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: Keep it clean

Download and save an original copy of your data – direct from the source, with the name as it is, and no alterations.

Whenever you make changes, make and save a COPY of the file. I like to add a suffix to the filename that describes the changes. (-deduped.csv, -sorted.csv, -cleaned.csv)

If my project involves a lot of changes to the data, I'll add a date to the suffix.

Step 2: Read the data dictionary

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://irsa.ipac.caltech.edu/applications/DDGEN/Doc/dd_tbl.html#examples

… Or something else entirely. (For example, maybe it's in an email to a government worker.) But you will want to know what each field means and how it's coded. Keep track of this.

Here's our PPP data dictionary.

Step 3: Create a data diary

Create a text document, if you’re working in Sheets or Excel, a Google doc works well.

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.

Step 4: Count!

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 4,7066 records. (exclude header!)

Step 5: Keep track of changes

Track your work.

Example:

The first thing I wanted to know was when the last loan was approved. So, I clicked on column O to select it.

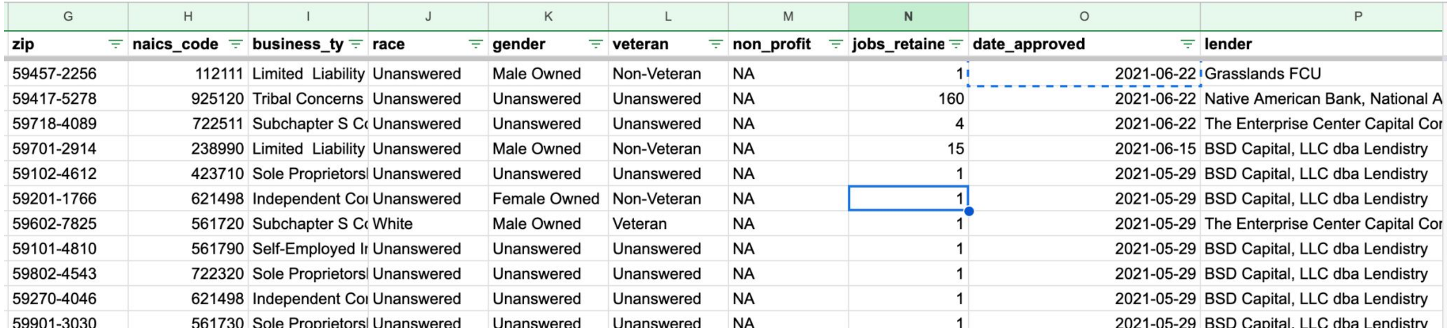
Data -> Sort sheet by column O, Z to A.

Step 6: Show your output

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

Example:

This is the top rows I got after the sort.



G	H	I	J	K	L	M	N	O	P
zip	naics_code	business_ty	race	gender	veteran	non_profit	jobs_retaine	date_approved	lender
59457-2256	112111	Limited Liability	Unanswered	Male Owned	Non-Veteran	NA	1	2021-06-22	Grasslands FCU
59417-5278	925120	Tribal Concerns	Unanswered	Unanswered	Unanswered	NA	160	2021-06-22	Native American Bank, National A
59718-4089	722511	Subchapter S Cc	Unanswered	Unanswered	Unanswered	NA	4	2021-06-22	The Enterprise Center Capital Cor
59701-2914	238990	Limited Liability	Unanswered	Male Owned	Non-Veteran	NA	15	2021-06-15	BSD Capital, LLC dba Lendistry
59102-4612	423710	Sole Proprietorsl	Unanswered	Unanswered	Unanswered	NA	1	2021-05-29	BSD Capital, LLC dba Lendistry
59201-1766	621498	Independent Coı	Unanswered	Female Owned	Non-Veteran	NA	1	2021-05-29	BSD Capital, LLC dba Lendistry
59602-7825	561720	Subchapter S Cc	White	Male Owned	Veteran	NA	1	2021-05-29	The Enterprise Center Capital Cor
59101-4810	561790	Self-Employed Iı	Unanswered	Unanswered	Unanswered	NA	1	2021-05-29	BSD Capital, LLC dba Lendistry
59802-4543	722320	Sole Proprietorsl	Unanswered	Unanswered	Unanswered	NA	1	2021-05-29	BSD Capital, LLC dba Lendistry
59270-4046	621498	Independent Coı	Unanswered	Unanswered	Unanswered	NA	1	2021-05-29	BSD Capital, LLC dba Lendistry
59901-3030	561730	Sole Proprietorsl	Unanswered	Unanswered	Unanswered	NA	1	2021-05-29	BSD Capital, LLC dba Lendistry

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 how many loans took place in a rural area:

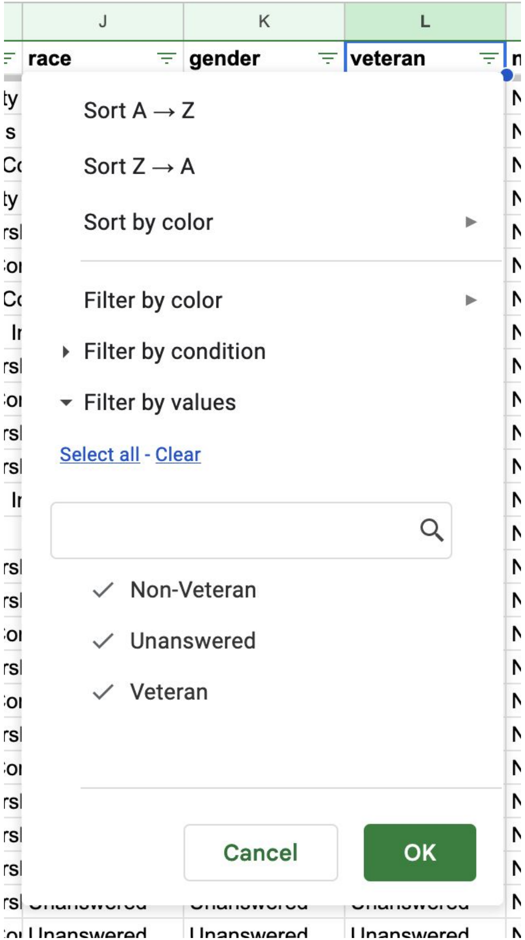
```
=countif(AH:AH, "R")
```

If can also be useful to note characteristics of fields.

(Column O, “date_approved,” ranges from 2020-04-03 to 2021-06-22.)

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