# Using Excel to do precision journalism


## By S. Doig and S. Cohen

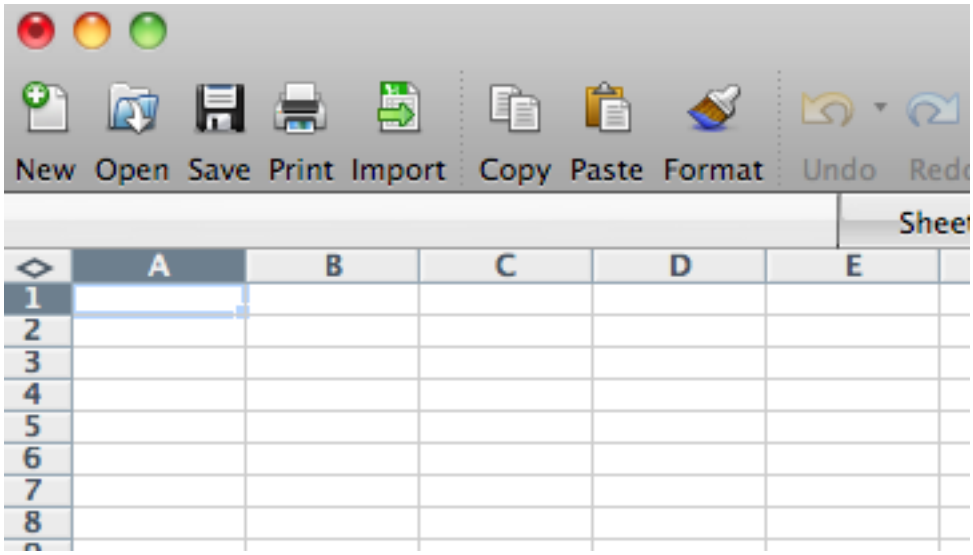

Materials create for a Workshop at the International Festival of Journalism, 26 April 2012.

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.

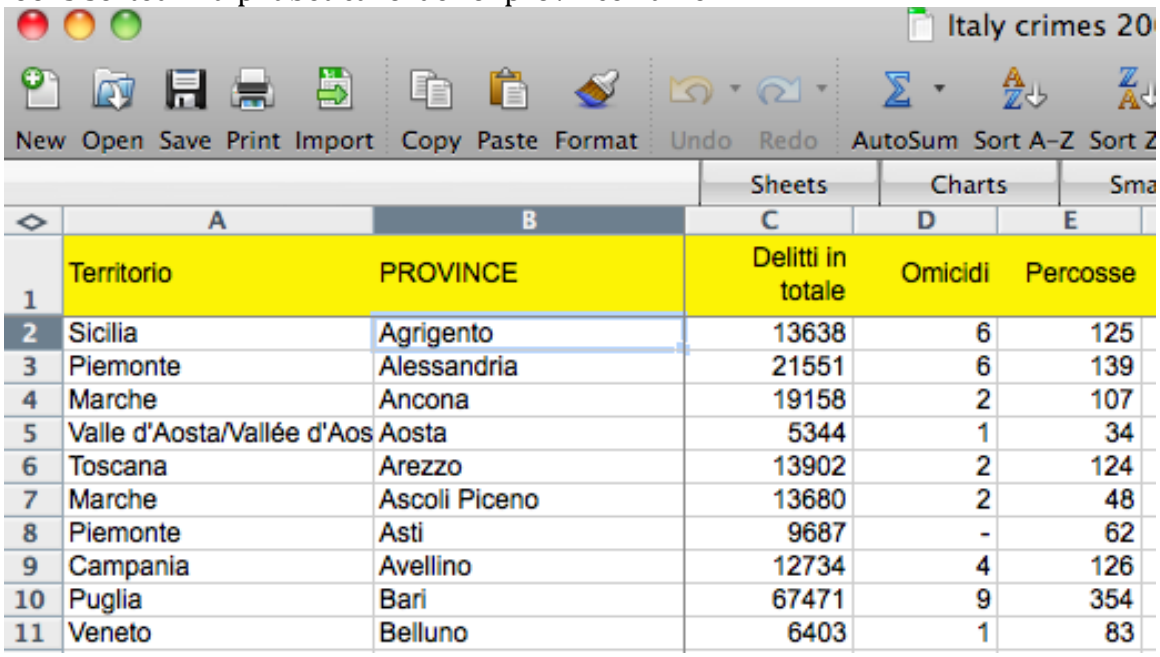



Modern versions of Excel will hold as many as 1,048,576 records with as many as 16,384 variables! An Excel spreadsheet also will hold multiple tables on separate sheets, which are tabbed on the bottom of the page.

SORTING

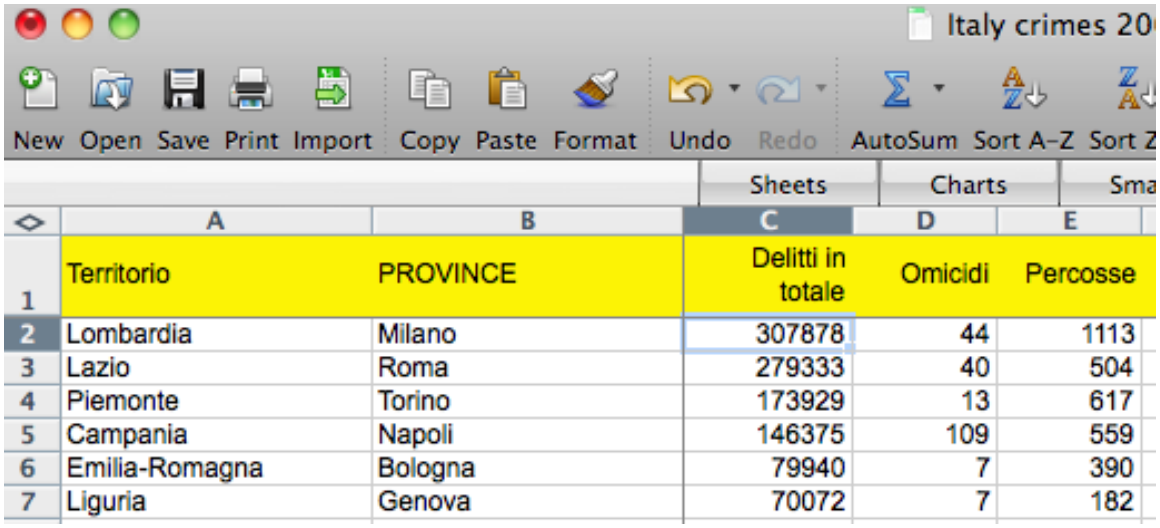
One of the most useful abilities of Excel is to sort the data into a more revealing order. Too often, we are given lists that are in alphabetical order, which is useful only for finding a particular record in a long list. In journalism, we usually are more interested in extremes: The most, the least,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



