Data Analysis with Spreadsheet Program

Faculty of DS & AI
Spring semester, 2026

Trong-Nghia Nguyen

Business AI Lab

Content

- Import from CSV / Excel / web
- Power Query basics
- Data cleaning
- Handling missing values

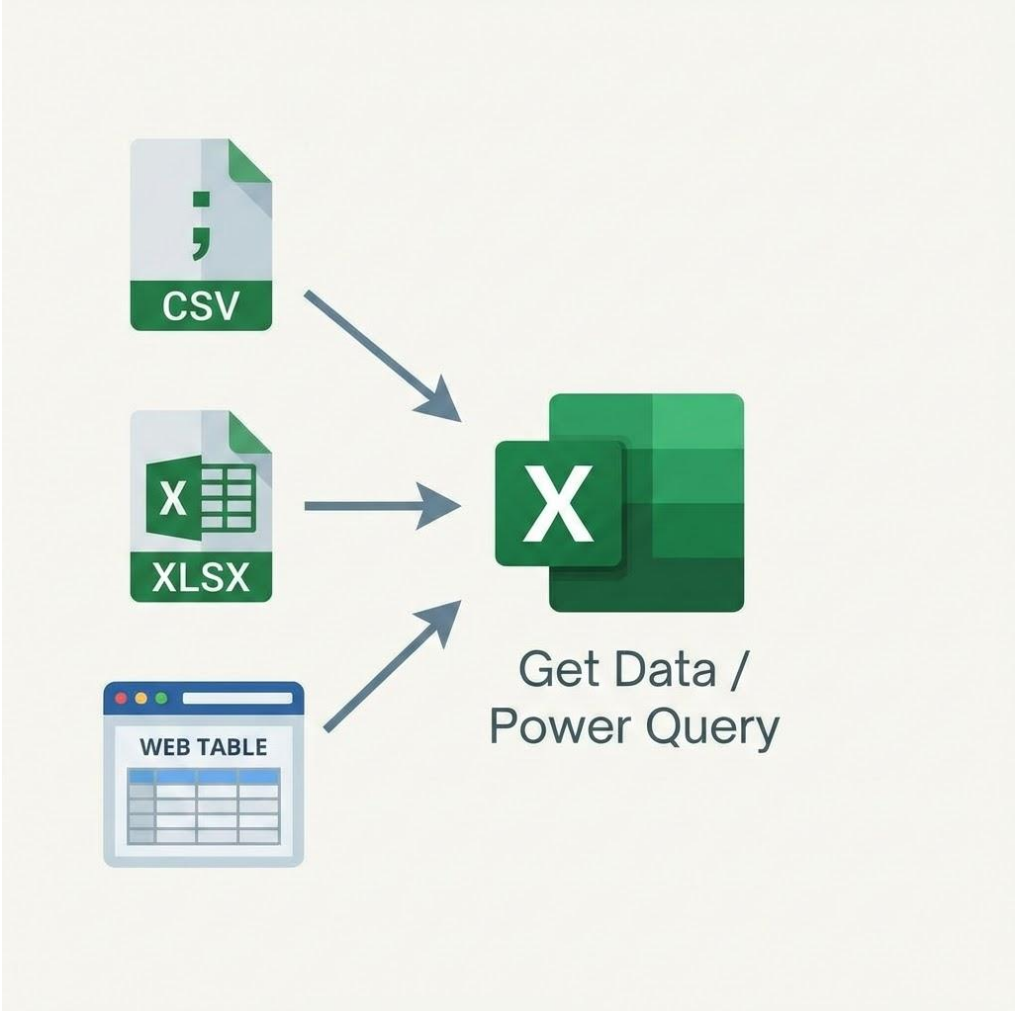
Content

- Import from CSV / Excel / web
- Power Query basics
- Data cleaning
- Handling missing values

Import from CSV / Excel / web

Importing Data into Excel

- Real-world data comes from multiple sources (CSV, Excel files, Web tables)
- Excel provides a unified import tool: Get Data (Power Query)
- Import ≠ Open file
- Goal: Preview → Detect data types → Load reusable connections



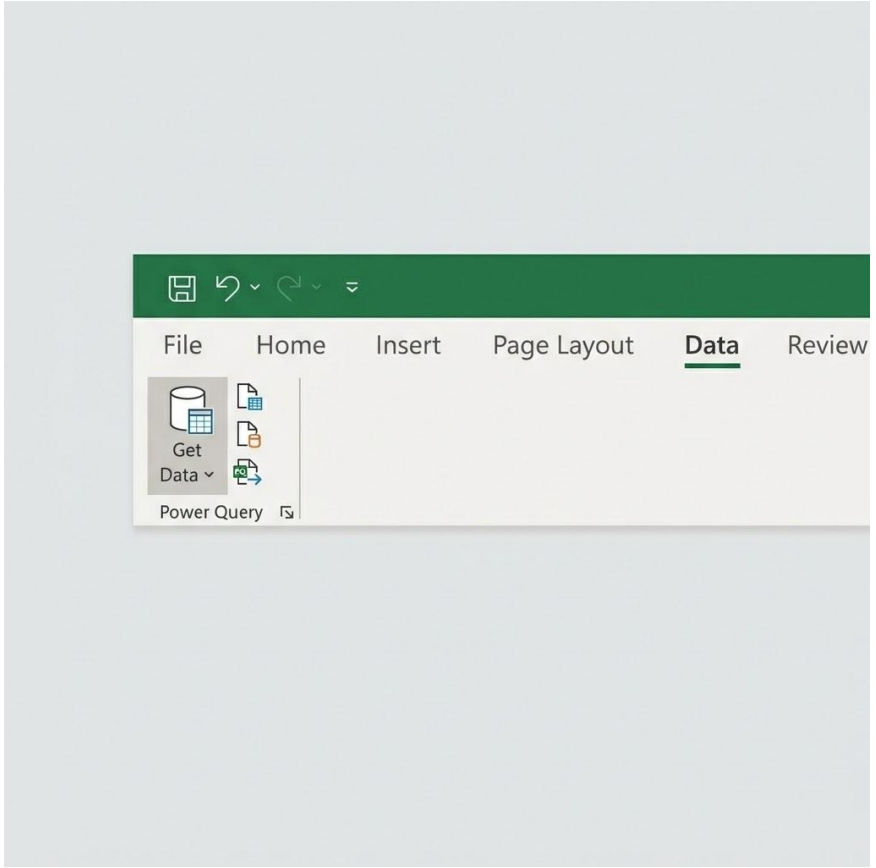
Import from CSV / Excel / web

Import from CSV & Excel Files

- Use Get Data, not “Open file”

