

Using Excel 2010 for Windows to do data journalism

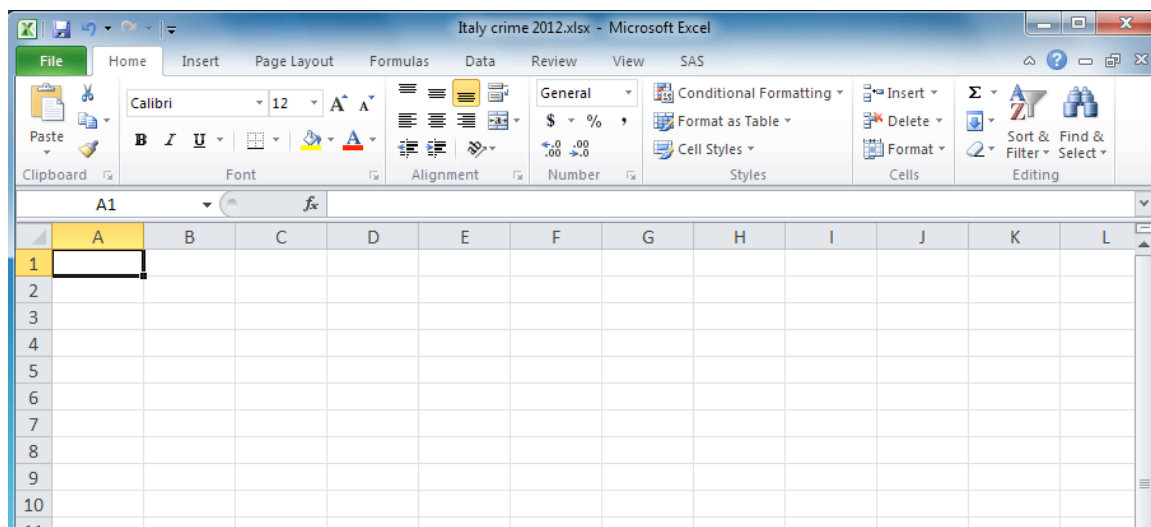
Steve Doig (steve.doig@asu.edu)

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 are tabbed on the bottom of the page.

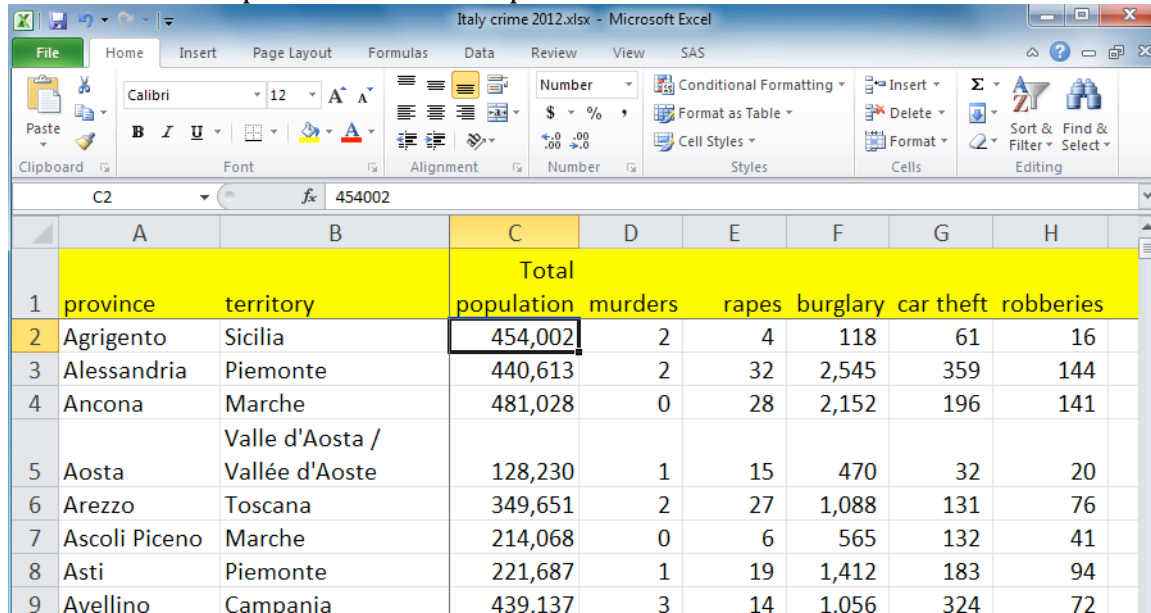
49	Lucca	Toscana	393,795	1
50	Macerata	Marche	325,362	4
51	Mantova	Lombardia	415,442	3
52	Massa-Carrara	Toscana	203,901	1
53	Matera	Basilicata	203,726	0
54	Messina	Sicilia	653,737	6
55	Milano	Lombardia	3,156,694	28

ITALLY CRIME PRISON DATA Sheet1

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.

