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.

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 2,124 records. (exclude header!)

Step 5: Keep track of changes

Track your work.

Example:

The first thing I wanted to know was when the last case was filed. So, I clicked on column D to select it.

Data -> Sort sheet by column D, 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.

	CaseNumber	Court	FileDate	CaseStatus	CaseStatusD	CaseType	CaseSubTyp	Style	IsActive	IsPublic	Parties	Attorneys	ShowWarrar
2	20D04-2010-SC-003065	Elkhart Superior Court 4	10/28/2020	Decided	10/28/20	SC - Small Claims		Goshen Commu	FALSE	TRUE	Goshen Commu	Davis	FALSE
3	70C01-2010-SC-000251	Rush Circuit Court	10/27/2020	Pending	10/27/20	SC - Small Claims		Rushville Consol	TRUE	TRUE	Rushville Consolidated High Scho		FALSE
4	50D02-2010-SC-000801	Marshall Superior Court 2	10/27/2020	Pending	10/27/20	SC - Small Claims		Plymouth Comm	TRUE	TRUE	Plymouth Comm	Houin	FALSE
5	50D02-2010-SC-000805	Marshall Superior Court 2	10/27/2020	Pending	10/27/20	SC - Small Claims		Plymouth Comm	TRUE	TRUE	Plymouth Comm	Houin	FALSE
6	50D02-2010-SC-000807	Marshall Superior Court 2	10/27/2020	Pending	10/27/20	SC - Small Claims		Plymouth Comm	TRUE	TRUE	Plymouth Comm	Houin	FALSE
7	50D02-2010-SC-000810	Marshall Superior Court 2	10/27/2020	Pending	10/27/20	SC - Small Claims		Plymouth Comm	TRUE	TRUE	Plymouth Comm	Houin	FALSE
8	50D02-2010-SC-000811	Marshall Superior Court 2	10/27/2020	Pending	10/27/20	SC - Small Claims		Plymouth Comm	TRUE	TRUE	Plymouth Comm	Houin	FALSE
9	50D02-2010-SC-000800	Marshall Superior Court 2	10/27/2020	Pending	10/27/20	SC - Small Claims		Plymouth Comm	TRUE	TRUE	Plymouth Comm	Houin	FALSE
10	50D02-2010-SC-000804	Marshall Superior Court 2	10/27/2020	Pending	10/27/20	SC - Small Claims		Plymouth Comm	TRUE	TRUE	Plymouth Comm	Houin	FALSE
11	50D02-2010-SC-000806	Marshall Superior Court 2	10/27/2020	Pending	10/27/20	SC - Small Claims		Plymouth Comm	TRUE	TRUE	Plymouth Comm	Houin	FALSE

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 cases were pending:

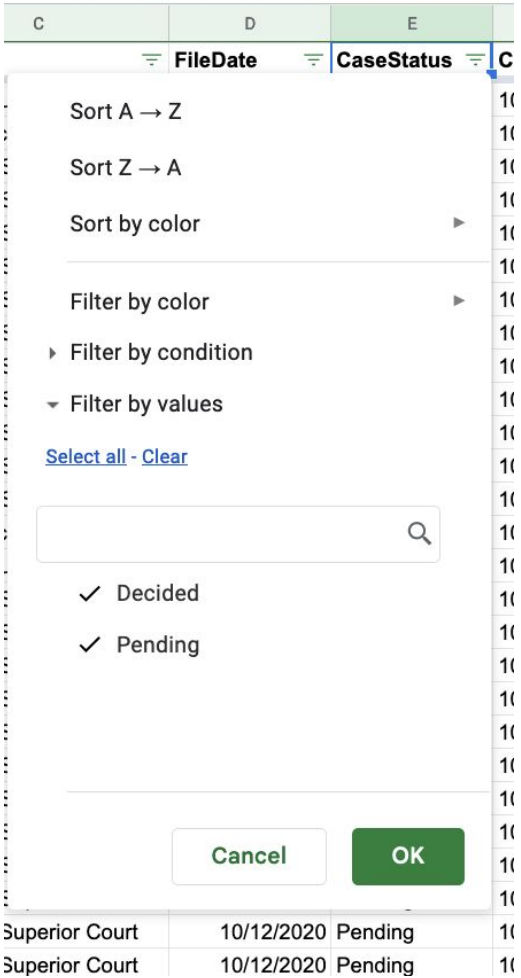
```
=countif(E:E, "Pending")
```

If can also be useful to note characteristics of calculated fields.

(Column D, “FileDate,” ranges from 1/2/2020 to 10/28/2020.)

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