- Supported sources:
  - From Text/CSV (CSV, TXT)
  - From Excel Workbook

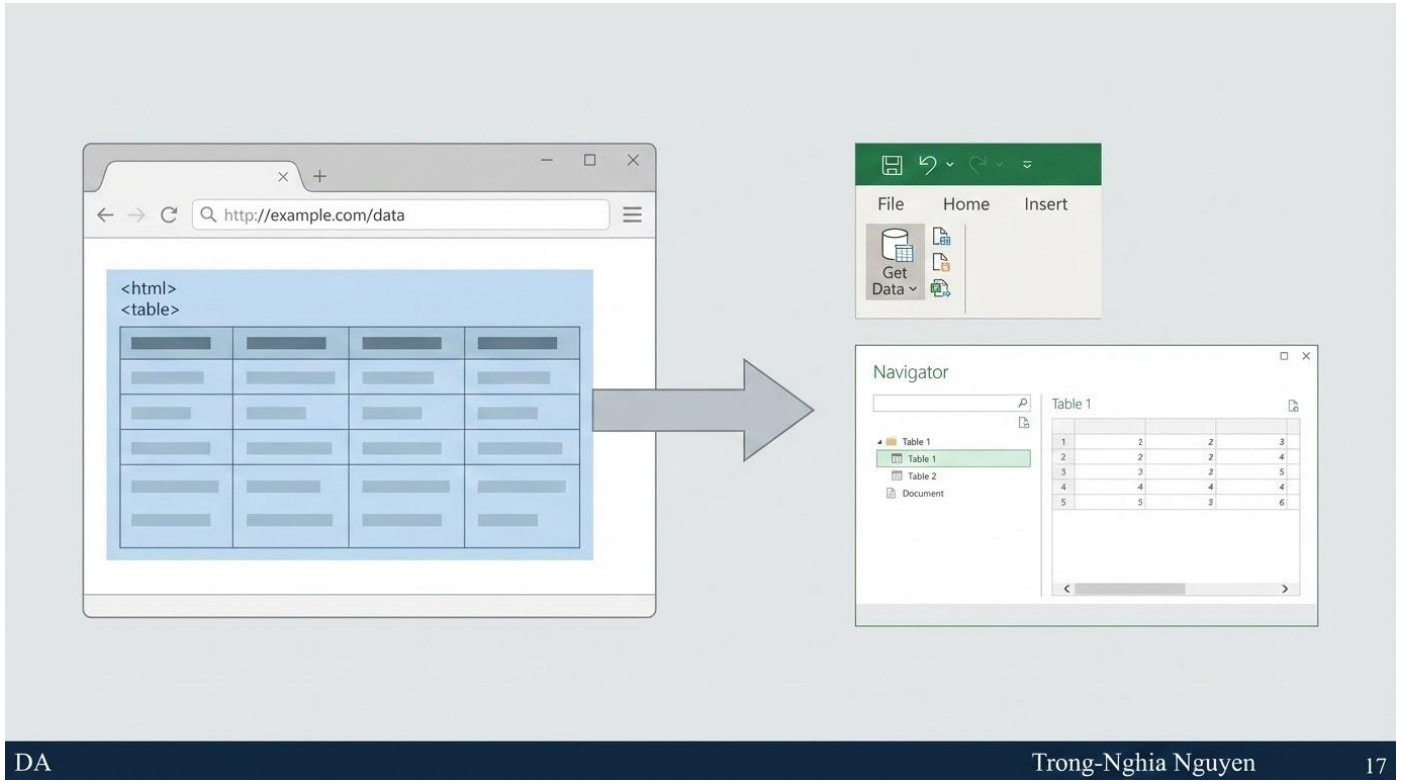
- Power Query workflow:
  - Preview data
  - Detect delimiters & data types
  - Select sheets / tables
  - Load as a reusable query



Import from CSV / Excel / web

Import from Web (Tables & Pages)

- Use Get Data → From Web
- Import structured data from:
  - HTML tables
  - Web pages with tabular data
- Power Query automatically:
  - Detects available tables
  - Previews content before loading
- Data can be refreshed when the web source updates



Import from CSV / Excel / web

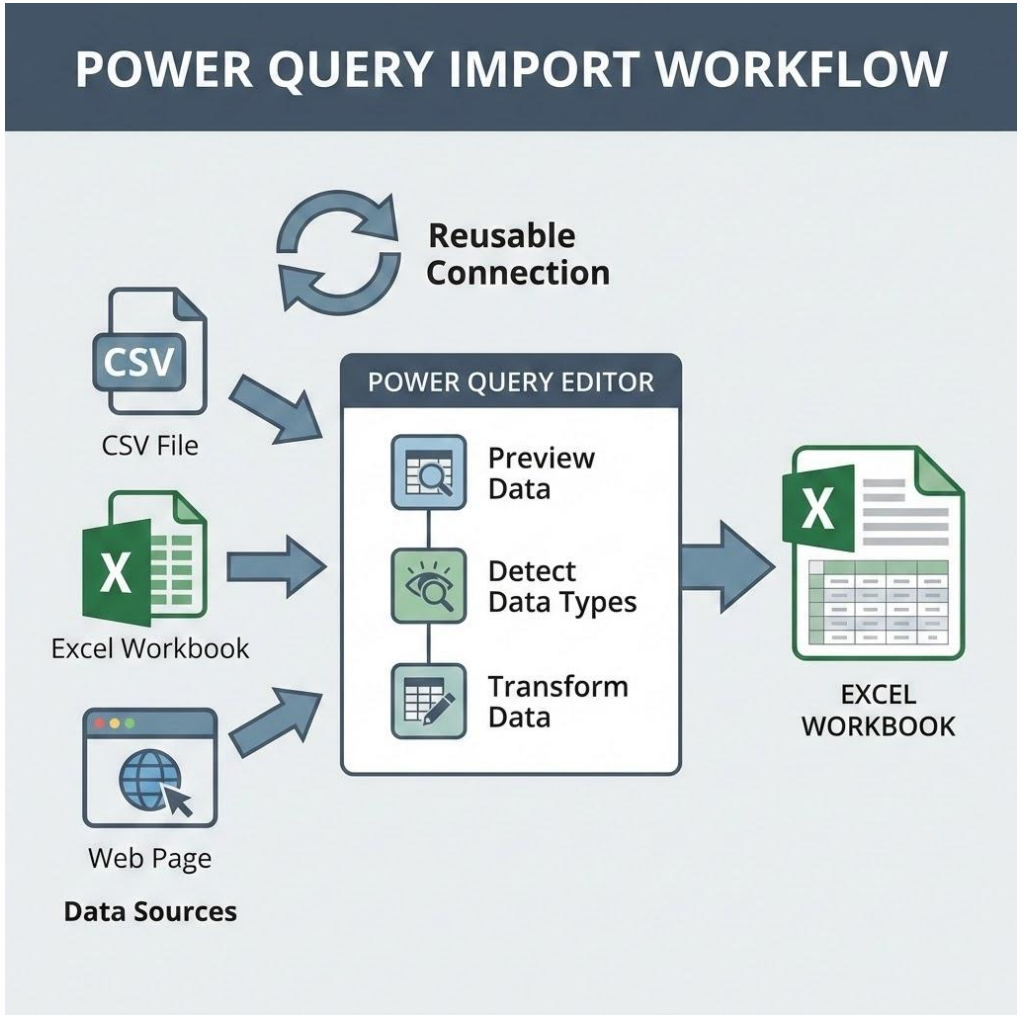
Power Query as the Import Tool

- Power Query is the default engine behind Get Data

- Key features during import:
  - Data preview before loading
  - Automatic data type detection
  - Non-destructive transformations

- Imported data is saved as a **connection**

- Data can be refreshed without re-importing



Import from CSV / Excel / web

Lab

Part A: Import Data

Exercise A1: Import from CSV

Question: Import the Store Sales Orders data from CSV and open the Power Query Editor. What do you see in the preview (column types, obvious issues)?

