

Chapter 4

Sorting and Filtering in a Spreadsheet

Summary: So we have all the cleaned-up data we need in our spreadsheet, either because we have created a table, copied and pasted data from the web as we saw in the previous tutorial, or simply opened a file from an open data website that is already compatible with Excel.

Before we even know what to do with the information, we must treat the table just like any other source: determine its strengths and weaknesses, and discern what it can tell us, and what it can't. If we were dealing with a human source, we would use probing questions to get this information, which then helps us determine the usefulness of the source. Well, with data, we can do something similar. That is, interview it to arrive at the same kind of information.

Sorting and filtering are two of the first methods we should use to accomplish this task. The former helps us to determine, which provincial bureaucrat earned the highest salary, which company emitted the highest level of carbon emissions, or which institution received the largest government assistance in the form of a loan or grant.

For this tutorial, we will use the dataset that the Atlantic Canada Opportunities Agencies uses to track the money it dispenses. ACOA, as it is also known, is one of six federal regional development agencies responsible for promoting economic growth in Atlantic Canada, Quebec, Northern Ontario, Southern Ontario, the North and the provinces west of Ontario.

What you will learn:

1. Sorting
2. Filtering

You'll find the ACOA data at the federal government's open data website:

<http://open.canada.ca/data/en/dataset/ad1f4897-3298-4d15-8e5e-094958be7388>

The screenshot shows the Open Data Canada website interface. At the top, there is a search bar and a button labeled 'Search...'. Below the search bar, it says '2 datasets found' and 'Order by Last Modified'. There is a filter button labeled 'Atlantic Canada Opportunities Agency X'. On the right side, there are three sidebar sections: 'Data Type' (Raw Data 2), 'Tags' (ACOA 2, Atlantic Canada 2, Atlantic Canada Opportunities Agency 2, proactive disclosure 1, procurement 1, Show more), and 'Subject' (Economics and Industry 1, Government and Politics 1). The main content area displays two datasets:

- ACOA - Disclosure of Contracts Over \$10,000**
This dataset provides information on contracts issued by or on behalf of the Atlantic Canada Opportunities Agency. On March 23, 2004, the Government announced a new policy on...
Organization: Atlantic Canada Opportunities Agency
Resource Formats: [CSV](#)
- ACOA Project Information**
This dataset contains information about projects that have been approved by the Atlantic Canada Opportunities Agency since 1995. Note: When the Atlantic Canada Opportunities...
Organization: Atlantic Canada Opportunities Agency
Resource Formats: [CSV](#) [TXT](#)

You can also access this registry using the API (see [API Docs](#)).

Click on the “ACOA Project Information” link.

ACOA Project Information

This dataset contains information about projects that have been approved by the Atlantic Canada Opportunities Agency since 1995. Note: When the Atlantic Canada Opportunities Agency or a provincial government department is listed as a client, it is because it has taken the lead on developing, evaluating and/or administering the project. Some of the funding figures available in this dataset may not be inclusive of funding received through joint programs such as the Atlantic Canada Cultural and Economic Partnership, the Atlantic Canada Tourism Partnership or other ACOA-administered programs such as Infrastructure Canada.

Licence:
[Open Government Licence - Canada](#)

Dataset Resources

Resource Name	Format	Language	Link
Dataset	CSV	Bilingual (English and French)	Download
Data Dictionary	TXT	French	Download
Data Dictionary	TXT	English	Download

Select the “Download” tab to the right of the csv file format, which opens a download folder, allowing you to browse to a location on your hard drive where you want to store the data.