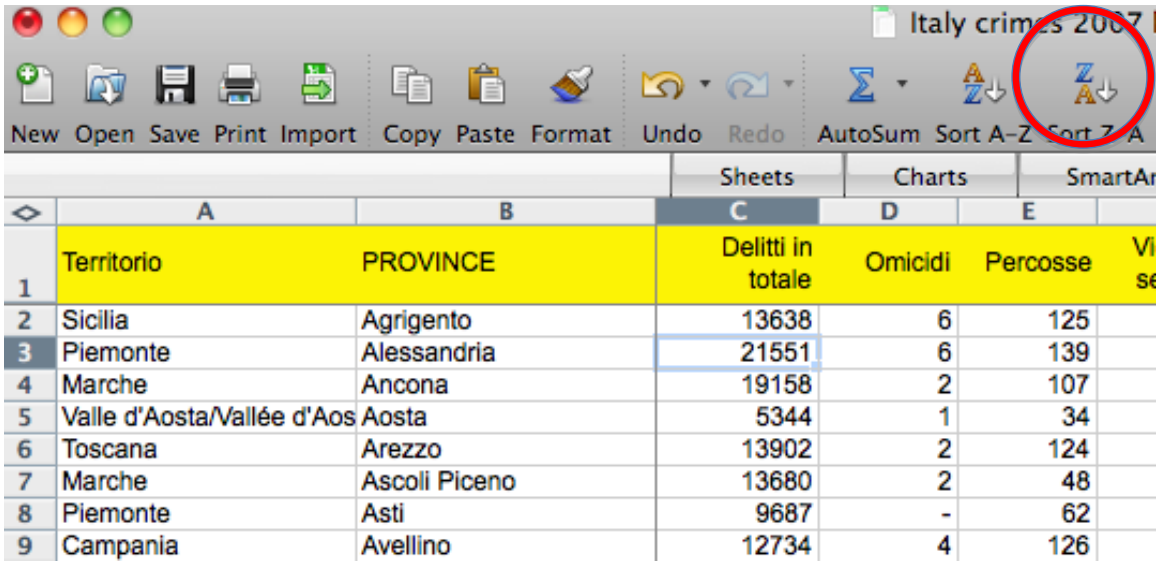

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in totale | Omicidi | Percosse |
| 2 | Sicilia | Agrigento | 13638 | 6 | 125 |
| 3 | Piemonte | Alessandria | 21551 | 6 | 139 |
| 4 | Marche | Ancona | 19158 | 2 | 107 |
| 5 | Valle d'Aosta/Vallée d'Aos | Aosta | 5344 | 1 | 34 |
| 6 | Toscana | Arezzo | 13902 | 2 | 124 |
| 7 | Marche | Ascoli Piceno | 13680 | 2 | 48 |
| 8 | Piemonte | Asti | 9687 | - | 62 |
| 9 | Campania | Avellino | 12734 | 4 | 126 |
| 10 | Puglia | Bari | 67471 | 9 | 354 |
| 11 | Veneto | Belluno | 6403 | 1 | 83 |

Far more interesting would be to sort it in descending order of the total number of crimes, with the most crime-ridden city at the top of the list:



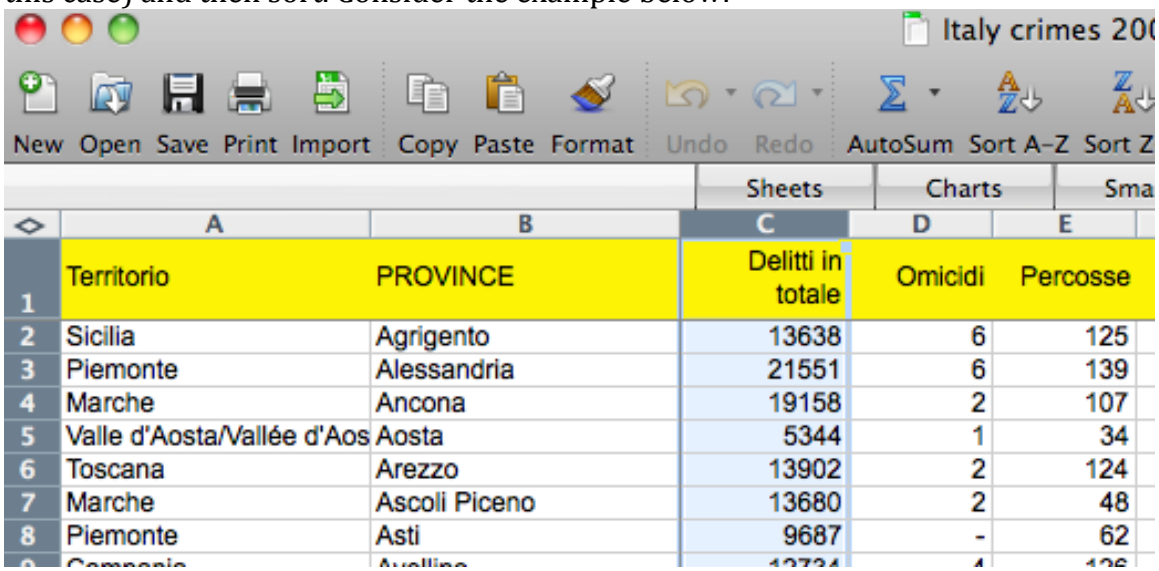

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in totale | Omicidi | Percosse |
| 2 | Lombardia | Milano | 307878 | 44 | 1113 |
| 3 | Lazio | Roma | 279333 | 40 | 504 |
| 4 | Piemonte | Torino | 173929 | 13 | 617 |
| 5 | Campania | Napoli | 146375 | 109 | 559 |
| 6 | Emilia-Romagna | Bologna | 79940 | 7 | 390 |
| 7 | Liguria | Genova | 70072 | 7 | 182 |

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 ("Delitti in totale" in this case) and then click the Z-A button:




| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in totale | Omicidi | Percosse |
| 2 | Sicilia | Agrigento | 13638 | 6 | 125 |
| 3 | Piemonte | Alessandria | 21551 | 6 | 139 |
| 4 | Marche | Ancona | 19158 | 2 | 107 |
| 5 | Valle d'Aosta/Vallée d'Aos | Aosta | 5344 | 1 | 34 |
| 6 | Toscana | Arezzo | 13902 | 2 | 124 |
| 7 | Marche | Ascoli Piceno | 13680 | 2 | 48 |
| 8 | Piemonte | Asti | 9687 | - | 62 |
| 9 | Campania | Avellino | 12734 | 4 | 126 |

