**Data Cleaning 101: Avoid These 5 Traps in Your Data**

*Data cleaning is a foundational skill in data science. Learn the five most common traps your data hides and how to recognize and handle them like a pro.*

[[Nathan Rosidi](https://nathanrosidi.medium.com/?source=post_page-----c31d5cb08a56--------------------------------)](https://nathanrosidi.medium.com/?source=post_page-----c31d5cb08a56--------------------------------)

[[Stackademic](https://blog.stackademic.com/?source=post_page-----c31d5cb08a56--------------------------------)](https://blog.stackademic.com/?source=post_page-----c31d5cb08a56--------------------------------)

[Nathan Rosidi](https://nathanrosidi.medium.com/?source=post_page-----c31d5cb08a56--------------------------------)

·

[Follow](https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fsubscribe%2Fuser%2Fab636cbf3611&operation=register&redirect=https%3A%2F%2Fblog.stackademic.com%2Fdata-cleaning-101-avoid-these-5-traps-in-your-data-c31d5cb08a56&user=Nathan+Rosidi&userId=ab636cbf3611&source=post_page-ab636cbf3611----c31d5cb08a56---------------------post_header-----------)

Published in

[Stackademic](https://blog.stackademic.com/?source=post_page-----c31d5cb08a56--------------------------------)

·

5 min read

·

Nov 28, 2023

412

4

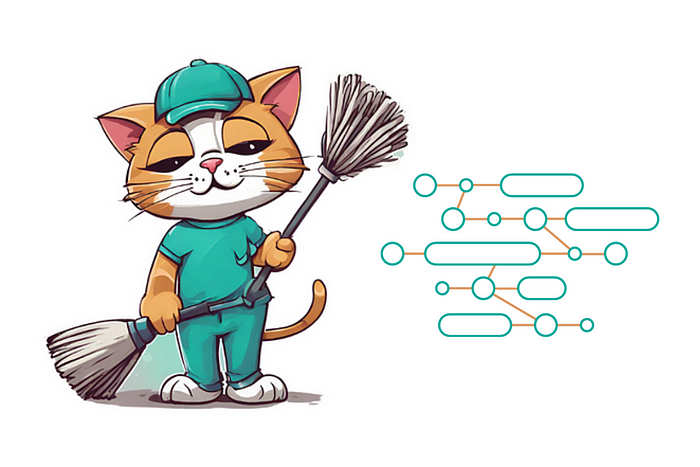


Image created by the author on [Canva](https://www.canva.com/)

Not all data scientists are created equal — some are amazing at creating machine learning models, while others are great at creating visuals.

But one thing all — literally all — data scientists need to be good at is cleaning data. So how do you get good? You should start with not sucking at it first!

And you stop sucking when you stop ignoring these five mistakes in your data. Let’s first see what they are. At the end of the article, I’ll propose one solution that can solve all five of these mistakes.

If you like hearing my voice while you read this article, here’s the video where I talk about these data-cleaning mistakes.

**Mistake 1: Handling Missing Values**

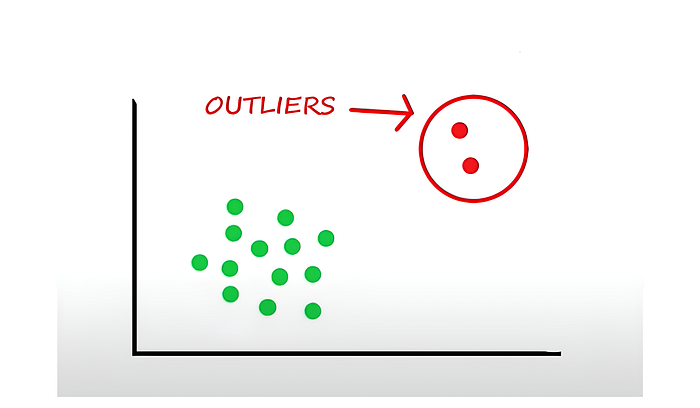
One thing you’ll notice when you’re cleaning data is that there are almost always missing values. Most people don’t know why data is missing. A higher power at work? Data being lost in the laundry? It’s not up to you to find an answer. What you need to do is handle these missing values.

Not doing it can introduce bias and lead to crazy, inaccurate conclusions. There are many ways of handling missing data. These three are the most common:

1. **Removing** **rows or columns** that have the missing data
2. **Imputing** **missing values** using proxies like [mean, median, or mode](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/)
3. **Using algorithms** that can handle these missing values, e.g., [random forest](https://scikit-learn.org/stable/modules/ensemble.html#forest)

**Mistake 2: Handling Outliers**

The opposite of missing values are outliers. It’s just a fancy word for data that makes no sense.



Not handling these outliers correctly can skew your view and your conclusions. But also, how do you know an outlier is an outlier? For example, a $2 million house might be an outlier. Unless you live in New York or San Francisco, that is. Then, $2 million is just the average price.

How do you, then, find outliers, and how do you deal with them? There are two ways:

1. **Visual approach** — Creating [box plots](https://en.wikipedia.org/wiki/Box_plot) and [scatter plots](https://en.wikipedia.org/wiki/Scatter_plot) so you can *see* the outlier
2. **Statistical approach** — Using [z-score](https://en.wikipedia.org/wiki/Standard_score) or [IQR methods](https://en.wikipedia.org/wiki/Interquartile_range) to find these outliers statistically

**Mistake 3: Handling Data Inconsistency**

This is all fine, but quite often, you can’t even read the data to begin with. It’s a common issue in data cleaning: data inconsistency. In other words, you might have different date formats or case usages, making it very difficult to analyze data.