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 (# #####) # #####)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Project Nu	Client Nan	Client Addi	Client City	Post Project De	Lo	Project Ge	Program T	Assistance ACOA	Assi: Total	Gove Eligible	Arr Total Proj	Public Acci	Estimated	
2	197595	(ISIS) Immi Halifax, B3 Halifax	B3L 4P1	IBDS fundi HALIFAX	1209034	Business D Non-Repa	#####	#####	#####	#####	#####	#####	#####	#####	
3	198632	(ISIS) Immi Halifax, B3 Halifax	B3L 4P1	Immigrant ATLANTIC	1400000	BDP - Atlai Non-Repa	14,040.00	14,040.00	18,720.00	30,720.00	#####	#####	#####	#####	
4	200029	(ISIS) Immi Halifax, B3 Halifax	B3L 4P1	The Immigr ATLANTIC	1400000	BDP - Atlai Non-Repa	79,607.00	79,607.00	93,607.00	93,607.00	#####	#####	#####	#####	
5	157116	* Quidi Vî St. John's, St. John's	A1B 2Z2	Environme ST. JOHN'S	1001519	Partnershi Non-Repa	35,000.00	50,000.00	69,378.00	#####	#####	#####	#####	#####	
6	154136	* Burin Pei Marystow Marystow AOK 2M0		implement MARYSTO	1002024	ACOA - Ac: Grant	6,800.00	6,800.00	13,600.00	13,600.00	#####	#####	#####	#####	
7	154135	* Burin Pei Marystow Marystow AOK 2M0		Youth Inte MARYSTO	1002024	Community Non-Repa	15,001.00	15,001.00	33,600.00	33,600.00	#####	#####	#####	#####	
8	113504	042691 N. Petit Roch Petit Roch E8J 2K7		To acquire PETIT-ROC	1315014	ACOA - Ac: Unconditio	9,520.00	9,520.00	23,800.00	23,800.00	#####	#####	#####	#####	
9	185785	047100 N. Bouchouch Bouchouch E4S 2J2		Lean Analy BOUCTOU	1308005	AIP - ISDI - Non-Repa	22,413.00	22,413.00	29,884.00	29,884.00	#####	#####	#####	#####	
10	203111	051996 N. Mundlevill Mundlevill E4W 2N5		Building ex RICHIBUC	1308018	Business D Unconditio	74,618.00	#####	#####	#####	#####	#####	#####	#####	

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, save the csv file in an Excel format, and work with that one. Rule number one when working with data: ALWAYS, ALWAYS, ALWAYS, ALWAYS, ALWAYS work from the back-up copy. (NOTE: Excel and any other spreadsheet allows you to work with multiple worksheets that contained smaller excerpts of your original data. A csv file only accommodates one worksheet.)

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. 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.

As we explained in the previous tutorial, 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.

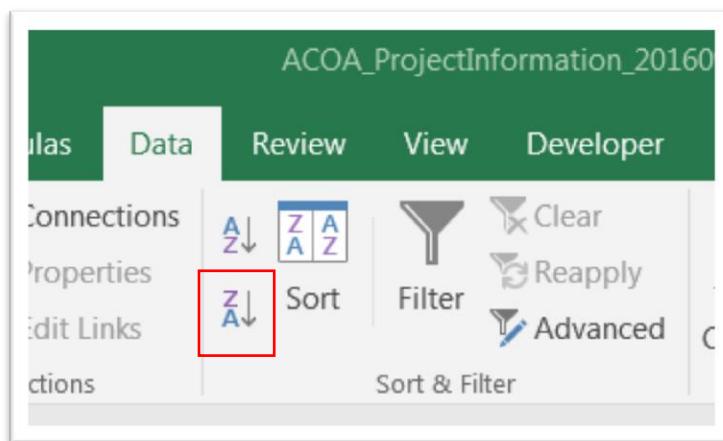
There are two ways to determine the number of rows or records in your table. Highlight a column to activate the number count feature on the border below the table. If a number is absent, click on the border to obtain a menu and select “COUNT”, which adds up the number of rows in the table. Some versions 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 in subsequent tutorials.

Sorting

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. To do so, place your cursor on O2. Then go to the “Sort” icon in the “Data” section of your menu (NOTE: This should be the same for PCs and Macs), and select Z to A, which is the newest or highest to the oldest or lowest.



L	M	N	O	
I	Gove	Eligible An	Total Proj	Public Access Date
00.00	36,000.00	36,000.00		26/08/2016
00.00	15,600.00	15,600.00		26/08/2016
67.00	40,400.00	40,400.00		26/08/2016
20.00	43,241.00	43,241.00		26/08/2016
#####	#####	#####		25/08/2016
#####	#####	#####		24/08/2016
#####	#####	#####		23/08/2016
#####	#####	#####		23/08/2016
#####	#####	#####		23/08/2016
00.00	70,000.00	70,000.00		23/08/2016
#####	#####	#####		23/08/2016
#####	#####	#####		23/08/2016
#####	#####	#####		23/08/2016
#####	#####	#####		22/08/2016
42.00	61,884.00	61,884.00		22/08/2016
#####	#####	#####		22/08/2016
45.00	44,600.00	44,600.00		22/08/2016