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

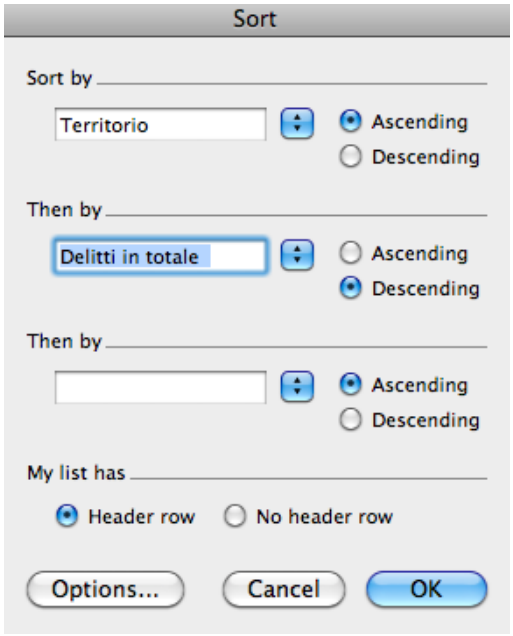



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in totale | Omicidi | Percosse |
| 2 | Sicilia | Agrigento | 13638 | 6 | 125 |
| 3 | Piemonte | Alessandria | 21551 | 6 | 139 |
| 4 | Marche | Ancona | 19158 | 2 | 107 |
| 5 | Valle d'Aosta/Vallée d'Aos | Aosta | 5344 | 1 | 34 |
| 6 | Toscana | Arezzo | 13902 | 2 | 124 |
| 7 | Marche | Ascoli Piceno | 13680 | 2 | 48 |
| 8 | Piemonte | Asti | 9687 | - | 62 |

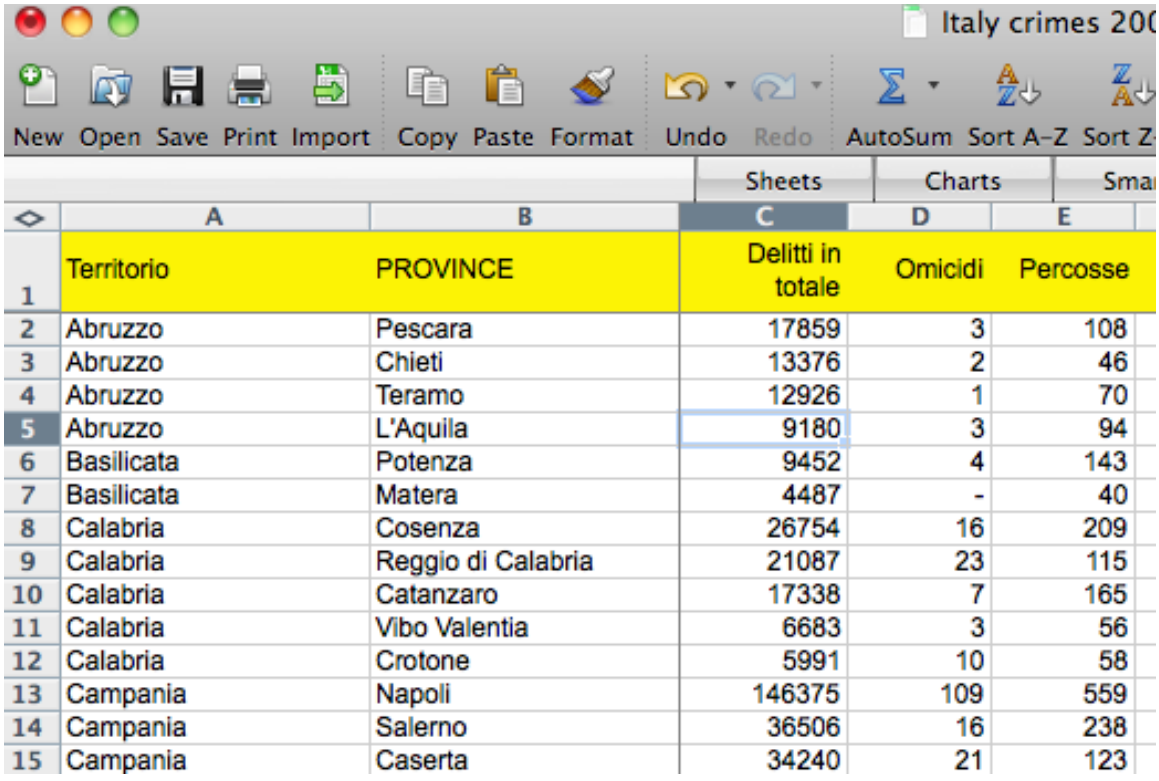
Doing that will sort ONLY the data in that column, thereby disordering your data! Notice well how this can happen!

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in totale | Omicidi | Percosse |
| 2 | Sicilia | Agrigento | 307878 | 6 | 125 |
| 3 | Piemonte | Alessandria | 279333 | 6 | 139 |
| 4 | Marche | Ancona | 173929 | 2 | 107 |
| 5 | Valle d'Aosta/Vallée d'Aos | Aosta | 146375 | 1 | 34 |
| 6 | Toscana | Arezzo | 79940 | 2 | 124 |

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 Territerio in alphabetical order, but then by "Delitti in Totale" in descending order within each Territerio. To do that, go to the toolbar, click on "Data" and then "Sort…", and then choose the variables by which you wish to sort. Then click "OK".



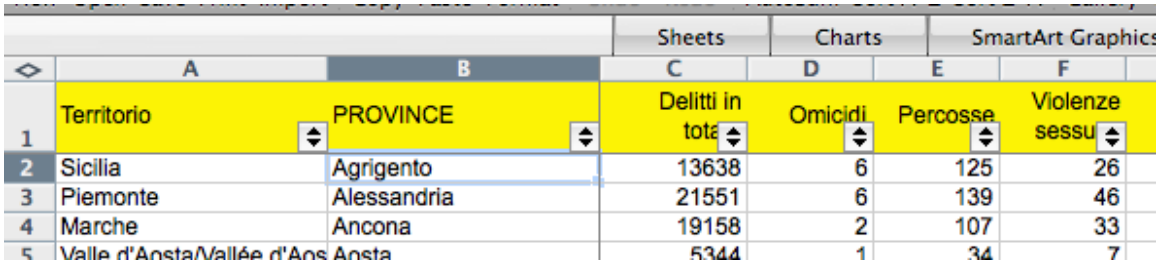

The result will be this:




| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in totale | Omicidi | Percosse |
| 2 | Abruzzo | Pescara | 17859 | 3 | 108 |
| 3 | Abruzzo | Chieti | 13376 | 2 | 46 |
| 4 | Abruzzo | Teramo | 12926 | 1 | 70 |
| 5 | Abruzzo | L'Aquila | 9180 | 3 | 94 |
| 6 | Basilicata | Potenza | 9452 | 4 | 143 |
| 7 | Basilicata | Matera | 4487 | - | 40 |
| 8 | Calabria | Cosenza | 26754 | 16 | 209 |
| 9 | Calabria | Reggio di Calabria | 21087 | 23 | 115 |
| 10 | Calabria | Catanzaro | 17338 | 7 | 165 |
| 11 | Calabria | Vibo Valentia | 6683 | 3 | 56 |
| 12 | Calabria | Crotone | 5991 | 10 | 58 |
| 13 | Campania | Napoli | 146375 | 109 | 559 |
| 14 | Campania | Salerno | 36506 | 16 | 238 |
| 15 | Campania | Caserta | 34240 | 21 | 123 |