Instructions:

1. **Excel:** Data → Get Data → From File → From Text/CSV.
2. Select `Lab2-StoreOrders.csv` (from this folder or where you saved it). Click **Transform Data** (not Load) so you open the Power Query Editor before loading.
3. In the Power Query Editor, check the data types shown in the column headers (ABC, 123, calendar icons). Scroll through the table and look for duplicates, blank rows, extra spaces in text, and missing values.

Explanation: Using **Transform Data** instead of **Load** lets you clean the data first. The preview shows how Excel interprets each column; types may be wrong (e.g. dates or numbers as Text) and need fixing.

Import from CSV / Excel / web

Lab

Exercise A2: Import from Excel

Question: Import the same dataset from the Excel workbook and open the Power Query Editor.

Instructions:

1. **Excel:** Data → Get Data → From File → From Excel Workbook.
2. Select `Lab2-StoreOrders.xlsx`, then choose the **Orders** sheet (or table). Click **Transform Data**.
3. Confirm the structure (columns and row count) matches the CSV import.

Explanation: Importing from Excel often preserves more type information than CSV, but you still need to verify. Use one source (CSV or Excel) consistently for Parts B–D.

Import from CSV / Excel / web

Lab

Exercise A3 (optional): Import from Web

Question: Import a small table from a web page using Get Data → From Web.

Instructions:

1. **Excel:** Data → Get Data → From Other Sources → From Web.
2. Enter a URL that contains an HTML table (e.g. a simple reference or demo table). Use **Transform Data** to preview and clean if needed.
3. Document the URL and the steps you used.

https://en.wikipedia.org/wiki/List_of_U.S._state_and_territory_abbreviations

Import from CSV / Excel / web

Lab

Exercise A4: Compare CSV vs Excel Import

Question: If you imported both CSV and Excel, compare the two queries. Do the initial detected data types or row counts differ? Why might they?

Instructions:

1. Create two queries: one from `Lab2-StoreOrders.csv`, one from `Lab2-StoreOrders.xlsx` (both with Transform Data).
2. In each, note the detected types for OrderDate, Amount, and Quantity, and the row count at the bottom.
3. Briefly note any differences.

Explanation: CSV has no type information; Excel infers from cell formats. Excel workbooks can preserve Number/Date formats, so Power Query may detect them more accurately. Row counts should match if the sources are the same.

Content

- Import from CSV / Excel / web
- Power Query basics
- Data cleaning
- Handling missing values

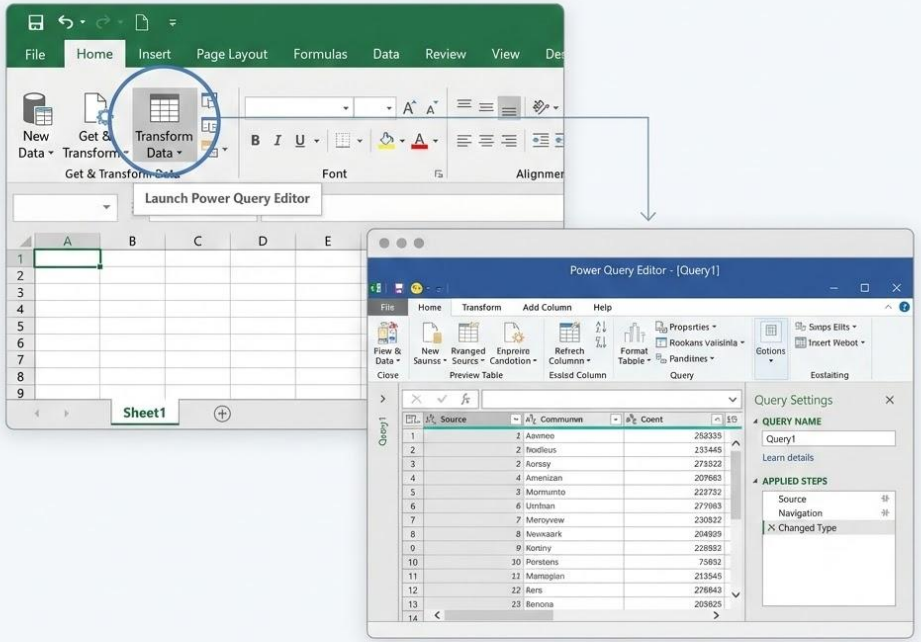
Power Query basics

Opening the Power Query Editor

- After importing data, choose Transform Data / Edit

- Power Query Editor is where:
  - Data is transformed, not manually edited
  - All steps are recorded and reusable

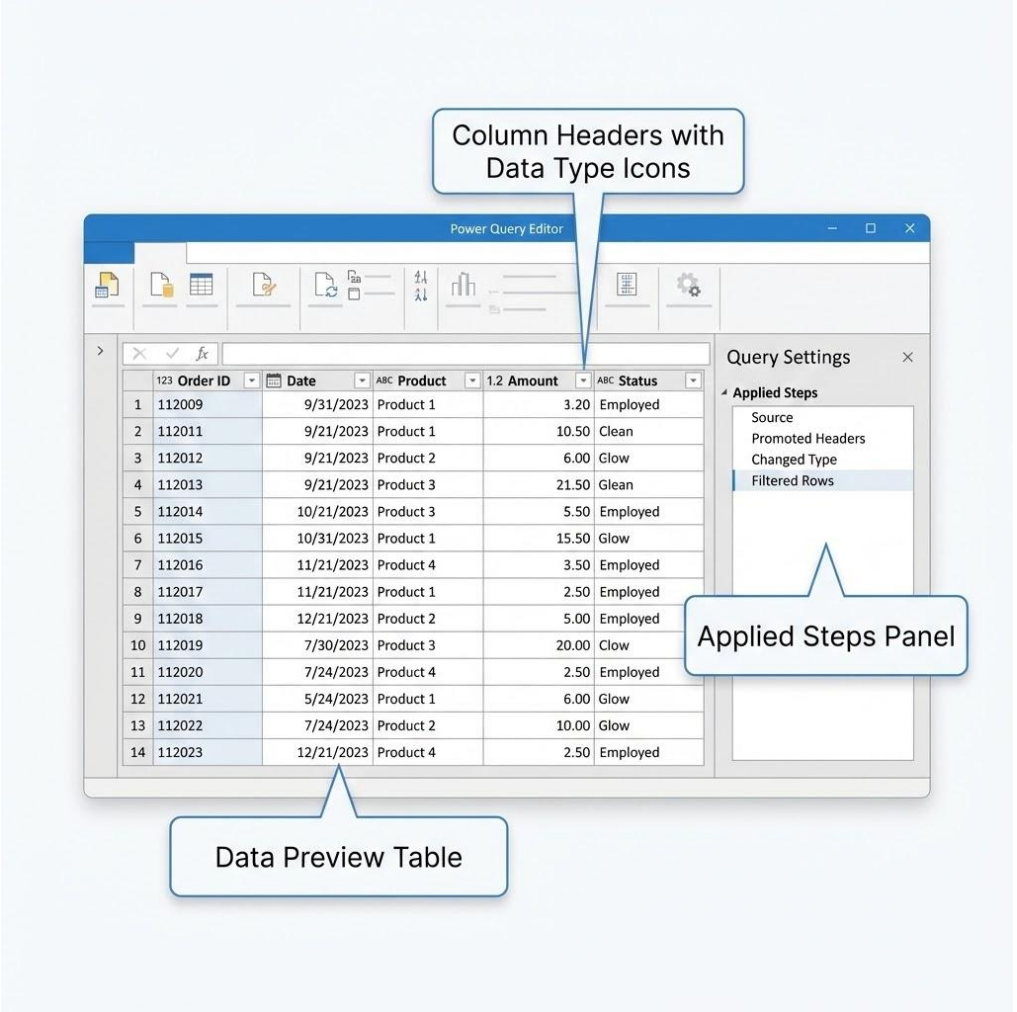
- One query = one data preparation workflow



