DUKE UNIVERSITY LIBRARIES

Powering up your data work with Excel's Power Query

ANGELA ZOSS

github.com/amzoss/power-query

LEARNING OUTCOMES

- Articulate the benefits of reproducible data work
- Find Power Query in modern versions of Excel
- Connect to a data file in Power Query
- Understand different types of Power Query data transformations
- Build a query with a multi-step data transformation sequence
- Push transformed data back to Excel for summaries and visualizations

github.com/amzoss/power-query

Schedule

Duration	Topic
10 minutes	Small group introductions
10 minutes	Brief presentation: Introduction to reproducibility and its connection to assessment work
15 minutes	Example data project: Aggregating event registration data from multiple files
5 minutes	Quick check-in with small groups
20 minutes	Example data project: Blend item metadata with circulation data through joins
10 minutes	Small group check-in, short break
20 minutes	Example data project: Transforming simple survey data (split multi-valued cells, unpivot)
15 minutes	Small group reflection
15 minutes	Full group Question & Answer session

Technical Requirements

To follow along with exercises, participants will need access to a laptop computer (Windows or Mac OS) with a recent version of Excel that supports Power Query

- Windows: Excel 2016 and later, Microsoft 365, Excel for the Web
- Mac: Excel for Microsoft 365 for Mac, Excel for the Web

Small-group introductions

10 minutes

Introductions

- Name
- Institution
- Position title
- Have you used Excel’s Power Query before?
- What are your biggest data analysis problems right now?

Reproducible data analysis

10 minutes

What is reproducibility?

"the ability to recompute data analytic results given an observed dataset and knowledge of the data analysis pipeline."

Leek, J. T., & Peng, R. D. (2015). Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, *112*(6), 1645–1646.

Why care about reproducibility?

- Open science makes review easier
- Increasingly a requirement by funding agencies
- Saves you a lot of time trying to figure out what you did last time!

*"Your closest collaborator is **you** six months ago, but you don't reply to emails."*

- Karl Broman, paraphrasing
Karen Cranston quoting Mark T. Holder

General tips for reproducible data work

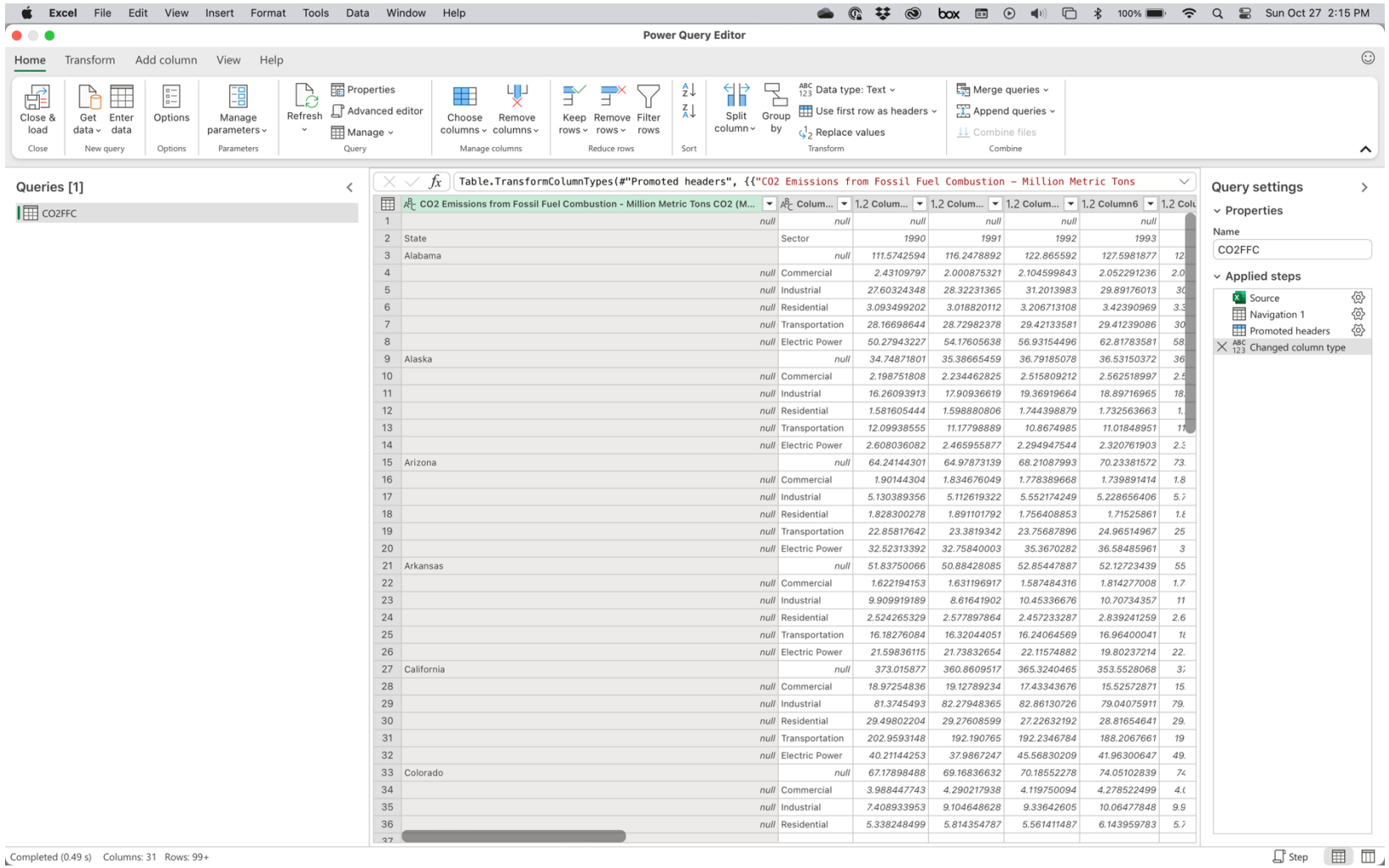
- Keep original (raw) data without modification, work from a copy
- Use non-proprietary, interoperable data formats (e.g., CSV)
- Use data processing tools that retain a record of the processing steps (e.g., programming scripts)
- Save processed data under a new file name
- Create documentation to keep track of everything you do to your data, including the sources of your data, storage locations, field properties, data cleaning steps, etc.

Reproducibility and Excel

Common ways of using Excel are not ideal for reproducibility

- Editing cells directly, with no record of the processing steps
- Injecting non-data notes and documentation into data
- Blending different data types in the same column
- Using position and formatting to indicate data values, instead of directly encoding data

Introducing: Power Query



Power Query makes Excel reproducible

- Keeps processed data separate from raw data
- Keeps track of processing steps
- Offers a wide range of data processing options, including ways to combine data from multiple sources
- After processing, loads data back into Excel for additional analysis or visualization
- Provides easy ways to refresh processed data when raw data changes

Example data project: Fix common data issues

15 minutes

CO_2 Emissions

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	**CO_2 Emissions from Fossil Fuel Combustion - Million Metric Tons CO_2 ($MMTCO_2$)***																		
2																			
3	**State**	**Sector**	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
4	**Alabama**		**111.57**	**116.25**	**122.87**	**127.60**	**125.72**	**133.45**	**139.81**	**136.71**	**134.45**	**136.86**	**143.09**	**134.24**	**139.11**	**140.10**	**143.49**	**145.11**	**147.26**
5		Commercial	2.43	2.00	2.10	2.05	2.06	1.96	2.16	2.42	1.90	2.19	2.25	2.17	1.99	2.10	2.16	1.85	2.22
6		Industrial	27.60	28.32	31.20	29.89	30.92	31.63	32.80	31.81	28.74	29.14	28.75	24.87	24.76	26.12	27.53	26.21	25.35
7		Residential	3.09	3.02	3.21	3.42	3.30	3.32	3.73	3.35	3.11	3.31	3.66	3.54	3.25	3.09	2.98	2.72	2.50
8		Transportation	28.17	28.73	29.42	29.41	30.63	32.14	31.55	31.16	31.77	32.15	33.53	31.49	33.49	32.45	34.98	34.88	35.44
9		Electric Power	50.28	54.18	56.93	62.82	58.82	64.41	69.58	67.97	68.93	70.06	74.90	72.17	75.62	76.34	75.85	79.46	81.75
10	**Alaska**		**34.75**	**35.39**	**36.79**	**36.53**	**36.37**	**41.01**	**42.03**	**42.13**	**43.19**	**43.79**	**44.57**	**44.26**	**43.83**	**43.76**	**47.04**	**48.36**	**46.01**
11		Commercial	2.20	2.23	2.52	2.56	2.59	2.50	2.71	2.56	2.67	2.83	2.69	2.49	2.07	1.94	2.21	2.12	2.42
12		Industrial	16.26	17.91	19.37	18.90	18.48	21.95	22.98	21.63	22.21	21.18	21.25	22.00	21.85	21.42	20.73	22.22	18.81
13		Residential	1.58	1.60	1.74	1.73	1.77	1.81	1.80	1.70	1.65	1.95	1.74	1.81	1.62	1.66	1.80	1.77	2.15
14		Transportation	12.10	11.18	10.87	11.02	11.19	12.34	12.01	13.49	13.75	14.82	15.72	14.73	15.03	15.75	19.14	19.02	19.17
15		Electric Power	2.61	2.47	2.29	2.32	2.33	2.41	2.53	2.74	2.91	3.01	3.16	3.24	3.26	2.97	3.16	3.23	3.47
16	**Arizona**		**64.24**	**64.98**	**68.21**	**70.23**	**73.09**	**68.31**	**69.84**	**73.05**	**78.07**	**82.11**	**87.60**	**89.52**	**89.36**	**91.21**	**98.85**	**98.83**	**101.80**
17		Commercial	1.90	1.83	1.78	1.74	1.83	1.79	1.88	1.98	2.28	2.21	2.19	2.09	2.17	2.05	2.02	2.01	2.04
18		Industrial	5.13	5.11	5.55	5.23	5.76	6.38	6.22	6.41	6.76	6.78	6.56	6.23	6.15	5.91	7.10	7.43	7.22
19		Residential	1.83	1.89	1.76	1.72	1.80	1.69	1.66	1.84	2.17	2.08	2.13	2.19	2.16	2.13	2.24	2.13	2.15
20		Transportation	22.86	23.38	23.76	24.97	25.68	26.15	27.80	27.81	29.98	31.65	32.46	33.51	34.34	35.44	36.25	36.91	38.09
21		Electric Power	32.52	32.76	35.37	36.58	38.01	32.30	32.29	35.00	36.87	39.39	44.26	45.51	44.54	45.67	51.24	50.36	52.30
22	**Arkansas**		**51.84**	**50.88**	**52.85**	**52.13**	**55.88**	**59.35**	**61.87**	**60.85**	**61.45**	**63.73**	**64.23**	**63.27**	**62.70**	**62.99**	**62.87**	**60.45**	**62.74**
23		Commercial	1.62	1.63	1.59	1.81	1.75	1.80	1.90	1.79	1.75	1.79	2.11	2.14	2.13	2.18	2.02	2.12	1.88
24		Industrial	9.91	8.62	10.45	10.71	11.34	12.39	12.03	12.43	12.22	11.88	12.08	11.98	12.24	11.87	11.08	10.50	11.01
25		Residential	2.52	2.58	2.46	2.84	2.63	2.71	2.86	2.65	2.35	2.66	2.91	2.65	2.62	2.49	2.25	2.15	2.07

Source:
https://www.epa.gov/statelocalenergy/state-co2-emissions-fossil-fuel-combustion

Remove empty, non-data rows

297		Electric Power	70.06	65.11	66.90	67.02	73.16	72.74	77.79	81.80	81.14	84.42	83.06	73.71	85.41
298	**Wisconsin**		**86.11**	**88.29**	**87.45**	**91.16**	**94.29**	**97.23**	**101.47**	**104.12**	**100.93**	**105.19**	**107.19**	**105.20**	**106.35**
299		Commercial	4.82	5.02	4.83	5.15	5.11	5.48	6.00	5.91	5.56	5.67	5.61	5.39	5.80
300		Industrial	14.32	14.59	14.52	15.33	15.81	15.86	16.07	16.61	15.44	17.01	17.80	17.00	17.26
301		Residential	9.45	10.14	9.87	10.61	10.15	10.29	11.50	10.37	8.94	10.05	10.24	9.75	10.46
302		Transportation	24.30	24.12	24.66	25.45	26.94	27.49	27.91	28.05	29.64	30.37	29.79	29.66	29.77
303		Electric Power	33.22	34.42	33.58	34.62	36.28	38.11	40.00	43.19	41.36	42.09	43.74	43.40	43.07
304	**Wyoming**		**57.84**	**56.06**	**61.77**	**58.95**	**61.17**	**58.84**	**60.29**	**59.74**	**63.36**	**61.91**	**62.72**	**62.99**	**61.71**
305		Commercial	0.86	0.93	0.75	0.97	1.01	0.94	1.30	0.93	0.94	0.91	1.00	1.00	0.94
306		Industrial	10.34	9.88	11.89	10.28	10.60	9.99	10.17	10.35	10.58	9.92	9.74	9.89	10.11
307		Residential	0.82	0.88	0.76	0.88	0.82	0.85	0.94	0.81	0.78	0.76	0.82	0.79	0.91
308		Transportation	5.77	5.30	5.57	5.96	5.85	6.77	6.78	6.89	7.09	8.09	7.61	7.77	7.71
309		Electric Power	40.05	39.08	42.80	40.86	42.89	40.27	41.09	40.76	43.97	42.23	43.55	43.53	42.04
310	*** Emission estimates are based on energy consumption data from EIA's State Energy Consumption, Price, and Expenditure Estimates (SEDS) rele**														
311	**Available online at: http://www.eia.gov/state/seds/seds-data-complete.cfm?sid=US#CompleteDataFile**														