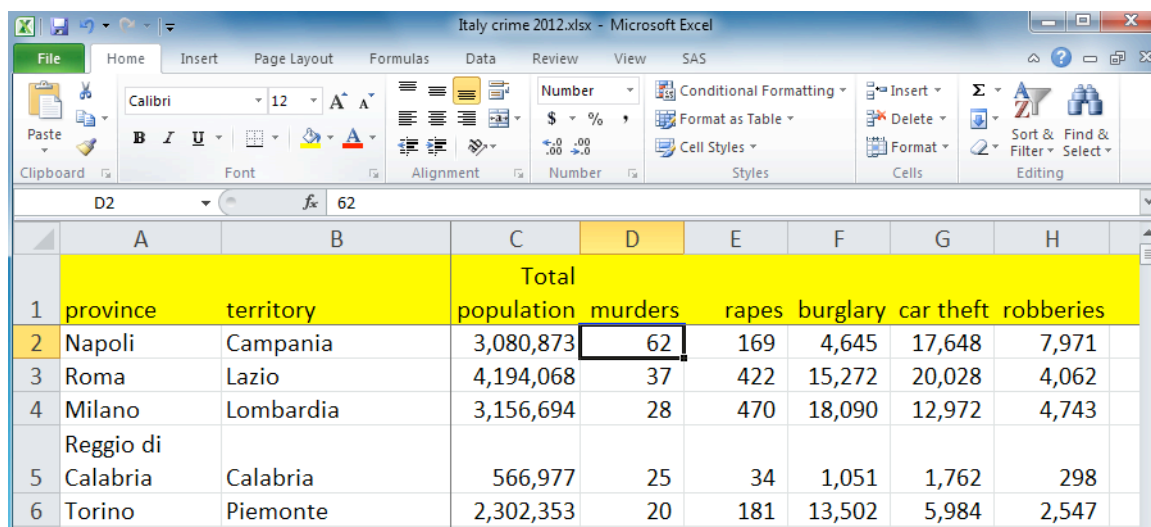
Consider the data used in this workshop, a list of the provinces of Italy showing the number of various kinds of crimes reported during a recent year. Here is how it looks sorted in alphabetical order of province name:



The screenshot shows the Microsoft Excel interface with the file 'Italy crime 2012.xlsx'. The 'Home' tab is active. The table data is as follows:

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

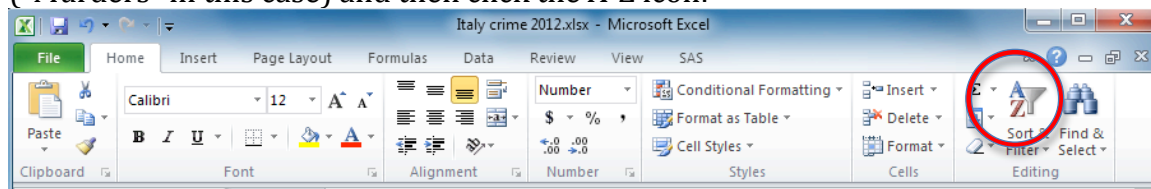
Far more interesting would be to sort it in descending order of the total number of murders, with the most violent city at the top of the list:



The screenshot shows the same Excel file, but the data is now sorted by the number of murders in descending order. The 'murders' column (D) is highlighted. The table data is as follows:

	A	B	C	D	E	F	G	H
1	province	territory	population	murders	rapes	burglary	car theft	robberies
2	Napoli	Campania	3,080,873	62	169	4,645	17,648	7,971
3	Roma	Lazio	4,194,068	37	422	15,272	20,028	4,062
4	Milano	Lombardia	3,156,694	28	470	18,090	12,972	4,743
5	Reggio di Calabria	Calabria	566,977	25	34	1,051	1,762	298
6	Torino	Piemonte	2,302,353	20	181	13,502	5,984	2,547

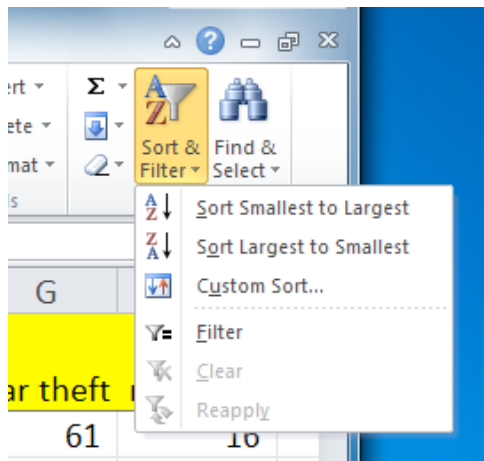
There are two methods of sorting. The first method is quick and can be used for sorting by a single variable. Put the cursor in the column you wish to sort by ("Murders" in this case) and then click the A-Z icon:



Italy crime 2012.xlsx - Microsoft Excel

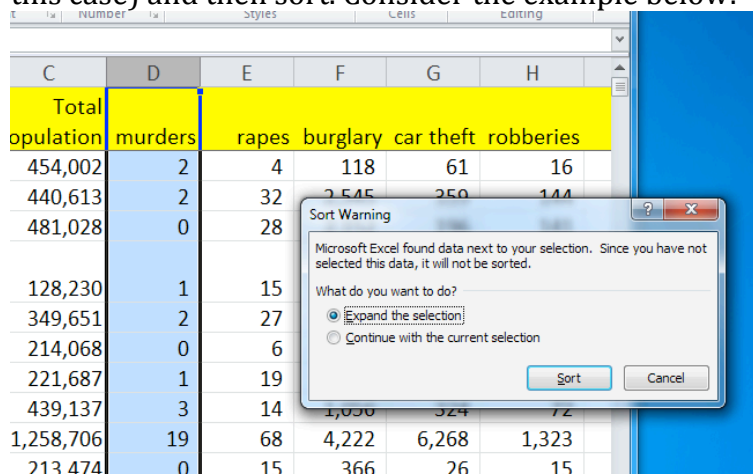
	A	B	C	D	E	F	G	H
			Total					
1	province	territory	population	murders	rapes	burglary	car theft	robberies
2	Agrigento	Sicilia	454,002	2	4	118	61	16
3	Alessandria	Piemonte	440,613	2	32	2,545	359	144
4	Ancona	Marche	481,028	0	28	2,152	196	141
5	Aosta	Valle d'Aosta / Vallée d'Aoste	128,230	1	15	470	32	20
6	Arezzo	Toscana	349,651	2	27	1,088	131	76
7	Ascoli Piceno	Marche	214,068	0	6	565	132	41

You'll get a window that looks like this:



Sort in whichever order you want.

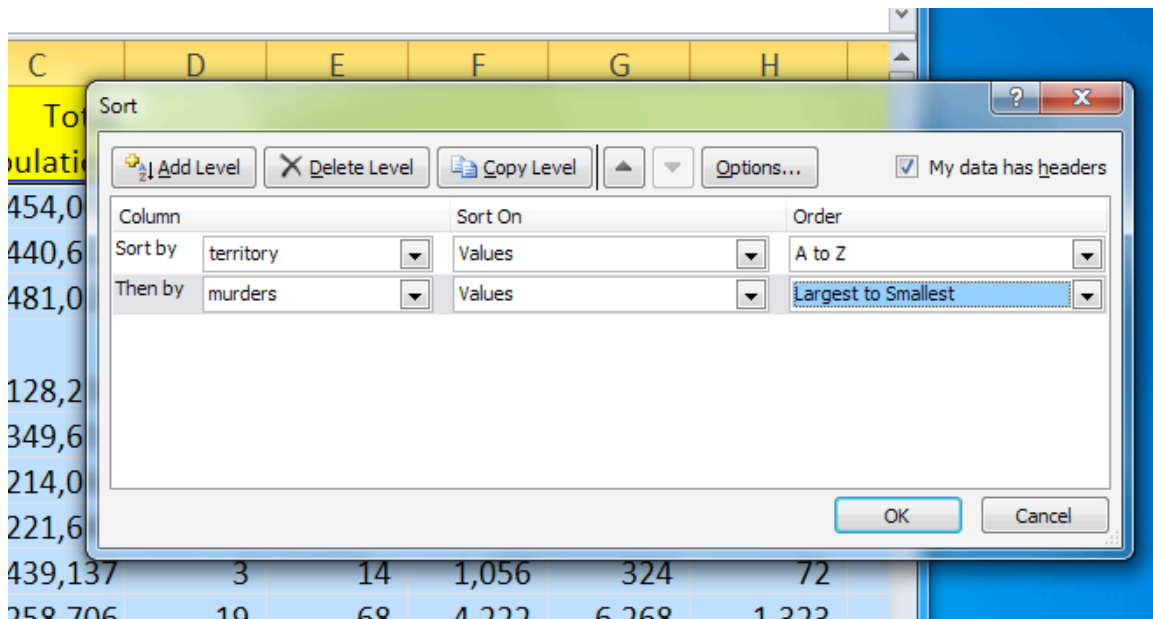
But beware! Put the cursor in the column, but DO NOT select the column letter (D, in this case) and then sort. Consider the example below:



C	D	E	F	G	H
Total	murders	rapes	burglary	car theft	robberies
454,002	2	4	118	61	16
440,613	2	32	2,545	359	144
481,028	0	28	2,152	196	141
128,230	1	15	470	32	20
349,651	2	27	1,088	131	76
214,068	0	6	565	132	41
221,687	1	19	470	32	20
439,137	3	14	1,088	131	76
1,258,706	19	68	4,222	6,268	1,323
213,474	0	15	366	26	15

You will get that warning message, but don't choose "Continue with the current selection". That 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 the "Data" tab and then the "Sort" icon, and then choose the variables by which you wish to sort. (Click "Add level" to add new levels.) Then click "OK".

