Data Modeling in Excel

CEE412/CET522

TRANSPORTATION DATA MANAGEMENT AND VISUALIZATION

WINTER 2020

Announcements

❑Extra credit for online quiz on next Wednesday.

❑Assignment will be published on Canvas by the end of the day.
  ❑Due on Jan. 22

❑Project group
  ❑Please Wait until the third week

❑Piazza
  ❑piazza.com/washington/winter2020/cee412cet522
  ❑Add code: 2020Winter

Summary of Survey
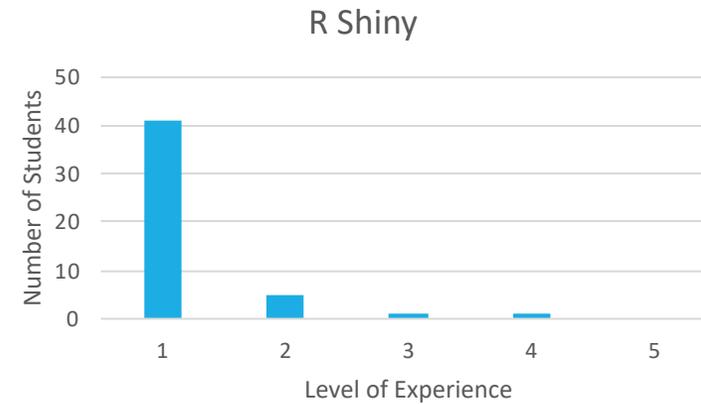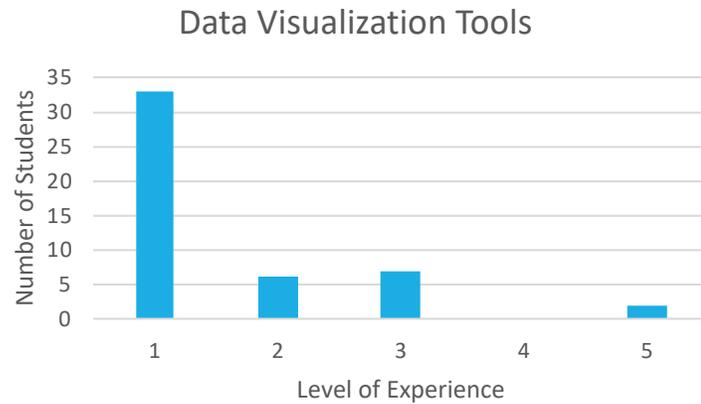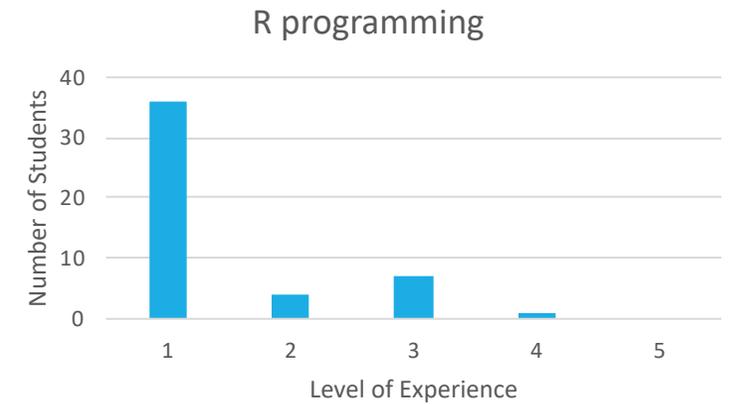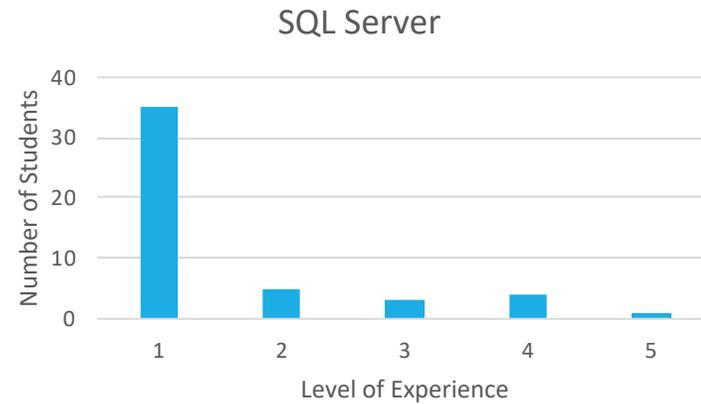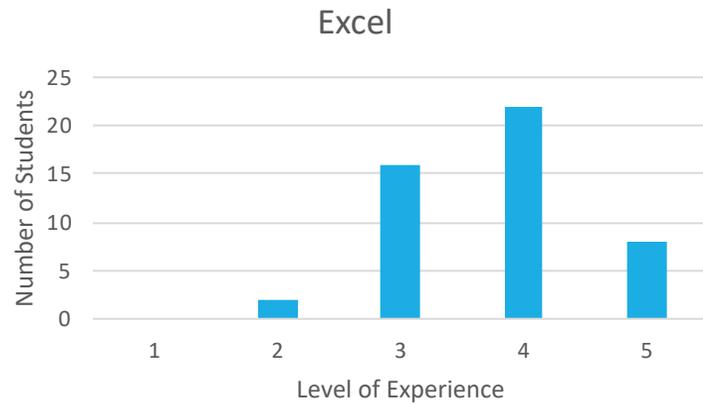
Topics

Data models

Data model Functionality in Excel
- Power Pivot
- Power View

Excel Exercise (Assignment 1)

Data Models

What is a data model?

◦ A logical structure or design for a collection of data.

◦ A data model is defined by the format and structure of the data elements, and the relationships between elements.

◦ Data represents things we want to store information about, so the data model represents some interpretation of reality.

Three levels of modeling:

Conceptual: Translating application needs into high-level conceptual data structure	**Logical:** Specification of data elements, domains, data types, and relationships	**Physical:** Physical design for implementation (more detailed than the logical level)

Data Models

Movie actor example: What kind of information do I want to store?

Maybe personal information, physical description, income, movie roles.

What is specified in the data model?

Actors
Name
Age
Height
Gender
Ethnicity

Play in

Roles
ActorName
Character
MovieTitle
Salary

Have

Movies
Title
Year
Budget
Genre

Data Models

Why do we need a data model?
- Design for efficiency in storage, access, management, and growth.
- Facilitate communication between different database users.
- Carefully plan HOW to manage data based partly on WHY you are storing it.

Accident database example (inefficient):

AccidentID	Vehicle1_Type	Vehicle2_Type	Vehicle3_Type	Milepost	Route	Pavement_type
144567	Sedan	SUV		114.2	005	asphalt
144568	Pickup			102.5	005	asphalt
144569	Compact	Pickup	SUV	120.1	005	asphalt
144570	SUV	Pickup		140.1	005	asphalt
…	…	…	…	…	…	…

Data Models

At the logical/physical level, we need to specify:

◦ What specific tables will be contained in the database?

◦ What specific fields each table will have?

◦ What data types will be used for each field?

◦ What values can a field take (domain constraints)?

◦ How the tables are related to each other?

◦ Access control and other practical issues – who can access what, etc.

