STRINGFEST ANALYTICS

FIRST STEPS FOR POWER QUERY WITH MICROSOFT EXCEL

George Mount

Data Analyst & Educator at Stringfest Analytics

George works as an independent analyst and data analytics educator with the goal to help clients manage their data so they think more creatively. He serves as a technical expert and lead curriculum developer for Thinkful's data analytics program and is the instructor of the DataCamp course "Survey and Measure Development in R."

George blogs about data, innovation, and career development at georgejmount.com. He holds a master's degree in information systems with a certificate of achievement in quantitative methods from Case Western Reserve University

OBJECTIVES

Load data from Excel workbooks and csv files into Power Query

Perform common data wrangling and cleaning tasks

Combine data from multiple sources

FOLLOWING ALONG

- Each section is a sub-folder
- Demos = follow along with me
- Drills = try it yourself
    - Refresh your memory with the demo notes

1. POWER QUERY AS EXCEL'S ETL TOOL

What the @%&! is ETL?

"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."

-- (where else but) Wikipedia

1. EXTRACT

2. TRANSFORM

3. LOAD

Power Query & Excel Myth-busting

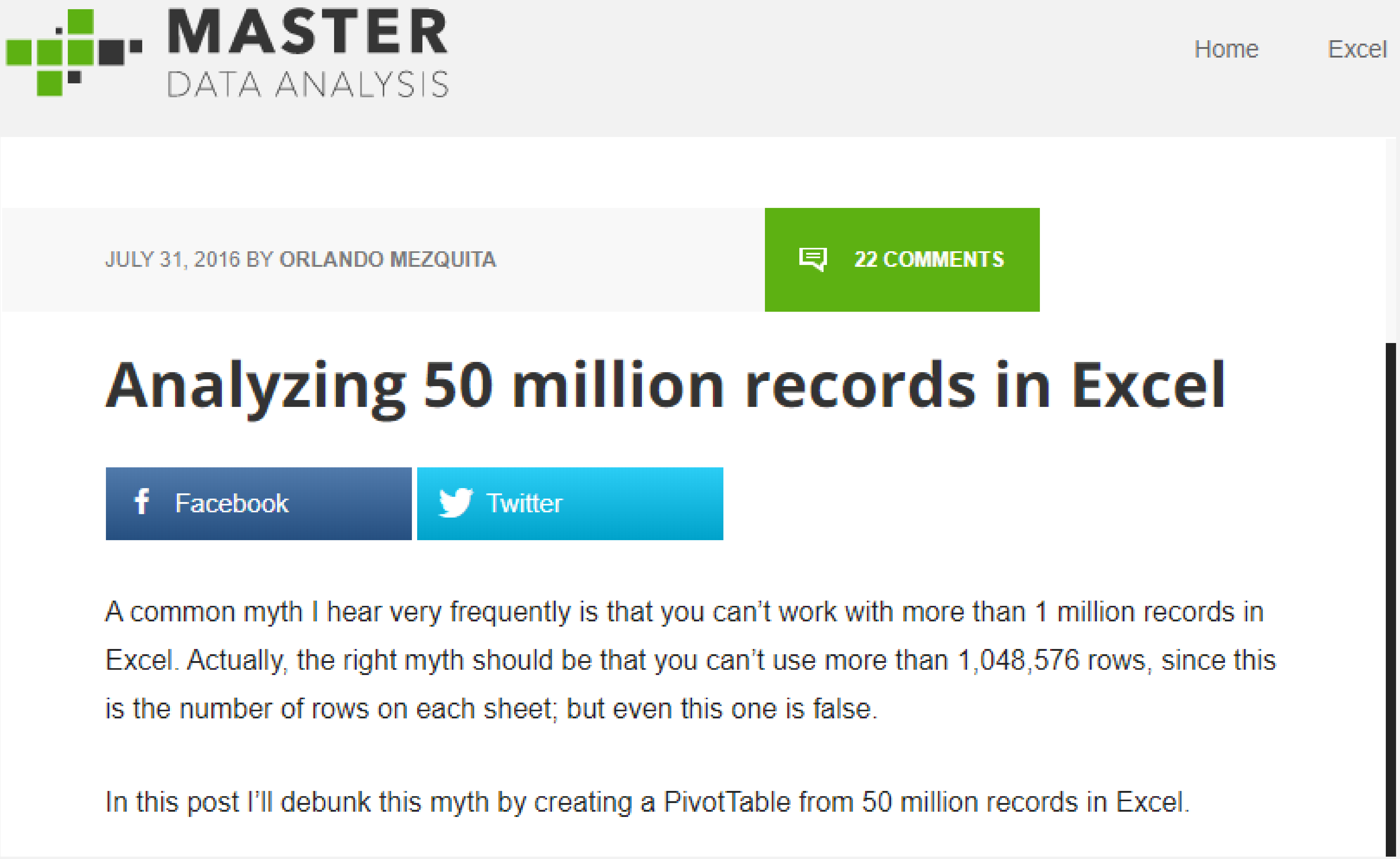
https://giphy.com/gifs/mythbusters-gif-qVuWpc7MxvGz6

1. “EXCEL IS NOT REPRODUCIBLE”

2. “EXCEL ONLY DOES STRUCTURED DATA”

- Access
- .txt and .csv files
- SQL Server & other relational databases
- XML, HTML & Web data
- SharePoint
- Hadoop
- oData
- *Combinations of the above…*

3. "EXCEL CAN'T HANDLE LARGE DATASETS"



JULY 31, 2016 BY ORLANDO MEZQUITA

Analyzing 50 million records in Excel

A common myth I hear very frequently is that you can't work with more than 1 million records in Excel. Actually, the right myth should be that you can't use more than 1,048,576 rows, since this is the number of rows on each sheet; but even this one is false.

In this post I'll debunk this myth by creating a PivotTable from 50 million records in Excel.

https://www.masterdataanalysis.com/ms-excel/analyzing-50-million-records-excel/

What did we do before Power Query?

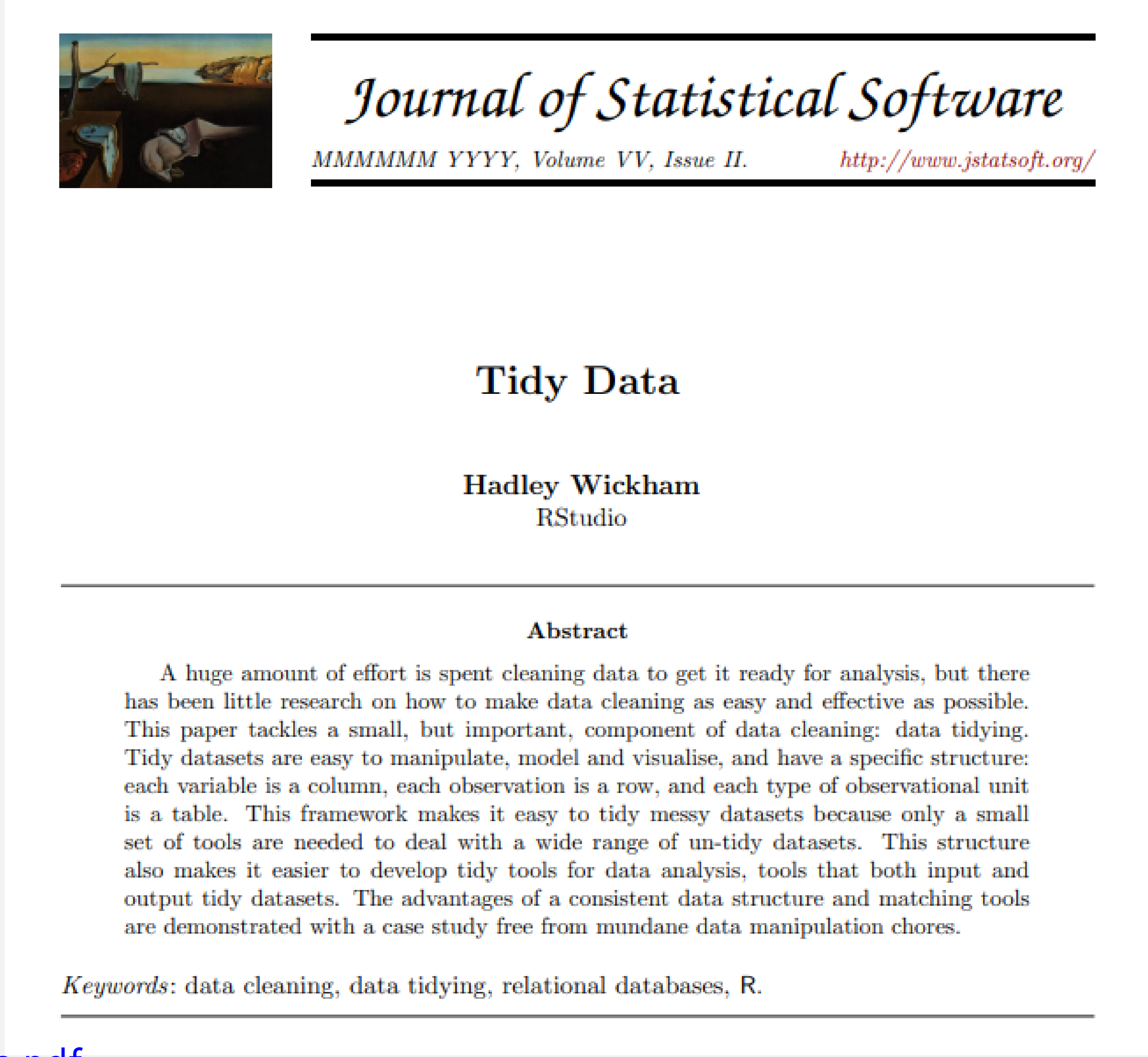
- File: `wholesale-customers.xlsx`
- How would you make this data “PivotTable-ready?”

