SPREADSHEETS

(Data for this tutorial at www.peteraldhous.com/Data)

Spreadsheets are great tools for sorting, filtering and running calculations on tables of data. Journalists who know the basics can "interview" data to find stories and trends that others may miss. In this tutorial, we will work with LibreOffice Calc, which you can download from here.

Once you know your way around a spreadsheet, it's fairly easy to pick up the skills needed to work with many other tools for data analysis.

TIP! **Whenever you get some new data, save it under a new name! Then if you mess things up, you can always go back to the original.**

So, let's open the file **City Budget.xls**, available online here, by selecting **File>Open** and save it under a new name by selecting **File>Save As**.

Formatting, basic calculations and sorting

Spreadsheets use a simple co-ordinate system, with letters for columns and numbers for rows. You can make calculations on data using these coordinates.

This data was provided some years ago to a reporter working on a small city paper.

1. Formatting data

First, let's format the columns as currency. Select columns **B** and **C** by clicking on the letters while pressing **Ctrl** (Windows) or **command** (Mac); then select **Format > Cells** from the top Menu; when the dialog box pops up, select **Currency**, **USD $** and **0** decimal places.
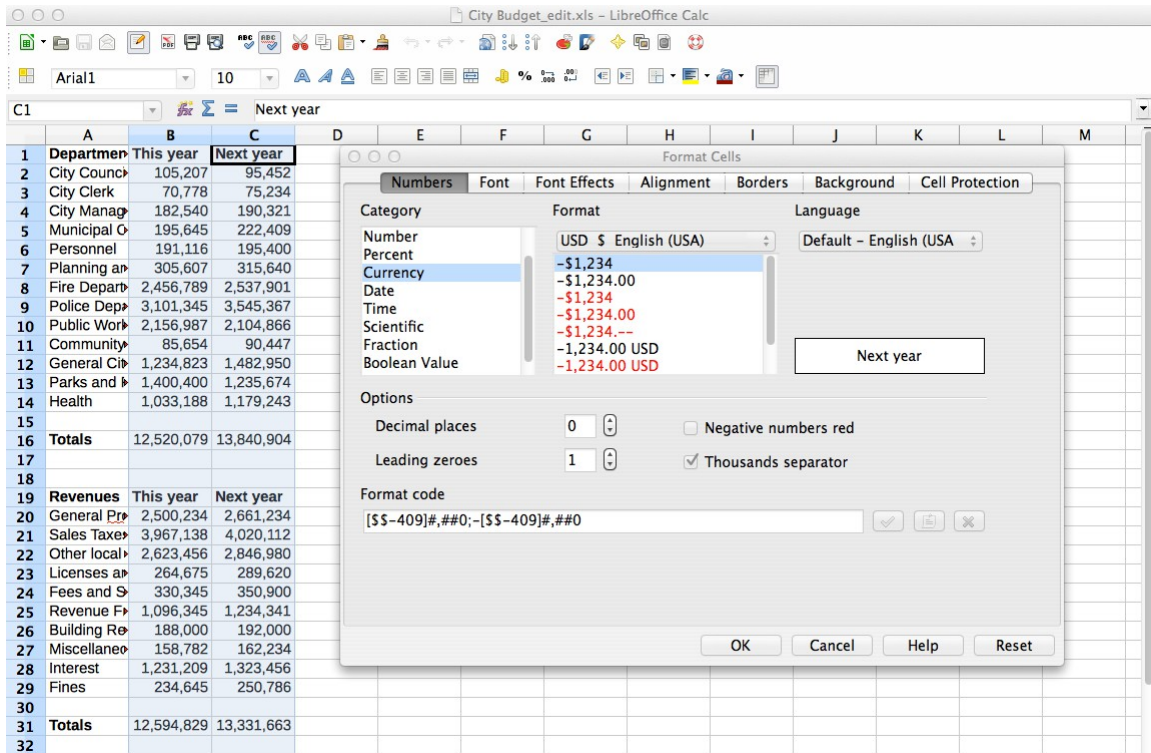
So, now we have everything in $, but we still can't see everything in column A, so let's format the column width.

Click the square at the top left, which selects all the data in the sheet. Then **Format>Column>Optimal Width** and **OK** at the next dialog box.

