Education Writers Association 2018 National Seminar

Matthew Kauffman, The Hartford Courant
mkauffman@courant.com

◆ All files at http://bit.ly/EWA18DATA ◆

APPLYING BASIC ARITHMETIC TO CELLS (TownBudget.xlsx)

All equations in Excel begin with the equal sign. From there, you can use cell addresses as well as numbers when building an equation. So **=B2+B3** will add the values in those two cells (assuming they are numeric). Remember that when building a formula, you can type out the cell addresses or simply click on the cell you want to include. And to add up a series of cells in a column, use the **SUM** function, followed by the beginning and ending cells, separated by a colon, within parentheses - for example: **=SUM(B2:B24)**. You can also type **=SUM(** and then click and hold the mouse to define the cells to include in the equation. Finally, you can click the Sigma icon – **Σ** – from the right side of the Home tab, to automatically build an equation.

CALCULATING PERCENT CHANGE (TownBudget.xlsx)

The three letters you must never forget in computer-assisted reporting: N...O...O.
To determine the percent increase or decline between two numbers, the formula is **=(New-Old)/Old**. So if a budget, for example, increases from $150,000 in 2017 to $180,000 in 2018, the calculation is: (180000-150000)/150000 which equals 0.2, or 20 percent. Remember that using the Fill Handle will allow you to write a formula like this once and replicate down an entire column or across a row.

CALCULATING PARTS OF THE WHOLE (TownBudget.xlsx)

If you have a list of departmental budget figures in a column and the total budget in a cell at the bottom of the column of data, calculate each department's share of the entire budget by dividing each department's budget by the total budget. Automate the process by building a formula, and using "anchors" – dollar signs placed in front of the column and/or row reference in a cell address – to freeze the address of the budget total. That will allow you to copy the formula down your spreadsheet using the Fill Handle and maintain the reference to the single cell with the budget total.

SORTING (TownBudget.xlsx)

Among the most valuable Excel functions is the ability to order data by various columns. To sort a spreadsheet, first select or define all the cells you want to rearrange – being careful not to leave data out (which can be disastrous) and careful not to include data (such as totals at the bottom of your spreadsheet) that you do not want rearranged. Then, from the Home or Data tabs, select "Custom Sort" or "Sort." From the windows that opens, choose the field or fields you want to sort by, and indicate if the results should be displayed from low to high, or high to low. If your data has "headers" in Row 1 that serve as labels for the columns, click the box for "My data has headers," which will make it easier to identify the rows to sort by.

FILTERING (TestScores.xlsx)

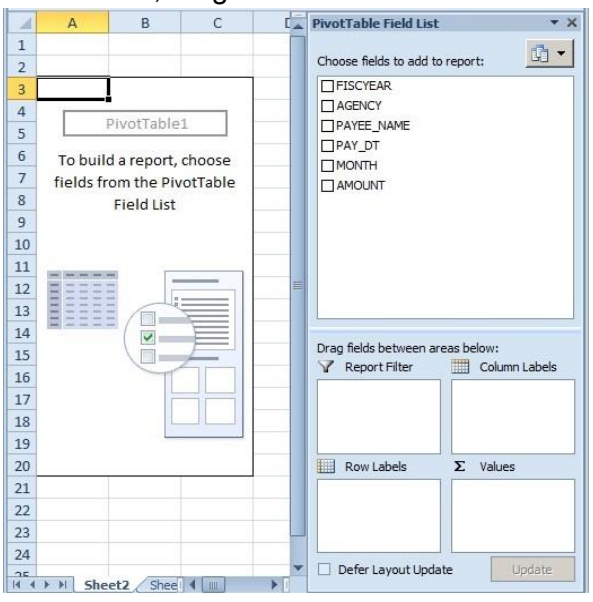
Activate Filtering from the Home tab or the Data tab, to temporarily restrict the worksheet to show only those rows that meet certain criteria. Remember that filters can be applied to numbers, text or dates, and that you can filter for individual values, a range of values, or a particular number or percent of the top or bottom values. And, of course, filters can be applied to multiple columns. It's a handy tool for seeing into your data.

