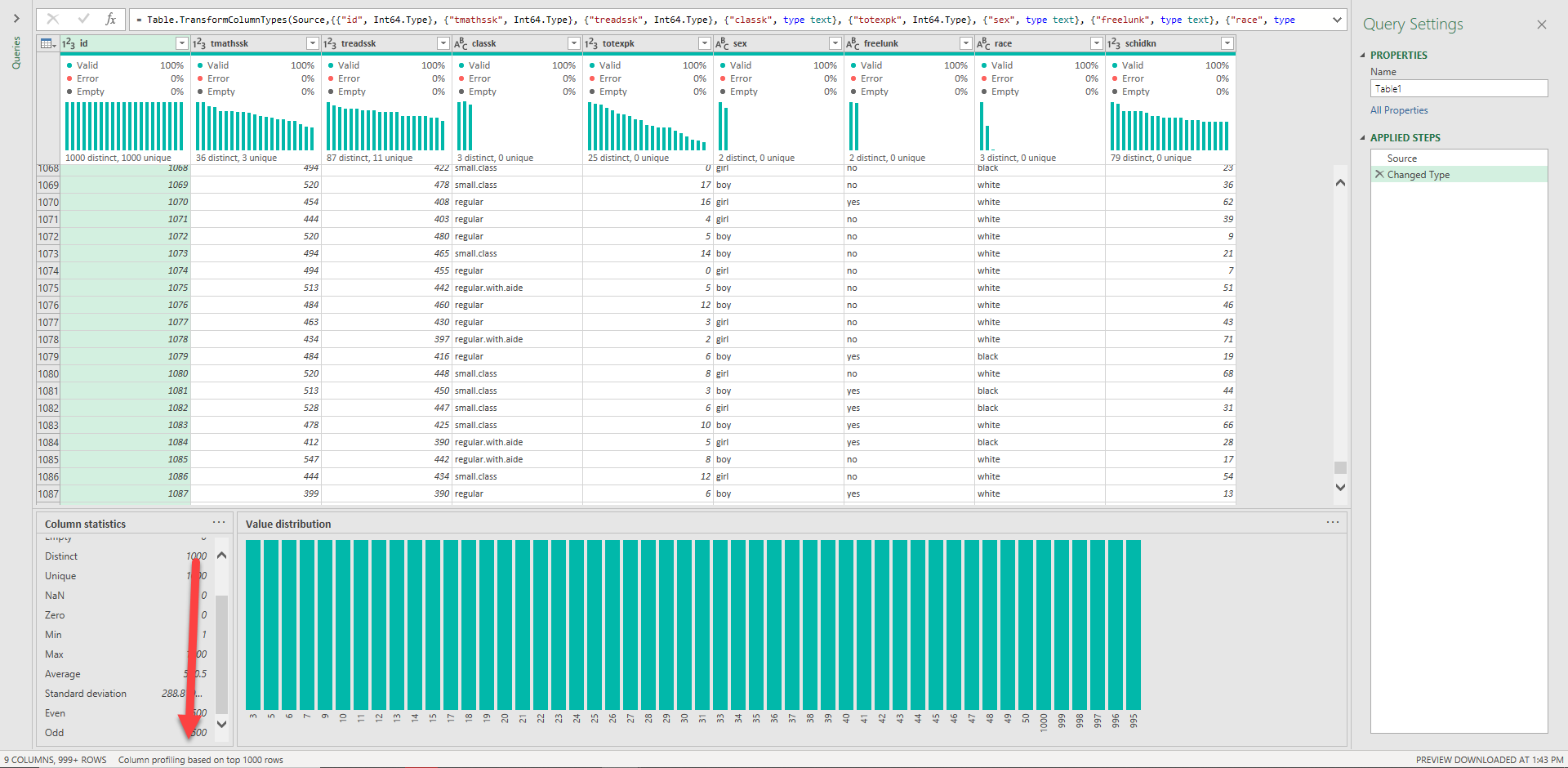
**What Makes Data “Tidy?” Why Does it Matter?** (25 minutes)