QUESTIONS?

2. WHAT MAKES DATA TIDY?

DATA CLEANING
GOLD & SILVER
PAWN
OPEN 24 HRS
HISTORY
YOU NEVER KNOW WHAT IS GOING TO COME THROUGH THAT ATTACHMENT.

TIDY'S ORIGINS



Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. http://www.jstatsoft.org/

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

http://vita.had.co.nz/papers/tidy-data.pdf

OBSERVATIONS AND VARIABLES

Observation: The unit that was measured

Variable: What was measured

DRILL

Which attributes are observations, and which are variables?

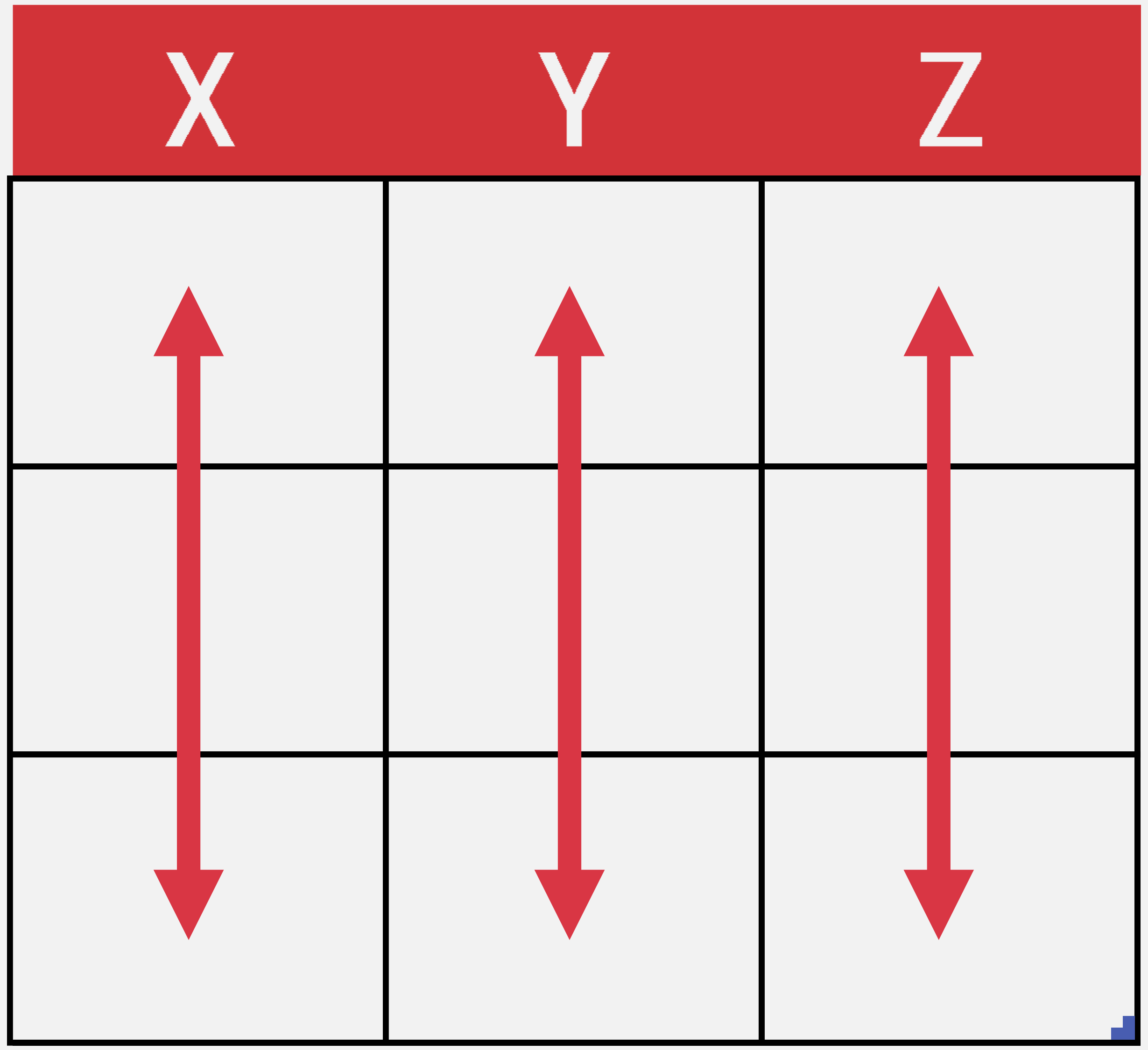
Attribute	Observation or variable?
Store #	
Month	
Sales	

DRILL

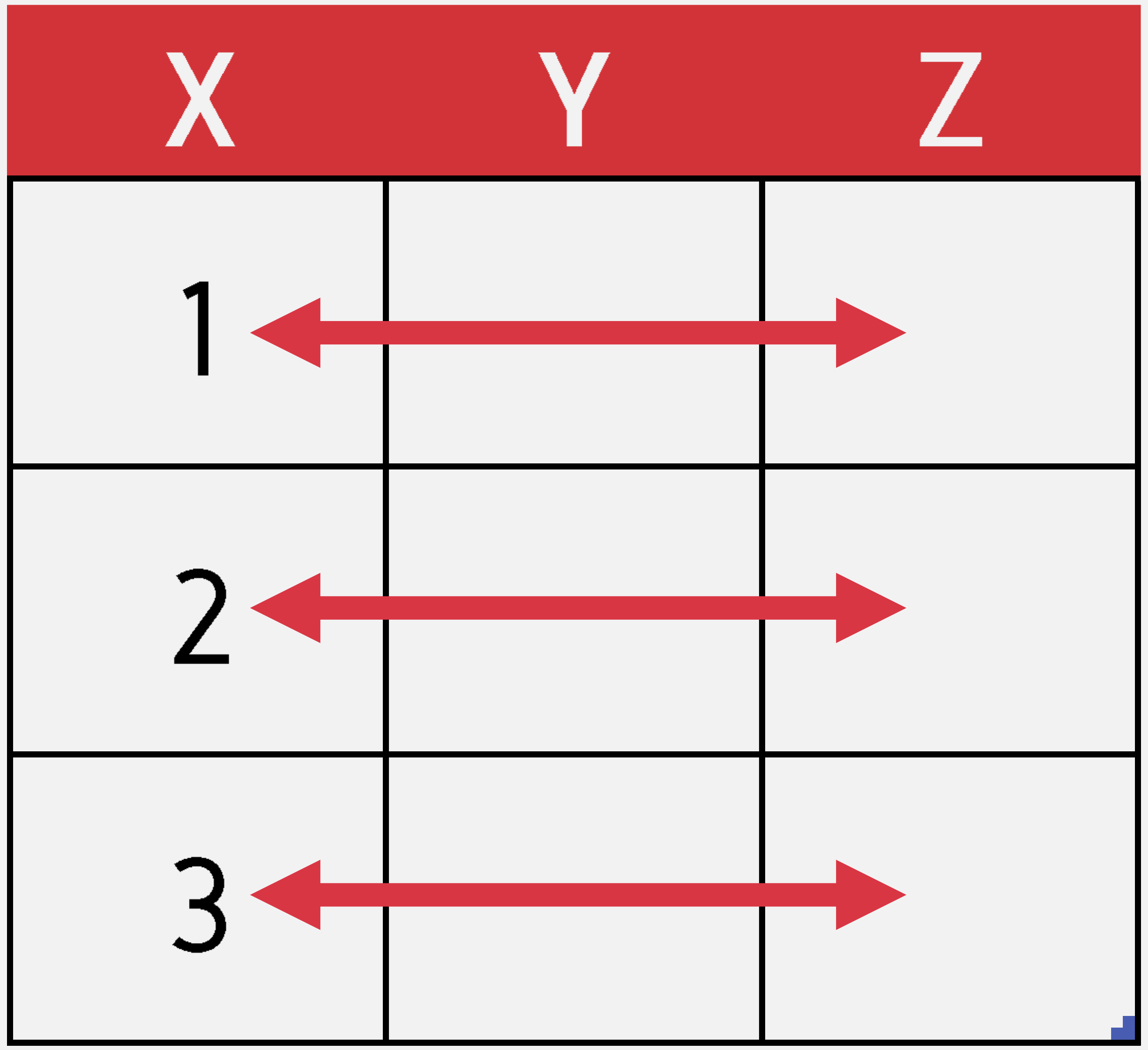
Which attributes are observations, and which are variables?

Attribute	Observation or variable?
Store #	Observation
Month	Observation
Sales	Variable

WHAT MAKES DATA TIDY?



Each variable in its own column



Each observation in its own row

Every cell is a populated observation-variable intersection…

DRILL

How would we tidy this dataset?