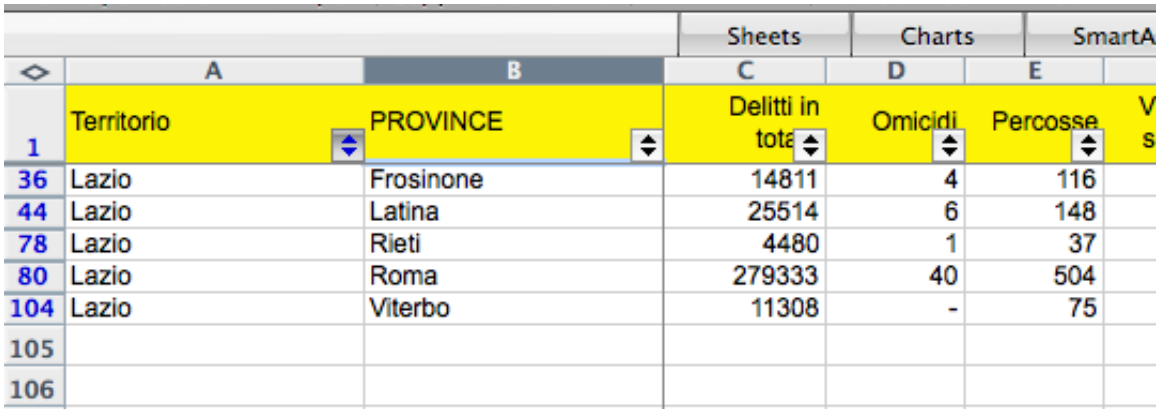
FILTERING

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…Autofilter". Small buttons will appear at the top of each column:



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in tota | Omicidi | Percosse | Violenze sessu |
| 2 | Sicilia | Agrigento | 13638 | 6 | 125 | 26 |
| 3 | Piemonte | Alessandria | 21551 | 6 | 139 | 46 |
| 4 | Marche | Ancona | 19158 | 2 | 107 | 33 |
| 5 | Valle d'Aosta/Vallée d'Aos | Aosta | 5344 | 1 | 34 | 7 |

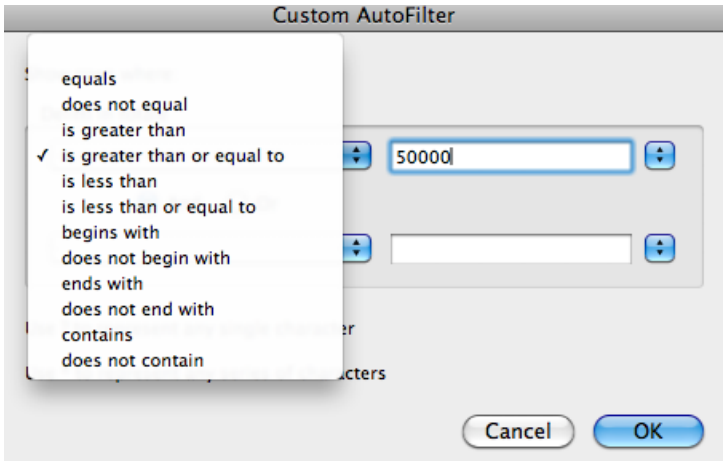
Suppose we wish to see only the records from the territerio of Lazio. Click on the button on the Territerio column and choose Lazio from the list. This is the result:



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in tota | Omicidi | Percosse |
| 36 | Lazio | Frosinone | 14811 | 4 | 116 |
| 44 | Lazio | Latina | 25514 | 6 | 148 |
| 78 | Lazio | Rieti | 4480 | 1 | 37 |
| 80 | Lazio | Roma | 279333 | 40 | 504 |
| 104 | Lazio | Viterbo | 11308 | - | 75 |
| 105 | | | | | |
| 106 | | | | | |

Notice that you now are seeing only rows 36, 44, 78, 80 and 104.

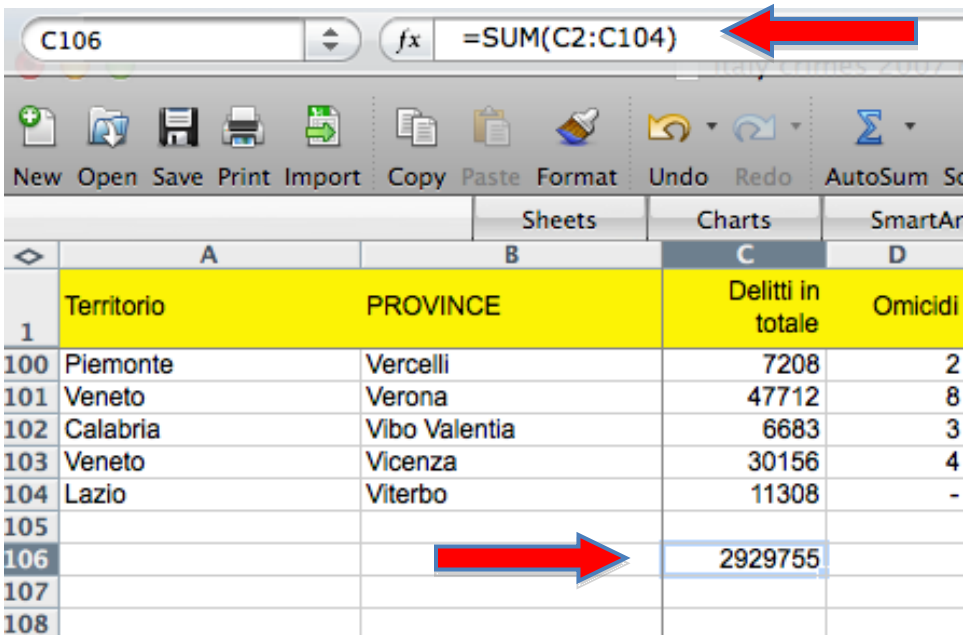
More complicated filters are possible. For instance, suppose you wish to see only records in which "Delitti in totale" is greater than or equal to 50,000. Click on the button and choose "Custom Filter…":




You could also, for instance, choose records in which "Delitti in totale" is greater than 50,000 and "Omicidi" is less than or equal to 25.

