

# Using Excel 2011 for Mac to do data journalism

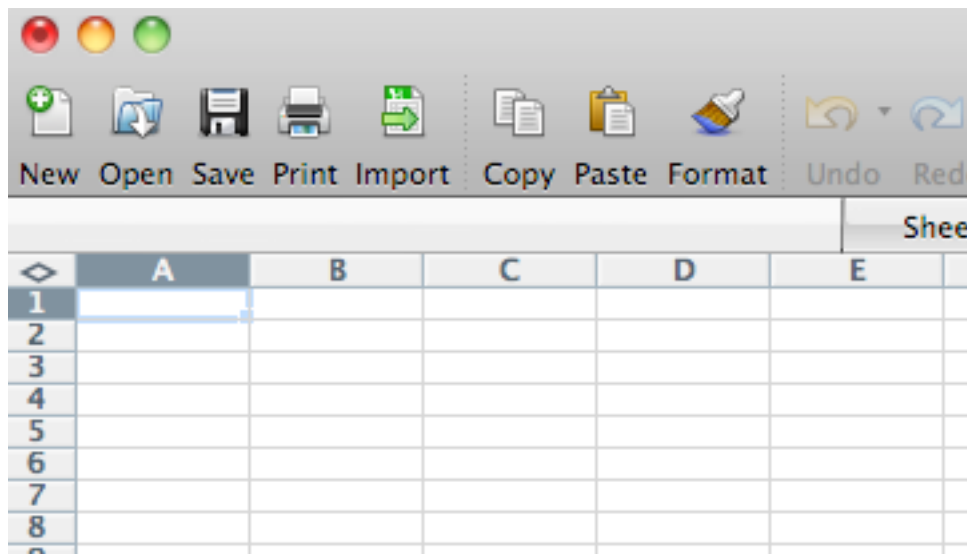
Steve Doig (steve.doig@asu.edu)

Microsoft Excel is a powerful tool that will handle most tasks that are useful for a journalist who needs to analyze data to discover interesting patterns. These tasks include:

- Sorting
- Filtering
- Using math and text functions
- Pivot tables

## INTRODUCTION TO EXCEL

Excel will handle large amounts of data that is organized in table form, with rows and columns. The columns (which are labeled A, B, C...) list the variables (like Name, Age, Number of Crimes, etc.) Typically, the first row holds the names of the variables. The rest of the rows are for the individual records or cases being analyzed. Each cell (like A1) holds a piece of data.



Modern versions of Excel will hold as many as 1,048,576 records with as many as 16,384 variables! An Excel spreadsheet also will hold multiple tables on separate sheets, which will be tabbed on the bottom of the page.

36	Friuli-Venezia Giulia	Trieste	10557
37	Liguria	Imperia	12616
38	Liguria	Savona	16952
39	Liguria	Genova	70072

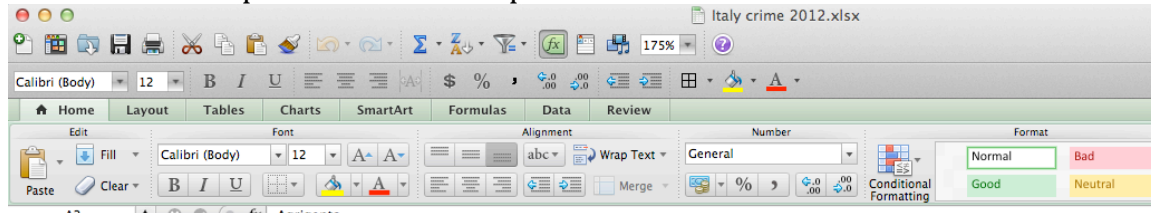
Summary	Ger 1 Tav 1a	Delitti denunci	Ger 1
---------	--------------	-----------------	-------

Normal View Ready

## SORTING

One of the most useful abilities of Excel is to sort the data into a more revealing order. Too often, we are given lists that are in alphabetical order, which is useful only for finding a particular record in a long list. In journalism, we usually are more interested in extremes: The most, the least, the biggest, the smallest, the best, the worst.

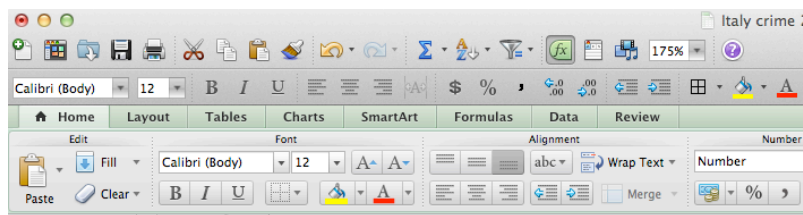
Consider the data used in this workshop, a list of the provinces of Italy showing the number of various kinds of crimes reported during a recent year. Here is how it looks sorted in alphabetical order of province name:



The screenshot shows the Excel interface with the 'Italy crime 2012.xlsx' file open. The ribbon includes Home, Layout, Tables, Charts, SmartArt, Formulas, Data, and Review. The 'Home' ribbon is active, showing Font, Paragraph, and Styles groups. The table data is as follows:

	A	B	C	D	E	F	G	H	I
1	province	territory	population	murders	rapes	burglary	car theft	robberies	drugs
2	Agrigento	Sicilia	454,002	2	4	118	61	16	28
3	Alessandria	Piemonte	440,613	2	32	2,545	359	144	213
4	Ancona	Marche	481,028	0	28	2,152	196	141	333
5	Aosta	Valle d'Aosta / Vallée d'Aoste	128,230	1	15	470	32	20	60
6	Arezzo	Toscana	349,651	2	27	1,088	131	76	193
7	Ascoli Piceno	Marche	214,068	0	6	565	132	41	105
8	Asti	Piemonte	221,687	1	19	1,412	183	94	45
9	Avellino	Campania	439,137	3	14	1,056	324	72	126
10	Bari	Puglia	1,258,706	19	68	4,222	6,268	1,323	553
11	Belluno	Veneto	213,474	0	15	366	26	15	66
12	Benevento	Campania	287,874	1	18	727	239	73	82

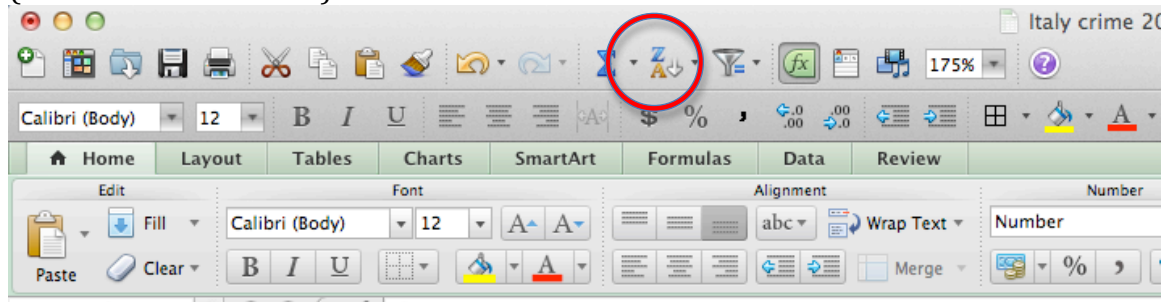
Far more interesting would be to sort it in descending order of the total number of murders, with the most violent city at the top of the list:



The screenshot shows the Excel interface with the 'Italy crime' file open. The ribbon includes Home, Layout, Tables, Charts, SmartArt, Formulas, Data, and Review. The 'Home' ribbon is active, showing Font, Paragraph, and Styles groups. The table data is as follows:

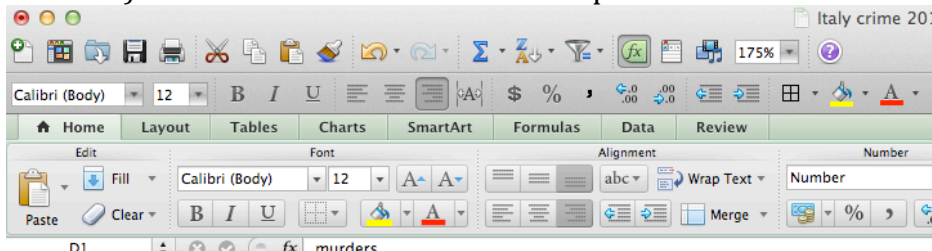
	A	B	C	D	E	F
1	province	territory	population	murders	rapes	burglary
2	Napoli	Campania	3,080,873	62	169	4,64
3	Roma	Lazio	4,194,068	37	422	15,27
4	Milano	Lombardia	3,156,694	28	470	18,09
5	Reggio di Calabria	Calabria	566,977	25	34	1,05
6	Torino	Piemonte	2,302,353	20	181	13,50
7	Bari	Puglia	1,258,706	19	68	4,22
8	Foggia	Puglia	640,836	16	33	1,63
9	Caserta	Campania	916,467	14	58	2,16
10	Catania	Sicilia	1,090,101	12	58	4,02

There are two methods of sorting. The first method is quick and can be used for sorting by a single variable. Put the cursor in the column you wish to sort by ("Murders" in this case) and then click the Z-A button:



	A	B	C	D	E	F
1	province	territory	Total population	murders	rapes	burglary
2	Agrigento	Sicilia	454,002	2	4	118
3	Alessandria	Piemonte	440,613	2	32	2,545
4	Ancona	Marche	481,028	0	28	2,152
5	Aosta	Valle d'Aosta / Vallée d'Aoste	128,230	1	15	470
6	Arezzo	Toscana	349,651	2	27	1,088
7	Ascoli Piceno	Marche	214,068	0	6	565
8	Asti	Piemonte	221,687	1	19	1,412
9	Avellino	Campania	439,137	3	14	1,056