* Presentation: Spotting messy data in the wild
  + Using the principles of “tidy data,” explicitly state how to reshape a dataset for ease of analysis
    - NYTimes: 50-80% of data science time is
    - You don’t want to re-invent the wheel every time you need to prepare data
  + How does data work?
    - Every value belongs to a variable and an observation.
      * If you are used to Tableau, these are dimensions and measures.
    - Drill: which variables in some dataset are dimensions and which are measures?
    - To get data tidy:
      * Each variable forms a column
      * Each observation forms a row
      * Each type of observational unit forms a table
    - “Are values fixed by the design of the data collection, or measured during the course of the experiment?”
    - (Find a better way to describe this) If you are measuring in the same unit in different columns, that’s probably a bad sign.
    - If you are measuring the same observation in the same row, that’s also probably a bad sign
* Exercise: Getting to tidy: what needs to change with these datasets?
  + Practice a couple more
    - Include things like totals and columns that should be split
      * Cuz then we could actually use the same datasets later on!
* Presentation: Power Query First Steps
  + Dataset: star.xlsx.
    - This is a clean dataset but there are still things we can do with PQ.
  + Load a first data source into Power Query, inspect it with data profiling, and begin the data cleaning process
  + Data -> From Table/Range
  + Tour of what you will see in Excel Power Query:

<https://people.highline.edu/mgirvin/AllClasses/348/MSPTDA/Content/PowerQuery/003-MSPTDA-IntroToPowerQuery.pdf>



* + Overview of the Home Ribbon – focus on the first three tabs, those are going to have the data cleaning functionality.
  + Then go into the dataset, show that we can click on the rows and cells and we can see their value at the bottom.
  + Then go to the header row, you will also see there are some shortcut options you can choose here.
  + Click on the column you’ll see you can filter it just like in native Excel.
  + You’ll also see that they have data types. Each of these columns have an explicit type.
  + Right click on the column and you’ll see you can get more options here.
    - Hit Control and page over columns to look for more.
  + Go to the View tab
    - Initially this is based on just the first 1,000 rows. There is a bottom waaay at the bottom, “Column Profiling Based on Top 1000 Rows,” click on that and it will expand through the other rows.



* Discard the results, we will go into that later.
* Q&A

Break (10 minutes)

* Okay so for a drill what I will have you do

**Transforming Rows in Power Query** (25 minutes)

* Presentation: Sorting, removing duplicates, filtering, filling a table (maybe grouping)
  + Demo: office characters
    - We have a list of RSVP’s to an office party and we want to clean up our dataset a little bit. (You are consulting with the Party Planning Committee).
      * Sorted A-Z
      * Remove duplicates
    - We could do this in basic Excel but we may have more people signing up for the list. And we may want to use it for other parties.

1. Create the connection (table)
2. Immediately you see a null. This is a special thing in Power Query. We want to get rid of them so we can hit our drop-down filter button and do so.
3. Now while we are up here we can sort from A-Z as well.
4. Check it out, now we have a running list of steps we have taken on this data.
   1. You’ll see that the first step is
5. If I hit the gear-box, we can get another look at what we did here.
6. If I remove it, we will see that things are still sorted from A-Z, but we have the nulls still.
   1. There is no undoing this!
7. So let’s filter out the NULL’s again.
8. Now let’s remove the duplicates, to do this I will go to Remove Rows -> Remove Duplicates.
   1. You’ll also see there is an option here to remove blank rows, this would have worked the same way to get rid of the NULLs.
9. Last but not least, there is a `Klevin` in here, we don’t want that, of course the data wouldn’t have been able to tell us that! .
10. Now let’s close and load this. I’m just going to hit the standard close and load.
    1. You’ll see we have a query here.
    2. We have the Queries & Connections tab here on the right, named Table1, that’s not a very descriptive name so let’s rename that to party\_rsvp.
    3. If you want to close this out, you can open it again, it’s Data -> Queries & Connections.
11. Check out that we have two tables.
12. We still have the original data, that is nice. Let’s say we add to it.
    1. I am going to add a Roy and then insert a blank row.
    2. Now right-click that and the resulting table is updated.

