# Powering up your data work with Excel's Power Query

Angela Zoss, Duke University Libraries

November 2024

<https://github.com/amzoss/power-query>

## Example 1: Fix common data issues

Data source: [State CO2 Emissions](https://www.epa.gov/statelocalenergy/state-co2-emissions-fossil-fuel-combustion)

1. Open a blank workbook
2. Under “Data”, click “Get Data”
3. Choose data source: Excel workbook
4. Select “state\_co2…xlsx” file
5. Select the “CO2FFC” sheet, check the preview, and click “Transform data”
6. Power Query Editor opens in new window
7. Remove non-data rows
   1. Click “Remove rows” 🡪 “Remove top rows” 🡪 “1”
   2. Click “Remove rows” 🡪 “Remove bottom rows” 🡪 “2”
8. Click “Use first row as column headers”
9. Fill in state name for all rows
   1. Right-click “State” column header
   2. Hover over “Fill”
   3. Select “Down…”
10. Remove subtotals
    1. Click on the filter menu for the “Sector” column
    2. Select “Remove empty”
11. Unpivot year columns
    1. Option 1: Select first two columns, unpivot other columns
    2. Option 2: Select all year columns, unpivot only selected columns
12. Rename new columns
    1. “Attribute” -> “Year”
    2. “Value” -> “Emissions”
13. Push data back into Excel with “Close and Load”

## Example 2: Additional data cleaning, appending data sources

Data source: [Sample training attendance data from Library Carpentry](https://librarycarpentry.org/lc-spreadsheets/#data)

Exercises inspired by Library Carpentry lesson: [Tidy Data for Librarians](https://librarycarpentry.org/lc-spreadsheets/)

1. In a new Excel file, get 2016 data from “training\_attendance.xlsx”
   1. Power Query 🡪 Get Data 🡪 Excel workbook 🡪 select file 🡪 select 2016 tab
2. Sheet has two data tables; we can duplicate the query to deal with each
   1. Right-click the 2016 query in the Queries sidebar
   2. Select “Duplicate”
3. Rename each query to match data tables: RDM and Open access
4. In RDM query, keep only the first 4 columns
   1. Select remaining columns
   2. Click “Remove columns”
5. Remove blank rows
   1. In “Remove rows” menu, select “Remove blank rows”
6. Split the third column
   1. Note: Column contains number of attendees of different types (post-graduate researcher (PGR), post-doctoral research associate (PDRA), and other)
   2. Select “PGR|PDRA|other” column
   3. Click “Split column”
   4. Select “By delimiter”
   5. For Separator, select “--Custom--”
   6. Type “|” as the separator
7. Use correct column headers
   1. Remove top 1 row
   2. Use first row as headers
8. Calculate total attendees
   1. Click on “Add column” in ribbon menu
   2. Select “Custom column”
   3. Name the column “Attendees”
   4. In the formula box:
      1. Insert “PGR” column from the right
      2. Add “+”
      3. Insert “PDRA” column
      4. Add “+”
      5. Insert “other” column
   5. Click OK
9. Add a column to indicate training type
   1. Note: Since we’ll be combining this table with the Open Access training table, we need the data in this table to be labelled with “RDM.”
   2. Add another “Custom column”
   3. Name the column “Training type”
   4. In the formula box, type "RDM"
10. Switch to “Open access” and clean that table as well
    1. Remove first 5 columns
    2. Use correct column headers
    3. In “Len” column, use “replace values” to replace “ hours” and “ hour” with nothing
    4. In “Len”, use Format  Trim to remove any remaining whitespace
    5. Change “Len” data type to “Decimal number”
    6. Rename column from “Len” to “Length (hours)”
    7. Delete empty 4th column
    8. Rename final column to “Cancelled”
    9. Add a custom column called “Training type” with the value “Open access”
11. Combine the two tables into one with “append”
    1. Under “Home”, click “Append queries”
    2. Select “Append queries as new”
    3. Select “RDM” as first table and “Open access” as second table
    4. Rename Query from “Append” to “2016 Trainings”
    5. Optional: remove the extra attendance columns
    6. Close and load data back into Excel

## Example 3: Merging data from multiple sources

Source: Simulated survey data based on the Duke University Libraries’ [biennial user satisfaction survey](https://library.duke.edu/about/depts/assessment-user-experience/biennial-surveys)

1. Clean responses data (try on your own)
   1. New blank workbook
   2. Load survey responses data (“Text/CSV”)
   3. Use first row as headers
   4. Split Q3 column (multi-select) on comma
   5. Trim whitespace for the resulting columns
   6. Unpivot all questions columns (Q1 to Q4)
   7. Rename columns:
      1. “Attribute” to “QuestionID”
      2. “Value” to “Response”
   8. In QuestionID column, replace values to clean up Q3 split column headers
      1. Hint: look at column filter to see all values in column
2. Create new query to bring in question details
   1. In “Home”, click “Get data”
   2. Select “Text/CSV”
   3. Select question details file
   4. Preview data and click “Create”
   5. Use first row as headers
3. Blend the tables with “Merge queries”
   1. Under “Merge queries”, select “Merge queries as new”
      1. A “merge” is like a database “join”
   2. For the “left” table, use survey responses
   3. In the data preview, select the QuestionID column
   4. For the “right” table, use question details
   5. In the data preview, select the QuestionID column again
   6. Leave the “join kind” as “Left outer”
   7. Click OK
   8. Rename new query as “responses-questions”
4. Decide what columns to bring in
   1. Scroll right to “survey-question-details” column
   2. Click on “expand” icon at the top
   3. Select columns to include, or leave all selected
   4. Can uncheck “QuestionID”, since it’s a duplicate
   5. Click OK
   6. New columns will be added to the right
5. Optional: repeat merge process for participant details
   1. Create new query to bring in participant details
   2. Use first row as headers
   3. Merge queries as new
   4. Use “responses-questions” as left table, select ResponseID
   5. Use participant details as right, select ResponseID
   6. Click OK
   7. Rename query as “fully-merged”
   8. Expand column at the end, include only demographics columns
   9. Close & Load