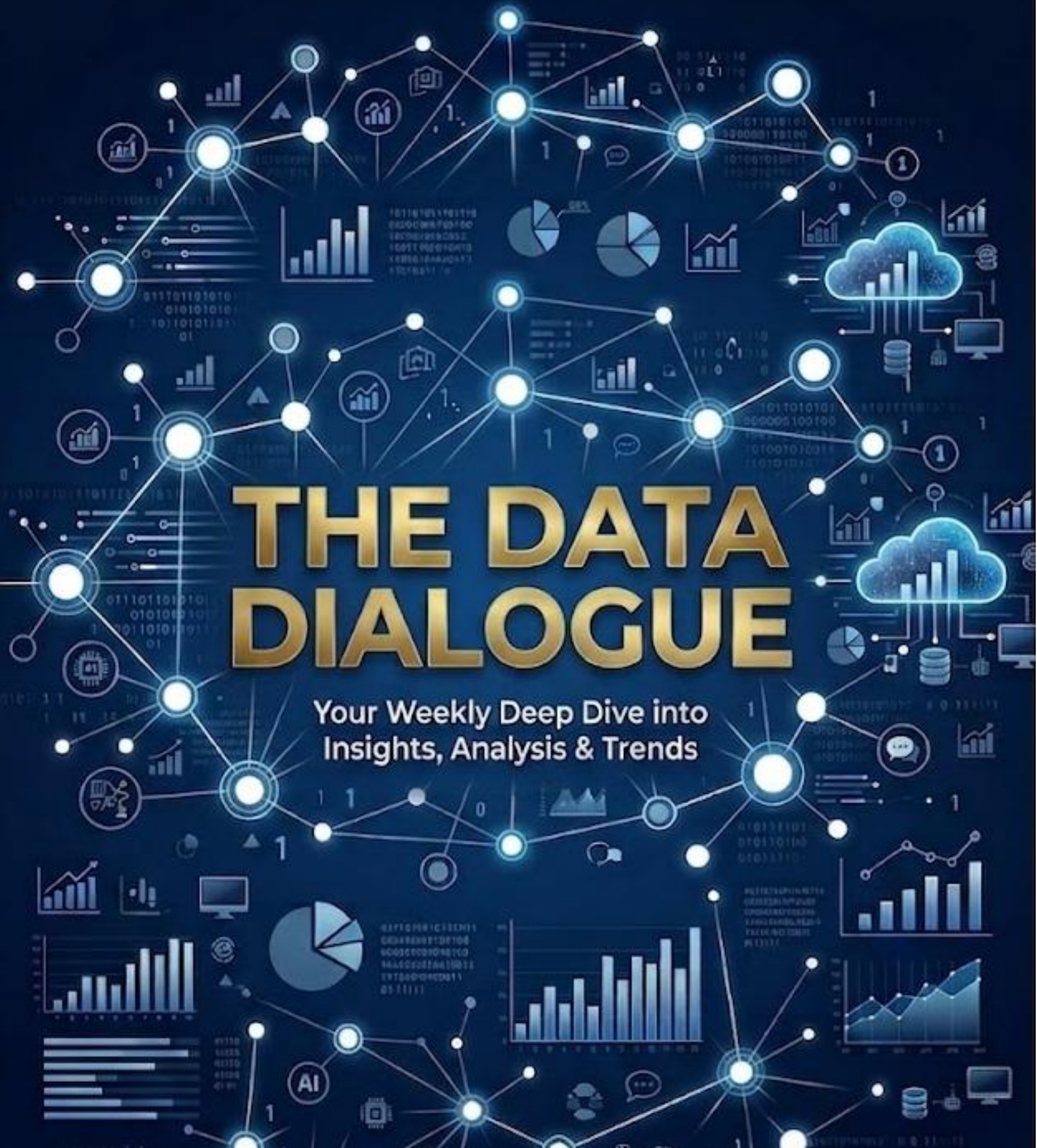


VOLUME 1| ISSUE 7| August 2026

THE DATA DIALOGUE

Your Weekly Deep Dive into
Insights, Analysis & Trends



The Data Dialogue.

Simplified. Not Simplistic

August 26, 2026 | Vol.1 | Issue 7 | Data Cleaning: Missing Data

In-Focus

Tinkering with Data-Cleaning Decisions: When Convenience Costs You the Game

It is rainy today and you decided to chill in and skip this day of training. You feel comfortable, cozy, and strongly convinced that missing out one day will not matter! Next day, your muscles ache, your coach is not happy and clearly said “if you keep skipping days, you will not make it to the tournament”. Few days later, the coach told you to cut off the carbs intake, but you decided to drive through McDonald’s to have some chips.... well, you love them! Next week, the trainer noticed that your weight did not go down as expected despite working out! He said, “if you did not cut carbs and get down to the required weight, you will not be eligible for the tournament”. Few weeks later, your tests recorded a spike in your blood sugar levels, and you shrugged it off and why not, “it is just a bad day”. A month later, your physician told you that your tests show a liver problem that gave you diabetes!