TIP! If ever you see a column containing **###,** it contains a number that is wider than the column. Use the same trick to fix the problem.

2. Performing basic calculations

Now let's calculate how the budget has changed from last year to this year. All calculations start by entering = into the cell. So, the formula for first cell is **=(C2-B2)**.

(At this stage, we may need to format the cells in the column so that they are also in $.)

Move the cursor to the bottom right hand corner of the cell until you see it change to a cross, then double click. Calc **autofills** the column, performing the same calculation for each row until it hits a space.

Calc can recognize other patterns in your data and autofill in the same way. To number a list of records, for instance, type 1, 2, 3, into the first three cells next to a column of data, select those three cells and autofill.

You can also **Copy (Ctrl-C or command-C)** the first cell and **Paste (Ctrl-V pr command-V)** all the way down the column. See how the "marching ants" form around the cell you've copied.

(This will give 0s where the cells are blank, and #VALUE! errors as Calc tries to perform a calculation on text. These can simply be deleted.)

You can lose your work if you later delete data that was used to calculate values that you are still working with. If this is a possibility, **Copy** the calculated data and paste it into a new column by selecting **Edit>Paste Special** and selecting **Numbers, Text** etc, as appropriate.

So let's do that here:

You may need to format our new column to currency once more. See how the cells now contain the numbers, not the formulas used to calculate them.

Then delete the old column D by selecting the cells, then **Edit>Delete Cells**.

Now let's calculate the percentage change in budget, rather than the absolute $ value.

The formula for the first cell is:
=((C2-B2)/B2) or **=(D2/B2)**.

So enter this, and copy down the column as before. Again, you may need to format to have the results displayed as a percentage, rather than a decimal.

3. Sorting data

Let's sort by % change. To do so, highlight the range of data you want to sort, then select **Data>Sort**. When the dialog box pops up, select **Sort key 1> % Change** and check **Descending** order.

TIP! When sorting, make sure Calc has recognized that your data has a header row if there is one present, otherwise that will be sorted as well. Look under the **Options** tab in the **Sort** dialog box and if necessary check **Range contains column labels**.

TIP! Make sure when sorting to select the full range of data you want to sort, and no more. Missing rows or columns from sorts, or including extraneous data are common ways to scramble your data. Use **Edit>Undo** if you make a mistake. In this case we want to sort separately on the Departments and Revenues tables, Taking care not to include the Totals.

Sorting by % change and Difference, for both Departments and Revenues can produce ideas for story angles. What might the reporter have pursued in this case?

TIP! **Always check any calculations that you haven't done yourself**

The mantra of journalism should be: if your mother says she loves you, check it out! We should be as skeptical of numbers as we are of human sources. So let's just check that the totals we were given add up correctly.

To check the Total for This Year, Departments, the formula is **=SUM(B2:B14)**.

Now **Copy** this cell, and **Paste** into columns C and D. What's our reporting plan now?

4. Performing anchored calculations

In each calculation we've performed so far, we've moved down or across the spreadsheet performing the equivalent calculation for that row or column. But sometimes that isn't what we want to do.

Let's calculate the increase in revenue from each source as a percentage of the total increase in revenue.

Use the formula:
=(D20/D31), and then a**utofill**.
The $ signs anchor the calculation. To anchor to a specific cell, you need a $ in front of both coordinates.

Importing data, filtering, subtotals and pivot tables

Now we're going to look at a larger dataset, used in reporting this story about the drug company Pfizer's payments to doctors, to show how you can use a spreadsheet to drill down to selections from the data, and to perform useful summary calculations.

Data is often provided as text (.txt) or comma separated values (.csv) files, so we'll first learn how to import a text file into Calc.

TIP! Databases should always be able to export data as .txt or .csv files, so bear this in mind whenever an organization says it can't provide data because it's in a special format.

1. Importing a text file

Find the file **Pfizer payments.txt**, available online here, and open it in a text editor See that individual entries are separated by Tabs and some of the entries are surrounded by double quote marks.

Open a new document in Calc by selecting **File>New>Spreadsheet.** Then select **Insert>Sheet From File**.

