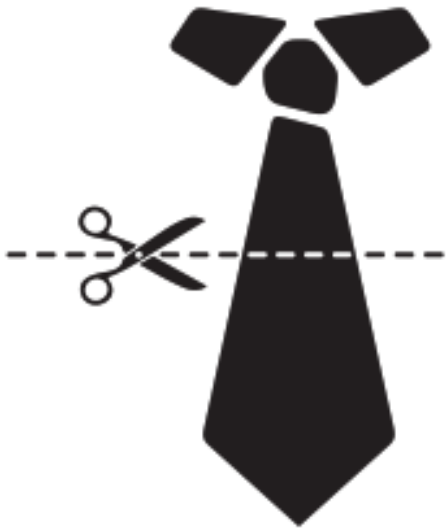
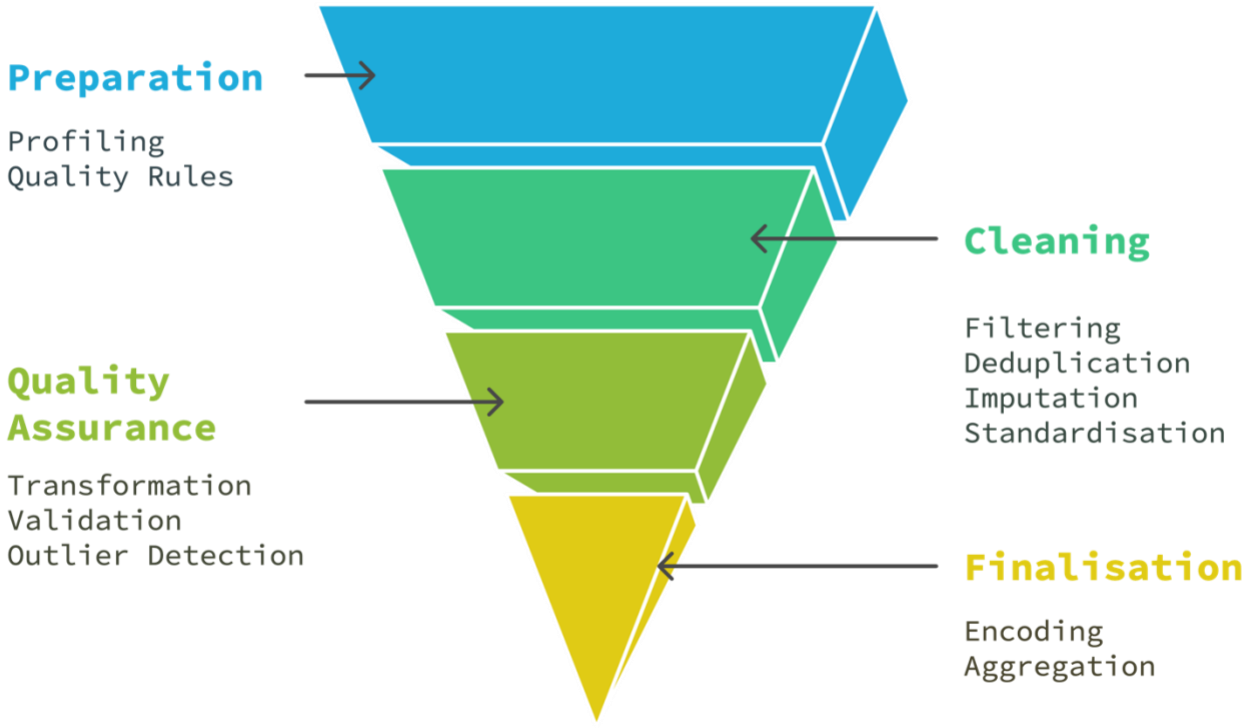


Beginner's Guide to Data Cleaning

Data Cleansing Process Funnel



If you've been asked to "**clean data**" but don't know where to start, don't worry—this guide is for you!

Raw data is often messy, incomplete, and inconsistent. Without proper cleaning, analysis can lead to **wrong conclusions** and **bad decisions**.

This **step-by-step guide** will walk you through the core principles of data cleaning, explain why each step matters, and help you avoid common mistakes.

Process Overview

1. **Preparation**
   Profiling and Quality Rules

2. **Cleaning**
   Filtering, Deduplication, Imputation and Standardisation

3. **Quality Assurance**
   Transformation, Validation and Outlier Detection

4. **Finalisation**
   Encoding and Aggregation

Why This Sequence?

✅ **Prepare before making changes**

✅ **Remove errors first** before restructuring data

✅ **Ensures accuracy** before encoding and aggregation

✅ **Supports iteration** as specific steps **feed back into earlier ones** as needed.

Profiling

Understand the Data Before Cleaning.