Ah, the tournament?! Sorry, you will not be able to participate.

Your journey ended before it started?! Shocked? Well, you should not be. Every single decision you made seemed convenient and harmless, deleting what you should have kept, keeping on what should have been deleted, and ignoring outliers, led eventually to this end.

This is exactly how data cleaning decisions might do to your research journey, if you did not think about them thoroughly. Data cleaning decisions are not about simply having complete data inputs for every observation/row. Neither is it about getting rid of what looks strange or does not fit into the general behavior. On the contrary, it is a very systematic process with clear purpose: capturing meaningful patterns, identifying areas of improvement and segmenting patterns if needed. All that can be achieved is the researcher posed to ask:

- What is the magnitude of missing data or outliers?
- Why is this the case? What does this say about the phenomenon?
- And if I just delete them, what will I miss out?
- Am I deleting an error? Or something inconvenient?

Censor your data-cleaning decision when no one is watching. Document every decision to weather the storm of peer-review process- and remember, every decision comes at a cost.

The Data Dialogue.

Simplified. Not Simplistic

August 26, 2026 | Vol.1 | Issue 7 | Data Cleaning: Missing Data

Tools: Before Asking What to Do with Missing Data

Step 1: Ask Why They Are Missing

1. **Missing completely at random (MCAR):** when missingness is the result of pure bad luck.
Example: you were conducting a survey when a number of computers crashed down. When they were restarted, the computer skipped the unanswered question and took the students to the next question! Pure bad luck- a technical glitch that has nothing to do with the participants.
2. **Missing at random (MAR)** When missingness is related to one of the observed/reported variables, but not to the missing value itself.
Example: In a clinical study, older patients are reported to higher percentage of missing follow-up tests compared to younger patients. Here, missingness is linked to an observed variable (age).
3. **Missing not at random (MNAR):** When missingness is related to the missing value itself. Example, in an economic study, it was noted that most wealthy people do not report their income levels. Here, missingness of income variable is linked to the value of the unreported income! The higher the income, the less likely the individual is going to report it.

And that's only Step 1. The next question is: What should you do about it?

To be continued in the coming issues...

Pitfall Patrol

 Immediately replacing missing data with a zero or the mean!

 keep them as they are (missing value) → Use imputation method depending on the type of variable and analysis you

Why? It might seem convenient, but zero is an answer, while missing is just missing.

And using the mean can also distort distributions and affect the overall findings.

The Data Dialogue.

Simplified. Not Simplistic

August 26, 2026 | Vol.1 | Issue 7 | Data Cleaning: Missing Data

Pandora's Box

The 5 % Deletion Rule: A Convenience Trap!

It is not known- at least for me- who came up with the rule that says: if missing values represent less than 5% of your sample size, delete them unapologetically, and move on. Getting rid of small percentage of missingness will not harm large samples, and it seems a practical numeric rule to follow. And let's be honest, by the end of the day, deleting 20 observations is much easier than imputation data to fill into just 20 missing cells!

On the other hand, this rule stumbles on the messy reality of data collection. It is easy to apply when all missing data are concentrated in one variable. What if you have more than a variable with each has less than 5% missing values- with no overlap. In this case, you go ahead and delete 4%- for instance- of the observations for Variable A and another 4% of the observations for Variable B; and you end up losing 8% of your sample size in two keyboard clicks! Imagine the impact these decisions have on the statistical power of your findings.

Now, imagine that in a best-case scenario, it is only about one variable, but it is linked to specific demographic group, and you decided to delete these rows. Your dataset looks cleaner? Congratulations, you have just wiped out an entire demographic group of the population. Can you imagine the impact it has on your findings' robustness and generalizability. The < 5% innocent rule is not so innocent after all.

In a worst-case scenario, imagine that your missing values happen to overlap with outlier values of another variable on the same observation; and you just delete them! Imagine that you throw off the same part of reality that your analysis should have revealed or flagged.

Following a numeric rule might be easy, but following it blindly is another form of convenience- which comes at a cost.

Before you Go

Did it happen that you made a mistake while conducting data analysis? If so, send us a punchy 250 words describing this mistake and how you fixed it.

If it's humanized, we'll publish it next week with your name, title and link to your profile.

Disclaimer: "We reserve the right to edit all submissions for conversational tone, length, and formatting to ensure it matches the Data Dialogue voice"

You will find the submission button on my website.