Practical Advice for Analysis of Large, Complex Datasets

by Patrick Riley

"https://www.unofficialgoogledatascience.com/2016/10/practical-advice-for-analysis-of-large.html"

Advice here is organized into three general areas:

- *Technical*: Ideas and techniques for how to manipulate and examine your data
- *Process*: Recommendations on how you approach your data, what questions to ask, what things to check
- *Social*: How to work with others and communicate about your data and insights

Technical

Look at your distributions
Although we typically use summary metrics (means, standard deviations, etc.) to communicate about distribution, you should usually be looking at a much richer representation of the distribution. Histograms and Q-Q plots will allow you to see if there are important interesting features of the data, such as multi-modal behaviour or a significant class of outliers.

Consider the outliers
Outliers can be canaries for more fundamental problems with your analysis. It can be fine to exclude them from your data or lump them together into an "unusual" category, but you should know why data ended up in that category. Some outliers you will never be able to explain, so be careful in how much time you devote this.

Report noise/confidence
Be aware that randomness exists and will fool us. If you're not careful, you will find patterns in the noise. Every estimator you produce should have a notion of your confidence in this estimate attached to it. Sometimes, this can be formal and precise (confidence intervals, p-values, Bayes factors for conclusions) and other times it will be more loose, i.e. a quick analysis that offers an estimate: "something between 10 and 12 million".

Look at examples
Anytime you produce new analysis code, you need to look at examples of the underlying data and how your code is interpreting those examples. Your analysis removes lots of features from the underlying data to produce useful summaries. By looking at the full complexity of individual examples, you gain confidence that your summarization is reasonable. You should be doing stratified sampling to look at a good sample across the distribution of values so you are not too focussed on the most common cases.

Slice your data
Separating your data into subgroups and look at the values of your metrics in those subgroups separately.

Consider practical significance
It can be tempting to focus solely on statistical significance or to hone in on the details of every bit of data. But you need to ask yourself, "even if it is true that value X is 0.1% more than value Y, does it matter?" This is especially important if you are unable to understand/categorise part of your data. If you can't make sense of some customer ID numbers, whether it's 0.1% or 10% makes a big difference in how much you should investigate those cases.

Check for consistency over time
You should almost always slice by units of time (days, weeks, whatever). Many disturbances to underlying data happen as our systems evolve over time. This can also give you a sense of the variation in the data that would eventually lead to confidence intervals or claims of statistical significance.

Process

Separate Validation, Description, and Evaluation
EDA has three interrelated stages:

1. *Validation or Initial Data Analysis*: Do I believe the data is self-consistent, that the data was collected correctly, and that data represents what I think it does? This is "sanity checking".
2. *Description*: What's the objective interpretation of this data? E.g. "A small % of users go to the next page of results" or "Users do fewer queries with 7 words in them?"
3. *Evaluation*: Given the description, does the data tell us that something good is happening for the user, the company, the world? E.g. "Users find results faster" or "The quality of the clicks is higher."

By separating these phases, you can more easily reach agreement with others. Descriptions should be things that everyone can agree on from the data. Evaluation is likely to have more debate because you are imbuing meaning and value to the data. If you do not separate Description and Evaluation, you are more likely to only see the interpretation of the data that you are hoping to see. Further. Evaluation tends to be harder because establishing the normative value of a metric, typically through rigorous comparisons with other features and metrics, takes significant investment. These stages do not progress linearly; as you explore the data, you may jump back and forth between the stages – but you should always be clear what stage you are in.

Checking vital signs
Before actually answering the question you are interested in (e.g. "Did users use my new feature?"), you need to check for other things that may not be related to what you are interested in, but may be useful in later analysis or indicate problems in the data. Did the number of users change? Did error rates change? Check your data vital signs to catch potential big problems.

Standard first, custom second
Always look at the standard metrics first, before jumping into the special metrics. This is because standard metrics are much better validated and more likely to be correct. If your new, custom metrics don't make sense with your standard metrics, your new, custom metrics are likely wrong.

Measure twice, or more
Especially if you are trying to capture a new phenomenon. Try to measure the same underlying thing in multiple ways. Check if these multiple measurements are consistent. This allows you to identify bugs in measurement or logging code and unexpected features of the underlying data.

Check for reproducibility
Slicing and consistency over time are particular examples of checking for reproducibility. If the phenomenon is important and meaningful, you should see it across different user populations and time. Furthermore, if you are building models of the data, you want those models to be stable across small perturbations in the underlying data. Using different time ranges or random sub-samples of your data will tell you how reliable/reproducible this model is. If it is not reproducible, you are probably not capturing something fundamental about the underlying process that produced this data.

Check for consistency with past measurements
You will probably be calculating a metric that is similar to things counted in the past. You should compare your metrics against those, even if these measurements are on different user populations. Your number may be right on this population, but you need to validate this. Most surprising data will turn out to be an error, not a fabulous new insight. New metrics should be applied to old data/features first.

If you gather new data and try to learn something new, you won't know if you got it right. You should first apply this data to a known feature or data. For example, if you have a new metric for user satisfaction, you should make sure it tells you your best features help satisfaction. Doing this provides validation for when you then go to learn something new.

Make hypotheses and look for evidence
EDA for a complex problem is iterative. You will discover anomalies, trends, etc. You will make hypotheses to explain this data. Look for evidence (inside or outside the data) to confirm or deny these theories. Good analysis will have a story to tell. To make sure it's the right story, you need to tell the story to yourself, predict what else you should see in the data if that hypothesis is true, then look for evidence that it's wrong. "What experiments would I run that would validate/invalidate the story I am telling?"

Exploratory analysis benefits from end to end iteration
You should strive to get as many iterations of the whole analysis as possible, i.e. multiple steps of signal gathering, processing, modelling, etc. You initial focus should not be on perfection, but on getting something reasonable all the way through. Leave notes for yourself and acknowledge things like filtering steps and data records that you can't understand.

Social

Data analysis starts with questions, not data or a technique
If you take the time to formulate your needs as questions or hypotheses, it will go a long way towards making sure that you are gathering the data you should be gathering and that you are thinking about the possible gaps in the data. Analysis without a question will end up aimless. Avoid the trap of finding some favourite technique and then only finding the parts of problems that this technique works on.

Acknowledge and count your filtering
Every large data analysis starts by filtering the data in various stages. Maybe you only want to consider UK users, or searches with a result click. Whatever the case you must acknowledge and clearly specify what filtering you are doing, and count how much is being filtered at each of your steps – what fraction of queries does your filtering remove?

Educate your consumers
You often present your analysis to non-technical audiences. You need to educate them on how to interpret and draw conclusions from your data. You are responsible for providing the context and a full picture of the data, not just the number the consumer asked for.

Share with peers first, external consumers second
A skilled peer reviewer can provide qualitatively different feedback and sanity-checking than the consumers of your data can. You can find out about gotchas they know about, suggestions for things to measure, and past research in this area. Near the end, peers are good at pointing out oddities, inconsistencies, or other confusions.

Expect and accept ignorance and mistakes

There are limits to what we can learn from data. We need to admit the limits of our certainty; this is a strength that is not usually immediately rewarded. Ultimately, it will earn you respect from colleagues and leaders who are data-wise. It feels bad making a mistake and discovering it later, but proactively owning up to your mistakes will translate into credibility. Credibility is the key social value for any data scientist.