Power Query basics

Power Query Editor: Key Components

- Data preview (central table view)
- Column headers with detected data types
- Applied Steps panel (all transformations)
- Each action = one recorded step
- Steps can be modified, reordered, or removed

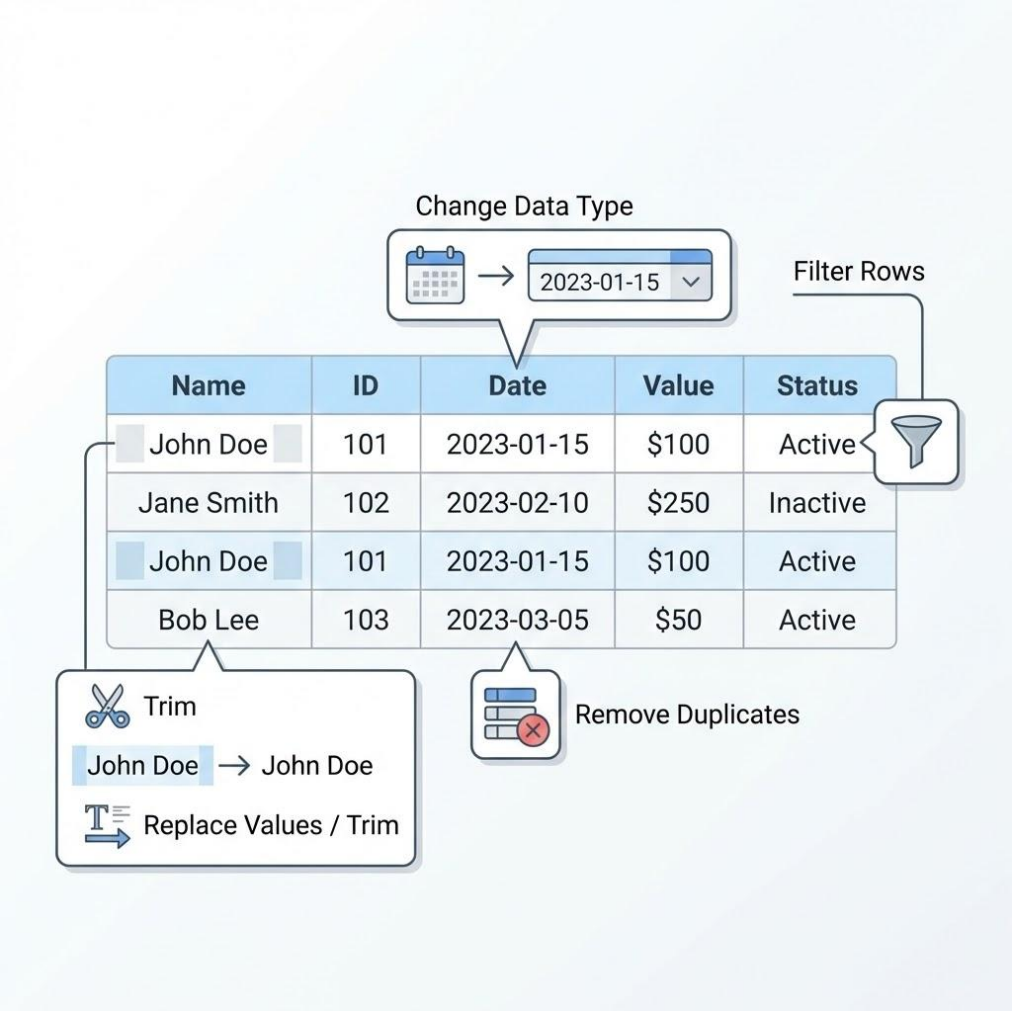


	Order ID	Date	Product	Amount	Status
1	112009	9/31/2023	Product 1	3.20	Employed
2	112011	9/21/2023	Product 1	10.50	Clean
3	112012	9/21/2023	Product 2	6.00	Glow
4	112013	9/21/2023	Product 3	21.50	Glean
5	112014	10/21/2023	Product 3	5.50	Employed
6	112015	10/31/2023	Product 1	15.50	Glow
7	112016	11/21/2023	Product 4	3.50	Employed
8	112017	11/21/2023	Product 1	2.50	Employed
9	112018	12/21/2023	Product 2	5.00	Employed
10	112019	7/30/2023	Product 3	20.00	Clow
11	112020	7/24/2023	Product 4	2.50	Employed
12	112021	5/24/2023	Product 1	6.00	Glow
13	112022	7/24/2023	Product 2	10.00	Glow
14	112023	12/21/2023	Product 4	2.50	Employed

Power Query basics

Core Transformations

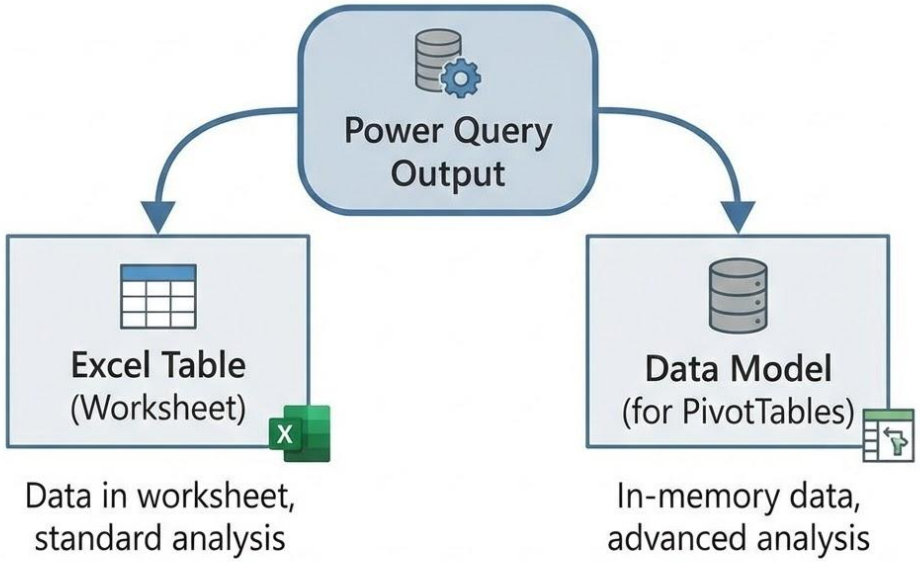
- Change data types (Text, Number, Date)
- Remove duplicates
  - All columns or selected key columns
- Filter rows
  - By values, blanks, or conditions
- Replace values and Trim text