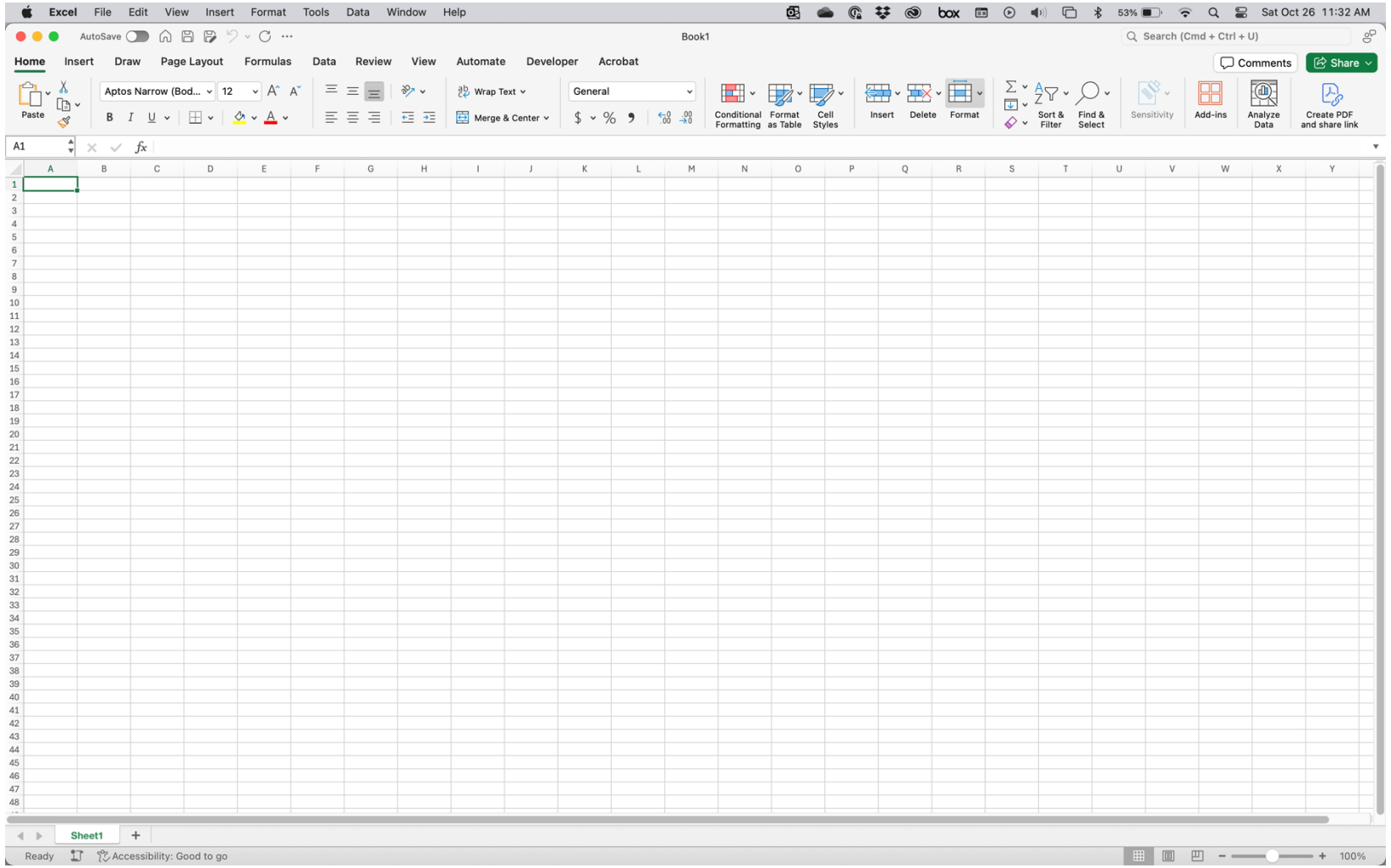
Complete incomplete rows, headers

	A	B	C	D
1	CO_2 Emissions from Fossil Fuel Combustion - M			
2				
3	**State**	**Sector**	**1990**	**1991**
4	**Alabama**		**109.05**	**113.22**
5		Commercial	2.43	2.00
6		Industrial	25.12	25.33
7		Residential	3.09	3.02
8		Transportation	28.14	28.70
9		Electric Power	50.28	54.18
10	**Alaska**		**34.25**	**34.86**
11		Commercial	2.20	2.23
12		Industrial	15.77	17.39
13		Residential	1.58	1.60
14		Transportation	12.09	11.17
15		Electric Power	2.61	2.46
16	**Arizona**		**62.94**	**63.75**
17		Commercial	1.90	1.83
18		Industrial	3.86	3.91
19		Residential	1.83	1.89
20		Transportation	22.83	23.36
21		Electric Power	32.52	32.76
22	**Arkansas**		**51.21**	**50.28**
23		Commercial	1.62	1.63
24		Industrial	9.32	8.04
25		Residential	2.52	2.58
26		Transportation	16.16	16.30
27		Electric Power	21.60	21.74
28	**California**		**362.43**	**350.77**
29		Commercial	18.96	19.11
30		Industrial	71.06	72.44
31		Residential	29.47	29.25

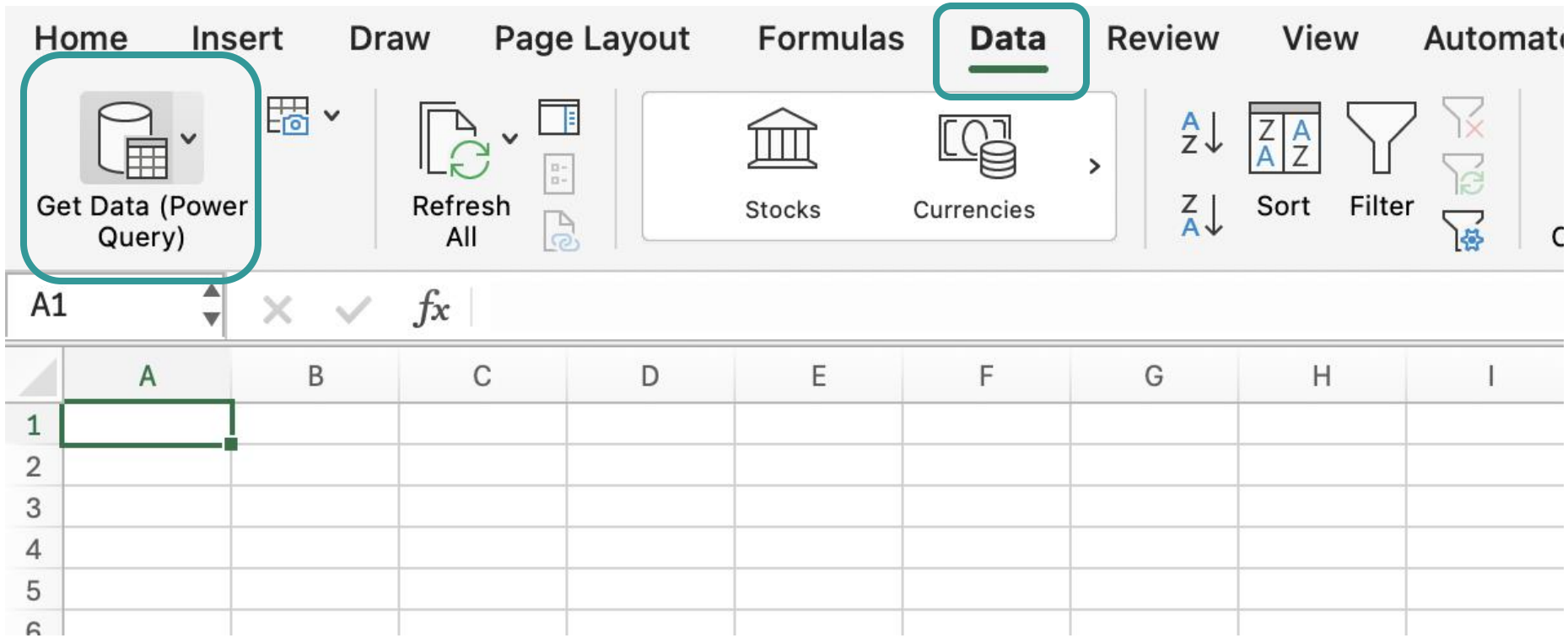
Remove sub-totals and totals

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CO_2 Emissions from Fossil Fuel Combustion - Million Metric Tons CO_2 ($MMTCO_2$)*													
2														
3	State	Sector	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
4	Alabama		109.05	113.22	119.96	123.87	121.93	129.64	135.66	133.57	131.24	133.42	139.43	131.07
5		Commercial	2.43	2.00	2.10	2.05	2.05	1.96	2.16	2.42	1.90	2.19	2.25	2.16
6		Industrial	25.12	25.33	28.33	26.15	27.10	27.80	28.65	28.67	25.52	25.70	25.07	21.69
7		Residential	3.09	3.02	3.20	3.42	3.30	3.31	3.73	3.34	3.11	3.31	3.66	3.53
8		Transportation	28.14	28.70	29.39	29.44	30.66	32.16	31.54	31.17	31.78	32.16	33.55	31.52
9		Electric Power	50.28	54.18	56.93	62.82	58.82	64.41	69.58	67.97	68.93	70.06	74.90	72.17
10	Alaska		34.25	34.86	36.24	36.08	35.91	40.45	41.51	41.59	42.61	43.18	43.85	42.89
11		Commercial	2.20	2.23	2.51	2.56	2.59	2.50	2.71	2.56	2.67	2.82	2.69	2.49
12		Industrial	15.77	17.39	18.83	18.45	18.03	21.39	22.46	21.11	21.63	20.58	20.54	20.63
13		Residential	1.58	1.60	1.74	1.73	1.77	1.81	1.80	1.70	1.65	1.95	1.74	1.80
14		Transportation	12.09	11.17	10.86	11.02	11.19	12.34	12.01	13.49	13.75	14.82	15.72	14.73
15		Electric Power	2.61	2.46	2.29	2.32	2.33	2.41	2.53	2.74	2.91	3.01	3.16	3.24
16	Arizona		62.94	63.75	66.55	68.96	71.69	66.62	68.48	71.61	76.01	80.13	85.82	88.17
17		Commercial	1.90	1.83	1.78	1.74	1.83	1.79	1.88	1.98	2.28	2.20	2.19	2.09
18		Industrial	3.86	3.91	3.92	3.94	4.34	4.67	4.85	4.96	4.69	4.79	4.76	4.86
19		Residential	1.83	1.89	1.75	1.71	1.80	1.69	1.65	1.84	2.17	2.08	2.13	2.19
20		Transportation	22.83	23.36	23.74	24.99	25.71	26.18	27.80	27.83	30.01	31.67	32.48	33.54
21		Electric Power	32.52	32.76	35.37	36.58	38.01	32.30	32.29	35.00	36.87	39.39	44.25	45.50
22	Arkansas		51.21	50.28	51.91	51.05	54.99	58.32	60.96	59.94	60.63	62.73	63.10	62.27
23		Commercial	1.62	1.63	1.59	1.81	1.75	1.80	1.90	1.79	1.75	1.79	2.11	2.14
24		Industrial	9.32	8.04	9.54	9.63	10.44	11.37	11.14	11.54	11.41	10.89	10.96	10.97
25		Residential	2.52	2.58	2.45	2.84	2.63	2.71	2.86	2.65	2.35	2.66	2.91	2.65
26		Transportation	16.16	16.30	16.22	16.97	18.13	18.49	18.85	19.43	19.65	20.88	20.97	19.81
27		Electric Power	21.60	21.74	22.11	19.80	22.04	23.96	26.21	24.54	25.47	26.52	26.16	26.70
28	California		362.43	350.77	355.05	344.64	361.77	350.88	351.72	354.73	364.99	368.15	383.20	387.36
29		Commercial	18.96	19.11	17.42	15.51	15.91	17.00	14.73	15.33	17.75	14.90	14.37	14.84
30		Industrial	71.06	72.44	72.83	70.04	70.41	69.22	72.38	77.15	75.01	73.16	72.58	74.27
31		Residential	29.47	29.25	27.20	28.79	29.35	26.74	26.94	26.73	32.25	31.92	27.53	28.63

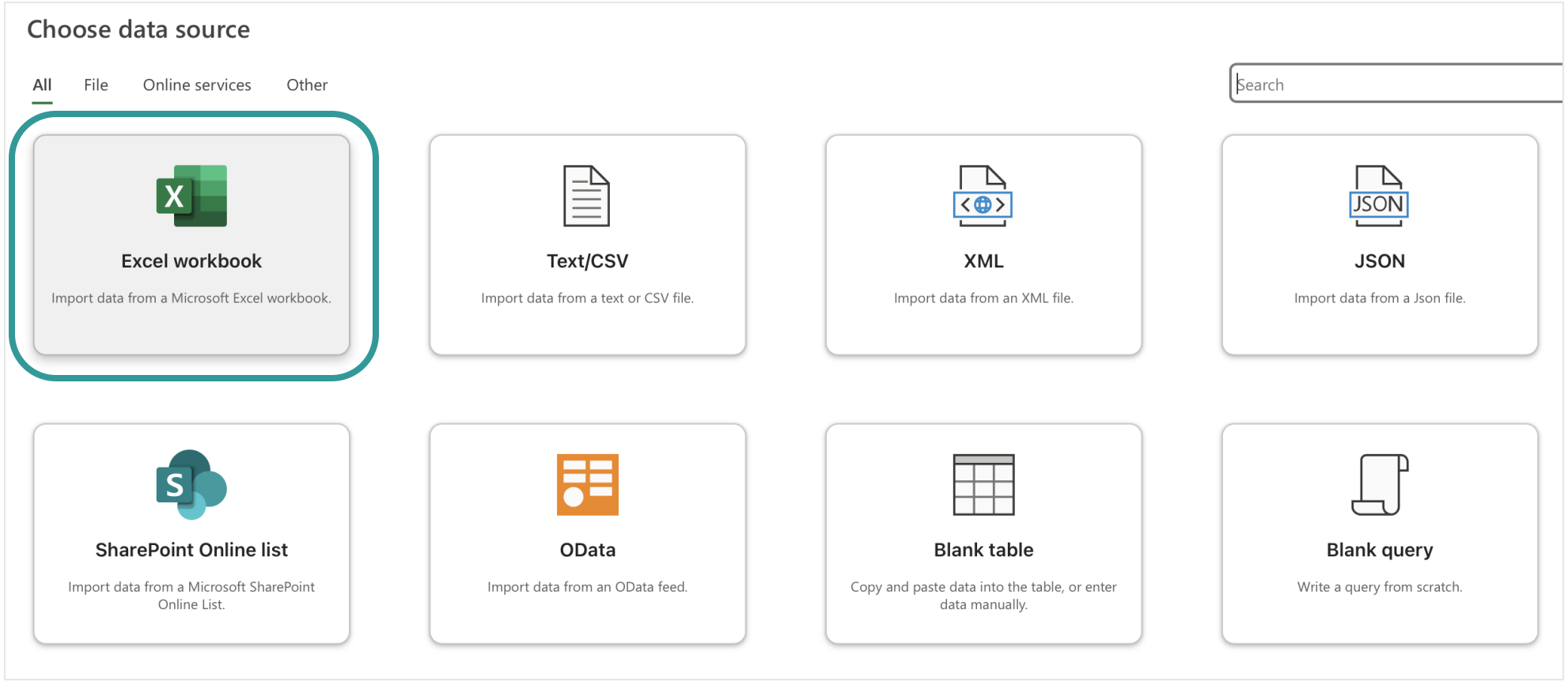
Open a blank workbook



Under “Data”, click “Get Data”

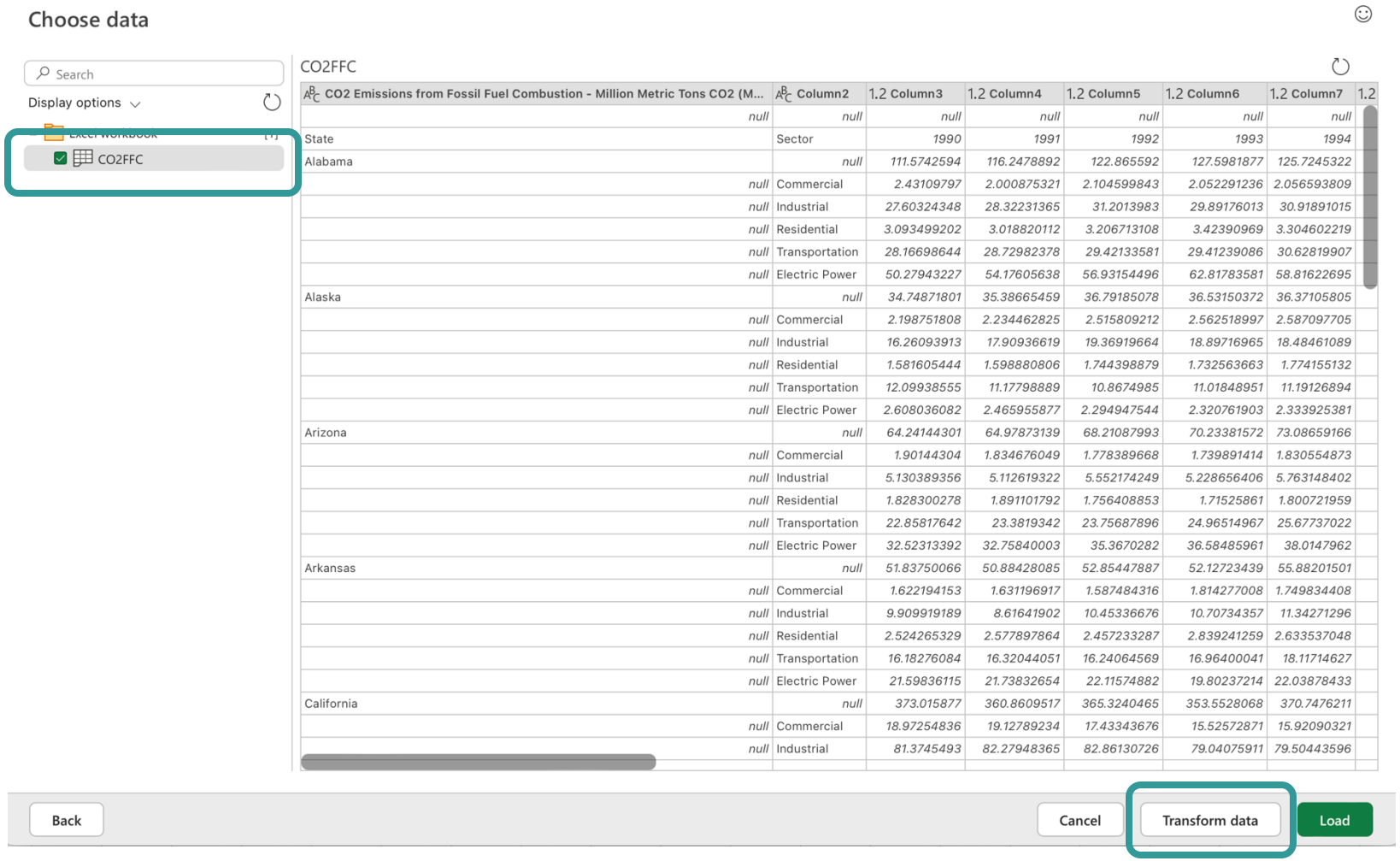


Choose data source: Excel workbook



Select “state_co2…xlsx” file

Select the “CO2FFC” sheet, check the preview, and click “Transform data”