(Don’t do, just think)

		North	South	East	West
Model A	January	76	59	66	60
	February	75	66	60	62
	March	96	60	74	72
	April	56	95	83	97
	May	93	85	72	80
	June	50	99	92	77
	July	50	56	54	96
	August	59	90	88	86
	September	79	55	92	67
	October	76	99	97	98
	November	75	83	57	92
	December	89	99	85	73
Model B	January	58	59	84	84
	February	53	94	71	99
	March	69	72	77	51
	April	91	95	70	74
	May	95	95	59	85
	June	54	55	64	83
	July	75	60	52	73
	August	58	98	94	63
	September	97	87	94	50
	October	51	71	94	81
	November	50	98	96	92
	December	54	62	55	84

SOLUTION

		North	South	East	West
Model A	January	76	59	66	60
	February	75	66	60	62
	March	96	60	74	72
	April	56	95	83	97
	May	93	85	72	80
	June	50	99	92	77
	July	50	56	54	96
	August	59	90	88	86
	September	79	55	92	67
	October	76	99	97	98
	November	75	83	57	92
	December	89	99	85	73
Model B	January	58	59	84	84
	February	53	94	71	99
	March	69	72	77	51
	April	91	95	70	74
	May	95	95	59	85
	June	54	55	64	83
	July	75	60	52	73
	August	58	98	94	63
	September	97	87	94	50
	October	51	71	94	81
	November	50	98	96	92
	December	54	62	55	84

Model	Month	Region	Amount
A	January	North	76
A	February	North	75
A	March	North	96
A	April	North	56
A	May	North	93
A	June	North	50
A	July	North	50
A	August	North	59
A	Septembei	North	79
A	October	North	76
A	November	North	75
A	December	North	89
B	January	North	58
B	February	North	53
B	March	North	69
B	April	North	91
B	May	North	95
B	June	North	54
B	July	North	75
B	August	North	58
B	Septembei	North	97
B	October	North	51
B	November	North	50
B	December	North	54
A	January	South	59

DRILL

How would we tidy this dataset?

(Don't do, just think)

Department/Division	sku	Q1	Q2	Q3	Q4
Writing					
Pens	YF7TVW	116	133	60	68
	HBHEPS	115	81	72	78
Paper	3BN7AS	138	86	107	122
	86LFIY	98	59	139	91
	AUM13Y	103	80	93	135
Electronics					
Computers	BTQQTS	82	118	121	58
	331Z5U	77	70	76	62
Printers	RUW2LX	109	81	75	96
	1QMXDT	71	133	63	131

SOLUTION

Department/Division	sku	Q1	Q2	Q3	Q4
Writing					
Pens	YF7TVW	116	133	60	68
	HBHEPS	115	81	72	78
Paper	3BN7AS	138	86	107	122
	86LFIY	98	59	139	91
	AUM13Y	103	80	93	135
Electronics					
Computers	BTQQTS	82	118	121	58
	331Z5U	77	70	76	62
Printers	RUW2LX	109	81	75	96
	1QMXDT	71	133	63	131

Department	Division	sku	Quarter	Amount
Writing	Paper	3BN7AS	Q1	138
Writing	Paper	86LFIY	Q1	98
Writing	Paper	AUM13Y	Q1	103
Writing	Paper	3BN7AS	Q2	86
Writing	Paper	86LFIY	Q2	59
Writing	Paper	AUM13Y	Q2	80
Writing	Paper	3BN7AS	Q3	107
Writing	Paper	86LFIY	Q3	139
Writing	Paper	AUM13Y	Q3	93
Writing	Paper	3BN7AS	Q4	122
Writing	Paper	86LFIY	Q4	91
Writing	Paper	AUM13Y	Q4	135
Writing	Pens	YF7TVW	Q1	116
Writing	Pens	HBHEPS	Q1	115
Writing	Pens	YF7TVW	Q2	133
Writing	Pens	HBHEPS	Q2	81
Writing	Pens	YF7TVW	Q3	60
Writing	Pens	HBHEPS	Q3	72
Writing	Pens	YF7TVW	Q4	68
Writing	Pens	HBHEPS	Q4	78
Electronics	Computers	BTQQTS	Q1	82
Electronics	Computers	331Z5U	Q1	77
Electronics	Computers	BTQQTS	Q2	118
Electronics	Computers	331Z5U	Q2	70
Electronics	Computers	BTQQTS	Q3	121

QUESTIONS?

3. FIRST STEPS IN POWER QUERY

DEMO

- File: `star.xlsx`
- Load into Power Query
- Explore via Data Preview

DRILL

- File: `computers.xlsx`
- Load into Power Query
- Explore via Data Preview
- *Don’t forget the Demo Notes if you get stuck*

QUESTIONS?

4. TRANSFORMING ROWS IN POWER QUERY

DEMO

- File: `office-rsvps.xlsx`

DEMO

- File: `regional-sales.xlsx`

DRILL

File: `state-populations.xlsx`

Worksheet: `states`

1. Name the query `State populations`.
2. Remove the `United States` row from the data.
3. Fill down blanks on the `Region` and `Division` columns
4. Sort by `Population` from high to low
5. Load results into a PivotTable

Worksheet: `midwest_cities`

1. Convert this data into a table where each city is in its own row.

QUESTIONS?

5. TRANSFORMING COLUMNS IN POWER QUERY, PART I

DEMO

- File: `dvdrentals.xlsx`

DRILL

File: orders.xlsx

1. Convert the Date column to a month data type.
2. Convert the Account column to proper case.
3. Split the Opportunity column into three columns:
   A. Vendor
   B. Status
   C. Order Type

QUESTIONS?

6. TRANSFORMING COLUMNS IN POWER QUERY, PART II

6A. Transforming Columns

DEMO

- File: `population-densities.xlsx`

DRILL

File: `wholesale-customers.xlsx`

1. Tidy this data!
2. Create a field calculating 10% of the sales called `Tax`

6B. Appending Tables

DEMO

- Files: `oscars_yes.csv, oscars_no.csv`
- Start with a blank Excel workboook

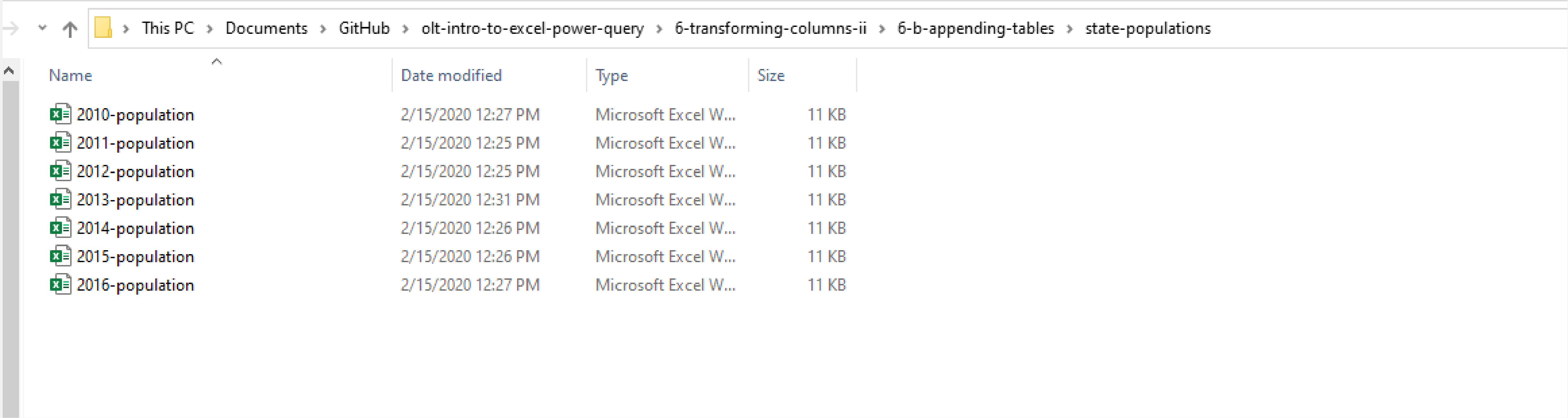
DRILL

File: `hof_inducted.csv, hof_not_inducted.csv`

1. Append these tables

DEMO

- Append from a folder of files:
  - Folder: `state-populations`



This PC › Documents › GitHub › olt-intro-to-excel-power-query › 6-transforming-columns-ii › 6-b-appending-tables › state-populations

Name	Date modified	Type	Size
2010-population	2/15/2020 12:27 PM	Microsoft Excel W...	11 KB
2011-population	2/15/2020 12:25 PM	Microsoft Excel W...	11 KB
2012-population	2/15/2020 12:25 PM	Microsoft Excel W...	11 KB
2013-population	2/15/2020 12:31 PM	Microsoft Excel W...	11 KB
2014-population	2/15/2020 12:26 PM	Microsoft Excel W...	11 KB
2015-population	2/15/2020 12:26 PM	Microsoft Excel W...	11 KB
2016-population	2/15/2020 12:27 PM	Microsoft Excel W...	11 KB

