Tutorial for downloading and analyzing data from the Atlantic Canada Opportunities Agency

The agency, which goes by the acronym ACOA, is one of many federal institutions that uploads data to the federal government's open data site, which is a repository for hundreds of datasets, some useful for journalists, some not. "Open data" is a policy adopted by developing countries and a growing number of Third World Nations. The policy, which has been dubbed a "movement", contains two important principles for journalists: a government's commitment to sharing data should be a default position; and that open data is equated to open government where citizens possess the right to gain easy access to downloadable data, and even the right to demand datasets absent from the website.

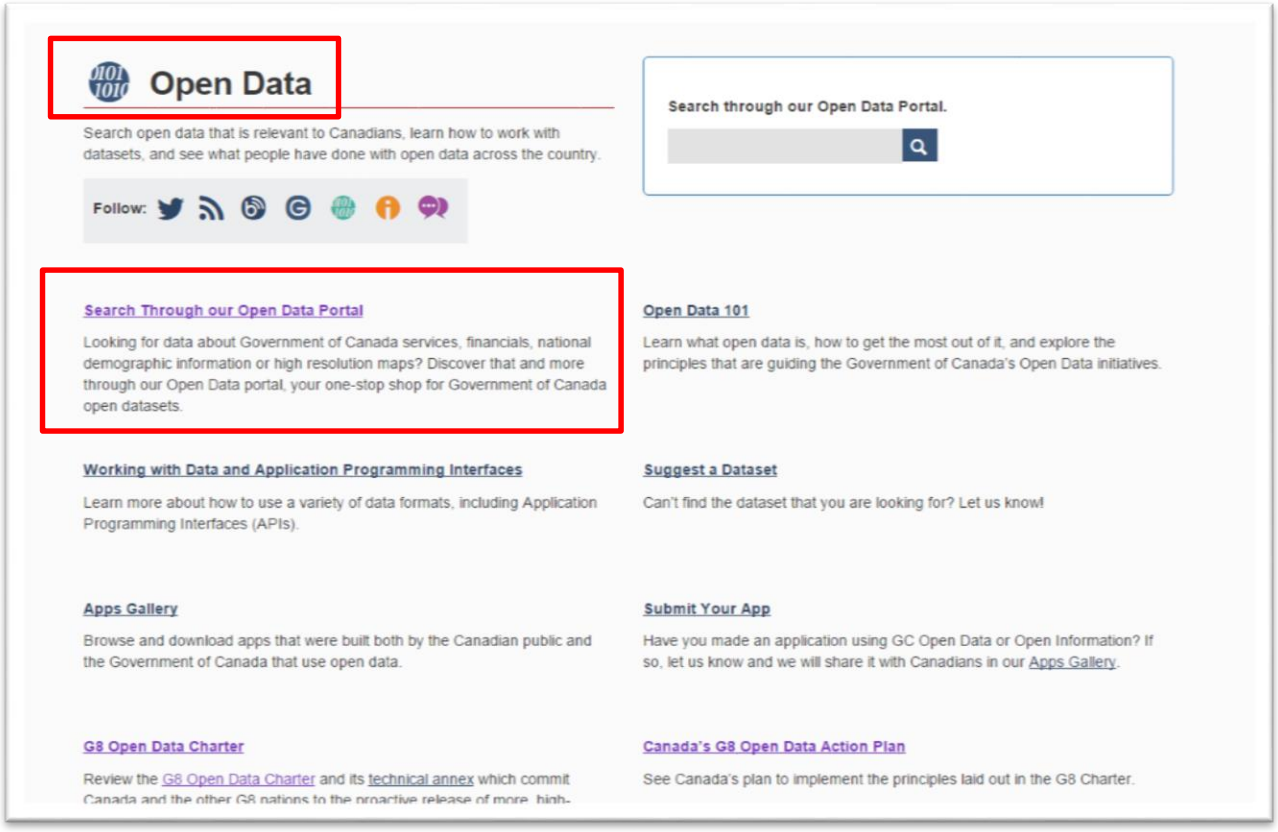
The good news for journalists is that open data is not only a global phenomenon, but one that has grown roots closer to home at the provincial and municipal levels. To date, news outlets like the Toronto Star have used it to tell stories about missed garbage complaints, and suicide calls.

You'll find a list of government open data websites by clicking here.