Microsoft Excel

What it is:

- A spreadsheet software for data storage and analysis.
- Good for simple analysis and visualization.
- Very widely used.
- Advanced features include VBA, data connection tools, and other nifty things.

What it isn't:

- Data management system.
- Powerful statistical analysis tool.
- The best solution for large or complex datasets.
- Legitimate software development tool.
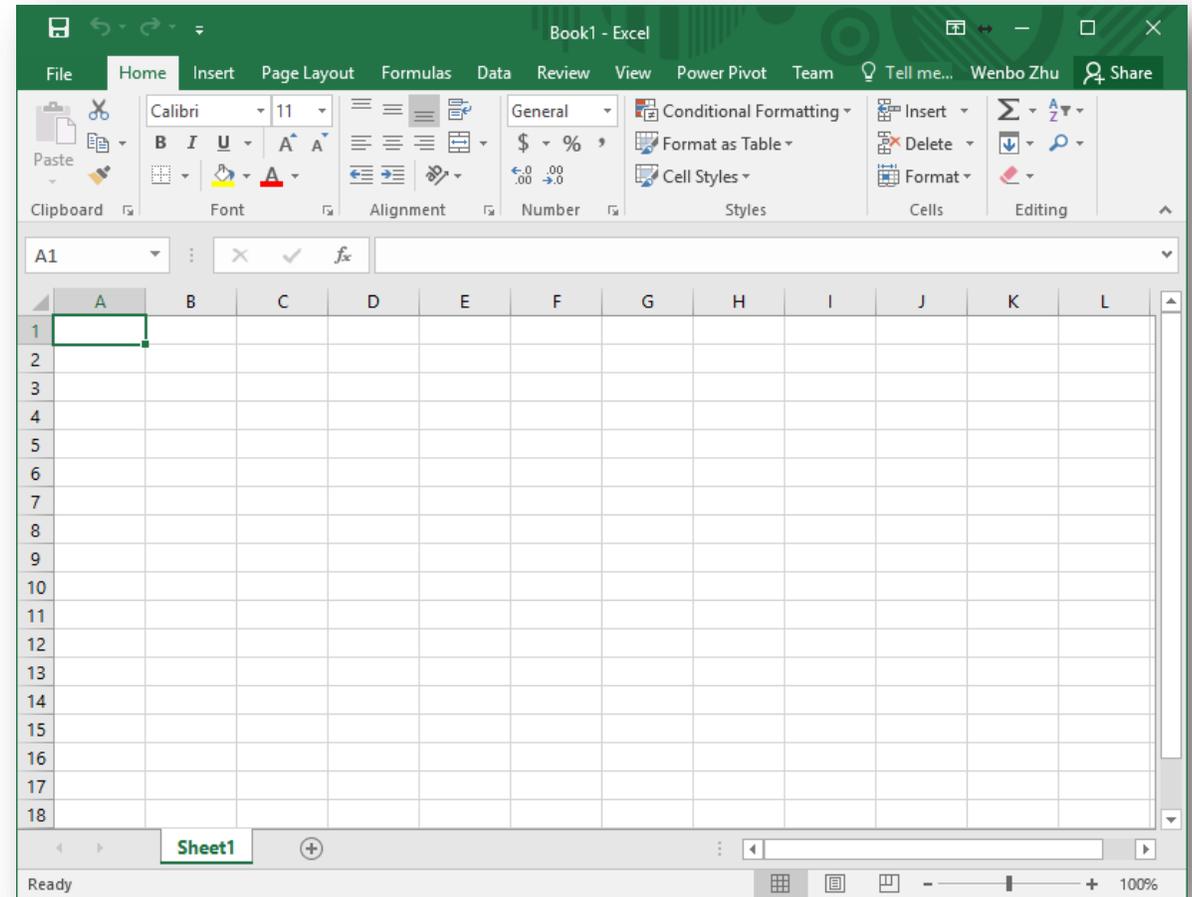
Microsoft Excel

Why are we talking about Excel?
- Because everyone has some experience with it.
- As a simple introduction to the broader topics in this class.
- To introduce less well known but useful functionality in Excel.

Data Models in Excel

Now on to data models in Excel

- Data model can reference Excel tables, as well as external data (e.g. databases, other workbooks).
- A data model is automatically created by adding tables to the workbook data model

Power Pivot (Excel Add-in)

- Based on powerful database technology.
- The only way to handle very large datasets in Excel (multiple millions of rows).
- Data modeling and schema editing in a special interface.
- Low memory usage, even compared to the source of the data.

Data Models in Excel

An example:

◦ Say we have an Excel workbook with two worksheets: Accidents and Vehicles.

◦ **Accidents**: A set of accident records for a road section and time period.

◦ **Vehicles**: a description of each vehicle/driver involved in a set of accident records.

◦ The two tables are related by the ReportNum field

Accidents
ReportNum
Date
Weather
Severity
Location

Involve

Vehicles
ReportNum
Make
Model
DriverAge
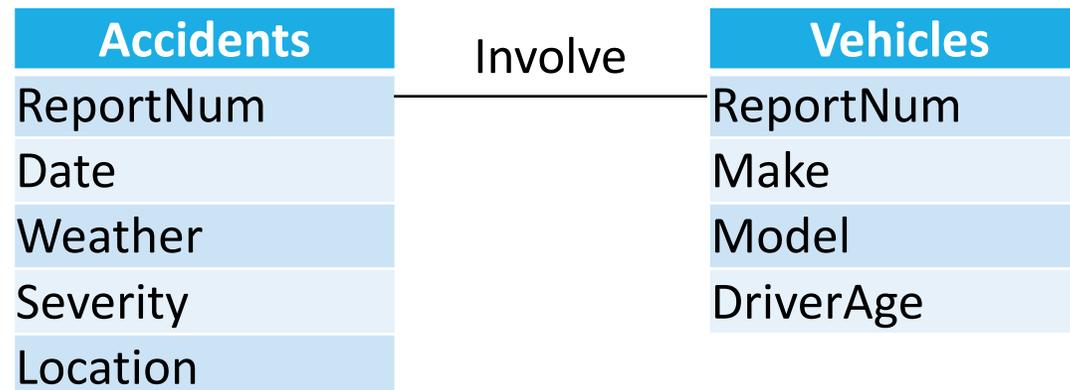
Data Models in Excel