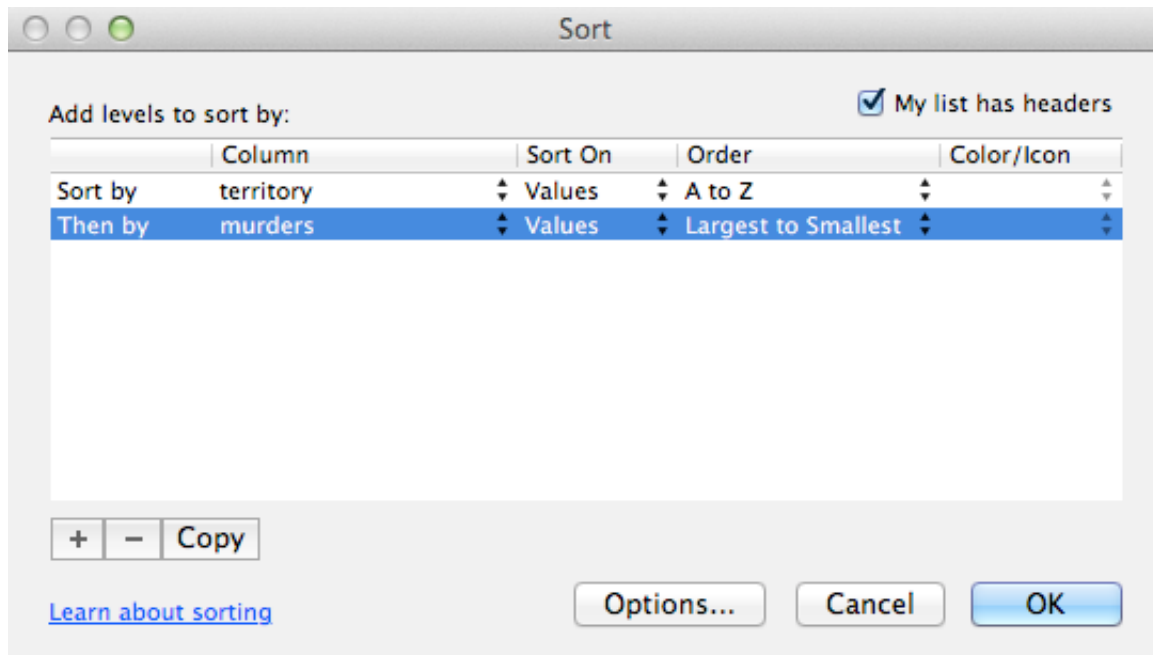
But beware! Put the cursor in the column, but DO NOT select the column letter (D, in this case) and then sort. Consider the example below:



	A	B	C	D	E	F
1	province	territory	Total population	murders	rapes	burglary
2	Agrigento	Sicilia	454,002	2	4	118
3	Alessandria	Piemonte	440,613	2	32	2,545
4	Ancona	Marche	481,028	0	28	2,152
5	Aosta	Valle d'Aosta / Vallée d'Aoste	128,230	1	15	470
6	Arezzo	Toscana	349,651	2	27	1,088
7	Ascoli Piceno	Marche	214,068	0	6	565
8	Asti	Piemonte	221,687	1	19	1,412
9	Avellino	Campania	439,137	3	14	1,056
10	Bari	Puglia	1,258,706	19	68	4,222
11	Belluno	Veneto	213,474	0	15	366

You will get a warning message, but going ahead will sort ONLY the data in that column, thereby disordering your data!

The other method of sorting is for when you want to sort by more than one variable. For instance, suppose we wish to sort the crime data first by Territory in alphabetical order, but then by “Murders” in descending order within each Territory. To do that, go to the toolbar, click on “Data” and then “Sort...”, and then choose the variables by which you wish to sort. (Click the plus sign to add new levels.) Then click “OK”.



The result will be this:

	A	B	C	D	E	F	G
1	province	territory	Total				
2	Chieti	Abruzzo	population	murders	rapes	burglary	car the
3	Pescara	Abruzzo	397,123	5	30	1,068	43:
4	L'Aquila	Abruzzo	323,184	4	28	1,033	80:
5	Teramo	Abruzzo	309,820	1	17	908	15:
6	Potenza	Basilicata	312,239	1	15	984	19:
7	Matera	Basilicata	383,791	3	21	607	20:
8	Reggio di	Calabria	203,726	0	10	308	12:
9	Vibo Valentia	Calabria	566,977	25	34	1,051	1,76:
10	Cosenza	Calabria	166,560	12	8	331	30:
11	Catanzaro	Calabria	734,656	6	55	1,695	1,44:
12	Crotone	Calabria	368,597	5	25	695	92:
13	Napoli	Campania	174,605	4	4	185	16:
14	Caserta	Campania	3,080,873	62	169	4,645	17,64:
15	Salerno	Campania	916,467	14	58	2,169	2,44:
			1,109,705	9	67	2,264	2,36:

## FILTERING

Sometimes you want to examine only particular records from a large collection of data. For that, you can use Excel's Filter tool. On the toolbar, go to "Data...Filter". Small buttons will appear at the top of each column:

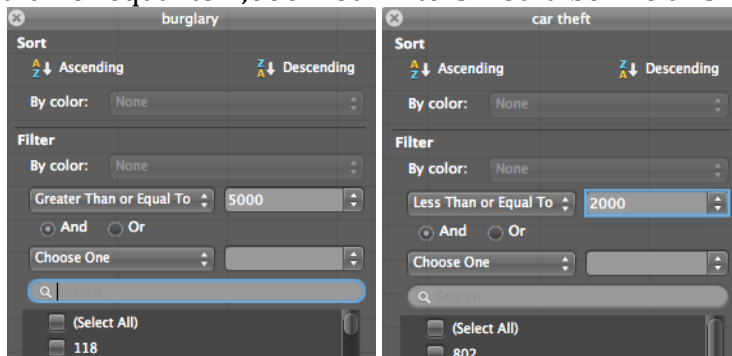
	A	B	C	D	E	F	G	
			Total					
1	province	territory	populatio	murde	rape	burglar	car the	rob
2	Agrigento	Sicilia	454,002	2	4	118	61	
3	Alessandria	Piemonte	440,613	2	32	2,545	359	
4	Ancona	Marche	481,028	0	28	2,152	196	
5	Aosta	Valle d'Aosta / Vallée d'Aoste	128,230	1	15	470	32	
6	Arezzo	Toscana	349,651	2	27	1,088	131	
7	Ascoli Piceno	Marche	214,068	0	6	565	122	

Suppose we wish to see only the records from the Territory of Lazio. Click on the button on the Territory column, uncheck the "Select All" box, and then choose Lazio from the list. This is the result:

The screenshot shows the Excel interface with the 'territory' column filter open. The filter menu displays a list of Italian regions: Friuli-Venezia Giulia, Lazio (checked), Liguria, and Lombardia. The spreadsheet view shows only the rows corresponding to the selected filter, which are rows 36, 44, 78, 80, and 104. The rows are highlighted in blue, indicating they are the only visible records after filtering.

Notice that you now are seeing only rows 36, 44, 78, 80 and 104. The rest are still there, but hidden.

More complicated filters are possible. For instance, suppose you wish to see only records in which “Burglaries” is greater than or equal to 5,000 AND car thefts is less than or equal to 2,000. Your filters would be like this



## FUNCTIONS

Excel has many built-in functions useful for performing math calculations and working with dates and text. For instance, assume that we wish to calculate the total number of murders in all the provinces. To do this, we would go to the bottom of Column D, skip a row, and then enter this formula in Cell D106: =SUM(D2:D104). The equals sign (=) is necessary for all functions. The colon (:) means “all the numbers from Cell D2 to Cell D104”. The result is this:

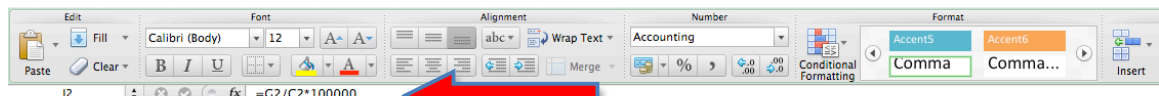
	A	B	C	D	E
1	province	territory	Total		
			population	murders	rapes
100	Vercelli	Piemonte	179,562	1	13
101	Verona	Veneto	920,158	3	73
102	Vibo Valentia	Calabria	166,560	12	8
103	Vicenza	Veneto	870,740	3	51
104	Viterbo	Lazio	320,294	1	36
105					
106				514	
107					

(The reason for skipping a row is to separate the sum from the main table so that the table can be sorted without pulling the sum into the table during the sorting operation. This way the sum will stay at the bottom of the column.

Often you will want to do a calculation on each row of your data table. For instance, you might want to calculate the auto theft rate (the number of cars stolen per 100,000 population), which would let you compare the auto theft problem in cities of different sizes.

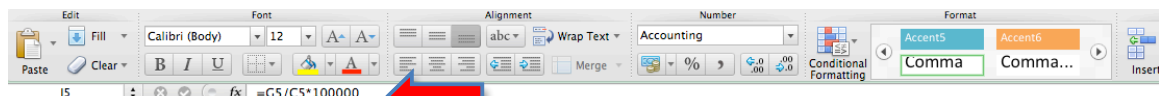


To do this, we would create a new variable called “Car Theft Rate per 100k” in Column J, the first empty column. Then, in Cell J2, we would enter this formula:  $= (G2/C2) * 100000$ . This divides the stolen cars by the population, then multiplies the result by 100,000. (Notice that there are no spaces and no thousands separators used in the formula.) Here is the result:



	A	B	C	D	E	F	G	H	I	J
	province	territory	Total population	murders	rapes	burglary	car theft	robberies	drugs	car theft rate
2	Agrigento	Sicilia	454,002	2	4	118	61	16	28	13.4
3	Alessandria	Piemonte	440,613	2	32	2,545	359	144	213	
4	Ancona	Marche	481,028	0	28	2,152	196	141	333	
5	Aosta	Valle d'Aosta / Vallée d'Aoste	128,230	1	15	470	32	20	60	
6	Arezzo	Toscana	349,651	2	27	1,088	131	76	193	
7	Ascoli Piceno	Marche	214,068	0	6	565	132	41	105	
8	Asti	Piemonte	221,687	1	19	1,412	183	94	45	
9	Avellino	Campania	439,137	3	14	1,056	324	72	126	
10	Bari	Puglia	1,258,706	19	68	4,222	6,268	1,323	553	

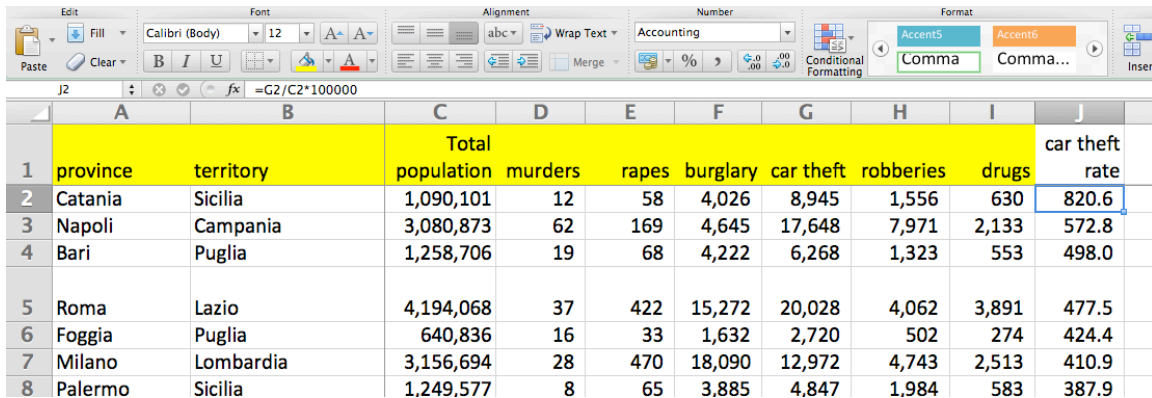
It would be very tedious to repeat writing that calculation in each of 103 rows of data. Happily, Excel has a way to rapidly copy a formula down a column of cells. To do that, you carefully move the cursor (normally a big fat white cross) to the bottom right corner of the cell containing the formula. When it is in the right spot, the cursor will change to a small black cross. At that point, you can double-click and the formula will copy down the column until it reaches a blank cell in the column to the left. This would be the result:



	A	B	C	D	E	F	G	H	I	J
	province	territory	Total population	murders	rapes	burglary	car theft	robberies	drugs	car theft rate
2	Agrigento	Sicilia	454,002	2	4	118	61	16	28	13.4
3	Alessandria	Piemonte	440,613	2	32	2,545	359	144	213	81.5
4	Ancona	Marche	481,028	0	28	2,152	196	141	333	40.7
5	Aosta	Valle d'Aosta / Vallée d'Aoste	128,230	1	15	470	32	20	60	25.0
6	Arezzo	Toscana	349,651	2	27	1,088	131	76	193	37.5
7	Ascoli Piceno	Marche	214,068	0	6	565	132	41	105	61.7
8	Asti	Piemonte	221,687	1	19	1,412	183	94	45	82.5
9	Avellino	Campania	439,137	3	14	1,056	324	72	126	73.8

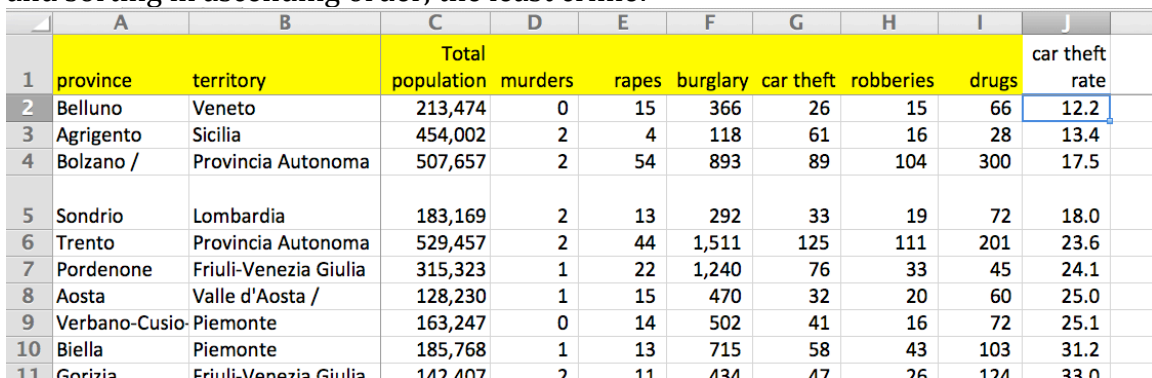
Notice that the formula changes for each row, so that Row 5 is  $=G5/C5*100000$ .