Power Query basics

Load Options in Power Query

- Load data into:
  - Excel Table (Worksheet)
  - Data Model
- Excel Table:
  - Easy to view and explore
  - Suitable for small to medium datasets
- Data Model:
  - Handles larger datasets
  - Used for PivotTables and advanced analysis



Power Query basics

Lab

Exercise B1: Change Data Types

Question: Set correct data types for OrderDate, Amount, and Quantity.

Instructions:

1. Select the **OrderDate** column. Transform → Data Type → Date (or Date/Time). Choose the correct locale if prompted (e.g. day/month/year vs month/day/year).
2. Select **Amount** → Transform → Data Type → Decimal Number (or Currency if you prefer).
3. Select **Quantity** → Transform → Data Type → Whole Number.
4. Check for any columns still detected as Text that should be numeric or date; fix them.

Explanation: Correct types ensure filtering, sorting, and calculations work properly. Wrong types (e.g. numbers as text) cause incorrect sums or sorts.

Exercise B2: Remove Duplicates

Question: Remove duplicate rows and note how many rows were removed.

Instructions:

1. Select all columns (or the key columns you use for duplicates, e.g. OrderID).
2. Home → Remove Rows → Remove Duplicates.
3. Check the row count before and after; record the number of rows removed.

Explanation: Duplicates inflate counts and can bias analysis. Remove them based on a meaningful key (e.g. OrderID) or on all columns, depending on your definition of a duplicate.

Power Query basics

Lab

Exercise B3: Filter Rows

Question: Use the column filter dropdowns to inspect or exclude certain values (e.g. blanks or a specific Category/Region).

Instructions:

1. Click the filter icon on a column header (e.g. Category or Region).
2. Use Text Filters or "Remove Empty" (or equivalent) to filter out blanks if needed, or filter to a single category for inspection.
3. Remove or adjust the filter as needed for later steps (e.g. you may remove blank rows via Home → Remove Rows → Remove Blank Rows instead).

Explanation: Filtering helps you explore subsets and confirm where blanks or odd values appear. Use it before standardizing categories or removing blank rows.

Exercise B4: Replace Values

Question: Standardize the Category column (e.g. replace "ELECTRONICS", " electronics " with "Electronics").

Instructions:

1. Select the **Category** column. Transform → Replace Values.
2. Replace each variant (e.g. "ELECTRONICS", " electronics ") with the standard form "Electronics". Repeat for other categories (Clothing, Food, Books, Sports) as needed.
3. Alternatively, use a Conditional Column or multiple Replace Values steps.

Explanation: Inconsistent categories split groups in PivotTables and reports. Standardizing them ensures correct aggregation and filters.

Power Query basics

Lab

Exercise B5: Trim and Clean

Question: Apply Trim to Customer, Category, and Region; use Clean if you encounter non-printable characters.

Instructions:

1. Select **Customer**. Transform → Format → Trim. Repeat for **Category** and **Region**.
2. If any column has non-printable characters (e.g. from imported web or text data), use Transform → Format → Clean on that column.

Explanation: Trim removes leading and trailing spaces; Clean removes non-printable characters (e.g. tab, vertical tab). Both improve matching and grouping.

Exercise B6: Duplicates by Key vs All Columns

Question: Try removing duplicates first by **OrderID only**, then (in a new query or after undoing) by **all columns**. Compare the number of rows removed in each case. When would you use each approach?

Instructions:

1. Remove duplicates based only on **OrderID**. Note the row count before and after.
2. Start again from the source (or use a duplicate of the query) and remove duplicates based on **all columns**. Note the row count before and after.
3. Compare and briefly explain when to use key-only vs all-column duplicate removal.

Explanation: Key-only (e.g. OrderID) treats two rows as duplicates if the key is the same, even if other columns differ. All-column duplicate removal keeps only rows that are identical in every column. Use key-only when the key uniquely identifies a record; use all-column when you want to drop only exact duplicates.

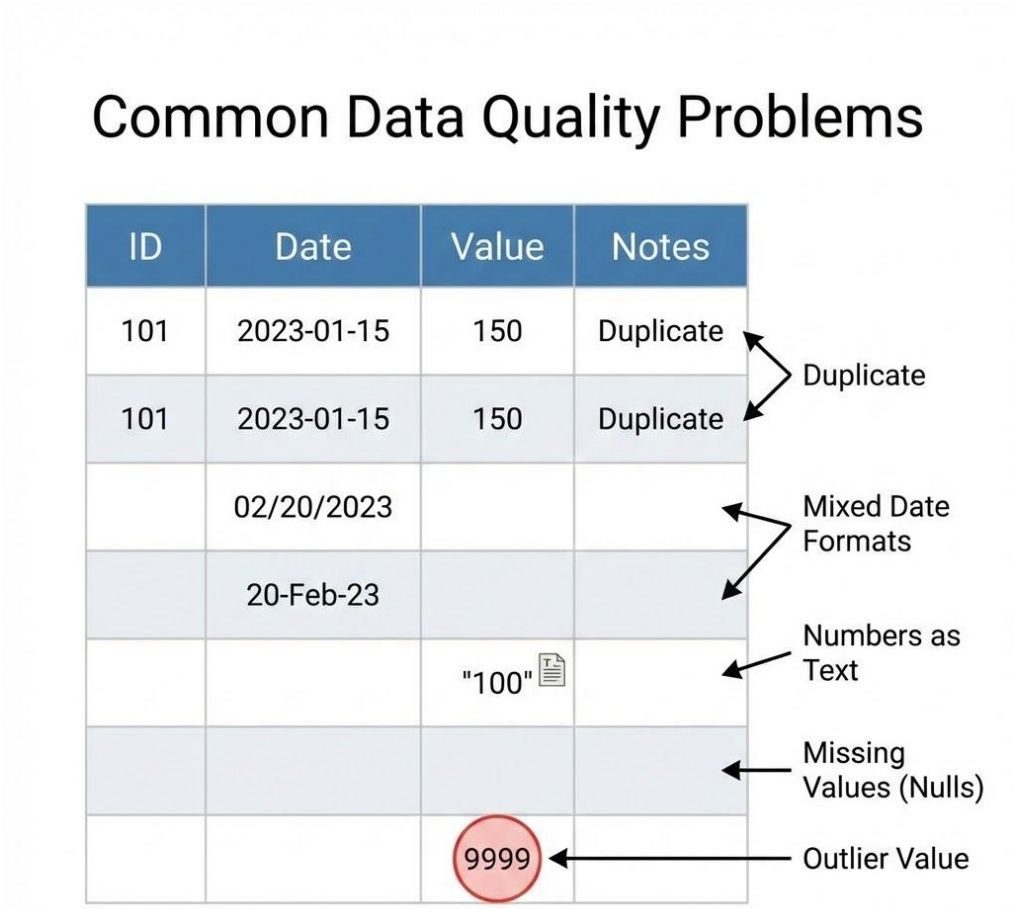
Content

- Import from CSV / Excel / web
- Power Query basics
- Data cleaning
- Handling missing values