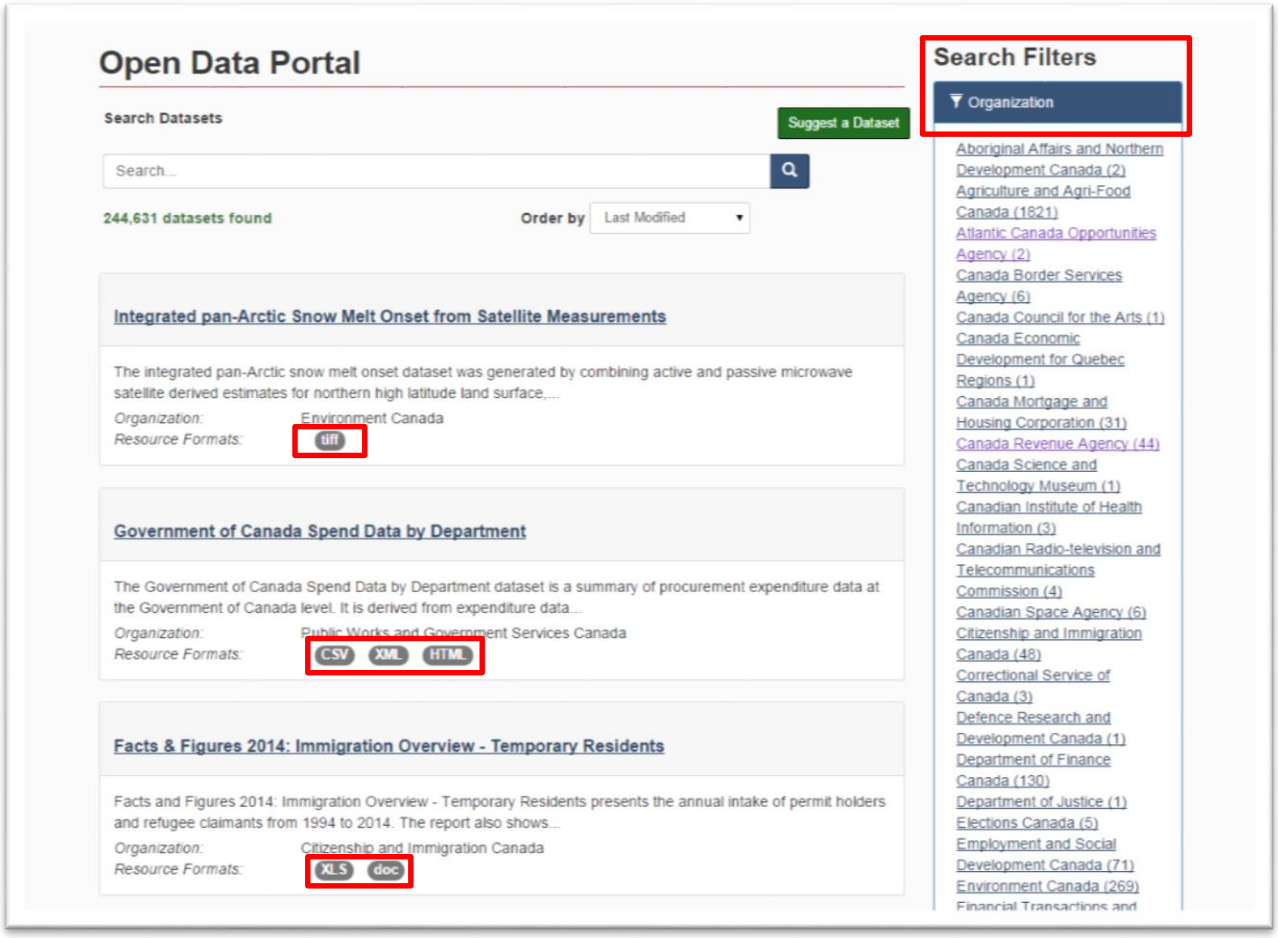
Before we shift to the Atlantic Opportunities Agency, a word about the datasets you'll encounter on these websites. The datasets in question are in what are called "machine readable format", meaning they can be downloaded and opened in a spreadsheet such as Excel.

For the most part, the files are in one of three formats: csv, or comma separated value; tab-delimited text file; or Excel).

Now, let's shift to our tutorial. Go to the federal open data site, and select the "Open Data" icon.



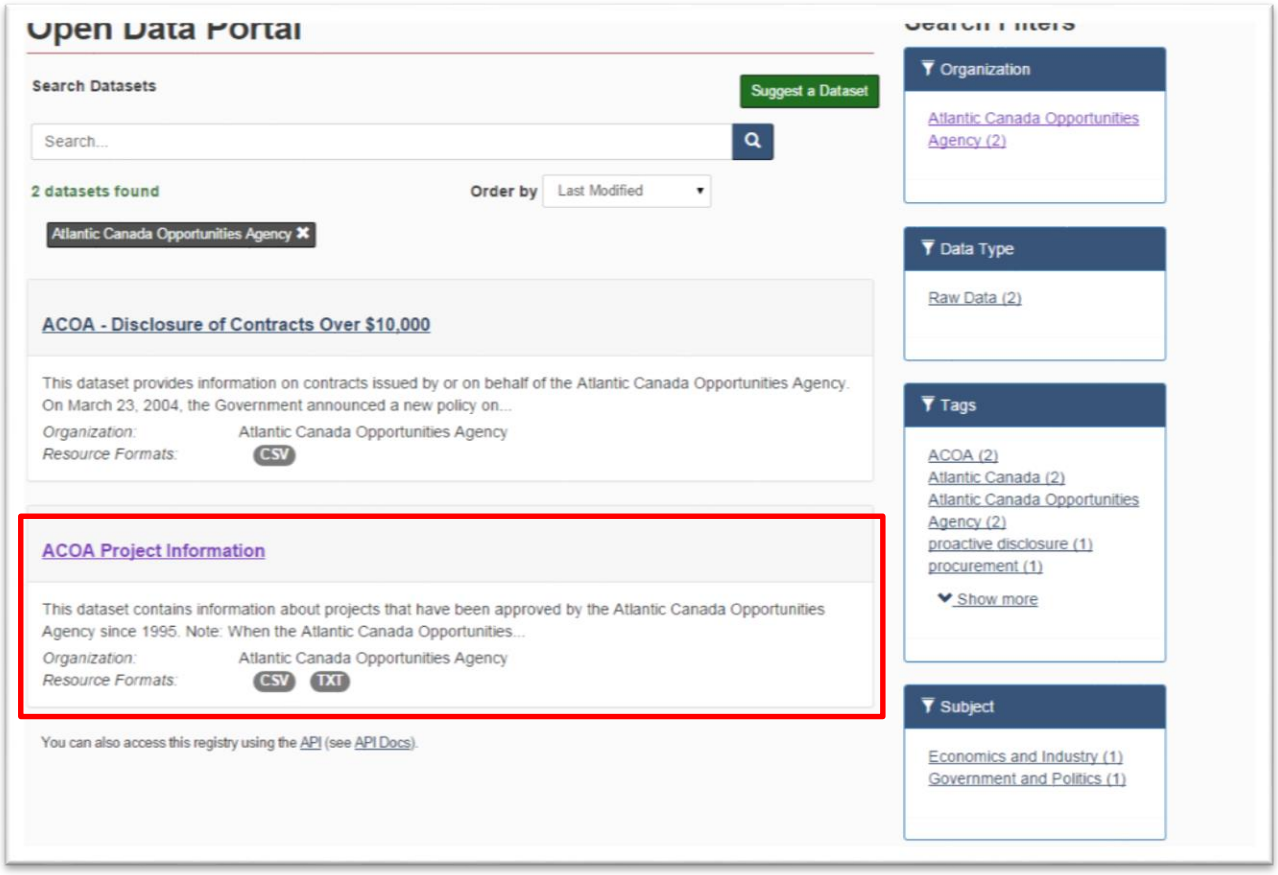
Select the first link to the top left. On the right-hand side, you'll notice a list of federal organizations, click on the "Show more" link to obtain a full list.