DRILL

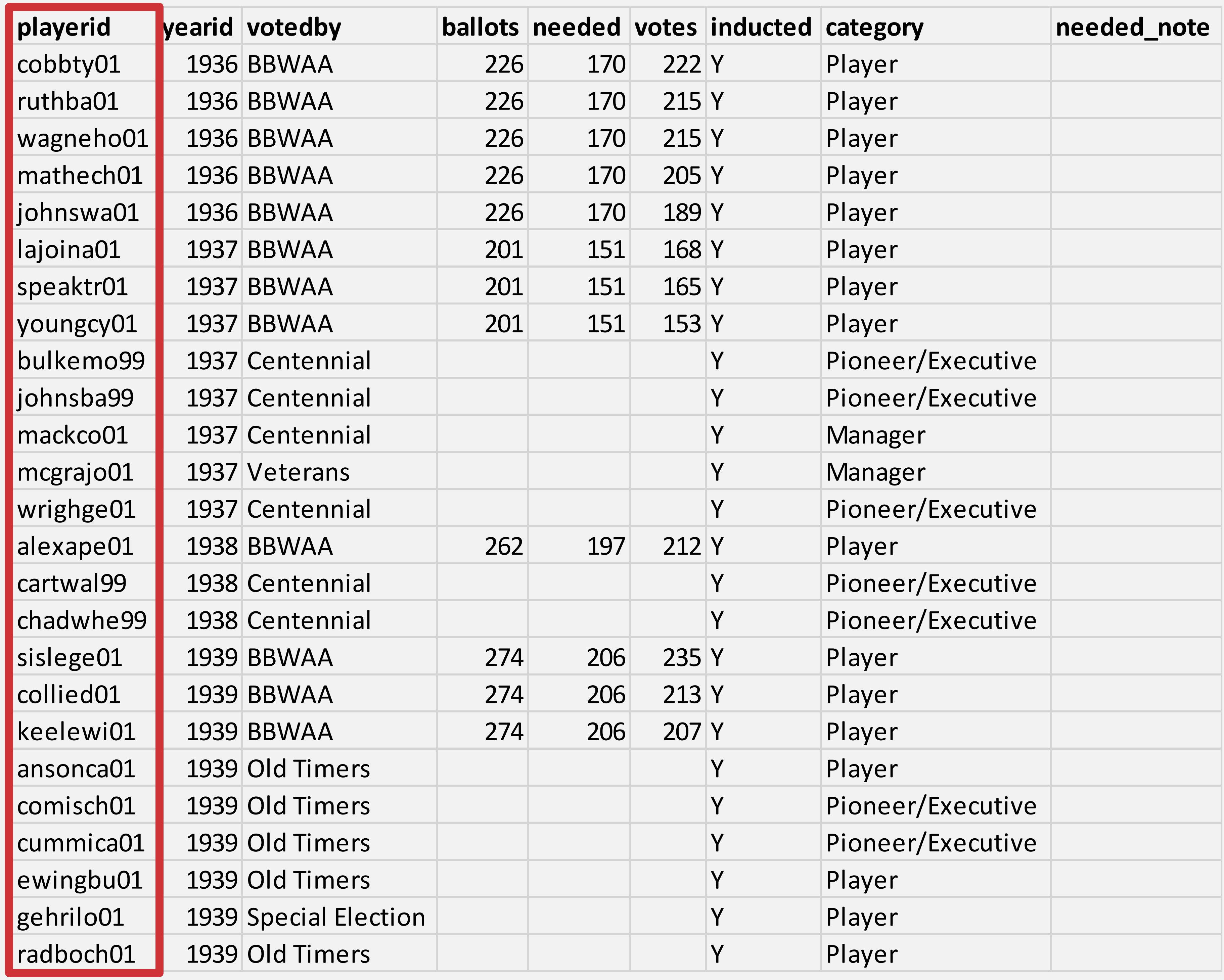
- `baseball` folder
    - This is a download of the csv version of the Lahman baseball database.
    - See if you can get a table of *all* files in this folder using Power Query.
        - In this case we *do not* want to transform the data, just load a table of the file metadata.

QUESTIONS?

QUESTIONS?

I have one:

Who are these people?



playerid	yearid	votedby	ballots	needed	votes	inducted	category	needed_note
cobbty01	1936	BBWAA	226	170	222	Y	Player	
ruthba01	1936	BBWAA	226	170	215	Y	Player	
wagneho01	1936	BBWAA	226	170	215	Y	Player	
mathech01	1936	BBWAA	226	170	205	Y	Player	
johnswa01	1936	BBWAA	226	170	189	Y	Player	
lajoina01	1937	BBWAA	201	151	168	Y	Player	
speaktr01	1937	BBWAA	201	151	165	Y	Player	
youngcy01	1937	BBWAA	201	151	153	Y	Player	
bulkemo99	1937	Centennial				Y	Pioneer/Executive	
johnsba99	1937	Centennial				Y	Pioneer/Executive	
mackco01	1937	Centennial				Y	Manager	
mcgrajo01	1937	Veterans				Y	Manager	
wrighge01	1937	Centennial				Y	Pioneer/Executive	
alexape01	1938	BBWAA	262	197	212	Y	Player	
cartwal99	1938	Centennial				Y	Pioneer/Executive	
chadwhe99	1938	Centennial				Y	Pioneer/Executive	
sislege01	1939	BBWAA	274	206	235	Y	Player	
collied01	1939	BBWAA	274	206	213	Y	Player	
keelewi01	1939	BBWAA	274	206	207	Y	Player	
ansonca01	1939	Old Timers				Y	Player	
comisch01	1939	Old Timers				Y	Pioneer/Executive	
cummica01	1939	Old Timers				Y	Pioneer/Executive	
ewingbu01	1939	Old Timers				Y	Player	
gehrilo01	1939	Special Election				Y	Player	
radboch01	1939	Old Timers				Y	Player	

6. VLOOKUP(), MEET JOIN

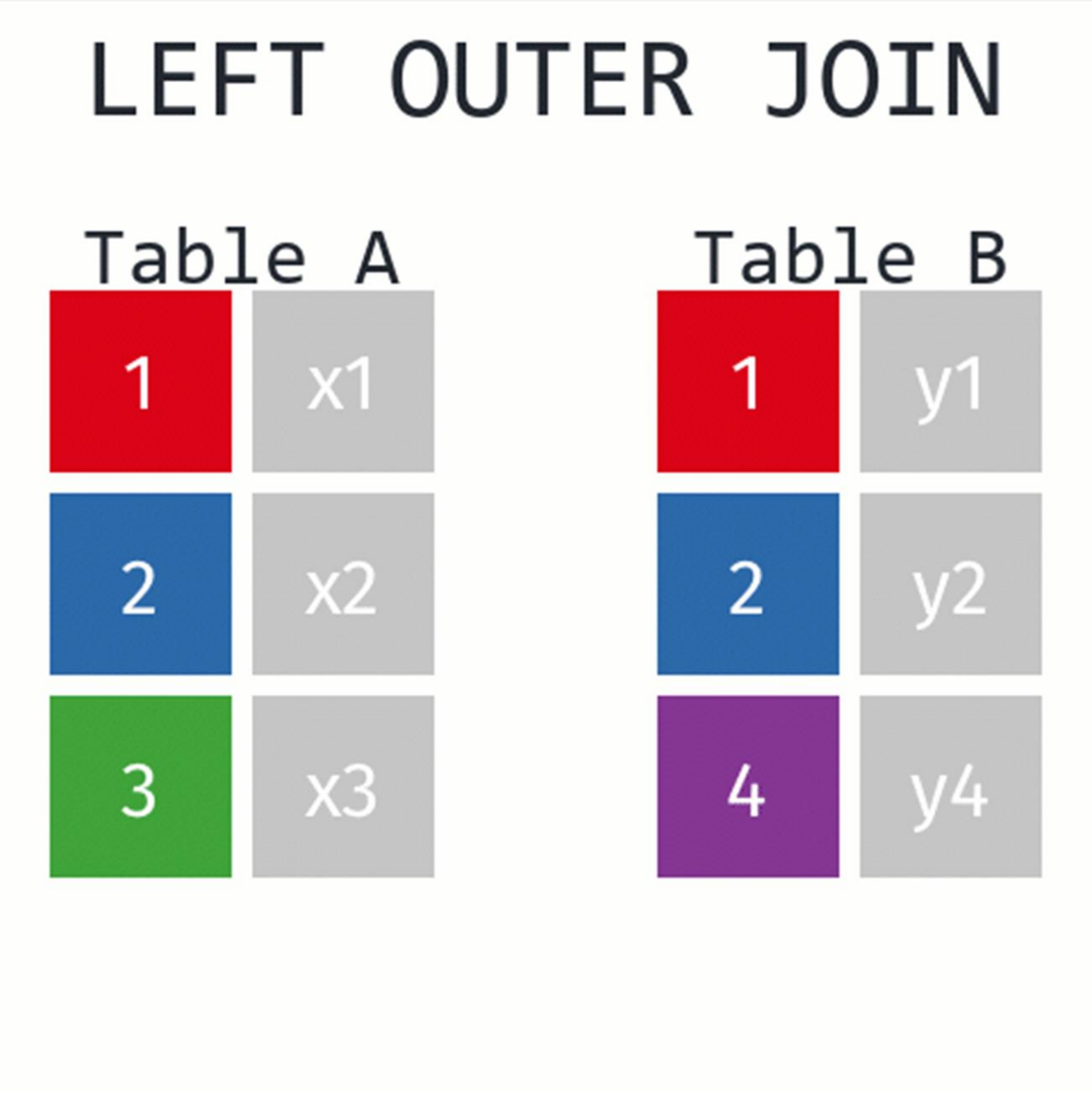
DUCT TAPE, MEET WELDER

VLOOKUP()

JOIN

DEMO

- File: `flights-and-planes.xlsx` (the VLOOKUP way)



