

Tidy up your spreadsheets

Sarah Cohen / Columbia University, Stabile Program / Spring 2014

The idea of “tidy” data is that each column means one thing, and each row is an instance of that one thing. It means you’d rarely type two separate things into one cell, such as two names or two addresses. It also means you’ll usually have a dataset that is tall and skinny with very few blanks, not short and fat. It’s not as nice a printout, but it will be possible to sort and filter rows and summarize groups.

More neatness

Tidy data also implies understanding what you’ll want to do with your records. It’s almost always easier to put tobits of information together in a computer than to pull out pieces.

Consider:

- Using the same method each time to enter names. This will usually be Last Name, First Middle Suffix. That is good enough to make it relatively easy to match against other sources and to sort.
- Let Excel anticipate your keystrokes to keep columns consistent.
- Use separate columns to flag items of interest. It doesn’t mean you can’t use colors to highlight, but it’s hard to use the colors as “data”. (Tip: conditional formatting lets you color rows based on a flag, but it’s hard to filter for rows of a certain color.)

Import the right way

You’ll sometimes get so-called “csv” files from sources, which are just text in columns separated by commas, like this:

```
School name, Grades, Year, City, State, Zip, latest review
"The Garden School", 3-4, 2012, New York, NY, 10024, 2012-01-23
"The Boston School", 9-12, 2012, Boston, MA, 01323, 1/14/2013
"The last school", K-12, Los Angeles, CA, 90123-0123, 2013
```

It’ll have an Excel icon next to it, suggesting that you can just double-click to open it. It’ll open, but... it’ll sometimes be messed up and you can’t do anything about it.

School name	Grades	Year	City	State	Zip	latest review
The Garden School	4-Mar	2012	New York	NY	10024	2012-01-23
The Boston School	12-Sep	2012	Boston	MA	1323	1/14/2013
The last school	K-12	2012	Los Angeles	CA	90123-0123	2013

Remember, that anything you see that’s left-justified is seen as text. Anything right-justified is seen as a number, including dates. There are several problems here. First, Excel thinks that grades 3 through 4 really means March 4th. It has turned zip codes into numbers when they are 4 digits, removing the leading zero on the Boston address. And it hated the mixed date formats (which it will always hate.)

Get used to importing the right way, and you'll find yourself with fewer headaches later on. There are two ways to import correctly. You can either use the File, Import command to bring up the parser, or you can use the data import command on an existing worksheet. I usually do the latter for two reasons. The main one is that you can always disconnect your worksheet from the underlying data, but you don't risk making changes to the original dataset by saving again as a CSV. The second is that you can refresh your import if the data changes or is updated.

Navigating your spreadsheet

Try to avoid using your mouse when you are trying to navigate your spreadsheet. Here are a few tricks for finding and editing more quickly and accurately.

Delete a column or row

Select the row by clicking on its heading, and type CTL-D. This is different than just blanking it out. This adjusts all the formulas in the spreadsheet and removes the blank column or row.

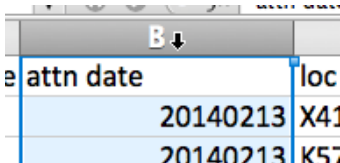
Insert a column or row

Same thing, just CTL-I.

Move a column or row

This is the process of cutting an entire row or column, then getting to an "insert cut cells" menu. If Excel doesn't understand what you want done, it will ask. It's safest, though, to only do this on entire columns and rows.

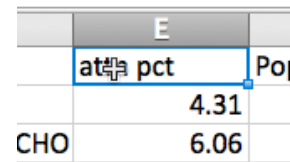
First select the entire column or row. As an example, click on the letter of the column header:



A screenshot of an Excel spreadsheet. Column B is highlighted in blue. The header row shows 'attn date' in column B and 'loc' in column C. Below the header, there are two rows of data: '20140213' in column B and 'X41' in column C, and '20140213' in column B and 'K57' in column C.

Use whatever method you want to cut the column – Ctl-X, cmd-X or right-click and choose "Cut". Select the cell at the top of the column where you want it to reside.