Please take a minute to notice the file formats we discussed above. In addition to the csv formats, the icons also indicate XML, HTML and doc. As long as you download a file with a “csv”, “txt”, “xls” (the older, pre-2007 version of Excel), or xlsx (any version from 2007 and on), you don’t have to worry about these other formats. However, it’s worth knowing that they exist.

The number beside each institution indicates the number of datasets. It’s also worth noting that this portal is a living document, meaning that departments continually add new datasets.

Select the Atlantic Canada Opportunities Agency.

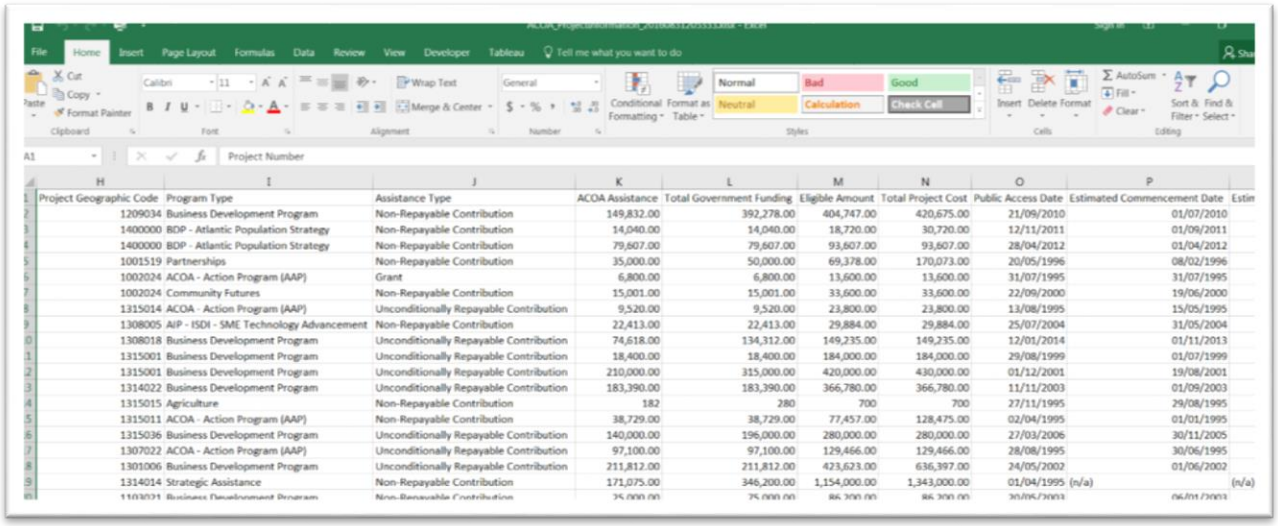


Most, if not all, federal institutions disclose the contracts they award that are worth more than $10,000. These, too, are datasets rich with story ideas about who's getting federal money and how much. There is also a lot of cash to follow in the second link, "ACOA Project Information". It contains data in the "csv" and "txt" format. While both can be opened in Excel, you're better off selecting the csv file.

Click on the "csv" download button, open your download folder, move the file to another folder and open it in Excel. Also be sure to download the "Data Dictionary", which is a text file. It explains the contents in the columns. If a dataset does not come with a data dictionary, "readme" or some sort of file that explains the contents and the frequency with which the dataset is updated, then be sure to

demand one. A dataset without a dictionary is practically useless, even if some of the column labels seem to be self-explanatory.

After you download the csv file, opening it should produce something that looks like this after re-adjusting the number columns to get rid of the hash marks (##########)



Project Geographic Code	Program Type	Assistance Type	ACOA Assistance	Total Government Funding	Eligible Amount	Total Project Cost	Public Access Date	Estimated Commencement Date
1209034	Business Development Program	Non-Repayable Contribution	149,832.00	392,278.00	404,747.00	420,675.00	21/09/2010	01/07/2010
1400000	BDP - Atlantic Population Strategy	Non-Repayable Contribution	14,040.00	14,040.00	18,720.00	30,720.00	12/11/2011	01/09/2011
1400000	BDP - Atlantic Population Strategy	Non-Repayable Contribution	79,607.00	79,607.00	93,607.00	93,607.00	28/04/2012	01/04/2012
1001519	Partnerships	Non-Repayable Contribution	35,000.00	50,000.00	69,378.00	170,073.00	20/05/1996	08/02/1996
1002024	ACOA - Action Program (AAP)	Grant	6,800.00	6,800.00	13,600.00	13,600.00	31/07/1995	31/07/1995
1002024	Community Futures	Non-Repayable Contribution	15,001.00	15,001.00	33,600.00	33,600.00	22/09/2000	19/06/2000
1315014	ACOA - Action Program (AAP)	Unconditionally Repayable Contribution	9,520.00	9,520.00	23,800.00	23,800.00	13/08/1995	15/05/1995
1308005	AIP - ISDI - SME Technology Advancement	Non-Repayable Contribution	22,413.00	22,413.00	29,884.00	29,884.00	25/07/2004	31/05/2004
1308018	Business Development Program	Unconditionally Repayable Contribution	74,618.00	134,312.00	149,235.00	149,235.00	12/01/2014	01/11/2013
1315001	Business Development Program	Unconditionally Repayable Contribution	18,400.00	18,400.00	184,000.00	184,000.00	29/08/1999	01/07/1999
1315001	Business Development Program	Unconditionally Repayable Contribution	210,000.00	315,000.00	420,000.00	430,000.00	01/12/2001	19/08/2001
1314022	Business Development Program	Unconditionally Repayable Contribution	183,390.00	183,390.00	366,780.00	366,780.00	11/11/2003	01/09/2003
1315015	Agriculture	Non-Repayable Contribution	182	280	700	700	27/11/1995	29/08/1995
1315011	ACOA - Action Program (AAP)	Non-Repayable Contribution	38,729.00	38,729.00	77,457.00	128,475.00	02/04/1995	01/01/1995
1315036	Business Development Program	Unconditionally Repayable Contribution	140,000.00	196,000.00	280,000.00	280,000.00	27/03/2006	30/11/2005
1307022	ACOA - Action Program (AAP)	Unconditionally Repayable Contribution	97,100.00	97,100.00	129,466.00	129,466.00	28/08/1995	30/06/1995
1301006	Business Development Program	Unconditionally Repayable Contribution	211,812.00	211,812.00	423,623.00	636,397.00	24/05/2002	01/06/2002
1314014	Strategic Assistance	Non-Repayable Contribution	171,075.00	346,200.00	1,154,000.00	1,343,000.00	01/04/1995	(n/a)

Expand the column widths to make sure you can see the information. Copy the website's URL, paste it into the first available cell in the first row, make a copy of the Excel file, and work with that one. Rule number one when working with data: ALWAYS, ALWAYS, ALWAYS, ALWAYS, ALWAYS work from the back-up copy.

