

The challenge of missing or duplicate data

The ins and outs of data outliers

- Video: Account for outliers 6 min
- Reading: Protect the people behind the data 20 min
- Video: Identify and deal with outliers in Python 13 min
- Reading: Reference guide: How to handle outliers 20 min
- Practice Quiz: Test your knowledge: The ins and outs of data outliers 3 questions

Change categorical data to numerical data

Input validation

Review: Clean your data

Protect the people behind the data

We've all been there...

Whether at work or at school, there's a moment of realization about an essay or a project you've been working on—you suddenly realize you've made a mistake, and it needs to be fixed.

"But think of all of the trouble that will cause," you think. "It will make the project late, and everyone will find out I made a mistake."

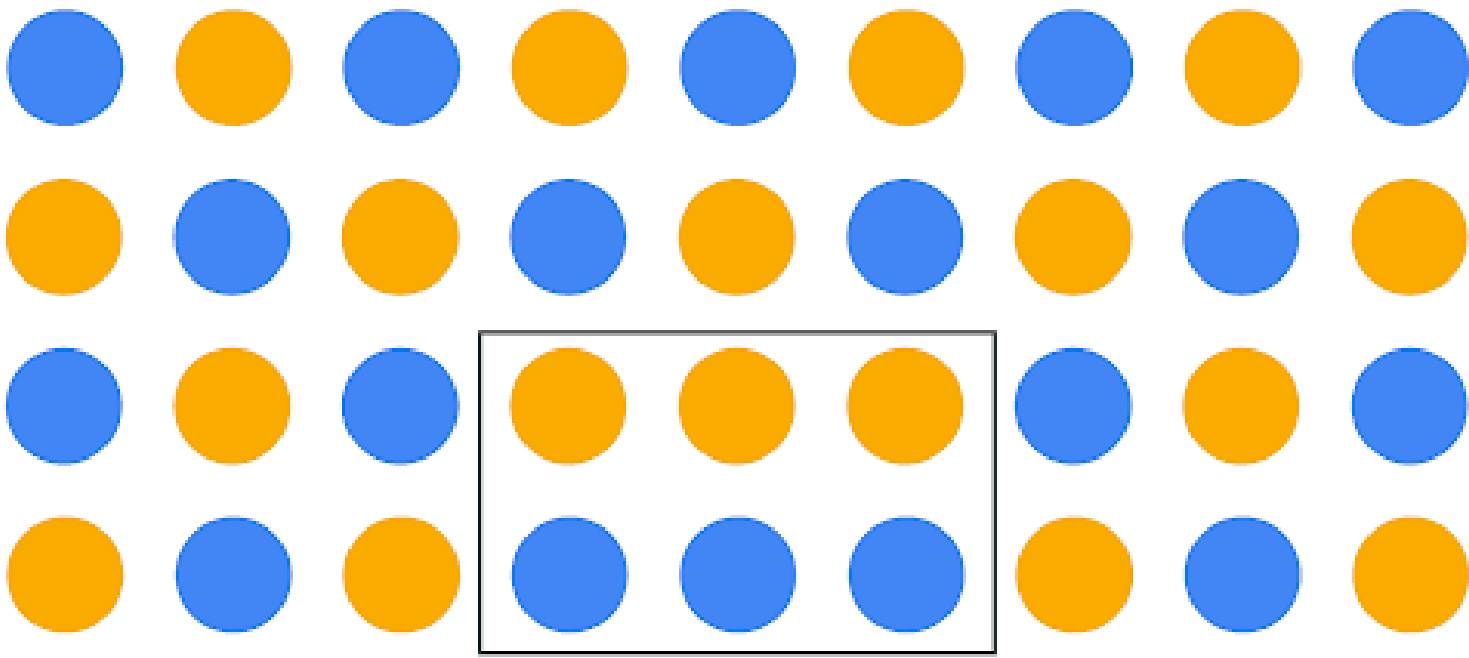
Is it really worth the time and effort to stop the process and fix it?

For data professionals, the answer needs to always be **YES**

The big picture

Even seasoned data professionals can fail to see the big picture. As intuitive as our brains can be, we often fail to see how a small error, a tiny change, or a seemingly insignificant choice about data analysis can have significant implications on a process, a business, or even a whole population.

To help illustrate this, imagine a grid of alternating orange and blue circles. Now, imagine there is a person who can only see a small section of six circles. What if this person decides that it makes more sense for the circles to be arranged by groups, with orange circles on top and blue on bottom? You'll find that even though the circles within the rectangle may be perceived as orderly, the reality is the change has actually disrupted the larger pattern. Now imagine the rectangle of six circles is a department and the bigger grid is an entire company.