CO2 Emissions from Fossil Fuel Combustion - Million Metric Tons CO2 (M...	Column2	1.2 Column3	1.2 Column4	1.2 Column5	1.2 Column6	1.2 Column7	1.2
null	null	null	null	null	null	null	
State	Sector	1990	1991	1992	1993	1994	
Alabama	null	111.5742594	116.2478892	122.865592	127.5981877	125.7245322	
null	Commercial	2.43109797	2.000875321	2.104599843	2.052291236	2.056593809	
null	Industrial	27.60324348	28.32231365	31.2013983	29.89176013	30.91891015	
null	Residential	3.093499202	3.018820112	3.206713108	3.42390969	3.304602219	
null	Transportation	28.16698644	28.72982378	29.42133581	29.41239086	30.62819907	
null	Electric Power	50.27943227	54.17605638	56.93154496	62.81783581	58.81622695	
Alaska	null	34.74871801	35.38665459	36.79185078	36.53150372	36.37105805	
null	Commercial	2.198751808	2.234462825	2.515809212	2.562518997	2.587097705	
null	Industrial	16.26093913	17.90936619	19.36919664	18.89716965	18.48461089	
null	Residential	1.581605444	1.598880806	1.744398879	1.732563663	1.774155132	
null	Transportation	12.09938555	11.17798889	10.8674985	11.01848951	11.19126894	
null	Electric Power	2.608036082	2.465955877	2.294947544	2.320761903	2.333925381	
Arizona	null	64.24144301	64.97873139	68.21087993	70.23381572	73.08659166	
null	Commercial	1.90144304	1.834676049	1.778389668	1.739891414	1.830554873	
null	Industrial	5.130389356	5.112619322	5.552174249	5.228656406	5.763148402	
null	Residential	1.828300278	1.891101792	1.756408853	1.71525861	1.800721959	
null	Transportation	22.85817642	23.3819342	23.75687896	24.96514967	25.67737022	
null	Electric Power	32.52313392	32.75840003	35.3670282	36.58485961	38.0147962	
Arkansas	null	51.83750066	50.88428085	52.85447887	52.12723439	55.88201501	
null	Commercial	1.622194153	1.631196917	1.587484316	1.814277008	1.749834408	
null	Industrial	9.909919189	8.61641902	10.45336676	10.70734357	11.34271296	
null	Residential	2.524265329	2.577897864	2.457233287	2.839241259	2.633537048	
null	Transportation	16.18276084	16.32044051	16.24064569	16.96400041	18.11714627	
null	Electric Power	21.59836115	21.73832654	22.11574882	19.80237214	22.03878433	
California	null	373.015877	360.8609517	365.3240465	353.5528068	370.7476211	
null	Commercial	18.97254836	19.12789234	17.43343676	15.52572871	15.92090321	
null	Industrial	81.3745493	82.27948365	82.86130726	79.04075911	79.50443596	