Now it's time to "interview" the data. In other words, study the information in the table to discover what it can and can't tell you, and what questions you need to ask the person in charge. As mentioned above, many datasets at open-data portals contain data dictionary or so-called "readme" files that explain the content in each column. In general, tables with these datasets contain three types of information: numbers, dates and text.

You'll know if a value is a number or a date if the information justifies to the right. If the justification is to the left, you're dealing with text. This is a crucial distinction because a spreadsheet cannot perform math on text. So if your spreadsheet is reading a value as text instead of a number, you may have to reformat it as a number or currency. Downloading data from the Internet also usually involves a lot of reformatting: numbers to currency; adding 1000 separators to numbers, etc. So develop a patience for formatting.

And speaking of formatting, let's reformat the columns with the dollar amounts as currency with no decimal points. The quickest way to do this is highlighting each simultaneously, right-clicking to obtain your short-cut menu, selecting the "format" option and choosing currency with no decimal points.

There are two ways to determine the number of rows or records in your table. Highlight a column to activate the number count feature on the border below the table. If a number is absent, click on the border to obtain a menu and select "COUNT", which adds up the number of rows in the table. Some versions of Excel allow you to select a number of these features. Others only allow one selection at a time.

The second way to determine the number of rows is to use the vertical scroll bar on your right to navigate to the bottom of the table and read the row number to the left.

Navigate up and down: write the names of the column names on a sheet of paper (it's good practice to take plenty of notes when interviewing your datasets), and describe the information they contain. Recording information about the data you've just downloaded is a good way to slow yourself down to find out what the data can tell you, what

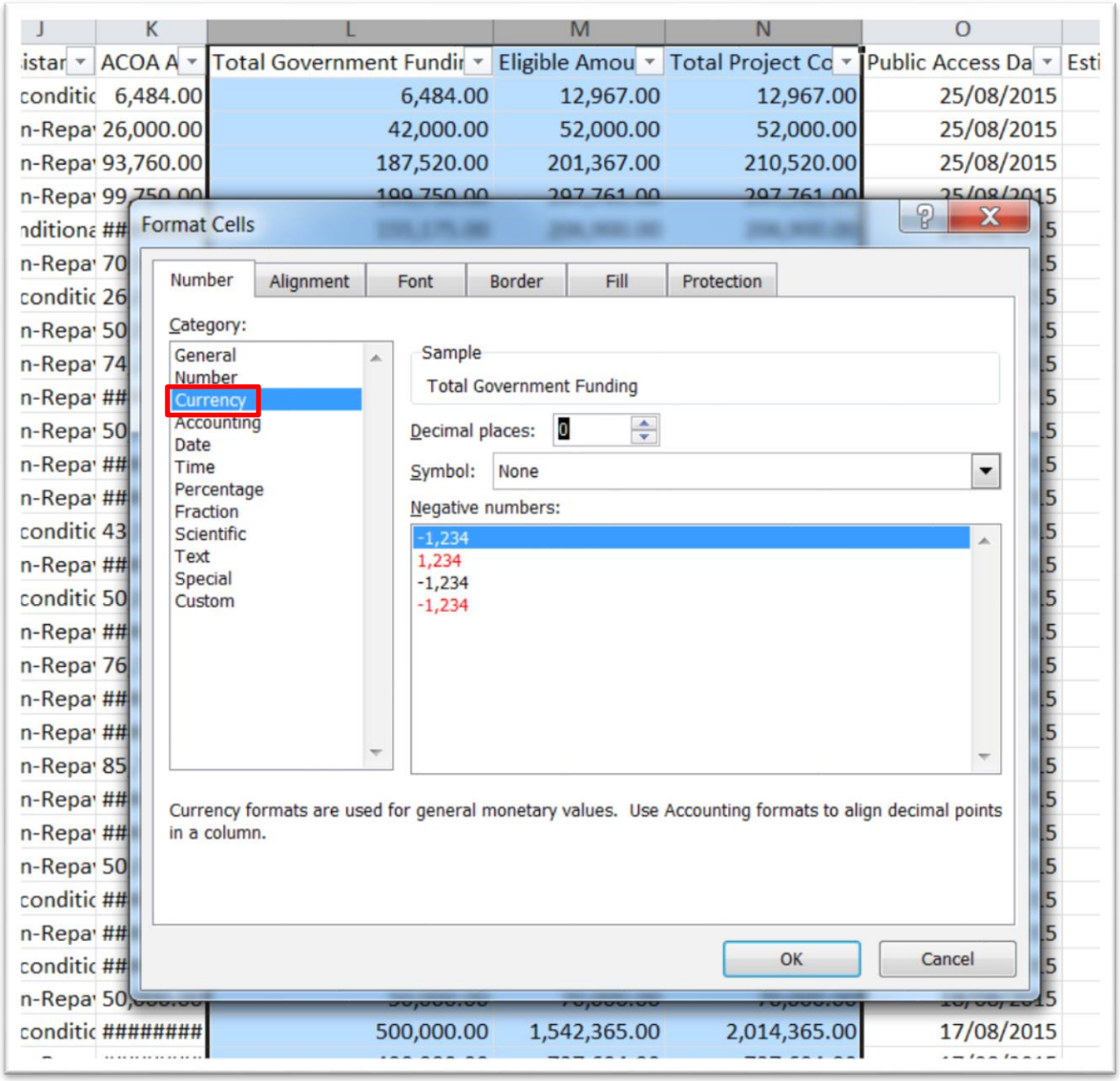
it can't, what's unclear and in need of follow-up. Also pay attention to whacky dates or other bits of information that appear to be mistakes, in large part because they usually are. To use the old saying that has become cliché among data journalist, "all data is dirty". Assume that it contains mistakes. Assume you'll have to do lots of cleaning, a skill that we will perfect throughout this course.