FUNCTIONS

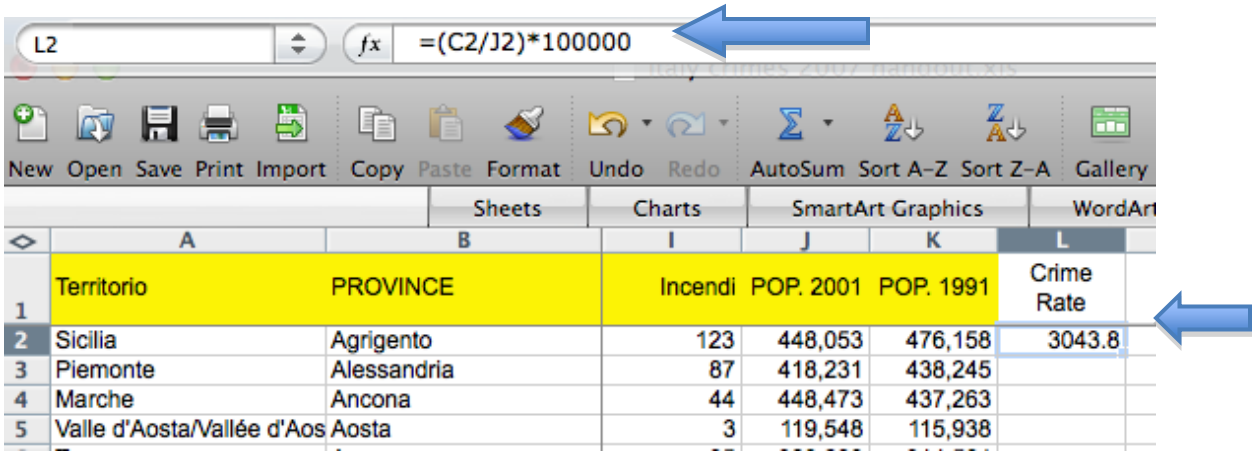
Excel has many built-in functions useful for performing math calculations and working with dates and text. For instance, assume that we wish to calculate the total number of crimes in all the provinces. To do this, we would go to the bottom of Column C, skip a row, and then enter this formula IN Cell C106: =SUM(C2:C104). The equals sign (=) is necessary for all functions. The colon (:) means "all the numbers from Cell C2 to Cell 104". The result is this:




| | A | B | C | D |
|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in totale | Omicidi |
| 100 | Piemonte | Vercelli | 7208 | 2 |
| 101 | Veneto | Verona | 47712 | 8 |
| 102 | Calabria | Vibo Valentia | 6683 | 3 |
| 103 | Veneto | Vicenza | 30156 | 4 |
| 104 | Lazio | Viterbo | 11308 | - |
| 105 | | | | |
| 106 | | | 2929755 | |
| 107 | | | | |
| 108 | | | | |

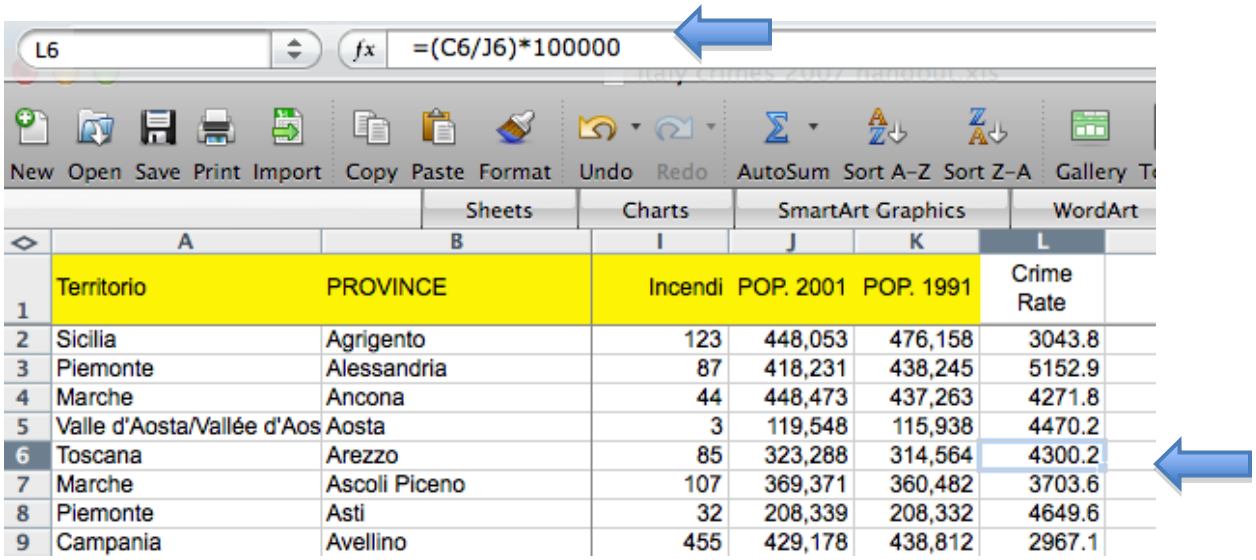
(The reason for skipping a row is to separate the sum from the main table so that the table can be sorted without pulling the sum into the table during the sorting operation. This way the sum will stay at the bottom of the column.

Often you will want to do a calculation on each row of your data table. For instance, you might want to calculate the crime rate (the number of crimes per 100,000 population), which would let you compare the crime problem in cities of different sizes. To do this, we would create a new variable called "Crime Rate" in Column L, the first empty column. Then, in Cell L2, we would enter this formula: =(C2/J2)*100000. This divides the total crimes 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 | I | J | K | L |
|---|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Incendi | POP. 2001 | POP. 1991 | Crime Rate |
| 2 | Sicilia | Agrigento | 123 | 448,053 | 476,158 | 3043.8 |
| 3 | Piemonte | Alessandria | 87 | 418,231 | 438,245 | |
| 4 | Marche | Ancona | 44 | 448,473 | 437,263 | |
| 5 | Valle d'Aosta/Vallée d'Aos | Aosta | 3 | 119,548 | 115,938 | |