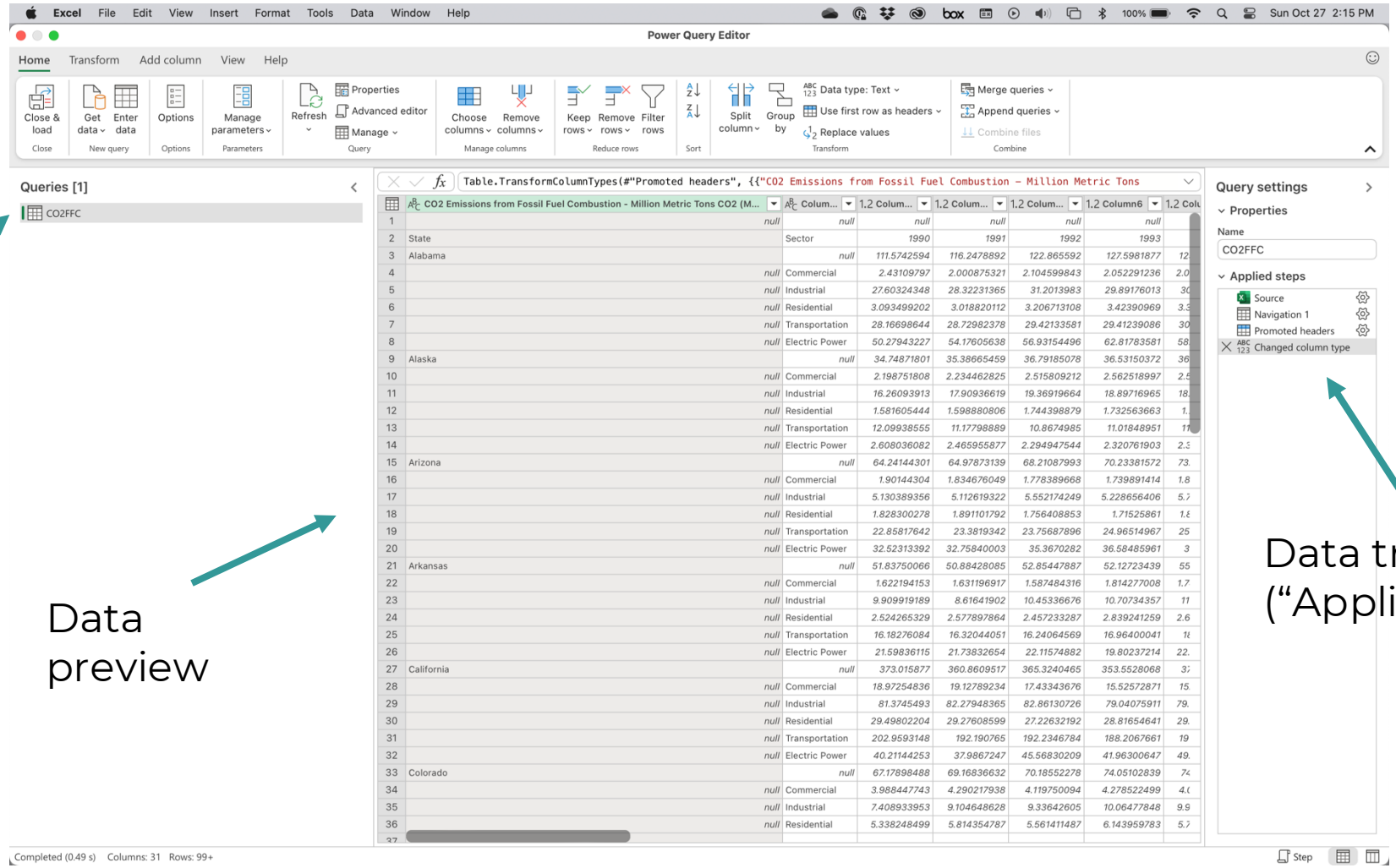
Power Query Editor opens in new window

Ribbon

Query we're working on

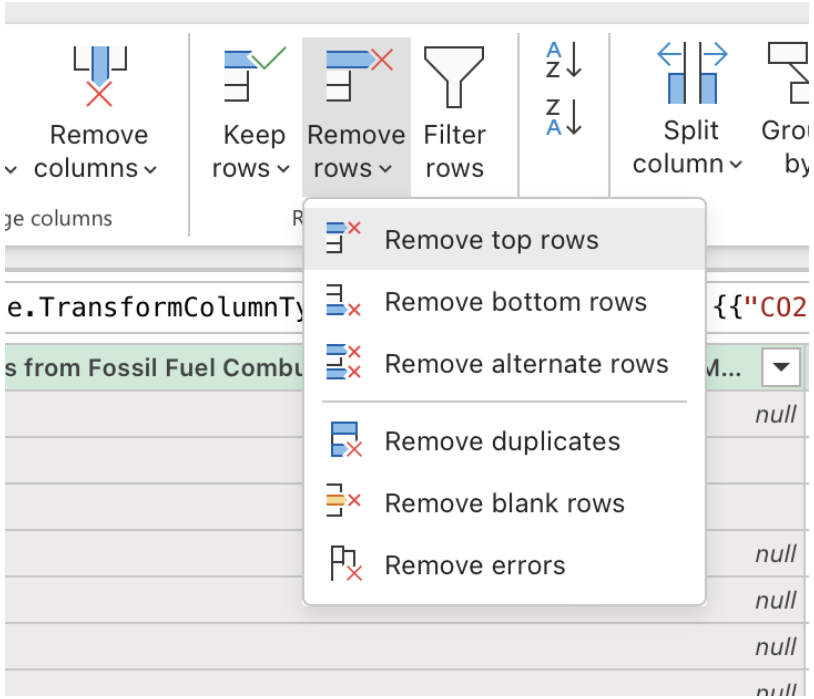
Data preview

Data transformation ("Applied steps") list

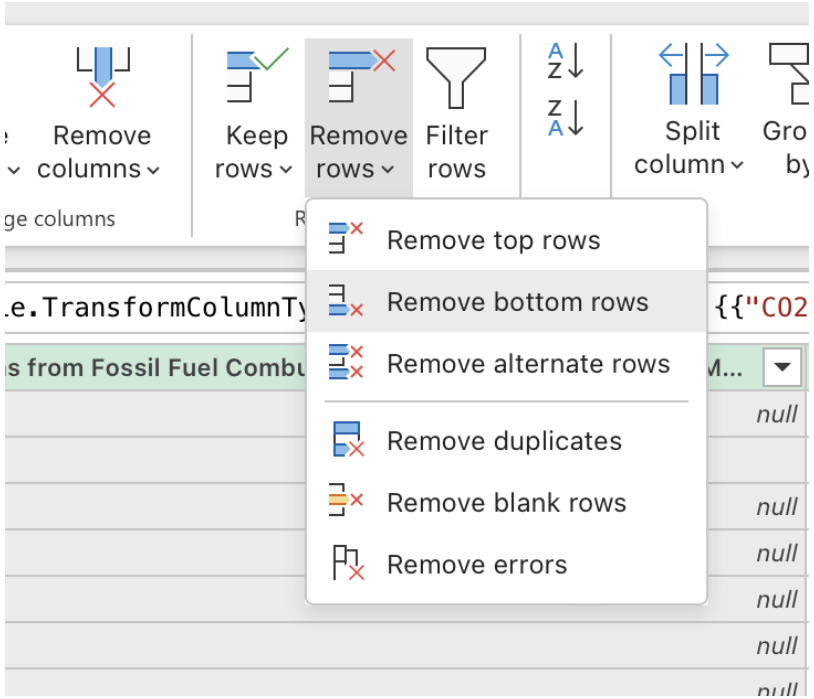


Remove non-data rows

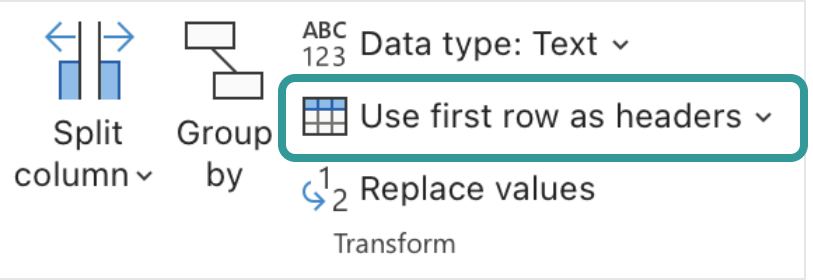
Remove top 1 row



Remove bottom 2 rows



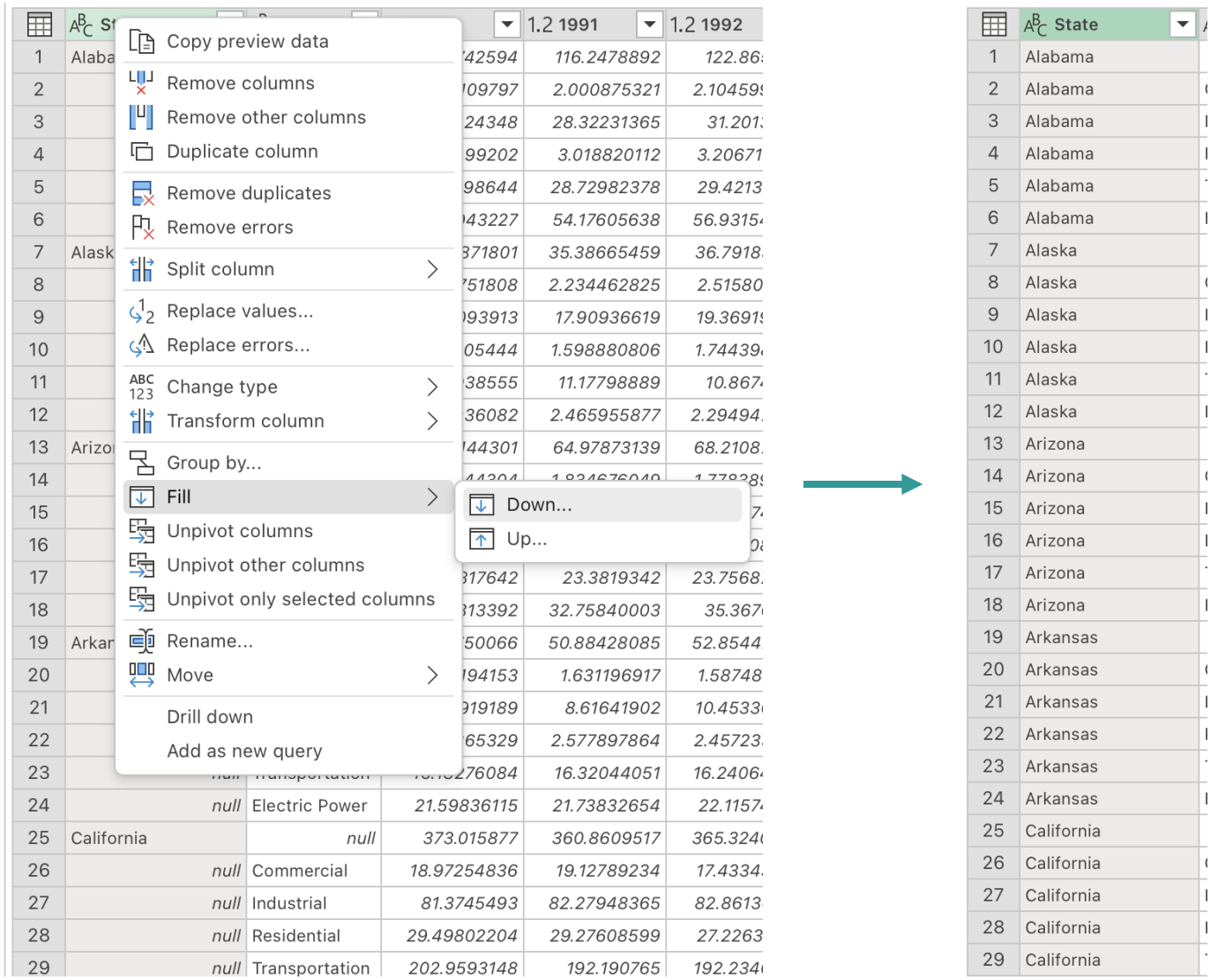
Use first row as column headers



	State	Sector	1.2 1990	1.2 1991	1.2 1992	1.2 1993	1.2 1994	1.2 1995	1.2 1996	1.2 1997	1.2 1998
1	Alabama	null	111.5742594	116.2478892	122.865592	127.5981877	125.7245322	133.4535393	139.8129389	136.710359	134.4
2	null	Commercial	2.43109797	2.000875321	2.104599843	2.052291236	2.056593809	1.964253397	2.163021568	2.419249089	1.904
3	null	Industrial	27.60324348	28.32231365	31.2013983	29.89176013	30.91891015	31.63089013	32.7968968	31.81206077	28.73
4	null	Residential	3.093499202	3.018820112	3.206713108	3.42390969	3.304602219	3.315088449	3.729836425	3.345841352	3.110
5	null	Transportation	28.16698644	28.72982378	29.42133581	29.41239086	30.62819907	32.13543484	31.54528487	31.16329022	31.76
6	null	Electric Power	50.27943227	54.17605638	56.93154496	62.81783581	58.81622695	64.40787249	69.57789919	67.96991755	68.93
7	Alaska	null	34.74871801	35.38665459	36.79185078	36.53150372	36.37105805	41.01413401	42.0320357	42.12872408	43.19
8	null	Commercial	2.198751808	2.234462825	2.515809212	2.562518997	2.587097705	2.503188907	2.711788547	2.564068617	2.668
9	null	Industrial	16.26093913	17.90936619	19.36919664	18.89716965	18.48461089	21.95067612	22.97610006	21.63241086	22.2
10	null	Residential	1.581605444	1.598880806	1.744398879	1.732563663	1.774155132	1.813549338	1.796780642	1.70389481	1.65
11	null	Transportation	12.09938555	11.17798889	10.8674985	11.01848951	11.19126894	12.33515018	12.01373146	13.49060762	13.75
12	null	Electric Power	2.608036082	2.465955877	2.294947544	2.320761903	2.333925381	2.411569465	2.533634983	2.737742181	2.909

Fill in state name for all rows

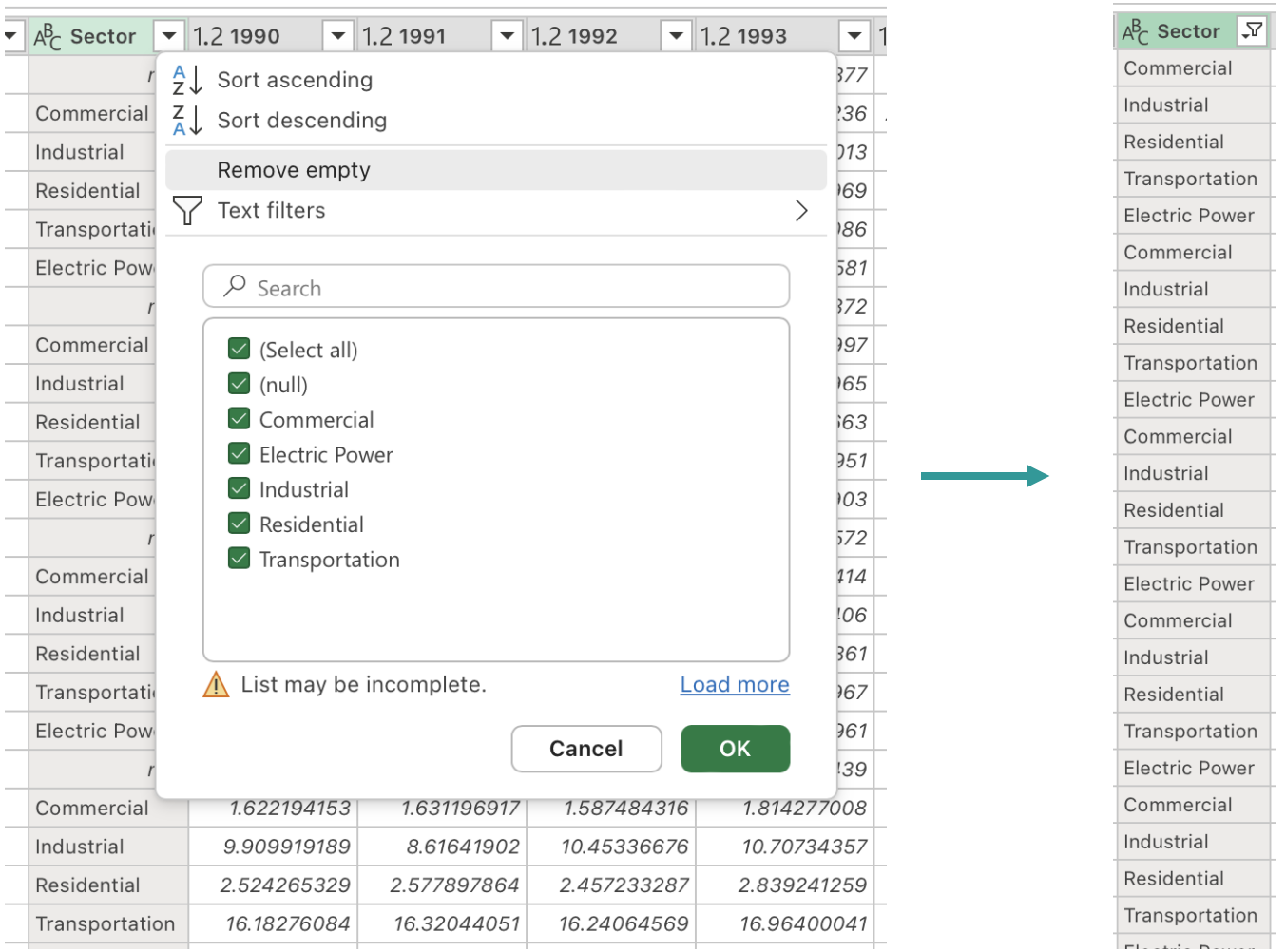
- Right-click “State” column header
- Hover over “Fill”
- Select “Down…”



Remove subtotals

Note: Subtotal rows all have "null" in the Sector column.

- Click on the filter menu for the "Sector" column
- Select "Remove empty"



What is “tidy data”?

Tidy data is a concept that helps us organize our data so that it's in a useful format for a bunch of different tools and operations.

See also: Tidy Data lecture from Hadley Wickham, circa 2011

What are the features of tidy data?

- Each variable is a column; each column is a variable.
- Each observation is a row; each row is an observation.
- Each value is a cell; each cell is a single value.

Tidy Data, from R for Data Science

Secret extra reason for tidy data: Pivot Tables

If you want to use Pivot Tables to summarize your data, you want to start with tidy data.

Wide data

(e.g., one column per year)

good for comparing two or more measurements of the same entity

State	Sector	1990	1991	1992	1993
Alabama	Commercial	2.43	2.00	2.10	2.05
Alabama	Electric Power	50.28	54.18	56.93	62.82
Alabama	Industrial	25.15	25.36	28.36	26.18
Alabama	Residential	3.09	3.02	3.20	3.42
Alabama	Transportation	28.13	28.70	29.39	29.51
Alaska	Commercial	2.20	2.23	2.51	2.56
Alaska	Electric Power	2.61	2.46	2.29	2.32
Alaska	Industrial	15.83	17.44	18.87	18.51
Alaska	Residential	1.58	1.60	1.74	1.73
Alaska	Transportation	12.09	11.17	10.86	11.03
Arizona	Commercial	1.90	1.83	1.78	1.74
Arizona	Electric Power	32.52	32.76	35.37	36.58
Arizona	Industrial	3.86	3.92	3.93	3.94
Arizona	Residential	1.83	1.89	1.75	1.71
Arizona	Transportation	22.83	23.36	23.74	25.05
Arkansas	Commercial	1.62	1.63	1.59	1.81
Arkansas	Electric Power	21.60	21.74	22.11	19.80
Arkansas	Industrial	9.35	8.07	9.56	9.66
Arkansas	Residential	2.52	2.58	2.45	2.84
Arkansas	Transportation	16.16	16.30	16.22	17.02
California	Commercial	18.96	19.11	17.42	15.51
California	Electric Power	40.18	37.95	45.53	41.93
California	Industrial	71.62	73.00	73.29	70.53

Long ("tidy") data

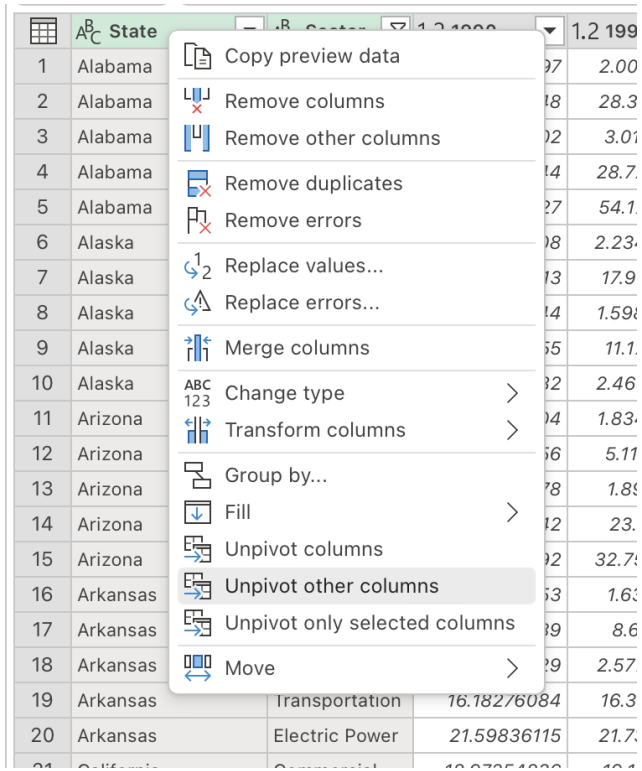
(e.g., many rows per entity)

good for filtering and subsetting on the fly, aggregating across measurements

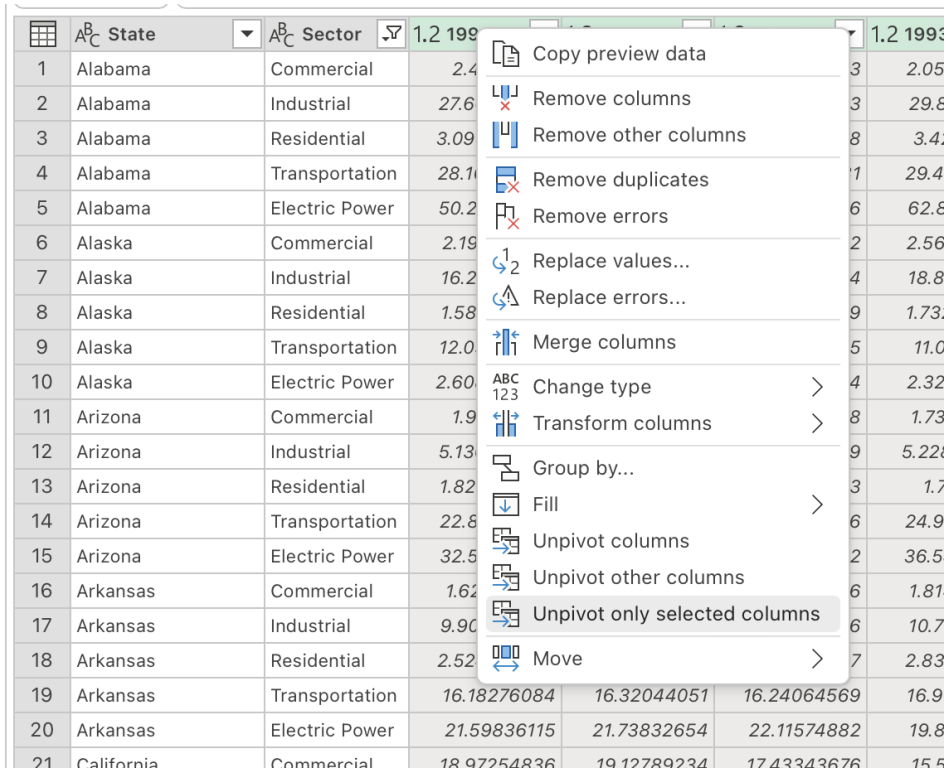
State	Sector	Year	Emissions
Alabama	Commercial	1990	2.42906
Alabama	Commercial	1991	1.999039
Alabama	Commercial	1992	2.10271
Alabama	Commercial	1993	2.05046
Alabama	Commercial	1994	2.054954
Alabama	Commercial	1995	1.962636
Alabama	Commercial	1996	2.16117
Alabama	Commercial	1997	2.417254
Alabama	Commercial	1998	1.903054
Alabama	Commercial	1999	2.187003
Alabama	Commercial	2000	2.249259
Alabama	Commercial	2001	2.163411
Alabama	Commercial	2002	1.99245
Alabama	Commercial	2003	2.089619
Alabama	Commercial	2004	2.15691
Alabama	Commercial	2005	1.845364
Alabama	Commercial	2006	2.221402
Alabama	Commercial	2007	1.987872
Alabama	Commercial	2008	2.005586
Alabama	Commercial	2009	1.911421
Alabama	Commercial	2010	2.106