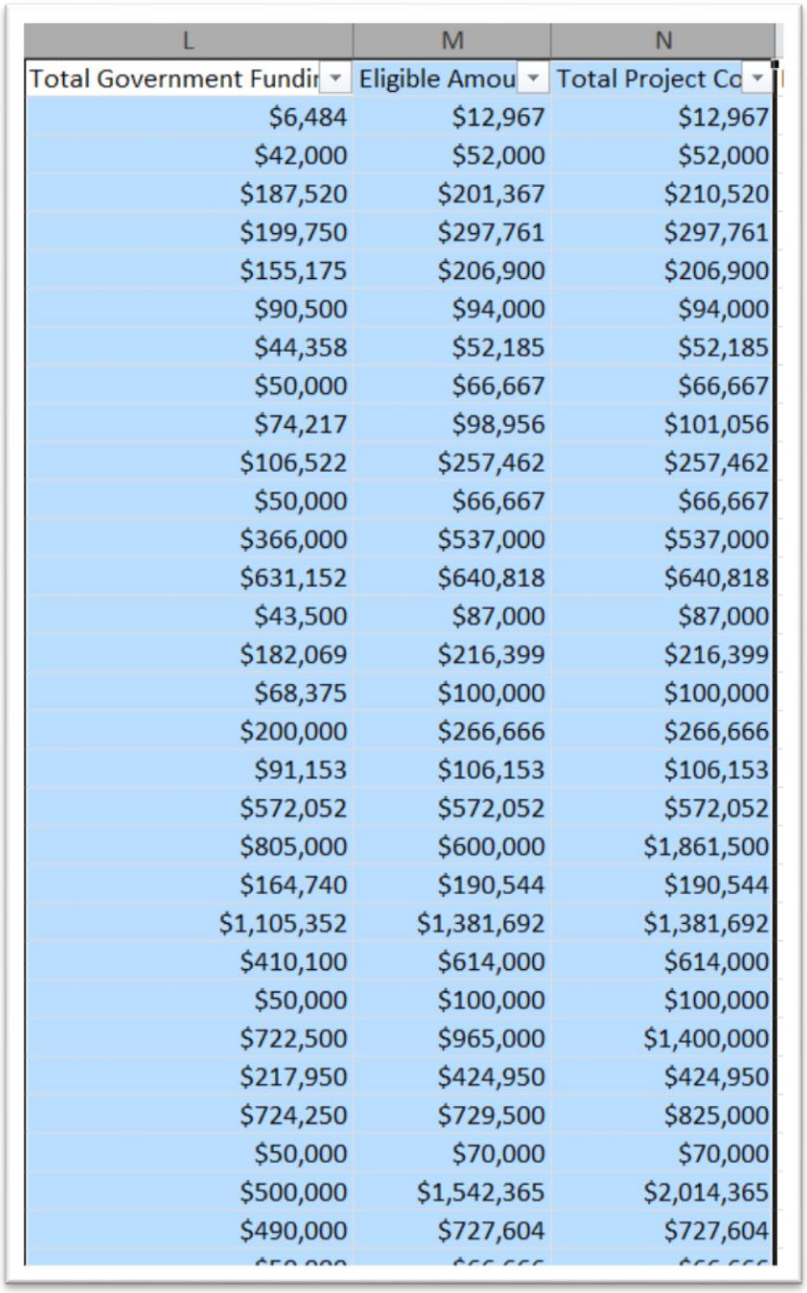
Now let's sort the data fields to determine the dataset's age. There are three date columns. Sort column O, the "Public Access Date", in descending order.

O	P	Q
Public Access Date	Estimated Commencement Date	Estimated Completion Date
26/08/2016	01/07/2016	30/06/2017
26/08/2016	03/05/2016	30/06/2016
26/08/2016	31/05/2016	31/08/2016
26/08/2016	31/05/2016	30/11/2017
25/08/2016	01/06/2016	01/06/2017
24/08/2016	01/04/2016	31/03/2017
23/08/2016	01/08/2016	31/03/2017
23/08/2016	19/09/2016	31/12/2016
23/08/2016	01/07/2016	30/06/2019
23/08/2016	01/07/2016	30/04/2017
23/08/2016	31/08/2016	31/03/2018
23/08/2016	30/06/2016	30/06/2017
23/08/2016	01/06/2016	31/05/2017
22/08/2016	30/04/2016	31/03/2017
22/08/2016	30/06/2016	31/12/2016
22/08/2016	30/06/2016	30/06/2017
22/08/2016	30/07/2016	31/03/2017
21/08/2016	01/07/2016	27/01/2017
21/08/2016	15/05/2016	31/03/2017
21/08/2016	01/07/2016	31/03/2017
21/08/2016	31/08/2016	31/12/2016
21/08/2016	31/08/2016	31/12/2016
21/08/2016	16/05/2016	31/12/2016
21/08/2016	30/06/2016	31/12/2017
21/08/2016	01/05/2016	30/06/2018
20/08/2016	01/03/2016	30/06/2016
20/08/2016	30/05/2016	31/03/2017

Since we downloaded this data on August 31, 2016, the dataset is fairly recent, though it's worth asking how frequently it is updated. This is another important query to make before ever using data.