An ethical mindset

Hopefully, this illustration helps you understand the importance of context and scope when dealing with data. One principle that can help data professionals keep data in context is ethics and developing an ethical mindset.

Here are three important concepts to help you foster an ethical mindset as you seek to tell stories with data:

- Models have lives beyond code blocks and data strings.
- Data science needs regulatory compliance.
- Customers should choose how their data is used.

Models have lives beyond the code blocks and data strings

There can be a tendency to think of data as strictly numbers and math. The reality is that data consists of people-driven inputs. Data professionals should never lose sight of the fact that analysis in all forms—visualizations, models, decisions, and strategies—impacts people.

The end goal of all data analysis and data science work should be to improve the lives of individuals and groups. Next time you are designing a data dashboard or coding a complex algorithm, take a moment to consider how the data might impact human beings, and whether the decisions you make alter that impact for the better. Remember that your own biases may prevent you from being able to anticipate many impacts, so gathering input from a diverse group of team members helps you mitigate your personal biases.



Here is a hypothetical data career example to help you see how this concept may someday impact your work:

- Imagine a data professional is analyzing traffic flow data in order to help a city planner for a major metropolitan area decide where to focus on expanding roads. At first glance, using data for road construction appears to be a purely financial decision between a city and a road construction company. The professional's analysis first led them to select a more hilly, uneven terrain as the most cost-effective place for the expansion of roads. But after analyzing the data with concern for the citizens and families who will be driving on those roads in all seasons, the data professional's analysis determined the hilly area to have a higher accident risk, and decided to recommend a flatter location for road expansion.

Data science needs regulatory compliance

If you are unfamiliar with the laws and regulations of the industry you work in, your job will not only be much more difficult, but also could land you in legal trouble. A data professional should always remain up to date and in compliance with all regulations in their particular field. As an example, when you begin a new job in the data career space, you can do the following:

- Ask your manager for compliance guidance.
- Take time to research the data governance policies.
- Take time to research the regulatory body of your particular industry and its relevant policy documents.

Keep in mind that the data governance field is quite young and that regulations can and will change at a quick pace. Be sure to keep up with the changes!

It's important to follow the rules ... but not **just** for rules' sake. Here are a few important benefits of remaining compliant with regulatory bodies and data governance policies:

- Keeps client and company data safe from security threats
- Bolsters trust with clients, peer groups, companies, and public
- Lessens likelihood of lost or mismanaged data
- Ensures business critical data remains usable, accessible, and available

Customers should choose how their data is used

With each passing day, technology gets more advanced and capable. As data becomes increasingly involved in our day-to-day lives, it stands to reason the data itself becomes more and more important and valuable. Because of this, data privacy and security is critical to a company's clients and a government's citizens.

As data professionals, it is our ethical responsibility to treat client and customer data with respect and dignity. Companies that store, analyze, and utilize data as part of their business processes should make sure those processes are transparent and compliant. Customers who divulge confidential information in order to procure a service should be able to trust that the data will remain confidential and not compromised due to breaches or cyber security threats.

You'll find that many companies have begun to treat customer data with more internal care and security. More companies are providing extensive employee training regarding the handling and securing of customer data. Many are adding multi-factor authentication systems and auditing their third-party vendors, requiring high levels of digital security systems and platforms.

As an example, because data breaches like Home Depot in 2014, Uber in 2017, and Instagram in 2020 have been commonplace for decades, corporations are increasingly ramping up cybersecurity, employee and vendor training, and transparency to customers and the public regarding data gathering and data use methods. Specifically, Home Depot has added data security into its risk management plan, and built a data security web page that describes what data they gather and how they use it.

Data transparency is not only good for business, it is the right thing to do—the ethical thing to do. Data professionals need to understand that more than anyone.

Key Takeaways

Data ethics is an expanding field and remains an integral part of a data professional's daily work. While manipulating, analyzing, or using data in any capacity, remember that data models have lives beyond code blocks and data strings. Data science needs regulatory compliance and customers should always get to choose how their data is used.

Resources for more information

Data governance and data ethics are a growing field of work and study. Here are some resources should you like to learn more:

- Harvard Business School: 5 principles of Data Ethics for Business
- Dataiversity: Data Governance and Data Quality

Mark as completed

Like Dislike Report an issue