Browse for the file **Pfizer payments.txt** and open it. Fill in the dialog box as follows:
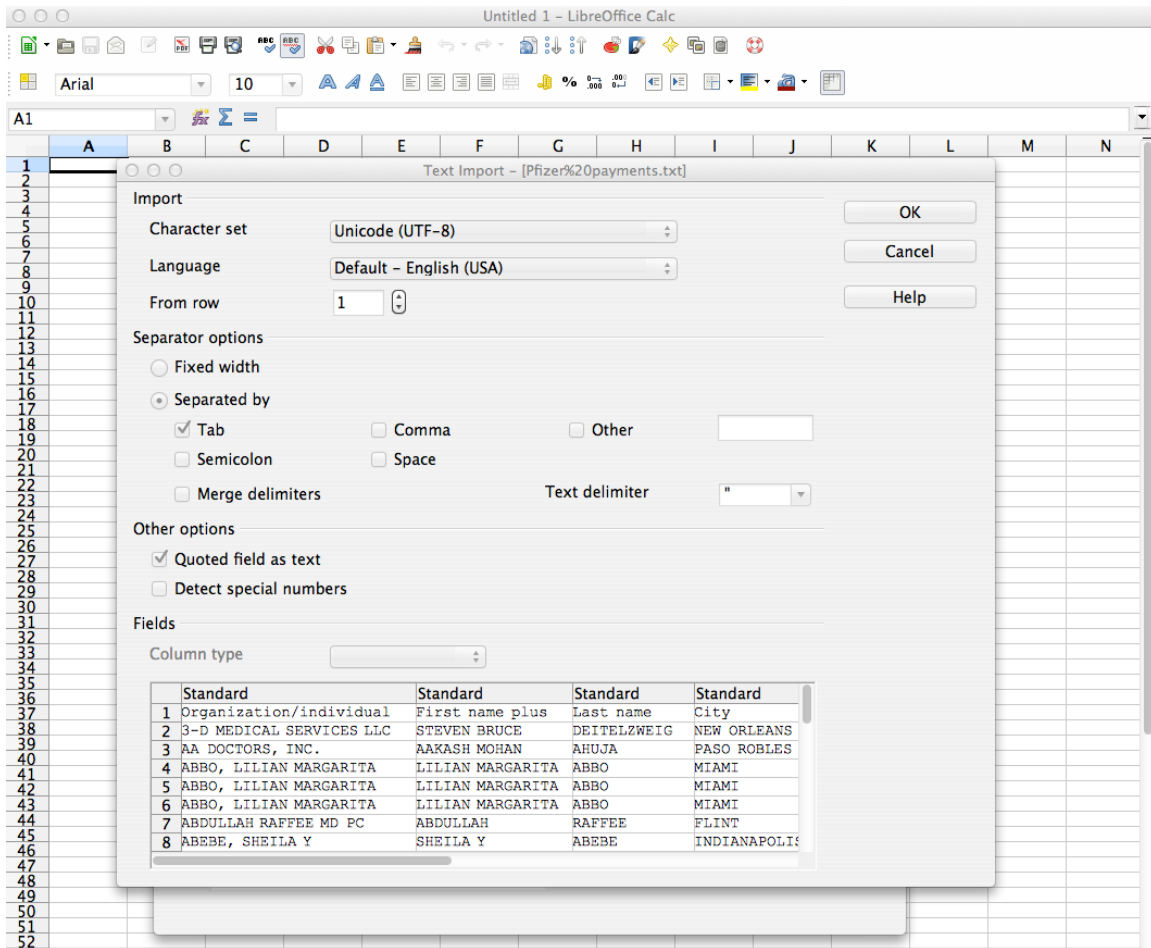
Click OK at the next dialog box, and the data should import.

Format the columns as desired, making the last three currency, and the header row bold.

<mark>TIP!</mark> To keep the header row present when you scroll through the spreadsheet, select the row beneath it and then select **Window>Freeze**.

Save this file as **Pfizer payments**, and then again under a new name, by selecting **File>Save As.** You can save as an ODF spreadsheet, or in Excel format.

Notice that we have a list of more than 10,000 payments, which in this form doesn't tell us very much.

2. Filtering data

So let's set up the spreadsheet so we can filter for what we're interested in. Let's say we want a list of all doctors in California who were paid $10,000 or more to run expert-led forums.

Select the entire spreadsheet be clicking the square at top left, then select **Data>Filter>Standard Filter**, and fill in the dialog box as follows:

Click **OK**, and you should have a list of 31 doctors.