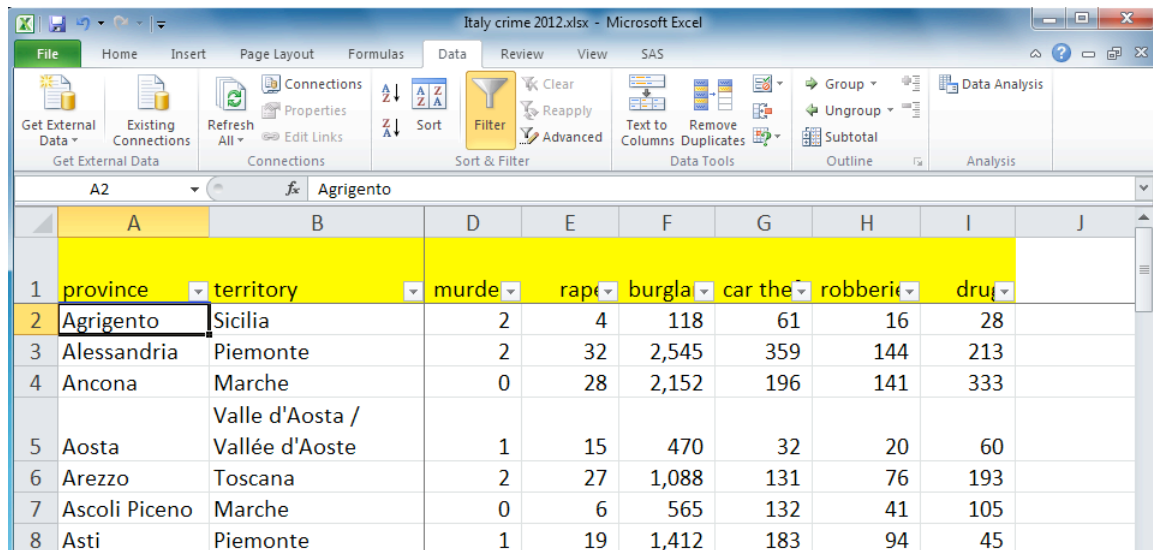


The result will be this:

	A	B	C	D	E	F	G	H
			Total					
1	province	territory	population	murders	rapes	burglary	car theft	robberies
2	Chieti	Abruzzo	397,123	5	30	1,068	431	112
3	Pescara	Abruzzo	323,184	4	28	1,033	809	213
4	L'Aquila	Abruzzo	309,820	1	17	908	157	85
5	Teramo	Abruzzo	312,239	1	15	984	195	101
6	Potenza	Basilicata	383,791	3	21	607	208	51
7	Matera	Basilicata	203,726	0	10	308	122	26
8	Reggio di Calabria	Calabria	566,977	25	34	1,051	1,762	298
9	Vibo Valentia	Calabria	166,560	12	8	331	301	68
10	Cosenza	Calabria	734,656	6	55	1,695	1,445	207
11	Catanzaro	Calabria	368,597	5	25	695	923	92

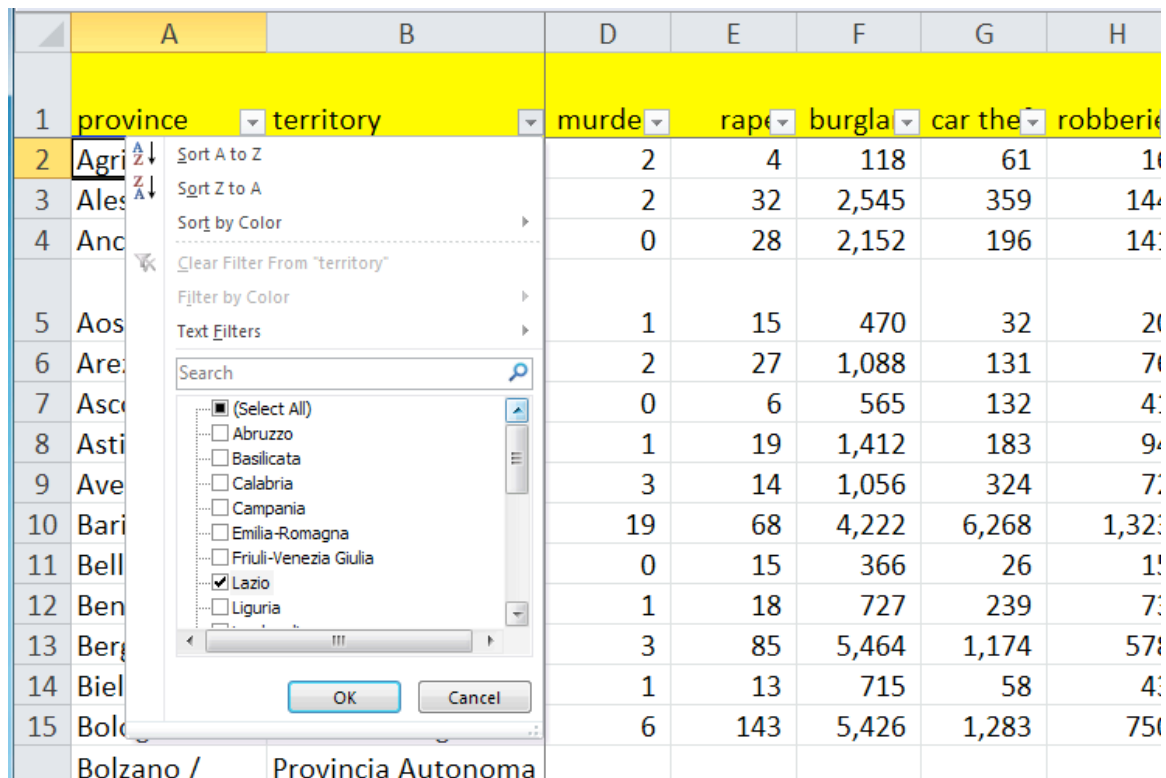
FILTERING

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 the "Data" tab, then click "Filter". Small buttons will appear at the top of each column:



	A	B	D	E	F	G	H	I	J
1	province	territory	murde	rape	burgla	car the	robberie	drug	
2	Agrigento	Sicilia	2	4	118	61	16	28	
3	Alessandria	Piemonte	2	32	2,545	359	144	213	
4	Ancona	Marche	0	28	2,152	196	141	333	
5	Aosta	Valle d'Aosta / Vallée d'Aoste	1	15	470	32	20	60	
6	Arezzo	Toscana	2	27	1,088	131	76	193	
7	Ascoli Piceno	Marche	0	6	565	132	41	105	
8	Asti	Piemonte	1	19	1,412	183	94	45	

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, like this:



	A	B	D	E	F	G	H
1	province	territory	murde	rape	burgla	car the	robberie
2	Agrigento	Sicilia	2	4	118	61	16
3	Alessandria	Piemonte	2	32	2,545	359	144
4	Ancona	Marche	0	28	2,152	196	141
5	Aosta	Valle d'Aosta / Vallée d'Aoste	1	15	470	32	20
6	Arezzo	Toscana	2	27	1,088	131	76
7	Ascoli Piceno	Marche	0	6	565	132	41
8	Asti	Piemonte	1	19	1,412	183	94
9	Avezzano	Lazio	3	14	1,056	324	76
10	Bari	Puglia	19	68	4,222	6,268	1,323
11	Belluno	Friuli-Venezia Giulia	0	15	366	26	15
12	Benvento	Campania	1	18	727	239	76
13	Bergamo	Lombardia	3	85	5,464	1,174	578
14	Biel	Valle d'Aosta / Vallée d'Aoste	1	13	715	58	45
15	Bologna	Emilia-Romagna	6	143	5,426	1,283	750
	Bolzano /	Provincia Autonoma					

This is the result:

	A	B	D	E	F	G	H	I
1	province	territory	murde	rape	burgla	car the	robberie	drug
36	Frosinone	Lazio	1	25	1,114	331	129	220
44	Latina	Lazio	8	48	2,644	1,022	307	341
78	Rieti	Lazio	0	7	539	108	31	92
80	Roma	Lazio	37	422	15,272	20,028	4,062	3,891
104	Viterbo	Lazio	1	36	966	189	59	199
105								
106								
107								

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 2,000. You start by filtering Burglaries like this:

	murde	rape	burgla	car the	robberie	drug
	61	16	28			
	359	144	213			
	196	141	333			
					60	
					193	
					105	
					45	
					126	
					553	
					66	
					82	
					517	
					103	
	1,283	750	782			

then this...

Custom AutoFilter

Show rows where:

burglary

is greater than or equal to

5000

☒ And
 ☐ Or

Use ? to represent any single character

Use * to represent any series of characters

OK

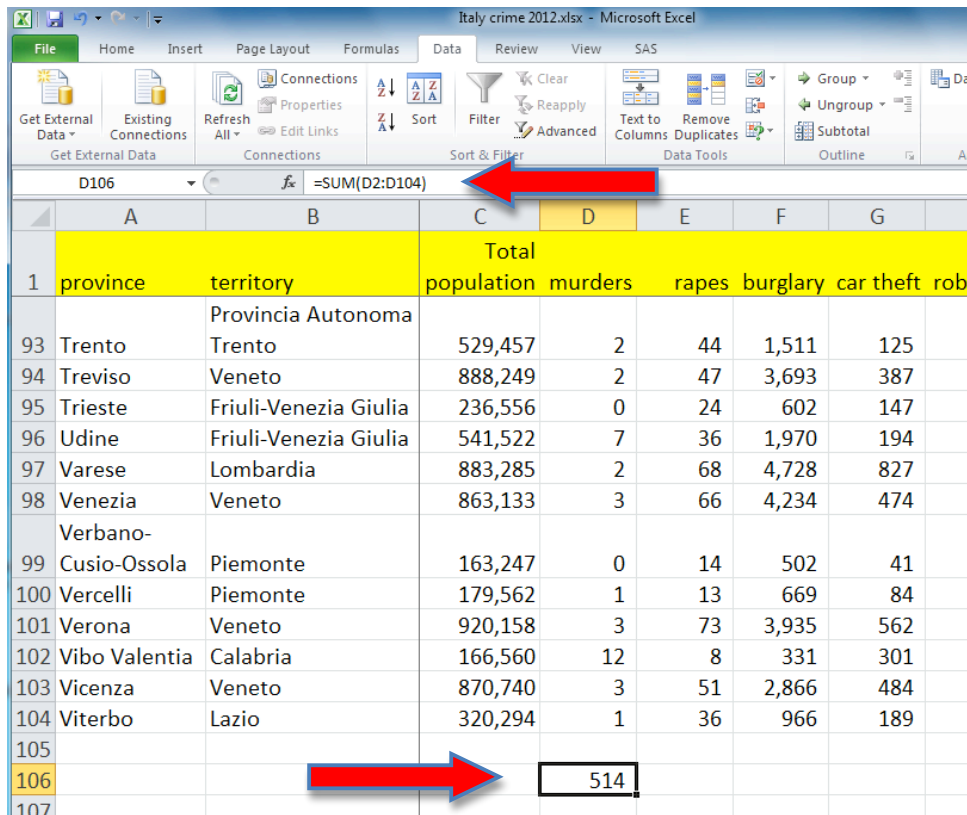
Cancel

Do the same for Car Thefts, and you get this:

	A	B	D	E	F	G	H	I	J
1	province	territory	murde	rape	burgla	car the	robberic	dru	
13	Bergamo	Lombardia	3	85	5,464	1,174	578	517	
15	Bologna	Emilia-Romagna	6	143	5,426	1,283	750	782	
17	Brescia	Lombardia	11	112	6,251	1,716	692	808	
33	Firenze	Toscana	3	134	5,672	802	617	778	
105									
106									

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”. After you hit Enter, the result is this:

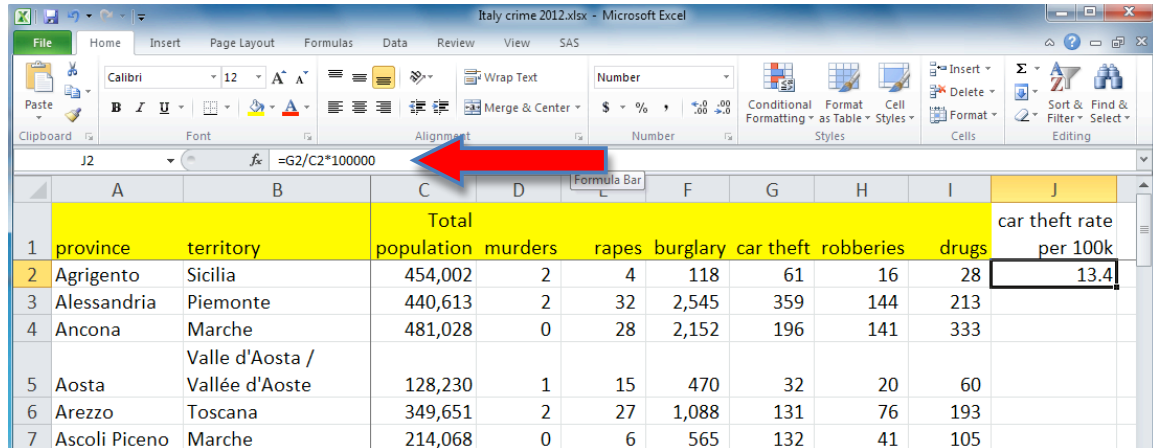


	A	B	C	D	E	F	G	H
1	province	territory	Total					
93	Trento	Provincia Autonoma	population	murders	rapes	burglary	car theft	rob
93	Trento	Trento	529,457	2	44	1,511	125	
94	Treviso	Veneto	888,249	2	47	3,693	387	
95	Trieste	Friuli-Venezia Giulia	236,556	0	24	602	147	
96	Udine	Friuli-Venezia Giulia	541,522	7	36	1,970	194	
97	Varese	Lombardia	883,285	2	68	4,728	827	
98	Venezia	Veneto	863,133	3	66	4,234	474	
99	Verbania	Piemonte	163,247	0	14	502	41	
100	Vercelli	Piemonte	179,562	1	13	669	84	
101	Verona	Veneto	920,158	3	73	3,935	562	
102	Vibo Valentia	Calabria	166,560	12	8	331	301	
103	Vicenza	Veneto	870,740	3	51	2,866	484	
104	Viterbo	Lazio	320,294	1	36	966	189	
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.)

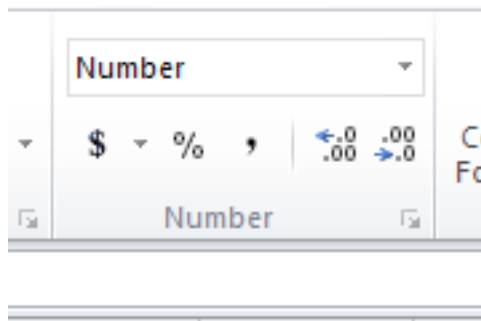
Often you will want to do a calculation on each row of your data table. For instance, you might want to calculate the auto theft rate (the number of cars stolen per 100,000 population), which would let you compare the auto theft problem in cities of different sizes.

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	population	murders	rapes	burglary	car theft	robberies	drugs	car theft rate per 100k
1	Agrigento	Sicilia	454,002	2	4	118	61	16	28	13.4
2	Alessandria	Piemonte	440,613	2	32	2,545	359	144	213	
3	Ancona	Marche	481,028	0	28	2,152	196	141	333	
4	Aosta	Valle d'Aosta / Vallée d'Aoste	128,230	1	15	470	32	20	60	
5	Arezzo	Toscana	349,651	2	27	1,088	131	76	193	
6	Ascoli Piceno	Marche	214,068	0	6	565	132	41	105	

You can format your numbers using various choices in this box under the "Home" tab:



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 dot on 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:

Italy crime 2012.xlsx - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View SAS

Clipboard Font Alignment Number

Calibri 12 A A

B I U

Wrap Text

Number

\$ % , .00 .00

Conditional Formatting as Table Cell Styles

Insert Delete Format

Sort & Find & Filter Select Editing

J5 fx =G5/C5*100000

	A	B	C	D	E	F	G	H	I	J
	province	territory	population	murders	rapes	burglary	car theft	robberies	drugs	car theft rate per 100k
1	Agrigento	Sicilia	454,002	2	4	118	61	16	28	13.4
2	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
10	Bari	Puglia	1,258,706	19	68	4,222	6,268	1,323	553	498.0
11	Belluno	Veneto	213,474	0	15	366	26	15	66	12.2

Notice that the formula changes for each row, so that the Row 5 formula is =G5/C5*100000, and so on. That's what makes Excel so powerful -- the ability to change formulas as you copy down or across.

