**Data Analysis Roadmap and Best Practices**

@MaryJoWebster  
Created September 2019

**Roadmap:**

1. Save a **backup copy** of your data and any documentation that came with it.
2. Start a **data journal.** If you made a public records request, copy and paste into your journal a copy of your request and any correspondence with the agency that followed. Write down a thorough description of the data that should include, but not be limited to:
   * Agency where you got the data including name and contact information of the people you dealt with, if applicable
   * how you obtained it (public records request; online data portal, etc). If it came from an online source, include a link to the site.
   * official name of the dataset, if applicable
   * the universe of the data it includes. This might include geography such as whether it covers an entire state or just a single city. This might include what kinds of incidents. Be sure to note if there’s something that’s not here that you might expect to find. (inspections of all schools in your state? All fatal car accidents, but not injury accidents in the whole U.S.?)
   * the time frame the data represents (all incidents from 2018? A snapshot as of Dec. 31, 2018? All inspections from the past 10 years?). If you got the data from an online data portal, be sure to note the date it was last updated on the site.
   * Anything else that is important to note. Are there limitations on how you can use this data? Did the agency warn you about the consistency or validity of anything in the data? Were there any changes in how this data is collected or reported during the time period your data represents?
   * If you have a record layout and/or codebook, make a note about where those items are stored on your computer.
3. If you haven’t done this already, **talk to someone at the agency to ask some questions** and then write the answers in your journal. Your goal is to understand how the system works.
   1. How is this data collected? (via a paper form? An electronic submission?) When is it collected? How often is it updated?
   2. Does anybody review it for accuracy?
   3. How is it used by the agency? Are there any reports generated on a regular basis? (if so, get a copy of one)
   4. You will likely have other questions that are specific to the topic. For example, when requesting an agency’s expense records, you’d want to know whether each record is a single expense (i.e. a hotel stay) or whether it’s a compilation of expenses (i.e. all expenses related to a trip). If you were requesting police dispatch data, you’d want to know what ends up in there – does a record get created each time an officer calls in to report what he or she is doing? (I worked with a dispatch dataset once that included including each time the officer took their car to the car wash)
4. **Get to know your data.** Open it up in whatever software you plan to use and then write down the answers to these questions in your data journal.

* Review record layout and other documentation provided. If you didn’t get a record layout (i.e. they said they didn’t have one), try making your own. It’s a good exercise for understanding what each column represents.
* What does each row represent? This can sometimes be not what you think. I worked with a fire calls dataset once that had one record for each vehicle dispatched, so you frequently had multiple records for each incident.
* What’s in each column? Are there any columns you don’t understand and that aren’t explained in the documentation?
* Are there any codes you can’t decipher or that don’t have descriptions in the documentation?
* Try pulling out a single record to try to understand what it’s telling you. For example, let’s say you’re looking at restaurant inspection data. Find all the records that pertain to a single inspection and try to figure out what it shows. Then expand out and find all inspections for that same restaurant and see what you can discern over time. This would help you understand, for example, if it’s possible to see whether violations were later corrected. If it’s not clear what you’re seeing, ask the agency to walk you through it using this one example. Another approach might be to ask for the paper documentation related to that inspection. Sometimes the paper documents are more thorough and include narrative that’s not in the database.
* By exploring the fields that are included in the dataset and understanding what each row/record represents you can get a better idea of the breadth of this source’s knowledge. For example, if the dataset includes only NFL games played in outdoor stadiums, then you know that its knowledge is limited to only outdoor games (you can’t ask any questions about all games). Or if a dataset of foster parent license revocations does NOT include a field indicating why the family lost its foster care license, then you know you can’t ask any questions about why the licenses were revoked.

1. **Talk to other people** who are already familiar with this data and/or topic, such as academics or other experts. Tell them what you’re hoping to figure out from this data and get their opinions about whether the data can help you, and any suggestions they might have for how to do that or what questions to ask or what to be wary of. Write summaries of these conversations in your data journal.
2. 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.

1. Make a **list of the questions** you want to ask the data, as if you’re preparing to interview a human source. Start by just doing a brain dump and writing down everything you can think of. Then rearrange the list starting with simple questions that you are certain you know how to answer, up to the harder ones.

To spur your thought process, here are some common questions we often ask:

How often has this happened in the past?

How does this compare to....?

What's the worst...?

What's the best....?

Where is there the most...?

Where is there the least...?

Is this higher or lower than last year?

How has this changed since...?

1. In addition to the questions, try to **write some statements** that you hope to make in your story. For example, for a project I did on how police handle sex assault investigations, we knew upfront that we wanted to be able to say something like “Acquaintance rapes are less likely to be solved than stranger rapes.” So, we knew that our data needed to include a column about the relationship between the assailant and victim and that we needed to have questions in our list that would address this.
2. Look at those **questions you aren’t sure you know how to answer.** Is it possible that you need to merge in some other data? If you think you need some other data, now is the time to start hunting that down.
   * For example, let’s say your question requires an answer expressed as a per capita rate, but your data doesn’t have population counts to use for that.
   * Or let’s say you are looking at school test scores and you want to know how the high poverty schools fared compared to each other, but the test score data doesn’t identify the poverty rate (i.e. free/reduced lunch percentage).
   * You might want to marry two datasets together. For example, I worked on a story about the Minnesota Twins baseball team and why they were struggling to hit homeruns. The baseball beat reporter suspected it was the weather. So, we married game-level hitting data with weather recordings from a weather station near the stadium that matched the date and start time of each game.
3. Look at whether you need to **categorize your data into buckets**, then start adding those new field(s). This is something that is necessary more often than not.

