How to avoid rookie mistakes:

- Think through your data request before filing – make sure you are requesting a broad enough set of fields and records. Is it a long enough time frame? Am I getting all the potential fields? Will the scope of my request be too limiting in terms of the types of questions I can ask or the analysis I can do?
- Always request the record layout, data dictionary, codebook – any documentation that will help explain the data.
- Make a backup copy of your data first. Then save versions as you go (or use a versioning control system like Github).
- Figure out a file naming system (for all those backups and versions) to make it easier to remember which is the most recent version. And then stick to that system!
- Keep a "trail" of your data analysis and cleaning
- Expect that there will be flaws, errors, missing information, etc. in your data.
- Use sort, filter and Pivot Tables (in Excel) and filtering and group by queries (in Access)to look for anomalies/outliers in your data.
- If you ever get a dataset with exactly 65,536 rows (or 1,048,576), don't believe it. Those are the maximum limits for a the various versions of Excel– it's possible data got cut off.
- Don't assume or guess what anything means.
- Before analyzing, write down some questions and/or working theories on what you're looking for. Wandering aimlessly in the data usually results in nothing or garbage.
- If you're doing something that seems unusually difficult, perhaps there is an easier way.
- When importing into Excel, beware of fields (like zip codes) that have leading zeros and make sure you don't lose those on the import.
- In Excel, keep your formulas as a trail for what you did, but don't delete the original fields that you based the formula on (you'll lose the answers!)
- Try to avoid slicing apart your data into separate worksheets or tables. It's okay to do this to some extent and under certain circumstances, but sometimes this kind of slicing makes more work for yourself (i.e. you have to run the same formulas, pivot tables or queries multiple times) or limits what you can do with the data because it's all separated.
- When cleaning or standardizing data, make the changes in a new field (a copy of the existing one) so that you preserve the original information.
- Don't delete fields. You'll regret it later. In Excel you can "hide" fields to get them out of your way. In Access you can create a query on your table to just view/work with the fields you want.
- Find and replace can be dangerous. Learn other data-cleaning techniques like regular expressions and OpenRefine.
- When sorting in Excel, make sure you're sorting all fields together, not just one. A paper in Florida had to write a correction because an Excel sorting error led to them publishing a story saying the Lhasa Apso was the most popular dog breed in their county.
- In Excel, don't use headers that have multiple rows. It causes problems for sorting and filtering.

- In Excel, don't insert blank rows in the middle of your data
- Don't put extraneous information in the same column as your data, or put two pieces of data in the same column (i.e. putting a note next to a value in some/all of the records)
- Don't use color-shading in Excel to categorize your data. Instead, create a new field and put a label there.
- Repeat your analysis, start to finish, at least one time to ensure you get the same results
- Apply the "smell test" or your "crap detector" to your findings. If something seems too good to be true, it probably is.
- Always check your results against other sources, human and documents or other data.
- Recruit a colleague and talk through your analysis with them. They help bring some much-needed skepticism and will tell you if something doesn't sound right.

Created by @MaryJoWebster
August 2015