Start by:
- Highlighting the data in each worksheet.
- Create tables by clicking the **Format as Table** button on the **Home** tab.
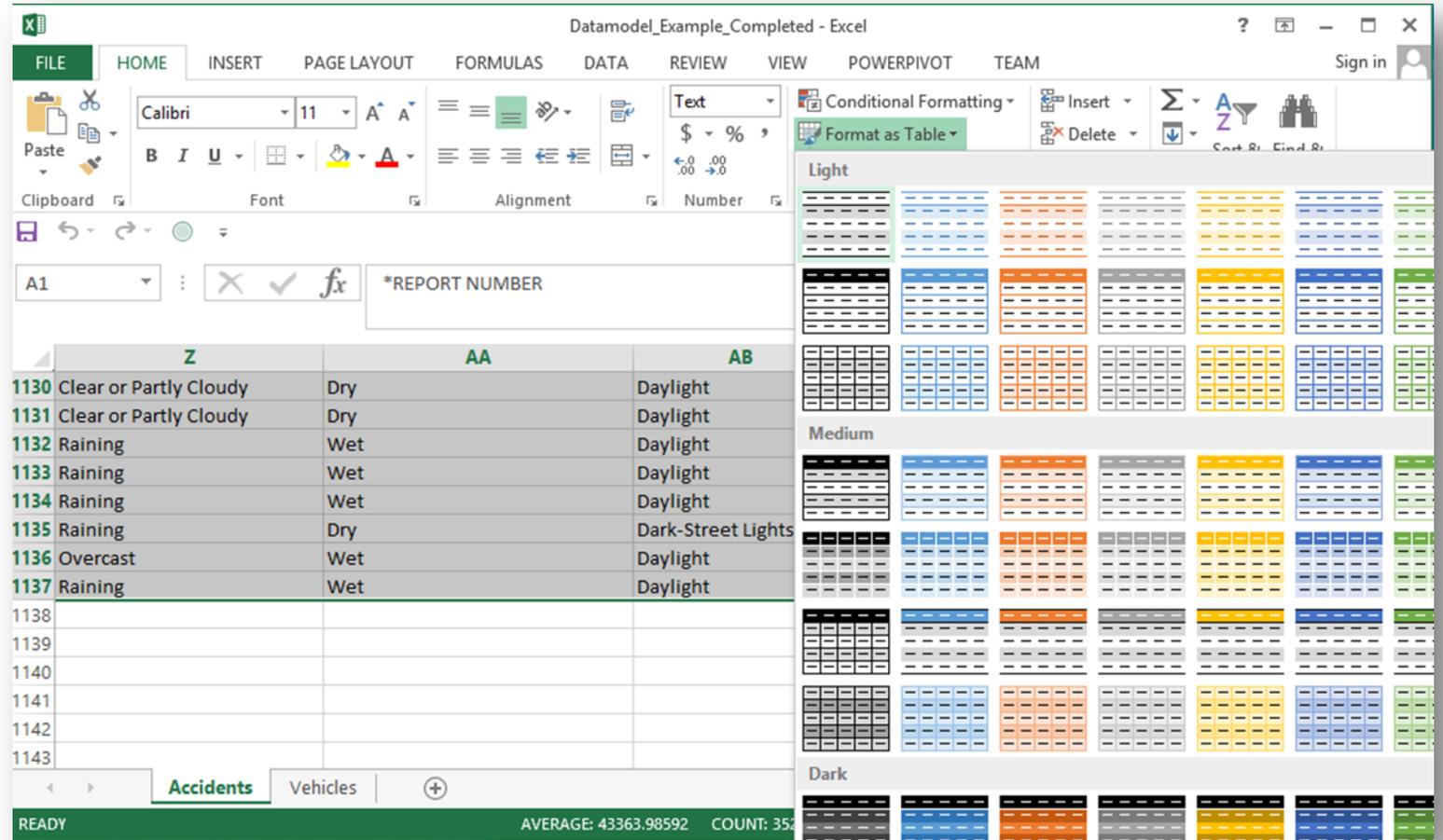
Data Models in Excel

○ When both tables are created, click the **Add to Data Model** button on the **Power Pivot** tab.

○ A workbook data model is automatically generated when tables are added.

○ Note: Power Pivot is a COM Add-in, and will need to be enabled the first time you use it. We will go through the process in the exercise later.

○ Next, define relationships…
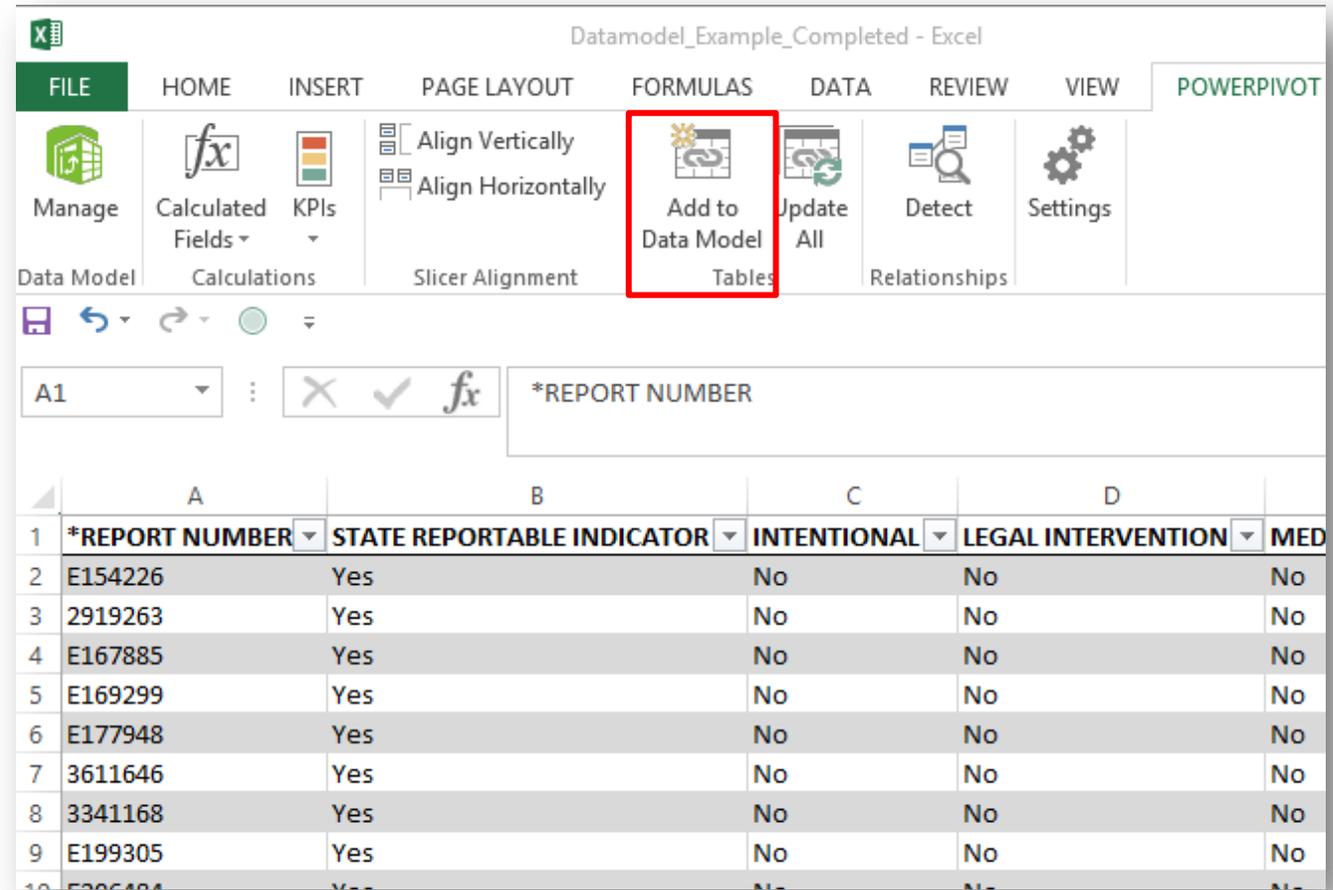
Data Models in Excel