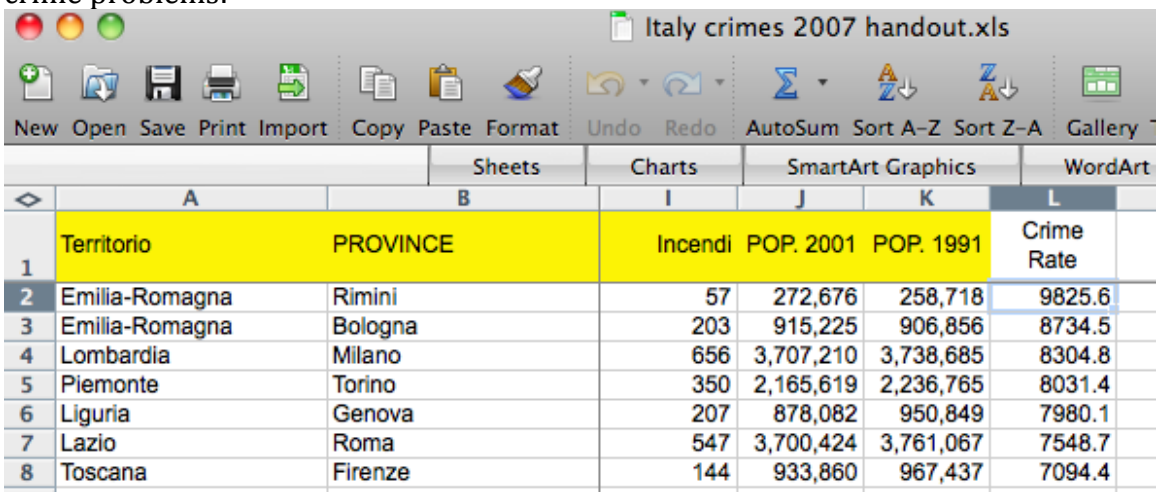
It would be very tedious to repeat writing that calculation in each of 103 rows of data. Happily, Excel has a way to rapidly copy a formula down a column of cells. To do that, you careful move the cursor (normally a big fat white cross) to the bottom right corner of the cell containing the formula. When 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 | I | J | K | L |
|---|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Incendi | POP. 2001 | POP. 1991 | Crime Rate |
| 2 | Sicilia | Agrigento | 123 | 448,053 | 476,158 | 3043.8 |
| 3 | Piemonte | Alessandria | 87 | 418,231 | 438,245 | 5152.9 |
| 4 | Marche | Ancona | 44 | 448,473 | 437,263 | 4271.8 |
| 5 | Valle d'Aosta/Vallée d'Aos | Aosta | 3 | 119,548 | 115,938 | 4470.2 |
| 6 | Toscana | Arezzo | 85 | 323,288 | 314,564 | 4300.2 |
| 7 | Marche | Ascoli Piceno | 107 | 369,371 | 360,482 | 3703.6 |
| 8 | Piemonte | Asti | 32 | 208,339 | 208,332 | 4649.6 |
| 9 | Campania | Avellino | 455 | 429,178 | 438,812 | 2967.1 |

Notice that the formula changes for each row, so that Row 6 is =(C6/J6)*100000.

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




| | A | B | I | J | K | L |
|---|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Incendi | POP. 2001 | POP. 1991 | Crime Rate |
| 2 | Emilia-Romagna | Rimini | 57 | 272,676 | 258,718 | 9825.6 |
| 3 | Emilia-Romagna | Bologna | 203 | 915,225 | 906,856 | 8734.5 |
| 4 | Lombardia | Milano | 656 | 3,707,210 | 3,738,685 | 8304.8 |
| 5 | Piemonte | Torino | 350 | 2,165,619 | 2,236,765 | 8031.4 |
| 6 | Liguria | Genova | 207 | 878,082 | 950,849 | 7980.1 |
| 7 | Lazio | Roma | 547 | 3,700,424 | 3,761,067 | 7548.7 |
| 8 | Toscana | Firenze | 144 | 933,860 | 967,437 | 7094.4 |

and sorting in ascending order, the least crime:

| | A | B | I | J | K | L |
|---|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Incendi | POP. 2001 | POP. 1991 | Crime Rate |
| 2 | Basilicata | Matera | 21 | 204,239 | 208,985 | 2196.9 |
| 3 | Sicilia | Enna | 69 | 177,200 | 186,182 | 2209.4 |
| 4 | Basilicata | Potenza | 75 | 393,529 | 401,543 | 2401.9 |
| 5 | Sardegna | Oristano | 8 | 153,082 | 156,970 | 2558.8 |

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.

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, 2010, the age in years of someone whose birth date is in cell B2, you could use this formula: =(DATE(2010,4,24)-B2)/365.25. The first part of the formula calculates the number of days between the two dates, then that is divided by 362.25 (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.

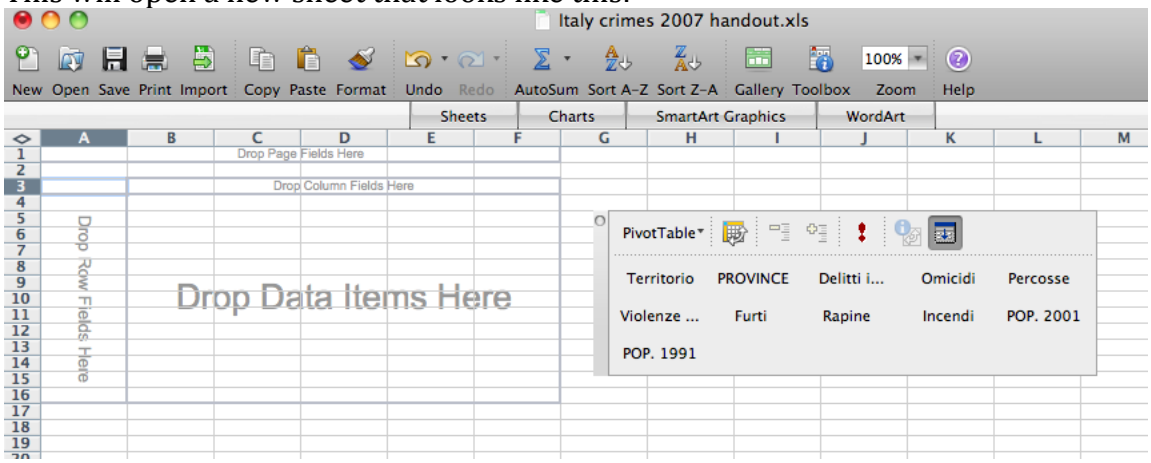
Excel offers well over 200 functions in a variety of categories beyond just math, dates and text: Financial, engineering, database, logical, statistical, etc. But it is unlikely that you will need to be familiar with more than a dozen or so functions, unless you are a journalist with a very specialized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

