Entity Resolution and Data Quality Assessment for Supplier Database Digitalisation

Data Analyst Challenge- Veridion

Dataset: `presales_data_sample.csv` **Inputs:** 592 suppliers **Candidates:** 2,951 rows

February 2026

Abstract. For this project, I worked with a sample of 592 suppliers from a manufacturing company's procurement database. My goal was to clean up their messy supplier data by matching each entry to the right company in our system.
For each supplier in their list, our entity resolution engine gave me up to 5 possible matches. I had to pick the best one by looking at a few things: how similar the company names were (using normalized name similarity), whether the locations matched up (geographic agreement), and how close the cities were to each other. I weighted these factors to decide which candidate was the best fit.
After going through all the data, I was able to resolve about 89.7% of the suppliers with either HIGH or MEDIUM confidence. The remaining 10.3% were trickier - maybe the names were too different or the locations didn't match well - so I flagged those for a human analyst to review manually.

1. The Problem

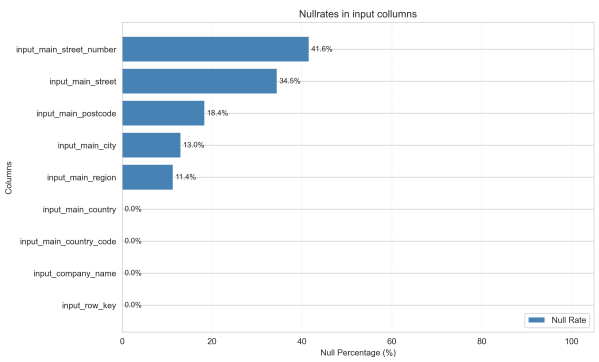
A large manufacturing company's Procurement department is kicking off a digitalization journey. Their category managers have hit a wall – they can't properly analyze spend because their supplier database is cluttered with messy, duplicate, and outdated entries. Meanwhile, leadership is pushing hard for a clear cost-saving strategy for next year. On top of that, there's interest in exploring sustainability in the supply chain, but they just don't have the resources to prioritize it right now.

2. Dataset EDA: First Look

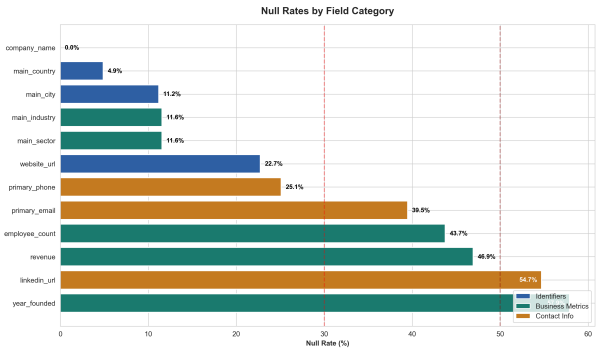
The file contains 2,951 rows: up to five candidates per input, giving 592 unique input companies. Four inputs received fewer than five candidates, meaning the retrieval engine found limited plausible matches for those.

2.1 Candidate field completeness

I split the 76 columns into two groups: the 9 that came from the client and the 66 that Veridion returned for each candidate. This split is useful later when auditing quality separately for each side.



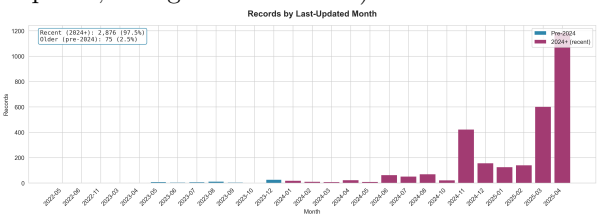
Null Rates in input fields are quite good, except for geographical records,as expected.



Null rates on key business fields: revenue 47 %, email 40 %, employee count 44 %, website 23 %, city 11 %. Identity fields (company name, country code) are quite complete.

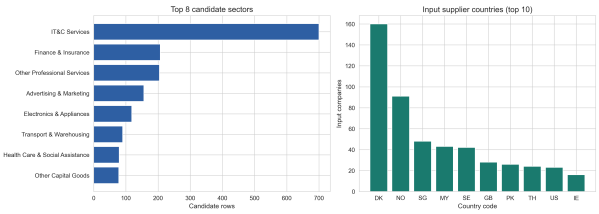
2.2 Freshness

Data freshness is strong: 97.5 % of records were updated in 2024 or later. I believe stale, old data should be automatically flagged for review (suppliers disappear, get acquired, change addresses etc.)