Now, if we sort by Crime Rate in descending order, we see the cities with the worst auto theft problems:



	A	B	C	D	E	F	G	H	I	J
	province	territory	population	murders	rapes	burglary	car theft	robberies	drugs	car theft rate
2	Catania	Sicilia	1,090,101	12	58	4,026	8,945	1,556	630	820.6
3	Napoli	Campania	3,080,873	62	169	4,645	17,648	7,971	2,133	572.8
4	Bari	Puglia	1,258,706	19	68	4,222	6,268	1,323	553	498.0
5	Roma	Lazio	4,194,068	37	422	15,272	20,028	4,062	3,891	477.5
6	Foggia	Puglia	640,836	16	33	1,632	2,720	502	274	424.4
7	Milano	Lombardia	3,156,694	28	470	18,090	12,972	4,743	2,513	410.9
8	Palermo	Sicilia	1,249,577	8	65	3,885	4,847	1,984	583	387.9

and sorting in ascending order, the least crime:



	A	B	C	D	E	F	G	H	I	J
	province	territory	population	murders	rapes	burglary	car theft	robberies	drugs	car theft rate
2	Belluno	Veneto	213,474	0	15	366	26	15	66	12.2
3	Agrigento	Sicilia	454,002	2	4	118	61	16	28	13.4
4	Bolzano /	Provincia Autonoma	507,657	2	54	893	89	104	300	17.5
5	Sondrio	Lombardia	183,169	2	13	292	33	19	72	18.0
6	Trento	Provincia Autonoma	529,457	2	44	1,511	125	111	201	23.6
7	Pordenone	Friuli-Venezia Giulia	315,323	1	22	1,240	76	33	45	24.1
8	Aosta	Valle d'Aosta /	128,230	1	15	470	32	20	60	25.0
9	Verbanio-Cusio	Piemonte	163,247	0	14	502	41	16	72	25.1
10	Biella	Piemonte	185,768	1	13	715	58	43	103	31.2
11	Gorizia	Friuli-Venezia Giulia	142,407	2	11	434	47	26	124	33.0

Here are some other useful Excel functions that can be used in similar ways:

(You can add, subtract, multiply or divide by using the symbols + - \* and /)

=AVERAGE – calculates the arithmetic mean of a column or row of numbers

=MEDIAN – finds the middle value of a column or row of numbers

=COUNT – tells you how many items there are in a column or row

=MAX – tells you the largest value in a column or row

=MIN – tells you the smallest value in a column or row

There are also a variety of text functions that can join and cut apart text strings. For instance:

If “Steve” is in Cell B2 and “Doig” is in Cell C2, then =B2&” “&C2 will produce “Steve Doig”. And =C2&”, “&B2 will produce “Doig, Steve”. Other text functions include:

=SEARCH – this will find the start of a desired string of text in a larger string.

=LEN – this will tell you how many characters are in a text string.

=LEFT – this will extract however many characters you specify starting from the left.

=RIGHT -- this will extract characters starting from the right.

=MID -- this will start extract where you tell it to start, and get as many characters as you specify.

You can also do date arithmetic, such as calculating the number of days or years between two dates, or hours, minutes and/or seconds between two times. For



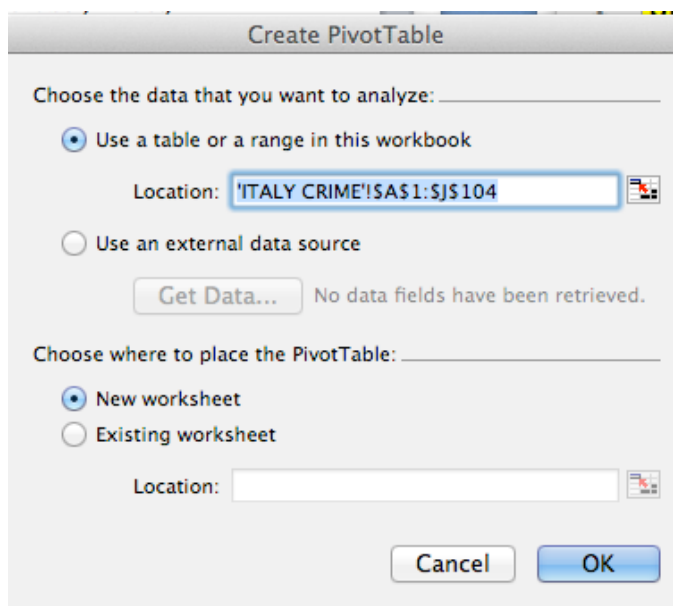
instance, to calculate on April 24, 2014, the age in years of someone whose birth date is in cell B2, you could use this formula:  $\text{=(DATE(2014,4,24)-B2)/365.25}$ . The first part of the formula calculates the number of days between the two dates, then that is divided by 365.25 (the .25 accounts for leap years) to produce the years. Another useful date function is =WEEKDAY, which will tell you on which day of the week a chosen date falls. For instance =WEEKDAY(DATE(1948,4,21)) returns a 4, which means I was born on a Wednesday.