In PQ, want to “Unpivot” year columns

Option 1:
Select first two columns,
unpivot other columns

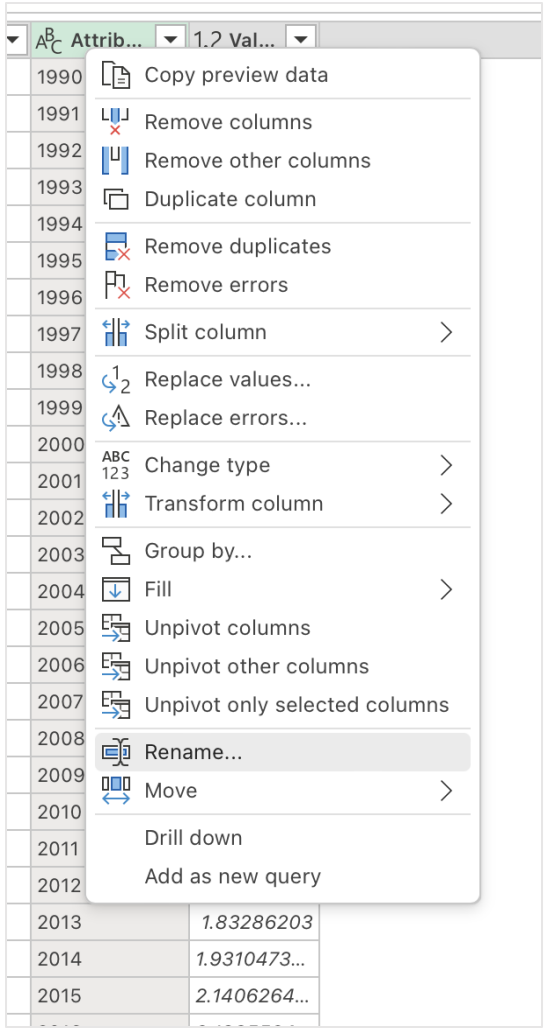


Option 2:
Select all year columns,
unpivot only selected columns

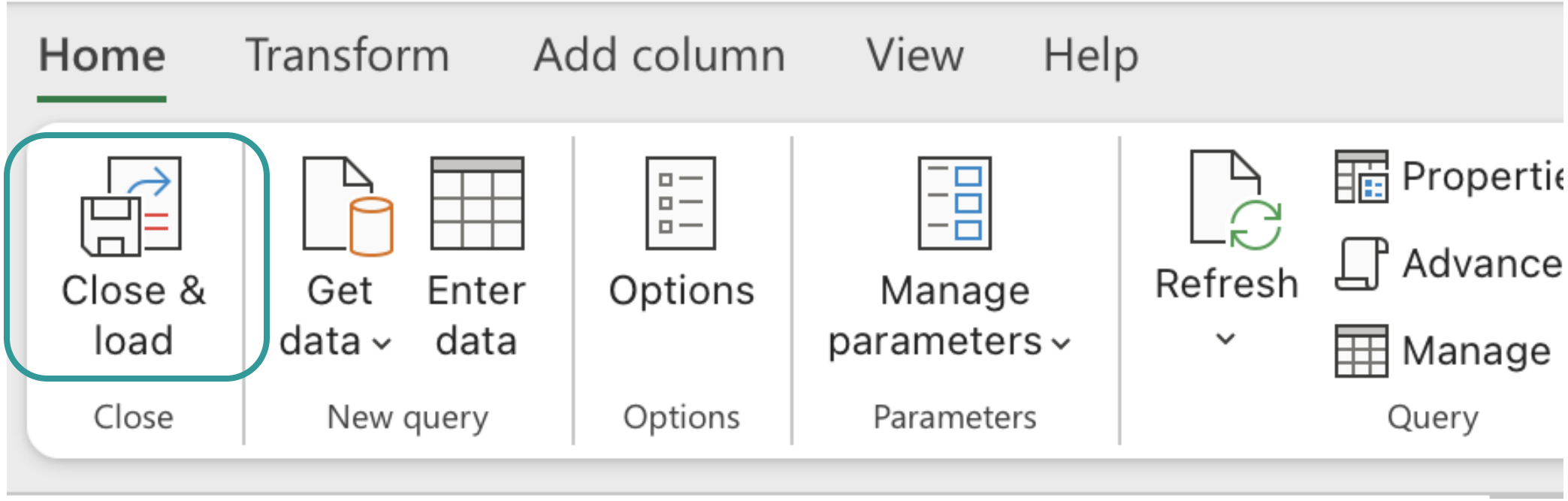


Rename new columns

- “Attribute” -> “Year”
- “Value” -> “Emissions”



Push data back into Excel with "Close and Load"



Final data in Excel

	A	B	C	D
1	State	Sector	Year	Emissions
2	Alabama	Commercial	1990	2.43109797
3	Alabama	Commercial	1991	2.000875321
4	Alabama	Commercial	1992	2.104599843
5	Alabama	Commercial	1993	2.052291236
6	Alabama	Commercial	1994	2.056593809
7	Alabama	Commercial	1995	1.964253397
8	Alabama	Commercial	1996	2.163021568
9	Alabama	Commercial	1997	2.419249089
10	Alabama	Commercial	1998	1.904669755
11	Alabama	Commercial	1999	2.188891283
12	Alabama	Commercial	2000	2.25100503
13	Alabama	Commercial	2001	2.165063571
14	Alabama	Commercial	2002	1.994078494
15	Alabama	Commercial	2003	2.104990205
16	Alabama	Commercial	2004	2.15836532
17	Alabama	Commercial	2005	1.846707982
18	Alabama	Commercial	2006	2.220878823
19	Alabama	Commercial	2007	1.985776915
20	Alabama	Commercial	2008	2.001873569
21	Alabama	Commercial	2009	1.892934184
22	Alabama	Commercial	2010	2.115442602
23	Alabama	Commercial	2011	2.051315328
24	Alabama	Commercial	2012	1.782148493
25	Alabama	Commercial	2013	1.83286203

Quick check-in with small groups

5 minutes

Example data project: Additional data cleaning, appending data sources

20 minutes

Sample training attendance data

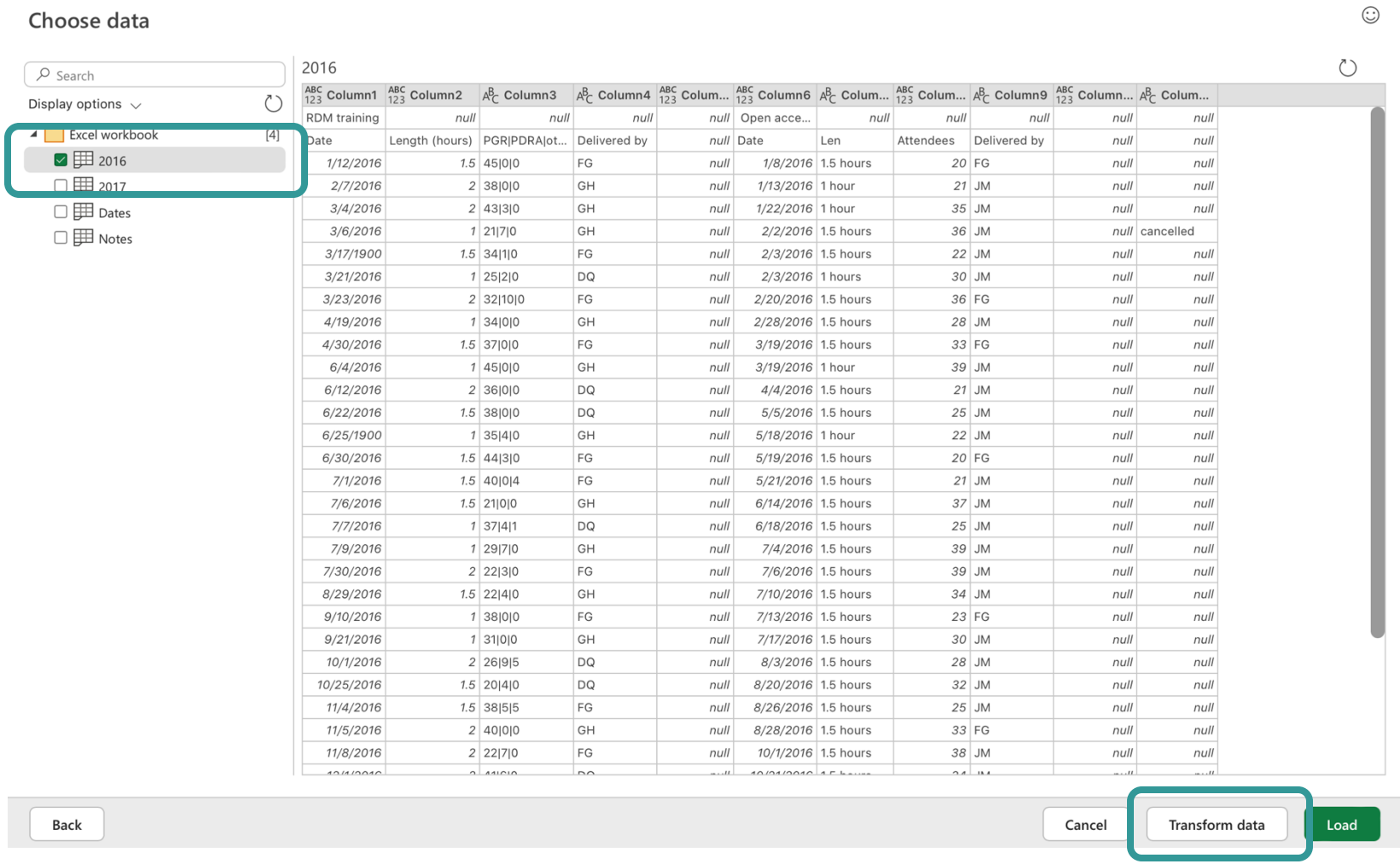
	RDM training			
Date	Length (hours)	PGR|PDRA|other	Delivered by
12 Jan	1.5	45|0|0	FG
7 Feb	2	38|0|0	GH
4 Mar	2	43|3|0	GH
6 Mar	1	21|7|0	GH
17 Mar	1.5	34|1|0	FG
21 Mar	1	25|2|0	DQ
23 Mar	2	32|10|0	FG
19 Apr	1	34|0|0	GH
30 Apr	1.5	37|0|0	FG
4 Jun	1	45|0|0	GH
12 Jun	2	36|0|0	DQ
22 Jun	1.5	38|0|0	DQ
25 Jun	1	35|4|0	GH
30 Jun	1.5	44|3|0	FG
1 Jul	1.5	40|0|4	FG
6 Jul	1.5	21|0|0	GH
7 Jul	1	37|4|1	DQ
9 Jul	1	29|7|0	GH
30 Jul	2	22|3|0	FG
29 Aug	1.5	22|4|0	GH
10 Sep	1	38|0|0	FG
21 Sep	1	31|0|0	GH
1 Oct	2	26|9|5	DQ
25 Oct	1.5	20|4|0	DQ
4 Nov	1.5	38|5|5	FG
5 Nov	2	40|0|0	GH
8 Nov	2	22|7|0	FG
1 Dec	2	41|6|0	DQ
19 Dec	2	39|9|1	GH

	Open access		
Date	Len	Attendees	Delivered by
8 Jan	1.5 hours	20	FG
13 Jan	1 hour	21	JM
22 Jan	1 hour	35	JM
2 Feb	1.5 hours	36	JM
3 Feb	1.5 hours	22	JM
3 Feb	1 hours	30	JM
20 Feb	1.5 hours	36	FG
28 Feb	1.5 hours	28	JM
19 Mar	1.5 hours	33	FG
19 Mar	1 hour	39	JM
4 Apr	1.5 hours	21	JM
5 May	1.5 hours	25	JM
18 May	1 hour	22	JM
19 May	1.5 hours	20	FG
21 May	1.5 hours	21	JM
14 Jun	1.5 hours	37	JM
18 Jun	1.5 hours	25	JM
4 Jul	1.5 hours	39	JM
6 Jul	1.5 hours	39	JM
10 Jul	1.5 hours	34	JM
13 Jul	1.5 hours	23	FG
17 Jul	1.5 hours	30	JM
3 Aug	1.5 hours	28	JM
20 Aug	1.5 hours	32	JM
26 Aug	1.5 hours	25	JM
28 Aug	1.5 hours	33	FG
1 Oct	1.5 hours	38	JM
21 Oct	1.5 hours	34	JM
9 Nov	1.5 hours	32	JM
15 Nov	1.5 hours	35	JM
15 Nov	1.5 hours	27	JM
2 Dec	1.5 hours	35	FG
7 Dec	1.5 hours	23	JM
11 Dec	1.5 hours	38	FG
19 Dec	1.5 hours	20	FG

cancelled

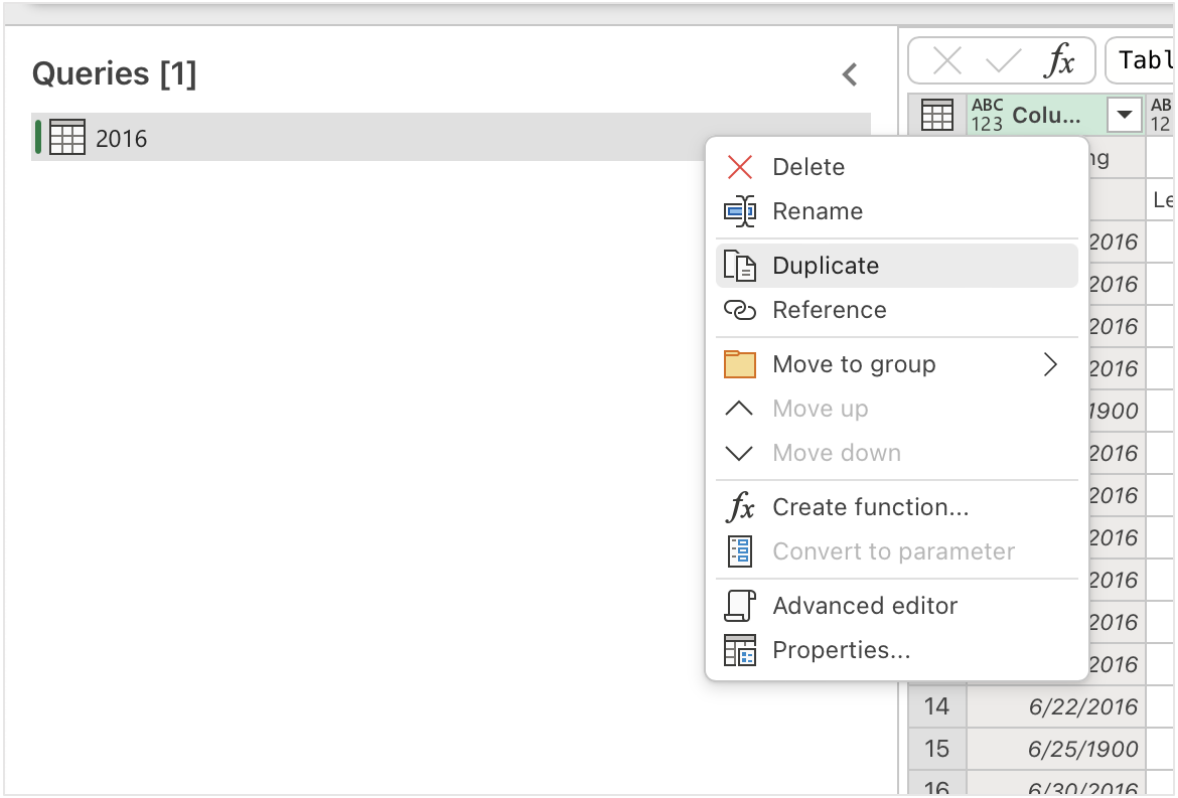
Source:
https://librarycarpentry.org/lc-spreadsheets/#data

In a new Excel file, get 2016 data from “training_attendance.xlsx”

