

Data Cleaning

Computing

The checks worth running on any dataset before analysing it.

In roughly this order — each step can change what the next one finds.

1

Check the shape

Row count, column count, types. Does it match what you expected?

2

Find the missing

How many nulls per column, and whether they are random or patterned.

3

Check ranges

Min, max and distribution per column. Negative ages, dates in 1900, impossible values.

4

Find duplicates

Exact and near-duplicates. Decide which copy is authoritative.

5

Standardise

Dates, units, casing, whitespace, encodings, categories that mean the same thing.

MISSING DATA OPTIONS

Drop the rows — only if they are few and random

Impute — with the median, or a model

Flag it — add an `is_missing` column

Missing not at random is the dangerous case: if the missingness correlates with the outcome, dropping rows biases the result.