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

	A	B	C	D	E	F	G	H	I	J
1	province	territory	population	murders	rapes	burglary	car theft	robberies	drugs	car theft rate per 100k
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
9	Reggio di Calabria	Calabria	566,977	25	34	1,051	1,762	298	226	310.8

and sorting in ascending order, the least crime:

	A	B	C	D	E	F	G	H	I	J
1	province	territory	population	murders	rapes	burglary	car theft	robberies	drugs	car theft rate per 100k
2	Belluno	Veneto	213,474	0	15	366	26	15	66	12.2
3	Agirgento	Sicilia	454,002	2	4	118	61	16	28	13.4
4	Bolzano / Bozen	Provincia Autonoma Bolzano / Bozen	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 Trento	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 Verbano-	Valle d'Aosta / Vallée d'Aoste	128,230	1	15	470	32	20	60	25.0

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: =(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.

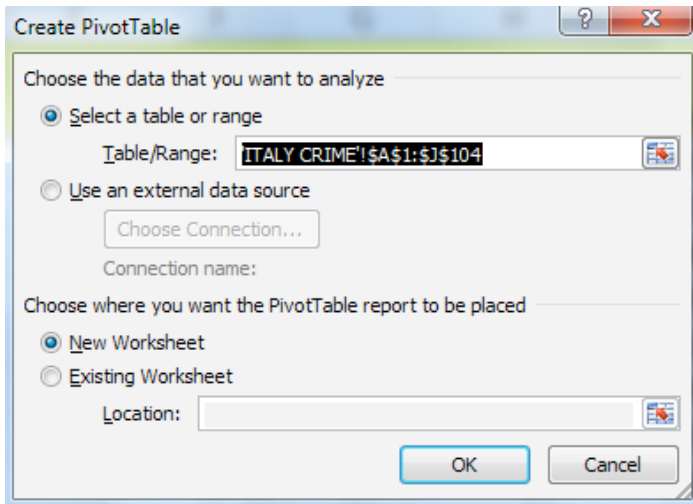
Excel offers well over 200 functions in a variety of categories beyond just math, dates and text: Financial, engineering, database, logical, statistical, etc. But it is unlikely that you will need to be familiar with more than a dozen or so functions, unless you are a journalist with a very specialized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

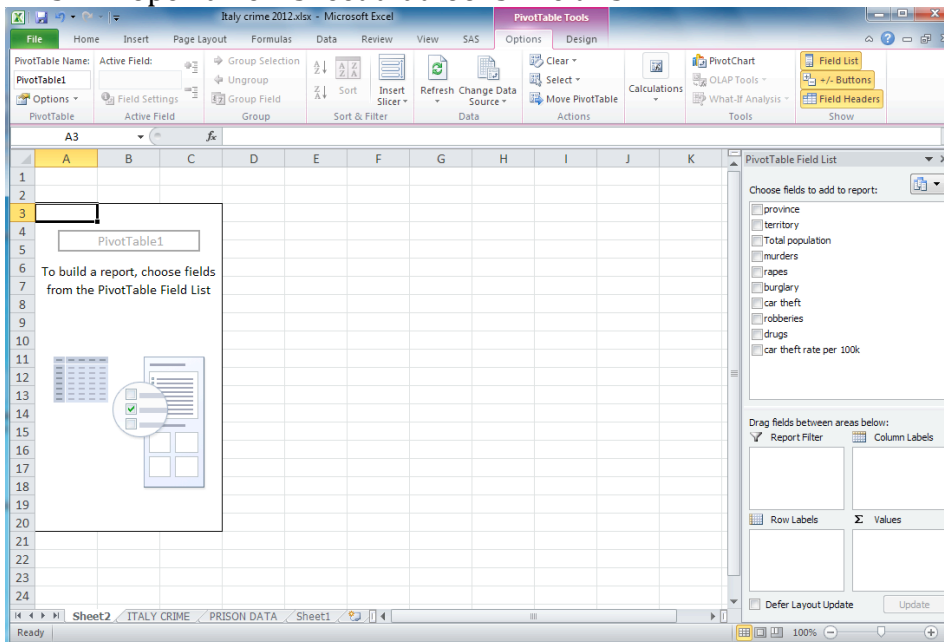
To create a pivot table, every column of your data must have a variable label; in fact, it is always good practice to put in a variable label as soon as you insert or add a new column.