2.3 Sector Distribution and geographical spread

I have also represented the main sectors and locations of companies.



3. Scoring Model

3.1 Prerequisites

To ensure robustness, I take some precautionary measures regarding problems that new data usually has:

- I ensure all nulls have the same format (this runs on every string column before anything else)
- I standardize country codes valid ISO 3166-1 and implement a generic corrector dictionary
- Normalize company names, strip them of punctuation and legal suffixes found on Wikipedia ("Accenture Services AS" and "ACCENTURE" should match)
- normalize city names, handle common english abbreviations

3.2 Features

Three most indicative columns are selected for each input/candidate pair:

Name Similarity $s_n \in [0, 100]$. I used token-sort ratio rather than plain string similarity. The idea is to split both names into individual words, sort those words alphabetically, then compare the sorted versions. This makes the metric order-invariant: "Floreasca Media SRL" and "SRL Floreasca Media" become the same string after sorting, so they score 100

Country similarity δ_c**.** Exact ISO 3166-1 alpha-2 match gives $+30$; mismatch gives -20, or -10 when $s_n > 85$ (to avoid penalising multinationals whose legal entity is registered in a different jurisdiction than their operating address).

City proximity $\delta_k \in \{0, +10\}$**.**

3.3 Composite score and confidence tiers

$$S = w_n \cdot s_n \; + \; \delta_c \; + \; \delta_k \tag{1}$$

Tiers are assigned as: **HIGH** if $S \geq 75$; **MEDIUM** if $50 \leq S < 75$; **LOW** otherwise. Additionally, any record with $> 60\,\%$ of key fields empty, or last updated before 2023, is escalated to `REVIEW_NEEDED` regardless of score (see **Table 1:** Entity Resolution Scoring Scenarios).

4. Weight Selection

4.1 Educated guess for name weight

The weight w_n controls how strongly the similarity of the name can pull a candidate across a tier boundary. Obviously, $100 * w_n$ (perfect name match) should be more than $\delta_c + \delta_k, \forall \delta_c, \delta_k$. Moreover, I want to have a MEDIUM score for perfect name similarity, in order to get some wiggle room depending on location and other rogue factors. Name similarity alone is not enough to auto-match, because generic names like "Global Solutions" can score high against completely unrelated companies.
I decided to make $w_n = 0.6$ (60 is medium, more than half)

4.2 Deriving country penalties

Country match should only reach HIGH if name also agrees. So it should be less than 50, but more than 25 = 100-75. It should still be much more important than city, so with the 40 points left until max confidence score 100. I chose a round number of +30.

4.2.1 The asymmetric country penalty

A symmetric ± 20 rule was also tested. This caused multinationals whose legal entity is registered in a holding jurisdiction (e.g., Ireland for "Accenture", Mauritius for regional offices) to be under-ranked even when the name similarity was near-perfect. The reduced penalty (-10 when $s_n > 85$) recovers those cases without materially increasing false positives: the relaxed rule only fires when an independent, strong name-similarity signal already corroborates the match.

The graph and heatmap below both help visualize this scoring process.

4.3 Deriving city penalties

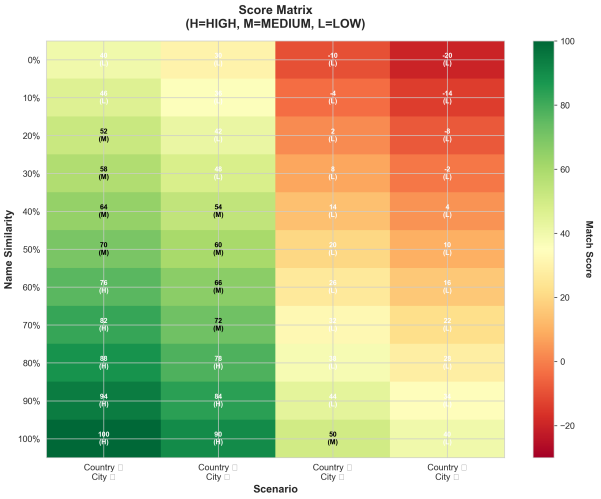
The remaining 10 if the city matches, whyle nothing is penalized if not.