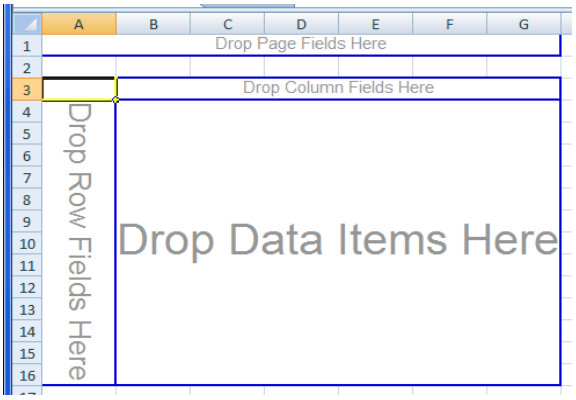
PIVOT TABLES (SDE_FY17.xlsx)

Pivot tables give Excel some of the power of a database manager by quickly performing aggregate functions on large datasets, allowing you to group and count and total your data in various ways that help you see into the numbers. In our example, we had a database of checks written by a government agency, and used a pivot table to determine how much was paid to individual vendors. Pivot tables can also have two axes, showing, as in our example, how much each company received in total – broken down by month.

To create a pivot table, define the entire worksheet and select PivotTable from the Insert tab. In contemporary versions of Excel, that will bring up an empty pivot table template on the left and the Pivot Table Field List on the right (as shown in the image at left below). As noted in the presentation, particularly if you're new to pivot tables, I recommend right-clicking on the pivot table template, choosing PivotTable Options, and from the Display tab, put a checkmark in the box for Classic PivotTable layout. That will give you the design shown at bottom-right here, which I think is a more intuitive layout that allows you to visualize how you want to build the table.

To build the table, drag the fields from the Field List either into the pivot table or into the boxes below the field list that correspond to various sections of the table. At a minimum, remember that you will typically have at least a field in the first column (the "Row Labels" box in the default layout, or the "Drop Row Fields Here" section in the classic layout) and values for those elements in the large data section of the pivot table (the "Values" box or the "Drop Data Items Here" area). Remember also that you can change the calculation of the value – displaying a sum, a count, or other choices. To create a second axis, drag a field either to the "Column Labels" box or to the thin section in Row 3 of the classic layout that is labeled "Drop Column Fields Here." To add yet another dimension, place a field in the "Report Filter" box or the "Drop Page Fields Here" section of the classic layout, and use that to filter the data that are included in the table.

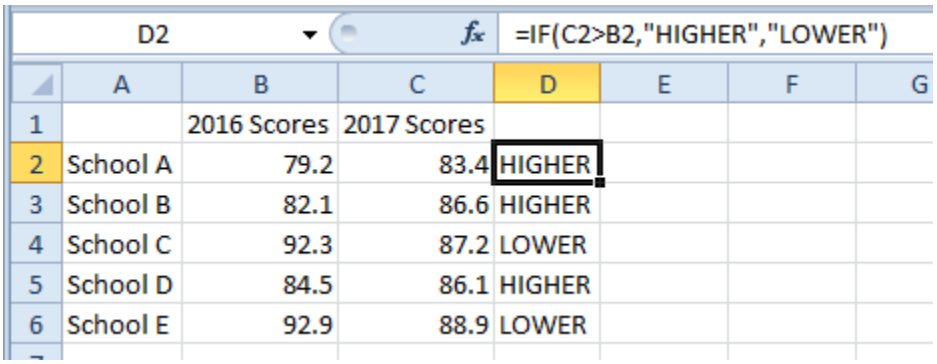




IF STATEMENTS (CountyPopulations.xlsx)

Excel formulas built with IF statements allow you to ask a question of the contents of a cell, and fill a new cell with data based on the answer to that question. The basic formula is:

=IF(*your question*,*value if the answer is yes*,*value if the answer is no*)



	A	B	C	D	E	F	G
1		2016 Scores	2017 Scores				
2	School A	79.2	83.4	HIGHER			
3	School B	82.1	86.6	HIGHER			
4	School C	92.3	87.2	LOWER			
5	School D	84.5	86.1	HIGHER			
6	School E	92.9	88.9	LOWER			

In the example above, the formula in D2 (which you can see in the formula box at the top of the spreadsheet) "asks" whether the score in C2 is greater than the score in B2. If it is, Excel places "HIGHER" in D2. If it isn't, it places "LOWER" in the cell. Then drag the formula down the sheet. (Note that this particular example doesn't account for ties, nested IF statements can be written in cases with more than two possible answers.)