For this filter, we connected the **Filter criteria** by **AND** operators, which ensured that rows were selected only if all the stated criteria were met. Filters obey Boolean logic; see what happens if you replace the first **AND** with **OR**.

TIP! If you need to keep, or do further calculations on a filtered subset of the data select the data, **Copy** it and **Paste** into a new worksheet. To rename the sheet for future reference, right click on the sheet tab and select **Rename Sheet**

Select **Data>Filter>Remove Filter** to return to the entire spreadsheet.

3. Calculating subtotals

Let's now find out how much money went to each state. Select **Data>Subtotals** and fill in the dialog box as follows:
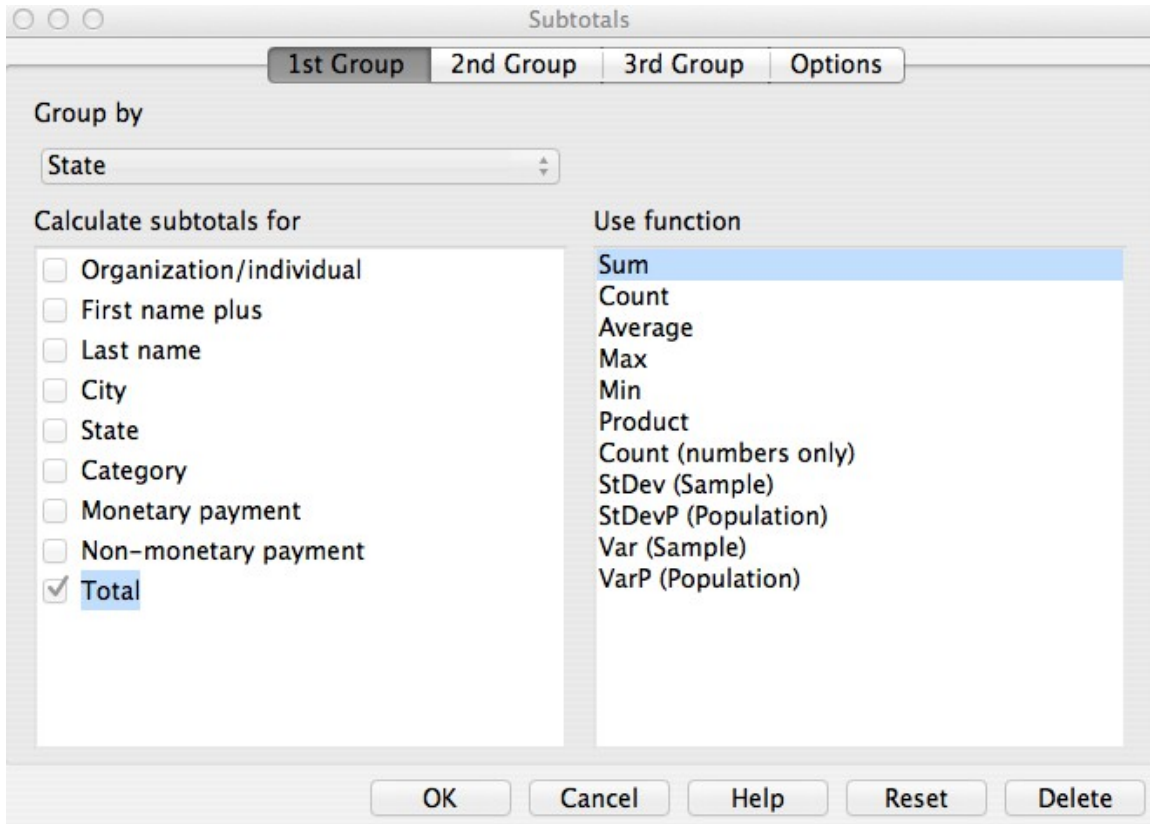
Click **OK**, and then click on the **2** at top left to hide the data and just leave the subtotals (**1** hides everything except the grand total.)

Notice that we added together the totals by state using the function **Sum**, but we could have used other functions to aggregate the data in other ways.

Select **Data>Subtotals>Delete** to clear the subtotals.

4. Making pivot tables

What if we want to know subtotals by state and by category of payment? We can do that in one step, using a pivot table.