* + Demo: regional sales

Here we have a table, we don’t have a header row, and we need to fill down the Region, and we would like be able to slice and dice on this data a little bit, for example filtering, aggregating and so forth.

1. Select everything, our table does not have headers, but we are going to name them, just double-click on the column name. See that those come across as steps in the query as well.
2. Now we want to fill down the blanks for the Regions, we can do that by going to the Transform tab, you will select Fill, Fill Down to fill the nulls down with blanks.
3. Now we’ve got everything. Close and load but I’m really not interested in this table per se, I’d like to be able to aggregate and summarize it, and to do that I can automatically add to a PivotTable. It says PivotTable report but same idea.
   1. See there are some other options here, we can create a Table, PivotTable, or Only Create Connection where the data is not loaded into the workbook, we just have a connection to that data.
   2. Also there is the option to add this data to the Data Model, this would be used if you are using Power Pivot.
   3. And now we can use it like any other PivotTable.

* Exercise: drills on transforming rows
  + Drill: Census regions

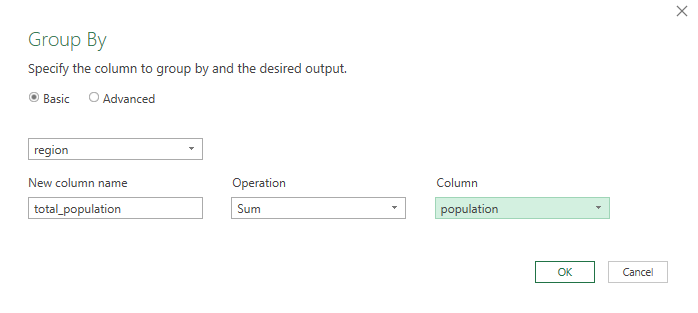
1. Name this query “State populations”
2. Get rid of the Total row.
3. Fill down the region and division
4. Sort by State Population, from high to low.
5. Then load this into a PivotTable to analyze.

Things to show during the demo:

1. If I hold down the Control key I can fill down two columns at the same time.
2. I can name the query over in Properties when I am in the Query Editor.

After the demo, show that you can group:

1. Let’s say I know that I am always going to want to analyze this by Region. I can actually group this data in Power Query rather than throwing it together in PQ:
   1. Right-click on query name, Edit.
   2. Right-click on region column, go to Group By. I am now going to group by total state population by region.



* Q&A

**Transforming Columns in Power Query, Part I** (25 minutes)

* Presentation: Changing data types, splitting columns, and re-formatting text.
  + You’ll want to show how to remove columns too. And duplicate them, etc. etc.
* Demo: dvd-rentals.

Start from left to right

1. Convert Title and Artist Name to Proper case this is likely to be more effective on the Artist Name column.
2. Also let’s may as well add a space between any colons or spaces here
3. Now let’s Text to Columns this thing.
4. UPC and ISBN code are probably better classified as Strings than numbers so let’s go ahead and change those.
5. We don’t need BTKey so delete it.
6. Can go ahead and change Retail to Currency if we want.
7. Finally Release Date, looks weird. So we have some options here.
   1. Right-click and transform. Here we can just have the date but let’s say for example that we want to split this into three separate columns, month day and year.
   2. If we just transform one of them it’s going to transform our raw date, so we want to duplicate it first before transforming it.

