

# Module 6

## AI-Assisted Data Cleaning

Transforming Chaos into Analysis-Ready Insights

---



# The Messy Spreadsheet Problem

Identifying the bottlenecks that stall data science projects before they even begin.

# | Common Data Culprits



**Missing Values:** Detecting "NaN" or empty strings and deciding on removal vs. imputation.



**Duplicate Entries:** Spotting identical records across multiple customer ID columns.



**Incorrect Dates:** Reconciling DD/MM/YY with YYYY-MM-DD for time-series consistency.



**String Noise:** Stripping trailing spaces, fixing case sensitivity, and removing junk symbols.

# | The Janitorial Reality

80%

Time Spent Cleaning

## The Data Science Bottleneck

Industry surveys consistently show that 4/5ths of a data professional's time is consumed by preparation. Poor data quality costs organizations billions in missed opportunities and technical debt.



# Prompting for Cleanup

Transitioning from manual find-and-replace to AI-architected  
automation.

# | AI as a Code Architect

## Strategic Scripting

AI doesn't just "fix" your data; it writes the instructions to fix it. By defining a **Target Schema**, you can prompt an LLM to generate Python code that:

- </> Implements **Pandas** logic for bulk updates.
- 🔹 Applies **IQR filters** to detect extreme outliers.
- ✓✓ Performs **Fuzzy Matching** for deduplication.



# | Why Python Beats Excel



## Reproducibility

Scripts are a "receipt" of every change. Run the same cleaning pipeline on next month's data with one click.



## Scalability

Spreadsheets crash at 1M rows. Python (and Pandas) processes millions of records in seconds without breaking.



# Feature Engineering

Mining new value from raw, existing data  
columns.

# | Creating New Data Value



## Time Extraction

Pulling "Month Name," "Week Number," or "Is\_Weekend" flags from raw date stamps for seasonality analysis.



## Binning & Mapping

Grouping low-frequency categories into an "Other" bucket to simplify complex categorical distributions.



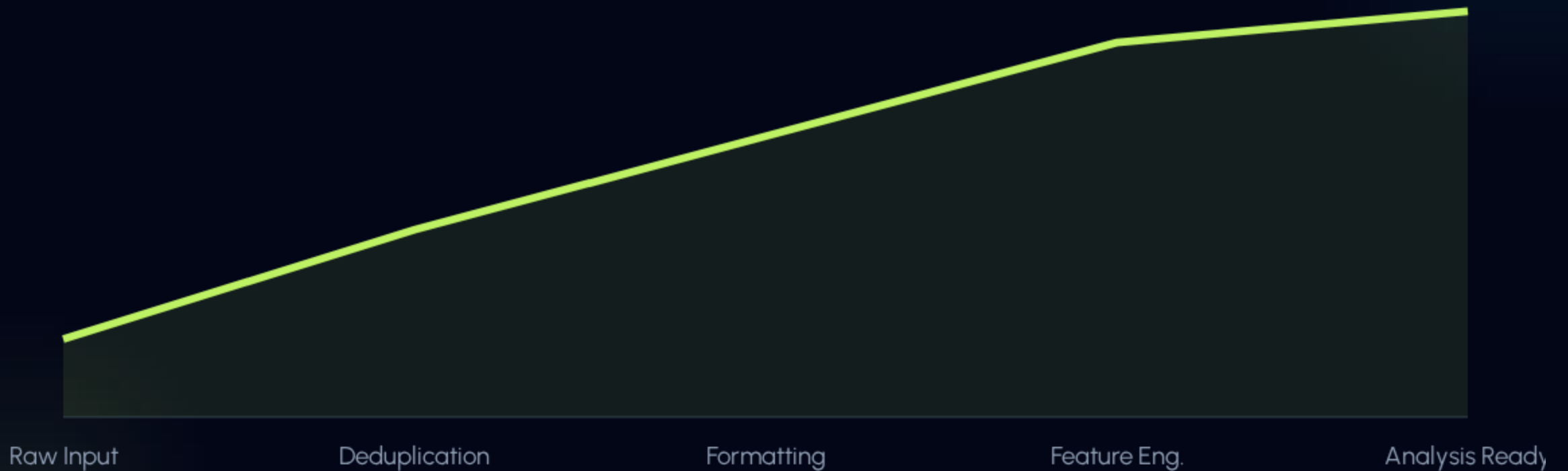
## Calculated Ratios

Synthesizing metrics like **LTV/CAC** or **Profit Margin** from raw revenue and cost columns.

# Exercise: The Marketing Dataset

| Issue Identified     | Raw Data Example            | Cleaning Strategy (AI Instruction)            |
|----------------------|-----------------------------|-----------------------------------------------|
| Format Inconsistency | "Oct-20-23" vs "2023/10/20" | Standardize to pd.to_datetime ISO format.     |
| Numeric Noise        | "\$1,200.50 USD" (String)   | Strip symbols and convert to Float.           |
| Duplication          | ID 1024 (appears twice)     | Drop duplicates based on 'Transaction_ID'.    |
| Empty Fields         | Status: [Empty]             | Impute with 'Unknown' or most frequent value. |

# | Impact of Automated Cleaning



*Visual representation of data usability growth through iterative AI cleaning steps.*

# Ready to Automate?

Master Module 6 by applying these prompts today.

 [ai-data-cleaning.io](https://ai-data-cleaning.io) |  [help@module6.edu](mailto:help@module6.edu)

# | Image Sources



[https://media.istockphoto.com/id/2022774836/photo/coding-concept-digital-background-with-an-abstract-particles-and-parts-of-code-strings.jpg?s=612x612&w=0&k=20&c=\\_Z2ezeHvcByzraT2CMPaid9EAkMI5iZxdwRgXoolBEU=](https://media.istockphoto.com/id/2022774836/photo/coding-concept-digital-background-with-an-abstract-particles-and-parts-of-code-strings.jpg?s=612x612&w=0&k=20&c=_Z2ezeHvcByzraT2CMPaid9EAkMI5iZxdwRgXoolBEU=)

Source: [www.istockphoto.com](https://www.istockphoto.com)