Select **Data>Pivot Table>Create**, click **OK** at the initial dialog box, and then drag and drop in the dialog box as follows:
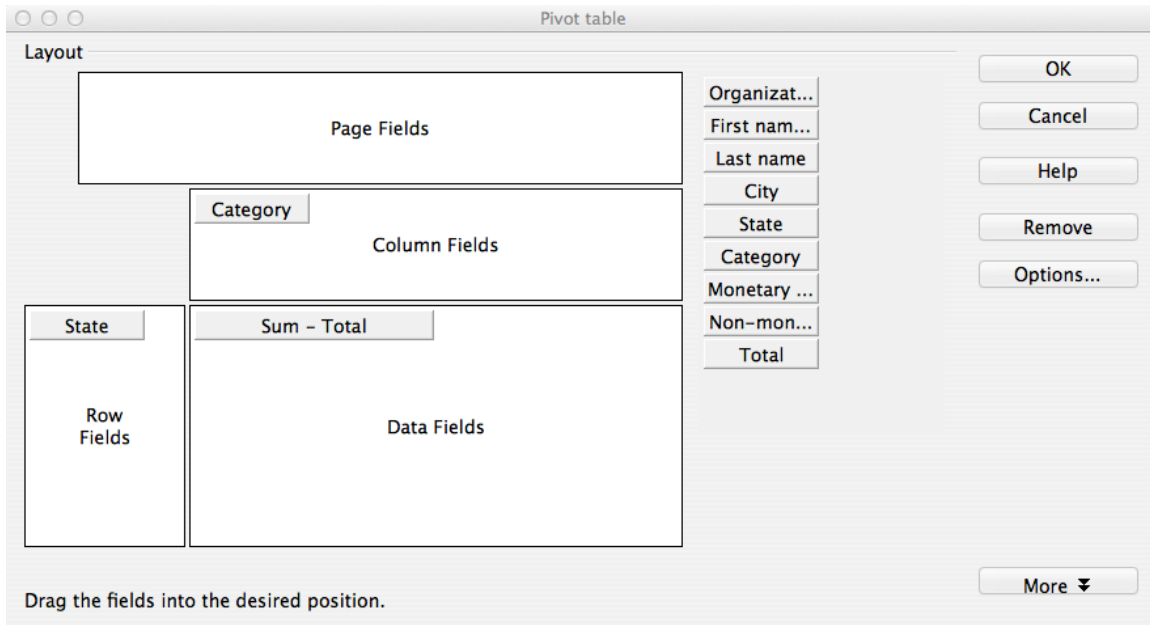
Notice that the calculation for Totals has defaulted to **Sum**, which is what we want, but we can select other aggregate functions by double clicking on **Sum – Total**.

Click **OK**, and the pivot table will be inserted into a new sheet. Format the column widths as necessary.

The final pivot table should look like this:

	A	B	C	D	E	F	G	H	I	J	K
1	Filter										
2											
3	Sum - Total	Category ▼									
4	State ▼	Business Related Tra▸	Educational Items	Expert-Led Forums	Investigator-Initiated P▸	Meals	Pfizer Sponsored Res▸	Pfizer Sponsored Res▸	Professional Advising	(empty)	Total Result
5	AK			$1,750							$1,750
6	AL	$29,123	$431	$129,850	$55,937	$10,217	$279,485	$66,096	$110,560		$681,699
7	AR	$3,783	$318	$72,825		$3,700	$77,670		$2,636		$160,932
8	AZ	$15,101	$173	$134,600	$76,102	$13,409	$225,615	$107,979	$68,872		$641,851
9	CA	$248,627	$947	$1,460,650	$173,563	$156,716	$1,125,624	$328,149	$1,243,531		$4,737,807
10	CO	$18,207	$144	$176,550		$12,433	$596,437		$111,467		$915,238
11	CT	$20,747	$79	$94,325		$9,133	$252,213	$90,000	$165,785		$632,282
12	DC	$31,098		$48,500		$7,774	$22,371		$140,798		$250,541
13	DE	$2,961	$135	$40,425		$1,344	$622		$8,500		$53,987
14	FL	$87,865	$495	$525,875	$41,187	$74,806	$306,600	$1,267,809	$259,410		$2,564,047
15	GA	$39,915		$252,800		$25,410	$180,632	$7,500	$112,388		$618,645
16	HI	$5,834		$21,200		$3,336	$6,247		$6,000		$42,617
17	IA	$7,025	$78	$56,225		$4,671	$32,207	$123,375	$20,125		$243,706
18	ID	$3,617		$20,725		$1,064			$12,250		$37,656
19	IL	$51,222	$1,465	$341,775	$90,000	$38,810	$393,395	$79,922	$260,236		$1,256,825
20	IN	$29,810	$291	$145,375		$15,018	$19,876	$3,333	$135,886		$349,589
21	KS	$16,763	$210	$89,025		$7,566	$7,442	$117,220	$68,979		$307,205
22	KY	$24,084	$228	$95,850	$4,530	$9,707	$211,385	$4,800	$86,354		$436,938
23	LA	$18,819	$463	$159,875		$14,575	$21,171		$47,018		$261,921
24	MA	$137,898		$154,875	$129,500	$21,498	$795,327		$525,673		$1,764,771
25	MD	$33,693	$99	$134,367	$90,288	$17,202	$391,499		$203,757		$870,905
26	ME	$1,757		$15,225		$1,249			$500		$18,731
27	MI	$50,940	$235	$299,925	$85,266	$26,190	$259,748	$184,966	$239,015		$1,146,285
28	MN	$35,747	$47	$185,036		$11,527	$155,872		$181,071		$569,300
29	MO	$34,272	$517	$331,950		$18,032	$225,788	$84,411	$278,616		$973,586
30	MS	$4,570	$345	$59,750		$6,111			$14,500		$85,276
31	MT	$2,623		$8,100		$485					$11,208
32	NC	$61,072	$31	$534,150		$38,175	$1,205,370	$167,818	$321,819	$0	$2,328,435
33	ND	$121		$15,825		$200					$16,146
34	NE	$5,616		$67,300		$2,772	$109,242	$1,071	$14,249		$200,250
35	NH	$10,404		$109,375		$3,245	$36,346		$12,999		$172,369
36	NJ	$21,924	$409	$269,625		$39,425	$137,154	$1,500	$130,805		$600,842
37	NM	$2,272		$24,700	$18,750	$1,683	$5,000	$1,500	$9,925		$63,830
38	NV	$3,632	$135	$46,000		$2,932			$20,325		$73,024
39	NY	$139,120	$506	$792,992	$94,138	$85,004	$322,747		$630,535		$2,065,042
40	OH	$80,428	$744	$362,625	$18,734	$42,998	$213,955		$299,966		$1,019,450
41	OK	$3,424	$242	$72,250		$4,067	$11,753		$19,787		$111,523
42	OR	$12,823		$55,000		$4,105	$186,490		$45,322		$303,740
43	PA	$63,983	$357	$301,300	$150,778	$44,034	$1,271,285	$15,944	$636,824		$2,484,505
44	PR	$5,770		$91,775		$11,460	$19,464		$1,925		$130,394
45	RI	$14,200		$25,250		$8,494	$104,420	$19,000	$38,840		$210,204
46	SC	$14,093		$101,275	$70,041	$11,798	$159,177		$65,107		$421,491
47	SD	$3,751	$79	$20,150		$1,773			$3,750		$29,503
48	TN	$35,033	$325	$353,200		$31,186	$136,401	$177,786	$115,294		$849,225
49	TX	$142,218	$1,955	$680,125	$302,641	$70,976	$1,013,070	$6,000	$585,211		$2,802,196
50	UT	$18,209		$91,650	$11,000	$7,856	$224,997		$27,180		$380,892
51	VA	$24,198	$85	$127,625	$22,000	$15,344	$85,646	$3,000	$90,094		$367,992
52	VT	$2,993		$5,000		$242			$21,653		$29,888
53	WA	$27,583	$184	$145,100		$15,470	$89,007	$37,732	$80,990		$396,066
54	WI	$31,547	$570	$189,500	$7,400	$13,405	$209,075		$58,625		$510,122
55	WV	$12,786		$69,675		$9,483	$9,749		$26,679		$128,372
56	WY	$2,028		$34,325		$1,309			$2,300		$39,962
57	Total Result	$1,695,329	$12,322	$9,643,220	$1,441,855	$979,419	$11,137,574	$2,896,911	$7,564,131	$0	$35,370,761
58											