STRING AND DATE FUNCTIONS (StringFunctions.xlsx)

Parse text and dates using common functions:

- LEFT, MID, RIGHT for text. (Example: **=MID(A2,8,3)** will return three characters, starting with the eighth character, from the text in Cell A2.)
  Also check out TEXT(), LEN() and FIND()
- YEAR, MONTH, DAY for dates.
  Also check out DATEDIF() and WEEKDAY().

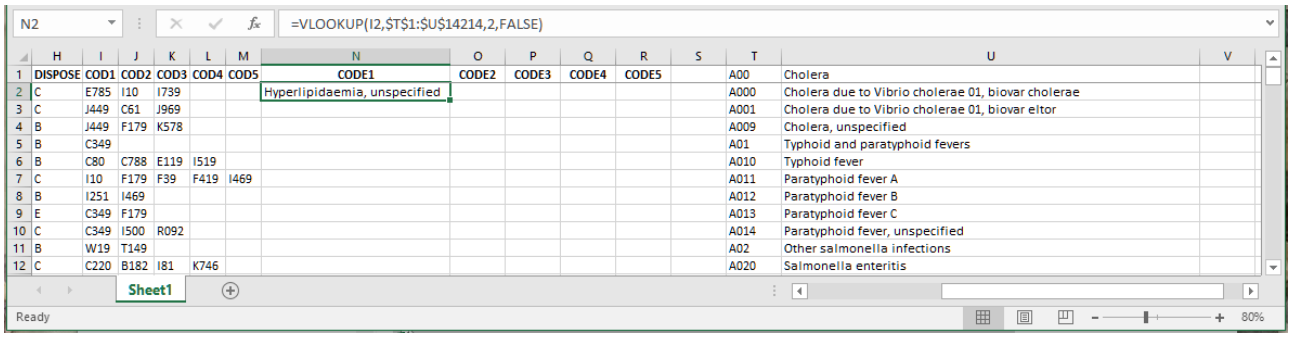
Also parse text strings with "Text to Columns" from the Data tab – splitting a column of text across multiple columns, using either a delimiter in the text, or breaking the column into fixed-width strings.

MERGING LOOKUP TABLES (DeathData.xlsx)

Join data stored in two Excel files by using VLOOKUP.
With death-certificate records, for example, we had multiple columns with disease codes, all of which are further explained in a separate translation table with the code in one column and the disease in another column.
Add the translations this way:
1. Create new column headers where the code translations will go
2. Paste the two columns from the translation table into the original table, to the right of the new columns.
3. Translate the codes into their text equivalent with this formula:
=VLOOKUP(*cell in the original spreadsheet containing the code,range of cells containing the code and translations,number indicating which column in the range contains the translations*). Then add a

comma and “FALSE” to indicate that you only want VLOOKUP to return a value where the codes match exactly.

So for Cell N2 below, the formula we have (which is visible in the formula box at the top of the sheet in this image) contains four elements:
=VLOOKUP – the function name
I2 – the cell address containing the untranslated code
$T$1:$U$14214 – the range of data containing a) the code we’ll match to our original data; and b) the translation associated with that code.
2 – indicating the data we want is the second of the two columns in our range
FALSE – An optional indicator that Excel should return a value only an exact match.
Use the fill handle to drag the formula across the screen to the other Code columns and then down the spreadsheet. (The anchors for the cells from Rows X and Y assure that the range with the lookup codes and diseases remains the same as we copy the formula throughout the spreadsheet. Then define the newly created columns and use Copy→Paste Special→Values to lock in the values.



	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	DISPOSE	COD1	COD2	COD3	COD4	COD5	CODE1	CODE2	CODE3	CODE4	CODE5		A00	Cholera	
2	C	E785	I10	I739			Hyperlipidaemia, unspecified						A000	Cholera due to Vibrio cholerae 01, biovar cholerae	
3	C	J449	C61	J969									A001	Cholera due to Vibrio cholerae 01, biovar eltor	
4	B	J449	F179	K578									A009	Cholera, unspecified	
5	B	C349											A01	Typhoid and paratyphoid fevers	
6	B	C80	C788	E119	I519								A010	Typhoid fever	
7	C	I10	F179	F39	F419	I469							A011	Paratyphoid fever A	
8	B	I251	I469										A012	Paratyphoid fever B	
9	E	C349	F179										A013	Paratyphoid fever C	
10	C	C349	I500	R092									A014	Paratyphoid fever, unspecified	
11	B	W19	T149										A02	Other salmonella infections	
12	C	C220	B182	I81	K746								A020	Salmonella enteritis	