Data cleaning

Common Data Quality Issues

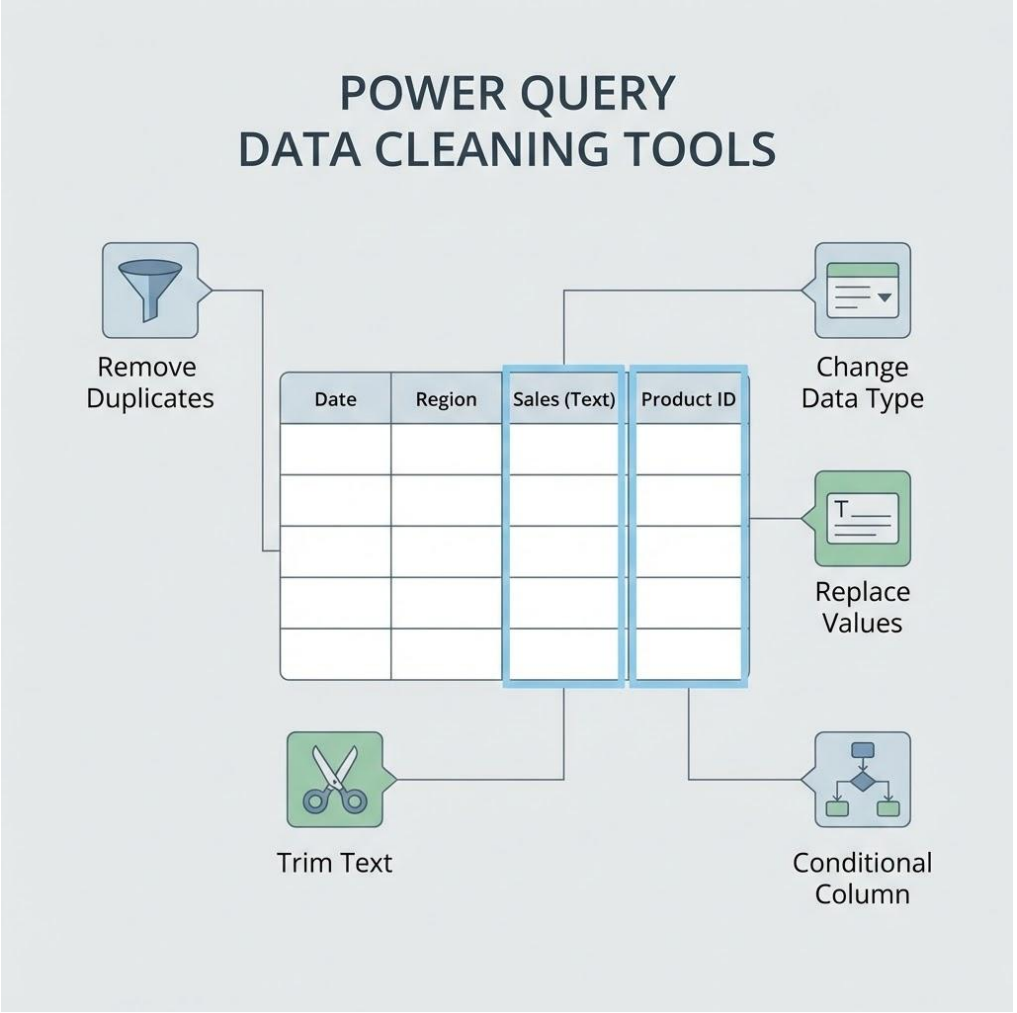
- Duplicate records
- Inconsistent formats
  (dates, numbers as text, regional settings)
- Wrong data types
- Missing values
- Obvious outliers or data entry errors



Data cleaning

Data Cleaning Tools

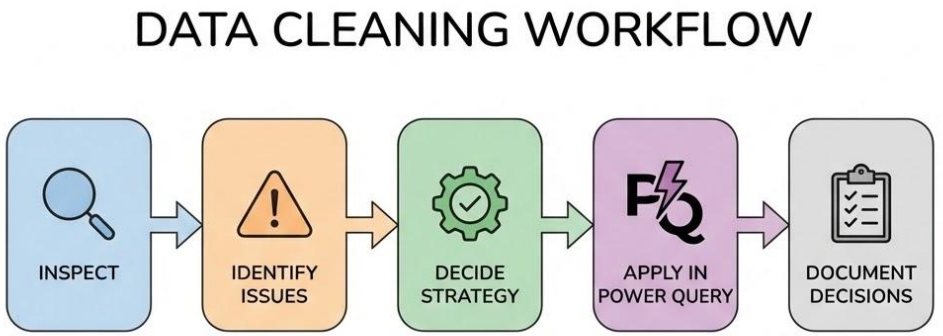
- Remove Duplicates
- Change Data Type
- Replace Values
- Trim / Clean text
- Create Conditional Columns (when needed)



Data cleaning

Data Cleaning Workflow

- Inspect the data
- Identify data quality issues
- Decide an appropriate strategy
- Apply steps in Power Query
- Document cleaning decisions