- Relationships can be defined in multiple ways.
- One way is to view the **Diagram View** in the Power Pivot window.
- Simply drag the join field from one table to the other as shown
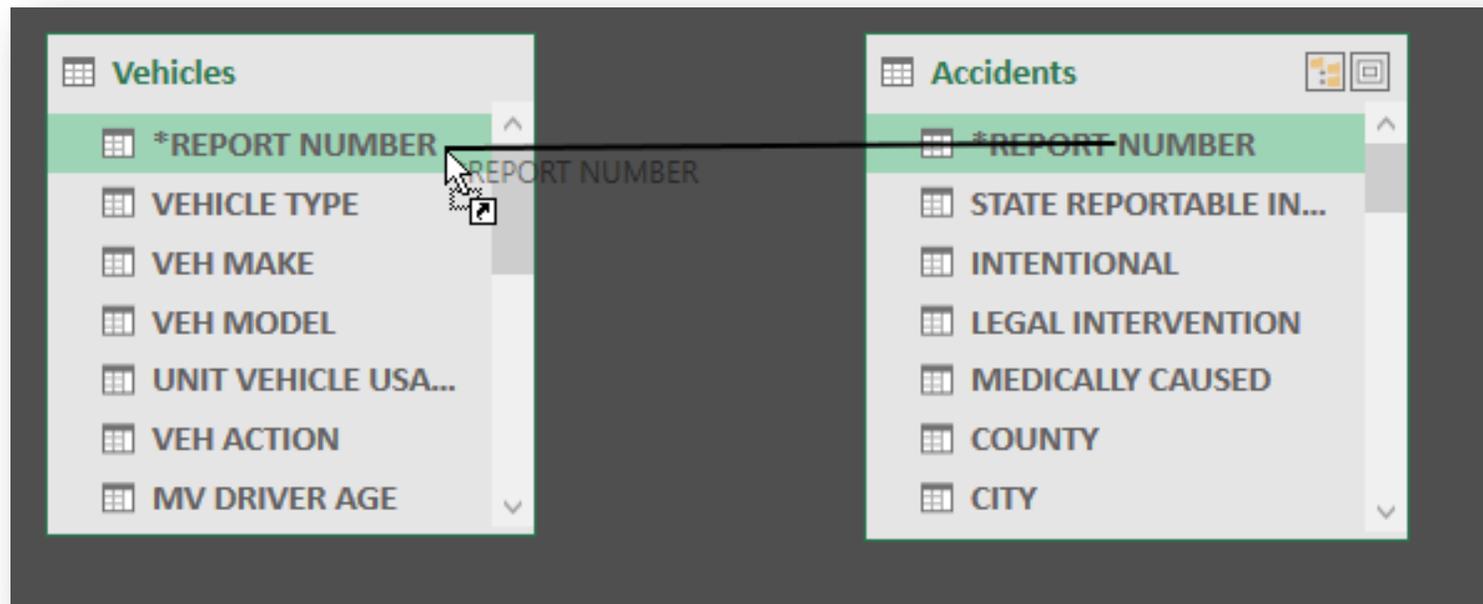
Data Models in Excel

In Power Pivot, relationships are directional. This requires that, in at least one of the two tables, the joining field only contains unique values.

What does this mean?

If you are trying to filter or aggregate a field in **table A** based on the values in **table B**, **table B** must have only unique values in the joining field. Example:

◦ If we are matching **A** and **B** by matching the values of a column named ID_number, **B** must have only a single row for each unique value of ID_number.

Data Models in Excel

Why the directionality?

In relational databases, there are three types of relationships:

- **One to one -** Each row in one table may only be matched to a single row in another table, and vice versa
- **One to many** – Each row in table A may correspond to any number of rows in table B BUT: each row in table B can only be matched to a single row in table A
- **Many to many –** Any row in one table may be matched to any number of rows in the other, and vice versa. This is **not** supported in Power Pivot/Power View

Power Pivot and Power View work on relational concepts, but are greatly simplified compared to a true database management system

Power Pivot Examples

- Got a giant text file that will not open in Excel?
- Need to take a look at the data?
- You can get data from external sources in Power Pivot
- For example, we import data from a text file in our local disk
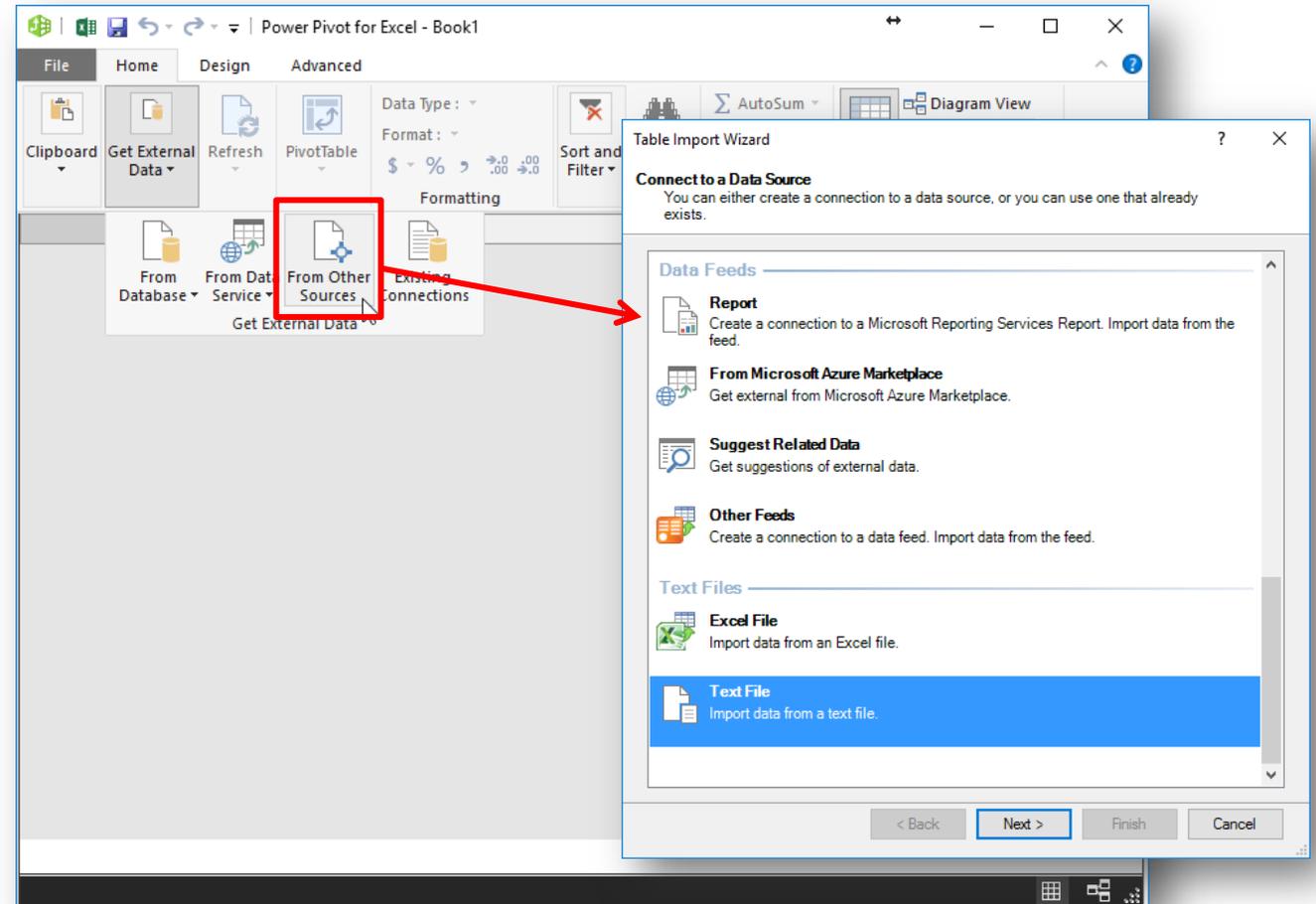
Power Pivot Examples