CLEANING DATA

Many of these tools – IF statements, filtering, string functions – as well as sorting and auto-filling, can be used to wrangle dirty data. For a powerful alternative, spend some time with Open Refine. It has a bit of a learning curve, but particularly for large datasets with lots of problems, it has some intuitive features that isolate problem areas and allow you to fix them efficiently.

And for a quick way to normalize names – often the messiest of data fields – Mr. People, from Matt Ericson of the New York Times, is surprisingly good.

IMPORTING DATA (CT-Payroll-FY2017.txt and Alaska_Weather_Stations.txt)

Sometimes we work with home-grown databases, but at least as often, we’re working with datasets created by others. Sometimes those databases come in a format that Excel (or a database manager) can read, such as a comma-separated value (.csv) file. But other times – typically when a file contains straight text, or for a table we find on the Web, or for data contained in a pdf – we’ll need help getting the data into Excel.

For a text file, examine the file to determine the best way to identify column breaks, either with a "delimiter" or as "fixed-width." Then launch Excel and choose "From Text" from the "Get External Data" box on the "Data" tab. Choose a delimiter or place separators for a fixed-width file, and import.

There are a variety of ways to import data from a table on the web. Sometimes, simply copying and pasting works. Excel also has a built-in Web query tool (click "From Web" in the "Get External Data" box on the Data tab), but that method – once the only option – gets less use now that most browsers have either built-in programs or add-ons that can copy tables in a format compatible with Excel.
With Internet Explorer, right-click on an Internet table and choose "Export to Microsoft Excel."
With Firefox, try the extension "Table to Excel" (it's free).
With Chrome, try the extensions "ColumnCopy" or "Table Capture" (also free).

The IE method automatically exports the data to Excel (when it works, anyway). The Firefox and Chrome versions copy the table to the clipboard, for pasting into an Excel worksheet or other spreadsheet program.

WRANGLING PDFs (schoollunch.pdf)

To extract data in a pdf table:
First: Remember that most data in pdfs exists in a database somewhere, so consider asking for the data behind the annoying pdf format.
If that doesn't work, there are a variety of programs that convert pdfs to other file formats, and some have the ability to apply optical-character recognition, or OCR, to image pdfs. There are also a number of online pdf converters, including www.cometdocs.com and www.onlineocr.net.
For times when that doesn't work, consider pdftotext, using this method:
Download pdftotext from http://www.foolabs.com/xpdf/download.html (choosing one of the "precompiled binaries") and place in a folder directly on the C: Drive.
Unzip and install the program.
Copy the pdf into the folder containing the pdftotext program.
Go into DOS mode by typing "cmd" in the search box at the bottom of the Windows menu
In the command-prompt box, change to the correct folder by typing: cd*name_of_folder.*
Type: pdftotext *pdf_file_name* -layout. (Include the .pdf extension in the file name.)
That will create a text file in the same folder. Use the import tips above to bring the file into Excel, and use string functions to clean up any messy data.

CREATING AN INTENSITY MAP WITH GOOGLE FUSION TABLES

(These are step-by-step notes for the map we built, which you can retrace on your own as a refresher)

1. Download PerCapitaIncome.xlsx from http://bit.ly/EWA18DATA
2. Import into Google Drive
    Launch/log in to Google Drive (at drive.google.com)
    New→More→Google Fusion Tables
        If you don't have "Google Fusion Tables" as an option:
        - *Click "+ Connect more apps"*
        - *Search "Fusion Tables"*
        - *From the choice "Fusion Tables (experimental)," click "+ Connect"*

