Basic steps for every data analysis

1) Make sure you have an extra copy of your data somewhere in the original form it came in. Store the working copy of your data in OneDrive or some place where it is automatically backed up (Do NOT store in your "documents" folder on your C: drive)

2) Get to know your data
    - Review record layout and other documentation provided
    - What does each row represent?
    - What's in each column?
    - Are there any columns you don't understand?
    - Are there any codes you can't decipher?
    - Try pulling out a single record to try to understand what it's telling you

3) If you have any unanswered questions from step 1, contact the agency that you got the data from. Don't ever ASSUME anything.

4) Evaluate whether you need to do any data cleaning.
    - Use filters or Pivot Tables to look at all the values in each column – anything strange?
    - Are there any numeric values or dates that are out of bounds? (i.e. a date that is in the future or one that is far earlier than you'd expect to find.) Are there inconsistencies in values? (I.e. Minneapolis spelled as "Mpls" and "Minneapolis" and "Mineapolis"). Keep in mind that periods (or lack of) can make inconsistencies such as "St. Paul" and "St Paul".
    - Remember that the only columns/fields you'll need to clean will be ones that you plan to use in your analysis
    - As part of this, also get a sense of whether there are any columns you can't use because of too many missing values. Less than 10% empty is generally safe to proceed.

5) Make a list of the questions you want to ask the data, as if you're preparing for an interview with a human source. When you're ready to start analysis, this gives you a road map to work from.

6) Look at whether you need to categorize your data into any buckets, then start adding those new field(s).

    For example, let's say one question you have is to see how many incidents occurred in the metro versus outside the metro. Is there a column that identifies in metro and

out of metro? If not, is there a column – such as county name – that you could use to create these buckets?

Another example – perhaps you have data about people and it gives their age. Do you want your analysis to be based on age groups instead?

Do you need to merge other data? For example, maybe you want to calculate a per capita rate – do you need population data?

7) Start a journal where you document the steps you make on your data. Be sure to include the source of your data (agency name, contact person, contact phone/email, when you got it, what the universe of data is, etc)

8) Some other best practices as you go along:

   Label all columns. Try to avoid spaces and symbols (you can use an underscore). Good labels might look like: “School_Name” or “SchoolName” or “Pct_chg_2017_to_2018”

   Name all the worksheets in your Excel file with something you’ll understand. If necessary, add some space at the top of your worksheet to write a note explaining what this data is and where you got it from.

   Don’t delete data, even if it’s something you think you won’t need (You might discover later that you do need it!). Use “hide columns” to get things out of your way.

@MaryJo Webster
mjwebster71@gmail.com
August 2018