LEFT OUTER JOIN

Table A

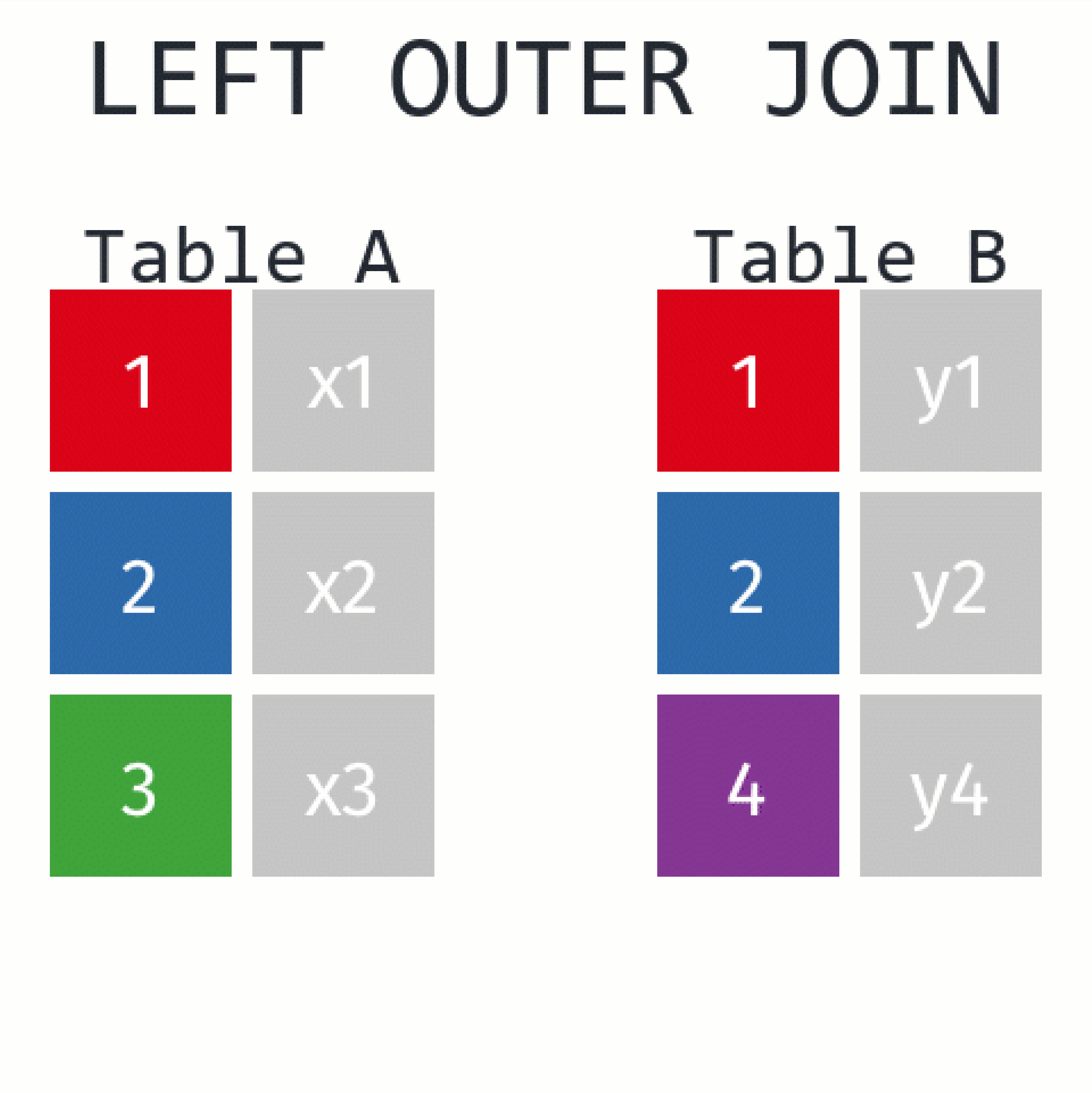
1	x1
2	x2
3	x3

Table B

1	y1
2	y2
4	y4

Returns ALL records from Table A and matching records from Table B. Results with no match are **null**.

https://github.com/gadenbuie/tidyexplain#mutating-joins



LEFT OUTER JOIN

Table A

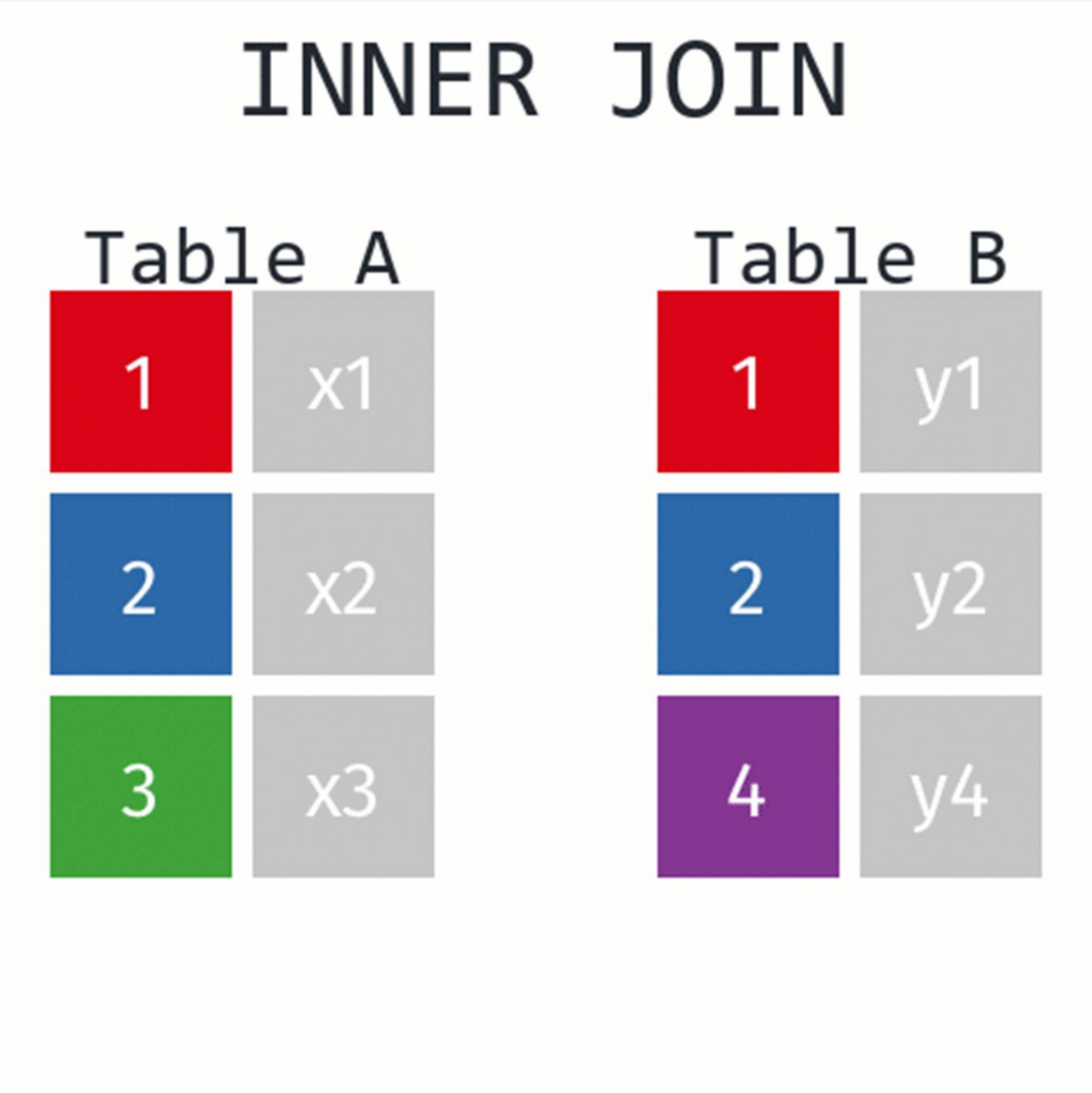
1	x1
2	x2
3	x3

Table B

1	y1
2	y2
4	y4

Returns ALL records from Table A and matching records from Table B. Results with no match are **null**.