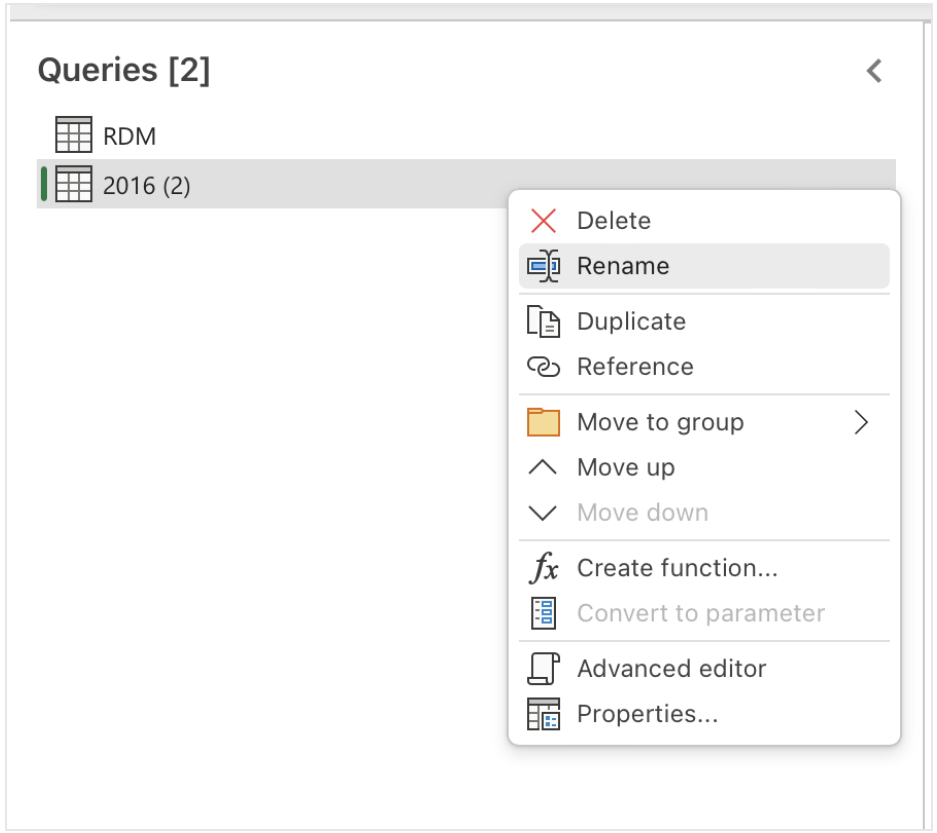


Column1	Column2	Column3	Column4	Colum...	Column6	Colum...	Colum...	Column9	Column...	Colum...
RDM training	null	null	null	null	Open acce...	null	null	null	null	null
Date	Length (hours)	PGR|PDRA|ot...	Delivered by	null	Date	Len	Attendees	Delivered by	null	null
1/12/2016	1.5	45|0|0	FG	null	1/8/2016	1.5 hours	20	FG	null	null
2/7/2016	2	38|0|0	GH	null	1/13/2016	1 hour	21	JM	null	null
3/4/2016	2	43|3|0	GH	null	1/22/2016	1 hour	35	JM	null	null
3/6/2016	1	21|7|0	GH	null	2/2/2016	1.5 hours	36	JM	null	cancelled
3/17/1900	1.5	34|1|0	FG	null	2/3/2016	1.5 hours	22	JM	null	null
3/21/2016	1	25|2|0	DQ	null	2/3/2016	1 hours	30	JM	null	null
3/23/2016	2	32|10|0	FG	null	2/20/2016	1.5 hours	36	FG	null	null
4/19/2016	1	34|0|0	GH	null	2/28/2016	1.5 hours	28	JM	null	null
4/30/2016	1.5	37|0|0	FG	null	3/19/2016	1.5 hours	33	FG	null	null
6/4/2016	1	45|0|0	GH	null	3/19/2016	1 hour	39	JM	null	null
6/12/2016	2	36|0|0	DQ	null	4/4/2016	1.5 hours	21	JM	null	null
6/22/2016	1.5	38|0|0	DQ	null	5/5/2016	1.5 hours	25	JM	null	null
6/25/1900	1	35|4|0	GH	null	5/18/2016	1 hour	22	JM	null	null
6/30/2016	1.5	44|3|0	FG	null	5/19/2016	1.5 hours	20	FG	null	null
7/1/2016	1.5	40|0|4	FG	null	5/21/2016	1.5 hours	21	JM	null	null
7/6/2016	1.5	21|0|0	GH	null	6/14/2016	1.5 hours	37	JM	null	null
7/7/2016	1	37|4|1	DQ	null	6/18/2016	1.5 hours	25	JM	null	null
7/9/2016	1	29|7|0	GH	null	7/4/2016	1.5 hours	39	JM	null	null
7/30/2016	2	22|3|0	FG	null	7/6/2016	1.5 hours	39	JM	null	null
8/29/2016	1.5	22|4|0	GH	null	7/10/2016	1.5 hours	34	JM	null	null
9/10/2016	1	38|0|0	FG	null	7/13/2016	1.5 hours	23	FG	null	null
9/21/2016	1	31|0|0	GH	null	7/17/2016	1.5 hours	30	JM	null	null
10/1/2016	2	26|9|5	DQ	null	8/3/2016	1.5 hours	28	JM	null	null
10/25/2016	1.5	20|4|0	DQ	null	8/20/2016	1.5 hours	32	JM	null	null
11/4/2016	1.5	38|5|5	FG	null	8/26/2016	1.5 hours	25	JM	null	null
11/5/2016	2	40|0|0	GH	null	8/28/2016	1.5 hours	33	FG	null	null
11/8/2016	2	22|7|0	FG	null	10/1/2016	1.5 hours	38	JM	null	null

Sheet has two data tables; we can duplicate the query to deal with each

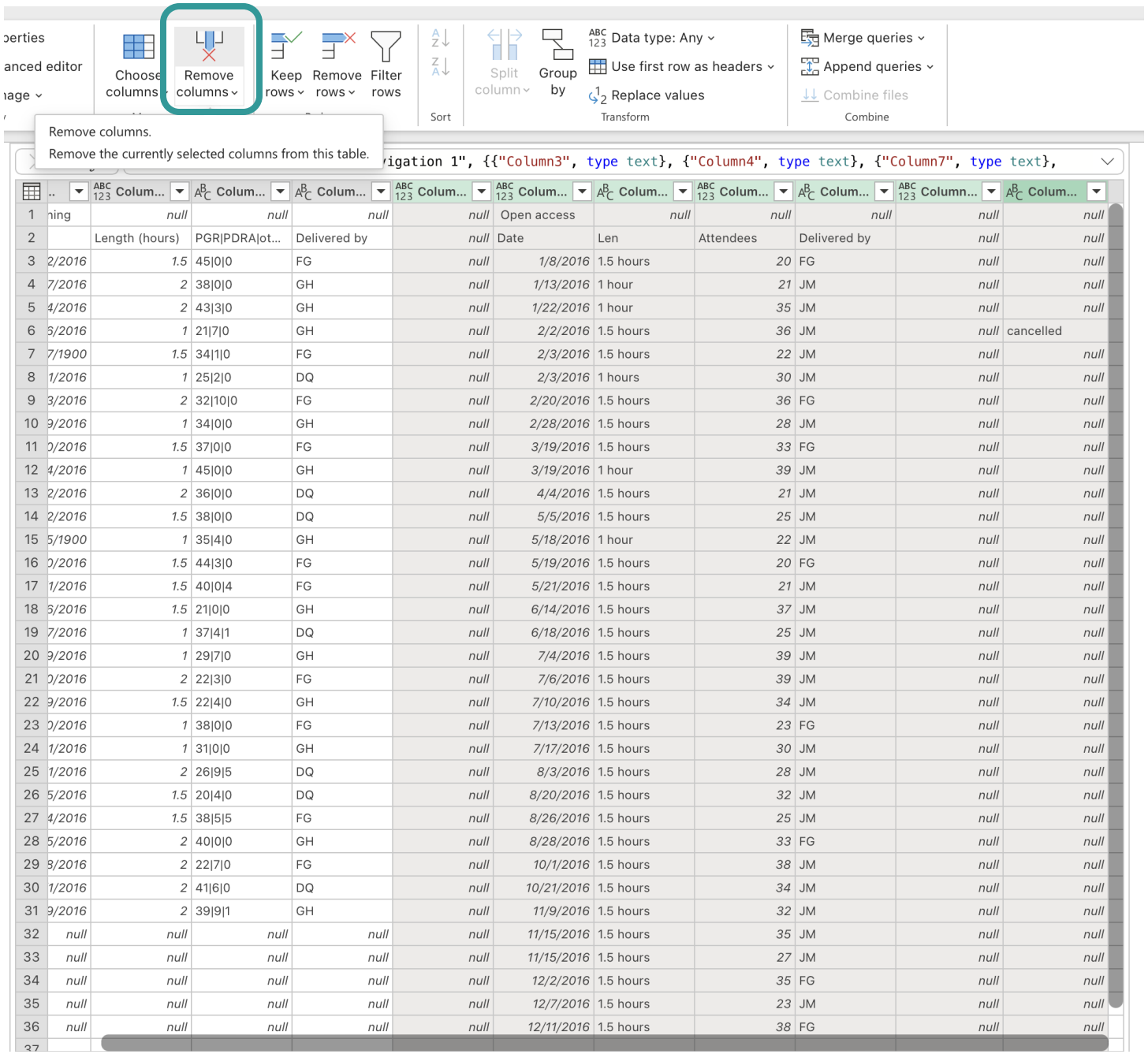


Rename each query to match data tables: RDM and Open access



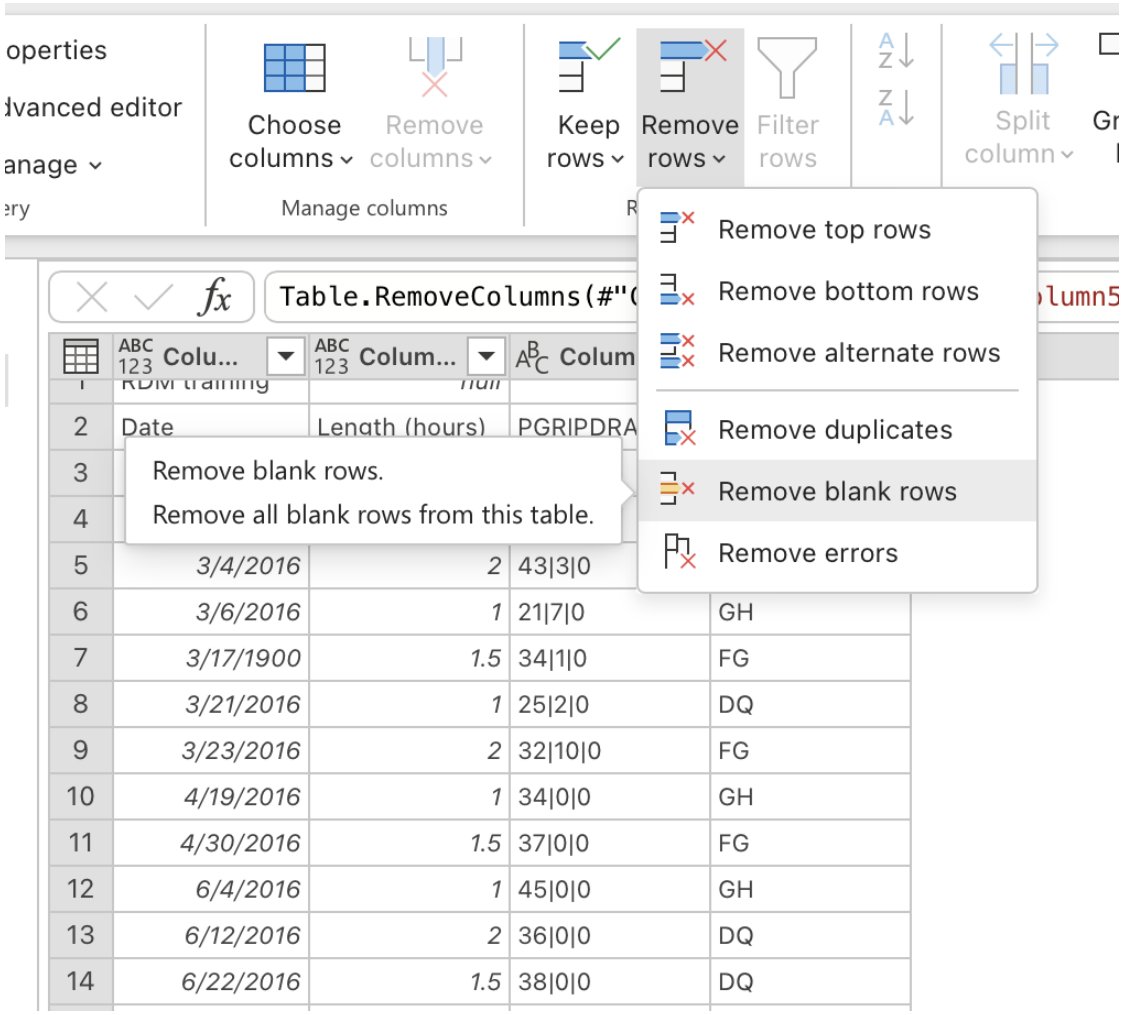
In RDM query, keep only the first 4 columns

- Select remaining columns
- Click “Remove columns”



Remove blank rows

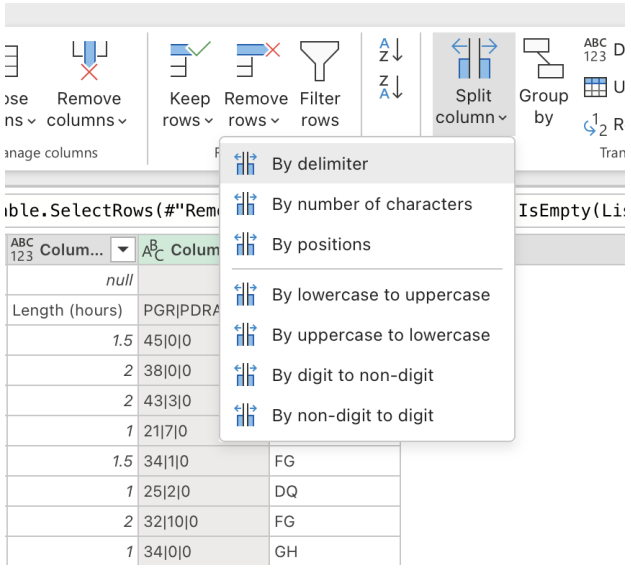
In “Remove rows” menu, select “Remove blank rows”

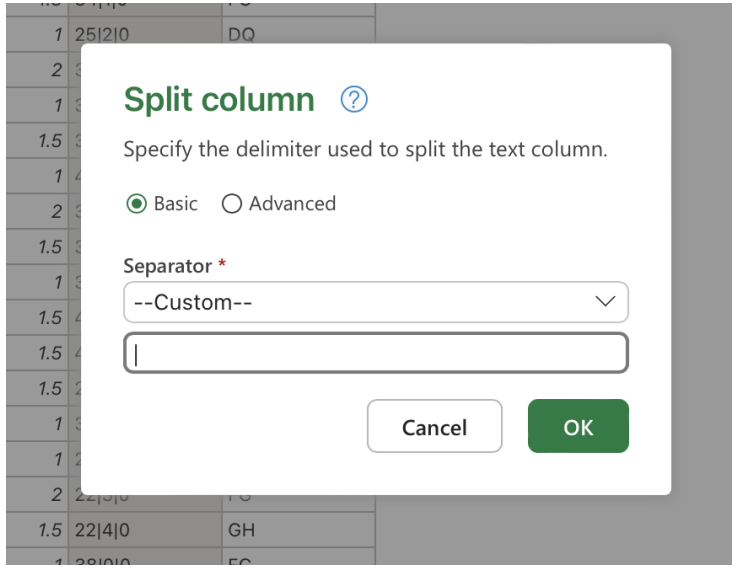


Split the third column

Column contains number of attendees of different types (post-graduate researcher (PGR), post-doctoral research associate (PDRA), and other)

- Select "PGR|PDRA|other" column
- Click "Split column"
- Select "By delimiter"
- For Separator, select "--Custom--"
- Type "|" as the separator

