Excel Data Cleaning Handbook

Beginner to Advanced with Real Examples

Author: ALU DATATOK

1. Introduction

Data cleaning is one of the most critical steps in data analysis. It ensures accuracy, reliability, and usability of data before insights are drawn. This handbook provides practical, step-by-step approaches to cleaning numerical, textual, and date/time data in Excel. Each section includes detailed examples, formulas, and a quick reference cheat sheet.

Workflow for Data Cleaning: Identify → Diagnose → Clean → Validate

2. Numerical Data Cleaning

Beginner

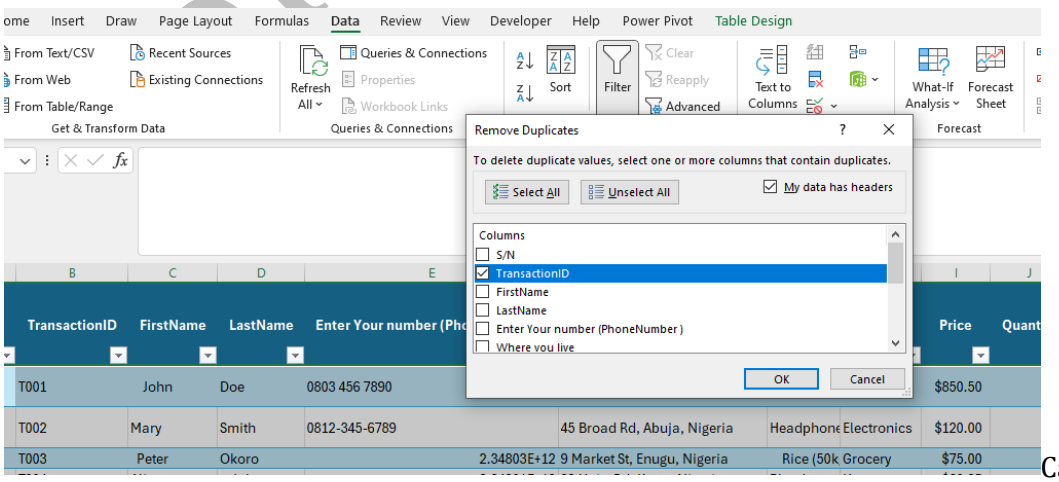
Case 1: Manually Scan the data and understand it

Case 2: Delete rows/unnecessary columns

Case 3: Removing Duplicates

Problem: A table has duplicates across rows identified with repeated ID numbers.

Steps: Select the table→ Data → Remove Duplicates →Unselect all →select unique identifier →check the box 'my table has headers' → click 'ok'



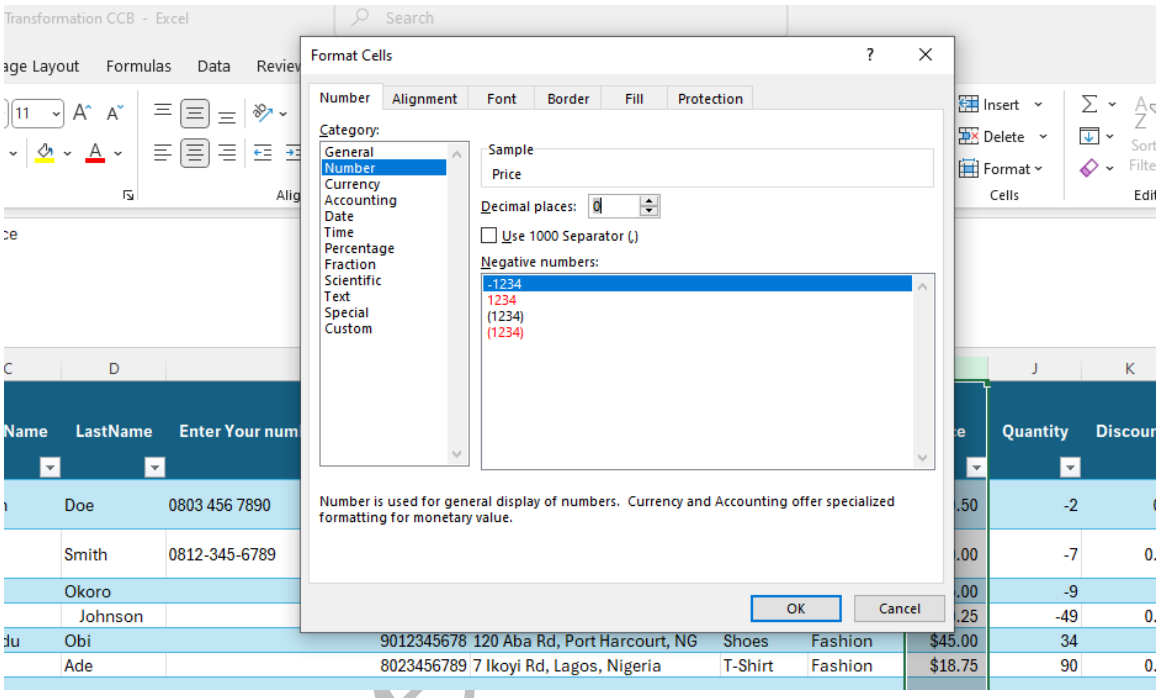
Case 2:

Case 4: Converting Text to Numbers

Problem: Numbers are stored as text (green triangle error).

