Business Insights

Harvard Business School Online's Business Insights Blog provides the career insights you need to achieve your goals and gain confidence in your business skills.

+ TOPICS

+ COURSES

DATA WRANGLING: WHAT IT IS & WHY IT’S IMPORTANT

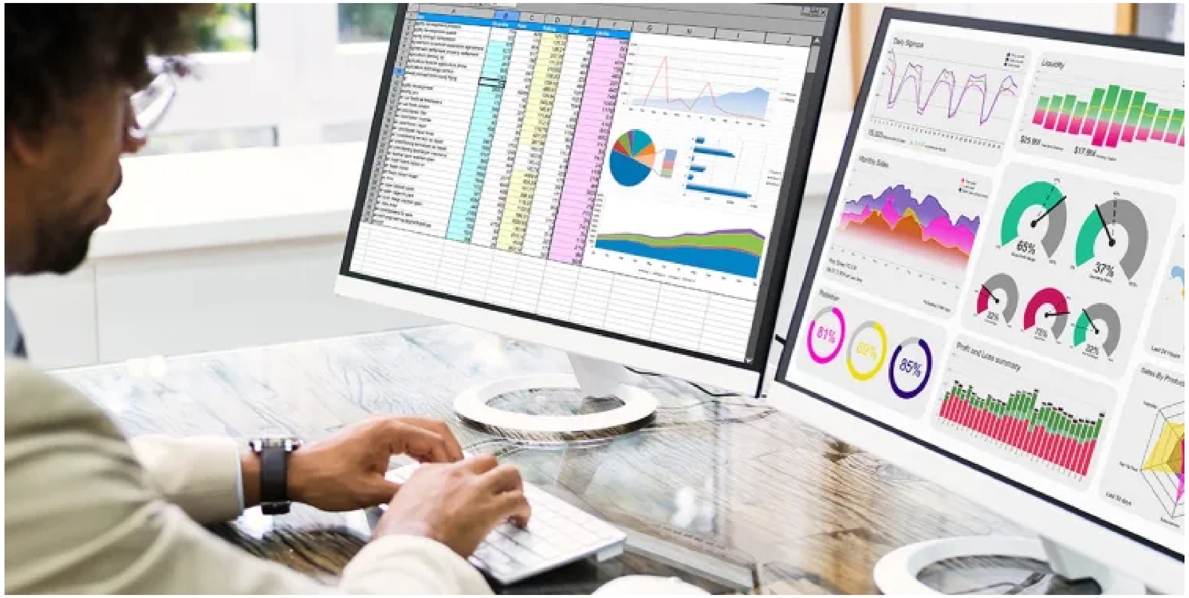
Subscribe to the Blog

Email*

Using assistive technology? Get more details on using this form.

SUBMIT

RSS feed



19 JAN 2021

Tim Stobierski | Contributors

Analytics

Email Print Share

Businesses have long relied on professionals with data science and analytical skills to understand and leverage information at their disposal. With the proliferation of data, due to the development of smart devices and other technological advancements, this need has accelerated.

It’s impossible to choose a single data science skill that’s most important for business professionals. One thing that's certain, however, is that insights are only as good as the data that informs them. This means it’s vital for organizations to employ individuals who understand what clean data looks like and how to shape raw data into usable forms. This is where data wrangling comes into play.

Below is an overview of what data wrangling is, its key steps, and why it’s crucial for business.

WHAT IS DATA WRANGLING?

Data wrangling—also called **data cleaning**, **data remediation**, or **data munging**—refers to a variety of processes designed to transform raw data into more readily used formats. The exact methods differ from project to project depending on the data you’re leveraging and the goal you’re trying to achieve.