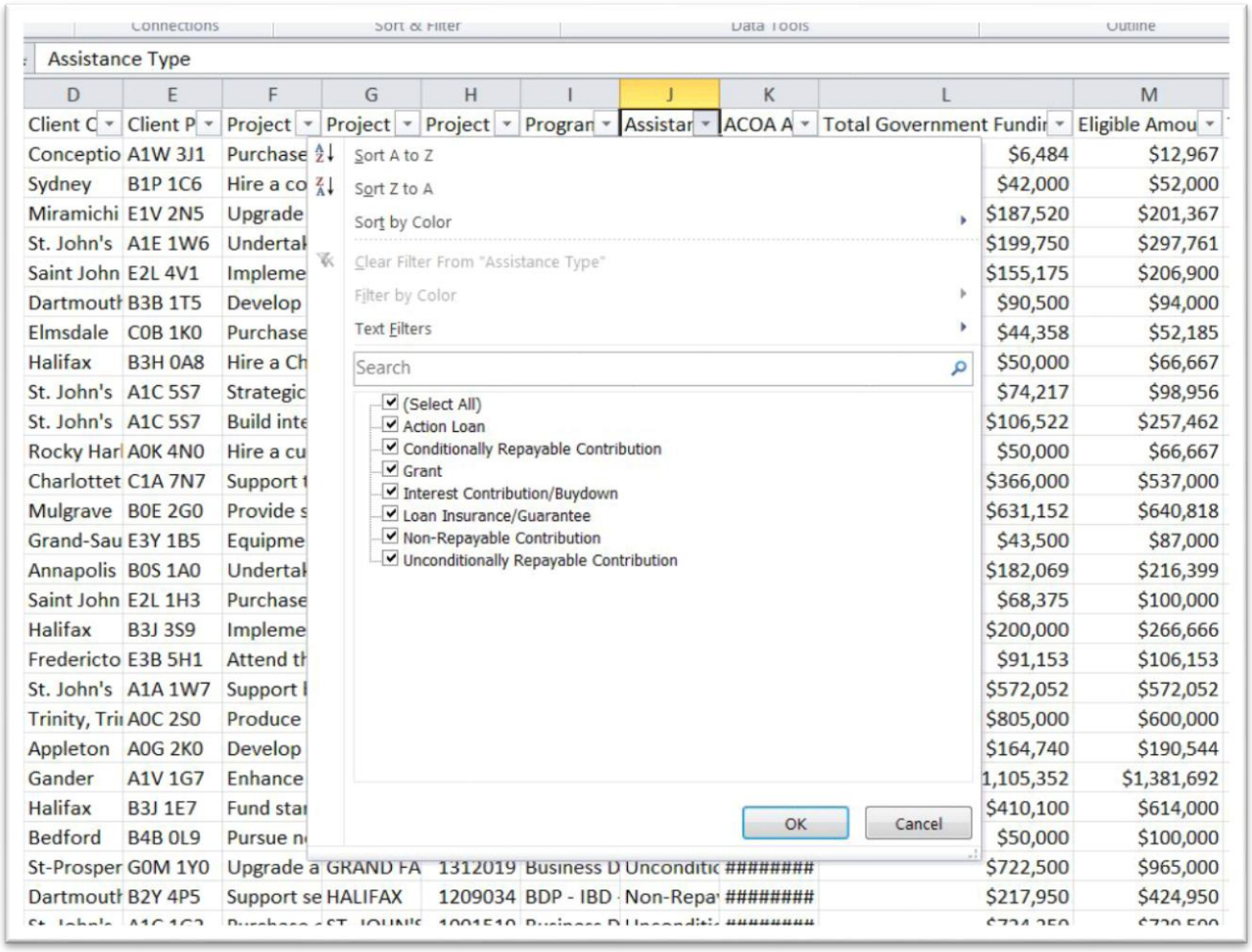
Now apply filter by clicking on the filter icon – the funnel -- to the far left located on the menu's "Data" section. You will get a dialog box that looks something like this. Select currency.





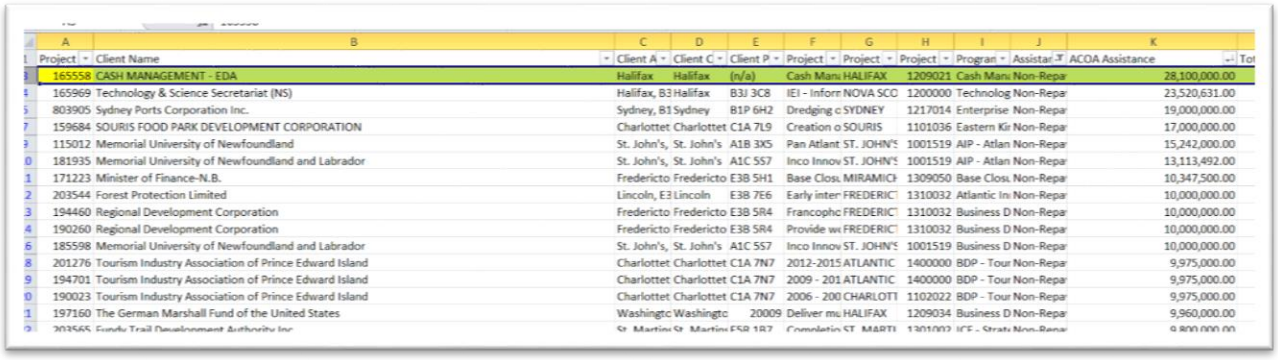
L	M	N
Total Government Fundir	Eligible Amou	Total Project Co
$6,484	$12,967	$12,967
$42,000	$52,000	$52,000
$187,520	$201,367	$210,520
$199,750	$297,761	$297,761
$155,175	$206,900	$206,900
$90,500	$94,000	$94,000
$44,358	$52,185	$52,185
$50,000	$66,667	$66,667
$74,217	$98,956	$101,056
$106,522	$257,462	$257,462
$50,000	$66,667	$66,667
$366,000	$537,000	$537,000
$631,152	$640,818	$640,818
$43,500	$87,000	$87,000
$182,069	$216,399	$216,399
$68,375	$100,000	$100,000
$200,000	$266,666	$266,666
$91,153	$106,153	$106,153
$572,052	$572,052	$572,052
$805,000	$600,000	$1,861,500
$164,740	$190,544	$190,544
$1,105,352	$1,381,692	$1,381,692
$410,100	$614,000	$614,000
$50,000	$100,000	$100,000
$722,500	$965,000	$1,400,000
$217,950	$424,950	$424,950
$724,250	$729,500	$825,000
$50,000	$70,000	$70,000
$500,000	$1,542,365	$2,014,365
$490,000	$727,604	$727,604

This is a rich dataset that contains many options for filtering. We chose column J, “Assistance Type”.