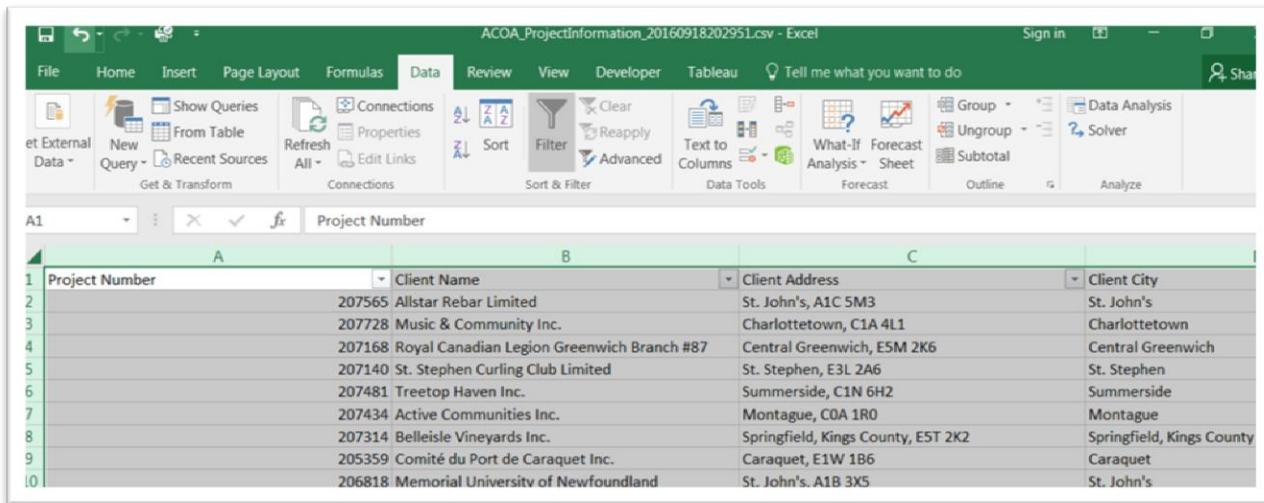
Your dates will look different, depending on when you downloaded the data. You can also sort the columns that contain the dollar values, such as column K which reveals the amount of ACOA assistance.

While sorting allows us to go from highest to lowest, it can be used in conjunction with filtering that allows for a deeper dive into the data. For instance, this dataset goes all the way back to 1988. Unless we are conducting historical research, we probably

don't want to go back that far. The last five years will probably suffice. Filtering allows us to accomplish this task.

Filtering

Filtering adds downward arrows at the top of each column, allowing for the option of filtering on any one of the columns. To apply the filter, click on the icon that looks like a funnel, which you'll find in the same location in the menu that contains the sorting option.



The screenshot shows a Microsoft Excel spreadsheet titled "ACOA_ProjectInformation_20160918202951.csv - Excel". The Data tab is selected in the ribbon. In the "Sort & Filter" group, the "Filter" icon (a funnel) is highlighted. The spreadsheet contains four columns of data: Project Number, Client Name, Client Address, and Client City. The first column, "Project Number", has a filter icon at the top. The data includes entries such as 207565 Allstar Rebar Limited, 207728 Music & Community Inc., 207168 Royal Canadian Legion Greenwich Branch #87, etc. The "Client Address" and "Client City" columns also have filter icons at the top.

A	B	C	D
1 Project Number	Client Name	Client Address	Client City
2 207565	Allstar Rebar Limited	St. John's, A1C 5M3	St. John's
3 207728	Music & Community Inc.	Charlottetown, C1A 4L1	Charlottetown
4 207168	Royal Canadian Legion Greenwich Branch #87	Central Greenwich, E5M 2K6	Central Greenwich
5 207140	St. Stephen Curling Club Limited	St. Stephen, E3L 2A6	St. Stephen
6 207481	Treetop Haven Inc.	Summerside, C1N 6H2	Summerside
7 207434	Active Communities Inc.	Montague, COA 1R0	Montague
8 207314	Belleisle Vineyards Inc.	Springfield, Kings County, E5T 2K2	Springfield, Kings County
9 205359	Comité du Port de Caraquet Inc.	Caraquet, E1W 1B6	Caraquet
10 206818	Memorial University of Newfoundland	St. John's, A1B 3X5	St. John's

Clicking on each filter produces a drop-down menu which organizes numbers from smallest to largest regardless of how you've sorted the data, and arranges text such as names in alphabetical order.