○ Find your text file, can be .csv, .txt, or .tab

○ You can filter rows out before importing
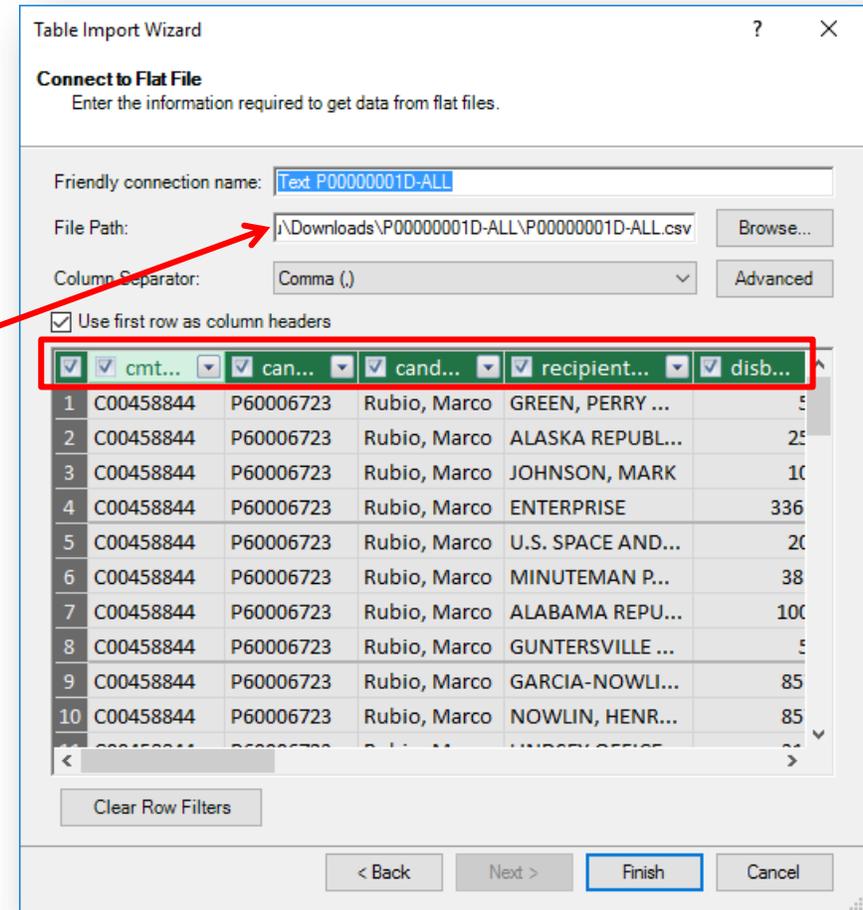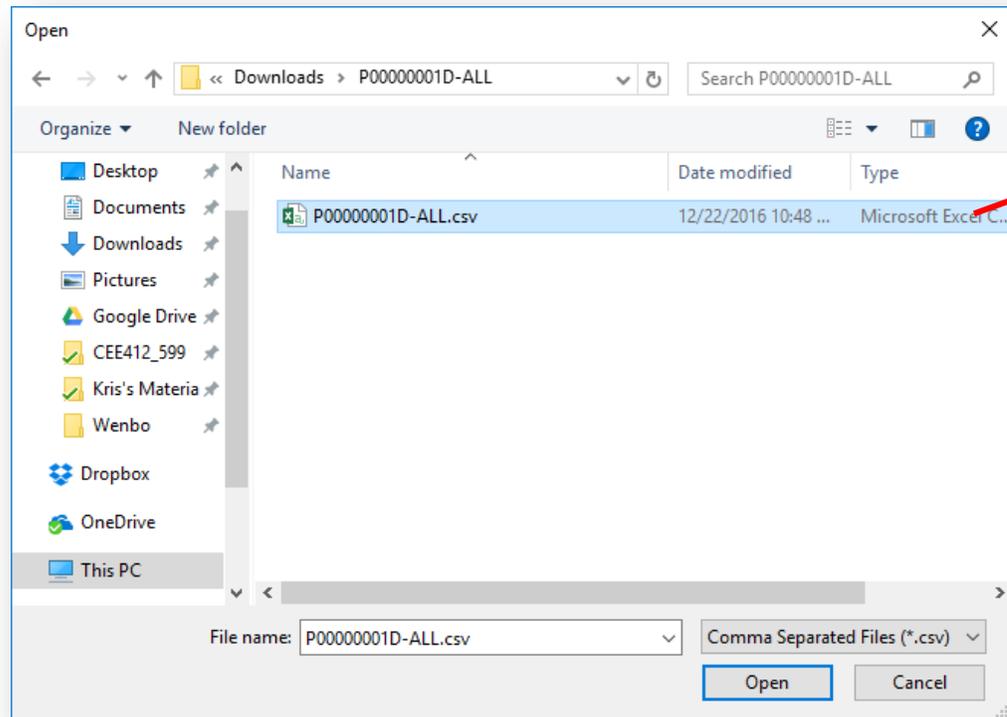
Power Pivot Examples