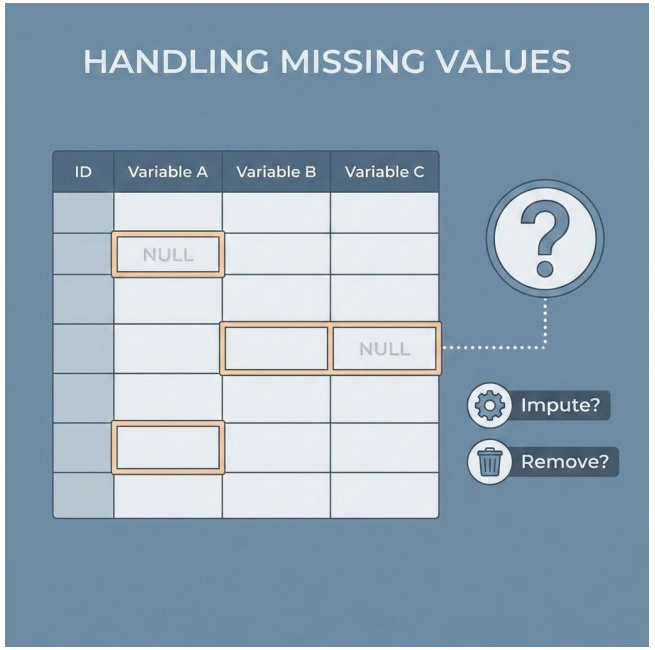
Content

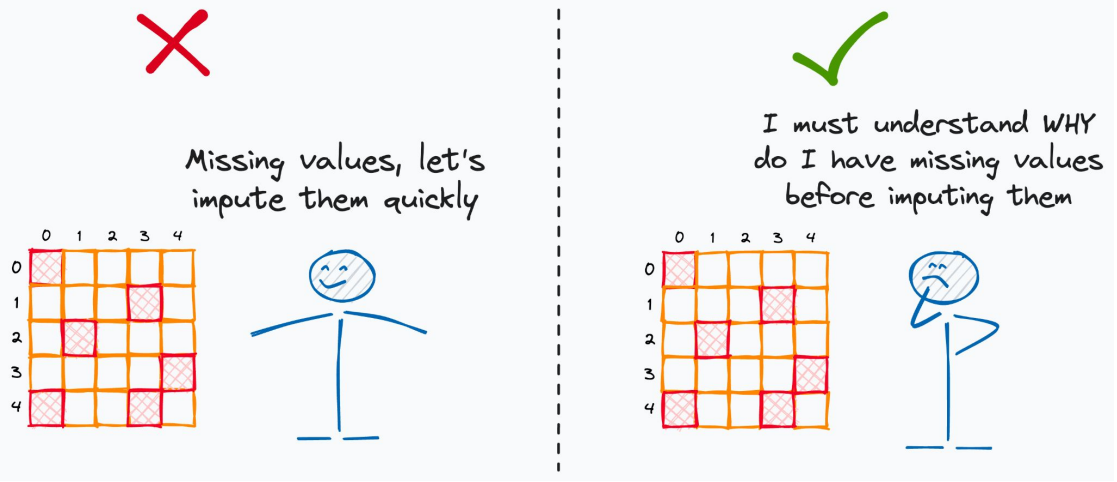
- Import from CSV / Excel / web
- Power Query basics
- Data cleaning
- Handling missing values

Handling missing values

Missing Values

- Missing values are common in real-world data
- Missing ≠ error by default
- There is no single correct rule
- Handling depends on:
  - Amount of missing data
  - Pattern of missingness
  - Role of the variable in the analysis

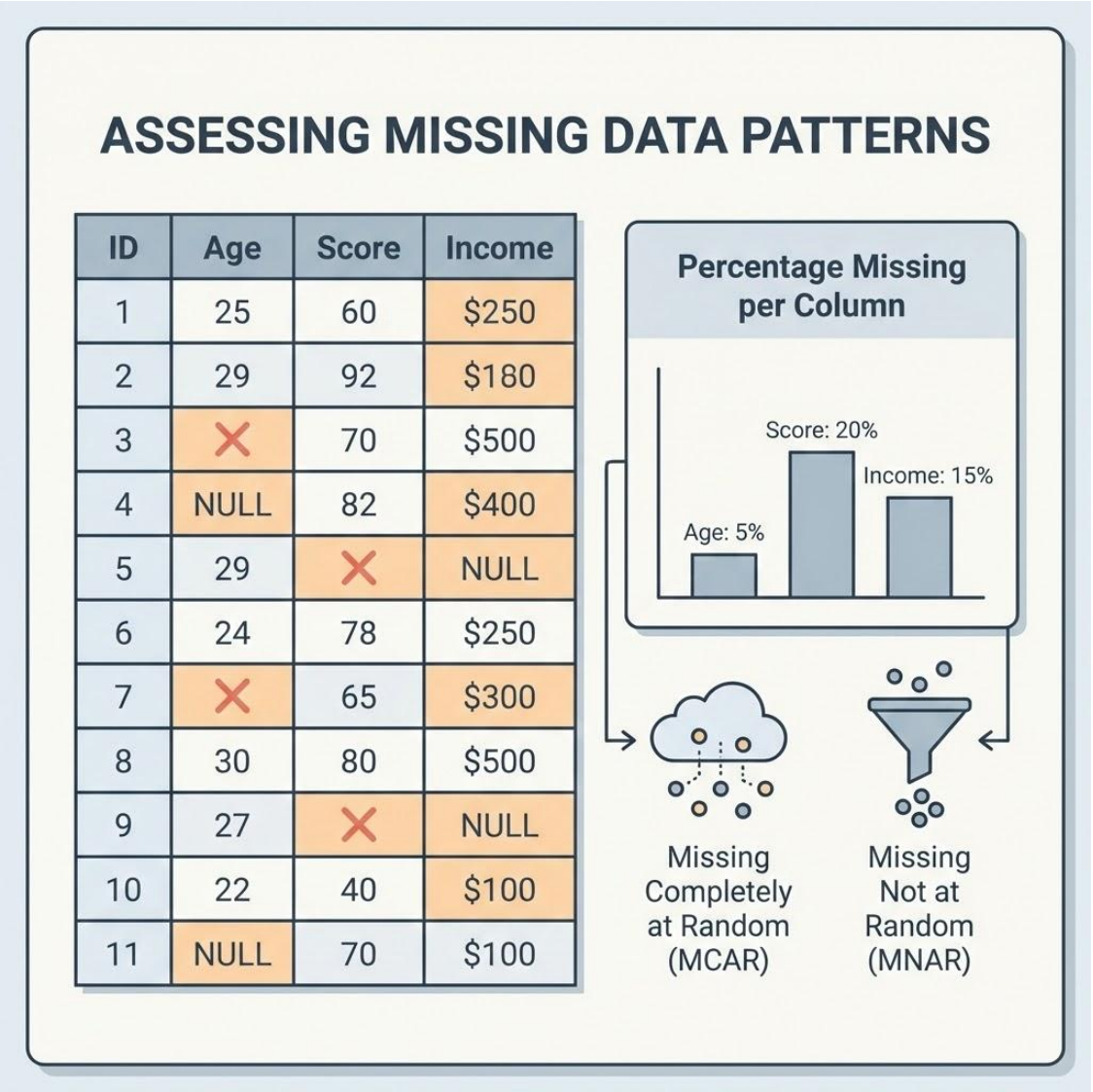




Handling missing values

Assessing Missing Data

- How much data is missing?
  - Percentage per column
- Is missingness:
  - Random?
  - Systematic (specific groups, time periods)?
- Does the variable matter for the analysis?