* Exercise: drills

File: orders

1. Convert the Date column to month
2. Convert account name to Proper case.
3. Turn Opportunity name into three columns:
   1. Vendor
   2. Status
   3. Order Type

* Q&A

Break (5 minutes)

**Transforming Columns in Power Query, Part II** (30 minutes)

* Presentation: Concatenating columns, creating calculated fields, and un-pivoting tables
  + Demo: census-population-density.xlsx

1. First concatenate name and abbreviation
   1. Works kinda like PivotTable calculated fields
   2. Click link to [Learn about Power Query formulas](https://docs.microsoft.com/en-us/powerquery-m/power-query-m-function-reference) – this is M code.
   3. If I make a mistake I can go back to that gear box.
   4. Move this new column to the front of the dataset
   5. Delete the other columns, no way there will be issues with formula errors!
   6. Now I want to calculate population density. This is where ordering the query steps can really matter. Because we could calculate each column at a time but it would make more sense to reshape this dataset.
   7. Right-click and Pivot
   8. Now we can rename these fields
   9. Now we can calculate population density.
   * Exercise: Wholesale customers
2. Tidy this data
3. Create a field calculating 10% of the sales.
   1. You can do this with your own calculated field.
   2. There is also a way to transform this -> Number -> Multiply.

* Presentation: Appending several tables
  + Demo dataset: Oscars

1. This time we have two datasets, they are in csv format. We want to stack them into one.
   1. Could happen if you are pulling data from multiple places
2. Open a new Excel workbook
   1. Data -> From text/csv
   2. Open Oscars\_yes. We have load this in to our workbook session.
      1. We have the option to load it in as-is or transform it. Let’s load it as a table just to take a gander at it.
   3. Do the same for Oscars\_no.
   4. Now we have two queries. So I want to append the “no’s” to the end of the “yes’s.” To do that I will click on that query name and here if we go to the Query tab, I can click Append.
      1. And we can append Oscars\_no to Oscars\_yes.
   5. Now we have created a new table.
      1. Again see how we have the yes and no files intact. And now a third query which has been loaded into a Table.
      2. By default PQ will append the columns based on having the same names. If columns are not the same names and you want to append them
      3. And if you have extra columns then they will just be NULLs.
3. Drill – do the same to the hof\_inducted and hof\_not\_inducted tables
4. This will work the same as the above
5. Notice that this doesn’t have the player’s name, just an ID field.
   1. Show the Master table, hey guess what we do have these names, how would we do this in Excel? With VLOOKUP(). Well we will transcend this in Power Query.

**VLOOKUP(), meet JOIN** (30 minutes)

* Presentation: Illustrate the differences between left and inner joins

1. First show the typical Venn Diagrams
2. Then show the animated joins: <https://github.com/gadenbuie/tidyexplain#mutating-joins>

* Presentation: Conducting joins in Power Query
  + Demo: Nycflights13 flights & planes table.

1. Pull up flights-and-planes.xlsx to really explain what it is you are doing.
   1. We have this list of flights and we would like to look up some information about these planes. But we don’t have information about *all* these planes.
      1. So, do we want to keep the information about the records without a match? Therein lies the main distinction between a left outer and an inner join.
2. So we can load these into our workbooks and now this time we will go to MERGE them rather than append them.
   1. Now our “point of reference” is flights. For the sake of what we are doing now imagine that the table which you pick first is going to be the one that you would write a VLOOKUP() on.
      1. Righ-click, merge and now we choose what “tables and matching columns” we should use.
         1. So we also need to click on the matcing rows.
      2. See from the resulting match that we are NOT going to get a match in every record by doing this!
      3. Hit OK
   2. Now we have a new column called planes. See that we have a tiny left<->right button.
      1. Click that and now we can select all of the “lookup” columns that we want.
      2. And now they are joined.
   3. Close and load. It’s a big file so it’s gonna take some time.
      1. Name it left join
      2. Check it out, we have the same number of rows in this file as we do in the flights file!
   4. Now let’s make a brand new query, do everything the same this time except it’s going to be an innerjoin.

* Exercise: Drills
* We have the hall of fame, and we have a list of people. But our list only includes people, A through M.
* We will join our combined hall of fame table onto this players A through M table. What is the result?
* Q&A