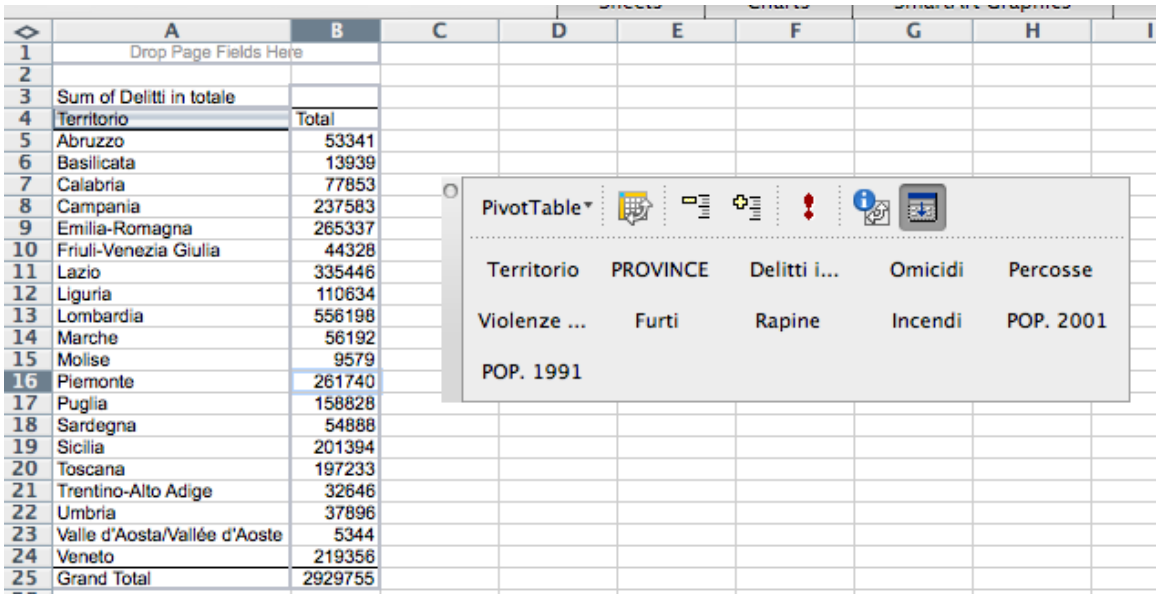
To create a pivot table, every column of your data must have a variable label; in fact, it is always good practice to put in a variable label any time you insert or add a new column. First, you make sure your cursor is on some cell in the table. Then go to the tool bar and click on "Data…Pivot Table Report". A window will pop up called the "Pivot Table Wizard". Just hit "Next…Next…Finish" on the three steps of the wizard.

This will open a new sheet that looks like this:

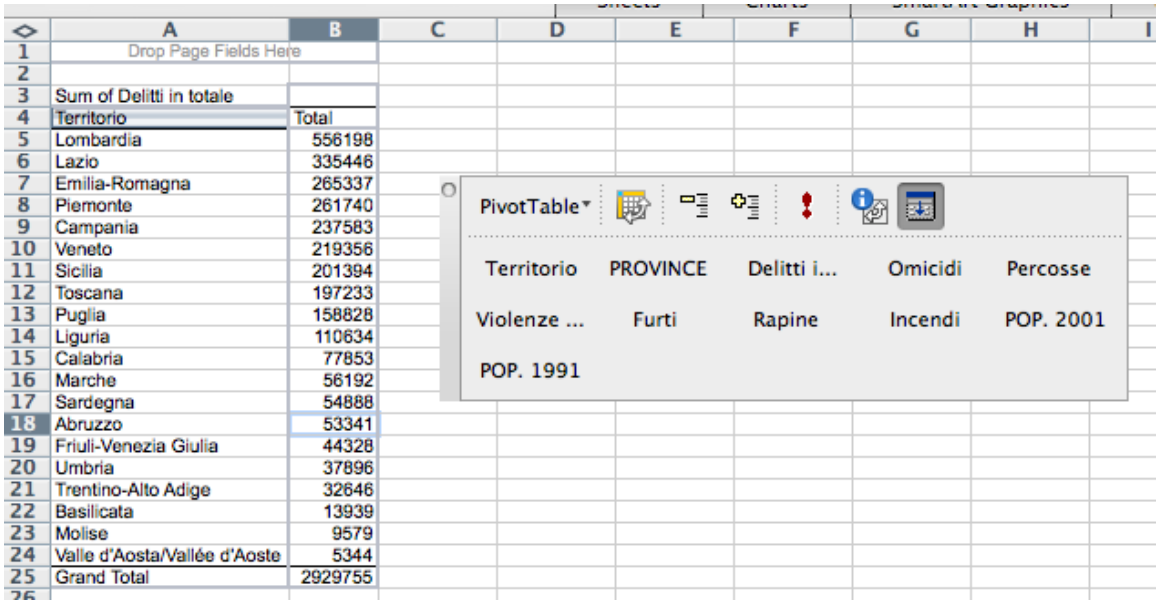



To build a pivot table, you should visualize the piece of paper that would answer your question. Our example data shows 103 provinces in the 20 Territorios of Italy. Imagine that you wanted to know the total number of crimes in each Territorio. The piece of paper that would answer that question would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

To build this pivot table, we would use the mouse to pick up "Territorio" from the list of variables in the floating box to the right, and place it in the "Drop Row Fields Here" box. We would then take the "Delitti in totale" variable and put it in the "Drop Data Items Here" 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:

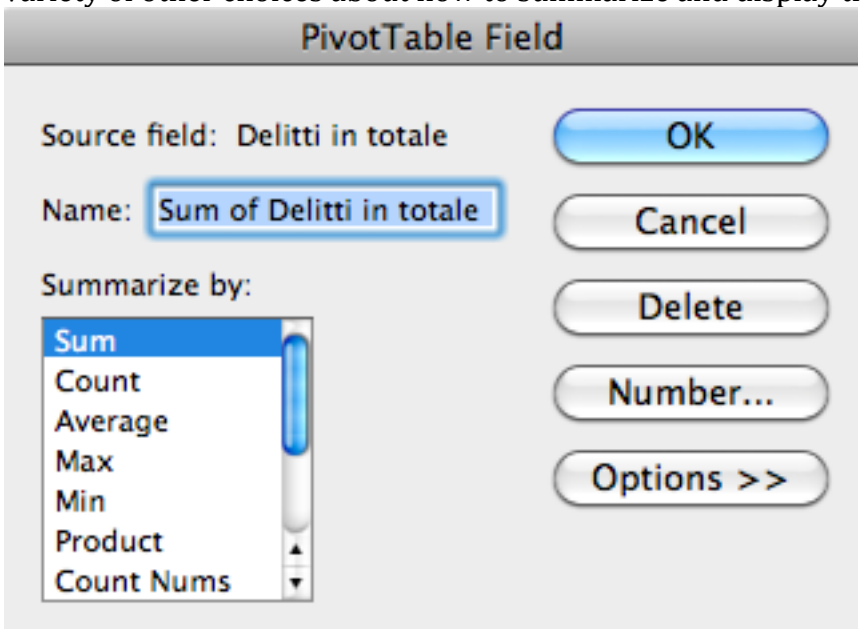


| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Drop Page Fields Here | | | | | | | | |
| 2 | | | | | | | | | |
| 3 | Sum of Delitti in totale | | | | | | | | |
| 4 | Territorio | Total | | | | | | | |
| 5 | Lombardia | 556198 | | | | | | | |
| 6 | Lazio | 335446 | | | | | | | |
| 7 | Emilia-Romagna | 265337 | | | | | | | |
| 8 | Piemonte | 261740 | | | | | | | |
| 9 | Campania | 237583 | | | | | | | |
| 10 | Veneto | 219356 | | | | | | | |
| 11 | Sicilia | 201394 | | | | | | | |
| 12 | Toscana | 197233 | | | | | | | |
| 13 | Puglia | 158828 | | | | | | | |
| 14 | Liguria | 110634 | | | | | | | |
| 15 | Calabria | 77853 | | | | | | | |
| 16 | Marche | 56192 | | | | | | | |
| 17 | Sardegna | 54888 | | | | | | | |
| 18 | Abruzzo | 53341 | | | | | | | |
| 19 | Friuli-Venezia Giulia | 44328 | | | | | | | |
| 20 | Umbria | 37896 | | | | | | | |
| 21 | Trentino-Alto Adige | 32646 | | | | | | | |
| 22 | Basilicata | 13939 | | | | | | | |
| 23 | Molise | 9579 | | | | | | | |
| 24 | Valle d'Aosta/Vallée d'Aoste | 5344 | | | | | | | |
| 25 | Grand Total | 2929755 | | | | | | | |



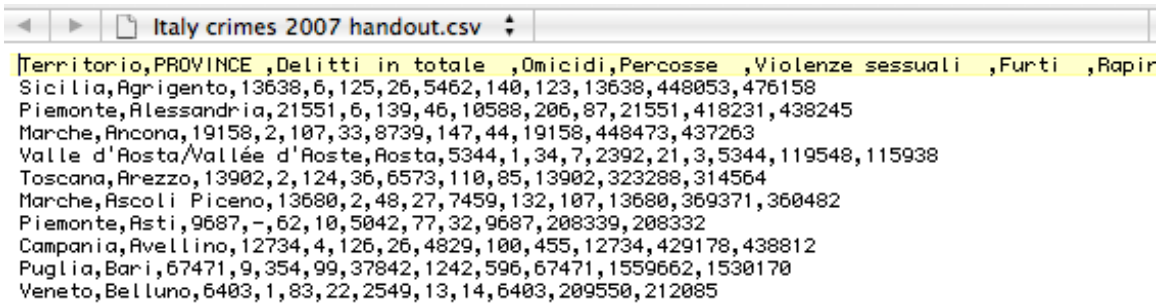
It is possible to make very complicated pivot tables, with multiple subtotals. But I recommend making a new pivot table for each question you want to answer; several simple tables are easier to understand than one very complicated table that tries to answer many questions at once.

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



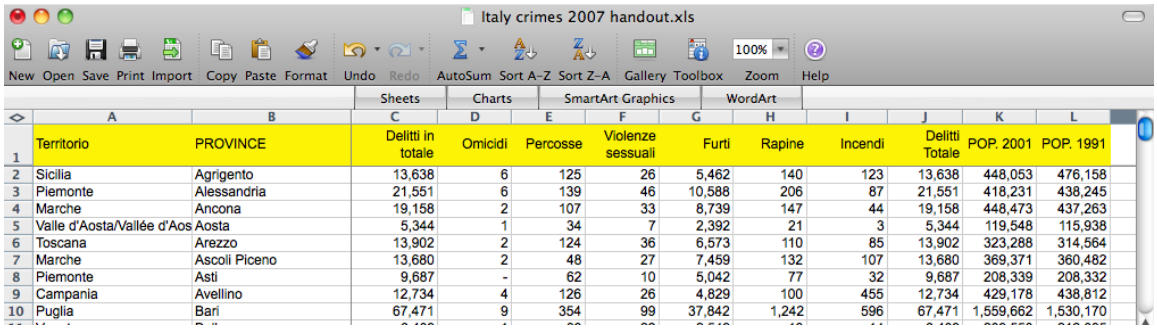
OTHER EXCEL TIPS

Excel will import data that comes in a variety of formats other than the native *.xls 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:




If you find a web page with data in table format (rows and columns), Excel can open it as a spreadsheet.

Excel also will let you format your data to make it more readable. For instance, “Format…Cells…Number” will allow you to put thousands separators in your numbers, like this:




| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Territorio | PROVINCE | Delitti in totale | Omicidi | Percosse | Violenze sessuali | Furti | Rapine | Incendi | Delitti Totale | POP. 2001 | POP. 1991 |
| 2 | Sicilia | Agrigento | 13,638 | 6 | 125 | 26 | 5,462 | 140 | 123 | 13,638 | 448,053 | 476,158 |
| 3 | Piemonte | Alessandria | 21,551 | 6 | 139 | 46 | 10,588 | 206 | 87 | 21,551 | 418,231 | 438,245 |
| 4 | Marche | Ancona | 19,158 | 2 | 107 | 33 | 8,739 | 147 | 44 | 19,158 | 448,473 | 437,263 |
| 5 | Valle d'Aosta/Vallée d'Aos | Aosta | 5,344 | 1 | 34 | 7 | 2,392 | 21 | 3 | 5,344 | 119,548 | 115,938 |
| 6 | Toscana | Arezzo | 13,902 | 2 | 124 | 36 | 6,573 | 110 | 85 | 13,902 | 323,288 | 314,564 |
| 7 | Marche | Ascoli Piceno | 13,680 | 2 | 48 | 27 | 7,459 | 132 | 107 | 13,680 | 369,371 | 360,482 |
| 8 | Piemonte | Asti | 9,687 | - | 62 | 10 | 5,042 | 77 | 32 | 9,687 | 208,339 | 208,332 |
| 9 | Campania | Avellino | 12,734 | 4 | 126 | 26 | 4,829 | 100 | 455 | 12,734 | 429,178 | 438,812 |
| 10 | Puglia | Bari | 67,471 | 9 | 354 | 99 | 37,842 | 1,242 | 596 | 67,471 | 1,559,662 | 1,530,170 |

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:.it filetype:xls.

NEED HELP?
Feel free to send us an email at **steve.doig@asu.edu** or **sarah.cohen@duke.edu**. We will be glad to give you advice if we can.