You can’t fix what you don’t understand. Before making any changes, analyse the data to see what’s wrong.

Identifies missing values, duplicates, incorrect formats, and unusual distributions before fixing anything.

- Scan the file for missing values, duplicate entries, and formatting inconsistencies.
- Use histograms and box plots to spot extreme values and patterns.
- Excel: Use "Summary Statistics" or Pivot Tables.
- Python: df.info() and df.describe() provide insights into distributions and anomalies.

Quality Rules

Set Clear Standards for “Good” Data.

Cleaning data without clear rules leads to inconsistencies and guesswork.

It prevents confusion by defining acceptable values, formats, and relationships before making changes.

- Define correct formats (e.g., all dates should be YYYY-MM-DD).
- Set rules for valid values (Order Status must be "Open" or "Closed"—not "Oepn").
- Validate relationships (Shipping Date must be after Order Date).

Filtering

Remove Unnecessary or Irrelevant Data.

Cleaning irrelevant data wastes time. Get rid of unnecessary records first.

Reduces file size and complexity, making the next steps faster.

- Remove test data, inactive records, or old transactions.
- Excel: Use filters to hide/remove unnecessary data.
- SQL: Run `DELETE FROM orders WHERE status = 'Inactive'`.

Deduplication

Remove Duplicate Entries.

Duplicates distort results and cause errors in reporting and analysis.

Ensures each row represents a unique entry, avoiding double-counting.

- Check if customers/products appear more than once with slightly different names.
- Excel: Use "Remove Duplicates" under the Data tab.
- SQL: Run SELECT DISTINCT to see unique values before deleting.

Imputation

Fill in Missing Values (But Do It Wisely).

Missing values can cause errors and miscalculations if not handled properly.

Prevents incorrect analysis and improves data completeness.

- Fill missing values using mean, median, mode, or forward fill.
- Use predictive models (e.g., k-nearest neighbours or regression imputation).
- Beware of bias! Imputation can introduce false patterns—always document changes.
- Excel: Use =IF(ISBLANK(A2), "Unknown", A2).
- SQL: Use COALESCE(column, 'Default Value').

Standardisation

Ensure Consistency in Formats.

Without standardisation, data from different sources won’t match.

Prevents issues like misaligned customer names, mismatched currencies, or incorrect date formats.

- Convert all dates to YYYY-MM-DD.
- Standardise names ("UK" vs. "United Kingdom").
- Use consistent units (e.g., all weights in kg, all currencies in USD).

Transformation

Reshape Data for Analysis.

Sometimes, data isn’t stored in a format that makes sense for reporting.

Creates new fields and structures that make data easier to use.

- Create calculated fields (e.g., TotalPrice = Quantity × UnitPrice).
- Break apart or merge fields (e.g., splitting "Full Name" into "First Name" and "Last Name").

Validation

Ensure the Data is Now Correct.

Double-check your work to ensure errors weren't introduced during cleaning.

Confirms that all rules were followed and no mistakes were made.

- Verify that all required fields are filled in.
- Check for incorrect values (e.g., negative prices where they shouldn't exist).
- SQL: SELECT * FROM Orders WHERE ShipDate < OrderDate (to catch logic errors).

Outlier Detection

Identify Extreme or Unusual Values.

Some extreme values are errors, but others might be valid.

Prevents incorrect insights by spotting suspicious values.

- Use box plots and histograms to identify extreme values.
- Apply Interquartile Range (IQR): Q1 - 1.5*IQR and Q3 + 1.5*IQR.
- Consult a domain expert before removing outliers—not all unusual values are errors!

Encoding

Convert Data for Systems or Analysis.

Some tools (like machine learning models) only accept numerical data.

Prepares data for machine learning, reports, and migrations.

- Convert "Yes/No" to 1/0 for analysis.
- Turn categories into numbers (e.g., "Gold Member" = 1, "Silver Member" = 2).
- SQL Example: `CASE WHEN Gender = 'Male' THEN 1 ELSE 0 END`.

Aggregation

Summarise for Reports.

Aggregation is meaningless if raw data isn't clean.

Creates useful reports and dashboards.

- Excel: Use Pivot Tables to group and summarise data.
- SQL: Use `GROUP BY` to summarise by category.

Key Takeaways

✅ **Data Profiling is essential:**
Know what you're working with before making changes.

✅ **Define clear rules:**
Cleaning should follow a plan, not guesswork.

✅ **Cleaning is iterative:**
Sometimes, you'll need to revisit previous steps.

✅ **Think before removing outliers:**
Not all extreme values are errors.

✅ **Validate before using the data:**
Double-check your work before analysis.

Please **SHARE** if you found this helpful!

Click **HERE** for more information.

Isard Haasakker
No Tie Generation Limited