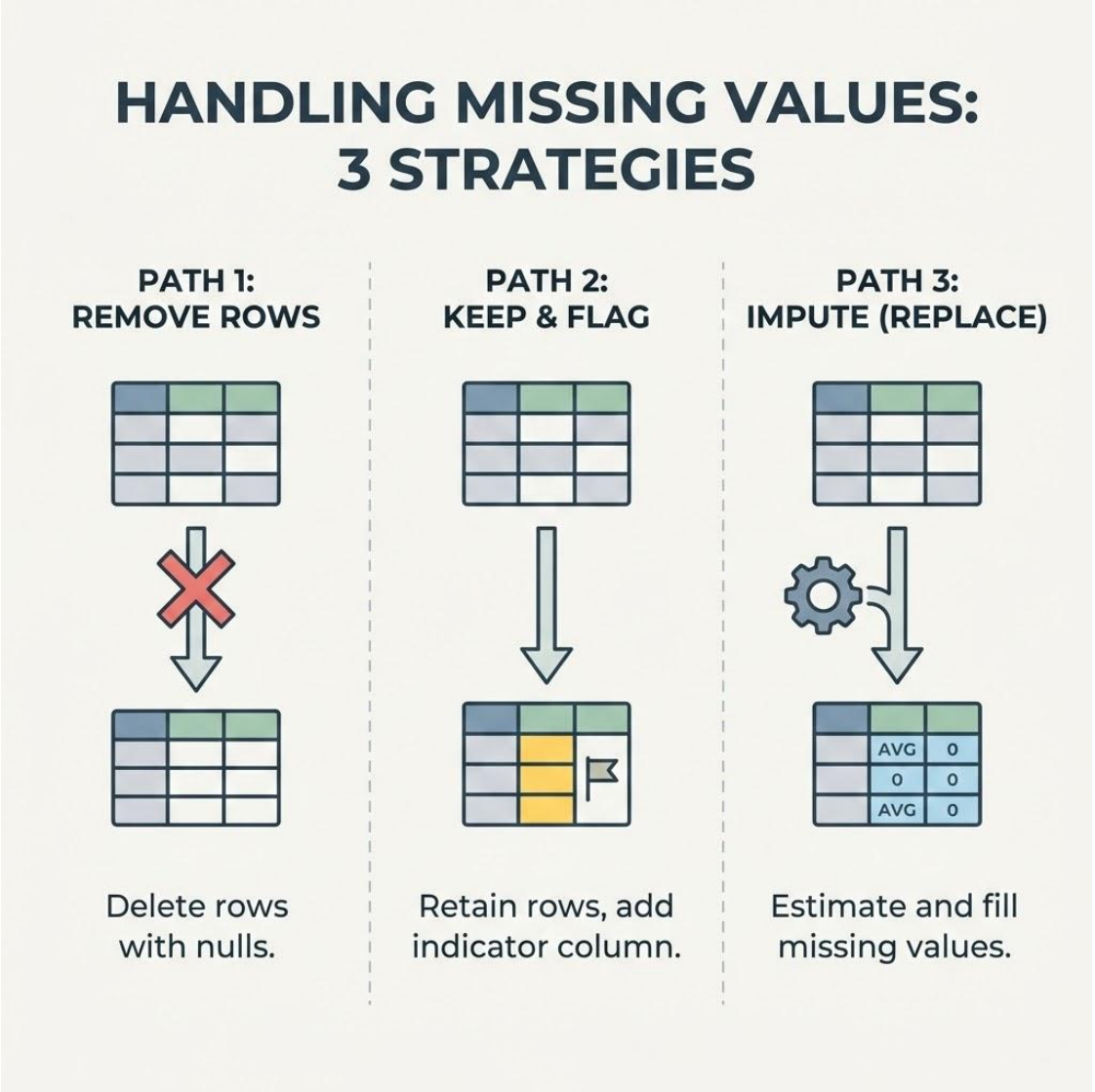


Handling missing values

Strategies for Handling Missing Values

- Remove rows (listwise deletion)
    - When missing is small and random
- Keep rows and flag missing values:
    - Preserve information
- Simple imputation:
    - Mean / median / mode (when appropriate)

Always state and justify the chosen strategy



Handling missing values

Common Mistakes & Key Takeaways

- Common mistakes:
  - Blindly deleting rows with missing values
  - Filling missing values without understanding the data
  - Ignoring missing data patterns

- Key takeaways:
  - Always explore missing values first
  - Understand why data is missing
  - Choose a strategy that fits the data and the task

Handling missing values

Lab

Exercise C1: Remove Blank Rows

Question: Remove fully blank rows and note the change in row count.

Instructions:

1. Home → Remove Rows → Remove Blank Rows.
2. Compare the row count before and after.

Explanation: Blank rows can break PivotTables and formulas. Remove them when they add no information.

Exercise C2: Flag Missing Notes

Question: Add a conditional column that flags rows where Notes is null or blank, without deleting those rows.

Instructions:

1. Add Column → Conditional Column.
2. Set **Column name** (e.g. `Notes_Missing`). If **Notes** is null or equals `""`, output `Yes` (or `True`); otherwise `No` (or `False`).
3. Load the query and confirm the new column appears.

Explanation: Flagging missing values preserves all rows for analysis while making it easy to filter or report on missingness.

Handling missing values

Lab

Exercise C3: Document Your Strategy

Question: In one or two sentences, state how you handled missing values in Region and Notes (e.g. removed blank rows, flagged missing, left as-is).

Exercise C4: When to Impute, Flag, or Delete

Question: Briefly explain when you would (a) delete rows with missing values, (b) flag missing values with a new column, or (c) impute (e.g. fill with mean/median/mode). Give one example each.

Instructions:

1. Consider amount of missing data, whether it's random or systematic, and how the variable is used.
2. Write 1–2 sentences for each of (a), (b), and (c) with a concrete example.

Explanation: Deletion is simple but loses data; use when missing is small and random. Flagging keeps all rows and supports analysis of missingness. Imputation preserves sample size but can bias results if done poorly; use when justified and documented.

Handling missing values

Lab

Part D: Load and Refresh

Exercise D1: Load to Worksheet

Question: Load the cleaned data into an Excel table in a new sheet.

Instructions:

1. In the Power Query Editor, **Close & Load**.
2. Ensure the query has a clear name (e.g. `StoreOrders_Cleaned`). Rename it in Queries & Connections if needed.

Exercise D2 (optional): Load to Data Model

Question: Create a connection that loads the data into the Data Model only (no worksheet).

Instructions:

1. Duplicate the query or create a new one from the same source. **Close & Load To...**
2. Choose **Only Create Connection** and check **Add this data to the Data Model**.
3. Confirm in Data → Queries & Connections that the query exists and is connected to the Data Model.

Handling missing values

Lab

Exercise D3: Refresh Data

Question: Refresh the query after changing the source file and confirm the change appears.

Instructions:

1. Close the workbook or ensure the source CSV/Excel is not locked. Add a new row to the source file (or change an existing value), then save.
2. In Excel, Data → Refresh All (or right-click the query → Refresh).
3. Verify that the updated data appears in the loaded table.

Exercise D4: Connection and Refresh Settings

Question: Where can you change the source file path or refresh settings for a query? What happens if you move or rename the source file?

Instructions:

1. Open Data → Queries & Connections. Right-click the query → **Edit** or **Properties** (depending on your Excel version).
2. Look for **Source** or **Connection** settings. Note how the path is stored.
3. Move or rename the source file, then try Refresh. Observe the result.

Explanation: The connection stores the full path to the source. If you move or rename the file, Refresh fails until you update the path (e.g. via Data Source Settings or by editing the query and changing the source step).

Thank you!