Some examples of data wrangling include:

- Merging multiple data sources into a single dataset for analysis
- Identifying gaps in data (for example, empty cells in a spreadsheet) and either filling or deleting them
- Deleting data that’s either unnecessary or irrelevant to the project you’re working on
- Identifying extreme outliers in data and either explaining the discrepancies or removing them so that analysis can take place

Data wrangling can be a manual or automated process. In scenarios where datasets are exceptionally large, automated data cleaning becomes a necessity. In organizations that employ a full data team, a data scientist or other team member is typically responsible for data wrangling. In smaller organizations, non-data professionals are often responsible for cleaning their data before leveraging it.

DATA WRANGLING STEPS

Each data project requires a unique approach to ensure its final dataset is reliable and accessible. That being said, several processes typically inform the approach. These are commonly referred to as data wrangling steps or activities.

1. Discovery

Discovery refers to the process of familiarizing yourself with data so you can conceptualize how you might use it. You can liken it to looking in your refrigerator before cooking a meal to see what ingredients you have at your disposal.

During discovery, you may identify trends or patterns in the data, along with obvious issues, such as missing or incomplete values that need to be addressed. This is an important step, as it will inform every activity that comes afterward.

2. Structuring

Raw data is typically unusable in its raw state because it's either incomplete or misformatted for its intended application. **Data structuring** is the process of taking raw data and transforming it to be more readily leveraged. The form your data takes will depend on the analytical model you use to interpret it.

3. Cleaning

Data cleaning is the process of removing inherent errors in data that might distort your analysis or render it less valuable. Cleaning can come in different forms, including deleting empty cells or rows, removing outliers, and standardizing inputs. The goal of data cleaning is to ensure there are no errors (or as few as possible) that could influence your final analysis.

4. Enriching

Once you understand your existing data and have transformed it into a more usable state, you must determine whether you have all of the data necessary for the project at hand. If not, you may choose to **enrich** or **augment** your data by incorporating values from other datasets. For this reason, it's important to understand what other data is available for use.

If you decide that enrichment is necessary, you need to repeat the steps above for any new data.

5. Validating

Data validation refers to the process of verifying that your data is both consistent and of a high enough quality. During validation, you may discover issues you need to resolve or conclude that your data is ready to be analyzed. Validation is typically achieved through various automated processes and requires programming.

6. Publishing

Once your data has been validated, you can **publish** it. This involves making it available to others within your organization for analysis. The format you use to share the information—such as a written report or electronic file—will depend on your data and the organization's goals.

Harvard Business School Online
A Beginner's Guide to Data & Analytics
Access Your Free E-Book
DOWNLOAD NOW

THE IMPORTANCE OF DATA WRANGLING

Any analyses a business performs will ultimately be constrained by the data that informs them. If data is incomplete, unreliable, or faulty, then analyses will be too—diminishing the value of any insights gleaned.

Data wrangling seeks to remove that risk by ensuring data is in a reliable state *before* it's analyzed and leveraged. This makes it a critical part of the analytical process.

It's important to note that data wrangling can be time-consuming and taxing on resources, particularly when done manually. This is why many organizations institute policies and best practices that help employees streamline the data cleanup process—for example, requiring that data include certain information or be in a specific format before it's uploaded to a database.

For this reason, it's vital to understand the steps of the data wrangling process and the negative outcomes associated with incorrect or faulty data.

Do you want to further your data literacy? Download our Beginner's Guide to Data & Analytics to learn how you can leverage the power of data for professional and organizational success.

About the Author

Tim Stobierski is a marketing specialist and contributing writer for Harvard Business School Online.

Top FAQs

All FAQs

How are HBS Online courses delivered? +

Do I need to come to campus to participate in HBS Online programs? +

How do I enroll in a course? +

Does Harvard Business School Online offer an online MBA? +

What are my payment options? +

What are the policies for refunds and deferrals? +

Sign up for News & Announcements

Subject Areas

Business Essentials
Leadership & Management
Entrepreneurship & Innovation
Strategy
Finance & Accounting
Business & Society

Quick Links

FAQs
Contact Us
Request Info
Apply Now

About

About Us
Media Coverage
Founding Donors
Leadership Team
Careers @ HBS Online

Legal

Legal
Policies

Site Map Trademark Notice Digital Accessibility