Excel offers well over 200 functions in a variety of categories beyond just math, dates and text: Financial, engineering, database, logical, statistical, etc. But it is unlikely that you will need to be familiar with more than a dozen or so functions, unless you are a journalist with a very specialized beat such as economics.

## PIVOT TABLES

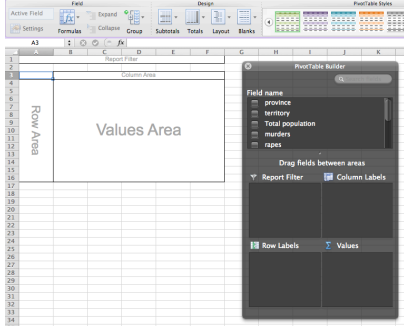
One of Excel's best tricks is the ability to summarize data that is in categories. The tool that does this is called a pivot table, which creates an interactive cross-tabulation of the data by category.

To create a pivot table, every column of your data must have a variable label; in fact, it is always good practice to put in a variable label any time you insert or add a new column. First, you make sure your cursor is on some cell in the table. Then go to the tool bar and click on "Data...Pivot Table". A window will pop up that looks like this:



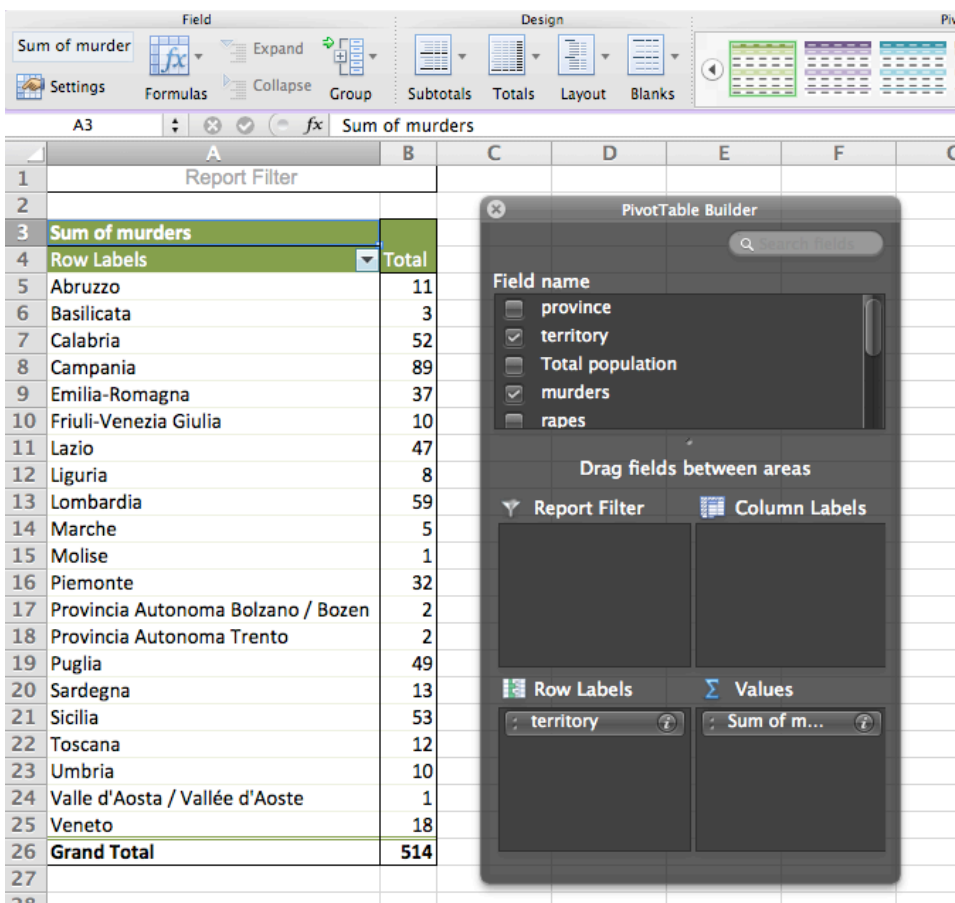
Just hit "OK"

This will open a new sheet that looks like this:



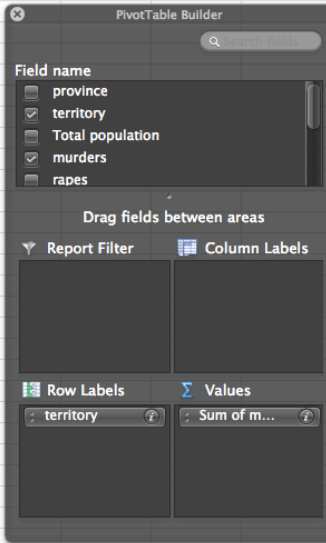
To build a pivot table, you should visualize the piece of paper that would answer your question. Our example data shows 103 provinces in the 20 Territories of Italy. Imagine that you wanted to know the total number of murders in each Territorio. The piece of paper that would answer that question would list each Territory, with the total number of murders next to each name.