Now insert the cut cells by right-clicking and choosing that option, or using the shortcut keys. On a Mac, it's Ctl+I. On Windows, it's CTL+Shft+I

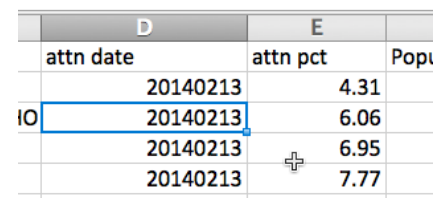


A screenshot of an Excel spreadsheet. Column E is highlighted in blue. The header row shows 'attn pct' in column E and 'Pop' in column F. Below the header, there are two rows of data: '4.31' in column E and '6.06' in column F, and 'CHO' in column E and '6.06' in column F.

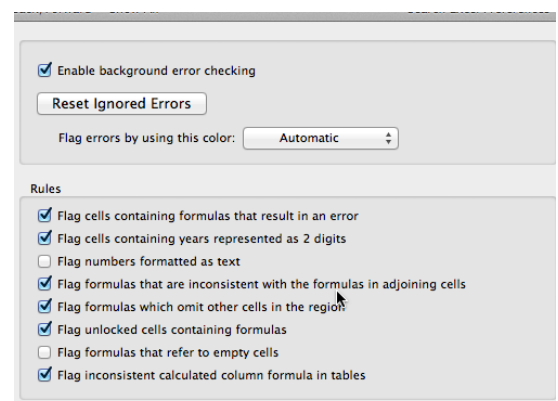
There are two advantages to this method rather than manually inserting a row and copying and pasting. First, all of the formulas automatically adjust to the new position. Second, it automatically shifts all of the data left or right, adjusting their formulas as well.

(Yes, it's now in column D not column E, but it's in the position you wanted. It's just that you've removed the old column B and shifted everything to the left.)

The same thing works with rows, though it's rarely useful.



A screenshot of an Excel spreadsheet showing the result of moving column B to column D. The header row shows 'attn date' in column D, 'attn pct' in column E, and 'Pop' in column F. Below the header, there are four rows of data: '20140213' in column D and '4.31' in column E, '20140213' in column D and '6.06' in column E, '20140213' in column D and '6.95' in column E, and '20140213' in column D and '7.77' in column E.



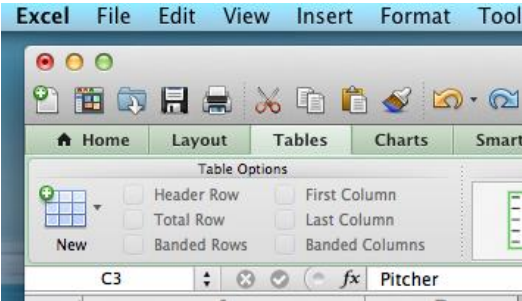
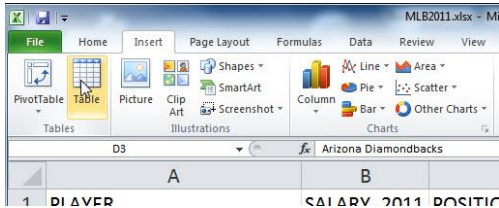
A screenshot of the Excel 'Formulas' tab, specifically the 'Error Checking' section. The 'Enable background error checking' checkbox is checked. Below it is a 'Reset Ignored Errors' button. The 'Flag errors by using this color:' dropdown is set to 'Automatic'. Under the 'Rules' section, several checkboxes are checked: 'Flag cells containing formulas that result in an error', 'Flag cells containing years represented as 2 digits', 'Flag formulas that are inconsistent with the formulas in adjoining cells', 'Flag formulas which omit other cells in the region', 'Flag unlocked cells containing formulas', and 'Flag inconsistent calculated column formula in tables'. The 'Flag numbers formatted as text' checkbox is unchecked.

Turn off some annoying error flags

Excel has two warning flags that don't really help you and can be distracting. In one, it says that you've treated a number as text. That's good in many cases, and you don't want it yelling at you about it. The other is when there are blank cells in a region. You can choose which errors to show in Preferences (Mac) or Options (Windows).