Steps: Select the cells → ctrl + 1 → format to Number.

Formula: =VALUE(A2)



Case 5: Cleaning the phone number

Problem: Spaces in numbers

Trim, Ctrl + H to find and replace values.

Problem: Standardize the numbers to start with +

=IF(LEFT(E2, 3) = "234", "+" & E2, IF(LEFT(E2, 1) = "0", "+234" & RIGHT(E2, LEN(E2) - 1), "+234" & E2))

Meaning in words:

1. Look at the value in **cell E2** (a phone number).
2. **If the first 3 characters are "234":**
   → Add a **"+"** sign in front.
   Example: 2348030001111 → +2348030001111

3. **Else, if the first character is "0":**
   → Replace the "0" with **"+234"**.
   Example: 08030001111 → +2348030001111

4. **Else (if it does not start with 234 or 0):**
   → Just add **"+234"** in front of the number.
   Example: 8030001111 → +2348030001111

Case 6: Handling blank cells (replace with 0, Unknown, or mean/median).

Wrong moves: Deleting missing values. This is one of the most controversial topics in the DA industry. Some people will say if missing values are less, you should delete them. But in real life you are just losing data. Please understand why these values are missing. For instance, in the case of promo, the amount for the giveaway will not be written. **So just replace with zero.**

In other cases of say phone number missing, use 'Unknown'. Values that are non-aggregatable

For numbers that can be aggregated use mean or median.

Use mean when there is an outlier. And either of them when there is no outlier.

Why: Outliers skew the numbers either to the extreme left or extreme right. And makes your value unrealistic.

How to spot outliers:

1. *Manually searching through*
2. *If your mean is far off from the median, then using conditional formatting to spot through*
3. *Using min and max: if the mean is somewhat in the middle, then yes, no outlier.*

Case 7: Handling negative numbers where not necessary like **quantity: Use 'ABS' or =IF(I2 < 0, I2 * -1, I2)**

Case 8: Calculated fields

1. Total Cost per Transaction: =IFERROR(Price * Quantity, 0)
2. Total Discount applied: =IFERROR(Quantity * Discount, 0)
3. Discount in price: =IFERROR(Price * Discount, 0)
4. Final Price: =IFERROR((Price * Quantity) - (Price * Discount), 0)
5. Bulky Order Indicator: =IF(Quantity > 50, "Bulky", "Regular")
6. Discount Applied?=IF(Discount>0, "Yes", "No")

Intermediate

Case 3: Identifying Outliers

Problem: Spot unusual values in a sales dataset.

Steps: Use Conditional Formatting → Highlight Cells → Greater than Average ± 2*STDEV.

Formula: =IF(ABS(A2-AVERAGE($A$2:$A$100))>2*STDEV($A$2:$A$100),"Outlier","OK")

[Insert Screenshot Here]

3. Textual Data Cleaning

Beginner

Case 1: Removing Extra Spaces

Formula: =TRIM(A2)

Case 2: Changing Text Case

UPPER: =UPPER(A2)
LOWER: =LOWER(A2)
Proper Case: =PROPER(A2)

Case 3: Proper and descriptive headers

Intermediate

Case 4: Merge columns: Concat / &

Split columns: Text-to-columns : Address

4. Date & Time Data Cleaning

Beginner

Case 1: Converting Text Dates to Real Dates

Steps: Select column → Data → Text to Columns → Finish.

Formula: =DATEVALUE(A2)

Filter and correct the remaining inconsistencies manually

Intermediate

Case 2: Extracting Date Parts

DAY: =DAY(A2)
MONTH: =MONTH(A2)
YEAR: =YEAR(A2)
WEEKDAY: =TEXT(A2,"dddd")

4. PRACTICE PRACTICE

ID	Name	Age	Date	Amount
1	john doe	18	12/3/2025	1230USD
2	Jane		1/4/2025	$12345
3	JON DOE	25		
4	john d.	20		$12367

Step-by-step cleaning pipeline:

1. Standardize names (PROPER, lookup).
2. Fix ages.
3. Correct dates.
4. Clean 'USD' and convert to numbers.
5. Remove duplicates.

Final Clean Dataset Example:

6. Cheat Sheet (Quick Reference)

Task	Formula	Example Use
Remove spaces	=TRIM(A2)	Fix names with extra spaces
Extract year	=YEAR(A2)	Get year from date
Handle errors	=IFERROR(A2/B2,"")	Replace error with blank
Convert to number	=VALUE(A2)	Convert '100' stored as text

Proper Case	=PROPER(A2)	Standardize names

7. Practice Exercises

1. Remove duplicates from a list of product IDs.
2. Convert '200USD', '500USD' into numbers.
3. Standardize names with inconsistent capitalization.
4. Fix wrong date formats (e.g., 12/11/25 vs 11/12/25).
5. Identify outliers in sales data.
6. Replace blank ages with average.
7. Extract first names from a full name column.
8. Remove trailing spaces from employee IDs.
9. Flag sales records from the last 30 days.
10. Normalize a set of salary values between 0 and 1.