https://github.com/gadenbuie/tidyexplain#mutating-joins



INNER JOIN

Table A

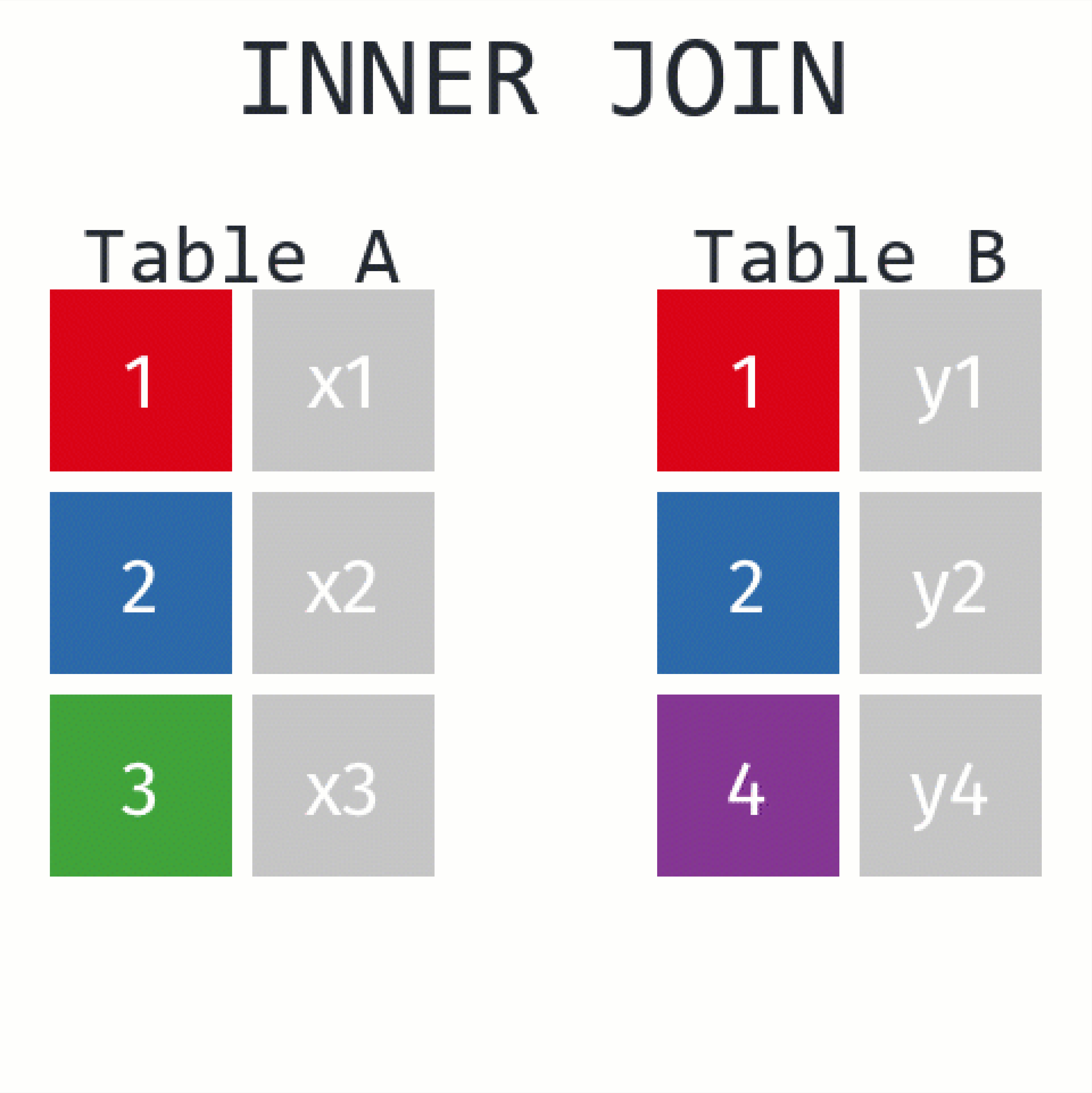
1	x1
2	x2
3	x3

Table B

1	y1
2	y2
4	y4

Returns records that have matching values in Tables A and B

https://github.com/gadenbuie/tidyexplain#mutating-joins



INNER JOIN

Table A

1	x1
2	x2
3	x3

Table B

1	y1
2	y2
4	y4

Returns records that have matching values in Tables A and B

https://github.com/gadenbuie/tidyexplain#mutating-joins

DEMO

- File: `flights-and-planes.xlsx` (the Power Query way)

DRILL

Files: `hof.csv, people-a-thru-m.csv`

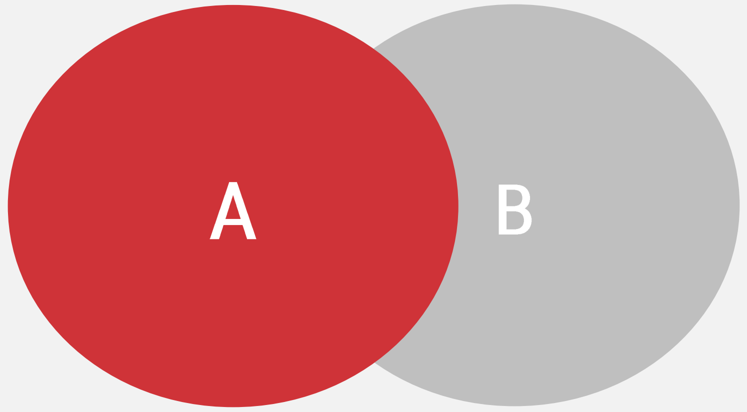
1. What is the result of a left outer join of `hof` on `people-a-thru-m`?
2. What about an inner join?

QUESTIONS?