To build this pivot table, we would use the mouse to pick up “Territory” from the list of variables in the floating box to the right, and place it in the “Row Labels” box below. We would then take the “Murders” variable and put it in the “Values” box. This would be the result:



If you click the cursor into the "Total" Column and hit the Z-A button to sort, you will get this:

	A	B	C	D	E	F	G
1	Report Filter						
2							
3	Sum of murders						
4	Row Labels	Total					
5	Campania	89					
6	Lombardia	59					
7	Sicilia	53					
8	Calabria	52					
9	Puglia	49					
10	Lazio	47					
11	Emilia-Romagna	37					
12	Piemonte	32					
13	Veneto	18					
14	Sardegna	13					
15	Toscana	12					
16	Abruzzo	11					
17	Friuli-Venezia Giulia	10					
18	Umbria	10					
19	Liguria	8					
20	Marche	5					
21	Basilicata	3					
22	Provincia Autonoma Bolzano / Bozen	2					
23	Provincia Autonoma Trento	2					
24	Molise	1					
25	Valle d'Aosta / Vallée d'Aoste	1					
26	Grand Total	514					
27							
28							



It is possible to make very complicated pivot tables, with multiple subtotals. But I recommend making a new pivot table for each question you want to answer; several simple tables are easier to understand than one very complicated table that tries to answer many questions at once.



The little button on the "Values" variable opens up a box that will let you make a variety of other choices about how to summarize and display the result:

PivotTable Field

Source field: murders

Field Name: Sum of murders

Summarize by:

Sum
Count
Average
Max
Min
Product
Count Numbers

OK

Cancel

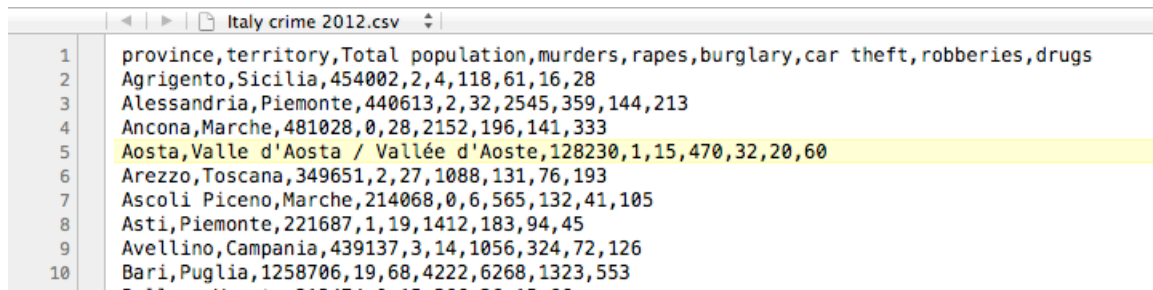
Delete

Number...

Options >>

## OTHER EXCEL TIPS

Excel will import data that comes in a variety of formats other than the native \*.xls or \*.xlsx that Excel uses. For instance, Excel can readily import text files in which the data columns are separated by commas, tabs, or other characters, like this:



	province,territory	Total population	murders	rapes	burglary	car theft	robberies	drugs
1	Agrigento,Sicilia	454002	2	4	118	61	16	28
2	Alessandria,Piemonte	440613	2	32	2545	359	144	213
3	Ancona, Marche	481028	0	28	2152	196	141	333
4	Aosta,Valle d'Aosta / Vallée d'Aoste	128230	1	15	470	32	20	60
5	Arezzo,Toscana	349651	2	27	1088	131	76	193
6	Ascoli Piceno, Marche	214068	0	6	565	132	41	105
7	Asti,Piemonte	221687	1	19	1412	183	94	45
8	Avellino,Campania	439137	3	14	1056	324	72	126
9	Bari,Puglia	1258706	19	68	4222	6268	1323	553
10								

If you find a web page with data in table format (rows and columns), Excel can open it as a spreadsheet. Copy the table and then paste it into Excel; very often it will flow properly into the correct columns.

## FINDING DATA

Government agencies are starting to make some of their data available in Excel or other formats. For instance, ISTAT.IT has very comprehensive data about Italian demographics, economy, crime, etc. Many of their tables can be downloaded directly as Excel files.

One trick to find interesting data would be to use Google and add these search terms: `site:.gov filetype:xls`.

## NEED HELP?

Feel free to send me an email at [steve.doig@asu.edu](mailto:steve.doig@asu.edu). I will be glad to give you advice if I can.