Data Best Practices

Why best practices?

- Be able to explain what you did
- Lets someone else (or you) reproduce/bulletproof your work
- For longer projects, you might not actually remember all of your steps
- Your process might be useful to you (or someone else) again someday

Step 1

Save (at least one) clean copy of your data. Never touch it again. Whatever you do, don’t work in your only data copy.

Better: Save copies in multiple locations. E.g. on your hard drive and in Google Docs.

Step 2

Read the documentation that goes with the data file.

This could be a data dictionary: https://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/LimitedDataSets/Downloads/InpatientVersionJ2011.pdf

Or a website:

http://datadictionary.cityofchicago.org/database_info.php?database_id=61

… Or something else entirely. But you will want to know what each field means and how it's coded.

Step 3

Create a text document.

Name it something useful and put it where you can find it.

Put your name, the date and a short description of what the project is at the top.

If you don’t have a complete analysis plan from the outset, that’s okay. You can circle back to this description at the end and make it more accurate and descriptive.

Example

Olga Pierce

5/24/2016

Description: These are the steps I used to calculate the NYC response times for various types of 3-1-1 calls.

Data: I used the city's 3-1-1 call database, as of 5/23/2016. It can be found here: www.mydatalink.com OR C://Users/olga/myfile.csv. The documentation is here: C://Users/olga/mydatadictionary.txt.

Step 4

Count the number of rows/records in your data. Make sure this matches the number (hopefully) described in the documentation. If so, then enter the record count in your text document.

Example:

Record count: This raw dataset contains 12,357 records.

Step 5

Track your work.

Example:

The first thing I wanted to know was who the biggest campaign contributors were. So I did:

Clicked on column B to select it.

Data -> Sort sheet by column B, Z to A.

Step 6

Where applicable include the output as well. (It's often OK to truncate).

Example:

This is the top rows I got after the sort.

Contributor Name	Employer	Occupation	City	State	Zip	Receipt Date	Amount
ABBASI, SOHAIB	INFORMATICA	CHAIRMAN	ATHERTON	CA	940275431	9/10/2015	$2,700
AARONSON, PAULA	RETIRED	ENGINEER	AUSTIN	TX	787031026	5/26/2015	$2,700
AARON, STEWART	ARNOLD & PORTER LLP	ATTORNEY	LARCHMONT	NY	105381728	10/18/2015	$2,000
ABBEY, MICHELLE	TMP WORLDWIDE	CEO	NEW YORK	NY	100171739	6/16/2015	$250
AAB, LISA	ZF TRW	CAD DESIGNER	FLINT	MI	485322321	6/13/2015	$143
CONARE, PURNA RODMAN	SELF-EMPLOYED	NONPROFIT CONSULTING	HARMONY	RI	28290387	2/5/2016	$100
ABBITT, ALICE	DHL	PROJECT MANAGER	ANTHEM	AZ	850863638	2/10/2016	$100
ABBEY-MAGEE, JAMES	MACKAY SHIELDS LLC	DIRECTOR	NEW YORK	NY	100041021	3/31/2016	$50
?LEARY, JEANNETTE	JOHNSON & JOHNSON	RETIRED BUYER AND SENIOR BU	DUNEDIN	FL	346987317	8/13/2015	$50
AARONSON, LAURA	RETIRED	R.N.	BETHESDA	MD	208161839	10/16/2015	$50

When including results is impractical, note the filename and spreadsheet tab. It can also be useful to include row counts. The goal is for someone else to make sure they can come up with the same answer as you.

More Step 6

If you used a function, copy and paste the exact text of the function you used.

Example:

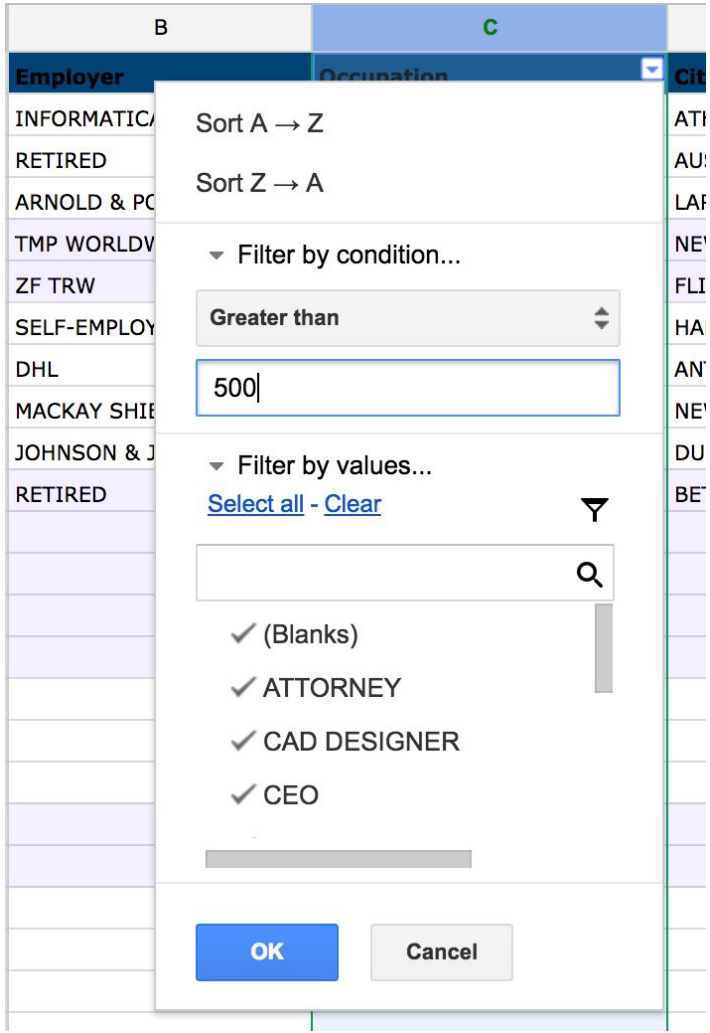
I used this function to calculate column J, called Big_Donations:

```
=IF(H2>1000,1,0)
```

If can also be useful to note characteristics of calculated fields. Here I could add that there were 926 1s and 512 0s.

Even more Step 6

If you are using dropdown menus, it can also be useful to take a screenshot.



Review: Why we are bothering with this

- You will know what you did
- You can easily do it again
- Someone can check your work
- If necessary, you can share what you did with experts
- It will make it easy to share what you did with the subject of a story
- It will make it easy to share what you did with readers
- If your work comes under question, you can show that you did due diligence