Practical big data analysis

Curated practical advice for analysis of large, complex, chaotic datasets

Author: Morales, J.A.
Last update: January 6, 2023

Contents

Introduction

Curated practical advice for analysis of large, complex, chaotic datasets.

Visualize distribution in different ways

We frequently see histograms when exploring data. They show the distribution of data over a continuous interval. A bar in a histogram corresponds to the tabulated frequency in that bin. Histograms highlight the presence of gaps, concentrated values, and extreme values.

Other visualizations that can help enrich our understanding of the distribution include box & whisker plot, bubble chart, density plot, and violin plot.

Outliers aren’t just for throwing away

It’s often beneficial to quickly check how outliers came to be.

Measure the trust you put in numbers

Estimates are useful but it is equally important to measure the confidence we put in them.

Sampling can see what summaries can’t

Stratified sampling will reflect a more randomized data. By doing this, we avoid the pitfall of seeing only the most common cases.

Spot-checking is another simple yet non-trivial practice for complementing summaries.

Slice and dissect in different ways

We divide and measure the data into different ways. We compare and contrast the metrics to help confirm a phenomenon, if it exists in different dimensions, if there’s bad data within a subgroup, or if there’s any fundamental difference. This will help

Prioritize practical significance

We can figure out all the mysteries in our data, if we had all the time in the world. But we don't. Spend time on what matters.

Don’t stop at launch, monitor

It's not uncommon for something to break after a thorough initial analysis. Check for consistency over time to see any breakage or get a

feel of the variation in the data

Which phase are you on?

Phases in data analysis are never linear. It's alright to jump back and forth between the stages but at any time, we should be clear what stage we're in.

Background-check your data

Who captured your data? What did they attempt to capture? Where, when, and how was the data captured? Confirm the experiment and data collection set-up so that you can have a richer understanding for your future choices.

Health-check your data

Regular check-ups could help catch potentially big problems in the health of the data. Check out the vital signs, even if they may not seem related to the main interest.

Take the standard path

Customisation is tempting because of the endless possibilities. However, there is some predictability in standard metrics and configurations that customisation doesn't offer. Take the standard path when given the choice. Customize only when necessary.

Measure more than once

Measure in different ways and then check for consistency across these measurements. This is one way to identify bugs in the metrics and unexpected features of underlying data.

Reproduce the model to check for stability

Is the model reproducible? When building models, we want them to be stable across small perturbations in the underlying data. If not, did we miss capturing something fundamental?

Compare and contrast with past metrics

Are the current measurements consistent with past measurements? Exact agreement is not necessary but a similar ballpark is a good sign. It contributes to the validation our numbers and new data.

Back your hypothesis

Find evidence that will support your hunches and hypothesis. It can come from within the data or from external sources.

Explore to the end

The focus from the start should be to get something reasonable completed, all the way to the end. It's a trap to think that one phase can be completed to perfection. We will never finish that way. Instead, we can iterate and do the phases better each time. Taking it all the way to the end provides analytical insights that would have been missed otherwise.

What problem are you trying to solve?

In a business set-up, it's usually productive to start with the question, not the technique. Do you have a favourite technique? Do you catch yourself defaulting to that technique time and again? A particular technique will be effective in exposing problems that it is good for but it might not be effective in exposing other kinds of problems.

Don't forget what's been filtered

With many iterations and details, some manipulation made on the data are at risk of being forgotten. Acknowledge and clearly specify the filtering done and count how many are filtered out in each step.

Clarify the context of what you're counting

The story behind the data is easier to follow and closer to the truth when we have clear context. Ideally, your ratios have clear numerators and denominators and you zoom in and out of the timeline to show how the numbers fit into the big picture.

Communicate with your consumer

Educated consumers are empowered consumers. They are empowered to make informed decisions, ask intelligent questions we might have overlooked, and partner with us in our mission to find insights. Update them about progress and blockers. Be honest about the short-term and long-term impact of the findings, the limitations and caveats, and possible alternatives.

Advocate your insights but stay skeptical

Be both skeptic and champion of the insights you find.

Peer review first, present to consumers second

Share with peers from the start to gather suggestions for what to measure. Throughout the process, their fresh eyes and wealth of experiences could help point out oddities, inconsistencies, or other confusions. This might also lead to earned support from internal/technical team, which is crucial. This gives us a chance to address or be aware of issues before presenting to external consumers

Failing is ok

Expect and accept ignorance and mistakes. A conducive workplace will reward honesty and integrity, including team members who own up to their failures. This helps create a culture of growth and innovation, which has significant impact beyond the project.

References

Riley, P. (2016) Practical advice for analysis of large, complex data sets. Available at: https://www.unofficialgoogledatascience.com/2016/10/practical-advice-for-analysis-of-large.html (Accessed: April 6, 2019).