- You can preview the data in the Power Pivot window, or create a pivot table to summarize the data.
- In a blank Excel worksheet, click **Pivot Table** under the **Insert** menu.
- Select **Use this workbook's Data Model** as the table has already been imported. Click **OK**.
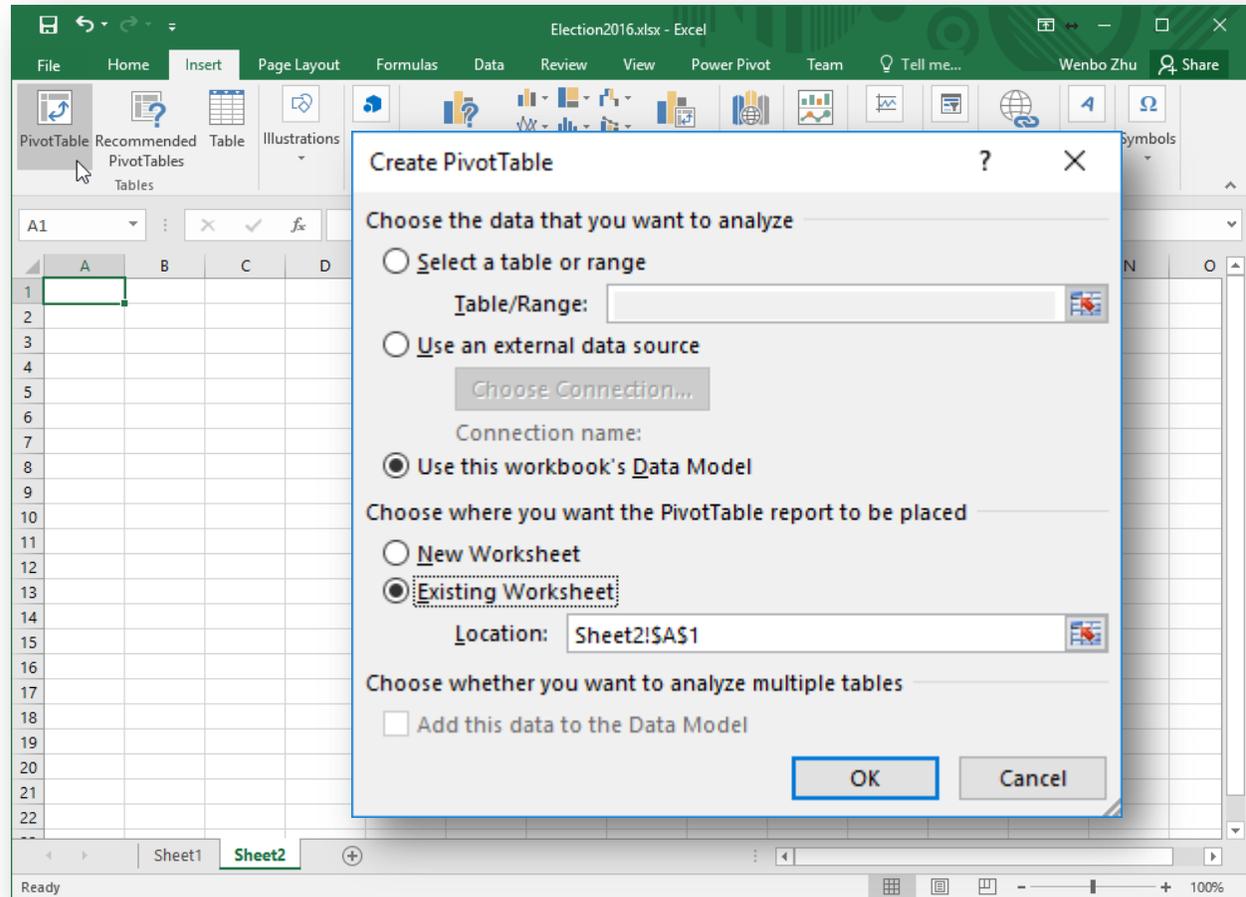
Power Pivot Examples

○ Edit the pivot table to show the desired summary statistics.

○ We will go into pivot tables in the exercise, this is just to show what is possible.

○ My dataset contained spending data for all 2016 presidential candidates. The Pivot table summarizes total spending by candidate.

○ You can also download the data from this website:

http://www.fec.gov/disclosurep/PDownload.do

○ This is a great way to view and summarize a dataset that is too large to open in Excel.

	A	B
3	**Row Labels**	**Sum of disb_amt**
4	Bush, Jeb	34,720,234.25
5	Carson, Benjamin S.	61,443,126.19
6	Christie, Christopher J.	8,563,486.21
7	Clinton, Hillary Rodham	575,881,347.57
8	Cruz, Rafael Edward 'Ted'	86,962,787.47
9	Fiorina, Carly	10,900,905.67
10	Gilmore, James S III	384,426.79
11	Graham, Lindsey O.	5,535,679.66
12	Huckabee, Mike	4,801,625.08
13	Jindal, Bobby	1,413,404.74
14	Johnson, Gary	11,328,086.34
15	Kasich, John R.	18,658,050.66
16	Lessig, Lawrence	437,679.17
17	McMullin, Evan	1,644,334.63
18	O'Malley, Martin Joseph	5,700,449.63
19	Pataki, George E.	525,604.47
20	Paul, Rand	11,862,667.26
21	Perry, James R. (Rick)	2,017,068.43
22	Rubio, Marco	45,980,545.90
23	Sanders, Bernard	226,527,256.41
24	Santorum, Richard J.	2,084,162.40
25	Stein, Jill	3,403,673.79
26	Trump, Donald J.	348,868,138.92
27	Walker, Scott	11,268,635.47
28	Webb, James Henry Jr.	821,424.81
29	**Grand Total**	**1,481,734,801.92**

PivotTable Fields

ACTIVE **ALL**

Choose fields to add to report:

Search

▲ ⊞ P00000001D-ALL
  ☐ cmte_id
  ☐ cand_id
  ☑ **cand_nm**
  ☐ recipient_nm
  ☑ **disb_amt**
  ☐ disb_dt

Drag fields between areas below:

▼ FILTERS | ▥ COLUMNS

≡ ROWS | Σ VALUES

cand_nm ▼ | Sum of disb_amt ▼

☐ Defer Layout Update | UPDATE

Power Pivot Examples

Pivot Tables are a great way to summarize and perform simple descriptive analysis.

Each column in a dataset can be used in the following way, depending on the type and format of the data:

- **Filter – set a value or range of values to include in the summary table**
  - Example: date/month can often serve as a filter, to show only data corresponding to a given time range
- **Column/Row – Each unique field will correspond to a single column/row in the summary table**
  - Example: candidate names
  - This is similar to Column, depending on how you like the layout of the Pivot Table.
- **Value – The values to summarize, take the mean, variance, count, or other aggregation**
  - Example: sum of sales or average customer spending if we are looking at the sales data of a store.

Power Pivot Examples

Power Pivot Examples

◦ Still on the candidate expenditure data, here I have added the state field into the **COLUMNS** area.

◦ The result table is a summary of candidate expenditures in different states, with candidate names on rows and states on columns (shown on next slide).

◦ Note that you can add multiple fields into COLUMNS or ROWS area to aggregate your results at different levels. But this may also produce a Pivot Table that is not readily interpretable.
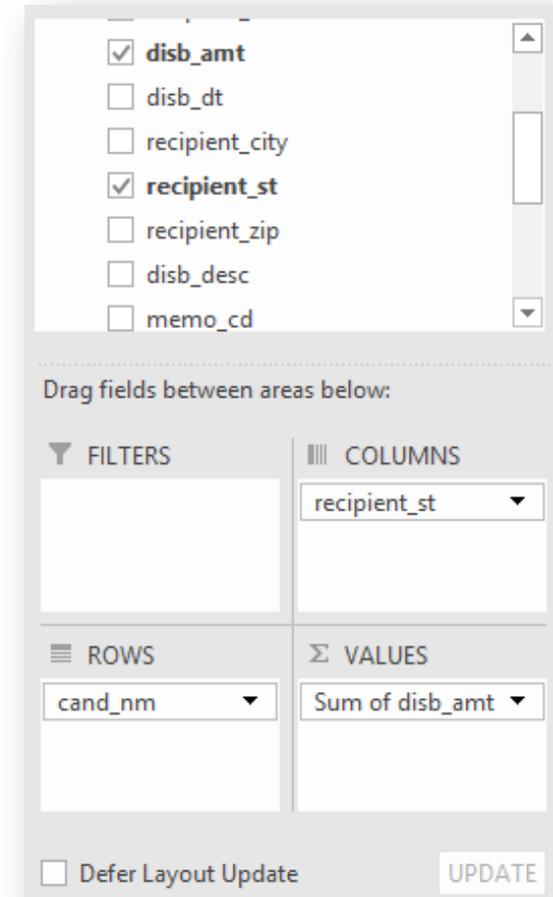
2016 Presidential Candidate Expenditure by State

