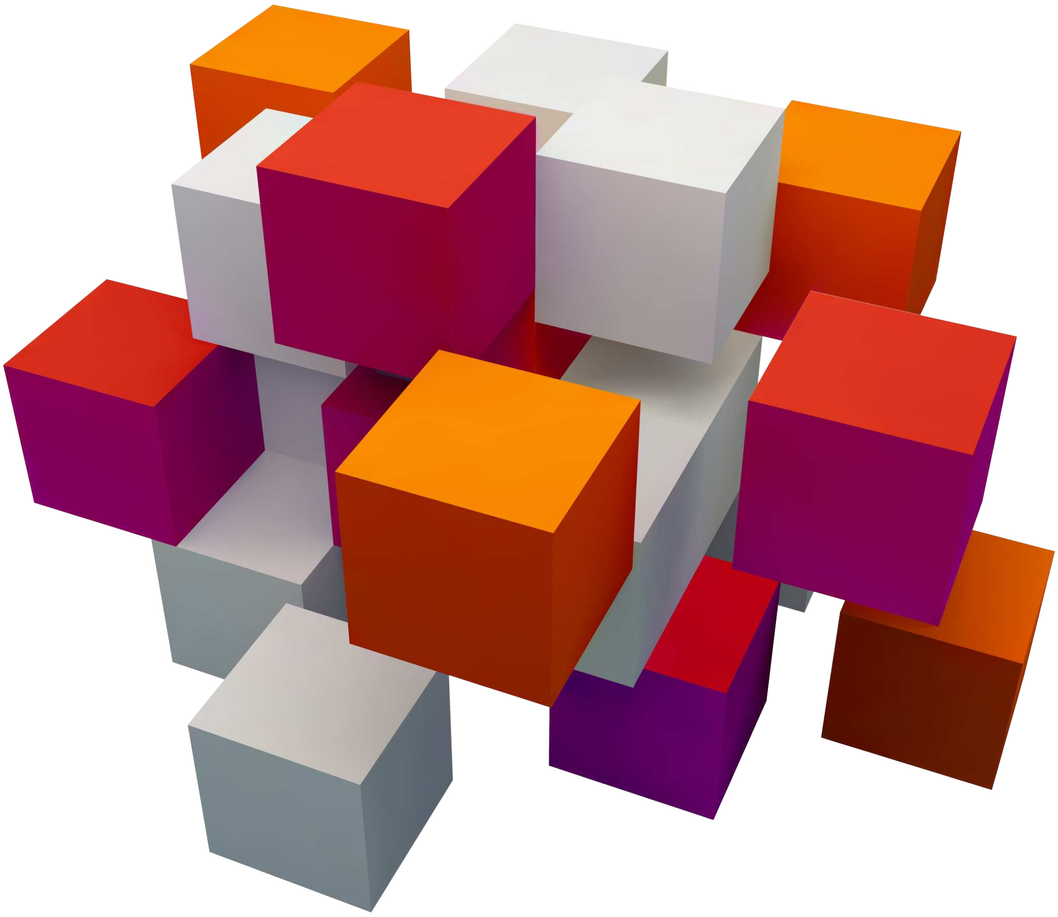
Data Wrangling Steps



DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

What is Data Wrangling?

Data wrangling - also called data remediation, or data munging - refers to various processes designed to transform raw data into more readily used formats.

Wrangling the data is crucial and is considered as the backbone of the entire analysis part.

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

Examples of Data Wrangling

> Merging multiple data sources into a single dataset for analysis.

> Identifying gaps in data 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.