We were interested in the “Non-Repayable Contribution” category. De-select by clicking on the “(Select All), and then select the one you want. You can also make many selections such as all the assistance types that appear to be grants that institutions don’t have to pay back.

Now sort column K in descending order to find out which institution received the largest loan that it didn’t have to repay.



Project	Client Name	Client A	Client C	Client P	Project	Project	Project	Program	Assistar	ACOA Assistance
165558	CASH MANAGEMENT - EDA	Halifax	Halifax	(n/a)	Cash Manı	HALIFAX	1209021	Cash Manı	Non-Repa	28,100,000.00
165969	Technology & Science Secretariat (NS)	Halifax, B3	Halifax	B3J 3C8	IEI - Inforn	NOVA SCO	1200000	Technolog	Non-Repa	23,520,631.00
803905	Sydney Ports Corporation Inc.	Sydney, B1	Sydney	B1P 6H2	Dredging c	SYDNEY	1217014	Enterprise	Non-Repa	19,000,000.00
159684	SOURIS FOOD PARK DEVELOPMENT CORPORATION	Charlottet	Charlottet	C1A 7L9	Creation o	SOURIS	1101036	Eastern Kir	Non-Repa	17,000,000.00
115012	Memorial University of Newfoundland	St. John's,	St. John's	A1B 3X5	Pan Atlant	ST. JOHN'S	1001519	AIP - Atlan	Non-Repa	15,242,000.00
181935	Memorial University of Newfoundland and Labrador	St. John's,	St. John's	A1C 5S7	Inco Innov	ST. JOHN'S	1001519	AIP - Atlan	Non-Repa	13,113,492.00
171223	Minister of Finance-N.B.	Fredericto	Fredericto	E3B 5H1	Base Closu	MIRAMICH	1309050	Base Closu	Non-Repa	10,347,500.00
203544	Forest Protection Limited	Lincoln, E3	Lincoln	E3B 7E6	Early inter	FREDERIC	1310032	Atlantic In	Non-Repa	10,000,000.00
194460	Regional Development Corporation	Fredericto	Fredericto	E3B 5R4	Francophc	FREDERIC	1310032	Business D	Non-Repa	10,000,000.00
190260	Regional Development Corporation	Fredericto	Fredericto	E3B 5R4	Provide wi	FREDERIC	1310032	Business D	Non-Repa	10,000,000.00
185598	Memorial University of Newfoundland and Labrador	St. John's,	St. John's	A1C 5S7	Inco Innov	ST. JOHN'S	1001519	Business D	Non-Repa	10,000,000.00
201276	Tourism Industry Association of Prince Edward Island	Charlottet	Charlottet	C1A 7N7	2012-2015	ATLANTIC	1400000	BDP - Tour	Non-Repa	9,975,000.00
194701	Tourism Industry Association of Prince Edward Island	Charlottet	Charlottet	C1A 7N7	2009 - 201	ATLANTIC	1400000	BDP - Tour	Non-Repa	9,975,000.00
190023	Tourism Industry Association of Prince Edward Island	Charlottet	Charlottet	C1A 7N7	2006 - 200	CHARLOTT	1102022	BDP - Tour	Non-Repa	9,975,000.00
197160	The German Marshall Fund of the United States	Washingtc	Washingtc	20009	Deliver mu	HALIFAX	1209034	Business D	Non-Repa	9,960,000.00

Of course, we could filter the data in a number of ways. If you are happy with this subset of the data, select and copy it, create a new work sheet by clicking on the blank worksheet at the bottom of the workbook.

You can call it something like “Non-repayable_loans”. You’ll notice that names in databases are devoid of spaces for reasons we’ll explore later. If you want to separate words in a title, use an underscore, which Excel reads as a character. As well, keep the titles concise because you can only use a certain number characters when naming worksheets.

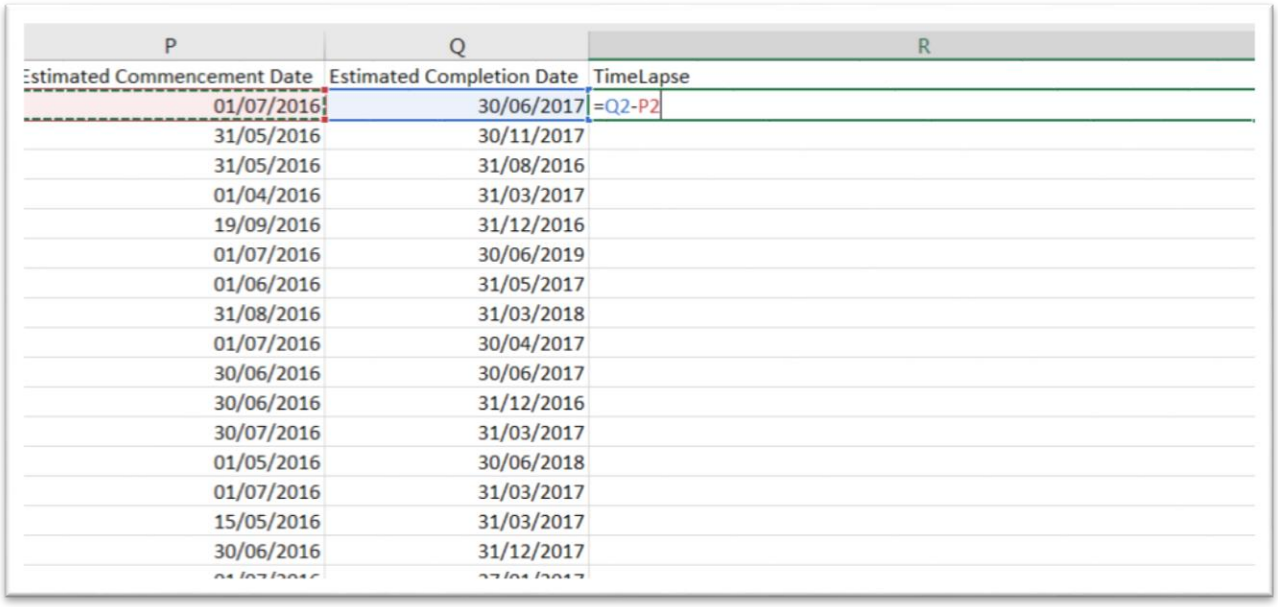
Return to our original worksheet, and get rid of the squiggly lines -- otherwise known as marching ants -- that border the table by scrolling to the far right, activating a cell outside the table and tapping your space bar.