Sum of disb_amt	Column Labels													
Row Labels	AB	AK	AL	AM	AR	AS	AU	AZ	BC	CA	CO	CT	D.	DC
Bush, Jeb		3,000.00	270,775.60		44,284.14			169,144.14		1,808,473.52	18,470.91	100,279.29		1,927,765.14
Carson, Benjamin S.	760.00	3,000.00	2,365,407.36	2,886.75	35,094.21			90,377.28	14,491.57	1,009,471.71	31,484.98	7,265.49		1,383,116.62
Christie, Christopher J.		2,840.70	10,000.00		25,000.00			17,739.26		909,750.68	25,507.29	12,145.97		174,978.30
Clinton, Hillary Rodham		25,974.93	2,592,226.80		16,162,879.58			222,938.23	4,432.96	16,373,746.63	5,040,136.07	10,966,897.52	2,500.00	352,970,213.18
Cruz, Rafael Edward 'Ted'		3,454.93	21,953.12	2,506.35	121,309.24		97.00	66,982.77	568.10	7,440,767.64	127,907.08	11,687.79		1,053,306.58
Fiorina, Carly		3,044.00	60,299.84		25,000.00			1,101,630.67		495,236.84	151,433.68	26,333.55		2,189,378.68
Gilmore, James S III														52,749.04
Graham, Lindsey O.			10,000.00		27,079.13			69,136.99		290,745.71	4,258.73	17,970.62		365,306.93
Huckabee, Mike		2,500.00	20,923.56		1,364,877.18			10,356.52		242,470.22	6,897.73	978.67		16,136.85
Jindal, Bobby			10,000.00		26,009.88			3,292.00		4,531.33	3,468.15			37,882.13
Johnson, Gary		8,521.30	8,417.98					273,025.53		1,688,357.82	50,508.91	8,137.82		127,061.06
Kasich, John R.		3,000.00	21,450.95		50,560.05			101,001.06		1,423,971.26	13,621.77	23,471.68		682,326.74
Lessig, Lawrence			7,500.00							146,552.10				101,001.06
McMullin, Evan								12,324.00		268,346.37	48,520.00	806.02		117,835.84
O'Malley, Martin Joseph		1,000.00	3,985.53		3,064.71			18,780.62		138,107.33	1,238.76	33,818.55		1,302,277.62
Pataki, George E.										7,456.64				5,822.50
Paul, Rand		3,960.00	25,364.98		32,206.26			47,274.83		528,226.67	17,560.00	7,791.57		514,982.23
Perry, James R. (Rick)								4,689.06		150,114.10	7,105.04	61,602.92		6,696.48
Rubio, Marco		4,000.00	20,696.09		158,905.54			330,097.89		1,415,077.92	141,988.93	23,711.79		6,783,446.97
Sanders, Bernard		77,705.20	218,911.92		193,350.55	2,500.00		735,197.23		10,634,309.55	1,003,969.27	195,343.61		51,816,796.03
Santorum, Richard J.		2,500.00	26,113.87		30,750.00			7,293.39		154,568.67	7,373.49	26,811.18		55,883.53
Stein, Jill		11,710.00	2,415.03					66,639.30		760,665.41	20,488.10	3,805.09		17,450.00
Trump, Donald J.		19,563.51	424,299.72		157,865.77			3,347,969.35		8,902,950.10	2,752,175.08	1,071,862.37		6,884,637.09
Walker, Scott			810.60		568.04			44,248.02		262,649.90	18,315.48	14,368.17		489,923.94
Webb, James Henry Jr.								5,072.02		118,356.93				174,822.26
Grand Total	760.00	175,774.57	6,121,552.95	5,393.10	18,458,804.28	2,500.00	97.00	6,745,210.16	19,492.63	55,174,905.05	9,492,429.45	12,615,089.67	2,500.00	429,251,796.80

What is Power View?

◦ Microsoft's interactive "data exploration, visualization, and presentation experience" –Microsoft.

◦ Tools for investigating and visualizing data.

◦ Can automate many of the data aggregation tasks, and has visualization capabilities beyond that of Power Pivot.

◦ Not the same as Power Pivot, but the two tools are closely related in Excel (Power View is based on data models).

◦ Browser-based, a Power View report is not an excel worksheet.

◦ Important note: save your work frequently in Power View.

Getting Started With Power View

◦ Once I have created a data model, I can click **Insert → Power View Report** to get started with Power View

◦ Note: if you have not used Power View before, you will be prompted to enable it as an add on, and possibly install Silverlight if you have not already done so.
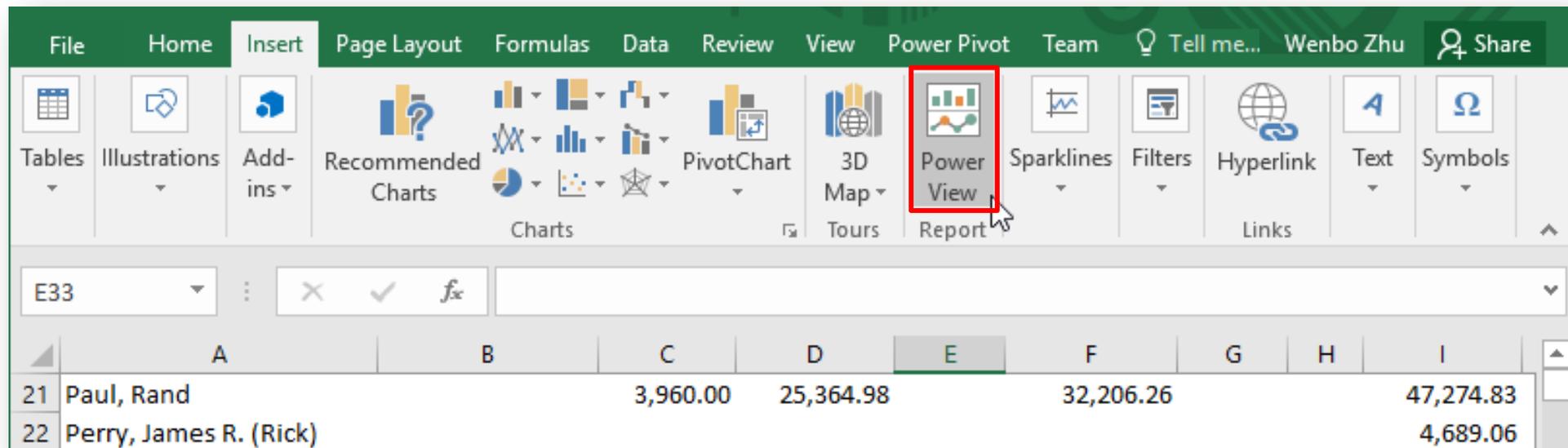
Getting Started With Power View