    Click "Choose File," browse to PerCapitaIncome.xlsx, double-click and hit Next
    Confirm column names ("State", "Income") are in Row 1, by clicking Next
    Accept defaults by clicking Finish
    File will be imported. Note yellow highlight for geographical data
3. Show on map
    Click "Map of STATE" tab
    Wait for geocoding to complete and for map to appear
        If there's a problem with Georgia, we can fix that
        - *Click the "Rows 1" tab*
        - *Click on "Georgia" then click the pencil icon to bring up the Edit box*
        - *Click "edit geocode…"*
        - *In the "Location description" box, type "Georgia, USA" and hit Enter*
        - *Click the red symbol over Georgia and click "Use this location"*
        - *Click Save*
        - *Click "Map of STATE" to confirm the dot is now over Georgia*

    Note that a data window appears when dots in individual states are clicked
    You now have a map with points – good if you want to pinpoint individual locations (homicides, banks, fast food joints), but we want a "polygon" map showing the outline of the states that we can shade in different colors.
4. Build a customizable "intensity map" with these steps:

4a. Import, Merge KML data
    Select File→Find a table to merge with…
    In the pop-up box, type US States kml boundaries
    Scroll through the choices, looking for an option that notes a 100% match
    Click "view table" for your selection to preview the table. Look for a column of matching state names and a column labeled "geometry."
    Close the browser tab with the new table, to return to the popup box
    Click the check box for the table you have selected and click Next
    Confirm that the columns with matching state names have been selected
    Click Next
    Select the columns that will appear in the merged table. At a minimum,

keep the original columns from the PerCapitaIncome file, plus the geometry column from the new table.

Click Merge

Click on “View table” or the link for the merged table

Note that the new table now includes a “geometry” column for each state.

Click the “Map of geometry” tab

Note that instead of dots, there are now filled red polygons for each state

(If you still see dots, zoom into the map)

Click on a state to see “info window” information

4b. Customize polygons

Click “Change feature styles…”

Click Polygons→Fill color

Two choices. First: Click “Gradient” tab

Click radio button to “Show a gradient”

Confirm that “Income” is selected in the box next to “Column”

Click “use this range” next to the dollar figures noting the high and low values

Click Save and intensity map appears

For another option, click again on “Change feature styles…”

Click “Buckets” tab

Click radio button and use pull-down arrow to divide into 3 buckets

Click “use this range” (but note that bucket ranges can be customized)

If desired, select custom color for each bucket

Click Save

New map appears, reflecting your choices

4c. Customize borders

Click “Change feature styles…”

Click Polygons→Border color

Choose border color (black makes the states stand out)

Click Save to view map

4d. Customize Info Window

Select “Change info window…”

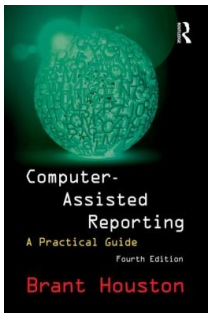
Select Automatic tab to delete particular fields, or

Select Custom tab to edit individual fields with basic html

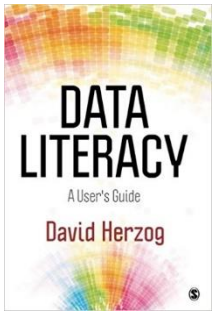
Click Save to view map

5. Click Tools → Publish… to retrieve embeddable link for publishing to web

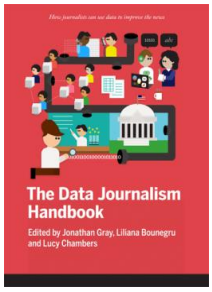
A SHORT LIST OF RESOURCES FOR DATA JOURNALISM



Computer Assisted Reporting: A Practical Guide
http://www.ire.org/carbook/
For years, this was the top resource for understanding key CAR skills. But it's very pricey.



Data Literacy: A User's Guide
https://www.amazon.com/Data-Literacy-David-L-Herzog/dp/1483333469
A little wonky in places, but comprehensive. It's used as a textbook, so it's also on the expensive side.



The Data Journalism Handbook
http://datajournalismhandbook.org/1.0/en/
More theoretical than hands-on, and tilts European, but worth a look. (Plus, it's free!)

The Data Journalism Blog
http://www.datajournalismblog.com/
Blog covering various CAR issues. Global perspective. Heavy on data visualization.

NICAR-L listserv
http://www.ire.org/resource-center/listservs/subscribe-nicar-l/
The premiere hangout for data nerds. Ask a question; you'll get an answer.

Matthew Kauffman
The Hartford Courant
mkauffman@courant.com
@matthewkauffman