This is a rich dataset that contains lots of information. However, we might want to begin by filtering the dataset for instances where institutions are not required to pay back any money. Or put another way. Institutions that received what critics might characterize as handouts.

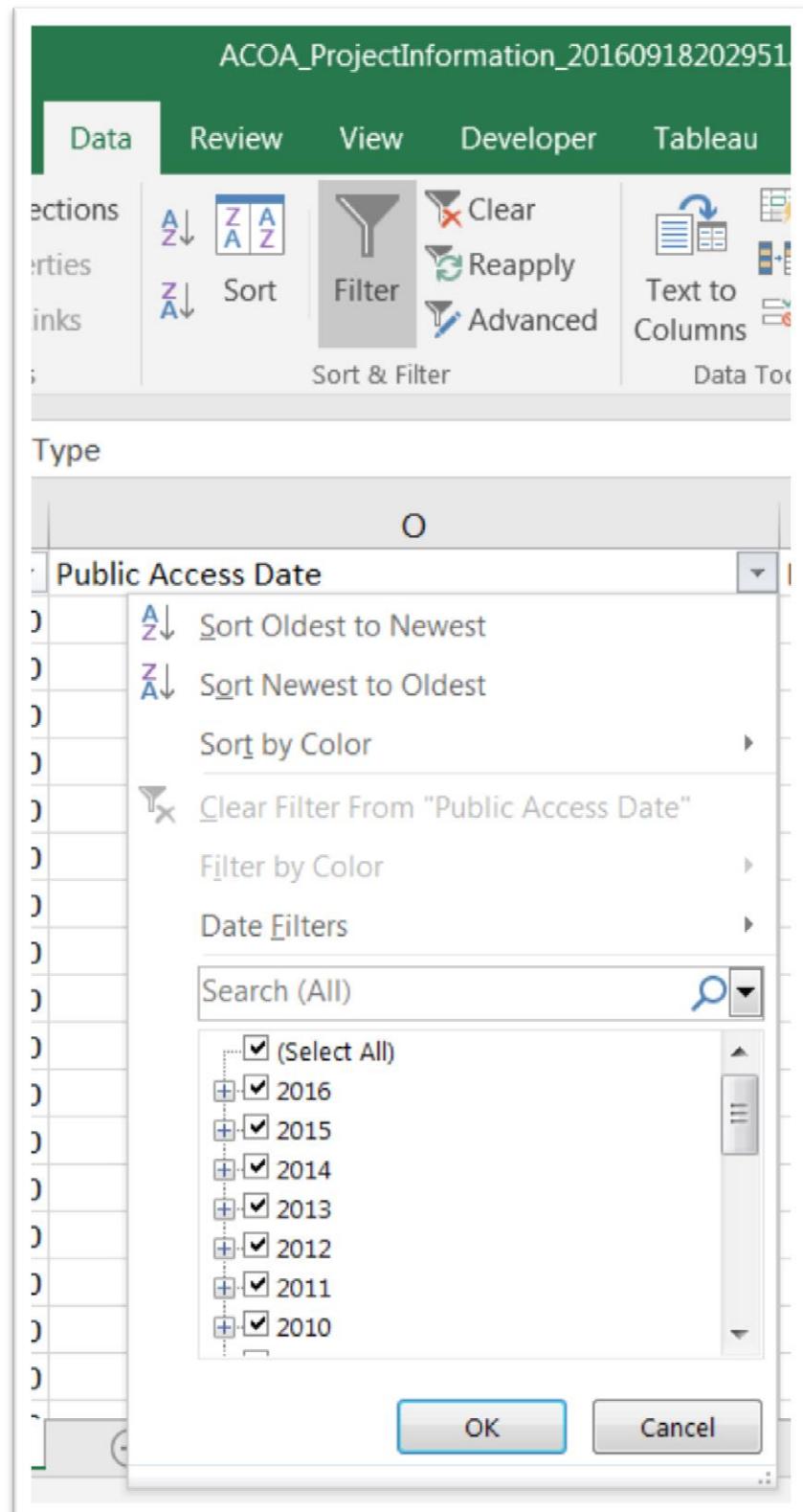
Activate the filter on column J, “Assistance Type.”