○ I can manage relationships within Power View, and choose visual themes for my Power View Report
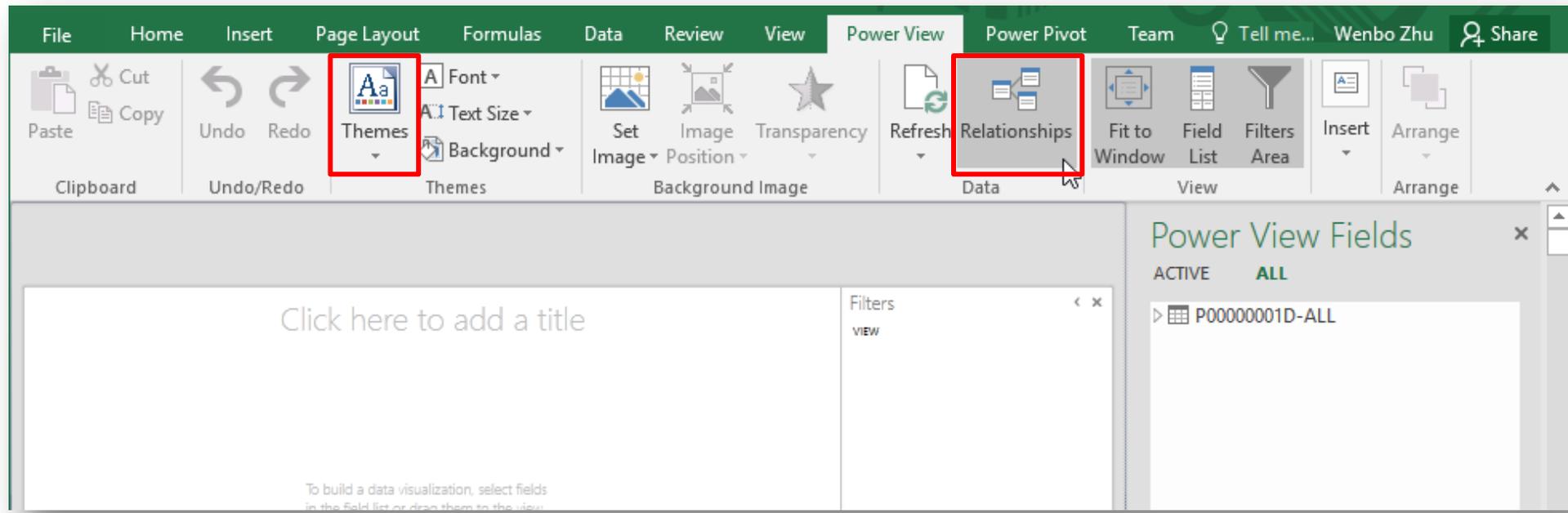
Getting Started With Power View

- You will note that, as with Pivot Tables, you can add fields from multiple tables in your data model
- Power View tried to guess what type of data is in each column (numeric, categorical, geospatial, etc.).
- You can customize the aggregation that is applied to each field after it is added.
- Once data has been added, you can insert various type of plots which will replace the data table.

Choose which fields to include in the report

Power View Report

Filter area

Getting Started With Power View

○ As an example graph, here I created a map showing the candidate spending as bar chart in each state.

○ I also filtered candidates names to only show the data with Clinton and Trump.
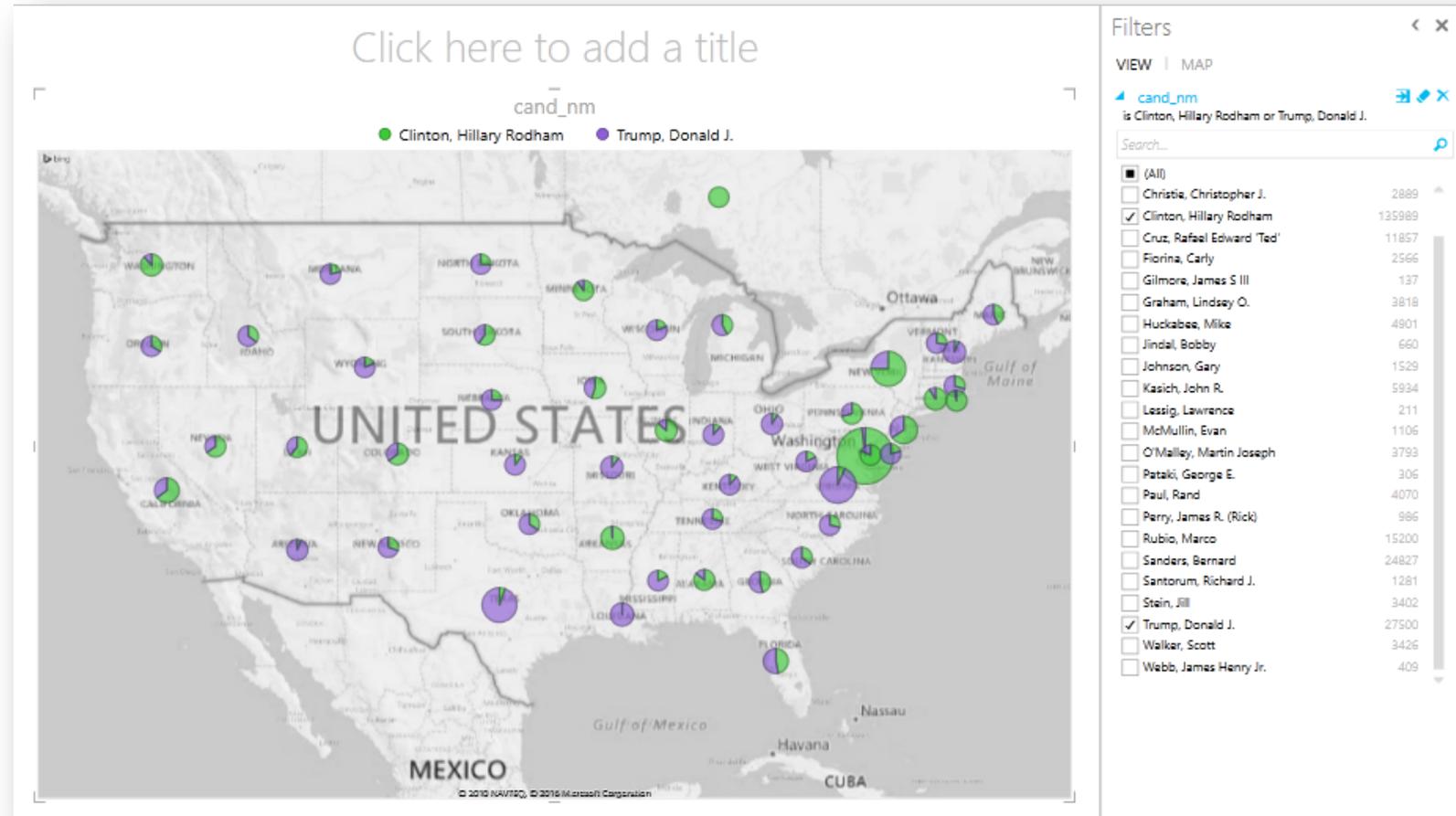
Getting Started With Power View

- When completing the exercises, do some exploring and see what you can create.

- The interface is pretty intuitive for Excel users

- Note: if you have difficulty working with the data modeling tools beyond the exercises, don't sweat it. We have not yet covered the details of how/why/when data models work.

Excel Summary

Excel is not a great solution for data management or complex analysis

It is a great tool for basic analysis, and Excel 2016 has expanded data management functionality

The real point(s):
◦ Introduced topics of data models, Power Pivot, and Power View.
◦ We showed how to do some things in Excel that are relatively new.
◦ There are much more powerful tools available to do these things and many more, which we will start talking about next lecture.