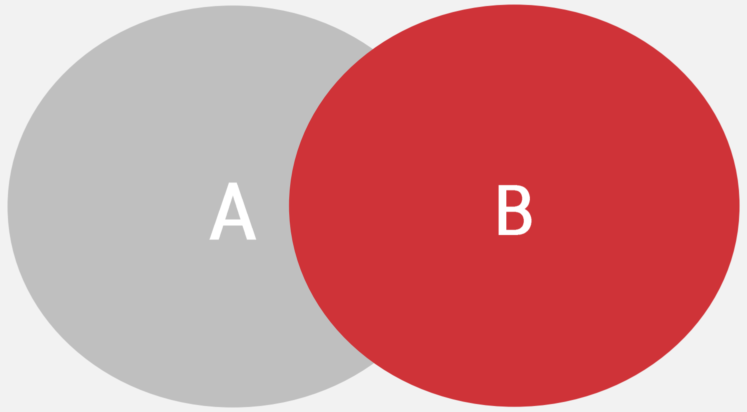
JOIN Beyond the basics

DATABASE JOIN TYPES

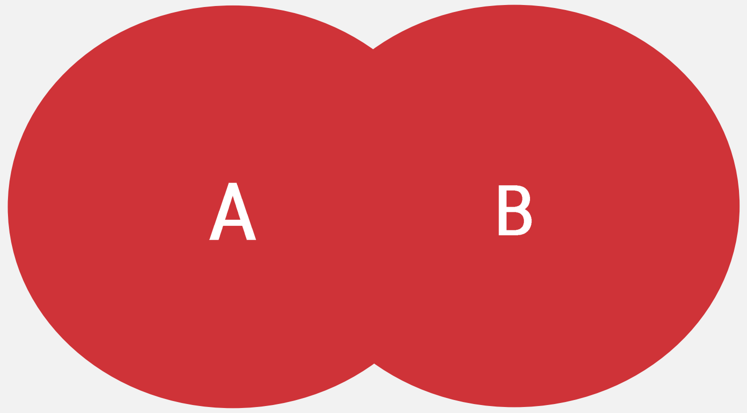
LEFT OUTER



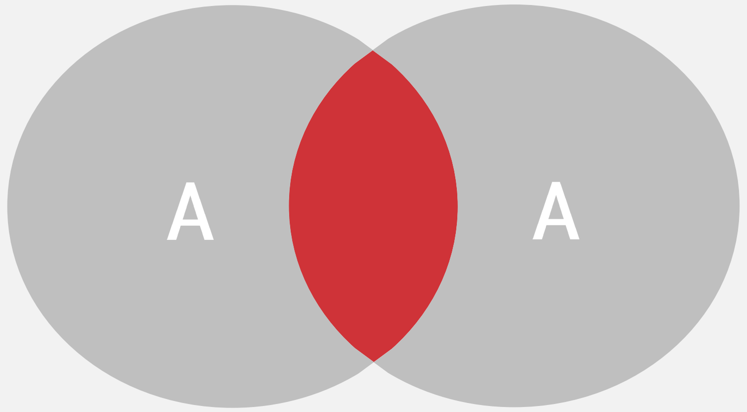
RIGHT OUTER



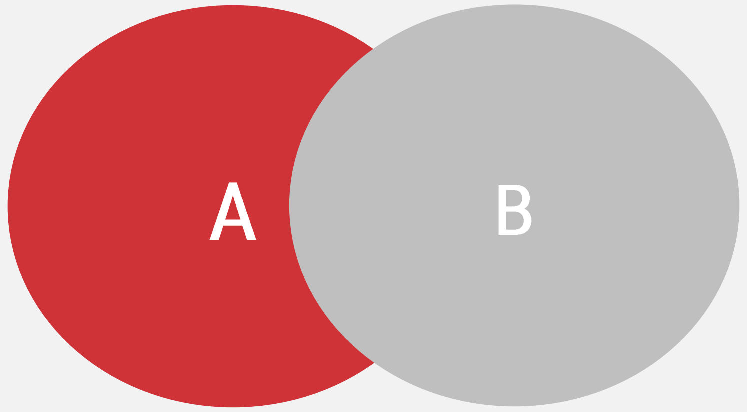
FULL OUTER



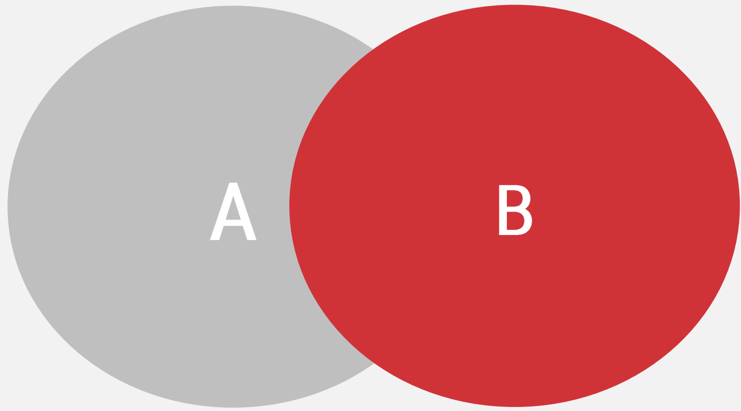
INNER



LEFT ANTI

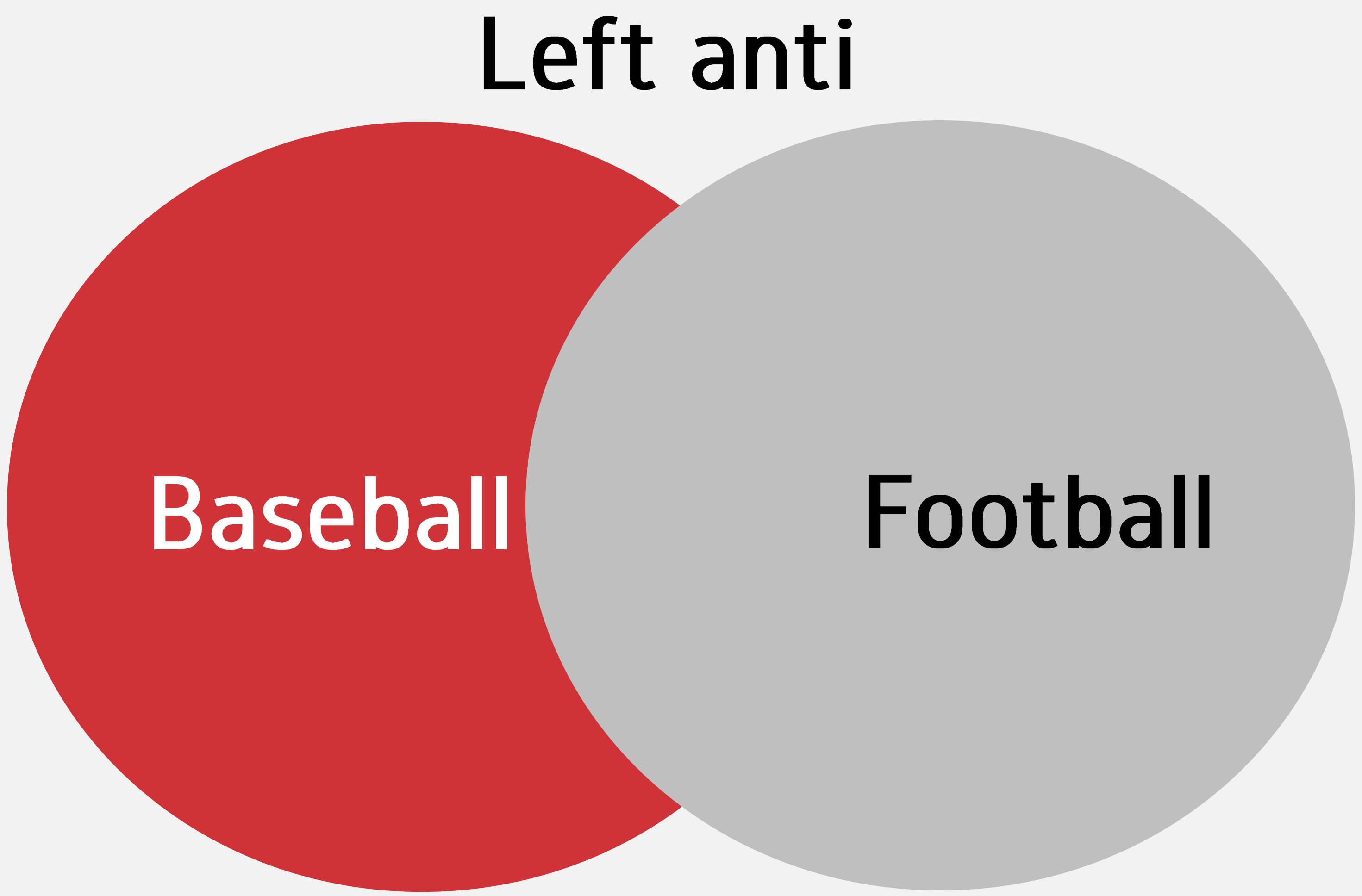


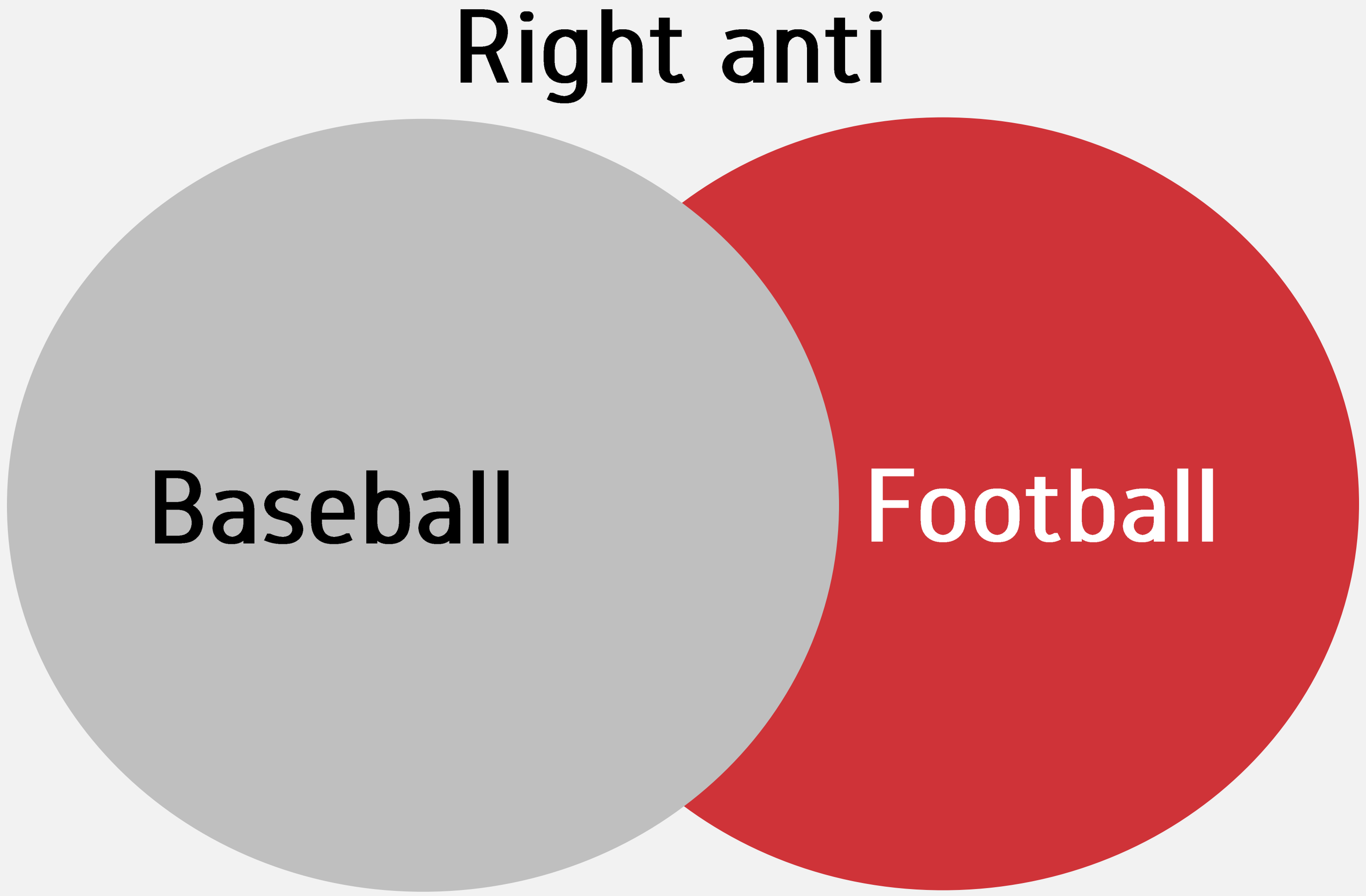
RIGHT ANTI



DEMO

- File: `championships-demo.xlsx`
- Which cities can claim *only* a baseball or football championship?

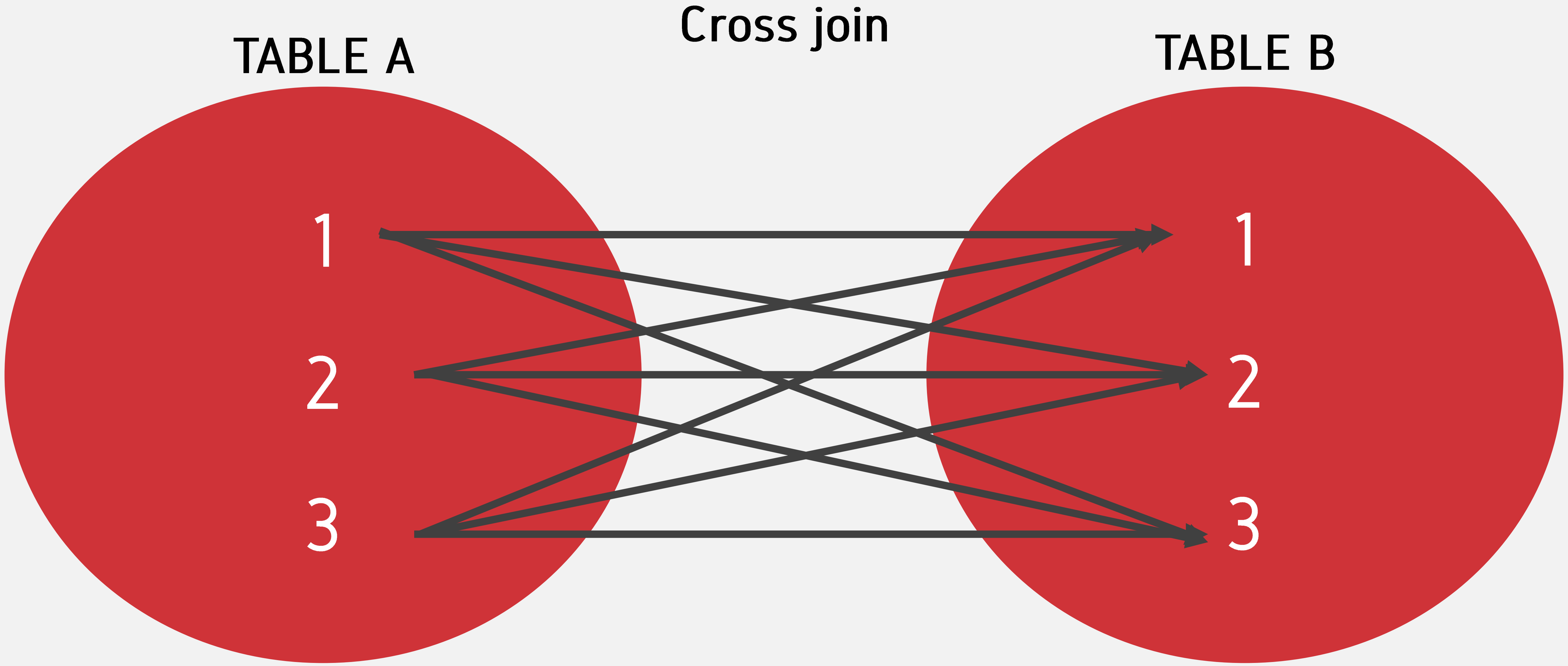




DRILL

- File: `championships-drill.xlsx`
- Which cities can claim *only* a hockey or basketball championship?