The screenshot shows an Excel spreadsheet titled "ACOA_ProjectInformation_20160918202951.csv - Excel". The Data tab is selected. A context menu is open over the "Assistance Type" column header in row J. The menu includes options like "Sort A to Z", "Sort Z to A", "Sort by Color", "Clear Filter From 'Assistance Type'", "Filter by Color", "Text Filters", and "Search". A list of filters is displayed, with the first item, "(Select All)", having a checked radio button. Other items listed include "Action Loan", "Conditionally Repayable Contribution", "Grant", "Interest Contribution/Buydown", "Loan Insurance/Guarantee", "Non-Repayable Contribution", and "Unconditionally Repayable Contribution". At the bottom of the dialog are "OK" and "Cancel" buttons.

Click the radio button to the left of “Select All”, and then select “Grant.”

The screenshot shows the same Excel spreadsheet after applying the filter. The "Assistance Type" column now only displays the value "Grant" for all rows. The rows are numbered from 156 down to 1317. The "Program Type" column also shows various entries such as "Business Development Program", "Community Infrastructure Program", and "Strategic Community Capacity".

And since we're only interested in the last five years, we'll need to filter a date field. Let's do so with column O, "Public Access Date."



As we did in the previous step, de-select all the dates by clicking “Select All”, and choosing the years 20122016.

The screenshot shows a Microsoft Excel spreadsheet titled "ACOA_ProjectInformation_20160918202951". The Data tab is selected in the ribbon. In the "Get & Transform" section, the "Query" dropdown is set to "Recent Sources". The "Connections" section shows a connection named "ACOA_ProjectInformation_20160918202951". The "Sort & Filter" section has a "Filter" button highlighted. The main table has two columns: "Total Project Cost" (Column N) and "Public Access Date" (Column O). The data is sorted by column O in descending order. The first few rows show values like 156, 569, 625, etc., with corresponding dates 12/07/2016, 13/03/2016, 15/01/2016, etc. The status bar at the bottom indicates "Ready 120 of 31927 records found".

	N	O
1	Total Project Cost	Public Access Date
156	12,995.00	12/07/2016
569	727,623.00	13/03/2016
625	2,925.00	15/01/2016
631	2,275.00	12/01/2016
644	2,501.00	03/01/2016
748	65,000.00	28/11/2015
816	65,000.00	10/11/2015
982	24,570.00	10/10/2015
987	47,496.00	09/10/2015
1068	165,250.00	22/09/2015
1222	21,063.00	14/08/2015
1247	43,923.00	07/08/2015
1250	29,375.00	07/08/2015
1265	39,959.00	02/08/2015
1276	180,000.00	28/07/2015
1277	12,790.00	28/07/2015
1317	16,011.00	14/07/2015

Now you can sort column O in descending order, which, as we can see in the screengrab, it already is.

If you're happy with this filtered dataset, you can copy and paste it into a new worksheet to keep working with it.

There are other ways to filter such as the use of colours, which can also be isolated using the option in your drop-down menu. Just use colours to select the values you want, and then isolate your selection from the drop-down menu.

The screenshot shows the Microsoft Excel ribbon with the 'Data' tab selected. A context menu is open over a column of numerical data. The menu includes options for sorting (Sort Oldest to Newest, Sort Newest to Oldest), color filtering (Sort by Color, Filter by Color), and date filtering (Date Filters). The 'Date Filters' option is selected, opening a dropdown menu where the year 2016 is checked. Other years from 2010 to 2015 are also listed with checkboxes. At the bottom of the filter menu are 'OK' and 'Cancel' buttons.

	Public Access Date	Estimated Commence
	12,995.00	
	727,623.00	
	2,925.00	
	2,275.00	
	2,501.00	
	65,000.00	
	65,000.00	
	24,570.00	
	47,496.00	
	165,250.00	
	21,063.00	
	43,923.00	
	29,375.00	
	39,959.00	
	180,000.00	
	12,790.00	
	16,011.00	

Sorting and filtering are excellent tools for an initial analysis of your dataset. In some instances, you need not do anything more in order to see a pattern worth exploring. For instance, the type of institutions receiving ACOA grants.

The beauty of Excel, though, is that it can do so much more, which we will learn in subsequent tutorials. However, this is a good start that involved absolutely no math.