For example, let’s say one question you have is to see how many incidents occurred in the metro area 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? If so, then you would need to create a new column that identifies which age group each person falls into.

1. **Does your data need to be re-arranged?** Sometimes if you’re struggling to figure out how to answer your questions, the problem might be that the data isn’t in the right format. Some examples:
   * Let’s say you have enrollment data at the local university, broken down into sub groups (international students, student groups by race or gender, etc), but each year of data is stored in a separate worksheet in Excel. If you want to look at trends over time, you’d need to have all the records in one worksheet, with a column identifying what year the data came from.
   * Another common issue is that your data might be in a “wide” format and you need it in a “long” format, or vice versa.

Wide data has more than one value in each row, like this:

A picture containing crossword puzzle

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Here’s that same data converted to long format. Notice that you’re seeing all the values from the 1st row of the wide data, now displayed as separate rows.

* + Another example is the fire calls data I mentioned above. With each row representing a vehicle dispatched to the scene, it was hard to do any counts or analysis based on each incident. I needed to boil my data down to one row for each incident. That isn’t always a straightforward thing. In the case of the fire trucks, which record do I keep? Sometimes a dataset will have a column that identifies a “primary” record in a situation like this, but I don’t see that too often. Or you might need to summarize your data and make a new table that has the basic information about each incident (date, time, location) and then a count column (i.e. the number of trucks dispatched).

1. If you’re still having trouble figuring out how to answer a question, consider the **possibility that your data might be missing something** that you expected it would have, or that you don’t fully understand how the data is set up. Or perhaps you didn’t ask for multiple years of data, but now you realize that you want to look at trends over time. Might be worth going back to the agency and asking more specific questions or asking for more data. Depending on your relationship with the agency, you might want to flat out tell them what question you’re trying to answer and see if they have suggestions.
2. Keep in mind that once you start analyzing your data, **more questions will jump to mind.** You will also end up asking far more questions than you truly need – that’s ok. You want to fully understand your data so that when you choose which analysis results to include in your story, you know that it’s the most important findings.
3. When you’re ready to start doing your analysis, go through each of your questions – starting with the ones you think will be the easiest. Keep in mind that **your question often gives you hints** about what columns you will need in your analysis.

For example, let’s say your question is: Are elementary schools more racially diverse than high schools in my state? That tells you that you need data with one row for each school. And that there needs to be a column identifying each school by type (elementary, middle, high school, etc), with other columns indicating the number of students of color, the number of white students and the total students. The universe of your data needs to encompass the whole state. Then you would need to calculate a combined diversity rate (percentage of students who are people of color) for each type of school. This would be done via a Pivot Table or a group-by query, depending on what tool you are using, because you are summarizing/grouping your data into buckets.

Another example: How many schools in my state are “majority minority”? Note that here we are just asking for a single number. We aren’t calculating new totals or averages or rates for groups. So, this might be as simple as a filtering your data to only those schools that have more than 50% students of color. Keep in mind that some questions don’t need the fancy tools.

**Best practices:**

Follow “tidy” data practices. Neat rectangle of data, with one row of headers across the top. No empty rows, no empty columns in the middle of your data. Each row is one observation. Each column contains one piece of information about that observation. Don’t bunch multiple pieces of information in a single column (for example, you probably want a last name and a first name in their own columns) or put extraneous information in the same column as a value (for example putting this in a date field: “approximately 1/1/2008”).

Name all your columns with clear names. If you add new columns, be sure to add them to your record layout, along with a description. Get in the habit of using database-friendly names (even if you’re in Excel), such as using underscores (my\_field\_name) or Camel case (MyFieldName) and avoiding spaces and symbols. When you’re ready to move out of Excel and to something else, you’ll be glad you had standardized field names.

Make sure your file names and worksheet names are clear. If you’re in an Excel workbook, don’t have “Sheet1”, “Sheet2” etc. Eventually you’ll just be confused. Name each one.

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

Never assume you know a field means in a database or what a code or value represents. Check with the agency where you got the data.

Include as many details in your data journal as possible. As you go through the analysis, document each step. It’s especially important to note instances where you filter out records (did you possibly leave out something that you shouldn’t have?) and instances where you make new tables. These are common spots where you could make a wrong turn.

Expect that there will be flaws, errors, missing information, etc. in your data. Find the problems before they find you!

If you’re doing something that seems unusually difficult, perhaps there is an easier way.

When cleaning or standardizing data, make the changes in a new field/column (a copy of the existing one) so that you preserve the original information.

Find and replace can be dangerous. Learn other data-cleaning techniques like regular expressions and OpenRefine.

If your analysis includes decisions you made (i.e. which records to include/exclude), start writing a rough draft of a methodology right away. Make sure you can defend your decisions to the public and writing out a methodology helps you figure out if you can do that. Why did you exclude those records? Why did you calculate it as a rate and not a raw number? Why did you only include hospitals of a particular size?

Repeat your analysis, start to finish, at least one time to ensure you get the same results.

Make sure your results pass the “common sense” test. If it seems too good to be true, it might be wrong.

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.