To fix this, you need to make data formats consistent by:

1. **Standardizing** the data to maintain consistency in terms of format unit and scales
2. **Automating** the work by implementing data validation checks, aka unit tests software developers use for testing features.

**Mistake 4: Handling Datatype Issues**

It’s not just the format that can cause problems. Data types can, too. Not checking them until it’s too late is also one of the common mistakes in data cleaning. The solution to this is straightforward:

1. **Inspect, cast, and convert** your data types
2. **Automate** by implementing checks to validate data types at various stages of your project.

For example, suppose a DateTime variable like order date is stored as a string. In that case, you will obviously need to cast that as a DateTime data type to enable your calculations and analysis.

Knowing what data types your database can use and which type can be converted to, like in the below matrix, becomes very useful knowledge.

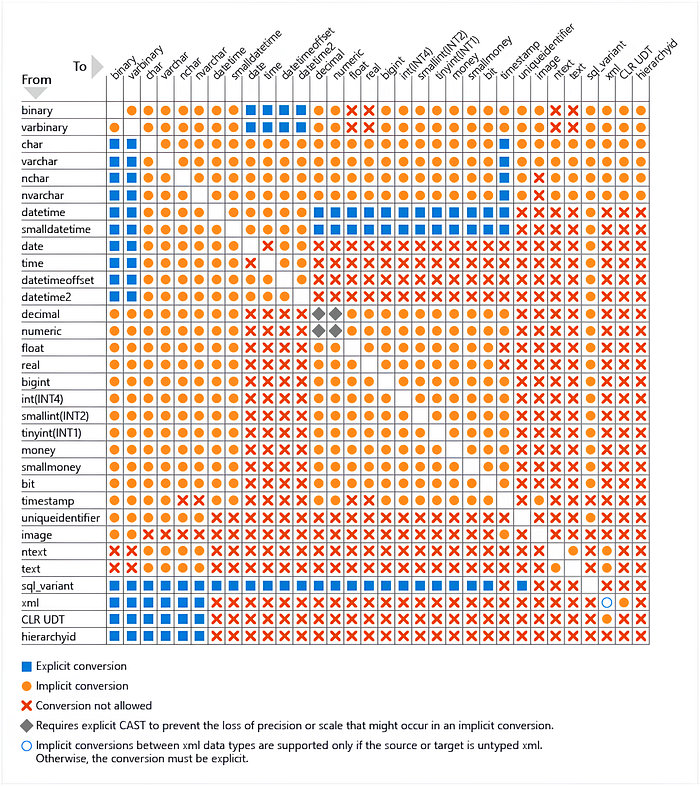


Image Source: [learn.microsoft.com](https://learn.microsoft.com/en-us/sql/t-sql/data-types/data-type-conversion-database-engine?view=sql-server-ver16)

**Mistake 5: Handling Duplicate Data**

Having duplicates in your data is the most common mistake. It often comes from bad joins or merges, and it ends up blowing your record count. Many data scientists don’t do a post-check or post-inspection to ensure no duplicates in the data. Don’t be one!

To dedupe your data:

1. **Group it** so that it’ll deduplicate automatically
2. **Add checks** at every stage of your analysis to prevent unwanted duplicates

If you’re working in SQL, [here’s more about data deduplication and other common SQL coding errors](https://www.stratascratch.com/blog/top-most-common-sql-coding-errors-in-data-science/?utm_source=blog&utm_medium=click&utm_campaign=stackademic+data+cleaning+traps).

**A Solution to Avoiding All Five Mistakes at Once**

What do all five above mistakes have in common? They mainly involve manual inspection of the data and then a manual fix. The solution now becomes apparent: Why not automate as much as possible?

As I said before, software developers have this concept called [unit testing](https://en.wikipedia.org/wiki/Unit_testing). It ensures that their work passes the requirements of the feature that they’re trying to build. You should do the same for data science. To be more specific:

1. **Reduce mistakes** — Create scripts that automate all of these data cleaning tasks that are always on repeat at every step of the process.
2. **Make your process reproducible** — Document the data cleaning process to ensure everyone on your team (including you!) knows what business rules logic is being used and can repeat it.

**Conclusion**

Implement what you learned here, and you’re on a path to success in data cleaning. There’s really no need to keep repeating these mistakes. They are embarrassing but also easy to avoid.

I know data cleaning is not as exciting as creating sexy graphs or ML models. But all data scientist spend their time (and a significant part, to tell the truth!) doing it. Without it, every other stage of a [data science project](https://www.stratascratch.com/blog/19-data-science-project-ideas-for-beginners/?utm_source=blog&utm_medium=click&utm_campaign=stackademic+data+cleaning+traps) doesn’t exist. So, first, do the groundwork and master data cleaning. Only after that should you learn other, fancier data science tasks.

*Originally published at*[*https://www.stratascratch.com*](https://www.stratascratch.com/blog/data-cleaning-101-avoid-these-5-traps-in-your-data/?utm_source=blog&utm_medium=click&utm_campaign=stackademic+data+cleaning+traps)*.*

**Stackademic**

*Thank you for reading until the end. Before you go:*

* *Please consider****clapping****and****following****the writer! 👏*
* *Follow us on*[***Twitter(X)***](https://twitter.com/stackademichq)*,*[***LinkedIn***](https://www.linkedin.com/company/stackademic)*, and*[***YouTube***](https://www.youtube.com/c/stackademic)***.***
* *Visit*[***Stackademic.com***](http://stackademic.com/)*to find out more about how we are democratizing free programming education around the world.*