First, you make sure your cursor is on some cell in the table. Then go to the tool bar, click on the "Insert" tab, and then click on the "Pivot Table" icon. A window will pop up that looks like this:



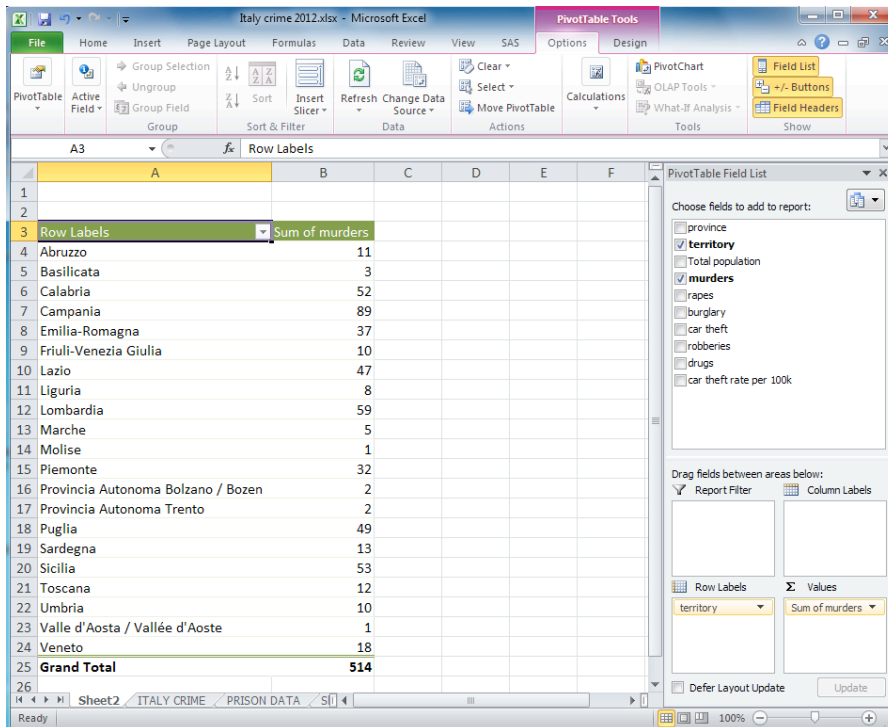
Normally, all you need to do is hit "OK"

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 Territorio, 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:



The screenshot shows the Excel interface with a PivotTable and the PivotTable Field List task pane. The PivotTable has 'territory' as the Row Label and 'Sum of murders' as the Value. The task pane shows the following fields:

- Choose fields to add to report:
 - ☐ province
 - ☒ territory
 - ☐ Total population
 - ☒ murders
 - ☐ rapes
 - ☐ burglary
 - ☐ car theft
 - ☐ robberies
 - ☐ drugs
 - ☐ car theft rate per 100k
- Drag fields between areas below:
 - Report Filter: (empty)
 - Column Labels: (empty)
 - Row Labels: territory
 - Σ Values: Sum of murders

A red arrow points to the task pane.

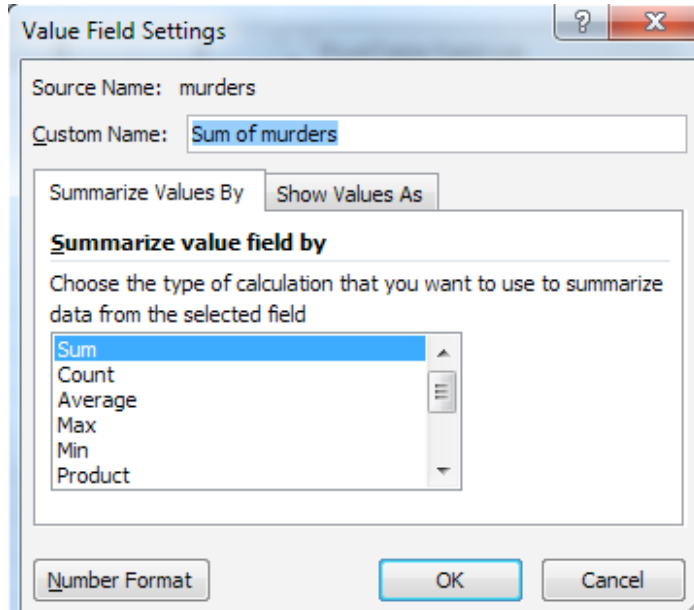
territory	Sum of murders
Abruzzo	11
Basilicata	3
Calabria	52
Campania	89
Emilia-Romagna	37
Friuli-Venezia Giulia	10
Lazio	47
Liguria	8
Lombardia	59
Marche	5
Molise	1
Piemonte	32
Provincia Autonoma Bolzano / Bozen	2
Provincia Autonoma Trento	2
Puglia	49
Sardegna	13
Sicilia	53
Toscana	12
Umbria	10
Valle d'Aosta / Vallée d'Aoste	1
Veneto	18
Grand Total	514

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

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 black down-arrow 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. Click on "Value Field Settings" and you get this:



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:

	Italy crime 2012.csv
1	province,territory>Total population,murders,rapes,burglary,car theft,robberies,drugs
2	Agrigento,Sicilia,454002,2,4,118,61,16,28
3	Alessandria,Piemonte,440613,2,32,2545,359,144,213
4	Ancona,Marce,481028,0,28,2152,196,141,333
5	Aosta,Valle d'Aosta / Vallée d'Aoste,128230,1,15,470,32,20,60
6	Arezzo,Toscana,349651,2,27,1088,131,76,193
7	Ascoli Piceno,Marce,214068,0,6,565,132,41,105
8	Asti,Piemonte,221687,1,19,1412,183,94,45
9	Avellino,Campania,439137,3,14,1056,324,72,126
10	Bari,Puglia,1258706,19,68,4222,6268,1323,553

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 example, 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**. I will be glad to give you advice if I can.