Group Submission — Human-Authored Synthesis

Investigation question:
When importing a messy CSV, what hidden assumptions do import tools make (about missing values, types, dates, and text), and how can those assumptions silently change the dataset before analysis?

What we found (using the provided file):
This dataset looks like a standard CSV, but several columns encode values in ways that import tools cannot safely interpret without guidance. Importing is therefore not a neutral step: the tool must *decide* what missingness looks like, what a number looks like, what a date looks like, and how text is encoded.

Hidden assumption 1: "Missing values are standardized."
Many import workflows assume missingness appears as blank cells or a small set of tokens (often "NA"). In our file, missing values appear in multiple forms. For example, host_response_rate contains blanks (28), "NA" (12), "N/A" (8), a literal period "." (11), and even a single space (11). If we do not declare all of these as missing, we will treat some missing values as valid strings. That distorts summaries and can create misleading categories (e.g., "." becomes a level).

Hidden assumption 2: "Dates have one consistent format."
Date parsing often fails silently when formats mix. In our dataset, host_since has two formats: 480 ISO dates (YYYY-MM-DD) and 84 US-style dates (M/D/YYYY) among 564 non-missing entries. last_scraped shows the same pattern (510 ISO, 90 US). If we parse using only one format, the other format may become NA, reducing sample size and potentially biasing downstream analysis (e.g., if older hosts are more likely to appear in one format).

Hidden assumption 3: "Numeric-looking fields are numeric."
Two key fields are not numeric at all in raw form: host_response_rate/host_acceptance_rate are percentages like "99%", and price is currency like "$1,027.00". Converting these requires deliberate parsing and a choice of scale (e.g., 99 vs 0.99). Additionally, price includes commas in 9 values, including an extreme value $40,000.00. Even with correct parsing, we must flag outliers because they can dominate plots and models.

Hidden assumption 4: "Text encoding is harmless."
We observed non-ASCII names (Zoë, María, Ömer, Tuğba, Le Méridien, Swissôtel). If encoding is mishandled, these can become corrupted strings, causing subtle join/deduplication problems and poor data quality.

Recommended import-and-validate routine (generalizable):

1. First import risky columns as text to inspect tokens and formats.

2. Explicitly define all missing tokens you observe (here: blank, NA, N/A, “.”, single space).
3. Explicitly parse dates with a multi-format strategy and report failures.
4. Convert percent and currency with safe parsers; document scaling choices.
5. Produce an “import report”: row/column counts, column classes, unique-value checks, and a count of parsing failures or suspicious values.

Bottom line:
Even if an AI’s advice is “correct,” trust comes from **evidence**—checking assumptions against the actual file—not from confidence.