Brace yourselves, now it’s time for a bit of math: Subtraction.

Sort the column 0, the “Public Access Date” in descending order. Create a new column that calculates the number of days between

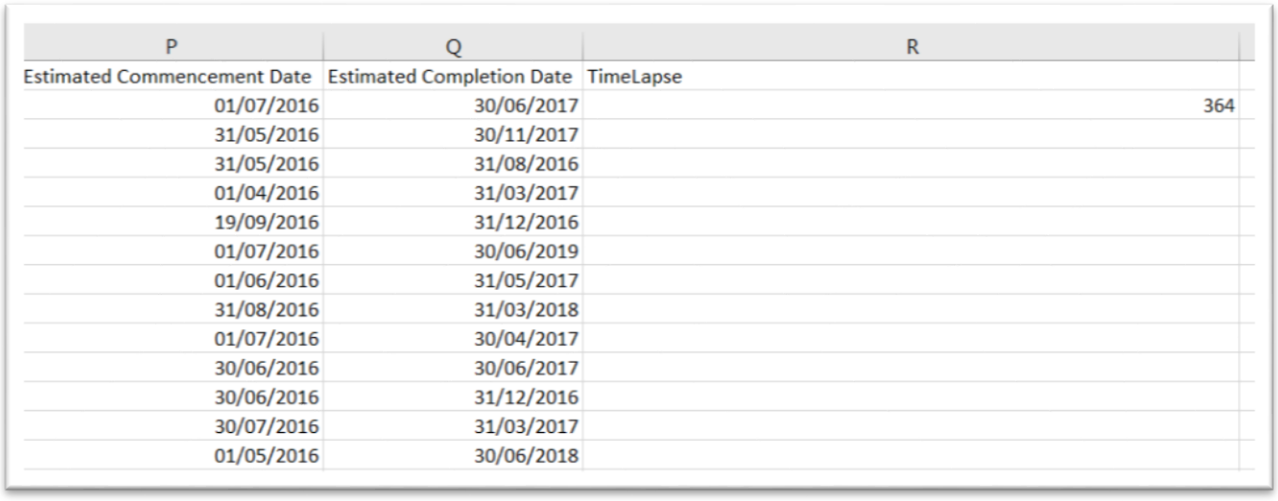
column P, the “Estimated Commencement Date”, and column Q, the “Estimated Completion Date”. We’ll perform this task in column R, which you call something like “Time-lapse”.

Remember, every calculation in Excel begins with an “=” sign. To find out the number of days between the two dates we simply subtract the estimated completion date from the estimated commencement date.



P	Q	R
Estimated Commencement Date	Estimated Completion Date	TimeLapse
01/07/2016	30/06/2017	=Q2-P2
31/05/2016	30/11/2017	
31/05/2016	31/08/2016	
01/04/2016	31/03/2017	
19/09/2016	31/12/2016	
01/07/2016	30/06/2019	
01/06/2016	31/05/2017	
31/08/2016	31/03/2018	
01/07/2016	30/04/2017	
30/06/2016	30/06/2017	
30/06/2016	31/12/2016	
30/07/2016	31/03/2017	
01/05/2016	30/06/2018	
01/07/2016	31/03/2017	
15/05/2016	31/03/2017	
30/06/2016	31/12/2017	

Hit enter.



P	Q	R
Estimated Commencement Date	Estimated Completion Date	TimeLapse
01/07/2016	30/06/2017	364
31/05/2016	30/11/2017	
31/05/2016	31/08/2016	
01/04/2016	31/03/2017	
19/09/2016	31/12/2016	
01/07/2016	30/06/2019	
01/06/2016	31/05/2017	
31/08/2016	31/03/2018	
01/07/2016	30/04/2017	
30/06/2016	30/06/2017	
30/06/2016	31/12/2016	
30/07/2016	31/03/2017	
01/05/2016	30/06/2018	

So the time lapse for the most recent project is 364 days. Now copy the formula to the bottom of the table.

P	Q	R
Estimated Commencement Date	Estimated Completion Date	TimeLapse
01/07/2016	30/06/2017	364
31/05/2016	30/11/2017	548
31/05/2016	31/08/2016	92
01/04/2016	31/03/2017	364
19/09/2016	31/12/2016	103
01/07/2016	30/06/2019	1094
01/06/2016	31/05/2017	364
31/08/2016	31/03/2018	577
01/07/2016	30/04/2017	303
30/06/2016	30/06/2017	365
30/06/2016	31/12/2016	184
30/07/2016	31/03/2017	244
01/05/2016	30/06/2018	790
01/07/2016	31/03/2017	273
15/05/2016	31/03/2017	320
30/06/2016	31/12/2017	549
01/07/2016	27/01/2017	210
31/08/2016	31/12/2016	122
31/08/2016	31/12/2016	122
30/05/2016	31/03/2017	305
11/04/2016	28/02/2017	323
01/03/2016	30/06/2016	121
30/06/2016	31/12/2016	184

Questions to answer when working with specific datasets from open-data websites:

1. How often is the data set updated?

2. How many records or rows does it contain?
3. How does the institution use the data?
4. Who inputs the data?
5. What time period does it cover?
6. What information does it exclude, and is it possible to obtain that information in the spirit of open-data-equals-open-government philosophy?
7. What stories have been done using this data?
8. Is there at least one story idea in this dataset?

To-do list:

Now that you've mastered the steps outlined above. Now it's time to look for story ideas beyond the ones we may have discussed in class.

1. Continue to navigate the dataset, perusing information in the columns we did not discuss in class.
2. Apply the filter once again to narrow the selections in the columns that were not part of the tutorial above.
3. Copy and paste subsets of the data in two more worksheets and think about possible stories the data could tell.
4. Be prepared to discuss the ideas in class.