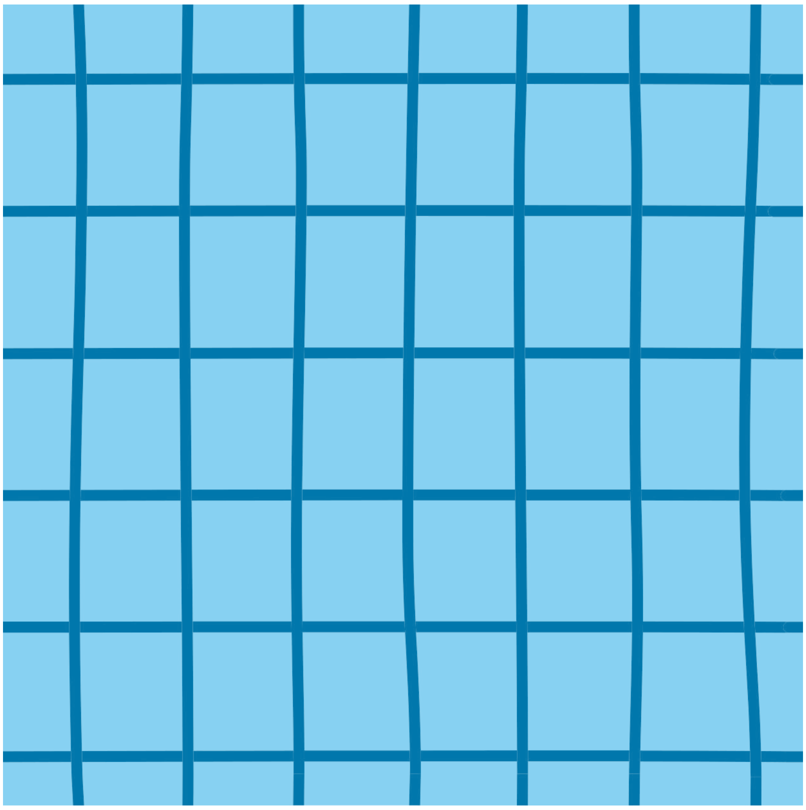
DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Step 1: Discovery

Discovery refers to the process of familiarizing yourself with data so you can conceptualize how you might use it.

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.

Step 2: Structuring



Data structuring is the process of taking raw data and transforming it to be more readily leveraged. Raw data is typically unusable in its raw state because it’s either incomplete or misformatted for its intended application.

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

Step 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.

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Step 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.

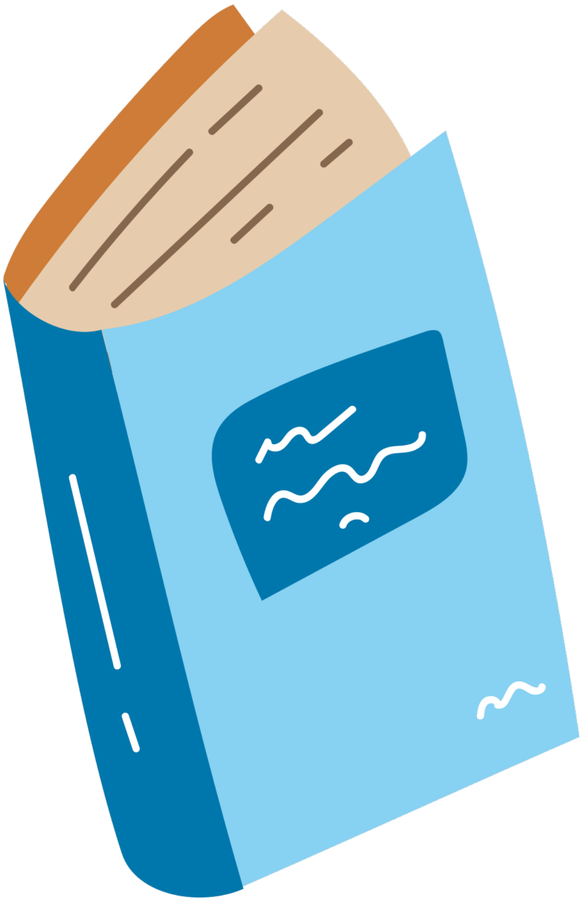
DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Step 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.

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Step 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.

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

✉ info@dataranch.org

in linkedin.com/company/dataranch