Treat your data area as a table

Excel doesn't really tell you that formatting your data as a table comes with many consequences. Many of them are good.

Mac	Windows
<p>In the Tables area in the ribbon, choose "New" next to the picture of a table.</p> 	<p>On the Home menu, look for "Format as Table". Alternatively, under Insert, choose "Table".</p> 

To create a table, choose a cell in your data region and look for the Tables tab on your ribbon. Choose one of the formats, and it will automatically find the area bounded by a blank column and a blank row.

This spreadsheet lists some schools' attendance from the snowstorm of Feb. 13, 2014. (It isn't the whole file, just a piece for example purposes.) If you wanted to create an average by district, and borough, you'd need formulas: the first two digits of the code (district), the third character (the borough) and the number of students attending. Here's what these three formulas look like:

	A	B	C	D	E	F	G	H	I
1	full code	loc cod	School name	attn date	attn pct	Population	attendance	district	borough
2	32K403	K403	ACADEMY FOR ENVIRONMENTAL LEADERSHIP	20140213	21.15	352	= (E2/100)*F2	=LEFT(A2,2)	=MID(A2,3,1)
3	11X270	X270	ACADEMY FOR SCHOLARSHIP AND ENTREPRENEUR	20140213	19.3	437			

But there is some missing data in the population column, so when we double-click on the black cross, it only copies the first few rows:

	A	B	C	D	E	F	G	H	I
1	full code	loc cod	School name	attn date	attn pct	Population	attendance	district	borough
2	32K403	K403	ACADEMY FOR ENVIRONMENTAL LEADERSHIP	20140213	21.15	352	74.448	32	K
3	11X270	X270	ACADEMY FOR SCHOLARSHIP AND ENTREPRENEUR	20140213	19.3	437	84.341	11	X
4	18K589	K589	ARTS MEDIA PREPARATORY ACADEMY	20140213	19.34	298	57.6332	18	K
5	27Q400	Q400	AUGUST MARTIN HIGH SCHOOL	20140213	18.45	838	154.611	27	Q
6	08X530	X530	BANANA KELLY HIGH SCHOOL	20140213	19.12	364	69.5968	08	X
7	27Q410	Q410	BEACH CHANNEL HIGH SCHOOL	20140213	14.77				
8	13K575	K575	BEDFORD STUYVESANT PREPARATORY HIGH SCHO	20140213	6.06	120			

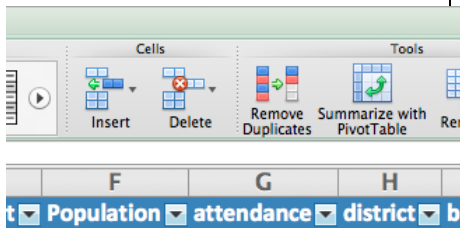
What happens, though, when we format the area as a table first?

	D	E	F	G	H
	attn date	attn pct	Population	attendance	district
P	20140213	21.15	352	$= (e2/100)*f2$	

	D	E	F	G
	attn date	attn pct	Population	attendance
SHIP	20140213	21.15	352	74.448
PRENEUR	20140213	19.3	437	84.341
	20140213	19.34	298	57.6332
	20140213	18.45	838	154.611
	20140213	19.12	364	69.5968
	20140213	14.77		0
GH SCHO	20140213	6.06	120	7.272
CHOO	20140213	16.46	274	45.1004

Once you hit Enter, Excel knew to copy it to the rest of the table. You also see warnings that some calculations have been made on missing values.

There are some things you can't do with tables, so if you want to convert it back to a normal range, go to the Tables tab or ribbon and look for "Convert to Range." You can also use this tab to create pivot tables, insert and delete table columns, and remove duplicates.

Mac	Windows
<p>In the Tables area in the ribbon, choose "New" next to the picture of a table.</p> 	<p>Look all the way to the right for "Table Tools". (Pivot tables are under "Insert" and "Data", but not here.)</p> 