JOINS CAN GET EXOTIC



https://www.tutorialgateway.org/sql-cross-join/

DEMO

- File: `office-employees.xlsx`

DRILL

Files: `states.xlsx`

1. Create a table to record each state’s bird, flower and capital

QUESTIONS?

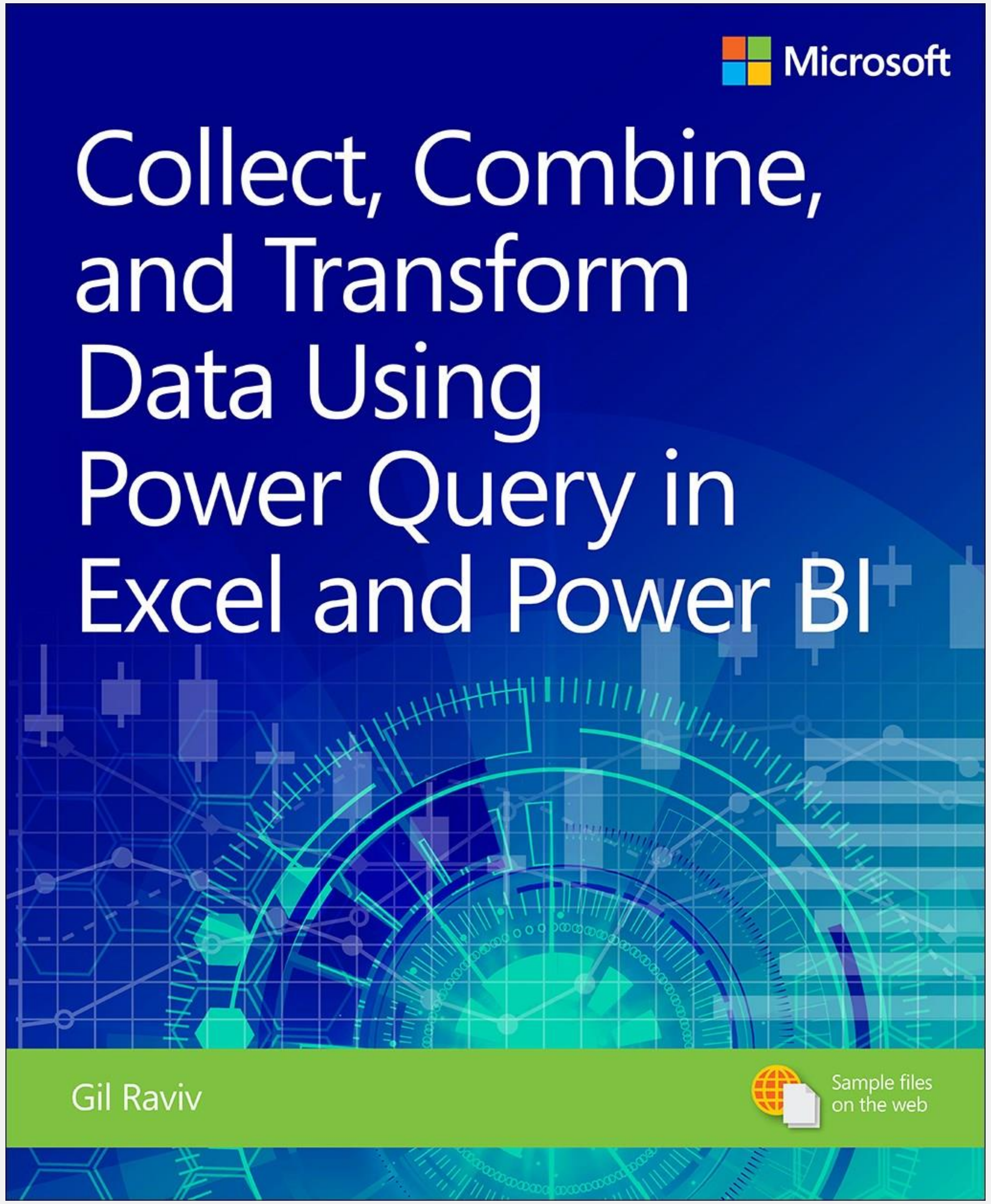
7. CONCLUSION

What to learn next?

- M language
- PowerPivot
- Power BI

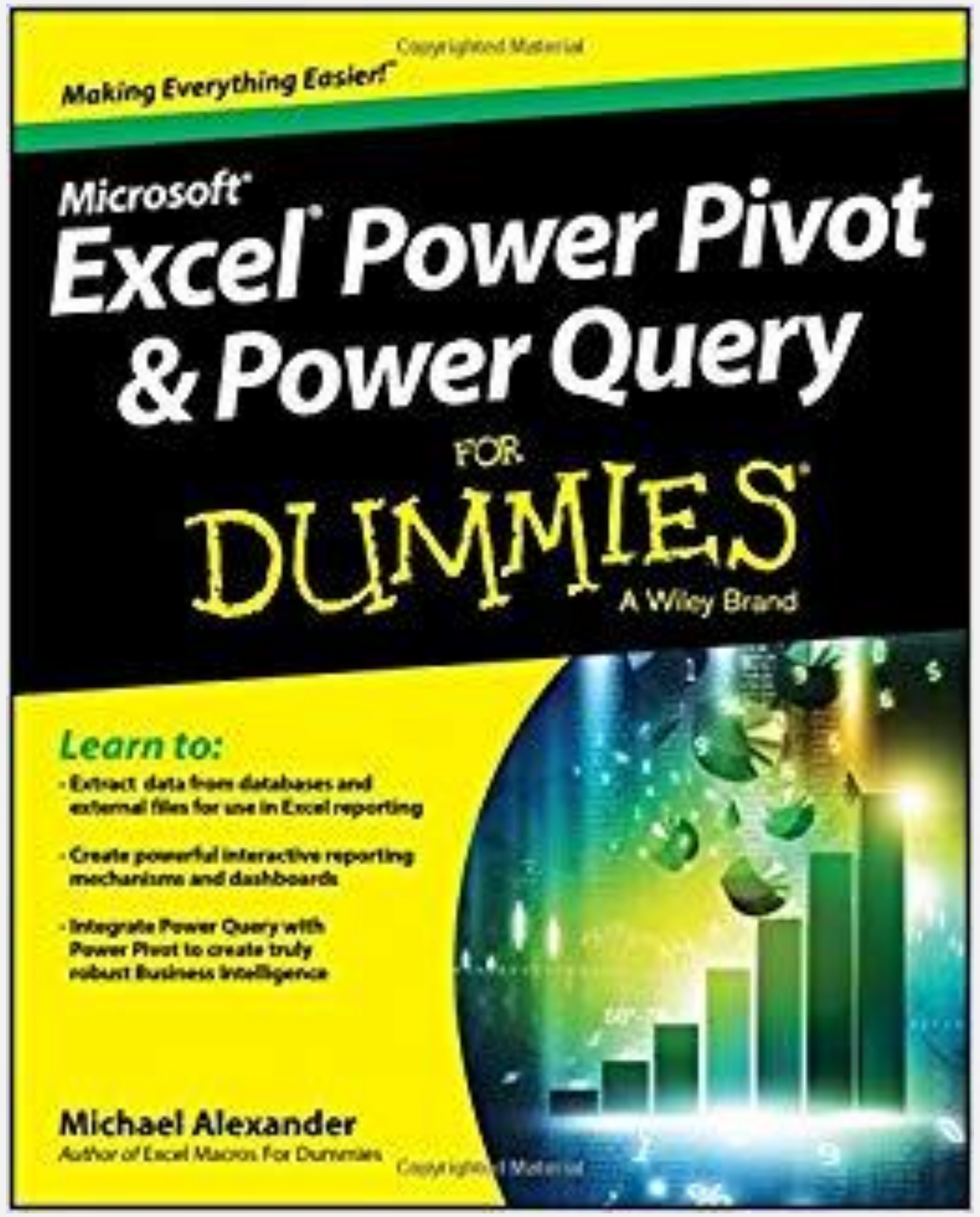
Collect, Combine, and Transform Data Using Power Query in Excel and Power BI, 1st Edition by Gil Raviv

- On O'Reilly Learning at https://learning.oreilly.com/library/view/collect-combine-and/9781509307982/


Microsoft
Collect, Combine, and Transform Data Using Power Query in Excel and Power BI
Gil Raviv
Sample files on the web

Excel Power Pivot and Power Query For Dummies, by Mike Alexander

- On O’Reilly Learning at https://learning.oreilly.com/library/view/excel-power-pivot/9781119210641/


Copyrighted Material
Making Everything Easier!
Microsoft Excel Power Pivot & Power Query
FOR DUMMIES
A Wiley Brand
Learn to:
Extract data from databases and external files for use in Excel reporting
Create powerful interactive reporting mechanisms and dashboards
Integrate Power Query with Power Pivot to create truly robust Business Intelligence
Michael Alexander
Author of Excel Macros For Dummies
Copyrighted Material

STRINGFEST ANALYTICS

LET'S TALK

LINKEDIN

linkedin.com/in/gjmount

EMAIL ADDRESS

george@stringfestanalytics.com

WEBSITE

stringfestanalytics.com

GITHUB

github.com/summerofgeorge

QUESTIONS?