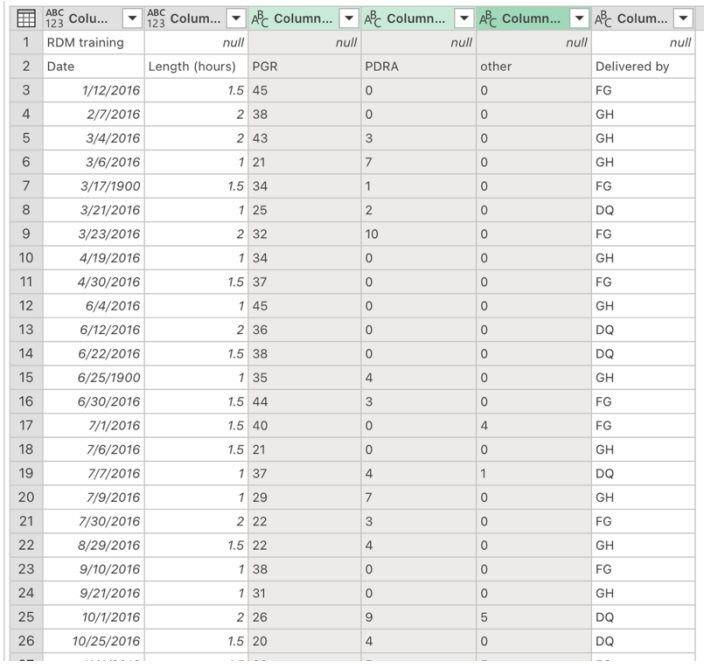




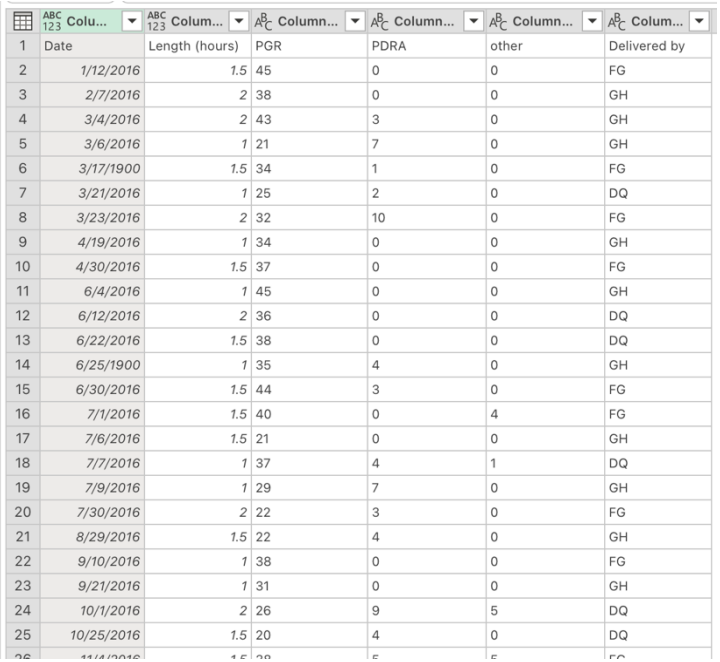
	Colu...	Colum...	Column...	Column...	Column...	Colum...
1	RDM training	*null*	*null*	*null*	*null*	*null*
2	Date	Length (hours)	PGR	PDRA	other	Delivered by
3	*1/12/2016*	*1.5*	45	0	0	FG
4	*2/7/2016*	*2*	38	0	0	GH
5	*3/4/2016*	*2*	43	3	0	GH
6	*3/6/2016*	*1*	21	7	0	GH
7	*3/17/1900*	*1.5*	34	1	0	FG
8	*3/21/2016*	*1*	25	2	0	DQ
9	*3/23/2016*	*2*	32	10	0	FG
10	*4/19/2016*	*1*	34	0	0	GH

Use correct column headers

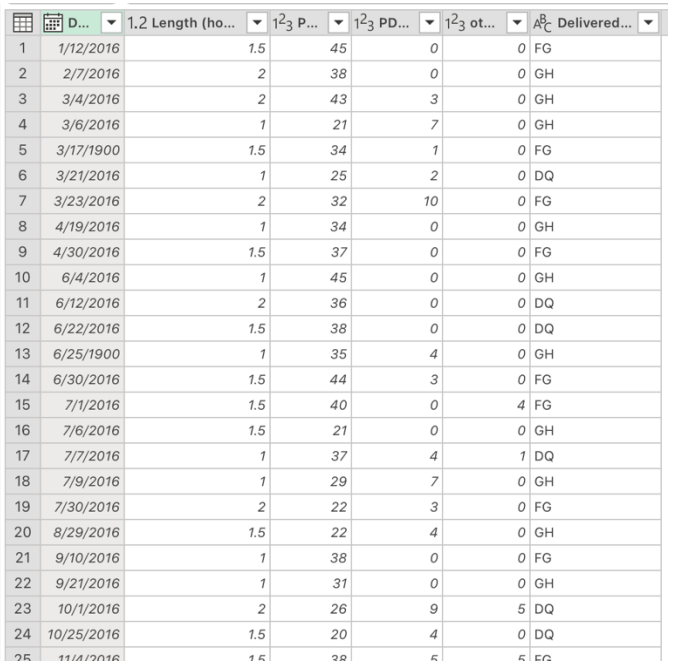
- Remove top 1 row
- Use first row as headers



	Colu...	Colum...	Column...	Column...	Column...	Colum...
1	RDM training	null	null	null	null	null
2	Date	Length (hours)	PGR	PDRA	other	Delivered by
3	1/12/2016	1.5	45	0	0	FG
4	2/7/2016	2	38	0	0	GH
5	3/4/2016	2	43	3	0	GH
6	3/6/2016	1	21	7	0	GH
7	3/17/1900	1.5	34	1	0	FG
8	3/21/2016	1	25	2	0	DQ
9	3/23/2016	2	32	10	0	FG
10	4/19/2016	1	34	0	0	GH
11	4/30/2016	1.5	37	0	0	FG
12	6/4/2016	1	45	0	0	GH
13	6/12/2016	2	36	0	0	DQ
14	6/22/2016	1.5	38	0	0	DQ
15	6/25/1900	1	35	4	0	GH
16	6/30/2016	1.5	44	3	0	FG
17	7/1/2016	1.5	40	0	4	FG
18	7/6/2016	1.5	21	0	0	GH
19	7/7/2016	1	37	4	1	DQ
20	7/9/2016	1	29	7	0	GH
21	7/30/2016	2	22	3	0	FG
22	8/29/2016	1.5	22	4	0	GH
23	9/10/2016	1	38	0	0	FG
24	9/21/2016	1	31	0	0	GH
25	10/1/2016	2	26	9	5	DQ
26	10/25/2016	1.5	20	4	0	DQ



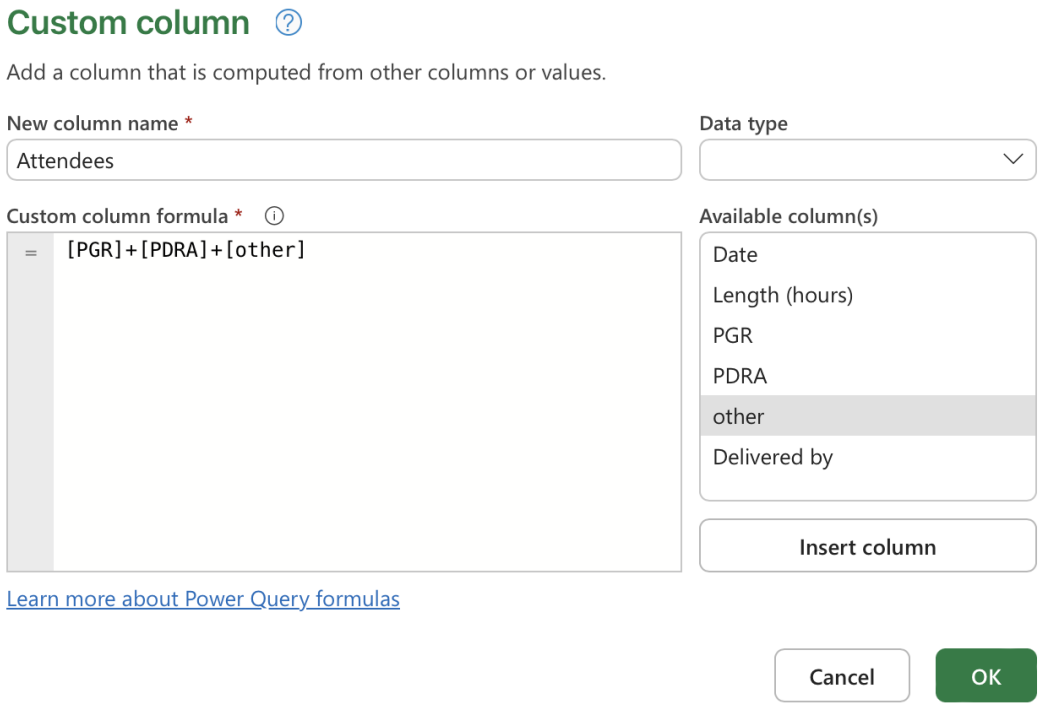
	Colu...	Colum...	Column...	Column...	Column...	Colum...
1	Date	Length (hours)	PGR	PDRA	other	Delivered by
2	1/12/2016	1.5	45	0	0	FG
3	2/7/2016	2	38	0	0	GH
4	3/4/2016	2	43	3	0	GH
5	3/6/2016	1	21	7	0	GH
6	3/17/1900	1.5	34	1	0	FG
7	3/21/2016	1	25	2	0	DQ
8	3/23/2016	2	32	10	0	FG
9	4/19/2016	1	34	0	0	GH
10	4/30/2016	1.5	37	0	0	FG
11	6/4/2016	1	45	0	0	GH
12	6/12/2016	2	36	0	0	DQ
13	6/22/2016	1.5	38	0	0	DQ
14	6/25/1900	1	35	4	0	GH
15	6/30/2016	1.5	44	3	0	FG
16	7/1/2016	1.5	40	0	4	FG
17	7/6/2016	1.5	21	0	0	GH
18	7/7/2016	1	37	4	1	DQ
19	7/9/2016	1	29	7	0	GH
20	7/30/2016	2	22	3	0	FG
21	8/29/2016	1.5	22	4	0	GH
22	9/10/2016	1	38	0	0	FG
23	9/21/2016	1	31	0	0	GH
24	10/1/2016	2	26	9	5	DQ
25	10/25/2016	1.5	20	4	0	DQ



	D...	1.2 Length (ho...	P...	PD...	ot...	Delivered...
1	1/12/2016	1.5	45	0	0	FG
2	2/7/2016	2	38	0	0	GH
3	3/4/2016	2	43	3	0	GH
4	3/6/2016	1	21	7	0	GH
5	3/17/1900	1.5	34	1	0	FG
6	3/21/2016	1	25	2	0	DQ
7	3/23/2016	2	32	10	0	FG
8	4/19/2016	1	34	0	0	GH
9	4/30/2016	1.5	37	0	0	FG
10	6/4/2016	1	45	0	0	GH
11	6/12/2016	2	36	0	0	DQ
12	6/22/2016	1.5	38	0	0	DQ
13	6/25/1900	1	35	4	0	GH
14	6/30/2016	1.5	44	3	0	FG
15	7/1/2016	1.5	40	0	4	FG
16	7/6/2016	1.5	21	0	0	GH
17	7/7/2016	1	37	4	1	DQ
18	7/9/2016	1	29	7	0	GH
19	7/30/2016	2	22	3	0	FG
20	8/29/2016	1.5	22	4	0	GH
21	9/10/2016	1	38	0	0	FG
22	9/21/2016	1	31	0	0	GH
23	10/1/2016	2	26	9	5	DQ
24	10/25/2016	1.5	20	4	0	DQ

Calculate total attendees

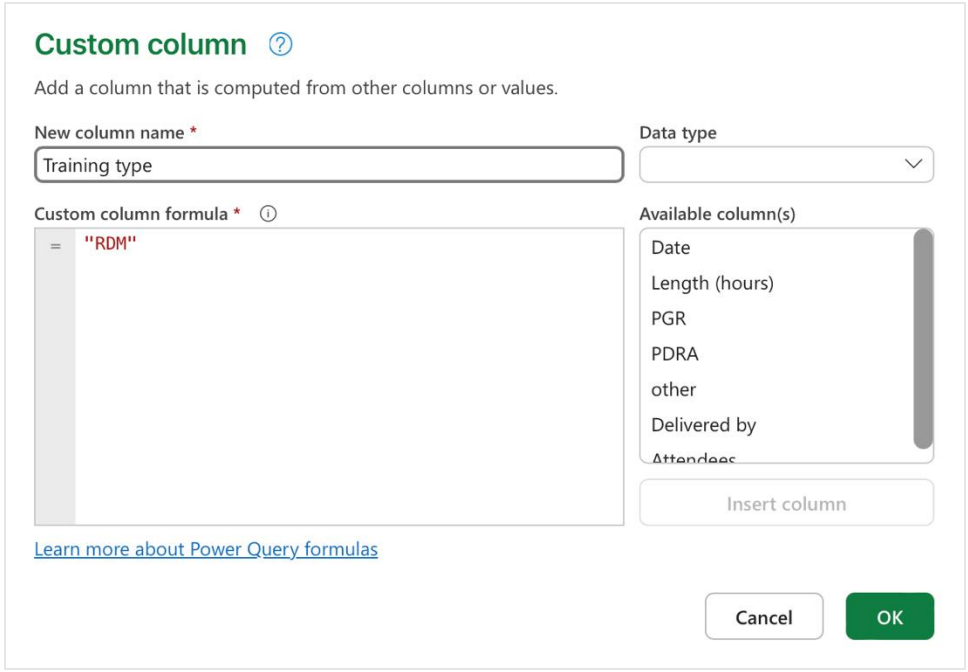
- Click on “Add column” in ribbon menu
- Select “Custom column”
- Name the column “Attendees”
- In the formula box:
  - Insert “PGR” column from the right
  - Add “+”
  - Insert “PDRA” column
  - Add “+”
  - Insert “other” column
- Click OK



Add a column to indicate training type

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."

- Add another "Custom column"
- Name the column "Training type"
- In the formula box, type "RDM"



Switch to “Open access” and clean that table as well

- Remove first 5 columns
- Use correct column headers
- In “Len” column, use “replace values” to replace “ hours” and “ hour” with nothing
- In “Len”, use Format → Trim to remove any remaining whitespace
- Change “Len” data type to “Decimal number”
- Rename column from “Len” to “Length (hours)”
- Delete empty 4th column
- Rename final column to “Cancelled”
- Add a custom column called “Training type” with the value “Open access”

Combine the two tables into one with “append”

- Under “Home”, click “Append queries”
- Select “Append queries as new”
- Select “RDM” as first table and “Open access” as second table
- Rename Query from “Append” to “2016 Trainings”
- Optional: remove the extra attendance columns
- Close and load data back into Excel

Quick check-in with small groups, Short break

10 minutes

Example data project: Merging data from multiple sources

20 minutes

Simulated survey data

survey-responses.csv

	A	B	C	D	E	F	G	H	I
1	ResponseID	Q1	Q2a	Q2b	Q2c	Q2d	Q2e	Q3	Q4
2	01JB6YVDDMTYG1SWA	David M. Rubenstein R	A little	No opinion	A little	A lot	Not at all	Library signage, Library	I am often frustrated by th
3	01JB6YVDDMQECCX7C	Perkins & Bostock Libra	No opinion	No opinion	No opinion	No opinion	No opinion	I don't want to be inforn	DUL is doing great. The on
4	01JB6YVDDNK4B96XK5	Perkins & Bostock Libra	Not at all	A little	A little	A lot	A little	Directly from library sta	I strongly suggest that libr
5	01JB6YVDDNVD8WCHS	Music Library	A lot	A lot	No opinion	Not at all	Not at all	Library signage, Library	I like having a hot water di
6	01JB6YVDDPQW6KMNI	Perkins & Bostock Libra	A lot	A lot	A little	A lot	A little	Library signage, Directl	I just want to reiterate hov
7	01JB6YVDDPN2ZAE20K	Lilly Library	A little	A little	A little	A little	No opinion	Directly from library sta	I think that making more s
8	01JB6YVDDQEGQAM38	Lilly Library	A little	Not at all	A lot	Not at all	No opinion	Library signage, Library	PLEASE consider keep the
9	01JB6YVDDRDZEJW54V	Music Library	No opinion	A lot	A little	Not at all	A lot	Library signage, Library	The printer paper cabinets
10	01JB6YVDDRFAGWJ3SI	David M. Rubenstein R	A lot	Not at all	No opinion	A little	A lot	Directly from library sta	Better mobile app
11	01JB6YVDDS33X9V9JD	Perkins & Bostock Libra	No opinion	A little	A lot	Not at all	No opinion	I don't want to be inforn	I wish the library had more
12									

survey-question-details.csv

	A	B	C	D
1	QuestionID	Question Long Text	Question Short Text	Question Type
2	Q1	Which library do you visit most frequently?	Primary Library	Select
3	Q2a	Library staff are considering offering or expanding the fo	Expanded Services: Individual Study	Likert
4	Q2b	Library staff are considering offering or expanding the fo	Expanded Services: Collaborative Study	Likert
5	Q2c	Library staff are considering offering or expanding the fo	Expanded Services: Textbook Lending	Likert
6	Q2d	Library staff are considering offering or expanding the fo	Expanded Services: Signage	Likert
7	Q2e	Library staff are considering offering or expanding the fo	Expanded Services: Lockers	Likert
8	Q3	How would you like to be informed about library service	Communication Preferences	Multi-Select
9	Q4	Any additional comments about Duke University Librari	Additional Comments	Free-Text
10				