I chose the numbers that best work with normal thresholds in the range [0, 100]. I noticed that the minimum final

Table 1: Entity Resolution Scoring Scenarios

Scenario	Name	Country	City	Score	Tier	Decision
Scenario	Name	Country	City	Score	Tier	Decision
Perfect match	100%	✓	✓	100	HIGH	Auto-accept
Strong match	85%	✓	×	81	HIGH	Auto-accept
Threshold case	75%	✓	×	75	HIGH	Auto-accept
Good match	70%	✓	×	72	MEDIUM	Review
Subsidiary (perfect)	100%	×	✓	50	MEDIUM	Review
Subsidiary (strong)	90%	×	✓	44	LOW	Research
Weak match	60%	✓	×	66	MEDIUM	Review
Very weak	50%	✓	×	60	MEDIUM	Review
Wrong company	60%	×	×	16	LOW	Reject

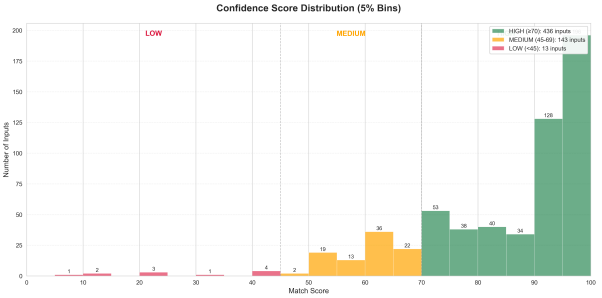
score is in the range [-20, 100], but that makes my formula even more strict – and after pulling the results, still a very large proportion of entities will be resolved, so that does not bother me.

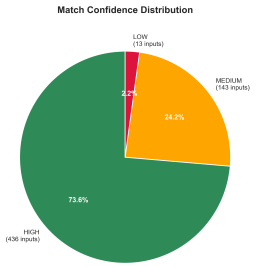


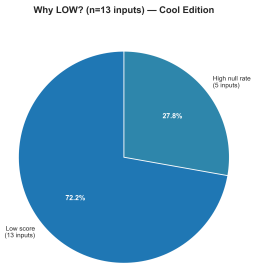
5. Results

Score statistics: mean = 84.4, median = 90, min = 6.7, max = 100.

The distribution is heavily right-skewed: most inputs sit in the 85–100 range where a strong name match combines with a country agreement. The smaller MEDIUM cluster (60–74) consists mainly of multinationals where the candidate is registered in a holding jurisdiction. Only 13 inputs fall below 50.







5.1 Review-needed breakdown

52 of the 61 flagged cases are sparse-record escalations rather than matching failures. The pipeline correctly identified the entity; we simply have limited data for those companies. This is an important distinction: the match is likely correct, but the enriched record would add little value without more sourcing

6. Extra Data Quality Observations for Client

Revenue. 47 % missing. Of populated records, 76 % are modelled (estimated) rather than extracted. Revenue should be segmented by `revenue_type` and labelled accordingly for category managers.

Email and phone. 40 % and 25 % missing respectively. LinkedIn coverage is better and is a more reliable B2B contact channel.

Country mismatches. 29 % of candidate rows show a different country code from the input. Three causes: holding entity registration (legitimate), surface-name coincidence (false candidate), and client data errors. The scoring model handles the first case; the other two end up in the review queue.

7. Final Thoughts

7.1 Possible future improvements

-Additional signals MOre signals could be incorporated and tested for a finer search

-Better weight selection, less 'ansatz' Right now the weights are hand-tuned with a logical justification, but there's no feedback loop.

-Input deduplication: The pipeline resolves each input independently, but the client's actual problem is that their supplier database has duplicates. A blocking on the input names before matching would directly address that — right now I only flagged the duplicate I found, I didn't solve it.

-Sustainability enrichment: The client mentioned interest in supply chain sustainability but didn't have resources to prioritize it.

-Production ready The current pipeline runs on a static CSV. If this were to go live: incremental processing (only new or changed suppliers, not the full database every time), a simple drift monitor (if the mean confidence score drops week-over-week, something changed badly and it should alert someone), and ideally a thin API wrapper so procurement systems can call it directly rather than running a notebook manually.

7.2 Conclusion

The pipeline resolves 89.7 % of 592 input suppliers with HIGH or MEDIUM confidence (mean score 84.4, median 90). The 10.3 % flagged for review are driven by record sparsity rather than matching failures. The weight $w_n = 0.60$ is justified both analytically — it is the midpoint of the valid range [0.45, 0.78) — and factually: it minimises both same-country demotions and cross-border false promotions relative to adjacent values.

All analysis in Python 3.11; Full resolution notebook in `Veridion Challenge FINALFINAL + cleaned.ipynb`.***NOTE:*** *All the code for plots in the notebook, plus the LaTEX scripts for tables are AI generated.*