Data Cleaning Steps

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

“Garbage in, garbage out"

if you start with bad data (garbage), you’ll only get “garbage” results.

Data cleaning is often a tedious process, but it’s absolutely essential to get top results and powerful insights from your data.

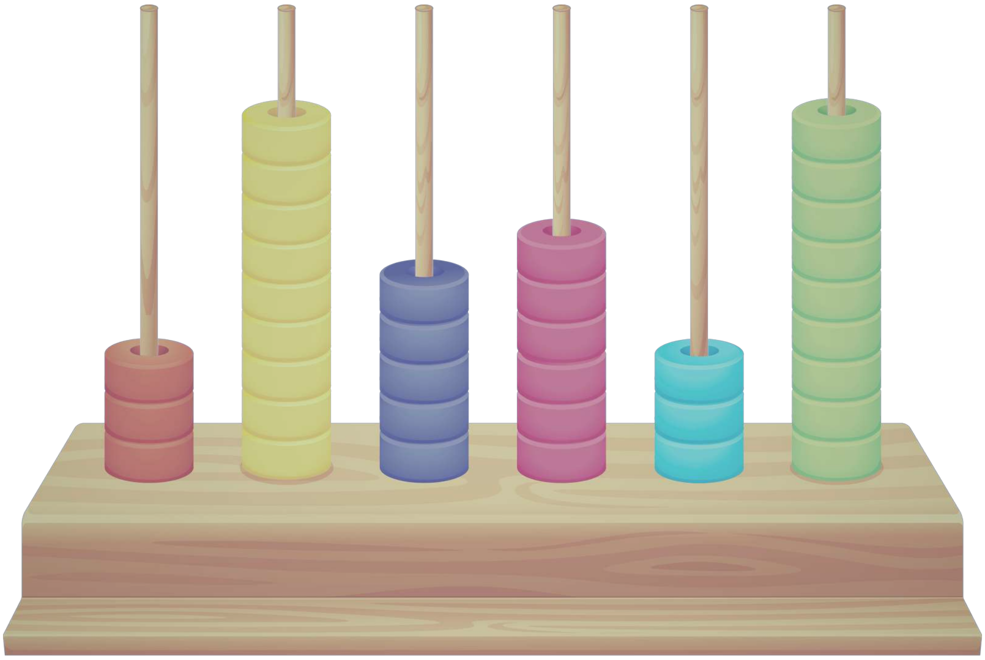
DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Step 1:
Remove irrelevant data

Take a good look at your data and get an idea of what is relevant and what you may not need. Filter out data or observations that aren't relevant to your downstream needs.

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Step 2:
Deduplicate your data



Duplicate records slow down analysis. Even more importantly, if you train a machine learning model on a dataset with duplicate results, the model will likely give more weight to the duplicates thus generating an incorrect model.

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Step 3:
Fix structural errors

Structural errors include things like misspellings, incongruent naming conventions, incorrect word use, etc. These can affect analysis because, while they may be obvious to humans, most machine learning applications wouldn't recognize the mistakes and your analyses would be skewed.

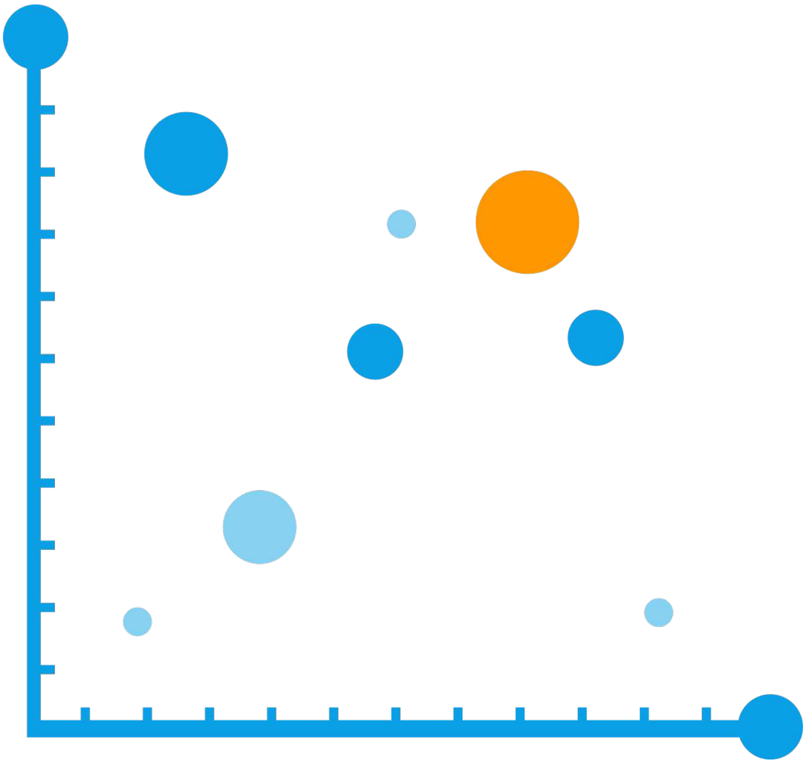
Step 4:
Deal with missing data

Scan your data to locate missing cells, blank spaces in text, etc. You'll need to determine whether everything connected to this missing data (an entire column or row, a whole survey, etc.) should be completely discarded, individual cells entered manually, or left as is.

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

Step 5:
Filter out data outliers



Outliers are data points that fall far outside of the norm and may skew your analysis too far in a certain direction. You'll have to consider what kind of analysis you're running and what effect removing or keeping an outlier will have on your results.

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Step 6:
Validate your data

Data validation is the final data cleaning technique used to authenticate your data and confirm that it’s high quality, consistent and properly formatted for downstream processes. Validate that your data is regularly structured and sufficiently clean for your needs.

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

info@dataranch.org

linkedin.com/company/dataranch