survey-participant-details.csv

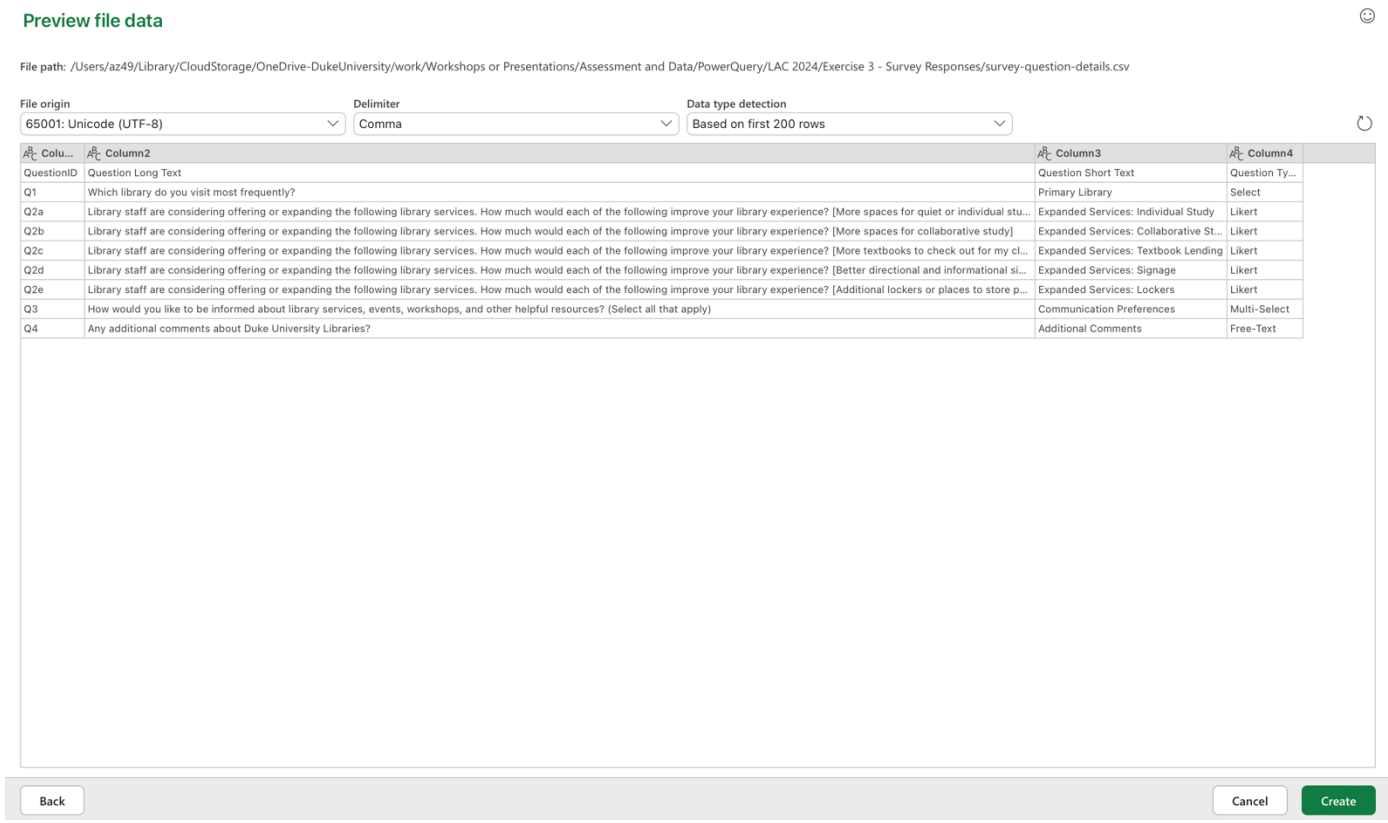
	A	B	C	D	E	F	G
1	ResponseID	First Name	Last Name	Gender	Class	Major	Discipline
2	01JB6YVDDMTYG1SWAD7HMZAQNH	Maggie	Jeffray	Female	First Year	Mathematics	STEM
3	01JB6YVDDMQECCX7CQH9Z82G2S	Tanny	Cromley	Male	Sophomore	Anthropology	Social Sciences
4	01JB6YVDDNK4B96XK5KYEQFH2F	Corrie	Moreman	Female	First Year	Sociology	Social Sciences
5	01JB6YVDDNVD8WCHSAAHMQ6RRT	Ursula	Lace	Female	Junior	Computer Sciences	STEM
6	01JB6YVDDPQW6KMNKNT39VWH8G	Ranee	Hebbard	Female	Senior	Literature	Humanities
7	01JB6YVDDPN2ZAE20KB7YS3Z9C	Pace	Banat	Male	First Year	Theatre	Humanities
8	01JB6YVDDQEGQAM38QERTRHMX2	Cilka	Crewes	Female	Senior	Statistics	STEM
9	01JB6YVDDRDZEJW54VX4P8085S	Yorgos	Mitford	Male	Sophomore	Computer Sciences	STEM
10	01JB6YVDDRFAGWJ3SMYDD313A9	Waylan	Theuss	Male	Junior	Biology	STEM
11	01JB6YVDDS33X9V9JD8NFZWZBE	Natasha	Sloan	Female	First Year	Economics	Social Sciences

Clean responses data (try on your own)

- New blank workbook
- Load survey responses data (“Text/CSV”)
- Use first row as headers
- Split Q3 column (multi-select) on comma
- Trim whitespace for the resulting columns
- Unpivot all questions columns (Q1 to Q4)
- Rename columns:
  - “Attribute” to “QuestionID”
  - “Value” to “Response”
- In QuestionID column, replace values to clean up Q3 split column headers
  - Hint: look at column filter to see all values in column

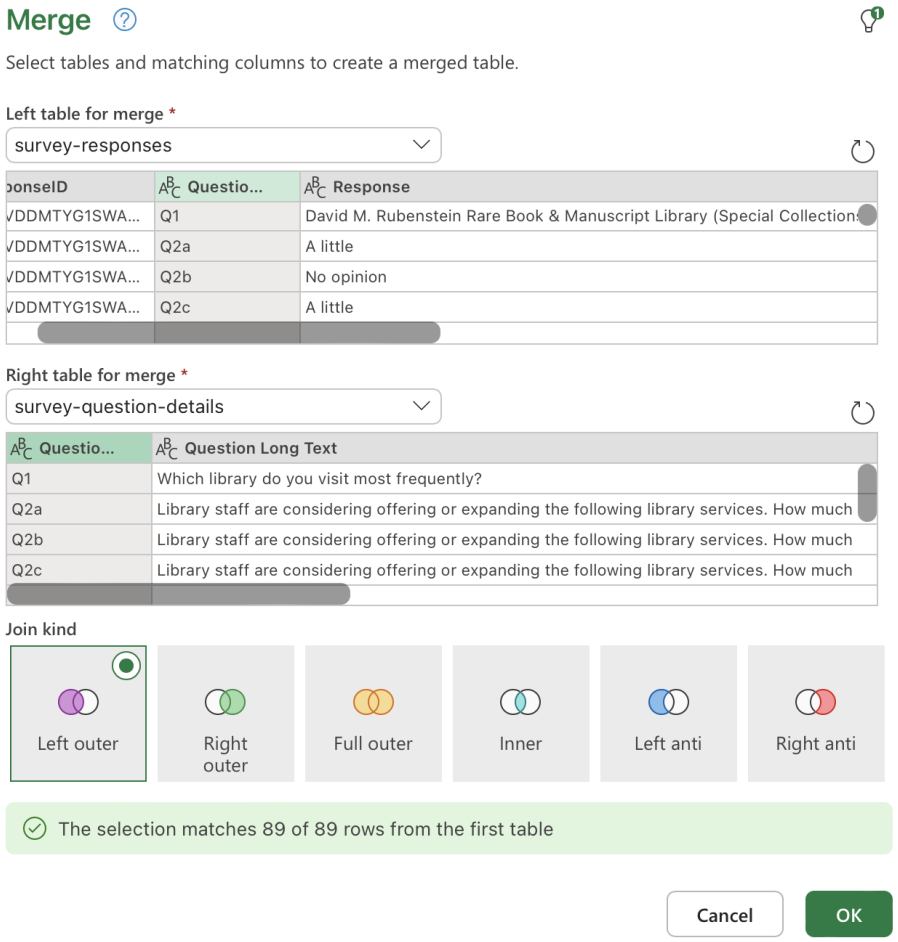
Create new query to bring in question details

- In “Home”, click “Get data”
- Select “Text/CSV”
- Select question details file
- Preview data and click “Create”
- Use first row as headers



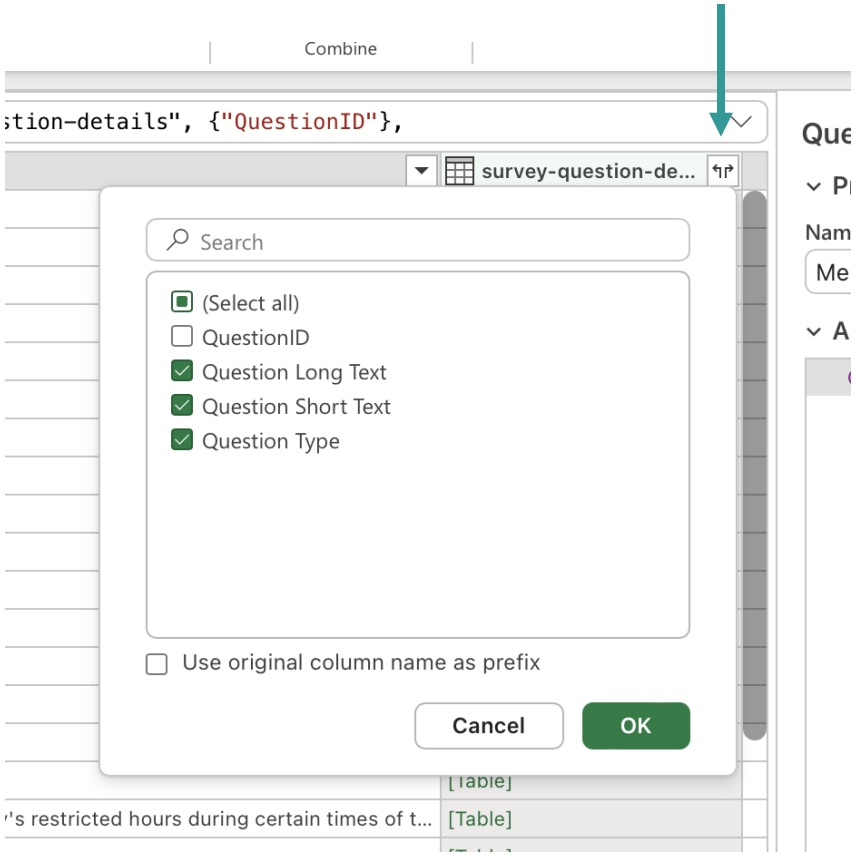
Blend the tables with “Merge queries”

- Under “Merge queries”, select “Merge queries as new”
  - A “merge” is like a database “join”
- For the “left” table, use survey responses
- In the data preview, select the QuestionID column
- For the “right” table, use question details
- In the data preview, select the QuestionID column again
- Leave the “join kind” as “Left outer”
- Click OK
- Rename new query as “responses-questions”



Decide what columns to bring in

- Scroll right to “survey-question-details” column
- Click on “expand” icon at the top
- Select columns to include, or leave all selected
  - Can uncheck “QuestionID”, since it’s a duplicate
- Click OK
- New columns will be added to the right

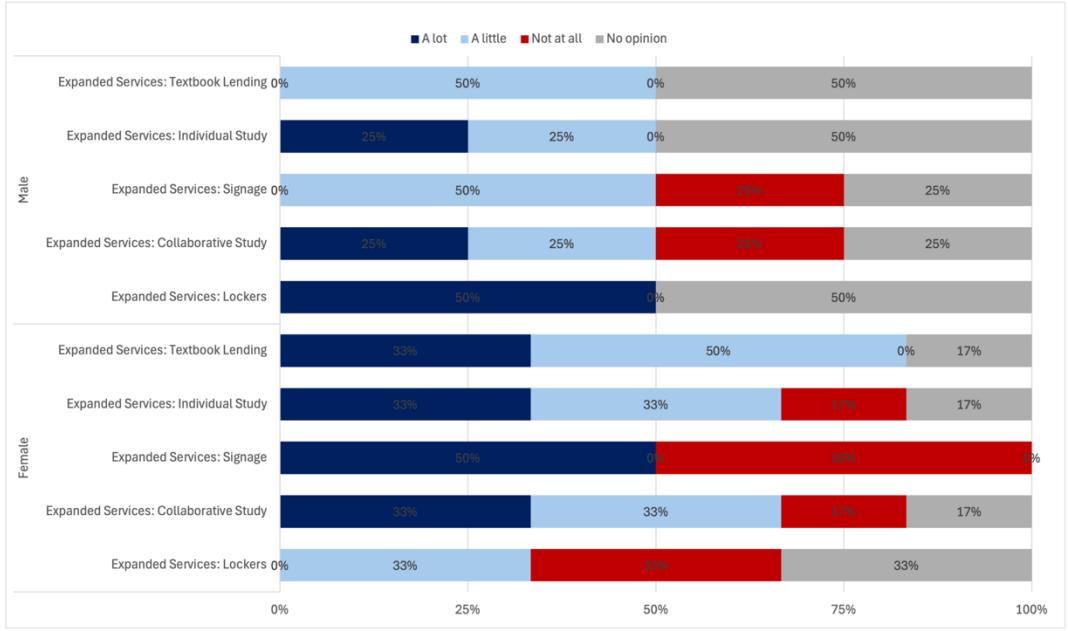


Optional: repeat merge process for participant details

- Create new query to bring in participant details
- Use first row as headers
- Merge queries as new
- Use “responses-questions” as left table, select ResponseID
- Use participant details as right, select ResponseID
- Click OK
- Rename query as “fully-merged”
- Expand column at the end, include only demographics columns
- Close & Load

Examples of what is now possible

	A	B	C	D	E	F
1	Question Type	Likert				
2						
3	**Count of ResponseID**	**Column Labels**				
4	**Row Labels**	**A lot**	**A little**	**Not at all**	**No opinion**	**Grand Total**
5	**Female**	**30%**	**30%**	**23%**	**17%**	**100%**
6	Expanded Services: Lockers	0%	33%	33%	33%	100%
7	Expanded Services: Collaborative Study	33%	33%	17%	17%	100%
8	Expanded Services: Signage	50%	0%	50%	0%	100%
9	Expanded Services: Individual Study	33%	33%	17%	17%	100%
10	Expanded Services: Textbook Lending	33%	50%	0%	17%	100%
11	**Male**	**20%**	**30%**	**10%**	**40%**	**100%**
12	Expanded Services: Lockers	50%	0%	0%	50%	100%
13	Expanded Services: Collaborative Study	25%	25%	25%	25%	100%
14	Expanded Services: Signage	0%	50%	25%	25%	100%
15	Expanded Services: Individual Study	25%	25%	0%	50%	100%
16	Expanded Services: Textbook Lending	0%	50%	0%	50%	100%
17	**Grand Total**	**26%**	**30%**	**18%**	**26%**	**100%**



Small group reflection

15 minutes

Discussion questions

- Are any of these examples similar to tasks you've had to do in the past?
- How do these methods compare with your usual methods?
- What questions do you still have about Power Query?

Full group Q&A session

15 minutes

DUKE UNIVERSITY LIBRARIES

Questions?

Angela Zoss, angela.zoss@duke.edu